VXLAN 技术白皮书

目 录

1 概述 1

1.1 产生背景 1

1.2 技术优点 1

2 VXLAN 技术实现 2

2.1 网络模型 2

2.2 VXLAN 支持 IPv6 3

2.3 报文封装格式 4

2.4 运行机制 5

2.4.1 运行机制概述 5

2.4.2 建立 VXLAN 隧道并将其与 VXLAN 关联 5

2.4.3 识别报文所属的 VXLAN 6

2.4.4 学习 MAC 地址 6

2.4.5 转发单播流量 7

2.4.6 转发泛洪流量 9

2.4.7 ARP/ND 泛洪抑制 12

2.5 VXLAN IP 网关 14

2.5.1 独立的 VXLAN IP 网关 14

2.5.2 集中式 VXLAN IP 网关 14

2.5.3 集中式 VXLAN IP 网关保护组 16

2.5.4 分布式 VXLAN IP 网关 17

3 Comware 实现的技术特色 22

3.1 VXLAN 支持 M-LAG 22

4 典型组网应用 24

4.1 VXLAN 二层互通组网 24

4.2 集中式 VXLAN IP 网关组网 24

4.3 分布式 VXLAN IP 网关组网 25

4.4 VXLAN 数据中心互联组网 26

4.5 VXLAN 与 SDN 控制器配合组网 27

5 参考文献 28

1 概述

1.1 产生背景

随着虚拟化技术的快速发展，数据中心的规模不断扩大，数据中心租户和虚拟机的数量呈爆发式增长，传统的二层网络面临着巨大的挑战：

- VLAN 资源不足

  传统的二层网络隔离技术 VLAN，因其标识相互隔离的虚拟二层网络的 Tag 域只有 12 比特，仅能划分出 4096 个相互隔离的虚拟二层网络，远远无法满足大二层网络中隔离大量租户的需求。

- 虚拟机迁移

  为了实现网络业务和资源的灵活调配，虚拟机跨设备甚至跨数据中心的迁移越来越频繁。为了保证虚拟机迁移过程中业务不中断，虚拟机迁移前后的 IP 地址和 MAC 地址需要保持不变，而传统网络技术无法实现虚拟机迁移前后的 IP、MAC 不变。

同时，随着数据中心多中心的部署，虚拟机的跨数据中心迁移、灾备，跨数据中心业务负载分担等需求，使得二层网络的扩展不仅是在数据中心的边界为止，还需要考虑跨越数据中心机房的区域，延伸到同城备份中心、远程灾备中心。一般情况下，多数据中心之间是通过路由连通的，天然是一个三层网络。而要实现通过三层网络连接的两个二层网络互通，就必须实现“L2 over L3”。

VXLAN（Virtual eXtensible LAN，可扩展虚拟局域网络）是基于 IP 网络、采用“MAC in UDP”封装形式的二层 VPN 技术。VXLAN 可以基于已有的服务提供商或企业 IP 网络，为分散的物理站点提供二层互联，并能够为不同的租户提供业务隔离。VXLAN 主要应用于数据中心网络和园区接入网络。

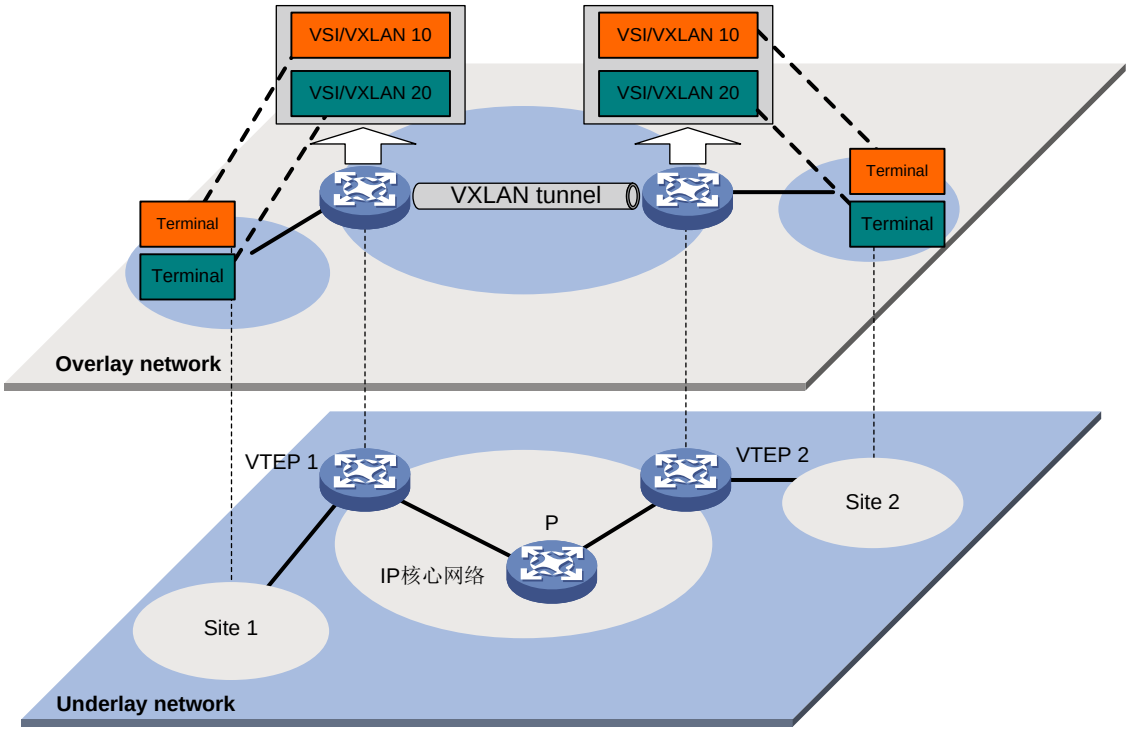
1.2 技术优点

- 支持大量的租户：使用 24 位的标识符，最多可支持 2 的 24 次方（16777216）个 VXLAN，支持的租户数目大规模增加，解决了传统二层网络 VLAN 资源不足的问题。
- 虚拟机迁移 IP、MAC 不变：采用了 MAC in UDP 的封装方式，实现原始二层报文在 IP 网络中的透明传输，保证虚拟机迁移前后的 IP 和 MAC 不变。
- 易于维护：基于 IP 网络组建大二层网络，使得网络部署和维护更加容易，并且可以充分地利用现有的 IP 网络技术，例如利用等价路由进行负载分担等；只有 IP 核心网络的边缘设备需要进行 VXLAN 处理，网络中间设备只需根据 IP 头转发报文，降低了网络部署的难度和费用。

2 VXLAN技术实现

2.1 网络模型

图1 VXLAN 网络模型示意图



如图 1 所示，VXLAN 的典型网络模型中包括如下几部分：

- 用户终端（Terminal）：用户终端设备可以是 PC 机、无线终端设备、服务器上创建的 VM（Virtual Machine，虚拟机）等。不同的用户终端可以属于不同的 VXLAN。属于相同 VXLAN 的用户终端处于同一个逻辑二层网络，彼此之间二层互通；属于不同 VXLAN 的用户终端之间二层隔离。

  说明

  本文档中如无特殊说明，均以 VM 为例介绍 VXLAN 工作机制。采用其他类型用户终端时，VXLAN 工作机制与 VM 相同，不再赘述。

- VTEP（VXLAN Tunnel End Point，VXLAN 隧道端点）：VXLAN 的边缘设备。VXLAN 的相关处理都在 VTEP 上进行，例如识别以太网数据帧所属的 VXLAN、基于 VXLAN 对数据帧进行二层转发、封装和解封装报文等。VTEP 可以是一台独立的物理设备，也可以是虚拟机所在的服务器。VTEP 可以划分为 VTEP 和 GW 两种角色：
  - VTEP：只支持 VXLAN 二层转发功能的设备，即只能在相同 VXLAN 内进行二层转发。
  - GW：可以进行跨 VXLAN 或者访问外部 IP 网络等三层转发的设备。根据部署方式，GW 可以分为集中式网关和分布式网关两种。

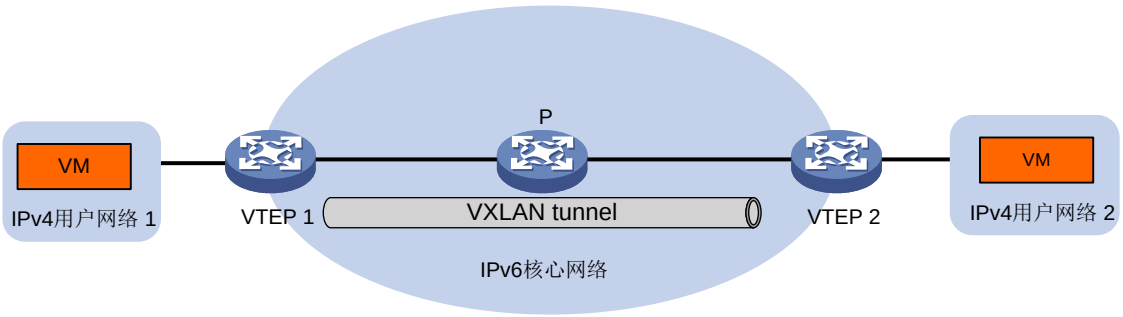
- VXLAN 隧道：两个 VTEP 之间的点到点逻辑隧道。VTEP 为数据帧封装 VXLAN 头、UDP 头和 IP 头后，通过 VXLAN 隧道将封装后的报文转发给远端 VTEP，远端 VTEP 对其进行解封装。
- 核心设备：IP 核心网络中的设备（如图 1 中的 P 设备）。核心设备不参与 EVPN 处理，仅需要根据封装后报文的外层目的 IP 地址对报文进行三层转发。
- VXLAN 网络：用户网络可能包括分布在不同地理位置的多个站点内的用户终端。在骨干网上可以利用 VXLAN 隧道将这些站点连接起来，为用户提供一个逻辑的二层 VPN。这个二层 VPN 称为一个 VXLAN 网络。VXLAN 网络通过 VXLAN ID 来标识，VXLAN ID 又称 VNI（VXLAN Network Identifier，VXLAN 网络标识符），其长度为 24 比特。不同 VXLAN 网络中的用户终端不能二层互通。
- VSI（Virtual Switch Instance，虚拟交换实例）：VTEP 上为一个 VXLAN 提供二层交换服务的虚拟交换实例。VSI 可以看作是 VTEP 上的一台基于 VXLAN 进行二层转发的虚拟交换机。它具有传统以太网交换机的所有功能，包括源 MAC 地址学习，MAC 地址老化，泛洪等。VSI 与 VXLAN 一一对应。
- VSI-Interface（VSI 的虚拟三层接口）：作为 VXLAN 内虚拟机的网关，用于处理跨 VXLAN 网络的报文转发。一个 VXLAN 网络对应一个 VSI-Interface。

2.2 VXLAN支持IPv6

VXLAN 支持用户网络和 IP 核心网络为 IPv6 网络。当用户网络或 IP 核心网络是 IPv6 网络时，可以通过部署 VXLAN 实现网络互通。

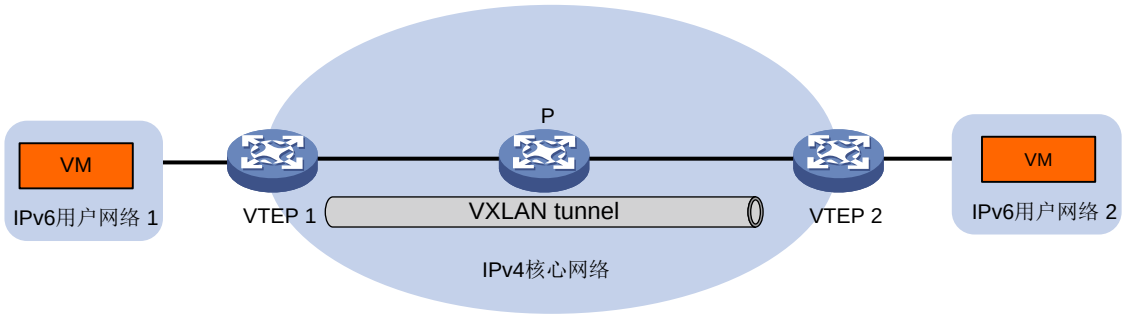
如图 2 所示，若核心网络已升级为 IPv6 网络，用户网络仍为 IPv4 网络，可以通过在 VTEP 之间建立 IPv6 VXLAN 隧道，实现 IPv4 用户网络跨 IPv6 网络的互通。

图2 IPv4 over IPv6 VXLAN 网络示意图



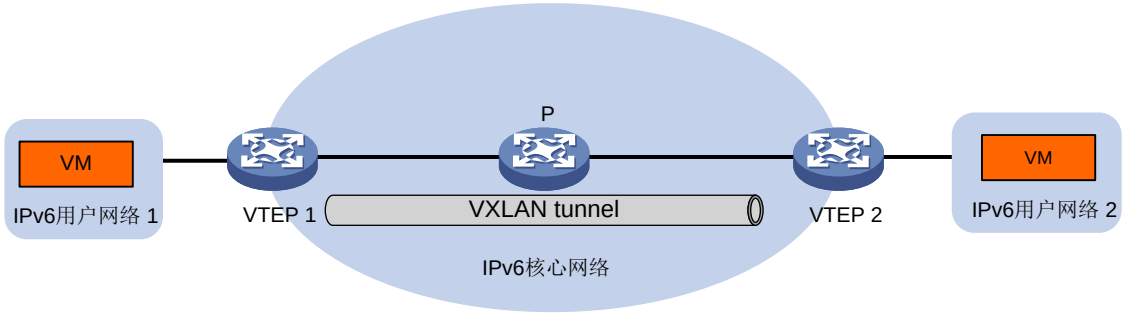
如图 3 所示，若用户网络已升级为 IPv6 网络，核心网络依然为 IPv4 网络，可以通过在 VTEP 之间建立 IPv4 VXLAN 隧道，实现 IPv6 用户网络跨 IPv4 网络的互通。

图3 IPv6 over IPv4 VXLAN 网络示意图



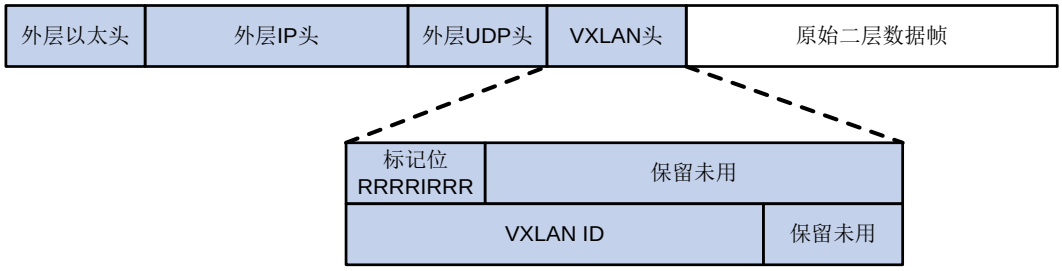
如图 4 所示，若用户网络、核心网络均已为 IPv6 网络，则可以通过在 VTEP 之间建立 IPv6 VXLAN 隧道，实现 IPv6 用户网络的互通。

图4 IPv6 over IPv6 VXLAN 网络示意图



2.3 报文封装格式

图5 VXLAN 报文封装示意图



如图 5 所示，VXLAN 报文的封装格式为：在原始二层数据帧外添加 VXLAN 头、UDP 头、IP 头和外层以太头。

- 外层以太头：长度为 14 字节，若包含 VLAN Tag 则为 18 字节。其中，源 MAC 地址为源 VM 所属 VTEP 的 MAC 地址，目的 MAC 地址为到达目的 VTEP 的路径上下一跳设备的 MAC 地址。

- 外层 IP 头：既可以是 IPv4 报文头，也可以是 IPv6 报文头。IPv4 报文头长度为 20 字节，IPv6 报文头长度为 40 字节。其中，源 IP 地址为源 VM 所属 VTEP 的 IP 地址，目的 IP 地址为目的 VM 所属 VTEP 的 IP 地址。
- 外层 UDP 报文头：长度为 8 字节。其中，UDP 目的端口号缺省为 4789，表示内层封装报文为 VXLAN 报文。UDP 源端口号（UDP Source Port）为本地随机选取的数值，可以用于 VTEP 之间多路径负载分担的计算。
- VXLAN 头长度为 8 字节，主要包括以下部分：
  - 标记位：“I”位为 1 时，表示 VXLAN 头中的 VXLAN ID 有效；为 0，表示 VXLAN ID 无效。其他位保留未用，设置为 0。
  - VXLAN ID：用来标识一个 VXLAN 网络，长度为 24 比特。
  - Reserved：当前协议保留位。
- 原始二层数据帧：虚拟机发送的原始以太网报文。

从报文的封装可以看出，VXLAN 头和原始二层数据帧是作为 UDP 报文的载荷存在的。VTEP 之间的网络设备只需要根据外层以太头和外层 IP 头进行转发、利用源 UDP 端口号进行负载分担。在这一过程中，VXLAN 报文的处理与普通的 IP 报文完全相同。因此，除了 VTEP 设备，现网的大量设备无需更换或升级即可支持 VXLAN 网络。

2.4 运行机制

2.4.1 运行机制概述

VXLAN 运行机制可以概括为：

(1) 发现远端 VTEP，在 VTEP 之间建立 VXLAN 隧道，并将 VXLAN 隧道与 VXLAN 关联。

(2) 识别接收到的报文所属的 VXLAN，以便将报文的源 MAC 地址学习到 VXLAN 对应的 VSI，并在该 VSI 内转发该报文。

(3) 学习终端的 MAC 地址。

(4) 根据学习到的 MAC 地址表项转发报文。

2.4.2 建立 VXLAN 隧道并将其与 VXLAN 关联

为了将 VXLAN 报文传递到远端 VTEP，需要创建 VXLAN 隧道，并将 VXLAN 隧道与 VXLAN 关联。

1. 创建 VXLAN 隧道

VXLAN 隧道的建立方式有如下两种：

- 手工方式：手工配置 Tunnel 接口，并指定隧道的源和目的 IP 地址分别为本端 VTEP 和远端 VTEP 的 IP 地址。
- 自动方式：通过 EVPN（Ethernet Virtual Private Network，以太网虚拟专用网络）发现远端 VTEP 后，自动在本端 VTEP 和远端 VTEP 之间建立 VXLAN 隧道。

2. 关联 VXLAN 隧道与 VXLAN

VXLAN 隧道与 VXLAN 关联的方式有如下两种：

- 手工方式：手工将 VXLAN 隧道与 VXLAN 关联。
- 自动方式：通过 EVPN 协议自动将 VXLAN 隧道与 VXLAN 关联。

2.4.3 识别报文所属的 VXLAN

1. 本地站点内接收到数据帧的识别

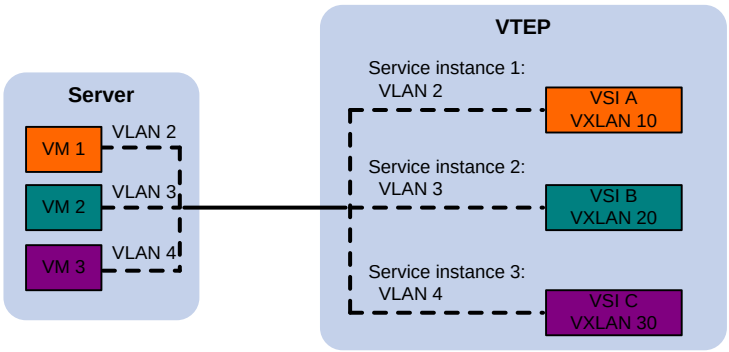
VTEP 采用如下几种方式在数据帧和 VXLAN 之间建立关联：

- 将三层接口与 VSI 关联：从该三层接口接收到的数据帧均属于指定的 VSI。VSI 内创建的 VXLAN 即为该数据帧所属的 VXLAN。
- 将以太网服务实例与 VSI 关联：以太网服务实例定义了一系列匹配规则，如匹配指定 VLAN 的报文、匹配接口接收到的所有报文等。从二层以太网接口上接收到的、与规则匹配的数据帧均属于指定的 VSI/VXLAN。
- 将 VLAN 与 VXLAN 关联：VTEP 接收到的该 VLAN 的数据帧均属于指定的 VXLAN。

VTEP 从指定 VLAN、三层接口或以太网服务实例接收到数据帧后，根据关联方式判断报文所属的 VXLAN。

如图 6 所示，VM 1 属于 VLAN 2，在 VTEP 上配置以太网服务实例 1 匹配 VLAN 2 的报文，将以太网服务实例 1 与 VSI A 绑定，并在 VSI A 内创建 VXLAN 10，则 VTEP 接收到 VM 1 发送的数据帧后，可以判定该数据帧属于 VXLAN 10。

图6 二层数据帧所属 VXLAN 识别



2. VXLAN 隧道上接收报文的识别

对于从 VXLAN 隧道上接收到的 VXLAN 报文，VTEP 根据报文中携带的 VXLAN ID 判断该报文所属的 VXLAN。

2.4.4 学习 MAC 地址

MAC 地址学习分为本地 MAC 地址学习和远端 MAC 地址学习两部分。

1. 本地 MAC 地址学习

本地 MAC 地址学习是指 VTEP 对本地站点内 VM 的 MAC 地址的学习。本地 MAC 地址的学习方式有如下几种：

- 静态配置：手工指定本地 MAC 地址所属的 VSI（即 VXLANI），及其对应的以太网服务实例（即 AC）。
- 通过报文中的源 MAC 地址动态学习：VTEP 接收到本地 VM 发送的数据帧后，判断该数据帧所属的 VSI，并将数据帧中的源 MAC 地址（本地站点内 VM 的 MAC 地址）添加到该 VSI 的 MAC 地址表中，该 MAC 地址对应的接口为接收到数据帧的接口。

2. 远端 MAC 地址学习

远端 MAC 地址学习是指 VTEP 对远端站点内 VM 的 MAC 地址的学习。远端 MAC 地址的学习方式有如下几种：

- 静态配置：手工指定远端 MAC 地址所属的 VSI（即 VXLAN），及其对应的 VXLAN 隧道接口。
- 通过报文中的源 MAC 地址动态学习：VTEP 从 VXLAN 隧道上接收到远端 VTEP 发送的 VXLAN 报文后，根据 VXLAN ID 判断报文所属的 VXLAN，对报文进行解封装，还原二层数据帧，并将数据帧中的源 MAC 地址（远端站点内 VM 的 MAC 地址）添加到所属 VXLAN 对应 VSI 的 MAC 地址表中，该 MAC 地址对应的接口为 VXLAN 隧道接口。
- 通过 BGP EVPN 学习：在 VTEP 上运行 BGP EVPN，通过 BGP EVPN 将本地 MAC 地址及其所属的 VXLAN 信息通告给远端 VTEP。远端 VTEP 接收到该信息后，在 VXLAN 对应 VSI 的 MAC 地址表中添加 MAC 地址表项，该 MAC 地址对应的接口为 VXLAN 隧道接口。
- 通过 OpenFlow 下发：OpenFlow 控制器以流表的形式向 VTEP 设备下发远端 MAC 地址表项。
- 通过 OVSDB 下发：控制器通过 OVSDB 协议向 VTEP 设备下发远端 MAC 地址表项。

通过不同方式学习到的远端 MAC 地址优先级由高到低依次为：

(1) 静态配置、OpenFlow 下发、OVSDB 下发的 MAC 地址优先级相同，且优先级最高。

(2) 通过 BGP EVPN 学习的 MAC 地址优先级次之。

(3) 动态学习的 MAC 地址优先级最低。

2.4.5 转发单播流量

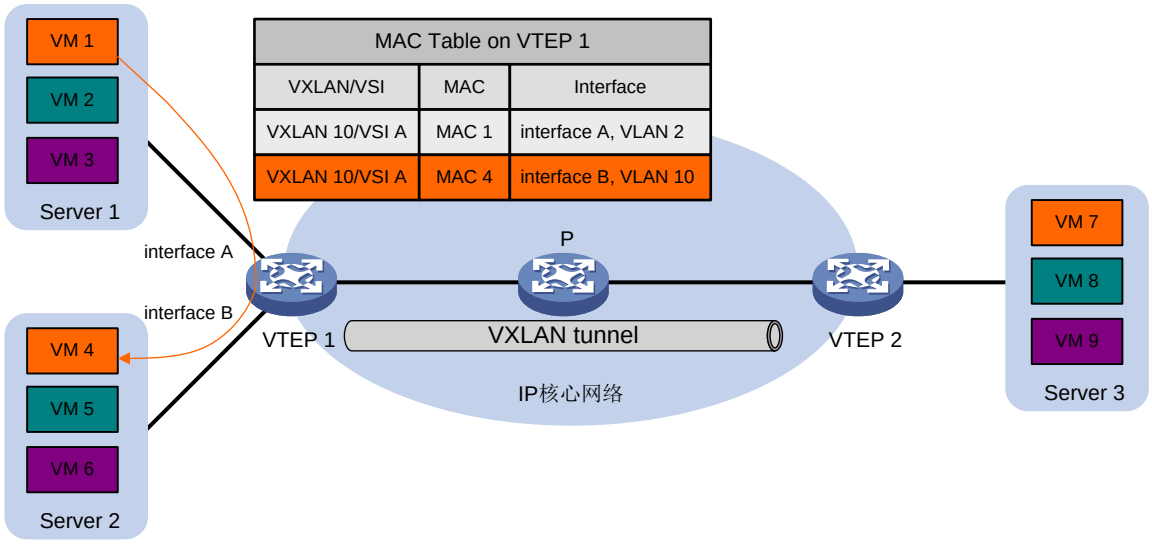
完成本地和远端 MAC 地址学习后，VTEP 在 VXLAN 内转发单播流量的过程如下所述。

1. 站点内流量

对于站点内流量，VTEP 判断出报文所属的 VSI 后，根据目的 MAC 地址查找该 VSI 的 MAC 地址表，从相应的本地接口转发给目的 VM。

如图 7 所示，VM 1（MAC 地址为 MAC 1）发送以太网帧到 VM 4（MAC 地址为 MAC 4）时，VTEP 1 从接口 Interface A 收到该以太网帧后，判断该数据帧属于 VSI A（VXLAN 10），查找 VSI A 的 MAC 地址表，得到 MAC 4 的出接口为 Interface B，所在 VLAN 为 VLAN 10，则将以太网帧从接口 Interface B 的 VLAN 10 内发送给 VM 4。

图7 站点内单播流量转发



2. 站点间流量

如图 8 所示，以 VM 1（MAC 地址为 MAC 1）发送以太网帧给 VM 7（MAC 地址为 MAC 7）为例，站点间单播流量的转发过程为：

(1) VM 1 发送以太网数据帧给 VM 7，数据帧的源 MAC 地址为 MAC 1，目的 MAC 为 MAC 7，VLAN ID 为 2。

(2) VTEP 1 从接口 Interface A（所在 VLAN 为 VLAN 2）收到该数据帧后，判断该数据帧属于 VSI A（VXLAN 10），查找 VSI A 的 MAC 地址表，得到 MAC 7 的出端口为 Tunnel1。

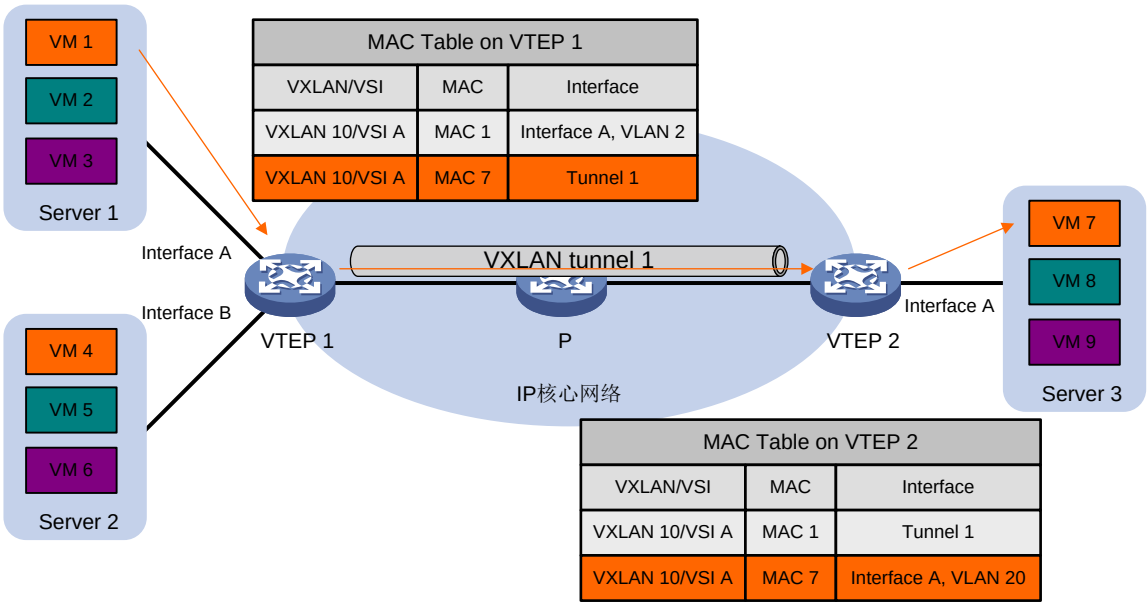
(3) VTEP 1 为数据帧封装 VXLAN 头、UDP 头和 IP 头后，将封装好的报文通过 VXLAN 隧道 Tunnel1、经由 P 设备发送给 VTEP 2。

(4) VTEP 2 接收到报文后，根据报文中的 VXLAN ID 判断该报文属于 VXLAN 10，并剥离 VXLAN 头、UDP 头和 IP 头，还原出原始的数据帧。

(5) VTEP 2 查找与 VXLAN 10 对应的 VSI A 的 MAC 地址表，得到 MAC 7 的出端口为 Interface A（所在 VLAN 为 VLAN 20）。

(6) VTEP 2 从接口 Interface A 的 VLAN 20 内将数据帧发送给 VM 7。

图8 站点间单播流量转发



2.4.6 转发泛洪流量

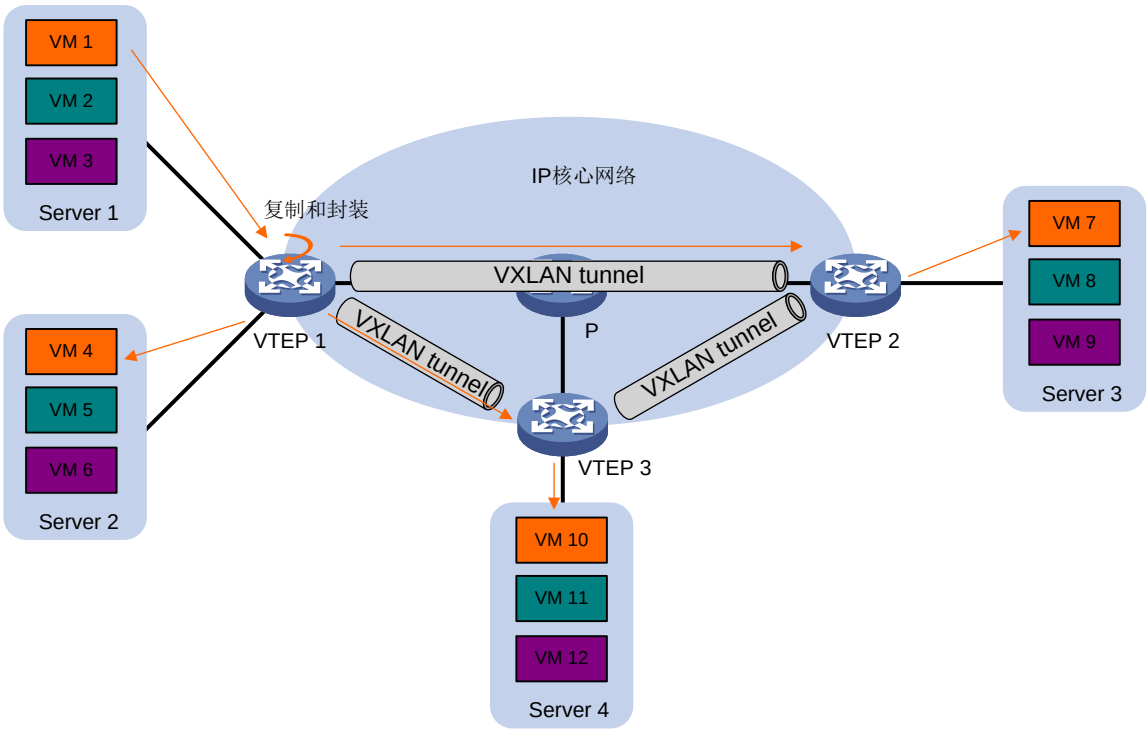
VTEP 从本地站点接收到泛洪流量（组播、广播和未知单播流量）后，将其转发给除接收接口外的所有本地接口和 VXLAN 隧道。为了避免环路，VTEP 从 VXLAN 隧道上接收到报文后，不会再将其泛洪到其他的 VXLAN 隧道，只会转发给所有本地接口。

根据复制方式的不同，流量泛洪方式分为单播路由方式（头端复制）、组播路由方式（核心复制）和泛洪代理方式（服务器复制）。

1. 单播路由方式（头端复制）

如图 9 所示，VTEP 负责复制报文，采用单播方式将复制后的报文通过本地接口发送给本地站点，并通过 VXLAN 隧道发送给 VXLAN 内的所有远端 VTEP。

图9 单播路由方式转发示意图



2. 组播路由方式（核心复制）

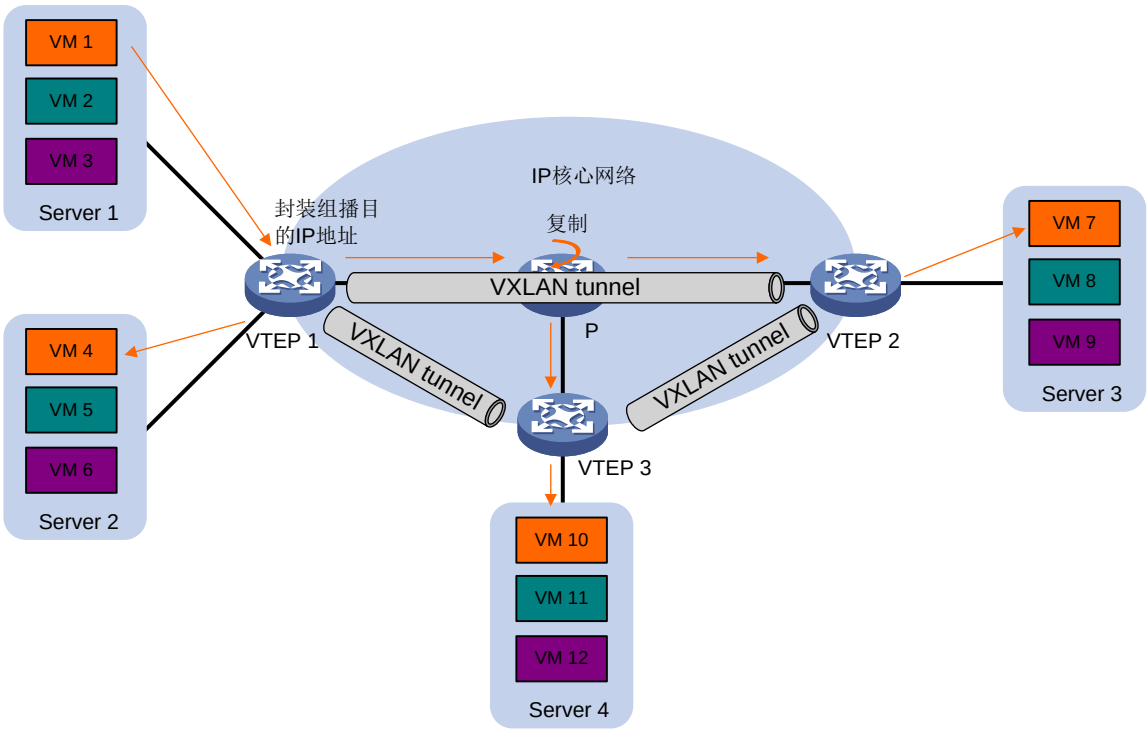
说明

IPv6 网络作为核心网络时不支持组播路由方式转发泛洪流量。

数据中心网络中需要通过 IP 核心网络进行二层互联的站点较多时，采用组播路由方式可以节省泛洪流量对核心网络带宽资源的占用。

如图 10 所示，在组播路由方式下，同一个 VXLAN 内的所有 VTEP 都加入同一个组播组，利用组播路由协议（如 PIM）在 IP 核心网上为该组播组建立组播转发表项。VTEP 接收到泛洪流量后，不仅在本地站点内泛洪，还会为其封装组播目的 IP 地址，封装后的报文根据已建立的组播转发表项转发到远端 VTEP。

图10 组播路由方式转发示意图

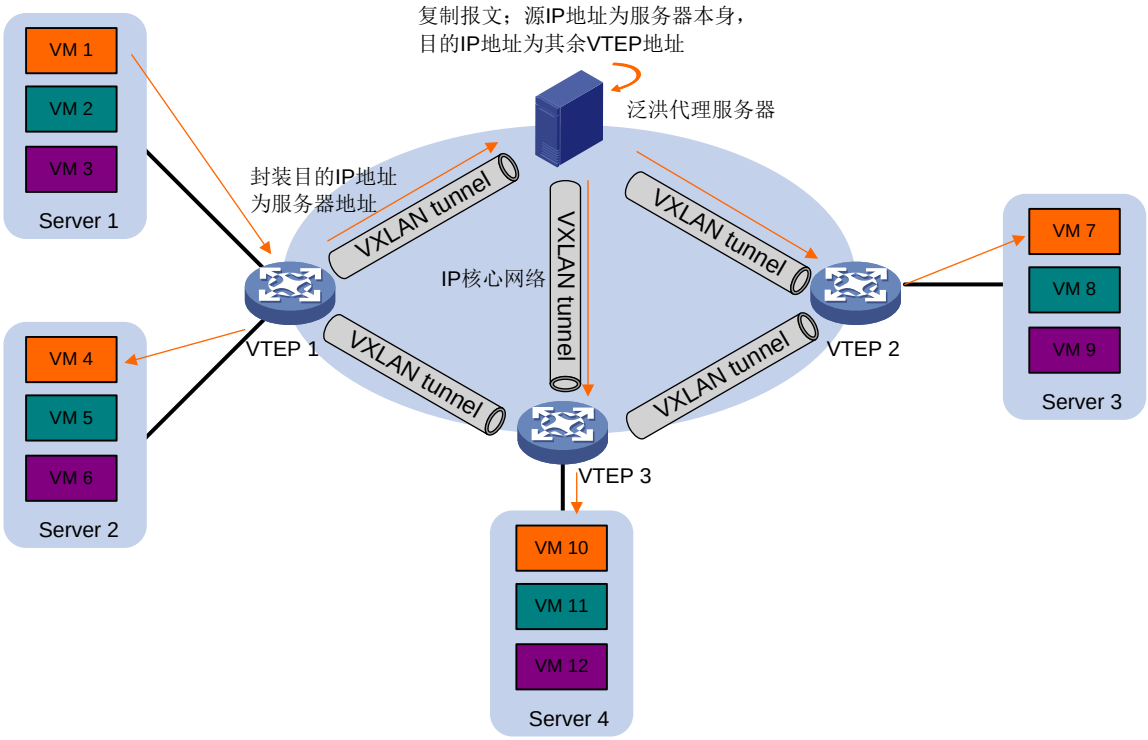


3. 泛洪代理方式（服务器复制）

数据中心网络中需要通过 IP 核心网络进行二层互联的站点较多时，采用泛洪代理方式可以在没有组播协议参与的情况下，节省泛洪流量对核心网络带宽资源的占用。

如图 11 所示，在泛洪代理方式下，同一个 VXLAN 内的所有 VTEP 都通过手工方式与代理服务器建立隧道。VTEP 接收到泛洪流量后，不仅在本地站点内泛洪，还会将其发送到代理服务器，由代理服务器转发到其他远端 VTEP。

图11 泛洪代理方式转发示意图



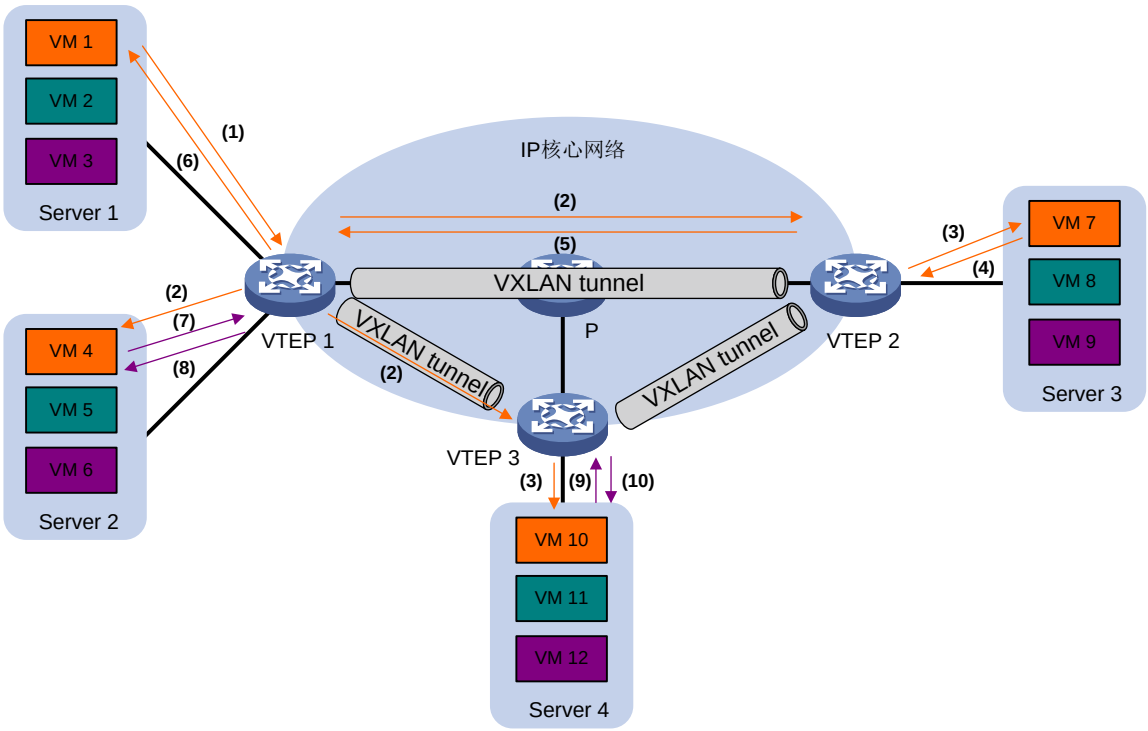
目前泛洪代理方式主要用于 SDN 网络，使用虚拟服务器作为泛洪代理服务器。采用泛洪代理方式时，需要注意如下几点：

- 在 VTEP 上关闭远端 MAC 地址自动学习功能，采用 SDN 控制器下发的 MAC 地址表项进行流量转发。
- 在 VTEP 网络侧接口上需要关闭报文入接口与静态 MAC 地址表项匹配检查功能。当 VTEP 设备为 IRF 设备时，成员设备间互连的 IRF 端口上也需要关闭报文入接口与静态 MAC 地址表项匹配检查功能。

2.4.7 ARP/ND 泛洪抑制

为了避免广播发送的 ARP 请求或组播发送的 ND 请求报文占用核心网络带宽，VTEP 从本地站点或 VXLAN 隧道接收到 ARP/ND 请求和 ARP/ND 应答报文后，根据该报文在本地建立 ARP/ND 泛洪抑制表项。后续当 VTEP 收到本站点内虚拟机请求其它虚拟机 MAC 地址的 ARP/ND 请求时，优先根据 ARP/ND 泛洪抑制表项进行代答。如果没有对应的表项，则将 ARP/ND 请求泛洪到核心网。ARP/ND 泛洪抑制功能可以大大减少 ARP/ND 泛洪的次数。

图12 ARP 泛洪抑制示意图



如图 12 所示，以 ARP 为例，泛洪抑制的处理过程如下：

(1) 虚拟机 VM 1 发送 ARP 请求，获取 VM 7 的 MAC 地址。

(2) VTEP 1 根据接收到的 ARP 请求，建立 VM 1 的 ARP 泛洪抑制表项，并在 VXLAN 内泛洪该 ARP 请求（图 12 以单播路由泛洪方式为例）。

(3) 远端 VTEP（VTEP 2 和 VTEP 3）解封装 VXLAN 报文，获取原始的 ARP 请求报文后，建立 VM 1 的 ARP 泛洪抑制表项，并在本地站点的指定 VXLAN 内泛洪该 ARP 请求。

(4) VM 7 接收到 ARP 请求后，回复 ARP 应答报文。

(5) VTEP 2 接收到 ARP 应答后，建立 VM 7 的 ARP 泛洪抑制表项，并通过 VXLAN 隧道将 ARP 应答发送给 VTEP 1。

(6) VTEP 1 解封装 VXLAN 报文，获取原始的 ARP 应答，并根据该应答建立 VM 7 的 ARP 泛洪抑制表项，之后将 ARP 应答报文发送给 VM 1。

(7) 在 VTEP 1 上建立 ARP 泛洪抑制表项后，虚拟机 VM 4 发送 ARP 请求，获取 VM 1 或 VM 7 的 MAC 地址。

(8) VTEP 1 接收到 ARP 请求后，建立 VM 4 的 ARP 泛洪抑制表项，并查找本地 ARP 泛洪抑制表项，根据已有的表项回复 ARP 应答报文，不会对 ARP 请求进行泛洪。

(9) 在 VTEP 3 上建立 ARP 泛洪抑制表项后，虚拟机 VM 10 发送 ARP 请求，获取 VM 1 的 MAC 地址。

(10) VTEP 3 接收到 ARP 请求后，建立 VM 10 的 ARP 泛洪抑制表项，并查找本地 ARP 泛洪抑制表项，根据已有的表项回复 ARP 应答报文，不会对 ARP 请求进行泛洪。

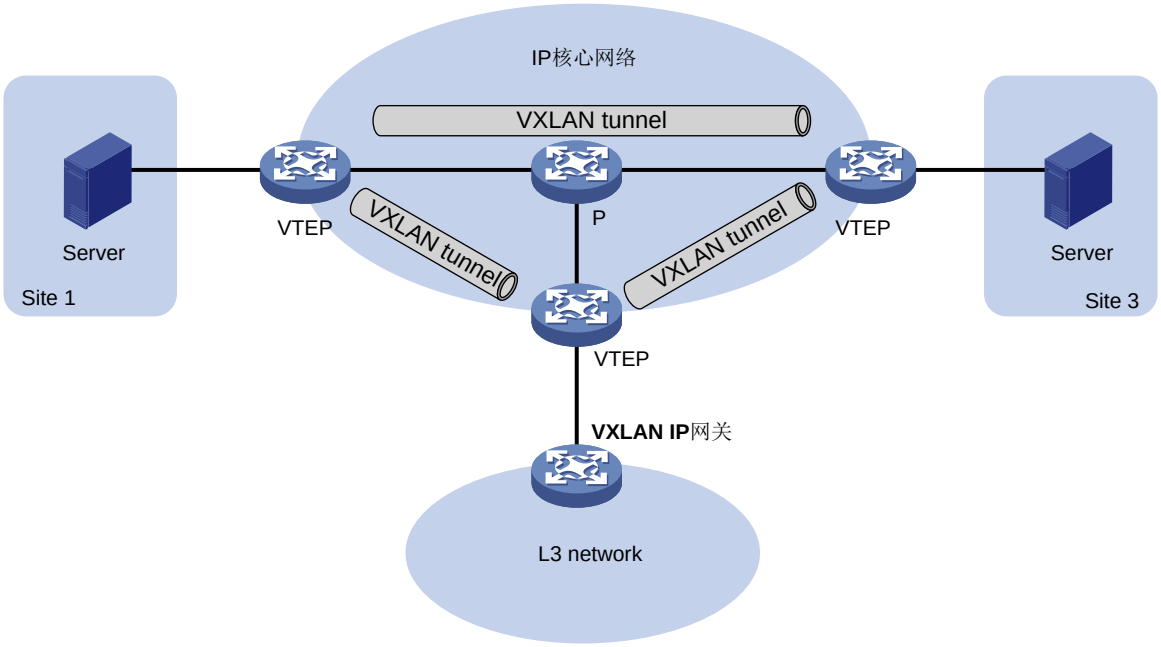
2.5 VXLAN IP网关

VXLAN 可以为分散的物理站点提供二层互联。如果要为 VXLAN 站点内的虚拟机提供三层业务，则需要在网络中部署 VXLAN IP 网关，以便站点内的虚拟机通过 VXLAN IP 网关与外界网络或其他 VXLAN 网络内的虚拟机进行三层通信。VXLAN IP 网关既可以部署在独立的物理设备上，也可以部署在 VTEP 设备上。VXLAN IP 网关部署在 VTEP 设备上时，又分为集中式 VXLAN IP 网关和分布式 VXLAN IP 网关两种方式。

2.5.1 独立的 VXLAN IP 网关

如图 13 所示，VXLAN IP 网关部署在独立的物理设备上时，VXLAN IP 网关作为物理站点接入 VTEP，VXLAN 业务对于网关设备透明。虚拟机通过 VXLAN IP 网关与三层网络中的节点通信时，虚拟机将三层报文封装成二层数据帧发送给 VXLAN IP 网关。VTEP 对该数据帧进行 VXLAN 封装，并在 IP 核心网络上将其转发给远端 VTEP（连接 VXLAN IP 网关的 VTEP）。远端 VTEP 对 VXLAN 报文进行解封装，并将原始的二层数据帧转发给 VXLAN IP 网关。VXLAN IP 网关去掉链路层封装后，对报文进行三层转发。

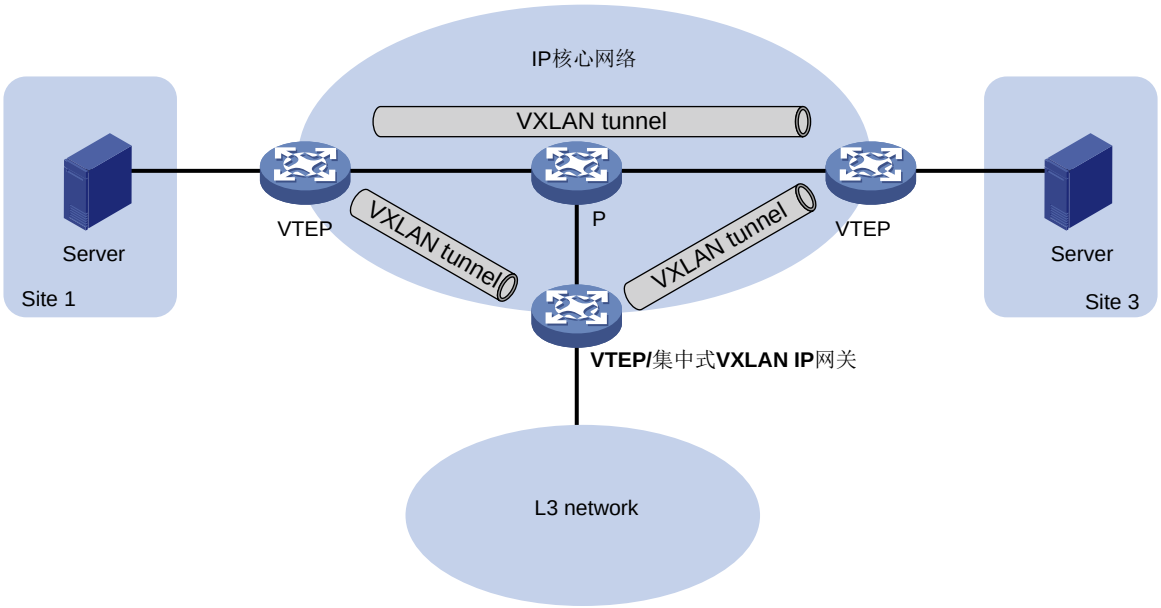
图13 独立的 VXLAN IP 网关示意图



2.5.2 集中式 VXLAN IP 网关

如图 14 所示，集中式 VXLAN IP 网关进行二层 VXLAN 业务终结的同时，还对内层封装的 IP 报文进行三层转发处理。与独立的 VXLAN IP 网关相比，该方式除了能够节省设备资源外，VXLAN IP 网关功能由 VXLAN 对应的三层虚接口（VSI 虚接口）承担，三层业务的部署和控制也更加灵活和方便。

图14 集中式 VXLAN IP 网关示意图



如图 15 所示，以地址为 10.1.1.11 的虚拟机为例，虚拟机与外界网络进行三层通信的过程为：

(1) 虚拟机（10.1.1.11）跨网段进行三层通信时，先广播发送 ARP 请求消息，解析 VXLAN IP 网关（10.1.1.1）的 MAC 地址。

(2) VTEP 1 收到 ARP 请求消息后，添加 VXLAN 封装并发送给所有的远端 VTEP。

(3) VTEP 3 解封装 VXLAN 报文后，发现 ARP 请求的目的 IP 为 VXLAN 对应的本地网关 IP 地址，即与 VXLAN 关联的 VSI 虚接口的 IP 地址，则学习 10.1.1.11 的 ARP 信息，并向虚拟机回应 ARP 应答消息。

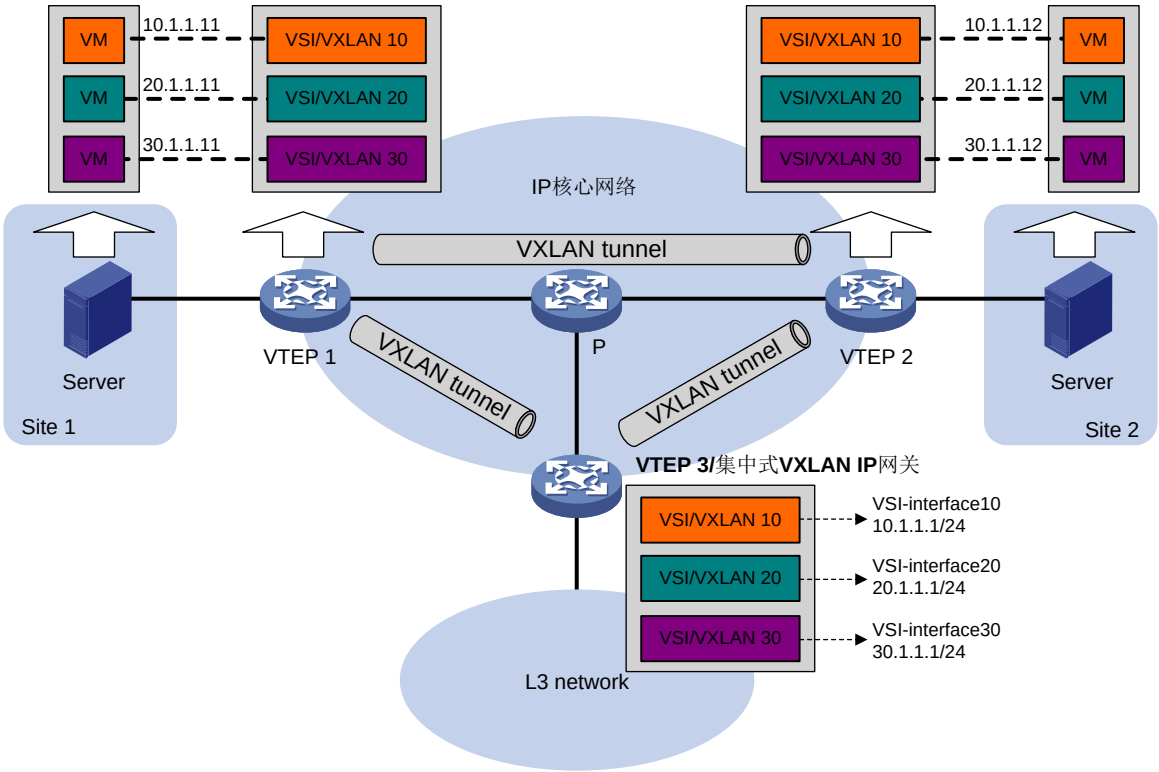
(4) VTEP 1 收到 ARP 应答消息后，将该消息转发给虚拟机。

(5) 虚拟机获取到网关的 MAC 地址后，为三层报文添加网关的 MAC 地址，通过 VXLAN 网络将二层数据帧发送给 VTEP 3。

(6) VTEP 3 解封装 VXLAN 报文，并去掉链路层头后，对内层封装的 IP 报文进行三层转发，将其发送给最终的目的节点。

(7) 目的节点回复的报文到达网关后，网关根据已经学习到的 ARP 表项，为报文封装链路层头，并通过 VXLAN 网络将其发送给虚拟机。

图15 集中式 VXLAN IP 网关的三层通信过程

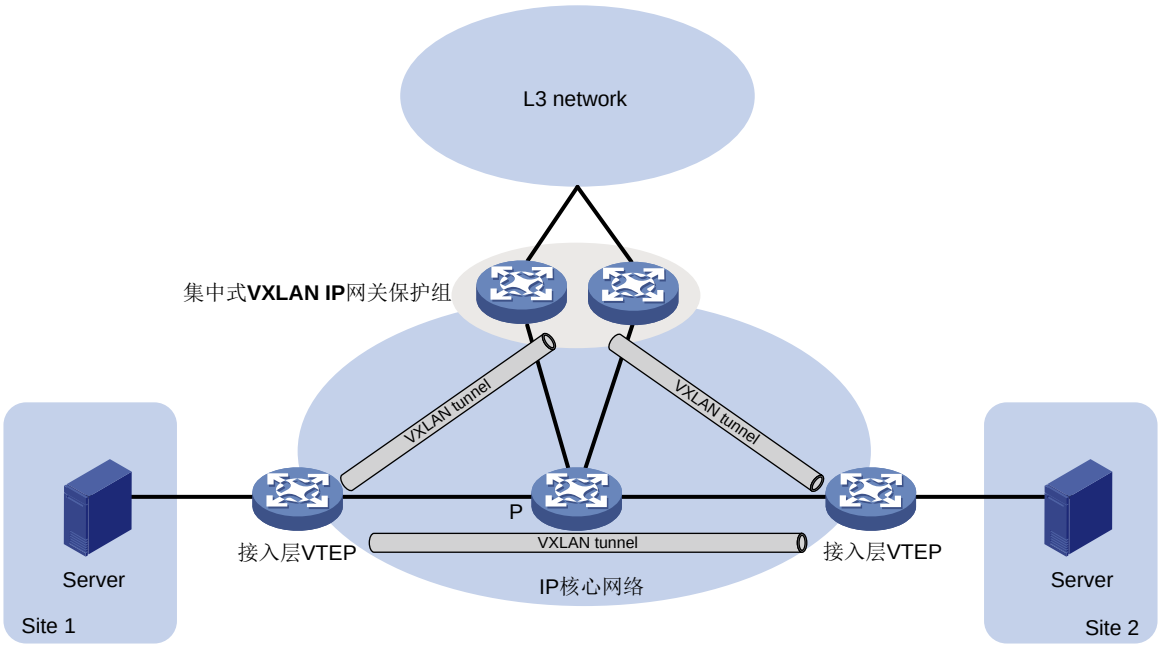


属于不同 VXLAN 网络的虚拟机之间的通信过程与上述过程类似，不同之处在于一个 VXLAN 网络的集中式网关需要将报文转发给另一个 VXLAN 网络的集中式网关，再由该集中式网关将报文转发给本 VXLAN 内对应的虚拟机。

2.5.3 集中式 VXLAN IP 网关保护组

由单台设备承担站点内大量虚拟机的集中式 VXLAN IP 网关功能，对设备的处理资源占用较高，并且对于网关的单点故障没有保护措施。通过集中式 VXLAN IP 网关保护组，可以实现多台设备同时承担网关功能，在提供单点故障保护机制的同时，还可以实现上下行流量的负载分担。

图16 集中式 VXLAN IP 网关保护组示意图



如图 16 所示，两台集中式 VXLAN IP 网关形成保护组，两台设备上存在相同的 VTEP IP，称为保护组的 VTEP IP。接入层 VTEP 与保护组的 VTEP IP 建立 VXLAN 隧道，将虚拟机发送至其它网络的报文转发至保护组，保护组中的两台网关设备均可以接收并处理虚拟机发往其它网络的流量。

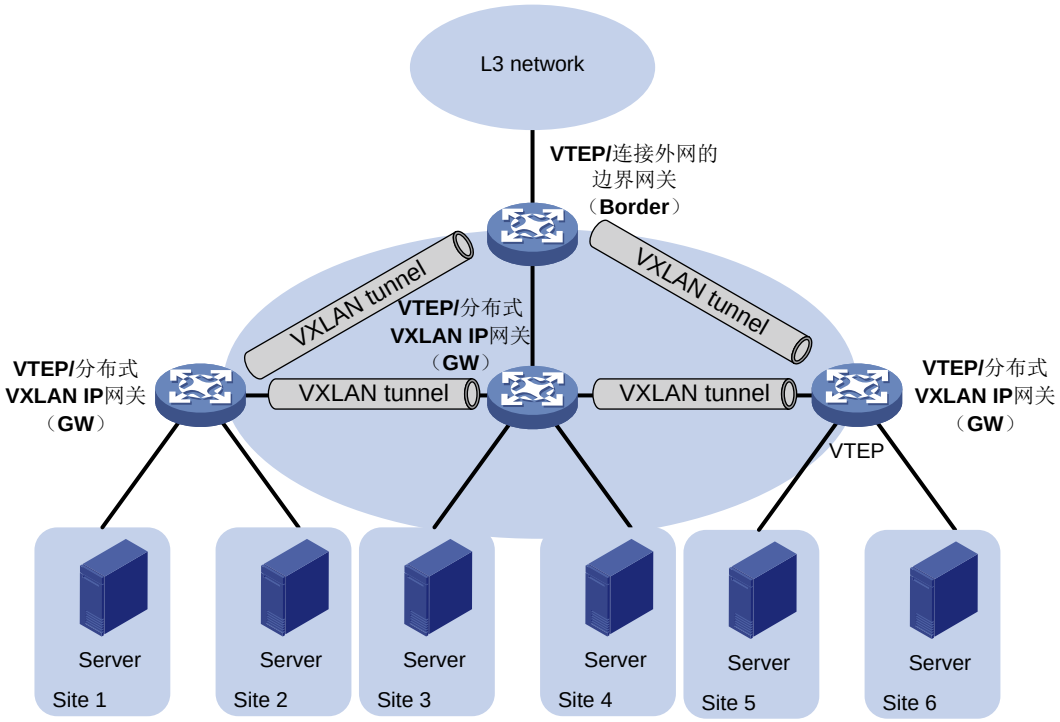
在接入层 VTEP 上，该 VTEP 会与保护组中每个成员 VTEP 的自身 IP 地址自动建立 VXLAN 隧道，泛洪流量（组播、广播和未知单播）通过该隧道转发给所有的成员 VTEP，以确保成员 VTEP 上的表项信息一致。

2.5.4 分布式 VXLAN IP 网关

1. 简介

采用集中式 VXLAN IP 网关方案时，不同 VXLAN 之间的流量以及 VXLAN 访问外界网络的流量全部由集中式 VXLAN IP 网关处理，网关压力较大，并加剧了网络带宽资源的消耗。如图 17 所示，在分布式 VXLAN IP 网关方案中，每台 VTEP 设备都可以作为 VXLAN IP 网关，对本地站点的流量进行三层转发，很好地缓解了网关的压力。

图17 分布式 VXLAN IP 网关示意图

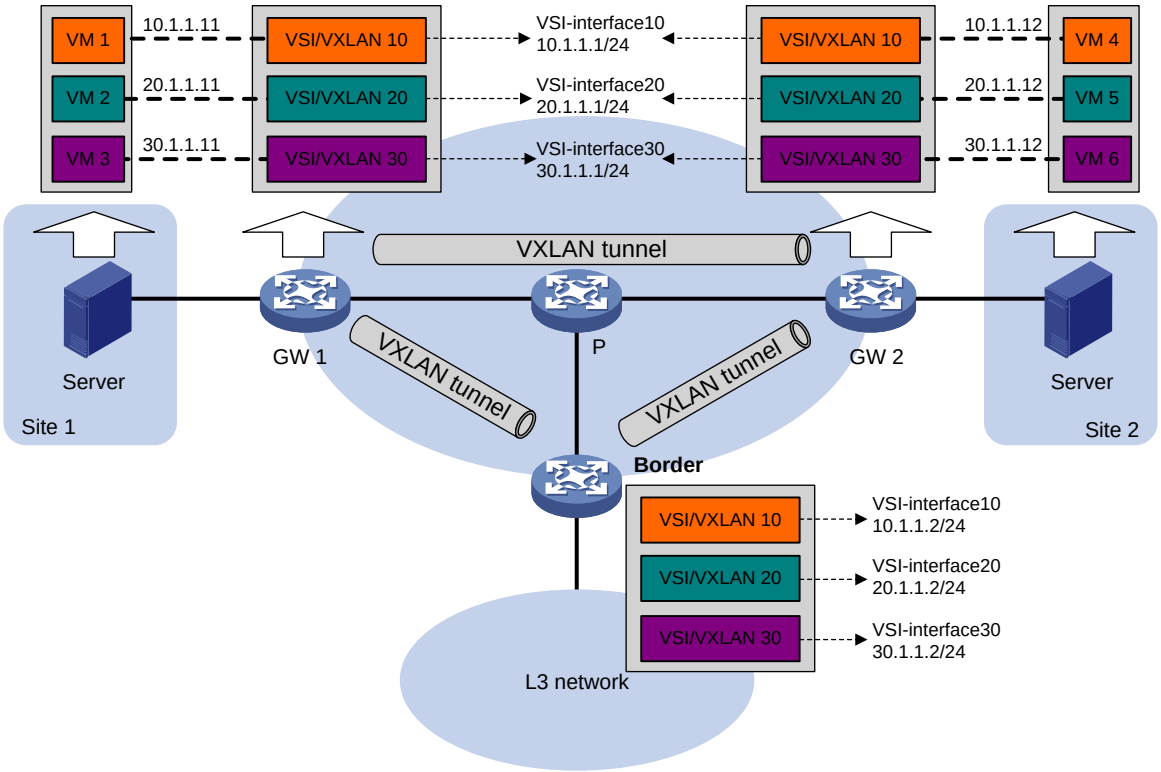


如图 18 所示，在分布式 VXLAN IP 网关组网中，所有的分布式 VXLAN IP 网关（GW）上都需要创建 VSI 虚接口，并为不同 GW 上的相同 VSI 虚接口配置相同的 IP 地址，作为 VXLAN 内虚拟机的网关地址。边界网关（Border）上也需要创建 VSI 虚接口，并配置 IP 地址。在分布式 VXLAN IP 网关上还需要开启以下功能中的一种：

- ARP/ND 泛洪抑制功能：开启本功能后，二层流量查找 MAC 地址表进行转发，三层流量查找 ARP/ND 表项进行转发。
- 本地代理 ARP 功能或本地 ND 代理功能：开启本功能后，所有流量都通过查找 ARP 表项或 ND 表项进行三层转发。下文均以此功能为例，介绍分布式 VXLAN IP 网关中的通信过程。

网关可以通过多种方式生成 ARP 表项和 ND 表项，下文以根据 ARP 协议和 ND 协议动态学习表项来介绍分布式 VXLAN IP 网关中的通信过程。

图18 分布式 VXLAN IP 网关部署示意图



2. 相同 VXLAN 内不同站点的虚拟机通信过程

如图 18 所示，以 VM 1 访问 VM 4 为例，相同 VXLAN 内不同站点的虚拟机的通信过程为：

(1) VM 1 广播发送 ARP 请求消息，获取 VM 4 的 MAC 地址。

(2) GW 1 收到 ARP 请求消息后，学习 VM 1 的 ARP 信息，并代理应答该 ARP 请求，即：向 VM 1 发送 ARP 应答消息，应答的 MAC 地址为 VSI 虚接口 10 的 MAC 地址。

(3) VM 1 学习到 VM 4 的 MAC 地址为 GW 1 上 VSI 虚接口 10 的 MAC 地址。

(4) GW 1 将接收到的 ARP 请求消息中的源 MAC 地址修改为 VSI 虚接口 10 的 MAC 地址，在 VXLAN 10 内向本地站点和远端站点广播发送该 ARP 请求。

(5) GW 2 对 VXLAN 报文进行解封装后，学习 VM 1 的 ARP 信息（IP 为 10.1.1.11、MAC 为 GW 1 上 VSI 虚接口 10 的 MAC、出接口为接收该 VXLAN 报文的 Tunnel 接口），并将 ARP 请求消息中的源 MAC 修改为本地 VSI 虚接口 10 的 MAC 地址，在 VXLAN 10 的本地站点内进行广播。

(6) VM 4 收到 ARP 请求后，学习 VM 1 的 ARP 信息（IP 为 10.1.1.11、MAC 为 GW 2 上 VSI 虚接口 10 的 MAC），并发送 ARP 应答消息给本地网关 GW 2。

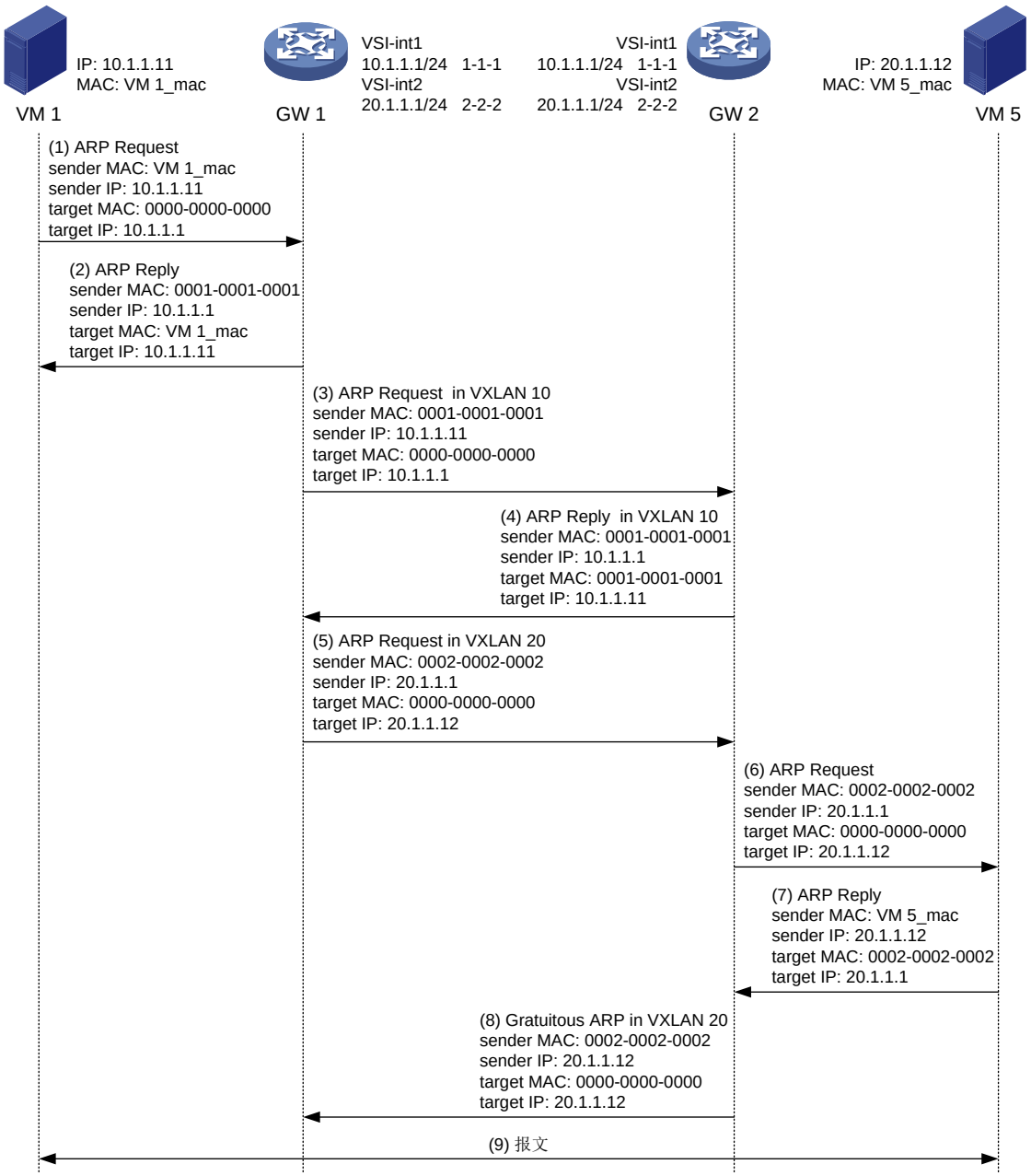
(7) GW 2 从 VM 4 收到 ARP 应答消息后，学习 VM 4 的 ARP 信息，将 ARP 应答消息中的源 MAC 修改为本地 VSI 虚接口 10 的 MAC 地址，并根据已经学习到的 ARP 表项，为 ARP 应答消息添加 VXLAN 封装后发送给 GW 1。

(8) GW 1 对 VXLAN 报文进行解封装后，根据收到的 ARP 应答消息学习 VM 4 的 ARP 信息（IP 为 10.1.1.12、MAC 为 GW 2 上 VSI 虚接口 10 的 MAC、出接口为接收该 VXLAN 报文的 Tunnel 接口）。

(9) 通过上述步骤完成 ARP 信息的学习后，VM 1 发送给 VM 4 的报文，根据已经学习到的 ARP 信息进行转发：首先发送给 GW 1；GW 1 对其进行 VXLAN 封装后，将其发送给 GW 2；GW 2 解封装后，将其发送给 VM 4。

3. 不同 VXLAN 间不同站点的虚拟机通信过程

图19 不同 VXLAN 间不同站点的虚拟机通信过程示意图



如图 19 所示，以 VM 1（VXLAN 10）访问 VM 5（VXLAN 20）为例，不同 VXLAN 的虚拟机的通信过程为：

(1) VM 1 广播发送 ARP 请求消息，获取网关 10.1.1.1 的 MAC 地址。

(2) GW 1 收到 ARP 请求消息后，学习 VM 1 的 ARP 信息，并向 VM 1 发送 ARP 应答消息，应答的 MAC 地址为 VSI 虚接口 10 的 MAC 地址。这样，VM 1 会将访问 VM 5 的报文发送给 GW 1。

(3) GW 1 在 VXLAN 10 内向本地站点和远端站点广播发送 ARP 请求。ARP 请求消息中的源 IP 地址为 10.1.1.11、源 MAC 地址为本地 VSI 虚接口 10 的 MAC 地址。

(4) GW 2 从 VXLAN 隧道上接收到 VXLAN 报文，对其进行解封装后，学习 VM 1 的 ARP 信息（IP 为 10.1.1.11、MAC 为 GW 1 上 VSI 虚接口 10 的 MAC、出接口为接收该 VXLAN 报文的 Tunnel 接口），并将 ARP 请求消息中的源 MAC 修改为本地 VSI 虚接口 10 的 MAC 地址，在 VXLAN 10 的本地站点内广播该 ARP 请求消息。GW 2 发送 ARP 应答消息（IP 为 10.1.1.1、MAC 为 GW 2 上 VSI 虚接口 10 的 MAC）给 GW 1。

(5) GW 1 在 VXLAN 10 内发送 ARP 请求的同时，也会在 VXLAN 20 内向本地站点和远端站点广播发送 ARP 请求，获取 VM 5 的 MAC 地址。ARP 请求消息中的源 IP 地址为 20.1.1.1、源 MAC 地址为本地 VSI 虚接口 20 的 MAC 地址。

(6) GW 2 从 VXLAN 20 内收到 ARP 请求后，将 ARP 请求消息中的源 MAC 修改为本地 VSI 虚接口 20 的 MAC 地址，在 VXLAN 20 的本地站点内广播该 ARP 请求消息。

(7) VM 5 收到 ARP 请求后，学习 GW 2 的 ARP 信息（IP 为 20.1.1.1、MAC 为 GW 2 上 VSI 虚接口 20 的 MAC），并发送 ARP 应答消息给本地网关 GW 2。

(8) GW 2 从 VM 5 收到 ARP 应答消息后，学习 VM 5 的 ARP 信息，并向本地站点和远端站点发送免费 ARP。免费 ARP 消息中的源 IP 地址为 20.1.1.12、源 MAC 地址为本地 VSI 虚接口 20 的 MAC 地址。GW 1 从 VXLAN 隧道上接收到 VXLAN 报文，对其进行解封装后，根据收到的免费 ARP 消息学习 VM 5 的 ARP 信息（IP 为 20.1.1.12、MAC 为 GW 2 上 VSI 虚接口 20 的 MAC、出接口为接收该 VXLAN 报文的 Tunnel 接口）。

(9) 通过上述步骤完成 ARP 信息的学习后，VM 1 发送给 VM 5 的报文，根据已经学习到的 ARP 信息进行转发：首先发送给 GW 1；GW 1 对其进行 VXLAN 封装后，将其发送给 GW 2；GW 2 解封装后，将其发送给 VM 5。

4. 虚拟机与外部网络的三层通信过程

虚拟机要想与外部网络进行三层通信，需要在接入虚拟机的本地分布式 VXLAN IP 网关上指定流量的下一跳为 Border，可以通过如下方式来实现：

- 在本地分布式 VXLAN IP 网关上配置静态路由，指定路由下一跳为 Border 上同一个 VXLAN 对应 VSI 虚接口的 IP 地址。
- 在本地分布式 VXLAN IP 网关上配置策略路由，设置报文的下一跳为 Border 上同一个 VXLAN 对应 VSI 虚接口的 IP 地址。

如图 18 所示，以 VM 1 访问外部网络内的主机 50.1.1.1 为例，虚拟机访问外部网络的三层通信过程为：

(1) VM 1 广播发送 ARP 请求消息，获取网关 10.1.1.1 的 MAC 地址。

(2) GW 1 收到 ARP 请求消息后，学习 VM 1 的 ARP 信息，并向 VM 1 发送 ARP 应答消息，应答的 MAC 地址为 VSI 虚接口 10 的 MAC 地址。

(3) VM 1 将访问外部网络的报文发送给 GW 1。

(4) GW 1 接收到报文后，根据策略路由判断报文的下一跳地址为 10.1.1.2。GW 1 在 VXLAN 10 内向本地站点和远端站点广播发送 ARP 请求消息，获取 10.1.1.2 对应的 MAC 地址。

(5) Border 对 VXLAN 报文进行解封装，学习 GW 1 的 ARP 信息，并通过 VXLAN 隧道回复 ARP 应答消息。

(6) GW 1 对 VXLAN 报文进行解封装，并获取到 10.1.1.2 的 ARP 信息。

(7) GW 1 根据获取到的信息为 VM 1 发送的报文封装链路层地址（10.1.1.2 对应的 MAC 地址），并通过 VXLAN 隧道将报文发送给 Border。

(8) Border 对接收到的报文进行解封装后，对报文进行三层转发。

3 Comware 实现的技术特色

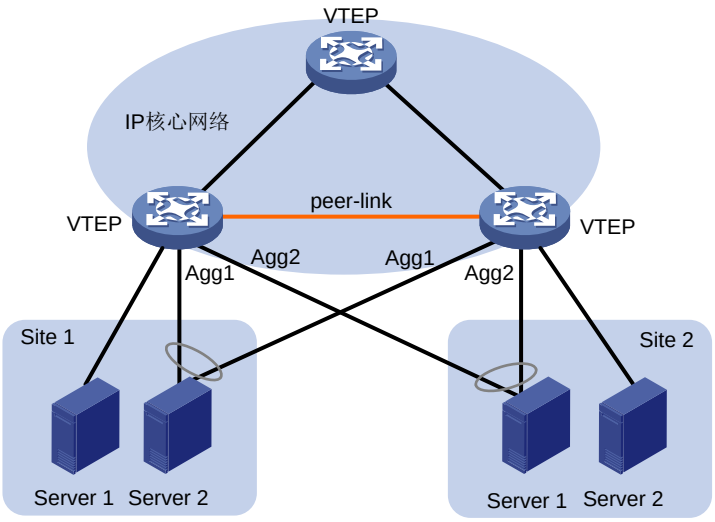
3.1 VXLAN支持M-LAG

说明

目前，本功能仅支持 IPv4 站点网络和 IPv4 核心网络。

VXLAN 利用 M-LAG 功能（Multichassis link aggregation，跨设备链路聚合）将两台物理设备连接起来虚拟成一台设备，使用该虚拟设备作为 VTEP（既可以是仅用于二层转发的 VTEP，也可以是 VXLAN IP 网关），可以避免 VTEP 单点故障对网络造成影响，从而提高 VXLAN 网络的可靠性。

图20 VXLAN 支持 M-LAG 组网图



如图 20 所示，VXLAN 支持 M-LAG 功能的工作机制包括：

- 同步 MAC 地址和 ARP 信息

  作为 M-LAG 成员设备的两台 VTEP 通过 peer-link 接口连接，在 peer-link 链路上同步 MAC 地址和 ARP 信息，以确保两台 VTEP 上的 MAC 地址和 ARP 信息保持一致。peer-link 链路既可以是以太网聚合链路，也可以是 VXLAN 隧道。采用以太网聚合链路作为 peer-link 链路称为直连模式 peer-link 链路，采用 VXLAN 隧道作为 peer-link 链路称为隧道模式 peer-link 链路。

说明

作为 peer-link 链路的 VXLAN 隧道自动与设备上的所有 VXLAN 关联。

- 使用相同的隧道源 IP 地址

  作为 M-LAG 设备的两台 VTEP 使用相同的隧道源 IP 地址，与其他 VTEP 设备建立 VXLAN 隧道。

- 备份双挂 AC 的用户侧链路

  在用户侧，两台 VTEP 均通过以太网链路接入同一台虚拟机，跨设备在两条链路间建立二层聚合接口，将该聚合接口配置为 AC（在聚合接口上创建以太网服务实例、配置报文匹配规则并关联以太网服务实例与 VSI），则该 AC 称为双挂 AC。VXLANM-LAG 组网中采用双挂 AC，来避免单条以太网链路故障导致虚拟机无法访问网络。

  - 采用直连模式 peer-link 链路时，用户侧链路备份机制为：将二层聚合接口配置为 AC 后，VTEP 会在 peer-link 链路上自动创建具有相同报文匹配规则、关联相同 VSI 的 AC。当一台 VTEP 上的 AC 故障后，从 VXLAN 隧道上接收到的、发送给该 AC 的报文将通过 peer-link 链路转发到另一台 VTEP，该 VTEP 根据 peer-link 链路上配置的 AC 判断报文所属 VSI，并转发该报文，从而保证转发不中断。
  - 采用隧道模式 peer-link 链路时，用户侧链路备份机制为：如果一台 VTEP 上的 AC 故障，则该 VTEP 从 VXLAN 隧道上接收到发送给故障 AC 的报文后，为报文添加 VXLAN 封装，封装的 VXLAN ID 为故障 AC 所属 VSI 对应的 VXLAN ID，并通过作为 peer-link 链路的 VXLAN 隧道将其转发到另一台 VTEP。该 VTEP 根据 VXLAN ID 判断报文所属的 VSI，并转发该报文。

- 单挂 AC 互通

  在 VXLAN M-LAG 组网中，组成 M-LAG 系统的两台 VTEP 上 AC 配置可能不一致，若某个 AC 仅连接到其中一台 VTEP，则该 AC 称为单挂 AC。组成 M-LAG 系统的两台 VTEP 下不同单挂 AC 的互通通过 peer-link 链路来实现。

  - 采用直连模式 peer-link 链路时，单挂 AC 互通机制为：将接口配置为单挂 AC 后，VTEP 会在 peer-link 链路上自动创建具有相同报文匹配规则、关联相同 VSI 的 AC。当从单挂 AC 上收到报文后，将通过 peer-link 链路转发到另一台 VTEP，该 VTEP 根据 peer-link 链路上配置的 AC 判断报文所属 VSI，并转发该报文。
  - 采用隧道模式 peer-link 链路时，单挂 AC 互通机制为：当从单挂 AC 上收到报文后，为报文添加 VXLAN 封装，封装的 VXLAN ID 为单挂 AC 所属 VSI 对应的 VXLAN ID，并通过作为 peer-link 链路的 VXLAN 隧道将其转发到另一台 VTEP。该 VTEP 根据 VXLAN ID 判断报文所属的 VSI，并转发该报文。

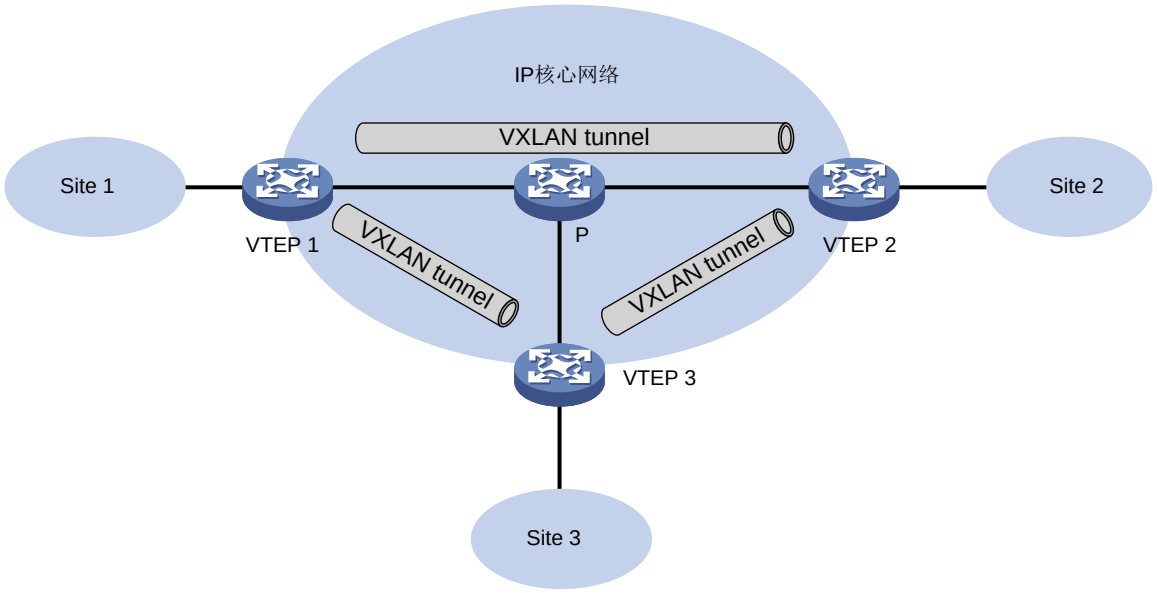
4 典型组网应用

4.1 VXLAN二层互通组网

在 VXLAN 二层互通组网中，属于相同 VXLAN 的虚拟机处于同一个逻辑二层网络，彼此之间二层互通；属于不同 VXLAN 的虚拟机之间二层隔离。接入 VXLAN 网络的租户可以规划自己的虚拟网络，不需要考虑物理网络 IP 地址和广播域的限制，降低了网络管理的难度。

VXLAN 二层互通组网如图 21 所示。VTEP 为 VXLAN 网络的边缘设备；VXLAN Tunnel 为两个 VTEP 之间点到点的逻辑隧道，用于不同 VTEP 之间的流量转发。

图21 VXLAN 二层互通组网示意图

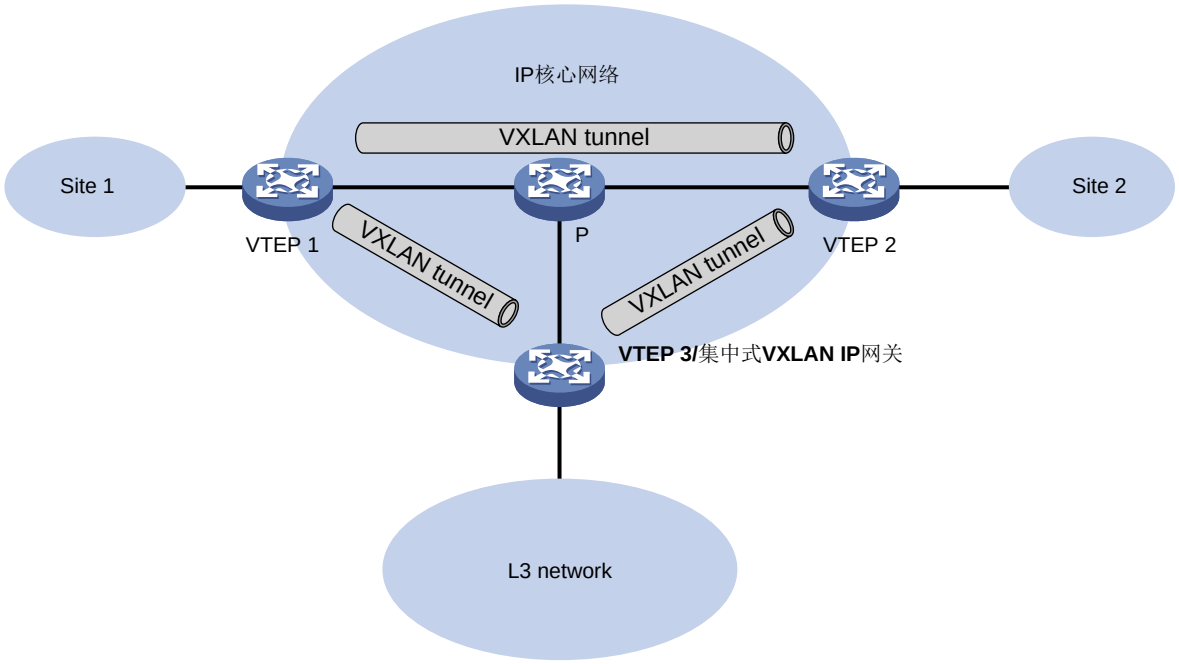


4.2 集中式VXLAN IP网关组网

集中式 VXLAN IP 网关组网中，网关部署在 Spine 设备。集中式网关的优点是流量均会经过 Spine 设备，容易实现流量控制、自动引流等功能。缺点是 Spine 设备处理所有三层流量，压力较大，不适于在大规模网络中部署。

集中式 VXLAN IP 网关的典型组网如图 22 所示。VTEP 为 VXLAN 网络边缘设备，Spine 为与广域网连接的边界网关设备，虚拟机通过 VXLAN 实现不同站点间的二层互联，并通过 VXLAN IP 网关实现与广域网的三层互联。

图22 集中式 VXLAN IP 网关组网示意图

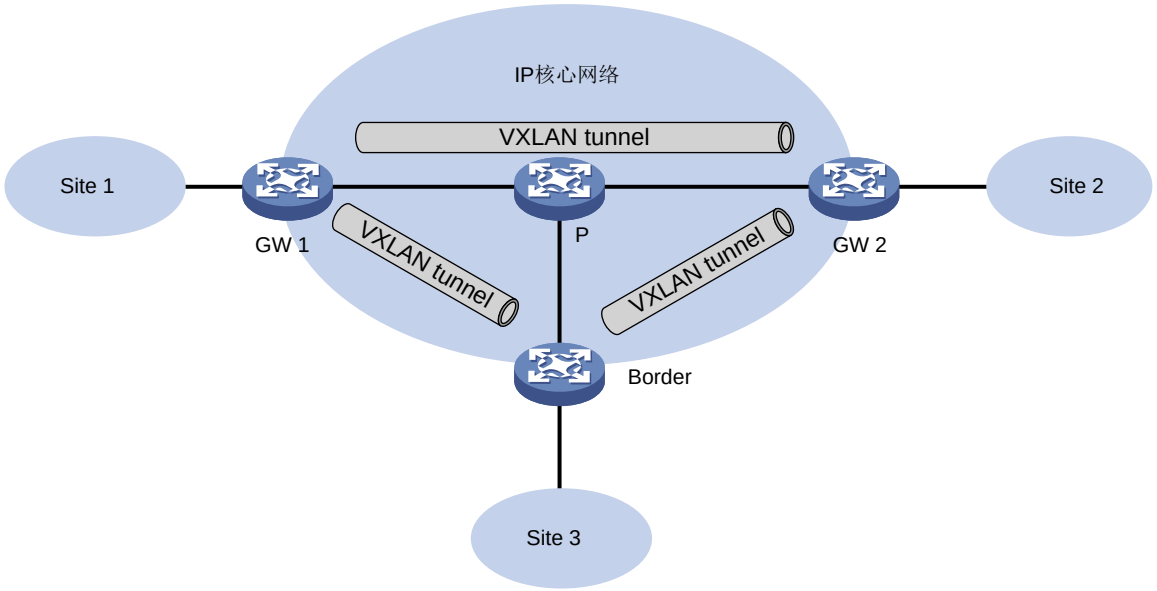


4.3 分布式VXLAN IP网关组网

分布式 VXLAN IP 网关组网中，每台 VTEP 设备都可以作为 VXLAN IP 网关，对本地站点的流量进行三层转发，很好地缓解了网关的压力。

分布式 VXLAN IP 网关的典型组网如图 23 所示。Leaf 为分布式 VXLAN IP 网关设备，Border 为与广域网连接的边界网关设备，虚拟机通过分布式 VXLAN IP 网关实现不同 VXLAN 网络的三层互联，并通过边界网关实现与广域网的三层互联。

图23 分布式 VXLAN IP 网关组网示意图

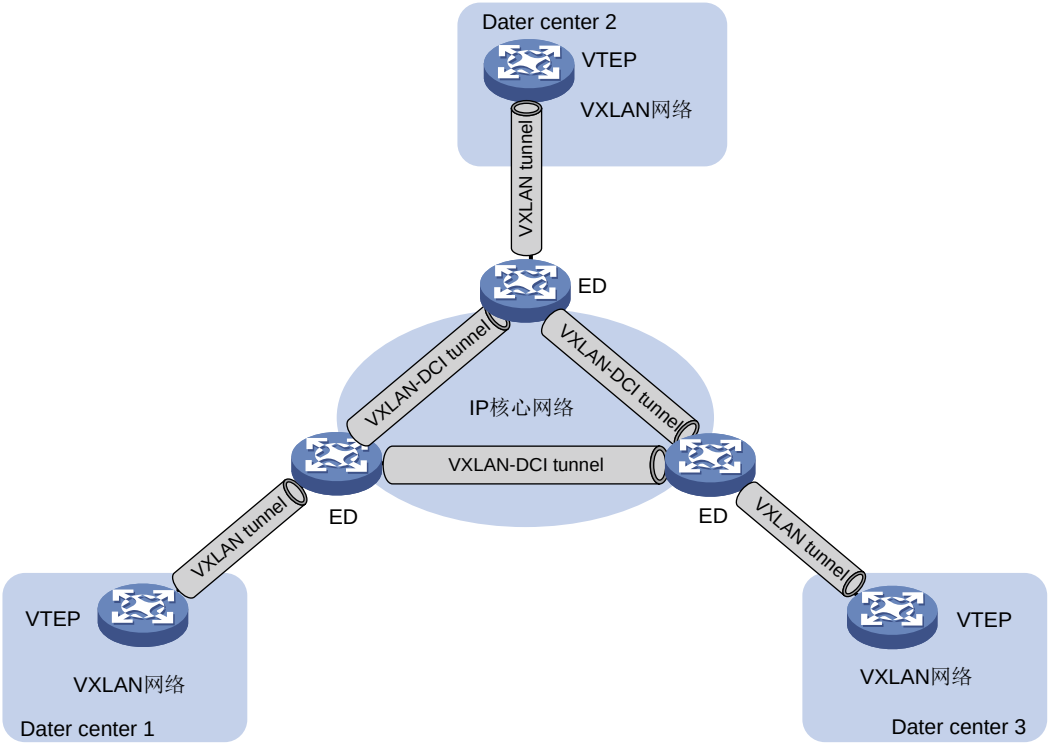


4.4 VXLAN数据中心互联组网

VXLAN 数据中心互联技术通过在数据中心之间建立VXLAN-DCI（VXLAN Data Center Interconnect，VXLAN 数据中心互联）隧道，实现不同数据中心之间虚拟机的互通。

如图 24 所示，数据中心的边缘设备为 ED（Edge Device，边缘设备）。ED 之间建立 VXLAN-DCI 隧道，该隧道采用 VXLAN 封装格式。ED 与数据中心内部的 VTEP 建立 VXLAN 隧道。ED 从 VXLAN 隧道或 VXLAN-DCI 隧道上接收到报文后，解除 VXLAN 封装，根据目的 IP 地址重新对报文进行 VXLAN 封装，并将其转发到 VXLAN-DCI 隧道或 VXLAN 隧道，从而实现跨数据中心之间的互通。

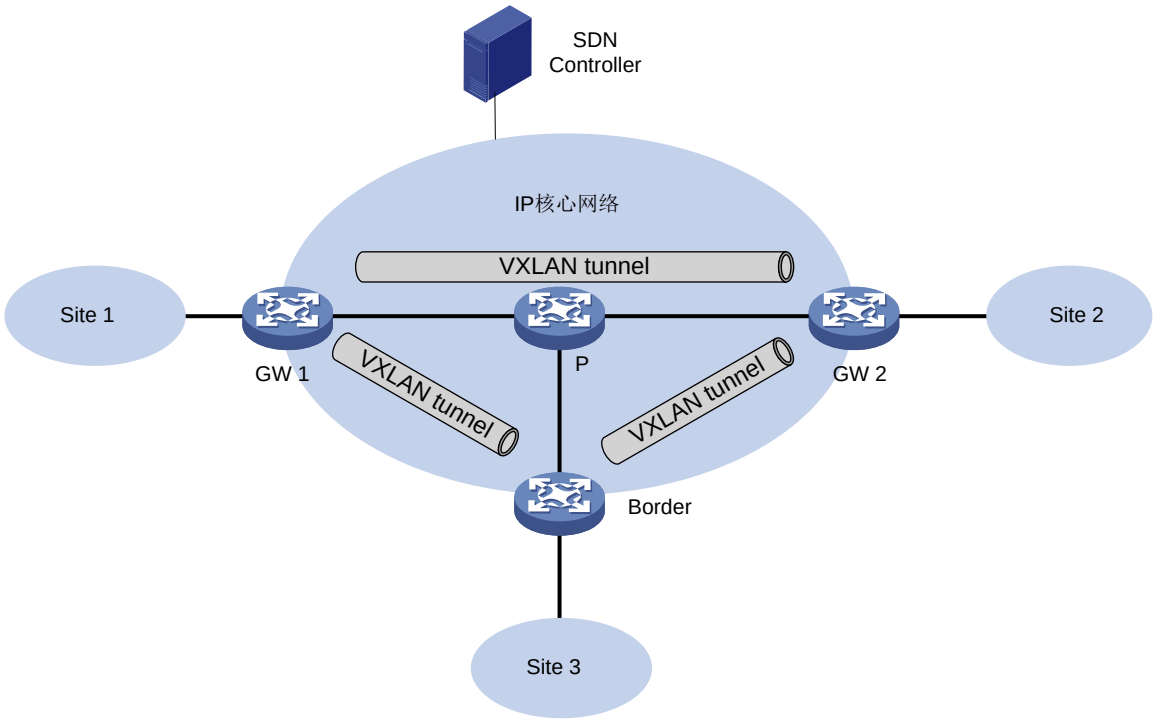
图24 VXLAN 数据中心互联典型组网示意图



4.5 VXLAN与SDN控制器配合组网

SDN（Software Defined Network，软件定义网络）是一种新型的网络架构，它将控制平面与转发平面分离，由 SDN 控制器集中控制和管理整个网络的设备。如图 25 所示，VXLAN 可以与 SDN 控制器配合使用，VXLAN 网络中的所有设备均由 SDN 控制器通过标准协议集中管理，减少了传统设备管理的复杂性。同时，当用户业务扩展时，通过集中管理，用户可以方便快速地部署网络设备，便于网络的扩展和管理。

图25 VXLAN 与 SDN 控制器配合组网示意图



5 参考文献

RFC 7348: Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks