Configuring VXLANs

This chapter contains the following sections:

Overview

VXLAN Overview

The Cisco Nexus 3600 platform switches are designed for a hardware-based Virtual Extensible LAN (VXLAN) function. These switches can extend Layer 2 connectivity across the Layer 3 boundary and integrate between VXLAN and non-VXLAN infrastructures. Virtualized and multitenant data center designs can be shared over a common physical infrastructure.

VXLANs enable you to extend Layer 2 networks across the Layer 3 infrastructure by using MAC-in-UDP encapsulation and tunneling. In addition, you can use a VXLAN to build a multitenant data center by decoupling tenant Layer 2 segments from the shared transport network.

When deployed as a VXLAN gateway, the Cisco Nexus 3600 platform switches can connect VXLAN and classic VLAN segments to create a common forwarding domain so that tenant devices can reside in both environments.

A VXLAN has the following benefits:

- Flexible placement of multitenant segments throughout the data center.

  It extends Layer 2 segments over the underlying shared network infrastructure so that tenant workloads can be placed across physical pods in the data center.

- Higher scalability to address more Layer 2 segments.

  A VXLAN uses a 24-bit segment ID called the VXLAN network identifier (VNID). The VNID allows a maximum of 16 million VXLAN segments to coexist in the same administrative domain. (In comparison, traditional VLANs use a 12-bit segment ID that can support a maximum of 4096 VLANs.)

- Utilization of available network paths in the underlying infrastructure.

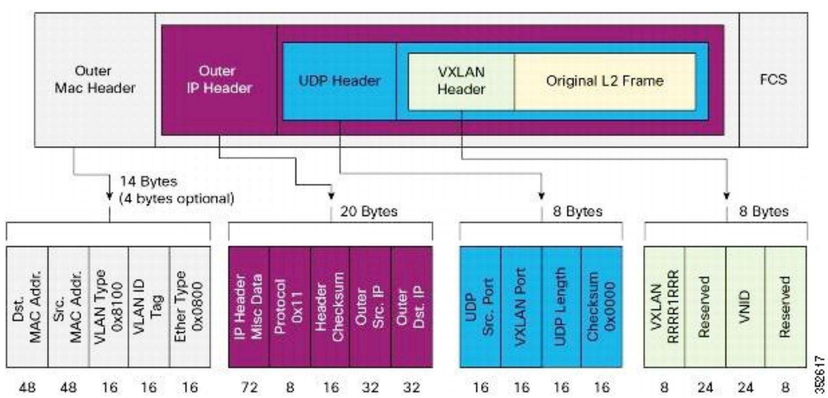
  VXLAN packets are transferred through the underlying network based on its Layer 3 header. It uses equal-cost multipath (ECMP) routing and link aggregation protocols to use all available paths.

VXLAN Encapsulation and Packet Format

A VXLAN is a Layer 2 overlay scheme over a Layer 3 network. It uses MAC-in-UDP encapsulation to extend Layer 2 segments across the data center network. The transport protocol over the physical data center network is IP plus UDP.

A VXLAN defines a MAC-in-UDP encapsulation scheme where the original Layer 2 frame has a VXLAN header added and is then placed in a UDP-IP packet. With this MAC-in-UDP encapsulation, VXLAN tunnels Layer 2 network over the Layer 3 network. The VXLAN packet format is shown in the following figure.

Figure 1: VXLAN Packet Format



A VXLAN uses an 8-byte VXLAN header that consists of a 24-bit VNID and a few reserved bits. The VXLAN header and the original Ethernet frame are in the UDP payload. The 24-bit VNID identifies the Layer 2 segments and maintains Layer 2 isolation between the segments. A VXLAN can support 16 million LAN segments.

VXLAN Tunnel Endpoints

A VXLAN uses VXLAN tunnel endpoint (VTEP) devices to map tenants' end devices to VXLAN segments and to perform VXLAN encapsulation and deencapsulation. Each VTEP device has two types of interfaces:

- Switch port interfaces on the local LAN segment to support local endpoint communication through bridging

- IP interfaces to the transport network where the VXLAN encapsulated frames will be sent

A VTEP device is identified in the IP transport network by using a unique IP address, which is a loopback interface IP address. The VTEP device uses this IP address to encapsulate Ethernet frames and transmits the encapsulated packets to the transport network through the IP interface. A VTEP device learns the remote VTEP IP addresses and the remote MAC address-to-VTEP IP mapping for the VXLAN traffic that it receives.

The VXLAN segments are independent of the underlying network topology; conversely, the underlying IP network between VTEPs is independent of the VXLAN overlay. The IP network routes the encapsulated packets based on the outer IP address header, which has the initiating VTEP as the source IP address and the terminating VTEP or multicast group IP address as the destination IP address.

VXLAN Packet Forwarding Flow

A VXLAN uses stateless tunnels between VTEPs to transmit traffic of the overlay Layer 2 network through the Layer 3 transport network.

ECMP and LACP Load Sharing with VXLANs

Encapsulated VXLAN packets are forwarded between VTEPs based on the native forwarding decisions of the transport network. Most data center transport networks are designed and deployed with multiple redundant paths that take advantage of various multipath load-sharing technologies to distribute traffic loads on all available paths.

A typical VXLAN transport network is an IP-routing network that uses the standard IP equal cost multipath (ECMP) to balance the traffic load among multiple best paths. To avoid out-of-sequence packet forwarding, flow-based ECMP is commonly deployed. An ECMP flow is defined by the source and destination IP addresses and optionally, the source and destination TCP or UDP ports in the IP packet header.

All the VXLAN packet flows between a pair of VTEPs have the same outer source and destination IP addresses, and all VTEP devices must use one identical destination UDP port that can be either the Internet Allocated Numbers Authority (IANA)-allocated UDP port 4789 or a customer-configured port. The only variable element in the ECMP flow definition that can differentiate VXLAN flows from the transport network standpoint is the source UDP port. A similar situation for Link Aggregation Control Protocol (LACP) hashing occurs if the resolved egress interface that is based on the routing and ECMP decision is an LACP port channel. LACP uses the VXLAN outer-packet header for link load-share hashing, which results in the source UDP port being the only element that can uniquely identify a VXLAN flow.

In the Cisco Nexus 3600 platform switches implementation of VXLANs, a hash of the inner frame's header is used as the VXLAN source UDP port. As a result, a VXLAN flow can be unique. The IP address and UDP port combination is in its outer header while the packet traverses the underlay transport network.

Advertising Primary IP Address

On a vPC-enabled leaf or border leaf switch, by default all Layer-3 routes are advertised with the secondary IP address (VIP) of the leaf switch VTEP as the BGP next-hop IP address. Prefix routes and leaf switch generated routes are not synced between vPC leaf switches. Using the VIP as the BGP next-hop for these types of routes can cause traffic to be forwarded to the wrong vPC leaf or border leaf switch and black-holed. The provision to use the primary IP address (PIP) as the next-hop when advertising prefix routes or loopback interface routes in BGP on vPC-enabled leaf or border leaf switches allows users to select the PIP as BGP next-hop when advertising these types of routes so that traffic will always be forwarded to the right vPC-enabled leaf or border leaf switch.

The configuration command for advertising the PIP is **advertise-pip**.

The following is a sample configuration:

```
switch(config)# router bgp 65536
  address-family 12vpn evpn
    advertise-pip
interface nve 1
    advertise virtual-rmac
```

The **advertise-pip** command lets BGP use the PIP as next-hop when advertising prefix routes or leaf-generated routes if vPC is enabled.

VMAC (virtual-mac) is used with VIP and system MAC is used with PIP when the VIP/PIP feature is enabled.

With the **advertise-pip** and **advertise virtual-rmac** commands enabled, type 5 routes are advertised with PIP and type 2 routes are still advertised with VIP. In addition, VMAC will be used with VIP and system MAC will be used with PIP.

Note The **advertise-pip** and **advertise-virtual-rmac** commands must be enabled and disabled together for this feature to work properly. If you enable or disable one and not the other, it is considered an invalid configuration.

Guidelines and Limitations for VXLANs

VXLAN has the following guidelines and limitations:

- IGMP snooping is supported on VXLAN VLANs.
- VXLAN Layer 2 Gateway functionality is supported.
- VXLAN Flood and Learn functionality is not supported.
- Ensure that the network can accommodate an additional 50 bytes for the VXLAN header.
- Only one Network Virtualization Edge (NVE) interface is supported on a switch.
- Layer 3 VXLAN uplinks are not supported in a nondefault virtual and routing forwarding (VRF) instance.
- Switched Port Analyzer (SPAN) for ports carrying VXLAN-encapsulated traffic is not supported.
- VXLAN with Layer 3 VPN is not supported.
- VXLAN with ingress replication is not supported.
- MLD snooping is not supported on VXLAN VLANs.
- ACLs and QoS policies are not supported on VXLAN VLANs.
- DHCP snooping is not supported on VXLAN VLANs.

Considerations for VXLAN Deployment

The following are some of the considerations while deploying VXLANs:

- A loopback interface IP is used to uniquely identify a VTEP device in the transport network.
- To establish IP multicast routing in the core, an IP multicast configuration, PIM configuration, and Rendezvous Point (RP) configuration are required.
- You can configure VTEP-to-VTEP unicast reachability through any IGP protocol.
- VXLAN multicast traffic should always use the RPT shared tree.
- An RP for the multicast group on the VTEP is a supported configuration. However, you must configure the RP for the multicast group at the spine layer/upstream device. Because all multicast traffic traverses the RP, it is more efficient to have this traffic directed to a spine layer/upstream device.

Enabling a VXLAN

Enabling VXLANs involves the following:

- Enabling the VXLAN feature
- Enabling VLAN to VN-Segment mapping

Before you begin

Ensure that you have installed the VXLAN Enterprise license.

Procedure

	Command or Action	Purpose
Step 1	switch# **configure terminal**	Enters global configuration mode.
Step 2	switch(config)# **[no] feature nv overlay**	Enables the VXLAN feature.
Step 3	switch (config)# **[no] feature vn-segment-vlan-based**	Configures the global mode for all VXLAN bridge domains. Enables VLAN to VN-Segment mapping. VLAN to VN-Segment mapping is always one-to-one.
Step 4	(Optional) switch(config)# **copy running-config startup-config**	Saves the change persistently through reboots and restarts by copying the running configuration to the startup configuration.

Example

This example shows how to enable a VXLAN and configure VLAN to VN-Segment mapping:

```
switch# configure terminal
switch(config)# feature nv overlay
switch(config)# feature vn-segment-vlan-based
switch(config)# copy running-config startup-config
```

Mapping a VLAN to a VXLAN VNI

Procedure

	Command or Action	Purpose
Step 1	switch# **configure terminal**	Enters global configuration mode.
Step 2	switch(config)# **vlan** *vlan-id*	Specifies a VLAN.
Step 3	switch(config-vlan)# **vn-segment** *vnid*	Specifies the VXLAN virtual network identifier (VNID). The range of values for vnid is 1 to 16777214.

Example

This example shows how to map a VLAN to a VXLAN VNI:

```
switch# configure terminal
switch(config)# vlan 3100
switch(config-vlan)# vn-segment 5000
```

Configuring a Routing Protocol for NVE Unicast Addresses

Configuring a routing protocol for unicast addresses involves the following:

- Configuring a dedicated loopback interface for NVE reachability.
- Configuring the routing protocol network type.
- Specifying the routing protocol instance and area for an interface.
- Enabling PIM sparse mode in case of multicast replication.

Note Open shortest path first (OSPF) is used as the routing protocol in the examples.

Procedure

	Command or Action	Purpose
Step 1	switch# **configure terminal**	Enters global configuration mode.

	Command or Action	Purpose
Step 2	switch(config)# **interface loopback** *instance*	Creates a dedicated loopback interface for the NVE interface. The instance range is from 0 to 1023.
Step 3	switch(config-if)# **ip address** *ip-address/length*	Configures an IP address for this interface.
Step 4	switch(config-if)# **ip ospf network** {**broadcast** | **point-to-point**}	Configures the OSPF network type to a type other than the default for an interface.
Step 5	switch(config-if)# **ip router ospf** *instance-tag* **area** *area-id*	Specifies the OSPF instance and area for an interface.
Step 6	switch(config-if)# **ip pim sparse-mode**	Enables PIM sparse mode on this interface. The default is disabled. Enable the PIM sparse mode in case of multicast replication.

Example

This example shows how to configure a routing protocol for NVE unicast addresses:

```
switch# configure terminal
switch(config)# interface loopback 10
switch(config-if)# ip address 222.2.2.1/32
switch(config-if)# ip ospf network point-to-point
switch(config-if)# ip router ospf 1 area 0.0.0.0
```

Creating and Configuring an NVE Interface

An NVE interface is the overlay interface that initiates and terminates VXLAN tunnels. You can create and configure an NVE (overlay) interface.

Procedure

	Command or Action	Purpose
Step 1	switch# **configure terminal**	Enters global configuration mode.
Step 2	switch(config)# **interface nve** *instance*	Creates a VXLAN overlay interface that initiates and terminates VXLAN tunnels. **Note** Only one NVE interface is allowed on the switch.
Step 3	switch(config-if-nve)# **source-interface loopback** *instance*	Specifies a source interface. The source interface must be a loopback interface that is configured on the switch with a valid /32 IP address. This /32 IP address must

	Command or Action	Purpose
		be known by the transit routers in the transport network and the remote VTEPs.

Example

This example shows how to create and configure an NVE interface:

```
switch# configure terminal
switch(config)# interface nve 1
switch(config-if-nve)# source-interface loopback 10
```

Configuring a VXLAN VTEP in vPC

Procedure

Step 1 Enter global configuration mode.

```
switch# configure terminal
```

Step 2 Enable the vPC feature on the device.

```
switch(config)# feature vpc
```

Step 3 Enable the interface VLAN feature on the device.

```
switch(config)# feature interface-vlan
```

Step 4 Enable the LACP feature on the device.

```
switch(config)# feature lacp
```

Step 5 Enable the PIM feature on the device.

```
switch(config)# feature pim
```

Step 6 Enables the OSPF feature on the device.

```
switch(config)# feature ospf
```

Step 7 Define a PIM RP address for the underlay multicast group range.

```
switch(config)# ip pim rp-address 192.168.100.1 group-list 224.0.0/4
```

Step 8 Create the VLAN to be used as a backup link.

```
switch(config)# vlan 10
```

Step 9 Create the SVI used for the backup routed path over the vPC peer-link.

```
switch(config)# interface vlan 10
switch(config-if)# ip address 10.10.10.1/30
switch(config-if)# ip router ospf UNDERLAY area 0
switch(config-if)# ip pim sparse-mode
switch(config-if)# no ip redirects
```

```
switch(config-if)# mtu 9216
```

Step 10 Create primary and secondary IP addresses.

```
switch(config)# interface loopback 0
switch(config-if)# description Control_plane_Loopback
switch(config-if)# ip address x.x.x.x/32
switch(config-if)# ip address y.y.y.y/32 secondary
switch(config-if)# ip router ospf process tag area area id
switch(config-if)# ip pim sparse-mode
switch(config-if)# no shutdown
```

Step 11

```
switch(config)# interface loopback 1
switch(config-if)# description Data_Plane_loopback
switch(config-if)# ip address z.z.z.z/32
switch(config-if)# ip router ospf process tag area area id
switch(config-if)# ip pim sparse-mode
switch(config-if)# no shutdown
```

Step 12 Create a vPC domain.

```
switch(config)# vpc domain 10
```

Step 13 Configure the IPv4 address for the remote end of the vPC peer-keepalive link.

```
switch(config-vpc-domain)# peer-keepalive destination 172.28.x.x
```

Note The system does not form the vPC peer link until you configure a vPC peer-keepalive link

The management ports and VRF are the defaults.

Note We recommend that you configure a separate VRF and use a Layer 3 port from each vPC peer device in that VRF for the vPC peer-keepalive link. For more information about creating and configuring VRFs, see the Cisco Nexus 3600 Series NX-OS Unicast Routing Configuration Guide.

Step 14 Enable Peer-Gateway on the vPC domain.

```
switch(config-vpc-domain)# peer-gateway
```

Note Disable IP redirects on all interface-vlans of this vPC domain for correct operation of this feature.

Step 15 Enable Peer-switch on the vPC domain.

```
switch(config-vpc-domain)# peer-switch
```

Note Disable IP redirects on all interface-vlans of this vPC domain for correct operation of this feature.

Step 16 Enable IP ARP synchronize under the vPC domain to facilitate faster ARP table population following device reload.

```
switch(config-vpc-domain)# ip arp synchronize
```

Step 17 (Optional) Enable IPv6 nd synchronization under the vPC domain to facilitate faster nd table population following device reload.

```
switch(config-vpc-domain)# ipv6 nd synchronize
```

Step 18 Create the vPC peer-link port-channel interface and add two member interfaces.

```
switch(config)# interface port-channel 1
switch(config-if)# switchport
switch(config-if)# switchport mode trunk
switch(config-if)# switchport trunk allowed vlan 1,100-200
switch(config-if)# mtu 9216
switch(config-if)# vpc peer-link
switch(config-if)# no shutdown
switch(config-if)# interface Ethernet 1/1, 1/20
switch(config-if)# switchport
switch(config-if)# mtu 9216
switch(config-if)# channel-group 1 mode active
switch(config-if)# no shutdown
```

Step 19 Modify the STP hello-time, forward-time, and max-age time.

As a best practice, we recommend changing the **hello-time** to four seconds to avoid unnecessary TCN generation when the vPC role change occurs. As a result of changing the **hello-time**, it is also recommended to change the **max-age** and **forward-time** accordingly.

```
switch(config)# spanning-tree vlan 1-3967 hello-time 4
switch(config)# spanning-tree vlan 1-3967 forward-time 30
switch(config)# spanning-tree vlan 1-3967 max-age 40
```

Step 20 (Optional) Enable the delay restore timer for SVI's.

We recommend that you tune this value when the SVI or VNI scale is high. For example, when the SVI count is 1000, we recommend setting the delay restore for interface-vlan to 45 seconds.

```
switch(config-vpc-domain)# delay restore interface-vlan 45
```

Configuring Replication for a VNI

Replication for VXLAN network identifier (VNI) can be configured in one of two ways:

- Multicast replication

Configuring Multicast Replication

Before you begin

- Ensure that the NVE interface is created and configured.
- Ensure that the source interface is specified.

Procedure

	Command or Action	Purpose
Step 1	switch(config-if-nve)# **member vni** {*vnid* **mcast-group** *multicast-group-addr* | *vnid-range* **mcast-group** *start-addr* [*end-addr*]}	Maps VXLAN VNIs to the NVE interface and assigns a multicast group to the VNIs.

Example

This example shows how to map a VNI to an NVE interface and assign it to a multicast group:

```
switch(config-if-nve)# member vni 5000 mcast-group 225.1.1.1
```

Configuring IGMP Snooping Over VXLAN

Overview of IGMP Snooping Over VXLAN

Starting with Cisco NX-OS Release 7.0(3)F3(4), you can configure IGMP snooping over VXLAN. The configuration of IGMP snooping is same in VXLAN as in configuration of IGMP snooping in regular VLAN domain. For more information on IGMP snooping, see the *Configuring IGMP Snooping* chapter in the Cisco Nexus 3600 NX-OS Multicast Routing Configuration Guide, Release 7.x.

Guidelines and Limitations for IGMP Snooping Over VXLAN

See the following guidelines and limitations for IGMP snooping over VXLAN:

- For IGMP snooping over VXLAN, all the guidelines and limitations of VXLAN apply.
- IGMP snooping over VXLAN is not supported on any FEX enabled platforms and FEX ports.

Configuring IGMP Snooping Over VXLAN

Procedure

	Command or Action	Purpose
Step 1	switch(config)#**ip igmp snooping vxlan**	Enables IGMP snooping for VXLAN VLANs. You have to explicitly configure this command to enable snooping for VXLAN VLANs.
Step 2	switch(config)#**ip igmp snooping disable-nve-static-router-port**	Configures IGMP snooping over VXLAN to not include NVE as static mrouter port using this global CLI command. IGMP snooping over VXLAN has the NVE interface as mrouter port by default.

Verifying the VXLAN Configuration

Use one of the following commands to verify the VXLAN configuration, to display the MAC addresses, and to clear the MAC addresses:

Command	Purpose
show nve interface nve *id*	Displays the configuration of an NVE interface.
show nve vni	Displays the VNI that is mapped to an NVE interface.
show nve peers	Displays peers of the NVE interface.
show nve vxlan-params	Displays the VXLAN UDP port configured.
show mac address-table	Displays both VLAN and VXLAN MAC addresses.
clear mac address-table dynamic	Clears all MAC address entries in the MAC address table.

Example

This example shows how to display the configuration of an NVE interface:

```
switch# show nve interface nve 1
Interface: nve1, State: up, encapsulation: VXLAN
Source-interface: loopback10 (primary: 111.1.1.1, secondary: 0.0.0.0)
```

This example shows how to display the VNI that is mapped to an NVE interface for multicast replication:

```
switch# show nve vni
Interface          VNI        Multicast-group   VNI State
----------------   --------   ---------------   ---------
nve1               5000       225.1.1.1          Up
```

This example shows how to display the VNI that is mapped to an NVE interface for ingress replication:

```
switch# show nve vni
Interface          VNI        Multicast-group   VNI State
----------------   --------   ---------------   ---------
nve1               5000       0.0.0.0           Up
```

This example shows how to display the peers of an NVE interface:

```
switch# show nve peers
Interface          Peer-IP           Peer-State
----------------   ---------------   -------------
nve1               111.1.1.1          Up
```

This example shows how to display the VXLAN UDP port configured:

```
switch# show nve vxlan-params
VxLAN Dest. UDP Port: 4789
```

This example shows how to display both VLAN and VXLAN MAC addresses:

```
Added draft comment: hidden contentswitch# show mac address-table
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
```

```
        age - seconds since first seen,+ - primary entry using vPC Peer-Link
   VLAN     MAC Address      Type      age     Secure NTFY   Ports/SWID.SSID.LID
---------+-----------------+--------+---------+------+----+------------------
* 109      0000.0410.0902    dynamic   470        F    F  Po2233
* 109      0000.0410.0912    dynamic   470        F    F  Po2233
* 109      0000.0410.0912    dynamic   470        F    F  nve1(1.1.1.200)
* 108      0000.0410.0802    dynamic   470        F    F  Po2233
* 108      0000.0410.0812    dynamic   470        F    F  Po2233
* 107      0000.0410.0702    dynamic   470        F    F  Po2233
* 107      0000.0410.0712    dynamic   470        F    F  Po2233
* 107      0000.0410.0712    dynamic   470        F    F  nve1(1.1.1.200)
* 106      0000.0410.0602    dynamic   470        F    F  Po2233
* 106      0000.0410.0612    dynamic   470        F    F  Po2233
* 105      0000.0410.0502    dynamic   470        F    F  Po2233
* 105      0000.0410.0512    dynamic   470        F    F  Po2233
* 105      0000.0410.0512    dynamic   470        F    F  nve1(1.1.1.200)
* 104      0000.0410.0402    dynamic   470        F    F  Po2233
* 104      0000.0410.0412    dynamic   470        F    F  Po2233
```

This example shows how to clear all MAC address entries in the MAC address table:

```
switch# clear mac address-table dynamic
switch#
```