Deployment and Best Practices Guide for Big Switch Networks' Big Cloud Fabric™ with VMware NSX

Dell EMC Networking Infrastructure Solutions
March 2018

Revisions

Date	Rev.	Description	Authors
March 2018	1.0	Initial release	Jim Slaughter, Shree Rathinasamy, Andrew Waranowski

.

Table of contents

1 Introduction

Applications are the engines for modern businesses. They drive innovation, operational efficiency, and revenue generation. They demand an infrastructure that is highly agile and easy to manage while reducing costs. These applications, which include mission-critical Enterprise Resource Planning (ERP) systems, multi-tier web applications, and big data, have placed new constraints on the networking infrastructure. Support for high east-west traffic bandwidth, virtual machine mobility, and multitenancy is critical.

Infrastructure teams have struggled to respond to these requirements. Unlike the rest of the portfolio, legacy networks remain highly static and require extensive manual intervention and operational overhead. To overcome these challenges, Software Defined Networking (SDN) is garnering due attention. SDN decouples the control plane from the data plane, allowing for dynamic management of the network. The advantages of SDN include agility, scalability, and superior network management. Open standards prevent lock-in with a single vendor and allow for financial flexibility. With such benefits, SDN solves emerging networking problems in the data center and helps keep up with virtualized environments. By providing various open networking hardware platforms and your choice of networking OS, Dell EMC Networking is an excellent choice for future-ready data centers.

This guide covers a Software Defined Data Center (SDDC) deployment based on the Dell EMC Ready Bundle for Virtualization. It includes a best practice leaf-spine network topology with a step-by-step configuration of a Big Cloud Fabric SDN solution integrated with VMware vCenter Server. It also provides the settings used in this environment for VMware distributed switches, vSAN clusters, and NSX components, following guidance from VMware Validated Design Documentation (VVD), release 4.1.

The goal of this guide is to enable a network administrator or engineer with traditional networking and VMware ESXi experience to build a scalable network using the hardware and software outlined in this guide.

1.1 Big Cloud Fabric

Dell EMC is working closely with Big Switch Networks to introduce the industry's first data center leaf-spine IP fabric solution using Dell EMC Open Networking switches and Big Cloud Fabric (BCF). This joint solution applies the hardware-software disaggregation enabled by Dell EMC and Big Switch Networks.

With built-in integration for VMware, BCF is ideal for virtual environments, network virtualization, and Hyper-Converged Infrastructure (HCI). It is the industry's first SDN-based fabric, using Dell EMC Open Networking switch hardware that provides intelligent, agile, and flexible networking for the VMware SDDC.

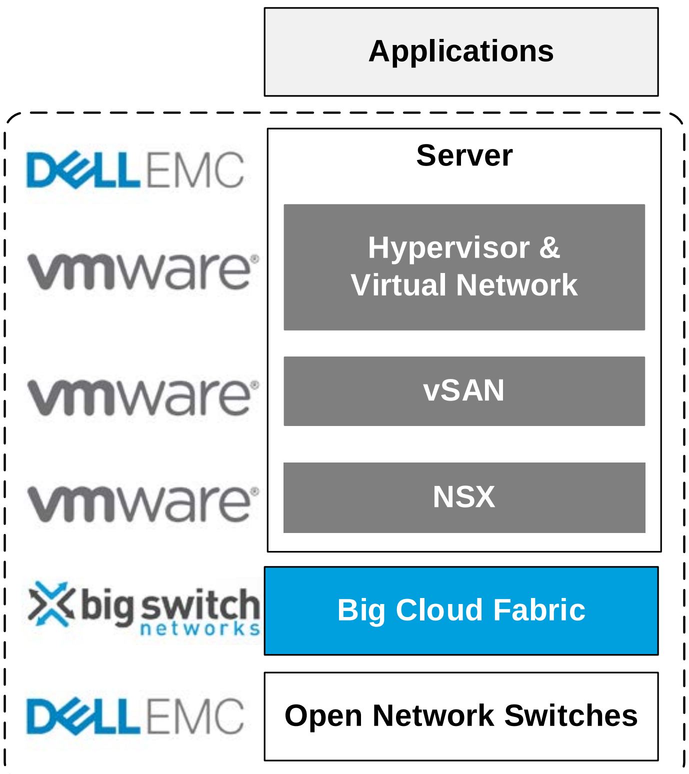


Figure 1 Dell EMC Open Networking with BCF and VMware

BCF utilizes SDN to provide scalability, improved management, visibility, flexibility, and intelligence to networks. Using redundant controllers, BCF delivers a "single logical switch" to add improved management and visibility to the network. Network agility is achieved through automation, zero-touch fabric, quicker troubleshooting, and controller-coordinated upgrades.

BCF allows hardware flexibility and prevents vendor lock-ins. Apart from open network hardware, BCF also helps in scaling seamlessly as per your workload needs. BCF accommodates the SDDC of the future by working in tandem with VMware vSphere, NSX, vSAN, OpenStack, VDI workloads, big data, and Software Defined Storage (SDS).

1.2 VMware vSAN

VMware vSAN combines the local physical storage resources of the ESXi hosts in a single cluster into a vSAN datastore. The vSAN datastore is used as the shared storage resource for creating virtual disks used by virtual machines in the cluster. vSAN is implemented directly in the ESXi hypervisor. It eliminates the need for external shared storage and simplifies storage configuration and virtual machine provisioning activities.

VMware vSphere features such as Distributed Resource Scheduling (DRS) and High Availability (HA) require shared storage. vMotion also integrates with vSAN. vSAN provides the performance and security needed for SDDCs at a lower cost. vSAN benefits include higher performance, higher storage efficiency, scalability, ease of management, security, and automation capability.

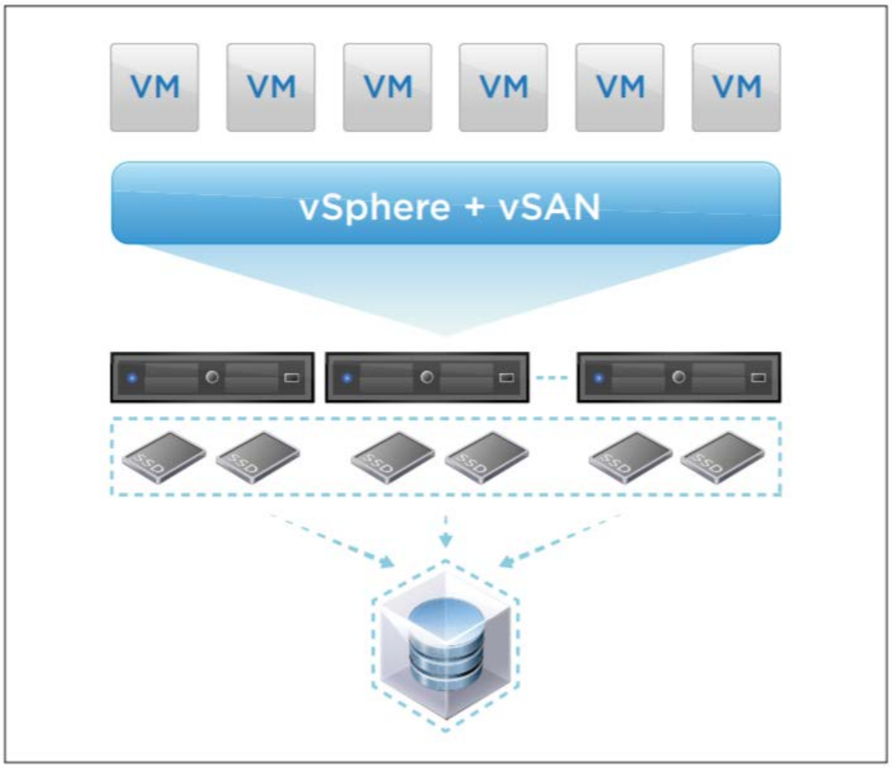


Figure 2 VMware vSAN

vSAN 6.6 features include a native HCI security solution with data-at-rest-encryption. The maintenance of vSAN is simplified with the aid of real-time support notifications and recommendations. Options for automation include the vSAN SDK and PowerCLI.

Note: For a list of VMware vSAN resources, see Appendix D.3.2.

1.3 VMware NSX

VMware NSX is a network virtualization technology. It allows for the decoupling of network services from the physical infrastructure. NSX creates logical networks on top of existing physical networks. 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 straightforward hardware migration and interoperability

Note: A list of VMware NSX resources is provided in Appendix D.3.3.

1.3.1 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 an existing 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 more 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:

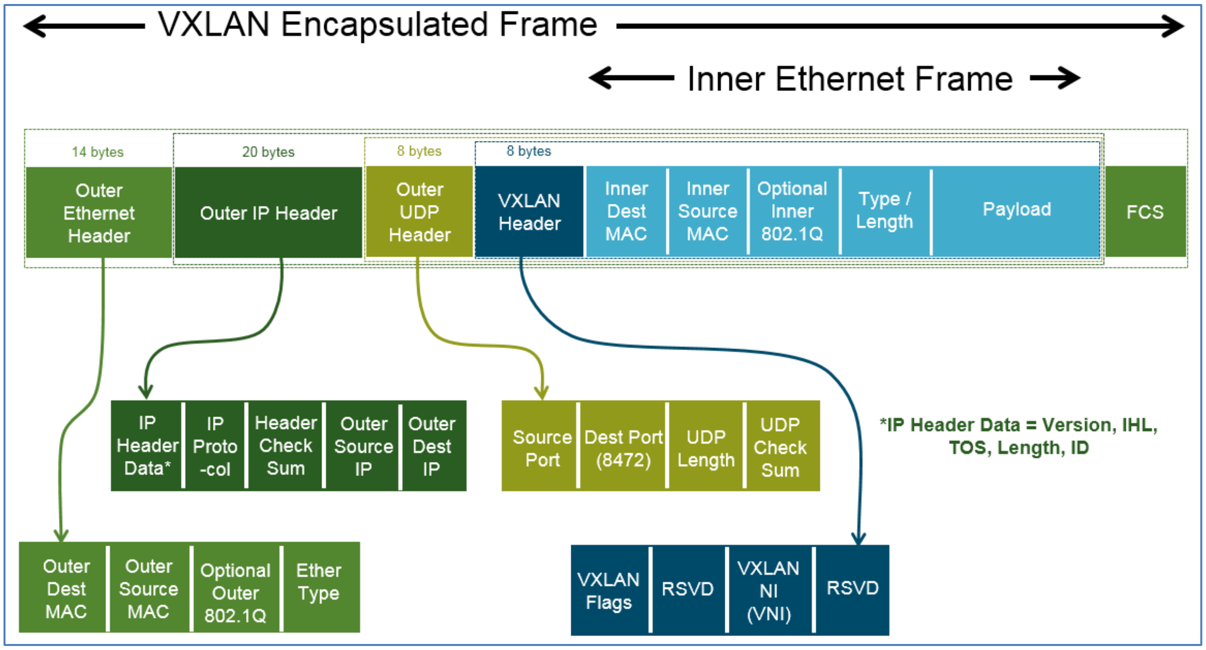


Figure 3 VXLAN encapsulated frame

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, ESXi hosts are configured with VXLAN tunnel endpoints (VTEPs). A VTEP is a VMkernel interface where VXLAN encapsulation and de-encapsulation occurs.

1.3.2 Micro-segmentation

Micro-segmentation, enabled by VMware NSX, allows granular security controls within the data center. Data centers are typically secured at the perimeter using various methods and systems, while security controls within the data center are minimal. Micro-segmentation handles this issue by providing granular controls that help prevent unauthorized lateral movements within the data center.

Increasing east-west traffic within data centers can be plagued by bottlenecks in firewalls and hairpinning of traffic. Micro-segmentation helps to logically manage the data center by assigning security controls and policies to each segment.

NSX reproduces the complete set of layer 2-7 networking services, such as switching, routing, and firewalling in software. This enables chaining of security technologies as controls are implemented, adding a higher level of security within the data center. Networking and security services are now tied to the hypervisors and individual VMs, which enable mobility of services when a VM is moved.

1.4 Typographical conventions

This document uses the following typographical conventions:

`Monospaced text`	Command Line Interface (CLI) examples
`Bold monospaced text`	Commands entered at the CLI prompt
`Italic monospaced text`	Variables in CLI examples
Bold text	Graphical User Interface (GUI) fields and information entered in the GUI

2 Hardware overview

This section briefly describes the hardware used to validate the deployment example in this guide. Appendix B provides a complete listing of hardware and components used.

2.1 Dell EMC Networking S3048-ON

The Dell EMC Networking S3048-ON is a 1-Rack Unit (RU) switch with forty-eight 1GbE Base-T ports and four 10GbE SFP+ ports. In this guide, one S3048-ON supports out-of-band (OOB) management traffic in each rack.

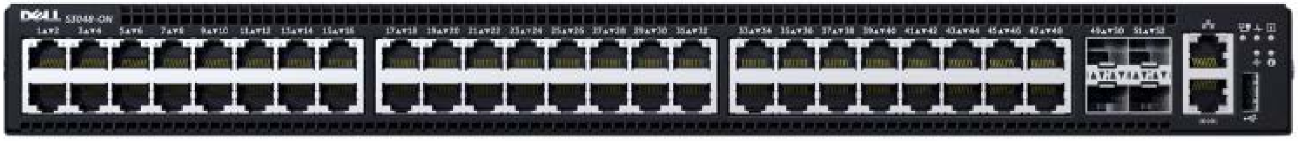


Figure 4 Dell EMC Networking S3048-ON

2.2 Dell EMC Networking S4048-ON

The Dell EMC Networking S4048-ON is a 1-RU, multilayer switch with forty-eight 10GbE SFP+ ports and six 40GbE QSFP+ ports. Four S4048-ON switches (two per rack) are used as leaf switches in this guide.

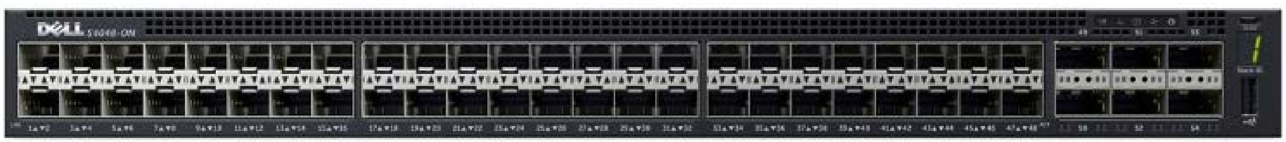


Figure 5 Dell EMC Networking S4048-ON

2.3 Dell EMC Networking Z9100-ON

The Dell EMC Networking Z9100-ON is a 1-RU, multilayer switch with thirty-two ports supporting 10/25/40/50/100GbE plus two 10GbE ports. Two Z9100-ON switches are used as spines in this guide.

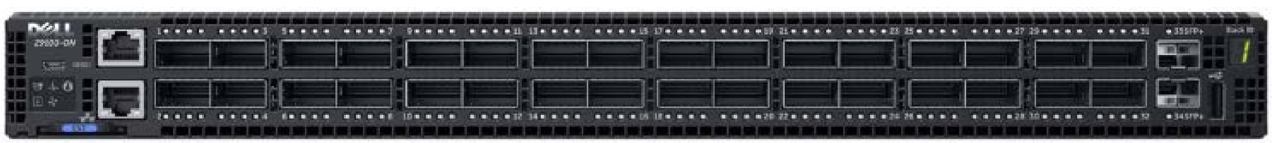


Figure 6 Dell EMC Networking Z9100-ON

2.4 Dell EMC PowerEdge R740xd

The Dell EMC PowerEdge R740xd is a 2-RU, two-socket server platform. It allows up to 32 x 2.5" SSDs or HDDs with SAS, SATA, and NVMe support. Ideal workloads include VMware vSANs, big data services, and data analysis. In this guide, four R740xd servers are in the Compute-Edge cluster.

Figure 7 Dell EMC PowerEdge R740xd

Note: VMware recommends each vSAN cluster contain a minimum of 10% flash storage capacity. For more information, see the *VMware vSAN Design and Sizing Guide*. The R740xd systems used in this deployment each contain twenty SSDs (100% flash storage).

2.5 Dell EMC PowerEdge R630

The Dell EMC PowerEdge R630 is a 1-RU, two-socket platform. In this guide, four R630 servers are in the Management cluster and four are in the Compute cluster.

Figure 8 Dell EMC PowerEdge R630

Note: For new deployments, the Dell EMC PowerEdge R640 is the latest generation 1-RU, two-socket platform. For existing deployments, the Dell EMC PowerEdge R630 or other supported hardware listed on the *VMware Compatibility Guide* web site may be used.

2.6 Big Switch Networks' BCF Controller Appliance

The Big Switch Networks' BCF Controller Appliance is a 1-RU, two-socket platform designed to deliver the right combination of performance, redundancy, and value. For fault tolerance, two appliances comprise a BCF Controller cluster.

Figure 9 Big Switch Networks' BCF Controller Appliance

3 Pod architectures

A pod is a combination of computing, network, and storage capacity designed to be deployed as a single unit. As a result, a pod is the largest unit of failure in the SDDC. Carefully engineered services ensure each pod has little to no shared vulnerability between pods.

There are two different types of pods used in this deployment:

- Big Cloud Fabric pod
- VMware vSphere pod

3.1 Big Cloud Fabric pod

A BCF pod contains two BCF Controllers and a leaf-spine network spanning up to sixteen racks with two leaf switches per rack.

The sixteen-rack limit is dictated by the port count of the spine switches used in the leaf-spine network. In this deployment, Z9100-ON switches with thirty-two 40GbE interfaces per switch are used as spines.

Note: Big Switch Networks has tested a maximum of 128 leaf switches in one pod. See the *Big Cloud Fabric Verified Scale* document on the Big Switch Networks' support site for more information. Big Switch Networks' documentation requires a customer account to access. Contact your Big Switch Networks account representative for assistance.

In this example, the BCF pod contains two Z9100-ON spine switches and four S4048-ON leaf switches distributed over two racks. Two BCF Controller Appliances are deployed in an active/standby configuration.

Each fabric device can switch at layer 2 or route at layer 3, while the BCF Controller centrally provides the intelligence required to make full use of redundant links. Incremental upgrades of the forwarding tables are dynamically pushed to each switch to ensure a stable and dynamic network operation. Spanning Tree Protocol is not required, and all links are in forwarding mode.

The networking architecture used by BCF is a leaf-spine design that increases server-to-server bandwidth. The leaf-spine architecture creates a high-performance backplane that can be extended by simply adding more switches.

The entire BCF pod can be thought of as a single, larger modular switch as shown:

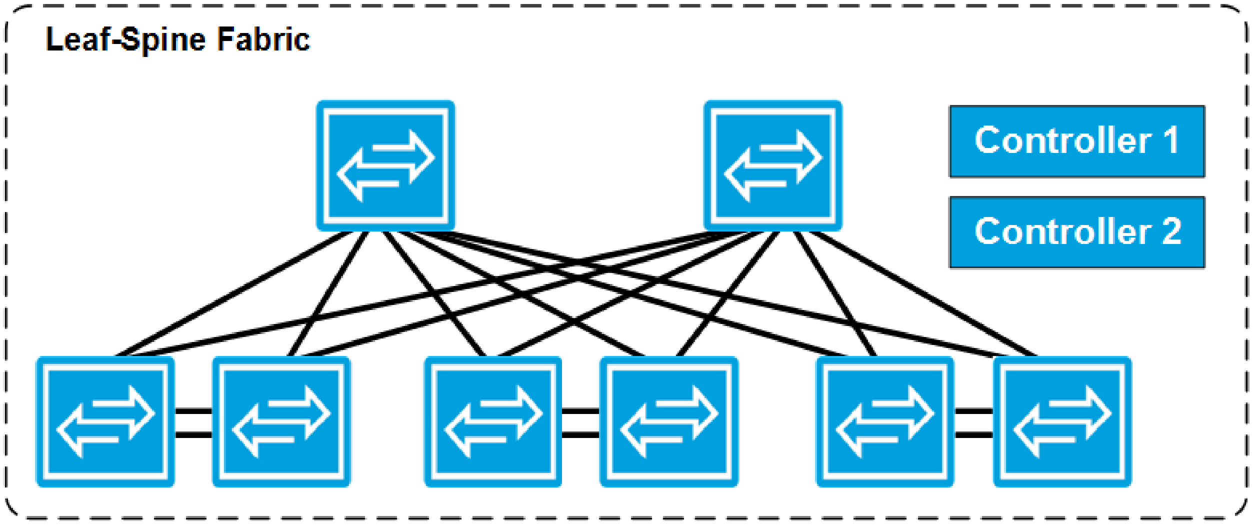


Figure 10 BCF pod

A pair of BCF Controllers provides functionality similar to the dual supervisors on a modular chassis. The spine switches provide the functionality of the backplane, while the leaf switches are similar in function to line cards.

The leaf-spine architecture provides a simple and efficient design in response to challenges inherent in the hierarchical data center architecture. The 2-layer leaf-and-spine architecture optimizes bandwidth between switch ports within the data center by creating a high-capacity fabric using multiple spine switches that interconnect the edge ports of each leaf switch. This design provides consistent latency and minimizes the hops between servers in different racks.

The design lends itself well to the creation of an independent, replicable pod that scales without disrupting network traffic. The addition of more leaf switches increases the number of switch edge ports for connecting to servers. Adding spine switches increases the fabric bandwidth and decreases oversubscription ratios.

3.2 VMware vSphere pods

VMware Validated Design Documentation (VVD), release 4.1 defines the concept of VMware pods. VVD 4.1 also contains numerous best practices for VMware vSphere deployment.

There are four types of VMware vSphere pods:

- Management pod
- Shared edge and compute pod
- Compute pod
- Storage pod

The management pod runs the virtual machines that manage the SDDC. These virtual machines host vCenter Server, Platform Services Controllers (PSCs), NSX Manager, vRealize Log Insight, and other shared management components. All management, monitoring, and infrastructure services are provisioned to a vSphere cluster which provides high availability for these critical services.

The shared edge and compute pod runs the required NSX services (NSX Controllers and Edge Services Gateways) to enable north-south routing between the SDDC and the external network and east-west routing inside the SDDC. This pod also hosts the SDDC tenant compute resources and virtual machines. The shared pod combines the characteristics of typical edge and compute pods into a single pod.

The compute pod hosts the SDDC tenant compute resources and virtual machines. The pod scales by adding nodes and racks as needed, which increases computing and storage capacity linearly. As the SDDC grows, additional compute-only pods are added as needed.

The storage pod provides secondary storage using NFS, iSCSI or Fibre Channel.

Note: vSAN datastores are considered primary storage and do not reside in the storage pod. vSAN datastores reside with their compute clusters in the management, compute, and shared edge and compute pods. Deployment of a storage pod for secondary storage is beyond the scope of this document. See the [VMware Validated Design Documentation] for storage pod deployment information.

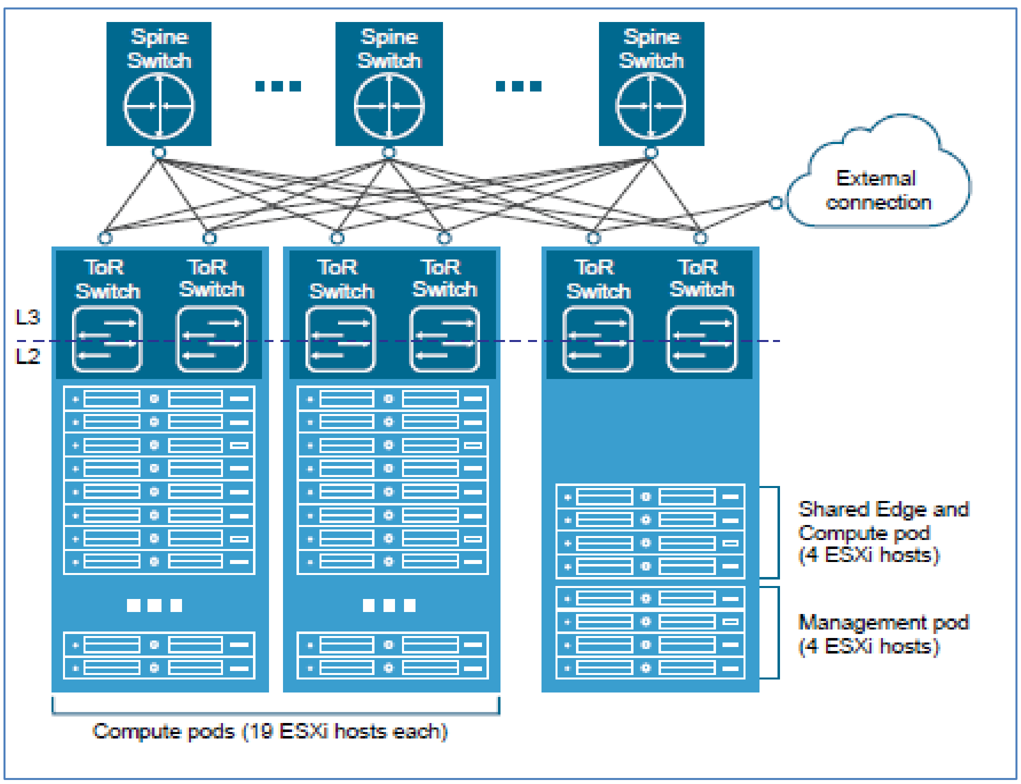


Figure 11 VMware vSphere pods

While a pod usually occupies one rack, it is possible to aggregate multiple pods into a single rack or to span a single pod across multiple racks.

This document covers deployment of three pods in two racks as listed in Table 1. Each pod is configured as a VMware cluster using the cluster names shown.

Table 1 Pods and corresponding VMware clusters

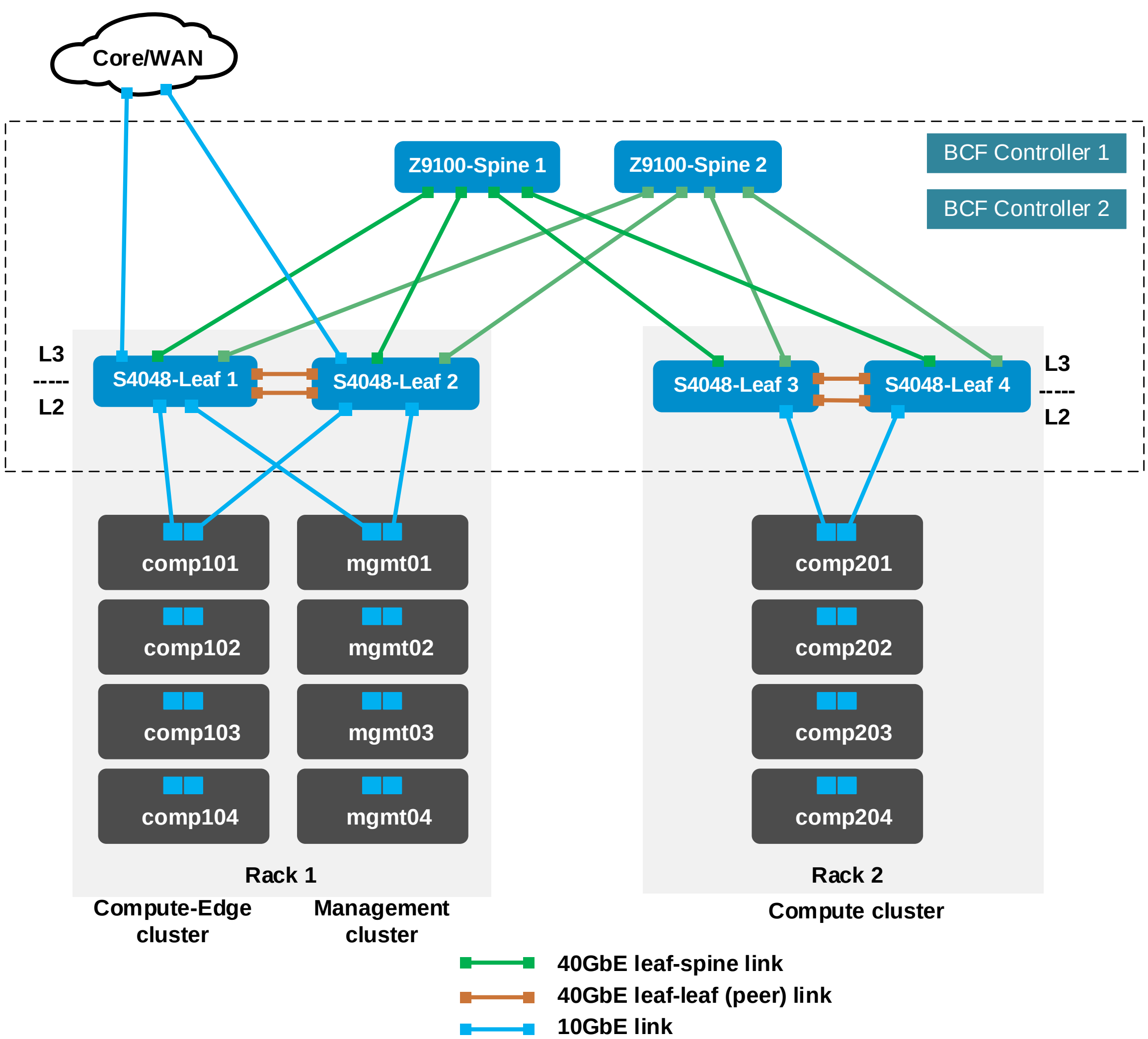
Rack	Pod	VMware cluster name
Rack 1	Management	Management
Rack 1	Shared edge and compute	Compute-Edge
Rack 2	Compute	Compute

4 Topology

The topology used in this deployment consists of a BCF leaf-spine network for in-band production traffic, an out-of-band (OOB) management network, and an OOB physical switch (P-switch) network for BCF Controller-to-switch communication.

4.1 Production network

The in-band production network in this guide is used for VXLAN (NSX), vSAN, and vMotion traffic. The production network connections, ESXi hosts, and clusters in this topology are shown in Figure 12. The BCF leaf-spine topology is shown inside the dashed line.



Note: For the 10GbE links downstream from the leafs, only connections to the first host in each cluster are shown for clarity.

Figure 12 Production topology with leaf-spine network and ESXi hosts

The leaf-spine topology includes two S4048-ON leaf switches at the top of each rack and two Z9100-ON spine switches. The layer 2/layer 3 (L2/L3) boundary is at the leaf switches. Each leaf and spine switch runs Big Switch Networks' Switch Light OS and is managed by a redundant pair of BCF Controllers. Every leaf has one 40GbE connection to every spine, and two 40GbE connections to its peer leaf.

Each ESXi host has one 10GbE connection to each of the two leaf switches in the rack.

There are four hosts in each cluster in this deployment. This follows the VVD 4.1 recommendation of at least four hosts per vSAN cluster. This allows an ESXi host to be taken offline for maintenance without impacting the overall vSAN cluster health. vSANs are enabled on clusters in Section 9 of this guide.

4.2 Host-to-leaf switch connection details

The production network connection details for hosts and leaf switches are shown in the figures below. Additional hosts in each cluster, not shown, are connected in the same manner.

> **Note:** Specific network interface cards (NICs) used on ESXi hosts in this paper are listed in Appendix B. See the *VMware Compatibility Guide* for a complete list of supported NICs.

In Rack 1, two 10GbE SFP+ ports from each host in the Management and Compute-Edge clusters are connected to Leaf 1 and Leaf 2 as shown:

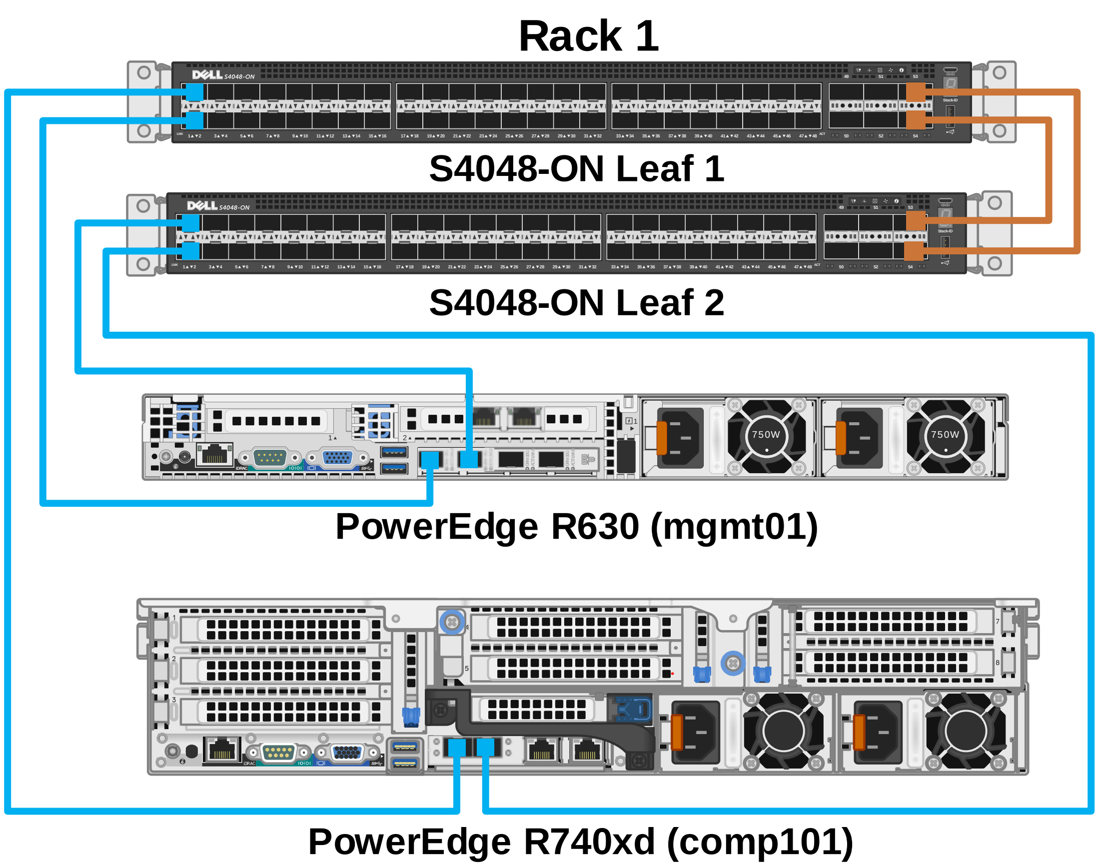


Figure 13 Production network port connection details – Rack 1

In Rack 2, two 10GbE SFP+ ports from each host in the Compute cluster are connected to Leaf 3 and Leaf 4 as shown:

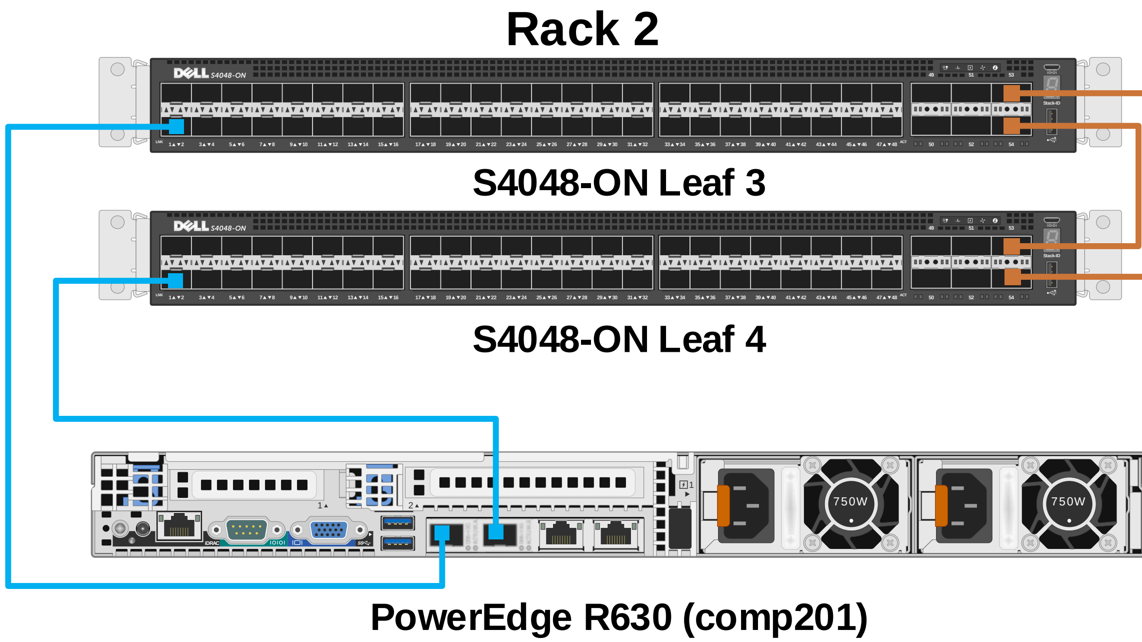


Figure 14   Production network port connection details – Rack 2

4.3 vSphere component locations

The management cluster contains all vSphere management components including vCenter Server Appliances (VCSAs), Platform Services Controllers (PSCs), and NSX Managers. The management cluster also contains NSX components dedicated to Management cluster NSX traffic. This includes an NSX Controller cluster, a Distributed Logical Router (DLR), and an Edge Services Gateway (ESG). Application VMs on NSX networks for management functions are also located in the management cluster. These VMs are named mgmt-*nn* in this guide.

The Compute-Edge cluster contains the NSX components dedicated to Compute-Edge and Compute cluster NSX traffic. This includes an NSX Controller cluster, a DLR, and an ESG. Application VMs on NSX networks for compute functions are located in the Compute-Edge and Compute clusters. These VMs are named app-*nn* and web-*nn* in this guide.

Note: Application VMs on NSX networks are commonly referred to as "NSX VMs" in this guide.

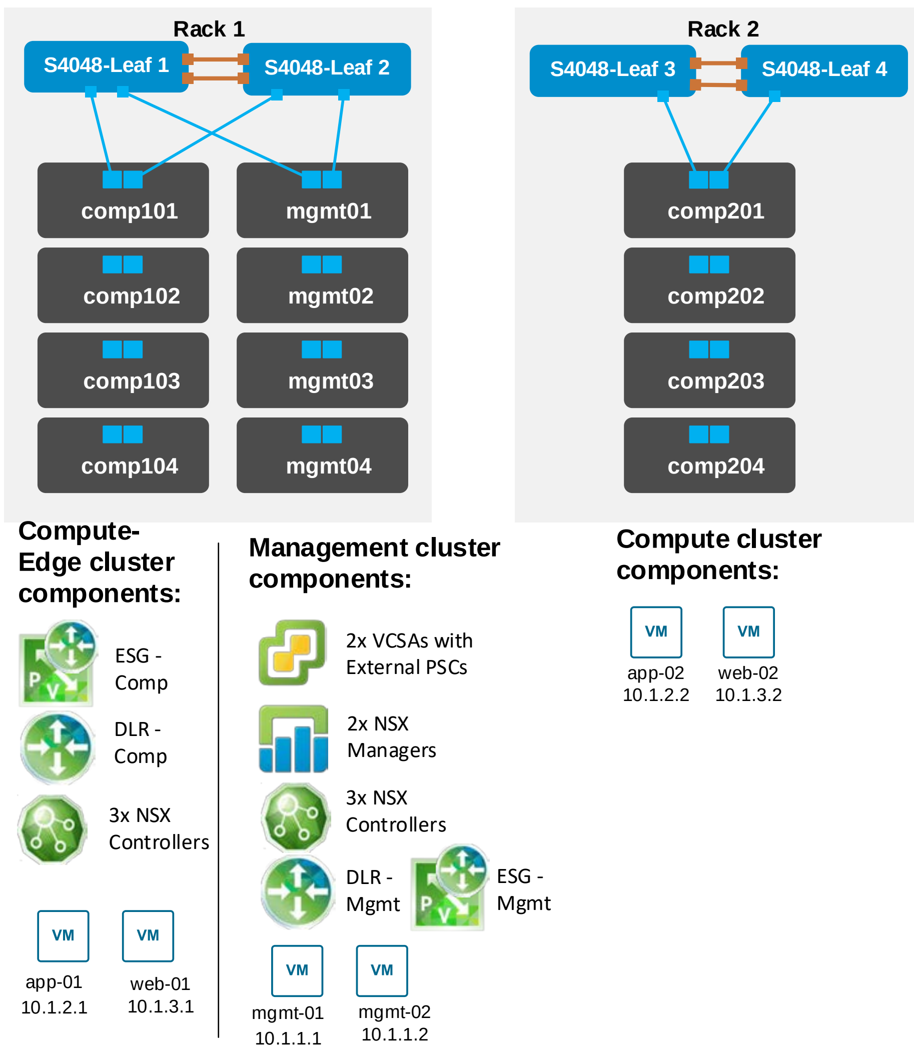


Figure 15 vSphere component locations in clusters

4.4 OOB networks

There are two administrative OOB networks in this deployment that are isolated from the production network:

- **OOB Management network** – used for server iDRAC, ESXi, and BCF Controller management
- **Physical switch (p-switch) control network** – used by BCF Controllers for leaf and spine switch management

One S3048-ON switch is installed as a top-of-rack (ToR) switch in each rack for OOB network traffic. All connections from hosts and switches in the rack to the S3048-ON ToR switch are 1GbE Base-T.

> **Note**: Using redundant OOB network switches in each rack is optional. Failure of OOB networks does not affect traffic on the production networks. This deployment example uses a single S3048-ON switch in each rack. See the *BCF Deployment Guide* for redundant OOB network switch information.

To separate the management and p-switch networks, two VLANs are configured on each S3048-ON: VLAN 100 and VLAN 200.

Ports assigned to VLAN 100, shown in red, are used for OOB management network connections and ports assigned to VLAN 200, shown in blue, are used for p-switch control network connections.

> **Note:** Big Switch Networks recommends using a dedicated broadcast domain for the p-switch control network. Using a separate VLAN accomplishes this.

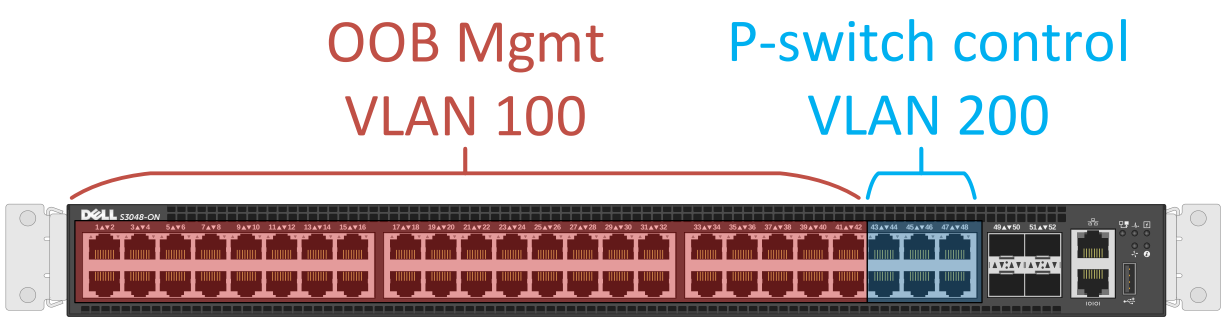


Figure 16 S3048-ON ports and administrative network VLANs

4.4.1 OOB management network connections

The OOB management network is used for PowerEdge server and BCF Controller configuration and monitoring. In this deployment guide, devices on the OOB management network use IPv4 addresses in the 100.67.0.0/16 address range.

As shown in Figure 17, each PowerEdge server has two connections to this network. One is for ESXi host management and one is for the server's iDRAC. Each BCF Controller appliance has one connection to the management network.

Notes: See your Dell EMC PowerEdge Server documentation for iDRAC features and instructions.
BCF Controllers have two ports available for management connections. For redundancy, both controller management ports may be used, and a second S3048-ON may be added in Rack 1. See the *BCF Deployment Guide* for more information.

All connections from hosts and switches in the rack to the S3048-ON ToR switch are 1GbE Base-T. 10GbE SFP+ ports are available on S3048-ON switches for uplinks to the management core.

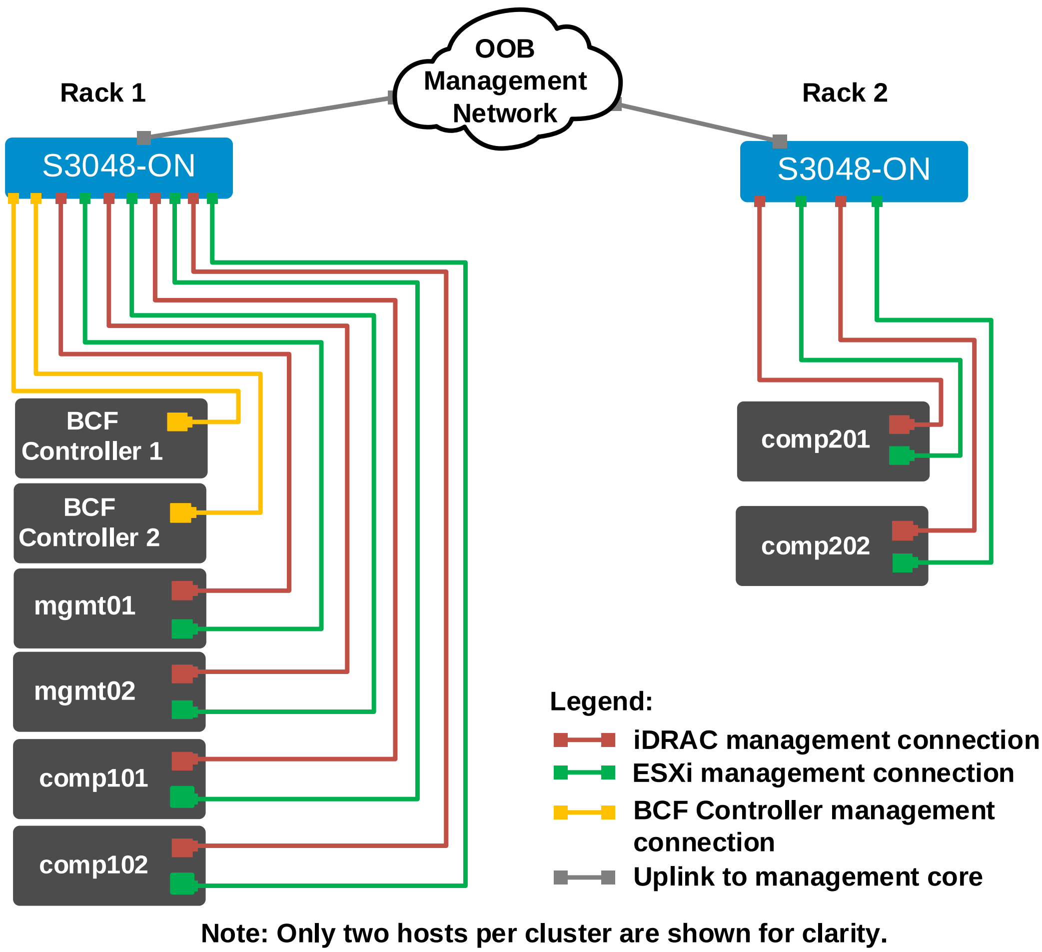


Figure 17 OOB management network connections

The OOB connection details for each unique device used in this deployment are shown in Figure 18. All connections are to ports in the OOB management VLAN on the S3048-ON switch. Additional systems in each cluster, not shown, are connected in an identical manner.

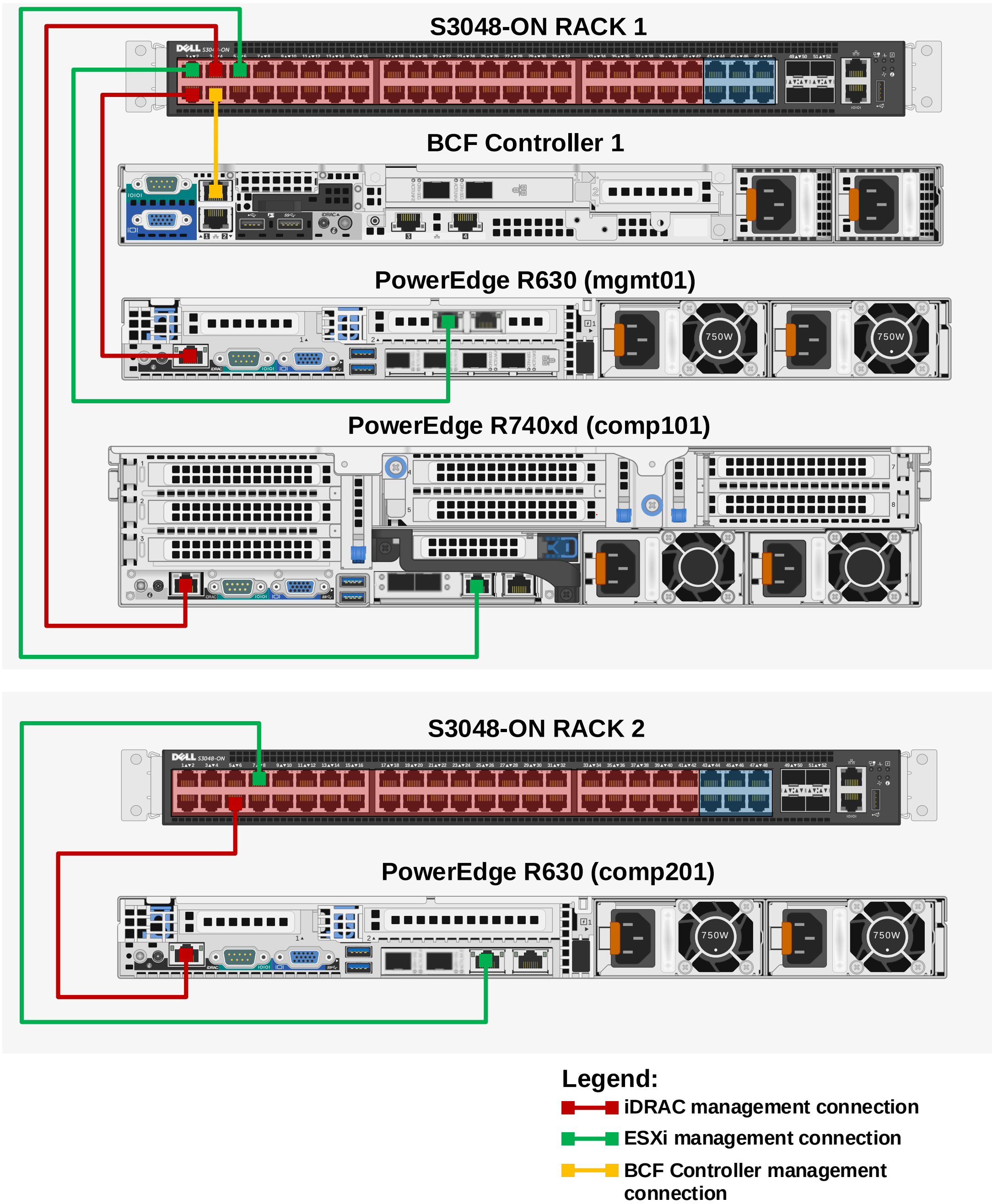


Figure 18 OOB management network port connection details

4.4.2 BCF p-switch control network connections

The physical switch, or p-switch, control network contains the BCF Controllers and all leaf and spine switches. This network is used by the BCF Controllers for leaf and spine switch configuration and management.

> **Note**: BCF Controllers have two ports available for p-switch control network connections. For redundancy, both ports may be used. A second S3048-ON should be added to Rack 1 for this case. See the *BCF Deployment Guide* for more information.

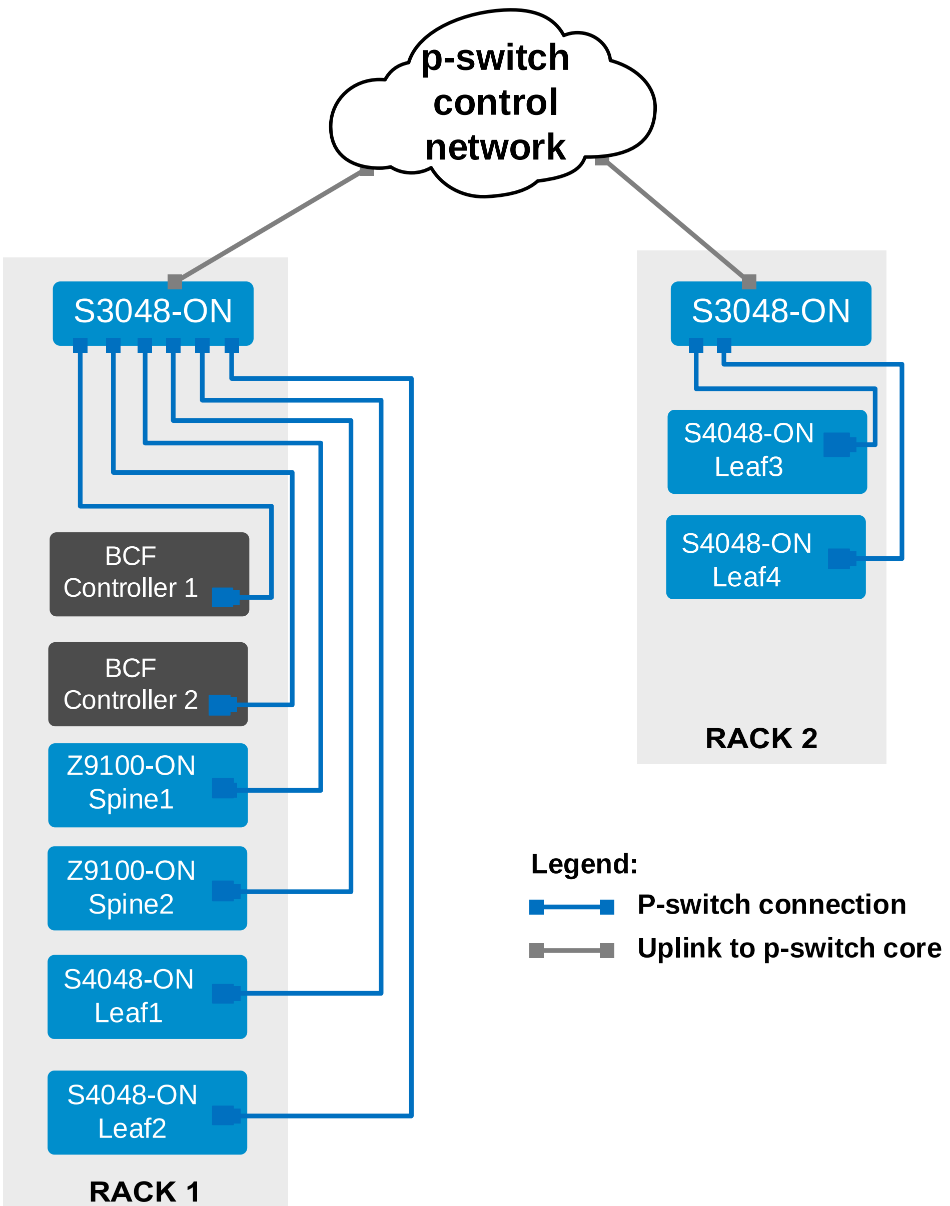


Figure 19 P-switch control network connections

In Rack 1, leaf and spine switch OOB management ports and BCF Controller p-switch ports are connected to S3048-ON ports in the p-switch VLAN as shown:

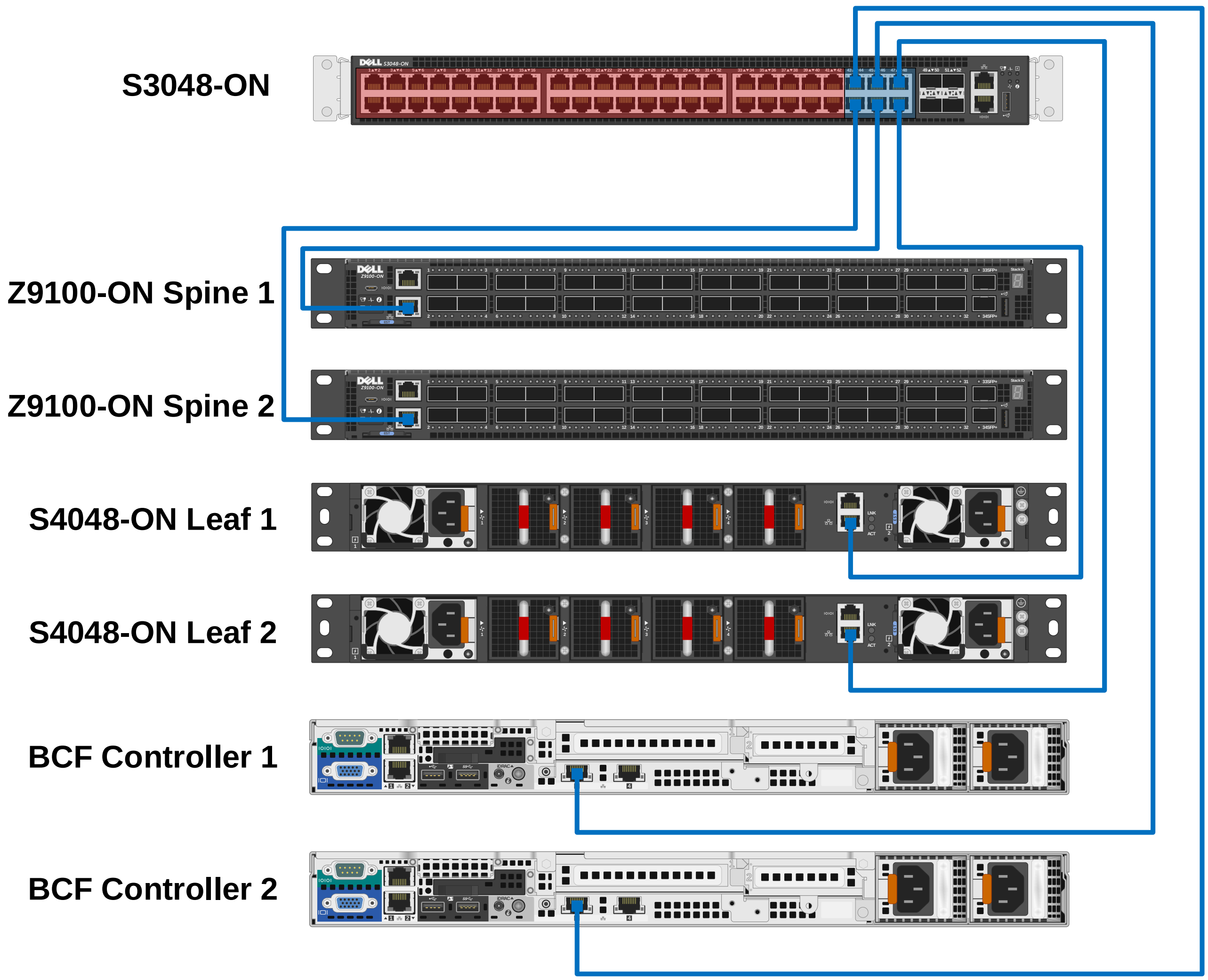


Figure 20 Rack 1 p-switch control network port connection details

The leaf switches in Rack 2, not shown, are connected in the same manner to the S3048-ON in Rack 2. The S3048-ON in Rack 2 must be able to reach the BCF Controllers via the p-switch network as shown in Figure 19.

Note: For small deployments or testing purposes, the leaf switches in Rack 2 may be connected directly to the S3048-ON in Rack 1. Ensure these connections are to ports in the p-switch VLAN on the S3048-ON.

4.5 BCF Controller in-band connections

The BCF Controller in-band connections enable the use of BGP on the leaf-spine network and connections to the core router. The BCF in-band connections are made by connecting each controller to both leaf switches in Rack 1.

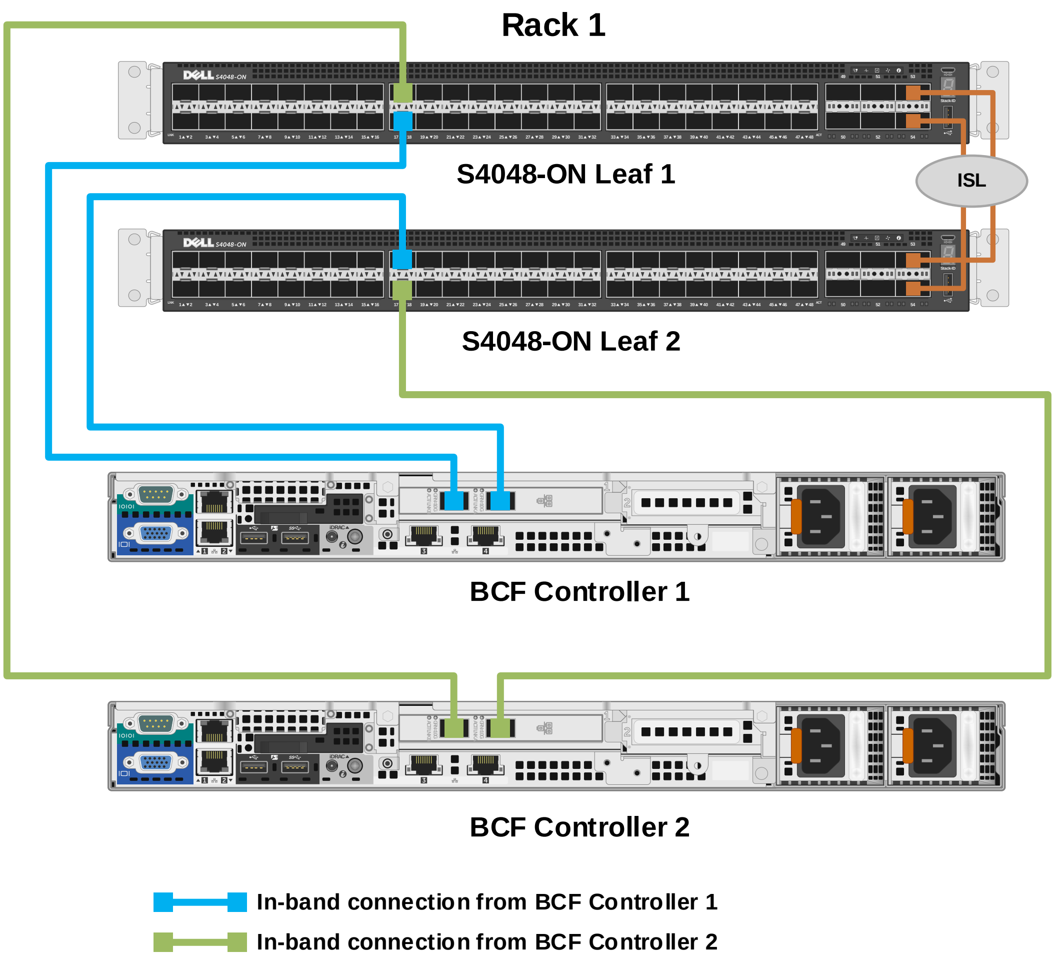


Figure 21 In-band connections for BCF Controllers 1 and 2

5 BCF deployment

This section provides steps to deploy the BCF Controller cluster and how to resolve common warning and error messages.

Note: For more information on BCF deployment, see the *Big Cloud Fabric Deployment Guide*.

5.1 BCF Controller overview

The BCF Controller provides a "single pane of glass" for management of all leaf and spine switches. The BCF Controller supports a familiar CLI and a web-based GUI. Any custom orchestration can be executed by using the industry-standard RESTful application programming interface (API).

BCF supports traditional tools for debugging, including ping, traceroute, show commands, and redirecting packets using port mirroring for fault analysis. The BCF Controller also supports unique troubleshooting tools, such as Fabric Test Path and Fabric Analytics to quickly isolate, identify, and resolve forwarding and application faults.

Figure 22 BCF GUI dashboard page

5.2 Deployment overview

Two BCF Controllers are deployed as a cluster for redundancy. The cluster is created when the first controller is deployed as the active controller. The second controller is joined to the cluster in the standby role during deployment. If the active controller fails, the standby controller automatically becomes the active controller.

The OOB management network settings used during deployment of each controller in this example are shown in Table 2.

Note: IPv4, IPv6, or both may be used on the OOB Management network. This deployment uses IPv4 only.

Table 2 BCF Controller initial configuration settings

Hostname	IP address	IPv4 prefix length	Default gateway	DNS server address	DNS search domain
bcfctrl01	100.67.187.201	24	100.67.187.254	100.67.189.33	dell.local
bcfctrl02	100.67.187.202	24	100.67.187.254	100.67.189.33	dell.local

Cluster settings used during deployment are shown in Table 3. A new cluster is created during deployment of the first controller. The second controller is added to the existing cluster by using the IP address of the first controller. The second controller imports the cluster name and NTP server information from the first.

Table 3 BCF Controller cluster settings

Hostname	Controller clustering	Existing controller IP	Cluster name	NTP server
bcfctrl01	Start a new cluster	NA	bcf-cluster-01	100.67.10.20
bcfctrl02	Join an existing cluster	100.67.187.201	NA	NA

5.3 Deployment steps

This section walks through each step to set up both controllers and the cluster. The values shown in Table 2 and Table 3 are used.

5.3.1 Deploy the first BCF Controller

The steps to deploy the first BCF Controller are as follows:

1. Connect to the console of the first BCF Controller. The login prompt displays.

```
Big Cloud Fabric 4.6.0 (bcf-4.6.0 #31)
Log in as 'admin' to configure

controller login: _
```

Figure 23 BCF login screen

2. Log in as admin (no password). **Do you accept the EULA for this product? (Yes/No/View)** displays.
3. Review the contents of the EULA if desired, and enter **Yes** to continue.
4. Enter and confirm the **Emergency recovery user password** for the controller. The screen appears as shown:

```
This product is governed by an End User License Agreement (EULA).
You must accept this EULA to continue using this product.

You can view this EULA by typing 'View', or from our website at:
http://www.bigswitch.com/eula

Do you accept the EULA for this product? (Yes/No/View) [Yes] >

Running system pre-check


Finished system pre-check


Starting first-time setup


Local Node Configuration
------------------------

Emergency recovery user password >
Emergency recovery user password (retype to confirm) >
```

Figure 24 EULA and Emergency recovery password prompts

5. At the **Hostname>** prompt, enter the first controller's hostname, **bcfctrl01.**
6. Under **Management network options:**, select **[1] IPv4 only**.
7. Enter the values from Table 2 at the corresponding prompts, for example:
    a. **IPv4 address > 100.67.187.201**
    b. **IPv4 prefix length > 24**
    c. **IPv4 gateway (Optional) > 100.67.187.254**
    d. **DNS server 1 (Optional) > 100.67.189.33**
    e. **DNS server 2 (Optional) >** not used in this example
    f. **DNS search domain (Optional) > dell.local**

After completing the steps above, the screen appears as shown:

```
Hostname > bcfctrl01

Management network options:

[1] IPv4 only
[2] IPv6 only
[3] IPv4 and IPv6

> 1
IPv4 address [0.0.0.0/0] > 100.67.187.201
IPv4 prefix length [24] > 24
IPv4 gateway (Optional) > 100.67.187.254
DNS server 1 (Optional) > 100.67.189.33
DNS server 2 (Optional) >
DNS search domain (Optional) > dell.local
```

Figure 25 Hostname and OOB management network settings

8. Under Controller cluster options:, select **[1] Start a new cluster**.
9. Enter the cluster name, **bcf-cluster-01.**
10. The next prompt is **Cluster description (Optional).** A description is not used in this example.
11. At the **Cluster administrator password** prompt, enter a password and retype to confirm.

```
Controller Clustering
---------------------

Controller cluster options:

[1] Start a new cluster
[2] Join an existing cluster

> 1
Cluster name > bcf-cluster-01
Cluster description (Optional) >
Cluster administrator password >
Cluster administrator password (retype to confirm) >
```

Figure 26 Create a new cluster

12. Under **NTP server options:**, select your preferred option. In this example, **[2] Use Custom NTP servers** is selected, and the **NTP server 1** address used is **100.67.10.20. NTP server 2 (Optional)** is not used in this example.

```
System Time
-----------

Default NTP servers:

 -  0.bigswitch.pool.ntp.org
 -  1.bigswitch.pool.ntp.org
 -  2.bigswitch.pool.ntp.org
 -  3.bigswitch.pool.ntp.org

NTP server options:

[1] Use default NTP servers
[2] Use custom NTP servers

[1] > 2
NTP server 1 > 100.67.10.20
```

Figure 27 NTP server selection

13. A summary of the configuration settings displays. Review the settings and select **[1] Apply settings**.

```
Menu
----

Please choose an option:

[ 1] Apply settings
[ 2] Reset and start over
[ 3] Update Recovery Password     (*****)
[ 4] Update Hostname              (bcfctrl01)
[ 5] Update IP Option             (IPv4 only)
[ 6] Update IPv4 Address          (100.67.187.201)
[ 7] Update IPv4 Prefix Length    (24)
[ 8] Update IPv4 Gateway          (100.67.187.254)
[ 9] Update DNS Server 1          (100.67.189.33)
[10] Update DNS Server 2          (<none>)
[11] Update DNS Search Domain     (dell.local)
[12] Update Cluster Option        (Start a new cluster)
[13] Update Cluster Name          (bcf-cluster-01)
[14] Update Cluster Description   (<none>)
[15] Update Admin Password        (*****)
[16] Update NTP Option            (Use custom NTP servers)
[17] Update NTP Server 1          (100.67.10.20)
[18] Update NTP Server 2          (<none>)

[1] >
```

Figure 28 Configuration summary – first BCF Controller

14. The settings are applied and the message **First-time setup is complete!** displays.

```
[1] > 1
[Stage 1] Initializing system
[Stage 2] Configuring controller
  Waiting for network configuration
  IP address on bond0 is 100.67.187.201
  Generating cryptographic keys
[Stage 3] Configuring system time
  Initializing the system time by polling the NTP server:
    100.67.10.20
[Stage 4] Configuring cluster
  Cluster configured successfully.
  Current node ID is 13684
  All cluster nodes:
    Node 13684: fe80::1618:77ff:fe5b:3cc3:6642

First-time setup is complete!

Press enter to continue >
```

Figure 29 Configuration settings applied

15. Press **Enter**. The controller hostname and login prompt displays.

```
Big Cloud Fabric 4.6.0 (bcf-4.6.0 #31)
Log in as 'admin' to configure

bcfctrl01 login:
```

Figure 30 Controller login prompt

5.3.2 Deploy the second BCF Controller

Setting up the second controller is similar to the first, except that the second controller joins the existing cluster configured on the first controller.

1. Connect to the console of the second BCF Controller. The login prompt displays as shown in Figure 23 in the previous section.
2. Log in as **admin** (no password), accept the EULA, and provide/confirm the **Emergency recovery user password.**
3. At the **Hostname>** prompt, enter the second controller's hostname, **bcfctrl02.**
4. Under Management network options:, select [1] IPv4 only.
5. Enter the values from Table 2 at the corresponding prompts, for example:
    a. **IPv4 address > 100.67.187.202**
    b. **IPv4 prefix length > 24**
    c. **IPv4 gateway (Optional) > 100.67.187.254**
    d. **DNS server 1 (Optional) > 100.67.189.33**
    e. **DNS server 2 (Optional) >** not used in this example
    f. **DNS search domain (Optional) > dell.local**

After completing the steps above, the screen appears as shown:

```
Hostname > bcfctrl02

Management network options:

[1] IPv4 only
[2] IPv6 only
[3] IPv4 and IPv6

> 1
IPv4 address [0.0.0.0/0] > 100.67.187.202
IPv4 prefix length [24] > 24
IPv4 gateway (Optional) > 100.67.187.254
DNS server 1 (Optional) > 100.67.189.33
DNS server 2 (Optional) >
DNS search domain (Optional) > dell.local
```

Figure 31 Management network configuration – second BCF Controller

Next, the **Controller Clustering** section is displayed.

6. Under **Controller cluster options:**, select **[2] Join an existing cluster**.
7. Enter the **Existing controller address, 100.67.187.201**. This is the IP address of the first controller.
8. Enter the **Cluster administrator password** previously configured on the first controller and retype to confirm.

```
Controller Clustering
---------------------

Controller cluster options:

[1] Start a new cluster
[2] Join an existing cluster

> 2
Existing controller address > 100.67.187.201
Cluster administrator password >
Cluster administrator password (retype to confirm) >
```

Figure 32 Joining an existing cluster

9. The configuration settings summary for the second controller displays. Review the settings and select **[1] Apply settings**.

```
Menu
----

Please choose an option:

[ 1] Apply settings
[ 2] Reset and start over
[ 3] Update Recovery Password      (*****)
[ 4] Update Hostname               (bcfctrl02)
[ 5] Update IP Option              (IPv4 only)
[ 6] Update IPv4 Address           (100.67.187.202)
[ 7] Update IPv4 Prefix Length     (24)
[ 8] Update IPv4 Gateway           (100.67.187.254)
[ 9] Update DNS Server 1           (100.67.189.33)
[10] Update DNS Server 2           (<none>)
[11] Update DNS Search Domain      (dell.local)
[12] Update Cluster Option         (Join an existing cluster)
[13] Update Existing Controller    (100.67.187.201)
[14] Update Admin Password         (*****)

[1] >
```

Figure 33 Configuration summary on second BCF Controller

10. Once the settings are applied, the screen appears as shown in Figure 34. The message **Please verify that: Secure control plane is NOT configured** displays. By default, the secure control plane is not configured.

```
[1] > 1
[Stage 1] Initializing system
[Stage 2] Configuring controller
  Waiting for network configuration
  IP address on bond0 is 100.67.187.202
  Generating cryptographic keys

Please verify that:

  Secure control plane is NOT configured.

You can verify the above by running "show secure control plane"
on the existing controller 100.67.187.201.

Options:

[1] Continue connecting (the above info is correct)
[2] Cancel and review parameters

>
```

Figure 34 Applying settings on second controller

Note: The secure control plane is a feature where certificates are issued by a trusted Certificate Authority (CA) to each controller and switch that will participate in the fabric. When enabled, controllers and switches cannot join the fabric without a valid certificate from the trusted CA. The secure control plane is not configured by default. See the *BCF User Guide* for more information on this feature.

11. Select option **[1] Continue connecting (the above info is correct)** to proceed. When done, the message **First-time setup is complete!** is displayed.

```
Options:

[1] Continue connecting (the above info is correct)
[2] Cancel and review parameters

> 1

[Stage 3] Configuring system time
  Initializing the system time by polling the NTP server:
    100.67.10.20
[Stage 4] Configuring cluster
  Cluster configured successfully.
  Current node ID is 20702
  All cluster nodes:
    Node 13684: fe80::1618:77ff:fe5b:3cc3:6642
    Node 20702: fe80::b283:feff:fed6:3e8f:6642

First-time setup is complete!

Press enter to continue >
```

Figure 35 Settings applied on second controller

12. Press the **Enter** key. The controller login screen for the second controller displays.
13. Log in as **admin**. The command prompt displays and indicates this controller is in the **standby** role.

```
Big Cloud Fabric 4.6.0 (bcf-4.6.0 #31)
Log in as 'admin' to configure

bcfctrl02 login: admin
Password:
Login: admin, on Fri 2018-03-02 16:49:50 UTC, from localhost
Last login:   on Fri 2018-03-02 16:46:21 UTC, from localhost
standby bcfctrl02> _
```

Figure 36 Login prompt and command prompt on second (standby) controller

Note: The standby controller is read only. Configuration commands made from the command line must be run on the active controller.

5.3.3 Configure the cluster virtual IP address

As a best practice, set a virtual IP (VIP) address for the cluster. This allows you to connect to the management port of the active node using an IP address that does not change even if the active controller fails over and the role of the standby controller changes to active.

To configure the cluster VIP address:

1. Log in to the console of the active controller locally or remotely using secure shell (SSH).
2. Use the set of commands shown in Figure 37 to set the cluster VIP address.

```
bcfctrl01> enable
bcfctrl01# configure
bcfctrl01(config)# controller
bcfctrl01(config-controller)# virtual-ip 100.67.187.200
```

Figure 37 Setting the cluster virtual IP address

3. To verify the cluster settings, enter the **`show controller`** command from either the active or standby controller. Verify that the **Cluster Virtual IP** address is correct and that **Redundancy status** is **redundant**.

```
bcfctrl01(config-controller)# show controller
Cluster Name          : bcf-cluster-01
Cluster Virtual IP    : 100.67.187.200
Redundancy Status     : redundant
Last Role Change Time : 2017-10-16 18:38:37.414000 UTC
Failover Reason       : Changed connection state: cluster configuration changed
Cluster Uptime        : 2 hours, 22 minutes
# IP             @ State   Uptime
-|--------------|-|-------|-------------------|
1 100.67.187.201 * active  2 hours, 22 minutes
2 100.67.187.202   standby 50 minutes
```

Figure 38 Show controller command output

5.3.4 Access the BCF GUI

The BCF GUI is accessible from a browser by navigating to the VIP address of the cluster.

Note: For more information, see the *Big Cloud Fabric GUI Guide*.

1. Enter the cluster VIP address in a web browser. You are redirected to a secure login page.

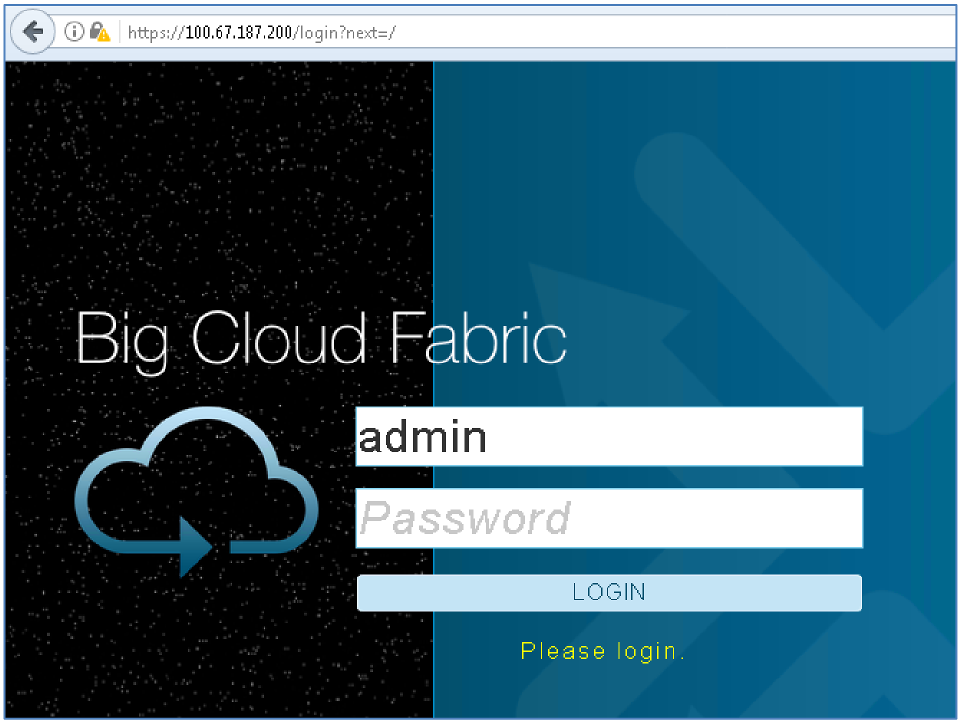


Figure 39 Connecting to the BCF GUI

Note: The BCF Controller uses a self-signed certificate by default. See the *Big Cloud Fabric User Guide* to install a certificate from a trusted CA.

2. Log in as **admin** using the password created during controller setup. The BCF dashboard displays similar to the following:

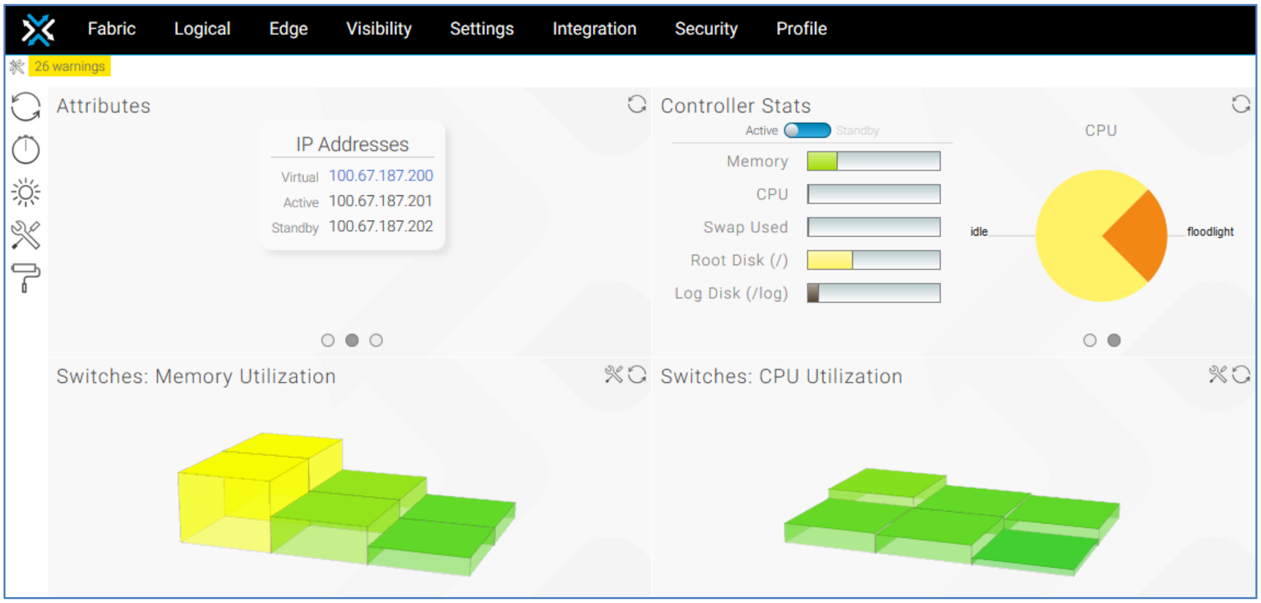


Figure 40 BCF dashboard

5.4 Switch deployment

This section covers BCF switch deployment. Before proceeding, make sure all leaf switches, spine switches, and BCF Controllers are physically connected to the p-switch network as shown in Figure 19 and Figure 20.

5.4.1 Zero Touch Fabric overview

Big Switch Zero Touch Fabric (ZTF) uses the Open Networking Install Environment (ONIE) boot loader to automate switch installation and configuration. ONIE makes deploying many switches in a data center easier and less prone to errors. The ZTF process uses ONIE to automatically install the correct version of Switch Light OS on each switch when the switch is powered on and connected to the BCF Controller.

The Dell EMC Networking switches used in this example do not have BCF Switch Light OS installed initially. In the following steps, the BCF Controller deploys the OS to the leaf and spine fabric switches.

Switch Light OS is a complete SDN operating system based on Open Network Linux (ONL) and is bundled with the BCF software distribution. This ensures that the software running on the switch is compatible with the controller software version.

Figure 41 provides an overview of the switch registration and OS deployment steps.

Note: For more information about this process, see Chapter 4 of the *Big Cloud Fabric User Guide*.

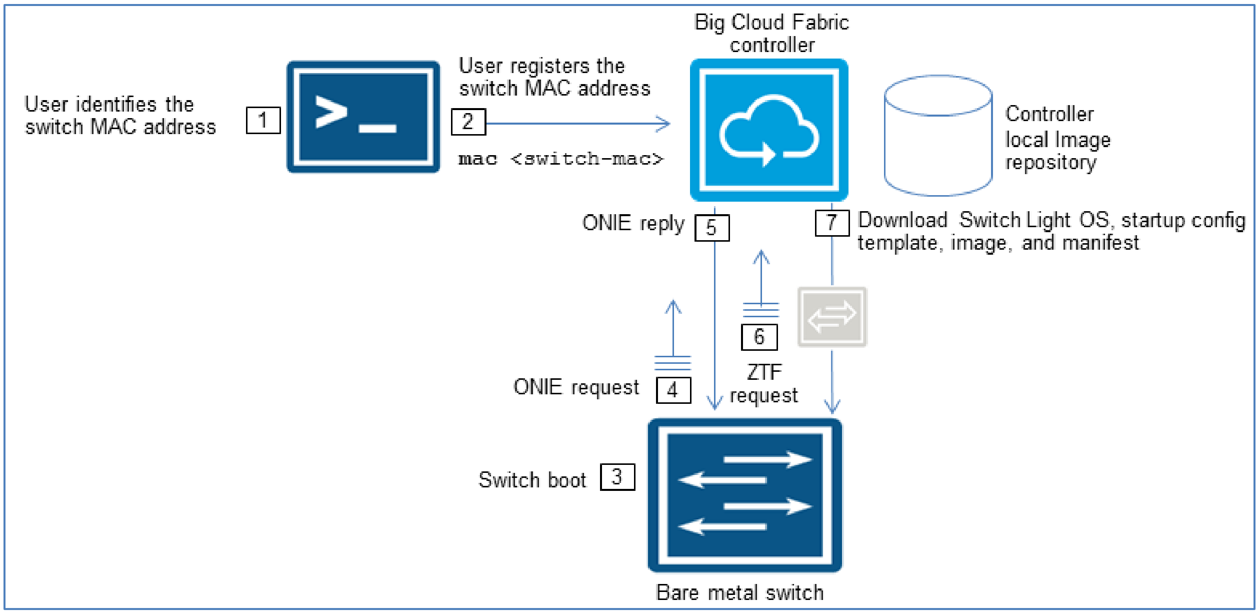


Figure 41 BCF switch registration and OS deployment workflow

The switch registration and OS deployment steps are listed in Table 4.

Table 4 BCF switch provisioning summary

Step	Description
1	Collect switch MAC address from the Dell EMC express service tag on the switch
2	Register switch MAC address using the BCF GUI or CLI
3	Switch is rebooted or power cycled
4	The switch ONIE loader generates an IPv6 neighbor discovery message on the local network segment
5	If the MAC is registered, the controller responds to the ONIE request from the switch and instructs it to download the Switch Light OS loader to begin installation
6	After installing the Switch Light OS loader and rebooting, the loader broadcasts a ZTF request
7	The ZTF server on the active BCF Controller sends the Switch Light OS image, manifest, and startup configuration to the switch

The startup configuration file provided to each switch by the BCF Controller includes the following information:

- Hostname
- Switch MAC address
- Controller IPv6 addresses
- NTP, logging, and Simple Network Management Protocol (SNMP) configuration

5.4.2 Collect switch MAC addresses

Record the MAC address of each leaf and spine switch. The MAC address is printed on the plastic express service tag labeled “EST” on each switch. The tag is located on the front of Z9100-ON switches, and the back of S4048-ON switches as shown in Figure 42.

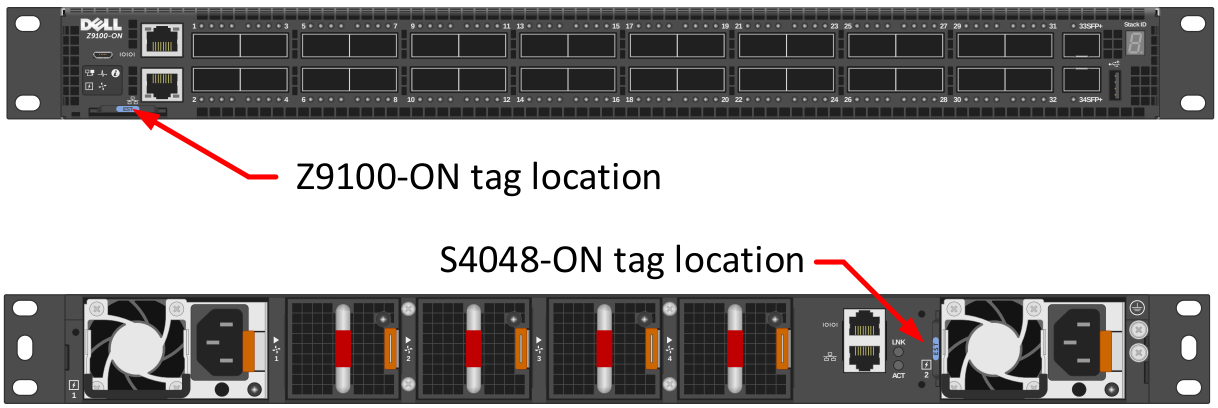


Figure 42 EST tag location on Z9100-ON and S4048-ON switches

5.4.3 Provision switches in the BCF Controller

Table 5 lists the MAC addresses, switch names, roles, and leaf groups used for provisioning in this section.

Table 5 Switch provisioning details

Model	MAC address	Switch name	Fabric role	Leaf group
S4048-ON	f4:8e:38:20:37:29	Leaf1	Leaf	Rack1
S4048-ON	f4:8e:38:20:54:29	Leaf2	Leaf	Rack1
S4048-ON	64:00:6a:e4:cc:3e	Leaf3	Leaf	Rack2
S4048-ON	64:00:6a:e7:24:14	Leaf4	Leaf	Rack2
Z9100-ON	4c:76:25:e7:41:40	Spine1	Spine	NA
Z9100-ON	4c:76:25:e7:3b:40	Spine2	Spine	NA

The BCF Controller CLI or GUI may be used to provision the switches. The GUI is used in this example.

1. Using a browser, navigate to the VIP address of the BCF Controller cluster and log in.
2. In the BCF GUI, navigate to **Fabric > Switches**.
3. To provision the leaf switches, click the + icon to open the **Provision Switch** dialog box.
4. Complete the fields as follows:
   a. **Name** - enter the switch name to use for the leaf switch as listed in Table 5.
   b. **MAC Address** - enter or paste the MAC address corresponding to the leaf switch.

Note: Depending on the state of the switch, its MAC address may appear in the drop down menu and may be selected from the menu if present.

c. **Fabric Role -** select the **Leaf** box.
d. **Leaf Group -** enter the appropriate **Leaf Group** for the switch as listed in Table 5.
e. Defaults are used for the remaining items.

After information for the first leaf switch is entered, the **Provision Switch** dialog box appears as shown:

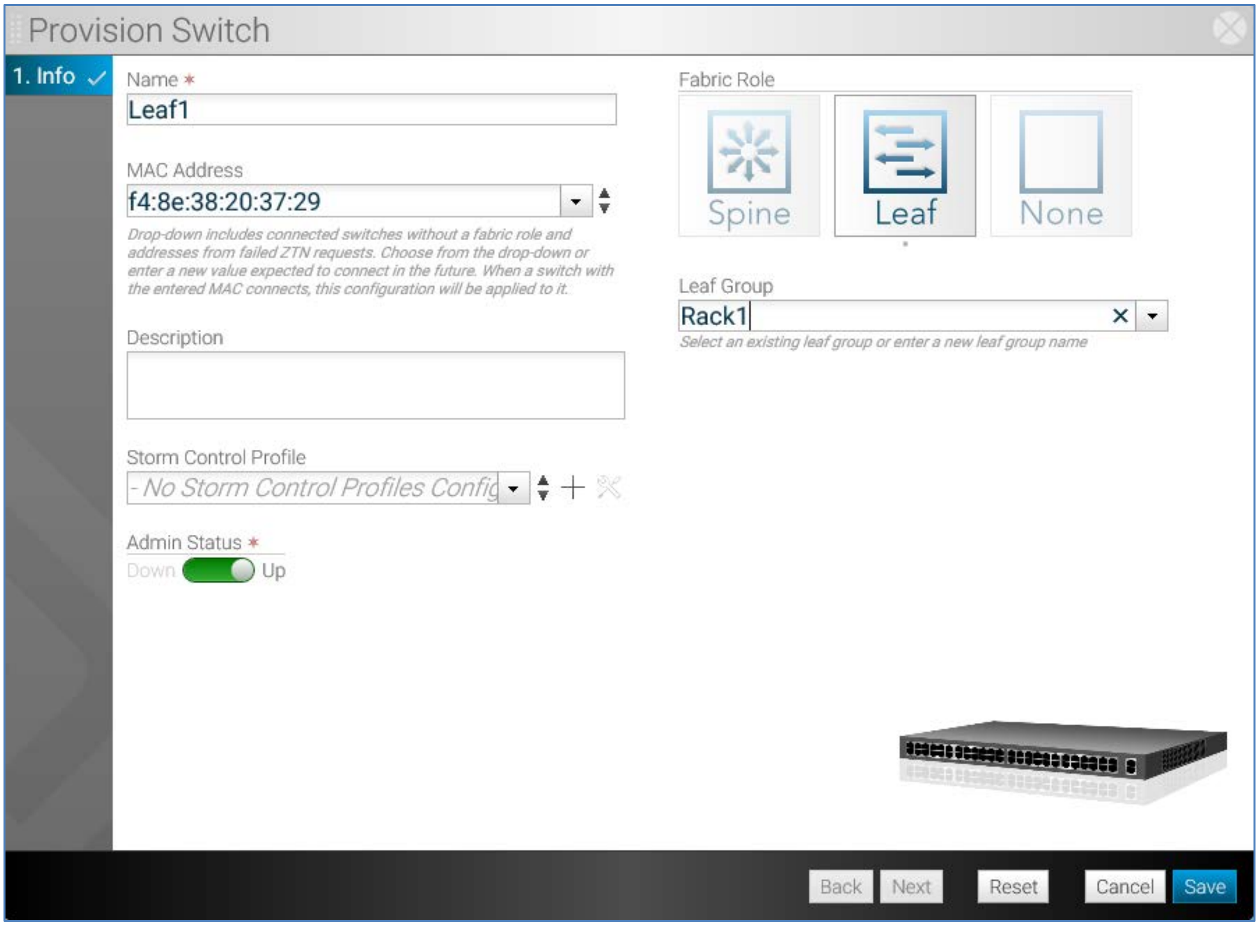


Figure 43 Provision the first leaf switch

f. Click **Save** and repeat steps 3 and 4 for the remaining leaf switches.

5. To provision the spine switches, click the + icon to open the **Provision Switch** dialog box:
6. Complete the fields as follows:
   a. **Name** – enter the switch name to use for the spine switch as listed in Table 5.
   b. **MAC Address** – select the MAC address corresponding to the first spine switch from the drop-down box.
   c. **Fabric Role** – select the **Spine** box.
   d. Defaults are used for the remaining items.

After information for the first spine switch is entered, the **Provision Switch** dialog box appears as shown:

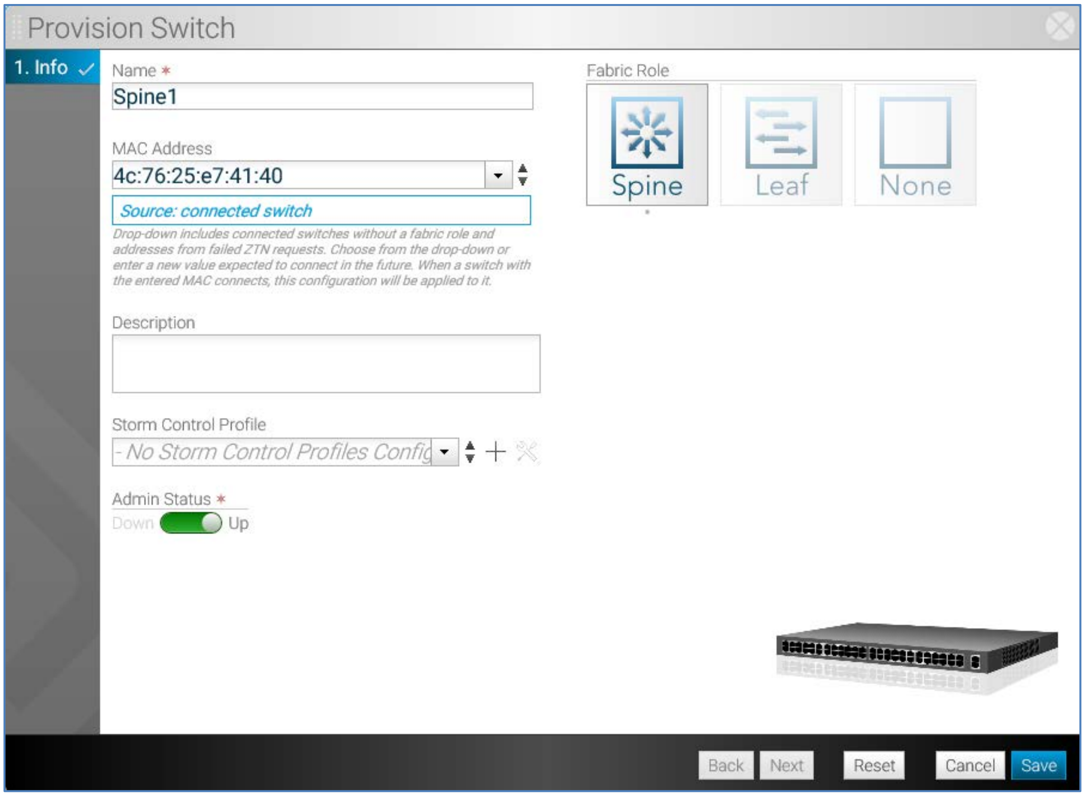


Figure 44 Provision the first spine switch

e. Click **Save** and repeat steps 5 and 6 for the remaining spine switch.

When complete, the list of switches will appear similar to that shown below. The **MAC** address, **Name**, **Fabric Role,** and **Leaf Group** is shown for each switch.

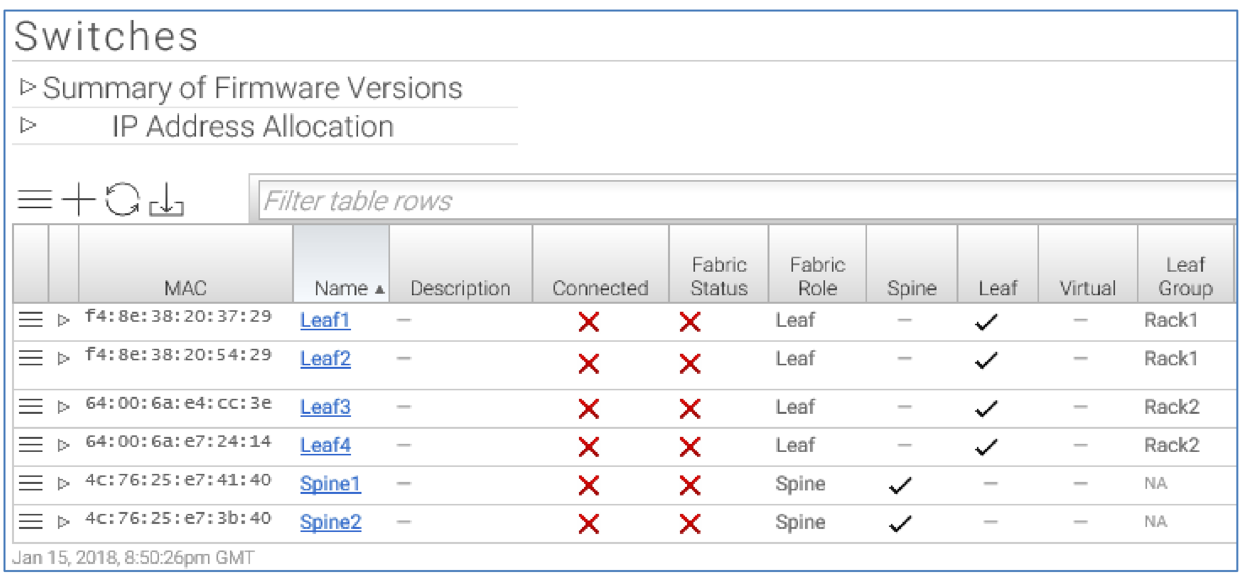


Figure 45 Switches ready for provisioning

5.4.4 Boot switches in ONIE install mode

To place switches in ONIE install mode, do the following on each leaf and spine switch:

1. Power on or reboot the switch.
2. If **press Esc to stop autoboot** is shown during boot, press **Esc**.

Note: Step 2 is required if switches are running the Dell Networking Operating System (DNOS) 9.x.

3. The Grub menu displays.

```
Grub 1.99~rc1 (Dell Inc)
Built by root at ubuntu on Tue_Feb_23_04:09:41_UTC_2016
S4000 Boot Flash Label 3.21.2.3 NetBoot Label 3.21.2.3
+--------------------------------------------------------------------------+
|FTOS                                                                      |
|FTOS-Boot Line Interface                                                  |
|DELL-DIAG                                                                 |
|ONIE                                                                      |
|                                                                          |
|                                                                          |
|                                                                          |
|                                                                          |
|                                                                          |
|                                                                          |
+--------------------------------------------------------------------------+

     Use the ^ and v keys to select which entry is highlighted.
     Press enter to boot the selected OS, 'f' to boot FTOS, 'b' to go to
     BLI, 'o' to boot ONIE, 'd' to boot DELL-DIAG, 'e' to edit the
     commands before booting or 'c' for a command-line.
```

Figure 46 Grub menu on S4048-ON switch

4. In the Grub menu, select **ONIE** and press **Enter.**
5. In the next window, select **ONIE: Install OS** and press **Enter.**

This starts the Switch Light OS installation and configuration process, listed in steps 4-7 of Table 4. Allow a few minutes for the BCF Controller to install the Switch Light OS to each switch.

Optionally, provisioning progress may be monitored at the switch consoles. When provisioning is complete, the console of each switch appears with its hostname and login prompt as shown:

Figure 47 Leaf 1 console view after provisioning

5.4.5 Verify Switch Light OS installation

To verify successful installation of the Switch Light OS, use the BCF Controller GUI to navigate to **Fabric > Switches** to view all switches, MAC addresses, names, connection status, fabric status, and fabric roles.

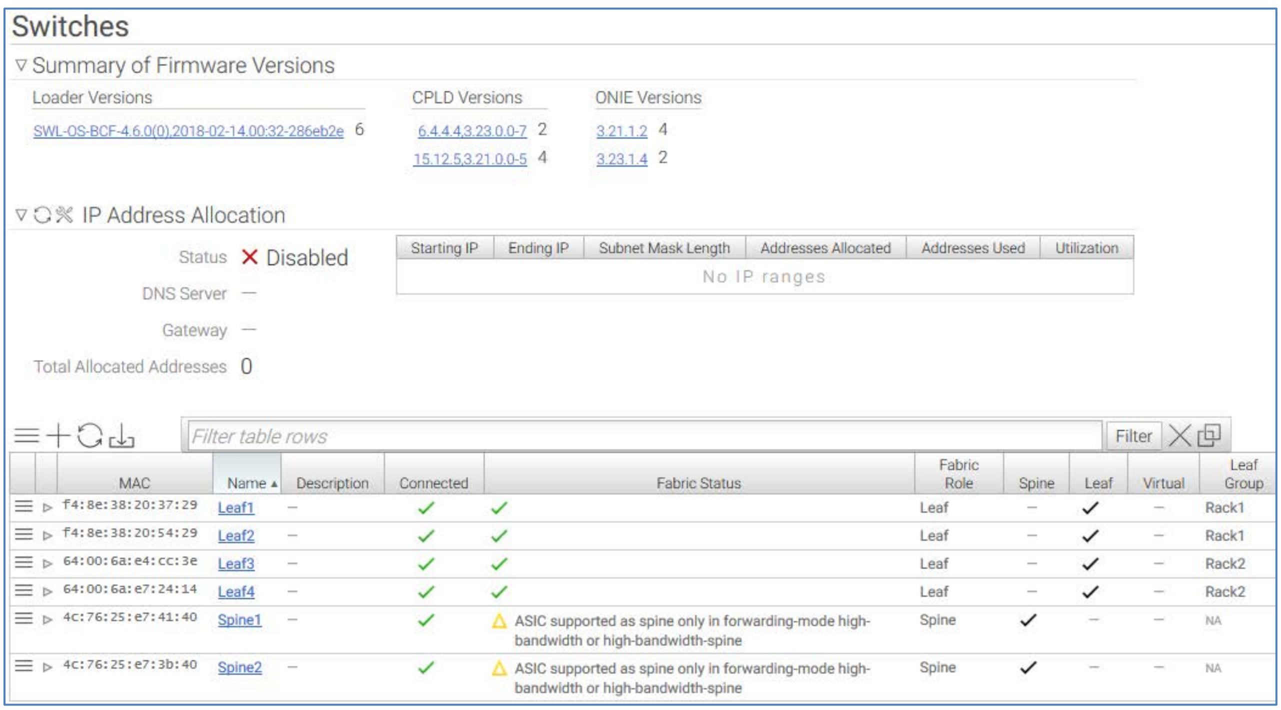


Figure 48 Switch summary in BCF Controller GUI

5.5 Resolve common warnings and errors

Current warnings and errors may be viewed in the BCF Controller GUI by going to **Visibility > Fabric Summar**y, or by clicking on the errors/warnings message in the upper left corner of the GUI.

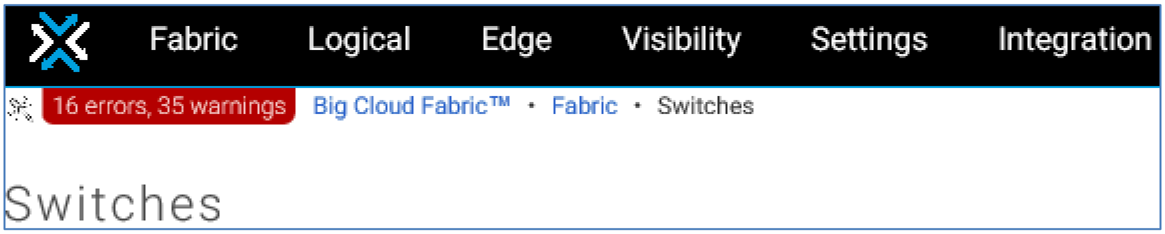


Figure 49 Errors/warnings message in upper left corner of GUI

On the **Fabric Summary** page, errors and warnings may be shown/hidden by selecting/deselecting the category in the left pane.

5.5.1 Suspended Switches

Z9100-ON spine switches may appear under **Suspended Switches** with the message **ASIC supported as spine only in forwarding-mode high-bandwidth or high-bandwidth-spine**.

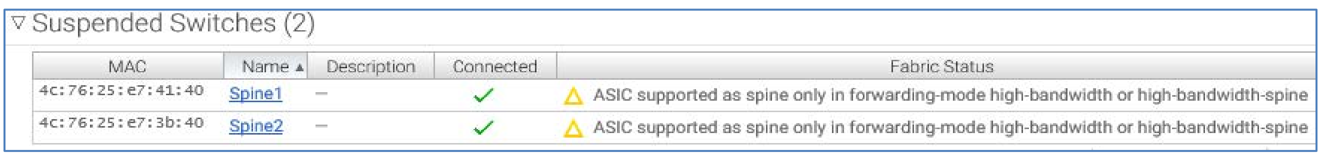


Figure 50 Suspended switches error

Z9100-ON switches are classified as high bandwidth spines in BCF. To set the forwarding mode to high bandwidth spine for these two switches, do the following:

1. Go to **Settings > Fabric Settings** and select the icon.
2. In the left pane of the **Fabric Settings** dialog box, select **Forwarding Mode.**
3. In the right pane, move the **High Bandwidth Spine** slider to the right. All other sliders are moved to the left.
4. Click **Submit.**
5. Return to the **Visibility > Fabric Summary** page and verify there are no suspended switches listed.

5.5.2 Switches with mismatched ONIE and CPLD

Some switches may be listed with mismatched ONIE and/or Complex Programmable Logic Device (CPLD) firmware as shown in Figure 51 and Figure 52.

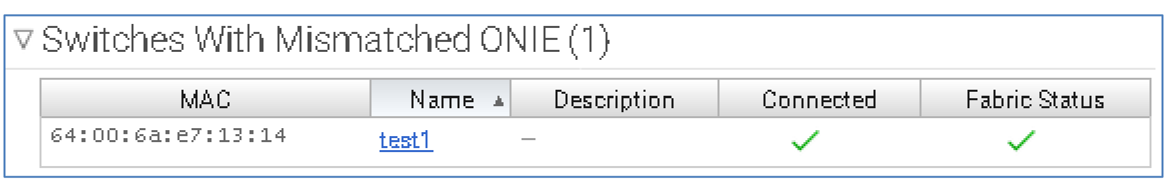


Figure 51 Switch with mismatched ONIE

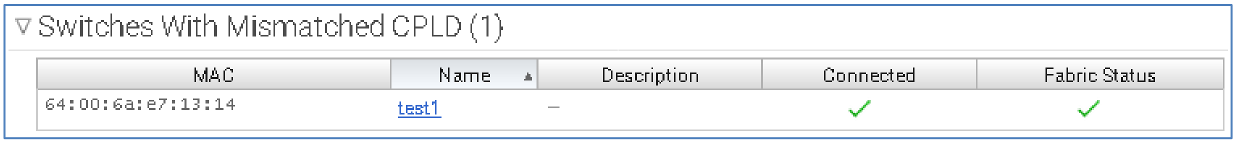


Figure 52 Switch with mismatched CPLD

Resolve switch ONIE mismatches as follows:

1. Scroll down to **Switches With Mismatched ONIE** and click on the switch name. This example uses a single switch named **test1**.
2. In the switch page that opens, select the **Actions** tab.
3. On the left side of the page, select **Manage Firmware.** The **Manage Switch Firmware** dialog box displays.

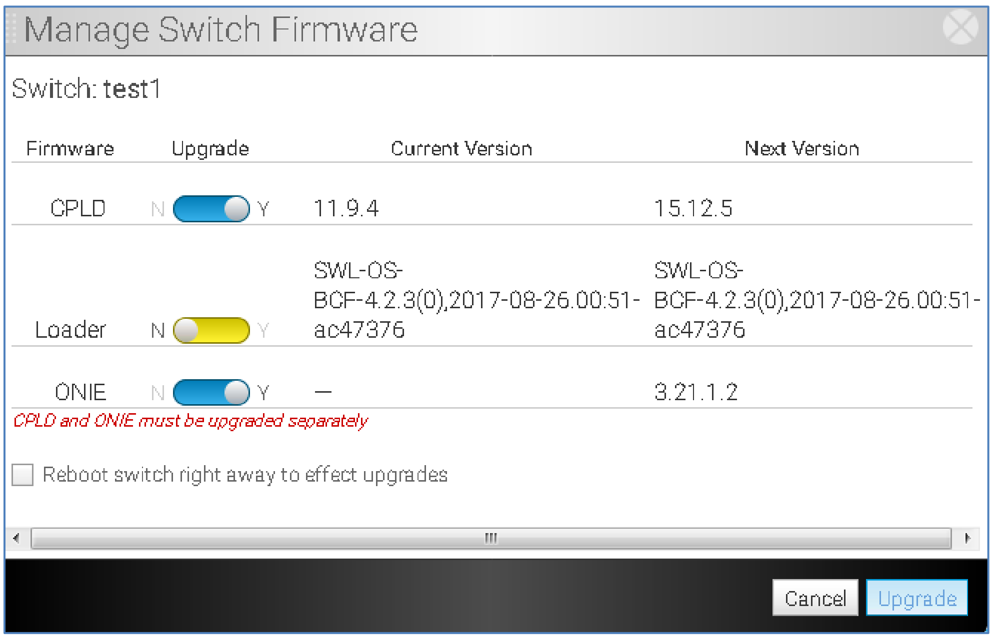


Figure 53 Manage switch firmware dialog box

4. Move the **CPLD** slider to **N**, and the **ONIE** slider to **Y**.

Note: CPLD and ONIE must be upgraded separately. Upgrade ONIE first.

5. Check the **Reboot switch right away** box and click **Upgrade**.
6. The switch reboots and ONIE firmware is updated. This can be observed at the switch console.
7. Repeat for remaining switches listed in the **Switches With Mismatched ONIE** table.
8. Refresh the **Visibility > Fabric Summary** page by clicking the icon to verify all ONIE issues are resolved.

After ONIE mismatches are resolved, resolve switch CPLD mismatches as follows:

1. Scroll down to **Switches With Mismatched CPLD** and click on the switch name.
2. In the switch page that opens, select the **Actions** tab.
3. On the left side of the page, select **Manage Firmware** to open the **Manage Switch Firmware** dialog box shown in Figure 53.
4. Ensure the **CPLD** slider is set to **Y**, and that the other sliders are set to **N**. Check the **Reboot switch right away** box and click **Upgrade**.
5. The switch reboots and CPLD firmware is updated. This can be observed at the switch console. The process may take 10-20 minutes.
6. After the switch has rebooted and CPLD firmware installation is complete, power cycle the switch by removing the power cable(s), waiting until all LEDs are off (5-10 seconds), then reconnecting the power cable(s).
7. Repeat for remaining switches listed in the **Switches With Mismatched CPLD** table.
8. Refresh the **Visibility > Fabric Summary** page by clicking the icon to verify all CPLD issues are resolved.

5.5.3 Switches without management address

Switches communicate with the controller using IPv6 on the p-switch control network. IPv6 addresses are automatically assigned to the fabric switches by the controller, but IPv4 management addresses are required if switches will connect to services that are not configured for IPv6 such as NTP, syslog, and SNMP.

Note: Switches connect to NTP, syslog, and SNMP servers using IPv4 addresses via the p-switch network. These connections are optional. See the *Big Cloud Fabric User Guide* for more information.

The BCF Controller automatically assigns IPv4 management addresses from a defined address pool. Until this pool is configured, **Switches Without Management Addresses** are listed under **Errors** on the **Visibility > Fabric Summary** page as shown in Figure 54.

▽ Switches Without Management Address (6)

MAC	Name ▲	Description	Connected	Fabric Status
f4:8e:38:20:37:29	Leaf1	–	✓	✓
f4:8e:38:20:54:29	Leaf2	–	✓	✓
64:00:6a:e4:cc:3e	Leaf3	–	✓	✓
64:00:6a:e7:24:14	Leaf4	–	✓	✓
4c:76:25:e7:41:40	Spine1	–	✓	✓
4c:76:25:e7:3b:40	Spine2	–	✓	✓

Figure 54 Switches without management address

The IPv4 address pool is configured as follows:

1. In the BCF GUI, go to **Fabric > Switches.**
2. Next to **IP Address Allocation**, click the icon.
3. In the **Configure Switch IP Allocation** dialog box, move the slider to **Enabled**.
4. Click the icon to open the **Create IP Range** dialog box.

Note: The **DNS Server Address** and **Gateway Address** fields are optional and not used in this example.

5. Specify a **Start IP, End IP**, and **Subnet Mask Length** to use for the pool. This example uses the range **192.168.200.1-100** with a subnet mask length of **24** as shown:

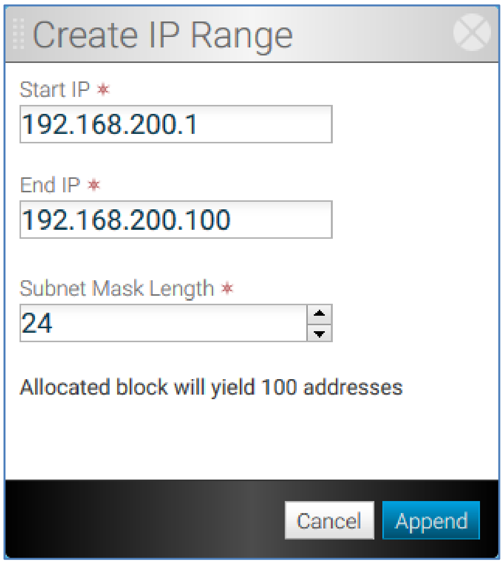


Figure 55 IP address range

6. Click **Append > Submit.**

When complete, the **IP Address Allocation** section of the **Fabric > Switches** page displays similar to Figure 56. Six addresses are used, one for each leaf and spine switch in the topology.

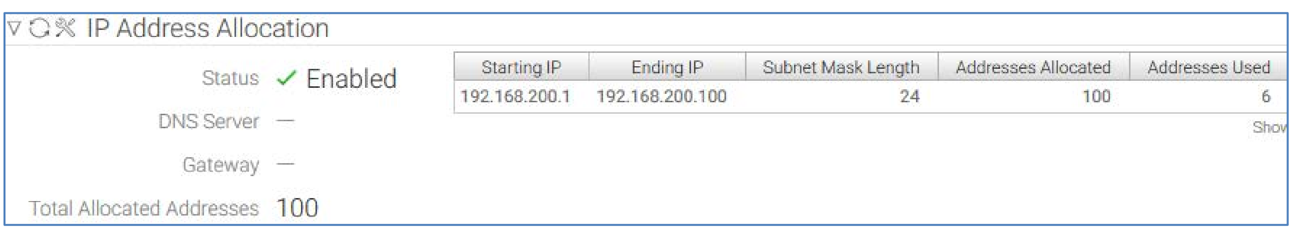


Figure 56 IP address pool configured

This resolves the **Switches Without Management Addresses** errors listed on the **Visibility > Fabric Summary** page.

5.5.4 Leaf interfaces not in interface groups

Interface groups can provide active-active load balancing and failover among members of the group. Connected leaf edge ports that are not configured in groups display in the **Leaf Interfaces Not in Interface Groups** section.

The VMware integration process covered in Section 7 automatically configures the interface groups and resolves these warnings.

Leaf Interfaces Not in Interface Groups (26)

Switch ▲	Switch MAC	Interface Name	Description	Status
Leaf1	f4:8e:38:20:37:29	ethernet1	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet16	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet2	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet3	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet4	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet5	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet6	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet7	–	✓ Up
Leaf1	f4:8e:38:20:37:29	ethernet8	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet1	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet16	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet2	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet3	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet4	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet5	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet6	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet7	–	✓ Up
Leaf2	f4:8e:38:20:54:29	ethernet8	–	✓ Up
Leaf3	64:00:6a:e4:cc:3e	ethernet20	–	✓ Up
Leaf3	64:00:6a:e4:cc:3e	ethernet21	–	✓ Up

Figure 57 Leaf interfaces not in interface groups

5.6 BCF validation commands from the CLI

The following commands help validate the fabric configuration. Run these commands from the active or standby controller.

> **Note:** See the *Big Cloud Fabric CLI Reference Guide* for a complete listing of commands.

5.6.1 show fabric error

The **`show fabric error`** command displays fabric errors. These items also appear in the GUI on the **Visibility > Fabric Summary** page under **Errors**. This command should return **None** at this point as shown below.

```
bcfctrl01> show fabric error
None.
```

> **Note:** To see items shown in the GUI on the **Visibility > Fabric Summary** page under **Warnings**, run the command **`show fabric warning.`** At this stage of deployment, there are warnings shown for interfaces not configured in interface groups as shown in section 5.5.4.

5.6.2 show link

The **`show link`** command returns all inter-switch links that are operational. This includes leaf-to-leaf (peer links) and leaf-spine links. Links are discovered using Link Layer Discovery Protocol (LLDP).

For the topology used in this deployment, shown in Figure 12, there are twelve inter-switch links: four peer links and eight leaf-spine links. All twelve inter-switch links should appear in the output as shown below:

```
bcfctrl01> show link
#  Switch Name IF Name    Switch Name IF Name    Link Type
--|-----------|----------|-----------|----------|----------|
1  Leaf1       ethernet53 Leaf2       ethernet53 peer
2  Leaf1       ethernet54 Leaf2       ethernet54 peer
3  Leaf3       ethernet53 Leaf4       ethernet53 peer
4  Leaf3       ethernet54 Leaf4       ethernet54 peer
5  Spine1      ethernet1  Leaf1       ethernet49 leaf-spine
6  Spine1      ethernet2  Leaf2       ethernet49 leaf-spine
7  Spine1      ethernet3  Leaf3       ethernet49 leaf-spine
8  Spine1      ethernet4  Leaf4       ethernet49 leaf-spine
9  Spine2      ethernet1  Leaf1       ethernet50 leaf-spine
10 Spine2      ethernet2  Leaf2       ethernet50 leaf-spine
11 Spine2      ethernet3  Leaf3       ethernet50 leaf-spine
12 Spine2      ethernet4  Leaf4       ethernet50 leaf-spine
```

5.6.3 show switch *switch name* interface

The command **`show switch`** *`switch name`* **`interface`** is used to check operational status of switch interfaces. Like most commands, it is run from the controller console instead of the switch console. The switch name is specified in the command.

```
bcfctrl01> show switch Spine1 interface

#  Switch IF Name    IF Type Phy. State Op. State LACP State Curr Features
--|------|----------|-------|----------|---------|----------|--------------|
1  Spine1 ethernet1  leaf    up         up        inactive   fiber, 40gb-fd
2  Spine1 ethernet2  leaf    up         up        inactive   fiber, 40gb-fd
3  Spine1 ethernet3  leaf    up         up        inactive   fiber, 40gb-fd
4  Spine1 ethernet4  leaf    up         up        inactive   fiber, 40gb-fd
```

With BCF, the Z9100-ON interfaces are automatically configured for 40GbE as shown when connected to 40GbE interfaces on S4048-ON leaf switches.

Note: The command output above is truncated; the remaining Spine1 interfaces are down.

6 VMware vSphere deployment

VMware vSphere is a critical component of the deployment of the SDDC. This section provides an overview of the vSphere configuration settings used for this deployment. Design decisions follow guidance outlined in VVD 4.1. Big Switch Networks recommends certain vSphere settings for integration with BCF, and those are included in this section where applicable.

Note: For detailed deployment instructions, refer to the *vSphere Installation and Setup* guide. Software versions used in this guide are listed in Appendix C.1.

6.1 Deploy and configure ESXi

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

All hosts are connected to the OOB management network shown in Figure 17 and Figure 18.

6.1.1 Deployment

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 on each host 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.

Note: Installing ESXi to redundant SD cards is preferred if vSAN clusters will be enabled. See VMware KB Article 2129050 for information on mixing vSAN and non-vSAN disks with the same storage controller.

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

6.1.2 Initial configuration

1. Via the iDRAC virtual console, log in to the ESXi console and select **Configure Management Network > Network Adapters.**
2. Select the correct vmnic for the server's OOB 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. Under **Configure Management Network > DNS Configuration,** enter the IP address of your DNS server(s). For the ESXi system hostname, enter a fully qualified domain name (FQDN), e.g. comp101.dell.local.

Note: BCF requires all ESXi hosts have unique hostnames and that the domain name field not be empty. BCF recommends hosts use FQDNs.

5. Press **Esc** to exit and answer **Y** to apply the changes.
6. From the ESXi main menu, select **Test Management Network** and verify the tests are successful.
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 and repeat the steps above for the remaining hosts.

6.2 vCenter Server deployment and design

In this deployment, two vCenter Server appliances are deployed as recommended in VVD 4.1:

- mgmtvc01.dell.local – supports the ESXi hosts that comprise the Management cluster
- compvc01.dell.local – supports the ESXi hosts that comprise the Compute-Edge and Compute clusters

Each vCenter Server is deployed using the Linux-based vCenter Server Appliance (VCSA). The VCSA is a prepackaged VM that is easy to deploy and supports up to 2000 hosts or 35,000 VMs.

Each vCenter Server is deployed with an external Platform Services Controller (PSC). A vCenter single sign-on (SSO) domain is created when the first PSC is deployed. When the second PSC is deployed with the second vCenter Server, it is joined to the first SSO domain. With both PSCs joined to a single SSO domain, the controllers function as a cluster and provide authentication to all components, and infrastructure data between the PSCs is replicated.

vCenter Servers and PSCs are initially deployed to local datastores on the management ESXi hosts listed in Table 6 (vSANs are configured later in Section 9 of this guide).

The default network, VM Network, is the OOB Management network and is selected during deployment. The appliances are assigned static IP addresses and hostnames during installation and include valid DNS registrations with reverse lookups.

Table 6 shows the configuration information for the two vCenter Servers and their associated PSCs.

Table 6 vCenter Servers and PSCs

Component	Deployment target	Network	System name (FQDN)	Static IP address
Management PSC	mgmt01.dell.local	VM Network	mgmtpsc.dell.local	100.67.187.170
Management vCenter	mgmt02.dell.local	VM Network	mgmtvc01.dell.local	100.67.187.171
Compute PSC	mgmt03.dell.local	VM Network	comppsc.dell.local	100.67.187.172
Compute vCenter	mgmt04.dell.local	VM Network	compvc01.dell.local	100.67.187.173

Additional settings used in this deployment:

- The Management and Compute PSCs are both given the same site name.
- The Management vCenter is built using the small appliance size (up to 100 hosts/1000 VMs).
- The Compute vCenter is built using the medium appliance size (up to 400 hosts/4000 VMs).

Note: See the *vSphere Installation and Setup* guide for vCenter sizing information.

After vCenter Servers and PSCs are deployed, data centers and clusters are created using the VMware Web Client. The vSphere vCenter Servers are identified by icons. A data center is created in each vCenter. The clusters are created in the data centers, and hosts are added to the clusters.

Figure 58 shows the **Hosts and Clusters** tab in the Web Client **Navigator** pane for this deployment after completing this section.

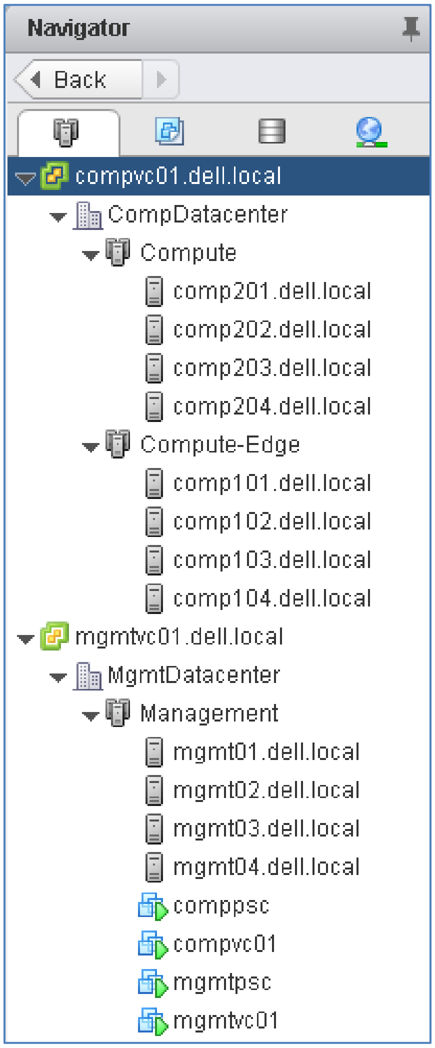


Figure 58 VMware Web Client - Hosts and Clusters tab

The data center, cluster, and host names shown are used for the remainder of this guide.

Note: The two vCenter VMs and two PSC VMs at the bottom of Figure 58 appear in the tree after their respective hosts are added to the Management cluster.

6.3 Virtual network design

When building the VMware virtual network counterpart to BCF, a few principles are followed to ensure that the design meets a diverse set of requirements while keeping operational complexity to a minimum:

- Different network services are assigned to different VLANs to achieve greater security and performance
- Network I/O control and traffic shaping is used to allocate bandwidth to critical workloads
- VMXNET3 virtual NIC drivers are used on all VMs
- The MTU size is set to 9000 bytes (jumbo frames) on all physical and virtual ports for best performance

VLANs and IP addresses used in this deployment are listed in Table 7. The L2/L3 boundary is at the leaf switches, and each VLAN is contained within a single rack.

Table 7 Production VLANs and IP addresses

Cluster	Function	VLAN ID	Network	Gateway
Management	vMotion	1612	172.16.12.0/24	172.16.12.254
	vSAN	1613	172.16.13.0/24	None
	VXLAN	1614	172.16.14.0/24	172.16.14.254
Compute-Edge	vMotion	1622	172.16.22.0/24	172.16.22.254
	vSAN	1623	172.16.23.0/24	None
	VXLAN	1624	172.16.24.0/24	172.16.24.254
Compute	vMotion	1632	172.16.32.0/24	172.16.32.254
	vSAN	1633	172.16.33.0/24	None
	VXLAN	1634	172.16.34.0/24	172.16.34.254

The vMotion VLANs have gateways configured, enabling VMs to be migrated across racks as needed.

Each vSAN is on an isolated VLAN for best performance as recommended in VVD 4.1. No gateways are configured and vSAN traffic is contained within each rack.

Note: VXLAN/NSX networks are configured in Section 10 of this guide.

The IP address and VLAN numbering scheme is similar to that used in VVD 4.1. Subnet-to-VLAN mapping uses the RFC1918 defined private IP address space 172.16.0.0/12 as the base for all addresses. The second and third octets represent the VLAN ID.

For example, 172.16.12.0/24 has an associated VLAN ID of 1612. This algorithm ensures that each subnet and VLAN pairing is unique and easily identified.

6.3.1 vDS configuration

This section provides details regarding VMware vSphere Distributed Switch (vDS) configuration for the vMotion and vSAN networks. One vDS is created for each cluster in the data center that contains the cluster.

Table 8 vDS names

Data center	vDS Name
MgmtDatacenter	vDS-Mgmt
CompDatacenter	vDS-CompEdge
CompDatacenter	vDS-Compute

The load balancing setting for all non-VXLAN port groups, regardless of the vDS, is **Route Based on Physical NIC Load** as recommended in VVD 4.1.

Route Based on Physical NIC Load tests vDS uplinks every 30 seconds. If an uplink's load exceeds 75 percent of usage, the port ID of the virtual machine with the highest usage is moved to a different uplink.

Note: The load balancing/teaming policy for VXLAN is set to **Route Based on Source ID**. This is configured in Section 10.5 of this document.

Note: Big Switch Networks supports the load balancing settings above, but recommends **IP Hash** for load balancing. **IP Hash** provides the best performance during BCF controller upgrades. **Route Based on Physical NIC Load** (and Route **Based on Source ID** for VXLAN) are used for the deployment in this guide as recommended by VVD.

6.3.1.1 vDS-Mgmt configuration

Configuration settings used for vDS-Mgmt are listed in Table 9.

Table 9 vDS-Mgmt settings

Distributed switch name	Version	Number of uplinks	Network I/O control	Discovery Protocol Type / Operation	MTU (Bytes)
vDS-Mgmt	6.5.0	2	Enabled	LLDP / Both	9000 Bytes

Note: For the discovery protocol type, BCF supports CDP and LLDP. Setting the MTU to its maximum value of 9000 is recommended for best performance.

The port group settings used for vDS-Mgmt are shown in Table 10.

Table 10 vDS-Mgmt port group settings

Port group	VLAN type	VLAN ID	Teaming and failover settings				
			Load balancing	Network failure detection	Notify switches	Failback	Active uplinks
vmotion-mgmt	VLAN	1612	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2
vsan-mgmt	VLAN	1613	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2

Attached hosts and physical adapters for vDS-Mgmt are listed in Table 11.

Table 11 vDS-Mgmt hosts and physical adapters

Host	Physical adapters
mgmt01.dell.local	vmnic0, vmnic1
mgmt02.dell.local	vmnic0, vmnic1
mgmt03.dell.local	vmnic0, vmnic1
mgmt04.dell.local	vmnic0, vmnic1

Note: Actual vmnic numbering may vary depending on network adapters installed in the host.

6.3.1.2 vDS-CompEdge configuration details

Configuration settings used for vDS-CompEdge in the Compute-Edge cluster are listed in Table 12.

Table 12 vDS-CompEdge settings

Distributed switch name	Version	Number of uplinks	Network I/O control	Discovery Protocol Type / Operation	MTU (Bytes)
vDS-CompEdge	6.5.0	2	Enabled	LLDP / Both	9000 Bytes

The port group settings used for vDS-CompEdge are shown in Table 13.

Table 13 vDS-CompEdge port group settings

Port group	VLAN type	VLAN ID	Teaming and failover settings				
			Load balancing	Network failure detection	Notify switches	Failback	Active uplinks
vmotion-compedge	VLAN	1622	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2
vsan-compedge	VLAN	1623	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2

Attached hosts and physical adapters for vDS-CompEdge are listed in Table 14.

Table 14 vDS-CompEdge hosts and physical adapters

Host	Physical adapters
comp101.dell.local	vmnic2, vmnic3
comp102.dell.local	vmnic2, vmnic3
comp103.dell.local	vmnic2, vmnic3
comp104.dell.local	vmnic2, vmnic3

Note: Actual vmnic numbering may vary depending on network adapters installed in the host.

6.3.1.3 vDS-Comp configuration details

Configuration settings used for vDS-Comp in the Compute cluster are listed in Table 15.

Table 15 vDS-Comp settings

Distributed switch name	Version	Number of uplinks	Network I/O control	Discovery Protocol Type / Operation	MTU (Bytes)
vDS-Comp	6.5.0	2	Enabled	LLDP / Both	9000 Bytes

The port group settings used for vDS-Comp are shown in Table 16.

Table 16 vDS-Comp port group settings

Port group	VLAN type	VLAN ID	Teaming and failover settings				
			Load balancing	Network failure detection	Notify switches	Failback	Active uplinks
vmotion-comp	VLAN	1632	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2
vsan-comp	VLAN	1633	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2

Attached hosts and physical adapters for vDS-Comp are listed in Table 17.

Table 17 vDS-Comp hosts and physical adapters

Host	Physical adapters
comp201.dell.local	vmnic0, vmnic1
comp202.dell.local	vmnic0, vmnic1
comp203.dell.local	vmnic0, vmnic1
comp204.dell.local	vmnic0, vmnic1

6.3.1.4 vDS summary

The Networking tab in the VMware Web Client Navigator pane is shown in Figure 59 after initial vDS configuration is complete. vDS-Mgmt, vDS-Comp, and vDS-CompEdge appear under their applicable data centers with port groups and uplinks configured.

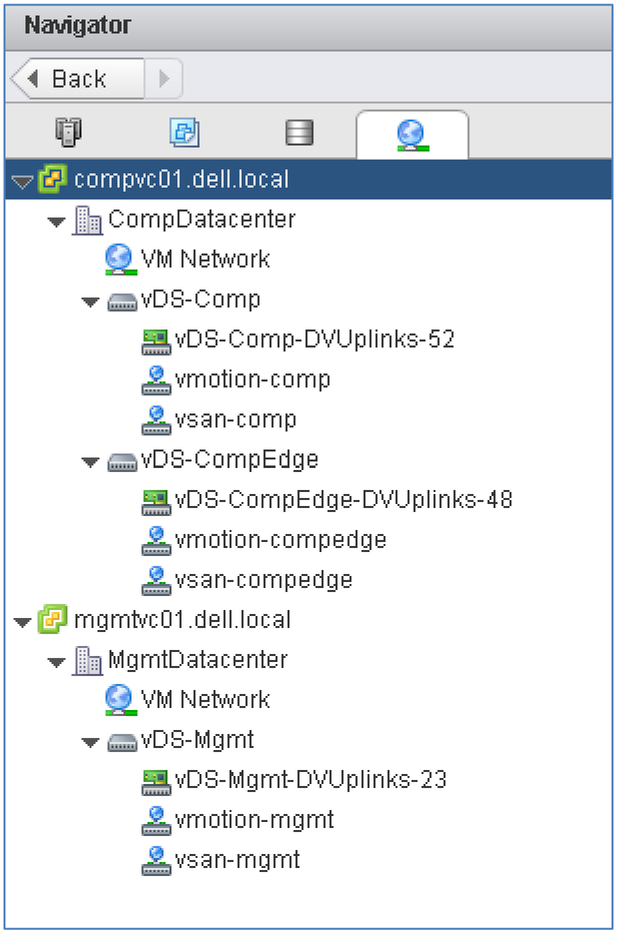


Figure 59 VMware Web Client - Networking tab

6.3.2 Network I/O Control

In VMware vSphere, Network I/O Control (NIOC) allows the assignment of share values for different traffic types. When bandwidth contention occurs, NIOC applies the share values to each traffic type. As a result, less important traffic is throttled allowing more bandwidth for critical traffic.

NIOC allows either shares or limits for bandwidth allocation restriction. It is a best practice to use shares instead of limits. Limits impose hard restrictions on the amount of bandwidth traffic flows utilize, even when network bandwidth is available.

To locate the configuration page, navigate to each vDS and select **Configure > Resource Allocation > System Traffic**.

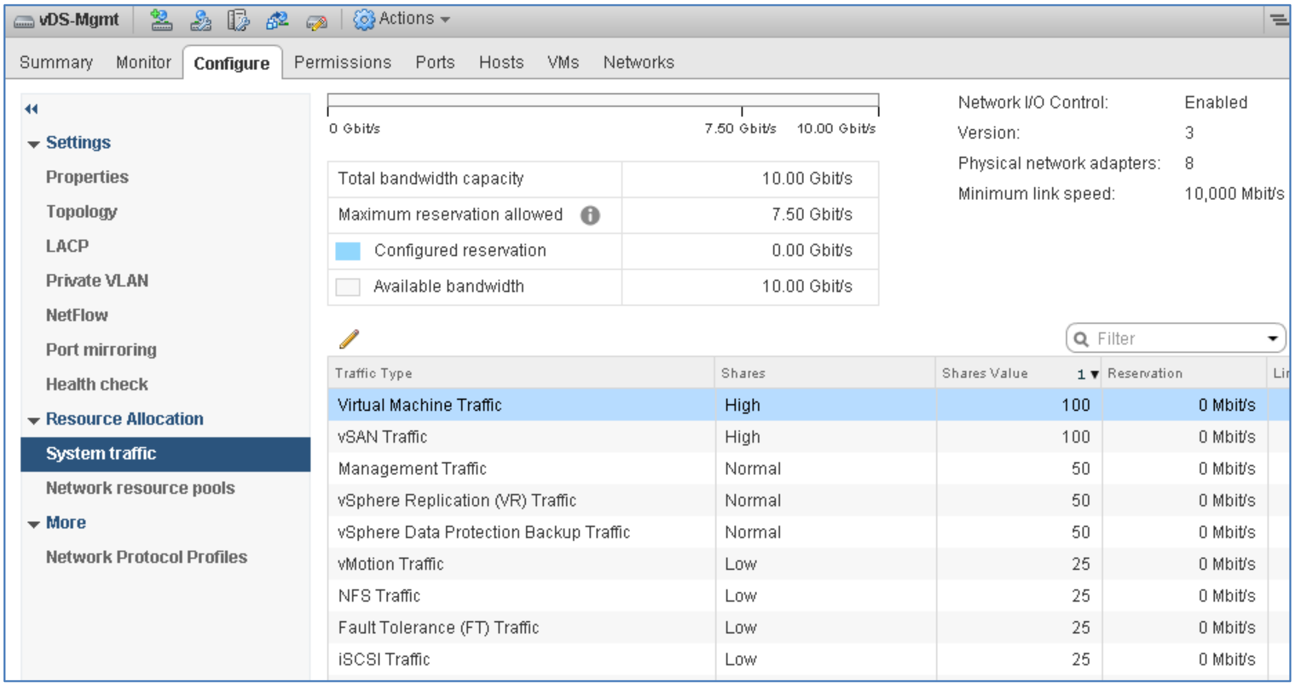


Figure 60 Resource allocation configured on vDS-Mgmt

The recommended share allocation by traffic type per VVD 4.1 is shown in Figure 60. Configure the settings shown on each vDS.

Note: VVD does not provide share setting recommendations for **Replication Traffic** or **Data Protection Backup Traffic**. The default share setting, **Normal**, is used for these two traffic types.

6.3.3 VMkernel adapter configuration

VMkernel adapters provide connectivity to ESXi hosts and handle management, vMotion, vSAN, and VXLAN traffic. In this section, VMkernel adapters are created and associated with vDS vMotion and vSAN port groups.

During ESXi installation, VMkernel adapter vmk0 is automatically created on each host for the OOB management network. It is on a VMware standard switch (VSS) named vSwitch0. No further configuration is needed for this connection. Two additional VMkernel adapters, vmk1 and vmk2, are manually added to each ESXi host for vMotion and vSAN traffic. These adapters are connected to the vDS for each cluster.

Note: VMkernel adapters for NSX VTEPs are configured in Section 10 of this guide.

The vSAN VMkernel adapter is used for vSAN traffic within the cluster and uses the default TCP/IP stack.

The vMotion VMkernel adapter is configured to allow VM mobility within and across clusters and it uses the vMotion TCP/IP stack. The vMotion TCP/IP stack allows a dedicated default gateway to be specified. This enables vMotion traffic to be routed between clusters and racks.

Note: The default gateway for the vMotion stack is configured by first selecting the host in the **Navigator** pane. In the center pane, select **Configure > Networking > TCP/IP Configuration**. Under TCP/IP stacks, select **vMotion**. Click the icon to edit. Select **Routing** and enter the default gateway.

6.3.3.1 Management cluster hosts

Table 18 shows the VMkernel configuration details for the four hosts in the Management cluster:

Table 18 vDS-Mgmt VMkernel adapters

vDS	Existing network	TCP/IP stack	Enabled services	Host VMkernel IP addresses	TCP/IP stack gateway address	MTU
vDS-Mgmt	vmotion-mgmt	vMotion	vMotion	172.16.12.1-4 /24	172.16.12.254	9000
vDS-Mgmt	vsan-mgmt	Default	vSAN	172.16.13.1-4 /24	Default (Not Used)	9000

When configuration is complete, the VMkernel adapters page for each host in the Management cluster appears similar to Figure 61. Adapters vmk1 and vmk2 are added with the appropriate service, vMotion or vSAN, enabled as shown in the two columns on the right.

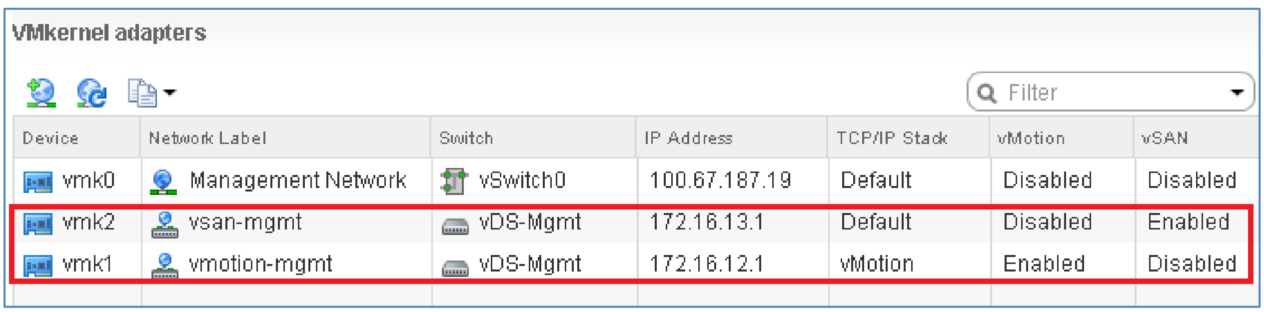


Figure 61 VMkernel adapters for host mgmt01 in the Management cluster

The **TCP/IP configuration** page for each host shows the default gateway for the vMotion stack is configured as shown:

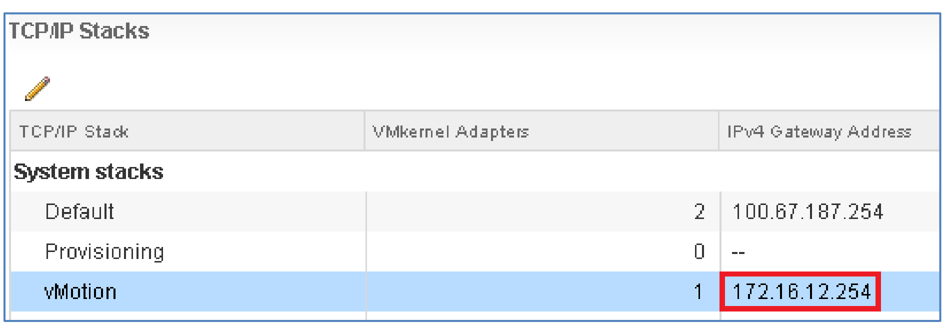


Figure 62 vMotion gateway configured for host mgmt01 in the Management cluster

Figure 63 shows the completed topology of vDS-Mgmt for the Management cluster. Port groups, VLAN assignments, VMkernels, IP addresses, and physical NIC uplinks are shown.

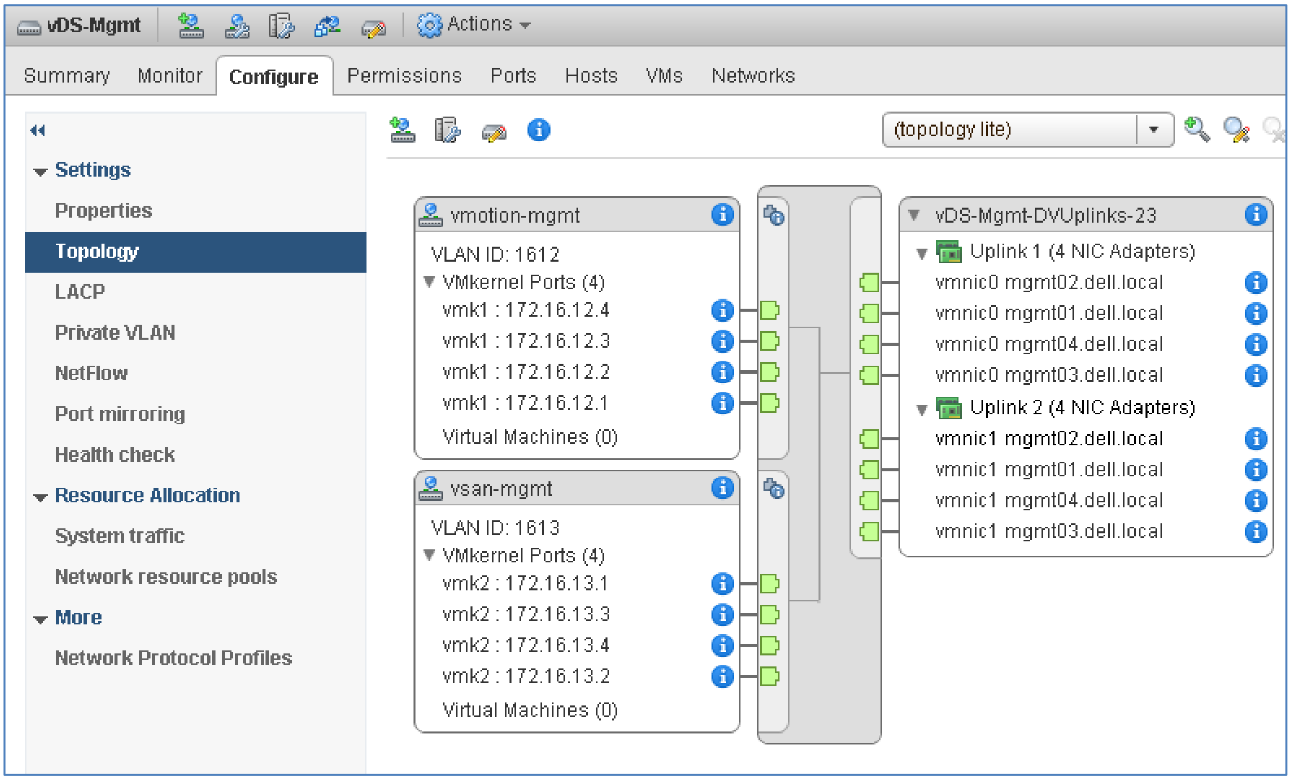


Figure 63 vDS-Mgmt topology after VMkernel adapter configuration

6.3.3.2 Compute-Edge cluster hosts

The VMkernel configuration details for the four hosts in the Compute-Edge cluster are listed in Table 19.

Table 19 vDS-Comp VMkernel adapters

vDS	Existing network	TCP/IP stack	Enabled services	Host VMkernel IP addresses	TCP/IP stack gateway address	MTU
vDS-CompEdge	vmotion-compedge	vMotion	vMotion	172.16.22.1-4 /24	172.16.22.254	9000
vDS-CompEdge	vsan-compedge	Default	vSAN	172.16.23.1-4 /24	Default (Not Used)	9000

When configuration is complete, the **VMkernel adapters** page for each host in the Compute-Edge cluster appears similar to Figure 64. Adapters vmk1 and vmk2 are added with the appropriate service, vMotion or vSAN, enabled as shown in the two columns on the right.

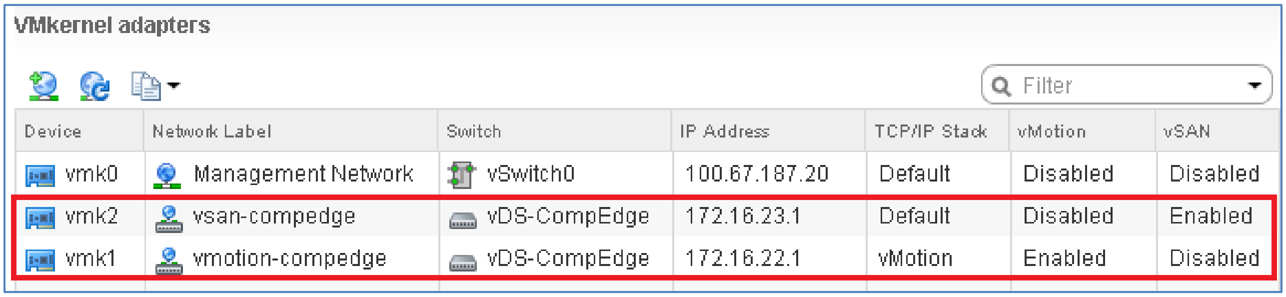


Figure 64 VMkernel adapters for host comp101 in the Compute-Edge cluster

The **TCP/IP configuration** page for each host shows the default gateway for the vMotion stack is configured as shown:

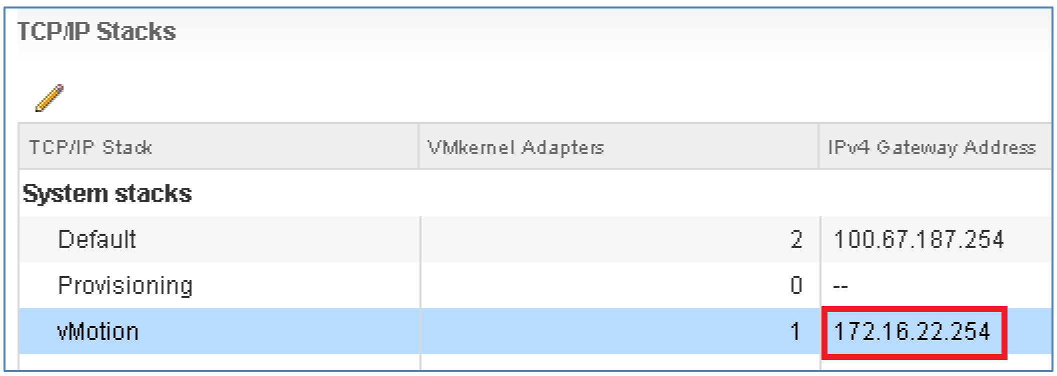


Figure 65 vMotion gateway configured on host comp101 in the Compute-Edge cluster

Figure 66 shows the completed topology of vDS-CompEdge for the Compute cluster. Port groups, VLAN assignments, VMkernels, IP addresses, and physical NIC uplinks are shown.

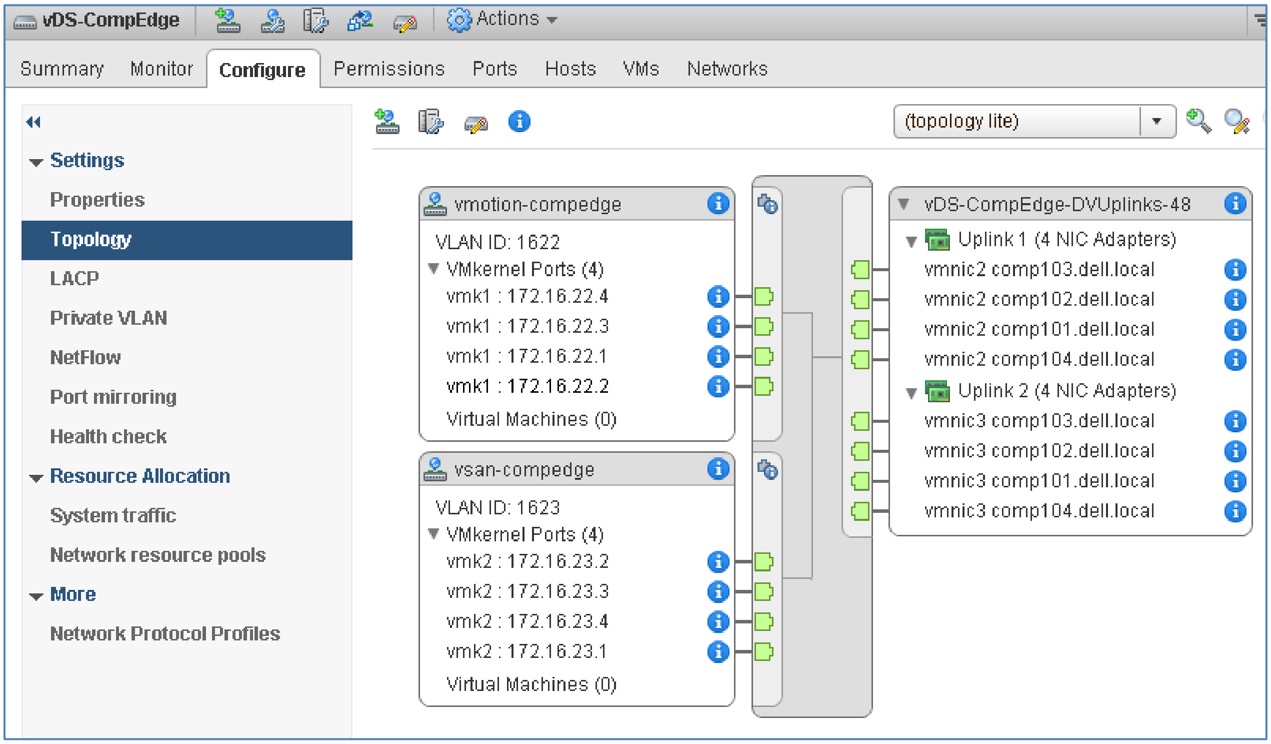


Figure 66 vDS-CompEdge topology after VMkernel adapter configuration

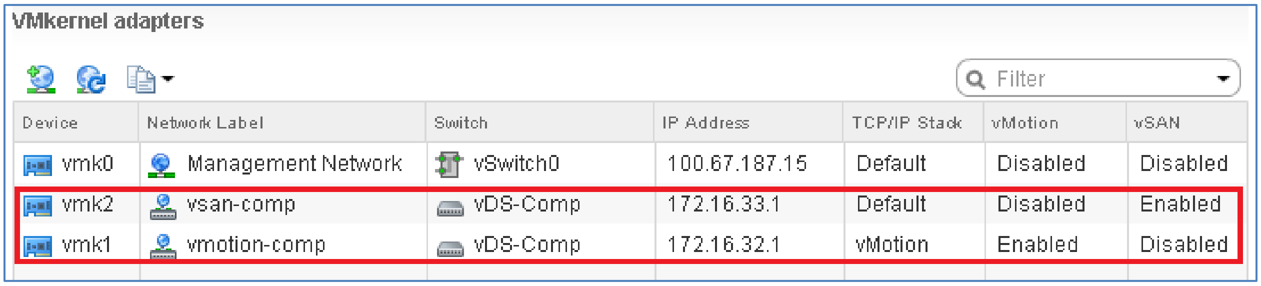
6.3.3.3 Compute cluster hosts

The VMkernel configuration details for the four hosts in the Compute cluster are listed in Table 20.

Table 20 vDS-Comp VMkernel adapters

vDS	Existing network	TCP/IP stack	Enabled services	Host VMkernel IP addresses	TCP/IP stack gateway address	MTU
vDS-Comp	vmotion-mgmt	vMotion	vMotion	172.16.32.1-4 /24	172.16.32.254	9000
vDS-Comp	vsan-mgmt	Default	vSAN	172.16.33.1-4 /24	Default (Not Used)	9000

When configuration is complete, the **VMkernel adapters** page for each host in the Compute cluster appears similar to Figure 67. Adapters vmk1 and vmk2 are added with the appropriate service, vMotion or vSAN, enabled as shown in the two columns on the right.

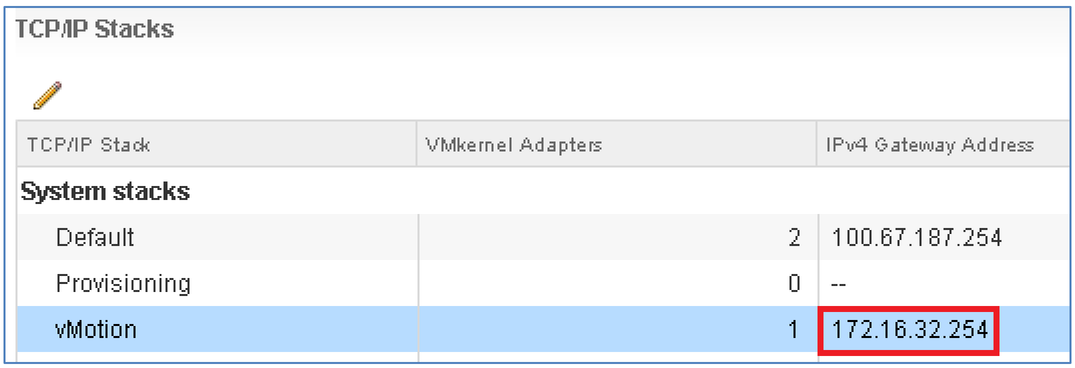


VMkernel adapters

Device	Network Label	Switch	IP Address	TCP/IP Stack	vMotion	vSAN
vmk0	Management Network	vSwitch0	100.67.187.15	Default	Disabled	Disabled
vmk2	vsan-comp	vDS-Comp	172.16.33.1	Default	Disabled	Enabled
vmk1	vmotion-comp	vDS-Comp	172.16.32.1	vMotion	Enabled	Disabled

Figure 67 VMkernel adapters for host comp201 in the Compute cluster

The **TCP/IP configuration** page for each host shows the default gateway for the vMotion stack is configured as shown:



TCP/IP Stacks

TCP/IP Stack	VMkernel Adapters	IPv4 Gateway Address
System stacks		
Default	2	100.67.187.254
Provisioning	0	--
vMotion	1	172.16.32.254

Figure 68 vMotion gateway configured on host comp201 in the Compute cluster

Figure 69 shows the completed topology of vDS-Comp for the Compute cluster. Port groups, VLAN assignments, VMkernels, IP addresses, and physical NIC uplinks are shown.

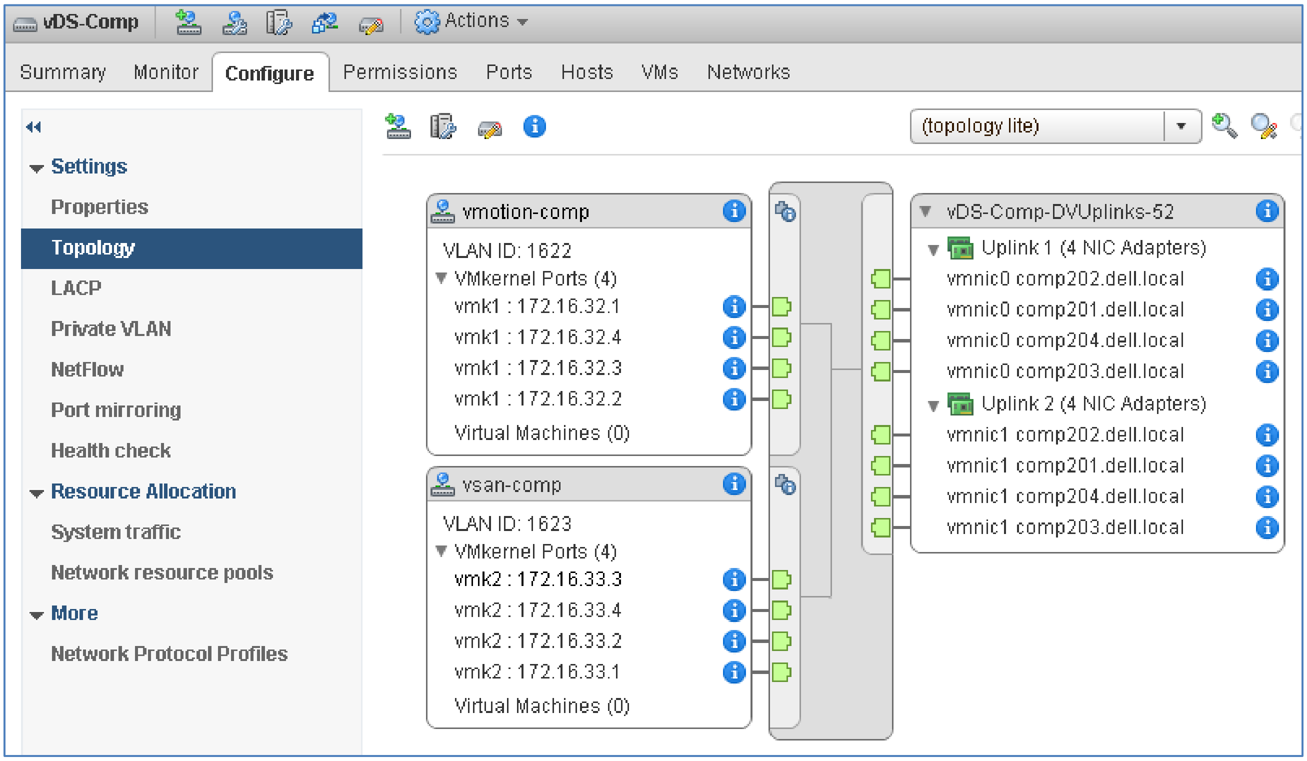


Figure 69 vDS-Comp topology after VMkernel adapter configuration

Note: Hosts will not be able to communicate with each other until the vCenters are integrated with BCF in the next section.

7 VMware integration with BCF

Integrating VMware vSphere with BCF provides an integrated solution that uses BCF as the underlying physical network. Integration benefits include:

- Automatic BCF ToR-to-host link detection and interface group formation
- Automatic BCF L2 network creation and VM learning
- Network policy migration for vMotion / DRS
- Improved VM network visibility and troubleshooting, especially in regard to mapping between virtual and physical network resources

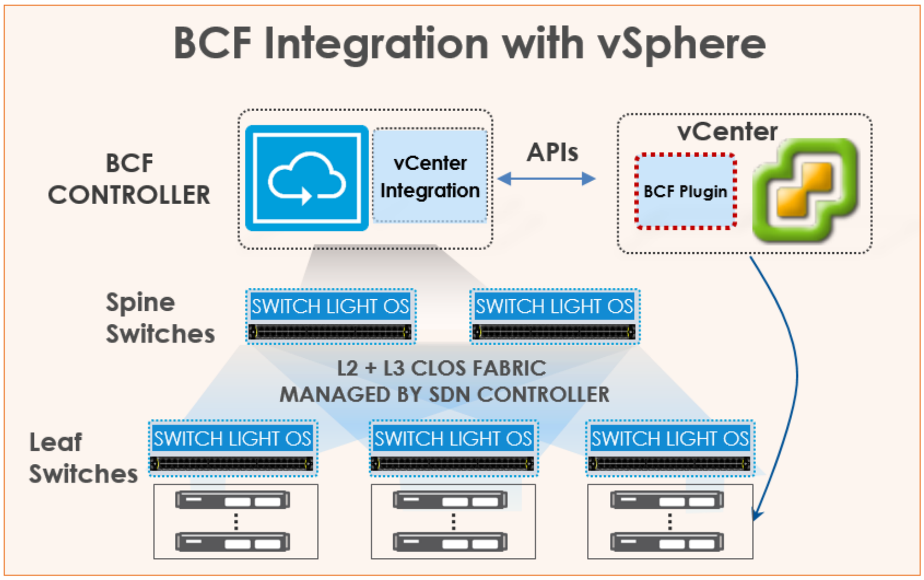


Figure 70 BCF integration with VMware vSphere

The information used to integrate both vCenter Server appliances with BCF is shown in Table 21. With BCF automation set to **Full**, the BCF configuration is automatically updated in response to changes on vCenter.

Table 21 VMware vCenter connection details

vCenter Name	Hostname	Tenant	BCF configuration automation level	vCenter plugin access right
mgmtvc01	mgmtvc01.dell.local	mgmtvc01	Full	Read-Write
compvc01	compvc01.dell.local	compvc01	Full	Read-Write

Note: The BCF automation level options are **Full** and **None**. Refer to the *BCF User Guide* for more information.

The vCenter plugin access right sets the permission level for the vCenter BCF plugin. The **Read-Write** option allows the plugin to be used similarly to the BCF GUI from within vCenter. It may be set to **Read-Only** to prevent changes to BCF from the vCenter plugin.

7.1 Add vCenter Servers to BCF

In this section, both vCenter Servers are added to BCF and the automatic configuration level is set to full.

1. In the BCF GUI, navigate to **Integration > Orchestration > VMware vCenters**.
2. Select the + icon to open the **Create vCenter** dialog box.
3. In the **Create vCenter** dialog box, complete the following:
    a. **Name:** Provide the name of the first vCenter, **mgmtvc01.**
    b. **Hostname**: Enter the FQDN of the first vCenter, **mgmtvc01.dell.local**
    c. **Username/Password:** Provide the vCenter login credentials.
    d. Leave **Operational Mode** set to **Normal** and **Preserve Auto-Generated BCF Configuration** set to **No.**
    e. In this example, the **vCenter Plugin Access right** is set to **Read-Write**.
    f. Set the **BCF Config Automation Level** to **Full**. This enables changes in vCenter to automatically update the BCF Controller without manual intervention.

The **Create vCenter** dialog box appears as shown:

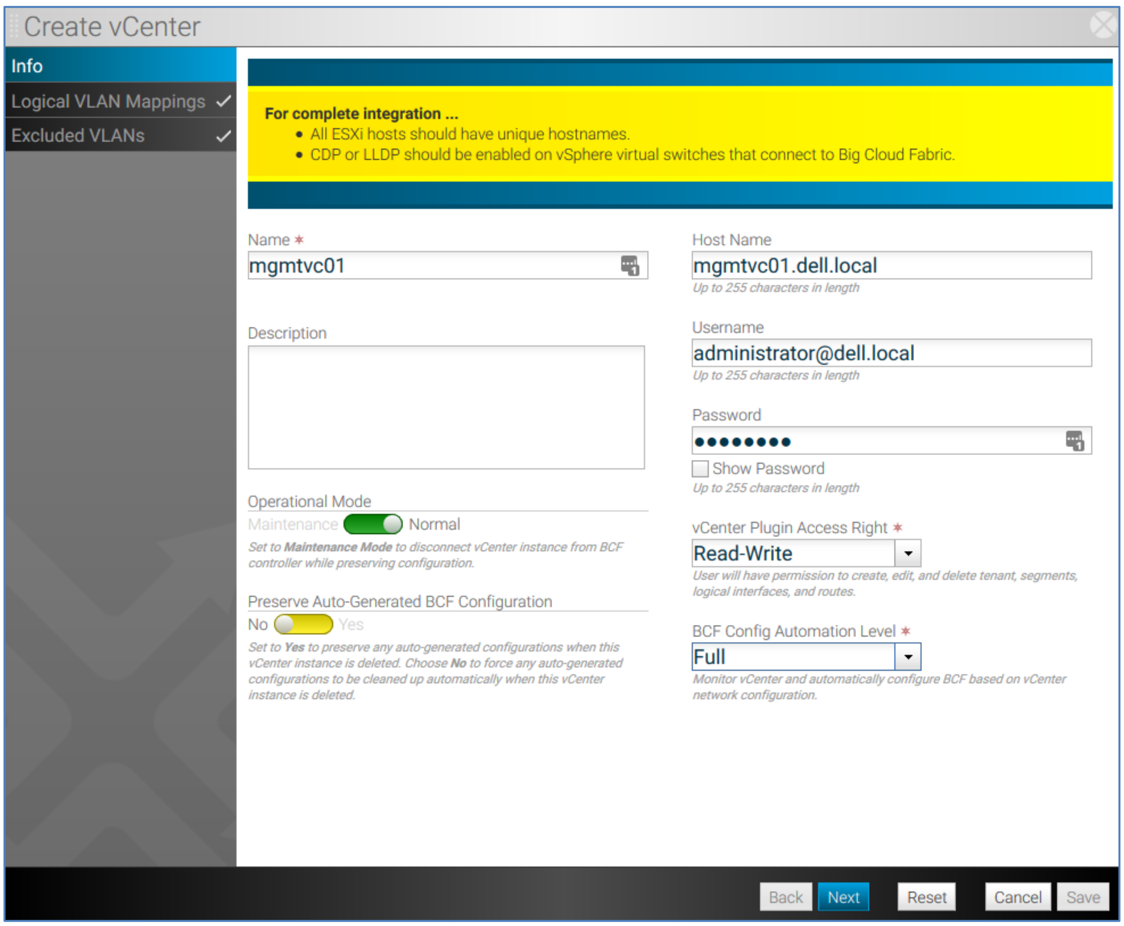


Figure 71 Create vCenter dialog box

4. Click **Next** to open the **Logical VLAN Mappings** page.
    a. Ensure the slider is set to **Single Tenant.**
    b. Next to the **Tenant** box, **c**lick the + icon and create a new tenant named **mgmtvc01**. Move the **Multicast** slider to **Enabled** and click **Submit.**

Note: As of vSAN 6.6, vSAN communication is done via unicast. However, multicast is required for NSX with hybrid mode replication, so multicast is enabled here. This is covered in Section 10 of this document.

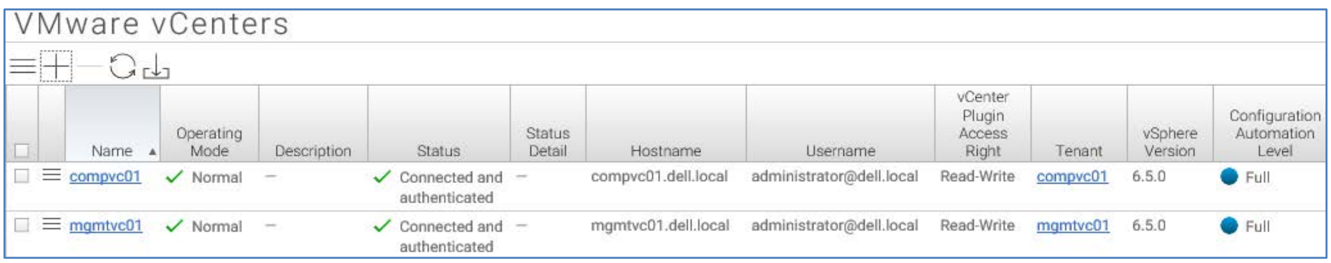
5. Click **Next** to open the **Excluded VLANs** page**.**
    a. Next to **VLAN Ranges to Exclude**, click the + icon.
    b. Set the range from **0** to **0** to exclude the untagged OOB management VLAN. (All other VLANs in this deployment are included).

Note: Specific VLANs defined in VMware virtual switch port groups may be excluded to prevent corresponding BCF segments from being automatically created. In this deployment, the OOB management network, VM Network, is not managed by BCF. Its untagged VLAN is excluded to prevent BCF from creating a segment for it. See the Big Cloud Fabric User Guide for more information.

6. Click **Save.**

Repeat the steps above for the second vCenter, **compvc01,** and create a new tenant named **compvc01.**

When complete, the two vCenters appear on the **Integration > Orchestration > VMware vCenters** page as shown in Figure 72:


VMware vCenters

Name	Operating Mode	Description	Status	Status Detail	Hostname	Username	vCenter Plugin Access Right	Tenant	vSphere Version	Configuration Automation Level
compvc01	Normal	–	Connected and authenticated	–	compvc01.dell.local	administrator@dell.local	Read-Write	compvc01	6.5.0	Full
mgmtvc01	Normal	–	Connected and authenticated	–	mgmtvc01.dell.local	administrator@dell.local	Read-Write	mgmtvc01	6.5.0	Full

Figure 72 VMware vCenters integrated with BCF

BCF imports the vCenter configuration and automatically configures the switches. Clicking the vCenter name displays imported configuration information for the vCenter.

For example, clicking vCenter **compvc01** displays the page shown in Figure 73.

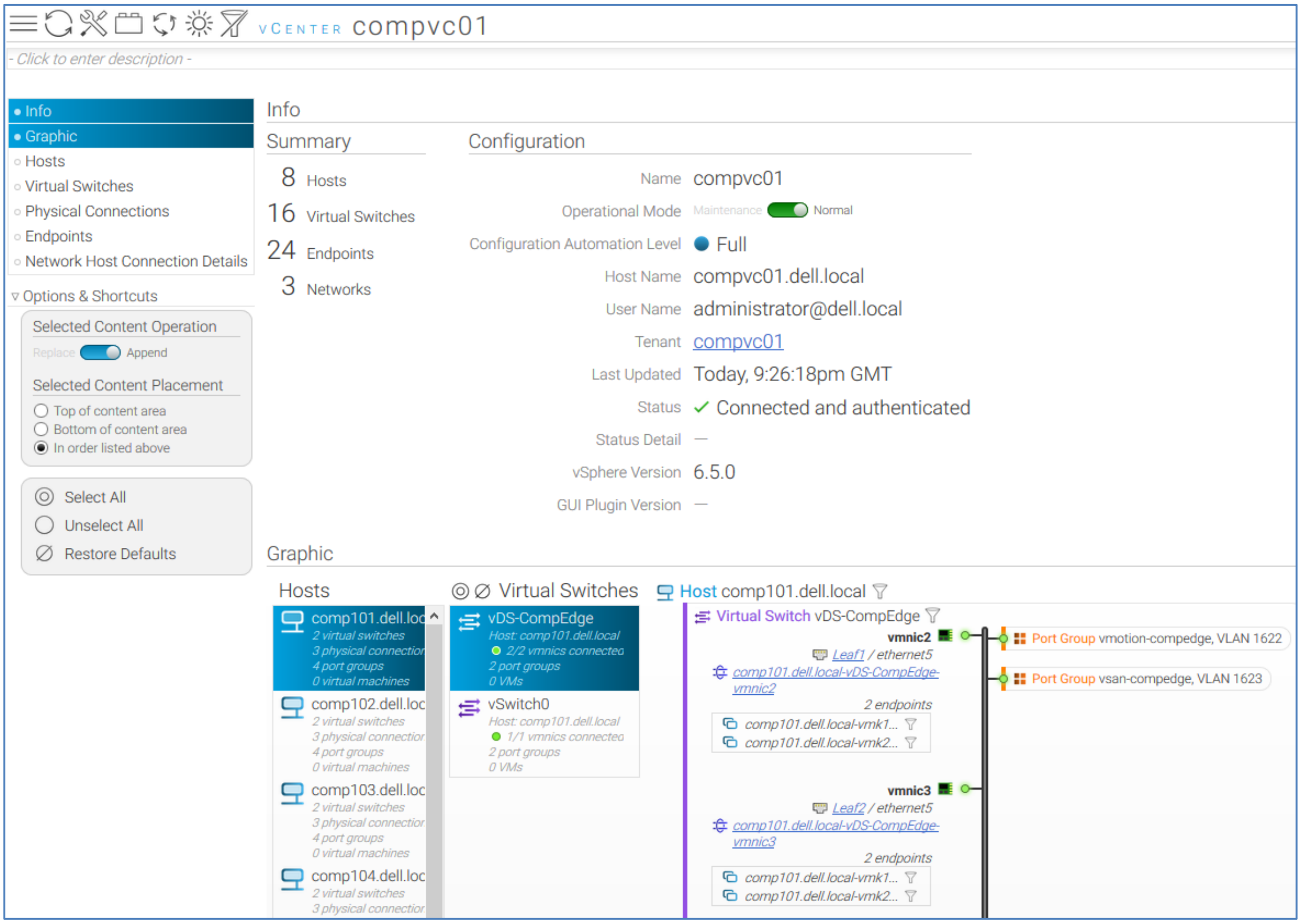


Figure 73 vCenter details for compvc01

The **Info** and **Graphic** items are selected in the left pane. In the right pane, the **Info** section provides an overview of the configuration.

In the **Graphic** section of the right pane, host **comp101** and **vDS-CompEdge** are selected. This displays information such as the mapping of vmnics to physical leaf switch ports and VMware port groups.

Automatically configured interface groups are viewed in BCF by going to **Fabric > Interface Groups** as shown in Figure 74.

Interface Groups

Name	Description	Leaf Group	State	Mode	Auto-Discovered	Backup Mode	Preempt Backups when Primaries Available	Total Member Interfaces	Member Interface Status
mgmt04.dell.local-vDS-Mgmt-vmnic1	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt04.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
mgmt04.dell.local-vDS-Mgmt-vmnic0	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt04.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
mgmt03.dell.local-vDS-Mgmt-vmnic1	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt03.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
mgmt03.dell.local-vDS-Mgmt-vmnic0	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt03.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
mgmt02.dell.local-vDS-Mgmt-vmnic1	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt02.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
mgmt02.dell.local-vDS-Mgmt-vmnic0	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt02.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
mgmt01.dell.local-vDS-Mgmt-vmnic1	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt01.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
mgmt01.dell.local-vDS-Mgmt-vmnic0	Interface group for virtual switch vDS-Mgmt in ESXi host mgmt01.dell.local	Rack1	✓ Up	LLDP	–	Static	–	1	✓ All Up
controller-a0369faccdd8	–	Rack1	✓ Up	Static Auto Controller Inband	✓	NA	–	2	✓ All Up
controller-a0369facc9c4	–	Rack1	✓ Up	Static Auto Controller Inband	✓	NA	–	2	✓ All Up
comp204.dell.local-vDS-Comp-vmnic1	Interface group for virtual switch vDS-Comp in ESXi host comp204.dell.local	Rack2	✓ Up	LLDP	–	Static	–	1	✓ All Up
comp204.dell.local-vDS-Comp-vmnic0	Interface group for virtual switch vDS-Comp in ESXi host comp204.dell.local	Rack2	✓ Up	LLDP	–	Static	–	1	✓ All Up

Figure 74 Interface Groups page

All configured host-to-leaf switch connections are listed and their status is **Up**. In this deployment, each vmnic is in a separate interface group since the teaming method is **Route based on physical NIC load**. BCF Controller in-band connections are also shown.

> **Note:** BCF also reads OOB management network (vSwitch0) information from vCenter. These connections are shown as **Down** on the **Interface Groups** page as they are connected to the S3048-ON switches and are not managed by BCF. They also appear under warnings on the **Visibility > Fabric Summary** page as **Interface Group With Members Disabled in Forwarding State**. This information may be disregarded for these ports.

Interface group details are viewed by clicking the ▷ next to the hostname as shown in Figure 75:

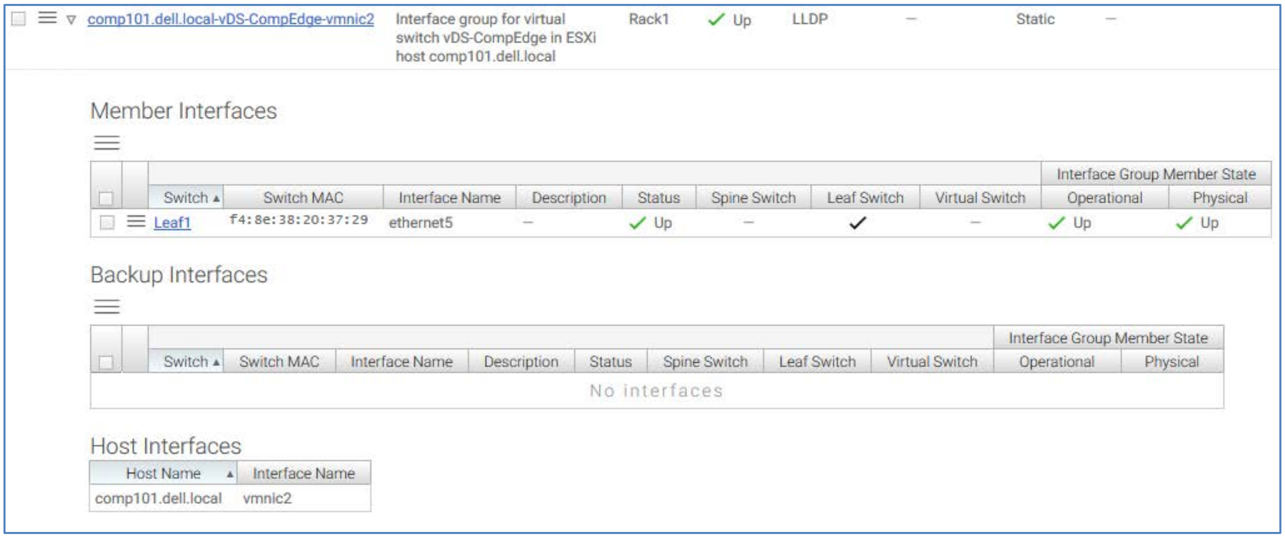


Figure 75 Interface group details for comp01 in BCF

Information includes leaf port and host vmnic connection information.

7.2 Add BCF Plugin to vCenter

Adding the BCF plugin to vCenter is optional. The plugin enables monitoring and configuration of certain BCF components from the vSphere Web Client as an alternative to using the BCF GUI.

Note: For more information about the BCF plugin, see the *Big Cloud Fabric User Guide*. The plugin is not used for configuration in this deployment guide.

The installation wizard is accessed in the BCF GUI by going to **Integration > Orchestration > VMware vCenters.** Select the ☰ icon next to the first vCenter name and click **Deploy vCenter GUI Plugin**.

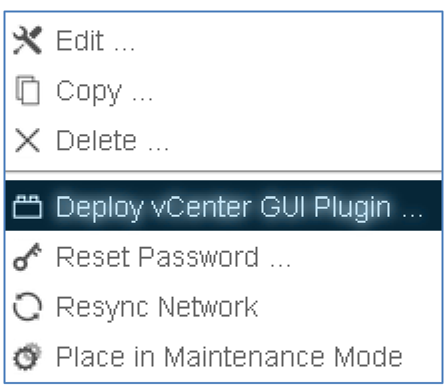


Figure 76 Deploy BCF plugin to vCenter

Enter the vCenter **Username** and **Password** in the dialog box and click **Submit.** Repeat for the second vCenter.

After installation is complete, log out and log back in to the vSphere Web Client for the vCenter. The vCenter **Home** page displays the **Big Cloud Fabric** icon.

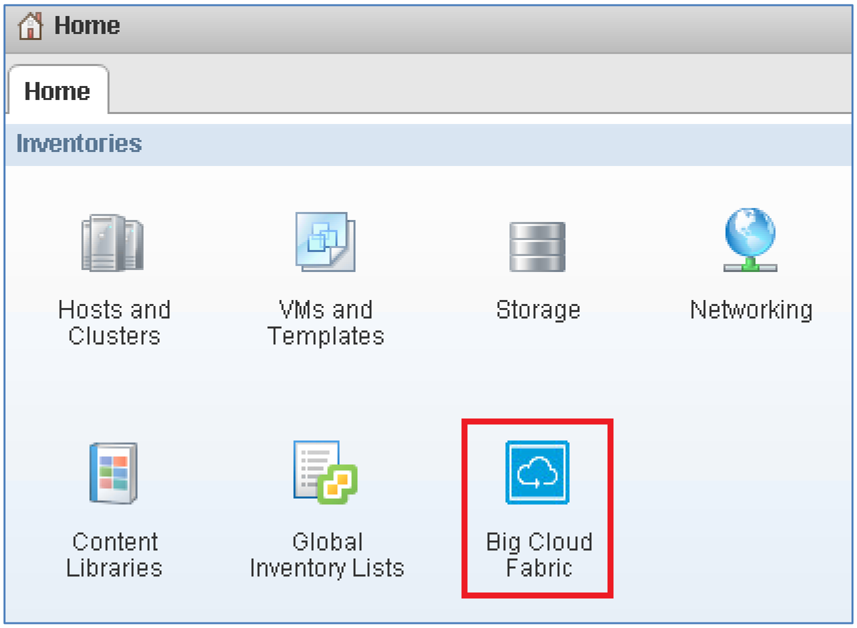


Figure 77 Big Cloud Fabric icon in vSphere Web Client

In the vSphere Web Client, double-click on the **Big Cloud Fabric** icon to open the page. In the left pane of the page, click on the **BCF Pod** address, **100.67.187.200**. The **Overview** page displays.

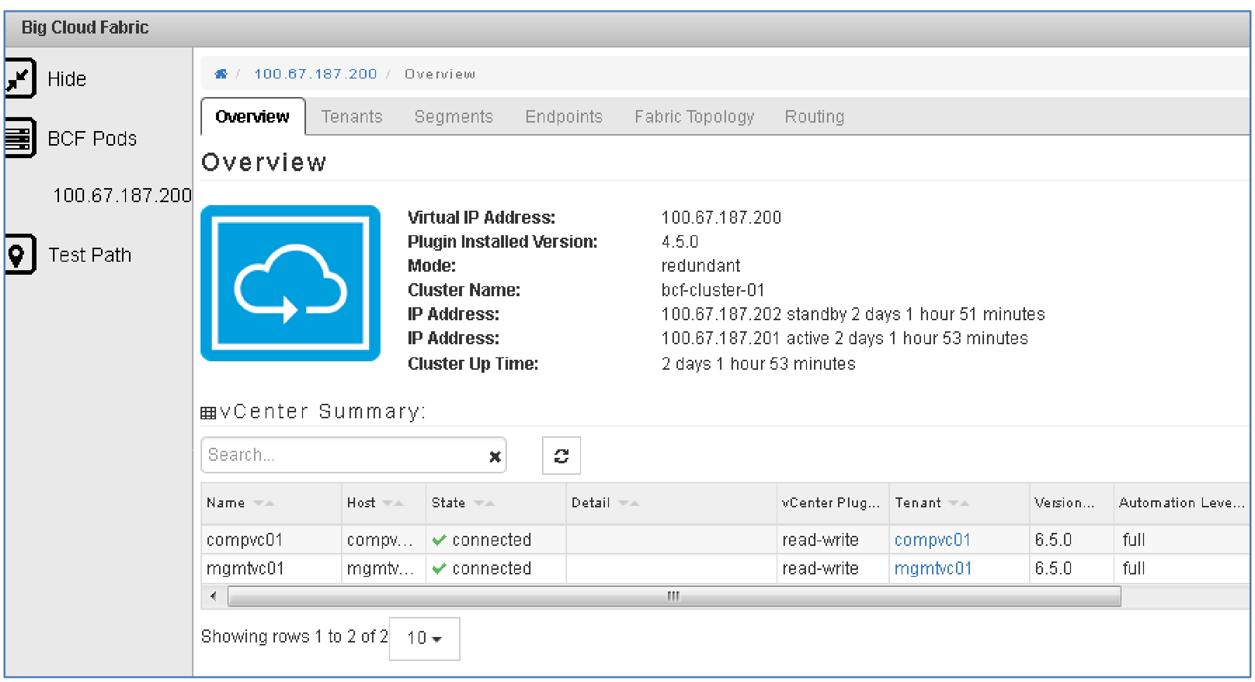


Figure 78 BCF plugin for vSphere web client

8 BCF tenant and segment configuration

8.1 Overview

A Big Cloud Fabric is organized into logical tenants and segments.

Tenants provide a logical grouping of layer 2 and layer 3 networks and are similar in function to a Virtual Routing and Forwarding (VRF) entity. Each tenant establishes a layer 3 boundary that separates traffic from other tenants through a logical router.

Segments are similar to VLANs. They are layer 2 networks consisting of logical ports and endpoints. Within each tenant, separate segments establish layer 2 boundaries for each tier.

In BCF, an endpoint is any host or virtual machine that terminates traffic. VMkernel adapters and VM vNICs are endpoints.

8.2 View tenants and segments

One tenant is manually created for each vCenter during the VMware integration process (completed in section 7.1). The two vCenter tenants are visible in the BCF GUI by selecting **Logical > Tenants.**

Tenants

Filter table rows

Name	Description	Multicast Enabled	Router MAC Address	Applied Policy List	Applied QoS List	System Tenant Interface
compvc01	vCenter default tenant	✓	NA	–	–	Not Configured
mgmtvc01	vCenter default tenant	✓	NA	–	–	Not Configured

Jan 18, 2018, 5:22:17pm GMT

Figure 79 Tenants screen with vCenter tenants compvc01 and mgmtvc01

One segment is automatically created in BCF for each vCenter VLAN. To view the list of segments, select **Logical > Segments**.

Segments

Filter table rows

Name	Tenant	Description	Member VNI	Member VLAN	Interface Group Membership Rules
compvc01-1622	compvc01	1 vSphere portgroups: vmotion-compedge	–	–	8
compvc01-1623	compvc01	1 vSphere portgroups: vsan-compedge	–	–	8
compvc01-1632	compvc01	1 vSphere portgroups: vmotion-comp	–	–	8
compvc01-1633	compvc01	1 vSphere portgroups: vsan-comp	–	–	8
mgmtvc01-1612	mgmtvc01	1 vSphere portgroups: vmotion-mgmt	–	–	8
mgmtvc01-1613	mgmtvc01	1 vSphere portgroups: vsan-mgmt	–	–	8

Figure 80 Segments automatically created through vCenter integration

At this point, hosts on the same segments can communicate with each other. For communication between segments, logical router interfaces are configured.

8.3 Configure logical router interfaces

In this section, logical router interfaces are configured to enable communication between segments and tenants.

Figure 81 shows the BCF logical view with the two vCenter tenants in this deployment, mgmtvc01 and compvc01, and their respective segments and hosts.

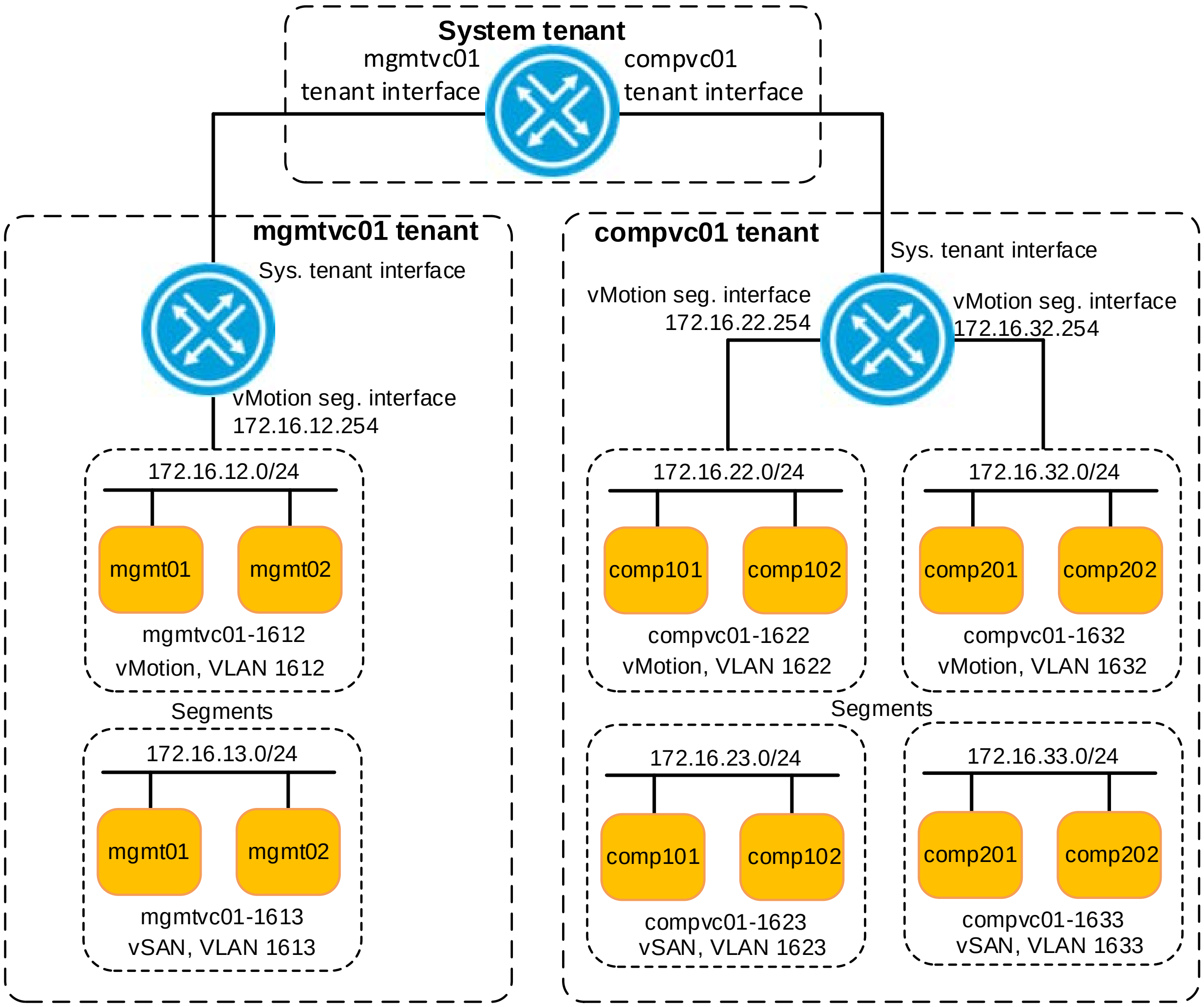


Figure 81 BCF tenants with vMotion and vSAN segments

Note: Only two of the four hosts from each cluster are shown in Figure 81 for clarity.

A logical router is automatically created on each tenant when it is defined. The logical router has two types of interfaces: tenant interfaces, and segment interfaces. Tenant interfaces are used to connect tenants together via the built in System tenant. Segment interfaces act as gateways for forwarding between segments within a tenant and routing traffic to other tenants through the System tenant.

Tenants, segments, and segment interface addresses used in this deployment are shown in Table 22:

Table 22 BCF tenant and segment configuration

Tenant	Logical segment name	Function	VLAN ID	Subnet	Segment interface address
mgmtvc01	mgmtvc01-1612	vMotion	1612	172.16.12.0/24	172.16.12.254
	mgmtvc01-1613	vSAN	1613	172.16.13.0/24	Not Used
compvc01	compvc01-1622	vMotion	1622	172.16.22.0/24	172.16.22.254
	compvc01-1623	vSAN	1623	172.16.23.0/24	Not Used
	compvc01-1632	vMotion	1632	172.16.32.0/24	172.16.32.254
	compvc01-1633	vSAN	1633	172.16.33.0/24	Not Used

Note: Only segment interfaces for vMotion are configured at this point. Since vSAN traffic is limited to a single segment, segment interfaces are not used for the vSAN networks. VXLAN segment interfaces are configured in Section 10.

To configure segment interfaces, do the following:

1. From the BCF GUI, go to **Logical > Tenants.**
2. Select a tenant, **mgmtvc01** in this example, to open the tenant configuration page.
3. In the left pane, scroll down and select **Segment Interfaces**. This adds **Segment Interfaces** to the right pane as shown in Figure 82.

Note: If additional items are selected in the left pane, you may need to scroll down in the right pane to view the **Segment Interfaces** section.

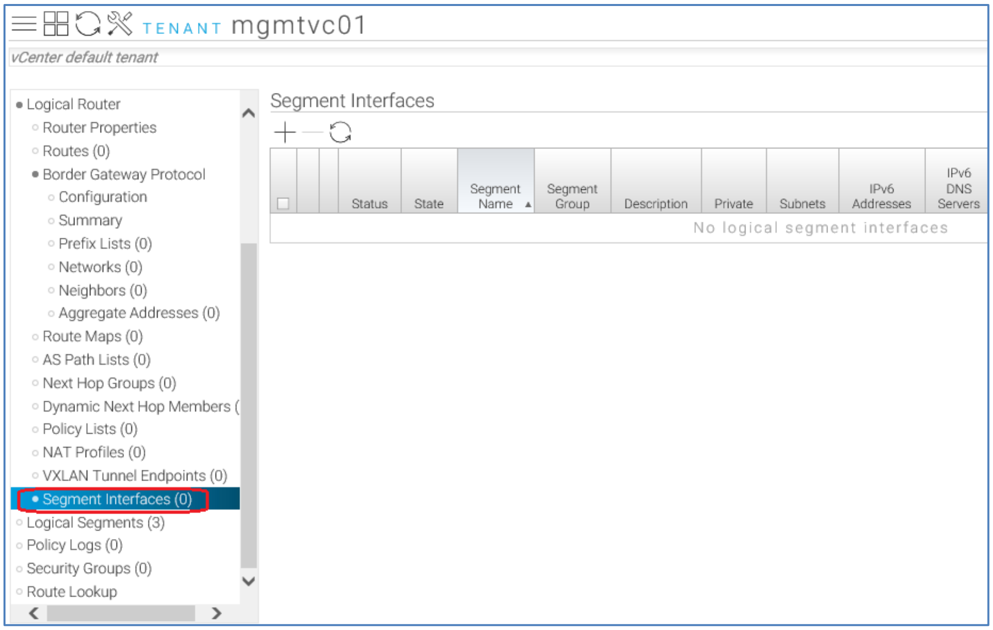


Figure 82 Segment Interfaces selected

4. In the right pane under **Segment Interfaces**, click the + icon. The **Create Logical Segment Interface** dialog box displays:

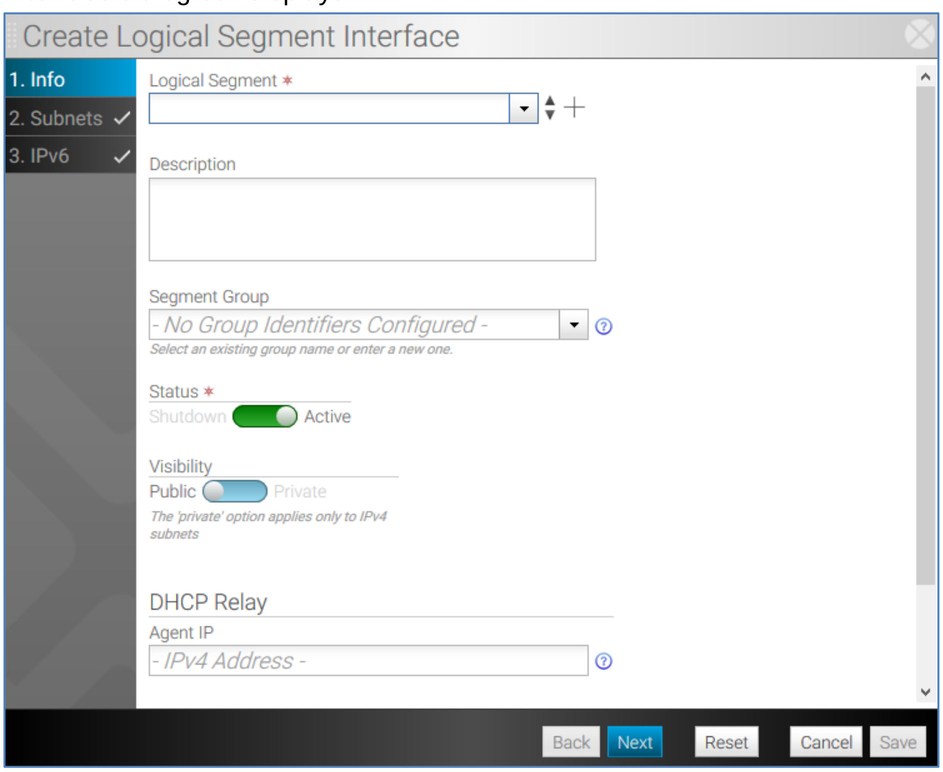


Figure 83 Create segment interface dialog box

5. Under **Logical Segment**, select the name of the first logical segment from the drop-down menu, **mgmtvc01-1612** in this example. This is for vMotion traffic in the management cluster. Leave other settings at their defaults and click **Next**.
6. Click the + icon to open the **Create Subnet** dialog box.
7. Provide the segment interface IP address and prefix per Table 22, **172.16.12.254 /24**. The subnet mask in dotted decimal form is automatically completed.

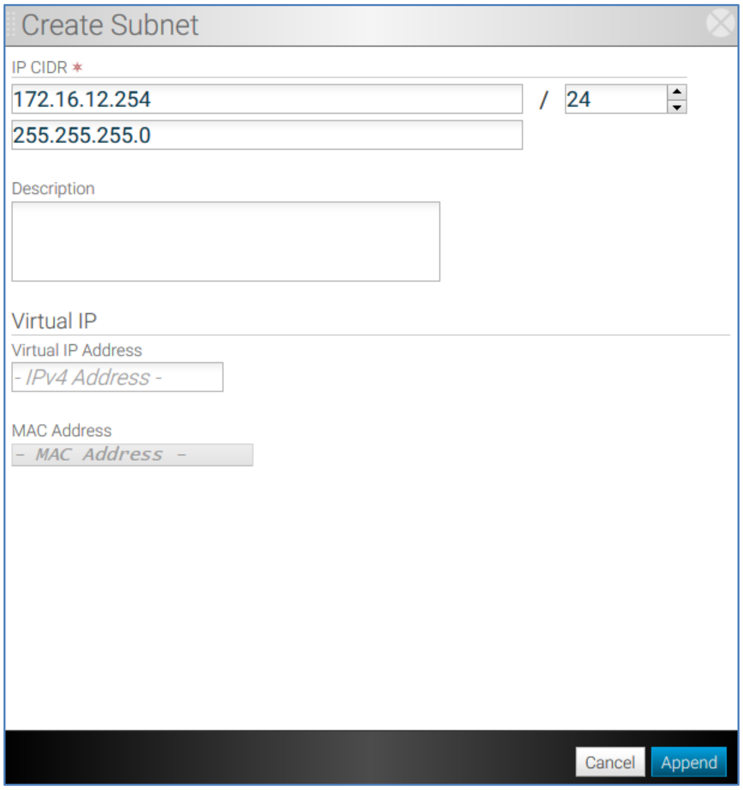
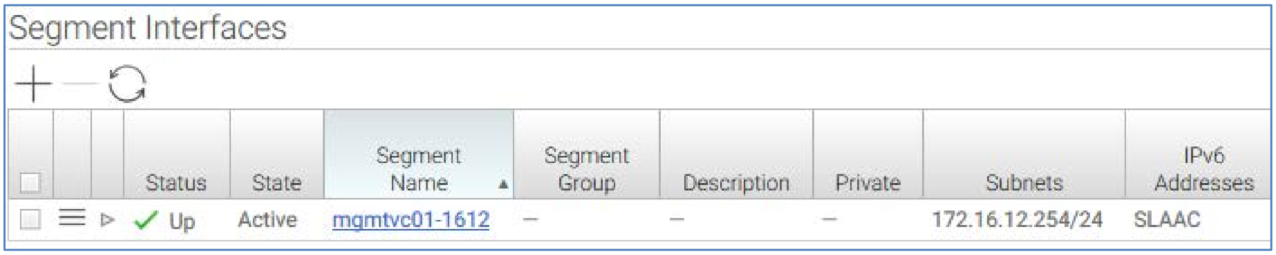


Figure 84 Create subnet dialog box

8. Click **Append > Save.**

The first segment interface is created. Repeat steps 1-8 above to create segment interfaces for the two vMotion segments in the compvc01 tenant per Table 22.

When complete, **Segment Interfaces** for mgmtvc01 and compvc01 appear as shown in Figure 85 and Figure 86.



Segment Interfaces

			Status	State	Segment Name	Segment Group	Description	Private	Subnets	IPv6 Addresses
			Up	Active	mgmtvc01-1612	–	–	–	172.16.12.254/24	SLAAC

Figure 85 Mgmtvc01 segment interfaces configured

Segment Interfaces

			Status	State	Segment Name	Segment Group	Description	Private	Subnets	IPv6 Addresses
			Up	Active	compvc01-1622	–	–	–	172.16.22.254/24	SLAAC
			Up	Active	compvc01-1632	–	–	–	172.16.32.254/24	SLAAC

Figure 86 Compvc01 segment interfaces configured

8.4 Configure System tenant interfaces and logical routers

At this point, hosts on different segments within the same tenant can communicate with each other, but not in different tenants. To enable communication between tenants, System tenant interfaces and logical routers are configured.

To configure the System tenant interfaces:

1. In the BCF GUI, go to **Logical > Tenants.**
2. Select a tenant, **mgmtvc01** in this example, to open its **Tenant** page.
3. In the left pane under **Logical Router**, select **Router Properties** and **Routes** to display them in the right pane as shown:

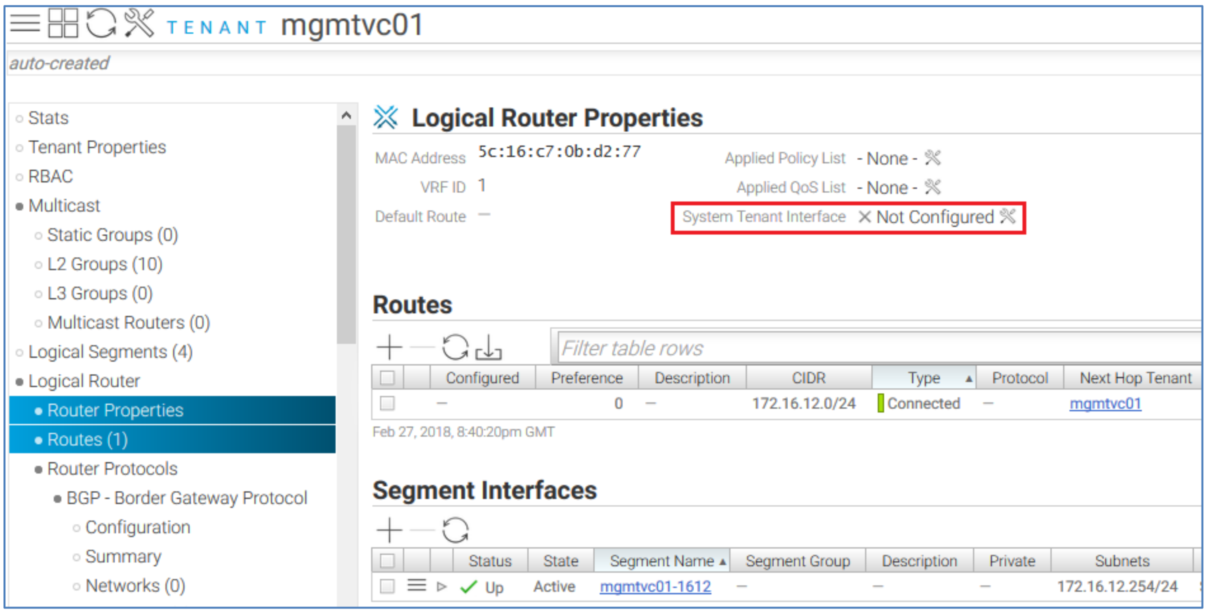


Figure 87 Tenant mgmtvc01 logical router properties and routes

4. In the right pane under **Logical Router Properties**, click the icon next to **System Tenant Interface** (outlined in red in Figure 87). The **Manage Tenant Interfaces** dialog box displays:

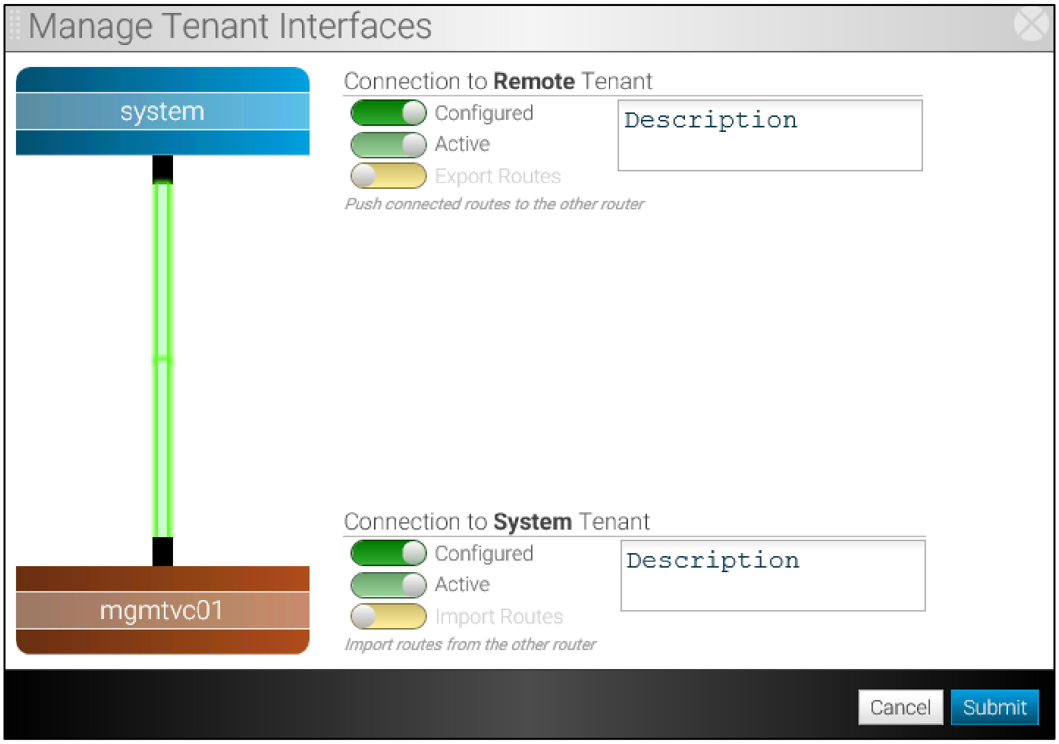


Figure 88 Enabling System tenant interfaces

5. Move both sets of **Configured** and **Active** sliders to the right to enable the interfaces.
6. Click **Submit** to apply the changes.
7. On the **Tenant** page, the **System Tenant Interface** is now **Up** as shown:

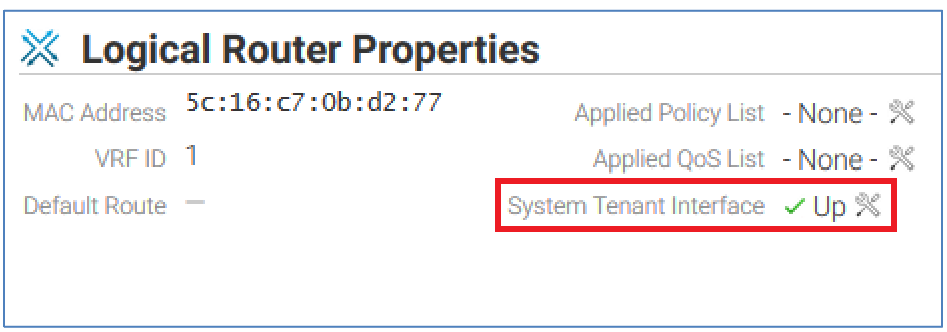


Figure 89 Mgmtvc01 default route set to System tenant

8. Configure the default route by scrolling down to the **Routes** section. Click the + icon to open the **Create Route** dialog box.

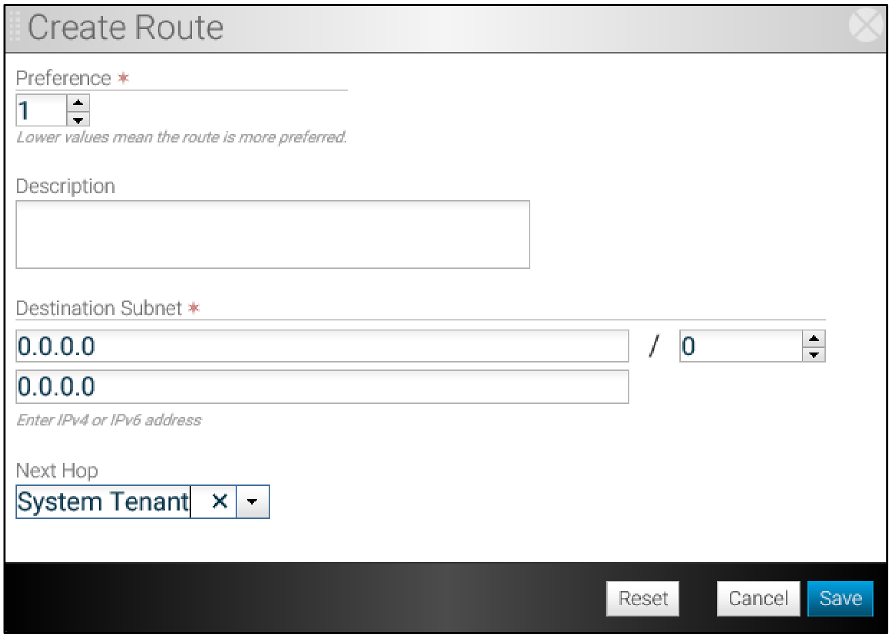


Figure 90 Create route dialog box

9. In this example, the destination subnet is set to **0.0.0.0/0** (any). Set **Next Hop** to **System Tenant** and click **Save.**

For the **mgmtvc01** tenant**, Logical Router Properties** and **Routes** appears as shown in Figure 91:

Routes

Configured	Preference ▴	Description	CIDR	Type	Protocol	Next Hop Tenant	Next Hop Group	Next Hop IP Address	Status
–	0	–	172.16.12.2/32	Host	–	mgmtvc01	Segment Iface mgmtvc01-1612	–	Active
–	0	–	172.16.12.1/32	Host	–	mgmtvc01	Segment Iface mgmtvc01-1612	–	Active
–	0	–	172.16.12.0/24	Connected	–	mgmtvc01	Segment Iface mgmtvc01-1612	–	Active
✓	1	–	0.0.0.0/0	Static	–	system	Tenant iface system	–	Active

Figure 91 mgmtvc01 tenant logical router properties and routes

The **Routes** table should include:

- The Management vMotion segment interface, **172.16.12.0/24, Type: Connected, Status: Active**
- The default route, **0.0.0.0/0, Type: Static, Status: Active**

Note: Host routes shown in Figure 91, such as 172.16.12.2/32, are discovered automatically and may or may not appear in the list. This is dependent on host network activity.

Repeat the steps above for the **compvc01** tenant. When complete, **Logical Router Properties** and **Routes** sections of the compvc01 Tenant page appear as shown in Figure 92.

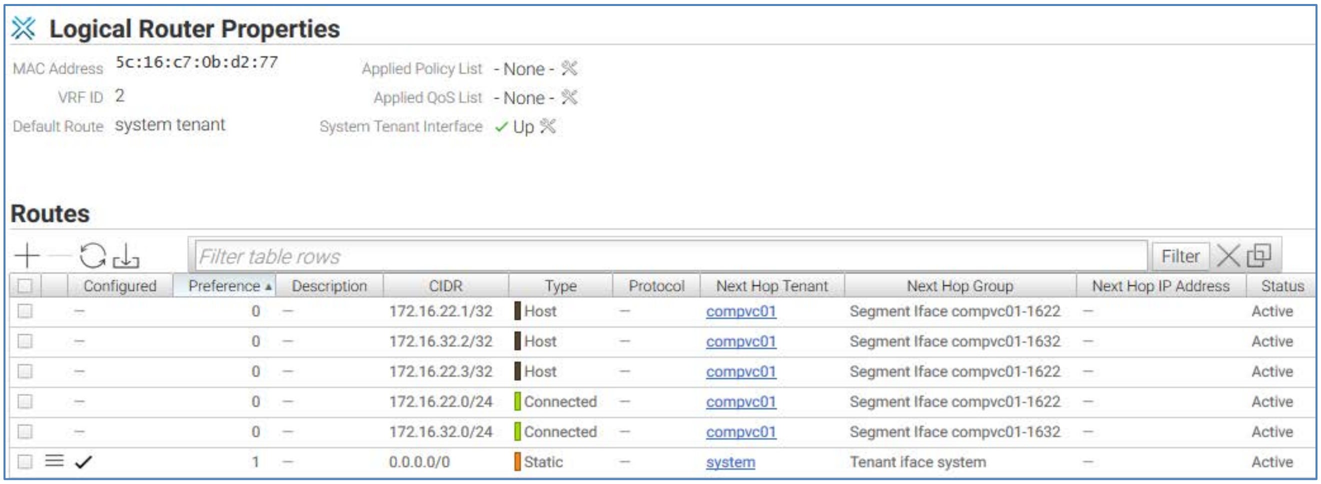


Figure 92 compvc01 tenant logical router properties and routes

The **Routes** table on the compvc01 tenant should include:

- The Compute-Edge vMotion segment interface, **172.16.22.0/24, Type: Connected, Status: Active**
- The Compute vMotion segment interface, **172.16.32.0/24, Type: Connected, Status: Active**
- The default route, **0.0.0.0/0, Type: Static, Status: Active**

Note: Host routes shown in Figure 92, such as 172.16.32.1/32, are discovered automatically and may or may not appear in the list. This is dependent on host network activity.

8.5 Verifying connectivity

This section shows the ping syntax used to validate connectivity between hosts. These commands are run from the ESXi CLI.

8.5.1 vSAN networks

On vSAN networks in this deployment, ESXi hosts must be able to reach other ESXi hosts on the same cluster/segment. Hosts should not be able to reach hosts in other clusters/segments. The vSAN network should be able to support 9000 byte (jumbo) frames.

In the following example, host comp101 successfully pings host comp102's vSAN VMkernel IP address:

```
[root@comp101:~] ping 172.16.23.2 –d -s 8950
PING 172.16.23.2 (172.16.23.2): 8950 data bytes
8958 bytes from 172.16.23.2: icmp_seq=0 ttl=64 time=0.495 ms
8958 bytes from 172.16.23.2: icmp_seq=1 ttl=64 time=0.537 ms
8958 bytes from 172.16.23.2: icmp_seq=2 ttl=64 time=0.507 ms
```

The `–d` argument prevents packet fragmentation and the `–s 8950` argument sets the ICMP data size in the packet. (This size does not include IP headers, so it is set to slightly under 9000 bytes).

Note: If pings fail with jumbo frames, check the MTU size settings on the associated vDS and VMkernel adapters in vCenter.

8.5.2 vMotion networks

On vMotion networks in this deployment, ESXi hosts must be able to reach all other ESXi hosts on all segments/clusters, including other tenants. The vMotion network should be able to support 9000 byte frames.

In the following example, host comp101 successfully pings the vMotion VMkernel IP address of host mgmt01:

```
[root@comp101:~] ping 172.16.12.1 -S vmotion -s 8950 -d
PING 172.16.12.1 (172.16.12.1): 8950 data bytes
8958 bytes from 172.16.12.1: icmp_seq=0 ttl=61 time=0.445 ms
8958 bytes from 172.16.12.1: icmp_seq=1 ttl=61 time=0.469 ms
8958 bytes from 172.16.12.1: icmp_seq=2 ttl=61 time=0.426 ms
```

The `–S vmotion` argument instructs the utility to use the vMotion TCP/IP stack. This argument is required for the command to succeed on vMotion networks.

The `–d` argument prevents packet fragmentation and the `–s 8950` argument sets the ICMP data size in the packet. (This size does not include IP headers, so it is set to slightly under 9000).

Note: If pings fail with jumbo frames, check the MTU size settings on the associated vDS and VMkernel adapters in vCenter.

9 Enable vSAN on clusters

This section provides a brief outline of the steps to enable vSAN on clusters in this deployment. For a list of vSAN resources, see Appendix D.3.2.

Servers used in vSAN clusters must either have a mix of flash (SSD) and magnetic (HDD) drives or be all-flash. See Appendix B.2 for the servers and disks used in this deployment and the *VMware vSAN Design and Sizing Guide* for storage requirements and guidance.

For redundancy, vSANs employ software RAID. With the exception of single drive RAID-0 configurations, vSANs do not support hardware RAID. The PowerEdge R740xd servers used in this deployment each have HBA330 controllers which do not support hardware RAID. The PowerEdge R630 servers used have PERC H730 controllers which support hardware RAID, but the controllers are set to HBA (non-RAID) mode. Some controllers may refer to this as pass-through mode. See your system documentation for storage controller settings.

Note: For systems using H730 storage controllers, see VMware KB article 213674 for more information. For other controllers, see VMware KB article 2129050.

vSAN is enabled on each cluster by following the instructions in the *VMware vSAN Operations Guide*. When configuring a vSAN for the Compute-Edge cluster in this deployment, the network validation page appears as shown in Figure 93. This confirms that VMware and BCF are properly configured for vSAN network functionality for this cluster. Network validation pages for the remaining two clusters are similar.

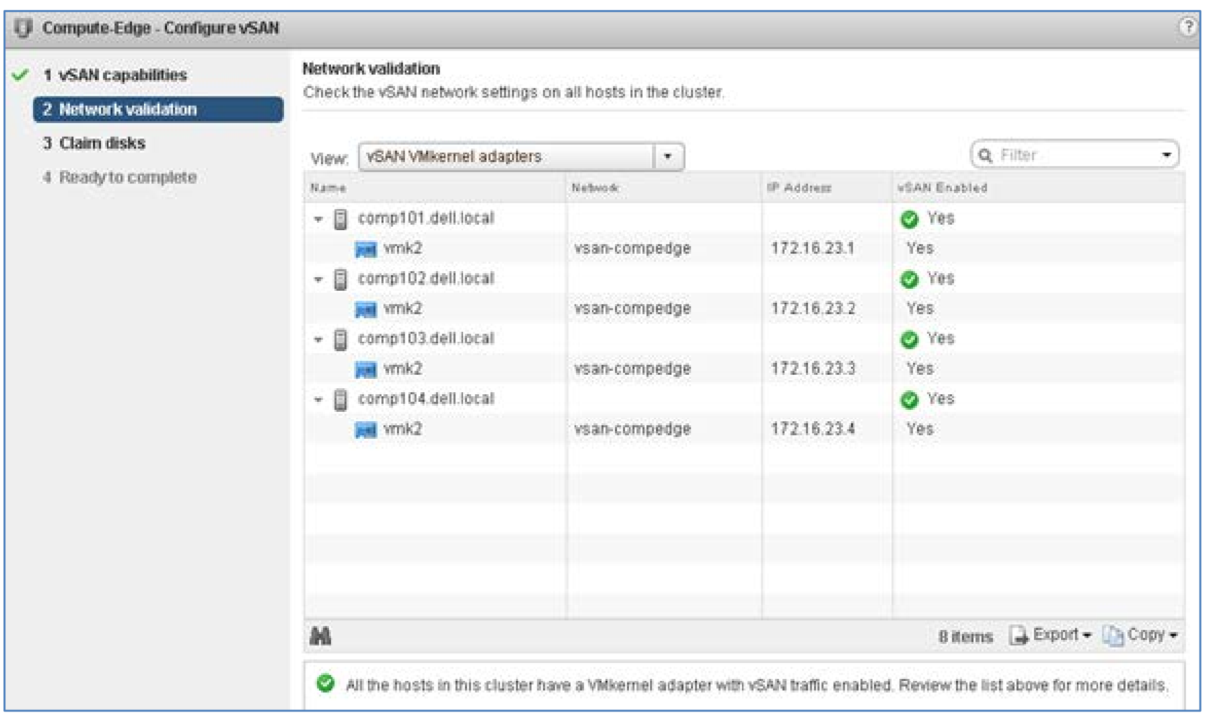


Figure 93 vSAN network validation page for the Compute-Edge cluster

Note: As of vSAN 6.6, vSAN communication is done via unicast. However, multicast is required for earlier versions of vSAN. This is accomplished by enabling multicast in BCF as covered in Section 7.1 of this guide.

After vSAN is configured on a cluster, a datastore named vsanDatastore appears on the **Navigator** pane **> Storage** tab under its corresponding data center**.** In Figure 94, the three vSAN datastores created have been renamed to **CompEdgevsanDatastore, CompvsanDatastore** and **MgmtvsanDatastore** respectively to correspond with their cluster names for usability.

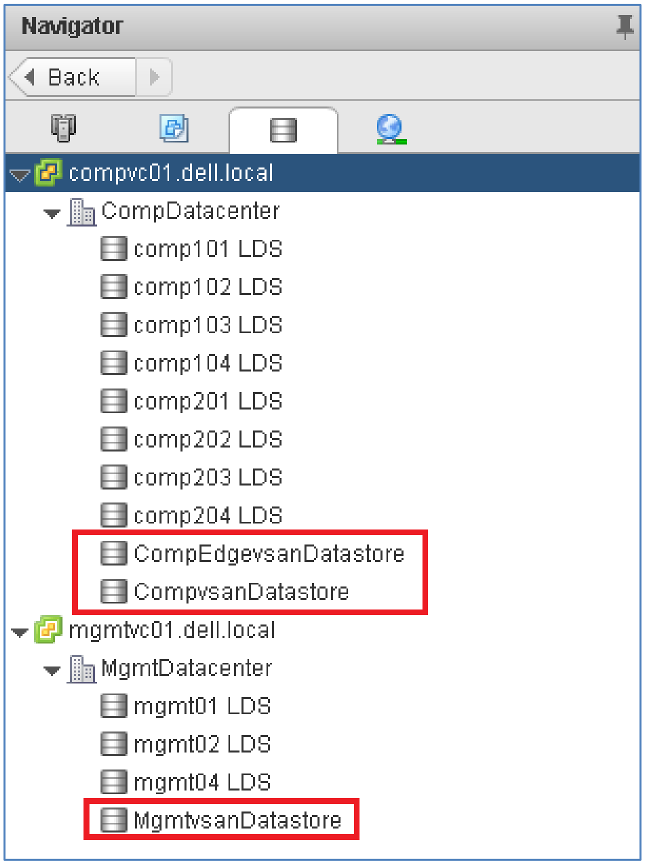


Figure 94 vSAN datastores created

Other datastores shown in Figure 94, named *hostname##* LDS, are single-disk local datastores on each host. Until vSANs are created, VMs initially reside on local datastores.

After configuring vSAN clusters, Dell EMC recommends:

1. Monitoring vSAN Health and following VMware remediation steps when applicable.
2. Enabling vSphere DRS and HA on each cluster.
3. Migrating VMs from local datastores to vSAN datastores using vMotion to take advantage of DRS and HA features.

Note: Refer to [VMware Documentation]() online for more information on DRS and HA features as well as VM migration.

10 Deploy VMware NSX

This section provides an overview of NSX deployment and the settings used for the example in this guide.

> **Note:** For detailed NSX deployment steps, see the *VMware NSX for vSphere 6.3 Installation Guide*. For NSX design considerations, see VMware Validated Design Documentation.

The NSX deployment process is as follows:

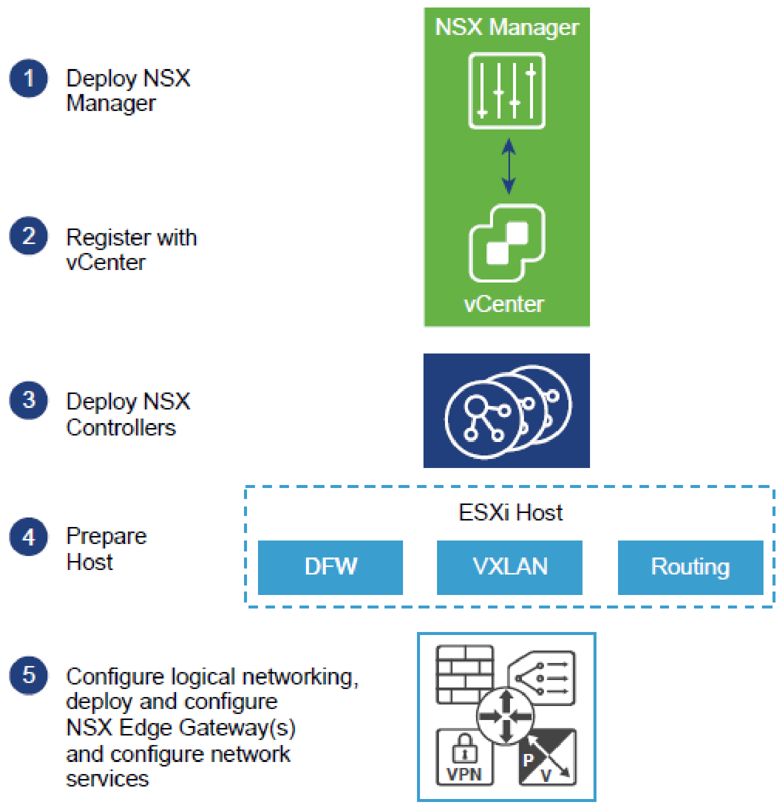


Figure 95 NSX deployment process

NSX components deployed include NSX Managers, NSX Controllers, and NSX Edges. NSX Edges include DLRs and ESGs. They are deployed to the Management and Compute-Edge clusters as shown earlier in Figure 15.

10.1 Deploy NSX Managers

An 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 NSX Edges.

Two NSX Managers are deployed, one for each vCenter. The Management NSX Manager serves the Management vCenter, and the Compute NSX Manager serves the compute vCenter.

Both NSX Managers are installed as virtual appliances in the Management cluster. NSX Manager is available from VMware as an Open Virtualization Appliance file named VMware-NSX-Manager-*version#*.ova.

The settings shown in the table are used during NSX Manager deployment:

Table 23 NSX Manager deployment settings

	Management NSX Manager	**Compute NSX Manager**
Name	nsxmgr-mgmt	nsxmgr-comp
Location	MgmtDatacenter	MgmtDatacenter
Resource	Management cluster	Management cluster
Storage	MgmtvsanDatastore	MgmtvsanDatastore
Destination Network	VM Network	VM Network
DNS server list	100.67.189.33	100.67.189.33
Domain search list	dell.local	dell.local
Default Gateway	100.67.187.254	100.67.187.254
Hostname	nsxmgr-mgmt.dell.local	nsxmgr-comp.dell.local
IP Address	100.67.187.180	100.67.187.181
Netmask	255.255.255.0	255.255.255.0
NTP Server	100.67.10.20	100.67.10.20

Note: Even though they serve different vCenters, both NSX Managers are installed in the Management cluster of the Management vCenter because NSX Manager is considered a management component.

Note: The default VMware network, named VM Network, is the OOB management network in this guide.

After NSX Managers are deployed, their VMs appear on the **Navigator** pane **> Hosts and Clusters** tab as shown:

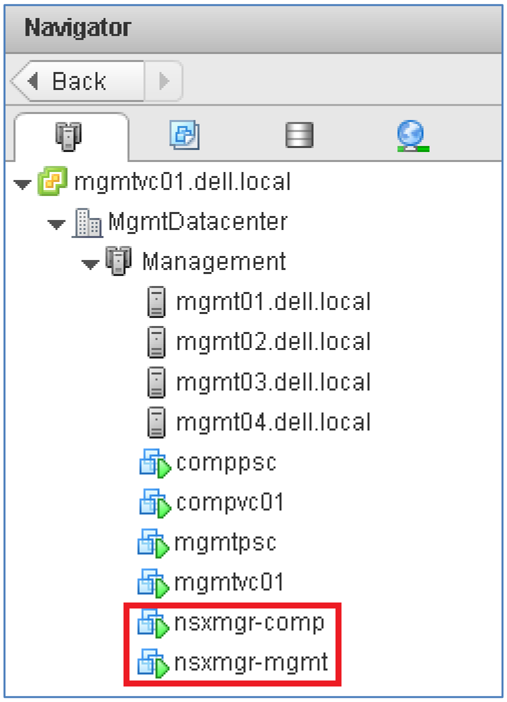


Figure 96 NSX Manager VMs installed and powered on

10.2 Register NSX Managers with vCenter Severs

After NSX Managers are installed, each NSX Manager is registered with its corresponding vCenter Server per Table 24.

Table 24 NSX Managers and vCenter Servers

NSX Manager	vCenter Server
nsxmgr-mgmt.dell.local	mgmtvc01.dell.local
nsxmgr-comp.dell.local	compvc01.dell.local

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

10.3 Deploy NSX Controller clusters

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

NSX Controllers communicate on the OOB management network, named VM Network in this guide, and do not have any data plane traffic passing through them. Therefore, data forwarding will continue even if all NSX Controllers are offline.

As a best practice, each NSX Controller should be deployed to a different ESXi host so that a single host failure will not bring down more than one controller. After deployment, VM/host rules (a.k.a. affinity rules) are created to keep NSX Controller VMs on different hosts.

NSX Controllers for the Management NSX Manager are deployed to hosts in the Management cluster and NSX Controllers for the Compute NSX Manager are deployed to hosts in the Compute-Edge cluster. The settings used are shown in the following table:

Table 25 NSX Controller deployment settings

Controller name	NSX Manager	Data center	Cluster	Datastore	Connected To	IP Pool
mgmt-controller-1	100.67.187.180	MgmtDatacenter	Management	Mgmt vsanDatastore	VM Network	mgmt nsx pool (Table 26)
mgmt-controller-2	100.67.187.180	MgmtDatacenter	Management	Mgmt vsanDatastore	VM Network	mgmt nsx pool
mgmt-controller-3	100.67.187.180	MgmtDatacenter	Management	Mgmt vsanDatastore	VM Network	mgmt nsx pool
comp-controller-1	100.67.187.181	CompDatacenter	Compute-Edge	CompEdge vsanDatastore	VM Network	comp nsx pool (Table 27)
comp-controller-2	100.67.187.181	CompDatacenter	Compute-Edge	CompEdge vsanDatastore	VM Network	comp nsx pool
comp-controller-3	100.67.187.181	CompDatacenter	Compute-Edge	CompEdge vsanDatastore	VM Network	comp nsx pool

A controller IP pool is created during deployment of the first controller for each NSX Manager. The settings used are shown in Table 26 and Table 27.

Table 26 IP pool settings for the Management NSX Manager

Field	Value
Name	mgmt nsx pool
Gateway	100.67.187.254
Prefix Length	24
Static IP Pool	100.67.187.182-100.67.187.184

Table 27 IP pool settings for the Compute NSX Manager

Field	Value
Name	comp nsx pool
Gateway	100.67.187.254
Prefix Length	24
Static IP Pool	100.67.187.185-100.67.187.187

After NSX Controllers are deployed, the **NSX Controller nodes** section of the **Home > Networking & Security > Installation > Management** page appears as shown:

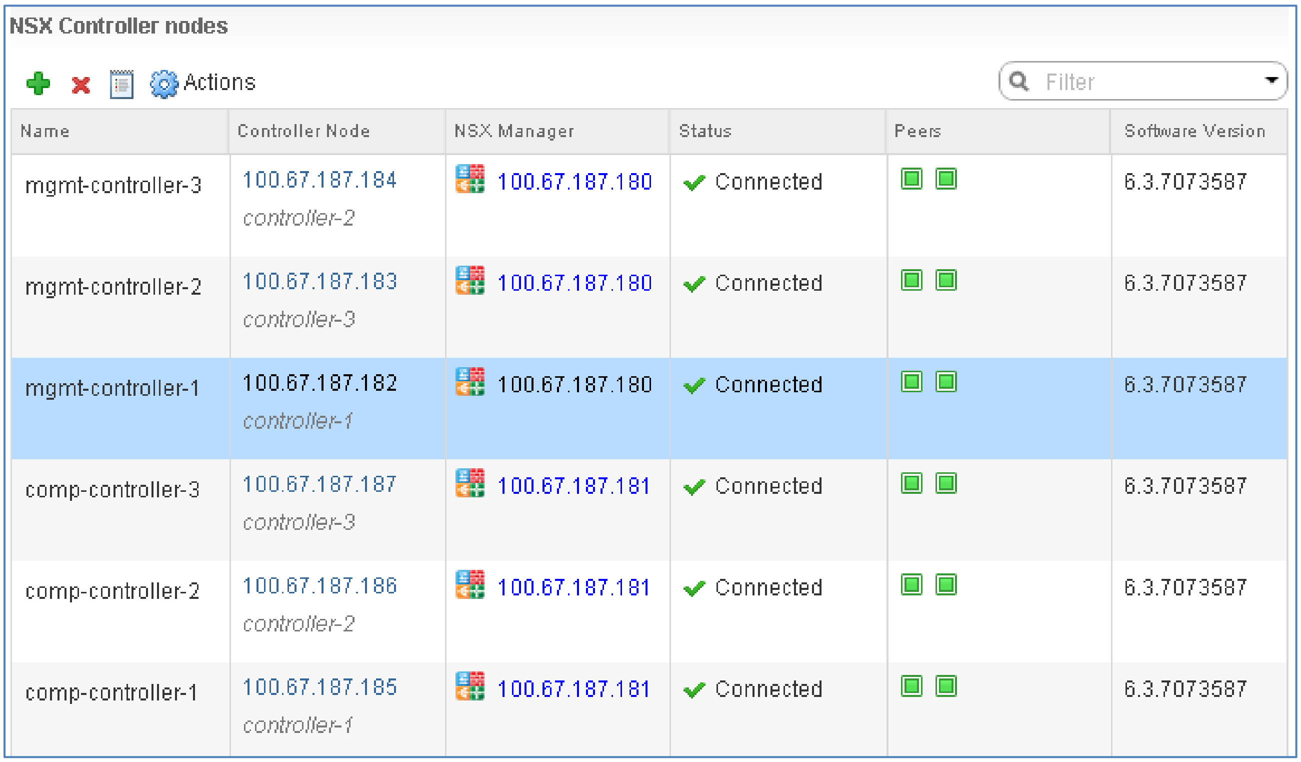


Figure 97 NSX Controllers deployed

For fault tolerance, create **VM/Host** rules to keep NSX Controller VMs on separate hosts. One rule is created for each cluster containing NSX Controllers. The rule for the Compute-Edge cluster is shown in Figure 98:

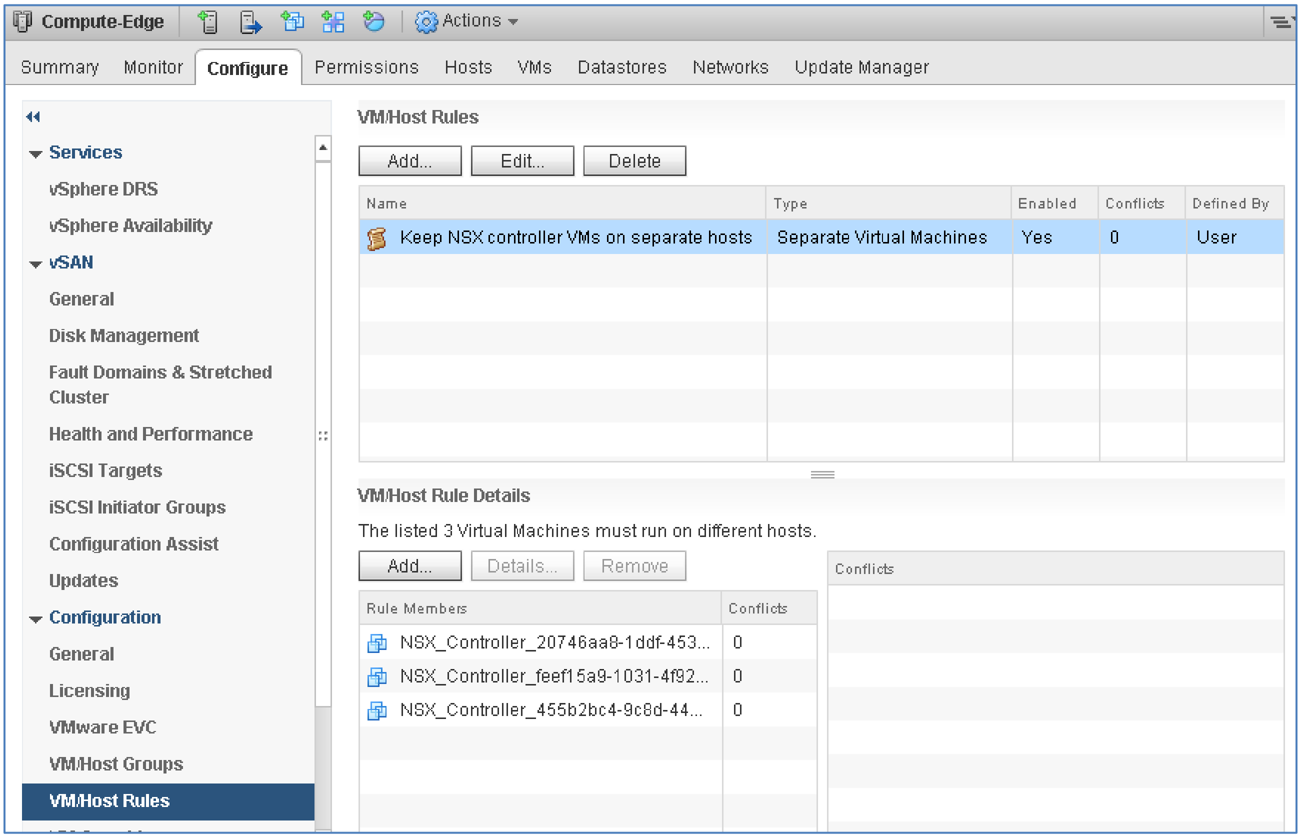


Figure 98 VM/Host rule for NSX Controllers in the Compute-Edge cluster created

10.4 Prepare host clusters for NSX

Host preparation is the process where the NSX Manager installs NSX kernel modules on each ESXi host in a cluster. This is performed on clusters that will participate on NSX networks.

When complete, the **Home > Networking & Security > Installation > Host Preparation** page for each NSX Manager appears as shown in Figure 99 and Figure 100. Notice that the IP address of the applicable NSX Manager, outlined in red, is selected to switch between NSX Managers.

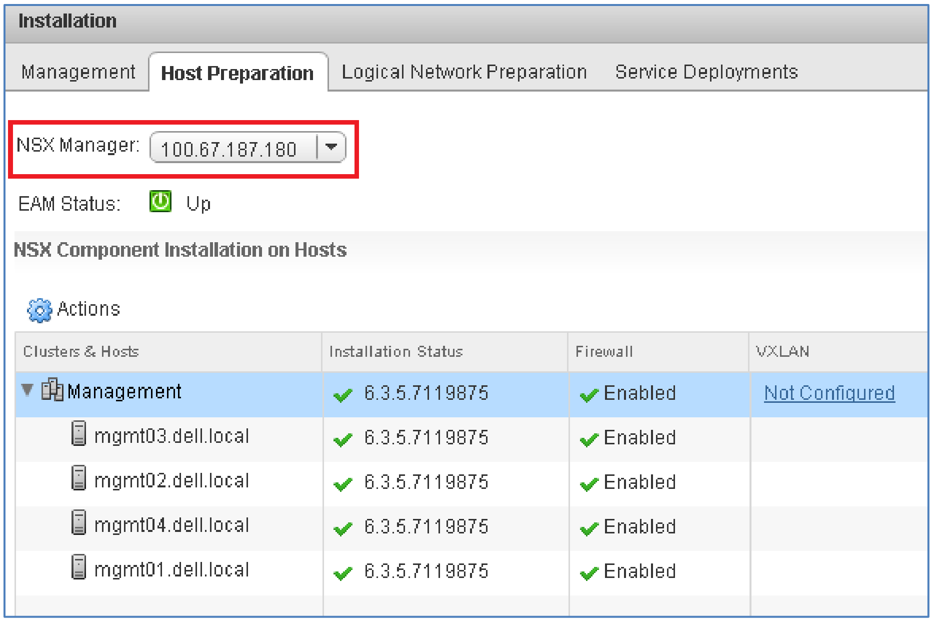


Figure 99 Hosts in Management cluster configured

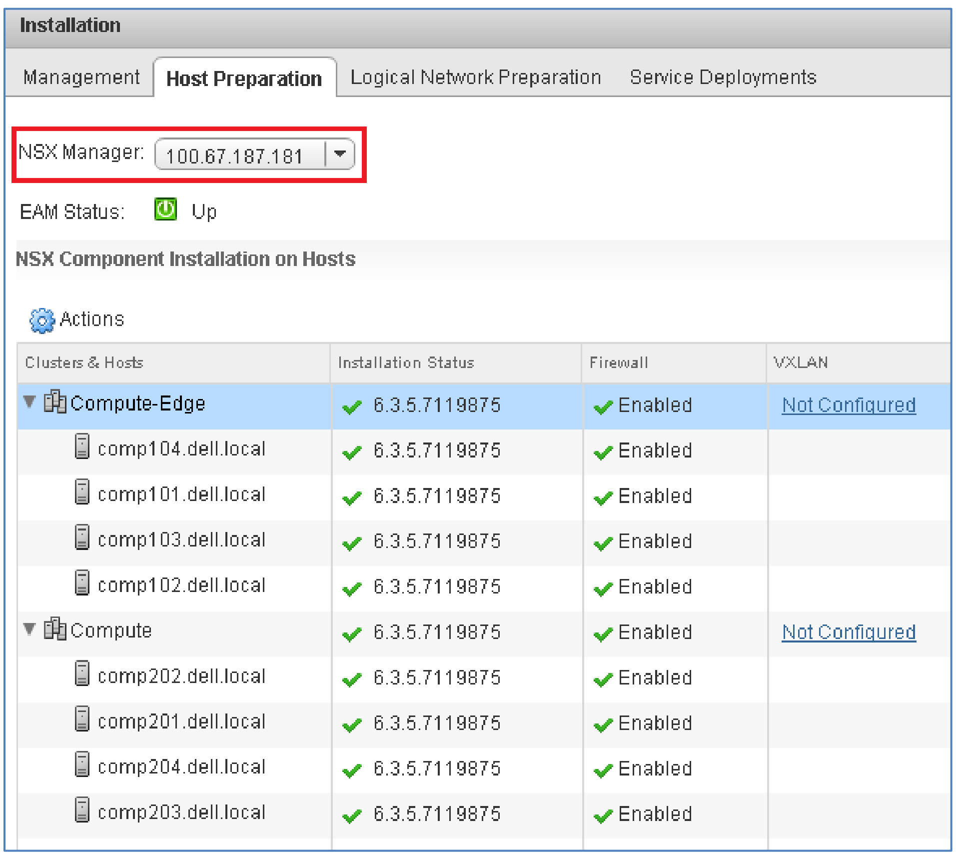


Figure 100 Hosts in Compute-Edge and Compute clusters configured

The host preparation process installs a vSphere Installation Bundle (VIB) to each host in the cluster named esx-nsxv. This may be confirmed by running the following command on hosts in a configured cluster:

```
[root@comp101:~] esxcli software vib list | grep nsx
esx-nsxv        6.5.0-0.0.7119877     VMware   VMwareCertified   2018-01-23
```

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

10.5 Configure VXLAN transport parameters

VXLAN is configured on a per-cluster basis with each cluster mapped to a vDS. VXLAN configuration creates VMkernel interfaces that serve as VTEPs on each host. This enables virtual network functionality on each host in the cluster.

The VVD-recommended teaming policy recommended for VXLAN is **Route Based on Source ID** (displays as **Load Balance – SRCID** in NSX). This teaming policy creates 2 VTEPs per host. Each VTEP is assigned a VMkernel IP address from an IP pool. Therefore, the number of addresses in the pool must be enough to cover 2 VTEPs per host in the cluster.

The VXLAN networking settings used for this deployment are shown in Table 28.

Table 28 VXLAN networking settings

Cluster	Switch	VLAN	MTU	IP Pool Settings			VMKNic teaming policy
				Gateway	Prefix length	Address range	
Management	vDS-Mgmt	1614	9000	172.16.14.254	24	172.16.14.1-172.16.14.20	Load Balance - SRCID
Compute-Edge	vDS-CompEdge	1624	9000	172.16.24.254	24	172.16.24.1-172.16.24.100	Load Balance - SRCID
Compute	vDS-Comp	1634	9000	172.16.34.254	24	172.16.34.1-172.16.34.100	Load Balance - SRCID

Note: The MTU value is increased from its default of 1600 bytes to 9000 bytes for best performance per VVD. The gateway addresses shown are configured as BCF segment interfaces in Section 11.2.

When complete, the **VXLAN** column indicates **Configured** on the **Home > Networking & Security > Installation > Host Preparation** page for each NSX Manager.

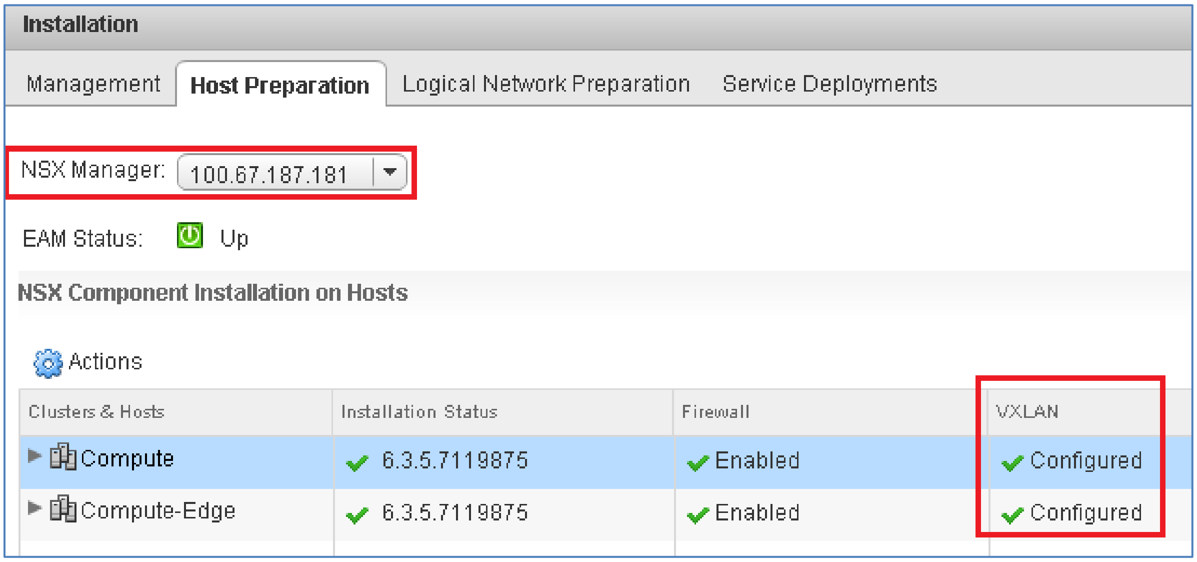


Figure 101 VXLAN configured for Compute and Compute-Edge clusters (Management cluster is similar)

The **Logical Network Preparation** page for each NSX Manager in this deployment appears as shown in Figure 102 and Figure 103. Each host has two VMkernel adapters added that act as VTEPs. IP addresses shown are allocated from the configured pools.

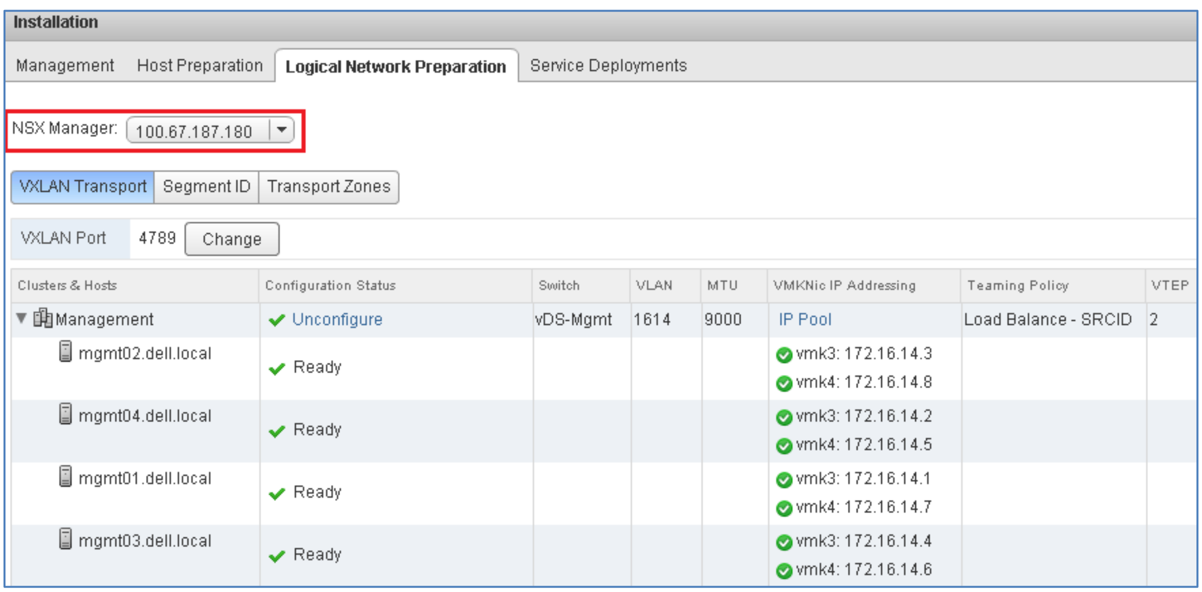


Figure 102 VXLAN transport configuration complete for the Management cluster

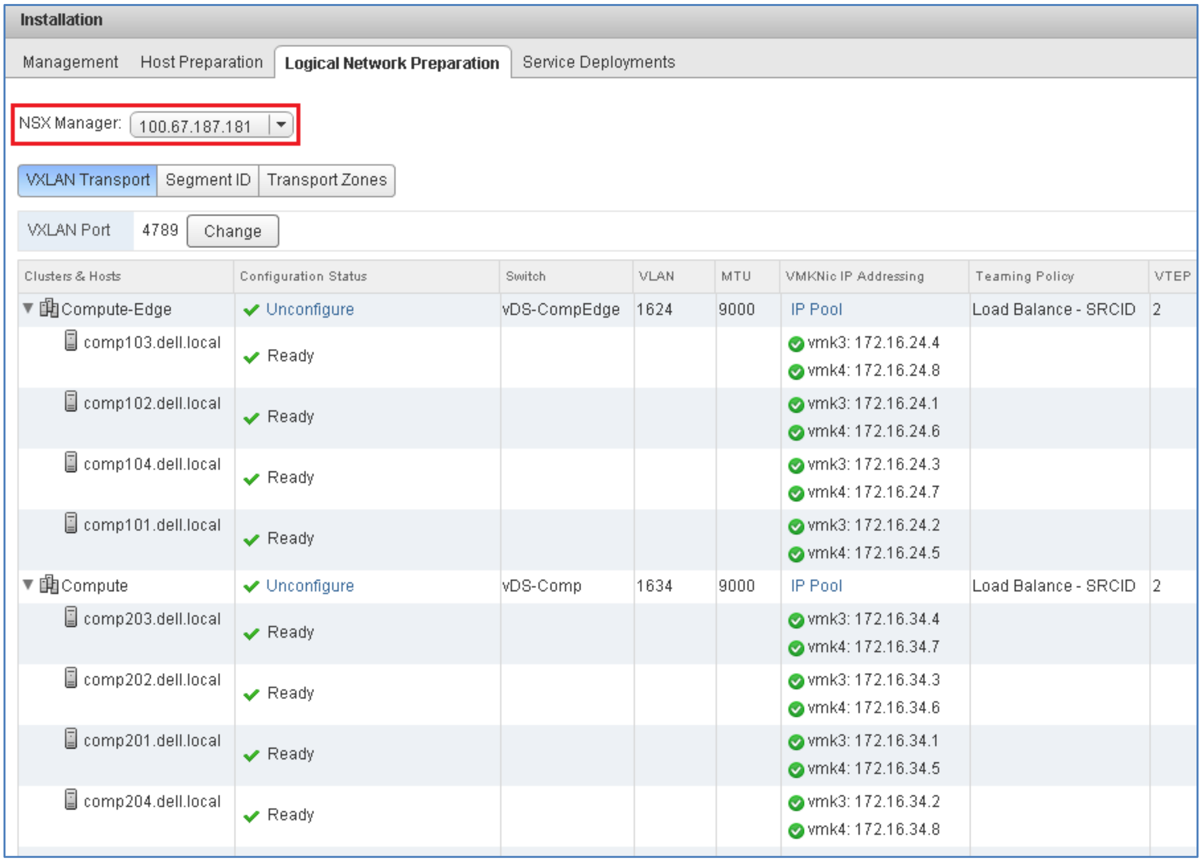


Figure 103 VXLAN transport configuration complete for the Compute and Compute-Edge clusters

10.6 Configure segment ID pools and multicast addresses

VXLAN tunnels are established between VTEPs. Each VXLAN tunnel must have a segment ID, which is pulled from a segment ID pool. Segment IDs are used as VNIs, and each logical switch created receives a segment ID from the pool. The range of valid segment IDs is 5000-16777215.

It is a best practice that segment ID numbers in pools on different NSX Managers do not overlap. Using non-overlapping segment ID ranges helps with tracking and ensures deployments are ready for a cross-vCenter environment if configured at a later date.

Multicast addressing is enabled for hybrid control plane replication mode. Hybrid replication mode offloads broadcast, unknown unicast, and multicast (BUM) traffic to the physical network which reduces pressure on VTEPs as the environment scales out. This is recommended by VVD for best performance in large environments.

Note: Hybrid mode requires physical switches have IGMP snooping enabled and that an IGMP querier is present on the network. Enabling multicast in the vCenter tenants in BCF accomplishes this. This is accomplished by enabling multicast in BCF in section 7.1 of this guide.

The segment ID pool and multicast settings for both NSX Managers used in this deployment are shown in Table 29.

Table 29 Segment ID pool settings

NSX Manager	NSX Manager IP	Segment ID pool	Multicast	Multicast address range
Management	100.67.187.180	5000-5999	Enabled	239.5.0.0-239.5.255.255
Compute	100.67.187.181	6000-6999	Enabled	239.6.0.0-239.6.255.255

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

When complete, the **Home > Networking & Security > Installation > Logical Network Preparation > Segment ID** page for the Compute NSX Manager appears as shown in Figure 104. The page for the Management NSX Manager is similar.

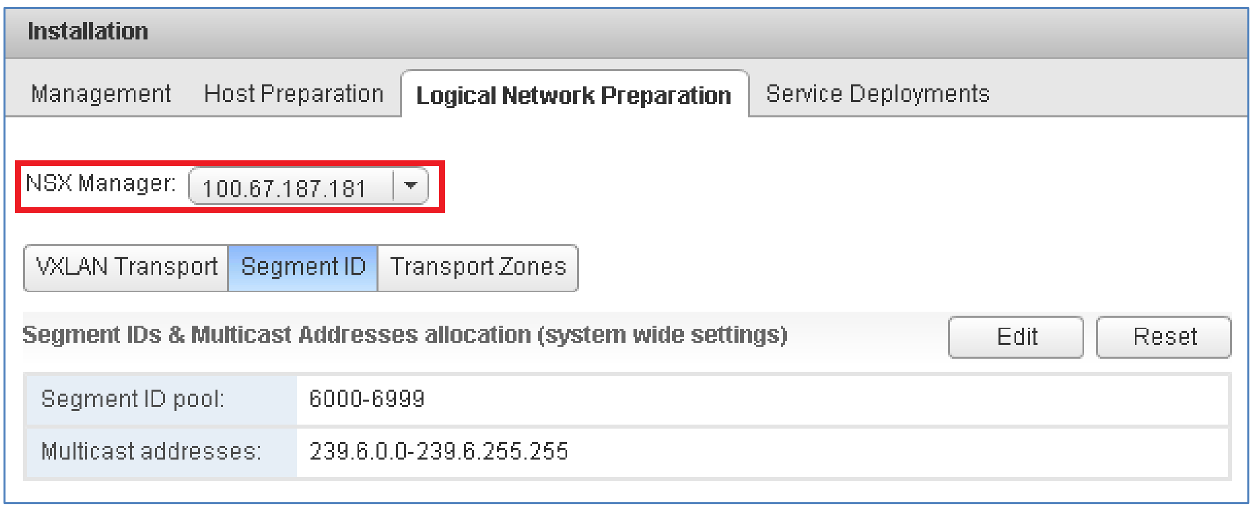


Figure 104 Segment ID page, Compute NSX Manager selected

10.7 Configure transport zones

A transport zone dictates which clusters and, therefore, which VMs can use a particular virtual network. It can span one or more clusters within one vCenter Server domain. NSX does not allow connection of VMs that are in different transport zones, and DLRs cannot connect to logical switches located in different transport zones.

A cluster can belong to multiple transport zones while a logical switch can belong to only one.

This deployment uses a single transport zone for each NSX Manager. Each transport zone is set to Hybrid replication mode as covered in the preceding section. After creating each transport zone, enable Controller Disconnected Operation (CDO) mode as recommended in VVD. CDO mode ensures that data plane connectivity is unaffected when hosts lose connectivity with the controller.

The transport zone settings used in this deployment are shown in Table 30.

Table 30 Transport zone settings

NSX Manager	NSX Manager IP	Transport zone name	Replication Mode	Cluster	CDO Mode
Management	100.67.187.180	Mgmt	Hybrid	Management	Enabled
Compute	100.67.187.181	Comp	Hybrid	Compute, Compute-Edge	Enabled

When complete, the **Home > Networking & Security > Installation > Logical Network Preparation > Transport Zones** page for the Management NSX Manager appears as shown in Figure 105. The page for the Compute NSX Manager is similar.

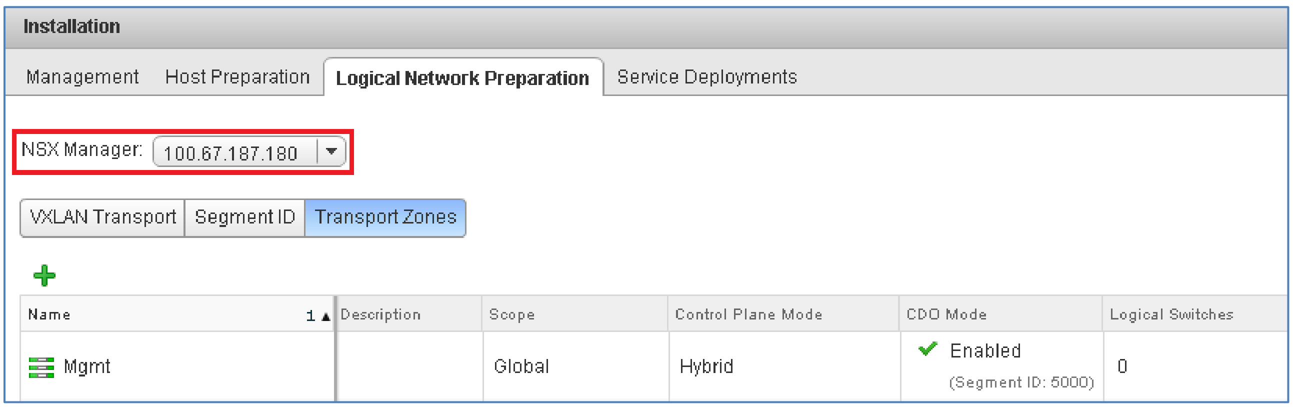


Figure 105 Transport zone configured for NSX Manager 100.67.187.180

10.8 Configure logical switches

An NSX logical switch reproduces switching functionality in a virtual environment that is completely decoupled from underlying hardware. An NSX logical switch creates a broadcast domain similar to a physical switch or a VLAN. A single logical switch is mapped to a unique VXLAN segment and is distributed across the ESXi hypervisors within a transport zone. The logical switch allows line-rate switching in the hypervisor without the constraints of VLAN sprawl or Spanning Tree Protocol issues.

This deployment creates two logical switches in the Management transport zone and three logical switches in the Compute transport zone.

The settings shown in Table 31 are used in this deployment. The replication mode is set to Hybrid as covered in section 10.6. The IP Discovery and MAC Learning settings are logical switch default values.

Table 31 Logical switch configuration settings

NSX Manager	NSX Manager IP	Logical switch name	Transport zone	Replication mode	IP Discovery	MAC Learning
Management	100.67.187.180	Transit	Mgmt	Hybrid	Enabled	Disabled
Management	100.67.187.180	Mgmt	Mgmt	Hybrid	Enabled	Disabled
Compute	100.67.187.181	Transit	Comp	Hybrid	Enabled	Disabled
Compute	100.67.187.181	App	Comp	Hybrid	Enabled	Disabled
Compute	100.67.187.181	Web	Comp	Hybrid	Enabled	Disabled

Note: There is a logical switch named Transit in each of the Transport zones, Mgmt and Comp. These are not the same switch.

When complete, the **Home > Networking & Security > Logical Switches** page for each NSX Manager in this deployment appears as shown in Figure 106 and Figure 107.

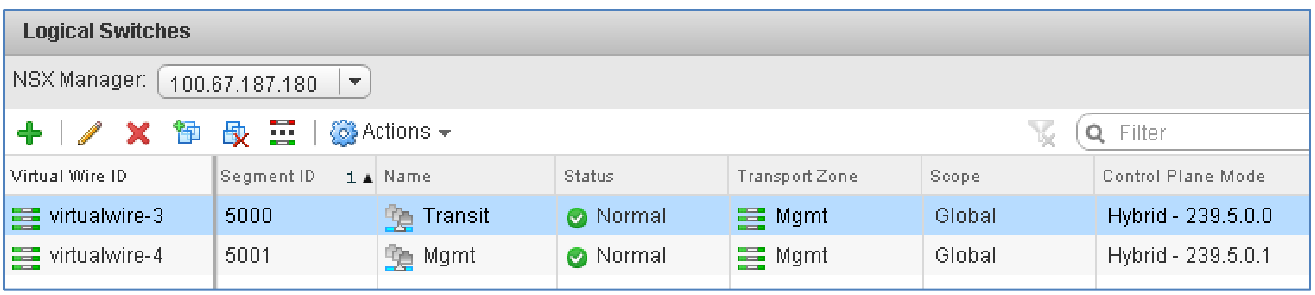


Figure 106 Logical switches created in Mgmt transport zone

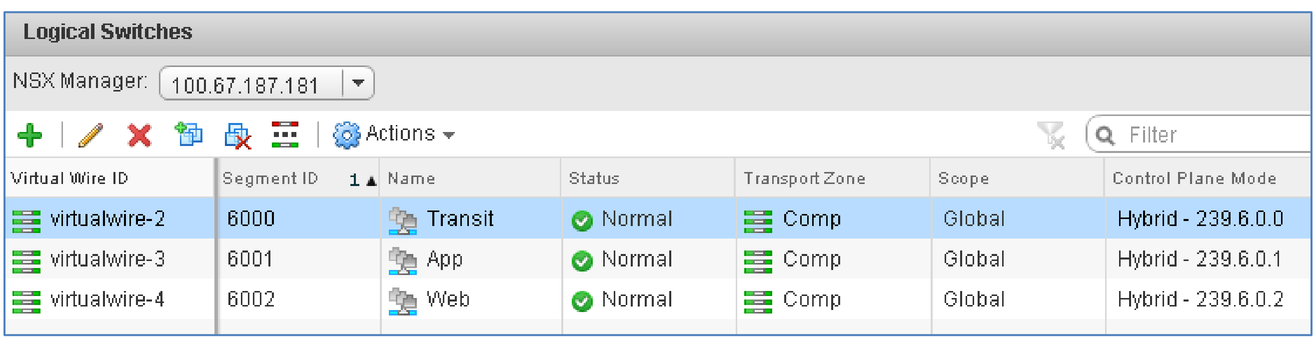


Figure 107 Logical switches created in Comp transport zone

The **Home > Networking** tab in the **Navigator** pane now appears as shown in Figure 108. A port group, or virtual wire, is automatically created for each logical switch. These are outlined in red.

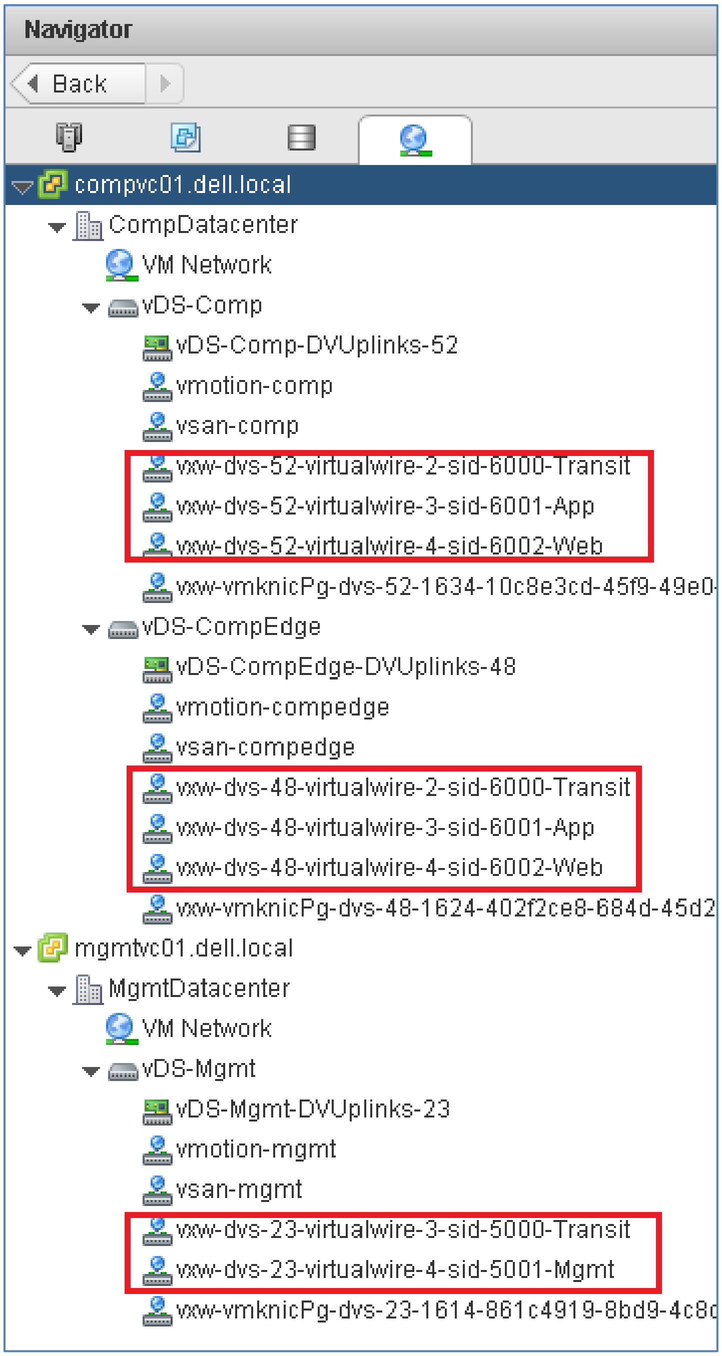


Figure 108 Port groups created for logical switches

10.9 Connect VMs to logical switches

The IP address and gateway settings shown in Table 32 are configured in the guest OS on each VM. Each VM is connected to the logical switch listed in the right column.

Note: Gateway addresses shown are DLR interfaces configured in the next section.

Table 32 VM network configuration

VM Name	Cluster	VM IP Address	Gateway	NSX Manager	Logical switch
mgmt-01	Management	10.1.1.1/24	10.1.1.254	100.67.187.180	Mgmt
mgmt-02	Management	10.1.1.2/24	10.1.1.254	100.67.187.180	Mgmt
app-01	Compute-Edge	10.1.2.1/24	10.1.2.254	100.67.187.181	App
app-02	Compute	10.1.2.2/24	10.1.2.254	100.67.187.181	App
web-01	Compute-Edge	10.1.3.1/24	10.1.3.254	100.67.187.181	Web
web-02	Compute	10.1.3.2/24	10.1.3.254	100.67.187.181	Web

Figure 109 shows the VMs app-01 and app-02 connected to the App logical switch. This page is accessed by navigating to **Home > Networking & Security > Logical Switches**. Select the **NSX Manager** and double click on the logical switch. Select **Related Objects > Virtual Machines**.

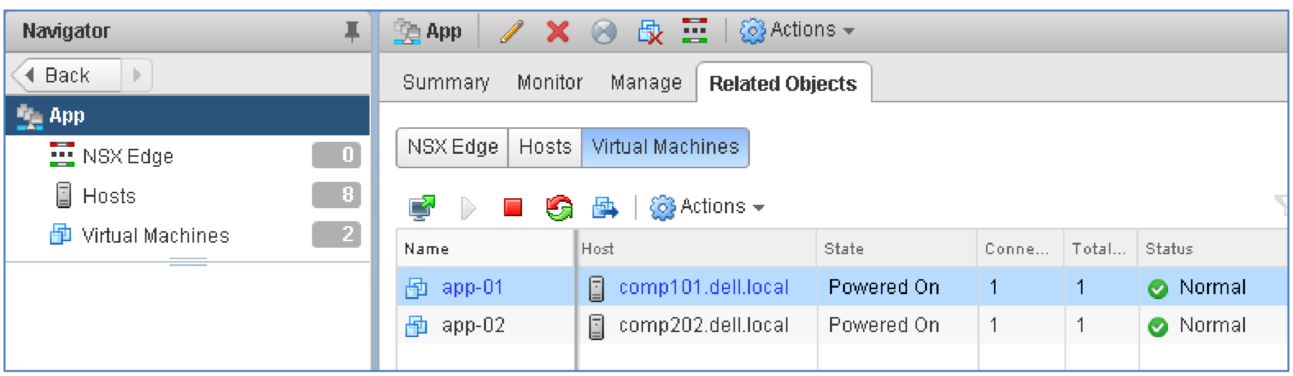


Figure 109 App VMs connected to App logical switch

10.10 Deploy DLRs

A distributed logical router (DLR) is an NSX Edge appliance that contains the routing control plane, while distributing the data plane in kernel modules to each hypervisor host. DLRs enable connectivity between virtual machines connected to different logical switches in the same transport zone.

DLRs provide high performance, low overhead first hop routing, scale with the number of hosts, and allow up to 1,000 Logical Interfaces (LIFs).

In this deployment, two DLRs are deployed: DLR-Mgmt and DLR-Comp. The DLRs and their associated connections are shown in Figure 110:

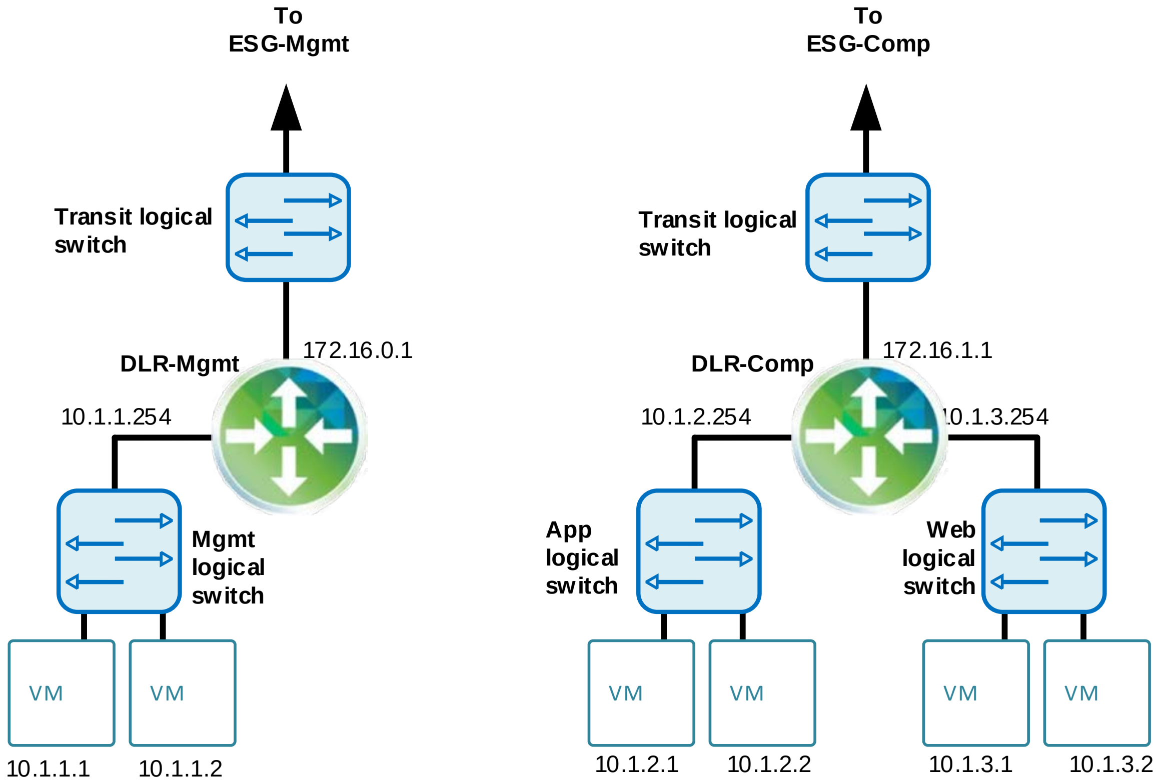


Figure 110 DLR-Mgmt and DLR-Comp

10.10.1 Deployment settings

The DLR deployment settings used are shown in Table 33.

Table 33 DLR deployment settings

NSX Manager	Install Type	DLR Name	Deploy Edge Appliance	Enable High Availability	Data center	Cluster/ Resource Pool	Datastore
100.67.187.180	Logical Router	DLR-Mgmt	Checked	Checked	Mgmt Data center	Management	Mgmtvsan Datastore
100.67.187.181	Logical Router	DLR-Comp	Checked	Checked	Comp Data center	Compute-Edge	CompEdgevsan Datastore

Each DLR is deployed with High Availability (HA) enabled. With HA enabled, NSX Manager deploys two identically configured VMs for each DLR instance. One VM is active and the other is standby. For example, DLR-Mgmt is deployed as two VMs named DLR-Mgmt-0 and DLR-Mgmt-1.

Note: The **Enable High Availability** option is required for dynamic routing.

During deployment of DLR-Mgmt and DLR-Comp, connect the HA interface to the **Transit** logical switch as shown in Figure 111. This is the interface that provides the heartbeat between the active and standby DLR VMs. No IP address needs to be specified for this interface.

Note: An IP address for each of the redundant VMs is automatically chosen from the link local address space, 169.250.0.0/16. No further configuration is necessary to configure the HA service.

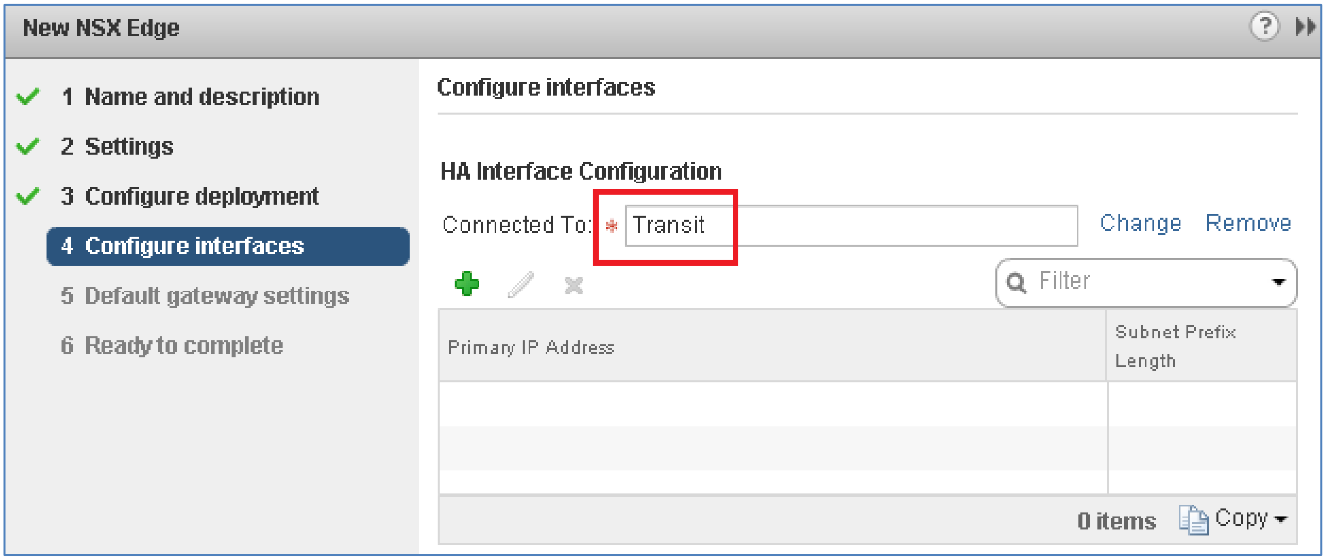


Figure 111 Connect HA interface to Transit logical switch

The settings in Table 34 and Table 35 are used to configure DLR interfaces:

Table 34 DLR-Mgmt interface settings

Interface name	Type	Connected to	IP address/ prefix	MTU
Transit	Uplink	Logical Switch - Transit	172.16.0.1/24	9000
Mgmt	Internal	Logical Switch - Mgmt	10.1.1.254/24	9000

Table 35 DLR-Comp interface settings

Interface name	Type	Connected to	IP address/ prefix	MTU
Transit	Uplink	Logical Switch - Transit	172.16.1.1/24	9000
App	Internal	Logical Switch - App	10.1.2.254/24	9000
Web	Internal	Logical Switch - Web	10.1.3.254/24	9000

Note: Default gateways are not configured on DLRs at this time. They are configured in the next section.

Figure 112 and Figure 113 show the DLR interface configuration settings when complete. The pages shown are accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the DLR name. Select **Manage > Settings > Interfaces.**

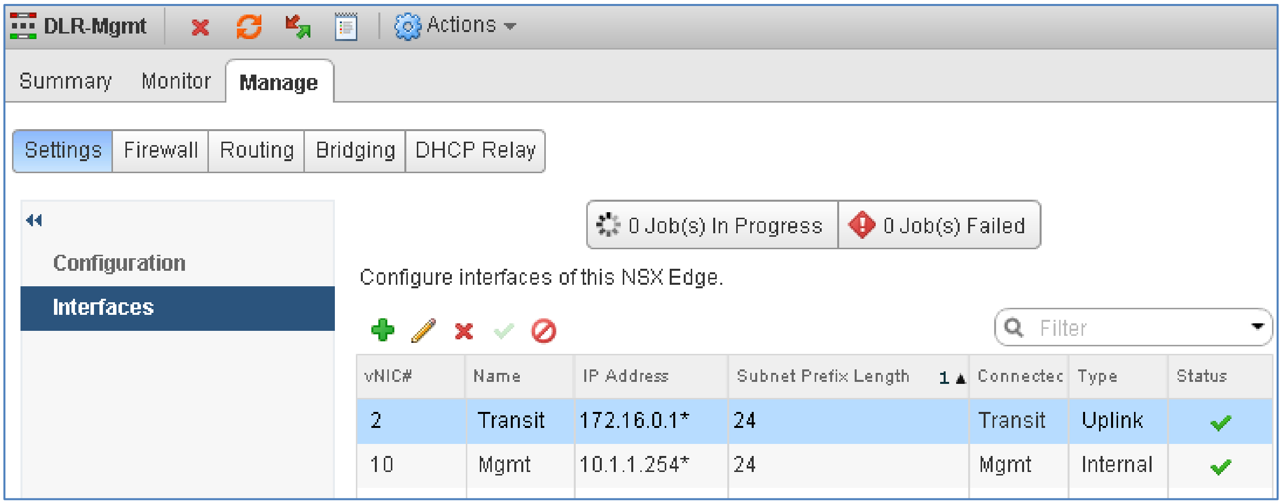


Figure 112 DLR-Mgmt interface configuration settings

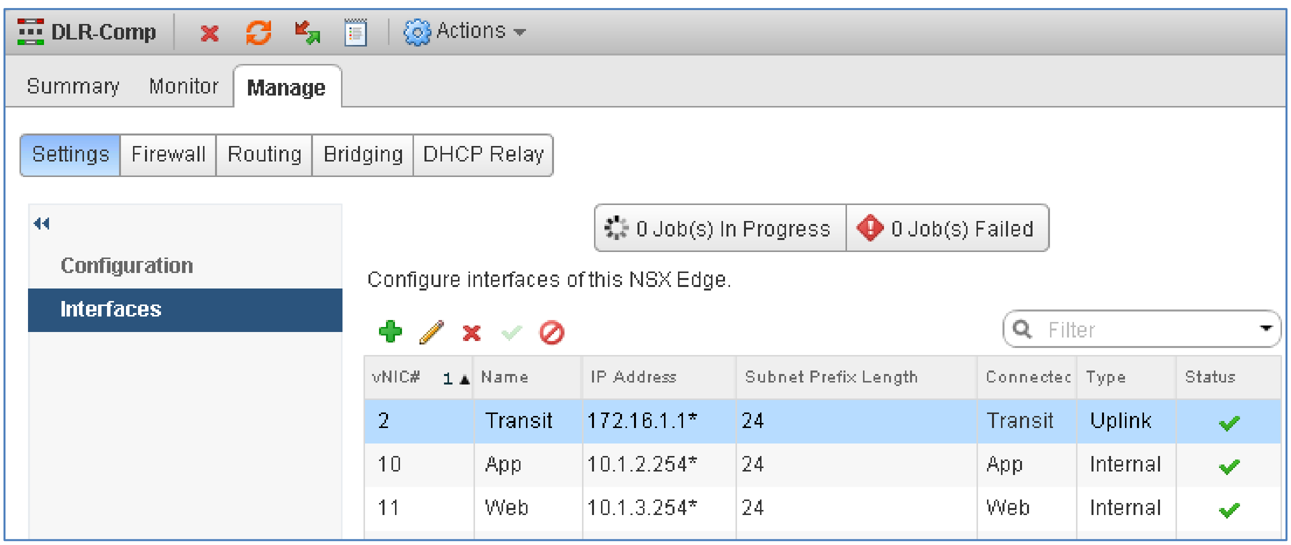


Figure 113 DLR-Comp interface configuration settings

With HA enabled, the bottom of the **Manage > Settings > Configuration** page for DLR-Comp appears as shown. The page for DLR-Mgmt is similar.

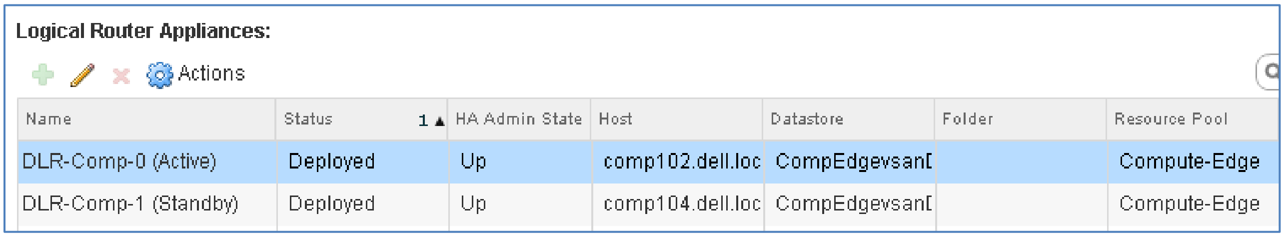


Figure 114 Active and Standby VMs deployed for DLR-Comp

On the **Hosts and Clusters** tab, redundant VMs for DLR-Comp and DLR-Mgmt are visible in their respective clusters:

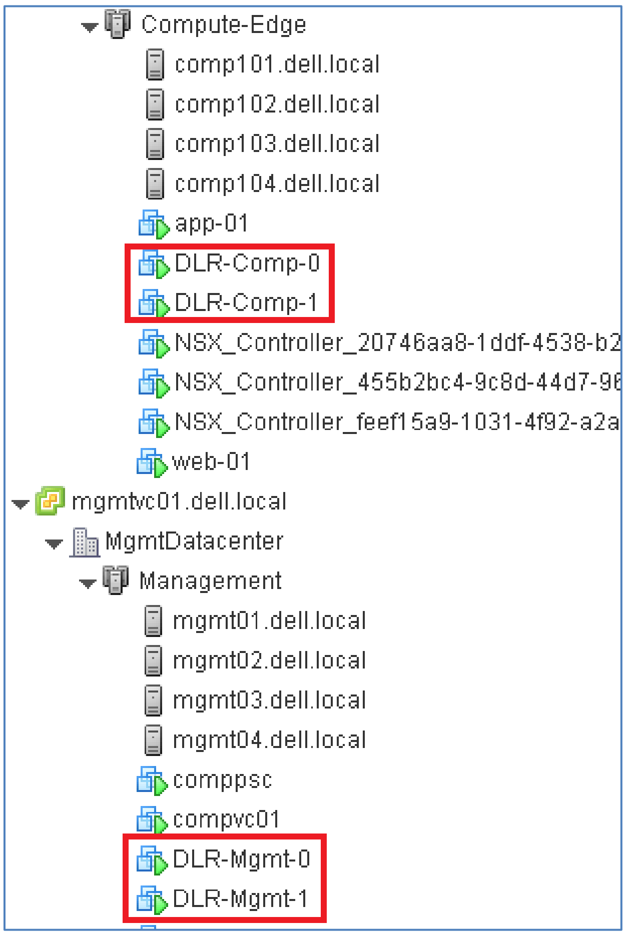


Figure 115 DLRs deployed with HA enabled

10.10.2 DLR global configuration settings

After deployment, global configuration settings are applied to each DLR per the following table:

Table 36 DLR global configuration settings

NSX Manager	DLR Name	ECMP	Default Gateway Interface	Default Gateway IP	MTU	Admin Distance	Router ID
100.67.187.180	DLR-Mgmt	Enabled	Transit	172.16.0.2	9000	1	172.16.0.1
100.67.187.181	DLR-Comp	Enabled	Transit	172.16.1.2	9000	1	172.16.1.1

Note: The default gateway IP is the ESG interface address configured in the next section.

The Global Configuration page is accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the DLR name. Select **Manage > Routing > Global Configuration.**

Figure 116 shows the global configuration settings for DLR-Comp when complete. The page for DLR-Mgmt is similar.

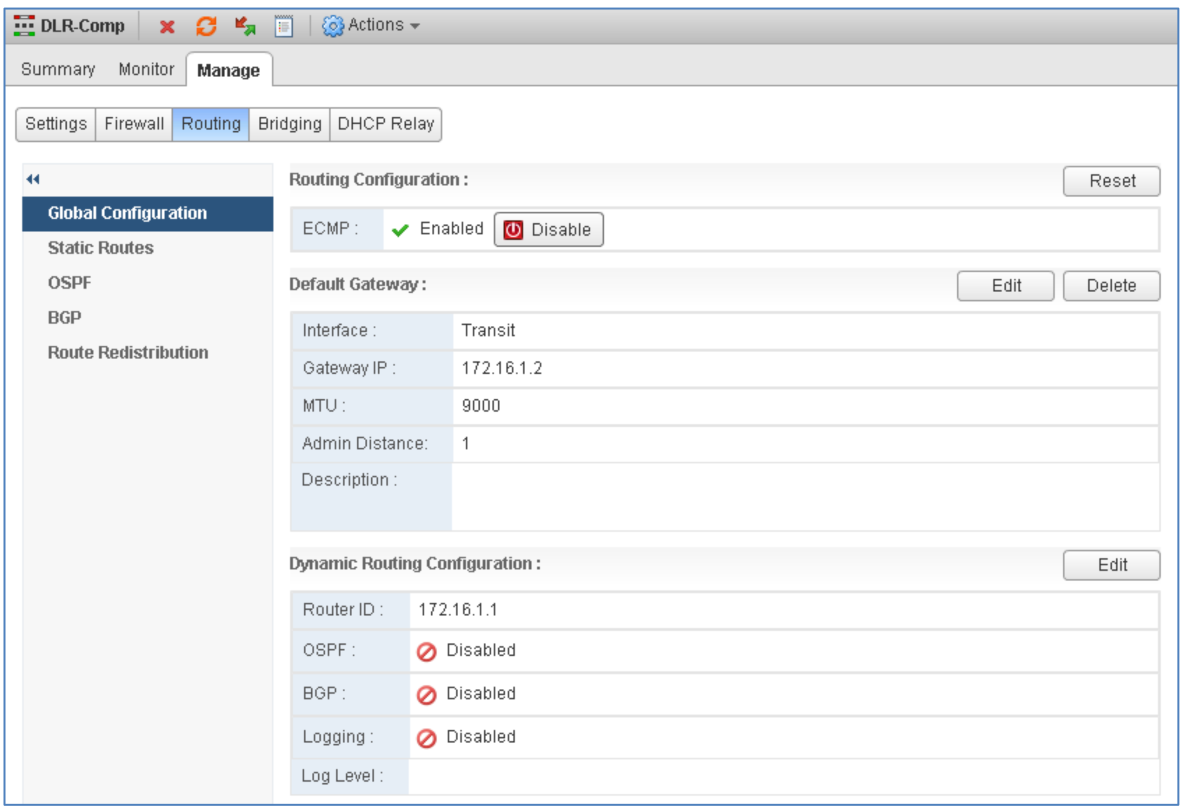


Figure 116 DLR-Comp global configuration settings complete

10.11 Deploy ESGs

Gateway services between VXLAN and non-VXLAN hosts (for example, a physical server) are performed by the NSX Edge Services Gateway (ESG) appliance. The ESG translates VXLAN segment IDs to VLAN IDs, so that non-VXLAN hosts can communicate with virtual machines on a VXLAN network. The NSX ESG's primary function is north-south communication, but it also offers firewall, load balancing and other services.

In this deployment, two ESGs are deployed: ESG-Mgmt and ESG-Comp. The ESGs and their associated connections are as shown:

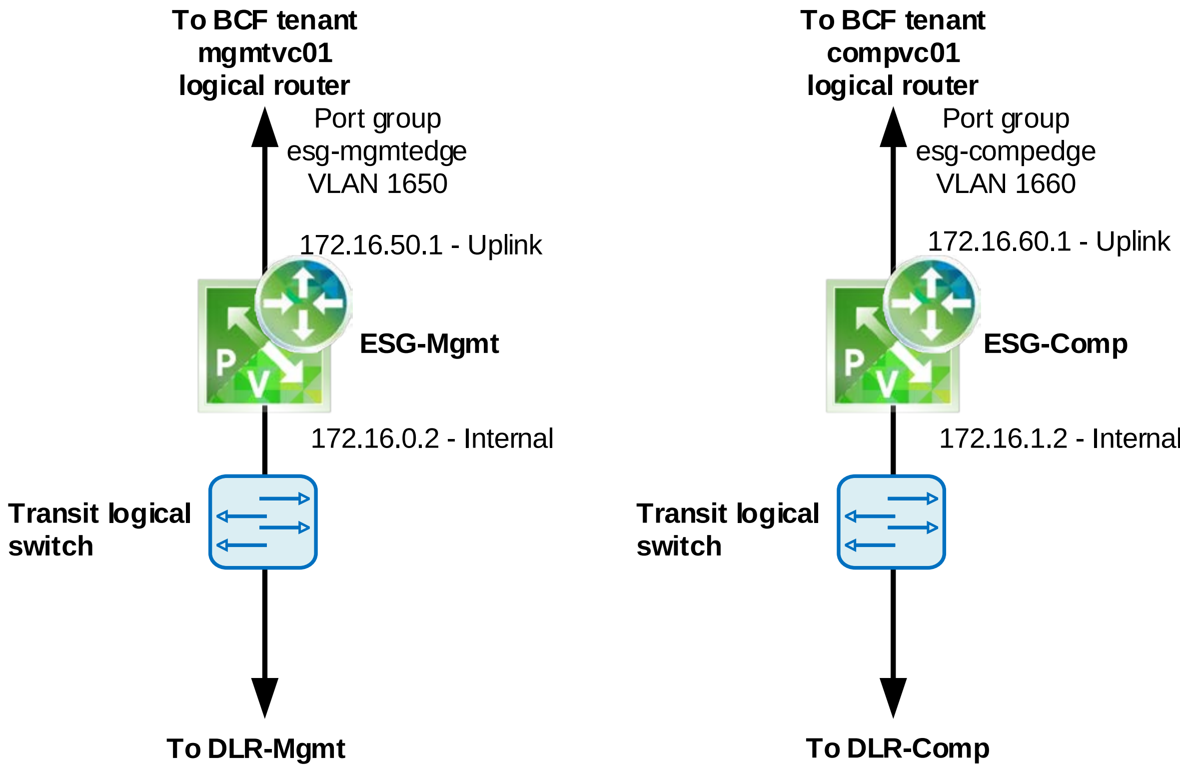


Figure 117 ESG-Mgmt and ESG-Comp

As with DLRs, both ESGs are deployed with HA enabled. This means for each ESG instance, two VMs are deployed: one VM is active and the other is standby.

10.11.1 ESG port group settings

Before deploying ESGs, port groups for ESG connections are created on vDS-Mgmt and vDS-CompEdge. These port groups handle all north-south traffic between the NSX environment and the network core.

ESG port group settings used are shown in Table 37.

Table 37 ESG port group settings

vDS	Port group name	VLAN ID	Teaming and failover settings				
			Load balancing	**Network failure detection**	**Notify switches**	**Failback**	**Active uplinks**
vDS-Mgmt	esg-mgmtedge	1650	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2
vDS-CompEdge	esg-compedge	1660	Route Based on Physical NIC Load	Link status only	Yes	Yes	1,2

10.11.2 ESG deployment settings

ESG deployment settings used are shown in the following tables:

Table 38 ESG configuration settings

NSX Manager	Install Type	ESG Name	Deploy NSX Edge	Enable HA	Data-center	Appl. Size	Cluster	Datastore
100.67.187.180	ESG	ESG-Mgmt	Checked	Checked	Mgmt Data center	Compact	Mgmt	Mgmtvsan Datastore
100.67.187.181	ESG	ESG-Comp	Checked	Checked	Comp Data center	Large	Compute -Edge	CompEdgevsan Datastore

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

The following settings in Table 39 and Table 40 are used to configure ESG interfaces:

Table 39 ESG-Mgmt interface settings

Interface Name	Type	Connected To	IP Address/prefix	MTU
mgmt-esg-uplink	Uplink	Distributed port group: esg-mgmtedge	172.16.50.1/24	9000
mgmt-esg-internal	Internal	Logical switch: Transit	172.16.0.2/24	9000

Table 40 ESG-Comp interface settings

Interface Name	Type	Connected To	IP Address/prefix	MTU
comp-esg-uplink	Uplink	Distributed port group: esg-compedge	172.16.60.1/24	9000
comp-esg-internal	Internal	Logical switch: Transit	172.16.1.2/24	9000

Default gateways are not configured on ESGs at this time. They are configured in the next section.

For HA parameters, default settings are used on the ESGs:

- **vNIC** is set to **any**
- HA management IP addresses are not specified

Note: HA management IP addresses are automatically chosen from the link local address space, 169.250.0.0/16.

Figure 118 and Figure 119 show the ESG interface configuration settings when complete. The pages shown are accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the ESG name. Select **Manage > Settings > Interfaces.**

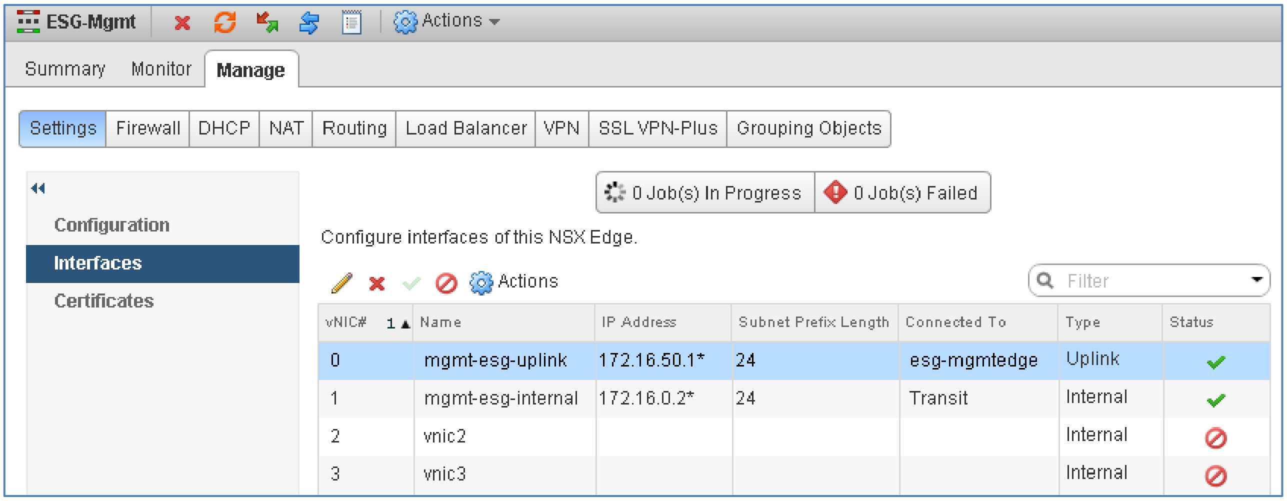


Figure 118 ESG-Mgmt interface configuration settings

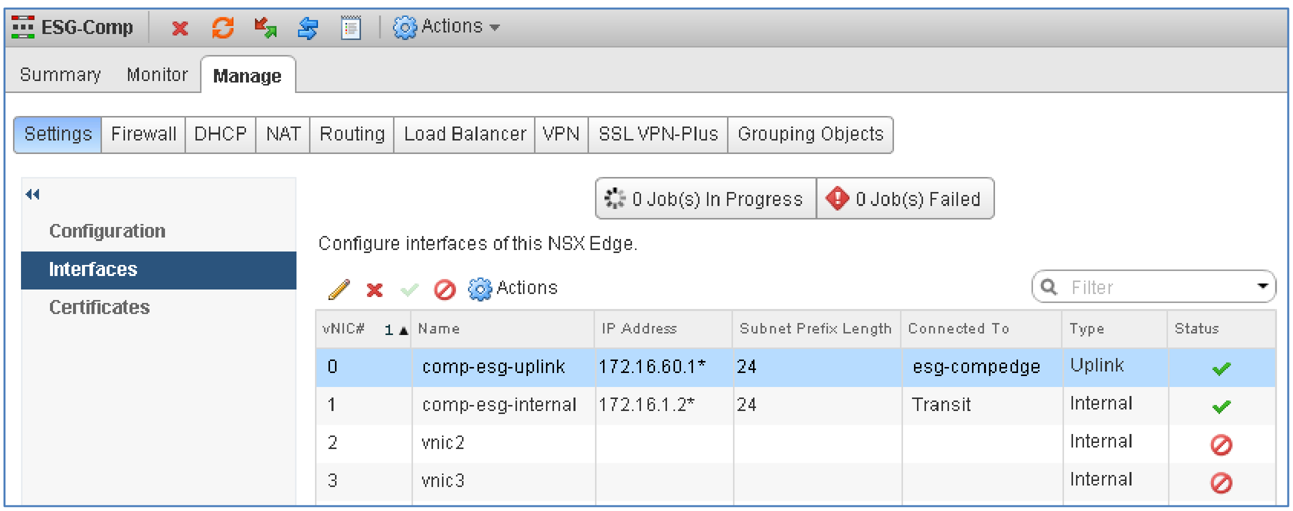
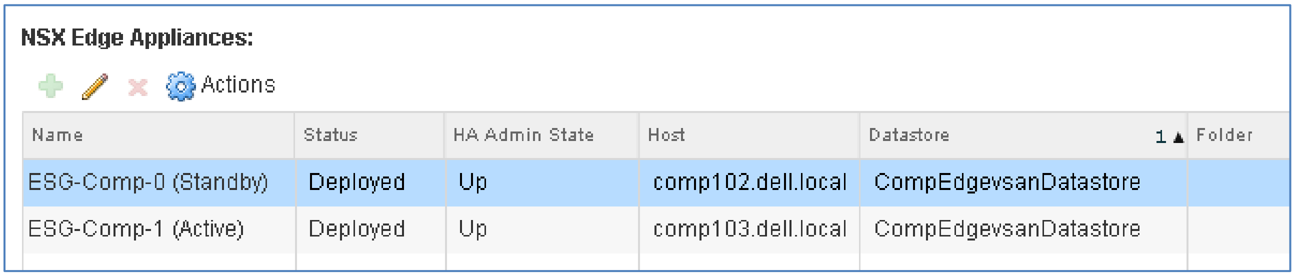


Figure 119 ESG-Comp interface configuration settings

With HA enabled, the bottom of the **Manage > Settings > Configuration** page for ESG-Comp appears as shown. The page for ESG-Mgmt is similar.



NSX Edge Appliances:

Actions

Name	Status	HA Admin State	Host	Datastore 1 ▲	Folder
ESG-Comp-0 (Standby)	Deployed	Up	comp102.dell.local	CompEdgevsanDatastore	
ESG-Comp-1 (Active)	Deployed	Up	comp103.dell.local	CompEdgevsanDatastore	

Figure 120 Active and Standby VMs deployed for ESG-Comp

On the **Hosts and Clusters** tab, redundant VMs for ESG-Comp and ESG-Mgmt are visible in their respective clusters:

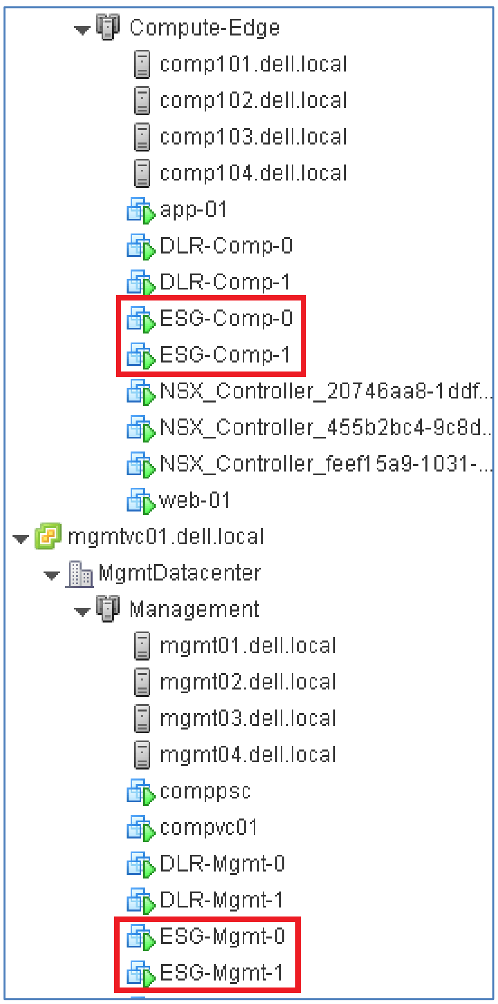


Figure 121 ESGs deployed with HA enabled

10.11.3 ESG global configuration settings

After deployment, global configuration settings are applied to each ESG per the following table:

Table 41 ESG global configuration settings

NSX Manager	ESG Name	ECMP	Default Gateway vNIC	Default Gateway IP	MTU	Admin dis.	Router ID
100.67.187.180	ESG-Mgmt	Enabled	mgmt-esg-uplink	172.16.50.2	9000	1	172.16.50.1
100.67.187.181	ESG-Comp	Enabled	comp-esg-uplink	172.16.60.2	9000	1	172.16.60.1

Note: The default gateway IP is the BCF segment interface address and is configured in Section 13.

The Global Configuration page is accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the ESG name. Select **Manage > Routing > Global Configuration.**

Figure 122 shows the global configuration settings for ESG-Comp when complete. The page for ESG-Mgmt is similar.

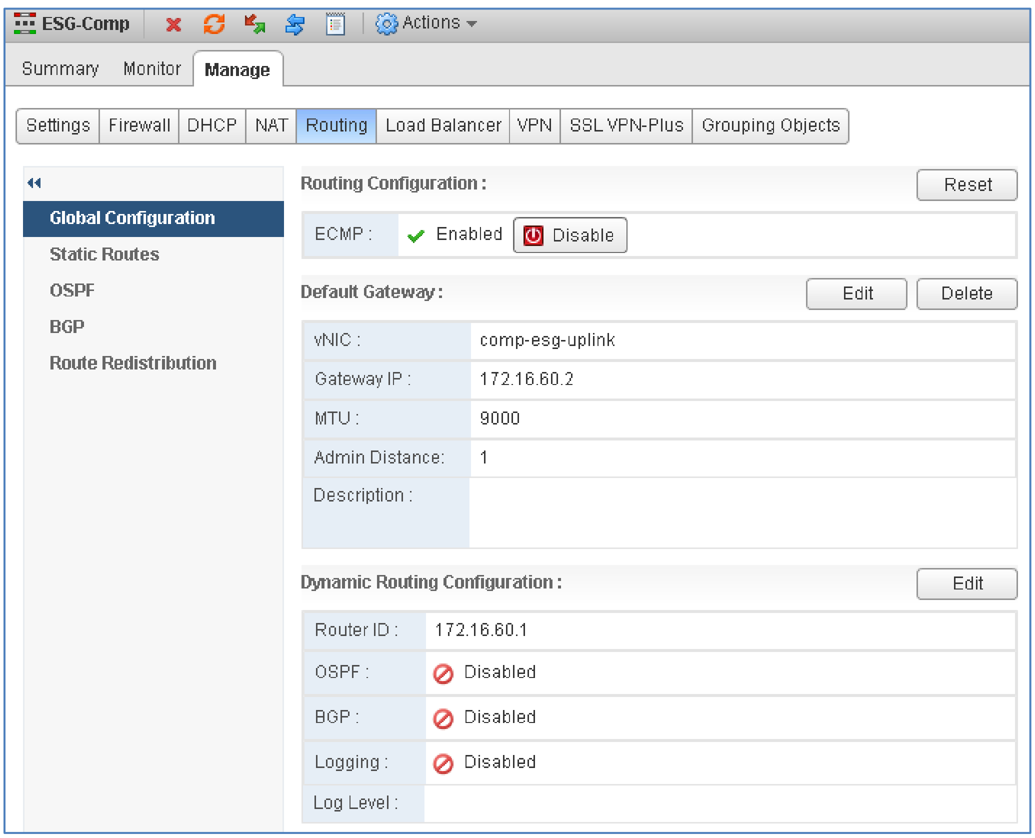


Figure 122 ESG-Comp global configuration settings

11 Configure BCF for VXLAN and verify connectivity

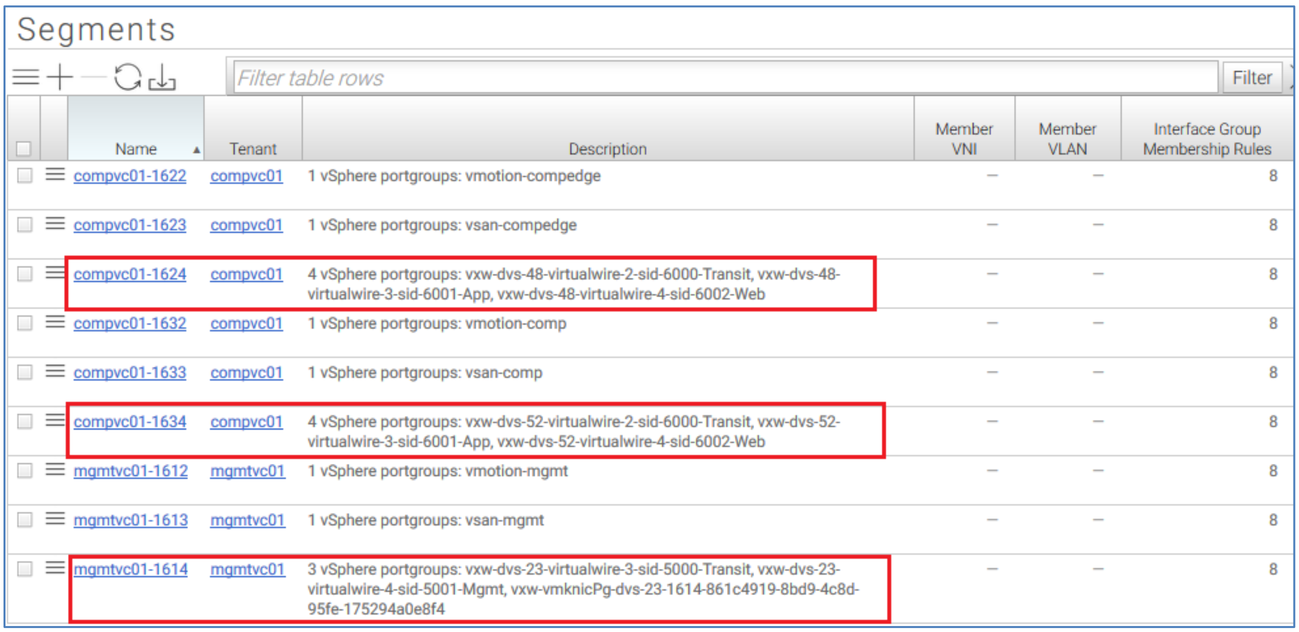
BCF is configured to allow communication between ESXi host VTEPs on the three VXLAN VLANs: 1614, 1624, and 1634. These VLANs were added for VXLAN traffic during VXLAN configuration in Section 10.5.

> **Note**: BCF is not used to manage connections between VMs on NSX virtual networks. This is done in NSX Manager.

11.1 View VXLAN segments

With vCenter integration and full automation enabled in BCF, one segment is automatically created for each VLAN as it is created in vCenter.

To view the current list of segments in the BCF GUI, select **Logical > Segments**.


Segments

Filter table rows Filter

Name	Tenant	Description	Member VNI	Member VLAN	Interface Group Membership Rules
compvc01-1622	compvc01	1 vSphere portgroups: vmotion-compedge	–	–	8
compvc01-1623	compvc01	1 vSphere portgroups: vsan-compedge	–	–	8
compvc01-1624	compvc01	4 vSphere portgroups: vxw-dvs-48-virtualwire-2-sid-6000-Transit, vxw-dvs-48-virtualwire-3-sid-6001-App, vxw-dvs-48-virtualwire-4-sid-6002-Web	–	–	8
compvc01-1632	compvc01	1 vSphere portgroups: vmotion-comp	–	–	8
compvc01-1633	compvc01	1 vSphere portgroups: vsan-comp	–	–	8
compvc01-1634	compvc01	4 vSphere portgroups: vxw-dvs-52-virtualwire-2-sid-6000-Transit, vxw-dvs-52-virtualwire-3-sid-6001-App, vxw-dvs-52-virtualwire-4-sid-6002-Web	–	–	8
mgmtvc01-1612	mgmtvc01	1 vSphere portgroups: vmotion-mgmt	–	–	8
mgmtvc01-1613	mgmtvc01	1 vSphere portgroups: vsan-mgmt	–	–	8
mgmtvc01-1614	mgmtvc01	3 vSphere portgroups: vxw-dvs-23-virtualwire-3-sid-5000-Transit, vxw-dvs-23-virtualwire-4-sid-5001-Mgmt, vxw-vmknicPg-dvs-23-1614-861c4919-8bd9-4c8d-95fe-175294a0e8f4	–	–	8

Figure 123 VXLAN segments automatically added with vCenter integration

The three segments outlined in red verifies BCF successfully imported the VXLAN VLAN information from vCenter.

11.2 Configure VXLAN segment interfaces

Figure 124 shows the BCF logical view with the two vCenter tenants in this deployment, mgmtvc01 and compvc01, and the three VXLAN segments imported from vCenter. The VXLAN segment interfaces on the tenant logical routers are configured in this section.

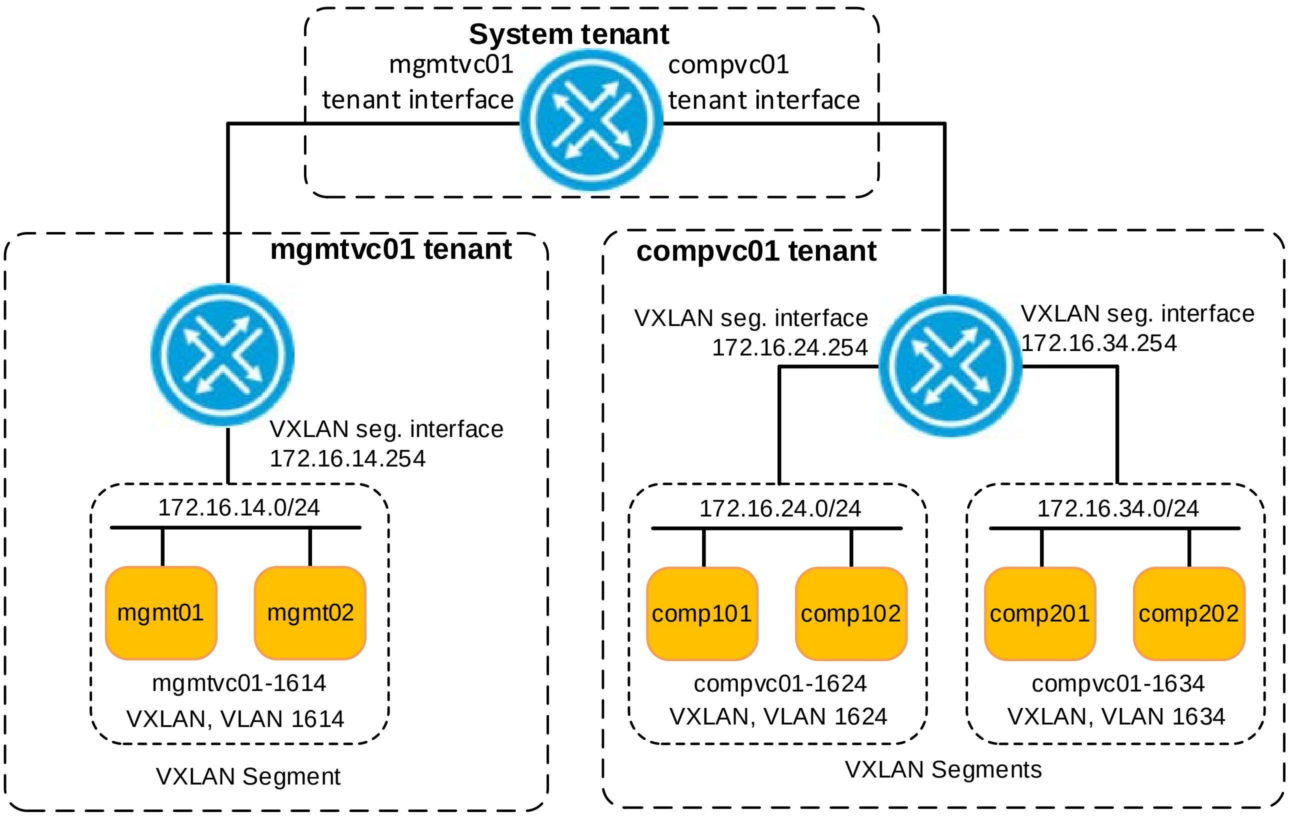


Figure 124 BCF tenants and VXLAN segments

Note: In Figure 124, only two of the four hosts in each cluster are shown, and the vMotion and vSAN segments are not shown for clarity. The vMotion and vSAN segments are shown in Figure 81.

The two tenant logical routers shown in Figure 124 are already connected to the System tenant's logical router as covered in section 8.3. No additional configuration of the tenant interfaces is required.

At this point, VXLAN hosts on the same segments can communicate with each other. For communication between segments, VXLAN segment interfaces are configured uisng the information listed in Table 42.

Table 42 BCF tenant and segment configuration

Tenant	Logical segment name	Function	VLAN ID	Subnet	Segment interface address
mgmtvc01	mgmtvc01-1614	VXLAN	1614	172.16.14.0/24	172.16.14.254
compvc01	compvc01-1624	VXLAN	1624	172.16.24.0/24	172.16.24.254
compvc01	compvc01-1634	VXLAN	1634	172.16.34.0/24	172.16.34.254

To configure segment interfaces, do the following:

1. In the BCF GUI, navigate to **Logical > Tenants.**
2. Select a tenant, **mgmtvc01** in this example, to open its tenant configuration page.
3. In the left pane, scroll down and select **Segment Interfaces**. This adds **Segment Interfaces** to the right pane as shown.

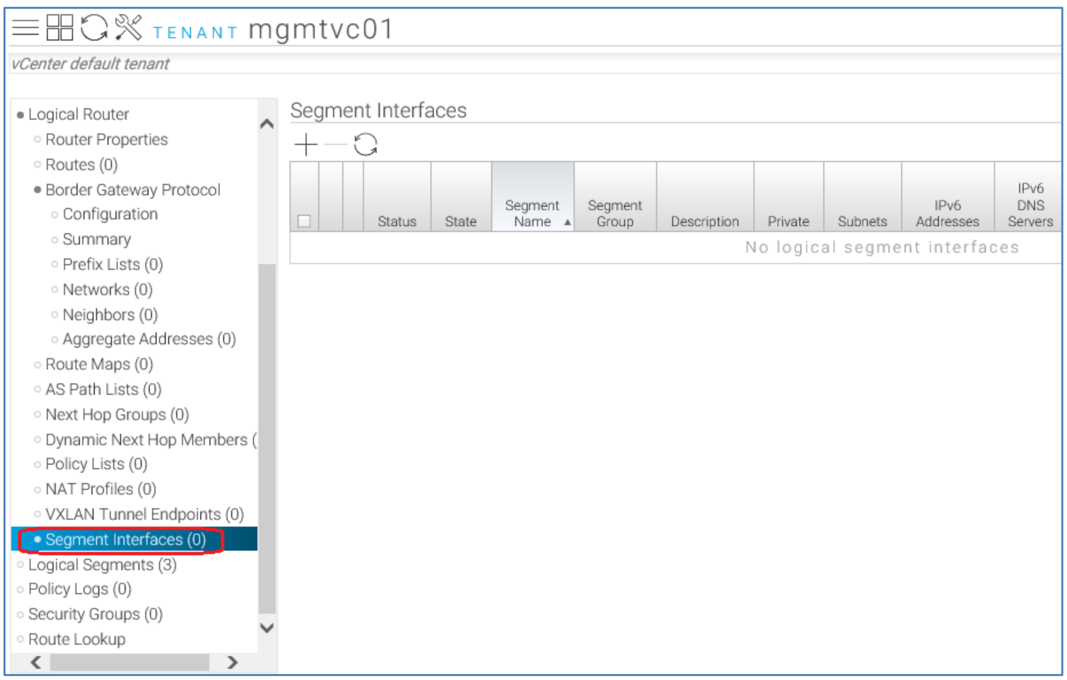


Figure 125 Segment Interfaces selected

4. In the right pane under **Segment Interfaces**, click the + icon. The **Create Logical Segment Interface** dialog box displays.

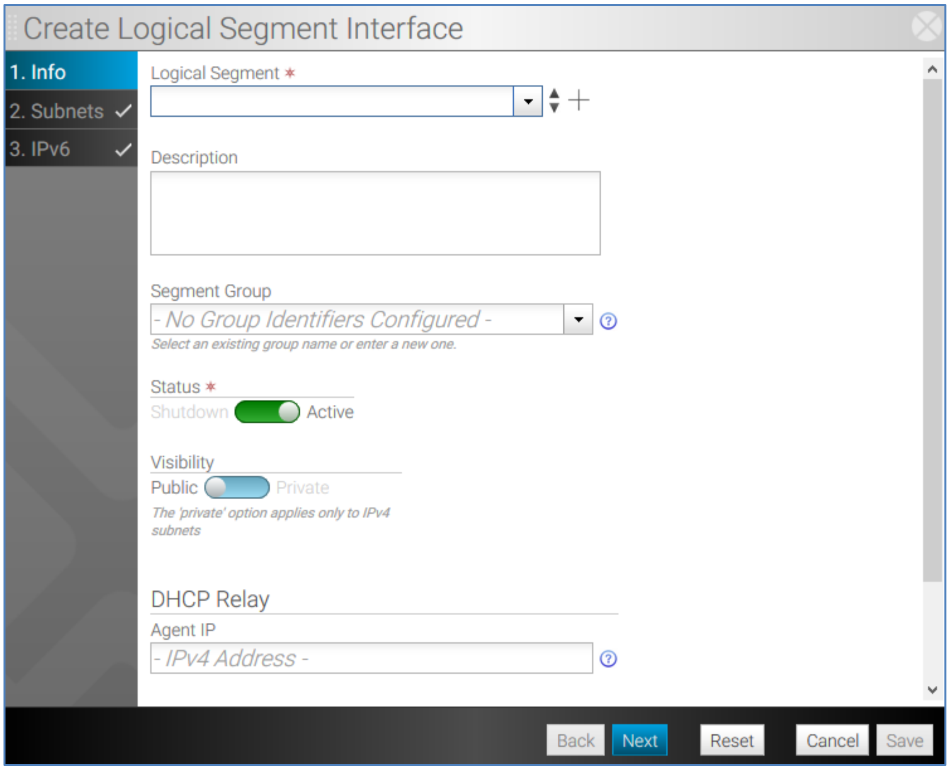


Figure 126 Create segment interface dialog box

5. Under **Logical Segment**, select the name of the VXLAN logical segment from the drop-down menu, **mgmtvc01-1614** in this example. This is for VXLAN traffic in the management cluster. Leave other settings at their defaults and click **Next**.
6. Click the + icon to open the **Create Subnet** dialog box.
7. Provide the segment interface IP address and prefix per Table 42, **172.16.14.254 /24**. The subnet mask in dotted decimal form is automatically completed.

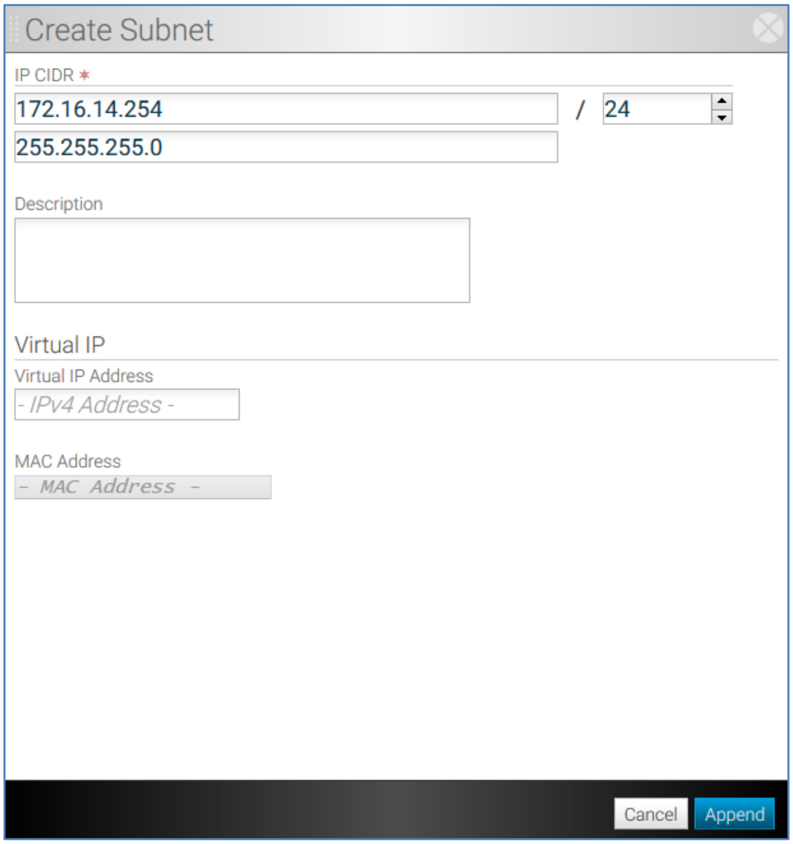
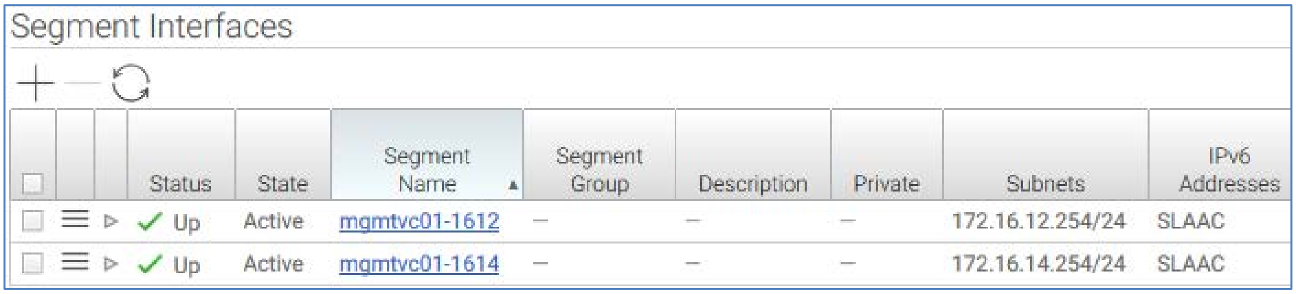


Figure 127 Create subnet dialog box

8. Click **Append > Save.**

The VXLAN segment interface is created in the mgmtvc01 tenant. Repeat steps 1-8 above to create segment interfaces for the two VXLAN segments in the compvc01 tenant using the information listed in Table 42.

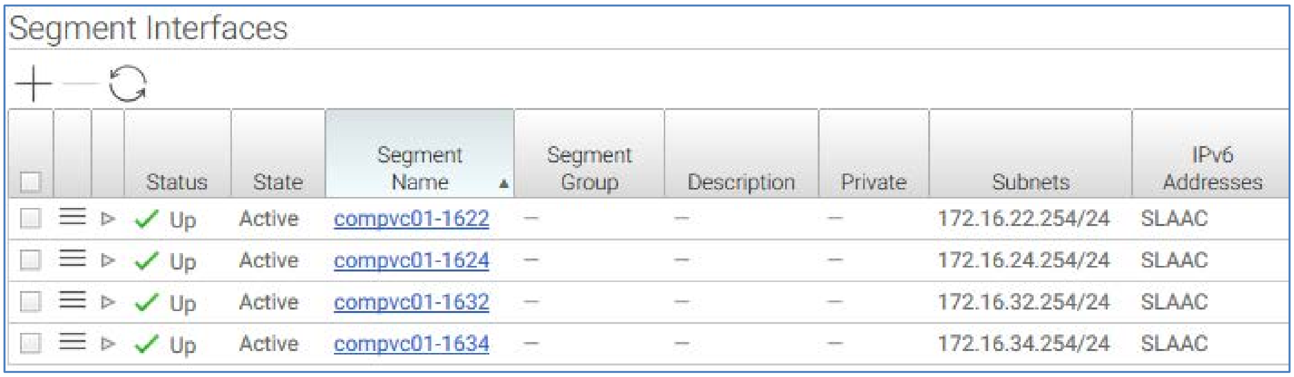
When complete, **Segment Interfaces** for mgmtvc01 and compvc01 appear as shown in Figure 128 and Figure 129. The segment interfaces created for VLANs 1614, 1624, and 1634 are now listed.



Segment Interfaces

			Status	State	Segment Name	Segment Group	Description	Private	Subnets	IPv6 Addresses
			Up	Active	mgmtvc01-1612	—	—	—	172.16.12.254/24	SLAAC
			Up	Active	mgmtvc01-1614	—	—	—	172.16.14.254/24	SLAAC

Figure 128 Mgmtvc01 segment interfaces configured



Segment Interfaces

			Status	State	Segment Name	Segment Group	Description	Private	Subnets	IPv6 Addresses
			Up	Active	compvc01-1622	—	—	—	172.16.22.254/24	SLAAC
			Up	Active	compvc01-1624	—	—	—	172.16.24.254/24	SLAAC
			Up	Active	compvc01-1632	—	—	—	172.16.32.254/24	SLAAC
			Up	Active	compvc01-1634	—	—	—	172.16.34.254/24	SLAAC

Figure 129 Compvc01 segment interfaces configured

11.3 Test VXLAN Connectivity

On the VXLAN networks in this deployment, ESXi hosts configured as VTEPs must be able to reach all other VTEP ESXi hosts in all other clusters. The VXLAN network should be able to support 9000 byte frames.

In the following example, host comp101 in the Compute-Edge cluster successfully pings the VXLAN VMkernel IP address of a host in the Compute cluster.

```
[root@comp101:~] ping 172.16.34.1 -S vxlan -s 8950 -d
PING 172.16.34.1 (172.16.34.1): 8950 data bytes
8958 bytes from 172.16.34.1: icmp_seq=0 ttl=63 time=0.266 ms
8958 bytes from 172.16.34.1: icmp_seq=1 ttl=63 time=0.227 ms
8958 bytes from 172.16.34.1: icmp_seq=2 ttl=63 time=0.338 ms
```

The `–S vxlan` argument instructs the utility to use the VXLAN TCP/IP stack. This argument is required for the command to succeed on VXLAN networks.

The `–d` argument means do not fragment the packet and the `–s 8950` argument sets the ICMP data size in the packet. (This size does not include IP headers, so it is set to slightly under 9000).

Note: If pings fail with jumbo frames, check the MTU size settings on the associated vDS and VMkernel adapters in vCenter.

11.4 Deploy VMs to validate NSX

In this example, six VMs that will communicate using NSX over VXLAN are deployed to ESXi host clusters as shown in Figure 130.

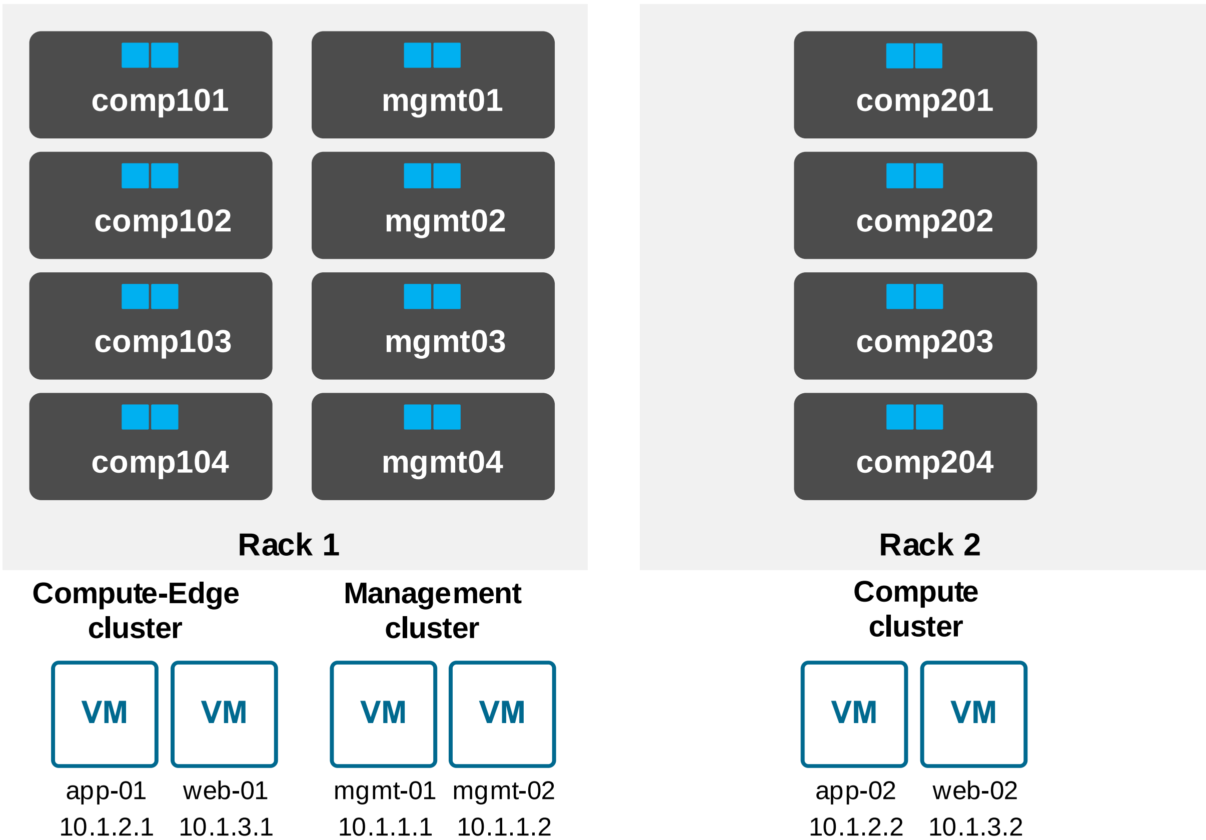


Figure 130 VM locations in physical topology

Each of these VMs runs a Microsoft Windows Server 2016 guest OS and has a single VMXNET3 virtual network adapter (vNIC).

> **Note:** Installation of guest operating systems such as Microsoft Windows Server, Red Hat Linux, etc. and configuration of vNICs is outside the scope of this document. See the *Installing a Guest Operating System* section of the [VMware vSphere Documentation] for information.

11.5 Validate NSX VM connectivity

Connectivity between VMs on the NSX virtual networks is validated in this section. Refer to Figure 130 for the locations of the VMs in the topology.

Within the guest operating system of the source VM, ping the IP address of the destination VMs using Table 43 as a guide. Successful pings validate the segment tested is configured properly.

Note: Guest operating system firewalls will 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.

Table 43 Test examples to validate connectivity

Source	Destination	Validates
mgmt-01 / 10.1.1.1	mgmt-02 / 10.1.1.2	Connectivity within the cluster on same segment.
app-01 / 10.1.2.1	app-02 / 10.1.2.2	Connectivity between clusters on the same segment.
app-01 / 10.1.2.1	web-01 / 10.1.3.1	Connectivity within the cluster on different segments.
app-01 / 10.1.2.1	web-02 / 10.1.3.2	Connectivity between clusters on different segments.

Note: Mgmt VMs in this deployment cannot communicate with App or Web VMs at this point because they are in different tenants.

12 Configure BCF connections to core

There are a number of options for connecting the Big Cloud Fabric to the network core. These are described in the *BCF User Guide*. The connections used in this deployment are covered in this section.

12.1 Physical connections

Note: Configuration of redundant core routers is outside the scope of this document. For this guide, a single S4048-ON switch running DNOS 9.11 is used as the core router to verify the topology. Its configuration is covered in Section 15.1.

The connections to the external network for this deployment example are shown in black in Figure 131. The port numbers used in this deployment are shown.

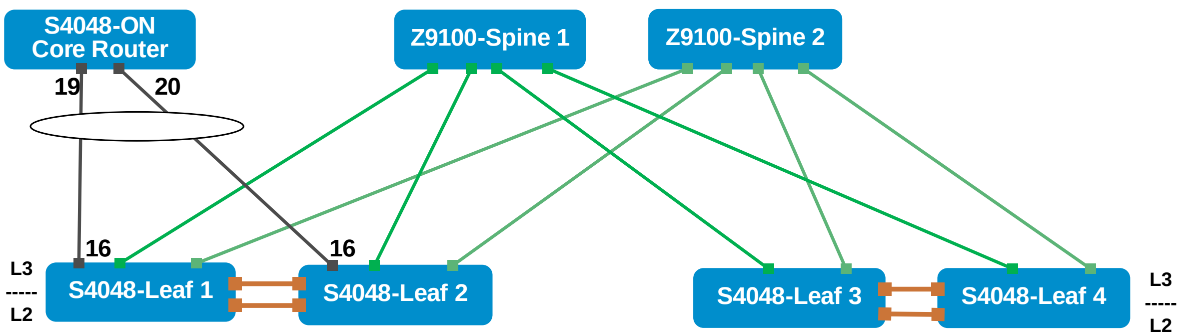


Figure 131 Physical connections to core router

12.2 Logical connections

To enable logical connections to the core router, an External tenant is created in BCF. It is connected to the System tenant and core router as shown. Configuration of the External tenant and its interfaces is covered in the following sections.

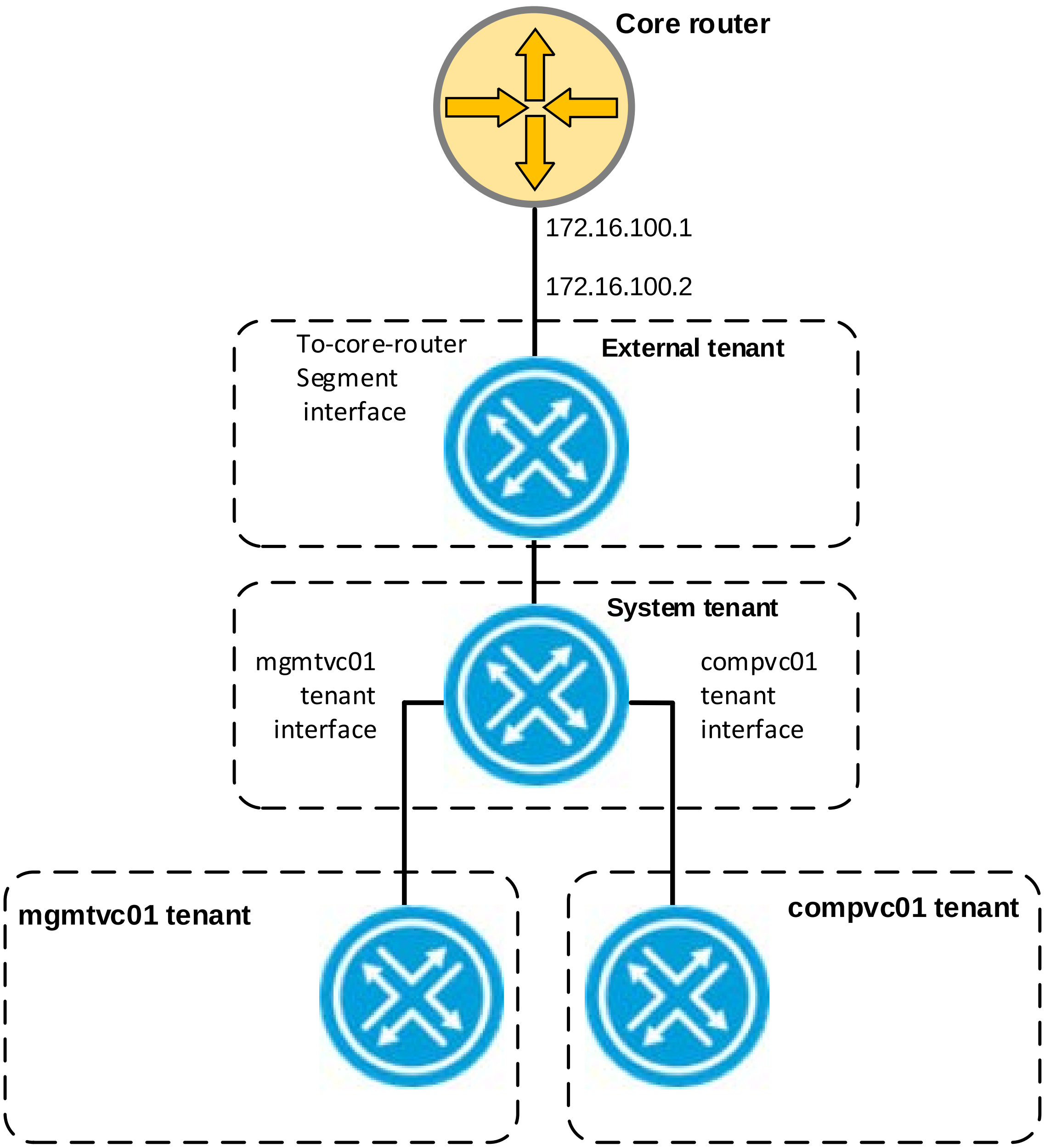


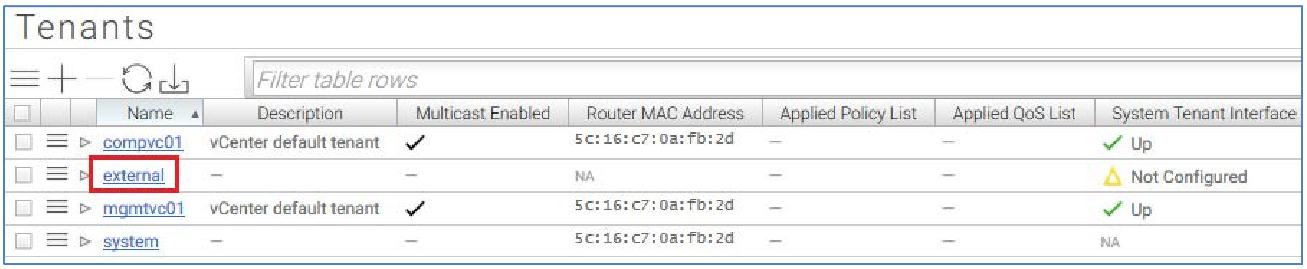
Figure 132 BCF logical connections

12.3 Create the External tenant

To create the External tenant, do the following:

1. In the BCF GUI, navigate to **Logical > Tenants**.
2. Click the + icon to open the **Create Tenant** dialog box.
3. Provide the **Name**, **external,** and leave the **Multicast** slider set to **Disabled.**
4. Click **Save.**

The External tenant is created as shown:


Tenants

Filter table rows

Name	Description	Multicast Enabled	Router MAC Address	Applied Policy List	Applied QoS List	System Tenant Interface
compvc01	vCenter default tenant	✓	5c:16:c7:0a:fb:2d	–	–	✓ Up
external	–	–	NA	–	–	Not Configured
mgmtvc01	vCenter default tenant	✓	5c:16:c7:0a:fb:2d	–	–	✓ Up
system	–	–	5c:16:c7:0a:fb:2d	–	–	NA

Figure 133 External tenant created

12.4 Connect the External tenant to the System tenant

On the External tenant, its System tenant interface is enabled and connected to the System tenant as follows:

1. On the **Logical > Tenants** page, click the ▷ next to the **External tenant** to view the **Logical Router** settings.
2. Under **Logical Router**, next to **System Tenant Interface**, click the icon to open the **Manage Tenant Interfaces** dialog box.

3. Move the **Configured** and **Active** sliders to the right to enable the interfaces as shown:

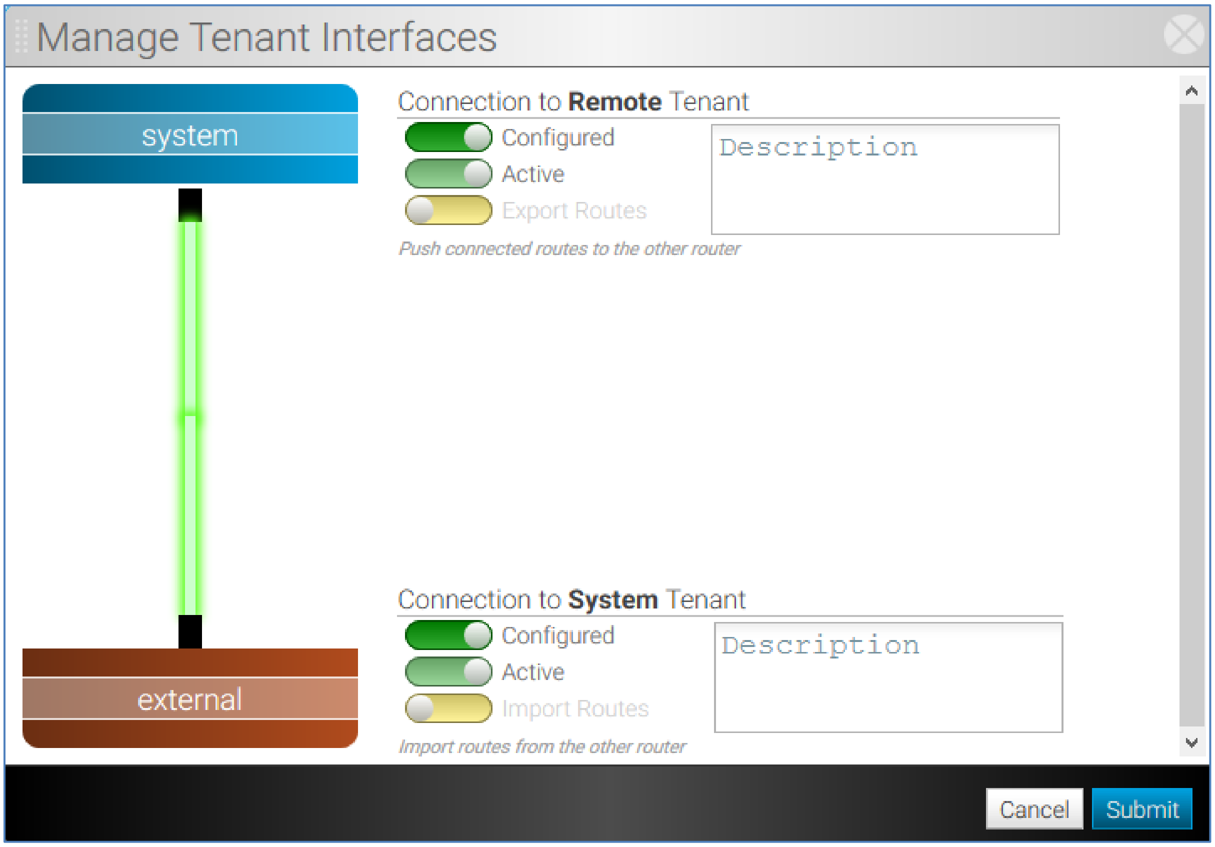


Figure 134 External tenant connected to System tenant

Note: The Export/Import Routes feature applies to directly connected routes. These sliders are left in the off position on the External tenant for this deployment.

4. Click **Submit.**

When complete, the External tenant's **System Tenant Interface** is **Up** as shown:

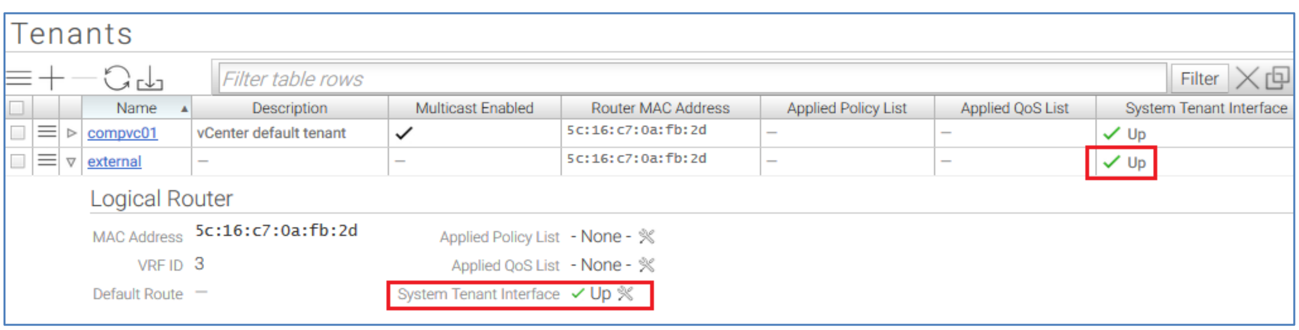


Figure 135 System tenant interface is up

12.5 Connect External tenant to core router

12.5.1 Create an interface group to core router

The **ethernet16** interfaces from Leafs 1 and Leaf 2 are each physically connected to a single S4048-ON which is acting as the core router as shown earlier in Figure 131.

These two connections to the core router are configured in an LACP interface group in BCF as follows:

1. In the BCF GUI, navigate to **Fabric > Interface Groups**.
2. On the **Interface Groups** page, click + to open the **Create Interface Group** dialog box.
3. In the **Create Interface Group** dialog box, the following values are used in this deployment example:
   a. **Name: to-core-router**
   b. **Leaf Group: Rack 1**
   c. **Members: Ethernet16** is selected under Leaf 1 and Leaf 2. These are the ports physically connected to the core router in this example.
   d. **Mode: LACP**

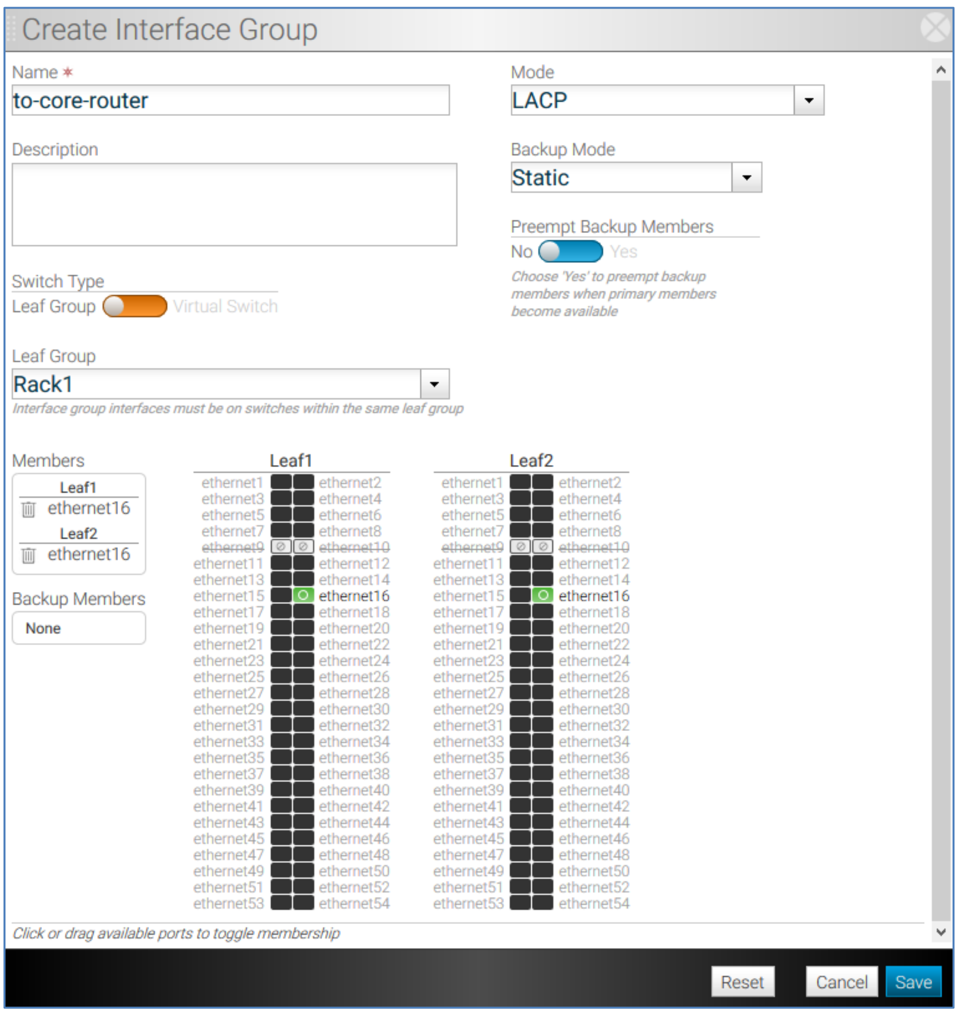


Figure 136 Create interface group dialog box

4. Leave all remaining items at their defaults and click **Save**.

12.5.2 Create a segment to the core router

To create the segment to the core router, do the following:

1. In the BCF GUI, navigate to **Logical > Segments**.
2. Click + to open the **Create Segment** dialog box:

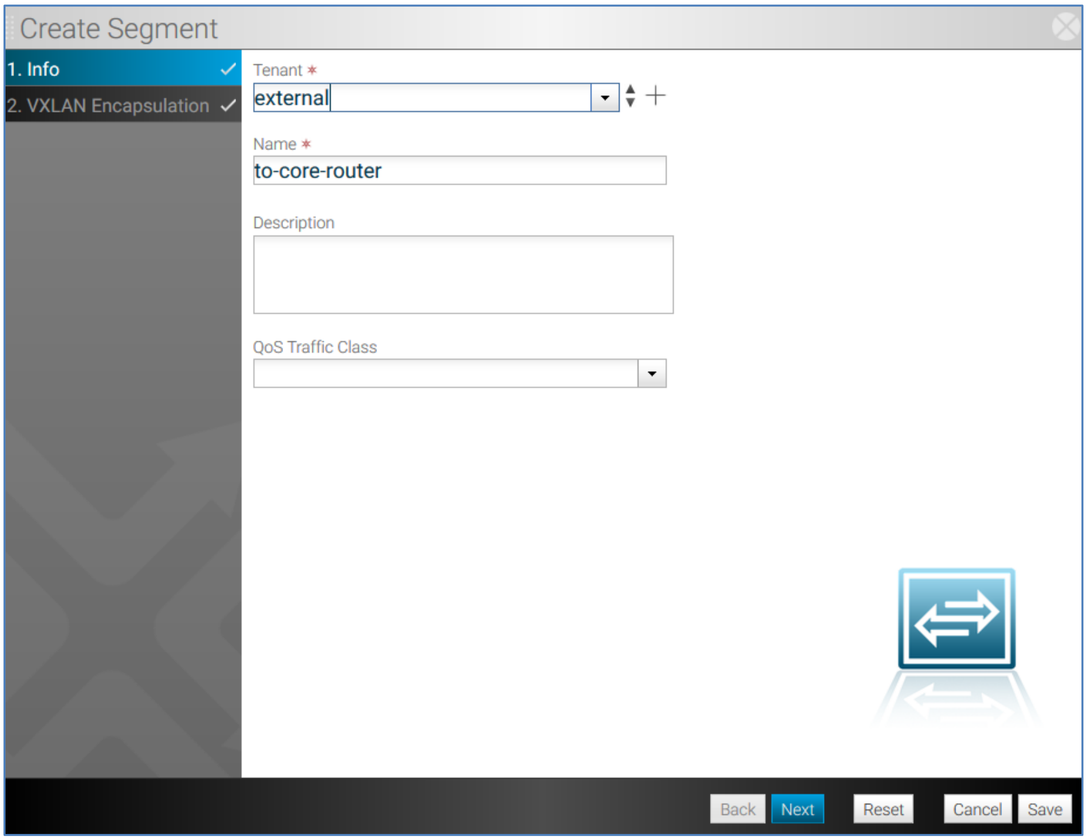


Figure 137 Create segment dialog box

3. The following values are set:
   a. Next to **Tenant**, select **external** from the drop-down menu.
   b. Name: to-core-router
4. Click **Save.**

12.5.3 Configure the External tenant's core router interface

The interface to the core router is configured as follows:

1. In the BCF GUI, go to **Logical > Tenants.**
2. Select **external** to open the **External tenant** configuration page.
3. In the left pane under **Logical Router**, select **Segment Interfaces**.
4. In the right pane under **Segment Interfaces**, click the + icon. The **Create Logical Segment Interface** dialog box displays:

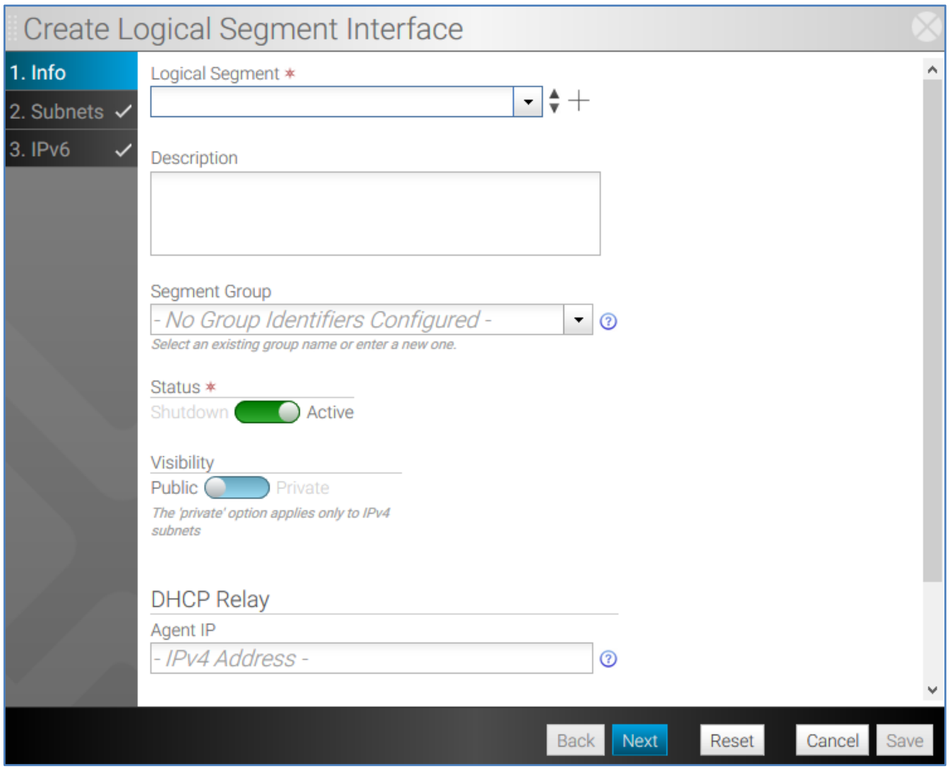


Figure 138 Create segment interface dialog box

5. Under **Logical Segment**, select the **to-core-router** segment from the drop-down menu. Leave other settings at their defaults and click **Next**.
6. On the **Subnets** page, click the + icon to open the **Create Subnet** dialog box.

7. Provide the segment interface IP address and prefix, **172.16.100.2 /24**, as shown earlier in Figure 132. The subnet mask in dotted decimal form is automatically completed:

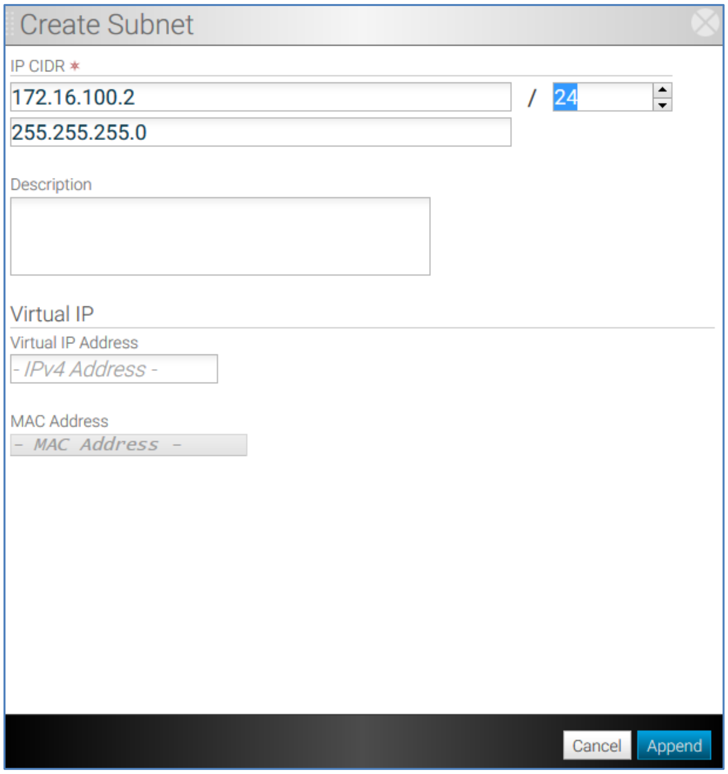


Figure 139 Create subnet dialog box

8. Click **Append > Save.**

12.5.4 Add the interface group to the core router segment

1. In the BCF GUI, go to **Logical > Segments.**
2. Select the segment named **to-core-router**.
3. On the segment to-core-router page, under **Interface Group Membership**, click +.
4. From the drop down menu, select the **to-core-router** interface group.
5. Leave the remaining values at their defaults and click **Save**.

When complete, the **Interface Group Membership** section of the page appears as shown. The **State** may be **Down** until the core router is configured. Core router configuration is covered in Section 15.

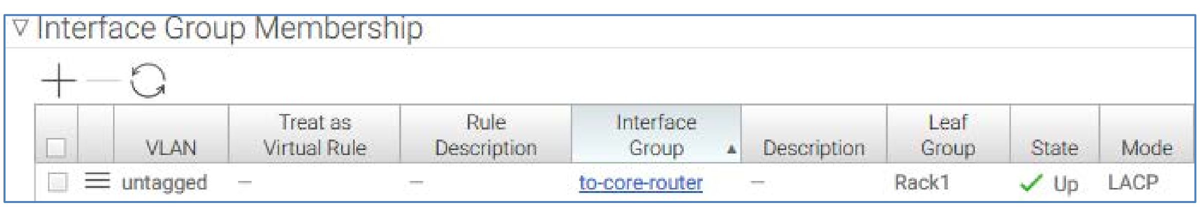


Figure 140 To-core-router interface group added to segment

13 Connect BCF logical routers to ESGs

In this section, a segment interface connecting to the ESG is created on each of the tenant logical routers. This enables VMs on NSX networks to communicate with devices on external networks and between tenants.

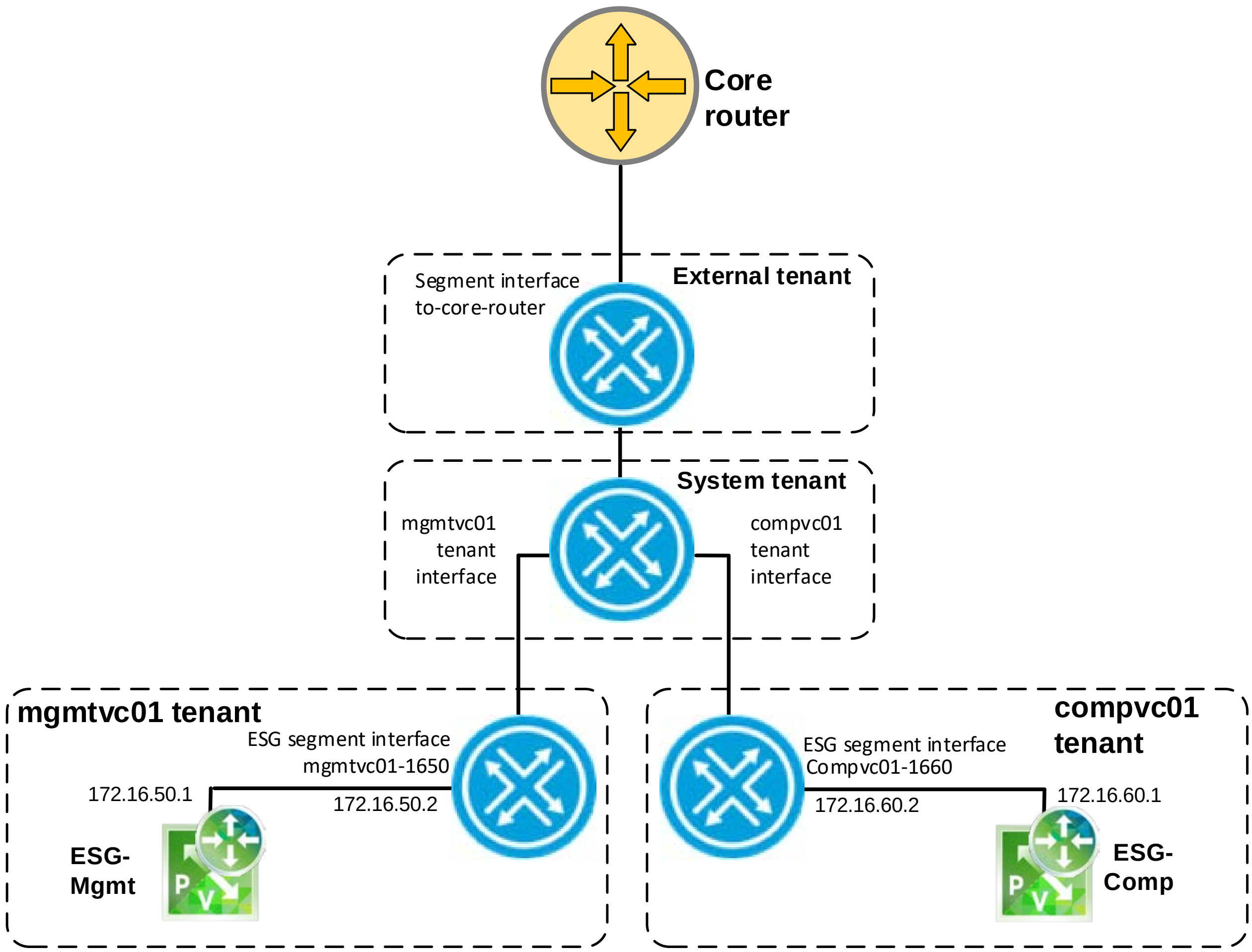


Figure 141 Tenant logical router connections to ESG-Mgmt and ESG-Comp

ESG segment interfaces are configured using the information listed in Table 44.

Table 44 Segment interface settings

Tenant	ESG segment name	Segment IP address
mgmtvc01	mgmtvc01-1650	172.16.50.2/24
compvc01	compvc01-1660	172.16.60.2/24

To configure segment interfaces, do the following:

1. From the BCF GUI, go to **Logical > Tenants.**
2. Select the first tenant from Table 44 to open the tenant configuration page.
3. In the left pane under **Logical Router,** scroll down and select **Segment Interfaces**.
4. In the right pane under **Segment Interfaces**, click the + icon. The **Create Logical Segment Interface** dialog box displays:

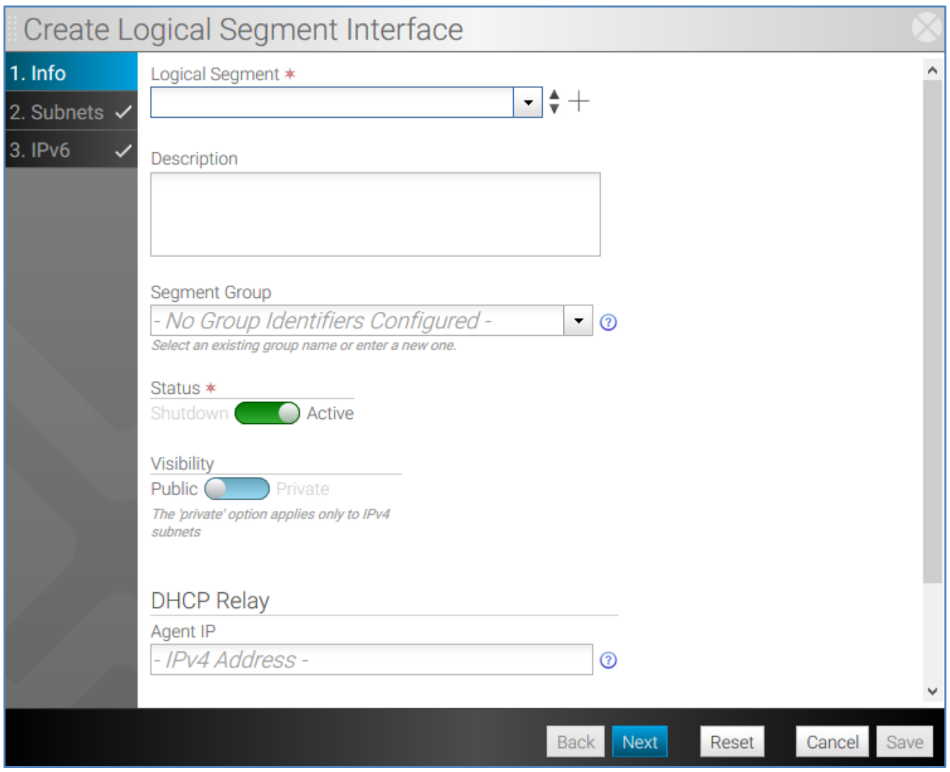


Figure 142 Create segment interface dialog box

5. Under **Logical Segment**, select the name of the ESG segment from the drop-down menu per Table 44. Leave other settings at their defaults and click **Next**.
6. Click the + icon to open the **Create Subnet** dialog box.
7. Provide the segment interface IP address and prefix from Table 44. The subnet mask in dotted decimal form is automatically completed as shown:

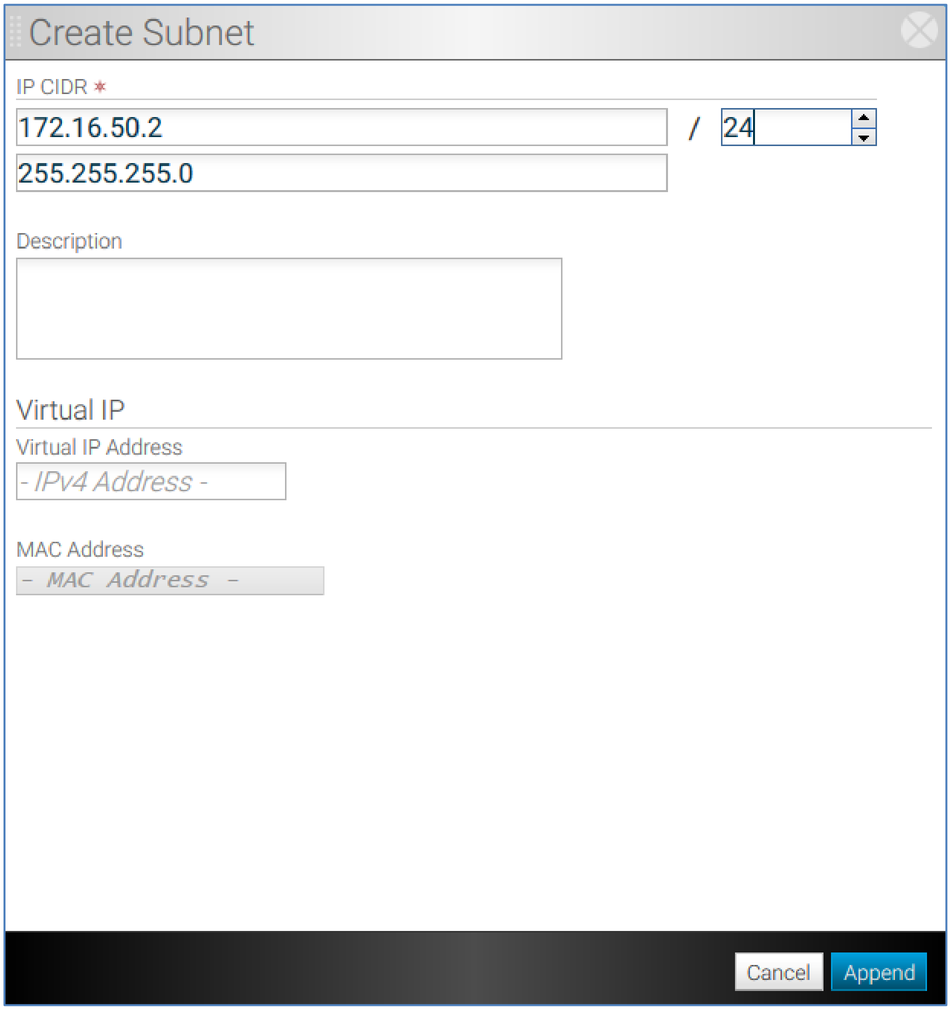


Figure 143 Create subnet dialog box

8. Click **Append > Save.**

Repeat the steps above for the remaining tenant using the data in Table 44.

To verify connectivity to this point, open the active ESG consoles and ensure:

- ESG-Mgmt is able to ping the mgmtvc01-1650 segment interface, 172.16.50.2.
- ESG-Comp is able to ping the compvc01-1660 segment interface, 172.16.60.2.

14 Configure routing on the virtual networks

In this deployment, BGP is used as the dynamic routing protocol to advertise routes to the NSX VMs beyond the DLRs. BGP is supported by BCF and is the routing protocol specified in VVD.

The end-to-end topology with all logical BCF and NSX routers and switches in this deployment is shown in Figure 144. BGP Autonomous System (AS) numbers and protocol IP addresses used are highlighted in yellow.

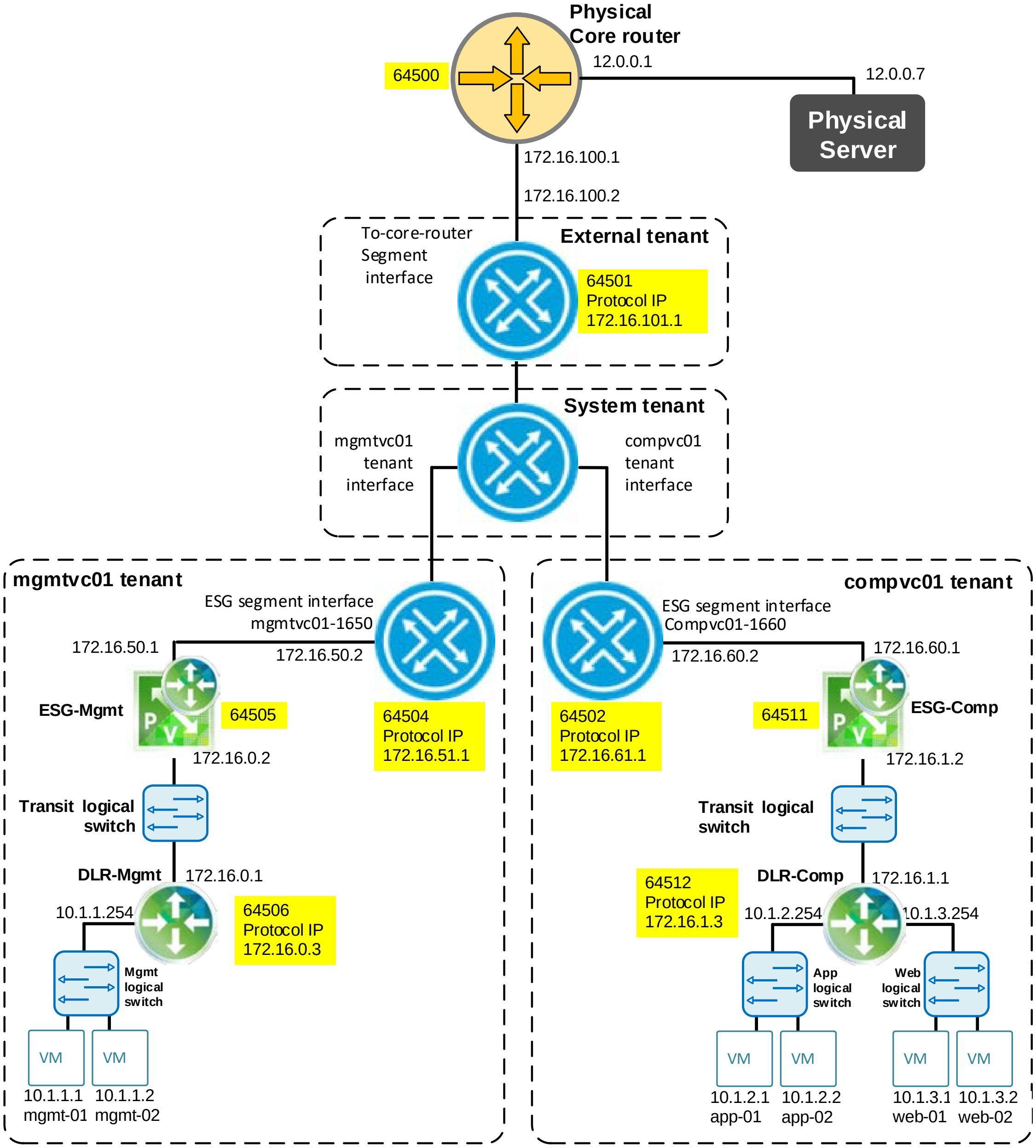


Figure 144 End-to-end logical router topology

14.1 Configure static routes on System and External tenants

Static routes to NSX VMs are configured on the System and External tenant logical routers.

Note: The routing examples in this section enable routing across the System and External tenants. This enables VMs to access VMs in other tenants and to access systems on external networks. Traffic is restricted as needed using the ESG and DLR firewalls.

14.1.1 System tenant static routes

On the System tenant, set its default route to the External tenant as follows:

1. In the BCF GUI, navigate to **Logical > Tenants**.
2. Click the ▷ next to **system** to view the System tenant's **Routes** table.
3. Under **Routes**, click the + icon to open the **Create Route** dialog box and make the following settings:
   a. **Destination Subnet: 0.0.0.0/0**
   b. **Next Hop: Tenant**
   c. **Tenant: external**
   d. Click **Save.**

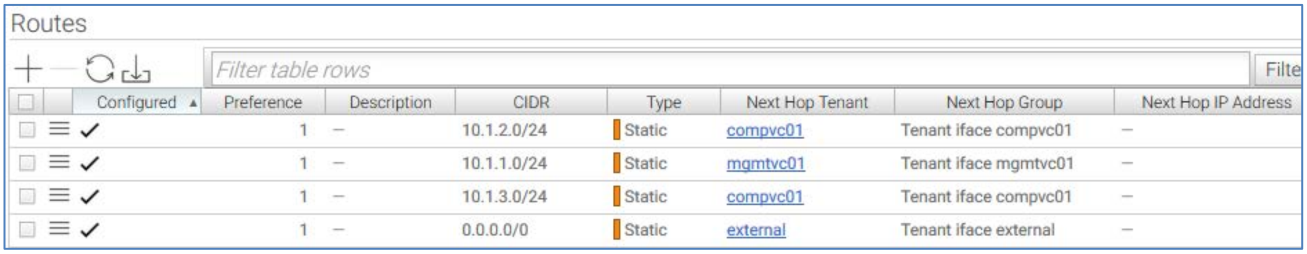
Create a static route on the System tenant pointing to the Mgmt VMs as follows:

4. Under **Routes**, click the + icon to open the **Create Route** dialog box and make the following settings:
   a. **Destination Subnet: 10.1.1.0/24**
   b. **Next Hop: Tenant**
   c. **Tenant: mgmtvc01**
   d. Click **Save.**

Create static routes on the System tenant pointing to the App and Web VMs as follows:

5. Under **Routes**, click the + icon to open the **Create Route** dialog box and make the following settings:
   a. **Destination Subnet: 10.1.2.0/24** (App VMs)
   b. **Next Hop: Tenant**
   c. **Tenant: compvc01**
   d. Click **Save.**
6. Repeat step 5 for the Web VMs on **10.1.3.0/24.**

When complete, static routes on the System tenant appear as shown:



Configured	Preference	Description	CIDR	Type	Next Hop Tenant	Next Hop Group	Next Hop IP Address
✓	1	–	10.1.2.0/24	Static	compvc01	Tenant iface compvc01	–
✓	1	–	10.1.1.0/24	Static	mgmtvc01	Tenant iface mgmtvc01	–
✓	1	–	10.1.3.0/24	Static	compvc01	Tenant iface compvc01	–
✓	1	–	0.0.0.0/0	Static	external	Tenant iface external	–

Figure 145 Static routes configured on System tenant

14.1.2 External tenant

On the External tenant, set its default route to the core router as follows:

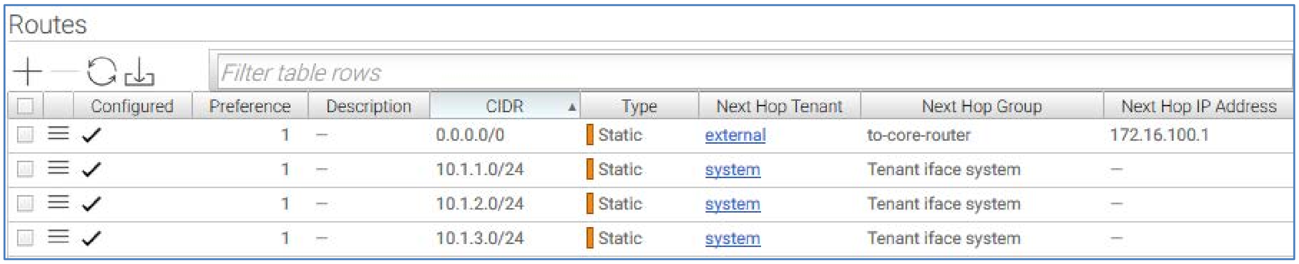
1. In the BCF GUI, navigate to **Logical > Tenants**.
2. Click the ▷ next to **external** to view the External tenant's **Routes** table.
3. Under **Routes**, click the + icon to open the **Create Route** dialog box and make the following settings:
   - **a. Destination Subnet: 0.0.0.0/0**
   - **b.** Under **Next Hop**, select **Next Hop Group.**
   - **c.** Next to **Next Hop Group**, click the + icon to open the **Create Next Hop Group** dialog box and make the following settings:
     - **i. Name: to-core-router**
     - **ii. IP Address type:** leave the slider set to **IPv4**
     - iii. Click the + icon to add the IPv4 address of the core router, **172.16.100.1.**
     - iv. Click **Append.**
   - **d.** Click **Submit > Save.**

Create static routes on the External tenant pointing to the Mgmt, App, and Web VMs as follows:

4. Under **Routes**, click the + icon to open the **Create Route** dialog box and make the following settings:
   - **a. Destination Subnet: 10.1.1.0/24** (Mgmt VMs)
   - **b. Next Hop: System Tenant**
   - **c.** Click **Save.**

5. Repeat step 4 for the App VMs on **10.1.2.0/24**, and the Web VMs on **10.1.3.0/24.**

When complete, static routes on the External tenant appear as shown:


Routes

Filter table rows

Configured	Preference	Description	CIDR	Type	Next Hop Tenant	Next Hop Group	Next Hop IP Address
✓	1	–	0.0.0.0/0	Static	external	to-core-router	172.16.100.1
✓	1	–	10.1.1.0/24	Static	system	Tenant iface system	–
✓	1	–	10.1.2.0/24	Static	system	Tenant iface system	–
✓	1	–	10.1.3.0/24	Static	system	Tenant iface system	–

Figure 146 Static routes configured on External tenant

14.2 Configure BGP

In this section, BGP is configured on the BCF and NSX logical routers.

Note: Before proceeding, ensure the BCF Controller in-band connections are made as shown in Section 4.5. The in-band connections are required for BGP to function on BCF.

14.2.1 Configure BGP on BCF tenants

In this section, BGP is configured on the BCF tenant logical routers for adjacency with their respective routers. This enables the tenant routers, compvc01 and mgmtvc01, to dynamically learn routes to VMs from the ESGs. It also enables the core router to dynamically learn routes to VMs from the External tenant.

The settings shown in Table 45 are used to configure BGP on the mgmtvc01, compvc01, and External tenant routers.

Table 45 BGP configuration settings

Tenant	Local AS	Protocol IP address	Neighbor Name	Neighbor IP address	Remote AS
mgmtvc01	64504	172.16.51.1	ESG-Mgmt	172.16.50.1	64505
compvc01	64502	172.16.61.1	ESG-Comp	172.16.60.1	64511
external	64501	172.16.101.1	Core-router	172.16.100.1	64500

Note: In BCF, the protocol IP address must be on a different network than that used for existing router interfaces. Otherwise, an error message is displayed. The protocol address is also used as the router ID when a router IP address is not configured.

1. In the BCF GUI, navigate to **Logical > Tenants** and click the name of the first tenant listed in Table 45 to open the **Tenant** page.
2. In the left pane under **Logical Router**, ensure that **Routes** is selected**.** Under **Border Gateway Protocol**, ensure that **Configuration** and **Neighbors** are selected as shown:

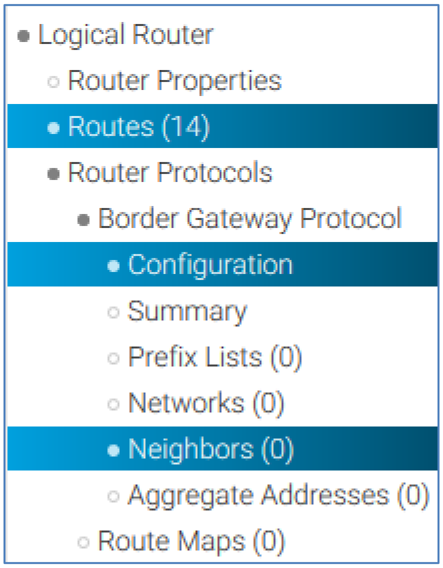


Figure 147 Items selected for BGP configuration

3. In the right pane under **BGP Configuration**, click the  icon to open the **BGP Configuration** dialog box.
   a. Under **Local Autonomous System ID**, enter the AS number from Table 45.
   b. Enter the **Protocol IP Address** from Table 45 and click **Next.**
   c. On the **Options** tab, set **Max Parallel Routes Installed Per Route** to 2.

d. Enable the features outlined in red by moving the sliders as shown:

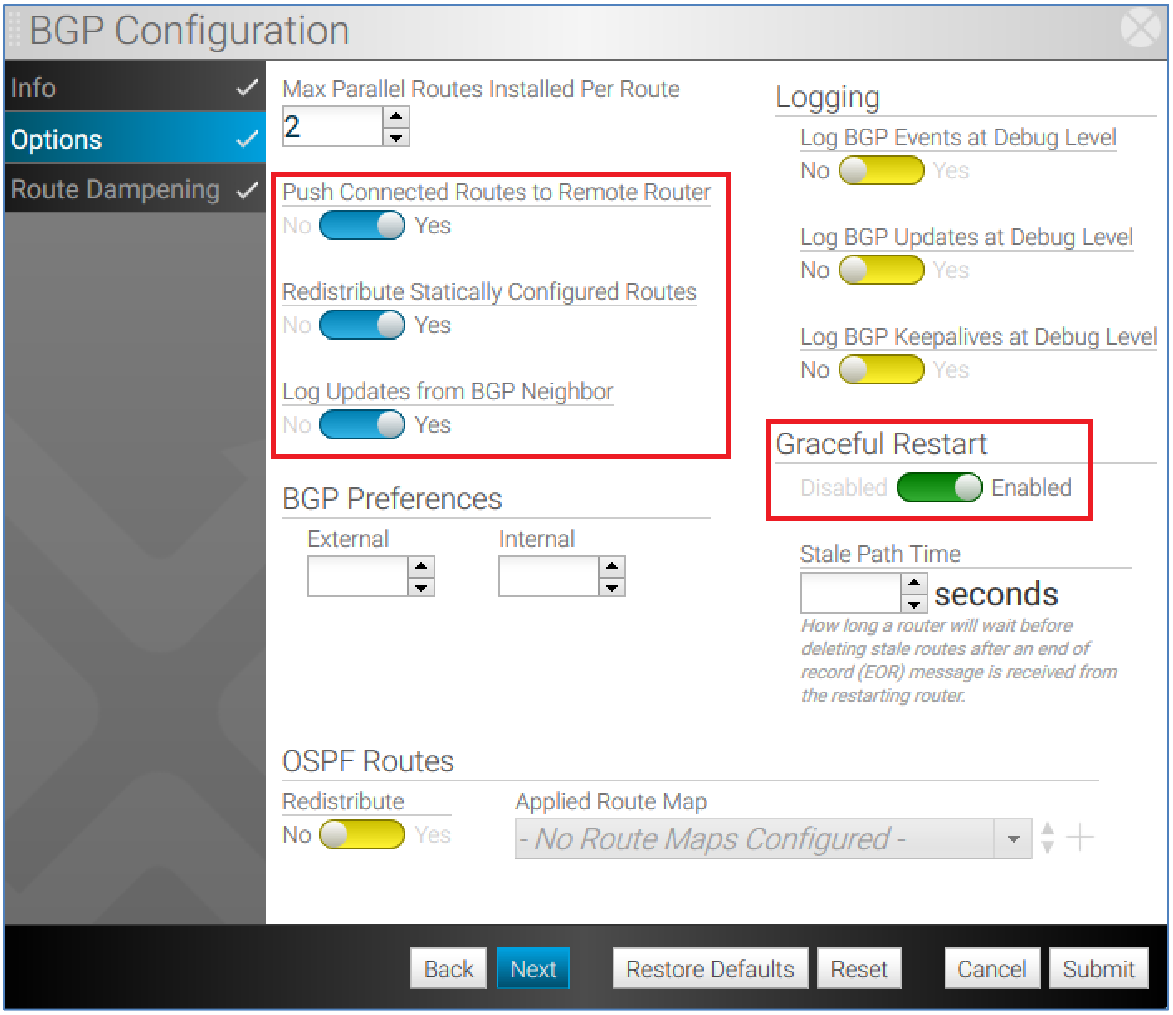


Figure 148 BGP configuration dialog, options tab settings

e. Leave the remaining items at defaults and click **Submit.**

4. On the same tenant page, scroll down to **BGP Neighbors** and click the + button to open the **Create BGP Neighbor** dialog box.
   a. Enter the neighbor **Name, IP Address, and Remote AS ID** number listed in Table 45.
   b. Leave the **Status** slider set to **Up**.

When complete, the dialog box appears as shown:

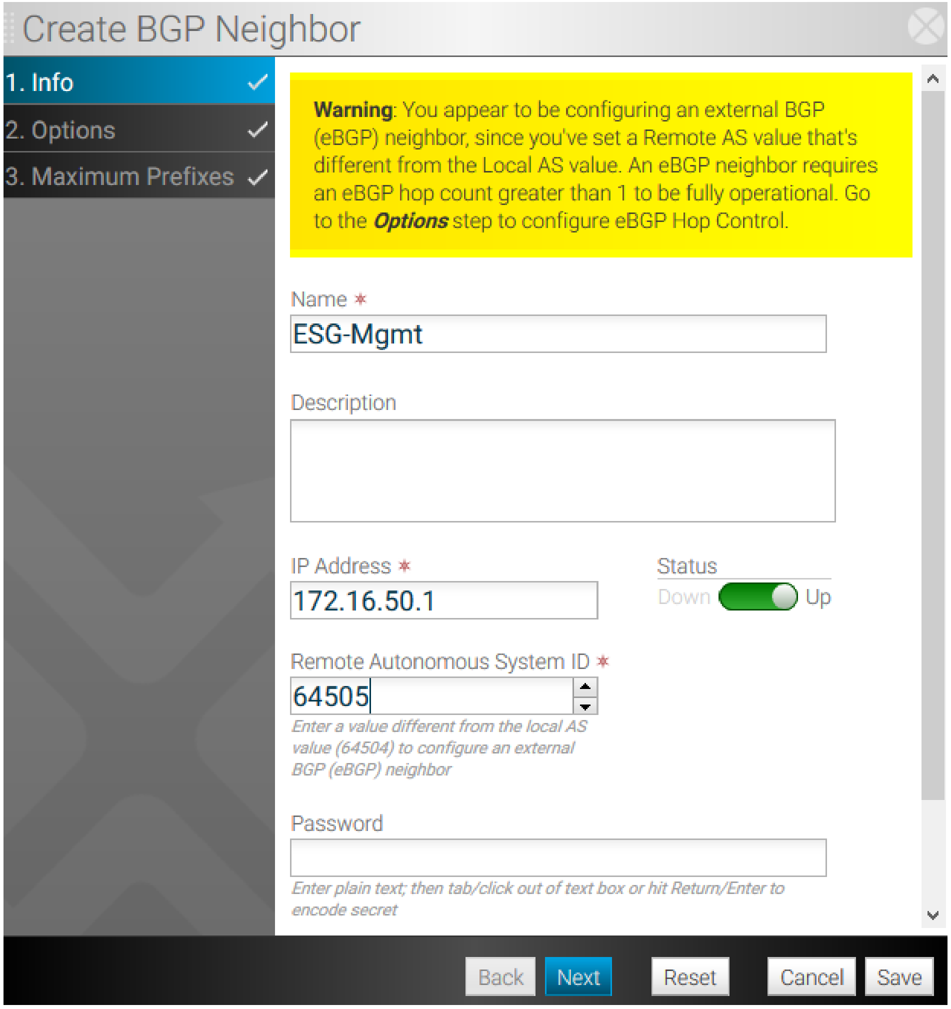


Figure 149 Create BGP neighbor settings – mgmtvc01 router

5. Click **Next** to go to the **Options** tab and complete the settings as follows:
   a. Move the **Store Inbound Routing Table Updates** slider to **Yes**.
   b. Under **eBGP Hop Control**, select **eBGP Multihop** and set the **Hop Count** to **2**.

Note: For the **Hop Count** setting, one hop **i**s added to the actual number of hops to the BGP neighbor.

When complete, the dialog box appears as shown:

Figure 150 BGP neighbor settings

6. Leave the remaining settings at their defaults and click **Save**.

Repeat steps 1-6 above for the remaining tenants in Table 45.

14.2.2 Configure BGP on DLRs

BGP configuration settings for the two NSX DLRs are shown in the following tables:

Table 46 DLR BGP configuration settings

NSX Manager	DLR	BGP	BGP Graceful Restart	Local AS
100.67.187.180	DLR-Mgmt	Enabled	Enabled	64506
100.67.187.181	DLR-Comp	Enabled	Enabled	64512

Table 47 DLR BGP neighbor settings

NSX Manager	DLR	IP Address	Forwarding Address	Protocol Address	Remote AS
100.67.187.180	DLR-Mgmt	172.16.0.2	172.16.0.1	172.16.0.3	64505
100.67.187.181	DLR-Comp	172.16.1.2	172.16.1.1	172.16.1.3	64511

- The **IP Address** is the address of the ESG that faces the DLR.
- The **Forwarding Address** is the DLR's interface address that faces the ESG.
- The **Protocol Address** is an address used by BGP.

Note: On the DLR, the **Protocol Address** must be on the same subnet as the forwarding address. This differs from BCF.

- Settings not shown in the tables remain at their default values

The BGP configuration page is accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the DLR name. Select **Manage > Routing > BGP.**

The figure below shows the BGP configuration settings for DLR-Comp when complete. DLR-Mgmt is similar.

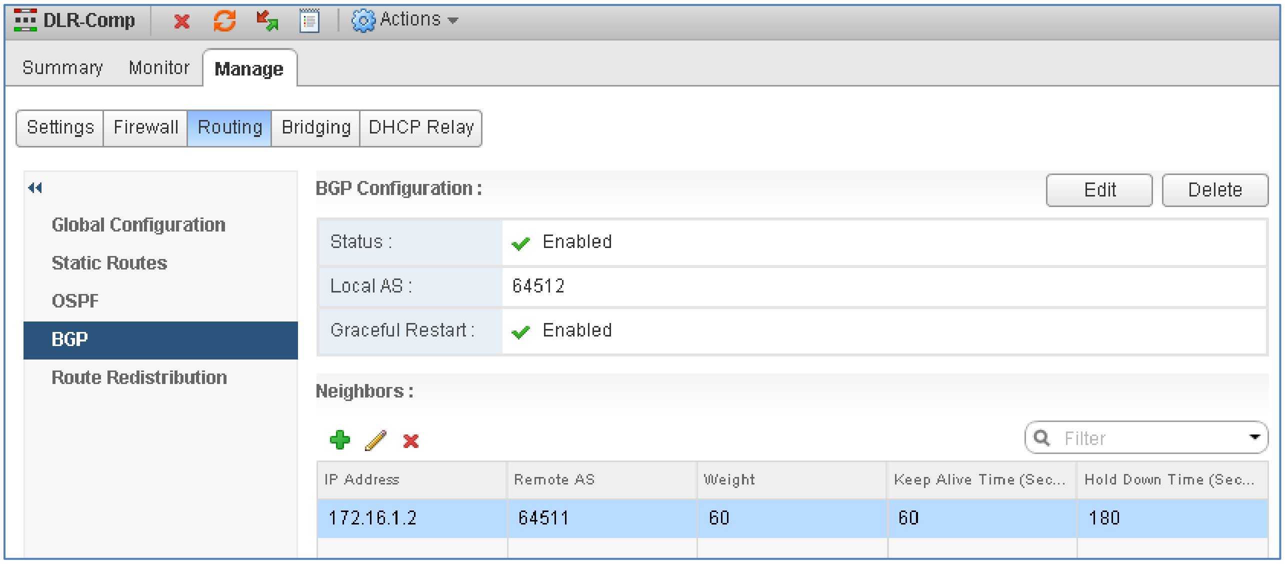


Figure 151 DLR-Comp BGP settings

Route redistribution is configured on both DLRs using the following settings:

Table 48 DLR route redistribution settings

Enable redistribution for	Learner protocol	Allow learning from	Prefix	Action
BGP	BGP	Connected	Any	Permit

The **Route Redistribution** page is accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the DLR name. Select **Manage > Routing > Route Redistribution.**

The figure below shows the Route Redistribution settings for DLR-Comp when complete. Settings are identical on DLR-Mgmt.

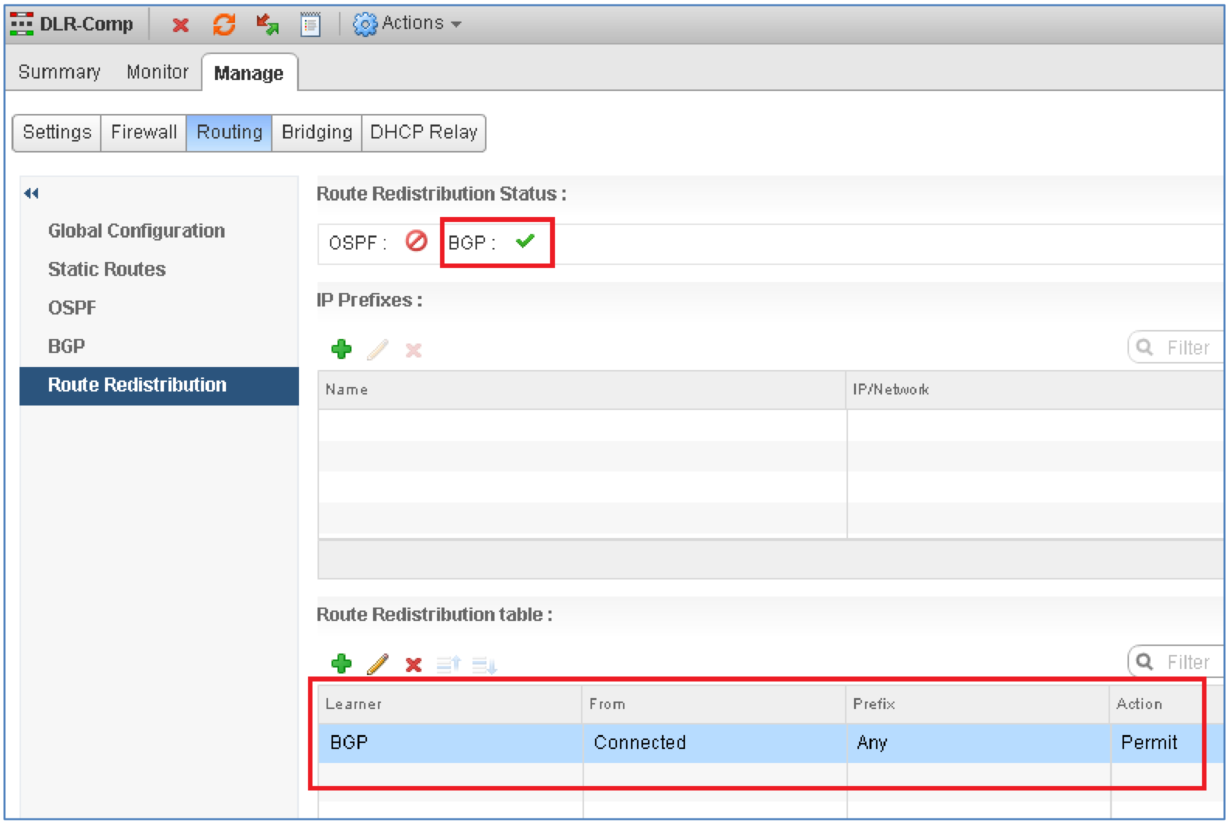


Figure 152 DLR-Comp route redistribution settings

14.2.3 Configure BGP on ESGs

BGP configuration and neighbor settings for the two ESGs are shown in the tables below. Each ESG has two neighbors: the DLR downstream and the BCF logical router upstream.

Table 49 ESG BGP configuration settings

NSX Manager	ESG	BGP	BGP Graceful Restart	Local AS
100.67.187.180	ESG-Mgmt	Enabled	Enabled	64505
100.67.187.181	ESG-Comp	Enabled	Enabled	64511

Table 50 ESG BGP neighbor settings

NSX Manager	ESG	Neighbor	IP Address	Remote AS
100.67.187.180	ESG-Mgmt	DLR-Mgmt	172.16.0.3	64506
		BCF: mgmtvc01 tenant	172.16.51.1	64504
100.67.187.181	ESG-Comp	DLR-Comp	172.16.1.3	64512
		BCF: compvc01 tenant	172.16.61.1	64502

- The neighbor name is shown for reference only.
- The IP Address column in Table 50 is the protocol address of the neighbor.
- Settings not shown in the tables remain at their default values.

The BGP Configuration page is accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the ESG name. Select **Manage > Routing > BGP.**

The figure below shows the BGP configuration settings for ESG-Comp when complete. ESG-Mgmt is similar.

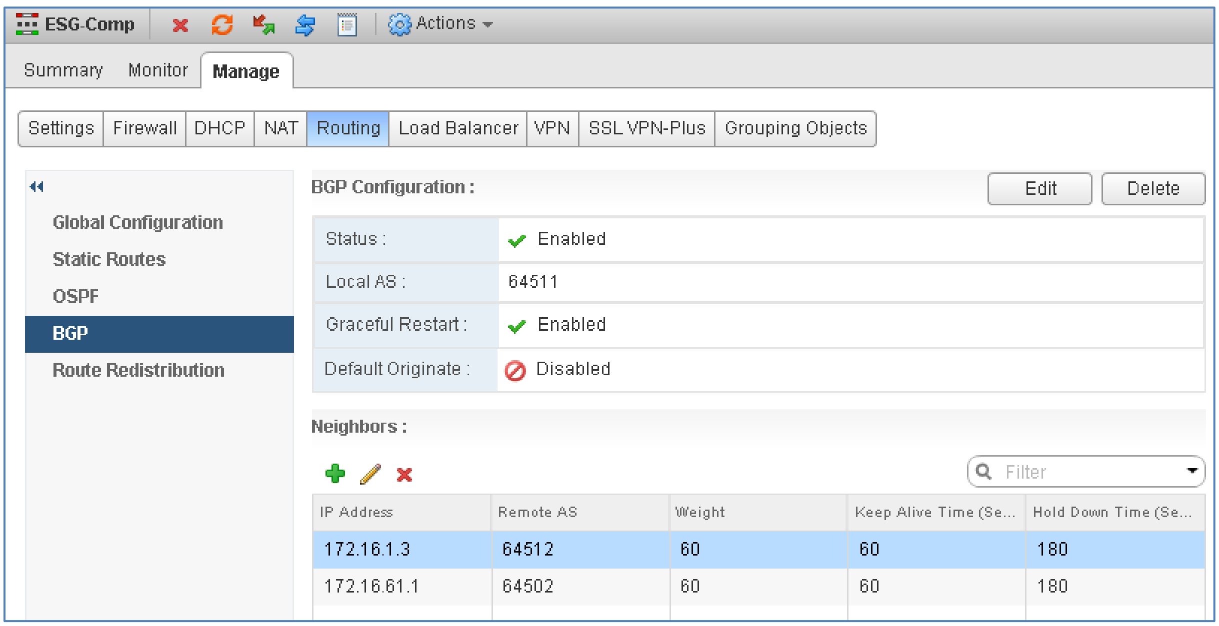


Figure 153 BGP configured on ESG-Comp

Route redistribution is configured on both ESGs using the following settings:

Table 51 ESG route redistribution settings

Enable redistribution for	Learner protocol	Allow learning from	Prefix	Action
BGP	BGP	Connected	Any	Permit

The **Route Redistribution** page is accessed by navigating to **Home > Networking & Security > NSX Edges**. Select the **NSX Manager** and double click on the ESG name. Select **Manage > Routing > Route Redistribution.**

The figure below shows the Route Redistribution settings for ESG-Comp when complete. Settings are identical on ESG-Mgmt.

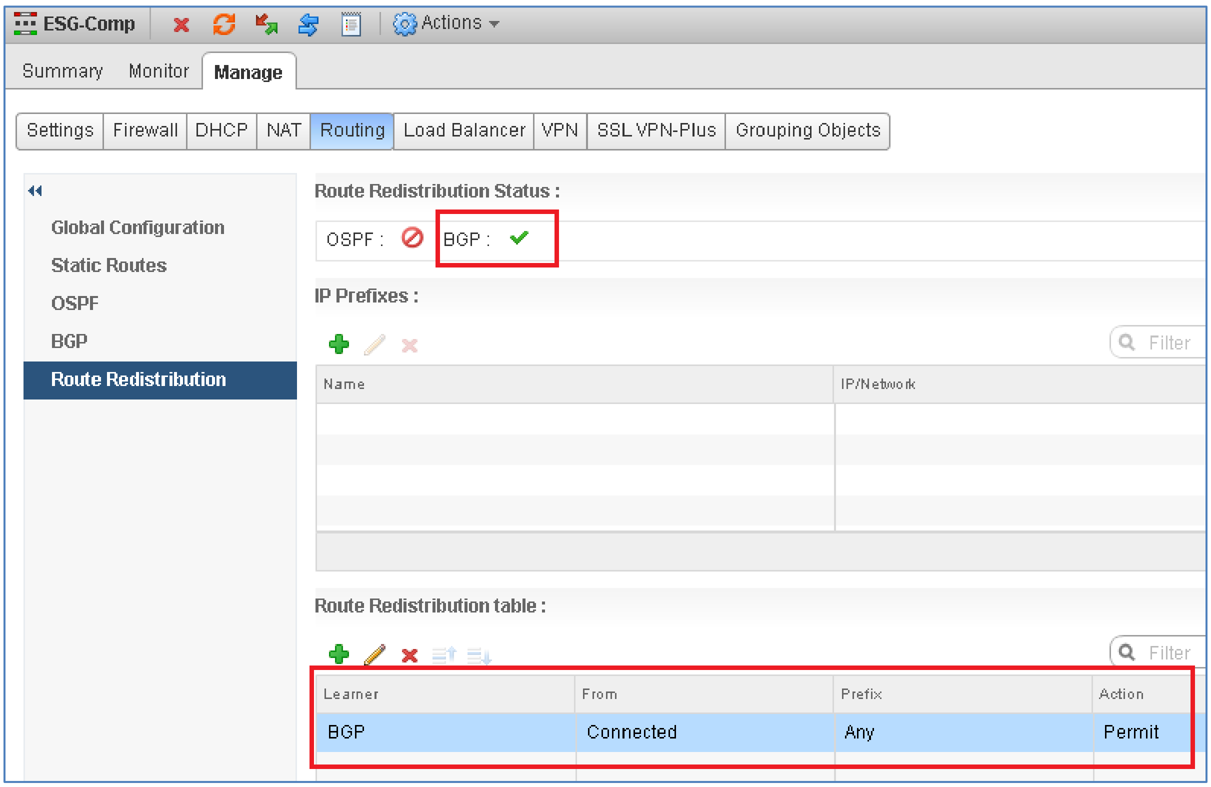


Figure 154 ESG-Comp route redistribution settings

DELL EMC

14.2.4 Validate BGP connections

The commands shown in the following sections are run to verify BGP is functioning properly between neighbors. Be sure the commands are run in the console of the active DLR and ESG VMs. The commands will not succeed on the standby VMs.

> **Note:** NSX Edge active/standby state may be determined by navigating to **Home > Networking & Security > NSX Edges.** Double click on the NSX Edge device, then go to **Manage > Settings > Configuration**. Active/standby state is shown at the bottom of the page under **NSX Edge Appliances**. In this example, the active devices are DLR-Comp-0 and ESG-Comp-0.

14.2.4.1 Verify BGP neighbors on NSX Edges

In the figures below, the command **`show ip bgp neighbors`** is run in the console of the active VMs for DLR-Comp and ESG-Comp to verify BGP connections are established.

In the command output on DLR-Comp, its neighbor is 172.16.1.2 (ESG-Comp). The BGP state is `Established, up`.

```
DLR-Comp-0> show ip bgp neighbors

BGP neighbor is 172.16.1.2,   remote AS 64511,
BGP state = Established, up
Hold time is 180, Keep alive interval is 60 seconds
Neighbor capabilities:
         Route refresh: advertised and received
         Address family IPv4 Unicast:advertised and received
         Graceful restart Capability:advertised and received
                 Restart remain time: 0
Received 1274 messages, Sent 1277 messages
Default minimum time between advertisement runs is 30 seconds
For Address family IPv4 Unicast:advertised and received
         Index 1 Identifier 0x686cd42c
         Route refresh request:received 0 sent 0
         Prefixes received 2 sent 3 advertised 3
Connections established 2, dropped 1
Local host: 172.16.1.3, Local port: 179
Remote host: 172.16.1.2, Remote port: 19177

DLR-Comp-0> _
```

Figure 155 Command output of `show ip bgp neighbors` on DLR-Comp

In the command output on ESG-Comp, its neighbors are 172.16.1.3 (DLR-Comp), and 172.16.61.1 (the BCF compvc01 tenant logical router). In both cases, the BGP state is `Established, up`.

```
ESG-Comp-0> show ip bgp neighbors

BGP neighbor is 172.16.1.3,   remote AS 64512,
BGP state = Established, up
Hold time is 180, Keep alive interval is 60 seconds
Neighbor capabilities:
         Route refresh: advertised and received
         Address family IPv4 Unicast:advertised and received
         Graceful restart Capability:advertised and received
                 Restart remain time: 0
Received 6453 messages, Sent 6460 messages
Default minimum time between advertisement runs is 30 seconds
For Address family IPv4 Unicast:advertised and received
         Index 2 Identifier 0x2b1a23dc
         Route refresh request:received 0 sent 0
         Prefixes received 3 sent 7 advertised 7
Connections established 3, dropped 2
Local host: 172.16.1.2, Local port: 24694
Remote host: 172.16.1.3, Remote port: 179


BGP neighbor is 172.16.61.1,   remote AS 64502,
BGP state = Established, up
Hold time is 180, Keep alive interval is 60 seconds
Neighbor capabilities:
         Route refresh: advertised and received
         Address family IPv4 Unicast:advertised and received
         Graceful restart Capability:advertised and received
                 Restart remain time: 0
Received 60 messages, Sent 67 messages
Default minimum time between advertisement runs is 30 seconds
For Address family IPv4 Unicast:advertised and received
         Index 3 Identifier 0x2b1a23dc
         Route refresh request:received 1 sent 0
         Prefixes received 6 sent 4 advertised 4
Connections established 1, dropped 70
Local host: 172.16.60.1, Local port: 179
Remote host: 172.16.61.1, Remote port: 49294

ESG-Comp-0>
```

Figure 156 Command output of **`show ip bgp neighbors`** on ESG-Comp

The command output on DLR-Mgmt and ESG-Mgmt is similar to the examples shown above, using the neighbors on the management side of the topology.

14.2.4.2 Verify BGP neighbors in BCF

BGP Neighbor state may also be viewed in BCF by going to the **BGP Neighbors** section of the applicable tenant page. The figure below from the compvc01 tenant page shows the BGP state is established between the compvc01 logical router and ESG-Comp.

BGP Neighbors

			Name ▲	State	Type	IP Address	Remote Autonomous System ID	Admin Status	Max Prefix Warn Only	Connection Hold Time After Restart (s)	Store Inbound Routing Table Updates
	≡	▷	ESG-Comp	Established	External	172.16.60.1	64511	✓ Up	✗ No	90	✓ Yes

Figure 157 BGP Neighbors section of tenant compvc01 in BCF

14.2.4.3 Verify route tables on NSX Edges

In the figures below, the command **`show ip route`** is run in the console of the active VMs for DLR-Comp and ESG-Comp.

```
DLR-Comp-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: 10

S       0.0.0.0/0            [1/1]         via 172.16.1.2
C       10.1.2.0/24          [0/0]         via 10.1.2.254
C       10.1.3.0/24          [0/0]         via 10.1.3.254
C       169.254.1.0/30       [0/0]         via 169.254.1.1
C       172.16.1.0/24        [0/0]         via 172.16.1.3
B       172.16.22.0/24       [20/0]        via 172.16.1.2
B       172.16.24.0/24       [20/0]        via 172.16.1.2
B       172.16.32.0/24       [20/0]        via 172.16.1.2
B       172.16.34.0/24       [20/0]        via 172.16.1.2
B       172.16.60.0/24       [20/0]        via 172.16.1.2
DLR-Comp-0>
```

Figure 158 Command output of **`show ip route`** on DLR-Comp

The DLR-Comp output above shows its default gateway is set to the IP address of ESG-Comp, 172.16.1.2. Its directly connected routes are shown with a “C” in the left column. The output on DLR-Mgmt is similar.

> **Note:** 172.16.x.x routes shown with a “B” in the left column have been learned from the ESG via BGP. These routes are to networks not used by NSX VMs and may be filtered as shown in Appendix E. The directly connected 169.254.1.0 network shown is for the automatically configured DLR HA interface.

```
ESG-Comp-0> sh 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: 10

S       0.0.0.0/0            [1/1]        via 172.16.60.2
B       10.1.2.0/24          [20/0]       via 172.16.1.1
B       10.1.3.0/24          [20/0]       via 172.16.1.1
C       169.254.1.4/30       [0/0]        via 169.254.1.5
C       172.16.1.0/24        [0/0]        via 172.16.1.2
B       172.16.22.0/24       [20/0]       via 172.16.60.2
B       172.16.24.0/24       [20/0]       via 172.16.60.2
B       172.16.32.0/24       [20/0]       via 172.16.60.2
B       172.16.34.0/24       [20/0]       via 172.16.60.2
C       172.16.60.0/24       [0/0]        via 172.16.60.1
ESG-Comp-0> _
```

Figure 159 Command output of **`show ip route`** on ESG-Comp

The ESG-Comp output shows its default gateway is set to the IP address of the BCF segment interface, 172.16.60.2. Its directly connected routes are shown with a “C” in the left column. ESG-Comp has learned the routes to the NSX VM networks, 10.1.2.0 and 10.1.3.0 from DLR-Comp via BGP as noted by the “B” in the first column.

The output for ESG-Mgmt is similar using the routes on the Management side of the topology.

> **Note:** 172.16.x.x routes shown with a “B” in the left column have been learned from the Compvc01 tenant logical router. These routes are not used by the NSX VMs and may be filtered as shown in Appendix E. The directly connected 169.254.1.0 network shown is for the automatically configured ESG HA interface.

14.2.4.4 Verify route tables in BCF

To verify routes to the NSX VM networks have been learned by the BCF tenant logical routers, do the following:

1. In the BCF GUI, navigate to **Logical > Tenants**.
2. Click the ▷ next to **compvc01** to view the compvc01 tenant’s **Routes** table.

3. Under **Routes**, routes learned from ESG-Comp are classified as **Dynamic** and appear as outlined in red below:

Routes

	Configured	Preference	Description	CIDR	Type	Next Hop Tenant	Next Hop Group	Next Hop IP Address
☐	–	0	–	172.16.34.0/24	Connected	compvc01	Segment Iface compvc01-1634	–
☐	–	0	–	172.16.24.0/24	Connected	compvc01	Segment Iface compvc01-1624	–
☐	–	0	–	172.16.32.0/24	Connected	compvc01	Segment Iface compvc01-1632	–
☐	–	0	–	172.16.22.0/24	Connected	compvc01	Segment Iface compvc01-1622	–
☐	–	0	–	172.16.60.0/24	Connected	compvc01	Segment Iface compvc01-1660	–
☐	–	20	–	172.16.1.0/24	Dynamic	compvc01	DRNH--0	172.16.60.1
☐	–	20	–	10.1.3.0/24	Dynamic	compvc01	DRNH--0	172.16.60.1
☐	–	20	–	10.1.2.0/24	Dynamic	compvc01	DRNH--0	172.16.60.1

Figure 160 Dynamic routes – compvc01 tenant

The **Routes** table for the mgmtvc01 tenant, is as shown below:

Routes

	Configured	Preference	Description	CIDR	Type	Next Hop Tenant	Next Hop Group	Next Hop IP Address
☐	–	0	–	172.16.14.0/24	Connected	mgmtvc01	Segment Iface mgmtvc01-1614	–
☐	–	0	–	172.16.12.0/24	Connected	mgmtvc01	Segment Iface mgmtvc01-1612	–
☐	–	0	–	172.16.50.0/24	Connected	mgmtvc01	Segment Iface mgmtvc01-1650	–
☐	–	20	–	10.1.1.0/24	Dynamic	mgmtvc01	DRNH--0	172.16.50.1
☐	–	20	–	172.16.0.0/24	Dynamic	mgmtvc01	DRNH--0	172.16.50.1

Figure 161 Dynamic routes – mgmtvc01 tenant

14.2.5 Connectivity test

Since the ESGs have learned the routes to VMs from the DLRs, ESG-Comp is able to ping the App and Web VMs at this point, and ESG-Mgmt is able to ping the Mgmt VMs.

```
ESG-Comp-0> ping 10.1.3.2
PING 10.1.3.2 (10.1.3.2) 56(84) bytes of data.
64 bytes from 10.1.3.2: icmp_seq=1 ttl=127 time=0.373 ms
64 bytes from 10.1.3.2: icmp_seq=2 ttl=127 time=0.459 ms
64 bytes from 10.1.3.2: icmp_seq=3 ttl=127 time=0.446 ms
```

Figure 162 ESG-Comp pings IP address of VM web-02

Note: VM guest operating system firewalls must be temporarily disabled or configured to allow ICMP traffic for the pings above to succeed. DLR firewalls allow this traffic by default.

If ESG firewalls are disabled, or configured to allow such traffic, ESG-Comp can also ping the Mgmt VMs and ESG-Mgmt can ping the App and Web VMs. Likewise, all NSX VMs can ping each other at this point. This is because BCF has learned the routes to the VMs, and BCF routes traffic from compute VMs to management VMs through the System tenant.

Configure firewalls to meet your needs and refer to the *VMware NSX for vSphere 6.3 Administration Guide* for NSX Edge firewall configuration.

15 S4048-ON core router

In this section, the S4048-ON core router is configured and communication is validated from a physical server on the external network to the NSX VMs.

Note: Configuration of redundant core routers is outside the scope of this document. For this guide, a single S4048-ON switch running DNOS 9.11 is used as the core router.

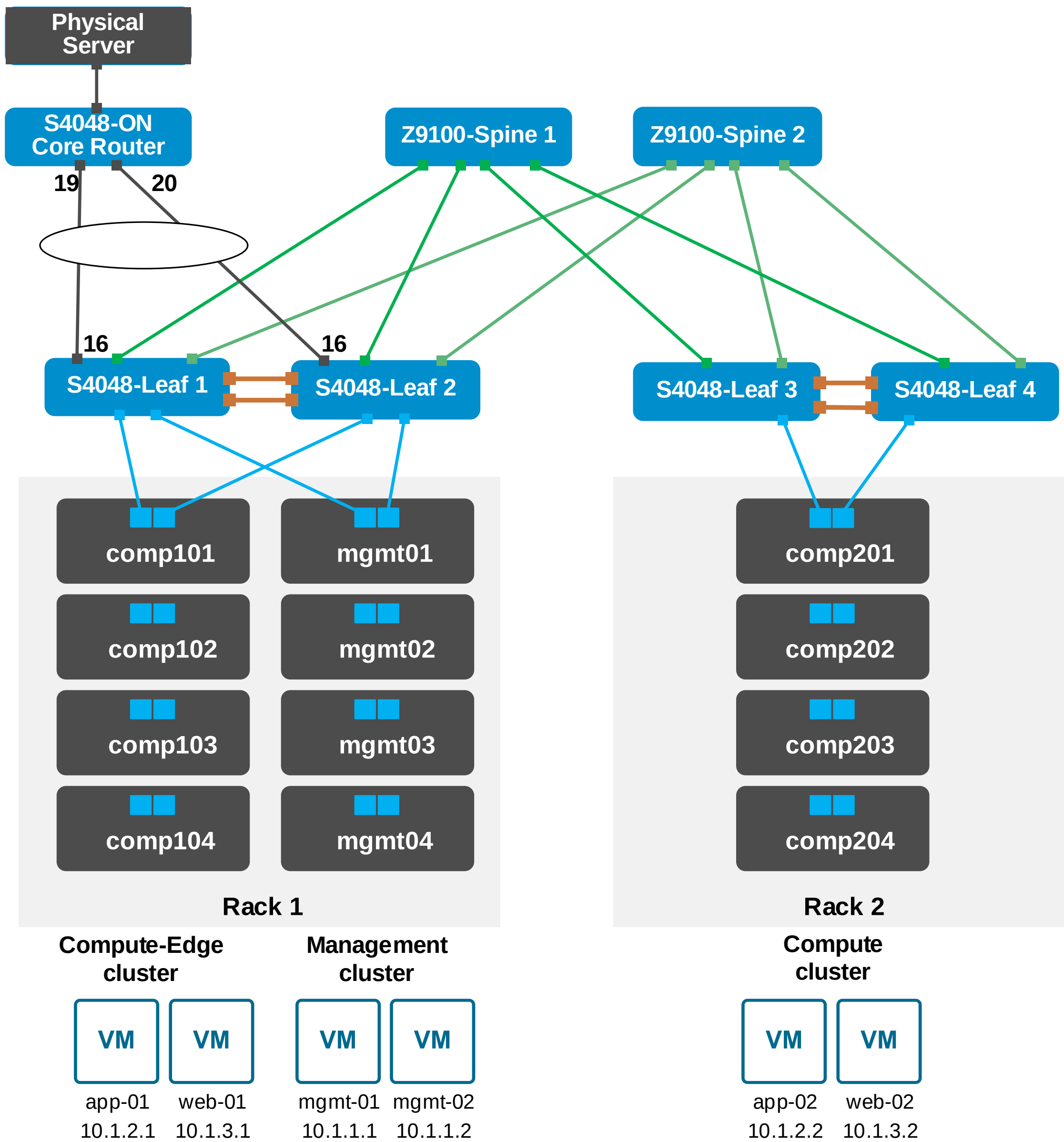


Figure 163 Physical server on external network and NSX VMs

15.1 S4048-ON configuration

The downstream connections from the S4048-ON to the leaf switches are configured in LACP port channel 1. This is connected to the LACP port channel configured on the leaf switches in Section 12.5.1. Setting the MTU to 9216 as shown is optional.

```
interface TenGigabitEthernet 1/19
 no ip address
 mtu 9216
!
 port-channel-protocol LACP
  port-channel 1 mode active
 no shutdown
!
interface TenGigabitEthernet 1/20
 no ip address
 mtu 9216
!
 port-channel-protocol LACP
  port-channel 1 mode active
 no shutdown
```

The port channel is put in layer 3 mode by assigning it an IP address. This is the same address configured in Section 14.1.2 as the next hop IP address on the BCF External tenant.

```
interface Port-channel 1
 ip address 172.16.100.1/24
 mtu 9216
 no shutdown
```

BGP is configured to form an adjacency with the BCF External tenant using the information shown in Figure 164:

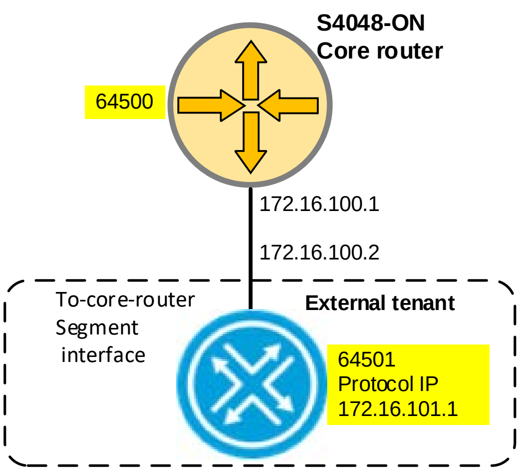


Figure 164 Core router BGP information

```
router bgp 64500
 bgp bestpath as-path multipath-relax
 redistribute connected
 bgp graceful-restart
 neighbor 172.16.101.1 remote-as 64501
 neighbor 172.16.101.1 ebgp-multihop 2
 neighbor 172.16.101.1 no shutdown
```

A static route is configured to the BCF External tenant's BGP protocol IP address downstream. A default gateway is set to the IP address of the next upstream router (12.0.0.2 for example). Routes to NSX VMs are learned via BGP.

```
ip route 172.16.101.1/32 172.16.100.2
ip route 0.0.0.0/0 12.0.0.2
```

Note: Upstream connections to the physical server and upstream router(s) are not included in this configuration.

15.2 Core router validation

15.2.1 show ip bgp neighbors

The command **`show ip bgp neighbors`** is run on the core router to verify a BGP connection is established with the BCF External tenant logical router:

```
S4048-Core#show ip bgp neighbors
BGP neighbor is 172.16.101.1, remote AS 64501, external link
  BGP remote router ID 172.16.101.1
  BGP state ESTABLISHED, in this state for 00:01:30
  Last read 00:00:00, Last write 00:00:30
  Hold time is 90, keepalive interval is 30 seconds
  Received 8 messages, 0 in queue
```

(output truncated)

15.2.2 show ip route

The command **`show ip route`** is run on the core router to verify routes to NSX VM networks, 10.1.x.0/24, are learned via BGP:

```
S4048-Core#show ip route

Codes: C - connected, S - static, R - RIP,
       B - BGP, IN - internal BGP, EX - external BGP,LO - Locally
Originated,
       O - OSPF, IA - OSPF inter area, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type 2, E1 - OSPF external type 1,
       E2 - OSPF external type 2, i - IS-IS, L1 - IS-IS level-1,
       L2 - IS-IS level-2, IA - IS-IS inter area, * - candidate default,
       > - non-active route, + - summary route

Gateway of last resort is 12.0.0.2 to network 0.0.0.0

       Destination        Gateway                  Dist/Metric Last Change
       -----------        -------                  ----------- -----------
 *S    0.0.0.0/0          via 12.0.0.2, Vl 12              1/0       3d17h
  B EX 10.1.1.0/24        via 172.16.101.1                20/0       1d18h
  B EX 10.1.2.0/24        via 172.16.101.1                20/0       1d18h
  B EX 10.1.3.0/24        via 172.16.101.1                20/0       1d18h
  C    12.0.0.0/24        Direct, Vl 12                    0/0        4w2d
  C    172.16.100.0/24    Direct, Po 1                     0/0       1d18h
  S    172.16.101.1/32    via 172.16.100.2, Po 1           1/0       1d18h
```

15.3 End-to-end Validation

The end-to-end topology shown in Figure 144 can now be validated. Provided firewalls are configured to allow this traffic, the physical server on an external network is now able to send traffic to any of the NSX VMs.

To see the path taken through the physical and virtual networks, the **`tracert`** command is run on the physical server to the IP address of the NSX VM named web-02:

```
C:\Windows\system32>tracert 10.1.3.2

Tracing route to 10.1.3.2 over a maximum of 30 hops

  1    <1 ms    <1 ms    <1 ms  12.0.0.1
  2    <1 ms    <1 ms    <1 ms  172.16.100.2
  3     *        *        *     Request timed out.
  4     *        *        *     Request timed out.
  5    <1 ms    <1 ms    <1 ms  172.16.60.1
  6    <1 ms    <1 ms    <1 ms  172.16.1.1
  7     1 ms    <1 ms    <1 ms  10.1.3.2

Trace complete.
```

Figure 165 Tracing route from physical server through the virtual networks to NSX VM

16 BCF 4.6 NSX visibility enhancements

BCF 4.6 was released shortly before publication of this guide, and includes NSX visibility improvements that we were unable to include in detail in the preceding sections.

After entering NSX Manager credentials in BCF, graphical views of the NSX topology become available. For example, the following is a view of the App logical switch used in the deployment covered in this guide:

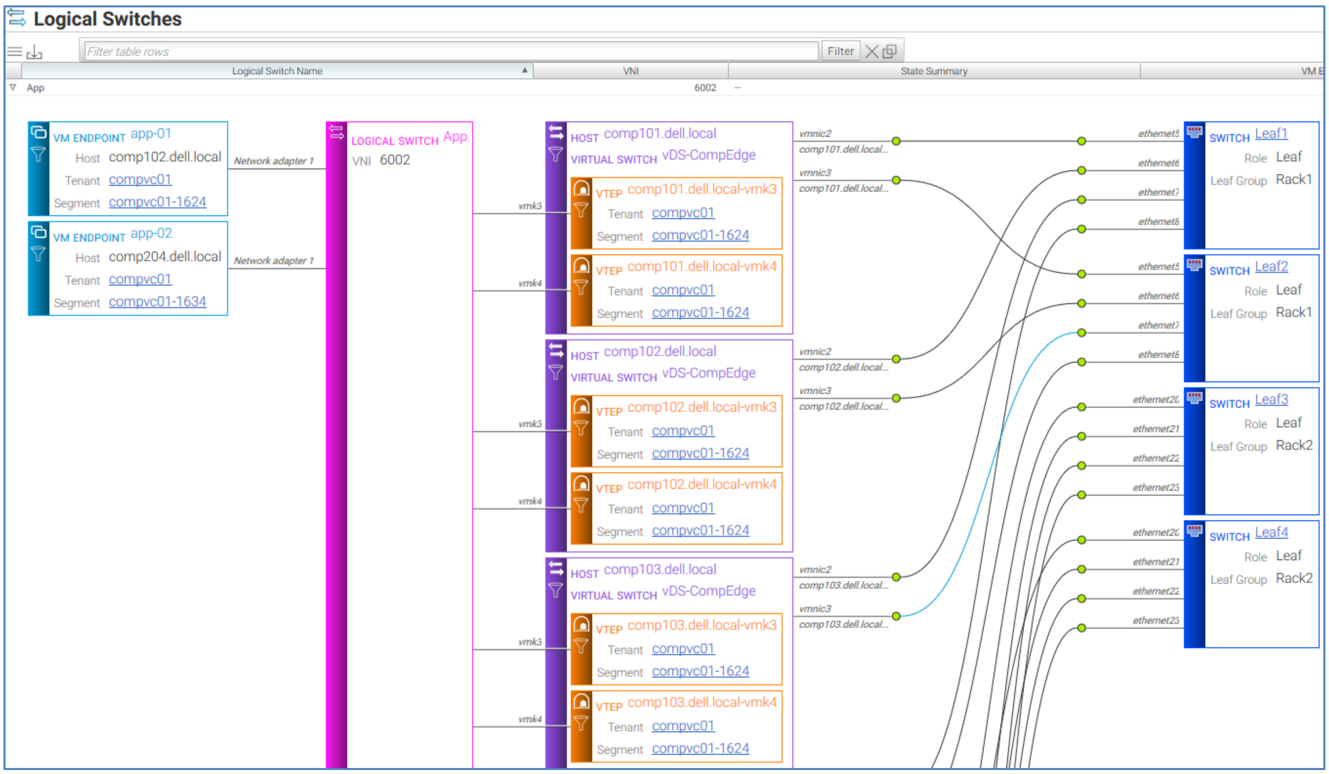


Figure 166 BCF graphical view of NSX logical switch

The view shown includes the NSX VMs connected to the App logical switch (app-01 and app-02), the ESXi hosts and their VTEPs, and connections from vmnics to the leaf switches.

See the *Big Cloud Fabric 4.6 GUI Guide* for more information on configuring and using NSX visibility features available in BCF 4.6.

A Rack diagrams

The racks and equipment used in this deployment guide are shown in Figure 167.

Each rack contains one S3048-ON OOB management switch and two S4048-ON leaf switches. Rack 1 also contains the Z9100-ON spine switches, two BCF Controllers, the Management cluster (four R630 hosts), and the Compute-Edge cluster (four R740xd hosts). Rack 2 contains the Compute cluster (four R630 hosts).

Rack 2 has total capacity for up to:

- 38 1-RU hosts (R630, R640, etc.) with the addition of a 2nd S3048-ON management switch, or
- 19 2-RU hosts (R730, R740, R740xd, etc.) with no additional switches required

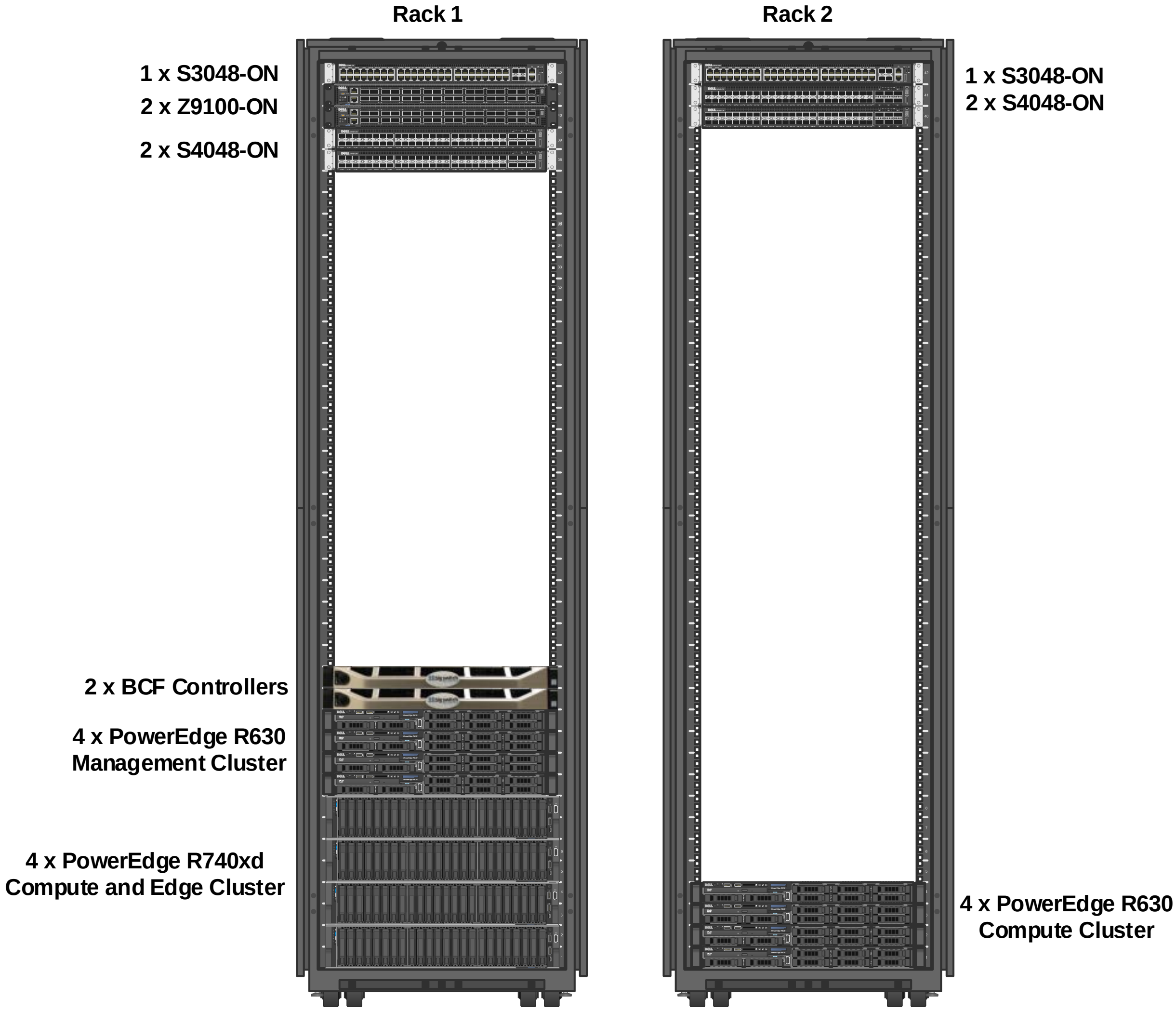


Figure 167 Racks and equipment used in this guide

B Dell EMC validated hardware and component versions

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

B.1 Switches

Qty	Item	OS/Firmware version
2	S3048-ON Management switch	**OS:** DNOS 9.13.0.0
		System CPLD: 9
		Module CPLD: 7
4	S4048-ON Leaf switch Switch Light OS and CPLD/ONIE firmware are provided by controller running BCF 4.6.0	**OS**: Switch Light 4.6.0
		CPLD: 15.12.5,3.21.0.0-5
		ONIE: 3.21.1.2
2	Z9100-ON Spine switch Switch Light OS and CPLD/ONIE firmware are provided by controller running BCF 4.6.0	**OS:** Switch Light 4.6.0
		CPLD: 6.4.4.4,3.23.0.0-7
		ONIE: 3.23.1.4
1	S4048-ON Core router	**OS**: DNOS 9.13.0.0
		System CPLD: 15.2
		Master CPLD: 12
		Slave CPLD: 5

B.2 PowerEdge Servers

Note: VVD 4.1 recommends all server nodes have uniform configurations across a given cluster. A balanced cluster delivers more predictable performance and impact during resync/rebuild is minimal.

B.2.1 PowerEdge R740xd servers – Compute-Edge cluster

Qty per server	Item	Firmware version
2	Intel Xeon Gold 6130 CPU @ 2.10GHz, 16 cores	-
64	GB RAM	-
20	400GB SAS SSD	-
1	Dell HBA330 Storage Controller	13.17.03.00
1	Broadcom QP rNDC: 5720 DP 1GbE Base-T + 57412 DP 10GbE SFP+	5720: 20.6.16 57412: 20.06.05.06
-	BIOS	1.0.7
-	iDRAC with Lifecycle Controller	3.00.00.00

B.2.2 PowerEdge R630 servers – Compute cluster

Qty per server	Item	Firmware Version
2	Intel Xeon E5-2695 v3 2.3GHz CPU, 14 cores	-
128	GB RAM	-
2	400GB SATA SSD	-
6	600GB SAS HDD	-
1	PERC H730 Mini Storage Controller	25.5.3.0005
1	Intel 2P X520 10GbE/2P I350 1GbE Base-T rNDC	18.0.17
-	BIOS	2.5.5
-	iDRAC with Lifecycle Controller	2.50.50.50

B.2.3 PowerEdge R630 servers – Management cluster

Qty per server	Item	Firmware Version
1	Intel Xeon E5-2650 v3 2.3GHz CPU, 10 cores	-
64	GB RAM	-
2	400GB SATA SSD	-
6	300GB SAS HDD	-
1	PERC H730 Mini Storage Controller	25.5.3.0005
1	QLogic 57840S 10GbE QP rNDC	10.01.00
1	Intel I350-T 1GbE Base-T DP PCIe adapter	18.0.17
-	BIOS	2.5.5
-	iDRAC with Lifecycle Controller	2.50.50.50

C Validated software and required licenses

The Software table lists the software components used to validate the configurations in this guide. The VMware Licenses section lists the VMware licenses required for the configurations used in this guide.

C.1 Software

Item	Version
Big Cloud Fabric	4.6.0
VMware ESXi	6.5 U1 - Dell EMC customized image version A07, build 7388607
VMware vCenter Server Appliance	6.5 Update 1d – build 7312210
VMware vSAN	6.6.1 Patch 02 (provided with ESXi 6.5 build 7388607)
VMware NSX for vSphere	6.3.5 build 7119875

C.2 VMware Licenses

vCenter Servers are 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
- VMware vCenter 6 Server Standard – 2 instances
- VMware NSX Enterprise - 20 CPU sockets
- VMware vSAN Standard – 20 CPU sockets

VMware product licenses are centrally managed by going to the **vSphere Web Client Home** page and selecting **Licensing** in the center pane.

D Product manuals and technical guides

D.1 Dell EMC

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

Dell EMC TechCenter Networking Guides

Dell EMC Ready Bundle for Virtualization Datasheet

Manuals and documentation for Dell EMC Networking S3048-ON

Manuals and documentation for Dell EMC Networking S4048-ON

Manuals and documentation for Dell EMC Networking Z9100-ON

D.2 Big Switch Networks

Big Switch Networks customer support site – Contains support information and BCF user guides, including:

Big Cloud Fabric 4.6 Deployment Guide

Big Cloud Fabric 4.6 User Guide

Big Cloud Fabric 4.6 GUI Guide

Big Cloud Fabric 4.6 CLI Reference Guide

Note: An account is required to use the support site. Contact your Big Switch Networks account representative for information.

Big Cloud Fabric Datasheet

Big Cloud Fabric: A Next-Generation Data Center Switching Platform

Big Switch Networks + Dell: Ideal SDN Fabric for VMware SDDC

D.3 VMware

D.3.1 General

VMware vSphere Documentation

vSphere Installation and Setup – This document includes ESXi 6.5 and vCenter Server 6.5.

VMware Compatibility Guide

VMware Validated Design Documentation – Release 4.1

D.3.2 VMware vSAN

VMware vSAN Technical Resources

VMware vSAN Design and Sizing Guide

VMware vSAN Operations Guide

D.3.3 VMware NSX

VMware NSX for vSphere Documentation

VMware NSX for vSphere 6.3 Installation Guide

VMware NSX for vSphere 6.3 Administration Guide

VMWare NSX for vSphere 6.3 – Recommended Configuration Maximums

VMware NSX for vSphere 6.3 - Troubleshooting Guide

E BGP route filtering

The compvc01 and mgmtvc01 tenant networks in this guide are stub networks, and downstream devices need only default routes to reach networks upstream. BGP is used in this deployment guide to advertise NSX VM networks to upstream routers.

By default, routes not used by NSX VMs (such as routes to the physical vMotion and VXLAN networks) are advertised by BCF logical routers downstream to the ESGs which in turn advertise them to the DLRs. These routes may be suppressed using BGP route filtering.

In BCF, route filtering is done using a route map and applying it outbound to the downstream BGP neighbor. This is done as follows:

1. In the BCF GUI, select **Logical > Tenants > mgmtvc01** to open the **mgmvc01** Tenant page.
2. Select **Neighbors** and **Route Maps** in the left pane as shown:

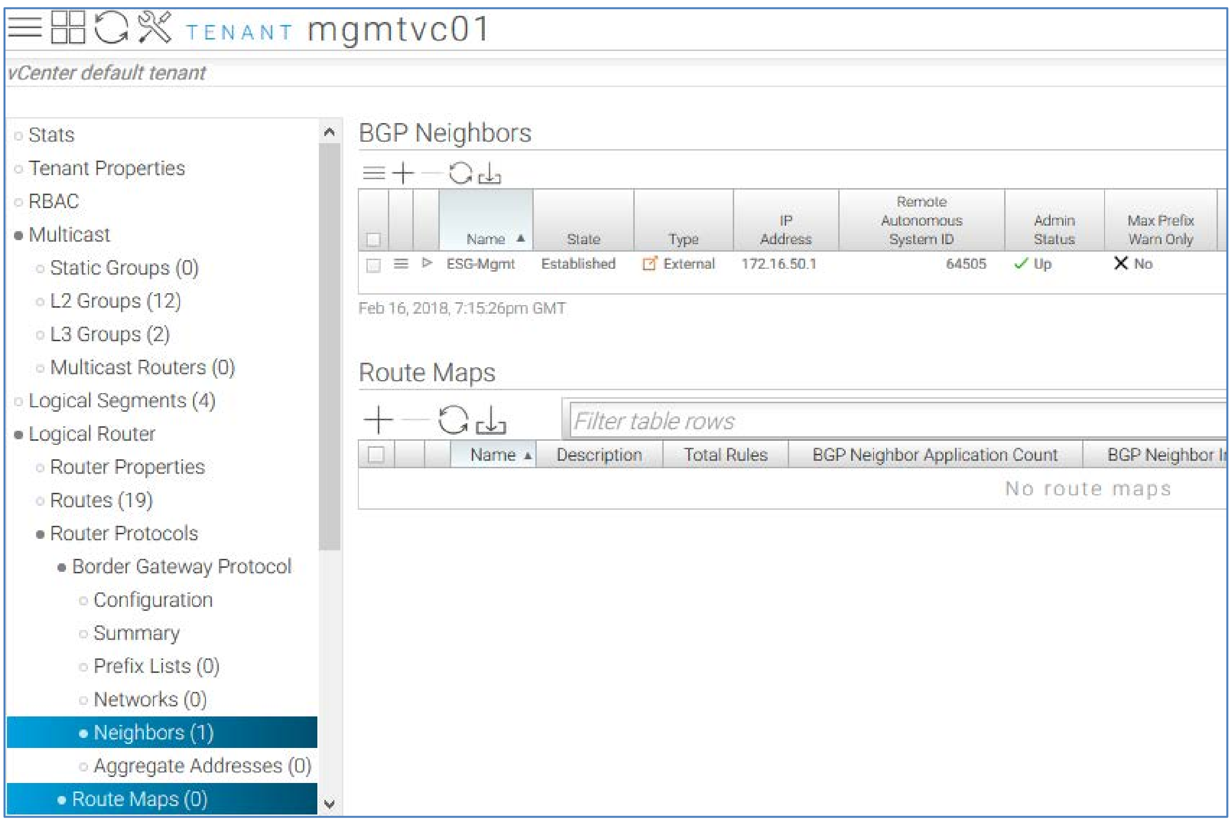


Figure 168 Route maps and BGP neighbors selected

3. In the right pane under **Route Maps,** click the + icon to open the **Create Route Map** dialog box.
4. Enter a **Name** for the route map such as **suppressAdvert** and click **Next** to go to the **Rules** tab.
5. On the **Rules** tab, click the + icon to open the **Configure Route Map Rule** dialog box, and move the **Action** slider to **Deny** as shown:

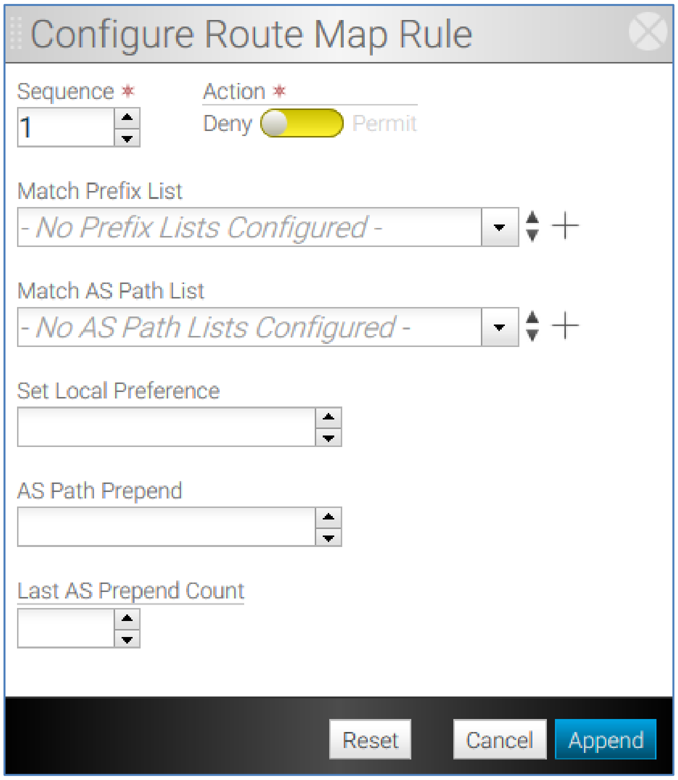


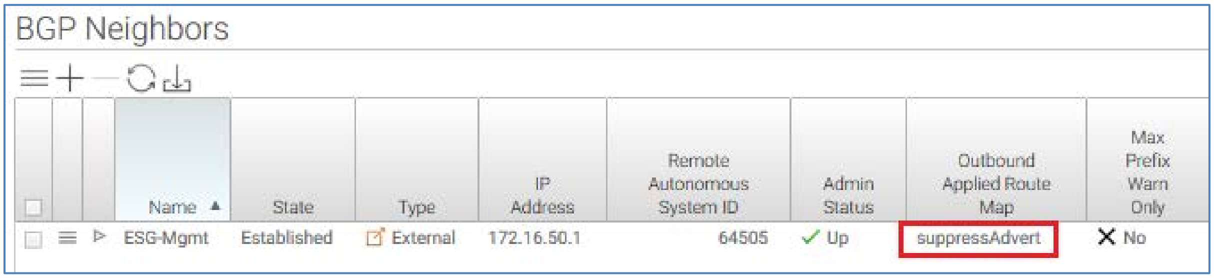
Figure 169 Create route map rule dialog box

6. Leave all other settings at their defaults and click **Append**.
7. In the **Configure Route Map Rule** dialog box, Click **Save** to create the route map.

Apply the route map to the BGP neighbor as follows:

1. On the **mgmtvc01 tenant** page under **BGP Neighbors**, click the ≡ icon next to the BGP neighbor name, **ESG-Mgmt** in this example.
2. Select **Edit** to open the **Edit BGP Neighbor** dialog box.
   a. Click **Next** to go to the **Options** tab.
   b. Under **Outbound Route Map**, select the route map that was just created.
   c. Click **Save.**

The route map is applied to the BGP neighbor as shown:


BGP Neighbors

			Name ▲	State	Type	IP Address	Remote Autonomous System ID	Admin Status	Outbound Applied Route Map	Max Prefix Warn Only
			ESG-Mgmt	Established	External	172.16.50.1	64505	Up	suppressAdvert	No

Figure 170 Outbound route map applied on mgmtvc01 logical router

Repeat the above on the compvc01 tenant.

To verify that route map is functioning properly, verify that the dynamic routes learned from the ESG are still in the tenant **Routes** table, and that the ESG is not learning routes from the BCF tenant router.

Note: The DLR will continue to learn the ESG's directly connected route to the BCF tenant router, 172.16.50.0 on DLR-Mgmt for example, via BGP. This can be filtered in a similar manner by creating a BGP filter on the ESG. See the *VMware NSX for vSphere 6.3 Administration Guide* for more information.

F 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.