Troubleshooting Cisco Nexus Switches and NX-OS

Vinit Jain, CCIE® No. 22854
Brad Edgeworth, CCIE® No. 31574
Richard Furr, CCIE® No. 9173

ciscopress.com

Troubleshooting Cisco Nexus Switches and NX-OS

Vinit Jain, CCIE No. 22854

Brad Edgeworth, CCIE No. 31574

Richard Furr, CCIE No. 9173

Cisco Press

800 East 96th Street

Indianapolis, Indiana 46240 USA

00c1.5c00.0011 (NX-2's VLAN 100 interface) and cannot forward the packet toward NX-1. The packet is dropped because packets received on a vPC member port cannot be forwarded across the peer link, as a loop-prevention mechanism.

Enabling a vPC peer-gateway on NX-2 and NX-3 allows NX-3 to route packets destined for NX-2's MAC addresses, and vice versa. The vPC peer-gateway feature is enabled with the command **peer-gateway** under the vPC domain configuration. The vPC peer-gateway functionality is verified with the **show vpc** command. Example 5-27 demonstrates the configuration and verification of the peer-gateway feature.

Example 5-27 *Configuration and Verification of vPC Peer-Gateway*

```
NX-2(config)# vpc domain 100
NX-2(config-vpc-domain)# peer-gateway
```

```
NX-2# show vpc
! Output omitted for brevity
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link

vPC domain id                     : 100
..
vPC role                          : primary
Number of Cs configured           : 1
Peer Gateway                      : Enabled
```

Note In addition, NX-OS automatically disables IP redirects on SVIs where the VLAN is enabled on a vPC trunk link.

Note Packets that are forwarded by the peer-gateway feature have their time to live (TTL) decremented. Packets carrying a TTL of 1 thus might get dropped in transit because of TTL expiration.

vPC ARP Synchronization

The previous section demonstrated how traffic becomes asymmetric depending on the hash calculated by the device with the regular port-channel interface. During normal operations, the device builds the Address Resolution Protocol (ARP) table (IP to MAC) in normal manner, but this is not fast enough when a node comes online after a reload. NX-OS includes an ARP synchronization feature that keeps the table synchronized between both vPC peers, thereby drastically speeding up the process for a vPC peer that was just restarted.

ARP synchronization is enabled with the command **ip arp synchronize** under the vPC domain configuration. Example 5-28 demonstrates enabling ARP synchronization on NX-2.

Example 5-28 *Enabling vPC ARP Synchronization*

```
NX-2# conf t
Enter configuration commands, one per line. End with CNTL/Z.
NX-2(config)# vpc domain 100
NX-2(config-vpc-domain)# ip arp synchronize
```

Backup Layer 3 Routing

Give special consideration to network designs when Nexus switches act as a gateway while providing vPC to hosts on that network segment. Certain failure scenarios can occur, or a Type 1 consistency check might trigger and shut down vPC ports.

Figure 5-7 demonstrates a simple topology in which NX-2 and NX-3 have an SVI interface for VLAN 200 that acts a gateway for the web server. NX-2, NX-3, and R4 are all running OSPF so that NX-2 and NX-3 can forward packets to R4. NX-3 is the operational primary Nexus switch.

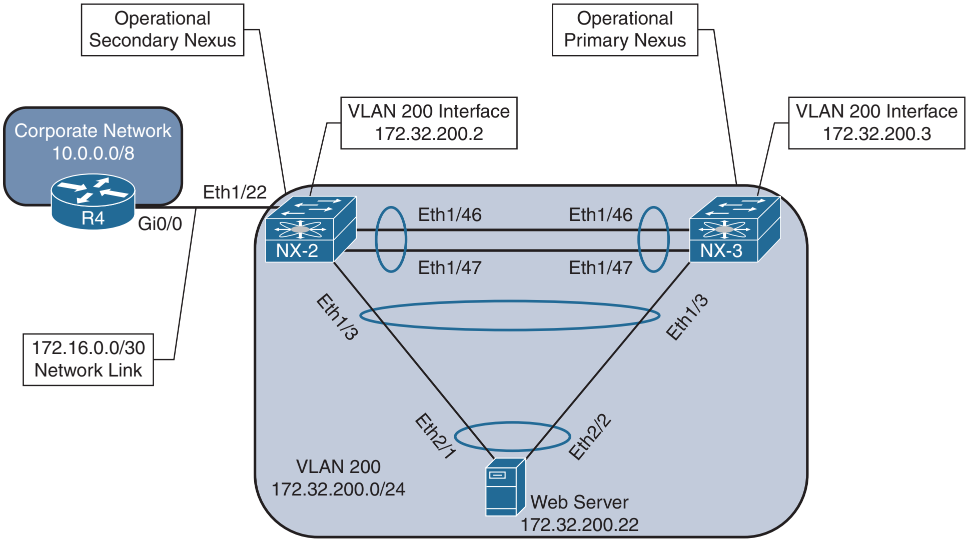
Operational Secondary Nexus
Operational Primary Nexus
VLAN 200 Interface 172.32.200.2
VLAN 200 Interface 172.32.200.3
Corporate Network 10.0.0.0/8
R4
Gi0/0
Eth1/22
NX-2
NX-3
Eth1/46
Eth1/46
Eth1/47
Eth1/47
Eth1/3
Eth1/3
172.16.0.0/30 Network Link
Eth2/1
Eth2/2
VLAN 200 172.32.200.0/24
Web Server 172.32.200.22

Figure 5-7 *Bad Layer 3 Routing Design*

If the vPC peer link is broken (physically or through an accidental change that triggers a Type 1 consistency checker error), NX-2 suspends activity on its vPC member port and shuts down the SVI for VLAN 200. NX-3 drops its routing protocol adjacency with NX-2 and then cannot provide connectivity to the corporate network for the web server. Any packets from the web server for the corporate network received by NX-3 are dropped.

This scenario is overcome by deploying a dedicated L3 connection between vPC peers. These are either individual links or an L3 port-channel interface.

Note Remember that the vPC peer link does not support the transmission of routing protocols as transient traffic. For example, suppose that Eth1/22 on NX-2 is a switch port that belongs to VLAN 200 and R4's Gi0/0 interface is configured with the IP address of 172.32.200.5. R4 pings NX-3, but it does not establish an OSPF adjacency with NX-3 because the OSPF packets are not transmitted across the vPC peer link. This is resolved by deploying the second solution listed previously.

Layer 3 Routing over vPC

vPC interfaces only forward traffic, from an L2 perspective. An IP address cannot be assigned directly to the vPC interface. The vPC devices provide gateway service to downstream devices by assigning IP addresses to the switched virtual interface (SVI) on the vPC devices.

However, vPC functionality was never meant to provide a logical L2 link to be used to form routing protocol adjacencies. However, the release of NX-OS version 7.3 provides the capability for the SVIs to form a routing protocol adjacency using a vPC interface with a router.

Note L3 Routing over vPC is specific only to unicast and does not include support for multicast network traffic.

Figure 5-8 demonstrates the concept in which NX-2 and NX-3 want to exchange routes using OSPF with R4 across the vPC interface. NX-2 and NX-3 enable Layer 3 routing over vPC to establish an Open Shortest Path First (OSPF) neighborship with R4. In essence, this design places NX-2, NX-3, and R4 on the same LAN segment.

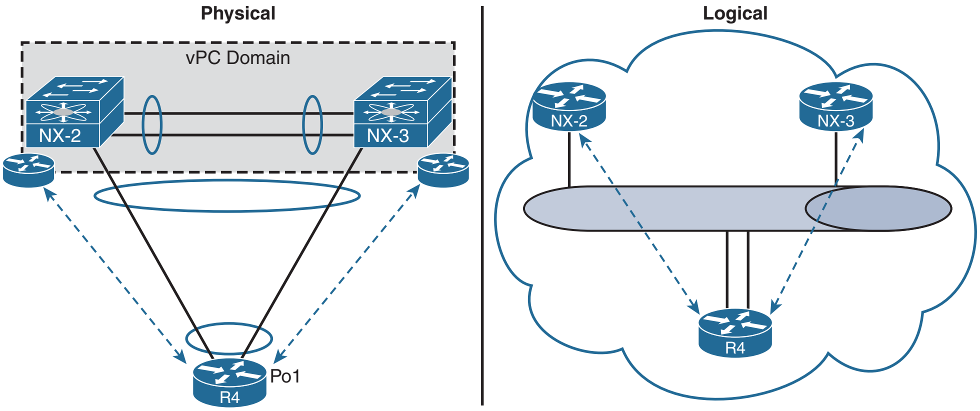

Figure 5-8 *Layer 3 Routing over vPC*

Layer 3 routing over vPC is configured under the vPC domain with the command **layer3 peer-router**. The peer-gateway is enabled when using this feature. The feature is verified with the command **show vpc**.

Example 5-29 demonstrates the configuration and verification of Layer 3 routing over vPC.

Example 5-29 *Configuration and Verification of Layer 3 Routing over vPC*

```
NX-2# configure terminal
Enter configuration commands, one per line. End with CNTL/Z.
NX-2(config)# vpc domain 100
NX-2(config-vpc-domain)# layer3  peer-router
```

```
NX-2# show vpc
! Output omitted for brevity
..
Delay-restore SVI status          : Timer is off.(timeout = 10s)
Operational Layer3 Peer-router    : Enabled
```

Note If vPC peering is not being established or vPC inconsistencies result, collect the **show tech vpc** command output and contact Cisco technical support.

FabricPath

Until recently, all L2 networks traditionally were enabled with STP to build a loop-free topology. However, the STP-based L2 network design introduces some limitations. One limitation is the inability of STP to leverage parallel forwarding paths. STP blocks additional paths, forcing the traffic to take only one path as the STP forms a forwarding tree rooted at a single device, even though redundant paths are physically available. Other limitations include the following:

- STP convergence is disruptive.
- MAC address tables don't scale.
- The tree topology provides limited bandwidth.
- The tree topology introduces suboptimal paths.
- Host flooding impacts the whole network.
- Local problems have a network-wide impact, making troubleshooting difficult.

To overcome these challenges, vPC was introduced in 2008. An Ethernet device then could connect simultaneously to two discrete Nexus switches while bundling these links into a logical port-channel. vPC provided users with active-active forwarding paths, thus overcoming the limitation of STP. Still, although vPC overcame most of the challenges, others remained. For example, no provision was made for adding third or fourth aggregation layer switches to further increase the density or bandwidth on the downstream switch. In addition, vPC doesn't overcome the traditional STP design limitation of extending the VLANs.

The Cisco FabricPath feature provides a foundation for building a simplified, scalable, and multipath-enabled L2 fabric. From the control plane perspective, FabricPath uses a shortest path first (SPF)–based routing protocol, which helps with best path selection to reach a destination within the FabricPath domain. It uses the L2 IS-IS protocol, which provides all IS-IS capabilities for handling unicast, broadcast, and multicast packets. Enabling a separate process for the L2 IS-IS is not needed; this is automatically enabled on the FabricPath-enabled interfaces.

FabricPath provides Layer 3 routing benefits to flexible L2 bridged Ethernet networks. It provides the following benefits of both routing and switching domains:

- Routing
  - Multipathing (ECMP), with up to 256 links active between any two devices
  - Fast convergence
  - High scalability
- Switching
  - Easy configuration
  - Plug and Play
  - Provision flexibility

Because the FabricPath core runs on L2 IS-IS, no STP is enabled between the spine and the leaf nodes, thus providing reliable L2 any-to-any connectivity. A single MAC address lookup at the ingress edge device identifies the exit port across the fabric. The traffic is then switched using the shortest path available.

FabricPath-based design allows hosts to leverage the benefit of multiple active Layer 3 default gateways, as Figure 5-9 shows. The hosts see a single default gateway. The fabric provides forwarding toward the active default gateways transparently and simultaneously, thus extending the multipathing from inside the fabric to the Layer 3 domain outside the fabric.

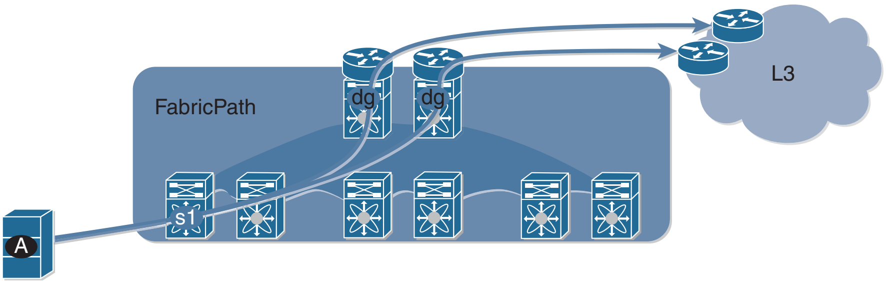

Figure 5-9 *Access to Multiple Active Default Gateways*

The fabric also is used to extend Layer 3 networks. An arbitrary number of routed interfaces can be created at the edge or within the fabric. The attached Layer 3 devices peer with those interfaces, thus providing a seamless Layer 3 network integration.

FabricPath Terminologies and Components

Before understanding the packet flow within the FabricPath (FP)–enabled network, it is important to understand the various terminologies and components that collectively form the FabricPath architecture. Figure 5-10 examines a standard FabricPath-enabled spine-leaf topology, also known as the Clos fabric. The leaf or edge switches in the topology have two different interfaces:

- FabricPath (FP) core ports
- Classical Ethernet (CE) edge ports

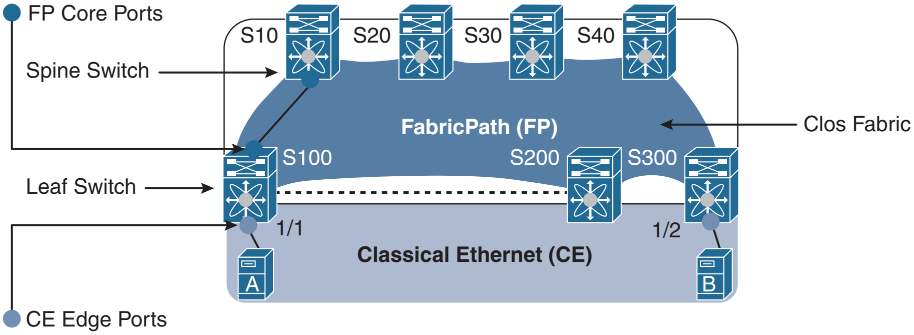

Figure 5-10 *FabricPath and Clos Fabric*

The FP core ports provide connectivity to the spine and are FabricPath-enabled interfaces. The FP core network is used to perform the following functions:

- Send and receive FP frames
- Avoid STP, require no MAC learning, and require no MAC address table maintained by FP Core ports
- Decide the best path by using a routing table computed by IS-IS

The CE edge ports are regular trunk or access ports that provide connectivity to the hosts or other classical switches. The CE ports perform the following functions:

- Send and receive regular Ethernet frames
- Run STP, perform MAC address learning, and maintain a MAC address table

The FP edge device maintains the association of MAC addresses and switch-IDs (which IS-IS automatically assigns to all switches). FP also introduces a new data plane encapsulation by adding a 16-byte FP frame on top of the classical Ethernet header. Figure 5-11