Network Virtualization with Dell Infrastructure and VMware NSX

A Dell-VMware Reference Architecture

September 2014

A Dell-VMware Reference Architecture

Revisions

Date	Description	Authors
9/19/14	Version 1.2	Humair Ahmed, Dell Networking
8/20/14	Version 1.1	Humair Ahmed, Dell Networking
7/31/14	Version 1 Initial release	Humair Ahmed, Dell Networking Reviewed and validated by VMware

Table of contents

Revisions 2

1 Overview 4

2 NSX Consumption Models with Dell Networking 5

2.1 Dell Infrastructure with VMware NSX-vSphere (Solution 1) 5

2.2 Dell Infrastructure with VMware NSX-MH (Solution 2) 5

3 Dell Open Networking Differentiators 6

4 VMware vSphere-only NSX Component Definitions 7

5 Network Virtualization with VMware NSX 9

5.1 NSX Component Interaction 11

5.2 Broadcast, Unknown Unicast, and Multicast Traffic (BUM) 14

6 A Dell End-to-End Converged Infrastructure for VMware NSX 16

6.1 Converged Underlay Infrastructure with DCB and iSCSI 16

6.2 Management, Edge, and Compute Clusters and VDS 19

6.3 Infrastructure VLANs 23

7 Dell Network Design Options for VMware NSX 29

8 Additional Network Design and Scaling Considerations 32

9 Conclusion 33

1 Overview

Server virtualization has been one of the key drivers that has transformed the Data Center and Enterprise environments by decoupling the OS from the server hardware. The flexibility, automation, high-availability, and efficiency gains generated by doing so has been well received by customers and the industry with Gartner reporting in 2013 that two-thirds of x86 server workloads were virtualized. Dell has been at the forefront of this transformation providing powerful PowerEdge blade and rack servers to enable incredibly dense virtualized environments while also providing the end-to-end solution with network and storage infrastructure.

The success of server virtualization over the last decade has now brought into the forefront the need to also virtualize the network or decouple the network services from the underlying physical infrastructure. Software Defined Data Center (SDDC) is the term given to the ability to represent a physical infrastructure and its network services logically within software. The same benefits that made server virtualization incredibly popular and successful is now also driving network virtualization and the SDDC. Some of these drivers include:

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

To meet this new transformation and demand in the industry, Dell and VMware have partnered to provide a solution for the SDDC using VMware NSX network virtualization technology running on top of reliable Dell servers, networking, and storage infrastructure.

VMware NSX expands the virtualization layer from the physical layer of the server to the physical layer of the network and in consequence creates a logical framework above the physical infrastructure providing for several benefits:

- Decoupling of network services from the physical infrastructure
- Ease of network service deployment/migration/automation. Reduced provisioning/deployment time.
- Scalable multi-tenancy across datacenter
- Overlays allow for logical switches spanning across physical hosts and network services; can also span a layer 2 logical switch over L3 infrastructure allowing for better VM mobility.
- Distributed routing and distributed firewall at the hypervisor allow for better East-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 don't require changes/modification on the physical network
- Normalizes underlying hardware enabling for ease of hardware migration/interoperability

The Dell-VMware partnership allows for a complete network virtualization solution with a robust physical end-to-end infrastructure that helps enable Data Centers and Enterprises transition to a SDDC.

2 NSX Consumption Models with Dell Networking

Dell and VMware offer two solutions for network virtualization and the SDDC as outlined below in Figure 1. One model represents a vSphere-only implementation and the other represents a multi-hypervisor implementation. Both implementations are supported with Dell servers, networking, and storage, however, this whitepaper will discuss in detail only the VMware NSX-vSphere model.

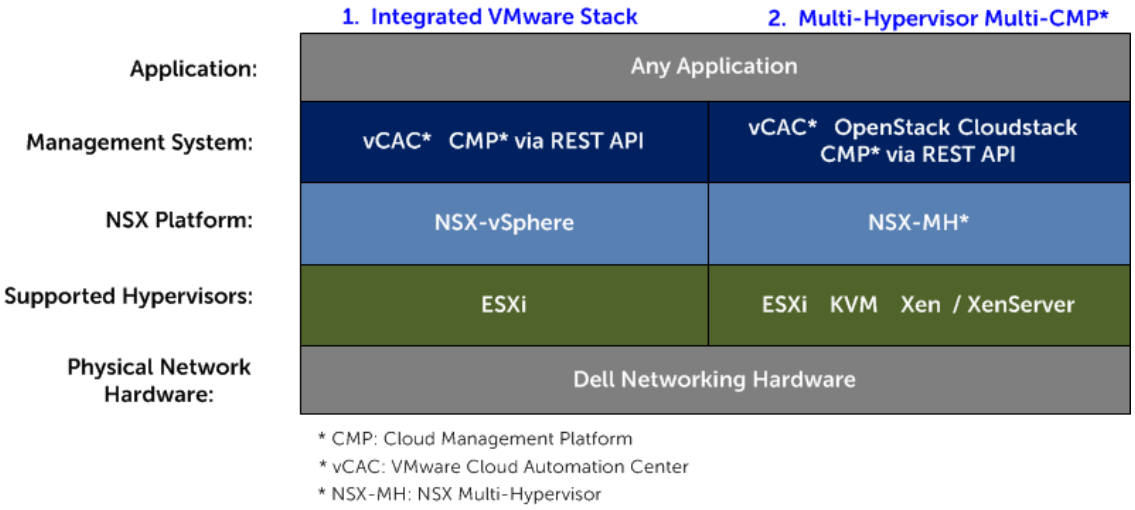

Figure 1 Two NSX Consumption models with Dell Networking

2.1 Dell Infrastructure with VMware NSX-vSphere (Solution 1)

- intended for traditional VMware customers who have only vSphere ESXi hypervisors
- only VMware ESXi hypervisor is supported
- allows for only software-based NSX L2 Gateway
- provides distributed routing and distributed firewall
- provides logical load balancer, VPN, and NAT capabilities
- provides for static and dynamic routing

2.2 Dell Infrastructure with VMware NSX-MH (Solution 2)

- intended for multi-hypervisor environments
- supports VMware ESXi, Xen, and KVM hypervisors
- supports software and hardware-based NSX L2 Gateway such as the Dell S6000, which allows for VXLAN header encapsulation/de-encapsulation at the hardware level in the ASIC at line-rate providing for better performance when bridging between traditional VLAN and logical VXLAN environments
- provides distributed routing
- provides for static routing

Figure 2 Dell S6000 switch

To stay concise and focused on the application of network virtualization, this whitepaper will focus on solution #1: **Dell Infrastructure with VMware NSX-vSphere**. Solution #2: **Dell Infrastructure with VMware NSX-MH** is a viable option for those who have multi-hypervisor environments or desire to leverage hardware-based NSX L2 Gateway.

For specific installation requirements and detailed deployment steps please see the Dell Infrastructure with VMware NSX Deployment Guide.

3 Dell Open Networking Differentiators

Dell believes in an open model for networking and solutions. Thus, Dell offers customers choice of OS on supported Dell Networking switches like the Dell S4810 and Dell S6000. Customers can choose either the typical robust, feature-full Dell Networking Operating System or Cumulus Linux. Both Dell and Cumulus operating systems are supported within a Dell VMware NSX solution.

Cumulus is a Linux software distribution that runs on Dell Networking hardware to allow for ease of integration in a Linux environment, use of Linux applications and automation tools, and an open source approach to networking applications.

The Dell Networking Operating System is a resilient, hardened OS and proven in the industry. Dell continues to innovate and incorporate the latest technology demands by customers into its products. Some of these technologies include but are not limited to:

- Data Center Bridging provides a lossless fabric for a converged LAN/SAN network
- VRF-lite provides for mult-tenancy and use of overlapping IP addresses at the switch level
- Virtual Link Trunking (VLT) for Active-Active links from a server/switch to two different switches
- VXLAN protocol support for creating logical network overlays
- Hardware-based NSX L2 Gateway support for bridging between logical and physical networks
- OpenFlow protocol support for OpenFlow based SDN solutions

Additionally, Dell provides a complete end-to-end infrastructure whether converged or non-converged and allows for the management of such infrastructure with Dell Active Fabric Manager (AFM) and Dell OpenManage Network Manager (OMNM).

Dell OMNM automates labor-intensive tasks while monitoring and managing Dell network switches. Dell AFM provides simplified configuration, management and monitoring of Dell networking components. Starting with AFM 2.0, Dell introduced the command line interface functionality for VMware vSphere Distributed Switch (VDS). Users can now access and view VDS information from the AFM CLI; with this functionality, Dell is helping bridge visibility between the virtual and physical environments.

4 VMware vSphere-only NSX Component Definitions

Before continuing with a more detailed discussion of network virtualization with VMware NSX, a few important components need to be defined and understood.

vSphere ESXi:
VMware's hypervisor installed on bare-metal servers and decouples the server operating system from the underlying hardware. The hypervisor manages the resources of the server.

vSphere vCenter:
VMware's centralized platform for managing a vSphere environment. vSphere vCenter can manage up to 1000 hosts and allows for a centralized location for ESXi configuration. vCenter is required for advanced operations, VDS, and NSX. vCenter is offered as both standalone software to be installed on a Window server or 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.

NSX Manager:
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.

vSphere Web Client:
vSphere Web Client allows for connecting to vCenter Server via a web browser. Once connected ESXi host configuration and operations can be done.

Virtual Distributed Switch (VDS):
In the NSX-vSphere implementation the **NSX vSwitch** is the **VDS switch**. The VDS switch is implemented via vCenter and ensures all hosts that are part of the VDS switch have the same virtual networking configuration. ESXi hosts are added to the VDS switch and all port groups and network configurations on the VDS switch are applied to each host. When NSX is installed, additional hypervisor kernel modules are installed in the VDS: **VXLAN, Distributed Logical Router (DLR)**, and **Distributed Firewall**.

NSX Controller Cluster:
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.

NSX API:
The NSX API is provided by the NSX Manager as a web service interface. The control cluster is responsible for the intake of network configuration directives which are sent via NSX API calls. The NSX API can also be

used by cloud management systems such as vCloud Automation Center for additional automation, monitoring, and management.

NSX Perimeter Edge (NSX Services Gateway):
A logical component installed on an edge server that provides connectivity to external uplinks and allows for logical networks to connect and peer with external networks while also providing network services such as DHCP, NAT, VPN, firewall, dynamic routing, and load balancing.

NSX Distributed Logical Router (DLR):
A logical component that provides routing and distributed forwarding within the logical network. Routing modules are installed within the kernel on each hypervisor and provide for East-West distributed routing. The DLR is managed via special purpose VM called the **DLR Control VM** that can also be used to bridge between logical (VXLAN) and physical (VLAN) networks.

Logical Switch:
Distributed logical broadcast domain/segment that can span across multiple clusters and to which a VM can be logically connected/wired to. This allows for VM mobility without concern for traditional physical layer 2 boundaries.

Logical Firewall:
The distributed firewall allows for a logical firewall that can segment virtual entities within the logical network; this is a hypervisor kernel-embedded distributed firewall. The perimeter edge firewall allows for perimeter security while also allowing for services on the perimeter edge such as DHCP, NAT, VPN, dynamic routing, and load balancing.

Transport Zone:
A transport zone defines the span of a logical network and is installed at the cluster level. A logical switch can span only across the hosts/clusters which are part of the transport zone.

VXLAN:
Standard network overlay technology where MAC frames are encapsulated into a VXLAN and UDP header and communication occurs between two endpoints called **Virtual Tunnel Endpoints (VTEPs)**. VMware NSX uses VXLAN to build logical L2 networks over any L2/L3 physical IP infrastructure.

Virtual Tunnel Endpoints (VTEPs):
vmkernel IP interfaces that are the endpoints for VXLAN communication. During the VXLAN configuration process, a VTEP is configured on every vSphere ESXi host that will be participating within a logical network. The IP addresses of the source and destination VTEP are used in the outer header of the VXLAN encapsulated packet; these are the IP addresses used for routing the packet through the physical underlay network.

VXLAN Network Identifier (VNI):
A logical switch is defined with a VXLAN unique identifier called a VNI which represents a logical segment.

5 Network Virtualization with VMware NSX

VMware NSX is VMware's network virtualization technology and allows for the decoupling of network services from the physical infrastructure. Overlay logical networks are created on top of a basic L2 or L3 physical infrastructure (underlay). Overlay logical networks are created using VXLAN, a L2 in L3 tunneling protocol (overlay).

In addition to the ability to decouple networking from the physical underlying hardware via logical networks, the NSX platform also provides for network services in the logical space as shown in Figure 3 below. Some of these logical services include switching, distributed routing, firewall, load balancing, DHCP, NAT, and VPN capabilities. Further, NSX also provides the capabilities of third party integration via NSX Service Composer. NSX Service Composer is provided via NSX Manager vCenter plug-in and allows for automating the consumption of services and their mapping to virtual machines using a logical policy.

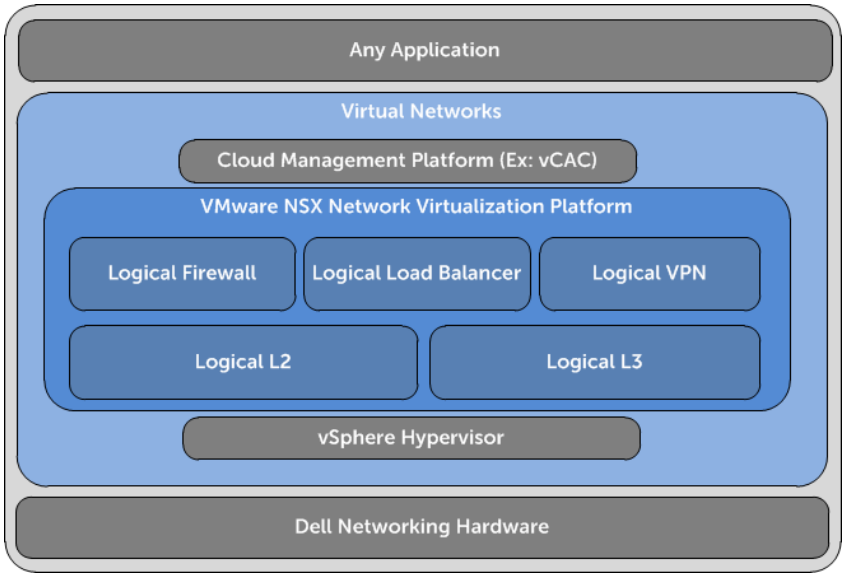

Figure 3 VMware NSX virtualization platform on Dell hardware (vSphere-only environment)

A logical switch is an abstraction of a physical switch. It is important to understand that a logical switch is not the same as a standard vSwitch or Distributed Virtual Switch (VDS). A logical switch is implemented via network overlay technology making it independent of the underlying physical infrastructure and results in an overlay logical network. A traditional standard vSwitch or VDS results in a bridged logical network that uses VLANs. Both overlay logical networks and bridged logical networks can co-exist within the same environment.

Once an overlay logical switch/network is created, VMs can be directly connected and completely separated from the physical network (underlay). Another advantage of overlay logical switches/networks is that it enables multi-tenancy and provides address isolation; tenants can be separated via different logical networks as shown below in Figure 4 where a VNI identifies a logical switch. To keep wording simplified and to cut down on verbosity, in this whitepaper, logical switch/network refers to an overlay logical switch/network.

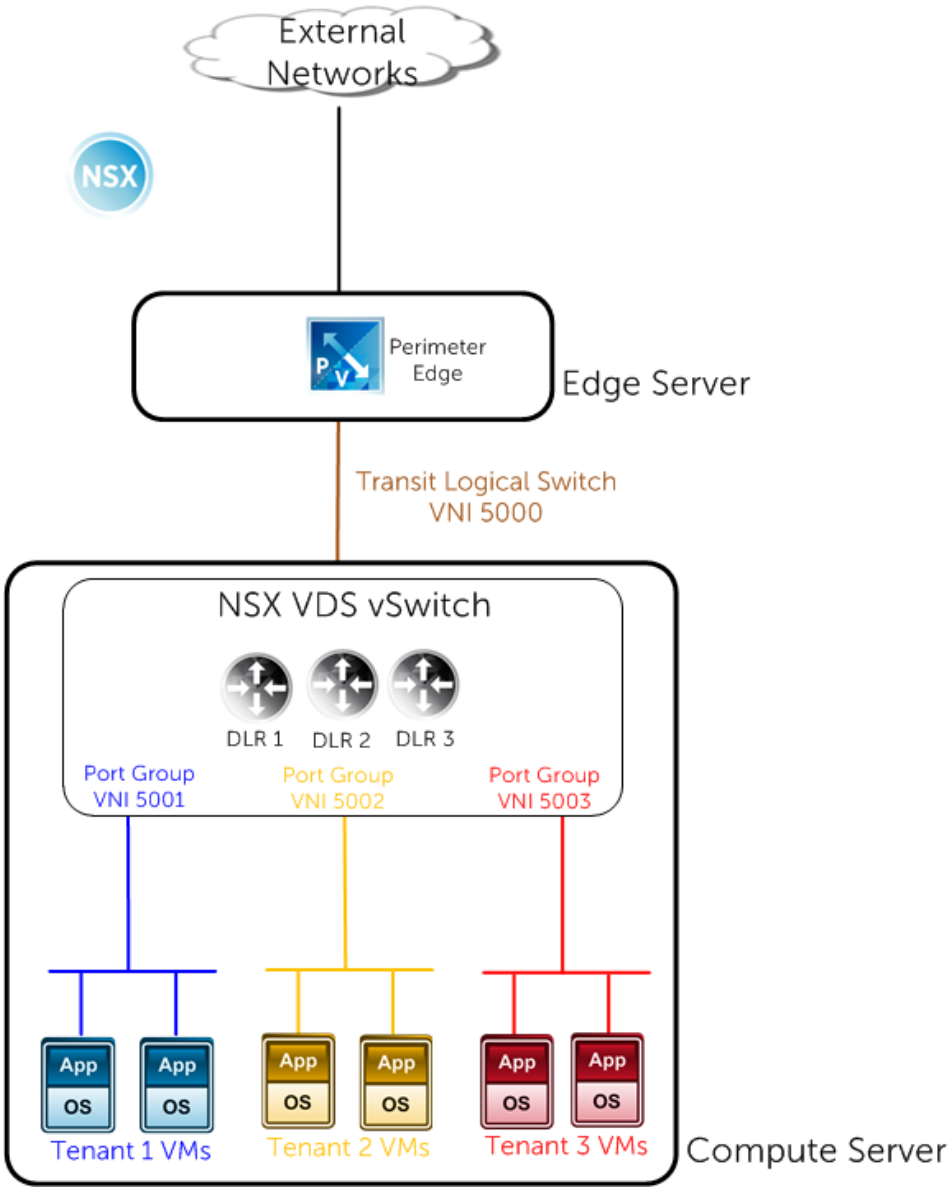

Figure 4 Multi-tenancy via creation of multiple logical networks and routers

Further, multi-tiered applications can be built completely in the virtual space allowing for better East-West traffic flow. Prior to this ability offered by VMware NSX, traffic from VMs on different subnets would need to leave the virtual space and be routed by an external physical router/L3 switch; this was true even if the VMs were on the same physical host as shown on the left side of Figure 5 below. On the right side of Figure 5, with NSX installed and configured, routing between the two subnets can be handled without ever leaving the host via the kernel-level DLR within the hypervisor.

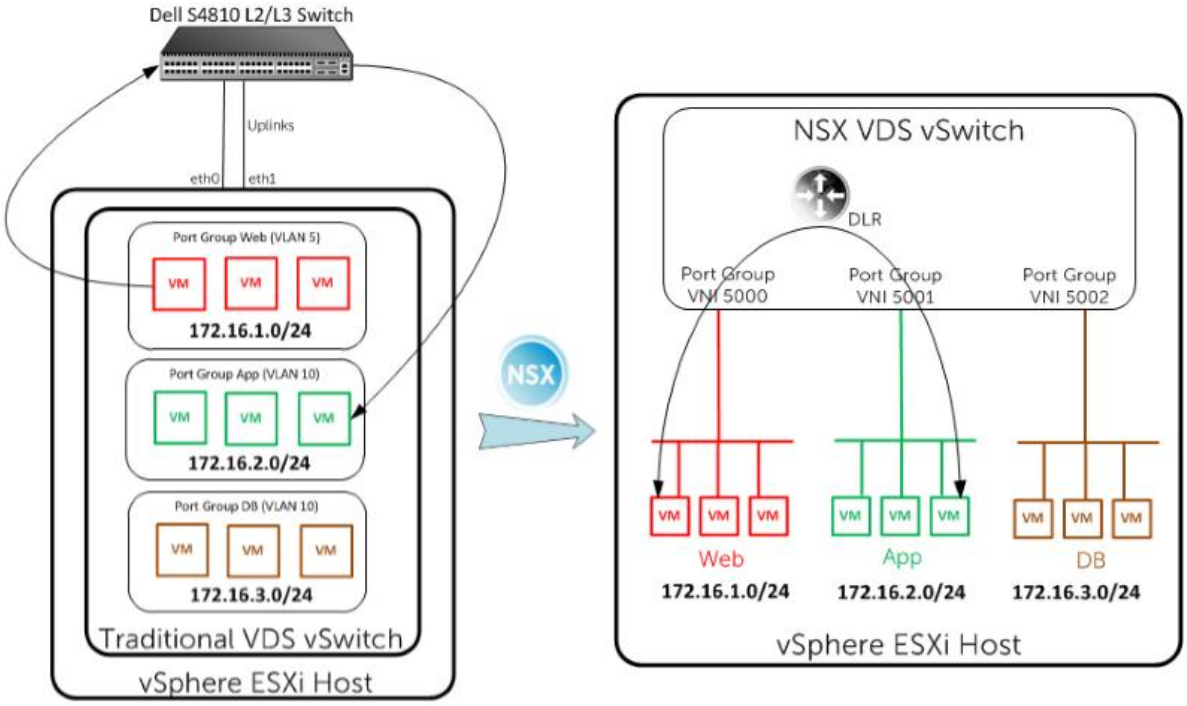

Figure 5 VMware NSX allows for routing within the virtual space without leaving the vSphere ESXi host

5.1 NSX Component Interaction

NSX components can be categorized as shown below in Figure 6. The NSX platform consists of a separate management, control, and data plane. The NSX vSwitch with its hypervisor kernel modules and the NSX Edge components sit in the data plane which is used for all data forwarding. **The control plane is utilized to distribute VXLAN and logical routing network information to ESXi hosts; at no point do data packets ever get forwarded to the control cluster**. All vSphere and NSX components have connectivity to the management network which is utilized for components to be able to interact with each other; this is also how the controllers are able to communicate to the NSX VDS switches and NSX edge devices.

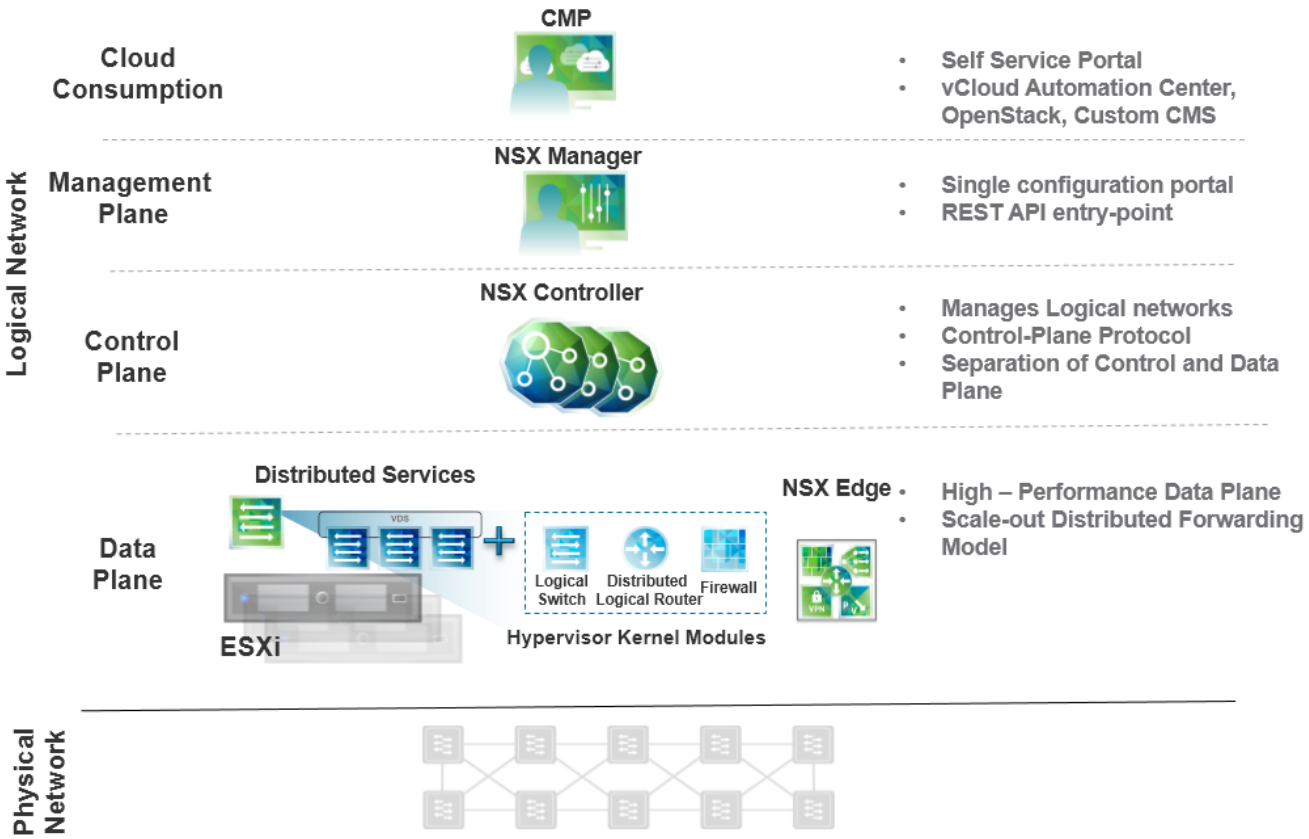

Figure 6 VMware vSphere-only NSX components

A cluster is prepared for NSX in a two-step process.

1. The kernel modules for VXLAN, DLR, and firewall are installed on the compute and edge clusters.
2. VXLAN is configured on the compute and edge clusters. As part of the configuration, each host is configured with a VTEP (virtual tunnel end point). This is where VXLAN encapsulation/de-encapsulation occurs as packets are sent across the network.

Once the NSX and VXLAN configuration is completed, logical switches/networks can be created via NSX Manger, NSX REST API, or a cloud management platform (CMP) such as VMware Cloud Automation Center (vCAC). A logical switch is defined with a VXLAN unique identifier called a VNI which represents a logical segment.

VMs on different hosts but on the same logical network are encapsulated with a VXLAN and UDP header and sent over the physical underlay network. The outer IP header is all that the physical underlay network sees; it is unaware of what VMs are talking to each other. Figure 7 below displays what a VXLAN encapsulated frame looks like. It's also important to note that any QoS bits within the encapsulated frame are copied to the external header so the physical underlay network can respect the DSCP or CoS QoS settings for different types of traffic. VMs on different logical networks can also communicate with each other via a logical DLR or Perimeter Edge router without leaving the virtual environment.

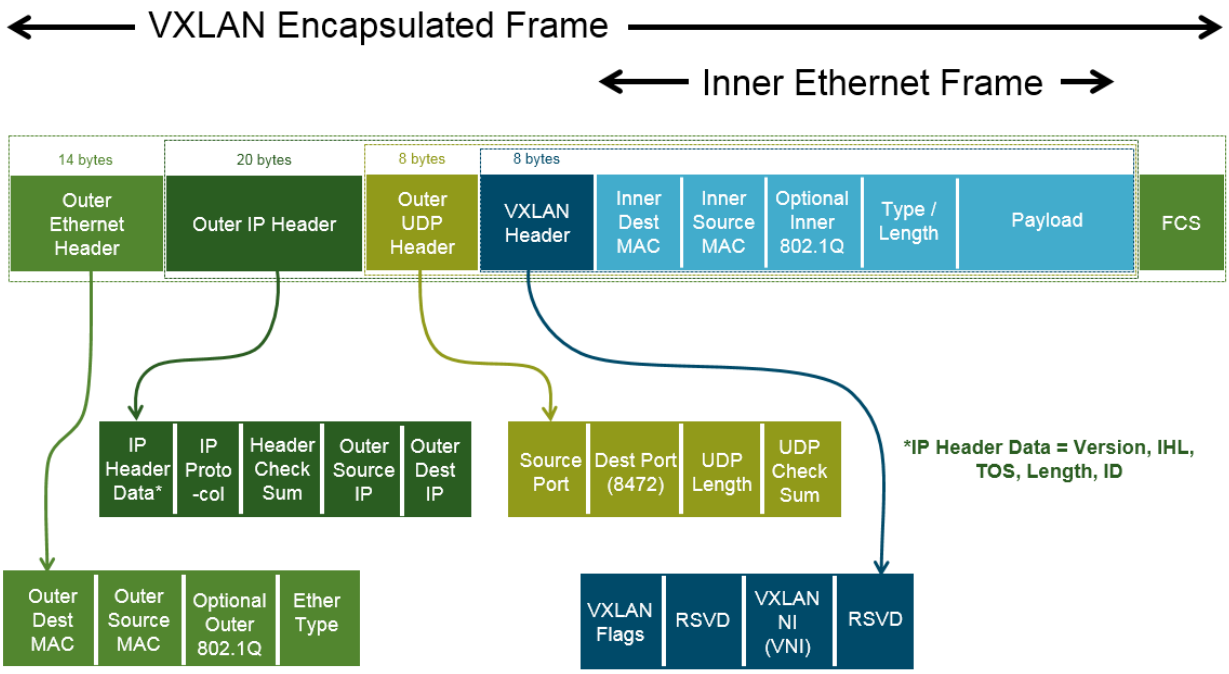

Figure 7 VXLAN Encapsulated Frame

The NSX controllers are all active and clustered for scalability and high availability; network information is distributed across the nodes in the cluster, and each control node is elected as the master for different roles and tasks.

The NSX control plane communication occurs over the management network. Each controller node has a unique IP address and nodes monitor each other via heart-beat; failed nodes are detected within ten seconds. A supported configuration consists of a cluster of at least three controllers.

The controller high availability is based on a majority concept in which a failure of one controller does not impact the control plane operation of the control cluster. Also, the network will continue forwarding traffic even if all three controllers fail, however, any changes or learning of new network information will not be possible until at least one controller is brought back to operation.

The NSX Manager communicates with the vCenter server and NSX controllers via the management plane. NSX Manager utilizes REST API calls to configure the NSX components on the logical network. Once the control cluster receives the NSX API calls, it's responsible for propagating the network state, configurations, and updates to all hosts in the respective transport zone. This communication is done via connection between controller and hosts. This connectivity from the host viewpoint is managed by the **User World Agent (UWA)** which is a is a TCP (SSL) client that sits on each host and communicates with the controller using the control plane protocol.

Note, once the network state, configuration, and update information has been pushed down to the host, data can move directly between hypervisors.

The vCenter server also utilizes the management plane to manage the ESXi hosts via the **Message Bus Agent (MBA)**. It's important to note the NSX Manager, vCenter, and NSX controllers all sit on the management network.

Additionally, a cloud management platform such as **vCloud Automation Center (vCAC)** can be utilized to more easily configure, manage, monitor, and automate the network. It will utilize the northbound REST API available via the NSX Manager, and the controller cluster is responsible for the intake of the respective network configuration directives from the NSX Manager. Figure 8 below displays a diagram showing the NSX component interaction.

Note here that NSX Manager achieves control plane security for communication to controllers and vSphere ESXi hosts via SSL certificates. The NSX VDS switches and NSX edge devices register with the NSX controller cluster using the NSX API; security is provided via SSL certificate. The NSX Manager generates self-signed certificates for each of the ESXi hosts and controllers; these certificates are then pushed to the respective ESXi hosts and controllers.

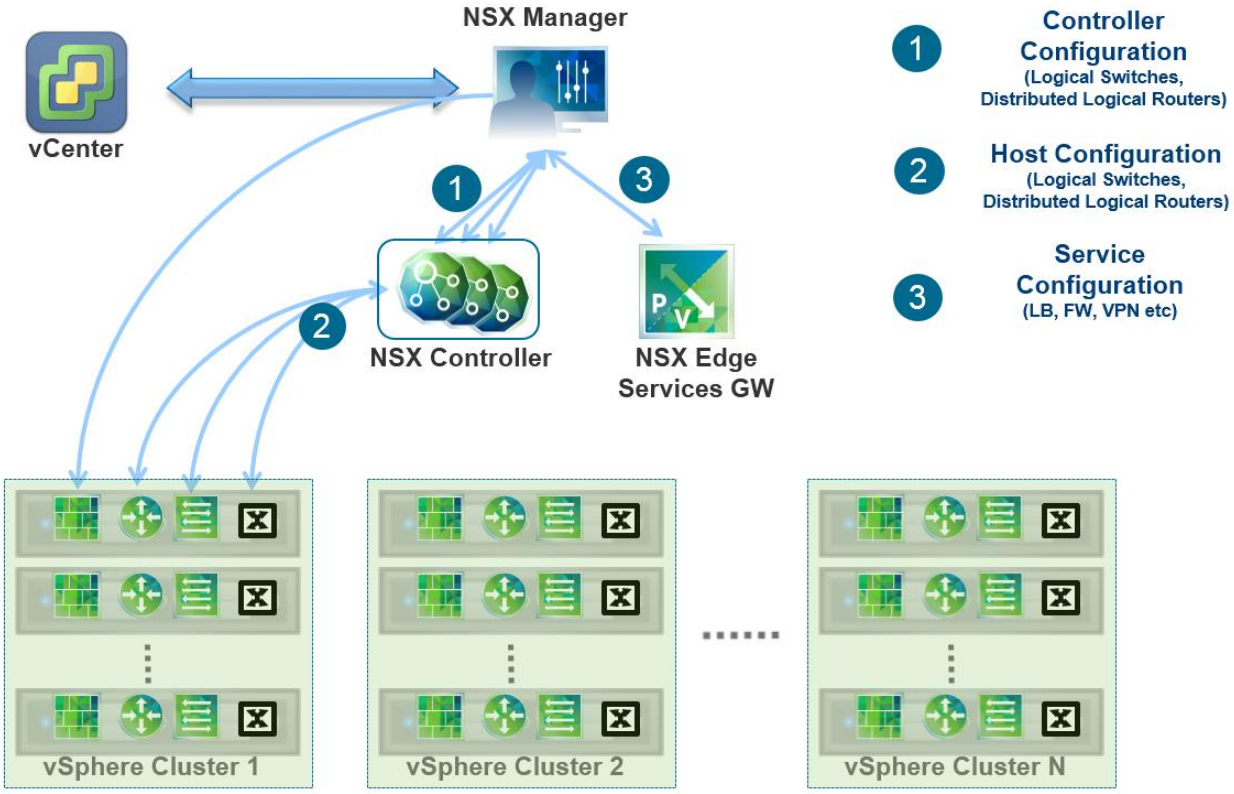

Figure 8 VMware NSX component interaction

5.2 Broadcast, Unknown Unicast, and Multicast Traffic (BUM)

The VXLAN standard resolves broadcast, unknown unicast, and multicast (BUM) traffic via multicast enabled physical network. NSX provides flexibility in how VXLAN replication is handled. It offers three control plane modes for the handling of BUM traffic: unicast, hybrid, and multicast.

In unicast mode, the multicast support on the underlay network switches is no longer required by VXLAN to handle BUM traffic. A great advantage 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.

In hybrid mode, some BUM traffic on a given local L2 logical segment is offloaded to the first hop physical switch, which then does the replication for other hosts on the same logical network. Replication of BUM traffic destined for the same host is still done by the source host. Although, this will provide some better performance, it requires IGMP Snooping configuration on the first-hop Dell switch. IGMP Querier on the network is also recommended for generating IGMP queries.

Multicast mode is also still available and offloads all BUM traffic replication to the underlay network; this requires PIM and IGMP configuration on the physical underlay network.

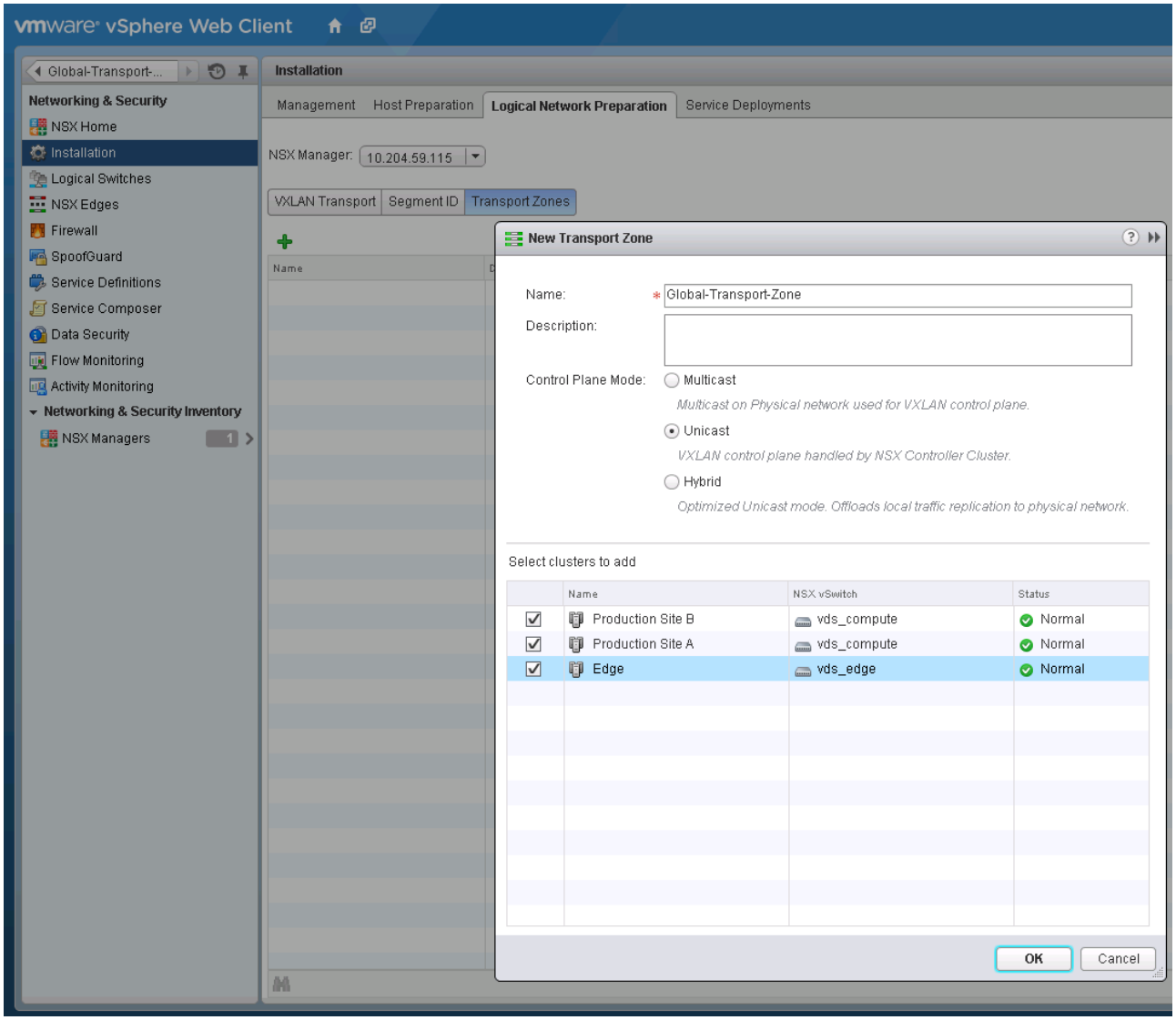

Figure 9 VMware vSphere Web Client: NSX Manager vCenter Plug-in allows for NSX configuration within VMware vCenter. Here, a transport zone is being created with control plane mode as unicast.

Figure 9 above shows that one of the three control plane modes must be selected when creating a transport zone via NSX Manager vCenter plug-in. As mentioned prior, a transport zone defines the boundaries of a logical network/switch. Hosts are added to the transport zone at the cluster level. Below, the three clusters being added to the transport zone **are Edge, Productions Site A (Compute)**, and **Production Site B (Compute).**

Defining of the transport zone, replication mode selection, and VXLAN configuration is the second step of NSX deployment. Once this configuration is done all other configuration of defining logical components (logical switch, DLR, DFW, logical load balancers, etc.) can be done repetitively via manual process or API calls. The next section details the physical network architecture as well as the logical components that are utilized to achieve three tier application provisioning in a flexible, consistent, and dynamic method.

6 A Dell End-to-End Converged Infrastructure for VMware NSX

As we now have a basic understanding of VMware NSX and the respective logical components and interactions, we next take a look at a complete Dell end-to-end converged infrastructure used to deploy VMware vSphere-only NSX. For specific installation requirements and detailed deployment steps please see the Dell Infrastructure with VMware NSX Deployment Guide.

6.1 Converged Underlay Infrastructure with DCB and iSCSI

Figure 10 below displays a Dell converged infrastructure underlay for VMware NSX. Depending on requirements and the desire to use an already existing infrastructure, Dell and VMware provide flexibility to support many different L2/L3 or hybrid L2/L3 underlay deployments.

Although VMware NSX is agnostic to the type of storage utilized in the network, it must be taken into account that the storage traffic is not virtualized and depending on what type of storage is used (iSCSI, FCoE, FC, NAS), this may affect the network underlay design of the overall solution. Note, at present ESXi iSCSI multi-pathing requires a non-routed network between the host and the iSCSI storage (for details, see the following VMware Knowledge Base article: http://kb.vmware.com/kb/1009524). In the setup shown in Figure 10, iSCSI storage to host multi-path connectivity is achieved via Layer 2 connectivity which also take advantage of DCB to provide for a lossless converged fabric.

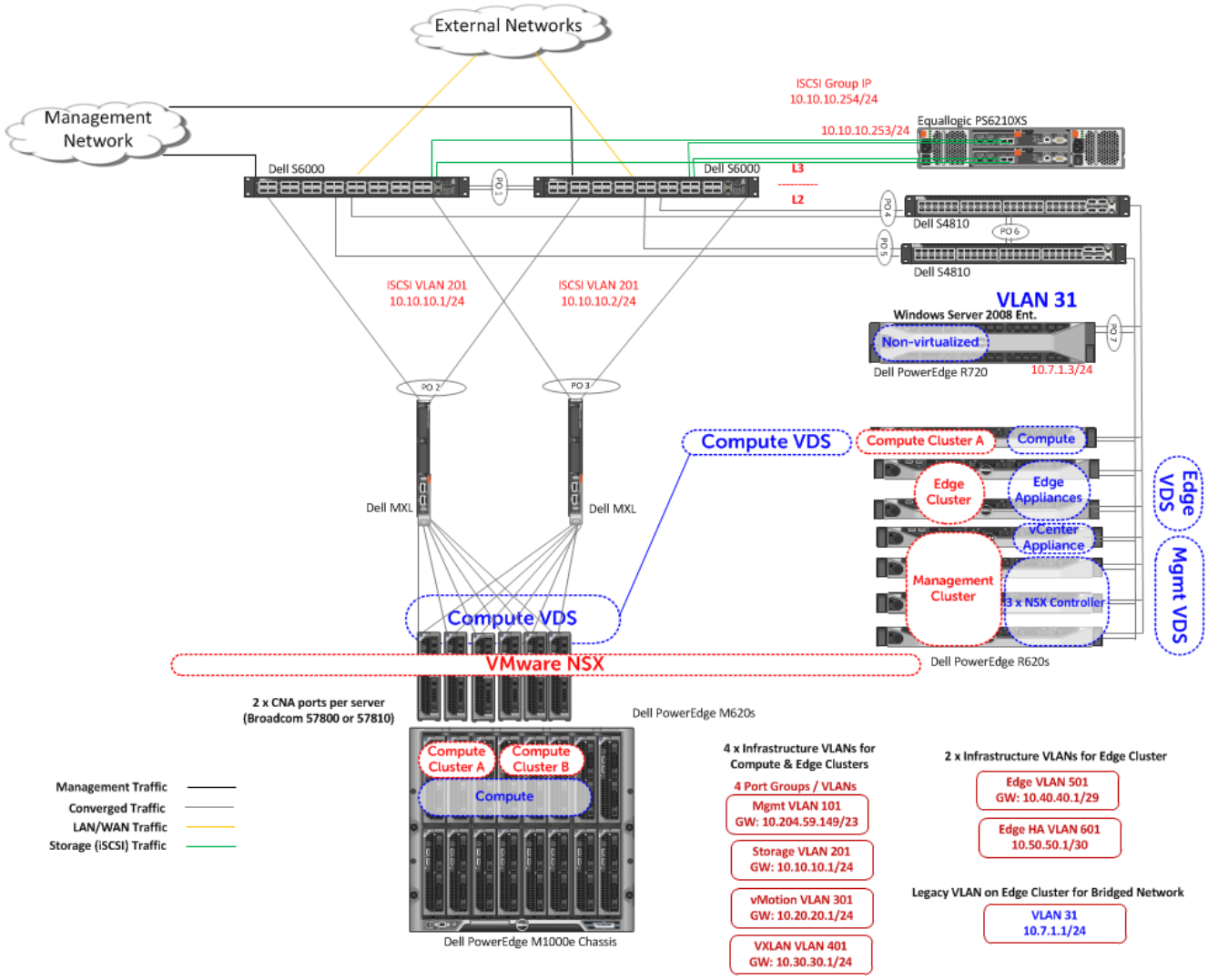

Figure 10 Dell End-to-End Converged Infrastructure with VMware NSX

In this end-to-end converged design, Data Center Bridging (DCB) is utilized to provide a lossless medium for converged LAN and SAN traffic. iSCSI storage, LAN, Management, vMotion, and VXLAN transport traffic is carried over 2 x 10 GbE ports per each compute and edge server. At least two edge servers are deployed for high availability, and the edge servers carry additional traffic for edge and high availability.

Important characteristics about this physical infrastructure:

- In this collapsed core architecture, Dell MXL blade access switches act as ToR for blade servers and Dell S4810 switches act as ToR for rack servers.

- This is a completely converged architecture leveraging DCB. Broadcom 57810 and 57800 CNAs were utilized and all iSCSI offload is done on the CNA; this implies that all priority tagging is handled by the CNA. There are 2 x 10 GbE Broadcom CNA ports on each server that carry LAN and

SAN traffic, and each CNA port per server is connected to a different ToR/access switch as shown in Figure 10 above.

- Dell EqualLogic PS6210 array is used and is DCB enabled for use with iSCSI. It has 2 controllers with 2 x 10 GbE copper and fiber ports (either the 2 x copper or the 2 x fiber ports can be utilized but not both).

- A 40 GbE to 4 x 10 GbE breakout cable is utilized to connect the S6000 to the Dell EqualLogic storage array; unused 10 GbE ports can be utilized for additional EqualLogic storage arrays (only one EqualLogic array is shown above in Figure 10).

- The Dell EqualLogic array controllers are active/standby with vertical failover employed, meaning that corresponding physical ports in each controller (vertical pairs) are combined into a single logical port from the point of view of the active controller. This allows for continuous full bandwidth to the array even if there is a link or switch failure.

- Utilizing DCB and having the Dell EqualLogic array connected centrally to the core allows for ease of scale while providing remote storage access to all blade and rack servers. Additional arrays can easily be connected to the Dell S6000 switches.

- Proxy ARP on Virtual Link Trunking (VLT) peer nodes or VRRP can be used to provide high availability for infrastructure VLAN gateways on the L2/L3 boundary on the S6000s.

  VRRP provides a virtual IP that can be used by two switches for redundancy purposes. One switch will be active in responding to ARP requests for the virtual IP while the other switch will be standby.

  A proxy ARP-enabled L3 switch/router answers the ARP requests that are destined for another L3 switch/router. The local host forwards the traffic to the proxy ARP-enabled device, which in turn transmits the packets to the destination.

VLT is Dell's L2 multi-pathing technology which allows for active-active connectivity from one host/switch to two different switches. In this network design, it is utilized on the Dell S6000 core switches down to the Dell MXL switches and ToR Dell S4810 switches. VLT is not used on the Dell MXL or S4810 switches to connect to any vSphere ESXi hosts. Active-Active connectivity is achieved from the vSphere ESXi server to the access/ToR switches via VMware vSphere's teaming options. VLT is only used on the ToR S4810s for connecting to the non-virtualized Dell PowerEdge R720 server running Windows Server 2008 Enterprise with local storage.

6.2 Management, Edge, and Compute Clusters and VDS

Note, in the hybrid blade/rack server deployment in Figure 10, all management and edge servers are rack servers; this allows for ease of management and scale for management and edge servers which will typically be located in management / edge racks. Also, by having all edge devices on rack servers in a hybrid environment, traffic flow will always be predictable when bridging from logical to physical and routing North-South traffic. Although rack layout can vary depending on size, space, and topology, a typical rack layout with VMware NSX is shown in Figure 11 below.

Note, the separate racks for compute and the combined racks for management, edge, and storage. Often, the management, edge, and storage racks are combined as shown or edge is separated from management and storage. In a converged fabric as displayed in Figure 10, it is typical to connect the storage centrally for better distribution and scale across all hosts/clusters; as ToR switches are added, they simply connect to the core S6000 switches and have access to all centralized storage.

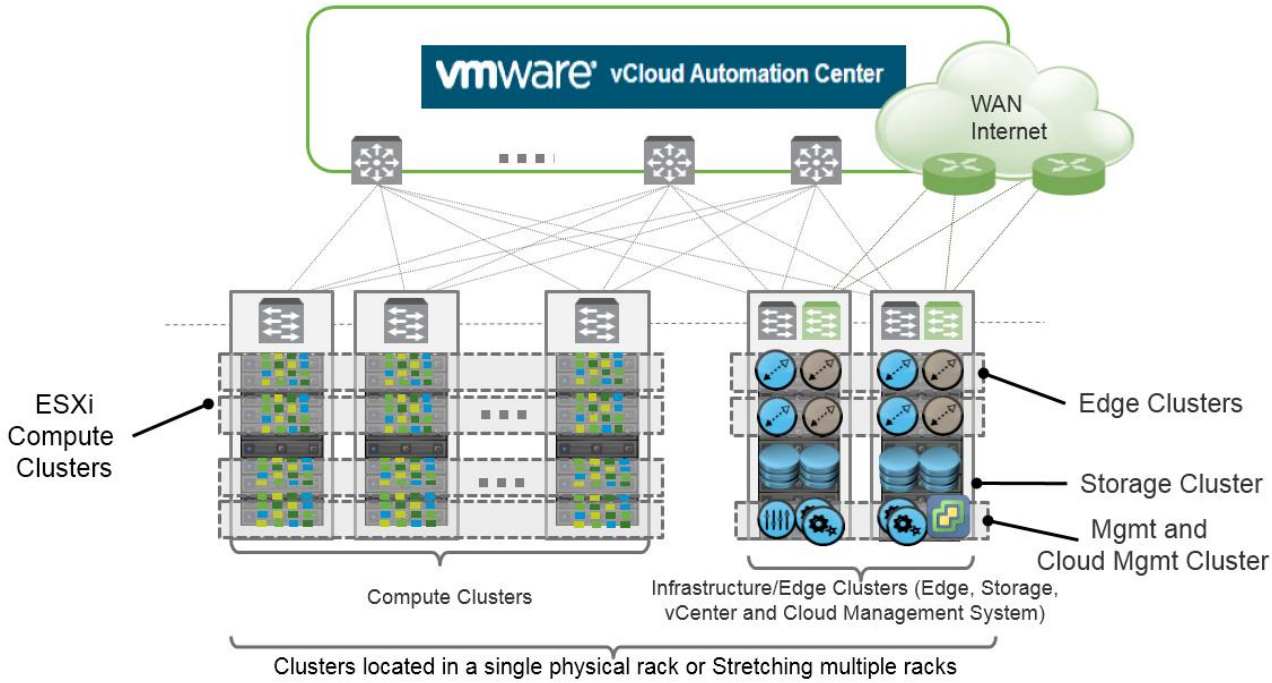

Figure 11 Typical rack layout for compute, management, and edge servers/clusters. Storage can be centrally connected to the core as in a completely converged fabric design or connected to ToR dedicated storage switches at EoR for a typical non-converged design as often shown with Fibre Channel or dedicated iSCSI storage.

As mentioned prior, the transport zone defines the span of a logical switch/network. Correspondingly, management hosts never have to be part of the transport zone as logical networks should not span across management hosts. However, logical networks will span across compute cluster(s) as the VMs and workloads will be on the hosts in compute cluster(s).

Additionally, edge cluster(s) need to also be part of the transport zone as edge components (NSX Perimeter Edge & DLR Control VM) that connect to and interact with the logical switches forward traffic/interact with the physical infrastructure. The Distributed Logical Router Control VM sits on the edge server and provides the control plane and configuration functions for the kernel-embedded DLR on the vSphere ESXi hosts. A minimum of two edge servers allows for high availability in an active/active configuration for edge devices. Scalability is achieved by each tenant DLR having a pair of active/active DLR Control VMs and Perimeter Edge devices. Starting with NSX 6.1 release, ECMP is supported on both the DLR and the Perimeter Edge. Up to eight active/active edges are possible per each tenant DLR.

Although not required, by having a separate cluster respectively for compute, management, and edge, it allows for cluster-level configuration for each cluster without affecting hosts in other clusters. Also, it allows for some predictability of Distributed Resource Scheduling (DRS) at the cluster-level; this can help ensure for example that a NSX controller (installed as a VM on a host in the management cluster) does not end up being vMotioned by DRS to a host where NSX Edge Services (contained within the edge cluster) is running.

You can see separate edge, management, and compute **(Production Site A & B)** clusters in Figure 12 below. The servers that are part of the respective clusters are labeled in the prior diagram in Figure 10. Note that each of these clusters has its own VDS switch. This also is not a requirement but is a design choice so NSX bits and infrastructure VLANs are only installed/configured on the hosts/clusters they need to be installed/configured on during the deployment process. There are different designs possible if desired such as having one VDS span across all clusters, or having one VDS for edge and compute clusters and another for the management cluster. For more detailed information see the Dell Infrastructure with VMware NSX Deployment Guide.

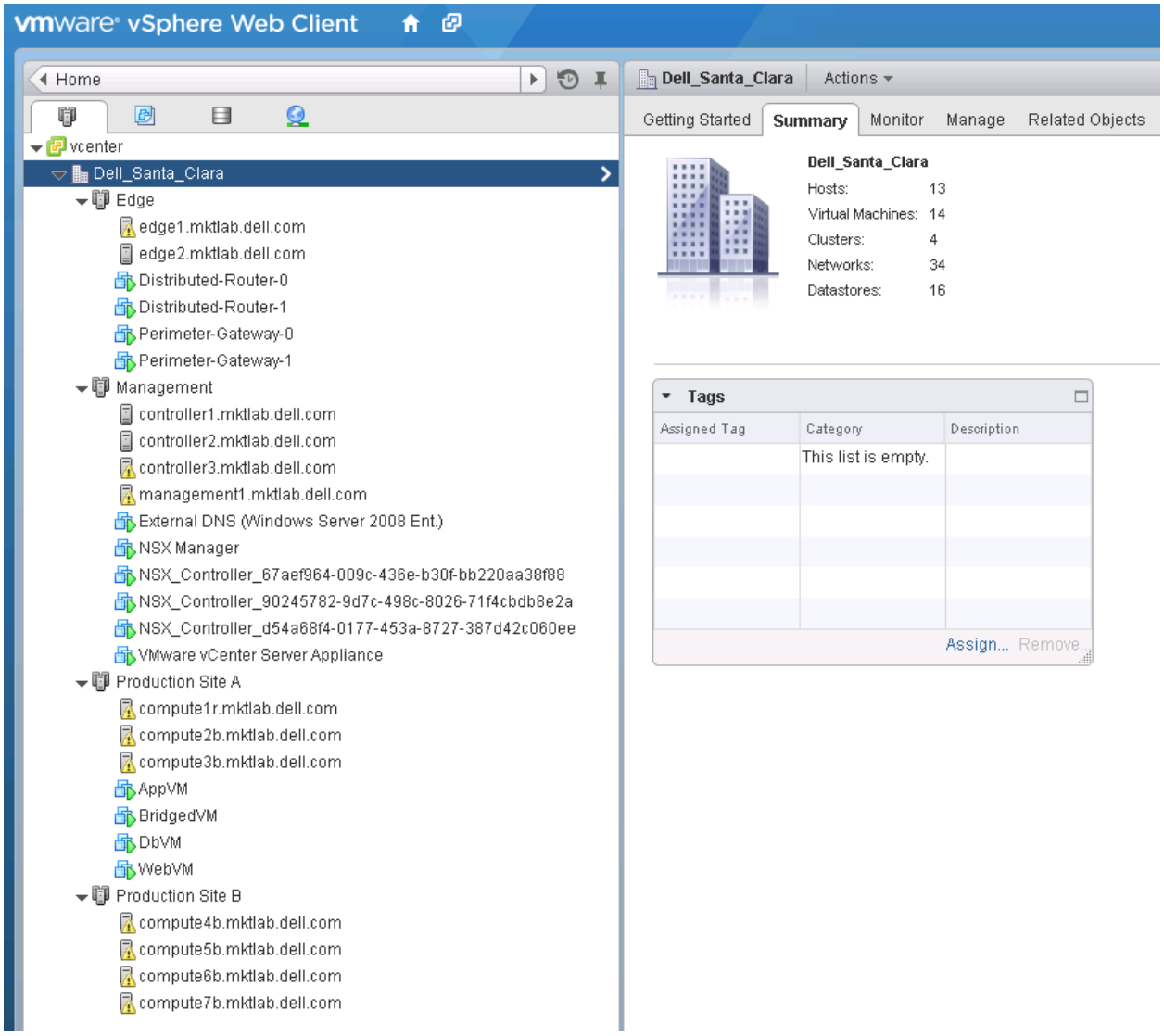

Figure 12 Edge, management, and compute clusters within the Dell_Santa_Clara datacenter

In Figure 13 below, you can see that the NSX bits have been installed into the compute and edge clusters and VXLAN has been enabled. However, the management cluster which hosts the vCenter Virtual Appliance, DNS, NSX Manager, and NSX Controllers does not have the NSX bits installed or VXLAN enabled as it is not required. No logical network components need to connect to anything within the management cluster.

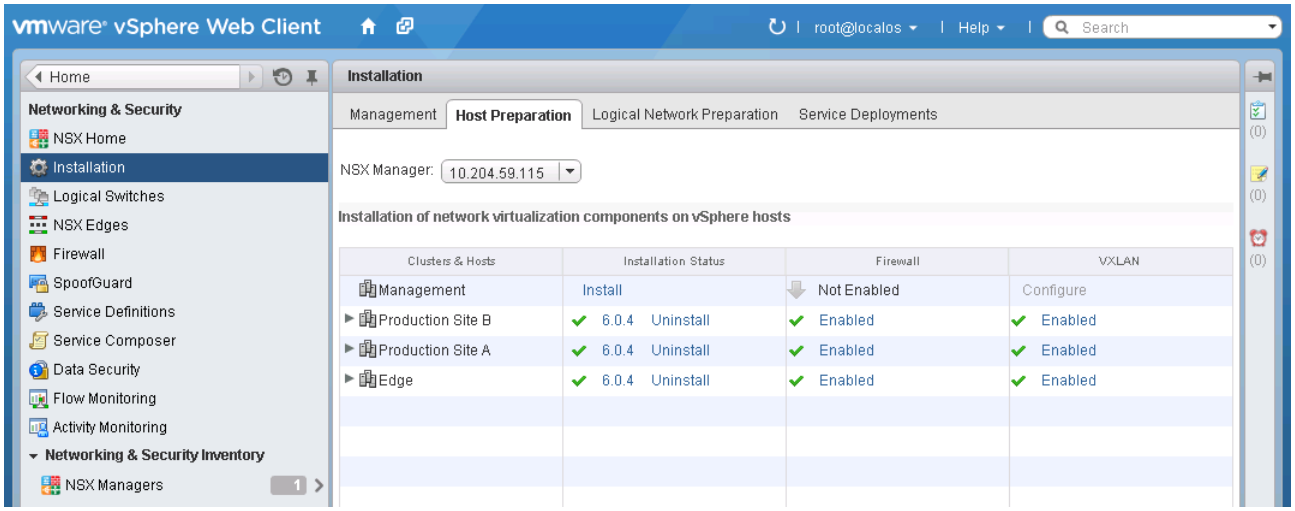

Figure 13 VMware vCenter NSX Plug-in: NSX bits and VXLAN configured at the cluster-level

You can see from Figure 14 below that when VXLAN is configured, a VDS switch is selected. Here VXLAN is being configured on the compute cluster Production Site A. It also needs to be configured on the edge cluster. If the management and edge clusters were on the same VDS, the transport VLAN 401, would be automatically created on all management and edge hosts even though it is only needed on the compute and edge hosts, thus the decision to have three separate VDS switches, one for each cluster.

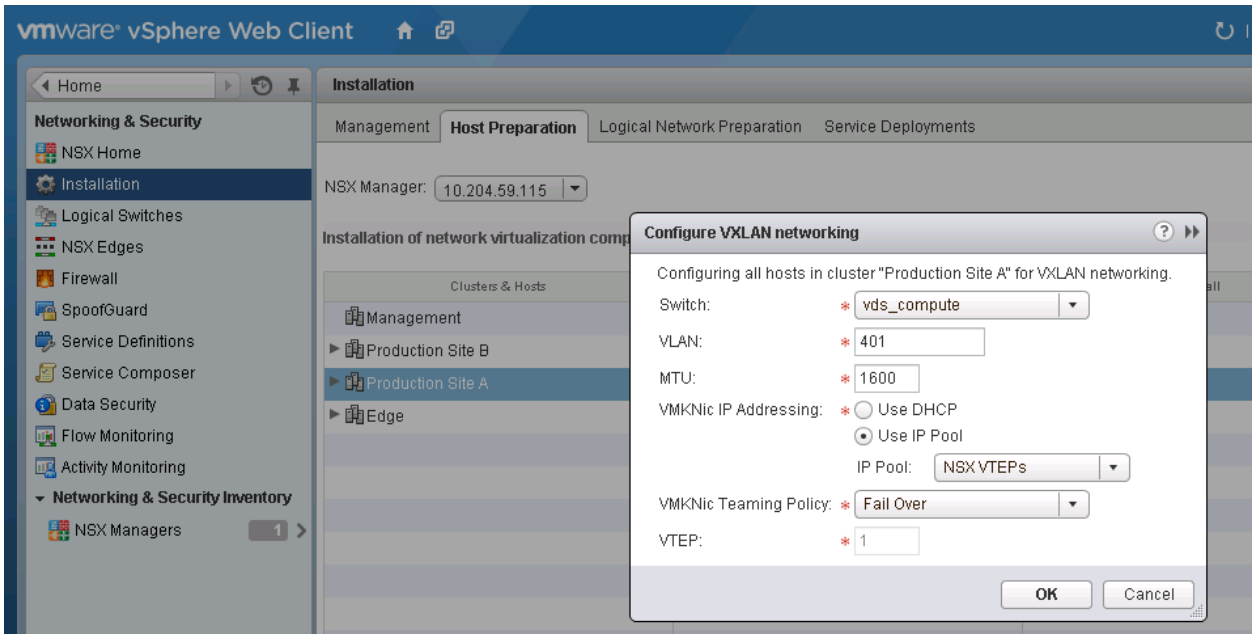

Figure 14 VMware vCenter NSX Plug-in: VXLAN configuration on compute cluster 'Production Site A'

6.3 Infrastructure VLANs

Infrastructure VLANs are still utilized on the physical underlay for management, storage, vMotion, VXLAN transport, and edge traffic. The type of network underlay design one chooses (L2, L3, or hybrid L2/L3) will determine where these infrastructure VLANs need to span on the physical infrastructure. Regardless, once these infrastructure VLANs are in place you will never have a need to configure additional VLANs on the physical network again. These infrastructure VLANs must also be present on the corresponding VDS switch. As an example, below in Figure 15 is the VDS port group configuration for the **compute_vds** VDS switch which is utilized for the compute clusters **Production Site A** and **Production Site B** in Figure 10.

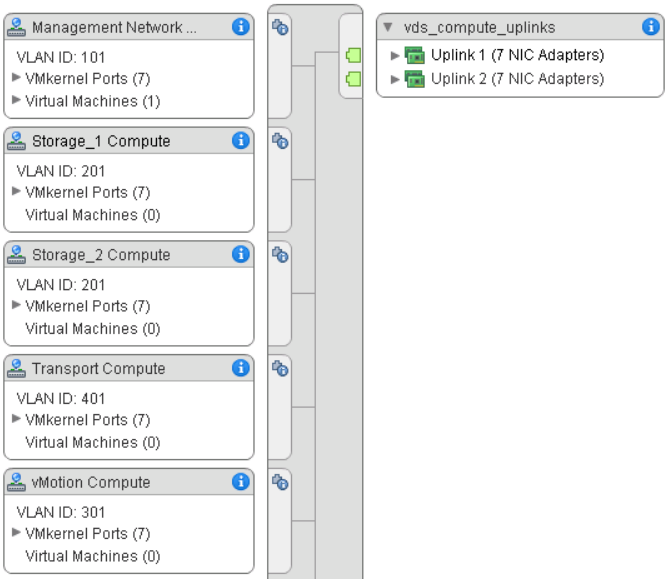

Figure 15 'compute_vds' VDS switch

The infrastructure Port Groups/VLANs in the setup are as follows:

Table 1 Infrastructure Port Groups/VLANS

Port Group Name	VLAN	Use
Management Network	101	Management traffic
Storage_1 Compute	201	Multi-path iSCSI storage traffic (Port 1)
Storage_2 Compute	201	Multi-path iSCSI storage traffic (Port 2)
Transport Compute	401	VXLAN transport traffic
vMotion	301	vMotion traffic

If we expand the **Transport Compute** port group's VMkernel ports as shown in Figure 16, we see exactly seven VMkernel interfaces or VTEPs with corresponding IP addresses for the seven servers we have in the compute clusters as shown in Figure 10.

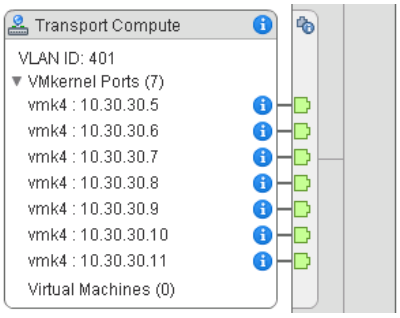

Figure 16 'Transport Compute' port group displaying seven VTEPs for all seven servers in the compute clusters.

As logical networks are created, port groups representing logical wires are automatically created on the VDS switch.

As an example, in Figure 17 directly below five logical switches are created. Figure 18 displays the port groups or logical wires automatically created on the **compute_vds** VDS switch in consequence; note, all port groups are tagged on **VLAN 401**, the transport VLAN. The reason for this is because all traffic originating from workloads connected to the logical switch and leaving the ESXi host will be VXLAN encapsulated and tagged with **VLAN 401**.

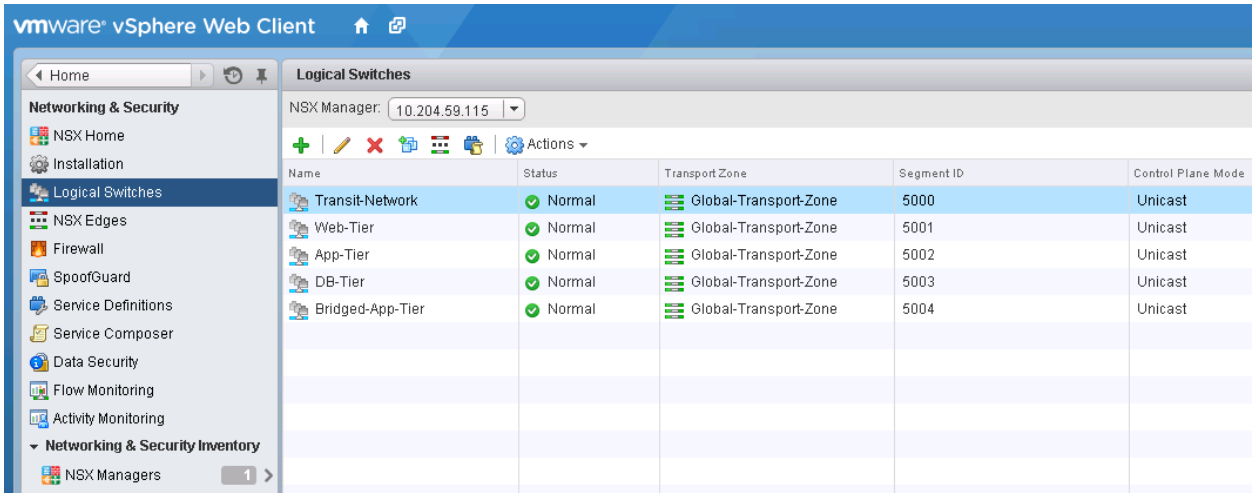

Figure 17 Five logical networks created

In Figure 18, VMs have also already been added to the respective logical networks.

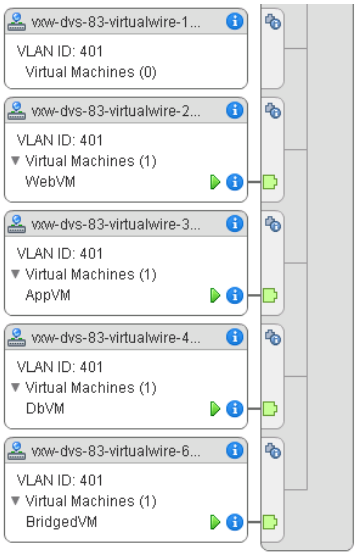

Figure 18 Corresponding port groups or virtual wires automatically created when the five logical networks in Figure 17 were created (VMs were added to the respective logical networks later)

Figure 19 below again shows the complete physical design. This corresponds to the complete logical network design shown further below in Figure 20.

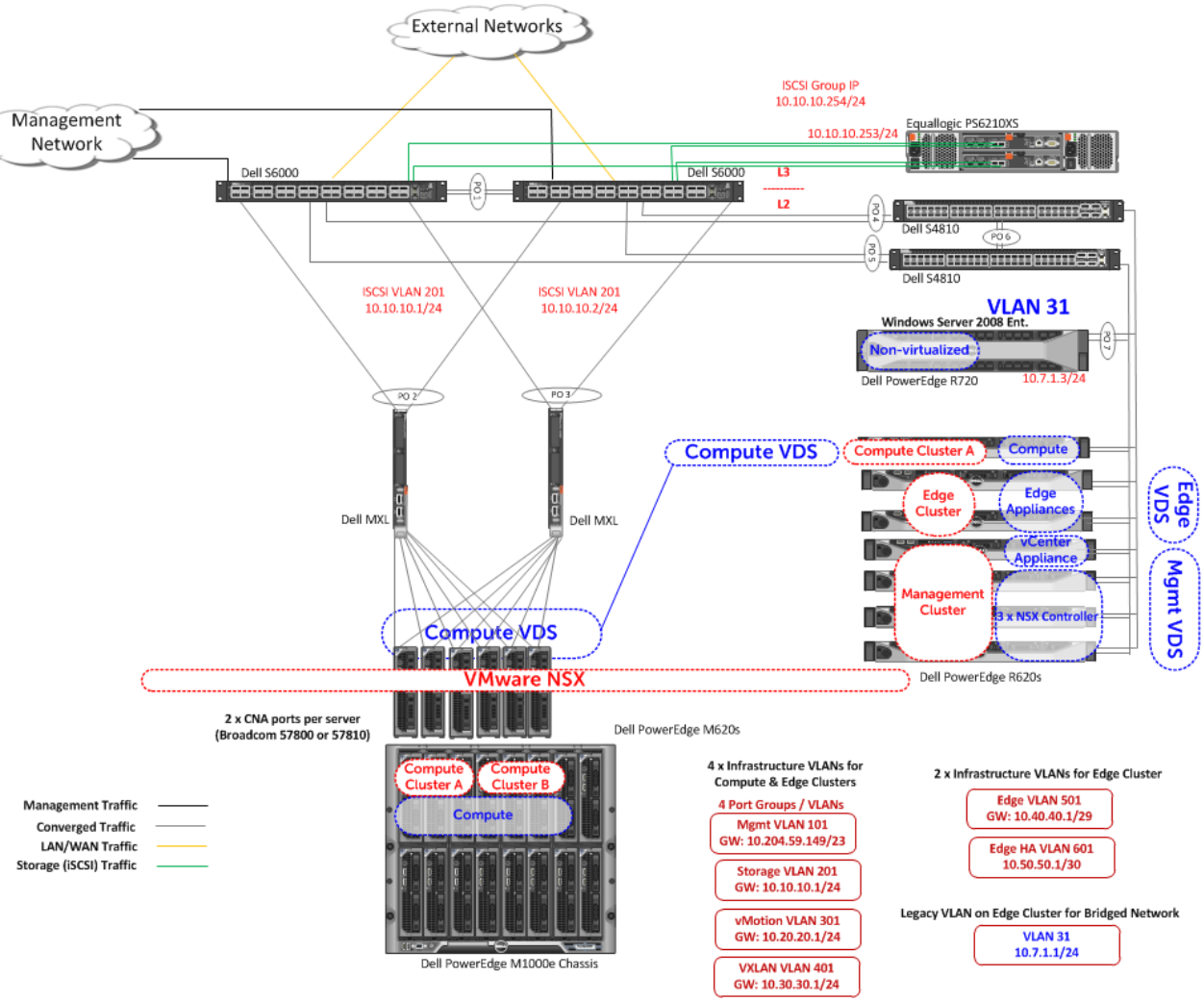

Figure 19 Dell End-to-End Converged Infrastructure with VMware NSX

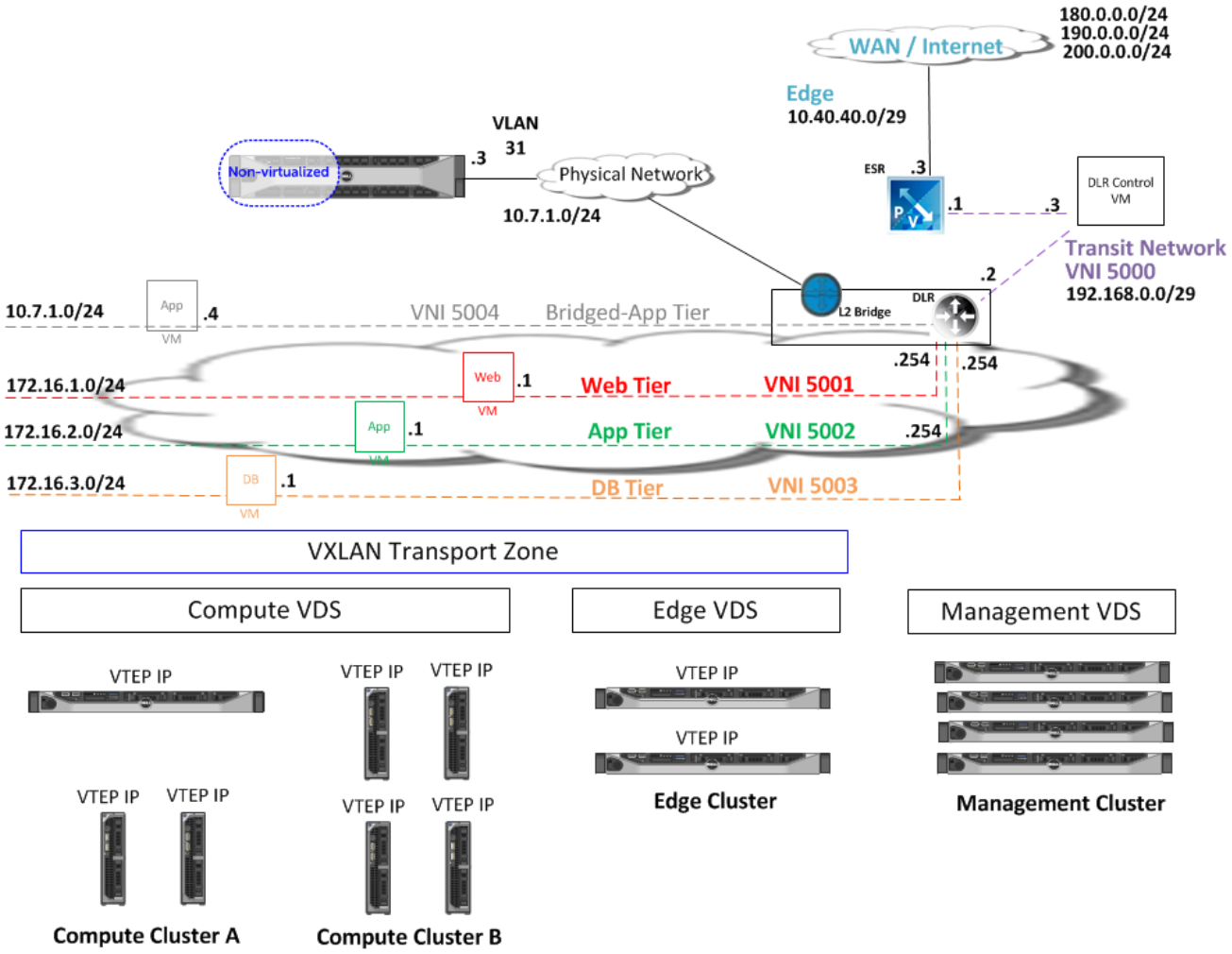

Figure 20 Logical View of Dell End-to-End Converged Infrastructure with VMware NSX

In Figure 20 and Table 2, there are five logical networks as shown in Figure 16 prior: Web Tier, App Tier, DB Tier, Bridged-App Tier, and Transit Network.

Table 2 Logical Networks

Logical Network	Use
Web Tier (VNI 5001)	Hosts web servers.
App Tier (VNI 5002)	Hosts applications with business logic. Web servers in the Web Tier talk to the applications in the App Tier.
DB Tier (VNI 5003)	Hosts databases that store the data relevant to the web apps. The applications on the App Tier talk to the databases in the DB Tier to store and retrieve information.
Bridged-App Tier (VNI 5004)	Logical network (VNI 5004) that bridges to a non-virtualized host on the physical network (VLAN 31). The DLR Control VM is used to provide this L2 Gateway bridging service.
Transit Network (VNI 5000)	Used to peer with the external networks and allows the NSX Edge Services Router and DLR to exchange routing information.

NSX Edge Gateway Services can provide an additional firewall boundary between the Web Tier logical switch, the other logical switches, and the external networks. This is represented by a perimeter firewall service in addition to the distributed firewall capabilities provided on every logical switch. Figure 21 shows how this would logically look if it was desired to utilize the perimeter firewall capabilities.

Note, the NSX Perimeter Edge provides services as a VM as compared to the NSX DLR, the VXLAN, and the Distributed Firewall ESXi kernel modules which provide capabilities within the hypervisor kernel at close to line-rate.

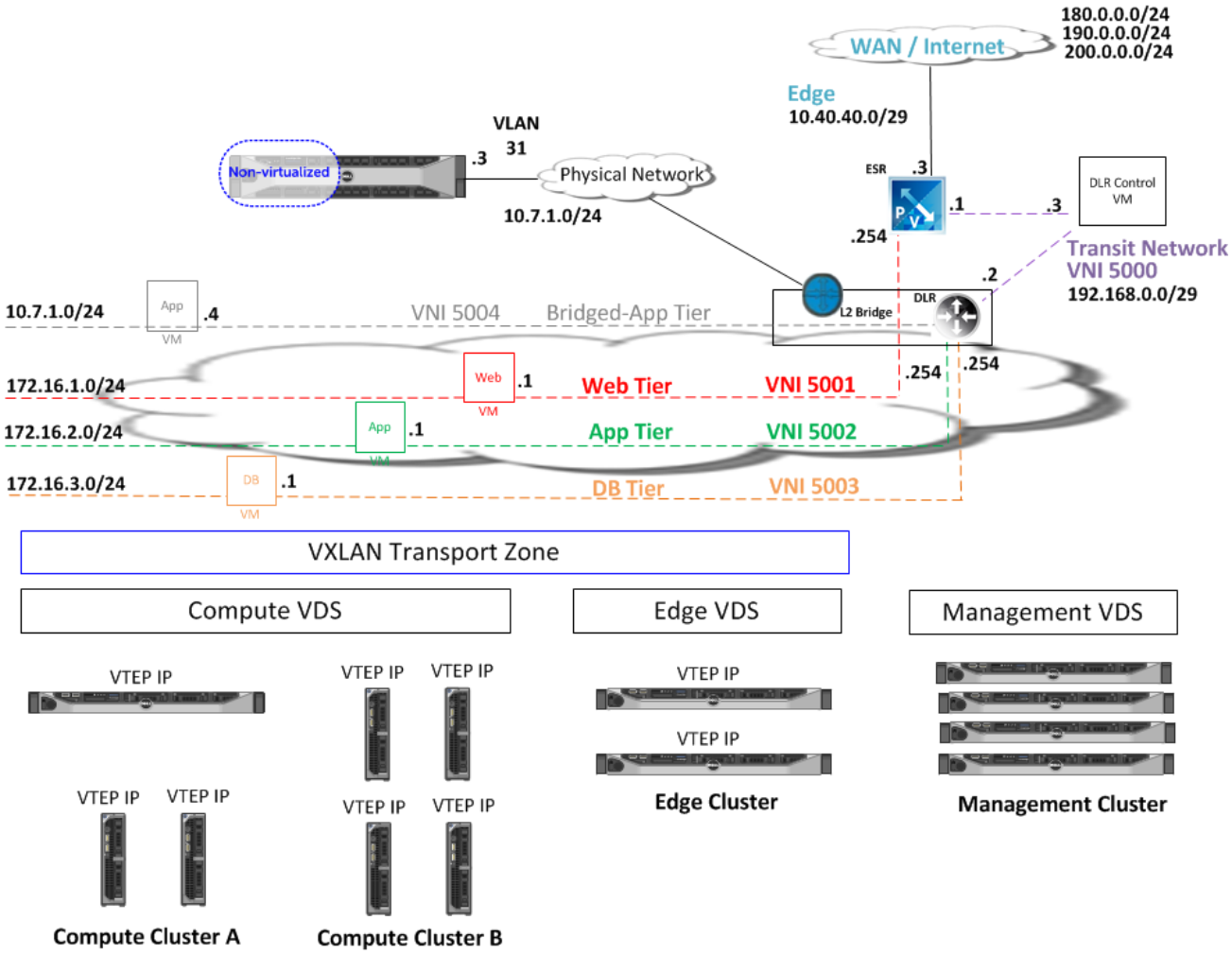

Figure 21 Logical View with the Web Tier logical switch connected to the NSX Perimeter Edge

7 Dell Network Design Options for VMware NSX

In this whitepaper we looked at a completely converged network infrastructure utilizing DCB and iSCSI for storage, and, as such, a layer 2 physical underlay was used. Thus, in this topology, the layer 3 boundary exists on the spine/core switches and not the access/ToR switches. Note, only blade servers are shown from the original diagram to keep the diagram less cluttered while displaying the concept.

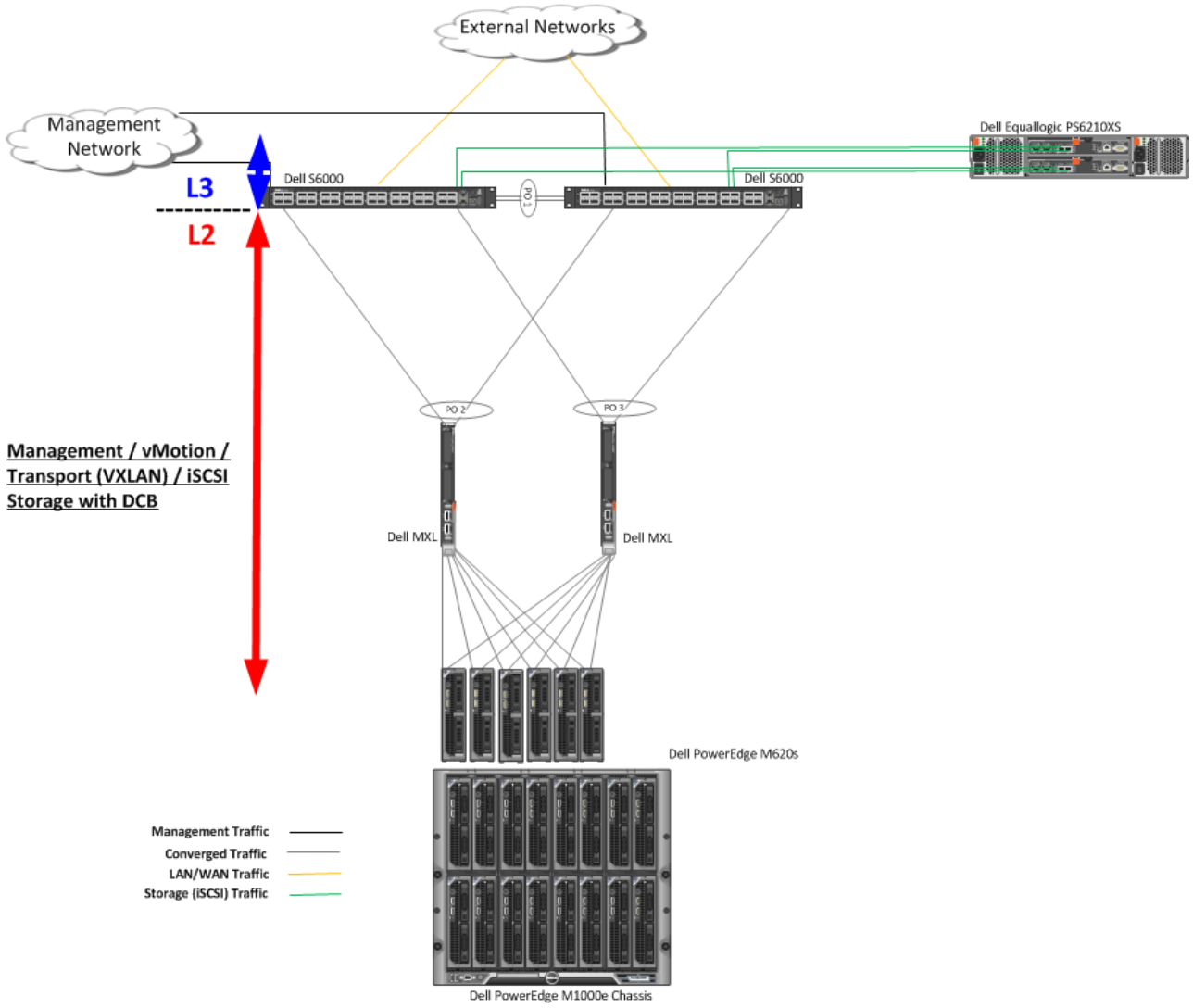

Figure 22 Dell Networking L2 Underlay with iSCSI Storage and DCB

Although a converged network is used as an example and discussed in detail in this whitepaper, if desired, a non-converged network, for example, using a separate Fibre Channel (FC) or iSCSI network, can also be used.

If desired or per requirements, a L3 or hybrid L2/L3 physical underlay can just as easily be used instead of L2. A hybrid L2/L3 approach may use layer 2 just for the DCB and iSCSI with multi-pathing requirements. Note, as mentioned prior, currently ISCSI multipathing with vSphere ESXi requires a non-routed network (see VMware Knowledge Base article: http://kb.vmware.com/kb/1009524). In the hybrid L2/L3 design, all infrastructure traffic except for iSCSI storage traffic is terminated at the access/ToR switch via switched virtual interface (SVI) and routed from that point on. Such a design would look like that shown in Figure 23 below.

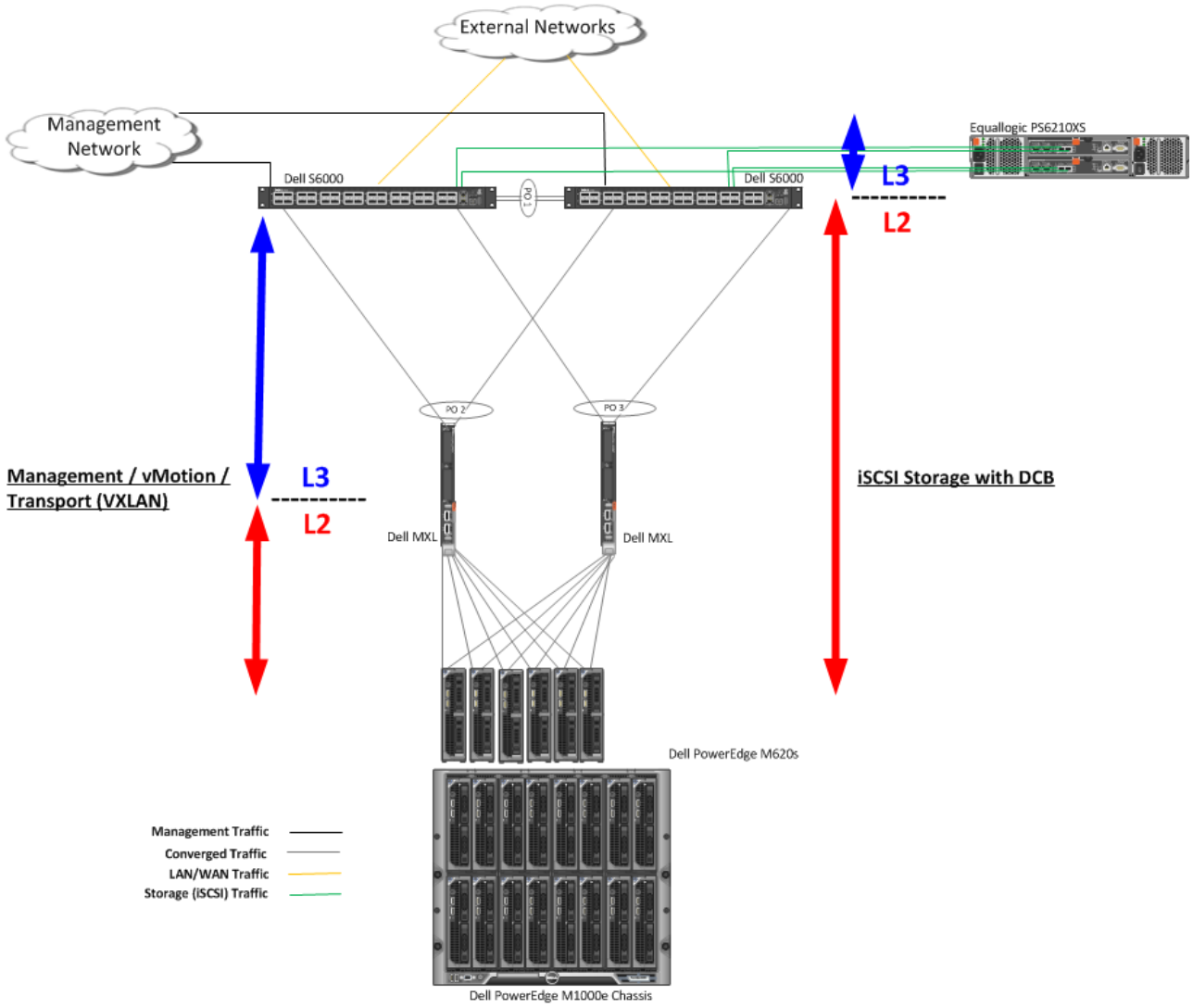

Figure 23 Dell Networking Hybrid L2/L3 Underlay with iSCSI Storage and DCB

A completely L3 physical underlay design with NFS storage provided by Dell Compellent FS8600 is shown in Figure 24 below. In this design choice, the multi-path iSCSI storage option is currently not possible.

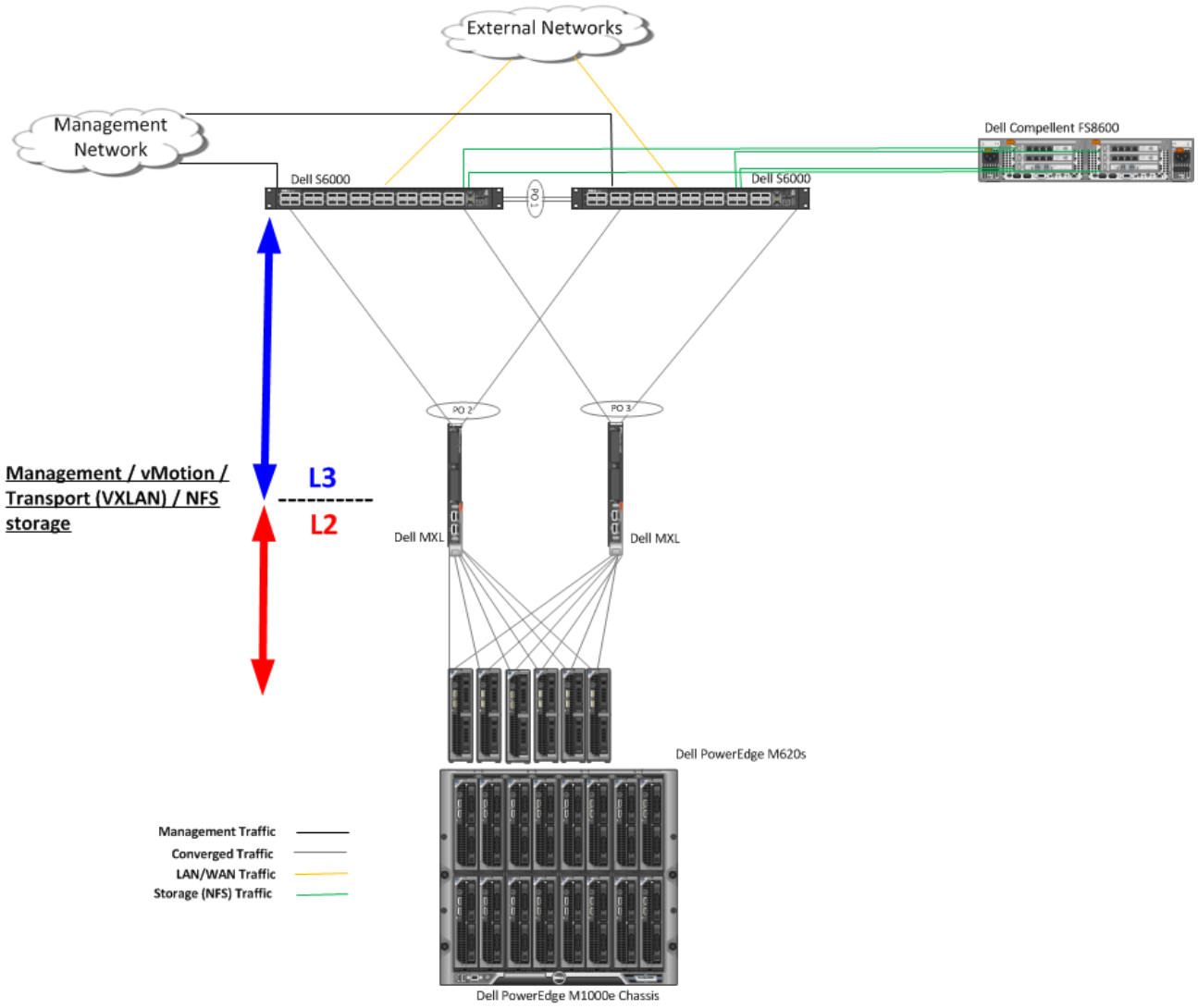

Figure 24 Dell Networking Hybrid L2/L3 Underlay with iSCSI Storage and DCB

8 Additional Network Design and Scaling Considerations

In the example network designs shown in this whitepaper, Dell S6000 switches are used at the spine/core. A Dell S6000 switch provides 32 x 40 GbE ports. With the design shown in this whitepaper, assuming 26 x 40 GbE ports are left after subtracting ports used for management, storage, VLTi, and uplinks, and assuming only blade servers are used, it is possible to scale up to 208 blade servers with a 2:1 oversubscription (using 2 x MXLs with 2 x 40 GbE ports per MXL for uplinks). Using just rack servers, it is possible to scale up to 624 servers with a 3:1 oversubscription (using 4 x 40 GbE ports for uplinks).

Depending on application requirements, if oversubscription is increased, it is possible to scale higher. The number of VMs that can be supported per server will depend on the resources available on the server. Assuming 20-50 VMs per server, the scalability can surpass several thousand VMs.

Figure 25 Dell z9500 switch

A higher density switch at the spine/core such as the Dell Z9500 with 128 x 40 GbE ports will allow for even greater scalability, at which point a complete L3 network underlay design for better scalability should be considered.

9 Conclusion

In this whitepaper, we looked at the benefits of network virtualization with VMware NSX on a reliable and proven Dell infrastructure including Dell servers, networking, and storage. With Dell networking providing a solid physical network underlay, we demonstrated a completely converged infrastructure running VMware NSX network virtualization technology. The unique innovation VMware provides with NSX and the proven and complete end-to-end converged infrastructure Dell provides offers the best-of-class solution for Data Centers and Enterprises looking to save costs and increase productivity with this transformative technology. For more detailed information on deployment of VMware NSX on Dell infrastructure see the Dell Infrastructure with VMware NSX Deployment Guide.