



Virtual eXtensible LAN (VXLAN)



What is VXLAN?

- ▶ Virtual eXtensible Local Area Network
- ▶ A layer 2 in layer 3 overlay tunnel
 - Specifically an Ethernet in UDP tunnel
- ▶ Standards based
 - [RFC 7348 - Virtual eXtensible Local Area Network \(VXLAN\): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks](#)
 - [RFC 7432 - BGP MPLS-Based Ethernet VPN](#)
 - Technically agnostic to the data plane encapsulation

Why use VXLAN?

▶ Expands VLAN name space

- VLANs use 12 bit space – 4096 values
- VXLAN uses 24 bit space – 16,777,216 values

▶ Allows layer 2 multipathing

- Doesn't need STP for loop prevention
- Uses layer 3 ECMP over CLOS fabric
 - Similar logic to FabricPath

Why use VXLAN?

▶ Includes scaling enhancements

- Optimizes control plane, e.g. MAC learning, ARP tables, BUM replication, etc.

▶ Doesn't break layer 2 adjacency requirements

- Allows for any to any stateless layer 2 and layer 3 transport
- E.g. vMotion

▶ Allows for multi-tenancy

- Separation of customer traffic over shared underlay fabric
- Allows for overlapping layer 2 and layer 3 addresses
 - E.g. VLANs & IPs are locally significant

VXLAN Terminology

- ▶ Underlay Network – provides the transport for VXLAN
 - OSPF/EIGRP/IS-IS routed fabric
- ▶ Overlay Network – uses the service provided by VXLAN
 - E.g. web server in rack 1 calling the database server in rack 2

VXLAN Terminology

- ▶ VXLAN - Virtual eXtensible Local Area Network
- ▶ VNI / VNID - VXLAN Network Identifier
 - Replaces the VLAN ID
- ▶ VTEP - VXLAN Tunnel End Point
 - Box that performs VXLAN encap/decap
 - Could be hardware or software
 - E.g. Nexus 5600 vs. Nexus 1000v

VXLAN Terminology

▶ VXLAN Segment

- The resulting Layer 2 overlay network

▶ VXLAN Gateway

- Device that forwards traffic between VXLANs
- Can be both layer 2 and layer 3 forwarding

▶ NVE – Network Virtualization Edge

- Logical representation of the VTEP
- I.e. NVE is the tunnel interface

VXLAN Encapsulation

```
Frame 22: 124 bytes on wire (992 bits), 124 bytes captured (992 bits) on interface 0
Ethernet II, Src: CiscoInc_12:1a:7c (00:de:fb:12:1a:7c), Dst: CiscoInc_d7:60:42 (68:bd:ab:d7:60:42)
Internet Protocol Version 4, Src: 10.1.1.51, Dst: 10.1.1.52
User Datagram Protocol, Src Port: 39965 (39965), Dst Port: 4789 (4789)
  Source Port: 39965
  Destination Port: 4789
  Length: 90
  Checksum: 0x0000 (none)
  [Stream index: 1]
Virtual eXtensible Local Area Network
  Flags: 0x0800, VXLAN Network ID (VNI)
    0... .. = GBP Extension: Not defined
    .... .0.. .. = Don't Learn: False
    .... 1... .. = VXLAN Network ID (VNI): True
    .... .... 0... = Policy Applied: False
    .000 .000 0.00 .000 = Reserved(R): False
  Group Policy ID: 0
  VXLAN Network Identifier (VNI): 100010
  Reserved: 0
Ethernet II, Src: IntelCor_8d:3d:98 (00:1b:21:8d:3d:98), Dst: IntelCor_88:80:75 (00:1b:21:88:80:75)
  Destination: IntelCor_88:80:75 (00:1b:21:88:80:75)
  Source: IntelCor_8d:3d:98 (00:1b:21:8d:3d:98)
  Type: IPv4 (0x0800)
Internet Protocol Version 4, Src: 10.0.0.11, Dst: 10.0.0.12
Internet Control Message Protocol
```

Basic VXLAN Workflow

- ▶ Receive ARP from local host
 - Assume a miss occurs
- ▶ Find the remote VTEP...
 - Multicast flood and learn
 - Ingress replication
 - MP-BGP L2VPN EVPN
- ▶ Unicast encap frame towards the VTEP
 - Throw away the VLAN
 - Replace it with the VNID

Recommended Resources

▶ Documentation

- [VXLAN Overview: Cisco Nexus 9000 Series Switches](#)
- [VXLAN Design with Cisco Nexus 9300 Platform Switches](#)
- [VXLAN Network with MP-BGP EVPN Control Plane Design Guide](#)
- [Cisco Programmable Fabric with VXLAN BGP EVPN Configuration Guide](#)

▶ Cisco Live

- [BRKRST-3045 - Introduction to LISP and VXLAN - Scalable Technology Overlays for Switching](#)
- [BRKDCT-3378 - Building Data Centre Networks with VXLAN BGP-EVPN](#)

Q&A

A person is seen from behind, sitting in front of a large window. The entire image is covered with a warm, orange-toned overlay, creating a soft, hazy atmosphere. The window has a grid pattern, and the person's silhouette is visible against the bright light coming through the glass.

In This Section

- ▶ VXLAN Configuration Prerequisites
- ▶ Implementing VXLAN Flood & Learn on NX-OS

VXLAN Prerequisites

- ▶ Prerequisites are hardware/software specific
- ▶ For Nexus 5600 as hardware VTEP...
 - Set switching mode to store-and-forward
 - hardware ethernet store-and-fwd-switching
 - Requires a reboot
 - Establish IP unicast reachability between VTEPs
 - Establish PIM BIDIR reachability between VTEPs
 - Spines can be phantom RPs for redundancy
 - Enable features
 - feature vn-segment-vlan-based
 - feature nv overlay

VXLAN Flood & Learn Config Steps

- ▶ Map VLAN to VXLAN
 - `vn-segment` under `vlan config` mode
- ▶ Create Network Virtualization Edge (NVE) interface
 - `interface NVE1`
- ▶ Specify VTEP source
 - `source interface loopback0`
- ▶ Specify VNI membership
 - `member vni [vnid]`
- ▶ Specify multicast group for BUM replication
 - `mcast-group [group]`

VXLAN Flood & Learn Verifications

▷ Common verifications...

- `show interface nve id`
- `show platform fwm info nve peer [all]`
- `show mac address-table`
- `show nve peer`
- `show nve vni`
- `show platform fwm info nve vni`

Q&A

A person is seen from behind, sitting in front of a large window. The entire image is covered with a warm, orange-toned overlay, creating a soft, hazy atmosphere. The window has a grid pattern, and the person's silhouette is visible against the bright light coming through the glass.

In This Section

- ▶ Implementing VXLAN BGP EVPN on NX-OS

VXLAN BGP EVPN Prerequisites

- ▶ Prerequisites are hardware/software specific
- ▶ For Nexus 5600 as hardware VTEP...
 - Set switching mode to store-and-forward
 - hardware ethernet store-and-fwd-switching
 - Requires a reboot
 - Establish IP unicast reachability between VTEPs
 - Establish PIM BIDIR reachability between VTEPs
 - Spines can be phantom RPs for redundancy

VXLAN BGP EVPN Prerequisites (cont.)

- For Nexus 5600 as hardware VTEP...
 - Enable features
 - install feature-set virtualization
 - install feature-set fabric
 - feature-set fabric
 - feature fabric forwarding
 - nv overlay evpn
 - feature nv overlay
 - feature vn-segment-vlan-based

VXLAN BGP EVPN Config Steps

- ▶ Map VLAN to VXLAN
 - `vn-segment` under `vlan config` mode
- ▶ Create Network Virtualization Edge (NVE) interface
 - `interface NVE1`
- ▶ Specify VTEP source
 - `source interface loopback0`
- ▶ Specify VNI membership
 - `member vni [vnid]`
- ▶ Specify multicast group for BUM replication
 - `mcast-group [group]`
- ▶ Specify BGP as control plane protocol
 - `host-reachability protocol bgp`

VXLAN BGP EVPN Config Steps

- ▶ Establish BGP EVPN peerings
 - address-family l2vpn evpn
 - Extended communities required
- ▶ Generate BGP advertisement
 - evpn
 - vni [vnid] l2
 - rd auto
 - route-target import auto
 - route-target export auto

VXLAN BGP EVPN Verifications

▶ Common verifications...

- `show interface nve id`
- `show platform fwm info nve peer [all]`
- `show mac address-table`
- `show nve peer`
- `show nve vni`
- `show platform fwm info nve vni`
- `show bgp l2vpn evpn summary`
- `show bgp l2vpn evpn`
- `show bgp l2vpn evpn neighbor [neighbor] advertised-routes`

Q&A

A person is seen from behind, sitting in front of a large window. The entire image is covered with a warm, orange-toned overlay, creating a soft, hazy atmosphere. The window has a grid pattern, and the person's silhouette is visible against the bright light coming through the glass.

In Today's Class

- ▶ VXLAN Configuration & Verification Review
- ▶ Symmetric vs. Asymmetric Integrated Routing & Bridging
- ▶ Unicast & Multicast Underlay Fabric High Availability
- ▶ VXLAN Anycast vPC
- ▶ Multi-fabric VXLAN with EBGP L2VPN EVPN



VXLAN Configuration & Verification Review



VXLAN Prerequisites

▶ Prerequisites are hardware/software specific

▶ For Nexus 5600 as hardware VTEP...

- Set switching mode to store-and-forward
 - hardware ethernet store-and-fwd-switching
 - Requires a reboot
- Establish IP unicast reachability between VTEPs
- Establish PIM BIDIR reachability between VTEPs
 - Spines can be phantom RPs for redundancy
- Enable features
 - feature vn-segment-vlan-based
 - feature nv overlay

VXLAN Flood & Learn Config Steps

- ▶ Map VLAN to VXLAN
 - `vn-segment` under `vlan config` mode
- ▶ Create Network Virtualization Edge (NVE) interface
 - `interface NVE1`
- ▶ Specify VTEP source
 - `source interface loopback0`
- ▶ Specify VNI membership
 - `member vni [vnid]`
- ▶ Specify multicast group for BUM replication
 - `mcast-group [group]`

VXLAN Flood & Learn Verifications

▷ Common verifications...

- `show interface nve id`
- `show platform fwm info nve peer [all]`
- `show mac address-table`
- `show nve peer`
- `show nve vni`
- `show platform fwm info nve vni`



VXLAN BGP EVPN

Configuration & Verification

Review



VXLAN Overlay with BGP L2VPN EVPN

▶ What are we trying to accomplish?

- Map VLANs to VXLAN Network Identifiers (VNIs/VNIDs)
- Advertise information into BGP
 - MAC to L2 VNI to VTEP mapping
 - IP to L3 VNI to VTEP mapping
- Import MAC addresses into the CAM table for bridging
- Route traffic through SVIs to remote segments

VXLAN BGP EVPN Config Steps

- ▶ Map VLAN to VXLAN
 - `vn-segment` under `vlan` config mode
- ▶ Create Network Virtualization Edge (NVE) interface
 - `interface NVE1`
- ▶ Specify VTEP source
 - `source interface loopback0`
- ▶ Specify VNI membership
 - `member vni [vnid]`
- ▶ Specify multicast group for BUM replication
 - `mcast-group [group]`
- ▶ Specify BGP as control plane protocol
 - `host-reachability protocol bgp`

VXLAN BGP EVPN Config Steps

- ▶ Establish BGP EVPN peerings
 - address-family l2vpn evpn
 - Extended communities required
- ▶ Generate BGP advertisement
 - evpn
 - vni [vnid] l2
 - rd auto
 - route-target import auto
 - route-target export auto

VXLAN BGP EVPN Verifications

▶ Common verifications...

- `show interface nve id`
- `show platform fwm info nve peer [all]`
- `show mac address-table`
- `show nve peer`
- `show nve vni`
- `show platform fwm info nve vni`
- `show bgp l2vpn evpn summary`
- `show bgp l2vpn evpn`
- `show bgp l2vpn evpn neighbor [neighbor] advertised-routes`

Q&A

A person is seen from behind, sitting in front of a large window. The entire image is covered with a warm, orange-toned overlay. The text "Q&A" is centered in white.



Inter-VXLAN Routing



Asymmetric vs. Symmetric IRB

▶ EVPN Integrated Routing and Bridging (IRB) has two options:

- Asymmetric IRB
- Symmetric IRB

▶ Asymmetric IRB

- Ingress VTEP does both L2 and L3 lookup
- Egress VTEP does L2 lookup only
- I.e. Bridge - Route - Bridge

▶ Symmetric IRB

- Ingress VTEP does both L2 and L3 lookup
- Egress VTEP does both L3 and L2 lookup
- I.e. Bridge - Route - Route - Bridge

Asymmetric IRB Issues

- ▶ VTEP must have all VNIs configured that require routing
 - Result is increased ARP cache and CAM table sizes
 - I.e. control plane scaling issue

How Symmetric IRB Works

- ▶ New concept called Layer-3 VNI
- ▶ Each tenant VRF is mapped to a unique Layer-3 VNI
 - Mapping must match on all VTEPs
- ▶ All VXLAN routed traffic is encapsulated with L3 VNI in VXLAN header
 - Allows for single shared VNI among all VTEPs
- ▶ L2 VNIs only need be configured where access ports exist
 - Result is savings of ARP and CAM table spaces

Q&A

A person is seen from behind, sitting in front of a large window. The entire image is covered with a warm, orange-toned overlay. The text "Q&A" is centered in white.



vPC and VXLAN BGP EVPN



vPC & VXLAN BGP Traffic Flow Problems

- ▶ VXLAN traffic is tunneled over the underlay network using the BGP next-hop address of the remote VTEP
 - NVE source interface (i.e. Loopback0) is the default BGP next-hop for advertised routes
 - In a vPC, both vPC peers advertise duplicate EVPN MAC/IP routes to spine RRs
 - With other attributes equal, next-hop is tie breaker in BGP Best Path selection
 - Implies that one vPC peer is always preferred for dual attached hosts
 - Result is that egress traffic from vPC Member is load balanced, but return ingress traffic is polarized
- ▶ Workaround is to use Anycast VTEP address

vPC Anycast VTEP

- ▶ vPC Peers share duplicate IP address on NVE source interface
 - Peer 1 – interface Loopback0 ; ip address 1.1.1.51/32
 - Peer 2 – interface Loopback0 ; ip address 1.1.1.52/32
 - Both Peers – interface Loopback0 ; ip address 1.1.1.111/32 secondary
- ▶ BGP next-hop is automatically set to secondary address for locally originated routes
 - I.e. L2VPN EVPN MAC/IP routes for vPC Member Ports
- ▶ Result is that ingress flows from spines are load balanced
 - Other leafs use IGP ECMP to reach shared secondary IP address

Nexus 5600 & NVE Peer-Link-VLAN

- ▶ On Nexus 5600, all traffic across the vPC Peer Link must be VXLAN encapsulated due to ASIC implementation
- ▶ Normal vPC Peer Link is a Classical Ethernet trunk
 - Result is that East/West flows over vPC Peer Link are broken by default
 - i.e. the VNI number is lost when packet is sent out the Peer Link
- ▶ Peer Link is normally only used for orphans or in failure scenarios
 - Result is that everything looks fine until the failure occurs
 - Traffic to orphans & single attached members black-holed over vPC Peer Link
- ▶ Workaround is to maintain VXLAN encapsulation across Peer Link
 - Implemented as `vpc nve peer-link-vlan`

Configuring NVE Peer-Link-VLAN

- ▶ Create new VLAN and specify as NVE Peer Link VLAN
 - `vlan 999`
 - `vpc nve peer-link-vlan 999`
- ▶ Establish layer 3 peering across NVE Peer Link VLAN
 - `interface vlan 999`
 - `ip router ospf 1 area 0`
 - `ip router isis 1`
- ▶ Traffic engineer so other vPC Peer's VTEP Loopback is preferred over vPC Peer Link
 - `ip ospf cost 10`
 - `isis metric 10 level-2`
- ▶ Verify that traffic flows are successful for orphans and single attached members

Q&A

A person is seen from behind, sitting in front of a large window. The entire image is covered with a warm, orange-toned overlay, creating a soft, hazy atmosphere. The window has a grid pattern, and the person's silhouette is visible against the bright light coming through the glass.



VXLAN Underlay Fabric Unicast High Availability



VXLAN Underlay Fabric Convergence

- ▶ VXLAN underlay fabric convergence is based on three factors
 - IGP convergence
 - PIM convergence
 - BGP convergence
- ▶ Factors must be addressed separately to achieve High Availability for VXLAN overlay flows

VXLAN Underlay – Tuning IGP Convergence

▷ Generally four factors affect IGP convergence time

- Failure Detection Time
 - Is the neighbor down?
- Event Propagation Time
 - Tell neighbors about the change
- Recalculation Time
 - Run SPF/DUAL/etc. calculation
- Forwarding Table Update Time
 - Install new paths

Failure Detection Time

- ▶ How long does it take me to realize there is a failure?
- ▶ Example failure detections
 - Link up/down event
 - Routing protocol hello/dead timers
 - IP SLA & EEM
 - Bidirectional Forwarding Detection (BFD)

Event Propagation Time

- ▶ How long does it take me to tell everyone else?
- ▶ Example event propagation
 - EIGRP Query/Reply
 - OSPF LSA Flooding Procedure
 - BGP Update/Withdraw

Recalculation Time

- ▶ How long does it take me to decide on the new topology?
- ▶ Example recalculation time
 - EIGRP DUAL
 - OSPF SPF
 - BGP Best Path Selection

Forwarding Table Update Time

▷ How long does it take me to install the changes?

▷ Example update time

- EIGRP topology to RIB download
- RIB to software FIB download
- Software FIB to hardware TCAM download

Example OSPF Reconvergence

▷ OSPF Failure Detection

- Neighbor dead interval expires

▷ OSPF Event Propagation

- LSA flooding procedure

▷ OSPF Recalculation Time

- Shortest Path First (SPF) runtime

▷ Forwarding Table Update Time

- OSPF database to RIB installation, RIB to FIB, FIB to TCAM

How Do We Affect Convergence?

- ▶ Some factors are software & configuration dependent
 - E.g. smaller EIGRP Query domain is better
 - E.g. OSPF stub areas are better
 - E.g. unnumbered fabric links or prefix-suppression is better
- ▶ Some factors are hardware dependent
 - E.g. SPF runtime is a function of CPU size
 - E.g. TCAM download is a function of the linecard

Methods of Modifying Convergence Time

▷ Can be both reactive and proactive

▷ Reactive optimizations

- E.g. carrier delay & link debounce timer
- E.g. Fast Hellos & BFD
- E.g. OSPF LSA & SPF Pacing
- E.g. FIB prefix prioritization

▷ Proactive optimizations

- EIGRP Feasible Successors
- OSPF Loop Free Alternate (LFA)
- BGP Prefix Independent Convergence (PIC)
- MPLS Traffic Engineering Fast Reroute (TE FRR)



Bidirectional Forwarding Detection (BFD)



What is BFD?

▷ [RFC 5880 - Bidirectional Forwarding Detection](#)

- Lightweight keepalive used for path failure detection

▷ Multiple implementations

- [Bidirectional Forwarding Detection \(BFD\) for IPv4 and IPv6 \(Single Hop\)](#)
- [Generic Application of Bidirectional Forwarding Detection \(BFD\)](#)
- [Bidirectional Forwarding Detection \(BFD\) for Multihop Paths](#)
- [Bidirectional Forwarding Detection \(BFD\) for MPLS Label Switched Paths \(LSPs\)](#)

Why Use BFD?

▶ Fast failure detection

- Typically 150ms, can be faster depending on hardware

▶ Independent of...

- media type, encapsulation, topology, & routing protocol

▶ BFD can be hardware offloaded

- IGP & BGP is process switched
- Fast IGP hellos can thrash the CPU

▶ Multiple protocols can register to BFD

- IGP, PIM, BGP, HSRP, MPLS TE, etc. can all use the same single BFD session
- Cuts down on control plane load significantly

How BFD Works

▶ BFD is a UDP-based ping

- Source IP – Me
- Destination IP – Me
- BFD Control packets at UDP destination port 3784
- BFD Echo packets at UDP destination port 3785

▶ Goal is to loop the packet back

- You receive packet with destination of Me
- You route packet back to Me out the same link to received it in
- If I receive my packet back in, link is bidirectional

▶ ICMP Redirects must be disabled globally or per link

- BFD is hardware offloaded but ICMP is not
- Redirects can cause CPU spikes and false positives
 - E.g. router/switch thinks neighbor is down but its really up, and reconvergence starts

Configuring BFD on NX-OS

▶ Enable feature

- `feature bfd`

▶ Set BFD interval globally or per-link

- `bfd interval milliseconds min_rx milliseconds multiplier interval-multiplier`

▶ Register upper-layer protocol with BFD

- `router ospf 1 ; bfd / router isis 1 ; bfd / router eigrp 1 ; bfd / ip pim bfd / (config-if)# hsrp bfd / etc.`

▶ Verify neighbor reachability

- `show bfd neighbors`
- `show bfd clients`

Q&A

A person is seen from behind, sitting in front of a large window. The entire image is covered with a warm, orange-toned overlay. The text "Q&A" is centered in white.