VXLAN Design Using Dell EMC S and Z series Switches

Standard based Data Center Interconnect using Static VXLAN.

Dell Networking – Data Center Technical Marketing
March 2017

Date	Revision	Description	Author(s)
March 2017	1.0	Initial Release	Mario Chow
July 2017	1.1	Second Release	Mario Chow

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

Table of contents

Introduction....5
Spanning Tree and VLAN Range Limitations....5
Multi-tenancy....5
Objective....6
VTEP (VXLAN Tunnel EndPoint) Overview....6
Software Based VTEP....7
Hardware-Based VTEP....7
VTEP Packet Forwarding Flow....10
Dell EMC Static VXLAN Architecture on the S and Z series Data Center Switches....11
BUM Data Packet Handling....12
Unicast Data Packet Handling....12
MAC learning....12
Caveats....13
Configuring static VXLAN on Dell EMC S and Z Series Switches....13
Dell EMC HW-Based VXLAN Deployment Options....15
Layer 2 Inter-Rack Domain Extension....16
Layer 3 Inter-Rack Data Center....17
Layer 3 VLXAN Option....20
Looking ahead....22
Conclusion....22
Appendix A: Sample Dell EMC Static VXLAN Configuration....23

Figure 1 VTEP Function Diagram....7
Figure 2 Layer 2 VTEP Gateway – VXLAN....8
Figure 3 Layer 2 VTEP Bridge – VXLAN Bridging....9
Figure 4 Layer 3 VTEP – VXLAN Routing....9
Figure 5 VXLAN Unicast Packet Flow....10
Figure 6 Dell EMC Static VXLAN Implementation....11
Figure 7 Local MAC address learning....13
Figure 8 Remote MAC address learning....13
Figure 9 Dell EMC S6000-ON Static VXLAN Example....14
Figure 10 Legacy Layer 2 Data Center Design....16
Figure 11 Enhanced Layer 2 Domain Extension Data Center Design....17
Figure 12 Layer 3 (Spine-to-Leaf) with Single Broadcast Domain per Rack....18
Figure 13 Layer 3 (Spine-to-Leaf) and Layer 2 Domain stretch – Option 1....19
Figure 14 Layer 3 (Spine-to-Leaf) and Layer 2 Domain stretch – Option 2....20
Figure 15 VXLAN Routing – VXLAN 3000-to-VXLAN 4000, or VXLAN 3000-to-VLAN 40....21

Figure 16 Static VXLAN Setup..23

Table 1 Software vs. Hardware VTEP comparison...10
Table 2 Dell EMC Static VXLAN guidelines...13

Introduction

In the information technology field often when a new technology is created to address a need, an enhancement or complement technology is created.

The best example of this is *virtualization* (network and resources) in the data center. Prior to virtualization, traditional network segmentation was provided through VLANs (Virtual Local Area Networks) where a set of hosts or a single host would be assigned to a different VLAN ID in order to segment them from each other if these hosts wanted to communicate with each other; inter-vlan routing would be required.

There are inherent shortcomings when using VLANs, such as inefficient network inter-links, spanning tree limitations, physical location of devices, limited number of VLANs (4,094), multi-tenant environments, and ToR (Top of Rack) switch scalability. VLANs have become a limiting factor for IT departments and providers as they look to build efficient and highly scalable multitenant data centers where *server virtualization* is a key critical component.

Spanning Tree and VLAN Range Limitations

A typical data center uses Layer 2 design guidelines to facilitate inter-VM communication. This always results in the presence of spanning tree protocol to avoid any sort of network loops due to potential duplicate network paths. The use of spanning tree though renders half of the data center network fabric useless since it blocks half of the links to avoid network traffic replication that could cause looping of frames. This increases TCO (Total Cost Ownership), effectively paying for ports that will not be used.

In addition to spanning tree, the typical Layer 2 network will always use VLANs to provide basic broadcast isolation. The traditional VLAN implementation uses a 12-bit field ID used to separate a large layer 2 domain into separate broadcast domains. This range is between 1 – 4,096 IDs. For a small data center environment, this range is usually not an issue; however, with the increased virtualization adoption rate in today's data center and multi-tenant requirements the traditional VLAN range simply cannot scale.

Multi-tenancy

It wasn't long ago when the word cloud computing was just a "buzz" word or a "concept" found in niche areas around service providers willing to try out new bleeding edge technologies. Today, cloud computing is as normal of a concept to most enterprise organizations as the internet is to the average individual. It provides a list of benefits that when properly leveraged it becomes a valued asset.

One of the primary drivers of cloud computing is the concept of "***elasticiticy***". This on-demand elastic provisioning of resources for multi-tenant applications makes cloud computing an opportunity for organizations to offload their IT infrastructure to a cloud service provider. Unfortunately, the use of traditional VLANs creates a scalability problem due to the typically large number of tenants supported within the same data center. Since VLANs are used to isolate traffic between the different tenants the usual 12 bit range is often inadequate.

Although Layer 3 networks can sometimes be leveraged to address the VLAN range limitation, it is not atypical to find two different tenants requiring the same Layer 3 IP addressing scheme due to legacy applications, network design requirements, or inherited requirements.

The IEEE has proposed potential initiatives such as TRILL (Transparent Interconnection of Lots of Links), or SPB (Shortest Path Bridge) to address traditional spanning tree shortcomings however these initiatives have received very little interest or have not been adopted at all.

What then can be done to resolve these issues? Enter VXLAN (Virtual Extensible LAN), a standard officially documented by the IETF as an open standards solution to resolve these shortcomings.

Dell EMC with its S and Z Series data center networking product portfolio switches have been designed for the next-generation data center with hardware-based VXLAN function providing Layer 2 connectivity extension across a Layer 3 boundary while keeping seamless integration between VXLAN and non-VXLAN environments. Together, these switches form the physical network underlay building blocks of a scalable virtualized and multitenant data center.

Objective

The objective of this design guide is to provide a deeper understanding of Dell EMC's Static VXLAN design and deployment options. The reader should first read the introductory VXLAN document found here at Dell Tech Center.

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

VTEP (VXLAN Tunnel EndPoint) Overview

VXLAN as noted before is an encapsulation and tunneling scheme and as such an entity to encapsulate/decapsulate and originate/terminate the tunnel is needed. The encapsulation requirement has been discussed in the ***VXLAN Overview*** section of the document available at Dell Tech Center (see link under "Objective"). Now we need to discuss how this encapsulated traffic is handled as it traverses the IP infrastructure?

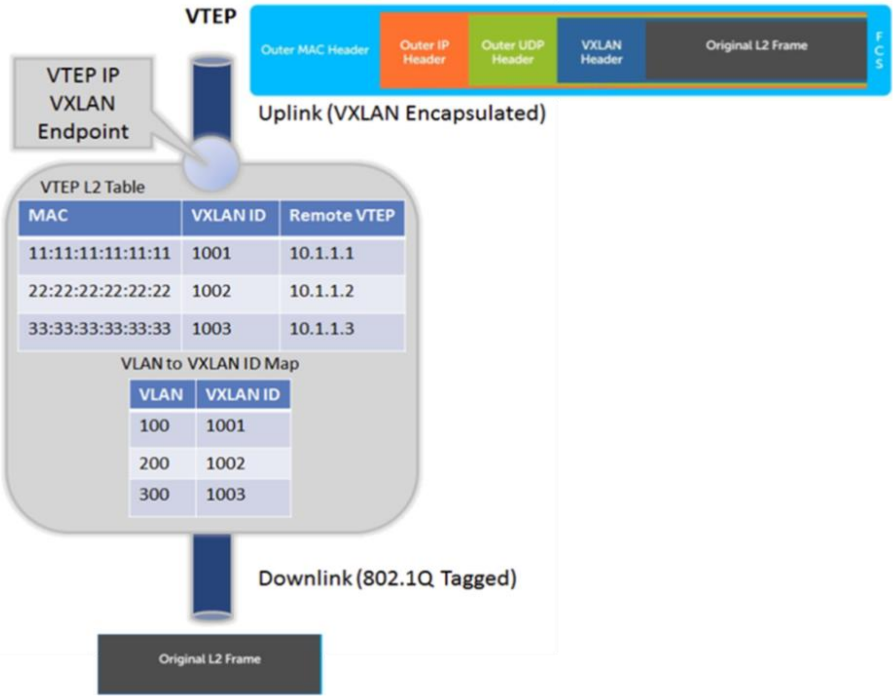
We know, every data center consists of a mixture of virtualized and non-virtualized compute resources. We also know that these virtualized environments MUST be able to communicate with non-virtualized environments. It is not possible to have a homogeneous data center from the applications, resources, or infrastructure point of view.

In a typical data center, a virtualized server has several VMs and they communicate with each other through what is called a vSwitch (virtual Switch). This vSwitch is the first hop for all the VMs and implements the network virtualization part of a virtual environment. Unfortunately, this vSwitch construct does not exist on a non-virtualized server or bare-metal server and therefore establishing any sort of communication between a virtualized and non-virtualized compute resource is not possible unless some type of appliance or gateway device can serve as a bridge between these two different environments.

To bridge this gap, the VXLAN RFC introduced the concept of a VXLAN Tunnel Endpoint or VTEP. This entity has two logical interfaces: an uplink and downlink. The uplink interface receives VXLAN frames and acts as the tunnel endpoint with an IP address used for routing VXLAN encapsulated frames. The IP address assigned to this uplink interface is part of the network infrastructure and completely separate from the VMs or tenants IP addressing using the VXLAN fabric.

Packets received on the uplink interface are mapped from the VXLAN ID to a VLAN and the Ethernet payload is then sent as a typical 802.1Q Ethernet frame on the downlink to the final destination. As this packet passes through the VTEP a local table is created with the inner MAC SA and VXLAN ID. Packets that arrive at the downlink interface are used to create a VXLAN ID to regular VLAN map.

Figure 1 VTEP Function Diagram



VTEP L2 Table

MAC	VXLAN ID	Remote VTEP
11:11:11:11:11:11	1001	10.1.1.1
22:22:22:22:22:22	1002	10.1.1.2
33:33:33:33:33:33	1003	10.1.1.3

VLAN to VXLAN ID Map

VLAN	VXLAN ID
100	1001
200	1002
300	1003

There are two types of VTEPs that have been introduced in the industry, a software and hardware based VTEP Gateway:

Software Based VTEP

The software based VTEP runs as a separate appliance and it typically runs on a standard x86 hardware with an instance of Open vSwitch. This VTEP is under a controller and it maps physical ports and VLANs on those ports to logical networks. Through this map, VMs that are part of the same logical network can now communicate with the physical device that belongs to the same logical network. This is the approach taken by several overlay architecture such as VMware NSX, Midokura, and others.

Software based VTEPs are a great solution for fairly moderate amounts of traffic between VMs and physical devices.

Hardware-Based VTEP

When a full rack of physical servers running database applications need to connect to logical networks containing VMs, ideally these bare metal servers will need a high-density and high-performance switch that could bridge/switch these traffic patterns between the physical servers and logical segments. This is where hardware based VTEPs make sense. The hardware based VTEP is no different than the software based VTEP in terms of functionality and controller interaction. A controller which has visibility into the virtualized environment is used to integrate the hardware VTEP and thus creates the gateway functionality required between the virtual and non-virtual environments.

With hardware-based VTEP, the VTEP functionality is implemented in the hardware ASIC resulting in a performance and scalability edge.

There are two types of VTEP functionalities: Layer 2 and Layer 3. As a Layer 2 it can be a gateway or a bridge. It provides both encapsulation and de-capsulation of legacy or traditional Ethernet and VXLAN packets respectively and allows the stretch of a Layer 2 domain across a Layer 3 domain. Tagged packets configured with a VLAN ID are mapped to a respective VNI entry encapsulated with a VXLAN header.

As a Layer 2 VTEP, VMs or hosts from VXLAN segment X cannot communicate with a different VM or host in VXLAN segment Y in order for this communication to take place, a router or Layer 3 device is needed where all routing functions will be performed based on the outer most IP header (see VXLAN Packet Format).

Figures 2 and 3 show the VTEP as a gateway and bridge.

Figure 2 Layer 2 VTEP Gateway – VXLAN

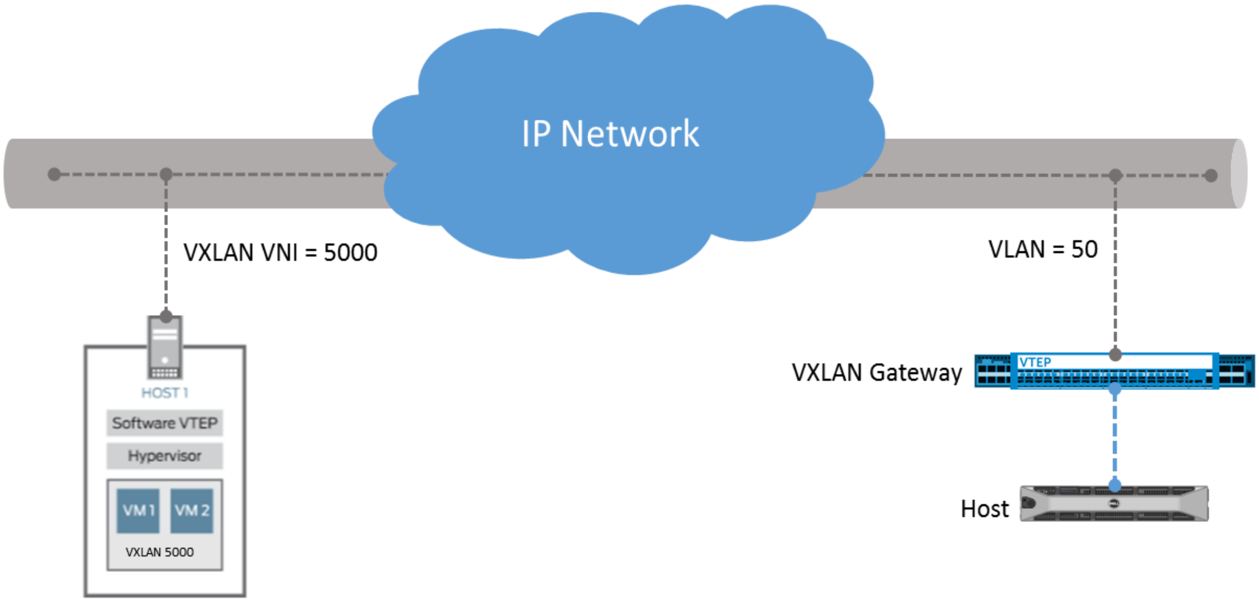
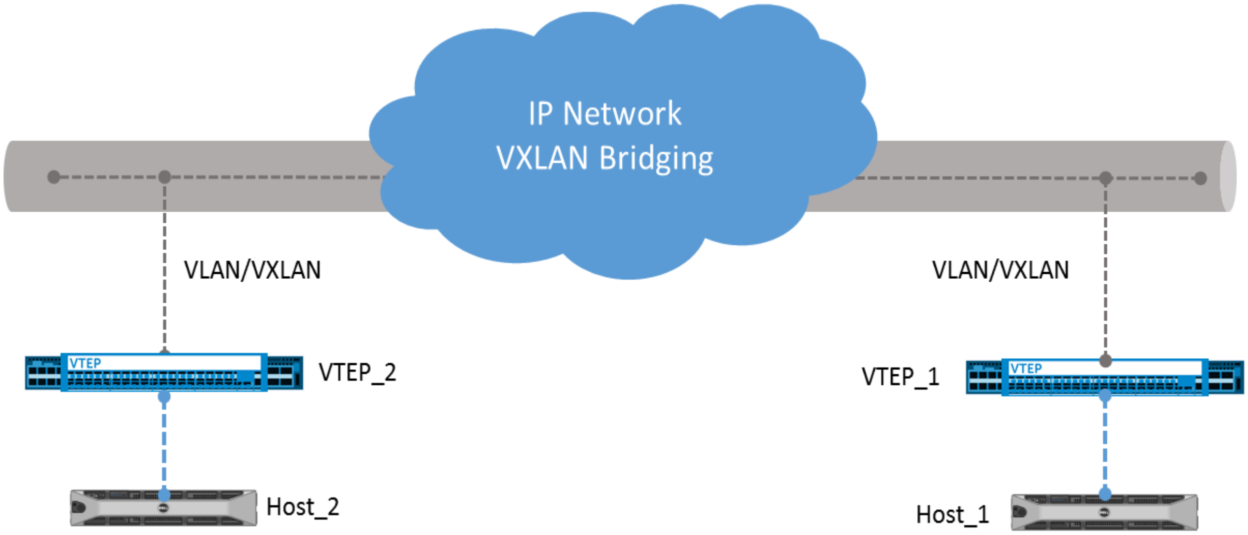
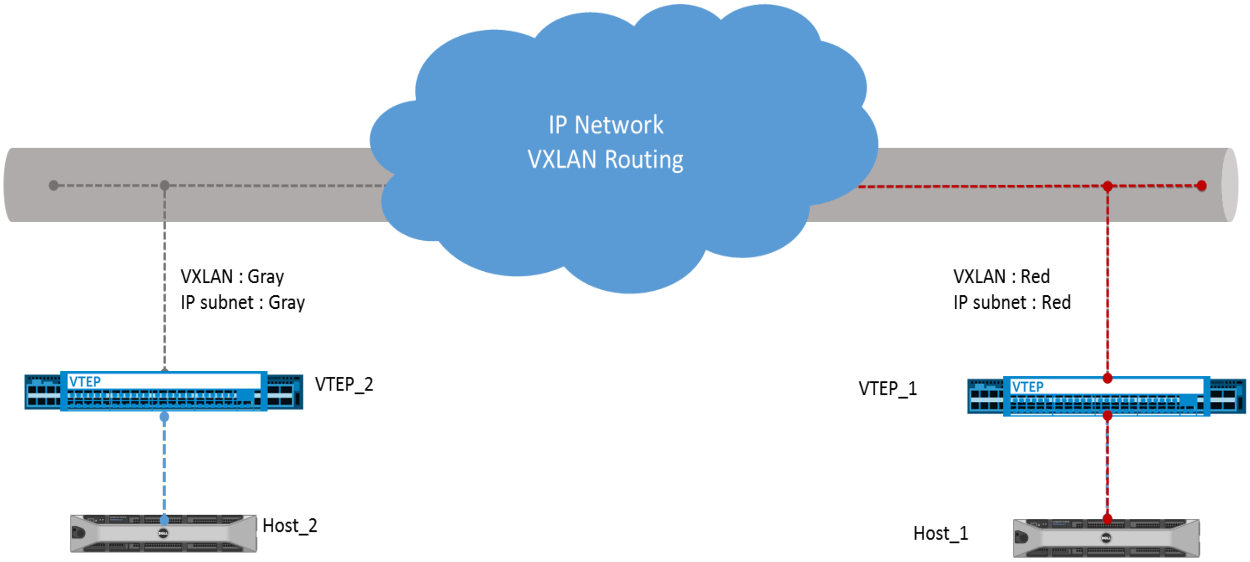


Figure 3 Layer 2 VTEP Bridge – VXLAN Bridging



As a Layer 3 VTEP, it performs straight VXLAN segment X to VXLAN segment Y routing similar to the traditional inter-VLAN routing providing communication between different VXLAN segments. Contrary to how a Layer 2 VTEP gateway functions (VXLAN to VLAN mapping), a Layer 3 VTEP gateway, performs routing using the VXLAN ID, there is no mapping dependency.

Figure 4 Layer 3 VTEP – VXLAN Routing



VXLAN routing is currently not widely supported with the current silicon. Currently, the most common type of a VTEP GW deployment is as a Layer 2 VTEP gateway with VXLAN routing support to take place with later software releases and newer silicon ASIC.

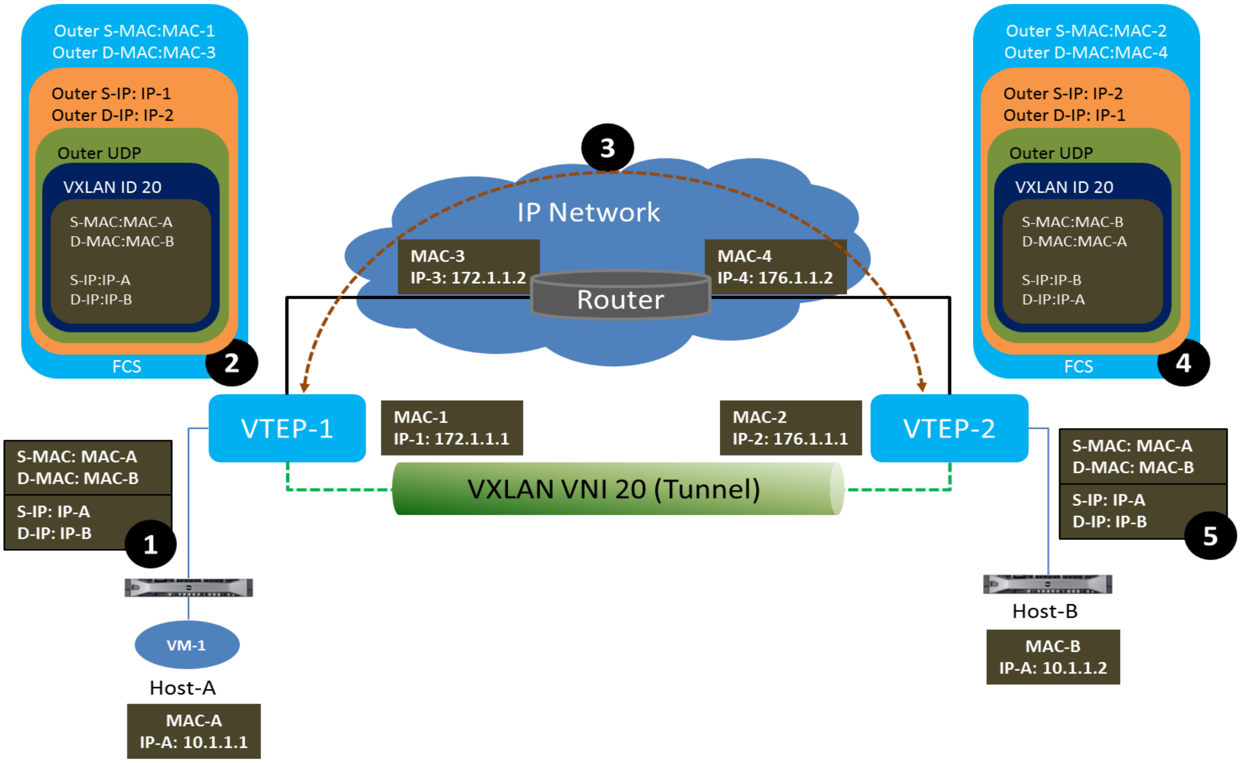
Table 1 Software vs. Hardware VTEP comparison

Key Points	Software VTEP	Hardware VTEP
Virtual, physical, or both	Virtual	Physical
Scalability	Moderate traffic between VMs	Moderate to heavy traffic between VMs and physical servers
Orchestration	Controller based	Controller and non-controller based
Performance	CPU driven	ASIC driven (Line rate)

VTEP Packet Forwarding Flow

Figure 5, shows an example of a VXLAN packet forwarding flow. Notice how the outer headers are used at the respective phases as the packet flows from VTEP 1 to VTEP 2.

Figure 5 VXLAN Unicast Packet Flow



Host-A and Host-B belong to the same VXLAN segment ID 20 and communicate with each other through the VXLAN tunnel created between VTEP-1 and VTEP-2.

Step 1 – Host-A sends traffic to Host-B, it creating an Ethernet frame with Host-B's destination MAC and IP address and sends it towards VTEP-1.

Step 2 – VTEP-1 upon receiving this Ethernet frame has a map of Host-B's MAC address to VTEP-2. VTEP-1 performs the VXLAN encapsulations by adding the respective headers such as VXLAN, UDP, and outer IP headers.

Step 3 – Using the information in the outer IP headers, VTEP-1 performs an IP address lookup for VTEP-2's address to resolve the next-hop in the transit or IP network. Using the outer MAC header information, VTEP-1 sees the router's MAC address as the next-hop device.

Step 4 – The Ethernet frame is routed towards VTEP-2 based on the outer IP header which has VTEP-2's IP address as the destination address.

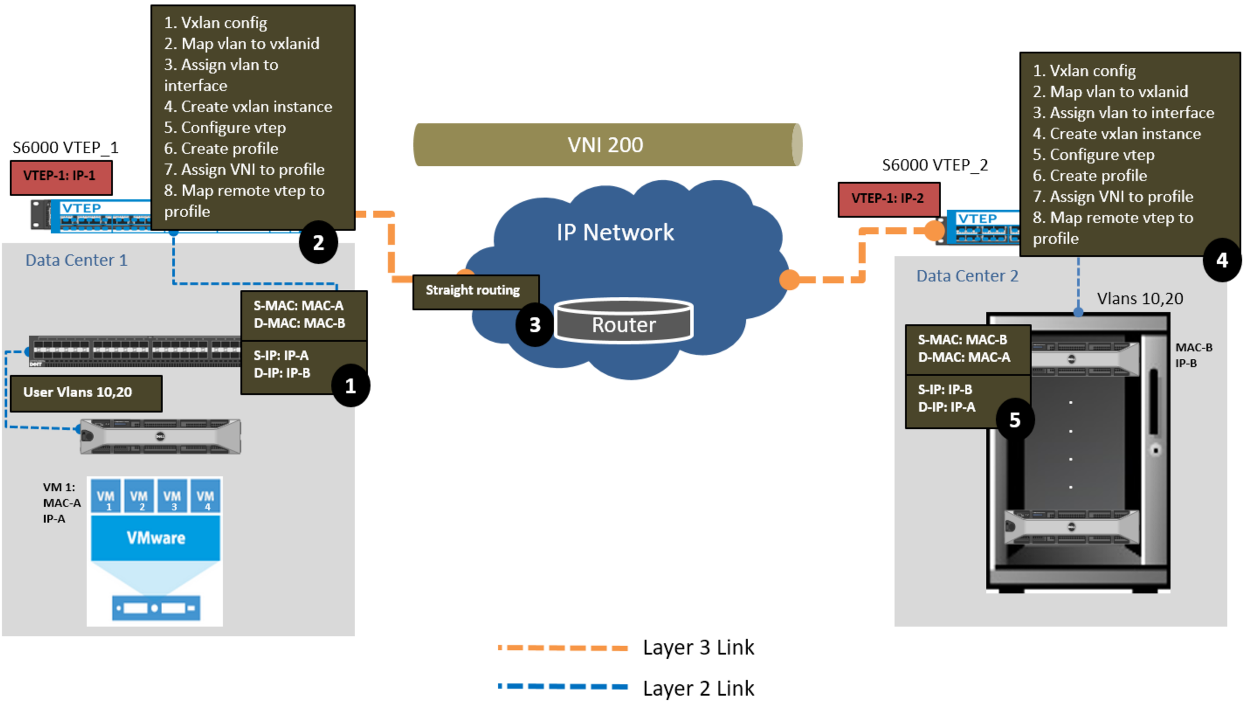
Step 5 – After VTEP-2 receives the Ethernet frame, it de-capsulates all the headers and forwards the packet to Host-B using the original destination MAC address.

Dell EMC Static VXLAN Architecture on the S and Z series Data Center Switches

Dell EMC's static vxlan implementation is simple and efficient. Unlike what others have done where multicast is heavily leveraged to create static VTEP tunnels between the different VTEP switches, Dell EMC's implementation does not use multicast.

The architecture creates profiles containing VNIs defined by the user that map to standard 802.1Q vlans, it then assigns these VNIs to the profile. The profiles are then statically mapped to the specific remote VTEP. Figure 6, shows the implementation steps and points where the specific implementation takes place.

Figure 6 Dell EMC Static VXLAN Implementation



There are several key areas of the implementation that need some highlighting. These are:

- BUM Data Packet handling
- Unicast Data Packet handling
- And Source Address learning methodology

BUM Data Packet Handling

When a packet arrives on an access port with a destination MAC address being either a broadcast, unknown unicast, or multicast (BUM) for which there is no knowledge on the VTEP switch which remote VTEP tunnel to be used for the destination MAC address, the following steps are taken by the local VTEP switch:

Step 1 - Capture the logical network ID configured on this port or vlan interface on which the packet has been received.

Step 2 - Learn the SA MAC address on this port or vlan interface.

Step 3 - Send a unicast copy of the packet to each remote VTEP over the vxlan tunnel which is configured to be part of this logical network. Also send a copy to any local access port that are also part of this logical network.

Unicast Data Packet Handling

When a data packet arrives on an access port with a unicast destination MAC address, the following steps are taken by the VTEP.

Step 1 - The logical network ID is derived from the VLAN or port on which the packet has been received.

Step 2 - A lookup is performed in the Layer 2 entry table with the VFI and a unicast destination MAC address.

- If the lookup is successful and if the packet is destined to a local access port(s), the native packet will be sent on the access port(s), however, if the packet is destined to a remote VTEP tunnel, the packet is encapsulated with a VXLAN header with a VNID based VFI and the destination IP address based on the remote VTEP tunnel and is sent out to the next hop along the path to the remote VTEP tunnel

- If the lookup is unsuccessful, the same steps that apply to a BUM traffic is performed.

MAC learning

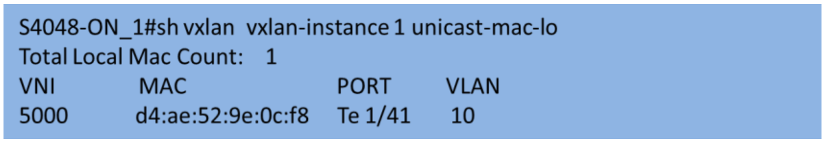
Local and remote SA MAC learning happen at different places during data forwarding. Local learning takes place when a data packet is received on the access ports. Whereas remote learning takes place at the VTEP tunnel.

Learning is enabled by default and it cannot be disabled. The same applies with remote MAC learning which is enabled on the VTEP tunnels.

For example, when a data packet is received on an access port, this SA MAC address is mapped to a specific VNID and ingress VTEP tunnel IP address. This piece of information (VNID, SA MAC, and VTEP tunnel IP) is programmed in the switch's Layer 2 entry or forwarding table.

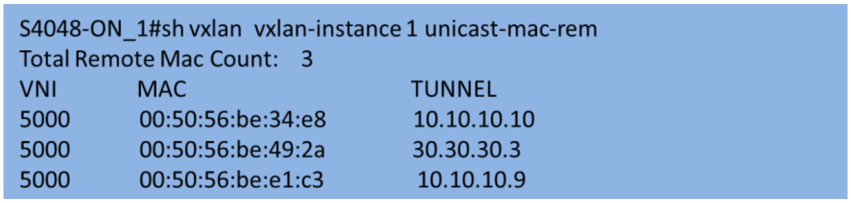
Figures 7 & 8 show an output of both local and remote address learning per vxlan-instance. Note the tunnel ID or IP address of the tunnel. This is the vtep tunnel on which the remote MAC address is learned. It is also the static tunnel used to reach the MAC address.

Figure 7 Local MAC address learning



```
S4048-ON_1#sh vxlan  vxlan-instance 1 unicast-mac-lo
Total Local Mac Count:    1
VNI          MAC                  PORT       VLAN
5000         d4:ae:52:9e:0c:f8    Te 1/41     10
```

Figure 8 Remote MAC address learning



```
S4048-ON_1#sh vxlan  vxlan-instance 1 unicast-mac-rem
Total Remote Mac Count:    3
VNI          MAC                        TUNNEL
5000         00:50:56:be:34:e8          10.10.10.10
5000         00:50:56:be:49:2a          30.30.30.3
5000         00:50:56:be:e1:c3           10.10.10.9
```

The current hardware supports up to 8,000 unique VNIDs and 511 tunnels, with 4,000 VNIDs and 256 tunnels tested in the lab. Each tunnel can carry multiple VNIDs

Note: Although VXLAN allows up to 16M unique VNIDs or VXLAN segments, the current silicon supports up to 8,000 unique VNIDs with 16M logical IDs available. The current number of VNIDs supported covers the majority of today's VXLAN to non-VXLAN environments deployments.

Caveats

There are some guidelines to follow when deploying static vxlan. Table 2 describes these guidelines.

Table 2 Dell EMC Static VXLAN guidelines

Guidelines
• No stacking or VLT supported • One vxlan instance supported • No Layer 3 over VXLAN • No SNMP or REST API supported • 4,000 VNIs • No Load balancing

Configuring static VXLAN on Dell EMC S and Z Series Switches

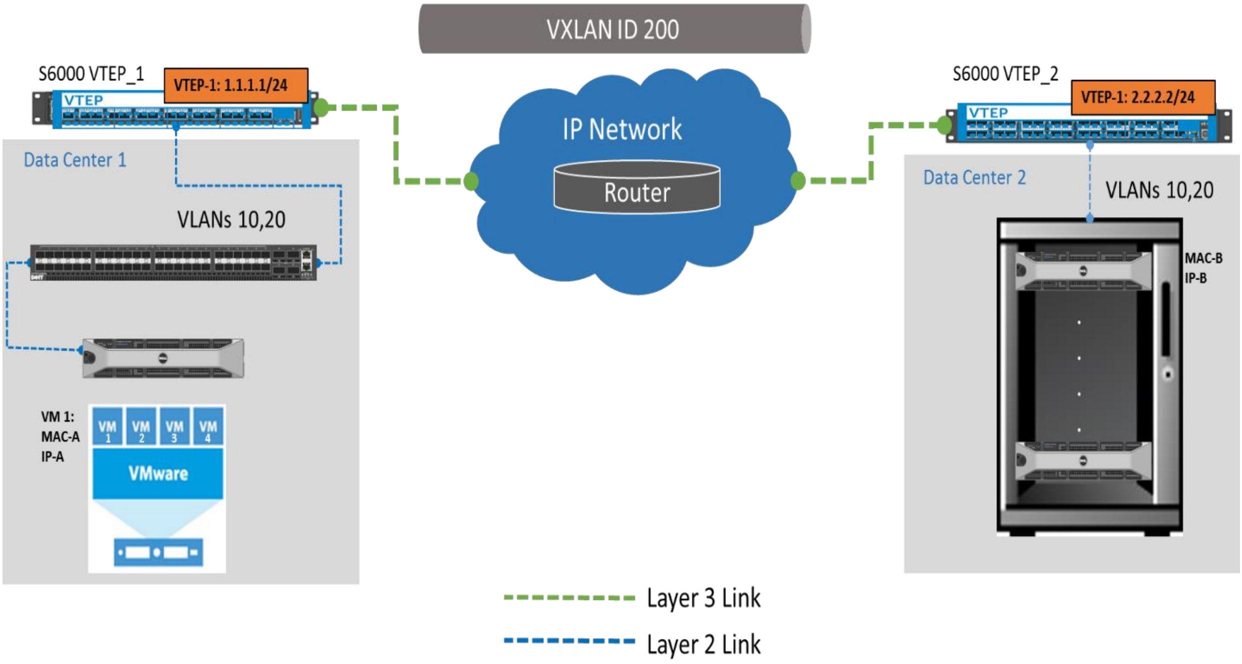
Dell EMC's static VXLAN configuration is divided into 6 short steps:

1. Enable VXLAN feature

2. Map standard VLAN to VXLAN ID (VNI)
3. Configure vxlan instance on participating VXLAN switchport
4. Create a static VXLAN instance
5. Create a VNI profile and associate VNID to VNI profile
6. Associate a remote VTEP to the VNID

Figure 9 is used to demonstrate the steps required to configure Dell EMC's static VXLAN feature. The commands described below apply to **S6000 VTEP_1** (left side of the diagram).

Figure 9 Dell EMC S6000-ON Static VXLAN Example



Step 1 – Enable VXLAN feature

```
Switch# config
Switch(conf)# feature vxlan
```

Step 2 – Map standard VLAN to VXLAN ID (VNI)

```
Switch# config
Switch(conf)# interface vlan 20
Switch(conf-if-vlan-20)# vxlan-vnid 200
Switch(conf-if-vlan-20)# tagged te0/0
```

Step 3 – Apply vxlan-instance on participating switchport

```
Switch# config
Switch(conf)# interface te0/0
Switch(conf-if-te-0/0)# vxlan-instance 1
Switch(conf-if-te-0/0)# switchport
```

Interface te0/0 is connected to an end host interested in communicating with another end host that belongs also to VLAN 20 but is being stretched across an IP infrastructure.

Step 4 – Create the actual VTEP tunnel starting with the static vxlan instance
Here the actual vxlan configuration starts. The keyword "static" **must** be configured in order to enforce the static vxlan feature. If the keyword "static" is not defined, the subsequent configuration commands will result in an "**ERROR: CLI not compatible with vxlan instance mode**"

```
Switch# config
Switch(conf)# vxlan-instance 1 static
Switch(conf-vxlan-inst-1)# local-vtep-ip 1.1.1.1
```

Step 5 – Create the VNI profiles under the same vxlan-instance 1
Continuing with step 4, the vni profile(s) are created and associated to the respective vni created in step 2

```
Switch(conf-vxlan-inst-1)# vni-profile profile1
Switch(conf-vxlan-inst-1-vniProfile-profile1)# vni 200
```

Step 6 – Associate the remote VTEP with the locally created vni-profile
The remote VTEP is the other end of the tunnel that is to carry all VNI traffic that have been defined by the profile. This tunnel can carry multiple VNIs which in turn are mapped to standard VLANs.

```
Switch(conf-vxlan-inst-1-vniProfile-profile1)# remote-vtep-ip 2.2.2.2 vniprofile profile1
Switch(conf-vxlan-inst-1-vniProfile-profile1)# end
Switch#
```

In addition to steps 1 – 6, the green link (Layer 3) towards the router needs to be configured. This link is a standard Layer 3 configuration with an IP address, a routing protocol and whatever else is needed in order to establish ip connectivity with the router. For detailed information on configuring Layer 3 please follow this link http://downloads.dell.com/manuals/all-products/esuprt_ser_stor_net/esuprt_networking/esuprt_net_fxd_prt_swtchs/force10-s6000-on_user%27s%20guide5_en-us.pdf

Dell EMC HW-Based VXLAN Deployment Options

There is no denying that virtualization (compute and network) play a key role in today's data centers. It has transformed the data center from a sunken expense into a revenue generating asset. It has provided enterprises business agility, competitive advantage, and most importantly efficient usage of resources.

One of VXLAN's main benefit is the extension of a Layer 2 domain across an IP infrastructure, however, there will always be a need for VXLAN traffic to communicate with standard VLAN environment as well as a direct VXLAN routing.

There are two supported HW-based VTEP GW solutions.

- Dynamic using an NVP (Network Virtualization Platform) controller such as VMware's NSX to establish the VTEP tunnels between the different VTEP points within or across data centers.

- Static using manually configured VTEP tunnels between local and remote VTEPs within or across data centers.

In this document, the static VXLAN deployment options are discussed. For more information on the dynamic deployment option (VMware NSX) please see the following documents.

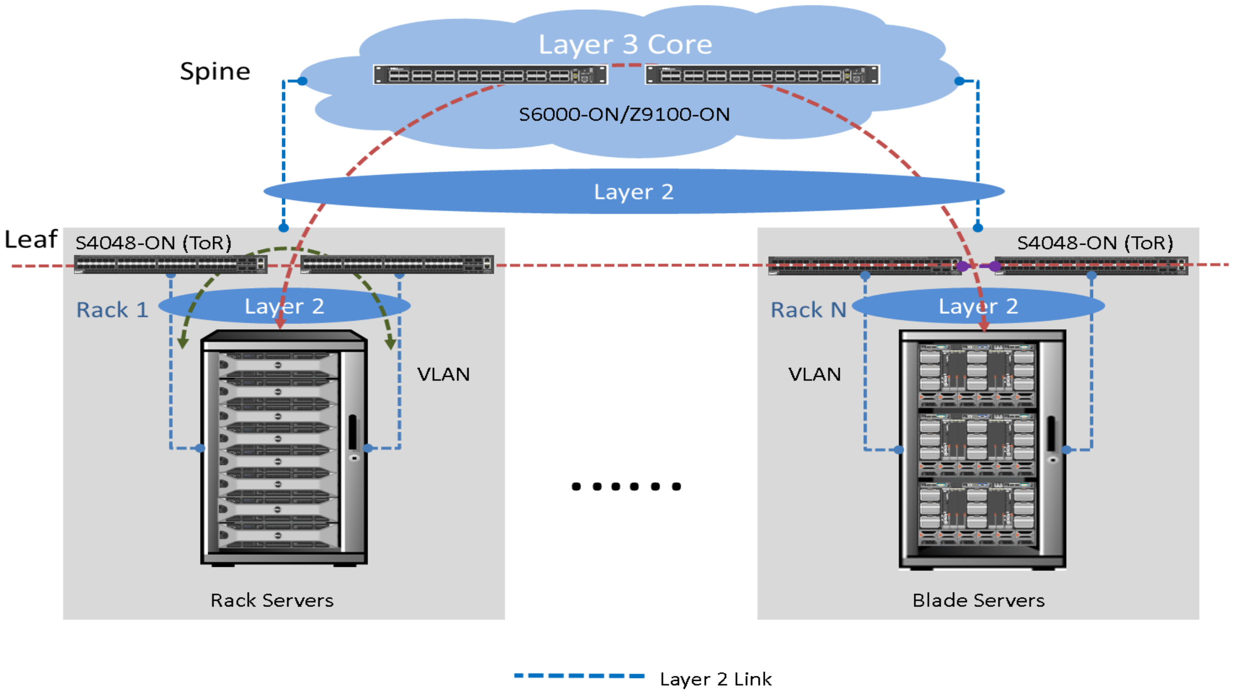
http://en.community.dell.com/techcenter/networking/m/networking_files/20443568

http://en.community.dell.com/techcenter/networking/m/networking_files/20443568

Layer 2 Inter-Rack Domain Extension

Before virtualization became widely adopted, a typical data center design would consists of racks filled with compute resources. These racks would then be connected to a set of upstream Top-of-Rack (ToR) switches providing a common Layer 2 adjacency point, if an application needs to exit the Layer 2 domain, then the ToRs are then connected a set of core Layer 3 devices providing inter-rack communication (see Figure 10.)

Figure 10 Legacy Layer 2 Data Center Design



As long as applications do not need to extend or exist across different racks, the design in figure 10 works just fine, unfortunately with the popularization of virtualization in today’s data centers, applications are no longer isolated to a single rack. The benefits of mobility, scalability, and simply increased application availability have created a need for a more efficient data center.

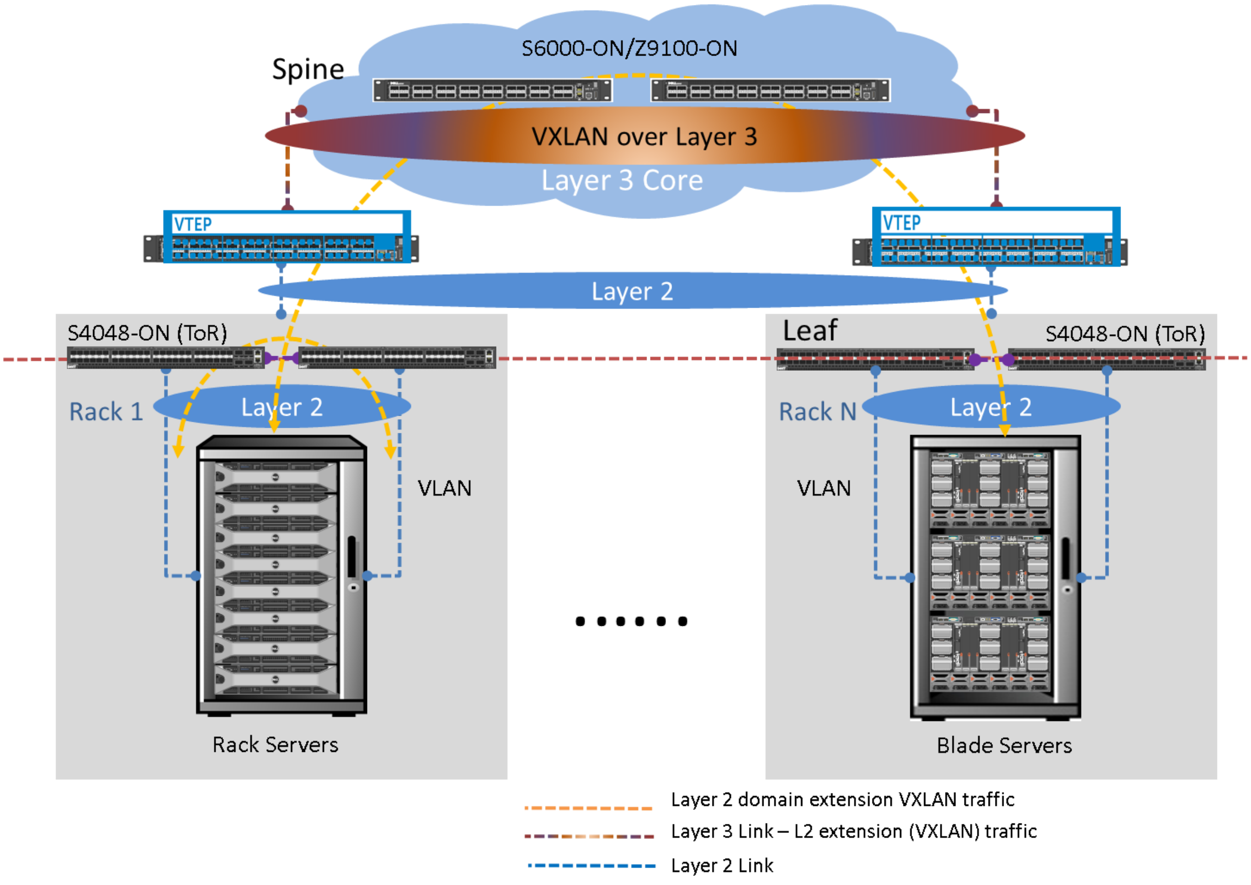
To address this new requirement, an overlay technology that could stretch a Layer 2 domain across an IP infrastructure was created. Figure 11 shows a Layer 2 intra/inter-rack application resource domain being extended across an IP infrastructure (see yellow arrows). In this design, the Dell EMC S or Z Series switch is configured as a VXLAN VTEP switch.

The VTEP switch performs the following functions:

1. Establish a dynamic or static tunnel between each VTEP switch.
2. Encapsulate and de-capsulate VXLAN header from all Layer 2 packets
    a. The encapsulation and de-capsulation process creates a VXLAN to VLAN mapping function and it is this process that allows Layer 2 domain stretch.

The links between the VTEP device and spine are straight Layer 3 links but carrying VXLAN traffic. These links can also carry regular Layer 3 traffic if needed.

Figure 11 Enhanced Layer 2 Domain Extension Data Center Design



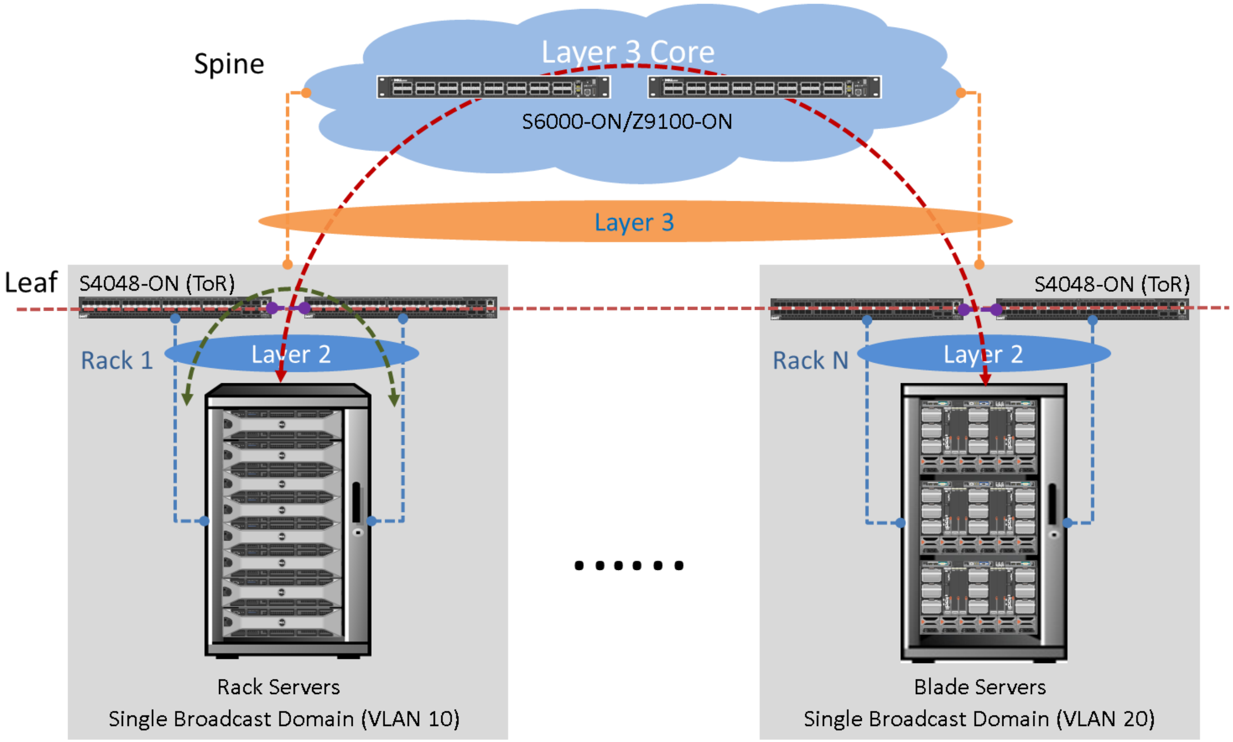
Layer 3 Inter-Rack Data Center

The traditional Layer 2 design (figure 10) provides full Layer 2 connectivity within a rack, unfortunately every Layer 2 based design will always come with inherent challenges such as scalability, and stability.

To address these challenges, a Layer 3 based design extending from the spine to the leaf layer is better suited. With a Layer 3 based design, scalability issues such as limited number of VLANs is not an issue since routing protocols and IP addresses simply scale better. As far as stability is concerned, a Layer 3 based design allows for a large unmanageable Layer 2 domain to be broken down into several smaller broadcast domains and still remain interconnected.

Figure 12 shows a typical Layer 3 based design. In this topology, the spine layer consists of a pair of Dell EMC S6000-ON or Z9100-ON. The leaf layer consists of Dell EMC S4048-ON switches acting as default gateways creating connectivity within a rack as well as Layer 3 connection to the spine layer for Layer 2 traffic that needs to exit a rack.

Figure 12 Layer 3 (Spine-to-Leaf) with Single Broadcast Domain per Rack



This design is both scalable, and stable. It is based on proven methodology before virtualization became widely deployed.

However, this design does not take into account the emerging requirement created by virtualization. Because of virtualization, applications tend to be spread across different racks but still maintain Layer 2 adjacency, creating a need for Layer 2 traffic to stretch across different racks within a data center or across data centers interconnected through an IP cloud.

Figures 13 and 14 show two variations where Layer 3 is extended from the spine to the leaf layer. Each topology show how Dell EMC's VXLAN is used to address the need of an overlay technology while keeping Layer 3's familiarity.

Figure 13 shows a topology where Layer 3 is extended from the spine to the leaf layer. At the leaf layer, the Dell EMC switches have been configured as VTEP devices providing the much needed Layer 2 stretch should an application ever needed to grow or stretch across a single rack. Furthermore, this deployment offers VTEP redundancy with dual links from the end-host.

At the VTEP level, the Dell EMC switches are not configured as a virtual VTEP switch. These VTEP switches are connected to the end-hosts via single dedicated links. It is the end-host's job to implement redundancy via software.

The Dell EMC VTEP switches will have their own respective static VXLAN tunnels configured as shown in Figure 6 and section "Configuring static VXLAN on Dell EMC S and Z Series Switches".

Figure 13 Layer 3 (Spine-to-Leaf) and Layer 2 Domain stretch – Option 1

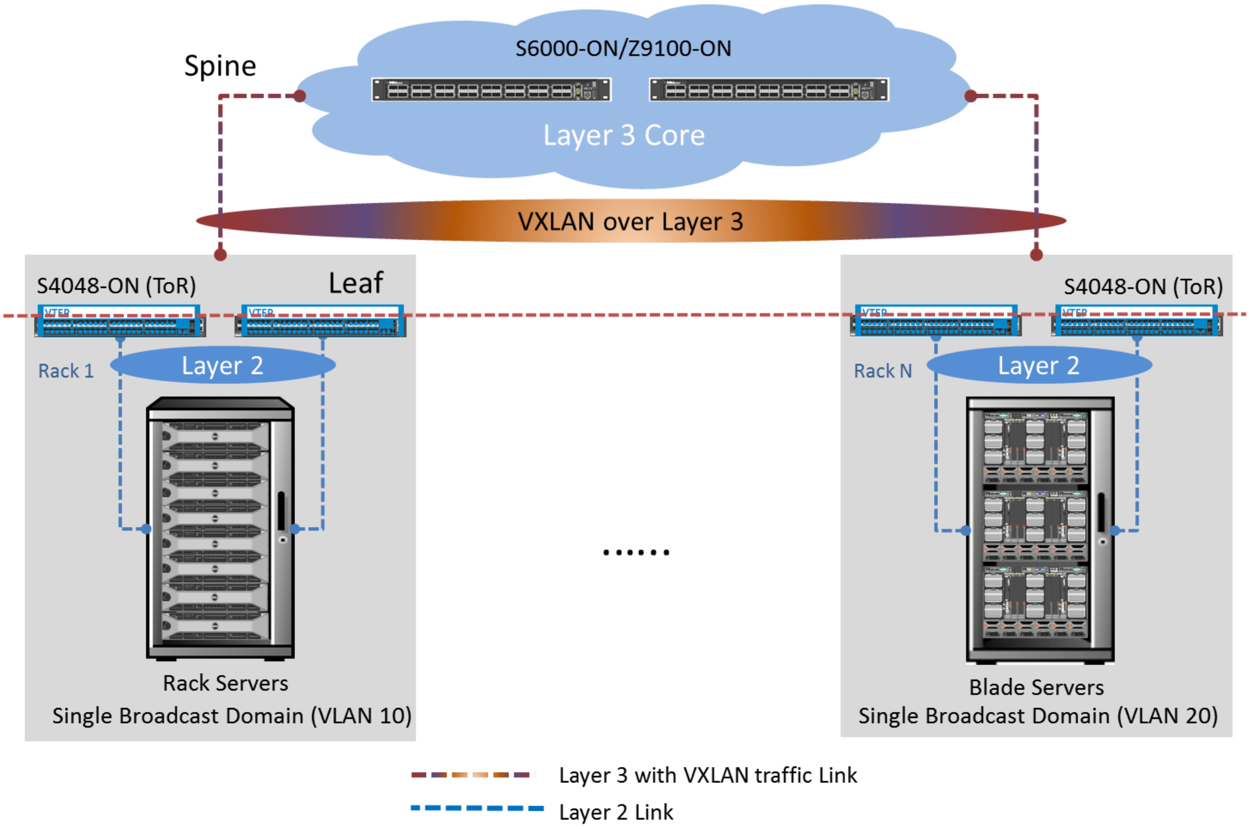
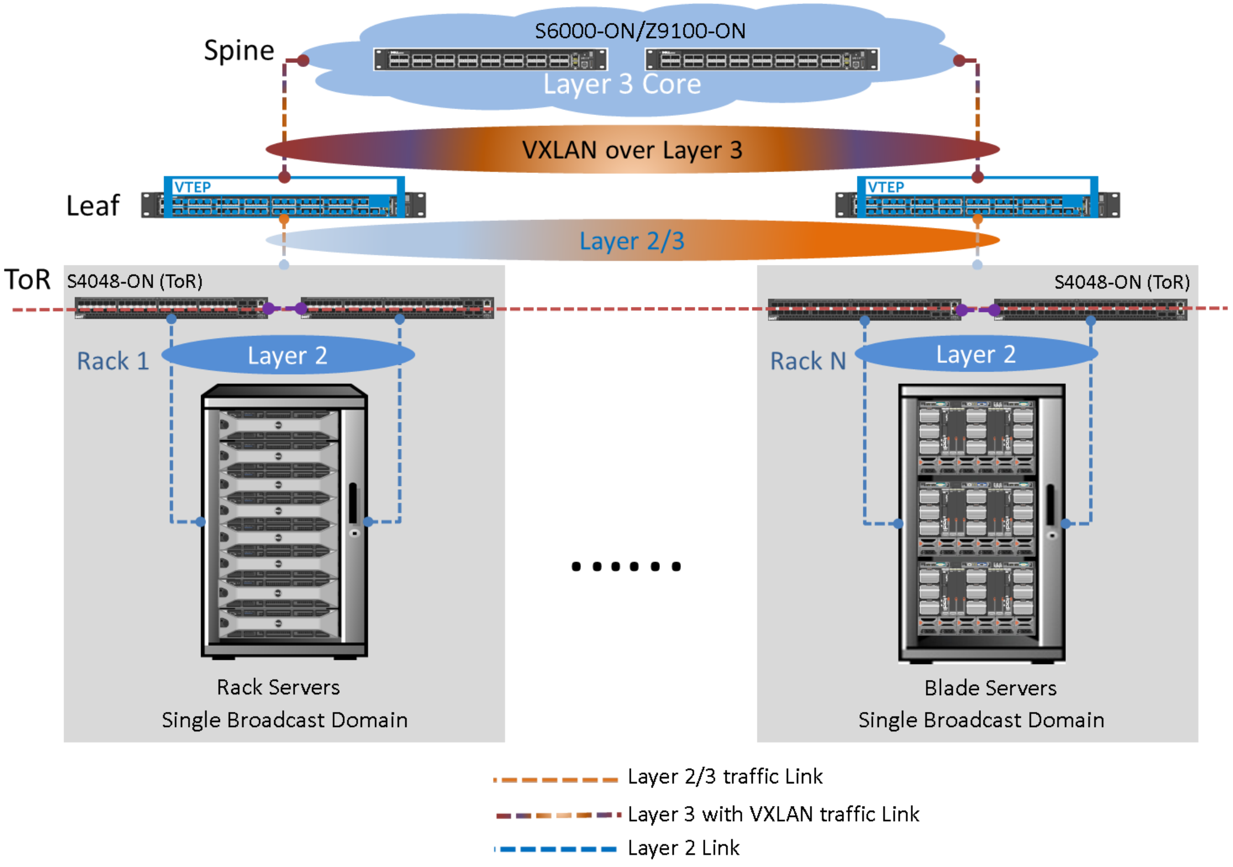


Figure 14 is a different option and moves the VTEP function from the leaf/ToR layer to a new layer. This variation achieves:

- Insertion of a simpler ToR switch without the need of VXLAN functionality at the ToR
- It potentially allows for additional horizontal growth. Depending on the type of switch being deployed as a VTEP device, the number of downlink ports (towards ToR switch) could range between tens of 100GigabitEthernet ports to hundreds of 10GigabitEthernet ports
- It keeps Layer 3 extended from the spine to the ToR while adding a central VXLAN processing layer

- It also allows for non-VXLAN-VLANs (VLANs that don't' need stretching) and VXLAN-VLANs (VLANs that need stretching) to coexist on the same device (VTEP)

Figure 14 Layer 3 (Spine-to-Leaf) and Layer 2 Domain stretch – Option 2



Layer 3 VLXAN Option

So far only Layer 2 VXLAN-to-VLAN mapping or bridging methodologies have been discussed, and to be fair it is often the most common deployment. However, similar to inter-VLAN routing, inter-VXLAN routing, and VXLAN to VLAN routing is sometimes needed.

An application residing in VXLAN 3000 may need to communicate with VXLAN 4000, and since VXLAN routing is currently not supported, a workaround using current Layer 2 and Layer 3 design methods can be used.

To address these needs, a different topology or design is needed. It still leverages previous designs such as figure 14 with one slight addition. This new addition consists of a routing or Layer 3 functional block.

Each VTEP switch is connected to the Layer 3 core in order to create a full VTEP connectivity mesh.

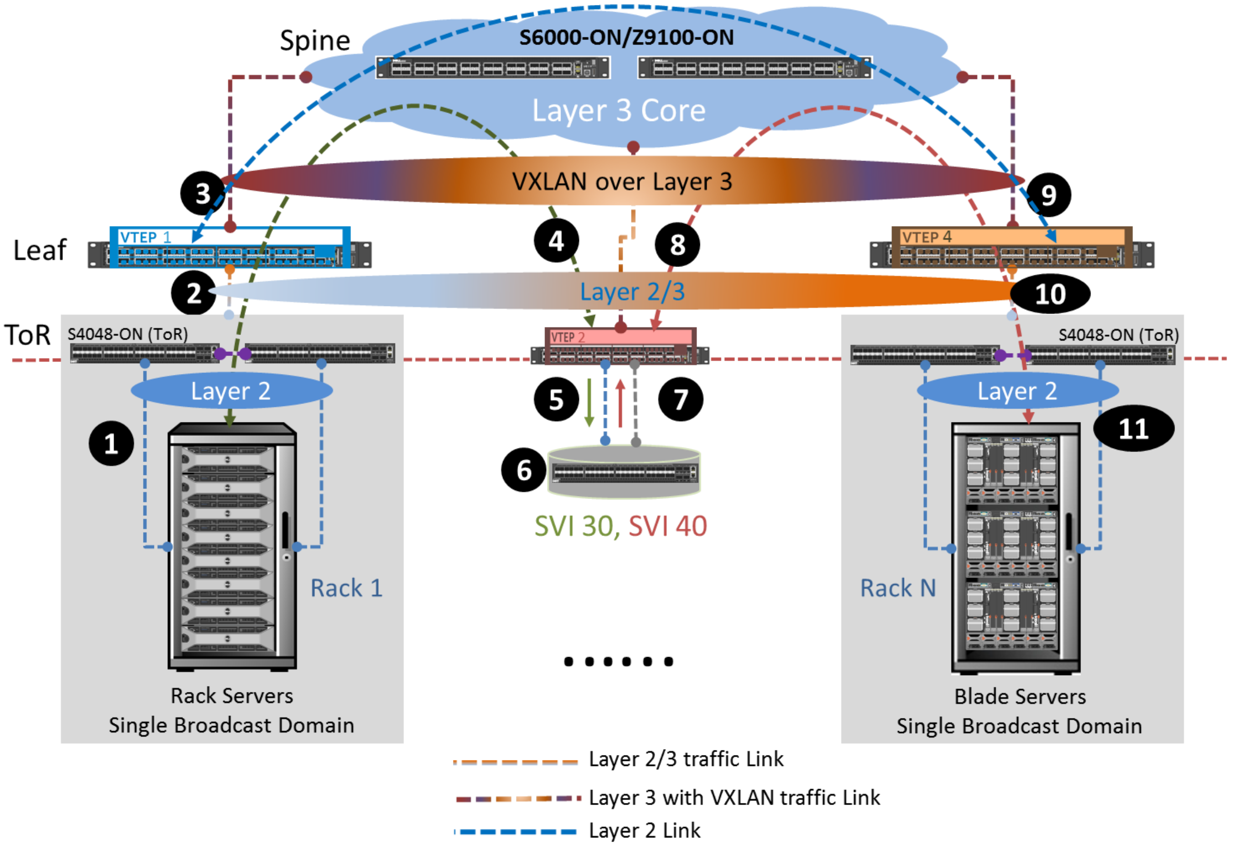
In this particular deployment, there are several static VXLAN tunnels configured. Tunnels 1 - 3 are considered intermediary tunnels needed to perform VXLAN routing. Tunnel 4, is considered a straight tunnel used to stretch an existing Layer 2 domain from either compute rack.

VXLAN Tunnel 1 (Green) – Tunnel 1 is between the first VTEP_1 and VTEP_2 (Red). This tunnel is carrying VXLAN 3000 (VLAN 30) traffic.

VXLAN Tunnel 2 (Red) – Tunnel 2 is between VTEP_2 (Red) and VTEP_4 (Orange). This tunnel is carrying VXLAN 4000 (VLAN 40) traffic.

VXLAN Tunnel 3 (Blue) – Tunnel 3 is between VTEP_1 and VTEP_4 (Orange). This tunnel is used to stretch VLANs 30 or 40 across an IP infrastructure. Tunnel 3 is straight VXLAN bridging where the routing functional block does not come in the picture.

Figure 15 VXLAN Routing – VXLAN 3000-to-VXLAN 4000, or VXLAN 3000-to-VLAN 40



Below are the steps followed as an end-host that belongs to VLAN 30 on the left hand side, pings another end-host that belongs to VLAN 40 on the right hand side. The setup in figure 15 is but one of several options that can be used to address VXLAN routing if needed.

Step 1 - Layer 2 packet(s) tagged with VLAN ID 30, source MAC_A, and destination MAC_B are sent from the end host, or VMs towards the ToR switches. The ToR switches receive these packets and perform Layer 2 switching operations. Destination MAC_B is not in the ToR switching table. This results in the packets being pushed towards the leaf switch.

Step 2 - Leaf switch hw VTEP receives Layer 2 packets with VLAN ID 30.

Step 3 - Outgoing packet(s) are encapsulated in a VXLAN header with VXLAN ID 3000 mapped to VLAN 30. The packet(s) leaving the local VTEP switch arrive at the middle VTEP part of the routing block.

Steps 4 & 5 – Incoming VXLAN packet is de-capsulated (VXLAN header info is removed), packet with VLAN ID 30 is sent to the router.

Step 6 – The router receives VLAN 30 packet whose default gateway is configured on the router. The router sees the source IP address belonging to 30.30.30.x subnet and destination subnet 40.40.40.x. The packet is routed via 40.40.40.1 configured on the router as SVI 40.

Step 7 – Traffic leaving the router is then encapsulated by VTEP_2 with VNID 4000 (**Step 8**). Tunnel 2 is used to carry this traffic.

Step 9 – Traffic is received at VTEP_4 encapsulated with VNIID 4000.

Step 10 – Egress traffic has been de-capsulated by VTEP_4 and regular VLAN 40 traffic is sent downstream towards the final destination.

Steps 1 – 11 depict how an application or VM that is part of a VXLAN domain achieves routing at the VXLAN level. The same setup can be used if VXLAN to VLAN routing is desired.

Looking ahead

Currently only VXLAN gateway and bridging functions are supported. Routing can be achieved, but as Figure 15 describes it requires additional components and data redirection. With later software releases, Dell EMC will introduce support for VXLAN routing, as well as VTEP redundancy making things simpler as far as inter-VXLAN routing and VXLAN traffic redundancy are concerned.

On top of that, Dell EMC will bring in BGP EVPN, a standards based control plane for VXLAN that overcomes the limitations of using multicast-based flood and learn control plane. Through the use of multicast-based approach, remote VTEP peer discovery as well as remote end-host learning is achieved, unfortunately, this approach can present potential scalability issues with large deployments.

BGP EVPN addresses this issue plus more such ash:

- Reduction of flooding traffic in the data center
- Full interoperability deployment across different vendors
- Better traffic engineering abilities

Conclusion

Dell EMC Networking continues to innovate and offer key solutions to our customers and partners leveraging an open standards approach.

VXLAN is a great solution for today’s data center. It builds on the traditional and well understood 802.1Q VLAN technology to address the challenges presented by today’s requirements and applications found in the data center.

Being able to bridge the connectivity gap between the ubiquitous virtual and physical environments as well as provide data center interconnect (DCI) functionality was critical in order for the typical data center to become a valued asset of any organization. Before VXLAN, organizations were limited to building isolated infrastructures where virtual environments and physical environments could only communicate with their respective counterparts. Unfortunately the typical organization’s business infrastructure runs on a mixed set of applications - virtualized and physical - and therefore a solution was needed.

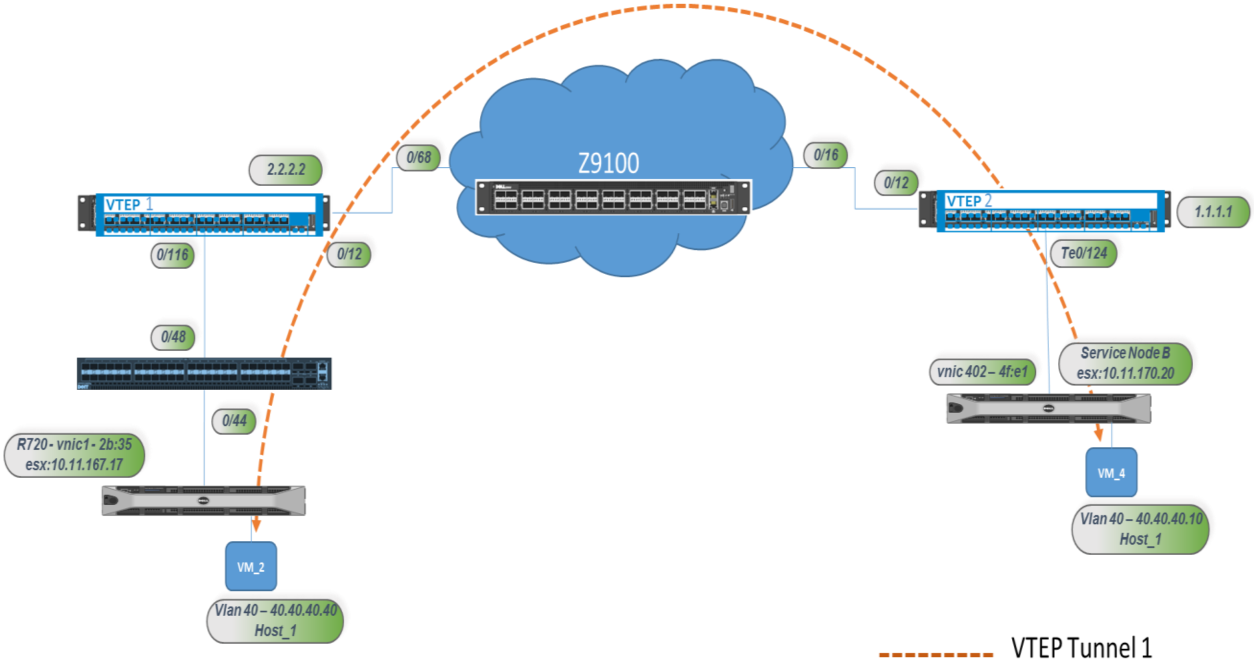
Dell EMC Networking with its data center product portfolio supports VXLAN in the hardware so customers can leverage the predictable high-performance and density expected from any large enterprise solution provider.

With Dell EMC Networking, VXLAN segments whether virtualized or physical can connect with the traditional VLAN segments allowing for multi-tenants to reside in either domain while keeping flexibility, scalability, and familiarity as part of its solution to the customer.

Appendix A: Sample Dell EMC Static VXLAN Configuration

A simple setup highlighting how simple it is to create a VXLAN tunnel between racks inside a data center or across a date center.

Figure 16 Static VXLAN Setup



VTEP_1

```
==========
VEP_1 (conf-if-range-vl-30,vl-40)#show config
feature vxlan
!
interface Vlan 30
 description vlan_2_vxlan_3000
 vxlan-vnid 3000
 no ip address
 tagged fortyGigE 0/116
 no shutdown
!
interface Vlan 40
 description vlan_2_vxlan_4000
 vxlan-vnid 4000
 no ip address
 tagged fortyGigE 0/116
 no shutdown
!
interface fortyGigE 0/116
 description Trunk_link_2_S4810
 vxlan-instance 1
 no ip address
 portmode hybrid
 switchport
!
 protocol lldp
  advertise management-tlv system-capabilities system-description system-name
  advertise interface-port-desc
 no shutdown
!
interface Loopback 0
 ip address 2.2.2.2/24
 no shutdown
!
interface fortyGigE 0/12
 description Link_2_Core_0/68
 ip address 192.168.2.1/24
!
 protocol lldp
  advertise management-tlv system-capabilities system-description system-name
  advertise interface-port-desc
 no shutdown
!
router ospf 1
```

```
 network 192.168.2.0/24 area 0
 network 2.2.2.0/24 area 0
 redistribute connected
!
vxlan-instance 1 static
  local-vtep-ip 2.2.2.2
  no shutdown
  vni-profile profile2
   vnid 3000,4000
  remote-vtep-ip 1.1.1.1 vni-profile profile2
  !
end
```

S4048_2_VTEP_1

```
==============
S4048_Rack17_Bottom(conf-if-range-vl-30,vl-40)#show config
!
interface Vlan 30
 no ip address
 tagged TenGigabitEthernet 0/44
 tagged fortyGigE 0/48
 no shutdown
!
interface Vlan 40
 no ip address
 tagged TenGigabitEthernet 0/44
 tagged fortyGigE 0/48
 no shutdown
!
interface TenGigabitEthernet 0/44
 description link_2_node_staticvxlan
 no ip address
 portmode hybrid
 switchport
 no shutdown
!
interface fortyGigE 0/48
 description link_2_S6K_1_116
 no ip address
 portmode hybrid
 switchport
!
 protocol lldp
  advertise management-tlv system-description system-name
  advertise interface-port-desc
 no shutdown
```

```
!
end
```

VTEP_2

```
==========
VTEP_2 (conf-if-range-vl-30,vl-40)#show config
!
interface Vlan 30
 description vlan_2_vxlan_3000
 vxlan-vnid 3000
 no ip address
 tagged TenGigabitEthernet 0/124
 no shutdown
!
interface Vlan 40
 description vlan_2_vxlan_4000
 vxlan-vnid 4000
 no ip address
 tagged TenGigabitEthernet 0/124
 no shutdown
!
interface TenGigabitEthernet 0/124
 description link_2_servicenodeB2_1
 vxlan-instance 1
 no ip address
 switchport
 no shutdown
!
protocol lldp
  advertise management-tlv system-capabilities system-description system-name
  advertise interface-port-desc
 no shutdown
!
interface Loopback 0
 description loopback for vtep
 ip address 1.1.1.1/24
 no shutdown
!
interface fortyGigE 0/12
 description Link_2_Z95K_0/60
 ip address 192.168.1.1/24
 no shutdown
!
protocol lldp
  advertise management-tlv system-capabilities system-description system-name
  advertise interface-port-desc
 no shutdown
!
router ospf 1
 network 192.168.1.0/24 area 0
 network 1.1.1.0/24 area 0
 redistribute connected
!
vxlan-instance 1 static
  local-vtep-ip 1.1.1.1
  no shutdown
  vni-profile profile1
   vnid 3000,4000,5000
  remote-vtep-ip 2.2.2.2 vni-profile profile1
  !
end
```