Dell EMC Validated System for Virtualization - NSX Reference Architecture

A step-by-step VMware NSX deployment on a leaf-spine data center network with FC430 compute nodes and VSAN storage.

Dell Networking Solutions Engineering
November 2016

Revisions

Date	Revision	Description	Authors
November 2016	1.0	Initial Release	Jim Slaughter, Curtis Bunch

Table of contents

1 Introduction

This guide covers an NSX deployment for the data center based on the Dell EMC Validated System for Virtualization.

The goal of this guide is to enable a network administrator or engineer with traditional networking and ESXi experience to build a scalable NSX virtual network using the Dell EMC Validated System for Virtualization hardware and software outlined in this guide.

This document provides a best practice leaf-spine topology with configuration steps for all physical switches in the topology. It includes step-by-step configuration of a virtual network using VMware NSX that overlays the physical network. This includes configuration of logical switches, routers, and options for communicating with external traditional networks using software and hardware solutions. It also includes steps to deploy ESXi on PowerEdge servers and deployment of a vSphere vCenter Server Appliance.

Note: See the appendices for product versions validated.

1.1 Validated System for Virtualization

The Dell EMC Validated System for Virtualization is the industry's most flexible converged system to date, with choice in building blocks of compute, storage and networking tested and validated to integrate and operate together in support of a virtualized environment. The system incorporates a wide range of form factors, technology choices and deployment options, right-sized to fit each customer's needs. A fully-validated system can be configured, quoted and ordered in minutes, while automated lifecycle management tools allow customers to easily deploy, scale, and update the system.

1.1.1 Addressing the need for flexibility

With increasing business demands and decreasing IT budgets, customers face unprecedented pressures to improve efficiency and lower costs. The current operational model of delivering IT services, which involves procuring technology from best of breed technology providers and managing them in silos, proves to not only be time consuming but problematic. In this approach, customers are typically burdened to manually make design decisions, validate various components, set up and configure components and manage the environment in an ongoing fashion by engaging multiple vendors for assistance. Across the end-to-end infrastructure lifecycle, these elements increase complexity and cost for customers.

Existing integrated solutions that aim to solve these challenges are either pre-integrated and prepackaged offers that optimize time to production and simplify ongoing operations, with customers making a tradeoff on flexibility and choice, or traditional reference architectures that provide some degree of flexibility but do not offer manageability or scalability benefits.

The Dell EMC Validated System for Virtualization bridges this gap by offering a tested and validated integrated system that is highly flexible, scalable, and driven using end-to-end automation throughout the infrastructure lifecycle.

1.1.2 Differentiated approach addresses challenges and limitations

To provide IT services faster, while lowering costs and streamlining operations, Dell EMC engineered the Validated System for Virtualization. This groundbreaking system enables you to achieve greater operational efficiencies and savings, and unparalleled management simplicity, by giving you more power than ever to define and design it.

The system can be deployed with options ranging from “do-it-yourself” using a deployment guide, a system integrated on-site by Dell EMC or by using your own integration vendor.

The Dell EMC Validated System for Virtualization is:

- Built on our best-of-breed products that are designed for virtualization across the ecosystem. By offering various design choices and guidance on choosing the right components, the Dell EMC Validated System for Virtualization takes the guesswork out of solution design and reduces the enormous time it takes to procure, validate, and integrate components. They can be designed to start small, based on the customer’s initial requirements, and grow, based on customer’s ongoing requirements, which reduces the initial investment required for infrastructure deployment.
- Tested, validated, and fully integrated, yet flexible enough to be tailored for your organization, removing risk and accelerating your time to value.
- Delivered with the Dell EMC Active System Manager (ASM) to simplify ongoing management. Whether it is configuring the system based on Dell EMC and VMware best practices, scaling the system as business demands grow, updating the system as new hardware updates come along, ensuring compliance of the system by managing drift, or repurposing the system as business needs change, ASM, when combined with the Dell EMC Validated System for Virtualization, drives IT agility and reduces ongoing IT costs.
- Delivered with Dell EMC’s global reach, exceptional execution and delivery, providing consistent deployment, management, and maintenance in every region of the world.
- Delivered with a single point of support for the complete system including hardware and software through Dell ProSupport Plus. ProSupport Plus resolves issues faster when they occur and reduces the risk of severe issues and outages.

More information about the Dell EMC Validated System for Virtualization is available here.

1.2 VMware NSX

VMware NSX is a network virtualization technology. It allows for the decoupling of network services from the physical infrastructure. With NSX, logical networks are created on top of a basic layer 2 (switched) or layer 3 (routed) physical infrastructure. This allows the physical and virtual environments to be decoupled, enabling agility and security in the virtual environment while allowing the physical environment to focus on throughput.

The NSX platform also provides for network services in the logical space. Some of these logical services include switching, routing, firewalling, load balancing and Virtual Private Network (VPN) services.

NSX benefits include the following:

- Simplified network service deployment, migration, and automation

- Reduced provisioning and deployment time
- Scalable multi-tenancy across one or more data centers
- Distributed routing and a distributed firewall at the hypervisor allow for better east-to-west traffic flow and an enhanced security model
- Provides solutions for traditional networking problems, such as limited VLANs, MAC address, FIB and ARP entries
- Application requirements do not require modification to the physical network
- Normalization of underlying hardware, enabling easier hardware migration and interoperability

1.3 The VXLAN protocol

NSX creates logical networks using the Virtual Extensible Local Area Network (VXLAN) protocol. The VXLAN protocol is described in Internet Engineering Task Force document RFC 7348. VXLAN allows a layer 2 network to scale across the data center by overlaying a layer 3 network. Each overlay is referred to as a VXLAN segment and only virtual machines (VMs) within the same segment can communicate with each other.

Each segment is identified through a 24-bit segment ID referred to as a VXLAN Network Identifier (VNI). This allows up to 16 Million VXLAN segment IDs, far greater than the traditional 4,094 VLAN IDs allowed on a physical switch.

VXLAN is a tunneling scheme that encapsulates layer 2 frames in User Datagram Protocol (UDP) segments, as shown in Figure 1:

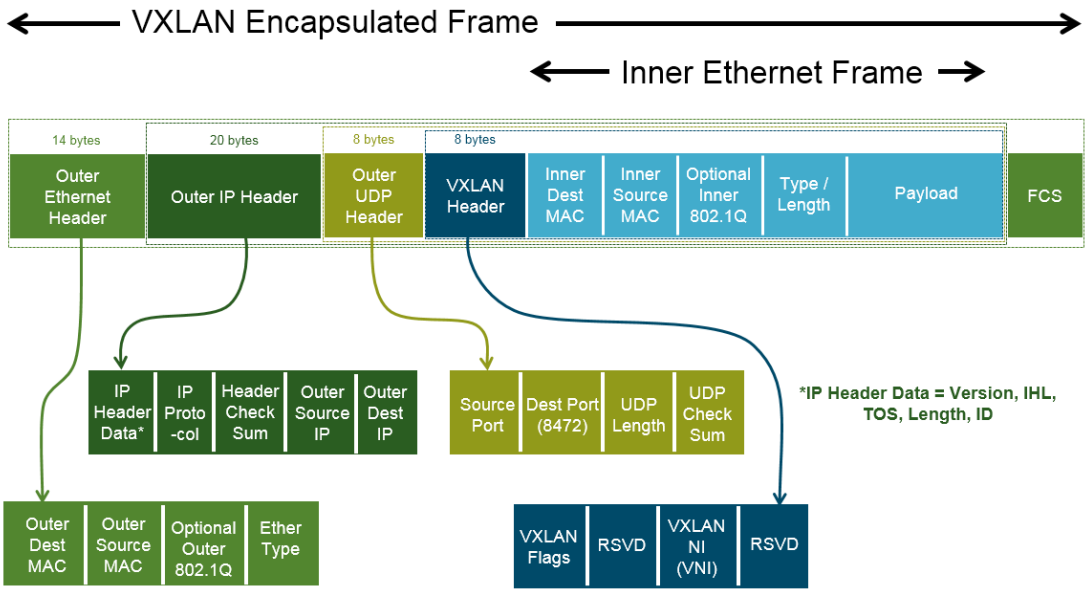


Figure 1 VXLAN encapsulated frame

VXLAN encapsulation adds approximately 50 bytes of overhead to each Ethernet frame. As a result, all switches in the underlay (physical) network must be configured to support an MTU of at least 1600 bytes on all participating interfaces.

As part of the VXLAN configuration, each ESXi host is configured with a software VXLAN tunnel end point (VTEP). A software VTEP is a VMkernel interface where VXLAN encapsulation and de-encapsulation occurs.

A physical switch that supports VXLAN can act as a hardware VTEP, also referred to as a VXLAN Gateway (Section 13.2). This allows communication with servers inside the data center that are outside of the virtual network.

1.4 Typographical conventions

This document uses the following typographical conventions:

`Monospace text`	Command Line Interface (CLI) examples
`Bold monospace text`	Commands entered at the CLI prompt
`Italic monospace text`	Variables in CLI examples

2 Hardware overview

While the Dell EMC Validated System for Virtualization has flexibility and choice across servers, storage and networking, this guide is focused on a single instance of the system. This section briefly describes the primary hardware used to validate this deployment. A complete listing of hardware validated for this guide is provided in Appendix A.

2.1 Dell PowerEdge FX2s enclosure and supported modules

The PowerEdge FX2s enclosure is a 2-rack unit (RU) computing platform. It has capacity for two FC830 full-width servers, four FC630 half-width servers or eight FC430 quarter-width servers. The enclosure is also available with a combination of servers and storage sleds. The FX2s enclosure used in this guide contains four FC430 servers (Section 2.1.1) and two FD332 storage sleds (Section 2.1.2).

Figure 2 Dell PowerEdge FX2s (front) with four PowerEdge FC430 servers and two FD332 storage sleds

The back of the FX2s enclosure includes two I/O networking modules (IOMs) and eight PCIe expansion slots.

Figure 3 Dell PowerEdge FX2s (back) with two PowerEdge FN410S IOMs installed

2.1.1 PowerEdge FC430 server

The PowerEdge FC430 server is a quarter-width, 2 socket server. Four FC430 servers in the top half of the FX2s enclosure combine with the FD332 storage sleds to form the compute cluster for this deployment.

Figure 4 PowerEdge FC430

2.1.2 PowerEdge FD332 storage sled

The PowerEdge FD332 is a half-width, direct-attached storage sled with up to 16 drives. It combines with FC-series servers to build flexible storage solutions. This deployment includes two FD332 storage sleds installed in the bottom half of the FX2s enclosure.

Figure 5 PowerEdge FD332

2.1.3 PowerEdge FN410S I/O Module

The PowerEdge FN410S IOM is a multilayer switch with eight internal, server-facing ports and four external, 10GbE SFP+ ports. Two FN410S IOMs installed in the FX2s enclosure provide fault tolerance.

Figure 6 PowerEdge FN410S

2.2 PowerEdge R630 server

The PowerEdge R630 server is a 2-socket, 1 RU server. This server functions in the management and edge clusters in this guide.

Figure 7 PowerEdge R630

2.3 Z9100-ON

The Z9100-ON is a 1 RU, multilayer switch with thirty-two ports supporting 10/25/40/50/100GbE. Two Z9100-ON switches are used as spines in the leaf-spine topology covered in this guide.

Figure 8 Dell Networking Z9100-ON

2.4 S4048-ON

The S4048-ON is a 1RU, layer 2/3 switch with forty-eight 10GbE SFP+ ports and six 40GbE QSFP+ ports. Six S4048-ON switches are used as leaf switches in the leaf-spine topology covered in this guide.

Figure 9 Dell Networking S4048-ON

2.5 S3048-ON

The S3048-ON is a 1RU switch with forty-eight 1GbE base-T ports. In this guide, one S3048-ON switch supports management traffic in each rack.

Figure 10 Dell Networking S3048-ON

3 Topology

This section provides an overview of the physical and virtual topology used in this deployment.

3.1 Servers

The servers are grouped into three VMware vCenter clusters, with one cluster per physical rack:

- Rack 1 Management Cluster – contains three PowerEdge R630 servers
- Rack 2 Compute FC430 Cluster – contains one PowerEdge FX2s chassis with four FC430 servers and two FD332 storage sleds
- Rack 3 Edge Cluster – contains three PowerEdge R630 servers

The three clusters have been spread across three physical racks as shown in Figure 11 to illustrate the scalability of this design as additional servers and switches are added.

3.2 Production network

The production network used in this guide has two major components:

- The physical, or underlay, network as shown in Figure 11.
- The NSX virtual, or overlay, network as shown in Figure 12.

3.2.1 Physical network (underlay)

On the production network, a spine and leaf topology is used for performance and scalability. Two leaf switches (S4048-ONs) are used in each rack for redundancy and increased performance. Dell Virtual-Link Trunking (VLT) connects each pair of leaf switches.

Each leaf switch has point-to-point connections to both spine switches (Z9100-ONs). Traffic between the leaf switches and spine switches is routed and Equal-Cost Multi-Path routing (ECMP) is leveraged to utilize all available bandwidth.

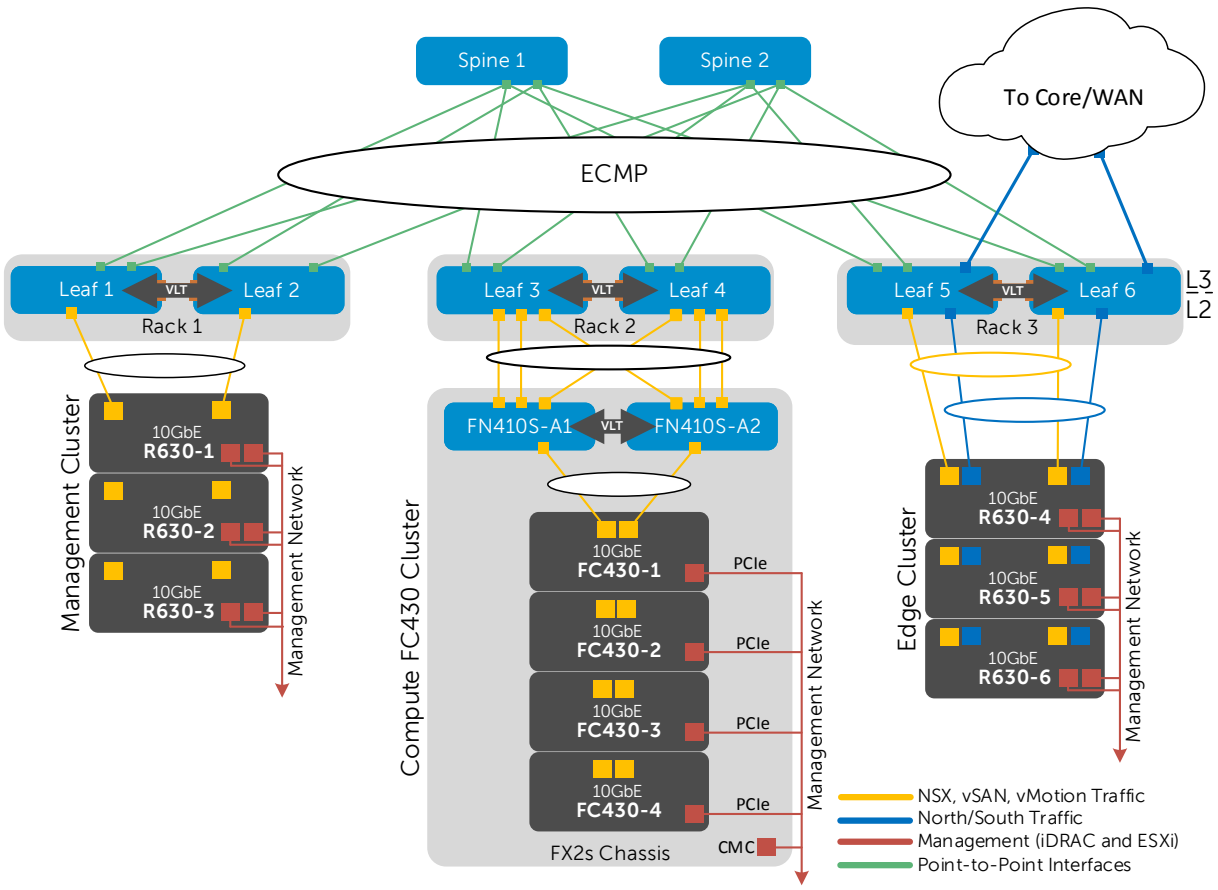


Figure 11 Physical production network

3.2.2 NSX virtual network (overlay)

The virtual network, built with VMware NSX, overlays the physical production network. All servers participating in the virtual network run VMware ESXi. VM-to-VM traffic is contained within the virtual network.

Traffic from the data center's virtual network to the network core or WAN (Wide Area Network) can be configured to pass through an Edge Services Gateway (ESG). This takes advantage of additional services provided by NSX, such as firewalling, load balancing and VPN services. ESG configuration is covered in Section 13.1.

Note: All networking in the data center does not require virtualization. Traffic from the virtual network to the physical data center network can use a hardware VTEP. For details, see Section 13.2.

Figure 12 shows the virtual network built for this guide.

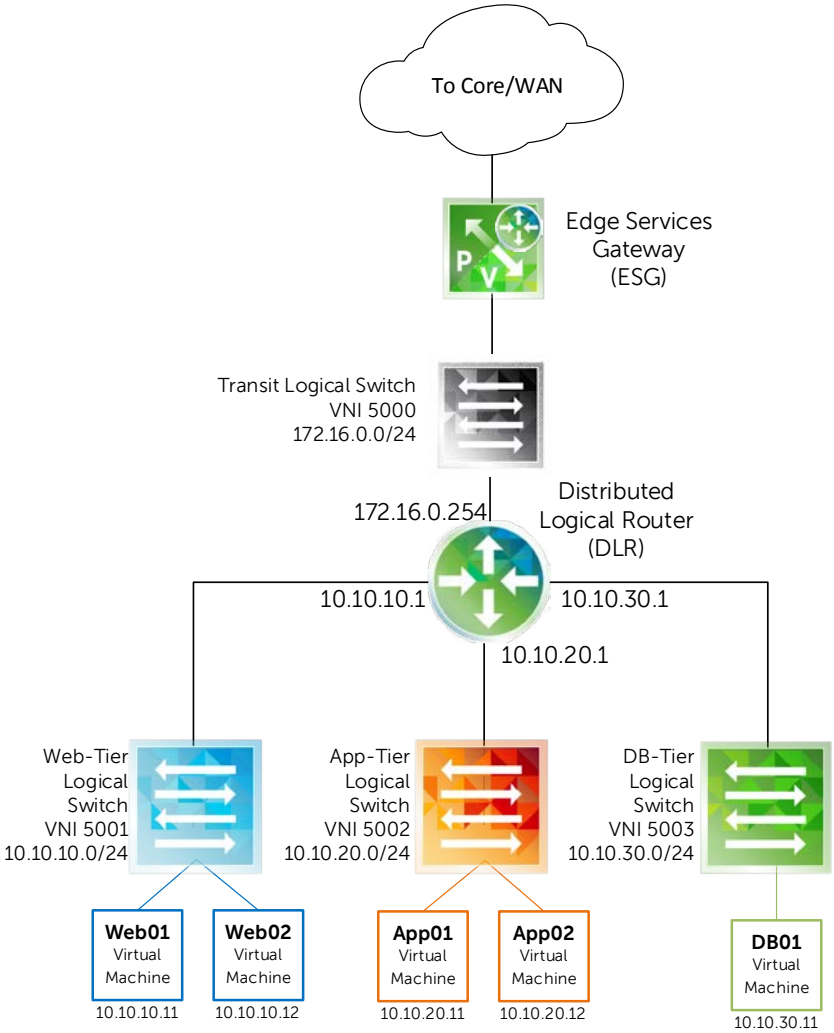


Figure 12 NSX virtual network

3.2.3 Combined physical and virtual networks

Figure 13 shows the combined networks. All servers are running ESXi.

The management cluster in Rack 1 contains the vCenter Server Virtual Appliance (VCVA), NSX Manager, and NSX controllers.

The compute cluster in Rack 2 contains the production virtual machines. In this guide, the compute cluster includes VMs deployed to different virtual networks to represent web servers, application servers, and database servers.

The edge cluster in Rack 3 contains the ESG for connectivity to the network core or WAN. The edge cluster also contains the distributed logical router (DLR) for routing NSX traffic between networks.

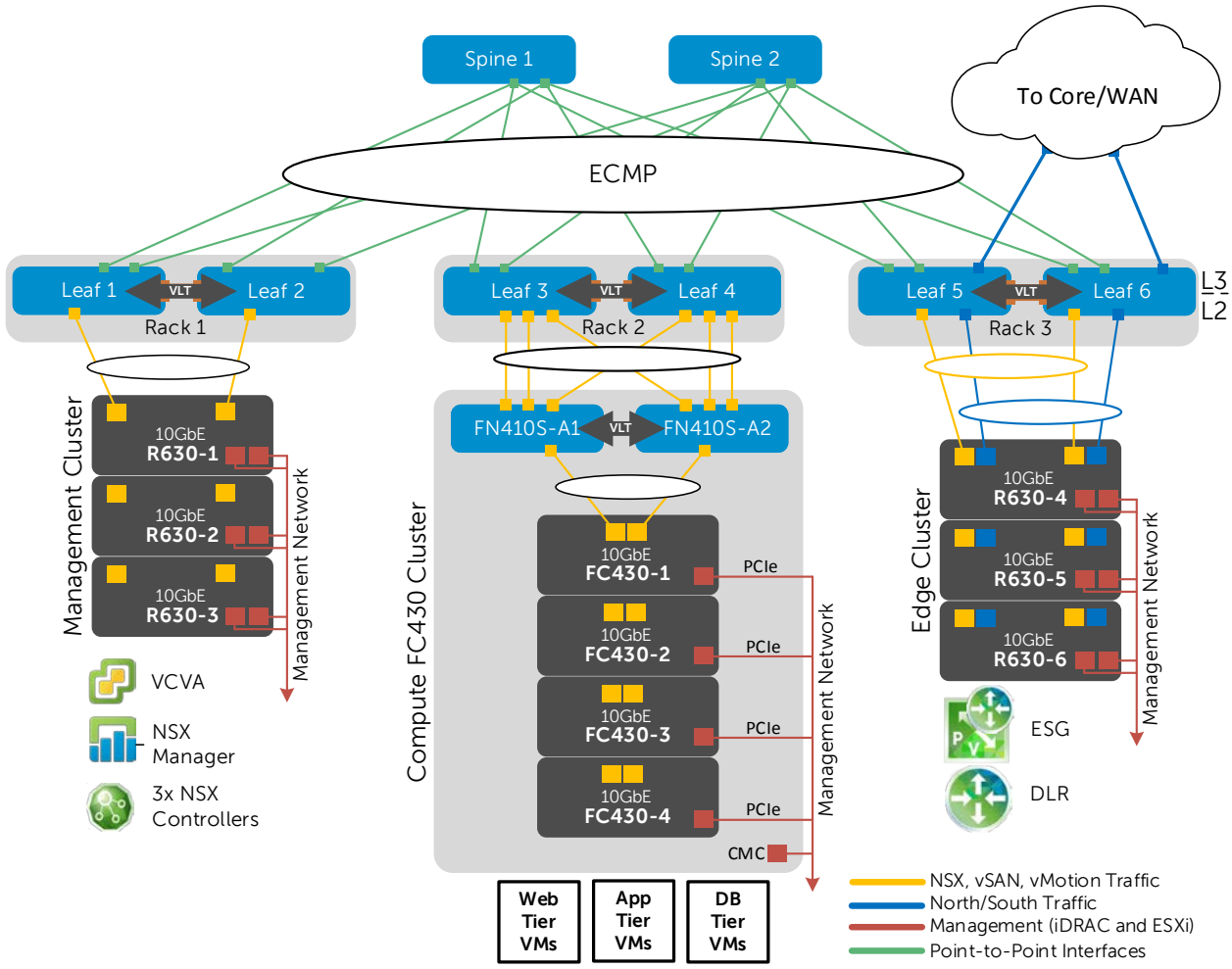


Figure 13 Combined physical and virtual networks

3.3 Management network

This guide uses a single management traffic network that is isolated from the production network. An S3048-ON switch installed in each rack provides connectivity to the management network.

Each R630 server has a 1GbE add-in PCIe network adapter installed for ESXi host management and a built-in iDRAC for out-of-band (OOB) server management. Each FX2s chassis has four 1GbE add-in PCIe network adapters (each connected internally to an FC430 server) for ESXi host management and a built-in CMC for OOB management.

These devices, in addition to the S3048-ON, S4048-ON and Z9100-ON switch management ports, are all connected to the management network as shown in Figure 14.

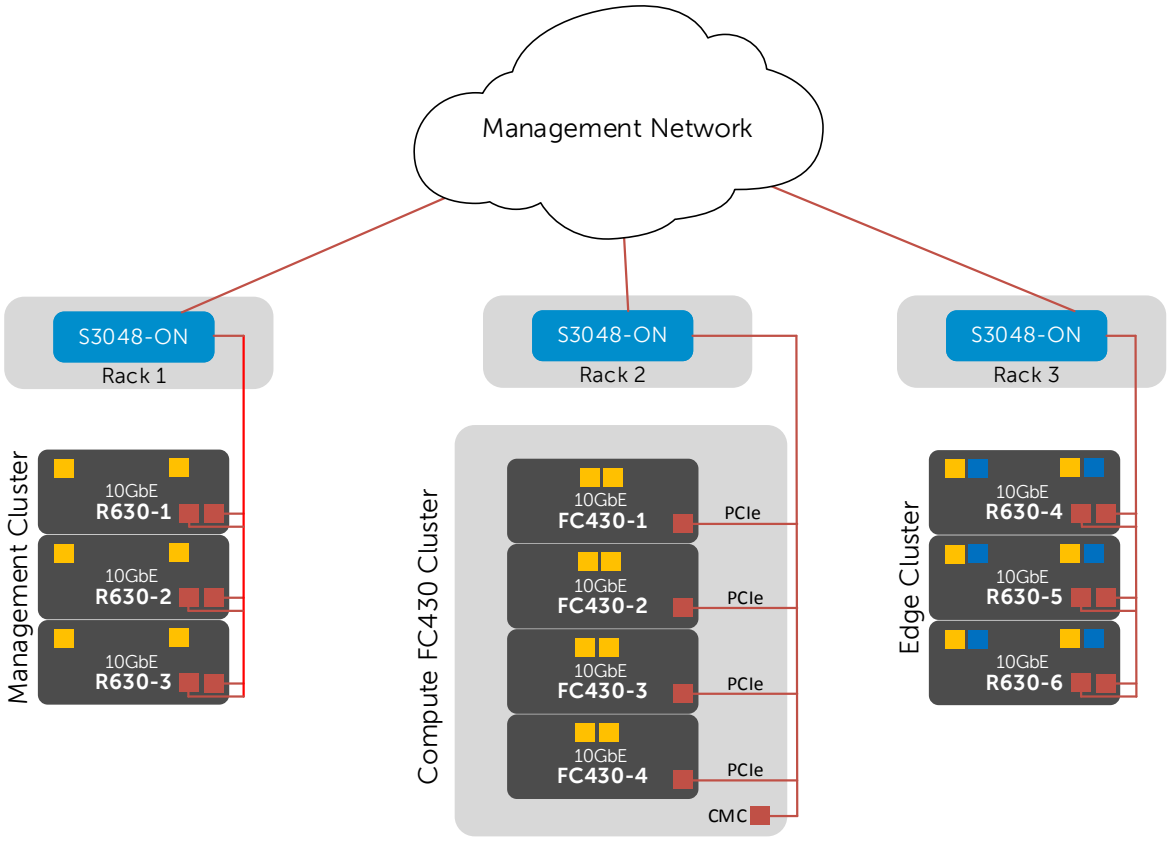


Figure 14 Physical layout of iDRAC, CMC and ESXi management interfaces

4 Network connections

This section details the physical network connections in each cluster.

4.1 Production network connections

4.1.1 Management cluster

Figure 15 shows three PowerEdge R630 servers in Rack 1 connected to two S4048-ON switches (Leaf 1 and Leaf 2) via QLogic 57810 dual-port NDCs. The leaf switches are VLT peers and one NDC port from each server connects to each leaf.

Note: Optionally, QLogic 57840 quad-port NDCs may be used in R630 servers. Two NDC ports are used in management cluster servers in this guide.

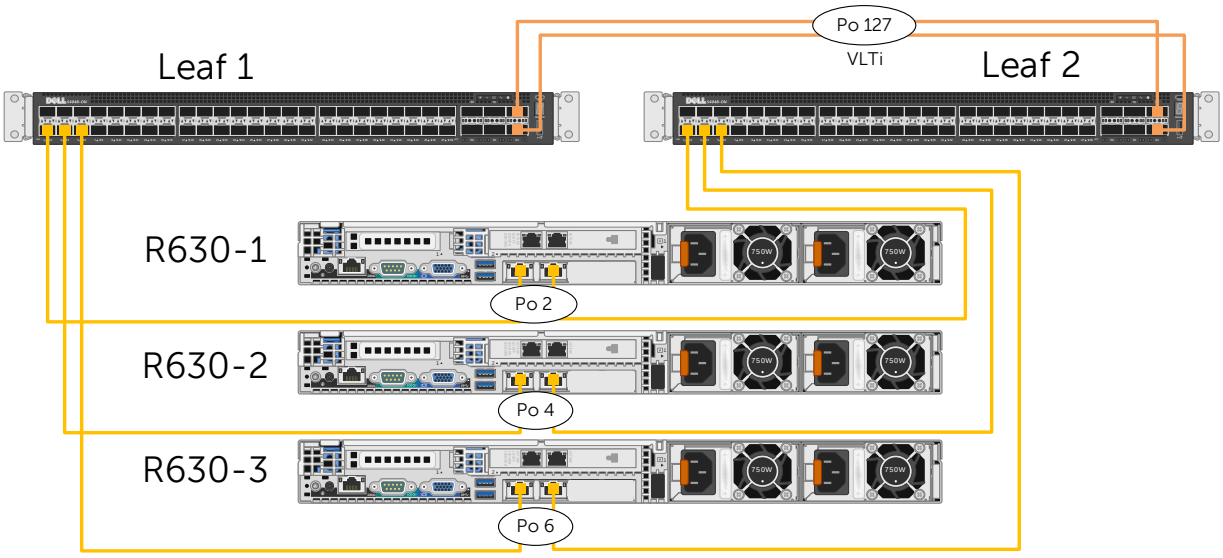


Figure 15 Production network connections for the management cluster

4.1.2 Compute cluster

Figure 16 shows the compute cluster connections from the FN410S switches in the FX2s chassis to Leaf 3 and Leaf 4 in Rack 2. The leaf switches are VLT peers and three FN410S ports connect to each leaf switch. The FN410S switches are also VLT peers. The fourth FN410S port functions as the VLTi (VLT interconnect) between the switches.

Inside the FX2s chassis (not shown), four PowerEdge FC430 servers connect via QLogic 57810 dual-port network adapters to FN410S-A1 and A2. For each server, one link connects internally to FN410S-A1 and the other connects to FN410S-A2.

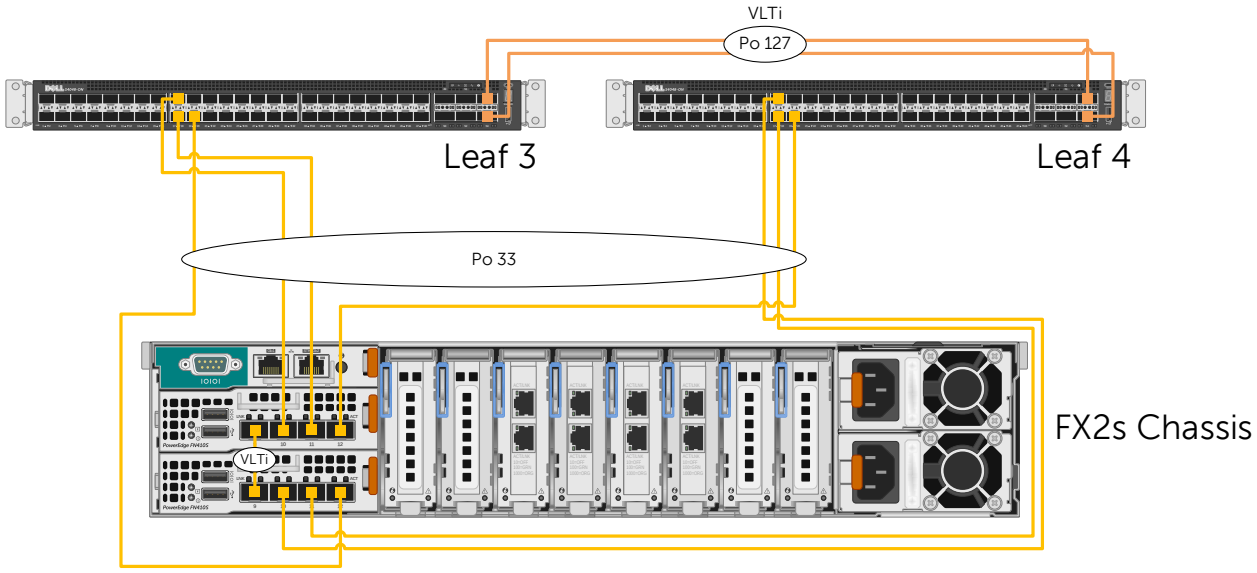


Figure 16 Production network connections for the compute cluster

4.1.3 Edge cluster

As Figure 17 shows, in Rack 3, three PowerEdge R630 servers connect to S4048-ON switches, Leaf 5 and Leaf 6, via QLogic 57840 quad-port NDCs. The yellow connections are used for East-West connections to Racks 1 and 2, and the blue connections are used for North-South connections to the network core or WAN.

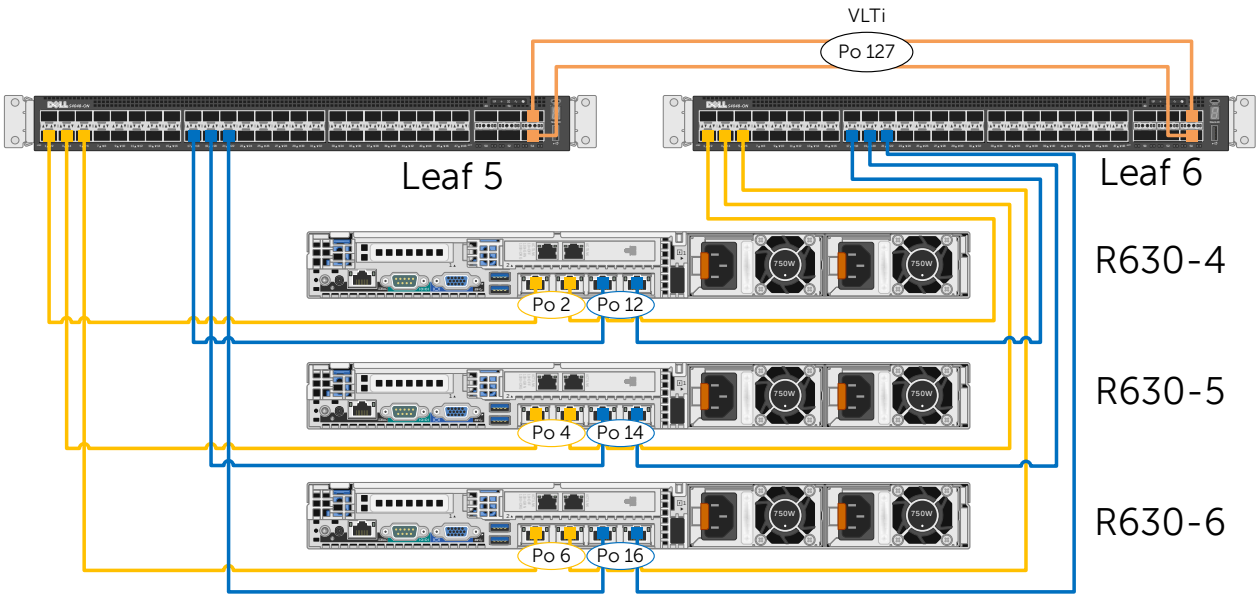


Figure 17 Production network connections for the edge cluster

4.2 Management network connections

These connections are used for non-production, management traffic.

4.2.1 Management and edge clusters

In the management cluster, servers R630-1 through R630-3 are connected to an S3048-ON switch via add-in Intel I350-T dual port PCIe adapters. The R630 server iDRACs are connected to the same switch as shown in Figure 18. The edge cluster is identical and uses servers R630-4 through R630-6.

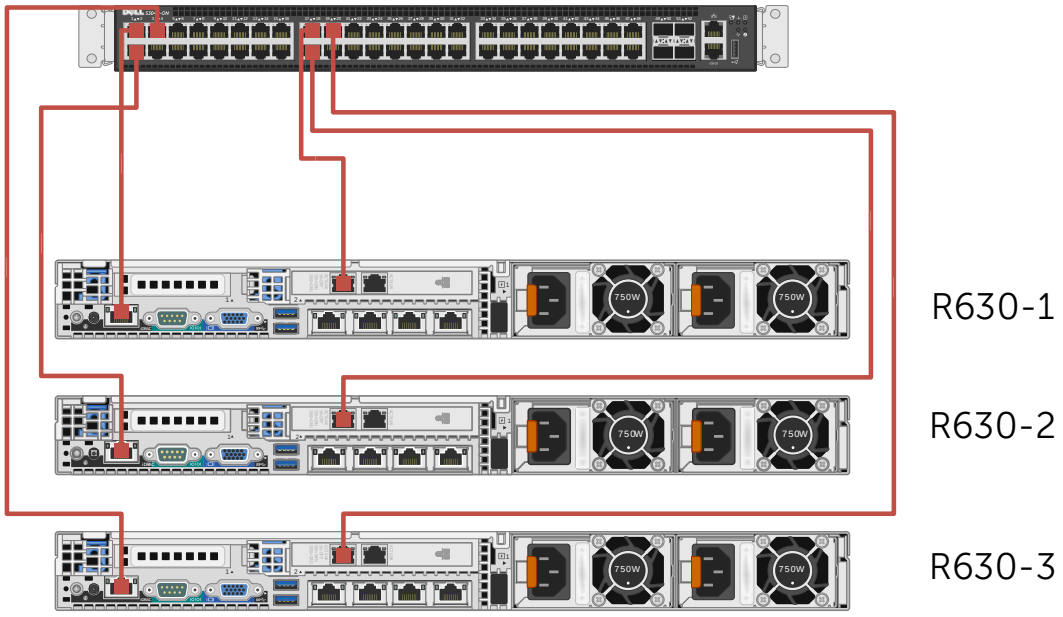


Figure 18 Management cluster – management network connections (edge cluster is identical)

4.2.2 Compute cluster

For management traffic, four FC430 servers are connected to an S3048-ON switch via four Intel I350-T add-in adapters in the FX2s chassis. The FX2s CMC is also connected for OOB management.

For scalability, S3048-ON ports 1–9 are allocated for FX2s chassis CMC ports. Ports 14–48 are available for Intel I350-t management interfaces (four for each FX2s). This allows up to nine FX2s chassis per S3048-ON switch.

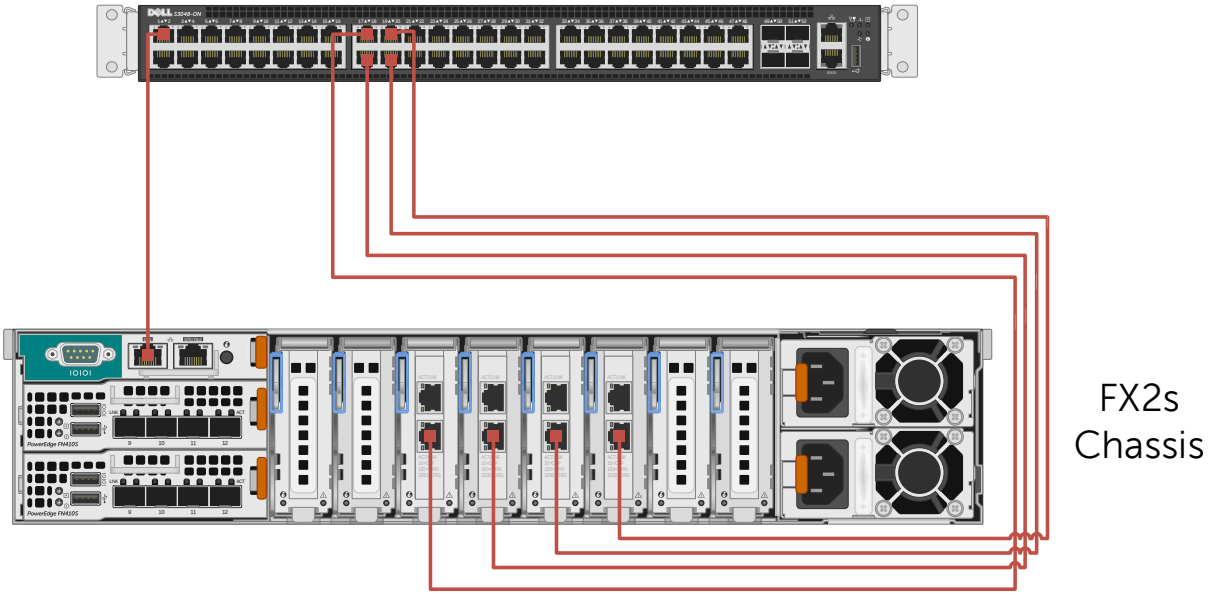


Figure 19 Compute cluster – management network connections

5 Spine and leaf topology

In a spine and leaf architecture, a series of access layer (top-of-rack) switches form the leaf switches. These switches are fully meshed to a series of spine switches. Each leaf connects to each spine, but the spines do not connect to one another. The total number of connections is equal to the number of leaf switches multiplied by the number of spine switches.

The mesh ensures that leaf switches are no more than one hop away from one another, minimizing latency and the likelihood of bottlenecks between leaf switches. Given any single-link failure scenario, all leaf switches retain connectivity to one another through the remaining links.

The connections between spine switches and leaf switches can be layer 2 or layer 3. The deployment scenario in this guide uses layer 3 connections. This limits layer 2 broadcast domains, resulting in improved network stability and scalability.

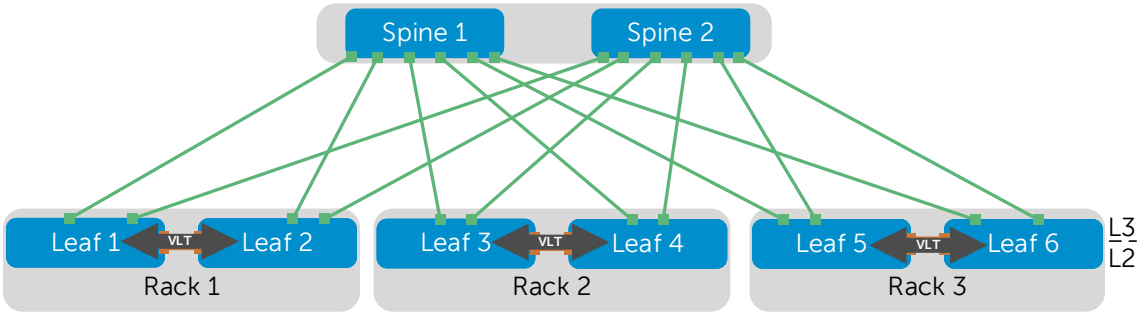


Figure 20 Spine and leaf topology example

Figure 20 shows a high-level diagram of the spine and leaf topology used in this guide with Z9100-ON switches as spines and S4048-ON switches as leaf switches.

The Z9100-ON supports a maximum number of 32 leaf switches. The example in this document uses six leaf switches in three racks. Two leaf switches are used in each rack for redundancy. The first rack contains the management cluster, the second rack contains the compute cluster and the edge cluster is in the third rack.

As administrators add racks to the data center, two leaf switches are added to each new rack. As bandwidth requirements increase, spine switches are added as needed. Scaling guidance is covered in Section 14.

5.1 Routing protocol selection

Choose from any of the following three routing protocols when designing a spine and leaf network:

- Border gateway protocol (External or Internal BGP)
- Open Shortest Path First (OSPF)
- Intermediate System to Intermediate System (IS-IS).

BGP was selected for this guide for scalability. BGP can be configured as *External* BGP (EBGP) to route between autonomous systems or *Internal* BGP (IBGP) to route within a single autonomous system.

EBGP excels at prefix filtering, traffic engineering and traffic tagging. This allows BGP to match on any attribute or prefix and prune prefixes between switches. Unlike EBGP, IBGP requires BGP add-path to support ECMP. To handle peering, IBGP requires route reflectors to mitigate the protocol's full-mesh requirement.

For scalability and the reasons described above, an EBGP deployment is used in this guide.

5.2 BGP ASN configuration

BGP has a reserved, private, 2-byte Autonomous System Number (ASN) range from 64,512 to 65,535. For this EBGP configuration, each switch is assigned a separate ASN. Figure 21 below shows the ASN assignments used in this guide.

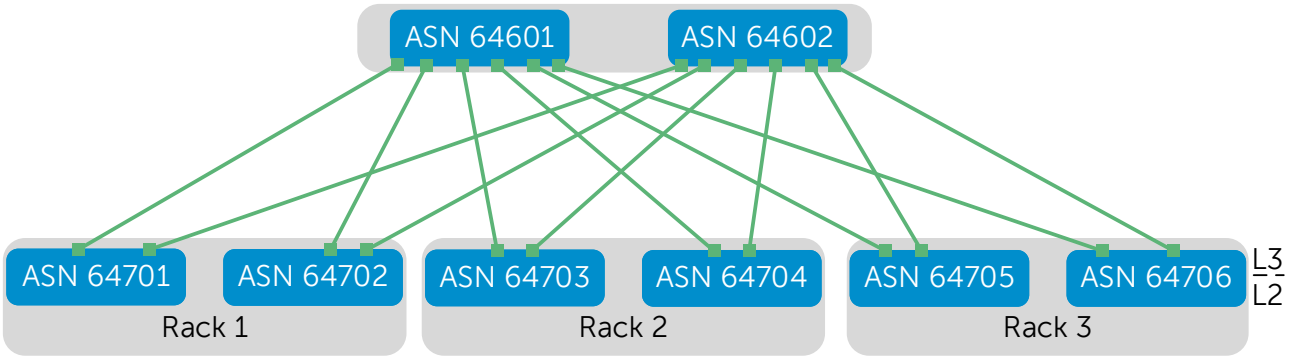


Figure 21 BGP ASN assignments

5.3 BGP fast fall-over

BGP tracks IP reachability to the peer remote address and the peer local address. Whenever either address becomes unreachable (for example, no active route exists in the routing table for the peer IPv4 destination/local address), BGP brings down the session with the peer. This feature is called fast fall-over. Dell EMC recommends enabling fast fall-over for EBGP settings.

5.4 IP Address Management

Proper IP Address Management (IPAM) is critical before deploying a spine and leaf topology. This section covers the IP addressing used on the physical network in this guide.

5.4.1 Loopback addresses

Figure 22 shows the loopback addresses used as router IDs. All loopback addresses are part of the 10.0.0.0/8 address space with each switch using a 32-bit mask.

In this scheme, 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.

This address scheme helps with establishing BGP neighbor adjacencies as well as troubleshooting connectivity.

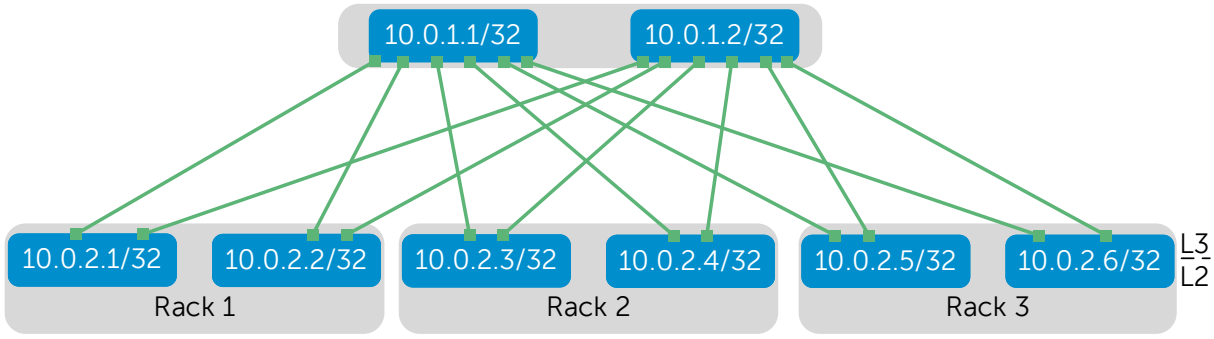


Figure 22 Loopback addressing

5.4.2 Point-to-point addresses

Table 1 below lists layer 3 connection details for each leaf and spine switch. The IP scheme below can be easily extended to account for additional spine and leaf switches.

All addresses come from the same base IP prefix, 192.168.0.0/16 with the 3rd octet representing the spine number. For instance 192.168.1.0/31 is a two host subnet that ties to Spine 1 while 192.168.2.0/31 ties to Spine 2.

Table 1 Interface and IP configuration

Source switch	Rack	Source interface	Source IP	Network	Destination switch	Destination interface	Destination IP	Label
Leaf 1	1	fo1/49	.1	192.168.1.0/31	Spine 1	fo1/1/1	.0	A
Leaf 1	1	fo1/50	.1	192.168.2.0/31	Spine 2	fo1/1/1	.0	B
Leaf 2	1	fo1/49	.3	192.168.1.2/31	Spine 1	fo1/2/1	.2	C
Leaf 2	1	fo1/50	.3	192.168.2.2/31	Spine 2	fo1/2/1	.2	D
Leaf 3	2	fo1/49	.5	192.168.1.4/31	Spine 1	fo1/3/1	.4	E
Leaf 3	2	fo1/50	.5	192.168.2.4/31	Spine 2	fo1/3/1	.4	F
Leaf 4	2	fo1/49	.7	192.168.1.6/31	Spine 1	fo1/4/1	.6	G
Leaf 4	2	fo1/50	.7	192.168.2.6/31	Spine 2	fo1/4/1	.6	H
Leaf 5	3	fo1/49	.9	192.168.1.8/31	Spine 1	fo1/5/1	.8	I
Leaf 5	3	fo1/50	.9	192.168.2.8/31	Spine 2	fo1/5/1	.8	J
Leaf 6	3	fo1/49	.11	192.168.1.10/31	Spine 1	fo1/6/1	.10	K
Leaf 6	3	fo1/50	.11	192.168.2.10/31	Spine 2	fo1/6/1	.10	L

Figure 23 shows the links from Table 1:

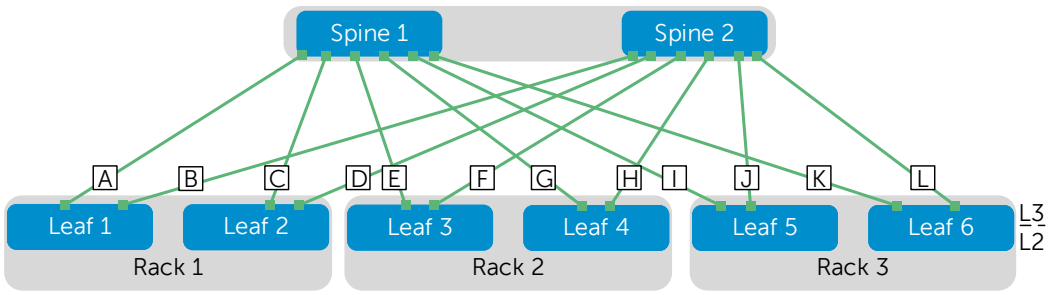


Figure 23 Point-to-point IP addressing

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 S4048-ON. The warning message can be safely ignored on point-to-point interfaces.

```
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.
```

5.4.3 VLANs and IP addressing

Table 2 outlines the VLAN IDs, network and gateway addresses used. The "x" in each network address is replaced by the rack number to create a different network for each rack. The gateway address is the Virtual Router Redundancy Protocol (VRRP) group address, described in the next section. The VLANs and networks are advertised through the BGP instance at the same cost.

Table 2 VLAN and network examples

VLAN ID	Network	Gateway	Used For
22	10.22.x.0/24	10.22.x.254	vMotion
44	10.44.x.0/24	10.44.x.254	VSAN
55	10.55.x.0/24	10.55.x.254	NSX

5.5 VRRP

VRRP is designed to eliminate a single point of failure in a routed network. VRRP is used to create a virtual router which is an abstraction of the two physical leaf switches. The virtual router is assigned an IP address that is used as the gateway address by the compute nodes. In the event that one of the leaf switches fails, the remaining leaf acts as the gateway until the failed unit recovers.

As illustrated in Figure 24, Node 1 is participating in VLAN 55 in Rack 2. The node has an IP address of 10.55.2.1. The node's gateway address is set to 10.55.2.254. This is the Virtual IP (VIP) provided by the VRRP instance running between leaf switches 3 and 4.

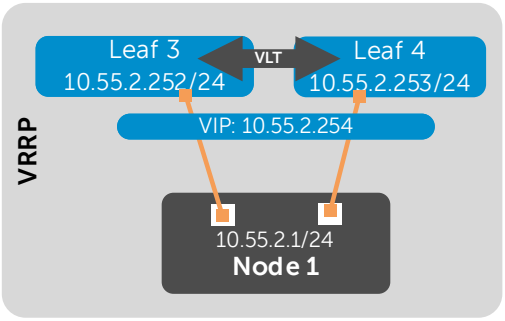


Figure 24 VRRP configuration example – VLAN 55 in Rack 2

A VRRP instance is created for each VLAN in each pair of leaf switches at the top of each rack.

Table 3 shows the VRRP IP addressing scheme for VLAN 55 as an example. The numbering scheme is also used for the other VLANs (VLAN 22 and VLAN 44), with the 2nd octet in the IP addresses replaced with the VLAN number.

Table 3 VRRP interface configuration for VLAN 55 – Racks 1-3

Rack ID	VLAN	First Leaf VLAN IP	Second Leaf VLAN IP	Virtual IP
Rack 1	55	10.55.1.252/24	10.55.1.253/24	10.55.1.254
Rack 2	55	10.55.2.252/24	10.55.2.253/24	10.55.2.254
Rack 3	55	10.55.3.252/24	10.55.3.253/24	10.55.3.254

5.6 ECMP

ECMP is the core protocol enabling the deployment of a layer 3 spine and leaf topology. ECMP gives each spine and leaf switch the ability to load balance flows across a set of equal next-hops. For example, when using two spine switches, each leaf has a connection to each spine. For every flow egressing a leaf switch, there exists two equal next-hops: one to each spine.

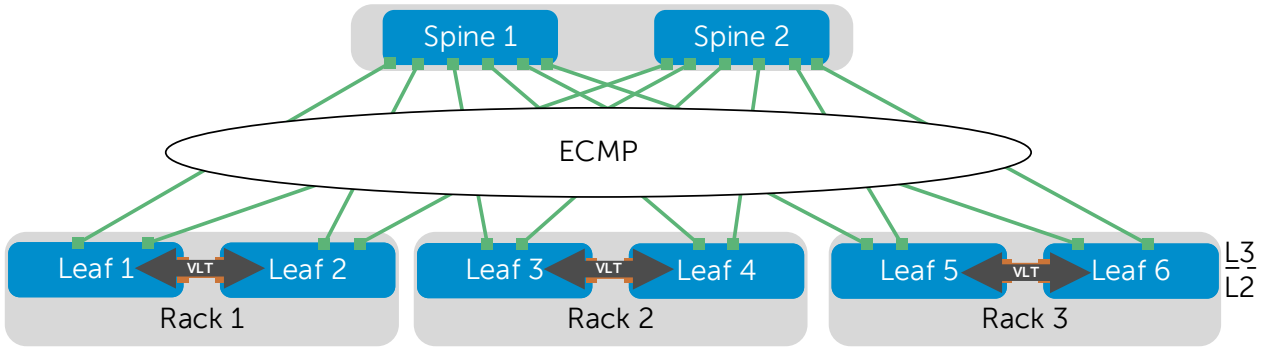


Figure 25 ECMP

5.7 VLT

A pair of leaf switches at the top of each rack provides redundancy. These switches' configurations include the Dell Networking Virtual Link Trunking (VLT) feature.

VLT reduces the role of spanning tree protocols (STPs) by allowing link aggregation group (LAG) terminations on two separate switches and supporting a loop-free topology. VLT provides Layer 2 multipathing and load-balances traffic where alternative paths exist. Virtual Link Trunking offers the following additional benefits:

- Allows a single device to use a LAG across two upstream devices
- Eliminates STP-blocked ports
- Uses all available uplink bandwidth
- Provides fast convergence if either the link or a device fails
- Provides link-level resiliency
- Assures high availability

5.8 Uplink Failure Detection

If a leaf switch loses connectivity to the spine layer, the attached hosts continue to send traffic 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 spine uplinks and the downlink interfaces. An uplink-state group is also configured on each FN410S.

In the event of an uplink failure, UFD automatically shuts down the corresponding downstream interfaces. This propagates down to the hosts attached to the leaf or FN410S switch. The host then uses its remaining Link Aggregation Control Protocol (LACP) port member to continue sending traffic across the leaf-spine network.

6 Configure physical switches

This section contains switch configuration details with explanations for one switch in each major role on the production network. This chapter details the following switches:

- FN410S-A1
- S4048-ON: Leaf 1
- S4048-ON: Leaf 5 with edge configuration
- Z9100-ON: Spine 1

The remaining switches use configurations very similar to one of the four configurations above, with the applicable switches specified in each section. Complete configuration files for all switches used on the production network in this guide are provided as attachments.

Notes:
MTU - The MTU is set to 9216 bytes on all switches in the production network in this guide. VXLAN protocol requirements require setting the MTU to at least 1600 bytes on all switches that will handle NSX traffic on your network.
Port Channel Numbering – LACP port channel numbers may be any number in the range 1-128.

6.1 Factory default settings

The configuration commands in the sections below assume switches are at their factory default settings. All switches in this guide can be reset to factory defaults as follows:

```
switch#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.

6.2 FN410S switch configuration

The compute cluster includes a PowerEdge FX2s chassis with four FC430 servers and two FN410S switches.

Each FC430 server has an LACP-enabled port channel connected to internal interfaces of each FN410S. For clarity, only port channel 1 (for server FC430-1) is shown in Figure 26. The remaining port channels are numbered 2-4.

The two FN410S switches are configured as VLT peers. Three of the four FN410S external interfaces, tengigabitethernet 0/10–12, are configured in port channel 33 which is connected to leaf switches 3 and 4. The 4th external interface, tengigabitethernet 0/9, is used as the VLT interconnect between FN410S-A1 and FN410S-A2.

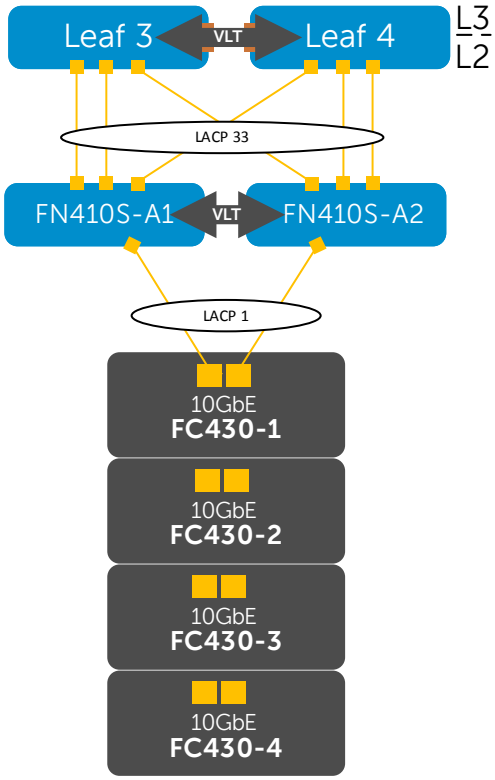


Figure 26 FN410S network connections (internal LACP connections to FC430-2 through 4 not shown)

The following section outlines the configuration commands issued to the FN410S switches. The switches start at their factory default settings per Section 6.1.

After FN410S switches boot to their default settings, place them in full-switch mode as follows:

```
Dell>enable
Dell#configure
Dell(conf)#stack-unit 0 iom-mode full-switch

% You are about to configure the Full Switch Mode.
```

```
Please reload to effect the changes

Dell(conf)#do reload

System configuration has been modified. Save? [yes/no]: yes
Proceed with reload [confirm yes/no]: yes
```

After FN410S switches boot to full-switch mode, enter the following commands to configure the FN410S-A1.

Note: Ensure FN410S switches have been placed in full-switch mode before proceeding. The following configuration details are specific to switch FN410S-A1. The configuration for FN410S-A2 is similar. See the FN410S-A1.txt and FN410S-A2.txt attachments.

Initial configuration involves setting the hostname, enabling Link Layer Discovery Protocol (LLDP) and disabling Data Center Bridging (DCB). LLDP is useful for troubleshooting (see Section 9.8). DCB is enabled by default on FN410S but is not used in this environment.

Enable Internet Group Messaging Protocol (IGMP) snooping for VSAN traffic. Finally, configure the management interface and default gateway.

Note: IGMP snooping is enabled, but IGMP snooping querier is not enabled on the FN410S switches. The querier role is enabled on the S4048-ON leaf switches upstream.

```
enable
configure

hostname FN410S-A1
protocol lldp
advertise management-tlv management-address system-description system-name
advertise interface-port-desc
no dcb enable
ip igmp snooping enable

interface ManagementEthernet 0/0
ip address 172.25.187.151/24
no shutdown

management route 0.0.0.0/0 172.25.187.254
```

Next, the VLT interface between the two switches is configured. In this configuration, interface tengigabitethernet 0/9 is used for the VLTi interface. It is added to static port-channel 127. The backup destination is the management IP address of the VLT peer switch, FN410S-A2. The VLT unit-id is set to 0 (and is set to 1 on FN410S-A2).

```
interface port-channel 127
description VLTi
```

```
mtu 9216
channel-member tengigabitethernet 0/9
no shutdown

interface tengigabitethernet 0/9
description VLTi
no shutdown

vlt domain 127
peer-link port-channel 127
back-up destination 172.25.187.152
unit-id 0
```

The upstream interfaces to the two leaf switches are configured in this section. External interfaces tengigabitethernet 0/10-12 are used and placed in LACP-enabled port channel 33. The port channel is configured for VLT and jumbo frames are enabled for VXLAN traffic.

```
interface range tengigabitethernet 0/10-12
description To Leaf switches 3 and 4 te 1/41-43
mtu 9216
port-channel-protocol LACP
port-channel 33 mode active
no shutdown

interface port-channel 33
description To Leaf switches 3 and 4 te 1/41-43
mtu 9216
portmode hybrid
switchport
vlt-peer-lag port-channel 33
no shutdown
```

The downstream interfaces are configured in the next set of commands. Each internal interface is added to a numerically corresponding port channel. The port channels are configured for VLT and jumbo frames are enabled on all interfaces for VXLAN traffic.

```
interface tengigabitethernet 0/1
description To FC430-1 172.25.187.53
mtu 9216
port-channel-protocol LACP
port-channel 1 mode active
no shutdown

interface port-channel 1
description To FC430-1 172.25.187.53
mtu 9216
portmode hybrid
switchport
```

```
vlt-peer-lag port-channel 1
no shutdown

interface tengigabitethernet 0/2
description To FC430-2 172.25.187.54
mtu 9216
port-channel-protocol LACP
port-channel 2 mode active
no shutdown

interface port-channel 2
description To FC430-2 172.25.187.54
mtu 9216
portmode hybrid
switchport
vlt-peer-lag port-channel 2
no shutdown

interface tengigabitethernet 0/3
description To FC430-3 172.25.187.55
mtu 9216
port-channel-protocol LACP
port-channel 3 mode active
no shutdown

interface port-channel 3
description To FC430-3 172.25.187.55
mtu 9216
portmode hybrid
switchport
vlt-peer-lag port-channel 3
no shutdown

interface tengigabitethernet 0/4
description To FC430-4 172.25.187.56
mtu 9216
port-channel-protocol LACP
port-channel 4 mode active
no shutdown

interface port-channel 4
description To FC430-4 172.25.187.56
mtu 9216
portmode hybrid
switchport
vlt-peer-lag port-channel 4
```

Finally, the three required VLAN interfaces are created. All downstream and upstream port channels are tagged in each VLAN.

```
interface Vlan 22
description vMotion
mtu 9216
tagged Port-channel 1-4,33
no shutdown

interface Vlan 44
description VSAN
mtu 9216
tagged Port-channel 1-4,33
no shutdown

interface Vlan 55
description NSX
mtu 9216
tagged Port-channel 1-4,33
no shutdown
```

UFD is configured. This shuts 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.

```
uplink-state-group 1
description Disable downstream ports in event all uplinks fail
downstream TenGigabitEthernet 0/1-8
upstream TenGigabitEthernet 0/10-12
```

Save the configuration.

```
end
write
```

6.3 S4048-ON leaf switch configuration

Each S4048-ON leaf switch has an LACP-enabled port channel connected to each of the downstream R630 servers, or in the case of the compute cluster, the downstream FN410S switches.

There are a total of six leaf switches in this guide, with two in each rack configured as VLT peers.

The following section outlines the configuration commands issued to S4048-ON leaf switches. The switches start at their factory default settings per Section 6.1.

Note: The following configuration details are specific to Leaf 1. The remaining leaf switches, 2-6, are similar. Leaf switches 3-4 have a different downstream port channel configuration. Leaf switches 5-6 have additional edge configuration steps that are covered in the next section. Complete configuration details for all six leaf switches are provided in the attachments named leaf1.txt through leaf6.txt.

For VSAN traffic, IGMP snooping and IGMP snooping querier are enabled on leaf switches. IGMP snooping is enabled globally, and IGMP querier is enabled on VLAN 44.

Initial configuration involves setting the hostname, and enabling LLDP and IGMP snooping. LLDP is useful for troubleshooting (see Section 9.8). IGMP snooping is enabled for VSAN traffic. Finally, the management interface and default gateway are configured.

```
enable
configure

hostname Leaf-1
protocol lldp
advertise management-tlv management-address system-description system-name
advertise interface-port-desc
ip igmp snooping enable

interface ManagementEthernet 1/1
ip address 172.25.187.35/24
no shutdown

management route 0.0.0.0/0 172.25.187.254
```

Next, the VLT interfaces between Leaf-1 and Leaf-2 are configured. In this configuration, interfaces fortyGigE 1/53-54 are used for the VLT interconnect. They are added to static port-channel 127. The backup destination is the management IP address of the VLT peer switch, Leaf-2.

```
interface port-channel 127
description VLTi
mtu 9216
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 172.25.187.34
unit-id 0
exit
```

The downstream interfaces, to the R630 servers in this case, are configured in the next set of commands. Each interface is added to a numerically corresponding port channel. The port channels are configured for VLT and jumbo frames are enabled on all interfaces for VXLAN traffic.

```
interface tengigabitethernet 1/2
description To R630-1 172.25.187.19
mtu 9216
port-channel-protocol LACP
port-channel 2 mode active
no shutdown

interface port-channel 2
description To R630-1 172.25.187.19
mtu 9216
portmode hybrid
switchport
vlt-peer-lag port-channel 2
no shutdown

interface tengigabitethernet 1/4
description To R630-2 172.25.187.18
mtu 9216
port-channel-protocol LACP
port-channel 4 mode active
no shutdown

interface port-channel 4
description To R630-2 172.25.187.18
mtu 9216
portmode hybrid
switchport
vlt-peer-lag port-channel 4
no shutdown

interface tengigabitethernet 1/6
description To R630-3 172.25.187.17
mtu 9216
port-channel-protocol LACP
port-channel 6 mode active
no shutdown

interface port-channel 6
description To R630-3 172.25.187.17
mtu 9216
portmode hybrid
switchport
```

```
vlt-peer-lag port-channel 6
no shutdown
```

The three required VLAN interfaces are created. All downstream port channels are tagged in each VLAN. Each interface is assigned to a VRRP group and a VRRP address is assigned. VRRP priority is set to 254 to make this switch the master. (On the VRRP peer switch, priority is set to 1).

```
interface Vlan 22
description vMotion
ip address 10.22.1.252/24
mtu 9216
tagged Port-channel 2,4,6
vrrp-group 22
description vMotion
priority 254
virtual-address 10.22.1.254
no shutdown

interface Vlan 44
description VSAN
ip address 10.44.1.252/24
mtu 9216
tagged Port-channel 2,4,6
ip igmp snooping querier
vrrp-group 44
description VSAN
priority 254
virtual-address 10.44.1.254
no shutdown

interface Vlan 55
description NSX
ip address 10.55.1.252/24
mtu 9216
tagged Port-channel 2,4,6
vrrp-group 55
description NSX
priority 254
virtual-address 10.55.1.254
no shutdown
```

The upstream layer 3 interfaces connected to the spines are configured. A loopback interface is configured as the router ID for BGP.

```
interface fortyGigE 1/49
description To Spine-1
ip address 192.168.1.1/31
```

```
mtu 9216
no shutdown

interface fortyGigE 1/50
description To Spine-2
ip address 192.168.2.1/31
mtu 9216
no shutdown

interface loopback 0
description Router ID
ip address 10.0.2.1/32
```

BGP is configured to allow routing to the IP fabric. Additionally, an IP prefix and route map are created to automatically redistribute all leaf subnets and loopback addresses from the spine and leaf switches.

```
route-map spine-leaf permit 10
match ip address spine-leaf

ip prefix-list spine-leaf
description BGP redistribute loopback and leaf networks
seq 5 permit 10.0.0.0/23 ge 32
seq 10 permit 10.0.0.0/8 ge 24
router bgp 64701
bgp bestpath as-path multipath-relax
maximum-paths ebgp 64
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 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
```

An ECMP group is created that includes the point-to-point interfaces to the two spine switches.

```
ecmp-group 1
interface fortyGigE 1/49
interface fortyGigE 1/50
link-bundle-monitor enable
```

UFD is configured. This shuts 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.

```
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
```

Save the configuration.

```
end
write
```

6.3.1 S4048-ON Edge configuration

The following section contains additional configuration steps required on leaf switches 5 and 6 connected to the core/WAN shown in Figure 27 below.

Note: Only the north/south (blue) links to the core/WAN are configured in this section. The remaining links were configured in the previous section. The following configuration details are specific to Leaf 5. Leaf 6 is similar. Complete configuration details are provided in the attachments named leaf5.txt and leaf6.txt.

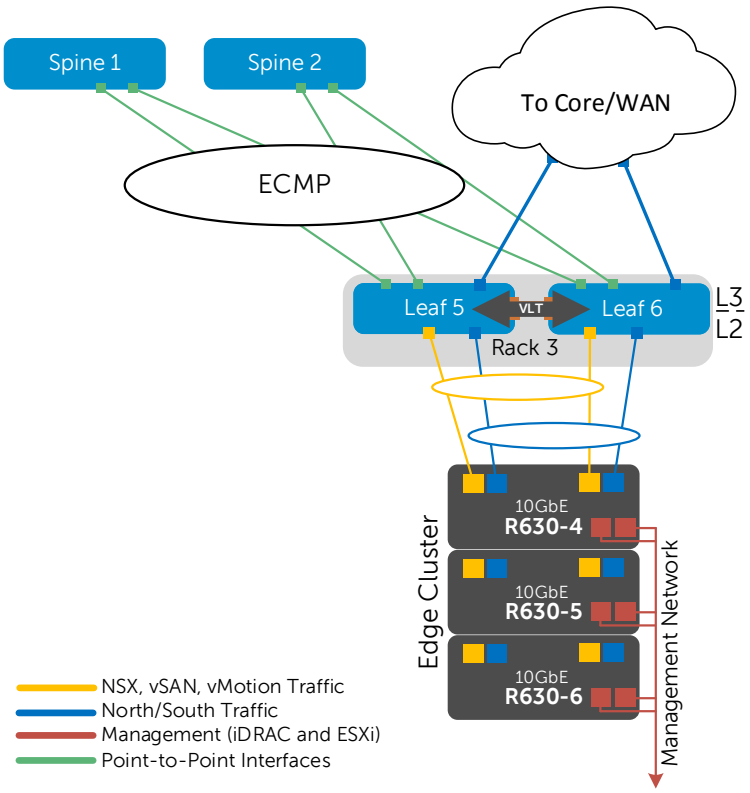


Figure 27 Edge cluster leaf switch configuration

Enable layer 3 VLT peer-routing. This will allow leaf 5 and leaf 6 to create an OSPF neighbor adjacency across the VLTi link.

```
vlt domain 127
peer-routing
```

Add the edge/wan/core links to the R630 edge servers

```
interface tengigabitethernet 1/18
description Edge To R630-4 172.25.187.14
port-channel-protocol LACP
port-channel 12 mode active
no shutdown

interface port-channel 12
description Edge To R630-4 172.25.187.14
portmode hybrid
switchport
vlt port-channel 12
no shutdown

interface tengigabitethernet 1/20
description Edge To R630-5 172.25.187.15
port-channel-protocol LACP
port-channel 14 mode active
no shutdown

interface port-channel 14
description Edge To R630-5 172.25.187.15
portmode hybrid
switchport
vlt port-channel 14
no shutdown

interface tengigabitethernet 1/22
description Edge To R630-6 172.25.187.16
port-channel-protocol LACP
port-channel 16 mode active
no shutdown

interface port-channel 16
description Edge To R630-6 172.25.187.16
portmode hybrid
switchport
vlt port-channel 16
no shutdown
```

Add VLAN 66. This VLAN is dedicated to handling north/south traffic and does not use VRRP. The interface is used to create an Open Shortest Path First (OSPF) router adjacency and does not require a VRRP group address for forwarding.

```
interface vlan 66
description Edge
ip address 10.66.3.252/24
tagged Port-channel 12,14,16
no shutdown
```

Create an OSPF routing process to handle north to south traffic. A specific router ID is specified here to separate the neighbor relationship tables. This OSPF instance will created a neighbor relationship with Leaf 6 as well as the ESG, configured in Section 12.

```
router ospf 1
network 10.66.3.0/24 area 0
router-id 10.66.3.252
```

Save the configuration.

```
end
write
```

6.4 Z9100-ON spine switch configuration

Note: The following configuration details are specific to Spine 1. Spine 2 is similar. Complete configuration details are provided in the attachments named spine1.txt and spine2.txt.

Set the hostname, enable LLDP, and configure the management interface. Set the interface speed to 40GbE for all interfaces used for point-to-point links with the six leaf switches.

```
enable
configure

hostname Spine-1
protocol lldp
advertise management-tlv management-address system-description system-name
advertise interface-port-desc

interface ManagementEthernet 1/1
ip address 172.25.187.39/24
no shutdown
management route 0.0.0.0/0 172.25.187.254

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
stack-unit 1 port 5 portmode single speed 40G no-confirm
stack-unit 1 port 6 portmode single speed 40G no-confirm
```

The point to point interfaces and loop back interface are configured.

```
interface fortyGigE 1/1/1
description To Leaf 1 fo1/49
ip address 192.168.1.0/31
mtu 9216
no shutdown

interface fortyGigE 1/2/1
description To Leaf 2 fo1/49
ip address 192.168.1.2/31
mtu 9216
no shutdown

interface fortyGigE 1/3/1
description To Leaf 3 fo1/49
ip address 192.168.1.4/31
mtu 9216
no shutdown

interface fortyGigE 1/4/1
description To Leaf 4 fo1/49
ip address 192.168.1.6/31
mtu 9216
no shutdown

interface fortyGigE 1/5/1
description To Leaf 5 fo1/49
ip address 192.168.1.8/31
mtu 9216
no shutdown

interface fortyGigE 1/6/1
description To Leaf 6 fo1/49
ip address 192.168.1.10/31
mtu 9216
no shutdown

interface loopback 0
description Router ID
ip address 10.0.0.1/32
```

BGP is configured to allow routing to the IP fabric. Additionally, an IP prefix and route map are created to automatically redistribute all leaf subnets as well as loopback addresses from the spine and leaf switches.

```
route-map spine-leaf permit 10
match ip address spine-leaf

ip prefix-list spine-leaf
description BGP redistribute loopback and leaf networks
seq 5 permit 10.0.0.0/23 ge 32
seq 10 permit 10.0.0.0/8 ge 24
router bgp 64601
bgp bestpath as-path multipath-relax
maximum-paths ebgp 64
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 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
neighbor 192.168.1.9 remote-as 64705
neighbor 192.168.1.9 peer-group spine-leaf
neighbor 192.168.1.9 no shutdown
neighbor 192.168.1.11 remote-as 64706
neighbor 192.168.1.11 peer-group spine-leaf
neighbor 192.168.1.11 no shutdown
```

Create an ECMP group and include the point to point interfaces from the two spine switches.

```
ecmp-group 1
interface fortyGigE 1/1/1
interface fortyGigE 1/2/1
interface fortyGigE 1/3/1
interface fortyGigE 1/4/1
interface fortyGigE 1/5/1
```

```
interface fortyGigE 1/6/1
link-bundle-monitor enable
```

Save the configuration.

```
end
write
```

6.5 S3048-ON management switch configuration

For the S3048-ON management switches, all ports used are in layer 2 mode and are in the default VLAN. No additional configuration is required.

6.6 Verify switch configuration

The following sections show commands and output to verify switches are configured and connected properly. Except where there are key differences, only output from one spine switch, one leaf switch, and one FN410S switch is shown to avoid repetition. Output from remaining devices will be similar.

6.6.1 Z9100-ON Spine Switch

6.6.1.1 show ip bgp summary

This command verifies each BGP session to each of the six leaf switches is connected and sharing prefixes.

```
Spine-1#show ip bgp summary
BGP router identifier 10.0.0.1, local AS number 64601
BGP local RIB : Routes to be Added 0, Replaced 0, Withdrawn 0
16 network entrie(s) using 1216 bytes of memory
26 paths using 2808 bytes of memory
BGP-RIB over all using 2834 bytes of memory
41 BGP path attribute entrie(s) using 6880 bytes of memory
39 BGP AS-PATH entrie(s) using 390 bytes of memory
6 neighbor(s) using 49152 bytes of memory

Neighbor      AS     MsgRcvd  MsgSent    TblVer  InQ  OutQ Up/Down  State/Pfx
192.168.1.1   64701    5032     5013          0    0     0 3d:00:50:29 5
192.168.1.3   64702   17469    17469          0    0     0 00:00:00 Idle
192.168.1.5   64703    5031     5032          0    0     0 3d:00:50:44 5
192.168.1.7   64704    5030     5028          0    0     0 3d:00:50:29 5
192.168.1.9   64705      64       69          0    0     0 00:50:42 5
192.168.1.11  64706      62       66          0    0     0 00:49:33 5
```

6.6.1.2 show ip route bgp

This command is used to verify the BGP instance entries in the Routing Information Base (RIB) and ECMP. The first set of routes with a subnet mask of /32 are the IPs configured for router IDs.

The second set of routes with a /24 mask represents the 3 networks used, vMotion, VSAN, and NSX. Note that for each of these networks there are two routes. For example, 10.22.1.0/24 is reachable via 192.168.1.1 and 192.168.1.3, Leaf 1 and Leaf 2 respectively.

```
Spine-1#show ip route bgp
       Destination        Gateway            Dist/Metric Last Change
       -----------        -------           ----------- -----------
  B EX 10.0.1.2/32        via 192.168.1.1          20/0    00:00:22
                          via 192.168.1.3
                          via 192.168.1.9
                          via 192.168.1.7
                          via 192.168.1.5
                          via 192.168.1.11
  B EX 10.0.2.1/32        via 192.168.1.1          20/0    00:13:57
  B EX 10.0.2.2/32        via 192.168.1.3          20/0    00:05:36
  B EX 10.0.2.3/32        via 192.168.1.5          20/0    00:13:18
  B EX 10.0.2.4/32        via 192.168.1.7          20/0    00:12:37
  B EX 10.0.2.5/32        via 192.168.1.9          20/0    00:12:06
  B EX 10.0.2.6/32        via 192.168.1.11         20/0    00:11:47
  B EX 10.22.1.0/24       via 192.168.1.1          20/0    00:05:36
                          via 192.168.1.3
  B EX 10.22.2.0/24       via 192.168.1.5          20/0        3d1h
                          via 192.168.1.7
  B EX 10.22.3.0/24       via 192.168.1.9          20/0    01:41:59
                          via 192.168.1.11
  B EX 10.44.1.0/24       via 192.168.1.1          20/0    00:05:36
                          via 192.168.1.3
  B EX 10.44.2.0/24       via 192.168.1.5          20/0        3d1h
                          via 192.168.1.7
  B EX 10.44.3.0/24       via 192.168.1.9          20/0    01:41:59
                          via 192.168.1.11
  B EX 10.55.1.0/24       via 192.168.1.1          20/0    00:05:36
                          via 192.168.1.3
  B EX 10.55.2.0/24       via 192.168.1.5          20/0        3d1h
                          via 192.168.1.7
  B EX 10.55.3.0/24       via 192.168.1.9          20/0    01:41:59
                          via 192.168.1.11
```

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

6.6.1.3 show ip route <network>

This command is used to verify that routes leading to the appropriate leaf switches are being propagated from BGP to the RIB. The commands for the 10.55.x.0 network are shown below as an example.

```
Spine-1#show ip route 10.55.1.0/24
Routing entry for 10.55.1.0/24
    Known via "bgp 64601", distance 20, metric 0
    Last update 00:23:33 ago
    Tag value 64701
    Routing Descriptor Blocks:
    * via 192.168.1.1
    * via 192.168.1.3

Spine-1#show ip route 10.55.2.0/24
Routing entry for 10.55.2.0/24
    Known via "bgp 64601", distance 20, metric 0
    Last update 3d2h ago
    Tag value 64703
    Routing Descriptor Blocks:
    * via 192.168.1.5
    * via 192.168.1.7

Spine-1#show ip route 10.55.3.0/24
Routing entry for 10.55.3.0/24
    Known via "bgp 64601", distance 20, metric 0
    Last update 02:00:32 ago
    Tag value 64705
    Routing Descriptor Blocks:
    * via 192.168.1.9
    * via 192.168.1.11
```

6.6.1.4 Ping VRRP addresses

Both spine switches must be able to ping the VRRP addresses configured on each leaf switch pair.

```
Spine-1#ping 10.22.1.254
Sending 5, 100-byte ICMP Echos to 10.22.1.254, timeout is 2 seconds:
!!!!!
Success rate is 100.0 percent (5/5), round-trip min/avg/max = 0/0/0 (ms)
```

Repeat for other VRRP addresses such as:

10.22.3.254
10.44.1.254
10.55.3.254
… etc.

Note: VRRP addresses in this document use the format 10.*vlan#.rack#*.254.

6.6.2 S4048-ON Leaf Switch

6.6.2.1 show vlt brief

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

```
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 :                Disabled
 Peer-Routing-Timeout timer:   0 seconds
 Multicast peer-routing timeout: 150 seconds
```

6.6.2.2 show vlt detail

On leaf switches 1 and 2, downstream LAGs (port channels 2,4, and 6) will all be down until LAGs are configured on the directly connected ESXi hosts (covered in Section 9.4). VLANs 1, 22, 44, and 55 are active.

```
Leaf-1#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
2             2            DOWN          DOWN          1, 22, 44, 55
4             4            DOWN          DOWN          1, 22, 44, 55
6             6            DOWN          DOWN          1, 22, 44, 55
```

Leaf switches 5 and 6 have three additional downstream lags (port channels 12, 14 and 16) for edge traffic on VLAN 66. Port channels 12, 14, and 16 will be down until edge-lag2 is configured in Section 13.1.2.

```
Leaf-5#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
2             2            DOWN          DOWN          1, 22, 44, 55
4             4            DOWN          DOWN          1, 22, 44, 55
```

```
6               6               DOWN            DOWN            1, 22, 44, 55
12              12              DOWN            DOWN            1, 66
14              14              DOWN            DOWN            1, 66
16              16              DOWN            DOWN            1, 66
```

On leaf switches 3 and 4, downstream port channel 33 is up because they are connected to properly configured FN410S switches. VLANs 1, 22, 44, and 55 are active.

```
Leaf-3#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
33            33           UP            UP            1, 22, 44, 55
```

6.6.2.3 show vrrp brief

The output from the `show vrrp brief` command should be similar to that shown below. The priority (`Pri` column) of the master router in the pair is 254 and the backup router (not shown) is assigned priority 1.

```
Leaf-1#show vrrp brief

Interface Group    Pri Pre State  Master addr   Virtual addr(s)     Description
-------------------------------------------------------------------------------
Vl 22     IPv4 22  254 Y   Master 10.22.1.252   10.22.1.254         vMotion
Vl 44     IPv4 44  254 Y   Master 10.44.1.252   10.44.1.254         VSAN
Vl 55     IPv4 55  254 Y   Master 10.55.1.252   10.55.1.254         NSX
```

6.6.3 FN410S I/O Module

6.6.3.1 show vlt brief

Like the S4048-ON switches above, the ICL Link Status, Heat Beat Status, and VLT Peer Status must all be up. One switch is primary and the peer (not shown) is the secondary.

```
FN410S-A1#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:     f8:b1:56:6e:fc:5b
 Remote System MAC address:    f8:b1:56:76:b9:b5
 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
```

6.6.3.2 show vlt detail

Downstream LAGs (port channels 1-4) are down until LAGs are configured on the directly connected ESXi hosts running on the FC430 servers. This is covered in Section 9.4.

The upstream LAG (port channel 33) is currently up because it is connected to properly configured leaf switches (Leaf 3 and Leaf 4).

VLANs 1, 22, 44, and 55 are active on all LAGs.

```
FN410S-A1#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
1             1            DOWN          DOWN          1, 22, 44, 55
2             2            DOWN          DOWN          1, 22, 44, 55
3             3            DOWN          DOWN          1, 22, 44, 55
4             4            DOWN          DOWN          1, 22, 44, 55
33            33           UP            UP            1, 22, 44, 55
```

7 Prepare Servers

This section covers basic PowerEdge server preparation and ESXi hypervisor installation. Installation of guest operating systems (Microsoft Windows Server, Red Hat Linux, etc.) is outside the scope of this document.

Note: Exact iDRAC console steps in this section may vary slightly depending on hardware, software and browser versions used. See your PowerEdge server documentation for steps to connect to the iDRAC virtual console.

7.1 Confirm CPU virtualization is enabled in BIOS

Note: CPU virtualization is typically enabled by default in PowerEdge server BIOS. These steps are provided for reference in case this required feature has been disabled.

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the **Next Boot** menu, select **BIOS Setup**.
3. Reboot the server.
4. From the **System Setup Main Menu**, select **System BIOS**, and then select **Processor Settings**.
5. Verify **Virtualization Technology** is set to **Enabled**.
6. To save the settings, click **Back, Finish**, and **Yes** if prompted to save changes.
7. If resetting network adapters to defaults, proceed to step 4, **System Setup Main Menu**, in the next section. Otherwise, reboot the server.

7.2 Confirm network adapters are at factory default settings

Note: These steps are only necessary if installed network adapters have been modified from their factory default settings.

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the **Next Boot** menu, select **BIOS Setup**.
3. Reboot the server.
4. From the **System Setup Main Menu**, select **Device Settings**.
5. From the **Device Settings** page, select the first port of the first NIC in the list.
6. From the **Main Configuration Page**, click the **Default** button followed by **Yes** to load the default settings. Click **OK**.
7. To save the settings, click **Finish** then **Yes** to save changes. Click **OK**.
8. Repeat for each NIC and port listed on the **Device Settings** page.
9. Reboot the server.

7.3 Confirm storage controllers for VSAN disks are in HBA mode

For redundancy, VSANs employ software RAID. With the exception of single drive RAID-0 configurations, VSANs do not support hardware RAID.

Note: For more information see the VMware Virtual SAN Hardware Guidance white paper.

Storage controllers used in VSANs should therefore be set to HBA mode (also referred to as pass-through mode). For the deployment used in this guide, this applies to all PERC FD33xD and PERC H730 controllers in FC430 and R630 servers respectively.

To verify storage controllers are in HBA mode:

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the **Next Boot** menu, select **BIOS Setup**.
3. Reboot the server.
4. From the **System Setup Main Menu**, select **Device Settings**.
5. From the list of devices, select the PERC controller. This opens the **Modular RAID Controller Configuration Utility Main Menu.**
6. Select Controller Management. Scroll down to **Controller Mode** and verify it is set to **HBA**. If set to **RAID**, select **Advanced Controller Management > Switch to HBA Mode > OK**.

Note: If unable to switch to HBA mode, configured RAID virtual disks may need to be deleted first. See your system documentation for more information.

7. Click **Back** and **Finish** as needed to exit **System Setup**.

7.4 Install ESXi

Dell EMC recommends using the latest Dell EMC customized ESXi .iso image available on support.dell.com. The correct drivers for your PowerEdge hardware are built into this image.

Install ESXi on all servers that will be part of your deployment. For the example in this guide, ESXi is installed to redundant internal SD cards in the PowerEdge servers. This includes six R630 servers (in the management and edge clusters) and four FC430 servers (in the compute cluster).

A simple way to install ESXi on a PowerEdge server remotely is by using the iDRAC to boot the server directly to the ESXi .iso image. This is done as follows:

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, select **Virtual Media** > **Connect Virtual Media**.
3. Select **Virtual Media** > **Map CD/DVD** > browse to the Dell EMC customized ESXi .iso image > **Open** > **Map Device**.
4. Select **Next Boot > Virtual CD/DVD/ISO > OK.**
5. Select **Power** > **Reset System (warm boot)**. Answer **Yes** to reboot the server.
6. The server reboots to the ESXi .iso image and installation starts.
7. Follow the prompts to install ESXi. Select the server's Internal Dual SD Module (IDSDM) when prompted for a location.

8. After installation is complete, click **Virtual Media** > **Disconnect Virtual Media** > **Yes.**
9. Reboot the system when prompted.

7.5 Configure the ESXi management network connection

Be sure the host is physically connected to the management network. For this deployment, the Intel I350-T 1GbE add-in PCIe adapter provides this connection for R630 servers and FC430 servers.

1. Log in to the ESXi console and select **Configure Management Network > Network Adapters**.
2. Select the correct vmnic for the management network connection. Follow the prompts on the screen to make the selection.
3. Go to **Configure Management Network > IPv4 Configuration**. If DHCP is not used, specify a static IP address, mask, and default gateway for the management interface.
4. Optionally, configure DNS settings from the **Configure Management Network** menu if DNS is used on your network.
5. Press **Esc** to exit and answer **Y** to apply the changes.
6. From the ESXi main menu, select **Test Management Network.** Verify pings are successful. If there is an error, be sure you have configured the correct vmnic.
7. Optionally, under **Troubleshooting Options,** enable the ESXi shell and SSH to enable remote access to the CLI.
8. Log out of the ESXi console.

8 Deploy VMware vCenter Server and add hosts

8.1 Deploy VMware vCenter Server

VMware vCenter Server is required for managing clusters and NSX, as well as many other advanced vSphere features. vCenter Server can be installed as a Windows-based application or as a prepackaged SUSE Linux-based VM.

This guide uses the prepackaged VM, called the vCenter Server Appliance (VCSA) and its built in PostgreSQL database. VCSA supports up to 1000 hosts and 10,000 VMs. VCSA is available for download at my.vmware.com.

In this guide, VCSA is installed on a PowerEdge R630 server running ESXi. The server will be part of the management cluster.

Note: This section provides simplified VCSA installation instructions. Detailed instructions and information are provided in the VMware vCenter Server 6.0 Deployment Guide available at the following location: https://www.vmware.com/files/pdf/techpaper/vmware-vcenter-server6-deployment-guide.pdf

1. On a Windows workstation connected to the management network, mount the VCSA .iso image.
2. Install the Client Integration Plugin by running **\vcsa\VMWare-ClientIntegrationPlugin-6.0.0.exe**.
3. Open **\vcsa-setup.html** in a browser and accept the related warning prompts. Click **Install**.
   a. Accept the license agreement and click **Next**.
   b. Provide the ESXi host destination IP address, ESXi host username (root) and password. Click **Next.** Click **Yes** to accept the SSL certificate warning if prompted.
   c. Provide a vCenter **Appliance name** (vctr01 for example), and **password**. Click **Next.**
   d. Keep the default selection: **Install vCenter Server with an Embedded Platform Services Controller.** Click **Next.**
   e. Select **Create a new SSO domain** > **Next**.
   f. Provide an **SSO Password**, **SSO Domain name** (pct.lab for example), and **SSO Site name** (site for example).
   g. Select an **Appliance size** depending on your requirements. For this guide **Medium (up to 400 hosts, 4000 VMs)** is selected.
   h. Select a **datastore**. Optionally, if space is limited, check the **Enable Thin Disk Mode** box. Click **Next.**
   i. Keep the default selection: **Use an embedded database (PostgreSQL).** Click **Next.**
   j. Under **Network Settings**:
      i. Keep the default network, **VMNetwork**.
      ii. Select **IPv4** and the network type (**static or DHCP**). A static address is used in this guide.
      iii. If **static** was selected, provide a **Network address**, **System name** (if not using a fully qualified domain name, retype the Network address), **Subnet mask**, **Network gateway**, and **DNS server**.
      iv. Under **Configure time sync**, select **Synchronize appliance time with ESXi host**.

Note: If you select Use NTP (Network Time Protocol) servers, a warning appears at the bottom of the screen indicating deployment will fail if the ESXi host clock is not in sync with the NTP server. Since the

ESXi hosts are not yet configured for NTP, select Synchronize appliance time with ESXi host. ESXi hosts are configured for NTP in Section 0.

v. Checking **Enable SSH** is optional. Click **Next.** Click **OK** if a fully qualified domain name (FQDN) recommendation box is displayed.

k. Joining the **VMWare Customer Experience Improvement Program** is recommended but optional. Select an option and click **Next.**

l. Review the summary page and click **Finish** if all settings are correct.

vCenter Server is installed as a virtual machine on the ESXi host. When complete, the message shown in Figure 28 is displayed.

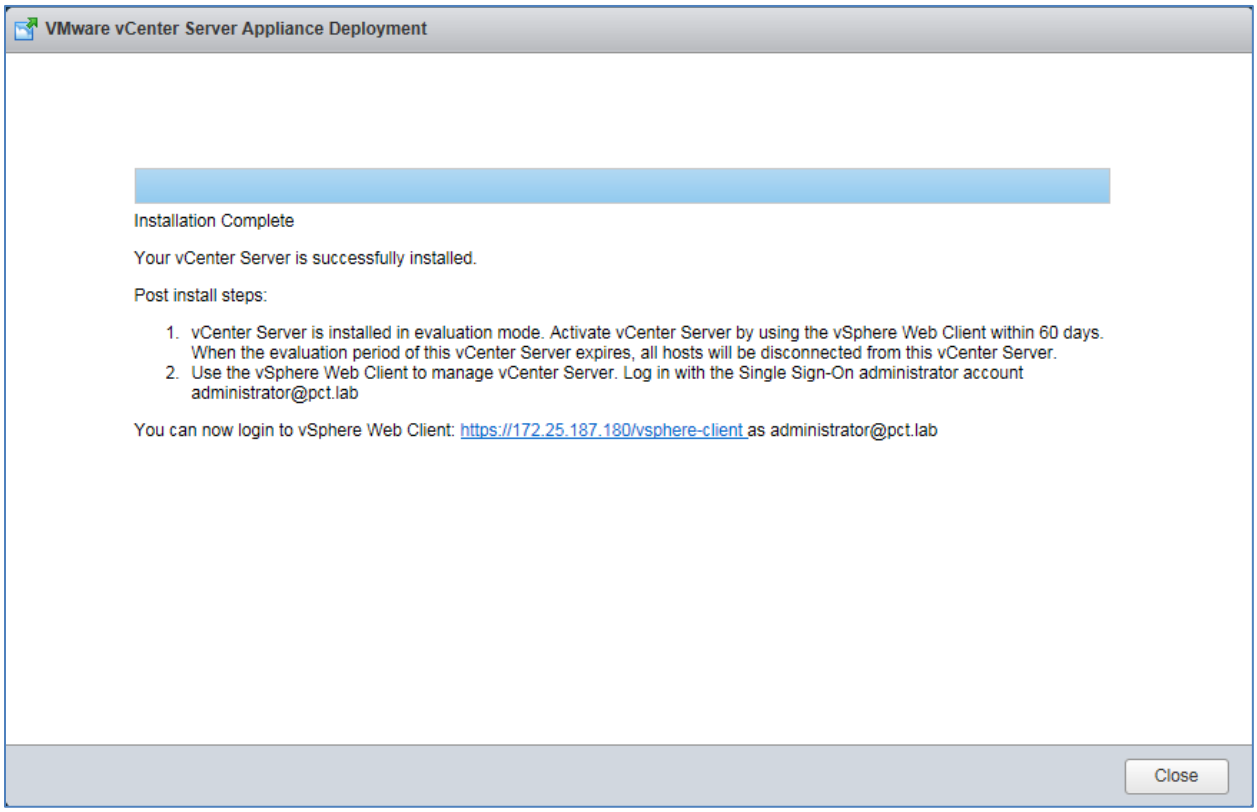


Figure 28 vCenter Server installation complete

8.2 Connect to the vSphere web client

Note: The vSphere Web Client is a service running on vCenter Server.

Connect to the vSphere Web Client in a browser by entering the following in the address bar:

https://*<ip-address-or-hostname-of-vCenter-appliance>*/vsphere-client

Log in with your vCenter credentials. After log in, the web client home page is displayed as shown in Figure 29.

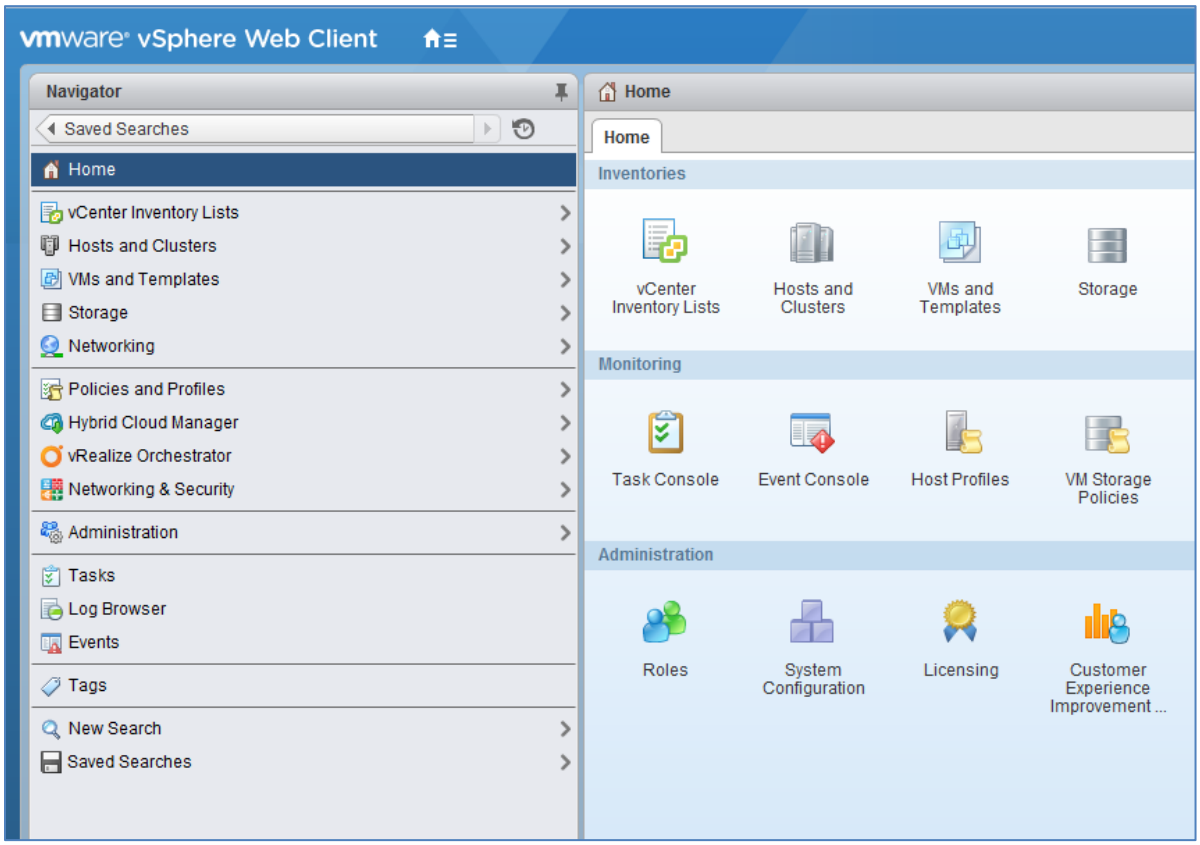


Figure 29 vSphere Web Client home page

The vast majority of management, configuration, and monitoring of your vSphere and NSX environment is done in the web client.

8.3 Install VMware licenses

The VMware licenses required for this deployment are listed in Appendix B.2. All VMware products used in this guide come with evaluation licenses that can be used for up to 60 days.

To install one or more product licenses:

1. From the web client **Home** page, select **Licensing** in the center pane.
2. Click the + icon, and type or paste license keys into the box provided. Click **Next**.
3. Provide **License names** for the keys or use the defaults. Click **Next > Finish**.

8.4 Create a datacenter object and add hosts

A datacenter object needs to be created before hosts can be added. This guide uses a single datacenter object named Datacenter.

To create a datacenter object:

1. On the web client **Home** screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, right click the vCenter Server object and select **New Datacenter**.
3. Provide a name (Datacenter) and click **OK.**

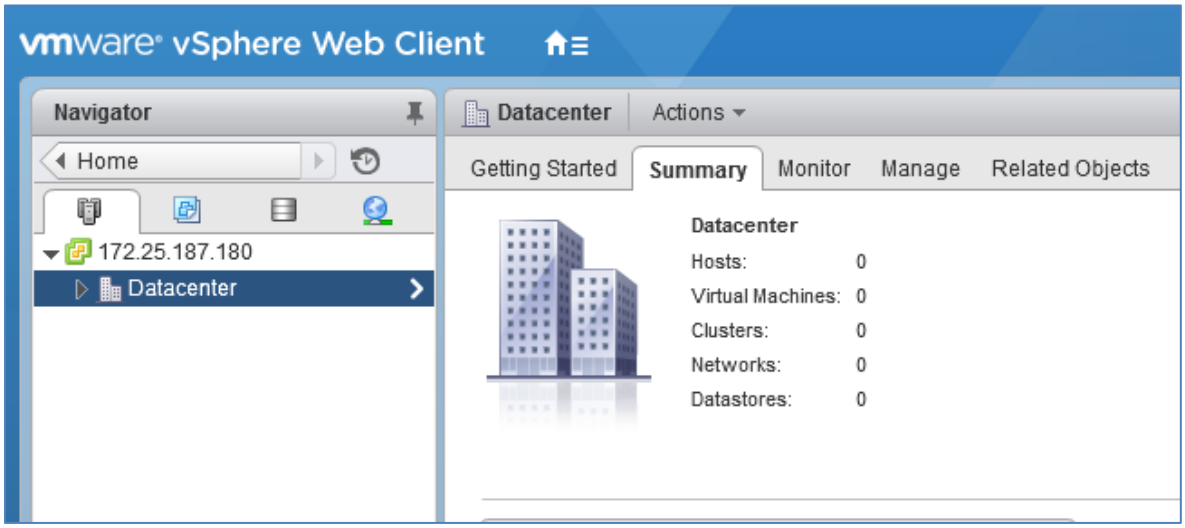


Figure 30 Datacenter created

To add ESXi hosts to the datacenter:

1. On the web client **Home** screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, right click on **Datacenter** and select **Add Host**.
3. Specify the **IP address** of an ESXi host (or the **host name** if DNS is configured on your network). Click **Next.**
4. Enter the credentials for the ESXi host and click **Next.** If a security certificate warning box is displayed, click **Yes** to proceed.
5. On the **Host summary** screen, click **Next.**
6. Assign a license or select the evaluation license. This guide uses a VMware vSphere 6 Enterprise Plus license for ESXi hosts. Click **Next.**
7. Select a **Lockdown mode**. This guide uses the default setting, **Disabled**. Click **Next.**
8. For the **VM location,** select **Datacenter** and click **Next.**
9. On the **Ready to complete** screen, select **Finish**.

Repeat for all servers running ESXi that will be part of the NSX environment. This deployment example uses four FC430 servers and six R630 servers for a total of ten hosts running ESXi. When complete, all ESXi hosts will be added to the datacenter as shown in Figure 31.

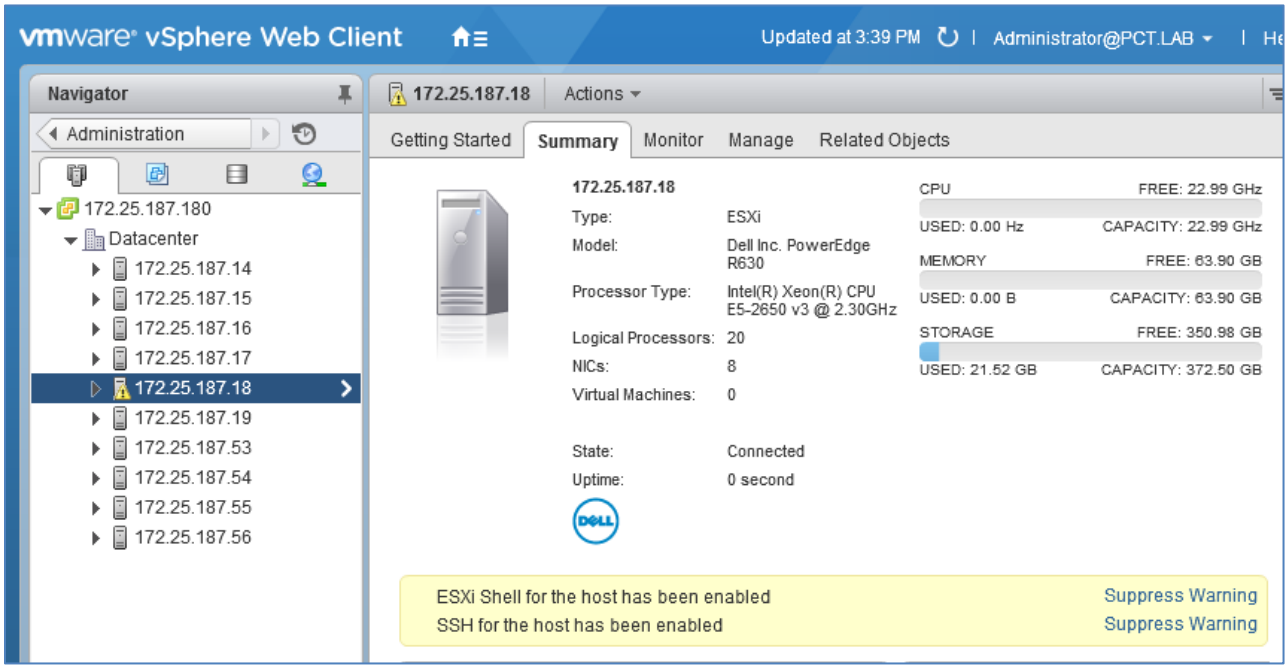


Figure 31 ESXi hosts added to the datacenter

Some hosts may have a warning icon ( ) as shown in Figure 31. By selecting the host and going to the **Summary** tab, the warning message can be viewed. In this case the warning indicates the ESXi Shell and SSH have been enabled (as described in Section 7.5). If the behavior is desired, click **Suppress Warning**.

The following warning messages may also appear:

- *No datastores have been configured*. This message will be resolved when either a local datastore or VSAN datastore is configured. VSAN datastore configuration is covered in section 10 or see your ESXi documentation to create a local datastore.
- *System logs on host are stored on non-persistent storage.* This message may appear when ESXi is installed to the redundant internal SD cards. This can be resolved by moving the system logs to either a local datastore or VSAN datastore. VSAN datastore configuration is covered in section 10 or see your ESXi documentation to create a local datastore. Resolution is documented in VMware Knowledge Base article 2032823.

8.5 Ensure hosts are configured for NTP

It is a best practice to use NTP on the management network to keep time synchronized in an NSX environment. Ensure NTP is configured on ESXi hosts as follows:

1. On the web client **Home** screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, select a host.
3. In the center pane, go to **Manage > Settings > Time Configuration.** If the information shown is correct (see Figure 32), skip to step 7. Otherwise, continue to step 4.
4. If NTP has not been configured properly, click **Edit.**
5. In the **Edit Time Configuration** dialog box:
   a. Select **Use Network Time Protocol** radio button.
   b. Next to **NTP Service Startup Policy**, select **Start and stop with host.**
   c. Next to **NTP servers**, enter the IP address or FQDN of the NTP server.
   d. Click **Start** to start the NTP client followed by **OK** to close the dialog box.
6. The **Time Configuration** page for the host should appear similar to Figure 32.

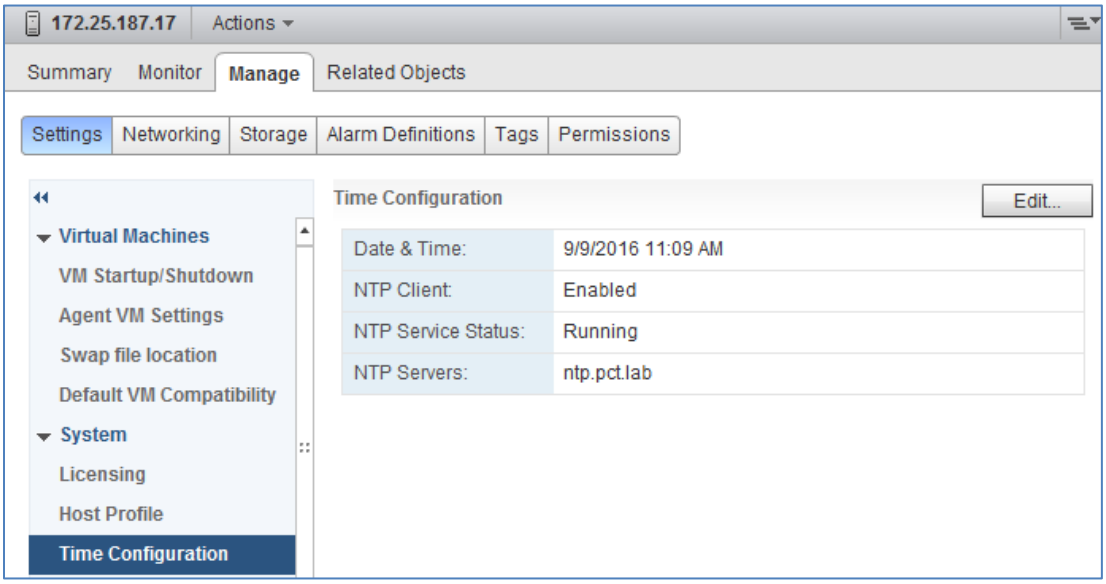


Figure 32 Proper NTP configuration on ESXi host

7. Repeat for remaining ESXi hosts as needed.

8.6 Create clusters and add hosts

When a host is added to a cluster, the host's resources become part of the cluster's resources. The cluster manages the resources of all hosts within it. Clusters enable features such as High Availability (HA), Distributed Resource Scheduler (DRS), and Virtual SAN (VSAN). For this guide, three clusters are created, with one cluster per rack:

- Rack 1 Management
- Rack 2 Compute FC430
- Rack 3 Edge

All ESXi hosts are added to one of the above clusters.

To add clusters to the datacenter:

1. On the web client **Home** screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, right click the datacenter object and select **New Cluster**.
3. Name the cluster. For this example, the first cluster is named **Rack 1 Management**. Leave **DRS, vSphere HA, EVC** and **Virtual SAN** at their default settings (**Off/Disabled**). Click **OK.**

Note: vSphere DRS, HA, and EVC cluster features are outside the scope of this guide. For more information on these features, see the VMware vSphere 6.0 Documentation. Virtual SAN configuration is covered in Section 10.

Repeat for the remaining two clusters:

- Rack 2 Compute FC430
- Rack 3 Edge

In the Navigator pane, drag and drop ESXi hosts into the appropriate clusters. The three ESXi hosts on R630 servers in Rack 1 are placed in the **Rack 1 Management** cluster, the four ESXi hosts on FC430 servers in Rack 2 are placed in the **Rack 2 Compute FC430** cluster, and the three ESXi hosts on R630 servers in Rack 3 are placed in the **Rack 3 Edge** cluster.

When complete, each cluster ( ) should contain its assigned hosts ( ) as shown in Figure 33:

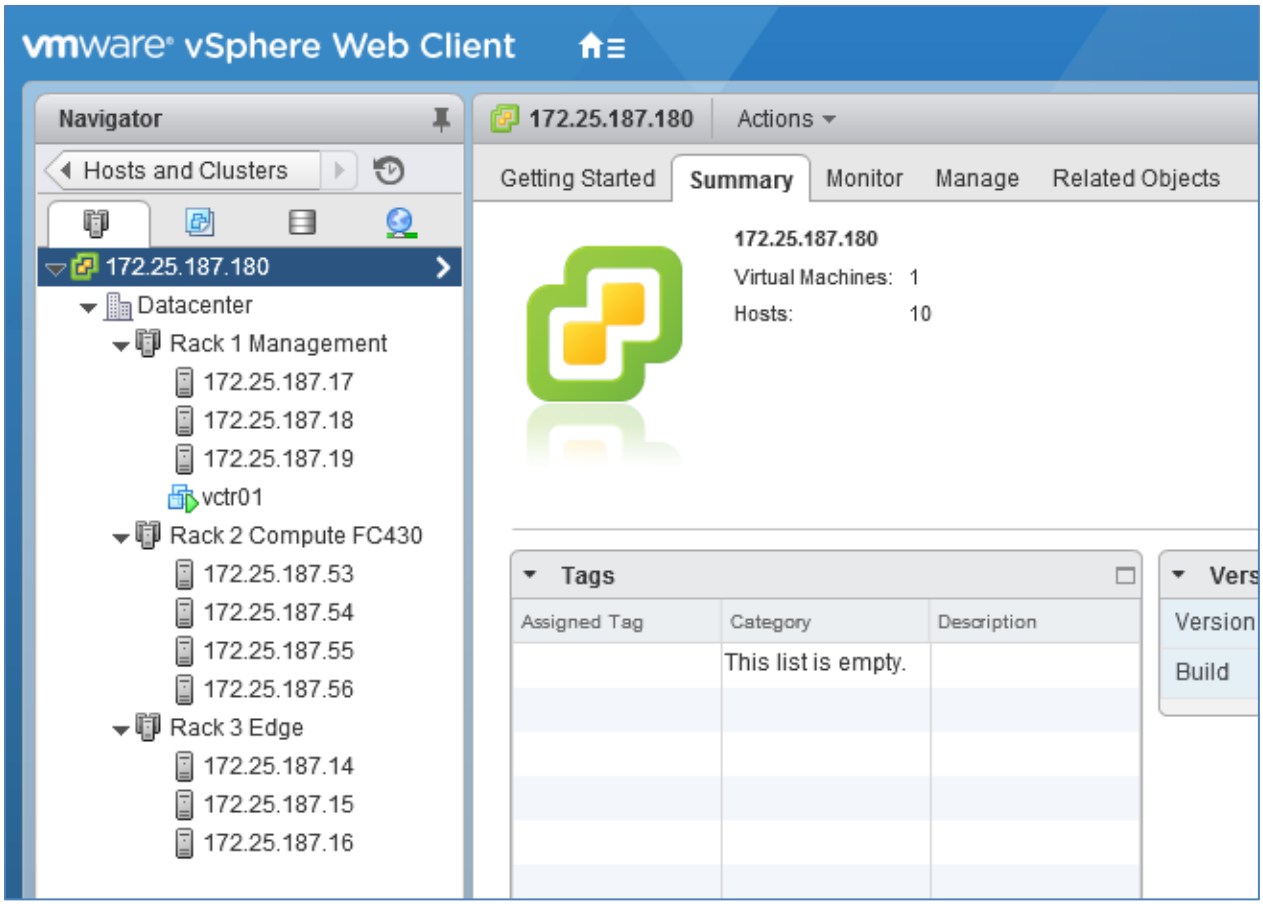


Figure 33 Clusters and hosts after initial configuration

Note: The vCenter Server Appliance, vctr01, is also shown in the Rack 1 Management cluster in Figure 33.

8.7 Information on vSphere standard switches

A vSphere standard switch (also referred to as a VSS or a standard switch) is a virtual switch that handles network traffic at the host level in a vSphere deployment. Standard switches provide network connectivity to hosts and virtual machines.

A standard switch named vSwitch0 is automatically created on each ESXi host during installation to provide connectivity to the management network.

Standard switches may be viewed, and optionally configured, as follows:

1. Go to the web client **Home** page, select **Hosts and Clusters**, and select a host in the **Navigator** pane.
2. In the center pane, select **Manage > Networking > Virtual switches**.

3. Standard switch **vSwitch0** appears in the list. Click on it to view details as shown in Figure 34.

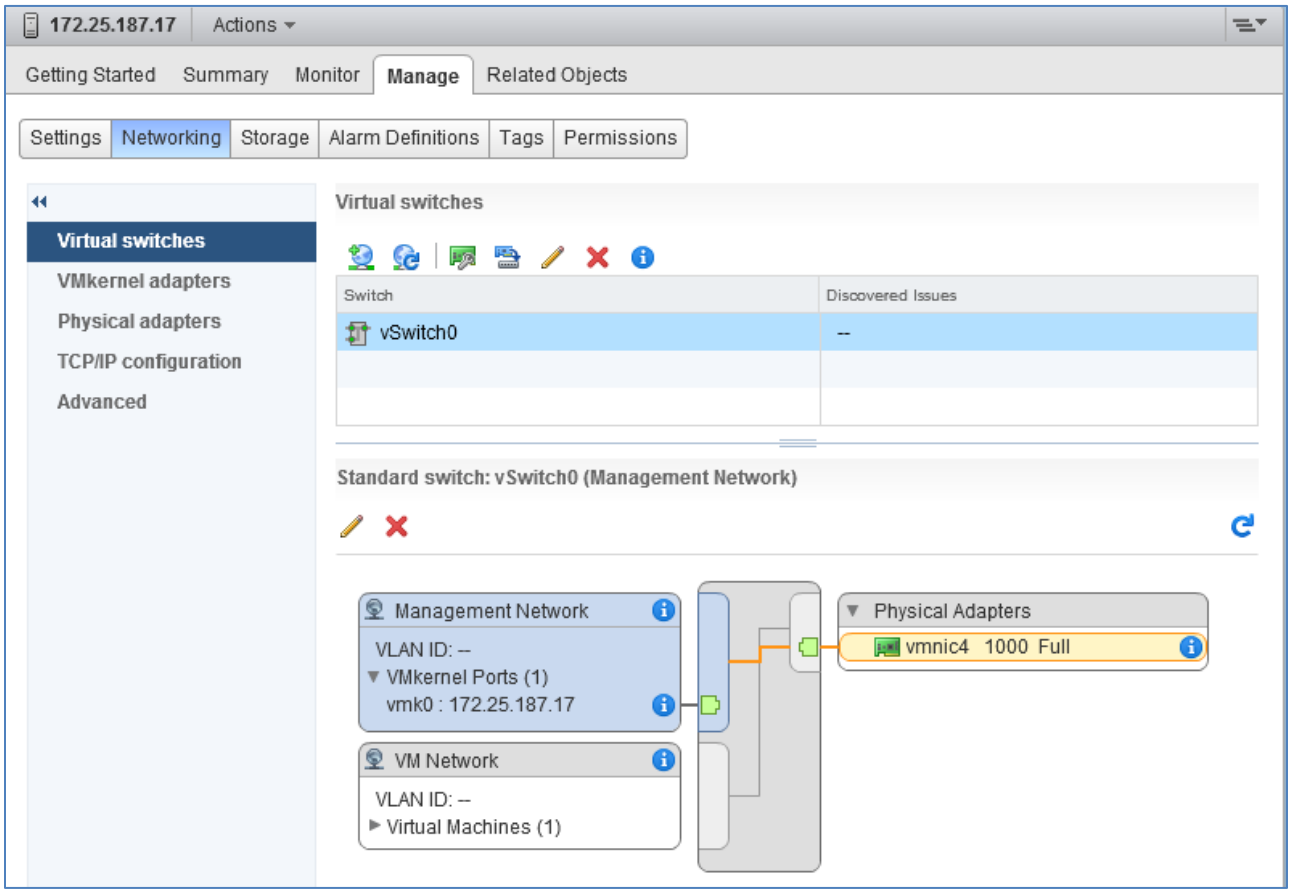


Figure 34 vSphere standard switch

Note: For this guide, only the default configuration is required on the standard switches. Standard switches are only used in this deployment for connectivity to the management network. Distributed switches, covered in the next section, are used for connectivity to the production network.

9 Deploy vSphere distributed switches

A vSphere Distributed Switch (also referred to as a VDS or a distributed switch) is a virtual switch that provides network connectivity to hosts and virtual machines. Unlike vSphere standard switches, distributed switches act as a single switch across multiple hosts in a cluster. This lets virtual machines maintain consistent network configurations as they migrate across multiple hosts.

Distributed switches are configured in the web client and the configuration is populated across all hosts associated with the switch. They are used for connectivity to the Production network in this guide.

Distributed Switches contain two different port groups:

- **Uplink port group** – an uplink port group maps physical NICs on the hosts (vmnics) to uplinks on the VDS. Uplink port groups act as trunks and carry all VLANs by default.

Note: For consistent network configuration, you can connect the same physical NIC port on every host to the same uplink port on the distributed switch. For example, if you are adding two hosts, connect vmnic1 on each host to Uplink1 on the distributed switch.

- **Distributed port group** - Distributed port groups define how connections are made through the VDS to the network. In this guide, one distributed port group is created for each VLAN, and one for each VXLAN Network ID (VNI).

For this guide, one VDS is created for each of the three clusters, and each VDS is shared by all hosts in the cluster. The three distributed switches used in this deployment are named:

- Rack 1 Management VDS
- Rack 2 Compute FC430 VDS
- Rack 3 Edge VDS

9.1 Create a VDS for each cluster

Create the first VDS named **Rack 1 Management VDS**:

1. On the web client **Home** screen, select **Networking.**
2. Right click on **Datacenter**. Select **Distributed switch > New Distributed Switch**.
3. Provide a name for the first VDS, **Rack 1 Management VDS**. Click **Next**.
4. On the Select version page, select **Distributed switch: 6.0.0 > Next**.
5. On the **Edit settings** page:
   a. Leave the **Number of uplinks** set to **4** (this field to be replaced by LAGs later).
   b. Leave **Network I/O Control** set to **Enabled**.
   c. **Uncheck** the **Create a default port group** box.
6. Click **Next** followed by **Finish.**
7. The VDS is created with the uplink port group shown beneath it.

Repeat steps 1-7 above, substituting the switch name in step 3 to create the remaining two distributed switches, **Rack 2 Compute FC430 VDS** and **Rack 3 Edge VDS**.

When complete, the Navigator pane should look similar to Figure 35.

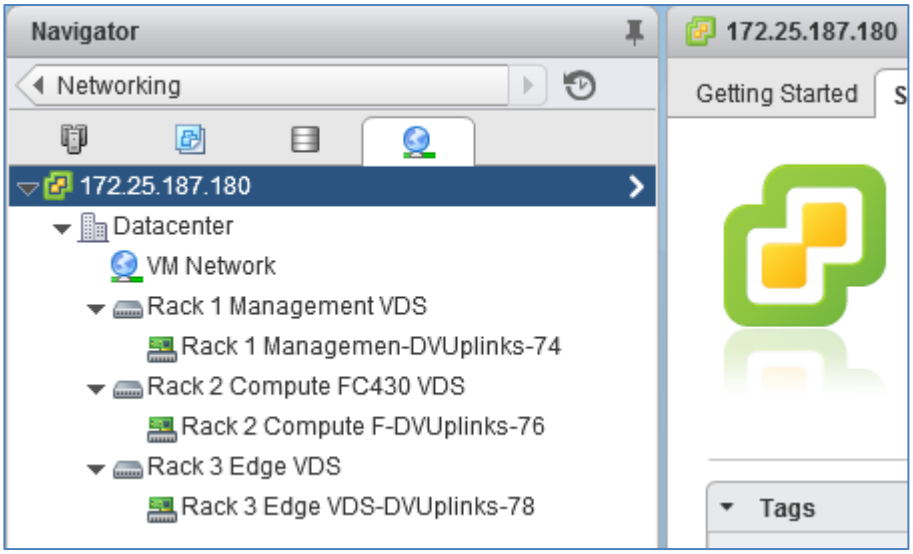


Figure 35 VDS created for each cluster

9.2 Add distributed port groups

In this section, separate distributed port groups for vMotion and VSAN traffic are added to each VDS.

To create the port group for vMotion traffic on the **Rack 1 Management VDS**:

1. On the web client **Home** screen, select **Networking**.
2. Right click on **Rack 1 Management VDS**. Select **Distributed Port Group > New Distributed Port Group.**
3. On the **Select name and location** page, provide a name for the distributed port group, for example, **R1 Management vMotion**. Click **Next**.
4. On the **Configure settings** page, next to **VLAN type**, select **VLAN**. Set the **VLAN ID** to **22** for the vMotion port group. Leave other values at their defaults as shown in Figure 36.
5. Click **Next > Finish**.

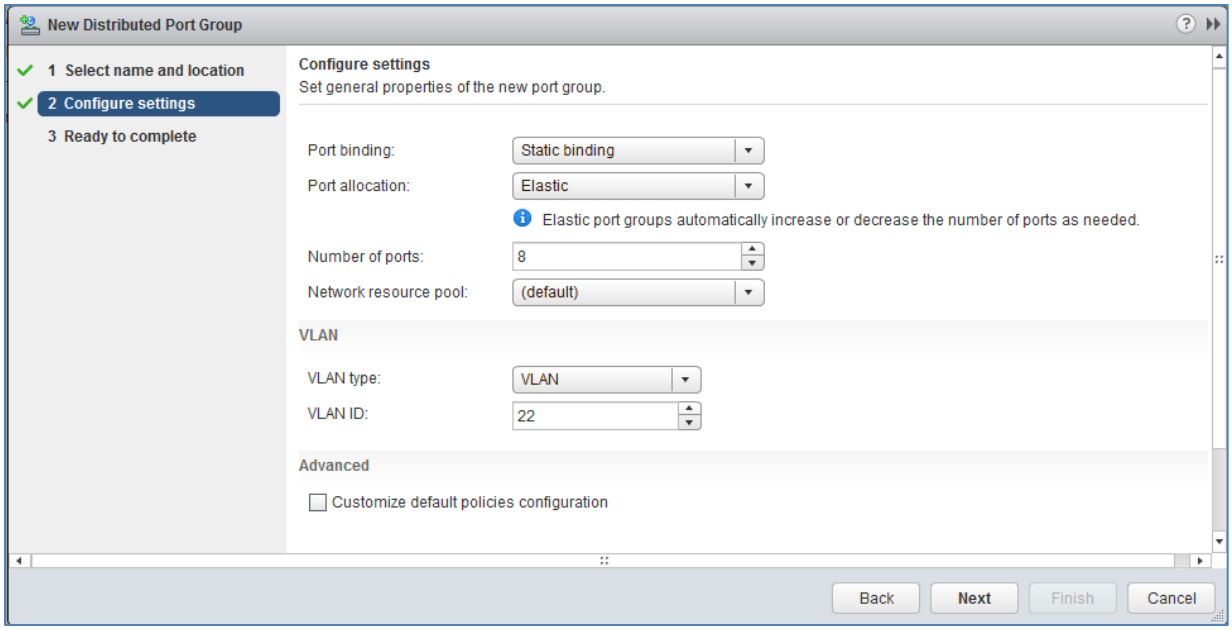


Figure 36 Distributed port group settings page – vMotion port group

Repeat steps 1-5 above to create the distributed port group for VSAN traffic, except replace "vMotion" with "VSAN" in the **port group name** and set the **VLAN ID** to **44** for the VSAN port group.

Repeat the above for the remaining two distributed switches, **Rack 2 Compute FC430 VDS** and **Rack 3 Edge VDS.**

When complete, the Navigator pane will appear similar to Figure 37.

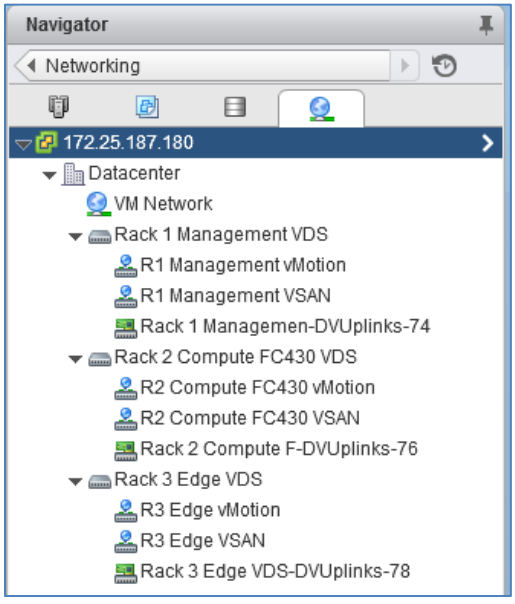


Figure 37 Distributed switches with vMotion and VSAN port groups created

9.3 Create LACP LAGs

Since Link Aggregation Control Protocol (LACP) LAGs are used in the physical network between ESXi hosts and physical switches, LACP LAGs are also configured on each VDS.

To enable LACP on **Rack 1 Management VDS:**

1. On the web client **Home** screen, select **Networking**.
2. In the **Navigator** pane, select Rack 1 Management VDS.
3. In the center pane, select **Manage > Settings > LACP**.
4. Click the + icon. The **New Link Aggregation Group** dialog box opens.
5. Set the number of ports equal to the number of physical uplinks on each ESXi host. In this deployment, R630 and FC430 hosts use two ports in a LAG to connect to the upstream switches so this number is set to **2**.
6. Set the **Mode** to **Active**. The remaining fields can be set to their default values as shown in Figure 38.

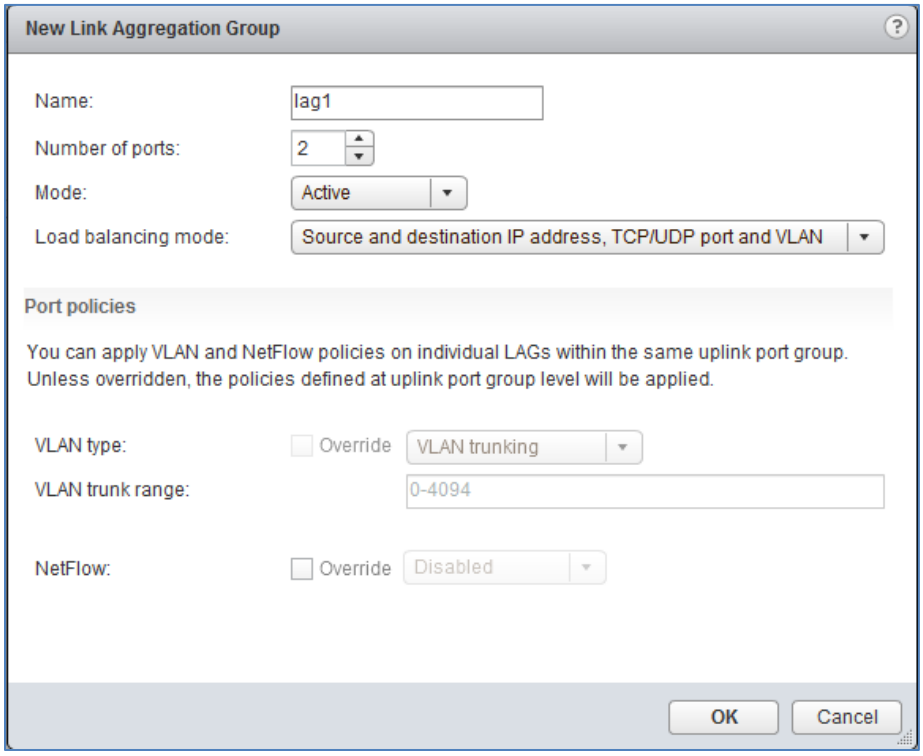


Figure 38 LAG configuration

7. Click **OK** to close the dialog box.

This creates **lag1** on the VDS. The refresh icon ( ) at the top of the screen may need to be clicked for the lag to appear in the table as shown in Figure 39.

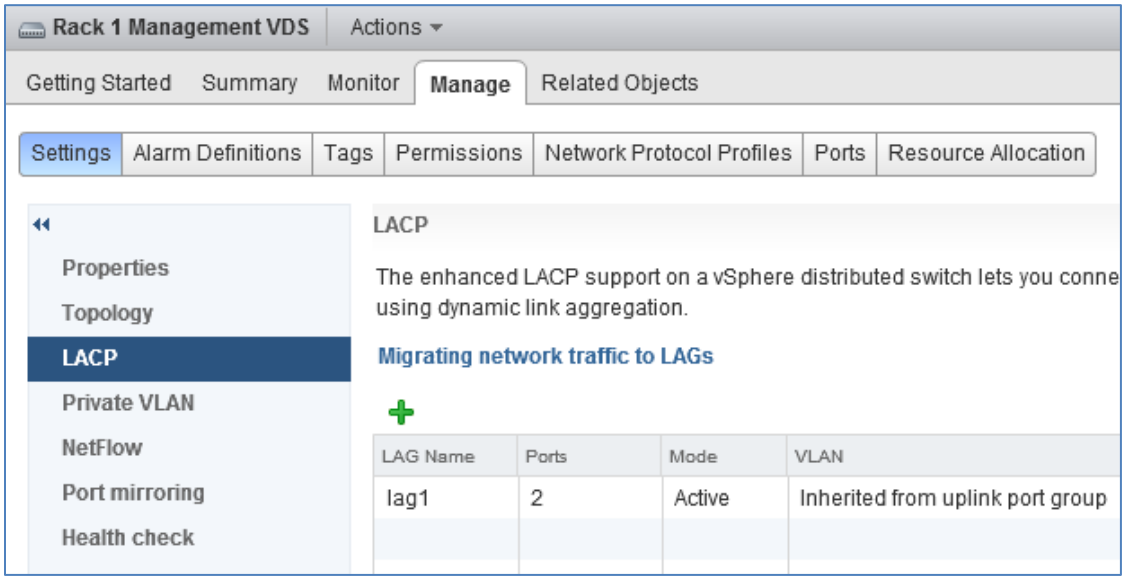


Figure 39 Lag1 created on Rack 1 Management VDS

Repeat steps 1-7 above for the remaining two distributed switches, **Rack 2 Compute FC430 VDS** and **Rack 3 Edge VDS.**

9.4 Associate hosts and assign uplinks to LAGs

Hosts and their vmnics must be associated with each vSphere distributed switch.

Note: Before starting this section, be sure you know the vmnic-to-physical adapter mapping for each host. This can be determined by going to **Home > Hosts and Clusters** and selecting the host in the **Navigator** pane. In the center pane select **Manage > Networking > Physical adapters**. In this example, vmnics used are numbered vmnic1 and vmnic3. Vmnic numbering will vary depending on adapters installed in the host.

To add hosts to Rack 1 Management VDS:

1. On the web client **Home** screen, select **Networking**.
2. Right click on Rack 1 Management VDS and select **Add and Manage Hosts**.
3. In the Add and Manage Hosts dialog box:
   a. On the **Select task** page, make sure **Add hosts** is selected. Click **Next**.
   b. On the **Select hosts** page, Click the **New hosts** icon. Select the check box next to each host in the **Rack 1 Management** cluster. Click **OK > Next**.
   c. On the **Select network adapters tasks** page, be sure the **Manage physical adapters** box is checked. Be sure all other boxes are unchecked. Click **Next**.
   d. On the **Manage physical network adapters** page, each host is listed with its vmnics beneath it.
      i. Select the first vmnic (vmnic1 in this example) on the first host and click Assign uplink .

ii. Select **lag1-0 > OK**.

iii. Select the second vmnic (vmnic3 in this example) on the first host and click Assign uplink.

iv. Select **lag1-1 > OK**.

e. Repeat steps i – iv for the remaining hosts. Click **Next** when done.

f. On the **Analyze impact** page, **Overall impact status** should indicate No impact.

g. Click **Next > Finish**.

When complete, the **Manage > Settings > Topology** page for Rack 1 Management VDS should look similar to Figure 40.

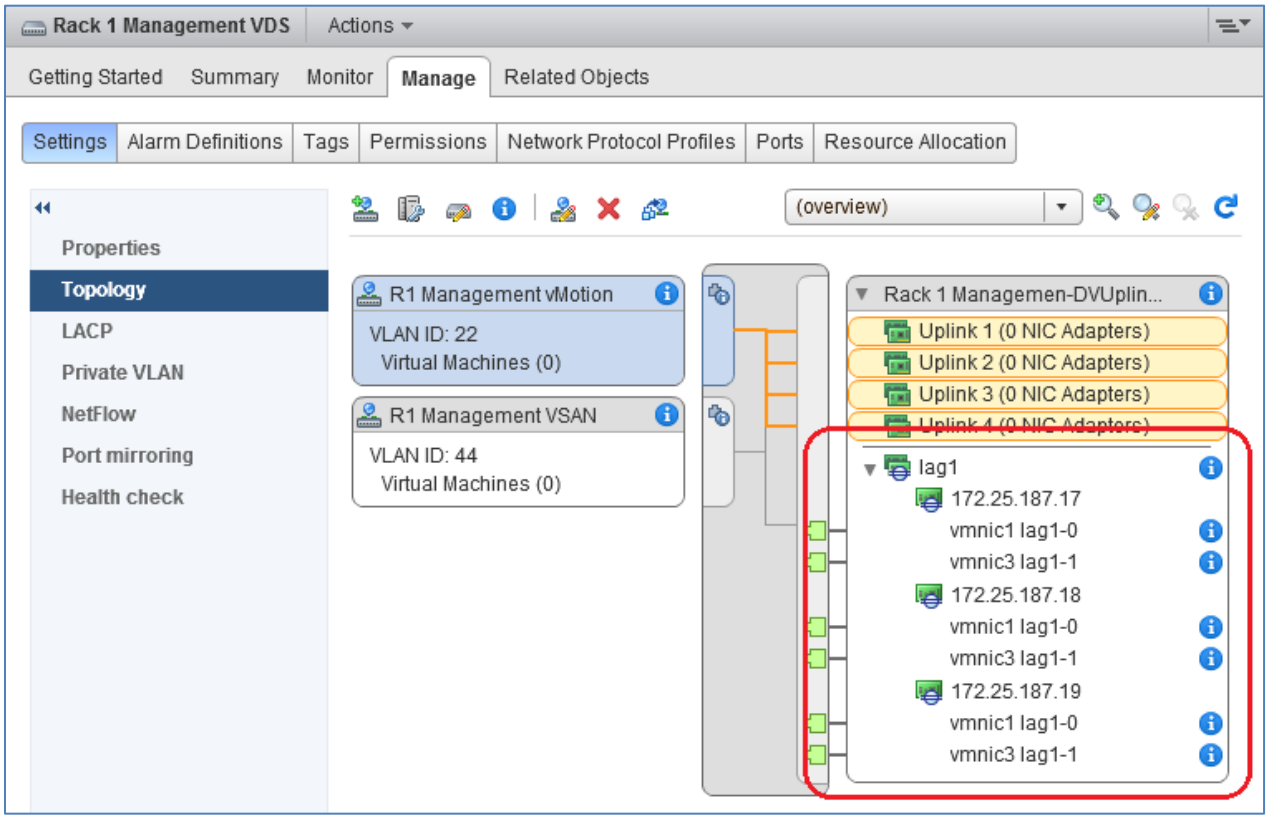


Figure 40 LAGs configured on Rack 1 Management VDS

Repeat steps 1-4 above for the remaining two distributed switches, **Rack 2 Compute FC430 VDS** and **Rack 3 Edge VDS.**

This configuration brings up the LAGs on the upstream switches. This can be confirmed by running the `show vlt detail` command on the upstream switches as shown in the examples from Leaf-1 (Management Cluster) and FN410S-A1 (Compute Cluster) below. The Local and Peer Status columns now indicate all LAGs are UP.

```
Leaf-1#show vlt detail

Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
2             2            UP            UP            1, 22, 44, 55
4             4            UP            UP            1, 22, 44, 55
6             6            UP            UP            1, 22, 44, 55

FN410S-A1#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
1             1            UP            UP            1, 22, 44, 55
2             2            UP            UP            1, 22, 44, 55
3             3            UP            UP            1, 22, 44, 55
4             4            UP            UP            1, 22, 44, 55
33            33           UP            UP            1, 22, 44, 55
```

9.5 Configure teaming and failover on LAGs

1. On the web client **Home** screen, select **Networking**.
2. Right click on Rack 1 Management VDS. Select **Distributed Port Group > Manage Distributed Port Groups**.
3. Select only the **Teaming and failover** checkbox. Click **Next**.
4. Click **Select distributed port groups**. Check the top box to select all port groups (vMotion and VSAN). Click **OK > Next.**
5. On the **Teaming and failover** page, click **lag1** and move it up to the **Active uplinks** section by clicking the up arrow. Move **Uplinks 1-4** down to the **Unused uplinks** section. Leave other settings at their defaults. The **Teaming and failover** page should look similar to Figure 41 when complete.

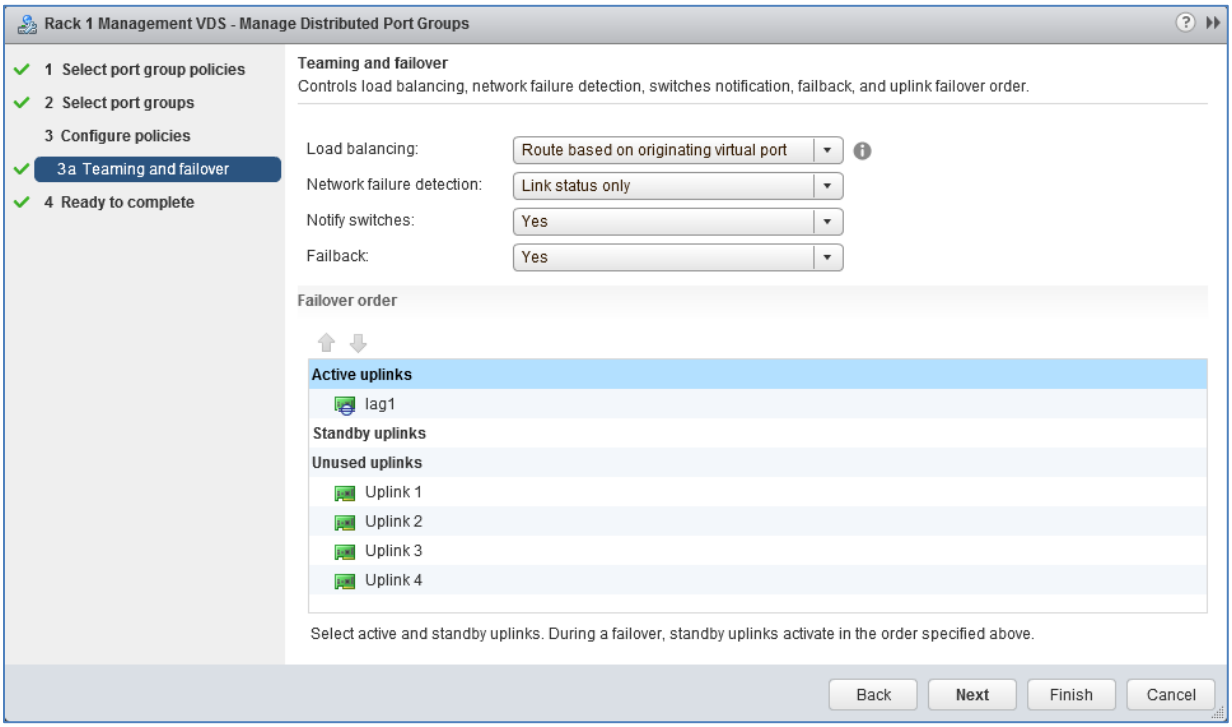


Figure 41 Teaming and failover settings

6. Click **Next** followed by **Finish** to apply the settings.

Repeat steps 1-6 above for the remaining two distributed switches, Rack 2 Compute FC430 VDS and Rack 3 Edge VDS.

9.6 Add VMkernel adapters for vMotion and VSAN

In this section, vMotion and VSAN VMkernel adapters (also referred to as VMkernel ports) will be added to each ESXi host to allow for vMotion and VSAN traffic.

IP addresses can be statically assigned to VMkernel adapters upon creation, or DHCP may be used. Static IP addresses are used in this guide.

This deployment uses the following addressing scheme for the vMotion and VSAN networks, where "x" represents the rack number:

Table 4 VLAN and network examples

VLAN ID	Network	Used For
22	10.22.x.0/24	vMotion
44	10.44.x.0/24	VSAN

To add VMkernel adapters to all hosts connected to the Rack 1 Management VDS:

1. On the web client **Home** screen, select **Networking**.
2. Right click on Rack 1 Management VDS, and select **Add and Manage Hosts**.
3. In the Add and Manage Hosts dialog box:
   a. On the **Select task** page, make sure **Manage host networking** is selected. Click **Next**.
   a. On the **Select hosts** page, click + **Attached hosts**. Select all hosts. Click **OK > Next.**
   b. On the **Select network adapter tasks** page, make sure the **Manage VMkernel adapters** box is checked and all other boxes are unchecked. Click **Next**.
   c. The **Manage VMkernel network adapters** page opens.

      vMotion adapter

      i. To add the vMotion adapter, select the first host and click + **New Adapter**.
      ii. On the **Select target device** page, click the radio button next to **Select an existing network** and click **Browse**.
      iii. Select the port group created for vMotion **> OK**. Click **Next**.
      iv. On the **Port properties** page, leave **IPv4** selected and check only the **vMotion traffic** box. Click **Next**.
      v. On the **IPv4 settings page**, if DHCP is not used, select **Use static IPv4 settings**. Set the IP address, for example 10.22.1.17, and subnet mask for the host on the vMotion network. Click **Next > Finish.**

      VSAN adapter

      vi. To add the VSAN adapter, select the first host and click + **New Adapter.**
      vii. On the **Select target device** page, click the radio button next to **Select an existing network** and click **Browse**.
      viii. Select the port group created for VSAN **> OK**. Click **Next**.

ix. On the **Port properties** page, leave IPv4 selected and check only the **Virtual SAN traffic** box. Click **Next**.

x. On the **IPv4 settings page**, if DHCP is not used, select **Use static IPv4 settings**. Set the IP address, for example 10.44.1.17, and subnet mask for the host on the VSAN network. Click **Next > Finish**.

d. Repeat steps i-x for the remaining hosts, then click **Next.**

e. On the **Analyze impact** page, **Overall impact status** should indicate No impact.

f. Click **Next > Finish**.

Repeat the steps above for the remaining two distributed switches, Rack 2 Compute FC430 VDS and Rack 3 Edge VDS.

When complete, the VMkernel adapters page for each ESXi host in the vSphere datacenter should look similar to Figure 42. This page is visible by going to **Hosts and Clusters**, selecting a host in the **Navigator** pane, then selecting **Manage > Networking > VMkernel adapters** in the center pane.

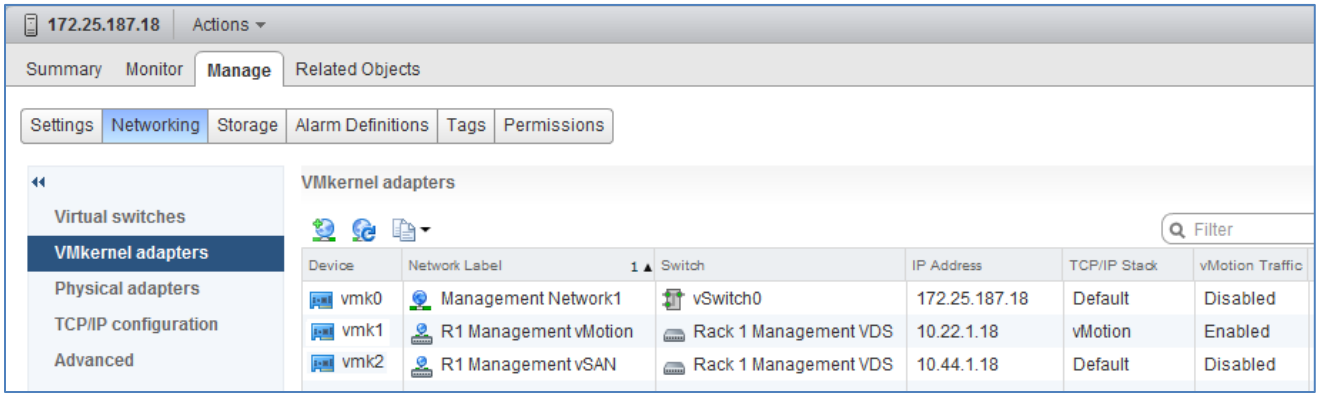


Figure 42 Host VMkernel adapters page

Adapter vmk0 was installed by default for host management. Adapters vmk1 and vmk2 were created in this section.

To verify the configuration, ensure the vMotion adapter, **vmk1** in this example, is shown as **Enabled** in the **vMotion Traffic** column, and the VSAN adapter, **vmk2** in this example, is shown as **Enabled** in the **Virtual SAN Traffic** column. Verify the VMkernel adapter IP addresses are correct.

Verify the information is correct on other hosts as needed.

9.7 Verify VDS configuration

To verify the distributed switches have been configured correctly, the **Topology** page for each VDS provides a summary.

To view the **Topology** page for the **Rack 1 Management VDS**:

1. On the web client **Home** screen, select **Networking**.
2. In the Navigator pane, select **Rack 1 Management VDS**.
3. In the center pane, select **Manage > Settings > Topology** and click the ▶ icon next to **VMkernel Ports** (2 places) to expand. The screen should look similar to Figure 43.

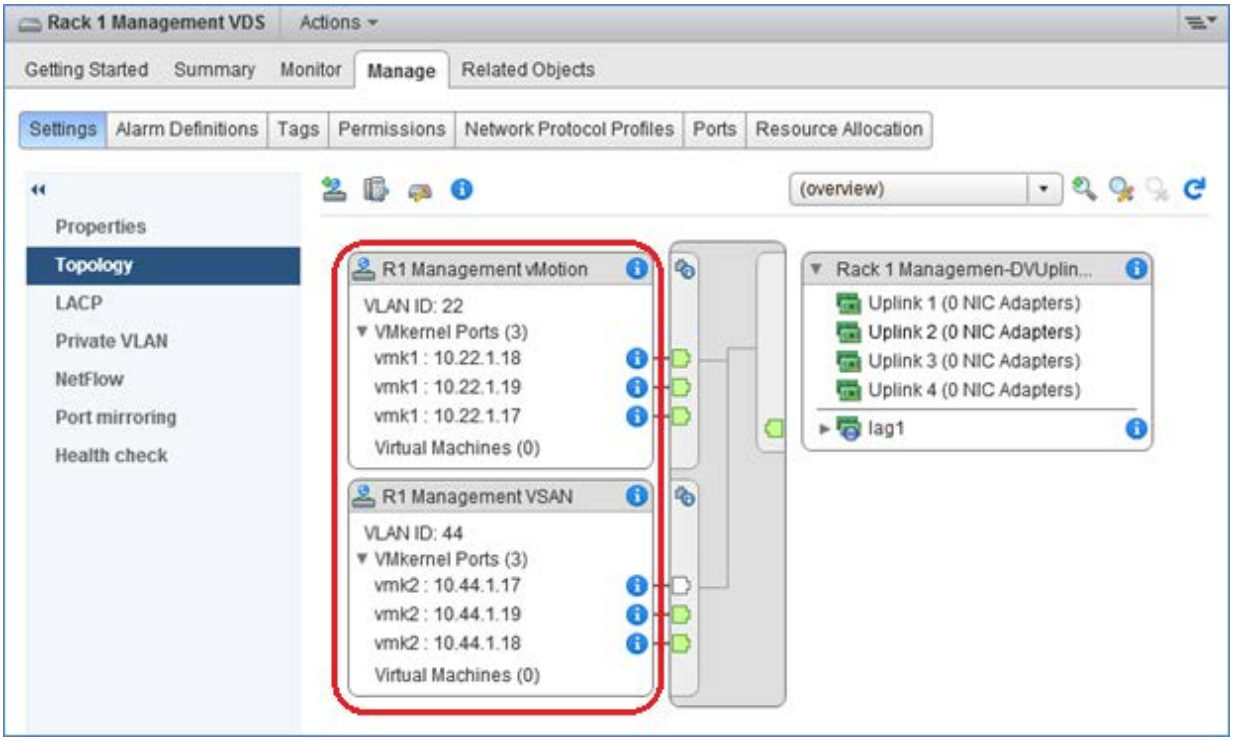


Figure 43 Rack 1 Management VDS VMkernel ports, VLANs, and IP addresses

Notice the two distributed port groups, **R1 Management vMotion** and **R1 Management VSAN** are shown in Figure 43 with their configured VLAN IDs and VMkernel ports. Since VMkernel ports were configured for all three ESXi hosts in the Management cluster, there are three VMkernel ports in each distributed port group.

Repeat steps 1-3 above for the remaining two distributed switches, **Rack 2 Compute FC430 VDS** and **Rack 3 Edge VDS,** to verify they are properly configured.

9.8 Enable LLDP

Enabling Link Layer Discovery Protocol (LLDP) on vSphere distributed switches is optional but can be helpful for link identification and troubleshooting.

Note: LLDP works as described in this section with QLogic 57810 or QLogic 57840 adapters specified in Appendix A. LLDP functionality may vary with other adapters. LLDP must also be configured on the physical switches per the switch configuration instructions provided earlier in this guide.

9.8.1 Enable LLDP on each VDS and view information sent

Enabling LLDP on vSphere distributed switches enables them to send information such as vmnic numbers and MAC addresses to the physical switch connected to the ESXi host.

To enable LLDP on each VDS:

1. On the web client **Home** screen, select **Networking**.
2. Right click on a VDS, and select **Settings > Edit Settings.**
3. In the left pane of the **Edit Settings** page, click **Advanced**.
4. Under **Discovery protocol**, set **Type** to **Link Layer Discovery Protocol** and **Operation** to **Both**.
5. Click **OK.**

Repeat for remaining distributed switches.

To view LLDP information sent from the ESXi host adapters, run the following command from the CLI of a directly connected switch:

```
Leaf-1#show lldp neighbors
 Loc PortID     Rem Host Name     Rem Port Id                  Rem Chassis Id
 ------------------------------------------------------------------------------
 Te 1/2         -                 00:0a:f7:38:88:12          00:0a:f7:38:88:12
 Te 1/2         localhost         00:50:56:18:88:12          vmnic1
 Te 1/4         -                 00:0a:f7:38:96:62          00:0a:f7:38:96:62
 Te 1/4         localhost         00:50:56:18:96:62          vmnic1
 Te 1/6         -                 00:0a:f7:38:94:32          00:0a:f7:38:94:32
 Te 1/6         localhost         00:50:56:18:94:32          vmnic1
 Fo 1/49        Spine-1           fortyGigE 1/1/1            4c:76:25:e7:41:40
 Fo 1/50        Spine-2           fortyGigE 1/1/1            4c:76:25:e7:3b:40
 Fo 1/53        Leaf-2            fortyGigE 1/53             f4:8e:38:20:54:29
 Fo 1/54        Leaf-2            fortyGigE 1/54             f4:8e:38:20:54:29
```

The output above shows Leaf 1 is connected to vmnic1 of each host via interfaces Te 1/2, Te 1/4, and Te 1/6.

9.8.2 View LLDP information received from physical switch

LLDP configuration is part of the physical switch configurations covered in Section 6. The switches are configured to send information (host name, port number, etc.) via LLDP to the ESXi host network adapters.

To view LLDP information sent from the physical switch:

1. On the web client **Home** screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, select a host.
3. In the center pane, select **Manage > Networking > Physical adapters**.
4. Select a connected physical adapter, **vmnic1** for example.
5. Below the adapter list, select the **LLDP** tab. Information similar to that shown in Figure 44 is provided by the switch.

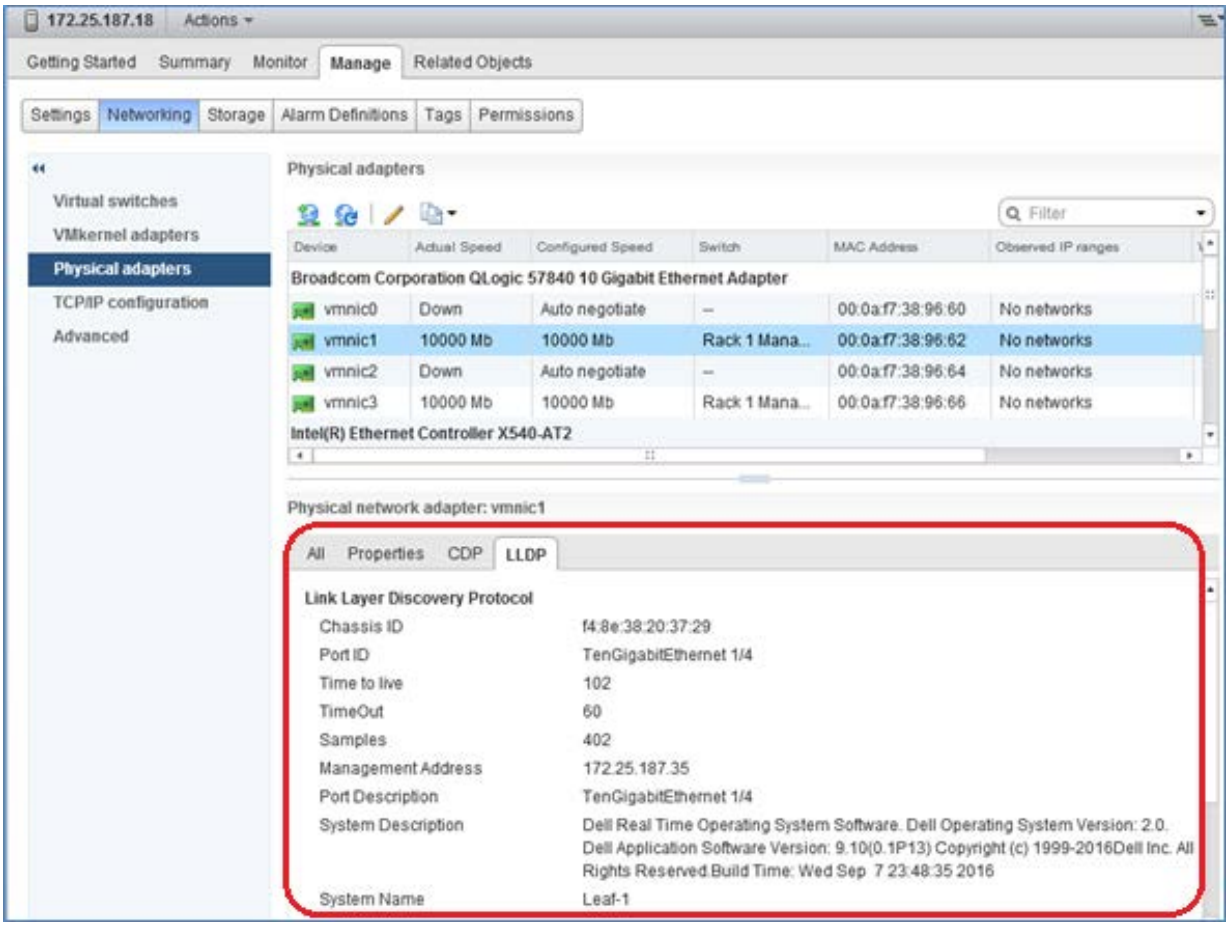


Figure 44 Information sent from physical switch to vmnic via LLDP

10 Configure a VSAN datastore in each cluster

10.1 VSAN Overview

VMware Virtual SAN virtualizes the local physical storage resources of ESXi hosts in a single cluster and turns them into pools of storage that can be divided and assigned to virtual machines and applications. VSAN is implemented directly in the ESXi hypervisor.

VSAN eliminates the need for external shared storage and simplifies storage configuration and virtual machine provisioning activities. VMware features such as HA, vMotion and DRS require shared storage.

Hosts must meet the following criteria to participate in a VSAN:

- A Virtual SAN cluster must contain a minimum of 3 and a maximum of 64 hosts that contribute capacity to the cluster.
- A host that resides in a Virtual SAN cluster must not participate in other clusters.
- If a host contributes its local capacity devices to the Virtual SAN datastore, it must provide at least one device for flash cache and at least one device for capacity, also called a data disk.
- All storage devices, drivers, and firmware versions in the Virtual SAN configuration must be certified and listed in the Virtual SAN section of the VMware Compatibility Guide

10.2 Configure VSAN

Before proceeding, ensure each host in the cluster has a properly configured VMkernel adapter enabled for VSAN traffic (covered in Section 9.6).

To configure VSAN on a cluster:

1. Go to **Home > Hosts and Clusters**.
2. In the **Navigator** pane, select a cluster such as Rack 2 Compute FC430.
3. In the center pane, select **Manage > Settings.** Under **Virtual SAN**, select **General**.
4. Click the **Configure** button to launch the **Configure Virtual VSAN** wizard.
   a. Set **Add disks to storage** to **Manual**. Leave the remaining options at their defaults as shown in Figure 45 and click **Next.**

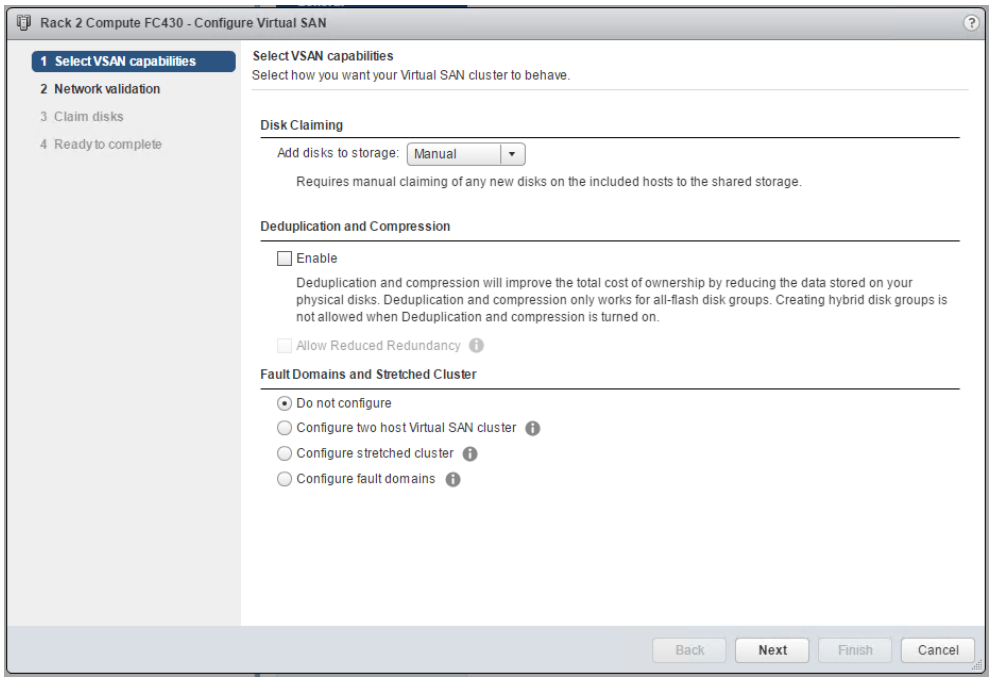


Figure 45 Configure Virtual VSAN – Select VSAN capabilities page

b. On the **Network validation** page, the VMkernel ports configured for VSAN traffic are shown with their IP addresses and a green check in the VSAN enabled column. Click **Next.**

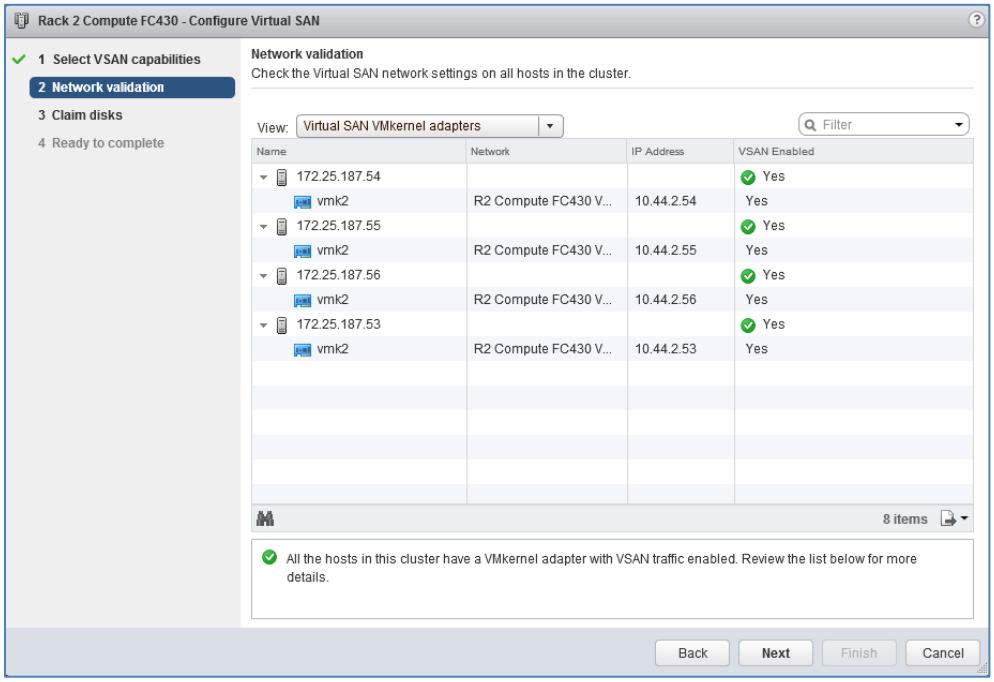


Figure 46 Configure Virtual SAN - VMkernel adapter confirmation

c. On the **Claim disks** page, set **Group by** to **Host** and expand the hosts to view available disks. One disk group will be configured for each host. A disk group should have 1 disk claimed for the **Cache Tier** and the remaining disks claimed for the **Capacity Tier**.

Note: A maximum of eight disks per disk group are allowed. Up to five disk groups can be configured per host.

In Figure 47, four disk groups are created, one for each host.

For the first disk on the first host, **Cache Tier** is selected. **Capacity Tier** is selected for the remaining seven disks. When all groups have been configured, make sure there is a green checkmark in the **Configuration validation** box as shown in Figure 47. Click **Next > Finish** to apply the configuration.

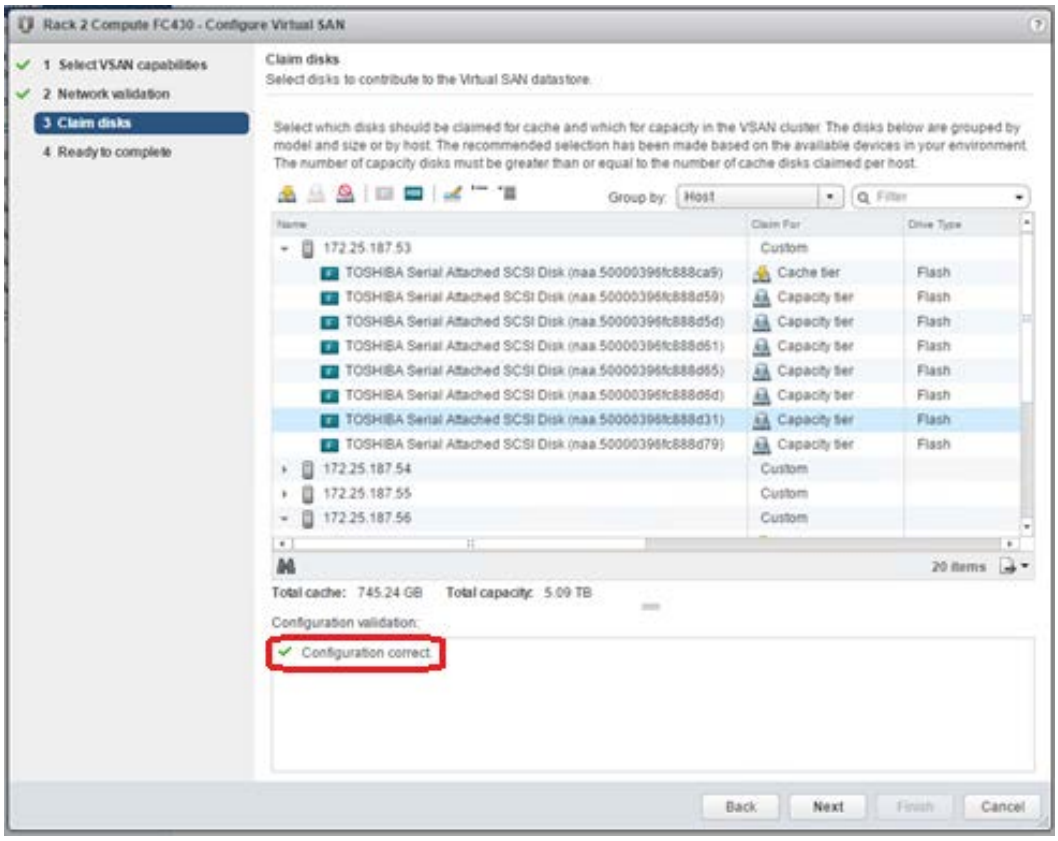


Figure 47 Configure Virtual SAN disk group management

Once this has been completed, a VSAN datastore is automatically created and attached to all participating hosts in the cluster.

Repeat steps 1-3 above for the remaining clusters to create a VSAN datastore on each cluster.

Note: For more information see the Designing and Sizing a Virtual SAN Cluster section of the vSphere 6.0 online documentation.

10.3 Verify VSAN configuration

VSANs are viewed by going to the web client **Home** page and selecting **Storage**. In the **Navigator** pane, in addition to local storage on each host, a vsanDatastore will be listed for each VSAN created.

The default names are vsanDatastore, vsanDatastore (1), etc. The hosts associated with each VSAN can be viewed by selecting a **vsanDatastore** in the Navigator pane. In the center pane, select **Related Objects > Hosts.**

It is a good idea to give datastores more user-friendly names to make them easier to work with. This is done by right clicking on the datastore name in the **Navigator** pane and selecting **Rename**.

For this guide the vsanDatastores are renamed to:

- Rack 1 Management VSAN
- Rack 2 Compute FC430 VSAN
- Rack 3 Edge VSAN

Note: Local datastores, if used, should also be renamed for usability. In this example, local datastores are renamed LDS (local datastore) plus the last octet # of the host address. Local datastore configuration and use is not covered in this guide.

In Figure 48, the **Rack 2 Compute FC430 VSAN** has been selected in the left pane. In the right pane, the **Related Objects > Hosts** tab shows the hosts and cluster associated with this VSAN.

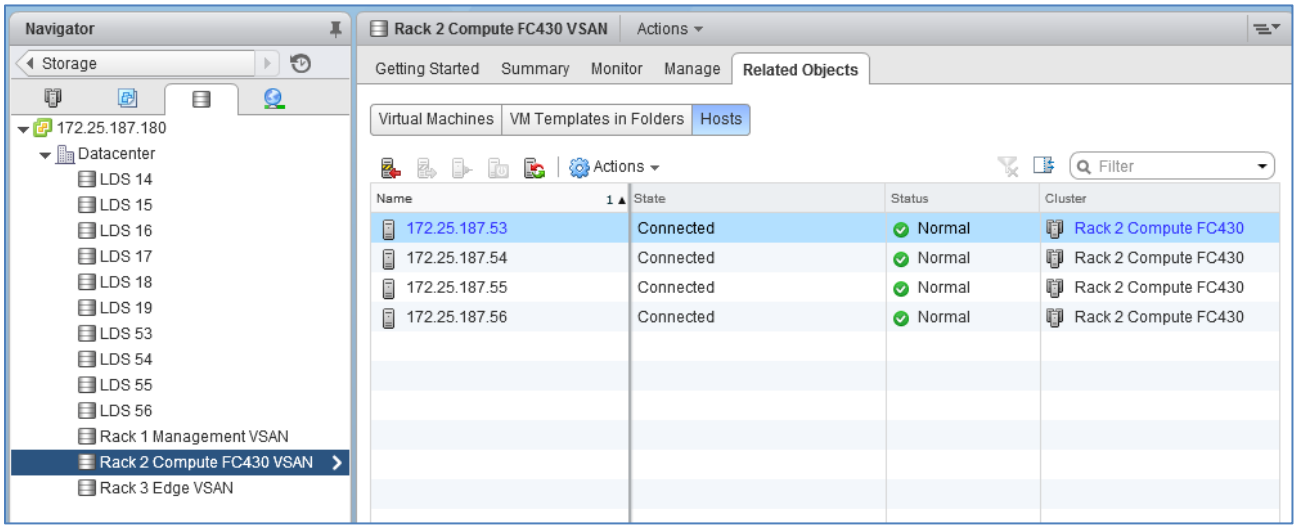


Figure 48 VSAN Hosts page

Additional VSAN monitoring and management is done on the **Summary, Monitor**, and **Manage** tabs.

10.4 Check VSAN health and resolve issues

A VSAN health check is run as follows:

1. On the web client **Home** screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, select a cluster such as **Rack 2 Compute FC430**.
3. In the center pane, select **Monitor > Virtual SAN > Health > Retest**.
4. Verify all health tests pass as shown in Figure 49.

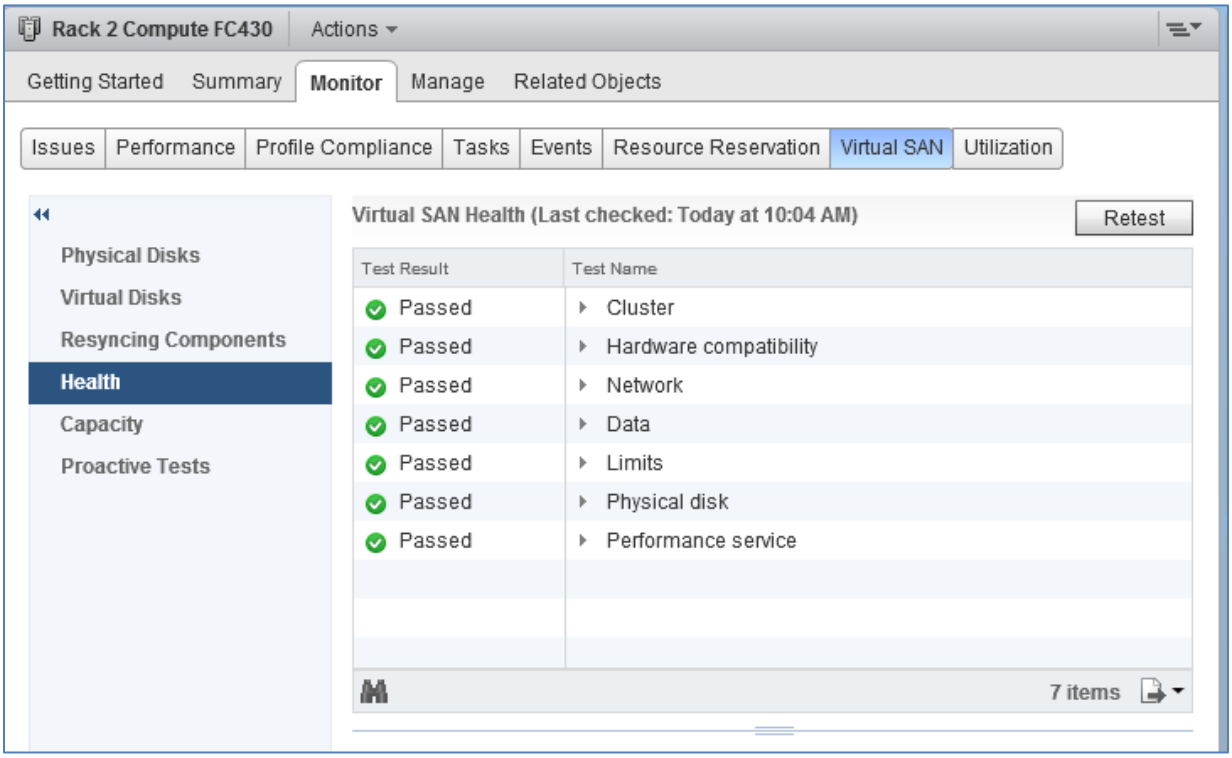


Figure 49 Virtual SAN health monitoring test results

If all tests pass, repeat for remaining clusters that have a VSAN configured. If there are warnings or failures, see the following sections to resolve common issues.

After all tests have passed on all VSANs, proceed to Section 10.5.

10.4.1 Failure: Virtual SAN HCL DB up-to-date

If this error is seen, select the failed test, and do one of the following to update the VMware VSAN Hardware Compatibility List (HCL):

- Online option: Click the **Get latest version online button** that appears when the failed HCL test is selected. When the file has been installed, click **Retest.** The test should pass.
- Local File option: If unable to connect online, you can upload from a local file. The HCL DB is a .json file available at http://partnerweb.vmware.com/service/vsan/all.json. Download the file to a

workstation. In the web client, click the **Upload from file** button and follow the prompts. Click **Retest.** This test should pass.

10.4.2 Warning: Controller Driver / Controller Release Support

This error may be seen with the PERC H730 (in R630 servers) or PERC FD33xD (in FC430 servers). If so, the drivers need to be updated per the following VMware Knowledge Base article: Best practices for VSAN implementations using Dell PERC H730 or FD332-PERC storage controllers (2109665)

Note: A step-by-step driver update video is available here: Updating FD332 PERC driver in VMware ESXi on PowerEdge FX2 chassis. The same procedure applies to the PERC H730 in R630 servers.

Install the updated driver on all hosts in the VSAN cluster and reboot the hosts. When the hosts have come back online, click **Retest**. This test should pass.

10.4.3 Warning: Performance Service / Stats DB object

The warning should disappear after the VSAN performance service is enabled.

To enable the VSAN performance service:

1. In the web client, go to **Hosts and Clusters** and select a cluster containing a VSAN.
2. In the center pane, go to **Manage > Settings > Virtual SAN > Health and Performance**.
3. Next to **Performance Service is Turned Off**, click **Edit**. Check the **Turn On** box and click **OK**.

Repeat as needed for other VSAN-enabled clusters, then click **Retest.** This test should pass.

10.5 Verify IGMP snooping functionality

On directly connected physical switches, the `show ip igmp snooping groups vlan 44` command can be issued to verify IGMP snooping is functioning properly.

There should be three groups at any given time. Multicast address group 224.1.2.3 is the default VSAN group for master nodes. Multicast group 224.2.3.4 is the member group and should contain all ESXi host-connected interfaces (port channels 1-4 in this example).

```
FN410S-A1#show ip igmp snooping groups vlan 44
IGMP Connected Group Membership
Group Address    Interface       Mode          Uptime    Expires   Last Reporter
224.1.2.3        Vlan 44         IGMPv2        1w0d      00:01:36  10.44.2.54
  Member Ports: Po 2, Po 3
224.2.3.4        Vlan 44         IGMPv2        1w0d      00:01:41  10.44.2.53
  Member Ports: Po 1, Po 2, Po 3, Po 4
239.255.255.253  Vlan 44         IGMPv2        1w0d      00  :01:39  10.44.2.53
  Member Ports: Po 1, Po 2, Po 3, Po 4
```

11 Configure the NSX virtual network

This section covers the configuration steps and best practices to build the NSX topology used in this guide. For more information, refer to the VMware NSX 6.2 Documentation Center.

11.1 NSX Manager

The NSX Manager is the centralized network management component of NSX. A single NSX Manager serves a single vCenter Server environment. It provides the means for creating, configuring, and monitoring NSX components such as controllers, logical switches and edge services gateways.

In this guide, NSX Manager is installed as a virtual appliance on an ESXi host in the management cluster. It is available from VMware as an Open Virtualization Appliance (.ova) file and is available for download at my.vmware.com.

To install NSX Manager:

1. On the web client **Home** screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, right click on the target ESXi host in the Rack 1 Management cluster and select **Deploy OVF Template**.
3. Select **Local file** and **Browse** to the .ova file (the current naming format is VMware-NSX-Manager-*version#*.ova). Select the file, click **Open > Next.**
4. Check the **Accept extra configuration options** box and click **Next**.

Note: The extra configuration options include IP address, default gateway, DNS, NTP, and SSH.

5. Click **Accept** on the **Accept license agreements** page. Click Next.
6. Keep the default name, **NSX Manager.** Select **Datacenter** and click **Next**.
7. On the **Select storage** screen, select the previously configured VSAN datastore (covered in Section 10). Click **Next.**

Note: It is a best practice to use VSAN storage to allow for a High Availability (HA) cluster configuration, so that the NSX Manager appliance can be restarted on another host if the original host fails. See the VMware vSphere 6.0 Documentation Center for HA cluster configuration instructions.

8. On the **Setup networks** screen, select the management network, named **VM Network** by default. Click **Next.**
9. On the **Customize template** screen:
   a. Enter the **CLI admin** and **CLI Privilege Mode** passwords to be used at the NSX Manager CLI.
   a. Expand **Network properties.** Provide a **hostname** (for example, nsxmanager). The IP address and gateway information may be filled out or supplied by a DHCP server on your network.
   b. Fill out the **DNS** section if used on your network and if not provided by DHCP.
   c. Under **Services Configuration**, Provide the **NTP server** host name or IP address. It is a best practice to use NTP on your NSX management network. Optionally, check the box to **Enable SSH.**
   d. Click **Next.**
10. The **Ready to complete** screen provides a summary of the installation as shown in Figure 50. Review your settings, check the **Power on after deployment** box, and click **Finish**.

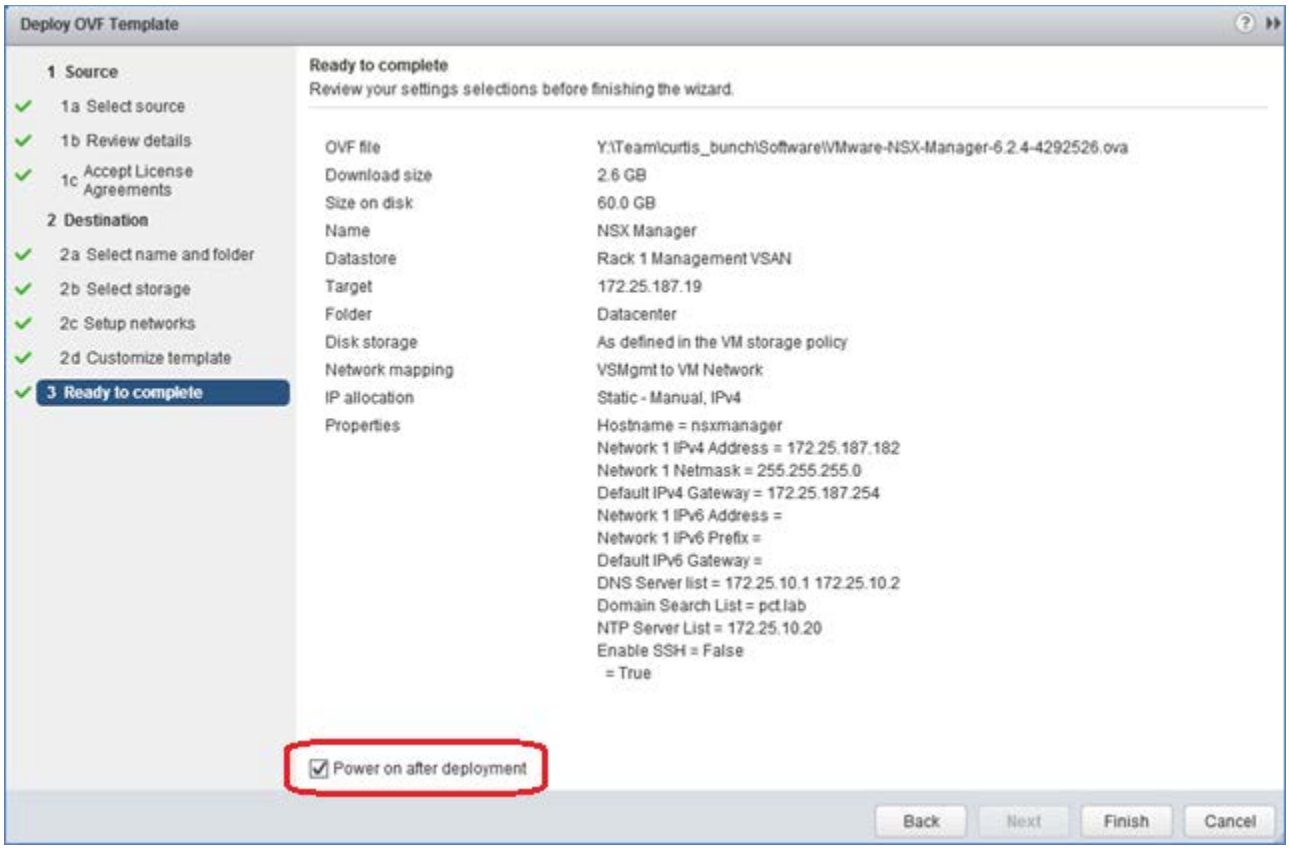


Figure 50 NSX Manager summary screen

The NSX Manager appliance is deployed and boots its Linux OS.

Note: If needed, the NSX Manager appliance console is accessible by going to Hosts and Clusters and expanding the Rack 1 Management cluster. Right click on the NSX Manager virtual machine and select Open Console.

11.2 Register NSX Manager with vCenter Server

Note: Only one NSX Manager can be registered with a vCenter Server.

To register the vCenter Server with NSX Manager:

1. After the NSX Manager appliance has booted, go to **https://*<ip_address_or_nsx_manager_hostname>*** in a web browser.
2. Login as **admin** with the NSX Manager password specified in the previous section.
3. Click the **Manage vCenter Registration** button.
4. Next to **vCenter Server**, click the **Edit** button.

5. Enter the IP address or host name of the vCenter Server, the vCenter Server user name (for example, administrator@pct.lab), password, and click **OK.**
6. Click **Yes** to trust the certificate when prompted.
7. Verify the vCenter Server status is **Connected** as shown in Figure 51.

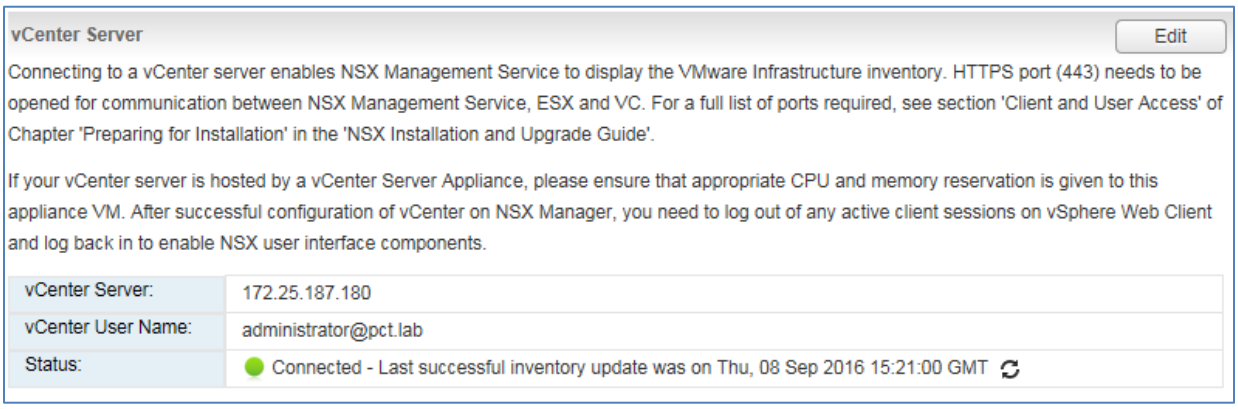


Figure 51 NSX Manager is connected to vCenter Server

The following steps are done in the web client:

1. Log out of the web client if logged in.
2. Log into the web client using the same credentials used to register NSX Manager.
3. The **Networking & Security** icon now appears on the **Home** page as shown in Figure 52.

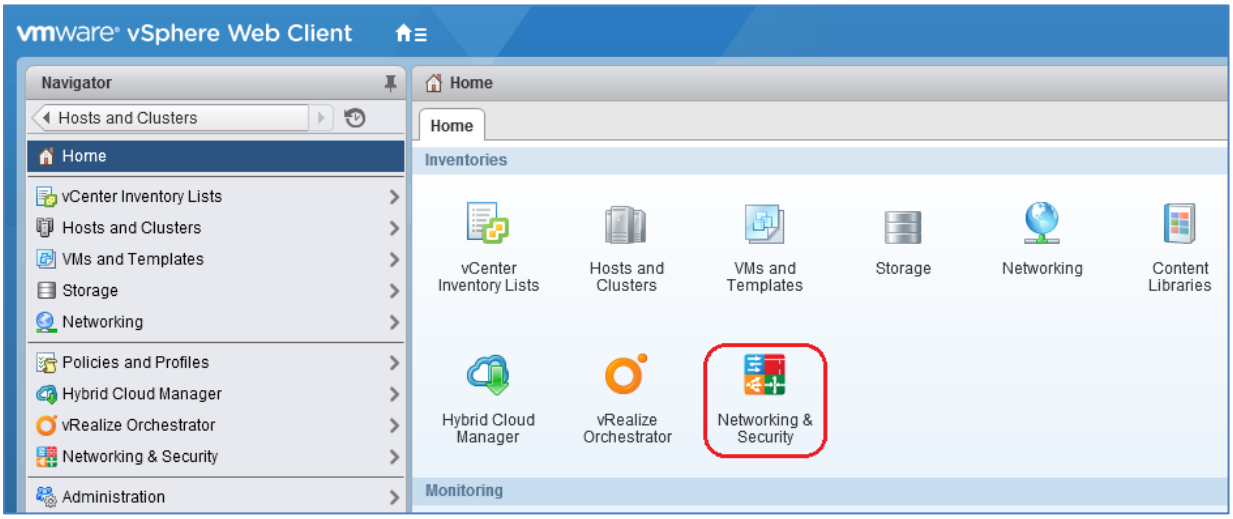


Figure 52 Networking & Security icon

11.3 Deploy NSX controllers

NSX controllers are responsible for managing the distributed switching and routing modules in the hypervisors. Three controllers are required in a supported configuration and can tolerate one controller failure while still providing for controller functionality.

NSX controllers communicate on the management network, and do not have any data plane traffic passing through them. Therefore, data forwarding will continue even if all NSX controllers are off line.

As a best practice, each NSX controller should be deployed on a different ESXi host so that a single host failure will not bring down more than one controller. In this guide, there are three hosts in the management cluster with one NSX controller deployed to each host.

To deploy the NSX controllers:

1. In the web client, go to **Home > Networking & Security**.
2. In the **Navigator** pane, select **Installation**.
3. On the **Management** tab under **NSX Controller nodes**, select ✚ to open the **Add Controller** dialog box.
4. In the **Add Controller** dialog box:
   a. Provide a **name** for the first controller, such as NSX Controller 1.
   b. The **NSX Manager** and **Datacenter** should be selected by default. If not, select them.
   c. Next to **Cluster/Resource Pool**, select the **Rack 1 Management** cluster.
   d. Next to **Datastore**, select a previously configured VSAN datastore (covered in Section 10).
   e. Next to **Host,** select the ESXi host where the NSX controller VM will reside. Use a different host for each controller.
   f. The **Folder** is optional and is skipped for this guide.
   g. Next to **Connected To**, click **Select**. Next to **Object Type** select **Network**. Select the management network, **VM Network,** and click **OK.**
   h. Next to **IP Pool**, click **Select.** When creating the first controller, an IP pool will need to be created. Click **New IP Pool**, and fill out the **Add Static IP Pool** fields similar to the example in Figure 53. Use an address range containing at least 3 available addresses on your management network (one address will be used for each NSX controller).

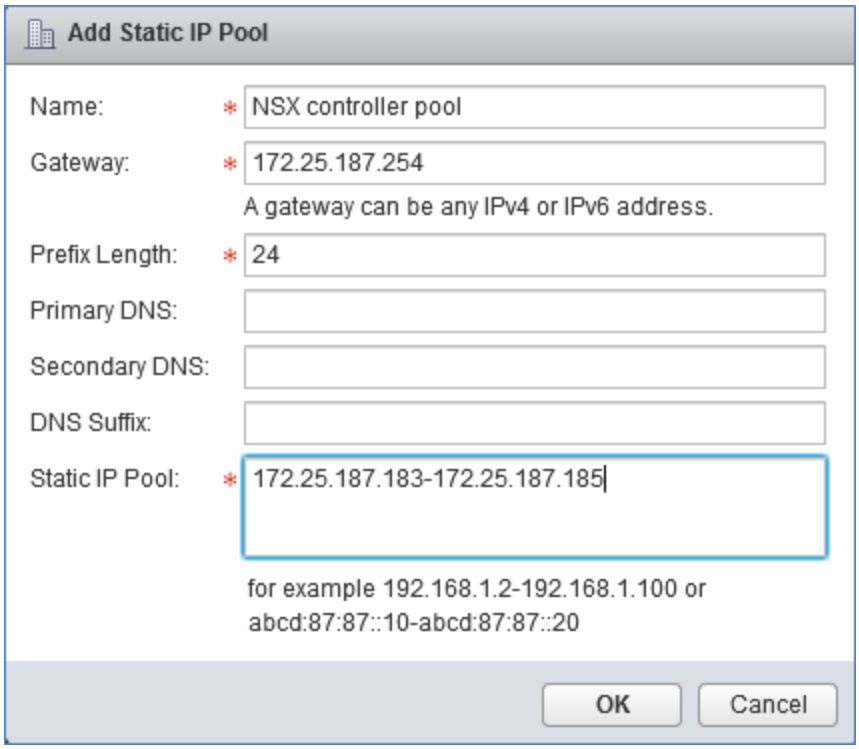


Figure 53 Add Static IP Pool dialog box

i. Click **OK** to create the pool.
j. In the **Select IP Pool** dialog box, select the pool and click **OK.**
k. Type and confirm a complex password for the controller cluster. This field only appears when setting up the first controller.
l. Click **OK** again to deploy the NSX controller**.**

Note: Wait for the deployment to complete as shown in the Status column under NSX Controller nodes before deploying the next controller.

5. Repeat Steps 3 and 4 above for the remaining two controllers, except use the existing IP pool instead of creating a new one in step h.

When all three controllers are deployed, the **Networking & Security > Installation > Management** page appears similar to Figure 54. Each controller's status is shown as **Connected** and each controller has two green boxes (representing status of each controller peer) in the **Peers** column.

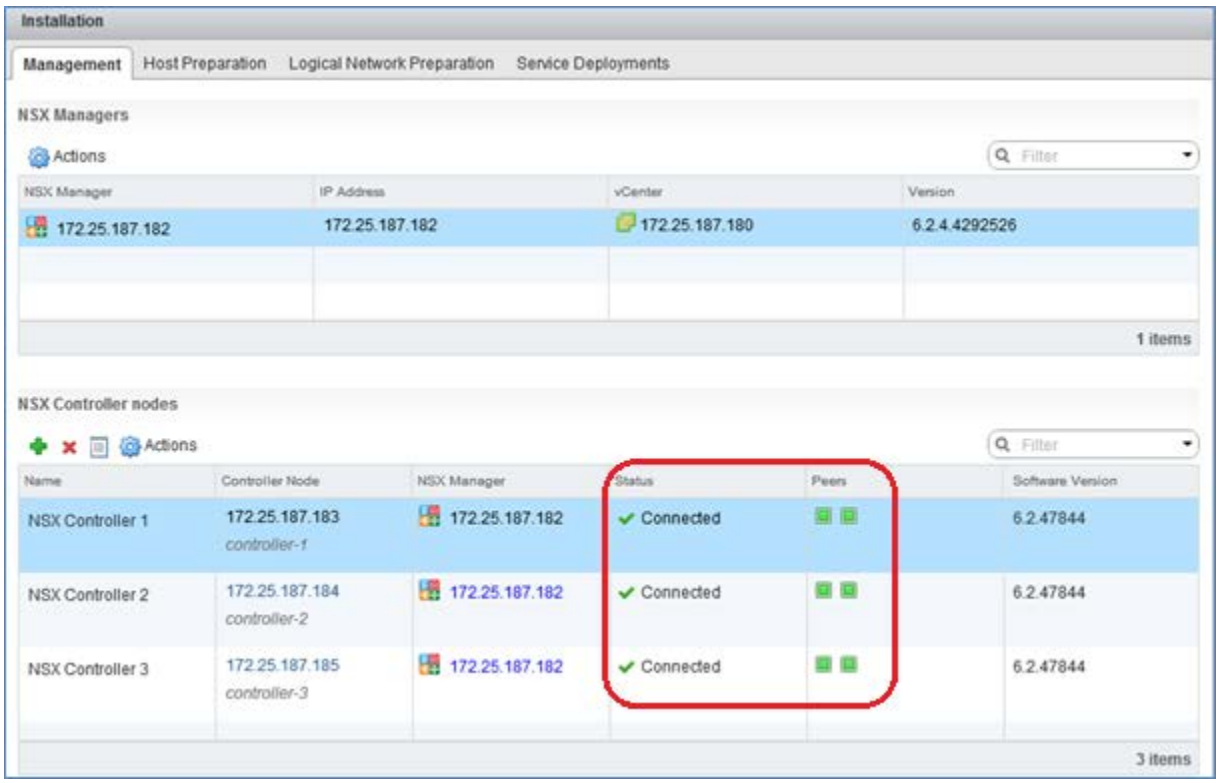


Figure 54 NSX controllers deployed

11.4 Prepare host clusters for NSX

Host preparation is the process in which the NSX Manager installs NSX kernel modules on each ESXi host in a cluster. This only needs to be done on clusters that will send and receive traffic on the NSX (virtual) network. In this deployment example, this includes the Rack 2 Compute FC430 and Rack 3 Edge clusters. The Rack 1 Management cluster will not be part of the virtual network.

To prepare the Compute cluster:

1. On the web client **Home** screen, select **Networking & Security**.
2. Select **Installation > Host Preparation**.
3. Click on the row containing **Rack 2 Compute FC430** cluster. Click **Actions** and click **Install > Yes**.
4. When complete, the cluster's **Installation Status** and **Firewall columns** will both show a green check mark.

Note: The VXLAN column will still indicate Not Configured. VXLAN will be configured in the next section.

Repeat steps 1-4 above for the Edge cluster.

When complete, the host preparation tab will appear similar to Figure 55.

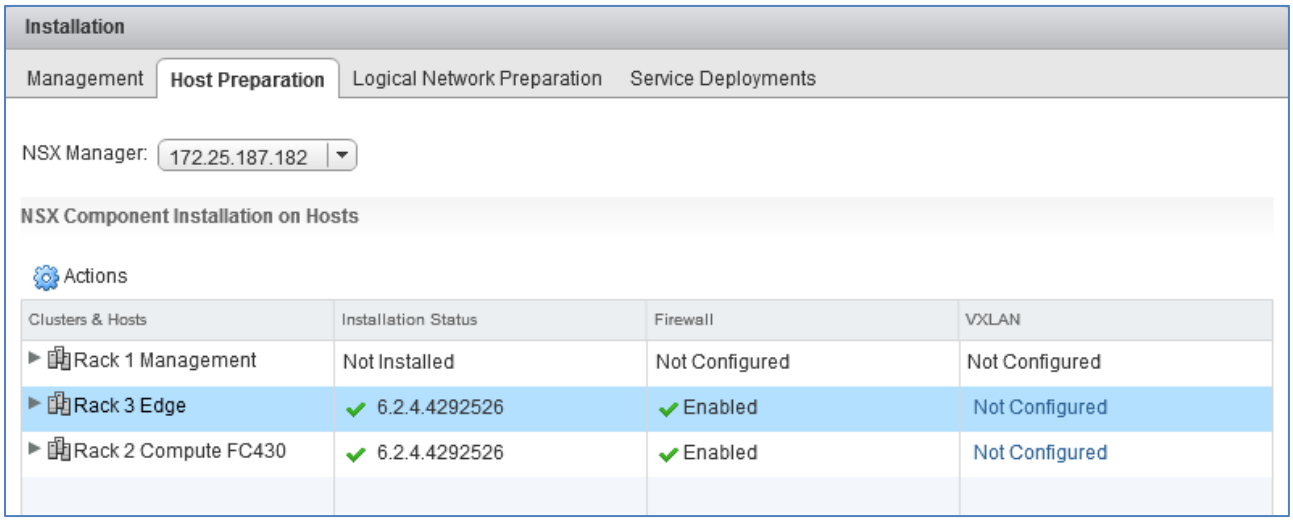


Figure 55 Host Preparation status

The host preparation process installs two vSphere Installation Bundles (VIBs) to each host in the cluster, esx-vsip and esx-vxlan. This can be confirmed running the following command from an ESXi host CLI:

```
esxcli software vib list | grep esx-v
```

The command output will include esx-vsip and esx-vxlan if installed successfully:

```
esx-vsip      6.0.0-0.0.4249023      VMware  VMwareCertified   2016-09-02
esx-vxlan     6.0.0-0.0.4249023      VMware  VMwareCertified   2016-09-02
```

Note: If additional hosts are later added to the prepared clusters, the required NSX components will be automatically deployed to those hosts.

11.5 Configure clusters for VXLAN

VXLAN is configured on a per-cluster basis with each cluster mapped to a VDS. VXLAN configuration creates a VMkernel interface on each host that serves as the software VTEP. This enables virtual network functionality on each host in the cluster.

Before starting, plan an IP addressing scheme for the VTEPs. The number of IP addresses in the pool should be enough to cover all ESXi hosts in the cluster participating in NSX. In this guide, VLAN 55 is used for VXLAN traffic and the IP addressing scheme is shown in Table 5.

Table 5 VTEP IP pool addresses

Cluster Name	IP Pool Name	Network	Gateway	IP Pool Range
Rack 2 Compute FC430	R2 VTEP Pool	10.55.2.0/24	10.55.2.254	10.55.2.1-100
Rack 3 Edge	R3 VTEP Pool	10.55.3.0/24	10.55.3.254	10.55.3.1-100

Note: The Rack 1 Management cluster is not configured because its hosts are not part of the virtual network in this guide.

To configure the Rack 2 Compute FC430 cluster for VXLAN:

1. Go to **Home > Networking & Security > Installation** and select the **Host Preparation** tab.
2. In the center pane, select the **Rack 2 Compute FC430** cluster. Click **Actions** > **Configure VXLAN.**
3. In the **Configure VXLAN Networking** dialog box:
   a. Next to **Switch**, ensure the correct VDS is selected (for example, Rack 2 Compute FC430 VDS).
   b. Set the **VLAN** to **55.**
   c. Leave the **MTU** set to **1600.**
   d. Next to **VMKNic IP Addressing**, select **Use IP Pool** and select **New IP Pool** from the drop-down menu. This opens the **Add Static IP Pool** dialog box.
      i. Next to **Name**, enter **R2 VTEP Pool.**
      ii. Set the **Gateway** to **10.55.2.254.**
      iii. Set the **Prefix Length** to **24** (number of bits in the subnet mask).
      iv. Fill out the DNS information if used on your network.
      v. Set the **Static IP Pool** to **10.55.2.1-10.55.2.100.**
      vi. Click **OK.**
4. On the **Configure VXLAN Networking** window, set the **VMKNic Teaming Policy** to **Enhanced LACP.**
5. Click **OK.**

It may take a few minutes for VXLAN configuration to complete. When done, the VXLAN column should indicate **Configured** with a green check mark as shown in Figure 56:

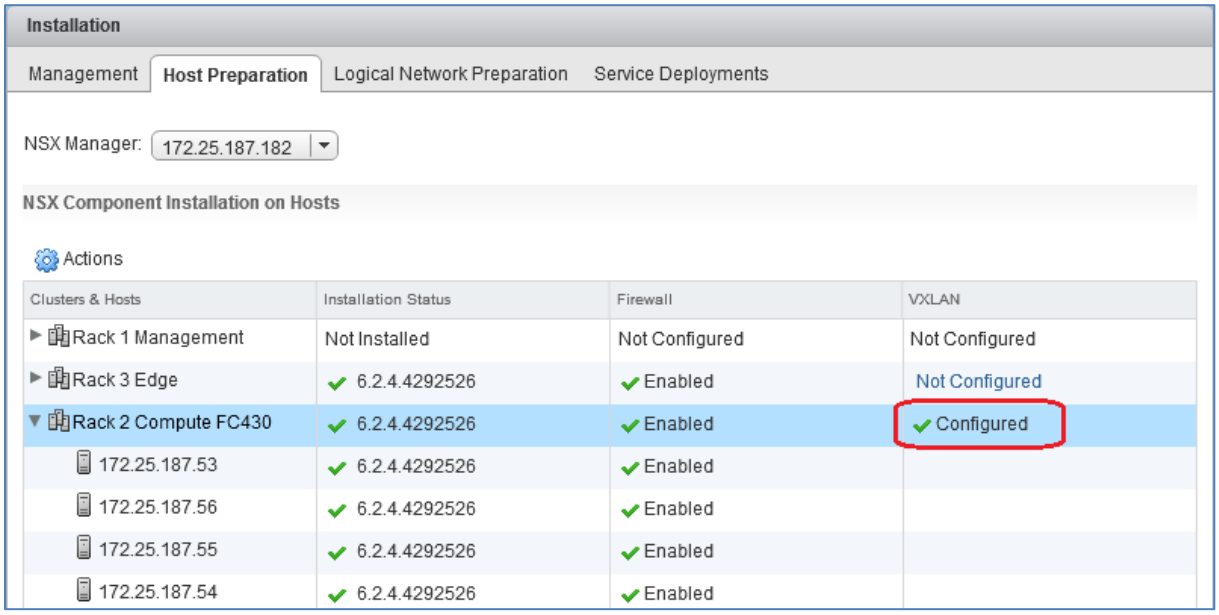


Figure 56 VXLAN successfully configured on Rack 2 Compute FC430 cluster

Repeat steps 1-5 above for the Rack 3 Edge cluster with the following changes:

- Step 3.a. - ensure the Rack 3 Edge VDS is selected.
- Step 3.d. - Replace the pool name and IP addressing as needed per Table 5.

When VXLAN configuration is complete, verify the configuration by viewing the network topology as follows:

1. Go to **Home > Networking**.
2. Select a VDS in a cluster configured for VXLAN, such as **Rack 2 Compute FC430 VDS**.
3. In the center pane, select **Manage > Settings > Topology**.
4. In the topology diagram, there is a new port group with the prefix **vxw-vmknicPg-dvs.** It is on VLAN 55 and has one VMkernel port for each host in the cluster. Each port has an IP address from the VTEP pool for the cluster, as shown in Figure 57.

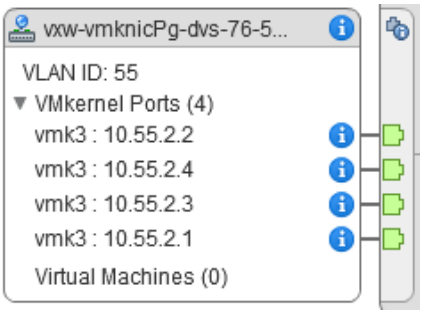


Figure 57 VXLAN VMkernel ports with VTEP IP addresses assigned

11.6 Create a segment ID pool

VXLAN tunnels are built between VTEPs. An ESXi host is an example of a typical software VTEP. Each VXLAN tunnel must have a segment ID, which is pulled from a segment ID pool that you create. Segment IDs are used as VNI's. The range of valid IDs is 5000-16777215.

Note: Do not configure more than 10,000 VNIs in a single vCenter. vCenter limits the number of distributed port groups to 10,000.

To create a segment ID pool:

On the web client **Home** screen, select **Networking & Security > Installation** and select the **Logical Network Preparation** tab.

1. Click **Segment ID** and click the **Edit** button.
2. Enter a contiguous range for **Segment ID pool**, for example **5000-5999.**
3. Leave the remaining items at their defaults, as shown in Figure 58, and click **OK.**

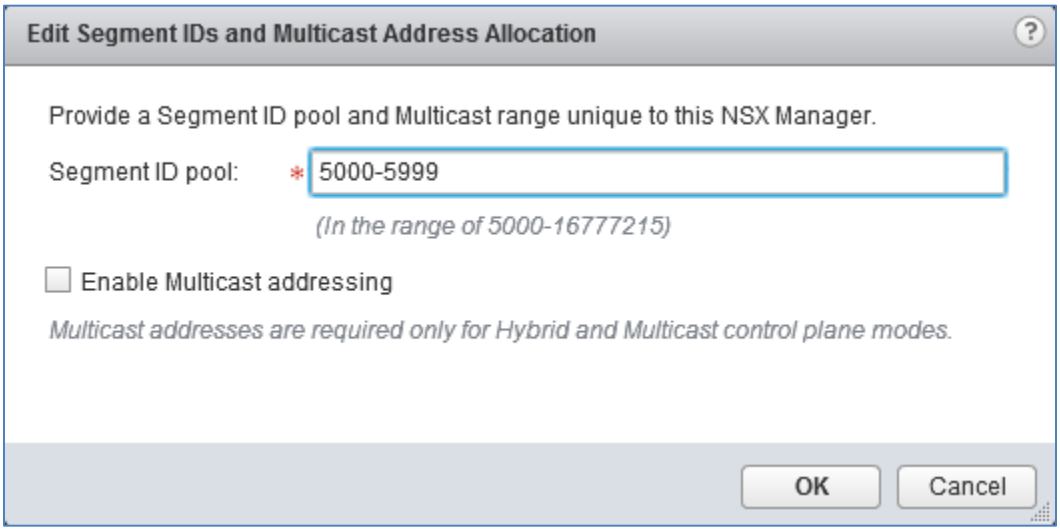


Figure 58 Segment ID pool dialog box

11.7 Add a transport zone

A transport zone controls which hosts a logical switch can reach. It can span one or more clusters. Transport zones dictate which clusters and, therefore, which VMs can use a particular virtual network.

An NSX environment can contain one or more transport zones. A cluster can belong to multiple transport zones while a logical switch can belong to only one transport zone. A single transport zone is used in this guide for all NSX-enabled clusters.

To create a transport zone:

1. Go to **Home > Networking & Security > Installation** and select the **Logical Network Preparation** tab.
2. Select **Transport Zones** and click the + icon.
3. Name the zone Transport Zone 1.
4. Leave **Replication mode** set to **Unicast**.
5. Select the clusters to add to the transport zone. **Rack 2 Compute FC430** and **Rack 3 Edge** are selected as shown in Figure 59:

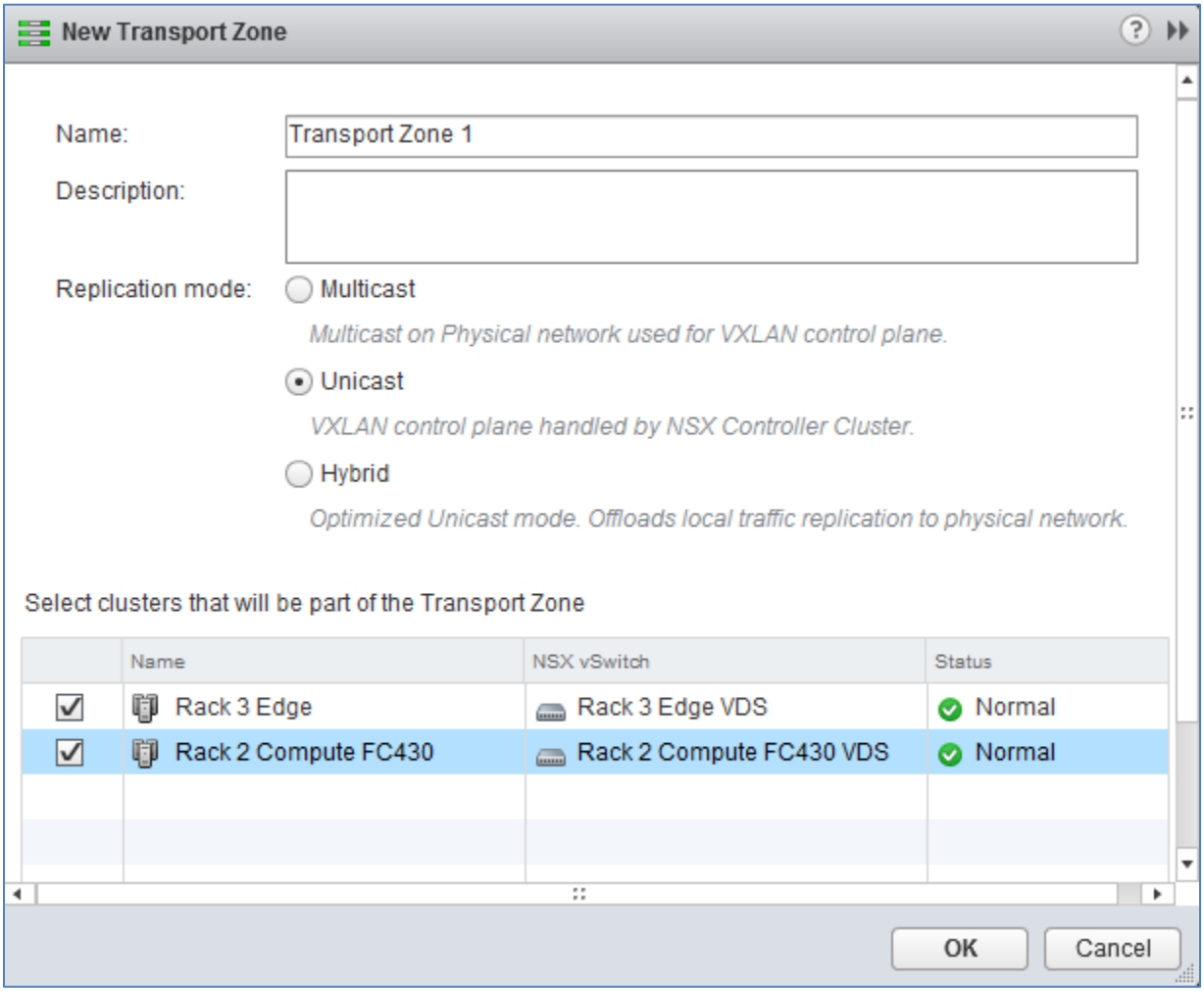


Figure 59 Transport zone with two attached clusters

6. Click **OK** to create the zone.

11.8 Logical switch configuration

An NSX logical switch creates a broadcast domain similar to a physical switch or a VLAN. This deployment creates four logical switches, each of which each is associated with a unique VNI. The VNI is automatically assigned from the segment ID pool created in Section 11.6.

Table 6 shows the four logical switches used in this deployment:

Table 6 Logical Switch and VNI Assignment

Logical switch name	VXLAN network ID (VNI)	Network	Used for
Transit Network	5000	172.16.0.0/24	Transit network
Web-Tier	5001	10.10.10.0/24	Web network
App-Tier	5002	10.10.20.0/24	Application network
DB-Tier	5003	10.10.30.0/24	Database network

Figure 60 shows the logical connectivity between the three logical switches used for VM traffic (Web-Tier, App-Tier, and DB-Tier), the Distributed Logical Router (DLR) and the transit logical switch. The DLR acts as the default gateway for each VM connected to its respective logical switch and is configured in Section 11.9.

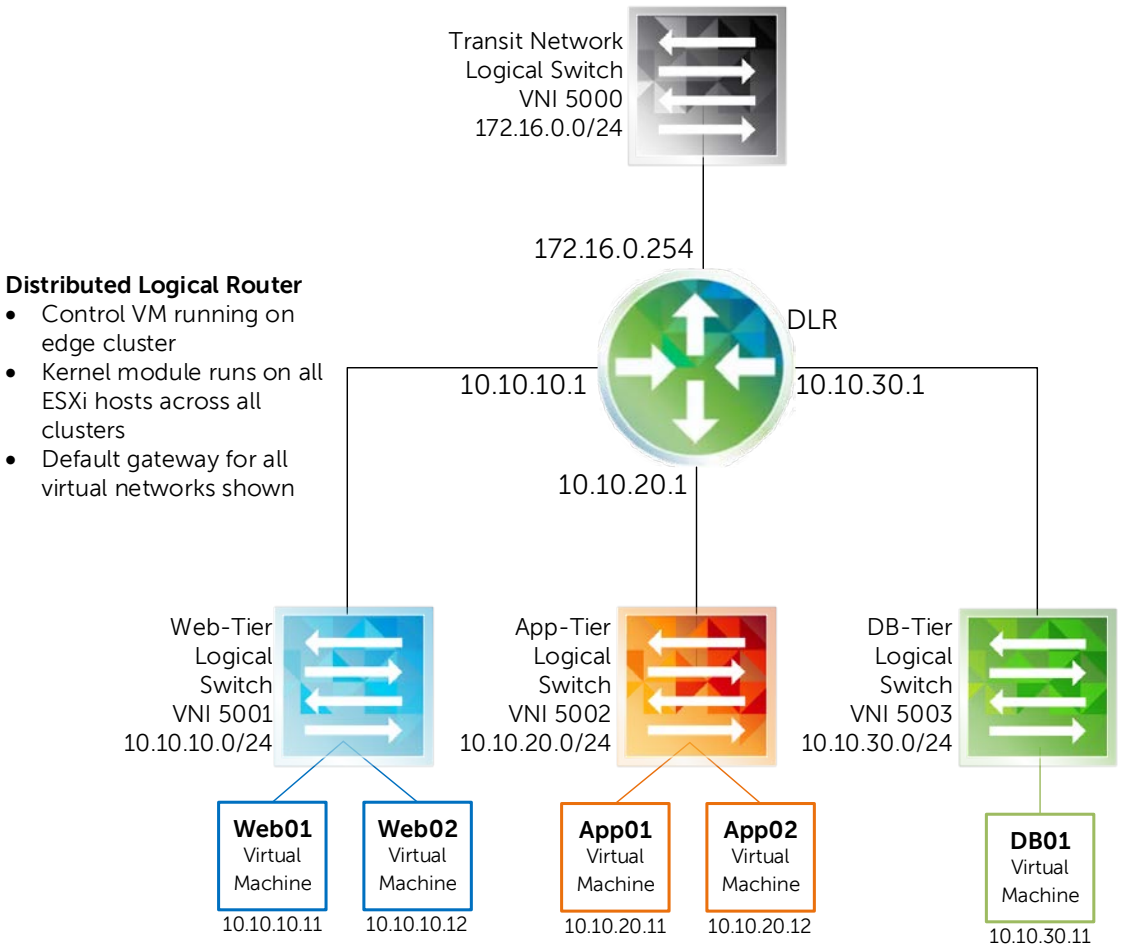


Figure 60 Logical switches and DLR topology

To deploy the four logical switches:

1. On the web client Home screen, select **Networking & Security > Logical Switches**.
2. Click the ✚ icon to add a new logical switch.
3. In the **New Logical Switch** dialog box:
   a. Type the first switch name, **Transit Network.**
   b. Next to **Transport Zone**, **Transport Zone 1** should already be selected. If not, click **Change** and select it.
   c. Leave **Replication mode** set to **Unicast, Enable IP Discovery** checked and **Enable MAC Learning** unchecked**.**
   d. Click **OK**.

Repeat the steps above for the remaining logical switches and substitute the proper switch name in step 3a. Ensure that all logical switches are placed in Transport Zone 1.

Figure 61 shows the four logical switches after creation:

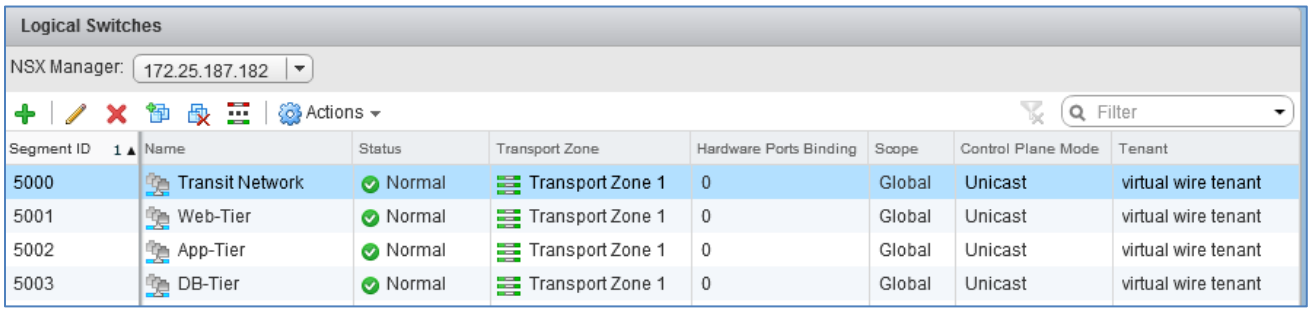


Figure 61 Logical switches after creation

11.9 Distributed Logical Router configuration

A Distributed Logical Router (DLR) is a virtual appliance that provides routing between VXLAN networks. It is installed on a host in the Rack 3 Edge cluster.

Figure 60 shows the DLR's location in the virtual network. All four logical switches connect to it. Table 7 shows the DLR interface IP addresses used in this guide:

Table 7 DLR IP addressing

Interface name	IP address/Subnet prefix
Transit Network	172.16.0.254/24
Web-Tier	10.10.10.1/24
App-Tier	10.10.20.1/24
DB-Tier	10.10.30.1/24

To configure the DLR:

1. Go to **Home > Networking & Security > NSX Edges** and click the + icon.
2. In the **New NSX Edge** dialog box:
   a. Select **Logical (Distributed) Router,** provide a name (**DLR1,** for example**).** Verify that the **Deploy Edge Appliance** box is checked and click **Next.**
   b. Provide CLI credentials for the DLR, leave other values at their defaults and click **Next.**
   c. Under **NSX Edge Appliances**, click the + icon to create an edge appliance.

d. In the **Add NSX Edge Appliance** dialog box:
   i. Set **Cluster/Resource Pool** to **Rack 3 Edge.**
   ii. Set **Datastore** to the VSAN configured in Section 10, **Rack 3 Edge VSAN** as shown in Figure 62. **Host** and **Folder** may be left blank. The host is automatically assigned from the cluster.

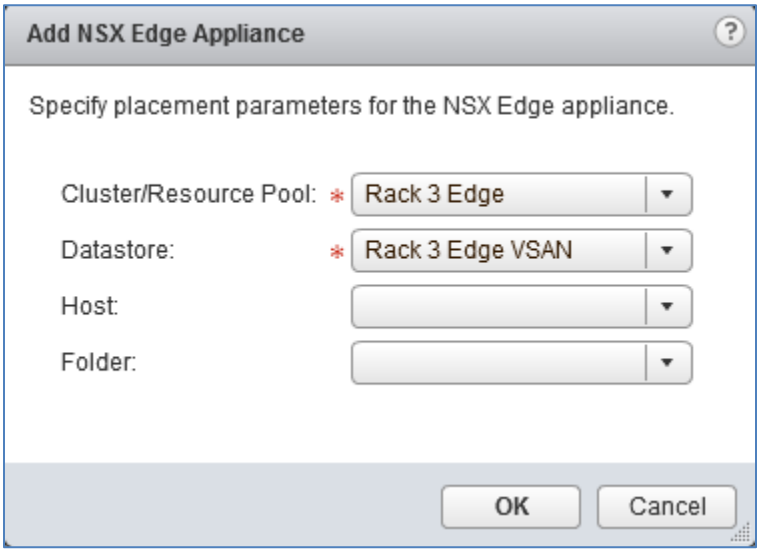


Figure 62 Add NSX Edge Appliance dialog box

e. Click **OK** to close the **Add NSX Edge Appliance** dialog box and click **Next.**
f. On the **Configure Interfaces** page, next to **Connected To,** click **Select**.
g. Be sure **Logical Switch** is selected at the top, and select **Transit Network**. Click **OK**.
h. To create the DLR uplink interface:
   i. Under **Configure interfaces of this NSX Edge**, click the icon to open the **Add Interface** dialog box.
   ii. Name the interface **Transit Network.**
   iii. Set **Type** to **Uplink.**
   iv. Next to **Connected To** click **Select**.
   v. Be sure **Logical Switch** is selected at the top and select **Transit Network**. Click **OK.**
   vi. Under **Configure Subnets,** click the icon.
   vii. Type **172.16.0.254** for the **Primary IP Address** and set the **Subnet Prefix Length** to **24.**
   viii. Leave the remaining values at their defaults and click **OK** to close.
i. To create the DLR internal interfaces:
   i. Under **Configure interfaces of this NSX Edge**, click the icon to open the **Add Interface** dialog box.
   ii. Name the interface **Web-Tier.**
   iii. Set **Type** to **Internal.**
   iv. Next to **Connected To** click **Select**.
   v. Be sure **Logical Switch** is selected at the top and select **Web-Tier**. Click **OK.**
   vi. Under **Configure Subnets, c**lick the icon.

vii. Type **10.10.10.1** for the **Primary IP Address** and set the **Subnet Prefix Length** to **24.**
viii. Leave the remaining values at their defaults and click **OK.**

j. Repeat steps i-viii under letter i above to create the remaining DLR internal interfaces (App-Tier and DB-Tier in this example). Substitute **Name, Connected To**, and **IP Address** values accordingly as Figure 63 shows:

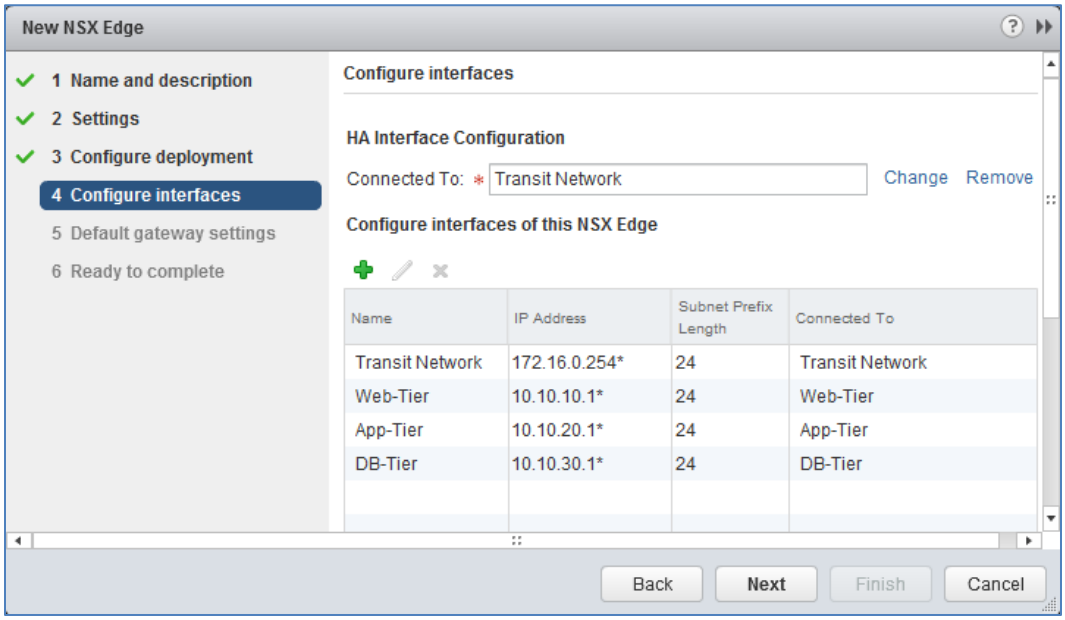


Figure 63 DLR interfaces configured

k. Click **Next** when complete.
l. Uncheck **Configure Default Gateway** and click **Next**.
m. Click **Finish** to deploy the DLR. It may take a few minutes to complete.

To validate DLR settings and status:

1. Go to **Home > Networking & Security > NSX Edges**.
2. In the center pane, double click on the DLR to open the DLR summary and management page.
3. Select **Manage** > **Settings > Interfaces**. Verify all settings are correct as shown in Figure 64. If any changes need to be made, click the pencil icon to edit.

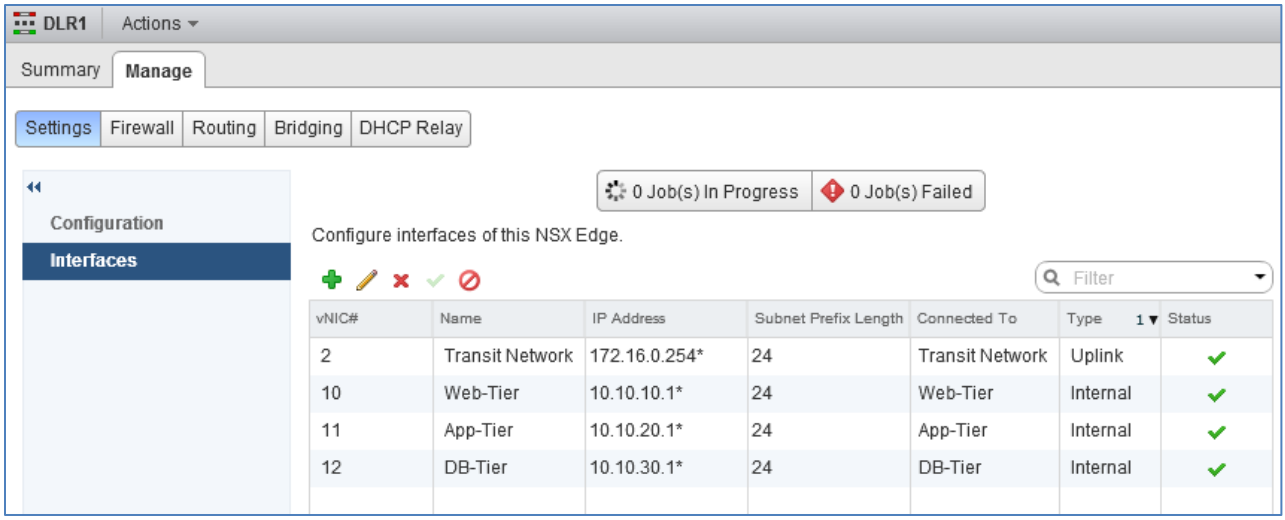


Figure 64 Configured interfaces on the distributed logical router

11.9.1 Configure OSPF on the DLR

This topology uses OSPF to provide dynamic routing to the ESG. BGP can be used instead, but for this guide OSPF was selected to provide a distinct demarcation between the physical underlay and the virtual overlay. The ESG serves as the next-hop router in this environment, and is configured in Section 13.1. The NSX default Area 51, which is a not-so-stubby area (NSSA), will be used between the DLR and the ESG.

Configure the Router ID:

1. Go to **Home > Networking & Security > NSX Edges**.
2. In the center pane, double click on the DLR to open the DLR summary and management page.
3. Select **Manage > Routing > Global Configuration**. Click **Edit** in the **Dynamic Routing Configuration** section.
4. Next to **Router ID**, choose the default (**Transit Network – 172.16.0.254**) and click **OK**.
5. Click **Publish Changes** near the top of the screen.

Enable OSPF and configure OSPF features:

1. On the **Routing** page, select **OSPF.**
2. Click **Edit** at the top right corner of the window and check the **Enable OSPF** box.
3. Set the **Protocol Address** to 172.16.0.253.
4. Set the **Forwarding Address** to 172.16.0.254.
5. Leave the **Enable Grateful Restart** box checked and click **OK**.
6. Click **Publish Changes.**

Note: The protocol and forwarding addresses should be from the same subnet. The protocol address is used to form OSPF adjacencies. The forwarding address is the DLR interface IP address.

Enable interfaces to participate in their respective OSPF areas:

1. Click the icon under **Area to Interface Mapping**.
2. Next to **Interface**, select **Transit Network**.
3. Set the **Area** to **51** (default).
4. Leave all other values at their defaults and click **OK.**
5. Click **Publish Changes**.

When complete, the OSPF page for the DLR should appear similar to Figure 65.

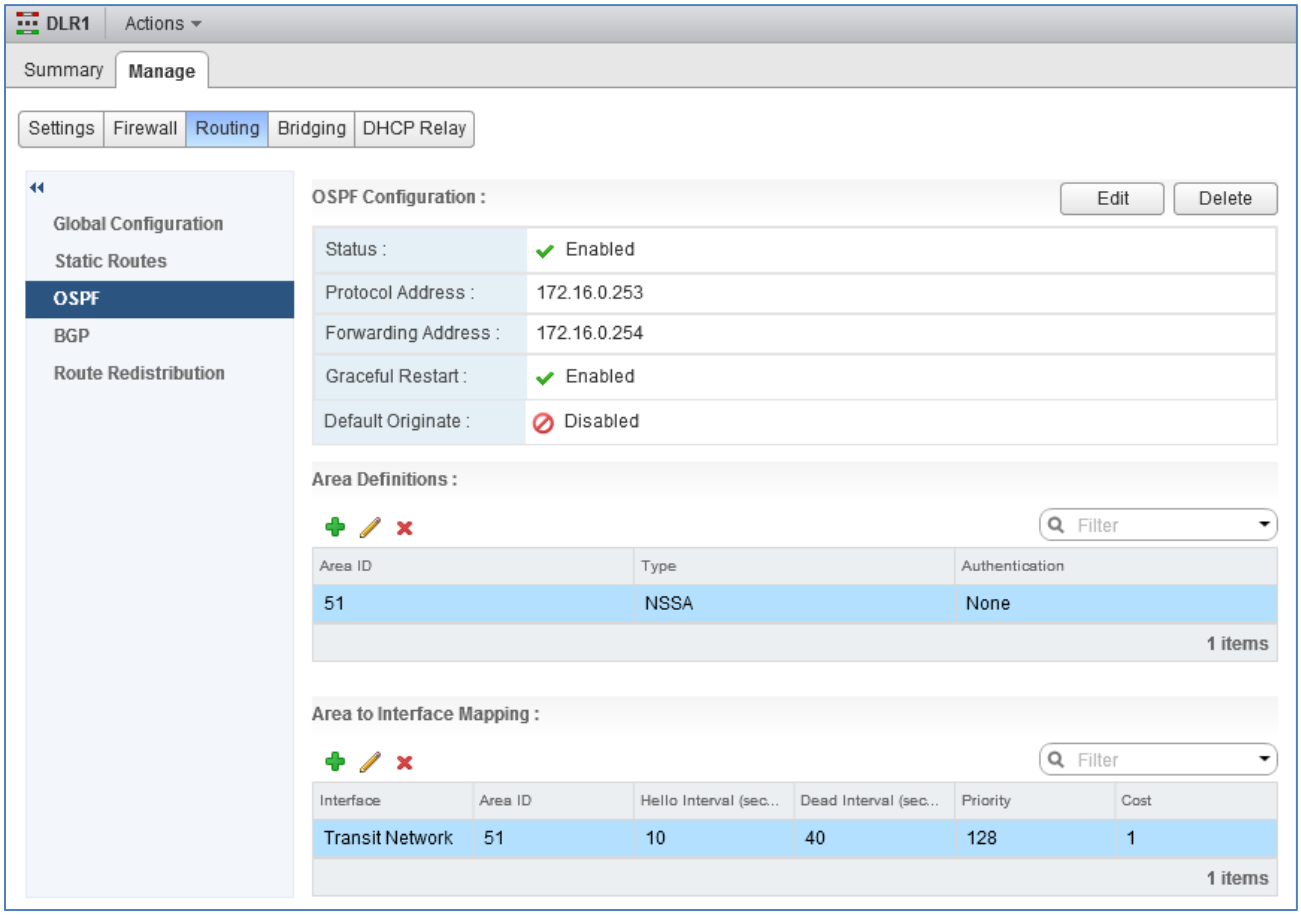


Figure 65 OSPF configuration complete on the DLR

11.9.2 Firewall information

The DLR firewall can be accessed by going to **Home > Networking & Security > NSX Edges**. Double click on the DLR and go to **Manage > Firewall**.

Note: Configuration of firewall rules is outside the scope of this document. For more information, refer to the VMware NSX 6.2 Documentation Center.

12 Verify NSX network functionality

In this section, a small number of virtual machines are deployed to different clusters to verify connectivity within the NSX network.

Note: Virtual machine/guest operating system deployment steps are not included in this document. For instructions, see the Deploying Virtual Machines section of the vSphere 6.0 online documentation. Guest operating systems can be any supported by ESXi 6.0. Microsoft Windows Server 2012 R2 was used as the guest operating system for each virtual machine deployed in this section.

12.1 Deploy virtual machines

For this example, three VMs are deployed in the Rack 2 Compute FC430 cluster. The first two represent application servers and are named App-VM1 and App-VM2. The third represents a web server and is named Web-VM1.

A fourth VM, App-VM3, is deployed in the Rack 3 Edge cluster to validate communication between clusters. The added VMs are shown in Figure 66.

Note: The Rack 1 Management cluster is not configured for VXLAN traffic and therefore is not part of the virtual network validation.

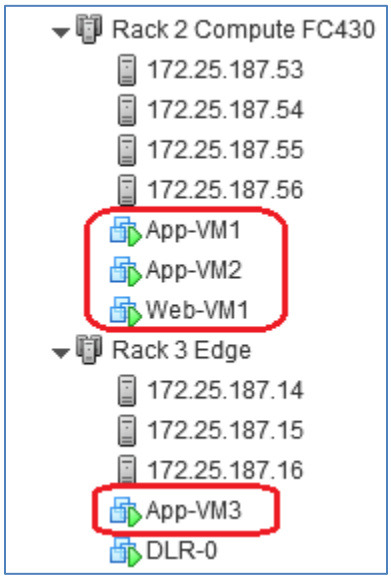


Figure 66 Hosts and Clusters view - virtual machines deployed

12.2 Connect virtual wires

A virtual wire is a distributed port group that is automatically created on each VDS as logical switches are created. The virtual wire descriptor contains the name of the logical switch and the logical switch's segment ID.

To connect a VM to a virtual wire:

1. Go to **Home > Hosts and Clusters**.
2. Right click on the first VM, **App-VM1**, and select **Edit Settings**.
3. Next to **Network adapter 1**, select the virtual wire on the **App-Tier** network as shown in Figure 67.

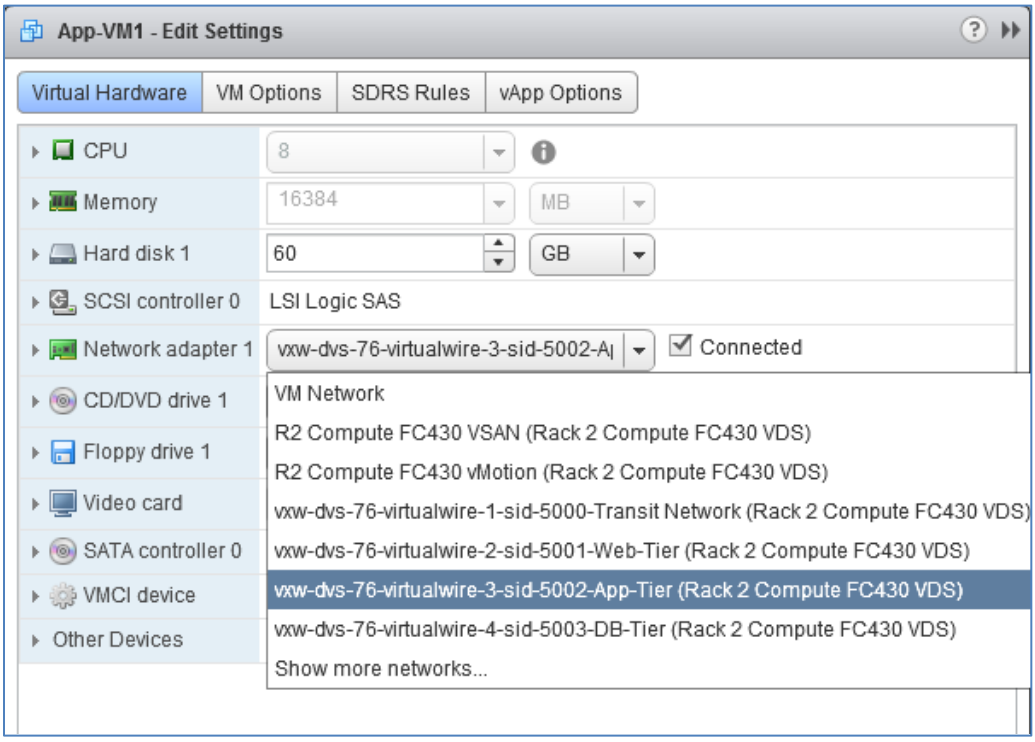


Figure 67 Virtual wire on App-Tier network selected

Repeat steps 1-3 for remaining VMs to be placed on the **App-Tier** segment (**App-VM2** and **App-VM3** in this example.)

Repeat for the VM named **Web-VM1**, except select the virtual wire on the **Web-Tier** segment in step 3.

12.3 Configure networking in the guest OS

Power on the virtual machines. Log in to a guest OS by right clicking on the VM and selecting **Open Console**. Use the normal procedure in the guest OS to configure networking.

Using the virtual networking IP address scheme covered in Section 11.8, the IP addresses, subnet masks, and gateway are configured on each VM per Table 8. The gateway addresses are the DLR internal interfaces.

Table 8 Virtual machine IP addressing

Virtual Machine	Cluster	IP Address	Gateway
Web-VM1	Rack 2 Compute FC430	10.10.10.11/24	10.10.10.1
App-VM1	Rack 2 Compute FC430	10.10.20.11/24	10.10.20.1
App-VM2	Rack 2 Compute FC430	10.10.20.12/24	10.10.20.1
App-VM3	Rack 3 Edge	10.10.20.13/24	10.10.20.1

12.4 Test connectivity

Note: Guest operating system firewalls may need to be temporarily disabled or modified to allow responses to ICMP ping requests for this test. By default, the firewall settings on the DLR allow this type of internal traffic.

Within the source guest operating system, ping the destination VMs using Table 9 as a guide. Successful pings validate the segment tested is configured properly.

Table 9 Test examples to validate connectivity

Source	Destination	Validates
App-VM1	App-VM2	Connectivity within the cluster on same segment.
App-VM1	App-VM3	Connectivity between clusters on the same segment.
App-VM1	Web-VM1	Connectivity within the cluster on different segments.
Web-VM1	App-VM3	Connectivity between clusters on different segments.

Note: The 2nd test in Table 9 (App-VM1 to App-VM3) is a good example of virtual layer 2 (switched) traffic over a physical layer 3 (routed) network. Both VMs are on the 10.10.20.0/24 virtual network but they are physically located on ESXi hosts in different racks. Therefore, traffic between these VMs is routed through the spine switches in the physical network.

13 Communicate outside the virtual network

In most cases, some virtual machines on the NSX network need to communicate with machines on external traditional networks. Two devices designed to handle this traffic are Edge Services Gateways (ESGs) and hardware VTEPs.

13.1 Edge Services Gateway

An ESG is an NSX virtual appliance similar to a DLR. Dell EMC recommends using an ESG to handle north-south traffic between the data center's virtual network and the WAN or network core. This allows the administrator to take advantage of additional features provided by the ESG, such as load balancing and VPN services.

The physical topology for the edge cluster is shown in Figure 68. The edge cluster contains the DLR and ESG virtual appliances.

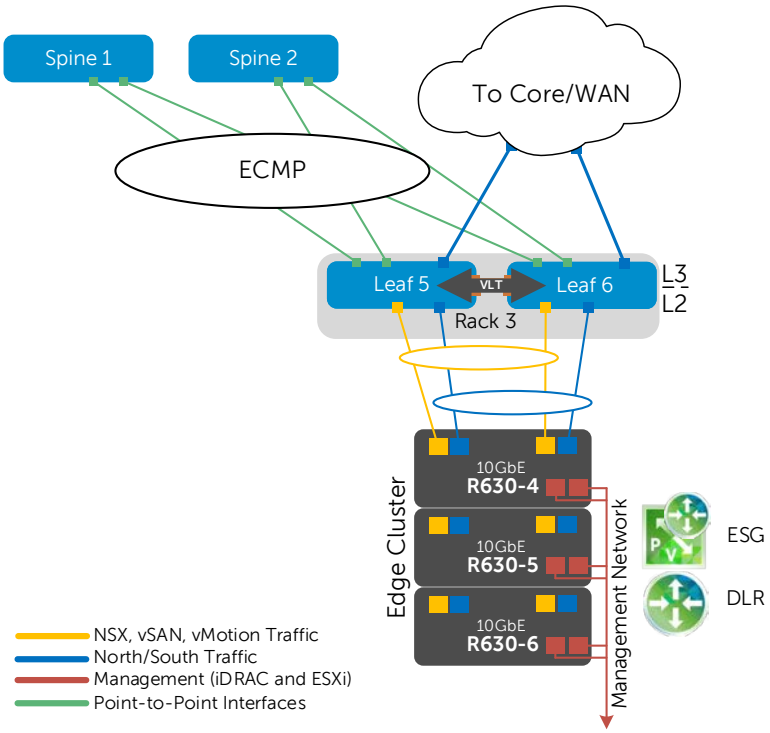


Figure 68 Rack 3 Edge Cluster physical topology

In Figure 68, the yellow connections leading into the edge cluster represent two 10GbE NICs per host, with each pair configured as a port channel. These connections handle NSX, VSAN, and vMotion traffic within the leaf-spine network.

The blue links are in a separate port channel that handles north and southbound traffic. Leaf switches 5 and 6 use OSPF to create an adjacency with the ESG virtual machine. Leaf switch edge configuration was covered in Section 6.3.1.

13.1.1 Add a distributed port group

Before deploying an ESG, an additional, VLAN-backed, distributed port group needs to be created on the edge cluster VDS. This additional port group handles all north and southbound traffic between the NSX environment and the network core or WAN.

Note: VLAN 66 was configured on Leaf 5 and Leaf 6 in Section 6.3.1.

To create the port group:

1. In the web client, go to **Home > Networking**.
2. Right click on Rack 3 Edge VDS. Select **Distributed Port Group > New Distributed Port Group**.
3. In the **New Distributed Port Group** wizard:
   a. Provide the name **R3 Edge VLAN 66** and click **Next.**
   b. On the **Configure Settings** page, set **VLAN type** to **VLAN** and set the **VLAN ID** to **66**.
   c. Click **Next > Finish**.

This creates the port group named R3 Edge VLAN 66.

13.1.2 Create second LACP LAG

On the Rack 3 Edge VDS, two LAGs are needed as shown in Figure 68 above. One is for traffic within the data center shown in yellow (**lag1,** created earlier in Section 9.3), and one for edge traffic to the WAN/network Core shown in blue (**lag2-edge,** to be created here).

To configure the edge LAG on **Rack 3 Edge VDS:**

1. Go to **Home > Networking**.
2. In the **Navigator** pane, select Rack 3 Edge VDS.
3. In the center pane, select **Manage > Settings > LACP.**
4. On the **LACP** page, click the + icon. The **New Link Aggregation Group** (LAG) dialog box opens.
5. Set the **Name** to **lag2-edge**.
6. Set the **Number of ports** equal to the number of physical uplinks in the LAG on each ESXi host. In this deployment, R630 hosts use two links for the edge LAG so this number is set to **2**.
7. Set the **Mode** to **Active**. The remaining fields can be set to their default values as shown in Figure 69.

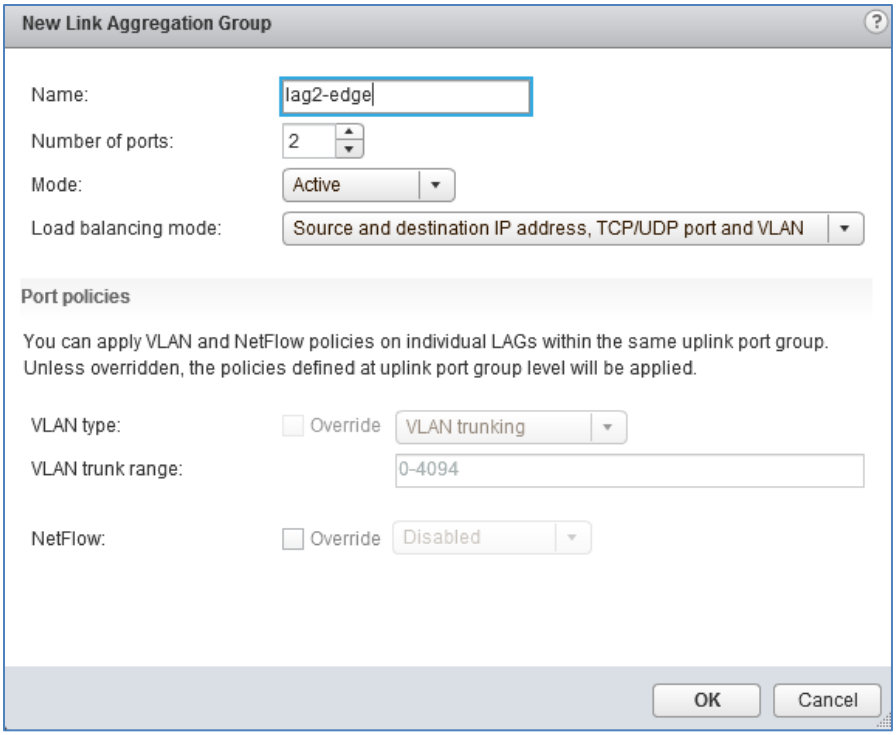


Figure 69 Lag2-edge configuration

8. Click **OK** to close the dialog box. This creates **lag2-edge** on the VDS.

The refresh icon ( ) at the top of the screen may need to be clicked for the lag to appear in the table as shown in Figure 70.

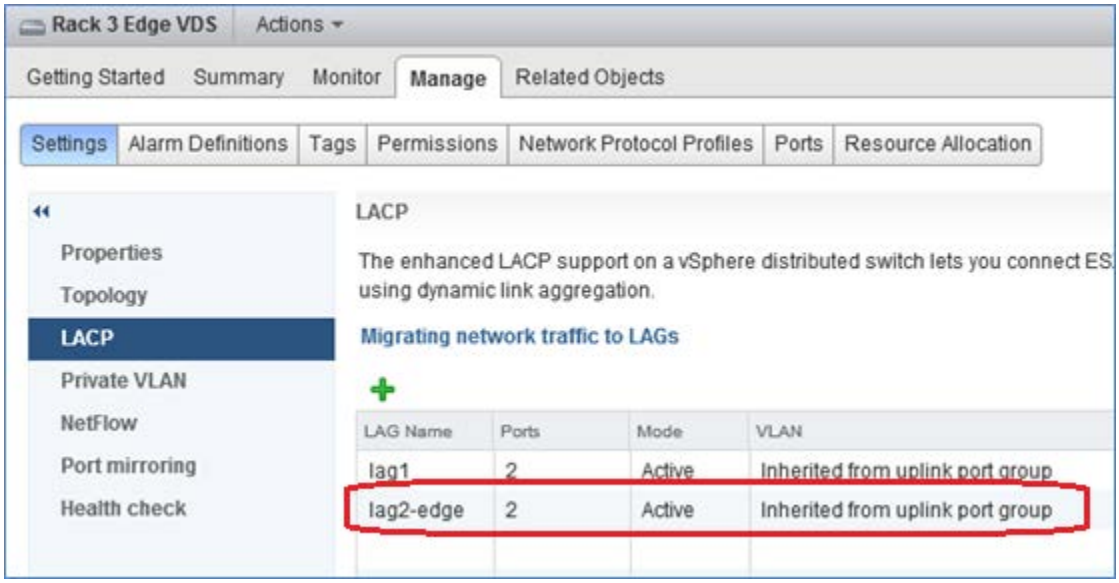


Figure 70 Lag2-edge created on Rack 3 Edge VDS

13.1.3 Assign uplinks to the second LAG

Note: Before starting this section, be sure you know the vmnic-to-physical adapter mapping for each host. This can be determined by going to Home > Hosts and Clusters and selecting the host in the Navigator pane. In the center pane select Manage > Networking > Physical adapters. These are the vmnics connected to port channels 12, 14 and 16 on Leaf Switches 5 and 6.

To assign uplinks to lag2-edge:

1. Go to **Home > Networking**.
2. Right click on Rack 3 Edge VDS, and select **Add and Manage Hosts**.
3. In the Add and Manage Hosts dialog box:
   a. On the **Select task** page, make sure **Manage host networking** is selected. Click **Next**.
   b. On the **Select hosts** page, click the **Attached hosts** icon. Select all hosts in the Rack 3 Edge cluster. Click **OK > Next**.
   c. On the **Select network adapters tasks** page, be sure the **Manage physical adapters** box is checked. Be sure all other boxes are unchecked. Click **Next**.
   d. On the **Manage physical network adapters** page, each host is listed with its vmnics beneath it.
      i. Select the first vmnic on the first host and click Assign uplink.
      ii. Select **lag2-edge-0 > OK**.
      iii. Select the second vmnic on the first host and click Assign uplink.
      iv. Select **lag2-edge-1 > OK**.
   e. Repeat steps i – iv for the remaining hosts. Click **Next** when done.
   f. On the **Analyze impact** page, **Overall impact status** should indicate No impact.
   g. Click **Next > Finish**.

When complete, the **Manage > Settings > Topology** page for **Rack 3 Edge VDS** should look similar to Figure 71.

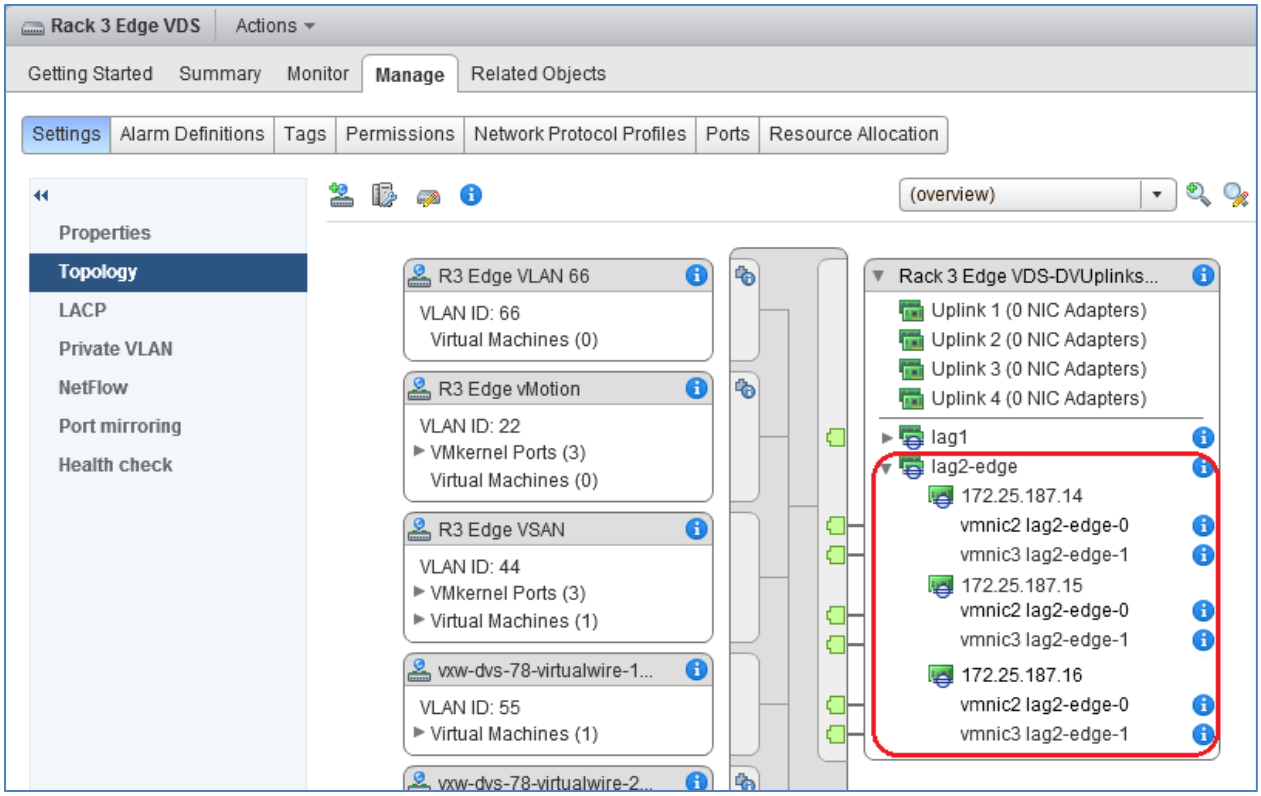


Figure 71 Lag2-edge configured on Rack 3 Edge VDS

This configuration brings up the edge LAGs (port channels 12, 14 and 16) on the upstream switches. This can be confirmed by running the `show vlt detail` command on the upstream switches (Leaf 5 and Leaf 6).

The Local and Peer Status columns indicate UP for all port channels.

```
Leaf-5#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
2             2            UP            UP            1, 22, 44, 55
4             4            UP            UP            1, 22, 44, 55
6             6            UP            UP            1, 22, 44, 55
12            12           UP            UP            1, 66
14            14           UP            UP            1, 66
16            16           UP            UP            1, 66
```

13.1.4 Configure port groups for teaming and failover

1. On the web client **Home** screen, select **Networking**.
2. Right click on Rack 3 Edge VDS. Select **Distributed Port Group > Manage Distributed Port Groups**.
3. Select the **Teaming and failover** checkbox. Click **Next**.
4. Click **Select distributed port groups**.
5. Check the box next to the **R3 Edge VLAN 66** port group. Click **OK > Next.**
6. On the **Teaming and failover** page, move **lag2-edge** up to the **Active uplinks** section. Move Uplinks 1-4 down to the **Unused uplinks** section as shown in Figure 72.

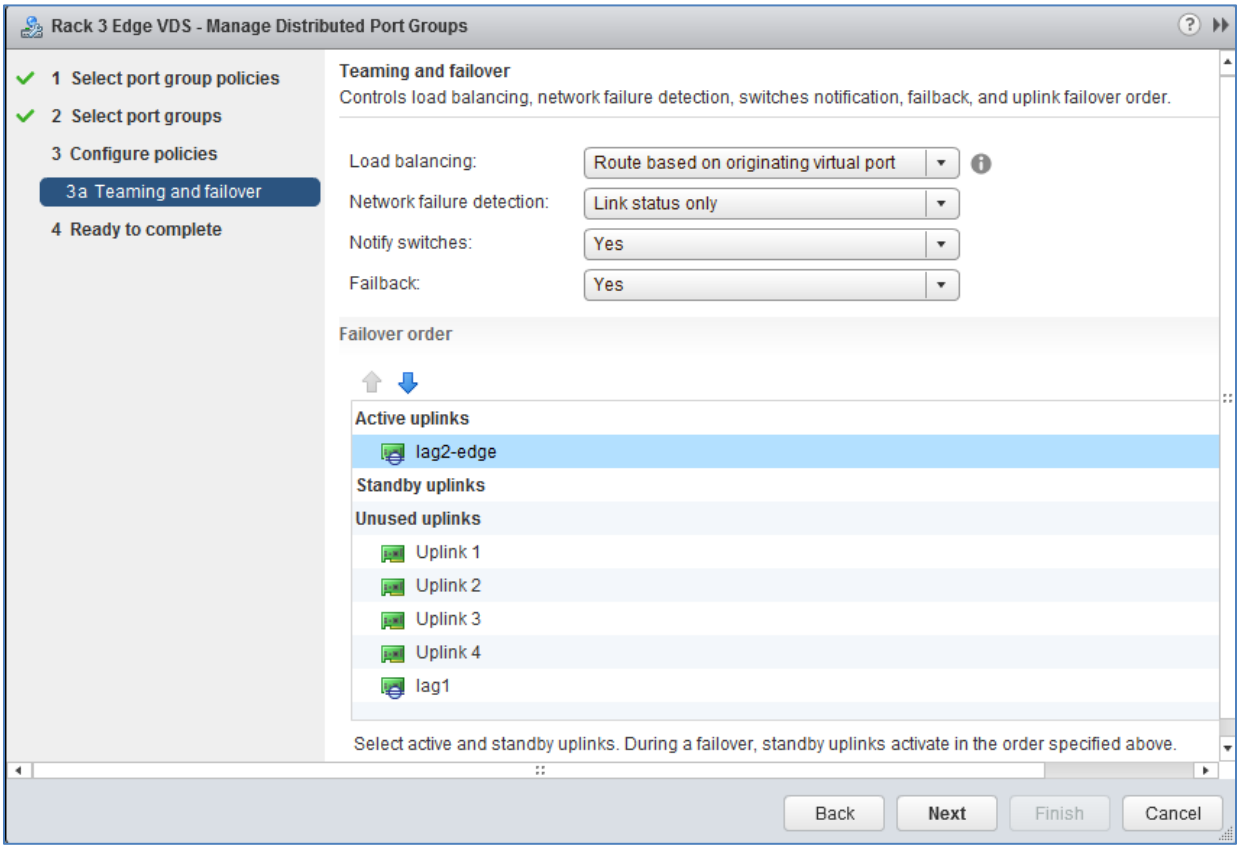


Figure 72 Rack 3 Edge VDS teaming and failover settings

7. Click **Next > Finish**.

13.1.5 Deploy the Edge Services Gateway

Now that layer 2 connectivity between the edge cluster and the two leaf switches has been established, the ESG appliance can be added and configured.

To deploy the ESG:

1. Go to **Home > Networking & Security > NSX Edges** and click the icon.
2. In the **New NSX Edge** dialog box:
   a. Select **Edge Services Gateway** and name it **ESG**.
   b. Verify the **Deploy NSX Edge** box is checked and click **Next.**
   c. Provide CLI credentials for the ESG, leave other settings at defaults, and click **Next.**
   d. Next to **Appliance Size,** select a size. **Large** is selected for this guide.

Note: See System Requirements for NSX for ESG sizing specifications.

   e. Click the icon to create an edge appliance
   f. In the **Add NSX Edge Appliance** dialog box:
      i. Set **Cluster/Resource Pool** to **Rack 3 Edge.**
      ii. Set Datastore to **Rack 3 Edge VSAN** as shown in Figure 73 and leave the **Host** and **Folder** boxes blank. Click **OK > Next**.

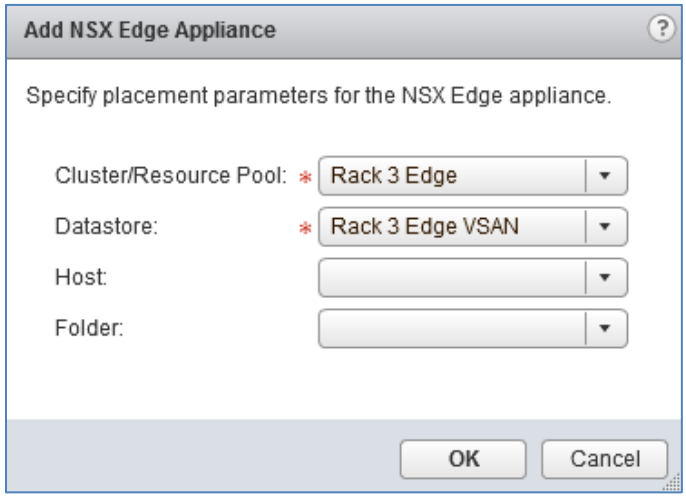


Figure 73 Add NSX edge appliance dialog box

   g. On the **Configure Interfaces** page, click the icon to open the **Add NSX Edge Interface** dialog box.
      i. **Name** the interface **Edge VLAN 66**
      ii. Set **Type** to **Uplink**
      iii. Next to **Connected To,** click **Select**.
      iv. Be sure **Distributed Portgroup** is selected at the top and select **R3 Edge VLAN 66.** Click **OK.**

v. Click the icon above **Primary IP Address**.
vi. Enter **10.66.3.1** for the **Primary IP Address** and set the **Subnet Prefix Length** to **24.**
vii. Leave the remaining values at their defaults and click **OK** to close.

h. Click the icon to open the **Add NSX Edge Interface** dialog box.
i. **Name** the interface **Transit Network.**
ii. Set **Type** to **Internal**
iii. Next to **Connected To,** click **Select**.
iv. Be sure **Logical Switch** is selected and select **Transit Network (5000).** Click **OK.**
v. Click the icon above **Primary IP Address**.
vi. Type **172.16.0.1** for the **Primary IP Address** and set the **Subnet Prefix Length** to **24.**
vii. Leave the remaining values at their defaults and click **OK > Next.**

i. Uncheck **Configure Default Gateway** and click **Next**.
j. Leave **Configure Firewall default policy** unchecked and click **Next.**
k. Click **Finish** to deploy the ESG. It may take a few minutes to complete.

To validate ESG settings:

1. Go to **Home > Networking & Security > NSX Edges**.
2. In the center pane, double click on the ESG to open the ESG summary and management page.
3. Select **Manage > Settings > Interfaces**. Verify settings are correct as shown in Figure 74.

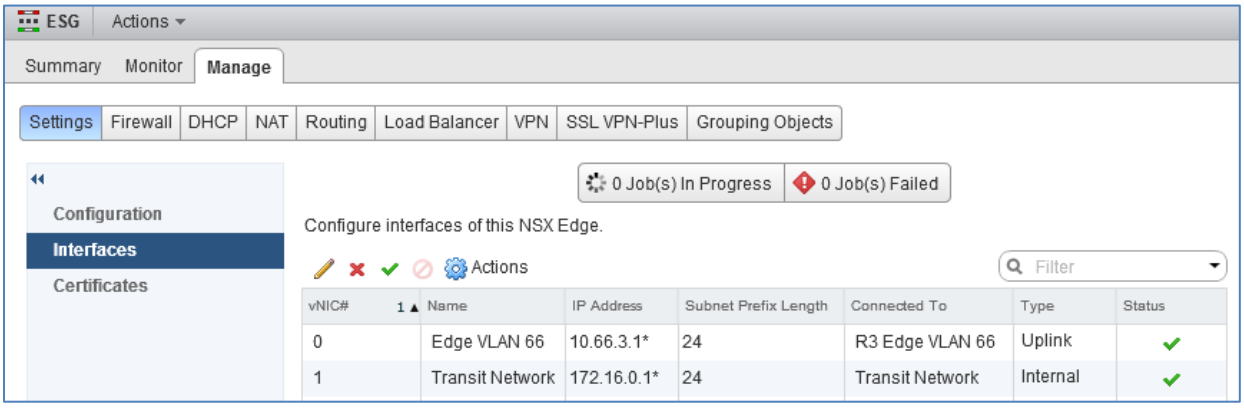


Figure 74 Configured interfaces on the ESG

13.1.6 Configure OSPF on the ESG

Configuring OSPF on the ESG enables the ESG to learn and advertise routes from the core network/WAN upstream. This deployment defines two tasks for OSPF participation. The first task defines Area 0 and creates route adjacencies to the two leaf switches, Leaf 5 and Leaf 6. The second task adds the NSX default Area 51 for use between the ESG and DLR.

Configure the Router ID:

1. Go to **Home > Networking & Security > NSX Edges**.
2. In the center pane, double click on the ESG to open the ESG summary and management page.
3. Select **Manage > Routing > Global Configuration**. Next to **Dynamic Routing Configuration** click **Edit.**
4. Next to **Router ID**, keep the default, **Edge VLAN 66 – 10.66.3.1,** and click **OK.**
5. Click **Publish Changes** near the top of the screen

Enable OSPF:

1. On the **Routing** page, select **OSPF.**
2. Next to **OPSF Configuration,** click **Edit**.
3. Check the **Enable OSPF** box and leave the **Enable Grateful Restart** box checked. Click **OK**.
4. Click **Publish Changes.**

Enable interfaces to participate in their respective OSPF areas:

1. Click the  icon under **Area to Interface Mapping**.
2. Next to **vNIC**, select **Edge VLAN 66**.
3. Set the **Area** to **0**.
4. Leave all other values at their defaults and click **OK.**
5. Click the  icon under **Area to Interface Mapping**.
6. Next to **vNIC**, select **Transit Network**.
7. Set the **Area** to **51** (default).
8. Leave all other values at their defaults and click **OK.**
9. Click **Publish Changes**.

When complete, the OSPF page for the ESG should be similar to Figure 75. Two interfaces are mapped to two separate OSPF areas. The external interface is mapped to Area 0 and the internal interface is mapped to Area 51.

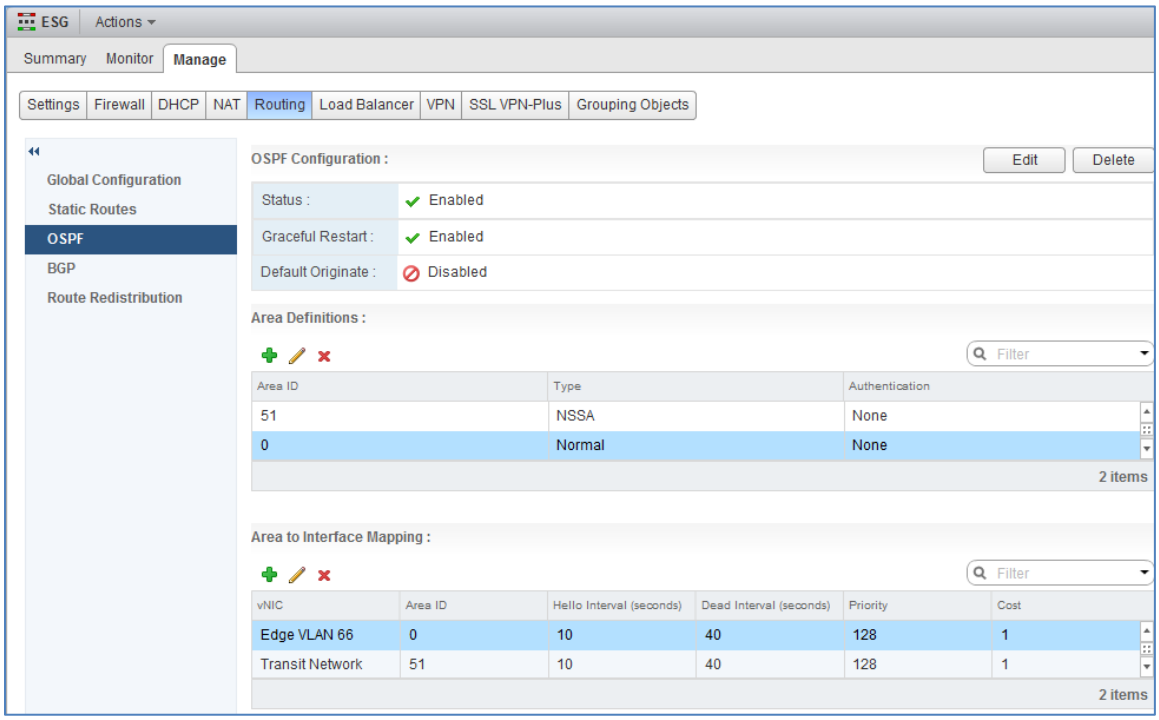


Figure 75 ESG OSPF configuration complete

At this point, all OSPF area 0 adjacencies are established. Run the command `show ip ospf neighbor` on leaf switches 5 and 6 to validate this.

```
Leaf-5#show ip ospf neighbor

Neighbor ID     Pri   State           Dead Time Address         Interface   Area
10.66.3.1       128   FULL/DR         00:00:38  10.66.3.1       Vl 66       0
10.66.3.253     1     FULL/DROTHER    00:00:30  10.66.3.253     Vl 66       0
```

13.1.7 High Availability Configuration

Enabling HA deploys a backup copy of the ESG (as an additional VM) to another host in the Edge Cluster to act as a standby ESG. The standby provides backup in case of a failure with the active ESG.

To enable HA for the ESG:

1. Go to **Home > Networking & Security > NSX Edges**.
2. In the center pane, double click on the ESG to open the ESG summary and management page.
3. Select **Manage > Settings > Configuration**.
4. Next to **HA Configuration**, click **Change**.

a. In the **Change HA configuration** box, set **HA Status** to **Enable.**
b. Set the **vNIC** to **Transit Network.**
c. Leave the remaining values at their defaults and click **OK.**

After a few minutes, the standby ESG is deployed. The ESG **Configuration** page appears similar to Figure 76.

Under **NSX Edge Appliances**, ESG-0 and ESG-1 are shown with ESG-0 active.

Note: It may take several minutes for (Active) to appear next to the active ESG.

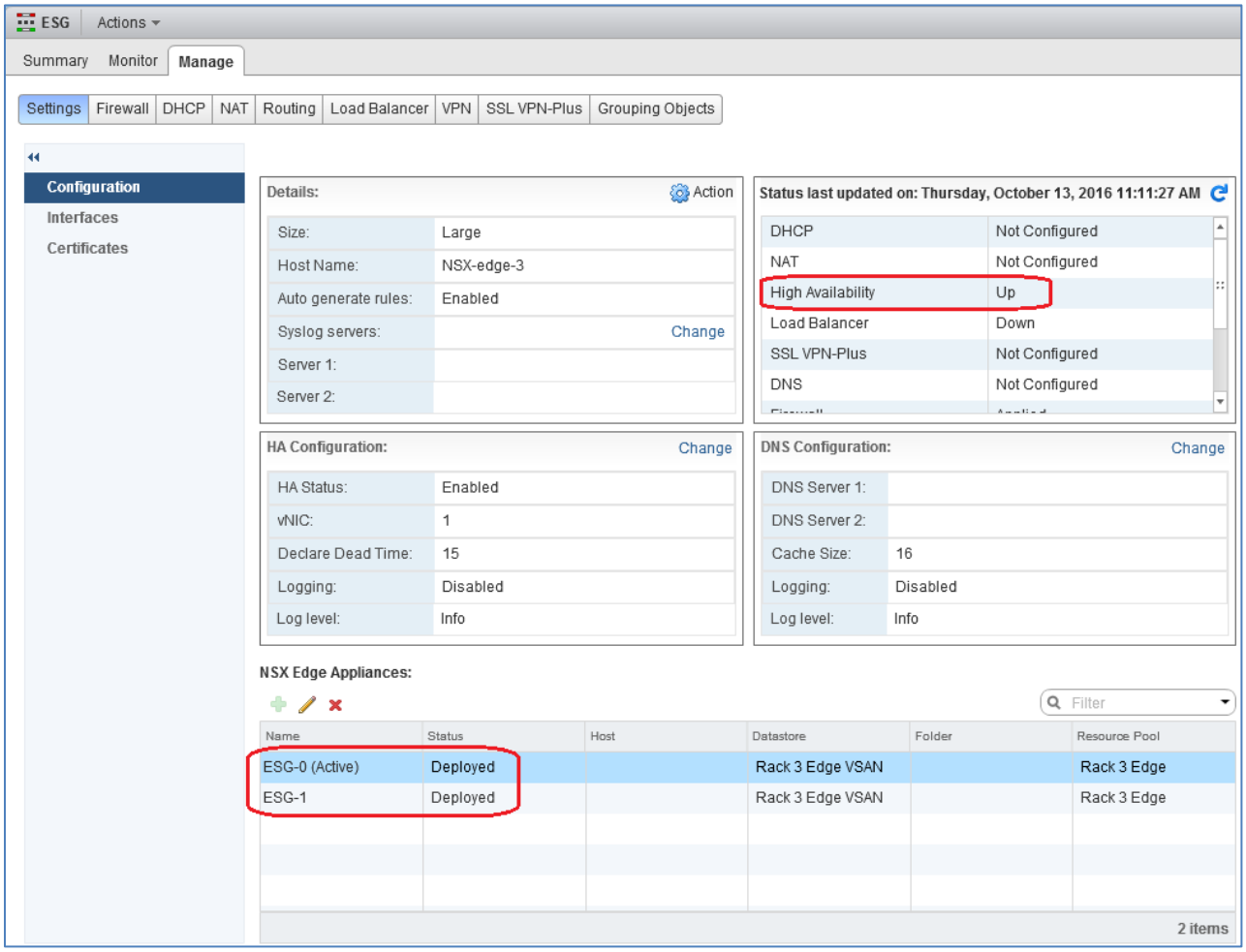


Figure 76 ESG high availability enabled

North-south access has been established using OSPF as the dynamic routing protocol. Figure 77 illustrates this configuration.

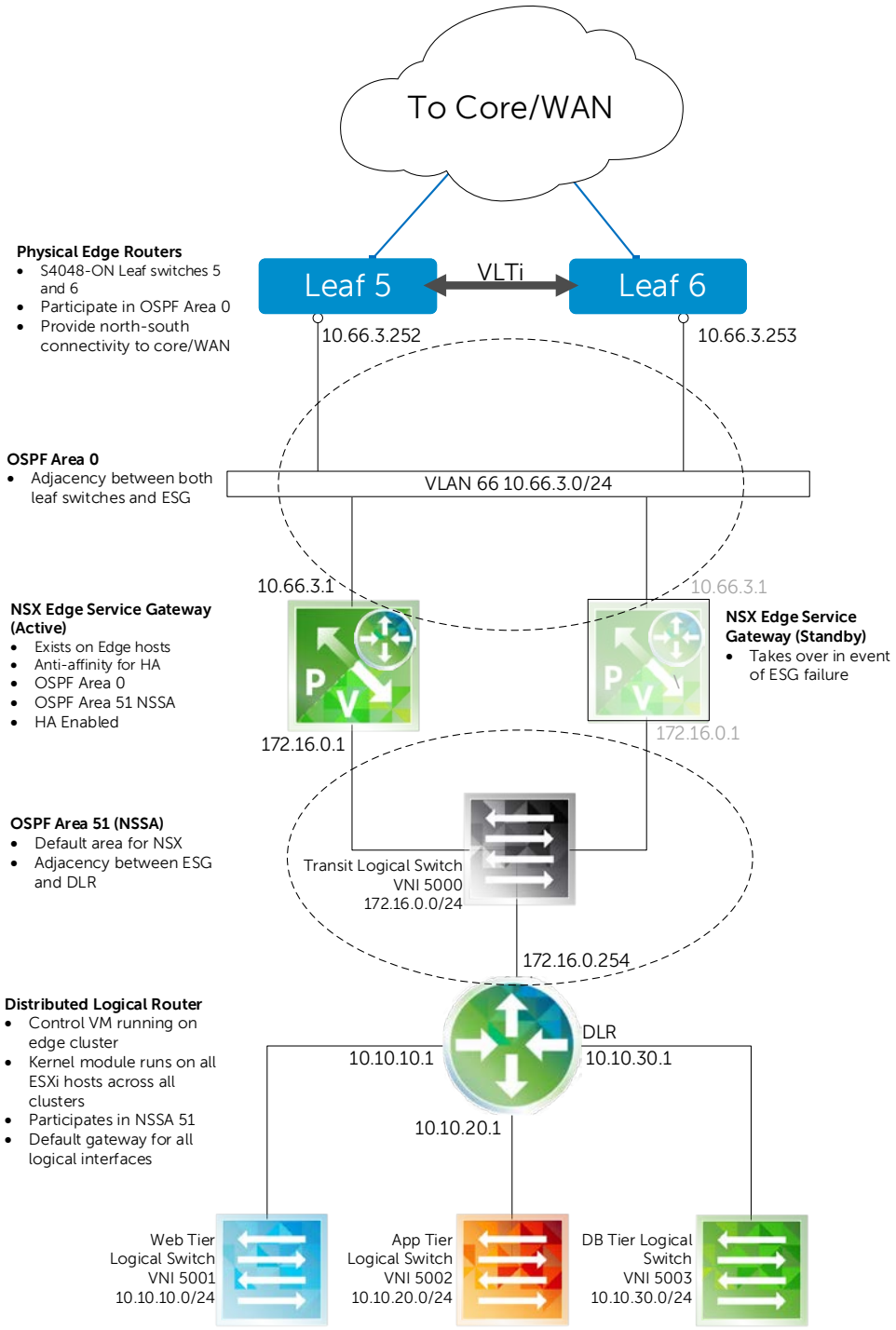


Figure 77 Logical overview of NSX edge

Note: Extending leaf switches 5 and 6 to the network core or WAN is outside the scope of this document.

13.1.8 ESG Validation

13.1.8.1 Commands and output

Access the ESG console by going to **Hosts & Clusters,** right clicking on the active ESG VM (**ESG-0**), and selecting **Open Console**. Login using the credentials specified when the ESG was created in Section 13.1.5 (default user name is **admin**).

Basic troubleshooting commands and output run from the ESG CLI are as follows:

```
NSX-edge-3-0> traceroute 10.10.10.11    ← IP address of Web-VM1
traceroute to 10.10.10.11 (10.10.10.11), 30 hops max, 60 byte packets
 1  172.16.0.254 (172.16.0.254)  0.059 ms  1002.063 ms  1002.069 ms
 2  10.10.10.11 (10.10.10.11)  0.534 ms * *

NSX-edge-3-0> show ip route
Codes: O - OSPF derived, i - IS-IS derived, B - BGP derived,
C - connected, S - static, L1 - IS-IS level-1, L2 - IS-IS level-2,
IA - OSPF inter area, E1 - OSPF external type 1, E2 - OSPF external type 2,
N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2

Total number of routes: 7

O   N2  10.10.10.0/24        [110/1]       via 172.16.0.254    ←Web-Tier Network
O   N2  10.10.20.0/24        [110/1]       via 172.16.0.254    ←App-Tier Network
O   N2  10.10.30.0/24        [110/1]       via 172.16.0.254    ←DB-Tier Network
C       10.66.3.0/24         [0/0]         via 10.66.3.1
C       169.254.1.0/30       [0/0]         via 169.254.1.1
C       172.16.0.0/24        [0/0]         via 172.16.0.1

NSX-edge-3-0> show ip ospf neighbors
Neighbor ID       Priority  Address         Dead Time  State           Interface
10.66.3.252 ←Leaf 5   1     10.66.3.252     31         Full/BDR        vNic_0
10.66.3.253 ←Leaf 6   1     10.66.3.253     34         Full/DROTHER    vNic_0
172.16.0.254 ←DLR     128   172.16.0.253    39         Full/DR         vNic_1
```

13.1.8.2 Traffic test

To validate functionality of the ESG, send traffic between a system on the network core/WAN network and the VMs on the NSX network.

For a simplified test, a compute node (running a Windows OS in this example) with an IP address 8.0.0.1/24 and gateway set to 8.0.0.2 is directly connected to Leaf 6, port tengigabitethernet 1/48. The following configuration is added to Leaf 6:

```
interface TenGigabitEthernet 1/48
description To Compute Node
ip address 8.0.0.2/24
```

```
no shutdown
exit

router ospf 1
network 8.0.0.0/24 area 0
```

Provided the compute node, ESG, and VM firewalls are properly configured, the VMs can now be pinged from the compute node connected to the leaf switch.

Note: The ESG Firewall denies external traffic by default and must be configured or temporarily disabled for traffic to pass. Access the ESG firewall by going to **Home > Networking & Security > NSX Edges**. Double-click on the ESG and go to **Manage > Firewall**.

The compute node at 8.0.0.1 pings Web-VM1 at 10.10.10.11 as follows:

```
C:\Windows\system32>ping 10.10.10.11

Pinging 10.10.10.11 with 32 bytes of data:
Reply from 10.10.10.11: bytes=32 time<1ms TTL=124
Reply from 10.10.10.11: bytes=32 time<1ms TTL=124
```

A trace route command issued from the compute node at 8.0.0.1 to Web-VM1 at 10.10.10.11 returns the following:

```
C:\Windows\system32>tracert 10.10.10.11

Tracing route to WIN-U3U892VR1IJ [10.10.10.11]
over a maximum of 30 hops:

  1    <1 ms    <1 ms    <1 ms  8.0.0.2                      ←Leaf 6
  2    <1 ms    <1 ms    <1 ms  10.66.3.1                    ←ESG
  3    <1 ms    <1 ms    <1 ms  172.16.0.254                 ←DLR
  4     1 ms    <1 ms    <1 ms  WIN-U3U892VR1IJ [10.10.10.11] ←Web-VM1

Trace complete.
```

13.2 Hardware VTEP

For virtual-to-physical network traffic within the data center (east-west traffic), a hardware VTEP provides the best performance. It is not considered a best practice to use an ESG for east-west traffic within the data center's leaf-spine network; it can become a bottleneck under heavy loads.

Note: The hardware VTEP feature requires VMware NSX Enterprise Edition.

The switch acting as the hardware VTEP must support VXLAN, such as Dell Networking S4048-ON or S6000-ON switches. This guide uses an S4048-ON.

The hardware VTEP connects upstream to the same two spine switches used in the NSX network. This enables communication between the virtual and physical networks. The added hardware VTEP and physical server are outlined in red in Figure 78:

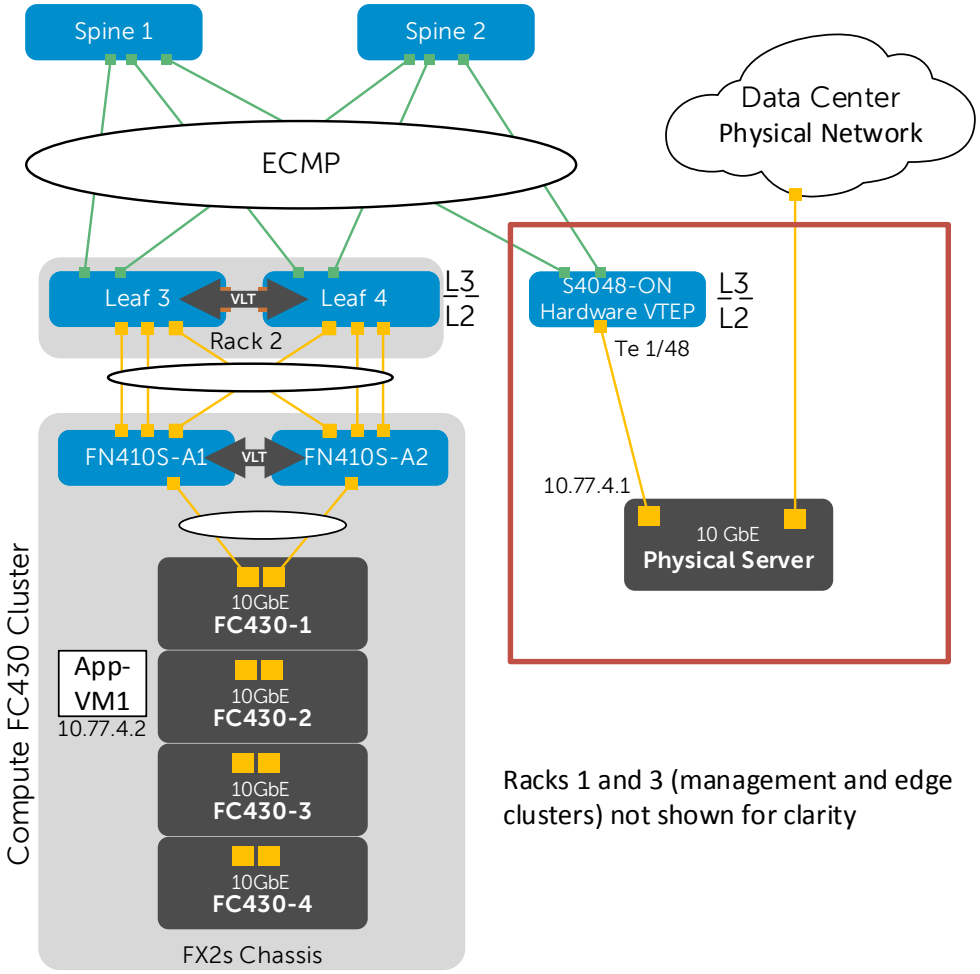


Figure 78 Hardware VTEP and physical server location in leaf-spine network

Note: Interface tengigabitethernet 1/48 is connected to a single server in this example. Any available interfaces on the hardware VTEP may be configured and connected to physical servers as needed.

13.2.1 Configure additional connections on spine switches

Note: These configuration steps are in addition to the spine switch configurations provided in the attachments named spine1.txt and spine2.txt.

On each spine switch, BGP and an additional interface, fortyGigE 1/7/1, are configured to connect to the hardware VTEP as shown in the following two sections.

13.2.1.1 Spine 1 – additional configuration steps

```
enable
configure

stack-unit 1 port 7 portmode single speed 40G no-confirm

interface fortyGigE 1/7/1
description To HW VTEP fo1/49
ip address 192.168.1.12/31
mtu 9216
no shutdown

router bgp 64601
neighbor 192.168.1.13 remote-as 64707
neighbor 192.168.1.13 peer-group spine-leaf
neighbor 192.168.1.13 no shutdown

ecmp-group 1
interface fortyGigE 1/7/1

end
write
```

13.2.1.2 Spine 2 – additional configuration steps

```
enable
configure

stack-unit 1 port 7 portmode single speed 40G no-confirm

interface fortyGigE 1/7/1
description To HW VTEP fo1/50
ip address 192.168.2.12/31
mtu 9216
no shutdown

router bgp 64602
neighbor 192.168.2.13 remote-as 64707
```

```
neighbor 192.168.2.13 peer-group spine-leaf
neighbor 192.168.2.13 no shutdown

ecmp-group 1
interface fortyGigE 1/7/1

end
write
```

13.2.2 Configure the hardware VTEP and connect to NSX

Note: The S4048-ON starts at its factory default settings. To reset to factory defaults, see Section 6.1. The switch configuration is provided in the hw-vtep.txt attachment.

Initial configuration involves setting the hostname, enabling LLDP and configuring the management interface and default gateway as follows:

```
enable
configure
hostname HW-VTEP
protocol lldp
advertise management-tlv management-address system-description system-name
advertise interface-port-desc

interface ManagementEthernet 1/1
ip address 172.25.187.36/24
no shutdown
management route 0.0.0.0/0 172.25.187.254
```

Next, configure the upstream layer 3 interfaces connected to the spines. Configure a loopback interface as the router ID for BGP. Complete these actions as follows:

```
interface fortyGigE 1/49
description To Spine-1
ip address 192.168.1.13/31
mtu 9216
no shutdown

interface fortyGigE 1/50
description To Spine-2
ip address 192.168.2.13/31
mtu 9216
no shutdown

interface loopback 0
description Router ID
ip address 10.0.2.7/32
```

Enable the BGP processes to allow routing to the IP fabric. Additionally, create an IP prefix and route map to automatically redistribute all leaf subnets and loopback addresses from the spine and leaf switches as follows:

```
route-map spine-leaf permit 10
match ip address spine-leaf

ip prefix-list spine-leaf
description BGP redistribute loopback and leaf networks
seq 5 permit 10.0.0.0/23 ge 32
seq 10 permit 10.0.0.0/8 ge 24
router bgp 64707
bgp bestpath as-path multipath-relax
maximum-paths ebgp 64
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 192.168.1.12 remote-as 64601
neighbor 192.168.1.12 peer-group spine-leaf
neighbor 192.168.1.12 no shutdown
neighbor 192.168.2.12 remote-as 64602
neighbor 192.168.2.12 peer-group spine-leaf
neighbor 192.168.2.12 no shutdown
```

Create an ECMP group and include the interfaces to the two spine switches as follows:

```
ecmp-group 1
interface fortyGigE 1/49
interface fortyGigE 1/50
link-bundle-monitor enable
```

Enable the VXLAN feature and BFD. Create a loopback interface and assign an address to be used as the HW VTEP address as follows:

```
feature vxlan
bfd enable

interface Loopback 77
ip address 10.77.4.254/32
no shutdown
```

Create a VXLAN instance. The gateway address is the hardware VTEP address (same address as loopback 77 above). For controller 1, use the IP address of NSX Controller 1.

Note: NSX controller addresses can be determined in the web client by going to Home > Networking & Security > Installation > Management.

```
vxlan-instance 1
gateway-ip 10.77.4.254
fail-mode secure
controller 1 172.25.187.183 port 6640 ssl
no shutdown
```

Configure an interface connected to a physical server and place it in the VXLAN instance as follows:

```
interface te 1/48
description To Physical Server
vxlan-instance 1
no shutdown

end
write
```

Create a secure management connection between the S4048-ON and VMware NSX by generating a certificate on the switch:

```
HW-VTEP#crypto cert generate cert-file flash://vtep-cert.pem key-file
flash://vtep-privkey.pem
Generating self signed certificate. This might take a few minutes.
! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! !
! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! !
Certificate generated successfully.
```

View the certificate from the CLI by running the following command:

```
S4048#show file flash://vtep-cert.pem
-----BEGIN CERTIFICATE-----
MIIDmTCCAoGgAwIBAgICAKswDQYJKoZIhvcNAQEFBQAwfjELMAkGA1UEBhMCVVMx
HjAcBgNVBAMMFVIxODdVMzAtUzQwNDgtUjMtVG9SMjENMAsGA1UECgwERGVsbDEY
MBYGA1UECwwPRGVsbCBOZXR3b3JraW5nMREwDwYDVQQHDAhTQU4gSm9zZTETMBEG
A1UECAwKQ2FsaWZvcm5pYTAeFw0xNjA5MDMxNzM0MzhaFw0yNjA5MDExNzM0Mzha
MH4xCzAJBgNVBAYTAlVTMR4wHAYDVQQDDBVSMTg3VTMwLVM0MDQ4LVIzLVRvUjIx
DTALBgNVBAoMBERlbGwxGDAWBgNVBAsMD0RlbGwgTmV0d29ya2luZzERMA8GA1UE
BwwIU0FOIEpvc2UxEzARBgNVBAgMCkNhbGlmb3JuaWEwggEiMA0GCSqGSIb3DQEB
AQUAA4IBDwAwggEKAoIBAQC+DF9S0vHVbUv0ZuxY5rO8nEqxXiUXYmJCyzhlW06I
LHYs3UBO/dFAqdxPh8ddRNL0zGXNoAYTUlQ6YeIou46xKgriWLCAWlCbk2QluiVn
5DeuvmbDt4JCssSzUj5jCCeX7sdjy3CVhzmL+pHUy1+FDDlyVi9cs5KapOqHHRDI
MDt0ZCFp9q8hdpmT6xfMtD2/Ml7DaUrmymGNNWh3xt+YewkYOBvJuydR2czUosRy
qCUhylIRcB+RhFlsdD9kKTqIJNqE7ouxG9Ona94+KuofOVFcZho3dHSpIUv1fhNz
Q307EfJIIpFufsxdRZWfhrOMtpJka9Qxp1GWCVjrPnkPAgMBAAGjITAfMB0GA1Ud
DgQWBBTaOaPuXmtLDTJVv++VYBiQr9gHCTANBgkqhkiG9w0BAQUFAAOCAQEAO9GD
DfipIcOfi+/L01lV63x6eXuVaLp1SPAyrgAIxFbepj7sHWWGe2UZixmEEhSmY9of
+AkqXY4sC1C4GQER4EaokF3FW0j/n35onNKeXY/78CkJx+lp60qD0DATe7/L5Ew+
OGhXDVCrI5y2Cw4sZ3gYoVXgQasAv8QGRU7tBV67ezdTu6Ur55DlOf+PpHpY8Dl1
fGjdTbzpN90HpZyFDoBaQbXgpuW9riVS8xNUj/M4sPTm81p4GdqgivilMo4CR/6Z
i1whD2I++gm+VJa0xVnUf1nmjKnpX/YHxpdHvsVxs1ZLdh7uMfGhEsTsrvpxqZTX
9IEsmi3+H+O4FW2FGQ==
-----END CERTIFICATE-----
```

Copy the certificate output above including the BEGIN and END CERTIFICATE statements.

1. In the web client, go to **Home > Networking & Security > Service Definitions > Hardware Devices** and click the + icon.
2. In the **Edit Hardware Device** box:
   a. Next to **Name**, enter **Hardware VTEP**.
   b. In the **Certificate** box, paste the certificate output from the S4048-ON as shown in Figure 79.

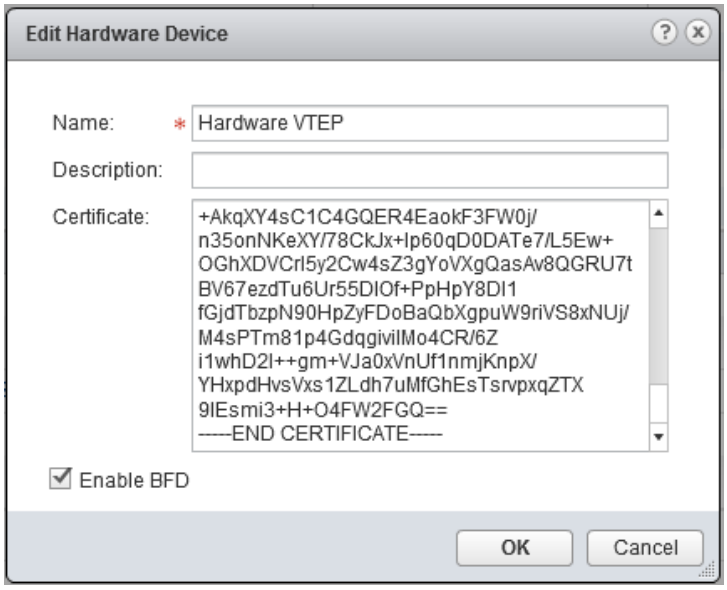


Figure 79 Creating a Hardware Device in VMware NSX

c. Leave the **Enable BFD** box checked and click **OK**.

After a minute or two, the following output is logged on the S4048-ON:

```
Oct 14 19:08:34: %STKUNIT1-M:CP %OVSDBSVR-5-SESSION_CONNECTED: Instance 1
session 172.25.187.185 is connected
Oct 14 19:08:04: %STKUNIT1-M:CP %OVSDBSVR-5-SESSION_CONNECTED: Instance 1
session 172.25.187.183 is connected
Oct 14 19:08:03: %STKUNIT1-M:CP %OVSDBSVR-5-SESSION_CONNECTED: Instance 1
session 172.25.187.184 is connected
```

This confirms that the hardware VTEP is connected to NSX and an Open vSwitch Database (OVSDB) session is established.

Note: The following optional debug command can be issued to view additional VXLAN connection information:
`S4048#debug vxlan ovsdb-json-rpc packet-type all vxlan-instance 1`

To confirm that the hardware device has been added to NSX and has proper connectivity, refresh the **Hardware Devices** screen in the web client by clicking the refresh icon ( ).

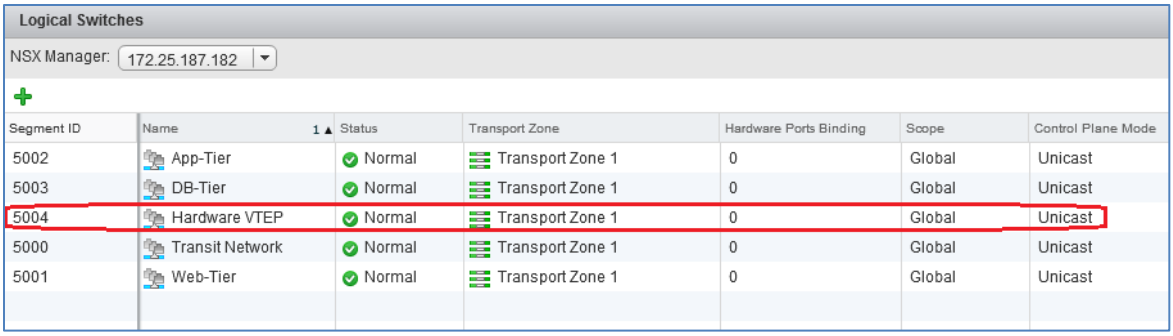
The screen should appear similar to Figure 80, with **Connectivity Up** and a **green checkmark** under **BFD Enabled**. The **Management IP Address** shown is the management address of the S4048-ON.

Figure 80 Hardware Device Status

13.2.3 Create a logical switch

Create a logical switch as follows:

1. In the web client, go to **Home > Networking & Security > Logical Switches**.
2. Click the icon to add a new logical switch.
3. In the **New Logical Switch** dialog box:
   a. For **Name**, enter **Hardware VTEP.**
   b. Next to **Transport Zone**, **Transport Zone 1** should already be selected. If not, click **Change** and select it.
   c. Leave the **Replication mode** set to **Unicast, Enable IP Discovery** checked and **Enable MAC Learning** unchecked**.**
   d. Click **OK**. A new logical switch named Hardware VTEP is created as shown in Figure 81:



Segment ID	Name	Status	Transport Zone	Hardware Ports Binding	Scope	Control Plane Mode
5002	App-Tier	Normal	Transport Zone 1	0	Global	Unicast
5003	DB-Tier	Normal	Transport Zone 1	0	Global	Unicast
5004	Hardware VTEP	Normal	Transport Zone 1	0	Global	Unicast
5000	Transit Network	Normal	Transport Zone 1	0	Global	Unicast
5001	Web-Tier	Normal	Transport Zone 1	0	Global	Unicast

Figure 81 Hardware VTEP logical switch created

4. Select the new logical switch, Hardware VTEP, and select **Actions > Manage Hardware Bindings**.
   a. Expand **Hardware VTEP (0 Bindings)** and click the icon. The IP address of the S4048 is automatically filled out (172.25.187.36 in this example).
   b. In the **Port** column, click **Select**. Ports that are assigned to vxlan-instance 1 on the S4048-ON switch appear as shown in Figure 82. In this example, it is port **Te 1/48**.

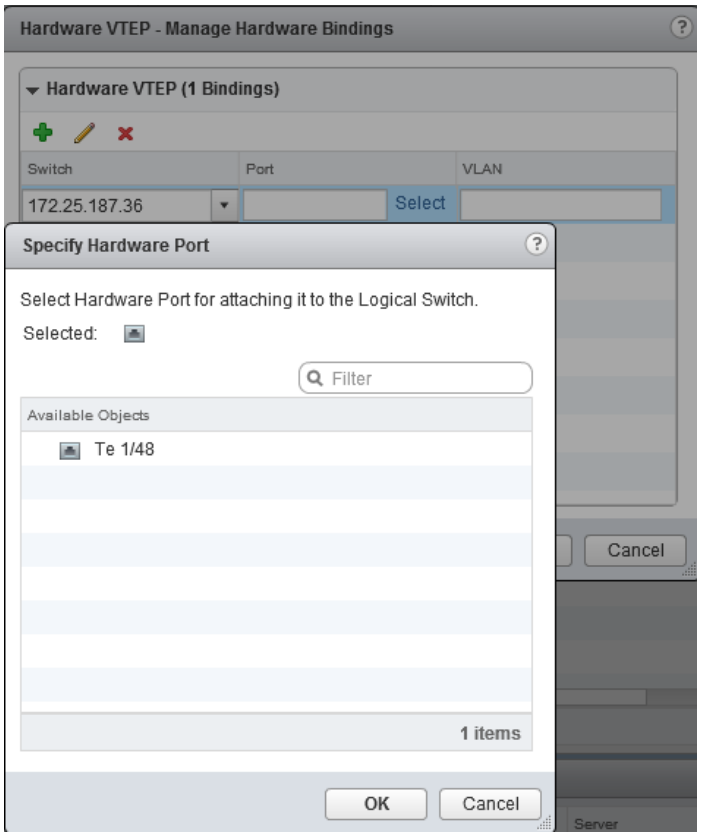
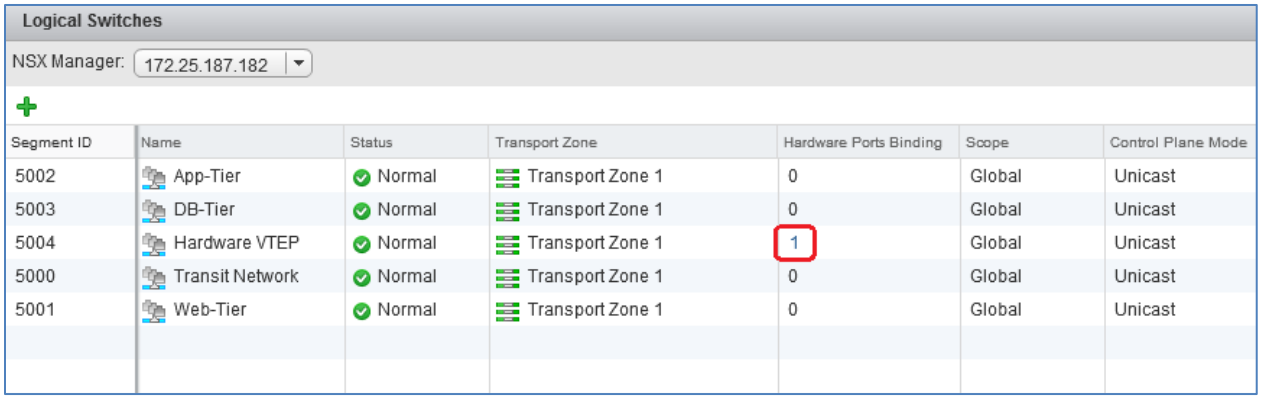


Figure 82 Manage Hardware Bindings – Specify Hardware Port window

c. Select the port and click **OK.**
d. Enter **0** in the VLAN box and click **OK**.

When complete, the **Logical Switches** page will look similar to **Error! Reference source not found.**. The **Hardware Ports Binding** column indicates one port, Te 1/48 in this example, is configured.



Segment ID	Name	Status	Transport Zone	Hardware Ports Binding	Scope	Control Plane Mode
5002	App-Tier	Normal	Transport Zone 1	0	Global	Unicast
5003	DB-Tier	Normal	Transport Zone 1	0	Global	Unicast
5004	Hardware VTEP	Normal	Transport Zone 1	1	Global	Unicast
5000	Transit Network	Normal	Transport Zone 1	0	Global	Unicast
5001	Web-Tier	Normal	Transport Zone 1	0	Global	Unicast

Figure 83 Hardware port bound to logical switch

13.2.4 Configure a replication cluster

The hardware VTEP is not capable of handling Broadcast, Unknown unicast and Multicast (BUM) traffic and requires at least one NSX-enabled host to process these requests.

On the **Home > Networking & Security > Service Definitions > Hardware Devices** page, next to **Replication Cluster,** click **Edit** and select up to 10 hosts. Only one host is active and the rest serve as backups. Figure 84 shows the four hosts in the compute cluster, Rack 2 Compute FC430, are selected.

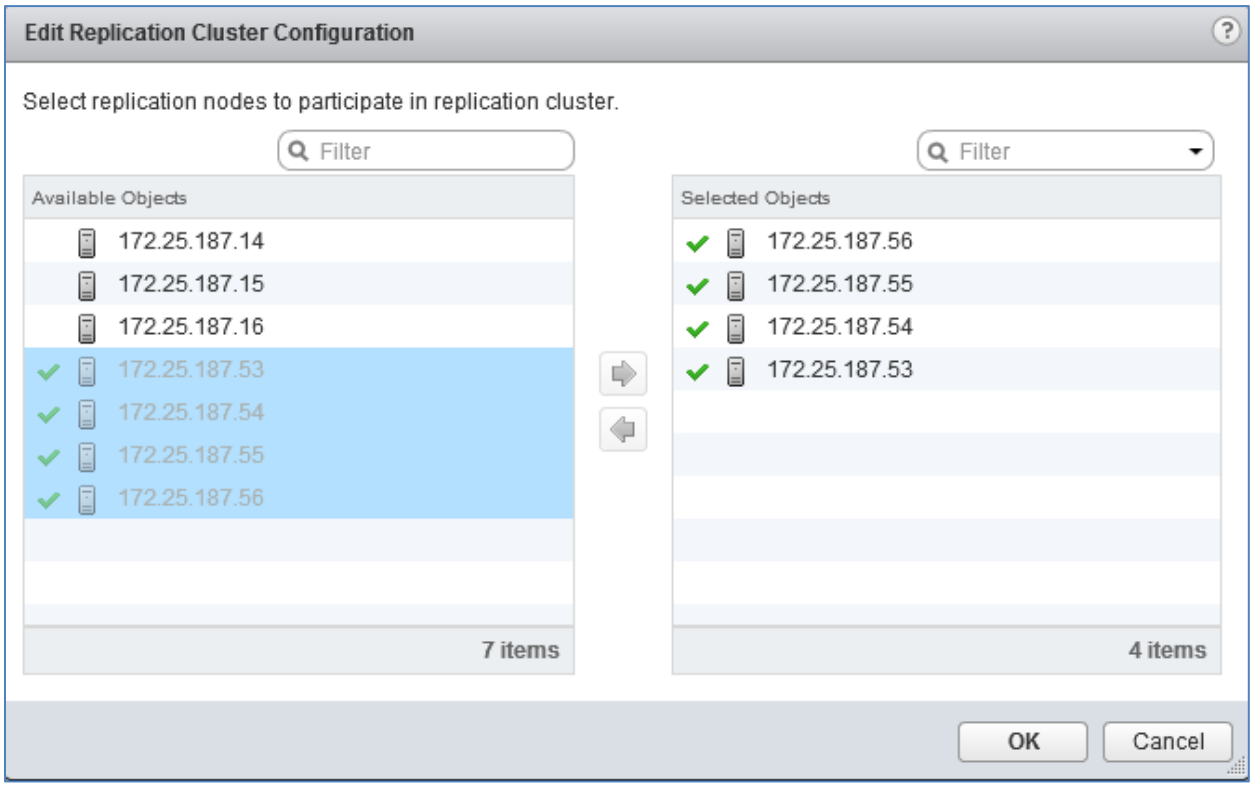


Figure 84 Creating a Replication Cluster

Click **OK** to add the hosts.

13.2.5 Hardware VTEP Validation

13.2.5.1 Switch commands and output

Use the following commands and output to verify the hardware VTEP configuration on the S4048-ON.

The `show vxlan vxlan-instance 1` command should return the information shown below. The managers shown in the output below are the three NSX controllers. All three should be connected.

```
HW-VTEP#show vxlan vxlan-instance 1
Instance        : 1
Mode            : Controller
Admin State     : enabled
Management IP   : 172.25.187.36
Gateway IP      : 10.77.4.254
MAX Backoff     : 30000
Controller 1    : 172.25.187.183:6640 ssl
Managers        :
                : 172.25.187.183:6640 ssl (connected)
                : 172.25.187.184:6640 ssl (connected)
                : 172.25.187.185:6640 ssl (connected)
Fail Mode       : secure
Port List       : Te 1/48
```

Use the `show vxlan vxlan-instance 1 logical-network` command to obtain the logical network name, to be used in the subsequent command.

```
HW-VTEP#show vxlan vxlan-instance 1 logical-network
Instance        : 1
Total LN count  : 1

 * - No VLAN mapping exists and yet to be installed
  Name                                    VNID
  3202111c-f90e-3c81-aa47-2aaceb72b0df     5004
```

The `show vxlan vxlan-instance 1 logical-network name <name>` command indicates the establishment of a MAC tunnel for each of the four hosts in the replication cluster, along with the software VTEP IP address of each host and the configured hardware port (Te 1/48).

```
HW-VTEP#show vxlan vxlan-instance 1 logical-network name 3202111c-f90e-3c81-
aa47-2aaceb72b0df
Name            : 3202111c-f90e-3c81-aa47-2aaceb72b0df
Description     :
Type            : ELAN
Tunnel Key      : 5004
VFI             : 28674
Unknown Multicast MAC Tunnels:
    10.55.2.1 : vxlan_over_ipv4 (up)
    10.55.2.2 : vxlan_over_ipv4 (up)
    10.55.2.3 : vxlan_over_ipv4 (up)
    10.55.2.4 : vxlan_over_ipv4 (up)
Port Vlan Bindings:
    Te 1/48: VLAN: 0 (0x80000001),
```

13.2.5.2 Traffic test

To validate functionality, send traffic between a physical server in the data center (running a Linux or Windows Server operating system for example) and a VM on the NSX network.

Connect the physical server to the configured port on the hardware VTEP (interface tengigabitethernet 1/48 in this example, shown in Figure 78 at the beginning of this section). The server's network adapter is assigned the address **10.77.4.1/24**.

In the web client, add a second network adapter to App-VM1 in the Rack 2 Compute FC430 cluster and connect it to the Hardware VTEP logical switch as follows:

1. In the web client, go to **Home > Hosts and clusters**.
2. Right click on the VM, **App-VM1**, and click **Edit Settings**.
3. Next to **New device**, select **Network** and click **Add**.

4. Next to **New Network**, expand the drop-down menu and select **Show more networks**. This opens the **Select Network** box shown in Figure 85:

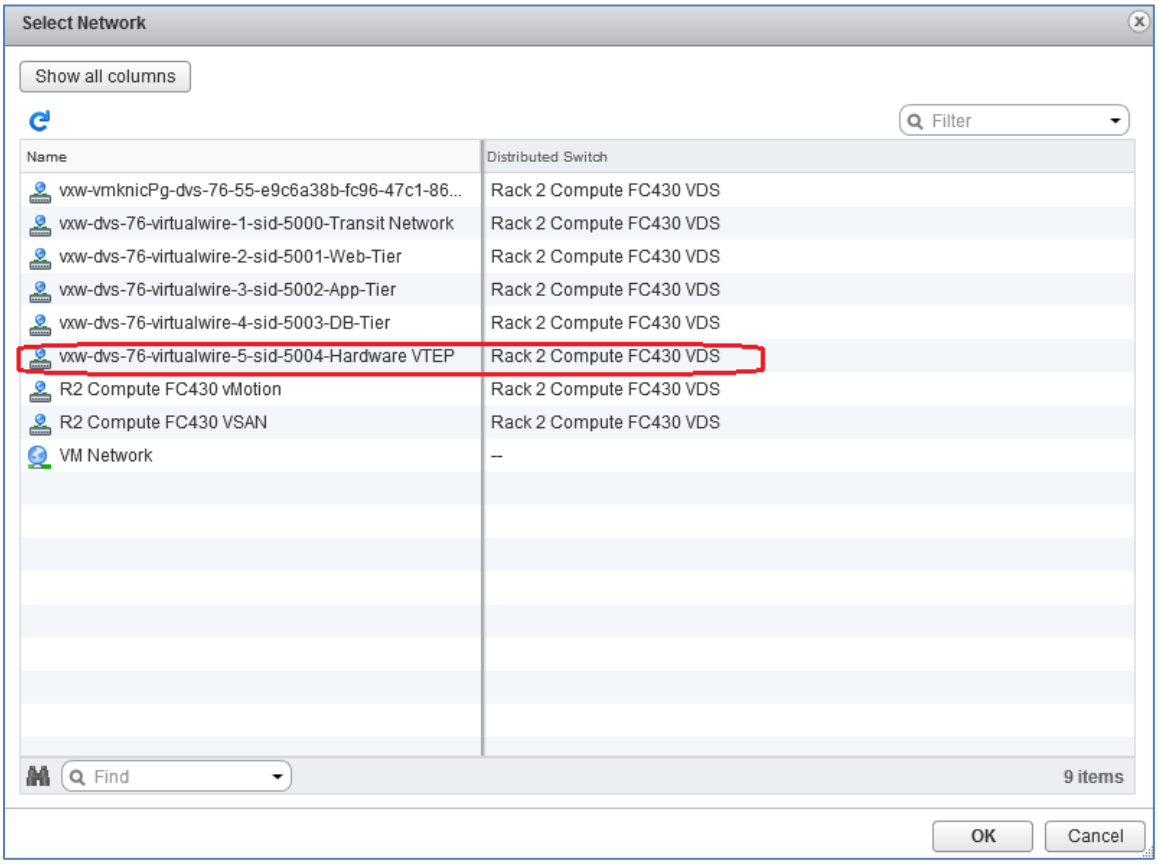


Figure 85 Select Network dialog box

5. Select the virtual wire labeled **Hardware VTEP** and click **OK** to return to the **Edit Settings** box.
6. In the **Edit Settings** box, expand **New Network**. Change the **Adapter Type** from E1000E to **VMXNET3** (since 10GbE adapters are used). Click **OK**.
7. Right click on **App-VM1** and select **Open Console**.
8. Log in to App-VM1 and set the IP address of the newly added adapter to **10.77.4.2/24**. The default gateway is not set or changed for this configuration. App-VM1's original adapter with IP address 10.10.20.11 and gateway remains configured for connectivity with other VMs.

Provided operating system firewalls are properly configured, App-VM1 (10.77.4.2) successfully pings the physical server (10.77.4.1) through the hardware VTEP using the new adapter.

App-VM1 continues to have connectivity to other VMs (Web-VM1, App-VM3, etc.) on their virtual networks (10.10.10.0, 10.10.20.0, etc.) as before.

14 Scaling guidance

14.1 Switch selection

The leaf layer in this deployment uses the Dell Networking S4048-ON because of its ability to provide a low-latency, non-blocking, layer-2 network architecture. It provides for growth and performance with 48 x 10GbE and six 40GbE ports.

The spine layer uses the Z9100-ON because it provides for substantial growth and outstanding performance with thirty-two 40/100GbE ports per switch. The solution outlined in this document provides scalability out to 16 racks without adding additional spine switches.

14.2 VSAN sizing

The largest single VSAN cluster size recommended for this deployment is 32 Dell FC430 servers installed in 8 PowerEdge FX2s enclosures per rack.

While the current maximum size for a VSAN cluster is 64 nodes, this deployment uses a 32-node cluster for the following reasons:

- Limiting a VSAN cluster to a single rack avoids the increased complexity of a layer 3 multicast deployment.
- Eight FX2s chassis per rack allows for moderate power consumption.

Note: See the VMware Virtual SAN Design and Sizing Guide and the VMware VSAN Compatibility Guide for more information.

14.3 Example – scale out to 3000 virtual machines

The goal of this section is to extend the solution outlined in this deployment guide to accommodate approximately 3000 virtual machines in compute clusters.

This is done as a mathematical exercise and there are many variables to consider when determining hardware requirements for a required number of VMs. To help estimate hardware needs based on the number of VMs required, VM specifications and the VSAN storage policy, see the scaling calculator attachment, **scaling_calc_fc430.xlsx.**

The tables below include the virtual machine profile and PowerEdge FC430 hardware used in this example:

Table 10 Virtual machine profile

Virtual CPUs	Virtual Memory (GB)	Virtual Disk Size (GB)
2	8	100

Table 11 Hardware per PowerEdge FC430

Sockets	Cores per socket	DIMM count	DIMM size (GB)	Total memory (GB)	Storage cache disk count	Storage cache disk size (GB)	Storage capacity disk count	Storage capacity disk size (GB)	Storage capacity total (GB)
2	14	8	32	256	1	800	7	800	5600

The virtual machine requirements in Table 10 and the FC430 server hardware in Table 11 are entered into scaling_calc_fc430.xlsx. For the number of VMs required, 3000 is entered into the spreadsheet.

Note: scaling_calc_fc430.xlsx uses the default VSAN storage policy (number of failures to tolerate = 1) and allows for 30% overhead per VMware VSAN scaling best practices. 30% overhead is also applied to CPU and memory resources. These are the recommended settings but are adjustable as indicated in the spreadsheet.

After entering the data above, the **Final Counts** section at the bottom of the spreadsheet indicates a requirement of 38 FX2s chassis containing 153 FC430 servers. These numbers are rounded up to 40 FX2s chassis containing 160 FC430 servers. At eight chassis per rack, 40 FX2s chassis divide equally across five racks.

Table 12 shows the final numbers for the compute clusters in a 3000 VM deployment.

Table 12 3000+ compute node example

FX2s chassis	FC430 servers	Racks	VMs	vCPUs	Total memory (TB)	Storage cache (TB)	Raw Storage capacity (TB)	Usable storage (TB)	Storage cache/ capacity ratio (10% min.)
40	160	5	3038	8960	40.96	128	896	434	14.29%

With five racks for five compute clusters and allowing one rack per management and edge cluster, this 3000 VM solution example uses seven racks. Its leaf-spine network consists of fourteen S4048-ON leaf switches (two per rack) and two Z9100-ON spine switches.

14.4 Port count and oversubscription

The following table outlines the connections for seven racks with two spine switches with 40Gb interconnect speeds.

Table 13 Oversubscription Information

	PowerEdge FC430	FN410S IOM (2 per FX2s)	IOM links to leaf switches (8 FX2s per rack)	Leaf links to spine switches per rack	Total links for leaf switches to two spine switches
Connections	2 NIC ports	6 uplink interfaces	8 chassis * 6 = 48 uplinks	2 per leaf switch, 2 leaf switches per rack = 4 links	7 racks * 4 links = 28 uplinks
Port bandwidth	10Gb	10Gb	10Gb	40Gb	40Gb
Total theoretical bandwidth	2 * 10 = 20Gb	60Gb	48 * 10Gb = 480Gb per rack (240Gb per leaf switch)	4 * 40Gb = 160Gb per rack (80Gb per leaf switch)	28 * 40Gb = 1120 Gb

This example provides for an oversubscription rate of 3:1 for 40Gb connectivity. To lower the subscription rate, make additional connections from the leaf switches to the spine switches as needed.

14.5 Rack Diagrams

Figure 86 shows the management cluster in Rack 1 and its related switches, plus the two spine switches used in this guide. The edge cluster in Rack 3 is identical, with the spine switches located in either rack.

Adequate space is available to allow for additional spine switches to be added as bandwidth requirements dictate. The management and edge clusters can also be combined in the same rack if preferred.

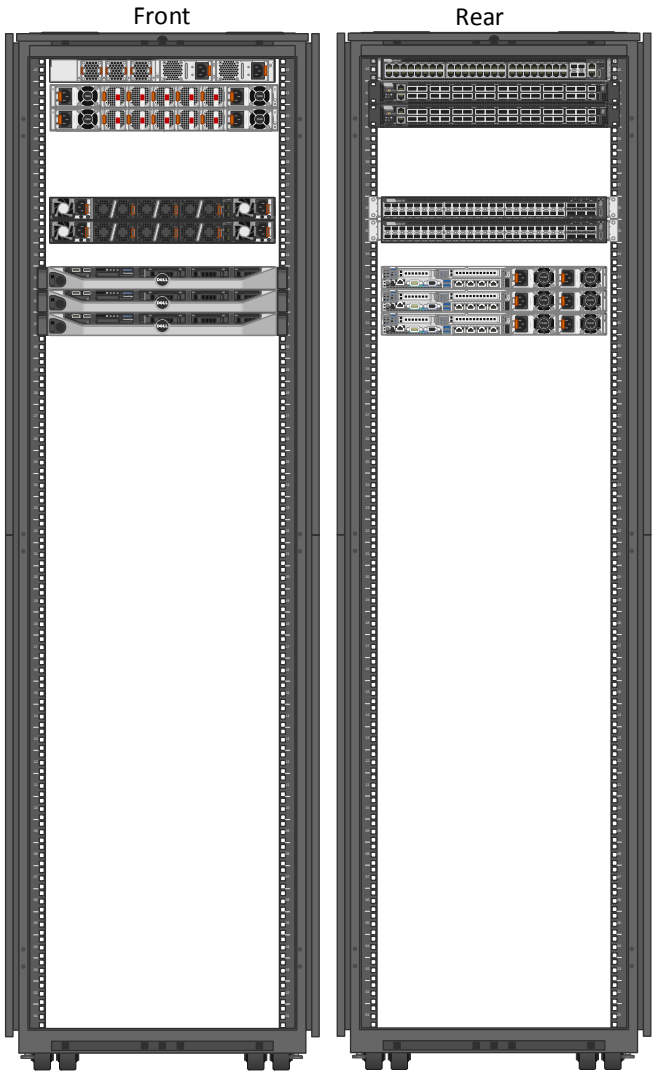


Figure 86 Rack containing management or edge cluster and spine switches

Figure 87 illustrates a rack containing a compute cluster with eight PowerEdge FX2s chassis, two leaf switches and a single management switch.

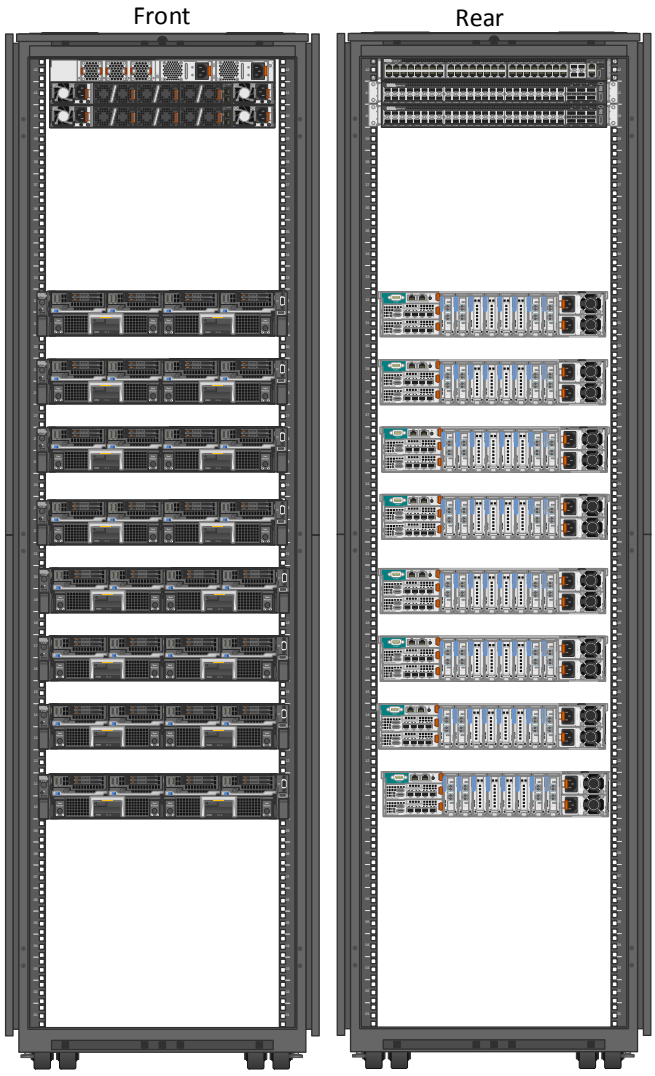


Figure 87 Rack containing compute cluster and switches

A Dell EMC validated hardware and components

The following tables present the hardware and components used to configure and validate the example configurations in this guide.

A.1 Switches

Qty	Item	Firmware Version
2	Z9100-ON Spine switch	DNOS 9.10.0.1 P3
6	S4048-ON Leaf switch	DNOS 9.10.0.1 P13
1	S4048-ON Hardware VTEP	DNOS 9.10.0.1 P13
3	S3048-ON Management switch	DNOS 9.10.0.1 P5

A.2 PowerEdge R630 servers

This guide uses six PowerEdge R630 servers, three in the Management cluster and three in the Edge cluster.

Qty per server	Item	Firmware Version
2	Intel Xeon E5-2695 v3 2.3GHz CPU, 14 cores	-
128	GB RAM	-
8	400 GB SAS SSD	-
1	PERC H730 Mini Storage Controller	25.2.1.0037
2	16 GB Internal SD Cards	-
1	QLogic 57840 10GbE QP rNDC (Required for Edge cluster, may substitute with QLogic 57810 10GbE DP rNDC in Mgmt. cluster)	7.12.19
1	Intel I350-T 1GbE DP PCIe adapter	17.5.10
-	R630 BIOS	2.1.7
-	iDRAC with Lifecycle Controller	2.30.30.30

A.3 PowerEdge FX2s chassis and components

This guide uses one FX2s chassis with four FC430 servers and two FD332 storage sleds in the Compute cluster.

Qty per chassis	Item	Firmware Version
1	FX2s Chassis Management Controller	1.32.200
4	FC430 servers. Each server contains: • 2 - Intel Xeon E5-2695 v3 2.3GHz CPU, 14 cores • 8 - 32GB DIMMS (256 GB total) • 8 - 800 GB SAS SSD (provided by FD332 storage sled) • 2 - 16 GB Internal SD Cards • 1 - QLogic 57810 10GbE DP LOM • 1 - PERC FD33xD Dual Storage Controller • FC430 BIOS • FC430 iDRAC with Lifecycle Controller	• - • - • - • - • 7.12.19 • 25.4.0.0015 • 2.1.5 • 2.30.30.30
2	FD332 storage sled with 16 x 800 GB SAS SSD	-
2	FN410S IOM	DNOS 9.10.0.1 P13
4	Intel I350-T 1GbE DP LP PCIe adapter	17.5.10

B Dell EMC validated software and required licenses

The Software table presents the versions of the software components used to validate the example configurations in this guide. The Licenses section presents the licenses required for the example configurations in this this guide.

B.1 Software

Item	Version
VMware ESXi	6.0.0 Update 2 - Dell EMC customized image version A00
VMware vSphere Desktop Client	6.0.0 build 3562874
VMware vCenter Server Appliance	6.0.0 Update 2 - build 3634788
vSphere Web Client	6.0.0 build 3617395 (included with VCSA above)
VMware NSX Manager	6.2.4 build 4292526

B.2 Licenses

The vCenter Server is licensed by instance. The remaining licenses are allocated based on the number of CPU sockets in the participating hosts.

Required licenses for the topology built in this guide are as follows:

- VMware vSphere 6 Enterprise Plus – 20 CPU sockets
- vCenter 6 Server Standard – 1 instance
- VSAN Advanced – 20 CPU sockets
- NSX for vSphere Enterprise - 14 CPU sockets

VMware product licenses can be centrally managed by going to the vSphere web client **Home** page and clicking **Licensing** in the center pane.

C Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

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

C.1 Dell EMC product manuals and technical guides

Manuals and documentation for Dell Networking S3048-ON

Manuals and documentation for Dell Networking S4048-ON

Manuals and documentation for Dell Networking Z9100-ON

Manuals and Documentation for PowerEdge FX2/FX2s and Modules

Manuals and documentation for PowerEdge R630

Dell TechCenter Networking Guides

PowerEdge FX2 – FN I/O Module – VLT Deployment Guide

Network Virtualization with Dell Infrastructure and VMware NSX

Dell Networking - VXLAN Technical Brief

Dell Network Virtualization Handbook for VMware NSX

Switch Configurations for Hardware VTEP

C.2 VMware product manuals and technical guides

VMware vSphere 6.0 Documentation Center

VMware NSX 6.2 Documentation Center

VMware vCenter Server 6.0 Deployment Guide

VMware Virtual SAN Design and Sizing Guide

VMware Compatibility Guide

VMware VSAN Compatibility Guide

VMware KB Article – Dell Networking VXLAN Hardware Gateway with NSX

D Support and Feedback

Contacting Technical Support

Support Contact Information	Web: http://Support.Dell.com/
	Telephone: USA: 1-800-945-3355

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.