Building Data Center Networks with VXLAN EVPN Overlays – Part II

Lukas Krattiger, Principal Engineer
Questions?
Use Cisco Spark to communicate with the speaker after the session

How
1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install Spark or go directly to the space
4. Enter messages/questions in the space

cs.co/ciscolivebot#BRKDCT-3378
Session Objective

- A short Overview on **Overlays**
- **Standards and Implementation** on VXLAN BGP EVPN
- A short repeat on **Control- & Data-Plane**
- Details around **Tenant Routed Multicast (TRM)**
- Overview and Details around EVPN **Multi-Site**
- **VXLAN OAM** – Operation, Administration and Management
Agenda

• Introduction to Overlays
• VXLAN with BGP EVPN
  • Standards and Implementation
  • Control & Data Plane
• Tenant Routed Multicast (TRM)
• Multi-Site
• VXLAN OAM
Introduction to Overlays
Overlay Taxonomy - Underlay

Layer-3 Interface Peering

Edge Device

LAN Segment

Virtual Server

Physical Server
Overlay Taxonomy - Overlay

VTEP: VXLAN Tunnel End-Point
VNI/VNID: VXLAN Network Identifier
Overlay Taxonomy - Overlay

Tunnel Encapsulation (VNI Namespace)

VTEP: VXLAN Tunnel End-Point
VNI/VNID: VXLAN Network Identifier
Understanding Overlay Technologies

**Overlay Services**
- Layer-2
- Layer-3
- Layer-2 and Layer-3

**Control-Plane**
- Peer-Discovery
- Route Learning and Distribution
  - Local Learning
  - Remote Learning

**Tunnel Encapsulation**

**Underlay Transport Network**

**Data-Plane**
- Overlay Layer-2/Layer-3 Unicast Traffic
- Overlay Broadcast, Unknown Unicast, Multicast traffic (BUM traffic) forwarding
  - Ingress Replication (Unicast)
  - Multicast
Agenda

- Introduction to Overlays
- VXLAN with BGP EVPN
  - Standards and Implementation
  - Control & Data Plane
- Tenant Routed Multicast (TRM)
- Multi-Site
- VXLAN OAM
Standards and Implementation
What is … ?

- **VXLAN**
  - Standards based Encapsulation
    - RFC 7348
    - Uses UDP-Encapsulation
  - Transport Independent
    - Layer-3 Transport (Underlay)
  - Flexible Namespace
    - 24-bit field (VNID) provides ~16M unique identifier
    - Allows Segmentations

- **EVPN**
  - Standards based Control-Plane
    - RFC 7432
    - Uses Multiprotocol BGP
  - Uses Various Data-Planes
    - VXLAN (EVPN-Overlay), MPLS, Provider Backbone (PBB)
  - Many Use-Cases Covered
    - Bridging, MAC Mobility, First-Hop & Prefix Routing, Multi-Tenancy (VPN)
Introducing Ethernet VPN (EVPN)

- EVPN MP-BGP – RFC 7432
- MPLS (draft-ietf-l2vpn-evpn)
- Provider Backbone Bridges (draft-ietf-l2vpn-pbb-evpn)
- Overlay (NVO3) (draft-ietf-bess-evpn-overlay)

- EVPN over NVO Tunnels (i.e. VXLAN) for Data Center Fabric Encapsulation
- Provides Layer-2 and Layer-3 Overlay Service over simple IP Network
## VXLAN and EVPN related RFCs & Drafts (IETF)

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFC 7348</td>
<td>Virtual Extensible Local Area Network</td>
<td>Data Plane</td>
</tr>
<tr>
<td>RFC 7432</td>
<td>BGP MPLS based Ethernet VPNs</td>
<td>Control Plane</td>
</tr>
<tr>
<td>draft-ietf-bess-evpn-overlay</td>
<td>A Network Virtualization Overlay Solution using EVPN</td>
<td>Control Plane</td>
</tr>
<tr>
<td>draft-ietf-bess-evpn-inter-subnet-forwarding</td>
<td>Integrated Routing and Bridging in EVPN</td>
<td>Control Plane</td>
</tr>
<tr>
<td>draft-ietf-bess-l2vpn-evpn-prefix-advertisement</td>
<td>IP Prefix Advertisement in E-VPN</td>
<td>Control Plane</td>
</tr>
<tr>
<td>draft-tissa-nvo3-oam-fm</td>
<td>NVO3 Fault Management / OAM</td>
<td>Management Plane</td>
</tr>
</tbody>
</table>
Different Type of EVPN Use-Cases

- Layer-2
  - Asymmetric IRB (VLAN-Aware)
- Layer-2 & Layer-3
  - Symmetric IRB (VLAN-Based) VRF to VRF
Different Type of EVPN Use-Cases

- Layer-2
- Layer-2 & Layer-3

- Asymmetric IRB (VLAN-Aware)
- Symmetric IRB (VLAN-Based) VRF-to-VRF
• Symmetric Inter-Subnet Forwarding
  • Bridge->Route/Route->Bridge
  • Symmetric VNI in both directions
  • Adjacency contains Remote VTEP,VRF
  • Optimal for Scale
  • Flexible Configuration

VTEP = VXLAN Tunnel End-Point
VRF = Virtual Routing and Forwarding
VNI = VXLAN Network Identifier
Integrated Routing and Bridging in EVPN

- **Asymmetric Inter-Subnet Forwarding**
  - Bridge->Route->Bridge
  - Different (Asymmetric) VNI depending on directions
  - Adjacency contains Remote VTEP,VRF and End-Points
  - Potential Sub-Optimal for Scale
  - Consistent Configuration

VTEP = VXLAN Tunnel End-Point
VRF = Virtual Routing and Forwarding
VNI = VXLAN Network Identifier
Integrated Routing and Bridging in EVPN

- **Symmetric Inter-Subnet Forwarding**
  - Bridge->Route/Route->Bridge
  - Symmetric VNI in both directions
  - Adjacency contains Remote VTEP,VRF
  - Optimal for Scale
  - Flexible Configuration

- **Asymmetric Inter-Subnet Forwarding**
  - Bridge->Route->Bridge
  - Different (Asymmetric) VNI depending on directions
  - Adjacency contains Remote VTEP,VRF and End-Points
  - Potential Sub-Optimal for Scale
  - Consistent Configuration

VTEP = VXLAN Tunnel End-Point
VRF = Virtual Routing and Forwarding
VNI = VXLAN Network Identifier
Operational Models for Asymmetric Inter-Subnet Forwarding
(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 4)

• Asymmetric IRB

Bridge -> Route -> Bridge
Operational Models for Asymmetric Inter-Subnet Forwarding
ARP and Adjacency Table

- Asymmetric IRB

**ARP table V1**
- a.a.a, 192.168.22.33, VLAN 30
- b.b.b, 192.168.33.44, VNI 40000

**ARP table V2**
- b.b.b, 192.168.33.44, VLAN 40
- a.a.a, 192.168.22.33, VNI 30000
Operational Models for Asymmetric Inter-Subnet Forwarding
Routing Table

- Asymmetric IRB

Routing table V1
192.168.22.33, local, VLAN 30
192.168.33.44, local, VNI 40000

Routing table V2
192.168.33.44, local, VLAN 40
192.168.22.33, local, VNI 30000
Operational Models for Asymmetric Inter-Subnet Forwarding
(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 4.1)

- Asymmetric IRB

![Diagram showing operational models for asymmetric inter-subnet forwarding]
4.1 Among EVPN NVEs within a DC

When an EVPN MAC/IP advertisement route is received by an NVE, the IP address associated with the route is used to populate the IP-VRF table, whereas the MAC address associated with the route is used to populate both the MAC-VRF table, as well as the adjacency associated with the IP route in the IP-VRF table (i.e., ARF table).

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated MAC-VRF for that EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup identifies an adjacency that contains a MAC rewrite and in turn the next-hop (i.e., egress) NVE to which the packet must be forwarded and the associated MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the EVPN MAC route), instead of the MAC address of the next-hop NVE. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE then forwards the frame to the next-hop (i.e., egress) NVE after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the EVPN label that was advertised by the egress NVE. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVPN label to identify the MAC-VRF table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 2 below depicts the packet flow, where NVE1 and NVE2 are the ingress and egress NVEs, respectively.

![Diagram](image-url)

Figure 2: Inter-Subnet Forwarding Among EVPN NVEs within a DC

Note that the forwarding behavior on the egress NVE is similar to EVPN intra-subnet forwarding. In other words, all the packet processing associated with the inter-subnet forwarding semantics is confined to the ingress NVE and that is why it is called Asymmetric IRB.
Operational Models for Symmetric Inter-Subnet Forwarding
(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 5)

• Symmetric IRB

Bridge -> Route -> Route -> Bridge
Operational Models for Symmetric Inter-Subnet Forwarding
ARP and Adjacency Table

- Symmetric IRB

![Diagram showing symmetric IRB and ARP tables]

**ARP table V1**
- a.a.a, 192.168.22.33, VLAN 30

**ARP table V2**
- b.b.b, 192.168.33.44, VLAN 40
Operational Models for Symmetric Inter-Subnet Forwarding
Routing Table

- Symmetric IRB

Routing table V1
192.168.22.33, V2, VNI 50000

Routing table V2
192.168.33.44, V1, VNI 50000
Operational Models for Symmetric Inter-Subnet Forwarding
(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 5.1.1)

- Symmetric IRB

![Diagram showing Symmetric IRB with IP addresses and MAC addresses]

- 192.168.22.33
- 192.168.33.44

VNI 50000 (L3VNI)

10.22.22.34 (Next-Hop)

L2VNI: 30000
L3VNI: 50000

BGP Update

[2]:[0]:[0]:[48]:[0050.569f.d495]:[32]:[192.168.33.44]
Operational Models for Symmetric Inter-Subnet Forwarding
(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 5.1.1)

- Symmetric IRB

5.1.1 Control Plane Operation

Each NVE advertises a Route Type-2 (RT-2, MAC/IP Advertisement Route) for each of its TS’s with the following field set:

- RD and ESI per [EVPN]
- Ethernet Tag = 0; assuming VLAN-based service
- MAC Address Length = 48
- MAC Address = Mi ; where i = 1,2,3,4, or 5 in the above example
- IP Address Length = 32 or 128
- IP Address = iPi ; where i = 1,2,3,4, or 5 in the above example
- Label-1 = MPLS Label or VNID corresponding to MAC-VRF
- Label-2 = MPLS Label or VNID corresponding to IP-VRF

Each NVE advertises an RT-2 route with two Route Targets (one corresponding to its MAC-VRF and the other corresponding to its IP-VRF). Furthermore, the RT-2 is advertised with two BGP Extended Communities. The first BGP Extended Community identifies the tunnel type per section 4.5 of [TUNNEL-ENCAP] and the second BGP Extended Community includes the MAC address of the NVE (e.g., MACx for NVE1 or MACy for NVE2) as defined in section 6.1. This second Extended Community (for the MAC address of NVE) is only required when Ethernet NVO tunnel type is used. If IP NVO tunnel type is used, then there is no need to send this second Extended Community.
Different Type of EVPN Use-Cases

Layer-2

Layer-2 & Layer-3

Asymmetric IRB (VLAN-Aware)

Symmetric IRB (VLAN-Based) VRF to VRF
EVPN Layer-2 Service Interface

- Single Subnet per EVI
  - VLAN-based

  - Per EVI BGP Route Distinguisher / Router Target per EVI / VNI
    - BGP Route-Target constrain mechanism to limit propagation (import/export)

- 1:1 mapping
  - EVI to Single Broadcast Domain (Bridge Domain)
  - Ethernet Tag ID must be 0

VID = VLAN ID
VNI = VXLAN Network Identifier
EVI = EVPN Virtual Instance
EVPN Layer-2 Service Interface

- Single Subnet per EVI
  - VLAN-based
  - Per EVI BGP Route Distinguisher / Router Target per EVI / VNI
    - BGP Route-Target constrain mechanism to limit propagation (import/export)
  - 1:1 mapping
    - EVI to Single Broadcast Domain (Bridge Domain)
    - Ethernet Tag ID must be 0

- Multiple Subnets per EVI
  - VLAN-aware
  - 1:N mapping
    - EVI to Multiple Broadcast Domains (Bridge Domains)
    - Ethernet Tag ID is to differentiate Bridge Domains

VID = VLAN ID
VNI = VXLAN Network Identifier
EVI = EVPN Virtual Instance
Virtual Identifiers to EVI Mapping
(draft-ietf-bess-evpn-overlay – Section 5.1.2)

- VLAN-based

Route Target: 65000:30000

VID = VLAN ID
VNI = VXLAN Network Identifier
EVI = EVPN Virtual Instance
Virtual Identifiers to EVI Mapping (draft-ietf-bess-evpn-overlay – Section 5.1.2)

• VLAN-based

5.1.2 Virtual Identifiers to EVI Mapping

When the EVPN control plane is used in conjunction with VXLAN (or NVGRE encapsulation), just like [RFC7432] where two options existed for mapping broadcast domains (represented by VLAN IDs) to an EVI, here there are also two options for mapping broadcast domains represented by VXLAN VNIs (or NVGRE VSIDs) to an EVI:

1. Option 1: Single Broadcast Domain per EVI

In this option, a single Ethernet broadcast domain (e.g., subnet) represented by a VNI is mapped to a unique EVI. This corresponds to the VLAN Based service in [RFC7432], where a tenant-facing interface, logical interface (e.g., represented by a VLAN ID) or physical, gets mapped to an EVPN instance (EVI). As such, a BGP RD and RT are needed per VNI on every NVE. The advantage of this model is that it allows the BGP RT constraint mechanisms to be used in order to limit the propagation and import of routes to only the NVEs that are interested in a given VNI. The disadvantage of this model may be the provisioning overhead if RD and RT are not derived automatically from VNI.
VLAN-Based Service Interface (RFC7432 – Section 6.1)

- VLAN-based

[2]:[0]:[0]:[48]:[0050.569f.d495]:[32]:[192.168.22.33]

6.1. VLAN-Based Service Interface

With this service interface, an EVPN instance consists of only a single broadcast domain (e.g., a single VLAN). Therefore, there is a one-to-one mapping between a VID on this interface and a MAC-VRF. Since a MAC-VRF corresponds to a single VLAN, it consists of a single bridge table corresponding to that VLAN. If the VLAN is represented by multiple VIDs (e.g., a different VID per Ethernet segment per PE), then each PE needs to perform VID translation for frames destined to its Ethernet segment(s). In such scenarios, the Ethernet frames transported over an MPLS/IP network SHOULD remain tagged with the originating VID, and a VID translation MUST be supported in the data path and MUST be performed on the disposition PE. The Ethernet Tag ID in all EVPN routes MUST be set to 0.
VLAN-Aware Bundle Service Interface
(RFC 7432 – Section 6.3)

- VLAN-aware

Route Target: 65000:30000

[2]:[0]:[20]:[48]:[0050.569f.d495]:[32]:[192.168.22.33]
VLAN-Aware Bundle Service Interface

With this service interface, an EVPN instance consists of multiple broadcast domains (e.g., multiple VLANs) with each VLAN having its own bridge table — i.e., multiple bridge tables (one per VLAN) are maintained by a single MAC-VRF corresponding to the EVPN instance.

Broadcast, unknown unicast, or multicast (BUM) traffic is sent only to the CEs in a given broadcast domain; however, the broadcast domains within an EVI either MAY each have their own P-Tunnel or MAY share P-Tunnels — e.g., all of the broadcast domains in an EVI MAY share a single P-Tunnel.

In the case where a single VLAN is represented by a single VID and thus no VID translation is required, an MPLS-encapsulated packet MUST carry that VID. The Ethernet Tag ID in all EVPN routes MUST be set to that VID. The advertising PE MAY advertise the MPLS Label in the MAC/IP Advertisement route representing ONLY the EVI or representing both the Ethernet Tag ID and the EVI. This decision is only a local matter by the advertising PE (which is also the disposition PE) and doesn’t affect any other PEs.

In the case where a single VLAN is represented by different VIDs on different CEs and thus VID translation is required, a normalized Ethernet Tag ID (VID) MUST be carried in the EVPN BGP routes. Furthermore, the advertising PE advertises the MPLS Label in the MAC/IP Advertisement route representing both the Ethernet Tag ID and the EVI, so that upon receiving an MPLS-encapsulated packet, it can identify the corresponding bridge table from the MPLS EVI label and perform Ethernet Tag ID translation ONLY at the disposition PE — i.e., the Ethernet frames transported over the MPLS/IP network MUST remain tagged with the originating VID, and VID translation is performed on the disposition PE. The Ethernet Tag ID in all EVPN routes MUST be set to the normalized Ethernet Tag ID assigned by the EVPN provider.
VLAN-Based / VLAN-Aware Bundle Service Interface

- VLAN-based
- VLAN-aware

[2]:[0]:[0]:[48]:[0050.569f.d495]:[32]:[192.168.22.33]

[2]:[0]:[20]:[48]:[0050.569f.d495]:[32]:[192.168.22.33]
Different Type of EVPN Use-Cases

- Layer-2
  - Asymmetric IRB (VLAN-Aware)

- Layer-2 & Layer-3
  - Symmetric IRB (VLAN-Based)
    - VRF-to-VRF
IP-VRF-to-IP-VRF Model in EVPN

- **Interface-Less Model**
  
  - Route-Type 5 only
    - Next-Hop is remote VTEP
    - Two extended communities
      - Encapsulation Extended Community
      - Router’s MAC Address (remote VTEP)

Route Type 2 = MAC/IP Route
Route Type 5 = IP Prefix Route
IP-VRF-to-IP-VRF Model in EVPN

• Interface-Less Model
  • Route-Type 5 only
    • Next-Hop is remote VTEP
    • Two extended communities
      • Encapsulation Extended Community
      • Router’s MAC Address (remote VTEP)

• Interface-Ful Model (2 Modes)
  • Route-Type 5
    • Next-Hop is remote IRB
    • One or two extended communities
      • Encapsulation Extended Community
      • Router’s MAC Address (remote VTEP)
  • Route-Type 2
    • Containing Router MAC or MAC/IP

Route Type 2 = MAC/IP Route
Route Type 5 = IP Prefix Route
Interface-less IP-VRF-to-IP-VRF Model
(draft-ietf-bess-evpn-prefix-advertisement – Section 4.4.1)

• Interface-Less

EVPN

[5]:[0]:[0]:[24]:[192.168.22.0]:[0.0.0.0]
10.22.22.34 (Next-Hop)
Encap:8 (VXLAN)
Router MAC:0200.0ade.de22

BGP Update

NVE IP: 10.22.22.34
Router MAC: 0200.0ade.de22

VTEP

VTEP
Interface-less IP-VRF-to-IP-V (draft-ietf-bess-evpn-prefix-advertisement)

- Interface-Less

**EVPN**

```
[5]:[0]:[0]:[24]:[192.168.22.0]:[0.0.0.0]  
10.22.22.34 (Next-Hop)  
Encap: 8 (VXLAN)  
Router MAC: 0200.0ade.de22
```

```

<table>
<thead>
<tr>
<th>Internet-Draft</th>
<th>EVPN Prefix Advertisement</th>
<th>November 21, 2017</th>
</tr>
</thead>
</table>

**Figure 6 Interface-less IP-VRF-to-IP-VRF model**

In this case:

a) The NVRs and DGMs must provide connectivity between hosts in SN1, SN2, IP1 and hosts sitting at the other end of the WAN, for example, H1. We assume the DGMs import/export IP and/or VPN-IP routes from/to the WAN.

b) The IP-VRF instances in the NVE/DGMs are directly connected through NVO tunnels, and no IRBs and/or BD instances are instantiated to connect the IP-VRFs.

c) The solution must provide layer-3 connectivity among the IP-VRFs for Ethernet NVO tunnels, for instance, VXLAN or NVGRE.

d) The solution may provide layer-3 connectivity among the IP-VRFs for IP NVO tunnels, for example, VXLAN GPE (with IP payload).

In order to meet the above requirements, the EVPN route type 5 will be used to advertise the IP prefixes, along with the Router’s MAC Extended Community as defined in [EVPN-INTERSUBNET] if the advertising NVE/DGM uses Ethernet NVO tunnels. Each NVE/DGM will advertise an RI-5 for each of its prefixes with the following fields:
Interface-ful IP-VRF-to-IP-VRF with SBD IRB
(draft-ietf-bess-evpn-prefix-advertisement – Section 4.4.2)

- Interface-Ful (Core-facing IRB)
4.4.2 Interface-ful IP-VRF-to-IP-VRF with SBD IRB

Figure 7 will be used for the description of this model.


Internet-Draft EVPN Prefix Advertisement November 21, 2017

![Diagram of Interface-Ful (Core-facing IRB)]

In this model:

a) As in section 4.4.1, the NVEs and DGWs must provide connectivity between hosts in SN1, SN2, IP1 and hosts sitting at the other end of the WAN.

b) However, the NVE/DGWs are now connected through Ethernet NVO tunnels terminated in the SBD instance. The IP-VRFs use IRB interfaces for their connectivity to the SBD.

c) Each SBD IRB has an IP and a MAC address, where the IP address must be reachable from other NVEs or DGWs.

d) The SBD is attached to all the NVE/DGWs in the tenant domain B/Ds.

e) The solution must provide layer-3 connectivity for Ethernet NVO tunnels, for instance, VXLAN or NVGRE.

EVPN type 5 routes will be used to advertise the IP Prefixes, whereas EVPN RT-2 routes will advertise the MAC/IP addresses of each SBD IRB interface. Each NVE/DGW will advertise an RT-5 for each of its prefixes with the following fields:
Interface-ful IP-VRF-to-IP-VRF with Unnumbered SBD IRB
(draft-ietf-bess-evpn-prefix-advertisement – Section 4.4.3)

- Interface-Ful (Unnumbered Core-facing IRB)
4.4.3 Interface-ful IP-VRF to IP-VRF with Unnumbered SBD IRB

Figure 8 will be used for the description of this model. Note that this model is similar to the one described in section 4.4.2, only without IP addresses on the SBD IRB interfaces.

- **Figure 8 Interface-ful with unnumbered SBD IRB model**

In this model:

a) As in section 4.4.1 and 4.4.2, the NVEs and DGWs must provide connectivity between hosts in SN1, SN2, IP1 and hosts sitting at the other end of the WAN.

b) As in section 4.4.2, the NVE/DGWs are connected through Ethernet NVO tunnels terminated in the SBD instance. The IP-VRFs use IRB interfaces for their connectivity to the SBD.

c) However, each SBD IRB has a MAC address only, and no IP address (that is why the model refers to an ‘unnumbered’ SBD IRB). In this model, there is no need to have IP reachability to the SBD IRB interfaces themselves and there is a requirement to save IP addresses on those interfaces.

d) As in section 4.4.2, the SBD is composed of all the NVE/DGW BDs of the tenant that need inter-subnet-forwarding.

e) As in section 4.4.2, the solution must provide layer-3 connectivity for Ethernet NVO tunnels, for instance, VXLAN or nvGRE.

This model will also make use of the RT-5 recursive resolution. EVPN type 5 routes will advertise the IP Prefixes along with the Router’s MAC Extended Community used for the recursive lookup, whereas EVPN RT-2 routes will advertise the MAC addresses of each SBD IRB.
Control- & Data-Plane
Host Advertisements

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0000.3001.1101 / 48</td>
<td>3001, 65500:3001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0000.3001.1102 / 48</td>
<td>3001, 65500:3001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0000.3002.2101 / 48</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Host MAC (Route Type 2)
  - MAC
  - MPLS Label1 (L2VNI*)
  - Route Target for MAC-VRF
- MAC attributes are Mandatory

*L2VNI: VNI for all Bridging operation ("VLAN-VNI")*
V2# show bgp l2vpn evpn

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.10.10.101:32777
BGP routing table entry for [2]:[0]:[0]:[48]:[0000.3001.1101]:[0]:[0.0.0.0]/216, version 4
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is locked

Advertised path-id 1
Route Type: internal, is valid, is best path, no labeled next-hop
AS-Path: NONE, path sourced internal to AS
10.200.200.101 (metric 3) from 10.10.10.201 (10.10.10.201)
Origin IGP, MED not set, localpref 100, weight 0
Received label 3001
Extcommunity: RT:65500:3001 ENCAP:8
Originator: 10.10.10.101 Cluster list: 10.10.10.201
Host Advertisements

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0000.3001.1101 / 48</td>
<td>3001, 65500:3001</td>
<td>192.168.10.101 /32</td>
<td>5000, 65500:5000</td>
</tr>
<tr>
<td>2</td>
<td>0000.3001.1102 / 48</td>
<td>3001, 65500:3001</td>
<td>192.168.10.102 /32</td>
<td>5000, 65500:5000</td>
</tr>
<tr>
<td>2</td>
<td>0000.3002.2101 / 48</td>
<td>3002, 65500:3002</td>
<td>192.168.20.101 /32</td>
<td>5000, 65500:5000</td>
</tr>
</tbody>
</table>

- Host MAC+IP (Route Type 2)
  - MAC and IP
  - MPLS Label1 (L2VNI)
  - Route Target for MAC-VRF
  - MPLS Label2 (L3VNI*)
  - Route Target for IP-VRF
  - Router MAC

- IP Attributes are Optional
- Populated through ARP/ND

*VNI: VNI for all Routing operation ("VRF-VNI")
V2# show bgp l2vpn evpn

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.10.10.101:32777
BGP routing table entry for [2]:[0]:[0]:[48]:[0000.3001.1101]:[32]:[192.168.10.101]/272, version 4
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn,
Precedence: 15, advertised path-id 1
Route Type: MAC/IP
AS Path: NONE, path sourced internal to AS
10.200.200.101 (metric 3) from 10.10.10.201 (10.10.10.201)
Origin IGP, MED not set, localpref 100, weight 0
Received label 3001 5000
Extcommunity: RT:65500:3001 RT:65500:5000 ENCAP:8 Router MAC:0200.0ade.de01
Originator: 10.10.10.201 Cluster list: 10.10.10.201
L2VNI Route Target
L3VNI Route Target
Internal route
Next-Hop IP Address
L2VNI (MPLS Label1)
L3VNI (MPLS Label2)
Encap:8 VXLAN
Encapsulation: 8
Router MAC
IP Address Length
MAC Address
IP Address
MAC Address
MAC Address Length
Ethernet Tag Identifier (Ethtag)
Ethernet Segment Identifier (ESI)
Route Type: MAC/IP
IP Address
L2VNI Route Target
L3VNI Route Target
Subnet Route Advertisements

<table>
<thead>
<tr>
<th>Type</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>192.168.10.0/24</td>
<td>5000, 65500:5000</td>
<td>10.200.200.101</td>
<td></td>
</tr>
</tbody>
</table>

- Internal and External Subnet Prefixes (Route Type 5)
  - IP Prefix
  - MPLS Label (L3VNI)
  - Route Target for IP-VRF
  - Router MAC

- Populated through External Routing Protocol
V2# show bgp l2vpn evpn

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.10.10.1:3
BGP routing table entry for [5]:[0]:[0]:[24]:[192.168.10.101]/224, version 4
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is locked
Path: internal, path is valid, is best path, no labeled nexthop
IP Address Length
IP Address
Next-Hop IP Address
L3VNI (MPLS Label)
L3VNI Route Target
Encap:8 VXLAN
Router MAC
Extcommunity: RT:65500:5000 ENCAP:8 Router MAC:0200.0ade.de01
Originator: 10.10.10.101 Cluster list: 10.10.10.201
Routing and the Router MAC – Ethernet

Router MAC

<table>
<thead>
<tr>
<th>SMAC</th>
<th>DMAC</th>
<th>SIP</th>
<th>DIP</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>0200.0ade.de01</td>
<td>0200.0ade.de07</td>
<td>192.168.10.101</td>
<td>192.168.20.101</td>
<td></td>
</tr>
</tbody>
</table>

Switch

interface: Eth2/1
MAC: 0200.0ade.de01
IP: 10.200.200.1

Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

SVI10
192.168.10.1

Switch

interface: Eth2/1
MAC: 0200.0ade.de07
IP: 10.200.200.7

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

SVI20
192.168.20.1

SMAC 0000.3001.1101
DMAC 2020:0000:AAAA
SIP 192.168.10.101
DIP 192.168.20.101

Payload

SMAC 0000.3002.2101
DMAC 2020:0000:AAAA
SIP 192.168.10.101
DIP 192.168.20.101

Payload
Routing and the Router MAC – VXLAN

Router MAC

<table>
<thead>
<tr>
<th>SMAC</th>
<th>DMAC</th>
<th>SIP</th>
<th>DIP</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000.3001.1101</td>
<td>2020:0000:AAAA</td>
<td>192.168.10.101</td>
<td>192.168.20.101</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SMAC</th>
<th>DMAC</th>
<th>SIP</th>
<th>DIP</th>
<th>Payload</th>
</tr>
</thead>
</table>

Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

SIP: 10.200.200.101
DIP: 10.200.200.107
VXLAN: 5000
SMAC: 0200.0ade.de01
DMAC: 0200.0ade.de07
SIP: 192.168.10.101
DIP: 192.168.20.101

Payload

SMAC: 0200.0ade.de01
DMAC: 0200.0ade.de07

SMAC: 0000.3002.2101
DMAC: 0000.3002.2101

SIP: 192.168.10.101
DIP: 192.168.20.101

SIP: 192.168.10.101
DIP: 192.168.20.101

Payload

SIP: 0000.3001.1101
DIP: 2020:0000:AAAA

SIP: 2020:0000:AAAA
DIP: 0000.3002.2101

Payload

SIP: 0200.0ade.de01
DIP: 0200.0ade.de07

SIP: 0200.0ade.de07
DIP: 0200.0ade.de01

Payload

SIP: 192.168.10.101
DIP: 192.168.20.101

SIP: 192.168.20.101
DIP: 192.168.10.101

Payload

SMAC: 0200.0ade.de01
DMAC: 0200.0ade.de07

SMAC: 0000.3002.2101
DMAC: 0000.3002.2101

SIP: 192.168.10.101
DIP: 192.168.20.101

SIP: 192.168.20.101
DIP: 192.168.10.101

Payload

SIP: 0000.3001.1101
DIP: 2020:0000:AAAA

SIP: 2020:0000:AAAA
DIP: 0000.3002.2101

Payload

SIP: 0200.0ade.de01
DIP: 0200.0ade.de07

SIP: 0200.0ade.de07
DIP: 0200.0ade.de01

Payload

SIP: 192.168.10.101
DIP: 192.168.20.101

SIP: 192.168.20.101
DIP: 192.168.10.101

Payload
Packet Walk – Symmetric IRB (A to C)

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
</table>
Packet Walk – Symmetric IRB (C to A)

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
</table>

**Payload**

- **SMAC**: 2020.0000.AAAA
- **DMAC**: 0000.3001.1101
- **SIP**: 192.168.20.101
- **DIP**: 192.168.10.101

**Host A**
- MAC: 0000.3001.1101
- IP: 192.168.10.101

**Host B**
- MAC: 0000.3001.1102
- IP: 192.168.10.102

**Host C**
- MAC: 0000.3002.2101
- IP: 192.168.20.101
Packet Walk – Asymmetric IRB (A to C)

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
</table>

**Diagram:**

- **SIP**: 10.200.200.101
- **DIP**: 10.200.200.107
- **VXLAN**: 3002
- **SMAC**: 2020.0000.AAAA
- **DMAC**: 0000.3002.2101
- **SIP**: 192.168.10.101
- **DIP**: 192.168.20.101

**Payload:**

- **SMAC**: 0000.3001.1101
- **DMAC**: 2020.0000.AAAA
- **SIP**: 192.168.10.101
- **DIP**: 192.168.20.101

**Host A**

- MAC: 0000.3001.1101
- IP: 192.168.10.101

**Host B**

- MAC: 0000.3001.1102
- IP: 192.168.10.102

**Host C**

- MAC: 0000.3002.2101
- IP: 192.168.20.101
Packet Walk – Asymmetric IRB (C to A)

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
</table>
Agenda

- Introduction to Overlays
- VXLAN with BGP EVPN
  - Standards and Implementation
  - Control & Data Plane
- Tenant Routed Multicast (TRM)
- Multi-Site
- VXLAN OAM
Multicast Forwarding
Tenant Routed Multicast (TRM)
Same Subnet Forwarding **no** IGMP Snooping

Traditional Forwarding in VXLAN Overlays

VXLAN EVPN

<table>
<thead>
<tr>
<th>SRC-10</th>
<th>224.10.10.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10.10.10.100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RCVR-10</th>
<th>10.10.10.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10.10.10.100</td>
</tr>
</tbody>
</table>

Not interested
Same Subnet Forwarding **no** IGMP Snooping

Traditional Forwarding in VXLAN Overlays

- "Single Copy" in Core – Treated as BUM
- Same Subnet Only
- No Pruning on Local Interface or Remote VTEP Interface

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

Baremetal

Not interested

Baremetal

Not interested

Baremetal

Baremetal

VXLAN EVPN

VTEP

Spine

VTEP

VTEP
Same Subnet Forwarding **with** IGMP Snooping

Traditional Forwarding in VXLAN Overlays
Same Subnet Forwarding with IGMP Snooping

Traditional Forwarding in VXLAN Overlays

- "Single Copy" in Core – Treated as BUM
- Same Subnet Only
- Pruning on Local Interface
- VXLAN is "pruned off" if no interest Receiver exists behind any Remote VTEP
Same Subnet Forwarding **with** IGMP Snooping

Traditional Forwarding in VXLAN Overlays

**VXLAN EVPN**

- **SRC-10**
  - 224.10.10.10
  - 10.10.10.100

- **RCVR-10**
  - 10.10.10.10

- **RCVR-14**
  - 10.10.10.14

**Not interested**
Same Subnet Forwarding with IGMP Snooping

Traditional Forwarding in VXLAN Overlays

- "Single Copy" in Core – Treated as BUM
- Same Subnet Only
- Pruning on Local Interface
- VXLAN is NOT pruned if interest Receiver exists behind one Remote VTEP
Different Subnet Forwarding – Router on-a-Stick

Traditional Forwarding in VXLAN Overlays

- SRC-10
  - 224.10.10.10
  - 10.10.10.100

- RCVR-10
  - 10.10.10.10

- RCVR-20
  - 20.20.20.20

- RCVR-14
  - 10.10.10.14

VXLAN EVPN

Spine

Baremetal

RCVR-14
  - 10.10.10.14

10.10.10.254
20.20.20.254
Different Subnet Forwarding – Router on-a-Stick

Traditional Forwarding in VXLAN Overlays

- Multiple Copy in Core – Treated as BUM
- Different Subnet possible – RPF Challenges
- Pruning on Local Interface
- VXLAN is NOT pruned if interest Receiver exists behind one Remote VTEP

 SRC-10 224.10.10.10 10.10.10.100
 RCVR-10 10.10.10.10
 RCVR-20 20.20.20.20 10.10.10.254 20.20.20.254
 RCVR-14 10.10.10.14
Functional Components
Tenant Routed Multicast (TRM)

Functional Components
Tenant Routed Multicast (TRM)

**Underlay:**
- PIM-based Underlay Transport (PIM ASM)
- Separate Multicast Groups from Layer-2 VNI
- Leveraging same redundant Underlay Rendezvous-Point (i.e. PIM Anycast-RP) Single Packet in Core

Functional Components
Tenant Routed Multicast (TRM)

**Underlay:**
- PIM-based Underlay Transport (PIM ASM)
- Separate Multicast Groups from Layer-2 VNI
- Leveraging same redundant Underlay Rendezvous-Point (i.e. PIM Anycast-RP) Single Packet in Core

**Overlay:**
- BGP-based Control-Plane using ngMVPN (Next-Generation Multicast VPN)
- Using existing BGP Route-Reflector
- Rendezvous-Point-less
- Efficient Single Copy in Multicast Underlay
- Always-Route approach (per-VLAN config)
- Distributed Anycast Designated Router (DR)
- VPC – Virtual Port-Channel
- Integration with non-TRM VTEP

Forwarding Behavior
Tenant Routed Multicast (TRM)
Same Subnet Forwarding – Local and Remote Snooping
TRM Forwarding (Layer-2 only mode)
Same Subnet Forwarding – Local and Remote Snooping

TRM Forwarding (Layer-2 only mode)
Same Subnet Forwarding – Local and Remote Snooping

TRM Forwarding (Layer-2 only mode)

- Local IGMP Termination (needs Querier)
- Bridge approach only – in Layer-2 VNI
- "Single Copy" in Core
- Local and Remote IGMP Snooping
- Uses BGP EVPN Route-Type 6 (SMET)
Different and Same Subnet Forwarding
TRM Forwarding (Layer-3 Mode)
Different and Same Subnet Forwarding

TRM Forwarding (Layer-3 Mode)

- Distributed Designated Router (DR)
- Always Route approach – in Layer-3 VNI (VRF)
- "Single Copy" in Core
- Egress replication - closest to the fan-out.
- Single Default MDT (I-PMSI)
Local and Remote Forwarding
TRM Forwarding (Layer-3 Mode)

VXLAN EVPN

TTL Decrement

No TTL Decrement (bridged)

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14

TTL Decrement (routed)
Tenant Routed Multicast routes between different IP Subnets on the same Switch (Leaf). As a result, the TTL is decremented in this routed forwarding operation.

Local to the same Switch (Leaf), Multicast is bridged within the same IP Subnet. During this forwarding operation, TTL is not decremented.

Tenant Routed Multicast uses an always route, single copy approach across the VXLAN EVPN Fabric. This is true if Sources and Receivers reside in the same or in different IP Subnets. As a result, the TTL is decremented during this forwarding operation.

All IP Subnets must be in the same Routing Domain or Tenant (= VRF)
Overlay Rendezvous Point

- RP-less
  - Distributed Anycast RP (NGMVPN-based)
  - Shortest Path Tress (SPT only)
  - Requires per-Tenant Loopback, Multicast enabled

- External RP
  - Centralized RP (PIM-based)
  - Shared Tree and Shortest Path Tree (cut over)
  - Requires External PIM-based RP

VXLAN EVPN

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14

VTEP

DR

Spine
TRM Control- & Data-Plane
Layer-3 VNI: 50001
Default MDT: 239.1.1.2
Route-Target: 65502:50001

*,G – (*, 239.1.1.2/32)
IIF: Uplink (Underlay)
OIF: NVE1 (Underlay)

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14

S,G – (10.0.0.1, 239.1.1.2/32)
IIF: NVE-Loopback (Underlay)
OIF: Uplink (Underlay)

S,G – (10.0.0.2, 239.1.1.2/32)
IIF: NVE-Loopback (Underlay)
OIF: Uplink (Underlay)

S,G – (10.0.0.3, 239.1.1.2/32)
IIF: NVE-Loopback (Underlay)
OIF: Uplink (Underlay)

S,G – (10.0.0.4, 239.1.1.2/32)
IIF: NVE-Loopback (Underlay)
OIF: Uplink (Underlay)
Underlay Multicast Tree – PIM ASM

- PIM ASM required for Underlay
- Separate Groups for BUM and MDT
- Default MDT initiates on VTEP startup (*. G)
- Per-VTEP (S, G) imitated on VTEP startup

Layer-3 VNI: 50001
Default MDT: 239.1.1.2
Route-Target: 65502:50001

*S, G – (*. 239.1.1.2/32)
IIF: Uplink (Underlay)
OIF: NVE1 (Underlay)

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14
NGMVPN – Source Active Advertisement (MVPN Type 5)

Layer-3 VNI: 50001
Default MDT: 239.1.1.2
Route-Target: 65502:50001

*,G – (*, 239.1.1.2/32)
IIF: Uplink (Underlay)
OIF: NVE1 (Underlay)

Source Active (NGMVPN Type 5)
Originator: Leaf #1
Route-Target: 65502:50001
S,G: 10.10.10.100, 224.10.10.10

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14
MRoute – Overlay Multicast Tree (Source Join)

Layer-3 VNI: 50001
Default MDT: 239.1.1.2
Route-Target: 65502:50001

*,G – (*, 239.1.1.2/32)
IIF: Uplink (Underlay)
OIF: NVE1 (Underlay)

S,G – (10.10.10.100, 224.10.10.10)
IIF: VLAN100 (Host-facing)
OIF: None

S,G – (10.10.10.100, 224.10.10.10)
IIF: VRF-L3VNI (Overlay)
OIF: None

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14
NGMVPN – Source Tree Join (MVPN Type 7)

Layer-3 VNI: 50001
Default MDT: 239.1.1.2
Route-Target: 65502:50001

*,G – (*, 239.1.1.2/32)
IIF: Uplink (Underlay)
OIF: NVE1 (Underlay)

Source Tree Join (NGMVPN Type 5)
Originator: Leaf #4
S,G: 10.10.10.100, 224.10.10.10

VRF Tenant1

VXLAN EVPN

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14
MRoute – Overlay Multicast Tree (Receiver Join)

Layer-3 VNI: 50001
Default MDT: 239.1.1.2
Route-Target: 65502:50001

*S,G – (*, 239.1.1.2/32)
IIF: Uplink (Underlay)
OIF: NVE1 (Underlay)

S,G – (10.10.10.100, 224.10.10.10)
IIF: VLAN100 (Host-facing)
OIF: VRF-L3VNI (Overlay), VLAN100 (Host-facing)

SRC-10
224.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
20.20.20.20

RCVR-14
10.10.10.14
Agenda

• Introduction to Overlays
• VXLAN with BGP EVPN
  • Standards and Implementation
  • Control & Data Plane
• Tenant Routed Multicast (TRM)
• Multi-Site
• VXLAN OAM
Inter-X Connectivity

VXLAN Multi-Pod

- Single Fabric with End-to-End Encapsulation
- Build Hierarchy in the Underlay – Flatten it in the Overlay

VXLAN Multi-Fabric

- Multiple Fabrics – Normalized through Ethernet
- Multiple Fabrics Interconnect using DCI (Layer 2 and Layer 3)

VXLAN Multi-Site

- Multiple Fabrics with Integrated DCI
- Integrated DCI – Scaling within and between Fabrics
Functional Components

Border Gateways
(Key Functional Components of VXLAN Multi-Site Architecture)

Site-Internal Fabric
(Common VXLAN and BGP-EVPN Functions)

Hierarchical Overlay Domains
Anycast Border Gateway (1)

Anycast Border Gateway
- Up to 4 Border Gateways
- Border Gateway
  - Deploying at Leaf – 7.0(3)I7(1)
  - Deploying at Spine – 7.0(3)I7(2)
Anycast Border Gateway (2)

Anycast Border Gateway

- Common Virtual IP (VIP) across BGW
  - VIP for communication between the Border Gateways in different Sites
  - VIP for communication between Border Gateway and Leaf within a Site

- Individual Primary IP (PIP) per BGW
  - Used for Broadcast, Unknown Unicast and Multicast (BUM) replication
  - PIP for communication with Single-Homed endpoints (routed only), intra- and inter-Site
Anycast Border Gateway (3)

Anycast Border Gateway

- Per-VNI Designated Forwarder (DF) election
  - Each BGW can serve as DF for a single or a set of Layer-2 VNIs
  - DF election and assignment is automatic
- Using BGP EVPN Route Type 4 for DF election
  - Operator Managed Assignment (Type: 03)
  - Six Octet Site Identifier (System MAC: 00:00:00:00:00:01)
  - Multi-Site Discriminator (Ethernet-Segment: 00:03:09)
  - Originators IP Address (PIP): 10.1.1.101
  - Layer-2 VNI: 30010
Failure Detection on BGWs – Fabric Isolation (1)

- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status ('**evpn multisite fabric-tracking**' command)
Failure Detection on BGWs – Fabric Isolation (2)

- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status ('evpn multisite fabric-tracking' command)
- If all the Site-Internal interfaces are detected as down:
  1. The isolated BGW stops advertising PIP/VIP addresses toward the Site-External network
  2. The remaining BGWs perform new DF elections for the L2VNIs owned by the isolated BGW
Failure Detection on BGWs – Fabric Isolation (3)

- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status (‘evpn multisite fabric-tracking’ command).

- If all the Site-Internal interfaces are detected as down:
  1. The isolated BGW stops advertising PIP/VIP addresses toward the Site-External network.
  2. The remaining BGWs perform new DF elections for the L2VNIs owned by the isolated BGW.

- As a result, the BGW becomes isolated from both the Site-Internal and Site-External networks.

- Seamless BGW node reinsertion using a “delay-restore” timer for the VIP address.
Failure Detection on BGWs – DCI Isolation

- The Site-External interfaces on BGW nodes are also tracked to determine their status (‘evpn multisite dci-tracking’ command)

- If all the Site-External interfaces are detected as down, the isolated BGW node:
  1. Stops advertising VIP VTEP address toward the Site-Internal network
  2. Withdraws BGP EVPN Type-4 advertisements (triggering a new DF election between other BGWs)
  3. Starts functioning as a regular VTEP (PIP still up)

- As a result, the BGW continues to operate as a Site-Internal VTEP

- Seamless BGW node reinsertion using a “delay-restore” timer for the VIP address
Multi-Site Control Plane Deployment Considerations

- **MP-eBGP EVPN only inter-Sites**
  - Next-hop behavior (VXLAN tunnel termination and reorigination) and loop protection (as-path attribute)

- **Two main options for underlay and overlay control plane deployment**
  1. **I-E-I (Recommended)**
     - Intra-Site: IGP (OSPF, IS-IS) as underlay CP, iBGP as overlay CP
     - Inter-Sites: eBGP for both underlay and overlay CPs
  2. **E-E-E**
     - Intra-Site and Inter-Sites: eBGP for both underlay and overlay CPs

- **Full mesh of MP-eBGP EVPN adjacencies across sites**
  - Recommended to deploy a couple of *Route-Servers* with 3 or more sites
  - RS in a separate AS only perform **control plane** functions (“eBGP Route-Reflectors”, IETF RFC 7947)
  - RS functions: EVPN routes reflection, next-hop-unchanged, route-target rewrite
Multi-Site Overlay Control Plane – back-to-back
Multi-Site Overlay Control Plane – Route-Server
Multi-Site Overlay Control Plane – Tenants

VRF Tenant1

L3VNI: 50001
Route-Target: 65501:50001

DC Core
(Layer-3 Unicast)

VRF Tenant1

L3VNI: 50001
Route-Target: 65502:50001

DCI

Fabric

VIP 10.1.1.111

VTEP

BGW

Host1
0000.3010.1101
192.168.10.101

Host2
0000.3020.2101
192.168.20.101

Host3
0000.3010.1102
192.168.10.102

VIP 10.2.2.222

VTEP

BGW

RS

iBGP-EVPN

VX

iBGP-EVPN

PN

RR

Spine

Spine

L2VNI: 30010 (VLAN 10)
L3VNI: 50001 (Tenant1)

L2VNI: 30020 (VLAN 20)
L3VNI: 50001 (Tenant1)

L2VNI: 30010 (VLAN 10)
L3VNI: 50001 (Tenant1)
Multi-Site Overlay Control Plane – Site1

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0000.3010.1101/48</td>
<td>30010, 65501:30010</td>
<td>192.168.10.101/32</td>
<td>50001, 65501:50001</td>
<td>10.1.1.1</td>
<td>10.1.1.111</td>
</tr>
</tbody>
</table>

VIP1: 10.1.1.111
VRF Tenant1
L3VNI: 50001
Route-Target: 65502:50001

VIP2: 10.2.2.222
VRF Tenant1
L3VNI: 50001
Route-Target: 65501:50001

VTEP VTEP VTEP VTEP VTEP VTEP VTEP
Spine Spine Spine Spine Spine Spine Spine
VXLAN EVPN VXLAN EVPN
RS RS
50001 50001
Route-Target: 65502:50001 65501:50001

L2VNI: 30010 (VLAN 10)
L3VNI: 50001 (Tenant1)
Host1 0000.3010.1101
192.168.10.101

L2VNI: 30020 (VLAN 20)
L3VNI: 50001 (Tenant1)
Host2 0000.3020.2101
192.168.20.101

L2VNI: 30010 (VLAN 10)
L3VNI: 50001 (Tenant1)
Host3 0000.3010.1102
192.168.10.102
Multi-Site Overlay Control Plane – Site2

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0000.3010.1101/48</td>
<td>30010, 65502:30010</td>
<td>192.168.10.101/32</td>
<td>50001, 65502:50001</td>
<td>10.2.2.222</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0000.3020.2101/48</td>
<td>30020, 65502:30020</td>
<td>192.168.20.101/32</td>
<td>50001, 65502:50001</td>
<td>10.2.2.1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0000.3010.1102/48</td>
<td>30010, 65502:30010</td>
<td>192.168.10.102/32</td>
<td>50001, 65502:50001</td>
<td>10.2.2.3</td>
<td></td>
</tr>
</tbody>
</table>
Multi-Site Overlay Control Plane – Between Sites

<table>
<thead>
<tr>
<th>Type</th>
<th>MAC / Length</th>
<th>L2VNI / RT</th>
<th>IP / Length</th>
<th>L3VNI / RT</th>
<th>Next-Hop</th>
<th>Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0000.3020.2101/48</td>
<td>30020, 65599:30020</td>
<td>192.168.20.101/32</td>
<td>50001, 65599:50001</td>
<td>10.2.2.222</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0000.3010.1102/48</td>
<td>30010, 65599:30010</td>
<td>192.168.10.102/32</td>
<td>50001, 65599:50001</td>
<td>10.2.2.222</td>
<td></td>
</tr>
</tbody>
</table>
Multi-Site Overlay Data Plane – Overview

- De-capsulation and Re-encapsulation on BGW (L2 or L3 lookup)
- VXLAN EVPN
- Site1
  - VTEP
  - VIP1: 10.1.1.111
  - BGW
  - Host1: 0000.3010.1101 192.168.10.101
- VXLAN EVPN
  - Site2
  - VTEP
  - VIP2: 10.2.2.222
  - BGW
  - Host2: 0000.3020.2101 192.168.20.101
  - Host3: 0000.3010.1102 192.168.10.102

- DC Core (Layer-3 Unicast)
- Inter-site VXLAN Data Plane
- Intra-site VXLAN Data Plane
- De-capsulation and Re-encapsulation on BGW (L2 or L3 lookup)
Multi-Site Packet Walk (BUM)
Multi-Site – BUM Traffic Distribution
Multi-Site – BUM Replication Modes (Multicast Sites)
Multi-Site – BUM Replication Modes (All Ingress Replication)
Multi-Site – BUM Replication Modes (Mixed Site)
Multi-Site – BUM Traffic Enforcement

Overlay Multi-Site

Storm Control
Broadcast 0-100%
Unknown Unicast 0-100%
Multicast 0-100%

Overlay Site 1

Site 1

Overlay Site n

Site n
Leaf10 replicates traffic intra-Site

Host 1 sends a L2 BUM frame

Layer 2 (BUM) – Site 1
Layer 2 (DF and Split Horizon) – Site 1

<table>
<thead>
<tr>
<th>SIP</th>
<th>DIP</th>
<th>VXLAN</th>
<th>SMAC</th>
<th>DMAC</th>
<th>SIP</th>
<th>DIP</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>L10</td>
<td>DGROUP</td>
<td>30010</td>
<td>H1-MAC</td>
<td>ALL-F</td>
<td>H1-IP</td>
<td>ALL-255</td>
<td></td>
</tr>
</tbody>
</table>
Layer 2 (BUM) – DCI

<table>
<thead>
<tr>
<th>SIP</th>
<th>DIP</th>
<th>VXLAN</th>
<th>SMAC</th>
<th>DMAC</th>
<th>SIP</th>
<th>DIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGW11</td>
<td>BGW21</td>
<td>30010</td>
<td>H1-MAC</td>
<td>ALL-F</td>
<td>H1-IP</td>
<td>ALL-255</td>
</tr>
<tr>
<td>BGW11</td>
<td>BGW22</td>
<td>30010</td>
<td>H1-MAC</td>
<td>ALL-F</td>
<td>H1-IP</td>
<td>ALL-255</td>
</tr>
</tbody>
</table>

BGW11 replicates traffic inter-Sites toward remote BGW nodes.

Host 1
0000.3010.1101
192.168.10.101

Host 2
0000.3010.1102
192.168.10.102
Layer 2 (DF and Split Horizon) – DCI

<table>
<thead>
<tr>
<th>SIP</th>
<th>DIP</th>
<th>VXLAN</th>
<th>SMAC</th>
<th>DMAC</th>
<th>SIP</th>
<th>DIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGW11</td>
<td>BGW21</td>
<td>30010</td>
<td>H1-MAC</td>
<td>ALL-F</td>
<td>H1-IP</td>
<td>ALL-255</td>
</tr>
<tr>
<td>BGW11</td>
<td>BGW22</td>
<td>30010</td>
<td>H1-MAC</td>
<td>ALL-F</td>
<td>H1-IP</td>
<td>ALL-255</td>
</tr>
</tbody>
</table>

Baremetal Host 1
0000.3010.1101
192.168.10.101

Baremetal Host 2
0000.3010.1102
192.168.10.102

Drop due to Designated Forwarder (DF) rule
Drop due to Split-Horizon rule

BUM Forward
Layer 2 (BUM) – Site 2

<table>
<thead>
<tr>
<th>SIP</th>
<th>DIP</th>
<th>VXLan</th>
<th>SMAC</th>
<th>DMAC</th>
<th>SIP</th>
<th>DIP</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGW22</td>
<td>DGROUP</td>
<td>30010</td>
<td>H1-MAC</td>
<td>ALL-F</td>
<td>H1-IP</td>
<td>ALL-255</td>
<td></td>
</tr>
</tbody>
</table>
Layer 2 (DF and Split Horizon) – Site 2

Drop due to Designated Forwarder (DF) rule
Drop due to Split-Horizon rule

SIP | DIP | VXLAN | SMAC | DMAC | SIP | DIP | Payload
---|---|---|---|---|---|---|---
BGW22 | DGROUP | 30010 | H1-MAC | ALL-F | H1-IP | ALL-255 | Payload

Bridge

VXLAN EVPN Site1

VTEP

Leaf10

VTEP

BGW11

VTEP

BGW12

VXLAN EVPN Site2

VTEP

Leaf20

VXLAN EVPN DCI

Payload

Host 1
0000.3010.1101
192.168.10.101

Host 2
0000.3010.1102
192.168.10.102
Layer 2 (BUM) – Site 2

Bridge

Leaf20 sends traffic to local Host 2

Host 1
0000.3010.1101
192.168.10.101

Host 2
0000.3010.1102
192.168.10.102
Multi-Site and External Layer 3 Connectivity

- The BGW nodes can also be used to provide Layer-3 external connectivity to each site
- Different connectivity models are supported
  - VRF-Lite peering with external WAN Edge routers
  - MP-BGP EVPN peering with external WAN Edge routers (Shared Border deployment model, aka GOLF)
  - Dedicated or shared pair of WAN Edge routers across sites
- External Layer-3 network may be different from the DCI network used for inter-site communication
Border Gateway and VRF-Lite

Dedicated physical interfaces / sub-interfaces* for each VRF, separate from interfaces used for Multi-Site traffic

Separate routing peering for each VRF (IGP or eBGP)

*No current SVIs support on BGWs
Border Gateway and Shared External Connectivity

- **Single MP-BGP EVPN instance to exchange routes for all VRFs**
- **Various hand-off options depending on deployed HW (VRF-Lite, MPLS-VPN, LISP)**
- **Shared Border operates like a traditional VXLAN EVPN VTEP (Layer 3 only)**
- **VXLAN Data Plane between BGW and WAN Edge Router**

**Site-External**

**Site-Internal**
Agenda

• Introduction to Overlays

• VXLAN with BGP EVPN
  • Standards and Implementation
  • Control & Data Plane
  • Multicast Forwarding
  • Multi-Site

• VXLAN OAM
Operations, Administration and Management (OAM)

- OAM – processes, activities, tools and standards
- Various Mode of Operation
  - Pro-Active
    - Controlling a Situation
  - Re-Active
    - Responding to a Situation
# VXLAN OAM - OAM Model of Operation

**Endpoint Locator**
- Locate End-Host and Segment Identifier
- Track History of End-Host
- Provide Fabric Host-Count and Activity

**Ping / Path MTU**
- Check liveliness of End-Host
- Option to specify Payload Parameters

**Pathtrace**
- Trace paths to End-Host and Tunnel-Endpoint
- Get Path, Interface and Error statistics along path
- Specify Payload Parameters for Path Selection

**Pro-Active Monitoring**
- Proactive Monitoring with Threshold and State Notifications
NGOAM or VXLAN OAM

- Next Generation OAM for Data Center Fabrics
- Running on Nexus 9000, Nexus 7000 and Nexus 5600
  - VXLAN Today
  - All IP Tomorrow
- Various Methods to Execute and Retrieve Data
  - Command Line Interface (CLI)
  - NX-API
  - DCNM (using NX-API)
Endpoint Reachability – VXLAN OAM

- **Endpoint Reachability**
  - Uses ICMP
  - VTEP to Endpoint reachability
  - VTEP to VTEP reachability

- **Validates ECMP Path**
  - Single Random Path
  - Multiple, Random/Specified Path

- **Provides VXLAN Outer UDP Source Port (SPORT) as output**

Is Host A alive?

### Host A
- MAC: 0000.3001.1101
- IP: 192.168.10.101

### Host B
- MAC: 0000.3001.1102
- IP: 192.168.10.102

### Host C
- MAC: 0000.3002.2101
- IP: 192.168.20.101
Endpoint Traceroute – VXLAN OAM

- Endpoint Reachability
  - Uses ICMP
  - VTEP to Endpoint reachability
  - VTEP to VTEP reachability

- Validates Overlay Path
  - Single Specified Path
  - Multiple, Specified Path

- Provides Overlay to Underlay correlation

Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101
How would a normal Traceroute look alike?

Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Which Path did my Traceroute take?

What is the Path to Host A?
How would a normal Traceroute look alike?

Which Path did my Traceroute take?

L15# traceroute 192.168.10.101 source 10.50.1.15 vrf BLUE
traceroute to 192.168.10.101 (192.168.10.101) from 10.50.1.15 (10.50.1.15), 30 hops max, 40 byte packets
1 10.50.1.18 (10.50.1.18)  0.96 ms  0.817 ms  0.746 ms
2 2 192.168.10.101 (192.168.10.101)  4.751 ms  0.69 ms  0.697 ms
Endpoint Traceroute – VXLAN OAM – Close-Up

L15# traceroute nve ip 192.168.10.101 vrf BLUE source 10.50.1.15 sport 35977 verbose

Codes: '!' - success, 'Q' - request not sent, '.' - timeout, 'D' - Destination Unreachable, 'X' - unknown return code, 'm' - malformed request(parameter problem), 'c' - Corrupted Data/Test, '#' - Duplicate response

Traceroute Request to peer ip 10.200.200.18 source ip 10.200.200.15
Sender handle: 94
1 !Reply from 10.1.1.17,time = 1 ms
2 !Reply from 10.200.200.18,time = 1 ms
3 !Reply from 192.168.10.101,time = 4 ms

Spine Ingress Interface and Destination VTEP IP Address are Underlay Information – additions vs. standard Traceroute
Pathtrace for Enhanced Network Visibility

What is the Path from Host C to Host A for HTTP?

- Application Specific Pathtrace
  - Uses “draft-tissa-nvo3-oam-fm”
  - Endpoint to Endpoint Pathtrace
  - Adds Interface Load and Error Statistic of the Path
  - Uses Protocol Information
- Validates Specific or All Path
- Provides Overlay to Underlay correlation
- Superset of NVE Ping/Traceroute
Pathtrace – VXLAN OAM – Close-Up

L15# pathtrace nve ip unknown vrf BLUE
  payload
  ip 192.168.10.101 192.168.20.101
  port 54321 80
  proto 6
  payload-end

Codes: '!' - success, 'Q' - request not sent, '.' - timeout,
'D' - Destination Unreachable, 'X' - unknown return code,
'm' - malformed request(parameter problem),
'c' - Corrupted Data/Test, '#' - Duplicate response

Path trace Request to peer ip 10.200.200.18 source ip 10.200.200.15
Sender handle: 142

<table>
<thead>
<tr>
<th>Hop</th>
<th>Code</th>
<th>ReplyIP</th>
<th>IngressI/f</th>
<th>EgressI/f</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>!</td>
<td>Reply from 10.1.1.17, Eth1/5 Eth1/8</td>
<td>UP / UP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>!</td>
<td>Reply from 10.200.200.18, Eth1/54 Unknown</td>
<td>UP / DOWN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Why are we specifying Payload information?

- VXLAN provides variable UDP Source Port in Outer Header
- Hash of the inner Layer-2/Layer-3/Layer-4 Headers of the original Ethernet Frame.
- Enables entropy for ECMP Load balancing in the Network

Which Path did your Application Traffic took?
Pathtrace – VXLAN OAM – Close-Up

L15# pathtrace nve ip unknown vrf BLUE payload ip 192.168.10.101 ...

Codes: '!' - success, 'Q' - request not sent, '.' - timeout, 'D' - Destination Unreachable, 'X' - unknown return code, 'm' - malformed request (parameter problem), 'c' - Corrupted Data/Test, '#' - Duplicate response

Path trace Request to peer ip 10.200.200.18 source ip 10.200.200.15
Sender [20.20.20.20]

<table>
<thead>
<tr>
<th>Hop</th>
<th>Code</th>
<th>ReplyIP</th>
<th>IngressI/i</th>
<th>EgressI/f</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>!</td>
<td>Reply from 10.1.1.17, Eth1/5 Eth1/8</td>
<td>UP / UP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>!</td>
<td>Reply from 10.200.200.18, Eth1/54 Unknown</td>
<td>UP / DOWN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Spine Ingress Interface, Egress Interface and Destination VTEP IP Address are Underlay Information – additions vs. standard and NVE Traceroute

Spine Ingress Interface IP | Spine Ingress Interface | Spine Egress Interface

Interface Status

Destination VTEP IP | Destination Leaf Ingress Interface

Spine Ingress Interface, Egress Interface and Destination VTEP IP Address are Underlay Information – additions vs. standard and NVE Traceroute
L15# show ngoam pathtrace database session 168 detail

Pathtrace entry for session id 168
================================

Start time: Tue Jun 13 01:18:39.710 PDT
End time: Tue Jun 13 01:18:39.735 PDT

Last Clear of Summary Statistics: Never
Pathtrace Requests: sent (2)/received (0)/timeout (0)/unsent (0)
Pathtrace Replies: sent (0)/received (2)/unsent (0)/Duplicate (0)

! Reply from 10.1.1.17 on Eth1/5, state UP. Sent on Eth1/8, state UP.
  Interface stats for interface: Eth1/5
-------------------------------------------
Rx Len  : 84
Rx Bytes: 66113123
Rx Pkt rate: 0
Rx Byte rate: 0
Rx Load: 0
Summary

- Overview on VXLAN Overlay
- Standards and Implementation
- Control- and Data-Plane interactions
- Some info around Multicast forwarding
- How Multi-Site enhanced VXLAN EVPN
- Operations is key – VXLAN OAM
If you haven’t had enough VXLAN BGP EVPN
Links & Resources

• VXLAN Multi-Site Intro

• VXLAN Multi-Site @ Cisco Live online
  • https://www.ciscolive.com/global/on-demand-library/?search=BRKDCN-2035#

• ”eBGP” for EVPN

• Configuration Example
  • https://communities.cisco.com/community/technology/datacenter/data-center-networking/blog/2015/05/19/vxlanevpn-configuration-example
Cisco Spark

Questions?
Use Cisco Spark to communicate with the speaker after the session

How
1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install Spark or go directly to the space
4. Enter messages/questions in the space

cs.co/ciscolivebot#BRKDCT-3378
• Please complete your Online Session Evaluations after each session

• Complete 4 Session Evaluations & the Overall Conference Evaluation (available from Thursday) to receive your Cisco Live T-shirt

• All surveys can be completed via the Cisco Live Mobile App or the Communication Stations

Don’t forget: Cisco Live sessions will be available for viewing on-demand after the event at www.ciscolive.com/global/on-demand-library/.
Continue Your Education

• Demos in the Cisco campus
• Walk-in Self-Paced Labs
• Tech Circle
• Meet the Engineer 1:1 meetings
• Related sessions
Thank you
You're it