Dell EMC Network Virtualization Handbook for VMware NSX Enterprise 6.2.x

A Dell EMC-VMware Reference Architecture

September 2016

Table of Contents

Table of Figures

Overview

The following document covers the standard reference architecture of networking virtualization with Dell infrastructure and provides a configuration guide of the Dell S6000 Layer 2 hardware VTEP gateway as a handbook.

Thanks to the success of server virtualization over the last decade, new needs have now come to the forefront: to virtualize the network or to decouple the network services from the underlying physical infrastructure. Software-Defined Data Center (SDDC) is the term given to the ability to logically represent a physical infrastructure and its network services within software. The same benefits that made server virtualization so successful are now also driving network virtualization and the SDDC. Key drivers include:

- Speed of deployment/migration, flexibility, and agility
- Automation
- Minimized downtime
- Normalization of underlying hardware

vCenter Suite has been the traditional control plane of virtual machines running on a pool of server hardware. The control plane benefits from network virtualization using VMware NSX, the marketing leading implementation of network virtualization. NSX delivers the operational model of a virtual machine with networking and security functionality embedded directly in the hypervisor. NSX offers a centralized fine-grained policies to provision and configure multiple isolated logical networks that run on a single physical network.

Logical networks are decoupled from physical network services like routing and switching, giving service providers and enterprises the flexibility to place a VM anywhere in the data center. Along with this connectivity between VMs and different tiers, NSX helps add layer 4-7 network services to secure different logical domains. Micro-segmentation is a concept based on the premise of isolation and the zero-trust environment created by the combination of different network services.

This paper will discuss the physical layer components of server connectivity inside the rack to the networking infrastructure. Next, it addresses the Layer 2 and Layer 3 network setup in the underlying physical network.

Building a scalable networking fabric is one of the key steps in the Software-Defined Data Center. This document discusses how to build a scalable underlying network for VMware NSX-based network virtualization.

Network virtualization, though sometimes considered a new trend in the networking discussion, is really not a new approach or methodology. What *is* new is the correlation between server virtualization and physical network virtualization.

The definition of network virtualization changes depending on who is in the conversation. However, if used in the same context as server virtualization, the most relevant definition would likely be:

> "…the ability to create logical, virtual networks that are decoupled from the underlying network hardware to ensure the network can better integrate with and support increasingly virtual environments…" (SDX Central).

The key concept of virtualization is the software aspect. Software is agile, scalable and repeatable when needed. Network virtualization aims to reproduce all the network services from Layer 2 – Layer 7 of the OSI model. Similar to how server virtualization reproduces a virtual CPU, RAM, and NIC at the virtual layer, network virtualization does the same with a logical switch, router, firewall, etc.

With this abstraction, the necessary network configuration is no longer driven through a CLI (Command Line Interface). Rather, all provisioning is driven and delivered via APIs at the software virtual switch interface.

Two major components make up network virtualization:

- Network Underlay
- Network Overlay

This paper covers the basics of physical network underlay and sets a basic networking infrastructure in place before the VMware NSX is installed and operationalized for overlay network capabilities.

1 Network virtualization components

1.1 Network underlay

The underlay's main role is to provide an IP transport highway for end hosts to communicate with each other with the **overlay** riding on top like a simple payload. The underlay must be:

Non-blocking - the fabric created by the underlay must be able to switch data at line-rate and make use of all fabric interlinks ranging from 1GE-100GE speeds.

Dynamic - the underlay must be able to scale on demand, depending on the type of overlays being created on top of it. As new services are added or deleted, the overlay infrastructure should be able to grow or shrink.

Open – the dynamic aspect of the underlay requires it to be an open architecture, defined as a mixed set of products from different vendors and different operating systems. Without this, the underlay created is a monolithic entity. Completeness can only be achieved by seamlessly integrating other components into the underlay.

1.2 ESXi Hypervisor

VMware vSphere includes a hypervisor called ESXi, with a version that gets installed on bare-metal servers and decouples the server operating system from the underlying hardware. The hypervisor manages and allocates the resources of the server between VMs. For configurable maximums, please refer to the vSphere 6.0 guide here. Specific VIBs (vSphere Installation Bundles) are kernel based pack functions, such as VXLAN bridging, distributed routing, and distributed firewall.

1.3 vCenter

vCenter is the main pane of glass from which all servers and clusters are managed. It manages multiple ESXi hosts (depending on the version, up to 1000 hosts), and allows for a centralized location for all ESXi

host configurations. vCenter is required for advanced operations, VDS, and NSX. vCenter is offered as both standalone software to be installed on a Window server and as a virtual appliance where vCenter is installed on top of SUSE Linux Enterprise and provided as a VM form factor. The virtual appliance version of vCenter is installed as an Open Virtualization Appliance (OVA) on a vSphere ESXi host. vSphere Web Client allows for connecting to vCenter Server via a web browser. Once connected, ESXi host configuration and operations can be done.

1.4 vCenter Clusters

A group of servers performing similar functions are managed together as a cluster. Clusters help configure, manage the servers and maintain uniformity within the cluster. Typically clusters are classified as Management and compute resource clusters. Management cluster hosts VMs responsible for management, configuration, monitoring and troubleshooting the servers. Compute cluster hosts the application workloads. In the case of NSX, a dedicated cluster known as Edge cluster is configured to perform various edge connectivity services between the the physical and logical networks. Section 4.2 discuss the configured clusters in this handbook in detail.

1.5 NSX Manager

NSX Manager is the centralized network management component of NSX for configuration and operation. A NSX Manager installation maps to a single vCenter Server environment. NSX Manager is installed as an OVA on a vSphere ESXi host. Once installed, NSX Manager allows for installation, configuration, and management of other NSX components via a GUI management plugin for vCenter.

1.6 NSX Controller

The NSX Controller maintains communication with the hypervisor to establish and manage the virtual networks, which consist of overlay transport tunnels. The NSX controller cluster is an advanced distributed state management system that manages virtual networks and overlay transport tunnels. The cluster is a group of VMs that run on any x86 server; each controller can be installed on a different server. Controllers are installed as a virtual appliance using the NSX Manager plug-in via vCenter. Three controllers are required in a supported configuration and can tolerate one controller failure while still providing for controller functionality. Data forwarding is not affected by controller cluster failure.

1.7 Virtual Distributed Switch (VDS)

A virtual distributed switch (VDS) is the basic building block of overall NSX architecture. VDS helps to uniformly configure ESXi host networking and to manage the configuration of ESXi in a given cluster situated across different racks from a single place. VDS helps establish communication between NSX manager and ESXi host (management plane) for information exchange, with no SX controllers and ESXi (control plane) over a SSL communication channel to populate various host tables (e.g., MAC address and ARP tables).

1.8 VXLAN

VXLAN is an overlay technology based on RFC 7348, using UDP for transporting L2 MAC frames. VXLAN is a MAC-in-UDP encapsulation method. In VXLAN, the original Layer 2 frame is encapsulated inside an IP-UDP packet by adding a VXLAN header. Please refer to the link below for more details on VXLAN.

http://en.community.dell.com/techcenter/networking/m/networking_files/20442272

By encapsulating the IP packet inside a UDP packet, 24-bit VXLAN ID helps solve a number of problems in modern data centers, including 12-bit VLAN range limits, multi-tenancy in cloud computing environments, and limits to exponential MAC address table growth in TOR switches.

1.9 VXLAN Tunnel End Point – software and hardware

VXLAN uses VXLAN Tunnel End Point (VTEP) to perform underlay-to-overlay network destination mapping for encapsulation and de-capsulation purposes. Typically, a software VTEP is a physical server with ESXi hypervisor installed to host multiple VMs, which then talk to NSX controllers over an SSL-authenticated TCP connection. VMware ESXi hypervisors running a distributed vSwitch with NSX User World Agents (UWAs) is an example. VTEPs primarily perform two important functions: first, VTEPs encapsulate the Layer 2 frame into the VXLAN header and then transport that encapsulated frame over an IP network (also called Transport zone) to other VTEPs. The destination VTEP de-capsulates the outer header & VXLAN header and sends the inner Layer 2 frame to the appropriate destination VM. In addition to forwarding the packets, VTEPs learn the MAC addresses of the VMs in the respective VNI (VXLAN Network ID), which is equivalent to the legacy VLAN, and maintain a table with the NSX controller to limit the BUM traffic in the network.

1.9.1 Software VTEP

A software VTEP typically is a physical server with hypervisor installed in it to host multiple VMs that can talk to NSX controllers over an SSL-authenticated TCP connection. VMware ESXi hypervisors running distributed vSwitch with NSX User World Agents is a typical example of software VTEP.

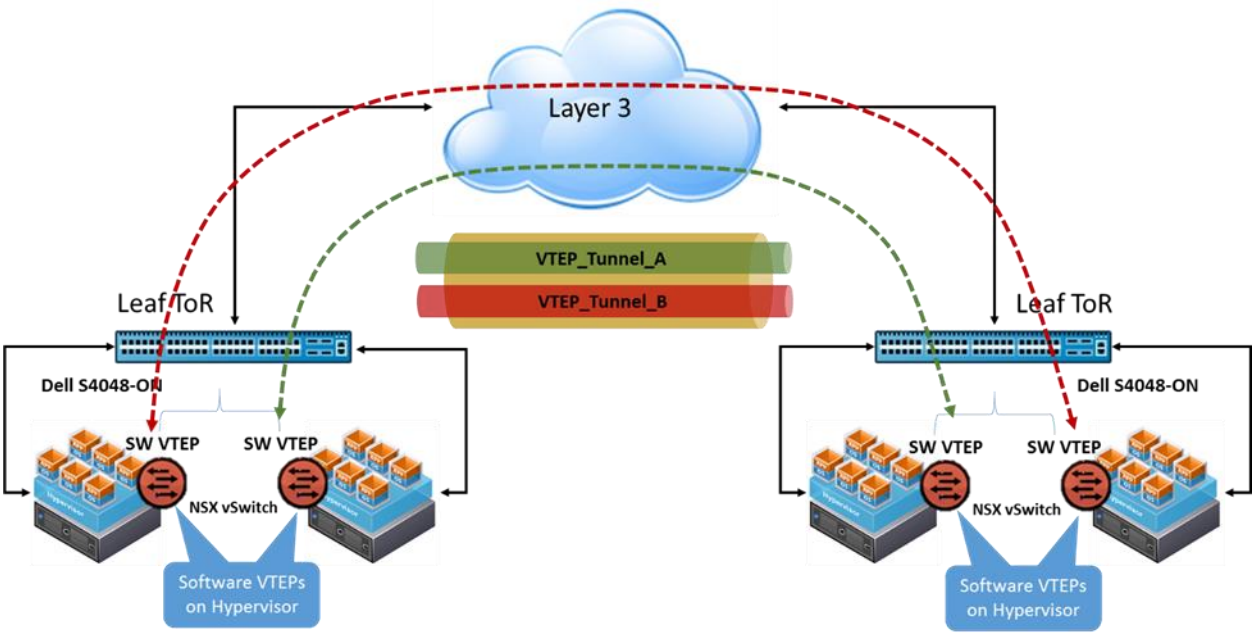


Figure 1. Software-based VTEP implementation and deployment

An important challenge posed by software VTEPs is that, at some point, the logical networks need to communicate with legacy physical networks. Because not all servers will be virtualized in a DC at the same time, virtualized applications need to talk to other non-virutalized applications as well. VTEP hardware bridges that gap.

1.9.2 Hardware VTEP

The solution to the challenges posed by software VTEP has been to develop a hardware VTEP device managed by a NSX controllers. The physical switch will act as a L2 gateway on a ToR switch to connect physical servers/datacenters to VMs in the logical networks.

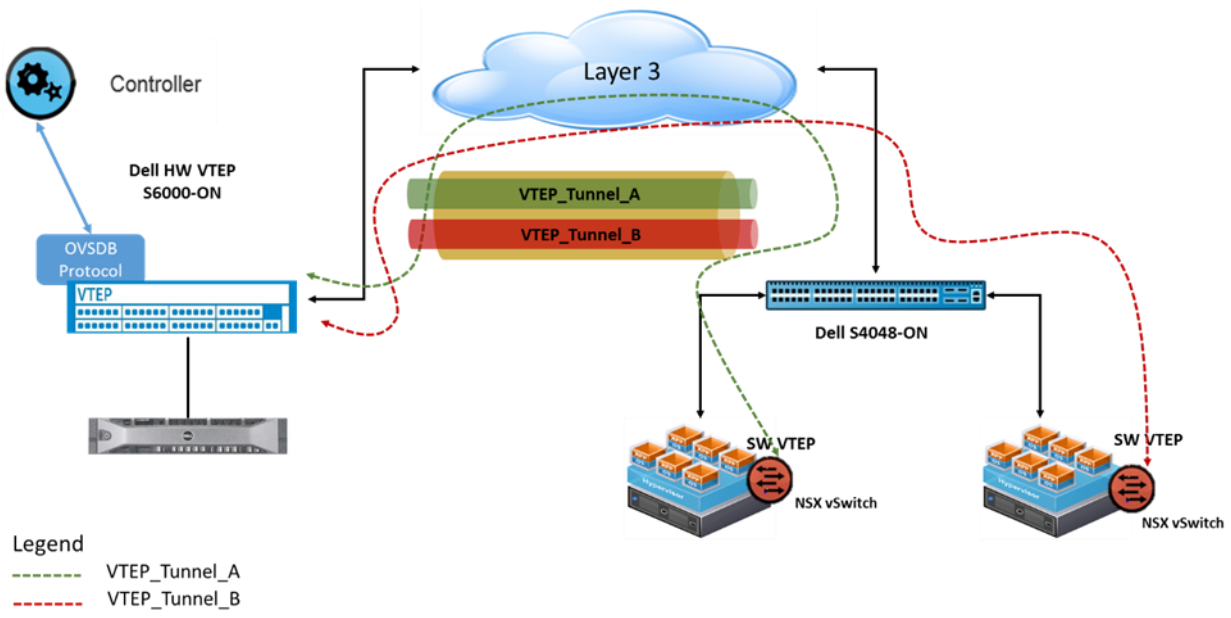


Figure 2. Hardware VTEP (Left side) talking to Software VTEP (Right side)

The hardware VTEP device registers with NSX controllers to establish a connection. Once the registration is successful, NSX controllers then configure the physical ports and VLANs in the hardware VTEP to map them to logical networks.

1.9.3 Hardware VTEP with Dell Network Operating System

DNOS (Dell Network Operating System 9.10) integrates with NSX 6.2.x and provides a Layer2 hardware VTEP gateway for terminating VXLAN tunnels controlled by the NSX controller. DNOS establishes an SSL-encrypted TCP connection to communicate with the controllers via OVSDB protocol. For connection reliability, BFD (Bi-directional Forwarding Detection) is enabled between switch running DNOS and NSX controllers. Once communication is established, NSX controllers configure the DNOS switch to program flow tables to provide connectivity between physical and logical networks. NSX controllers update the MAC table, ARP table and VTEP table for hardware VTEPs, similar to software VTEPs, to provide seamless connectivity between VMs in logical networks to servers in physical networks.

The figure below shows three controllers in a high-availability (HA) cluster:

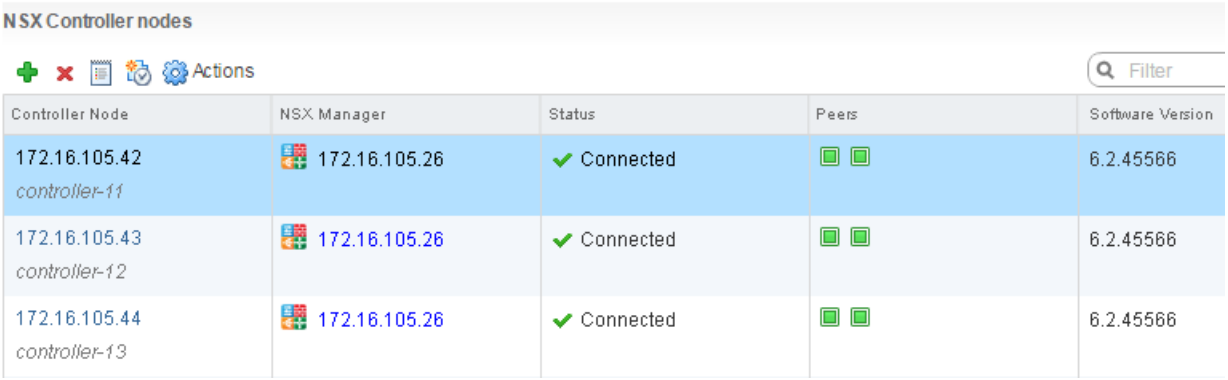


Figure 3. NSX Controller cluster

The following snapshot shows the successful connection of a hardware VTEP connected to the controller cluster.

Table 1. HW VTEP Output

```
VTEP#show vxlan vxlan-instance 1
Instance       : 1
Admin State    : enabled
Management IP  : 172.16.105.33 ←Switch IP on management interface
Gateway IP     : 172.17.6.4 ← Hardware VTEP IP on Loopback interface
MAX Backoff    : 30000
Controller 1   : 172.16.105.42:6640 ssl ←Configured Controller IP
Managers       :
               : 172.16.105.42:6640 ssl (connected)
               : 172.16.105.43:6640 ssl (connected)
               : 172.16.105.44:6640 ssl (connected)
Fail Mode      : secure
Port List      :
 Te 1/50/1  Te 1/50/2 ← Baremetal Server Ports

VTEP#show vxlan vxlan-instance 1 logical-network
Instance       : 1
Total LN count  : 1
  Name                                VNID
  0342c9a7-b544-3083-8e97-27b98f8e3cb7    5011 ← Logical Network ID

VTEP#show vxlan vxlan-instance 1 unicast-mac-local
Total Local Mac Count:    2
VNI           MAC               PORT      VLAN
5011          00:50:56:a8:82:81     Te 1/50/1 100
5014          00:50:56:a8:c1:c3     Te 1/50/2 200


VTEP#show vxlan vxlan-instance 1 unicast-mac-remote
Total Remote Mac Count:    1
VNI           MAC               TUNNEL
5011          00:50:56:92:ec:2b     172.17.1.15
################################################
```

In the above output, the MAC address belongs to the VM that is part of VNID 5011 and IP belongs to the Server Host VTEP this VM resides on.

The hardware VTEP can be configured under the service definitions TAB. All VTEPs are added after a SHA5 certificate.

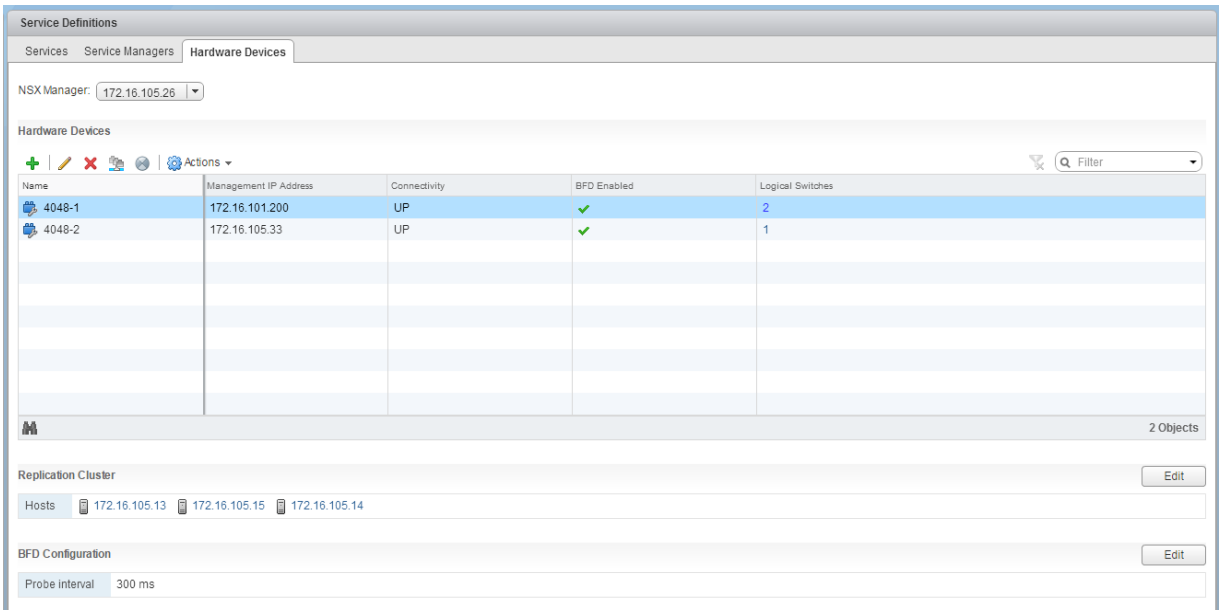


Figure 4. Hardware Device TAB under Service Definition

2 NSX virtual components

In this section, we will discuss the NSX configurations related to VXLAN overlay network.

2.1 Transport zone

A transport zone defines the span of a logical switch/network, and defines a collection of ESXi hosts that can communicate with each other in the physical infrastructure. Communication between ESXi hosts in the underlay happens with VTEP IPs as source and destination. It is important to understand the relationship between VTEPs, VDS, transport zone and logical switch to understand the VMware NSX VXLAN-based overlay networking.

Each ESXi host can be identified by the NSX manager with the help of a unique VTEP IP assigned during the host preparation process of the NSX installation. A VDS is a group of VTEPs and uplink ports, part of a given cluster. VDS can be centrally configured and managed through vCenter networking. A transport zone combines compute and edge VDS, and typically has a logical switch associated with it. Broadcast domain of the L2 logical switch is limited by the scope of the transport zone. By this definition, the scope of a logical switch extends across clusters, typically between compute and edge. Hosts in the management cluster (see Section 4.2.1) never have to be part of the transport zone, as logical networks should not span across management hosts.

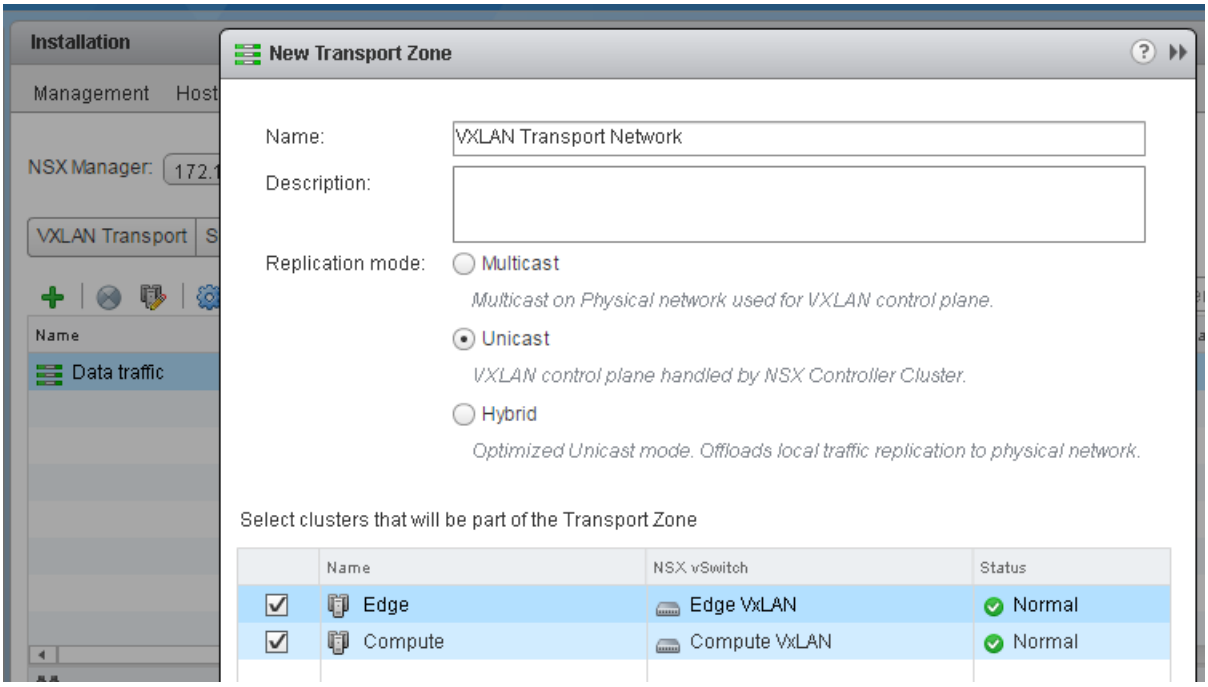


Figure 5 - Transport Zoneconfiguration in NSX Manager

When two VMs connected to different ESXi hosts needs to communicate directly, unicast VXLAN encapsulated traffic is exchanged between VTEP IP addresses associated with the respective ESXi host hypervisors. In some scenarios, traffic originated by a VM may need to be sent to all the other VMs belonging to the same logical networks. This type of traffic is called multi-destination traffic.

In hardware VTEPs to handle BUM traffic, hosts in the compute cluster are configured to act as a replicators. One of the hosts from the configured group will actually do the replication job, while the remaining servers will act as a backup.

The VXLAN standard resolves multi-destination traffic like broadcast, unknown unicast and multicast (BUM) traffic via a multicast-enabled physical underlay network. VMware NSX provides flexibility in how VXLAN replication is handled by the logical switches. It offers 3 control plane modes for handling BUM traffic: multicast, unicast and hybrid. In a non-multicast application environment, it is enough to configure the transport zone in unicast mode to handle ARP/DHCP traffic. Unicast mode helps handle BUM traffic without touching the underlay network configuration.

2.2 Replication modes

When two VMs connected to different ESXi hosts needs to communicate directly, unicast VXLAN encapsulated traffic is exchanged between VTEP IP addresses associated with the respective ESXi host hypervisors. In some scenarios, traffic originated by a VM may need to be sent to all the other VMs belonging to the same logical networks. This type of traffic is called multi-destination traffic.

The VXLAN standard resolves multi-destination traffic like broadcast, unknown unicast and multicast (BUM) traffic via multicast enabled physical underlay network. VMware NSX provides flexibility in how VXLAN replication is handled by the logical switches. It offers 3 control plane modes for handling BUM traffic.

- Multicast
- Unicast
- Hybrid

2.2.1 Multicast mode

In multicast mode, NSX relies on Layer 2/Layer 3 multicast capability of the underlay network to ensure that the VXLAN encapsulated BUM traffic is replicated and reaching all the VTEPs in the given logical switch. Multicast mode is same as the VXLAN RFC way of handling BUM traffic, and does not leverage any of the enhancements brought by the NSX controller cluster.

IGMP snooping should be configured on the physical switches to optimize the delivery of the L2 multicast traffic. To ensure multicast traffic is delivered to VTEPs across different subnets, L3 multicast routing should be enabled and PIM should be configured in the physical switches. Using multicast mode offloads the replication load on the hypervisors.

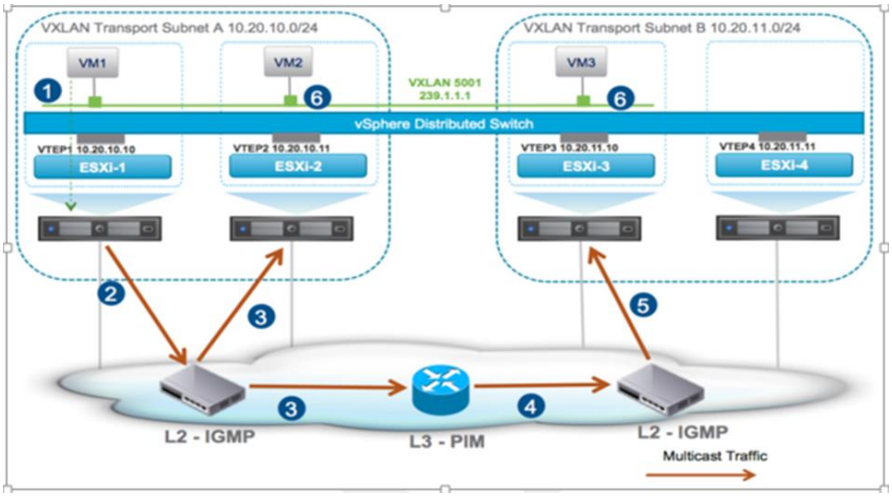


Figure 6. Multicast Mode

2.2.2 Unicast mode

In unicast mode, the multicast support on the underlay network switches is no longer required by VXLAN to handle BUM traffic. In unicast mode, ESXi hosts in a transport zone are divided into separate groups (VTEP segments) based on the IP subnet of the VTEP interfaces. In each segment, a unicast tunnel end point (UTEP) is selected and it is responsible for replicating and sending the traffic to other VTEPs in the given segment. To optimize the replication behavior, traffic to the remote segment is sent only to the remote UTEP, and the remote UTEP in the given segment is responsible for replicating the traffic in that segment.

The benefit of this mode is that no network configuration is required on the physical underlay switches, as all the BUM traffic is replicated locally on the host and sent via unicast packets to respective VTEPs.

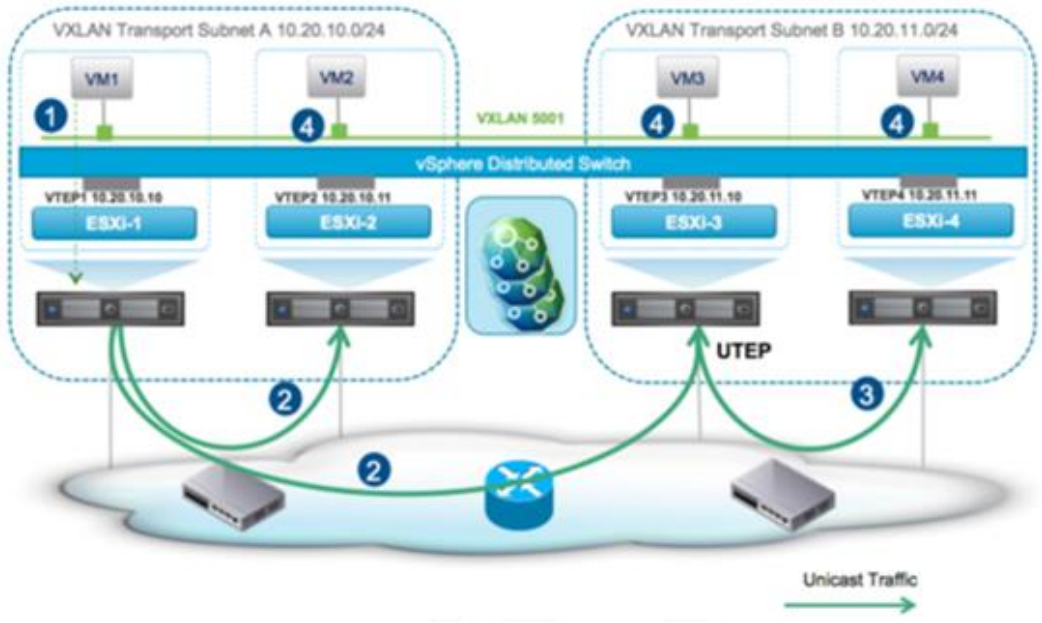


Figure 7. Unicast Mode

2.2.3 Hybrid mode

Hybrid mode is operationally similar to unicast mode. IP multicast routing configuration is not required in the physical network; however, hybrid mode leverages the L2 multicast capability of the physical network. Similar to unicast mode, in hybrid mode a multicast tunnel end point (MTEP) is created in each VTEP segment. Traffic in the same VTEP segment is replicated by the IGMP snooping-enabled physical switch. Traffic to the hosts in the remote VTEP segment are sent to the remote MTEP and the remote MTEP uses the physical switch in the remote segment to replicate and send packets to all the VTEPs. If the traffic has to be sent to multiple remote segments, the MTEP at the source VTEP segment is responsible for replicating and sending the packets to each of the remote MTEPs.

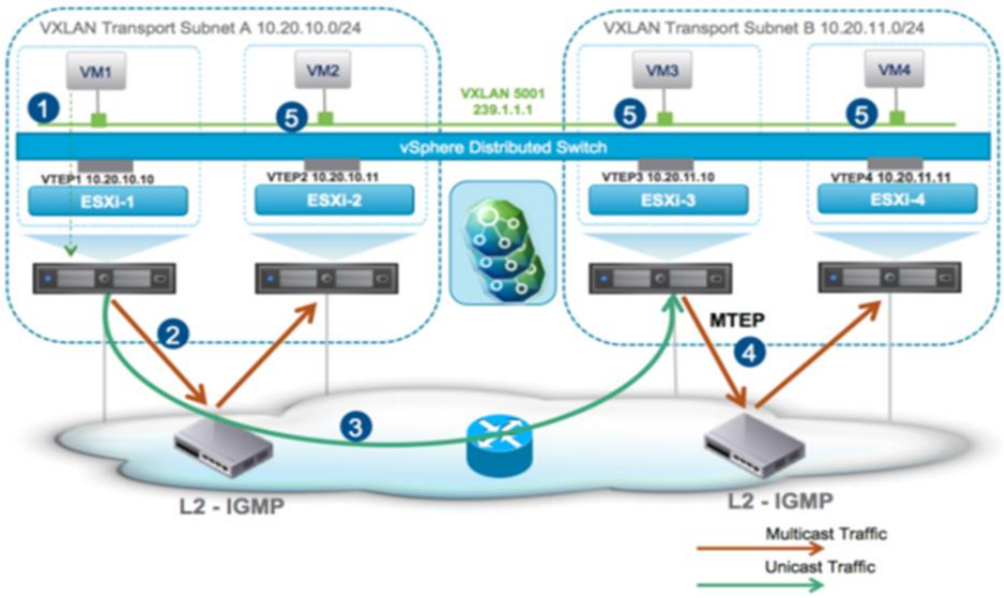


Figure 8. Hybrid mode

Hybrid mode helps offload the replication of ESXi hosts in each VTEP segment and simplifies the physical network configuration by relying only on L2 multicast configurations such as IGMP. This helps effectively scale the VXLAN BUM traffic in a large scale design without overloading the ESXi servers or increasing the complexity of the physical underlay networking with L3 multicast configurations.

2.3 Logical switching

VMware NSX provides isolated L2 logical switching capability with the help of the construct known as logical switches. Logical switches help connect VMs located across different ESXi hosts and situated across different VTEP segments in the given cluster. Each logical switch is assigned a VXLAN Network ID (VNID), similar to a VLAN ID. Any packet sent by the VM part of a logical switch will be sent out with MAC-in-UDP encapsulated with VXLAN header. This abstraction helps to decouple the logical network (i.e. VXLAN encapsulated overlay network) from the underlay network.

There are two types of logical switches:

- Local – this type of logical switch is deployed within a single vCenter/single NSX domain. The limitation of this deployment is the limited scope of the all the resources to a single vCenter/ NSX manager.
- Universal – to address the limitation imposed by a single vCenter/ NSX manager, the universal switch/router was introduced. With a universal logical switch or router, the switch or router can span multiple vCenter domains.

For this reference architecture, multi-site architecture is not used; therefore, nor is Universal Logical Switching.

2.4 Logical routing

Logical routing in VMware NSX provides the routing capabilities to interconnect different logical switches, as well as to interconnect logical networks with the physical networks. The logical routing can be performed independently of the physical underlay networking because the overlay networking is totally decoupled from the underlay and managed by NSX controllers using VXLAN.

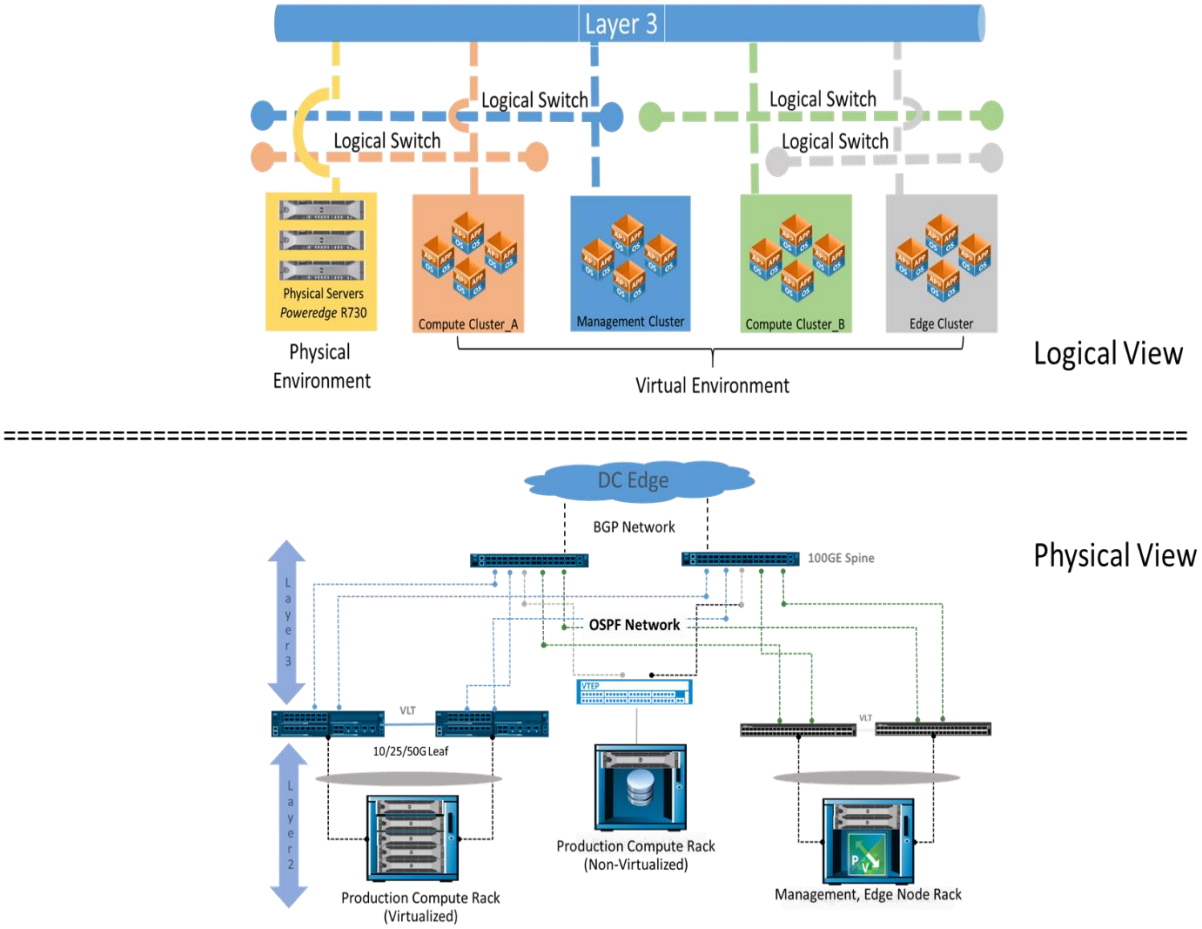


Figure 9 - Logical Routingand switching

Traffic in a data center can be broadly classified as east-west traffic and north-south traffic. Network traffic within the data centers and within logical switches or across different tiers is considered east-west traffic. Traffic to and from the outside world and the data center is considered north-south traffic. VMware NSX provides two distinct appliances (VMs) to cater to these types of traffic. A distributed logical router (DLR) is a control-plane VM that enables routing capabilities to the ESXi host VTEPs to send traffic between logical switches. An edge services gateway (ESG) is a data-plane appliance that acts like a L3 hop between the physical and logical networks. Both the DLR and the ESG support OSPF and BGP routing protocols to enable connectivity by exchanging route information between them or to the outside world.

2.4.1 Distributed routing with DLR

As described, the DLR is only a control plane VM. The DLR installs VIBs across ESXi hosts to push routing information (RIB) through NSX controllers. The DLR enables ESXi hosts to route traffic across logical switches without sending traffic all the way to the ESG in the edge cluster. By making distributed routing decisions within the individual hosts, the DLR helps avoid hairpinning of traffic across compute and edge clusters during east-west communication between the application network tiers.

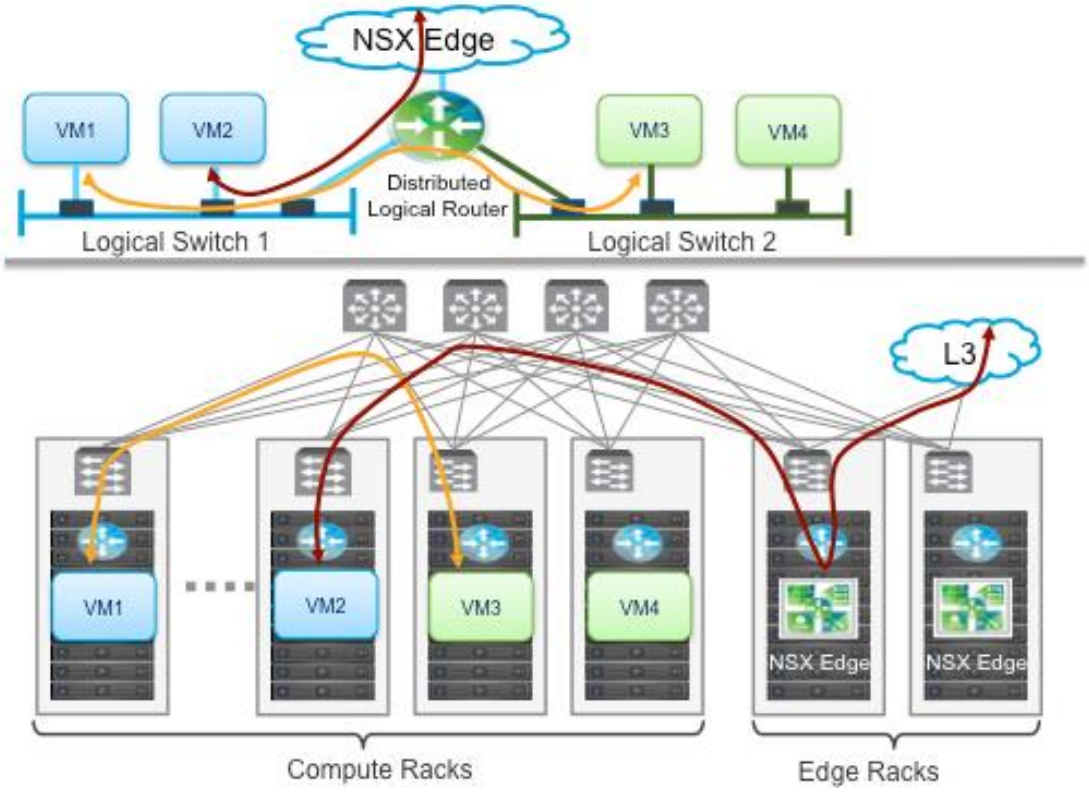


Figure 10 - Distributed Logical Router

2.4.2 Centralized routing with ESG

Logical switches can be directly connected to the ESG to enable both east-west and north-south communication. The ESG is a data plane virtual appliance performing routing operations. In modern data center networks, the east-west traffic ratio is much higher than the north-south traffic ratio. If the edge cluster is deployed without the DLR, the ESG becomes a core router, responsible for all routing decisions. This potentially creates bottlenecks in the network for traffic in both east-west and north-south directions. To avoid this, centralized routing for traffic in both directions should be separated out. The ESG should be made responsible for connecting the physical network to the logical network using routing protocols.

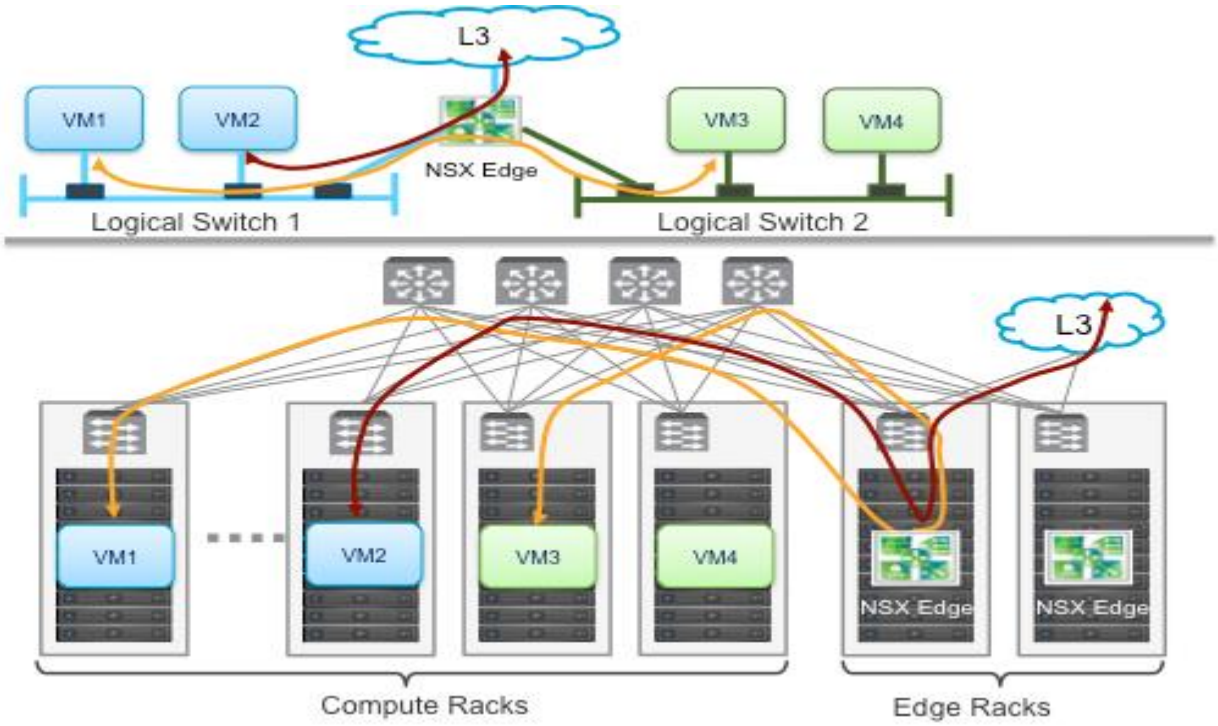


Figure 11 - Edge Services Gateway

In addition, the ESG also supports other networking such as NAT, firewall, load balancing and VPN services to the logical network.

2.4.2.1 HA Mode – Stateful Edge Services

This is a redundancy model where a pair of ESG appliances are deployed with one ESG in active state and a second ESG in standby state. Keepalives (KA) packets are exchanged every second between active and standby using an internal dedicated link. Missing 15 heartbeats will result in standby to transition to active state. The default value of 15 can be reduced up to 6 heartbeats to improve downtime. However, the VMware recommended safe lower limit for heartbeat failure is 9 seconds. The standby ESG appliance continuously exchanges information with the active ESG to synchronize NAT, firewall, load balancer and interface configuration. The standby ESG will take over the active ESG with all the synchronized information whenever it detects an active ESG failure.

While deploying the ESG in HA mode, it is important to consider that the DLR control VMs are also running in HA mode with a similar kind of heartbeat exchange between active-standby control VMs. If both the active ESG and the active DLR control VMs reside in a single physical host of an edge cluster, a failure to the host machine will lead to double failure of DLR and ESG. This kind of failure will bring down routing protocol adjacencies between the VMs, resulting in traffic outage for longer durations. It is recommended to follow the anti-affinity rule to place the ESG and DLR in different physical hosts to avoid double failures.

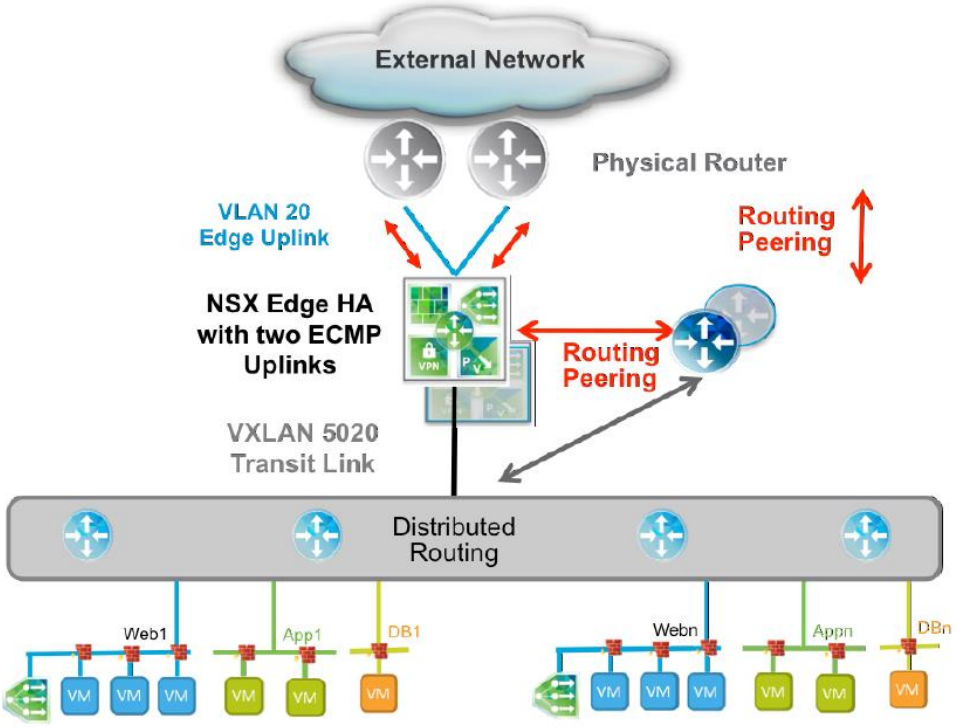


Figure 12 - ESG in HA Mode

2.4.2.2 ECMP Mode – better throughput

VMware NSX 6.2 or later supports ECMP capabilities in the edge appliance. ESG can support ECMP up to 8 edge appliances. When ESG appliances are deployed in ECMP, there are two important advantages: first, increased throughput capacity for north-south traffic with more ESG deployed in the edge cluster (upper limit: 8); and second, reduced downtime, because traffic flowing through a failed ESG will automatically get routed to other active ESGs in the edge cluster. To maintain minimum availability, it is necessary to deploy at least 2 ESG VMs in the cluster with anti-affinity rule enforced.

Due to the increased throughput capability and resiliency offered in ECMP design for ESG, it is recommended to deploy VMware NSX 6.2 or later this way compared to HA mode. With ECMP, there is a high likelihood of asymmetric traffic flow. To avoid traffic black-holing, stateful services like firewalls, if needed, should be disabled or pushed down one level to a tenant specific ESG.

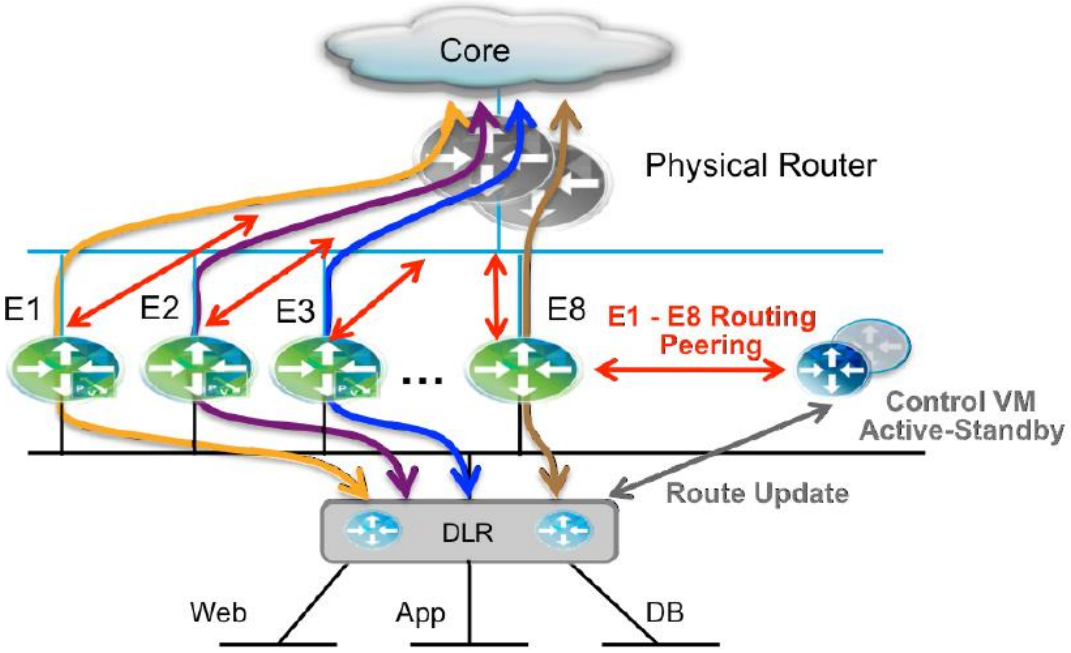


Figure 13 - ESG in ECMP Mode

3 Getting started

The Dell networking underlay provides a high-performance, highly scalable, and non-blocking architecture. This dynamic underlay is based on the highly distributed Clos architecture, based on defining a two layer (Spine and Leaf) switching architecture that provides full non-blocking switching fabric, where the leaf and spine switches are interconnected but not connected within the layers.

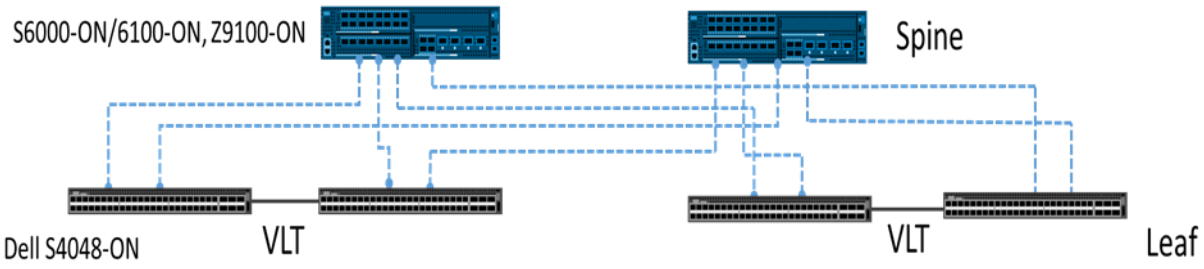


Figure 14 A simple Spine-Leaf architecture with HA

In order to build a reference architecture that meets the needs and limitations of each customer, design options for each component discussed here should be carefully studied. There are two primary requirements for VMware NSX.

1. IP based Underlay
2. Jumbo MTU support

An IP-based underlay is needed to establish end-to-end logical network communication across different host machines. To send traffic across VMs in a logical network or to resolve ARP of the

destination VM, the ESXi host looks up its VTEP table to find the destination VTEP IP, then encapsulates the host VTEP as source IP and the destination host VTEP as destination IP. If a match is not found, the ESXi communicates with the NSX controller to find the destination VTEP IP to encapsulate the packet.

Network virtualization using VXLAN is achieved by encapsulating the original L2 frame coming from the VM with the VXLAN header and underlay IP header, adding up to 50 bytes of additional information. Assuming the default MTU in the VMs to be 1500 and adding the 50 byte VXLAN encapsulation, there is a chance that a switch without jumbo frame support might drop the packet. Underlay network design should take this additional data encapsulation into consideration and configure accordingly.

3.1 Dell Underlay Network design options

The ability of new network technologies to succeed lies in their ability to co-exist and inter-operate with legacy network infrastructure. VMware NSX using VXLAN technology provides greater visibility and ease of management without increasing complexity or compromising flexibility. A DNOS-based networking underlay for VMware NSX, based on legacy network design and long-term application and network growth considerations, can be deployed in the following ways:

1. Pure L2 underlay networks using VLT
2. Leaf-Spine Hybrid underlay with mix of L2 and L3 networks

3.1.1 Pure Layer 2 with HA – VLT (Virtual Link Trunk)

By decoupling the application network provisioning using the VXLAN overlay from the physical network infrastructure, the underlay design becomes simple, easy to maintain and scalable. ESXi hosts need to be provisioned for management, storage, vMotion, VXLAN guest data traffic (VTEPs) and the external network. Based on need and availability, a set of ports separate from ESXi hosts can be provisioned to handle each type of traffic as well.

At the leaf level, the ESXi host's NIC ports are connected to the VLT pair leaf switches. The ESXi host can be configured to form bonding using NIC teaming or LAG. To enable NIC redundancy and efficient operation, it is recommended to configure static or LACP LAG from the server to leaf switches to form VLT.

Multiple uplinks from the VLT pair of leaf switches across racks can be connected to create an aggregate VLT pair of spine switches to provide connectivity across different racks. By configuring L2 spine switches with VLT to eliminate loops, full uplink bandwidth from the switches can be utilized without blocking any of the uplink ports. VLANs configured at the leaf level need to be configured at the spine and advertised to routing protocols to provide reachability to the underlay networks as well as to the edge gateway.

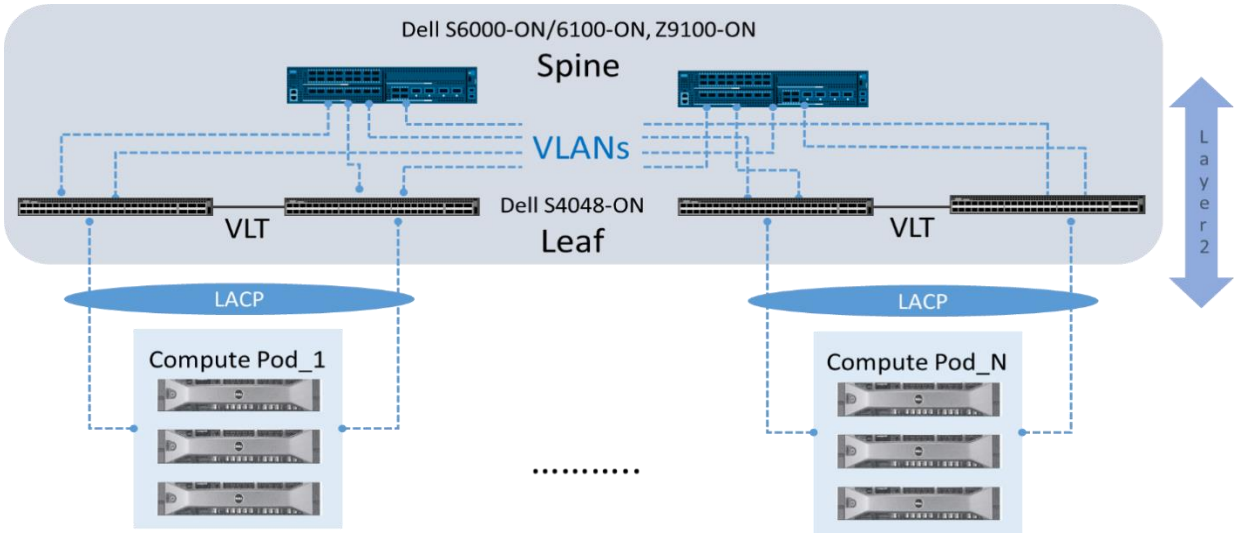


Figure 15 - Dell Networking Underlay in Layer-2 mode using VLT for redundancy

3.1.2 Hybrid – Layer 3 and Layer 2

Modern data centers can deploy L3 CLOS architecture with a mix of leaf-spine switches with oversubscription ranging from 1:1 to N:1, based on the network needs. OSPF or EBGP can be used as a L3 protocol of choice to provide connectivity between VTEPs. An underlay network using L3 CLOS architecture provides non-blocking connection between VTEPs across different racks.

For small-scale data centers, the ease of configuration of OSPF may make it the preferred option over BGP. All leaf-spine devices can be configured to operate under the single OSPF area 0. In slightly larger deployment scenarios, spine switches can be configured to operate in Area 0, with leaf switches from each TOR in different areas to limit the LSA flooding to individual racks.

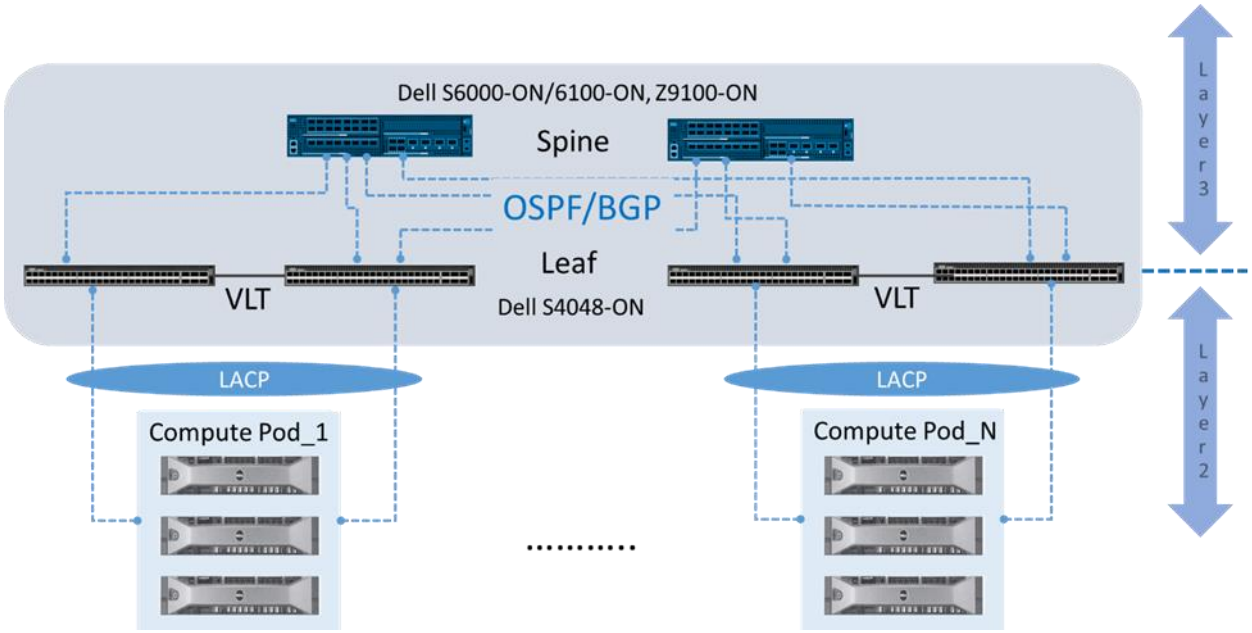


Figure 16 Routed Network underlay

With EBGP as the routing protocol between leaf-spine, L3 CLOS architecture can scale to a considerably higher degree for large data center environments. EBGP is capable of handling big routing tables with incremental updates and, with AS path attribute built-in loop prevention, offers a solution for data center scalability needs. With EBGP L3 CLOS architecture, all the spine nodes will be configured to be part of single AS. The leaf nodes from each rack will be assigned to a different unique AS number.

Similar to the VLT L2 underlay design, leaf nodes can be configured to work in VLT mode to provide network switch level redundancy. Hosts in each rack can be configured to be part of a cluster, and VTEPs from each host will be advertised to BGP from the leaf node by the particular rack VTEP subnet.

NSX, being an overlay technology, uses an underlay foundation which it builds upon to deliver its ultimate benefit. The Dell networking foundation is simple, scalable, and easy to manage.

As discussed earlier, Dell networking switches fully support IP based underlay and can support jumbo frames up to 9252 bytes at line-rate. Dell networking switches can be used to create a VLT based L2 or Hybrid underlay. In this RA, we will use OSPF in underlay to enable VTEP reachability.

3.1.3 Underlay Network Isolation with VRF

In L3 CLOS architecture using the DNOS VRF feature, network segmentation between tenants can be achieved to create to complete network isolation in the underlay network. By using VRF in the leaf switches during interconnection of the physical and logical networks with ESG, each VRF in the underlay should be mapped to a tenant-specific ESG to achieve total isolation. Using OSPF-MP with VRF in the leaf switches will help exchange logical network information of each tenant independently. With the VRF-lite option, multiple-tenant networks can use the same IP address scheme without worrying about duplication between the tenants. For more information on how to configure underlay using VRF, please refer to this link.

3.2 NSX virtualization platform deployment

When building a new VMware NSX based environment, it is essential to choose an architecture that allows for future growth. The reference architecture we choose to deploy should give the flexibility to grow horizontally without affecting overall design as the network starts to scale out.

To achieve modular growth, it is necessary to logically separate and group the ESXi hosts providing separate functions. We can broadly classify the role of different resource functions as follows:

- **Compute Cluster**
- **Edge Cluster**
- **Management Cluster**

This logical separation offers the flexibility to grow individual logical functions as per the network growth needs. In smaller deployments, management and edge functions can sometimes be consolidated into a single cluster. The Dell Server portfolio provides various rack and blade server models to run different application workloads as per the application need. R430/R630/R730 servers can be used to build these clusters.

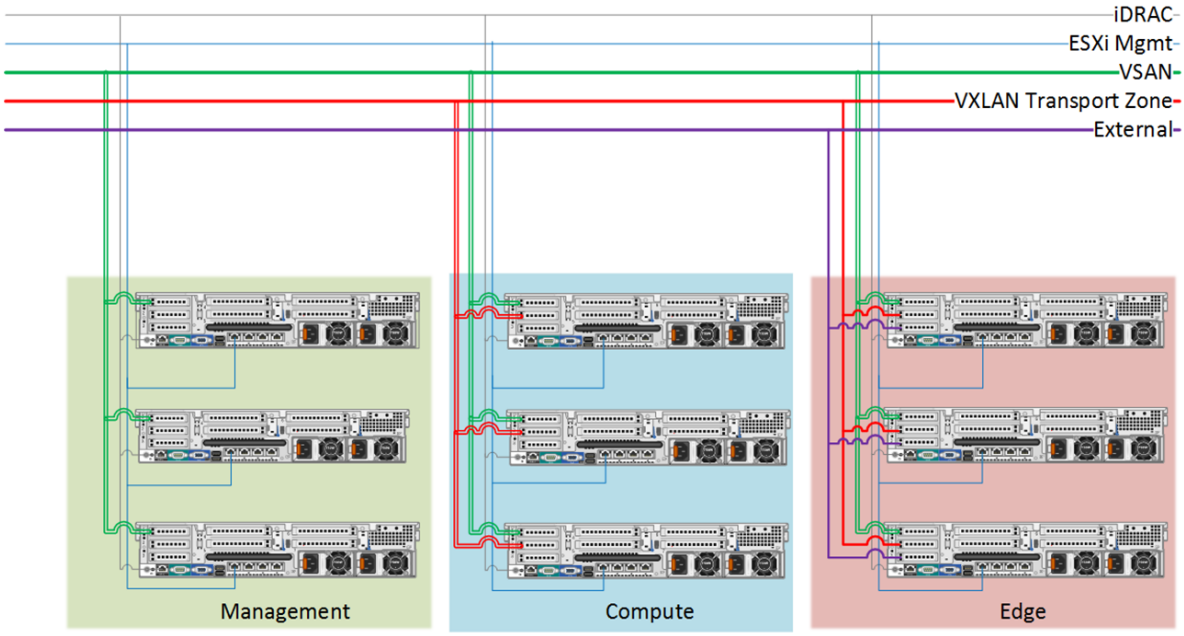


Figure 17 Cluster types and underlay VLAN network

In the above picture, each server has two network adapters with 2x10G interfaces each. The iDRAC interface on each server is used for server management. The ESXi management network uses the LOM 1G ports. The management network is an out-of-band network running mostly 1G interfaces.

3.2.1 Management cluster

The management resource cluster typically hosts the infrastructure components that are required to manage ESXi environment in production. Typical management resource components include the external platform services controller (EPSC), vCenter, NSX manager, vRealize Log Insights, vRealize Operations manager, and other shared components. Compute and memory requirements to host these components for required scale are pre-identified and hosts in this cluster are provisioned accordingly. ESXi hosts provisioned in this cluster separately without consolidating the edge cluster do not need to provision VXLAN network. So the ESXi hosts connecting to the leaf can be provisioned only for Management, vMotion and Storage.

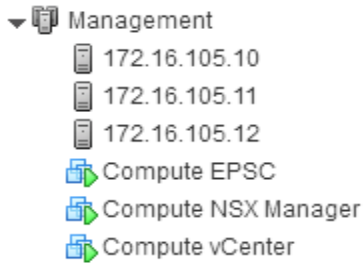


Figure 18 - Management Cluster – vCenter screenshot

3.2.2 Compute cluster

The compute resource cluster consists of hosts in which the tenant virtual machine running application workloads are hosted. The VMs are connected to the logical networks, independent of the physical infrastructure using VXLAN overlay. Compute resource clusters based on scalability needs can span across multiple racks. Application VMs hosted in a compute cluster can grow or shrink dynamically based on the workload requirements. Communication requirements of the VMs in compute resource cluster include the need to communicate with other VMs in the same logical network or different logical network for east-west communication. Furthermore, VMs in logical networks may need to communicate with outside world via an edge cluster. To accommodate these requirements, ESXi hosts connecting to the leaf switches need to be provisioned for management, vMotion, storage and VXLAN data networks.

VMs in logical networks may need to communicate with legacy hosts in the compute cluster as well. Hardware VTEP-enabled switches should be configured to enable communication between them.

Multiple application clusters can be hosted in isolated fashion within a single compute cluster as well.

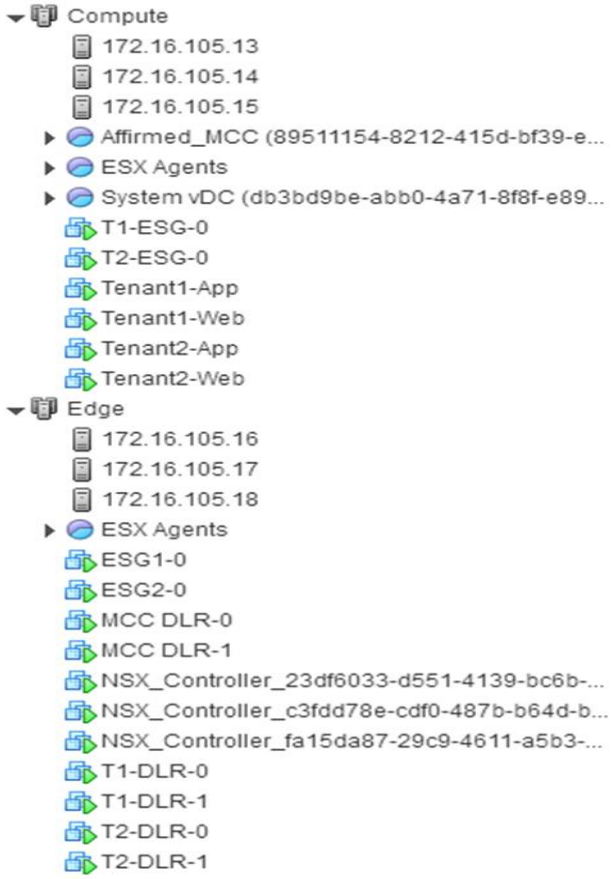


Figure 19- Resource Clusters (Compute, Edge) – vCenter screenshot

3.2.3 Edge cluster

The main function provided by an edge resource cluster is that it acts like a bridge between VXLAN-enabled overlay networks and physical world. The VM providing this functionality is the edge services gateway. The ESG acts like a L3 hop between logical networks and physical networks. The ESG establishes routing protocol adjacencies between the physical world and itself. The ESG also learns the logical network information by forming an adjacency with the distributed logical router (DLR). The ESG supports ECMP functionality, so as a best practice it is recommended to deploy at least 2 ESGs in the edge cluster to provide redundancy.

To accommodate the connectivity requirements of the edge resource cluster, we need to provision management, vMotion, storage and VXLAN data networks. ESG enables connectivity to the VMs in a logical network either by advertising logical networks through a routing protocol or by using NAT and using the underlay IP network scheme. It is recommended to configure a separate set of ports towards WAN or underlay network to enable north-south communications.

The edge resource cluster also hosts the DLR control VM. The DLR control VM helps avoid hairpinning of traffic between east-west traffic across logical networks to the edge cluster by enabling distributed logical routing functionality in the ESXi hypervisors. The DLR is a control VM only, so it does not participate in forwarding any data traffic packets like ESG. The DLR also communicates with hosts via the management network, so a separate network for DLR to ESXi communication must be provisioned. For a given application there can be only one DLR. To avoid a single point of failure, DLRs are typically deployed in HA mode.

The table below summarizes the network configuration requirements for each type of resource cluster. These network types are nothing but VMKernel adapters, part of a single (uplinks are shared) or multiple distributed vSwitch with vmknic IP addresses provisioned for each the respective port groups.

Table 2 Underlay VLAN networks for each cluster

Network type	**Resource cluster type**		
	Management	Compute	Edge
Management	Yes	Yes	Yes
vMotion	Yes	Yes	Yes
Storage	Yes	Yes	Yes
VXLAN	No	Yes	Yes
External	No	No	Yes

3.3 vSphere networking

Configuring ESXi host networking is an important aspect of NSX design. It ensures connectivity across hosts, hardware gateways, storage and the capability to handle broadcast, unknown-unicast and

multicast traffic also known as BUM. The ESXi hosts have physical NICs configured to be part of different VDS to perform the following tasks.

I. Management VDS
II. VSAN VDS
III. VXLAN VDS

3.3.1 Management VDS

To manage the ESXi hosts using OOB management, management VDS is created. Management IP address of each ESXi host is the VMkernel adapter's IP address. This IP address is made part of the management VDS. To provide necessary isolation for management traffic from storage & data traffic, we recommend creating this separate VDS for each cluster to manage the ESXi hosts over a typically 1G interface.

3.3.2 VMware Virtual SAN (VSAN) VDS

To consolidate host datastores to a single VSAN cluster and to enable storage services like vMotion, Fault tolerance, replication, etc., across hosts in a given cluster, a pair of dedicated 10G ports can be used to create a VSAN VDS. VMware Virtual SAN uses disk groups to pool together SSD and SAS/SATA hard disks in each server to create a consolidated datastore. This datastore is presented to the ESXi hosts in each cluster as a single shared location for storage. VSAN needs to synchronize host datastores across location that could be in different racks. For this synchronization it is recommended to use Layer-2 multicast in the network. In this RA, we have used IGMP in the network to build multicast for the VSAN VLAN connectivity across hosts.

VSAN VDS uplink ports are configured in a NIC team with explicit failover mode to handle link failure scenarios. Two VMKernel adapter IPs are created in the VDS and the adapters are associated with each of the uplink interfaces as primary interface. One NIC is dedicated to enable VSAN communication while the second one is configured to handle the rest of the storage related services.

To enable layer-2 multicast in the Leaf switches, enable IGMP snooping globally. This configuration enables IGMP in all the VLANs of the switch. If a particular VLAN need not have to participate in IGMP, it could be disabled explicitly. Each VSAN cluster should be configured in separate VLANs to avoid unnecessary traffic from a different cluster reaching ESXi hosts in one cluster.

3.3.3 VXLAN VDS

For application data traffic (i.e., for east-west and north-south communication of VMs within the data center and outside the data center), VXLAN VDS is configured with a pair of dedicated 10 ports. Instead of separate VDS for storage and data traffic, a converged data center could also be created with a single VDS. In this reference architecture, we built two VDS to handle storage and data traffic on separate physical ports.

The transport zone is a collection of VDS ports carrying application data traffic. This design shows separate VDS to handle different types of traffic, though it is enough to configure VXLAN VDS alone to be part of the transport zone. A 3-Tier web architecture does not require multicast in the underlying network. We take advantage of the transport zone unicast mode offered by NSX to handle BUM traffic with no multicast configuration required on the underlay network.

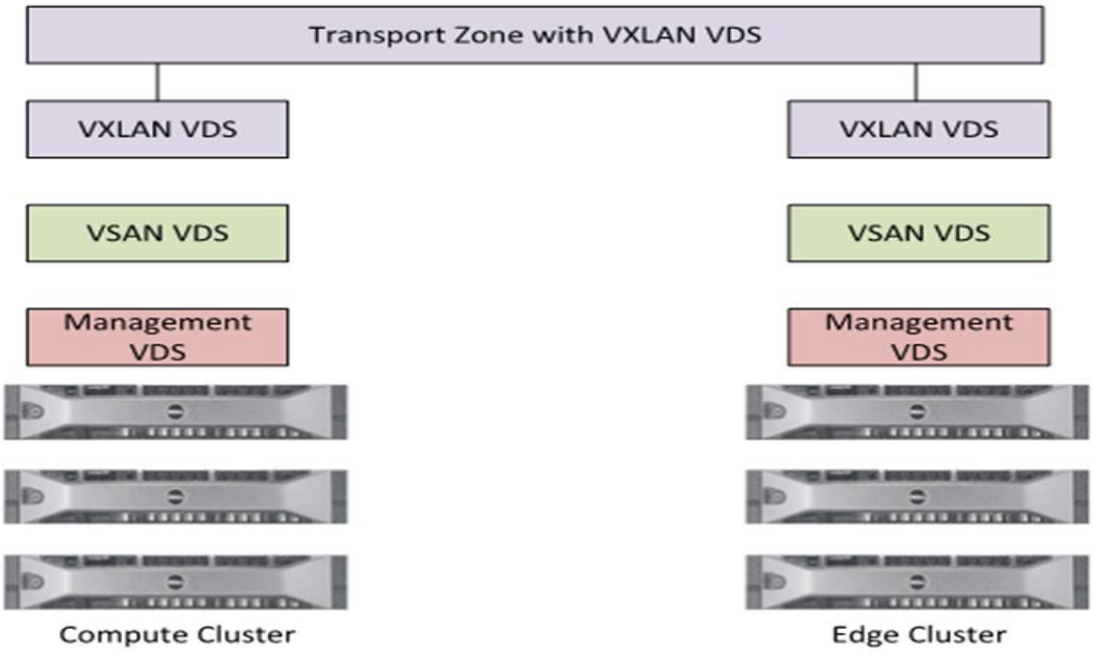


Figure 20 - VDS & Tranport Zone

3.4 Overlay network design

VMware NSX uses the VXLAN overlay to decouple the application network provisioning and simplifies underlay networking in the physical switches. Therefore, there is no need to touch any physical devices to provision these networks. In this reference architecture, we can take a simple example of deploying a web based application deployed for two different tenants.

A classic 3-tier web application reference architecture requires three different network segments: web, application and database tiers. VMs running in each tier should have the ability to communicate between VMs in the same tier and across tiers as per the application need.

Typically, in a 3-tier web based application, only the web network must be routable to the outside world (that is, requiring north-south communication). Only the web tier subnet must be advertisted through DLR via routing protocols. Application and database traffic are confined within the data center and predominantly contribute to east-west traffic. Apart from switching within the tier, routing between the tiers within the data center is also a very common requirement. This can be achieved by associating the networks with DLR without advertising them through routing protocols to provide the necessary isolation.

3.4.1 Creating logical network tiers

In the chosen 3-tier architecture (web, app and database) for each tenant, we need to create three logical switches for each tier. Apart from this, we have to create two more logical switches in each tenant. The reason for creating these two logical switches will be discussed in the subsequent section. One switch is for providing the heartbeat between the DLRs depoyed in HA mode. Another logical switch is for providing connectivity between the tenant DLR and the tenant-specific ESG. Finally, another logical switch provides connectivity between the tenant ESG and perimeter ESG.

Tenant1-App	Normal	Data traffic	Global	5010	Unicast
Tenant1-DB	Normal	Data traffic	Global	5011	Unicast
Tenant1-DLR_HA	Normal	Data traffic	Global	5015	Unicast
Tenant1-Transit	Normal	Data traffic	Global	5017	Unicast
Tenant1-Web	Normal	Data traffic	Global	5009	Unicast
Tenant2-App	Normal	Data traffic	Global	5013	Unicast
Tenant2-DB	Normal	Data traffic	Global	5014	Unicast
Tenant2-DLR_HA	Normal	Data traffic	Global	5016	Unicast
Tenant2-Transit	Normal	Data traffic	Global	5018	Unicast
Tenant2-Web	Normal	Data traffic	Global	5012	Unicast
Tenant Edge Transit	Normal	Data traffic	Global	5008	Unicast

Figure 21 – Logical Switch screenshot

The drawing below shows the logical network of Tenant1. Compare the previous picture with logical switches and locate them in the picture below for Tenant1.

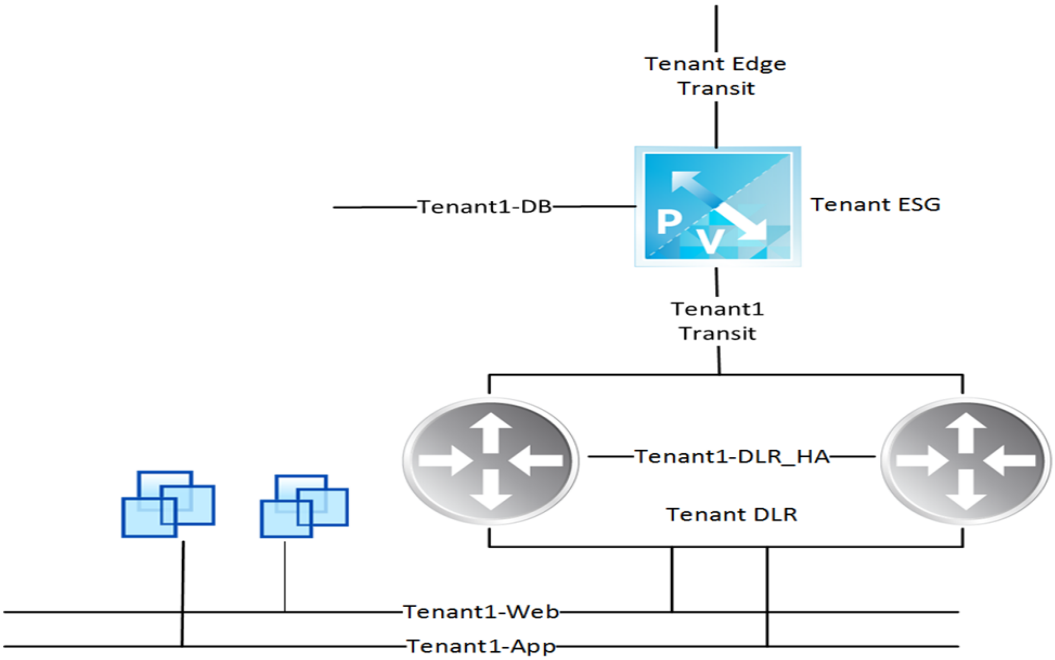


Figure 22 Tenant1 logical network

3.4.2 Enabling routing between tiers

In order to provide a multi-tenant environment with the necessary isolation, logical routers like DLR, tenant ESG and perimeter ESG need to configured with a routing protocol. DLR and ESG support both OSPF and BGP as routing protocol options. To maintain uniformity and a single protocol to administer, in this RA we will choose EBGP to implement between the logical routers similar to underlay networks. The following reference architecture is designed to avoid a single point of failure, protect the logical routers from the node failures, and to provide maximum throughput and the option of future scalability.

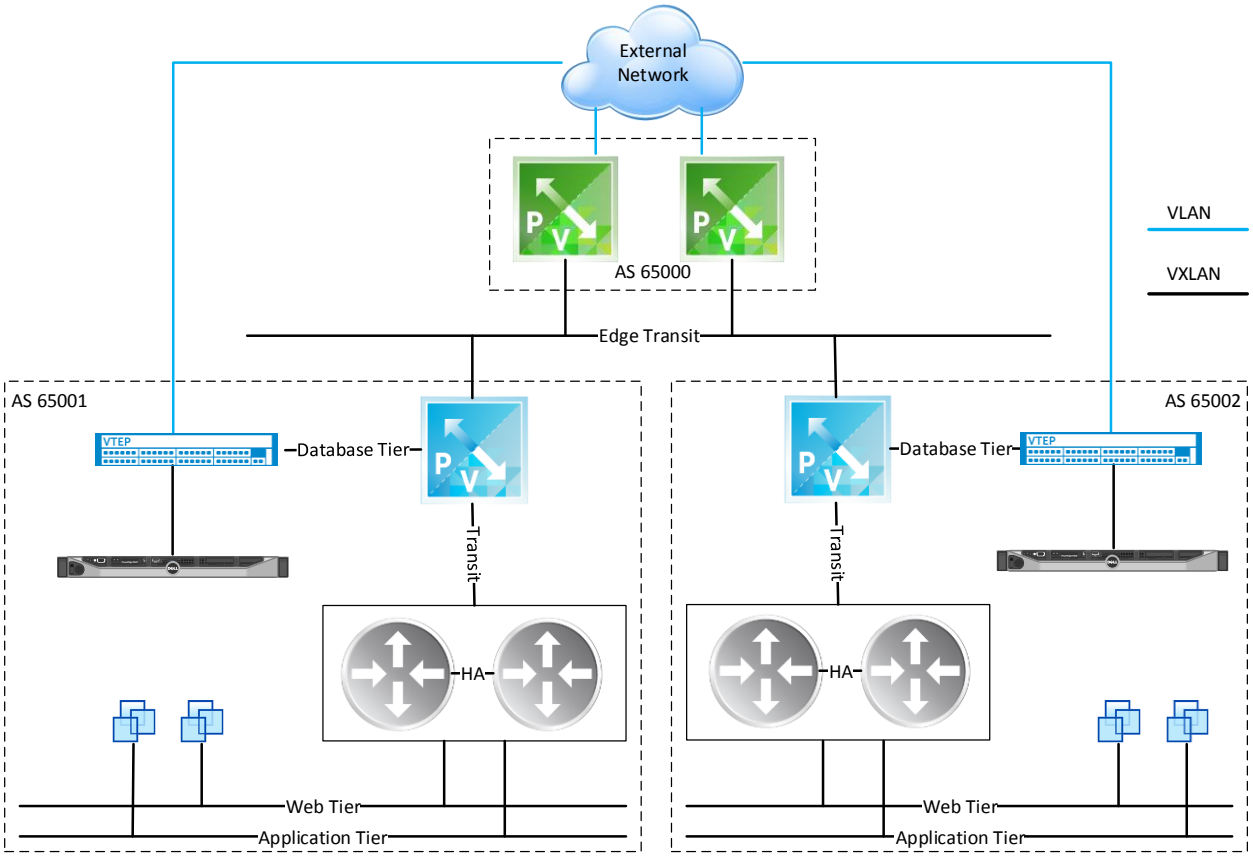


Figure 23 Full logical network view

Different logical routers deployed in the system are:

1. DLR in HA Mode
2. Tenant ESG as a standalone appliance
3. Perimeter ESG in ECMP mode

3.4.2.1 Distributed Logical Router (DLR)

The DLR, described earlier, is only a control VM that enables ESXi hosts to route between the logical networks. It does not pass any traffic through the appliance. To protect the control VM from physical host failure, the DLR is typically installed in HA mode. To ensure connectivity between the active-standby link, a dedicated logical network is configured to synchronize data. DLR can be placed in the edge cluster since it is a edge cluster related service. The DLR interface IP is configured as the gateway IP to the VMs connected to the respective logical swtich.

3.4.2.2 Tenant Edge Services Gateway

A tenant ESG is needed to perform several important operations. Tenant ESGs help bridge non-virtualized servers to logical networks using a Layer 2 gateway, typically database servers. Stateful firewall at per-tenant level, NAT and load-balancing can be performed for each tenant at the tenant ESG. The tenant ESG establishes an IBGP connection with the DLR to learn tenant logical networks. A dedicated logical network is configured between the tenant ESG and the DLR to exchange routing information. The tenant ESG uses EBGP to advertise the tenant network information to outside via perimeter ESG. Using route redistribution, only the web network is advertised to the outside world, providing necessary

isolation to internal tenant networks. To avoid traffic hairpinning with L2 gateway for east-west traffic, the tenant ESG is placed in the compute cluster

For a multi-tier architectecture that does not require multi-tenancy, a per-tenant level ESG appliance is not required. In this case, the hardware VTEP would directly connect to the perimeter ESG. DLRs would also directly connect to the perimeter ESG to provide external network connectivity.

3.4.2.3 Perimeter Edge Services Gateway

The perimeter ESG acts like a aggregate router for multiple tenant ESGs. Tenant ESGs communicate with the perimeter ESG using a dedicated logical network. To provide maximum throughput for north-south traffic, multiple perimeter ESGs can be deployed in ECMP mode. Up to 8 ECMP ESGs are supported by VMware NSX.

The perimeter ESG acts like a L3 hop between the logical networks and physical networks. Perimeter ESGs establish routing adjacencies with physical routers to exchange tenant VMs' reachability information to the outside world, as well as providing reachability to the virtual infrastructure. If the routing protocols configured in the logical network and physical network are different, ESGs support route redistribution between routing protocols to exchange logical network information and vice versa.

3.5 Hardware VTEP Creation and configuration

Bridging the connectivity gap between the virtualized and non-virtualized environments is key to the overall network virtualization effort. In order to do this, a device capable of encapsulating and decapsulating VXLAN headers is needed, and a hw VTEP gateway does this.

Once the virtual environment (VDS, ESG, DLRs, logical switches, etc.) has been configured the hardware VTEP must be added to the network overlay through NSX manager.

3.5.1 Generate a certificate file on the hw VTEP device

Create a certificate file using *'crypto cert generate'* command. We could use the show file command to see the contents of the certificate file. The public key generated with this command is copied over to NSX service definition while adding the device to authenticate and establish a secure channel of communication between NSX and the HW VTEP switch

DNOS#crypto cert generate cert-file **flash://vtep-cert.pem** key-file flash://vtep-privkey.pem

```
DNOS#show file vtep-cert.pem
-----BEGIN CERTIFICATE-----
MIIDkDCCAnigAwIBAgIBYzANBgkqhkiG9w0BAQUFADB6MQswCQYDVQQGEwJVUzEa
MBgGA1UEAwwRY29uZG9yLW1nbXQtcsQwNDgxDTALBgNVBAoMBER1bGwxGDAWBgNV
BAsMD0R1bGwgTmV0d29ya2luZzERMA8GA1UEBwwIU0FOIEpvc2UxEzARBgNVBAgM
CkNhbGlmb3JuaWEwHhcNMTYwNDA3MDkyMjEyWhcNMjYwNDA1MDkyMjEyWjB6MQsw
CQYDVQQGEwJVUzEaMBgGA1UEAwwRY29uZG9yLW1nbXQtcsQwNDgxDTALBgNVBAoM
BER1bGwxGDAWBgNVBAsMD0R1bGwgTmV0d29ya2luZzERMA8GA1UEBwwIU0FOIEpv
c2UxEzARBgNVBAgMCkNhbGlmb3JuaWEwggEiMA0GCSqGSIb3DQEBAQUAA4IBDwAw
ggEKAoIBAQDY5tjxtbykEA5NpIhA92ctA0vA3/icFVWEKE80MQqU8u4h1KfrZ2of
GRTQj8apigoLhIYUIP5hhuFxW6KXac7Nm3G8TFme1HYa4K+3df+XVPCPArFTigqf
m6IwZGvNIxgmz4fWP8szKuCVscO+QTc8NoDf+Z23KgX1aCUh/+/eew9ir53ItGX8
iz23ZD9AB8tVH33+MZHdHtOQNbGQY23hAgMBAAGjITAfMB0GA1UdDgQWBBTaOaPu
XmtLDTJVv++VYBiQr9gHCTANBgkqhkiG9w0BAQUFAAOCAQEAz8DtyPgz2ozdUvcg
m9tIBGjLi6IAjBpt3L912c8GNCeeJOmWXz+ZAqj/4kQpjyrgdymh086JrwF7N/Te
JavHnyOMYXPpENCjTfAqAkzPncnHZUG8335R1VPQ8VqR2k0PJdG1b5TuGIs1HQUd
7qqybaK9/6vKRMbY8vMoQa14T0BFvCA7pr0sq4OrsTBKK3YGME3yOADDEpWPWrCq
P1U+JXzK6X30fToh+Kwvp128FCbfA7pkRENdhYOgKmpf777xZUutkJvgnZUPj5j0
YXyHpzWgszNxHrDTAHHIEs8V0An5HQ0UhdwXlcYm82KDwZswlnWwQ29MNg1Zjp8Q
8y22Vw==
-----END CERTIFICATE-----
```

3.5.2 Configure VXLAN feature and instance on HW VTEP device

Once the certificate has been generated, it will be used by the NSX platform in order for the HW VTEP to be part of the network overlay. Prior to adding the HW VTEP GW to the NSX platform, VXLAN feature must be enabled on the HW VTEP and a VXLAN instance must be configured.

```
Switch# conf
Switch(conf)# feature vxlan                          #Enable VXLAN feature
Switch(conf)# vxlan-instance 1                       #Configure VXLAN instance. 1 instance supported
Switch(conf-vxlan-instance-1)# gateway-ip <IP>     #VXLAN HW VTEP IP (loopback ideally)
Switch(conf-vxlan-instance-1)# fail-mode secure   #Specify mode on loss of connection
Switch(conf-vxlan-instance-1)# controller 1 <controller_IP> port 6640 ssl
Switch(conf-vxlan-instance-1)# no shut
Switch(conf-vxlan-instance-1)# end
Switch#conf
Switch(conf)#int range te1/50/1 , te1/50/2
Switch(conf-int-range-te-1/50/1-150/2)#vxlan-instance 1       #1/50/1-2 part of vxlan-instance 1
```

Currently one vxlan instance is supported with additional instances to be supported in future OS releases. Also, two controllers can be supported for redundancy purposes.

The gateway ip address parameter should be any reachable IP address defined or configured on the HW VTEP device. It is recommended to use a loopback interface with an ip address as the gateway.

Once the above configuration has been applied on the HW VTEP switch, IP connectivity between the HW VTEP and controller VM on the NSX platform should be up. If this connection does not come up the HW VTEP switch will be added to the network overlay.

3.5.3 Add the newly created certificate to NSX manager

Navigate to **NSX → Service Definitions → Hardware Devices** tabs to add the hardware VTEP device to the network overlay. Once the certificate has been added, connectivity between the hardware VTEP device and controllers should come up. Figure 24 shows the dialog window on NSX manager when adding the hardware VTEP certificate.

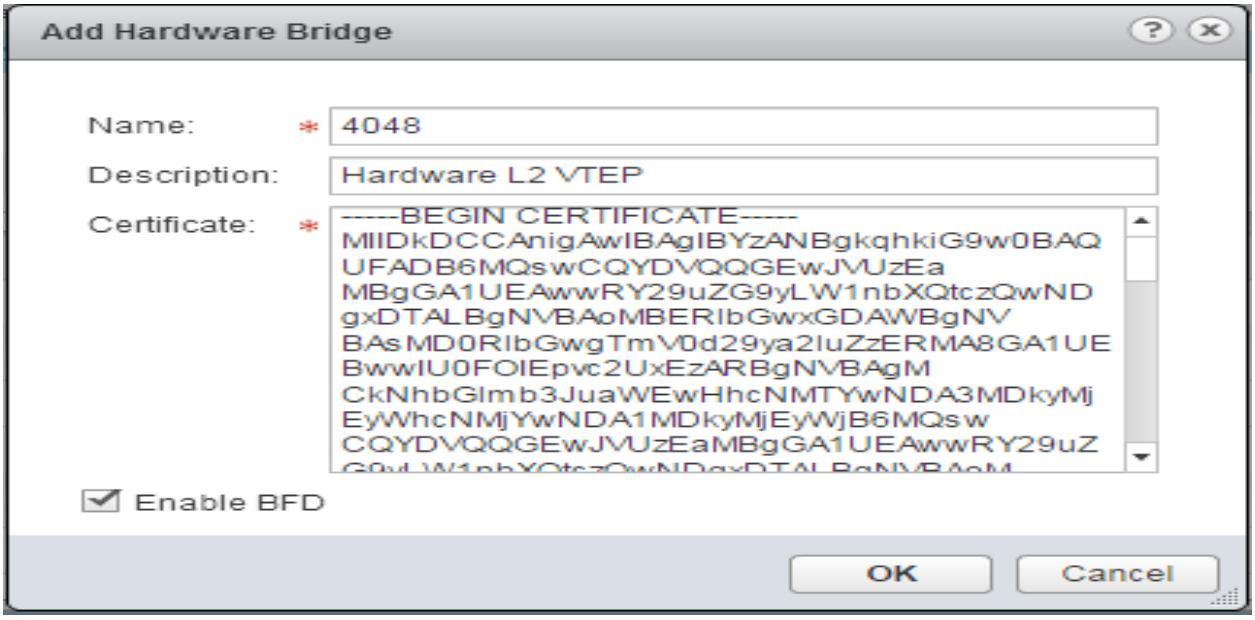


Figure 24 : HW VTEP Certificate addition on NSX manager

If the certificate copied to the NSX and switch certificate match, the authentication will succeed to create a secure OVSDB communication channel between NSX and HW VTEP switch**.** Figure 25 shows the status of the HW VTEP switch once the certificate has been added to the NSX platform and connectivity between the HW VTEP and controller(s) has been established.

Figure 25 : HW VTEP status

3.5.4 Adding HW VTEP ports to the Logical Switch

In this step, the ports 1/150/1 and 1/50/2 part of the HW VTEP switch are now configured to be part of the logical switches created in section 3.4.1. There are several logical switches or Tiers (App, DB, Web)

Ports 1/50/1 and 1/50/2 are configured to be part of logical switch *Database*. Following are the steps.

Step 1. NSX → Service Definitions → Hardware Devices. Click on "Add logical switch" (see circle in red)

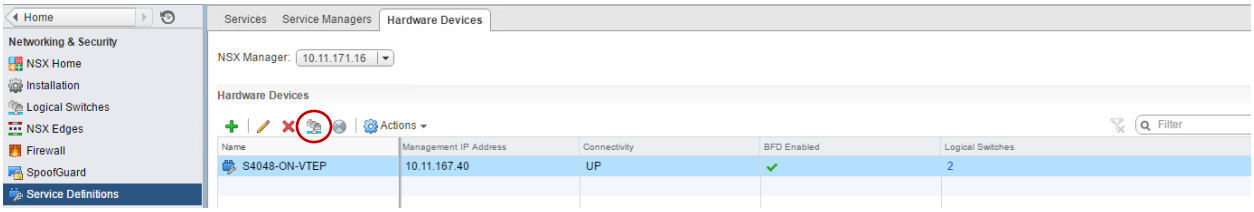
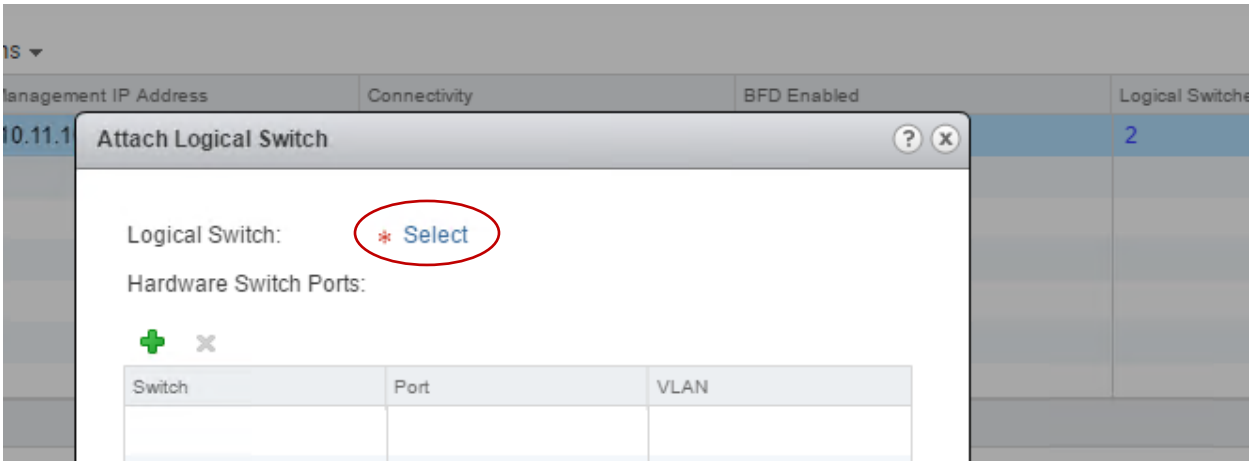


Figure 26: Adding logical switch to HW VTEP GW

Step 2. Next choose the logical switch to attach to. In this case, the logical switch is the *Database* logical switch. It this at this stage where the network virtualization configuration is taking place. Database VMs part of the logical switch (Database) connect with the physical database server connected to the HW VTEP GW.



Upon clicking on the "Select" button, a new window showing the available logical switches will come up. Click on the logical switch and a new window will come up with the selected logical switch.

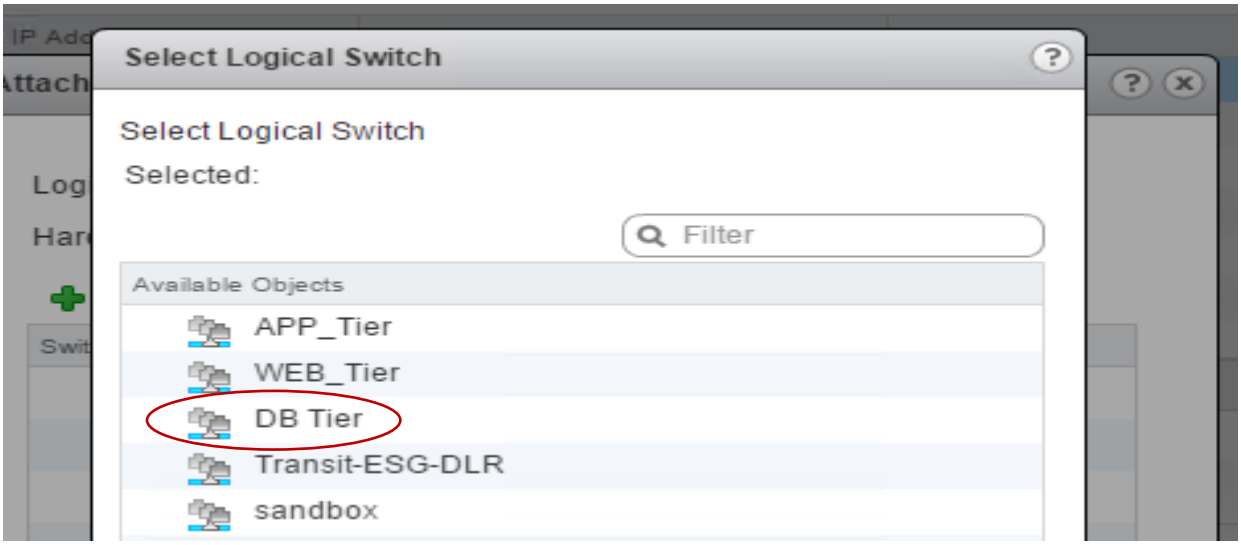
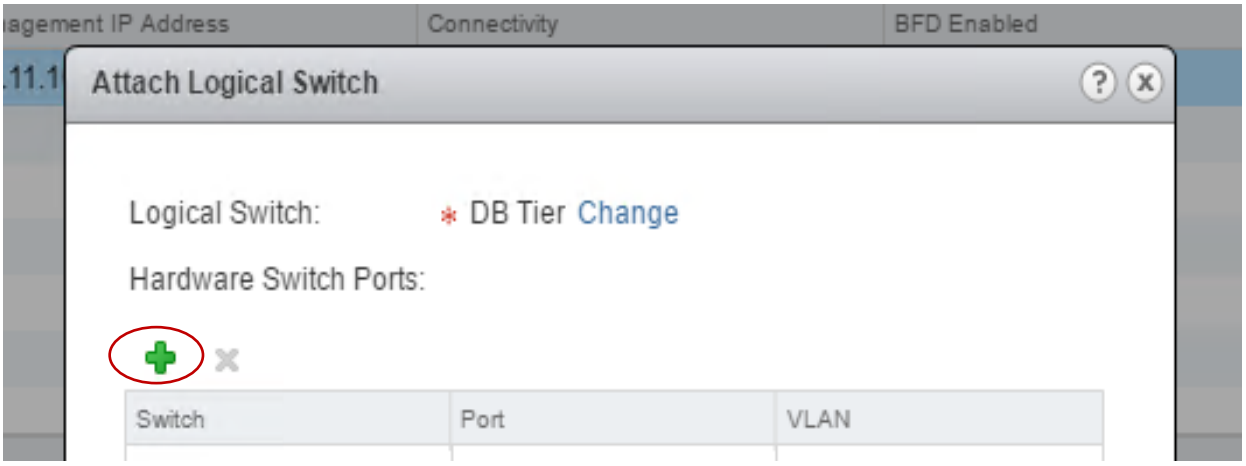


Figure 27: Selecting logical switch and attach to HW VTEP GW



Click on the "Plus" sign to add the switch ports connected to the physical database server to the logical switch and thus configuring network virtualization.

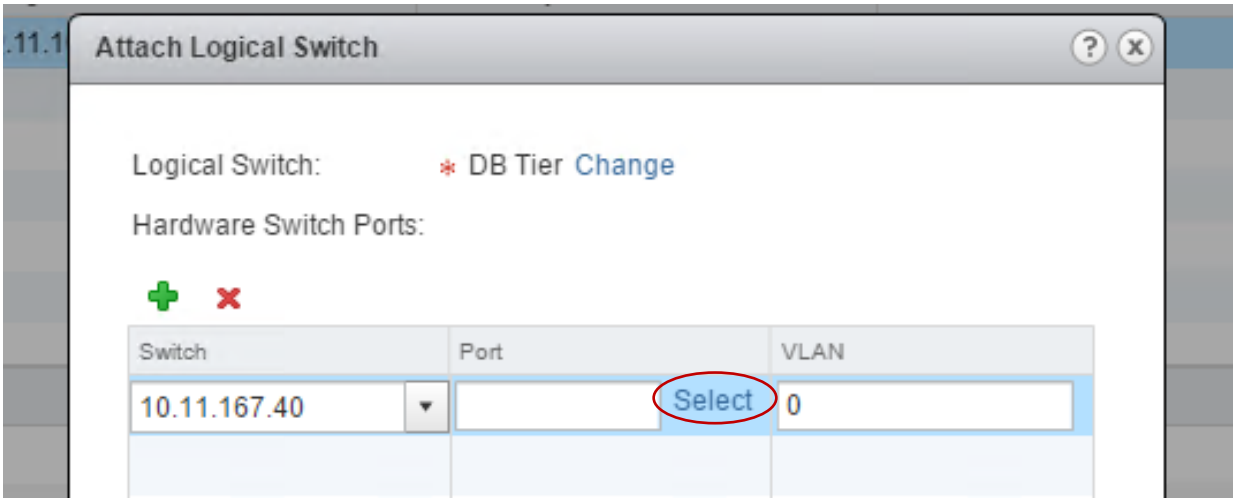
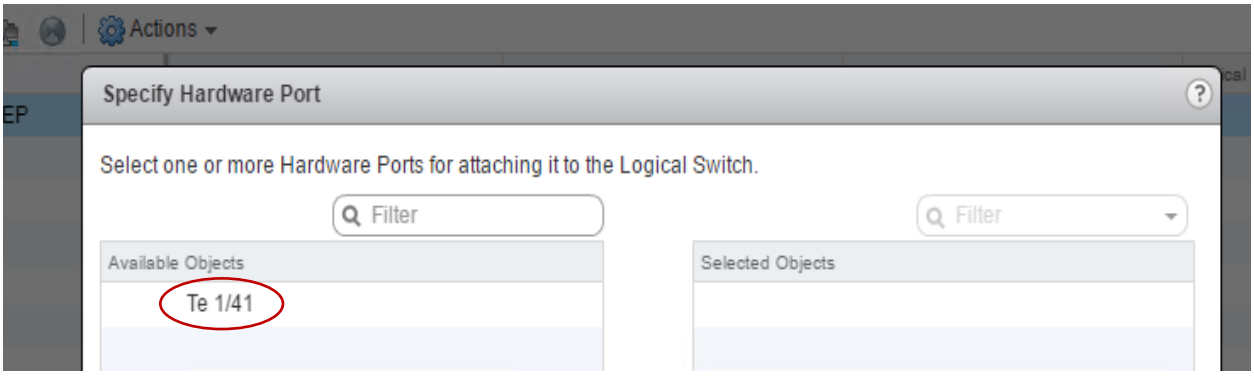


Figure 28: Attaching the HW VTEP GW switchport(s) to the logical switch

The ip address shown is the management ip address of the HW VTEP GW. Click on the"Select" button.



The switchports connected to the physical server are then attached to the logical switch "Database".

3.6 Summary of Dell end-to-end underlay infrastructure

Figure 29 shows the infrastructure (Dell end-to-end) used to create this handbook. Notice, although figure 29 does not show the second HW VTEP GW, the sample configuration section captures the configuration of both HW VTEP GWs.

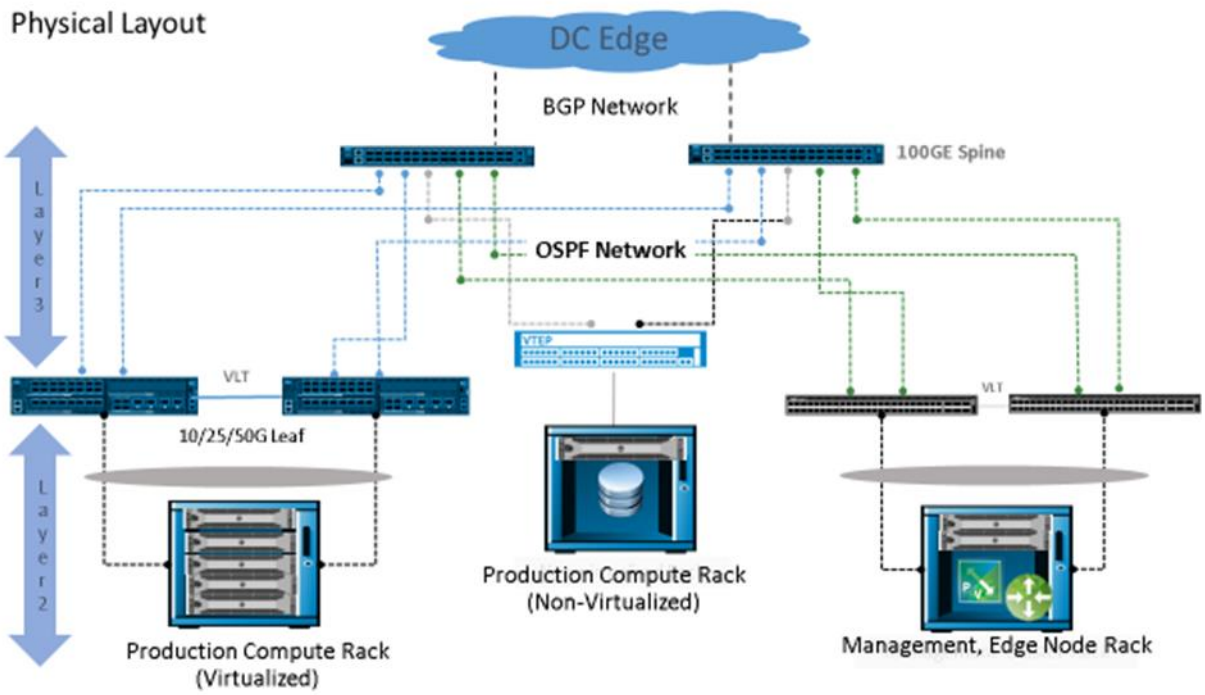


Figure 29 Full view of underlay network

A quick summary of Dell-VMware NSX reference architecture:

- The 3-tier multi-tenant reference architecture is deployed on top of Dell end-to-end infrastructure using Dell networking switches and Dell PowerEdge R730 servers.
- PowerEdge R730 rack servers to provide management, compute and edge cluster resources.
- The physical network fabric is deployed in Layer 3 CLOS architecture using S4048-ON and S6000-ON switches to create a Layer 3 underlay network.
- The S4048-ON and S6000-ON switches help bridge logical networks using Layer 2 HW VTEP gateway functionality connected to bare metal servers.
- Virtual Link Trunk (VLT) is used at the S4048-ON leaf layer to provide active-active NIC connectivity and leaf switch redundancy as well as providing uplink ECMP to spine switches.
- OSPF is used to enable VTEP connectivity across racks through leaf-spine switches.
- Using the Dell networking content pack for vRealize, logs of the phsycial switches and servers can be monitored from a single application.
- Three different logical clusters are created to perform management, compute and edge operations.
- The transport zone is configured to operate in unicast mode to handle BUM traffic.
- Multiple logical switches are created to host multi-tier network application.
- Per-tenant DLR and ESG are implemented to enable routing between tiers in each tenant.

- Combination of IBGP and EBGP protocols are used to advertise routes to the perimeter ESG.
- BGP routing policies are configured to provide necessary isolation to internal network tiers.
- Micro-segmentation rules are configured to provide security for east-west communication.
- Bare metal servers are bridged to logical networks using L2 hardware gateway functionality.

4 Conclusion

This handbook provided a basic introduction to VMware components and overlay network concepts related to VMware NSX as well as how to use Dell networking switches and servers to deploy a VMware NSX based network virtualization platform. With Dell networking providing a solid physical network underlay and a Layer 2 gateway to bridge logical networks, this guide demonstrated a multi-tenant multi-tier network running VMware NSX-vSphere network virtualization technology. The unique innovation VMware provides with NSX for network virtualization combined with Dell servers and NSX certified switches for a fully operational non-blocking underlay, provides businesses the power to harness the flexibility of virtualization and to seamlessly interoperate with legacy networks.

Appendix

Sample Configuration

Leaf1	Leaf2
leaf1-s6000#show run int vlan 3000 ! interface Vlan 3000 description VSAN ip address 172.17.0.7/24 untagged Port-channel 10-18 ! vrrp-group 1 virtual-address 172.17.0.1 no shutdown leaf1-s6000#show run int vlan 3001 ! interface Vlan 3001 description VxLAN Data ip address 172.17.1.7/24 mtu 9000 tagged Port-channel 20-28 ! vrrp-group 1 virtual-address 172.17.1.1 no shutdown leaf1-s6000#show run int vlan 3002 ! interface Vlan 3002 description To Spine Switch ip address 172.17.2.7/24 tagged Port-channel 1 no shutdown	leaf2-s6000#show run int vlan 3000 ! interface Vlan 3000 description VSAN ip address 172.17.0.8/24 untagged Port-channel 10-18 ! vrrp-group 1 virtual-address 172.17.0.1 no shutdown leaf2-s6000#show run int vlan 3001 ! interface Vlan 3001 description VxLAN Data ip address 172.17.1.8/24 mtu 9000 tagged Port-channel 20-28 ! vrrp-group 1 virtual-address 172.17.1.1 no shutdown leaf2-s6000#show run int vlan 3002 ! interface Vlan 3002 description To Spine Switch ip address 172.17.2.8/24 tagged Port-channel 1 no shutdown

leaf1-s6000# leaf1-s6000#show run ospf ! router ospf 1 network 172.17.2.0/24 area 0 network 172.17.1.0/24 area 0 passive-interface Vlan 3001 default-information originate redistribute static leaf1-s6000#	leaf2-s6000# leaf2-s6000#show run ospf ! router ospf 1 network 172.17.2.0/24 area 0 network 172.17.1.0/24 area 0 passive-interface Vlan 3001 default-information originate redistribute static leaf2-s6000#
Spine1	Spine2
spine1-s6000#show run int vlan 3002 ! interface Vlan 3002 description To Leaf Switch ip address 172.17.2.5/24 tagged Port-channel 1,3 no shutdown spine1-s6000#show run int fo 0/124 ! interface fortyGigE 0/124 description To S4048-VTEP1 ip address 172.17.3.5/24 no shutdown spine1-s6000#show run ospf ! router ospf 1 network 172.17.2.0/24 area 0 network 172.17.3.0/24 area 0 spine1-s6000#	spine2-s6000#show run int vlan 3002 ! interface Vlan 3002 description To Leaf Switch ip address 172.17.2.6/24 tagged Port-channel 1,3 no shutdown spine2-s6000#show run int fo 0/124 ! interface fortyGigE 0/124 description To S4048-VTEP2 ip address 172.17.4.5/24 no shutdown spine2-s6000#show run ospf ! router ospf 1 network 172.17.2.0/24 area 0 network 172.17.4.0/24 area 0 spine2-s6000#
HW VTEP1	HW VTEP2
S4048-VTEP1#show run vxlan ! feature vxlan ! vxlan-instance 1 gateway-ip 172.17.5.4 fail-mode secure controller 1 172.16.105.42 port 6640 ssl no shutdown ! interface TenGigabitEthernet 1/50/1 vxlan-instance 1 no ip address no shutdown ! interface TenGigabitEthernet 1/50/2 vxlan-instance 1 no ip address no shutdown ! interface TenGigabitEthernet 1/50/3 vxlan-instance 1	S4048-VTEP2#show run vxlan ! feature vxlan ! vxlan-instance 1 gateway-ip 172.17.6.4 fail-mode secure controller 1 172.16.105.42 port 6640 ssl no shutdown ! interface TenGigabitEthernet 1/50/1 vxlan-instance 1 no ip address no shutdown ! interface TenGigabitEthernet 1/50/2 vxlan-instance 1 no ip address no shutdown ! interface TenGigabitEthernet 1/50/3 vxlan-instance 1

```
 no ip address
no shutdown
!
interface TenGigabitEthernet 1/50/4
 vxlan-instance 1
 no ip address
 no shutdown
S4048-VTEP1#show run int fo 1/49
!
interface fortyGigE 1/49
 description To Spine1
 ip address 172.17.3.4/24
 no shutdown
S4048-VTEP1#show run int loop 0
!
interface Loopback 0
 description HW VTEP IP
 ip address 172.17.5.4/24
 no shutdown
S4048-VTEP1#show run ospf
!
router ospf 1
 network 172.17.3.0/24 area 0
 network 172.17.5.0/24 area 0
S4048-TOR-1#show run bfd
!
bfd enable
S4048-TOR-1#show bfd neighbors

*       - Active session role
Ad Dn   - Admin Down
B       - BGP
C       - CLI
I       - ISIS
O       - OSPF
O3      - OSPFv3
R       - Static Route (RTM)
M       - MPLS
V       - VRRP
VT      - Vxlan Tunnel

  LocalAddr                                RemoteAddr
Interface  State Rx-int Tx-int Mult Clients
* 172.17.6.4                               172.17.1.13
Fo 1/49    Up    300    300    3    VT
* 172.17.6.4                               172.17.1.14
Fo 1/49    Up    300    300    3    VT
* 172.17.6.4                               172.17.1.15
Fo 1/49    Up    300    300    3    VT
S4048-TOR-2#
```

```
 no ip address
 no shutdown
!
interface TenGigabitEthernet 1/50/4
 vxlan-instance 1
 no ip address
 no shutdown
S4048-VTEP2#show run int fo 1/49
!
interface fortyGigE 1/49
 description To Spine2
 ip address 172.17.4.4/24
 no shutdown
S4048-VTEP2#show run int loop 0
!
interface Loopback 0
 description HW VTEP IP
 ip address 172.17.6.4/24
 no shutdown
S4048-VTEP2#show run ospf
!
router ospf 1
 network 172.17.4.0/24 area 0
 network 172.17.6.0/24 area 0
S4048-TOR-2#show run bfd
!
bfd enable
S4048-TOR-2#show bfd neighbors

*       - Active session role
Ad Dn   - Admin Down
B       - BGP
C       - CLI
I       - ISIS
O       - OSPF
O3      - OSPFv3
R       - Static Route (RTM)
M       - MPLS
V       - VRRP
VT      - Vxlan Tunnel

  LocalAddr                                RemoteAddr
Interface  State Rx-int Tx-int Mult Clients
* 172.17.6.4                               172.17.1.13
Fo 1/49    Up    300    300    3    VT
* 172.17.6.4                               172.17.1.14
Fo 1/49    Up    300    300    3    VT
* 172.17.6.4                               172.17.1.15
Fo 1/49    Up    300    300    3    VT
S4048-TOR-2#
```

5 Reference

https://www.vmware.com/files/pdf/products/nsx/vmw-nsx-network-virtualization-design-guide.pdf