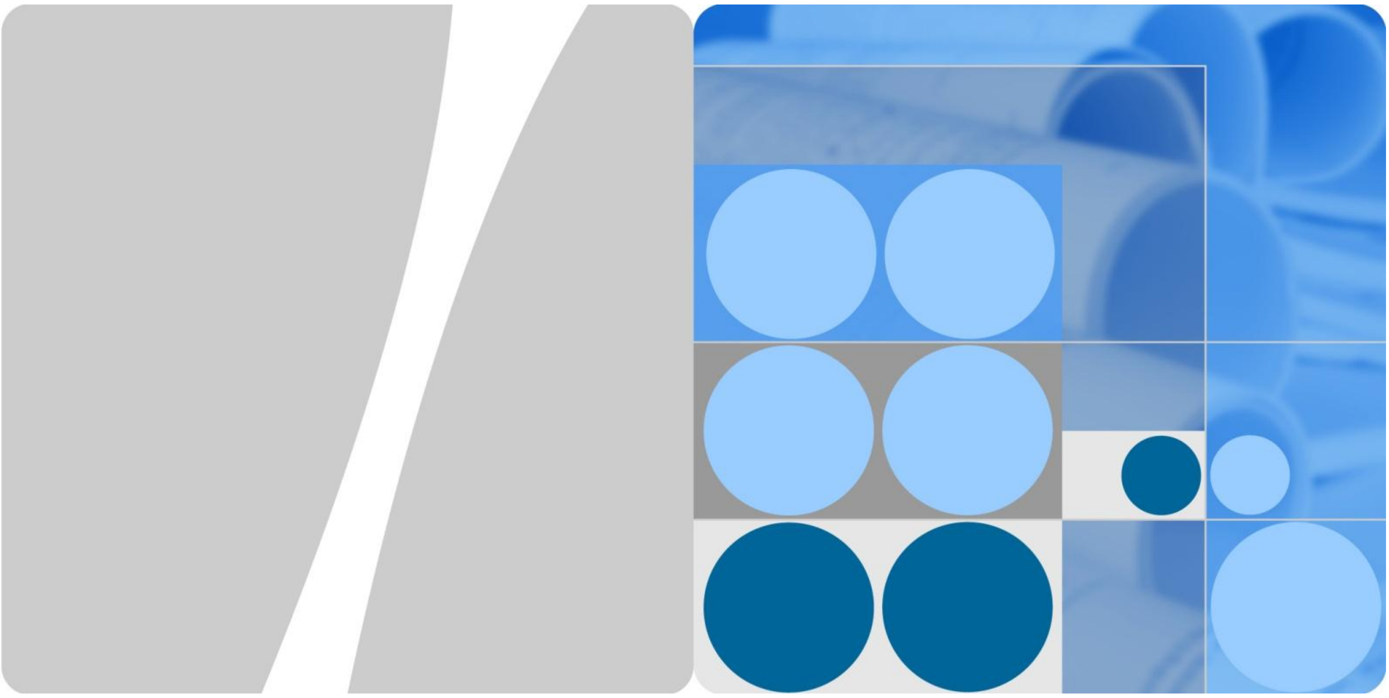


Huawei CloudEngine Series Switches

EVPN Solution White Paper

Issue 01

Date 2016-12-12

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://e.huawei.com

Contents

1 Executive Summary..1

2 Procedure..2

2.1 Automatic VXLAN Tunnel Discovery..2

2.2 Host MAC Address Advertisement 3

2.3 Host IP Address Advertisement...4

2.4 Host ARP Entry Advertisement..5

3 Solution 7

3.1 Centralized Gateway Deployment..7

3.1.1 Intra-subnet L2 Communication..7

3.1.2 Inter-subnet L3 Communication..10

3.1.3 Intra-subnet L2 BUM Traffic Forwarding...11

3.1.4 ARP Broadcast-to-Unicast Conversion 13

3.2 Distributed Gateway Deployment 14

3.2.1 Intra-subnet L2 Communication..14

3.2.2 Inter-subnet L3 Communication..14

3.2.3 Intra-subnet L2 BUM Traffic Forwarding...16

3.2.4 ARP Broadcast-to-Unicast Conversion 16

3.2.5 Virtual Machine Migration...16

3.2.6 Active-Active High Availability of Hosts...17

3.3 Use of Two BGP Processes 18

4 Typical Applications..19

4.1 Centralized Gateway Deployment..19

4.2 Distributed Gateway Deployment 20

4.3 BGP Deployment on Distributed Gateways 21

4.4 DCI...23

4.5 Underlay EBGP + Overlay IBGP Deployment...25

5 Conclusion 27

6 Acronyms and Abbreviations..28

1 Executive Summary

This document describes the solution that uses EVPN as the signaling protocol to advertise L2/L3 information of hosts on a data center network using hardware VXLAN gateways. This solution can be used in three scenarios: centralized gateway deployment, distributed VXLAN gateway deployment, and data center interconnection (DCI).

Keywords: Ethernet Virtual Private Network (EVPN), Virtual eXtensible Local Area Network (VXLAN), VXLAN Tunnel End Point (VTEP)

2 Procedure

2.1 Automatic VXLAN Tunnel Discovery

On current VXLAN networks, VXLAN tunnels are manually created by configuring VXLAN network identifiers (VNIs) and peer lists of VNIs. The configuration is difficult for users and requires heavy manual workload.

When EVPN is used as the control plane on a VXLAN network, L2 VNIs and VTEP addresses can be advertised using EVPN inclusive routes, enabling L2 VXLAN tunnels to be created dynamically. This solution greatly reduces manual configuration and simplifies network deployment.

-EVPN defines Type 3 routes, namely, inclusive multicast routes, through which local L2 VNI and L2 VTEP information can be advertised to implement automatic VXLAN tunnel discovery. An inclusive multicast route contains the following:

Prefix:

```
+--------------------------------------+
| RD (8 octets) |----Route distinguisher of an EVPN instance
+--------------------------------------+
| Ethernet Tag ID (4 octets) |
+--------------------------------------+
| IP Address Length (1 octet) |
+--------------------------------------+
| Originating Router's IP Address |----VTEP IP address
| (4 or 16 octets) |
+--------------------------------------+
```

Provider Multicast Service Interface (PMSI) attribute:

```
+--------------------------------+
| Flags (1 octet) |
+--------------------------------+
```

| Tunnel Type (1 octets) |

+--------------------------------+

| MPLS Label (3 octets)|----L2 VNI

+--------------------------------+

| Tunnel Identifier (variable) |----VTEP IP address

+--------------------------------+

2.2 Host MAC Address Advertisement

Hosts on the same subnet (same network segment) can communicate through known unicast packets after they learn MAC addresses of each other.

If EVPN is not used as the control plane on a VXLAN network, host MAC addresses are learned through flooding of data traffic. EVPN defines Type 2 routes, namely, MAC/IP routes, to advertise host MAC addresses, reducing traffic floods on the VXLAN network.

+--------------------------------------+

| RD (8 octets) |---RD of an EVPN instance

+--------------------------------------+

|Ethernet Segment Identifier (10 octets)|

+--------------------------------------+

| Ethernet Tag ID (4 octets) |

+--------------------------------------+

| MAC Address Length (1 octet) |

+--------------------------------------+

| MAC Address (6 octets) |---Host MAC address

+--------------------------------------+

| IP Address Length (1 octet) |

+--------------------------------------+

| IP Address (0, 4, or 16 octets) |

+--------------------------------------+

| MPLS Label1 (3 octets) |---L2 VNI

+--------------------------------------+

| MPLS Label2 (0 or 3 octets) |

+--------------------------------------+

2.3 Host IP Address Advertisement

EVPN defines Type 5 routes, namely, IP-prefix routes, to advertise host IP routes, enabling hosts on different subnets to communicate with each other.

```
+---------------------------------------+
| RD (8 octets) |
+---------------------------------------+
|Ethernet Segment Identifier (10 octets)|
+---------------------------------------+
| Ethernet Tag ID (4 octets) |
+---------------------------------------+
| IP Prefix Length (1 octet) |
+---------------------------------------+
| IP Prefix (4 or 16 octets) |----Host IP address
+---------------------------------------+
| GW IP Address (4 or 16 octets) |
+---------------------------------------+
| MPLS Label (3 octets) |----L3 VNI
+---------------------------------------+
```

Host IP routes can also be advertised through EVPN Type 2 routes, namely, MAC/IP routes.

```
+---------------------------------------+
| RD (8 octets) |---RD of an EVPN instance
+---------------------------------------+
|Ethernet Segment Identifier (10 octets)|
+---------------------------------------+
| Ethernet Tag ID (4 octets) |
+---------------------------------------+
| MAC Address Length (1 octet) |
+---------------------------------------+
| MAC Address (6 octets) |---Host MAC address
+---------------------------------------+
| IP Address Length (1 octet) |
+---------------------------------------+
| IP Address (0, 4, or 16 octets) |---Host IP address
+--------------------------------------+
| MPLS Label1 (3 octets)|---L2 VNI
+--------------------------------------+
| MPLS Label2 (0 or 3 octets) |---L3 VNI
+--------------------------------------+
```

It should be noticed that when Type 2 routes are used to advertise host IP routes, their **MPLS Label2** field must be set to the corresponding L3 VNI. Without an L3 VNI, Type 2 routes cannot be used as host IP routes for traffic forwarding over VXLAN tunnels, because the VXLAN encapsulation does not contain the VNI.

2.4 Host ARP Entry Advertisement

EVPN Type 2 routes (MAC/IP routes) can carry host MAC addresses and IP addresses simultaneously, which means that the routes can be used to advertise ARP entries of hosts.

```
+--------------------------------------+
| RD (8 octets) |---RD of an EVPN instance
+--------------------------------------+
|Ethernet Segment Identifier (10 octets)|
+--------------------------------------+
| Ethernet Tag ID (4 octets) |
+--------------------------------------+
| MAC Address Length (1 octet) |
+--------------------------------------+
| MAC Address (6 octets) |---Host MAC address
+--------------------------------------+
| IP Address Length (1 octet) |
+--------------------------------------+
| IP Address (0, 4, or 16 octets) |---Host IP address
+--------------------------------------+
| MPLS Label1 (3 octets) |---L2 VNI
+--------------------------------------+
| MPLS Label2 (0 or 3 octets) |
+--------------------------------------+
```

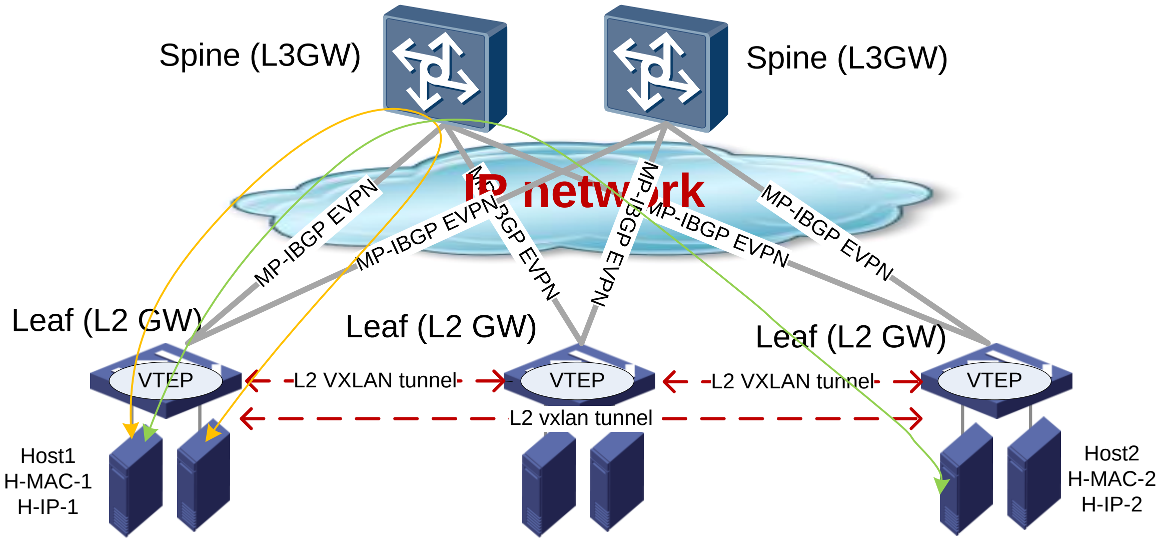
After a switch learns ARP entries of downstream hosts, it uses the Type 2 routes to advertise the ARP entries to its neighbors. The neighbors then cache the ARP entries so that they can convert broadcast ARP packets into unicast packets to reduce broadcast traffic on the network.

3 Solution

3.1 Centralized Gateway Deployment

As shown in Figure 3-1, leaf switches act as L2 gateways on the VXLAN network, and the L3 gateway is centrally deployed on the spine switches. The centralized L3 gateway implements inter-subnet communication between hosts and L3 traffic forwarding. The spine and leaf switches run the Multiprotocol EVPN (MP-EVPN) protocol and establish EVPN neighbor relationships.

Figure 3-1 Centralized gateway deployment



3.1.1 Intra-subnet L2 Communication

On a VXLAN network, communication between hosts on the same subnet is implemented following the L2 known unicast traffic forwarding process.

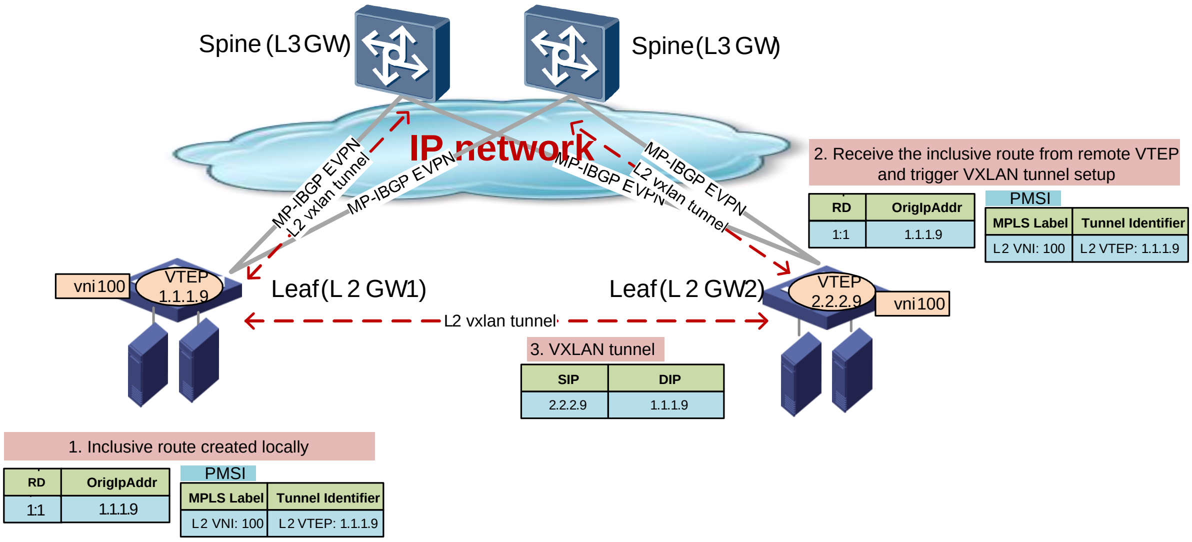
L2 gateways use EVPN to advertise host MAC addresses. After learning MAC addresses from each other, the L2 gateways can forward traffic of hosts on the same subnet as L2 known unicast traffic.

Control Plane

The workflow on the control plane includes two stages:

- Dynamic L2 VXLAN tunnel setup

Figure 3-2 Process of dynamic L2 VXLAN tunnel setup



2. Receive the inclusive route from remote VTEP and trigger VXLAN tunnel setup

RD	OrigIpAddr
1:1	1.1.1.9

PMSI

MPLS Label	Tunnel Identifier
L2 VNI: 100	L2 VTEP: 1.1.1.9

3. VXLAN tunnel

SIP	DIP
2.2.2.9	1.1.1.9

1. Inclusive route created locally

RD	OrigIpAddr
1:1	1.1.1.9

PMSI

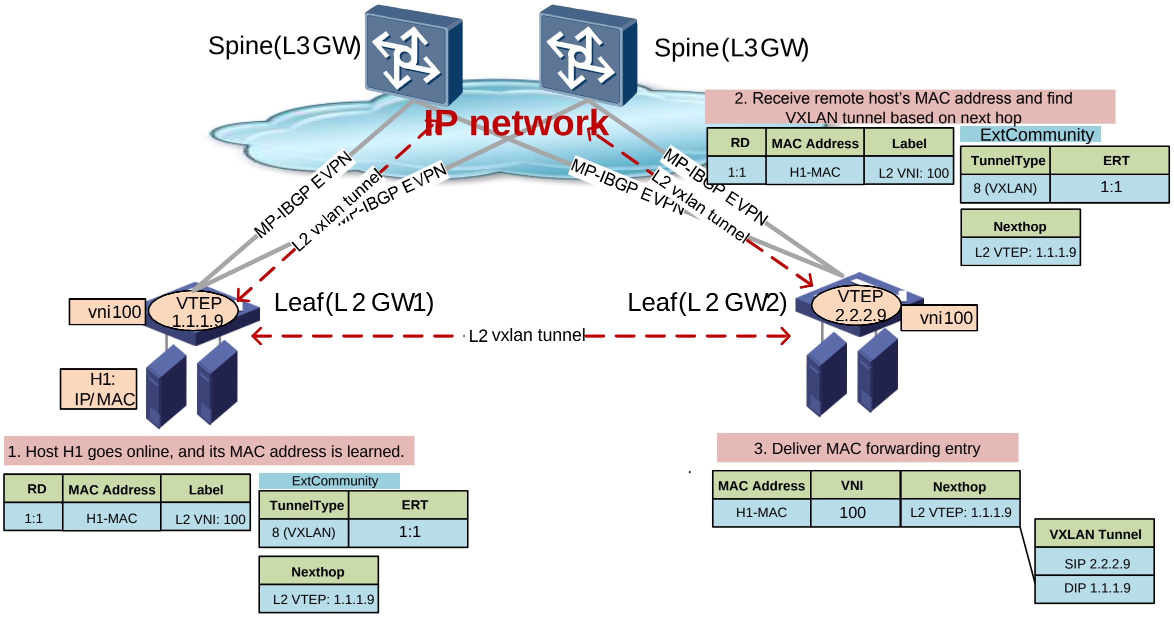
MPLS Label	Tunnel Identifier
L2 VNI: 100	L2 VTEP: 1.1.1.9

EVPN is used as the VXLAN control plane and advertises L2 VNIs and VTEP addresses using inclusive routes, enabling L2 VXLAN tunnels to be set up dynamically. A tunnel setup process as follows:

1. After the bridge domain (BD), L2 VNI, and VTEP address are configured on the local VTEP, the local VTEP advertise an EVPN inclusive route carrying these parameters to the remote VTEP.
2. When the remote VTEP receives the inclusive route, it finds that the BD and L2 VNI in the route are the same as those configured locally. Therefore, this VTEP sets up a VXLAN tunnel with the remote VTEP.

- Host MAC route learning through EVPN

Figure 3-3 Process of host MAC address learning through EVPN



1. Host H1 goes online, and its MAC address is learned.

RD	MAC Address	Label
1:1	H1-MAC	L2 VNI: 100

ExtCommunity

TunnelType	ERT
8 (VXLAN)	1:1

Nexthop
L2 VTEP: 1.1.1.9

2. Receive remote host's MAC address and find VXLAN tunnel based on next hop

RD	MAC Address	Label
1:1	H1-MAC	L2 VNI: 100

ExtCommunity

TunnelType	ERT
8 (VXLAN)	1:1

Nexthop
L2 VTEP: 1.1.1.9

3. Deliver MAC forwarding entry

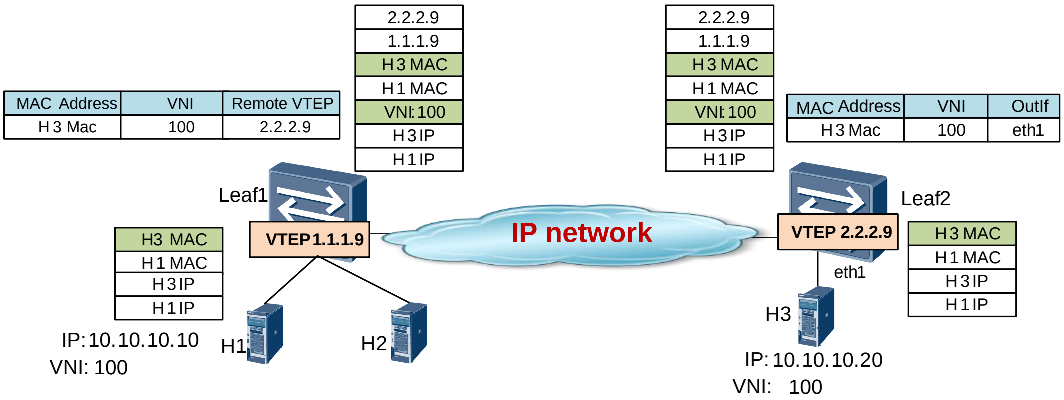
MAC Address	VNI	Nexthop
H1-MAC	100	L2 VTEP: 1.1.1.9

VXLAN Tunnel
SIP 2.2.2.9
DIP 1.1.1.9

1. After the local VTEP learns a local host's MAC address, the VTEP creates an EVPN Type 2 route based on the MAC address, and then advertises the route to remote VTEP over the EVPN control plane.
2. After receiving the MAC route, the remote VTEP delivers the route to the corresponding EVPN instance based on the export route target (ERT) of the route. The VTEP then finds the matching VXLAN tunnel based on the next hop of the route. If the tunnel is reachable, the VTEP delivers the MAC forwarding entry.

Forwarding Plane

Figure 3-4 L2 known unicast traffic forwarding process within a subnet



MAC Address	VNI	Remote VTEP
H3 Mac	100	2.2.2.9

MAC Address	VNI	OutIf
H3 Mac	100	eth1

1. When Leaf 1 receives a packet from host H1, it obtains the L2 bridge domain of the host based on the inbound interface and VLAN ID, and determines whether the destination MAC address is a known unicast MAC address. If the destination MAC address is a known unicast MAC address, Leaf 1 forwards the packet following the known unicast

traffic forwarding process. If not, Leaf 1 follows the broadcast, unknown unicast, and multicast (BUM) traffic forwarding process.

2. To forward the known unicast packet, the VTEP on Leaf 1 encapsulates the packet into a VXLAN packet based on the VNI and remote VTEP address found in the local forwarding table. The VXLAN packet is then forwarded over the IP network between Leaf 1 and Leaf 2.
3. After the VTEP on Leaf 2 receives the VXLAN packet, it checks the UDP destination port number, source and destination IP addresses, and VNI of the packet to determine the packet validity. The VTEP obtains the L2 bridge domain based on the VNI, and then decapsulates the VXLAN packet to obtain the original L2 packet.
4. Leaf 2 looks up the outbound interface and encapsulation information in its MAC address table based on the destination MAC address of the L2 packet, adds a VLAN tag to the packet, and then forwards the packet to host H3.

3.1.2 Inter-subnet L3 Communication

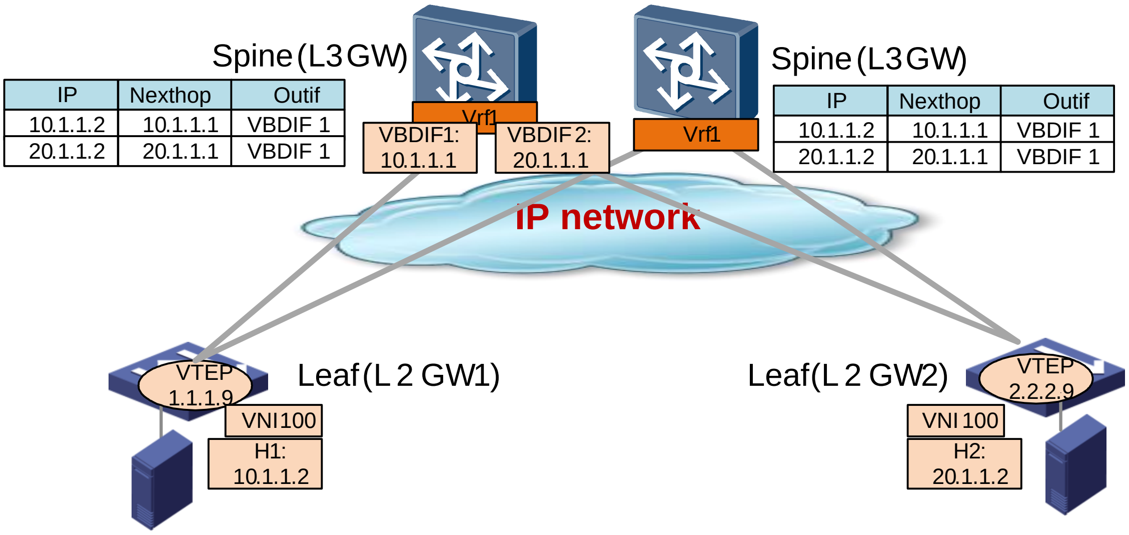
On a VXLAN network, hosts on different subnets cannot learn MAC addresses of each other, so traffic needs to be forwarded between them at L3.

In centralized gateway deployment, the L3 gateway node is responsible for route lookup. Therefore, L3 routes do not need to be advertised on the control plane.

After a host goes online, the L3 gateway learns the ARP entry of the host and creates an IP route for the host.

Figure 3-5 shows the known unicast traffic forwarding process. When the L3 gateway receives ARP entries of the hosts that have gone online, it can learn the IP routes of the hosts H1 and H2.

Figure 3-5 Centralized gateway: L3 known unicast traffic forwarding across subnets



IP	Nexthop	Outif
10.1.1.2	10.1.1.1	VBDIF 1
20.1.1.2	20.1.1.1	VBDIF 1

IP	Nexthop	Outif
10.1.1.2	10.1.1.1	VBDIF 1
20.1.1.2	20.1.1.1	VBDIF 1

The forwarding process is as follows:

1. After receiving a VXLAN packet, the L3 gateway decapsulates the packet to obtain the original L2 packet. Because the destination MAC address of the L2 packet is the MAC

address of the L3 gateway, the L3 gateway starts the L3 forwarding process and looks up a route to forward the packet based on its destination IP address.

2. The L3 gateway obtains the next-hop IP address from the route found in the routing table, and then looks up the ARP entry of this IP address. If the outbound interface in ARP entry is a VXLAN tunnel, the L3 gateway encapsulates the packet based on the VXLAN tunnel information in the ARP entry, and then sends the packet out.

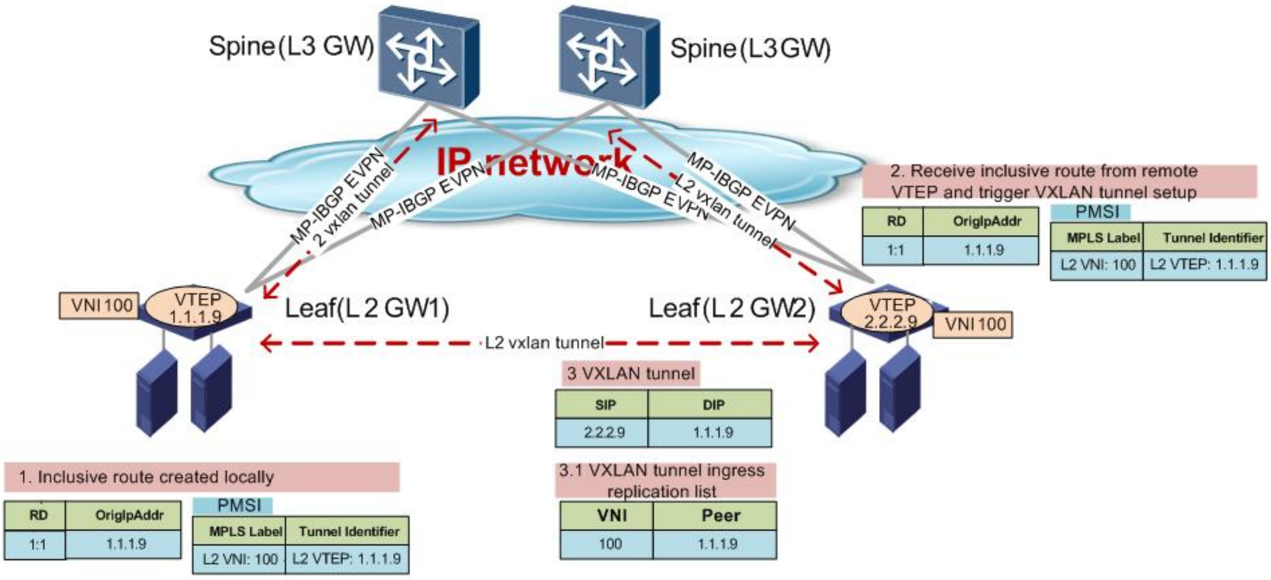
3.1.3 Intra-subnet L2 BUM Traffic Forwarding

On a VXLAN network, BUM packets are forwarded based on the ingress replication list of a VXLAN tunnel.

When VTEPs dynamically discover a VXLAN tunnel based on the L2 VNI and L2 VTEP advertised through an EVPN Type 3 route (inclusive route), they also create an ingress replication list for the VXLAN tunnel.

Control Plane

Figure 3-6 Control plane working mechanism for intra-subnet BUM traffic forwarding



2. Receive inclusive route from remote VTEP and trigger VXLAN tunnel setup

RD	OrigIpAddr
1:1	1.1.1.9

PMSI

MPLS Label	Tunnel Identifier
L2 VNI: 100	L2 VTEP: 1.1.1.9

3 VXLAN tunnel

SIP	DIP
2.2.2.9	1.1.1.9

3.1 VXLAN tunnel ingress replication list

VNI	Peer
100	1.1.1.9

1. Inclusive route created locally

RD	OrigIpAddr
1:1	1.1.1.9

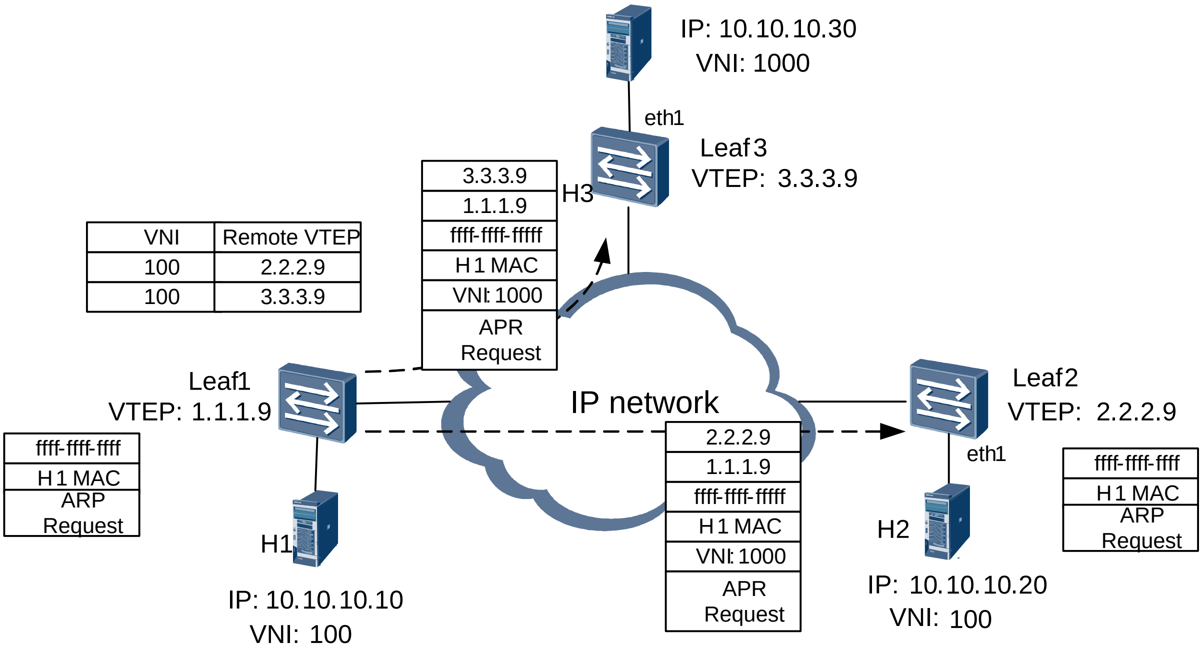
PMSI

MPLS Label	Tunnel Identifier
L2 VNI: 100	L2 VTEP: 1.1.1.9

When a VTEP receives an inclusive route from a remote VTEP and finds that the BD and L2 VNI in the route are the same as the locally configured ones, the VTEP establishes a VXLAN tunnel to the remote VTEP. In addition, it creates an ingress replication list for this VXLAN tunnel.

Forwarding Plane

Figure 3-7 BUM traffic forwarding process within a subnet

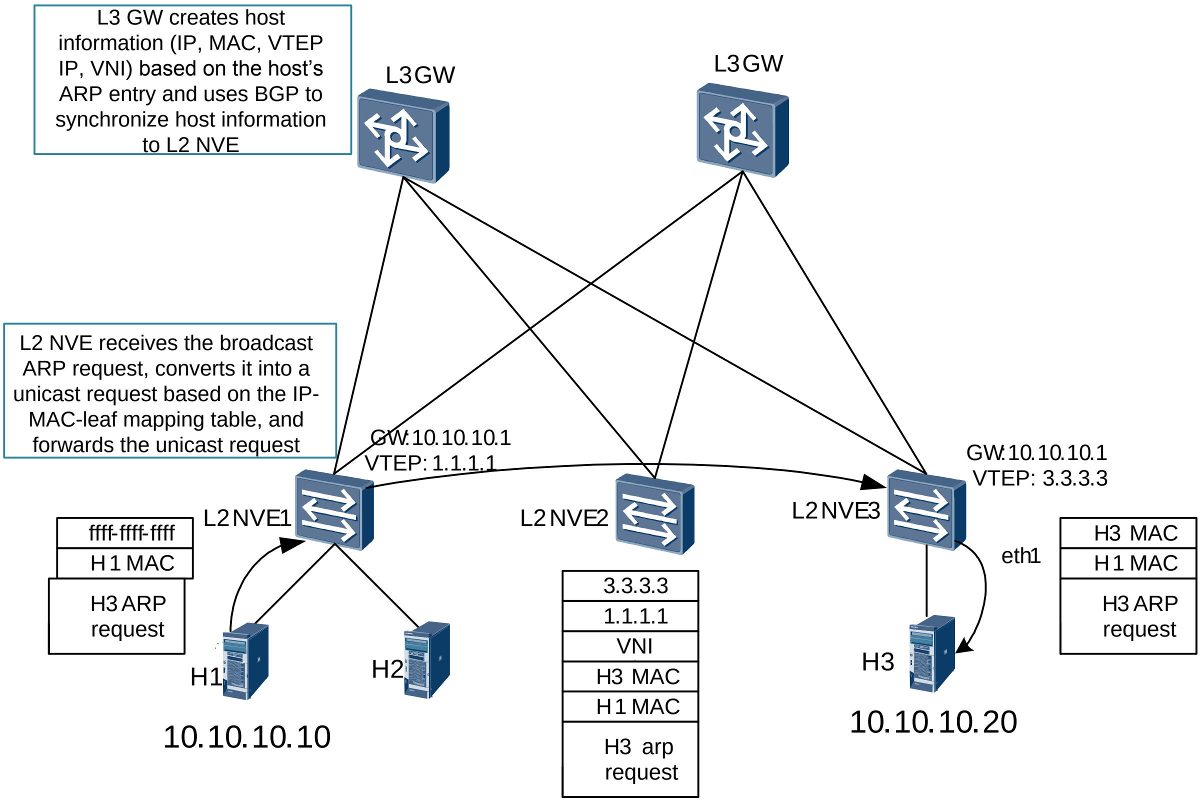


VNI	Remote VTEP
100	2.2.2.9
100	3.3.3.9

1. When Leaf 1 receives a packet from host H1, it obtains the L2 bridge domain of the host based on the inbound interface and VLAN ID, and determines whether the destination MAC address is a BUM MAC address. If the destination MAC address is a BUM MAC address, Leaf 1 forwards the packet following the BUM traffic forwarding process. If not, Leaf 1 follows the known unicast traffic forwarding process.
2. If Leaf 1 determines that the packet is a BUM packet, the VTEP on Leaf 1 obtains the ingress replication list for the VNI based on the Layer 2 bridge domain. The VTEP then replicates the packet based on the list, encapsulates the packet into a VXLAN packet, and forwards the packet through the outbound interface.
3. After the VTEP on Leaf 2 or Leaf 3 receives the VXLAN packet, it checks the UDP destination port number, source and destination IP addresses, and VNI of the packet to determine the packet validity. The VTEP obtains the L2 bridge domain based on the VNI, and then decapsulates the VXLAN packet to obtain the original L2 packet.
4. Leaf 2 or Leaf 3 forwards the original L2 packet based on its MAC address table. If the destination MAC address of the packet matches a MAC address entry, the leaf switch encapsulates and forwards the packet through the outbound interface in the MAC address entry. If the destination MAC address matches no MAC address entry, the leaf switch broadcasts the packet to user-side interfaces.
5. BUM packets received from a VXLAN tunnel can only be broadcast to user-side interfaces and cannot be broadcast to the VXLAN tunnel, as this will cause a loop.

3.1.4 ARP Broadcast-to-Unicast Conversion

Figure 3-8 Centralized gateway deployment: ARP broadcast suppression



L3 GW creates host information (IP, MAC, VTEP IP, VNI) based on the host's ARP entry and uses BGP to synchronize host information to L2 NVE

L2 NVE receives the broadcast ARP request, converts it into a unicast request based on the IP-MAC-leaf mapping table, and forwards the unicast request

On the figure:

- An L3GW stands for a VXLAN L3 gateway, which is a spine node.
- An L2NVE stands for a VXLAN L2 gateway, which is a leaf node.
- H1 is a virtual machine or server connected to a VXLAN L2 gateway.

The ARP broadcast-to-unicast conversion function can reduce ARP broadcast traffic on L2NVEs, thereby improving network performance.

The L3GW learns ARP entries of hosts connected to its bridge domain interface (BDIF) and creates host information (IP address, MAC address, VTEP address, and VNI) based on the ARP entries. The ARP protocol advertises host information to BGP, which then synchronizes the information to other L2NVEs, so that all BGP neighbors can learn the host information. On an L2NVE, BGP delivers host information to ARP, so that ARP broadcast packets can be converted to unicast packets for broadcast traffic suppression. (In distributed gateway deployment, L3GW and L2NVE are deployed on the same switch.)

The ARP broadcast-to-unicast conversion process is as follows:

1. H1 sends a broadcast ARP request packet to H2 or H3 in the subnet.
2. Leaf node L2NVE1 connected to H1 matches the received ARP request packet with the local IP-MAC-leaf mapping table. If the packet matches the table, L2NVE1 replaces the broadcast destination MAC address in the ARP request with the unicast MAC address in the mapping table, and the sends the unicast packet.

3. When the destination host receives the unicast ARP request packet, it sends an ARP reply packet.
4. When receiving the ARP reply packet, H1 caches the ARP entry and can communicate with the destination host.

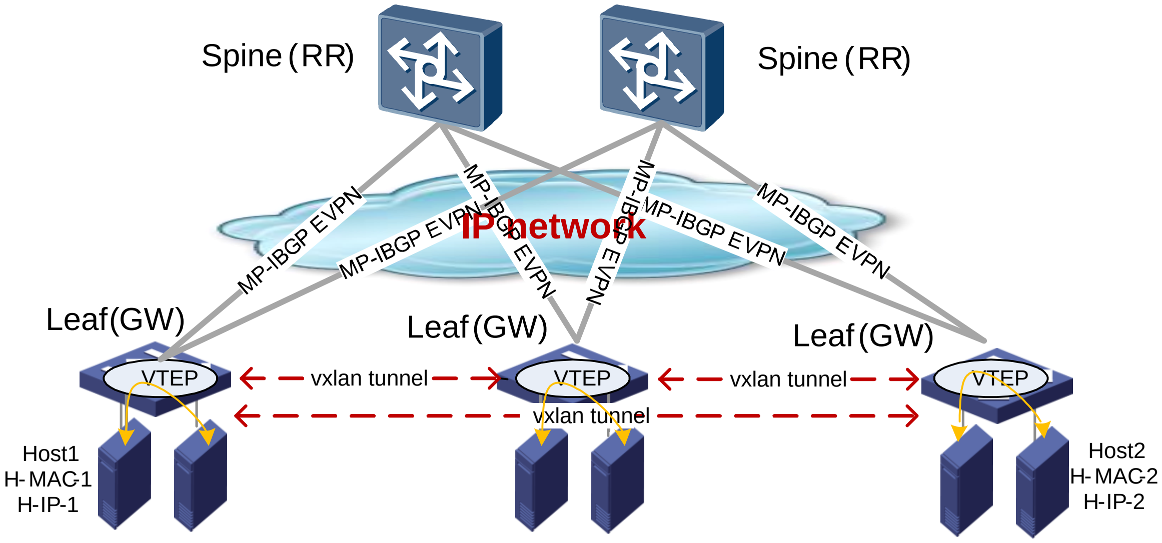
3.2 Distributed Gateway Deployment

Figure 3-9 shows a distributed gateway deployment scenario, where each leaf acts as both an L2 gateway and an L3 gateway.

Compared to centralized gateway deployment, distributed gateway deployment enables inter-subnet L3 traffic to be forwarded by nearby gateways. Additionally, host information is distributed on local leaf nodes. Therefore, distributed gateway deployment avoids the capacity bottleneck on a centralized gateway and provides better scalability of gateways.

Host MAC addresses and IP addresses are advertised through the EVPN control plane to enable communication between hosts under different leaf nodes.

Figure 3-9 Distributed gateway deployment



3.2.1 Intra-subnet L2 Communication

For communication between hosts under different leaf nodes, the working processes on the control plane and forwarding plane are the same as those in centralized gateway deployment. For details, see section 3.1.1 Intra-subnet L2 Communication.

3.2.2 Inter-subnet L3 Communication

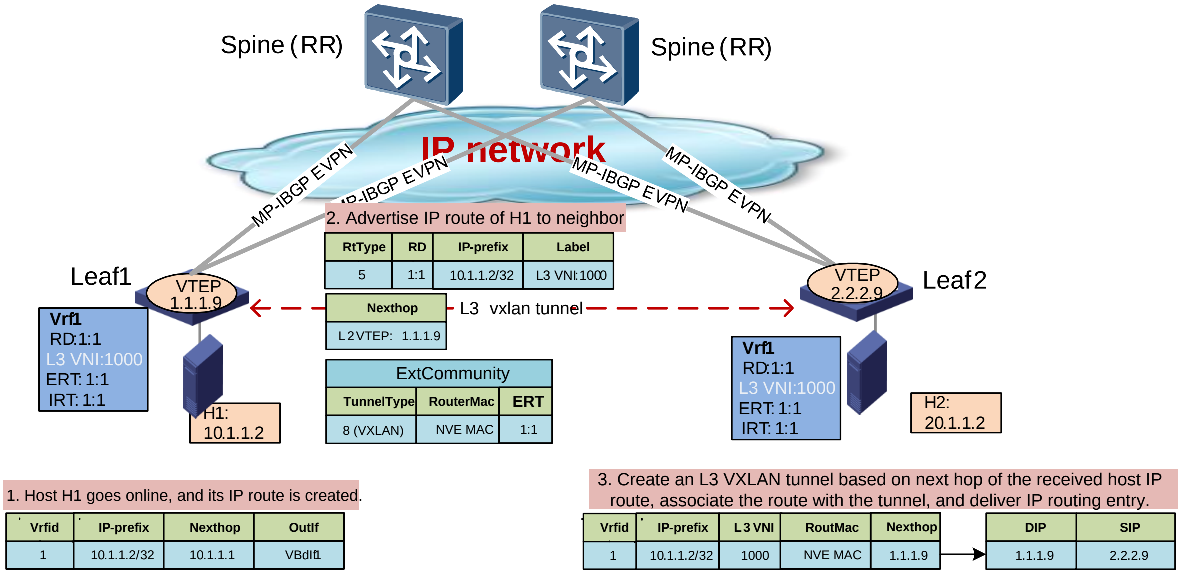
On a VXLAN network, hosts on different subnets cannot learn MAC addresses of each other, so traffic needs to be forwarded between them at L3.

In distributed gateway deployment, leaf nodes need to advertise host routes to each other before hosts on different subnets can communicate with each other. The EVPN control plane is used to advertise host routes.

Control Plane

- Host IP route learning through EVPN

Figure 3-10 Process of host IP route learning through EVPN



2. Advertise IP route of H1 to neighbor

RtType	RD	IP-prefix	Label
5	1:1	10.1.1.2/32	L3 VNI:1000

Nexthop
L2VTEP: 1.1.1.9

ExtCommunity		
TunnelType	RouterMac	ERT
8 (VXLAN)	NVE MAC	1:1

1. Host H1 goes online, and its IP route is created.

Vrfid	IP-prefix	Nexthop	OutIf
1	10.1.1.2/32	10.1.1.1	VBdIf1

3. Create an L3 VXLAN tunnel based on next hop of the received host IP route, associate the route with the tunnel, and deliver IP routing entry.

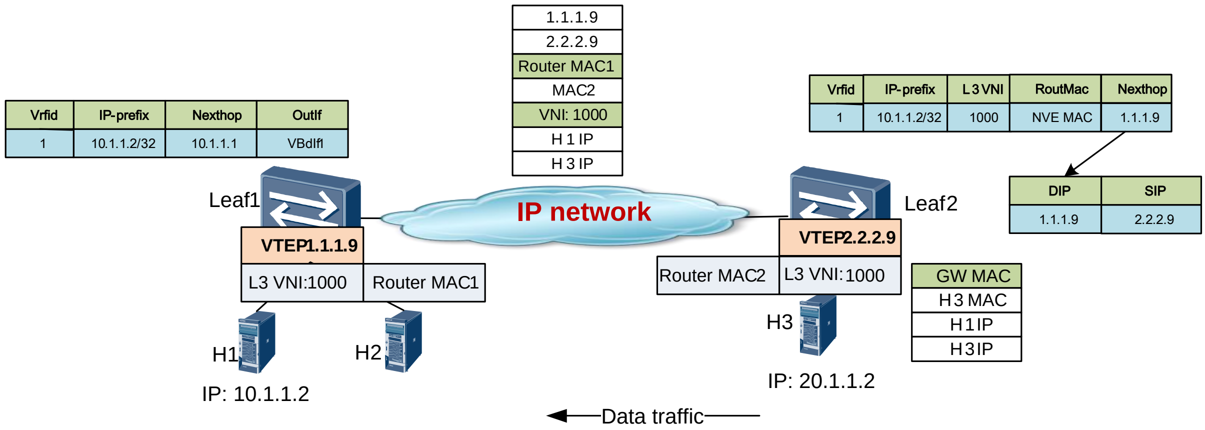
Vrfid	IP-prefix	L3 VNI	RoutMac	Nexthop
1	10.1.1.2/32	1000	NVE MAC	1.1.1.9

DIP	SIP
1.1.1.9	2.2.2.9

1. After a host goes online, the leaf node learns a host IP route and advertises this route to its neighbor using EVPN. When advertising the host IP route, the leaf node adds the L3 VNI of the tenant to the route advertisement packet and sets the next-hop address to the local VTEP address.
2. When the remote leaf node learns the host IP route, it delivers the route to the corresponding VPN instance based on the ERT. Then the remote leaf node creates a dynamic L3 VXLAN tunnel, associates the host IP route with the tunnel, and delivers the IP routing entry.

Forwarding Plane

Figure 3-11 Distributed gateway deployment: inter-subnet L3 forwarding process



Vrfid	IP-prefix	Nexthop	OutIf
1	10.1.1.2/32	10.1.1.1	VBdIf1

Vrfid	IP-prefix	L3 VNI	RoutMac	Nexthop
1	10.1.1.2/32	1000	NVE MAC	1.1.1.9

DIP	SIP
1.1.1.9	2.2.2.9

When H3 communicates with H1, the traffic needs to be forwarded across subnets at L3. The forwarding process is as follows:

1. When Leaf 2 receives a packet from H3, it checks whether the destination MAC address of the packet is the gateway MAC address (its own leaf MAC address). If so, Leaf 2 triggers the L3 forwarding process. First, Leaf 2 finds the VPN instance bound to the inbound interface of the packet.
2. It then looks up the destination IP address of the packet in the routing table of the VPN instance, encapsulates the packet with a VXLAN header carrying the L3 VNI and router MAC address, and forwards the packet to Leaf 1.
3. When receiving the VXLAN packet, Leaf 1 finds the bridge domain based on VNI in the packet and determines whether the destination MAC address is its own MAC address. If the destination MAC address is the system MAC address of its own, Leaf 1 starts the L3 forwarding process. In this case, Leaf 1 finds the VPN instance based on the VNI and looks up the matching route in the routing table of the VPIN route. As Leaf 1 finds that the next hop of the route is the gateway interface, it encapsulates the packet and sends it to H1.

3.2.3 Intra-subnet L2 BUM Traffic Forwarding

In distributed gateway deployment, the working processes on the control plane and forwarding plane for L2 BUM traffic forwarding are the same as those in centralized gateway deployment. For details, see section 3.1.3 Intra-subnet L2 BUM Traffic Forwarding.

3.2.4 ARP Broadcast-to-Unicast Conversion

In distributed gateway deployment, the ARP broadcast-to-unicast conversion process is the same as that in centralized gateway deployment. For details, see section 3.1.4 ARP Broadcast-to-Unicast Conversion.

3.2.5 Virtual Machine Migration

In centralized gateway deployment, when a virtual machine (VM) migrates from one leaf node to another, external routers and other leaf nodes need to update their routing tables to direct traffic of the VM to the destination leaf node. Other leaf nodes also need to update their host information tables to ensure that broadcast ARP request packets of the VM can be converted to unicast packets correctly. After the migration, the VM sends gratuitous ARP packets to trigger updates of its host route and ARP information on the external routes and leaf nodes.

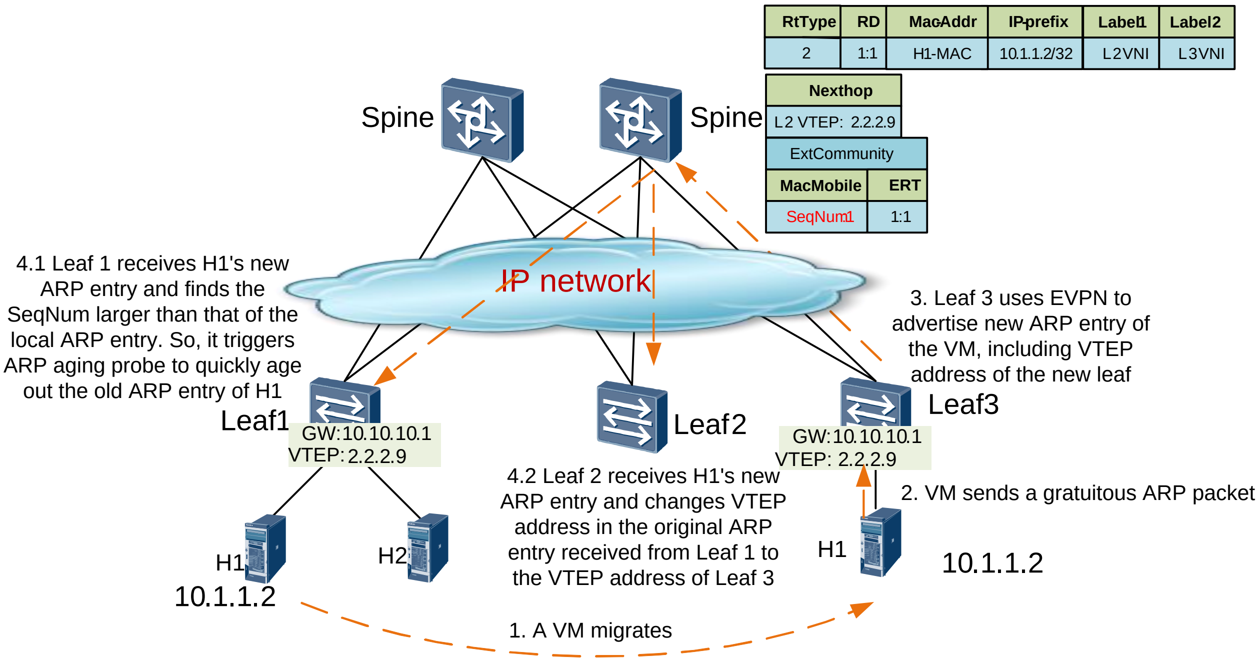
After a VM migrates, its ARP entry needs to be advertised through the EVPN control plane. The migration attribute of the host is advertised using the MAC Mobility community attribute of EVPN.

```
MAC Mobility

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| Type=0x06 | Sub-Type=0x00 |Flags (1 octet)| Reserved=0 |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| Sequence Number|

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

After a VM migration, the Sequence Number field of the MAC Mobility attribute increases by 1.

Figure 3-12 shows the VM migration process.

Figure 3-12 Virtual machine migration



RtType	RD	MacAddr	IP-prefix	Label1	Label2
2	1:1	H1-MAC	10.1.1.2/32	L2VNI	L3VNI

Nexthop
L2 VTEP: 2.2.2.9

ExtCommunity	
MacMobile	ERT
SeqNum1	1:1

1. A VM migrates

2. VM sends a gratuitous ARP packet

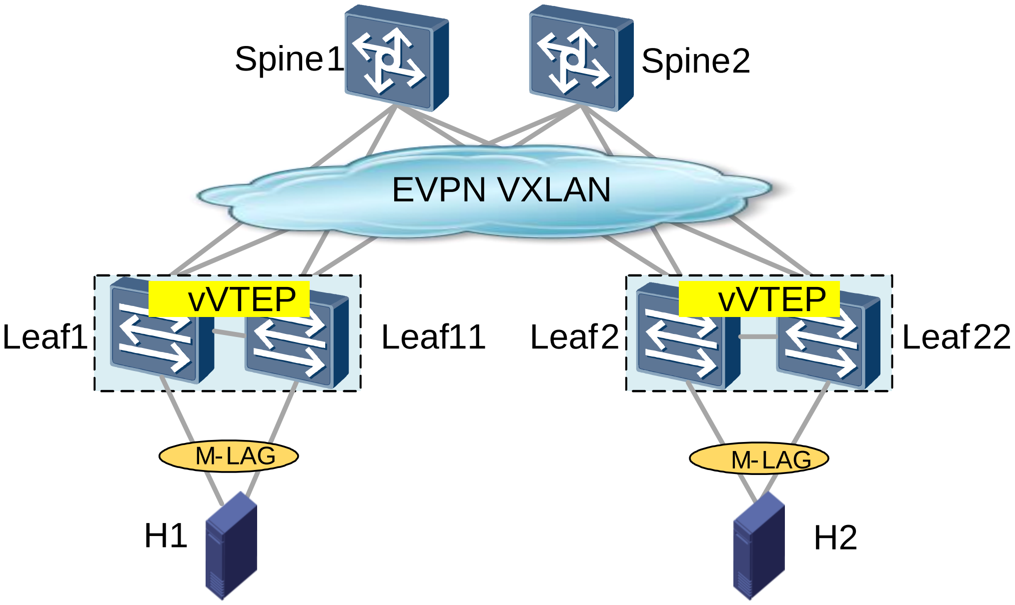
3. Leaf 3 uses EVPN to advertise new ARP entry of the VM, including VTEP address of the new leaf

4.1 Leaf 1 receives H1's new ARP entry and finds the SeqNum larger than that of the local ARP entry. So, it triggers ARP aging probe to quickly age out the old ARP of H1

4.2 Leaf 2 receives H1's new ARP entry and changes VTEP address in the original ARP entry received from Leaf 1 to the VTEP address of Leaf 3

3.2.6 Active-Active High Availability of Hosts

Figure 3-13 Active-active high availability of hosts



To ensure high availability of hosts, hosts can be dual homed to leaf nodes through multichassis link aggregation groups (M-LAGs). Two leaf nodes set up a an M-LAG and are configured with the same VTEP IP address (vVTEP) to set up VXLAN tunnels with other devices. An interconnect link is established between the two leaf nodes in an M-LAG system to provide path protection in case of a downlink failure.

Normally, when a leaf node receives packets from an interface connected to a host, it directly forwards the packets to another server or encapsulates the packets in the VXLAN tunnel and forwards them to the remote vVTEP. When receiving packets from the VXLAN tunnel, a leaf node encapsulates the packets and forwards them to the host.

If one of the LAG member interfaces connected to a host fails, traffic received from other hosts or the VXLAN tunnel is forwarded to the other link of the host through the interconnect link between the two leaf nodes.

The two leaf nodes that set up an M-LAG advertise the same VTEP IP address and router MAC address when they use BGP EVPN to advertise local host routes.

3.3 Use of Two BGP Processes

CloudEngine switches can run two BGP processes to decouple the overlay network from the underlay network. This allows for separate configuration on the overlay and underlay networks and facilitates network maintenance.

Two BGP instances can be configured on a switch. One BGP instance provides IP route connectivity on the underlay network, whereas the other BGP instance runs EVPN as the VXLAN control plane on the overlay network.

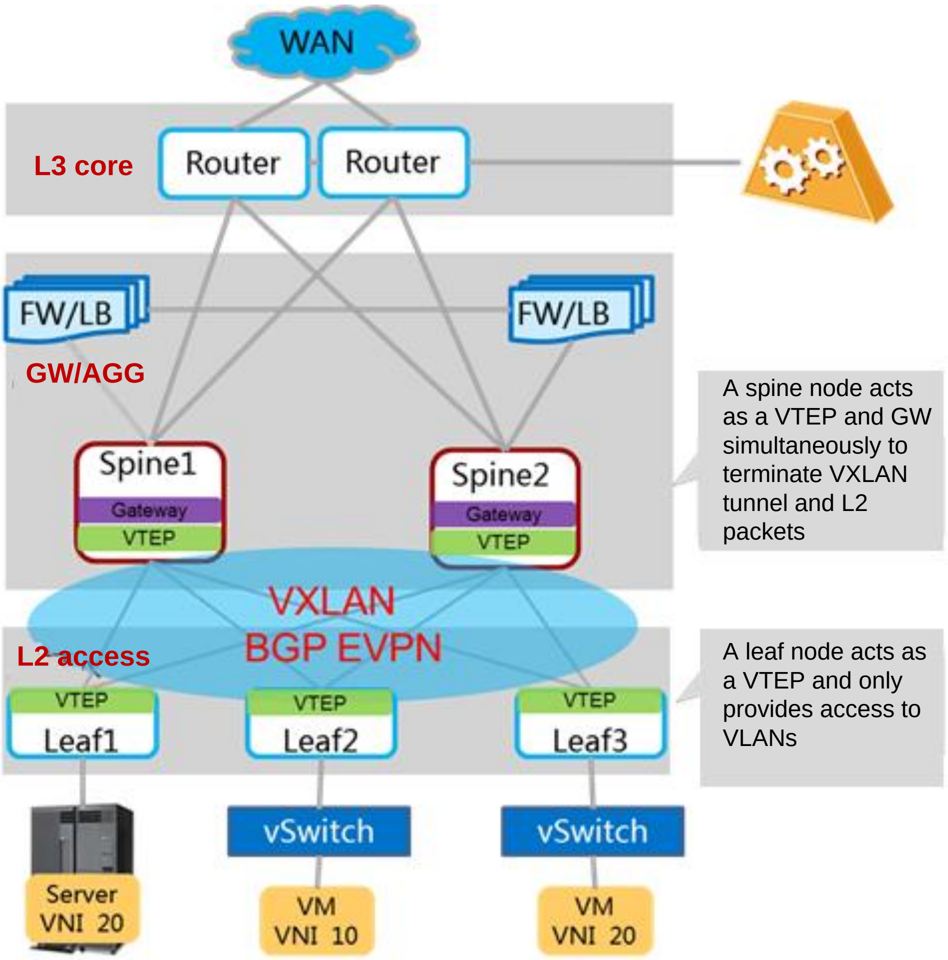
Typical configuration:

```
bgp 200
 #
 ipv4-family unicast
 #
 ipv4-family vpn-instance vrf1
  network 101.0.0.2 255.255.255.255
  import-route direct
  advertise l2vpn evpn
#
bgp 100 instance overlay
 peer 133.0.0.11 as-number 100
 peer 133.0.0.11 connect-interface LoopBack0
 #
 l2vpn-family evpn
  policy vpn-target
  peer 133.0.0.11 enable
 #
 l2vpn-family evpn
  policy vpn-target
  peer 133.0.0.11 enable
```

4 Typical Applications

4.1 Centralized Gateway Deployment

Figure 4-1 Centralized gateway deployment



A spine node acts as a VTEP and GW simultaneously to terminate VXLAN tunnel and L2 packets

A leaf node acts as a VTEP and only provides access to VLANs

1. The leaf and spine nodes set up full-mesh connections. At least two spine nodes need to be deployed on the network, which back up each other and implement load balancing.

Forwarding paths are established between spine and leaf nodes through L3 routing. Normally, ECMP paths exist between the spine and leaf nodes.

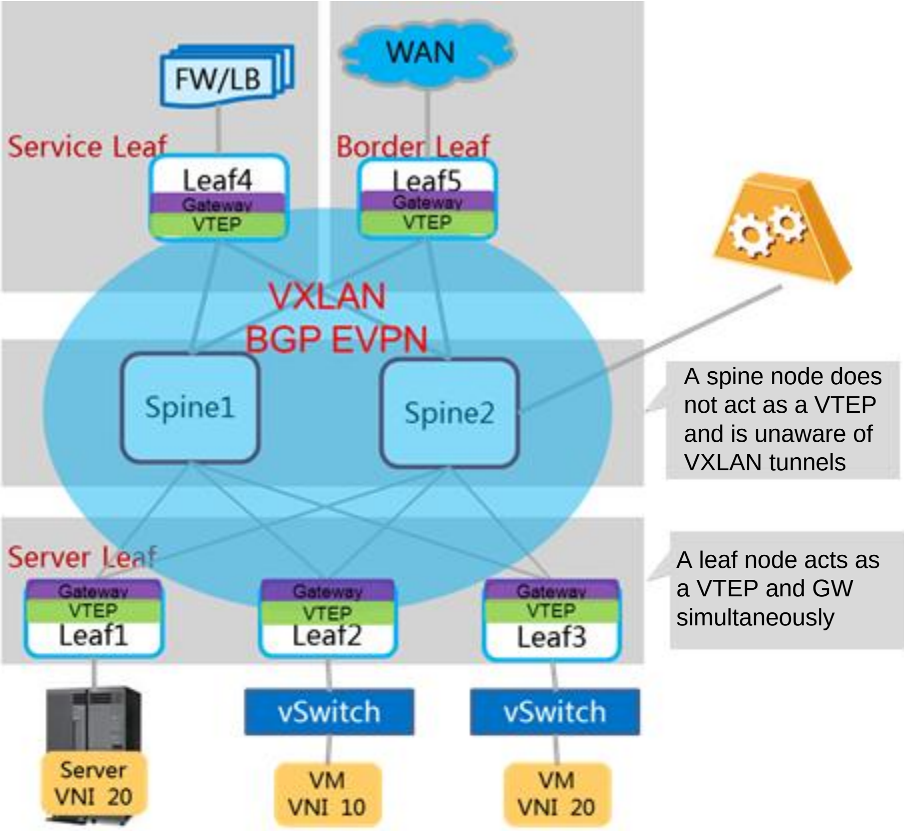
2. The spine and leaf nodes all act as VTEPs.
3. A leaf node only completes L2 traffic forwarding. It maps the VLAN tag of user packets to a VNI, encapsulates the packets in VXLAN packets, and forwards them to the destination node in the local bridge domain.
4. A spine node acts as a centralized L3 gateway to decapsulate VXLAN packets and forward them based on L3 routes, enabling communication between bridge domains on the network and communication with external networks.

BGP EVPN acts as the VXLAN control plane to provide the following function:

- Triggers automatic VXLAN tunnel setup between VTEPs to avoid the need to manually configure full-mesh tunnels.
- Advertises MAC address tables to prevent flooding of unknown traffic.

4.2 Distributed Gateway Deployment

Figure 4-2 Distributed gateway deployment



1. The leaf and spine nodes set up full-mesh connections. At least two spine nodes need to be deployed on the network, which back up each other and implement load balancing.

Forwarding paths are established between spine and leaf nodes through IGP. Normally, ECMP paths exist between the spine and leaf nodes.

2. Leaf nodes act as VTEPs, whereas spine nodes are not necessarily VTEPs.
3. Each leaf node acts as an L3 gateway to terminate L3 traffic for local users.
4. Different leaf nodes provide different functions. For example, some leaf nodes connect to external networks, and some leaf nodes connect to firewalls for traffic filtering.

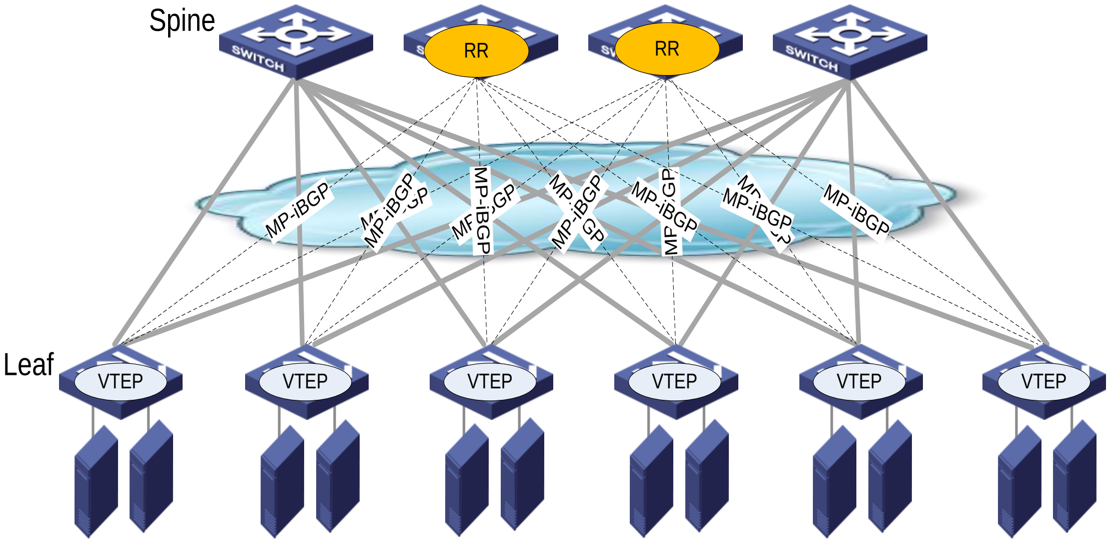
BGP EVPN acts as the VXLAN control plane to provide the following function:

- Triggers automatic VXLAN tunnel setup between VTEPs to avoid the need to manually configure full-mesh tunnels.
- Advertises host routes and MAC addresses to guide traffic forwarding.

4.3 BGP Deployment on Distributed Gateways

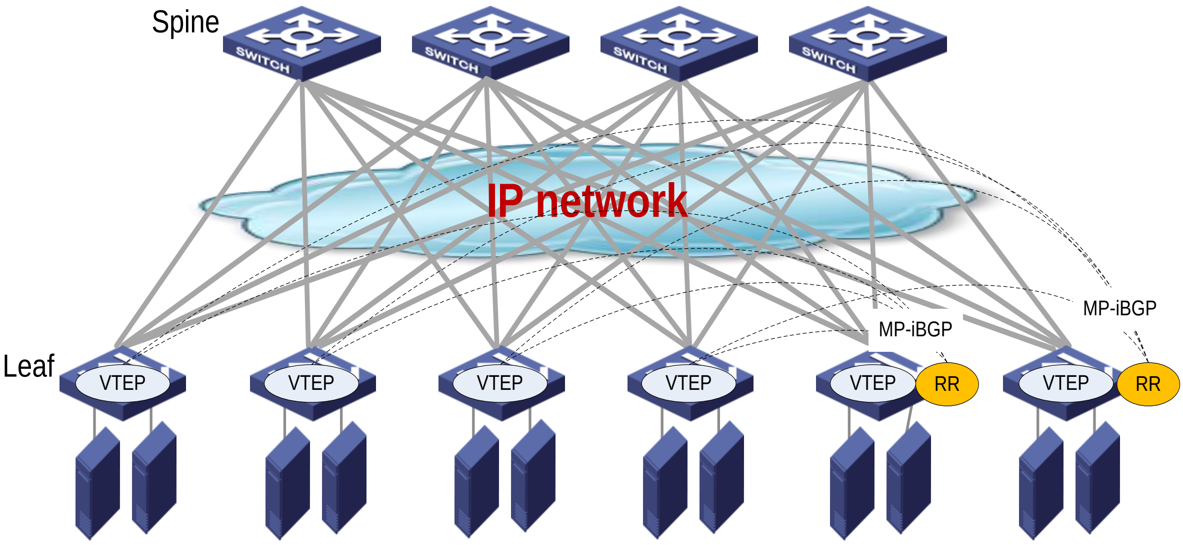
If VTEPs establish IBGP peer relationships, route reflectors (RR) can be configured on the network to simplify IBGP peer configuration.

Figure 4-3 RR deployment on distributed gateways



In Figure 4-3, some spine nodes act as RRs. The spine nodes must support EVPN. If spine nodes do not support EVPN, RRs can be deployed on VTEPs, as shown in Figure 4-4.

Figure 4-4 RR deployment on distributed gateway leaf nodes



If the underlay network runs EBGP, EGBP peers need also be established on the EVPN VXLAN network, as shown in Figure 4-5 and Figure 4-6.

Figure 4-5 EBGP deployment on distributed gateways (leaf nodes in different ASs)

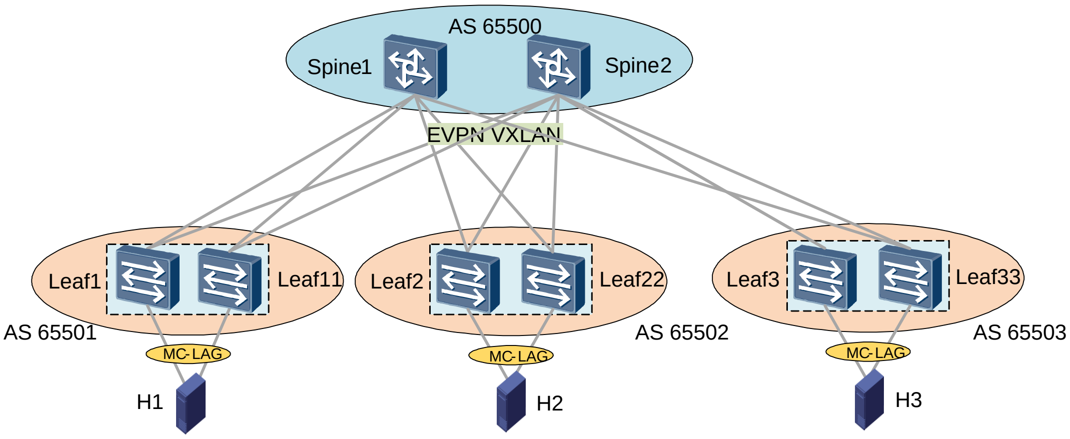
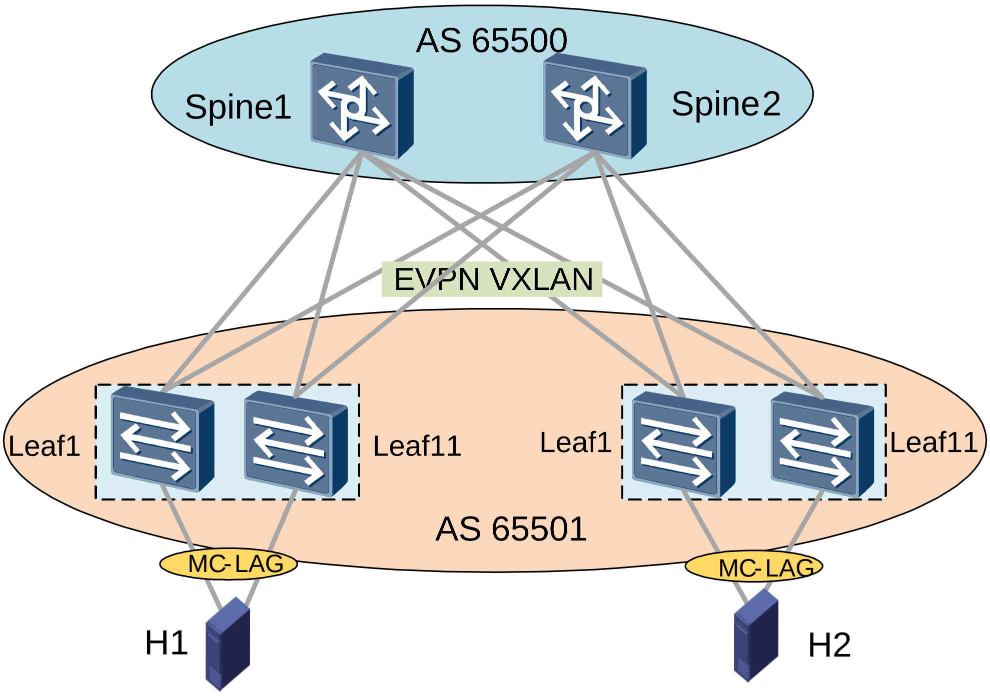


Figure 4-6 EBGP deployment on distributed gateways (leaf nodes in the same AS)



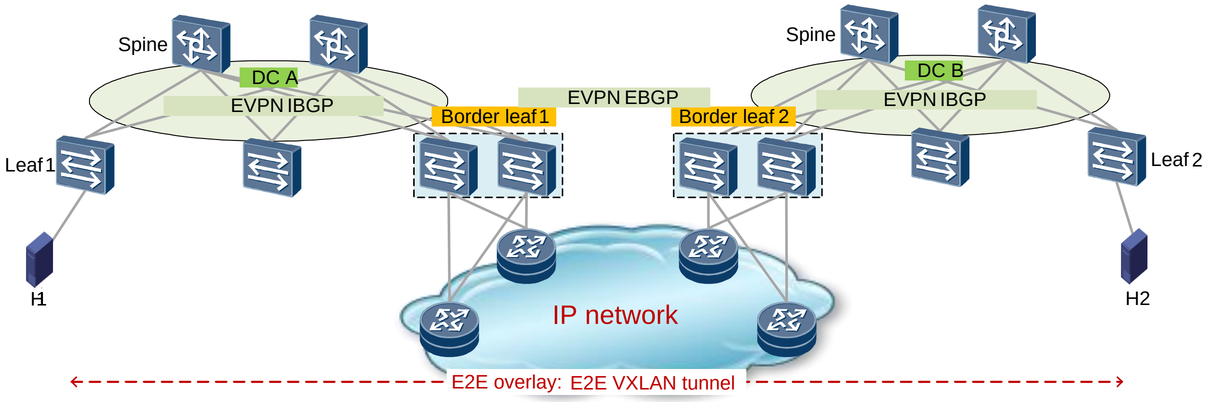
1. The spine nodes are configured not to change the next hops of EBGP routes.
2. If all leaf nodes are in the same AS, the leaf nodes need to be configured to ignore EBGP AS-path loop detection on inbound interfaces, so that they can learn host routes from other leaf nodes.

If the preceding configurations are not performed, each leaf node will receive a route advertised by itself. Because this route has a longer AS-path, it will not affect route selection on the leaf node. However, the redundant route wastes the capacity of the routing table.

4.4 DCI

1. End-to-end VXLAN tunnel mode: EVPN peers between DCI border leaf nodes

Figure 4-7 End-to-end VXLAN tunnel mode: EVPN peers between DCI border leaf nodes



As shown in Figure 4-7, MP-EBGP EVPN runs between DC A and DC B. MP-EBGP EVPN does not change the next-hop addresses (or change the VNI and router MAC address) when advertising MAC routes, host routes, MAC/IP routes. In this way, VTEPs in different data centers can set up end-to-end VXLAN tunnels.

The intermediate routing devices between the two data centers provide raw IP addresses to ensure IP reachability of the end-to-end VXLAN tunnels.

L2 forwarding and L3 forwarding are both implemented over end-to-end VXLAN tunnels. VTEPs in different data centers communicate directly through the VXLAN tunnels, and the underlay network provides transparent transmission of IP packets to implement DCI. Therefore, the two data centers form one large VXLAN domain logically.

Advantages:

- Multiple data centers from one logical DC to simplify management.

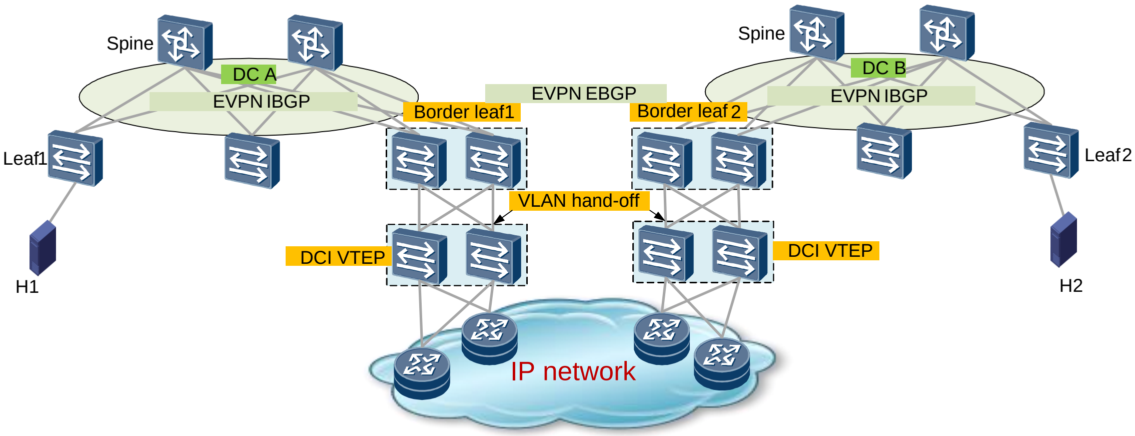
Disadvantages:

- The data centers must use the same protocols and encapsulation mode.

This mode is recommended for small-sized data centers.

2. VLAN hand-off mode

Figure 4-8 End-to-end VXLAN tunnel mode: DCI VLAN hand-off



This mode only allows service traffic of the same subnet to be transmitted across data centers.

As shown in Figure 4-8, VTEPs in a data center and DCI VTEPs are connected through VLANs, and DCI VTEPs set up VXLAN tunnels to implement DCI.

Advantages:

- Different data centers can run different protocols.
- Data centers do not need to use the same encapsulation mode and can have different architectures.
- Data centers are connected by P2P connections logically, which facilitate bandwidth control, policy control, and broadcast storm suppression at ingress.

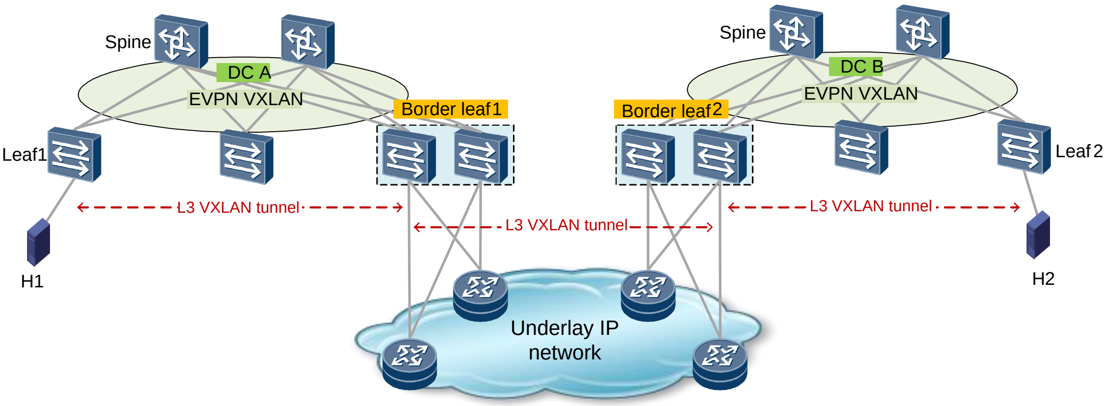
Disadvantages:

- DCI devices must provide high performance (save all entries of all data centers).

This mode is recommended for large-scale, modular data centers.

3. DCI L3 traffic over three-segment VXLAN

Figure 4-9 DCI L3 traffic over three-segment VXLAN

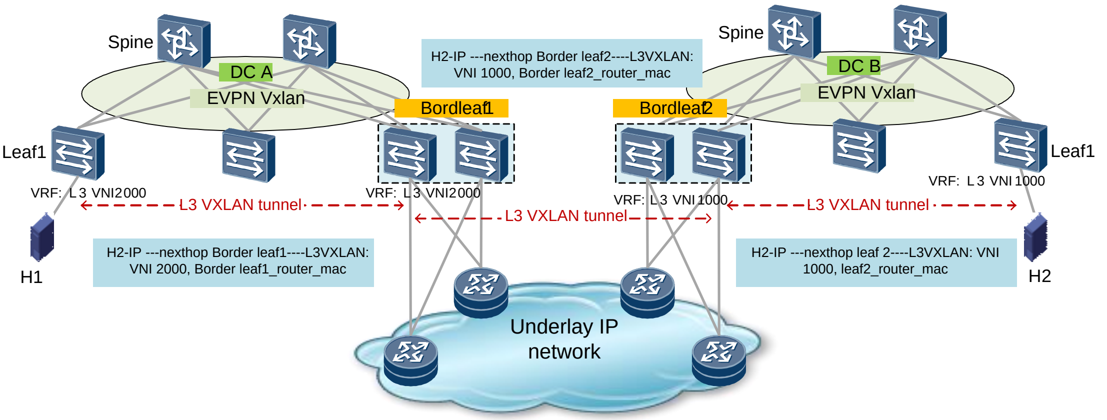


As shown in Figure 4-9, DC A and DC B are in different campuses and need to communicate with each other. To reduce L2 floods and prevent broadcast storms in one data center from spreading to the other data center, L3 DCI is deployed between the two data centers.

The IP network between DC A and DC B functions as an underlay network. Boarder leaf 1 and Border leaf 2 need to establish an L3 VXLAN tunnel.

Therefore, when hosts in different data centers communicate with each other, their traffic is actually forwarded through three VXLAN tunnels.

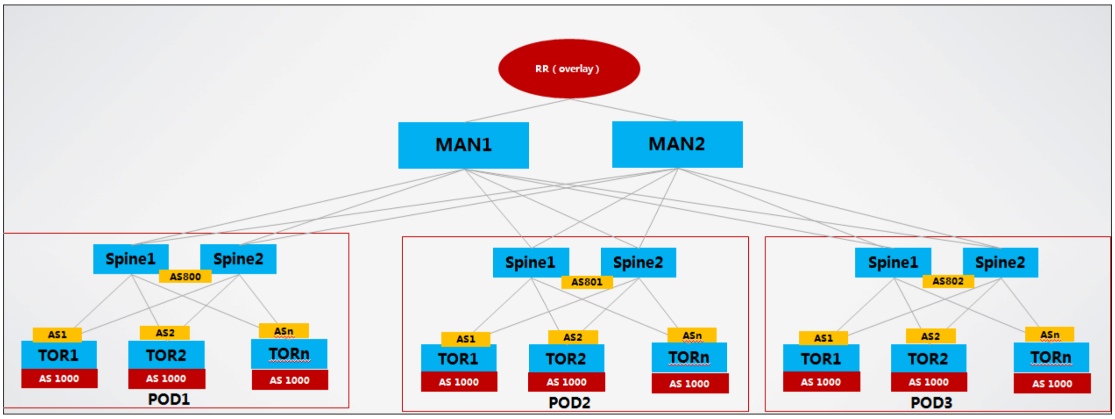
Figure 4-10 Control plane route advertisement process in three-segment VXLAN mode for DCI


Spine
DC A
EVPN Vxlan
Bordleaf1
Leaf1
VRF: L3 VNI2000
L3 VXLAN tunnel
VRF: L3 VNI2000
H1
H2-IP ---nexthop Border leaf1----L3VXLAN: VNI 2000, Border leaf1_router_mac
H2-IP ---nexthop Border leaf2----L3VXLAN: VNI 1000, Border leaf2_router_mac
L3 VXLAN tunnel
Bordleaf2
VRF: L3 VNI 1000
Spine
DC B
EVPN Vxlan
Leaf1
VRF: L3 VNI 1000
L3 VXLAN tunnel
H2-IP ---nexthop leaf 2----L3VXLAN: VNI 1000, leaf2_router_mac
H2
Underlay IP network

4.5 Underlay EBGP + Overlay IBGP Deployment

EVPN supports multiple BGP processes and allows for EBGP deployment on the underlay network and IBGP deployment on the overlay network. This deployment decouples the overlay network from the underlay network and enables the two networks to be configured separately. VXLAN EVPN is deployed on the overlay network, and the underlay network uses the direct EBGP peering deployment. When EGBP is replaced by OSPF on the underlay network, the configuration of the overlay the overlay network does not need to be changed.

Figure 4-11 Deployment of two BGP processes



1. EBGP runs on the underlay network.

Devices in each POD set up direct EBGP peer relationships. Leaf nodes in the same POD have different AS numbers, but leaf nodes in different PODs can have identical AS numbers. Spine nodes in the same POD have the same AS number. Spine nodes in different PODs have different AS numbers and set up direct EBGP peer relationships. The underlay network learns IP routes to enable overlay nodes to set up EVPN peer relationships.

2. IBGP runs on the overlay network.

The overlay network uses EVPN VXLAN L3 gateways, which can be assigned independent AS numbers. Leaf nodes set up IBGP peer relationships and the RRs support EVPN. IBGP EVPN peers advertise Type 2 or Type 5 routes to each other and set up VXLAN tunnels based on the routes.

3. The overlay and underlay networks are decoupled and configured separately. When EBGP is replaced by OSPF on the underlay network, the configuration of the overlay network does not need to be changed.
4. Each leaf node can be configured with two AS numbers, one for underlay EBGP and the other for overlay IBGP.

5 Conclusion

The EVPN solution supported by CloudEngine switches is Huawei's L2/L3 DCI solution. The solution combines the advantages of IETF EVPN on the control plane and VXLAN encapsulation on the forwarding plane. Additionally, the solution integrates innovative technologies that Huawei developed to simplify O&M. With all these advantages, the EVPN solution can meet the requirements and challenges for L2 and L3 DCI.

6 Acronyms and Abbreviations

Abbreviation	Full Name
BGP	Border Gateway Protocol
VXLAN	Virtual eXtensible Local Area Network
EVPN	Ethernet VPN
ARP	Address Resolution Protocol
VTEP	VXLAN Tunnel Endpoints
M-LAG	Multi-Chassis Link Aggregation Group