Data Center Switching

Deploying IPv4/IPv6 in Financial Enterprise Networks using Nexus 7000/7700

Cisco Validated Profile

April 2017

Contents

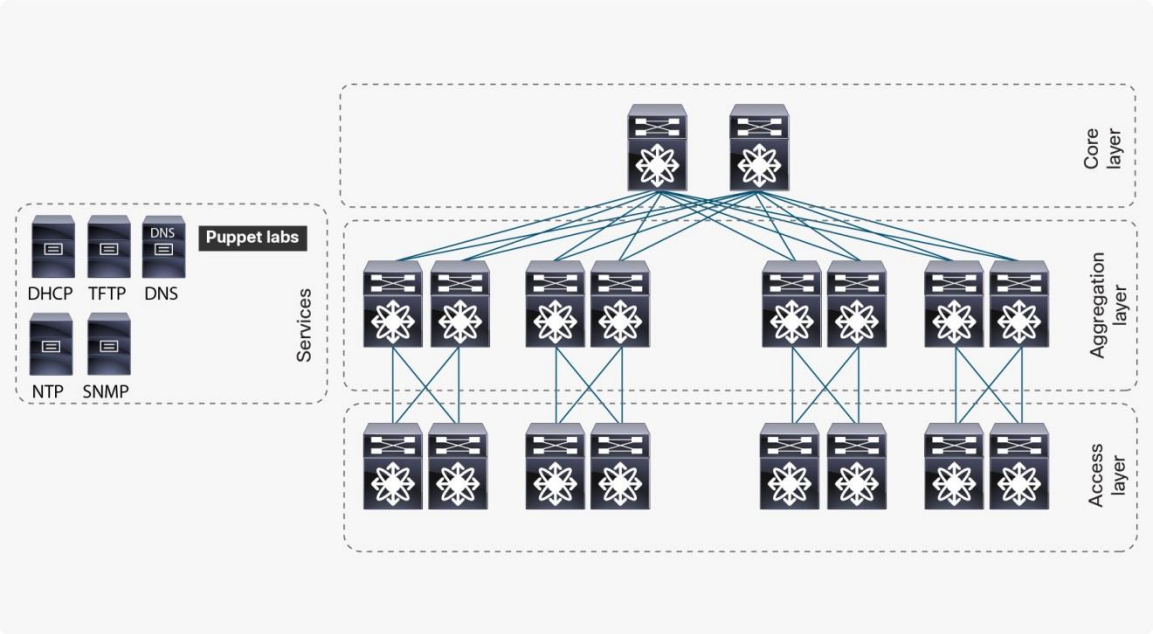
Profile Introduction **3**
Core Layer Design Overview 3
Aggregation Layer Design Overview 3
Access Layer Design Overview 3
Validated Network Profile **4**
Core Layer Detailed Deployment 4
Aggregation Layer Detailed Deployment 5
Access Layer Detailed Deployment 6
Hardware and Software Matrix 6
Scale 6
Operation **8**
Network Management 8
Validated Traffic Profile 9
Maintenance and Software Upgrade 10
Network Design Best Practices **10**

Profile Introduction

The Cisco Nexus line of data center hardware and software products must pass Cisco's comprehensive quality assurance process, which includes a multistage approach comprising extensive unit test, feature test, and system-level test. Each successive stage in the process adds increasingly higher levels of complexity in a multidimensional mix of features and topologies.

This document describes the Cisco Validated Profile for Financial Enterprise deployments with a typical 3-tiers network. The Nexus 7000/7700 was placed to act as access layer node (customer edge, or **CE**), as aggregation layer node (provider edge, or **PE**), and also as the Nexus 7700 as core layer node (P).

Figure 1. Logical three-tiers topology



Core Layer Design Overview

Borderline Gateway Protocol (BGP) allows the core layer to properly interconnect the different Sites for both IPv4 and IPv6 services. All sites belong to the same Autonomous System (AS), so in order to minimize the number of BGP sessions, two core routers (P1 and P2) have been configured to be Route-Reflectors (RRs) serving the whole AS. The other P routers and all of the PEs hence have a BGP connection to both RRs for both IPv4 and IPv6 Address-families (AFs).

Aggregation Layer Design Overview

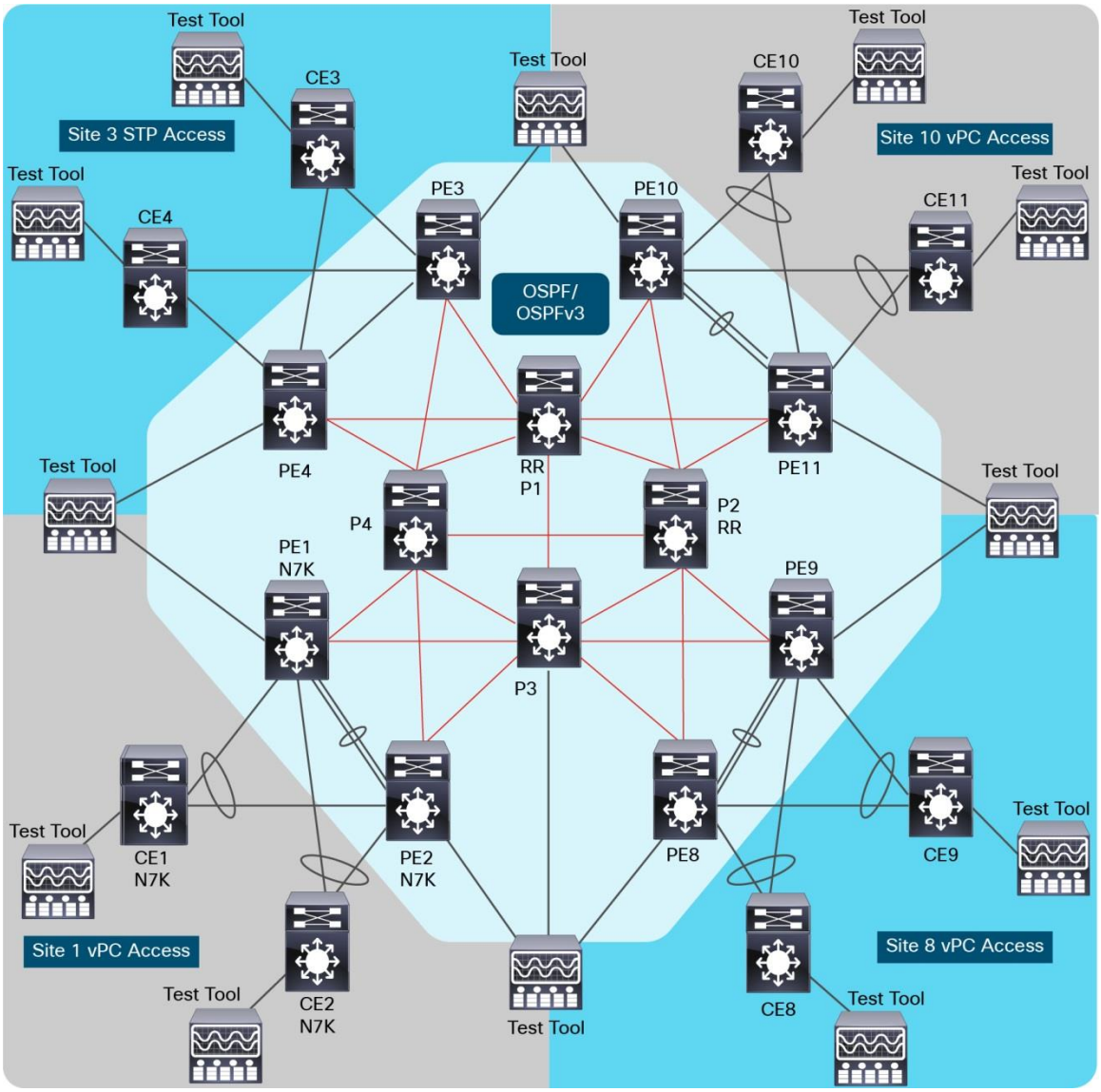
The aggregation layer allows the combination of different access layers on each site to communicate with the other sites throughout the core network, guaranteeing both IPv4 and IPv6 services.

Access Layer Design Overview

In the Financial Enterprise IPv4/IPv6 network, the access layer (CEs) is formed by L2 switches. The use of VLAN trunking in the access layer allows providing both security and segregation, whereas LACP is deployed as port-aggregation protocol to increase bandwidth and connection resiliency.

Validated Network Profile

Figure 2. Financial enterprise IPv4/IPv6 system test topology



In this diagram, each color represents a different N77 or N7k chassis in which the solid-lined chassis is the default Virtual Device Context (VDC) and the dotted-lined chassis are its non-default VDCs.

Core Layer Detailed Deployment

Interior Gateway Protocols (IGPs) guarantee L3 reachability among all the interfaces used by BGP for its peering, as well as propagate the Protocol-Independent Multicast (PIM) Rendezvous Point (RP) information across the whole AS.

Nonstop Forwarding (NSF) is a high-availability feature on modular switches running NX-OS with a redundant supervisor. On the Nexus 7000, data packets are forwarded by the hardware forwarding engines on the modules. These engines are programmed with information learned from the routing control plane running on the supervisors. If the active supervisor were to fail, the forwarding tables on the modules are preserved. All interface states are also preserved while the standby supervisor takes over active control of the system. This high-availability system prevents any drop in traffic during the failure of the active control plane.

BGP Graceful restart (GR) is a BGP feature that prevents disruption to the control and data plane. It allows for the graceful recovery of BGP sessions after a peer has failed. When combined with the NSF feature, any GR capable peers connected to a switch going through supervisor switchover will continue to forward traffic seamlessly.

NSF and GR for BGP are enabled by default on NX-OS.

From edge/core switches to public cloud, one PE has enabled eBGP to establish peering between data center autonomous systems and public cloud autonomous systems to exchange routing updates. BGP policies have been applied to the eBGP peering configuration to control route exchanges between the eBGP peers.

PIM is the multicast routing protocol widely deployed by financial enterprise customers. In this profile, PIM runs in ASM mode and Anycast RP with static RP is the chosen method to learn the RP information. Multicast Source Discovery Protocol (MSDP) is used to synchronize the PIM (S, G) entries among the different RP announcing routers.

Aggregation Layer Detailed Deployment

On the 2 PEs of each site, the Switched Virtual Interfaces (SVIs) are configured in dual-stack mode along with First-Hop Redundancy Protocol (FHRP) to provide a default gateway for all the IPv4 and IPv6 hosts directly connected to both PEs' orphan ports as well as the CEs. The configured dual-stack FHRP protocols are Hot Standby Router Protocol (HSRP) and Virtual Router Redundancy Protocol (VRRPv3).

All of the IPv4 and IPv6 networks are injected into internal BGP (iBGP) through either direct injection (with the BGP network statement) or via controlled redistribution.

PIM is enabled with Static RP pointing to the Anycast RP address configured in the core routers. All of the connections toward the core network are on M3 modules on both Site 3 and 10. On Site 8, the connections toward the core network are on F3 modules.

As for the connections facing the access layer, the interfaces are on an M2 module on Site 1, an M3 module on Site 3, an F3 module on Site 8, and an M3 module on Site 10. Given these different hardware combinations, it has been possible to verify the unicast and multicast data-path in the following VDC types: pure intra-module M3, pure inter-modules M3, mixed M3-F3, and the newly introduced M2-M3 (called **M2-M3 interop mode** specific for N7k chassis only). In particular, in a mixed M2-M3 VDC type, all of the M2 modules have to be programmed to be in interop mode to ensure that the M2 header is compatible with the M3 module(s). The first time this operation is performed, it causes the affected M2 modules to reload.

The aggregation layer on the different sites is designed differently based on the technology deployed and the hardware combinations. There are mainly two different technologies covered in the Financial Enterprise solution network:

- The classic Ethernet access network based on the spanning-tree protocol (RPVST) configured on Site 3
- The vPC access network used in the other three sites following the vPC Best Practice guidelines

In particular, on Site 1 the new M2-M3 VDC type has been tested on the N7k chassis. On Sites 3 and 10, there is a pure M3 VDC type, and on Site 8 the mixed M3-F3 VDC type.

Access Layer Detailed Deployment

On Site 1, the CEs are placed on M2 modules.

On Sites 3, 8 and 10, the CEs are placed on different variations of F3 modules.

Hardware and Software Matrix

Table 1. Hardware profile summary

Tier/layer	Hardware	Release
Core Layer	Chassis: N77-C7706 Supervisor: N77-SUP2E (2) I/O Module: N77-M324FQ-25L (1) I/O Module: N77-M348XP-32 (1) I/O Module: N77-F324FQ-25 (1)	NX-OS 8.0(1)
Aggregation Layer/ Access Layer - N77 Site 3 & 8	Chassis: N77-C7718 Supervisor: N77-SUP2E (2) I/O Module: N77-F324FQ-25 (2) I/O Module: N77-F348XP-23 (1) I/O Module: N77-M324FQ-25L (1)	NX-OS 8.0(1)
Aggregation Layer/ Access Layer - N77 Site 10	Chassis: N77-C7710 Supervisor: N77-SUP2E (2) I/O Module: N77-F324FQ-25 (1) I/O Module: N77-F348XP-23 (1) I/O Module: N77-M324FQ-25L (1) I/O Module: N77-M348XP-23L (1)	NX-OS 8.0(1)
Aggregation Layer/ Access Layer - N7k Site 1	Chassis: N7K-C7010 Supervisor: N7k-SUP2E (2) I/O Module: N7K-M348XP-25L (1) I/O Module: N7K-M224XP-23L (2)	NX-OS 8.0(1)

Scale

Table 2. IPv4/IPv6 core scale (N77)

Feature	Scale
OSPF neighbors	7
OSPF routes	20
OSPF paths	29
OSPF Timer Throttle	20-50-500
BGP Route Reflectors	2
iBGP IPv4 neighbors	11
iBGP IPv4 prefixes	4000
iBGP IPv4 paths	8000
iBGP IPv6 neighbors	11
iBGP IPv6 prefixes	4000
iBGP IPv6 paths	8000
eBGP IPv4 neighbors	0

Feature	Scale
eBGP IPv4 prefixes	80000 (portion of internet feed)
eBGP IPv4 paths	80000
LACP port-channels	8
LACP port-channel members	1 or 4
BFD sessions	10
MHBFD sessions	22
PIM neighbors	7
PIM (*,G)	1000
PIM Sources per group	16
PIM OIFs	3
PIM RP advertisement	Static with Anycast RP
MSDP mesh-groups	1
MSDP neighbors	3
MSDP SA-cache entries	12000
Number of VRFs	1

Table 3. IPv4/IPv6 aggregation layer scale (7000/7700)

Feature	Scale
VLANs	1001
SVIs	1001
HSRPv4	500
HSRPv6	500
VRRPv3 AFv4	500
VRRPv3 AFv6	0
Host per subnet	12
ARP entries per subnet	30
Virtual Ports (Rapid PVST)	4000
vPC legs	40
LACP port-channels	40
LACP port-channel members	1 or 2
BFD sessions	1003
MHBFD sessions	4
OSPF neighbors	4
OSPF routes	32
OSPF paths	48
OSPF Timer Throttle	20-50-500
OSPF Passive Interfaces	1001
iBGP IPv4 neighbors	3
iBGP IPv4 prefixes	44000
iBGP IPv4 paths	48000
iBGP IPv6 neighbors	4
iBGP IPv6 prefixes	3000
iBGP IPv6 paths	6000

Feature	Scale
eBGP IPv4 neighbors	1
eBGP IPv4 prefixes	40000 (subset of internet feed)
eBGP IPv4 paths	40000
ECMP	4
PIM neighbors	1007
PIM (*,G)	1000 (250 per Site)
PIM Sources per group	16
PIM OIFs	10 for each PE's orphan port
PIM RP Protocol	Static
IGMP Groups	1000
IGMP Snooping Entries	10000
L2 total mroutes	11003
L2 (*,G) mroutes	10000
PIM OMF routes	1003
Number of VRFs	3 globally [100 limited to Site 10 only]

Table 4. Access layer scale (7000/7700)

Feature	Scale
VLANs/CE	500
Virtual Ports (Rapid PVST)	2000
LACP port-channels/CE	20
LACP port-channel members	2
IGMP Snooping Entries/CE	5000
L2 total mroutes	5500
L2 (*,G) mroutes	5000
PIM OMF routes	500
MAC Address/CE	8000

It is important to mention that this unicast scale is well-below the actual supported N7k TCAM capacity.

Operation

Network Management

In the Financial Enterprise solution testbed, network management is limited to monitoring a few network parameters and to enabling all the nodes to send specific simple network management protocol (SNMP) traps. Generated traps are therefore directed toward an external collector. Periodically, an SNMP walk is performed to gather critical information about the network status and statistics.

System Provisioning through TFTP/DHCP and Puppet

When a N77/N7k system reboot without a startup-configuration file (for instance after issuing a write erase and a reload sequence), the Power on Auto Provisioning (PoAP) process is triggered to initiate the system provisioning process. The provisioning process can be done manually on the node's console in an interactive way or it can be done automatically by either Cisco Data Center Network Manager (DCNM) or through the use of TFTP/DHCP services. It allows the network administrator to automatically download images, configuration files, and scripts to each node in order to fully automate the provisioning process for each node based on a unique chassis identifier

(which can be the chassis S/N, the MAC address, the chassis name, the IP address, etc.). In essence, each chassis is then associated to its own template, which also includes the proper configuration script.

In the Financial Enterprise network, a Python script is used to install the virtual container on the N77/N7k node in order to subsequently install the Puppet Agent.

The Puppet Agent then securely handshakes with the Puppet Master and transfers the correct template(s) to the node itself to complete the provisioning process.

Table 5. Cisco PuppetLabs manifests

Node	PoAPmanifestname
N77 Core Node	Cisco PuppetLabs BGP, OSPF manifests
N77 Aggregation Node	Cisco PuppetLabs BGP, OSPF and interface-VLAN manifests
N7K Aggregation Node	Cisco PuppetLabs BGP, OSPF and interface-VLAN manifests
N77 Access Node	Cisco PuppetLabs VLAN manifest
N7k Access Node	Cisco PuppetLabs VLAN manifest

It is also possible to control and manage the nodes through Puppet, in order to perform network updates and/ or network upgrades guaranteeing ease-of-use, security, and consistency throughout the whole network from a centralized network resource, in this case the Puppet Master.

Validated Traffic Profile

Traffic has been simulated by traffic generating tools to both analyze traffic convergence upon network disruptions or network upgrades and to reach the proper L2 and L3 scales.

Intra-Site Unicast Traffic

Portions of overall IPv4 and IPv6 unicast traffic is confined within each site. In this case, simulated hosts attached to one CE are sending bidirectional traffic to simulated hosts attached to the other CE. This traffic is sent over all configured subnets and is meant to be either routed or switched at the aggregation layer without leaving the site. Some of the simulated hosts are connected to the CE with breakout cables.

Inter-Site Unicast Traffic

Another component of the IPv4 and IPv6 traffic is dedicated to inter-site. Here, simulated hosts attached to one CE are sending bidirectional traffic to simulated hosts attached to the CEs on a different site. This traffic is sent over all configured subnets, and it is meant to be routed across the entire network in the global context.

Multicast Traffic

On each site, multicast sources are attached to both CEs as well as the orphan ports attached to the aggregation layer. In particular, for each local source, there are 16 different sources. On the same subnets, there are also simulated multicast receivers that are sending IGMP joins for the locally sourced multicast traffic. The same receivers are then sending the IGMP joins for the remotely sourced multicast groups. The receivers directly attached to both CEs and PEs are on different subnets in order to scale the multicast replication at the aggregation layer.

Internet Traffic through Site 10

A few simulated hosts in each site, attached to the CEs, are sending IPv4 unicast traffic to some of the prefixes injected by the Internet feed on Site 10.

Maintenance and Software Upgrade

ISSU—In Service Software Upgrade

A software upgrade can be performed in a non-disruptive manner on the Nexus 7000/7700, if redundant Supervisor is present. This procedure is called In-Service Software Upgrade or ISSU. The upgrade will be non-disruptive for connected end-points. For more information, see the Cisco Nexus 7000 Series Software Upgrade and Downgrade Guide.

Software Upgrade through System Reload

A software upgrade can be performed in a disruptive manner on the Nexus 7000/7700, if the network is built with switch level redundancy or a network outage is not of concern (maintenance window). The procedure requires the copying of the kickstart and system images into the bootflash, followed by changing the boot variables and a reload of the switch.

For general information about upgrading the vPC peers, see the Cisco Nexus 7000 Series NX-OS Interfaces Configuration Guide.

If in the N7k chassis there is at least one VDC configured in interop-mode (that is, VDC type M2-M3), then there is a special procedure that must be followed to ensure that the configuration on the M2 module is not lost upon image upgrade. In most of the cases, the saved configuration file must be applied twice. For more information, see the Cisco Nexus 7000 Series Virtual Device Context Configuration Guide.

Network Design Best Practices

Core Layer:

- Deploy BGP RR functionality in the core to reduce the number of iBGP sessions and to improve the control of route advertisements among the iBGP peers; all the PEs eventually peer with both RRs for both IPv4 and IPv6 address families.
- Enable BGP ECMP to optimize bandwidth utilization on all the routed connections between the aggregation PEs and the core nodes. To fully implement iBGP ECMP, enable “additional-path send/receive/selection” in the RR routers as well as “additional-paths receive/install backup” on all the RR clients for both IPv4 and IPv6 address family.
- Deploy the use of BGP templates to reduce the BGP configuration and to group together similar BGP peers that share the same ingress/egress routing policies while reducing the probability of configuration mistakes.
- The use of loopback interfaces increases the resiliency of the iBGP connections.
- Configure BGP as a BFD client to improve network convergence with both SHBFD and MHBFD sessions.
- For MHBFD sessions hosted on a given module, configure “bfd multihop hosting-linecard add module <slot>”.
- Reduce the utilization of TCAM tables whenever possible by configuring route filtering and/or route summarization and/or injecting the default route (especially when the PEs are F3-based).
- Deploy BGP authentication to increase the security of the network.
- Where applicable, use Open Shortest Path First (OSPF)/OSPFv3 network type as point-to-point to minimize the time necessary to fully establish OSPF/OSPFv3 neighbor.

- Configure OSPF/OSPFv3 as a BFD client to improve network convergence. Alternatively, you can deploy aggressive timers for the same purpose.
- OSPF timer throttling can improve network convergence time.
- Deploy OSPF authentication on both interface and area levels to increase the security of the network.
- Deploy Rendezvous-Point functionally on more than one core node using Anycast RP and MSDP to properly sync the (S,G) entries among the different RP candidates.
- The use of MSDP mesh groups might help in reducing the amount of unnecessary SA messages.
- Configure PIM as a bidirectional forwarding detection client to improve network convergence. Alternatively, you can deploy aggressive timers for the same purpose.
- To increase the number of possible connections on the 40 Gb modules, it is possible to employ the interface breakout command that allows to split one 40 Gb port into 4 different 10 Gb ports.
- Use port-aggregation protocols to optimize bandwidth utilization and increase connection resiliency. In this profile, link aggregation control protocol is deployed.

Aggregation Layer:

- Enable iBGP Equal-Cost Multi-Path routing (ECMP) to optimize bandwidth utilization on all the routed connections between the aggregation PEs and the core nodes. To fully implement iBGP ECMP, enable “additional-path send/receive/selection” in the RR routers as well as “additional-paths receive/install backup” on all the RR clients for both IPv4 and IPv6 address family.
- Deploy route injections (through either network statements and/or IGP redistribution to advertise the local network (site routes) to the rest of the network.
- Deploy BGP authentication to increase the security of the network.
- Reduce the utilization of TCAM tables whenever possible when deploying ingress/egress route filtering, route summarization, and/or injecting only the default route from the core routers (especially when the PEs are purely F3 based or F3 mixed with M2 or M3 modules).
- Even though the unicast scale is quite limited in the deployed financial enterprise network, it is recommended that you install the SCALABLE_SERVICES_PKG for the N7k peers containing M2 modules.
- If needed, it is possible to change for each VDC the values of u4routes, u6routes, m4routes and m6routes to properly tune the TCAM allocation. Every time these values are altered and the new configuration saved, the system must either be reloaded or undergo a system switchover.
- Where applicable, use OSPF/OSPFv3 network type as point-to-point to minimize the time necessary to fully establish OSPF/OSPFv3 neighbor.
- Configure OSPF/OSPFv3 as a BFD client to improve network convergence. Alternatively, you can deploy aggressive timers for the same purpose.
- OSPF timer throttling can improve network convergence time.
- Deploy OSPF authentication on both interface and area levels to increase the security of the network.
- Use OSPF passive interfaces for the SVI on the PE peers to avoid having unnecessary OSPF peering.
- Configure FHRP on the SVIs on both PEs (either HSRP for both IPv4 and IPv6 or VRRP for both IPv4 and IPv6 AFs).
- Ensure that FHRP Active/Standby (or Master/Backup) roles are evenly distributed on the two PEs.

- Depending on the SVI scale, the use of aggressive timers might be recommended for FHRP messaging.
- Deploy FHRP authentication on both PEs to increase the security of the network.
- Using vPC, for End-Point connectivity can achieve reduced downtimes. For more information about best practices for improving convergence and availability, see the Design and Configuration Guide.
- vPC technology helps build a loop-free topology by leveraging port-channels from access devices to the vPC domain. A port-channel is seen as a logical link from the spanning- tree's standpoint, so a vPC domain with vPC-attached access devices forms a star topology at L2 (there are no STP-blocked ports in this type of topology). In this case, STP is used as a fail-safe mechanism to protect against any network loops caused by human error (like plugging a loopback cable across the 2 vPC peer device).
- In this profile, the use of Rapid-PVST, which is the default spanning tree protocol on NX-OS, has been chosen due to the limited number of virtual ports. For networks with larger virtual port counts, multiple spanning tree protocol is recommended.
- For mixed VDC type M2-M3 (that is, VDC having port on both M2 and M3 modules), use the command "system interop-mode m2-m3 module <slot>". When first issued, the command will cause the M2 modules to reload even when the system is loaded without a startup-config (that is, after issuing a write erase and reload sequence). Sometimes after image upgrade, it is necessary to apply the saved configuration twice to ensure the configuration is properly pushed onto the M2 modules in interop-mode.
- To increase the number of possible connections on the 40 Gb modules, it is possible to employ the interface breakout command that allows you to split one 40 Gb port into 4 different 10 Gb ports.
- To minimize the impact of both broadcast and multicast packets in the switched network, enable storm-control.
- Use port aggregation protocols to optimize bandwidth utilization and increase connection resiliency. In this profile, LACP is deployed.

Access Layer:

- Use port aggregation protocols to optimize bandwidth utilization and increase connection resiliency. In this profile, LACP is deployed.
- Enable STP portfast edge trunk on all the leaf connections.
- To increase the number of possible connections on the 40 Gb modules, it is possible to employ the interface breakout command that allows you to split one 40 Gb port into 4 different 10 Gb ports.

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 **https://www.cisco.com/go/offices**.

Printed in USA C17-739899-00 11/17