VxLAN BGP-EVPN

Vinit Jain
Twitter - @vinugenie
Email: vinijain@cisco.com
Agenda

• VxLAN Overview
  – Flood & Learn Mechanism
  – Ingress Replication

• Intro to VxLan BGP EVPN
  – Components / Features
  – BGP EVPN Route Types and Fields
VxLAN Overview

Simple Definition

• VXLAN Overlay
  – Layer 2 overlay on top of your Layer 3 underlay

• VxLAN Network Identifier
  – Each VxLAN segment is identified by a unique 24-bit segment ID
  – Only hosts on the same VNI are allowed to communicate with each other

• Benefits
  – Overcome 4094 VLAN Scale limitation
  – Better utilization of available network paths
  – Multi-Tenant with virtualization
VxLAN Overview

IP Network (Underlay)

- Edge Device
- Local LAN Segment
- IP Interface
- Physical Host
- Virtual Switch
- Virtual Hosts

Local LAN Segment
Physical Host
Virtual Hosts
Virtual Switch
Edge Device
VxLAN Overview

VTEP – VXLAN Tunnel End-Point
VNI/VNID – VXLAN Network Identifier
VxLAN Overview

VXLAN Concepts

• **VXLAN Overlay**
  - A VXLAN Overlay or VXLAN segment is a Layer-2 broadcast domain identified by the VNID that extends or tunnels traffic from one VTEP to another.

• **VXLAN Tunnel End Point (VTEP)**
  - A VTEP is a device that provides both encapsulation and de-capsulation of classical Ethernet and VXLAN packets to and from a VXLAN segment
  - Each VTEP may have the following types of interfaces:
    - Switchport interfaces on the local LAN segment to support local endpoints
    - Layer-3 interfaces to the transport IP network
    - SVI interfaces

• **VXLAN Gateway**
  - A VTEP that bridges traffic between VXLAN segments
The outer IP header has the source IP and destination IP of the VTEP endpoints.
The outer Ethernet header has the source MAC of the source VTEP and the destination MAC of the immediate Layer-3 next hop.

Outer UDP Destination Port = VXLAN (originally 8472, recently updated to 4789)
Outer UDP Source Port = Hash of Inner Frame Headers (optional)
VxLAN Overview

VxLAN Gateway Types

- **Layer 2 Gateway**
  - The layer 2 gateway is required when the layer 2 traffic (IEEE 802.1q tagged traffic) comes from VLAN into VxLAN segment (encapsulation) or
  - The Ingress VxLAN packet egresses out an 802.1q tagged interface (de-encapsulation), where the packet is bridged to a new VLAN.

- **Layer 3 Gateway**
  - A layer 3 gateway is used when there is a VxLAN to VxLAN routing
  - The ingress packet is a VxLAN packet on a routed segment but the packet egresses out on a tagged 802.1q interface and the packet is routed to a new VLAN
VxLAN – Flood and Learn

Overview

- Data Plane learning technique for VxLAN
- VNI’s are mapped to a multicast group on a VTEP
- Local MACs are learnt over a VLAN (VNI) on a VTEP
- Broadcast, Unknown Unicast, Multicast (BUM Traffic) is flooded to the delivery multicast group for that VNI
- Remote VTEPs part of same multicast group learn host MAC, VNI and source VTEP as the next-hop for the host MAC from flooded traffic
- Unicast packets to the host MAC are sent directly to source VTEP as VxLAN encapsulated packet
VxLAN Overview

Ingress Replication

- Some customers not comfortable deploying multicast in their core
- With *Ingress Replication (IR)*, BUM traffic ingress access side is replicated to remote VTEP as unicast
- Static IR VETP tunnel is kept alive as long as the route to the VTEP is available.
- Support multiple VTEPs per VNI and a VTEP in multiple VNIs
- Up to 16 static IR VTEPs recommended – on Cisco Platforms
- Multicast and IR config can co-exist on the same switch nodes but on different VNI’s
Problem Definition

- Host placement anywhere, and mobility
- Optimal east-west traffic
- Segmentation of tenant L2 and L3 tenant traffic
- Minimum flooding traffic
Elements of Solution

1. BGP Extension: Advertisement of Host MAC, IP routes
2. VxLAN Overlay
3. Distributed Anycase Gateway
4. ARP Suppression

L3 Underlay

L2, L3 traffic segmented via VxLAN Overlay Network

TOR 1

Host 1
Host 2

TOR 2

Host 3
Host 4

Anycast Gateways
BGP for VxLAN

WHY?

• Control plane learning for end host Layer-2 and Layer-3 reachability information to build more robust and scalable VXLAN overlay networks.

• Leverages the decade-long MP-BGP VPN technology to support scalable multi-tenant VXLAN overlay networks.

• EVPN address family carries both Layer 2 and Layer 3 reachability information. This provides integrated bridging and routing in VXLAN overlay networks.
### VxLAN BGP-EVPN Overview

#### VXLAN EVPN VNI Types

**Tenant 1 (VRF 1)**

- **SVI X**
  - Layer-3 VNI X'
  - VLAN X

  - 1 Layer-3 VNI per Tenant (VRF) for routing
  - VNI X' is used for routed packets

- **SVI A**
  - Layer-2 VNI A'
  - VLAN A

  - 1 Layer-2 VNI per Layer-2 segment
  - Multiple Layer-2 VNIs per tenant
  - VNI A' and B' are used for bridged packets

- **SVI B**
  - Layer-2 VNI B'
  - VLAN B
VxLAN BGP-EVPN Overview
BGP for VxLAN

Advantages

• Minimizes network flooding through protocol-driven host MAC/IP route distribution and ARP suppression on the local VTEPs.

• Provides optimal forwarding for east-west and north-south bound traffic with the distributed any-cast function.

• Provides VTEP peer discovery and authentication which mitigates the risk of rouge VTEPs in the VXLAN overlay network.
VxLAN BGP-EVPN Overview

Distributed Anycast Gateway

- All VTEPs has same IP address for an L2 VNI
- Anycast Gateway MAC is global to each VTEP for all VNI’s for all Tenants

```plaintext
fabric forwarding anycast-gateway-mac 0001.0001.0001
!
interface Vlan100
no shutdown
vrf context test-evpn-tenant
ip address 172.16.1.254/24
fabric forwarding mode anycast-gateway
```
VxLAN BGP-EVPN Overview

ARP Suppression

• Hosts send out G-ARP when they come online
• Local leaf node receives G-ARP, creates local ARP cache and advertises to other leaf by BGP as route type 2
• Remote leaf node puts IP-MAC info into remote ARP cache and suppresses incoming ARP request for this IP
• If IP info not found in ARP suppression cache table, VTEP floods the ARP request to other VTEPs
VxLAN BGP-EVPN Overview

Different Integrated Route/Bridge (IRB) Modes

Asymmetric IRB
- Uses different path from source to destination and back
- Required to configure the source VTEP with both the source and destination VNIs for both layer 2 and layer 3 forwarding

Symmetric IRB
- Uses same path from source to destination and back
- the ingress VTEP routes packets from source VNI to L3 VNI where the destination MAC address in the inner header is rewritten to egress VTEP’s router MAC address
EVPN Route Types

- **BGP EVPN Route Types**
  - Type 1 - Ethernet Auto-Discovery (A-D) route
  - Type 2 - MAC advertisement route $\rightarrow$ L2 VNI MAC/MAC-IP
  - Type 3 - Inclusive Multicast Route $\rightarrow$ EVPN IR, Peer Discovery
  - Type 4 - Ethernet Segment Route
  - Type 5 - IP Prefix Route $\rightarrow$ L3 VNI Route

- Route type 2 or MAC Advertisement route is for MAC and ARP resolution advertisement, MAC or MAC-IP

- Route type 5 or IP Prefix route will be used for the advertisement of prefixes, IP only
BGP EVPN Route Fields

Leaf1#show bgp l2vpn evpn 8c60.4f93.5ffc
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10000:1 (L2VNI 10000)
BGP routing table entry for [2]:[0]:[0]:[48]:[8c60.4f93.5ffc]:[0]:
[0.0.0.0]/216, version 8
Paths: (1 available, best #1)
Flags: (0x00010a) on xmit-list, is not in 12rib/evpn

Advertised path-id 1
Path type: local, path is valid, is best path, no labeled nexthop
AS-Path: NONE, path locally originated
192.168.1.1 (metric 0) from 0.0.0.0 (192.168.1.1)
Origin IGP, MED not set, localpref 100, weight 32768
Received label 10000
Extcommunity: RT:10000:1

Route Distinguisher – 8 byte
Ethernet Segment ID – 10 byte
Ethernet Tag ID – 4 byte
MAC Address Length – 1 byte
MAC Address – 6 byte
IP Address Length – 1 byte
IP Address – 0, 4, 16 byte
MPLS Label 1 – 3 byte, L2VNI
MPLS Label 2 – 3 byte, L3VNI
BGP EVPN Route Fields

<table>
<thead>
<tr>
<th>Field</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route Distinguisher</td>
<td>8 byte</td>
<td>Identifies the route in the AS</td>
</tr>
<tr>
<td>Ethernet Segment ID</td>
<td>10 byte</td>
<td>Identifies the Ethernet segment</td>
</tr>
<tr>
<td>Ethernet Tag ID</td>
<td>4 byte</td>
<td>Identifies the Ethernet tag</td>
</tr>
<tr>
<td>MAC Address Length</td>
<td>1 byte</td>
<td>Specifies the MAC address length</td>
</tr>
<tr>
<td>MAC Address</td>
<td>6 byte</td>
<td>Identifies the MAC address</td>
</tr>
<tr>
<td>IP Address Length</td>
<td>1 byte</td>
<td>Specifies the IP address length</td>
</tr>
<tr>
<td>IP Address</td>
<td>0, 4, 16 byte</td>
<td>Identifies the IP address</td>
</tr>
<tr>
<td>MPLS Label 1</td>
<td>3 byte</td>
<td>Identifies the L2VNI</td>
</tr>
<tr>
<td>MPLS Label 2</td>
<td>3 byte</td>
<td>Identifies the L3VNI</td>
</tr>
</tbody>
</table>

Leaf2# show bgp l2vpn evpn 100.1.1.1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 20000:1 (L3VNI 20000)
BGP routing table entry for [2]:[0]:[0]:[48]:[8c60.4f1b.e43c]:[32]:[100.1.1.1]/272, version 6
Paths: (1 available, best #1) Flags: (0x00021a) on xmit-list, is in l2rib/evpn, is not in HW, Advertised path-id 1 Path type: internal, path is valid, is best path, no labeled nexthop
Imported from 10000:1:[2]:[0]:[0]:[48]:[8c60.4f1b.e43c]:[32]:[100.1.1.1]/144 (VNI 10000)
AS-Path: NONE, path sourced internal to AS 192.168.1.1 (metric 5) from 192.168.10.10 (192.168.10.10)
Origin IGP, MED not set, localpref 100, weight 0 Received label 10000 20000 Extcommunity: RT:10000:1 RT:20000:1 ENCAP:8 Router MAC:f40f.1b6f.926f
Originator: 192.168.1.1 Cluster list: 192.168.10.10
Troubleshooting BGP
A Practical Guide To Understanding and Troubleshooting BGP

Vinit Jain, CCIE No. 22854
Brad Edgeworth, CCIE No. 31574
Q&A