Reference Design:

Deploying NSX for vSphere with Cisco UCS and Nexus 9000 Switch Infrastructure

TECHNICAL WHITE PAPER

Table of Contents

1 Executive Summary

Enterprise data centers are already realizing the tremendous benefits of server and storage virtualization solutions to consolidate infrastructure, reduce operational complexity, and dynamically scale application infrastructure. However, the data center network has not kept pace and remains rigid, complex, and closed to innovation—a barrier to realizing the full potential of virtualization and the software defined data center (SDDC).

VMware NSX network virtualization delivers for networking what VMware has already delivered for compute and storage. It enables virtual networks to be created, saved, deleted, and restored on demand without requiring any reconfiguration of the physical network. The result fundamentally transforms the data center network operational model, reduces network provisioning time from days or weeks to minutes and dramatically simplifies network operations.

This document provides guidance for networking and virtualization architects interested in deploying VMware NSX for vSphere for network virtualization with Cisco UCS (Unified Computing System) blade servers and Cisco Nexus 9000 Series switches. It discusses the fundamental building blocks of NSX with VMware ESXi (the enterprise-class hypervisor), recommended configurations with Cisco UCS, and the connectivity of Cisco UCS to Nexus 9000 switches.

2 Scope and Design Goals

This document assumes readers have a functional knowledge of NSX and deploying Cisco UCS and Nexus 9000 series infrastructure. Readers are strongly advised to read the design guide below for additional context; it provides a detailed characterization of NSX operations, components, design, and best practices for deploying NSX.

VMware® NSX for vSphere Network Virtualization Design Guide

Specifically, the goal of this document is to provide guidance for running NSX with Cisco UCS Blade Servers and Cisco Nexus 9000 series switches deployed as traditional switches with either layer-2 or layer-3 topology. The document covers three critical aspects of the design:

- Connectivity requirements for NSX including VMkernel networking, VLAN allocation and configuration, VXLAN Tunnel End-Point (VTEP) configuration, and layer-3 peering and routing configurations
- Cisco Nexus 9000 connectivity options with NSX in a virtual Port Channel (vPC) or non-vPC mode
- Cisco UCS blade servers running ESXi with NSX connectivity options, VTEP configurations and Virtual NIC (vNIC) configurations

2.1 NSX VMkernel Networking Requirements

In a traditional ESXi environment three infrastructure VLANs are provisioned. The VLANs that are defined on the VMkernel interface are shown in the table below:

Table 1: Infrastructure Traffic Types and VLAN

Traffic Types	Functions	VLAN ID
Management	ESXi and NSX management plane	100
vMotion	VM mobility	101
IP Storage VLAN	Applications & infrastructure data store connectivity	102

The NSX introduces an additional infrastructure VLAN that provides single bridge domain for VM guest traffic carried over physical network.

VXLAN Transport Zone VLAN: During the NSX configuration phase an additional VMkernel interface is created for VXLAN traffic. Overall, each host is prepared with four VMkernel networks that are presented to Cisco UCS as well as Nexus 9000 infrastructure. These VLANs are trunked to the Nexus 9000 access-layer switch. Configuring these four VLANs is a one-time task. This allows the logical networks to be created independently from the physical network, eliminating the need to define the VLAN every time a new logical segment is added to accommodate VM growth. The VLAN Switch Virtual Interface (SVI) termination is either at the aggregation layer or at the access layer, depending on the topology deployed with Nexus 9000 physical network.

Table 2: VXLAN VLAN for VM Guest Traffic

Traffic Type	Function	VLAN ID
VXLAN Transport Zone VLAN	Overlay VXLAN VTEP Connectivity	103

Additional VLANs are needed for:

- **L3 ECMP Connectivity:** Two VLANs are typically required for allowing north-south traffic from the NSX domain to the physical world.
- **Bridging:** Optionally, NSX supports VXLAN-to-VLAN bridging for P-V or V-V connectivity. The number of VLAN requirements will vary based on the instances of bridging desired.

2.1.1 Demystifying the VXLAN and IP Connectivity

VXLAN decouples the connectivity for the logical space from the physical network infrastructure. Devices connected to logical networks can leverage the entire set of network services (load balancer, firewall, NAT, etc.) independently from how the underlying physical infrastructure is configured. This helps solve many of the challenges of traditional data center deployments, such as agile & programmatic application deployment, vMotion across layer-3 boundaries, as well as multi-tenancy support to overcome the VLAN limitation of 4094 logical segments.

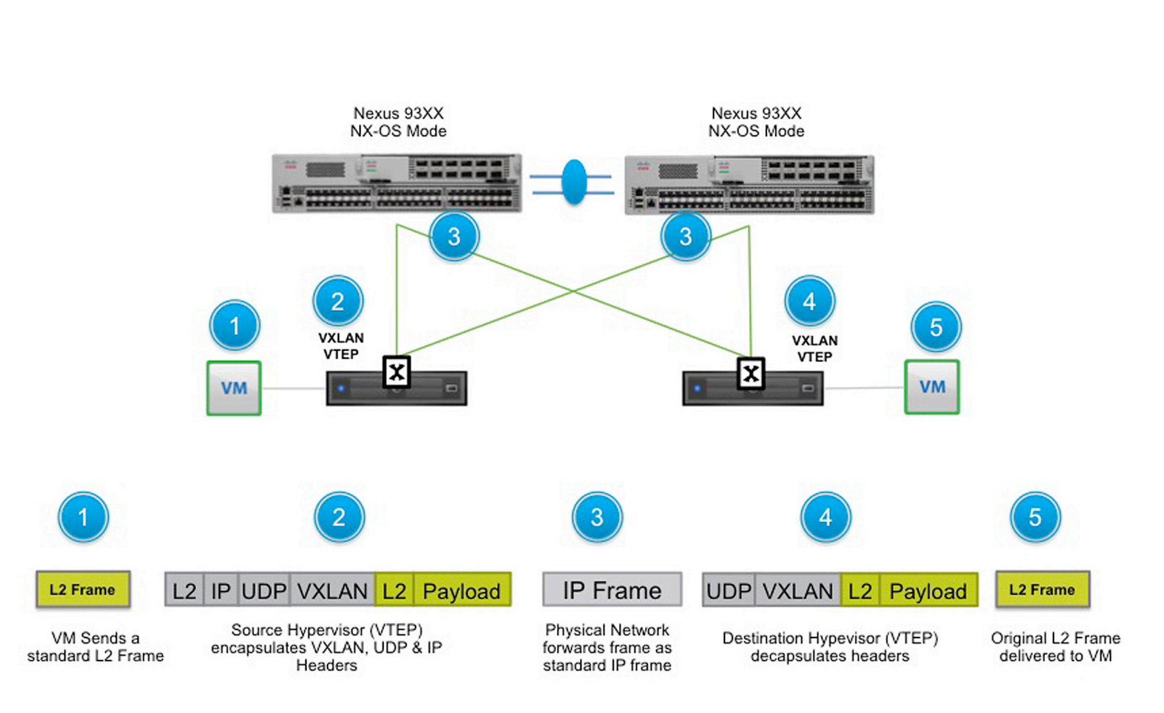
NSX VXLAN Capabilities

NSX VXLAN implementation offers critical enhancements to VXLAN:

- NSX enables multicast free VXLAN with the help of the controller. Removing multicast from the underlay network greatly simplifies physical network configuration. The distributed discovery and efficient management of control plane functions (MAC, VTEP and ARP table management) are relegated to highly available clustered controllers.
- NSX enables VXLAN encapsulation at the kernel level in the ESXi host. This VXLAN encapsulated frame is nothing but a generic IP packet, which is then routed by Nexus 9000 switch forwarding traffic, based on the destination VTEP IP address. In addition, the underlay does not see the explosion of MAC addresses or the intensive configuration requirement for ad-hoc connectivity in the virtual domain.

The net effect of these enhancement is shown in below figure where ESXi encapsulate/decapsulate the VXLAN header with multicast-free replication of BUM (Broadcast Unknown Multicast) traffic.

Figure 1: VXLAN Encapsulation & Decapsulation at ESXi Kernel


Nexus 93XX
NX-OS Mode
Nexus 93XX
NX-OS Mode
3
3
1
2
VXLAN
VTEP
4
VXLAN
VTEP
5
VM
VM
1
2
3
4
5
L2 Frame
L2 IP UDP VXLAN L2 Payload
IP Frame
UDP VXLAN L2 Payload
L2 Frame
VM Sends a standard L2 Frame
Source Hypervisor (VTEP) encapsulates VXLAN, UDP & IP Headers
Physical Network forwards frame as standard IP frame
Destination Hypevisor (VTEP) decapsulates headers
Original L2 Frame delivered to VM

These enhancements to VXLAN simplify the underlay physical network. For additional details about VXLAN, packet flow for various layer-2 control plane discovery, and connectivity, please refer to the **VMware® NSX for vSphere Network Virtualization Design Guide.**

2.1.2 VXLAN and VDS Connectivity with Cisco UCS and Nexus 9000

VXLAN connectivity consists of two components: transport zone and VTEP. The transport zone is a collection of ESXi clusters participating in a single VXLAN domain. VTEP (VXLAN Tunnel End Point) is a logical interface (VMkernel) that connects to the transport zone for encapsulation/decapsulation of VM guest traffic as shown in Figure 1.

For a given cluster, only one VDS is responsible for VXLAN connectivity. The cluster design in section below goes in details of VDS design recommendation. However, there are two critical design requirements for VXLAN connectivity: VLAN ID for VXLAN, and VDS uplink Configuration.

- **VLAN ID for VXLAN:** At the NSX configuration phase, the VTEP(s) are defined with transport zone VLAN ID; this VLAN port-group is dynamically created during the cluster VXLAN preparation. For a VLAN that supports VXLAN transport zone, a specific configuration for a VLAN ID is required based on the physical topology. NSX requires the VDS dvUplink configuration to be consistent per VDS and thus for VLAN ID for the VXLAN transport zone has to be the same regardless of layer-2 or layer-3 topology. The detailed configuration VLAN ID mapping to a specific topology is described in section 2.2.3.

- **VDS Uplink Configuration:** The NSX creates a dvUplink port-group for VXLAN that must be consistent for any given VDS and NIC teaming policy for VXLAN port-group must be consistent across all hosts belonging to the VDS. Typically one VTEP is sufficient; however multiple VTEPs are also supported. The number of VTEP(s) supported is determined by a combination of the number of host interfaces exposed to VDS and uplink teaming mode configurations as shown in the table below.

Table 3: VDS Teaming Mode and NSX Support

Teaming and Failover Mode	NSX Support	Multi-VTEP Support	Uplink Behavior 2 x 10G	Nexus 9xxx Port Configurations
Route Based on Originating Port	✓	✓	Both Active	Standard
Route Based on Source MAC Hash	✓	✓	Both Active	Standard
LACP	✓	×	Flow based	vPC Port-Channel - LACP
Route Based on IP Hash (Static EtherChannel)	✓	×	Flow based	vPC Port-Channel – LACP mode OFF
Explicit Failover Order	✓	×	Only one link is active	
Route Based on Physical NIC Load (LBT)	×	×	×	Standard

As one can notice from Table 3, selecting LACP or Static EtherChannel teaming mode limits the choice for selecting team-modes per traffic types (port-groups for management, vMotions, VXLAN). With LACP or Static EtherChannel, only one VTEP per host can be configured. Any other teaming modes allow the flexibility to choose the behavior of failover or load sharing per traffic type. The only exception is that LBT (Load Based Teaming) mode is not supported for VTEP VMkernel.

The table above also shows the port-configuration mode for Nexus 9000 switches relative to the uplink teaming mode. Notice that LACP and Static EtherChannel modes require VLAN based vPC (Virtual Port-Channel) and can only support a single VTEP. The LACP mode is also not possible with Cisco UCS blade server environment due lack of support on the server-side LACP on Fabric Interconnect.

2.2 Nexus 9000 Connectivity Options with NSX

This section covers the details of connectivity requirements for various NSX components and clusters required for integration with Cisco Nexus 9000 series switches. Supporting NSX for Cisco Nexus 9000 switches is as simple as following three basic requirements:

1. Support for jumbo frame
2. Configurations supporting IP forwarding, including SVI configuration and routing support
3. VLAN configuration requirements based on physical network topology

Advice for supporting VLANs, SVI, and jumbo frame can be found in the following Cisco 9000 configuration guide:

http://www.cisco.com/c/en/us/support/switches/nexus-9000-series-switches/products-installation-and-configuration-guides-list.html

2.2.1 Cisco Nexus 9000 Jumbo Frame Configurations

Cisco Nexus 9000 switches support jumbo frame; however it is not enabled by default. The jumbo frame configuration steps are different for layer-2 and layer-3 interfaces.

Configuration steps for layer-2 interface

Change the system jumbo MTU to 9214 with the "system jumbomtu 9214" global command. The reason: one can only set MTU to default value (1500 Bytes) or the system-wide configured value. Then change MTU to 9214 on each layer-2 interface with the "mtu 9214" interface command.

Configuration steps for layer-3 interface

Change MTU to 9214 on each layer-3 interface via the "mtu 9214" interface command. Sample CLI is show for each type of interface in below table:

Table 4: Nexus 900 MTU CLI Configurations

Layer 2 Interface	Layer 3 Interface
system jumbomtu 9214 ← Global configurations *interface Ethernet1/9* *description to esx-vmnic3-VMK* *switchport mode trunk* *switchport trunk allowed vlan 22-25* *spanning-tree port type Edge trunk* ***mtu 9214 ← Layer 2 MTU*** *channel-group 9 mode active*	***interface Vlan151 ← SVI Interface*** *no ip redirects* *ip address 10.114.221.34/27* *no ipv6 redirects* *hsrp 1* *ip 10.114.221.33* *description VXLAN Transport Zone* *no shutdown* ***mtu 9214*** *interface Ethernet2/12 ← Layer 3 point-to-point Interface* *description L3 Link to Spine* *no switchport* *speed 40000* *duplex full* ***mtu 9214*** *ip address 10.114.211.117/31* *no shutdown*

2.2.2 Configuration Support for IP Forwarding

The IP forwarding feature requires defining the SVI interface with an IP address and enabling the routing protocol of choice (OSPF - Open Shortest Path First or BGP - Border Gateway Protocol). The SVI configuration is also enabled with the First Hop Redundancy Protocol (FHRP) to provide redundancy for ToR failure. The routing protocol configuration and its interaction with NSX routing domains is further described in Edge cluster connectivity in section 2.3.3.

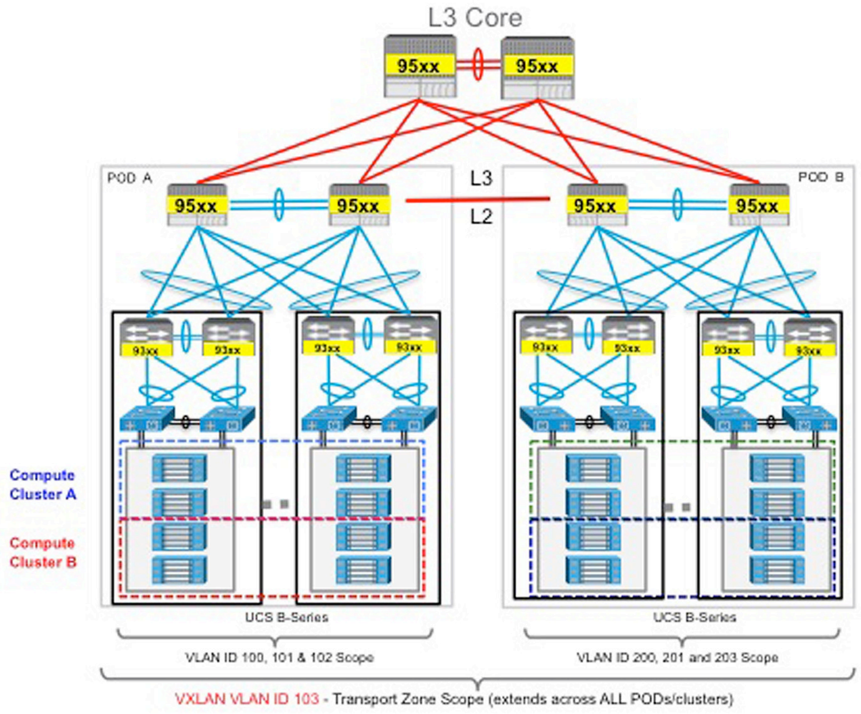
2.2.3 VLAN Connectivity Requirement

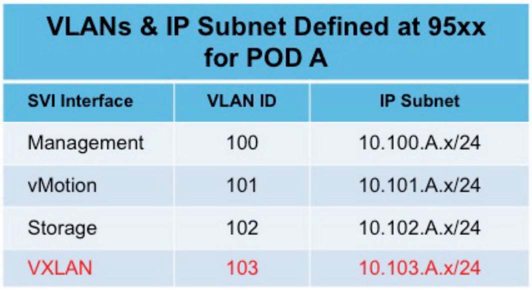
Each form of infrastructure traffic described in Table 1 requires a separate VLAN ID and SVI configuration. As discussed in section 2.1.2, VXLAN connectivity requires further clarification on VLAN ID configurations with layer-2 and layer-3 topology. Cisco Nexus 9000 series switches are capable of supporting multiple types of topologies. They can support classical network topologies as well spine-leaf topologies.

Classical Access/Aggregation/Core DC Network

As shown in the figure below, Nexus 93xx ToR is typically deployed in the access layer and Nexus 9500 switches are deployed as modular chassis at the aggregation layer. The access layer switches are connected to aggregation switch with a single vPC topology. The scale-out model requires a POD design where each POD is a distinct layer-2 domain. The demarcation of the layer-3 boundary starts at the aggregation layer. Without NSX, the VLANs are confined to each POD boundary, which is also the vMotion boundary. There is a unique VLAN-to-subnet mapping local to POD.

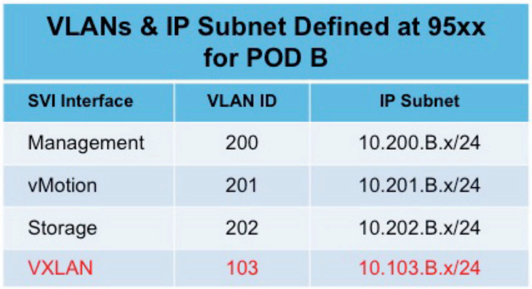
Figure 2: Layer-2 POD Design with VXLAN





VLANs & IP Subnet Defined at 95xx for POD A

SVI Interface	VLAN ID	IP Subnet
Management	100	10.100.A.x/24
vMotion	101	10.101.A.x/24
Storage	102	10.102.A.x/24
VXLAN	103	10.103.A.x/24



VLANs & IP Subnet Defined at 95xx for POD B

SVI Interface	VLAN ID	IP Subnet
Management	200	10.200.B.x/24
vMotion	201	10.201.B.x/24
Storage	202	10.202.B.x/24
VXLAN	103	10.103.B.x/24

Typically in a layer-2 topology, the VLAN ID only has to be unique to the layer-2 domain. In the diagram above, two distinct layer-2 PODs each have locally unique VLAN ID. However, the VXLAN transport zone, which defines the scope of VXLAN enabled cluster, spans both PODs. This implies that VLAN ID for VXLAN has to be the same for both the PODs. In other words one would map the VLAN designated for VXLAN with two different subnets for the same VLAN ID; however at a different aggregation boundary for each POD. This is depicted in the above figure with VLAN ID for VXLAN being 103 extending both pods however the subnet that it maps to is unique at aggregation layer. This multi-POD case is similar to a spine-leaf routed data center design. The only difference is that in spine-leaf routed DC layer-3 demarcation starts at the access layer, which is discussed next.

Leaf-Spine Design (Routed DC Network)

The **VMware® NSX for vSphere Network Virtualization Design Guide** goes into detail on leaf-spine topology attributes and hence is not discussed here since most of the recommendations apply to Nexus 9000 switches. In leaf-spine design, the layer-3 is terminated at the ToR and thus all the VLANs originating from ESXi hosts terminate on Nexus 9000 ToR. Typically in layer-3 design this means the VLAN ID is irrelevant and can be kept unique or the same for a type of traffic per rack, as long as the VLAN ID maps to the unique subnet.

Figure 3: Layer-3 POD Design with VXLAN

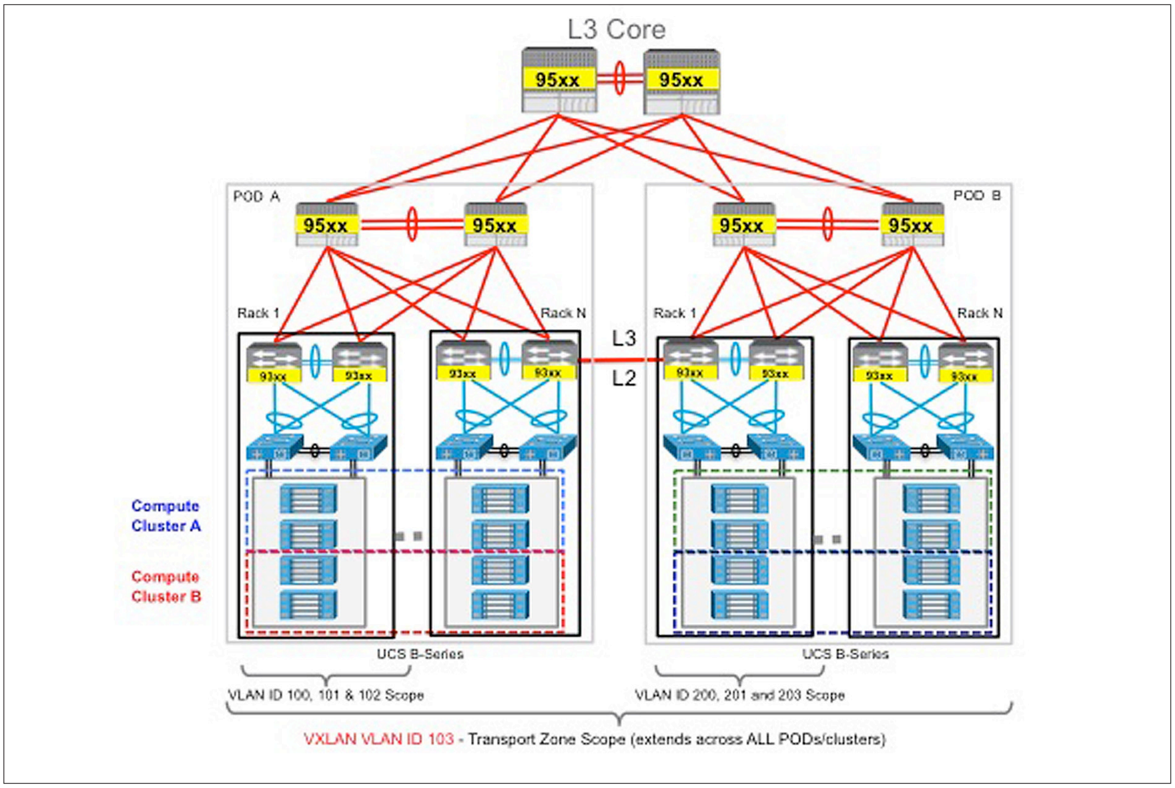


Table 5: SVI to VLAN Mapping for Layer-3 POD Design

VLANs & IP Subnet Defined at each ToR		
SVI Interface	VLAN ID	IP Subnet
Management	100	10.100.R_ID.x/24
vMotion	101	10.101.R_ID.x/24
Storage	102	10.102.R_ID.x/24
VXLAN	103	10.103.R_ID.x/24

However, the exception is the selection of the VLAN ID configured for the given VDS VXLAN transport zone. The VLAN ID must be the same for VXLAN VTEP for each rack/ToR and map to the unique subnet. In other words, for the VXLAN VTEP, the VLAN ID remains the same for every ToR; however, the subnet that maps to VLANs is unique per ToR. One can keep the VLAN ID for the rest of the traffic types to be the same for every rack. This simplifies the configuration for every rack and only requires configuration once. As an example, this is depicted in the table above, which can be repeated for each ToR configuration with unique subnet identified by rack ID.

2.3 NSX Components and Cluster Connectivity

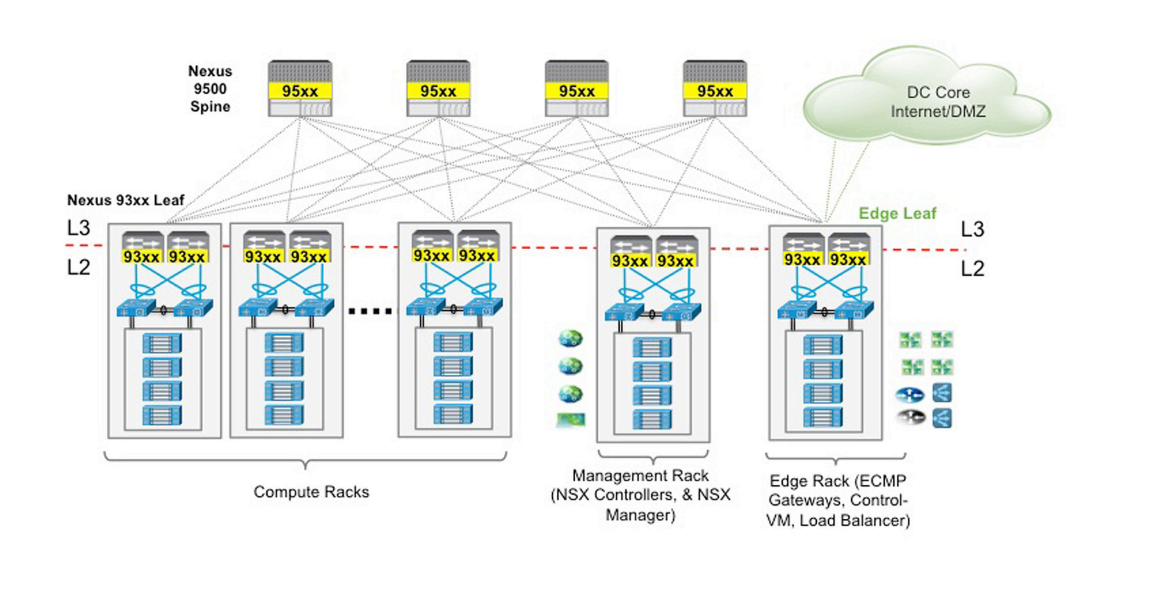
The NSX functions and component operation are defined in the **VMware® NSX for vSphere Network Virtualization Design Guide**. The reader is strongly advised to read the document in order to follow the rationale regarding connectivity to physical network. The NSX components are categorized in following table. The NSX components organization and functions are mapped to appropriate cluster. The **VMware® NSX for vSphere Network Virtualization Design Guide** calls for organizing NSX components, compute, and management of the virtualized environment. This organization principle is carried in the document and repeated to maintain ease of user readability.

Table 6: NSX Functions and Components Mapping to Cluster Type

Function	NSX Components	Recommended Clusters Designation
Management Plane	NSX Manager & vCenter Connectivity	Management Cluster
Control Plane	NSX Controller Cluster	Management Cluster* *Can be in Edge Cluster
	Logical Routers Control VM	Edge Cluster
Data Plane East-West	Compute and Edge VDS kernel components – VXLAN forwarding & DLR (Distributed Logical Router)	Compute & Edge Cluster
Data Plane North-South	Edge Service Gateway (ESG)	Edge Cluster
Bridging Traffic	DLR Control VM	Edge Cluster

The **VMware® NSX for vSphere Network Virtualization Design Guide** recommends building three distinct vSphere cluster types. The figure below shows an example of logical components of cluster design to the physical rack placement.

Figure 4: Mapping Cluster Types to Functions



As shown in the diagram, edge and management clusters are distributed to separate physical racks and connect to separate ToR switches. For management and edge clusters, the resources are shared or split between two racks to avoid any single rack failure. This also enables scaling.

Note that for even in smaller configurations a single rack can be used to provide connectivity for the edge and management cluster. The key concept is that the edge cluster configuration is localized to a ToR pair to reduce the span of layer-2 requirements; this also helps localize the egress routing configuration to a pair of ToR switches. The localization of edge components also allows flexibility in selecting the appropriate hardware (CPU, memory and NIC) and features based on network-centric functionalities such as firewall, NetFlow, NAT and ECMP routing.

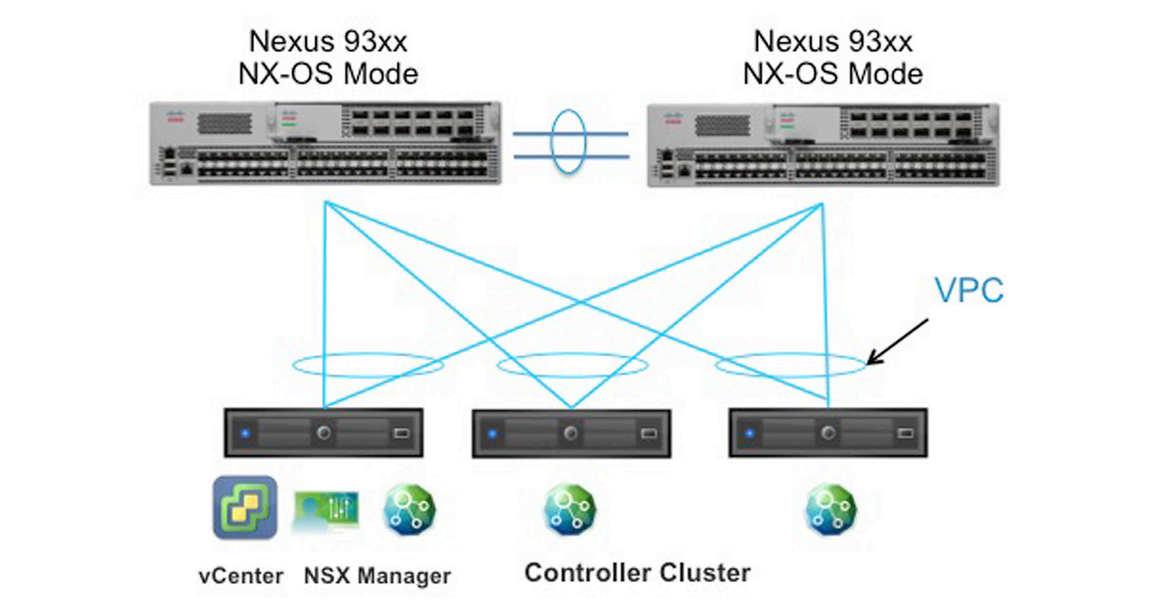
In order to provide a recommendation on connecting host belonging to different cluster types it is important to know the VDS uplink design option as well as the Nexus 9000 capability support. These capabilities are described in section 2.1.2.

The **VMware® NSX for vSphere Network Virtualization Design Guide** best practices document calls for a separate VDS for compute and edge cluster. This enables flexibility of choosing VDS uplink configuration mode per cluster type. It is important to note that the guidelines provided below supersede the **VMware® NSX for vSphere Network Virtualization Design Guide** guideline in some cases as these recommendations apply only to Cisco Nexus 9000 switches.

2.3.1 Management Cluster Connectivity

The management cluster consists of hosts supporting multiple critical virtual machines and virtual appliances. The NSX manager VM and controllers also typically deployed in management clusters requiring high availability (surviving the failure of the host or ToR/uplink). Typically, the management cluster is not prepared for VXLAN and thus connectivity from the ESXi host is VLAN based port-group on a separate VDS. In order to achieve maximum availability and load sharing, LACP teaming mode is typically recommended. Thus, the Nexus 9000 switch ports connecting to management hosts require LACP. For Nexus 9000 switches, this is achieved by enabling traditional layer-2 VLAN-based vPC (Virtual Port Channel). Typically, all the traffic types including management, vMotion, and IP storage are carried over LACP.

Figure 5: Nexus 9000 vPC Connectivity for Management Cluster



2.3.2 Compute Cluster Connectivity

NSX offers a clear departure from the traditional methods, in which the VLANs are only defined once for infrastructure traffic (VXLAN, vMotion, storage, management). The VM connectivity is defined programmatically without relying on the physical network as described in section 3 below. This decoupling enables a repeatable rack design where physical planning (power, space, cooling and cookie-cutter switch configuration) is streamlined. The physical network only requires robust forwarding and adequate bandwidth planning.

The compute cluster requires the most flexibility as it carries multiple types of traffic. Each type of traffic can have its own service level. For example, the storage traffic requires the lowest latency, as opposed to vMotion, which may require higher bandwidth.

Some workloads may have many sources and destinations and require granular load sharing by using multiple VTEPs. The flexibility of selecting teaming mode per traffic type and allowing multiple VTEPs for VXLAN (as described in the VDS uplink configuration section) are two primary reasons for *not* recommending LACP for the compute cluster host's connectivity to Nexus 9000 switches. This recommendation is also followed in Cisco UCS connectivity for NSX as described below.

UCS Connectivity with NSX

The UCS connectivity and configuration is well described in the NSX+Cisco Nexus 7000/UCS Design Guide. The key connectivity criteria are described below:

- The NSX deployed on a Cisco UCS server carries all the VLANs on both fabrics
- The Cisco Fabric Interconnect runs in end-host mode; fabric mode is fabric failover
- The host traffic types are mapped to an active-standby fabric failover vNIC connectivity
- The uplinks from Cisco UCS Fabric Interconnects have vPC connectivity to Nexus 9000 switches to provide loop-free topology
- vNICs (UCS logical interface) are either dedicated or shared based on bandwidth and performance isolation requirements
- VDS uplinks are equal to the number of VIC adapters installed per UCS blade
- The VDS uplink teaming mode cannot use LACP (its an orphaned connection) since fabric interconnect does not support server side LACP nor it support pass-through
- The NSX supports multiple VTEPs depending upon the uplink-teaming mode; typically one VTEP is sufficient, however for the Cisco UCS Blade Server multiple VTEPs are recommended with each VTEP mapping to different vNICs and fabrics with active-standby failover configurations

Figure 6: NSX VMkernel Mapping to UCS vNIC

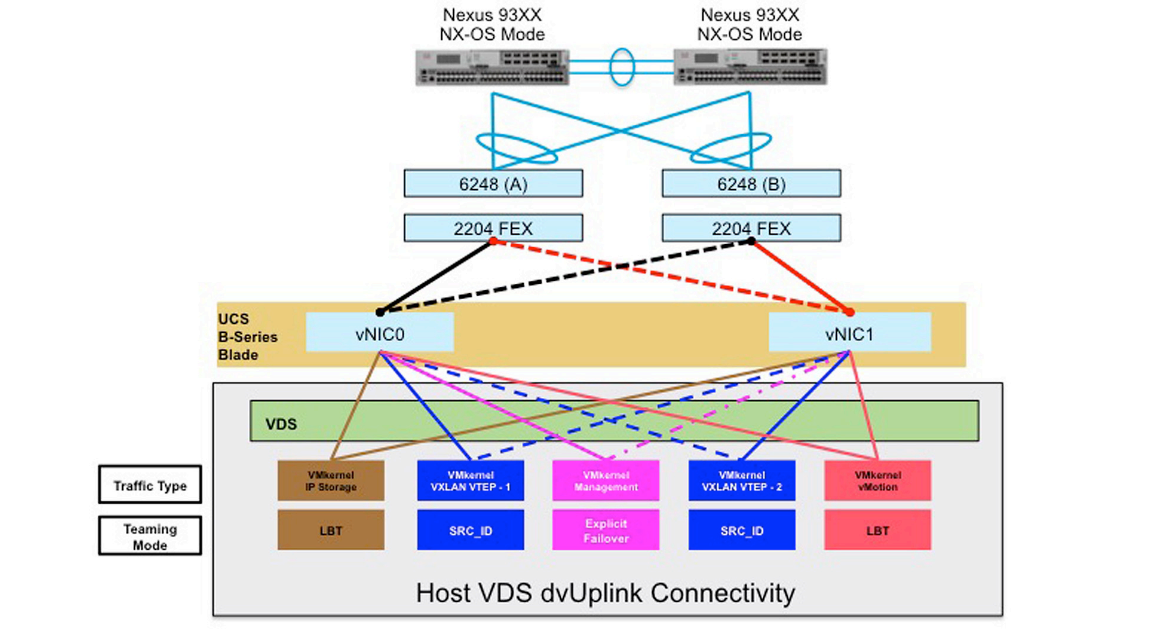


Table 5: VTEP Mapping to VLAN and dvUplink

Port Group	VLAN	dvUplink 1	dvUplink 2	Load Balancing
VTEP 1	301	Active	Standby	SRC_ID
VTEP 2	301	Standby	Active	SRC_ID

The connectivity recommendation described above also applies to edge cluster Cisco UCS blade servers, simplifying the connectivity as well as configuration variation.

2.3.3 Edge Cluster Connectivity

NSX ESG (Edge Services Gateway) is a multi-function VM, enabling services such as north-south routing, firewall, NAT, load balancing, and SSL-VPN. The capabilities and features are beyond the scope of this paper. Please refer to **VMware® NSX for vSphere Network Virtualization Design Guide**. This section covers necessary technical details that are pertinent to physical and logical connectivity required. The critical functions provided by the edge cluster hosting multiple edge VMs are:

- Providing on-ramp and off-ramp connectivity to physical networks (north-south L3 routing delivered by NSX edge virtual appliances)
- Allowing communication with physical devices connected to VLANs in the physical networks (NSX L2 bridging provided via logical control VMs)
- Supporting centralized logical or physical services (firewall, load-balancers, and logical router control VM, etc.)

The benefits of confining edge clusters to a pair of ToRs (or pair of racks) include:

- Reducing the need to stretch VLANs across compute clusters
- Localizing the routing configuration for N-S traffic, reducing the need to apply any additional configuration knobs for N-S routing on the compute ToRs
- Allowing network admins to manage the cluster workload that is network centric (operational management, BW monitoring and enabling network-centric features such as NetFlow, security, etc.)

2.3.3.1 Nexus 9000 and NSX Routing

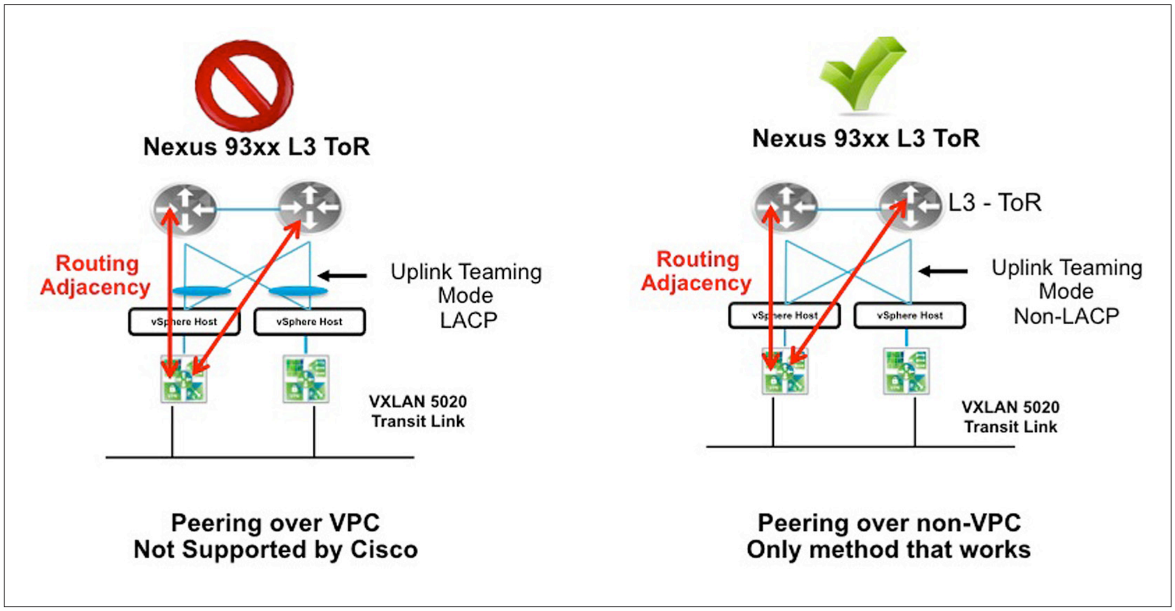
This paper addresses connectivity for north-south routing with the ECMP mode of the edge services gateway. The NSX edge gateway provides ECMP (Equal Cost Multi-path) based routing, which allows up to eight VMs presenting 8-way bidirectional traffic forwarding from NSX logical domain to the enterprise DC core or Internet. This represents up to 80 GB of traffic that can be offered from the NSX virtual domain to the external network in both directions. It's scalable per tenant, so the amount of bandwidth is elastic as on-demand workloads and/or multi-tenancy expand or contract. The configuration requirements to support the NSX ECMP edge gateway for N-S routing is as follows:

- VDS uplink teaming policy and its interaction with ToR configuration
- Requires two external VLAN(s) per pair of ToR
- Route Peering with Nexus 9000

2.3.3.2 VDS uplink design with ESXi Host in Edge cluster

The edge rack has multiple traffic connectivity requirements. First, it provides connectivity for east-west traffic to the VXLAN domain via VTEP; secondly, it provides a centralized function for external user/traffic accessing workloads in the NSX domain via dedicated VLAN-backed port-group. This later connectivity is achieved by establishing routing adjacencies with the next-hop L3 devices. The figure below depicts two types of uplink connectivity from host containing edge ECMP VM.

Figure 7: Nexus 9000 Layer 3 Peering over vPC Support



As of this writing, Nexus 9000 switches do not support routing over vPC. Therefore, the recommendation for edge clusters is to select either the Explicit Failover Order or the SRC-ID as teaming options for VDS dvUplink. This will allow the edge cluster to establish a routing peer over a selected dvUplink along with load sharing per ECMP edge VM to a dvUplink. Please refer to below URL or latest release URL for an additional topology and connectivity options with Nexus 9000 switches:

Configuring vPCs

In addition, the non-LACP uplink-teaming mode allows the multiple-VTEPs configuration recommended with Cisco UCS blade servers.

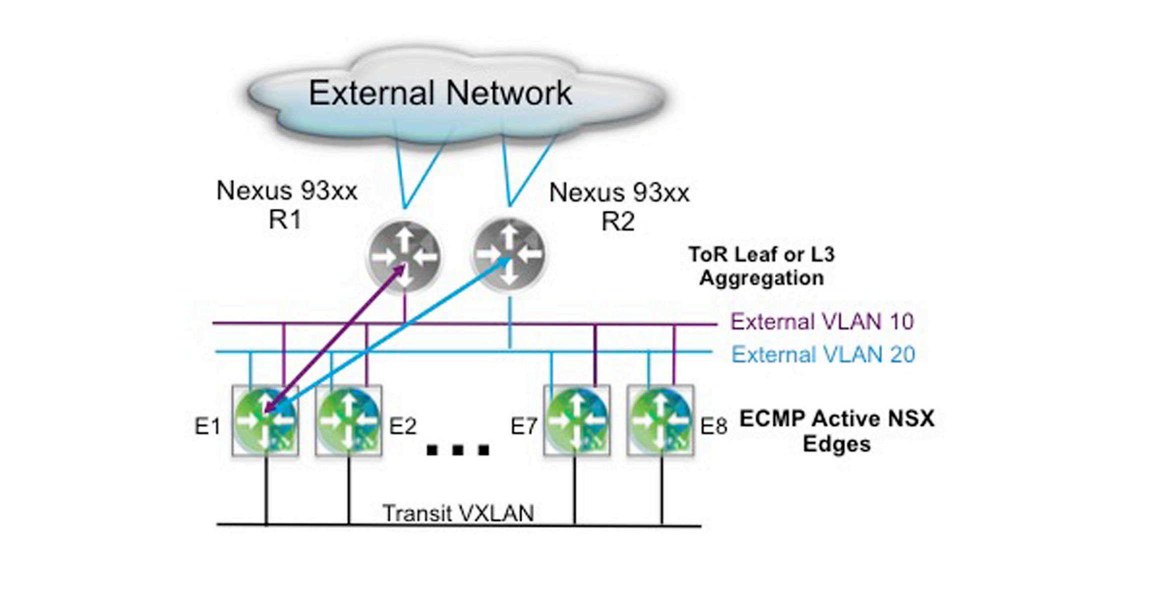
2.3.3.3 Edge ECMP Peering and VLAN Design

Once the uplink-teaming mode is determined, the next step is to provide design guidance around VLAN configuration and mapping to uplink as well peering to Nexus 9000 switches.

The first decision is how many logical uplinks should be deployed on each NSX edge. The recommended design choice is to always map the number of logical uplinks defined on NSX edge VM to the number of VDS dvUplinks available on the ESXi servers hosting the NSX edge VMs. This means always map a VLAN (port-group) to a VDS dvUplink, which then maps to a physical link on the ESXi host that connects to the Nexus 9000 switch, over which an edge VM forming a routing peer relationship with Nexus 9000 switch.

In the example shown in below, NSX edge ECMP VMs (E1-E8) are deployed on ESXi hosts with two physical uplinks connected to the Nexus 9000 ToR switches. Thus, the recommendation is to deploy two logical uplinks on each NSX edge. Since an NSX edge logical uplink is connected to a VLAN-backed port-group, it is necessary to use two external VLAN segments to connect the physical routers and establish routing protocol adjacencies.

Figure 8: VLAN, dvUplink and Routing Peering Mapping



As shown in the figure above, each ECMP node peers over its respective external VLANs to exactly one Nexus router. Each external VLAN is defined only on one ESXi uplink (in the figure above external VLAN10 is enabled on uplink toward R1 while external VLAN20 on the uplink toward R2). This is done so that under normal circumstances both ESXi uplinks can be concurrently utilized to send and receive north-south traffic, even without requiring the creation of a port-channel between the ESXi host and the ToR devices.

In addition, with this model a physical failure of an ESXi NIC would correspond to a logical uplink failure for the NSX edge running inside that host, and the edge would continue sending and receiving traffic leveraging the second logical uplink (the second physical ESXi NIC interface).

In order to build a resilient design capable of tolerating the complete loss of an edge rack, it is also recommended to deploy two sets of four edge gateways in two separate edge racks. The below table describes the necessary configuration with ECMP edge.

Table 6: Edge Cluster VDS Configuration

Port Group	VLAN	dvUplink 1	dvUplink 2	Load Balancing
VTEPs	XXX	Active	Active	SRC_ID
Edge-External-1	YYY	Active	NA	SRC_ID
Edge-External-2	ZZZ	NA	Active	SRC_ID

2.3.3.4 NSX Edge Routing Protocol Timer Recommendations

The NSX edge logical router allows dynamic as well as static routing. The recommendation is to use dynamic routing protocol to peer with Nexus 9000 switches in order to reduce the overhead of defining static routes every time the logical network is defined. The NSX edge logical routers support OSPF, BGP and IS-IS routing protocol. The NSX edge ECMP mode configuration supports reduction of the routing protocol “hello and hold” timer to improve failure recovery of traffic in the case of node or link failure. The minimum recommended timer for both OSPF and BGP is shown in table below.

Table 7: Edge Cluster VDS Configuration

Routing Protocol	Keep Alive or Hello Timer	Hold Down Timer
OPSF	1	3
BGP	1	3

3 Benefits of NSX Architecture with Cisco Nexus and UCS Infrastructure

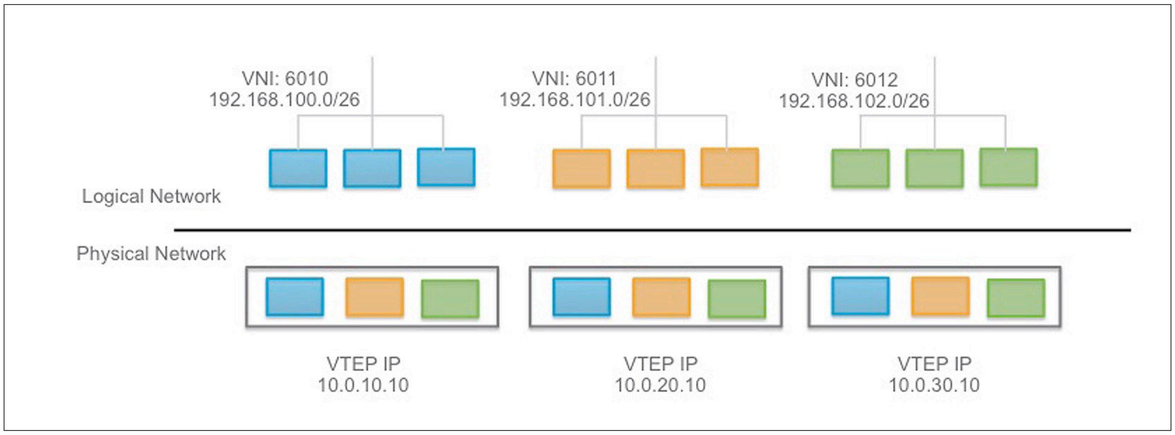
NSX enables users to build logical services for networking and security without having to make configuration changes to the physical infrastructure. In this case, once the Nexus 9000 switches and Cisco UCS systems are configured to provide IP connectivity and the routing configuration is provisioned as described above, we can continue to deploy new services with NSX.

Let us look at some examples that show how applications can be deployed with NSX for network virtualization.

3.1 Logical Layer Connectivity

The figure below shows how logical layer-2 segments can be built. Here we can observe that servers in the physical infrastructure can be in different subnets, yet an overlay network enables VMs to be in the same subnet and layer-2 adjacent, essentially providing topology-independent connectivity and mobility beyond the structured topology constraint imposed by physical networking.

Figure 9: Logical Layer 2

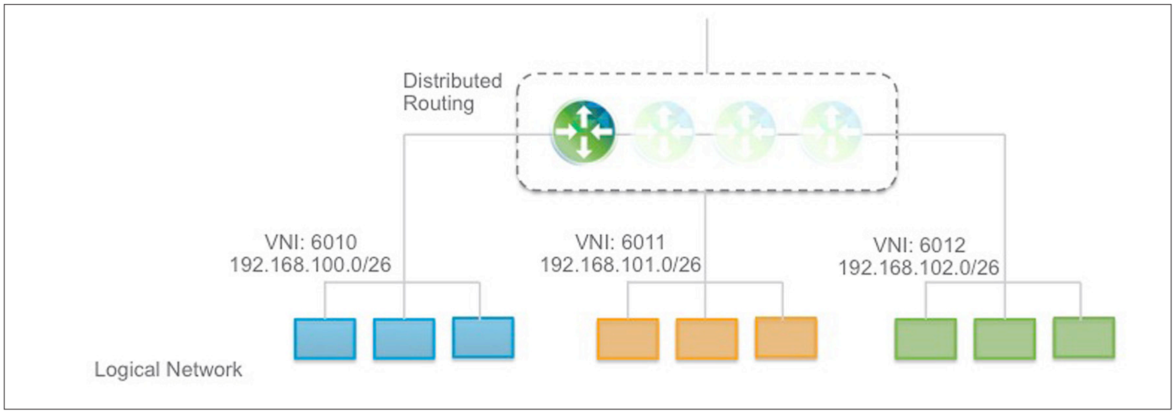


NSX builds multicast-free VXLAN based overlay networks. One can extend layer-2 and IP subnets across servers connected to different ToR Nexus 9000 switches in a layer-3 fabric. This layer-2 adjacency between the VMs can be established independently of the physical network configuration. New logical networks can be created on demand via NSX, decoupling the logical virtual network from the physical network topology.

3.2 Distributed Routing

NSX enables distributed routing and forwarding between logical segments within the ESXi hypervisor kernel. As shown in the figure below, three different logical networks are isolated in three different subnets. One can simply route between the subnets using the distributed routing functionality provided by NSX. OSPF, BGP and static routing are supported with distributed routing.

Figure 10: Distributed Routing with NSX



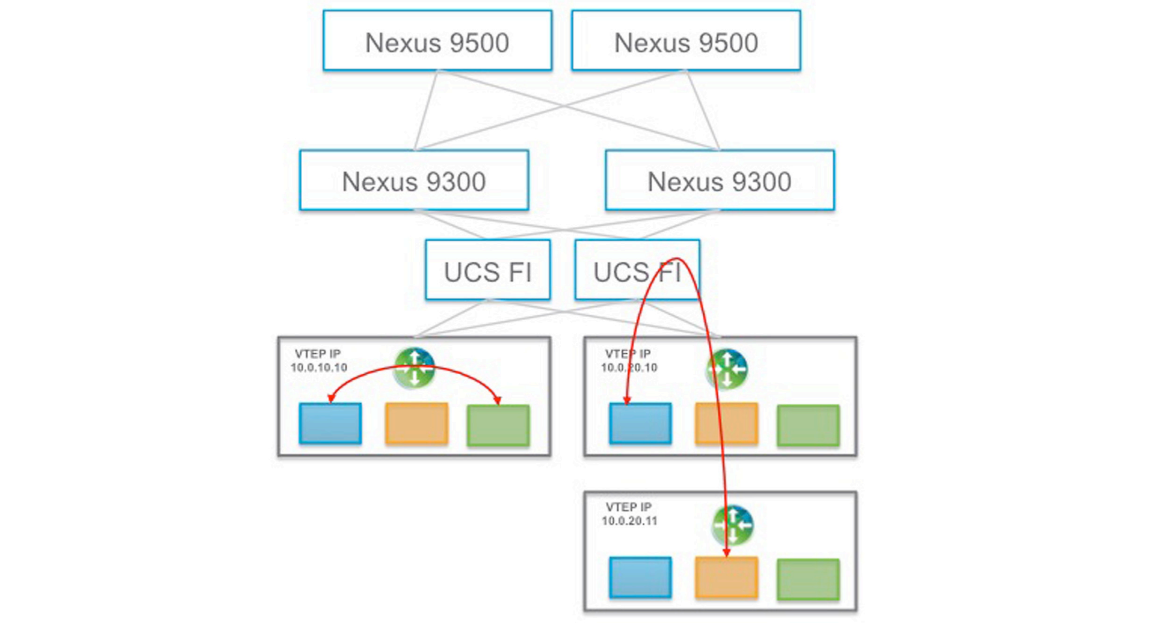
The key benefit of distributed routing is an optimal scale-out routing for east-west traffic between VMs. Each hypervisor has a kernel module that is capable of a routing lookup and forwarding decision. As shown in Figure 10 above, traffic within a single host can be routed optimally within the host itself—even if the VMs are part of a different logical switch. The localized forwarding reduces traffic to the ToR and potential for reduced latency as packets are switched locally in memory.

Traffic across hosts needs to go to the physical switch where NSX can make a forwarding decision based upon the destination VTEP IP. In Figure 11 below, traffic between two VMs on two different VTEP IP addresses is sent up to the UCS fabric interconnects. However, since VTEP IP 10.0.20.10 and VTEP IP 10.0.20.11 are in the same layer-2 domain, the UCS fabric interconnect can forward it without sending it up to the Nexus 9300 switch, reducing the physical number of hops needed—thereby improving latency, performance and oversubscription.

In a classic architecture all traffic would be forwarded to the switch with the SVI configuration; that is not necessary with the NSX distributed routing capability.

The distributed router scale-out capability supports multi-tenancy in which multiple distributed logical router instances can be invoked to provide routing-control plane separation within the shared infrastructure.

Figure 11: Distributed Routing Traffic Flow

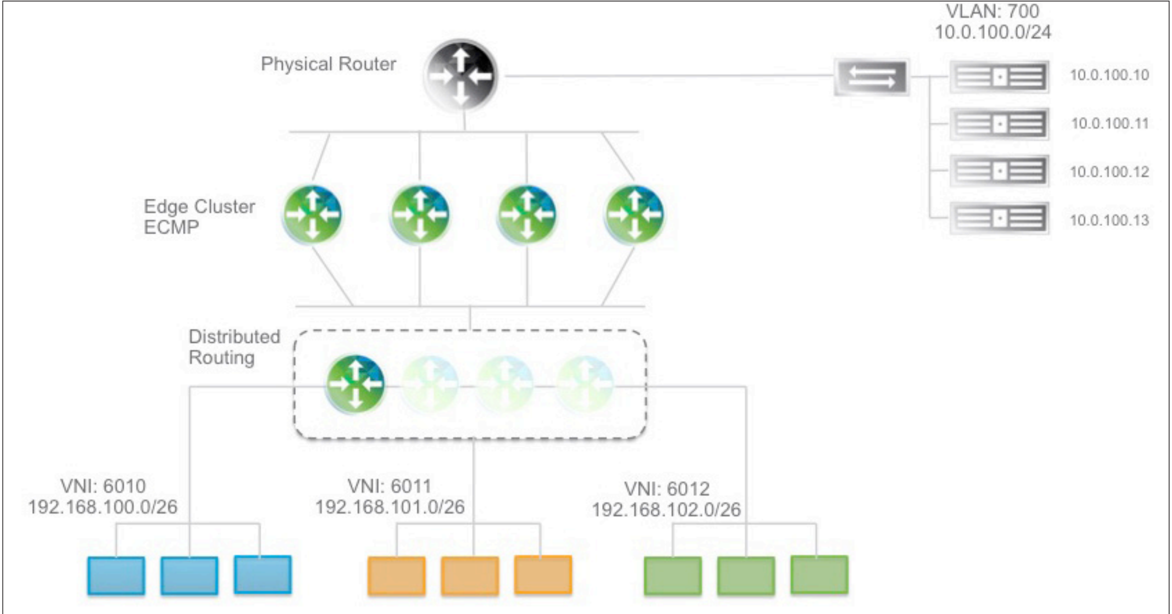


3.3 Routing to Physical Infrastructure

Distributed routing can meet the requirements of routing between virtual workloads. In order to route from the logical network to the physical network, NSX can learn and exchange routes with the physical infrastructure in order to reach resources such as a database server or a non-virtualized application, which could be located on different subnet on a physical network.

NSX provides a scale-out routing architecture with the use of ECMP between the NSX distributed router and the NSX Edge routing instances as shown in the figure below. The NSX Edges can peer using dynamic routing protocols (OSPF or BGP) with the physical routers and provide scalable bandwidth. In the case of a Nexus 9000 switch infrastructure, the routing peer could be a ToR Nexus 9300.

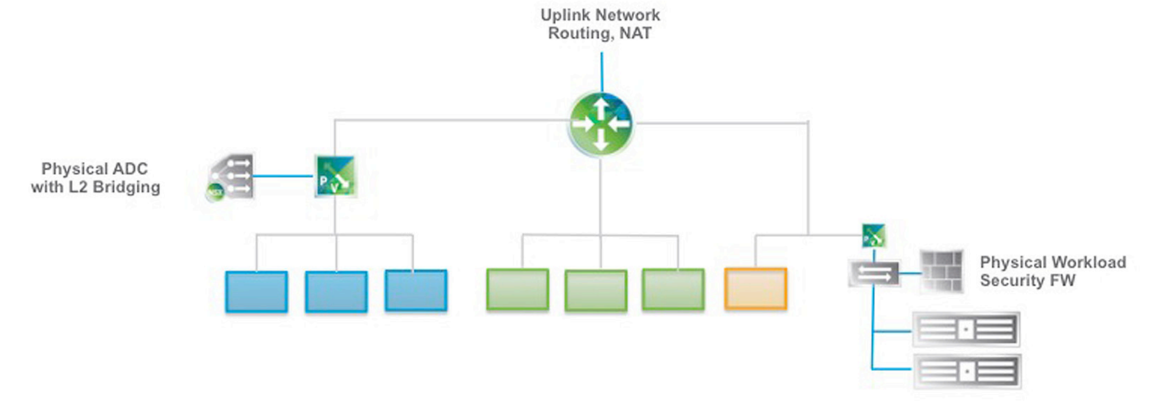
Figure 12: Routing from Logical to Physical Workloads



3.4 Layer-2 Bridging from Virtual to Physical Infrastructure

Some application and service integration may require connecting VMs to physical devices on the same subnet (layer-2 centric workload connectivity). Examples of this are migrations to virtual workloads; app-tiers have hard-coded IP addresses and some workloads reside in virtual and integration with ADC appliances. This can be accomplished by leveraging the native bridging functionality in NSX. The layer-2 bridge instance runs in a distributed manner and can bridge a VXLAN segment to a VLAN instance as shown in Figure 13 below.

Figure 13: Layer-2 Bridging from Virtual to Physical



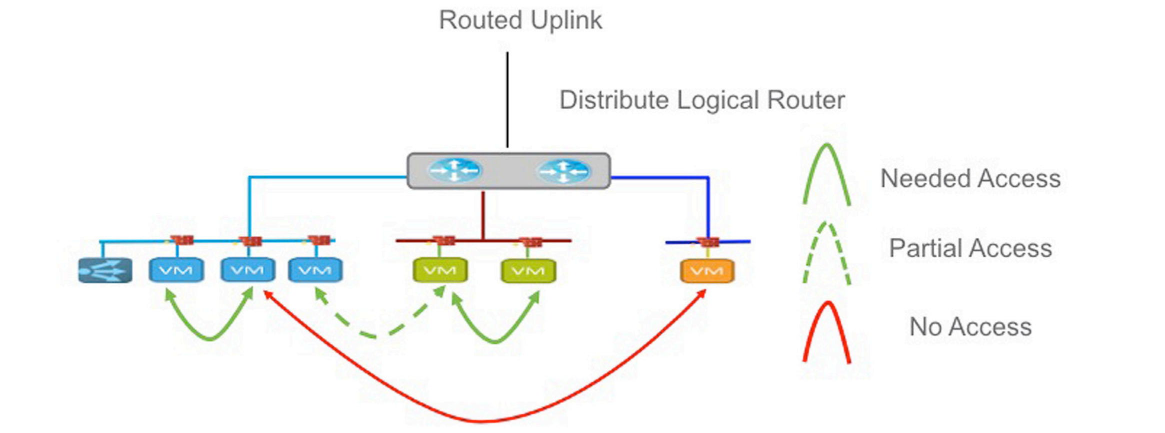
Layer-2 bridging design considerations are covered in the **NSX design guide**. Additionally, one can use multicast-based HW VTEP integration if needed, with additional design considerations.

3.5 Security with Distributed Firewall

NSX by default enables the distributed firewall on each VM at the vNIC level. The firewall is always in the path of the traffic to and from VM. The key benefit is that it can reduce the security exposure at the root for east-west traffic and not at the centralized location. Additional benefits of distributed firewall include:

- Eliminating the number of hops (helps reduce bandwidth consumption to and from the ToR) for applications traversing to a centralized firewall
- Flexible rules sets (rules sets can be applied dynamically, using multiple objects available in vSphere such as logical SW, cluster and DC)
- Allowing the policy and connection states to move with VM vMotion
- Developing an automated workflow with programmatic security policy enforcement at the time of deployment of the VM via cloud management platform, based on exposure criteria such as tiers of security levels per client or application zone

Figure 14: Micro-segmentation and Protection of Traffic



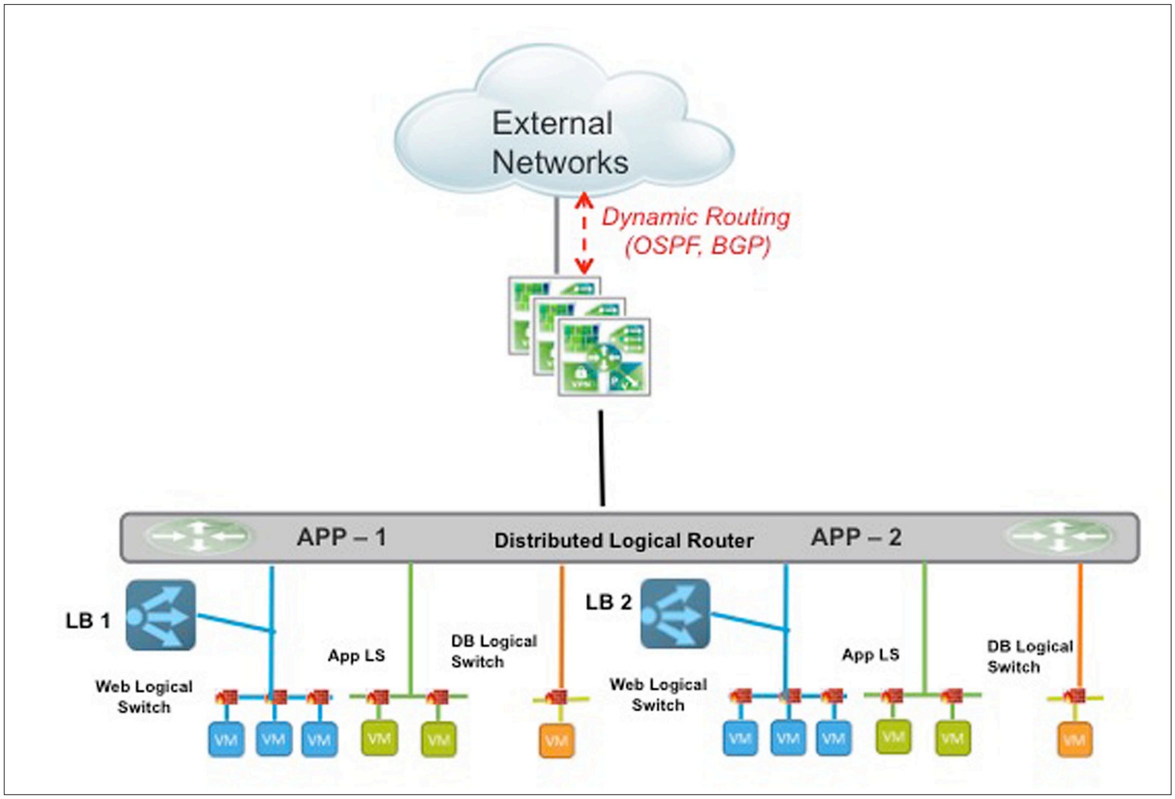
As shown in the figure above, the designer now has flexibility in building a sophisticated policy since policy is not tied to physical topology. The policy can be customized for inter- and intra-layer-2 segment(s), complete or partial access, as well as managing N-S rules sets that can be employed directly at the VM level with edge firewall being an option for the interdomain security boundary.

Micro-segmentation as shown in the figure above allows creating a PCI zone within a shared segment, allowing sophisticated security policies for desktops in a VDI environment as well as eliminating the scaling limitation of centralized access-control ACL management.

3.6 Flexible Application Scaling with Virtualized Load Balancer

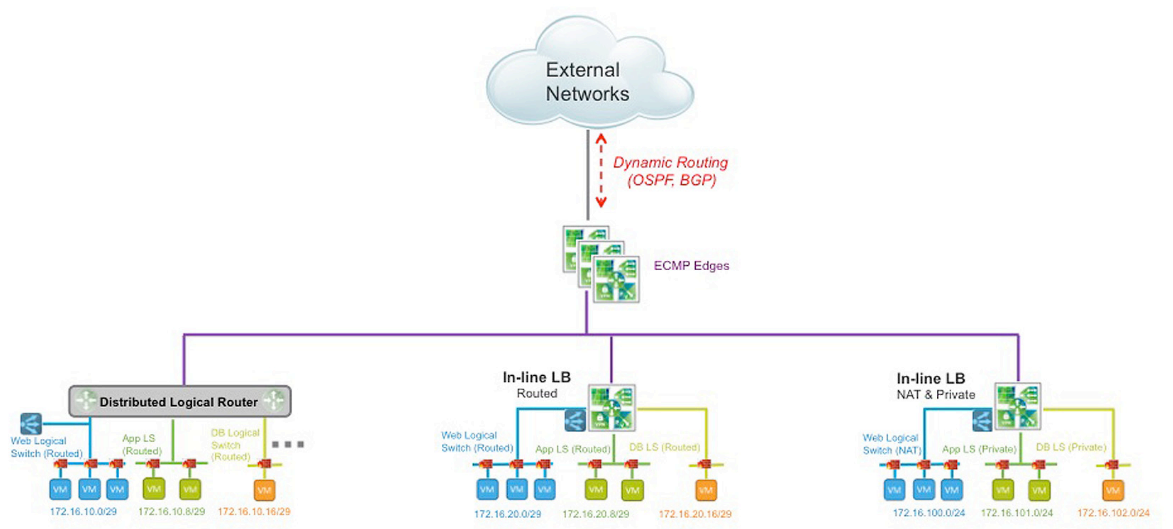
Elastic application workload scaling is one of the critical requirements in today's data center. Application scaling with a physical load balancer may not be sufficient given the dynamic nature of self-service IT and DevOps style workloads. The load-balancing functionality natively supported in the edge appliance covers most of the practical requirements found in deployments. It can be deployed programmatically based on application requirements with appropriate scaling and features. The scale and application support level determines whether the load balancer can be configured with layer-4 or layer-7 services. The topology wise the load balancer can be deployed either in-line or in single-ARM mode. The mode is selected based on specific application requirements, however the single-ARM design offers extensive flexibility since it can be deployed near the application segment and can be automated with the application deployment.

Figure 15: Logical Load Balancing per Application



The figure above shows the power of a software-based load-balancer in which multiple instances of the load-balancer serve multiple applications or segments. Each instance of the load-balancer is an edge appliance that can be dynamically defined via an API as needed and deployed in a high-availability mode. Alternatively, the load balancer can be deployed in an in-line mode, which can serve the entire logical domain. The in-line load-balancer can scale via enabling multi-tier edge per application such that each application is a dedicated domain for which first-tier edge is a gateway for an application, the second-tier edge can be an ECMP gateway to provide the scalable north-south bandwith.

Figure 16: Scaling Application and Services with NSX



As one can observe from the figure above, the first application block on the left is allowing a single-ARM load-balancer with distributed logical routing. The center and the right block of the application allow an in-line load-balancer with either routed or NAT capability respectively. The second-tier edge is enabled with ECMP mode to allow the application to scale on demand from 10GB to 80GB and more.

Conclusion

NSX, deployed on top of Nexus 9000 switches and Cisco UCS blade server infrastructure, enables best-of-breed design with flexibility and ease of deployment for a full stack of virtualized network services. The programmatic capability of software-based services opens the door for self-service IT, dynamic orchestration of workloads, and strong security policies with connectivity to the hybrid cloud.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com