

VXLAN Design with Cisco Nexus 9300 Platform Switches

Guide

October 2014

Contents

What You Will Learn	3
VXLAN Technology Overview	3
Terminology	5
Hardware and Software Support	6
Cisco Nexus 9300 Platform as VXLAN VTEP	7
Building and Managing a Cisco Nexus 9300 VTEP	7
Configuring a Cisco Nexus 9300 VTEP	7
Inter-VLAN Tag Handling and VLAN Translation	9
Host MAC Address Management on Cisco Nexus 9300 VTEP	10
Monitoring the VXLAN Status on a Cisco Nexus 9300 VTEP	11
VXLAN Statistics on Cisco Nexus 9300 VTEP	12
Per-VXLAN VTEP Peer Statistics	12
Per-VXLAN VNI Statistics	12
Multicast Handling on a Cisco Nexus 9300 VTEP	12
Building Redundant vPC VTEPs with Cisco Nexus 9300 Platform Switches	13
Operation Fundamentals of Cisco Nexus 9300 vPC VTEPs	15
vPC Anycast VTEP Address	15
VXLAN Multicast and Broadcast, Unknown Unicast, and Multicast Traffic Handling	15
VXLAN Unicast Traffic Handling	17
vPC Consistency Check for vPC VTEPs	17
VXLAN Design Considerations for Cisco Nexus 9300 VTEPs	20
Maximum Transmission Unit Adjustment in the Underlay Network	20
Multicast Considerations for the Underlay Network	20
Multicast Rendezvous Point Configuration	20
Multicast Group Sharing by VXLAN VNIs	20
ECMP Hashing Algorithm in the Underlay Network	21
Limitation in VXLAN Topology Support: Unsupported Bud-Node Topology	21
Design Options with Cisco Nexus 9300 Platform Switches as VXLAN VTEPs	22
Inter-Pod Layer 2 Extension Design	22
Layer 2 Extension in a Layer 3 Data Center Pod Design	23
Inter-VXLAN Routing Design	25
Inter-VXLAN Routing Design Option A: Routing Block Design	25
Inter-VXLAN Routing Design Option B: VTEP-on-a-Stick Design	27
What's Next	29
Conclusion	29
For More Information	29
Appendix A: Sample Cisco Nexus 9300 VTEP Switch Configuration	30
Appendix B: Routing Block Configuration	36
VXLAN Configuration on vPC VTEPs	36
Router Configuration	37

What You Will Learn

Starting with Cisco® NX-OS Software Release 6.1(2)I2(1), Cisco Nexus® 9300 platform switches support Virtual Extensible LAN (VXLAN) bridging and gateway functions. In its initial implementation, the Cisco Nexus 9300 platform supports multicast-based VXLAN: that is, the network uses the multicast function in the underlay network to transmit the broadcast, unknown unicast, and multicast traffic of the overlay VXLAN network. This document discusses VXLAN functions on the Cisco Nexus 9300 platform and the network virtualization designs enabled by the Cisco Nexus 9300 platform as VXLAN tunnel endpoints (VTEPs). The fundamentals of VXLAN technology and multicast-based VXLAN are not within the scope of this document. For more information about these topics, please refer to: <http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html>

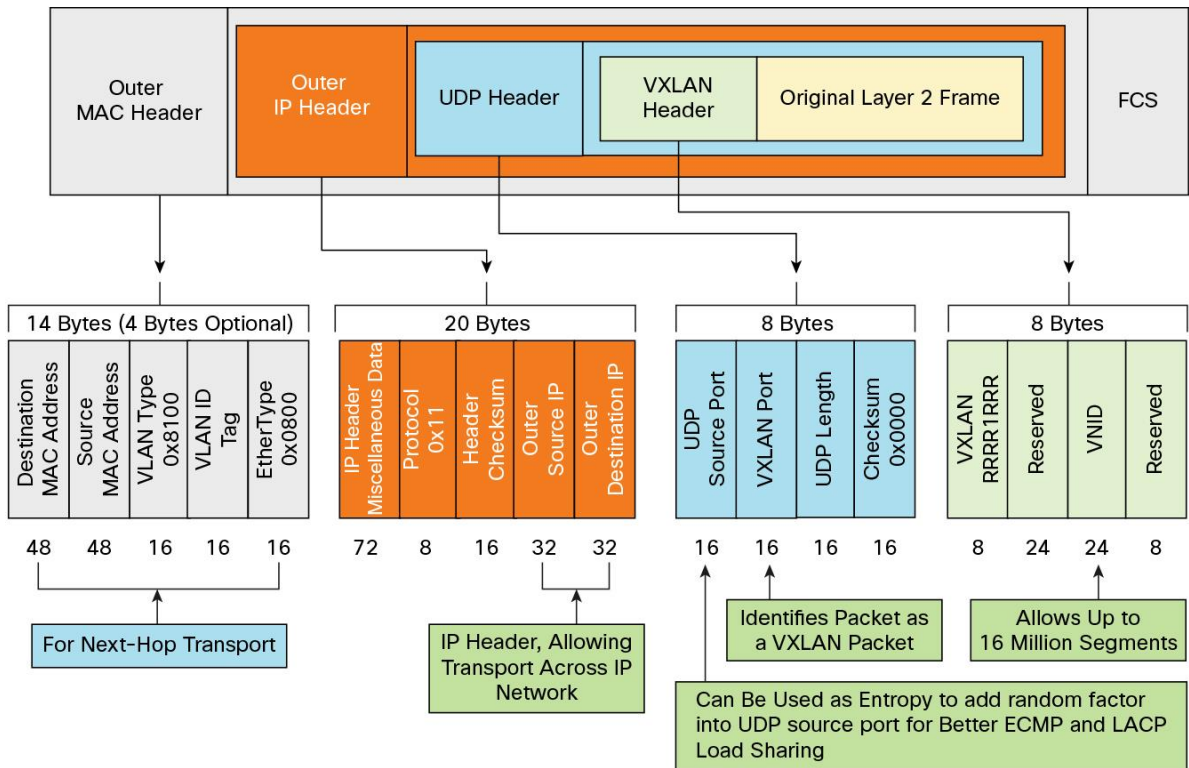
VXLAN Technology Overview

New demands are being placed on data centers every day that require them to be more efficient, optimized to reduce operating costs, scalable to support the growing demand for data, and more agile to support the applications that run on top of these environments. The industry has looked increasingly to virtualization technologies for these benefits, not only for computing and storage resources, but for network infrastructure as well.

VXLAN, one of many available network virtualization overlay technologies, offers several advantages. VXLAN is an industry-standard protocol and uses underlay IP networks. It extends Layer 2 segments over a Layer 3 infrastructure to build Layer 2 overlay logical networks. It encapsulates Ethernet frames into IP User Data Protocol (UDP) headers and transports the encapsulated packets through the underlay network to the remote VTEPs using the normal IP routing and forwarding mechanism.

Figure 1 shows the VXLAN packet format. The packet has an 8-byte VXLAN header, UDP header, outer IP header, and outer MAC header.

Figure 1. VXLAN Packet Format (MAC-in-UDP)



- **VXLAN Header:** The 24-bit VNID field in the VXLAN header identifies the VXLAN segments. It provides an expanded address spaces for Layer 2 networks.
- **UDP header:** The destination port in the UDP header indicates that the packet is a VXLAN encapsulated packet. VXLAN originally used the same UDP destination port 8472 as Overlay Transport Virtualization (OTV), based on the VXLAN IETF draft, until IANA assigned port 4789 to VXLAN. Therefore, both ports may be seen in different VXLAN implementations. The source UDP port is a hashing result based on the original Layer 2 frame head, so the source port number varies on a per-flow basis. This approach allows better per-flow load sharing of VXLAN traffic across the underlay network.
- **Outer IP header:** The source IP address in the outer IP header is the local VTEP address. The destination IP address is the remote VTEP address for known unicast traffic for the associated multicast group address in the event of broadcast, unknown unicast, and multicast traffic. The encapsulated packets will be routed through the underlay transport network based on the outer header IP addresses.
- **Outer MAC address or Layer 2 header:** This header is used to forward the encapsulated packets to the immediate next-hop device.

As a network virtualization overlay technology, VXLAN can potentially provide solutions to the following problems:

- **Layer 2 segment scalability:** VXLAN has a 24-bit virtual network identifier (VNI) field that allows up to 16 million unique Layer 2 segments in the same network. Although the current network software and hardware limitations reduce the usable VNI scale in actual deployments, the VXLAN protocol by design has at least lifted the 4096-VLAN limitation in the traditional IEEE 802.1q VLAN name space. This change enables organizations to build data center networks with more Layer 2 segments, such as in large multitenant network environments.
- **Layer 2 domain scalability:** Many data center applications have a simple view of the network world and require Layer 2 adjacency among their endpoint hosts. The growth of these applications necessitates the stretching of the Layer 2 domains within the data center. But large Layer 2 domains mean large broadcast and failure domains. To maintain network stability and control the impact of any network failure, Layer 2 domains can't be too big. This restriction, however, conflicts with the growth of applications. VXLAN can solve this dilemma by decoupling Layer 2 domains from the network infrastructure. The infrastructure is built as a Layer 3 fabric that doesn't rely on Spanning Tree Protocol for loop prevention or topology convergence. The Layer 2 domains reside on the overlay, with isolated broadcast and failure domains. This approach allows the data center network to grow without risking creation of too large a failure domain.
- **Layer 2 segment elasticity over Layer 3 boundary:** Data center networks are often built with multiple Layer 2 pods that are interconnected through a Layer 3 aggregation layer. The application workload is put into individual pods for Layer 2 connectivity. This approach imposes rigid limitations on application workload placement within the data center network. With the MAC-in-IP-UDP tunneling mechanism, VXLAN can build Layer 2 virtual networks across the underlay Layer 3 infrastructure. Application endpoint hosts can be flexibly placed in the data center network without concern for the Layer 3 boundary of the underlay infrastructure and while maintaining Layer 2 adjacency within a VXLAN overlay network.

Terminology

The following definitions will help you understand VXLAN.

- **Virtual network identifier (VNI) or VXLAN segment ID:** The system uses the VNI, also called the VXLAN segment ID, along with the VLAN ID to identify the Layer 2 segments in the VXLAN overlay network.
- **VXLAN segment:** A VXLAN segment is a Layer 2 overlay network over which endpoint devices, including physical devices and virtual machines, communicate through a direct Layer 2 adjacency.
- **VXLAN tunnel endpoint (VTEP):** The VTEP originates and terminates VXLAN tunnels. The VTEP encapsulates the end-host Layer 2 frames within an IP header to send them across the IP transport network and decapsulates VXLAN packets received from the underlay IP network to forward them to local end hosts. The end hosts are unaware of the VXLAN. There are two types of VTEPs:
 - **Virtual VTEP:** Software-based VTEP; an example is a VXLAN-capable virtual switch within a hypervisor host
 - **Physical VTEP:** Hardware-based VTEP; Cisco Nexus 9300 platform switches are physical VTEPsA physical VTEP provides hardware-based high performance and the capability to bridge VXLAN segments with traditional VLAN segments and to extend a Layer 2 segment over a Layer 3 infrastructure.

- **VXLAN gateway:** A VXLAN gateway connects VXLAN and traditional VLAN environments. A physical VTEP device can provide a hardware-based VXLAN gateway function. Figure 2 shows an example in which a hypervisor VTEP initiates VXLAN tunnels on one side and a physical VTEP device on the other side provides VXLAN gateway service to terminate the VXLAN tunnel and map the VXLAN VNI to a traditional VLAN.

Figure 2. VXLAN Gateway



- **VXLAN bridging:** VXLAN bridging is the function provided by VTEP devices to extend a VLAN or VXLAN VNI over the Layer 3 infrastructure. Figure 3 shows VLAN-to-VLAN and VXLAN-to-VXLAN bridging.

Figure 3. VXLAN Bridging



- **VXLAN routing:** VXLAN routing is also referred to as inter-VXLAN routing. It provides IP routing service between two VXLAN VNIs in the overlay network in a way similar to inter-VLAN routing. Figure 4 shows the logical concept of VXLAN routing.

Figure 4. VXLAN Routing



Hardware and Software Support

The solutions described in this document use Cisco Nexus 9300 platform switches as physical VXLAN VTEPs. They require the following hardware and software:

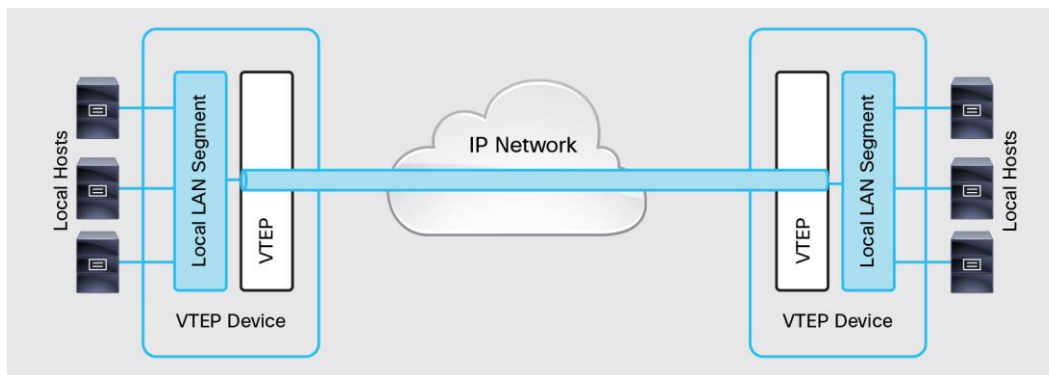
- Cisco Nexus 9300 platform switches must be used as the VTEP devices in a VXLAN topology.
- Cisco NX-OS Release 6.1(2)I2(2b) or later on the Cisco Nexus 9300 VTEP switches is recommended. Although the Cisco Nexus 9300 platform began supporting VXLAN functions in Cisco NX-OS Release 6.1(2)I2(1), many enhancements were added in Release 6.1(2)I2(2b).
- The VXLAN function doesn't require an additional license. However, the underlay network needs the appropriate licenses for Interior Gateway Protocol (IGP) routing and IP multicast functions.

- Cisco Nexus 9500 platform switches (or switches in other platforms that provide the same or similar 10 and 40 Gigabit Ethernet port density and performance) as the spine (or aggregation-layer) devices in the underlay network.

Cisco Nexus 9300 Platform as VXLAN VTEP

A physical VTEP device plays two roles: the normal Layer 2 switching function in the local VLAN, and extension of the local VLAN through VXLAN encapsulation to remote sites in the same Layer 2 segment. For hosts in the local VLAN, the VTEP is a normal Layer 2 switch. The remote VTEP devices, the device functions as a VTEP peer initiating and terminating VXLAN tunnels. Figure 5 shows the logical functions of a VTEP device.

Figure 5. Physical VTEP Functions



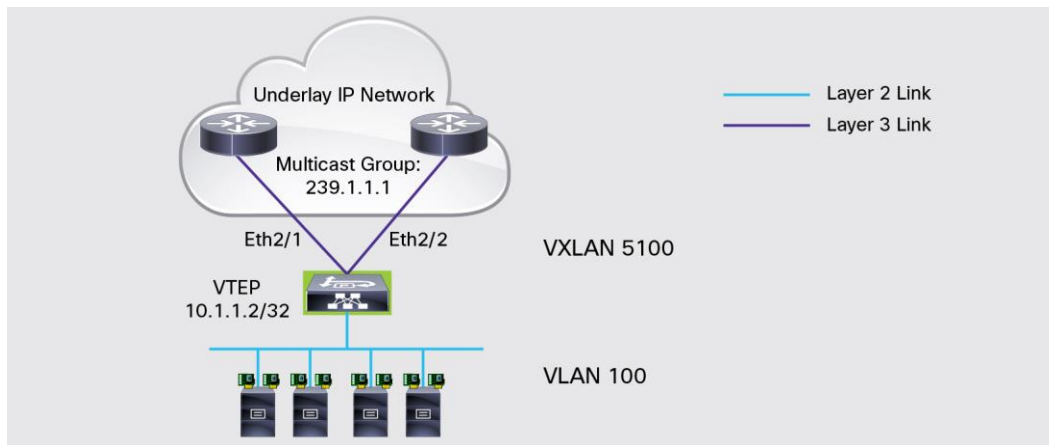
Building and Managing a Cisco Nexus 9300 VTEP

This section demonstrates how to configure and manage a Cisco Nexus 9300 VTEP.

Configuring a Cisco Nexus 9300 VTEP

This section uses the sample topology shown in Figure 6 to demonstrate the steps for configuring a Cisco Nexus 9300 platform switch as a VTEP.

Figure 6. Cisco Nexus 9300 VTEP Example



To configure a Cisco Nexus 9300 platform switch as a VTEP device, follow these steps:

Step 1. Enable the VXLAN feature.

```
feature nv overlay
feature vn-segment-vlan-based
```

Step 2. Map VLANs to VXLAN VNIs.

The following example maps VLAN 100 to VXLAN VNI 5100:

```
vlan 100
  vn-segment 5100
```

Step 3. Create a loopback interface with a/32 IP address.

This IP address will be used as the switch VTEP address. Currently, a Cisco Nexus 9300 platform switch can have only one VTEP address. The VTEP address needs to be advertised through the underlay IGP routing protocol so that VTEP devices can obtain IP reachability on their VTEP addresses. IP multicast routing needs to be enabled on the loopback interface.

The following example creates **interface loopback0** with IP address 10.1.1.2/32 and advertises it in **ospf 1 area 0**. Protocol-Independent Multicast (PIM) sparse-mode is enabled under the loopback interface.

```
interface loopback0
  ip address 10.1.1.2/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

Step 4. Create the network virtualization endpoint (NVE) interface that is used as the VXLAN tunnel interface.

Under the NVE interface, add VXLAN VNIs and associate them with underlay multicast groups.

The following example configures interface **nve1** with **loopback0** as the tunnel source interface. VNI 5100 is added under the **nve1** interface and associated with multicast group 239.1.1.1. This means VXLAN VNI 5100 will use multicast group 239.1.1.1 in the underlay network to carry overlay unknown unicast, multicast, and broadcast traffic.

```
interface nve1
  no shutdown
  source-interface loopback0
  overlay-encapsulation VXLAN
  member vni 5100 mcast-group 239.1.1.1
```

In addition to the preceding steps for VXLAN-related configuration, you need to configure IGP routing and multicast routing on the VTEP switch and in the underlay network. In this example, Open Shortest Path First (OSPF) is configured as the IGP, and IP PIM sparse-mode is enabled on the uplink interfaces and the **loopback0** interface. Automatic rendezvous point (**auto-rp**) is configured as the PIM rendezvous-point discovery protocol. The two upstream routers serve as anycast rendezvous points. The underlay network configuration follows the best practices for the selected IGP and for multicast routing, and so the specific configuration isn't shown here.

The following is the relevant configuration on the Cisco Nexus 9300 VTEP switch:

```
feature pim
feature ospf

interface Ethernet2/1
 ip address 192.168.1.6/30
 ip ospf network point-to-point
 ip router ospf 1 area 0.0.0.0
 ip pim sparse-mode
 no shutdown
interface Ethernet2/2
 ip address 192.168.1.10/30
 ip ospf network point-to-point
 ip router ospf 1 area 0.0.0.0
 ip pim sparse-mode
 no shutdown

interface loopback0
 ip address 10.1.1.2/32
 ip router ospf 1 area 0.0.0.0
 ip pim sparse-mode

router ospf 1
 router-id 10.1.1.2

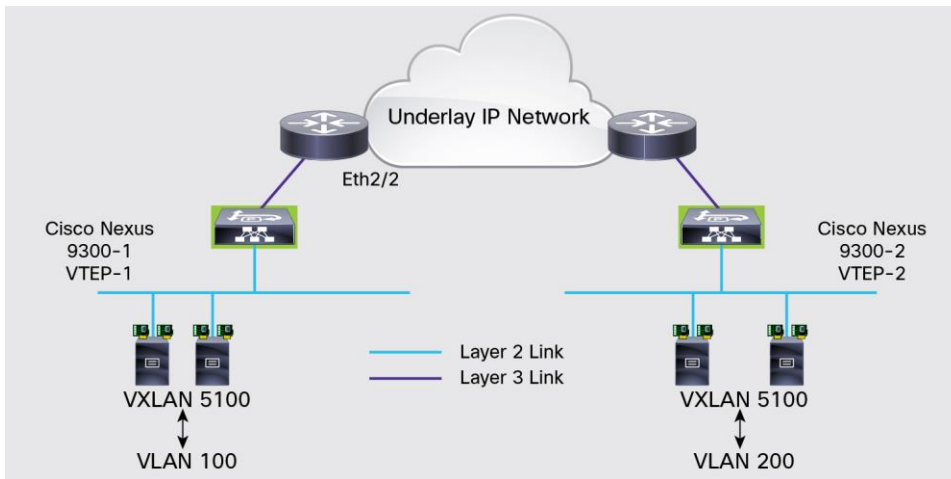
ip pim ssm range 232.0.0.0/8
ip pim auto-rp listen
```

Note: Appendix A shows the complete configuration for the sample Cisco Nexus 9300 VTEP switch.

Inter-VLAN Tag Handling and VLAN Translation

Cisco Nexus 9300 platform switches follow the VXLAN IETF standard for VXLAN implementation. According to the VXLAN IETF draft, the ingress VTEP device shall remove the IEEE 802.1Q VLAN tag in the original Layer 2 packet if there is one before encapsulating the packet into the VXLAN format to transmit it through the underlay network. The remote VTEP devices have information about the VLAN in which the packet will be placed based on their own VLAN-to-VXLAN VNI mapping configurations. With this mechanism, VTEP devices for the same VXLAN VNI could possibly map a VXLAN VNI to different VLANs. Figure 7 shows an example in which Cisco Nexus 9300 VTEP-1 maps VLAN 100 to VXLAN VNI 5100, whereas Cisco Nexus 9300 VTEP-2 maps VLAN 200 to VXLAN VNI 5100. As a result, VLAN 100 behind VTEP-1 and VLAN 200 behind VTEP-2 are bridged into one Layer 2 domain in the overlay network, and hosts within these two vlans gain direct Layer 2 adjacency. This VLAN translation through the VXLAN overlay can provide solutions for organizations that want to connect different traditional VLANs together: for instance, after a company acquisition or during internal merging or data center migration.

Figure 7. Inter-VLAN Tag Handling and VLAN Translation by Cisco Nexus 9300 VTEPs



Host MAC Address Management on Cisco Nexus 9300 VTEP

A Cisco Nexus 9300 VTEP learns and stores the MAC addresses of local hosts and remote hosts in the MAC address table. Local host MAC address entries are programmed with the physical port information. For remote hosts that the VTEP learned from remote VTEP peers, the NVE interface (VXLAN tunnel interface) with the peer VTEP address is listed as the port information.

The MAC address entries for both local and remote hosts are subject to the same aging mechanism. By default, the MAC address table aging timer is 1800 seconds. It is user configurable with the **mac address-table aging-time** configuration command. The range is 120 to 91800 seconds.

An example of the MAC address table configuration in a Cisco Nexus 9300 VTEP switch is shown here:

```
n9396-vtep-1# sh mac address-table
Legend:
```

```
* - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
age - seconds since last seen,+ - primary entry using vPC Peer-Link,
(T) - True, (F) - False
```

VLAN	MAC Address	Type	age	Secure	NTFY	Ports
* 100	0000.0c07.ac64	dynamic	0	F	F	nve1(10.1.1.100)
* 100	0000.ced0.41cf	dynamic	0	F	F	Eth1/1
* 100	0000.ced1.072f	dynamic	0	F	F	nve1(10.1.1.3)
* 100	6412.2574.6ae7	dynamic	0	F	F	nve1(10.1.1.100)
* 100	6412.2574.9eb7	dynamic	0	F	F	nve1(10.1.1.100)
* 101	0000.0c07.ac65	dynamic	0	F	F	nve1(10.1.1.100)
* 101	0000.ced0.41d1	dynamic	0	F	F	Eth1/1
* 101	0000.ced1.0731	dynamic	0	F	F	nve1(10.1.1.3)
* 101	6412.2574.6ae7	dynamic	0	F	F	nve1(10.1.1.100)
* 101	6412.2574.9eb7	dynamic	0	F	F	nve1(10.1.1.100)
G -	7c69.f6df.e597	static	-	F	F	sup-eth1(R)

Local hosts

Remote hosts

Monitoring the VXLAN Status on a Cisco Nexus 9300 VTEP

A Cisco Nexus 9300 VTEP maintains the information about VXLAN VNI segments and active VTEP peers. The following example shows a VTEP device that has two VXLAN VNIs, 5100 and 5101, provisioned, using multicast group 239.1.1.1 and 239.1.1.2, respectively. It also has two active remote VTEP peers.

```
n9396-vtep-1# sh nv vni
```

Interface	VNI	Multicast-group	VNI State
nve1	5100	239.1.1.1	up
nve1	5101	239.1.1.2	up

```
n9396-vtep-1# sh nv peers
```

Interface	Peer-IP	Peer-State
nve1	10.1.1.3	Up
nve1	10.1.1.100	Up

```
n9396-vtep-1# sh nv peers detail
```

```
Peer: 10.1.1.3
```

```
Interface      : nve1
Peer learnt VNI : 5100
Configured VNIs : 5100-5101
Provision State : add-complete
Route Update   : Yes
Uptime         : 10:41:46
```

```
Peer: 10.1.1.100
```

```
Interface      : nve1
Peer learnt VNI : 5100
Configured VNIs : 5100-5101
Provision State : add-complete
Route Update   : Yes
Uptime         : 10:41:35
```

Multicast-based VXLAN peer status is data based. A remote VTEP peer appears in the peer database only when traffic is received from that peer. After the traffic stops, the peer will age out over time. The peer aging timer is linked to the aging timer of the host MAC address entries. When the last MAC address entry associated with this peer ages out, the peer will be removed from the peer database as well.

VXLAN Statistics on Cisco Nexus 9300 VTEP

Cisco NX-OS provides per-VXLAN VNI segment statistics and per-VTEP peer statistics on a Cisco Nexus 9300 VTEP. The command-line interface (CLI) monitoring commands and sample output are shown here.

Per-VXLAN VTEP Peer Statistics

```
n9396-vtep-1# sh nv peers 10.1.1.3 interface nve 1 counters
Peer IP: 10.1.1.3
TX
    3954189 unicast packets 4136035670 unicast bytes
    0 multicast packets 0 multicast bytes
RX
    3941862 unicast packets 4123118616 unicast bytes
    0 multicast packets 0 multicast bytes
```

Per-VXLAN VNI Statistics

```
n9396-vtep-1# sh nv vni 5100 counters
VNI: 5100
TX
    3717075 unicast packets 3886820534 unicast bytes
    2 multicast packets 140 multicast bytes
RX
    3708372 unicast packets 3873416710 unicast bytes
    5 multicast packets 600 multicast bytes
```

Multicast Handling on a Cisco Nexus 9300 VTEP

Multicast-based VXLAN uses the IP multicast function in the underlay network to transport the broadcast, multicast, and unknown unicast traffic of the overlay Layer 2 segments. A VXLAN VNI is associated with a multicast group in the transport network. Multiple VNIs can share a multicast group. Each VTEP device is both a receiver and a source in the multicast group it joins. Following the usual IP PIM sparse-mode practice, if a VNI segment contains n number of VTEP devices, the multicast group for this VNI will have n (S, G) states and 1 (*, G) state. Therefore, as the number of VNI segments grows, and as the number of multicast groups associated with VXLAN VNIs grows, the IP multicast route table size may become very large, presenting challenges for multicast scalability and increasing the complexity of multicast operations and management.

To help reduce the growth of the multicast route table on a VTEP switch, Cisco NX-OS for Cisco Nexus 9300 platform switches is implemented with (S, G) shortest-path tree (SPT) suppression for multicast groups that are associated with VXLAN VNIs. With this enhancement, when a Cisco Nexus 9300 platform switch functions as a VTEP, it uses the shared rendezvous-point tree (*, G) entry to receive multicast encapsulated VXLAN traffic from other VTEP devices in the same VNI, instead of switching to the SPT for each remote VTEP source. This feature eliminates the need to maintain an (S, G) entry for every remote VTEP. As a result, a Cisco Nexus 9300 VTEP needs to maintain only two multicast states for each VNI-associated multicast group: the (*, G) entry for the shared tree with the rendezvous point as the root, and the local (S, G) entry for the local source tree.

The (*, G) entry is built toward the rendezvous point of the multicast group. Its outgoing interface is the VXLAN NVE1 interface. This entry is used to receive VXLAN multicast encapsulated packets and to perform decapsulation.

The local (S, G) entry has the local VTEP address (the loopback/32 address) as the source, and the uplink interfaces towards the rendezvous point as the outgoing interface. This entry is used to perform VXLAN multicast encapsulation and to send the encapsulated multicast packets to the underlay network toward the rendezvous point of the multicast group.

The following is the output of **show ip mroute** from the sample Cisco Nexus 9300 VTEP. It shows the multicast states for group 239.1.1.1, which is used for VXLAN VNIs. The local VTEP address is 10.1.1.2.

```
n9396-vtep-1# sh ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 11:31:35, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0, uptime: 11:31:35
  Outgoing interface list: (count: 0)

(*, 239.1.1.1/32), uptime: 11:31:34, nve ip pim
  Incoming interface: Ethernet2/1, RPF nbr: 192.168.1.1, uptime: 11:29:36
  Outgoing interface list: (count: 1)
    nve1, uptime: 11:31:34, nve

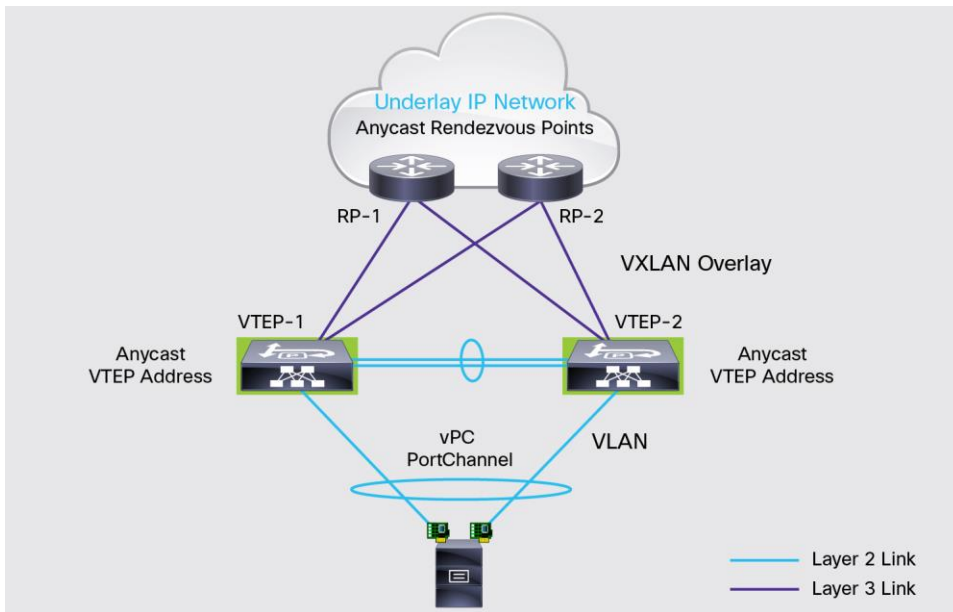
(10.1.1.2/32, 239.1.1.1/32), uptime: 11:31:34, nve ip mrib pim
  Incoming interface: loopback0, RPF nbr: 10.1.1.2, uptime: 11:30:38
  Outgoing interface list: (count: 1)
    Ethernet2/1, uptime: 11:30:29, pim
```

Note that multicast encapsulation is used only to transmit the broadcast, unknown unicast, and multicast traffic of the overlay VXLAN network. Known unicast traffic is unicast encapsulated and transmitted using the VTEP address.

Building Redundant vPC VTEPs with Cisco Nexus 9300 Platform Switches

Cisco Nexus 9300 platform switches support VTEP redundancy by allowing a pair of virtual PortChannel (vPC) switches to function as a logical VTEP device sharing an anycast VTEP address (shown in Figure 8). The vPC switches provide vPC for redundant host connectivity while individually running Layer 3 protocols with the upstream devices in the underlay network. They both join the multicast group for the same VXLAN VNI and use the same anycast VTEP address as the source to send VXLAN encapsulated packets. To the devices in the underlay network, including the multicast rendezvous point and the remote VTEP devices, the two vPC VTEP switches appear to be one logical VTEP entity.

Figure 8. Cisco Nexus 9300 Platform Switches as vPC VTEPs



To configure vPC VTEP, use the following steps:

Step 1. Configure vPC switches and vPC for host connectivity.

Use the standard vPC configuration procedure. Use Cisco recommended best practices for vPC.

Step 2. Enable the VXLAN feature.

```
feature nv overlay
feature vn-segment-vlan-based
```

Step 3. Configure a loopback interface with/32 as the secondary address.

The primary address on a loopback interface will likely be used as the router ID by network protocols, such as OSPF and Border Gateway Protocol (BGP). In this case, two switches can't have the identical primary loopback address. Therefore, vPC VTEP uses an identical secondary address on the loopback interface between the two switches as the anycast VTEP address. Some examples are shown here.

On vPC Switch-1:

```
interface loopback0
  no ip redirects
  ip address 10.1.1.4/32
  ip address 10.1.1.100/32 secondary ← anycast VTEP address
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

On vPC Switch-2:

```
interface loopback0
  no ip redirects
```

```
ip address 10.1.1.5/32
ip address 10.1.1.100/32 secondary ← anycast VTEP address
ip ospf network point-to-point
ip router ospf 1 area 0.0.0.0
```

Step 4. Configure VXLAN following the normal VTEP configuration steps.

```
vlan 100
  vn-segment 5100
interface nve1
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1
```

Note: Appendix B shows the complete configuration of the sample Cisco Nexus 9300 vPC VTEP switches.

vPC peers must have the following identical configurations:

- Consistent mapping of the VLAN to the virtual network segment (VN-segment)
- Consistent NVE binding to the same loopback secondary IP address (anycast VTEP address)
- Consistent VNI-to-group mapping.

Operation Fundamentals of Cisco Nexus 9300 vPC VTEPs

This section discusses the basic operations of Cisco Nexus 9300 vPC VTEPs.

vPC Anycast VTEP Address

Cisco Nexus 9300 vPC VTEP switches use a secondary IP address on the loopback interface bound to the VXLAN NVE tunnel as the anycast VTEP address. The two vPC switches need to have the exact same secondary loopback IP address. They both advertise this anycast VTEP address on the underlay network so that the upstream devices learn the/32 route from both vPC VTEPs and can load-share VXLAN unicast encapsulated traffic between them.

In the event of vPC peer-link failure, the vPC operational secondary switch will shut down its loopback interface bound to VXLAN NVE. This shutdown will cause the secondary vPC switch to withdraw the anycast VTEP address from its IGP advertisement so that the upstream devices in the underlay network start to send all traffic just to the primary vPC switch. The purpose of this process is to avoid a vPC active-active situation when the peer link is down. With this mechanism, the orphan devices connected to the secondary vPC switch will not be able to receive VXLAN traffic when the vPC peer link is down.

VXLAN Multicast and Broadcast, Unknown Unicast, and Multicast Traffic Handling

Both vPC VTEP switches independently send the IP PIM register to the rendezvous point for the multicast group of the VXLAN VNI. They both source the register packets from the anycast VTEP address. They each install the corresponding (*, G) entry in their multicast route tables with NVE1 in the output interface (OIF) list.

The rendezvous-point device will see at least two Equal-Cost Multipath (ECMP) routes to the anycast VTEP address: one to each VTEP switch. It will choose one of the paths to send an (S, G) join. S here is the anycast VTEP address. The VTEP switch that receives the (S, G) join from the rendezvous point will install the uplink interface toward the rendezvous point in its (S, G) OIF list. It will become the designated forwarder (DF) for the group. It'll use the (S, G) OIF list to encapsulate VXLAN multicast packets and send them out to the underlay network through the uplink interfaces.

Only the designated-forwarder (DF) VTEP switch performs VXLAN multicast encapsulation and decapsulation. For multicast encapsulated traffic coming from remote VTEP peers, the rendezvous point will always forward it to the designated-forwarder VTEP switch for decapsulation followed by local forwarding. The designated-forwarder VTEP switch uses its (*, G) OIF list for decapsulation. For local-host-generated broadcast, unknown unicast, and multicast traffic that needs to go out to remote VTEP devices, a local host can load-share the traffic between the two VTEP switches through a vPC. In this case, the non-designated-forwarder VTEP switch will forward the traffic across the vPC peer link to the designated-forwarder VTEP switch. The designated-forwarder VTEP switch will perform VXLAN multicast encapsulation using its local (S, G) OIF list and send the encapsulated multicast packets toward the rendezvous point.

The following example shows multicast entries for a VXLAN VNI multicast group on the vPC VTEP switches.

```
vtep-2# sh ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 1d05h, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0, uptime: 1d05h
  Outgoing interface list: (count: 0)

(*, 239.1.1.1/32), uptime: 1d05h, nve pim ip
  Incoming interface: Ethernet2/1, RPF nbr: 192.168.1.13, uptime: 1d05h
  Outgoing interface list: (count: 1)
    nve1, uptime: 1d05h, nve
```

```
(10.1.1.100/32, 239.1.1.1/32), uptime: 1d05h, nve pim mrrib ip
  Incoming interface: loopback0, RPF nbr: 10.1.1.100, uptime: 1d05h
  Outgoing interface list: (count: 1)
    Ethernet2/1, uptime: 1d05h, pim
```

```
vtep-2#
```

This VTEP is the designated forwarder for group 239.1.1.1. It performs VXLAN multicast encapsulation and decapsulation for the associated VXLAN VNIs.

```
vtep-1# sh ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 1d05h, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0, uptime: 1d05h
  Outgoing interface list: (count: 0)

(*, 239.1.1.1/32), uptime: 1d05h, nve ip pim
  Incoming interface: Ethernet2/1, RPF nbr: 192.168.1.9, uptime: 1d05h
  Outgoing interface list: (count: 1)
    nve1, uptime: 1d05h, nve
```



```
(10.1.1.100/32, 239.1.1.1/32), uptime: 1d05h, nve pim mrib ip
Incoming interface: loopback0, RPF nbr: 10.1.1.100, uptime: 1d05h
Outgoing interface list: (count: 0)
```

```
vtep-1#
```

This VTEP is not the designated forwarder for group (239.1.1.1). It won't perform VXLAN multicast encapsulation and decapsulation for the associated VXLAN VNIs.

The underlay network multicast is often designed with redundant rendezvous points, such as anycast rendezvous points. The redundant rendezvous-point devices may independently send (S, G) joins to either of the vPC VTEP devices. For example, in Figure 8, RP-1 could send its join to VTEP-1, and RP-2 could send to VTEP-2. In this case, an election process between the two switches will select one of them as the designated forwarder, the only one that performs VXLAN multicast encapsulation and decapsulation.

Note: This designated forwarder election process doesn't exist in Cisco NX-OS Release 6.1(2)I2(2a) or 6.1(2)I2(3) or earlier releases. In these releases, for organizations that want to deploy VXLAN with vPC VTEP, the underlay network multicast needs to have a single rendezvous point to avoid potential loops for VXLAN multicast encapsulated traffic.

VXLAN Unicast Traffic Handling

For known unicast traffic, both vPC VTEP switches perform encapsulation and decapsulation.

Traffic from a local host to a remote host will traverse the VTEP device in the direction from the access port to the underlay network uplink. The VTEPs will encapsulate unicast packets with the anycast VTEP address as the source and the remote VTEP address as the destination in the outer IP header.

Traffic from a remote host to a local host will arrive on the uplink interfaces of the VTEP devices. The encapsulated unicast packets have the remote VTEP address as the source and the local anycast VTEP address as the destination in the outer IP header. Both VTEP switches perform decapsulation and forward packets to the local destination host.

vPC Consistency Check for vPC VTEPs

The vPC consistency check is a mechanism used by the two switches configured as a vPC pair to exchange and verify their configuration compatibility. This check is essential for the correct operation of vPC functions. As Cisco Nexus 9300 platform switches support vPC VTEPs, the VXLAN configuration components are added to the vPC consistency check.

VLAN-to-VXLAN VN-segment mapping is a type-1 consistency check parameter. The two VTEP switches are required to have identical mappings. VLANs that have mismatched VN-segment mappings will be suspended. If graceful consistency check is enabled, the primary vPC switch will keep the problematic VLANs up while the secondary vPC switch suspends them. If graceful consistency check is disabled, both vPC switches will suspend the VLANs.

The following situations will be detected as inconsistencies:

- One switch has a VLAN mapped to a VN-segment (VXLAN VNI), and the other switch doesn't have a mapping for the same VLAN.
- The two switches have a VLAN mapped to different VN-segments.

The following is an example of **show** command output when two vPC switches have VLAN 300 mapped to different VN-segments.

```
n9396-vtep-4# sh vpc consistency-parameters global
```

Legend:

Type 1 : vPC will be suspended in case of mismatch

Name	Type	Local Value	Peer Value
Vlan to Vn-segment Map	1	4 Relevant Map(s)	4 Relevant Map(s)
STP Mode	1	Rapid-PVST	Rapid-PVST
STP Disabled	1	None	None
STP MST Region Name	1	""	""
STP MST Region Revision	1	0	0
STP MST Region Instance to	1		
VLAN Mapping			
STP Loopguard	1	Disabled	Disabled
STP Bridge Assurance	1	Enabled	Enabled
STP Port Type, Edge	1	Normal, Disabled,	Normal, Disabled,
BPDUFILTER, Edge BPDUGuard		Disabled	Disabled
STP MST Simulate PVST	1	Enabled	Enabled
Allowed VLANs	-	1,100-101,200,300	1,100-101,200,300
Local suspended VLANs	-	300	-

n9396-vtep-4#

n9396-vtep-4#

```
n9396-vtep-4# sh vpc consistency-parameters vlans
```

Name	Type	Reason Code	Pass Vlans
Vlan to Vn-segment Map	1	vPC type-1 configuration incompatible - vn_segment inconsistent	0-299,301-4095
STP Mode	1	success	0-4095
STP Disabled	1	success	0-4095
STP MST Region Name	1	success	0-4095
STP MST Region Revision	1	success	0-4095
STP MST Region Instance to	1	success	0-4095
VLAN Mapping			
STP Loopguard	1	success	0-4095
STP Bridge Assurance	1	success	0-4095
STP Port Type, Edge	1	success	0-4095

```
BPDUFilter, Edge BPDUGuard
STP MST Simulate PVST      1      success      0-4095
Pass Vlans                  -                0-299,301-4095
n9396-vtep-4#
```

```
n9396-vtep-4# sh vpc br
```

Legend:

(*) - local vPC is down, forwarding via vPC peer-link

```
vPC domain id           : 100
Peer status              : peer adjacency formed ok
vPC keep-alive status   : peer is alive
Configuration consistency status : success
Per-vlan consistency status : failed
Type-2 inconsistency reason : Consistency Check Not Performed
vPC role                 : primary
Number of vPCs configured : 1
Peer Gateway             : Enabled
Dual-active excluded VLANs : -
Graceful Consistency Check : Disabled
Auto-recovery status     : Disabled
```

vPC Peer-link status

```
-----
id  Port  Status Active vlans
--  ---  -----
1   Po1   up    1,100-101,200
```

VLAN 300 is suspended

vPC status

```
-----
id  Port  Status Consistency Reason
--  ---  -----
10  Po10  up    success      success
```

Active vlans

1,100-101,200

```
n9396-vtep-4#
```

VXLAN Design Considerations for Cisco Nexus 9300 VTEPs

This section presents points to consider when designing a VXLAN network for Cisco Nexus 9300 VTEPs.

Maximum Transmission Unit Adjustment in the Underlay Network

VXLAN adds a 50-byte overhead in total, including:

- 8-byte VXLAN header
- 8-byte UDP header
- 20-byte outer IP header
- 14-byte outer MAC header

To avoid exceeding the maximum transmission unit (MTU) size while sending VXLAN encapsulated packets through the underlay network, you should increase the MTU size in the underlay network by 50 bytes, or enable jumbo frames.

Multicast Considerations for the Underlay Network

Multicast Rendezvous Point Configuration

The current VXLAN support on Cisco Nexus 9300 platform switches uses multicast in the underlay network to transit broadcast, multicast, and unknown unicast traffic on the overlay Layer 2 segments. The best practices for common multicast configuration apply to multicast for VXLAN. Two considerations for multicast in VXLAN are rendezvous-point placement and the protocol of choice. Redundant rendezvous points in general are recommended, with rendezvous points placed in a central location for the data paths, such as aggregation layer or spine layer. The rendezvous-point protocol of choice depends on the protocols supported by all the devices in the underlay network that will be in the multicast forwarding data paths. On the Cisco Nexus 9000 platform, supported rendezvous-point protocols include Static-RP, Auto-RP, and Bootstrap Router (BSR). Both PIM-based anycast rendezvous points and Multicast Source Discovery Protocol (MSDP)-based anycast rendezvous points are supported on the Cisco Nexus 9000 Series Switches for rendezvous-point redundancy.

Multicast Group Sharing by VXLAN VNIs

The VXLAN implementation on Cisco Nexus 9300 platform switches uses multicast tunnels for broadcast, unknown unicast, and multicast traffic forwarding. Ideally, one-to-one mapping between VXLAN VNIs and IP multicast groups provides the optimal multicast forwarding. However, with this one-to-one mapping, an increase in the number of VXLAN VNIs or in the number of VTEP devices causes a parallel increase in the required multicast address space and the multicast forwarding table size in the underlay network. At some point, the underlay network multicast scalability and management can become challenging. In this case, mapping multiple VXLAN segments to a single multicast group can help conserve multicast control-plane resources on the transport network devices. This mapping provides an easier way to achieve the desired VXLAN scalability without adding too much burden to the underlay network for multicast scaling.

Note, however, that this multicast group sharing comes with the cost of suboptimal multicast forwarding. Packets forwarded to the multicast group for one VNI are now sent to the VTEPs of other VNIs that are sharing the same multicast group. This approach causes inefficient use of multicast data plane resources. Therefore, this solution is a trade-off between control-plane scalability and data-plane efficiency.

Despite the suboptimal multicast replication and forwarding, having multiple VNIs share a multicast group does not have any implications for the Layer 2 isolation between the VXLAN VNI networks. After receiving an encapsulated packet from the multicast group, a VTEP checks and validates the VNID in the VXLAN header of the packet. The VTEP discards the packet if the VNID is unknown to it. Only when the VNID matches one of the VTEP's local VXLAN VNIDs does the VTEP accept and decapsulate the packets for further lookup and forwarding to its local hosts within this VXLAN VNI. Other VNI networks will not receive the packet. Thus, the segregation between VXLAN VNI networks is not compromised.

ECMP Hashing Algorithm in the Underlay Network

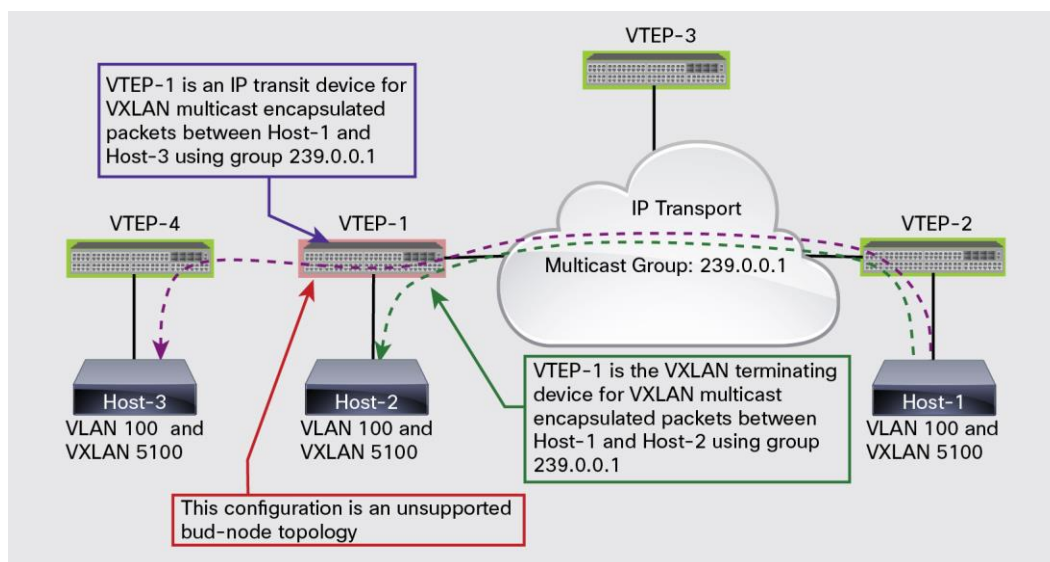
As described in a previous section, Cisco Nexus 9000 Series Switches form the source UDP port for a VXLAN traffic flow by hashing the original Layer 2 and Layer 3 packet header. This hashing helps ensure that each VXLAN encapsulated flow can be uniquely identified by the five tuples in its outer IP header (source address and port, destination address and port, and protocol). To achieve the best load sharing and distribution for VXLAN traffic transmission, the underlay network should use a five-tuple-based hashing algorithm for ECMP and Link Aggregation Control Protocol (LACP).

Limitation in VXLAN Topology Support: Unsupported Bud-Node Topology

Because of hardware limitations of the network forwarding engine (NFE), Cisco Nexus 9300 platform switches currently don't support the bud-node topology, illustrated in Figure 9.

A bud node is a device that is a VXLAN VTEP device and at the same time an IP transit device for the same multicast group used for VXLAN VNIs. In Figure 9, multicast group 239.0.0.1 is used for VXLAN VNIs. For VXLAN multicast encapsulated traffic from Host-1 to Host-2, VTEP-1 performs a multicast reverse-path forwarding (RPF) check in group 239.0.0.1 and then VXLAN decapsulation. For VXLAN multicast encapsulated traffic from Host-1 to Host-3 using the same group 239.0.0.1, VTEP-1 is an IP transit device for the multicast packets. It performs RPF check and IP forwarding based on the outer IP header that has 239.0.0.1 as the destination. When these two different roles collide on the same device, the device becomes a bud node. Because of hardware limitations on the NFE, Cisco Nexus 9300 platform switches can't perform the two roles simultaneously. Therefore, designers need to avoid the bud-node topology in VXLAN network design using Cisco Nexus 9300 platform switches as VTEPs.

Figure 9. VXLAN Bud-Node Topology



Notes:

- The NFE is based on well-adopted merchant silicon. The other switch platforms that use the same application-specific integrated circuit (ASIC) are subject to the same hardware limitations for bud-node support. Readers of this document should be aware of this limitation and should avoid the bud-node topology when designing a VXLAN network using these platforms.
- Cisco will soon introduce a software solution for this hardware limitation on the NFE by using the internally developed application leaf engine (ALE) ASIC that resides in the Cisco Nexus 9300 platform switches as well. When this software support becomes available, the bud-node topology will be supported by Cisco Nexus 9300 VTEP switches.

Design Options with Cisco Nexus 9300 Platform Switches as VXLAN VTEPs

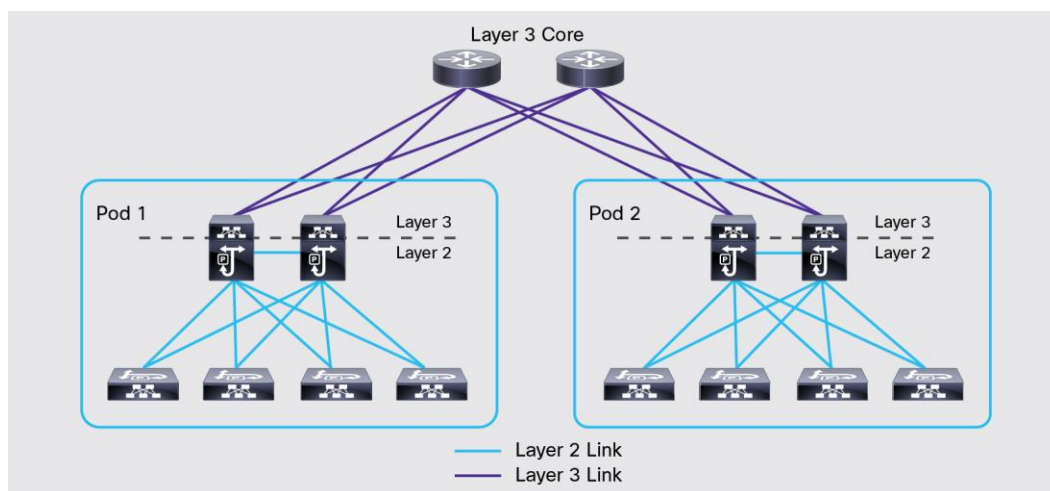
Cisco Nexus 9300 platform switches began supporting VXLAN gateway and bridging functions in Cisco NX-OS Release 6.1(2)I2(1). Cisco NX-OS Release 6.1(2)I2(b) is the latest release for the Cisco Nexus 9300 platform as of the writing of this document. Prior to this release, Cisco Nexus 9300 platform switches don't support VXLAN routing. The hardware is capable of supporting VXLAN routing, but the software implementation will be available in a future release.

The design discussions in this section are based on the currently available VXLAN functions on the Cisco Nexus 9300 platform: that is, the multicast-based VXLAN gateway and bridging functions. This section will be revised when the VXLAN routing function and Ethernet VPN (EVPN) control plane become available on Cisco Nexus 9300 platform switches.

Inter-Pod Layer 2 Extension Design

Traditionally for applications that require Layer 2 adjacency among end hosts, the typical Layer 2 pod design is used in the data center network (Figure 10). In a Layer 2 pod, a pair of aggregation switches serve as the Layer 2 and Layer 3 boundary in the network, and the access switches run Layer 2-only switching functions. Network services, such as firewalls and load balancers, are often attached to the aggregation switches. Either the aggregation switches or the service appliances function as the default IP gateway for application VLANs that reside in the pod. Traffic between different pods needs to pass through the Layer 3 boundary to be routed by the aggregation switches or service appliances.

Figure 10. Traditional Layer 2 Data Center Pods



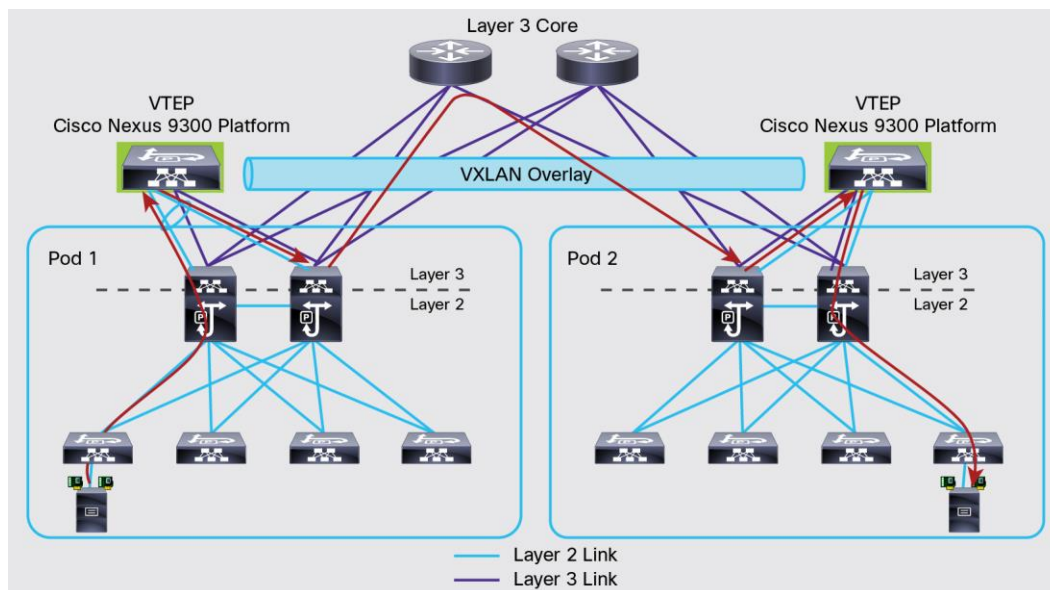
This design works for traditional data center applications. However, with application workloads increasingly virtualized, support for workload mobility and flexibility becomes a new requirement for data center networks. Now the Layer 2 segments need to stretch across the Layer 3 boundary between pods.

As an overlay technology for Layer 2 extension over a Layer 3 network, VXLAN is one solution that can meet this requirement. Figure 11 shows the inter-pod Layer 2 extension design using Cisco Nexus 9300 platform switches as the VXLAN VTEPs to interconnect application VLANs in the overlay network between different pods. The figure shows only one Cisco Nexus 9300 VTEP switch attached to the aggregation switches in each pod, but a pair of Cisco Nexus 9300 vPC VTEPs can be attached for redundancy.

In this design, the Cisco Nexus 9300 VTEP switches in the data center pods are configured as part of the local VLANs that need to be extended through VXLAN. They then map the local VLANs to VXLAN VNIs that are connected together by the VXLAN tunnels between them. The VXLAN tunnels traverse the Layer 3 part of the data center network, including the pod aggregation-layer switches and the data center core switches. Figure 16 shows the traffic forwarding path for the extended VLANs between two pods.

The aggregation switches continue to act as the default gateway for these VLANs. Therefore, routed traffic will traverse a pod directly through the aggregation switches without going through the VTEP switches.

Figure 11. Inter-Pod Layer 2 Extension

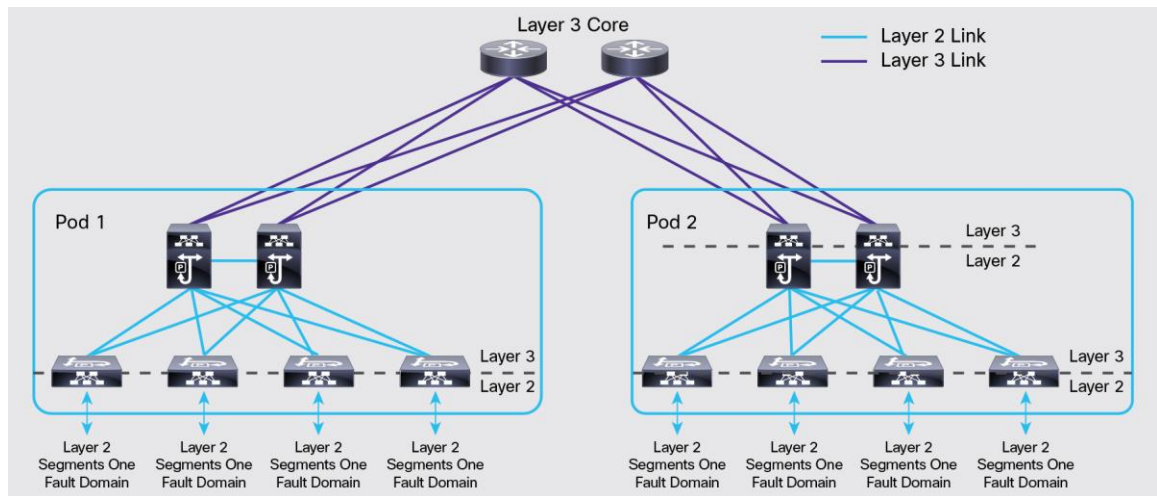


Layer 2 Extension in a Layer 3 Data Center Pod Design

The traditional Layer 2 pod design provides Layer 2 adjacency within a pod, but it presents several design challenges related to stability and scalability. A Layer 2 pod is a single Layer 2 broadcast and fault domain. The commonly deployed Layer 2 protocols, such as Spanning Tree Protocol, are not as stable or scalable as Layer 3 routing protocols. As a Layer 2 domain grows, its stability decreases, and the impact of a failure within a fault domain increases. You should try to keep the fault domain size under control or reduce the Layer 2 domain size while the data center network continues to grow.

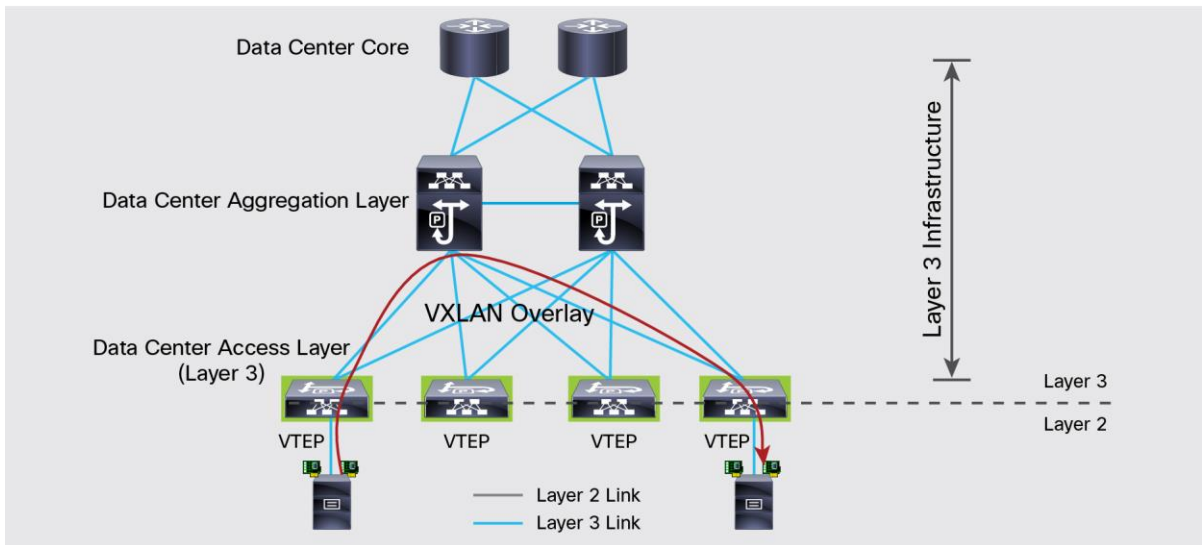
One approach that is increasingly common is a Layer 3 pod design that runs Layer 3 routing protocols all the way to the access switches. As a result, Layer 2 domains are well contained under each access switch, as shown in Figure 12. If the access layer consists of top-of-rack (ToR) access switches, a Layer 2 domain is just within a server rack. This design reduces the network fault domain size and greatly increases the stability of the data center network. Multiple instances of an application can be easily placed into separate network fault domains so that failure in one domain will not jeopardize the availability of the entire application. This approach enables a pod to grow beyond the size that Layer 2 protocols can stably sustain.

Figure 12. Layer 3 Pod Design



If part of an application in a Layer 3 pod needs Layer 2 adjacency among its hosts attached to different access switches, or if a Layer 2 domain needs to grow beyond a single server rack space, a Layer 2 extension technology such as VXLAN is needed to provide a Layer 2 overlay on top of the Layer 3 pod infrastructure. Figure 13 shows such a solution that uses the VXLAN bridging function on Cisco Nexus 9300 platform switches. In this design, Cisco Nexus 9300 platform switches are deployed as Layer 3 ToR switches for server-access connectivity and as VXLAN VTEP devices to extend the Layer 2 segments between racks.

Figure 13. Layer 2 Extension within a Layer 3 Pod



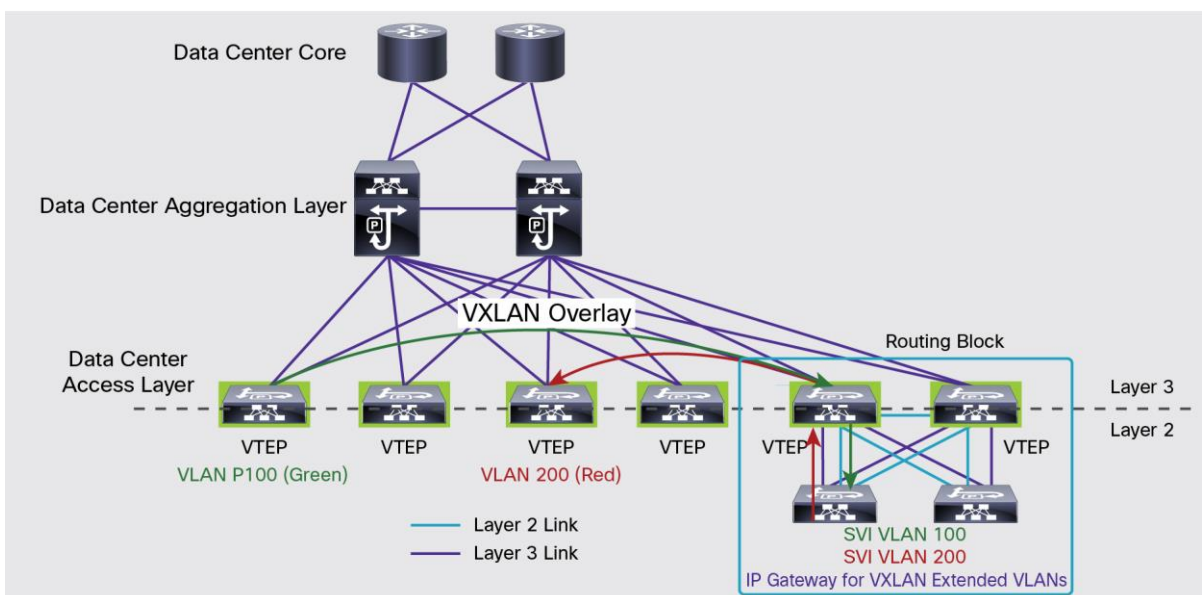
Inter-VXLAN Routing Design

As with traditional VLAN environment, routing between VXLAN segments or from VXLAN to VLAN segments is required in many situations. Because the current Cisco NX-OS releases (Release 6.1(2)I2(3) and earlier) don't support VXLAN routing, specific designs need to be applied to achieve this network function.

Inter-VXLAN Routing Design Option A: Routing Block Design

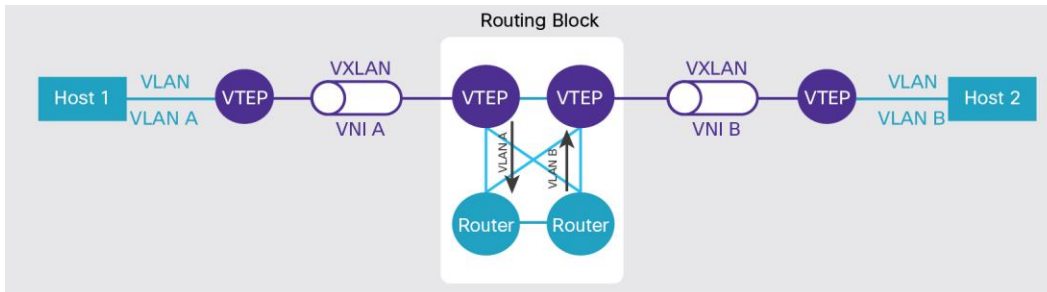
Figure 14 depicts a VXLAN routing solution by adding a routing block to the Layer 3 pod network. The routing block has a router-on-a-stick design consisting of a VTEP or a pair of vPC VTEPs to terminate VXLAN tunnels, and one or a pair of routers that serve as the IP gateway for the VXLAN-extended VLANs and perform routing functions for these VLANs.

Figure 14. Routing Block Design for VXLAN Routing



For Layer 2 traffic within a VXLAN VNI, the traffic will go directly between the local VTEP and the remote VTEPs. For Layer 3 routed traffic between VXLAN VNIs, the traffic will first reach the IP gateway of the source VXLAN VLAN IP subnet that is on the routers in the routing block and will be routed to the destination VXLAN VLAN IP subnet by the gateway router. The gateway router will then forward the packets back to the VTEP in the routing block for encapsulation in the destination VXLAN and forwarding toward the destination host. The logical traffic flow is shown in Figure 15.

Figure 15. Traffic Flow in Routing-Block VXLAN Routing Design



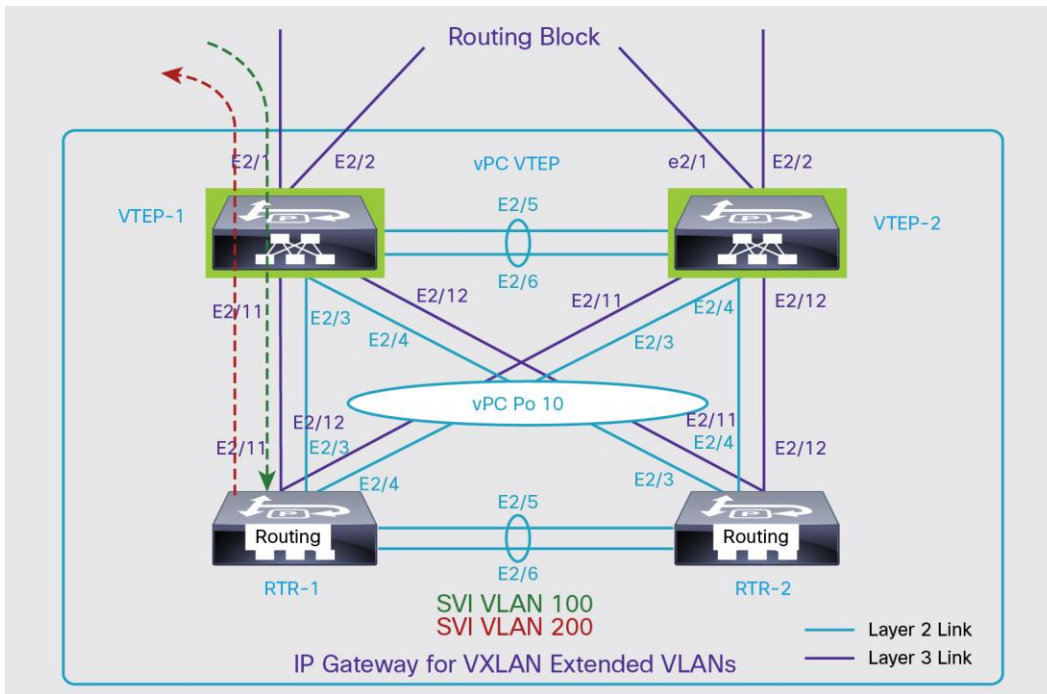
Routing Block Configuration

The routing block in the recommended design for VXLAN routing consists of a physical VTEP or vPC VTEP pair that converts VXLAN VNIs back to VLANs, and a router or a pair of routers that functions as the IP gateway for the VLAN IP subnets and routes between VLAN IP subnets. For device redundancy, redundant VTEP devices, such as a pair of Cisco Nexus 9300 as vPC VTEPs and a pair of routers running a first-hop redundancy protocol such as Hot Standby Router Protocol (HSRP), are recommended.

Figure 16 shows a sample VXLAN routing block that is designed with two pairs of Cisco Nexus 9300 platform switches. One pair of Cisco Nexus 9300 platform switches functions as a vPC VTEP that maps between the VXLAN and VLAN. The second pair is an IP gateway for the VXLAN-extended VLANs. There is a double-sized vPC between the two pairs of switches for Layer 2 connectivity. A separate set of Layer 3 links can be installed for routing between the VXLAN VLAN to non-VXLAN VLANs or an IP network. The relevant configuration of the devices in the routing block is provided in Appendix A.

Note: Because of a known software issue, the peer links of the vPC VTEPs and the Layer 2 links to the routers in the routing block can't be on the 40 Gigabit Ethernet links of Cisco Nexus 9300 platform switches before Cisco NX-OS Release 6.1(2)I2(2a). This problem is fixed in Cisco NX-OS Release 6.1(2)I2(2a).

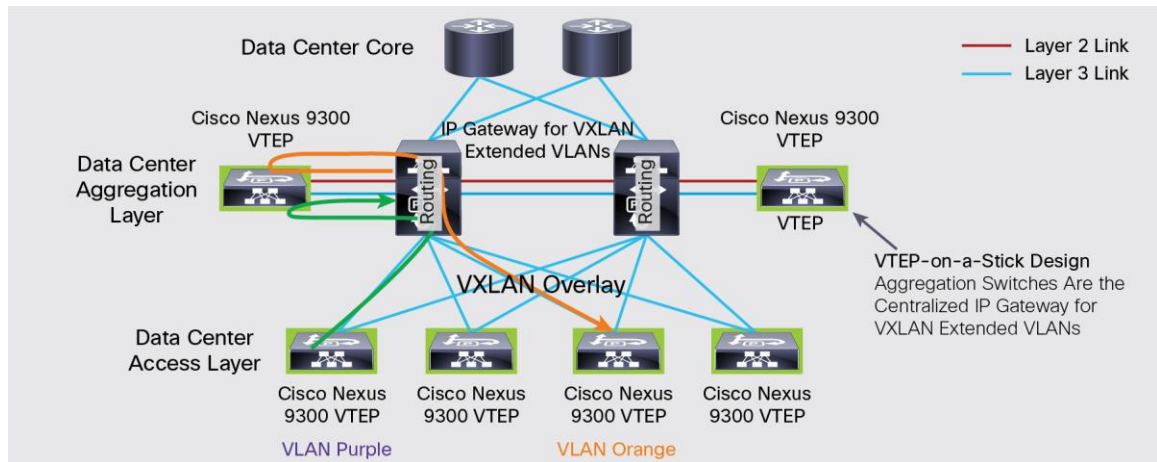
Figure 16. Routing Block Design with Cisco Nexus 9300 Platform Switches



Inter-VXLAN Routing Design Option B: VTEP-on-a-Stick Design

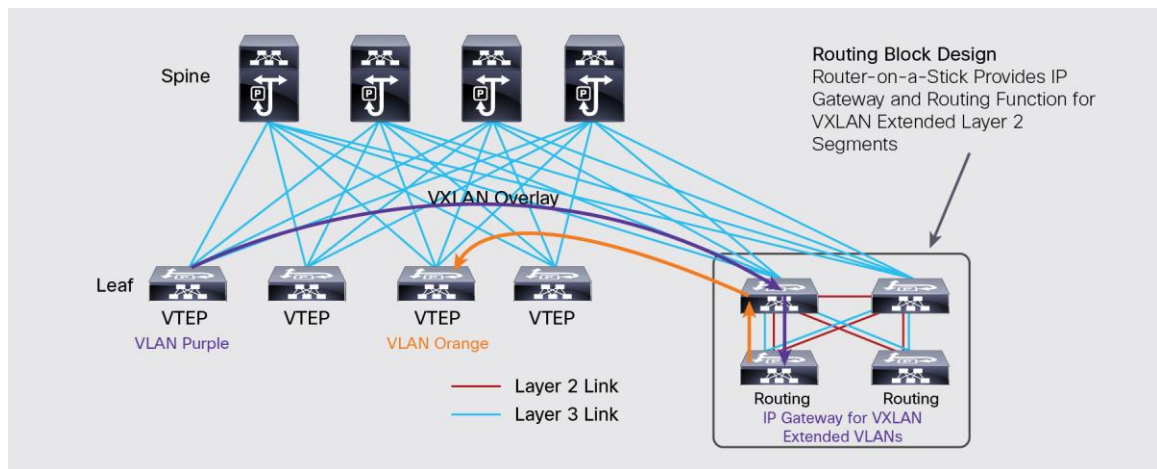
One alternative design for inter-VXLAN routing is shown in Figure 17. It has a VTEP-on-a-stick design, in which one or a pair of Cisco Nexus 9300 VTEPs is connected to the aggregation switches through a Layer 2 link and a Layer 3 link. The Layer 3 links are used to establish VXLAN tunnels with the in-rack VTEP access switches to extend the host VLANs across the Layer 3 network. The aggregation switches are configured with the host VLANs and switch virtual interfaces (SVIs) for their IP subnets. HSRP and Virtual Router Redundancy Protocol (VRRP) can be used to provide the first-hop redundancy with a Layer 2 link in place between the two aggregation switches. The Cisco Nexus 9300 VTEPs map the VXLAN VNIs back to VLANs and send the traffic over the Layer 2 links to the aggregation switches for inter-VLAN routing. After the packets are routed to the destination VLAN IP subnet, the aggregation switches will send the packets back to the Cisco Nexus 9300 VTEPs through the Layer 2 links for VXLAN encapsulation. The encapsulated packets will be forwarded to the destination rack through the underlay Layer 3 network. In this design, the added Cisco Nexus 9300 VTEPs extend the host VLAN segments and bring them onto the aggregation switches. The aggregation switches are the centralized IP gateway for the VXLAN-extended VLANs.

Figure 17. Inter-VLAN Routing Design: VTEP on a Stick



The VTEP-on-a-stick design keeps the IP gateway of the VXLAN-extended VLANs on the aggregation switches, which preserves the IP gateway placement of the traditional Layer 2 data center pod. However, it may create blocks for migrating the network to a spine-and-leaf fabric architecture in the future. The routing block design, by contrast, makes it easier to transform the existing aggregation- and access-layer architecture into a true spine-and-leaf fabric, as shown in Figure 18. This architecture truly enables Layer 2 adjacency across a routed (Layer 3) fabric.

Figure 18. Evolution to Spine-and-Leaf Architecture with VXLAN



What's Next

Currently Cisco Nexus 9300 platform switches support only VXLAN gateway and bridging functions. A planned future release of Cisco NX-OS will bring the VXLAN routing function to the Cisco Nexus 9300 platform, which will greatly simplify the network design for inter-VXLAN routing.

In addition, Cisco is working on a BGP EVPN control plane for VXLAN. The current multicast-based VXLAN lacks a control plane and has to rely on flooding and learning to build the Layer 2 forwarding information base in the overlay network. Multicast in the underlay network is used to support the overlay flood-and-learn behavior. The Cisco BGP EVPN control plane is standards based and does not depend on any fabric controllers. It will offer the following main benefits:

- Eliminate or reduce flooding in the data center
- Achieve optimal handling of multiple-destination traffic (broadcast, unknown unicast, and multicast) on overlay networks
- Provide reliable and quick address resolution and updates for hosts in VXLAN VNIs: essential to support workload mobility in the data center
- Provide a distributed anycast IP gateway for VXLAN overlay networks, enabling optimal VXLAN traffic routing across the Layer 3 network

Conclusion

VXLAN is a network virtualization technology. It uses MAC-in-UDP tunneling to build Layer 2 overlay networks across a Layer 3 infrastructure. This approach decouples the tenant network view from the shared common infrastructure, allowing organizations to build a scalable and reliable Layer 3 data center network while maintaining direct Layer 2 adjacency in the overlay network.

Cisco Nexus 9300 platform switches can be physical VTEPs, providing hardware-based high performance. VXLAN functions on Cisco Nexus 9300 platform switches are quickly evolving, with inter-VXLAN routing and EVPN control plane functions already planned. After these enhancements become available, the VXLAN overlay design with Cisco Nexus 9300 platform switches can be further optimized and simplified. This solution will provide the data center network design for a Layer 2 overlay across a Layer 3 fabric to help provide the application workload mobility and network virtualization required by multitenant environments.

For More Information

- <http://www.ietf.org/id/draft-mahalingam-dutt-dcops-VXLAN-09.txt>
- <https://datatracker.ietf.org/doc/draft-ietf-nvo3-arch/>
- <http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html>

Appendix A: Sample Cisco Nexus 9300 VTEP Switch Configuration

```
n9396-vtep-1# sh run

!Command: show running-config
!Time: Thu Jul 3 17:33:40 2014

version 6.1(2)I2(2a)
hostname n9396-vtep-1
vdc n9396-vtep-1 id 1
  allocate interface Ethernet1/1-48
  allocate interface Ethernet2/1-12
  limit-resource vlan minimum 16 maximum 4094
  limit-resource vrf minimum 2 maximum 4096
  limit-resource port-channel minimum 0 maximum 768
  limit-resource u4route-mem minimum 248 maximum 248
  limit-resource u6route-mem minimum 96 maximum 96
  limit-resource m4route-mem minimum 58 maximum 58
  limit-resource m6route-mem minimum 8 maximum 8

feature nxapi
feature bash-shell
feature scp-server
feature ospf
feature pim
feature interface-vlan
feature vn-segment-vlan-based
feature lacp
feature nv overlay

logging level aaa 6
username admin password 5 $1$e8no0GAX$ptdDp5VsZCXG3unIumghO/ role network-admin
no password strength-check
ip domain-lookup
copp profile strict
snmp-server user admin network-admin auth md5 0x66ec86927ebe7a1eac0d1642ba15c553
priv 0x66ec86927ebe7a1eac0d1642ba15c553 localizedkey
rmon event 1 log trap public description FATAL(1) owner PMON@FATAL
rmon event 2 log trap public description CRITICAL(2) owner PMON@CRITICAL
rmon event 3 log trap public description ERROR(3) owner PMON@ERROR
rmon event 4 log trap public description WARNING(4) owner PMON@WARNING
rmon event 5 log trap public description INFORMATION(5) owner PMON@INFO
snmp-server community public group network-admin

ip pim ssm range 232.0.0.0/8
```

```
ip pim auto-rp listen
vlan 1,13,80,90,100-102,110,200,300,999
vlan 100
    vn-segment 5100
vlan 101
    vn-segment 5101
vlan 200
    vn-segment 5200

vrf context management
    ip route 0.0.0.0/0 173.42.127.1

interface Vlan1

interface nve1
    source-interface loopback0
    member vni 5100 mcast-group 239.1.1.1
    member vni 5101 mcast-group 239.1.1.2
    no shutdown

interface Ethernet1/1
    switchport mode trunk
    switchport access vlan 100
    switchport trunk allowed vlan 1,100-101
    storm-control broadcast level 10.00

interface Ethernet1/2

interface Ethernet1/3
    shutdown

interface Ethernet1/4
    shutdown

interface Ethernet1/5

interface Ethernet1/6
    shutdown

interface Ethernet1/7
    shutdown

interface Ethernet1/8
    shutdown
```

```
interface Ethernet1/9
  shutdown

interface Ethernet1/10

interface Ethernet1/11
  shutdown

interface Ethernet1/12
  shutdown

interface Ethernet1/13
  shutdown

interface Ethernet1/14
  shutdown

interface Ethernet1/15
  shutdown

interface Ethernet1/16
  shutdown

interface Ethernet1/17
  shutdown

interface Ethernet1/18
  shutdown

interface Ethernet1/19
  shutdown

interface Ethernet1/20
  shutdown

interface Ethernet1/21
  shutdown

interface Ethernet1/22
  shutdown

interface Ethernet1/23
  shutdown

interface Ethernet1/24
```

```
shutdown

interface Ethernet1/25
shutdown

interface Ethernet1/26
shutdown

interface Ethernet1/27
shutdown

interface Ethernet1/28
shutdown

interface Ethernet1/29
shutdown

interface Ethernet1/30
shutdown

interface Ethernet1/31
shutdown

interface Ethernet1/32
switchport mode trunk
switchport trunk allowed vlan 1,80,90,100

interface Ethernet1/33

interface Ethernet1/34
shutdown

interface Ethernet1/35
no switchport

interface Ethernet1/36
shutdown

interface Ethernet1/37
shutdown

interface Ethernet1/38
shutdown

interface Ethernet1/39
```

```
shutdown

interface Ethernet1/40
shutdown

interface Ethernet1/41
shutdown

interface Ethernet1/42
shutdown

interface Ethernet1/43
shutdown

interface Ethernet1/44
shutdown

interface Ethernet1/45
shutdown

interface Ethernet1/46
shutdown

interface Ethernet1/47
shutdown

interface Ethernet1/48
shutdown

interface Ethernet2/1
no switchport
ip address 192.168.1.6/30
ip ospf network point-to-point
ip router ospf 1 area 0.0.0.0
ip pim sparse-mode
no shutdown

interface Ethernet2/2
no switchport
ip address 192.168.1.10/30
ip ospf network point-to-point
ip router ospf 1 area 0.0.0.0
ip pim sparse-mode
no shutdown
```

```
interface Ethernet2/3
  shutdown

interface Ethernet2/4
  shutdown

interface Ethernet2/5
  shutdown

interface Ethernet2/6
  shutdown

interface Ethernet2/7
  shutdown

interface Ethernet2/8
  shutdown

interface Ethernet2/9
  shutdown

interface Ethernet2/10
  shutdown

interface Ethernet2/11
  shutdown

interface Ethernet2/12
  shutdown

interface mgmt0
  vrf member management
  ip address 173.42.127.15/24

interface loopback0
  ip address 10.1.1.2/32
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
line console
line vty
boot nxos bootflash:/n9000-dk9.6.1.2.I2.2a.bin
router ospf 1
  router-id 10.1.1.2
no xml server exec-mode
```

```
logging server 173.42.127.175 7 use-vrf management
logging source-interface mgmt0
```

Appendix B: Routing Block Configuration

VXLAN Configuration on vPC VTEPs

VTEP-1

```
feature nv overlay
feature vn-segment-vlan-based

vlan 100
  vn-segment 5100
vlan 101
  vn-segment 5101

interface nve1
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1
  member vni 5101 mcast-group 239.1.1.2

interface loopback0
  no ip redirects
  ip address 10.1.1.4/32
  ip address 10.1.1.100/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

interface Ethernet2/1
  no switchport
  ip address 192.168.1.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet2/2
  no switchport
  ip address 192.168.2.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

```
feature vpc
```

VTEP-2

```
feature nv overlay
feature vn-segment-vlan-based

vlan 100
  vn-segment 5100
vlan 101
  vn-segment 5101

interface nve1
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1
  member vni 5101 mcast-group 239.1.1.2

interface loopback0
  no ip redirects
  ip address 10.1.1.5/32
  ip address 10.1.1.100/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

interface Ethernet2/1
  no switchport
  ip address 192.168.1.14/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet2/2
  no switchport
  ip address 192.168.2.14/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

```
feature vpc
```

```
vpc domain 100
  peer-switch
  peer-keepalive destination
172.21.128.79 source 172.21.128.104
  peer-gateway
```

```
interface port-channel1
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link
```

```
interface port-channel10
  switchport mode trunk
  vpc 10
```

```
interface Ethernet2/3
  switchport mode trunk
  channel-group 10 mode active
```

```
interface Ethernet2/4
  switchport mode trunk
  channel-group 10 mode active
```

```
interface Ethernet2/5
  switchport mode trunk
  channel-group 1 mode active
```

```
interface Ethernet2/6
  switchport mode trunk
  channel-group 1 mode active
```

Router Configuration

Router 1

```
feature vpc

vpc domain 1
  peer-switch
  peer-keepalive destination 5.5.5.5
source 5.5.5.4 vrf default
  peer-gateway
  ip arp synchronize

interface port-channel1
```

```
vpc domain 100
  peer-switch
  peer-keepalive destination
172.21.128.104 source 172.21.128.79
  peer-gateway
```

```
interface port-channel1
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link
```

```
interface port-channel10
  switchport mode trunk
  vpc 10
```

```
interface Ethernet2/3
  switchport mode trunk
  channel-group 10 mode active
```

```
interface Ethernet2/4
  switchport mode trunk
  channel-group 10 mode active
```

```
interface Ethernet2/5
  switchport mode trunk
  channel-group 1 mode active
```

```
interface Ethernet2/6
  switchport mode trunk
  channel-group 1 mode active
```

Router 2

```
feature vpc

vpc domain 1
  peer-switch
  peer-keepalive destination 5.5.5.4
source 5.5.5.5 vrf default
  peer-gateway
  ip arp synchronize

interface port-channel1
```

```

switchport mode trunk
spanning-tree port type network
vpc peer-link

interface port-channel10
switchport mode trunk
vpc 10

interface Ethernet2/3
switchport mode trunk
channel-group 1 mode active

interface Ethernet2/4
switchport mode trunk
channel-group 1 mode active

interface Ethernet2/5
switchport mode trunk
channel-group 10 mode active

interface Ethernet2/6
switchport mode trunk
channel-group 10 mode active

interface Vlan100
no shutdown
ip address 100.0.0.2/24
hsrp 100
preempt
priority 110
ip 100.0.0.1

interface Vlan101
no shutdown
no ip redirects
ip address 101.0.0.2/24
no ipv6 redirects
hsrp 101
preempt
priority 110
ip 101.0.0.1

switchport mode trunk
spanning-tree port type network
vpc peer-link

interface port-channel10
switchport mode trunk
vpc 10

interface Ethernet2/3
switchport mode trunk
channel-group 1 mode active

interface Ethernet2/4
switchport mode trunk
channel-group 1 mode active

interface Ethernet2/5
switchport mode trunk
channel-group 10 mode active

interface Ethernet2/6
switchport mode trunk
channel-group 10 mode active

interface Vlan100
no shutdown
ip address 100.0.0.3/24
hsrp 100
preempt
ip 100.0.0.1

interface Vlan101
no shutdown
no ip redirects
ip address 101.0.0.3/24
no ipv6 redirects
hsrp 101
preempt
ip 101.0.0.1

```




Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

 Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)