EVPN 技术白皮书

目 录

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。

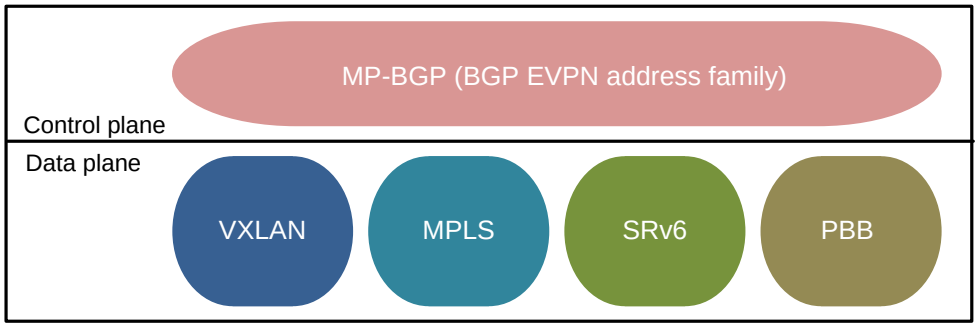
EVPN（Ethernet Virtual Private Network，以太网虚拟专用网络）是一种基于 Overlay 技术的二层网络互联技术，具有部署简单、扩展性强等优点。EVPN 采用 MP-BGP 协议通告 MAC/IP 的可达性和组播等信息，通过生成的 MAC 表项和路由表项进行二/三层报文转发，以实现二层网络互联，很好地满足了用户对于大型数据中心网络的需求。

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、运营商网络中也具有一定的应用。

1.2 协议框架

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、MPLS 和 IPv6 Segment Routing（SRv6）作为数据平面。

图1 EVPN 协议框架



不同数据平面对应的 EVPN 技术分别为：

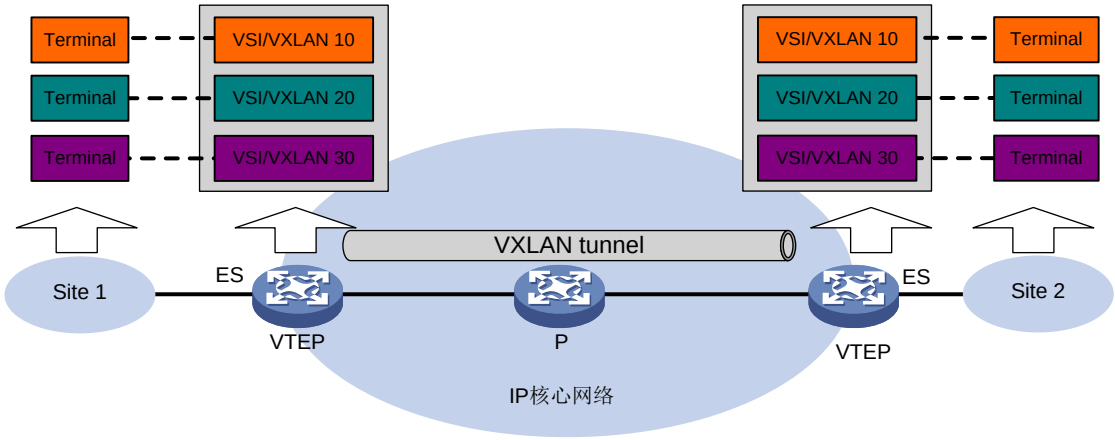
- EVPN VXLAN：数据平面采用 VXLAN 封装。

  EVPN VXLAN 网络的边缘设备称为 VTEP（VXLAN Tunnel End Point，VXLAN 隧道端点），EVPN 的相关处理均在 VTEP 上完成。EVPN VXLAN 通过在 VTEP 间建立 VXLAN 隧道，透明传输二层数据报文，实现不同站点间的二层互联。

  通过在 EVPN VXLAN 网络中部署 EVPN 网关，可以实现为同一租户的不同子网提供三层互联，并为其提供与外部网络的三层互联。

  EVPN VXLAN 的详细介绍请参见“2 EVPN VXLAN”。

图2 EVPN VXLAN 网络模型示意图

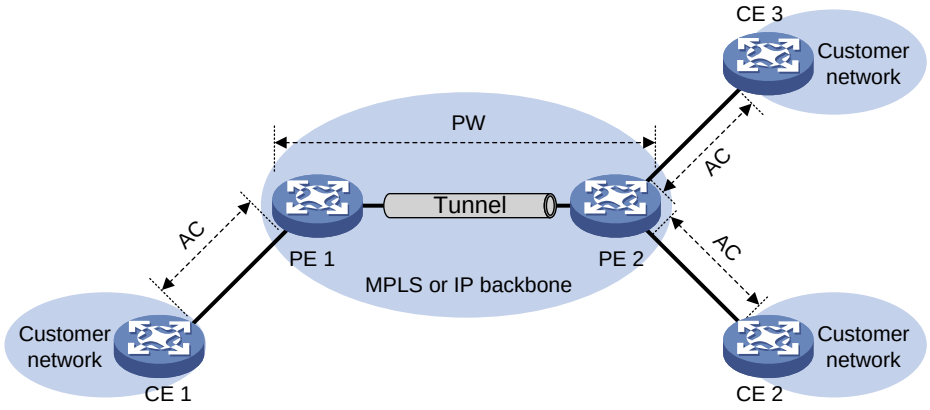


- EVPN VPLS：数据平面采用 MPLS 封装，用来实现用户的点到多点二层互通。

  EVPN VPLS 组网中，用户网络侧设备 CE 通过 AC 接入服务提供商网络侧设备 PE，PE 间通过 BGP EVPN 路由建立 PW，PE 通过查找 MAC 地址表转发报文，可以实现用户点对多点的二层互通。

  EVPN VPLS 的详细介绍请参见“3 EVPN VPLS”。

  图3 EVPN VPLS 网络模型示意图

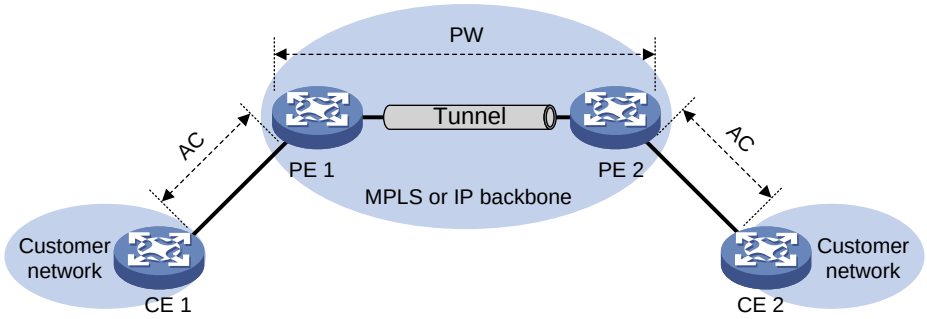


- EVPN VPWS：数据平面采用 MPLS 封装，用来实现用户的点到点二层互通。

  EVPN VPWS 组网中，用户网络侧设备 CE 通过 AC 接入服务提供商网络侧设备 PE，PE 间通过 BGP EVPN 路由建立 EVPN PW，在 PE 上使用交叉连接将 AC 与 EVPN PW 关联，即可实现用户点对点的二层互通。

  EVPN VPWS 的详细介绍请参见“4 EVPN VPWS”。

图4 EVPN VPWS 网络模型示意图



- EVPN VPLS over SRv6：数据平面采用 SRv6 封装，用来实现用户的点到多点二层互通。
- EVPN VPWS over SRv6：数据平面采用 SRv6 封装，用来实现用户的点到点二层互通。

本文仅介绍 EVPN VXLAN、EVPN VPLS 和 EVPN VPWS 三种 EVPN 技术，EVPN VPLS over SRv6 和 EVPN VPWS over SRv6 的详细介绍，请参见《SRv6 技术白皮书》。

1.3 技术优势

EVPN 不仅继承了 MP-BGP 和 VXLAN/MPLS 的优势，还提供了新的功能。EVPN 具有如下特点：

- 简化配置：通过 MP-BGP 实现 VTEP/PE 自动发现、VXLAN 隧道/PW 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。
- 提供点到点和点到多点的服务：将用户的二层数据封装成可以在 IP 或 MPLS 网络中传送的分组，从而实现用户二层数据跨越 IP 或 MPLS 网络在不同站点间透明地传送。

相对于传统 VPLS 技术，EVPN 存在如下优势：

- EVPN 支持完善的多归属接入应用场景，支持负载分担和主备备份两种工作模式。
- 二层网络间的 MAC/IP 地址学习和发布从数据平面转移到控制平面，采用 MP-BGP 协议通告 MAC/IP 的可达性，使设备可以像管理路由一样灵活地管理 MAC/IP 地址：
  - 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。
- 使用 BGP 作为控制协议，统一了二层和三层的控制信令协议。
- 通过部署路由反射器，避免设备全连接，降低网络部署的难度。

1.4 BGP EVPN路由

为了支持 EVPN，MP-BGP 在 L2VPN 地址族下定义了新的子地址族——EVPN 地址族，并为该地址族定义了 EVPN NLRI（Network Layer Reachability Information，网络层可达性信息），即 EVPN 路由。EVPN 子地址族使用的地址族编号为：AFI=25，SAFI=70。

在 EVPN 网络中，VTEP/PE 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/PE 都只和 RR 建立 BGP 邻居关系。RR 发现并接收 VTEP/PE 发起的 BGP 连接后形成客户机列表，将从某个 VTEP/PE 收到的路由反射给其他所有的 VTEP/PE。
- 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。

1.4.1 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，也可以用来在 EVPN VPWS 组网中通告 Service ID 信息。

以太网自动发现路由分为：

- Ethernet Auto-discovery Per ES 路由：主要用于多归属组网的快速收敛、冗余模式和水平分割。
- Ethernet Auto-discovery Per EVI 路由：主要用于多归属组网的别名（Aliasing）、备份路径功能。

图5 以太自动发现路由报文格式

Route type (1 octet): Ethernet Auto-discovery Route
Length (1 octet)
RD (8 octets)
Ethernet segment identifier (10 octets)
Ethernet tag ID (4 octets)
MPLS label (3 octets)

如图 5 所示，以太自动发现路由包括如下字段：

- RD（Route Distinguisher，路由标识符）：EVPN 实例的 RD 值。
- Ethernet segment identifier：VTEP/PE 与 CE 之间的以太网链路的段标识符。同一站点 CE 通过不同链路多归属到不同 PE 时，这些链路构成一个 ES，并以一个相同的 ESI 标识。
- Ethernet tag ID：
  - 对于 Ethernet Auto-discovery Per ES 路由，该字段为全 F。
  - 对于 Ethernet Auto-discovery Per EVI 路由，不同类型组网中，该字段对应不同的取值：
    - 在 EVPN VPLS 和 EVPN VXLAN 组网中，该字段为 VSI 实例的 Tag ID、接入 AC 对应的 VLAN 或全 0。
    - 在 EVPN VPWS 组网中中，本字段为本端的 Service ID。
- MPLS label：
  - 对于 Ethernet Auto-discovery Per ES 路由，该字段为 0。
  - 对于 Ethernet Auto-discovery Per EVI 路由，不同数据封装类型下，该字段对应不同的取值：
    - VXLAN 封装时，为 VXLAN ID。
    - MPLS 封装时，为 MPLS Label。

- SRv6 封装时，该字段与 SRv6 TLV 组合在一起表示 SID。

1.4.2 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）。

图6 MAC/IP 发布路由报文格式

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)
MPLS label1 (3 octets)
MPLS label2(0 or 3 octets)

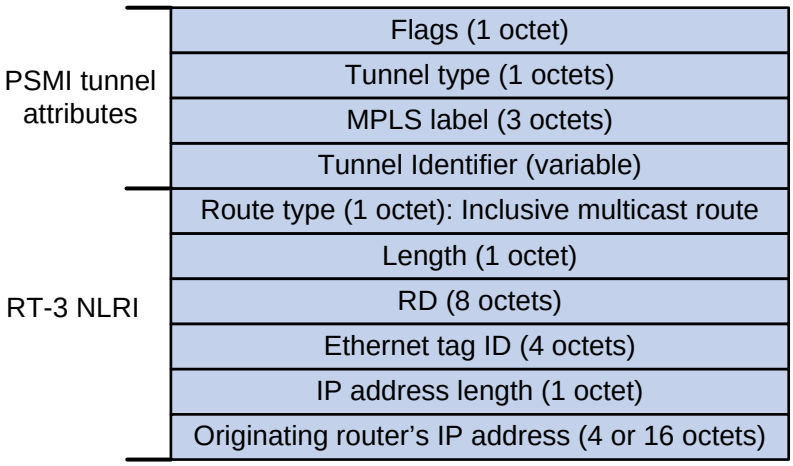
如图 6 所示，MAC/IP 发布路由包括如下字段：

- RD：EVPN 实例的 RD 值。
- Ethernet segment identifier：VTEP/PE 与 CE 之间的以太网链路的段标识符。
- Ethernet tag ID：该字段为 VSI 实例的 Tag ID、接入 AC 对应的 VLAN 或全 0。
- MAC address length：通告的 MAC 地址长度。
- MAC address：通告的 MAC 地址。
- IP address length：主机 IP 地址的掩码长度。
- IP address ：通告的主机 IP 地址。
- MPLS label1：不同的数据封装类型下，该字段对应不同的取值：
  - VXLAN 封装时，为 VXLAN ID。
  - MPLS 封装时，为 MPLS Label。
  - SRv6 封装时，该字段与 SRv6 TLV 组合在一起表示 SID。
- MPLS label2：三层业务流量转发时使用的标识。不同的数据封装类型下，该字段对应不同的取值：
  - VXLAN 封装时，为 L3VNI。
  - MPLS 封装暂不支持该字段。
  - SRv6 封装时，该字段与 SRv6 TLV 组合在一起表示 SRv6 分布式网关用于转发三层流量的 SRv6 SID。

1.4.3 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。在 EVPN VXLAN 组网中用来通告 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；在 EVPN VPLS 组网中用来通告 PE 信息，实现 PE 的自动发现、自动建立 PW。

图7 包含性组播以太网标签路由的报文格式



如图 7 所示，发布包含性组播以太网标签路由时，需要在该路由中携带 PSMI（Provider Multicast Service Interface，运营商组播业务接口）tunnel attributes，该属性中各字段的含义为：

- Flags：标记位。
- Tunnel type：隧道类型，取值如下：
  - 0：表示 No tunnel information present。
  - 1：表示 RSVP-TE P2MP LSP。
  - 2：表示 mLDP P2MP LSP。
  - 3：表示 PIM-SSM Tree。
  - 4：表示 PIM-SM Tree。
  - 5：表示 BIDIR-PIM Tree。
  - 6：表示 Ingress Replication。
  - 7：表示 mLDP MP2MP LSP。
- MPLS label：转发 BUM（Broadcast/Unknown unicast/Unknown Multicast，广播/未知单播/未知组播）流量时，封装的 MPLS 标签、VXLAN ID 或 SID。
- Tunnel Identifier：当隧道类型为 Ingress Replication 时，表示隧道对端的 IP 地址。

包含性组播以太网标签路由包含如下字段：

- RD：EVPN 实例的 RD 值。
- Ethernet tag ID：该字段为接入 AC 对应的 VLAN 或全 0。
- IP address length：始发该路由的 IP 地址的掩码长度。
- Originating router's IP address：始发该路由的 VTEP 或 PE 的 IP 地址，取值为 BGP 协议的 Router ID。

1.4.4 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 的多归属冗余备份组中的其他成员，以及在冗余组之间选举 DF 等。

图8 以太网段路由报文格式

Route type (1 octet): Ethernet segment route
Length (1 octet)
RD (8 octets)
Ethernet segment identifier (10 octets)
IP address length (1 octet)
Originating router's IP address (4 or 16 octets)

如图 8 所示，以太网段路由包含如下字段：

- RD：根据 VTEP/PE 的 IP 地址自动生成的 RD，例如 X.X.X.X:0。
- Ethernet segment identifier：VTEP/PE 与 CE 之间的以太网链路的段标识符。
- IP address length：始发该路由的 IP 地址的掩码长度。
- Originating router's IP address：始发该路由的 VTEP 或 PE 的 IP 地址，取值为 BGP 协议的 Router ID。

1.4.5 IP Prefix Advertisement Route（RT-5）

IP 前缀路由，用来以 IP 前缀的形式通告 BGP IPv4 单播路由或 BGP IPv6 单播路由。

图9 IP 前缀路由报文格式

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)
MPLS label (3 octets)

如图 9 所示，IP 前缀路由包含如下字段：

- RD：VPN 实例/公网实例下 EVPN 地址族的 RD 值。
- Ethernet segment identifier：VTEP/PE 与 CE 之间的以太网链路的段标识符。
- Ethernet tag ID：固定为全 0。
- IP prefix length：IP 前缀掩码长度。
- IP prefix：IP 前缀地址。

- GW IP address：默认网关地址。
- L3VNI：不同数据封装类型下，该字段对应不同的取值：
  - VXLAN 封装时，为转发三层业务流量时使用的 L3VNI。
  - MPLS 封装时，为 MPLS Label。
  - SRv6 封装时，为转发三层业务流量时使用的 SID。

1.4.6 Selective Multicast Ethernet Tag Route（RT-6）

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。

图10 选择性组播以太网标签路由报文格式

Route type (1 octet): Selective Multicast Ethernet Tag route
Length (1 octet)
RD (8 octets)
Ethernet tag ID (4 octets)
Multicast source length (1 octet)
Multicast source address (variable)
Multicast group length (1 octet)
Multicast group address (Variable)
Originator router length (1 octet)
Originator router address (variable)
Flags (1 octets) (optional)

如图 10 所示，选择性组播以太网标签路由包含如下字段：

- RD：EVPN 实例的 RD 值。
- Ethernet tag ID：该字段为全 0。
- Multicast source length：租户加入的组播源的 IP 地址长度，32 位代表 IPv4，128 位代表 IPv6。
- Multicast source address：租户加入的组播源的地址。
- Multicast group length：租户加入的组播组的 IP 地址长度，32 代表 IPv4，128 位代表 IPv6。
- Multicast group address：租户加入的组播组地址。
- Originator router length：始发该路由的 IP 地址的长度，32 代表 IPv4，128 位代表 IPv6。
- Originator router address：始发该路由的 VTEP 或 PE 的 IP 地址，取值为 BGP 协议的 Router ID。
- Flags：标记位。该字段表示的内容与 Multicast group address 字段有关：
  - 如果 Multicast group address 为 IPv4 地址：
    - bit 7 表示是否支持 IGMP version 1。
    - bit 6 表示是否支持 IGMP version 2。
    - bit 5 表示是否支持 IGMP version 3。

- bit 4 表示携带的(S，G)的模式，取值为 1，表示 Exclude 模式;取值为 0，表示 Include 模式。该 bit 位仅在 bit 5 取值为 1 时有效，bit 5 取值为 0 时忽略该 bit 位。

  - 如果 Multicast group address 为 IPv6 地址：
    - bit 7 表示是否支持 MLD version 1。
    - bit 6 表示是否支持 MLD version 2。
    - bit 5 目前固定值为 0。
    - bit 4 表示携带的(S，G)的模式，取值为 1，表示 Exclude 模式;取值为 0，表示 Include 模式。该 bit 位仅在 bit 6 取值为 1 时有效，bit 6 取值为 0 时忽略该 bit 位。

1.4.7 IGMP Join Synch Route（RT-7）

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。

图11 IGMP 加入同步路由报文格式

Route type (1 octet): IGMP join synch route
Length (1 octet)
RD (8 octets)
Ethernet segment identifier (10 octets)
Ethernet tag ID (4 octets)
Multicast source length (1 octet)
Multicast source address (variable)
Multicast group length (1 octet)
Multicast group address (Variable)
Originator router length (1 octet)
Originator router address (variable)
Flags (1 octets) (optional)

如图 11 所示，IGMP 加入同步路由包含如下字段：

- RD：EVPN 实例的 RD 值。
- Ethernet segment identifier：VTEP/PE 与 CE 之间的以太网链路的段标识符。
- Ethernet tag ID：接入 AC 对应的 VLAN。
- Multicast source length：租户加入的组播源的 IP 地址长度，32 位代表 IPv4，128 位代表 IPv6。
- Multicast source address：租户加入的组播源的地址。
- Multicast group length：租户加入的组播组的 IP 地址长度，32 代表 IPv4，128 位代表 IPv6。
- Multicast group address：租户加入的组播组地址。
- Originator router length：始发该路由的 IP 地址的长度，32 代表 IPv4，128 位代表 IPv6。
- Originator router address：始发该路由的 VTEP 或 PE 的 IP 地址，取值为 BGP 协议的 Router ID。
- Flags：标记位。该字段表示的内容与 Multicast group address 字段有关：

- 如果 Multicast group address 为 IPv4 地址：
  - bit 7 表示是否支持 IGMP version 1。
  - bit 6 表示是否支持 IGMP version 2。
  - bit 5 表示是否支持 IGMP version 3。
  - bit 4 表示携带的(S，G)的模式，取值为 1，表示 Exclude 模式；取值为 0，表示 Include 模式。该 bit 位仅在 bit 5 取值为 1 时有效，bit 5 取值为 0 时忽略该 bit 位。
- 如果 Multicast group address 为 IPv6 地址：
  - bit 7 表示是否支持 MLD version 1。
  - bit 6 表示是否支持 MLD version 2。
  - bit 5 目前固定值为 0。
  - bit 4 表示携带的(S，G)的模式，取值为 1，表示 Exclude 模式；取值为 0，表示 Include 模式。该 bit 位仅在 bit 6 取值为 1 时有效，bit 6 取值为 0 时忽略该 bit 位。

1.4.8 IGMP Leave Synch Route（RT-8）

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。

图12 IGMP 离开同步路由报文格式

Route type (1 octet): IGMP leave synch route
Length (1 octet)
RD (8 octets)
Ethernet segment identifier (10 octets)
Ethernet tag ID (4 octets)
Multicast source length (1 octet)
Multicast source address (variable)
Multicast group length (1 octet)
Multicast group address (Variable)
Originator router length (1 octet)
Originator router address (variable)
Leave group synchronization # (4 octets)
Maximum response time (1 octet)
Flags (1 octets) (optional)

如图 12 所示，IGMP 离开同步路由包含如下字段：

- RD：EVPN 实例的 RD 值。
- Ethernet segment identifier：VTEP/PE 与 CE 之间的以太网链路的段标识符。
- Ethernet tag ID：接入 AC 对应的 VLAN。
- Multicast source length：租户加入的组播源的 IP 地址长度，32 位代表 IPv4，128 位代表 IPv6。

- Multicast source address：租户加入的组播源的地址。
- Multicast group length：租户加入的组播组的 IP 地址长度，32 代表 IPv4，128 位代表 IPv6。
- Multicast group address：租户加入的组播组地址。
- Originator router length：始发该路由的 IP 地址的长度，32 代表 IPv4，128 位代表 IPv6。
- Originator router address：始发该路由的 VTEP 或 PE 的 IP 地址，取值为 BGP 协议的 Router ID。
- Leave group synchronization：租户离开组播组的序列号。
- Maximum response time：通告的最大响应时间。
- Flags：标记位。该字段表示的内容与 Multicast group address 字段有关：
  - 如果 Multicast group address 为 IPv4 地址：
    - bit 7 表示是否支持 IGMP version 1。
    - bit 6 表示是否支持 IGMP version 2。
    - bit 5 表示是否支持 IGMP version 3。
    - bit 4 表示携带的(S，G)的模式，取值为 1，表示 Exclude 模式;取值为 0，表示 Include 模式。该 bit 位仅在 bit 5 取值为 1 时有效，bit 5 取值为 0 时忽略该 bit 位。
  - 如果 Multicast group address 为 IPv6 地址：
    - bit 7 表示是否支持 MLD version 1。
    - bit 6 表示是否支持 MLD version 2。
    - bit 5 目前固定值为 0。
    - bit 4 表示携带的(S，G)的模式，取值为 1，表示 Exclude 模式;取值为 0，表示 Include 模式。该 bit 位仅在 bit 6 取值为 1 时有效，bit 6 取值为 0 忽略该 bit 位。

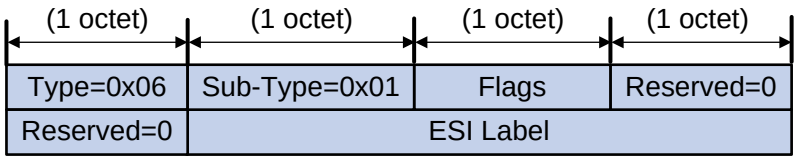
1.5 BGP EVPN路由的扩展团体属性

为了配合不同类型 BGP EVPN 路由实现不同的功能，BGP EVPN 定义了多种扩展团体属性。

1.5.1 ESI Label Extended Community

Ethernet Auto-discovery Route 路由中携带该扩展团体属性，用来实现水平分割和识别冗余备份模式。

图13 ESI Label Extended Community 报文格式



如图 13 所示，ESI Label Extended Community 包含如下字段：

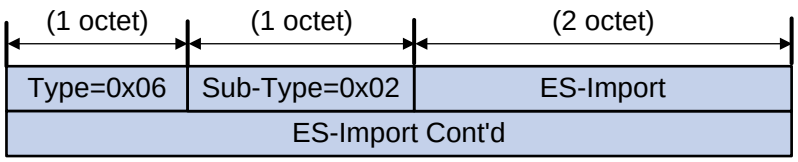
- Flags：该字段的最后一个 bit 位用于标识多归属的冗余备份模式。取值为 0，表示多活冗余模式；取值为 1，表示单活冗余模式。

- ESI Label：用于在 EVPN 多归属组网中实现水平分割。不同的数据封装类型下，该字段对应不同的取值：
  - MPLS 封装时，为 MPLS Label。
  - VXLAN 封装时，该字段无意义。
  - SRv6 封装时，为 SID 的 argument。

1.5.2 ES-Import Route Target Extended Community

Ethernet Segment Route 中携带该扩展团体属性，用于通告 ES 的 Route target 属性。

图14 ES-Import Route Target Extended Community 报文格式

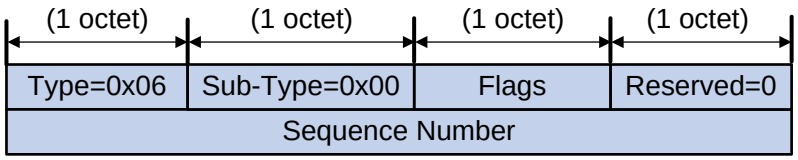


如图 14 所示，ES-Import Route Target Extended Community 中 ES-Import 和 ES-Import Cont'd 字段一起表示根据 ESI 自动生成的 Route target 属性值。

1.5.3 MAC Mobility Extended Community

当主机发生迁移时，携带在 MAC/IP Advertisement Route 路由中，用于标识主机发生迁移的次数。

图15 MAC Mobility Extended Community 报文格式



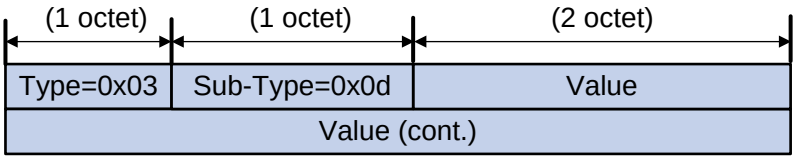
如图 15 所示，MAC Mobility Extended Community 包含如下字段：

- Flags：该字段的最后一个 bit 位用于标识是否为静态 MAC。取值为 1，表示该 MAC 地址为静态 MAC，不可迁移。
- Reserved：保留字段。
- Sequence Numbe：标记 MAC 迁移的次数。

1.5.4 Default Gateway Extended Community

EVPN VXLAN 分布式网关组网中，携带在 MAC/IP Advertisement Route 路由中，表示本地址是网关地址。

图16 Default Gateway Extended Community 报文格式

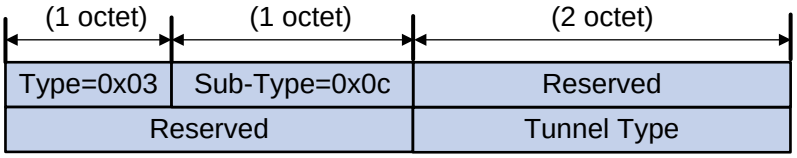


如图 16 所示，Default Gateway Extended Community 中 Value 和 Value (cont.)字段取值均为 0。

1.5.5 Encapsulation Type Extended Community

所有 BGP EVPN 路由均可以携带该扩展团体属性，用于标识报文的封装类型。默认报文封装类型为 MPLS 封装。因此，采用 MPLS 封装时，BGP EVPN 路由中可以不携带该属性。

图17 Encapsulation Type Extended Community 报文格式



如图 17 所示，Encapsulation Type Extended Community 包含如下字段：

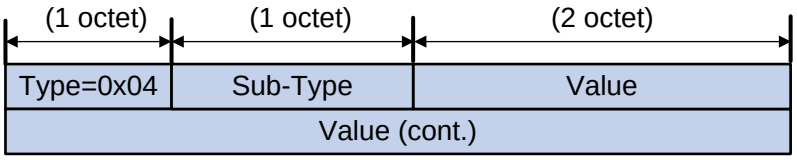
- Reserved：保留字段。
- Tunnel Type：封装类型。该字段不同取值代表不同的封装类型：
  - 8：VXLAN 封装。
  - 9：NVGRE 封装。
  - 10：MPLS 封装。
  - 11：MPLS in GRE 封装。
  - 12：VXLAN GPE 封装。

1.5.6 VPN Target Extended Community（也称为 Route Target）

所有 BGP EVPN 路由均需要携带 VPN Target 扩展团体属性，通过 VPN Target 属性来控制 EVPN 路由信息的发布与接收：

- 本地 VTEP 在通过 BGP 的 Update 消息将 EVPN 路由发送给远端 VTEP 时，在 Update 消息中携带 VPN Target 属性（该属性称为 Export target 属性）。
- 远端 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 VPN target 属性（该属性称为 Import target 属性）进行匹配，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。

图18 VPN Target Extended Community 报文格式



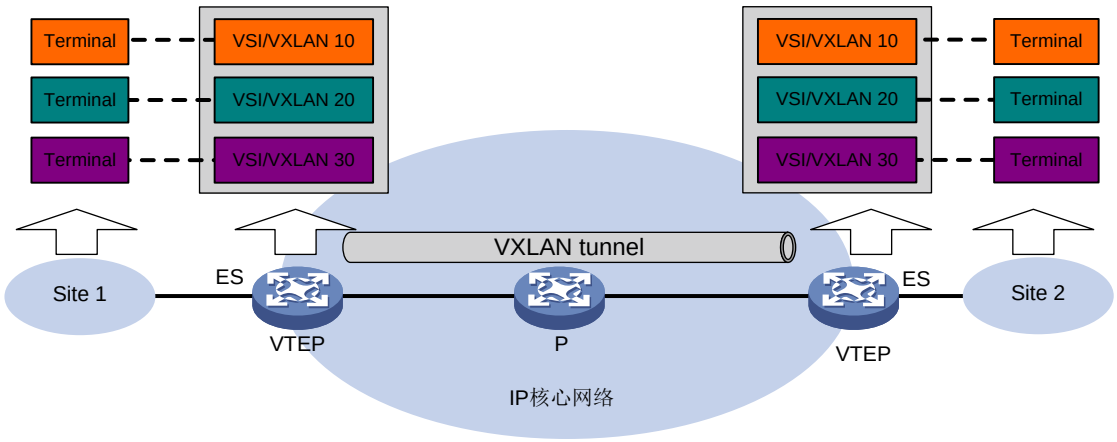
如图18所示，VPN Target Extended Community中Value和Value (cont.)字段一起表示Route Target。Route Target取值有如下三种格式：

- 16位自治系统号:32位用户自定义数，例如：101:3。
- 32位IP地址:16位用户自定义数，例如：192.168.122.15:1。
- 32位自治系统号:16位用户自定义数字，其中的自治系统号最小值为65536。例如：65536:1。

2 EVPN VXLAN

2.1 EVPN VXLAN网络模型

图19 EVPN VXLAN 网络模型示意图



如图 19 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。

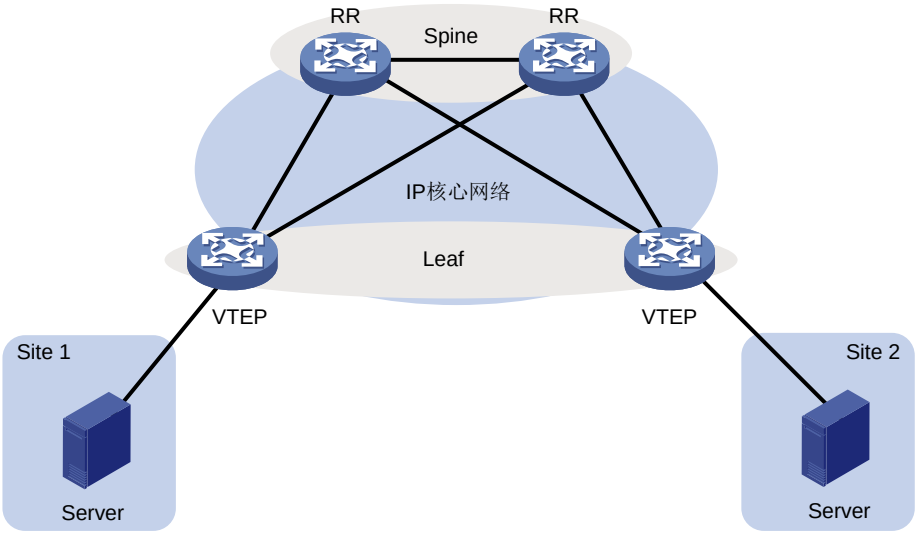
- VTEP（VXLAN Tunnel End Point，VXLAN 隧道端点）：EVPN VXLAN 的边缘设备。EVPN VXLAN 的相关处理都在 VTEP 上进行。根据 VTEP 功能，VTEP 可以划分为 L2 VTEP 和 GW 两种角色：
  - L2 VTEP：只支持二层 VXLAN 转发功能的设备，即只能在相同 VXLAN 内进行二层转发。
  - GW：可以进行跨 VXLAN 或者访问外部 IP 网络等三层转发的设备。EVPN VXLAN 网络根据 GW 的部署方式，可以分为集中式网关和分布式网关两种。
- VXLAN 隧道：两个 VTEP 之间的点到点逻辑隧道。VTEP 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 19 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 20 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20 EVPN VXLAN 分层组网模型



2.2 EVPN VXLAN控制平面工作机制

2.2.1 VXLAN 隧道及 BUM 广播表建立

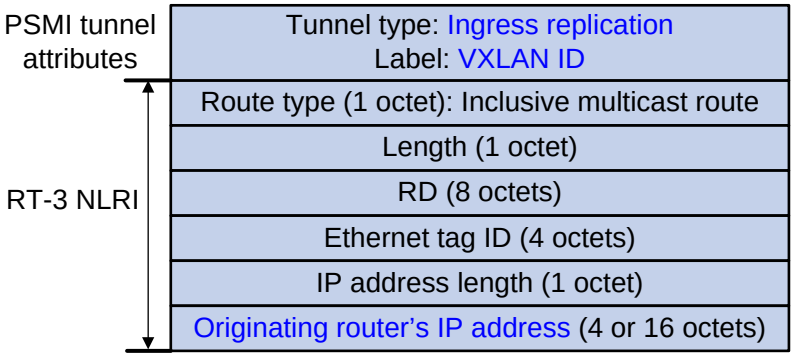
VXLAN 采用“MAC in UDP”封装，是一种在 IP 网络基础之上构建 Overlay 网络的技术。在 IP 网络上传输报文时，VXLAN 使用 Ingress Replication，即头端复制，来转发 BUM（Broadcast/Unknown unicast/Unknown Multicast，广播/未知单播/未知组播）流量。所谓头端复制，是指在 VXLAN 转发实体（VSI）中保存 BUM 流量需要通过哪些 VXLAN 隧道复制到远端 PE 设备，此 VXLAN 隧道列表叫做 BUM 广播表。

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 21 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21 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。

2.2.2 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 22 所示。

图22 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 23 所示，MAC 地址和 ARP/ND 信息的通告和学习过程为：

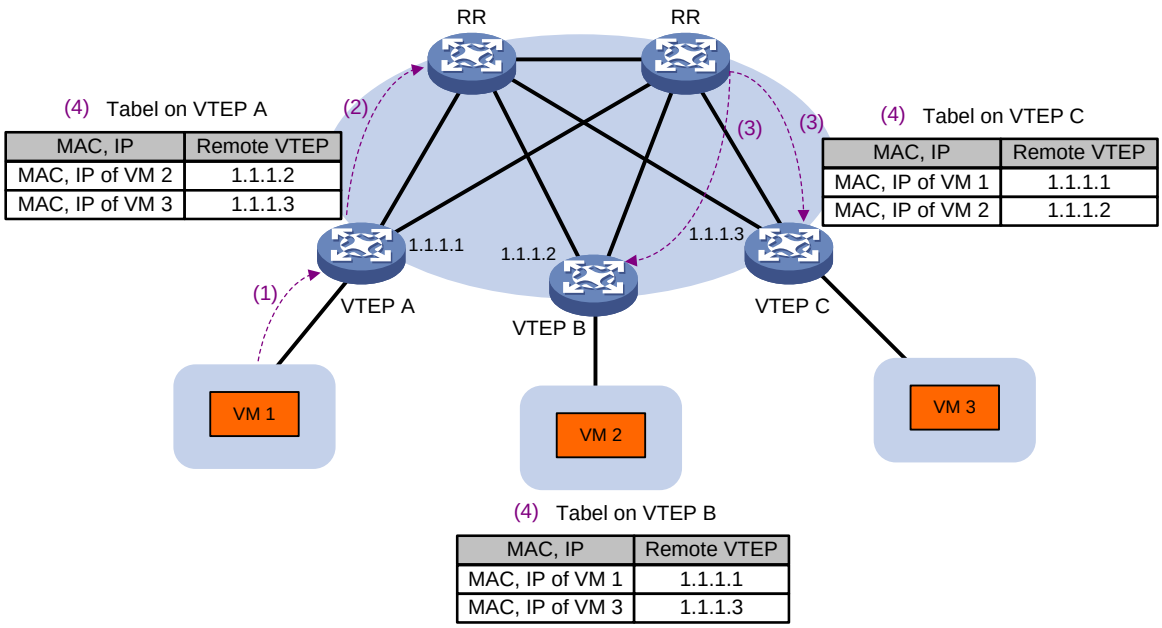
(1) VTEP 在数据平面完成本地 MAC 地址和 ARP/ND 信息的学习。本地 MAC 地址通过以太网报文的源 MAC 地址学习获得；ARP/ND 信息通过 ARP、免费 ARP、ND 等报文学习获得。

(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23 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。

2.2.3 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 24 所示。

图24 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 25 所示，外部路由通告与学习过程为：

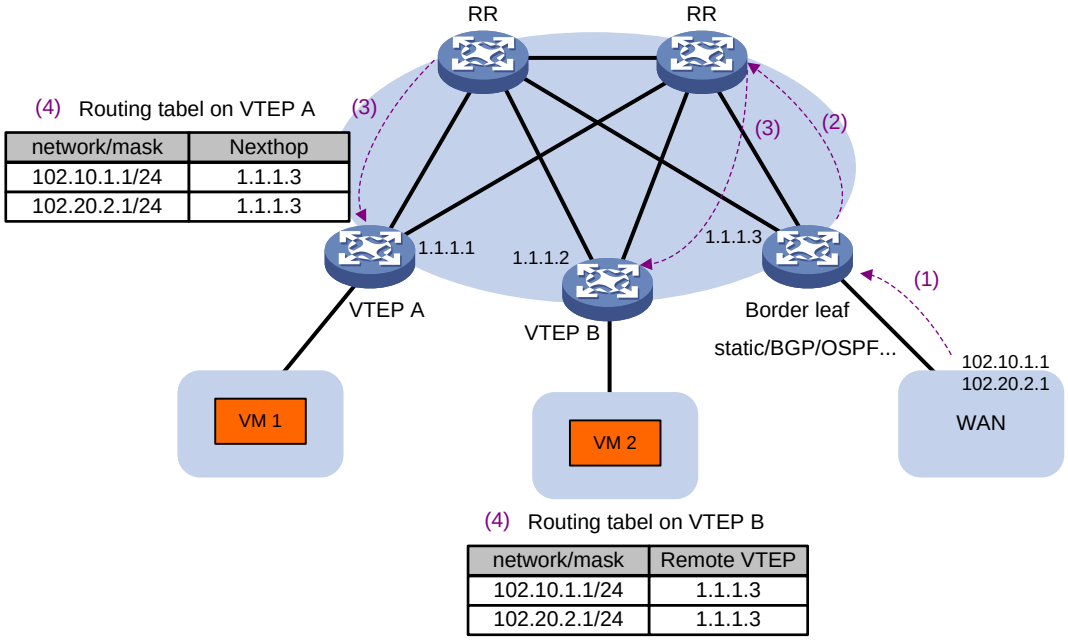
(1) Board leaf 与 WAN 网络之间配置静态路由，或运行 BGP、OSPF 等动态路由协议。Board leaf 学习到外网的路由。

(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25 外部路由通告与学习过程



2.2.4 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。

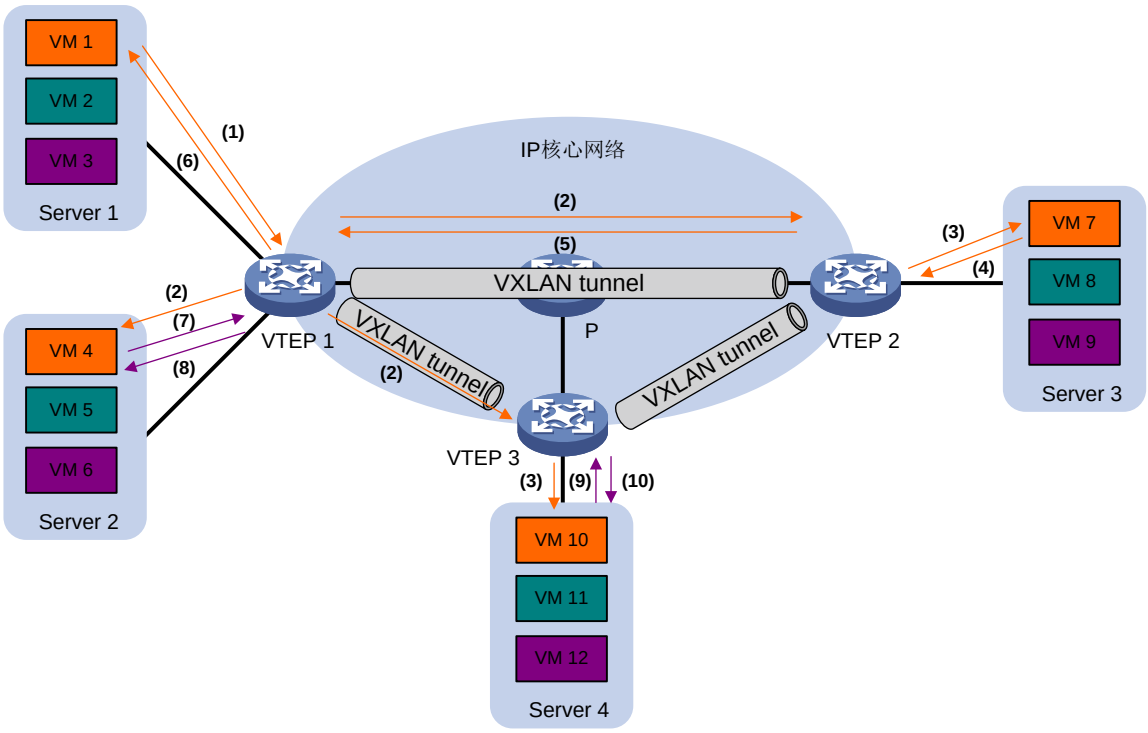
(3) 远端 VTEP 接收到比自己本地保存的序列号更大的 MAC/IP 路由时，更新自己的 MAC/IP 路由消息，下一跳指向迁移后通告此路由的 VTEP。

(4) 原 VTEP 在收到此路由更新后，撤销之前通告的路由。

2.2.5 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26 ARP 泛洪抑制示意图



如图 26 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 26 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

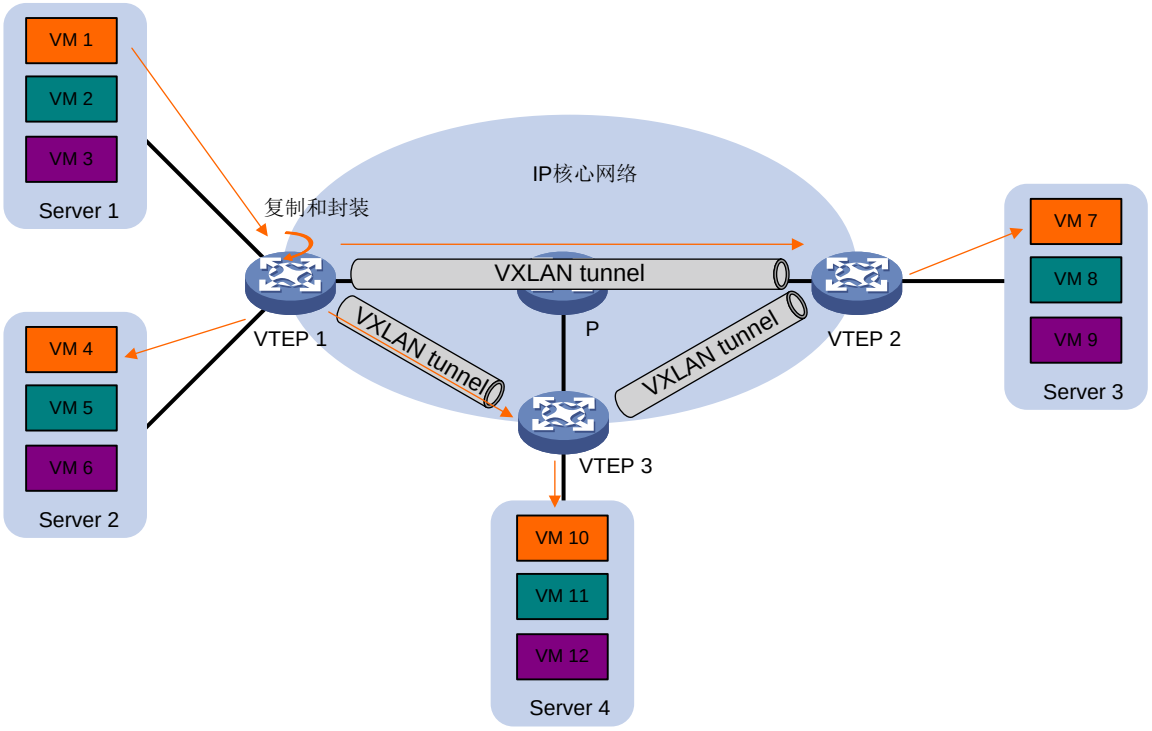
EVPN VXLAN 通过控制平面完成 MAC 地址表项的学习。VTEP 接收到二层数据帧后，判断其所属的 VSI，根据目的 MAC 地址查找该 VSI 的 MAC 地址表，通过表项的出接口转发该数据帧。如果出接口为本地接口，则 VTEP 直接通过该接口转发数据帧；如果出接口为 Tunnel 接口，则 VTEP 根据 Tunnel 接口为数据帧添加 VXLAN 封装后，通过 VXLAN 隧道将其转发给远端 VTEP。

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27 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

在分布式网关对称 IRB 转发方式中，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。

1. 基本概念

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。

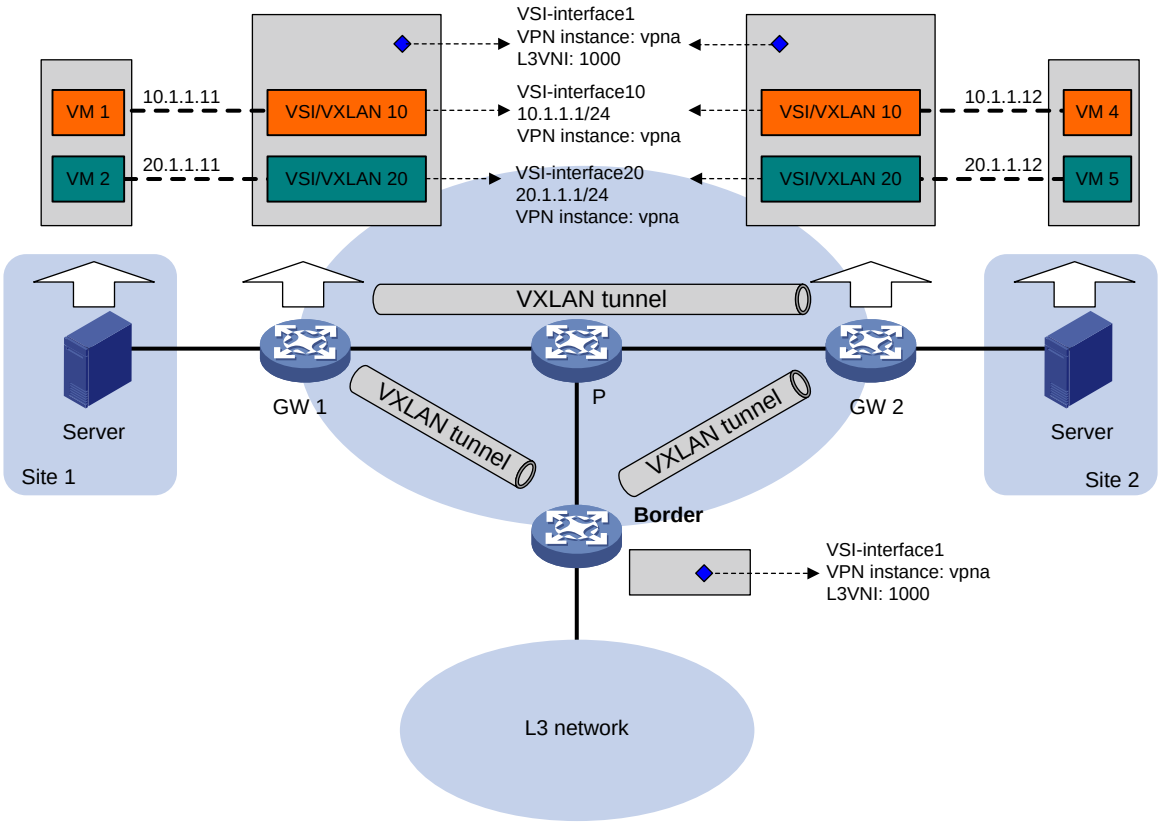
2. 分布式 EVPN 网关部署要求

如图 28 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28 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 29 所示；不同站点间三层流量转发过程如图 30 所示。

图29 相同站点间三层流量转发过程

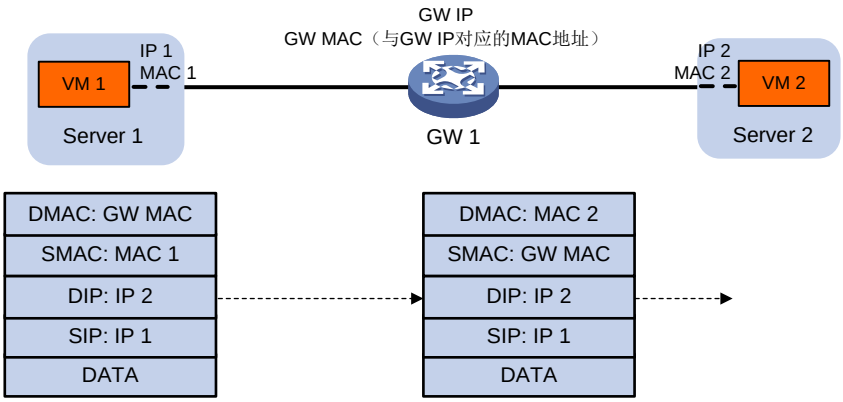
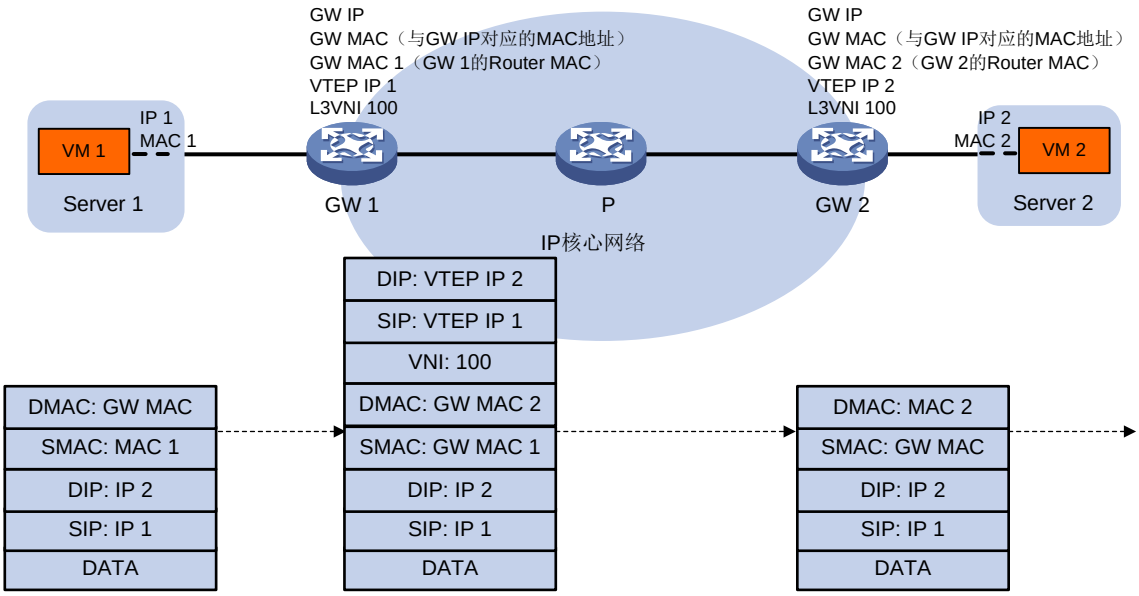


图30 不同站点间三层流量转发过程



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。

2.3.4 分布式网关非对称 IRB 转发

在分布式网关非对称 IRB 转发方式中，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。

1. 分布式 EVPN 网关部署要求

非对称 IRB 与对称 IRB 方式中，分布式 EVPN 网关的部署方式基本相同。

如图 28 所示，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 地址不能相同。
- 承载 L3VNI 的 VSI 虚接口。在非对称 IRB 转发方式中，L3VNI 用来实现 VXLAN 网络与外界网络的互通。当 VXLAN 内的虚拟机需要通过边界网关（Border）与外界通信时，GW 上必须部署该类 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。

边界网关上也需要存在承载 L3VNI 的 VSI 虚接口。

2. 三层流量转发过程

目前，非对称 IRB 转发方式仅支持通过分布式 EVPN 网关转发相同 VXLAN 的三层流量。

在非对称 IRB 转发方式中，GW 学习到本地虚拟机的 ARP 信息后，通过 MAC/IP 发布路由将其通告给其他 GW。其他 GW 学习 ARP 信息，并生成对应的 FIB 表项。

如图 31 所示，VM 1 和 VM 2 属于 VXLAN 10，通过分布式 EVPN 网关实现三层互通。分布式 EVPN 网关采用非对称 IRB 方式转发三层流量的过程为：

(1) GW 1 接收到 VM 1 发送的报文后，由于目的 MAC 地址为自己，GW 1 剥离二层帧头，根据目的 IP 地址查找 FIB 表。

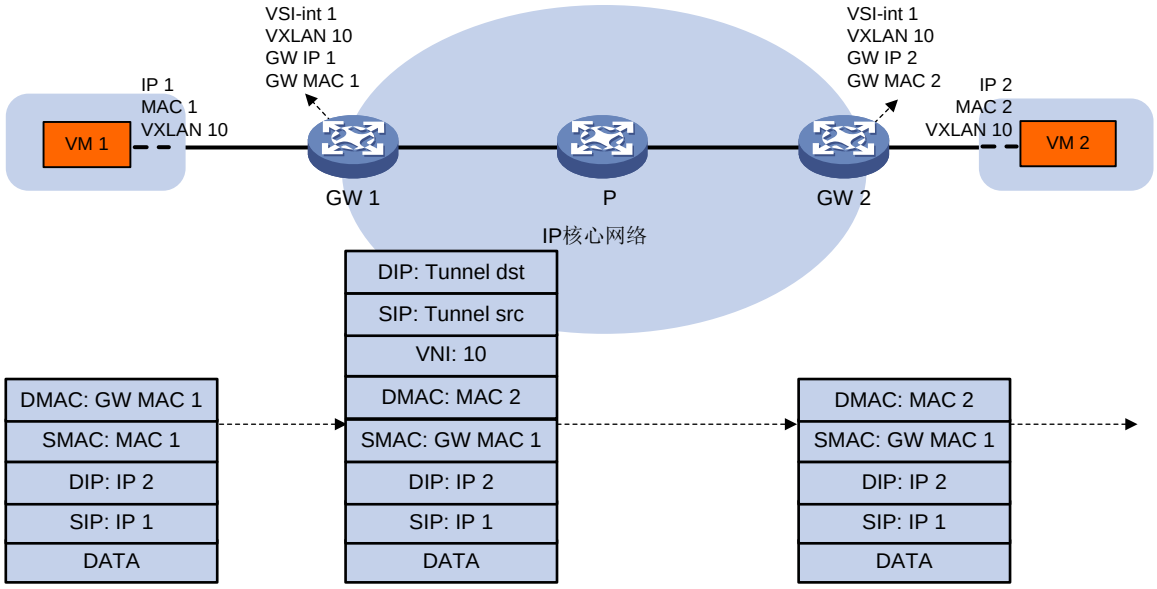
(2) GW 1 在 FIB 表中匹配到 VM 2 的 ARP 信息生成的 FIB 表项。

(3) GW 1 为报文封装源和目的 MAC 地址（分别为网关 MAC 地址和 VM 2 的 MAC 地址）、VXLAN 头后，通过 VXLAN 隧道将其转发到 GW 2。

(4) GW 2 接收到报文后，解除 VXLAN 封装，并在 VXLAN 10 内进行二层转发，即根据目的 MAC 地址查找 MAC 地址表。

(5) GW 2 根据 MAC 地址表查找结果，将报文转发给 VM 2。

图31 非对称 IRB 三层流量转发过程

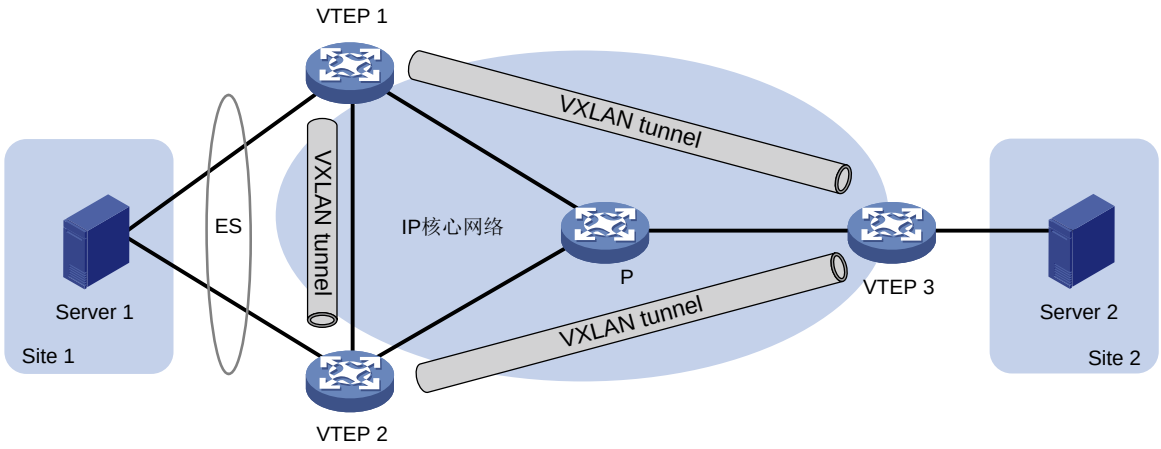


2.4 EVPN VXLAN多归属

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 网络的可靠性。

图32 多归属站点示意图



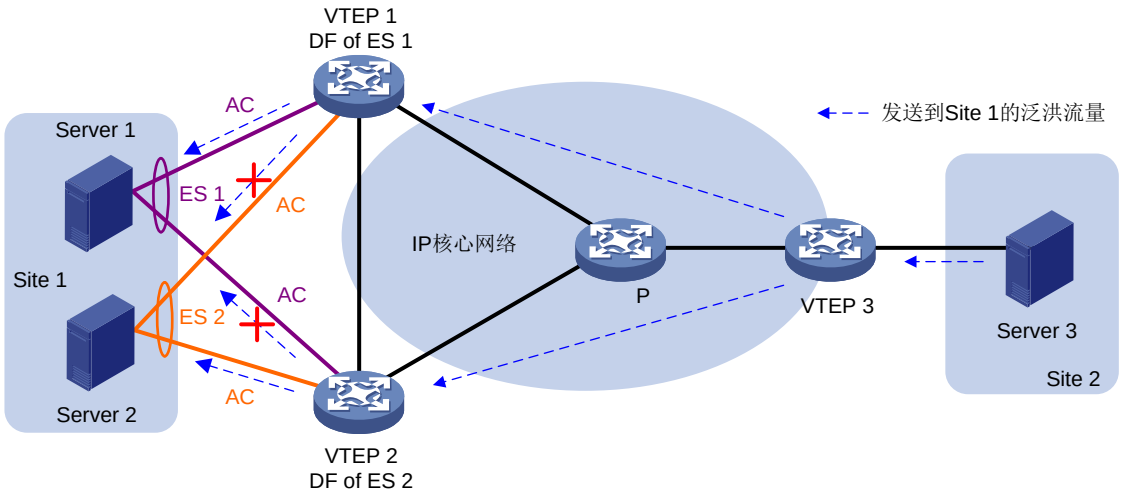
2.4.2 DF 选举

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

转发给本地站点。其他 VTEP 作为 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。

设备支持多种 DF 选举算法，用户可以根据业务需要灵活地选择 DF 选举算法，使组网中 DF 能够均匀分布，提高网络设备的使用率。

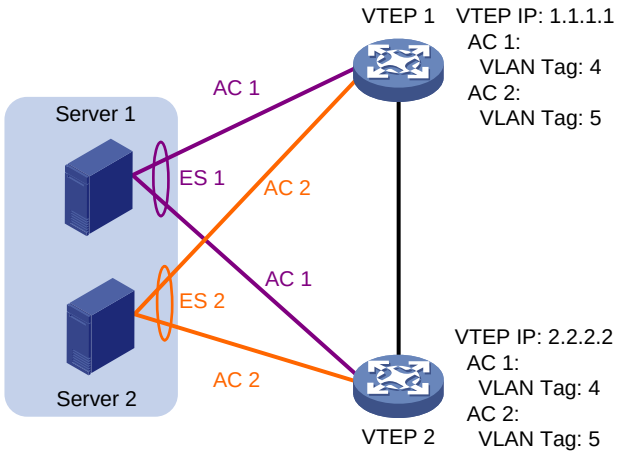
图33 DF 示意图



2. 基于 VLAN Tag 的 DF 选举算法

基于 VLAN Tag 的 DF 选举算法根据 VLAN Tag 和 VTEP 的 IP 地址为每个 AC 选举 DF。

图34 基于 VLAN Tag 的 DF 选举



如图 34 所示，以允许 VLAN Tag 4 通过的 AC 1 的 DF 选举为例，基于 VLAN Tag 的 DF 选举算法为：

(2) 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。

(3) 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。

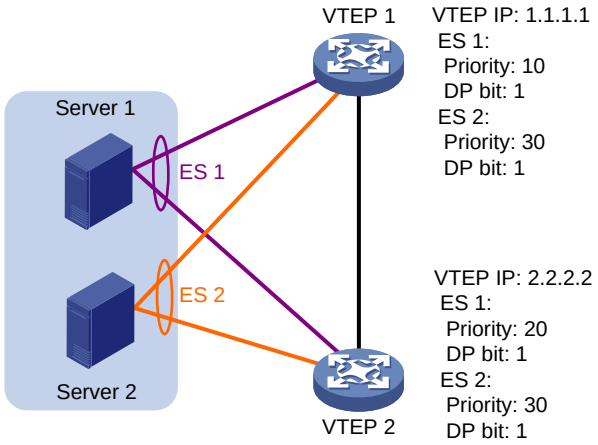
(4) 根据 VLAN Tag 除以 N 的余数 M 来选举 DF，N 代表冗余备份组中成员的数量，M 对应的编号为该 AC 的 DF。在本例中，4 除以 2 的余数为 0，即 AC 1 的 DF 为编号为 0 的 VTEP 1。

3. 基于优先级的 DF 选举算法

基于优先级的 DF 选举算法根据 DF 选举优先级、DP（Don't Preempt Me，不可回切）位和 VTEP 的 IP 地址为每个 ES 选举 DF。其中，DP 位的取值包括：

- 1：表示开启了基于优先级 DF 选举算法不回切功能。即当前设备被选举为 DF 后，即使后续选举出了新的设备作为 DF，依然使用当前设备作为 DF。
- 0：表示关闭了基于优先级 DF 选举算法不回切功能。即当前设备被选举为 DF 后，如果后续选举出了新的设备作为 DF，则直接使用新的设备作为 DF。

图35 基于优先级的 DF 选举



如图 35 所示，以 ES 1、ES 2 的 DF 选举为例，基于优先级的 DF 选举算法为：

(1) 同一 ES 内 DF 选举优先级（数值越大则优先级越高）最高的 VTEP 作为该 ES 的 DF。在本例中，选举 VTEP 2 作为 ES 1 的 DF。

(2) 若优先级相同，则 DP 位为 1 的 VTEP 作为 DF。

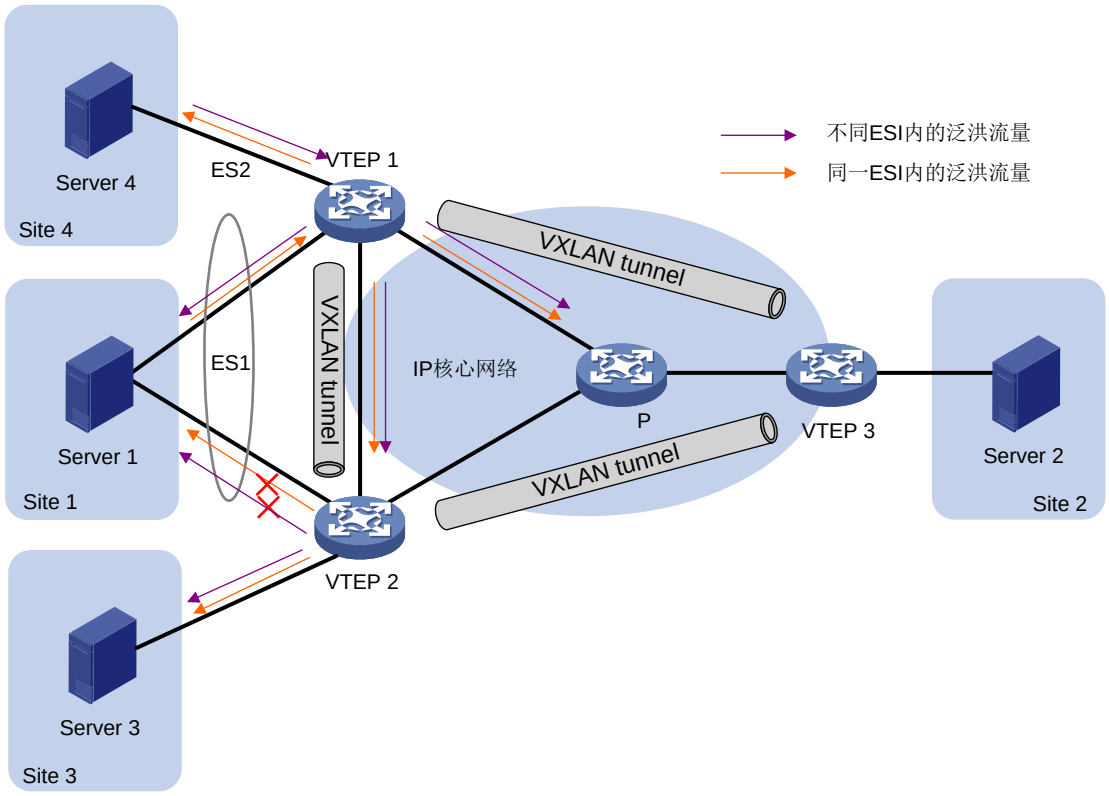
(3) 若 DP 位相同，则 IP 地址小的 VTEP 作为 DF。在本例中，选举 VTEP 1 作为 ES 2 的 DF。

2.4.3 协议报文交互过程

2.4.4 水平分割

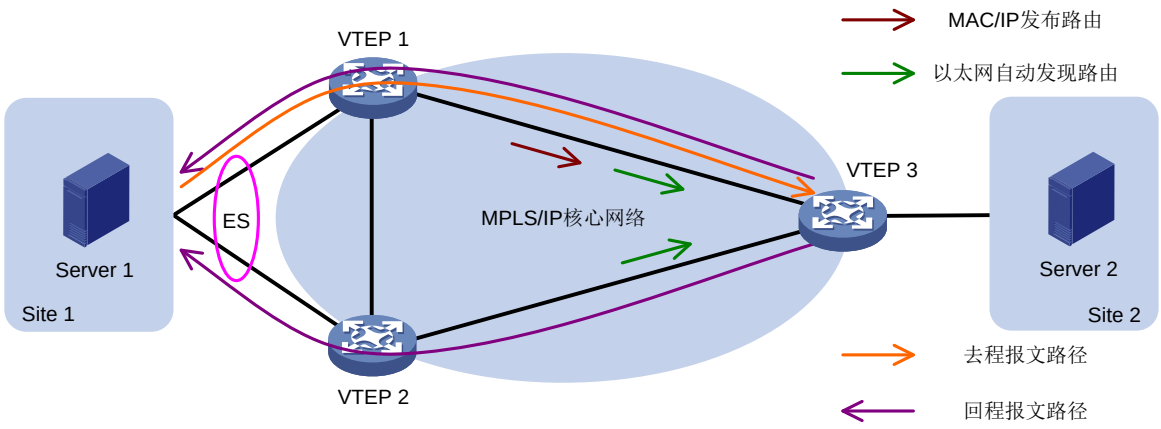
在多归属站点组网中，VTEP 接收到站点发送的组播、广播和未知单播数据帧后，判断数据帧所属的 VXLAN，通过该 VXLAN 内除接收接口外的所有本地接口和 VXLAN 隧道转发该数据帧。同一冗余备份组中的 VTEP 接收到该数据帧后会在本地所属的 VXLAN 内泛洪，这样数据帧会通过 AC 泛洪到本地站点，造成环路和站点的重复接收。EVPN 通过水平分割解决该问题。水平分割的机制为：VTEP 接收到同一冗余备份组中成员转发的广播、组播、未知单播数据帧后，不向具有相同 ESI 标识的 ES 转发该数据帧。

图36 水平分割



2.4.5 别名

图37 别名示意图



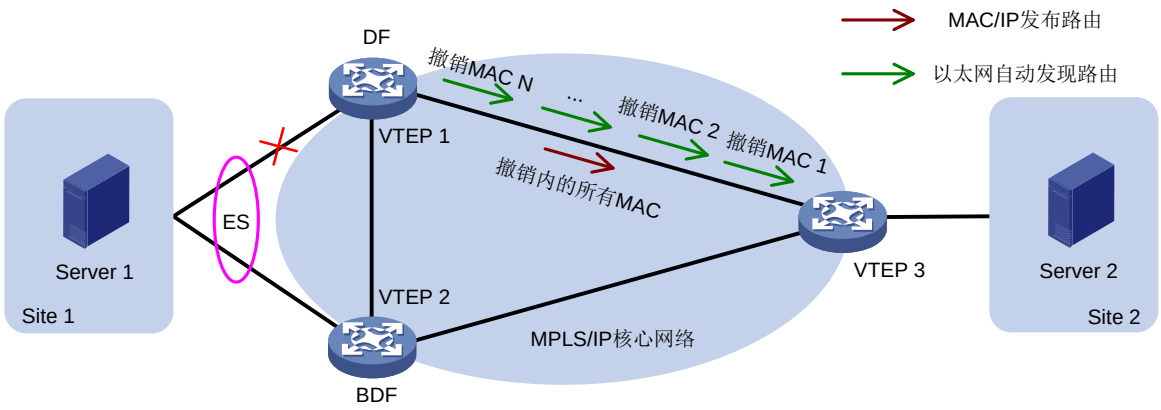
如图 37 所示，在多活冗余模式下，冗余备份组中可能仅有一台 VTEP 能学习到某些业务相关的 MAC 地址，这会导致远端 PE 仅能从这台 VTEP 收到这些 MAC 地址的 MAC/IP 发布路由，因此远端 VTEP 无法将访问这些 MAC 地址的流量负载分担到冗余备份组中的其它 VTEP 上。

为了解决这个问题，EVPN 多归属引入了别名机制，即当冗余备份组中仅有一台 VTEP 通过 MAC/IP 发布路由向远端 VTEP 通告了 Server 侧 MAC 地址的可达性时，远端 VTEP 能够根据冗余备份组内

VTEP 发送的以太网自动发现路由（携带 VTEP、ESI 等信息）感知到冗余备份组中其它 VTEP 与 MAC 地址的可达性，并生成对应的 MAC 表项，从而形成负载分担。

2.4.6 MAC 地址快速收敛

图38 MAC 地址快速收敛示意图



如图 38 所示，在 EVPN 网络中，MAC 地址可达性是通过 VTEP 之间发布 MAC/IP 发布路由通告的。因此，在 CE 1 与 VTEP 1 间链路故障时，VTEP 1 需要逐条撤销 MAC/IP 发布路由，在大规模的网络中会导致 MAC 地址收敛速度较慢。

EVPN 多归属组网提供了快速收敛机制，使得 VTEP 可以通过撤销一条以太网自动发现路由，通告对指定 ES 内所有 MAC 地址的不可达性，通知远端 VTEP 批量删除 MAC 地址表项，减少收敛时间。

2.5 EVPN VXLAN支持组播

2.5.1 功能简介

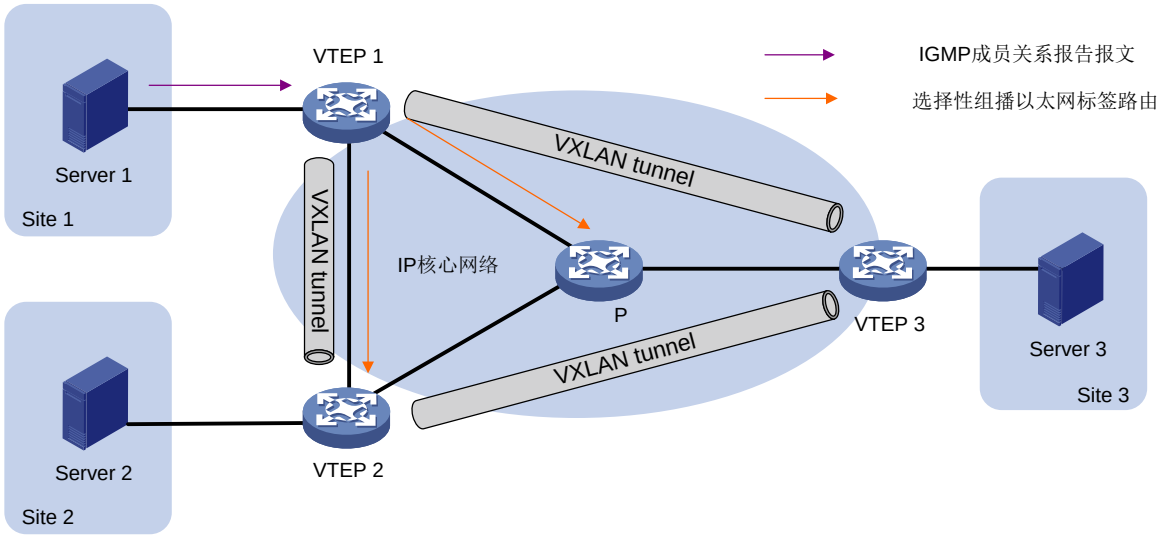
为了避免组播发送的 IGMP 报文占用核心网络带宽，VTEP 会根据接收到的报告报文和离开报文在本地建立或删除组播转发表项。通过 SMET（Selective Multicast Ethernet Tag Route，选择性组播以太网标签路由）路由将组播组信息通告给其他 VTEP，远端 VTEP 收到 SMET 路由后在本地建立组播转发表项。当 VTEP 再次收到属于同一 IGMP 版本加入同一组播组的报告报文时，将不再发送 SMET 路由。EVPN VXLAN 支持组播功能可以大大减少 IGMP 报文泛洪的次数。

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“1.4 BGP EVPN 路由”。

2.5.2 单归属站点组播

如图 39 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。

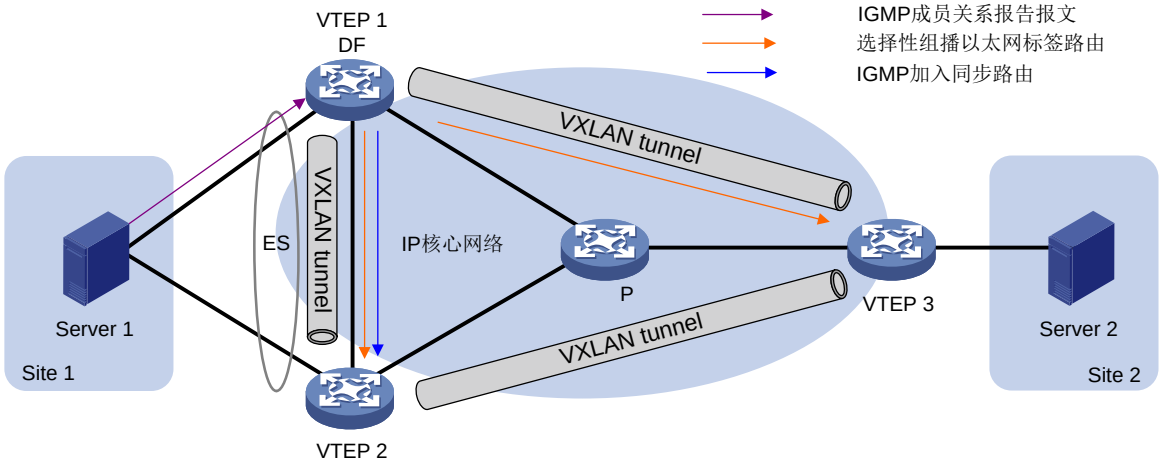
图39 单归属站点组播示意图



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40 多归属站点组播示意图



如图 40 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。

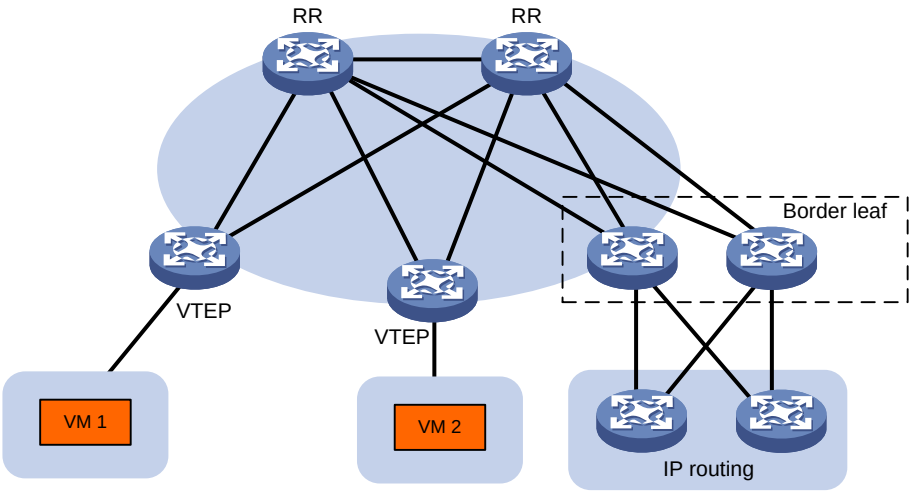
2.6 典型组网应用

2.6.1 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 41 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41 EVPN 分布式网关组网示意图

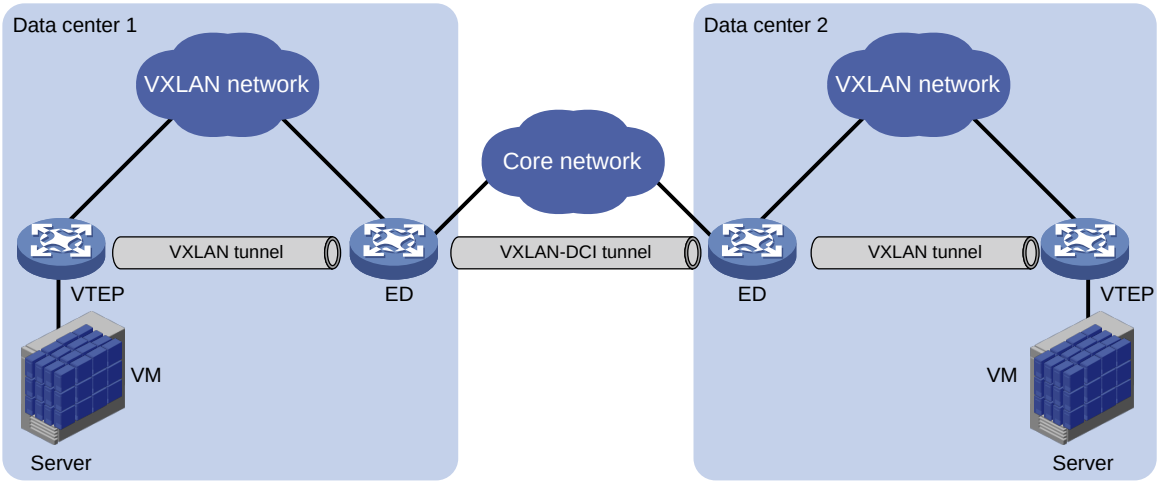


2.6.2 EVPN 数据中心互联组网

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 42 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。

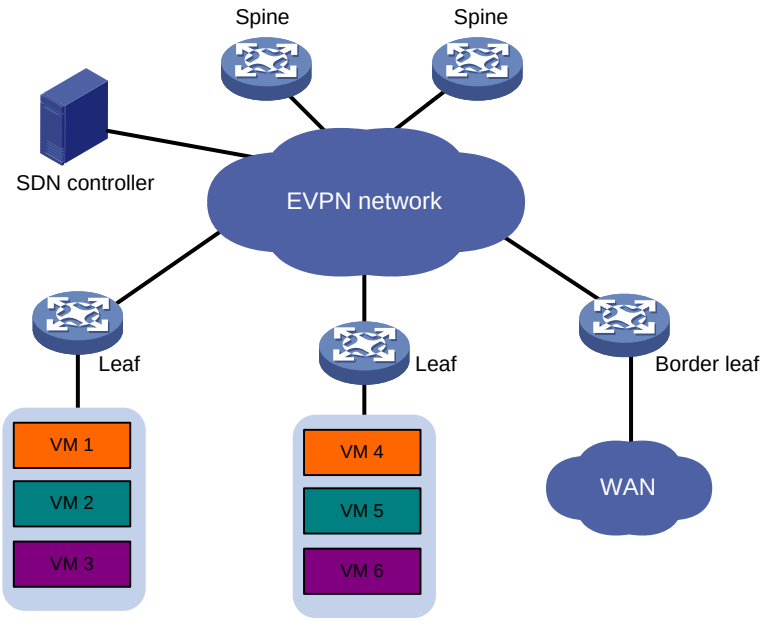
图42 VXLAN 数据中心互联典型组网图



2.6.3 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 43 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。

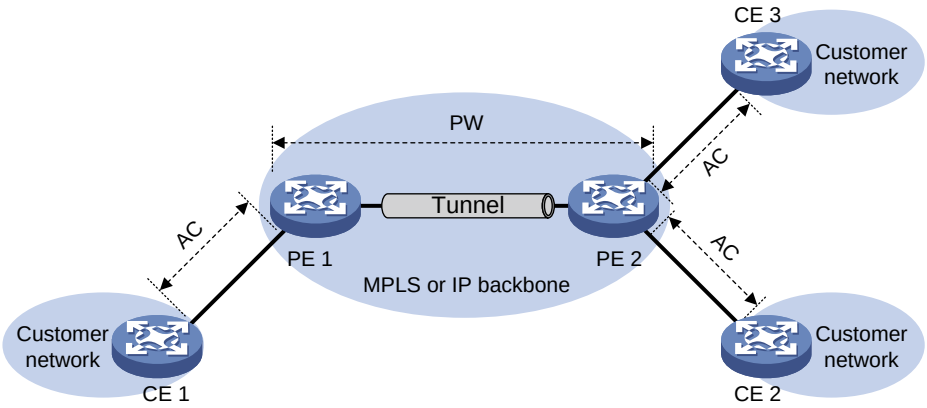
图43 EVPN 与 SDN 控制器配合组网



3 EVPN VPLS

3.1 EVPN VPLS 网络模型

图44 EVPN VPLS 网络模型示意图



如图 44 所示，EVPN VPLS 网络中主要包括如下几部分：

- CE（Customer Edge，用户网络边缘）：直接与服务提供商网络相连的用户网络侧设备。
- PE（Provider Edge，服务提供商网络边缘）：与 CE 相连的服务提供商网络侧设备。PE 主要负责 EVPN VPLS 业务的接入，完成报文从用户网络到公网隧道、从公网隧道到用户网络的映射与转发。
- AC（Attachment Circuit，接入电路）：连接 CE 和 PE 的物理电路或虚拟电路。
- PW（Pseudowire，伪线）：两个 PE 之间的虚拟双向连接。PW 由一对方向相反的单向虚拟连接构成。
- 公网隧道（Tunnel）：穿越 IP 或 MPLS 骨干网、用来承载 PW 的隧道。一条公网隧道可以承载多条 PW，公网隧道可以是 LSP、GRE 隧道或 MPLS TE 隧道。
- VSI（Virtual Switch Instance，虚拟交换实例）：VSI 是 PE 设备上为一个 VPLS 实例提供二层交换服务的虚拟实例。VSI 可以看作 PE 设备上的一台虚拟交换机，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。VPLS 通过 VSI 实现在 VPLS 实例内转发二层数据报文。

3.2 EVPN VPLS控制平面工作机制

3.2.1 建立 PW

EVPN VPLS 组网中，PW 的建立过程为：

(1) PE 为每个 VSI 实例分配两个 PW 标签，分别用于转发已知单播报文和 BUM（Broadcast/Unknown unicast/Unknown Multicast，广播/未知单播/未知组播）报文。

(2) 本端 PE 通过 MAC/IP 发布路由将转发已知单播报文的 PW 标签通告给远端 PE；通过 IMET 路由将转发 BUM 报文的 PW 标签通告给远端 PE。路由中携带 VPN Target 属性。

(3) 远端 PE 接收到 MAC/IP 发布路由或 IMET 路由后，将路由中的 VPN Target 属性与 EVPN 实例的 Import Target 进行匹配，如果一致则根据路由中携带的 PE 地址（对于 MAC/IP 发布路由，为路由的下一跳地址；对于 IMET 路由，为 PSMI tunnel attributes 中 Tunnel Identifier 字段携带的地址）、PW 标签等信息建立一条单向的虚拟连接。

(4) 当两端的 PE 间建立了两条方向相反的单向虚拟连接，则 PW 建立完成。

3.2.2 MAC 地址学习、老化和回收

1. MAC 地址学习

PE 根据学习到的 MAC 地址表项转发二层单播流量。PE 上 MAC 地址学习分为两部分：

- 本地 MAC 地址学习：PE 接收到本地 CE 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 CE 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。
- 远端 MAC 地址学习：PE 通过 MAC/IP 发布路由将本地学习的 MAC 地址通告给远端 PE。远端 PE 接收到该信息后，将其添加到对应的 VSI 的 MAC 地址表中，该 MAC 地址的出接口为两个 PE 之间 PW 的索引。

2. MAC 地址老化

- 本地 MAC 地址老化：PE 学习本地 MAC 地址后，如果 MAC 地址老化定时器超时，则删除该 MAC 地址表项，减少占用的 MAC 地址表资源。
- 远端 MAC 地址老化：PE 从 MAC/IP 发布路由中学习远端 MAC 地址，在接收到撤销该 MAC 地址的路由前，MAC 地址会一直存在 MAC 地址表中。

3. MAC 地址回收

AC 状态变为 down 时，EVPN 会向所有远端 PE 发送 MAC/IP 发布路由撤销消息来撤销该 AC 对应的 MAC 地址，远端 PE 根据撤销消息删除指定 VSI 内的指定 MAC 地址，以加快 MAC 地址表的收敛速度。

3.2.3 MAC 地址迁移

MAC 地址迁移是指主机/虚拟机从其接入的 PE 迁到另一台 PE 下。EVPN VPLS 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) PE 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 PE 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。

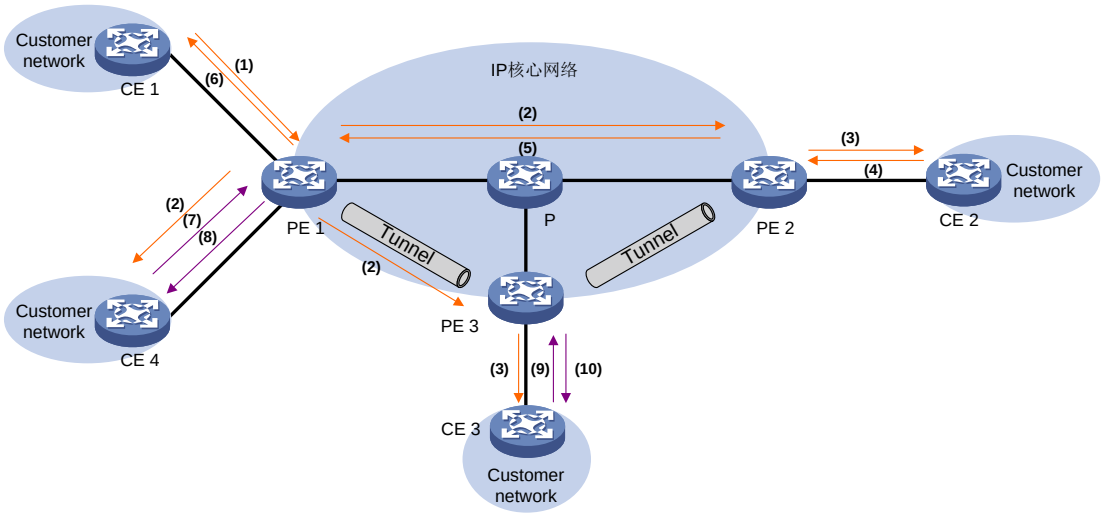
(3) 远端 PE 接收到比自己本地保存的序列号更大的 MAC/IP 路由时，更新自己的 MAC/IP 路由消息，下一跳指向迁移后通告此路由的 PE。

(4) 原 VTEP 在收到此路由更新后，撤销之前通告的路由。

3.2.4 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，PE 会根据接收到的 ARP 请求和 ARP 应答报文、BGP EVPN 路由在本地建立 ARP 泛洪抑制表项。当 PE 再收到本地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 ARP 泛洪抑制表项进行代答。如果没有对应的表项，则通过 PW 将 ARP 请求泛洪到其他站点。ARP 泛洪抑制功能可以大大减少 ARP 泛洪的次数。

图45 ARP 泛洪抑制示意图



如图 45 所示，ARP 泛洪抑制的处理过程如下：

(1) 虚拟机 CE 1 发送 ARP 请求，获取 CE 2 的 MAC 地址。

(2) PE 1 根据接收到的 ARP 请求，建立 CE 1 的 ARP 泛洪抑制表项，向 VSI 内的本地 CE 和远端 PE（PE 2 和 PE 3）泛洪该 ARP 请求（图 45 以单播路由泛洪方式为例）。PE 1 还会通过 BGP EVPN 将该表项同步给 PE 2 和 PE 3。

(3) 远端 PE 解封装报文，获取原始的 ARP 请求报文后，向 VSI 内的本地 CE 泛洪该 ARP 请求。

(4) CE 2 接收到 ARP 请求后，回复 ARP 应答报文。

(5) PE 2 接收到 ARP 应答后，建立 CE 2 的 ARP 泛洪抑制表项，通过 PW 将 ARP 应答发送给 PE 1。PE 2 通过 BGP EVPN 将该表项同步给 PE 1 和 PE 3。

(6) PE 1 解封装报文并获取原始的 ARP 应答，将 ARP 应答报文发送给 CE 1。

(7) 在 PE 1 上建立 ARP 泛洪抑制表项后，CE 4 发送 ARP 请求，获取 CE 1 的 MAC 地址。

(8) PE 1 接收到 ARP 请求后，建立 CE 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) CE 3 发送 ARP 请求，获取 CE 1 的 MAC 地址。

(10) PE 3 接收到 ARP 请求后，建立 CE 3 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（PE 1 通过 BGP EVPN 同步）回复 ARP 应答报文，不会对 ARP 请求进行泛洪。

3.3 EVPN VPLS数据平面工作机制

3.3.1 本地站点接入模式

本地站点可以通过以下几种方式接入 EVPN VPLS 网络：

- 端口模式

  本地站点通过三层以太网接口接入 EVPN VPLS 网络。从该接口收到的所有报文都属于三层以太网接口关联的 VSI。

  在这种接入模式下，三层以太网接口作为 AC。

- VLAN 模式

  本地站点通过三层以太网子接口接入 EVPN VPLS 网络。从三层以太网接口接收到的、所有被该子接口终结的 VLAN 的报文都属于三层以太网子接口关联的 VSI。

  在这种接入模式下，三层以太网子接口作为 AC。

- 灵活匹配模式

  本地站点通过二层以太网接口上的以太网服务实例接入 EVPN VPLS 网络。通过以太网服务实例的报文匹配规则（如匹配接口接收到的所有报文、所有携带 VLAN Tag 的报文和所有不携带 VLAN Tag 的报文等），灵活匹配来自用户网络的报文。从接口接收到的、符合报文匹配规则的报文，属于以太网服务实例关联的 VSI。

  在这种接入模式下，以太网服务实例作为 AC。

VTEP 从本地站点接收到报文后，根据接入模式判断报文所属的 VSI，以便在 VSI 内转发该报文。

3.3.2 流量转发

1. 转发已知单播流量

- PE 从 AC 接收到已知单播报文后，会在对应的 VSI 内查找 MAC 地址表，从而确定如何转发报文：
  - 表项的出接口为 PW 索引时，为报文封装 PW 标签（用于转发已知单播报文的 PW 标签），再添加公网隧道封装后，通过 PW 将该报文转发给远端 PE。如果公网隧道为 LSP 或 MPLS TE 隧道，则通过 PW 转发报文时将为报文封装两层标签。内层标签为 PW 标签，用来将报文转发给相应的 VSI；外层标签为公网 LSP 或 MPLS TE 隧道标签，用来保证报文在 PE 之间正确传送。
  - 表项的出接口为连接本地 CE 的接口时，直接通过出接口将报文转发给本地 CE。
- PE 从 PW 接收到已知单播报文后，在其所属的 VSI 内查找 MAC 地址表，出接口应为连接本地站点的接口，PE 通过该出接口将报文转发给本地站点。

2. 转发泛洪流量

PE 从 AC 上接收到泛洪流量后，向该 AC 关联的 VSI 内的所有其他 AC 泛洪该报文，并查找该 VSI 内所有用于转发 BUM 流量的 PW 标签，为报文分别封装这些 PW 标签后，将该报文泛洪给所有远端 PE。

PE 从 PW 上接收到泛洪流量后，向该 PW 所属 VSI 内的所有 AC 泛洪该报文。

3.3.3 全连接和水平分割

为避免环路，一般的二层网络都要求使用环路预防协议，比如 STP（Spanning Tree Protocol，生成树协议）。但在骨干网的 PE 上部署环路预防协议，会增加管理和维护的难度。因此，EVPN VPLS 采用如下方法避免环路：

- PE 之间建立全连接，即一个 EVPN 实例内的每两个 PE 之间必须都建立 PW。
- 采用水平分割转发规则，即从 PW 上收到的泛洪报文禁止向其他 PW 转发，只能转发到 AC。

3.4 EVPN VPLS多归属

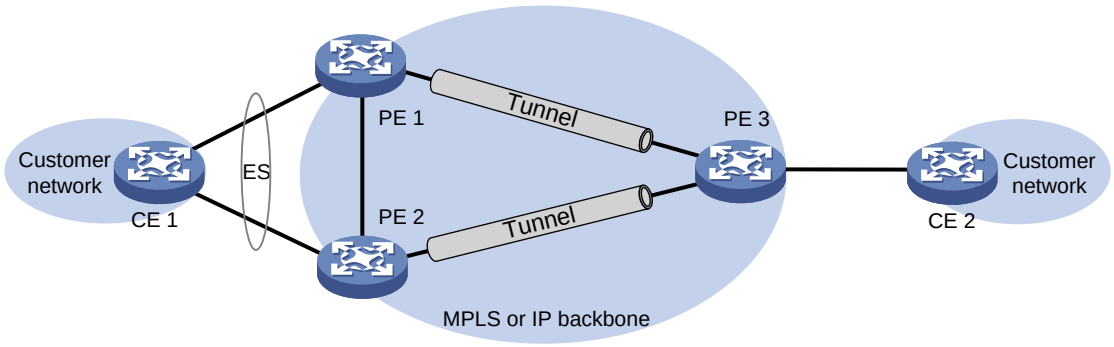
3.4.1 功能简介

EVPN 多归属是指一个站点通过不同的以太网链路接入 EVPN 网络中的多台 PE，接入的多台 PE 组成冗余备份组，该站点的流量在多台 PE 间进行负载分担。利用多归属技术可以避免 PE 单点故障造成 EVPN 网络通信中断，从而提高 EVPN 网络的可靠性。

EVPN 多归属的网络模型如图 46 所示，其中：

- 站点 CE 1 接入的多台 PE 组成冗余备份组。
- 接入冗余备份组中不同 PE 的一组链路，组成一个 ES（Ethernet Segment，以太网段），它们具有相同的 ESI （Ethernet Segment Identifier，以太网段标识）。
- 通过 ES 接入冗余备份组的站点，称为多归属站点。

图46 多归属站点示意图



3.4.2 DF 选举

当一个 CE 连接到多台 PE 时，为了避免冗余备份组中的 PE 均发送泛洪流量给该 CE，需要在冗余备份组中选举一个 PE 作为 DF（Designated Forwarder，指定转发者），负责将泛洪流量转发给本地站点。其他 PE 作为 BDF（Backup DF，备份 DF），不会向本地 CE 转发泛洪流量。

DF 的选举过程为：

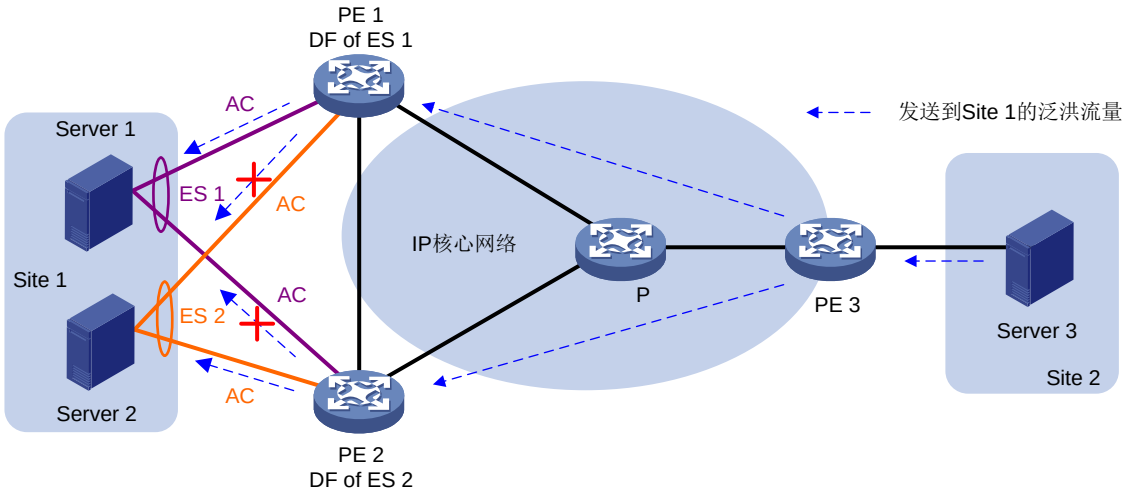
(1) 冗余备份组中的 PE 设备之间互相发送以太网段路由，通告 ES 的 ESI 值及其连接的 PE 信息（如 IP 地址、优先级等）。

(2) PE 接收到以太网段路由后，如果路由中携带的 ESI 值与本地相同，则 PE 记录发送该路由的 PE 信息，以便生成连接到同一 ES 的所有 PE 的列表。

(3) 冗余备份组中的 PE 设备根据以太网段路由中的 PE 信息选举出 DF。

设备支持多种 DF 选举算法，用户可以根据业务需要灵活地选择 DF 选举算法，使组网中 DF 能够均匀分布，提高网络设备的使用率。

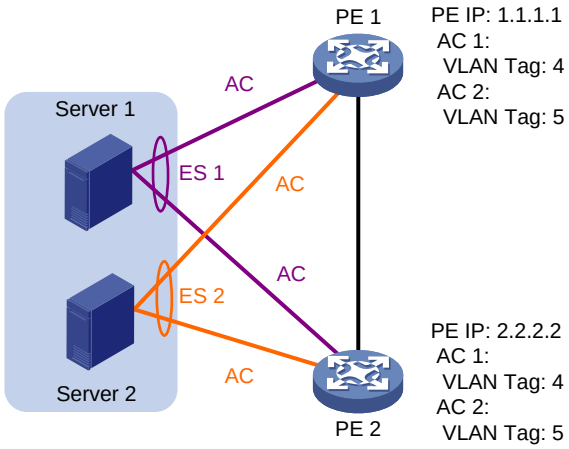
图47 DF 示意图



2. 基于 VLAN Tag 的 DF 选举算法

基于 VLAN Tag 的 DF 选举算法根据 VLAN Tag 和 VTEP 的 IP 地址为每个 AC 选举 DF。

图48 基于 VLAN Tag 的 DF 选举



如图 48 所示，以允许 VLAN Tag 4 通过的 AC 1 的 DF 选举为例，基于 VLAN Tag 的 DF 选举算法为：

(2) 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。

(3) 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。

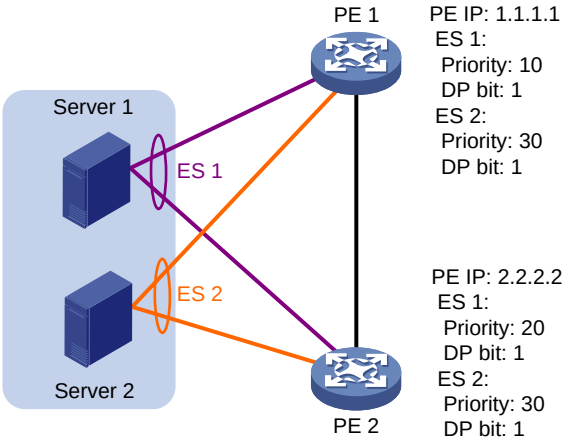
(4) 根据 VLAN Tag 除以 N 的余数 M 来选举 DF，N 代表冗余备份组中成员的数量，M 对应的编号为该 AC 的 DF。在本例中，4 除以 2 的余数为 0，即 AC 1 的 DF 为编号为 0 的 VTEP 1。

3. 基于优先级的 DF 选举算法

基于优先级的 DF 选举算法根据 DF 选举优先级、DP（Don't Preempt Me，不可回切）位和 VTEP 的 IP 地址为每个 ES 选举 DF。其中，DP 位的取值包括：

- 1：表示开启了基于优先级 DF 选举算法不回切功能。即当前设备被选举为 DF 后，即使后续选举出了新的设备作为 DF，依然使用当前设备作为 DF。
- 0：表示关闭了基于优先级 DF 选举算法不回切功能。即当前设备被选举为 DF 后，如果后续选举出了新的设备作为 DF，则直接使用新的设备作为 DF。

图49 基于优先级的 DF 选举



如图 49 所示，以 ES 1、ES 2 的 DF 选举为例，基于优先级的 DF 选举算法为：

(2) 同一 ES 内 DF 选举优先级（数值越大则优先级越高）最高的 VTEP 作为该 ES 的 DF。在本例中，选举 VTEP 2 作为 ES 1 的 DF。

(3) 若优先级相同，则 DP 位为 1 的 VTEP 作为 DF。

(4) 若 DP 位相同，则 IP 地址小的 VTEP 作为 DF。在本例中，选举 VTEP 1 作为 ES 2 的 DF。

3.4.3 冗余备份模式

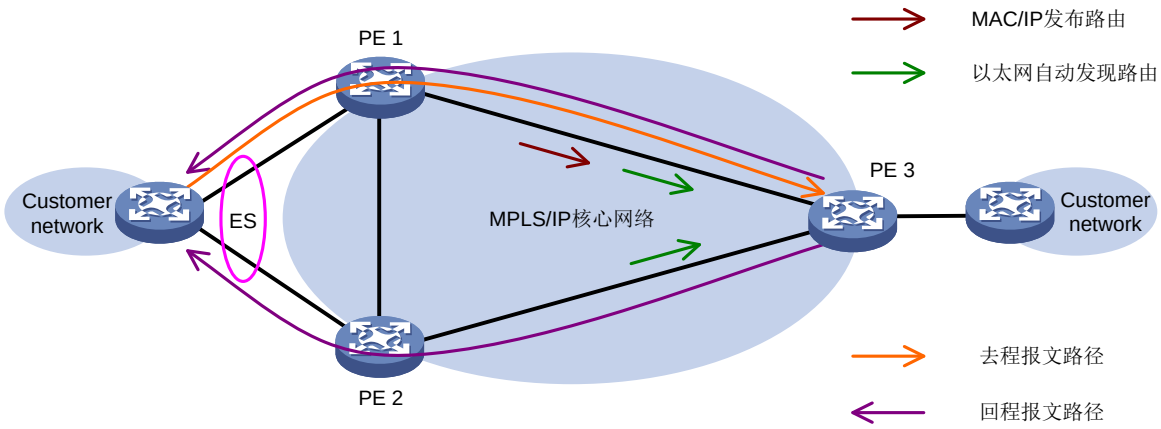
当前设备支持多活冗余模式，在该模式下：

- 出多归属站点方向流量：多归属站点可以通过冗余备份组中多台 PE 访问其它站点。
- 入多归属站点方向流量：其它站点的已知单播流量可以通过冗余备份组中多台 PE 访问多归属站点；其它站点的未知单播流量、广播流量和组播流量仅可以通过冗余备份组中作为 DF 的 PE 访问多归属站点。
- 负载分担：站点间可以通过冗余备份组中多台 PE 互相访问，CE 之间存在多条可达链路，可以形成负载分担。

3.4.4 协议报文交互过程

3.4.5 别名

图50 别名示意图

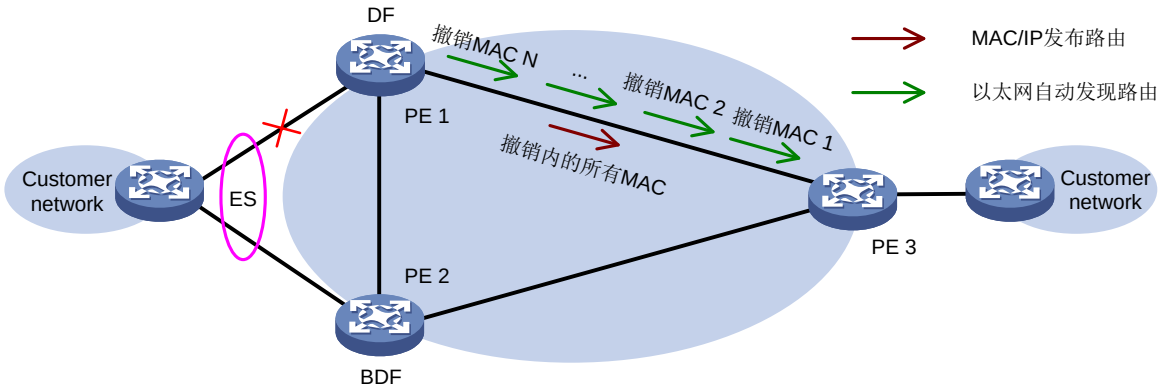


如图 50 所示，在多活冗余模式下，冗余备份组中可能仅有一台 PE 能学习到某些业务相关的 MAC 地址，这会导致远端 PE 仅能从这台 PE 收到这些 MAC 地址的 MAC/IP 发布路由，因此远端 PE 无法将访问这些 MAC 地址的流量负载分担到冗余备份组中的其它 PE 上。

为了解决这个问题，EVPN 多归属引入了别名机制，即当冗余备份组中仅有一台 PE 通过 MAC/IP 发布路由向远端 PE 通告了 CE 侧 MAC 地址的可达性时，远端 PE 能够根据冗余备份组内 PE 发送的以太网自动发现路由（携带 PE、ESI 等信息）感知到冗余备份组中其它 PE 与 MAC 地址的可达性，并生成对应的 MAC 表项，从而形成负载分担。

3.4.6 MAC 地址快速收敛

图51 MAC 地址快速收敛示意图



如图 51 所示，在 EVPN 网络中，MAC 地址可达性是通过 PE 之间发布 MAC/IP 发布路由通告的。因此，在 CE 1 与 PE 1 间链路故障时，PE 1 需要逐条撤销 MAC/IP 发布路由，在大规模的网络中会导致 MAC 地址收敛速度较慢。

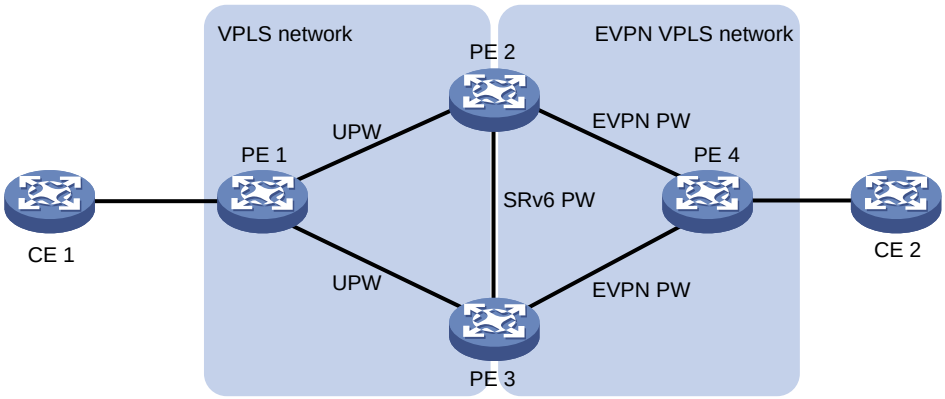
EVPN 多归属组网提供了快速收敛机制，使得 PE 可以通过撤销一条以太网自动发现路由，通告对指定 ES 内所有 MAC 地址的不可达性，通知远端 PE 批量删除 MAC 地址表项，减少收敛时间。

3.5 LDP PW或静态PW接入EVPN PW

在实际组网中，可能会存在传统的 VPLS 网络与 EVPN VPLS 网络共存的情况。LDP PW 或静态 PW 接入 EVPN PW 功能，通过将 VPLS 网络中的 LDP PW 或静态 PW 看作 EVPN VPLS 网络的 AC（该 PW 称为 UPW），实现报文在 EVPN PW 与 UPW 之间相互转发，从而实现 VPLS 网络与 EVPN VPLS 网络的互通。

本功能不仅支持一条 LDP PW 或静态 PW 接入一条 EVPN PW，还支持将两条 LDP PW 或静态 PW 多归属接入两条 EVPN PW。如图 52 所示，在 VPLS 网络中，PE 1 与 PE 2、PE 3 分别建立主备 LDP PW 或静态 PW，该 PW 称为 UPW；在 EVPN VPLS 网络中，PE 4 与 PE 2、PE 3 分别建立 EVPN PW。UPW 作为 EVPN VPLS 网络中的 AC，PE 2 或 PE 3 从 UPW 接收到报文后，会解除 MPLS 封装，查找 MAC 地址表获取到对应的 EVPN PW，为报文添加该 EVPN PW 对应的 MPLS 封装，并将其转发给 PE 4；PE 2 或 PE 3 从 EVPN PW 接收报文的处理方法与此类似。

图52 LDP PW 或静态 PW 接入 EVPN PW 组网示意图

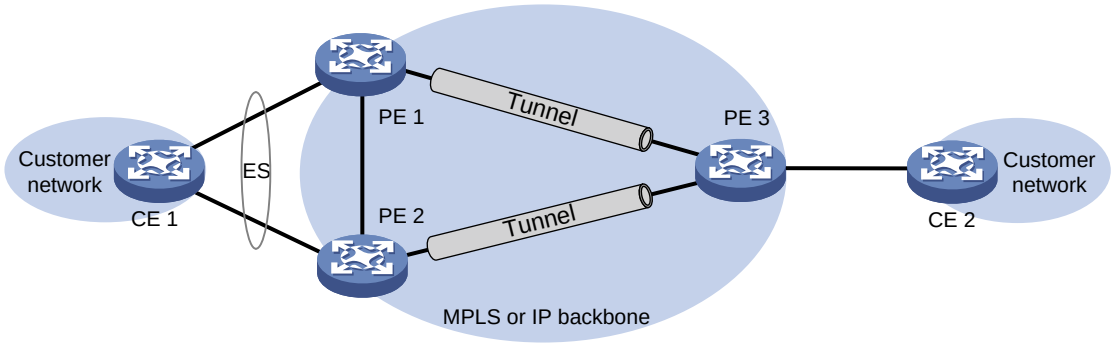


3.6 典型组网应用

3.6.1 多归属组网

为了避免 PE 单点故障造成报文转发中断，EVPN VPLS 通常采用多归属组网方式。多归属站点的流量在多台 PE 之间进行负载分担，即所有 PE 均转发流量，以提高网络的可靠性。

图53 多归属组网方式

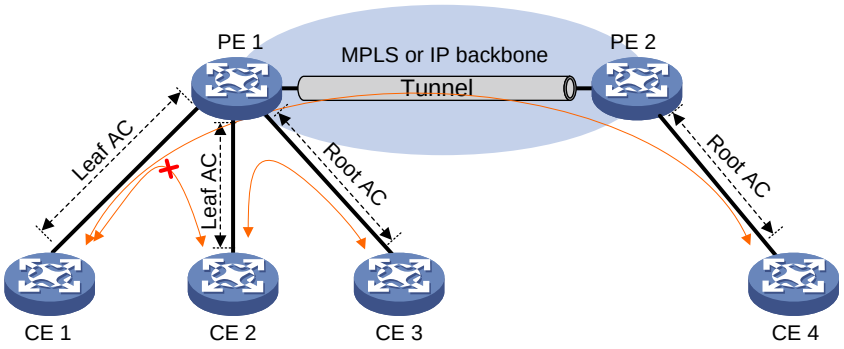


3.6.2 E-Tree 组网

在 EVPN VPLS 组网中，属于同一个 VSI 的所有 AC 均可以互相访问。在 EVPN VPLS 网络中，为了提高 AC 侧用户业务的安全性，减少用户业务之间的相互影响，网络管理员可能需要控制 AC 侧用户之间的相互访问。E-Tree 功能通过将 AC 分为 Root 和 Leaf 两种角色，实现了同一 VSI 内 AC 之间流量的隔离：

- Leaf AC 连接的用户只能和 Root AC 连接的用户相互访问。
- 不同 Leaf AC 连接的用户之间相互隔离。
- Root AC 连接的用户可以与 VSI 内所有 AC 连接的用户相互访问。

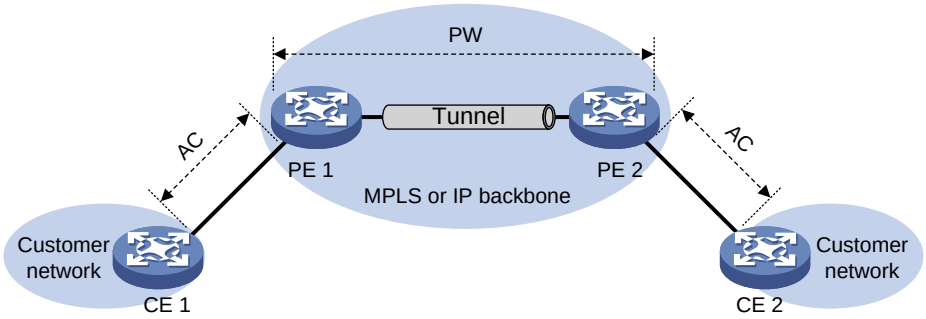
图54 EVPN E-Tree 示意图



4 EVPN VPWS

4.1 网络模型

图55 EVPN VPWS 网络模型示意图



如图 55 所示，EVPN VPWS 的典型网络模型中包括如下几部分：

- CE（Customer Edge，用户网络边缘）：直接与服务提供商网络相连的用户网络侧设备。
- PE（Provider Edge，服务提供商网络边缘）：与 CE 相连的服务提供商网络侧设备。PE 主要负责 EVPN 业务的接入，完成报文从用户网络到公网隧道、从公网隧道到用户网络的映射与转发。
- AC（Attachment Circuit，接入电路）：连接 CE 和 PE 的物理电路或虚拟电路，例如 Frame Relay 的 DLCI、ATM 的 VPI/VCI、Ethernet 接口、VLAN、物理接口上的 PPP 连接。
- PW（Pseudowire，伪线）：两个 PE 之间的虚拟双向连接。PW 由一对方向相反的单向虚拟连接构成。
- 公网隧道（Tunnel）：穿越 IP 或 MPLS 骨干网、用来承载 PW 的隧道。一条公网隧道可以承载多条 PW，公网隧道可以是 LSP、GRE 隧道或 MPLS TE 隧道。
- 交叉连接（Cross connect）：由两条物理电路或虚拟电路串连而成的一条连接，从一条物理、虚拟电路收到的报文直接交换到另一条物理、虚拟电路转发。交叉连接包括二种方式：AC 到 AC 交叉连接和 AC 到 PW 交叉连接。

4.2 EVPN VPWS控制平面工作机制

4.2.1 工作机制综述

EVPN VPWS 通过穿越 IP 或 MPLS 骨干网络的 PW 连接两端的用户网络，为用户提供点对点的二层服务。

EVPN VPWS 控制平面的工作机制为：

(1) 建立公网隧道，公网隧道用来承载 PE 之间的一条或多条 PW。

(2) 建立用来传送特定用户网络报文的 PW，PW 标签标识了报文所属的用户网络。

(3) 建立用来连接 CE 和 PE 的 AC，AC 的报文匹配规则（显式配置或隐含的规则）决定了从 CE 接收到的哪些报文属于一个特定的用户网络。

(4) 将 AC 和 PW 关联，以便 PE 确定从 AC 接收到的报文向指定 PW 转发，从 PW 接收到的报文向指定 AC 转发。

完成上述工作后，PE 从 AC 接收到用户网络的报文后，根据 AC 关联的 PW 为报文封装 PW 标签，并通过公网隧道将报文转发给远端 PE；远端 PE 从公网隧道接收到报文后，根据 PW 标签判断报文所属的 PW，并将还原后的原始报文转发给与该 PW 关联的 AC。

4.2.2 建立公网隧道

公网隧道用来承载 PW，可以是 LSP 隧道、MPLS TE 隧道和 GRE 隧道等。不同隧道的建立方式不同，详细介绍请参见相关手册。

当两个 PE 之间存在多条公网隧道时，可以通过配置隧道策略，确定如何选择隧道。

说明

如果 PW 建立在 LSP 或 MPLS TE 隧道之上，则 PW 上传送的报文将包括两层标签：内层标签为 PW 标签，用来决定报文所属的 PW，从而将报文转发给正确的 CE；外层标签为公网 LSP 或 MPLS TE 隧道标签，用来保证报文在 MPLS 网络正确传送。

4.2.3 建立 PW

如图 56 所示，PW 的建立过程为：

(1) 在 PE 1、PE 2 上均配置 Local service ID 来标识与其连接的 CE，配置 Remote service ID 来标识远端 PE 连接的 CE，并为每个 Local service ID 分配 MPLS 标签（即 PW 标签），该标签作为 PW 的入标签。

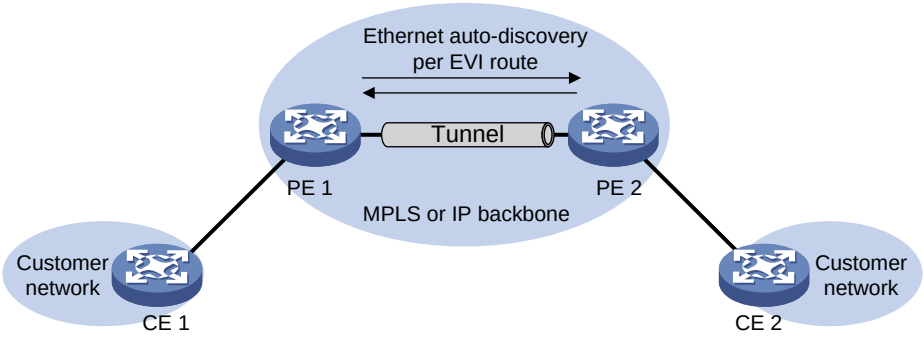
(2) 本地 PE（如 PE 1）通过 Ethernet Auto-discovery Per EVI 路由将 Local service ID 和为其分配的 PW 标签通告给远端 PE（如 PE 2）。

(3) 如果路由中的 Export target 属性与 PE 2 本地配置的 Import target 属性相同，则 PE 2 将接收到的 Local service ID 与本地配置的 Remote service ID 匹配。若二者相同，则建立一条从 PE 2 到 PE 1 的单向 LSP，PE 1 通告的 PW 标签作为该 LSP 的出标签。

(4) 同时，PE 2 也会向 PE 1 发送 Ethernet Auto-discovery Per EVI 路由。PE 1 将接收到的 Local service ID 与本地配置的 Remote service ID 匹配。若二者相同，则建立一条从 PE 1 到 PE 2 的单向 LSP。

(5) 当两端 PE 间建立了两条方向相反的单向 LSP 时，EVPN PW 建立完成。

图56 建立 PW 示意图



4.2.4 建立 AC

在 EVPN VPWS 中，AC 是与交叉连接关联的三层以太网接口、三层以太网子接口或以太网服务实例。以太网服务实例在二层以太网接口上创建，它定义了一系列匹配规则，用来匹配从该二层以太网接口上接收到的数据帧。

4.2.5 关联 AC 和 PW

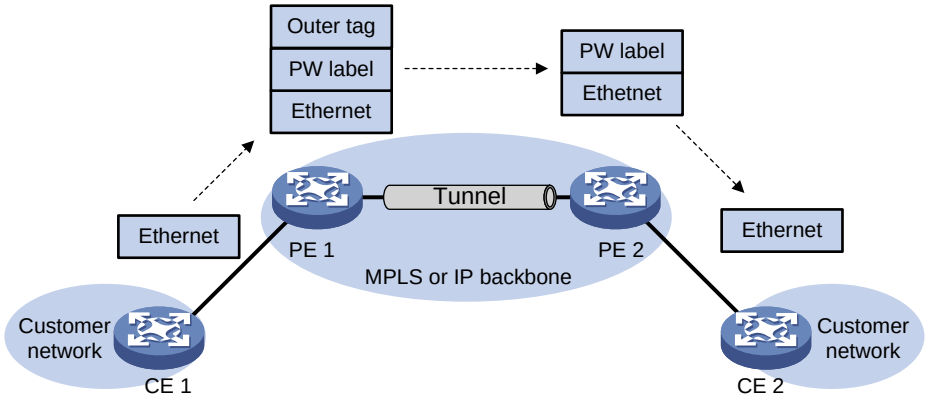
通过命令行将 AC 连接对应的三层以太网接口、三层以太网子接口或以太网服务实例与 PW 关联，即可实现从该 AC 接收到的报文通过关联的 PW 转发，从关联的 PW 上接收到的报文通过该 AC 转发。

4.3 EVPN VPWS数据平面工作机制

如图 57 所示，PE 从 AC/PW 接收到报文后，会在对应的交叉连接内查找出方向 PW 或 AC 信息，从而确定如何转发报文：

- 出接口为 PW 索引时，为报文封装 PW 标签，再添加公网隧道封装后，通过 PW 将该报文转发给远端 PE。如果公网隧道为 LSP 或 MPLS TE 隧道，则通过 PW 转发报文时将为报文封装两层标签。内层标签为 PW 标签，用来将报文转发给相应的 PW；外层标签为公网 LSP 或 MPLS TE 隧道标签，用来保证报文在 PE 之间正确传送。
- 出接口为连接本地 CE 的接口时，直接通过出接口将报文转发给本地 CE。

图57 EVPN VPWS 报文转发过程

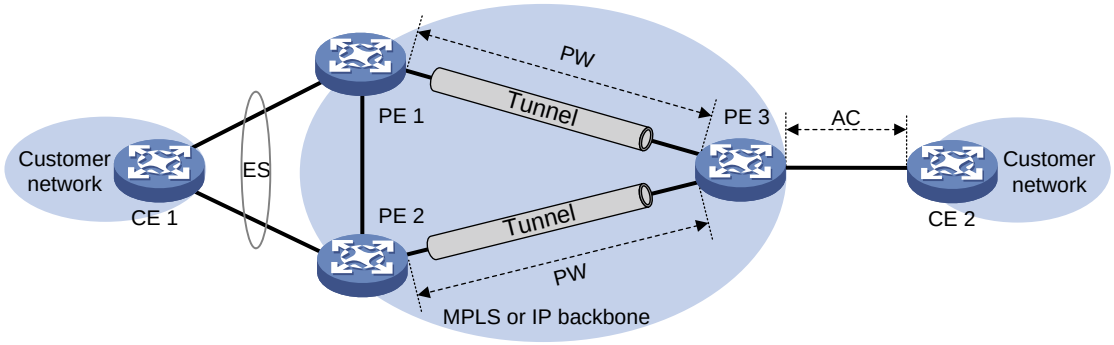


4.4 EVPN VPWS多归属

4.4.1 功能简介

当一个站点通过不同的以太网链路连接到多台 PE 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 PE 组成冗余备份组，可以避免 PE 单点故障对网络造成影响，从而提高网络的可靠性。目前仅支持双归属。

图58 多归属站点示意图



4.4.2 冗余备份模式

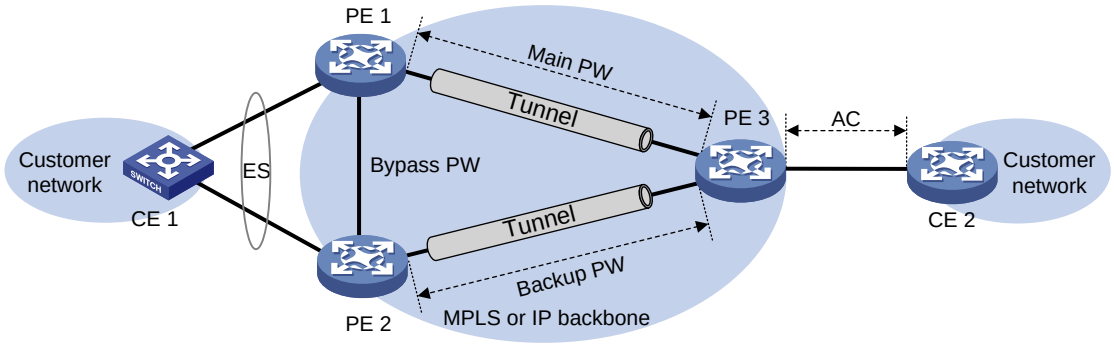
EVPN VPWS 组网场景支持的冗余备份模式包括：单活冗余模式和多活冗余模式。

- 单活冗余模式

  如图 59 所示，单活冗余模式下，PE 1 和 PE 2 中仅其中一台转发流量，PE 1 和 PE 2 上的两条 PW 为主备关系，实现当主 PW 出现故障后，将流量立即切换到备份 PW，使流量转发得以继续。通过 DF 选举可以确定主备 PW，DF 选举的详细介绍，请参见“4.4.3 DF 选举”。当 PE 1 的 PW 不可用（可能是 PE 1 节点故障，也可能是 PW 故障）时，PE 3 将启用备份 PW，通过备份 PW 将 CE 2 的报文转发给 PE 2，再由 PE 2 转发给 CE 1；同时建议在 PE 1

设备 AC 侧的物理接口与 PW 侧的物理接口（用于建立 EVPN PW 的接口）配置 EAA 和 Track 联动的 CLI 监控策略，使这两个接口联动，可以确保 PW 侧的 Underlay 网络断开时，将 AC 侧的接口置于 Down 状态，使 CE 1 到 CE 2 的流量通过 PE 2 转发。

图59 单活冗余模式示意图



- 多活冗余模式

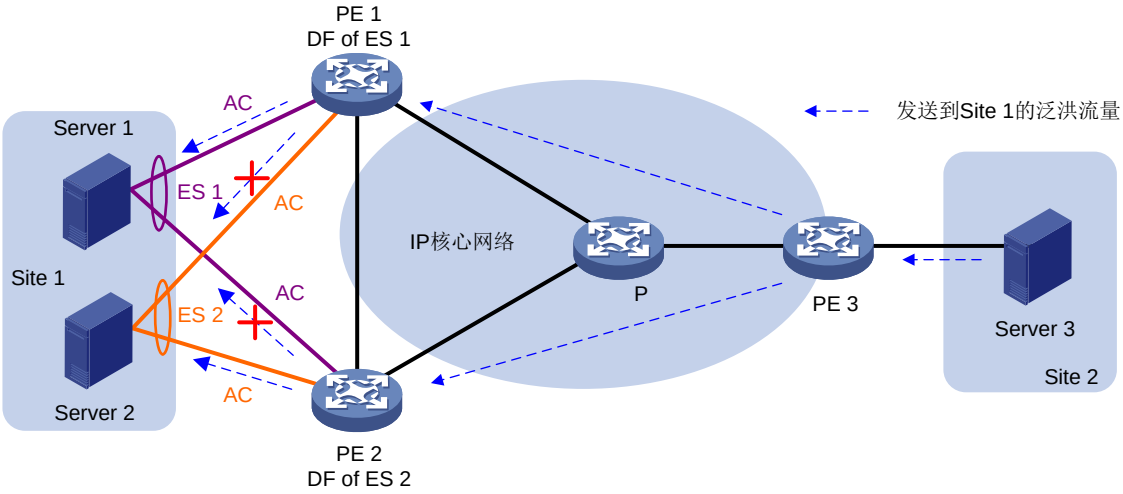
  多活冗余模式下，两条 PW 等价负载分担转发数据报文。该模式下也需要在 PE 设备 AC 侧的物理接口与 PW 侧的物理接口（用于建立 EVPN PW 的接口）配置 EAA 和 Track 联动的 CLI 监控策略，使这两个接口联动，提高网络可靠性。

4.4.3 DF 选举

在单活冗余模式下，数据报文仅通过一条 PW 转发，此时需要在冗余备份组中选举一个 PE 作为 DF（Designated Forwarder，指定转发者），该 PE 上创建的 PW 为主 PW。其他 PE 作为 BDF（Backup DF，备份 DF），其上创建的 PW 为备份 PW。多归属成员通过发送以太网段路由，向其它 PE 通告 ES 及 PE 信息，仅配置了 ESI 的 PE 会接收以太网段路由并根据其携带的 ES 和 PE 信息选举出 DF。

设备支持多种 DF 选举算法，用户可以根据业务需要灵活地选择 DF 选举算法，使组网中 DF 能够均匀分布，提高网络设备的使用率。

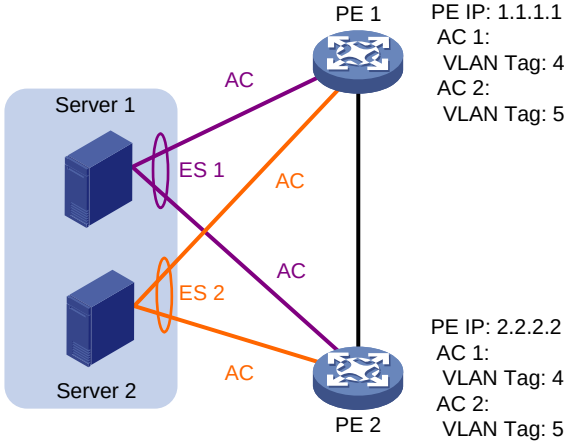
图60 DF 示意图



2. 基于 VLAN Tag 的 DF 选举算法

基于 VLAN Tag 的 DF 选举算法根据 VLAN Tag 和 VTEP 的 IP 地址为每个 AC 选举 DF。

图61 基于 VLAN Tag 的 DF 选举



如图 61 所示，以允许 VLAN Tag 4 通过的 AC 1 的 DF 选举为例，基于 VLAN Tag 的 DF 选举算法为：

(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。

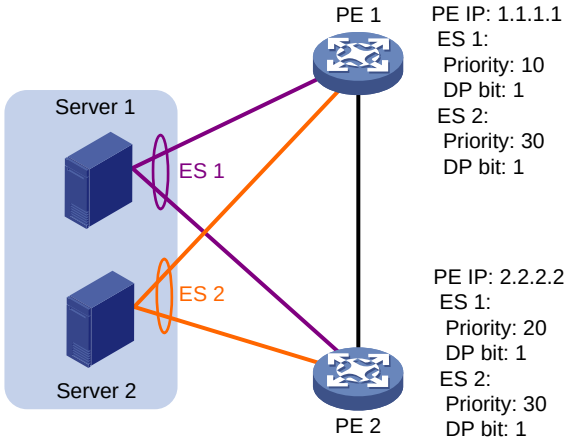
(3) 根据 VLAN Tag 除以 N 的余数 M 来选举 DF，N 代表冗余备份组中成员的数量，M 对应的编号为该 AC 的 DF。在本例中，4 除以 2 的余数为 0，即 AC 1 的 DF 为编号为 0 的 VTEP 1。

3. 基于优先级的 DF 选举算法

基于优先级的 DF 选举算法根据 DF 选举优先级、DP（Don't Preempt Me，不可回切）位和 VTEP 的 IP 地址为每个 ES 选举 DF。其中，DP 位的取值包括：

- 1：表示开启了基于优先级 DF 选举算法不回切功能。即当前设备被选举为 DF 后，即使后续选举出了新的设备作为 DF，依然使用当前设备作为 DF。
- 0：表示关闭了基于优先级 DF 选举算法不回切功能。即当前设备被选举为 DF 后，如果后续选举出了新的设备作为 DF，则直接使用新的设备作为 DF。

图62 基于优先级的 DF 选举



如图 62 所示，以 ES 1、ES 2 的 DF 选举为例，基于优先级的 DF 选举算法为：

(1) 同一 ES 内 DF 选举优先级（数值越大则优先级越高）最高的 VTEP 作为该 ES 的 DF。在本例中，选举 VTEP 2 作为 ES 1 的 DF。

(2) 若优先级相同，则 DP 位为 1 的 VTEP 作为 DF。

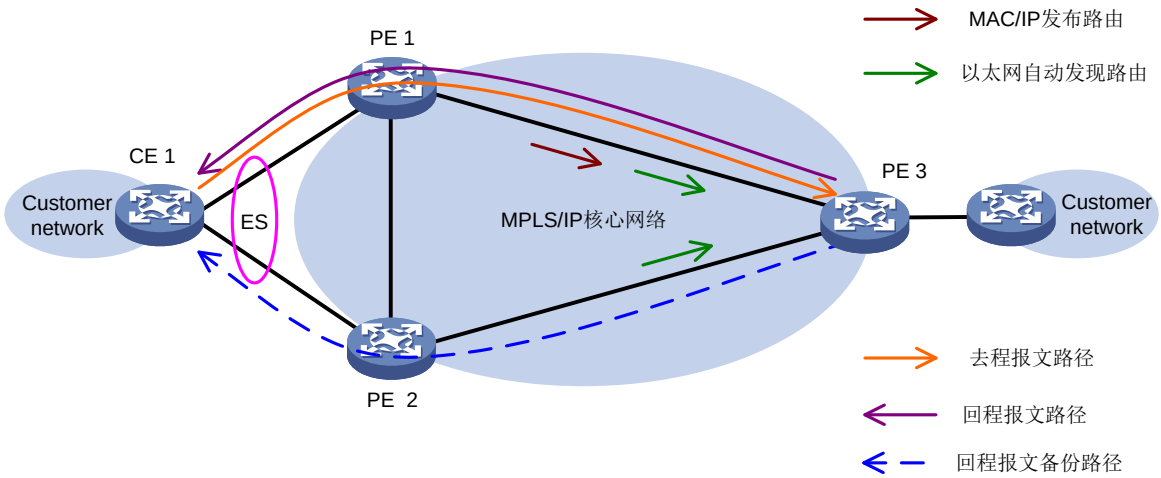
(3) 若 DP 位相同，则 IP 地址小的 VTEP 作为 DF。在本例中，选举 VTEP 1 作为 ES 2 的 DF。

4.4.4 协议报文交互过程

4.4.5 别名与备份路径

1. 单活冗余模式下的备份路径机制

图63 备份路径示意图

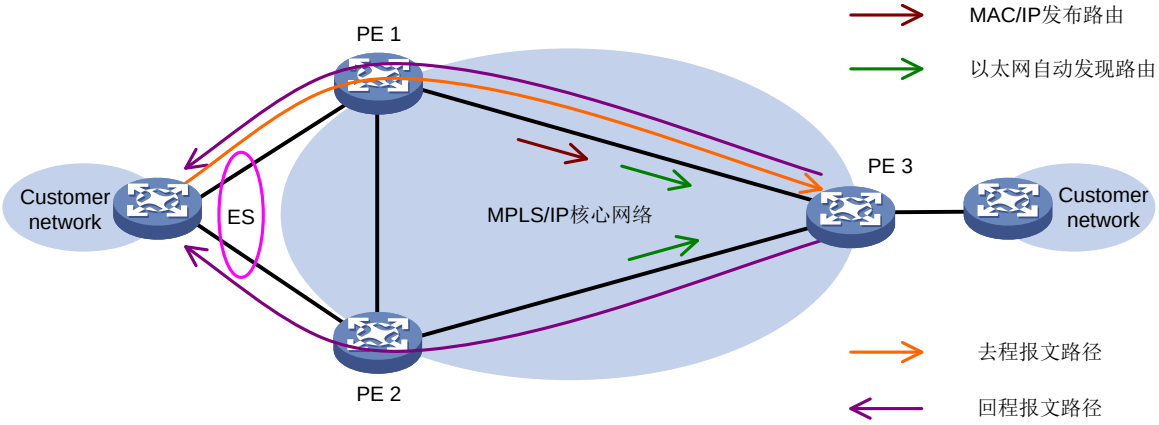


如图 63 所示，在单活冗余模式下，仅 DF 可以学习到多归属站点内的 MAC 地址，这会导致 PE 3 仅能从 DF 收到这些 MAC 地址的 MAC/IP 发布路由，若 DF 故障，PE 3 需要较长时间重新学习这些 MAC 地址的表项指导报文转发。

EVPN 多归属引入备份路径机制解决上述问题。备份路径机制是指冗余备份组中作为 DF 的 PE 通过 MAC/IP 发布路由向远端 PE 通告了 CE 侧 MAC 地址的可达性时，远端 PE 能够根据冗余备份组内 PE 发送的以太网自动发现路由（携带 PE、ESI 等信息）感知到冗余备份组中 BDF 与 MAC 地址的可达性，从而形成远端 PE 经过 BDF 到达与 CE 1 的备份路径。当 DF 故障时，冗余备份组直接将转发路径切换到通过 BDF 的备份路径上，而不需重新学习 MAC 表项。

2. 多活冗余模式下的别名机制

图64 别名示意图



如图 64 所示，在多活冗余模式下，冗余备份组中可能仅有一台 PE 能学习到某些业务相关的 MAC 地址，这会导致远端 PE 仅能从这台 PE 收到这些 MAC 地址的 MAC/IP 发布路由，因此远端 PE 无法将访问这些 MAC 地址的流量负载分担到冗余备份组中的其它 PE 上。

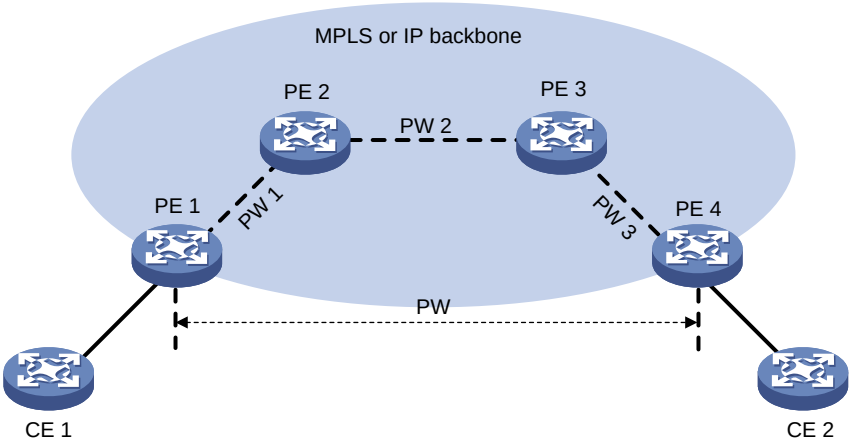
为了解决这个问题，EVPN 多归属引入了别名机制，即当冗余备份组中仅有一台 PE 通过 MAC/IP 发布路由向远端 PE 通告了 CE 侧 MAC 地址的可达性时，远端 PE 能够根据冗余备份组内 PE 发送的以太网自动发现路由（携带 PE、ESI 等信息）感知到冗余备份组中其它 PE 与 MAC 地址的可达性，并生成对应的 MAC 表项，从而形成负载分担。

4.5 多段PW

多段 PW 是指将两条或多条 PW 串连（concatenated）起来，形成一条端到端的 PW。通过在同一个交叉连接下创建两条 PW，可以实现将该交叉连接下的两条 PW 串连。PE 从一条 PW 接收到报文后，剥离报文的隧道标识和 PW 标签，封装上与该 PW 串连的另一条 PW 的 PW 标签，并通过承载该 PW 的公网隧道转发该报文，从而实现报文在两条 PW 之间的转发。

如图 65 所示，通过在 PE 2 上将 PW 1 和 PW 2 串连、在 PE 3 上将 PW 2 和 PW 3 串连，可以建立从 PE 1 到 PE 4 的端到端 PW，实现报文沿着 PW 1、PW 2 和 PW 3 形成的多段 PW 在 PE 1 和 PE 4 之间转发。

图65 多段 PW 示意图



多段 PW 分为：

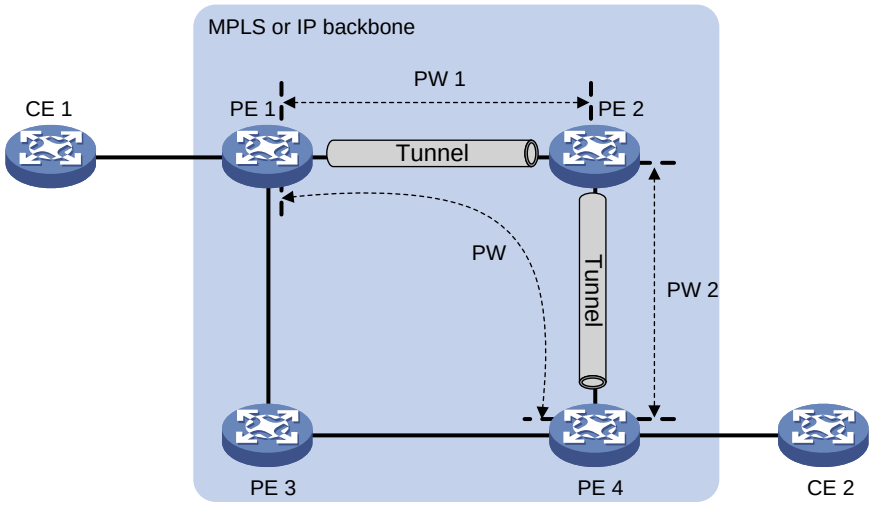
- 域内多段 PW：即在一个自治系统内部署多段 PW。

  在一个自治系统内部署多段 PW，可以实现两个 PE 之间不存在端到端公网隧道的情况下，在这两个 PE 之间建立端到端 PW。

  如图 66 所示，PE 1 和 PE 4 之间没有建立公网隧道，PE 1 和 PE 2、PE 2 和 PE 4 之间已经建立了公网隧道。通过在 PE 1 与 PE 2、PE 2 与 PE 4 之间分别建立一条 PW（PW 1 和 PW 2），在 PE 2 上将这两条 PW 串连，可以实现在 PE 1 和 PE 4 之间建立一条由两段 PW 组成的端到端域内多段 PW。

  通过建立域内多段 PW 可以充分利用已有的公网隧道，减少端到端公网隧道数量。

  图66 域内多段 PW



- 域间多段 PW：即跨越自治系统部署多段 PW。关于域间多段 PW 的详细介绍，请参见“4.6.2 跨域-Option B”。

4.6 EVPN VPWS跨域

实际组网应用中，不同 Site 间可能会通过使用不同 AS 号的多个服务提供商通信，或者跨越一个服务提供商的多个 AS 通信。这种跨越多个自治系统的应用方式被称为 EVPN VPWS 跨域。

EVPN VPWS 跨域解决方案分为以下几种：

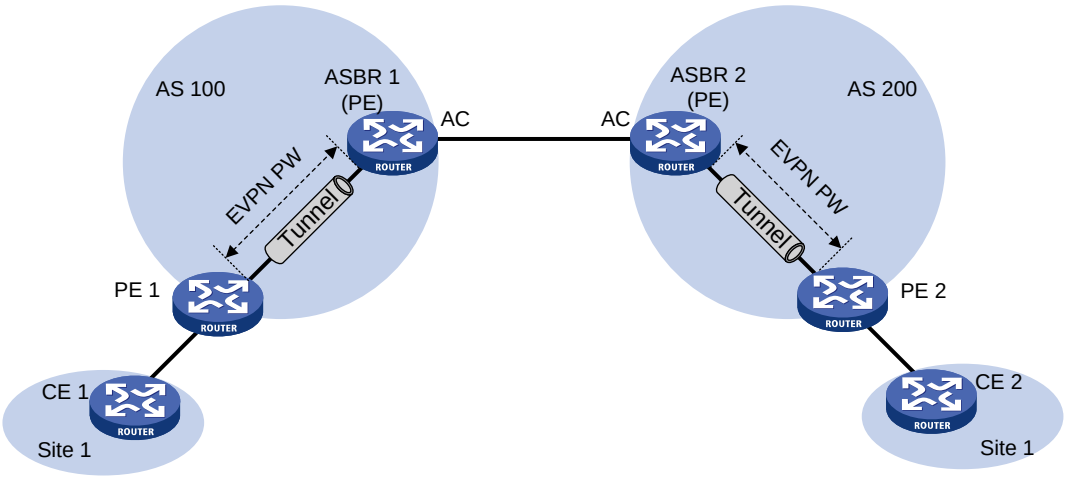
- PE 与 ASBR 间通过 MP-IBGP（IBGP redistribution of EVPN routes between PE and ASBR）发布 EVPN 路由建立 EVPN PW，ASBR 互为 CE，在 ASBR 上将 AC 与 EVPN PW 关联，也称为 Inter-Provider Option A。
- PE 与 ASBR 间通过 MP-IBGP 发布 EVPN 路由建立 EVPN PW，ASBR 间通过 MP-EBGP（EBGP redistribution of EVPN routes between ASBRs）发布 EVPN 路由建立 EVPN PW，也称为 Inter-Provider Option B；
- PE 间通过 MP-EBGP（Multi-hop EBGP redistribution of EVPN routes between PE routers）发布 EVPN 路由建立 EVPN PW，也称为 Inter-Provider Option C。

4.6.1 跨域-Option A

如图 67 所示，这种方式下，两个 AS 的 PE 路由器直接相连，并且作为各自所在自治系统的边界路由器 ASBR。两个 ASBR 均把对方当作自己的 CE 设备，并将与对端 ASBR 相连的接口与 EVPN PW 关联实现报文的跨域转发。

这种方式的优点是实现简单，两个作为 ASBR 的 PE 之间不需要为跨域进行特殊配置。缺点是可扩展性差：需要在两端的 ASBR 上为每个跨域站点配置 AC 并与 EVPN PW 绑定，配置复杂且管理难度大。

图67 ASBR 互为 CE 连接组网图

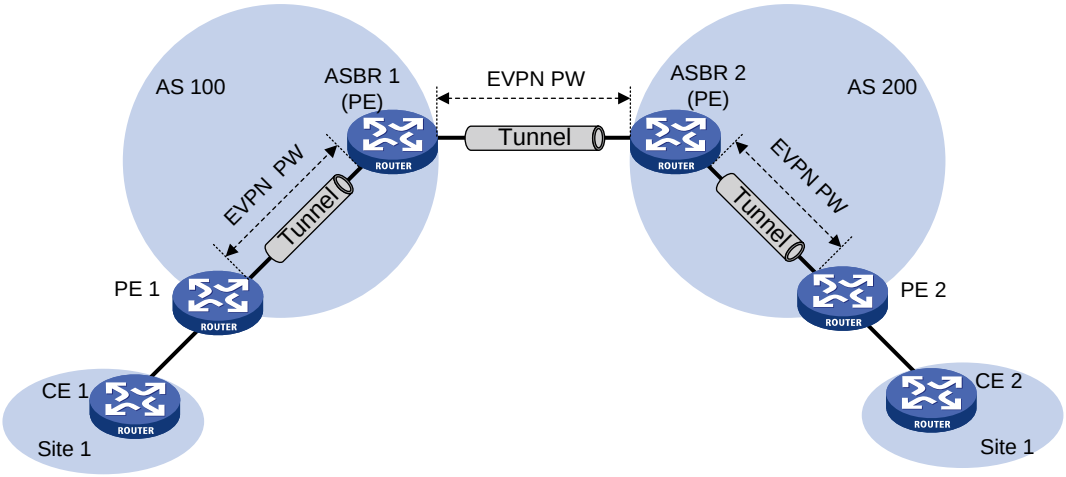


4.6.2 跨域-Option B

如图 68 所示，这种方式下，在 PE 1 与 ASBR 1、ASBR 2 与 PE 2 之间分别通过 MP-IBGP 发布 EVPN 路由建立 EVPN PW，ASBR 1 与 ASBR 2 间通过 MP-EBGP 发布 EVPN 路由建立 EVPN PW。通过多条 EVPN PW 的串联，即可实现报文的跨域传送。

这种方式的扩展性优于 Inter-Provider Option A。缺点是 ASBR 仍然需要为每个跨域站点配置多段 PW。

图68 多段 PW 跨域组网图

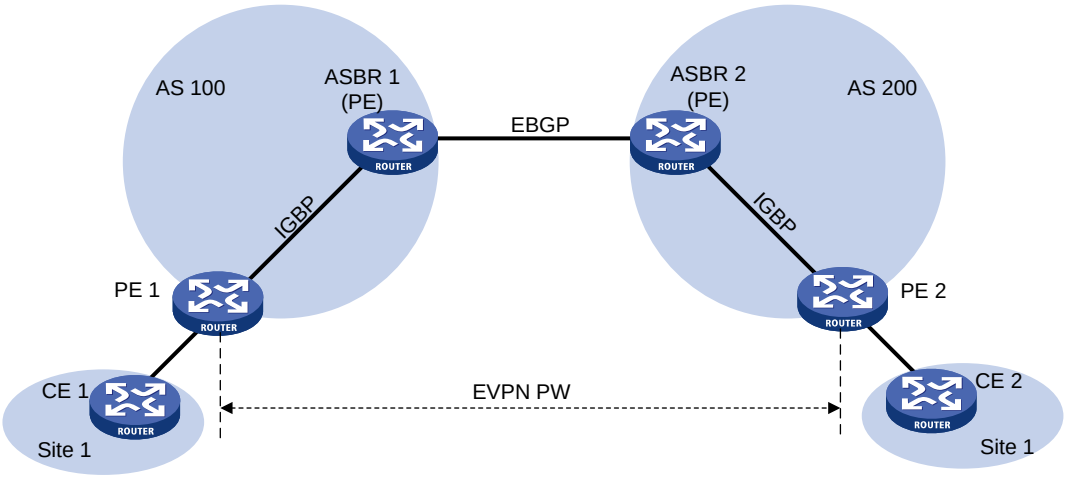


4.6.3 跨域-Option C

这种方式下，不同 AS 的 PE 之间建立多跳 MP-EBGP 会话，通过该会话直接在 PE 之间发布 EVPN 路由创建 EVPN PW。此时，一端 PE 上需要具有到达远端 PE 的路由以及该路由对应的标签，以便在两个 PE 之间建立跨越 AS 的公网隧道。Inter-Provider Option C 通过如下方式建立公网隧道：

- 利用 LDP 等标签分发协议在 AS 内建立公网隧道；
- ASBR 通过 BGP 发布带标签的 IPv4 单播路由，实现跨越 AS 域建立公网隧道。带标签的 IPv4 单播路由是指为 IPv4 单播路由分配 MPLS 标签，并同时发布 IPv4 单播路由和标签，以便将路由和标签关联。

图69 PE 间通过 Multi-hop MP-EBGP 发布 EVPN 路由组网图



如图 69所示，Inter-Provider Option C 的难点是建立跨越 AS 域的公网隧道。以 PE 2 到 PE 1 为例，公网隧道建立过程为：

(1) 在 AS 100 内，通过 LDP 等标签分发协议建立从 ASBR 1 到 PE 1 的公网隧道。假设 ASBR 1 上该公网隧道的出标签为 L1。

(2) ASBR 1 通过 EBGP 会话向 ASBR 2 发布带标签的 IPv4 单播路由，将 PE 1 地址对应的路由及 ASBR 1 为其分配的标签（假设为 L2）发布给 ASBR 2，路由的下一跳地址为 ASBR 1。这样，就建立了从 ASBR 2 到 ASBR 1 的公网隧道，ASBR 1 上公网隧道的入标签为 L2。

(3) ASBR 2 通过 IBGP 会话向 PE 2 发布带标签的 IPv4 单播路由，将 PE 1 地址对应的路由及 ASBR 2 为其分配的标签（假设为 L3）发布给 PE 2，路由的下一跳地址为 ASBR 2。这样，就建立了从 PE 2 直接到 ASBR 2 的公网隧道，ASBR 2 上公网隧道的入标签为 L3，出标签为 L2。

(4) MPLS 报文不能直接从 PE 2 转发给 ASBR 2，在 AS 200 内，还需要通过 LDP 等标签分发协议逐跳建立另一条从 PE 2 到 ASBR 2 的公网隧道。假设 PE 2 上该公网隧道的出标签为 Lv。

公网隧道建立后，PE 1 和 PE 2 间通过多跳 MP-EBGP 会话发布 EVPN 路由建立 EVPN PW，在 PE 1 和 PE 2 上将 EVPN PW 与 AC 关联，即可实现报文的跨域转发。

为减少 IBGP 连接数，可以在每个 AS 中指定一个 RR（Route Reflector，路由反射器），与同一 AS 的 PE 交换 EVPN 路由信息，由 RR 保存所有 EVPN 路由。两个 AS 的 RR 之间建立多跳 MP-EBGP 会话，通告 EVPN 路由。

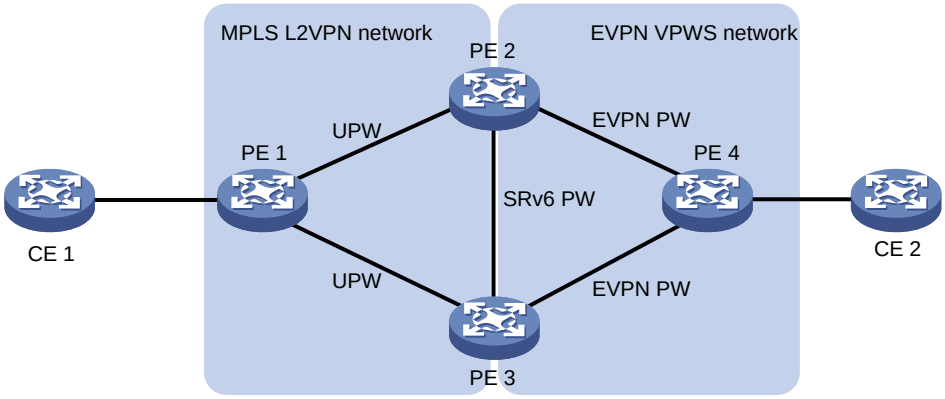
Inter-Provider Option A 和 Inter-Provider Option B 都需要 ASBR 参与 EVPN 路由的维护和发布。当每个 AS 都有大量的 EVPN 路由需要交换时，ASBR 很可能成为阻碍网络进一步扩展的瓶颈。Inter-Provider Option C 中 PE 之间直接交换 EVPN 路由，具有很好的可扩展性。

4.7 LDP PW或静态PW接入EVPN PW

在实际组网中，可能会存在传统的 MPLS L2VPN 网络（也称为 VPWS 网络）与 EVPN VPWS 网络共存的情况。LDP PW 或静态 PW 接入 EVPN PW 功能，通过将 MPLS L2VPN 网络中的 LDP PW 或静态 PW 看作 EVPN VPWS 网络的 AC（该 PW 称为 UPW），实现报文在 EVPN PW 与 UPW 之间相互转发，从而实现 MPLS L2VPN 网络与 EVPN VPWS 网络的互通。

本功能不仅支持一条 LDP PW 或静态 PW 接入一条 EVPN PW，还支持将两条 LDP PW 或静态 PW 多归属接入两条 EVPN PW。如 3.5 图 52 所示，在 MPLS L2VPN 网络中，PE 1 与 PE 2、PE 3 分别建立主备 LDP PW 或静态 PW，该 PW 称为 UPW；在 EVPN VPWS 网络中，PE 4 与 PE 2、PE 3 分别建立 EVPN PW。UPW 作为 EVPN VPWS 网络中的 AC，PE 2 或 PE 3 从 UPW 接收到报文后，会解除 MPLS 封装，查找与 UPW 关联的 EVPN PW，为报文添加该 EVPN PW 对应的 MPLS 封装，并将其转发给 PE 4；PE 2 或 PE 3 从 EVPN PW 接收报文的处理方法与此类似。

图70 LDP PW 或静态 PW 接入 EVPN PW 组网示意图

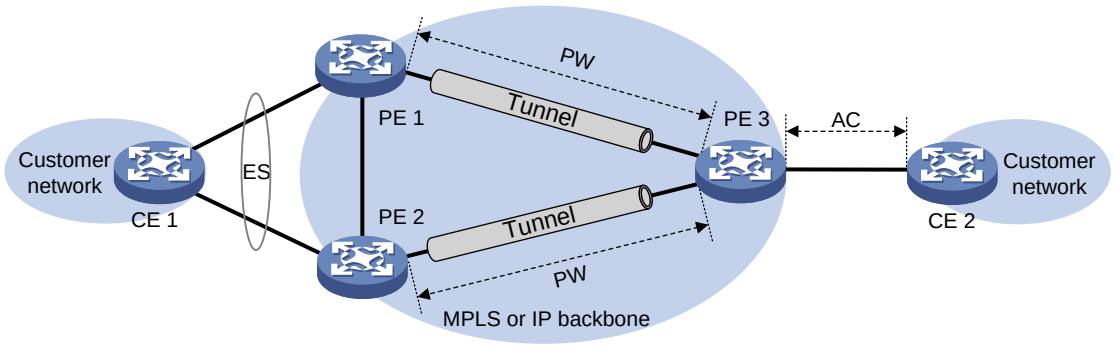


4.8 典型组网应用

4.8.1 多归属组网

为了避免 PE 单点故障造成报文转发中断，EVPN VPWS 通常采用多归属组网方式。多归属站点的流量可以在多台 PE 之间形成主备备份，即同一时间仅有其中一台 PE 转发流量；也可以在多台 PE 之间进行负载分担，即所有 PE 均转发流量。

图71 多归属组网方式

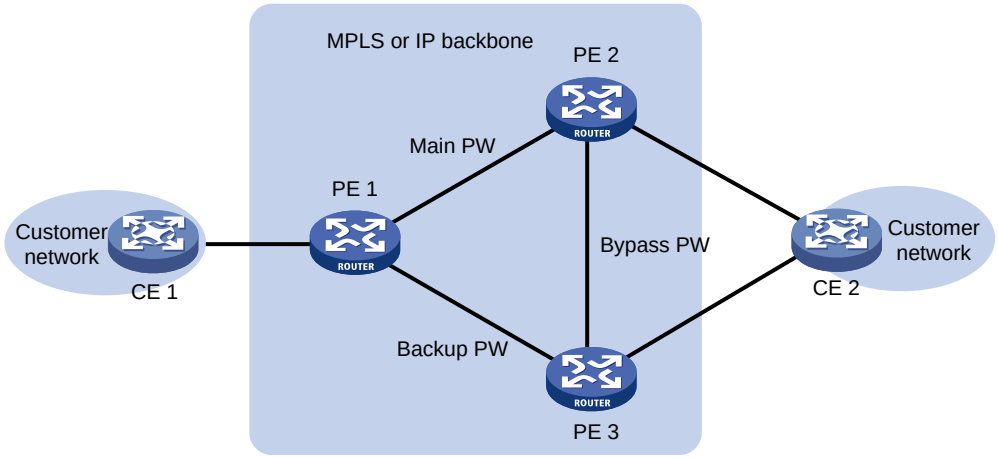


4.8.2 FRR 组网

为了减少 AC 链路或 PW 链路故障对网络造成的影响，提升网络的可靠性和稳定性，可以在 EVPN VPWS 组网中部署 FRR（Fast Reroute，快速重路由）功能。FRR 功能包含如下两种类型：

- Bypass PW：即旁路 PW。该功能可以减少 AC 链路故障导致的丢包。如 PE 2 侧 AC 链路故障时，PE 2 通过 Bypass PW 临时将流量到转发到 PE 3，再由 PE 3 转发到 CE 2。
- 主备 PW：即 PE 间建立的两条互为备份的 EVPN PW，其中一条为主 PW，一条为备份 PW。如 PE 1 与 PE 2 间的主 PW 故障时，PE 1 将流量切换到备 PW 上转发给 PE 3，再由 PE 3 转发到 CE 2。

图72 FRR 组网示意图

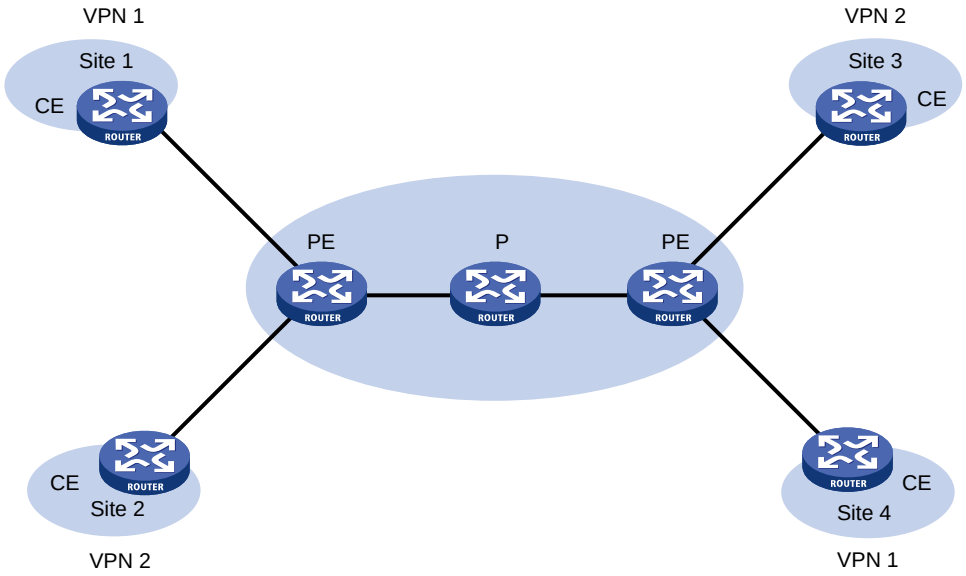


5 EVPN L3VPN

5.1 EVPN L3VPN网络模型

EVPN 的 IP 前缀路由可以用来发布 VPN 私网路由信息，以实现 MPLS L3VPN 组网，该网络称为 EVPN L3VPN。与 BGP/MPLS L3VPN 网络相比，EVPN L3VPN 组网中，在 EVPN 的基础上可以快速部署大二层网络，使得网络同时承载二层 VPN 和三层 VPN 业务。

图73 EVPN L3VPN 典型组网图



如图 73 所示，EVPN L3VPN 网络中主要包括如下几部分：

- CE（Customer Edge，用户网络边缘）：直接与服务提供商网络相连的用户网络侧设备。
- PE（Provider Edge，服务提供商网络边缘）：与 CE 相连的服务提供商网络侧设备。PE 主要负责 EVPN L3VPN 业务的接入，完成报文从用户网络到公网隧道、从公网隧道到用户网络的映射与转发。

5.2 EVPN L3VPN控制平面工作机制

在 EVPN L3VPN 组网中，VPN 路由信息的发布涉及 CE 和 PE。P 路由器只维护骨干网的路由，不需要了解任何 VPN 路由信息。PE 路由器只维护与它直接相连的 VPN 的路由信息，不维护所有 VPN 路由。

VPN 路由信息的发布过程包括三部分：本地 CE 到入口 PE、入口 PE 到出口 PE、出口 PE 到远端 CE。完成这三部分后，本地 CE 与远端 CE 之间将建立可达路由。

5.2.1 本地 CE 到入口 PE 的路由信息交换

CE 使用静态路由、RIP、OSPF、IS-IS、EBGP 或 IBGP，将本站点的 VPN 路由发布给 PE。CE 发布给 PE 的是标准的 IPv4 或 IPv6 路由。

5.2.2 入口 PE 到出口 PE 的路由信息交换

PE 从 CE 学到 VPN 路由信息后，将其存放到相应的 VPN 实例的路由表中。PE 为这些标准 IPv4 或 IPv6 路由增加 RD 和 Export Target 属性，并为这些路由分配 MPLS 私网标签，形成 EVPN 的 IP 前缀路由（包括 RD、Export Target 属性和 MPLS 私网标签）发布给出口 PE。出口 PE 将 IP 前缀路由的 Export Target 属性与自己维护的 VPN 实例的 Import Target 属性进行匹配。如果出口 PE 上某个 VPN 实例的 Import Target 属性与路由的 Export Target 属性中存在相同的属性值，则接收该 IP 前缀路由并将其添加到 VPN 路由表中。

5.2.3 出口 PE 到远端 CE 的路由信息交换

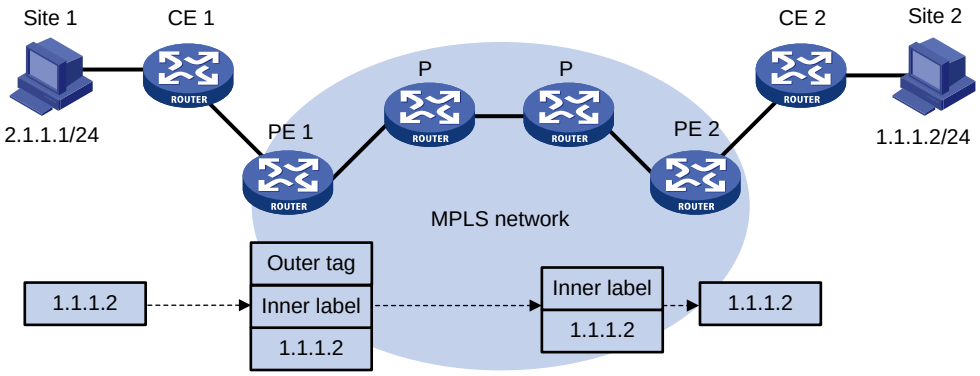
与本地 CE 到入口 PE 的路由信息交换相同，远端 CE 可以通过多种方式从出口 PE 学习 VPN 路由，包括静态路由、RIP、OSPF、IS-IS、EBGP 和 IBGP。

5.3 EVPN L3VAN数据平面工作机制

在 EVPN L3VPN 组网中，PE 转发 VPN 报文时为报文封装如下内容：

- 外层标记：又称为公网标记。VPN 报文在骨干网上沿着公网隧道从一端 PE 传送到另一端 PE。公网隧道可以是 LSP 隧道、MPLS TE 隧道和 GRE 隧道。当公网隧道为 LSP 隧道或 MPLS TE 隧道时，公网标记为 MPLS 标签，称为公网标签；当公网隧道为 GRE 隧道时，公网标记为 GRE 封装。
- 内层标签：又称为私网标签，用来指示报文应被送到哪个 Site。对端 PE 根据私网标签可以确定报文所属的 VPN 实例，通过查找该 VPN 实例的路由表，将报文正确地转发到相应的 Site。PE 之间在发布 EVPN 路由时，将为私网路由分配的私网标签通告给对端 PE。

图74 EVPN L3VPN 报文转发示意图



如图 74 所示，VPN 报文的转发过程为：

(1) Site 1 发出一个目的地址为 1.1.1.2 的 IP 报文，由 CE 1 将报文发送至 PE 1。

(2) PE 1 根据报文到达的接口及目的地址查找对应 VPN 实例的路由表，根据匹配的路由表项为报文添加私网标签，并查找到报文的下一跳为 PE 2。

(3) PE 1 在公网路由表内查找到达 PE 2 的路由，根据查找结果为报文封装公网标签或进行 GRE 封装，并沿着公网隧道转发该报文。

(4) MPLS 网络内，P 根据报文的公网标记转发报文，将报文转发到 PE 2。如果公网标记为 MPLS 标签，则报文在到达 PE 2 的前一跳时剥离公网标签，仅保留私网标签；如果为 GRE 封装，则由 PE 2 剥离报文的 GRE 封装。

(5) PE 2 根据私网标签确定报文所属的 VPN 实例，通过查找该 VPN 实例的路由表，确定报文的出接口，剥离私网标签后将报文转发至 CE 2。

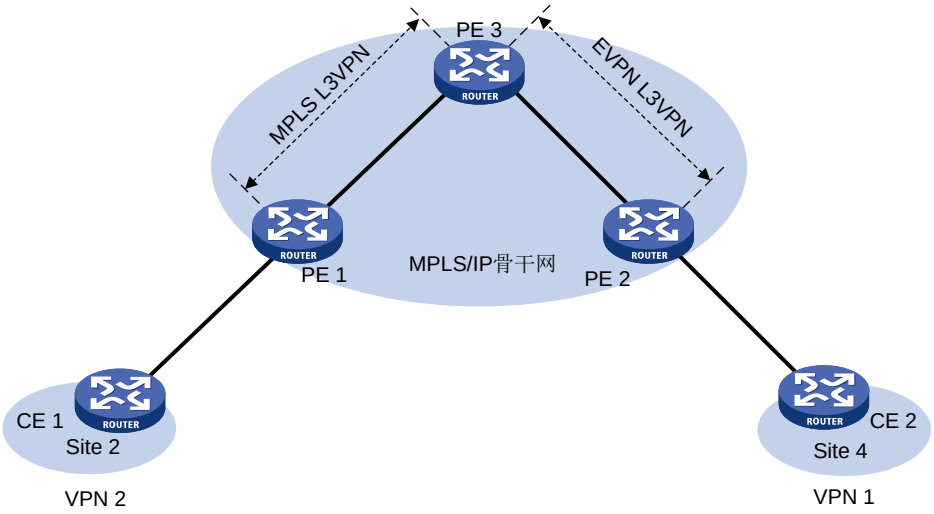
(6) CE 2 根据正常的 IP 转发过程将报文转发给目的主机。

属于同一个 VPN 的两个 Site 连接到同一个 PE 时，PE 不需要为 VPN 报文封装外层标记和内层标签，只需查找对应 VPN 实例的路由表，找到报文的出接口，将报文转发至相应的 Site。

5.4 BGP/MPLS L3VPN与EVPN L3VPN对接

将现网 L3VPN 网络改造成 EVPN L3VPN 网络的过程中，会存在两种类型网络对接的情况。通过在 PE 3 上部署 BGP VPNv4 或 BGP VPNv6 路由通过 BGP EVPN 的 IP 前缀路由发布给邻居功能和 EVPN 路由通过 BGP VPNv4 或 BGP VPNv6 地址族发布给邻居功能，可实现在 CE 1、CE 2 间跨越 MPLS L3VPN 和 EVPN L3VPN 网络建立可达路由，并进行通信。

图75 MPLS L3VPN 与 EVPN L3VPN 对接示意图



BGP/MPLS L3VPN 与 EVPN L3VPN 对接分为两部分：

- 在 PE 3 上部署 BGP VPNv4 或 BGP VPNv6 路由通过 BGP EVPN 的 IP 前缀路由发布给邻居功能，从而实现将站点 1 的路由通过 MPLS L3VPN 网络发布到 EVPN L3VPN 网络，进而发布给站点 2。具体过程为：
  a. PE 1 从 CE 1 学到 VPN 路由信息后，将其保存到 VPN 实例的路由表中。同时，为这些 IPv4 或 IPv6 路由增加 RD，形成 VPNv4 或 VPNv6 路由。
  b. PE 1 通过 MP-BGP 把 VPNv4 或 VPNv6 路由发布给 PE 3。路由中携带 VPN Target 属性及 MPLS 私网标签。
  c. PE 3 收到 VPNv4 或 VPNv6 路由后，将路由中的 Export target 与本地 VPN 实例的 Import target 进行匹配。如果二者中存在相同的值，则将该路由添加到该 VPN 实例的路由表中。

d. PE 3 上将 VPN 实例路由表中的 IPv4 或 IPv6 路由转换为 EVPN 的 IP Prefix 路由，路由的下一跳地址为 PE 3，且路由中携带 VPN Target 属性及该 VPN 的 MPLS 私网标签等信息。

e. PE 3 将该 IP Prefix 路由发送给 PE 2。

f. PE 2 收到 IP Prefix 路由后，如果路由通过 VPN Target 属性匹配，则将路由添加到 VPN 实例的路由表中。

g. PE 2 将 IPv4 或 IPv6 路由发布给 CE 2。

- 在 PE 3 上部署 EVPN 路由通过 BGP VPNv4 或 BGP VPNv6 地址族发布给邻居功能，从而实现将站点 2 的路由通过 EVPN L3VPN 网络发布到 MPLS L3VPN 网络，进而发布给站点 1。具体过程为：

  a. PE 2 从 CE 2 学到 VPN 路由信息后，将其保存到 VPN 实例的路由表中。

  b. PE 2 将 VPN 实例路由表中的 IPv4 或 IPv6 路由转换为 EVPN 的 IP Prefix 路由，该路由的下一跳地址为 PE 2，且路由中携带 VPN Target 属性及该 VPN 的 MPLS 私网标签等信息。

  c. PE 2 将该路由发布给 PE 3。

  d. PE 3 收到 IP Prefix 路由后，如果路由通过 VPN Target 属性匹配，则将路由添加到 VPN 实例的路由表中。

  e. PE 3 将 VPN 实例路由表中的 IPv4 或 IPv6 路由转换为 VPNv4 或 VPNv6 路由，并发布给 PE 1。路由中携带 VPN Target 属性及 MPLS 私网标签。

  f. PE 1 收到 VPNv4 或 VPNv6 路由后，如果路由通过 VPN Target 属性匹配，则将该路由加入到 VPN 实例的路由表。

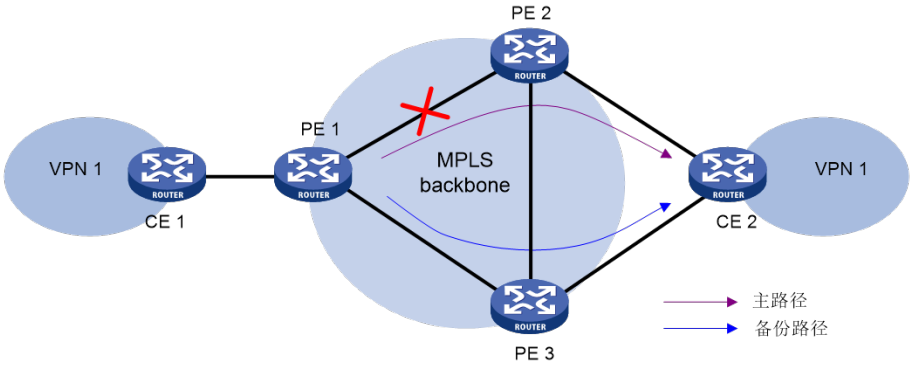
  g. PE 1 将 IPv4 或 IPv6 路由发布给 CE 1。

5.5 BGP EVPN快速重路由

当 EVPN 网络中的链路或某台路由器发生故障时，需要通过故障链路或故障路由器传输才能到达目的地的报文将会丢失或产生路由环路，数据流量将会被中断。直到根据新的网络拓扑路由收敛后，被中断的流量才能恢复正常的传输。

通过 BGP EVPN 快速重路由功能，可以尽可能地缩短网络故障导致的流量中断时间。在 BGP EVPN 地址族下开启快速重路由功能后，BGP 会为 EVPN 地址族的所有路由自动计算备份路由，即只要从不同 BGP 对等体学习到了到达同一目的网络的路由，且这些路由不等价，就会生成主备两条路由。当主路由不可达时，BGP 会使用备份路由来指导报文的转发，从而大大缩短了流量中断时间。在使用备份路由转发报文的同时，BGP 会重新进行路由优选，优选完毕后，使用新的最优路由来指导报文转发。

图76 EVPN L3VPN 快速重路由示意图



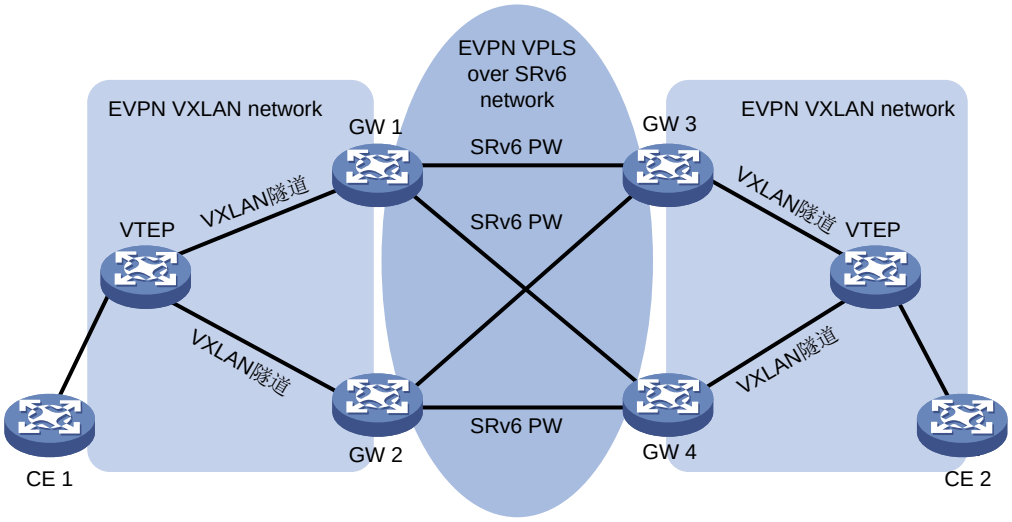
如图 76 所示，在入节点 PE 1 上配置 FRR 后，PE 1 将计算出 PE 3 为 PE 2 的备份下一跳。当 PE 1 接收到 PE 2 和 PE 3 发布的到达 CE 2 的 BGP EVPN IP 前缀路由后，PE 1 会记录这两条 BGP EVPN IP 前缀路由，并将 PE 2 发布的路由当作主路径，PE 3 发布的路由当作备份路径。

在 PE 1 上配置 BFD 检测 LSP 或 MPLS TE 隧道功能，通过 BFD 检测 PE 1 到 PE 2 之间公网隧道的状态。当公网隧道正常工作时，CE 1 和 CE 2 通过主路径 CE 1—PE 1—PE 2—CE 2 通信。当 PE 1 检测到该公网隧道出现故障时，PE 1 将通过备份路径 CE 1—PE 1—PE 3—CE 2 转发 CE 1 访问 CE 2 的流量。在此过程中，PE 1 负责主路径检测和流量切换。

6 EVPN VXLAN 与 EVPN VPLS over SRv6 网络互通

6.1 网络模型

图77 EVPN VXLAN 网络与 EVPN VPLS over SRv6 网络互通典型应用场景



EVPN VXLAN 网络与 EVPN VPLS over SRv6 网络互通功能的典型应用场景如图 77 所示。两个 EVPN VXLAN 网络通过 EVPN VPLS over SRv6 网络互联。GW 1、GW 2、GW 3 和 GW 4 为 EVPN VXLAN 网络和 EVPN VPLS over SRv6 网络的边界设备，通过在 GW 上重生成 EVPN 路由，实现 EVPN VXLAN 网络和 EVPN VPLS over SRv6 网络之间互通。

6.2 网络互通原理

EVPN VXLAN 网络和 EVPN VPLS over SRv6 网络之间实现互通，依赖于网络的边界设备上进行 MAC/IP 发布路由的重生成，具体过程为如下两种：

- GW 从 EVPN VXLAN 网络接收到 MAC/IP 发布路由后，GW 重生成该路由，即为该路由分配并添加 SRv6 SID、将路由的封装类型修改为 SRv6 封装、修改路由的 RD 和 RT，之后将该路由发送到 EVPN VPLS over SRv6 网络。
- GW 从 EVPN VPLS over SRv6 网络接收到 MAC/IP 发布路由后，GW 重生成该路由，即查找该路由对应的 VXLAN ID 并为路由添加 VXLAN ID、将路由的封装类型修改为 VXLAN 封装、修改路由的 RD 和 RT，之后将该路由发送到 EVPN VXLAN 网络。

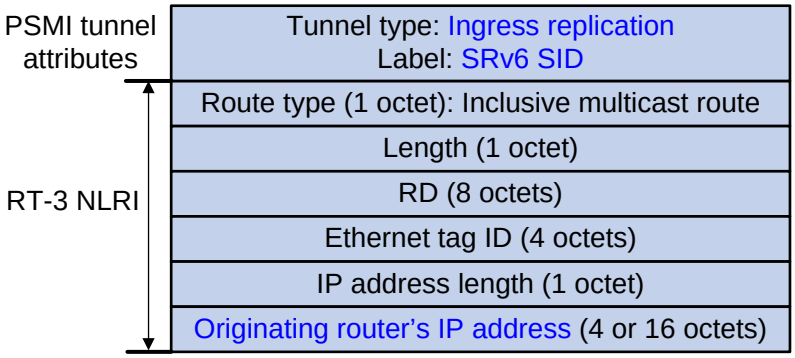
完成路由重生成后，边界网关上会形成 SRv6 SID、VXLAN ID 与 VSI 实例的映射关系。在报文转发过程中，当 GW 收到 VXLAN 封装的报文时，对报文解封装，然后重新查表，再使用 VXLAN ID 对应的 SRv6 SID 对报文进行重新封装。同样的，当 GW 收到 IPv6 或 SRv6 封装的报文时，对报文解封装，然后重新查表，再使用 SRv6 SID 对应的 VXLAN ID 对报文进行重新封装。

6.3 控制平面工作机制

6.3.1 SRv6 PW 及 BUM 广播表建立

如图 78 所示，EVPN VPLS over SRv6 组网中 PE 1 和 PE 2 依靠 RT-3（Inclusive Multicast Ethernet Tag Route）路由自动发现 PE 站点、建立 SRv6 PW 和 BUM 广播表。

图78 RT-3 路由消息格式



如图 79 所示，SRv6 PW 及 BUM 广播表建立的具体过程为：

(1) PE 1 通过 RT-3 向 PE 2 通告自身的 IP 地址，以及 PE 1 为 VSI 实例 A 分配的用于转发 BUM 流量的 End.DT2M SID。

(2) PE 2 收到 PE 1 发送的 RT-3 路由后，若路由中 VPN Target 属性与 PE 2 本地配置的 Import Target 属性匹配，则 PE 2 建立一条到达 PE 1 的 IP 地址的单向 SRv6 隧道，其出 SID 为 PE 1 携带的 End.DT2M SID；若路由中 VPN Target 属性与 PE 2 本地配置的 Import Target 属性不匹配，则不建立 SRv6 PW。

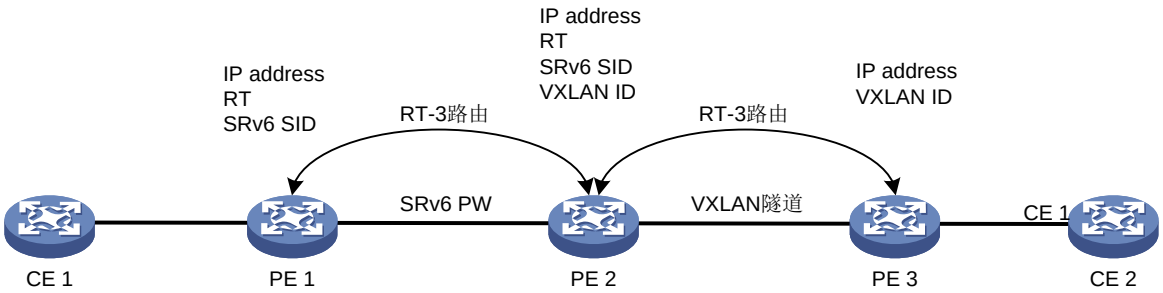
(3) PE 2 通过 RT-3 向 PE 1 通告自身的 IP 地址，以及 PE 2 为 VSI 实例 A 分配的用于转发 BUM 流量的 End.DT2M SID。

(4) PE 1 收到 PE 2 发送的 RT-3 路由后，若路由中 VPN Target 属性与 PE 1 本地配置的 Import Target 属性匹配，则 PE 1 建立一条到达 PE 1 的 IP 地址的单向 SRv6 隧道，其出 SID 为 PE 2 携带的 End.DT2M SID；若路由中 VPN Target 属性与 PE 1 本地配置的 Import Target 属性不匹配，则不建立 SRv6 PW。

(5) 两条方向相反的 SRv6 隧道，构成一条 SRv6 PW。

(6) 对于每个 VSI 而言，所有这些建立并关联的 SRv6 PW 就形成了 BUM 广播表。

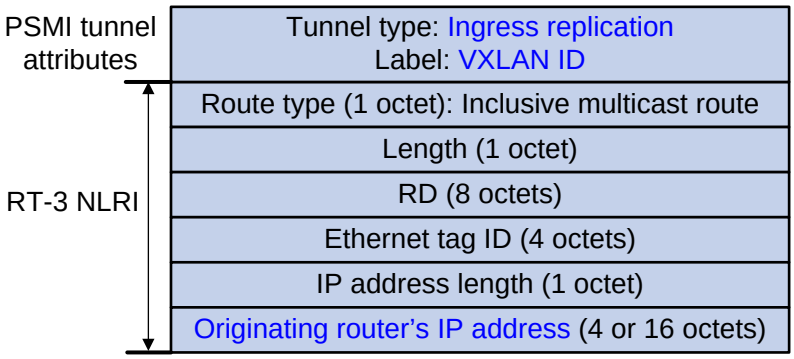
图79 SRv6 PW 及 BUM 广播表建立过程示意图



6.3.2 VXLAN 隧道及 BUM 广播表建立

如图80所示，EVPN VXLAN组网中PE 2和PE 3依靠RT-3（Inclusive Multicast Ethernet Tag Route）路由自动发现 PE 站点、建立 VXLAN 隧道和 BUM 广播表。

图80 RT-3 路由消息格式



如图 79所示，VXLAN 隧道及 BUM 广播表建立的具体过程为：

(1) PE 2 通过 RT-3 向 PE 3 通告自己所属的 VXLAN ID 及自身的 IP 地址。

(2) PE 3 收到 PE 2 发送的 RT-3 路由后，若路由中 VPN Target 属性与 PE 3 本地配置的 Import Target 属性匹配，则获取 PE 2 的 VXLAN 信息以及 VXLAN 和下一跳的关系；若路由中 VPN Target 属性与 PE 3 本地配置的 Import Target 属性不匹配，则丢弃该报文。

(3) PE 3 查看自己的 VXLAN 信息，若存在与 PE 2 相同的 VXLAN，则与 RT-3 路由中携带的下一跳自动建立 VXLAN 隧道，并将此 VXLAN 隧道与 VXLAN 关联。

(4) PE 3 通过 RT-3 向 PE 2 通告自己所属的 VXLAN ID 及自身的 IP 地址。

(5) PE 2 收到 PE 3 发送的 RT-3 路由后，若路由中 VPN Target 属性与 PE 2 本地配置的 Import Target 属性匹配，则获取 PE 3 的 VXLAN 信息以及 VXLAN 和下一跳的关系；若路由中 VPN Target 属性与 PE 2 本地配置的 Import Target 属性不匹配，则丢弃该报文。

(6) PE 2 查看自己的 VXLAN 信息，若存在与 PE 3 相同的 VXLAN，则与 RT-3 路由中携带的下一跳自动建立 VXLAN 隧道，并将此 VXLAN 隧道与 VXLAN 关联。

(7) 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。

6.3.3 MAC 地址学习

如图 81 所示，EVPN VXLAN 与 EVPN VPLS over SRv6 互通组网中，PE 之间通过 MAC/IP 发布路由（即 RT-2，二类路由）通告和学习站点的 MAC 地址。

图81 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 82 所示，MAC 地址通告和学习过程为：

(1) PE 1 接收到 CE 1 发送的报文后，通过以太网报文的源 MAC 地址学习获得 CE 1 的 MAC 地址。即将 CE 1 的 MAC 地址添加到接收报文的 AC 口关联的 VSI 实例 A 的 MAC 地址表中，其出接口为连接 CE 1 的 AC 口。

(2) PE 1 通过 RT-2 路由将 CE 1 的 MAC 地址以及 PE 1 为 VSI 实例 A 分配的 End.DT2U SID 发布给 PE 2。

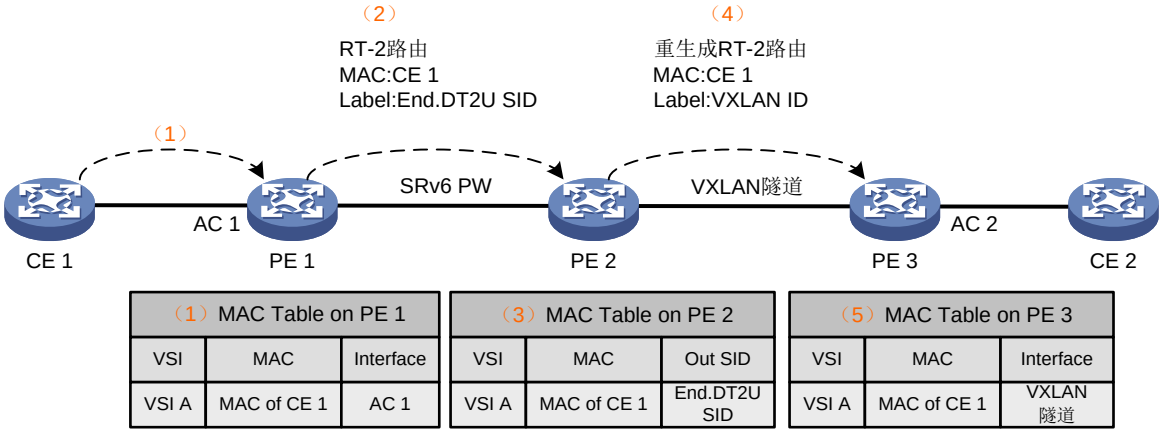
(3) PE 2 接收到 RT-2 路由后，如果路由中的 VPN Target 属性与 PE 2 本地配置的 Import target 属性匹配，则 PE 2 接收该路由，并将 CE 1 的 MAC 地址加入到本地 VSI 实例 A 的 MAC 地址表中，其出 SID 为 PE 1 通告的 End.DT2U SID；如果路由中的 VPN Target 属性与 PE 2 本地配置的 Import target 属性不匹配，则 PE 2 丢弃该路由。

(4) PE 2 将从 PE 1 接收到 RT-2 路由的下一跳地址修改为自身的地址、将报文封装类型修改为 VXLAN 方式，并添加 VSI 实例 A 关联的 VXLAN ID，然后通过 RT-2 路由发送给 PE 3。此外，PE 2 上还会建立 End.DT2U SID、VXLAN ID 与 VSI 实例的映射关系。

(5) PE 3 接收到 RT-2 路由后，如果路由中的 VPN Target 属性与 PE 1 本地配置的 Import target 属性匹配，则将 MAC 地址加入到本地 VSI 实例 A 的 MAC 地址表中，其出接口为接收该路由的 VXLAN 隧道。

完成 RT-2 路由发布后，PE 1、PE 2 和 PE 3 上均有 CE 1 的 MAC，PE 2 上还具有 End.DT2U SID、VXLAN ID 与 VSI 实例的映射关系。用户二层数据报文可以通过查找 MAC 地址表在 EVPN VXLAN 网络和 EVPN VPLS over SRv6 网络之间转发。

图82 MAC 地址通告和学习过程



（1）MAC Table on PE 1

VSI	MAC	Interface
VSI A	MAC of CE 1	AC 1

（3）MAC Table on PE 2

VSI	MAC	Out SID
VSI A	MAC of CE 1	End.DT2U SID

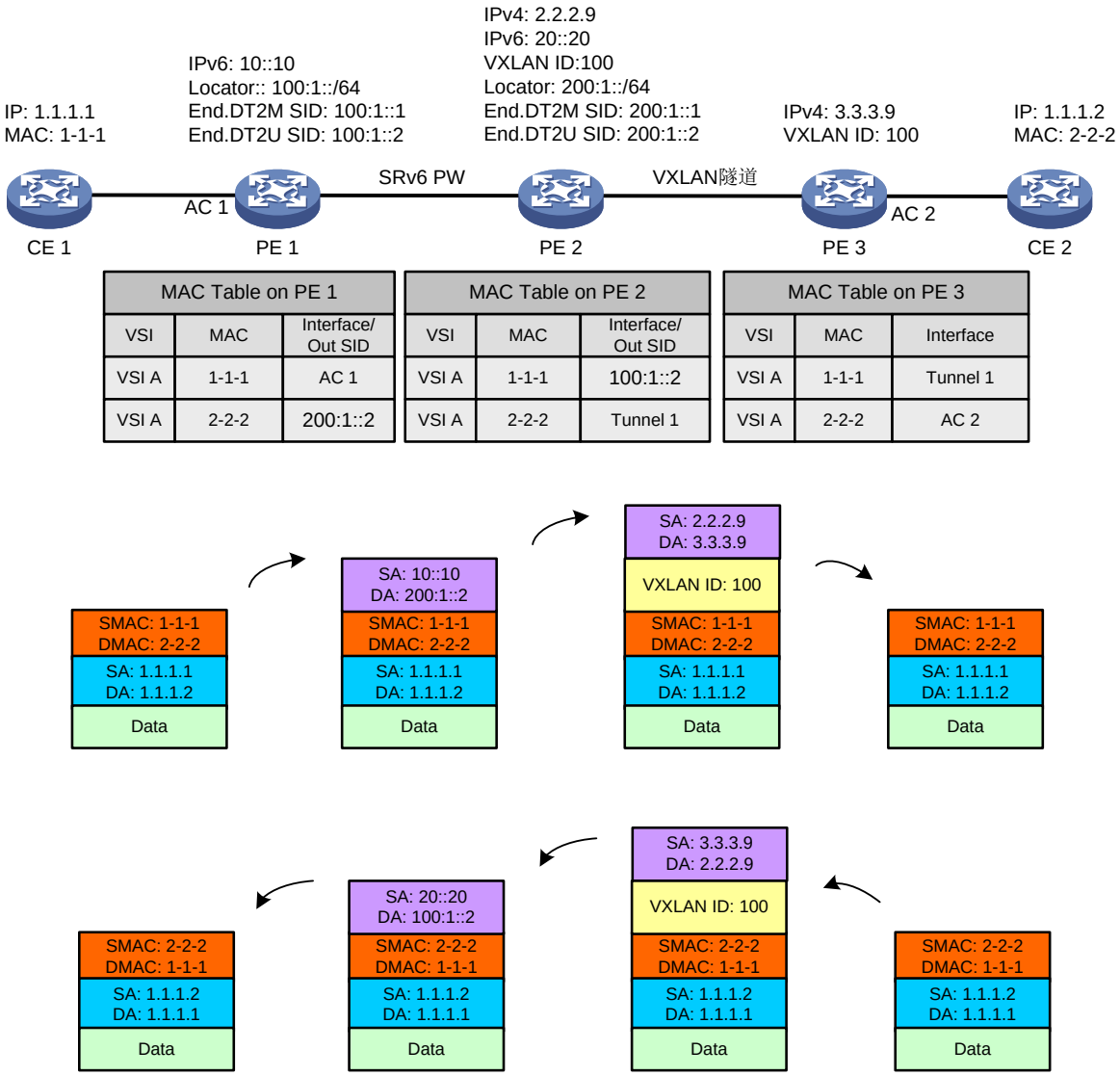
（5）MAC Table on PE 3

VSI	MAC	Interface
VSI A	MAC of CE 1	VXLAN 隧道

6.4 数据平面工作机制

6.4.1 转发已知单播流量

图83 转发已知单播流量



如图 83所示，PE 完成 MAC 地址表项的学习后，以 CE 1 访问 CE 2 为例，已知单播报文从 EVPN VPLS over SRv6 网络转发到 EVPN VXLAN 网络的转发过程为：

(1) PE 1 从 AC 接收到 CE 1 发送的报文后，在 AC 关联的 VSI 实例内查找 MAC 地址表，找到对应的出接口 SRv6 PW，并获取该 VSI 实例的 End.DT2U SID。

(2) PE 1 根据 SRv6 BE 或 SRv6 TE 方式为报文封装 IPv6 报文头，然后查找 IPv6 路由表将报文发送到 PE 2。

(3) PE 2 接收到报文后，对报文解封装，去掉 IPv6 报文头，根据 End.DT2U SID 找到关联的 VSI 实例，在 VSI 实例内查找 MAC 地址表，找到对应的 VXLAN 隧道，然后为报文进行 VXLAN 封装，通过 VXLAN 隧道将报文发送给 PE 3。

(4) PE 3 接收到报文后，对报文解封装，根据 VXLAN 报头中的 VXLAN ID 找到关联的 VSI，并在该 VSI 实例内查找 MAC 地址表，根据查表结果将报文转发给 CE 2。

以 CE 2 访问 CE 1 为例，已知单播报文从 EVPN VXLAN 网络转发到 EVPN VPLS over SRv6 网络的转发过程为：

(1) PE 3 从 AC 接收到 CE 2 发送的报文后，在 AC 关联的 VSI 实例内查找 MAC 地址表，找到对应的出接口 VXLAN 隧道，并获取 VSI 实例关联的 VXLAN ID。

(2) PE 3 为报文进行 VXLAN 封装，并通过 VXLAN 隧道将报文发送给 PE 2。

(3) PE 2 接收到报文后，对报文进行解封装，根据 VXLAN 报头中的 VXLAN ID 找到关联的 VSI，并在该 VSI 实例内查找 MAC 地址表，获取对应的 End.DT2U SID。

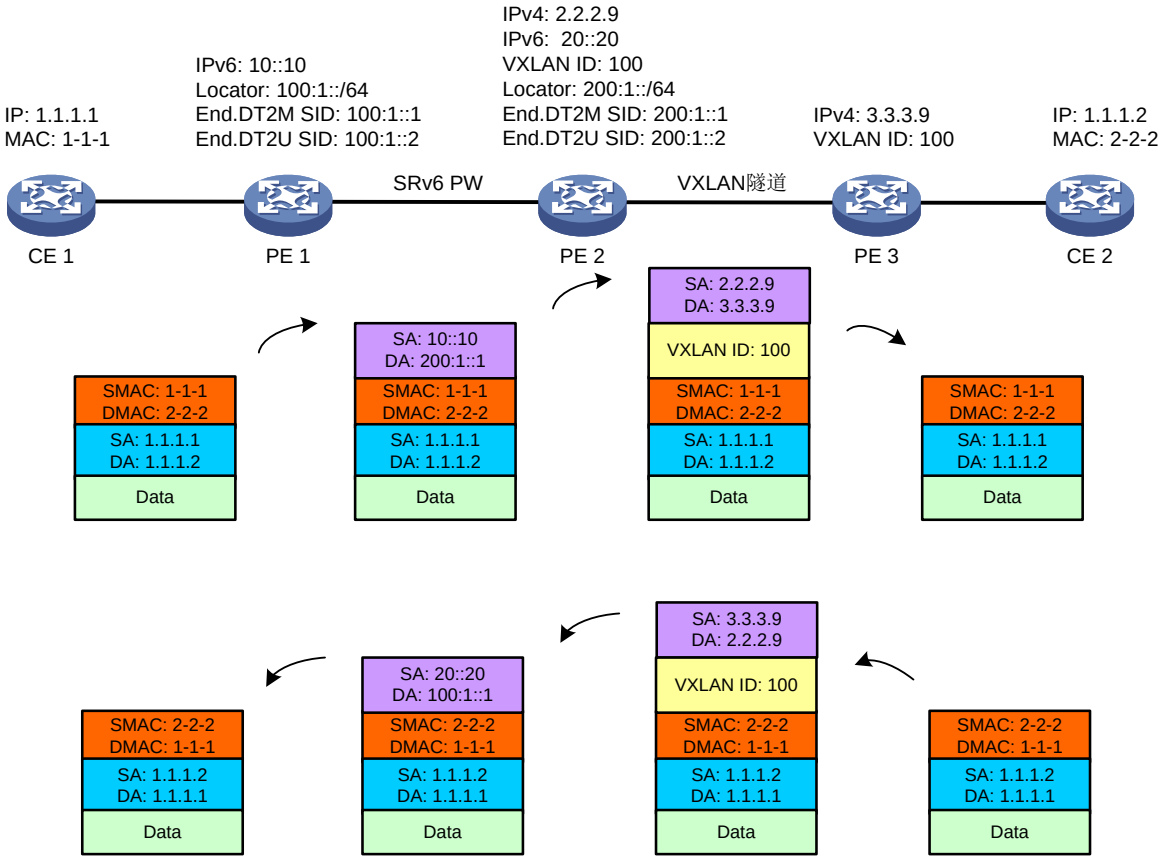
(4) PE 2 根据 SRv6 BE 或 SRv6 TE 方式为报文封装 IPv6 报文头，然后查找 IPv6 路由表将报文发送到 PE 1。

(5) PE 1 接收到报文后，对报文解封装，去掉 IPv6 报文头，根据 End.DT2U SID 找到关联的 VSI 实例，在该 VSI 实例内查找 MAC 地址表，根据查表结果将报文转发给 CE 1。

6.4.2 转发 BUM 流量

除了单播流量转发，EVPN VPLS over SRv6 与 EVPN VXLAN 网络互通组网中还会转发广播、未知组播与未知单播流量，即 BUM 流量。

图84 转发 BUM 流量



如图 84 所示，以 CE 1 侧发送未知单播报文为例，报文转发过程为：

(1) PE 1 从 AC 接收到 CE 1 发送的 BUM 报文后，在 AC 关联的 VSI 的 MAC 地址表中未找到匹配的 MAC 地址表项，则在 VSI 中查找所有远端 PE 分配的所有 End.DT2M SID，此例中只有 PE 2。

(2) PE 1 根据 SRv6 BE 或 SRv6 TE 方式为报文封装 IPv6 报文头，然后查找 IPv6 路由表将报文发送到 PE 2。如果存在多个远端 PE，则会将报文复制多份并为报文封装不同的 IPv6 报文头，然后查表发给多个远端 PE。

(3) PE 2 接收到报文后，对报文解封装，去掉 IPv6 报文头，根据 End.DT2M SID 找到关联的 VSI 实例，在 VSI 实例内查找绑定的 VXLAN 隧道，然后为报文进行 VXLAN 封装，通过 VXLAN 隧道发送给 PE 3。如果存在多个远端 PE，则会对报文进行复制并进行 VXLAN 封装，然后通过多个 VXLAN 隧道将报文转发到远端 PE。

(4) PE 3 接收到报文后，对报文解封装，根据 VXLAN 报头中的 VXLAN ID 找到关联的 VSI，在该 VSI 实例内进行广播，即通过 VSI 实例关联的所有 AC 将报文转发到 CE。

以 CE 2 侧发送未知单播报文为例，报文转发过程为：

(1) PE 3 从 AC 收到 CE 2 发送的 BUM 报文后，在 AC 关联的 VSI 的 MAC 地址表中未找到匹配的 MAC 地址表项，则在该 VXLAN 内除接收接口外的所有本地 AC 口和 VXLAN 隧道转发该报文。

(2) PE 3 为报文进行 VXLAN 封装，并通过 VXLAN 隧道将报文发送给远端 PE 2。如果存在多个远端 PE，则会对报文进行复制并进行 VXLAN 封装，然后通过多个 VXLAN 隧道将报文转发到远端 PE。

(3) PE 2 接收到报文后，对报文进行解封装，根据 VXLAN 报头中的 VXLAN ID 找到关联的 VSI，并在该 VSI 实例内查找所有远端 PE 分配的所有 End.DT2M SID，此例中只有 PE 1。

(4) PE 2 根据 SRv6 BE 或 SRv6 TE 方式为报文封装 IPv6 报文头，然后查找 IPv6 路由表将报文发送到 PE 1。如果存在多个远端 PE，则会将报文复制多份并为报文封装不同的 IPv6 报文头，然后查表发给多个远端 PE。

(5) PE 1 接收到报文后，对报文解封装，去掉 IPv6 报文头，根据 End.DT2M SID 找到关联的 VSI 实例，在该 VSI 实例内进行广播，即通过 VSI 实例关联的所有 AC 将报文转发到 CE。

7 参考文献

- RFC 7432: BGP MPLS-Based Ethernet VPN
- draft-ietf-bess-evpn-overlay: A Network Virtualization Overlay Solution using EVPN
- draft-ietf-bess-evpn-prefix-advertisement: BGP MPLS-Based Ethernet VPN
- draft-ietf-bess-evpn-igmp-mld-proxy: IGMP and MLD Proxy for EVPN
- draft-boutros-l2vpn-vxlan-evpn: VXLAN DCI Using EVPN
- draft-ietf-bess-srv6-services: SRv6 BGP based Overlay services