Dell EMC Switch Configuration Guide for iSCSI and Software-Defined Storage

Dell EMC Networking Infrastructure Solutions
November 2017

A Dell EMC Technical White Paper

Revisions

Date	Description	Authors
November 2017	Initial release	Colin King, Gerald Myres

Table of contents

1 Introduction 5
1.1 Dedicated storage networks 5
1.2 Shared leaf-spine networks 5
1.3 Typographical conventions 6
2 Objective 7
2.1 Navigating this document 7
3 Dedicated storage network 9
3.1 Isolated switch pair 10
3.2 Isolated switch pair configuration 11
3.3 Switch pair with switch interconnect 12
3.4 Switch pair with switch interconnect configuration 13
4 Leaf-Spine architecture with Dell EMC spine and leaf 15
4.1 Layer 3 switch configuration 16
4.2 Layer 3 topology protocols 16
4.3 Layer 3 configuration planning 19
4.4 Layer 3 configuration with Dell EMC leaf and spine switches 22
4.5 Layer 2 switch configuration 41
4.6 Layer 2 topology protocols 42
4.7 Layer 2 configuration with Dell EMC leaf and spine switches 43
5 Leaf-Spine architecture with Cisco spine and Dell EMC leaf 55
5.1 Layer 3 switch configuration 56
5.2 Layer 3 topology protocols 56
5.3 Layer 3 configuration planning 59
5.4 Layer 3 with Dell EMC leaf and Cisco Nexus spine switches 62
5.5 Layer 2 switch configuration 73
5.6 Layer 2 topology protocols 73
5.7 Layer 2 with Dell EMC leaf and Cisco Nexus spine switches 75
6 Networking features and guidelines for storage 88
6.1 iSCSI optimization 88
6.2 Link-level flow control 90
6.3 Data Center Bridging 91
6.4 QoS with DSCP 94

6.5 Frame size 96
6.6 Multicast 96
6.7 Storage network connections 96
A Storage topologies 99
A.1 Leaf-Spine with iSCSI storage array topology 99
A.2 Leaf-Spine with software-defined storage topology 100
B Dell EMC Networking switches 101
B.1 Dell EMC Networking switch factory default settings 101
C Overview of leaf-spine architecture 102
C.1 Layer 3 leaf-spine topology 102
C.2 Layer 2 leaf-spine topology 103
C.3 Design considerations 103
D Management network 105
D.1 iDRAC server interface 105
E Validated hardware and operating systems 106
F Technical support and resources 107
G Support and Feedback 108

1 Introduction

Storage networks are constantly evolving. From traditional Fibre Channel to IP-based storage networks, each technology has its place in the data center.

IP-based storage solutions have two main network topologies to choose from based on the technology and administration requirements.

- Dedicated storage network topology
  - iSCSI SAN
- Shared leaf-spine network topology
  - Software defined storage (SDS)
  - iSCSI SAN

1.1 Dedicated storage networks

Fibre Channel storage has imparted a traditional network implementation to IP-based storage. The dedicated storage network has a proven design that provides performance, predictability, and manageability to storage deployments. The storage traffic is isolated from the application traffic to allow each network to be optimized for its own purpose.

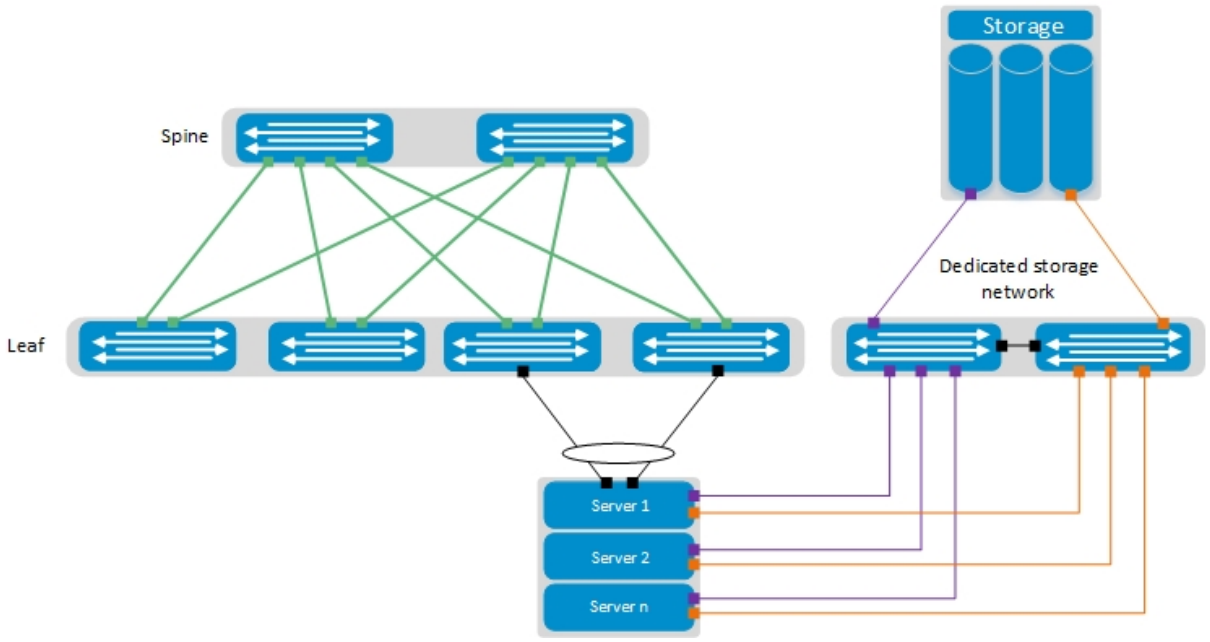


Figure 1 Dedicated storage network with application network

1.2 Shared leaf-spine networks

Data center networks have traditionally been built in a three-layer hierarchical tree consisting of access, aggregation, and core layers as shown in Figure 2.

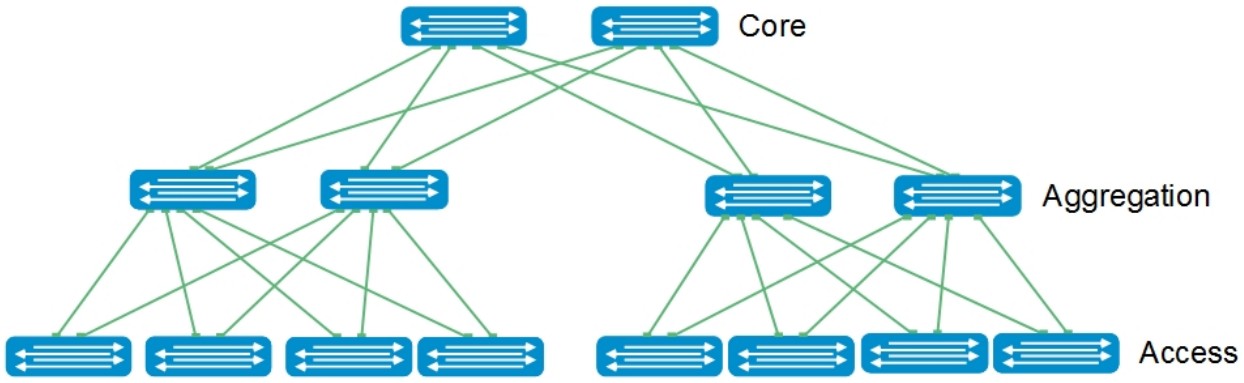


Figure 2 Traditional network architecture

Due to increasing east-west traffic within the data center (server-server, server-storage, etc.), an alternative to the traditional access-aggregation-core network model is becoming more widely used. This architecture, shown in Figure 3, is known as a Clos or leaf-spine network and is designed to minimize the number of hops between hosts.

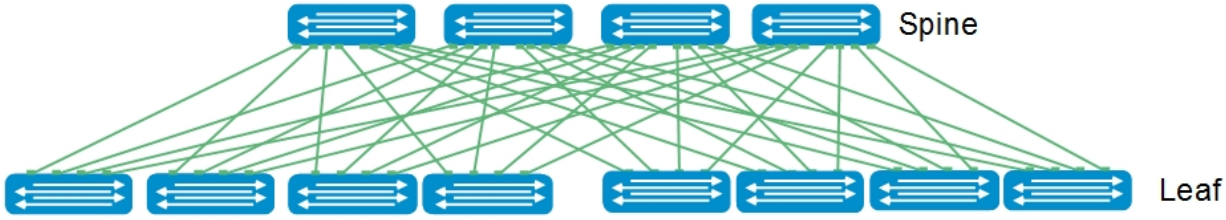


Figure 3 Leaf-spine architecture

In a leaf-spine architecture, the access layer is referred to as the leaf layer. Servers and storage devices connect to leaf switches at this layer. At the next level, the aggregation and core layers are condensed into a single spine layer. Every leaf switch connects to every spine switch to ensure that all leaf switches are no more than one hop away from one another. This minimizes latency and the likelihood of bottlenecks in the network.

A leaf-spine architecture is highly scalable. As administrators add racks to the data center, a pair of leaf switches are added to each new rack. Spine switches may be added as bandwidth requirements increase.

1.3 Typographical conventions

This document uses the following typographical conventions:

`Monospaced text`	Command Line Interface (CLI) examples
`Bold monospaced text`	Commands entered at the CLI prompt
`Italic monospaced text`	Variables in CLI examples
`Underlined Monospace Text`	CLI examples that wrap the page. This text is entered as a single command.

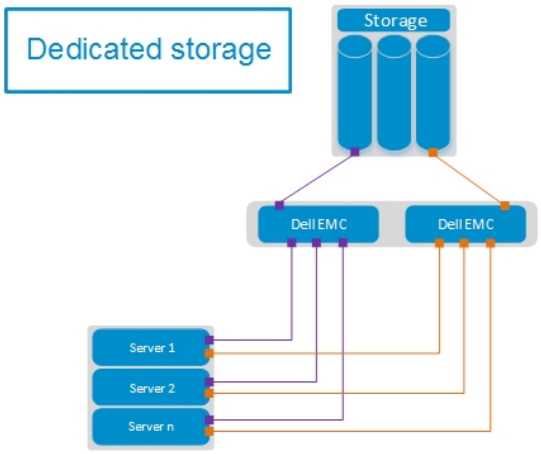
2 Objective

This document provides universal guidelines for design and configuration of a both dedicated storage networks for iSCSI SAN and leaf-spine networks for both iSCSI and SDS. The content includes full configuration steps for each architecture in addition to recommendations for storage based features that can be implemented across a variety of storage technologies.

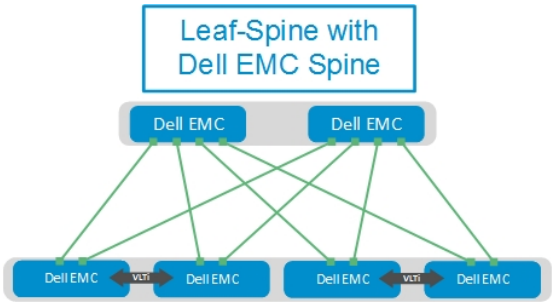
This deployment guide provides step-by-step configuration examples. For dedicated storage, both isolated and switch interconnect topologies are documented. For the shared leaf-spine network, layer 2 (switched) and layer 3 (routed) topologies are detailed to include examples using either Dell EMC Networking switches or Cisco switches at the spine layer.

2.1 Navigating this document

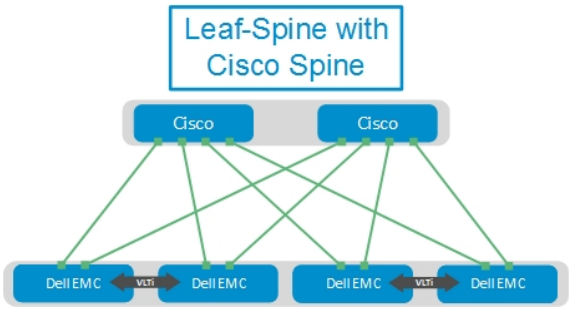
Each choice below contains the full switch configuration within their respective sections.



Dedicated storage architecture



Leaf-Spine architecture with Dell EMC spine and leaf



Leaf-Spine architecture with Cisco spine and Dell EMC leaf

3 Dedicated storage network

A dedicated storage network has been a popular standard for iSCSI based storage. The two main reasons to deploy a dedicated storage network are performance and administration.

Performance:

- Low latency
- Predictable performance
  - No competition with application traffic

Administration:

- Simplify management
  - Storage administrators manage storage network
- Streamline storage troubleshooting
  - Simple fault isolation

The following figure shows the dedicated storage network on the right and a leaf-spine network for application traffic on the left.

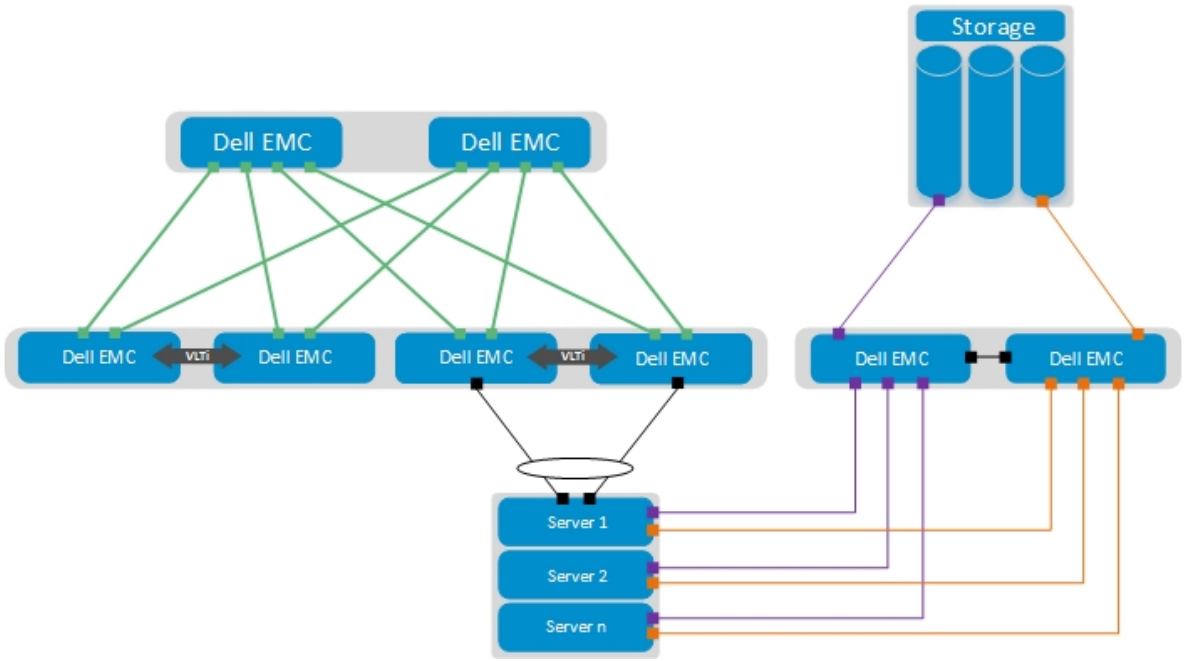
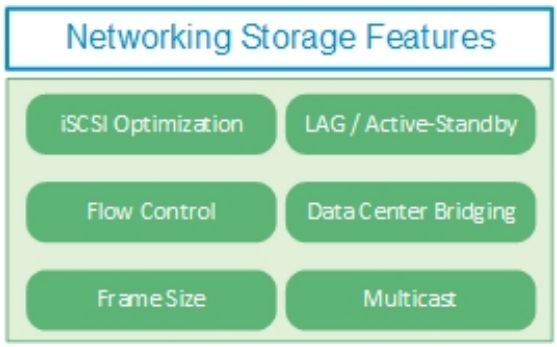


Figure 4 Dedicated storage network diagram

There are two main topologies for switch pairs in the dedicated storage network. The type of topology you choose should be based on the storage appliance being deployed.

Note: The recommendations and information in this section are not written for any specific storage device. Consult the user guide to determine the best features and settings to implement for your specific storage model.

To learn more about storage related features on Dell EMC Networking switches, use the link below to navigate to the following section:



Networking features and guidelines for storage

3.1 Isolated switch pair

Many storage appliances prefer to use an isolated pair of switches for the storage network. In this type of deployment, different subnets are used on each switch.

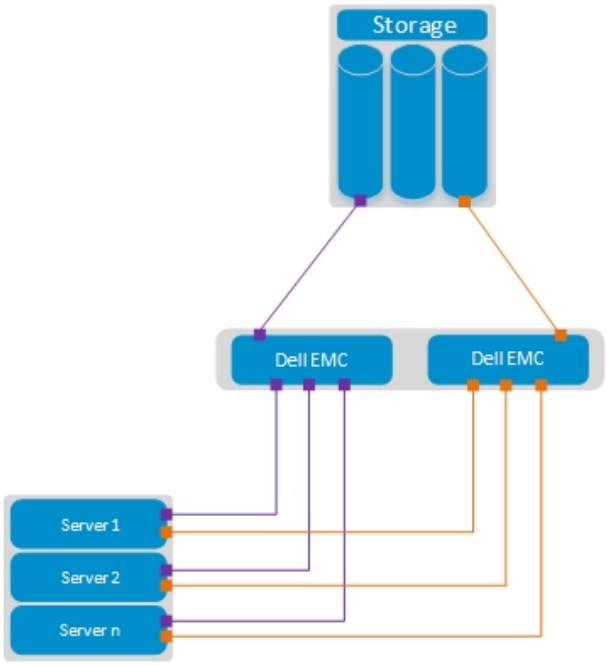


Figure 5 Isolated switch pair

3.2 Isolated switch pair configuration

The following sections contain the configuration examples for an isolated switch pair using Dell Networking OS 9.

3.2.1 Enable switch ports

```
Dell#configure
Dell(conf)#interface range tengigabitethernet 1/1-48
Dell(conf-if-range-te-1/1-48)#switchport
Dell(conf-if-range-te-1/1-48)#no shutdown
Dell(conf-if-range-te-1/1-48)#exit
Dell(conf)#exit
```

3.2.2 Enable jumbo frames

```
Dell#configure
Dell(conf)#interface range tengigabitethernet 1/1–48
Dell(conf -if-range-te-1/1-48)#mtu 9216
```

3.2.3 Configure flow control

```
Dell(conf -if-range-te-1/1-48)#flowcontrol rx on tx off
```

3.2.4 Configure Spanning Tree Protocol on edge ports

Note: Make sure that the following command is used only on server-connected and storage-connected edge ports.

```
Dell(conf-if-range-te-1/1-48)#spanning-tree rstp edge-port
Dell(conf-if-range-te-1/1-48)#exit
Dell(conf)#protocol spanning-tree rstp
Dell(conf-rstp)#no disable
Dell(conf-rstp)#exit
```

3.2.5 VLAN configuration example

Note: Dell recommends assigning a unique VLAN ID, between 2 and 4094, to each switch fabric. For example, assign VLAN 100 on the first switch and VLAN 200 on the second switch. The following example assigns all the ports to the VLAN, however, you may also assign individual ports to the VLAN after they are enabled, as shown in section 4.2.1. If you prefer to use the default VLAN, then you may skip this section entirely. In addition, edge devices, such as server NIC ports and the storage NIC ports, need to be configured with the corresponding VLAN tag.

```
Dell(config)#interface vlan vlan_id
Dell#(config-if-vl-###)#no shutdown
Dell#(config-if-vl-###)#tagged tengigabitethernet 1/1-1/48
Dell#exit
```

3.2.6 Configure second switch

Repeat the commands from sections 3.2.1 through 3.2.5 to configure the second switch. Be sure to use a different VLAN number for the second switch.

3.3 Switch pair with switch interconnect

Many storage appliances prefer to use a pair of switches with a switch interconnect for the storage network. In this type of deployment, all subnets are used on each switch.

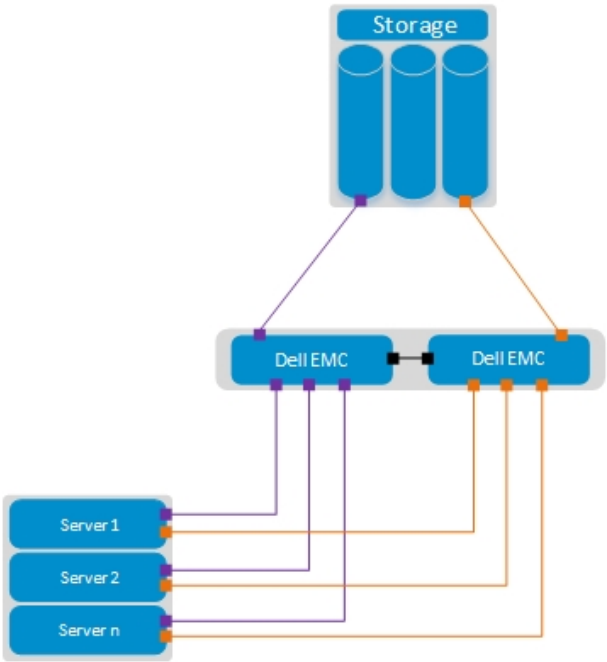


Figure 6 Switch pair with switch interconnect

3.4 Switch pair with switch interconnect configuration

The following sections contain the configuration examples for a switch pair with switch interconnect.

3.4.1 Enable switch ports

```
Dell(conf)#interface range tengigabitethernet 1/1-48
Dell(conf-if-range-te-1/1-48)#switchport
Dell(conf-if-range-te-1/1-48)#no shutdown
Dell(conf-if-range-te-1/1-48)#exit
Dell(conf)#exit
```

3.4.2 Enable jumbo frames

```
Dell#configure
Dell(conf)# interface range tengigabitethernet 1/1–48
Dell(conf-if-range-te-1/1-48)#mtu 9216
```

3.4.3 Configure flow control

```
Dell(conf-if-range-te-1/1-48)#flowcontrol rx on tx off
```

3.4.4 Configure Spanning Tree Protocol on edge ports

```
Dell(conf-if-range-te-1/1-48)#spanning-tree rstp edge-port
Dell(conf-if-range-te-1/1-48)#exit
Dell(conf)# protocol spanning-tree rstp
Dell(conf-rstp)#no disable
Dell(conf-rstp)#exit
```

3.4.5 Configure port channel for the switch interconnect

The following commands configure the switch interconnect:

```
Dell(conf)#interface Port-channel 1
Dell(conf-if-po-1)#mtu 9216
Dell(conf-if-po-1)#switchport
Dell(conf-if-po-1)#no shutdown
Dell(conf-if-po-1)#exit
```

3.4.6 Configure QSFP ports for LAG

The following commands assign 40Gb QSFP ports to the switch interconnect:

```
Dell(conf)#interface range fortyGigE 1/49 , fortyGigE 1/50
Dell(conf-if-range-fo-1/49,fo-1/50)#no ip address
Dell(conf-if-range-fo-1/49,fo-1/50)#mtu 9216
Dell(conf-if-range-te-1/49,fo-1/50)#no shutdown
Dell(conf-if-range-fo-1/49,fo-1/50)#flowcontrol rx on tx off
Dell(conf-if-range-fo-1/49,fo-1/50)#port-channel-protocol lacp
Dell(conf-if-range-fo-1/49,fo-1/50-lacp)#port-channel 1 mode active
Dell(conf-if-range-fo-1/49,fo-1/50-lacp)#exit
Dell(conf-if-range-fo-1/49,fo-1/50)#exit
Dell(conf)#exit
```

3.4.7 Configure second switch

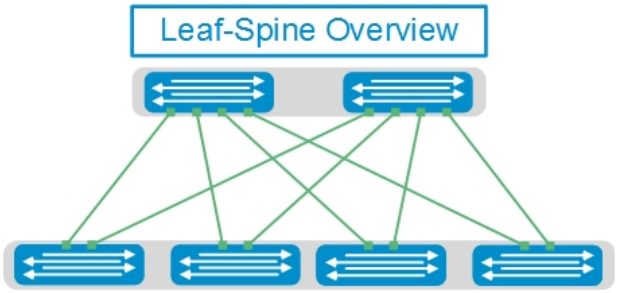
Repeat the commands from sections 3.4.1 through 3.4.6, to configure the second switch.

Note: The preceding procedure places all switch ports in the default VLAN. To place ports in a non-default VLAN, refer to the switch documentation.

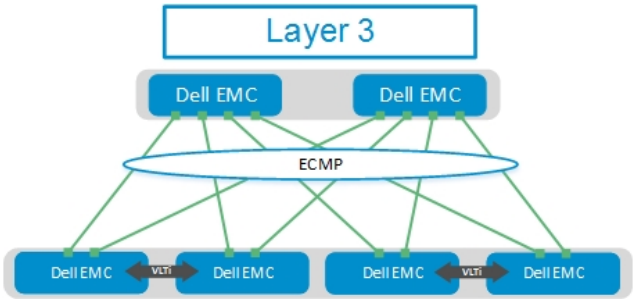
4 Leaf-Spine architecture with Dell EMC spine and leaf

This section includes configuration information for both the layer 3 and layer 2 topologies.

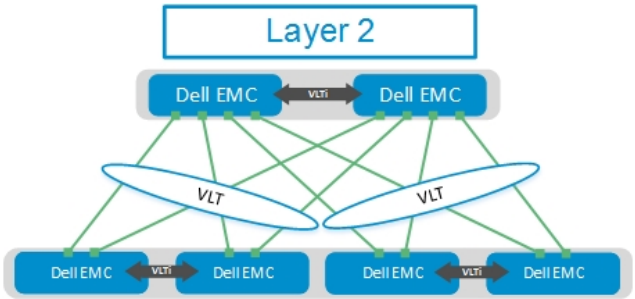
Use the following hyperlinks to navigate to the appropriate sections:



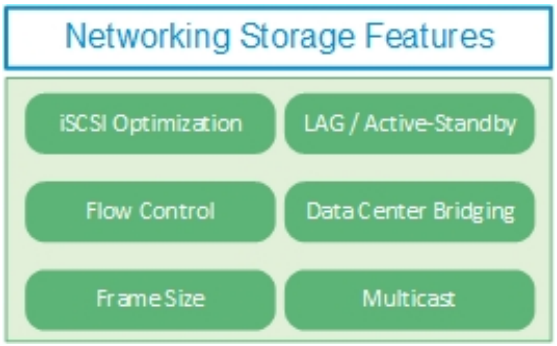
Overview of leaf-spine architecture



Layer 3 topology



Layer 2 topology



Networking features and guidelines for storage

4.1 Layer 3 switch configuration

This section provides an overview of the protocols used in constructing the leaf-spine network examples in this guide.

The first three protocols are used in all layer 2 and layer 3 topology examples:

- Virtual Link Trunking (VLT)
- Uplink Failure Detection (UFD)
- Rapid Spanning Tree Protocol (RSTP)
- Link Aggregation Protocol (LACP) / Link Aggregation Group (LAG)

The remaining protocols are only used in the layer 3 topology examples:

- Routing protocols
  - Border Gateway Protocol (BGP)
  - Open Shortest Path First (OSPF)
- Bidirectional Forwarding Detection (BFD)
- Equal-cost multipath routing (ECMP)

4.2 Layer 3 topology protocols

4.2.1 VLT

VLT allows link aggregation group (LAG) terminations on two separate switches and supports a loop-free topology. The two switches are referred to as VLT peers and are kept synchronized via an inter-switch link called the VLT interconnect (VLTi). A separate backup link maintains heartbeat messages across the OOB management network.

VLT provides layer 2 multipathing and traffic load-balancing. VLT offers the following additional benefits:

- Eliminates blocked ports from STP
- Uses all available uplink bandwidth
- Provides fast convergence if either a link or device fails
- Assures high availability

In layer 2 leaf-spine topologies, VLT is used at both the leaf and spine layers.

In layer 3 topologies, VLT is only used at the leaf layer. An additional feature called VLT peer routing is enabled on the leaf switches for connections to layer 3 networks. VLT peer routing:

- Enables one VLT node to act as the default gateway for its VLT peer
- Eliminates the need to use Virtual Router Redundancy Protocol (VRRP)
- Enables active-active load sharing

With peer routing enabled, traffic is routed through either VLT peer and is passed directly to the next hop without needing to traverse the VLTi.

Note: Downstream connections from leaf switches configured for VLT do not have to be configured as LAGs if other fault tolerant methods, such as multipath IO, are preferred.

4.2.2 LACP/LAG

Link Aggregation Group (LAG) bundles multiple links into a single interface to increase bandwidth between two devices. LAGs also provide redundancy via the multiple paths. In a leaf-spine network, LAGs are typically used to attach servers or storage devices to the VLT leaf pairs.

Link Aggregation Control Protocol (LACP) is an improvement over static LAGs in that the protocol will automatically failover if there is a connectivity issue. This is especially important if the links traverse a media converter where it is possible to lose Ethernet connectivity while links remain in an Up state.

4.2.3 UFD

If a leaf switch loses all connectivity to the spine layer, by default the attached hosts continue to send traffic to that leaf without a direct path to the destination. The VLTi link to the peer leaf switch handles traffic during such a network outage, but this is not considered a best practice.

Dell EMC recommends enabling uplink failure detection (UFD), which detects the loss of upstream connectivity. An uplink-state group is configured on each leaf switch, which creates an association between the uplinks to the spines and the downlink interfaces.

In the event that all uplinks on a switch fail, UFD automatically shuts down the downstream interfaces. This propagates to the hosts attached to the leaf switch. The host then uses its link to the remaining switch to continue sending traffic across the leaf-spine network.

4.2.4 RSTP

As a precautionary measure, Dell EMC recommends enabling Rapid Spanning Tree Protocol (RSTP) on all switches that have layer 2 interfaces. Because VLT environments are loop-free, simultaneously running spanning tree is optional though considered a best practice in case of switch misconfiguration or improperly connected cables. In properly configured and connected leaf-spine networks, there are no ports blocked by STP.

4.2.5 Routing protocols

The following routing protocols may be used on layer 3 connections when designing a leaf-spine network:

- BGP
- OSPF

4.2.5.1 BGP

Border Gateway Protocol (BGP) may be selected for scalability and is well suited for very large networks. BGP can be configured as External BGP (EBGP) to route between autonomous systems or Internal BGP (IBGP) to route within a single autonomous system.

Layer 3 leaf-spine networks use ECMP routing. EBGP and IBGP handle ECMP differently. By default, EBGP supports ECMP without any adjustments. IBGP requires a BGP route reflector and the use of the AddPath feature to fully support ECMP. To keep configuration complexity to a minimum, Dell EMC recommends EBGP in leaf-spine fabric deployments.

BGP tracks IP reachability to the peer remote address and the peer local address. Whenever either address becomes unreachable, BGP brings down the session with the peer. To ensure fast convergence with BGP, Dell EMC recommends enabling fast fall-over with BGP. Fast fall-over terminates external BGP sessions of any directly adjacent peer if the link to reach the peer goes down without waiting for the hold-down timer to expire.

Examples using EBGP (BGPv4) are provided in the layer 3 topology examples in this guide.

4.2.5.2 OSPF

Open Shortest Path First (OSPF) is an interior gateway protocol that provides routing inside an autonomous network. OSPF routers send link-state advertisements to all other routers within the same autonomous system areas. While generally more memory and CPU intensive than BGP, OSPF offers faster convergence and is often used in smaller networks.

Examples using OSPF (OSPFv2 for IPv4) are provided in the layer 3 topology examples in this guide.

4.2.6 BFD

Bidirectional Forwarding Detection (BFD) is a protocol used to rapidly detect communication failures between two adjacent systems over a layer 3 link. It is a simple and lightweight replacement for existing routing protocol link state detection mechanisms. Though optional, use of BFD is considered a best practice for optimizing a leaf-spine network.

BFD provides forwarding path failure detection times on the order of milliseconds rather than seconds as with conventional routing protocols. It is independent of routing protocols and provides a consistent method of failure detection when used across a network. Networks converge faster because BFD triggers link state changes in the routing protocol sooner and more consistently.

Dell EMC Networking has implemented BFD at layer 3 with user datagram protocol (UDP) encapsulation. BFD is supported with routing protocols including BGP and OSPF.

4.2.7 ECMP

The nature of a leaf-spine topology is that leaf switches are no more than one hop away from each other. As shown in Figure 7, Leaf 1 has two equal cost paths to Leaf 4, one through each spine. The same is true for all leafs.

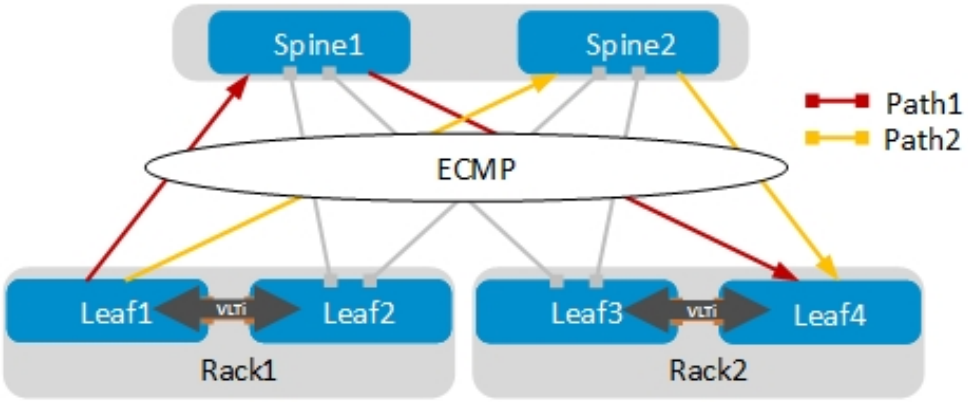


Figure 7 Use of ECMP in a layer 3 topology

Equal-cost multipath routing, or ECMP, is a routing technique used in a layer 3 leaf-spine topology for load balancing packets along these multiple equal cost paths. ECMP is enabled on all leaf and spine switches, allowing traffic between leafs to be load balanced across the spines.

4.3 Layer 3 configuration planning

4.3.1 BGP ASN configuration

When EBGP is used, an autonomous system number (ASN) is assigned to each switch. Valid private, 2-byte ASNs range from 64512 through 65534. Figure 8 shows the ASN assignments used for leaf and spine switches in the BGP examples in this guide.

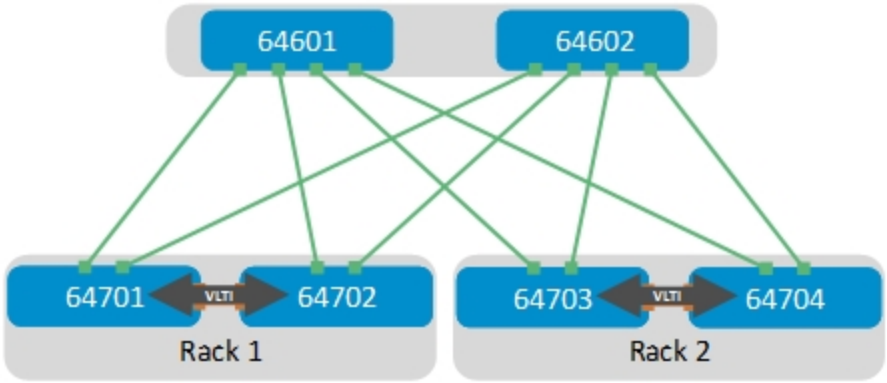


Figure 8 BGP ASN assignments

ASNs should follow a logical pattern for ease of administration and allow for growth as additional leaf and spine switches are added. In this example, an ASN with a "6" in the hundreds place, such as 64601, represents a spine switch and an ASN with a "7" in the hundreds place, such as 64701, represents a leaf switch.

4.3.2 IP addressing

Establishing a logical, scalable IP address scheme is important before deploying a leaf-spine topology. This section covers the IP addressing used in the layer 3 examples in this guide.

4.3.2.1 Loopback addresses

When configuring routing protocols, loopback addresses can be used as router IDs. As with ASNs, loopback addresses should follow a logical pattern that will make it easier for administrators to manage the network and allow for growth. Figure 9 shows the loopback addresses used as router IDs in the BGP and OSPF examples in this guide.

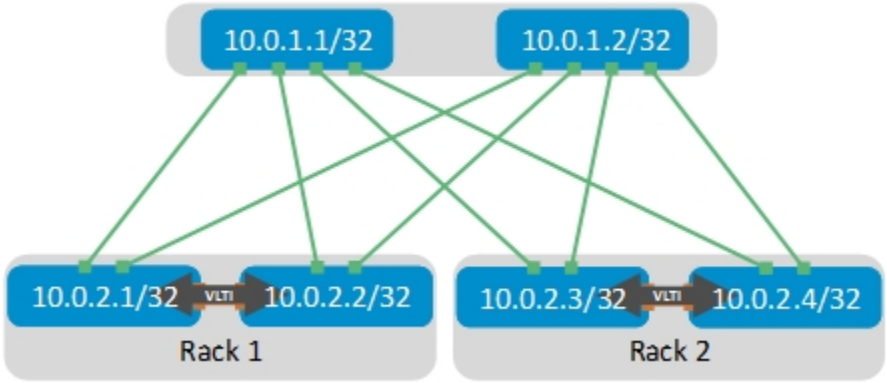


Figure 9 Loopback addressing

All loopback addresses used are part of the 10.0.0.0/8 address space with each address using a 32-bit mask. In this example, the third octet represents the layer, “1” for spine and “2” for leaf. The fourth octet is the

counter for the appropriate layer. For example, 10.0.1.1/32 is the first spine switch in the topology while 10.0.2.4/32 is the fourth leaf switch.

4.3.2.2 Point-to-point addresses

Table 1 lists layer 3 connection details for each leaf and spine switch. All addresses come from the same base IP prefix, 192.168.0.0/16 with the third octet representing the spine number. For example, 192.168.1.0/31 is a two host subnet connected to Spine 1 while 192.168.2.0/31 is connected to Spine 2. This IP scheme is easily extended as leaf and spine switches are added to the network.

Link labels are provided in the table for quick reference with Figure 10.

Table 1 Interface and IP configuration

Link Label	Source switch	Source interface	Source IP	Network	Destination switch	Destination interface	Destination IP
A	Leaf 1	fo1/49	.1	192.168.1.0/31	Spine 1	fo1/1/1	.0
B	Leaf 1	fo1/50	.1	192.168.2.0/31	Spine 2	fo1/1/1	.0
C	Leaf 2	fo1/49	.3	192.168.1.2/31	Spine 1	fo1/2/1	.2
D	Leaf 2	fo1/50	.3	192.168.2.2/31	Spine 2	fo1/2/1	.2
E	Leaf 3	fo1/49	.5	192.168.1.4/31	Spine 1	fo1/3/1	.4
F	Leaf 3	fo1/50	.5	192.168.2.4/31	Spine 2	fo1/3/1	.4
G	Leaf 4	fo1/49	.7	192.168.1.6/31	Spine 1	fo1/4/1	.6
H	Leaf 4	fo1/50	.7	192.168.2.6/31	Spine 2	fo1/4/1	.6

The point-to-point IP addresses used in this guide are shown in Figure 10:

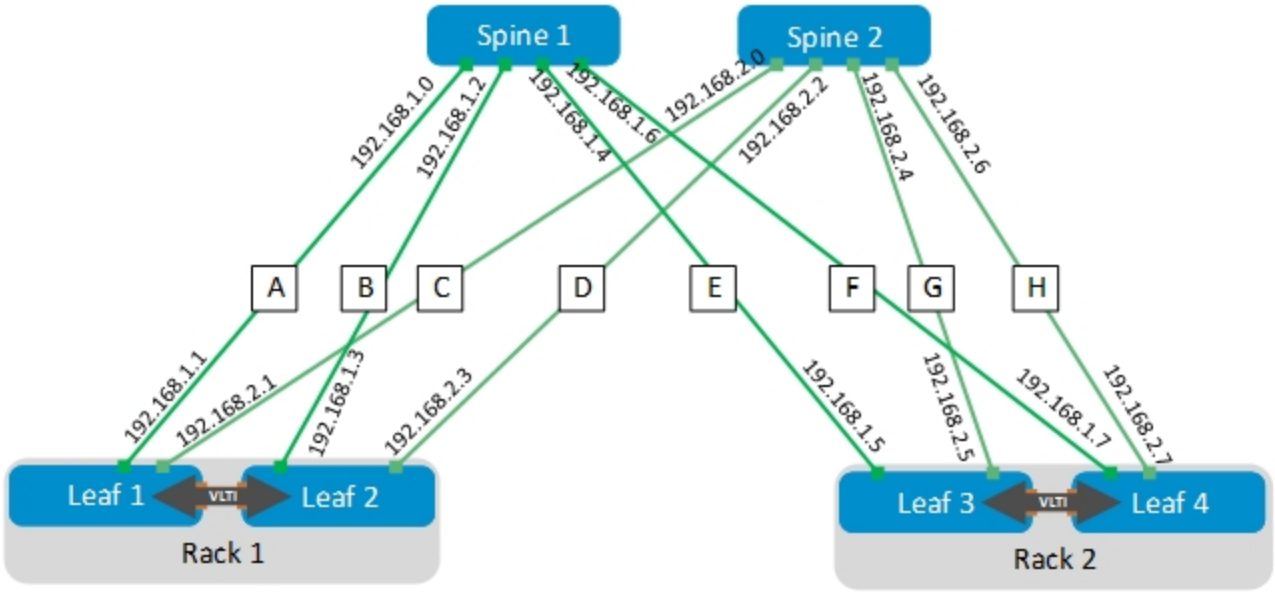


Figure 10 Point-to-point IP addresses

Note: The example point-to-point addresses use a 31-bit mask to save address space. This is optional and covered in RFC 3021. The example uses IP address when setting a 31-bit mask on a Dell EMC S4048-ON. The warning message can be safely ignored on point-to-point interfaces:

```
S4048-Leaf-1(conf-if-fo-1/49)#ip address 192.168.1.1/31
% Warning: Use /31 mask on non point-to-point interface cautiously.
```

4.4 Layer 3 configuration with Dell EMC leaf and spine switches

This section provides BGP and OSPF configuration examples to build the layer 3 leaf-spine topology shown in Figure 11. Dell EMC Networking S4048-ON switches are used at the leaf layer and Dell EMC Networking Z9100-ON switches are used at the spine layer.

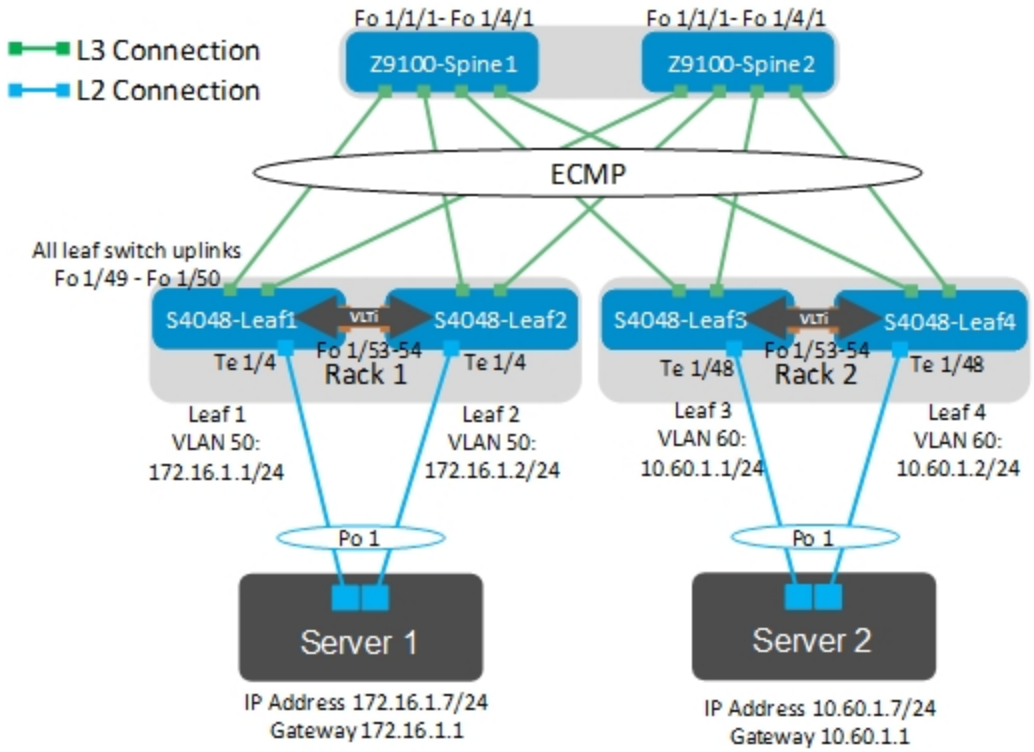


Figure 11 Layer 3 leaf-spine topology with Dell EMC switches

In this topology, there is one broadcast domain in each rack.

In Rack 1, VLAN 50 is used and devices in VLAN 50 are assigned IP addresses on the 172.16.1.0/24 network. With VLT peer routing enabled on S4048-Leaf1 and S4048-Leaf2, Server 1 may specify the IP address assigned to VLAN 50 on either leaf, 172.16.1.1 or 172.16.1.2, as its default gateway. Traffic is load-balanced across both leafs.

Rack 2 is configured in an identical manner, except VLAN 60 is used and devices in VLAN 60 are assigned IP addresses on the 10.60.1.0/24 network. Server 2 may specify the VLAN 60 IP address of either leaf, 10.60.1.1 or 10.60.1.2, as its default gateway.

4.4.1 S4048-ON leaf switch configuration

The following configuration details are for S4048-Leaf1 and S4048-Leaf2 in Figure 11. The configuration commands for S4048-Leaf3 and S4048-Leaf4 are similar.

Note: On S4048-ON switches, Telnet is enabled and SSH is disabled by default. Both services require the creation of a non-root user account to login. If needed, it is a best practice to use SSH instead of Telnet for security. SSH can optionally be enabled with the command:

```
(conf)#ip ssh server enable
```

A user account can be created to access the switch via SSH with the command:

```
(conf)#username ssh_user sha256-password ssh_password
```

1. Using the following commands to configure the serial console, enable password, and disable Telnet:

Table 2 Console, password, and Telnet configuration commands

S4048-Leaf1	S4048-Leaf2
enable configure enable sha256-password *enable_password* no ip telnet server enable	enable configure enable sha256-password *enable_password* no ip telnet server enable

2. Enter the commands in Table 3 below to:
   a. Set the hostname, configure the OOB management interface and default gateway
   b. Enable LLDP and BFD
   c. Enable RSTP as a precaution
   d. Configure S4048-Leaf1 as the primary RSTP root bridge using the bridge-priority 0 command
   e. Configure S4048-Leaf2 as the secondary RSTP root bridge using the bridge-priority 4096 command.

Table 3 Leaf1 and Leaf2 configuration commands

S4048-Leaf1	S4048-Leaf2
hostname S4048-Leaf1 interface ManagementEthernet 1/1 ip address 100.67.187.35/24 no shutdown management route 0.0.0.0/0 100.67.187.254 protocol lldp advertise management-tlv management-address system- description system-name advertise interface-port-desc bfd enable protocol spanning-tree rstp no disable bridge-priority 0	hostname S4048-Leaf2 interface ManagementEthernet 1/1 ip address 100.67.187.34/24 no shutdown management route 0.0.0.0/0 100.67.187.254 protocol lldp advertise management-tlv management-address system- description system-name advertise interface-port-desc bfd enable protocol spanning-tree rstp no disable bridge-priority 4096

3. Configure the VLT interconnect between S4048-Leaf1 and S4048-Leaf2. In this configuration, add interfaces fortyGigE 1/53-54 to static port channel 127 for the VLT interconnect. The backup destination is the management IP address of the VLT peer switch. Enable peer routing.

Note: Dell EMC recommends that the VLTi is configured as a static LAG, without LACP, per the commands shown below.

Table 4 VLT interconnect configuration

S4048-Leaf1	S4048-Leaf2
interface port-channel 127 description VLTi channel-member fortyGigE 1/53 - 1/54 no shutdown interface range fortyGigE 1/53 - 1/54 description VLTi no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.187.34 unit-id 0 peer-routing exit	interface port-channel 127 description VLTi channel-member fortyGigE 1/53 - 1/54 no shutdown interface range fortyGigE 1/53 - 1/54 description VLTi no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.187.35 unit-id 1 peer-routing exit

4. Configure each downstream server-facing interface with an LACP port channel. Configure each port channel for VLT. Port channel 1 connects downstream to Server 1 and is configured as an RSTP edge port.

Table 5 Port channel configuration

S4048-Leaf1	S4048-Leaf2
interface tengigabitethernet 1/4 description Server 1 port-channel-protocol LACP port-channel 1 mode active no shutdown interface port-channel 1 description Server 1 portmode hybrid switchport spanning-tree rstp edge-port vlt-peer-lag port-channel 1 no shutdown	interface tengigabitethernet 1/4 description Server 1 port-channel-protocol LACP port-channel 1 mode active no shutdown interface port-channel 1 description Server 1 portmode hybrid switchport spanning-tree rstp edge-port vlt-peer-lag port-channel 1 no shutdown

5. Create a VLAN interface containing the server-facing port channel(s). Use the same VLAN ID on both leafs. Create a switched virtual interface (SVI) by assigning an IP address to the VLAN interface. The address must be unique but on the same network on both leaf switches.

Table 6 VLAN interface configuration

S4048-Leaf1	S4048-Leaf2
interface Vlan 50 ip address 172.16.1.1/24 untagged Port-channel 1 no shutdown	interface Vlan 50 ip address 172.16.1.2/24 untagged Port-channel 1 no shutdown

6. The two upstream layer 3 interfaces connected to the spine switches are configured. Assign IP addresses per Table 1. Configure a loopback interface to be used as the router ID. This is used with BGP or OSPF.

Note: If multiple loopback interfaces exist on a system, the interface with the highest numbered IP address is used as the router ID. This configuration only uses one loopback interface.

Table 7 Interface and loopback configuration

S4048-Leaf1	S4048-Leaf2
interface fortyGigE 1/49 description Spine-1 ip address 192.168.1.1/31 no shutdown interface fortyGigE 1/50 description Spine-2 ip address 192.168.2.1/31 no shutdown interface loopback 0 description Router ID ip address 10.0.2.1/32 no shutdown	interface fortyGigE 1/49 description Spine-1 ip address 192.168.1.3/31 no shutdown interface fortyGigE 1/50 description Spine-2 ip address 192.168.2.3/31 no shutdown interface loopback 0 description Router ID ip address 10.0.2.2/32 no shutdown

7. Configure a route map and IP prefix-list to redistribute all loopback addresses and leaf networks via BGP or OSPF.

Note: The command seq 10 permit 10.0.0.0/8 ge 24 includes all addresses in the 10.0.0.0/8 address range with a mask greater than or equal to 24. This includes all loopback addresses used as router IDs as well as the 10.60.1.0/24 network used on Leafs 3 and 4 as shown in Figure 11.
The command seq 20 permit 172.16.0.0/16 ge 24 includes the 172.16.1.0/24 network used on Leafs 1 and 2 as shown in Figure 11.

Table 8 Route map and loopback configuration

S4048-Leaf1	S4048-Leaf2
route-map spine-leaf permit 10 match ip address spine-leaf ip prefix-list spine-leaf description Redistribute loopback and leaf networks seq 10 permit 10.0.0.0/8 ge 24 seq 20 permit 172.16.0.0/16 ge 24	route-map spine-leaf permit 10 match ip address spine-leaf ip prefix-list spine-leaf description Redistribute loopback and leaf networks seq 10 permit 10.0.0.0/8 ge 24 seq 20 permit 172.16.0.0/16 ge 24

8. Include the point-to-point interfaces to each leaf pair in an ECMP group. Enable link bundle monitoring to report when traffic is unevenly distributed across multiple links.

Note: ECMP is not enabled until BGP or OSPF is configured.

Table 9 Enable link bundle configuration

S4048-Leaf1	S4048-Leaf2
ecmp-group 1 interface fortyGigE 1/49 interface fortyGigE 1/50 link-bundle-monitor enable	ecmp-group 1 interface fortyGigE 1/49 interface fortyGigE 1/50 link-bundle-monitor enable

9. Configure UFD to shut down the downstream interfaces if all uplinks fail. The hosts attached to the switch use the remaining LACP port member to continue sending traffic across the fabric.

Table 10 UFD configuration

S4048-Leaf1	S4048-Leaf2
uplink-state-group 1 description Disable downstream ports in event all uplinks fail downstream TenGigabitEthernet 1/1-1/48 upstream fortyGigE 1/49,1/50 end write	uplink-state-group 1 description Disable downstream ports in event all uplinks fail downstream TenGigabitEthernet 1/1-1/48 upstream fortyGigE 1/49,1/50 end write

10. Exit configuration mode and save the configuration.

4.4.1.1 S4048-ON BGP configuration

Use the following commands to configure BGP:

Note: If OSPF is used, skip to section 4.4.1.2.

1. Enable BGP using the `router bgp` *ASN* command.

Note: The ASN is from Figure 8.

2. Use the `bgp bestpath as-path multipath-relax` to enable ECMP. The `maximum-paths ebgp 2` command specifies the maximum number of parallel paths to a destination to add to the routing table. This number should be equal to or greater than the number of spines, up to 64.
3. Configure BGP neighbors and enable fast fall-over.
4. Configure BFD settings to 100 millisecond send/receive intervals.

Note: The multiplier is the number of packets that must be missed to declare a session down.

5. Use the `end` and `write` commands to exit the configuration mode and to save the configuration.

Table 11 Dell EMC Networking S4048-ON BGP configuration

S4048-Leaf1	S4048-Leaf2
enable configure router bgp 64701 bgp bestpath as-path multipath-relax maximum-paths ebgp 2 redistribute connected route-map spine-leaf bgp graceful-restart neighbor spine-leaf peer-group neighbor spine-leaf fall-over neighbor spine-leaf advertisement- interval 1 neighbor spine-leaf no shutdown neighbor spine-leaf bfd neighbor 192.168.1.0 remote-as 64601 neighbor 192.168.1.0 peer-group spine- leaf neighbor 192.168.1.0 no shutdown neighbor 192.168.2.0 remote-as 64602 neighbor 192.168.2.0 peer-group spine- leaf neighbor 192.168.2.0 no shutdown bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write	enable configure router bgp 64702 bgp bestpath as-path multipath-relax maximum-paths ebgp 2 redistribute connected route-map spine-leaf bgp graceful-restart neighbor spine-leaf peer-group neighbor spine-leaf fall-over neighbor spine-leaf advertisement- interval 1 neighbor spine-leaf no shutdown neighbor spine-leaf bfd neighbor 192.168.1.2 remote-as 64601 neighbor 192.168.1.2 peer-group spine- leaf neighbor 192.168.1.2 no shutdown neighbor 192.168.2.2 remote-as 64602 neighbor 192.168.2.2 peer-group spine- leaf neighbor 192.168.2.2 no shutdown bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write

4.4.1.2 S4048-ON OSPF configuration

Use the following commands to configure OSPF:

Note: Skip this section if BGP is used.

1. Enable OSPF using the `router ospf process-id` command.

Note: The valid range is 1 to 65535.

2. Add the connected networks to OSPF area 0.
3. Use the `maximum-paths 2` command to enable ECMP and to specify the maximum number of parallel paths to a destination to add to the routing table. This number should be equal to or greater than the number of spines, up to 64.
4. Configure the BFD settings to 100 millisecond send/receive intervals.

Note: The multiplier is the number of packets that must be missed to declare a session down.

5. Use the `end` and `write` commands to exit the configuration mode and to save the configuration.

Table 12 OSPF configuration commands

S4048-Leaf1	S4048-Leaf2
enable configure router ospf 1 log-adjacency-changes network 192.168.1.0/31 area 0 network 192.168.2.0/31 area 0 maximum-paths 2 redistribute connected route-map spine-leaf bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write	enable configure router ospf 1 log-adjacency-changes network 192.168.1.2/31 area 0 network 192.168.2.2/31 area 0 maximum-paths 2 redistribute connected route-map spine-leaf bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write

4.4.2 Z9100-ON spine switch configuration

The following configuration details are for Z9100-Spine1 and Z9100-Spine2 in Figure 11.

Note: On Z9100-ON switches, Telnet is enabled and SSH is disabled by default. Both services require the creation of a non-root user account to login. If needed, it is a best practice to use SSH instead of Telnet for security. SSH can optionally be enabled with the command:

```
(conf)#ip ssh server enable
```

A user account can be created to access the switch via SSH with the command:

```
(conf)#username ssh_user sha256-password ssh_password
```

1. Use the following commands to configure the serial console, enable the password, and disable Telnet:

Table 13 Console, password, and Telnet commands

Z9100-Spine1	Z9100-Spine2

`enable`
`configure`
 	
`enable sha256-password enable_password`
`no ip telnet server enable`
	`enable`
	`configure`
	
	`enable sha256-password enable_password`
	`no ip telnet server enable`

2. Enter the commands in the Table 14 to perform the following:
   a. Set the hostname
   b. Configure the OOB management interface and default gateway
   c. Set the hostname
   d. Configure the OOB management interface and default gateway
   e. Enable LLDP and BFD
   f. Set the port speed of the four ports connected to the leaf switches to 40GbE

Table 14 Spine1 and Spine2 configuration commands

Z9100-Spine1	Z9100-Spine2
hostname Z9100-Spine1 interface ManagementEthernet 1/1 ip address 100.67.187.39/24 no shutdown management route 0.0.0.0/0 100.67.187.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc bfd enable stack-unit 1 port 1 portmode single speed 40G no-confirm stack-unit 1 port 2 portmode single speed 40G no-confirm stack-unit 1 port 3 portmode single speed 40G no-confirm stack-unit 1 port 4 portmode single speed 40G no-confirm	hostname Z9100-Spine2 interface ManagementEthernet 1/1 ip address 100.67.187.38/24 no shutdown management route 0.0.0.0/0 100.67.187.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc bfd enable stack-unit 1 port 1 portmode single speed 40G no-confirm stack-unit 1 port 2 portmode single speed 40G no-confirm stack-unit 1 port 3 portmode single speed 40G no-confirm stack-unit 1 port 4 portmode single speed 40G no-confirm

3. Configure the four point-to-point interfaces connected to leaf switches.
4. Assign IP addresses per Table 1.
5. Configure a loopback interface to be used as the router ID. This is used with BGP or OSPF.

Note: If multiple loopback interfaces exist on a system, the interface with the highest numbered IP address is used as the router ID. This configuration only uses one loopback interface.

Table 15 Point-to-point and loopback configuration commands

Z9100-Spine1	Z9100-Spine2
interface fortyGigE 1/1/1 description Leaf 1 fo1/49 ip address 192.168.1.0/31 no shutdown interface fortyGigE 1/2/1 description Leaf 2 fo1/49 ip address 192.168.1.2/31 no shutdown interface fortyGigE 1/3/1 description Leaf 3 fo1/49 ip address 192.168.1.4/31 no shutdown interface fortyGigE 1/4/1 description Leaf 4 fo1/49 ip address 192.168.1.6/31 no shutdown interface loopback 0 description Router ID ip address 10.0.1.1/32 no shutdown	interface fortyGigE 1/1/1 description Leaf 1 fo1/50 ip address 192.168.2.0/31 no shutdown interface fortyGigE 1/2/1 description Leaf 2 fo1/50 ip address 192.168.2.2/31 no shutdown interface fortyGigE 1/3/1 description Leaf 3 fo1/50 ip address 192.168.2.4/31 no shutdown interface fortyGigE 1/4/1 description Leaf 4 fo1/50 ip address 192.168.2.6/31 no shutdown interface loopback 0 description Router ID ip address 10.0.1.2/32 no shutdown

6. Configure a route map and IP prefix-list to redistribute all loopback addresses and leaf networks via BGP or OSPF.

Note: The command `seq 10 permit 10.0.0.0/8 ge 24` includes all addresses in the 10.0.0.0/8 address range with a mask greater than or equal to 24. This includes all loopback addresses used as router IDs as well as the 10.60.1.0/24 network used on Leafs 3 and 4 as shown in Figure 11. The command `seq 20 permit 172.16.0.0/16 ge 24` includes the 172.16.1.0/24 network used on Leafs 1 and 2 as shown in Figure 11.

Table 16 Route map and loopback configuration commands

Z9100-Spine1	Z9100-Spine2
route-map spine-leaf permit 10 match ip address spine-leaf ip prefix-list spine-leaf description Redistribute loopback and leaf networks seq 10 permit 10.0.0.0/8 ge 24 seq 20 permit 172.16.0.0/16 ge 24	route-map spine-leaf permit 10 match ip address spine-leaf ip prefix-list spine-leaf description Redistribute loopback and leaf networks seq 10 permit 10.0.0.0/8 ge 24 seq 20 permit 172.16.0.0/16 ge 24

7. Add the point-to-point interfaces to each leaf pair in an ECMP group.
8. Enable link bundle monitoring to report when traffic is unevenly distributed across multiple links.

Note: ECMP is not actually enabled until BGP or OSPF is configured.

9. Use the `end` and `write` commands to exit configuration mode and save the configuration.

Table 17 Addition of point-to-point interface commands

Z9100-Spine1	Z9100-Spine2

`ecmp-group 1`
`interface fortyGigE 1/1/1`
`interface fortyGigE 1/2/1`
`link-bundle-monitor enable`
 	
`ecmp-group 2`
`interface fortyGigE 1/3/1`
`interface fortyGigE 1/4/1`
`link-bundle-monitor enable`
 	
`end`
`write`
	`ecmp-group 1`
	`interface fortyGigE 1/1/1`
	`interface fortyGigE 1/2/1`
	`link-bundle-monitor enable`
	
	`ecmp-group 2`
	`interface fortyGigE 1/3/1`
	`interface fortyGigE 1/4/1`
	`link-bundle-monitor enable`
	
	`end`
	`write`

4.4.2.1 Z9100-ON BGP configuration

Use the following commands to configure BGP:

Note: If OSPF is used, skip to section 4.4.2.2.

1. Enable BGP using the `router bgp` *ASN* command.

Note: The ASN is from Figure 8.

2. Use the `bgp bestpath as-path multipath-relax` command to enable ECMP and the `maximum-paths ebgp 2` command to specify the maximum number of parallel paths to a destination to add to the routing table. In this topology, there are two equal cost best paths from a spine to a host, one to each leaf that the host is connected.
3. Ensure that the BGP neighbors are configured and fast fall-over is enabled.

Note: BFD settings are configured to 100 millisecond send/receive intervals. The multiplier is the number of packets that must be missed to declare a session down.

4. Use the `end` and `write` commands to exit the configuration mode and to save the configuration.

Table 18 BGP configuration commands

Z9100-Spine1	Z9100-Spine2
enable configure router bgp 64601 bgp bestpath as-path multipath-relax maximum-paths ebgp 2 redistribute connected route-map spine-leaf bgp graceful-restart neighbor spine-leaf peer-group neighbor spine-leaf fall-over neighbor spine-leaf advertisement- interval 1 neighbor spine-leaf no shutdown neighbor spine-leaf bfd neighbor 192.168.1.1 remote-as 64701 neighbor 192.168.1.1 peer-group spine- leaf neighbor 192.168.1.1 no shutdown neighbor 192.168.1.3 remote-as 64702 neighbor 192.168.1.3 peer-group spine- leaf neighbor 192.168.1.3 no shutdown neighbor 192.168.1.5 remote-as 64703 neighbor 192.168.1.5 peer-group spine- leaf neighbor 192.168.1.5 no shutdown neighbor 192.168.1.7 remote-as 64704 neighbor 192.168.1.7 peer-group spine- leaf neighbor 192.168.1.7 no shutdown bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write	enable configure router bgp 64602 bgp bestpath as-path multipath-relax maximum-paths ebgp 2 redistribute connected route-map spine-leaf bgp graceful-restart neighbor spine-leaf peer-group neighbor spine-leaf fall-over neighbor spine-leaf advertisement- interval 1 neighbor spine-leaf no shutdown neighbor spine-leaf bfd neighbor 192.168.2.1 remote-as 64701 neighbor 192.168.2.1 peer-group spine- leaf neighbor 192.168.2.1 no shutdown neighbor 192.168.2.3 remote-as 64702 neighbor 192.168.2.3 peer-group spine- leaf neighbor 192.168.2.3 no shutdown neighbor 192.168.2.5 remote-as 64703 neighbor 192.168.2.5 peer-group spine- leaf neighbor 192.168.2.5 no shutdown neighbor 192.168.2.7 remote-as 64704 neighbor 192.168.2.7 peer-group spine- leaf neighbor 192.168.2.7 no shutdown bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write

4.4.2.2 Z9100-ON OSPF configuration

Use the following commands to configure OSPF:

Note: Skip this section if BGP is used.

1. Enable OSPF using the `router ospf process-id` command.

Note: A valid range is 1 to 65535.

2. Add the connected networks to OSPF area 0.
3. Enter the `maximum-paths 2` command to enable ECMP that specifies the maximum number of parallel paths to a destination to add to the routing table. In this topology, there are two equal cost best paths from a spine to a host, one to each leaf that the host is connected.
4. Configure the BFD settings to 100 millisecond send/receive intervals.

Note: The multiplier is the number of packets that must be missed to declare a session down.

5. Use the `end` and `write` commands to exit configuration mode and save the configuration.

Table 19 OSPF configuration commands

Z9100-Spine1	Z9100-Spine2
enable configure router ospf 1 log-adjacency-changes network 192.168.1.0/31 area 0 network 192.168.1.2/31 area 0 network 192.168.1.4/31 area 0 network 192.168.1.6/31 area 0 maximum-paths 2 redistribute connected route-map spine-leaf bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write	enable configure router ospf 1 log-adjacency-changes network 192.168.2.0/31 area 0 network 192.168.2.2/31 area 0 network 192.168.2.4/31 area 0 network 192.168.2.6/31 area 0 maximum-paths 2 redistribute connected route-map spine-leaf bfd all-neighbors interval 100 min_rx 100 multiplier 3 role active end write

4.4.3 Layer 3 example validation

In addition to sending traffic between hosts, the configuration shown in Figure 11 can be validated with the commands shown in this section. For more information on commands and output, see the *Command Line Reference Guide* for the applicable switch (links to documentation are provided in Appendix F.

Command and output examples are provided for one spine and one leaf. Command output on other switches is similar.

4.4.3.1 show ip bgp summary

When BGP is configured, the `show ip bgp summary` command shows the status of all BGP connections. Each spine has four neighbors (the four leafs) and each leaf has two neighbors (the two spines). The following commands confirm that BFD is enabled on the 6th line of output:

```
Z9100-Spine-1#show ip bgp summary
BGP router identifier 10.0.1.1, local AS number 64601
BGP local RIB : Routes to be Added 0, Replaced 0, Withdrawn 0
8 network entrie(s) using 608 bytes of memory
```

```
13 paths using 1404 bytes of memory
BGP-RIB over all using 1417 bytes of memory
BFD is enabled, Interval 100 Min_rx 100 Multiplier 3 Role Active
29 BGP path attribute entrie(s) using 4816 bytes of memory
27 BGP AS-PATH entrie(s) using 270 bytes of memory
4 neighbor(s) using 32768 bytes of memory

Neighbor      AS       MsgRcvd  MsgSent     TblVer  InQ  OutQ Up/Down  State/Pfx
192.168.1.1   64701    3014     3014          0    0     0 1d:19:31:07 3
192.168.1.3   64702    3013     3011          0    0     0 1d:19:31:11 3
192.168.1.5   64703    3014     3012          0    0     0 1d:19:30:59 3
192.168.1.7   64704    3014     3012          0    0     0 1d:19:31:06 3

S4048-Leaf-1#show ip bgp summary
BGP router identifier 10.0.2.1, local AS number 64701
BGP local RIB : Routes to be Added 0, Replaced 0, Withdrawn 0
8 network entrie(s) using 608 bytes of memory
12 paths using 1296 bytes of memory
BGP-RIB over all using 1308 bytes of memory
BFD is enabled, Interval 100 Min_rx 100 Multiplier 3 Role Active
17 BGP path attribute entrie(s) using 2752 bytes of memory
15 BGP AS-PATH entrie(s) using 150 bytes of memory
2 neighbor(s) using 16384 bytes of memory

Neighbor        AS            MsgRcvd  MsgSent     TblVer  InQ  OutQ Up/Down
State/Pfx
192.168.1.0     64601            15       17          0    0     0 00:03:41 5
192.168.2.0     64602            13       13          0    0     0 00:03:41 5
```

4.4.3.2 show ip ospf neighbor

When OSPF is configured, the `show ip ospf neighbor` command shows the state of all connected OSPF neighbors. In this configuration, each spine has four neighbors (the four leafs) and each leaf has two neighbors (the two spines).

```
Z9100-Spine-1#show ip ospf neighbor

Neighbor ID    Pri      State        Dead Time Address        Interface   Area
10.0.2.1         1      FULL/DR      00:00:32  192.168.1.1    Fo 1/1/1    0
10.0.2.2         1      FULL/DR      00:00:34  192.168.1.3    Fo 1/2/1    0
10.0.2.3         1      FULL/DR      00:00:35  192.168.1.5    Fo 1/3/1    0
10.0.2.4         1      FULL/DR      00:00:35  192.168.1.7    Fo 1/4/1    0

S4048-Leaf-1#show ip ospf neighbor

Neighbor ID    Pri      State        Dead Time Address        Interface   Area
10.0.1.1         1      FULL/BDR     00:00:38  192.168.1.0    Fo 1/49     0
10.0.1.2         1      FULL/BDR     00:00:39  192.168.2.0    Fo 1/50     0
```

4.4.3.3 show ip route bgp

On switches with BGP configured, the `show ip route bgp` command is used to verify the BGP entries in the Routing Information Base (RIB). Entries with multiple paths shown are used with ECMP. The two server networks in this example, 10.60.1.0 and 172.16.1.0, each have two paths from Z9100-Spine1, one through each leaf.

The first set of routes with a subnet mask of /32 are the IPs configured for router IDs.

```
Z9100-Spine1#show ip route bgp
       Destination        Gateway                      Dist/Metric Last Change
       -----------        -------                      ----------- -----------
  B EX 10.0.1.2/32        via 192.168.1.1                    20/0    00:00:37
                          via 192.168.1.3
  B EX 10.0.2.1/32        via 192.168.1.1                    20/0    00:00:37
  B EX 10.0.2.2/32        via 192.168.1.3                    20/0    00:03:37
  B EX 10.0.2.3/32        via 192.168.1.5                    20/0    00:03:31
  B EX 10.0.2.4/32        via 192.168.1.7                    20/0    00:03:23
  B EX 10.60.1.0/24       via 192.168.1.5                    20/0    00:03:19
                          via 192.168.1.7
  B EX 172.16.1.0/24      via 192.168.1.1                    20/0    00:00:37
                          via 192.168.1.3
```

S4048-Leaf1 has two paths to all other leafs and two paths to Server 2's network, 10.60.1.0. There is one path through each spine.

Note: If all paths do not appear, make sure the maximum-paths statement in the BGP configuration is equal to or greater than the number of spines in the topology.

```
S4048-Leaf1#show ip route bgp
       Destination        Gateway                      Dist/Metric Last Change
       -----------        -------                      ----------- -----------
  B EX 10.0.1.1/32        via 192.168.1.0                    20/0    00:03:56
  B EX 10.0.1.2/32        via 192.168.2.0                    20/0    00:07:02
  B EX 10.0.2.2/32        via 192.168.1.0                    20/0    00:03:56
                          via 192.168.2.0
  B EX 10.0.2.3/32        via 192.168.1.0                    20/0    00:03:56
                          via 192.168.2.0
  B EX 10.0.2.4/32        via 192.168.1.0                    20/0    00:03:56
                          via 192.168.2.0
  B EX 10.60.1.0/24       via 192.168.1.0                    20/0    00:03:56
                          via 192.168.2.0
```

Note: The `show ip route <cr>` command can be used to verify the information above as well as static routes and direct connections.

4.4.3.4 show ip route ospf

On switches with OSPF configured, the `show ip route ospf` command is used to verify the OSPF entries in the Routing Information Base (RIB). Entries with multiple paths shown are used with ECMP. The two server networks in this example, 10.60.1.0 and 172.16.1.0, each have two paths from Z9100-Spine1, one through each leaf.

The first set of routes with a subnet mask of /32 are the IPs configured for router IDs.

```
Z9100-Spine1#show ip route ospf
       Destination        Gateway                     Dist/Metric Last Change
       -----------        -------                     ----------- -----------
  O E2 10.0.1.2/32        via 192.168.1.3, Fo 1/2/1         110/20    16:46:28
                          via 192.168.1.5, Fo 1/3/1
  O E2 10.0.2.1/32        via 192.168.1.1, Fo 1/1/1         110/20    17:20:59
  O E2 10.0.2.2/32        via 192.168.1.3, Fo 1/2/1         110/20    17:20:59
  O E2 10.0.2.3/32        via 192.168.1.5, Fo 1/3/1         110/20    17:20:59
  O E2 10.0.2.4/32        via 192.168.1.7, Fo 1/4/1         110/20    17:20:59
  O E2 10.60.1.0/24       via 192.168.1.5, Fo 1/3/1         110/20    16:46:28
                          via 192.168.1.7, Fo 1/4/1
  O E2 172.16.1.0/24      via 192.168.1.1, Fo 1/1/1         110/20    16:46:28
                          via 192.168.1.3, Fo 1/2/1
  O    192.168.2.0/31     via 192.168.1.1, Fo 1/1/1          110/2    17:20:59
  O    192.168.2.2/31     via 192.168.1.3, Fo 1/2/1          110/2    17:20:59
  O    192.168.2.4/31     via 192.168.1.5, Fo 1/3/1          110/2    17:20:59
  O    192.168.2.6/31     via 192.168.1.7, Fo 1/4/1          110/2    17:20:59
```

S4048-Leaf1 has two paths to all other leafs and two paths to the Server 2 network, 10.60.1.0. There is one path through each spine. If all paths do not appear, make sure that the maximum-paths statement in the OSPF configuration is equal to or greater than the number of spines in the topology.

```
S4048-Leaf1#show ip route ospf
       Destination        Gateway                     Dist/Metric Last Change
       -----------        -------                     ----------- -----------
  O E2 10.0.1.1/32        via 192.168.1.0, Fo 1/49          110/20    17:30:11
  O E2 10.0.1.2/32        via 192.168.2.0, Fo 1/50          110/20    18:18:43
  O E2 10.0.2.2/32        via 192.168.1.0, Fo 1/49          110/20    17:30:11
                          via 192.168.2.0, Fo 1/50
  O E2 10.0.2.3/32        via 192.168.1.0, Fo 1/49          110/20    17:30:11
                          via 192.168.2.0, Fo 1/50
  O E2 10.0.2.4/32        via 192.168.1.0, Fo 1/49          110/20    17:30:11
                          via 192.168.2.0, Fo 1/50
  O E2 10.60.1.0/24       via 192.168.1.0, Fo 1/49          110/20    17:30:11
                          via 192.168.2.0, Fo 1/50
  O    192.168.1.2/31     via 192.168.1.0, Fo 1/49           110/2    17:30:11
  O    192.168.1.4/31     via 192.168.1.0, Fo 1/49           110/2    17:30:11
  O    192.168.1.6/31     via 192.168.1.0, Fo 1/49           110/2    17:30:11
  O    192.168.2.2/31     via 192.168.2.0, Fo 1/50           110/2    18:18:43
```

```
O    192.168.2.4/31     via 192.168.2.0, Fo 1/50          110/2    18:18:43
O    192.168.2.6/31     via 192.168.2.0, Fo 1/50          110/2    18:18:43
```

Note: The `show ip route <cr>` command can be used to verify the information above as well as static routes and direct connections.

4.4.3.5 show bfd neighbors

The `show bfd neighbors` command verifies that BFD is properly configured and that the sessions are established, as indicated by `Up` in the `State` column.

Note: The output shown below is for BGP configurations as indicated by a `B` in the `Clients` column. On OSPF configurations, the output is identical except there is an `O` in the `Clients` column.

```
Z9100-Spine-1#show bfd neighbors

*       - Active session role
B       - BGP
O       - OSPF

  LocalAddr      RemoteAddr          Interface  State Rx-int Tx-int Mult Clients
* 192.168.1.0    192.168.1.1         Fo 1/1/1   Up    100    100    3    B
* 192.168.1.2    192.168.1.3         Fo 1/2/1   Up    100    100    3    B
* 192.168.1.4    192.168.1.5         Fo 1/3/1   Up    100    100    3    B
* 192.168.1.6    192.168.1.7         Fo 1/4/1   Up    100    100    3    B

S4048-Leaf-1#show bfd neighbors

*       - Active session role
B       - BGP
O       - OSPF

  LocalAddr       RemoteAddr         Interface  State Rx-int Tx-int Mult Clients
* 192.168.1.1     192.168.1.0        Fo 1/49    Up    100    100    3    B
* 192.168.2.1     192.168.2.0        Fo 1/50    Up    100    100    3    B
```

4.4.3.6 show vlt brief

The `show vlt brief` command validates the VLT configuration status on leaf switches in this topology. The Inter-chassis link (ICL) Link Status, HeartBeat Status, and VLT Peer Status must all be shown as `Up`. The role for one switch in the VLT pair is primary and its peer switch (not shown) is assigned the secondary role. Verify that Peer-Routing shows as `Enabled`.

```
S4048-Leaf-1#show vlt brief
 VLT Domain Brief
------------------
 Domain ID:                    127
 Role:                         Primary
```

```
Role Priority:                  32768
ICL Link Status:                Up
HeartBeat Status:               Up
VLT Peer Status:                Up
Local Unit Id:                  0
Version:                        6(7)
Local System MAC address:       f4:8e:38:20:37:29
Remote System MAC address:      f4:8e:38:20:54:29
Remote system version:          6(7)
Delay-Restore timer:            90 seconds
Delay-Restore Abort Threshold:  60 seconds
Peer-Routing :                  Enabled
Peer-Routing-Timeout timer:     0 seconds
Multicast peer-routing timeout: 150 seconds
```

4.4.3.7 show vlt detail

The `show vlt detail` command validates the VLT LAG status on leaf switches in this topology. This command shows the status and active VLANs of all VLT LAGs (Port channel 1 in this example). The local and peer status must both be up.

```
S4048-Leaf-1#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
1             1            UP            UP            50
```

4.4.3.8 show vlt mismatch

The `show vlt mismatch` command highlights configuration issues between VLT peers. Mismatch examples include incompatible VLT configuration settings, VLAN differences, different switch operating system versions, and spanning-tree inconsistencies.

Note: There should be no output to this command on any switch configured for VLT. If there is, resolve the mismatch.

```
S4048-Leaf1#show vlt mismatch
S4048-Leaf1#
```

4.4.3.9 show uplink-state-group

The `show uplink-state-group` command validates the UFD status on leaf switches in this topology. Status: `Enabled, Up` indicates UFD is enabled and no interfaces are currently disabled by UFD.

```
S4048-Leaf1#show uplink-state-group
Uplink State Group: 1   Status: Enabled, Up
```

If an interface happens to be disabled by UFD, the `show uplink-state-group` command output will appear as follows:

```
Uplink State Group: 1   Status: Enabled, Down
```

Note: When an interface has been disabled by UFD, the show interfaces *interface* command for affected interfaces indicates it is error-disabled as follows:

```
S4048-Leaf-1#show interfaces te 1/4
TenGigabitEthernet 1/4 is up, line protocol is down(error-disabled[UFD])
-- Output truncated --
```

4.4.3.10 show spanning-tree rstp brief

The `show spanning-tree rstp brief` command ensures that spanning tree is enabled on the leaf switches. All interfaces are forwarding (`Sts` column shows `FWD`). One of the leaf switches (S4048-Leaf1 in this example) is the root bridge and sever-facing interfaces (Po 1 in this example) are edge ports.

```
S4048-Leaf1#show spanning-tree rstp brief
Executing IEEE compatible Spanning Tree Protocol
Root ID    Priority 0, Address f48e.3820.3729
Root Bridge hello time 2, max age 20, forward delay 15
Bridge ID    Priority 0, Address f48e.3820.3729
We are the root
Configured hello time 2, max age 20, forward delay 15

Interface                                                  Designated
 Name      PortID   Prio Cost    Sts         Cost        Bridge ID        PortID
---------- -------- ---- ------- ----------- ----- ------------------- --------
Po 1       128.2    128  1800    FWD(vlt)    0     0     f48e.3820.3729 128.2
Po 127     128.128  128  600     FWD(vltI)   0     0     f48e.3820.3729 128.128

Interface
 Name      Role   PortID   Prio Cost    Sts         Cost    Link-type Edge
---------- ------ -------- ---- ------- ----------- ------- --------- ----
Po 1       Desg   128.2    128  1800    FWD         0       (vlt) P2P Yes
Po 127     Desg   128.128  128  600     FWD         0       (vltI)P2P No
```

4.5 Layer 2 switch configuration

This section provides an overview of the protocols used in constructing the leaf-spine network examples in this guide.

These protocols are used in layer 2 topology examples:

- Virtual Link Trunking (VLT)
- Uplink Failure Detection (UFD)
- Rapid Spanning Tree Protocol (RSTP)
- Link Aggregation Protocol (LACP) / Link Aggregation Group (LAG)

In layer 2 leaf-spine topologies, VLT is used at both the leaf and spine layers.

4.6 Layer 2 topology protocols

4.6.1 VLT

Virtual Link Trunking (VLT) allows link aggregation group (LAG) terminations on two separate switches and supports a loop-free topology. The two switches are referred to as VLT peers and are kept synchronized via an inter-switch link called the VLT interconnect (VLTi). A separate backup link maintains heartbeat messages across the OOB management network.

VLT provides layer 2 multipathing and load-balances traffic. VLT offers the following additional benefits:

- Eliminates spanning tree-blocked ports
- Uses all available uplink bandwidth
- Provides fast convergence if either a link or device fails
- Assures high availability

In layer 2 leaf-spine topologies, VLT is used at both the leaf and spine layers.

Note: Downstream connections from leaf switches configured for VLT do not necessarily have to be configured as LAGs if other fault tolerant methods, such as multipath IO, are preferred.

4.6.2 UFD

If a leaf switch loses all connectivity to the spine layer, by default, the attached hosts continue to send traffic to that leaf without a direct path to the destination. The VLTi link to the peer leaf switch handles traffic during such a network outage, but this is not considered a best practice.

Dell EMC recommends enabling Uplink Failure Detection (UFD), which detects the loss of upstream connectivity. An uplink-state group is configured on each leaf switch, which creates an association between the uplinks to the spines and the downlink interfaces.

In the event that all uplinks fail on a switch, UFD automatically shuts down the downstream interfaces. This propagates to the hosts attached to the leaf switch. The host then uses its link to the remaining switch to continue sending traffic across the leaf-spine network.

4.6.3 RSTP

As a precautionary measure, Dell EMC recommends enabling Rapid Spanning Tree Protocol (RSTP) on all switches that have layer 2 interfaces. Because VLT environments are loop-free, simultaneously running spanning tree is optional though considered a best practice in case of switch misconfiguration or improperly connected cables. In properly configured and connected leaf-spine networks, there are no ports blocked by spanning tree.

4.6.4 LACP/LAG

Link Aggregation Group (LAG) bundles multiple links into a single interface to increase bandwidth between two devices. LAGs also provide redundancy via the multiple paths. In a leaf-spine network, LAGs are typically used to attach servers or storage devices to the VLT leaf pairs.

Link Aggregation Control Protocol (LACP) is an improvement over static LAGs in that the protocol will automatically failover if there is a connectivity issue. This is especially important if the links traverse a media converter where it is possible to lose Ethernet connectivity while links remain in an Up state.

4.7 Layer 2 configuration with Dell EMC leaf and spine switches

This section provides configuration information to build the layer 2 leaf-spine topology shown in Figure 12. Dell EMC Networking S4048-ON switches are used at the leaf layer and Dell EMC Networking S6010-ON switches are used at the spine layer.

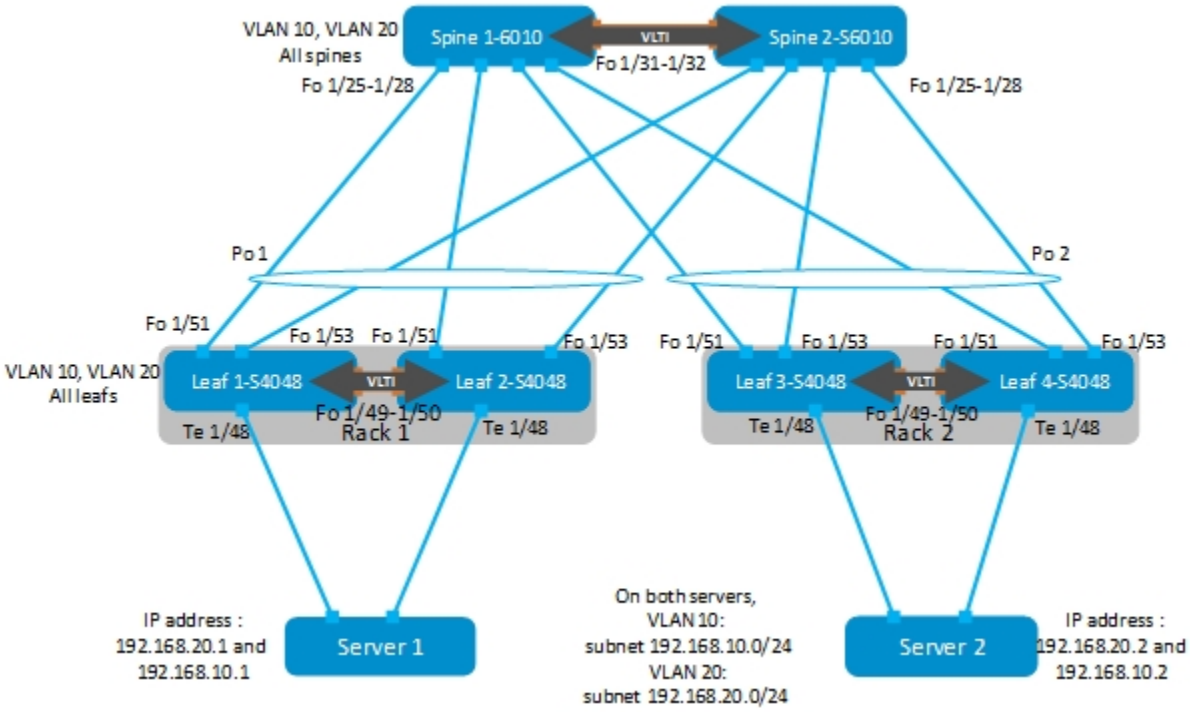


Figure 12 Layer 2 leaf-spine topology with Dell EMC leaf and spine switches

4.7.1 S4048-ON leaf switch configuration

The following sections outline the configuration commands issued to the S4048-ON leaf switches to build the topology in Figure 12. The commands detailed below are for L2-Leaf1-S4048 and L2-Leaf2-S4048. The configuration commands for L2-Leaf3-S4048 and L2-Leaf4-S4048 are similar.

Note: On S4048-ON switches, Telnet is enabled and SSH is disabled by default. Both services require the creation of a non-root user account to login. If needed, it is a best practice to use SSH instead of Telnet for security. Optionally, SSH can be enabled using the following command:
(conf)#ip ssh server enable
A user account can be created to access the switch via SSH using the following command:
(conf)#username *ssh_user* sha256-password *ssh_password*

Configure the serial console, enable the password, and disable Telnet.

Table 20 Serial console, password, and Telnet configuration commands

L2-Leaf1-S4048	**L2-Leaf2-S4048**

`enable`
`configure`
 	
`enable sha256-password` *`enable_password`*
`no ip telnet server enable`
	`enable`
	`configure`
	
	`enable sha256-password` *`enable_password`*
	`no ip telnet server enable`

1. Use the following commands to:
   a. Set the hostname
   b. Configure the OOB management interface and default gateway
   c. Enable LLDP
   d. Enable RSTP as a precaution

Note: In this layer 2 topology, the RSTP root bridge is configured at the spine level.

Table 21 Hostname, management, LLDP, and RSTP configuration commands

L2-Leaf1-S4048	L2-Leaf2-S4048
hostname L2-Leaf1-S4048 interface ManagementEthernet 1/1 ip address 100.67.194.1/24 no shutdown management route 0.0.0.0/0 100.67.194.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc protocol spanning-tree rstp no disable	hostname L2-Leaf2-S4048 interface ManagementEthernet 1/1 ip address 100.67.194.2/24 no shutdown management route 0.0.0.0/0 100.67.194.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc protocol spanning-tree rstp no disable

2. Configure the VLT interconnect between Leaf1 and Leaf2. In this configuration, add interfaces fortyGigE 1/49-50 to static port channel 127 for the VLT interconnect. The backup destination is the management IP address of the VLT peer switch.

Note: Dell EMC recommends that the VLTi is configured as a static LAG (without LACP) per the commands shown below.

Table 22 VLT interconnection configuration commands

L2-Leaf1-S4048	L2-Leaf2-S4048
interface Port-channel 127 description VLTi Port-Channel no ip address channel-member fortyGigE 1/49,1/50 no shutdown interface range fortyGigE 1/49 - 1/50 description VLTi no ip address no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.194.2 unit-id 0	interface Port-channel 127 description VLTi Port-Channel no ip address channel-member fortyGigE 1/49,1/50 no shutdown interface range fortyGigE 1/49 - 1/50 description VLTi no ip address no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.194.1 unit-id 1

3. Verify that interface Te 1/48 connects downstream to Server 1 and is configured as an RSTP edge port.
4. Ensure that interfaces Fo 1/51 and Fo 1/53 connect to the spines upstream and are configured in LACP port channel 1. The port channel is configured for VLT.

Table 23 Interface connection commands

L2-Leaf1-S4048	L2-Leaf2-S4048
interface TenGigabitEthernet 1/48 description Server 1 no ip address portmode hybrid switchport spanning-tree rstp edge-port no shutdown interface fortyGigE 1/51 description Spine1-Port25 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/53 description Spine2-Port25 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface Port-channel 1 description To Spines no ip address portmode hybrid switchport vlt-peer-lag port-channel 1 no shutdown	interface TenGigabitEthernet 1/48 description Server 1 no ip address portmode hybrid switchport spanning-tree rstp edge-port no shutdown interface fortyGigE 1/51 description Spine1-Port26 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/53 description Spine2-Port26 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface Port-channel 1 description To Spines no ip address portmode hybrid switchport vlt-peer-lag port-channel 1 no shutdown

5. Configure VLANs 10 and 20 on each switch and verify that `Port-channel 1` is tagged in both VLANs.

Note: The shutdown/no shutdown commands on a VLAN have no effect unless the VLAN is assigned an IP address, (configured as an SVI).

Table 24 VLAN configuration command

L2-Leaf1-S4048	L2-Leaf2-S4048
interface Vlan 10 no ip address tagged TenGigabitEthernet 1/48 tagged Port-channel 1 shutdown interface Vlan 20 no ip address tagged TenGigabitEthernet 1/48 tagged Port-channel 1 shutdown	interface Vlan 10 no ip address tagged TenGigabitEthernet 1/48 tagged Port-channel 1 shutdown interface Vlan 20 no ip address tagged TenGigabitEthernet 1/48 tagged Port-channel 1 shutdown

6. Use the following commands to configure UFD to shut down the downstream interfaces if all uplinks fail. The hosts attached to the switch use the remaining LACP port member to continue sending traffic across the fabric.
7. Use the `end` and `write` commands to exit configuration mode and save the configuration.

Table 25 UFD configuration commands

L2-Leaf1-S4048	L2-Leaf2-S4048
uplink-state-group 1 description Disable all edge ports in event all spines uplinks fail downstream TenGigabitEthernet 1/1-1/48 upstream Port-channel 1 end write	uplink-state-group 1 description Disable all edge ports in event all spines uplinks fail downstream TenGigabitEthernet 1/1-1/48 upstream Port-channel 1 end write

4.7.2 S6010-ON spine configuration

The following sections outline the configuration commands issued to the S6010-ON spine switches to build the topology in Figure 12.

Note: On S6010-ON switches, Telnet is enabled and SSH is disabled by default. Both services require the creation of a non-root user account to login. If needed, it is a best practice to use SSH instead of Telnet for security. SSH can optionally be enabled with the command:

```
(conf)#ip ssh server enable
```

A user account can be created to access the switch via SSH with the command:

```
(conf)#username ssh_user sha256-password ssh_password
```

1. Configure the serial console, enable password, and disable Telnet.

Table 26 Serial console, password, and Telnet configuration commands

L2-Spine1-S6010	L2-Spine2-S6010
enable configure enable sha256-password *enable_password* no ip telnet server enable	enable configure enable sha256-password *enable_password* no ip telnet server enable

2. Using the commands in the following table:
   a. Set the hostname
   b. Configure the OOB management interface and default gateway
   c. Enable LLDP
   d. Enable RSTP as a precaution

Note: L2-Spine1-S6010 is configured as the primary RSTP root bridge using the `bridge-priority 0` command. L2-Spine2-S6010 is configured as the secondary RSTP root bridge using the `bridge-priority 4096` command.

Table 27 Hostname, management, LLDP, and RSTP configuration commands

L2-Spine1-S6010	L2-Spine2-S6010
hostname L2-Spine1-S6010 interface ManagementEthernet 1/1 ip address 100.67.194.15/24 no shutdown management route 0.0.0.0/0 100.67.194.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc protocol spanning-tree rstp no disable bridge-priority 0	hostname L2-Spine2-S6010 interface ManagementEthernet 1/1 ip address 100.67.194.16/24 no shutdown management route 0.0.0.0/0 100.67.194.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc protocol spanning-tree rstp no disable bridge-priority 4096

3. Configure the VLT interconnect between Spine1 and Spine2. In this configuration, add interfaces fortyGigE 1/31-32 to static port channel 127 for the VLT interconnect. The backup destination is the management IP address of the VLT peer switch.

Note: Dell EMC recommends that the VLTi is configured as a static LAG, without LACP, per the commands shown below.

Table 28 VLT configuration commands

L2-Spine1-S6010	L2-Spine2-S6010
interface Port-channel 127 description VLTi Port-Channel no ip address channel-member fortyGigE 1/31,1/32 no shutdown interface range fortyGigE 1/31 - 1/32 description VLTi no ip address no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.194.16 unit-id 0	interface Port-channel 127 description VLTi Port-Channel no ip address channel-member fortyGigE 1/31,1/32 no shutdown interface range fortyGigE 1/31 - 1/32 description VLTi no ip address no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.194.15 unit-id 1

4. Verify that interfaces Fo 1/25-28 connect to the leaf switches downstream via LACP port channels.
5. Ensure that port-channel 1 shows members Fo 1/25 and Fo 1/26, and that port channel 2 shows members Fo 1/27 and Fo 1/28. The port channels are configured for VLT.

Table 29 Interface connection commands

L2-Spine1-S6010	L2-Spine2-S6010
interface fortyGigE 1/25 description Leaf1-Port51 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/26 description Leaf2-Port51 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/27 description Leaf3-Port51 no ip address port-channel-protocol LACP port-channel 2 mode active no shutdown interface fortyGigE 1/28 description Leaf4-Port51	interface fortyGigE 1/25 description Leaf1-Port53 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/26 description Leaf2-Port53 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/27 description Leaf3-Port53 no ip address port-channel-protocol LACP port-channel 2 mode active no shutdown interface fortyGigE 1/28 description Leaf4-Port53

no ip address port-channel-protocol LACP port-channel 2 mode active no shutdown interface Port-channel 1 description Leaf 1 & 2 no ip address portmode hybrid switchport vlt-peer-lag port-channel 1 no shutdown interface Port-channel 2 description Leaf 3 & 4 no ip address portmode hybrid switchport vlt-peer-lag port-channel 2 no shutdown	no ip address port-channel-protocol LACP port-channel 2 mode active no shutdown interface Port-channel 1 description Leaf 1 & 2 no ip address portmode hybrid switchport vlt-peer-lag port-channel 1 no shutdown interface Port-channel 2 description Leaf 3 & 4 no ip address portmode hybrid switchport vlt-peer-lag port-channel 2 no shutdown

6. Verify that VLANs 10 and 20 are configured on each switch and that port-channels 1 and 2 are tagged in both VLANs.

Note: The shutdown/no shutdown commands on a VLAN have no effect unless the VLAN is assigned an IP address, configured as an SVI.

7. Use the end and `write` commands to exit configuration mode and save the configuration.

Table 30 VLAN configuration commands

L2-Spine1-S6010	**L2-Spine2-S6010**
interface Vlan 10 no ip address tagged Port-channel 1-2 shutdown interface Vlan 20 no ip address tagged Port-channel 1-2 shutdown end write	interface Vlan 10 no ip address tagged Port-channel 1-2 shutdown interface Vlan 20 no ip address tagged Port-channel 1-2 shutdown end write

4.7.3 Layer 2 example validation

In addition to sending traffic between hosts, the configuration shown in Figure 12 can be validated with the commands shown in this section. For more information on commands and output, see the *Command Line Reference Guide* for the applicable switch (links to documentation are provided in Appendix F).

Command and output examples are provided for one spine and one leaf. Command output on other switches is similar.

4.7.3.1 show vlt brief

The Inter-chassis link (ICL) Link Status, HeartBeat Status and VLT Peer Status must all be up. The role for one switch in the VLT pair is primary and its peer switch (not shown) is assigned the secondary role.

```
L2-Spine1-S6010#show vlt brief
 VLT Domain Brief
------------------
 Domain ID:                    127
 Role:                         Primary
 Role Priority:                32768
 ICL Link Status:              Up
 HeartBeat Status:             Up
 VLT Peer Status:              Up
 Local Unit Id:                0
 Version:                      6(7)
 Local System MAC address:     f4:8e:38:2b:08:69
 Remote System MAC address:    f4:8e:38:2b:36:e9
 Remote system version:        6(7)
 Delay-Restore timer:          90 seconds
 Delay-Restore Abort Threshold:  60 seconds
 Peer-Routing :                Disabled
 Peer-Routing-Timeout timer:   0 seconds
 Multicast peer-routing timeout: 150 seconds

L2-Leaf1-S4048#show vlt brief
 VLT Domain Brief
------------------
 Domain ID:                    127
 Role:                         Primary
 Role Priority:                32768
 ICL Link Status:              Up
 HeartBeat Status:             Up
 VLT Peer Status:              Up
 Local Unit Id:                0
 Version:                      6(7)
 Local System MAC address:     f4:8e:38:20:c5:29
 Remote System MAC address:    64:00:6a:e6:cc:14
 Remote system version:        6(7)
```

```
Delay-Restore timer:            90 seconds
Delay-Restore Abort Threshold:  60 seconds
Peer-Routing :                  Disabled
Peer-Routing-Timeout timer:     0 seconds
Multicast peer-routing timeout: 150 seconds
```

4.7.3.2 show vlt detail

The `show vlt detail` command shows the status and active VLANs of all VLT LAGs (port channels 1 and 2 in this example). The local and peer status must both be up.

```
L2-Spine1-S6010#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  ----------  ------------  -----------  -------------
1             1           UP            UP            1, 10, 20
2             2           UP            UP            1, 10, 20

L2-Leaf1-S4048#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  ----------  ------------  -----------  -------------
1             1           UP            UP            1, 10, 20
```

4.7.3.3 show vlt mismatch

The `show vlt mismatch` command highlights configuration issues between VLT peers. Mismatch examples include incompatible VLT configuration settings, VLAN differences, different switch operating system versions and spanning-tree inconsistencies.

Note: There should be no output to this command on any switch configured for VLT. If there is, resolve the mismatch.

```
L2-Spine1-S6010#show vlt mismatch
L2-Spine1-S6010#

L2-Leaf1-S4048#show vlt mismatch
L2-Leaf1-S4048#
```

4.7.3.4 show uplink-state-group

The `show uplink-state-group` command is used to validate the UFD status on leaf switches in this topology. `Status: Enabled, Up` indicates UFD is enabled and no interfaces are currently disabled by UFD.

```
S4048-Leaf1#show uplink-state-group
Uplink State Group: 1   Status: Enabled, Up
```

If an interface happens to be disabled by UFD, the `show uplink-state-group` command output will appear as follows:

```
Uplink State Group: 1   Status: Enabled, Down
```

Note: When an interface has been disabled by UFD, the show interfaces interface command for affected interfaces indicates it is error-disabled as follows:

```
S4048-Leaf-1#show interfaces te 1/48
TenGigabitEthernet 1/48 is up, line protocol is down(error-disabled[UFD])
-- Output truncated --
```

4.7.3.5 show spanning-tree rstp brief

The `show spanning-tree rstp brief` command validates spanning tree is enabled. All interfaces are forwarding (`Sts` column shows `FWD`) because VLT is configured at the leaf and spine layers, eliminating the need for blocked ports. One of the spine switches (L2-Spine1-S6010 in this example) is the root bridge. Sever-facing interfaces on leaf switches (L2-Leaf1-S4048 interface Te 1/48 in this example) are edge ports.

```
L2-Spine1-S6010#show spanning-tree rstp brief
Executing IEEE compatible Spanning Tree Protocol
Root ID    Priority 0, Address f48e.382b.0869
Root Bridge hello time 2, max age 20, forward delay 15
Bridge ID    Priority 0, Address f48e.382b.0869
We are the root
Configured hello time 2, max age 20, forward delay 15

Interface                                                  Designated
 Name      PortID   Prio Cost    Sts         Cost        Bridge ID        PortID
---------- -------- ---- ------- ----------- ----- -------------------- --------
Po 1       128.2    128  188     FWD(vlt)    0     0     f48e.382b.0869 128.2
Po 2       128.3    128  188     FWD(vlt)    0     0     f48e.382b.0869 128.3
Po 127     128.128  128  600     FWD(vltI)   0     0     f48e.382b.0869 128.128

Interface
 Name      Role   PortID   Prio Cost    Sts         Cost    Link-type Edge
---------- ------ -------- ---- ------- ----------- ------- --------- ----
Po 1       Desg   128.2    128  188     FWD         0       (vlt) P2P No
Po 2       Desg   128.3    128  188     FWD         0       (vlt) P2P No
Po 127     Desg   128.128  128  600     FWD         0       (vltI)P2P No
```

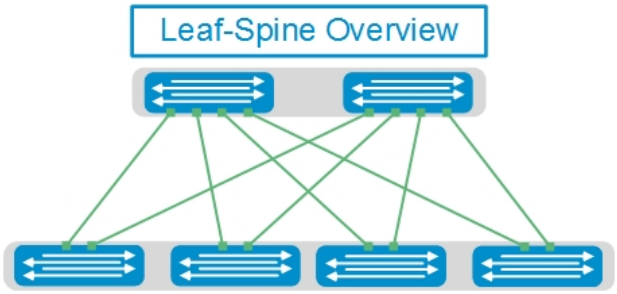
```
L2-Leaf1-S4048#show spanning-tree rstp brief
Executing IEEE compatible Spanning Tree Protocol
Root ID    Priority 0, Address f48e.382b.0869
Root Bridge hello time 2, max age 20, forward delay 15
Bridge ID    Priority 32768, Address f48e.3820.c529
Configured hello time 2, max age 20, forward delay 15

Interface                                                  Designated
 Name      PortID   Prio Cost    Sts          Cost        Bridge ID         PortID
---------- -------- ---- ------- ----------- ----- -------------------- --------
Po 1       128.2    128  188     FWD(vlt)    788   32768 f48e.3820.c529 128.2
Po 127     128.128  128  600     FWD(vltI)   788   32768 6400.6ae6.cc14 128.128
Te 1/48    128.249  128  2000    FWD         788   32768 f48e.3820.c529 128.249

Interface
 Name      Role   PortID   Prio Cost    Sts          Cost    Link-type Edge
---------- ------ -------- ---- ------- ----------- ------- --------- ----
Po 1       Root   128.2    128  188     FWD          788     (vlt) P2P No
Po 127     Root   128.128  128  600     FWD          788     (vltI)P2P No
Te 1/48    Desg   128.249  128  2000    FWD          788     P2P       Yes
```

5 Leaf-Spine architecture with Cisco spine and Dell EMC leaf

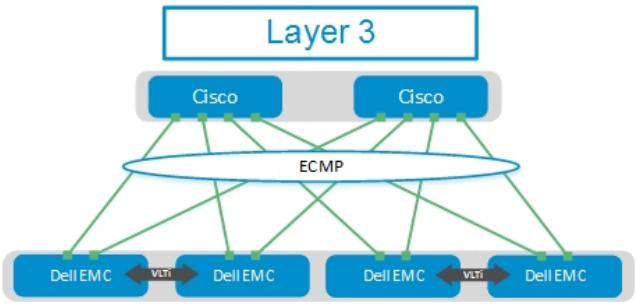
Connecting Dell Networking leaf switches to a Cisco spine will integrate similarly. In some cases this will be a brownfield installation and conforming to existing protocols and numbering plans is required. The same benefits exist as with Dell Networking spines.

This section includes configuration information for both the layer 3 and layer 2 topologies.

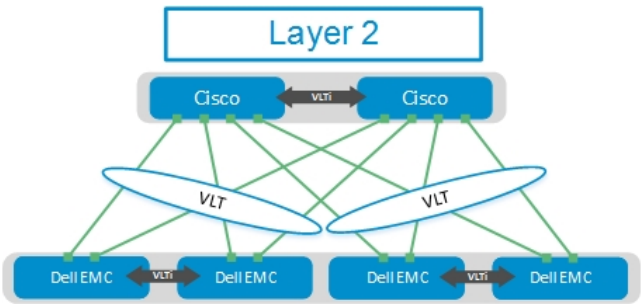
Use the following hyperlinks to navigate to the appropriate sections:



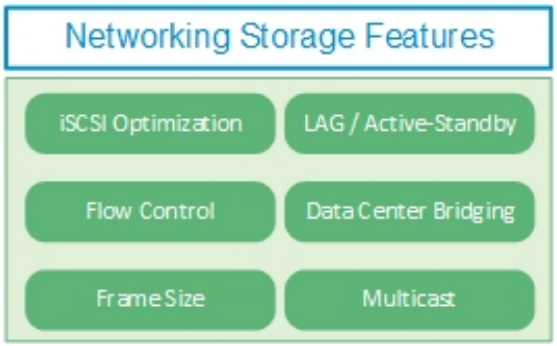
Overview of leaf-spine architecture



Layer 3 topology



Layer 2 topology



Networking features and guidelines for storage

5.1 Layer 3 switch configuration

This section provides an overview of the protocols used in constructing the leaf-spine network examples in this guide.

The first three protocols are used in all layer 2 and layer 3 topology examples:

- Virtual Link Trunking (VLT)
- Uplink Failure Detection (UFD)
- Rapid Spanning Tree Protocol (RSTP)
- Link Aggregation Protocol (LACP) / Link Aggregation Group (LAG)

The remaining protocols are only used in the layer 3 topology examples:

- Routing protocols
  - Border Gateway Protocol (BGP)
  - Open Shortest Path First (OSPF)
- Bidirectional Forwarding Detection (BFD)
- Equal-cost multipath routing (ECMP)

5.2 Layer 3 topology protocols

5.2.1 VLT

Virtual Link Trunking (VLT) allows link aggregation group (LAG) terminations on two separate switches and supports a loop-free topology. The two switches are referred to as VLT peers and are kept synchronized via an inter-switch link called the VLT interconnect (VLTi). A separate backup link maintains heartbeat messages across the OOB management network.

VLT provides layer 2 multipathing and load-balances traffic. VLT offers the following additional benefits:

- Eliminates blocked ports from STP
- Uses all available uplink bandwidth
- Provides fast convergence if either a link or device fails

- Assures high availability

In layer 2 leaf-spine topologies, VLT is used at both the leaf and spine layers.

In layer 3 topologies, VLT is only used at the leaf layer. An additional feature called VLT peer routing is enabled on the leaf switches for connections to layer 3 networks. VLT peer routing:

- Enables one VLT node to act as the default gateway for its VLT peer
- Eliminates the need to use Virtual Router Redundancy Protocol (VRRP)
- Enables active-active load sharing

With peer routing enabled, traffic is routed through either VLT peer and is passed directly to the next hop without needing to traverse the VLTi.

Note: Downstream connections from leaf switches configured for VLT do not necessarily have to be configured as LAGs if other fault tolerant methods, such as multipath IO, are preferred.

5.2.2 LACP/LAG

Link Aggregation Group (LAG) bundles multiple links into a single interface to increase bandwidth between two devices. LAGs also provides redundancy via the multiple paths. In a leaf-spine network, LAGs are typically used to attach servers to the VLT leaf pairs.

Link Aggregation Control Protocol (LACP) is an improvement over static LAGs in that the protocol will automatically failover if there is a connectivity issue. This is especially important if the links traverse a media converter where it is possible to lose Ethernet connectivity while links remain in an Up state.

5.2.3 UFD

If a leaf switch loses all connectivity to the spine layer, by default the attached hosts continue to send traffic to that leaf without a direct path to the destination. The VLTi link to the peer leaf switch handles traffic during such a network outage, but this is not considered a best practice.

Dell EMC recommends enabling Uplink Failure Detection (UFD), which detects the loss of upstream connectivity. An uplink-state group is configured on each leaf switch, which creates an association between the uplinks to the spines and the downlink interfaces.

In the event all uplinks fail on a switch, UFD automatically shuts down the downstream interfaces. This propagates to the hosts attached to the leaf switch. The host then uses its link to the remaining switch to continue sending traffic across the leaf-spine network.

5.2.4 RSTP

As a precautionary measure, Dell EMC recommends enabling Rapid Spanning Tree Protocol (RSTP) on all switches that have layer 2 interfaces. Because VLT environments are loop-free, simultaneously running spanning tree is optional though considered a best practice in case of switch misconfiguration or improperly connected cables. In properly configured and connected leaf-spine networks, there are no ports blocked by spanning tree.

5.2.5 Routing protocols

Any of the following routing protocols may be used on layer 3 connections when designing a leaf-spine network:

- BGP
- OSPF

5.2.5.1 BGP

Border Gateway Protocol (BGP) may be selected for scalability and is well suited for very large networks. BGP can be configured as External BGP (EBGP) to route between autonomous systems or Internal BGP (IBGP) to route within a single autonomous system.

Layer 3 leaf-spine networks use ECMP routing. EBGP and IBGP handle ECMP differently. By default, EBGP supports ECMP without any adjustments. IBGP requires a BGP route reflector and the use of the AddPath feature to fully support ECMP. To keep configuration complexity to a minimum, Dell EMC recommends EBGP in leaf-spine fabric deployments.

BGP tracks IP reachability to the peer remote address and the peer local address. Whenever either address becomes unreachable, BGP brings down the session with the peer. To ensure fast convergence with BGP, Dell EMC recommends enabling fast fall-over with BGP. Fast fall-over terminates external BGP sessions of any directly adjacent peer if the link to reach the peer goes down without waiting for the hold-down timer to expire.

Examples using EBGP (BGPv4) are provided in the layer 3 topology examples in this guide.

5.2.6 OSPF

Open Shortest Path First, or OSPF, is an interior gateway protocol that provides routing inside an autonomous network. OSPF routers send link-state advertisements to all other routers within the same autonomous system areas. While generally more memory and CPU intensive than BGP, OSPF may offer faster convergence. OSPF is often used in smaller networks. Examples using OSPF (OSPFv2 for IPv4) are provided in the layer 3 topology examples in this guide.

5.2.7 BFD

Bidirectional Forwarding Detection (BFD) is a protocol used to rapidly detect communication failures between two adjacent systems over a layer 3 link. It is a simple and lightweight replacement for existing routing protocol link state detection mechanisms. Though optional, use of BFD is considered a best practice for optimizing a leaf-spine network.

BFD provides forwarding path failure detection times on the order of milliseconds rather than seconds as with conventional routing protocols. It is independent of routing protocols and provides a consistent method of failure detection when used across a network. Networks converge faster because BFD triggers link state changes in the routing protocol sooner and more consistently.

Dell EMC Networking has implemented BFD at layer 3 with user datagram protocol (UDP) encapsulation. BFD is supported with routing protocols including BGP, and OSPF.

5.2.8 ECMP

The nature of a leaf-spine topology is that leaf switches are no more than one hop away from each other. As shown in Figure 13, Leaf 1 has two equal cost paths to Leaf 4, one through each spine. The same is true for all leafs.

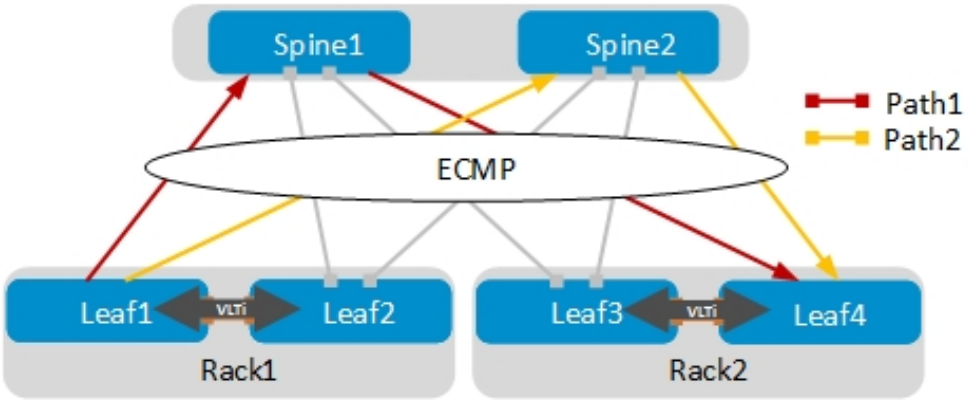


Figure 13 Use of ECMP in a layer 3 topology

Equal-cost multipath routing, or ECMP, is a routing technique used in a layer 3 leaf-spine topology for load balancing packets along these multiple equal cost paths. ECMP is enabled on all leaf and spine switches, allowing traffic between leafs to be load balanced across the spines.

5.3 Layer 3 configuration planning

5.3.1 BGP ASN configuration

When EBGP is used, an autonomous system number (ASN) is assigned to each switch. Valid private, 2-byte ASNs range from 64512 through 65534. Figure 14 shows the ASN assignments used for leaf and spine switches in the BGP examples in this guide.

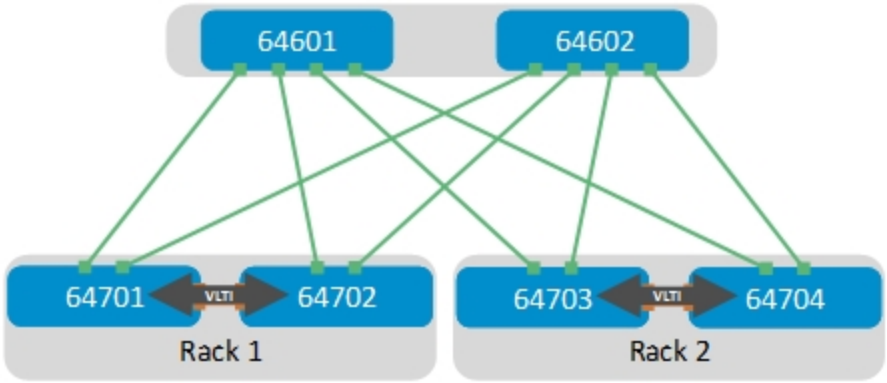


Figure 14 BGP ASN assignments

ASNs should follow a logical pattern for ease of administration and allow for growth as additional leaf and spine switches are added. In this example, an ASN with a "6" in the hundreds place, such as 64601, represents a spine switch and an ASN with a "7" in the hundreds place, such as 64701, represents a leaf switch.

5.3.2 IP addressing

Establishing a logical, scalable IP address scheme is important before deploying a leaf-spine topology. This section covers the IP addressing used in the layer 3 examples in this guide.

5.3.2.1 Loopback addresses

Loopback addresses may be used as router IDs when configuring routing protocols. As with ASNs, loopback addresses should follow a logical pattern that will make it easier for administrators to manage the network and allow for growth. Figure 15 shows the loopback addresses used as router IDs in the BGP and OSPF examples in this guide.

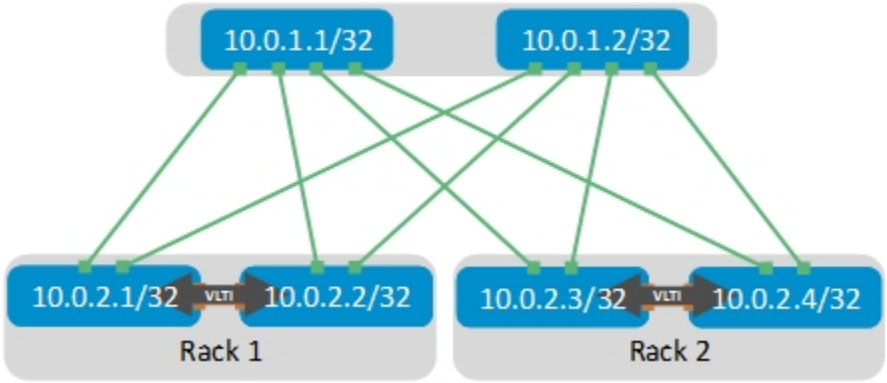


Figure 15 Loopback addressing

All of the loopback addresses used are part of the 10.0.0.0/8 address space with each address using a 32-bit mask. In this example, the third octet represents the layer, “1” for spine and “2” for leaf. The fourth octet is the

counter for the appropriate layer. For example, 10.0.1.1/32 is the first spine switch in the topology while 10.0.2.4/32 is the fourth leaf switch.

5.3.2.2 Point-to-point addresses

Table 2 lists layer 3 connection details for each leaf and spine switch.

All addresses come from the same base IP prefix, 192.168.0.0/16 with the third octet representing the spine number. For example, 192.168.1.0/31 is a two host subnet connected to Spine 1 while 192.168.2.0/31 is connected to Spine 2. This IP scheme is easily extended as leaf and spine switches are added to the network.

Link labels are provided in the table for quick reference with Figure 16.

Table 31 Interface and IP configuration

Link Label	Source switch	Source interface	Source IP	Network	Destination switch	Destination interface	Destination IP
A	Leaf 1	fo1/49	.1	192.168.1.0/31	Spine 1	fo1/1/1	.0
B	Leaf 1	fo1/50	.1	192.168.2.0/31	Spine 2	fo1/1/1	.0
C	Leaf 2	fo1/49	.3	192.168.1.2/31	Spine 1	fo1/2/1	.2
D	Leaf 2	fo1/50	.3	192.168.2.2/31	Spine 2	fo1/2/1	.2
E	Leaf 3	fo1/49	.5	192.168.1.4/31	Spine 1	fo1/3/1	.4
F	Leaf 3	fo1/50	.5	192.168.2.4/31	Spine 2	fo1/3/1	.4
G	Leaf 4	fo1/49	.7	192.168.1.6/31	Spine 1	fo1/4/1	.6
H	Leaf 4	fo1/50	.7	192.168.2.6/31	Spine 2	fo1/4/1	.6

The point-to-point IP addresses used in this guide are shown in Figure 16:

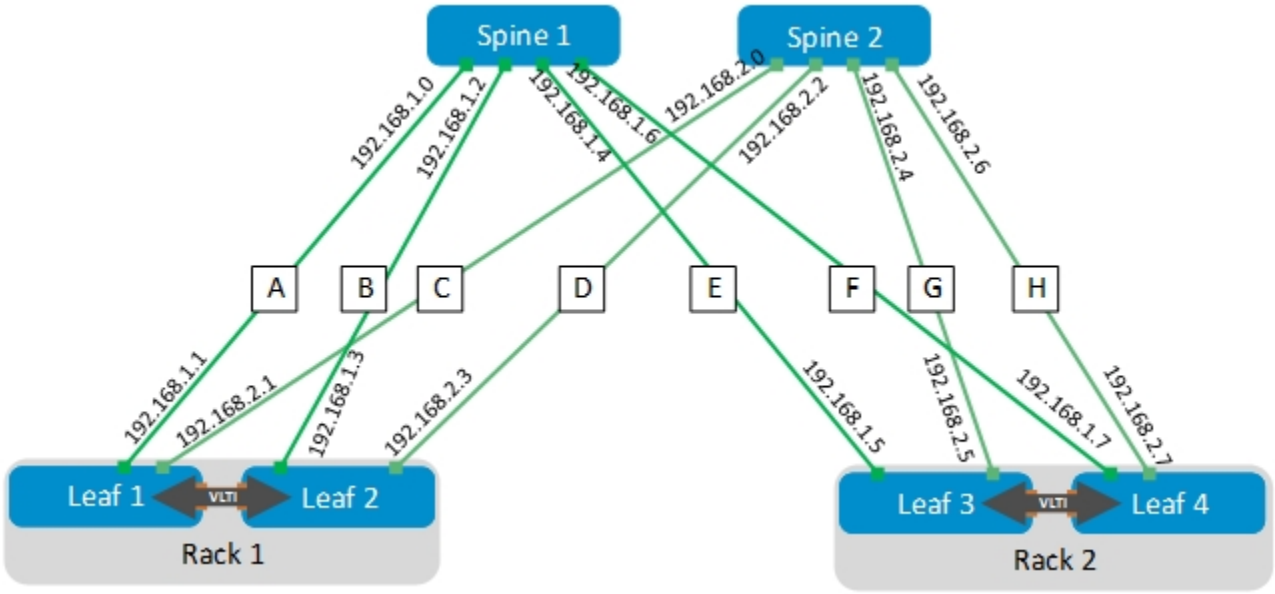


Figure 16 Point-to-point IP addresses

Note: The example point-to-point addresses use a 31-bit mask to save address space. This is optional and covered in RFC 3021. Below is an example when setting an IP address with a 31-bit mask on a Dell EMC S4048-ON. The warning message can be safely ignored on point-to-point interfaces:

```
S4048-Leaf-1(conf-if-fo-1/49)#ip address 192.168.1.1/31
% Warning: Use /31 mask on non point-to-point interface cautiously.
```

5.4 Layer 3 with Dell EMC leaf and Cisco Nexus spine switches

In this section, the Dell EMC Networking Z9100-ON spines used in the previous example are replaced with Cisco Nexus 5600 series spines as shown in Figure 17. BGP and OSPF configuration examples are included. S4048-ON leaf switch configuration is identical to that covered in section 4.4.1 and is not repeated in this section.

Note: The BGP ASNs and IP addresses defined in section 5.3.1 are used here.

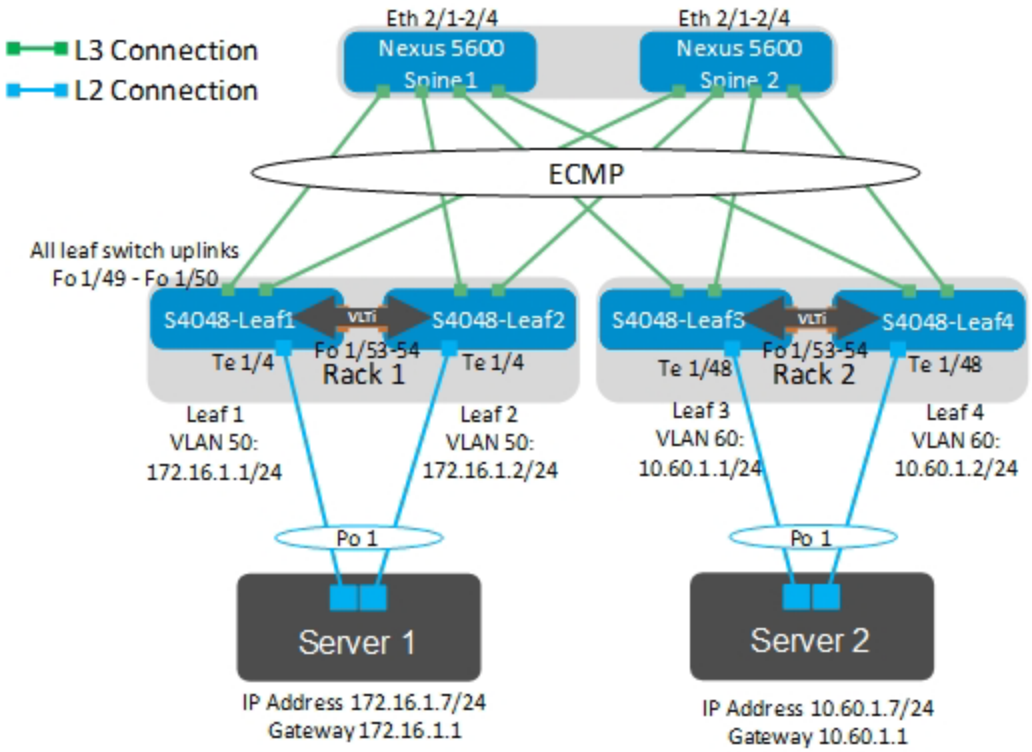


Figure 17 Layer 3 leaf-spine topology with Dell EMC leaf and Cisco Nexus spine switches

Note: Cisco Nexus switches in this example were reset to their factory default configurations by running write erase followed by reload. After reload, "Power on Auto Provisioning" was not used, the admin password was configured and the Nexus “basic configuration dialog” was not used. Refer to your Nexus system documentation for more information.

5.4.1 Nexus 5600 series spine switch configuration

The following configuration details are for Nexus5600-Spine1 and Nexus5600-Spine2 in Figure 17.

1. Use the following commands to:
   a. Set the hostname
   b. Enable LLDP and disable switchport as the default port type
   c. Configure the management interface and default management route

Table 32 Hostname, LLDP, and management interface configuration commands

Nexus5600-Spine1	Nexus5600-Spine2

`enable`
`configure`
 	
`hostname Nexus5600-Spine1`
`feature lldp`
 	
`no system default switchport`
 	
`interface mgmt0`
`vrf member management`
`ip address 100.67.219.34/24`
 	
`vrf context management`
`ip route 0.0.0.0/0 100.67.219.254`
	`enable`
	`configure`
	
	`hostname Nexus5600-Spine2`
	`feature lldp`
	
	`no system default switchport`
	
	`interface mgmt0`
	`vrf member management`
	`ip address 100.67.219.33/24`
	
	`vrf context management`
	`ip route 0.0.0.0/0 100.67.219.254`

2. Configure the four point-to-point interfaces connected to leaf switches.
3. Assign IP addresses per Table 31.

Note: Replace destination interfaces Fo 1/1/1-1/4/1 in Table 31 with Nexus interfaces Ethernet 2/1-2/4.

4. Configure a loopback interface to be used as the router ID. This is used with BGP or OSPF.
5. Use the `end` and `copy running-config startup-config` commands to exit configuration mode and save the configuration.

Table 33 Point-to-point and loopback configuration commands

Nexus5600-Spine1	Nexus5600-Spine2
interface ethernet 2/1 description Leaf 1 fo1/49 ip address 192.168.1.0/31 no shutdown interface ethernet 2/2 description Leaf 2 fo1/49 ip address 192.168.1.2/31 no shutdown interface ethernet 2/3 description Leaf 3 fo1/49 ip address 192.168.1.4/31 no shutdown interface ethernet 2/4 description Leaf 4 fo1/49 ip address 192.168.1.6/31 no shutdown interface loopback 0 description Router ID ip address 10.0.1.1/32 no shutdown end copy running-config startup-config	interface ethernet 2/1 description Leaf 1 fo1/50 ip address 192.168.2.0/31 no shutdown interface ethernet 2/2 description Leaf 2 fo1/50 ip address 192.168.2.2/31 no shutdown interface ethernet 2/3 description Leaf 3 fo1/50 ip address 192.168.2.4/31 no shutdown interface ethernet 2/4 description Leaf 4 fo1/50 ip address 192.168.2.6/31 no shutdown interface loopback 0 description Router ID ip address 10.0.1.2/32 no shutdown end copy running-config startup-config

5.4.1.1 Nexus 5600 series BGP configuration

Use these commands in this section to configure BGP and BFD.

Note: If OSPF is used, skip to section 5.4.1.2.

1. Use the commands in Table 34 to enable the BGP and BFD features.

Note: After running the feature bfd command, the message Please disable the ICMP redirects on all interfaces running BFD sessions using the command 'no ip redirects' may be displayed. This is done in the subsequent commands.

2. Run the `no ip redirects` command on the interfaces that will run BFD. BGP is enabled with the `router bgp` *ASN* command. The ASN is from Figure 14.
3. Use the `bestpath as-path multipath-relax` command to ECMP and the `maximum-paths 2` command to specify the maximum number of parallel paths to a destination to add to the routing table. In this topology, there are two equal cost best paths from a spine to a host, one to each leaf that the host is connected.

Note: BGP neighbors are configured and BFD is enabled for each neighbor connection.

4. Use the `end` and copy `running-config startup-config` commands to exit configuration mode and save the configuration.

Note: On Nexus 5600 series switches, `BGP graceful-restart`, `fast-external-fallover` and BFD interval commands are configured by default.

Table 34 BGP and BFD configuration commands

Nexus5600-Spine1	Nexus5600-Spine2
enable configure feature bgp feature bfd interface ethernet 2/1-4 no ip redirects router bgp 64601 bestpath as-path multipath-relax address-family ipv4 unicast maximum-paths 2 neighbor 192.168.1.1 remote-as 64701 address-family ipv4 unicast bfd neighbor 192.168.1.3 remote-as 64702 address-family ipv4 unicast bfd neighbor 192.168.1.5 remote-as 64703 address-family ipv4 unicast bfd neighbor 192.168.1.7 remote-as 64704 address-family ipv4 unicast bfd end copy running-config startup-config	enable configure feature bgp feature bfd interface ethernet 2/1-4 no ip redirects router bgp 64602 bestpath as-path multipath-relax address-family ipv4 unicast maximum-paths 2 neighbor 192.168.2.1 remote-as 64701 address-family ipv4 unicast bfd neighbor 192.168.2.3 remote-as 64702 address-family ipv4 unicast bfd neighbor 192.168.2.5 remote-as 64703 address-family ipv4 unicast bfd neighbor 192.168.2.7 remote-as 64704 address-family ipv4 unicast bfd end copy running-config startup-config

5.4.1.2 Nexus 5600 series OSPF configuration

Use these commands to configure OSPF and BFD. Skip this section if BGP is used.

First, enable the OSPF and BFD features.

Note: After running the feature bfd command, the following message may be displayed:
Please disable the ICMP redirects on all interfaces running BFD sessions using the command 'no ip redirects'. This is done in the subsequent commands.

OSPF is enabled with the `router ospf process-tag` command.

The `maximum-paths 2` command enables ECMP and specifies the maximum number of parallel paths to a destination to add to the routing table. In this topology, there are two equal cost best paths from a spine to a host, one to each leaf that the host is connected.

Run the `no ip redirects` command on the interfaces that will run BFD. Add the interfaces connected to the leaf switches to OSPF area 0. Enable BFD on the interfaces.

Finally, exit configuration mode and save the configuration with the `end` and `copy running-config startup-config` commands.

Table 35 OSPF and BFD commands

Nexus5600-Spine1	Nexus5600-Spine2
enable configure feature ospf feature bfd router ospf 1 log-adjacency-changes maximum-paths 2 interface ethernet 2/1-4 no ip redirects ip router ospf 1 area 0 ip ospf bfd end copy running-config startup-config	enable configure feature ospf feature bfd router ospf 1 log-adjacency-changes maximum-paths 2 interface ethernet 2/1-4 no ip redirects ip router ospf 1 area 0 ip ospf bfd end copy running-config startup-config

5.4.2 Validation

In addition to sending traffic between hosts, the configuration shown in Figure 17 can be validated with the commands shown in this section. For more information on commands and output, see the *Command Line Reference Guide* for the applicable switch (links to documentation are provided in Appendix F).

Command and output examples are provided for one spine and one leaf. Command output on other switches is similar.

5.4.2.1 show ip bgp summary

When BGP is configured, the `show ip bgp summary` command shows the status of all BGP connections. Each spine has four neighbors (the four leafs) and each leaf has two neighbors (the two spines). On Dell EMC switches, this command also confirms BFD is enabled on the 6th line of output.

```
Nexus5600-Spine1# show ip bgp summary
BGP summary information for VRF default, address family IPv4 Unicast
BGP router identifier 10.0.1.1, local AS number 64601
BGP table version is 59, IPv4 Unicast config peers 4, capable peers 4
6 network entries and 8 paths using 1024 bytes of memory
BGP attribute entries [4/576], BGP AS path entries [4/24]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
192.168.1.1     4 64701      89      91       59    0    0 00:46:58 2
192.168.1.3     4 64702      90      91       59    0    0 00:51:12 2
192.168.1.5     4 64703      91      94       59    0    0 00:47:07 2
192.168.1.7     4 64704      92      91       59    0    0 00:51:07 2

S4048-Leaf1#show ip bgp summary
BGP router identifier 10.0.2.1, local AS number 64701
BGP local RIB : Routes to be Added 0, Replaced 0, Withdrawn 0
6 network entrie(s) using 456 bytes of memory
11 paths using 1188 bytes of memory
BGP-RIB over all using 1199 bytes of memory
BFD is enabled, Interval 100 Min_rx 100 Multiplier 3 Role Active
13 BGP path attribute entrie(s) using 2064 bytes of memory
11 BGP AS-PATH entrie(s) using 110 bytes of memory
2 neighbor(s) using 16384 bytes of memory

Neighbor        AS            MsgRcvd  MsgSent     TblVer  InQ  OutQ Up/Down
State/Pfx
192.168.1.0     64601              54       58          0    0     0 00:47:30 5
192.168.2.0     64602              59       66          0    0     0 00:33:00 4
```

5.4.2.2 show ip ospf neighbor

When OSPF is configured, the `show ip ospf neighbor` command shows the state of all connected OSPF neighbors. In this configuration, each spine has four neighbors (the four leafs) and each leaf has two neighbors (the two spines).

```
Nexus5600-Spine1# show ip ospf neighbor
 OSPF Process ID 1 VRF default
 Total number of neighbors: 4
 Neighbor ID     Pri State            Up Time  Address         Interface
 10.0.2.1          1 FULL/DR          00:22:25 192.168.1.1     Eth2/1
 10.0.2.2          1 FULL/DR          00:22:05 192.168.1.3     Eth2/2
 10.0.2.3          1 FULL/DR          00:21:56 192.168.1.5     Eth2/3
 10.0.2.4          1 FULL/DR          00:21:47 192.168.1.7     Eth2/4

S4048-Leaf1#show ip ospf neighbor

Neighbor ID    Pri      State         Dead Time Address         Interface   Area
10.0.1.1         1      FULL/BDR      00:00:33  192.168.1.0     Fo 1/49     0
10.0.1.2         1      FULL/BDR      00:00:39  192.168.2.0     Fo 1/50     0
```

Note: All neighbor states should be FULL. If a neighbor is stuck in EXSTART or EXCHANGE, there may be an MTU setting mismatch between the two connected interfaces.

5.4.2.3 show ip route bgp

On switches with BGP configured, the `show ip route bgp` command is used to verify the BGP entries in the Routing Information Base (RIB). Entries with multiple paths shown are used with ECMP. The two server networks in this example, 10.60.1.0 and 172.16.1.0, each have two best paths from Nexus5600-Spine1, one through each leaf.

Note: The first set of routes with a subnet mask of /32 are the IPs configured for router IDs.

```
Nexus5600-Spine1# show ip route bgp-64601
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.0.2.1/32, ubest/mbest: 1/0
    *via 192.168.1.1, [20/0], 00:51:59, bgp-64601, external, tag 64701,
10.0.2.2/32, ubest/mbest: 1/0
    *via 192.168.1.3, [20/0], 00:56:12, bgp-64601, external, tag 64702,
10.0.2.3/32, ubest/mbest: 1/0
    *via 192.168.1.5, [20/0], 00:52:07, bgp-64601, external, tag 64703,
10.0.2.4/32, ubest/mbest: 1/0
    *via 192.168.1.7, [20/0], 00:56:08, bgp-64601, external, tag 64704,
```

```
10.60.1.0/24, ubest/mbest: 2/0
    *via 192.168.1.5, [20/0], 00:52:07, bgp-64601, external, tag 64703,
    *via 192.168.1.7, [20/0], 00:52:23, bgp-64601, external, tag 64704,
172.16.1.0/24, ubest/mbest: 2/0
    *via 192.168.1.1, [20/0], 00:51:59, bgp-64601, external, tag 64701,
    *via 192.168.1.3, [20/0], 00:53:42, bgp-64601, external, tag 64702,
```

S4048-Leaf1 has two paths to all other leafs and two paths to Server 2's network, 10.60.1.0. There is one path through each spine. If all paths do not appear, make sure the `maximum-paths` statement in the BGP configuration is equal to or greater than the number of spines in the topology.

```
S4048-Leaf1#show ip route bgp
       Destination        Gateway                      Dist/Metric Last Change
       -----------        -------                      ----------- -----------
  B EX 10.0.2.2/32        via 192.168.1.0                    20/0    00:39:04
                          via 192.168.2.0
  B EX 10.0.2.3/32        via 192.168.1.0                    20/0    00:39:04
                          via 192.168.2.0
  B EX 10.0.2.4/32        via 192.168.1.0                    20/0    00:39:03
                          via 192.168.2.0
  B EX 10.60.1.0/24       via 192.168.1.0                    20/0    00:39:04
                          via 192.168.2.0
```

5.4.2.4 show ip route ospf

On switches with OSPF configured, the `show ip route ospf` command is used to verify the OSPF entries in the Routing Information Base (RIB). Entries with multiple paths shown are used with ECMP. The two server networks in this example, 10.60.1.0 and 172.16.1.0, each have two best paths from Nexus5600-Spine1, one through each leaf.

The first set of routes with a subnet mask of /32 are the IPs configured for router IDs.

```
Nexus5600-Spine1# show ip route ospf
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.0.2.1/32, ubest/mbest: 1/0
    *via 192.168.1.1, Eth2/1, [110/20], 00:32:09, ospf-1, type-2
10.0.2.2/32, ubest/mbest: 1/0
    *via 192.168.1.3, Eth2/2, [110/20], 00:31:49, ospf-1, type-2
10.0.2.3/32, ubest/mbest: 1/0
    *via 192.168.1.5, Eth2/3, [110/20], 00:31:42, ospf-1, type-2
10.0.2.4/32, ubest/mbest: 1/0
    *via 192.168.1.7, Eth2/4, [110/20], 00:31:30, ospf-1, type-2
10.60.1.0/24, ubest/mbest: 2/0
```

```
    *via 192.168.1.5, Eth2/3, [110/20], 00:31:30, ospf-1, type-2
    *via 192.168.1.7, Eth2/4, [110/20], 00:31:30, ospf-1, type-2
172.16.1.0/24, ubest/mbest: 2/0
    *via 192.168.1.1, Eth2/1, [110/20], 00:31:49, ospf-1, type-2
    *via 192.168.1.3, Eth2/2, [110/20], 00:31:49, ospf-1, type-2
192.168.2.0/31, ubest/mbest: 1/0
    *via 192.168.1.1, Eth2/1, [110/2], 00:32:09, ospf-1, intra
192.168.2.2/31, ubest/mbest: 1/0
    *via 192.168.1.3, Eth2/2, [110/2], 00:31:49, ospf-1, intra
192.168.2.4/31, ubest/mbest: 1/0
    *via 192.168.1.5, Eth2/3, [110/2], 00:31:42, ospf-1, intra
192.168.2.6/31, ubest/mbest: 1/0
    *via 192.168.1.7, Eth2/4, [110/2], 00:31:30, ospf-1, intra
```

S4048-Leaf1 has two paths to all other leafs and two paths to the Server 2 network, 10.60.1.0. There is one path through each spine. If all paths do not appear, make sure the `maximum-paths` statement in the OSPF configuration is equal to or greater than the number of spines in the topology.

```
S4048-Leaf1#show ip route ospf
       Destination        Gateway                      Dist/Metric Last Change
       -----------        -------                      ----------- -----------
  O E2 10.0.2.2/32        via 192.168.1.0, Fo 1/49          110/20    00:34:45
                          via 192.168.2.0, Fo 1/50
  O E2 10.0.2.3/32        via 192.168.1.0, Fo 1/49          110/20    00:34:45
                          via 192.168.2.0, Fo 1/50
  O E2 10.0.2.4/32        via 192.168.1.0, Fo 1/49          110/20    00:34:35
                          via 192.168.2.0, Fo 1/50
  O E2 10.60.1.0/24       via 192.168.1.0, Fo 1/49          110/20    00:34:45
                          via 192.168.2.0, Fo 1/50
  O    192.168.1.2/31     via 192.168.1.0, Fo 1/49           110/2    00:35:16
  O    192.168.1.4/31     via 192.168.1.0, Fo 1/49           110/2    00:35:16
  O    192.168.1.6/31     via 192.168.1.0, Fo 1/49           110/2    00:35:16
  O    192.168.2.2/31     via 192.168.2.0, Fo 1/50           110/2    00:34:45
  O    192.168.2.4/31     via 192.168.2.0, Fo 1/50           110/2    00:35:16
  O    192.168.2.6/31     via 192.168.2.0, Fo 1/50           110/2    00:35:16
```

5.4.2.5 show bfd neighbors

The `show bfd neighbors` command may be used to verify BFD is properly configured and sessions are established as indicated by Up in the RH (Remote Heard) and State columns on the Nexus spine and Up in the State column on the Dell EMC leaf.

Note: The output for S4048-Leaf1 shown is for BGP configurations as indicated by a `B` in the `Clients` column. On OSPF configurations, the output is identical except there is an `O` in the `Clients` column. Nexus spine output is the same for either protocol.

```
Nexus5600-Spine1# show bfd neighbors

OurAddr       NeighAddr       RH/RS    Holdown(mult)   State   Int       Vrf
192.168.1.2   192.168.1.3     Up       219(3)          Up      Eth2/2    default
192.168.1.4   192.168.1.5     Up       211(3)          Up      Eth2/3    default
192.168.1.0   192.168.1.1     Up       276(3)          Up      Eth2/1    default
192.168.1.6   192.168.1.7     Up       202(3)          Up      Eth2/4    default

S4048-Leaf1#show bfd neighbors

*       - Active session role
B       - BGP
O       - OSPF

  LocalAddr       RemoteAddr         Interface  State Rx-int Tx-int Mult Clients
* 192.168.1.1     192.168.1.0        Fo 1/49    Up    100    100    3    B
* 192.168.2.1     192.168.2.0        Fo 1/50    Up    100    100    3    B
```

5.4.2.6 Dell EMC Networking leaf validation commands previously covered

The following commands previously covered may be run to validate the Dell EMC Networking leaf switches for this configuration. The output is the same or similar to that shown in the referenced sections.

- `show vlt brief` – see section 4.4.3.6
- `show vlt detail` – see section 4.4.3.7
- `show vlt mismatch` – see section 4.4.3.8
- `show uplink-state-group` – see section 4.4.3.9
- `show spanning-tree rstp brief` – see section 4.4.3.10

5.5 Layer 2 switch configuration

This section provides an overview of the protocols used in constructing the leaf-spine network examples in this guide.

These protocols are used in layer 2 topology examples:

- Virtual Link Trunking (VLT)
- Uplink Failure Detection (UFD)
- Rapid Spanning Tree Protocol (RSTP)
- Link Aggregation Protocol (LACP) / Link Aggregation Group (LAG)

In layer 2 leaf-spine topologies, VLT is used at both the leaf and spine layers.

5.6 Layer 2 topology protocols

5.6.1 VLT

Virtual Link Trunking (VLT) allows link aggregation group (LAG) terminations on two separate switches and supports a loop-free topology. The two switches are referred to as VLT peers and are kept synchronized via an inter-switch link called the VLT interconnect (VLTi). A separate backup link maintains heartbeat messages across the OOB management network.

VLT provides layer 2 multipathing and load-balances traffic. VLT offers the following additional benefits:

- Eliminates blocked ports from STP
- Uses all available uplink bandwidth
- Provides fast convergence if either a link or device fails
- Assures high availability

In layer 2 leaf-spine topologies, VLT is used at both the leaf and spine layers.

Note: Downstream connections from leaf switches configured for VLT do not necessarily have to be configured as LAGs if other fault tolerant methods, such as multipath IO, are preferred.

If a leaf switch loses all connectivity to the spine layer, by default the attached hosts continue to send traffic to that leaf without a direct path to the destination. The VLTi link to the peer leaf switch handles traffic during such a network outage, but this is not considered a best practice.

Dell EMC recommends enabling UFD, which detects the loss of upstream connectivity. An uplink-state group is configured on each leaf switch, which creates an association between the uplinks to the spines and the downlink interfaces.

In the event all uplinks fail on a switch, UFD automatically shuts down the downstream interfaces. This propagates to the hosts attached to the leaf switch. The host then uses its link to the remaining switch to continue sending traffic across the leaf-spine network.

5.6.2 RSTP

As a precautionary measure, Dell EMC recommends enabling Rapid Spanning Tree Protocol (RSTP) on all switches that have layer 2 interfaces. Because VLT environments are loop-free, simultaneously running spanning tree is optional though considered a best practice in case of switch misconfiguration or improperly connected cables. In properly configured and connected leaf-spine networks, there are no ports blocked by spanning tree.

In a Layer-2 leaf-spine network, it is best practice to enable RSTP on all leaf and spine switches prior to cabling. This prevents loops until configurations are in place and can protect network from future cabling loops. Also, the spine switches should have the highest bridge priorities assigned (Ex. primary = 0, secondary = 4096). This prevents leaf switches from unintentionally becoming the root bridge.

With RSTP is enabled on leaf-switches, host interfaces will need to have spanning-tree rstp edge-port enabled to allow for immediate forwarding when interface comes up.

5.6.3 LACP/LAG

Link Aggregation Group (LAG) bundles multiple links into a single interface to increase bandwidth between two devices. LAGs also provides redundancy via the multiple paths. In a leaf-spine network, LAGs are typically used to attach servers to the VLT leaf pairs.

Link Aggregation Control Protocol (LACP) is an improvement over static LAGs in that the protocol will automatically failover if there is a connectivity issue. This is especially important if the links traverse a media converter where it is possible to lose Ethernet connectivity while links remain in an Up state.

5.7 Layer 2 with Dell EMC leaf and Cisco Nexus spine switches

This section provides configuration information to build the layer 2 leaf-spine topology shown in Figure 18. Dell EMC Networking Z9100-ON switches are used at the leaf layer and Cisco Nexus 7000 series switches are used at the spine layer.

5.7.1 Cisco layer-2 spine switch

Cisco Virtual Port-Channel (vPC) technology operates with same benefits and functionality as Dell EMC Networking VLT. In some cases this will be a brownfield installation and conforming to existing protocols and numbering plans is required.

- Comparable features between vPC and VLT
- Limited to two spine switches
- Supports LACP and Static LAGs
- Supports orphan ports
- Supports active-active and active-standby non-LAG ports
- Automatically assigns VLANs to vPC peer-links when VLAN assigned
- vPC Port-channels cannot be L3; must assign port-channel to a L3 VLAN

Some differences in configuration of Dell EMC network leaf switches should be considered when attaching to Cisco vPC spines.

- Cisco default Spanning Tree Protocol is Rapid Per VLAN Spanning Tree (RPVST+) but will also support MSTP. Dell switches only support RSTP or PVST+ in VLT configurations. In the Dell VLT configuration, the Spanning Tree Protocol is only used to prevent loops and would not normally block any paths, so either RSTP or PVST+ can be used in the Dell switch. Dell recommends that RSTP be used on leaf switches when connected to Cisco Spines to avoid unnecessary complexity of the per VLAN instances if not needed.
- If RSTP is used on Dell leaf switch:
  - Received RPVST+ BDPUs for non-default VLANs will be flooded out all forwarding interfaces in that VLAN. In this environment, RSTP communicates with Nexus RPVST+ using the untagged default VLAN (VLAN 1).
  - Enable `spanning-tree rstp edge-port` on all server and storage device interfaces to allow immediate forwarding after link comes up.
- The Dell Networking OS implementation of PVST+ uses IEEE 802.1s costs as the default port costs. For a more consistent path cost calculation with Dell switches, run the `spanning-tree pathcost method long` command on the Nexus switches.
- Cisco recommends disabling LACP graceful-convergence option on port channels when connected to other switches (On by default in Cisco switch).
- On LACP links from Dell leaf to Cisco spine, Dell's default is short timeout while Cisco's default is long timeout (rate normal). Match accordingly.

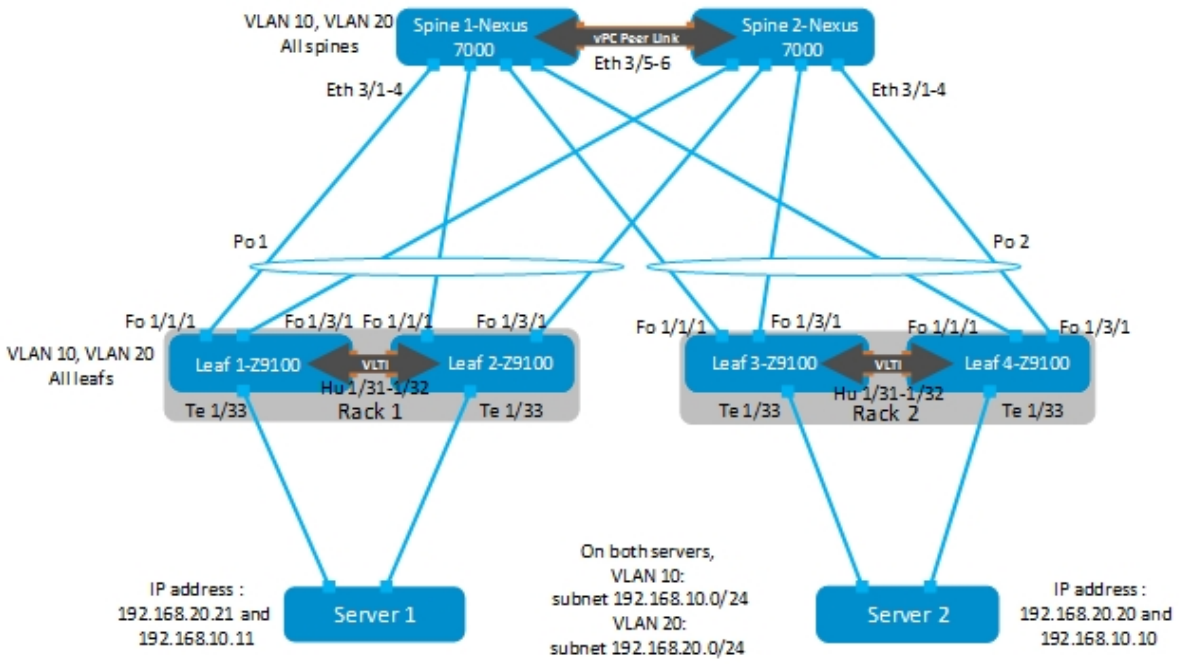


Figure 18 Layer 2 leaf-spine topology with Dell EMC leaf and Cisco Nexus spine switches

Note: Cisco Nexus switches in this example were reset to their factory default configurations by running write erase followed by reload. After reload, "Power on Auto Provisioning" was not used, the admin password was configured and the Nexus “basic configuration dialog” was not used. Refer to your Nexus system documentation for more information.

5.7.2 Z9100-ON leaf switch configuration

The following section outlines the configuration commands issued to the Z9100-ON leaf switches to build the topology in Figure 18. The commands detailed below are for L2-Leaf1-Z9100 and L2-Leaf2-Z9100. The configuration commands for L2-Leaf3-Z9100 and L2-Leaf4-Z9100 are similar.

Note: On Z9100-ON switches, Telnet is enabled and SSH is disabled by default. Both services require the creation of a non-root user account to login. If needed, it is a best practice to use SSH instead of Telnet for security. SSH can optionally be enabled with the command:

```
(conf)#ip ssh server enable
```

A user account can be created to access the switch via SSH with the command:

```
(conf)#username ssh_user sha256-password ssh_password
```

1. Use the following commands to configure the serial console, enable the password, and disable Telnet:

Table 36 Serial console, password, and Telnet configuration commands

L2-Leaf1-Z9100	L2-Leaf2-Z9100
enable configure enable sha256-password *enable_password* no ip telnet server enable	enable configure enable sha256-password *enable_password* no ip telnet server enable

2. Use the following commands to:
   a. Set the hostname
   b. Configure the OOB management interface and default gateway
   c. Enable LLDP
   d. Enable RSTP as a precaution

Note: In this layer 2 topology, the RSTP root bridge is configured at the spine level.

Table 37 Hostname, management interface, LLDP, and RSTP configuration commands

L2-Leaf1-Z9100	L2-Leaf2-Z9100
hostname L2-Leaf1-Z9100 interface ManagementEthernet 1/1 ip address 100.67.194.5/24 no shutdown management route 0.0.0.0/0 100.67.194.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc protocol spanning-tree rstp no disable	hostname L2-Leaf2-Z9100 interface ManagementEthernet 1/1 ip address 100.67.194.6/24 no shutdown management route 0.0.0.0/0 100.67.194.254 protocol lldp advertise management-tlv management- address system-description system-name advertise interface-port-desc protocol spanning-tree rstp no disable

3. Convert interfaces connected to the Nexus 5600 spines from their default speed of 100GbE to 40GbE.

Table 38 Interface conversion commands

L2-Leaf1-Z9100	L2-Leaf2-Z9100
stack-unit 1 port 1 portmode single speed 40G no-confirm stack-unit 1 port 3 portmode single speed 40G no-confirm	stack-unit 1 port 1 portmode single speed 40G no-confirm stack-unit 1 port 3 portmode single speed 40G no-confirm

4. Configure the VLT interconnect between Leaf1 and Leaf2. In this configuration, add interfaces hundredGigE 1/31 – 1/32 to static port channel 127 for the VLT interconnect. The backup destination is the management IP address of the VLT peer switch.

Table 39 VLT interconnect commands

L2-Leaf1-Z9100	L2-Leaf2-Z9100
interface Port-channel 127 description VLTi Port-Channel no ip address channel-member hundredGigE 1/31,1/32 no shutdown interface range hundredGigE 1/31-1/32 description VLTi no ip address no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.194.6 unit-id 0	interface Port-channel 127 description VLTi Port-Channel no ip address channel-member hundredGigE 1/31,1/32 no shutdown interface range hundredGigE 1/31-1/32 description VLTi no ip address no shutdown vlt domain 127 peer-link port-channel 127 back-up destination 100.67.194.5 unit-id 1

5. Ensure that the interface Te 1/33 connects downstream to Server 1 and is configured as an RSTP edge port.
6. Verify that the interfaces Fo 1/1/1 and Fo 1/3/1 connect to the spines upstream and are configured in LACP port channel 1. The port channel is configured for VLT.

Table 40 Interface connections configuration

L2-Leaf1-Z9100	L2-Leaf2-Z9100
interface TenGigabitEthernet 1/33 description Server-1 no ip address portmode hybrid switchport spanning-tree rstp edge-port no shutdown interface fortyGigE 1/1/1 description Spine1-Port1 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/3/1 description Spine2-Port1 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface Port-channel 1 description To Spines no ip address portmode hybrid switchport vlt-peer-lag port-channel 1 no shutdown	interface TenGigabitEthernet 1/33 description Server-1 no ip address portmode hybrid switchport spanning-tree rstp edge-port no shutdown interface fortyGigE 1/1/1 description Spine1-Port2 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface fortyGigE 1/3/1 description Spine2-Port2 no ip address port-channel-protocol LACP port-channel 1 mode active no shutdown interface Port-channel 1 description To Spines no ip address portmode hybrid switchport vlt-peer-lag port-channel 1 no shutdown

7. Verify that VLANs 10 and 20 are configured on each switch. Port-channel 1 is tagged in both VLANs.

Note: The shutdown/no shutdown commands on a VLAN have no effect unless the VLAN is assigned an IP address (configured as an SVI).

Table 41 VLAN configuration commands

L2-Leaf1-Z9100	L2-Leaf2-Z9100
interface Vlan 10 no ip address tagged TenGigabitEthernet 1/33 tagged Port-channel 1 shutdown interface Vlan 20 no ip address tagged TenGigabitEthernet 1/33 tagged Port-channel 1 shutdown	interface Vlan 10 no ip address tagged TenGigabitEthernet 1/33 tagged Port-channel 1 shutdown interface Vlan 20 no ip address tagged TenGigabitEthernet 1/33 tagged Port-channel 1 shutdown

8. Configure UFD to shut down the downstream interfaces if all uplinks fail. The hosts attached to the switch use the remaining LACP port member to continue sending traffic across the fabric.
9. Use the `end` and `write` commands to exit configuration mode and save the configuration.

Table 42 UFD configuration commands

L2-Leaf1-Z9100	L2-Leaf2-Z9100
uplink-state-group 1 description Disable edge port in event all spine uplinks fail downstream TenGigabitEthernet 1/33 upstream Port-channel 1 end write	uplink-state-group 1 description Disable edge port in event all spine uplinks fail downstream TenGigabitEthernet 1/33 upstream Port-channel 1 end write

5.7.3 Nexus 7000 series spine switch configuration

The following sections outline the configuration commands issued to the Nexus 7000 series switches to build the topology in Figure 18.

1. Using the commands in Table 42, enable the LACP and virtual port channel (vPC) features and configure the hostname, management IP address, and default management route.

Note: Cisco enables Rapid Per VLAN Spanning Tree Plus (RPVST+), its implementation of RSTP, on Nexus 7000 series switches by default.

Table 43 Nexus 7000 spine switch configuration

L2-Spine1-Nexus7K	L2-Spine2-Nexus7K
enable configure feature lacp feature vpc hostname L2-Spine1-Nexus7K interface mgmt0 vrf member management ip address 100.67.184.21/24 no shutdown vrf context management ip route 0.0.0.0/0 100.67.184.254	enable configure feature lacp feature vpc hostname L2-Spine2-Nexus7K interface mgmt0 vrf member management ip address 100.67.184.28/24 no shutdown vrf context management ip route 0.0.0.0/0 100.67.184.254

2. Create VLAN 10 and 20.

Note: All VLANs are added to RPVST+ as a precaution against loops. L2-Spine1-Nexus7K is configured as the primary spanning tree root bridge using the `spanning tree vlan vlan_numbers priority 0` command. L2-Spine1-Nexus7K is configured as the secondary spanning tree root bridge using the `spanning tree vlan vlan_numbers priority 4096` command.

Table 44 VLAN creation commands

L2-Spine1-Nexus7K	L2-Spine2-Nexus7K
vlan 10 vlan 20 spanning-tree vlan 1,10,20 spanning-tree vlan 1,10,20 priority 0	vlan 10 vlan 20 spanning-tree vlan 1,10,20 spanning-tree vlan 1,10,20 priority 4096

3. Create a vPC domain and vPC peer link between the two spine switches.
4. On spine 1, assign a role priority of 1 to make it the vPC primary.
5. Specify the management IP address of the vPC peer as the vPC peer-keepalive destination.

Note: In this example, interfaces Ethernet 3/5 and 3/6 are used to create the vPC peer link. Interfaces are configured as trunk ports and allow applicable VLANs.

Table 45 vPC domain and peer link configuration commands

L2-Spine1-Nexus7K	L2-Spine2-Nexus7K
vpc domain 1 role priority 1 peer-keepalive destination 100.67.184.28 source 100.67.184.21 auto-recovery interface port-channel 20 switchport switchport mode trunk switchport trunk allowed vlan 1,10,20 spanning-tree port type network vpc peer-link interface Ethernet3/5 switchport switchport mode trunk switchport trunk allowed vlan 1,10,20 channel-group 20 mode active no shutdown interface Ethernet3/6 switchport switchport mode trunk switchport trunk allowed vlan 1,10,20 channel-group 20 mode active no shutdown	vpc domain 1 role priority 65535 peer-keepalive destination 100.67.184.21 source 100.67.184.28 auto-recovery interface port-channel 20 switchport switchport mode trunk switchport trunk allowed vlan 1,10,20 spanning-tree port type network vpc peer-link interface Ethernet3/5 switchport switchport mode trunk switchport trunk allowed vlan 1,10,20 channel-group 20 mode active no shutdown interface Ethernet3/6 switchport switchport mode trunk switchport trunk allowed vlan 1,10,20 channel-group 20 mode active no shutdown

6. Configure port channels and member ports for downstream connectivity to the leaf switches.
7. Use the the `end` and `copy running-config startup-config` commands to exit configuration mode and save the configuration.

Table 46 Port channel and member port configuration commands

L2-Spine1-Nexus7K	L2-Spine2-Nexus7K
interface port-channel1 switchport switchport mode trunk vpc 1 interface port-channel2 switchport switchport mode trunk vpc 2 interface Ethernet3/1 switchport switchport mode trunk channel-group 1 mode active no shutdown interface Ethernet3/2 switchport switchport mode trunk channel-group 1 mode active no shutdown interface Ethernet3/3 switchport switchport mode trunk channel-group 2 mode active no shutdown interface Ethernet3/4 switchport switchport mode trunk channel-group 2 mode active no shutdown end copy running-config startup-config	interface port-channel1 switchport switchport mode trunk vpc 1 interface port-channel2 switchport switchport mode trunk vpc 2 interface Ethernet3/1 switchport switchport mode trunk channel-group 1 mode active no shutdown interface Ethernet3/2 switchport switchport mode trunk channel-group 1 mode active no shutdown interface Ethernet3/3 switchport switchport mode trunk channel-group 2 mode active no shutdown interface Ethernet3/4 switchport switchport mode trunk channel-group 2 mode active no shutdown end copy running-config startup-config

5.7.4 Validation

In addition to sending traffic between hosts, the configuration shown in Figure 18 can be validated with the commands shown in this section. For more information on commands and output, see the *Command Line Reference Guide* for the applicable switch (links to documentation are provided in Appendix F).

5.7.4.1 show vpc

The `show vpc` command displays the vPC status on the Nexus spine switches. Peer status and vPC keep-alive status must be as shown. The consistency status fields should all show `success`. If not, see the `show vpc consistency-parameters` command.

```
L2-Spine1-Nexus7K# show vpc
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link

vPC domain id                     : 1
Peer status                       : peer adjacency formed ok
vPC keep-alive status             : peer is alive
Configuration consistency status  : success
Per-vlan consistency status       : success
Type-2 inconsistency reason       : Consistency Check Not Performed
vPC role                          : primary
Number of vPCs configured         : 2
Peer Gateway                      : Disabled
Dual-active excluded VLANs        : -
Graceful Consistency Check        : Enabled
Auto-recovery status              : Enabled (timeout = 240 seconds)

vPC Peer-link status
---------------------------------------------------------------------
id   Port   Status Active vlans
--   ----   ------ --------------------------------------------------
1    Po20   up     1,10,20

vPC status
----------------------------------------------------------------------------
id   Port      Status Consistency Reason                Active vlans
--   ----      ------ ----------- ------                ------------
1    Po1       up     success     success                  1,10,20
2    Po2       up     success     success                  1,10,20
```

5.7.4.2 show vpc consistency-parameters

The `show vpc consistency-parameters` command pinpoints inconsistencies between vPC peers on the Nexus spine switches. Depending on the severity of the misconfiguration, vPC may either warn the user (Type-2 misconfiguration) or suspend the port channel (Type-1 misconfiguration). In the specific case of a VLAN mismatch, only the VLAN that differs between the vPC member ports is suspended on the port channels.

```
L2-Spine1-Nexus7K# show vpc consistency-parameters ?
  global     Global Parameters
  interface  Specify interface
  vlans      Vlans
  vpc        Virtual Port Channel configuration

L2-Spine1-Nexus7K# show vpc consistency-parameters global

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

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
STP Mode                    1     Rapid-PVST             Rapid-PVST
STP Disabled                1     None                   None
STP MST Region Name         1     ""                     ""
STP MST Region Revision     1     0                      0
STP MST Region Instance to  1
 VLAN Mapping
STP Loopguard               1     Disabled               Disabled
STP Bridge Assurance        1     Enabled                Enabled
STP Port Type, Edge         1     Normal, Disabled,      Normal, Disabled,
BPDUFilter, Edge BPDUGuard        Disabled               Disabled
STP MST Simulate PVST       1     Enabled                Enabled
Allowed VLANs               -     1,10,20                1,10,20
Local error VLANs           -     -                      -
```

5.7.4.3 show spanning-tree

The `show spanning-tree` command validates STP is enabled on all VLANs on the Nexus spine switches and all interfaces are forwarding (`Sts` column shows `FWD`). One of the spine switches (L2-Spine1-Nexus7K in this example) is the root bridge on all VLANs.

```
L2-Spine1-Nexus7K# show spanning-tree

VLAN0001
  Spanning tree enabled protocol rstp
  Root ID    Priority    1
             Address     8478.ac11.e341
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    1      (priority 0 sys-id-ext 1)
             Address     8478.ac11.e341
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
```

```
Po1              Desg FWD 1         128.4096 (vPC) P2p
Po2              Desg FWD 1         128.4097 (vPC) P2p
Po20             Desg FWD 1         128.4115 (vPC peer-link) Network P2p


VLAN0010
  Spanning tree enabled protocol rstp
  Root ID    Priority    10
             Address     8478.ac11.e341
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    10     (priority 0 sys-id-ext 10)
             Address     8478.ac11.e341
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Po1              Desg FWD 1         128.4096 (vPC) P2p
Po2              Desg FWD 1         128.4097 (vPC) P2p
Po20             Desg FWD 1         128.4115 (vPC peer-link) Network P2p


VLAN0020
  Spanning tree enabled protocol rstp
  Root ID    Priority    20
             Address     8478.ac11.e341
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    20     (priority 0 sys-id-ext 20)
             Address     8478.ac11.e341
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Po1              Desg FWD 1         128.4096 (vPC) P2p
Po2              Desg FWD 1         128.4097 (vPC) P2p
Po20             Desg FWD 1         128.4115 (vPC peer-link) Network P2p
```

5.7.4.4 Dell EMC Networking leaf validation commands previously covered

The following commands previously covered may be run to validate the Dell EMC Networking leaf switches for this configuration. The output is the same or similar to that shown in the referenced sections.

- `show vlt brief` – see section 4.7.3.1
- `show vlt detail` – see section 4.7.3.2
- `show vlt mismatch` – see section 4.7.3.3
- `show uplink-state-group` – see section 4.7.3.4
- `show spanning-tree rstp brief` – see section 4.7.3.5

6 Networking features and guidelines for storage

This section includes networking features that affect the performance and functionality of storage traffic. Implementing and configuration of these features are dependent on the capability and support of the storage device deployed in the data center.

Note: The recommendations and information in this section are not written for any specific storage device. Consult the user guide to determine the best features and settings to implement for your specific storage model.

iSCSI topics:

- iSCSI optimization
- Link-level flow control
- Data Center Bridging
- Frame size
- Storage network connections

Software defined storage topics:

- Multicast
- Storage network connections

6.1 iSCSI optimization

iSCSI is a TCP/IP-based protocol for establishing and managing connections between IP-based storage devices and initiators in a storage area network (SAN). iSCSI optimization enables the network switch to auto-detect Dell's iSCSI storage arrays and triggers self-configuration of several key network features that enable optimization of the network for better storage traffic throughput. iSCSI optimization is disabled by default.

iSCSI optimization also provides a means of monitoring iSCSI sessions and applying quality of service (QoS) policies on iSCSI traffic. When enabled, iSCSI optimization allows a switch to monitor (snoop) the establishment and termination of iSCSI connections. The switch uses the snooped information to detect iSCSI sessions and connections established through the switch.

iSCSI optimization reduces deployment time and management complexity in data centers. In a data center network, Dell EqualLogic and Compellent iSCSI storage arrays are connected to a converged Ethernet network using the data center bridging exchange protocol (DCBx) through stacked and/or non-stacked Ethernet switches.

Note: VLT and stacking are incompatible. Stacking is not used in the leaf-spine topology.

iSCSI session monitoring over virtual link trunking (VLT) synchronizes the iSCSI session information between the VLT peers, allowing session information to be available in both the VLT peers. You can enable or disable iSCSI when you configure VLT.

iSCSI optimization functions as follows:

- Auto-detection of EqualLogic storage arrays — the switch detects any active EqualLogic array directly attached to its ports.
- Manual configuration to detect Compellent storage arrays where auto-detection is not supported.
- Automatic configuration of switch ports after detection of storage arrays (EqualLogic or Compellent).
  - All switch interfaces changed to mtu 9192.
  - If DCB disabled (default), all interfaces assigned `flowcontrol rx on tx off`.
  - If DCB enabled, all interfaces assigned with DCB.
  - Storm control disabled on interfaces connected to storage .
  - Spanning Tree Protocol Portfast (edgeport) enabled on interfaces connected to storage.
- If you configure flow-control, iSCSI uses the current configuration. If you do not configure flow-control, iSCSI auto-configures flow control settings so that receive-only is enabled and transmit-only is disabled.
- iSCSI monitoring sessions — the switch monitors and tracks active iSCSI sessions in connections on the switch, including port information and iSCSI session information.
- iSCSI Quality of Service (QoS) — A user-configured iSCSI class of service (CoS) profile is applied to all iSCSI traffic. Classifier rules are used to direct the iSCSI data traffic to queues that can be given preferential QoS treatment over other data passing through the switch. Preferential treatment helps to avoid session interruptions during times of congestion that would otherwise cause dropped iSCSI packets.
- iSCSI DCBx time, length, and value elements (TLVs) are supported.

6.1.1 Information monitored in iSCSI traffic flows

iSCSI optimization examines the following data in packets and uses the data to track the session and create the classifier entries that enable QoS treatment.

- Initiator's IP Address
- Target's IP Address
- ISID (Initiator defined session identifier)
- Initiator's IQN (iSCSI qualified name)
- Target's IQN
- Initiator's TCP Port
- Target's TCP Port
- Connection ID
- Aging
- Up Time

If no iSCSI traffic is detected for a session during a user-configurable aging period, the session data is cleared.

Note: If you are using EqualLogic or Compellent storage arrays, more than 256 simultaneous iSCSI sessions are possible. However, iSCSI session monitoring is not capable of monitoring more than 256 simultaneous iSCSI sessions. If this number is exceeded, sessions may display as unknown in session monitoring output. Dell Networking recommends that you disable iSCSI session monitoring for EqualLogic and Compellent storage arrays or for installations with more than 256 simultaneous iSCSI sessions.

6.1.2 Configuring iSCSI optimization

To configure iSCSI optimization:

1. Allocate content addressable memory (CAM).

Note: See specific device Configuration guide. This is optional if session monitoring is not required.

2. Enable or disable DCB as required.
3. If Compellent or other storage attached, execute the `iscsi profile-compellent` command on all storage interfaces.

Note: If EqualLogic storage attached, no additional commands needed.

4. Execute `iscsi enable` command.
5. Execute write memory to save configuration.
6. Execute reload to activate CAM allocation and DCB.
7. After switch is reloaded, `show iscsi` to validate iscsi status.
8. If required, set iSCSI QoS policy (Optional; dot1p value unchanged).

Note: Content addressable memory (CAM) allocation is optional. If CAM is not allocated, the following features are disabled: session monitoring, session aging, and Class of Service (CoS) override.

6.2 Link-level flow control

In storage networks, it is sometimes recommended by storage vendors to activate link-level flow control on the servers and the switching network. Flow control temporarily stops the flow to the end device to prevent congestion and the resulting frame drops. Flow control is only recommended at the end point connections (server and/or storage) and not between switching nodes. This is known as asymmetric flow control and can be achieved by enabling `iscsi enable` globally or `flowcontrol rx on tx off` on all interfaces. A switch with asymmetric flow control enabled will only respond to pause frames and never send a pause frame. This allows the end point to momentarily stop the traffic flow to prevent a potential packet drop by the end device. When the switch receives a pause frame, any new packet destined to the end device will be queued in the switch's output buffer until the pause request expires or the pause request is rescinded by the end device. While flow control may slow some packets down, preventing a packet drop may optimize the traffic flow by preventing retransmissions.

Note: Dell OS9 does not allow simultaneous operation of DCB and link-level flow control. If enabling link-level flow control, ensure DCB is disabled with `no dcb enable` command.

While considering viable options of implementing flow control, it is very important to understand the following:

- Flow control is not intended to solve the problem of steady-state congested links or networks
- Flow control is not intended to address poor network capacity
- Flow control is not intended to provide end-to-end flow control

Note: Changes in the flow-control values may not be reflected automatically in show interface output. To display the change, apply the new flow control setting, perform a shutdown followed by a no shutdown command on the interface, and then check the show interface output again.

6.3 Data Center Bridging

Data Center Bridging (DCB) refers to a set of IEEE Ethernet enhancements that provide data centers with a single, robust, converged network to support multiple traffic types, including local area network (LAN), server, and storage traffic. Through network consolidation, DCB results in reduced operational cost, simplified management, and easy scalability by avoiding the need to deploy separate application-specific networks. DCB is only supported on layer 2 networks and must assigned on all devices in the storage path. DCB should be used with FCoE storage and when storage traffic is mixed on same paths as heavy LAN traffic to prevent congestion.

For example, instead of deploying separate Ethernet network for LAN traffic, with a storage area network (SAN) to ensure lossless Fibre Channel traffic, and a separate network for high-performance inter-processor computing within server clusters, only one DCB-enabled network is required in a data center. The Dell Networking switches that support a unified fabric and consolidate multiple network infrastructures use a single input/output (I/O) device called a converged network adapter (CNA).

A CNA is a computer input/output device that combines the functionality of a host bus adapter (HBA) with a network interface controller (NIC). Multiple adapters on different devices for several traffic types are no longer required.

OS9 supports the following DCB features:

- Data center bridging exchange protocol (DCBx)
- Priority-based flow control (PFC)
- Enhanced transmission selection (ETS)

To ensure lossless delivery and latency-sensitive scheduling of storage and service traffic and I/O convergence of LAN, storage, and server traffic over a unified fabric, IEEE data center bridging adds the following extensions to a classical Ethernet network:

- 802.1Qbb — Priority-based Flow Control (PFC)
- 802.1Qaz — Enhanced Transmission Selection (ETS)
- 802.1Qau — Congestion Notification
- Data Center Bridging Exchange (DCBx) protocol

Note: Dell Networking OS supports only the PFC, ETS, and DCBx features in data center bridging.

6.3.1 Priority-Based Flow Control

In a data center network, priority-based flow control (PFC) manages large bursts of one traffic type in multiprotocol links so that it does not affect other traffic types and no frames are lost due to congestion.

When PFC detects congestion on a queue for a specified priority, it sends a pause frame for the 802.1p priority traffic to the transmitting device. In this way, PFC ensures that PFC-enabled priority traffic is not dropped by the switch.

PFC enhances the existing 802.3x pause and 802.1p priority capabilities to enable flow control based on 802.1p priorities (classes of service). Instead of stopping all traffic on a link, as performed by the traditional Ethernet pause mechanism, PFC pauses traffic on a link according to the 802.1p priority set on a traffic type. You can create lossless flows for storage and server traffic while allowing for loss in case of LAN traffic congestion on the same physical interface.

6.3.2 Enhanced Transmission Selection

Enhanced transmission selection (ETS) supports optimized bandwidth allocation between traffic types in multiprotocol (Ethernet, FCoE, iSCSI) links.

ETS allows you to divide traffic according to its 802.1p priority into different priority groups (traffic classes) and configure bandwidth allocation and queue scheduling for each group to ensure that each traffic type is correctly prioritized and receives its required bandwidth. For example, you can prioritize low-latency storage or server cluster traffic in a traffic class to receive more bandwidth and restrict best-effort LAN traffic assigned to a different traffic class.

6.3.3 Data Center Bridging Exchange protocol

DCBx allows a switch to automatically discover DCB-enabled peers and exchange configuration information. PFC and ETS use DCBx to exchange and negotiate parameters with peer devices. DCBx capabilities include:

- Discovery of DCB capabilities on peer-device connections
- Determination of possible mismatch in DCB configuration on a peer link
- Configuration of a peer device over a DCB link

DCBx requires the link layer discovery protocol (LLDP) to provide the path to exchange DCB parameters with peer devices. Exchanged parameters are sent in organizationally specific TLVs in LLDP data units.

6.3.4 Enabling Data Center Bridging on DCB source switches

DCB is disabled by default and should be enabled on all devices carrying the storage traffic. Typically, the core layer switch will be the DCB source for all devices. For core switches in a VLT configuration, both would be assigned as DCB source.

1. Allocate content addressable memory (CAM) for either FCoE or iSCSI depending on desired storage access.

Note: See specific device Configuration Guide for CAM allocation configuration.

2. Switch must be reloaded after changing CAM.
3. Ensure link-layer flow control is disabled on all switch interfaces using `flowcontrol rx off tx off` command.
4. Enable DCB using global `dcb enable` command.
5. On the DCB source switches, create a DCB map.
   a. `dcb-map` *`dcb-map-name`*
6. Create priority flow control (PFC) groups within the created DCB map. For iSCSI and LAN traffic, assign 2 priority groups (PG) with desired bandwidth for each. In the following example, LAN traffic (PG 0) will be given 40% and iSCSI (PG 1) will be given 60%. iSCSI will be assigned with PFC.
   a. `priority-group 0 bandwidth 40 pfc off`
   b. `priority-group 1 bandwidth 60 pfc on`
7. Within the DCB map, assign the priority groups to the desired CoS priority. Typically, CoS 4 is used for iSCSI traffic and the remaining CoS (0-3, 5-7) are assigned to priority group 1.
   a. `priority-pgid 0 0 0 0 1 0 0 0`
8. Assign DCB map to all participating interfaces (downstream switches, servers and storage devices).
   a. `dcb-map` *`dcb-map-name`*

6.3.5 Enabling Data Center Bridging on downstream switches

1. Allocate content addressable memory (CAM) for either FCoE or iSCSI depending on desired storage access.

Note: See specific device Configuration Guide for CAM allocation configuration.

2. Switch must be reloaded after changing CAM.
3. Ensure link-layer flow control is disabled on all switch interfaces using `flowcontrol rx off tx off` command.
4. Enable DCB using global `dcb enable` command.
5. On each interface connected to the DCB source switch, enable the following:
   a. protocol lldp
   b. dcbx port-role auto-upstream
6. On each interface connected to downstream switches or end devices, enable the following:
   a. protocol lldp
   b. dcbx port-role auto-downstream
7. Verify DCB status of each interface using following commands:
   a. `show interface ten x/y dcbx detail`

b. `show interface ten x/y pfc detail`
c. `show interface ten x/y ets detail`

6.4 QoS with DSCP

The Dell EMC Networking switches use differentiated services code point (DSCP) marking to place the appropriate traffic into separate queues for prioritization. This section details a simple configuration to place a higher priority on storage traffic than application traffic.

Dell EMC Networking switches provide a high degree of customization for QoS. Some options available include bandwidth limitations, Weighted Random Early Detection (WRED) and Explicit Congestion Notification (ECN). There is not a generic, one-size-fits-all approach to QoS. The strict queuing used in this example could easily be substituted with explicit bandwidth assignments. Administrators can use this example as a starting point for their QoS strategy.

The configuration example below accomplishes the following:

- Uses the DSCP values as configured on the iSCSI storage appliance or software-defined storage VDS
- Maps DSCP input traffic to specified queues and DSCP color
- Prioritizes egress traffic on uplinks through strict queuing and WRED

Note: Consult the user guide of your iSCSI storage appliance to determine if the DSCP marking feature is available. Software-defined storage users should consult the appropriate user guide for DSCP support. An example with instructions for setting up DSCP marking on a VMware VDS can be seen within the *ScaleIO IP Fabric Best Practice and Deployment Guide*.

Leaf switch configuration procedure:

1. Access the command line and enter configuration mode.
2. Create a class to match traffic for each DSCP value.

```
class-map match-any class_compute
 match ip dscp 14

class-map match-any class_storage
 match ip dscp 46
```

3. Create an input policy map to map each class of traffic to a specific queue.

```
policy-map-input pmap_ingress
 service-queue 1 class-map class_compute
 service-queue 3 class-map class_storage
```

4. Create a DSCP color map profile for the application traffic.

```
qos dscp-color-map Colormap_DSCP
 dscp yellow 14
```

5. Apply to each input interface the input policy map with an input service policy and the DSCP color map profile with a QoS DSCP color map policy.

```
interface TenGigabitEthernet 1/1
 service-policy input pmap_ingress
 qos dscp-color-policy Colormap_DSCP
```

Note: Repeat step 4 for each input interface. (leaf interfaces to each node)

6. Create a QoS output policy for the storage traffic.

```
qos-policy-output qpol_egress
 scheduler strict
```

7. Create a WRED profile for the color used in Step 4.

```
wred-profile Yellow_profile
 threshold min X max X max-drop-rate X
```

Note: Replace the "**x**" in the command above with appropriate values for your deployment.

8. Create a QoS output policy for the application traffic.

```
qos-policy-output qpol_egress2
 wred yellow Yellow_profile
```

9. Create an output policy map for each QoS policy.

```
policy-map-output pmap_egress
 service-queue 1 qos-policy qpol_egress2
 service-queue 3 qos-policy qpol_egress
```

10. Apply the output policy with an output service policy to each switch uplink interface.

```
interface fortyGigE 1/49
 service-policy output pmap_egress
```

Repeat Step 10 for all leaf uplink interfaces.

Note: Interface numbers in the above QoS example do not correspond to the example configurations in Chapters 4 and 5. Use the appropriate input and uplink interfaces numbers for your specific deployment.

The configuration in the above steps can be applied to any input and uplink interface throughout the leaf-spine topology. This example only implements priority queuing at the uplink interfaces. Tagging or marking occurs at the distributed port groups and mapping occurs at the switch interfaces to the nodes.

6.5 Frame size

Standard frame size is 1500 bytes, but most devices can be expanded to 9000 bytes. This allows greater throughput with less utilization impact because of reducing the relative size of the headers in each frame. To function correctly, the storage initiator and target Maximum Transmission Unit (MTU) needs to be set to the same size. Also, each switch in the data path should be set to at least the packet size plus IP header size. Traffic with larger data block sizes, there can be a significant throughput improvement.

Note: iSCSI can support up to 9000 bytes. FCoE requires supporting up to 2148 bytes.

6.6 Multicast

Some HCI solutions with SDS require IPv4 and/or IPv6 multicast traffic. Multicast Listener Discovery (MLD) is required at switch to support VxRail and XC Series appliances. Enabling IGMP snooping allows the switch to distribute IPv4 multicast traffic only to end devices that have joined the multicast group. This prevents unwanted multicast traffic on interfaces that do not need the traffic.

For some virtual SAN connections, IGMP snooping and IGMP snooping querier are enabled on leaf switches:

1. Enable following IGMP snooping commands:
   a. `ip igmp snooping enable`
   b. `no ip igmp snooping flood`
2. Enable IGMP querier on the virtual SAN VLAN:
   a. `ip igmp snooping querier`

6.7 Storage network connections

Dell recommends that storage and servers have redundant network connections. The following sections describe networking features used for connecting server and storage devices.

Traditional SAN devices may utilize more than one of the redundancy features listed below. Consult the user guide for your specific storage model to determine the best features and settings to implement.

6.7.1 Active-standby

Active-standby NIC teaming is a method where multiple links are connected to the network, but only one link is active at a time. While this provides redundancy it leaves available bandwidth unused.

6.7.1.1 VMware VDS with active-standby

A Virtual Distributed Switch (VDS) in a VMware ESXi host can be set to use active-standby for those traffic types that prefer active-standby. The following diagram shows a representation of the network connection from the physical to the virtual in the ESXi host.

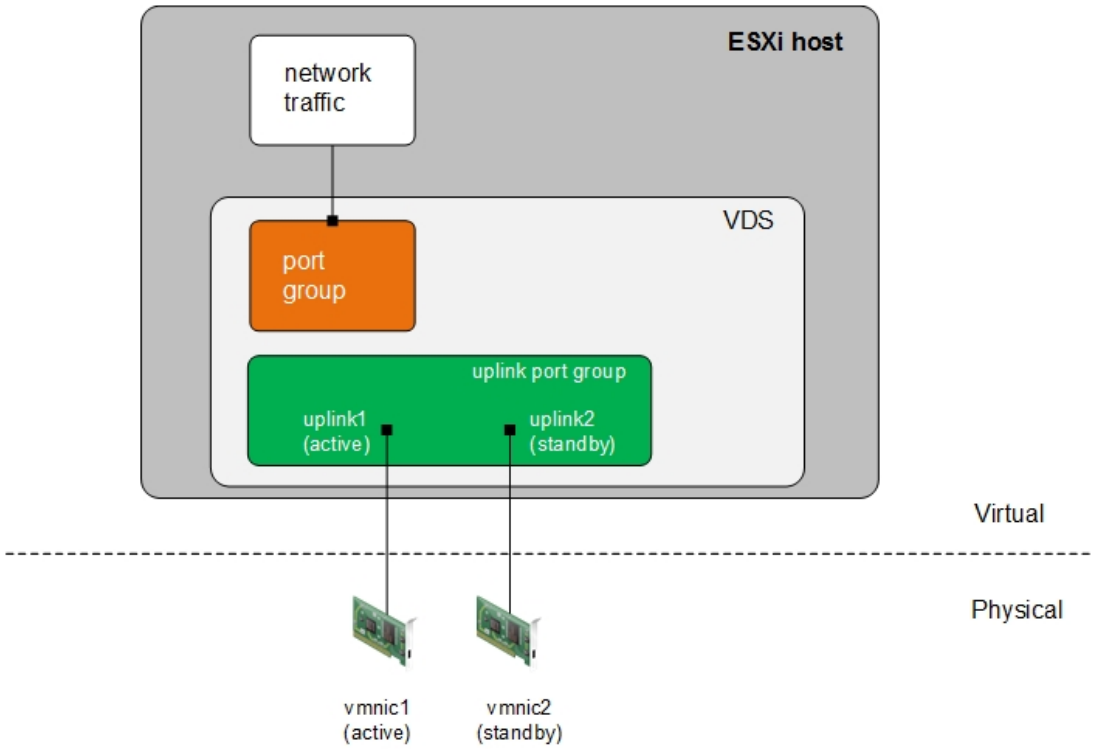


Figure 19 VDS with active-standby configuration

An example with instructions for setting up an active-standby connection within a VDS can be seen within the *ScaleIO IP Fabric Best Practice and Deployment Guide*.

6.7.2 LACP

Link Aggregation Control Protocol (LACP) is a method of NIC teaming that provides active-active network connections using the same protocol available in the switching network.

6.7.2.1 VMware VDS with LACP

A Virtual Distributed Switch (VDS) in a VMware ESXi host can be set to use LACP for those traffic types that prefer active-active. The following diagram shows a representation of the network connection from the physical to the virtual in the ESXi host.

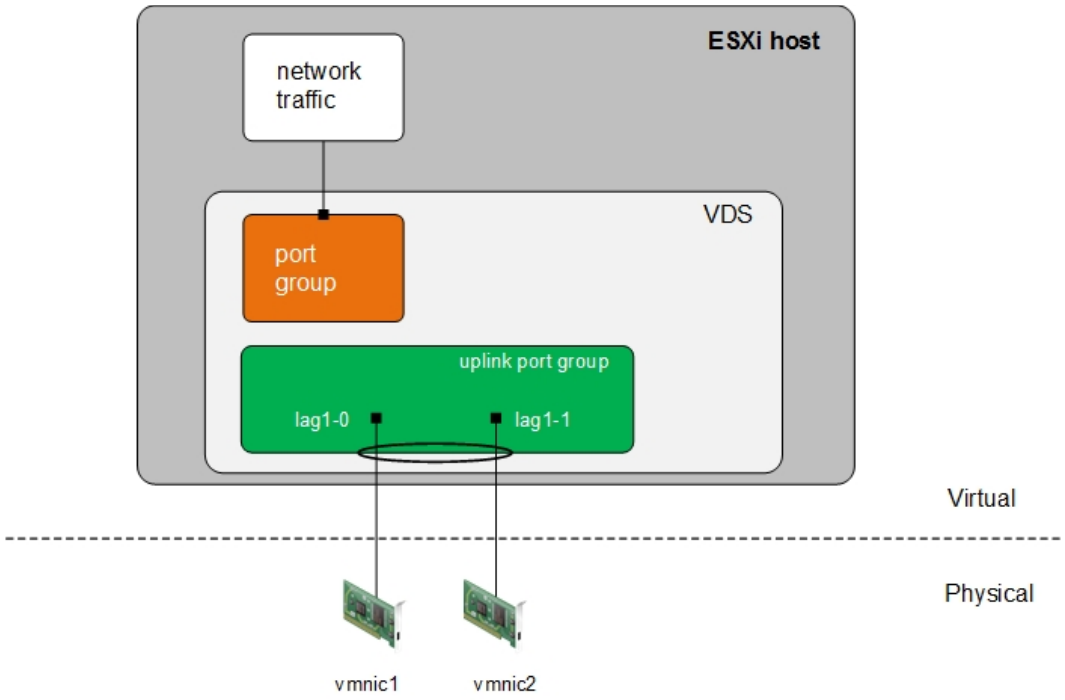


Figure 20 VDS with LACP

An example with instructions for setting up an active-active connection within a VDS can be seen within the *ScaleIO IP Fabric Best Practice and Deployment Guide*.

6.7.3 Multipath IO

Multipath IO (MPIO) is a Microsoft feature designed to provide alternate paths from multiple Host Bus Adapters (HBA) connecting to storage systems. MPIO provides up to 32 paths to add redundancy and load balancing for the storage traffic. A device specific module (DSM) is required from the storage vendor to integrate into the Microsoft Windows operating system.

Server interfaces using MPIO are not configured in a LAG. When connecting to VLT pairs, the server interfaces are orphan ports and will not take advantage of the layer-2 multipathing provided by VLT peering. Orphan ports are ports that do not belong to a VLT.

6.7.4 iSCSI offload

Some CNAs support offloading the iSCSI and TCP/IP stacks to hardware embedded on the network adapter reducing server CPU load so that more resources are allocated to the user applications. Choosing the right network adapters for storage access can increase performance.

A Storage topologies

A.1 Leaf-Spine with iSCSI storage array topology

In the modern leaf-spine data center, storage appliances are attached to a leaf pair. This topology removes the segregation of the LAN and SAN, allowing for a much more manageable and scalable solution.

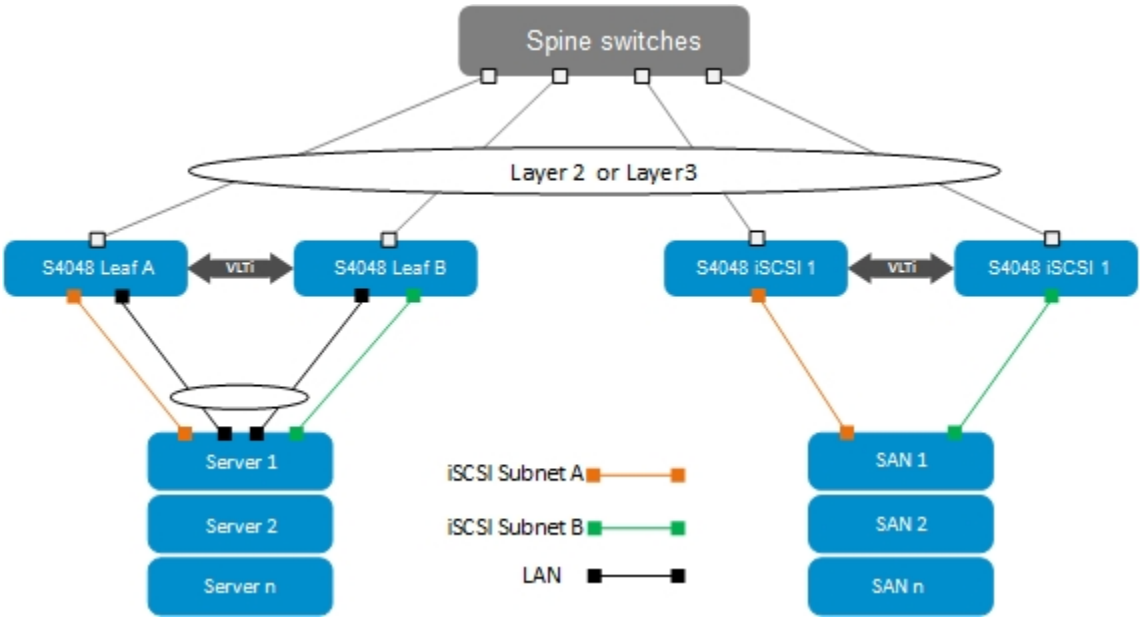


Figure 21 Leaf-spine data center with iSCSI storage array

A.2 Leaf-Spine with software-defined storage topology

In the modern leaf-spine data center, hyper-converged appliances are attached to a leaf pair. This topology allows for a much more manageable and scalable solution.

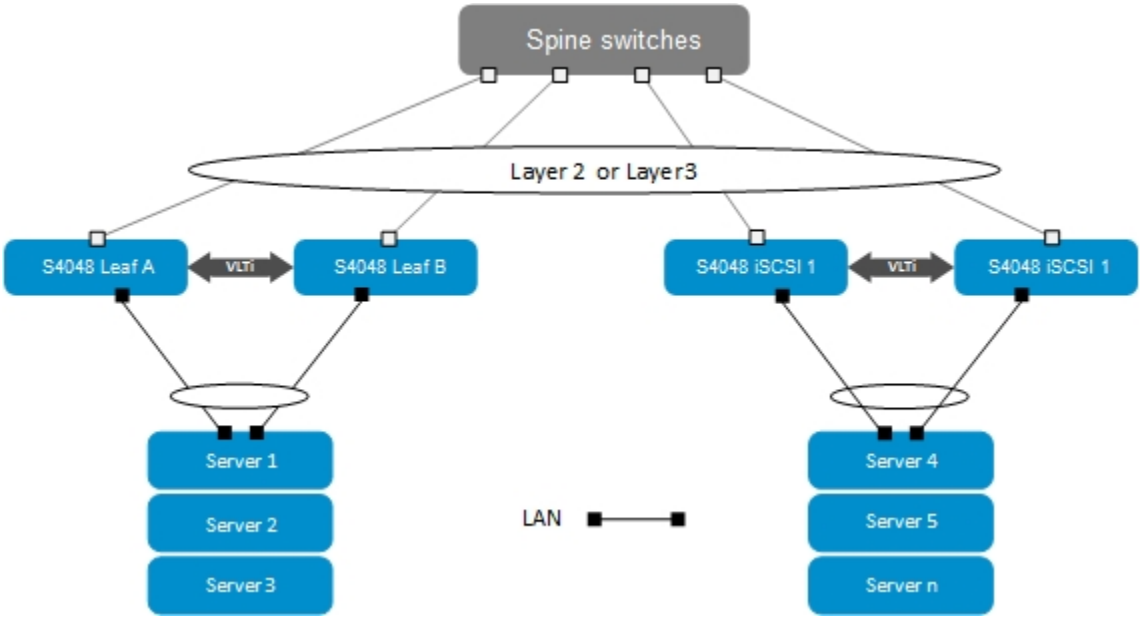


Figure 22 Leaf-spine data center with software-defined storage on hyper-converged appliances

B Dell EMC Networking switches

Dell EMC offers many networking switches to meet customer requirements. The following are some of the models that can be used with this guide:

- Spine switches
  - Dell EMC Networking Z9100-ON
  - Dell EMC Networking S6010-ON
- Leaf switches
  - Dell EMC Networking S4048-ON
  - Dell EMC Networking S3048-ON

For more information on Dell EMC data center switches see the *Leaf-Spine Deployment and Best Practices Guide* and product information at Dell.com.

B.1 Dell EMC Networking switch factory default settings

All Dell EMC Networking switches in this guide can be reset to factory defaults as follows:

```
Dell#restore factory-defaults stack-unit unit# clear-all
Proceed with factory settings? Confirm [yes/no]:yes
```

Factory settings are restored and the switch reloads. After reload, enter A at the [A/C/L/S] prompt as shown below to exit Bare Metal Provisioning mode.

```
This device is in Bare Metal Provisioning (BMP) mode.
To continue with the standard manual interactive mode, it is necessary to abort
BMP.

Press A to abort BMP now.
Press C to continue with BMP.
Press L to toggle BMP syslog and console messages.
Press S to display the BMP status.
[A/C/L/S]:A

% Warning: The bmp process will stop ...

Dell>
```

The switch is now ready for configuration.

C Overview of leaf-spine architecture

The connections between leaf and spine switches can be layer 2 (switched) or layer 3 (routed). The terms “layer 3 topology” and “layer 2 topology” in this guide refer to these connections. In both topologies, downstream connections to servers, storage and other endpoint devices within the racks are layer 2 and connections to external networks are layer 3.

The following concepts apply to layer 2 and layer 3 leaf-spine topologies:

- Each leaf switch connects to every spine switch in the topology.
- Servers, storage arrays, edge routers and similar devices always connect to leaf switches, never to spines.

The layer 2 and layer 3 topologies each use two leaf switches at the top of each rack configured as a Virtual Link Trunking (VLT) pair. VLT allows all connections to be active while also providing fault tolerance. As administrators add racks to the data center, two leaf switches configured for VLT are added to each new rack.

The total number of leaf-spine connections is equal to the number of leaf switches multiplied by the number of spine switches. Bandwidth of the fabric may be increased by adding connections between leafs and spines as long as the spine layer has capacity for the additional connections.

C.1 Layer 3 leaf-spine topology

In a layer 3 leaf-spine network, traffic between leafs and spines is routed. The layer 3/layer 2 boundary is at the leaf switches. Spine switches are never connected to each other in a layer 3 topology. Equal cost multi-path routing (ECMP) is used to load balance traffic across the layer 3 network. Connections within racks from hosts to leaf switches are layer 2. Connections to external networks are made from a pair of edge or border leafs as shown in Figure 23.

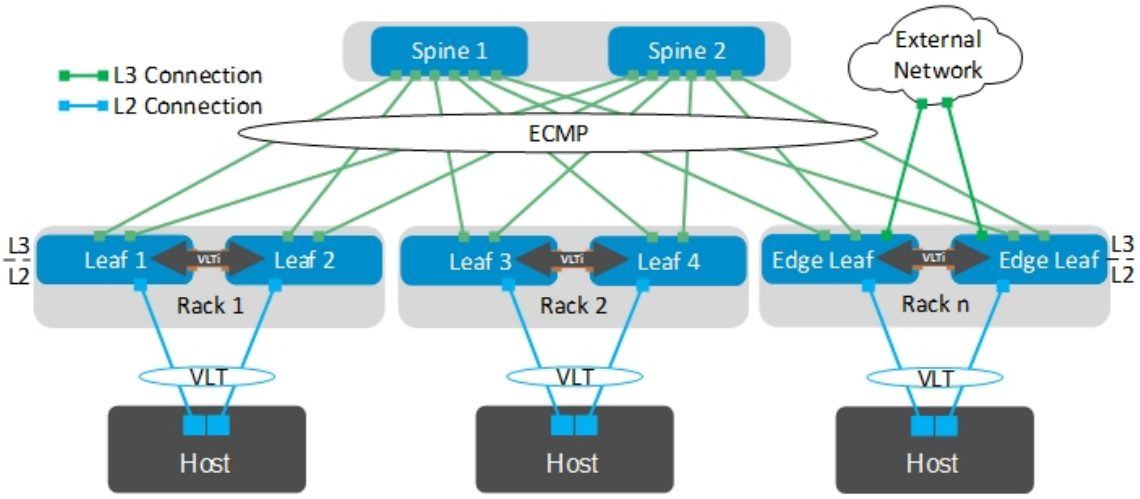


Figure 23 Layer 3 leaf-spine network

C.2 Layer 2 leaf-spine topology

In a layer 2 leaf-spine network, traffic between leafs and spines is switched (except for a pair of edge leafs) as shown in Figure 24. VLT is used for multipathing and load balancing traffic across the layer 2 leaf-spine fabric. Connections from hosts to leaf switches are also layer 2.

For connections to external networks, layer 3 links are added between the spines and a pair of edge leafs.

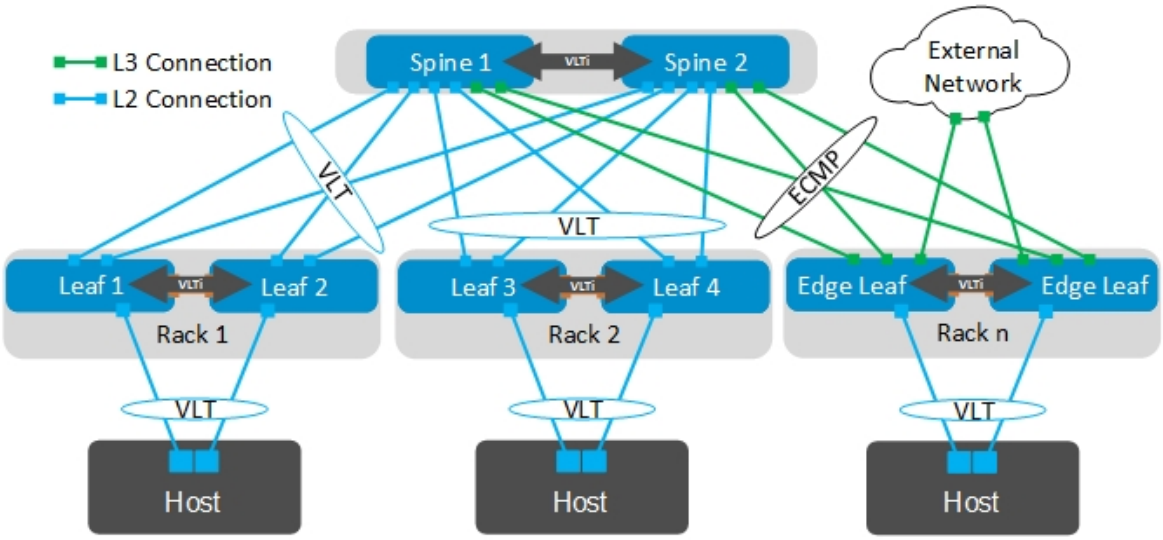


Figure 24 Layer 2 leaf-spine network

C.3 Design considerations

A layer 3 topology has some advantages over a comparable layer 2 topology:

- Fewer spanning-tree interactions and complexity
- Increase bandwidth scalability over layer 2 spine
- Broadcast domain is contained at rack level
- Avoid traffic polarization by using flow-based hash algorithms with ECMP
- Layer 3 is for required for VXLAN
- In very large deployments (>4000 nodes), eBGP preferred over OSPF due to scalability

A layer 2 topology is generally less complex but has some limitations that must be considered. These include:

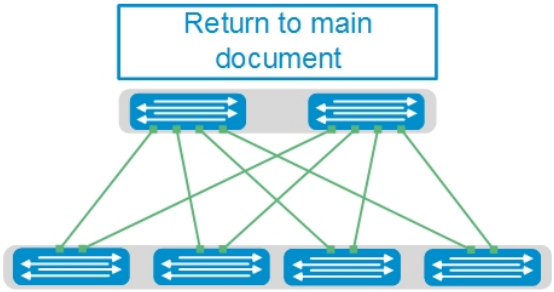
- For each VLAN, the layer 2 topology creates one large broadcast domain across the fabric. The layer 3 topology has the benefit of containing broadcast domains to each rack.
- The layer 2 topology is limited to 4094 VLANs across the fabric. The layer 3 topology allows up to 4094 VLANs per rack.

- The layer 2 topology is limited to two physical switches at the spine layer (configured as VLT peers). In a layer 3 topology, additional spines may be added as needed to provide additional paths and bandwidth. Therefore, a layer 3 topology is more scalable and is better suited for very large networks.
- Overlay networks utilizing VXLAN (such as VMware NSX) require a layer 3 underlay network.

Other design considerations:

- Layer 2 at spine is needed for supporting DCB protocols
- For best traffic balancing in L2 leaf-spine network, use source-destination IP, L4 port, and VLAN as port-channel hashing algorithm
- Although Static LAGs are supported, it is best practice to use LACP with VLT LAGs to allow graceful failover and to protect from misconfigurations

If none of the layer 2 limitations are a concern, it may ultimately come down to a matter of preference. This guide provides examples of both topologies.



Return to Leaf-Spine with Dell EMC spine and leaf

Return to Leaf-Spine with Cisco spine and Dell EMC leaf

D Management network

The OOB management network is isolated from the leaf-spine production network. It is the same for the layer 2 and layer 3 leaf-spine topologies.

In this example, a Dell EMC Networking S3048-ON switch installed in each rack provides 1GbE connectivity to the management network. The RJ-45 OOB management ports on each spine and leaf switch are connected to the S3048-ON switches as shown in Figure 25. PowerEdge server iDRACs and Chassis Management Controllers (CMCs) are also connected for server administration.

For the S3048-ON management switches, all ports used are in layer 2 mode and are in the default VLAN. Rapid Spanning Tree Protocol (RSTP) is enabled as a precaution against loops. No additional configuration is required.

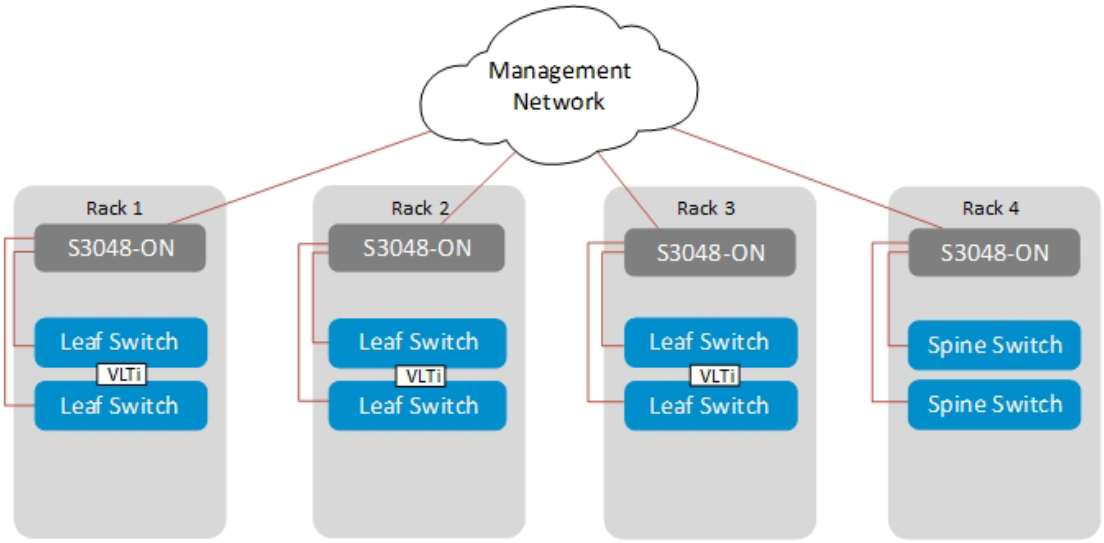


Figure 25 Management network

Note: A management network is not a requirement to configure or use a leaf-spine network, but is recommended to efficiently manage servers, switches and storage devices.

D.1 iDRAC server interface

iDRAC interfaces provide remote access over an out-of-band (OOB) management to the physical server via 1Gb RJ-45 interface. All server management functions can use this interface rather than physically connecting keyboard and screen. Dell EMC recommends setting up an out-of-band management switch to facilitate direct communication with each server node.

E Validated hardware and operating systems

The following table includes the hardware and operating systems used to validate the examples in this guide:

Table 47 Switches and operating systems used in this guide

Switch	OS / Version
Dell EMC Networking S3048-ON	DNOS 9.11.2.0 P0
Dell EMC Networking S4048-ON	DNOS 9.11.2.0 P0
Dell EMC Networking S6010-ON	DNOS 9.11.2.0 P0
Dell EMC Networking Z9100-ON	DNOS 9.11.2.0 P0

F Technical support and resources

Dell EMC TechCenter is an online technical community where IT professionals have access to numerous resources for Dell EMC software, hardware and services.

Dell EMC TechCenter Networking Guides

- *Leaf-Spine Deployment and Best Practices Guide*
- *ScaleIO IP Fabric Best Practice and Deployment Guide*

Manuals and documentation for Dell EMC Networking S3048-ON

Manuals and documentation for Dell EMC Networking S4048-ON

Manuals and documentation for Dell EMC Networking S6010-ON

Manuals and documentation for Dell EMC Networking Z9100-ON

G Support and Feedback

Contacting Technical Support

Support Contact Information

Web: http://dell.com/support

Feedback for this document

We encourage readers to provide feedback on the quality and usefulness of this publication by sending an email to Dell_Networking_Solutions@Dell.com.