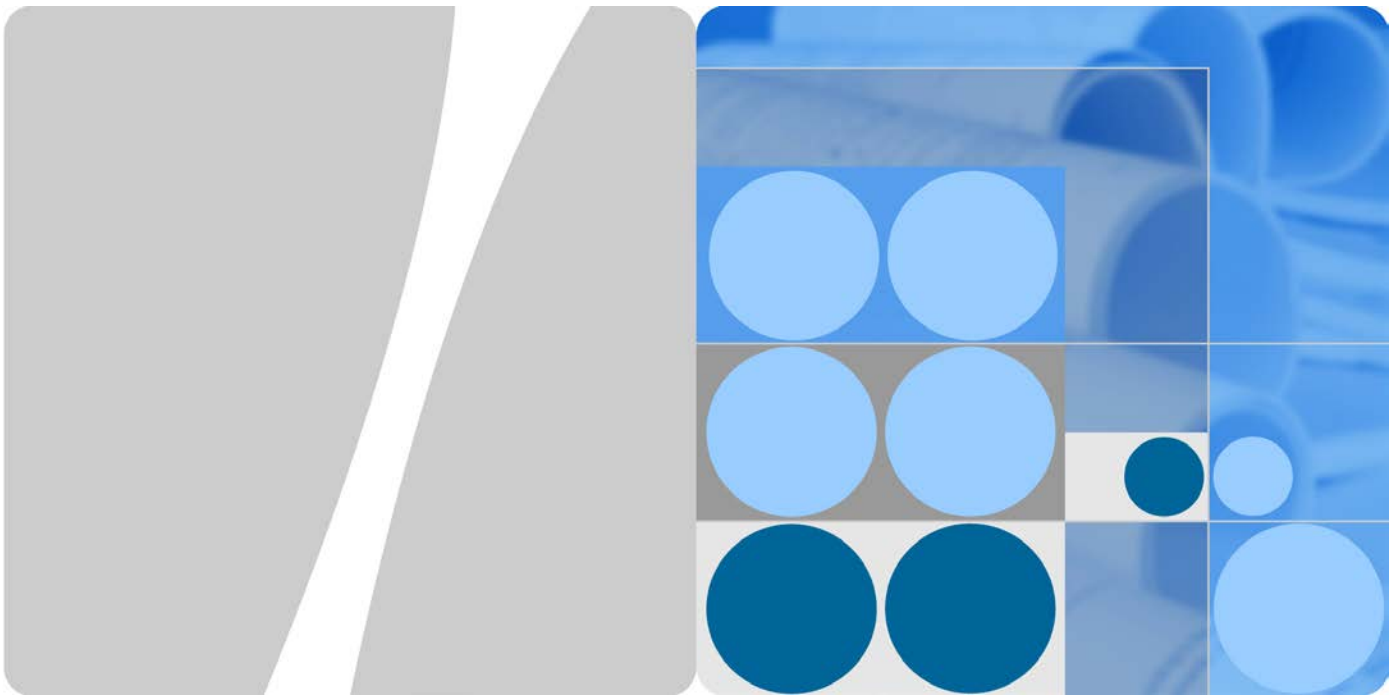


S1720&S2700&S5700&S6720 V200R012C00 Feature Description - VXLAN

Issue 01

Date 2018-07-10

Contents

1 Overview of VXLANs

Definition

As defined by RFC 7348, Virtual eXtensible Local Area Network (VXLAN) is a Network Virtualization over Layer 3 (NVO3) technology that uses the MAC in User Datagram Protocol (MAC-in-UDP) mode to encapsulate packets.

Purpose

Cloud computing has become the new trend in enterprise IT construction with its features such as high system utilization, low manpower and management costs, flexibility, and strong scalability. As a core technology of cloud computing, server virtualization has a wide range of applications.

NOTE
For detailed description about server virtualization, see 6.1 Server Virtualization.

The wide application of server virtualization technology greatly increases computing density in a data center. In addition, VMs need to freely migrate on the network to meet service change requirements. These bring challenges to traditional data center networks of the Layer 2 + Layer 3 architecture.

VXLAN addresses the preceding problems:

- For VM scale limitations imposed by table entry capacities

  Server virtualization leads to an exponential growth of the number of VMs, compared with physical servers. However, the MAC address table size of a Layer 2 device at the access side is incapable to meet this change.

  VXLAN encapsulates original data packets sent from VMs in the same domain into UDP packets, with the IP and MAC addresses used on the physical network in outer headers. The network is only aware of the encapsulated parameters. This greatly reduces the number of MAC address entries required on large Layer 2 networks.
- For limited network isolation capabilities

  While VLAN is the most commonly used network isolation technology, it has its own limitations. The VLAN field in packets is only 12 bits long, which means that at most 4096 VLANs can be used on a network. In public cloud or other cloud computing

scenarios involving tens of thousands or even more tenants, VLAN technology can no longer meet network isolation requirements.

NOTE

A tenant is a complete collection of logical resources deployed on a data center network, including network resources such as VLANs and IP address pools, as well as computing resources such as physical servers and VMs. Each tenant has its own tenant administrator to orchestrate and deploy network services.

VXLAN uses a VXLAN Network Identifier (VNI) field similar to the VLAN ID field to identify users. The VNI field has 24 bits and can identify up to 16 million VXLAN segments, effectively isolating massive tenants in cloud computing scenarios.

- For limited VM migration scope

  VM migration is a process in which a VM moves from one physical server to another. To ensure uninterrupted services during VM migration, the IP and MAC addresses of VMs must remain unchanged. To meet this requirement, server migration must occur in a Layer 2 domain. Layer 2 domains on a traditional network are small, limiting the VM migration scope.

  VXLAN encapsulates original packets sent by VMs over a VXLAN tunnel. VMs at two ends of a VXLAN tunnel do not need to know the physical architecture of the transmission network. In this way, VMs using IP addresses in the same network segment are in a Layer 2 domain logically, even if they are on different physical Layer 2 networks. VXLAN technology constructs a virtual large Layer 2 network over a Layer 3 network, so that VMs are on the same large Layer 2 network so long as there are reachable routes between them. The virtual large Layer 2 network enlarges the VM migration scope.

NOTE

For detailed description about large Layer 2 network, see 6.2 Large Layer 2 Network.

Benefits

When server virtualization is widely deployed in data centers based on physical network infrastructure, VXLAN offers the following benefits:

- As a Layer 2 VPN technology, VXLAN establishes a Layer 2 virtual network over any networks with reachable routes to implement communication within a VXLAN network through the VXLAN gateway as well as communication between a VXLAN network and a non-VXLAN network.
- VXLAN uses MAC-in-UDP encapsulation to extend Layer 2 networks. It encapsulates Ethernet packets into IP packets for these Ethernet packets to be transmitted over routes, and does not need to be aware of VMs' MAC addresses. Because there is no limitation on Layer 3 network architecture, Layer 3 networks are scalable capabilities. This allows for VM migration irrespective of the network architecture.

2 Understanding VXLANs

About This Chapter

2.1 VXLAN Network Architecture

VXLAN is an NVO3 network virtualization technology that encapsulates data packets sent from original hosts into UDP packets and encapsulates IP and MAC addresses used on the physical network in outer headers before sending the packets over an IP network. The virtual tunnel endpoint (VTEP) then decapsulates the packets and sends the packets to the destination host.

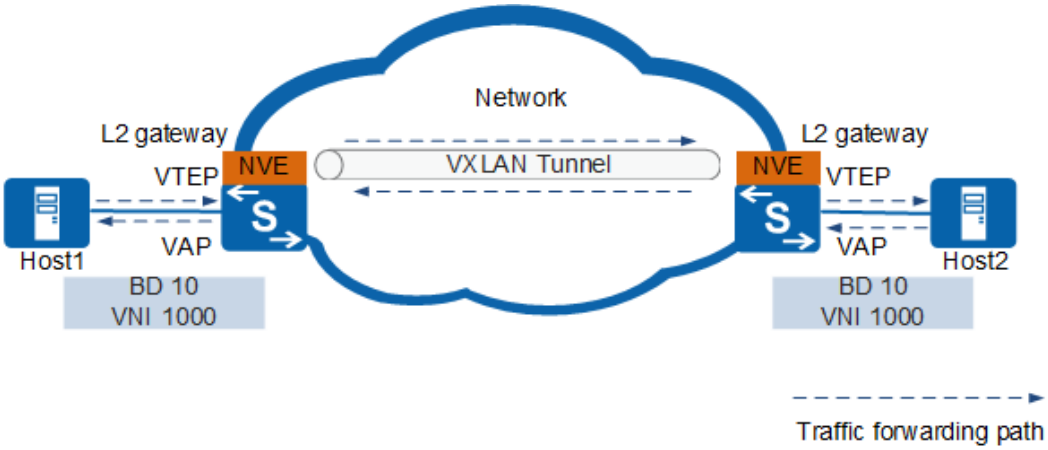
By leveraging VXLAN, a virtual network can accommodate a large number of tenants. Tenants can plan their own virtual networks without being limited by physical network IP addresses or broadcast domains. This technology significantly simplifies network management, allows VMs to migrate over a large Layer 2 network, and isolates tenants in a virtual.

Similar to a traditional VLAN, a VXLAN also allows for intra- and inter-VXLAN communication.

Intra-VXLAN Communication

VXLAN technology constructs a virtual Layer 2 network over a Layer 3 network, implementing Layer 2 communication between VMs. Figure 2-1 shows intra-VXLAN communication.

Figure 2-1 Intra-VXLAN Communication



Involved concepts

- VXLAN Network Identifier (VNI)

  A VNI is similar to a VLAN ID on a traditional network, and it identifies a VXLAN segment. Tenants on different VXLAN segments cannot communicate at Layer 2. One tenant may have one or more VNIs. A VNI consists of 24 bits and supports up to 16 million tenants.

- Broadcast Domain (BD)

  Similar to VLANs divided on a traditional network, BD is used for broadcast domain division on a VXLAN.

  On a VXLAN, to allow Layer 2 communication between VMs in a BD, VNIs and BDs are mapped in 1:1 mode.

- VXLAN VTEP

  A VTEP encapsulates and decapsulates VXLAN packets.

  The source and destination IP addresses in a VXLAN packet are the IP addresses of the local and remote VTEPs, respectively. A pair of VTEP addresses defines one VXLAN tunnel. A source VTEP encapsulates packets and selects a tunnel to forward them. The corresponding destination VTEP decapsulates the received packets.

- Virtual Access Point (VAP)

  A VAP is a VXLAN service access point used for service access based on VLANs or packet encapsulation modes. For more information, see 2.3 Packet Identification:

  - Service access based on VLANs: The 1:1 or N:1 mapping between VLANs and BDs is configured on VTEPs. When a VTEP receives a service packet, it forwards the packet in a BD based on the mapping between VLANs and BDs.
  - Service access based on packet encapsulation modes: Layer 2 sub-interfaces are created on a downlink physical interface of a VTEP, and different encapsulation modes are configured for these sub-interfaces to enable different interfaces to receive different data packets. The 1:1 mapping between Layer 2 sub-interfaces and BDs is also defined. Then service packets are sent to specific Layer 2 sub-interfaces after reaching the VTEP. That is, packets are forwarded in a BD based on the mapping between Layer 2 sub-interfaces and BDs.

- Network Virtualization Edge (NVE)

An NVE is a network entity used to implement network virtualization functions. After packets are encapsulated and decapsulated through NVEs, a Layer 2 VXLAN can be established between NVEs over the basic Layer 3 network.

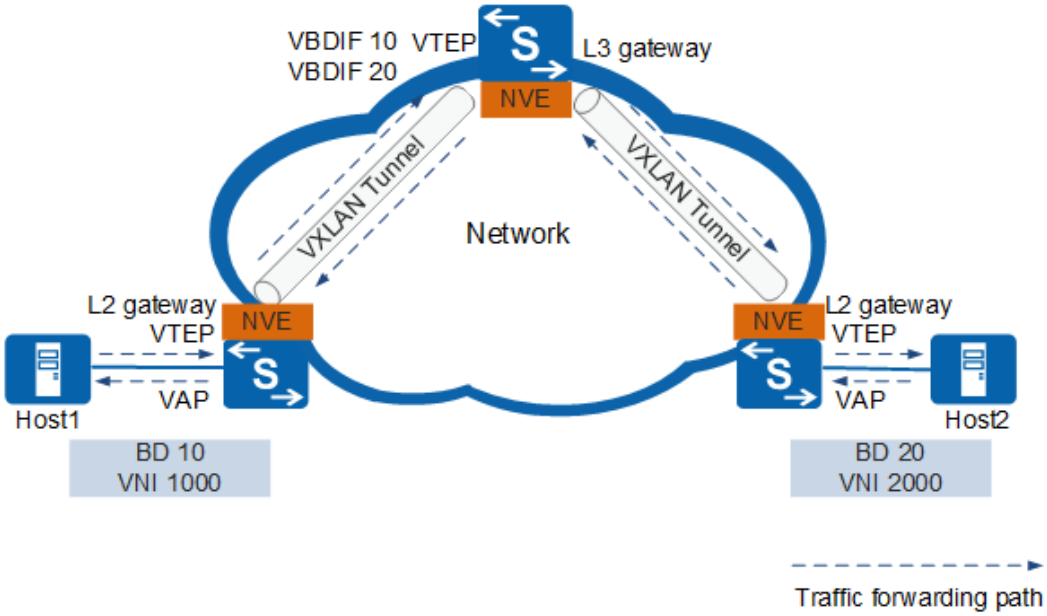
- Layer 2 gateway

  Similar to a Layer 2 access device on a traditional network, it allows tenant access to VXLANs and intra-subnet VXLAN communication in the same network segment.

Inter-VXLAN Network Communication (Centralized Gateway)

VMs in different BDs cannot directly communicate at Layer 2. VXLAN Layer 3 gateways need to be configured to implement Layer 3 communication between VMs. Figure 2-2 shows inter-VLAN communication.

Figure 2-2 Inter-VXLAN Communication



Involved concepts

- Layer 3 gateway

  On a traditional network, users in different VLANs cannot directly communication at Layer 2. Layer 2 communication is also not allowed between VXLANs identified by different VNIs or between VXLANs and non-VXLANs. To address these problems, the VXLAN Layer 3 gateway is introduced to enable data transmission between VXLANs or between VXLANs and non-VXLANs.

  The VXLAN Layer 3 gateway is used for cross-subnet communication on the VXLAN and external network access.

- VBDIF interface

  On a traditional network, VLANIF interfaces are used to enable communication between different BDs. Similarly, VBDIF interfaces are introduced in a VXLAN to implement such function.

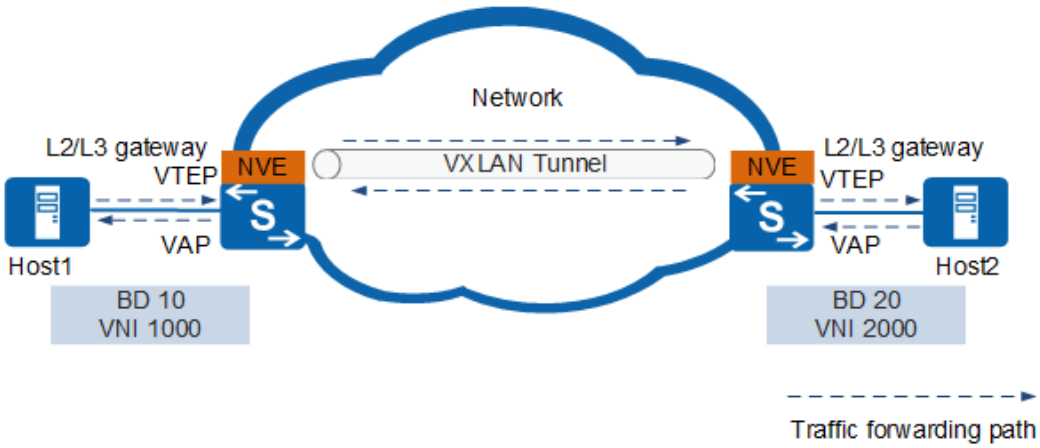
  The VBDIF interface is configured on the VXLAN Layer 3 gateway and is a Layer 3 logical interface based on BDs. After IP addresses are configured for VBDIF interfaces,

VXLANs on different network segments, VXLANs and non-VXLANs, and Layer 2 and Layer 3 networks can communicate with each other.

Inter-VXLAN Network Communication (Distributed Gateway)

A distributed gateway is the device that supports the functions of a VXLAN Layer 2 gateway and a Layer 3 gateway. As shown in Figure 2-3, the VTEP device work as a Layer 2 gateway on the VXLAN network and is connected to hosts, allowing terminal tenants to access the VXLAN network. The VTEP device can also work as a Layer 3 gateway on the VXLAN network, allowing terminal tenants across subnets to communicate with each other and access the extranet. The distributed gateway is supported only for the VXLAN network deployed in BGP EVPN mode.

Figure 2-3 Inter-VXLAN communication



The VXLAN distributed gateway has the following characteristics:

- One VTEP node can work as a VXLAN Layer 2 or 3 gateway, enabling flexible deployment.
- Unlike the centralized Layer 3 gateway which has to learn the ARP entries of all servers, the VTEP node only needs to learn the ARP entries of the connected server, solving the ARP entry problem of the centralized Layer 3 gateway and improving network scalability.

Comparison Between VXLAN and VLAN

The following table lists the differences between VXLAN and VLAN.

Table 2-1 Comparison between VXLAN and VLAN

Item	VLAN	VXLAN
Concept	Virtual local area network	Virtual extensible local area network
Implementation Method	A physical LAN is divided into multiple BDs logically to limit the network to a small geographic range.	Layer 2 virtual networks are established between networks with reachable routes. Such networks are not subject to geographical restrictions and can deliver a

Item	VLAN	VXLAN
		large-scale scalability.
Supported capacity	VLAN is the most commonly used network isolation technology. The VLAN field in packets is only 12 bits in length, which means that only a maximum of 4096 VLANs can be used on a network. In public cloud or other cloud computing scenarios involving tens of thousands or even more tenants, VLAN technology can no longer meet network isolation requirements.	VXLAN is a new network isolation technology defined in IETF RFC 7348. It has a 24-bit segment identifier (VNI) and can isolate up to 16 million tenants. This technology effectively enables isolation of mass tenants in cloud computing.
Network division mode	VLAN IDs are used to divide broadcast domains. Hosts within a BD can communicate at Layer 2.	BDs are used to divide broadcast domains. VMs within a BD can communicate at Layer 2.
Encapsulation mode	A VLAN tag is added to packets.	During VXLAN encapsulation, a VXLAN header, UDP header, IP header, and outer MAC header are added in sequence to an original packet. For details, see 2.2 Packet Encapsulation Format.
Network communication mode	Inter-VLAN communication is implemented by VLANIF interfaces. As Layer 3 logical interfaces, VLANIF interfaces enable Layer 3 communication between VLANs.	Communication between VXLANs or between VXLANs and non-VXLANs is implemented by VBDIF interfaces. VBDIF interfaces are configured on VXLAN Layer 3 gateways and are Layer 3 logical interfaces based on BDs.
Benefits	Limits broadcast domains: A broadcast domain is limited in a VLAN, which saves bandwidth and improves network processing capabilities. Enhances LAN security: Packets from different VLANs are separately transmitted. Hosts in a VLAN cannot directly communicate with hosts in another VLAN.	Location-independent capability: Services can be deployed flexibly at any location, solving network expansion issues related to server virtualization. Flexible network deployment: VXLANs are constructed over the traditional network. They are easy to deploy and highly scalable while preventing broadcast storms on a large Layer 2 network. Cloud service adaptation: A VXLAN is able to isolate ten millions of tenants and support large-scale deployment of cloud services. Technical advantage: VXLAN uses MAC-in-UDP encapsulation. Such encapsulation mode does not rely on MAC addresses of VMs, reducing the number of MAC address entries required on a large Layer 2 network.

2.2 Packet Encapsulation Format

During VXLAN encapsulation, a VXLAN header, UDP header, IP header, and Ethernet header are added in sequence to an original packet.

Figure 2-4 shows the packet encapsulation format.

Figure 2-4 VXLAN packet format

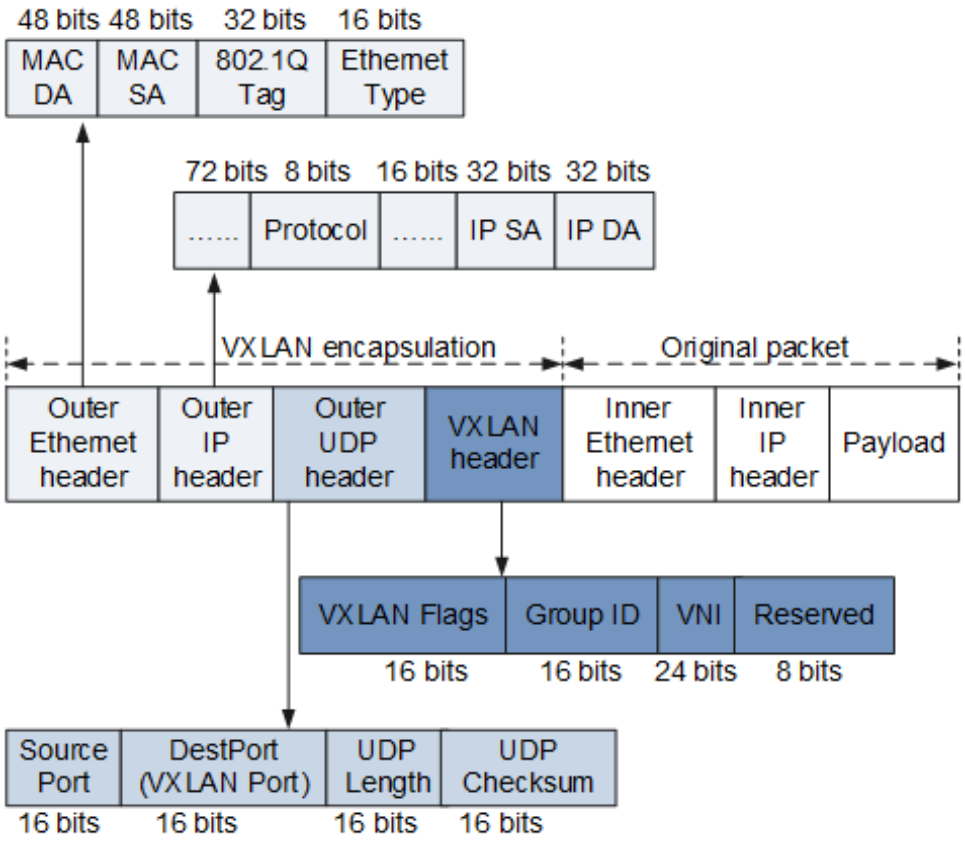


Table 2-2 describes headers added to an original packet during VXLAN encapsulation.

Table 2-2 Description of headers added to an original packet

Field	Description
VXLAN header	• VXLAN Flags: specifies flags (16 bits). • Group ID: indicates the user group ID (16 bits). When the first bit of **VXLAN Flags** is **1**, the value is the group ID. When the first bit of **VXLAN Flags** is **0**, the value is 16 zeros. • VNI: specifies an identifier (24 bits) used to identify a VXLAN segment, with up to 16M tenants. Users in

Field	Description
	different VXLAN segments cannot directly communicate at Layer 2. • Reserved: The 8-bit field is reserved and set to **0**.
Outer UDP header	• DestPort: specifies the destination UDP port number. The value is 4789. • Source Port: specifies the source port number. It is the hash value calculated using parameters in the inner Ethernet frame header.
Outer IP header	• IP SA: specifies the source IP address, which is the IP address of the source VTEP. • IP DA: specifies the destination IP address, which is the IP address of the destination VTEP.
Outer Ethernet header	• MAC DA: specifies the destination MAC address, which is the MAC address of the next-hop device on the route to the destination VTEP. • MAC SA: specifies the source MAC address, which is the MAC address of the source VTEP that sends the packet. • 802.1Q Tag (optional): specifies the VLAN tag in the packet. • Ethernet Type: specifies the type of the Ethernet frame. The value of this field is 0x0800 when an IP packet is transmitted.

2.3 Packet Identification

On a VXLAN network, VNIs are mapped to BDs in 1:1 mode. After a packet reaches a VTEP, the VTEP can identify the BD to which the packet belongs, then select a correct tunnel to forward the packet. Two methods are available for a VTEP to identify the VXLAN to which a packet belongs.

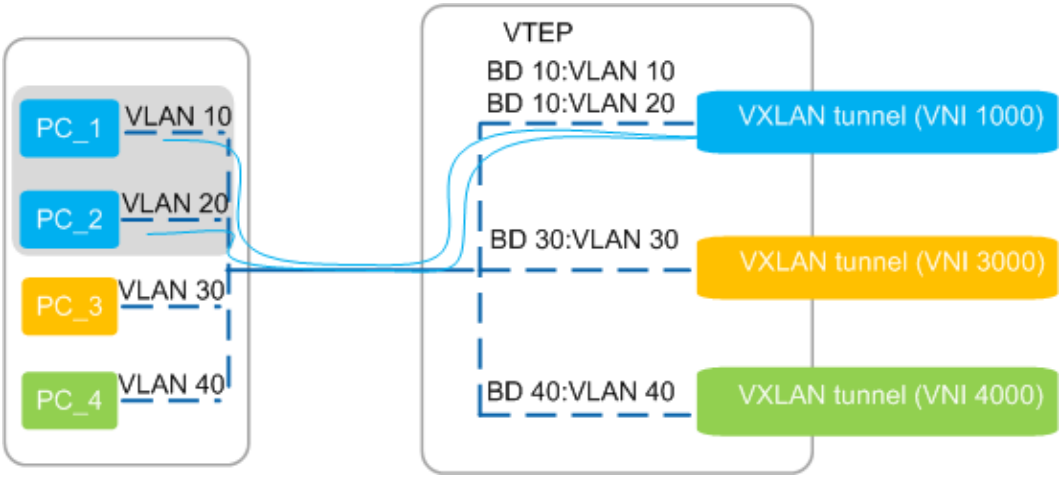
VXLAN Identification by VLAN

The 1:1 or N:1 mapping between VLANs and BDs is configured on VTEPs based on network planning. After a VTEP receives a service packet, it correctly selects a VXLAN tunnel to forward the packet based on the mapping between VLANs and BDs and the mapping between BDs and VNIs.

In Figure 2-5, VLAN 10 and VLAN 20 belong to BD 10. The mapping between VLANs 10 and 20 and BD 10, as well as the mapping between BD 10 and VNI 1000 are configured on

the VTEP. After the VTEP receives a packet from PC_1 or PC_2, the VTEP forwards the packet over the VXLAN tunnel for VNI 1000.

Figure 2-5 VXLAN identification by VLAN



VXLAN Identification by Encapsulation Mode

An encapsulation mode defines packet processing based on whether a packet contains VLAN tags. To implement VXLAN identification by encapsulation mode, Layer 2 sub-interfaces need to be configured on a downlink physical interface of a VTEP, and different encapsulation modes need to be configured for these sub-interfaces. The 1:1 mapping between Layer 2 sub-interfaces and BDs should also be defined. Then service packets are sent to specific Layer 2 sub-interfaces after reaching the VTEP. The VTEP selects a correct VXLAN tunnel to forward packets based on the mapping between Layer 2 sub-interfaces and BDs and the mapping between BDs and VNIs.

Table 2-3 lists four default packet processing methods of Layer 2 sub-interfaces that use different encapsulation modes.

Table 2-3 Packet processing in different encapsulation modes by default

Encapsulation Mode	Allowed Packet Type	Packet Encapsulation	Packet Decapsulation
dot1q	With specified VLAN tag	Removes the VLAN tag from original packets.	Adds a VLAN tag to packets based on the VLAN ID for Dot1q termination on the sub-interface after VXLAN decapsulation and then forwards them.
untag	Without VLAN tags	Does not perform any operation on the original packets.	Does not perform any operation, including adding, replacing, or removing the VLAN

Encapsulation Mode	Allowed Packet Type	Packet Encapsulation	Packet Decapsulation
			tag, on packets after VXLAN decapsulation is implemented.
default	All packets regardless of whether they contain VLAN tags	Does not perform any operation on the original packets.	Does not perform any operation, including adding, replacing, or removing the VLAN tag, on packets after VXLAN decapsulation is implemented.
qinq	With specified double VLAN tags	Removes all the VLAN tags from original packets.	After implementing VXLAN decapsulation: • S5730HI, S6720HI, and S5720HI: Add double VLAN tags to packets based on the outer and inner VLAN IDs for QinQ termination on the sub-interface configured using the **qinq termination pe-vid ce-vid** command before forwarding them. • Other models: If received packets do not carry any VLAN tag, add double VLAN tags to packets based on the outer and inner VLAN IDs for QinQ termination on the sub-interface configured using the **qinq termination pe-vid ce-vid** command before

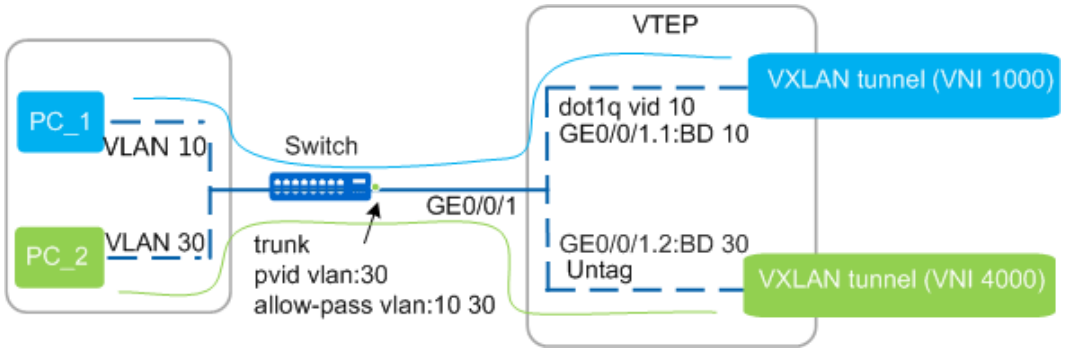
Encapsulation Mode	Allowed Packet Type	Packet Encapsulation	Packet Decapsulation
			forwarding them. If received packets carry VLAN tags, remove the outer VLAN tag and add double VLAN tags to packets based on the outer and inner VLAN IDs for QinQ termination on the sub-interface configured using the **qinq termination pe-vid ce-vid** command before forwarding them.

In Figure 2-6, the physical interface GE0/0/1.1 on the VTEP has two Layer 2 sub-interfaces, which are configured with different encapsulation modes and associated with different BDs. PC_1 and PC_2 belong to VLAN 10 and VLAN 30, respectively. An uplink interface on the Layer 2 switch connecting to the VTEP is configured as a trunk interface with the PVID 30 and is configured to allow packets from VLANs 10 and 30 to pass through. When a packet from PC_1 reaches this interface, the interface transparently transmits the packet to the VTEP because the VLAN ID of the packet is different from the default VLAN ID of the interface. When a packet from PC_2 reaches this interface, the interface removes the VLAN tag 30 from the packet before forwarding it to the VTEP because the VLAN ID of the packet is the same as the default VLAN ID of the interface. As a result, when the packets reach GE0/0/1.1 on the VTEP, the packet from PC_1 contains VLAN tag 10, while the packet from PC_2 does not contain a VLAN tag. To distinguish the two types of packets, Layer 2 sub-interfaces of the **dot1q** and **untag** types need to be configured on GE0/0/1.1:

- The encapsulation mode of the Layer 2 sub-interface GE0/0/1.1 is **dot1q**, allowing packets with VLAN tag 10 to enter the VXLAN tunnel.
- The encapsulation mode of the Layer 2 sub-interface GE0/0/1.1 is **untag**, allowing packets without a VLAN tag to enter the VXLAN tunnel.

After packets from PC_1 or PC_2 reach the VTEP, the VTEP sends the packets to different Layer 2 sub-interfaces based VLAN tags in the packets. Then, the VTEP chooses a correct VXLAN tunnel to forward the packets based on the mapping between sub-interface and BD, as well as the mapping between BD and VNI.

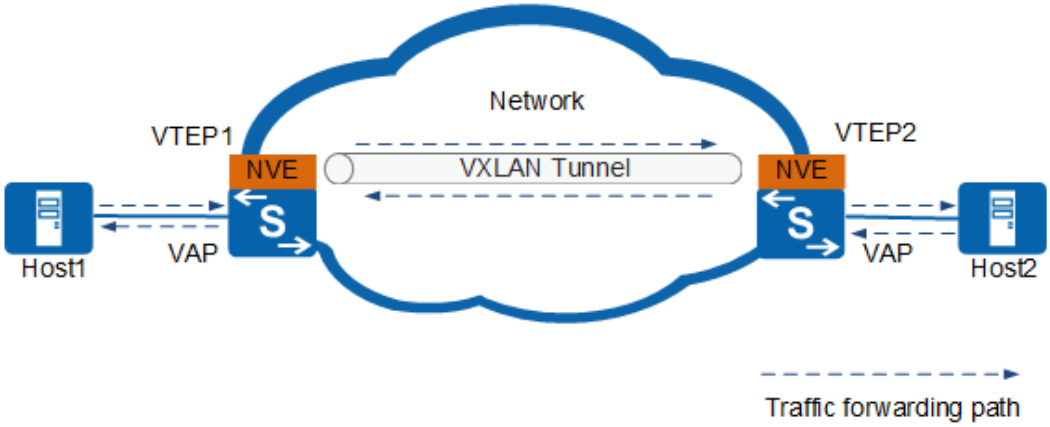
Figure 2-6 VXLAN identification by encapsulation mode



2.4 Establishing VXLAN Tunnels

A virtual extensible LAN (VXLAN) tunnel is determined by a pair of virtual tunnel end point (VTEP) IP addresses. Packets are encapsulated on the VTEP device and then transmitted through the VXLAN tunnel over routes. After the VXLAN tunnel is configured, when the IP addresses of VTEP devices on both ends of the tunnel are reachable at Layer 3, the VXLAN tunnel can be successfully established.

Figure 2-7 VXLAN networking



Based on tunnel creation modes, VXLAN tunnels can be divided into the following two types:

- Static VXLAN tunnels: completed by manually configuring local and remote VNIs, VTEP IP addresses, and headend replication lists. Static tunnel configuration is supported only for the VXLAN in centralized gateway mode. For details, see 3.1 Centralized VXLAN Gateway Deployment in Static Mode.
- Dynamic VXLAN tunnels: dynamically created in BGP EVPN mode. Specifically, create BGP EVPN peers on VTEP devices on both ends and enable the peers to exchange VNI and VTEP IP address information over the BGP EVPN route to dynamically create VXLAN tunnels. Dynamic tunnel configuration through BGP EVPN is supported for the VXLAN in both centralized and distributed gateway modes. For details, see 3.2

Centralized VXLAN Gateway Deployment Using BGP EVPN and 3.3 Distributed VXLAN Gateway Deployment Using BGP EVPN.

2.5 BGP EVPN Basic Principles

Introduction

Ethernet virtual private network (EVPN) is a VPN technology used for Layer 2 internetworking. EVPN is similar to BGP/MPLS IP VPN. EVPN defines a new type of BGP network layer reachability information (NLRI), called the EVPN NLRI. The EVPN NLRI defines new BGP EVPN routes to implement MAC address learning and advertisement between Layer 2 networks at different sites.

VXLAN does not provide the control plane, and VTEP discovery and host information (IP and MAC addresses, VNIs, and gateway VTEP IP address) learning are implemented by traffic flooding on the data plane, resulting in high traffic volumes on VXLAN networks. To address this problem, VXLAN uses EVPN as the control plane. EVPN allows VTEPs to exchange BGP EVPN routes to implement automatic VTEP discovery and host information advertisement, preventing unnecessary traffic flooding.

EVPN uses extended BGP and defines new BGP EVPN routes to transmit VTEP addresses and host information. As such, the application of EVPN on VXLANs moves VTEP discovery and host information learning from the data plane to the control plane.

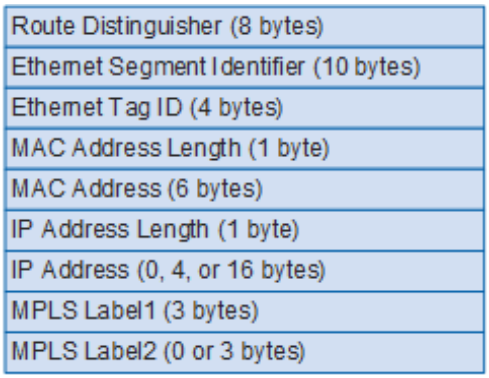
BGP EVPN Routes

EVPN NLRI defines the following BGP EVPN route types applicable to the VXLAN control plane:

Type 2 route—MAC/IP route

The following figure shows the format of MAC/IP routes.

Figure 2-8 MAC/IP route



The following table describes the fields.

Field	Description
Route Distinguisher	RD value of an EVPN instance

Field	Description
Ethernet Segment Identifier	Unique ID for defining the connection between local and remote devices
Ethernet Tag ID	VLAN ID configured on the device
MAC Address Length	Length of the host MAC address carried in the route
MAC Address	Host MAC address carried in the route
IP Address Length	Mask length of the host IP address carried in the route
IP Address	Host IP address carried in the route
MPLS Label1	Layer 2 VNI carried in the route
MPLS Label2	Layer 3 VNI carried in the route

MAC/IP routes function as follows on the VXLAN control plane:

- MAC address advertisement

  To implement Layer 2 communication between intra-subnet hosts, the source and remote VTEPs must learn the MAC addresses of the hosts. The VTEPs function as BGP EVPN peers to exchange MAC/IP routes so that they can obtain the host MAC addresses. The MAC Address Length and MAC Address fields identify the MAC address of a host.

- ARP advertisement

  A MAC/IP route can carry both the MAC and IP addresses of a host, and therefore can be used to advertise ARP entries between VTEPs. The MAC Address and MAC Address Length fields identify the MAC address of the host, whereas the IP Address and IP Address Length fields identify the IP address of the host. This type of MAC/IP route is called the ARP route. ARP advertisement applies to the following scenarios:

  a. ARP broadcast suppression. After a Layer 3 gateway learns the ARP entries of a host, it generates host information that contains the host IP and MAC addresses, Layer 2 VNI, and gateway's VTEP IP address. The Layer 3 gateway then transmits an ARP route carrying the host information to a Layer 2 gateway. When the Layer 2 gateway receives an ARP request, it checks whether it has the host information corresponding to the destination IP address of the packet. If such host information exists, the Layer 2 gateway replaces the broadcast MAC address in the ARP request with the destination unicast MAC address and unicasts the packet. This implementation suppresses ARP broadcast packets.

  b. VM migration in distributed gateway scenarios. After a VM migrates from one gateway to another, the new gateway learns the ARP entry of the VM (after the VM sends gratuitous ARP packets) and generates host information that contains the host IP and MAC addresses, Layer 2 VNI, and gateway's VTEP IP address. The new gateway then transmits an ARP route carrying the host information to the original gateway. After the original gateway receives the ARP route, it detects a VM location change and triggers ARP probe. If ARP probe fails, the original gateway withdraws the ARP and host routes of the VM.

- IP route advertisement

  In distributed VXLAN gateway scenarios, to implement Layer 3 communication between inter-subnet hosts, the source and remote VTEPs that function as Layer 3 gateways must learn the host IP routes. The VTEPs function as BGP EVPN peers to

exchange MAC/IP routes so that they can obtain the host IP routes. The IP Address Length and IP Address fields identify the destination address of the IP route. In addition, the MPLS Label2 field must carry the Layer 3 VNI. This type of MAC/IP route is called the integrated routing and bridging (IRB) route.

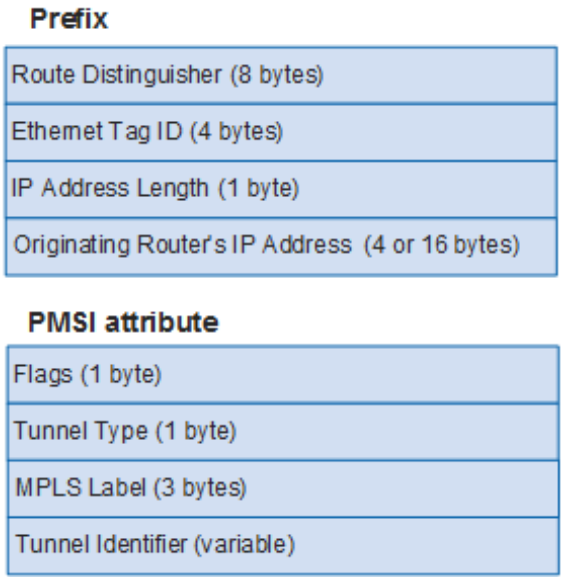
NOTE

An ARP route carries host MAC and IP addresses and a Layer 2 VNI. An IRB route carries host MAC and IP addresses, a Layer 2 VNI, and a Layer 3 VNI. Therefore, IRB routes carry ARP routes and can be used to advertise IP routes as well as ARP entries.

Type 3 route—inclusive multicast route

An inclusive multicast route comprises a prefix and a PMSI attribute.

Figure 2-9 Format of an inclusive multicast route



The following table describes the fields.

Field	Description
Route Distinguisher	RD value of an EVPN instance
Ethernet Tag ID	VLAN ID The value is all 0s in this type of route.
IP Address Length	Mask length of the local VTEP's IP address carried in the route
Originating Router's IP Address	Local VTEP's IP address carried in the route
Flags	Flags indicating whether leaf node information is required for the tunnel This field is inapplicable in VXLAN scenarios.
Tunnel Type	Tunnel type carried in the route

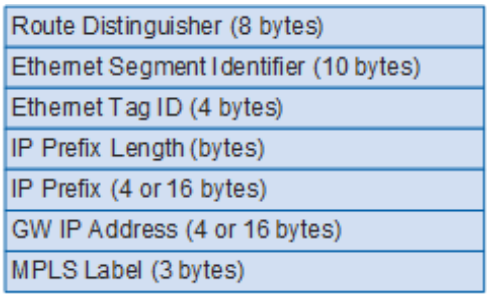
Field	Description
	The value can only be 6, representing Ingress Replication in VXLAN scenarios. It is used for BUM packet forwarding.
MPLS Label	Layer 2 VNI carried in the route
Tunnel Identifier	Tunnel identifier carried in the route This field is the local VTEP's IP address in VXLAN scenarios.

This type of route is used on the VXLAN control plane for automatic VTEP discovery and dynamic VXLAN tunnel establishment. VTEPs that function as BGP EVPN peers exchange inclusive multicast routes to transfer Layer 2 VNIs and VTEPs' IP addresses. The Originating Router's IP Address field identifies the local VTEP's IP address; the MPLS Label field identifies a Layer 2 VNI. If the remote VTEP's IP address is reachable at Layer 3, a VXLAN tunnel to the remote VTEP is established. If the remote VNI is the same as the local VNI, an ingress replication list is created for subsequent BUM packet forwarding.

Type 5 route—IP prefix route

The following figure shows the format of IP prefix routes.

Figure 2-10 IP prefix route



The following table describes the fields.

Field	Description
Route Distinguisher	RD value of an EVPN instance
Ethernet Segment Identifier	Unique ID for defining the connection between local and remote devices
Ethernet Tag ID	VLAN ID configured on the device
IP Prefix Length	Length of the IP prefix carried in the route
IP Prefix	IP prefix carried in the route
GW IP Address	Default gateway address This field is inapplicable in VXLAN scenarios.
MPLS Label	Layer 3 VNI carried in the route

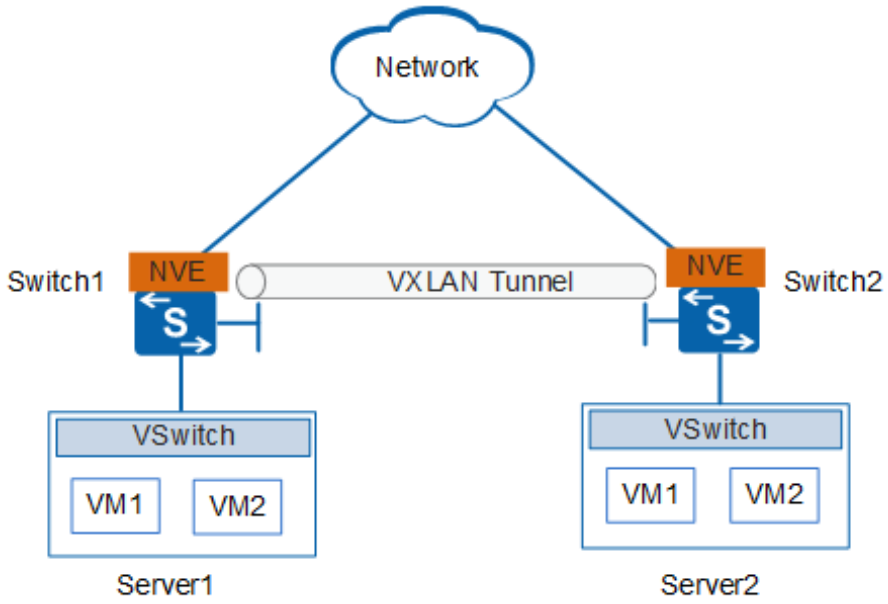
The IP Prefix Length and IP Prefix fields in an IP prefix route can identify a host IP address or network segment.

- If the IP Prefix Length and IP Prefix fields in an IP prefix route identify a host IP address, the route is used for IP route advertisement in distributed VXLAN gateway scenarios, which functions the same as an IRB route on the VXLAN control plane.
- If the IP Prefix Length and IP Prefix fields in an IP prefix route identify a network segment, the route allows external network access.

2.6 VXLAN QoS

VXLAN quality of service (QoS) provides differentiated service in VXLAN applications. A device implements mapping between QoS priorities in original packets, internal priorities (local precedence assigned by the device to differentiate service classes of packets), and priorities of encapsulated packets. In this way, the switch provides the differentiated QoS service based on original packets.

Figure 2-11 VXLAN networking



On the network as shown in Figure 2-11, VXLAN QoS implements mapping between QoS priorities in original packets, internal priorities, and priorities of encapsulated packets according to the following process:

1. An original packet arrives at a Layer 2 sub-interface on Switch1. Switch1 maps the 802.1p priority of the original packet to the internal priority (PHB and color) based on the DiffServ profile bound to the main interface, and then sends the packet to the specified queue.
2. After the packet enters the VXLAN tunnel, the VTEP encapsulates the packet and maps the packet's internal priority to the 802.1p priority or DSCP priority based on the default

profile in the DiffServ domain. The packet is then transmitted over the VXLAN tunnel based on the 802.1p priority or DSCP priority.

3. When the packet leaves the tunnel, its 802.1p priority or DSCP priority (depending on which priority is trusted on the tunnel interface) is mapped to the internal priority based on the default profile in the DiffServ domain. The packet then enters the queue matching the internal priority. An Ethernet interface working in Layer 3 mode trusts the DSCP priority only.
4. Finally, the internal priority is mapped to the 802.1p priority based on the profile in the DiffServ domain bound to the main interface. The packet is transmitted through the outbound interface based on the 802.1p priority.

When the S6720S-EI or S6720EI is used as the access device, it cannot map the DSCP priority of original packets in the outbound direction but can normally map the packet priority in the inbound direction. Table 2-4 lists other mapping rules.

Table 2-4 Mapping rules in a DiffServ domain applied to an interface when the S6720S-EI or S6720EI works as the access device

Access Mode to the VXLAN Network	Mapping Rule
By VLAN	• Inbound: Performs priority mapping based on the 802.1p priority of packets and the DiffServ domain configured on the interface. • Outbound: Modifies the 802.1p priority of packets based on the DiffServ domain configured on the interface.
By the flow encapsulation type **default**	• Inbound: Performs priority mapping based on the priority configured on the main interface by using the **port priority** *priority-value* command and the DiffServ domain configured on the interface, if packets do not carry VLAN tags. Performs priority mapping based on the 802.1p priority of packets and the DiffServ domain configured on the interface, if packets carry VLAN tags. • Outbound: Does not modify the 802.1p priority of packets.
By the flow encapsulation type **untag**	• Inbound: Performs priority mapping based on the priority configured on the main interface by using the **port priority** *priority-value* command and the DiffServ domain configured on the interface. • Outbound: Does not modify the 802.1p priority of packets.
By the flow encapsulation type **dot1q**	• Inbound: Performs priority mapping based on the 802.1p priority of packets and the DiffServ domain configured on the interface.

Access Mode to the VXLAN Network	Mapping Rule
	• Outbound: Modifies the 802.1p priority of packets based on the DiffServ domain configured on the interface.
By the flow encapsulation type **qinq**	• Inbound: Performs priority mapping based on the outer 802.1p priority of packets and the DiffServ domain configured on the interface. • Outbound: Modifies the outer 802.1p priority of packets based on the DiffServ domain configured on the interface.

For details, see Configuring Priority Mapping in "Priority Mapping Configuration" in the *S1720&S2700&S5700&S6720 V200R012C00 Configuration Guide - QoS*.

3 VXLAN Implementation

About This Chapter

3.1 Centralized VXLAN Gateway Deployment in Static Mode

In centralized VXLAN gateway deployment in static mode, the control plane is responsible for VXLAN tunnel establishment and dynamic MAC address learning; the forwarding plane is responsible for intra-subnet known unicast packet forwarding, intra-subnet BUM (Broadcast&Unknown-unicast&Multicast) packet forwarding, and inter-subnet packet forwarding.

Deploying centralized VXLAN gateways in static mode involves heavy workload and is inflexible, and therefore is inapplicable to large-scale networks. As such, 3.2 Centralized VXLAN Gateway Deployment Using BGP EVPN is recommended.

VXLAN Tunnel Establishment

A VXLAN tunnel is identified by a pair of VTEP IP addresses. A VXLAN tunnel can be statically created after you configure local and remote VNIs, VTEP IP addresses, and an ingress replication list, and the tunnel goes Up when the pair of VTEPs are reachable at Layer 3.

On the network shown in Figure 3-1, VTEP 2 connects to Host 1 and Host 3; VTEP 3 connects to Host 2; VTEP 1 functions as a Layer 3 gateway.

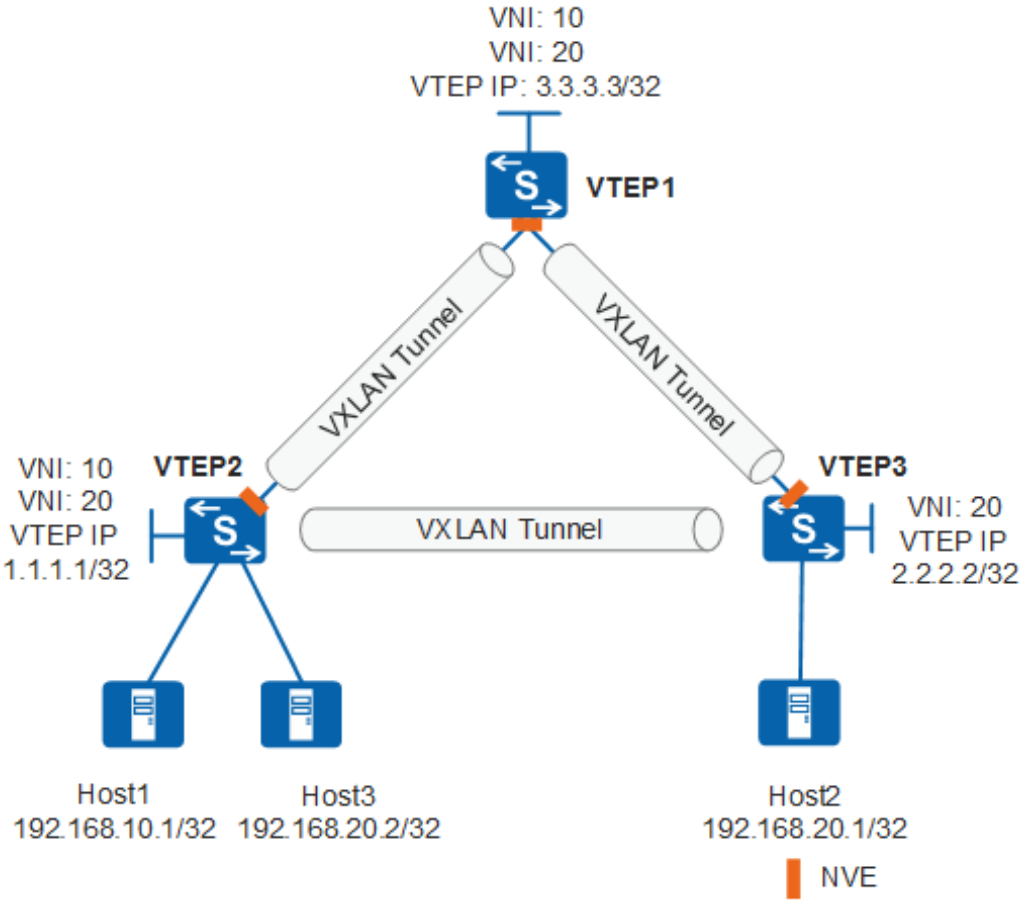
- To allow Host 3 and Host 2 to communicate, Layer 2 VNIs and an ingress replication list must be configured on VTEP 2 and VTEP 3. The peer VTEPs' IP addresses must be specified in the ingress replication list. A VXLAN tunnel can be established between VTEP 2 and VTEP 3 if their VTEPs have Layer 3 routes to each other.
- To allow Host 1 and Host 2 to communicate, Layer 2 VNIs and an ingress replication list must be configured on VTEP 2, VTEP 3, and also VTEP 1. The peer VTEPs' IP addresses must be specified in the ingress replication list. A VXLAN tunnel can be

established between VTEP 2 and VTEP 1 and between VTEP 3 and VTEP 1 if they have Layer 3 routes to the IP addresses of the VTEPs of each other.

NOTE

Although Host 1 and Host 3 both connect to VTEP 2, they belong to different subnets and must communicate through the Layer 3 gateway (VTEP 1). Therefore, a VXLAN tunnel is also required between VTEP 2 and VTEP 1.

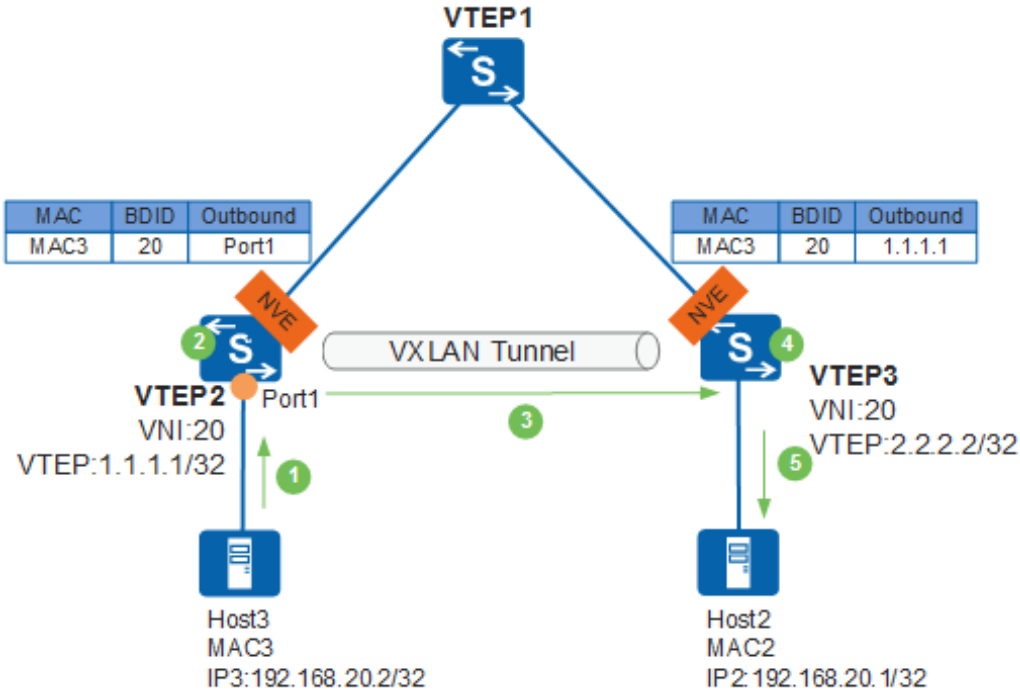
Figure 3-1 VXLAN tunnel networking



Dynamic MAC Address Learning

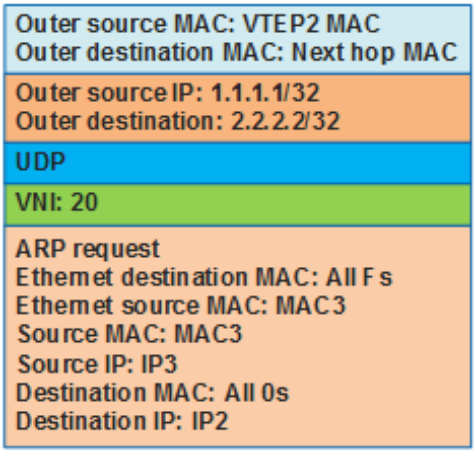
VXLAN supports dynamic MAC address learning to allow communication between tenants. MAC address entries are dynamically created and do not need to be manually maintained, greatly reducing maintenance workload. The following example illustrates dynamic MAC address learning for intra-subnet communication on the network shown in Figure 3-2.

Figure 3-2 Dynamic MAC Address Learning



1. Host 3 sends an ARP request for Host 2's MAC address. The ARP request carries the source MAC address being MAC3, destination MAC address being all Fs, source IP address being IP3, and destination IP address being IP2.
2. Upon receipt of the ARP request, VTEP 2 determines that the Port1 receiving the ARP request belongs to a BD that has been bound to a VNI (20), meaning that the ARP request packet must be transmitted over the VXLAN tunnel identified by VNI 20. VTEP 2 then learns the mapping between Host 3's MAC address, BDID (Layer 2 broadcast domain ID), and inbound interface (Port1) that has received the ARP request and generates a MAC address entry for Host 3. The MAC address entry's outbound interface is Port1.
3. VTEP 2 then performs VXLAN encapsulation on the ARP request, with the VNI being the one bound to the BD, source IP address in the outer IP header being the VTEP's IP address of VTEP 2, destination IP address in the outer IP header being the VTEP's IP address of VTEP 3, source MAC address in the outer Ethernet header being MAC address of VTEP 2, and destination MAC address in the outer Ethernet header being the MAC address of the next hop pointing to the destination IP address. Figure 3-3 shows the VXLAN packet format. The VXLAN packet is then transmitted over the IP network based on the IP and MAC addresses in the outer headers and finally reaches VTEP 3.

Figure 3-3 VXLAN packet format



4. After VTEP 3 receives the VXLAN packet, it decapsulates the packet and obtains the ARP request originated from Host 3. VTEP 3 then learns the mapping between Host 3's MAC address, BDID, and VTEP's IP address of VTEP 2 and generates a MAC address entry for Host 3. Based on the next hop (VTEP's IP address of VTEP 2), the MAC address entry's outbound interface is iterated to the VXLAN tunnel destined for VTEP 2.
5. VTEP 3 broadcasts the ARP request in the Layer 2 domain. Upon receipt of the ARP request, Host 2 finds that the destination IP address is its own IP address and saves Host 3's MAC address to the local MAC address table. Host 2 then responds with an ARP reply.

So far, Host 2 has learned Host 3's MAC address. Therefore, Host 2 responds with a unicast ARP reply. The ARP reply is transmitted to Host 3 in the same manner. After Host 2 and Host 3 learn the MAC address of each other, they will subsequently communicate with each other in unicast mode.

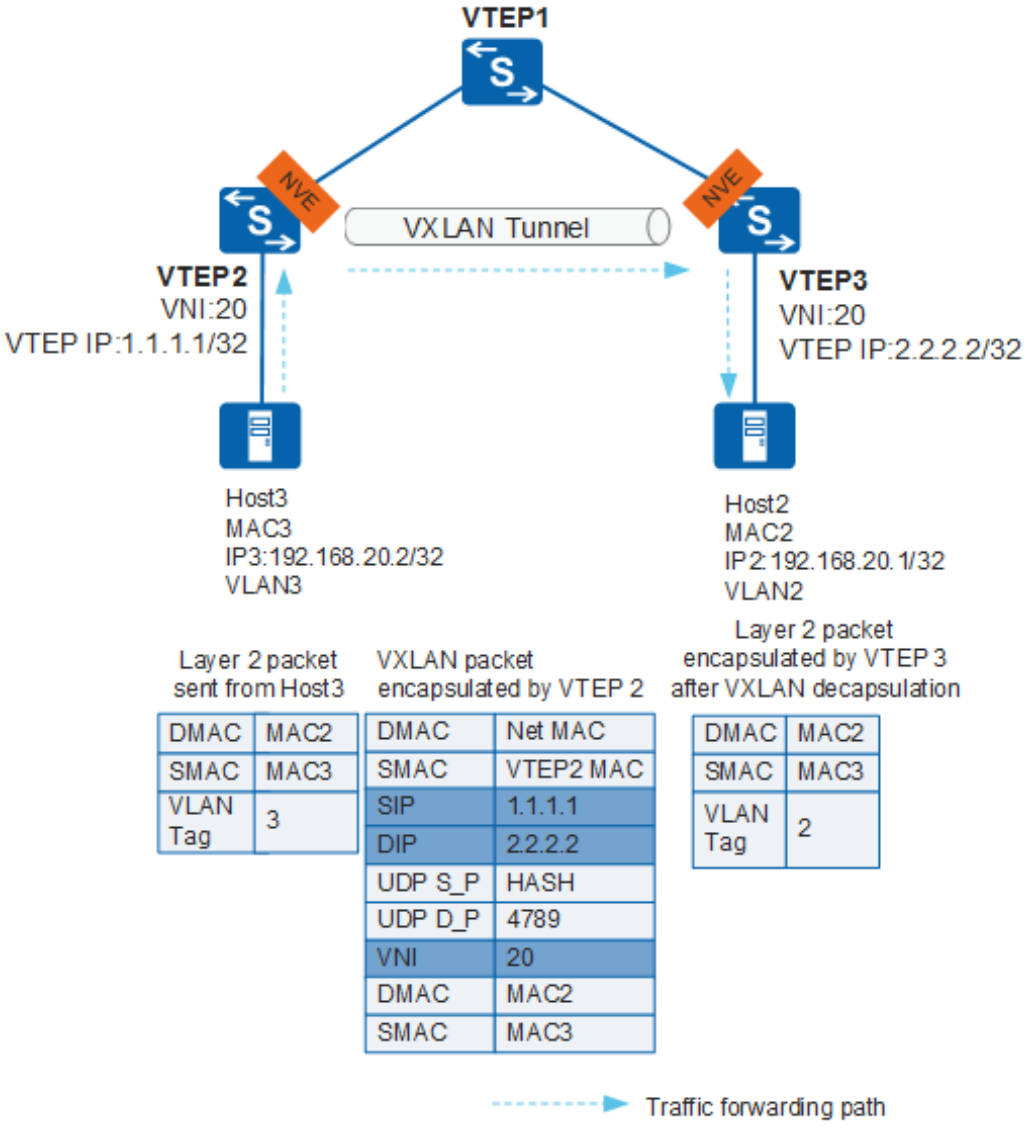
NOTE

Dynamic MAC address learning is required only between hosts and Layer 3 gateways in inter-subnet communication scenarios. The process is the same as that for intra-subnet communication.

Intra-Subnet Known Unicast Packet Forwarding

Intra-subnet known unicast packets are forwarded only through Layer 2 VXLAN gateways and are unknown to Layer 3 VXLAN gateways. Figure 3-4 shows the intra-subnet known unicast packet forwarding process.

Figure 3-4 Intra-subnet known unicast packet forwarding



1. After VTEP 2 receives Host 3's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information and searches for the outbound interface and encapsulation information in the BD.
2. VTEP 2 performs VXLAN encapsulation based on the encapsulation information obtained and forwards the packets through the outbound interface obtained.
3. Upon receipt of the VXLAN packet, VTEP 3 verifies the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 3 obtains the Layer 2 BD based on the VNI and performs VXLAN decapsulation to obtain the inner Layer 2 packet.
4. VTEP 3 obtains the destination MAC address of the inner Layer 2 packet, performs VLAN tags to the packets based on the outbound interface and encapsulation information in the local MAC address table, and forwards the packets to Host 2.

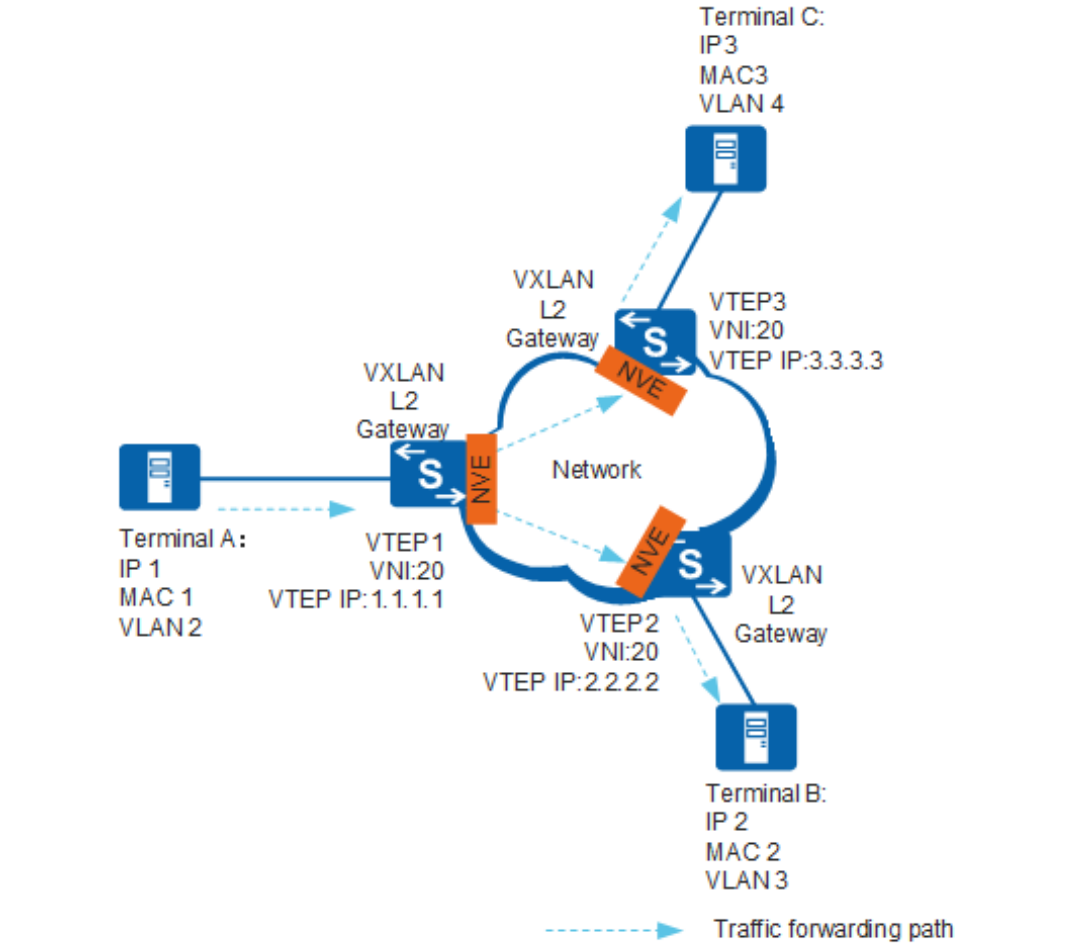
Host 2 sends packets to Host 3 in the same manner.

Intra-Subnet BUM Packet Forwarding

Intra-subnet BUM packet forwarding is completed between Layer 2 VXLAN gateways. Layer 3 VXLAN gateways do not need to be unaware of the process. Intra-subnet BUM packets can be forwarded in ingress replication mode.

In ingress replication mode, after a BUM packet enters a VXLAN tunnel, the ingress VTEP performs VXLAN encapsulation based on the ingress replication list and sends the packet to all the egress VTEPs in the list. When the BUM packet leaves the VXLAN tunnel, the egress VTEPs decapsulate the BUM packet. Figure 3-5 shows the forwarding process of a BUM packet in ingress replication mode.

Figure 3-5 Forwarding process of an intra-subnet BUM packet in ingress replication mode



Layer 2 packet sent from Terminal A

DMAC	All Fs
SMAC	MAC1
VLAN Tag	2

VXLAN packet encapsulated by VTEP 1

VTEP1→VTEP2

DMAC	Net MAC
SMAC	VTEP1 MAC1
SIP	1.1.1.1
DIP	2.2.2.2
UDP S_P	HASH
UDP D_P	4789
VNI	20
DMAC	All Fs
SMAC	MAC1

VTEP1→VTEP3

DMAC	Net MAC
SMAC	VTEP1 MAC1
SIP	1.1.1.1
DIP	3.3.3.3
UDP S_P	HASH
UDP D_P	4789
VNI	20
DMAC	All Fs
SMAC	MAC1

Layer 2 packet encapsulated by VTEP 2/ VTEP 3 after VXLAN decapsulation

DMAC	All Fs
SMAC	MAC1
VLAN Tag	3

DMAC	All Fs
SMAC	MAC1
VLAN Tag	4

1. After VTEP 1 receives Terminal A's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information.
2. VTEP 1 obtains the ingress replication list for the VNI, replicates packets based on the list, and performs VXLAN encapsulation by adding outer headers. VTEP 1 then forwards the VXLAN packet through the outbound interface.

3. Upon receipt of the VXLAN packet, VTEP 2's VTEP and VTEP 3's VTEP verify the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 2/VTEP 3 obtains the Layer 2 BD based on the VNI and performs VXLAN decapsulation to obtain the inner Layer 2 packet.
4. VTEP 2/VTEP 3 checks the destination MAC address of the inner Layer 2 packet and finds it a BUM MAC address. Therefore, VTEP 2/VTEP 3 broadcasts the packet onto the network connected to the terminals (not the VXLAN tunnel side) in the Layer 2 broadcast domain. Specifically, VTEP 2/VTEP 3 finds the outbound interfaces and encapsulation information not related to the VXLAN tunnel, performs VLAN tags to the packet, and forwards the packet to Terminal B/Terminal C.

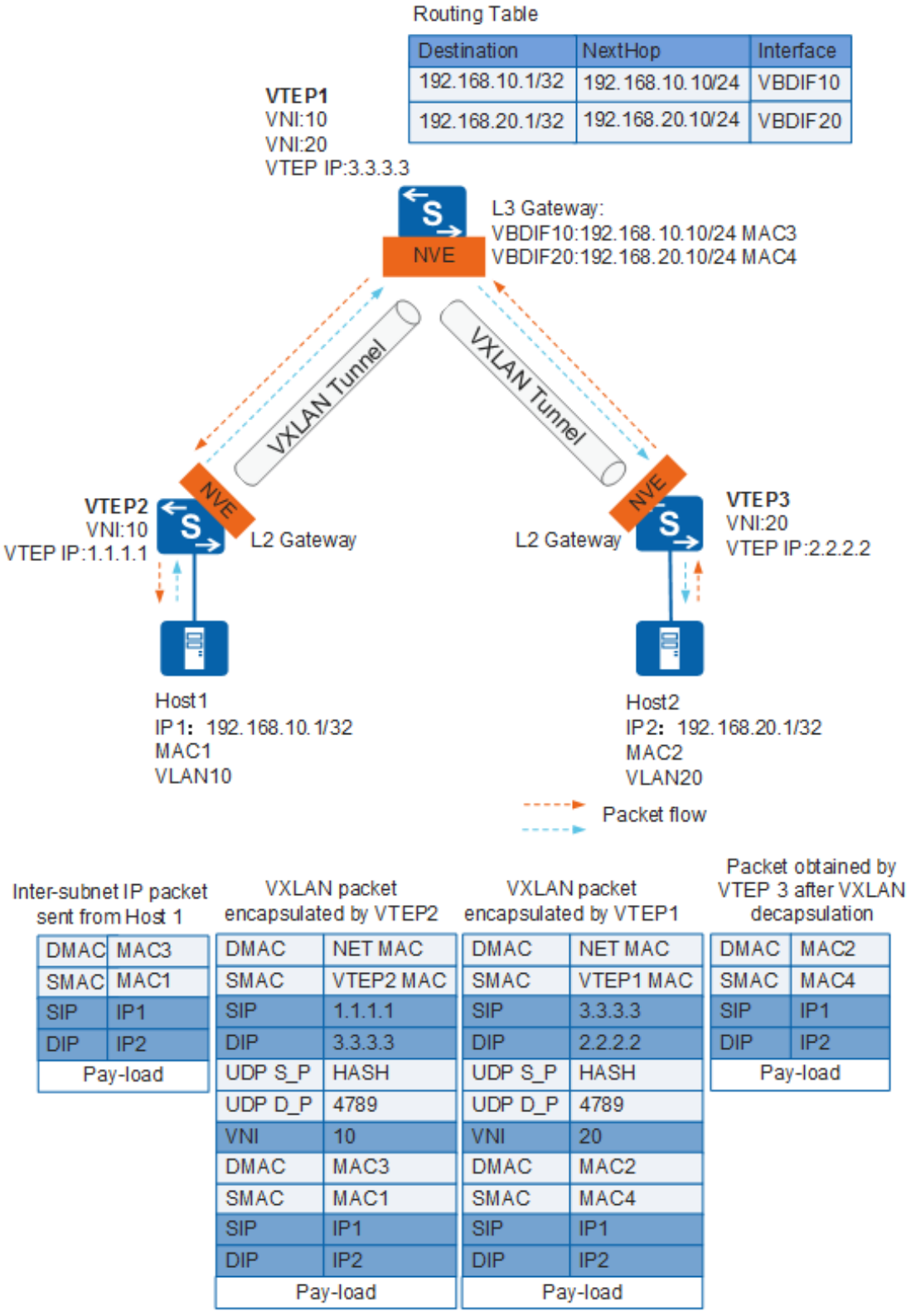
NOTE

Terminal B/Terminal C responds to Terminal A in the same process as intra-subnet known unicast packet forwarding.

Inter-Subnet Packet Forwarding

Inter-subnet packets must be forwarded through a Layer 3 gateway. Figure 3-6 shows inter-subnet packet forwarding in centralized VXLAN gateway scenarios.

Figure 3-6 Inter-subnet packet forwarding



1. After VTEP 2 receives Host 1's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information and searches for the outbound interface and encapsulation information in the BD.
2. VTEP 2 performs VXLAN encapsulation based on the outbound interface and encapsulation information and forwards the packets to VTEP 1.
3. After VTEP 1 receives the VXLAN packet, it decapsulates the packet and finds that the destination MAC address of the inner packet is the MAC address (MAC3) of the Layer 3 gateway interface (VBDIF10) so that the packet must be forwarded at Layer 3.
4. VTEP 1 removes the inner Ethernet header, parses the destination IP address, and searches the routing table for a next hop address. VTEP 1 then searches the ARP table based on the next hop address to obtain the destination MAC address, VXLAN tunnel's outbound interface, and VNI.
5. VTEP 1 performs VXLAN encapsulation on the inner packet again and forwards the VXLAN packet to VTEP 3, with the source MAC address in the inner Ethernet header being the MAC address (MAC4) of the Layer 3 gateway interface (VBDIF20).
6. Upon receipt of the VXLAN packet, VTEP 3 verifies the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 3 then obtains the Layer 2 broadcast domain based on the VNI and removes the outer headers to obtain the inner Layer 2 packet. It then searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain.
7. VTEP 3 performs VLAN tags to the packets based on the outbound interface and encapsulation information and forwards the packets to Host 2.

Host 2 sends packets to Host 1 in the same manner.

3.2 Centralized VXLAN Gateway Deployment Using BGP EVPN

In centralized VXLAN gateway deployment using BGP EVPN, the control plane is responsible for VXLAN tunnel establishment and dynamic MAC address learning; the forwarding plane is responsible for intra-subnet known unicast packet forwarding, intra-subnet BUM (Broadcast&Unknown-unicast&Multicast) packet forwarding, and inter-subnet packet forwarding. This deployment mode is flexible because EVPN allows dynamic VTEP discovery and VXLAN tunnel establishment, and is therefore applicable to large-scale networks. If centralized VXLAN gateway deployment is needed, using this mode is recommended.

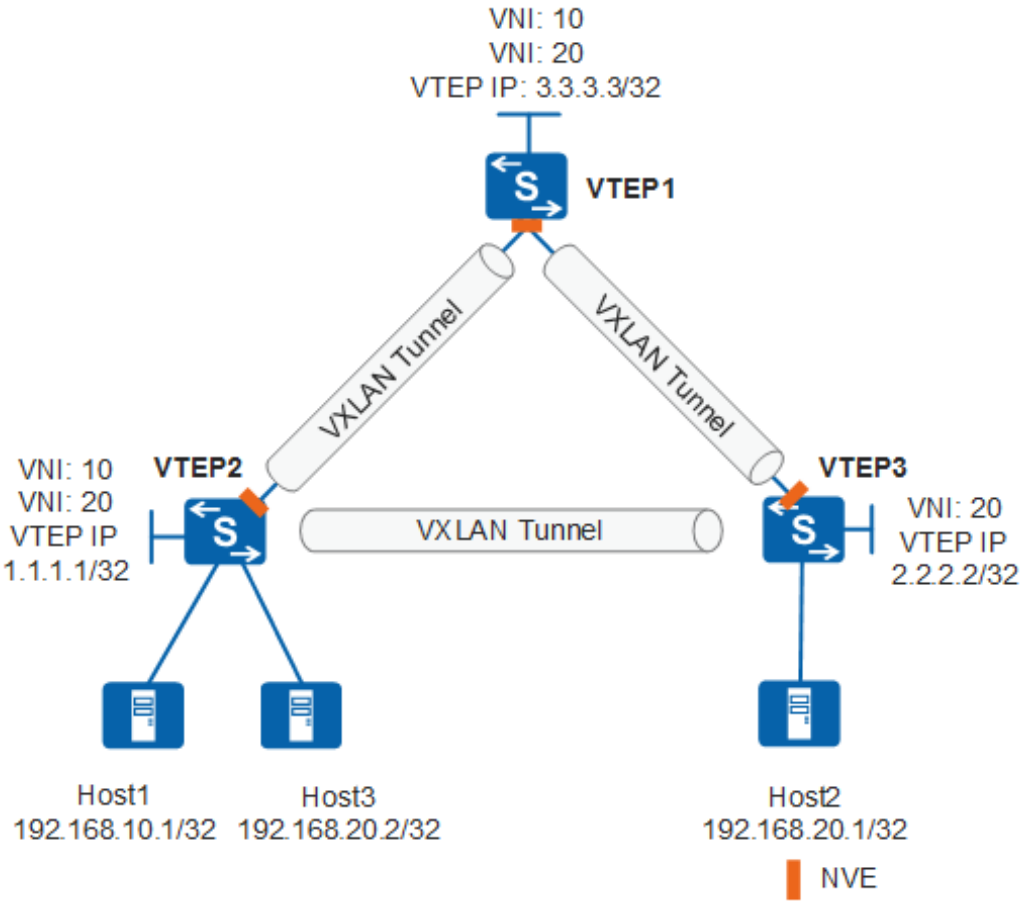
VXLAN Tunnel Establishment

A VXLAN tunnel is identified by a pair of VTEP IP addresses. During VXLAN tunnel establishment, the local and remote VTEPs attempt to obtain the IP addresses of each other. A VXLAN tunnel can be established if the IP addresses obtained are reachable at Layer 3. When BGP EVPN is used to dynamically establish a VXLAN tunnel, the local and remote VTEPs first establish a BGP EVPN peer relationship and then exchange BGP EVPN routes to transmit VNIs and VTEPs' IP addresses.

On the network shown in Figure 3-7, VTEP 2 connects to Host 1 and Host 3; VTEP 3 connects to Host 2; VTEP 1 functions as a Layer 3 gateway. To allow Host 3 and Host 2 to communicate, establish a VXLAN tunnel between VTEP 2 and VTEP 3. To allow Host 1 and Host 2 to communicate, establish a VXLAN tunnel between VTEP 2 and VTEP 1 and between VTEP 1 and VTEP 3. Although Host 1 and Host 3 both connect to VTEP 2, they

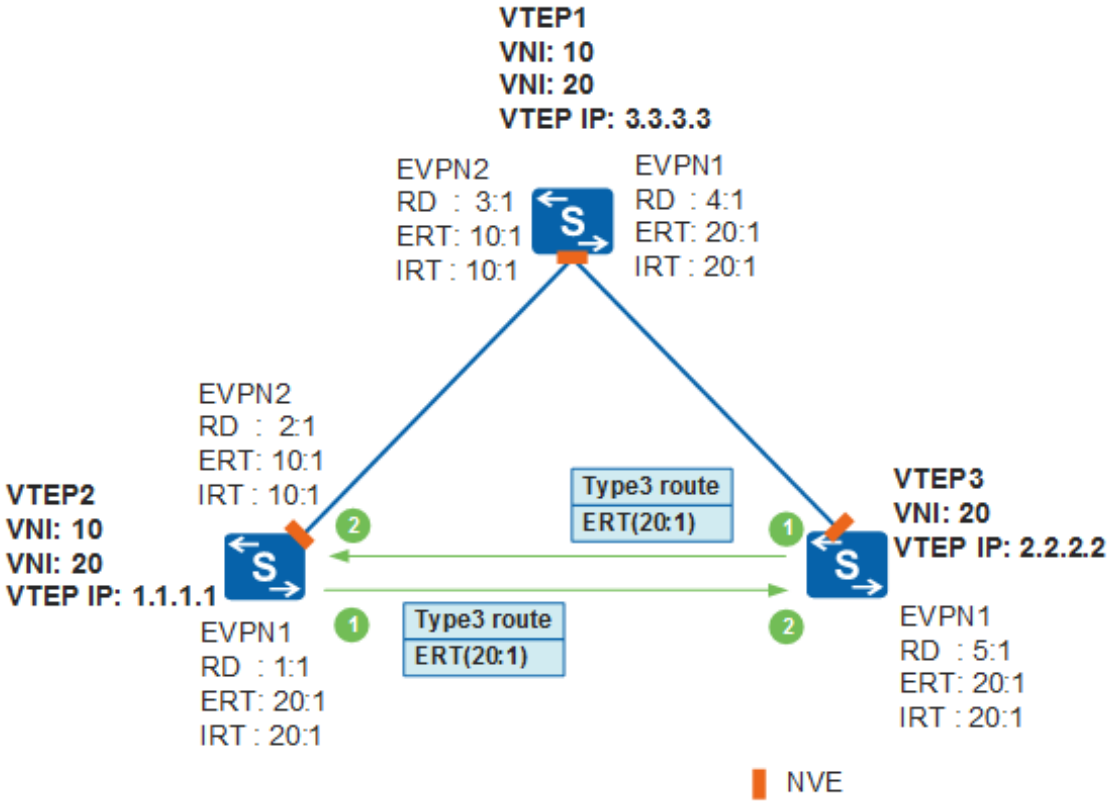
belong to different subnets and must communicate through the Layer 3 gateway (VTEP 1). Therefore, a VXLAN tunnel is also required between VTEP 2 and VTEP 1.

Figure 3-7 VXLAN tunnel networking



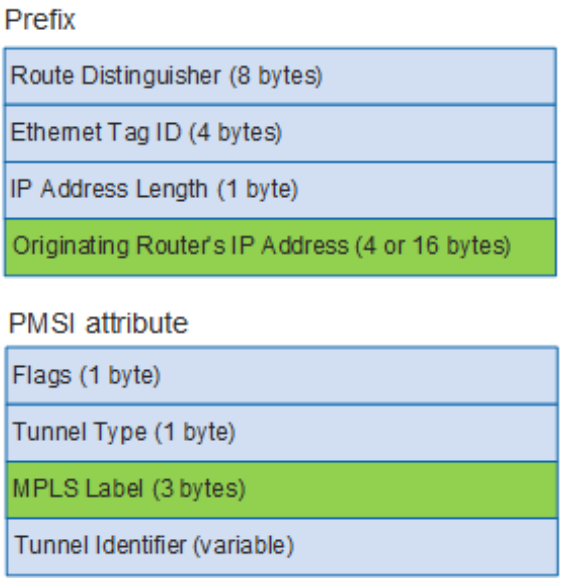
The following example illustrates how to use BGP EVPN to dynamically establish a VXLAN tunnel between VTEP 2 and VTEP 3.

Figure 3-8 Dynamic VXLAN tunnel establishment



1. VTEP 2 and VTEP 3 establish a BGP EVPN peer relationship. Then, Local EVPN instances are created on VTEP 2 and VTEP 3, and an RD, export VPN targets (ERT), and import VPN targets (IRT) are configured for the EVPN instance. Layer 2 broadcast domains are created and bound to VNIs and EVPN instances. After the local VTEP's IP address is configured on VTEP 2 and VTEP 3, they generate a BGP EVPN route and send it to each other. The BGP EVPN route carries the local EVPN instance's export VPN target and an inclusive multicast route (Type 3 route defined in BGP EVPN). Figure 3-9 shows the format of an inclusive multicast route, which comprises a prefix and a PMSI attribute. VTEP IP addresses are stored in the Originating Router's IP Address field in the inclusive multicast route prefix, and VNIs are stored in the MPLS Label field in the PMSI attribute.

Figure 3-9 Format of an inclusive multicast route



2. After VTEP 2 and VTEP 3 receive a BGP EVPN route from each other, they match the export VPN targets of the route against the import VPN targets of the local EVPN instance. If a match is found, the route is accepted. If no match is found, the route is discarded. If the route is accepted, VTEP 2/VTEP 3 obtains the remote VTEP's IP address and VNI carried in the route. If the remote VTEP's IP address is reachable at Layer 3, a VXLAN tunnel to the remote VTEP is established. If the remote VNI is the same as the local VNI, an ingress replication list is created for subsequent BUM packet forwarding.

The processes for dynamic VXLAN tunnel establishment using BGP EVPN between VTEP 2 and VTEP 1 and between VTEP 1 and VTEP 3 are the same.

NOTE

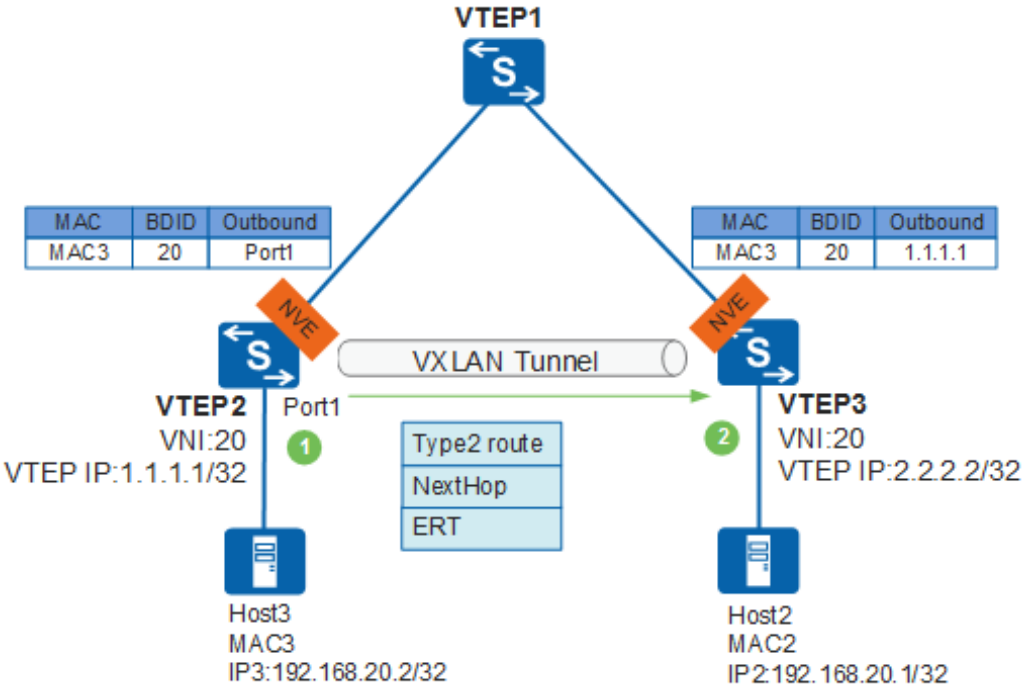
A VPN target is an extended community attribute of BGP for advertising VPN routes. An EVPN instance can have import and export VPN targets configured. The local EVPN instance's export VPN target must match the remote EVPN instance's import VPN target for EVPN route advertisement. If not, VXLAN tunnels cannot be dynamically established. If only one end can successfully accept the BGP EVPN route, this end can establish a VXLAN tunnel to the other end, but cannot exchange data packets with the other end. The other end drops packets after confirming that there is no VXLAN tunnel to the end that has sent these packets.

For details on VPN targets, see Concepts in "BGP/MPLS IP VPN Configuration" in the *S1720&S2700&S5700&S6720 Series Ethernet Switches Configuration Guide - VPN*.

Dynamic MAC Address Learning

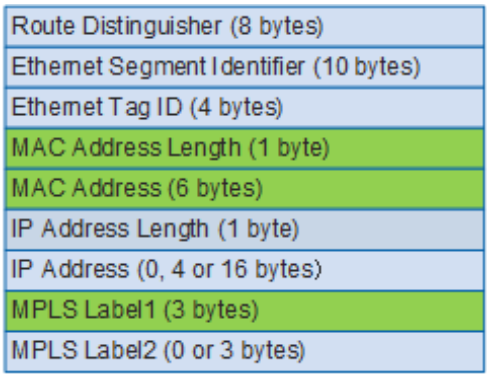
VXLAN supports dynamic MAC address learning to allow communication between tenants. MAC address entries are dynamically created and do not need to be manually maintained, greatly reducing maintenance workload. The following example illustrates dynamic MAC address learning for intra-subnet communication on the network shown in Figure 3-10.

Figure 3-10 Dynamic MAC address learning



1. When Host 3 communicates with VTEP 2 for the first time, VTEP 2 learns the mapping between Host 3's MAC address, BDID (Layer 2 broadcast domain ID), and inbound interface (Port1) that has received the dynamic ARP packet and generates a MAC address entry for Host 3. The MAC address entry's outbound interface is Port1. VTEP 2 generates and sends a BGP EVPN route based on the ARP entry of Host 3 to VTEP 3. The BGP EVPN route carries the local EVPN instance's export VPN targets, Next_Hop attribute, and a Type 2 route (MAC/IP route) defined in BGP EVPN. The Next_Hop attribute carries the local VTEP's IP address. The MAC Address Length and MAC Address fields identify Host 3's MAC address. The Layer 2 VNI is stored in the MPLS Label1 field. Figure 3-11 shows the format of a MAC/IP route.

Figure 3-11 MAC/IP route



2. After VTEP 3 receives a BGP EVPN route from VTEP 2, VTEP 3 matches the export VPN targets of the route against the import VPN targets of the local EVPN instance. If a match is found, the route is accepted. If no match is found, the route is discarded. If the route is accepted, VTEP 3 obtains the mapping between Host 3's MAC address, BDID, VTEP 2's VTEP IP address (Next_Hop attribute) and generates a MAC address entry for Host 3. Based on the next hop, the MAC address entry's outbound interface is iterated to the VXLAN tunnel destined for VTEP 2.

VTEP 2 learns the MAC address of Host 2 in the same process.

When Host 3 communicates with Host 2 for the first time, Host 3 sends an ARP request for Host 2's MAC address. The ARP request carries the destination MAC address being all Fs and destination IP address being IP2. By default, VTEP 2 broadcasts the ARP request onto the network segment after receiving it. To reduce broadcast packets, ARP broadcast suppression can be enabled on VTEP 2. In the case ARP broadcast suppression is enabled and VTEP 2 receives the ARP request, VTEP 2 checks whether it has Host 2's MAC address based on the destination IP address of the ARP request. If VTEP 2 has Host 2's MAC address, it replaces the destination MAC address of the ARP request with Host 2's MAC address and unicasts the ARP request to VTEP 3 through the VXLAN tunnel. Upon receipt, VTEP 3 forwards the ARP request to Host 2, which then learns Host 3's MAC address and responds with an ARP reply in unicast mode. After Host 3 receives the ARP reply, it learns Host 2's MAC address. So far, Host 2 and Host 3 have learned the MAC address of each other, and will subsequently communicate with each other in unicast mode.

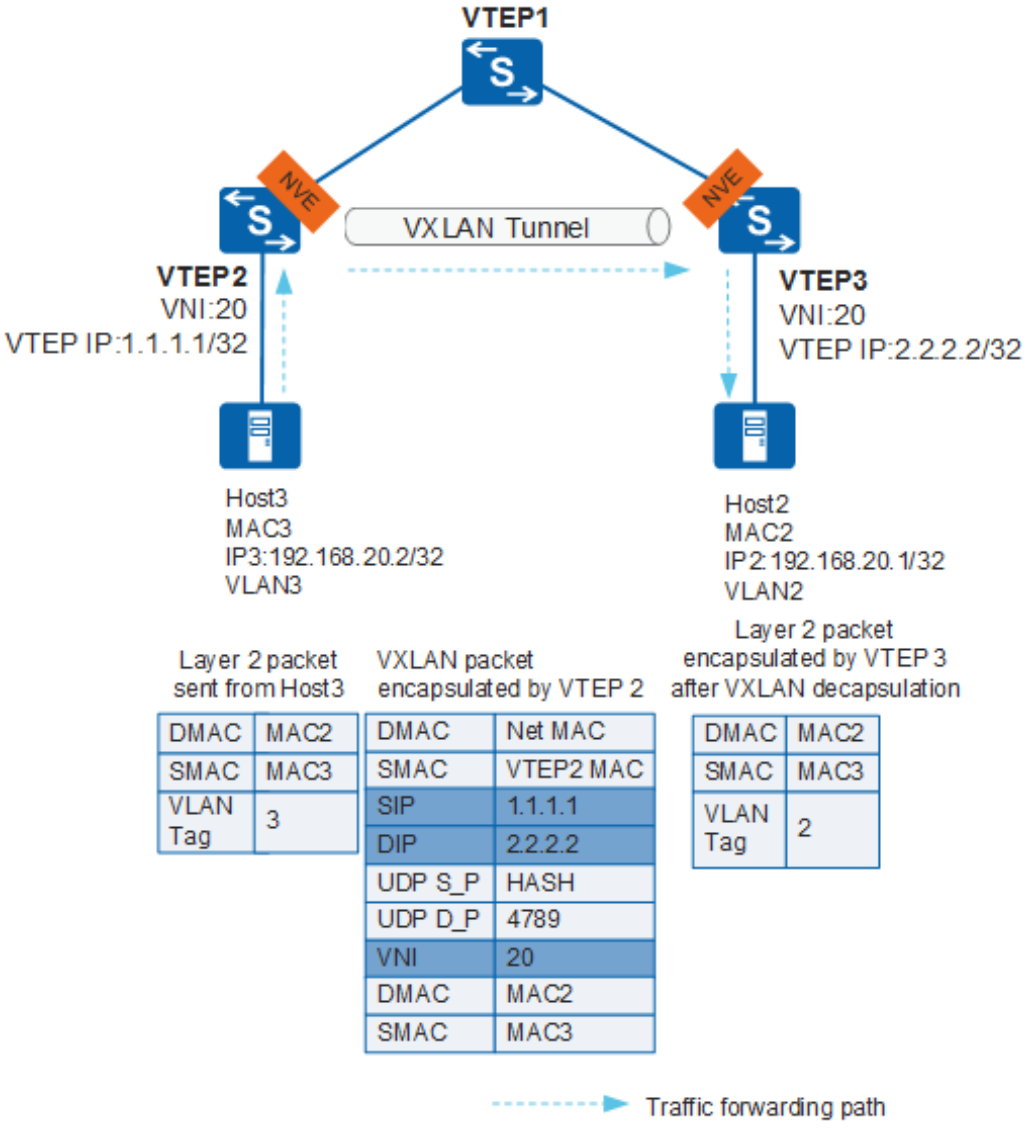
NOTE

- Dynamic MAC address learning is required only between hosts and Layer 3 gateways in inter-subnet communication scenarios. The process is the same as that for intra-subnet communication.
- VTEP nodes can learn the MAC addresses of hosts during data forwarding, if this capability is enabled. If VXLAN tunnels are established using BGP EVPN, VTEP nodes can dynamically learn the MAC addresses of hosts through BGP EVPN routes, rather than data forwarding.

Intra-Subnet Known Unicast Packet Forwarding

Intra-subnet known unicast packets are forwarded only through Layer 2 VXLAN gateways and are unknown to Layer 3 VXLAN gateways. Figure 3-12 shows the intra-subnet known unicast packet forwarding process.

Figure 3-12 Intra-subnet known unicast packet forwarding



1. After VTEP 2 receives Host 3's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information and searches for the outbound interface and encapsulation information in the BD.
2. VTEP 2 performs VXLAN encapsulation based on the encapsulation information obtained and forwards the packets through the outbound interface obtained.
3. Upon receipt of the VXLAN packet, VTEP 3 verifies the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 3 obtains the Layer 2 BD based on the VNI and performs VXLAN decapsulation to obtain the inner Layer 2 packet.
4. VTEP 3 obtains the destination MAC address of the inner Layer 2 packet, performs VLAN tags to the packets based on the outbound interface and encapsulation information in the local MAC address table, and forwards the packets to Host 2.

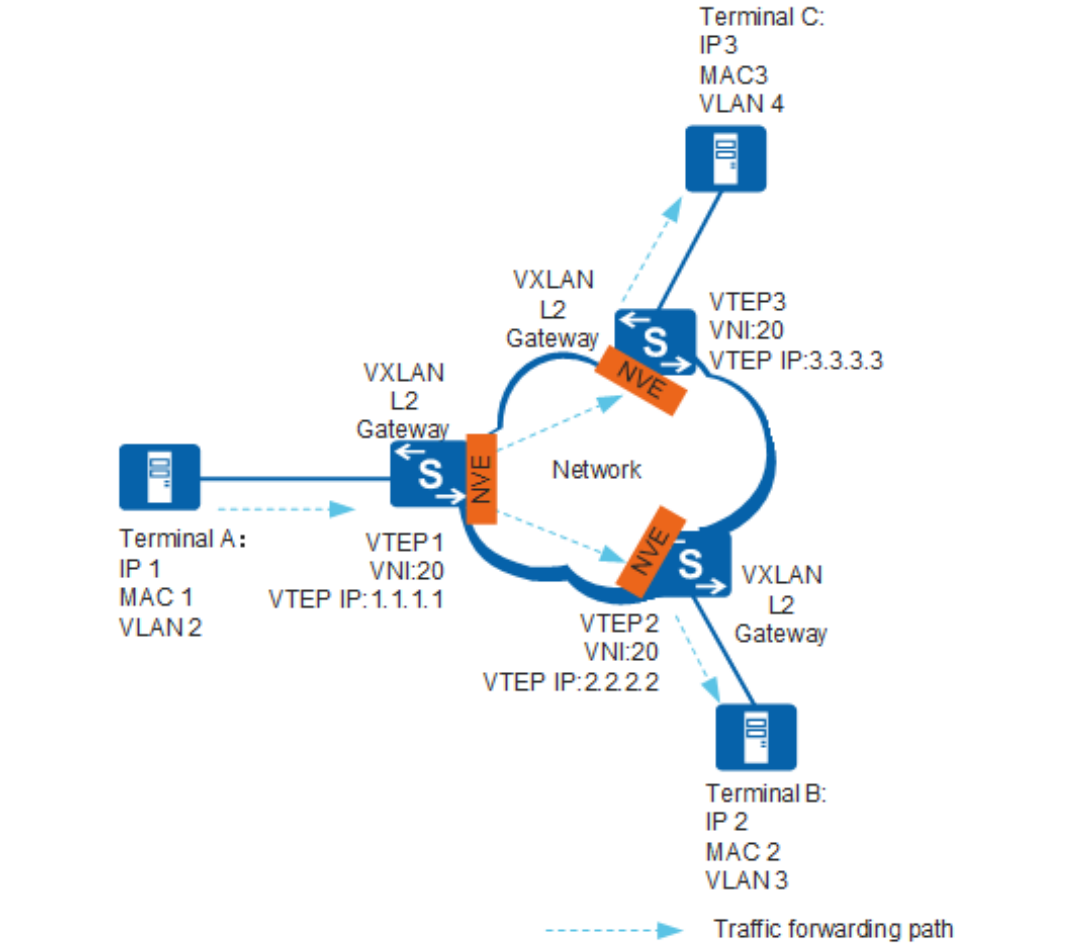
Host 2 sends packets to Host 3 in the same manner.

Intra-Subnet BUM Packet Forwarding

Intra-subnet BUM packet forwarding is completed between Layer 2 VXLAN gateways. Layer 3 VXLAN gateways do not need to be unaware of the process. Intra-subnet BUM packets can be forwarded in ingress replication mode.

In ingress replication mode, after a BUM packet enters a VXLAN tunnel, the ingress VTEP performs VXLAN encapsulation based on the ingress replication list and sends the packet to all the egress VTEPs in the list. When the BUM packet leaves the VXLAN tunnel, the egress VTEPs decapsulate the BUM packet. Figure 3-13 shows the forwarding process of a BUM packet in ingress replication mode.

Figure 3-13 Forwarding process of an intra-subnet BUM packet in ingress replication mode



Layer 2 packet sent from Terminal A

DMAC	All Fs
SMAC	MAC1
VLAN Tag	2

VXLAN packet encapsulated by VTEP 1

VTEP1—>VTEP2

DMAC	Net MAC
SMAC	VTEP1 MAC1
SIP	1.1.1.1
DIP	2.2.2.2
UDP S_P	HASH
UDP D_P	4789
VNI	20
DMAC	All Fs
SMAC	MAC1

VTEP1—>VTEP3

DMAC	Net MAC
SMAC	VTEP1 MAC1
SIP	1.1.1.1
DIP	3.3.3.3
UDP S_P	HASH
UDP D_P	4789
VNI	20
DMAC	All Fs
SMAC	MAC1

Layer 2 packet encapsulated by VTEP 2/ VTEP 3 after VXLAN decapsulation

DMAC	All Fs
SMAC	MAC1
VLAN Tag	3

DMAC	All Fs
SMAC	MAC1
VLAN Tag	4

1. After VTEP 1 receives Terminal A's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information.
2. VTEP 1 obtains the ingress replication list for the VNI, replicates packets based on the list, and performs VXLAN encapsulation by adding outer headers. VTEP 1 then forwards the VXLAN packet through the outbound interface.

3. Upon receipt of the VXLAN packet, VTEP 2 and VTEP 3 verify the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 2/VTEP 3 obtains the Layer 2 BD based on the VNI and performs VXLAN decapsulation to obtain the inner Layer 2 packet.
4. VTEP 2/VTEP 3 checks the destination MAC address of the inner Layer 2 packet and finds it a BUM MAC address. Therefore, VTEP 2/VTEP 3 broadcasts the packet onto the network connected to the terminals (not the VXLAN tunnel side) in the Layer 2 broadcast domain. Specifically, VTEP 2/VTEP 3 finds the outbound interfaces and encapsulation information not related to the VXLAN tunnel, performs VLAN tags to the packet, and forwards the packet to Terminal B/Terminal C.

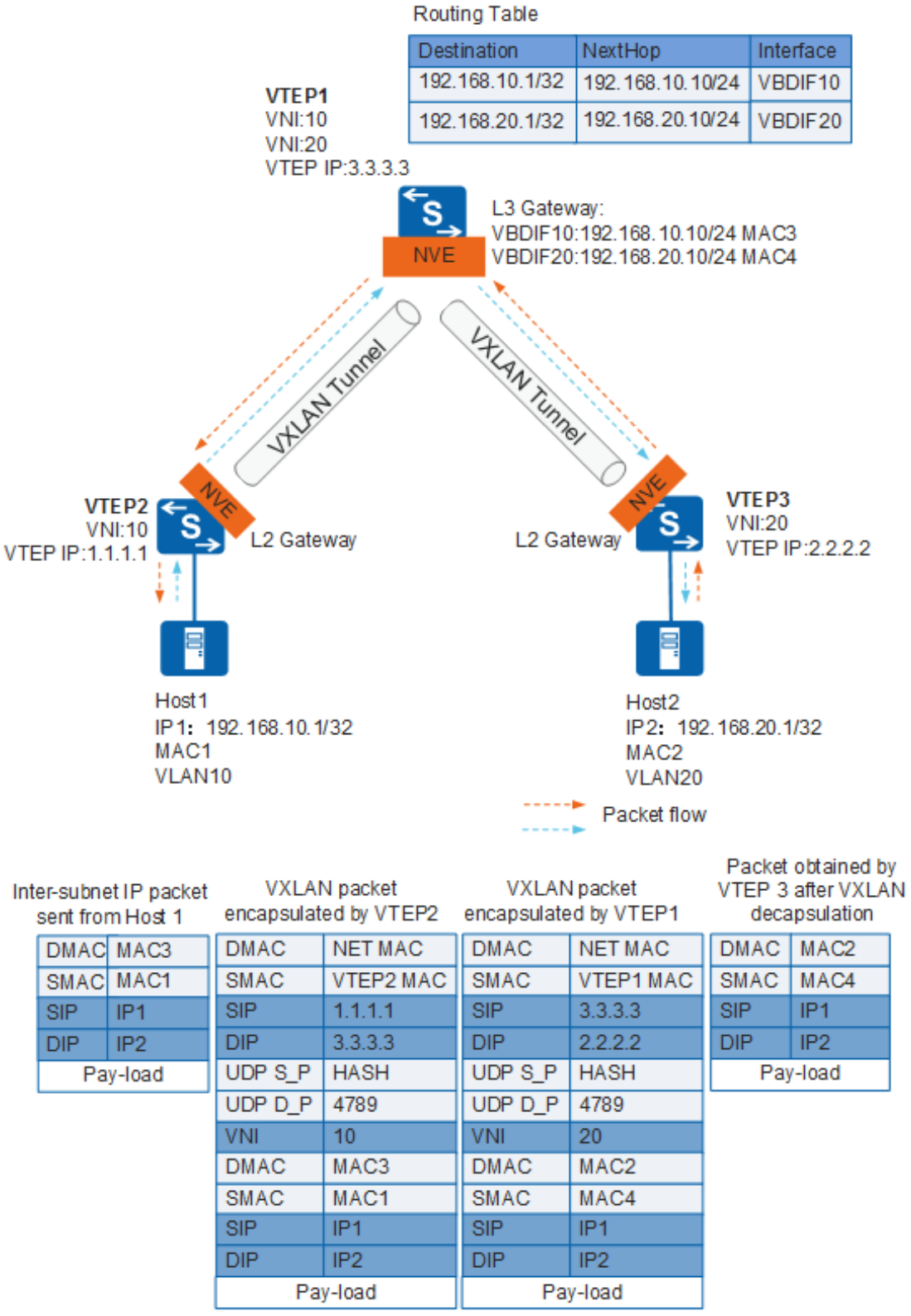
NOTE

Terminal B/Terminal C responds to Terminal A in the same process as intra-subnet known unicast packet forwarding.

Inter-Subnet Packet Forwarding

Inter-subnet packets must be forwarded through a Layer 3 gateway. Figure 3-14 shows the inter-subnet packet forwarding process.

Figure 3-14 Inter-subnet packet forwarding



1. After VTEP 2 receives Host 1's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information and searches for the outbound interface and encapsulation information in the BD.
2. VTEP 2 performs VXLAN encapsulation based on the outbound interface and encapsulation information and forwards the packets to VTEP 1.
3. After VTEP 1 receives the VXLAN packet, it decapsulates the packet and finds that the destination MAC address of the inner packet is the MAC address (MAC3) of the Layer 3 gateway interface (VBDIF10) so that the packet must be forwarded at Layer 3.
4. VTEP 1 removes the inner Ethernet header, parses the destination IP address, and searches the routing table for a next hop address. VTEP 1 then searches the ARP table based on the next hop address to obtain the destination MAC address, VXLAN tunnel's outbound interface, and VNI.
5. VTEP 1 performs VXLAN encapsulation on the inner packet again and forwards the VXLAN packet to VTEP 3, with the source MAC address in the inner Ethernet header being the MAC address (MAC4) of the Layer 3 gateway interface (VBDIF20).
6. Upon receipt of the VXLAN packet, VTEP 3 verifies the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 3 then obtains the Layer 2 broadcast domain based on the VNI and removes the outer headers to obtain the inner Layer 2 packet. It then searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain.
7. VTEP 3 performs VLAN tags to the packets based on the outbound interface and encapsulation information and forwards the packets to Host 2.

Host 2 sends packets to Host 1 in the same manner.

3.3 Distributed VXLAN Gateway Deployment Using BGP EVPN

In distributed VXLAN gateway deployment using BGP EVPN, the control plane is responsible for VXLAN tunnel establishment and dynamic MAC address learning; the forwarding plane is responsible for intra-subnet known unicast packet forwarding, intra-subnet BUM packet forwarding, and inter-subnet packet forwarding. This mode supports IP route advertisement, MAC address advertisement, and ARP advertisement, and ARP broadcast suppression can be directly enabled. For details on the functions, see 2.5 BGP EVPN Basic Principles.

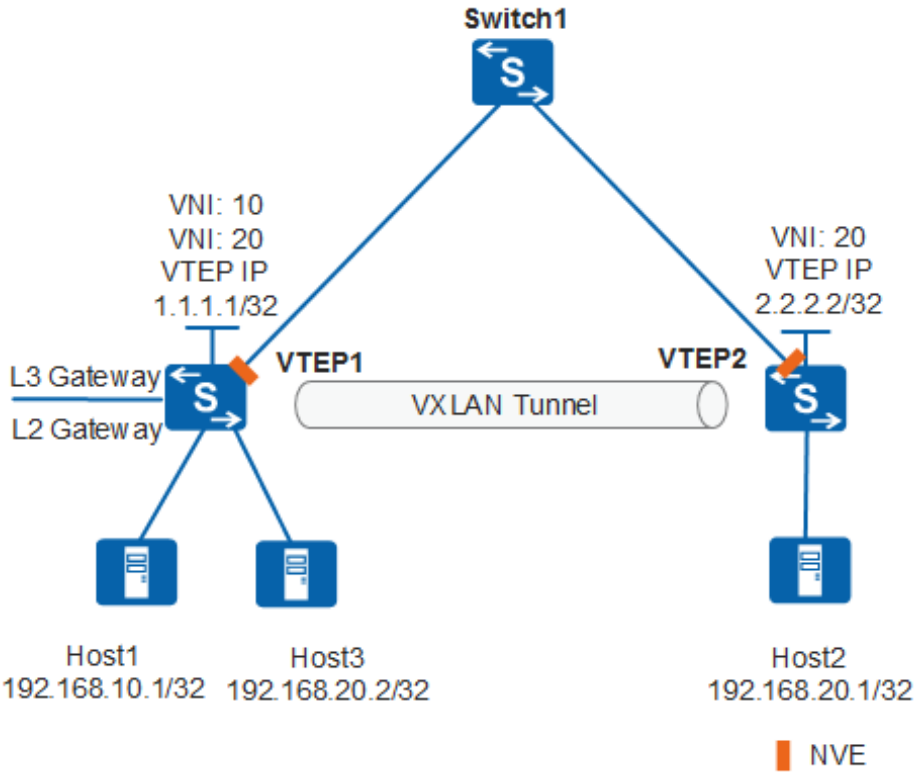
VXLAN Tunnel Establishment

A VXLAN tunnel is identified by a pair of VTEP IP addresses. During VXLAN tunnel establishment, the local and remote VTEPs attempt to obtain the IP addresses of each other. A VXLAN tunnel can be established if the IP addresses obtained are reachable at Layer 3. When BGP EVPN is used to dynamically establish a VXLAN tunnel, the local and remote VTEPs first establish a BGP EVPN peer relationship and then exchange BGP EVPN routes to transmit VNIs and VTEPs' IP addresses.

In distributed VXLAN gateway scenarios, VTEP nodes function as both Layer 2 and Layer 3 VXLAN gateways. Switch 1 are unaware of the VXLAN tunnels and only forward VXLAN packets between different VTEP 1 and VTEP 2. On the network shown in Figure 3-15, a VXLAN tunnel is established between VTEP 1 and VTEP 2 for Host 1 and Host 2 or Host 3 and Host 2 to communicate. Host 1 and Host 3 both connect to VTEP 1, and therefore

communication between Host 1 and Host 3 is allowed through VTEP 1, but not the VXLAN tunnel.

Figure 3-15 VXLAN tunnel networking

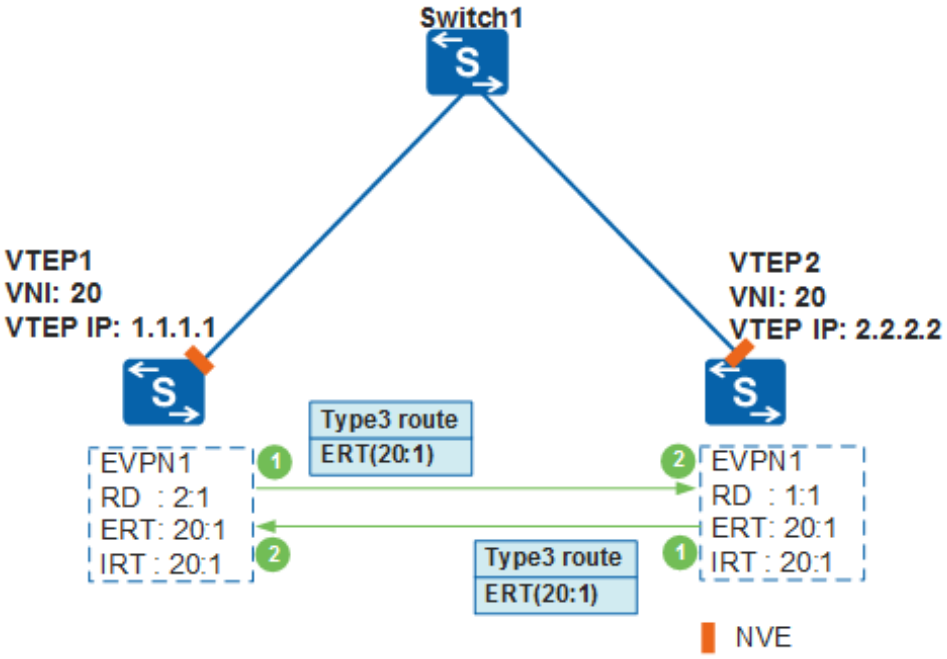


In distributed VXLAN gateway scenarios, VXLAN tunnels can be dynamically established using BGP EVPN for intra-subnet and inter-subnet communication.

Intra-subnet communication

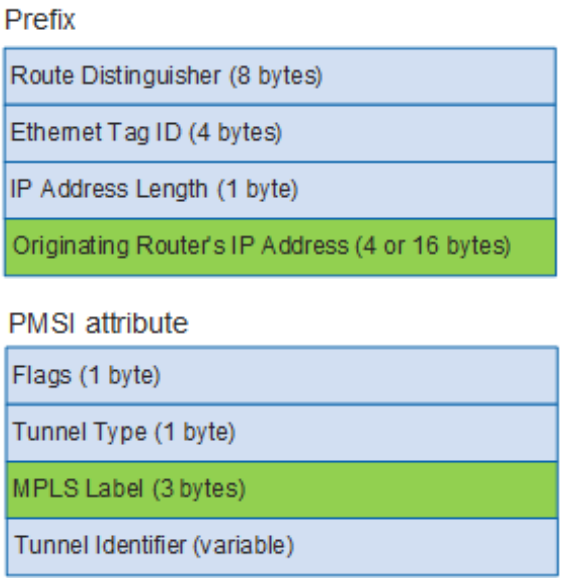
On the network shown in Figure 3-16, intra-subnet communication between Host 2 and Host 3 requires only Layer 2 forwarding. The process for establishing a VXLAN tunnel using BGP EVPN is as follows:

Figure 3-16 Dynamic VXLAN tunnel establishment 1



1. VTEP 1 and VTEP 2 establish a BGP EVPN peer relationship. Then, local EVPN instances are created on VTEP 2 and VTEP 3, and an RD, export VPN targets (ERT), and import VPN targets (IRT) are configured for the EVPN instance. Layer 2 broadcast domains are created and bound to VNIs and EVPN instances. After the local VTEP's IP address is configured on VTEP 1 and VTEP 2, they generate a BGP EVPN route and send it to each other. The BGP EVPN route carries the local EVPN instance's export VPN target and an inclusive multicast route (Type 3 route defined in BGP EVPN). Figure 3-17 shows the format of an inclusive multicast route, which comprises a prefix and a PMSI attribute. VTEP IP addresses are stored in the Originating Router's IP Address field in the inclusive multicast route prefix, and Layer 2 VNIs are stored in the MPLS Label field in the PMSI attribute.

Figure 3-17 Format of an inclusive multicast route



2. After VTEP 1 and VTEP 2 receive a BGP EVPN route from each other, they match the export VPN targets of the route against the import VPN targets of the local EVPN instance. If a match is found, the route is accepted. If no match is found, the route is discarded. If the route is accepted, VTEP 1/VTEP 2 obtains the remote VTEP's IP address and Layer 2 VNI carried in the route. If the remote VTEP's IP address is reachable at Layer 3, a VXLAN tunnel to the remote VTEP is established. If the remote Layer 2 VNI is the same as the local Layer 2 VNI, an ingress replication list is created for subsequent BUM (Broadcast&Unknown-unicast&Multicast) packet forwarding.

NOTE

A VPN target is an extended community attribute of BGP. An EVPN instance can have import and export VPN targets configured. The local EVPN instance's export VPN target must match the remote EVPN instance's import VPN target for EVPN route advertisement. If not, VXLAN tunnels cannot be dynamically established. If only one end can successfully accept the BGP EVPN route, this end can establish a VXLAN tunnel to the other end, but cannot exchange data packets with the other end. The other end drops packets after confirming that there is no VXLAN tunnel to the end that has sent these packets.

For details on VPN targets, see Concepts in "BGP/MPLS IP VPN Configuration" in the *S1720&S2700&S5700&S6720 Series Ethernet Switches Configuration Guide - VPN*.

Inter-subnet communication

Inter-subnet communication between Host 1 and Host 2 requires Layer 3 forwarding. When VXLAN tunnels are established using BGP EVPN, VTEP 1 and VTEP 2 must advertise the host IP routes. Generally, 32-bit host IP routes are advertised. Because different leaf nodes may connect to the same network segment on VXLANs, the network segment routes advertised by these leaf nodes may conflict. This conflict may cause host unreachability of some leaf nodes. Leaf nodes can advertise network segment routes in the following scenarios:

- The network segment that a leaf node connects is unique on a VXLAN, and a large number of specific host routes are available. In this case, the network segment routes to

which the host IP routes belong can be advertised so that leaf nodes do not have to store all these routes.

- When hosts on a VXLAN need to access external networks, leaf nodes can advertise routes destined for external networks onto the VXLAN to allow other leaf nodes to learn the routes.

Before establishing a VXLAN tunnel, perform the following configurations on VTEP 1 and VTEP 2.

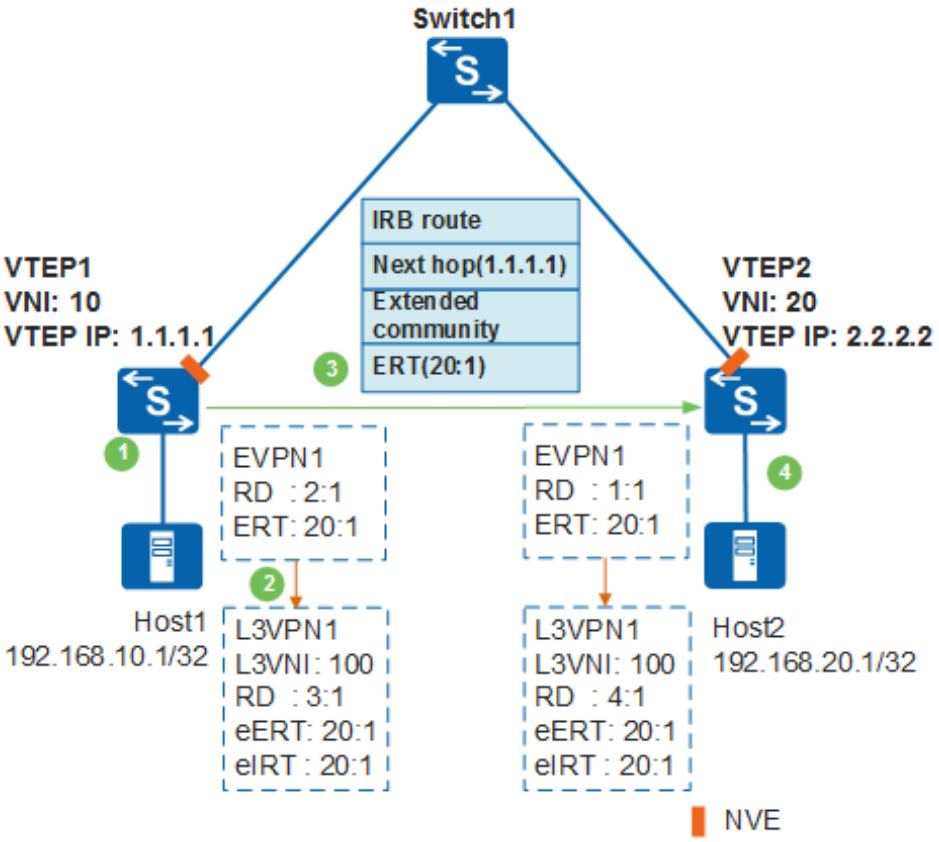
Configuration Task	Function
Create a Layer 2 BD and associate a Layer 2 VNI to the Layer 2 BD.	A BD functions as a VXLAN network entity to transmit VXLAN data packets.
Establish a BGP EVPN peer relationship between VTEP 1 and VTEP 2.	This configuration is used to exchange BGP EVPN routes.
Configure an EVPN instance and bound to a Layer 2 BD, and configure an RD, export VPN target (ERT), and import VPN target (IRT) for the EVPN instance.	This configuration is used to generate BGP EVPN routes.
Configure L3VPN instances for tenants and bind the L3VPN instances to the VBDIF interfaces of the Layer 2 BD.	This configuration is used to differentiate and isolate IP routing tables of different tenants.
Specify a Layer 3 VNI for an L3VPN instance.	This configuration allows the leaf nodes to determine the L3VPN routing table for forwarding data packets.
Configure export VPN targets (eERT) from an L3VPN instance to an EVPN instance and import VPN targets (eIRT) from an EVPN instance to an L3VPN instance.	This configuration controls advertisement and reception of BGP EVPN routes between the local L3VPN instance and remote EVPN instance.
Configure the type of route to be advertised between VTEP 1 and VTEP 2.	This configuration is used to advertise IP routes between Host 1 and Host 2. Two types of routes are available, IRB and IP prefix routes, which can be selected as needed. • IRB routes advertise only 32-bit host IP routes. IRB routes carry ARP routes, and therefore ARP broadcast suppression can be enabled on leaf nodes after IRB routes are advertised. This also facilitates VM migration. For details, see 2.5 BGP EVPN Basic Principles. If only 32-bit host IP route advertisement is needed, advertising IRB routes is recommended. • IP prefix routes can advertise both 32-bit host IP routes and network segment routes. However, before IP prefix routes advertise 32-bit host IP routes, direct routes to the host IP addresses must be generated. This will affect VM

Configuration Task	Function
	migration. If only 32-bit host IP route advertisement is needed, advertising IP prefix routes is not recommended. Advertise IP prefix routes only when network segment route advertisement is needed.

Dynamic VXLAN tunnel establishment varies depending on how host IP routes are advertised.

- Host IP routes are advertised through IRB routes. (Figure 3-18 shows the process.)

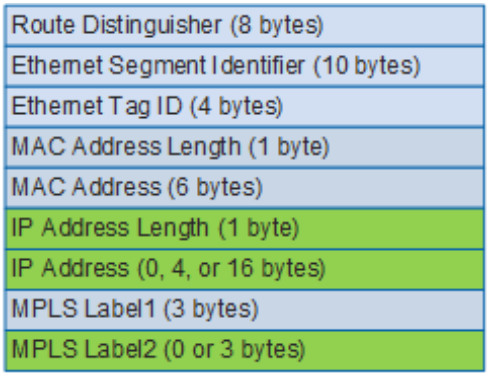
Figure 3-18 Dynamic VXLAN tunnel establishment 2



a. When Host 1 communicates with VTEP 1 for the first time, VTEP 1 learns the ARP entry of Host 1 after receiving dynamic ARP packets. VTEP 1 then finds the L3VPN instance bound to the VBDIF interface of the Layer 2 BD where Host 1 resides, and obtains the Layer 3 VNI associated with the L3VPN instance. The EVPN instance of VTEP 1 then generates an IRB route based on the information obtained. Figure 3-19 shows the IRB route. The host IP address is stored in the IP

Address Length and IP Address fields; the Layer 3 VNI is stored in the MPLS Label2 field.

Figure 3-19 IRB route



b. The EVPN instance of VTEP 1 obtains Host 1's IP address and Layer 3 VNI from the IRB route and sends it to the local L3VPN instance. The L3VPN instance then stores Host 1's IP route in the routing table. Figure 3-20 shows the host IP route.

Figure 3-20 Local host IP route

Destination	L3 VNI	Next Hop	Outbound Interface
192.168.10.1/32	100	Gateway address	VBDIF

c. VTEP 1 generates and sends a BGP EVPN route to VTEP 2. The BGP EVPN route carries the local EVPN instance's export VPN targets (ERT), extended community attribute, Next_Hop attribute, and the IRB route. The extended community attribute carries the tunnel type (VXLAN tunnel) and local VTEP MAC address; the Next_Hop attribute carries the local VTEP IP address.

d. After VTEP 2 receives the BGP EVPN route from VTEP 1, VTEP 2 processes the route as follows:

- Matches the ERT of the route against the import VPN targets (IRT) of the local EVPN instance. If a match is found, the route is accepted. After the EVPN instance obtains IRB routes, it can extract ARP routes from the IRB routes to implement ARP advertisement.
- Matches the ERT of the route against the import VPN targets (eIRT) of the local L3VPN instance. If a match is found, the route is accepted. The L3VPN instance obtains the IRB route, extracts Host 1's IP address and Layer 3 VNI, stores Host 1's IP route in the routing table. Based on the next hop, the IP route's outbound interface is iterated to the VXLAN tunnel destined for Leaf1. Figure 3-21 shows the host route.

NOTE

Only when the ERT in a BGP EVPN route is different from the local EVPN instance's IRT and local L3VPN instance's eIRT, the route is discarded.

Figure 3-21 Remote host IP route

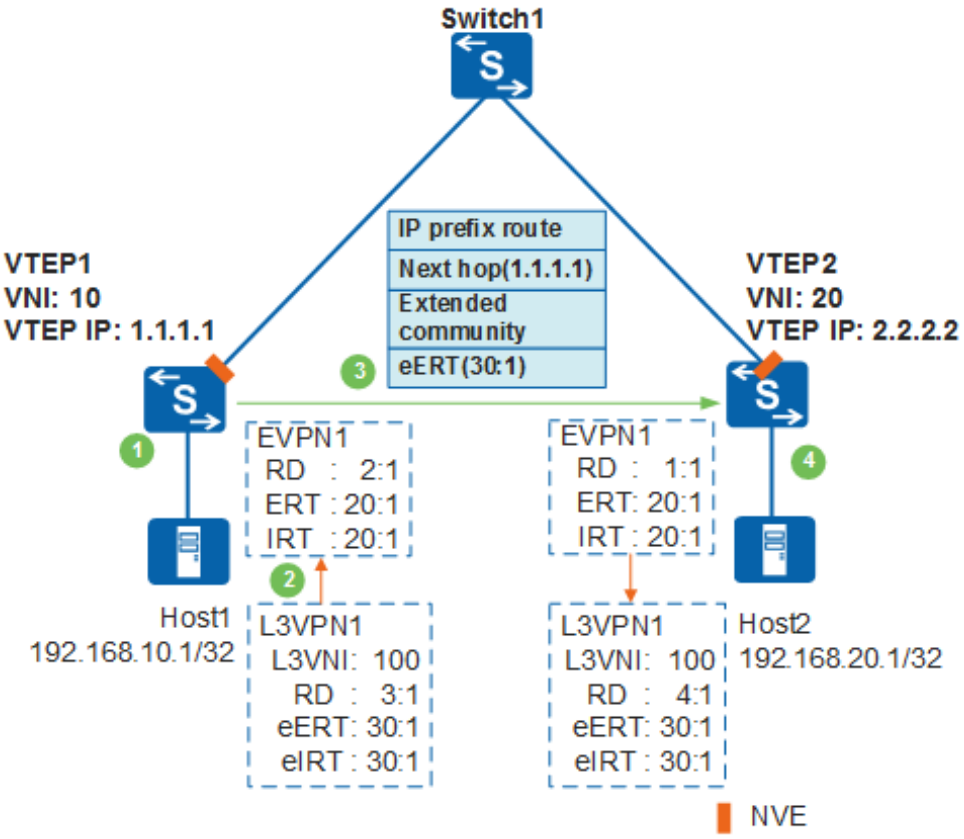
Destination	L3 VNI	Next Hop	Outbound Interface
192.168.10.1/24	100	1.1.1.1	VXLAN tunnel

- If the route is accepted by the EVPN instance or L3VPN instance, VTEP 2 obtains VTEP 1's VTEP IP address from the Next_Hop attribute. If the VTEP IP address is reachable at Layer 3, a VXLAN tunnel to VTEP 1 is established.

VTEP 1 establishes a VXLAN tunnel to VTEP 2 in the same process.

- Host IP routes are advertised through IP prefix routes. Figure 3-22 shows the process.

Figure 3-22 Dynamic VXLAN tunnel establishment 3



a. VTEP 1 generates a direct route to Host 1's IP address. Then, VTEP 1 has an L3VPN instance configured to import the direct route, so that Host 1's IP route is saved to the routing table of the L3VPN instance and the Layer 3 VNI associated with the L3VPN instance is added. Figure 3-23 shows the host IP route.

Figure 3-23 Local host IP route

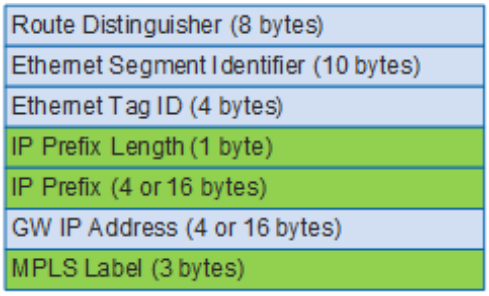
Destination	L3 VNI	Next Hop	Outbound Interface
192.168.10.1/32	100	Gateway address	VBDIF

NOTE

If network segment route advertisement is required, use a dynamic routing protocol, such as OSPF. Then, configure an L3VPN instance to import the routes of the dynamic routing protocol.

b. If VTEP 1 is configured to advertise IP routes in the L3VPN instance to the EVPN instance, VTEP 1 advertise Host 1's IP routes in the L3VPN instance to the EVPN instance. The EVPN instance then generates IP prefix routes. Figure 3-24 shows the IP prefix route. The host IP address is stored in the IP Prefix Length and IP Prefix fields; the Layer 3 VNI is stored in the MPLS Label field.

Figure 3-24 IP prefix route



c. VTEP 1 generates and sends a BGP EVPN route to VTEP 2. The BGP EVPN route carries the local L3VPN instance's export VPN targets (eERT), extended community attribute, Next_Hop attribute, and the IP prefix route. The extended community attribute carries the tunnel type (VXLAN tunnel) and local VTEP MAC address; the Next_Hop attribute carries the local VTEP IP address.

d. After VTEP 2 receives the BGP EVPN route from VTEP 1, VTEP 2 processes the route as follows:

- Matches the eERT of the route against the import VPN targets (eIRT) of the local L3VPN instance. If a match is found, the route is accepted. If no match is found, the route is discarded. The L3VPN instance obtains the IP prefix route, extracts Host 1's IP address and Layer 3 VNI, stores Host 1's IP route in the routing table, and sets the next hop's iterated outbound interface to the VXLAN tunnel interface. Figure 3-25 shows the host route.

Figure 3-25 Remote host IP route

Destination	L3 VNI	Next Hop	Outbound Interface
192.168.10.1/24	100	1.1.1.1	VXLAN tunnel

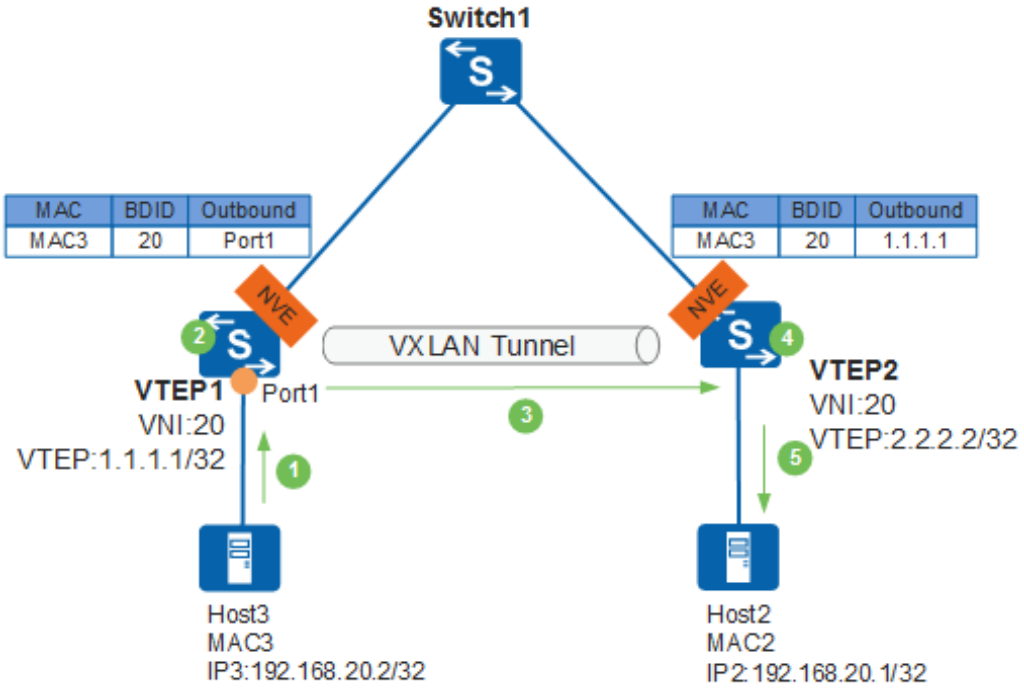
- If the route is accepted by the L3VPN instance, VTEP 2 obtains VTEP 1's VTEP IP address from the Next_Hop attribute. If the VTEP IP address is reachable at Layer 3, a VXLAN tunnel to VTEP 1 is established.

VTEP 1 establishes a VXLAN tunnel to VTEP 2 in the same process.

Dynamic MAC Address Learning

VXLAN supports dynamic MAC address learning to allow communication between tenants. MAC address entries are dynamically created and do not need to be manually maintained, greatly reducing maintenance workload. In distributed VXLAN gateway scenarios, inter-subnet communication requires Layer 3 forwarding; MAC address learning is implemented using ARP between the local host and gateway. The following example illustrates dynamic MAC address learning for intra-subnet communication on the network shown in Figure 3-26.

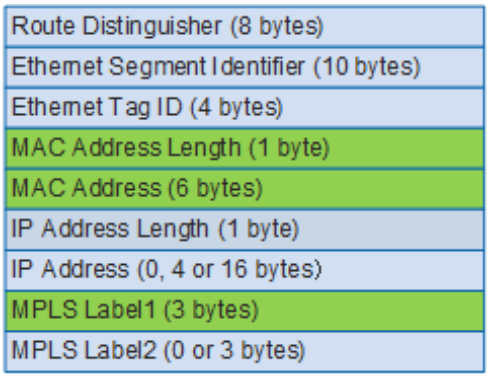
Figure 3-26 Dynamic MAC address learning



1. When Host 3 communicates with VTEP 1 for the first time, VTEP 1 learns the mapping between Host 3's MAC address, BDID (Layer 2 broadcast domain ID), and inbound interface (Port1) that has received the dynamic ARP packet and and generates a MAC address entry for Host 3. The MAC address entry's outbound interface is Port1. VTEP 1 generates and sends a BGP EVPN route based on the ARP entry of Host 3 to VTEP 2.

The BGP EVPN route carries the local EVPN instance's export VPN targets, Next_Hop attribute, and a Type 2 route (MAC/IP route) defined in BGP EVPN. The Next_Hop attribute carries the local VTEP's IP address. The MAC Address Length and MAC Address fields identify Host 3's MAC address. The Layer 2 VNI is stored in the MPLS Label1 field. Figure 3-27 shows the format of a MAC/IP route.

Figure 3-27 MAC/IP route



2. After VTEP 2 receives a BGP EVPN route from VTEP 1, VTEP 2 matches the export VPN targets of the route against the import VPN targets of the local EVPN instance. If a match is found, the route is accepted. If no match is found, the route is discarded. If the route is accepted, VTEP 2 obtains the mapping between Host 3's MAC address, BDID, VTEP 1's VTEP IP address (Next_Hop attribute) and generates a MAC address entry for Host 3. Based on the next hop, the MAC address entry's outbound interface is iterated to the VXLAN tunnel destined for VTEP 1.

VTEP 1 learns the MAC route of Host 2 in the same process.

When Host 3 communicates with Host 2 for the first time, Host 3 sends an ARP request for Host 2's MAC address. The ARP request carries the destination MAC address being all Fs and destination IP address being IP2. By default, VTEP 1 broadcasts the ARP request onto the network segment after receiving it. To reduce broadcast packets, ARP broadcast suppression can be enabled on VTEP 1. In the case ARP broadcast suppression is enabled and VTEP 1 receives the ARP request, VTEP 1 checks whether it has Host 2's MAC address based on the destination IP address of the ARP request. If VTEP 1 has Host 2's MAC address, it replaces the destination MAC address of the ARP request with Host 2's MAC address and unicasts the ARP request to VTEP 2 through the VXLAN tunnel. Upon receipt, VTEP 2 forwards the ARP request to Host 2, which then learns Host 3's MAC address and responds with an ARP reply in unicast mode. After Host 3 receives the ARP reply, it learns Host 2's MAC address. So far, Host 2 and Host 3 have learned the MAC address of each other, and will subsequently communicate with each other in unicast mode.

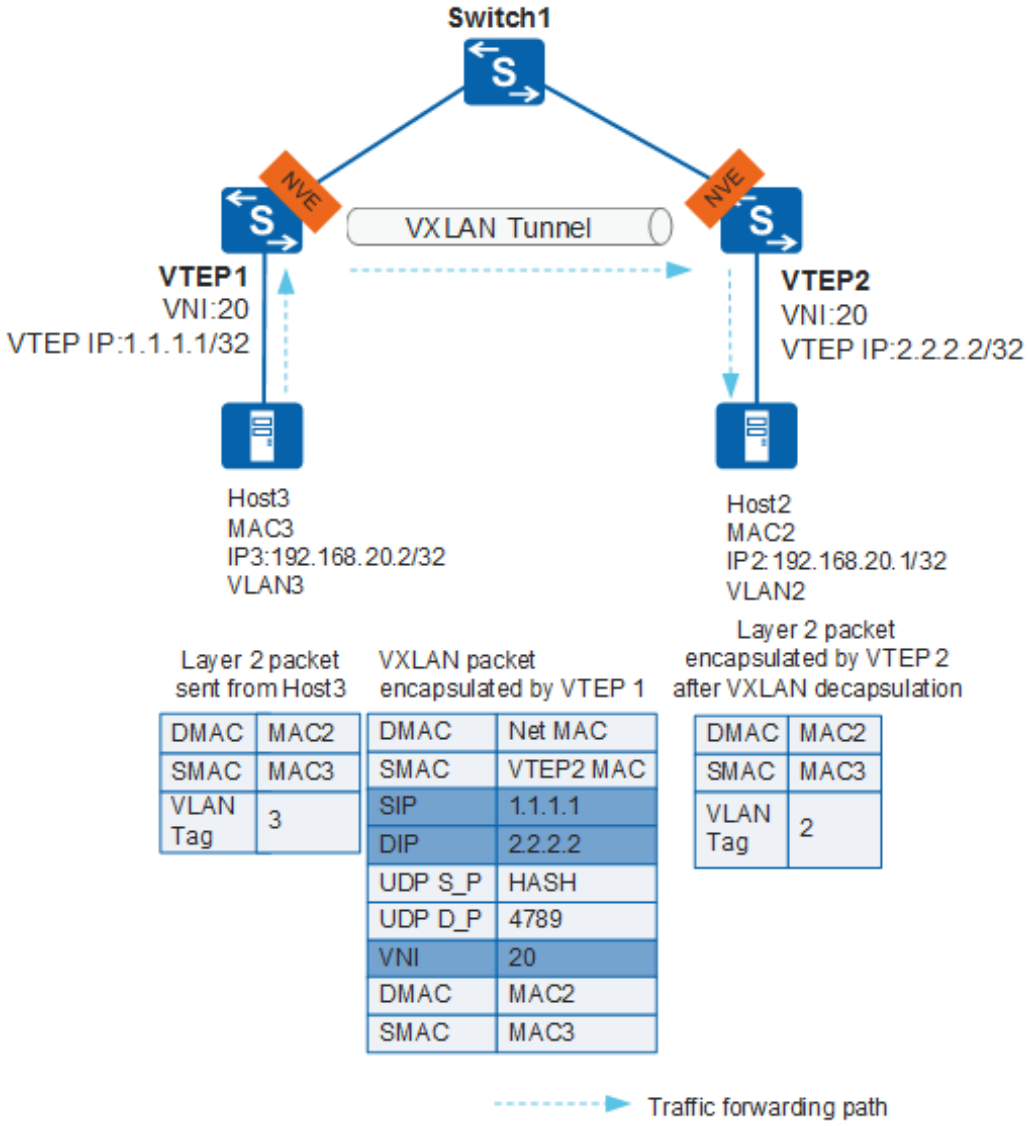
NOTE

Leaf nodes can learn the MAC addresses of hosts during data forwarding, if this capability is enabled. If VXLAN tunnels are established using BGP EVPN, leaf nodes can dynamically learn the MAC addresses of hosts through BGP EVPN routes, rather than data forwarding.

Intra-Subnet Known Unicast Packet Forwarding

Intra-subnet known unicast packets are forwarded only through Layer 2 VXLAN gateways and are unknown to Layer 3 VXLAN gateways. Figure 3-28 shows the intra-subnet known unicast packet forwarding process.

Figure 3-28 Intra-subnet known unicast packet forwarding



1. After VTEP 1 receives Host 3's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information and searches for the outbound interface and encapsulation information in the BD.
2. VTEP 1 performs VXLAN encapsulation based on the encapsulation information obtained and forwards the packets through the outbound interface obtained.
3. Upon receipt of the VXLAN packet, VTEP 2 verifies the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 2 obtains the Layer 2 BD based on the VNI and performs VXLAN decapsulation to obtain the inner Layer 2 packet.

4. VTEP 2 obtains the destination MAC address of the inner Layer 2 packet, performs VLAN tags to the packets based on the outbound interface and encapsulation information in the local MAC address table, and forwards the packets to Host 2.

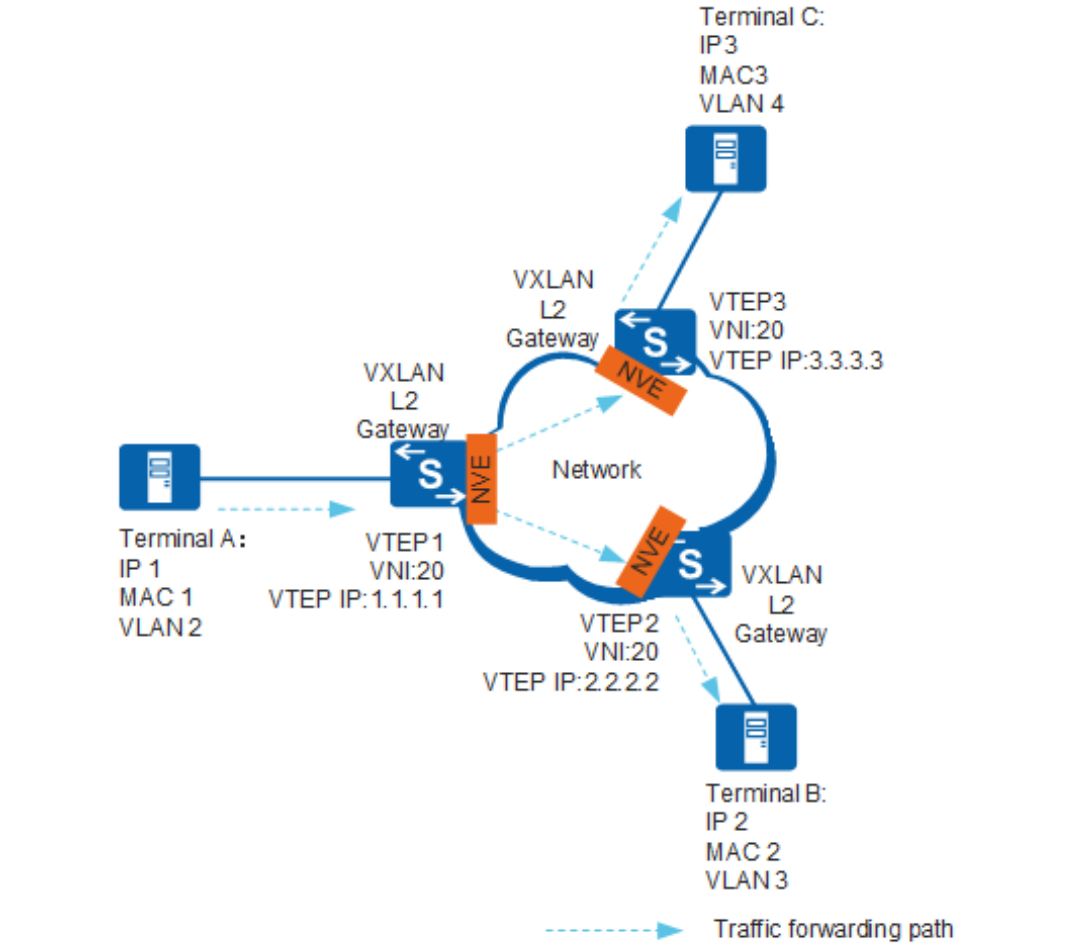
Host 2 sends packets to Host 3 in the same process.

Intra-Subnet BUM Packet Forwarding

Intra-subnet BUM packet forwarding is completed between Layer 2 VXLAN gateways. Layer 3 VXLAN gateways do not need to be unaware of the process. Intra-subnet BUM packets can be forwarded in ingress replication mode.

In ingress replication mode, after a BUM packet enters a VXLAN tunnel, the ingress VTEP performs VXLAN encapsulation based on the ingress replication list and sends the packet to all the egress VTEPs in the list. When the BUM packet leaves the VXLAN tunnel, the egress VTEPs decapsulate the BUM packet. Figure 3-29 shows the forwarding process of a BUM packet in ingress replication mode.

Figure 3-29 Forwarding process of an intra-subnet BUM packet in ingress replication mode



Layer 2 packet sent from Terminal A

DMAC	All Fs
SMAC	MAC1
VLAN Tag	2

VXLAN packet encapsulated by VTEP 1

VTEP1—>VTEP2

DMAC	Net MAC
SMAC	VTEP1 MAC1
SIP	1.1.1.1
DIP	2.2.2.2
UDP S_P	HASH
UDP D_P	4789
VNI	20
DMAC	All Fs
SMAC	MAC1

VTEP1—>VTEP3

DMAC	Net MAC
SMAC	VTEP1 MAC1
SIP	1.1.1.1
DIP	3.3.3.3
UDP S_P	HASH
UDP D_P	4789
VNI	20
DMAC	All Fs
SMAC	MAC1

Layer 2 packet encapsulated by VTEP 2/ VTEP 3 after VXLAN decapsulation

DMAC	All Fs
SMAC	MAC1
VLAN Tag	3

DMAC	All Fs
SMAC	MAC1
VLAN Tag	4

1. After VTEP 1 receives Terminal A's packet, it determines the Layer 2 BD of the packet based on the access interface and VLAN information.
2. VTEP 1 obtains the ingress replication list for the VNI, replicates packets based on the list, and performs VXLAN encapsulation by adding outer headers. VTEP 1 then forwards the VXLAN packet through the outbound interface.

3. Upon receipt of the VXLAN packet, VTEP 2 and VTEP 3 verify the VXLAN packet based on the UDP destination port number, source and destination IP addresses, and VNI. VTEP 2/VTEP 3 obtains the Layer 2 BD based on the VNI and performs VXLAN decapsulation to obtain the inner Layer 2 packet.
4. VTEP 2/VTEP 3 checks the destination MAC address of the inner Layer 2 packet and finds it a BUM MAC address. Therefore, VTEP 2/VTEP 3 broadcasts the packet onto the network connected to the terminals (not the VXLAN tunnel side) in the Layer 2 broadcast domain. Specifically, VTEP 2/VTEP 3 finds the outbound interfaces and encapsulation information not related to the VXLAN tunnel, performs VLAN tags to the packet, and forwards the packet to Terminal B/Terminal C.

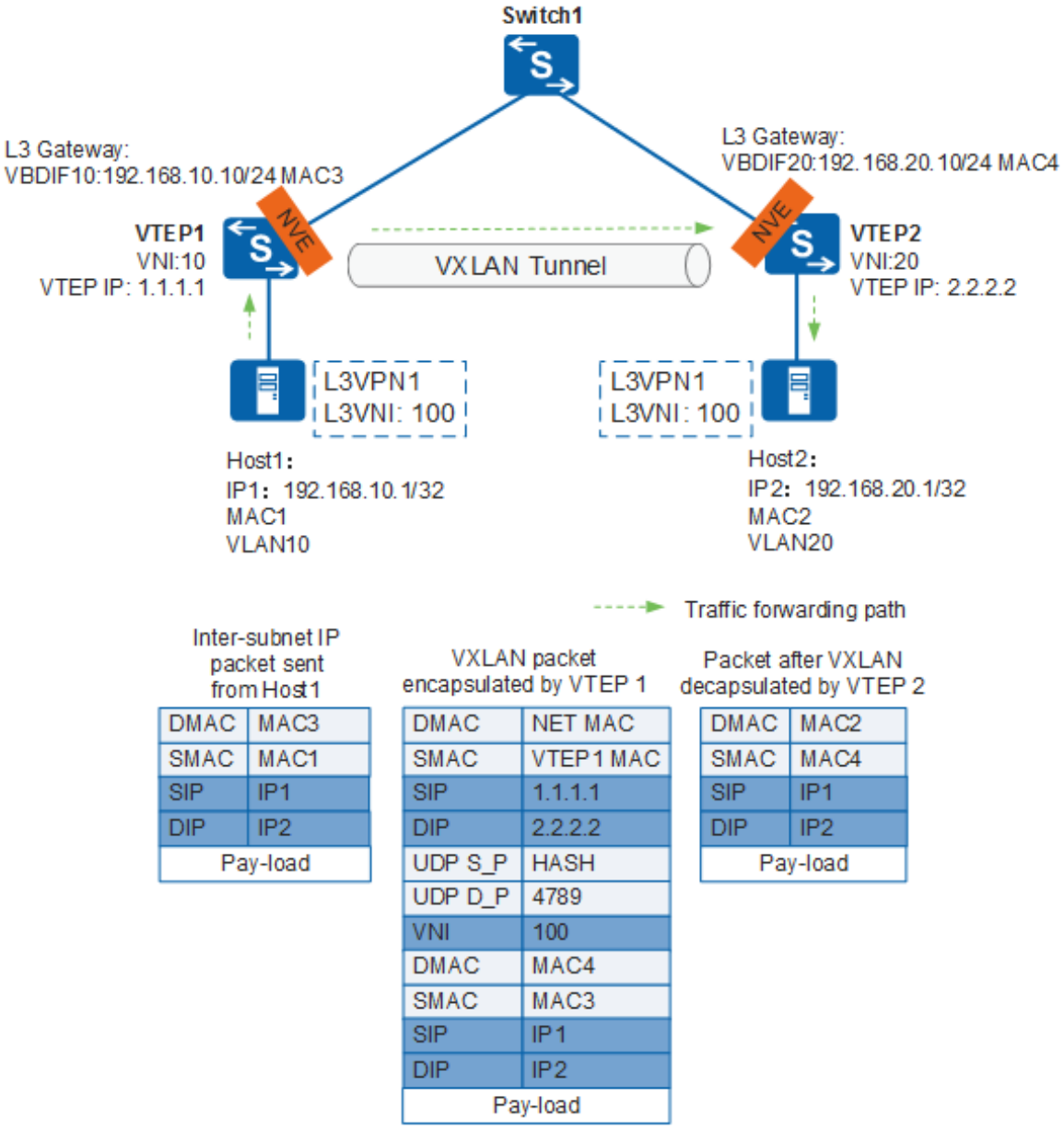
NOTE

Terminal B/Terminal C responds to Terminal A in the same process as intra-subnet known unicast packet forwarding.

Inter-Subnet Packet Forwarding

Inter-subnet packets must be forwarded through a Layer 3 gateway. Figure 3-30 shows the inter-subnet packet forwarding process in distributed VXLAN gateway scenarios.

Figure 3-30 Inter-subnet packet forwarding



1. After VTEP 1 receives a packet from Host 1, it finds that the destination MAC address of the packet is a gateway MAC address so that the packet must be forwarded at Layer 3.
2. VTEP 1 determines the Layer 2 broadcast domain of the packet based on the inbound interface and accordingly finds the L3VPN instance bound to the VBDIF interface of the Layer 2 broadcast domain. VTEP 1 then searches the L3VPN routing table and finds the destination address of packet. Figure 3-31 shows the host route in the L3VPN routing table. VTEP 1 obtains the Layer 3 VNI and next hop address of the host route and find that the iterated outbound interface is a VXLAN tunnel interface. Therefore, VTEP 1 determines that the packet must be transmitted through a VXLAN tunnel. Because the packet must be transmitted over a VXLAN tunnel, VTEP 1 performs VXLAN encapsulation as follows:

- Obtains the MAC address based on the VXLAN tunnel's source and destination IP addresses and replace the source and destination MAC addresses in the inner Ethernet header.
- Encapsulates the packet with the Layer 3 VNI.
- Encapsulates the VXLAN tunnels' source and destination IP addresses in the outer IP header, and VTEP 1's MAC address as the source MAC address and MAC address of the next hop pointing to the destination IP address as the destination MAC address in the outer Ethernet header.

Figure 3-31 Host route in the L3VPN routing table

L3VPN1:

Destination	L3 VNI	Next Hop	Outbound Interface
192.168.20.1/32	100	2.2.2.2	VXLAN tunnel

3. The VXLAN packet is then transmitted over the IP network based on the IP and MAC addresses in the outer headers and finally reaches VTEP 2.
4. After VTEP 2 receives the VXLAN packet, it decapsulates the packet and finds that the destination MAC address is its own MAC address so that the packet must be forwarded at Layer 3.
5. VTEP 2 determines the L3VPN instance bound to the Layer 3 VNI of the packet, searches the L3VPN routing table, and finds the next hop being the gateway IP address. VTEP 2 replaces the destination MAC address with Host 2's MAC address (MAC2) and source MAC address with VTEP 2's MAC address and sends the packet to Host 2. Figure 3-32 shows the host route in the L3VPN routing table.

Figure 3-32 Host route in the L3VPN routing table

L3VPN1:

Destination	L3 VNI	Next Hop	Outbound Interface
192.168.20.1/32	100	Gateway address	VBDIF

Host 2 sends packets to Host 1 in the same process.

NOTE

When a Huawei device communicates with a non-Huawei device, ensure that the non-Huawei device uses the same forwarding mode as that of the Huawei device. If they use different forwarding modes, the communication may fail.

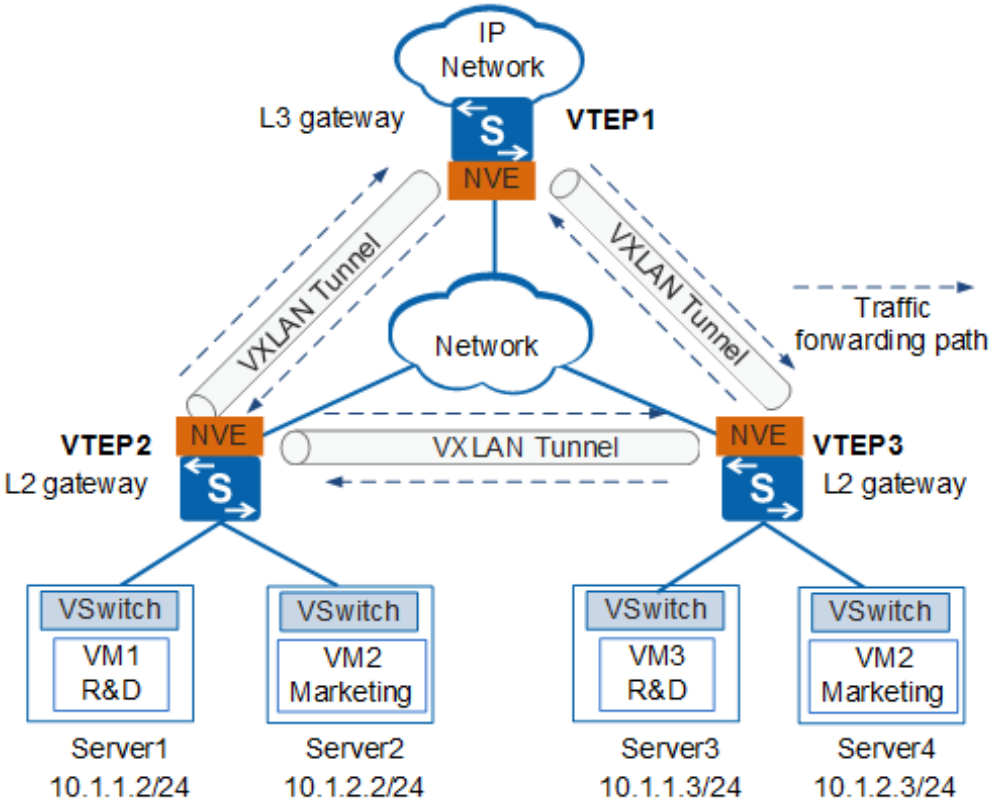
4 Application Scenarios for VXLANs

About This Chapter

4.1 Virtual Network Construction over a Campus Network

In Figure 4-1, an enterprise has constructed a mature campus network but does not have a dedicated data center network. All the servers of the enterprise are scattered in different departments, and they are interconnected through the campus network. The VXLAN technology can construct a virtual network over the campus network, realizing resource integration and flexible service deployment. To facilitate management and maintenance, VMs with the same service requirements are planned in the same network segment, while VMs with different service requirements are planned in different network segments. For example, VMs in the R&D department need to communicate in the same network segment; VMs in the R&D department and marketing department need to communicate across different network segments.

Figure 4-1 Virtual Network Construction over a Campus Network



VXLAN enables Layer 2 communication between virtual networks. For example, in Figure 4-1, VTEP2 and VTEP3 are Layer 2 VXLAN gateways, and they establish a VXLAN tunnel to enable VMs in the R&D department to communicate with each other in the same network segment.

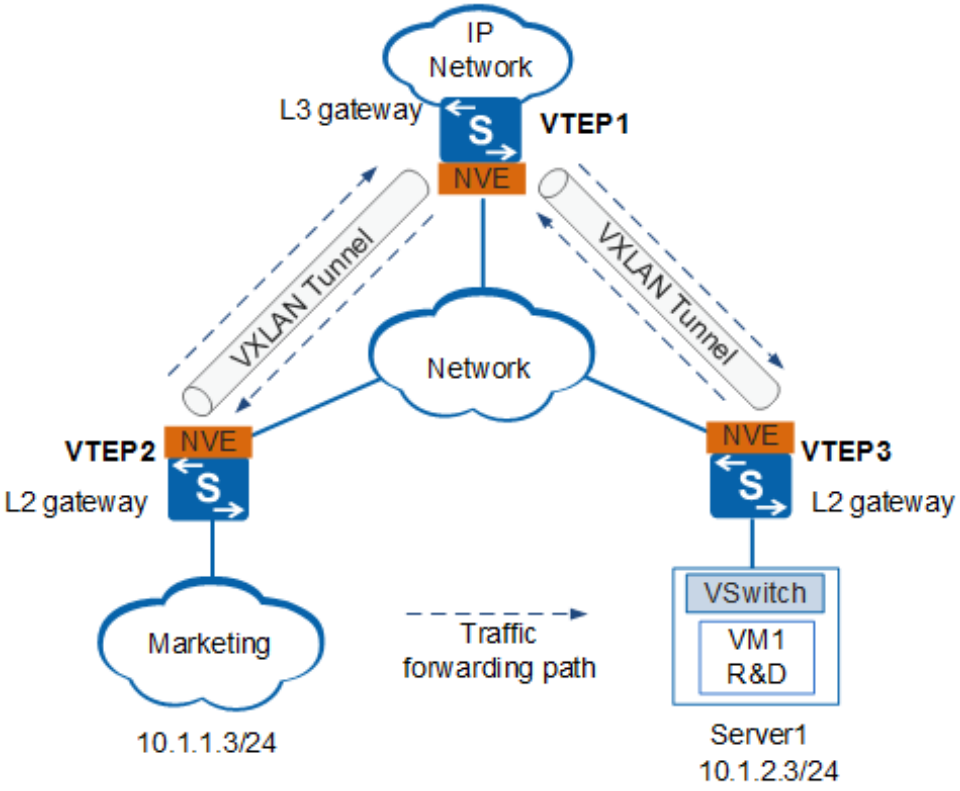
VXLAN enables Layer 3 communication between virtual networks. For example, when terminals in the R&D department want to communicate with terminals in the marketing department, VTEP1 functions as the Layer 3 VXLAN gateway to establish VXLAN tunnels with VTEP2 and VTEP3 respectively.

After static VXLAN tunnels are established between the switches, they dynamically learn flow table information, such as MAC address entries and ARP entries. After flow table information is learned, end users in the same or different network segments can communicate with each other over the VXLAN tunnels.

4.2 Mutual Access Between the Virtual Network and Campus Networks

In Figure 4-2, VMs are deployed for the R&D department of the enterprise and a traditional network is deployed for the marketing department. There are mutual access requirements between the R&D personnel on the virtual network and marketing personnel on the campus network in their daily work.

Figure 4-2 Mutual Access Between the Virtual Network and Campus Networks



In Figure 4-2, VTEP3 and VTEP2 are Layer 2 VXLAN gateways on the edge of the virtual network and traditional campus networks. VTEP1 serves as the Layer 3 VXLAN gateway, and it can establish VXLAN tunnels with VTEP2 and VTEP3 respectively to transmit VXLAN packets.

Packets from the marketing department to VM1 in the R&D department are processed in the following procedure:

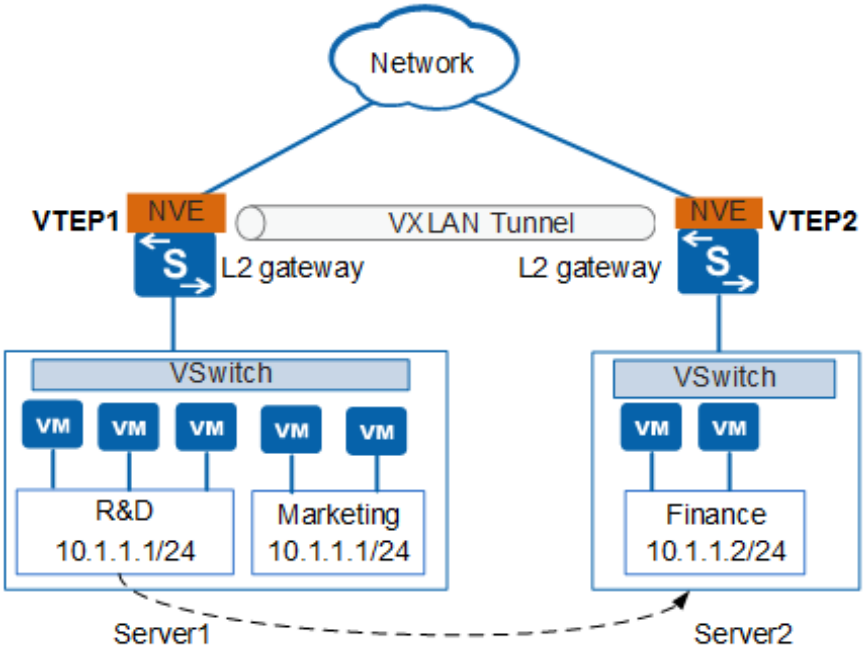
1. After receiving a packet from the traditional campus network, VTEP2 encapsulates the packet into a VXLAN packet and sends it to VTEP1.
2. VTEP1 decapsulates the received VXLAN packet, removes the Ethernet header from the inner packet, parses the destination IP address, searches the routing table for the next hop based on the destination IP address, searches ARP entries based on the next hop, and determines the destination MAC address, VXLAN tunnel outbound interface, and VNI.
3. VTEP1 re-encapsulates the VXLAN packet based on the obtained VXLAN tunnel outbound interface and VNI and sends it to VTEP3.
4. VTEP3 finds the outbound interface based on the destination MAC address in the packet and sends the packet to the correct VM.

4.3 Applying a Virtual Network in the VM Migration Scenario

Dynamic VM migration becomes a critical issue to meet flexible service changes. Dynamic VM migration is a process of moving VMs from one physical server to another, while ensuring normal running of the VMs. This process is also called smooth migration. End users are unaware of this process, so administrators can flexibly allocate server resources or maintain and upgrade servers without affecting server usage by end users.

The key of dynamic VM migration is to ensure uninterrupted services during the migration, so the IP and MAC addresses of VMs must remain unchanged. To meet this requirement, VM migration must occur within a Layer 2 domain but not across Layer 2 domains. In Figure 4-3, an enterprise has two servers deployed in a virtual network: Server1 providing services to the R&D and marketing departments, and Server2 providing services to the financial department. Because Server1 has insufficient computing space but Server2 is not fully utilized, the network administrator wants to migrate the R&D department to Server2 without service interruption.

Figure 4-3 Applying a Virtual Network to the VM Migration Scenario



The VXLAN technology can be used to establish a Layer 2 virtual network over any networks with reachable routes to implement Layer 2 interconnection. VXLAN encapsulates original packets sent by VMs over a VXLAN tunnel. VMs at two ends of a VXLAN tunnel do not need to know the physical architecture of the transmission network. In this way, VMs using IP addresses in the same network segment are in a Layer 2 domain logically, even if they are on different physical Layer 2 networks.

The VM migration process of the R&D department is as follows:

1. The R&D department is migrated from Server1 to Server2.
2. The VM in the R&D department sends gratuitous ARP or Reverse ARP (RARP) packets to instruct VTEP2 and other devices of the migration.

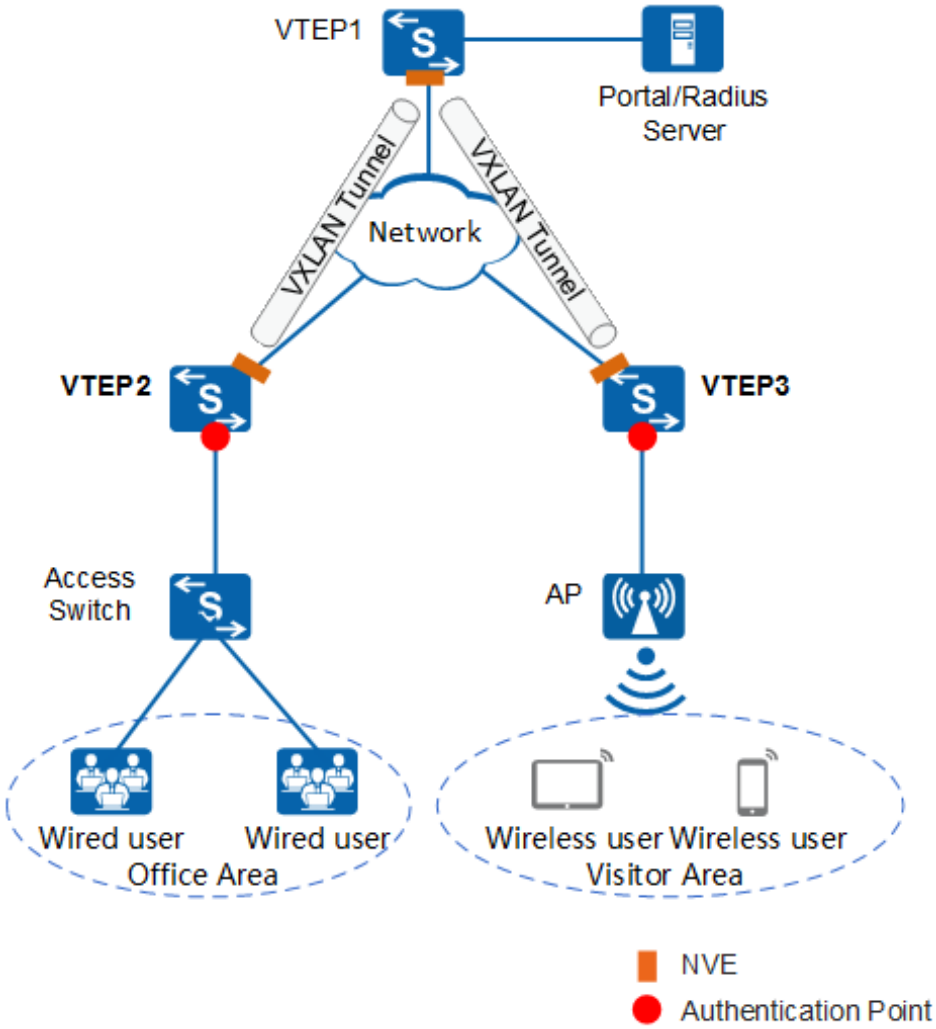
3. After learning the gratuitous ARP or RARP packets, VTEP1 updates the MAC address table and ARP table to that of the VM after the migration.

After the R&D department is migrated from Server1 to Server2, VM send gratuitous ARP or Reverse ARP (RARP) packets to update all gateways' MAC addresses and ARP entries of the original VMs to those of the VMs to which the R&D department is migrated.

4.4 Applying a Virtual Network in the User Access Authentication Scenario

In Figure 4-4, an enterprise builds a virtual network on a campus network using VXLAN for network planning. The office zone and guest zone connect to the virtual network through VTEP2 and VTEP3, respectively. Network access of the users needs to be controlled to ensure security of the enterprise intranet. Only the users who pass authentication are allowed to access authorized network resources.

Figure 4-4 Applying a virtual network to the user access authentication scenario



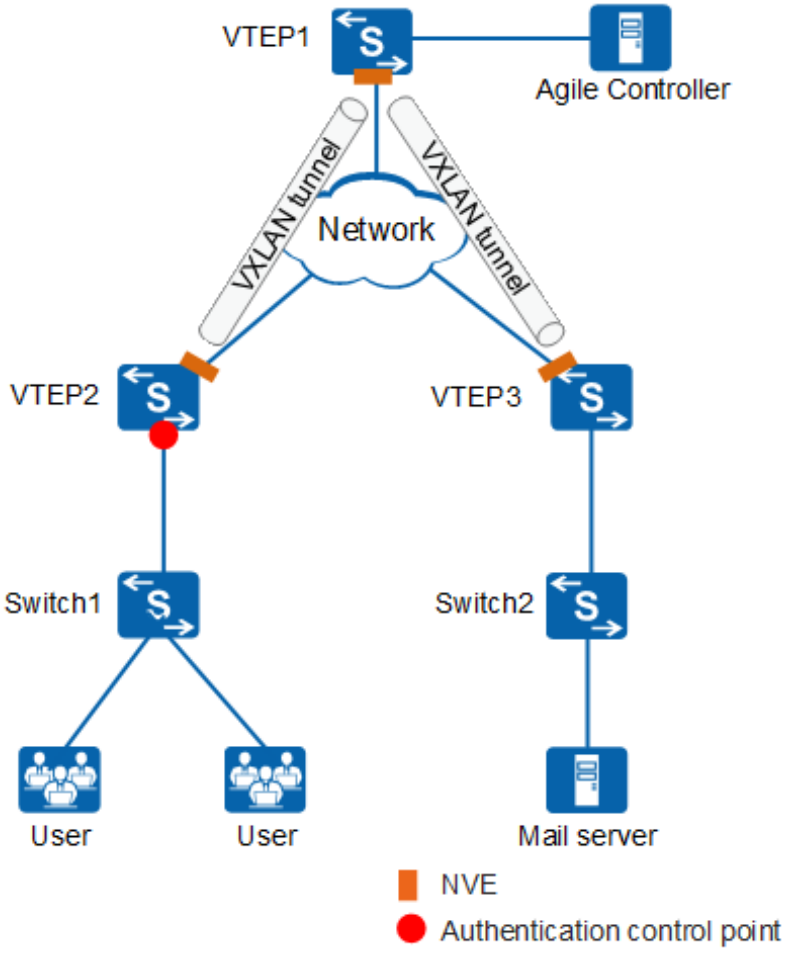
VXLAN-built virtual networks support two user access modes: VLAN and sub-interface. In the given scenario, user access can be authenticated only on physical interfaces. Therefore, in the user authentication scenario on a VXLAN network, user authentication and transparent VLAN transmission can be performed only on physical interfaces. The following describes how a wired user in Figure 4-4 is authenticated and accesses network resources:

1. After the user is connected to the access switch, a VLAN tag is added to the packet.
2. Upon arriving at VTEP2, the authentication request packet carrying the VLAN tag is VXLAN-encapsulated and sent to VTEP1 based on the binding relationship between the VLAN and BD on the device.
3. VTEP1 decapsulates the received packet and sends it to the Portal or RADIUS server for authentication based on the binding relationship between the VLAN and BD.
4. After the authentication is complete, the packet carrying the authentication response enters the VXLAN tunnel through VTEP1 and is then sent to VTEP2.
5. VTEP2 decapsulates the packet from VTEP1, obtains the authentication response from the Portal or RADIUS server, and completes user authentication.

4.5 Applying a Virtual Network in the Free Mobility Scenario

A security group will be assigned to a user when the user connects to the network through an authentication device. When the user attempt to access resources on other devices, no further authentication and authorization are required. A key technology in the free mobility solution is to synchronize the association between users and security groups from a device to other devices, eliminating the need of authentication and authorization when the users access these devices. This technology saves authentication and authorization resources.

Figure 4-5 Carrying user group information in VXLAN packets in the free mobility scenario



1. Establish VXLAN tunnels between VETP1 and VETP2 and between VTEP1 and VETP3.
2. A user accesses the network through Switch1. After the user is authenticated on VETP2, the RADIUS server (Agile Controller) associates the user with a user group.
3. Configure an ACL rule on VETP3 to forbid the user group to access the mail server.
4. When the user attempts to access the mail server, the request packet is sent to VETP3 through the VXLAN tunnels. The corresponding user group information is matched in the ACL rule on VETP3, so the user's access request is denied.

5 References for VXLANs

The following table lists the references for this document.

Document	Description	Protocol Compliance
RFC 7348	Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks	Partially compliant: Currently, broadcast of BUM packets is implemented through ingress replication but not multicast replication.

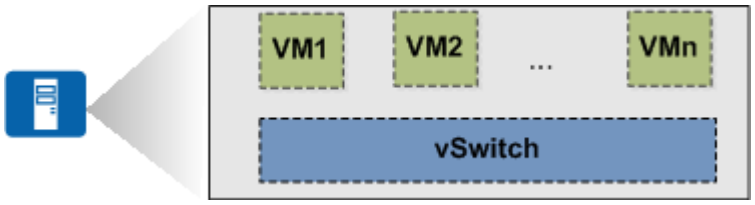
6 Further Reading

About This Chapter

6.1 Server Virtualization

Server virtualization virtualizes one physical server into multiple logical servers, that is virtual machines (VMs), as shown in Figure 6-1.

Figure 6-1 Basic architecture of server virtualization



- VM

  Each VM has its own operating system and application software, and has an independent MAC address and IP address. VMs can run independently.
- vSwitch

  A vSwitch provides Layer 2 communication, isolation, and QoS capabilities for VMs.

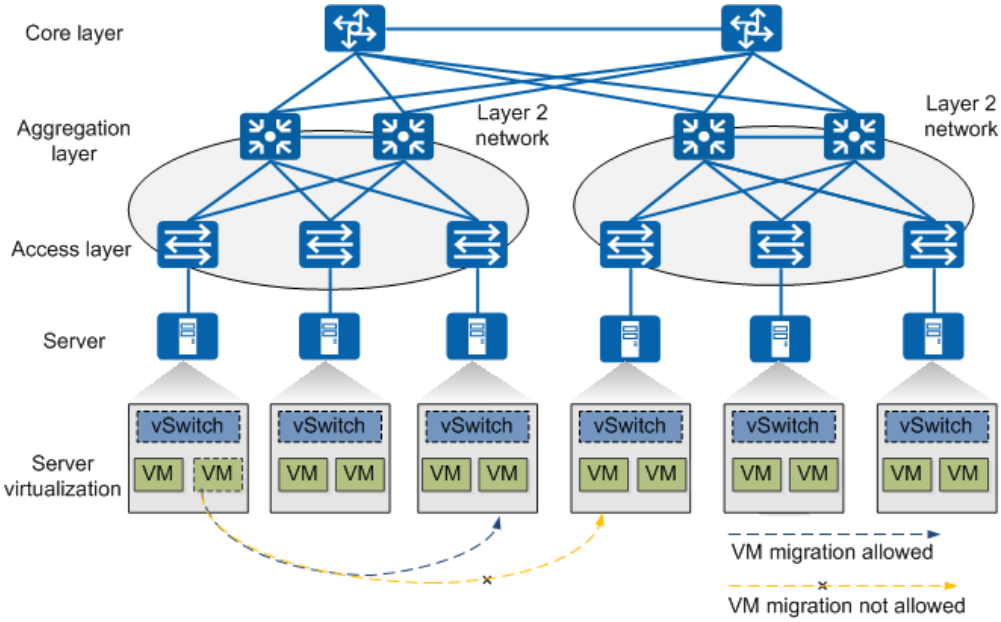
Server virtualization has the following advantages:

- Effectively improves server utilization.
- Provides services and resources on demand.
- Reduces energy consumption.
- Lowers customers' operations and maintenance (O&M) costs.

6.2 Large Layer 2 Network

Dynamic VM migration becomes a critical issue to meet flexible service changes. Dynamic VM migration is a process of moving VMs from one physical server to another, while ensuring normal running of the VMs. End users are unaware of this process, so administrators can flexibly allocate server resources or maintain and upgrade servers without affecting server usage by end users. The key of dynamic VM migration is to ensure uninterrupted services during the migration, so the IP and MAC addresses of VMs must remain unchanged. To meet this requirement, VM migration must occur within a Layer 2 domain but not across Layer 2 domains,as shown in Figure 6-2.

Figure 6-2 VM migration on a traditional network



In the traditional data center network architecture, the Layer 2 network uses redundant devices and links to improve reliability. This will inevitably result in physical loops during VM migration.

To prevent broadcast storms caused by physical loops, a loop prevention protocol such as Spanning Tree Protocol (STP) is required to block redundant links. Due to STP limitations, an STP-enabled Layer 2 network can contain no more than 50 network nodes, so dynamic VM migration can only occur in a limited scope.

To enable VM migration in a large scope or across domains, servers involved must be on the same Layer 2 network, which is called large Layer 2 network.

Generally, the following technologies can be used to provide a large Layer 2 network:

- Network device virtualization
- Transparent Interconnection of Lots of Links (TRILL)
- VXLAN
- Ethernet Virtual Network (EVN)

Network device virtualization, TRILL, and EVN technologies can construct a physical large Layer 2 network to enlarge the VM migration scope. However, a physical large Layer 2

network requires huge changes to the existing network structure, and still has many restrictions on the VM migration scope. VXLAN can solve the preceding problems.

A virtual large Layer 2 network can solve the problem and enable VM migration in a larger scope, as shown in Figure 6-3.

Figure 6-3 VM migration on a large Layer 2 network