# VPLS 技术白皮书

Copyright © 2019 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 VPLS 技术实现 ····· 2
2.1 概念介绍 ····· 2
2.2 VPLS 网络架构 ····· 3
2.3 PW 的创建 ····· 3
2.3.1 静态 PW ····· 4
2.3.2 LDP PW ····· 4
2.3.3 BGP PW ····· 4
2.3.4 BGP 自动发现 LDP 信令 PW ····· 5
2.4 MAC 地址管理 ····· 6
2.4.1 MAC 地址学习 ····· 6
2.4.2 MAC 地址老化 ····· 7
2.4.3 MAC 地址回收 ····· 7
2.5 流量转发与泛洪机制 ····· 7
2.5.1 单播流量转发与泛洪 ····· 7
2.5.2 组播和广播流量的泛洪 ····· 8
2.6 报文封装方式 ····· 8
2.6.1 AC 上的报文封装方式 ····· 8
2.6.2 PW 上的报文封装方式 ····· 8
2.7 报文转发过程中 VLAN TAG 的处理方式 ····· 9
2.7.1 Ethernet 接入 Ethernet 模式的 VLAN TAG 处理 ····· 9
2.7.2 Ethernet 接入 VLAN 模式的 VLAN TAG 处理 ····· 10
2.7.3 VLAN 接入 Ethernet 模式的 VLAN TAG 处理 ····· 10
2.7.4 VLAN 接入 VLAN 模式的 VLAN TAG 处理 ····· 11
2.8 环路避免 ····· 11
2.9 PW 冗余保护 ····· 11
2.9.1 主备 PW 的状态 ····· 11
2.9.2 PW 冗余保护模式 ····· 12
2.9.3 PW 主备切换 ····· 12
2.10 H-VPLS 实现方式 ····· 13
2.10.1 H-VPLS 的接入方式 ····· 13

2.10.2 H-VPLS 冗余保护 15
2.10.3 环路避免方法 15
2.11 流标签 16
2.11.1 流标签能力类型 16
2.11.2 流标签配置方式 16
2.11.3 流标签报文转发 16
2.12 应用限制 17
2.12.1 QinQ 配置和 PW 上的报文封装 17
2.12.2 H-VPLS 的 QinQ 接入方式 17
3 Comware 实现的技术特色 18
3.1 H-VPLS 组网中的技术特色 18
3.1.1 MAC 地址回收 18
3.1.2 BFD 检测和冗余保护 18
3.2 丰富的组网方式 19
3.2.1 Hub-Spoke 组网方式 19
3.2.2 本地 E-Tree 组网方式 20
4 典型组网应用 20
5 参考文献 20

# 1 概述

## 1.1 产生背景

随着社会发展，经济全球化的趋势越来越明显，越来越多的企业分布范围日益扩大，公司员工的移动性也不断增加。这就迫切需要电信运营商提供链路连接，以便企业将各分支机构囊括进来，组成自己的企业网，以方便公司员工在企业以外的地方很方便地访问企业内部网络。

最初，电信运营商是以租赁专线的方式为企业提供链路，这种方式的主要缺点是：不适应当前企业分支多、增加快的特点，且价格相对昂贵，难以管理。此后，随着 ATM（Asynchronous Transfer Mode，异步传输模式）和 FR（Frame Relay，帧中继）技术的兴起，电信运营商转而使用虚电路方式为客户提供点到点的二层连接，客户在其上建立自己的三层网络以承载 IP、IPX 等数据流。但是这些技术提供的都是点到点的二层连接，配置比较复杂，尤其当增加一个站点时，管理员需要进行大量的配置工作。

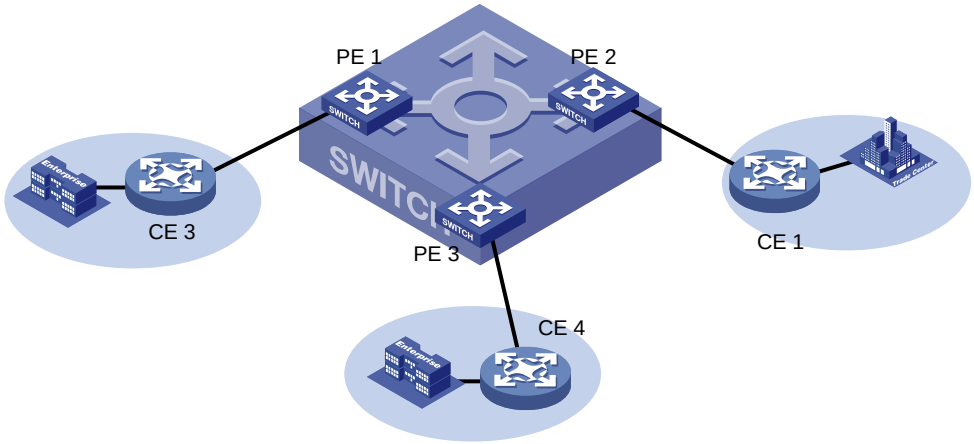
当今 IP 网络已经遍布全球，利用现有 IP 网络为企业提供低成本专网逐渐成为各大运营商的关注点。因此，一种在 IP 网上提供 VPN 服务、可方便设定速率、配置简单的技术应运而生，这种技术即 MPLS VPN 技术。基于 MPLS 的 VPN 技术有两种，分别是 MPLS L3VPN 和 MPLS L2VPN。MPLS L3VPN 需要介入用户的内部路由管理，运营商的管理比较复杂。传统 VLL（Virtual Leased Line，虚拟租用线路）方式的 MPLS L2VPN，在公网中提供一种点到点的 L2VPN 业务，可以让两个站点之间的连接效果像直接用链路连接一样，但它不能直接在服务提供者处进行多点间的交换。VPLS（Virtual Private LAN Service，虚拟专用局域网服务）在传统 MPLS L2VPN 方案的基础上发展而成，它可以实现多点到多点的 VPN 组网。VPLS 为运营商提供了一种更加完备的解决方案。

## 1.2 技术优点

VPLS 结合了以太网技术和 MPLS 技术的优势，是对传统 LAN 全部功能的仿真，其主要目的是通过运营商提供的 IP/MPLS 网络连接地域上隔离的多个由以太网构成的 LAN，使它们像一个 LAN 那样工作。

如图 1 所示，服务提供商利用 VPLS 技术在 MPLS 骨干网络上为用户网络模拟了一个以太网桥，基于 MAC 地址或者 MAC 地址加 VLAN 标识来做出转发决策。在最简单的情况下，一个 VPLS 实例包括连接到多个 PE 的多个站点，允许 CE 设备直接跟所有与该 VPLS 实例关联的其它 CE 通信。在 CE 设备看来，服务提供商网络是一个以太网桥（以太网交换机）。

图1 VPLS 模拟以太网桥




VPLS 具有以下优点：

- VPLS 在面向用户网络侧使用以太网接口，简化了 LAN/WAN 边界，可以支持快速和灵活的服务部署；
- VPLS 将用户网络的路由策略控制和维护权利交给了用户，简化了运营商网络的管理；
- VPLS 服务内的所有用户路由器 CE 是相同子网的一部分，简化了 IP 寻址规划；
- VPLS 服务既不需要感知，又不需要参与 IP 寻址和路由。

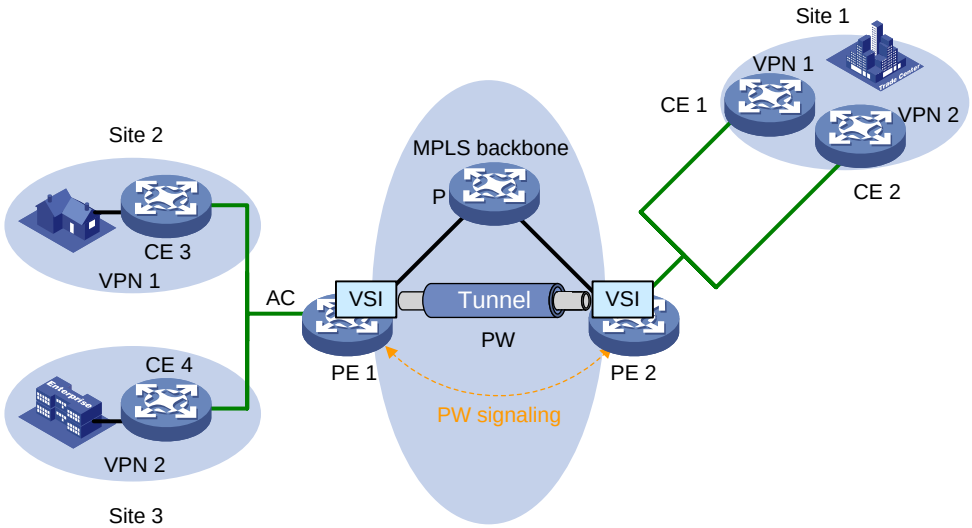
# 2 VPLS 技术实现

## 2.1 概念介绍

- CE（Customer Edge，用户网络边缘）设备：直接与服务提供商网络相连。
- PE（Provider Edge，服务提供商网络边缘）设备：与 CE 相连，主要负责 VPN 业务的接入。它完成报文从私网到公网隧道和报文从公网隧道到私网的映射与转发。在分层 VPLS（H-VPLS）体系结构下，PE 可以细分为 UPE 和 NPE。
- UPE（User facing-Provider Edge，靠近用户侧的 PE）：用于连接 CE 设备与服务商网络，主要作为用户接入 VPN 的汇聚设备，为用户提供接入服务。
- NPE（Network Provider Edge，网络核心侧 PE）：处于 VPLS 网络的核心域边缘，在核心网上提供 VPLS 透明传输服务。
- 服务界定符：服务提供商加在用户数据帧前用来标识特定 VPN 的报文标识符，服务界定符只具有本地意义，服务界定符的典型例子是 QinQ 的外层 TAG。
- QinQ（802.1Q in 802.1Q）：一种基于 802.1Q 封装的隧道协议，能够提供点到多点的 L2VPN 服务机制。它将用户网络的 VLAN TAG 封装在公网 VLAN TAG 中，最终报文带着两层 TAG 穿越服务提供商的骨干网络，从而为用户提供一种较为简单的二层 VPN 隧道。

## 2.2 VPLS网络架构

图2 VPLS 网络架构




如图 2 所示，VPLS 网络主要包括以下几个重要的组成部分：

- AC（Attachment Circuit，接入电路）：用户与服务提供商之间的连接，即连接 CE 与 PE 的链路。对应的接口只能是以太网接口。
- PW（Pseudowire，伪线）：两个 PE 设备上的 VSI 之间的一条双向虚拟连接。它由一对方向相反的单向 LSP 组成。
- 公网隧道（Tunnel）：穿越 IP 或 MPLS 骨干网、用来承载 PW 的隧道。一条公网隧道上可以承载多条 PW。公网隧道是一条本地 PE 与对端 PE 之间的直连通道，完成 PE 之间的数据透明传输，可以是 MPLS 或 GRE 隧道等。
- PW Signaling：PW 信令协议，VPLS 实现的基础，用于创建和维护 PW。PW 信令协议还可用于自动发现 VSI 的对端 PE 设备。目前，PW 信令协议主要有 LDP 和 BGP。
- VPLS 实例：用户网络可能包括分布在不同地理位置的多个站点，在骨干网上可以利用 VPLS 技术将这些站点连接起来，为用户提供一个二层 VPN。这个二层 VPN 称为一个 VPLS 实例。不同 VPLS 实例中的站点不能二层互通。
- VSI（Virtual Switch Instance，虚拟交换实例）：VPLS 实例在一台 PE 设备上的一个以太网桥功能实体，根据 MAC 地址和 VLAN TAG 进行二层报文转发。

## 2.3 PW的创建

PW 是 VPLS 在公网上的通信隧道，它建立在 MPLS（包括普通 LSP 和 CR-LSP）或 GRE 等隧道之上。创建 PW 需要：

(1) 首先在本端和对端 PE 之间建立 MPLS 或 GRE 等隧道。

(2) 确定对端 PE 的地址。对于同一个 VSI 内的 PE 设备，可以通过手工配置来指定对端 PE 地址，也可以通过信令协议自动发现对端 PE。

(3) 利用 LDP 或 BGP 信令协议为 PW 分配多路复用分离标记（PW 标签），并将分配的 PW 标签通告给对端 PE，建立单向的 LSP，从而创建 PW。如果 PW 建立在 MPLS 隧道之上，则 PW 上传输的报文将包括两层标签：内层标签为 PW 标签，用来判断报文属于的 PW，从而将报文转发给正确的 CE；外层标签为公网 MPLS 隧道标签，用来保证报文在 MPLS 隧道上的正确传输。

下面将分别介绍通过不同方式创建 PW 的过程。

### 2.3.1 静态 PW

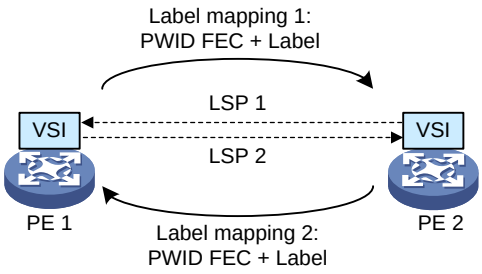
手工创建的 PW 称为静态 PW。创建静态 PW 时，需要手工指定远端 PE 的地址，并静态配置 PW 出、入两个方向的 PW 标签。两端 PE 上指定的出、入 PW 标签必须匹配。

静态方式配置简单，但无法适应网络变化，仅适用于比较稳定的小型网络。

### 2.3.2 LDP PW

采用手工指定远端 PE 的地址，并通过 LDP 信令协议将 PW 标签与 PW 的绑定关系等信息通告给远端 PE 的方式建立的 PW，称为 LDP PW。

图3 LDP PW 的建立过程




如图 3 所示，建立 LDP PW 的过程为：

(1) PE 和特定的 VSI 关联后，采用 LDP 的 DU（Downstream unsolicited，下游自主）方式主动向对端 PE 发送标签映射消息，该消息中包含 PWid FEC（FEC 128）和与该 FEC 绑定的 PW 标签，以及接口参数（如最大传输单元等）。

(2) 如果对端 PE 和这个特定的 PWID 关联，它将接受标签映射消息。

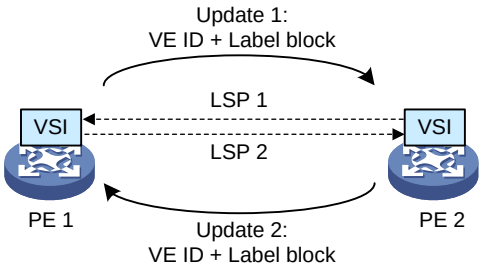
(3) 两端 PE 均接受对端的标签映射消息，一对单向的 LSP 建立成功后，它们组合起来形成双向的 PW，这个双向的 PW 可以看作是 VSI 上的一个虚拟以太网接口。

LDP 方式实现简单。但是 LDP 不能提供 VPLS 成员的自动发现机制，需要手工指定 PE 的各个对等体。新的 PE 加入时，每个 PE 上都要修改配置。

### 2.3.3 BGP PW

通过 BGP 协议将标签块等信息通告给对端 PE，两端 PE 根据标签块计算 PW 出、入两个方向标签的方式建立的 PW，称为 BGP PW。

图4 BGP PW 建立过程




如图 4 所示，建立 BGP PW 的过程为：

(1) PE 利用 BGP 的 Update 消息向所有对端 PE 设备发送 VE ID 和标签块信息。其中，VE ID 为与 PE 相连的每个 Site 在 VPN 内的唯一编号，由服务供应商统一规划；标签块包含一组连续的标签。

(2) 接收到 Update 消息的 PE 设备，根据自己的 VE ID 和报文中的标签块，计算出唯一的一个标签值，作为 PW 标签。同时，接收 Update 消息的 PE 设备，根据报文中的 VE ID 和本地的标签块，也可以得知对端 PE 的 PW 标签值等信息。

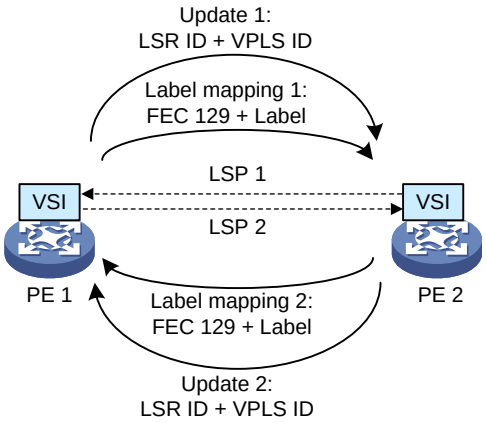
(3) 两个 PE 设备互相发送 Update 消息，并计算出 PW 标签后，两台设备间的 PW 创建成功。

BGP 方式中，通过配置 VPN Target 实现了 VPLS 成员的自动发现，增加或删除 PE 时，无需手工配置，具有较好的可扩展性，但 BGP 协议本身比较复杂。

### 2.3.4 BGP 自动发现 LDP 信令 PW

通过 BGP 协议自动发现远端 PE 后，利用 LDP 信令协议将 PW 标签与 PW 的绑定关系等信息通告给远端 PE 的方式建立的 PW，称为 BGP 自动发现 LDP 信令 PW。

图5 BGP 自动发现 LDP 信令 PW 建立过程




如图 5 建立 BGP 自动发现 LDP 信令 PW 过程为：

(1) PE 利用 BGP 的 Update 消息向所有对端 PE 设备发送 LSR ID 和 VPLS ID（用来标识本端 PE 所属的 VPLS 实例）信息。

(2) 接收到 Update 消息的 PE 设备，比较两端 PE 的 VPLS ID 是否相同。如果相同，则继续利用 LDP 信令协议在二者之间建立 PW；否则，不会建立 PW。

(3) PE 通过 BGP 协议发现可以建立 PW 的对端 PE 地址后，采用 LDP 的 DU（Downstream unsolicited，下游自主）方式主动向对端 PE 发送标签映射消息，该消息中包含 Generalized PWid FEC（FEC 129）和与该 FEC 绑定的 PW 标签等信息。其中 FEC 129 携带 VPLS ID、SAII（Source Attachment Individual Identifier，源转发实例本地标识符）和 TAII（Target Attachment Individual Identifier，目的转发实例本地标识符）等信息。SAII 用来标识本地 PE，为本地 PE 的 LSR ID；TAII 用来标识远端 PE，为远端 PE 通过 BGP 协议发布的 PE 标识。VPLS ID+SAII+TAII 可以唯一标识 VPLS 实例内的一条与 PW 标签绑定的 PW。

(4) 两端 PE 均接受对端的标签映射消息，建立 LSP 后，BGP 自动发现 LDP 信令 PW 即成功建立。

BGP 自动发现 LDP 信令 PW，结合了 LDP 方式和 BGP 方式的优点，通过 BGP 协议实现了 VPLS 成员的自动发现，通过 LDP 信令协议实现了 VPLS 成员之间的标签交换。

## 2.4 MAC地址管理

VPLS 为用户网络模拟了一个以太网桥，基于 MAC 地址或者 MAC 地址加 VLAN TAG 来做出转发决策。与一个特定的 VPLS 服务关联的每个 PE 设备都为该 VPLS 实例建立一个 VSI，每个 VSI 维护一张 MAC 地址表，并具有泛洪和转发、MAC 地址学习和老化的功能，以便实现报文的转发。

### 2.4.1 MAC 地址学习

VPLS 通过源 MAC 地址学习来提供可达性。PE 为每个 VSI 维护一张 MAC 地址表。

如图 6 所示，源 MAC 地址学习过程包含两部分：

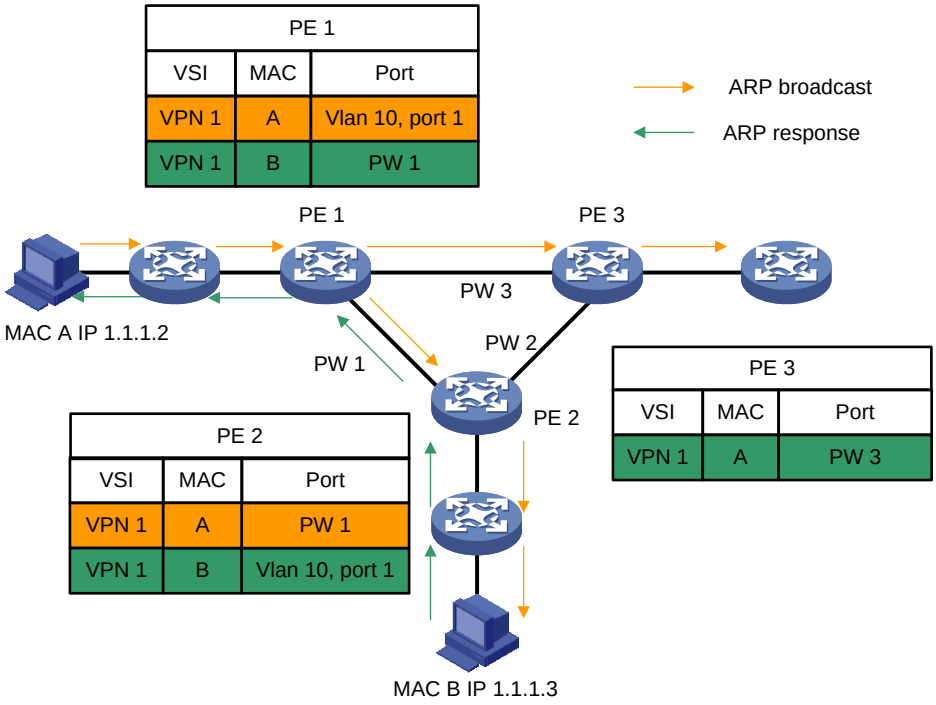
- 与 PE 直接相连的本地站点的源 MAC 地址学习

  本地站点的源 MAC 地址学习与传统以太网交换机相同。PE 从 CE 接收到报文后，如果 MAC 地址表中不存在报文源 MAC 地址，则将该报文的源 MAC 地址学习到 PE 连接 CE 的 AC 链路上。

- 通过 PW 连接的远端站点的源 MAC 地址学习

  VSI 将 PW 看作是逻辑以太网接口。PE 从 PW 上接收到报文后，如果 MAC 地址表中不存在报文源 MAC 地址，则将该报文的源 MAC 地址学习到 VSI 的 PW 逻辑以太网接口上。

图6 MAC 地址学习




### 2.4.2 MAC 地址老化

PE 学习到的 MAC 地址转发表项如果不再使用，需要有老化机制来移除。老化机制根据报文中的源 MAC 地址进行处理：PE 收到报文时，除了学习源 MAC 地址外，对于已经生成的 MAC 表项，还需要设置“激活”或者“有效”标记，一定时间内未被设置“激活”或者“有效”标记的 MAC 表项，将从 MAC 转发表中移除。

### 2.4.3 MAC 地址回收

在 AC 或 PW 状态变为 down 时，LDP 协议会发送地址回收消息通知 VPLS 实例内的所有远端 PE 删除指定 VSI 内的指定 MAC 地址，以加快 MAC 地址表的收敛速度。

## 2.5 流量转发与泛洪机制

### 2.5.1 单播流量转发与泛洪

PE 从 AC 接收到单播报文后，在与 AC 关联的 VSI 内查找 MAC 地址表，从而确定如何转发报文：

- 如果查找到目的 MAC 地址对应的表项，则根据该表项转发报文。
  - 表项的出接口为 PW 逻辑以太网接口时，为报文封装该 PW 的 PW 标签，并添加公网隧道封装后，通过 PW 将该报文转发给远端 PE。如果 PW 由 LSP 或 MPLS TE 隧道承载，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，用来决定报文所属的 PW，从而将报文转发给正确的 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站点的接口时，直接通过出接口将报文转发给本地站点。

- 如果没有找到目的 MAC 地址对应的表项，则向 VSI 内的所有其他 AC 对应的接口和所有 PW 逻辑以太网接口泛洪该报文。

PE 从 PW 接收到单播报文后，在 PW 所属的 VSI 内查找 MAC 地址表，从而确定如何转发报文：

- 如果查找到目的 MAC 地址对应的表项，则根据该表项转发报文。该表项的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。
- 如果没有找到目的 MAC 地址对应的表项，则向 VSI 内所有 AC 对应的接口泛洪该报文。

### 2.5.2 组播和广播流量的泛洪

PE 从 AC 上接收到组播或广播报文后，向该 AC 关联的 VSI 内的所有其他 AC 对应的接口和所有 PW 逻辑以太网接口泛洪该报文。

PE 从 PW 上接收到组播或广播报文后，向该 PW 所属的 VSI 内所有 AC 对应的接口泛洪该报文。

## 2.6 报文封装方式

### 2.6.1 AC 上的报文封装方式

AC 上的报文封装方式分为两种：VLAN 接入和 Ethernet 接入。其含义如下：

- VLAN 接入：CE 发送给 PE 或 PE 发送给 CE 的以太网帧头带有一个 VLAN TAG，该 TAG 是一个服务提供商网络为了区分用户而压入的“服务界定符”。服务界定符一般是服务提供商设备添加的，我们把这个作为服务界定符的 TAG 称为 P-TAG。
- Ethernet 接入：CE 发送给 PE 或 PE 发送给 CE 的以太网帧头中没有服务界定符，如果此时帧头中有 VLAN TAG，则说明它只是用户报文的内部 VLAN TAG，对于 PE 设备没有意义。这种用户内部 VLAN 的 TAG 称为 U-TAG。

### 2.6.2 PW 上的报文封装方式

PW 由 PWID 和 PW 封装类型唯一标识。两端 PE 设备通告的 PWID 和 PW 封装类型必须相同。

PW 上的报文封装方式分为两种：Ethernet 模式和 VLAN 模式。

- Ethernet 模式下，PW 上传输的帧不能带 P-TAG

  对于 CE 侧的报文：

  - 如果 PE 从 CE 收到带有 P-Tag 的报文，则将其去除后再压入 PW 标签和公网隧道封装转发；
  - 如果从 CE 收到不带 P-Tag 的报文，则直接压入 PW 标签和公网隧道封装后转发。

  对于 PE 发送给 CE 的报文：

  - 如果 AC 的接入模式为 VLAN，则添加 P-Tag 后转发给 CE；
  - 如果 AC 的接入模式为 Ethernet，则不添加 P-Tag，直接转发给 CE；
  - 不允许重写或去除已经存在的任何 Tag。
- VLAN 模式下，PW 上传输的帧必须带 P-TAG

  对于 CE 侧的报文：

- 从CE收到带有P-Tag的报文后，如果远端PE不要求Ingress改写P-Tag，则保留P-Tag，如果远端 PE 要求 Ingress 改写 P-Tag，则将 P-Tag 改写为远端 PE 期望的 VLAN Tag（Tag 可能是值为 0 的空 Tag），再压入 PW 标签和公网隧道封装后转发；
- 从 CE 收到不带 P-Tag 的报文后，如果远端 PE 不要求 Ingress 改写 P-Tag，则添加值为 0 的空 P-Tag，如果远端 PE 要求 Ingress 改写 P-Tag，则添加一个远端 PE 期望的 VLAN Tag（Tag 可能是值为 0 的空 Tag）后，再压入 PW 标签和公网隧道封装后转发。

对于 PE 发送给 CE 的报文：

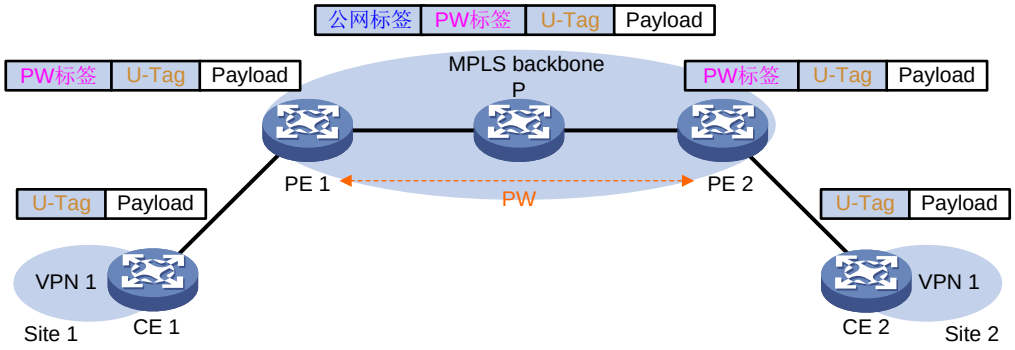
- 如果 AC 的接入模式为 VLAN，转发给 CE 时重写或保留 P-Tag；
- 如果 AC 的接入模式为 Ethernet，则去除 P-Tag 后转发给 CE。

## 2.7 报文转发过程中VLAN TAG的处理方式

根据 AC 接入方式和 PW 上报文封装模式的不同，报文转发中 VLAN TAG 的处理方式有所不同。

### 2.7.1 Ethernet 接入 Ethernet 模式的 VLAN TAG 处理

图7 Ethernet 接入 Ethernet 模式的 VLAN TAG 处理




如图 7 所示，AC 采用 Ethernet 接入方式，PW 上的报文封装模式为 Ethernet 模式时，报文的转发过程为：

(1) CE 1 将携带用户所属 VLAN 信息（U-Tag）的报文发送给 PE 1。

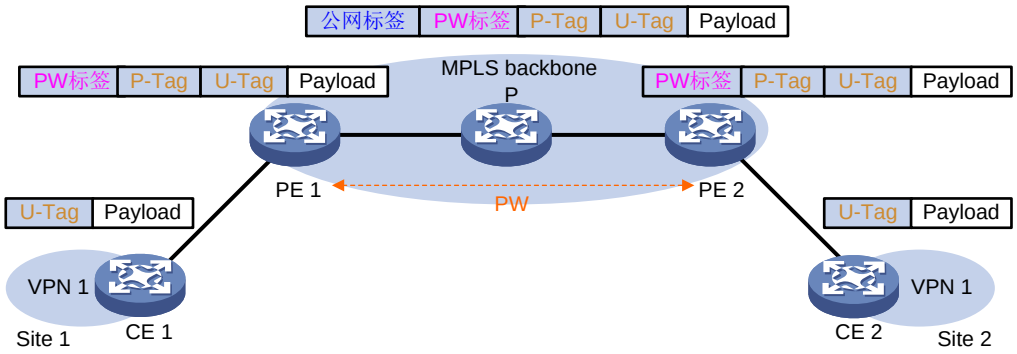
(2) PE 1 根据用户目的 MAC 地址（或用户所属的 VLAN 和目的 MAC 地址），选择合适的 PW，并在报文中添加 PW 对应的 PW 标签。

(3) 为了在公网上利用 MPLS 隧道转发报文，PE 1 在报文中添加公网隧道标签，通过公网隧道将报文传递给 PE 2。

(4) PE 2 收到报文后，根据 PW 标签找到报文所属的 VSI，并将携带 U-TAG 的报文发送给 CE 2。

### 2.7.2 Ethernet 接入 VLAN 模式的 VLAN TAG 处理

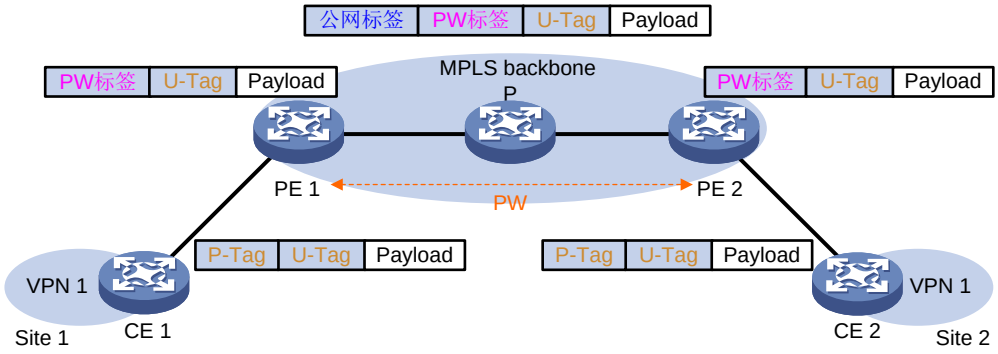
图8 Ethernet 接入 VLAN 模式的 VLAN TAG 处理




如图 8 所示，AC 采用 Ethernet 接入方式，PW 上的报文封装模式为 VLAN 模式时，报文的转发过程与 Ethernet 接入方式、Ethernet 模式的报文转发过程类似。所不同的是，PW 上传输的帧必须带 P-TAG。PE 1 从 CE 1 收到不带 P-TAG 的报文后，首先添加一个对端 PE 期望的 VLAN TAG 或空 TAG，再压入两层 MPLS 标签转发。PE 2 接收到报文后，去除 MPLS 两层标签和 P-TAG，再将报文转发给 CE 2。

### 2.7.3 VLAN 接入 Ethernet 模式的 VLAN TAG 处理

图9 VLAN 接入 Ethernet 模式的 VLAN TAG 处理




如图 9 所示，AC 采用 VLAN 接入方式，PW 上的报文封装模式为 Ethernet 模式时，报文的转发过程为：

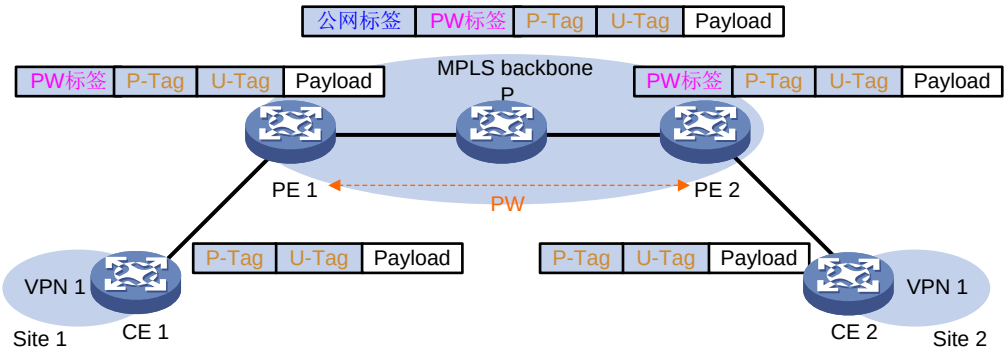
(1) CE 1 发送的报文携带服务界定符 P-TAG；

(2) PE 1 接收到报文后去除 P-TAG，并添加两层 MPLS 标签，通过公网 MPLS 隧道，将报文发送给 PE 2；

(3) PE 2 去除接收到报文中的两层 MPLS 标签，增加 P-TAG 后，将报文转发给 CE 2。

### 2.7.4 VLAN 接入 VLAN 模式的 VLAN TAG 处理

图10 VLAN 接入 VLAN 模式的 VLAN TAG 处理




如图 10 所示，AC 采用 VLAN 接入方式，PW 上的报文封装模式为 VLAN 模式时，报文的转发过程与 VLAN 接入方式 Ethernet 模式的报文转发过程类似。二者的区别是，PW 上传输的帧携带 P-TAG，PE 接收到报文后，修改或保留报文中的 P-TAG。

## 2.8 环路避免

为了避免环路，一般的二层网络都要求使能 STP 协议。但是对使用 VPLS 的用户来说，不会感知到 ISP 的网络，因此在私网侧使能 STP 的时候，不能把 ISP 的网络考虑进来。因而，VPLS 中使用 PW 全连接和水平分割转发来避免环路。

VPLS 环路避免的方法如下：

- PE 之间逻辑上全连接（PW 全连接），也就是每个 PE 必须为每一个 VPLS 转发实例创建一棵到该实例下的所有其他 PE 的树。
- 每个 PE 设备必须支持水平分割转发来避免环路。如果从 PW 上收到报文，那么这个报文将不再向这个 VSI 关联的其它 PW 上转发，也就是说要求任意两个 PE 之间通过直接相连的 PW 通信，而不能通过第三个 PE 设备中转报文，这也就是 VPLS 的 VSI 实例之间要求建立全连接 PW 的原因。

## 2.9 PW冗余保护

如果两个 CE 之间只存在一条 PW，则当 PE 节点、PE 与 CE 之间的链路、或 PE 之间的 PW 出现故障时，CE 之间将无法通信。PW 冗余保护功能通过部署主备两条 PW，实现当主 PW 出现故障后，将流量立即切换到备份 PW，使得流量转发得以继续。目前，只有静态 PW 和 LDP PW 支持 PW 冗余保护功能。

### 2.9.1 主备 PW 的状态

主备 PW 的状态分为 Active 和 Standby。

- Active：表示该 PW 可以用于业务传送。
- Standby：表示该 PW 处于备份状态，不能用于业务传送。

### 2.9.2 PW 冗余保护模式

对于 LDP PW，PW 两端的 PE 通过 LDP 通告消息协商主备 PW 的状态，协商方法由 PW 冗余保护模式决定：

- 独立操作模式

  PW 两端的 PE 独立操作，并把 PW 在本端的 Active/Standby 状态通过 LDP 通告消息告知远端 PE。每个 PE 分别比较 PW 在本端和远端的 Active/Standby 状态。一条 PW 只有在本端和远端都是可工作的，并且在本端和远端都同时处于 Active 状态时，该 PW 才可以用来传送客户业务。在仅一端处于 Active 状态、另一端处于 Standby 状态或两端都处于 Standby 状态的 PW 上不允许传送客户业务。需要由用户保证两端 PE 的 PW 主备配置一致。

- 主从操作模式

  PW 两端的 PE 不是独立操作的，其中一个 PE 作为主节点，另一个 PE 作为从节点。

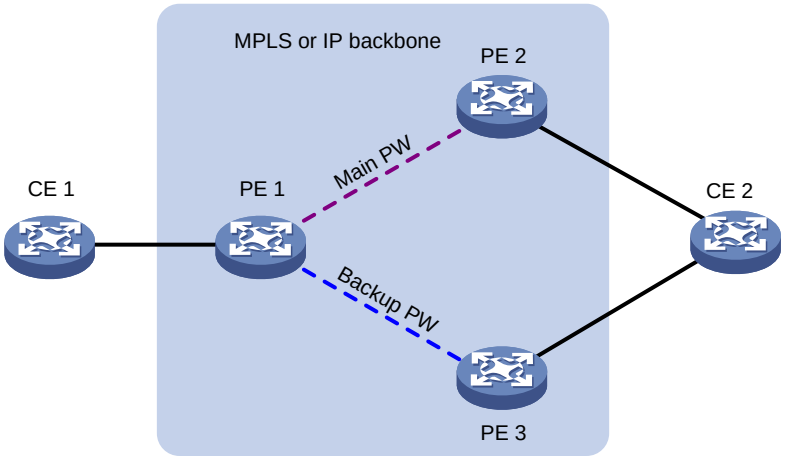
  主节点决定了本地 PW 的 Active 状态、Standby 状态后，通过 LDP 通知消息将该状态通告给从节点。从节点接收到主节点的 LDP 通知消息后，保持本地的 PW 状态与主节点一致，从而保证主、从节点均在相同的、处于 Active 状态的 PW 上传送客户业务。

  从节点无需向主节点通告 PW 在本端的 Active/Standby 状态，主节点也忽略来自于从节点的任何关于 PW 的 Active/Standby 状态的 LDP 通知消息。

### 2.9.3 PW 主备切换

如图 11 所示，在两个 CE 之间建立两条 PW 链路，正常情况下，CE 使用主 PW 与远端 CE 通信；当 PE 1 检测出到 PE 2 的 PW 不可用（可能是 PE 2 节点故障，也可能是 PW 故障，或 PE 2 与 CE 2 之间的链路故障），PE 1 将启用备份 PW，通过备份 PW 将 CE 1 的报文转发给 PE 3，再由 PE 3 转发给 CE 2。CE 2 接收到报文后，通过更新 MAC 地址表项等方式将发送给 CE 1 的报文切换到备份 PW 转发，从而保证通信不会中断。

图11 VPLS 的 PW 冗余保护




VPLS 根据控制平面的 LDP 会话状态，或者数据平面连通性检测结果等来判断 PW 是否可以继续使用。在以下情况下，设备认为 PW 不可用：

- 承载 PW 的公网隧道被拆除或通过 BFD 等检测机制检测到公网隧道出现故障，导致 PW 的状态变为 down；

- 控制平面拆除 PW（如 PW 两端 PE 之间的 LDP 会话 down 导致 PW 被删除），或利用 BFD 等链路检测机制检测到 PW 故障；
- 执行命令手工切换主备 PW。

对于静态 PW，PE 判断主 PW 不可用后，将流量切换到备份 PW，以确保流量转发不中断。

对于 LDP PW，PE 根据上述条件分别确定本地主、备 PW 的 Active/Standby 状态后，通过 LDP 消息中的 PW Preferential Forwarding 状态位携带本端 PW 的状态，将该状态通告给对端 PE：

- 该状态位置位时，表示本端 PW 处于 Standby 状态。
- 该状态位未置位时，表示本端 PW 处于 Active 状态。

两端 PE 根据本地 PW 状态和 PW 冗余保护模式，确定使用哪条 PW 转发流量。

## 2.10 H-VPLS实现方式

如上文所述，VPLS 要求 PE 之间全连接，因此一个 VPLS 实例的 PW 的条数跟 PE 设备的个数之间的关系是：PW 的条数＝PE 的个数×（PE 的个数－1）/2。在 VPLS 网络规模比较大的情况下，PW 的数目非常庞大，PW 信令开销很大，网络的管理和扩展都将变得复杂。为了简化网络管理和提高网络的扩展性，引出了 H-VPLS（Hierarchical VPLS，分层 VPLS）的组网方式。

H-VPLS 将 PE 划分为 UPE 和 NPE。UPE 主要作为用户接入 VPN 的 MTU（Multi-Tenant Unit，多租用单元，如路由器或交换机设备），用于连接 CE 和服务商网络；NPE 处于 VPLS 网络的核心域边缘，提供用户报文在核心网上的透明传输服务。UPE 不需要与所有的 NPE 建立全连接，只需在 NPE 之间建立全连接。H-VPLS 通过分级，减少了 PW 的数目和 PW 信令的负担。

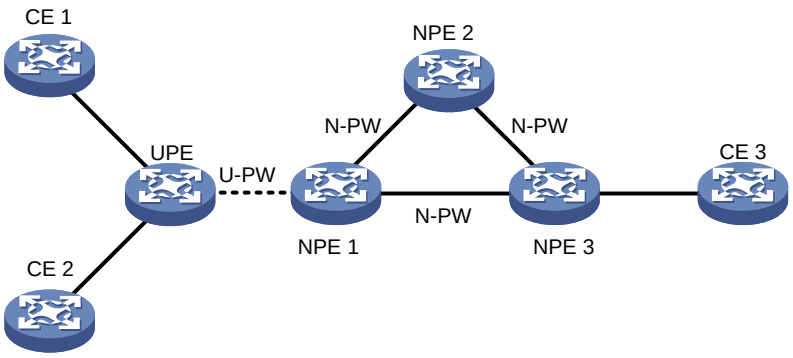
### 2.10.1 H-VPLS 的接入方式

根据 UPE 和 NPE 之间连接方式的不同，H-VPLS 分为：

- LSP 接入方式
- QinQ 接入方式

#### 1. LSP 接入方式

图12 LSP 接入方式




如图 12 所示，UPE 作为汇聚设备 MTU，它只跟 NPE 1 建立一条虚连接接入链路 U-PW（建立 U-PW 需要在 NPE 和 UPE 设备上创建 VSI 实例，指定对等体，并要求两台设备上的 PWID 必须相同），跟其他所有的对端都不建立虚连接。

数据转发流程如下：

(1) UPE 负责将 CE 上送的报文发给 NPE 1，同时打上 U-PW 对应的 PW 标签（NPE 1 分配的 PW 标签，作为多路 PW 复用分离标记）；

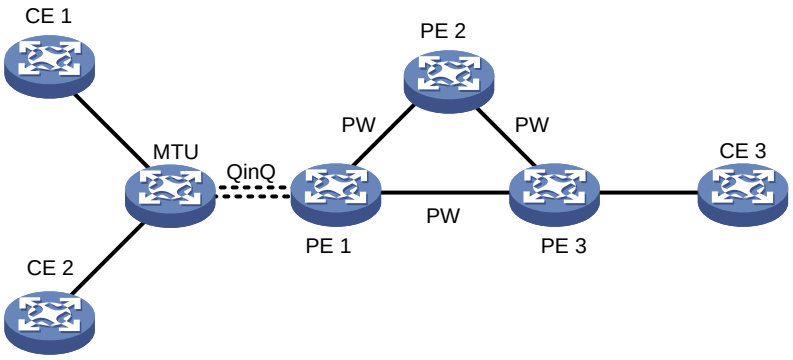
(2) NPE 1 收到报文后，先根据 PW 标签判断报文所属的 VSI，再根据该报文的目的 MAC 压入 N-PW 对应的 PW 标签，然后转发该报文；

(3) NPE 1 从 N-PW 侧收到报文后，打上 U-PW 对应的 PW 标签，将报文发送给 UPE，UPE 再将报文转发给 CE。

CE 1 与 CE 2 之间的数据交换为本地 CE 之间交换时，如果 UPE 本身具有桥接功能，UPE 将直接完成两者间的报文转发，而无需将报文上送给 NPE 1。不过对于目的 MAC 未知的第一个数据报文或广播报文，UPE 在将数据通过桥广播到 CE 2 的同时，仍然会通过 U-PW 转发给 NPE 1，由 NPE 1 来完成报文的复制并转发到各个对端 CE。

### 2. QinQ 接入方式

图13 QinQ 接入方式




如图 13 所示，MTU（也可以称为 UPE）为标准的桥接设备，在 MTU 和 PE 1 之间建立点到点的以太网 QinQ 连接（即在 MTU 面向 CE 的接口上使能 QinQ，在与 MTU 直连的 PE 1 上使用 VLAN 接入方式）。MTU 从 CE 上收到的报文都将被打上一层外层 VLAN 标记，报文转发到 PE 1 时，根据配置的 VLAN 接入方式，外层 VLAN 标记被解释为服务提供商 VLAN 标记，也就是服务提供商为用户分配的服务界定符，根据这个界定符，报文被映射到相应的 VSI 实例，由这个 VSI 实例决策出报文如何转发（单播或广播）。

数据转发流程如下：

(1) 在 CE 接入端口使能 QinQ，为收到的报文添加压入 VLAN TAG 作为多路复用分离标记，在 MTU 与 PE 1 之间通过 QinQ 隧道将报文透明传输到 PE 1；

(2) PE 1 先根据报文携带的 VLAN TAG 判断所属的 VSI，再根据该报文的目的 MAC 为其压入 PW 对应的 PW 标签，然后将其转发；

(3) PE 1 从 PW 侧收到报文后，根据 PW 标签，判断报文所属的 VSI，再根据用户报文的目的 MAC 打上 VLAN TAG 通过 QinQ 隧道将报文转发给 MTU，由 MTU 将报文转发给 CE。

如果 CE 1 与 CE 2 之间的数据交换为本地 CE 之间交换，由于 MTU 本身具有桥接功能，MTU 将直接完成两者间的报文转发，而无需将报文上送给 PE 1。不过对于目的 MAC 未知的第一个数据报文或广播报文，MTU 在通过桥广播到 CE 2 的同时，仍然会通过 QinQ 隧道转发给 PE 1，由 PE 1 来完成报文的复制并转发到各个对端 CE。

## 2.10.2 H-VPLS 冗余保护

UPE 与 NPE（或 MTU 与 PE）之间只有单条链路连接的方案具有明显的弱点，一旦该接入链路出现故障，汇聚设备连接的所有 VPN 都将丧失连通性。所以，对于 H-VPLS 的两种接入模式，都需要有冗余备份链路存在。在正常情况下，设备只使用一条链路（主链路）接入，一旦 VPLS 系统检测到接入链路失败，它将启用备用链路继续提供 VPN 业务。

对于 LSP 接入方式的 H-VPLS，由于 UPE 与 NPE 之间运行 LDP 会话，可以根据 LDP 会话的活动状态来判断主 PW 是否失效；对于 QinQ 接入方式的 H-VPLS，需要在 MTU 与它相连的 PE 设备之间运行 STP，保证在一条链路失败以后启用另一条链路。

图14 LSP 接入方式的冗余保护

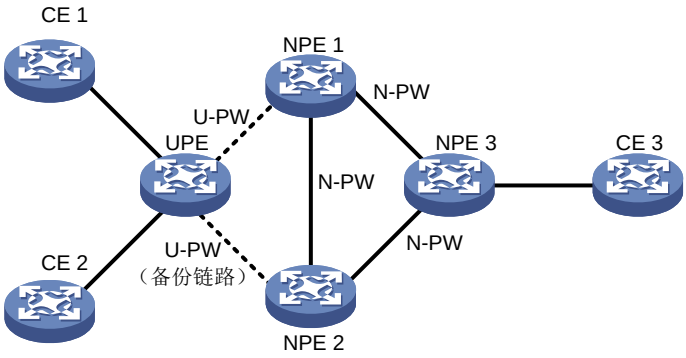
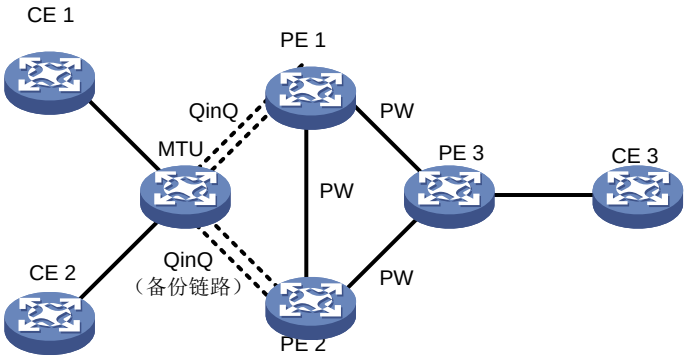



图15 QinQ 接入方式的冗余保护




## 2.10.3 环路避免方法

H-VPLS 中环路避免方法需要做如下调整：

- 只需要在 NPE 之间建立全连接（PW 全连接），UPE 和 NPE 之间不需要全连接。

- 每个 NPE 设备上，从与 NPE 连接的 PW 上收到的报文，不再向这个 VSI 关联的、与其它 NPE 连接的 PW 转发，但可以向与 UPE 连接的 PW 转发。
- 每个 NPE 设备上，从与 UPE 连接的 PW 上收到的报文，可以向这个 VSI 关联的所有与其它 NPE 连接的 PW 转发。

## 2.11 流标签

不同类型的数据流可能通过同一条 PW 来传输，这些数据流在 PE 节点上封装完全相同的 PW 标签。封装了 PW 标签的报文到达 P 节点时，尽管 P 节点上存在多条隧道可以进行负载分担，但同一条 PW 的多条数据流仍然只能选择同一条路径转发，不能针对不同的数据流进行负载分担。L2VPN 流标签功能可以实现在 P 设备上针对不同的数据流进行负载分担。

### 2.11.1 流标签能力类型

PE 节点上的 L2VPN 流标签能力分为：

- 流标签接收能力：PE 从 PW 上接收到报文后，能够识别报文中的流标签，并在解封装时删除流标签字段。
- 流标签发送能力：PE 在通过 PW 发送报文时，会在 PW 标签后增加流标签字段。

### 2.11.2 流标签配置方式

PE 节点上的 L2VPN 流标签配置方式分为：

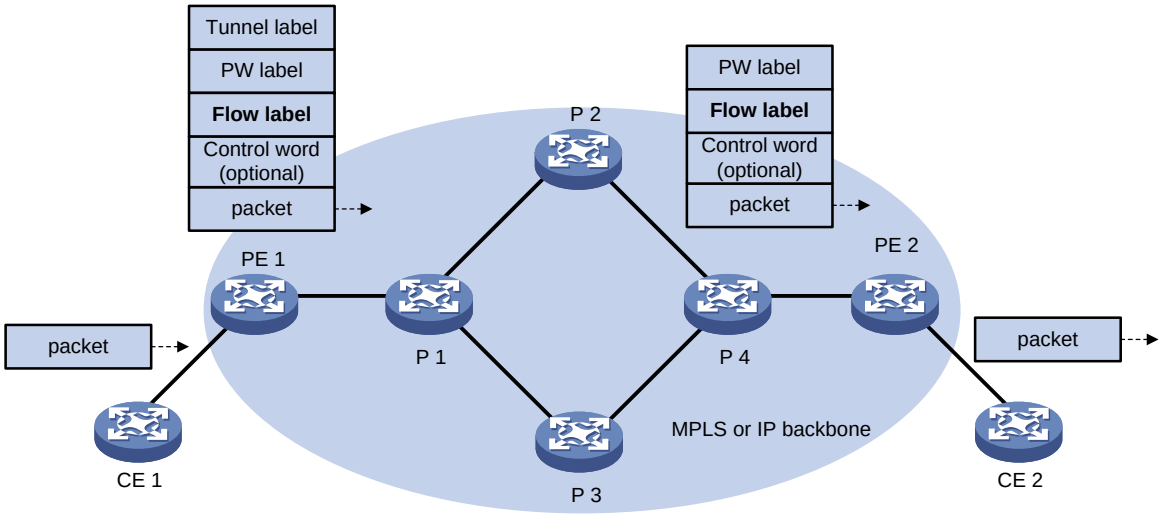
- 手工指定：由用户保证两端 PE 的 L2VPN 流标签能力匹配，即一端具有接收能力、另一端具有发送能力。否则，可能会导致流量处理失败。
- 动态协商：在信令协议报文（如 LDP 报文）中携带本端的流标签能力，以便进行流标签能力协商。只有两端 PE 的 L2VPN 流标签能力匹配，流标签功能才能协商成功。如果协商失败，则两端 PE 均不具备流标签能力，采用正常的转发流程转发报文。

静态 PW 两端 PE 的 L2VPN 流标签能力只能手工指定；动态创建 PW 两端 PE 的 L2VPN 流标签能力既可以手工指定，也可以动态协商。

### 2.11.3 流标签报文转发

如图 16 所示，在两端的 PE 节点上均开启 L2VPN 流标签功能后，入口 PE 对数据报文进行封装时，会在 PW 标签后加入一个流标签（Flow label）字段，不同类型的数据流可以添加不同的流标签。P 节点根据流标签进行负载分担。出口 PE 剥离 PW 标签和流标签后，将报文转发到本地站点。

图16 L2VPN 流标签示意图




## 2.12 应用限制

### 2.12.1 QinQ 配置和 PW 上的报文封装

某些产品上没有根据“2.6.1 AC 上的报文封装方式”来解析收到的报文携带的外层 TAG，而是根据私网接口是否使能 QinQ 和 PW 上的报文封装方式来确定 TAG 的处理方式。在这些产品上需要注意：

- VPLS 业务板处理 VPLS 业务流量时，业务处理板始终将外层 TAG 当作 P-TAG（服务界定符）来处理。在私网接口没有使能 QinQ 的情况下，如果 PW 工作在 Ethernet 模式，业务处理板处理 VPLS 业务流量报文时，直接将外层 TAG 剥离（即使这个 TAG 是 U-TAG）。因此应用时，如果 PW 工作在 Ethernet 模式又希望不剥离 U-TAG，则必须在私网接口上使能 QinQ，这样报文送到 VPLS 业务板处理时，只剥离 QinQ 加上的 TAG。
- 向对端通告的 Requested VLAN ID 与 VPLS 允许接入的 AC 数目（即与一个 VSI 绑定的接口数目）有关。如果 VSI 与单个 AC 绑定，那么 Requested VLAN ID 取值为本地私网接口对应的 VLAN ID，否则为 0。因此如果一个 VSI 可以跟多个 AC 绑定，PW 上的报文封装模式为 VLAN 模式时，PW 上传递的报文携带的 P-TAG 为 0 TAG。

### 2.12.2 H-VPLS 的 QinQ 接入方式

H-VPLS 的 QinQ 接入方式组网中，需要注意以下几点：

- MTU 设备上，与 NPE 相连的以太网接口和与 CE 相连的以太网接口上需要使能 STP 或者 MSTP，BPDU 报文在两台 NPE 之间传递，以避免出现环路。
- 在 NPE 设备上，与 MTU 连接的以太网接口上一定不能使能 MSTP，否则 BPDU 协议报文无法正常传递。
- 为了防止 BPDU 协议报文传送到 VPLS 网络域中的其他 PE 设备，BPDU 协议报文和用户数据报文需要映射到不同的 VPLS 实例。

# 3 Comware 实现的技术特色

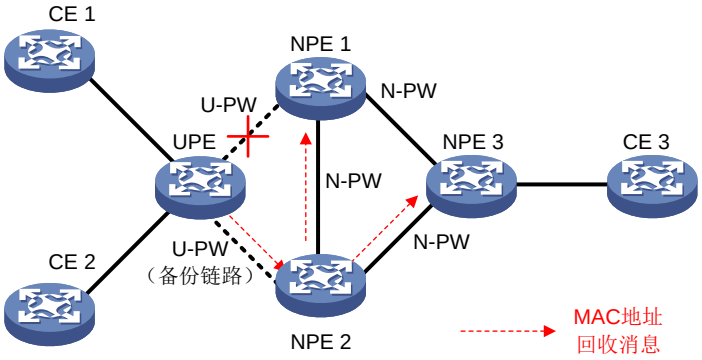
## 3.1 H-VPLS组网中的技术特色

### 3.1.1 MAC 地址回收

UPE 双归属于两个 NPE，当 UPE 与 NPE 之间使用的 PW（如主 PW）失效，则 UPE 发起 PW 切换，激活另一条 PW。但是在 PW 失效后的一段时间内，其它站点的 NPE 设备仍然向这条 PW 连接的 NPE 转发流量。当流量到这台 NPE 时，将无法继续转发。为了提高收敛速度，当发生 PW 切换时，应尽可能快地通知其它 NPE 清除相应 VSI 本地 MAC 表项、触发 MAC 地址的重新学习和 MAC 转发路径的重新建立。LDP 协议的地址回收消息提供了这个机制。

Comware 的实现方式为由 UPE 设备发送 MAC 地址回收消息。如图 17 所示，UPE 向新激活的 PW 连接的 NPE 设备发送 MAC 地址回收消息，NPE 收到地址回收消息后，向其它 NPE 转发地址回收消息。

图17 Comware 实现的 MAC 地址回收方式




地址回收消息中携带 MAC TLV，收到这个消息的设备根据 TLV 中指定的参数进行 MAC 地址的删除或重新学习这些 MAC 地址。当 MAC 地址数目巨大的时候，为了加快收敛速度，可以发送一个空的 MAC 地址列表。收到地址回收消息后，NPE 将移除指定 VSI 中的所有 MAC 地址（从发送此消息的 PE 处学习到的 MAC 地址除外）。Comware 只支持 MAC 地址列表为空的地址回收方式。

### 3.1.2 BFD 检测和冗余保护

为了实现冗余保护，H-VPLS 组网中，UPE 与两个 NPE 之间建立主备 PW 连接。当主用 PW 发生故障时，快速切换到备用 PW，以保证通信的连续性。

BFD 是一套全网统一的检测机制，用于快速检测、监控网络中链路连通状况。在 UPE 和 NPE 之间采用 BFD（报文周期最快可达 10ms）检测 UPE 和 NPE 之间的路由可达性。UPE 设备上，通过 BFD 检测到与对端 NPE 之间出现连通性故障时，UPE 设备发起 PW 的切换过程，将所有与 NPE 上的 VSI 连接的 PW，都切换到备用的 PW 上，以保证通信的连续性。

## 3.2 丰富的组网方式

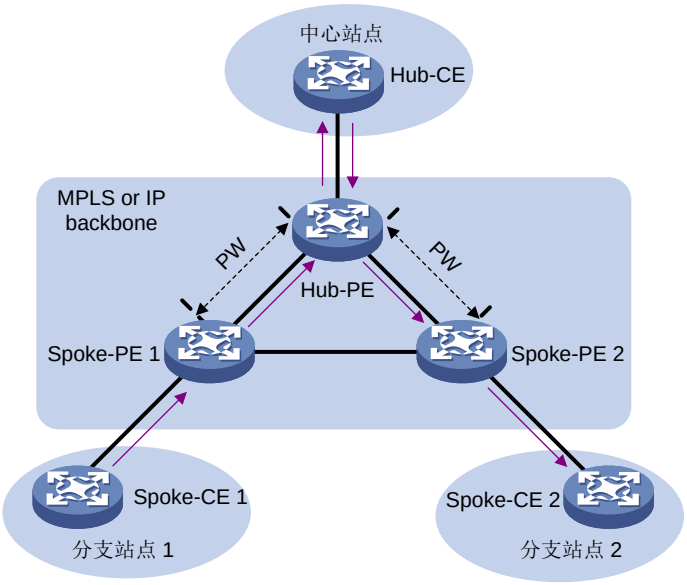
除了传统的 VPLS 组网方式外，Comware 还支持 Hub-Spoke 和本地 E-tree 组网方式，以满足用户的不同网络需求。

### 3.2.1 Hub-Spoke 组网方式

Hub-Spoke 是指网络中存在一个中心站点（Hub 站点）和多个分支站点（Spoke 站点）的组网方式。在 Hub-Spoke 组网方式中，分支站点之间不能直接通信，必须通过中心站点通信，以实现中心站点对数据流量进行统一管理。

目前，只有静态 PW 和 LDP PW 支持 Hub-Spoke 组网方式。

图18 VPLS 的 Hub-Spoke 组网方式




如图 18 所示，VPLS 支持 Hub-Spoke 组网方式中：

- 用户网络中心站点的 CE 称为 Hub-CE。
- 与用户网络中心站点连接的 PE 称为 Hub-PE。
- 用户网络分支站点的 CE 称为 Spoke-CE。
- 与用户网络分支站点连接的 PE 称为 Spoke-PE。
- 朝向中心站点方向的链路（AC 或 PW）称为 Hub 链路。用户需要手工指定 VSI 内的 Hub 链路。在一个 VSI 内只能存在一条 Hub 链路。
- 朝向分支站点方向的链路（AC 或 PW）称为 Spoke 链路。

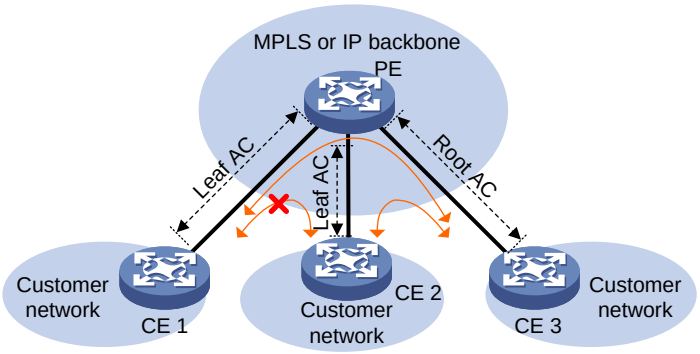
在 VPLS 的 Hub-Spoke 组网方式中，PE 设备按照如下规则进行 MAC 地址学习和报文转发：

- 从 Spoke 链路接收到报文后，进行 MAC 地址学习，并将报文转发到 Hub 链路。
- 从 Hub 链路接收到报文后，不进行 MAC 地址学习，查找 MAC 地址表后，根据查找结果将报文转发到 Spoke 链路。

### 3.2.2 本地 E-Tree 组网方式

E-Tree 组网用于点到多点业务。在本地 E-Tree 组网中，将用户的接入链路（AC）的角色分为 Root 和 Leaf，本地同一个 VSI 内 Leaf 之间相互隔离，Root 与 Leaf、Root 与 Root 之间可以互通，从而灵活地控制 AC 之间的通信。

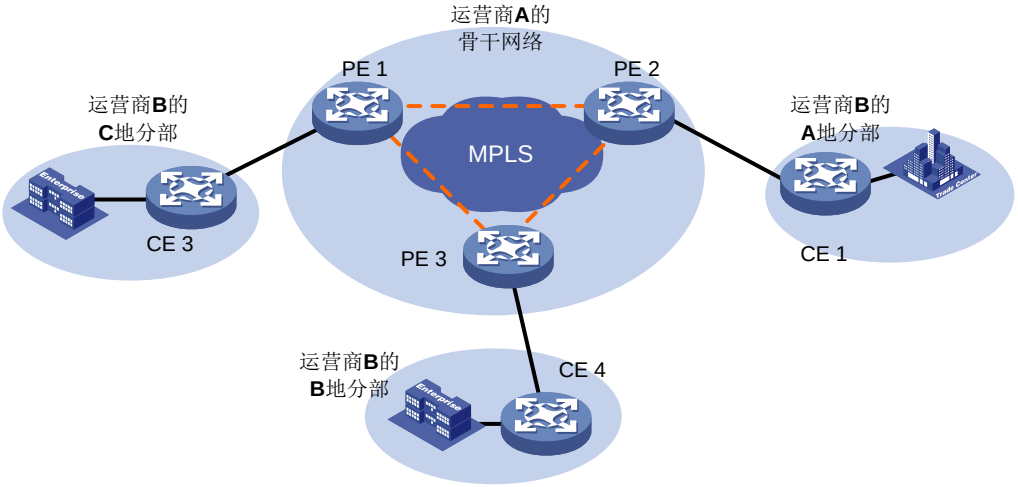
图19 VPLS 的本地 E-Tree 组网方式




# 4 典型组网应用

VPLS 适用于网络庞大、路由数目众多、有自己的网络维护力量、分支机构分布多个不同地域、对服务质量要求较高的大型企业。如图 20 所示，运营商 A 拥有全国骨干，运营商 B 在多个城市有自己的驻地网，希望租用运营商 A 的带宽，将各地的网络互连起来。运营商 B 具有足够的网络管理、维护能力。为了保证服务质量，并保持路由的私有性，运营商 B 采用了 VPLS 的组网方案。

图20 典型组网应用




# 5 参考文献


- RFC 4447: Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)
- RFC 4761: Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling

- RFC 4762: Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling