EVPN VXLAN 技术白皮书

Copyright © 2021 紫光恒越技术有限公司版权所有，保留一切权利。
非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。
除紫光恒越技术有限公司的商标外，本手册中出现的其它公司的商标、产品标识及商品名称，由各自权利人拥有。本文中的内容为通用性技术信息，某些信息可能不适用于您所购买的产品。

目 录

1 概述 ····· 1
1.1 产生背景 ····· 1
1.2 技术优点 ····· 1
2 技术实现 ····· 2
2.1 EVPN VXLAN 网络模型 ····· 2
2.2 EVPN VXLAN 控制平面工作机制 ····· 3
2.2.1 BGP EVPN 邻居建立 ····· 4
2.2.2 BGP EVPN 路由 ····· 4
2.2.3 VXLAN 隧道及 BUM 广播表建立 ····· 5
2.2.4 MAC/IP 路由通告与学习 ····· 6
2.2.5 外部路由通告与学习 ····· 7
2.2.6 MAC 地址迁移 ····· 9
2.2.7 ARP 泛洪抑制 ····· 9
2.3 EVPN VXLAN 数据平面工作机制 ····· 10
2.3.1 二层流量转发 ····· 10
2.3.2 集中式网关转发 ····· 11
2.3.3 分布式网关对称 IRB 转发 ····· 12
2.4 EVPN VXLAN 多归属 ····· 15
2.4.1 功能简介 ····· 15
2.4.2 DF 选举 ····· 15
2.4.3 水平分割 ····· 16
2.4.4 冗余备份模式 ····· 17
2.4.5 Aliasing ····· 17
2.5 EVPN VXLAN 支持组播 ····· 18
2.5.1 功能简介 ····· 18
2.5.2 单归属站点组播 ····· 18
2.5.3 多归属站点组播 ····· 18
3 典型组网应用 ····· 19
3.1 EVPN 分布式网关组网 ····· 19
3.2 EVPN 数据中心互联组网 ····· 20
3.3 EVPN 与 SDN 控制器配合组网 ····· 20

4 参考文献 21

1 概述

1.1 产生背景

随着数据中心业务日益增加，用户需求不断提高，数据中心的规模和功能日趋复杂，管理难度也越来越高。出于灾备，企业分支机构的多地部署，提升资源利用率等方面的考虑，企业可能在不同的物理站点部署自己的数据中心网络。于是，如何将这些数据中心站点互联起来，并降低数据中心的管理成本、灵活扩充数据中心业务等就成为企业数据中心的重要任务。

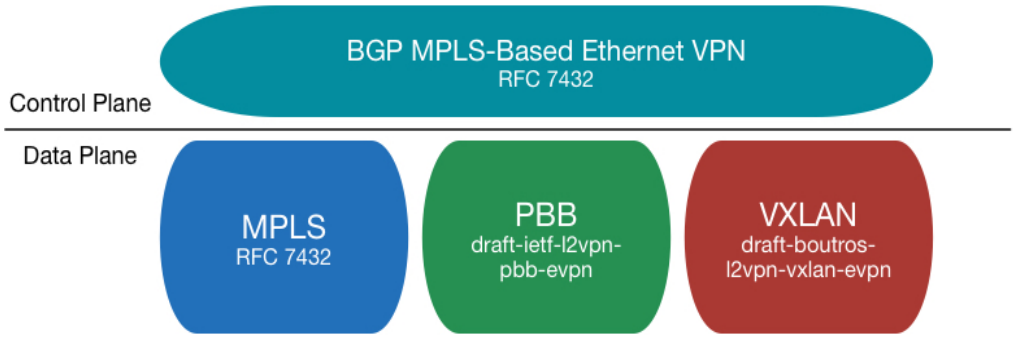
EVPN（Ethernet Virtual Private Network，以太网虚拟专用网络）是一种基于 Overlay 技术的二层网络互联技术，具有部署简单、扩展性强等优点。EVPN 采用 MP-BGP 协议通告 MAC/IP 的可达性和组播等信息，通过生成的 MAC 表项和路由表项进行二/三层报文转发，以实现二层网络互联的目的，很好地满足了用户对于大型数据中心网络的需求。

目前，EVPN 不仅广泛应用于数据中心网络，在园区接入网络、广域网中也具有一定的应用。

1.2 技术优点

EVPN 定义了一套通用的控制平面，数据平面可以使用不同的封装技术，他们的关系如图 1 所示。目前，Comware 支持 VXLAN 和 MPLS 作为数据平面。本文仅介绍 VXLAN 作为数据平面的 EVPN 技术，即 EVPN VXLAN。

图1 EVPN 协议框架



EVPN VXLAN 继承了 MP-BGP 和 VXLAN 的优势。具有如下特点：

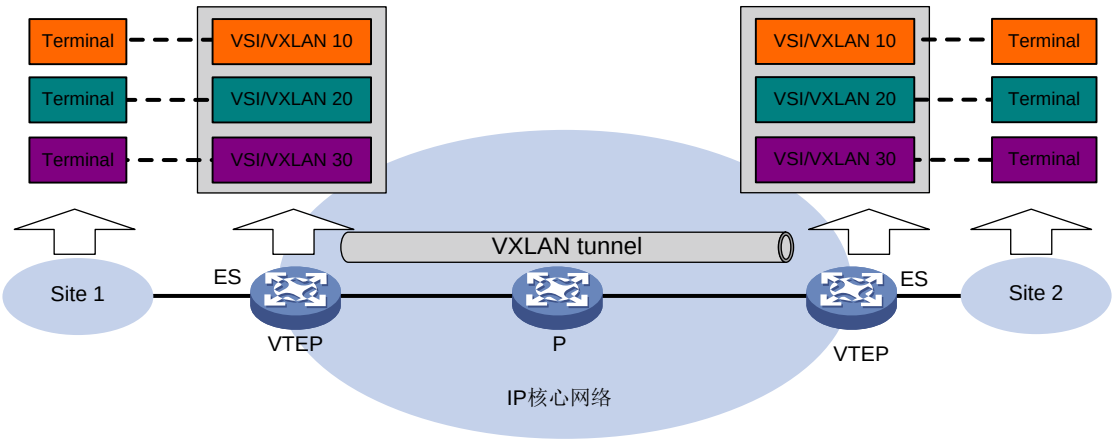
- 简化配置：通过 MP-BGP 实现 VTEP 自动发现、VXLAN 隧道自动建立、VXLAN 隧道与 VXLAN 自动关联，无需用户手工配置，降低网络部署难度。
- 分离控制平面与数据平面：控制平面负责发布路由信息，数据平面负责转发报文，分工明确，易于管理。
- 支持多归属：当同一个站点通过多台 VTEP 接入 VXLAN 网络时，连接该站点的多条路径均可以进行流量转发，以提高网络带宽利用率。
- 支持对称 IRB（Integrated Routing and Bridging，集成的路由和桥接）：MP-BGP 同时发布二层 MAC 地址和三层路由信息，VTEP 既可以进行二层转发，也可以进行三层路由。这样，不仅可以保证流量采用最优路径转发，还可以减少广播流量。

- 支持组播：通过 EVPN 路由传递组播表项，使得组播流量可以在 VTEP 之间正确转发。单归属和多归属组网中，均支持 EVPN 组播转发。

2 技术实现

2.1 EVPN VXLAN网络模型

图2 EVPN VXLAN 网络模型示意图



如图 2 所示，EVPN VXLAN 的典型网络模型中包括如下几部分：

- 用户终端(Terminal)：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

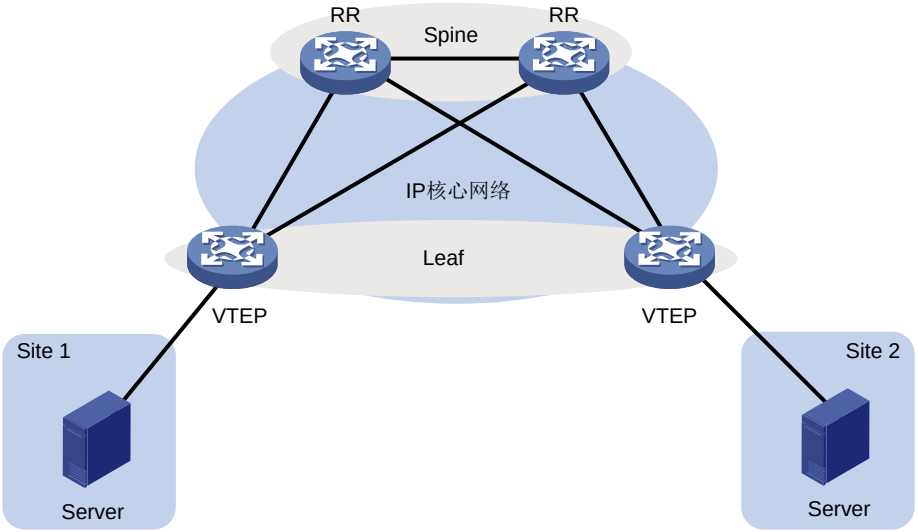
本文档中如无特殊说明，均以 VM 为例介绍 EVPN VXLAN 工作机制。采用其他类型用户终端时，EVPN VXLAN 工作机制与 VM 相同，不再赘述。

- VTEP（VXLAN Tunnel End Point，VXLAN 隧道端点）：EVPN VXLAN 的边缘设备。EVPN VXLAN 的相关处理都在 VTEP 上进行。根据 VTEP 功能，VTEP 可以划分为 L2 VTEP 和 GW 两种角色：
  - L2 VTEP：只支持二层 VXLAN 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。EVPN VXLAN 网络根据 GW 的部署方式，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 2 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/EVPN 实例：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，也称为 EVPN 实例。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。VSI 与 VXLAN 一一对应。
- ES（Ethernet Segment，以太网段）：用户站点连接到 VTEP 的链路，通过 ESI（Ethernet Segment Identifier，以太网段标识符）唯一标识。当一个站点通过多条链路接入到 EVPN VXLAN 网络时，这些链路构成一个 ES，以实现主备备份或负载分担。

如图 3 所示，EVPN VXLAN 通常采用 Spine（核心）—Leaf（分支）的分层结构。Leaf 层的设备作为 VTEP 对报文进行 EVPN 相关处理，Spine 层为核心设备，根据报文的目的 IP 地址转发报文。EVPN VXLAN 网络中的设备属于同一个 AS（Autonomous System，自治系统）时，为了避免在所有 VTEP 之间建立 IBGP 对等体，可以将核心设备配置为 RR（Route Reflector，路由反射器），以减轻网络的部署难度。通常情况下，在集中式网关组网中，VTEP 为 L2 VTEP，其中一台 RR 同时作为 GW；在分布式网关组网中，VTEP 作为 GW，RR 仅作为反射器发布、接收 EVPN 路由，不需要封装、解封装 VXLAN 报文。

图3 EVPN VXLAN 分层组网模型



2.2 EVPN VXLAN控制平面工作机制

在 EVPN VXLAN 中，VTEP 之间的 MAC/IP 地址学习是基于控制平面的，采用 MP-BGP 协议通告 MAC/IP 的可达性，其策略控制与 MPLS L3VPN 相似。这种基于控制平面的学习能够对 MAC/IP 的学习过程提供更强的控制能力，因此具有以下优势：

- 具有较好的扩展性。
- 能够维护主机或虚拟机彼此间的隔离性。
- 解决了设备多归属或网络多归属接入时的负载分担问题，并缩短了网络出现故障时的收敛时间。

2.2.1 BGP EVPN 邻居建立

BGP 新增了 EVPN 子地址族，用于协商并建立 BGP EVPN 邻居。EVPN 子地址族使用的地址族编号为：AFI=25，SAFI=70。

在 EVPN VXLAN 组网中，VTEP 之间既可以建立 IBGP 邻居，也可以建立 EBGP 邻居。

- 建立 IBGP 邻居时，为简化全连接配置，需要部署 RR 反射器。所有 VTEP 都只和 RR 建立 BGP 邻居关系。RR 发现并接收 VTEP 发起的 BGP 连接后形成客户机列表，将从某个 VTEP 收到的路由反射给其他所有的 VTEP。RR 的部署位置可以是 Spine，也可以是 Leaf，还可以是独立出来的设备。
- 建立 EBGP 邻居时，不需要部署 RR。BGP 自动将从 EBGP 邻居收到的 EVPN 消息发送给其他 EBGP 和 IBGP 邻居。

2.2.2 BGP EVPN 路由

为了支持 EVPN，BGP EVPN 地址族新增了如下几种 EVPN NLRI（Network Layer Reachability Information，网络层可达性信息），即 EVPN 路由：

- Ethernet Auto-discovery Route（RT-1）：以太网自动发现路由，用来在站点多归属组网中通告 ES 信息，以便实现水平分割、Aliasing 和主备备份等特性。
- MAC/IP Advertisement Route（RT-2）：MAC/IP 发布路由，用来通告 MAC 地址和主机路由信息（即 ARP 信息和 ND 信息）。
- Inclusive Multicast Ethernet Tag Route（RT-3）：包含性组播以太网标签路由，又称为 IMET 路由，用来通告 VTEP 及其所属 VXLAN 信息，以实现自动发现 VTEP、自动建立 VXLAN 隧道和自动关联 VXLAN 与 VXLAN 隧道。
- Ethernet Segment Route（RT-4）：以太网段路由，用来通告 ES 及其连接的 VTEP 信息，以便发现连接同一 ES 的 VTEP 冗余组其他成员，以及在冗余组之间选举指定转发器 DF 等。
- IP Prefix Advertisement Route（RT-5）：IP 前缀路由，用来以 IP 前缀的形式通告引入的外部路由。
- Selective Multicast Ethernet Tag Route：选择性组播以太网标签路由，用来通告租户的 IGMP 组播组信息。
- IGMP Join Synch Route：IGMP 加入同步路由，用来在多归属成员间同步租户的 IGMP 加入组播组信息。
- IGMP Leave Synch Route：IGMP 离开同步路由，用来在多归属成员间通告租户的 IGMP 离开组播组信息，以撤销相应的 IGMP 加入同步路由。

EVPN 路由中包含 RD（Route Distinguisher，路由标识符）字段，用来区分不同 VXLAN 的 EVPN 路由，以免 EVPN 路由冲突。

VTEP 在发布携带 EVPN 路由的 BGP Update 消息时，还需要携带 VPN Target 扩展团体属性（也称为 Route Target），通过 VPN Target 属性来控制 EVPN 路由信息的发布与接收。VPN target 属性分为以下两种，每一种都可以包括多个属性值：

- Export target 属性：本地 VTEP 在通过 BGP 的 Update 消息将 EVPN 路由发送给远端 VTEP 时，将 Update 消息中携带的 VPN target 属性设置为 Export target。

- Import target 属性：VTEP 收到其它 VTEP 发布的 Update 消息时，将消息中携带的 VPN target 属性与本地配置的 Import target 属性进行比较，只有二者中存在相同的属性值时，才会接收该消息中的 EVPN 路由。

2.2.3 VXLAN 隧道及 BUM 广播表建立

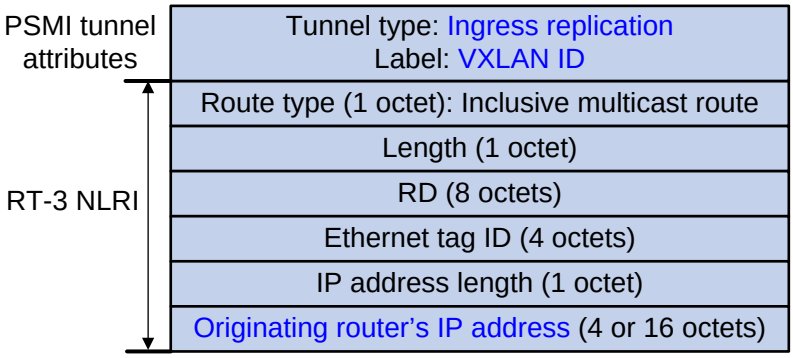
VXLAN 采用“MAC in UDP”封装，是一种在 IP 网络基础之上构建 Overlay 网络的技术。在 IP 网络上传输报文时，VXLAN 使用 Ingress Replication，即头端复制，来转发 BUM（Broadcast/Unknown unicast/Unknown Multicast，广播/未知单播/未知组播）流量。所谓头端复制，是指在 VXLAN 转发实体（VSI）中保存 BUM 流量需要通过哪些 VXLAN 隧道复制到远端 PE 设备，此 VXLAN 隧道列表叫做 BUM 广播表。

EVPN VXLAN 可以通过以下两种方式建立 VXLAN 隧道和 BUM 广播表：

- 在二层转发时，EVPN VXLAN 依靠 RT-3（Inclusive Multicast Ethernet Tag Route）自动发现 VTEP 站点、建立 VXLAN 隧道、建立 BUM 广播表。

  RT-3 路由的关键信息及路由格式如图 4 所示。每个 VTEP 都通过 RT-3 通告自己所属的 VXLAN ID 及自身的 IP 地址。这样，每个 VTEP 设备都有全网的 VXLAN 信息以及 VXLAN 和下一跳的关系。VTEP 设备会和那些跟自己具有相同 VXLAN 的下一跳自动建立 VXLAN 隧道，并将此 VXLAN 隧道与 VXLAN 关联。于是，对于每个 VXLAN 而言，所有这些建立并关联的 VXLAN 隧道就形成 BUM 广播表。

  图4 RT-3 路由消息格式



- 在分布式网关进行三层转发时，EVPN VXLAN 依靠 RT-2 或 RT-5 自动发现 VTEP 站点、建立 VXLAN 隧道。

  当分布式网关接收到远端网关通告的 RT-2 或 RT-5 路由，且该路由携带的 Export target 属性与本地某个 VPN 实例的 Import target 属性匹配时，本地 VTEP 会与远端 VTEP 建立 VXLAN 隧道，并将该 VXLAN 隧道与 VPN 实例对应的 L3VNI（Layer 3 VNI，三层 VXLAN ID）关联。此隧道用于三层转发时对报文进行封装。分布式网关的详细介绍，请参见“2.3.3 分布式网关对称 IRB 转发”。

如果通过上述两种方式发现同一个远端 VTEP，则只建立一条隧道，该隧道与不同的 VXLAN 关联，同时用于二层转发和三层转发，即两个 VTEP 之间最多只会建立一条 VXLAN 隧道。

2.2.4 MAC/IP 路由通告与学习

EVPN VXLAN 在控制平面学习 MAC 地址和 ARP/ND 信息。站点的 MAC 地址和 ARP/ND 信息通过 EVPN 的 MAC/IP 发布路由（即 RT-2，二类路由）通告。因此，在 EVPN VXLAN 网络中，不需要将 ARP/ND 请求泛洪到网络中。

RT-2 路由格式如图 5 所示。

图5 RT-2 路由格式

Route type (1 octet): MAC/IP advertisement route
Length (1 octet)
RD (8 octets)
Ethernet segment identifier (10 octets)
Ethernet tag ID (4 octets)
MAC address length (1 octet)
MAC address (6 octets)
IP address length (1 octet)
IP address (0, 4, or 16 octets)
L2VNI (3 octets)
L3VNI (0 or 3 octets)

如图 6 所示，MAC 地址和 ARP/ND 信息的通告和学习过程为：

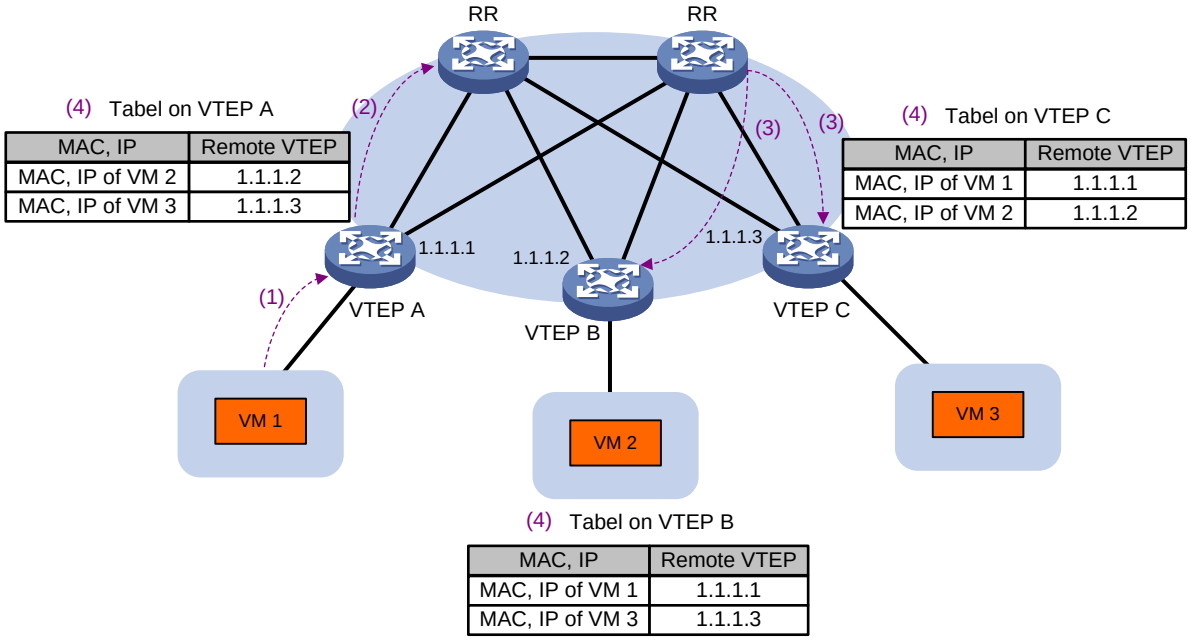
(1) VTEP 在数据平面完成本地 MAC 地址和 ARP/ND 信息的学习。本地 MAC 地址通过以太网报文的源 MAC 地址学习获得；ARP/ND 信息通过 ARP、免费 ARP、ND 等报文学习获得。

(2) VTEP 学习到本地 MAC 地址和 ARP/ND 信息后，在控制平面通过 BGP EVPN 的 RT-2 路由将该信息发布给 RR。

(3) RR 将接收到的 RT-2 路由同步给所有 BGP EVPN 邻居（远端 VTEP）。

(4) 远端 VTEP 接收到 RT-2 路由后，将 MAC 地址添加到 MAC 地址转发表，将 ARP/ND 信息添加到 ARP/ND 表和路由表。

图6 MAC/IP 路由通告与学习过程



在发布 RT-2 路由时，VTEP 可以选择是否携带 IP。为了抑制 ARP 请求泛洪到网络中，通常需要携带 IP，以便让远端 VTEP 学习到本端 VTEP 下挂的主机 ARP，使得远端 VTEP 可以直接代答回应远端主机发起的 ARP 请求。如果只是纯二层网络、不进行三层转发，则在 RT-2 中只携带 MAC 地址。由于在三层转发环境下远端 VTEP 能够从 ARP 信息中获取 MAC 地址，Comware 上可以禁止通告只包含 MAC 地址的 RT-2 路由，以减少通告的 EVPN 路由数量。

在集中式网关组网中，L2 VTEP 需要将学习到的 ARP 通告给 GW，GW 添加该 ARP 表项，并生成 32 位主机路由，路由的下一跳为路由的目的地址本身。

在分布式网关组网中，每一个分布式网关都会将学习到的 ARP 通告给其他网关。在远端 GW 上，RT-2 中的 IP 地址会下发到 VPN 实例的路由表形成 32 位主机路由，此路由的下一跳为通告此路由的 GW 设备。

2.2.5 外部路由通告与学习

EVPN VXLAN 网络构建的是一个私有网络，它也可以通过接入外网，实现与外网的通信。通常在 EVPN VXLAN 的 Spine-Leaf 架构中，会部署一台或多台专门接入外网的设备，称之为 Board leaf。Board leaf 通过普通接口与外网之间运行普通路由协议，学习路由；之后，在 Board leaf 上 EVPN VXLAN 可以引入这些外部路由，形成 EVPN RT-5（5 类）路由，进而通告到 EVPN VXLAN 网络中，使其他 VTEP 也能学到这些外部路由。这些路由的下一跳均指向通告此路由的 Board leaf。当网络中存在多台 Board leaf 时，这些 Board leaf 都可以通告此路由，从而形成等价路由，以达到负载分担的目的。

5 类路由的格式如图 7 所示。

图7 RT-5 路由格式

Route type (1 octet): IP prefix route
Length (1 octet)
RD (8 octets)
Ethernet segment identifier (10 octets)
Ethernet tag ID (4 octets)
IP prefix length (1 octet)
IP prefix (4 or 16 octets)
GW IP address (4 or 16 octets)
L3VNI (3 octets)

如图 8 所示，外部路由通告与学习过程为：

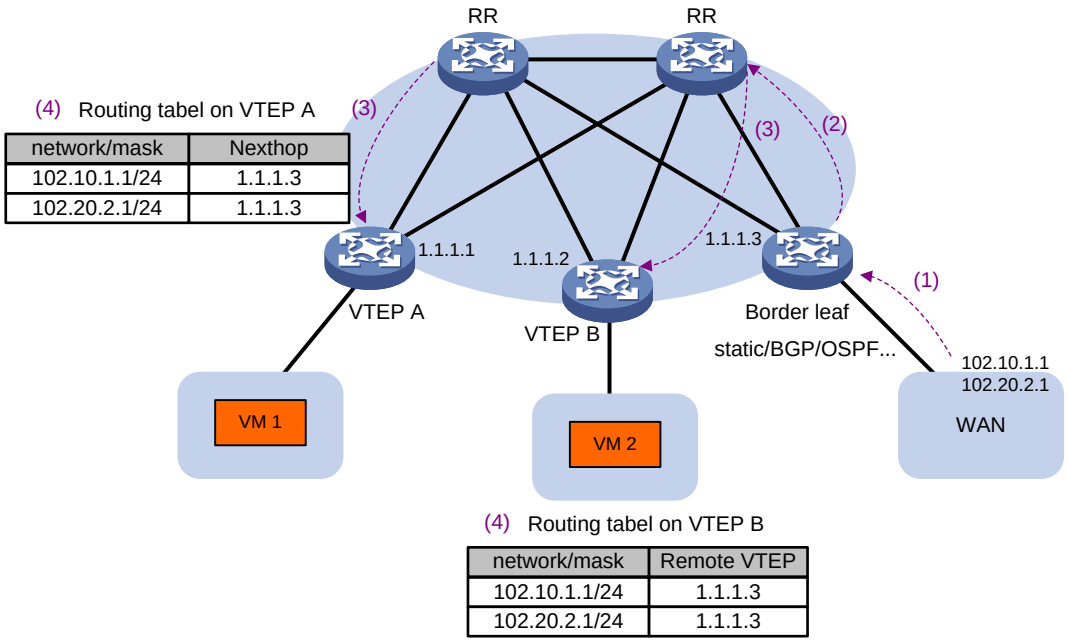
(1) Board leaf 与 WAN 网络之间配置静态路由，或运行 BGP、OSPF 等动态路由协议。Board leaf 学习到外网的路由。

(2) 在 Board leaf 上，将外部路由引入到 EVPN，形成 EVPN 的 5 类路由，并发布给 RR。

(3) RR 将 Board leaf 通告的 5 类路由反射给其他 VTEP。

(4) 远端 VTEP 收到 5 类路由后，如果该路由携带的 Export target 属性与本地某个 VPN 实例的 Import target 属性匹配，将此路由添加到该 VPN 实例的路由表中。

图8 外部路由通告与学习过程



2.2.6 MAC 地址迁移

MAC 地址迁移是指主机/虚拟机从其接入的 VTEP 迁到数据中心网络的另一台 VTEP 下。EVPN VXLAN 通过在 BGP update 消息中携带 MAC Mobility 扩展团体属性，来确保主机/虚拟机迁移后，VTEP 能够及时更新 MAC/IP 路由。

(1) VTEP 第一次发布某个 MAC/IP 路由时，BGP update 消息中不携带 MAC Mobility 扩展团体属性。

(2) 主机/虚拟机迁移后，新迁移到的 VTEP 感知到主机/虚拟机上线，重新通告该 MAC/IP 路由，并在路由中携带 MAC Mobility 扩展团体属性。此扩展团体包含一个序列号。每次迁移，迁移序列号将递增。

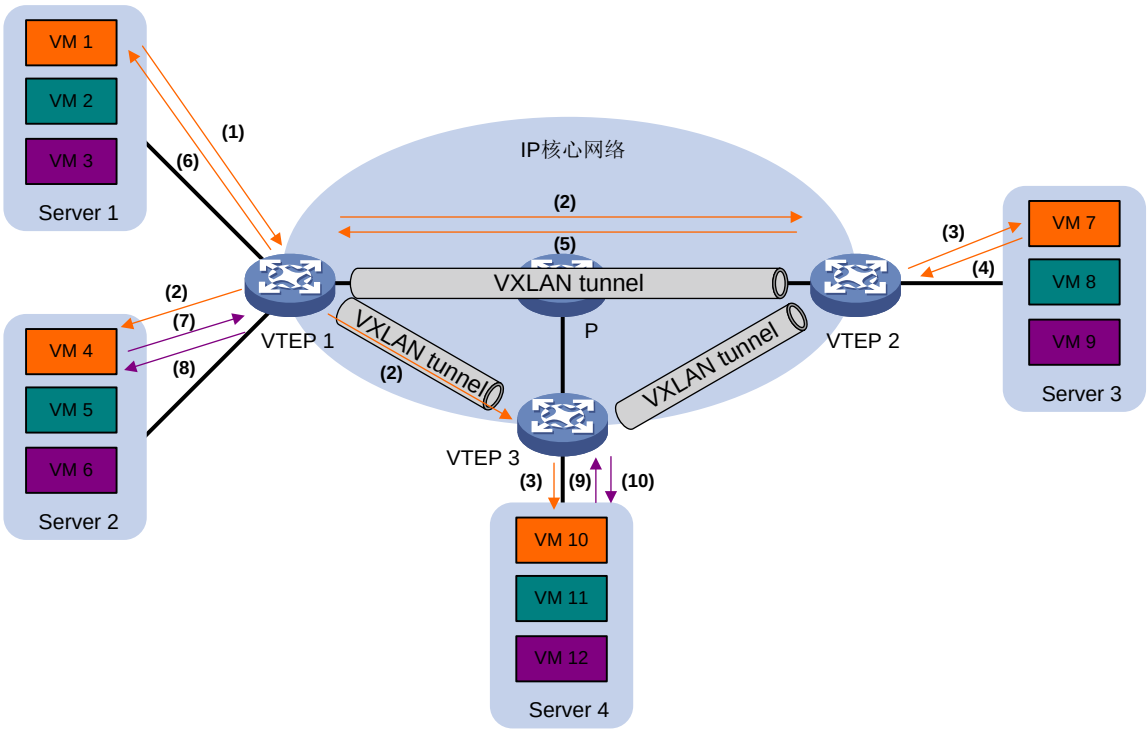
(3) 远端 VTEP 接收到比自己本地保存的序列号更大的 MAC/IP 路由时，更新自己的 MAC/IP 路由消息，下一跳指向迁移后通告此路由的 VTEP。

(4) 原 VTEP 在收到此路由更新后，撤销之前通告的路由。

2.2.7 ARP 泛洪抑制

为了避免广播发送的 ARP 请求报文占用核心网络带宽，VTEP 根据接收到的 ARP 请求和 ARP 应答报文、BGP EVPN 的 RT-2 路由在本地建立 ARP 缓存表项。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 请求时，优先根据本地存储的 ARP 表项进行代理回应。如果没有对应的表项，则将 ARP 请求泛洪到核心网。ARP 泛洪抑制功能可以大大减少 ARP 泛洪的次数。

图9 ARP 泛洪抑制示意图



如图 9 所示，ARP 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，在 VXLAN 内泛洪该 ARP 请求（图 9 以单播路由泛洪方式为例）。VTEP 1 还会通过 BGP EVPN 将该表项同步给 VTEP 2 和 VTEP 3。

(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，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，通过 VXLAN 隧道将 ARP 应答发送给 VTEP 1。VTEP 2 通过 BGP EVPN 将该表项同步给 VTEP 1 和 VTEP 3。

(6) VTEP 1 解封装 VXLAN 报文，获取原始的 ARP 应答，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 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) 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 ARP 泛洪抑制表项，根据已有的表项（VTEP 1 通过 BGP EVPN 同步）回复 ARP 应答报文，不会对 ARP 请求进行泛洪。

2.3 EVPN VXLAN数据平面工作机制

2.3.1 二层流量转发

1. 转发已知单播流量

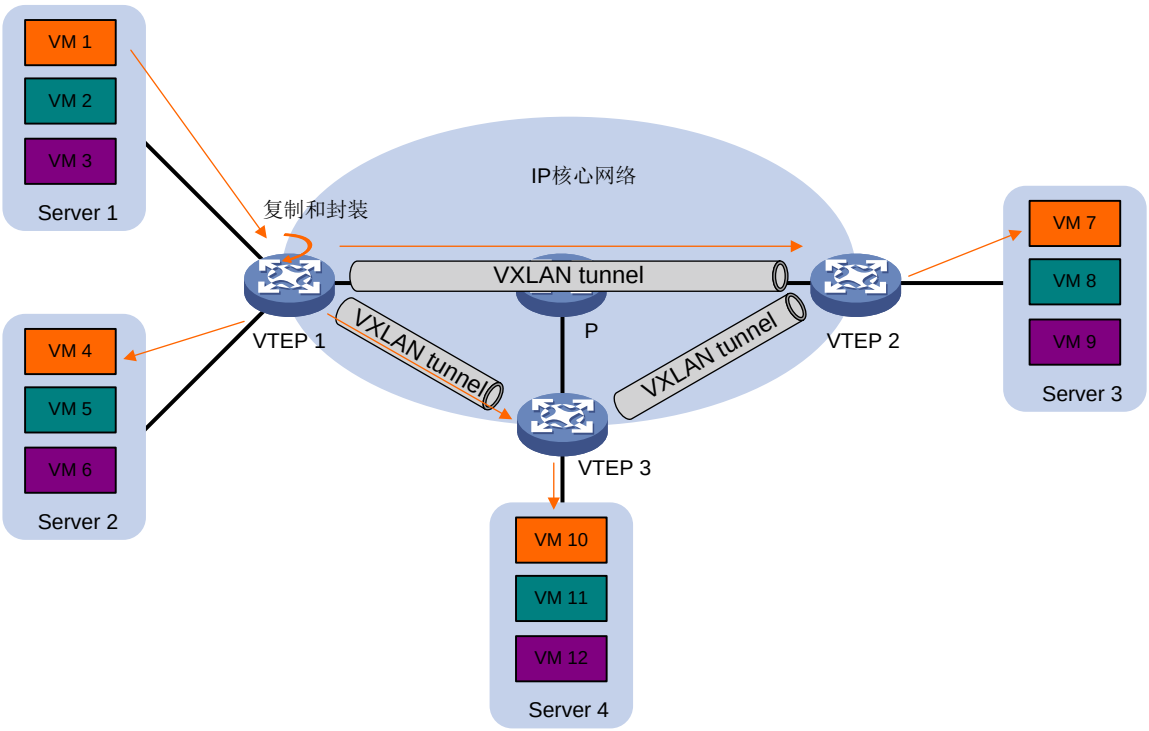
EVPN VXLAN 通过控制平面完成 MAC 地址表项的学习。VTEP 接收到二层数据帧后，判断其所属的 VSI，根据目的 MAC 地址查找该 VSI 的 MAC 地址表，通过表项的出接口转发该数据帧。如果出接口为本地接口，则 VTEP 直接通过该接口转发数据帧；如果出接口为 Tunnel 接口，则 VTEP 根据 Tunnel 接口为数据帧添加 VXLAN 封装后，通过 VXLAN 隧道将其转发给远端 VTEP。

2. 转发 BUM 流量

除了单播流量转发，EVPN VXLAN 网络中还需要转发广播，未知组播与未知单播流量，即 BUM 流量。EVPN VXLAN 采用头端复制方式转发 BUM 流量。

VTEP 接收到本地虚拟机发送的组播、广播和未知单播数据帧后，判断数据帧所属的 VXLAN，通过该 VXLAN 内除接收接口外的所有本地接口和 VXLAN 隧道转发该数据帧。通过 VXLAN 隧道转发数据帧时，需要为其封装 VXLAN 头、UDP 头和 IP 头，将泛洪流量封装在多个单播报文中，发送到 VXLAN 内的所有远端 VTEP。VXLAN 的头端复制列表（即 BUM 广播表）由 EVPN 自动发现并创建，不需要手工干预。

图10 BUM 流量头端复制转发示意图



2.3.2 集中式网关转发

在 EVPN 集中式网关组网中，L2 VTEP 将本地学到的 ARP 通过 EVPN 路由通告给 GW。GW 上创建 ARP 表项，ARP 表项的 MAC 地址为虚拟机的 MAC 地址。GW 还会根据 ARP 生成 32 位主机路由，路由的下一跳为路由的目的地址本身（即虚拟机的 IP 地址）。

集中式网关转发流量的方式为：

- 对于外网访问 EVPN VXLAN 网络内 VM 的流量，GW 接收到报文后，进行三层查表转发，根据 32 位主机路由获取到下一跳为虚拟机的 IP 地址。GW 查找虚拟机 IP 地址对应的 ARP 表项，将报文内层目的 MAC 地址封装为虚拟机的 MAC 地址，并添加 VXLAN 封装后发送给 L2 VTEP。L2 VTEP 解封装后，根据目的 MAC 地址进行二层转发，将报文发送给 VM。
- 对于 EVPN VXLAN 网络内 VM 访问外网的流量，VM 发送给 VTEP 的报文的目的 MAC 为 GW 的网关 MAC。VTEP 查找 MAC 地址表项，添加 VXLAN 封装后，将报文发送给 GW。GW 解封装后，根据内层报文的目的 IP 地址进行三层转发。此时，GW 充当的是 IP 网关角色。
- 对于 EVPN VXLAN 网络内不同 VM 之间的流量，如果 VM 属于同一个 VXLAN，则在 VTEP 上查找 MAC 地址表进行二层转发即可；如果 VM 属于不同的 VXLAN，则 VM 发送给 VTEP 的报文的目的 MAC 为 GW 的网关 MAC，需要经过 GW 进行三层转发，才能将报文转发到目的 VXLAN。此时，GW 充当的是 VXLAN 网关角色。

2.3.3 分布式网关对称 IRB 转发

1. 对称 IRB 转发方式

在 EVPN VXLAN 网络中，VTEP 既可以进行二层 Bridge 转发，也可以进行三层 Router 转发，因此 EVPN VXLAN 的转发称为 IRB（Integrated Routing and Bridging，集成的路由和桥接）。

在分布式网关里面，IRB 转发分为非对称 IRB 和对称 IRB 两种：

- 非对称 IRB：入口网关和出口网关上的处理方式不同。入口网关需要同时进行二层和三层转发，而出口网关只进行二层转发。
- 对称 IRB：入口网关和出口网关上的处理方式相同。对于二层流量，入口网关和出口网关都只进行二层转发；对于三层流量，入口网关和出口网关都只进行三层转发。

Comware 目前仅支持对称 IRB 转发。对称 IRB 转发引入了以下概念：

- L3VNI（Layer 3 VNI）：是指在分布式网关之间通过 VXLAN 隧道转发流量时，属于同一租户（VPN 实例）的流量通过 L3VNI 来标识。L3VNI 唯一关联一个 VPN 实例，通过 VPN 实例确保不同租户之间的业务隔离。
- Route MAC：网关的 Router MAC 地址，是指每个分布式网关拥有的唯一一个用来标识本机的本地 MAC 地址，此 MAC 用于在网关之间通过 VXLAN 隧道转发三层流量。报文在网关之间转发时，报文的内层 MAC 地址为出口网关的 Router MAC 地址。

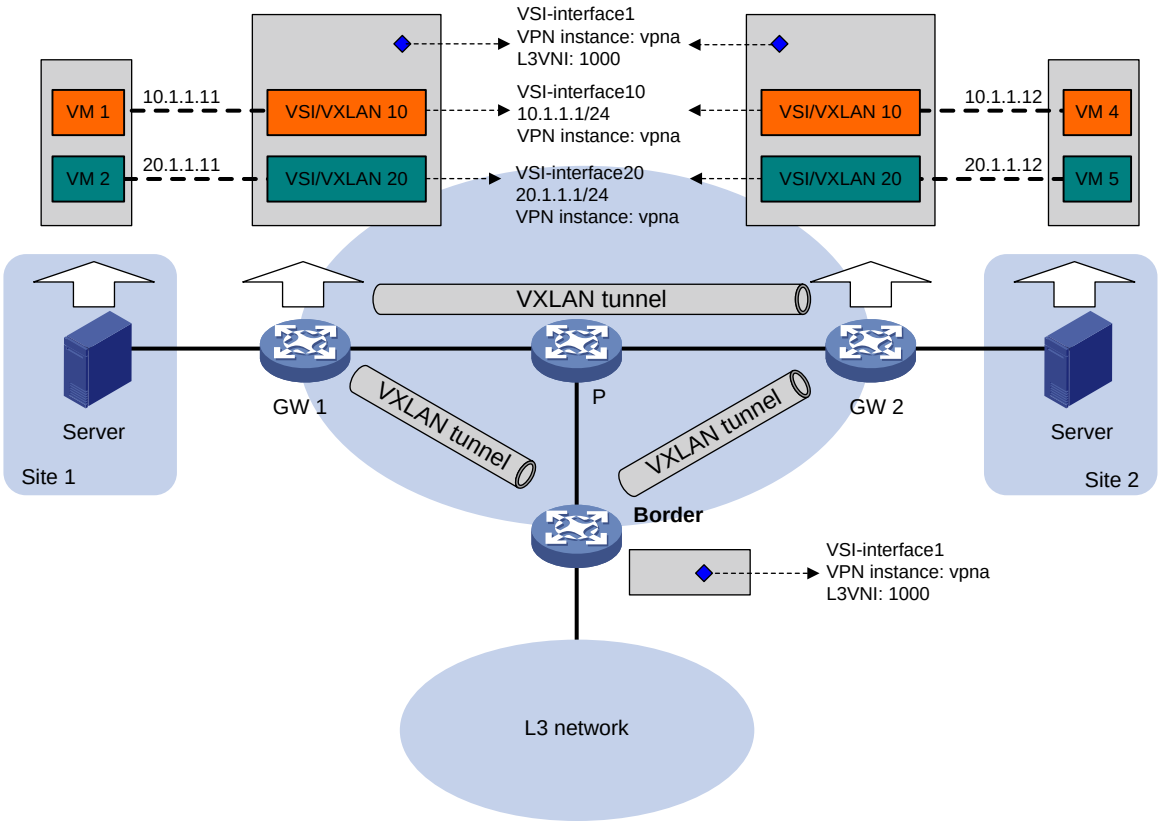
2. 分布式 EVPN 网关部署要求

如图 11 所示，在分布式 EVPN 网关组网中，所有的分布式 EVPN 网关（GW）上都存在以下类型的 VSI 虚接口：

- 作为分布式网关接口的 VSI 虚接口。该接口需要与 VSI、VPN 实例关联。不同 GW 上相同 VSI 虚接口的 IP 地址必须相同，该 IP 地址作为 VXLAN 内虚拟机的网关地址。
- 承载 L3VNI 的 VSI 虚接口。该接口需要与 VPN 实例关联，并需要指定 L3VNI。关联相同 VPN 实例的 VSI 虚接口共用该 L3VNI。

边界网关（Border leaf）上也需要存在承载 L3VNI 的 VSI 虚接口。

图11 分布式 EVPN 网关部署示意图



3. 流量转发过程

分布式网关对流量的转发方式分为两种：

- 区分二三层转发方式：对于二层流量，查找 MAC 地址表进行转发；对于三层流量，查找 FIB 表进行转发。在该方式下，建议在分布式网关上开启 ARP 泛洪抑制功能，以减少泛洪流量。
- 全三层转发方式：对于二层和三层流量，均查找 FIB 表进行转发。在该方式下，需要在分布式网关上开启本地代理 ARP 功能。

查找 MAC 地址表转发二层流量的过程，请参见“2.3.1 1. 转发已知单播流量”；相同站点间三层流量的转发过程如图 12 所示；不同站点间三层流量转发过程如图 13 所示。

图12 相同站点间三层流量转发过程

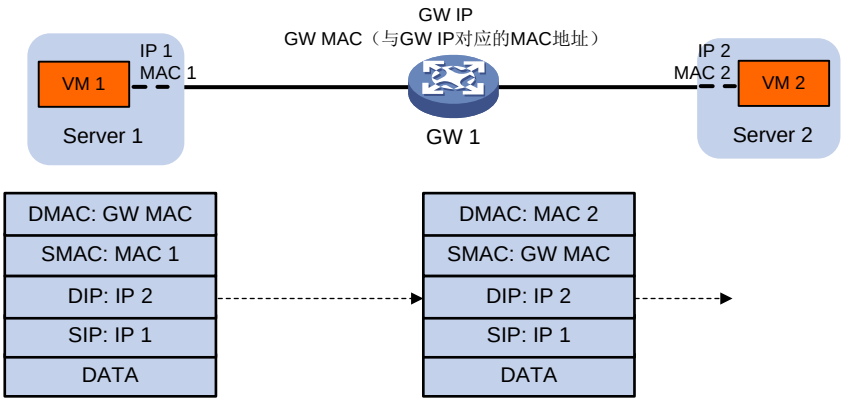
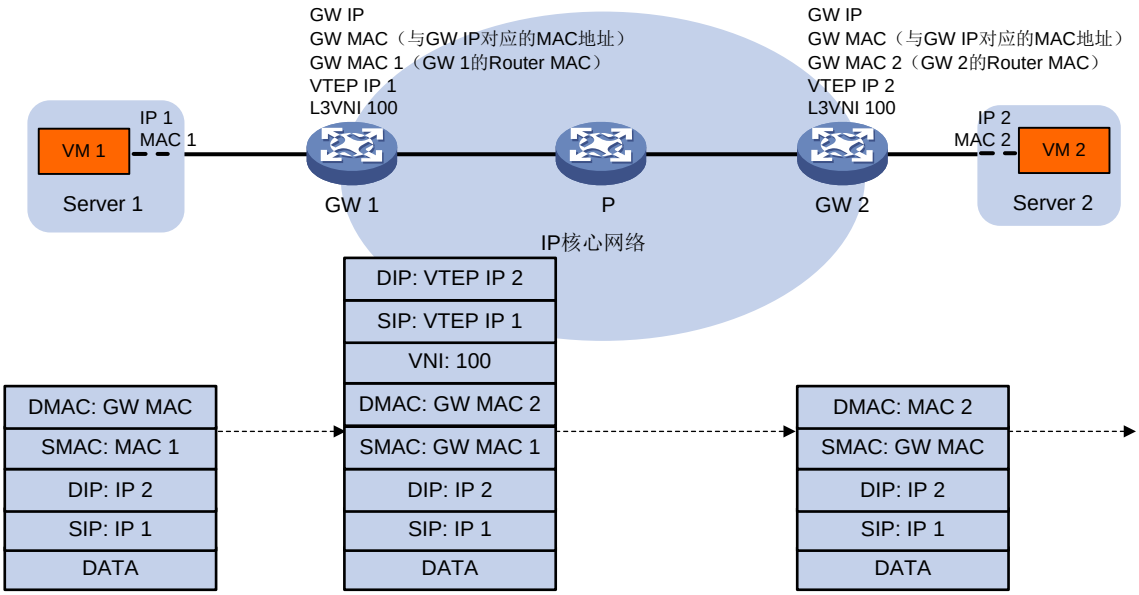


图13 不同站点间三层流量转发过程



以 IPv4 网络为例，查找 FIB 表转发流量的过程为：

(1) 虚拟机访问相同子网、不同子网内的其他虚拟机时，发送 ARP 请求获取 ARP 信息。

(2) GW 接收到 ARP 请求后，判断 ARP 请求所属 VSI，采用与该 VSI 关联的 VSI 虚接口 MAC 地址对其进行应答。

(3) 虚拟机将报文发送给 GW。

(4) GW 判断报文所属 VSI，并查找与该 VSI 关联的 VSI 虚接口，在与 VSI 虚接口关联的 VPN 实例内查找 FIB 表项，并根据匹配的 FIB 表项转发报文：

- 如果 FIB 表项的出接口为本地接口，则 GW 将目的 MAC 替换为目的虚拟机的 MAC 地址、源 MAC 替换为 VSI 虚接口的 MAC，并通过本地接口转发给目的虚拟机。
- 如果 FIB 表项的出接口为 VSI 虚接口，则 GW 将目的 MAC 替换为目的 GW 的 Router MAC 地址、源 MAC 替换为自己的 Router MAC，报文添加 VXLAN 封装后将其转发给目的 GW。其中，为报文封装的 VXLAN ID 为与 VPN 实例关联的 L3VNI。

(5) 目的 GW 接收到报文后，根据 L3VNI 判断报文所属的 VPN 实例，解除 VXLAN 封装后，在该 VPN 实例内查找 ARP 表项转发该报文。

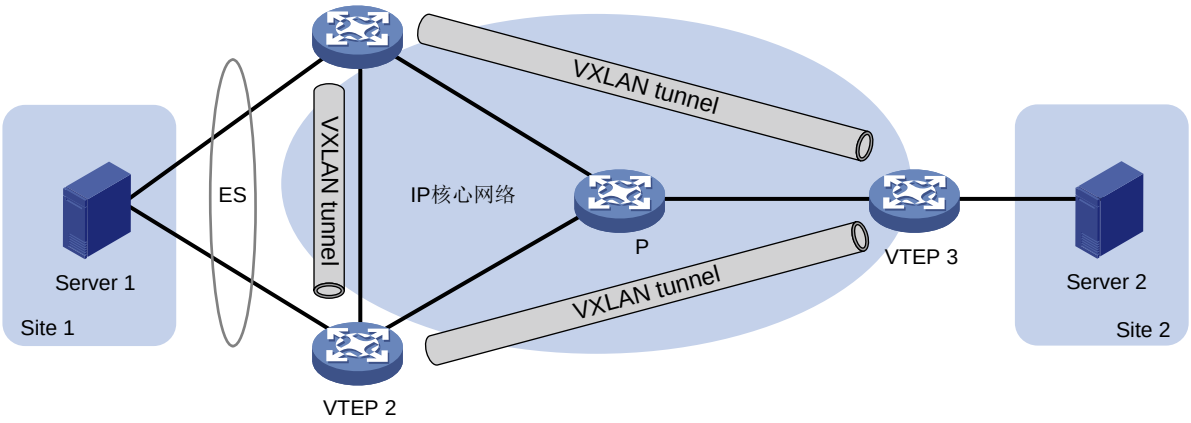
在分布式网关组网中，每一台分布式网关只需要配置下挂的主机/虚拟机所在的 VXLAN ID 即可，且分布式网关不需要维护本租户内所有主机/虚拟机的 ARP 信息，只需要维护少量的远端分布式网关的 ARP 信息即可。

2.4 EVPN VXLAN多归属

2.4.1 功能简介

当一个站点通过不同的以太网链路连接到多台 VTEP 时，这些链路就构成了一个 ES（Ethernet Segment，以太网段），并以一个相同的 ESI（ES Identifier）标识其属于同一个 ES。连接的多台 VTEP 组成冗余备份组，可以避免 VTEP 单点故障对网络造成影响，从而提高 EVPN VXLAN 网络的可靠性。

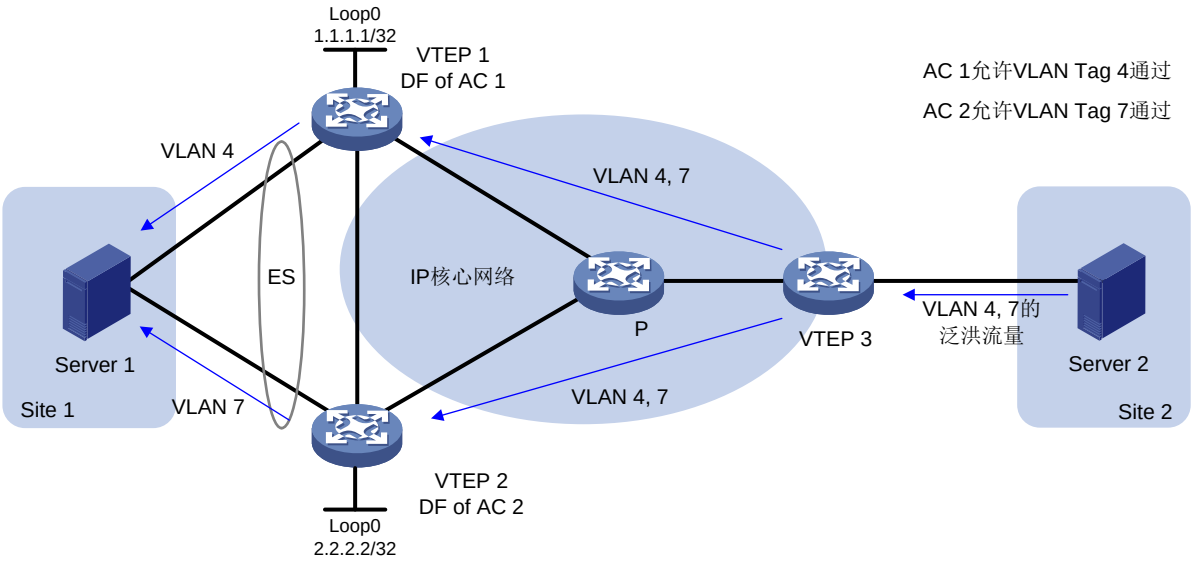
图14 多归属站点示意图



2.4.2 DF 选举

当一个站点连接到多台 VTEP 时，为了避免冗余备份组中的 VTEP 均发送泛洪流量给该站点，需要在冗余备份组中为每个 AC 选举一个 VTEP 作为 DF（Designated Forwarder，指定转发者），负责将泛洪流量转发给该 AC。其他 VTEP 作为该 AC 的 BDF（Backup DF，备份 DF），不会向本地站点转发泛洪流量。多归属成员通过发送以太网段路由，向其它 VTEP 通告 ES 及其连接的 VTEP 信息，仅配置了 ESI 的 VTEP 会接收以太网段路由并根据其携带的 ES、VTEP 信息选举出 DF。

图15 DF 选举方式



如图 15 所示，以 AC 1 的 DF 为例，Comware 支持的 DF 选举方式为：

(1) 选取 AC 内允许通过的最小 VLAN Tag 代表该 AC。在本例中，代表 AC 1 的 VLAN Tag 为 4。

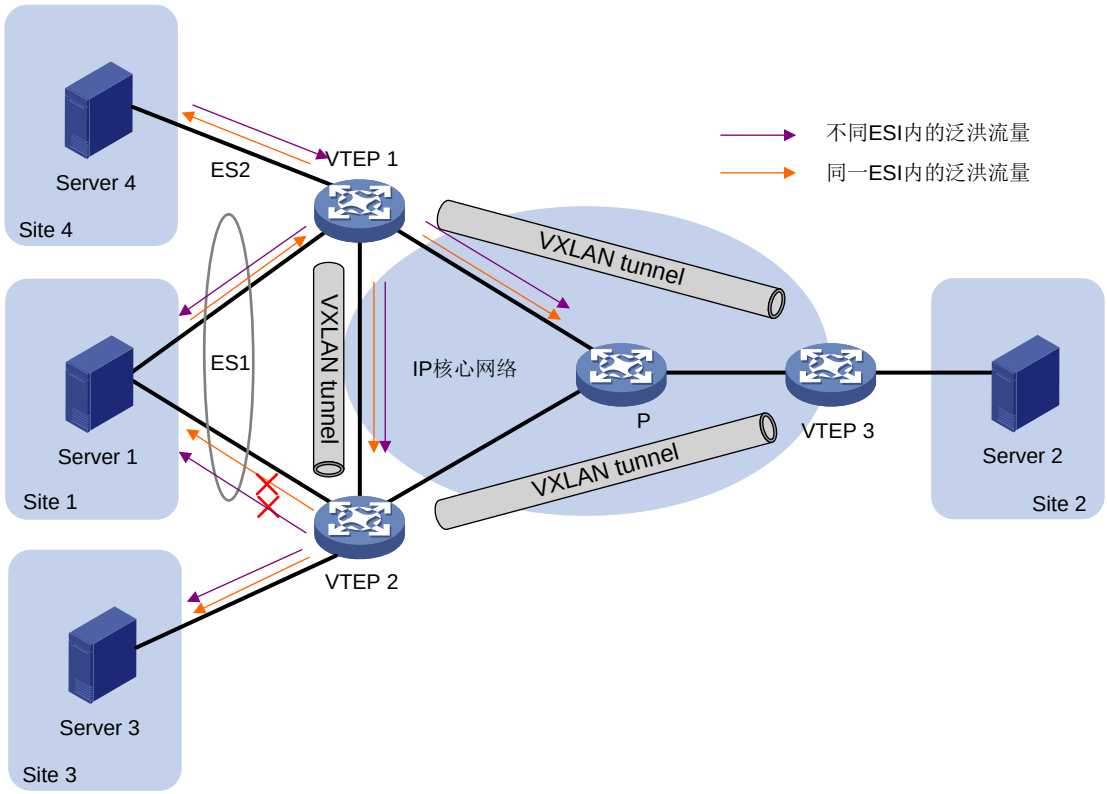
(2) VTEP 根据接收到的以太网段路由，对携带相同 ESI 的路由中的源 IP 地址按升序排列，编号从 0 开始。在本例中，源 IP 1.1.1.1、2.2.2.2 对应的编号依次为 0、1。

(3) 根据 VLAN Tag 除以 N 的余数 M 来选举 DF，N 代表冗余备份组中成员的数量，M 对应的编号为该 AC 的 DF。在本例中，4 除以 2 的余数为 0，即 AC 1 的 DF 为编号为 0 的 VTEP 1。

2.4.3 水平分割

在多归属站点组网中，VTEP 接收到站点发送的广播、未知单播和未知组播数据帧后，判断数据帧所属的 VXLAN，通过该 VXLAN 内除接收接口外的所有本地接口和 VXLAN 隧道转发该数据帧。同一冗余备份组中的 VTEP 接收到该数据帧后会在本地所属的 VXLAN 内泛洪，这样数据帧会通过 AC 泛洪到本地站点，造成环路和站点的重复接收。EVPN VXLAN 通过水平分割解决该问题。水平分割的机制为：VTEP 接收到同一冗余备份组中成员转发的广播、组播、未知单播数据帧后，不向具有相同 ESI 标识的 ES 转发该数据帧。

图16 水平分割



2.4.4 冗余备份模式

当前 Comware 支持的冗余备份模式为多活模式。

在该模式下，冗余备份组中的任何一台成员 VTEP 设备接收到本地站点发送的广播、未知单播和未知组播数据帧后，判断数据帧所属的 VXLAN，通过该 VXLAN 内除接收接口外的所有本地接口和 VXLAN 隧道转发该数据帧。通过 VXLAN 隧道转发数据帧时，需要为其封装 VXLAN 头、UDP 头和 IP 头，将泛洪流量封装在多个单播报文中，发送到 VXLAN 内的所有远端 VTEP。

冗余备份组中的成员 VTEP 接收到远端 VTEP 发送的组播、广播和未知单播数据帧时，只有 DF 可以将数据帧泛洪到本地站点，BDF 接收到后会将泛数据帧丢弃。

2.4.5 Aliasing

Aliasing 功能是指：在多活模式下，冗余备份组中的 VTEP 都会通过 MP-BGP 向远端 VTEP 通告同一 ES 的信息。如果只有一个 VTEP 学习到了 MAC/ARP 信息并发布给远端 VTEP，则远端 VTEP 不仅添加下一跳为该 VTEP 的 IP 地址，还会添加下一跳为冗余备份组中其他 VTEP 的 IP 地址，以实现自动在不同 VTEP 之间形成等价路径。

2.5 EVPN VXLAN支持组播

2.5.1 功能简介

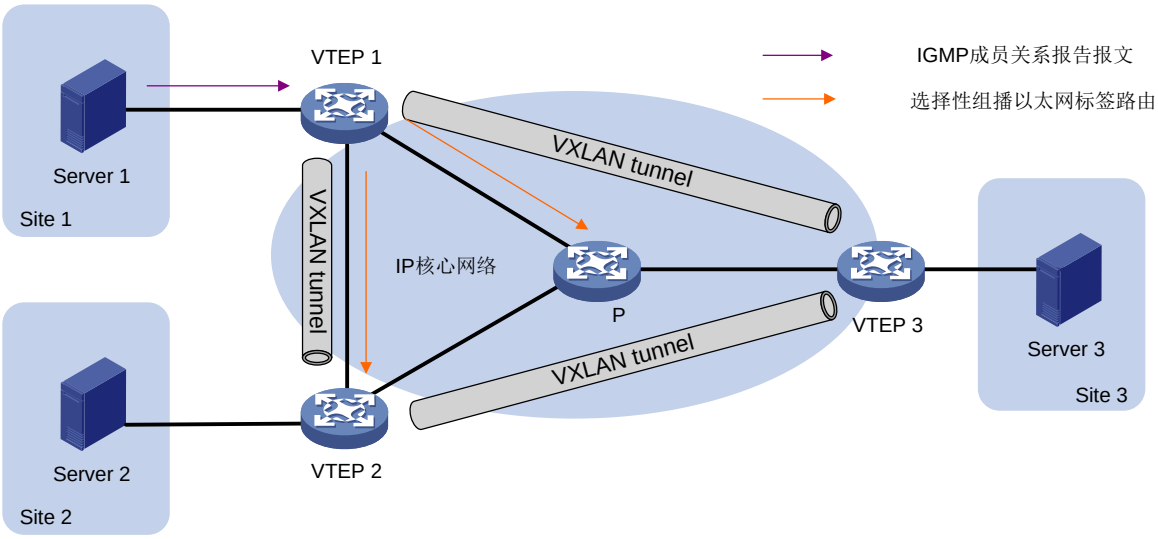
为了避免组播发送的 IGMP 报文占用核心网络带宽，VTEP 会根据接收到的报告报文和离开报文在本地建立或删除组播转发表项。通过 SMET（Selective Multicast Ethernet Tag Route，选择性组播以太网标签路由）路由将组播组信息通告给其他 VTEP，远端 VTEP 收到 SMET 路由后在本地建立组播转发表项。当 VTEP 再次收到属于同一 IGMP 版本加入同一组播组的报告报文时，将不再发送 SMET 路由。EVPN VXLAN 支持组播功能可以大大减少 IGMP 报文泛洪的次数。

为了支持组播，MP-BGP 在 EVPN 地址族新增了 SMET、IGMP-JS 和 IGMP-LS 三类 EVPN 路由，详细介绍请参见“2.2.2 BGP EVPN 路由”。

2.5.2 单归属站点组播

如图 17 所示，在单归属站点组网中，Server 1 发出 IGMP 成员关系报告报文至 VTEP 1。VTEP 1 上生成相应的组播表项，并发送 SMET 路由将组播信息通告给 VTEP 2 和 VTEP 3。VTEP 2 和 VTEP 3 收到 SMET 路由后形成下一跳为 VTEP 1 的组播表项。

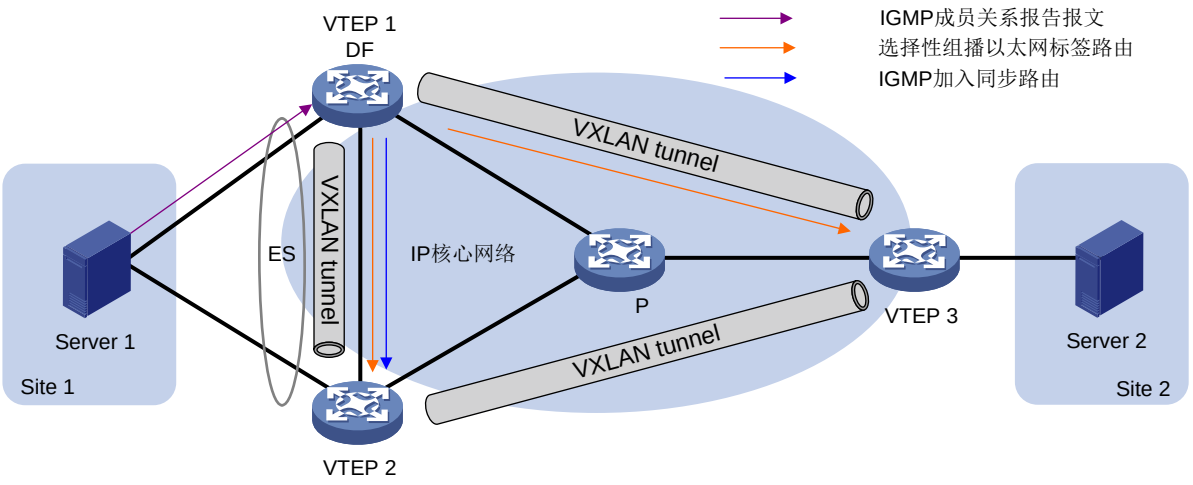
图17 单归属站点组播示意图



2.5.3 多归属站点组播

站点多归属组网中，站点侧发送的加入组播组报文和离开组播组报文，会被不同的 VTEP 接收。为了在多归属站点间管理站点的组播表项，收到加入和离开组播组报文的 VTEP 会发送 IGMP-JS 路由和 IGMP-LS 路由来告诉其他成员，保证同 ESI 成员 VTEP 间组播信息的同步。

图18 多归属站点组播示意图



如图 18 所示，多归属站点组播处理过程如下：

- 当接收报告报文的设备为 DF（VTEP 1）时，DF 通告 SMET 路由给 VTEP 2 和 VTEP 3，并通告 IGMP-JS 路由给 VTEP 2。当组播接收者离开组播组时：
  - 若接收离开报文的设备为 DF，则 DF 通告 IGMP-LS 路由并撤销 IGMP-JS 路由和 SMET 路由。
  - 若接收离开报文的设备为 BDF（VTEP 2），则 BDF 通告 IGMP-LS 路由给同一冗余备份中的其他成员。DF 收到 BDF 同步的 IGMP-LS 路由后，撤销 IGMP-JS 路由和 SMET 路由。
- 当接收报告报文的设备为 BDF 时，BDF 通告 IGMP-JS 路由给同一冗余备份中的其他成员，DF 收到 IGMP-JS 路由后生成 SMET 路同步给 VTEP 2 和 VTEP 3。当组播接收者离开组播组时：
  - 若接收离开报文的设备为 DF，则 DF 通告 IGMP-LS 路由给同一冗余备份中的其他成员。BDF 收到 IGMP-LS 路由后撤销 IGMP-JS 路由。DF 收到撤销 IGMP-JS 路由后，撤销由 IGMP-JS 路由生成的 SMET 路由。
  - 若接收离开报文的设备为 BDF，则 BDF 通告 IGMP-LS 路由并撤销 IGMP-JS 路由。DF 收到撤销 IGMP-JS 路由后，会撤销由该 IGMP-JS 路由生成的 SMET 路由。

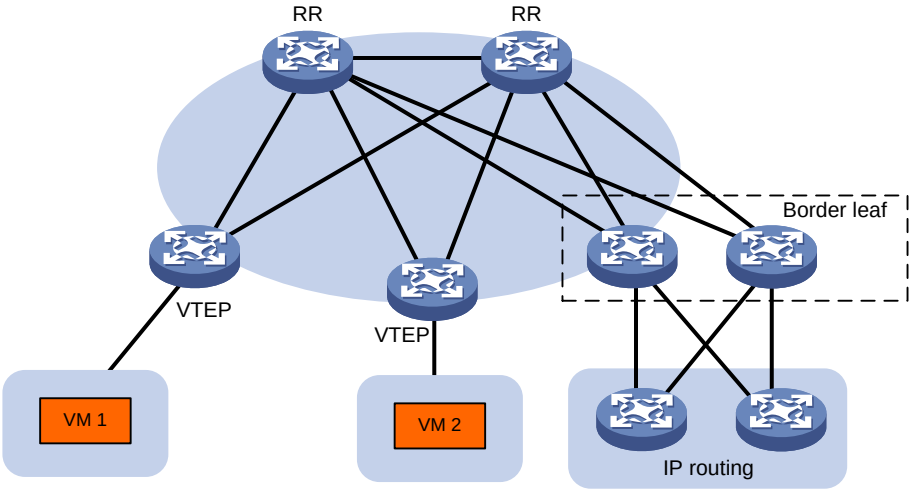
3 典型组网应用

3.1 EVPN分布式网关组网

EVPN 分布式网关组网中，对网关设备转发能力的要求没有集中式网关那么高，且在核心设备只需要支持普通的 IP 转发即可，因此，EVPN 分布式网关应用非常广泛。

EVPN 分布式网关的典型组网如图 19 所示。VTEP 为 EVPN 分布式网关设备；Border leaf 为与广域网连接的边界网关设备，部署两台 Border leaf，形成备份；RR 负责在交换机之间反射 BGP 路由。

图19 EVPN 分布式网关组网示意图

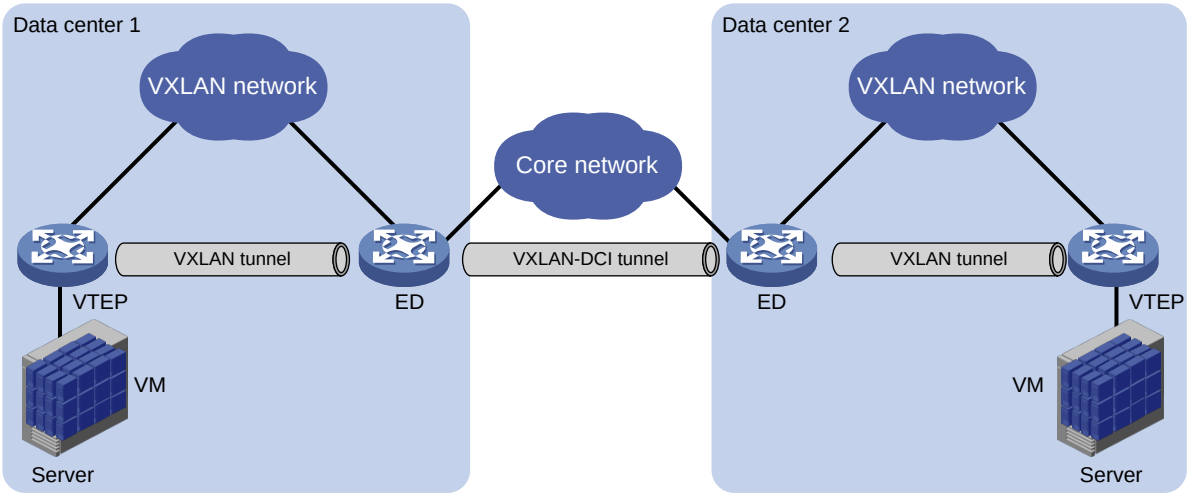


3.2 EVPN数据中心互联组网

EVPN 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 20 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20 VXLAN 数据中心互联典型组网图

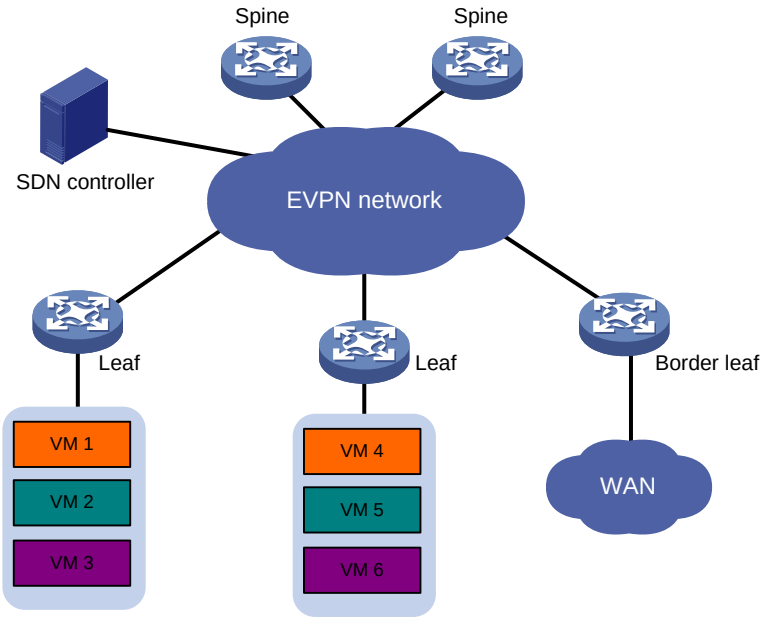


3.3 EVPN与SDN控制器配合组网

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网的设备。如图 21 所示，EVPN 可以与 SDN 控制器

配合使用，EVPN 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21 EVPN 与 SDN 控制器配合组网



4 参考文献

- RFC 7432: BGP MPLS-Based Ethernet VPN
- draft-ietf-bess-evpn-overlay-02: A Network Virtualization Overlay Solution using EVPN
- draft-ietf-bess-evpn-prefix-advertisement-02: BGP MPLS-Based Ethernet VPN
- draft-ietf-bess-evpn-igmp-mld-proxy-01: IGMP and MLD Proxy for EVPN