



Configuring VXLAN

This chapter contains the following sections:

- [Guidelines and Limitations for VXLAN, on page 1](#)
- [Considerations for VXLAN Deployment, on page 7](#)
- [vPC Considerations for VXLAN Deployment, on page 10](#)
- [Network Considerations for VXLAN Deployments, on page 14](#)
- [Considerations for the Transport Network, on page 15](#)
- [Considerations for Tunneling VXLAN, on page 16](#)
- [Configuring VXLAN, on page 17](#)

Guidelines and Limitations for VXLAN

VXLAN has the following guidelines and limitations:

Table 1: ACL Options That can be Used for VXLAN traffic, on Platforms that Include, Cisco Nexus 92300YC, 92160YC-X, 93120TX, 9332PO, and 9348GC-FXP Switches

ACL Direction	ACL Type	VTEP Type	Port Type	Flow Direction	Traffic Type	Supported
Ingress	PACL	Ingress VTEP	L2 port	Access to Network [GROUP:encap direction]	Native L2 traffic [GROUP:inner]	YES
	VACL	Ingress VTEP	VLAN	Access to Network [GROUP:encap direction]	Native L2 traffic [GROUP:inner]	YES
Ingress	RACL	Ingress VTEP	tenant L3 SVI	Access to Network [GROUP:encap direction]	Native L3 traffic [GROUP:inner]	YES

ACL Direction	ACL Type	VTEP Type	Port Type	Flow Direction	Traffic Type	Supported
Egress	RACL	Ingress VTEP	uplink L3/L3-PO/SVI	Access to Network [GROUP:encap direction]	VXLAN encap [GROUP:outer]	NO
Ingress	RACL	Egress VTEP	uplink L3/L3-PO/SVI	Network to Access [GROUP:decap direction]	VXLAN encap [GROUP:outer]	NO
Egress	PACL	Egress VTEP	L2 port	Network to Access [GROUP:decap direction]	Native L2 traffic [GROUP:inner]	NO
	VACL	Egress VTEP	VLAN	Network to Access [GROUP:decap direction]	Native L2 traffic [GROUP:inner]	NO
Egress	RACL	Egress VTEP	tenant L3 SVI	Network to Access [GROUP:decap direction]	Post-decap L3 traffic [GROUP:inner]	YES

Table 2: ACL options that can be used for VXLAN traffic, on platforms that include, Cisco Nexus 92160YC-X, 93108TC-EX, 93180LC-EX, and 93180YC-EX switches, Release 7.0(3)I6(1)

ACL Direction	ACL Type	VTEP Type	Port Type	Flow Direction	Traffic Type	Supported
Ingress	PACL	Ingress VTEP	L2 port	Access to Network [GROUP:encap direction]	Native L2 traffic [GROUP:inner]	YES (works only for base port PO)
Egress	PACL	Egress VTEP	L2 port	Network to Access [GROUP:decap direction]	Native L2 traffic [GROUP:inner]	NO
Ingress	VACL	Ingress VTEP	VLAN	Access to Network [GROUP:encap direction]	Native L2 traffic [GROUP:inner]	YES
Egress	VACL	Egress VTEP	VLAN	Network to Access [GROUP:decap direction]	Native L2 traffic [GROUP:inner]	YES

ACL Direction	ACL Type	VTEP Type	Port Type	Flow Direction	Traffic Type	Supported
Ingress	RACL	Ingress VTEP	tenant L3 SVI	Access to Network [GROUP:encap direction]	Native L3 traffic [GROUP:inner]	YES
Egress	RACL	Egress VTEP	tenant L3 SVI	Network to Access [GROUP:decap direction]	Post-decap L3 traffic [GROUP:inner]	YES
Ingress	RACL	Egress VTEP	uplink L3/L3-PO/SVI	Network to Access [GROUP:decap direction]	VXLAN encap [GROUP:outer]	NO
Egress	RACL	Ingress VTEP	uplink L3/L3-PO/SVI	Access to Network [GROUP:encap direction]	VXLAN encap [GROUP:outer]	NO

- Non-blocking Multicast (NBM) running on a VXLAN enabled switch is not supported. Feature nbm may disrupt VXLAN underlay multicast forwarding.
- For scale environments, the VLAN IDs related to the VRF and Layer-3 VNI (L3VNI) must be reserved with the **system vlan nve-overlay id** command.
- NLB in the unicast, multicast, and IGMP multicast modes is not supported on Cisco Nexus 9000 Series based VXLAN VTEPs. The work around is to move the NLB cluster behind intermediary device (which supports NLB in the respective mode) and inject the cluster IP address as external prefix into VXLAN fabric.
- Beginning with Cisco NX-OS Release 9.2(3), support added for MultiAuth Change of Authorization (CoA). For more information, see the [Cisco Nexus 9000 Series NX-OS Security Configuration Guide, Release 9.2\(x\)](#).
- The **lACP vpc-convergence** command can be configured in VXLAN and non-VXLAN environments that have vPC port channels to hosts that support LACP.
- Beginning with Cisco NX-OS Release 9.2(1), PIM BiDir for VXLAN underlay with and without vPC is supported.

The following is a list of what is not supported when the PIM BiDir for VXLAN underlay feature is configured:

- Flood and learn VXLAN
- Tenant Routed Multicast (TRM)
- VXLAN EVPN Multi-Site
- VXLAN EVPN Multihoming
- vPC attached VTEPs

For redundant RPs, use Phantom RP.

For transitioning from PIM ASM to PIM BiDir or from PIM BiDir to PIM ASM underlay, we recommend that you use the following example procedure:

```
no ip pim rp-address 192.0.2.100 group-list 230.1.1.0/8
clear ip mroute *
clear ip mroute date-created *
clear ip pim route *
clear ip igmp groups *
clear ip igmp snooping groups * vlan all
```

Wait for all tables to clean up.

```
ip pim rp-address 192.0.2.100 group-list 230.1.1.0/8 bidir
```

- When entering the **no feature pim** command, NVE ownership on the route is not removed so the route stays and traffic continues to flow. Aging is done by PIM. PIM does not age out entries having a VXLAN encap flag.
 - Fibre Channel over Ethernet (FCoE) N-port Virtualization (NPV) can co-exist with VXLAN on different fabric uplinks but on same or different front panel ports on the Cisco Nexus 93180YC-EX and 93180YC-FX switches.
Fibre Channel N-port Virtualization (NPV) can co-exist with VXLAN on different fabric uplinks but on same or different front panel ports on the Cisco Nexus 93180YC-FX switches. VXLAN can only exist on the Ethernet front panel ports, but not on the FC front panel ports.
 - VXLAN is supported on the Cisco Nexus 9348GC-FXP switch.
 - When SVI is enabled on a VTEP (flood and learn, or EVPN) regardless of ARP suppression, make sure that ARP-ETHER TCAM is carved using the **hardware access-list tcam region arp-ether 256 double-wide** command. This is not applicable to the Cisco Nexus 9200 and 9300-EX platform switches and Cisco Nexus 9500 platform switches with 9700-EX line cards.
 - For information regarding the **load-share** keyword usage for the PBR with VXLAN feature, see the [Guidelines and Limitations](#) section of the Configuring Policy -Based Routing chapter of the [Cisco Nexus 9000 Series NX-OS Unicast Routing Configuration Guide, Release 9.2\(x\)](#) .
 - For the Cisco Nexus 9504 and 9508 switches with -R line cards, VXLAN Layer 2 Gateway is supported on the 9636C-RX line card. VXLAN and MPLS cannot be enabled on the Cisco Nexus 9508 switch at the same time.
 - For the Cisco Nexus 9504 and 9508 switches with -R line cards, if VXLAN is enabled, the Layer 2 Gateway cannot be enabled when there is any line card other than the 9636C-RX.
 - For the Cisco Nexus 9504 and 9508 switches with -R line cards, PIM/ASM is supported in the underlay ports. PIM/Bidir is not supported. For more information, see the [Cisco Nexus 9000 Series NX-OS Multicast Routing Configuration Guide, Release 9.2\(x\)](#) .
 - For the Cisco Nexus 9504 and 9508 switches with -R line cards, IPv6 hosts routing in the overlay is supported.
 - For the Cisco Nexus 9504 and 9508 switches with -R line cards, ARP suppression is supported.
- load-share**
- The keyword has been added to the Configuring a Route Policy procedure for the PBR over VXLAN feature.

For more information, see the [Cisco Nexus 9000 Series NX-OS Unicast Routing Configuration Guide, Release 9.2\(x\)](#).

- A new CLI command **lACP vpc-convergence** is added for better convergence of Layer 2 EVPN VXLAN:

```
interface port-channel10
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1001-1200
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  lACP vpc-convergence
  vpc 10
```

```
interface Ethernet1/34 <- The port-channel member-port is configured with LACP-active
mode (for example, no changes are done at the member-port level.)
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1001-1200
  channel-group 10 mode active
  no shutdown
```

- Port-VLAN with VXLAN is supported on Cisco Nexus 9300-EX and 9500 Series switches with 9700-EX line cards with the following exceptions:
 - Only Layer 2 (no routing) is supported with port-VLAN with VXLAN on these switches.
 - No inner VLAN mapping is supported.
- The **system nve ipmc** CLI command is not applicable to the Cisco 9200 and 9300-EX platform switches and Cisco 9500 platform switches with 9700-EX line cards.
- Bind NVE to a loopback address that is separate from other loopback addresses that are required by Layer 3 protocols. A best practice is to use a dedicated loopback address for VXLAN. This best practice should be applied not only for the vPC VXLAN deployment, but for all VXLAN deployments.
- To remove configurations from an NVE interface, we recommend manually removing each configuration rather than using the **default interface nve** command.
- When SVI is enabled on a VTEP (flood and learn or EVPN), make sure that ARP-ETHER TCAM is carved using the **hardware access-list tcam region arp-ether 256** CLI command. This is not applicable to Cisco 9200 and 9300-EX Series switches and Cisco 9500 Series switches with 9700-EX line cards.
- **show** commands with the **internal** keyword are not supported.
- FEX ports do not support IGMP snooping on VXLAN VLANs.
- VXLAN is supported for the Cisco Nexus 93108TC-EX and 93180YC-EX switches and for Cisco Nexus 9500 Series switches with the X9732C-EX line card.
- DHCP snooping (Dynamic Host Configuration Protocol snooping) is not supported on VXLAN VLANs.
- RACLs are not supported on Layer 3 uplinks for VXLAN traffic. Egress VACLs support is not available for de-capsulated packets in the network to access direction on the inner payload.
As a best practice, use PACLS/VACLs for the access to the network direction.
- The QoS buffer-boost feature is not applicable for VXLAN traffic.
- SVI and subinterfaces as uplinks are not supported.

- VTEPs do not support VXLAN encapsulated traffic over Parent-Interfaces if subinterfaces are configured. This is regardless of VRF participation.
- VTEPs do not support VXLAN encapsulated traffic over subinterfaces. This is regardless of VRF participation or IEEE 802.1Q encapsulation.
- Mixing Sub-Interfaces for VXLAN and non-VXLAN enabled VLANs is not supported.
- Point to multipoint Layer 3 and SVI uplinks are not supported.
- A FEX HIF (FEX host interface port) is supported for a VLAN that is extended with VXLAN.
- In an ingress replication vPC setup, Layer 3 connectivity is needed between vPC peer devices. This aids the traffic when the Layer 3 uplink (underlay) connectivity is lost for one of the vPC peers.
- Rollback is not supported on VXLAN VLANs that are configured with the port VLAN mapping feature.
- The VXLAN UDP port number is used for VXLAN encapsulation. For Cisco Nexus NX-OS, the UDP port number is 4789. It complies with IETF standards and is not configurable.
- VXLAN is supported on Cisco Nexus 9500 platform switches with the following line cards:
 - 9500-R
 - 9564PX
 - 9564TX
 - 9536PQ
 - 9700-EX
 - 9700-FX
- Cisco Nexus 9300 Series switches with 100G uplinks only support VXLAN switching/bridging. Cisco Nexus 9200, Cisco Nexus 9300-EX, and Cisco Nexus 9300-FX platform switches do not have this restriction.



Note For VXLAN routing support, a 40G uplink module is required.

- MDP is not supported for VXLAN configurations.
- Consistency checkers are not supported for VXLAN tables.
- ARP suppression is supported for a VNI only if the VTEP hosts the First-Hop Gateway (Distributed Anycast Gateway) for this VNI. The VTEP and SVI for this VLAN must be properly configured for the Distributed Anycast Gateway operation (for example, global anycast gateway MAC address configured and anycast gateway with the virtual IP address on the SVI).
- ARP suppression is a per-L2VNI fabric-wide setting in the VXLAN fabric. Enable or disable this feature consistently across all VTEPs in the fabric. Inconsistent ARP suppression configuration across VTEPs is not supported.
- The VXLAN network identifier (VNID) 16777215 is reserved and should not be configured explicitly.
- VXLAN supports In Service Software Upgrade (ISSU).

- VXLAN does not support coexistence with the GRE tunnel feature or the MPLS (static or segment-routing) feature.
- VTEP connected to FEX host interface ports is not supported.
- If multiple VTEPs use the same multicast group address for underlay multicast but have different VNIs, the VTEPs should have at least one VNI in common. Doing so ensures that NVE peer discovery occurs and underlay multicast traffic is forwarded correctly. For example, leafs L1 and L4 could have VNI 10 and leafs L2 and L3 could have VNI 20, and both VNIs could share the same group address. When leaf L1 sends traffic to leaf L4, the traffic could pass through leaf L2 or L3. Because NVE peer L1 is not learned on leaf L2 or L3, the traffic is dropped. Therefore, VTEPs that share a group address need to have at least one VNI in common so that peer learning occurs and traffic is not dropped. This requirement applies to VXLAN bud-node topologies.
- VXLAN does not support co-existence with MVR and MPLS for Cisco Nexus 9504 and 9508 with -R line cards.
- Resilient hashing (port-channel load-balancing resiliency) and VXLAN configurations are not compatible with VTEPs using ALE uplink ports.



Note Resilient hashing is disabled by default.

- Native VLANs for VXLAN are not supported. All traffic on VXLAN Layer 2 trunks needs to be tagged. This limitation applies to Cisco Nexus 9300 and 9500 platform switches with 95xx line cards. This limitation does not apply to Cisco Nexus 9200, 9300-EX, 9300-FX, and 9500 platform switches with -EX or -FX line cards.
- NVE source interface loopback for VTEP should only be IPv4 address. Use of IPv6 address for NVE source interface is not supported.
- Next hop address in overlay (in `bgp l2vpn evpn address family updates`) should be resolved in underlay URIB to the same address family. For example, the use of VTEP (NVE source loopback) IPv4 addresses in fabric should only have BGP l2vpn evpn peering over IPv4 addresses.

Considerations for VXLAN Deployment

- For scale environments, the VLAN IDs related to the VRF and Layer-3 VNI (L3VNI) must be reserved with the `system vlan nve-overlay id` command.

This is required to optimize the VXLAN resource allocation to scale the following platforms:

- Cisco Nexus 9300 platform switches
- Cisco Nexus 9500 platform switches with 9500 line cards

The following example shows how to reserve the VLAN IDs related to the VRF and the Layer-3 VNI:

```
system vlan nve-overlay id 2000

    vlan 2000
      vn-segment 50000

    interface Vlan2000
```

```

vrf member MYVRF_50000
ip forward
ipv6 forward

vrf context MYVRF_50000
vni 50000

```



Note The **system vlan nve-overlay id** command should be used for a VRF or a Layer-3 VNI (L3VNI) only. Do not use this command for regular VLANs or Layer-2 VNIs (L2VNI).

- When configuring VXLAN BGP EVPN, the "System Routing Mode: Default" is applicable for the following hardware platforms:
 - Cisco Nexus 9200 platform switches
 - Cisco Nexus 9300 platform switches
 - Cisco Nexus 9300-EX platform switches
 - Cisco Nexus 9300-FX/FX2 platform switches
 - Cisco Nexus 9500 platform switches with X9500 line cards
 - Cisco Nexus 9500 platform switches with X9700-EX/FX line cards
- The "System Routing Mode: template-vxlan-scale" is not applicable.
- When using VXLAN BGP EVPN in combination with Cisco NX-OS Release 7.0(3)I4(x) or NX-OS Release 7.0(3)I5(1), the "System Routing Mode: template-vxlan-scale" is required on the following hardware platforms:
 - Cisco Nexus 9300-EX Switches
 - Cisco Nexus 9500 Switches with X9700-EX line cards
- Changing the "System Routing Mode" requires a reload of the switch.
- A loopback address is required when using the **source-interface config** command. The loopback address represents the local VTEP IP.
- During boot-up of a switch, you can use the **source-interface hold-down-time** *hold-down-time* command to suppress advertisement of the NVE loopback address until the overlay has converged. The range for the *hold-down-time* is 0 - 2147483647 seconds. The default is 300 seconds.



Note Though the loopback is still down, the traffic is encapsulated and sent to fabric.

- To establish IP multicast routing in the core, IP multicast configuration, PIM configuration, and RP configuration is required.
- VTEP to VTEP unicast reachability can be configured through any IGP protocol.

- In VXLAN flood and learn mode, the default gateway for VXLAN VLAN is recommended to be a centralized gateway on a pair of vPC devices with FHRP (First Hop Redundancy Protocol) running between them.

In BGP EVPN, it is recommended to use the anycast gateway feature on all VTEPs.

- For flood and learn mode, only a centralized Layer 3 gateway is supported. Anycast gateway is not supported. The recommended Layer 3 gateway design would be a pair of switches in vPC to be the Layer 3 centralized gateway with FHRP protocol running on the SVIs. The same SVI's cannot span across multiple VTEPs even with different IP addresses used in the same subnet.



Note When configuring SVI with flood and learn mode on the central gateway leaf, it is mandatory to configure **hardware access-list tcam region arp-ether size double-wide**. (You must decrease the size of an existing TCAM region before using this command.)

For example:

```
hardware access-list tcam region arp-ether 256 double-wide
```



Note Configuring the **hardware access-list tcam region arp-ether size double-wide** is not required on Cisco Nexus 9200 Series switches.

- When configuring ARP suppression with BGP-EVPN, use the **hardware access-list tcam region arp-ether size double-wide** command to accommodate ARP in this region. (You must decrease the size of an existing TCAM region before using this command.)



Note This step is required for Cisco Nexus 9300 switches (NFE/ALE) and Cisco Nexus 9500 switches with N9K-X9564PX, N9K-X9564TX, and N9K-X9536PQ line cards. This step is not needed with Cisco Nexus 9200 switches, Cisco Nexus 9300-EX switches, or Cisco Nexus 9500 switches with N9K-X9732C-EX line cards.

- VXLAN tunnels cannot have more than one underlay next hop on a given underlay port. For example, on a given output underlay port, only one destination MAC address can be derived as the outer MAC on a given output port.

This is a per-port limitation, not a per-tunnel limitation. This means that two tunnels that are reachable through the same underlay port cannot drive two different outer MAC addresses.

- When changing the IP address of a VTEP device, you must shut the NVE interface before changing the IP address.
- As a best practice, when migrating any sets of VTEP to a multisite BGW, NVE interface must be shut on all the VTEPs where this migration is being performed. NVE interface should be brought back up once the migration is complete and all necessary configurations for multisite are applied to the VTEPs.
- As a best practice, the RP for the multicast group should be configured only on the spine layer. Use the anycast RP for RP load balancing and redundancy.

The following is an example of an anycast RP configuration on spines:

```
ip pim rp-address 1.1.1.10 group-list 224.0.0.0/4
ip pim anycast-rp 1.1.1.10 1.1.1.1
ip pim anycast-rp 1.1.1.10 1.1.1.2
```



Note

- 1.1.1.10 is the anycast RP IP address that is configured on all RPs participating in the anycast RP set.
 - 1.1.1.1 is the local RP IP.
 - 1.1.1.2 is the peer RP IP.
-

- Static ingress replication and BGP EVPN ingress replication do not require any IP Multicast routing in the underlay.

vPC Considerations for VXLAN Deployment

- As a best practice when feature vPC is added or removed from a VTEP, the NVE interfaces on both the vPC primary and the vPC secondary should be shut before the change is made.
- Bind NVE to a loopback address that is separate from other loopback addresses that are required by Layer 3 protocols. A best practice is to use a dedicated loopback address for VXLAN.
- On vPC VXLAN, it is recommended to increase the **delay restore interface-vlan** timer under the vPC configuration, if the number of SVIs are scaled up. For example, if there are 1000 VNIs with 1000 SVIs, we recommend to increase the **delay restore interface-vlan** timer to 45 seconds.
- If a ping is initiated to the attached hosts on VXLAN VLAN from a vPC VTEP node, the source IP address used by default is the anycast IP that is configured on the SVI. This ping can fail to get a response from the host in case the response is hashed to the vPC peer node. This issue can happen when a ping is initiated from a VXLAN vPC node to the attached hosts without using a unique source IP address. As a workaround for this situation, use VXLAN OAM or create a unique loopback on each vPC VTEP and route the unique address via a backdoor path.
- The loopback address used by NVE needs to be configured to have a primary IP address and a secondary IP address.

The secondary IP address is used for all VXLAN traffic that includes multicast and unicast encapsulated traffic.

- vPC peers must have identical configurations.
 - Consistent VLAN to vn-segment mapping.
 - Consistent NVE1 binding to the same loopback interface
 - Using the same secondary IP address.
 - Using different primary IP addresses.

- Consistent VNI to group mapping.
- For multicast, the vPC node that receives the (S, G) join from the RP (rendezvous point) becomes the DF (designated forwarder). On the DF node, encap routes are installed for multicast.
Decap routes are installed based on the election of a decapper from between the vPC primary node and the vPC secondary node. The winner of the decap election is the node with the least cost to the RP. However, if the cost to the RP is the same for both nodes, the vPC primary node is elected.
The winner of the decap election has the decap mroute installed. The other node does not have a decap route installed.
- On a vPC device, BUM traffic (broadcast, unknown-unicast, and multicast traffic) from hosts is replicated on the peer-link. A copy is made of every native packet and each native packet is sent across the peer-link to service orphan-ports connected to the peer vPC switch.
To prevent traffic loops in VXLAN networks, native packets ingressing the peer-link cannot be sent to an uplink. However, if the peer switch is the encapper, the copied packet traverses the peer-link and is sent to the uplink.



Note Each copied packet is sent on a special internal VLAN (VLAN 4041 or VLAN 4046).

- When the peer-link is shut, the loopback interface used by NVE on the vPC secondary is brought down and the status is **Admin Shut**. This is done so that the route to the loopback is withdrawn on the upstream and that the upstream can divert all traffic to the vPC primary.



Note Orphans connected to the vPC secondary will experience loss of traffic for the period that the peer-link is shut. This is similar to Layer 2 orphans in a vPC secondary of a traditional vPC setup.

- When the vPC domain is shut, the loopback interface used by NVE on the VTEP with shutdown vPC domain is brought down and the status is Admin Shut. This is done so that the route to the loopback is withdrawn on the upstream and that the upstream can divert all traffic to the other vPC VTEP.
- When peer-link is no-shut, the NVE loopback address is brought up again and the route is advertised upstream, attracting traffic.
- For vPC, the loopback interface has two IP addresses: the primary IP address and the secondary IP address.

The primary IP address is unique and is used by Layer 3 protocols.

The secondary IP address on loopback is necessary because the interface NVE uses it for the VTEP IP address. The secondary IP address must be same on both vPC peers.

- The vPC peer-gateway feature must be enabled on both peers to facilitate NVE RMAC/VMAC programming on both peers. For peer-gateway functionality, at least one backup routing SVI is required to be enabled across peer-link and also configured with PIM. This provides a backup routing path in the case when VTEP loses complete connectivity to the spine. Remote peer reachability is re-routed over

peer-link in his case. In BUD node topologies, the backup SVI needs to be added as a static OIF for each underlay multicast group.

```
switch# sh ru int vlan 2

interface Vlan2
  description backup1_svi_over_peer-link
  no shutdown
  ip address 30.2.1.1/30
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  ip igmp static-oif route-map match-mcast-groups

route-map match-mcast-groups permit 1
  match ip multicast group 225.1.1.1/32
```



Note In BUD node topologies, the backup SVI needs to be added as a static OIF for each underlay multicast group.

The SVI must be configured on both vPC peers and requires PIM to be enabled.

- When the NVE or loopback is shut in vPC configurations:
 - If the NVE or loopback is shut only on the primary vPC switch, the global VXLAN vPC consistency checker fails. Then the NVE, loopback, and vPCs are taken down on the secondary vPC switch.
 - If the NVE or loopback is shut only on the secondary vPC switch, the global VXLAN vPC consistency checker fails. Then, the NVE, loopback, and secondary vPC are brought down on the secondary. Traffic continues to flow through the primary vPC switch.
 - As a best practice, you should keep both the NVE and loopback up on both the primary and secondary vPC switches.
- Redundant anycast RPs configured in the network for multicast load-balancing and RP redundancy are supported on vPC VTEP topologies.
- As a best practice, when changing the secondary IP address of an anycast vPC VTEP, the NVE interfaces on both the vPC primary and the vPC secondary must be shut before the IP changes are made.
- When SVI is enabled on a VTEP (flood and learn, or EVPN) regardless of ARP suppression, make sure that ARP-ETHER TCAM is carved using the **hardware access-list tcam region arp-ether 256 double-wide** command. This requirement does not apply to Cisco Nexus 9200, 9300-EX, and 9300-FX/FX2 platform switches and Cisco Nexus 9500 platform switches with 9700-EX line cards.
- The **show** commands with the **internal** keyword are not supported.
- DHCP snooping (Dynamic Host Configuration Protocol snooping) is not supported on VXLAN VLANs.
- RACLs are not supported on Layer 3 uplinks for VXLAN traffic. Egress VACLs support is not available for de-capsulated packets in the network to access direction on the inner payload.

As a best practice, use PACLS/VACLs for the access to the network direction.

See the [Cisco Nexus 9000 Series NX-OS Security Configuration Guide, Release 9.3\(x\)](#) for other guidelines and limitations for the VXLAN ACL feature.

- QoS classification is not supported for VXLAN traffic in the network to access direction on the Layer 3 uplink interface.
See the [Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.3\(x\)](#) for other guidelines and limitations for the VXLAN QoS feature.
- The QoS buffer-boost feature is not applicable for VXLAN traffic.
- VTEPs do not support VXLAN-encapsulated traffic over parent interfaces if subinterfaces are configured, regardless of VRF participation.
- VTEPs do not support VXLAN encapsulated traffic over subinterfaces. This is regardless of VRF participation or IEEE802.1Q encapsulation.
- Mixing subinterfaces for VXLAN and non-VXLAN VLANs is not supported.
- Point-to-multipoint Layer 3 and SVI uplinks are not supported.
- Using the **ip forward** command enables the VTEP to forward the VXLAN de-capsulated packet destined to its router IP to the SUP/CPU.
- Before configuring it as an SVI, the backup VLAN needs to be configured on Cisco Nexus 9200, 9300-EX, and 9300-FX/FX2 platform switches as an infra-VLAN with the **system nve infra-vlans** command.
- VXLAN is supported on Cisco Nexus 9500 platform switches with the following line cards:
 - 9700-EX
 - 9700-FX
- When Cisco Nexus 9500 platform switches are used as VTEPs, 100G line cards are not supported on Cisco Nexus 9500 platform switches. This limitation does not apply to a Cisco Nexus 9500 switch with 9700-EX or -FX line cards.
- Cisco Nexus 9300 platform switches with 100G uplinks only support VXLAN switching/bridging. Cisco Nexus 9200 and Cisco Nexus 9300-EX/FX platform switches do not have this restriction.



Note For VXLAN routing support, a 40 G uplink module is required.

- The VXLAN UDP port number is used for VXLAN encapsulation. For Cisco Nexus NX-OS, the UDP port number is 4789. It complies with IETF standards and is not configurable.
- For Cisco Nexus 9200 platform switches that have the Application Spine Engine (ASE2). There exists a Layer 3 VXLAN (SVI) throughput issue. There is a data loss for packets of sizes 99 - 122.
- The VXLAN network identifier (VNID) 16777215 is reserved and should not be configured explicitly.
- VXLAN supports In Service Software Upgrade (ISSU).
- VXLAN does not support coexistence with the GRE tunnel feature or the MPLS (static or segment routing) feature.
- VTEP connected to FEX host interface ports is not supported.
- Resilient hashing (port-channel load-balancing resiliency) and VXLAN configurations are not compatible with VTEPs using ALE uplink ports.



Note Resilient hashing is disabled by default.

- When ARP suppression is enabled or disabled in a vPC setup, a down time is required because the global VXLAN vPC consistency checker will fail and the VLANs will be suspended if ARP suppression is disabled or enabled on only one side.



Note For information about VXLAN BGP EVPN scalability, see the *Cisco Nexus 9000 Series NX-OS Verified Scalability Guide, Release 9.3(x)*.

Network Considerations for VXLAN Deployments

- MTU Size in the Transport Network

Due to the MAC-to-UDP encapsulation, VXLAN introduces 50-byte overhead to the original frames. Therefore, the maximum transmission unit (MTU) in the transport network needs to be increased by 50 bytes. If the overlays use a 1500-byte MTU, the transport network needs to be configured to accommodate 1550-byte packets at a minimum. Jumbo-frame support in the transport network is required if the overlay applications tend to use larger frame sizes than 1500 bytes.

- ECMP and LACP Hashing Algorithms in the Transport Network

As described in a previous section, Cisco Nexus 9000 Series Switches introduce a level of entropy in the source UDP port for ECMP and LACP hashing in the transport network. As a way to augment this implementation, the transport network uses an ECMP or LACP hashing algorithm that takes the UDP source port as an input for hashing, which achieves the best load-sharing results for VXLAN encapsulated traffic.

- Multicast Group Scaling

The VXLAN implementation on Cisco Nexus 9000 Series Switches uses multicast tunnels for broadcast, unknown unicast, and multicast traffic forwarding. Ideally, one VXLAN segment mapping to one IP multicast group is the way to provide the optimal multicast forwarding. It is possible, however, to have multiple VXLAN segments share a single IP multicast group in the core network. VXLAN can support up to 16 million logical Layer 2 segments, using the 24-bit VNID field in the header. With one-to-one mapping between VXLAN segments and IP multicast groups, an increase in the number of VXLAN segments causes a parallel increase in the required multicast address space and the amount of forwarding states on the core network devices. At some point, multicast scalability in the transport network can become a concern. In this case, mapping multiple VXLAN segments to a single multicast group can help conserve multicast control plane resources on the core devices and achieve the desired VXLAN scalability. However, this mapping comes at the cost of suboptimal multicast forwarding. Packets forwarded to the multicast group for one tenant are now sent to the VTEPs of other tenants that are sharing the same multicast group. This causes inefficient utilization 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-tenant VXLAN networks to share a multicast group does not bring any implications to the Layer 2 isolation between the tenant 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 it forward the packet to that VXLAN segment. Other tenant networks will not receive the packet. Thus, the segregation between VXLAN segments is not compromised.

Considerations for the Transport Network

The following are considerations for the configuration of the transport network:

- On the VTEP device:
 - Enable and configure IP multicast.*
 - Create and configure a loopback interface with a /32 IP address.
(For vPC VTEPs, you must configure primary and secondary /32 IP addresses.)
 - Enable IP multicast on the loopback interface.*
 - Advertise the loopback interface /32 addresses through the routing protocol (static route) that runs in the transport network.
 - Enable IP multicast on the uplink outgoing physical interface.*
- Throughout the transport network:
 - Enable and configure IP multicast.*

For Cisco Nexus 9200, 9300-EX, and 9300-FX/FX2, the use of the **system nve infra-vlans** command is required. Otherwise, VXLAN traffic (IP/UDP 4789) is actively treated by the switch. The following scenarios are a non-exhaustive list but most commonly seen, where the need for a **system nve infra-vlans** definition is required.

Every VLAN that is not associated with a VNI (vn-segment) is required to be configured as a **system nve infra-vlans** in the following cases:

In the case of VXLAN flood and learn as well as VXLAN EVPN, the presence of non-VXLAN VLANs could be related to:

- An SVI related to a non-VXLAN VLAN is used for backup underlay routing between vPC peers via a vPC peer-link (backup routing).
- An SVI related to a non-VXLAN VLAN is required for connecting downstream routers (external connectivity, dynamic routing over vPC).
- An SVI related to a non-VXLAN VLAN is required for per Tenant-VRF peering (L3 route sync and traffic between vPC VTEPs in a Tenant VRF).
- An SVI related to a non-VXLAN VLAN is used for first-hop routing toward endpoints (Bud-Node).

In the case of VXLAN flood and learn, the presence of non-VXLAN VLANs could be related to:

- An SVI related to a non-VXLAN VLAN is used for an underlay uplink toward the spine (Core port).

The rule of defining VLANs as **system nve infra-vlans** can be relaxed for special cases such as:

- An SVI related to a non-VXLAN VLAN that does not transport VXLAN traffic (IP/UDP 4789).
- Non-VXLAN VLANs that are not associated with an SVI or not transporting VXLAN traffic (IP/UDP 4789).



Note You must not configure certain combinations of infra-VLANs. For example, 2 and 514, 10 and 522, which are 512 apart. This is specifically but not exclusive to the “Core port” scenario that is described for VXLAN flood and learn.



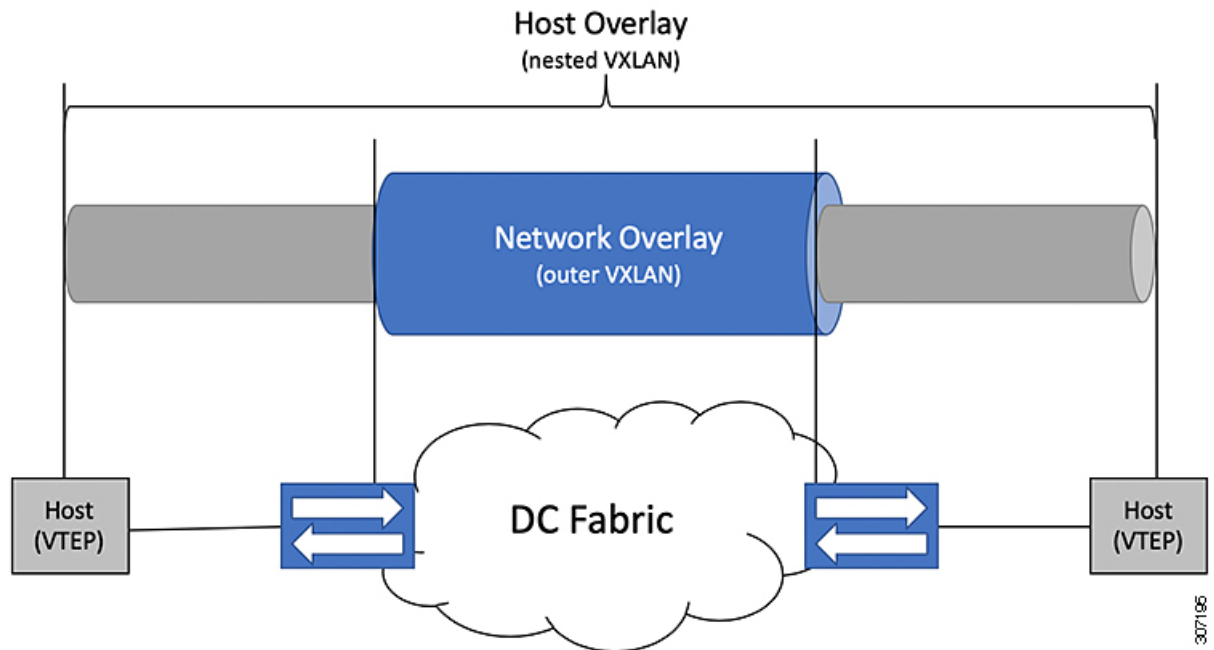
Note * Not required for static ingress replication or BGP EVPN ingress replication.

Considerations for Tunneling VXLAN

DC Fabrics with VXLAN BGP EVPN are becoming the transport infrastructure for overlays. These overlays, often originated on the server (Host Overlay), require integration or transport over the top of the existing transport infrastructure (Network Overlay).

Nested VXLAN (Host Overlay over Network Overlay) support has been added starting with Cisco NX-OS Release 7.0(3)I7(4) and Cisco NX-OS Release 9.2(2) on the Cisco Nexus 9200, 9300-EX, 9300-FX, 9300-FX2, 9500-EX, 9500-FX platform switches.

Figure 1: Host Overlay



To provide Nested VXLAN support, the switch hardware and software must differentiate between two different VXLAN profiles:

- VXLAN originated behind the Hardware VTEP for transport over VXLAN BGP EVPN (nested VXLAN)
- VXLAN originated behind the Hardware VTEP to integrated with VXLAN BGP EVPN (BUD Node)

The detection of the two different VXLAN profiles is automatic and no specific configuration is needed for nested VXLAN. As soon as VXLAN encapsulated traffic arrives in a VXLAN enabled VLAN, the traffic is transported over the VXLAN BGP EVPN enabled DC Fabric.

The following attachment modes are supported for Nested VXLAN:

- Untagged traffic (in native VLAN on a trunk port or on an access port)
- Tagged traffic (tagged VLAN on a IEEE 802.1Q trunk port)
- Untagged and tagged traffic that is attached to a vPC domain
- Untagged traffic on a Layer 3 interface of a Layer 3 port-channel interface

Configuring VXLAN

Enabling VXLANs

Procedure

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

Mapping VLAN to VXLAN VNI

Procedure

	Command or Action	Purpose
Step 1	configure terminal	Enters global configuration mode.
Step 2	vlan <i>vlan-id</i>	Specifies VLAN.
Step 3	vn-segment <i>vnid</i>	Specifies VXLAN VNID (Virtual Network Identifier)
Step 4	exit	Exit configuration mode.

Creating and Configuring an NVE Interface and Associate VNIs

An NVE interface is the overlay interface that terminates VXLAN tunnels.

You can create and configure an NVE (overlay) interface with the following:

Procedure

	Command or Action	Purpose
Step 1	<code>configure terminal</code>	Enters global configuration mode.
Step 2	<code>interface nve x</code>	Creates a VXLAN overlay interface that terminates VXLAN tunnels. Note Only 1 NVE interface is allowed on the switch.
Step 3	<code>source-interface src-if</code>	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 be known by the transient devices in the transport network and the remote VTEPs. This is accomplished by advertising it through a dynamic routing protocol in the transport network.
Step 4	<code>member vni vni</code>	Associate VXLAN VNIs (Virtual Network Identifiers) with the NVE interface.
Step 5	<code>mcast-group start-address [end-address]</code>	Assign a multicast group to the VNIs. Note Used only for BUM traffic

Configuring a VXLAN VTEP in vPC

You can configure a VXLAN VTEP in a 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 Define a non-VXLAN enabled VLAN as a backup routed path.

```
switch(config)# system nve infra-vlans 10
```

Step 9 Create the VLAN to be used as an infra-VLAN.

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

Step 10 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
(Optional) switch(config-if)# ip igmp static-oif route-map match-mcast-groups
switch(config-if)# no shutdown
(Optional) switch(config)# route-map match-mcast-groups permit 10
(Optional) switch(config-route-map)# match ip multicast group 225.1.1.1/32
```

Step 11 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 router ospf process tag area area id
switch(config-if)# ip pim sparse-mode
switch(config-if)# no shutdown
```

Step 12 Create a primary IP address for the data plane loopback interface.

```
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 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 13 Create a vPC domain.

```
switch(config)# vpc domain 5
```

Step 14 Configure the IPv4 address for the remote end of the vPC peer-keepalive link.

```
switch(config-vpc-domain)# peer-keepalive destination 172.28.230.85
```

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 9000 Series NX-OS Unicast Routing Configuration Guide](#).

Step 15 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 16 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 17 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 18 (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 19 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 1,10,100-200
switch(config-if) # mtu 9216
switch(config-if) # vpc peer-link
switch(config-if) # no shutdown
switch(config-if) # interface Ethernet 1/1 , 1/21
switch(config-if) # switchport
switch(config-if) # mtu 9216
switch(config-if) # channel-group 1 mode active
switch(config-if) # no shutdown
```

Step 20 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 21 (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 recommended setting the delay restore for interface-vlan to 45 seconds.

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

Configuring Static MAC for VXLAN VTEP

Static MAC for VXLAN VTEP is supported on Cisco Nexus 9300 Series switches with flood and learn. This feature enables the configuration of static MAC addresses behind a peer VTEP.



Note Static MAC cannot be configured for a control plane with a BGP EVPN-enabled VNI.

Procedure

	Command or Action	Purpose
Step 1	configure terminal	Enters global configuration mode.
Step 2	mac address-table static <i>mac-address</i> vni <i>vni-id</i> interface nve <i>x</i> peer-ip <i>ip-address</i>	Specifies the MAC address pointing to the remote VTEP.
Step 3	exit	Exits global configuration mode.
Step 4	(Optional) copy running-config startup-config	Saves the change persistently through reboots and restarts by copying the running configuration to the startup configuration.
Step 5	(Optional) show mac address-table static interface nve <i>x</i>	Displays the static MAC addresses pointing to the remote VTEP.

Example

The following example shows the output for a static MAC address configured for VXLAN VTEP:

```
switch# show mac address-table static interface nve 1
```

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
* 501	0047.1200.0000	static	-	F	F	nve1(33.1.1.3)
* 601	0049.1200.0000	static	-	F	F	nve1(33.1.1.4)

Disabling VXLANs

Procedure

	Command or Action	Purpose
Step 1	configure terminal	Enters global configuration mode.
Step 2	no feature vn-segment-vlan-based	Disables the global mode for all VXLAN bridge domains

	Command or Action	Purpose
Step 3	<code>no feature nv overlay</code>	Disables the VXLAN feature.
Step 4	(Optional) <code>copy running-config startup-config</code>	Saves the change persistently through reboots and restarts by copying the running configuration to the startup configuration.

Configuring BGP EVPN Ingress Replication

The following enables BGP EVPN with ingress replication for peers.

Procedure

	Command or Action	Purpose
Step 1	<code>configure terminal</code>	Enters global configuration mode.
Step 2	<code>interface nve x</code>	Creates a VXLAN overlay interface that terminates VXLAN tunnels. Note Only 1 NVE interface is allowed on the switch.
Step 3	<code>source-interface src-if</code>	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 be known by the transient devices in the transport network and the remote VTEPs. This is accomplished by advertising it through a dynamic routing protocol in the transport network.
Step 4	<code>member vni vni</code>	Associate VXLAN VNIs (Virtual Network Identifiers) with the NVE interface.
Step 5	<code>ingress-replication protocol bgp</code>	Enables BGP EVPN with ingress replication for the VNI.

Configuring Static Ingress Replication

The following enables static ingress replication for peers.

Procedure

	Command or Action	Purpose
Step 1	<code>configuration terminal</code>	Enters global configuration mode.
Step 2	<code>interface nve x</code>	Creates a VXLAN overlay interface that terminates VXLAN tunnels.

	Command or Action	Purpose
		Note Only 1 NVE interface is allowed on the switch.
Step 3	member vni [<i>vni-id</i> <i>vni-range</i>]	Maps VXLAN VNIs to the NVE interface.
Step 4	ingress-replication protocol static	Enables static ingress replication for the VNI.
Step 5	peer-ip <i>n.n.n.n</i>	Enables peer IP.

