Virtual Extensible LAN and Ethernet Virtual Private Network

Contents

Introduction

This document describes how Virtual Extensible LAN (VXLAN) helps data center operators support multitenancy, enables Equal Cost Multi-Pathing (ECMP) in order to make use of available paths, and avoid the perils of Spanning Tree. VXLAN works when you add a header to an Ethernet frame that makes it routable across an IP network. Also, how hosts inside a VXLAN network communicate with end points outside that network is discussed.

Contributed by Sabyasachi Kar, Cisco TAC Engineer.

Prerequisites

Requirements

Cisco recommends that you have knowledge of these topics:

- VLAN
- MULTICAST
- Border Gateway Protocol (BGP)

Components Used

The information in this document is based on these software and hardware versions:

- NX-OSv9K is a demo version of the Nexus Operating System Software
- BIOS
- NXOS: Version 7.0(3)I6(1)

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, ensure that you understand the potential impact of any command.

Background Information

Why you need a new extension for VLAN?

VLAN uses Spanning Tree Protocol (STP) for Loop prevention, which ends up not being able to use half of the network by blocking redundant paths. In contrast, VXLAN packets are transferred through the underlying network that is based on its Layer 3 header and takes complete advantage of layer 3 routing, ECMP and link aggregation protocols use all available paths.

VLAN has been running in the DC for many years but with the rapid growth of Virtualisation, On-demand VM, the increasing customer 4K VLAN is not sufficient.

Also, because of the limitation of STP such as link/path utilization convergence issues, MAC table size and some of the network links are under utilized.

VXLAN is a MAC-in-UDP encapsulation method that is used in order to extend a Layer 2 or Layer 3 overlay network over a Layer 3 infrastructure that already exists.

The VXLAN encapsulation provides a VNI that can be used to provide segmentation of Layer 2 and Layer 3 data traffic.

To facilitate the discovery of these VNI over the underlay Layer 3 network, virtual tunnel end point is used. VTEP is an entity that terminates the VXLAN tunnel end points.

It maps the Layer 2 frames to a VNI in order to be used in the overlay network. In order to encapsulate customer Layer 2 and Layer 3 traffic in VNI over the physical, the Layer 3 network provides.

Why do you chose EVPN over Virtual Private LAN Service (VPLS)?

- More efficient in the BGP table.
- Controls your information more completely, distribution of MAC addresses.
- Much simpler solution than VPLS.

EVPN isa next generation VPLS.

VPLS Highlights

- VPLS customer MAC addresses are learned through the data plane.
- Source MAC addresses are recorded based on Source address from both Attachment Circuit (AC) and Pseudowire (PW).
- In VPLS, in orderto balance active active flow-based load is not possible.
- Customer can be dual-homed to the same or different Provider Edge (PE) of service provider, but **e**ither those links can be used as active/standby for all VLAN or VLAN Based Load Balancing can be achieved.

EVPN Highlights

- EVPN can support active flow that is based on load balancing, so same VLAN can be used on both PE devices actively.
- This provides faster convergence in customer link, PE link, or node failure scenarios.
- Customer MAC addresses are advertised over the MP-BGP control plane. There is no data plane MAC learning over the core network in EVPN.
- But Customer MAC addresses from the AC are still learned through the data plane.

What is VXLAN?

As the name VXLAN implies, the technology is meant to provide the same service to connected ethernet end systems that VLAN do today, but in a more extensible manner.

Compared to VLAN, VXLAN are extensible with regards to scale the reach of their deployment.

802.1Q VLAN identifier space is only 12 bits. The VXLAN identifier space is 24 bits. This doubling in size allows the VXLAN ID space to increase over 400000 percent to 16 million unique identifiers.

VXLAN uses IP (both unicast and multicast) as the transport medium. The ubiquity of IP networks and equipment allows end-to-end reach of a VXLAN segment to be extended far beyond the typical reach of VLAN with the use of 802.1Q today.

One cannot deny that there are other technologies that can extend the reach of VLAN, however, none are as ubiquitously deployed as IP.

Terminology

EVI: An EVPN instance that spans across the PE’s that participate in that EVPN.

Ethernet Segment Identifier (ESI): The set of ethernet links that attach a CE to when CE is multi-homed to two or more PE's. Ethernet Segment must have a unique non-zero identifier, the Ethernet segment identifier.

Ethernet Tag: An ethernet TAG identifies a particular broadcast domain, e.g. a VLAN. An EVPN instance consists of one or more broadcast domains. Ethernet Tags are assigned to the broadcast domains of a given EVPN instance by the provider of that EVPN. Each PE in that EVPN instance performs a mapping between broadcast.

VLAN Tunnel EndPoint (VTEP)

VXLAN uses VTEP devices in order to map tenants end devices to VXLAN segment and in order to perform VXLAN encapsulation and decapsulation.

Each VTEP function has two interfaces:

- One is a switch interface on Local LAN Segment to support local endpoint communication through bridging.
- IP interface to the transport IP network.

The IP interface has a unique IP address that identifies the VTEP device on the transport IP network known as the infrastructure VLAN.

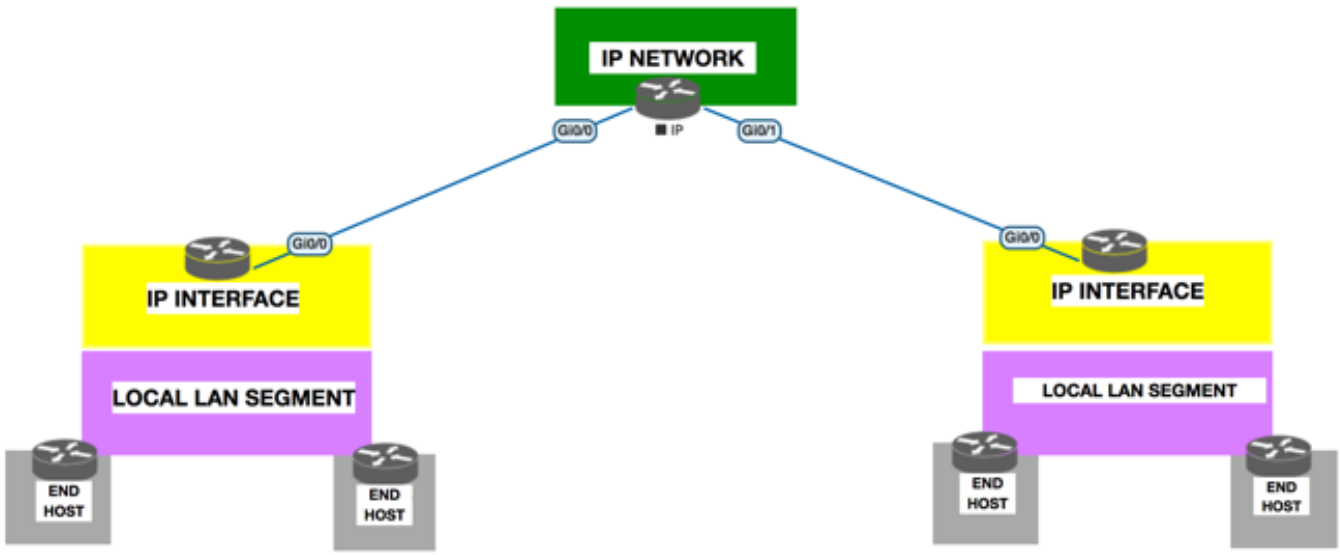
The VTEP device uses this IP address to encapsulate ethernet frames and transmits encapsulated packet to transport network through the IP interface.

A VTEP device also discovers the remote VTEP for its VXLAN segment and learns remote MAC address to VTEP mappings through the IP interface.

VTEP has Two Interfaces

Local LAN Interface: Provides a bridging function for local host connected to the VTEP.

IP Interface: The interface on the core network for VXLAN. The IP address on the IP interface helps to uniquely identify VTEP in the network.



IP intrasubnetwork or non-IP Layer 2 traffic is mapped to a VNI that is set aside for VLAN or bridge domain.

The routed traffic on the other hand is mapped to a VNI that is set aside Layer 3 VRF.

Because of the Layer 3 underlay network, VXLAN is capable to perform ECMP, link aggregation and other L3 functionalities.

STP is not required anymore, there is no more blocked path to make the network under-utilised.

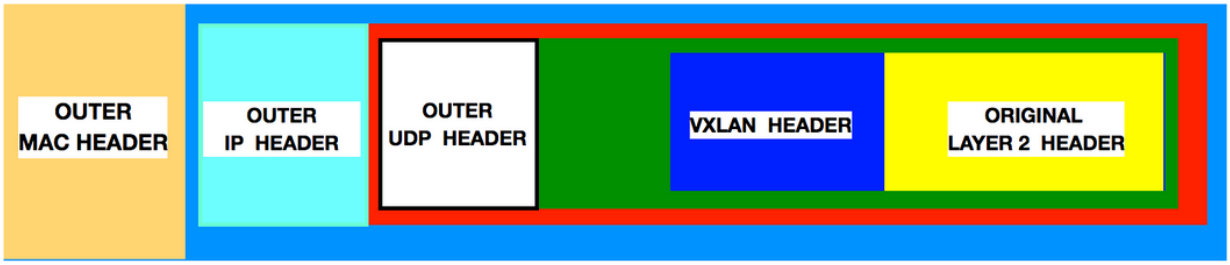
VXLAN provides multi-tenant solution where the network traffic is isolated by a tenant and the same VLAN can be used by different tenants.

VXLAN Encapsulation and Packet Format

VXLAN packet is nothing more than a MAC-in-UDP encapsulated packet. The VXLAN header is added to the original Layer 2 frame and then placed in a UDP-IP packet.

The VXLAN header is an 8 bytes header that consists of 24 but VXLAN Network Identifier (VNID) and few reserved bits.

The VNI uniquely identifies the layer 2 segment and helps to maintain isolation among them. Because the VNID is 24, VXLAN can support 16 million LAH segment.



Flags: 8 Bits in length, where the fifth bit (I Flag) is set to 1 and indicates valid VNI. The 7 bits (R bits) that remain are reserved fields and are set to zero.

VNI: 24 bits value that provide a unique identifier for the individual VXLAN segment.

Outer UDP Error: The Source port in the outer UDP header is dynamically assigned by the VTEP that is originated. The source port is calculated based on the hash of inner Layer 2/Layer 3/ Layer 4 headers of the original frame. The destination port assigned UDP port 4789 or the customer configured.

SRC Port: Dynamically Assigned Originating VTEP

DST PORT: 4789

Outer IP Error: The Source IP address in the outer IP header is the originating VTEP's IP interface. The IP address on the IP interface uniquely identifies a VTEP. The destination address of the outer IP header is the IP address of the destination VTEP's IP interface.

SRC IP: VTEP Interface IP

DST IP: IP Address of the Destination IP Interface

Outer Ethernet/ MAC Header: The Source MAC address is the source VTEP MAC address. The destination MAC address is the next-hop MAC address. The next hop is the interface that is used to reach the destination or remote VTEP.

SRC MAC: SRC VTEP ROUTER MAC

DST MAC: DST MAC is the Next Hop interface that re-reaches the destination or remote VTEP.

```
Frame 13: 164 bytes on wire (1312 bits), 164 bytes captured (1312 bits) on interface 0
Ethernet II, Src: 50:01:00:01:00:07 (50:01:00:01:00:07), Dst: 50:01:00:02:00:07 (50:01:00:02:00:07)
Internet Protocol Version 4, Src: 1.1.1.1, Dst: 3.3.3.3
   0100 .... = Version: 4
   .... 0101 = Header Length: 20 bytes
▶  Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
   Total Length: 150
   Identification: 0x8000 (32768)
▶  Flags: 0x00
   Fragment offset: 0
   Time to live: 254
   Protocol: UDP (17)
▶  Header checksum: 0x344f [validation disabled]
   Source: 1.1.1.1
   Destination: 3.3.3.3
   [Source GeoIP: Unknown]
   [Destination GeoIP: Unknown]
User Datagram Protocol, Src Port: 51377 (51377), Dst Port: 4789 (4789)
Virtual eXtensible Local Area Network
▶  Flags: 0x0800, VXLAN Network ID (VNI)
   Group Policy ID: 0
   VXLAN Network Identifier (VNI): 10020
   Reserved: 0
Ethernet II, Src: aa:bb:cc:80:51:00 (aa:bb:cc:80:51:00), Dst: aa:bb:cc:80:61:00 (aa:bb:cc:80:61:00)
Internet Protocol Version 4, Src: 10.0.0.1, Dst: 10.0.0.2
Internet Control Message Protocol
```

VXLAN is a Layer 2 overlay scheme over a Layer 3 network.

It uses MAC address in UDP encapsulation to provide a means to extend Layer 2 segment across the data centre network.

VXLAN is a solution to support a flexible, large-scale multi-tenant environment over a shared common physical infra.

The transport protocol over the physical data centre network is IP plus UDP.

VXLAN defines a MAC-in-UDP encapsulation scheme where the original Layer 2 frame has a VXLAN header added and is then placed in a UDP IP packet.

With this MAC-in-UDP encapsulation VXLAN tunnels Layer 2 network over Layer 3 network. The VXLAN packet format.

VXLAN introduces an 8 Bytes VXLAN header that consists of 24 bits VNID and few reserved bits. The VXLAN header together with the Original Ethernet Frame goes in UDP Payload. The 24 bit VNID is used to identify Layer 2 segment and to maintain Layer 2 isolation between the segment.

VXLAN Gateway Types:

Frames encapsulation and decapsulation is performed by the VTEP.

A VTEP originates and terminates VXLAN tunnels.

VXLAN gateway traffic between a VXLAN segment and another physical or logical Layer 2 domain (such as VLAN). There are two kinds of VXLAN Gateways.

Layer 2 Gateway:The Layer 2 gateway is required when the Layer 2 traffic comes from VXLAN segment (encapsulation) or the egress VXLAN packet egress out an 802.1q tagged interface (decapsulation) where the packet is bridge to a new VLAN.

Layer 3 Gateway:A Layer 3 gateway is used when there is a VXLAN to VXLAN routing, that is when the egress VXLAN packet is router to a new VXLAN segment. A Layer 3 gateway is also used when there is VXLAN to VLAN routing; that is 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 Maximum Transmission Unit (MTU):

VXLAN adds 50 bytes to the original Ethernet Frame.

VTEP must not fragment the VXLAN Packets

Intermediate routers may fragment encapsulated VXLAN packets due to the larger frame size.

The destination VTEP might silently discard such VXLAN fragments.

In order to ensure end-to-end traffic delivery without fragmentation, it is recommended that the MTU across the physical network infrastructure is set to a value that accommodates the large frame size due to the encapsulation.

VXLAN Overview

The VXLAN overlay mechanism requires that the VTEP peer be with each other so that the data can be forwarded to the relevant destination.

- Flood and Learn
- BGP EVPN
- Ingress Replication

VXLAN Flood and Learn Mechanism

This is a Data Plane learning technique for VXLAN, where a VNI is mapped to a multicast group on a VTEP.

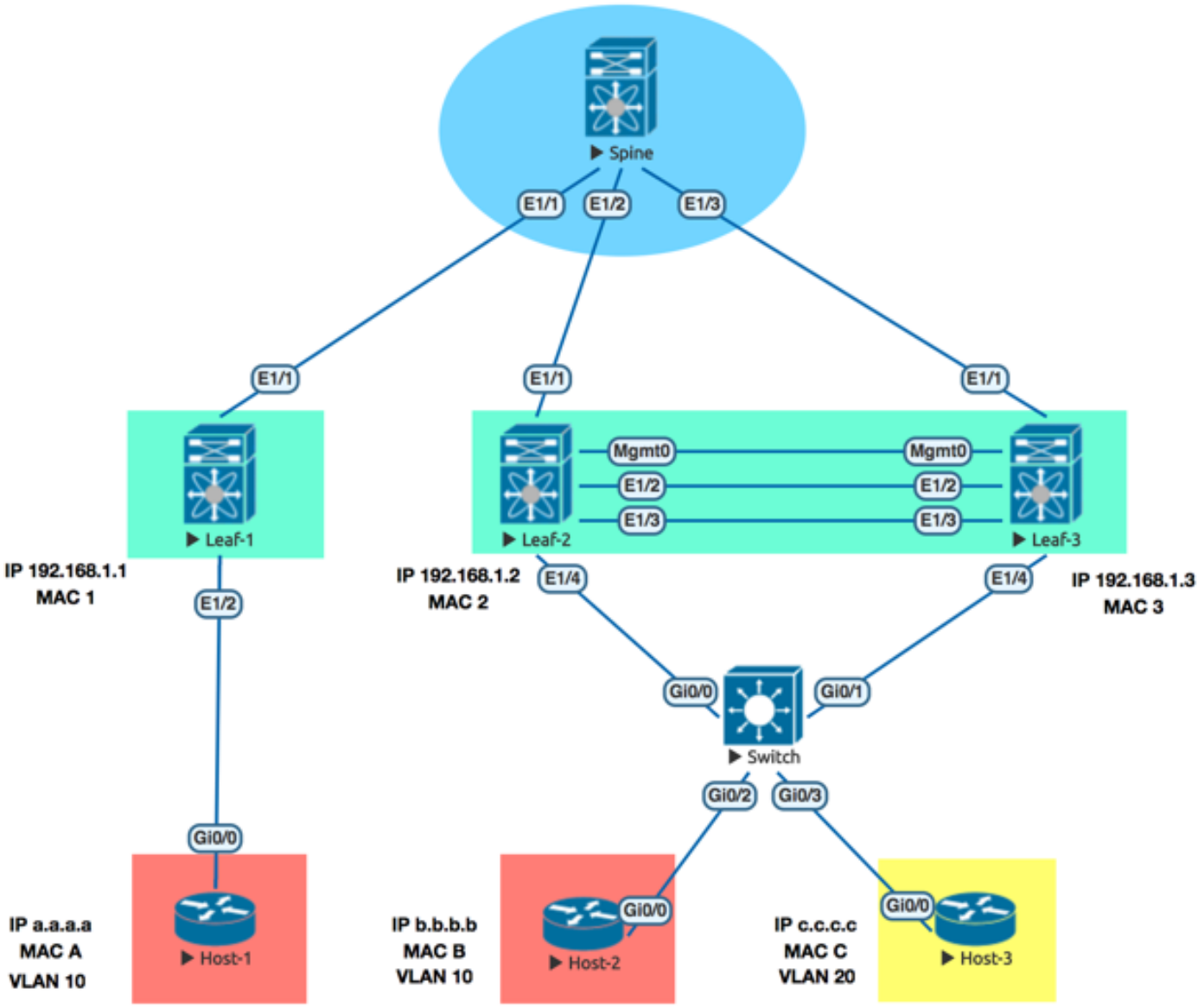
There is no control or signaling protocol defined, emulation of multidirectional traffic is handled through the VXLAN IP underlay through the use of segment control multicast group.

The Host traffic is always Broadcast/Unknown Unicast/Multicast (BUM) Format. BUM traffic is flooded to the multicast delivery group for the VNI that is sourcing the host packet. The remote VTEP that is a part of the multicastgroup learns about the remote host MAC, VNI and source

VTEP IP information from the flooded traffic.

The unicast packet to the Host MAC are sent directly to the destination VTEP as a VXLAN packet.

Note: Local MAC are learned over a VLAN (VNI) on a VTEP.



Packet Flow in Flood and Learn:

Step 1. The End System A with MAC-A and IP-A sends an ARP request for host with IP-B.

The source MAC address of the ARP packet is MAC-A and the destination MAC address is FF:FF:FF:FF:FF:FF.

Suppose the host is in VLAN 10. This packet is sent towards VTEP 1. VTEP 1 has VNID 10 mapped to VLAN 10.

Step 2.When the ARP request is received at the VTEP-1, the packet is encapsulated and forwarded tothe remote VTEP-2 and VTEP-3 with the source address as 192.168.1.1. and destination as 239.1.1.1. as a VXLAN packet. When the encapsulation is done, the VNID is set to 10, the source MAC of the packet is MAC 1, and the destination MAC is 0001.5E01.0101, which is

multicast MACaddress for 239.1.1.1.

> **Note**: VTEP that have subscribed to that particular multicast group received the multicast packet. The multicast group is configured to map to the VNI on each VTEP.

Step 3. Both the VTEP 2 and VTEP 3 receive the VXLAN packet and decapsulated it to forward it to the End-Systems connected to the respective VTEPS.

VTEP 2 and VTEP 3 update their MAC address table with this information:

```
 MAC address : MAC A

VxLAN ID  : 10

Remote VTEP : 192.168.1.1
```

Now, VTEP 2 and 3 knows the MAC address of MAC-A.

Step 4. After the ARP packet is forwarded to Host B after decapsulation, Host B responds back with the ARP reply.

Step 5. When the ARP reply reaches VTEP 2. VTEP 2 already knows that to reach MAC-A, it needs to go to VTEP-1. Thus VTEP 2 forwards the ARP reply from Host B as a unicast VXLAN packet.

Step 6. When the VXLAN packet reaches VTEP 1, it then updates its MAC address table with this information:

```
 MAC Address : MAC B

VxLAN ID : 10

Remote VTEP : 192.168.2.2
```

Step 7. After the MAC table is updated on VTEP 1, the ARP reply is forwarded to Host A.

Overview of VXLAN BGP EVPN

- Flexible Workload placement.
- Reduce flooding in the DC.
- Overlay setup with the use of Control Plane that is independent of specific fabric controller.
- Layer 2 and Layer 3 traffic segmentation.

The VXLAN Flood and Learn does not meet the requirements.

BGP MPLS based EVPN solution was developed in order to meet the limitation of the flood and learn mechanism.

In the BGP EVPN solution for VXLAN overlay, a VLAN is mapped to a VNI for the Layer 2 services and a VRF is mapped to VNI for the Layer 3 services on a VTEP.

An iBGP EVPN session is established between all the VTEPs or with the EVPN RR in order to provide the full mesh connectivity required by iBGP peering rules.

After the iBGP EVPN session is established, the VTEP exchanges MAC-VNI or MAC-IP bindings as part of BGP NLRI update.

Distributed Anycast Gateway:

Distributed anycast gateway refers to the use of any cast gateway addressing and an overlay network in order to provide a distributes control plane that governs the forwarding facilities of frames within and across a Layer 3 core network.

The distributed any cast gateway functionality facilitates transparent VM mobility and optimal east-west routing by configuring the leaf switches with same gateway IP and MAC address for each locally defined subnet.

The main benefit of the distributed any cast gateways is that the hosts or VM use the same default gateway IP and MAC address no matter which leaf they are connected to. Thus all VTEP have the same IP address and MAC address for the Switched Virtual Interface (SVI) in the same VNI.

Within the spine-and-leaf topology, there can be various traffic forwarding combinations. Based on the forwarding types, the distributed any cast gateway plays its role in oneof these manners:

Intra Subnet and Non IP Traffic: For the host-to-host communication that is intrasubnet or non IP, the destination MAC address in the ingress frame is the target end host's MAC address. This traffic is bridge from VLAN to VNI on the ingress/egress VTEP.

Inter Subnet IP Traffic: For host-to-host communication that is intersubnet, the destination MAC address in the ingress frame belongs to the default gateway MAC address. This traffic gets routed. But on the egress switch, there can be two possible forwarding behaviours, it can either get router or bridge.

If the inner destination MAC address belongs to the end host, then on the egress switch after VXLAN decapsulation, the traffic is bridge.

On the other hand, if the inner destination MAC address belongs to the egress switch, the traffic is routed.

In order to configure distributed any cast gateway, all the leaf switches or VTEP are required to be configured with the global command **Fabric Forwarding anycast-gateway-mac <MAC ADDRESS>** where MAC address is the statistically assigned address to be used across all switches by the anycast gateway.

The next step is to assign the fabric forwarding mode to any cast gateway with the use of the command **fabric forwarding mode anycast-gateway**.

ARP Supression:

ARP Request from a host is flooded in the VLAN.

It is possible to optimize the flooding behaviour and maintain an ARP cache locally on the attached VTEP and generate an ARP response from the information available from local cache.

This is achieved with the use of the ARP suppression feature.

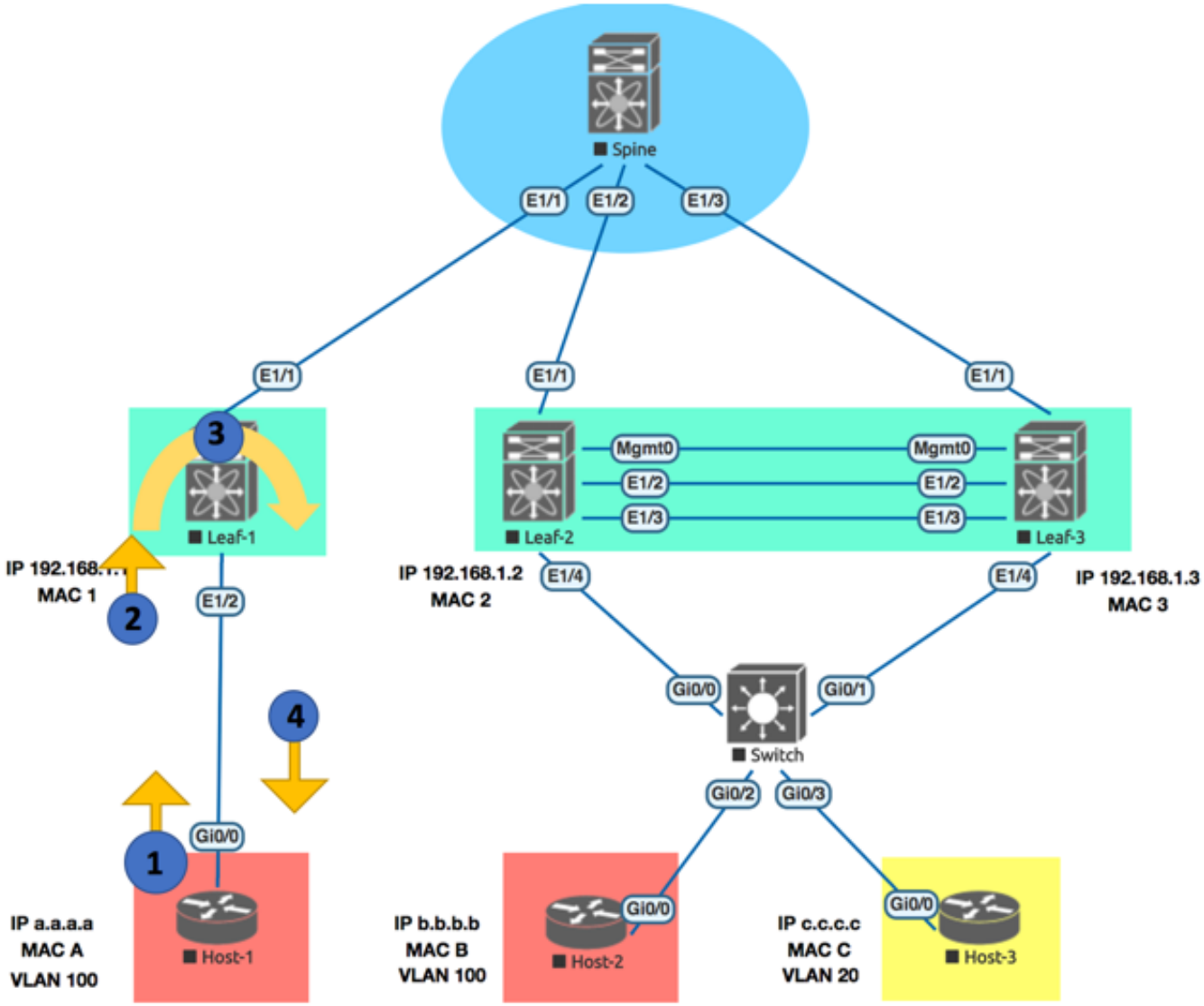
With the use of the ARP suppression, network flooding due to host learning can be reduced with

the use of G-ARP.

Typically, a host sends out a G-ARP message when its first comes online. When local VTEP device received the ARP, it creates an ARP cache entry and advertise to the remote leaf VTEP with the use of BGP Route Type 2. (BGP EVPN MAC route advertisement).

The remote leaf node puts the IP-MAC information into the remote ARP cache and surpasses the incoming ARP requests to that particular IP.

If a VTEP does not have a match for the IP address in its ARP cache table, it floods the ARP request to all other VTEP in the VNI.



Step 1. Host 1 in VLAN 100 sends an ARP request for Host 2 IP address.

Step 2.VTEP 1 on Leaf-1 intercepts the ARP request. Rather than forwarding it towards the core,it checks ARP suppression cache table. If it finds a match for Host 2 IP address in VLAN 100 in its ARP suppression cache. It is important to note that the BUM traffic is sent to other VTEPS.

Step 3.VTEP 1 sends the ARP response back to Host-1 with the MAC address of Host-2, this reduces the ARP flooding in the core network.

Step 4. Host 1 gets IP and MAC mapping for Host 2 and update the ARP cache.

Integrated Routed and Bridge Mode (IRB):

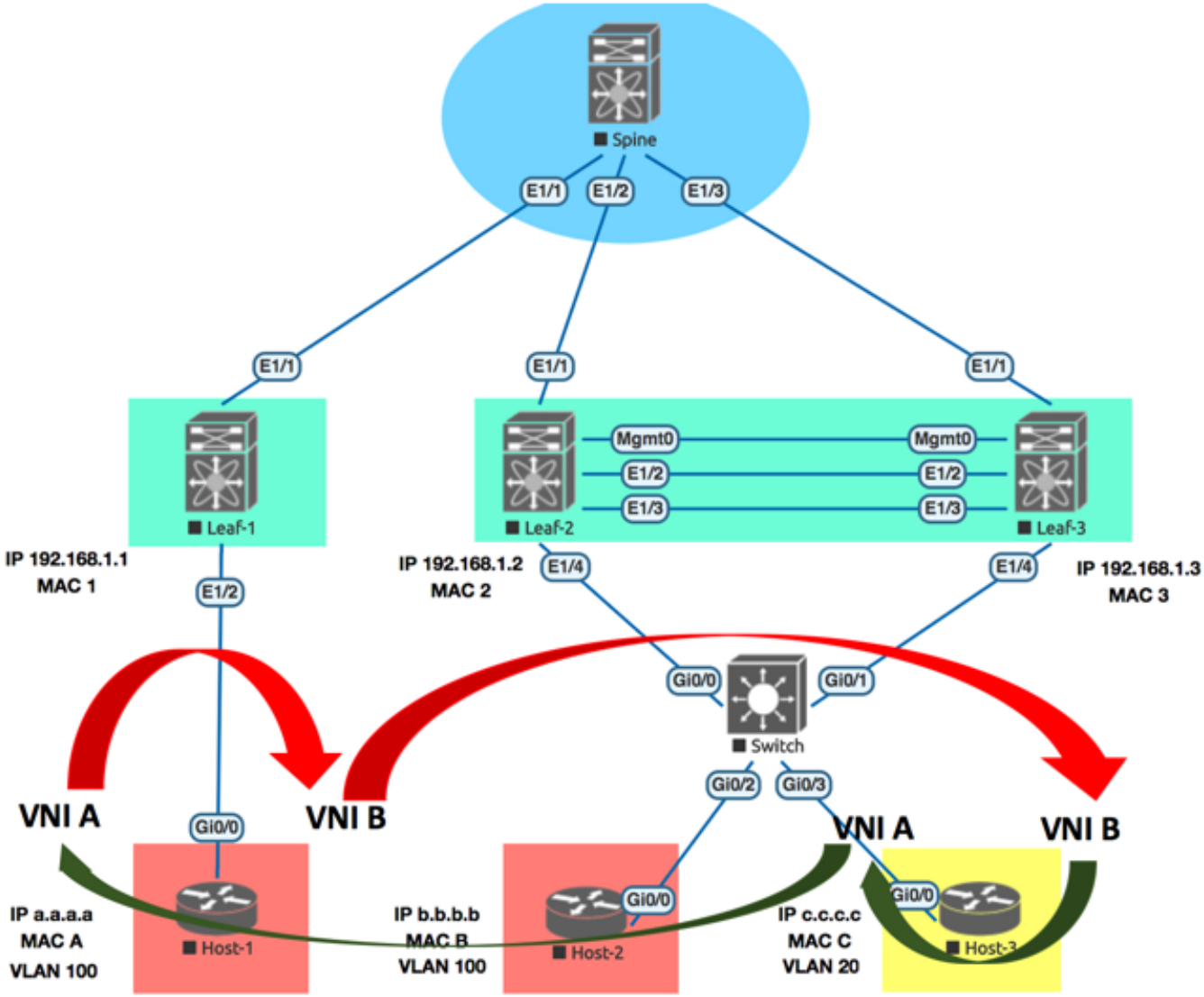
EVPN draft defines two IRB mechanisms:

1. Asymmetric IRB

2. Symmetric IRB

1. Asymmetric IRB

In this method VTEP performs both Layer 2 bridging and Layer 3 routing lookup, whereas the egress VTEP performs only Layer 2 bridging lookup. Asymmetric IRB requires the ingress VTEP to be configured with both the source and destination VNIs for both Layer 2 and Layer 3 forwarding. Essentially, it requires every VTEP to be configured with all VNIs in the VXLAN network and to learn ARP entries and MAC addresses for all the end hosts attached to those VNIs.



Step 1.Host 1 in VNI A sends a packet towards Host 2 with the source MAC address of Host 1 and the destination MAC address set to gateway MAC address set to gateway MAC.

Step 2.The ingress VTEP routes the packets from the source VNI to the destination VNI; that is, if the source packet was received in VNI-A the packet is routed to the destination VTEP VNI-B.

When the packet is sent, the source MAC of the inner packet is set to gateway MAC and the destination MAC as the Host 2 MAC address.

Step 3. When the packet reaches the destination VTEP, the egress VTEP bridges the packets in the destination VNI.

Step 4. The return packet also follows the same process.

Because the ingress VTEP device needs to be configured with both the source and destination VNI, it creates a scalability problem, because all the VTEP devices require to be configured with all VNI in the network so that they can learn about all the other hosts attached to those VNI.

Packet Flow

2. Symmetric IRB:

The symmetric IRB is more scalable and preferred option.

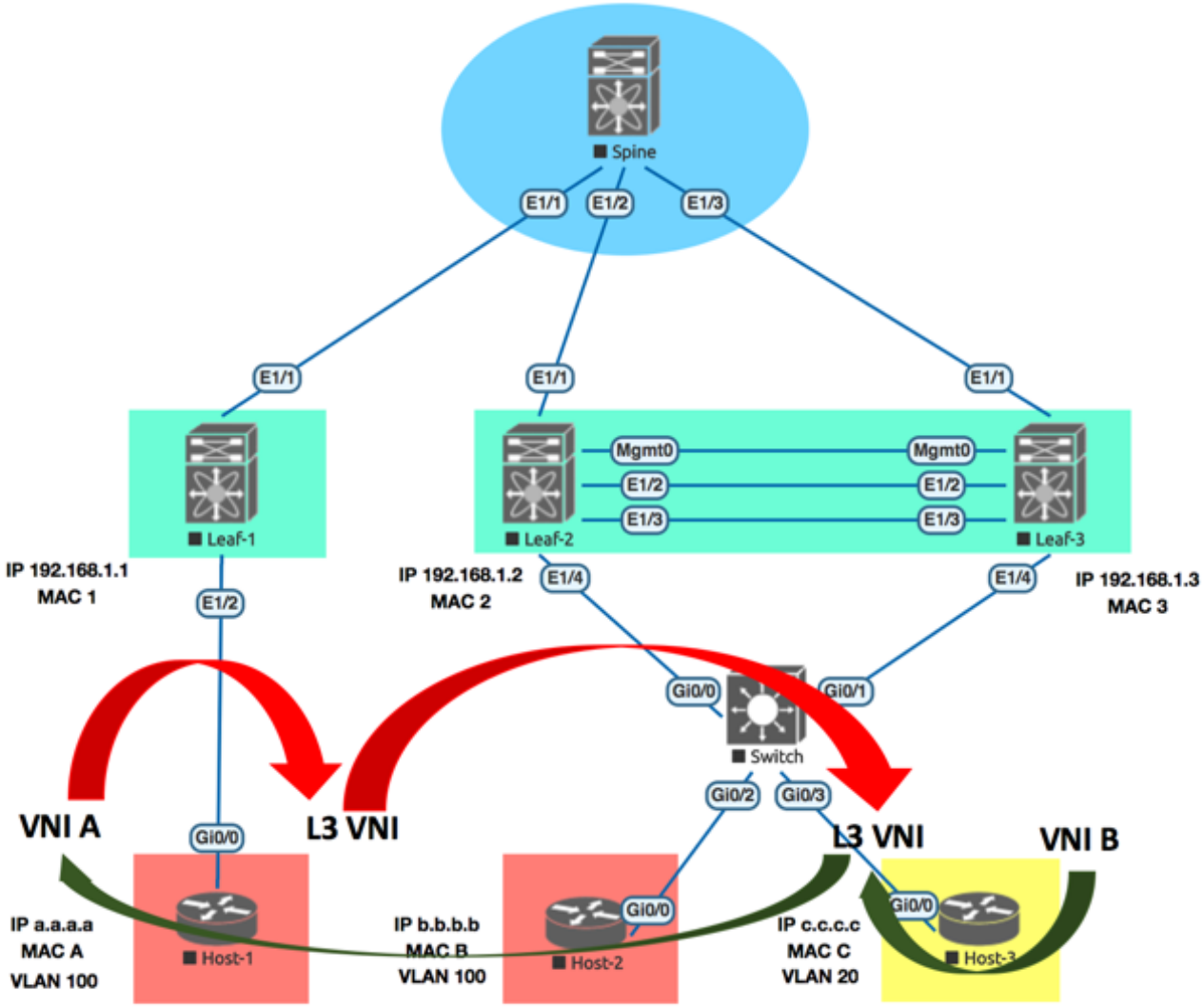
The VTEP is not required to be configured with all the VNI.

The symmetric IRB used the same path from the source to the destination and on the way back as well.

In this method the ingress VTEP routes packets form source VNI to L3 VNI where the destination MAC address in the inner header is rewritten to egress VTEP router MAC address.

On the egress side, the egress VTEP decapsulated the packet and looks at the inner packet header. Since the destination MAC address of the inner header is its own router MAC address, it performs the Layer 3 routing lookup.

Because the layer 3 VNI (in the VXLAN) provide the VRF context lookup, the packet are routed to the destination VNI and VLAN.



Step 1.Host 1 in VNI A sends a packet towards VNI B with the source MAC address of Host 1 and the destination MAC address set to gateway MAC address set to gateway MAC.

Step 2.Ingress VTEP routes packets form source VNI to L3 VNI where the destination MAC address in the inner header is rewritten to egress VTEP router MAC address.

Step 3. On the egress side, the egress VTEP decapsulated the packet and looks at the inner packet header. Since the destination MAC address of the inner header is its own router MAC address, it performs the Layer 3 routing lookup.

Step 4. Because the layer 3 VNI (in the VXLAN) provides the VRF context lookup, the packet are routed to the destination VNI and VLAN.

How VXLAN works?

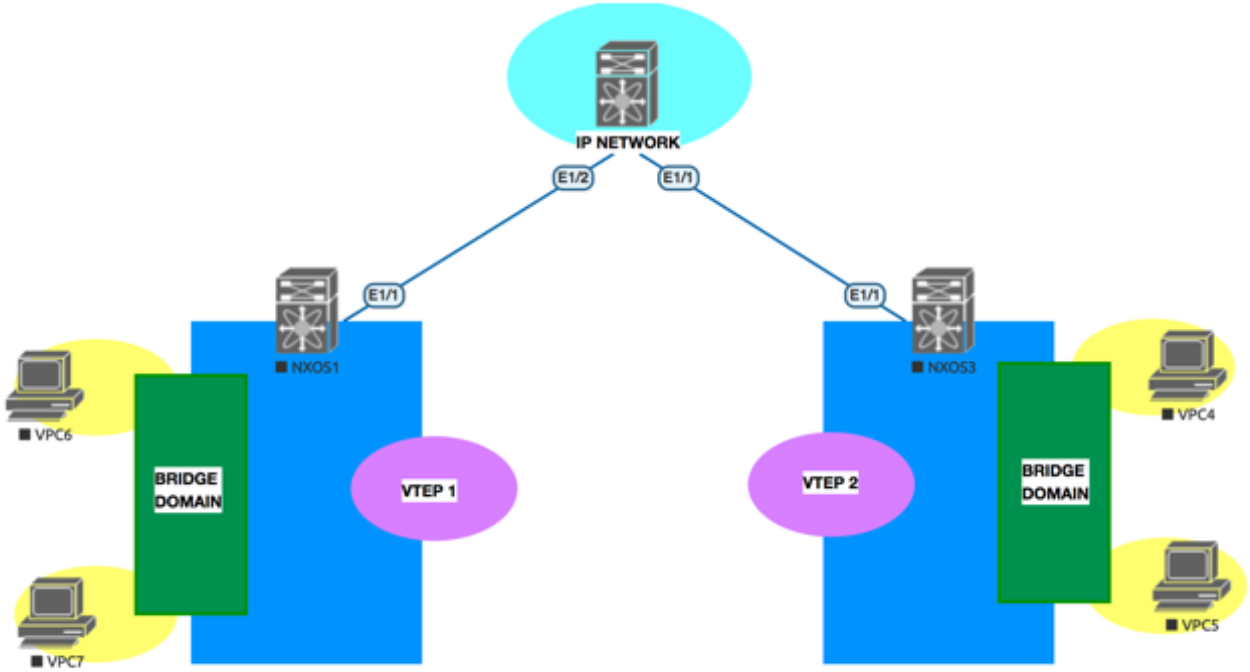
The VXLAN draft defines the VXLAN Tunnel End Point (VTEP) which contains all the functionality needed to provide ethernet Layer 2 services to connected end systems.

VTEP are intended to be at the edge of the network, typically connecting an access switch (virtual or physical) to an IP transport network.It is expected that the VTEP functionality would be built into the access switch, but it is logically separate from the the access switch.

Each end system connected to the same access switch communication through the access switch. The access switch acts as any learning bridge does, by flooding out its ports when it doesn't know the destination MAC or send out a single port when it has learned which direction leads to the end station as determined by source MAC learning.

Broadcast traffic is sent out all ports.

Further the access switch can support multiple bridge domain which are typically identified as VXLAN with as associated VLAN ID that is carried in the 802.1Q header on trunk port. In case of VXLAN enabled switch, the bridge domain would instead by associated with a VXLAN ID.



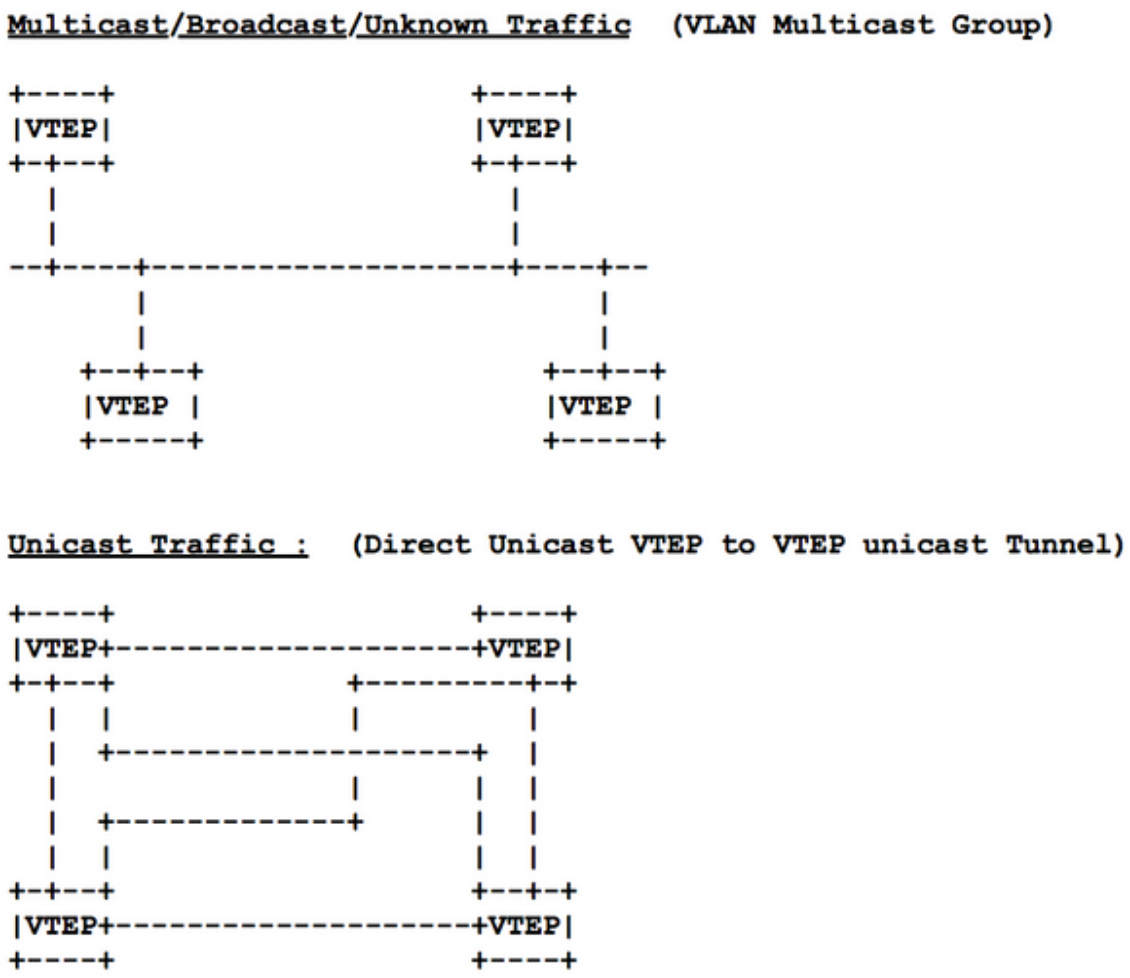
Each VXLAN has two interfaces.

One is a bridge domain trunk port to the access switch, and the other is an IP interface to the IP network.

The VTEP behaves as in IP host to the IP network. It is configured with an IP address based on the subnet its OP interface is connected to. The VTEP uses this IP interface to exchange IP packets carrying the encapsulated Ethernet Frame with other VTEPs.

A VTEP also acts as an IP host by using the IGMP to join IP multicast group.

In addition to a VXLAN ID to be carried over the IP interface between VTEP, each VXLAN is associated with an IP multicast group. The IP multicast group is used as communication bus between each VTEP to carry broadcast, multicast and unknown unicast frames to VTEP participating in the VXLAN.



The VTEP function also works the same way as a learning bridge, in that if it doesn't know where a given destination MAC is, it floods the frame, but it performs this flooding function and sends the frame to the VXLAN associated multicast group.

The VTEP function also work the same way as a learning bridge, in that if it does not know where the destination MAC is, it floods the frame, but it performs this flooding function and sends the frame to the VXLAN associated multicast group. Learning is similar except of learning the source interface associated with a frame source MAC, it learns the encapsulation source IP address. Once it has learned this MAC to remote IP associated, frames can be encapsulated within a unicast IP packet directly to the destination VTEP.

The initial use case for VXLAN enabled access switches are for access switches connected to the end system VM. These SW are tightly integrated with the hypervisor.

One benefit of this tight integration is that the virtual access switch knows exactly when a VM connect to or disconnect form the switch, and what VXLAN the VM is connected to, using this information, the VTEP can decide when to join or leave a VXLAN multicast group. When the first VM connects to a given VXLAN the VTEP joins the multicast group and starts to receive broadcast /multicast/ floods over that group.

Similarly, when the last VM connected to a VXLAN disconnects, the VTEP can see the IGMP leave the multicast group and also, it stops to receive traffic for the VXLAN which has no local receiver.

Note: Because of the potential number of VXLAN, (16M) could exceed the amount of

multicast state supported by IP network multiple and VXLAN could potentially map to the same IP multicast group.

While this could result in VXLAN traffic being sent needlessly to a VTEP that has no need systems connected to that VXLAN, inter VXLAN traffic isolation is still maintained.

The same VXLAN ID is carried in multicast encapsulated packets as is carried in unicast encapsulated packets. It is not the job of the IP Network to keep the traffic to the end system isolated, but the VTEP. Only the VTEP inserts and interprets/removes the VXLAN header within the IP/UDP payload. The IP network simply sees IP packets that carry UDP traffic with a well known dest UDP port.

Introduction to MP-BGP (EVPN)

Ethernet VPN introduces the concept of BGP MAC routing.

It uses MP-BGP for learning MAC addresses between provider edges.

Learning between PE and CE is still done in the data plane.

The BGP Control Plane has the advantage of scalability and flexibility for MAC routing, just as it does for IP routing.

EVPN provides seperation between the data plane and control plane, which allows it to use different encapsulation mechanism in the data plane while maintaining the same control plane.

IANA has allocated EVPN a new NLRI with an AFI of 25 and SAFI of 70.

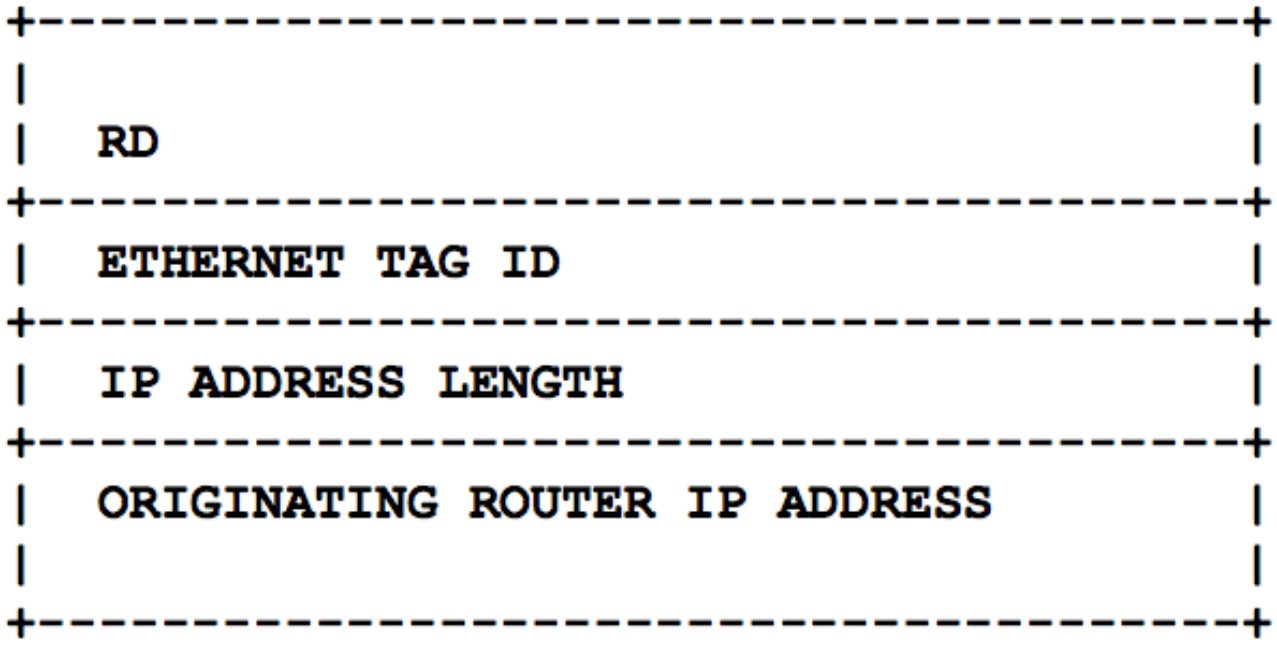
EVPN/PBB-EVPN introduces four new BGP route types and communities.

Various components are involved as part of BGP EVPN control Plane, these work together to implement the VXLAN functionality with the use of the control plane learning and discovery mechanism.

MP BGP plays an important role with the VXLAN BGP EVPN feature. The router distribution is carried out via MP-iBGP update message in the L2VPN EVPN family.

Generally MP-BGP (EVPN) uses route type 2 to advertise MAC and MAC+IP information of the hosts and router type 3 to carry the VTEP information.

The BGP EVPN overlay specifies the distribution and discovery of VTEP with the use of EVPN. The information is carried as EVPN Inclusive Multicast (IM) NLRI.



Encoding of the IM NLRI is based on Single Virtual Identifier per EVI, whereas the VPNID is mapped to a unique Ethernet VPN instance (EVI).

RD: Route Distingusher for the EVPN instance

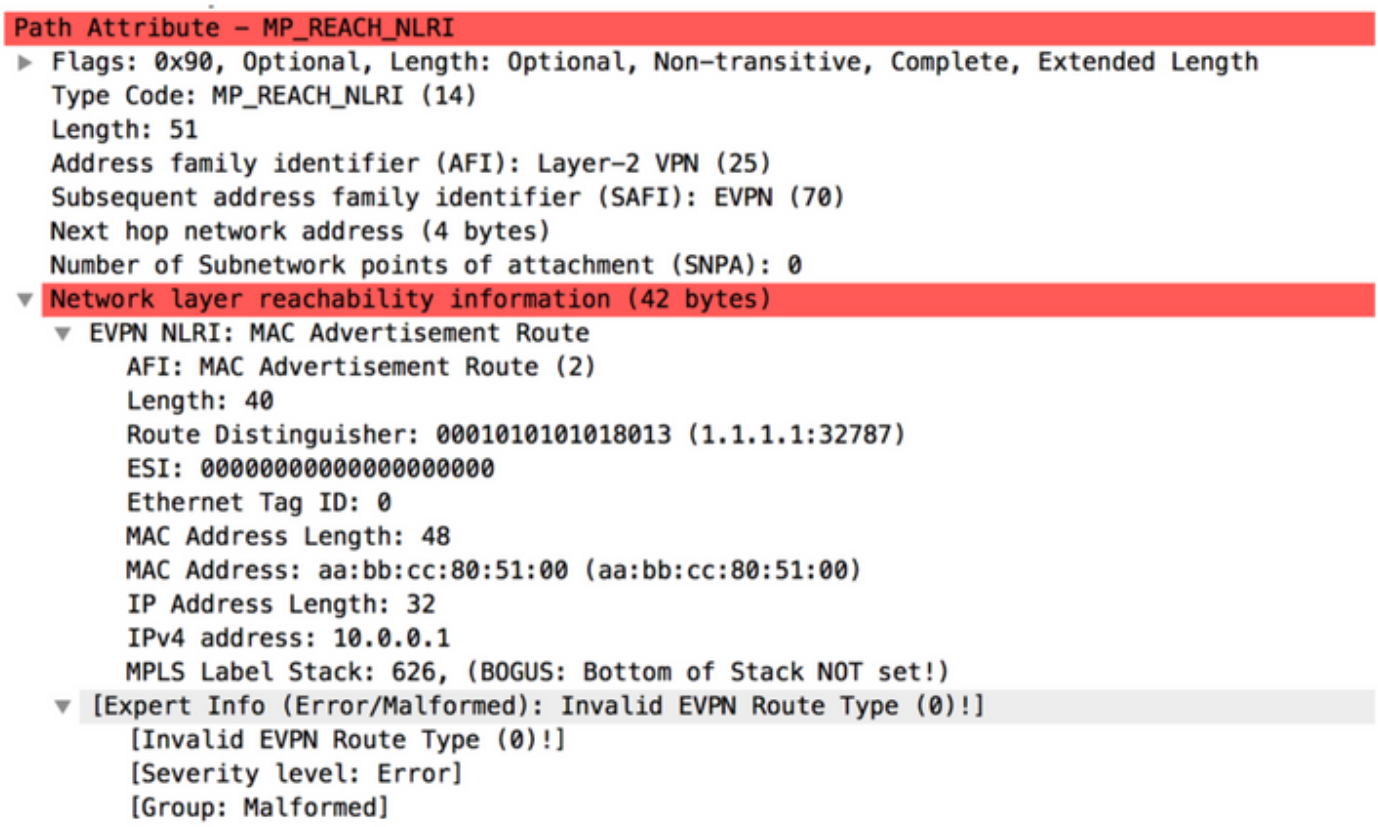
Ethernet Tag ID: VNI for the Bridge Domain

IP address Length: 1 Byte

Originating Routers IP address: VTEP IP address of the advertising endpoint

Advertisement and learning of IP host address associated with a VTEP is accomplished via BGP EVPN MAC advertisement NLRI.

The VTEP information is implicitly sent as the BGP Next hop associated with the IP host and also by providing the VTEP gateway MAC address in the MAC advertisement NLRI.



The RT value is manually configured or auto generated which is based on a 2 Bytes AS Number and the VNI value.

The route is imported into the correct VLAN or bridge domain based on the import route target configuration.

The design for the VXLAN deployment follows the spine and lead architecture. With VXLAN BGP EVPN solution, the spine nodes are usually configured as the RR and it only requires the nv overlayfeature to be enabled along with BGP.

The leaf nodes on the other hand require the nv overlay feature along with the vn-segment-vlan-based feature to be enabled.

The vn-segment-vlan-based feature is required to map the VLAN to the VNI.

BGB Route Type

Type 1.

Route Type: Ethernet Auto-Discovery Route

Usage: MAC MASS Withdraw, Aliasing, Advertising Split Horizon Labels

BGP Community: ESI MPLS Label Extended Community

In case of a Multi-homed CE device.

Route Type 1: Ethernet Auto Discovery Routes

Ethernet Auto-Discovery (A-D) routes are type 1 mandatory routes and are used for achieving split

horizon, fast convergence and aliasing.

Only EVPN uses Type 1 routes, PBB-EVPN uses B Mac in order to achieve the same function.

Multi-homed PE advertises an auto discovery route per Ethernet Segment with the newly introduced ESI MPLS label extended community.

- PE recognise other PE connected to the same Ethernet segment after the Type 4 E-S route exchange.
- All the multi-homed and remote PE routers that are part of the EVI import the auto discovery route.

All the multi-homed and remote PE routers that are part of the EVI import the auto discovery route.

The Ethernet AD route is not needed when ESI=0.

Example; When CE is single-homed, the ESI label extended community has an 8 Bit flag which indicates Single-Active or All-Active redundancy mode.

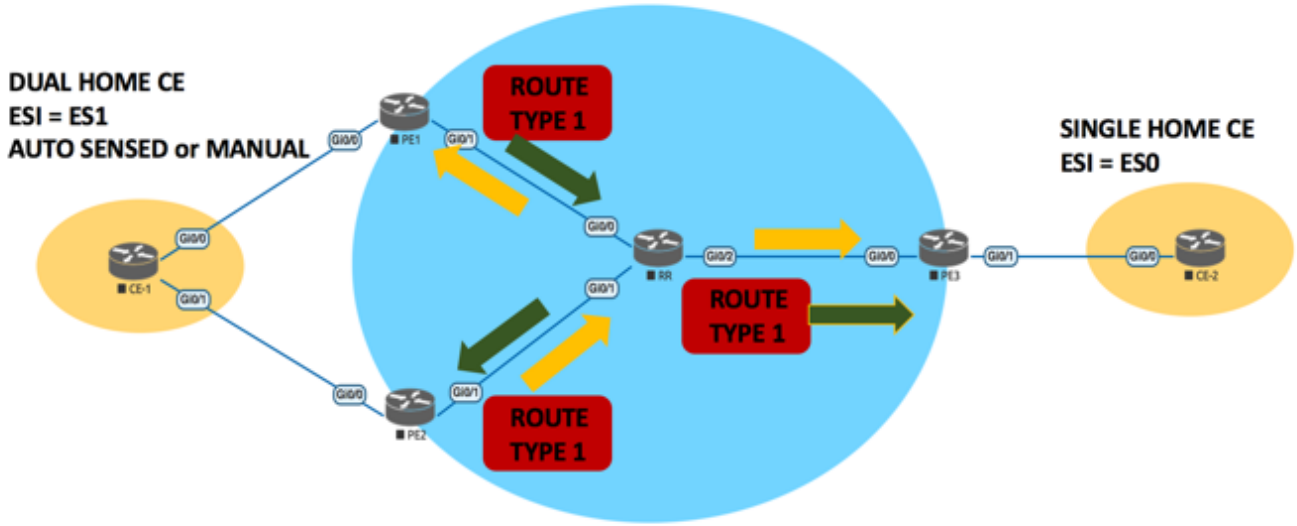
Split Horizon:

CE1 sends a BUM frame to a non-DF PE, lets say PE1 forwards the traffic to all other PE in the EVPN instance that includes the DF PE.

PE2 in this example.

PE2 must drop the packet and it cannot forward it to CE1. This is referred to as Spilt Horizon.

The ESI label is distributed by all the PE operating in A-S and A-A mode with the use of the Ethernet A-D route as ES. Ethernet A-D routes are imported by all PE that participate in the EVPN instance.



Type 2.

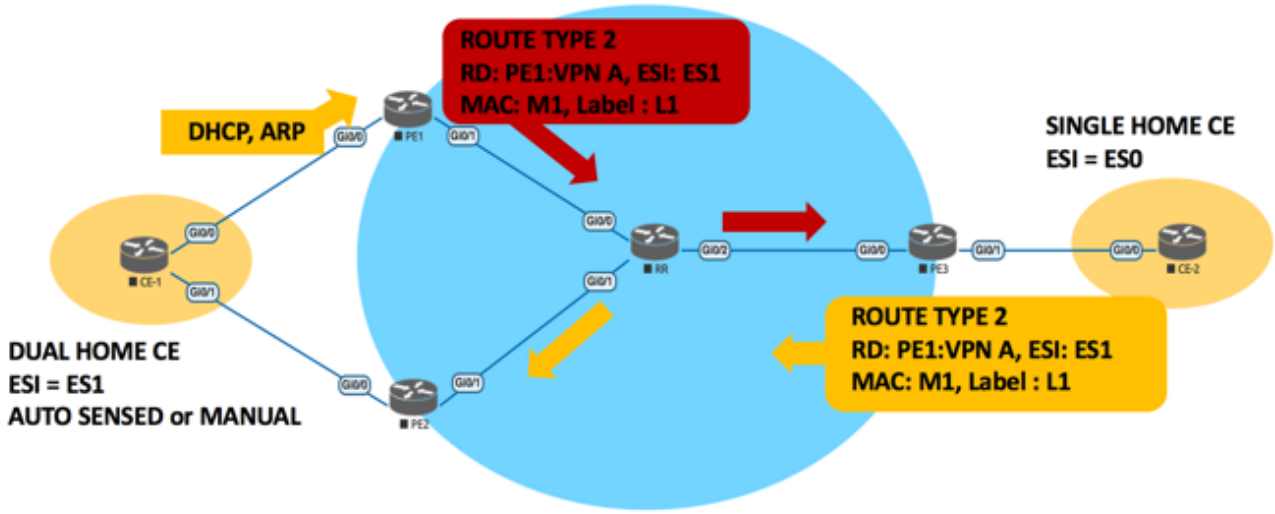
Route Type: MAC Advertisement Route

Usage: Advertising MAC Address reachability, Advertise IP/MAC bindings

BGP Community: MAC Mobility extended community, Default Gateway Extended Community

This is responsible for MAC advertisement routes which are responsible for advertising MAC address reachability via MP-BGP to all other PE in a given EVPN instance. MAC advertisement routes are Type 2 routes.

Here, learning the PE-CE is in the Data Plane, once PE1 learns the MAC of CE1, it advertises it to the other PE's through the BGP NLRI with the use of MAC advertisement route which contains RD, ESI (which could be zero or non zero value for multi homes cases), MAC address,NPS label associated with MAC and the IP address field which is
optional.



Per EVI Label Assignment: This is similar to Per VRF label allocation mode in IP world. A PE advertises single EVPN label for all the MAC addresses in a given EVI instance. Obviously, this is the most conservative way of allocating labels and the tradeoff is similar to per-VRF label assignment. This method required an additional lookup on the egress PE.

Per MAC Address Label Assignment: This is similar to per-prefix label allocation mode in IP. A PE advertise unique EVPN labels for every MAC address This is the most liberal way of allocating labels and the tradeoff is memory consumption and the possibility of running out label space.

Type 3.

Route Type: Inclusive Multicast Route

Usage: Multicast Tunnel End Point Discovery

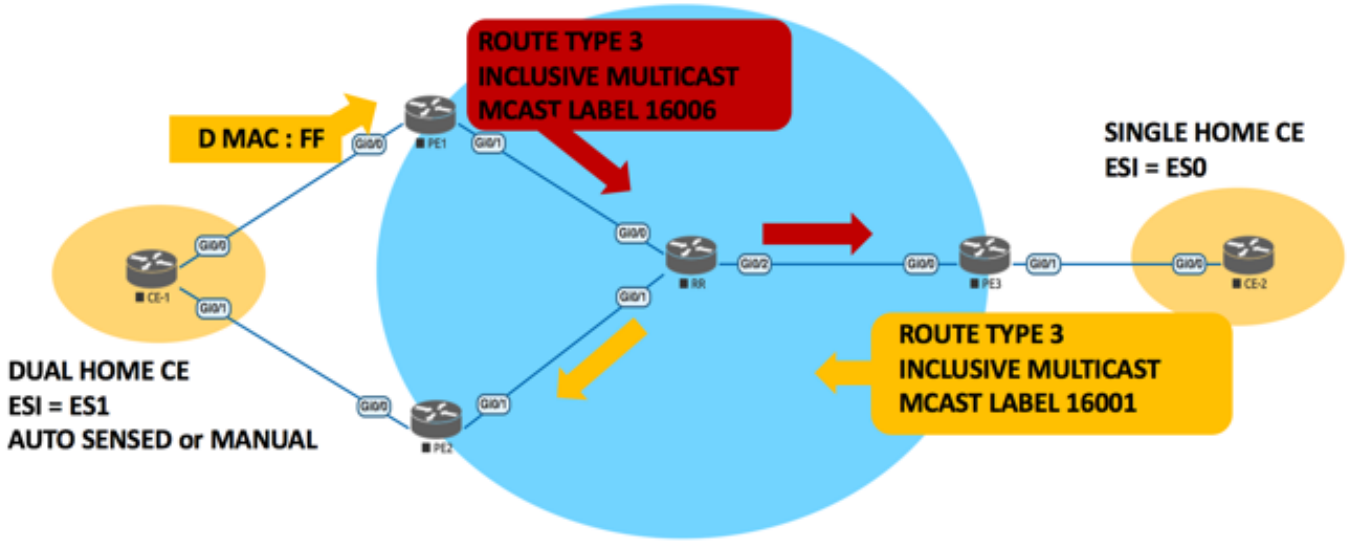
When you send BUM frames, PE can use ingress replication, P2MP or MP2MP (mLDP) LSP.

Every participating in an EVI advertises its mcast labels at the time of its startup sequence via inclusive Multicast Routes.

Inclusive Multicast Routes are BGP Route Type 3.

Once a PE has received mcast routes from all the other PE and a BUM frame arrives the PE will

do ingress replication by attaching PE Mcast label.



In these details, PE1 label 16006 and PE3 Label 16001 advertise their multicast label to PE3.

When PE2 receives a broadcast packet, it adds the mcast label 16001 + the label to reach PE3 and sends the packet to PE3.

PE2 also forwards the packet to PE2 and adds the ESI label + Label 16001 + label to reach PE1.

PE3 receives the packet and sees the mcast label, it treats the packet as a BUM frame. When PE1 receives the packet, it notices the ESI label which was advertised as part of Ethernet AD route and drops the packet.

Type 4.

Route Type: Ethernet Segment Route

Usage: Redundancy group discovery, DF election

BGP Community: ES-Import extended community

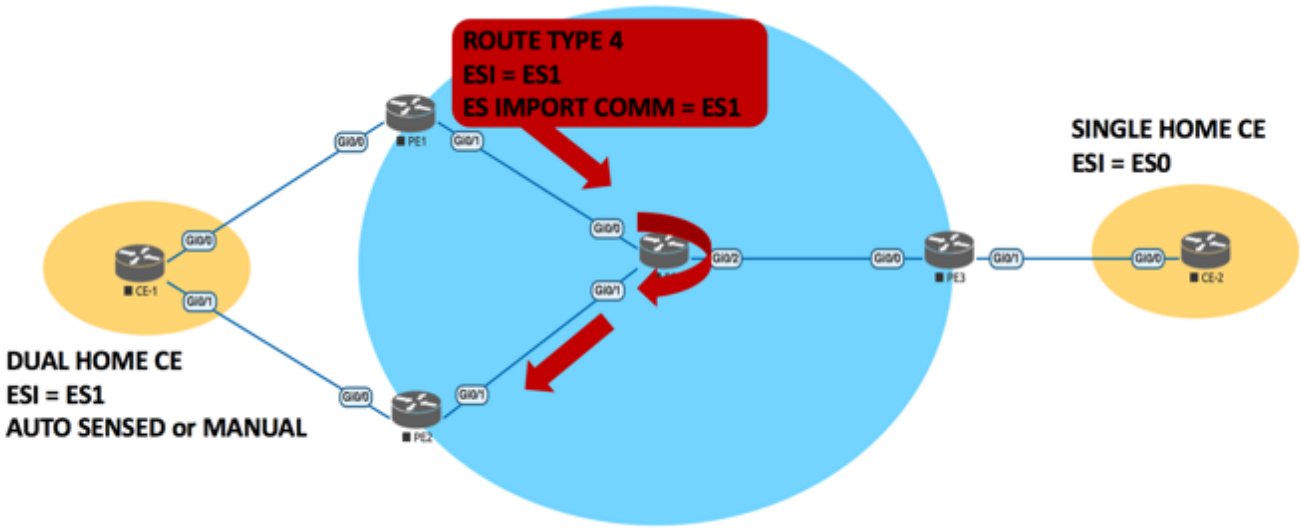
In case of multi homed CE device, a set of ethernet links comprises an Ethernet Segment. A unique ethernet segment identifier (ESI) number identifies this ethernet segment, which can be manually configured or automatically derived.

When a single homes CE is attached to an Ethernet segment, the ESI value is zero.

Route (BGP Route Type 4) with newly introduced ES-import extended community (=ESI value) along with the extended community.

All the PE automatically imports the route if their ESI value matches ESI Import community.

This process is also referred to as auto-discovery and allows PE connected to the same ethernet segment to auto discover each other.



PE2 and PE1 have the same EVI value (ES=1); PE1 advertises its ESI value in the ethernet segment route with ES-Import Community set to ES1.

PE2 and PE3 receives the route but only PE2 will import this route, since it has a Matching ESI value.

This ensure PE2 knows that PE1 is connected to the same CE device.

After Auto Discovery the Designated Forwarder (DF) election happens for Multi homes CE.

The PE which assumes the roles of DF, is responsible for forwarding BUM frames on a given segment to CE.

The DF election happens by the PE first building an ordered list of IP addresses of all the PE nodes in ascending order.

For example:

```
PE1 : 1.1.1.1

PE2 : 2.2.2.2


Position PE

0     PE1 1.1.1.1

1 PE2 2.2.2.2


Ethernet TAG Value Ethernet TAG ID

300 0

301 1
```

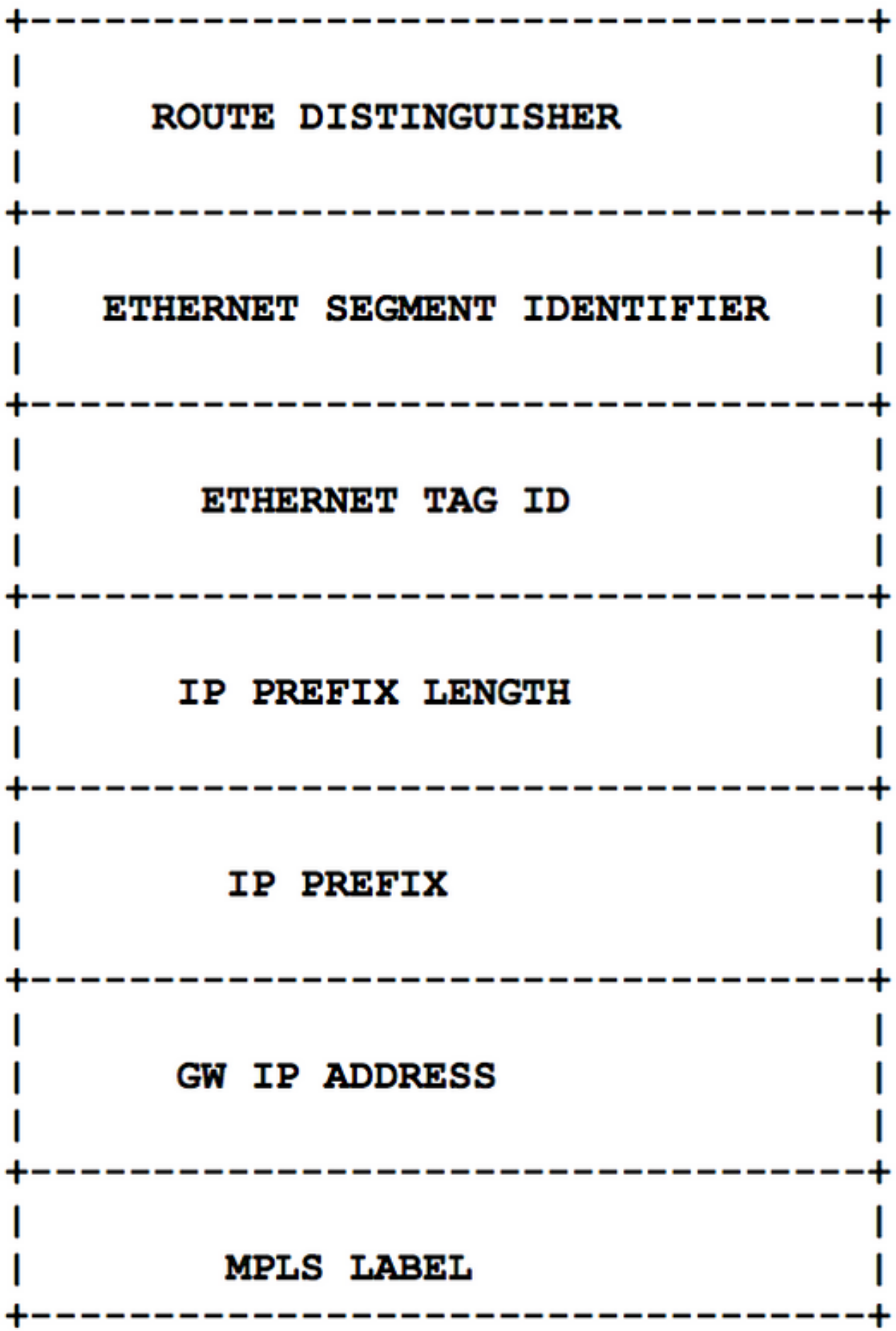
PE1 becomes DF for Ethernet tag 300 and PE2 becomes DF for Ethernet Tag 301

Type 5.

Route Type: IP PREFIX ADVERTISEMENT IN EVPN

It's a mechanism to carry IPv4 and IPv6 advertisement in EVPN only networks.

While EVPN Type 2 route allows to carry both MAC addresses and IP addresses, tight coupling of specific IP address with IP prefixes might not be desirable of the draft discusses different scenarios where such coupling is not desirable.



GW IP Address: Will be 32 or 128 bit field and will encode an overlay IP index for the IP prefix. The GW IP field should Zero it, it is not used as an overlay index.

MPLS Label: The MPLS label field is encoded as 3 octet where the high order 20 contain the label value. This should be null when the IP prefix route used for recursive lookup resolution.

Prefix Advertisement draft introduces the concept of overlay index. When an overlay index is present in the Route Type 5 advertisement, the receiving NVE PE needs to be performed to a recursive route resolution to find out to which egress NVE (PE) to forward the packet.

The route will contain a single overlay index at most. If the ESI field is different from Zero.

Reference: https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-05#page-7

VXLAN over EVPN Packet Flow

Understand the various components of Nexus Architecture.

VXLAN Manager: VXLAN Manager is the VXLAN control and management plane component that is responsible for VXLAN Local Tunnel Endpoint configuration, remote endpoint learning, management of ARP suppression and Platform Dependent Program.

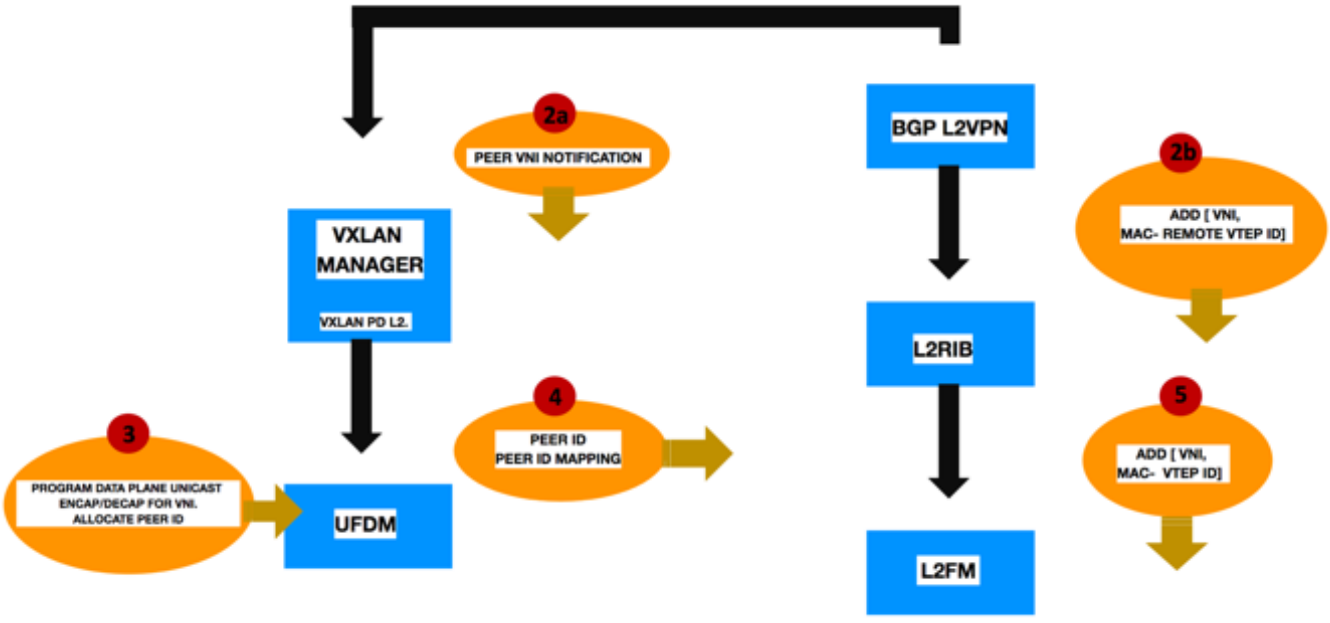
L2RIB: The L2RIB component manages the Layer 2 routing information. The L2RIB component interacts with VXLAN Manager, BGP, Layer2 Forwarding Manager (L2FM), ARP and Multicast forwarding Distribution Manager (MFDM).

MFIB: Multicast Distribution Information Base finds out all VXLAN VNI that share multicast group and program encapsulation/decapsulation entries for each of the VNI when the VTEP interface is in Outgoing Interface List (OIL) for a group.

Adjacency Manager (AM) performs tow tasks:

Host IP and MAC bindings for locally learned hosts.

Programs Routing Information Base and Forwarding Information Base for host route and adjacency binding, respectively.



Packet Forwarding Based on the Hardware

Step 1.The Host sends ARP request to the remote host. The MAC address is learned on the VLAN for the local host. The MAC address information is sent to the Layer 2 Forwarding Manager (L2FM) component of the system. The information can be viewed with the use of the command **show system internal l2fm event-history debugs | include mac-address**.

The MAC-Address variable is the MAC address of the locally attached host.

Step 2. The L2FM component then sends a notification about the L2VNI and MAC address to the Layer 2 Routing Information Base (L2RIB). The information on the L2RIB that is received from L2FM is viewed with the use of the command **show system internal l2rib event-history mac | i mac-address**.

Step 3. The L2RIB then sends the L2 VNI and MAC address information to BGP L2VPN component, which is then advertised to the remote VTEP. BGP builds the L2 NLRI information with the local host MAC received.

Prefix

MAC: HOST MAC ADDRESS

Label2: L2 VNI-ID

BD-RT: Configured RT

NH: VTEP-IP

The Type 2 NLRI is then sent to the remote VTEP as part of the update.

Step 4. When you receive those updates, the remote VTEP stores the information in the L2VPN EVPN table. Apart from viewing the information on the BGP EVPN table, the route import on the remote VTEP is verified with the use of the command **show bgp internal event-history events | i mac-address**.

The command is executed on the remote VTEP and mac-address of the host attached to the Local VTEP can be used as a filter option.

Step 5. The BGP process on the remote VTEP sends peer information and VNI notification to be VXLAN Manager. This information is verified with the use of the command **show eve bgp rnh database** and also from the event history logs with the use of the command **show eve internal event-history event**. It also adds VNI and the MAC address of the remote host learned from VTEP withe next hop set to Local VTEP.

Step 6.The VXLAN Manager then programs the hardware that is the data plane, and also allocates the Peer ID, which is then sent to L2RIB. The information from VXLAN manager is sent to the UnicastRIB/Unicast Forwarding Distribution Manager (UFDM) process which is used to program the FIB. The forwarding information can be viewed with the use of the command **show forwarding eve l3 peers** and the command **show forwarding nve l3 adjacency tunnel tunnel-id** where the tunnel-id is received from the first command.

The L2RIB on the other hand adds the VNI and the MAC address information in the L2FM table

which contains an entry that consists of the MAC address and the next-hop peer ID (Remote VTEP).

Advertise and Install L3 VNI Route

Step 1. A host attached to the VTEP sends an ARP request. The VTEP received the request and updated in the ARP table for the VLAN.

Step 2. After the ARP table is updated, the information is passed onto an AM, which installs the adjacency for the local host. This is viewed with the use of the command **show forwarding vrf vrf-name adjacency**.

Step 3. The AM then sends an adjacency notification to the Host Mobility Manager (HMM) with the MAC+IP also know as the combo routes.

Because the host MAC needs to be carried in a BGP update along with the host IP, HMM publishes the combo route into L2RIB.

Use the command in order to check **show system internal l2rib event-history mac-ip** in order to view the combo route in L2RIB.

Step 4. The L2RIB then sends the combo route along with the L3 VNI to BGP.

The BGP uses the information to prepare the L2+L3 NLRI which consists of:

Prefix: Host IP

MAC: Host MAC

Label 1: L3VNI

Label 2: L2VNI

VRF-RT

BD-DT

NH-VTEP-IP

Remote Next hop (RNH): Remote MAC (RMAC)

The L2 + L3 NLRI is sent as an update to the remote VTEP.

Step 5. The update received on the remote node is viewed with the use of the command **show bgp l2vpn even ip-address**. The IP address is of the host connected to VTEP node. The BGP update then encapsulated in the VXLAN, the value is 8.

BGP on receiving an update on remote VTEP, updates two components. First, it updates the URIB with VRF, Host-IP, L3 VNI and VTEP-IP information.

Second, it updates the VXLAN manager with Peer Information and VNI and RMAC notification.

Step 6. The information in the URIB is used by UFDM along with the information from VXLAN manager which programs the data place with the encapsulation and decapsulation information for the L2 VNI. The VXLAN manager also sends RMAC to UFDM and allocates the peer ID.

Step 7. VXLAN manager on the other hand sends the Peer-ID notification to the L2RIB. At this time, the VNI is set with the Next Hop of the Peer-ID.

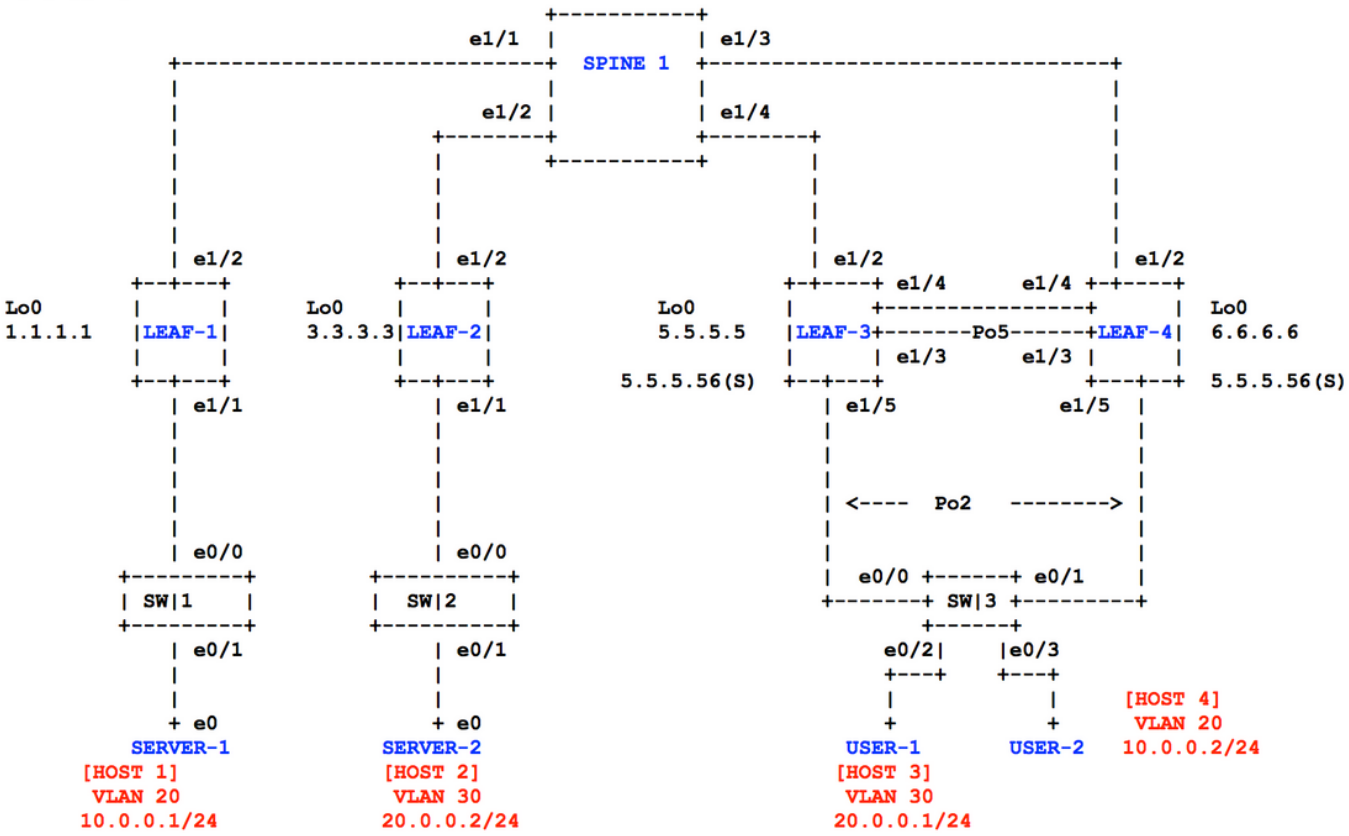
Step 8. The L2RIB then updates the L2FM in order to update the MAC address table.

Configure

Network Diagram

LAB REPLICATION DETAILS :

TOPOLOGY :



Configurations

> **Note**: The sequence of configuration is mandatory in order to make VXLAN work.

Step 1. Perform the initial configuration of each VTEP switch.

Enable the VXLAN and MP-BGP EVPN Control Plane.

```
feature nv overlay  ———> Enable VXLAN

feature vn-segment-vlan-based -——> Enabled VLAN based VXLAN (Currently the only mode)

feature bgp ———> Enable BGP

nv overlay evpn ——-> Enable the EVPN control plane for VxLAN
```

Other features might need to be enabled.

```
feature ospf  ——-> Enable OSPF if its choose as the underlay IGP routing protocol

feature pim   -——> Enable IP protocol-independent Multicast (PIM) routing

feature interface-vlan          ——> Enabled VLAN switch virtual interface (SVI) if the VTEP
needs to be the IP gateway and route for the VxLAN VLAN IP packets.
```

Step 2. Configure the EVPN tenant VRF instance.

```
vrf context ONE    ———> Create a VxLAN tenant VRF instance

 vni 30001 ———> Specify the Layer 3 VNI for VxLAN routing for this tenant VRF instance

 rd auto  ———> VRF Route Distinguisher

 address-family ipv4 unicast

   route-target import 64522:30001

   route-target import 64522:30001 evpn ———-> Defined the VRF route target import and export
policies in address-family ipv4 unicast

   route-target export 64522:30001       This is Manually configured Route Target. We can also
create Auto RD and RT.

   route-target export 64522:30001 evpn ———> Route Target Export Manually configured.
```

Step 3. Create a Layer-3 VNI for each tenant VRF instance.

```
vlan 3901                 ———> Create the VLAN for the Layer 3 VNI, create one Layer3 VNI for
each tenant VRF routing instance

name ONE

 vn-segment 30001 ———> Define the layer 3 VNI


interface Vlan3901

 no shutdown

 vrf member ONE     ———> Create the SVI for the Layer 3 VNI. Put this SVI in the tenant VRF
context.

 no ip redirects           The command “IP Forward” enables prefix-based routing for the VNI ip
subnet. Its needed to complete the initial routing to silent hosts in the

 ip forward       VNI network


vrf context ONE
```

```
 vni 30001     ———> Associate the Tenant VRF routing instance

 rd auto

 address-family ipv4 unicast

   route-target import 64522:30001

   route-target import 64522:30001 evpn

   route-target export 64522:30001

   route-target export 64522:30001 evpn
```

Step 4. Configure EVPN Layer-2 VNIs for Layer-2 networks.

This step involves how to map VLANs to Layer-2 VNIs and how to define their EVPN parameters.

```
vlan 20

 vn-segment 10020 ——> Map the VLAN to the VxLAN VNI



evpn ——> Under the EVPN configuration, define the route distinguisher and route target import
and export policies for each Layer 2 VNI

 vni 10020 l2

   rd auto

   route-target import 64522:10021

   route-target export 64522:10021
```

Step 5. Configure the SVI for Layer-2 VNIs and enable the anycast gateway under the SVI.

This step includes how to configure the anycast gateway virtual MAC address for each VTEP and the anycast gateway IP address for each VNI. All the VTEPs in the EVPN domain must have the same anycast gateway virtual MAC address and the same anycast gateway IP address for a given VNI for which they function as the default IP gateway.

fabric forwarding anycast-gateway-mac 0000.2222.3333 ————> Configure the distributed Virtual MAC address. Configure one Virtual MAC address per VTEP.

The any cast gateway MAC address must be same on all the switches that are part of distributed gateway.

Note: Create a SVI for a Layer 2 VNI. Associate twitch the tenant VRF instance.

All the VTEPs for this VLAN and VNI should have the same SVI IP address as the distributed IP gateway.

Enable the distributed any cast gateway for the VLAN and VNI.

```
interface Vlan20
```

```
no shutdown

  vrf member ONE           ———> Configured the virtual IP address

  ip address 10.0.0.3/24                    All VTEP for this VLAN must be the same virtual IP
address

  fabric forwarding mode anycast-gateway

   |

   |
```

Enable the distributed gateway for this VLAN.

Step 6. Configure VXLAN tunnel interface nve1 and associate Layer-2 VNIs and Layer-3 VNIs with it.

```
interface nve1

 no shutdown

 source-interface loopback0             ———> Specify loopback0 as the source interface

host-reachability protocol bgp           ———> Define BGP as the mechanism for host reachability
advertisement

source-interface hold-down-time 600

member vni 10020

   mcast-group 239.0.0.1       ———> Associate the Multicast group

member vni 30001 associate-vrf       ———> Add Layer 3 VNI one per tenant VRF
```

> **Note**: Also, you can configure Suppression-arp under the Layer 2 VNI. If the VNI is configured with SVI, only then ARP suppression will work.

```
interface loopback0       ———> This is the loopback interface to the source VxLAN tunnels

description VTEP Source Interface

  ip address 6.6.6.6/32           ———> Source interface Loopback for the NVE interface

  ip address 5.5.5.56/32 secondary

  ip router ospf UNDERLAY area 0.0.0.0

  ip pim sparse-mode
```

> **Note**: Secondary Loopback is only required when you have Redundancy for VTEP. Or VPC configured between VTEP. Then the VTEP address will be taken from the secondary address.

Step 7. Configure on MP-BGP on the VTEP and Configure iBGP route reflector in Spine.

Configure VTEP BGP:

```
router bgp 100

  router-id 111.111.111.111

 log-neighbor-changes

 address-family ipv4 unicast                 ———> Use address family ipv4 unicast for prefix
based routing

    nexthop trigger-delay critical 250 non-critical 1000

  address-family l2vpn evpn                                    ———> Use address family l2vpn
even for even host routes

    nexthop trigger-delay critical 250 non-critical 1000

  template peer spine-peer

    remote-as 100

    update-source loopback2

    address-family ipv4 unicast

      send-community

      send-community extended

      soft-reconfiguration inbound always

    address-family l2vpn evpn

      send-community

      send-community extended               ———> Send extended community in address-family
l2vpn even to distribute EVPN route attributes

      soft-reconfiguration inbound always

 neighbor 22.22.22.22                ———> Define the MP-BGP neighbours, under each neighbor,
define address-family ipv4 unicast and l2vpn evpn

    inherit peer spine-peer

    no shutdown

  vrf ONE                        ———> Under address family ipv4 unicast for each tenant VRF
instance, enable advertising for EVPN routes

    address-family ipv4 unicast

      advertise l2vpn evpn
```

Spine iBGP Configuration as RR:

```
router bgp 100

address-family ipv4 unicast               ———> Use address family ipv4 unicast for prefix based
routing
```

```
address-family l2vpn evpn

retain route-target all               ———> Use address-family l2vpn for EVPN VxLAN host routes.
Retain all the route-target attributes

template peer vtep-peer                    ———> Use iBGP RR client peer template

remote-as 100

update-source loopback0

address-family ipv4 unicast

send-community

send-community extended                 ———> Use both standard and extended communities in
address-family ipv4 unicast

route-reflector-client

soft-reconfiguration inbound always

address-family l2vpn evpn

send-community

send-community extended               ———> Send both standard and extended communities in
address-family l2vpn evpn

route-reflector-client

soft-reconfiguration inbound always

neighbor 1.1.1.1

inherit peer vtep-peer

no shutdown

neighbor 3.3.3.3

inherit peer vtep-peer

no shutdown



neighbor 5.5.5.5

inherit peer vtep-peer

no shutdown

neighbor 6.6.6.6

inherit peer vtep-peer

no shutdown
```

Verify and Troubleshoot

Use this section in order to confirm that your configuration works properly, also this section

provides information you can use in order to troubleshoot your configuration.

Verify Control Plane

```
Leaf-4# sh bgp l2vpn evpn

BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 7662, local router ID is 6.6.6.6
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-i
njected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 10020:100    (L2VNI 10020)
*>i[2]:[0]:[0]:[48]:[5000.0007.0000]:[0]:[0.0.0.0]/216
                      1.1.1.1                          100          0 i
*>l[2]:[0]:[0]:[48]:[aabb.cc00.d000]:[0]:[0.0.0.0]/216
                      5.5.5.56                         100      32768 i >>>>>>>>>>>>>>>>>>>>>>
LOCALLY LEARNT MAC ADDRESS
*>i[2]:[0]:[0]:[48]:[aabb.cc80.5000]:[0]:[0.0.0.0]/216
                      1.1.1.1                          100          0 i >>>>>>>>>>>>>>>>>>>>>>
Learnt from Leaf#1 Layer 2 information only (MAC INFO)
*>l[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/216
                      5.5.5.56                         100      32768 i
*>i[2]:[0]:[0]:[48]:[aabb.cc80.5000]:[32]:[10.0.0.1]/272
                      1.1.1.1                          100          0 i >>>>>>>>>>>>>>>>>>>>>>
Learnt from Leaf#1 layer 3 information (MAC-IP INFO)
*>l[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[32]:[10.0.0.2]/272
                      5.5.5.56                         100      32768 i

Route Distinguisher: 10030:100    (L2VNI 10030)
*>i[2]:[0]:[0]:[48]:[aabb.cc80.6000]:[0]:[0.0.0.0]/216
                      3.3.3.3                          100          0 i
```

***>l[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/216**

5.5.5.56 100 32768 i

*>i[2]:[0]:[0]:[48]:[aabb.cc80.6000]:[32]:[20.0.0.2]/272

3.3.3.3 100 0 i

***>l[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[32]:[20.0.0.1]/272**

5.5.5.56 100 32768 i

Route Distinguisher: 10500:100 (L3VNI 10500)

*>i[2]:[0]:[0]:[48]:[aabb.cc80.5000]:[32]:[10.0.0.1]/272

1.1.1.1 100 0 i

*>i[2]:[0]:[0]:[48]:[aabb.cc80.6000]:[32]:[20.0.0.2]/272

3.3.3.3 100 0 i

Leaf-4# sh nve interface

Interface: nve1, State: Up, encapsulation: VXLAN

VPC Capability: VPC-VIP-Only [notified]

Local Router MAC: 5000.0009.0007

Host Learning Mode: Control-Plane

Source-Interface: loopback0 (primary: 6.6.6.6, secondary: 5.5.5.56)

Leaf-4# sh interface nve1

nve1 is up

admin state is up, Hardware: NVE

MTU 9216 bytes

Encapsulation VXLAN

Auto-mdix is turned off

RX

ucast: 0 pkts, 0 bytes - mcast: 0 pkts, 0 bytes

TX

ucast: 0 pkts, 0 bytes - mcast: 0 pkts, 0 bytes

Note: If NVE Interface is down, then a no shut is performed on the interface.

Local MAC Learning on Leaf 4.

L2FM

```
Leaf-4# sh mac address-table vlan 20

Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link,
        (T) - True, (F) - False, C - ControlPlane MAC
   VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
G   20     5000.0009.0007   static   -         F      F    sup-eth1(R)    >>>>>>>>>>>>> MAC
Learnt on VLAN 20
G   20     5000.000a.0007   static   -         F      F    vPC Peer-Link(R)


Leaf-4# sh system internal l2fm event-history debugs | i 5000.0009.0007


    [102] l2fm_pss_insert_stat_mac(1726): Trying to insert gwmac into FU_PSS_TYP
E_CONFIG vlan_id = 500 if_index = 0x90101f4 MAC: 5000.0009.0007 caller1 = 0x101f
c12d caller2 = 0x101fda97 caller3 = 0x102004d8
    [104] l2fm_pss_stat_sec_mac(1640): sec_flag = 0, vlan_id = 500 ifindex = 0x9
0101f4  MAC: 5000.0009.0007 delete 0usr_cfg = 0
    [102] l2fm_macdb_delete(7301): Trying to delete an entry not present in MACD
B  5000.0009.0007
```

L2FM > L2RIB

```
Leaf-4# sh l2route evpn mac evi 20


Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
(Dup):Duplicate (Spl):Split (Rcv):Recv (AD):Auto-Delete(D):Del Pending (S):Stale
(C):Clear
(Ps):Peer Sync (O):Re-Originated
```

```
Topology    Mac Address    Prod   Flags         Seq No     Next-Hops
----------- -------------- ------ ------------- ---------- ----------------
20          aabb.cc00.d000 Local  L,            0          Po2
20          aabb.cc80.5000 BGP    SplRcv        0          1.1.1.1
20          aabb.cc80.b000 Local  L,            0          Po2
```

L2FM > L2RIB > BGP L2VPN

```
Leaf-4# sh bgp l2vpn evpn vni-id 10020
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 7737, local router ID is 6.6.6.6
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-i
njected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 10020:100    (L2VNI 10020)
*>l[2]:[0]:[0]:[48]:[aabb.cc00.d000]:[0]:[0.0.0.0]/216
                      5.5.5.56                           100      32768 i

*>l[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/216
                      5.5.5.56                           100      32768 i

*>l[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[32]:[10.0.0.2]/272
                      5.5.5.56                           100      32768 i

Leaf-4# sh bgp internal event-history events | i aabb.cc80.b000

2017 Aug  9 13:06:43.824949 bgp 100 [9604]: [9617]: (default) IMP: bgp_tbl_ctx_i
mport: 1971: [L2VPN EVPN] Importing 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:
[0]:[0.0.0.0]/112 to RD 10020:100
2017 Aug  9 13:06:43.824940 bgp 100 [9604]: [9617]: (default) IMP: [L2VPN EVPN]
```

```
Import of 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112 (EVI: 10
020) to RD 6.6.6.6:65534 (0) inhibited, not importing local paths
2017 Aug  9 13:06:43.824885 bgp 100 [9604]: [9617]: (default) IMP: bgp_tbl_ctx_i
mport: 1971: [L2VPN EVPN] Importing 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:
[0]:[0.0.0.0]/112 to RD 6.6.6.6:65534
2017 Aug  9 13:06:43.824513 bgp 100 [9604]: [9617]: (default) RIB: [L2VPN EVPN]
10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112 is not in rib, no
del
2017 Aug  9 13:06:43.824503 bgp 100 [9604]: [9617]: (default) RIB: [L2VPN EVPN]
For 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112, added 0 next
hops, suppress 0
2017 Aug  9 13:06:43.824453 bgp 100 [9604]: [9617]: (default) RIB: [L2VPN EVPN]
10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112 via 5.5.5.56 is lo
cal, no add
2017 Aug  9 13:06:43.824280 bgp 100 [9604]: [9617]: (default) RIB: [L2VPN EVPN]
Add/delete 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112, flags=
0x100, in_rib: no
```

Note: This shows that the route has been added to the default RIB table.

```
Leaf-4# sh bgp l2vpn evpn aabb.cc80.b000

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10020:100    (L2VNI 10020)
BGP routing table entry for [2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/216,
version 7807
Paths: (1 available, best #1)
Flags: (0x000102) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    5.5.5.56 (metric 0) from 0.0.0.0 (6.6.6.6)
```

```
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 10020
      Extcommunity:  RT:100:10020 SOO:5.5.5.56:0 ENCAP:8
```

Note: MAC got installed in the BGP L2VPN EVPN Table.

Remote L2 MAC Route Installation Via BGP EVPN.

BGP L2VPN

```
Leaf-1# show bgp l2vpn evpn aabb.cc80.b000

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10020:100    (L2VNI 10020)
BGP routing table entry for [2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/216,
version 3272
Paths: (1 available, best #1)
Flags: (0x000212) on xmit-list, is in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, imported same remote RD, received and used
, is best path, no labeled nexthop, in rib
  AS-Path: NONE, path sourced internal to AS
    5.5.5.56 (metric 81) from 2.2.2.2 (2.2.2.2)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10020
      Extcommunity:  RT:100:10020 SOO:5.5.5.56:0 ENCAP:8
      Originator: 5.5.5.5 Cluster list: 2.2.2.2
```

How Route Type 2 looks like with details:

```
Route Distinguisher: 10020:100    (L2VNI 10020)
ROUTE TYPE 2 : [2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/216

 [2]  ——> Route Type 2
 [0]  ——> Ethernet Segment ID
```

```
 [0]  ——> Ethernet Tag ID

 [48]:[aabb.cc80.b000] ——-> MAC Address Length

 [0]:[0.0.0.0]  ———> IP Address

 Received label 10020 ———> MPLS Label
```

Note: Ethernet TAG ID, MAC Address Length, MAC Address, IP Address Length and IP Address field are considered to be the part of the prefix of NLRI.

Ethernet Segment Identifier, MPLS Label 1 and MPLS Label 2 are treated as route attribute not the part of the route. The length of both the IP and MAC addresses are in bits.

BGP > L2RIB

```
Leaf-1# sh bgp internal event-history events | i aabb.cc80.b000


2017 Aug  9 15:14:58.682300 bgp 100 [31648]: [31660]: (default) IMP: bgp_tbl_ctx
_import: 1971: [L2VPN EVPN] Importing 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000
]:[0]:[0.0.0.0]/112 to RD 10020:100
2017 Aug  9 15:14:58.682174 bgp 100 [31648]: [31660]: (default) IMP: [L2VPN EVPN
] Import of 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112 (EVI:
10020) to RD 1.1.1.1:65534 (0) inhibited, no Type2 for EAD-ES import
2017 Aug  9 15:14:58.681994 bgp 100 [31648]: [31660]: (default) IMP: bgp_tbl_ctx
_import: 1971: [L2VPN EVPN] Importing 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000
]:[0]:[0.0.0.0]/112 to RD 1.1.1.1:65534
2017 Aug  9 15:14:58.671127 bgp 100 [31648]: [31660]: (default) RIB: [L2VPN EVPN
]: Send to L2RIB 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112
2017 Aug  9 15:14:58.658330 bgp 100 [31648]: [31660]: (default) RIB: [L2VPN EVPN
] For 10020:100:[2]:[0]:[0]:[48]:[aabb.cc80.b000]:[0]:[0.0.0.0]/112, added 1 nex
t hops, suppress 0
```

Note: BGP Imports the Route Type 2 and then installs in Default RIB.

Remote L2 MAC Route Installation via BGP EVPN.

VXLAN Manager Component:

```
Leaf-1# sh nve internal bgp rnh database
--------------------------------------------
```

```
Total peer-vni msgs recvd from bgp: 3

Peer add requests: 3

Peer update requests: 0

Peer delete requests: 0

Peer add/update requests: 3

Peer add ignored (peer exists): 0

Peer update ignored (invalid opc): 0

Peer delete ignored (invalid opc): 0

Peer add/update ignored (malloc error): 0

Peer add/update ignored (vni not cp): 0

Peer delete ignored (vni not cp): 0

------------------------------------------

Showing BGP RNH Database, size : 3 vni 0


Flag codes: 0 - ISSU Done/ISSU N/A        1 - ADD_ISSU_PENDING

            2 - DEL_ISSU_PENDING          3 - UPD_ISSU_PENDING



VNI    Peer-IP            Peer-MAC           Tunnel-ID  Encap     (A/S)  Flags

10020  5.5.5.56           0000.0000.0000     0x0        vxlan     (1/0)    0

10500  3.3.3.3            5000.0003.0007     0x3030303  vxlan     (1/0)    0

10500  5.5.5.56           5000.0009.0007     0x5050538  vxlan     (1/0)    0
```

L2RIB

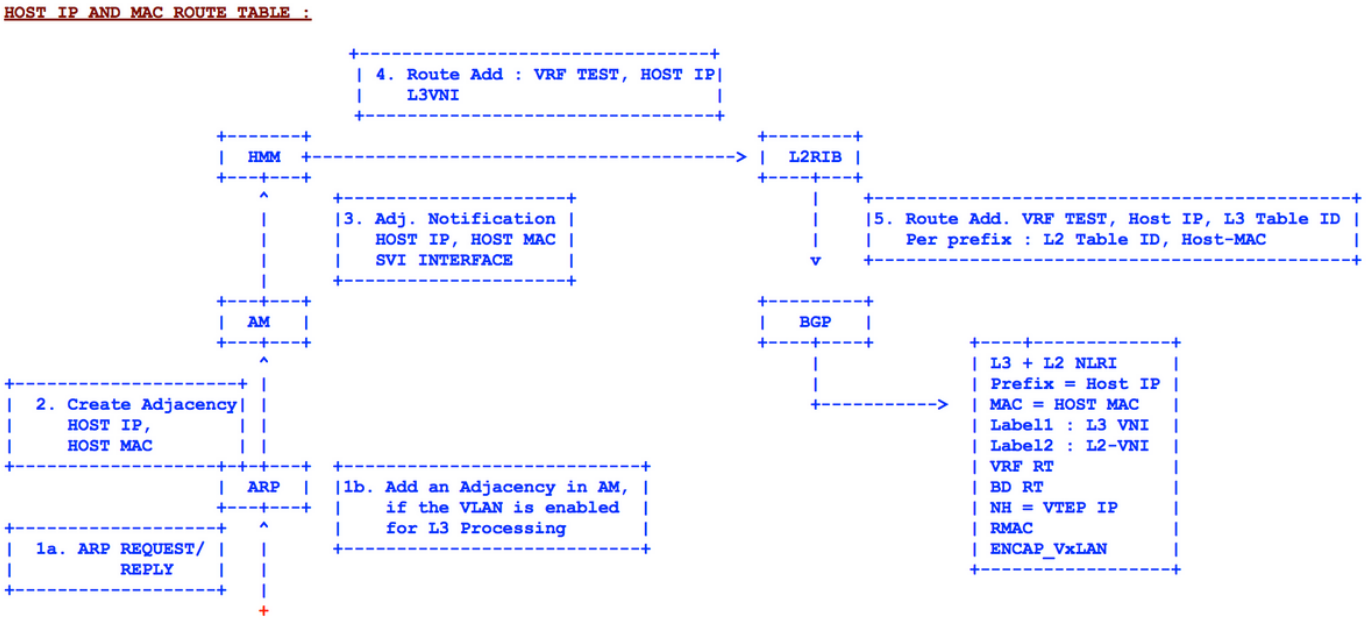
```
Leaf-1# show l2route evpn mac evi 20


Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link

(Dup):Duplicate (Spl):Split (Rcv):Recv (AD):Auto-Delete(D):Del Pending (S):Stale

(C):Clear

(Ps):Peer Sync (O):Re-Originated


Topology    Mac Address    Prod   Flags         Seq No     Next-Hops

----------- -------------- ------ ------------- ---------- ----------------
```

```
20          5000.0007.0000 Local  L,           0          Eth1/1
20          aabb.cc00.d000 BGP    Rcv          0          5.5.5.56
20          aabb.cc80.5000 Local  L,           0          Eth1/1
20          aabb.cc80.b000 BGP    SplRcv       0          5.5.5.56
```

VXLAN Manager:

```
Leaf-1# sh nve  peers  detail

Details of nve Peers:


Peer-Ip: 5.5.5.56   >>>>>>>>>>>>>>>>>>>>>> Use of Secondary Address as we have two Gateways

    NVE Interface       : nve1

    Peer State          : Up

    Peer Uptime         : 2d00h

    Router-Mac          : 5000.0009.0007 >>>>>>>>>>>>>>> Remote MAC Address Details

    Peer First VNI      : 10500

    Time since Create   : 2d00h

    Configured VNIs     : 10020,10500

    Provision State     : add-complete >>>>>>>>>>>>>>> Hardware Programmed

    Route-Update        : Yes

    Peer Flags          : RmacL2Rib, TunnelPD, DisableLearn

    Learnt CP VNIs      : 10020,10500

    Peer-ifindex-resp   : Yes

---------------------------------------
```

Note: Programs data plane with unicast encapsulation/decapsulation for VNI, allocated peer ID.



ARP > AM

Leaf-1# sh ip arp vrf L3VNI

```
IP ARP Table for context L3VNI

Total number of entries: 1

Address          Age       MAC Address     Interface       Flags

10.0.0.1         00:09:02  aabb.cc80.5000  Vlan20
```

Leaf-1# sh ip route vrf L3VNI

```
10.0.0.0/24, ubest/mbest: 1/0, attached

    *via 10.0.0.254, Vlan20, [0/0], 2d06h, direct

10.0.0.1/32, ubest/mbest: 1/0, attached

    *via 10.0.0.1, Vlan20, [190/0], 1d12h, hmm

10.0.0.254/32, ubest/mbest: 1/0, attached

    *via 10.0.0.254, Vlan20, [0/0], 2d06h, local
```

AM > HMM > L2RIB

Leaf-1# show l2route evpn mac-ip evi 20

```
Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
```

(Dup):Duplicate (Spl):Split (Rcv):Recv(D):Del Pending (S):Stale (C):Clear

(Ps):Peer Sync (Ro):Re-Originated

Topology Mac Address Prod Flags Seq No Host IP Next-

Hops

---------- -------------- ------ ---------- --------------- ---------------

20 **aabb.cc80.5000 HMM** -- 0 **10.0.0.1** **Local**

L2RIB > BGP

Leaf-1# sh bgp l2vpn evpn 10.0.0.1

BGP routing table information for VRF default, address family L2VPN EVPN

Route Distinguisher: **10020:100 (L2VNI 10020)**

BGP routing table entry for [2]:[0]:[0]:[48]:[**aabb.cc80.5000**]:[32]:[**10.0.0.1**]/27

2, version 1101

Paths: (1 available, best #1)

Flags: (0x000102) on xmit-list, is not in l2rib/evpn

Advertised path-id 1

Path type: local, path is valid, is best path, no labeled nexthop

AS-Path: NONE, path locally originated

1.1.1.1 (metric 0) from 0.0.0.0 (1.1.1.1)

Origin IGP, MED not set, localpref 100, weight 32768

Received label 10020 10500

Extcommunity: **RT:100:10020 RT:100:10500** ENCAP:8 Router MAC:5000.0001.0007

L2RIB > URIB

Leaf-4# show l2route evpn mac-ip evi 20

Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link

(Dup):Duplicate (Spl):Split (Rcv):Recv(D):Del Pending (S):Stale (C):Clear

```
(Ps):Peer Sync (Ro):Re-Originated

Topology    Mac Address    Prod   Flags         Seq No     Host IP         Next-

Hops

----------- -------------- ------ ---------- --------------- ---------------

20          aabb.cc80.5000 BGP    --            0          10.0.0.1       1.1.1.1

20          aabb.cc80.b000 HMM    --            0          10.0.0.2       Local
```

```
Leaf-4# sh ip route vrf L3VNI


10.0.0.1/32, ubest/mbest: 1/0

    *via 1.1.1.1%default, [200/0], 1d12h, bgp-100, internal, tag 100 (evpn) segi

d: 10500 tunnelid: 0x1010101encap: VXLAN
```

Remote Host Prefix-EVPN

```
Leaf-4# sh bgp l2vpn evpn 10.0.0.1


BGP routing table information for VRF default, address family L2VPN EVPN

Route Distinguisher: 10020:100    (L2VNI 10020)

BGP routing table entry for [2]:[0]:[0]:[48]:[aabb.cc80.5000]:[32]:[10.0.0.1]/27

2, version 2285

Paths: (1 available, best #1)

Flags: (0x000212) on xmit-list, is in l2rib/evpn, is not in HW, is locked


  Advertised path-id 1

  Path type: internal, path is valid, imported same remote RD, received and used

, is best path, no labeled nexthop, in rib

  AS-Path: NONE, path sourced internal to AS

    1.1.1.1 (metric 81) from 2.2.2.2 (2.2.2.2)

      Origin IGP, MED not set, localpref 100, weight 0

      Received label 10020 10500

      Extcommunity:  RT:100:10020 RT:100:10500 ENCAP:8 Router MAC:5000.0001.0007

      Originator: 1.1.1.1 Cluster list: 2.2.2.2
```

NVE Internal Platform Details:

```
Leaf-4# show nve internal platform interface nve1 detail

Printing Interface ifindex 0x49000001 detail

|======|=========================|===============|===============|=====|=====|
|Intf  |State                    |PriIP          |SecIP          |Vnis |Peers|
|======|=========================|===============|===============|=====|=====|
|nve1  |UP                       |6.6.6.6        |5.5.5.56       |3    |2    |
|======|=========================|===============|===============|=====|=====|

SW_BD/VNIs of interface nve1:

================================================

|======|======|=========================|======|====|======|========
|Sw BD |Vni   |State                    |Intf  |Type|Vrf-ID|Notified
|======|======|=========================|======|====|======|========
|20    |10020 |UP                       |nve1  |CP  |0     |Yes
|30    |10030 |UP                       |nve1  |CP  |0     |Yes
|500   |10500 |UP                       |nve1  |CP  |4     |Yes
|======|======|=========================|======|====|======|========

Peers of interface nve1:

=========================================

Peer_ip: 1.1.1.1
  Peer-ID   : 1
  State     : UP
  Learning  : Disabled
  TunnelID  : 0x1010101
  MAC       : 5000.0001.0007
  Table-ID  : 0x1
  Encap     : 0x1
```

Verify Data Plane

Step 1: Verify if NVE peers are UP:

```
Leaf-4# sh nve peers

Interface Peer-IP          State LearnType Uptime   Router-Mac

--------- ---------------  ----- --------- -------- -----------------

nve1      1.1.1.1          Up    CP        1d12h    5000.0001.0007

nve1      3.3.3.3          Up    CP        1d17h    5000.0003.0007
```

Step 2: Verify if peer-id is allocated:

```
Leaf-4# sh forwarding distribution peer-id


UFDM Peer-id allocations:

App: VXLAN   Vlan: 1     Id: 0x101010101  Peer-id: 0x1

App: VXLAN   Vlan: 1     Id: 0x103030303  Peer-id: 0x2
```

Step 3: Verify if MAC address is present in the TCAM Table:

```
Leaf-4# sh forwarding distribution peer-id


UFDM Peer-id allocations:

App: VXLAN   Vlan: 1     Id: 0x101010101  Peer-id: 0x1

App: VXLAN   Vlan: 1     Id: 0x103030303  Peer-id: 0x2
```