BGP-EVPN VxLAN Lab - Part 3 - Inter VNI Routing

CSE PRACTICALS

visit http://www.csepracticals.com **25+ Courses on System Programming and Network Development**

Udemy Link : https://www.udemy.com/user/abhishek-sagar-8/

Syllabus

Chapter 1 Intra-Vlan Routing Lab

Chapter 2 BGP-EVPN Control Plane & Route Advertisement

Chapter 3 Inter-VNI Routing (this)

Chapter 4 External Connectivity

Introduction

This is a detailed GNS3 LAB in which **BGP-EVPN VxLAN** is Implemented on Cisco NXOSv L3 Switches. We Explain the Concept involved in BGP-EVPN VxLAN, show the various output, and packet captures, and explain the data plane operations.

In this LAB , We will demonstrate Data Plane Operations involved in setting up communication between Two hosts within same Customer VRF but in different VLANs.

Why VxLAN is introduced, and what problem it solves comes under theory. You are recommended to cover the ***Why VxLAN part*** *from standard books or otherwise.*

Software Used: GNS3 version 2.2.39 (*You can use latest available on GNS3 website* **)**

Virtualization: VMWare workstation PRO (*You can use the latest available on the VMWare website.* VMWare WS PRO is a paid software, go for *Work Station Player* instead which is free **)**

System: My system has 64 GB RAM. At least 16 GB RAM is recommended to run this LAB.

Cisco NXOSv Image: NXOSv9k-93000v-10.1.1.1 (This is L3 Switch)

A Great Resource of GNS3 Image Collection is here

The NXOSv Image that I used in this lab is here .

For this LAB you are recommended to use the same version of NXOSv Image as mine to avoid any discrepancy during lab time.

Books: We use two Books as a reference :

Building Data Centers with VxLAN BGP EVPN (Cisco Book)

Virtual Extensible LAN - VxLA A Practical Guide (*Strongly Recommend this one , I followed this book Most of the time*)

Troubleshooting PPT (*Somebody created this nice PPT on VxLAN*)

Pre-Requisites

To do this Tutorial You must know the following :

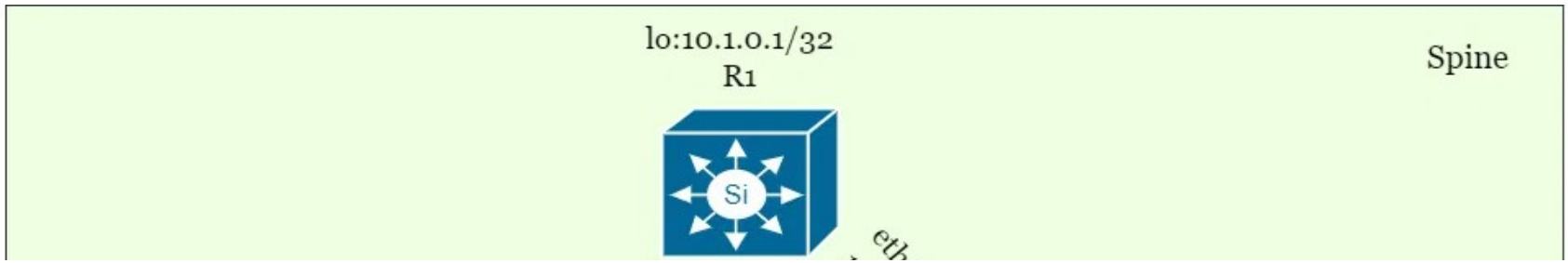
You Must Have Completed Part 1 and Part 2

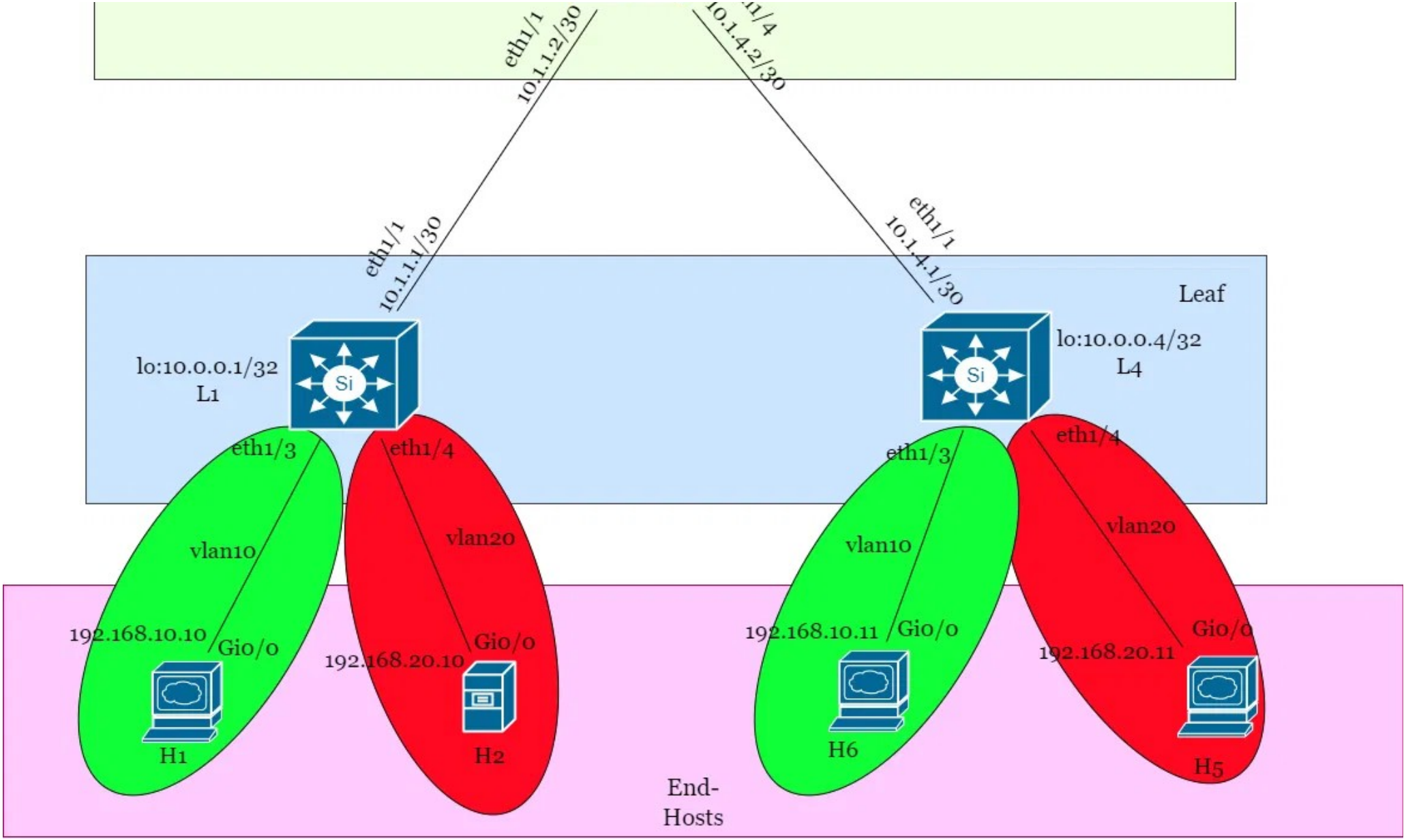
GNS3 LAB with all the configuration we have done in the previous two parts

Having a willingness to learn and the ability to ask questions promptly will greatly aid in your professional development.

Topology

We will resume with the same topology carrying forward from the previous lab.





Lab Topology

At Minimum, you must deploy **R1, L1,L4, H1, H2, H5, H6** nodes as shown in the above diagram. Spine R1 and Leaf Nodes (L1 and L4) are the same NXOSv

L3 Switches. So, Total 3 NXOSv nodes you need to deploy - R1, L1, and L4.

End-hosts could be any node that supports minimum Network functionality like ping, ARP, etc. You can use VPCS hosts (come inbuilt with GNS3, no need to install anything). In my lab, I prefer to use Cisco ASAv Firewall Images. The end host Image type doesn't matter.

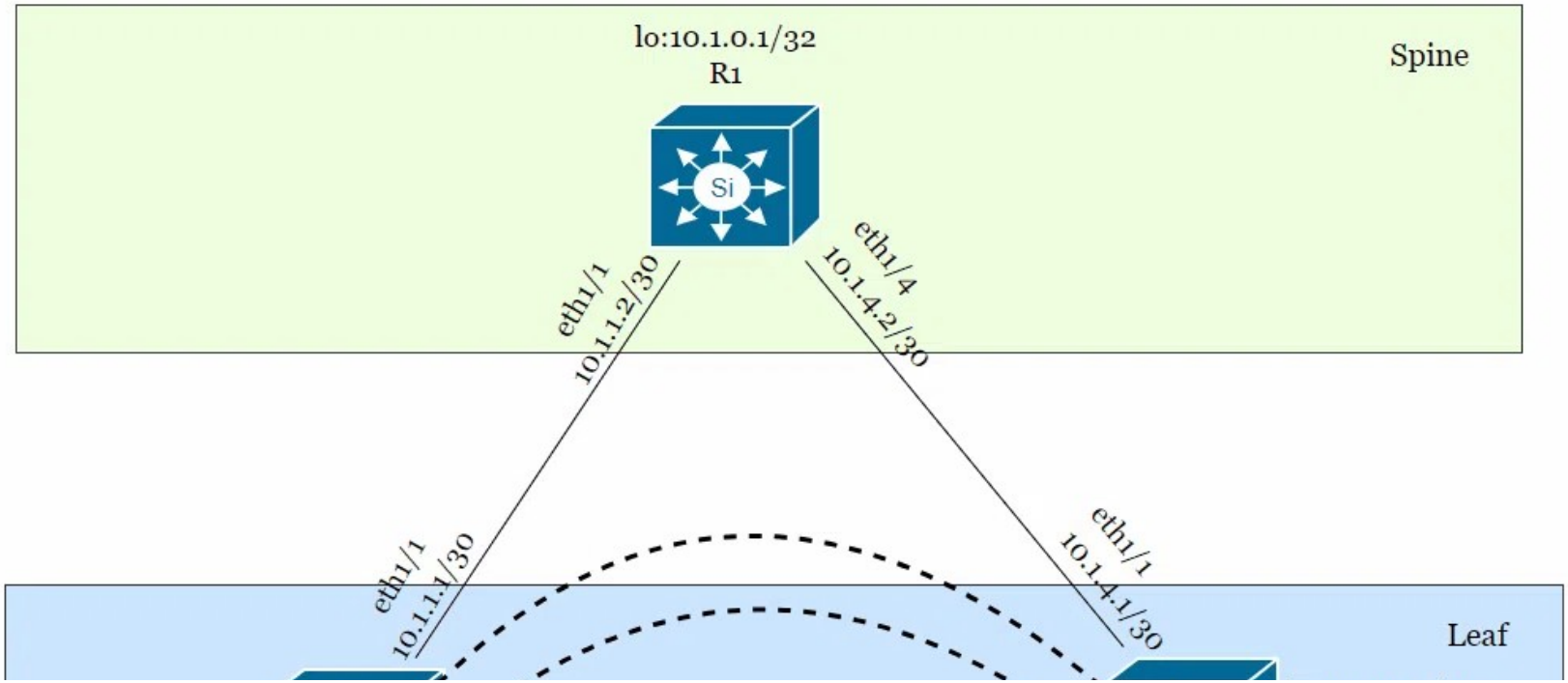
End Hosts H1 and H2 are sitting behind Leaf Node L1

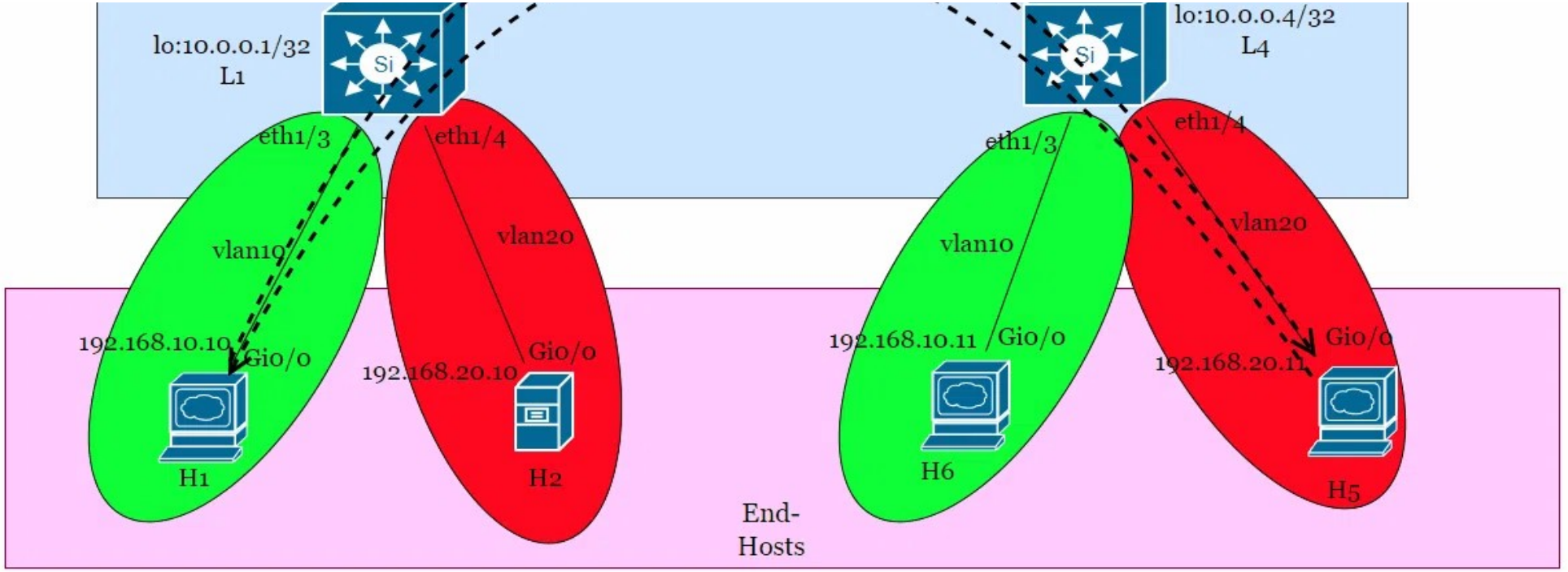
End Hosts H5 and H6 are sitting behind Leaf Node L4

Thankfully, we don't have to do any new configuration in this lab on any device.

End Goal of the LAB

In this lab, we will practice and understand the data plane operations involved in case of Inter-VNI routing i.e. when Host H1 is present in VLAN 10 (VNI 5010) pings the Host H5 is present in VLAN 20 (VNI 5020). Both Vlans are alooted to same customer VRF **Cust-A**. We will understand step by step how the Inter-VNI communication is feasible in a VxLAN network. We dont have to do any new configuration in this Chapter, Please carry forward all configuration from previous chapter.





Data Plane Operations

Initially, all Tables are Empty, Leaf L1 and L4 has just booted up. iBGP neighborship has been formed between R1 and L1 , R1 and L4, R1 being the Spine (Route reflector)

Initial Table States

Leaf L1	Leaf L2

Mac Address Table which is empty
`leaf1# show mac address-table`
`VLAN MAC Address Type age Secure NTFY Ports`
`---------+-----------------+--------+---------+------+----+------------------`
`G - 0002.0002.0002 static - F F sup-eth1(R)`
	Mac Address Table which is empty
	`leaf4# show mac address-table`
	`VLAN MAC Address Type age Secure NTFY Ports`
	`---------+-----------------+--------+---------+------+----+------------------`
	`G - 0002.0002.0002 static - F F sup-eth1(R)`

```
G - 0ca5.0000.1b08 static - F F sup-eth1(R)
G 10 0ca5.0000.1b08 static - F F sup-eth1(R)
G 20 0ca5.0000.1b08 static - F F sup-eth1(R)
G 999 0ca5.0000.1b08 static - F F sup-eth1(R)
leaf1#
```

```
G - 0c88.0000.1b08 static - F F sup-eth1(R)
G 10 0c88.0000.1b08 static - F F sup-eth1(R)
G 20 0c88.0000.1b08 static - F F sup-eth1(R)
G 999 0c88.0000.1b08 static - F F sup-eth1(R)
leaf4#
```

VRF Routing Table, containing only local and direct routes

```
leaf1# show ip route vrf Cust-A
IP Route Table for VRF "Cust-A"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>
192.168.10.0/24, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 00:07:59, direct
192.168.10.1/32, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 00:07:59, local
192.168.20.0/24, ubest/mbest: 1/0, attached
*via 192.168.20.1, Vlan20, [0/0], 00:07:59, direct
192.168.20.1/32, ubest/mbest: 1/0, attached
*via 192.168.20.1, Vlan20, [0/0], 00:07:59, local
```

VRF Routing Table, containing only local and direct routes

```
leaf4# show ip route vrf Cust-A
IP Route Table for VRF "Cust-A"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>
192.168.10.0/24, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 00:19:31, direct
192.168.10.1/32, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 00:19:31, local
192.168.20.0/24, ubest/mbest: 1/0, attached
*via 192.168.20.1, Vlan20, [0/0], 00:19:31, direct
192.168.20.1/32, ubest/mbest: 1/0, attached
*via 192.168.20.1, Vlan20, [0/0], 00:19:31, local
```

VRF ARP table, Empty

```
leaf1# show ip arp vrf Cust-A
Flags: * - Adjacencies learnt on non-active FHRP router
+ - Adjacencies synced via CFSoE
# - Adjacencies Throttled for Glean
CP - Added via L2RIB, Control plane Adjacencies
PS - Added via L2RIB, Peer Sync
RO - Re-Originated Peer Sync Entry
D - Static Adjacencies attached to down interface
```

VRF ARP table, Empty

```
leaf4# show ip arp vrf Cust-A
Flags: * - Adjacencies learnt on non-active FHRP router
+ - Adjacencies synced via CFSoE
# - Adjacencies Throttled for Glean
CP - Added via L2RIB, Control plane Adjacencies
PS - Added via L2RIB, Peer Sync
RO - Re-Originated Peer Sync Entry
D - Static Adjacencies attached to down interface
```

```
IP ARP Table for context Cust-A
Total number of entries: 0
Address Age MAC Address Interface Flags
leaf1#
```

```
IP ARP Table for context Cust-A
Total number of entries: 0
Address Age MAC Address Interface Flags
leaf4#
```

ARP Suppression Cache, Empty

```
leaf1# show ip arp suppression-cache detail
Flags: + - Adjacencies synced via CFSoE
L - Local Adjacency
R - Remote Adjacency
L2 - Learnt over L2 interface
PS - Added via L2RIB, Peer Sync
RO - Dervied from L2RIB Peer Sync Entry
Ip Address Age Mac Address Vlan Physical-ifindex Flags Remote Vtep Addrs
```

ARP Suppression Cache, Empty

```
leaf4# show ip arp suppression-cache de
Flags: + - Adjacencies synced via CFSoE
L - Local Adjacency
R - Remote Adjacency
L2 - Learnt over L2 interface
PS - Added via L2RIB, Peer Sync
RO - Dervied from L2RIB Peer Sync Entry
Ip Address Age Mac Address Vlan Physical-ifindex Flags Remote Vtep Addrs
```

L2 EVPN local RIB for VNI 5010, Empty

```
leaf1# show l2route evpn mac evi 10
Topology Mac Address Prod Flags Seq No Next-Hops
----------- -------------- ------ ------------- ----------
---------------------
------------------
leaf1#
```

L2 EVPN local RIB for VNI 5010, Empty

```
leaf4# show l2route evpn mac evi 10
Topology Mac Address Prod Flags Seq No Next-Hops
----------- -------------- ------ ------------- ----------
---------------------
------------------
leaf4#
```

L2 EVPN local RIB for VNI 5020, Empty

```
leaf1# show l2route evpn mac evi 20
Topology Mac Address Prod Flags Seq No Next-Hops
----------- -------------- ------ ------------- ----------
---------------------
------------------
leaf1#
```

L2 EVPN local RIB for VNI 5020, Empty

```
leaf4# show l2route evpn mac evi 20
Topology Mac Address Prod Flags Seq No Next-Hops
----------- -------------- ------ ------------- ----------
---------------------
------------------
leaf4#
```

NVE Known Peers

```
leaf1# show nve peers
```

NVE Known Peers

```
leaf4# show nve peers
```

leaf1#	leaf4#

I am assuming all multicast Route has already been setup as explained in Chapter 1.

Also, in this chapter we will not focus on inspecting BGP route exchange messages, as they are same as explained in previous chapter.

H1 pings 192.168.20.11

Now a Host H1, which is in vlan 10 behind leaf L1 pings host H5 which is in vlan 20 behind remote leaf L4. Following sequence of operations takes place.

ARP Broadcast

Host H1 issues ARP-B request for its default gateway which is `192.168.10.1`. ARP is resolved with MAC Address = Anycast Gateway which is **0002.0002.0002.**

In Steps 3 and 4 below, we show once again what all tables are populated as a result of processing ARP-B packet from Host H1 by Leaf L1.

Using ARP-B packet recvd from H1, L1 populates its ARP table, Routing Table, Mac Address Table, L2 EVPN local RIBs (MAC VRF + IP VRF)

ARP Broadcast packet recvd ———> ARP Table |———> Routing Table

. . . . . |———→ ARP Suppression Cache

. . . . . |———→ IP VRF Table (Mac-IP routes) ———→ BGP L2 EVPN Route Export

ARP Broadcast packet recvd ———> Mac Address Table ———> MAC VRF Table ———→ BGP L2 EVPN Route Export

MAC-IP Routes Learning

```
leaf1# show ip arp vrf Cust-A

IP ARP Table for context Cust-A

Total number of entries: 1

Address Age MAC Address Interface Flags

192.168.10.10 00:01:07 0c08.4e22.0001 Vlan10

leaf1#
```

3.a. L1 inserts Host route 192.168.10.10/32 in its VRF Routing table

```
leaf1# show ip route vrf Cust-A
IP Route Table for VRF "Cust-A"
<snipped>
192.168.10.10/32, ubest/mbest: 1/0, attached
*via 192.168.10.10, Vlan10, [190/0], 00:00:43, hmm
<snipped>
```

3.b. ARP Suppression Cache on L1 is also updated mapping IP of H1 to its Mac Address

```
<output not available>
```

3.c. L1 updates its IP VRF table and inserts MAC-IP route in it with L3VNI value

```
leaf1# show l2route evpn mac-ip evi 10 detail
Topology Mac Address Host IP Prod Flags Seq No Next-Hops
----------- -------------- ------------------------------------
10 0c08.4e22.0001 192.168.10.10 HMM L, 0 Local
L3-Info: 99999
Sent To: BGP
```

Mac-Only Routes Learning

4.a L1 also performs MAC learning using ARP-Broadcast Packet from H1 which in-tun updates MAC-VRF table

```
leaf1# show mac address-table
MAC Address Type age Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
* 10 0c08.4e22.0001 dynamic 0 F F Eth1/3 <<< New Entry inserted, learning the mac addr of Local Host H1
```

```
<snipped>
```

4.b MAC VRF Table Learning is triggered for VNI 5010/vlan 10

```
leaf1# show l2route evpn mac evi 10 detail

Topology Mac Address Prod Flags Seq No Next-Hops

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

10 0c08.4e22.0001 Local L, 0 Eth1/3 <<< New Entry inserted, Learning the Mac-only routes

Route Resolution Type: Regular

Forwarding State: Resolved

Sent To: BGP
```

H1 Recvs ARP reply and now it knows MAC Address of its gateway ip `192.168.10.1` which is **0002.0002.0002**

Host H1 sends ICMP echo packet to L1 with Src Mac = MAC(H1) = `0c08.4e22.0001`, Dst MAC = `0002.0002.0002`, Src Ip = `192.168.10.10`, Dst Ip = `192.168.20.11`

When pkt enters interface eth3 of VTEP L1, it gets tagged with VLAN id 10

Since Dst Mac = Anycast Gateway, L1 looks up its VRF Routing table of Cust-A (VLAN 10 is assigned to VRF Cust-A) for destination 192.168.20.11 as a key and find below matching entry :

```
192.168.20.0/24, ubest/mbest: 1/0, attached

*via 192.168.20.1, Vlan20, [0/0], 00:07:59, direct
```

Note that the above routing table entry is found because there is a VLAN 20 attached to Leaf L1. If there was no vlan 20 created on Leaf L1 this route wouldn't have existed and pkt would have dropped right here.

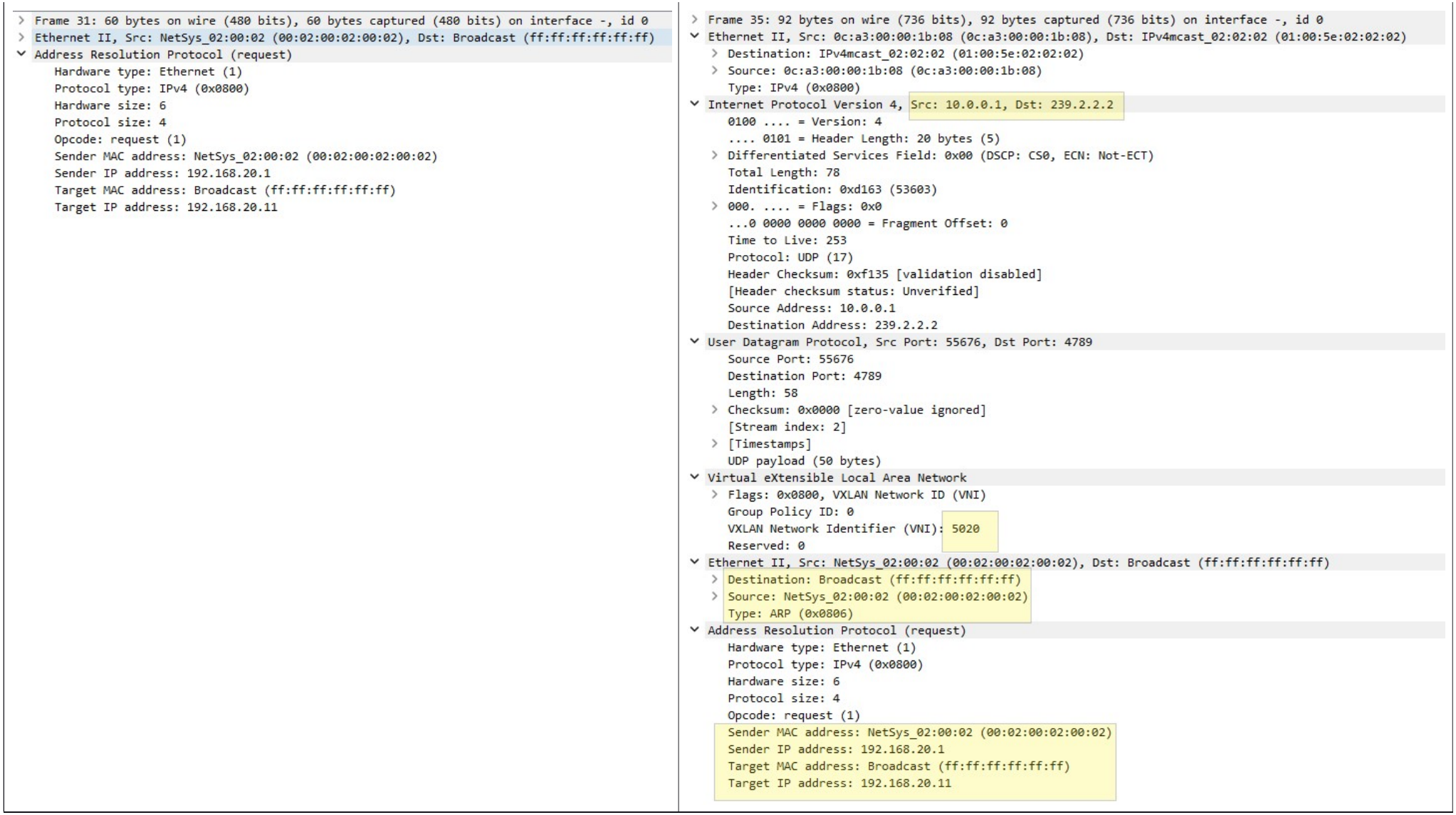
Due to the above route, the VTEP route packet from VLAN 10 to VLAN 20. Now Leaf L1 tries to find a destination within vlan 20 subnets.

VTEP L1 looks up its ARP suppression cache and ARP VRF table to look for the Mac address for destination 192.168.20.11. Assuming it dont found, VTEP L1 launches ARP-Broadcast request for 192.168.20.11 in vlan 20. Note that, it is VTEP L1 that is launching ARP-B request, therefore **Src Mac = Anycast Gateway** in the ethernet hdr of ARP-B packet. Dest Mac would obviously broadcast MAC. The SRC IP will be IP Address of **SVI 20 which is 192.168.20.1**, Dst IP will be

obviously the destination which is **192.168.20.11**. The ARP-B packet is flooded in local vlan 20 on leaf L1 (***Pkt1***) and is encapsulated as Mcast packet and is routed to Remote Leaf L4 (***Pkt2***). This Multicast packet contains VxLAN hdr with VNI 5020 (mapped to VLAN 20). The details regarding how this encapsulation is done are explained in Lab 1

Below is the Wireshark Capture of the local broadcast on VTEP L1 (pkt1) and the ARP-Broadcast packet captured on the link connecting Spine and Leaf L4 (Pkt2)

Pkt1 - Local ARP Broadcast in vlan 20	**Pkt2 - VxLAN Multicast Encasulated ARP-Broadcast in vlan 20**



Remote Leaf L4 receives the ARP-B packet as a multicast packet. The other copy of ARP-B pkt is silently discarded in local vlan 20 network on Leaf L1.

The multicast module on LEaf L4 receives the ARP-Broadcast packet since it is a multicast packet, and encapsulated by a special OIF which is nve1 as explained in Part 1

The Packet VxLAN header is decoded and the extracted VNI is 5020 which maps to vlan 20. The packet is flooded in vlan 20 on Leaf L4 i.e. pushed out of local

interface eth1/3 on Leaf L4.

This packet doesn't contribute anything to flood and learn, since it was routed packet on VTEP L1 (Routed from vlan 10 to vlan 20) . The Src MAC Address in the inner ethernet header is AnyCast Gateway which is not a switchable MAC address. The SRC IP Address is also 192.168.20.1 which is the gateway IP Address for vlan 20 on all VTEPs. We don't learn L3 information using the flood and learn approach.

ARP REPLY

The ARP Broadcast Request is eventually received by host H5. The Host H5 does local ARP learning - `192.168.20.1 ↔ 00:02:00:02:00:02.` Host H5 sends back an ARP reply.

In this ARP reply, In ethernet header, the **Src MAC = MAC(H5) , Dst MAC = Anycast Gateway MAC, and in ARP hdr Src IP = 192.168.20.11 and Dst IP = 192.168.20.1**

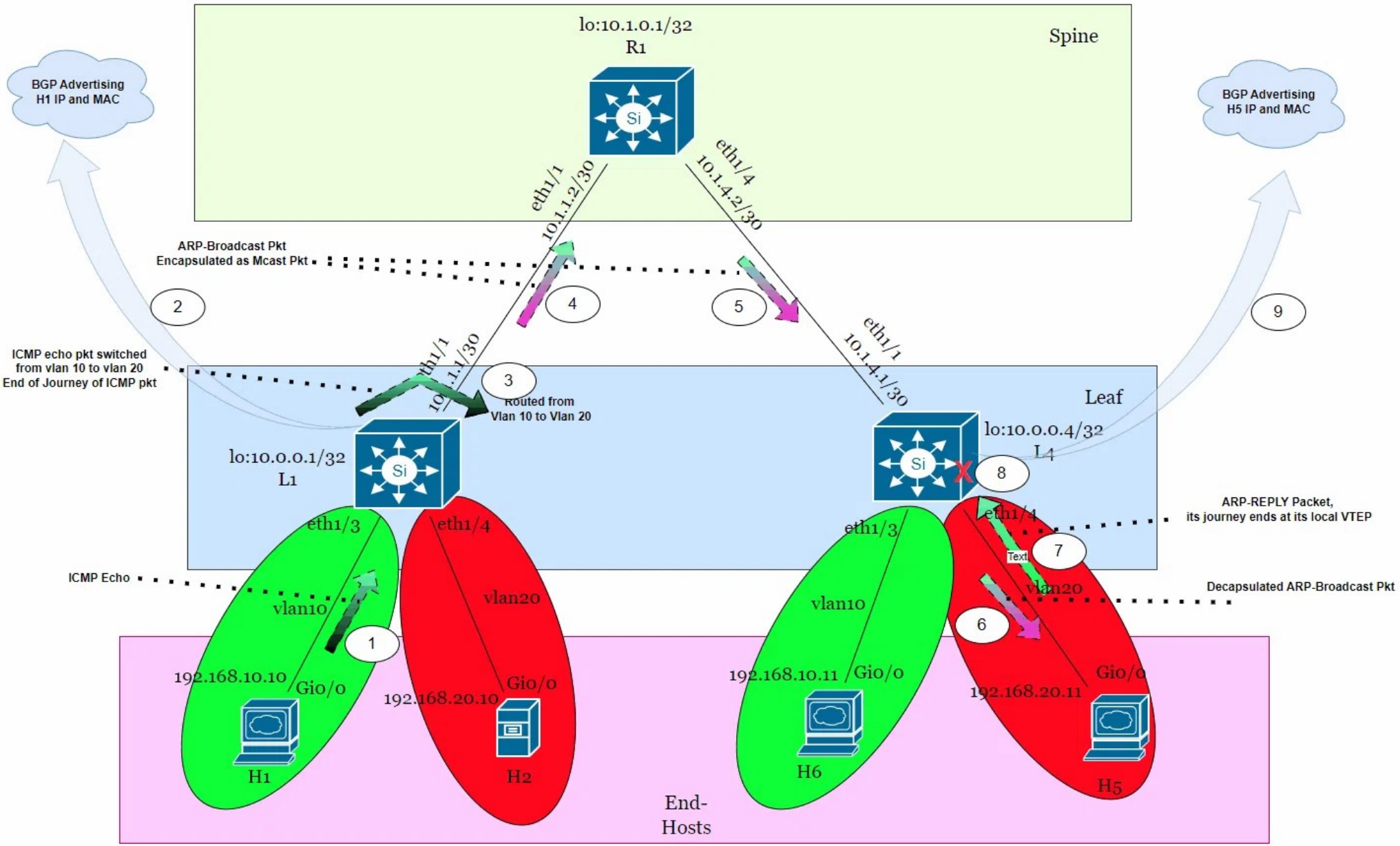
This ARP reply packet leads to learning the VTEP L4 about the presence of Host H5 and it populates all its required tables, the Same way as VTEP L1 learned about Host H1 using the ARP-B packet. Repeat Steps 3 and 4 above for this ARP reply packet.

ARP Reply packet received ———--> ARP Table |———> Routing Table

. . . . . |———→ ARP Suppression Cache

. . . . . |———→ IP VRF Table (Mac-IP routes) ———→ BGP L2 EVPN Route Export

ARP Reply packet recvd ———> Mac Address Table ———> MAC VRF Table ———→ BGP L2 EVPN Route Export



But what is the fate of this ARP-Reply packet once it reaches Leaf L4 ? **The packet is discarded/consumed because as per MAC in ethernet hdr and IP**

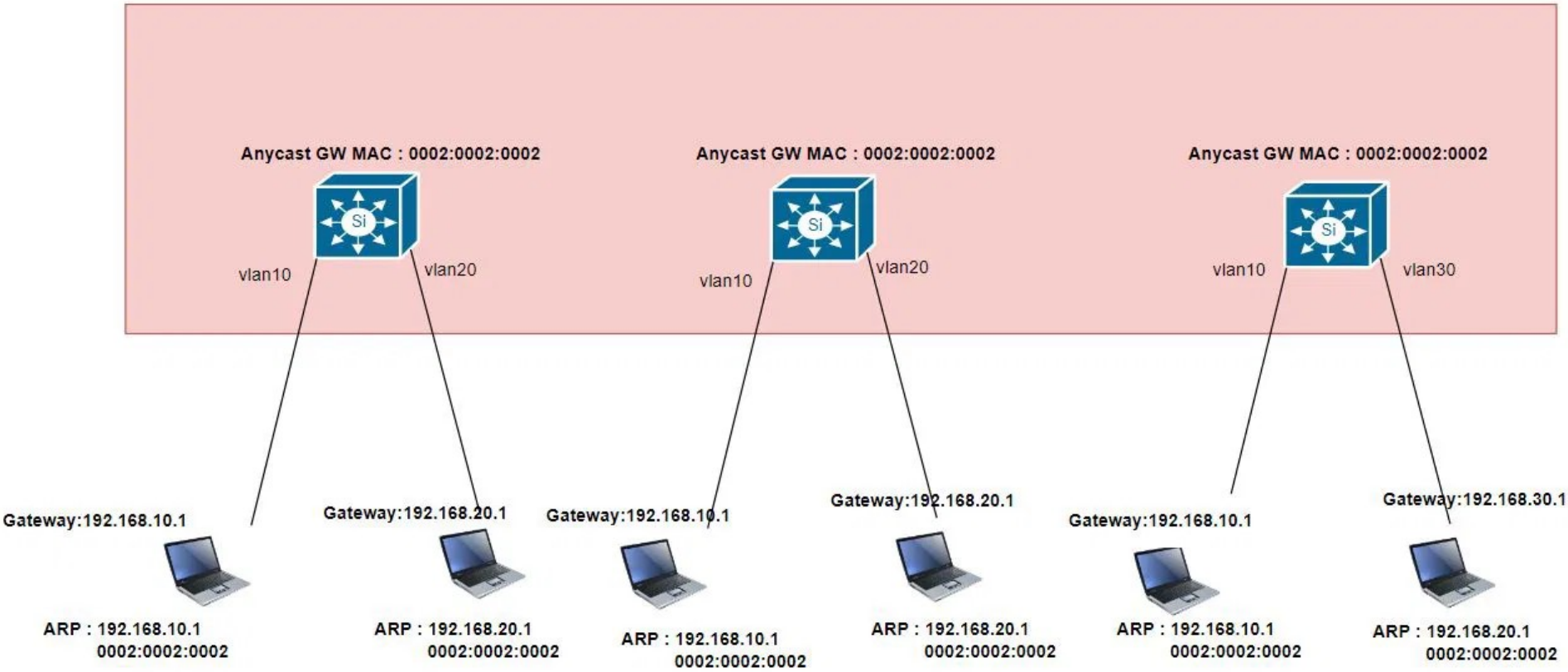
in ARP hdr this packet appears to be destined to VTEP L4 itself. Because the Dst Mac = Anycast MAC and Dst IP Address = Gateway IP 192.168.20.1 in this ARP reply packet which is the same configured on all VTEPs vlan 20 gateway. Thus the journey of ARP reply packet ends at its local VTEP L4.

ARP-Broadcast Pkt Delivered to H5

```
> Frame 16: 60 bytes on wire (480 bits), 60 bytes captured (480 bits) on interface -, id 0
> Ethernet II, Src: NetSys_02:00:02 (00:02:00:02:00:02), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
v Address Resolution Protocol (request)
    Hardware type: Ethernet (1)
    Protocol type: IPv4 (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (1)
    Sender MAC address: NetSys_02:00:02 (00:02:00:02:00:02)
    Sender IP address: 192.168.20.1
    Target MAC address: Broadcast (ff:ff:ff:ff:ff:ff)
    Target IP address: 192.168.20.11
```

ARP-Reply pkt generated by H5

```
> Frame 17: 60 bytes on wire (480 bits), 60 bytes captured (480 bits) on interface -, id 0
> Ethernet II, Src: 0c:b8:f4:23:00:01 (0c:b8:f4:23:00:01), Dst: NetSys_02:00:02 (00:02:00:02:00:02)
v Address Resolution Protocol (reply)
    Hardware type: Ethernet (1)
    Protocol type: IPv4 (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: reply (2)
    Sender MAC address: 0c:b8:f4:23:00:01 (0c:b8:f4:23:00:01)
    Sender IP address: 192.168.20.11
    Target MAC address: NetSys_02:00:02 (00:02:00:02:00:02)
    Target IP address: 192.168.20.1
```



In the above diagram, Anycast Gateway MAC is configured on each VTEP. This is global config. Each of the VLANs created on these VTEPs has IP address configured which serves as the Gateway IP address of local hosts present on those vlans. For example, vlan 10 is configured with GW IP 192.168.10.1 on all VTEPs. For vlan 20 and 30 it is 192.168.20.1 and 192.168.30.1 repectively. Hosts are configured with their default Gateway IPs. For eample, Host present in vlan 10 on all VTEPs are configured with GW IP 192.168.10.1. Host present in vlan 20 and 30 on all VTEPs are configured with GW IP 192.168.20.1 and 192.168.30.1 respectively. When any host initiate the communication destined outsode its local network, Host resolve ARP for their GW IP. For all Hosts, Gateway IP is resolved to same MAC Address ***0002.0002.0002****. VTEP sends ARP replies with this MAC address to all ARP-Broadcast queries generated by their local hosts.*

Significance of Anycast MAC In VxLAN :

Seamless Mobility*: In a VXLAN environment, Virtual Machines (VMs) or hosts can move across different physical locations without changing their default gateway. The anycast gateway MAC allows VMs to maintain the same gateway MAC address regardless of their location, facilitating seamless mobility.*

Redundancy and High Availability*: By configuring multiple devices with the same anycast gateway MAC address, the network ensures high availability. If one device fails, others can continue to provide gateway services without interruption.*

Simplified Configuration*: Network configuration and management are simplified since the same gateway MAC address is used across multiple devices, reducing the complexity associated with managing multiple unique gateway MAC addresses.*

The Generator of the ARP-B packet, the VTEP L1 never receives an ARP reply. The Ist ICMP packet generated by Host H1 is sacrificed in an attempt to provoke Host H5 to report its presence to the Leaf.

The below screenshot captures that Ist ping packet never received any reply, whereas the second one succeeded. The host used is the Cisco ASA Firewall.

```
ciscoasa#
ciscoasa# ping 192.168.20.11 repeat 1
Type escape sequence to abort.
Sending 1, 100-byte ICMP Echos to 192.168.20.11, timeout is 2 seconds:
?
Success rate is 0 percent (0/1)
ciscoasa# ping 192.168.20.11 repeat 1
Type escape sequence to abort.
Sending 1, 100-byte ICMP Echos to 192.168.20.11, timeout is 2 seconds:
!
Success rate is 100 percent (1/1), round-trip min/avg/max = 100/100/100 ms
ciscoasa#
```

Final Forwarding States in Data Plane Tables

Once, the BGP exchanges the presence of Host H1 and H5 across VxLAN fabric i.e. VTEP L1 learning about H5 and VTEP L4 learning about H1, bidirectional communication is seamlessly implemented between the two hosts.

Below is the state of Data plane tables (which contributes in packet routing and switching) after BGP has successfully exchanged the routes between VTEP L1 and L4 for Host H1 and H5.

Leaf L1	Leaf L2

```
leaf1# show mac address-table
VLAN MAC Address Type age Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
* 10 0c08.4e22.0001 dynamic 0 F F Eth1/3
C 20 0cb8.f423.0001 dynamic 0 F F nve1(10.0.0.4)
* 999 0c88.0000.1b08 static - F F nve1(10.0.0.4)
* 999 0ca5.0000.1b08 static - F F Vlan999
G - 0002.0002.0002 static - F F sup-eth1(R)
G - 0ca5.0000.1b08 static - F F sup-eth1(R)
G 10 0ca5.0000.1b08 static - F F sup-eth1(R)
G 20 0ca5.0000.1b08 static - F F sup-eth1(R)
G 999 0ca5.0000.1b08 static - F F sup-eth1(R)
```

```
leaf4# show mac address-table
VLAN MAC Address Type age Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
C 10 0c08.4e22.0001 dynamic 0 F F nve1(10.0.0.1)
* 20 0cb8.f423.0001 dynamic 0 F F Eth1/4
* 999 0c88.0000.1b08 static - F F Vlan999
* 999 0ca5.0000.1b08 static - F F nve1(10.0.0.1)
G - 0002.0002.0002 static - F F sup-eth1(R)
G - 0c88.0000.1b08 static - F F sup-eth1(R)
G 10 0c88.0000.1b08 static - F F sup-eth1(R)
G 20 0c88.0000.1b08 static - F F sup-eth1(R)
G 999 0c88.0000.1b08 static - F F sup-eth1(R)
```

```
leaf1# show ip arp vrf Cust-A
IP ARP Table for context Cust-A
Total number of entries: 1
Address Age MAC Address Interface Flags
192.168.10.10 00:01:15 0c08.4e22.0001 Vlan10
```

```
leaf4# show ip arp vrf Cust-A
IP ARP Table for context Cust-A
Total number of entries: 1
Address Age MAC Address Interface Flags
192.168.20.11 00:03:26 0cb8.f423.0001 Vlan20
```

```
leaf1# show ip route vrf Cust-A
IP Route Table for VRF "Cust-A"
192.168.10.0/24, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 04:08:22, direct
192.168.10.1/32, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 04:08:22, local
192.168.10.10/32, ubest/mbest: 1/0, attached
*via 192.168.10.10, Vlan10, [190/0], 00:02:13, hmm
192.168.20.0/24, ubest/mbest: 1/0, attached
*via 192.168.20.1, Vlan20, [0/0], 04:08:22, direct
192.168.20.1/32, ubest/mbest: 1/0, attached
*via 192.168.20.1, Vlan20, [0/0], 04:08:22, local
```

```
leaf4# show ip route vrf Cust-A
IP Route Table for VRF "Cust-A"
192.168.10.0/24, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 01:14:31, direct
192.168.10.1/32, ubest/mbest: 1/0, attached
*via 192.168.10.1, Vlan10, [0/0], 01:14:31, local
192.168.10.10/32, ubest/mbest: 1/0
*via 10.0.0.1%default, [200/0], 00:03:48, bgp-65501, internal, tag 65501, se
gid: 99999 tunnelid: 0xa000001 encap: VXLAN
192.168.20.0/24, ubest/mbest: 1/0, attached
*via 192.168.20.1, Vlan20, [0/0], 01:14:31, direct
192.168.20.1/32, ubest/mbest: 1/0, attached
```

192.168.20.11/32, ubest/mbest: 1/0 *via 10.0.0.4%default, [200/0], 00:02:03, bgp-65501, internal, tag 65501, se gid: 99999 tunnelid: 0xa000004 encap: VXLAN	*via 192.168.20.1, Vlan20, [0/0], 01:14:31, local 192.168.20.11/32, ubest/mbest: 1/0, attached *via 192.168.20.11, Vlan20, [190/0], 00:03:39, hmm
leaf1# show nve peers Interface Peer-IP State LearnType Uptime Route r-Mac --------- ------------------------------------ nve1 10.0.0.4 Up CP 00:02:08 0c88.0000.1b08	leaf4# show nve peers Interface Peer-IP State LearnType Uptime Route r-Mac --------- ------------------------------------ nve1 10.0.0.1 Up CP 00:05:01 0ca5.0000.1b08

ICMP PING Using L3 VNI

Henceforth, we show, once the route distribution is stabilized, how do H1 and H5 communicate. Now the procedure is completely different and this is the first time L3VNI will be used. Let say, H1 issues ping 192.168.20.11 again while all Data plane Table Entries are in place as listed above.

H1 generates ICMP echo packet. It resolved MAC for its default gateway 192.168.10.1 successfully. ICMP echo packet is received by VTEP L1 successfully on interface eth1/3, vlan tagged 10.

Since, `Dst MAC = AnyCast MAC`, pkt is promoted for L3 routing by VTEP L1

VTEP L1 checks it VRF routing table using Destination addr 192.168.20.11 as a key. It finds the matching route entry **(Routing)**

```
192.168.20.11/32, ubest/mbest: 1/0
*via 10.0.0.4%default, [200/0], 00:02:03, bgp-65501, internal, tag 65501, se
gid: 99999 tunnelid: 0xa000004 encap: VXLAN
```

The above entry tells VTEP L1 to encapsulate the packet using VxLAN hdr , using VNI 99999 to remove VTEP 10.0.0.4 using tunnelid `0xa000004` which is NVE1. The VTEP L1 Executes the below algorithm and prepare the VxLAN Packet.

```
nve1->encapsulate_and_send (remote_vtep_ip = 10.0.0.4 , VNI = 99999)
```

The packet is encapsulated as follows :

Inner Ethernet Pkt

Src Mac : Router MAC of L1

Dst Mac : Router Mac of L4 (see show nve peers output, VTEP knows other VTEP's Rmacs)

Inner IP Hdr

Src IP : 192.168.10.10

Dst IP : 192.168.20.11

VxLAN Hdr

VNI ": 99999

UDP HDr

Src Port : Random (60547)

Dst Port : 4789

Outer IP Hdr :

Src IP : 10.0.0.1

Dst IP : 10.0.0.4

Wireshark Capture below shows ICMP echo request packet values of all the fields as described above. The packet is captured on link **R1 — L4**

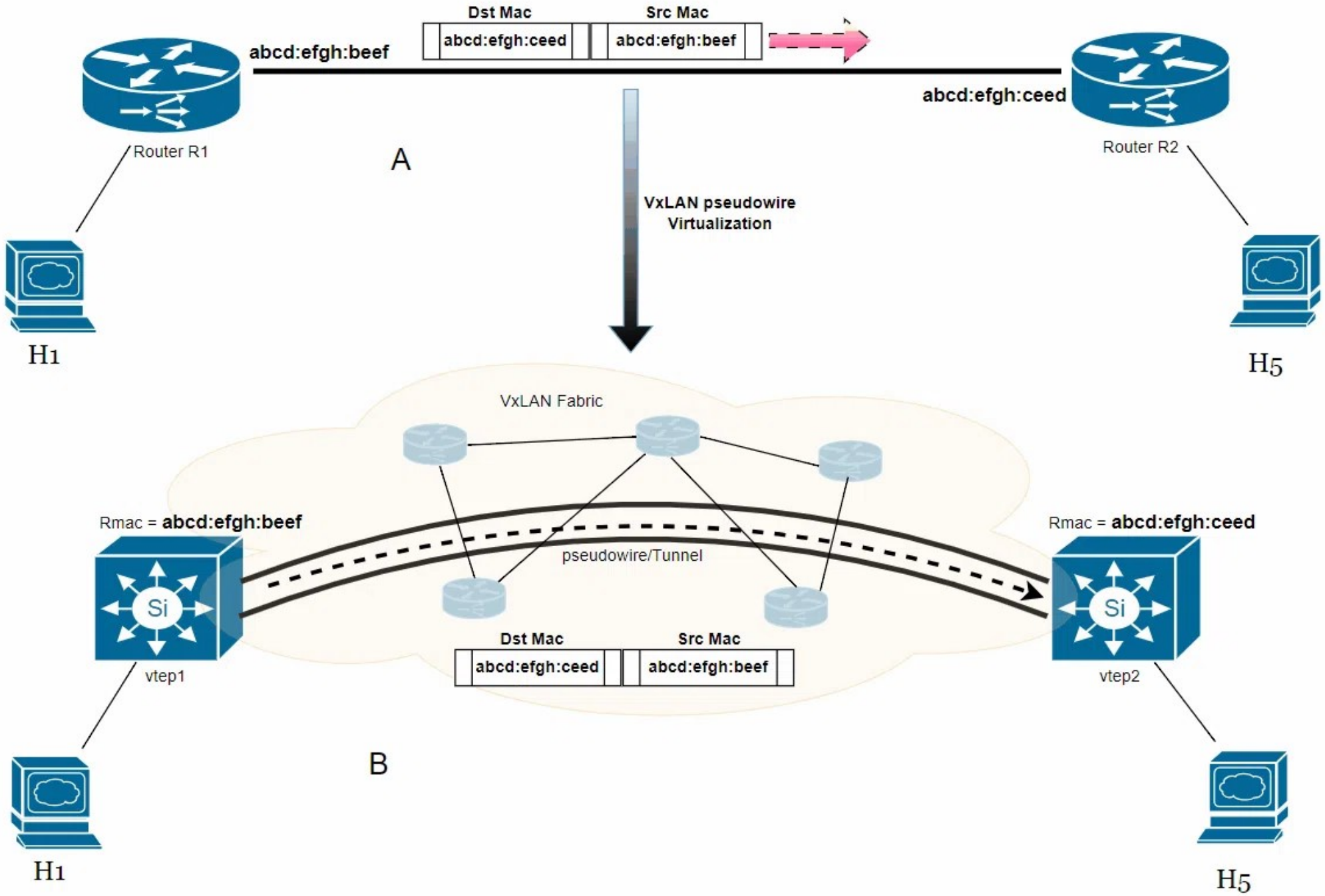
```
> Frame 11: 164 bytes on wire (1312 bits), 164 bytes captured (1312 bits) on interface -, id 0
> Ethernet II, Src: 0c:a3:00:00:1b:08 (0c:a3:00:00:1b:08), Dst: 0c:88:00:00:1b:08 (0c:88:00:00:1b:08)
> Internet Protocol Version 4, Src: 10.0.0.1, Dst: 10.0.0.4
> User Datagram Protocol, Src Port: 60547, Dst Port: 4789
v Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 99999
    Reserved: 0
> Ethernet II, Src: 0c:a5:00:00:1b:08 (0c:a5:00:00:1b:08), Dst: 0c:88:00:00:1b:08 (0c:88:00:00:1b:08)
> Internet Protocol Version 4, Src: 192.168.10.10, Dst: 192.168.20.11
> Internet Control Message Protocol
```

Note that, In inner Ethernet hdr, the Src MAC and Dst MAC addresses used are Router MACs. All VTEPs have their own personal unique Router MAC address. We cant use MAC Addresses of Actual Src and Dst (H1 and H5) here, since, we need to abide by Networking fundamental rules that MAC Addresses of Hosts must be invisible out-side their local subnets. Secondly, Router MAC helps to implement the tunneling of the packet using VxLAN. Here, VxLAN tunnel acts as a

pseudowire which connects the VTEPs. Therefore, it is desirable that they must have Some MAC Addresses which serves the purpose of Src and Dst MAC address for this Pseudowire. This is an attempt to virtualize P2P link. Thirdly, Router MACs helps in implemeting the VxLAN technology using traditional routing and switching without any (or minimal) special changes to support VxLAN based routing. Thirdly, Router MAC helps the Destination VTEP to subject the inner packet to L3 routing again so that it can be routed correctly to locally attached vlan segment.

In the below diagram part A, Every packet which leaves the router A and goes to Router B over a cable would have Dest and Src Mac as shown in the diagram. It is the requirement of the data link layer that Drc and Dst MAC addresses are collected set up in ethernet header of the frame.

Similarly, in the part B, the same behavior is implemented between VTEPs as if there is a CABLE (Pseudowire) between the two VTEPs, though it is virtual. The Inner packet hidden behind VxLAN header experiences the same routing environment as in Part A. Concerning the Inner Packet, the intermediate IP network/VxLAN fabric is invisible or does not exist at all. This presents the bigger picture of a renowned Tunneling Concept.


Dst Mac
Src Mac
abcd:efgh:ceed
abcd:efgh:beef
abcd:efgh:beef
abcd:efgh:ceed
Router R1
Router R2
A
VxLAN pseudowire
Virtualization
H1
H5
VxLAN Fabric
Rmac = abcd:efgh:beef
Rmac = abcd:efgh:ceed
pseudowire/Tunnel
Si
Si
vtep1
vtep2
Dst Mac
Src Mac
abcd:efgh:ceed
abcd:efgh:beef
B
H1
H5

The Packet is L3 routed using outer IP hdr by the VxLAN fabric to Destination VTEP 10.0.0.4

The Destination VTEP receives the packet

Ethernet Hdr is removed and pkt is promoted to L3 routing

Since, Dst ip address in outer hdr of the pkt is 10.0.0.4, IP hdr is removed and pkt is passed to UDP layer

Since, UDP Dst port is 47889, pks is passed to VxLAN module for decoding

VxLAN value is 99999, using VxLAN value and Dst Mac Address `0c88.0000.1b08` (Rmac) combined as the key, Mac Address table is looked up **(Switching)**

Matching entry : `* 999 0c88.0000.1b08 static - F F Vlan999`

Revise : Router MAC helps the Destination VTEP to subject the inner packet to L3 routing again so that it can be routed correctly to locally attached vlan segment.

Outgoing interface is Vlan999. This Vlan is a special interface created on VTEPs to implemen L3 routing. Any packet forwarded to this interface in data plane is subjected to L3 routing again using VRF routing table for Cust-A (since vlan 999 is assigned to VRF Cust-A) . This is virtual interface attached to L3 route functionality.

```
conf

interface vlan 999

ip forward <<< Implement L3 routing

no shutdown

vrf member Cust-A

end
```

Since, Dst IP Address of inner IP Hdr of the packet is 192.168.20.11, VTEP L7 looks up its VRF Routing table using 192.168.20.11 as a key **(Routing)**

The matching entry is below. This entry dictactes the packet needs to be forwarded to **vlan20** interface, and nexthop ip is `192.168.20.11`

```
192.168.20.11/32, ubest/mbest: 1/0, attached

*via 192.168.20.11, Vlan20, [190/0], 00:03:39, hmm
```

Packet is demoted for L2 routing again. VTEP looks for nexthop MAc address in its ARP cache for nexthop IP Address 192.168.20.11.

Matching entry : `192.168.20.11 00:03:26 0cb8.f423.0001 Vlan20`

VTEP L4 appends the ethernet hdr with Src Mac as its Router Mac and Dst MAC as MAC(H5). Now VTEP has to find the outgoing port in vlan 20.

VTEP L4 looks up its MAC table again using vlan 20 and Dst mac as the key. **(Switching)**

Matching entry : `* 20 0cb8.f423.0001 dynamic 0 F F Eth1/4`

Packet is pushed out of interface eth1/4.

Packet (ICMP echo) is delivered to Host H5.

ICMP reply generated by Host H5 destined to Host H1 follow exactly same procedure as above.

In the above procedure, how many times the packet is subjected to Routing and Switching ?

Whenever, VTEP looks up its mac address table to decide what to do with the packet, it is called Switching. On the contrary, if VTEP looks up the routing table then it is called Routing.

> Step 3 - Routing --- VTEP 1
>
> Step 8 - Switching --- VTEP 2
>
> Step 10 - Routing --- VTEP 2
>
> Step 13 - Switching --- VTEP 2

Conclusion

We Explained in this lab how L3VNI is used for Inter-vlan routing as opposed to L2VNI which is used for intra-vlan routing. ARP plays a crucial role when it comes to VxLAN which depends on BGP-EVPN control plane for route distribution. Now in the next lab, we will see how VxLAN network is leveraged by Customers. How Customer's network connnect ti Spine-Leaf VxLAN Architecture powered by BGP-EVPN and how different customer sites carry out data communication between them through VxLAN fabric.

visit : http://www.csepracticals.com for more courses and offers.

Abhishek Sagar