VXLAN BGP EVPN –Control Plane operation

Introduction

Figure 7-1 represents the logical structure of the example VXLAN fabric. BGP peering is established between the VTEP Leaf switches and the Spine-11 switch, which is BGP Route Reflector (not shown in figure 7-1). Both VTEP Leaf switches have a local VRF context TENANT77 that has VNI 10077 (L3VI) attached to it and used for routing between the hosts in different vlan/vn-segment. Hosts Café and Beef are connected to vlan 10 (192.168.11.0/24), which in turns is attached to vn-segment 10000 (L2VNI). Hosts Abba and Babe are connected to vlan 20 (192.168.12.0/24), which in turns is attached to vn-segment 20000 (L2VNI). We are using auto-generated RD and RT values in every VNI and ARP-suppression in both L2VNIs.Physical topology and the configurations of the switches is presented in Appendix 1 at the end of the document. For simplicity, I have used only one uplink in each VTEP switches.

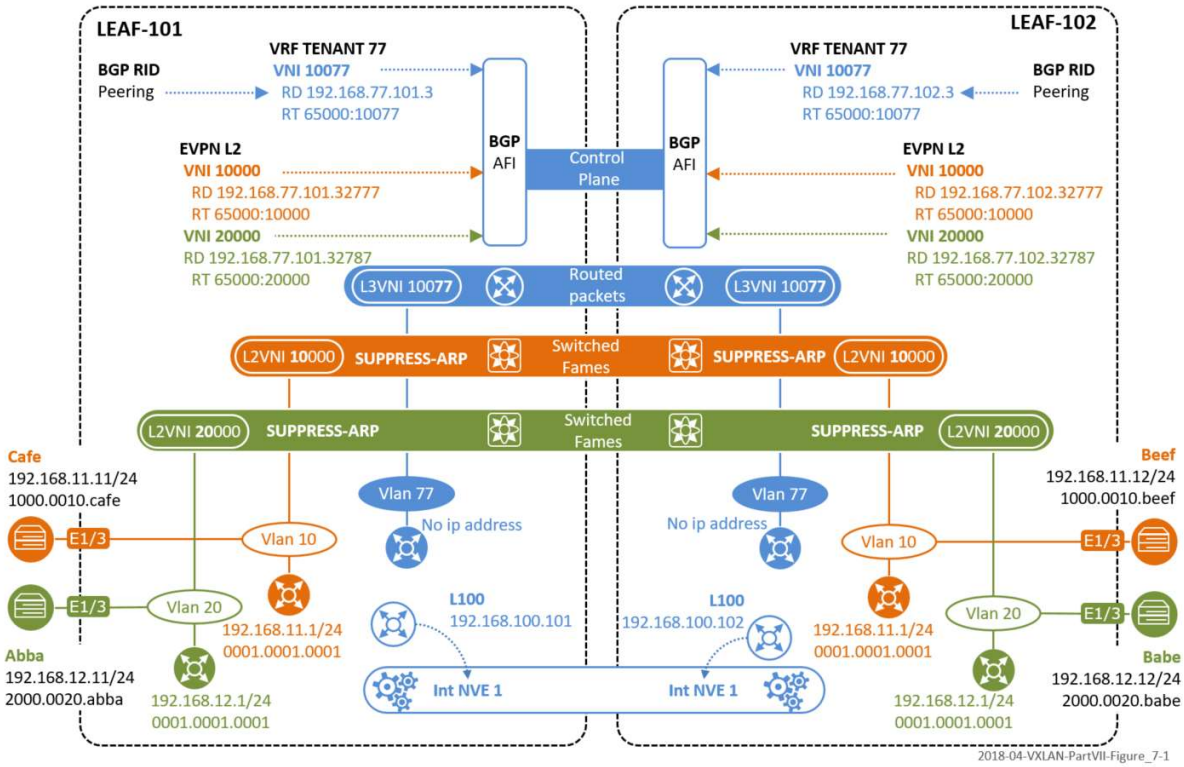


Figure 7-1: *VXLAN Fabric logical structure*

Control Plane operation

Starting point - All hosts are disconnected from the network. We are going to connect host Café to interface Ethernet 1/3 of Leaf-101. We will not generate any Data Plane traffic from Cafe.

Gratuitous ARP

Host Café (ip: 192.168.11.11/mac: 1000.0010.cafe) joins the network. It validates the uniqueness of its IP-address by sending a Gratuitous ARP (Figure 7-1). The Gratuitous ARP-message is actually an ARP-reply message, which is sent without receiving an ARP-request. The message is sent as an L2 broadcast. The ARP reply message itself is targeted to IP address 192.168.11.11, which is the IP address of the host Cafe. If host Café receives an ARP reply for this message, it knows that some other host is already using its IP address.

```
> Frame 73: 110 bytes on wire (880 bits), 110 bytes captured (880 bits)
> Ethernet II, Src: 5e:00:00:00:00:07 (5e:00:00:00:00:07), Dst: IPv4mcast_0a (01:00:5e:00:00:0a)
> Internet Protocol Version 4, Src: 192.168.100.101, Dst: 238.0.0.10
> User Datagram Protocol, Src Port: 61934, Dst Port: 4789
v Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 10000
    Reserved: 0
> Ethernet II, Src: Private_10:ca:fe (10:00:00:10:ca:fe), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
v Address Resolution Protocol (reply/gratuitous ARP)
    Hardware type: Ethernet (1)
    Protocol type: IPv4 (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: reply (2)
    [Is gratuitous: True]
    Sender MAC address: Private_10:ca:fe (10:00:00:10:ca:fe)
    Sender IP address: 192.168.11.11
    Target MAC address: Broadcast (ff:ff:ff:ff:ff:ff)
    Target IP address: 192.168.11.11
```

Capture 7-1: *Gratuitous ARP from host Café*

Even though we have a vni based suppress-arp configured under the NVE1 interface in VTEP Leaf-101, the Gratuitous ARP received from host Café is flooded as a VXLAN encapsulated packet to the Mcast Group 238.0.0.10 (Mcast is explained in Parts III – V.) This happens since VTEP Leaf-101 do not have information about the IP address of host Café in neither the ARP table nor ARP-Suppress cache. VTEP Leaf-101 will update the entries after it has processed the message (Figure 7-2).

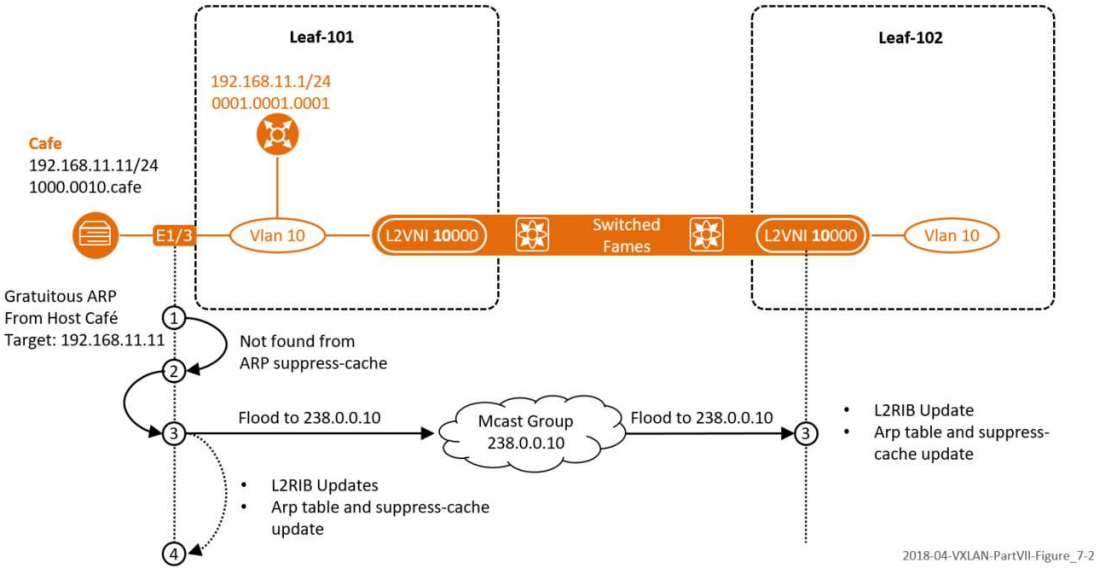


Figure 7-2: *Gratuitous ARP processing*

The process can also be seen from the debug output taken from the VTEP Leaf-101. It receives the GARP from the host Café. It has no cache entry for 192.168.11.11 so it has to flood the frame. After flooding, it updates its ARP cache and L2RIB.

```
Debug ip arp cache
Debug ip arp event
Debug ip arp suppression event

arp_process_packet_in_l3_mode: GARP:  Vlan: 10, Dest-ip: 192.168.11.11, Mac-Addr:
1000.0010.cafe, ifindex: 0x0

arp_cache_resolve_l3_addr: arp_cache_resolve_l3_addr

arp_cache_resolve_l3_addr: cache-entry not found for sw_bd: 10, ip: 192.168.11.11

arp_send_packet: Packet for ffff.ffff.ffff/192.168.11.11, iod 71(Vlan10), phy_iod
7(Ethernet1/3), phy_is_mct 0, flood_bd 1, flood port 1, skip_unnumbered_flood 0

arp_add_adj: Updating MAC on interface Vlan10, phy-interface Ethernet1/3, flags:0x1

arp_create_adj: Create adjacency, interface Vlan10, 192.168.11.11

arp_add_adj: Successful action on new entry Current State:0x10 Received entry:
192.168.11.11, 1000.0010.cafe, Vlan10, action to be taken send_to_am:TRUE,
arp_aging:TRUE

arp_add_adj: Successful action on new entry Current State:0x10 Received entry:
192.168.11.11, 1000.0010.cafe, Vlan10, action to be taken send_to_am:TRUE,
arp_aging:TRUE

arp: arp_cache_add_entry_to_cache_and_upd_l2rib: Post L2FM lookup MAC binding : for sw-
bd: 10, mac: 1000.0010.cafe ip: 192.168.11.11, uuid: 268, vlan_mode: 2, ifindex:
0x901000a, phyifindex 0x1a000400

arp_cache_create_cache_node: create node for uuid:268, sw-bd:10, ip:192.168.11.11,
mac:1000.0010.cafe, mode:2, flags:0x10 is_timer: 0
```

```
arp_cache_create_cache_node: New entry: create node 0x6c13eb8c 0x6c13ead4, uuid: 268,
sw-bd: 10, ip:192.168.11.11, mac: 1000.0010.cafe, is_local: TRUE, num-macs: 1

arp_cache_create_cache_node: New entry: create node 0x6c13eb8c 0x6c13ead4, uuid: 268,
sw-bd: 10, ip:192.168.11.11, mac: 1000.0010.cafe, is_local: TRUE, num-macs: 1

arp_add_adj: Entry added for 192.168.11.11, 1000.0010.cafe, state 2 on interface
Vlan10, physical interface Ethernet1/3, ismct 0. flags:0x10, Rearp (interval: 0, count:
0), TTL: 1500 seconds update_shm:TRUE

arp_add_adj: Adj info: iod: 71, phy-iod: 7, ip: 192.168.11.11, mac: 1000.0010.cafe,
type: 0, sync: FALSE, suppress-mode: L2/L3 ARP Suppression flags:0x10
```

Example 7-1: *Gratuitous ARP process in LEAF-101*

Remote VTEP Leaf-102 receives the flooded frame and updates its ARP cache (Example 7-2).

```
arp_cache_create_cache_node: Host IP 192.168.11.11, Remote vtep addr count = 1

arp_cache_create_cache_node: RNHs : 192.168.100.101

arp_cache_create_cache_node: New entry: create node 0x6c13eb8c 0x6c13ead4, uuid: 1290,
sw-bd: 10, ip:192.168.11.11, mac: 1000.0010.cafe, is_local: FALSE, num-macs: 1
```

Example 7-2: *Gratuitous ARP process in LEAF-102*

L2FWDER

L2FWDER in Nexus 9000v is a centralized forwarding component, which is responsible for mac learning, switching, VXLAN encapsulation and decapsulation and VXLAN BGP EVPN.

Mac learning

The ARP processes started by the host Cafe launches the mac address-learning process on the VTEP Leaf-101. (Figure 7-3).

L2FWDER component notices the incoming frame from the port eth1/3 (vlan 10 access port) with the source mac address 1000.0010.cafe. Note that the interface-index 0x1a000400 points to the interface eth1/3. Mac address 1000.0010.cafe is installed together with vlan and interface information to the mac address-table.

L2FWDER component installs the mac route to L2RIB. L2RIB mac entry is needed since we are also going to advertise the mac address information to the remote VTEP Leaf-102.

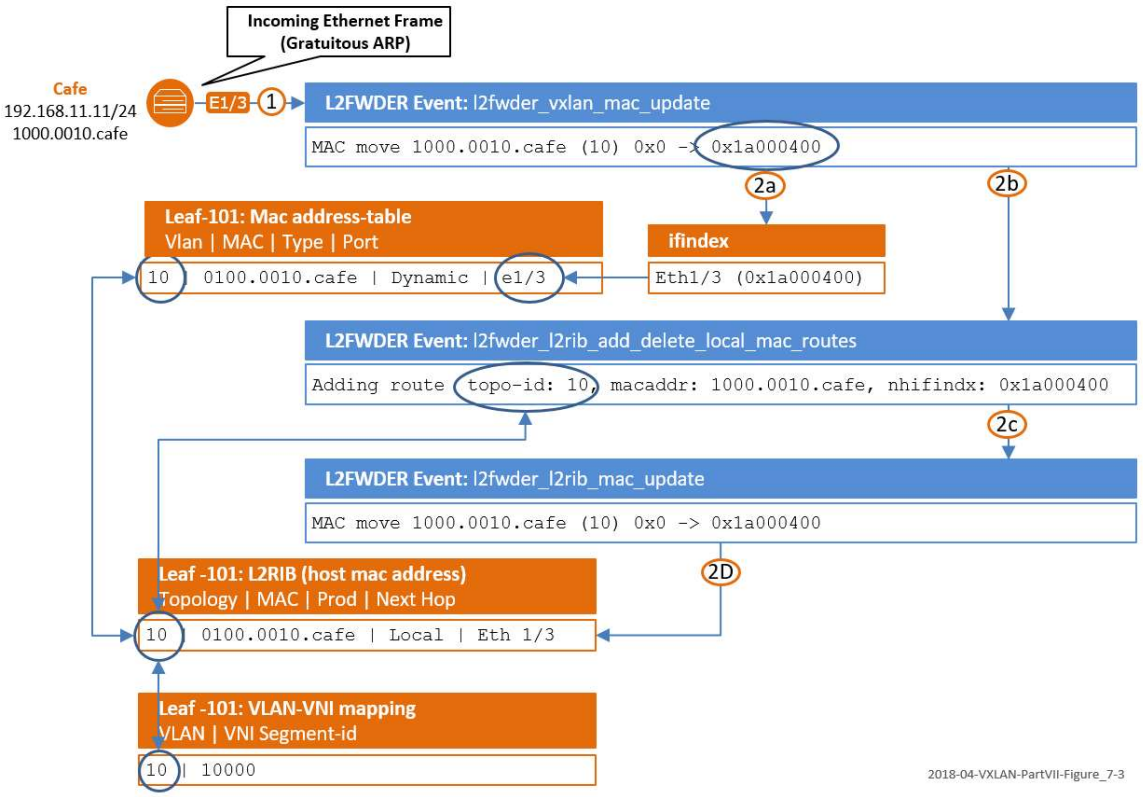


Figure 7-3: Local MAC learning process.

Note! I am using Nexus 9000v (Cisco VIRL). The process of the L2RIB update differs from the physical Nexus 9000 platform. In Nexus 9000v MAC routes are produced directly into L2RIB by L2FWRED (Example 7-3).

The mac address learning process and L2RIB update can be seen in example 7-3.

```
Leaf-101# sh sys internal l2fwder event-history events | i cafe
    [117] [25037]: l2fwder_dbg_ev, 690 l2fwder_vxlan_mac_update, 886MAC move
1000.0010.cafe (10) 0x0 -> 0x1a000400
    [117] [25037]: l2fwder_dbg_ev, 690
l2fwder_l2rib_add_delete_local_mac_routes, 154Adding route  topo-id: 10,
macaddr: 1000.0010.cafe, nhifindx: 0x1a000400
    [117] [25037]: l2fwder_dbg_ev, 690 l2fwder_l2rib_mac_update, 736MAC move
1000.0010.cafe (10) 0x0 -> 0x1a000400
```

Example 7-3: *Mac learning and L2RIB Update.*

We can see that mac address-table is correctly updated (Example 7-4).

```
Leaf-101# show system internal l2fwder mac
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
---------+-----------------+--------+---------+------+----+------------------
*    10    1000.0010.cafe   dynamic   00:00:28   F     F     Eth1/3
```

Example 7-4: *mac address-table on Leaf-101*

We can also see that the L2RIB is updated (Example 7-5). Note that the Topology field value 10 correspond vni 10000 (Example 7-6).

```
Leaf-101# show l2route mac all

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 (Nho):NH-Override
(Pf):Permanently-Frozen

Topology    Mac Address    Prod   Flags          Seq No     Next-Hops
----------- -------------- ------ -------------- ---------- -----------------
10          1000.0010.cafe Local  L,             0          Eth1/3
```

Example 7-5: *L2RIB on Leaf-101*

Previous outputs do not show the VLAN-to-VNI mapping or Physical interface to Interface-Index (ifindx) mapping. VLAN-to-VNI mapping is verified in in Example 7-6.

```
Leaf-101# show vlan id 10 vn-segment
VLAN Segment-id
---- -----------
10   10000
```

Example 7-6: *VLAN-to-VNI mapping*

In Example 7-7 there is an event "l2fwder_l2rib_add_delete_local_mac_routes, 154Adding route topo-id: 10, macaddr: 1000.0010.cafe, nhifindx: 0x1a000400" where 0x1a000400 refers to the interface Ethernet 1/3 (Example 7-7).

```
Leaf-101# show interface snmp-ifindex | i 0x1a000400
Eth1/3          436208640  (0x1a000400)
```

Example 7-7: *Physical Interface to Interface Index mapping*

BGP EVPN

The mac address-table, as well as L2RIB in the VTEP Leaf-101, are now up to date. Now the mac address 1000.0010.cafe needs to be advertised to Leaf-102 so it can switch frames from its connected host Babe to Café. First, we are going to verify that the mac address is advertised internally in VTEP Leaf-101 from the L2RIB by the L2FWDER to BGP EVPN instance and then we check that it has been sent to the correct BGP EVPN Address-Family for redistribution (Figure 7-4).

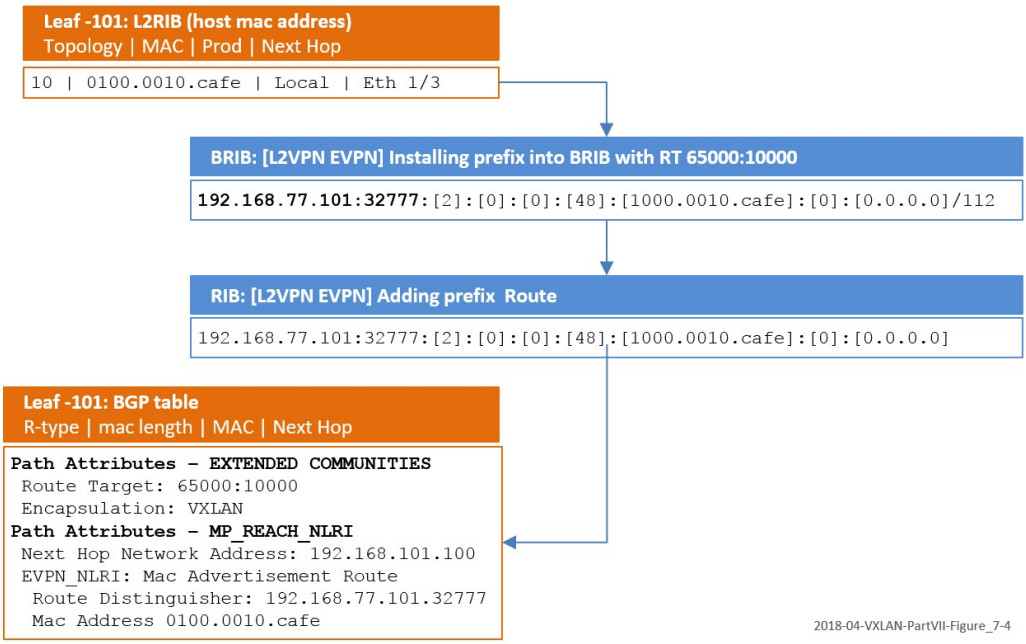


Figure 7-4: mac update from L2RIB to BGP table

We can see that the mac address 1000.0010.cafe is produced to the EVPN instance (Example 7-8).

```
Leaf-101# show l2route evpn mac evi 10

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 (Nho):NH-Override
(Pf):Permanently-Frozen

Topology    Mac Address    Prod   Flags         Seq No     Next-Hops
----------- -------------- ------ ------------- ---------- ----------------
10          1000.0010.cafe Local  L,            0          Eth1/3
```

Example 7-8: *mac address in EVPN instance (EVI) topology 10 (=vni 10000)*

We can also see that the mac address information is installed to the BGP EVPN AFI (Example 7-9).

```
Leaf-101# show bgp l2vpn evpn vni-id 10000
<snipped>

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 192.168.77.101:32777    (L2VNI 10000)
*>l[2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/216
                      192.168.100.101                   100      32768 i
*>l[2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272
                      192.168.100.101                   100      32768 i
```

Example 7-9: *BGP EVPN AFI*

Example 7-10 shows the BGP processes for mac address update to the BGP table. The VTEP Leaf-101 installs the mac route to BGP table and to the RIB.

```
Leaf-101# sh bgp internal event-history events | i cafe
2018 May  2 08:03:19.869831: (default) BRIB: [L2VPN EVPN] Installing prefix
192.168.77.101:32777:[2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/112
(local) via 192.168.100.101 label 10000 (0x0/0x0) into BRIB with extcomm
Extcommunity: RT:65000:10000 ENCAP:8

2018 May  2 08:03:19.869774: (default) RIB: [L2VPN EVPN] Adding prefix
192.168.77.101:32777:[2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0] Route
Length 12 Prefix Length
14:
```

Example 7-10: *BGP EVPN AFI*

We can see that the mac address 1000.0010.cafe is advertised from the VTEP Leaf-101 to remote VTEP Leaf-102 by BGP (Example 7-11). RD is derived from the vni configuration under the EVPN instance This update is used for L2VNI service (frame switching). The notification /216 specifies the bit count of the prefix. There is also mac-IP information as can be seen from the Example 7-9. The prefix mask for mac-IP is /272 since there are 32 bits for ip address and 24 bits for additional Label (L3VNI). This gives as a mask 216 + 32 + 24 = 272. The mac-IP is related to ARP and routing, but i will get back to this later.

```
Leaf-101# sh bgp l2vpn evpn 1000.0010.cafe
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 192.168.77.101:32777    (L2VNI 10000)
BGP routing table entry for
[2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/216, version 109
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
  AS-Path: NONE, path locally originated
    192.168.100.101 (metric 0) from 0.0.0.0 (192.168.77.101)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 10000
      Extcommunity: RT:65000:10000 ENCAP:8

  Path-id 1 advertised to peers:
    192.168.77.11
BGP routing table entry for
[2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272, version 4
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
  AS-Path: NONE, path locally originated
    192.168.100.101 (metric 0) from 0.0.0.0 (192.168.77.101)
```

```
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router
MAC:5e00.0000.0007

  Path-id 1 advertised to peers:
    192.168.77.11

Leaf-101#
```

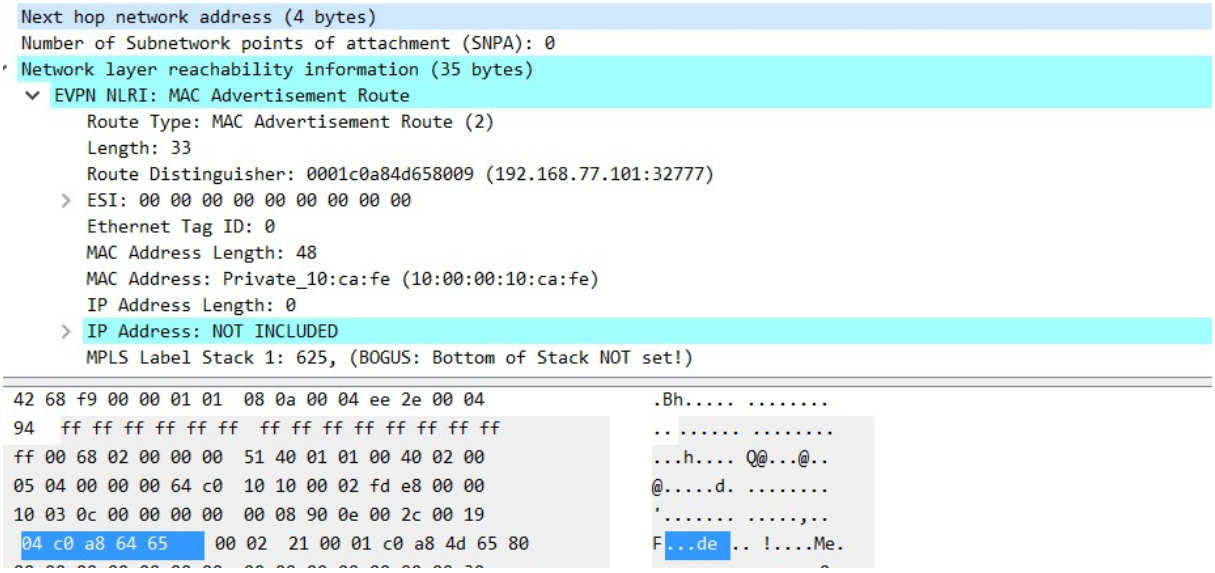
Example 7-11: *BGP EVPN advertisement*

BGP will advertise both mac and mac-IP routes as a separate BGP Route-type 2 update, but we will concentrate on the mac-only advertisement first (Capture 7-1). BGP Update source is BGP RID (Loopback 77). BGP Update has two extended community path attributes. First Extended Community, Route-Target is derived from the BGP AS number and VNI id. This gives us the RT 65000:10000. Second Extended Community defines the encapsulation type which is VXLAN. Under EVPN NLRI: Mac Advertisement Route, there is the Route Distinguisher which is derived from the BGP RID and 32767 + vlan id. This gives as RD 192.168.77.101:32777 for vlan 10.

```
> Frame 76: 170 bytes on wire (1360 bits), 170 bytes captured (1360 bits)
> Ethernet II, Src: 5e:00:00:00:00:07 (5e:00:00:00:00:07), Dst: 5e:00:00:02:00:07 (5e:00:00:02:00:07)
> Internet Protocol Version 4, Src: 192.168.77.101, Dst: 192.168.77.11
> Transmission Control Protocol, Src Port: 51460, Dst Port: 179, Seq: 185, Ack: 58, Len: 104
v Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 104
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 0
    Total Path Attribute Length: 81
  v Path attributes
    > Path Attribute - ORIGIN: IGP
    > Path Attribute - AS_PATH: empty
    > Path Attribute - LOCAL_PREF: 100
    v Path Attribute - EXTENDED_COMMUNITIES
      > Flags: 0xc0, Optional, Transitive, Complete
        Type Code: EXTENDED_COMMUNITIES (16)
        Length: 16
      v Carried extended communities: (2 communities)
        > Route Target: 65000:10000 [Transitive 2-Octet AS-Specific]
        > Encapsulation: VXLAN Encapsulation [Transitive Opaque]
    v Path Attribute - MP_REACH_NLRI
      > Flags: 0x90, Optional, Extended-Length, Non-transitive, Complete
        Type Code: MP_REACH_NLRI (14)
        Length: 44
        Address family identifier (AFI): Layer-2 VPN (25)
        Subsequent address family identifier (SAFI): EVPN (70)
        Next hop network address (4 bytes)
        Number of Subnetwork points of attachment (SNPA): 0
      v Network layer reachability information (35 bytes)
        v EVPN NLRI: MAC Advertisement Route
            Route Type: MAC Advertisement Route (2)
            Length: 33
            Route Distinguisher: 0001c0a84d658009 (192.168.77.101:32777)
          > ESI: 00 00 00 00 00 00 00 00 00 00
            Ethernet Tag ID: 0
            MAC Address Length: 48
            MAC Address: Private_10:ca:fe (10:00:00:10:ca:fe)
            IP Address Length: 0
          > IP Address: NOT INCLUDED
            MPLS Label Stack 1: 625, (BOGUS: Bottom of Stack NOT set!)
```

Capture 7-2: *BGP Update message from Leaf-101 > Spine-11 > Leaf-102*

As can be seen, there is no ip address NLRI specified in this update. The next hop address field is empty, but it can be seen as a HEX format in HEX window (Capture 7-1.1). C0.a8.64.65 = 192.168.100.101 > interface NVE1 IP address.



Capture 7-2.1: *Next Hop Address in BGP Update*

Now we have verified that the local VTEP Leaf-101 has learned the mac address 1000.0010.cafe and installed it to both mac address-table and L2RIB. From L2RIB it is advertised to BGP EVPN instance and from there to BGP EVPN AFI with the vni 10000 specific RD and RT values.

As next step, we are going to check the routing information from the remote VTEP Leaf-102. The process is the same than what we did with local VTEP Leaf-101 but in reversed order. First, we check that the remote VTEP Leaf-102 has received both of the BGP Updates. The Example 7-12 shows that remote VTEP Leaf-102 has received two BGP EVPN type-2 updates. One update regarding host Café mac 1000.0010.cafe/216 and the other one regarding host Café mac-IP 1000.0010.cafe:192.168.11.11/272 (Note that I have skipped some of the address fields). Next, we verify that routes are installed in correct EVPN instances (L2VNI). We can see that remote VTEP Leaf-102 is correctly imported both mac and mac-IP entries from the BGP table to the EVPN instance of VNI 10000. This is done based on Route-Target 65000:32777, which is carried in both routing updates. The last thing to verify from this output is the check that ip-MAC is also installed in VRF Context, otherwise routing between the subnets does not work. As can be seen at the end of the output, the mac-IP route is correctly installed and this is done based on Route-Target 65000:1007 carried only in mac-IP BGP EVPN update.

Note that L3VNI RD is derived from the BGP RID + VRF Id (vrf id can be seen from the output of "show vrf" command) while L2VNI RD is derived from BGP RID + [32767 + VLAN Id].

```
Leaf-102# sh bgp l2vpn evpn 1000.0010.cafe
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 192.168.77.101:32777
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/216,
version 271
```

```
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path
             Imported to 1 destination(s)
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000
      Extcommunity: RT:65000:10000 ENCAP:8
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272,
version 4
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path
             Imported to 3 destination(s)
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router MAC:5e00.0000.0007
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer

Route Distinguisher: 192.168.77.102:32777    (L2VNI 10000)
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/216,
version 272
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, is best path, in rib
             Imported from
192.168.77.101:32777:[2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/216
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000
      Extcommunity: RT:65000:10000 ENCAP:8
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer

BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272,
version 5
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, is best path, in rib
             Imported from
192.168.77.101:32777:[2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router MAC:5e00.0000.0007
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer
```

```
Route Distinguisher: 192.168.77.102:3    (L3VNI 10077)
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272,
version 6
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path
             Imported from
192.168.77.101:32777:[2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router MAC:5e00.0000.0007
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer
```

Example 7-12: *BGP EVPN l2vpn evpn information received from Leaf-101*

If we take a look at the whole BGP table in remote VTEP Leaf-102, we can see that routes concerning to host Cafe are correctly installed (Example 7-13). The output of "show bgp l2vpn evpn" correspond the regular "show bgp" command which shows the BGP table regarding IPv4 afi.

```
Leaf-102# sh bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 288, Local Router ID is 192.168.77.102
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-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 192.168.77.101:32777
*>i[2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/216
                      192.168.100.101                  100          0 i
*>i[2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272
                      192.168.100.101                  100          0 i

Route Distinguisher: 192.168.77.102:32777    (L2VNI 10000)
*>i[2]:[0]:[0]:[48]:[1000.0010.cafe]:[0]:[0.0.0.0]/216
                      192.168.100.101                  100          0 i
*>i[2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272
                      192.168.100.101                  100          0 i

Route Distinguisher: 192.168.77.102:3    (L3VNI 10077)
*>i[2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272
                      192.168.100.101                  100          0 i
```

Example 7-13: *BGP EVPN l2vpn evpn table on remote VTEP-102*

We can see that mac address routing information is produced from the BGP EVPN AFI to EVPN instance (Example 7-14).

```
Leaf-102# sh l2route evpn mac evi 10

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 (Nho):NH-Override
(Pf):Permanently-Frozen

Topology    Mac Address    Prod   Flags          Seq No     Next-Hops
----------- -------------- ------ -------------- ---------- ----------------
10          1000.0010.cafe BGP    SplRcv         0          192.168.100.101
```

Example 7-14: *mac address in EVPN instance (EVI) topology 10 (=vni 10000)*

And from EVPN Instance it is copied to L2RIB (Example 7-15). If we compare this information to local VTEP Leaf-101 L2RIB we can see that on remote VTEP Leaf-102 entry is produced by BGP while in local VTEP Leaf-101 it was locally produced by L2FWDER.

```
Leaf-102# sh l2route mac topology 10

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 (Nho):NH-Override
(Pf):Permanently-Frozen

Topology    Mac Address    Prod   Flags          Seq No     Next-Hops
----------- -------------- ------ -------------- ---------- ----------------
10          1000.0010.cafe BGP    SplRcv         0          192.168.100.101
```

Example 7-15: *L2RIB on remote VTEP Leaf-102*

Finally, we can see that the information is also in mac address table (Example 7-16).

```
Leaf-102# show system internal l2fwder mac | i cafe
*    10    1000.0010.cafe    static   -          F     F  (0x47000001) nve-
peer1 192.168
```

Example 7-16: *Mac address-table on remote VTEP Leaf-102*

Summary

In figure 7-5, the host Café joins to the network and validates its IP address uniqueness by sending a Gratuitous-ARP out of the interface. Local VTEP Leaf-101 learns the mac address 1000.0010.cafe and installs it to mac address-table. The newly created entry in mac address-table is also produced to the L2RIB by L2FWDER. Why? Because we need to advertise the mac route and just like in a regular ip address advertisement, only the routes that are installed in RIB (L2 or L3) could be advertised by routing a protocol. From L2RIB the mac route is sent to the BGP EVPN Address-Family via VNI 10000 EVPN instance. Why via EVPN Instance? We have defined the associated Route-Distinguisher and Route-Target under VNI Specific VNI instance and from there those are attached to the BGP EVPN Update message.

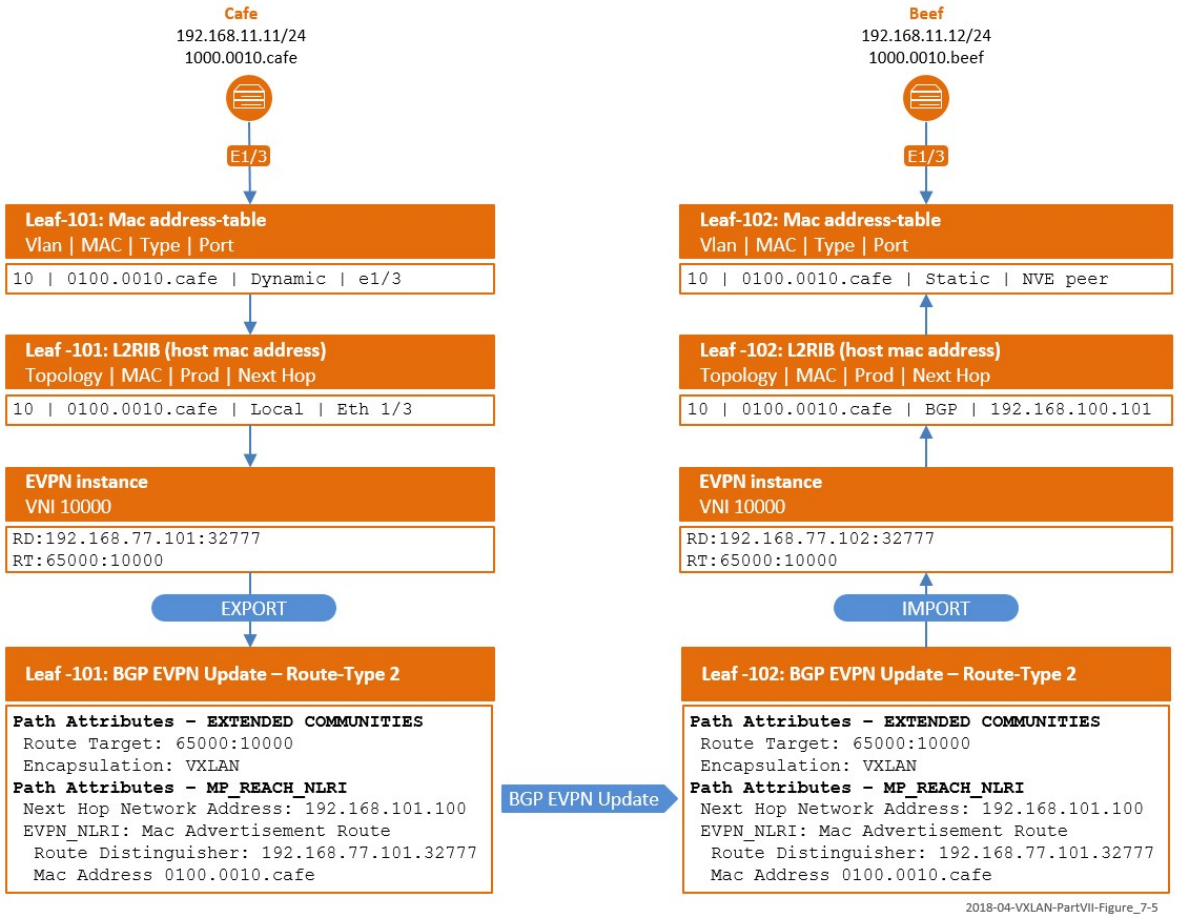


Figure 7-5: L2VNI Summary – Part 1

What have we achieved and verified at this point? Host Café mac address and IP address are now known by the local VTEP Leaf-101 and as well as remote VTEP Leaf-102. This means that inside the VN segment 1000 hosts Café and host Beef are now able to communicate with each other (of course we first have to connect host Beef to the network). The mac-IP information is also installed in ARP suppress-cache of both Leaf switches and switches are able to answer local ARP request messages sent my locally connected hosts without flooding the message over the VXLAN fabric to requested host.

Mac-IP Learning

The previous chapter describes the process of mac learning. Now we will take a closer look at the mac-IP learning process. The Local VTEP Leaf-101 has learned both the mac address and the IP address of host Café from the Gratuitous ARP message. The information is installed in ARP table (Example 7-17) and ARP-suppression cache (Example 7-18).

```
Leaf-101# sh ip arp vrf TENANT77

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 TENANT77
Total number of entries: 1
Address         Age       MAC Address     Interface       Flags
192.168.11.11   00:01:50  1000.0010.cafe  Vlan10
```

Example 7-17: *ARP table on local VTEP Leaf-101*

```
Leaf-101# sh 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

192.168.11.11   00:03:59 1000.0010.cafe   10 Ethernet1/3         L
```

Example 7-18: *ARP suppression-cache on local VTEP Leaf-101*

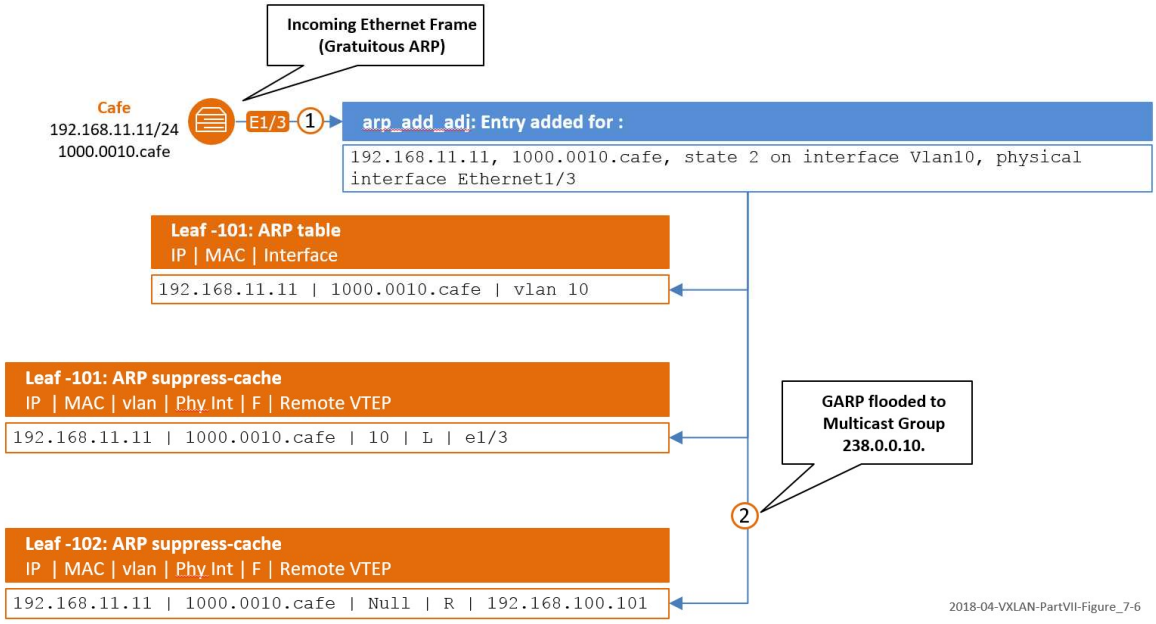


Figure 7-6: *ARP suppression update on Leaf-101 and Leaf-102*

In figure 7-2, we saw how the GARP reply-message was first flooded to Mcast Group 238.0.0.10 and after that to local ARP table and ARP-Suppression cache was updated. In this way, the information is also available for remote VTEP Leaf-102 (Example 7-19).

```
Leaf-102# sh 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

192.168.11.11   03:00:12 1000.0010.cafe   10 (null)              R
192.168.100.101
```

Example 7-19: *ARP suppression-cache on remote Leaf-102*

I am using Nexus 9000v (Cisco VIRL). The process of the L2RIB update differs from the physical Nexus 9000 platform. In Nexus 9000v MAC routes are produced directly into L2RIB by L2FWRED (Example 7-15, Local). MAC-IP routes are produced to the L2RIB and to the L3RIB by the Host Mobility Manager (Example 7-16, HMM) in the same way than in physical switch. Note that there is no Adjacency Manager (AM) component in Nexus 9000v, the command "show forwarding vrf [vrf name] adjacency" does not give any information.

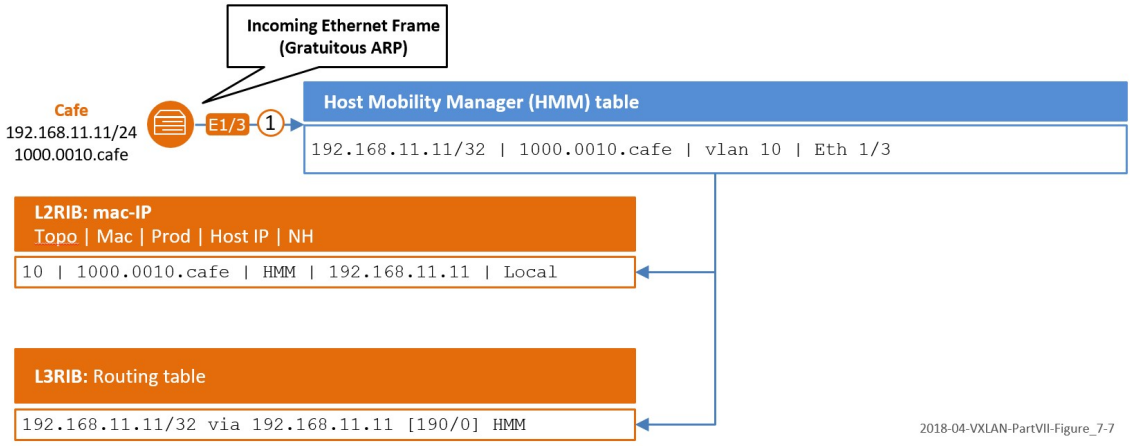


Figure 7-7: *L2RIB and L3RIB tables update by HMM on local VTEP Leaf-101*

Example 7-20 shows the Host Mobility Manager (HMM) table for all known mac-IP entries. Note that at this moment the host Café is the only connected host in our VXLAN fabric.

```
Leaf-101# show fabric forwarding ip local-host-db vrf TENANT77

HMM host IPv4 routing table information for VRF TENANT77
Status: *-valid, x-deleted, D-Duplicate, DF-Duplicate and frozen,
        c-cleaned in 00:02:04
    Host                MAC Address         SVI         Flags       Phy Int
*   192.168.11.11/32    1000.0010.cafe      Vlan10      0x420201    Ethernet1/3
```

Example 7-20: *HMM table on local VTEP Leaf-101*

We can also verify the host-specific information. Example 7-21 shows the Host Mobility Manager table regarding the host Café mac-IP entries.

```
Leaf-101# show fabric forwarding ip local-host-db vrf TENANT77
192.168.11.11/32
HMM routing table information for VRF TENANT77, address family IPv4
HMM routing table entry for 192.168.11.11/32
Hosts: (1 available)

  Host type: Local(Flags: 0x420201), in Rib
  mac: 1000.0010.cafe, svi: Vlan10, bd: 10, phy_intf: Ethernet1/3
```

Example 7-21: *HMM entry about host Café on local VTEP Leaf-101*

Example 7-22 shows the L2RIB mac-IP entry. Here we can see that the Host Mobility Manager (HMM) produces the mac-IP entry. For the mac-only table entry was locally produced by L2FWDER (Example 7-8).

```
Leaf-101# sh l2route evpn mac-ip evi 10 detail
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
----------- -------------- ------ ---------- -------------- --------------
10          1000.0010.cafe HMM    --            0          192.168.11.11  Local
            Sent To: BGP
            L3-Info: 10077
```

Example 7-22: *L2RIB mac entry on local VTEP Leaf-101*

If we take a look at the L3 routing table in Leaf-101, we can see that it is also updated with the host route of Cafe.

```
Leaf-101# sh ip route vrf TENANT | sec 192.168.11.11
192.168.11.11/32, ubest/mbest: 1/0, attached
    *via 192.168.11.11, Vlan10, [190/0], 02:24:14, hmm
```

Example 7-23: *L2RIB mac entry on local VTEP Leaf-101*

In addition to mac-only information, L2RIB is updated also with the mac-IP information. Next, the mac-IP information needs to be advertised to remote VTEP Leaf-102 so the mac-IP information from the L2RIB is sent to BGP VRF Afi (Example 7-24) where it is further is advertised to the remote VTEP Leaf-102 via Route Reflector Spine-11.

```
Leaf-101# sh bgp l2vpn evpn 192.168.11.11 vrf TENANT77
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 192.168.77.101:32777    (L2VNI 10000)
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272,
version 8
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
  AS-Path: NONE, path locally originated
    192.168.100.101 (metric 0) from 0.0.0.0 (192.168.77.101)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router MAC:5e00.0000.0007

  Path-id 1 advertised to peers:
    192.168.77.11
```

Example 7-24: *BGP table entry on local VTEP Leaf-101*

Note that Leaf-101 sends two separate BGP EVPN Updates regarding the mac and ip addresses of host Café. Mac-only update is sent with RT65000:10000 while mac-IP Update entry is sent with an additional RT 65000:10077.

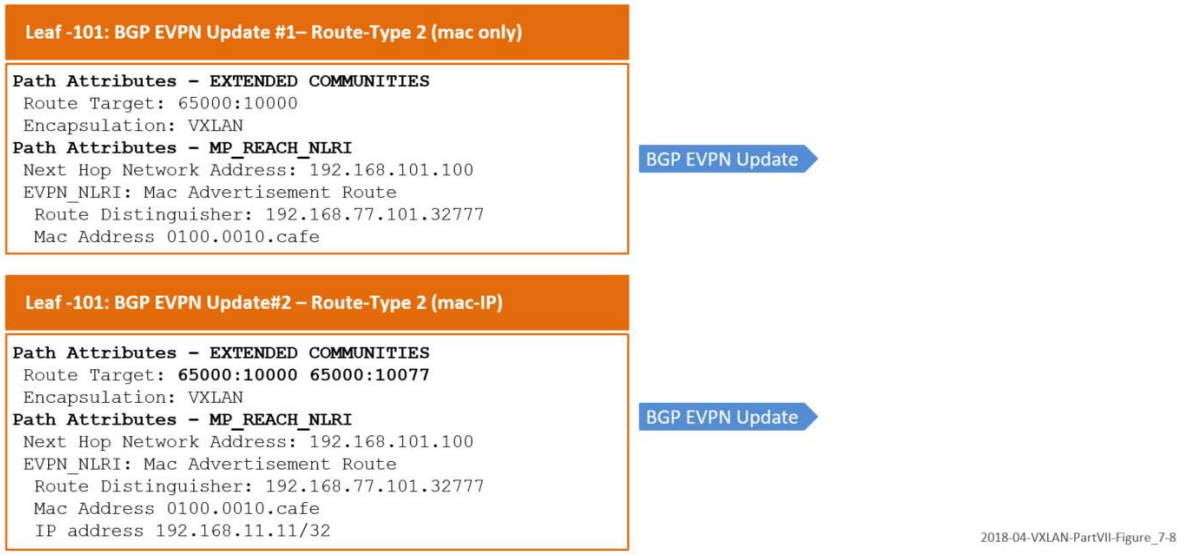


Figure 7-8: *BGP EVPN Updates*

We can see from the remote VTEP Leaf-102 BGP table that it has received BGP EVPN Update from the VTEP Leaf-101 (Example 7-25). Based on the Route-Target values, it imports these routes to correct tables (this was explained in the explanation regarding Example 7-12).

```
Leaf-102# sh bgp l2vpn evpn 192.168.11.11 vrf TENANT77
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 192.168.77.101:32777
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272,
version 4
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path
             Imported to 3 destination(s)
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router MAC:5e00.0000.0007
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer

Route Distinguisher: 192.168.77.102:32777    (L2VNI 10000)
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272,
version 5
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, is best path, in rib
             Imported from 192.168.77.101:32777:[2]:[0]:[0]:[48]:[1000.0010.cafe
]:[32]:[192.168.11.11]/272
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router MAC:5e00.0000.0
007
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer

Route Distinguisher: 192.168.77.102:3    (L3VNI 10077)
BGP routing table entry for [2]:[0]:[0]:[48]:[1000.0010.cafe]:[32]:[192.168.11.11]/272,
version 6
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path
             Imported from 192.168.77.101:32777:[2]:[0]:[0]:[48]:[1000.0010.cafe
]:[32]:[192.168.11.11]/272
  AS-Path: NONE, path sourced internal to AS
    192.168.100.101 (metric 81) from 192.168.77.11 (192.168.77.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10000 10077
      Extcommunity: RT:65000:10000 RT:65000:10077 ENCAP:8 Router MAC:5e00.0000.0007
      Originator: 192.168.77.101 Cluster list: 192.168.77.111

  Path-id 1 not advertised to any peer

Leaf-102#
```

Example 7-25: *BGP table entry*

The route is installed to the L2RIB produced by BGP (Example 7-26). Note also that information is sent to the ARP process.

```
Leaf-102# sh l2route evpn mac-ip evi 10 detail
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
----------- -------------- ------ ---------- -------------- ---------------
10          1000.0010.cafe BGP    --            0          192.168.11.11
192.168.100.101
            Sent To: ARP
```

Example 7-26: *L2RIB on remote VTEP Leaf 102*

However, the vrf specific ARP table on Leaf-102 does not have an ARP entry (Example 7-26).

```
Leaf-102# sh ip arp vrf TENANT77

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 TENANT77
```

```
Total number of entries: 0
Address         Age       MAC Address     Interface       Flags
```

Example 7-26: *ARP table on remote VTEP Leaf-102*

But information is installed to ARP suppression-cache (Example 7-27).

```
Leaf-102# sh 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

192.168.11.11   03:53:24 1000.0010.cafe   10 (null)              R
192.168.100.101
```

Example 7-27: *ARP Suppression-cache on remote VTEP Leaf-102*

Host-specific ip routing entry is also installed to RIB (Example 7-28).

```
Leaf-102# sh ip route vrf TENANT77 | sec 192.168.11.11
192.168.11.11/32, ubest/mbest: 1/0
    *via 192.168.100.101%default, [200/0], 03:03:26, bgp-65000, internal, tag 65000
(evpn) segid: 10077 tunnelid: 0xc0a86465 encap: VXLAN
```

Example 7-28: *RIB entry on remote VTEP Leaf-102*

Summary

Remote VTEP Leaf-102 receives two separate BGP EVPN Update (Figure 7-9). The first one is the mac –only update where we only have on RT value 65000:10000. Based on this RT the mac information is imported to L2VNI specific tables (mac, L2RIB) and the information is used for switching frames between the hosts in the same L2 VNI.
The other mac-IP BGP EVPN Update has two RT values: 65000:10000 and 65000:10077. Based on the RT 65000:10000 Mac-IP route is installed to the L2VNI specific table just like the previous one. This information is used for ARP process. Information is stored to ARP suppression-cache and if locally connected hosts try to resolve the mac address of specific IP with the ARP request, the local switch is able to reply with ARP reply message. This reduces BUM traffic.
Based on the RT value 65000-10077 in second BGP EVPN Update the route is installed to L3VNI specific L3RIB and it is used for routing packets between the hosts in different subnets inside a vrf/tenant.

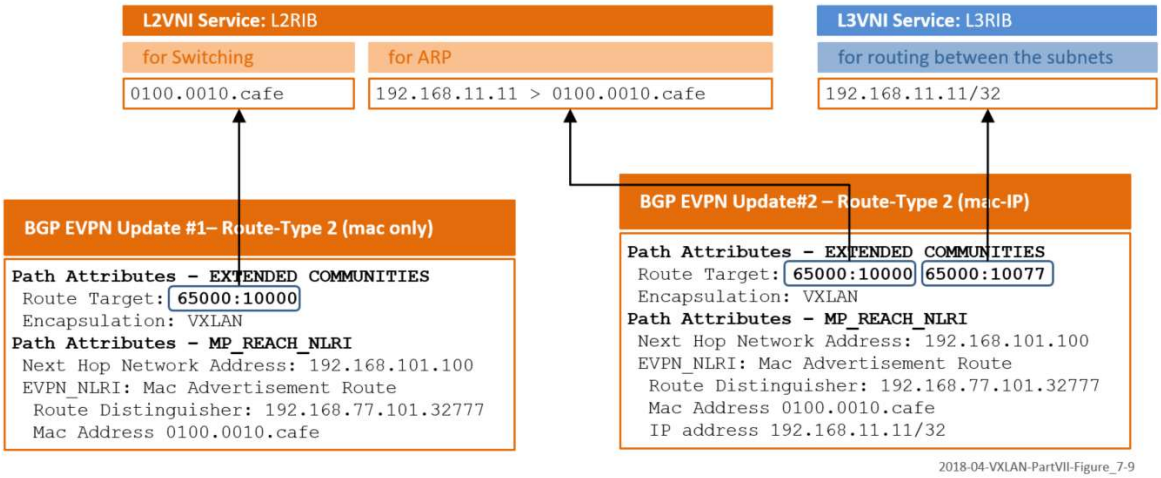


Figure 7-9: *Route import on remote VTEP Leaf-102*

Data Plane testing

Now I am going to connect host Beef (192.168.11.12) remote VTEP Leaf-102. Then both VTEP switches updates their ARP table and ARP Suppression-cache.

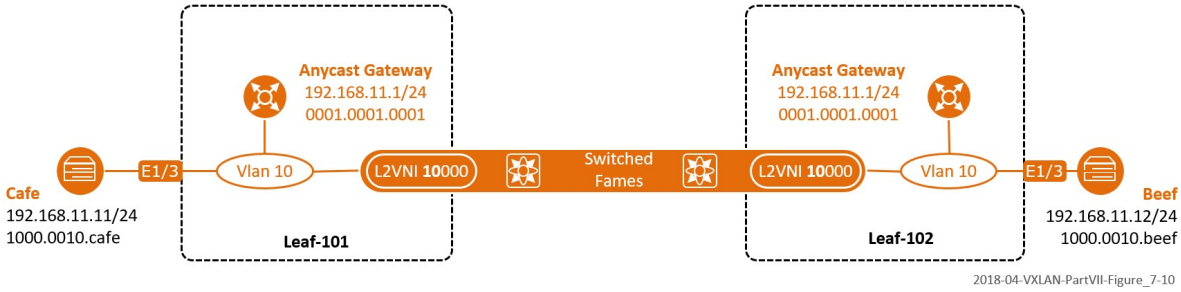


Figure 7-10: *Data Plane testing inside a vni 10000*

ARP suppression-cache verification from VTEP Leaf-101 (Example 7-24)

```
Leaf-101# sh ip arp suppression-cache vlan 10
<Snipped>
Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep
Addrs

192.168.11.11   00:01:01 1000.0010.cafe   10 Ethernet1/3         L
192.168.11.12   00:03:26 1000.0010.beef   10 (null)              R
192.168.100.102
```

Example 7-24: *ARP suppression-cache on Leaf-101*

ARP suppression-cache verification from VTEP Leaf-102 (Example 7-25)

```
Leaf-102# sh ip arp suppression-cache vlan 10
<Snipped>
Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep
Addrs

192.168.11.12   00:03:01 1000.0010.beef   10 Ethernet1/3         L
192.168.11.11   04:32:02 1000.0010.cafe   10 (null)              R
192.168.100.101
```

Example 7-25: *ARP suppression-cache on Leaf-102*

Now I am going to turn on the ARP debugs (events and packets) on Leaf-101 and then ping from the Host Café (192.168.11.11) to Host Beef (192.168.11.12).

Ping works fine, the first reply is missing because of ARP request (Example 7-26).

```
Cafe#ping 192.168.11.12
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.11.12, timeout is 2 seconds:
.!!!!
Success rate is 80 percent (4/5), round-trip min/avg/max = 14/19/23 ms
```

Example 7-26: *Ping from host Café to host Beef*

Now the local VTEP Leaf-101 is able to answer the ARP-request message since it has information stored in ARP suppression-cache (Example 7-27). So when the host joins the network for the first time, it sends a Gratuitous ARP message just to make sure that the IP address assigned to it is unique. This message is flooded to other VTEP leaf switches since neither the ARP-table or the ARP-suppression cache has no entry regarding asked IP-mac binding. After these tables are updated and there is no need for ARP request flooding.

```
arp: (context 3) Receiving packet from Vlan10, logical interface Vlan10 physical
interface Ethernet1/3

arp: Src 1000.0010.cafe/192.168.11.11 Dst 0000.0000.0000/192.168.11.12

arp_send_response_internal: ARP response from 192.168.11.12 to 192.168.11.11 on Vlan10,
phy iod Ethernet1/3, vlan 10

arp: Src 1000.0010.beef/192.168.11.12 Dst 1000.0010.cafe/192.168.11.11
```

Example 7-27: *ARP process in Leaf-101*

This phase we should also have ip connectivity between the hosts in different vlan. Now I am going to do some ping testing and while capturing the ip packets to see what VNI tag is used in between the hosts in the same vlan and between the hosts in different vlan.

First, ping test is again from host Café (192.168.11.11) to Host Beef (192.168.11.12)

```
Cafe#ping 192.168.11.12
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.11.12, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 16/18/21 ms
```

Example 7-28: *Ping from café to Beef (hosts in same subnet)*

As can be seen from capture 7-2, the VNI segment Id in VXLAN header is 10000 as it should be.

```
> Frame 4: 164 bytes on wire (1312 bits), 164 bytes captured (1312 bits)
> Ethernet II, Src: 5e:00:00:00:00:07 (5e:00:00:00:00:07), Dst: 5e:00:00:02:00:07 (5e:00:00:02:00:07)
> Internet Protocol Version 4, Src: 192.168.100.101, Dst: 192.168.100.102
> User Datagram Protocol, Src Port: 61763, Dst Port: 4789
v Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 10000
    Reserved: 0
> Ethernet II, Src: Private_10:ca:fe (10:00:00:10:ca:fe), Dst: Private_10:be:ef (10:00:00:10:be:ef)
> Internet Protocol Version 4, Src: 192.168.11.11, Dst: 192.168.11.12
> Internet Control Message Protocol
```

Capture 7-3: *Ping from café to Beef (hosts in the same subnet but in different VTEP)*

Second ping test is from host Café (192.168.11.11) to host Babe (192.168.12.12) in vlan 20 on remote VTEP Leaf-102 (192.168.12.12). Before that, I need to connect the hosts to the network. I also connect the host Abba (192.168.12.11) in vlan 20 on Leaf-101 to the network. After a very short period we should have updated ARP-Suppression-cache entries. I am going to verify that first.

Looks good on VTEP Leaf-101

```
Leaf-101# sh ip arp suppression-cache detail

<snip>

Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep
Addrs

192.168.11.11   00:09:21 1000.0010.cafe   10 Ethernet1/3         L
192.168.11.12   00:26:48 1000.0010.beef   10 (null)              R      192.168.100.102
192.168.12.11   00:02:51 2000.0020.abba   20 Ethernet1/4         L
192.168.12.12   00:02:34 2000.0020.babe   20 (null)              R      192.168.100.102
```

Example 7-29: *ARP suppression-cache in Leaf-101*

As well as in VTEP Leaf-102

```
Leaf-102# sh ip arp suppression-cache detail

<snip>

Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep
Addrs

192.168.11.12   00:03:03 1000.0010.beef   10 Ethernet1/3         L
192.168.11.11   04:57:08 1000.0010.cafe   10 (null)              R      192.168.100.101
192.168.12.12   00:03:53 2000.0020.babe   20 Ethernet1/4         L
192.168.12.11   00:04:10 2000.0020.abba   20 (null)              R      192.168.100.101
```

Example 7-30: *ARP suppression-cache in Leaf-102*

Now I am going to ping from Café (192.168.11.11) to Babe (192.168.12.12). The data path is can be seen in figure 7-11.

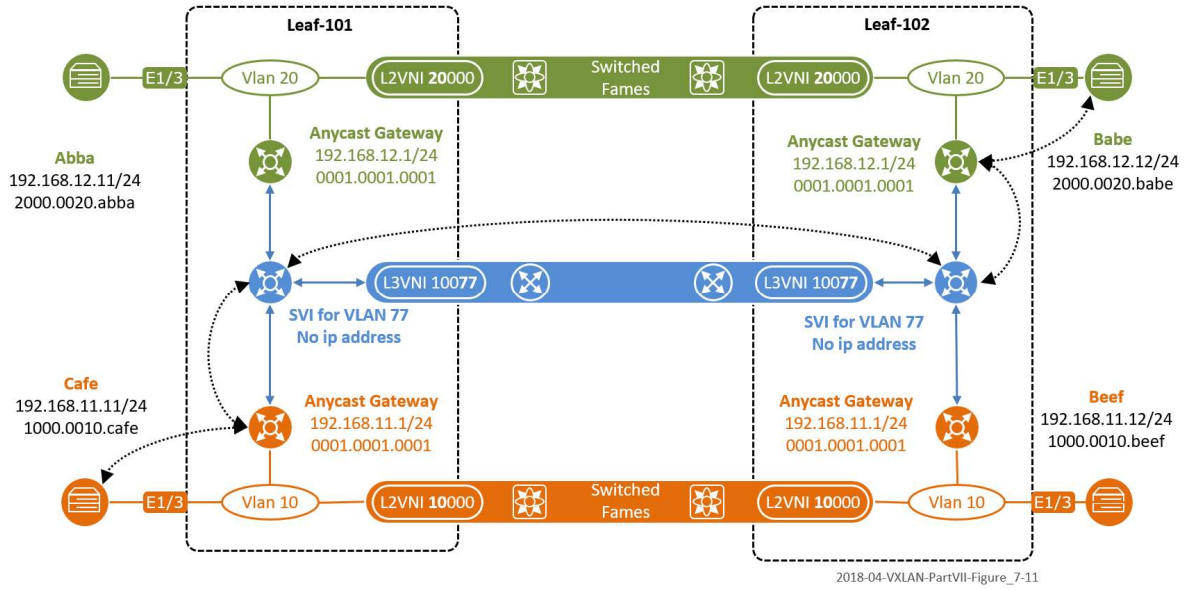


Figure 7-11: *Packet flow between the hosts in different subnets inside a VRF/Tenant*

The first ICMP reply is missed because of the ARP. The host does not send anything through the gateway, so first it has to resolve the mac address of its gateway 192.168.11.1.

```
Cafe#ping 192.168.12.12
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.12.12, timeout is 2 seconds:
.!!!!
Success rate is 80 percent (4/5), round-trip min/avg/max = 15/21/35 ms
```

Example 7-31: *Ping from host Café to babe*

As can be seen from capture 7-4, the VNI segment Id in VXLAN header is 10077 as it should be.

```
> Frame 368: 164 bytes on wire (1312 bits), 164 bytes captured (1312 bits)
> Ethernet II, Src: 5e:00:00:00:00:07 (5e:00:00:00:00:07), Dst: 5e:00:00:02:00:07 (5e:00:00:02:00:07)
> Internet Protocol Version 4, Src: 192.168.100.101, Dst: 192.168.100.102
> User Datagram Protocol, Src Port: 60963, Dst Port: 4789
v Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 10077
    Reserved: 0
> Ethernet II, Src: 5e:00:00:00:00:07 (5e:00:00:00:00:07), Dst: 5e:00:00:01:00:07 (5e:00:00:01:00:07)
> Internet Protocol Version 4, Src: 192.168.11.11, Dst: 192.168.12.12
> Internet Control Message Protocol
```

Capture 7-3: *Ping from Café to Babe (hosts in the different subnet in different VTEP)*

References:

Building Data Center with VXLAN BGP EVPN – A Cisco NX-OS Perspective
ISBN-10: 1-58714-467-0 – Krattiger Lukas, Shyam Kapadia, and Jansen Davis

Cisco Live 2018 - BRKDCN-3040: Troubleshooting VxLAN BGP EVPN – Vinit Jain

Cisco Nexus 9000v Guide:
Chapter: Troubleshooting the Cisco Nexus 9000v

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/nx-osv/configuration/guide/b_Cisco_Nexus_9000v/b_NX-OSv_9000_chapter_010.html#reference_34AB8A043C3E4177BEFEDF95F54386A4

Appendix 1.

Topology

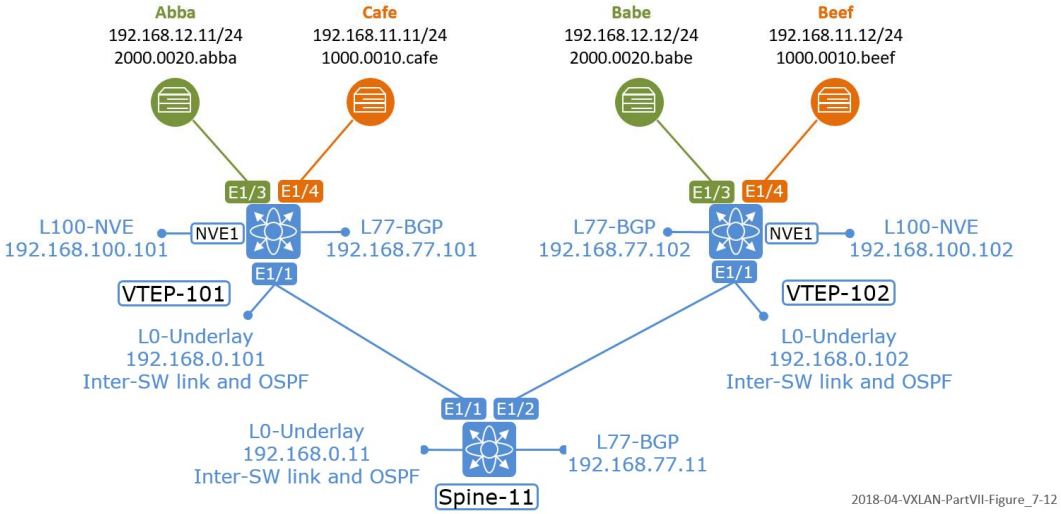


Figure 7-12: *Physical network topology*

The configuration of the VTEP Leaf-101

```
Leaf-101# sh run

!Command: show running-config
!Time: Mon Apr 16 12:48:51 2018

version 7.0(3)I7(1)
hostname Leaf-101
vdc Leaf-101 id 1
  limit-resource vlan minimum 16 maximum 4094
  limit-resource vrf minimum 2 maximum 4096
  limit-resource port-channel minimum 0 maximum 511
  limit-resource u4route-mem minimum 128 maximum 128
  limit-resource u6route-mem minimum 96 maximum 96
  limit-resource m4route-mem minimum 58 maximum 58
  limit-resource m6route-mem minimum 8 maximum 8

nv overlay evpn
feature ospf
feature bgp
feature pim
feature fabric forwarding
feature interface-vlan
feature vn-segment-vlan-based
feature nv overlay

no password strength-check
username admin password 5
$5$aV2kcO97$7ioNn2XTmsfuFj62MLL/wcMnEoJE9ifSY/AFfWPY2/
/  role network-admin
```

```
ip domain-lookup
ip host Spine-12 192.168.0.12
snmp-server user admin network-admin auth md5
0x223cfb63ca87c5b4856c960235329cff
 priv 0x223cfb63ca87c5b4856c960235329cff localizedkey
rmon event 1 description FATAL(1) owner PMON@FATAL
rmon event 2 description CRITICAL(2) owner PMON@CRITICAL
rmon event 3 description ERROR(3) owner PMON@ERROR
rmon event 4 description WARNING(4) owner PMON@WARNING
rmon event 5 description INFORMATION(5) owner PMON@INFO

fabric forwarding anycast-gateway-mac 0001.0001.0001
ip pim rp-address 192.168.238.1 group-list 238.0.0.0/24 bidir
ip pim ssm range 232.0.0.0/8
vlan 1,10,20,77
vlan 10
  name L2VNI-for-VLAN10
  vn-segment 10000
vlan 20
  name L2VNI-for-VLAN20
  vn-segment 20000
vlan 77
  name TENANT77
  vn-segment 10077

vrf context TENANT77
  vni 10077
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
vrf context management
hardware access-list tcam region racl 512
hardware access-list tcam region arp-ether 256 double-wide


interface Vlan1
  no shutdown

interface Vlan10
  no shutdown
  vrf member TENANT77
  ip address 192.168.11.1/24
  fabric forwarding mode anycast-gateway

interface Vlan20
  no shutdown
  vrf member TENANT77
  ip address 192.168.12.1/24
  fabric forwarding mode anycast-gateway

interface Vlan77
  no shutdown
  vrf member TENANT77
  ip forward
```

```
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback100
  member vni 10000
    suppress-arp
    mcast-group 238.0.0.10
  member vni 10077 associate-vrf
  member vni 20000
    suppress-arp
    mcast-group 238.0.0.10

interface Ethernet1/1
  no switchport
  medium p2p
  ip unnumbered loopback0
  ip ospf network point-to-point
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/2
  no switchport
  medium p2p
  ip unnumbered loopback0
  ip ospf network point-to-point
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/3
  switchport access vlan 10

interface Ethernet1/4
  switchport access vlan 20

<empty interfaces removed from configuration output>

interface mgmt0
  vrf member management

interface loopback0
  description ** RID/Underlay **
  ip address 192.168.0.101/32
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode

interface loopback77
  description ** BGP peering **
  ip address 192.168.77.101/32
  ip router ospf UNDERLAY-NET area 0.0.0.0

interface loopback100
  description ** VTEP/Overlay **
  ip address 192.168.100.101/32
  ip router ospf UNDERLAY-NET area 0.0.0.0
```

```
  ip pim sparse-mode
line console
line vty
router ospf UNDERLAY-NET
  router-id 192.168.0.101
  name-lookup
router bgp 65000
  router-id 192.168.77.101
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 192.168.77.11
    remote-as 65000
    description ** Spine-11 BGP-RR **
    update-source loopback77
    address-family l2vpn evpn
      send-community extended
  vrf TENANT77
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 10000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto


Leaf-101#
```

The configuration of the VTEP Leaf-102

```
Leaf-102# sh run

!Command: show running-config
!Time: Mon Apr 16 12:51:04 2018

version 7.0(3)I7(1)
hostname Leaf-102
vdc Leaf-102 id 1
  limit-resource vlan minimum 16 maximum 4094
  limit-resource vrf minimum 2 maximum 4096
  limit-resource port-channel minimum 0 maximum 511
  limit-resource u4route-mem minimum 128 maximum 128
  limit-resource u6route-mem minimum 96 maximum 96
  limit-resource m4route-mem minimum 58 maximum 58
  limit-resource m6route-mem minimum 8 maximum 8

nv overlay evpn
feature ospf
feature bgp
feature pim
feature fabric forwarding
```

```
feature interface-vlan
feature vn-segment-vlan-based
feature nv overlay

username admin password 5
$5$r25DfmPc$EvUgSVebL3gCPQ8e1ngSTxeKYIk4yuuPIomJKa5Lp/
3  role network-admin
ip domain-lookup
ip host Leaf-102 192.168.0.102
ip host Spine-11 192.168.0.11
snmp-server user admin network-admin auth md5
0x713961e592dd5c2401317a7e674464ac
 priv 0x713961e592dd5c2401317a7e674464ac localizedkey
rmon event 1 description FATAL(1) owner PMON@FATAL
rmon event 2 description CRITICAL(2) owner PMON@CRITICAL
rmon event 3 description ERROR(3) owner PMON@ERROR
rmon event 4 description WARNING(4) owner PMON@WARNING
rmon event 5 description INFORMATION(5) owner PMON@INFO

fabric forwarding anycast-gateway-mac 0001.0001.0001
ip pim rp-address 192.168.238.1 group-list 238.0.0.0/24 bidir
ip pim ssm range 232.0.0.0/8
vlan 1,10,20,77
vlan 10
  name L2VNI-for-VLAN10
  vn-segment 10000
vlan 20
  name L2VNI-for-VLAN20
  vn-segment 20000
vlan 77
  name TENANT77
  vn-segment 10077

vrf context TENANT77
  vni 10077
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
vrf context management
hardware access-list tcam region racl 512
hardware access-list tcam region arp-ether 256 double-wide


interface Vlan1
  no shutdown

interface Vlan10
  no shutdown
  vrf member TENANT77
  ip address 192.168.11.1/24
  fabric forwarding mode anycast-gateway

interface Vlan20
  no shutdown
  vrf member TENANT77
```

```
  ip address 192.168.12.1/24
  fabric forwarding mode anycast-gateway

interface Vlan77
  no shutdown
  vrf member TENANT77
  ip forward

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback100
  member vni 10000
    suppress-arp
    mcast-group 238.0.0.10
  member vni 10077 associate-vrf
  member vni 20000
    suppress-arp
    mcast-group 238.0.0.10

interface Ethernet1/1
  no switchport
  medium p2p
  ip unnumbered loopback0
  ip ospf network point-to-point
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/2
  no switchport
  medium p2p
  ip unnumbered loopback0
  ip ospf network point-to-point
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/3
  switchport access vlan 10

interface Ethernet1/4
  switchport access vlan 20

<empty interfaces removed from configuration output>

interface mgmt0
  vrf member management

interface loopback0
  description ** RID/Underlay **
  ip address 192.168.0.102/32
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode

interface loopback77
```

```
  description ** BGP peering **
  ip address 192.168.77.102/32
  ip router ospf UNDERLAY-NET area 0.0.0.0

interface loopback100
  description ** VTEP/Overlay **
  ip address 192.168.100.102/32
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
line console
line vty
router ospf UNDERLAY-NET
  router-id 192.168.0.102
  name-lookup
router bgp 65000
  router-id 192.168.77.102
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 192.168.77.11
    remote-as 65000
    description ** Spine-11 BGP-RR **
    update-source loopback77
    address-family l2vpn evpn
      send-community extended
  vrf TENANT77
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 10000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto


Leaf-102#
```

The configuration of the Core switch Spine-11

```
Spine-11# sh run

!Command: show running-config
!Time: Mon Apr 16 12:53:17 2018

version 7.0(3)I7(1)
hostname Spine-11
vdc Spine-11 id 1
  limit-resource vlan minimum 16 maximum 4094
  limit-resource vrf minimum 2 maximum 4096
  limit-resource port-channel minimum 0 maximum 511
  limit-resource u4route-mem minimum 128 maximum 128
  limit-resource u6route-mem minimum 96 maximum 96
```

```
  limit-resource m4route-mem minimum 58 maximum 58
  limit-resource m6route-mem minimum 8 maximum 8

nv overlay evpn
feature ospf
feature bgp
feature pim
feature vn-segment-vlan-based
feature nv overlay

no password strength-check
username admin password 5
$5$60DVUPIV$uZWPu6ufHQOJSG18SK5b9/5kpZnV5E4/EFapzQP5CI
/  role network-admin
ip domain-lookup
ip host Spine-12 192.168.0.12
ip host Leaf-102 192.168.0.102
snmp-server user admin network-admin auth md5
0xd177fd3448eab21dd2feb16d54938469
 priv 0xd177fd3448eab21dd2feb16d54938469 localizedkey
rmon event 1 description FATAL(1) owner PMON@FATAL
rmon event 2 description CRITICAL(2) owner PMON@CRITICAL
rmon event 3 description ERROR(3) owner PMON@ERROR
rmon event 4 description WARNING(4) owner PMON@WARNING
rmon event 5 description INFORMATION(5) owner PMON@INFO

ip pim rp-address 192.168.238.1 group-list 238.0.0.0/24 bidir
ip pim ssm range 232.0.0.0/8
vlan 1

vrf context management

interface Ethernet1/1
  no switchport
  medium p2p
  ip unnumbered loopback0
  ip ospf network point-to-point
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/2
  no switchport
  medium p2p
  ip unnumbered loopback0
  ip ospf network point-to-point
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
  no shutdown

<empty interfaces removed from configuration output>

interface mgmt0
  vrf member management

interface loopback0
```

```
  description ** RID/Underlay **
  ip address 192.168.0.11/32
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode

interface loopback77
  description ** BGP peering **
  ip address 192.168.77.11/32
  ip router ospf UNDERLAY-NET area 0.0.0.0

interface loopback238
  description ** Anycast-RP address **
  ip address 192.168.238.6/29
  ip ospf network point-to-point
  ip router ospf UNDERLAY-NET area 0.0.0.0
  ip pim sparse-mode
line console
line vty
router ospf UNDERLAY-NET
  router-id 192.168.0.11
  name-lookup
router bgp 65000
  router-id 192.168.77.111
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 192.168.77.101
    remote-as 65000
    update-source loopback77
    address-family l2vpn evpn
      send-community
      send-community extended
      route-reflector-client
  neighbor 192.168.77.102
    remote-as 65000
    update-source loopback77
    address-family l2vpn evpn
      send-community
      send-community extended
      route-reflector-client


Spine-11#
```