Cisco Support Community
Expert Series Webcast
Cisco Data Center Overlays with focus on VXLAN
Vishal Mehta, CCIE Data Center, SP, and R&S
October 20, 2015
Upcoming Events

https://supportforums.cisco.com/expert-corner/events
Become an Event Top Contributor

Participate in Live Interactive Technical Events and much more
http://bit.ly/1jl93B

https://supportforums.cisco.com/expert-corner/top-contributors
Rate Content

Encourage and acknowledge people who generously share their time and expertise

Now your ratings on documents, videos, and blogs count give points to the authors!!!

So, when you contribute and receive ratings you now get the points in your profile.

Help us to recognize the quality content in the community and make your searches easier. Rate content in the community.

https://supportforums.cisco.com/blog/154746
Cisco Support Community Expert Series Webcast

Vishal Mehta
CCIE Data Center SP and R&S #37139
Ask the Expert Event following the Webcast

Now through October 30th


Join the discussion for these Ask The Expert Events:
If you would like a copy of the presentation slides, click the PDF file link in the chat box on the right or go to:

Submit Your Questions Now!

Use the Q & A panel to submit your questions and the panel of experts will respond.

Please take a moment to complete the survey at the end of the webcast
Polling Question 1

Are you planning to implement VXLAN in your network?

- Yes
- No
- Still Evaluating
Agenda

• Overlays in Data Center
• Why VXLAN
• VXLAN Evolution
• Deployment Scenarios
• Comparison
Why Do We Need Overlays?

Location and Identity Separation

**Traditional Behaviour**
Loc/ID “Overloaded” Semantic

Device IPv4 or IPv6 Address Represents **Identity** and **Location**

When the Device Moves, It Gets a New IPv4 or IPv6 Address for Its New **Identity** and **Location**

**Overlay Behaviour**
Loc/ID “Split”

Device IPv4 or IPv6 Address Represents **Identity Only**. Its **Location** Is Here!

When the Device Moves, Keeps Its IPv4 or IPv6 Address. It Has the Same **Identity**

Only the **Location** Changes
Overlay Taxonomy

Service = Virtual Network Instance (VNI)
Identifier = VN Identifier (VNID)
NVE = Network Virtualization Edge
VTEP = VXLAN Tunnel End-Point
Overlay Attributes

Service
- Layer 2 Service
- Layer 3 Service

Edge Device
- Host Overlays
- Network Overlays

Signalling
- Data Plane Learning
- Control Plane Learning
Types of Overlay Service

Layer 2 Overlays
- Emulate a LAN segment
- Transport Ethernet Frames (IP and non-IP)
- Single subnet mobility (L2 domain)
- Exposure to open L2 flooding
- Useful in emulating physical topologies

Layer 3 Overlays
- Abstract IP based connectivity
- Transport IP Packets
- Full mobility regardless of subnets
- Contain network related failures (flooding)
- Useful in abstracting connectivity and policy

Hybrid L2/L3 Overlays offer the best of both domains
Overlay Edge Device & Data Plane Evolution

Service

Layer 2 Service
Layer 3 Service

Edge Device

Host Overlays
Network Overlays

Network DB

Virtual
Physical
Overlay Network Evolution: Edge Devices

**Network Overlays**
- Physical
- Router/switch end-points
- Protocols for resiliency/loops
- Traditional VPNs
- OTV, VPLS, LISP, FP

**Host Overlays**
- Virtual
- Virtual end-points only
- Single admin domain
- VXLAN, NVGRE, STT

**Hybrid Overlays**
- Virtual
- Physical
- Physical and Virtual - VXLAN
- Resiliency + Scale
- x-organization/federation
- Open Standards
Overlay Signalling

**Data Plane**

- Based on gleaning information from data plane events
  - Example: Source Learning on bridges, FabricPath, VXLAN (Multicast)
- Provides the following:
  - Address advertisement/mapping
  - Some tunnel management is possible
  - Does not provide Service Auto-discovery
- Requires a flood facility for data plane events to propagate:
  - Multicast tree
  - Unicast replication group at the head-end

**Control Plane**

- Provides:
  - Service Discovery
  - Address Advertising/Mapping
  - Tunnel Management
  - Extensions for multi-homing and advanced services can be provided

**Protocol or Controller:**

- **Routing Protocol** amongst Edge Devices
  - BGP, IS-IS, LISP
- **Central database on a Controller**
  - Distributed Virtual Switches (OVS, N1Kv/VSM)

**Push or Pull:**

- **Push** all information to all Edge Devices
  - BGP, IS-IS, Controllers
- **Pull** and cache on demand @ ED
  - LISP, DNS, Controllers
Modern DC Fabric

Seek well integrated best in class Overlays and Underlays

Robust Underlay/Fabric
- High Capacity Resilient Fabric
- Intelligent Packet Handling
- Programmable & Manageable

Flexible Overlay Virtual Network
- Mobility – Track end-point attach at edges
- Scale – Reduce core state
  - Distribute and partition state to network edge
- Flexibility/Programmability
  - Reduced number of touch points
Trend: Flexible Data Center Fabrics

Workload Mobility
Workload Placement
Segmentation
Scale
Automation & Programmability
L2 + L3 Connectivity
Physical + Virtual
Open
Network Virtualization
Data Center “Fabric” Journey

STP

VPC

FabricPath

VXLAN

FabricPath /BGP

VXLAN /EVPN

© 2015 Cisco and/or its affiliates. All rights reserved.
Which Encapsulation?

VXLAN

NVGRE

LISP

MPLS

FabricPath
Why VXLAN

Standards based (VXLAN-RFC7348) Overlay with 16M identifiers
Leverages Layer-3 ECMP – all links forwarding
Integration of Physical and Virtual Nodes

Limited Rack-wide VM Mobility

Virtual/Cloud Data Center
Network Virtualization with VXLAN

**Underlay Network:**
- IP routing – proven, stable, scalable
- Support any routing protocols --- OSFP, EIGRP, IS-IS, BGP, etc.
- ECMP – utilize all available network paths

**Overlay Network:**
- Standards-based overlay
- Layer-2 extensibility and mobility
- Expanded Layer-2 name space
- Scalable network domain
- Multi-Tenancy
VXLAN VTEP

VXLAN terminates its tunnels on VTEPs (Virtual Tunnel End Point).

Each VTEP has two interfaces, one is to provide bridging function for local hosts, the other has an IP identification in the core network for VXLAN encapsulation/decapsulation.
Normalization: The Encapsulation Doesn’t Matter

- Intelligence in the Control Plane
- Capabilities Exchange in Control Plane (negotiate encapsulation)
- Normalize to common encapsulation
- Pervasive Multi-encap Gateways for optimal traffic patterns
VXLAN Frame Format

MAC-in-IP Encapsulation

- **Outer MAC Header**
- **Outer IP Header**
- **UDP Header**
- **VXLAN Header**

**Original Layer-2 Frame**

50 (54) Bytes of Overhead

**Underlay**

14 Bytes (4 Bytes Optional)
- Ether Type 0x0800
- VLAN Type 0x8100
- VLAN ID Tag

**Overlay**

- **Src VTEP MAC Address**
- **Dest. MAC Address**

- **Src MAC Address**
- **VLAN ID**
- **VLAN Type**

- **UDP Header**
  - **Source Port**
  - **UDP Length**
  - **Checksum 0x0000**

- **VXLAN Header**
  - **Source Port**
  - **VXLAN Port**

- **Checksum**

- **Reserved**

- **VNI**

**Hash of the inner L2/L3/L4 headers of the original frame. Enables entropy for ECMP Load balancing in the Network.**

**Src and Dst addresses of the VTEPs**

- **Src VTEP MAC Address**
- **Dest. VTEP MAC Address**

- **Next-Hop MAC Address**

**Allowed for 15M possible Segments**

© 2015 Cisco and/or its affiliates. All rights reserved.
Terminology - Reference

• **Layer-2 VNI:**
  - VNI (VXLAN network identifier) carried in VXLAN packets bridged across VTEPs (VXLAN tunnel end point). This VNI is configured per VLAN.

• **Layer-3 VNI:**
  - VNI carried in the VxLAN packets routed across VTEPs. This VNI is linked per Tenant VRF.

• **Anycast GW:**
  - All L3 VTEPs are configured with same mac and same subnet for host facing SVI.

• **VRF overlay VLAN:**
  - Every Tenant VRF will need a Vlan to be configured for VXLAN routing.
  - This VLAN is configured with L3-VNI.

• **VXLAN L2 Gateway:**
  - VTEP capable of switching VLAN->VXLAN, VXLAN->VLAN packets with in same VNI.

• **VXLAN L3 Gateway:**
  - VTEP capable of routing packets across different VNIs.
VXLAN Overview (1)
VTEP – VXLAN Tunnel End-Point
VNI/VNID – VXLAN Network Identifier
VXLAN L2 and L3 Gateways
Connecting VXLAN to the broader network

L2 Gateway: VXLAN to VLAN Bridging

L3 Gateway: VXLAN to X Routing
- VXLAN
- VLAN
# Cisco VXLAN Portfolio

## Cisco VXLAN Solutions

<table>
<thead>
<tr>
<th>Scale</th>
<th>Secure Multi-tenancy</th>
<th>Workload Mobility</th>
<th>Workload Anywhere</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR1000 CSR1000</td>
<td>Nexus 1000</td>
<td>Nexus 5600</td>
<td>Nexus 7000</td>
</tr>
<tr>
<td>Nexus 1000</td>
<td>Nexus 3100</td>
<td>Nexus 7000</td>
<td>Nexus 9000</td>
</tr>
<tr>
<td>Nexus 9000</td>
<td></td>
<td></td>
<td>ASR9000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L2 Gateway</th>
<th>L3 Gateway</th>
<th>BGP EVPN Control Plane</th>
<th>Anycast Gateway</th>
<th>Head End Replication</th>
</tr>
</thead>
</table>
Common Building Blocks

EXISTING 3-TIER DESIGNS
- DC Core
- DC PODs
- VPC
- FEX

PROGRAMMABLE SDN OVERLAY MODEL
- VXLAN Bridging & Routing
- Integrated Network Virtualization
- SDN Controllers

APPLICATION PROFILES & POLICIES
- Application Centric Infrastructure
- Policy Model
- Automation

Common Building Blocks

Nexus 3000, 5600, 7000

Nexus 9000
VXLAN Evolution

• Yesterday: VXLAN, yet another Overlay
  • Data-Plane only (Multicast based Flood & Learn)

• Today: VXLAN for the creation of scalable DC Fabrics – Intra-DC
  • Control-Plane, active VTEP discovery, Multicast and Unicast (Head-End Replication)

• Tomorrow: VXLAN for DCI – Inter-DC
  • DCI Enhancements (ARP caching/suppress, Multi-Homing, Failure Domain isolation, Loop Protection etc.)
Multicast-Based VxLAN

- No VXLAN control plane
- Data driven flood-&-learn
- Multicast transport for VXLAN BUM (Broadcast, Unknown Unicast and Multicast) traffic.
VXLAN Flood & Learn

MAC  VNI  VTEP
MAC_A  30000  E1/12

MAC  VNI  VTEP
MAC_B  30000  E1/4

MAC  VNI  VTEP
MAC_C  30000  E1/8

Host A
MAC_A / IP_A

Host B
MAC_B / IP_B

Host C
MAC_C / IP_C

Destination Group
239.1.1.1
(00:01:00:01:01:01)
VTEP Peer Discovery & Address Learning (1)

**VXLAN Flood & Learn**

**Underlay**
- **Host A**
  - MAC: MAC_A / IP_A
  - ARP Request for IP_B
    - Src MAC: MAC_A

**SMAC:** MAC_V1
**DMAC:** 00:01:5E:01:01:01
**SIP:** IP_V1
**DIP:** 239.1.1.1
**UDP**
**VXLAN VNID:** 30000

**Overlay**
- **Virtual Switch**
- **Host C**
  - MAC: MAC_C / IP_C
  - ARP Request for IP_B
    - Src MAC: MAC_A

**MAC VNI VTEP**
- MAC_A 30000 E1/12

**Destination Group**
239.1.1.1
(00:01:5E:01:01:01)

**Overlay**
- **Host B**
  - MAC: MAC_B / IP_B
  - ARP Request for IP_B
    - Src MAC: MAC_A

**MAC VNI VTEP**
- MAC_B 30000 E1/4
- MAC_A 30000 IP_V1

**MAC VNI VTEP**
- MAC_A 30000 E1/8

© 2015 Cisco and/or its affiliates. All rights reserved.
VTEP Peer Discovery & Address Learning (2)

VXLAN Flood & Learn

<table>
<thead>
<tr>
<th>MAC</th>
<th>VNI</th>
<th>VTEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC_A</td>
<td>30000</td>
<td>E1/12</td>
</tr>
<tr>
<td>MAC_B</td>
<td>30000</td>
<td>IP_V2</td>
</tr>
</tbody>
</table>

ARP Response from IP_B
Src MAC: MAC_B  
Dst MAC: MAC_A

 ARP Response for IP_B
Src MAC: MAC_B  
Dst MAC: MAC_A

ARP Response from IP_B
Src MAC: MAC_B  
Dst MAC: MAC_A

Underlay
SIP: IP_V2  
DIP: IP_V1
UDP
VXLAN VNID: 30000

Overlay
ARP Response
SMAC: MAC_B  
DMAC: MAC_A
VTEP Peer Discovery & Address Learning (3)

VXLAN Flood & Learn

1. Host X
   MAC_X / IP_X

2. ARP Request for IP_Y
   Src MAC: MAC_X

3. ARP Request for IP_Y
   Src MAC: MAC_X

4. ARP Request for IP_Y
   Src MAC: MAC_X

MAC VNI VTEP
MAC_X 30001 E1/11

MAC VNI VTEP
MAC_Y 30001 E1/8
MAC_X 30001 V1

Virtual Switch
VTEP Peer Discovery & Address Learning (4)

VXLAN Flood & Learn

<table>
<thead>
<tr>
<th>MAC</th>
<th>VNI</th>
<th>VTEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC_X</td>
<td>30001</td>
<td>E1/11</td>
</tr>
<tr>
<td>MAC_Y</td>
<td>30001</td>
<td>V3</td>
</tr>
</tbody>
</table>

Host X
MAC_X / IP_X

ARP Response for IP_Y
Src MAC: MAC_Y
Dst MAC: MAC_X

SMAC: MAC_V3
DMAC: hop-by-hop

UDP

VXLAN VNID: 30001

ARP Response

SMAC: MAC_Y
DMAC: MAC_X

V1

Host Y
MAC_Y / IP_Y

ARP Response for IP_Y
Src MAC: MAC_Y
Dst MAC: MAC_X

SMAC: MAC_V3
DMAC: hop-by-hop

UDP

VXLAN VNID: 30001

ARP Response

SMAC: MAC_Y
DMAC: MAC_X

V3

Virtual Switch

Overlay

Underlay

SIP: IP_V3
DIP: IP_V1

SMAC: MAC_V3
DMAC: hop-by-hop

UDP

VXLAN VNID: 30001

ARP Response
VXLAN Packet Forwarding (1)

**VXLAN Flood & Learn**

1. **Host A**
   - MAC: MAC_A
   - IP: IP_A
   - SMAC: MAC_A
   - DMAC: MAC_B
   - SIP: IP_A
   - DIP: IP_B

2. **V1**
   - MAC: MAC_A
   - VNI: 30000
   - VTEP: E1/12
   - SMAC: MAC_A
   - DMAC: MAC_B
   - SIP: IP_A
   - DIP: IP_B
   - UDP
   - VXLAN VNID: 30000

3. **V2**
   - MAC: MAC_B
   - VNI: 30000
   - VTEP: E1/4
   - SMAC: MAC_B
   - DMAC: MAC_A
   - SIP: IP_A
   - DIP: IP_B
   - UDP
   - VXLAN VNID: 30000

4. **Host B**
   - MAC: MAC_B
   - IP: IP_B
   - SMAC: MAC_A
   - DMAC: MAC_B
   - SIP: IP_A
   - DIP: IP_B
VXLAN Evolution

• Leveraging the Control-Plane to avoid Flood & Learn VTEP discovery (pro-active learning)

• Head-End Replication to relax the requirement for Multicast in the Underlay

*Multicast Independence requires the usage of the Overlay Control-Plane or static configuration
### Head-End Replication

**Multicast Independent**

<table>
<thead>
<tr>
<th>Peer</th>
<th>VNI</th>
<th>VTEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>V2</td>
<td>30000</td>
<td>V2</td>
</tr>
<tr>
<td>V3</td>
<td>30000</td>
<td>V3</td>
</tr>
<tr>
<td>V3</td>
<td>30001</td>
<td>V3</td>
</tr>
</tbody>
</table>

**Underlay**

<table>
<thead>
<tr>
<th>Peer</th>
<th>VNI</th>
<th>VTEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>30000</td>
<td>V1</td>
</tr>
<tr>
<td>V3</td>
<td>30000</td>
<td>V3</td>
</tr>
<tr>
<td>V3</td>
<td>30001</td>
<td>V3</td>
</tr>
</tbody>
</table>

**Overlay**

1. ARP Request for IP_B
   - **Src MAC**: MAC_A
   - **DST MAC**: FF:FF:FF:FF:FF:FF

2. ARP Request
   - **SMAC**: MAC_V1
   - **DMAC**: hop-by-hop
   - **SIP**: IP_V1
   - **DIP**: IP_V2

3. **VXLAN VNID**: 30000

4. **SMAC**: MAC_A
   - **DMAC**: FF:FF:FF:FF:FF:FF

5. ARP Request for IP_B
   - **Src MAC**: MAC_A
   - **DST MAC**: FF:FF:FF:FF:FF:FF

**Host A**
- **MAC_A** / **IP_A**

**Host B**
- **MAC_B** / **IP_B**

**Host C**
- **MAC_C** / **IP_C**

© 2015 Cisco and/or its affiliates. All rights reserved.
Polling Question 2

Is the thought of using Layer4 BGP protocol for DC switching a scary one? 😊

Yes

No (I’m BGP Expert)
VXLAN Evolution

Protocol Learning

• Workload MAC and IP Addresses learnt by VXLAN Edge Devices (NVEs)
• Advertises Layer-2 and Layer-3 Address-to-VTEP Association (Overlay Control-Plane)
• Flood Prevention
• Optimized ARP forwarding

• Multi-Protocol BGP (MP-BGP) based Control-Plane using EVPN NLRI (Network Layer Reachability Information)
• Make Forwarding decisions at VTEPs for Layer-2 (MAC) and Layer-3 (IP); Integrated Route/Bridge (IRB)
• Reduce Flooding
• Reduce impact of ARP on the Network
• Standards Based (IETF draft)
BGP-EVPN VXLAN

- **Tunnel Endpoints Location**
- **Host Reachability Information**
  - Mac Address
  - IP address

Use Multi-Protocol BGP with EVPN Address family for:

IBGP Route Reflector* (on spine or different box)
EVPN – Ethernet VPN

VXLAN Evolution

Control-Plane

- EVPN MP-BGP
draft-ietf-l2vpn-evpn

Data-Plane

- Multi-Protocol Label Switching (MPLS)
draft-ietf-l2vpn-evpn
- Provider Backbone Bridges (PBB)
draft-ietf-l2vpn-pbb-evpn
- Network Virtualization Overlay (NVO)
draft-sd-l2vpn-evpn-overlay

- EVPN over NVO Tunnels (VXLAN, NVGRE, MPLSoE) for Data Center Fabric encapsulations
- Provides Layer-2 and Layer-3 Overlays over simple IP Networks
EVPN Solution Advantages

- Early ARP Termination
  Suppresses flooding for Unknown Unicast ARP

- Security
  Authenticate Tunnel Endpoints

- Distributed Anycast Gateway
  Seamless and Optimal vm-mobility

- Ingress Replication
  Unicast Alternative to Multicast underlay

- Active/Active Multipathing
  Active/Active and Resilient Multipathing using vPC on Nexus
Host and Subnet Route Distribution

**VXLAN/EVPN**

- Host Route Distribution decoupled from the Underlay protocol
- Use MultiProtocol-BGP (MP-BGP) on the Leaf nodes to distribute internal Host/Subnet Routes and external reachability information
- Route-Reflectors deployed for scaling purposes
Protocol Learning & Distribution (1)

VXLAN/EVPN

VTEPs advertise Host Routes (IP+MAC) for the Host within the Control-Plane
BGP propagates routes for the Host to all other VTEPs.

VTEPs obtain host routes for remote hosts and install in RIB/FIB.
Host Advertisement

**VXLAN/EVPN**

1. Host Attaches
2. VTEP V1 advertises Host A MAC (+IP) through BGP RR
3. Choice of Encapsulation is also advertised
Host Moves

VXLAN/EVPN

1. Host Moves to V3
2. V3 detects Host A and advertises it with Seq #1
3. V1 sees more recent route and withdraws its advertisement

<table>
<thead>
<tr>
<th>MAC, IP</th>
<th>VNI (L2)</th>
<th>VNI (L3)</th>
<th>NH</th>
<th>Encap</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC_A, IP_A</td>
<td>30000</td>
<td>50000</td>
<td>IP_V3</td>
<td>3::VXLAN</td>
<td>1</td>
</tr>
</tbody>
</table>

V1# sh bgp l2vpn evpn IP_A
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 30000:V3
BGP routing table entry for [2]:[0]:[0]:[48]:[MAC_A]:[32]:[IP_A]/272, version 28839
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
AS-Path: NONE, path sourced internal to AS
IP_V3 (metric 3) from RR (RR)
- Origin IGP, MED not set, localpref 100, weight 0
- Received label 30000 50000
- Extcommunity: RT:1000:30000 RT:1000:50000 ENCAP:3
- Originator: IP_V3
- Cluster list: RR
- Remote Next-hop Attribute: IP_V3
- encapsulation VXLAN VNID 50000 MAC MAC_V3

© 2015 Cisco and/or its affiliates. All rights reserved.
ARP Suppression

VXLAN/EVPN

ARP Request sent for IP_B sent from Host A

V1 knows about IP_B and can respond.

No need for ARP forwarding across the Network
ARP Handling on Lookup “Miss” (1)

VXLAN/EVPN

1. ARP Request sent for IP_B sent from Host A

2. Miss of IP_B. Forward ARP Request to all Ports except source-port (ARP snooping)
ARP Handling on Lookup “Miss” (2)

VXLAN/EVPN

<table>
<thead>
<tr>
<th>MAC, IP</th>
<th>VNI</th>
<th>NH</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC_C, IP_C</td>
<td>30000</td>
<td>IP_V3</td>
</tr>
<tr>
<td>MAC_Y, IP_Y</td>
<td>30001</td>
<td>IP_V3</td>
</tr>
<tr>
<td>MAC_B, IP_B</td>
<td>30000</td>
<td>IP_V2</td>
</tr>
</tbody>
</table>

3. ARP Response is sent to V2

V2 will populate this information in the control-plane (learn) and forward it subsequently

4. ARP Response from IP_B
Src MAC: MAC_B
Dst MAC: MAC_A

<table>
<thead>
<tr>
<th>MAC, IP</th>
<th>VNI</th>
<th>NH</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC_A, IP_A</td>
<td>30000</td>
<td>IP_V1</td>
</tr>
<tr>
<td>MAC_C, IP_C</td>
<td>30000</td>
<td>IP_V3</td>
</tr>
<tr>
<td>MAC_Y, IP_Y</td>
<td>30001</td>
<td>IP_V3</td>
</tr>
</tbody>
</table>

ARP Response for IP_B
Src MAC: MAC_B
Dst MAC: MAC_A

<table>
<thead>
<tr>
<th>MAC, IP</th>
<th>VNI</th>
<th>NH</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC_A, IP_A</td>
<td>30000</td>
<td>IP_V1</td>
</tr>
<tr>
<td>MAC_B, IP_B</td>
<td>30000</td>
<td>IP_V2</td>
</tr>
<tr>
<td>MAC_Y, IP_Y</td>
<td>30001</td>
<td>IP_V3</td>
</tr>
</tbody>
</table>

Host A
MAC_A / IP_A

Host B
MAC_B / IP_B

Host C
MAC_C / IP_C

Host Y
MAC_Y / IP_Y
Packet Forwarding (Bridge)

VXLAN/EVPN

<table>
<thead>
<tr>
<th>MAC, IP</th>
<th>VNI</th>
<th>NH</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC_A, IP_A</td>
<td>30000</td>
<td>Local</td>
</tr>
<tr>
<td>MAC_B, IP_B</td>
<td>30000</td>
<td>IP_V2</td>
</tr>
</tbody>
</table>

SMAC: MAC_A
DMAC: MAC_B
SIP: IP_A
DIP: IP_B

SMAC: MAC_B, IP_B
DMAC: MAC_A
SIP: IP_V1
DIP: IP_V2

VXLAN VNID: 30000
UDP

© 2015 Cisco and/or its affiliates. All rights reserved.
Packet Forwarding (Route)

**VXLAN/EVPN**

**MAC, IP | VNI | NH | VRF**

| MAC_A, IP_A | 30000 | Local | 50000 |
| MAC_F, IP_F | 30005 | IP_V2 | 50000 |

**MAC, IP | VNI | NH | VRF**

| MAC_A, IP_A | 30000 | Local | 50000 |
| MAC_F, IP_F | 30005 | E1/4 | 50000 |

Host A  
MAC_A / IP_A

SMAC: MAC_A  
DMAC: MAC_GW

SIP: IP_A  
DIP: IP_F

---

Host F  
MAC_F, IP_F

SMAC: MAC_GW  
DMAC: MAC_F

SIP: IP_A  
DIP: IP_F

---

UDP

**SMAC: MAC_V1**  
**DMAC: hop-by-hop**

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_GW  
DMAC: MAC_V2

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_A  
DMAC: MAC_V2

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_GW  
DMAC: MAC_A

SIP: IP_A  
DIP: IP_F

---

SMAC: hop-by-hop  
DMAC: MAC_V2

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_V1  
DMAC: hop-by-hop

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_A  
DMAC: MAC_GW

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_GW  
DMAC: MAC_F

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_GW  
DMAC: MAC_A

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_V1  
DMAC: hop-by-hop

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_GW  
DMAC: MAC_V2

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_A  
DMAC: MAC_V2

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_GW  
DMAC: MAC_A

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_V1  
DMAC: hop-by-hop

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_GW  
DMAC: MAC_V2

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_A  
DMAC: MAC_V2

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_GW  
DMAC: MAC_A

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_V1  
DMAC: hop-by-hop

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_GW  
DMAC: MAC_V2

SIP: IP_V1  
DIP: IP_V2

VXLAN VNID: 50000

---

SMAC: MAC_A  
DMAC: MAC_V2

SIP: IP_A  
DIP: IP_F

---

SMAC: MAC_GW  
DMAC: MAC_A

SIP: IP_A  
DIP: IP_F

---
EVPN Control Plane Advantages

A multi-tenant fabric solution with host-based forwarding

• Industry standard protocol for multi-vendor interoperability
• Build-in multi-tenancy support
  • Leverage MP-BGP to deliver VXLAN with L3VPN characteristics
• Truly scalable with protocol-driven learning
  • Host MAC/IP address advertisement through EVPN MP-BGP
• Fast convergence upon host movements or network failures
  • MP-BGP protocol driven re-learning and convergence
  • Upon host movement, the new VTEP will send out a BGP update to advertise
    the new location of the host
EVPN Control Plane Advantages (Cont’ed)

A multi-tenant fabric solution with host-based forwarding

- Optimal traffic forwarding supporting host mobility
  - Anycast IP gateway for optimal forwarding for host generated traffic
  - No need for hair-pinning to reach the IP gateway
- ARP suppression
  - Minimize ARP flooding in overlay
- Head-end Replication with dynamically learned remote-VTEP list
  - Head-end replication enables multicast-free underlay network
  - Dynamically learned remote-VTEP list minimizes the operational overhead of head-end replication
- VTEP peer authentication via MP-BGP authentication
  - Added security to prevent rogue VTEPs or VTEP spoofing
VXLAN Evolution

IP Services

- VXLAN Routing
- Distributed Anycast Gateway (requires Overlay Control-Plane)
- Multi-Tenancy

- Forward based on MAC or IP address learnt via Control-Plane (MP-BGP EVPN)
- Make routing decisions at VTEPs
- Scale and Multipathing (ECMP)
- Leverage Layer-3 Gateway capabilities along with Protocol Information
- LISP-ish / LISP-like approach for Host/IP Mobility
  - Location (VTEP), Identifier (MAC, IP of End-Host)
Distributed Gateway Function in L3 Overlays

Traditional L2 - centralised L2/L3 boundary
- Always bridge, route only at an aggregation point
- Large amounts of state converge
- Scale problem for large# of L2 segments
- Traditional L2 and L2 overlays

L2/L3 fabric (or overlay)
- Always route (at the leaves), bridge when necessary
- Distribute and disaggregate necessary state
- Optimal scalability
- Enhanced forwarding and L3 overlays
Distributed IP Anycast Gateway

The same “Anycast” SVI IP/MAC is used at all VTEPs/ToRs.
A host will always find its SVI anywhere it moves.

SVI IP Address
MAC: 0000.dead.beef
IP: 10.1.1.1

SVI IP Address
MAC: 0000.dead.beef
IP: 10.1.2.1
Distributed IP Anycast Gateway

Detailed View

Consistent Anycast SVI IP / MAC address at all leaves
VLAN-IDs are locally significant
VLANs are stretched over L2 VNIs

- VLANs (VLAN A) mapped to VNI (VNI A) at each VTEP: VLAN A' ↔ VNI A ↔ VLAN A
- Bridged traffic forwarded over the L2 VNIs
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Same Subnet aka Bridging (ARP)

1. PM1 sends an ARP request for Default Gateway –10.10.10.1
2. The ARP request is suppressed at TOR and punt to the Supervisor, where MAC and IP is learned and distributed
3. TOR response with Gateway MAC to PM1

Standard behavior of End-Host (virtual or physical) to ARP for the Default Gateway
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Same Subnet aka Bridging (ARP)

4. VM1 sends an ARP request for PM1 – 10.10.10.20

5. The ARP request is suppressed at TOR and punted to the Supervisor, where MAC and IP is learned and distributed

6. Assuming PM1 is known and a valid route does exist in the Unicast RIB, TOR responds to ARP with PM1 MAC as Source MAC. VM1 can build its ARP cache
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Same Subnet aka Bridging (Data Packet)

7. VM1 generates a data packet with PM1_MAC as destination MAC
8. TOR receives the packet and performs Layer-2 lookup for the destination
9. TOR adds VXLAN-Header information (Destination VTEP, VNI, etc) and forwards the packet across the Layer-3 fabric, picking one of the equal cost paths available via the multiple Spines
10. The destination TOR receives the packet, strips off the VXLAN header and performs lookup and forwarding toward PM1

In case of VM1 is not known to PM1, PM1 would ARP for VM1. Destination TOR would Proxy for VM1. No Silent-Host discovery problem.
A common VNI (VNI X) is provisioned amongst the different VTEPs to carry routed traffic.

Routed traffic between VTEPs will be encapsulated in VNI X.

Standard longest prefix match routing takes place:
- Host routes for all known remote hosts are installed at every VTEP ➔ Forward over VNI X.
- Local hosts are covered by directly connected prefix, a host route will not be present.
Distributed IP Anycast Gateway
Packet-Walk – IP Forwarding within the Different Subnet aka Routing (ARP)

1. VM1 sends ARP request for Default Gateway – 10.10.10.1

2. The ARP request will be received at TOR and punt to the Supervisor, where MAC and IP is learned and distributed

3. TOR acts as regular Default Gateway and sends ARP response with GW_MAC to VM1
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Different Subnet aka Routing (Data Packet)

4. VM1 generates a data packet destined to PM2 IP (20.20.20.20) with GW_MAC as destination MAC

5. TOR receives the packet and performs Layer-3 lookup for the destination (known)

6. TOR adds VXLAN-Header information (Destination VTEP, VNI, etc) and forwards the packet across the Layer-3 fabric, picking one of the equal cost paths available via the multiple Spines

7. The destination TOR receives the packet, strips off the VXLAN header and performs lookup and forwarding toward PM2
VXLAN Evolution

Multicast Independent
- Head-end replication enables unicast-only mode
- Control Plane provides dynamic VTEP discovery

Protocol Learning prevents floods
- Workload MAC addresses learnt by VXLAN NVEs
- Advertise L2/L3 address-to-VTEP association information in a protocol

External Connectivity
- VXLAN HW Gateways to other encaps/networks
- VXLAN HW Gateway redundancy
- Enable hybrid overlays

IP Services
- VXLAN Routing
- Distributed IP Gateways
VXLAN Designs
VXLAN Design Considerations

**VXLAN Mode:**
- Flood-and-Learn
- With EVPN control Plane

**BUM Traffic Handling:**
- Multicast replication
- Unicast/ingress replication

**Deployment Scenarios:**
- Brown field vs green field
- Investment protection
- Multi-vendor environment?

**Scalability:**
- The number of VXLAN VNIs
- The number of VTEP peers
- The number of EVPN tenants
- The number of VXLAN Host IP routes
- The number of VXLAN Host MAC addresses
- The number of IPv4/IPv6 LPM routes
- The number of Ingress replication peers
VXLAN Inter-PoD Extension
Brownfield: Connecting Two Data Center PODs

Pod 1
IP GW
Layer-2 VLAN Domain

L3 Core
VTEP

VXLAN Overlay (VLAN Extension)

Pod 2
IP GW
Layer-2 VLAN Domain

L2 Link
L3 Link
VXLAN in 3-Tier Network

**Brownfield**: Cross Layer 3 Boundaries

- **DC Core**
- **DC Aggregation**
- **DC Access**
- **VXLAN Overlay**
- **VTEP**
- **L2 Link**
- **L3 Link**
VXLAN Fabric Design with BGP EVPN

**Greenfield:** Multi-Tenancy with Mobility Support

- Tunnel Endpoints are on leaf layer
- Spine nodes are iBGP Route Reflectors
- Supports Multi-tenancy with seamless Host Mobility
Integrate VXLAN with WAN

- For Disaster Recovery, High Availability
- Integrate EVPN/VXLAN to MPLS-L3VPN or LISP
Data Center Interconnectivity with VXLAN EVPN (Option A)
DCI with VXLAN EVPN (Option B)

One EVPN Administrative Domain
Stretched Across Two Data Centers

Inter-DC EVPN eBGP (multi-hop)
## VXLAN: Flood-\&-Learn vs EVPN Control Plane

<table>
<thead>
<tr>
<th></th>
<th>Flood-&amp;-Learn</th>
<th>EVPN Control Plane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlay Services</td>
<td>L2+L3</td>
<td>L2+L3</td>
</tr>
<tr>
<td>Underlay Network</td>
<td>IP network with ECMP</td>
<td>IP network with ECMP</td>
</tr>
<tr>
<td>Encapsulation</td>
<td>MAC in UDP</td>
<td>MAC in UDP</td>
</tr>
<tr>
<td>Peer Discovery</td>
<td>Data-driven flood-&amp;-learn</td>
<td>MP-BGP</td>
</tr>
<tr>
<td>Peer Authentication</td>
<td>Not available</td>
<td>MP-BGP</td>
</tr>
<tr>
<td>Host Route Learning</td>
<td>Local hosts: Data-driven flood-&amp;-learn</td>
<td>Local Host: Data-driven</td>
</tr>
<tr>
<td></td>
<td>Remote hosts: Data-driven flood-&amp;-learn</td>
<td>Remote host: MP-BGP</td>
</tr>
<tr>
<td>Host Route Distribution</td>
<td>No route distribution.</td>
<td>MP-BGP</td>
</tr>
<tr>
<td>L2/L3 Unicast Forwarding</td>
<td>Unicast encap</td>
<td>Unicast encap</td>
</tr>
<tr>
<td>BUM Traffic forwarding</td>
<td>Multicast replication</td>
<td>Multicast replication</td>
</tr>
<tr>
<td></td>
<td>Unicast/Ingress replication</td>
<td>Unicast/Ingress replication</td>
</tr>
</tbody>
</table>
Polling Question 3

Since VXLAN w/BGP-EVPN is standard based, is multi-vendor integration a possibility?

Yes
No
VXLAN/EVPN - Interoperability & Feasibility

- VXLAN/EVPN interoperability demonstrated during MPLS/SDN World Congress in Paris
- Participating Vendors are Cisco, Juniper, Alcatel Lucent & Ixia

 Independently Tested at EANTC with public available Whitepaper

http://www.eantc.de/showcases/mpls_sdn_2015/intro.html
Resources
Resources

- **VXLAN Overview: Cisco Nexus 9000 Series Switches**

- **VXLAN Network with MP-BGP EVPN Control Plane**

- **Fundamentals of VXLAN**

- **Digging Deeper into VXLAN, Part 1**
  http://blogs.cisco.com/datacenter/digging-deeper-into-vxlan

- **Virtual Extensible LAN (VXLAN) Best Practices**
Submit Your Questions Now!

Use the Q & A panel to submit your questions and our expert will respond
Ask the Expert Event following the Webcast

Now through October 30th


Join the discussion for these Ask The Expert Events:
Collaborate within our Social Media

Learn About Upcoming Events

LinkedIn http://bit.ly/csc-linked-in

Cisco has support communities in other languages!

If you speak Spanish, Portuguese, Japanese, Russian or Chinese we invite you to participate and collaborate in your language.

Spanish
https://supportforums.cisco.com/community/spanish

Portuguese
https://supportforums.cisco.com/community/portuguese

Japanese
https://supportforums.cisco.com/community/csc-japan

Russian
https://supportforums.cisco.com/community/russian

Chinese
http://www.csc-china.com.cn
More IT Training Videos and Technical Seminars on the Cisco Learning Network

View Upcoming Sessions Schedule
https://cisco.com/go/techseminars
Please take a moment to complete the survey

Thank you for Your Time!
Thank you for participating!

- Redeem your 35% discount offer by entering code: **CSC** when checking out:

Visit Cisco Press at:

TOMORROW starts here.