

# 第 5 篇 BGP/MPLS VPN

---

第 15 章 MPLS 技术基础

第 16 章 BGP MPLS VPN 基本原理

第 17 章 BGP MPLS VPN 配置与故障排除

## 第15章 MPLS 技术基础

上世纪 90 年代初，随着 Internet 的快速普及，网络上的数据量日益增大，而由于当时硬件技术的限制，采用最长匹配算法、逐跳转发方式的传统 IP 转发路由器日益成为限制网络转发性能的一大瓶颈。因此快速转发路由器技术成为当时研究的热点。然而在解决该问题上被赋予众望的 ATM 技术却因为技术复杂、成本高昂让人望而却步。在这个情形下，迫切需要一种介于 IP 和 ATM 之间的技术，以适应网络发展的需要。

### 15.1 本章目标

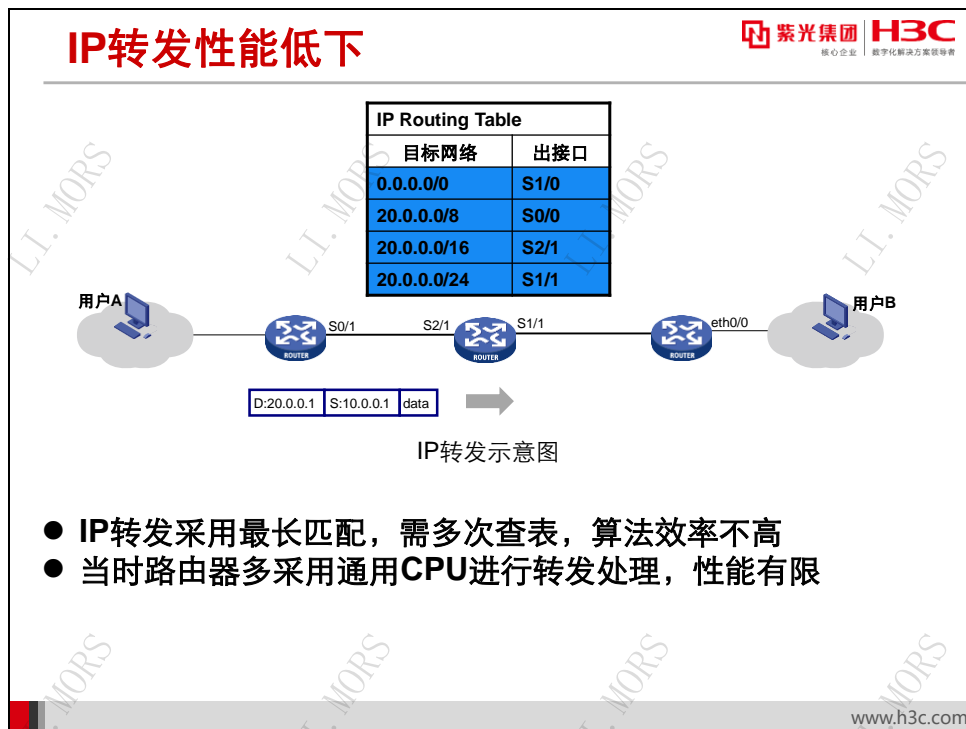
#### 课程目标

学习完本课程，您应该能够：

- 了解MPLS技术产生背景
- 掌握MPLS技术实现原理
- 理解MPLS标签分配、数据转发过程



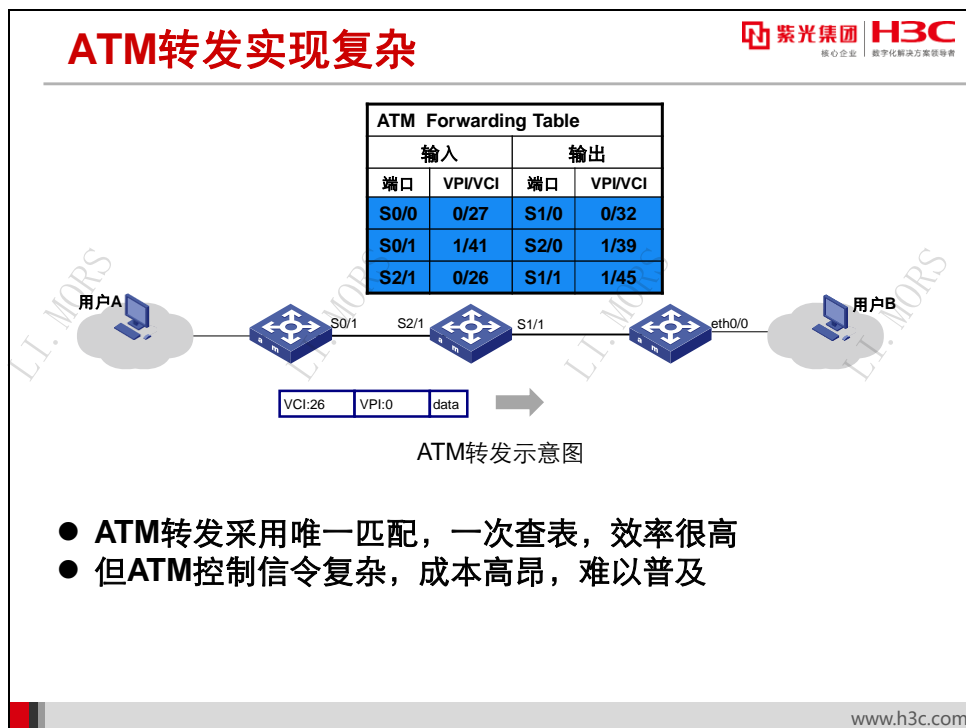
## 15.2 MPLS起源



随着 Internet 的迅速普及，传统的路由器设备因其转发性能低下，逐渐成为网络的瓶颈。

首先，路由器采用的转发算法效率不高。路由器普遍采用 IP 转发。IP 转发的原则是最长匹配算法。路由器在判断该如何转发一个数据包时，需要遍历整个路由表，找出最能精确表达该数据目的地址所在位置的那一条路由。如上图所示，报文目的地址为 20.0.0.1，在路由表中有四条路由，都能涵盖 20.0.0.1，但路由器在转发该报文时，需要遍历整个路由表，并对比确定最能精确说明 20.0.0.1 所在位置的 20.0.0.1/24 这条路由，才能确定如何转发该报文。随着网络规模的增大，路由表的规模也逐步增大，遍历路由表需要花费越来越多的时间；而数据量也随着网络规模的扩大逐步上升，路由器变得不堪重负。

此外，当时路由器多采用通用 CPU 进行转发处理，性能有限，对 IP 地址和路由的匹配运算需要耗费较多的处理器时间。



ATM (Asynchronous Transfer Mode, 异步传送模式) 协议采用定长的标签代替 IP 地址, 数据包抵达 ATM 交换机后只需要一次查表, 就能找出与其唯一匹配的表项, 确定报文的出接口。如上图所示, 整个 ATM 转发表中只有唯一的一个表项与图中抵达的数据包相匹配。ATM 转发算法可以大大的提高报文的转发效率, 然而 ATM 技术的控制信令实现复杂, 它独立于 IP 转发中各种路由协议计算出的路由表项, 采用一套复杂的专用表项建立机制形成 ATM 转发表。

因为实现复杂, 支持 ATM 技术的网络设备成本也相对高昂, 用户将原有采用 IP 转发的网络改造成 ATM 转发的网络, 需要投入较高的成本。也因为复杂, ATM 技术的普及度不高, 众多网络维护人员不能像维护 IP 网络一样熟练维护 ATM 网络。

所以 ATM 技术只得到了较小规模的应用, 没有能如设计者预期的那样替代 IP 转发。人们希望能在 IP 和 ATM 之间取一个平衡, 既可以提高转发效率, 还要容易实现。

## MPLS的概念



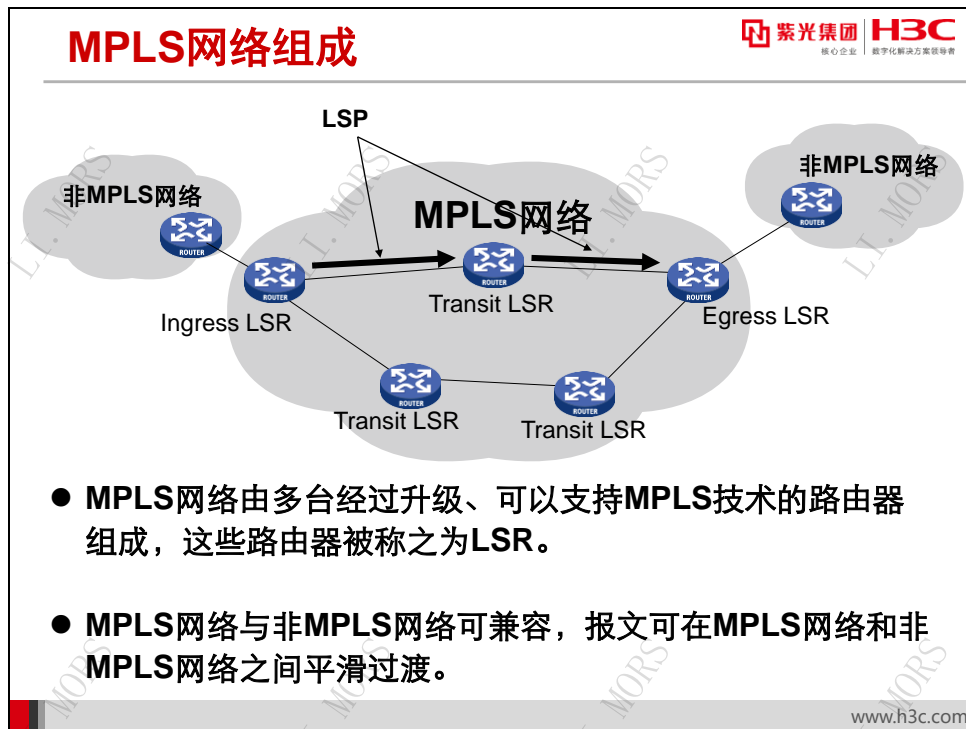
- **MPLS (Multiprotocol Label Switching, 多协议标签交换)**
- **MPLS用一个短而定长的标签来封装网络层分组，交换机或路由器根据标签值转发报文**
- **MPLS中的多协议是指：**
  - MPLS可以承载在各种链路层协议上，如PPP、ATM、帧中继、以太网等
  - 原理上各种报文也可以承载在MPLS之上，如IPv4/IPv6、也包括各种链路层报文如ATM等

www.h3c.com

MPLS (Multiprotocol Label Switching, 多协议标签交换) 的基本概念是用一个短而定长的标签来封装网络层分组，并将标签封装后的报文转发到已升级改进过的交换机或者路由器，交换机或路由器根据标签值转发报文。后文将详细阐述 MPLS 的具体实现过程。

MPLS 中的多协议有多层含义。一方面是指 MPLS 协议可以承载在多种二层协议之上，如常见的 PPP、ATM、帧中继 (FR)、以太网等等；另一方面多种报文也可以承载在 MPLS 之上，如 IPV4 报文、IPv6 报文等，甚至也包括各种二层报文。与多种协议的兼容性是 MPLS 协议得以普及的重要原因之一。

## 15.3 MPLS网络组成



MPLS 网络架构与普通的 IP 网络相比，并无任何特殊性。普通的 IP 网络，其路由器只要经过升级，支持 MPLS 功能，就成了 MPLS 网络。在 MPLS 网络中的路由器，具有标签分发能力和标签交换能力，被称之为 LSR。此外，MPLS 网络可以与非 MPLS 网络共存，报文可在非 MPLS 网络和 MPLS 网络之间进行转发。

## LSR



- **LSR (Label Switching Router) :** LSR是具有标签分发能力和标签交换能力的设备，是MPLS网络中的基本元素。
- **入节点Ingress:** 报文的入口LSR，负责为进入MPLS网络的报文添加标签。
- **中间节点Transit:** MPLS网络内部的LSR，根据标签沿着由一系列LSR构成的LSP将报文传送给出口LSR。
- **出节点Egress:** 报文的出口LSR，负责剥离报文中的标签，并转发给目的网络

www.h3c.com

MPLS 网络的基本构成单元是 LSR(Label Switching Router)，是具有标签分发能力和标签交换能力的设备。

MPLS 网络包括以下几个组成部分：

- **入节点 Ingress:** 报文的入口 LSR，负责为进入 MPLS 网络的报文添加标签。
- **中间节点 Transit:** MPLS 网络内部的 LSR，根据标签沿着由一系列 LSR 构成的 LSP 将报文传送给出口 LSR。
- **出节点 Egress:** 报文的出口 LSR，负责剥离报文中的标签，并转发给目的网络。

## FEC与LSP

紫光集团 H3C  
核心企业 数字化转型领导者

- **FEC (Forwarding Equivalence Class, 转发等价类)**：MPLS将具有相同特征（目的地相同或具有相同服务等级等）的报文归为一类，称为FEC。属于相同FEC的报文在MPLS网络中将获得完全相同的处理。
- **LSP (Label Switching Path, 标签交换路径)**：属于同一个FEC的报文在MPLS网络中经过的路径称为LSP

www.h3c.com

FEC (Forwarding Equivalence Class, 转发等价类) 是 MPLS 中的一个重要概念。MPLS 将具有相同特征（目的地相同或具有相同服务等级等）的报文归为一类，称为 FEC。属于相同 FEC 的报文在 MPLS 网络中将获得完全相同的处理。

属于同一个 FEC 的报文在 MPLS 网络中经过的路径称为 LSP (Label Switched Path, 标签交换路径)。LSP 是一条单向报文转发路径。在一条 LSP 上，沿数据传送的方向，相邻的 LSR 分别称为上游 LSR 和下游 LSR。




## 15.4 MPLS 标签

### 15.4.1 MPLS 标签基本概念

### MPLS 标签定义

- **MPLS 标签 (Label)** 是一个比较短的，定长的，通常只具有局部意义的标识
- **MPLS 标签**通常位于报文的数据链路层封装头和网络层封装之间
- 路由器可以根据标签决定如何转发报文



紫光集团 H3C  
核心企业 数字化解决方案领导者

www.h3c.com

MPLS 的标签是一个比较短的，定长的，通常只有局部意义的标识，这个标识通常位于报文的链路层头和网络层头之间，路由器可以根据标签来决定如何转发报文，而不需要再检查报文的网络层目的地址。

## MPLS 标签结构

紫光集团 H3C  
核心企业 数字化转型方案领导者



### ● MPLS 标签有4个字节，32个bits

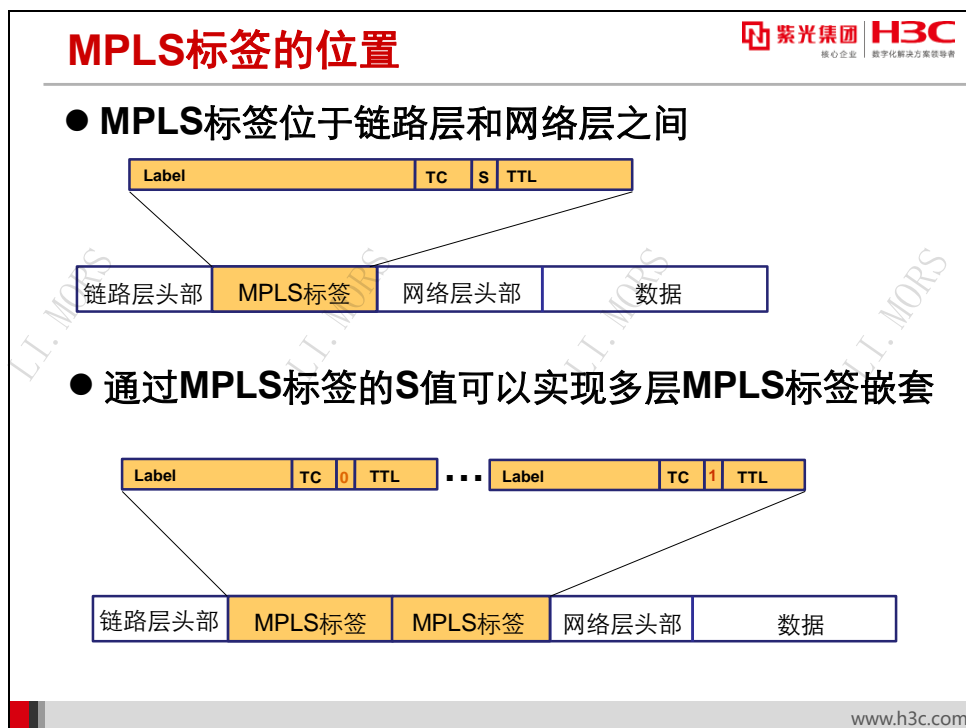
### ● MPLS 标签分为4个区域，含义如下：

- Label：标签值，长度20b，是标签转发表的关键索引；
- TC：用于QoS，长度3b，作用与Ethernet 802.1p值相似；
- S：栈底标识，长度1b，如果有多个Label时，在栈底的Label的S位置为“1”，其他为“0”，只有一个Label时S位置为“1”；
- TTL：存活时间，长度8b，与IP报文的TTL值相似，这个值从IP报文头的TTL域拷贝过来，每进行一次Label交换时，外层Label的TTL值就减“1”。

www.h3c.com

MPLS 标签的结构如上图所示，每个 MPLS 标签有 32 个 bits，分成四个区域，每个区域都有其独特的含义与作用。

- **Label:** 标签值区域，长度 20bits，是 MPLS 标签的核心内容，标签转发就是指根据 MPLS 标签的标签值查找标签转发表进行转发，它是标签转发表的关键索引；
- **TC:** 该区域用于标识报文 QoS 优先级，长度 3bits，该字段又称为 Exp 字段。作用与 Ethernet 802.1P 值或是 IP 包的 DSCP 值类似；
- **S:** 是 MPLS 标签的栈底标识，长度只有 1bit，当一个报文存在多个 MPLS 标签时，用它来标识紧接在该 MPLS 标签后面的是另一个 MPLS 标签还是 MPLS 载荷报文。当 S 位置为“1”时，标识已经是最后一个 MPLS 标签，紧接其后的是 MPLS 载荷报文；而相反当 S 位置为“0”时，表示该 MPLS 标签后面还有下一层 MPLS 标签。S 位使得 MPLS 标签可以实现多层嵌套，这也是 MPLS 技术后来被应用于隧道、VPN 等技术的一个重要基础。
- **TTL:** 存活时间，长度为 8bits，与 IP 报文的 TTL 值相似。TTL 值在报文进入 MPLS 网络时从报文的 IP 头的 TTL 域拷贝出来，每经过一台 LSR，外层 Label 的 TTL 值就减“1”。目的也与 IP 报文的 TTL 值相似，为了防止报文因为环路长期在网络里循环转发，浪费网络资源。



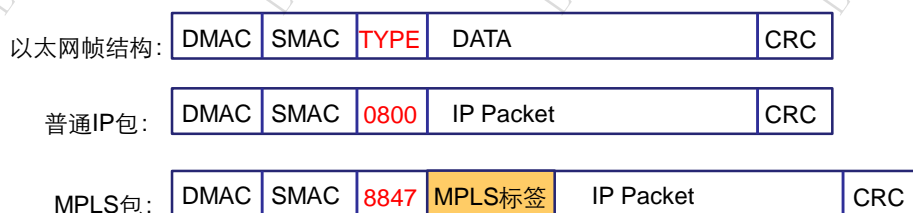
MPLS 标签位于报文的链路层头和网络层头之间。上图是以 IP 报文为例，MPLS 标签位于报文的链路层头部以后，IP 头部之前。

通过 MPLS 标签头部栈底标识的设计，MPLS 标签头可以实现多层嵌套。不同产品能够支持的嵌套层数有一定的限制，目前为止，常见的 MPLS 应用中，最多采用到三层 MPLS 标签嵌套使用。

## MPLS标签的识别

- **MPLS**协议通过在报文的链路层报文头中进行标识，使得路由器能够识别**MPLS**报文。

→ 如在以太网中：指示上层协议的TYPE字段使用0x8847来表示承载的是MPLS报文（0800是IP报文）



www.h3c.com

报文抵达路由器，设备解析完报文的链路层头部以后，需要区分出内部载荷是 **MPLS** 标签包还是普通的网络层封装包，才能正确的处理该报文。

在各种链路层协议中，都有一个标识位，指明报文网络层所采用的协议类型。下表所示就是常见的链路层协议为 **MPLS** 分配的标识：

二层封装协议	协议标识名称	值
PPP	PPP Protocol field	0x0281
Ethernet/802.3 LLC/SNAP	Ether type value	0x8847
HDLC	Protocol	0x8847
Frame Relay	NLPID(Network Level Protocol	0x80

如上图所示的以太网帧，当路由器接收到该报文，解析链路层头部发现 **Ether type** 值为 **0x8847** 时，就可以判断出该报文是一个 **MPLS** 报文，那么紧随链路头部之后的便是 **MPLS** 标签头。

## 15.4.2 MPLS 标签分配协议分类

## 标签分配协议的分类

 紫光集团   
核心企业 数字化转型方案领导者

- 标签分配协议，用于在LSR之间分配标签，建立LSP
  - LDP
  - CR-LDP
  - RSVP-TE
  - MP-BGP
  - MP-BGP (BGP4+)
- **LDP Label Distribution Protocol**标签分配协议，因为它实现简单可靠，逐渐成为MPLS网络中应用最为广泛的标签分配协议之一

www.h3c.com

MPLS 协议实现的重点是利用标签进行数据转发。IP 转发时报文是根据路由表进行转发的，而 MPLS 转发时报文根据 MPLS 标签转发表进行转发。路由表是由各种路由协议根据一定的路由算法计算出来的，指示抵达各个目网段的最优路径。标签转发表是由标签分配协议根据一定的规则生成的，而这些规则通常都有一个重要的特点，那就是它们依赖 IP 路由协议的计算结果，即路由转发表。这也正是 MPLS 标签转发与 ATM 转发实现的一个主要区别，这样的实现解决了 ATM 技术实现重要阻碍，也就是控制信令实现复杂问题。

标签分配的协议有很多种，目前应用较为广泛的有如下几种：

- **LDP (Label Distribution Protocol)**：标签分发协议，他是最为通用的标签分配协议之一。
- **CR-LDP (Constraint-Based Label Distribution Protocol)**：基于路由受限的标签分发协议，它对 LDP 进行了扩展，根据明确的路由约束、服务质量 (QoS) 约束及其它约束，建立一个 LSP，主要用于流量工程技术。
- **RSVP-TE (Resource Reservation Protocol-Traffic Engineering)**：基于流量工程扩展的资源预留协议，它是 RSVP 协议的一个补充协议，用于流量工程技术中进行 MPLS 标签分配。
- **MP-BGP (Multiprotocol BGP)**：多协议扩展 BGP 协议，该协议是对 BGP 协议的扩展，扩展的功能之一就是为 BGP 路由分配 MPLS 标签。

在这些标签分配协议当中，LDP 协议应用最为广泛，被较多厂家的路由器作为缺省的标签分配协议使用。LDP 协议的特点在于简单可靠，下文就以 LDP 协议为例，详细讲述标签分配协议的原理及其建立标签转发表的具体过程。

### 15.4.3 LDP 消息类型

## LDP消息类型

紫光集团 H3C  
核心企业 数字化转型领航者

- 在LDP协议中，存在4种类型的LDP消息：
  - 发现消息（Discovery messages）  
用于LDP邻居的发现和维持。
  - 会话消息（Session messages）  
用于LDP邻居会话的建立、维持和中止。
  - 通告消息（Advertisement messages）  
用于LDP实体向LDP邻居宣告Label、地址等信息。
  - 通知消息（Notification messages）  
用于向LDP邻居通知事件或者错误。

www.h3c.com

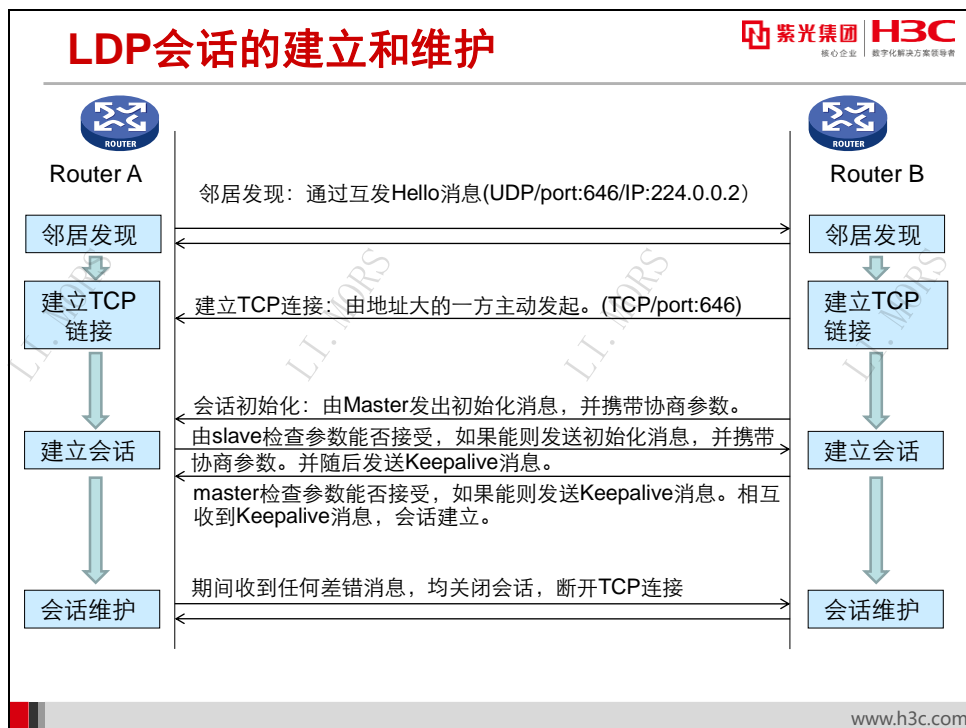
LDP 协议定义了以下四类消息：

- 发现消息（**Discovery messages**）：用于 LDP 邻居的发现和维持。
- 会话消息（**Session messages**）：用于 LDP 邻居会话的建立、维持和中止。
- 通告消息（**Advertisement messages**）：用于 LSR 向 LDP 邻居宣告 Label、地址等信息。
- 通知消息（**Notification messages**）：用于向 LDP 邻居通知事件或者错误。

所有的 LDP 消息都采用 TLV 结构，具有很强的扩展性。

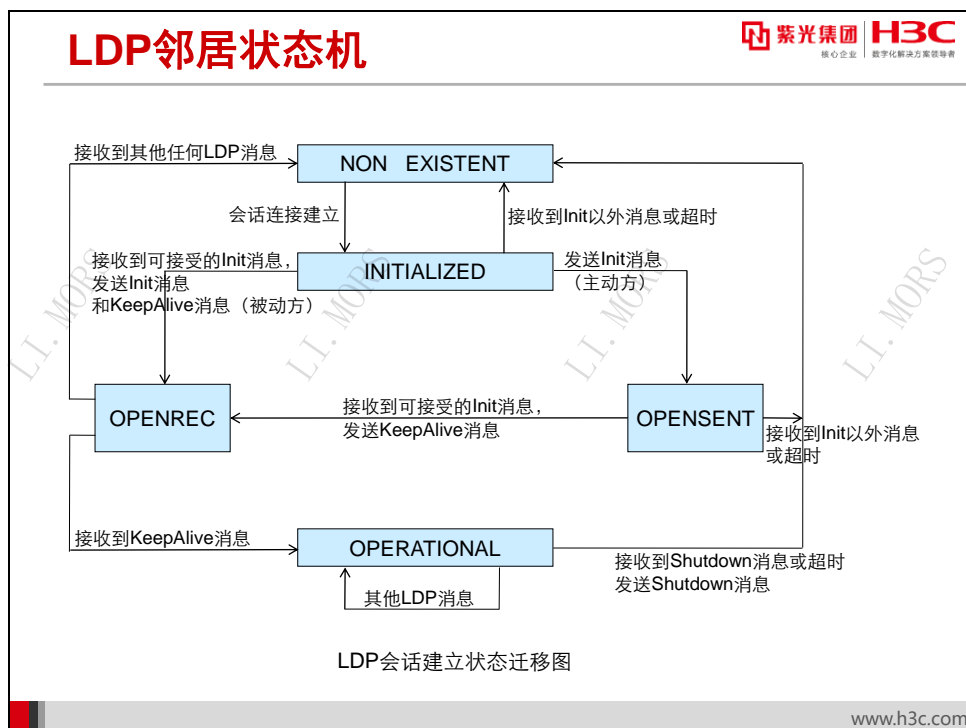
#### 注意：

具体的 LDP 消息有很多种，包括 Notification、Hello、Initialization、KeepAlive、Address、Address Withdraw、Label Mapping、Label Request、Label Abort Request、Label Withdraw、Label Release 等，这些消息都可以分别归入上述四类之中。



上图是 LDP 协议的会话建立和维护的过程,LSR 首先定期发送 Hello 消息发现其它的 LSR。如果两台 LSR 的 Hello 消息相关参数匹配,两 LSR 就会建立 TCP 会话,然后在 TCP 连接中交换 Initialization 消息协商 LDP 参数。LDP 参数协商成功以后完成 LDP Session 的建立,Session 建立后 LSR 要定期发送 Keep-alive 消息维持 Session。

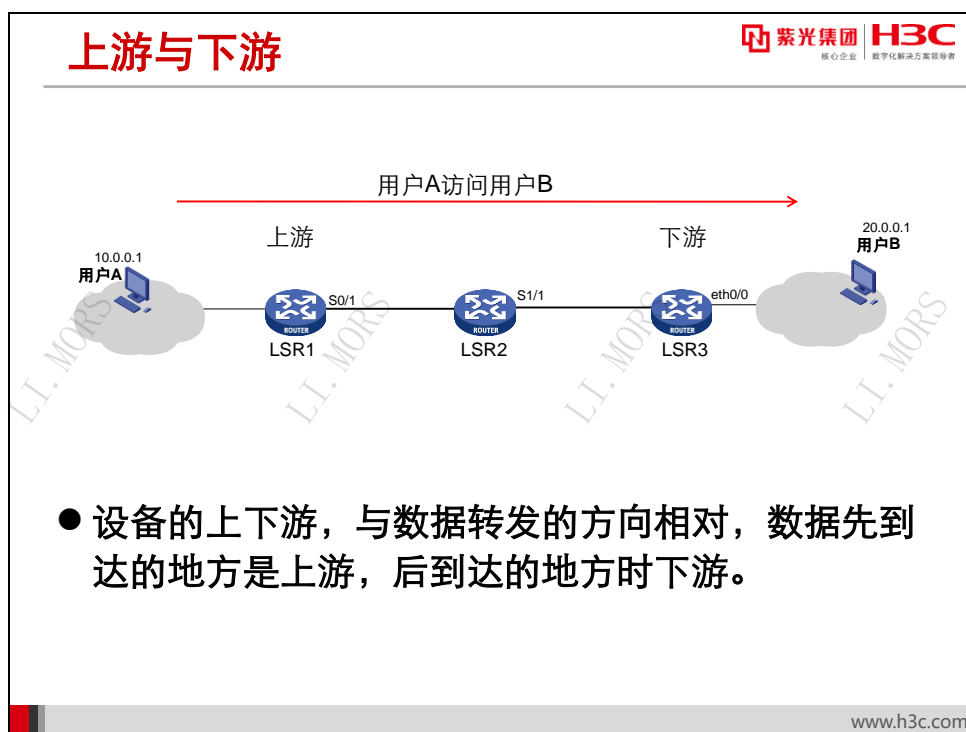
建立好 LDP 的 Session 以后,LSR 之间开始互相发送一个或者多个 Label mapping 消息,LSR 接收到邻居发来的 Label mapping 消息,再根据自身的路由状况,形成标签转发表项。



上图是 LDP 协议在会话建立和维护的过程中，LSR 设备上 LDP 状态机的转化过程。

两台 LDP 邻居之间建立起 LDP Session 后，状态会维持在 Operational。

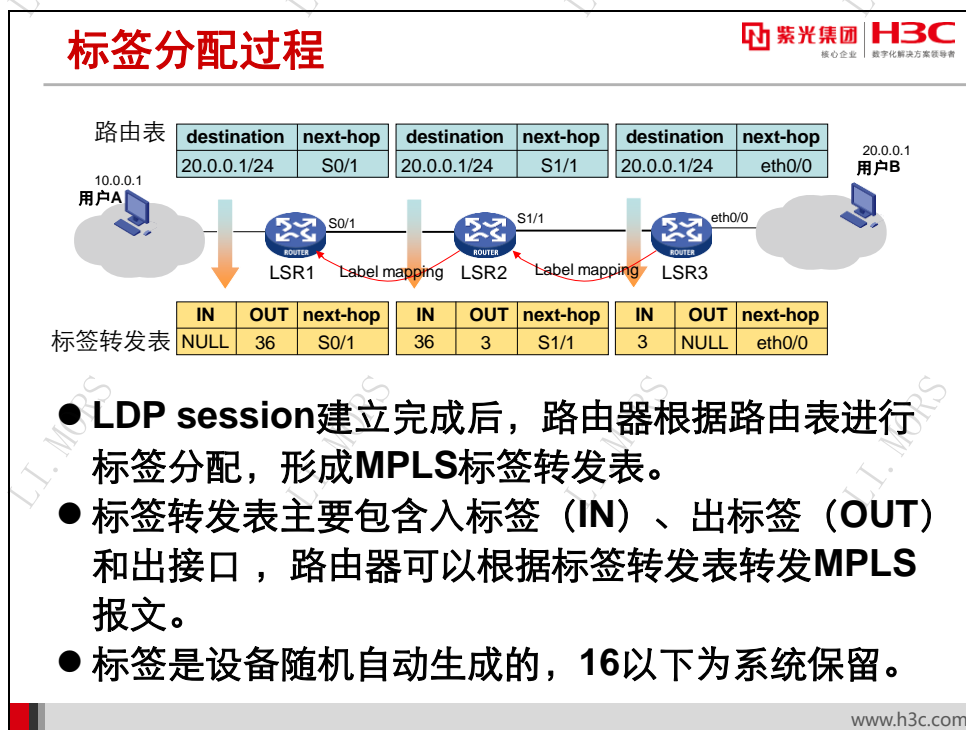
#### 15.4.4 标签分配过程





首先，在 MPLS 网络中，根据数据报文的传输方向，定义了 LSR 设备上下游概念。以上图为例，用户 A 要访问用户 B，报文会依次抵达 LSR1、LSR2、LSR3，那么 LSR3 就是 LSR2 的下游设备，LSR2 是 LSR1 的下游设备。LSR 的上游和下游是根据报文传输的方向来判断的，报文先抵达的 LSR 是上游 LSR，而后抵达的 LSR 是下游 LSR。因此，如果针对某一条路由来看，就以报文抵达该路由的目的网段的方向来判断设备的上下游。所以，第一个发现这条路由的 LSR 将是最下游 LSR。比如针对 20.0.0.1/24 这条路由，LSR3 与该网段直连，最先发现这条路由，它是最下游 LSR。理解上游和下游的概念是理解下文 LDP 标签分配过程的基础。

下文就已一个实际的案例来讲解 LDP 协议完成标签分配的过程。



如上图所示，用户 A 与用户 B 之间存在一个普通的 IP 网络，三台路由器 LSR1、LSR2、LSR3 之间运行某种路由协议。针对用户 B 所在网段，三台路由器都学习到这条路由，形成上图所示的路由转发表，进而用户 A 可以顺利访问用户 B。

在三台路由器之间运行 MPLS LDP 协议，LSR1 和 LSR2、LSR2 和 LSR3 之间建立 LDP 邻居关系。LDP 邻居建立完成以后，LDP 协议将按照用户的定义为每个 FEC 分配 MPLS 标签。常见的应用通常按照路由来划分 FEC，即所有匹配某一路由表项的报文属于同一个 FEC。在这种 FEC 划分方式下，LSR 就为每个路由转发表项分配 MPLS 标签，继而形成标签转发表。如上图所示，LSR3 设备为 20.0.0.1/24 这条路由表项分配了一个 MPLS 标签“3”，并将为 20.0.0.1/24 该路由分配了 MPLS 标签值“3”这条信息通过 Label mapping 消息发布给 LSR2。各个 LSR 设备间通过 Label mapping 消息的交互，最终在各台 LSR 设备上形成了如上图所示的标签转发表。

MPLS 的标签转发表包含入（IN）标签、出（OUT）标签和出接口三个主要部分：

- **入标签 (IN Label):** 本 LSR 为某一 FEC 分配的 MPLS 标签。报文抵达 LSR 后, LSR 将报文所携带的 MPLS 标签值与 MPLS 标签转发表项的入标签进行对比, 匹配到相同的值时, 就按照此表项转发该报文。与此同时, LSR 会将为某一 FEC 分配的入标签信息通过 Label mapping 消息通知给上游 LSR。如上图, LSR2 为 20.0.0.1/24 这条路由分配了 MPLS 标签 “36”, 对应 LSR2 标签转发表项的入标签值, 同时 LSR2 会将为 20.0.0.1/24 这条路由分配了 MPLS 标签 “36” 这条信息通过 Label mapping 消息通知给 LSR1。当然, 当该 FEC 在此 LSR 已经没有上游设备时, 其入标签为 “NULL”, 图中 LSR1 对应 20.0.0.1/24 这条路由的入标签为 “NULL”。
- **出标签 (OUT Label):** 下游的 LSR 为某一 FEC 分配的标签, 通过 Label mapping 消息发送到本 LSR, 在本 LSR 上记录为该 FEC 的出标签。LSR 在转发 MPLS 报文时, 将报文携带的标签值修改成对应 MPLS 标签转发表项的出标签值。如上图, LSR3 为 20.0.0.1/24 这条路由分配了标签值 “3”, 通过 Label mapping 消息发布给 LSR2, 对应 LSR2 标签转发表项的出标签值。当然, 当该 FEC 在此 LSR 已经是最下游设备时, 出标签为 “NULL”。如 LSR3 上对应 20.0.0.1/24 这条路由的出标签为 “NULL”。
- **出接口:** 某一标签转发表项的出接口就指向该 FEC 的下游一台设备。如 LSR2 为 20.0.0.1/24 这条路由生成的标签转发表项的出接口与 20.0.0.1/24 这条路由表项的出接口相同, 为 S1/1 接口。

MPLS 标签只有本地意义, 每台 LSR 设备都对自己分配的标签即入标签负责, 确保自己为不同的 FEC 所分配的入标签不会相同, 以确保属于不同 FEC 的报文抵达该设备后, 匹配到唯一的与此 FEC 对应的标签转发表项进行转发。所以在 LSR 上的标签转发表项的 IN 标签列均不会相同。相反, 某一 LSR 上标签转发表项的 OUT 标签值可能是来源与不同的下游 LSR 为不同的 FEC 分配的标签, 它们之间并无关联性, 也没有意义比较是否相同。

任何一台 LSR 设备只会将针对某一 FEC 的 Label mapping 消息发送给该 FEC 的上游设备, 而绝不会发送给下游设备。这样一来就可以确保只要路由表没有环路, MPLS 的标签转发表项就也不会有环路。

## 15.4.5 标签分配和管理方式

## 标签分配和管理

紫光集团 H3C  
核心企业 数字化转型方案领导者

- 标签通告模式
  - DOD (downstream-on-demand, 下游按需方式)
  - DU (downstream unsolicited, 下游自主方式)
- 标签控制模式
  - 有序方式 (Ordered)
  - 独立方式 (Independent)
- 标签保持方式
  - 保守模式 (Conservative)
  - 自由模式 (Liberal)

www.h3c.com

在标签分配的过程中，存在很多种方式，这些方式一般情况下可以在设备上配置，各种不同的方式适合不同的 MPLS 应用。

标签通告模式包括：

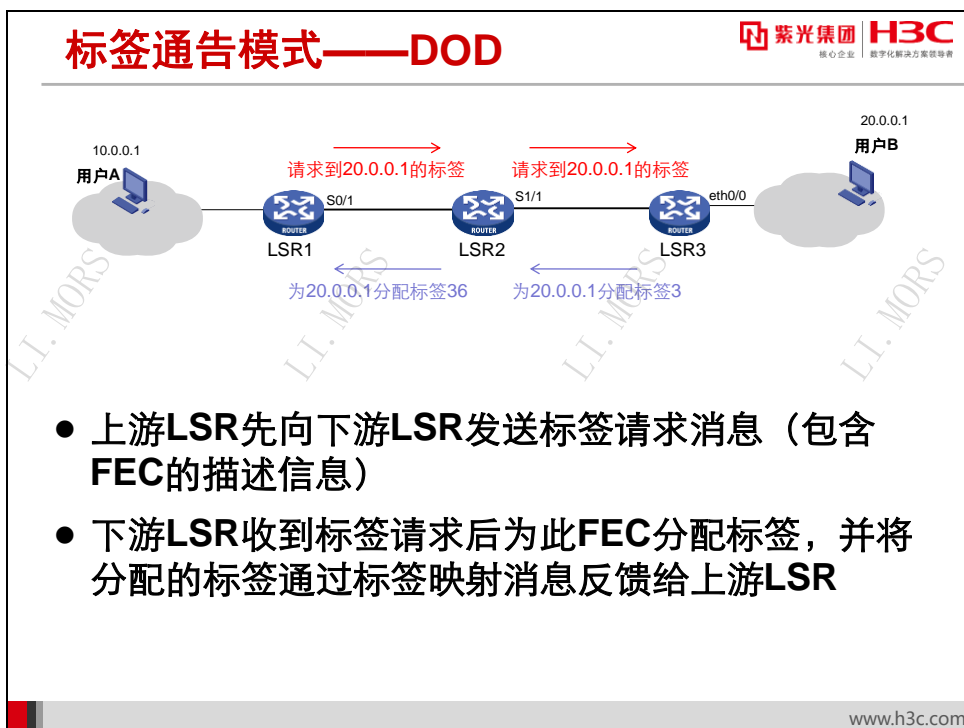
- DOD: downstream-on-demand 下游按需方式
- DU: downstream unsolicited 下游自主方式

标签控制模式包括：

- 有序方式 (Ordered)
- 独立方式 (Independent)

标签保持模式包括：

- 保守模式 (Conservative)
- 自由模式 (Liberal)



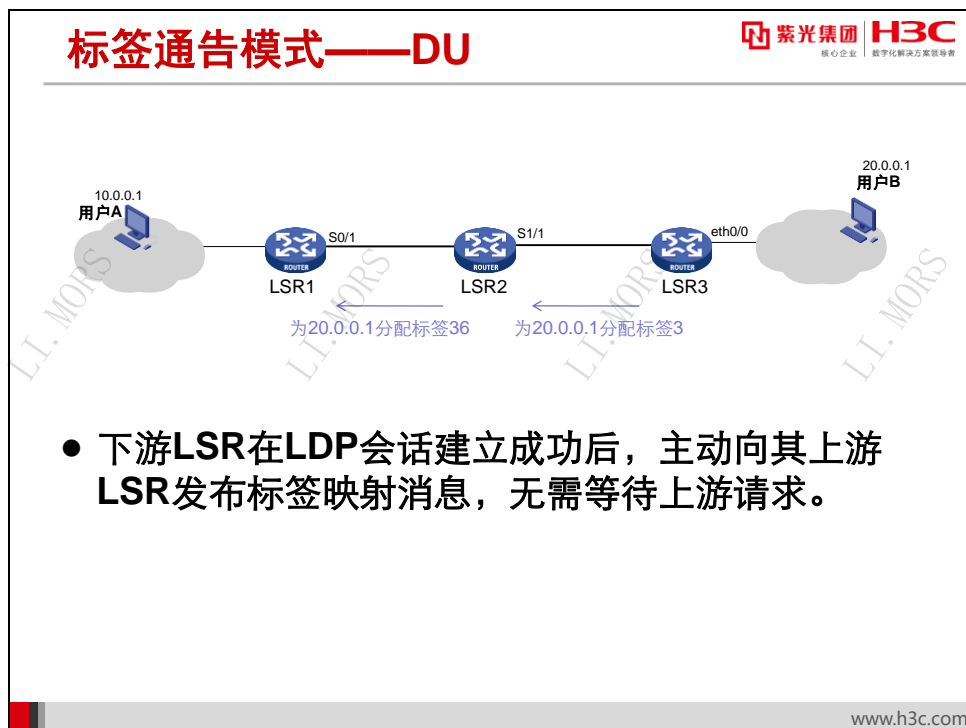
标签通告模式是指 LSR 设备何时可以分发标签所要遵循的原则。它分为两种模式，第一种叫做 DOD 模式，即下游按需标签通告方式，基本原则是下游设备需要等到上游设备的标签分配申请才可以分配标签。

DOD 方式通告标签的具体实现方式如下：

**第1步：**上游的 LSR 先向下游的 LSR 发送标签请求的消息，该消息主要包含了上游 LSR 需要下游 LSR 针对哪个 FEC 分配标签的具体要求。

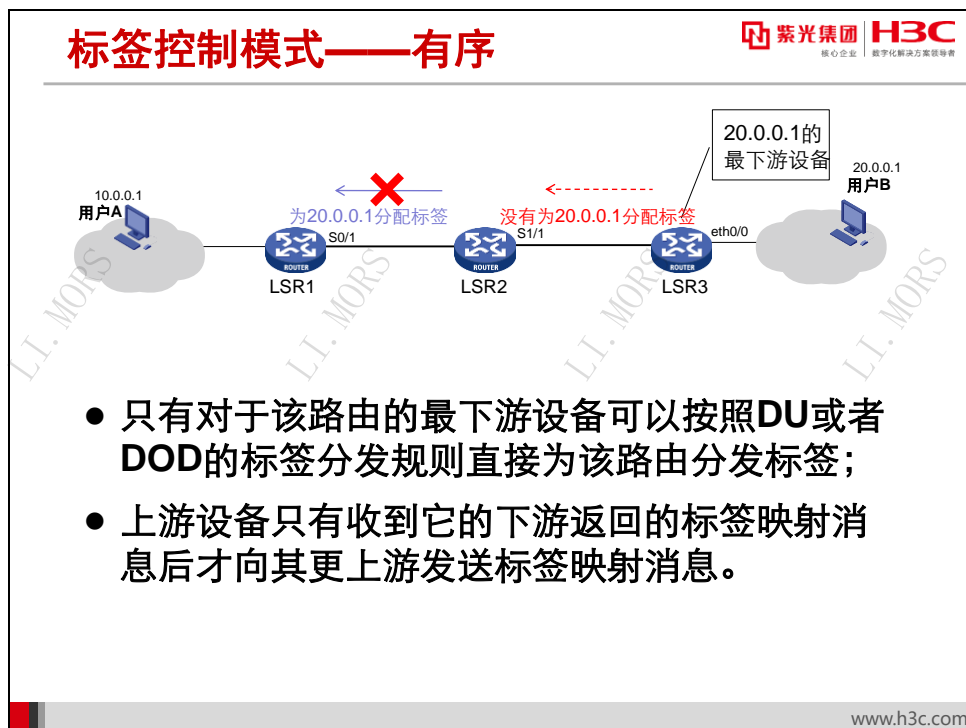
**第2步：**下游 LSR 收到标签请求后将为请求消息里要求的 FEC 分配标签，然后通过 Label mapping 消息发送给上游的 LSR，再形成标签转发表项。

DOD 的标签通告方式适用于那些需要由上游设备来动态决定 FEC 的划分方法，或者需要由上游设备来为各个 FEC 指定网络资源要求，等等类似的应用环境。DOD 标签分发模式目前主要应用在流量工程技术中。



另一种标签通告的模式叫做 DU 方式，即下游自主标签通告模式，这种模式相对 DOD 模式较为简单，下游设备不需要等待上游 LSR 的标签分配申请，主动将其为各个 FEC 分配标签的情况通过 Label mapping 消息发送给上游 LSR，并形成标签转发表项。

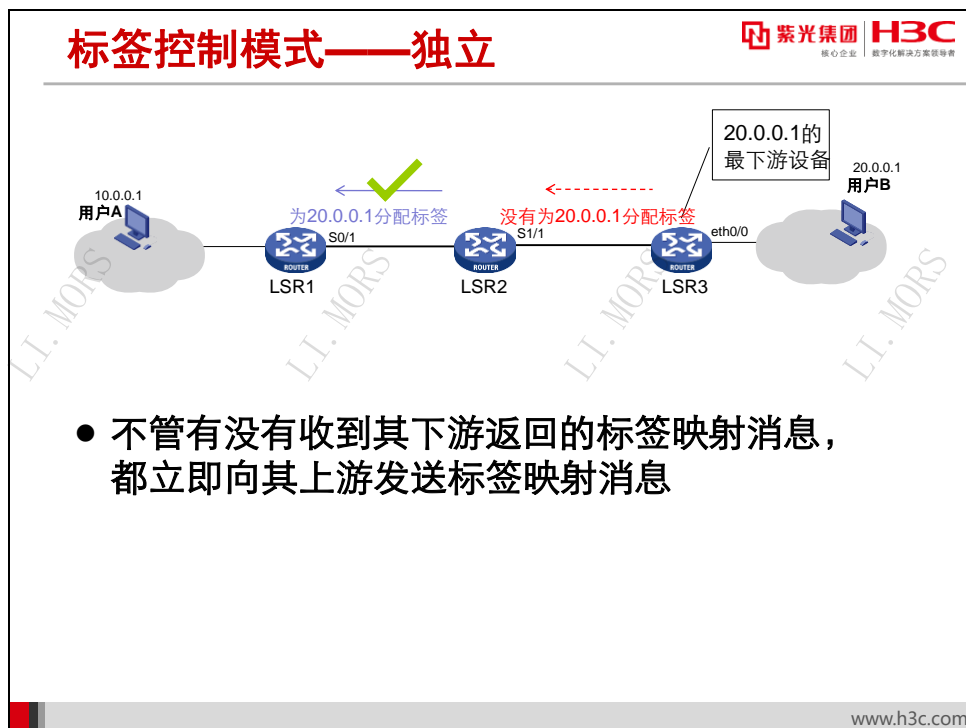
当前，DU 标签分配方式应用更为普遍。



标签的控制模式是指 LSR 设备可以为哪些 FEC 分配标签的原则。它分为有序和独立两种模式，其中有序的模式是指，LSR 只可以主动的为那些自己是最下游 LSR 的 FEC 分配标签。以上图为例，对于 20.0.0.1/24 这条路由，LSR2 不是最下游设备，就不能主动的为其分配标签。而 LSR3 设备是 20.0.0.1/24 这条路由的最下游设备，它才可以主动的为 20.0.0.1/24 分配标签。

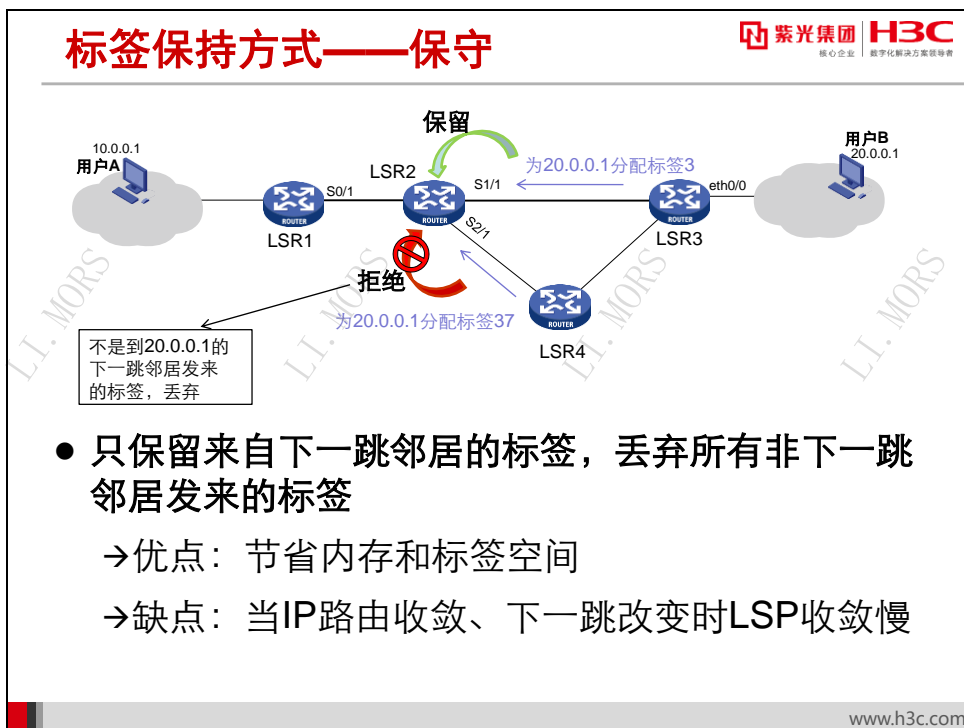
在有序的标签控制模式下，上游的 LSR 需要等待收到它下游 LSR 发送过来的 Label mapping 消息，才能为对应 FEC 分配标签，然后再发送给更上游的 LSR。可见，在这种标签控制模式下，针对某 FEC 的标签转发表将从其最下游的 LSR 立起，依次往上游逐台 LSR 形成该 FEC 的标签转发表项。

有序的标签分配模式使得 MPLS 的转发是端到端的，对那些不是直接接入 MPLS 网络的用户，将无法进行 MPLS 转发。



另一种标签控制的模式叫着独立模式，独立的标签控制模式相对有序控制模式的实现较为简单，LSR 不管自己是不是该 FEC 的最下游 LSR，也不管有没有收到该 FEC 的下游 LSR 发布过来的 Label mapping 消息，都可以直接为该 FEC 分配标签，并向上游设备发布 Label mapping 消息。如上图所示，LSR2 可以不用等待 LSR3 为 20.0.0.1/24 这条路由分配的标签消息，直接为其分配标签。

独立的标签控制模式，使得任何一个数据流在经过 MPLS 网络时都可以进行 MPLS 转发，其最终的目的可能在一个非 MPLS 网络里。

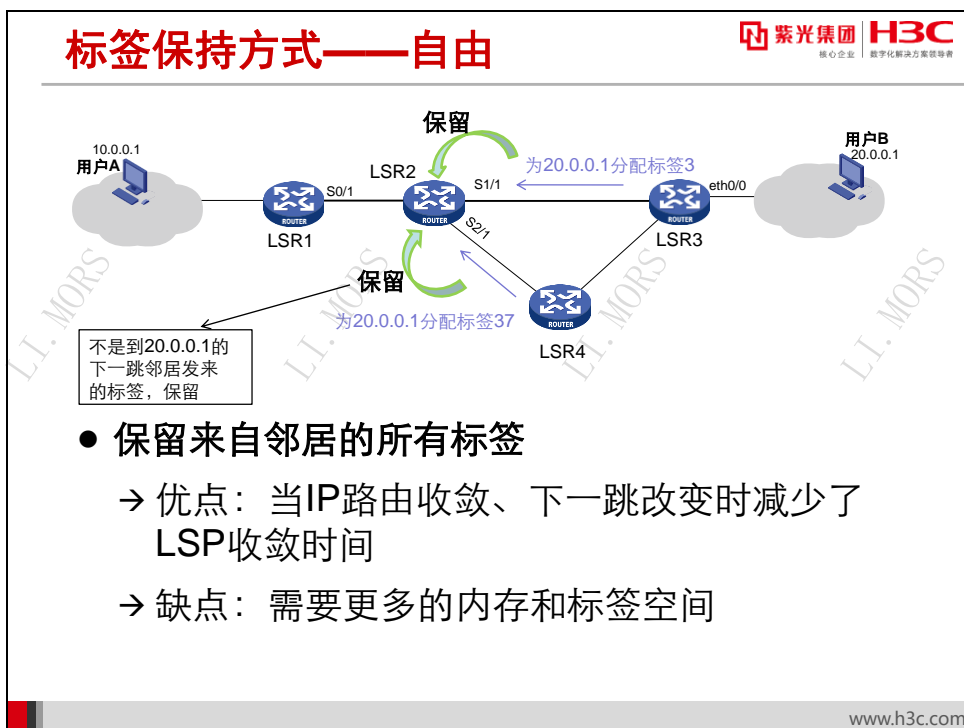


标签保持的方式是指 LSR 收到下游 LSR 的 Label mapping 消息以后，是否记录 Label mapping 消息所携带的标签信息的原则。标签保留的方式也分为两种，保守方式和自由方式。其中保守方式（Conservative retention mode）是指，LSR 只保留来自该 FEC 下一跳的 LSR 邻居发送过来的标签消息，而对于其他 LSR 邻居发送过来的标签消息不作记录。

如上图所示，在 LSR2 上，20.0.0.1/24 这条路由的下一跳是 LSR3，但是对于 LSR3 和 LSR4 来讲，LSR2 都是 20.0.0.1/24 这条路由的上游 LSR，它们都会向 LSR2 发送针对 20.0.0.1/24 这条路由的 Label mapping 消息。在保守的标签保持方式下，LSR2 只会将 LSR3 发布给它的针对 20.0.0.1/24 这条路由的标签记录下来，而将 LSR4 发布给它的针对 20.0.0.1/24 这条路由的标签直接丢弃不作记录。

保守的标签分配方式的优点是，可以节省 LSR 设备的内存和标签空间。但是相反，当网络发生故障，下一跳发生变化时，LSP 的收敛比较慢。如上图，当 LSR2 和 LSR3 之间的链路中断，到达 20.0.0.1/24 的路径切换至 LSR4 上时，LSR2 上因为没有保留 LSR4 发布给它的针对 20.0.0.1/24 这条路由的标签信息，需要等待 LSR4 重新发布标签分配协议周期性的通告消息，才能重新建立起 LSP，收敛比较慢。



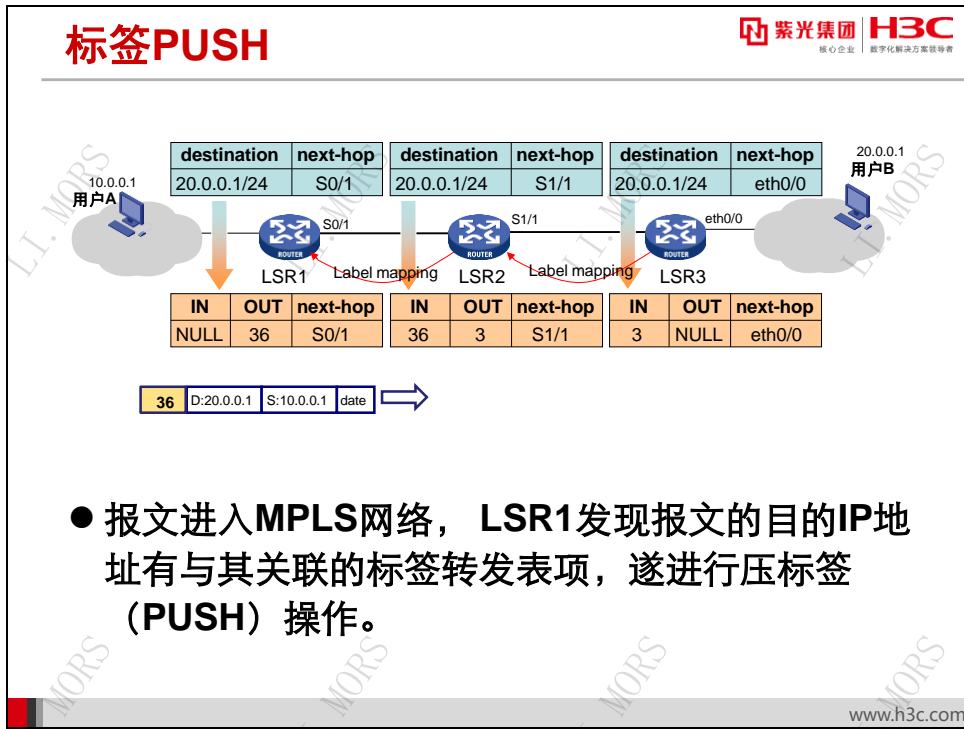


标签保持的另一种方式叫自由方式（Liberal retention mode），自由方式是指无论是否是该 FEC 的下一跳 LSR 发布过来的标签，LSR 都予以记录。

如上图所示，LSR2 将 LSR4 发送过来的针对 20.0.0.1/24 这条路由的标签也记录下来，这种标签保持的方式就是自由方式。自由方式的优缺点与保守方式刚好相反，它在网络发生故障，路由发生切换时，LSP 收敛较快，但是它需要占用更多的内存和标签空间。

标签的分配、控制、保持等方式在实际的应用中根据用户的需要可进行任意组合，通常情况下，LDP 协议缺省运行在 DU+有序+自由的方式下，这种组合也是 MPLS 在具体应用中最常用到的一种方式。

## 15.5 MPLS转发实现

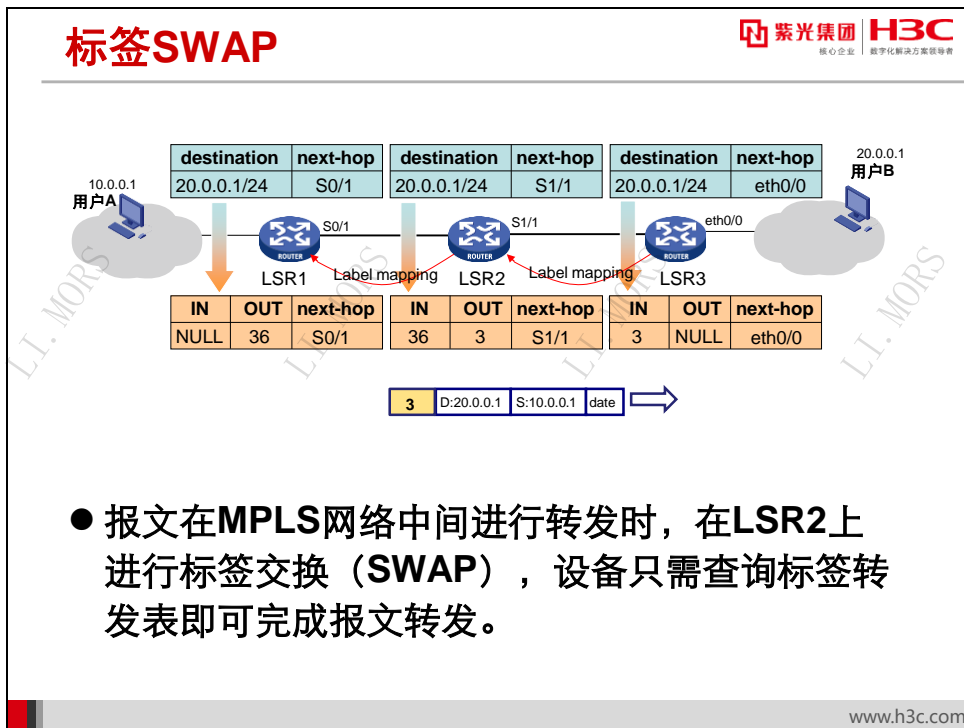


- 报文进入MPLS网络，LSR1发现报文的目的IP地址有与其关联的标签转发表项，遂进行压标签（PUSH）操作。

LSR 上建立起 MPLS 标签转发表以后，当报文进入 MPLS 网络时，就可以进行 MPLS 转发。报文进行 MPLS 转发的过程分为三个不同的阶段，第一阶段是报文从非 MPLS 网络进入 MPLS 网络时，在 ingress LSR 设备上上进行压标签动作，也被称之为标签 PUSH。

如上图所示，用户 A 访问用户 B，报文从 LSR1 进入 MPLS 网络。抵达 LSR1 设备时该报文还是一个普通的 IP 包，LSR1 按照普通的 IP 转发流程，首先检查 IP 路由表，找到 20.0.0.1/24 这条路由表项，此时 LSR1 发现对应该路由表项有一个与此关联的标签转发表，于是 LSR1 将对该报文启动 MPLS 转发。在 LSR1 上要完成的就是 MPLS 压标签操作，也就是给这个报文加上一个 MPLS 头，MPLS 头内的标签值就是对应的标签转发表项的 OUT 标签值“36”。完成压标签操作后，LSR1 再将报文按照标签转发表将报文从相应的出接口发出。

在 ingress LSR 设备上进行了压标签操作后，该报文就转变成一个 MPLS 报文，接下来在 MPLS 网络的转发过程中，就按照报文的 MPLS 头进行转发，LSR 设备无需再检查该报文的 IP 头，也就进入了 MPLS 转发的第二个阶段。



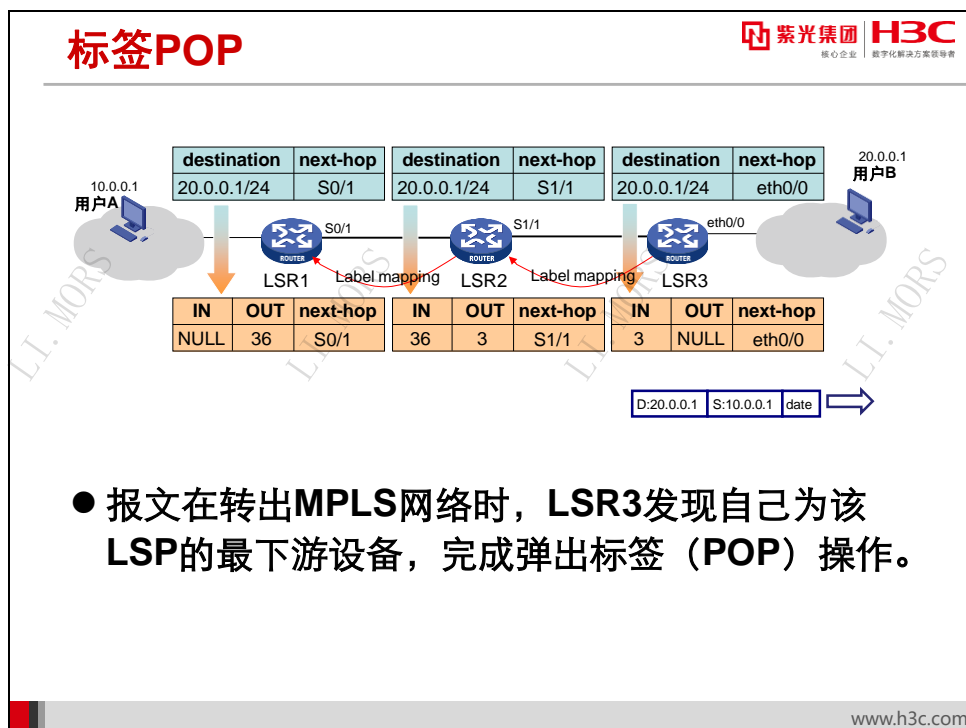
MPLS 转发的第二个阶段是标签交换，也被称之为标签 SWAP。报文在 MPLS 网络内部的所有 LSR 设备上，都按照这个阶段的操作方式进行转发。

如上图所示，报文带着在 LSR1 上压好的标签“36”进入 LSR2，LSR2 发现该报文并非一个普通的 IP 包，而是一个 MPLS 报文，于是进行 MPLS 转发。根据报文所携带的 MPLS 标签值“36”，检查本地的标签转发表，找到 IN 标签值等于“36”的那一个表项，并根据这一表项进行报文转发。

LSR2 成功的找到对应的标签转发表项后，首先将报文的 MPLS 头部所携带的标签值转换成该转发表项的 OUT 标签值，即将“36”改成“3”，然后将报文按该表项从相应的出接口发出，也就是从“S1/1”接口发出。

可以看出，LSR2 无需查看报文的 IP 头，或者查询 IP 路由表，直接按照报文的 MPLS 标签检查标签转发表即可完成转发。实际的 MPLS 网络中，通常报文会经过多跳的 LSR 设备，在所有的 LSR 设备上，报文都是进行类似的标签交换操作就能完成转发。

因为根据 MPLS 标签转发表建立的实现原理，MPLS 标签转发表中只会有唯一的一个标签转发表项，其入标签值与抵达的 MPLS 报文所携带的标签值相同。所以在 MPLS 转发中只需一次查表就能完成对该报文的转发，相对 IP 转发多次查表已达到最优匹配的转发方式，效率要高很多。



MPLS 转发的最后一个阶段是标签弹出，也被称之为标签 POP。这个阶段是报文最终离开 MPLS 网络时需要进行的操作，该操作也最为简单。当报文到达离开 MPLS 网络的 Egress LSR 设备时，首先报文依然与到达普通的 LSR 设备一样，根据标签转发表进行转发，但是对应的标签转发表项的出标签为“NULL”，这就表示该 LER 已经是该 FEC 的最下游设备，从此该报文将离开 MPLS 网络。于是该 Egress LSR 会直接将报文的 MPLS 头部去除，并按照标签转发表从对应的出接口将报文发出。可见此时发出的报文已经恢复成一个普通的 IP 包，在接下来的非 MPLS 网络中可以按照原先的法式正常的进行转发，这体现出 MPLS 网络可以与非 MPLS 网络平滑的进行兼容过渡。

如上图所示，报文离开 LSR2 后，携带标签值“3”抵达 LSR3 设备，LSR3 检查标签转发表，发现 IN 标签是 3 的标签转发表项显示出标签值为“NULL”，LSR3 就直接将该报文的 MPLS 头部去除，并按照该标签转发表项将报文从“eth0/0”接口发出。报文离开“eth0/0”接口后就按照正常的非 MPLS 网络转发方式成功的抵达用户 B，至此实现了用户 A 到用户 B 的访问。

当然如果考虑用户 B 访问用户 A，就需要针对 10.0.0.1/24 这一条路由在整个 MPLS 网络上首先形成标签转发表项，然后按照从 LSR3 到 LSR1 的方向进行转发，整个过程的实现完全相同，这里也就不做赘述了。

## 15.6 MPLS应用与发展

### MPLS的实际应用



- 随着硬件技术的进步，采用**ASIC**和**NP**进行转发的高速路由器和三层交换机得到广泛应用，使得**IP**转发性能大为提高，可以满足网络数据转发性能需求。**MPLS**技术在提高转发性能应用上未能发挥优势。
- 但**MPLS**支持多层标签嵌套和面向连接的特点，使得其在**VPN**、流量工程（**TE**）、**QoS**等方面得到广泛应用。

www.h3c.com

随着硬件技术的进步，采用 ASIC 和 NP 替代 CPU 进行转发的高速路由器和三层交换机得到了广泛的推广和应用，使得 IP 转发并没有像预期的那样无法满足网络数据转发的性能需求，至今为止 IP 转发仍然是网络数据转发的主流，MPLS 技术在提高转发性能的应用上没有能够真正发挥优势。

尽管如此，MPLS 技术并没有被抛弃，相反因为它设计上的很多优点，如支持多层标签嵌套、可以兼容多种二、三层协议等，它被人们应用在很多隧道和 VPN 技术中，如 BGP MPLS VPN、流量工程（TE）、QoS 等，尤其是其中 VPN 的应用，非常流行。

## 15.7 本章总结

### 本章总结

- MPLS的产生背景与基本概念
- 标签与标签分配协议
- MPLS标签分配与MPLS报文转发过程

www.h3c.com

## 15.8 习题和解答

### 15.8.1 习题

- MPLS 的标签有哪些字段？（ ）  
 A. Label 标签值                      B. TTL 存活时间  
 C. S 栈底标识                        D. TC QoS 优先级
- PPP 协议如何标识其承载的上层报文为 MPLS 报文？（ ）  
 A. 在 PPP 协议 LCP 协商阶段与对端设备协商好  
 B. 在 PPP 协议 NCP 协商阶段与对端设备协商好  
 C. 在报文的 PPP 头部 PPP Protocol field 位置填写 0x0281  
 D. 在报文的 PPP 头部 PPP Protocol field 位置填写 0x8847
- MPLS LDP 协议 Session 建立成功后，其状态维持在什么状态？（ ）  
 A. Full    B. Establish    C. Operational    D. Opensent
- 下图所示，用户 A 访问用户 B 时，哪一台设备是最下游设备？（ ）



- A. LSR1    B. LSR2    C. LSR3    D. 不确定
- 下列关于 MPLS 标签转发表的入标签和出标签的说法错误的是（ ）  
 A. 在某 LSR 的标签转发表里，每一表项的入标签一定各不相同  
 B. 在某 LSR 的标签转发表里，每一表项的入标签有可能相同  
 C. 在某 LSR 的标签转发表里，每一表项的出标签一定各不相同  
 D. 在某 LSR 的标签转发表里，每一表项的出标签有可能相同
- 某一 MPLS 网络选用了 DU+有序+自由的标签通告、控制和保持方式，那么下列说法正确的是（ ）  
 A. 需要等待上游的设备请求，下游的 LSR 才会为对应的 FEC 分配 MPLS 标签  
 B. LSR 如果是某一 FEC 的最下游设备，他可以直接为该 FEC 分配标签  
 C. LSR 需要保留所有 LDP 邻居发送过来的 MPLS 标签

D. LSR 在收到下游设备为某一 FEC 分配的标签后，才会为此 FEC 分配标签

7. MPLS 转发过程中，哪些阶段只需要查询标签转发表，无需查询路由表（ ）

A. PUSH

B. SWAP

C. POP

D. 所有阶段

### 15.8.2 习题答案

1. ABCD

2. C

3. C

4. C

5. BC

6. BCD

7. BC



## 第16章 BGP MPLS VPN 基本原理

BGP MPLS VPN 以其更合理的结构模型，更简单的维护需求，更灵活的业务控制方法，解决了传统 VPN 技术的一系列问题，逐渐成为当今应用最为广泛的 VPN 技术之一。

### 16.1 本章目标

#### 课程目标

● 学习完本课程，您应该能够：

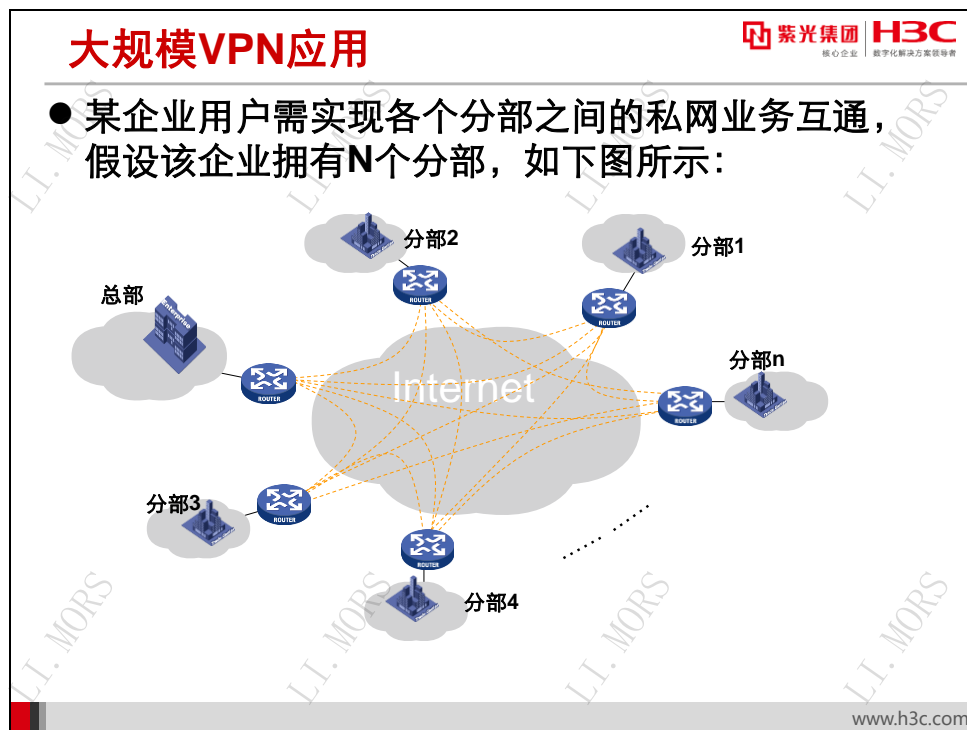
- 了解BGP MPLS VPN技术的产生背景
- 掌握BGP MPLS VPN技术的基本原理



www.h3c.com

## 16.2 BGP MPLS VPN技术背景

### 16.2.1 传统 VPN 的缺陷



当用户的各个部门或分支分散在不同的物理地点时，VPN 技术可以模拟实际的物理线路，将这些分支网络连接起来，实现用户私有网络之间的互通。VPN 技术大大节省了用户建立私有网络的成本，用户的各个分支部门只需要与公共的 Internet 网相连，就可以采用 VPN 技术建立起不同分支之间的模拟通道，在各个分支部门之间实现私网互通。

如上图所示，某企业用户，在不同的地理位置拥有 N 个分部，该用户通过 VPN 技术，建立总部及各个分布之间的隧道，将总部与各个分布相连，实现其各个分部及总部之间的私网互通。

## 传统VPN的缺陷

紫光集团 H3C  
核心企业 数字化转型领导者

- 静态隧道的可扩展性不强。传统VPN技术的隧道需要静态建立，随着用户网络规模的扩大，VPN隧道的数量成N平方增长。
- VPN维护和管理工作只能由用户自行完成。因不同的VPN用户，私网地址可能存在冲突，负责维护和管理公共网络的运营商因为不能区分开用户，无法接管用户的VPN。

因此传统VPN无法适应大规模VPN网络应用，需要一种新型的VPN技术解决上述的问题。

www.h3c.com

VPN 技术能够非常有效的节省用户建立私有网络的成本，成为用户建立私有网络的一个很好的选择，因此各种各样的 VPN 技术也纷纷出炉，如 GRE、IPSec、L2TP 等等。这些 VPN 技术被统称为传统 VPN 技术。随着网络的发展，用户私有网络的规模也逐渐扩大，传统 VPN 技术的一些缺点被暴露出来，最主要的体现为以下两点：

- 静态隧道的可扩展性不强。传统 VPN 技术的隧道需要静态建立，随着用户网络规模的扩大，VPN 隧道的数量成 N 平方增长，用户的分支部门的增减都将涉及到大量的静态隧道的配置或删除。
- VPN 只能由用户自行维护和管理。在公共的 Internet 网络上承载着很多的 VPN 用户，而各个 VPN 用户的私网地址空间通常是重叠的，VPN 的维护和管理工作只能由用户自行完成。负责维护和管理公共网络的运营商因为不能区分开用户，无法接管用户的 VPN。

这两个缺点，使得用户不可能选择采用传统的 VPN 技术建立大规模的私有网络。

## 16.2.2 BGP MPLS VPN 的优点

## BGP MPLS VPN的优点

紫光集团 H3C  
核心企业 数字化转型领导者

- **BGP MPLS VPN技术作为一种新的VPN技术，在传统VPN技术基础上解决了以下三个重大问题：**
  - 实现隧道的动态建立
  - 解决本地地址冲突问题
  - VPN私网路由易于控制

www.h3c.com

BGP MPLS VPN 技术是通过 BGP 和 MPLS 两种技术配合实现的一种新型的 VPN 技术，与传统 VPN 技术相比，它有如下的三个优点：

- 实现隧道的动态建立。传统 VPN 技术用户各个分部之间的 VPN 隧道需要由维护人员手工静态配置，而 BGP MPLS VPN 技术隧道是动态建立的。用户的各个分部之间会自动建立隧道，且分部的增加或减少，不需要用户再去添加或删除隧道配置。
- 解决了本地地址冲突问题。BGP MPLS VPN 技术实现了一台路由器可以同时处理多个不同的 VPN 用户数据的功能，最终使得多个地址冲突的用户都可以将建立和维护 VPN 的工作交给运营商，进一步降低 VPN 用户的负担。
- VPN 私网路由易于控制。私网路由的交互是 VPN 技术实现的关键，能够动态的交互私网路由，用户网络才能自动发现各个分部网络的所在位置。传统的 VPN 技术中可以支持私网路由的动态学习，然而 BGP MPLS VPN 技术在私网路由的动态交互的基础上，加入了互通或隔离的控制，可以更加灵活的控制用户各个分部，或者各个不同用户之间的互访关系。

## 16.3 MPLS隧道

### 16.3.1 隧道技术与 MPLS

### 隧道技术与MPLS

紫光集团 H3C  
核心企业 数字化解决方案领导者

- **隧道**：一个虚拟的点对点的连接。它提供了一条虚拟通路，使经过特殊封装的数据报能够在这个通路上传输。在隧道的两端分别对数据报进行封装及解封装。如**GRE**封装，隧道上的路由器根据报文外层的公网IP头进行数据转发。

GRE封装：

公网IP头部	GRE	私网IP头部	数据
--------	-----	--------	----

- **MPLS**是天然的隧道，隧道上的路由器可以根据报文的**MPLS**头进行报文转发：

MPLS封装：

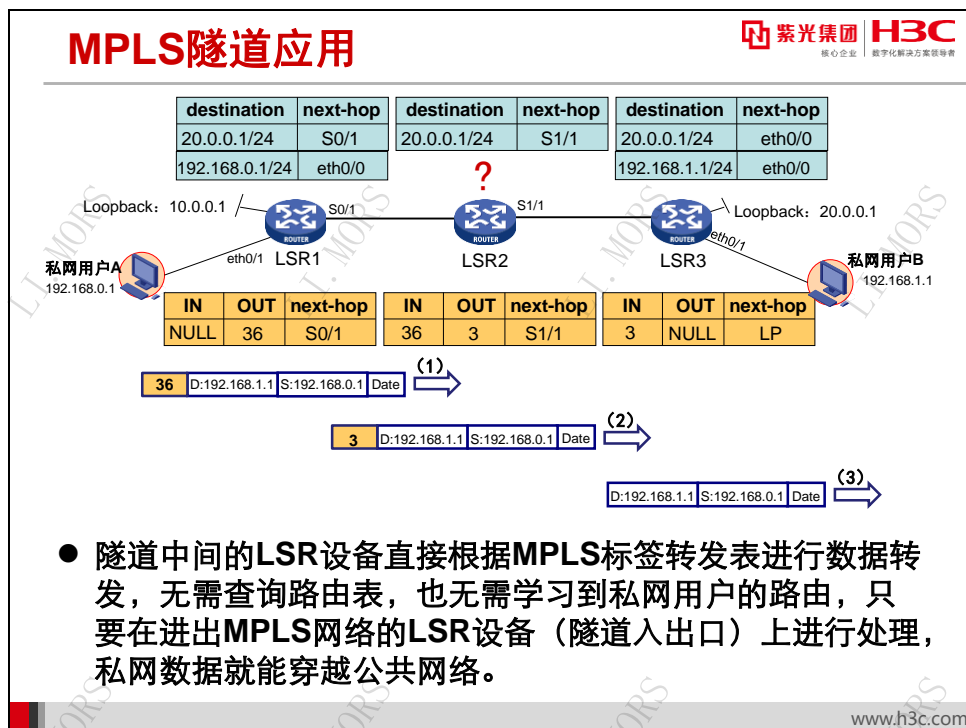
MPLS	私网IP头部	数据
------	--------	----

www.h3c.com

VPN 的实现依赖于一个很重要的技术，那就是隧道。通过隧道，用户私网在公网之间建立起一个逻辑通路，让用户的各个分部之间如有实际的物理线路相连。而隧道的实现是通过报文封装的方法。如 GRE 隧道，就是采用公网 IP 头部来封装私网报文，公网上的路由器根据报文的外层 IP 头进行转发，直到报文抵达目的私有网络，再去除外层的 IP 头，解封出私网报文。

MPLS 技术产生本意是为了加快报文的转发效率，但它其实也是一种隧道技术。从它的实现原理可以看出，它也是对报文进行封装，在 IP 报文的前面加上了 MPLS 标签，路由器直接根据标签进行转发，而无需检查内部的目的地址，这与 GRE 的封装非常类似。所以 MPLS 技术是一种天然的隧道技术，而且与已有的 VPN 所采用的隧道技术相比，它有着一个非常重要的优势，那就是动态性。

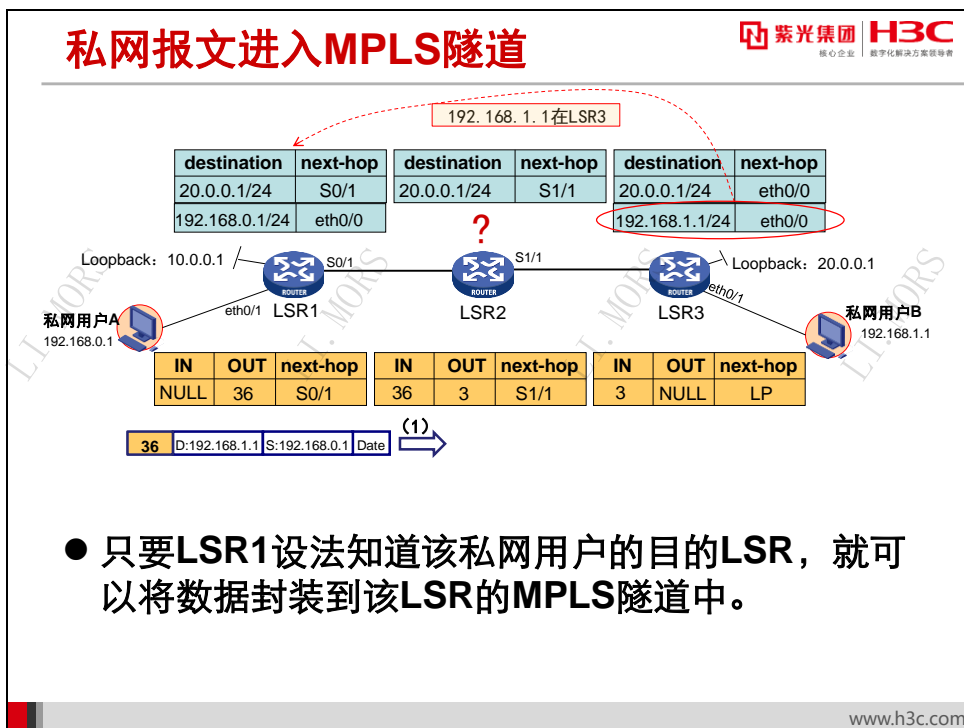
## 16.3.2 MPLS 隧道应用



如上图所示，用户 A 和用户 B 是某私网用户两个不同分部，它们需要穿过公网进行互相通信。LSR1 和 LSR3 分别是用户 A 和用户 B 连接公网的出口设备，LSR2 是公网内部的一台设备。

此时，LSR1、LSR2 和 LSR3 运行 MPLS 协议，并建立起了 LSR1 访问 LSR3 的标签转发路径。此时，如果一个报文从 LSR1 去访问 LSR3，例如访问 20.0.0.1 这个地址，那么，该报文将进行 MPLS 转发。在 MPLS 网络内部的 LSR 设备上，路由器只需要检查报文的 MPLS 标签值就可以完成转发。

设想，如果在 LSR1 设备上，将访问用户 B 的私网报文，也封装在为 20.0.0.1/24 分配的 MPLS 标签里面，并按照为 20.0.0.1/24 这条路由生成的标签转发表转发出去，如上图所示，结果私网报文在 MPLS 网络内部的 LSR 设备上只需要按照 MPLS 标签进行转发，直到抵达 LSR3 设备。这样也就在用户 A 和用户 B 所在分部连接公网的出口设备之间形成了一个隧道，隧道中间的所有 LSR 设备直接根据报文的 MPLS 标签进行数据转发，无需识别封装在 MPLS 标签内部的私网报文。可见，只要在进、出 MPLS 网络的 LSR 设备（隧道出入口）上进行特殊的处理，私网报文就能成功的穿越公共网络。

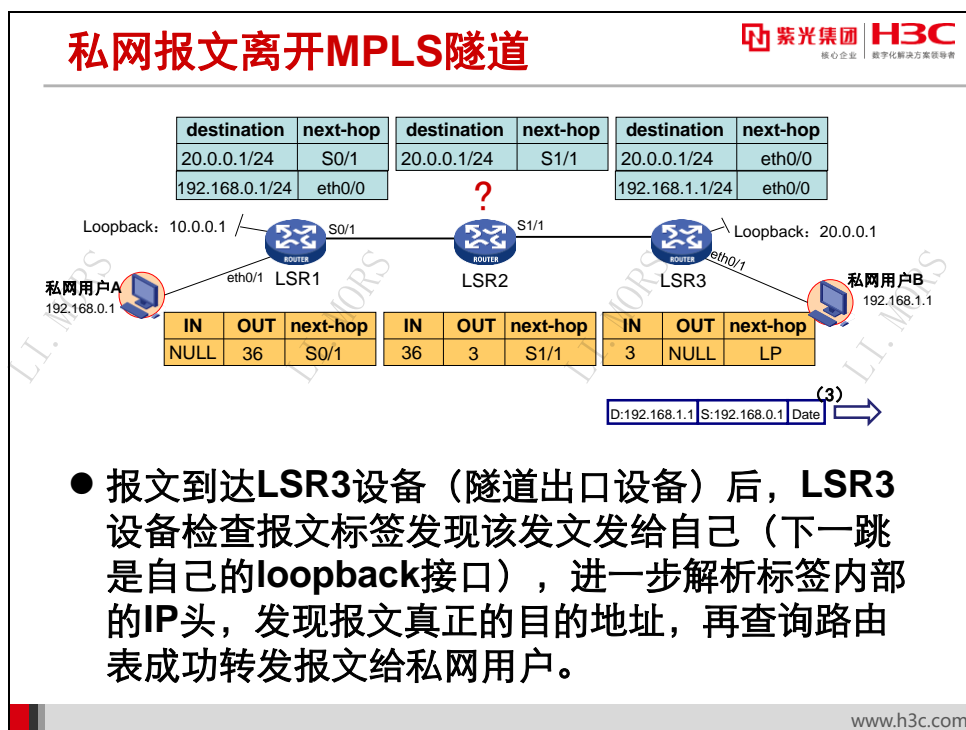


私网报文在进入 MPLS 隧道时，需要由 Ingress LSR 设备特殊进行处理，为私网报文压上合适的 MPLS 标签，报文才能进入正确的 MPLS 隧道。Ingress LSR 设备根据私网报文的目的地地址，判断出该报文需从哪一台 Egress LSR 设备离开 MPLS 网络，对端的这台 Egress LSR 设备应该是私网报文目的地地址所在网段的公网出口设备。如上图所示，当用户 A 访问用户 B 的报文抵达 LSR1 设备时，LSR1 设备需要判断出，与 192.168.1.1 这个目的网段相连的 Egress LSR 设备应该是 LSR3。判断目的私网地址所在位置，那就需要 LSR1 和 LSR3 设备之间能互相交互私网路由，而关于私网路由的交互方法，将在后续相关章节进行介绍。

当 LSR1 知道该私网报文对应的 MPLS 隧道 Egress LSR 设备是 LSR3 以后，需要进行的处理就是按照抵达 LSR3 的标签转发表压上对应的 MPLS 标签并进行转发。所以，在 MPLS 技术作为隧道来使用时，我们通常为 MPLS 网络中的每一个 LSR 设备都配置一个 Loopback 地址，来代表这台 LSR 设备。如上图所示，LSR3 的 Loopback 地址为 20.0.0.1/32，按照 MPLS 标签分配及控制原理，将在整个 MPLS 网络上形成抵达 20.0.0.1/32 这条路由的 MPLS 标签转发表。当 LSR1 设备判断出某私网报文的 MPLS 隧道出口是 LSR3 后，就会按照代表 LSR3 的 20.0.0.1/32 这条路由所对应的标签转发表项进行转发，如上图所示，为报文加上 36 的 MPLS 标签，并从 S0/1 接口转发出去。

MPLS 网络中间的 LSR 设备，如上图的 LSR2，在接收到 MPLS 标签为 36 的报文时，完全感知不到该报文 MPLS 标签后面封装着的是私网报文，而是直接按照标签转发表进行转发，报文可以成功的抵达 MPLS 隧道出口设备 LSR3。



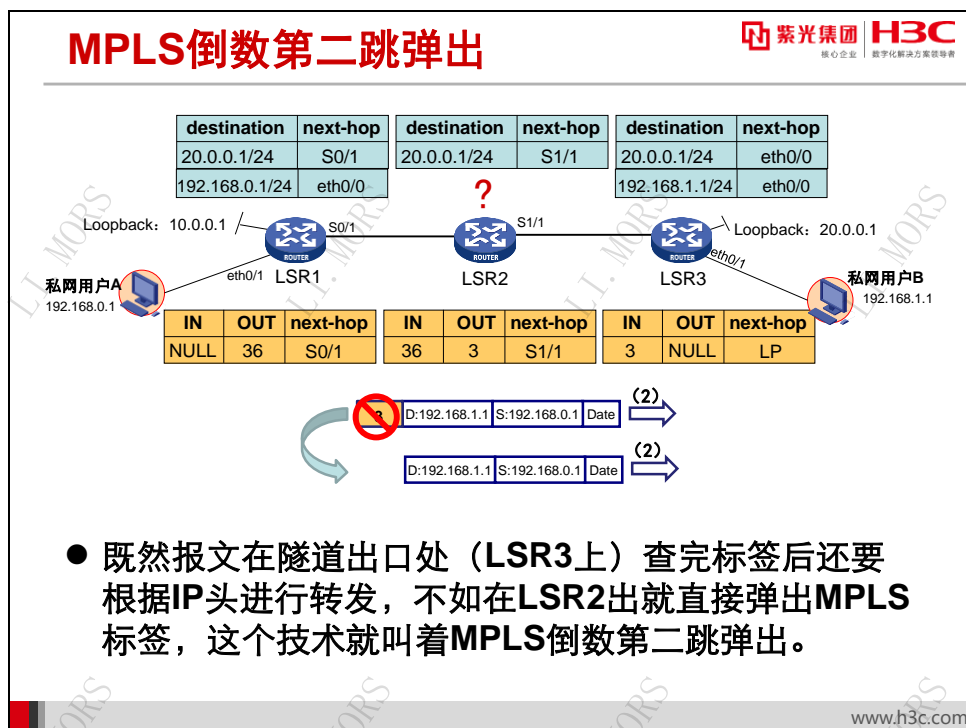


私网报文离开 MPLS 网络时的实现要比进入 MPLS 网络更为简单。如上图所示，私网报文成功的穿越所有的公网 LSR 设备后，抵达 LSR3 设备，LSR3 根据报文所携带的 MPLS 标签 3 查到对应的标签转发表，发现该报文的下一跳是 loopback0，表示该报文是发送给 LSR3 自身的，此时 LSR3 需要进一步解封封装该报文，检查报文 MPLS 标签后的内容再进行处理。而 LSR3 检查该报文 MPLS 标签内部的内容后，发现该报文是访问私网 192.168.1.1 的一个私网报文。LSR3 设备正是 192.168.1.1/24 这个私网的公网出口设备，拥有抵达 192.168.1.1/24 的路由，LSR3 设备只需要按照路由表转发该私网报文，报文就可以抵达最终的私网目的地址，也就是用户 B。可见 MPLS 技术与传统的隧道技术一样，可以在私网用户的出口设备之间建立起穿越公网的隧道。

与传统的隧道技术相比，MPLS 隧道技术有着一个非常大的优点，那就是传统的隧道技术需要在公网出口设备之间手工配置静态隧道，每建立一个隧道就要增加一组相关配置；而 MPLS 隧道技术，只需要在公共网络上运行 MPLS 协议，就可以依靠 MPLS 的动态标签分配原理，在所有的用户公网出口设备之间建立起抵达对方的标签转发路径。MPLS 隧道的这一优点，解决了传统 VPN 的一个重要缺陷，也就是 VPN 隧道静态性导致维护难度大的问题导致 VPN 的规模受限。采用 MPLS 技术作为隧道应用的 BGP MPLS VPN 因为能够支持动态建立隧道，从而可以支持更大规模的 VPN 应用。



## 16.3.3 MPLS 倒数第二跳弹出

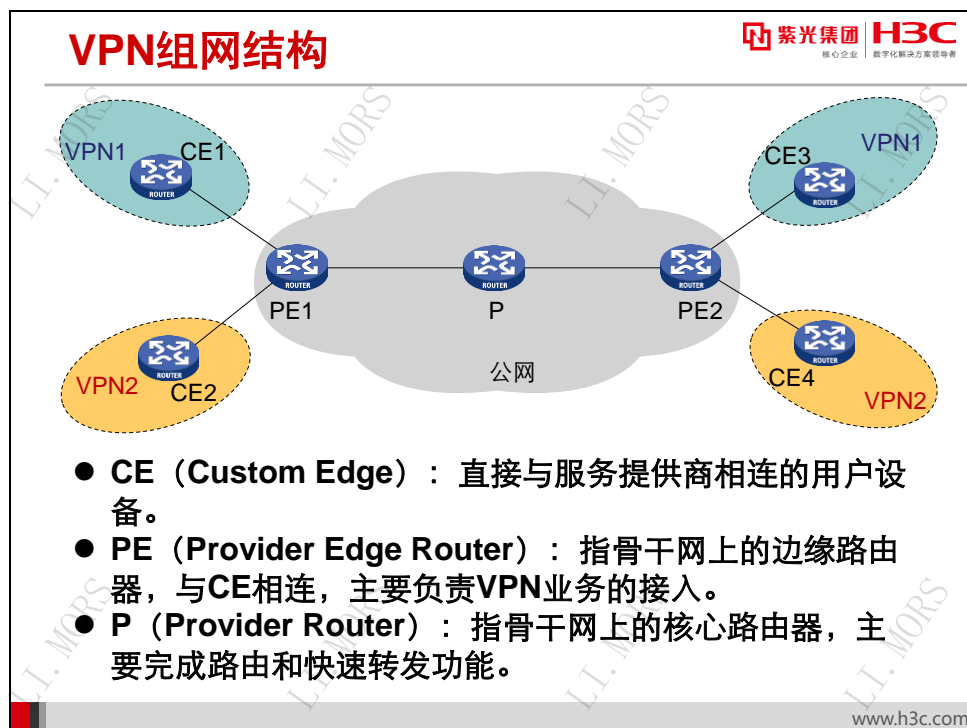


在私网报文抵达 Egress LSR 设备，如上图所示的 LSR3 设备时，LSR3 设备需要首先根据报文的 MPLS 标签检查 MPLS 标签转发表，找到对应的表项后，才发现该报文是发送给 LSR3 本身。此时 LSR3 设备需要再检查报文 MPLS 标签内部的目的 IP 地址，再根据 LSR3 上的路由表转发该报文。故在 MPLS 隧道的 Egress LSR 上需要两次查表（先查标签转发表，再查路由表）才能将报文发出。既然在隧道 Egress LSR 上最终要根据报文的 IP 头进行转发，不如在隧道 Egress LSR 上游的那一台 LSR 上，就直接将 MPLS 标签弹出。如上图所示，在 LSR2 上将本来该加上的标签 3 弹出，而直接将私网报文发送给 LSR3。此时 LSR3 将收到一个普通的 IP 报文，只需要查询路由表就可以转发该报文了。这样可以减少隧道出口 LSR 上的报文处理步骤，由两次查表转变成一次查表，提高了转发效率，这个技术就被称之为 MPLS 的倒数第二跳弹出。

在 MPLS 的倒数第二跳弹出技术的实现中，利用了一个特殊的标签值 3，使得 LSR 可以判断出它是否是报文的倒数第二跳设备。使能 MPLS 的倒数第二跳弹出技术后，在某 FEC 的最下游 LSR 上，将为该 FEC 分配一个特殊的标签值，通常为 3。当 LSR 转发数据时，检查标签转发表，发现该表项的 OUT 标签值是 3，就将 OUT 标签值 3 弹出再转发报文。如上图所示，LSR2 收到标签值为 36 的报文时，检查标签转发表，找到对应的标签转发表项，根据标签转发表项，发现 OUT 标签值是特殊值 3，此时 LSR2 路由器就会将标签弹出，直接将 IP 报文从 S1/1 接口发出。这样，抵达 LSR3 的报文就是不带 MPLS 标签的普通 IP 报文，LSR3 只需要检查 IP 路由表就可以转发该报文。

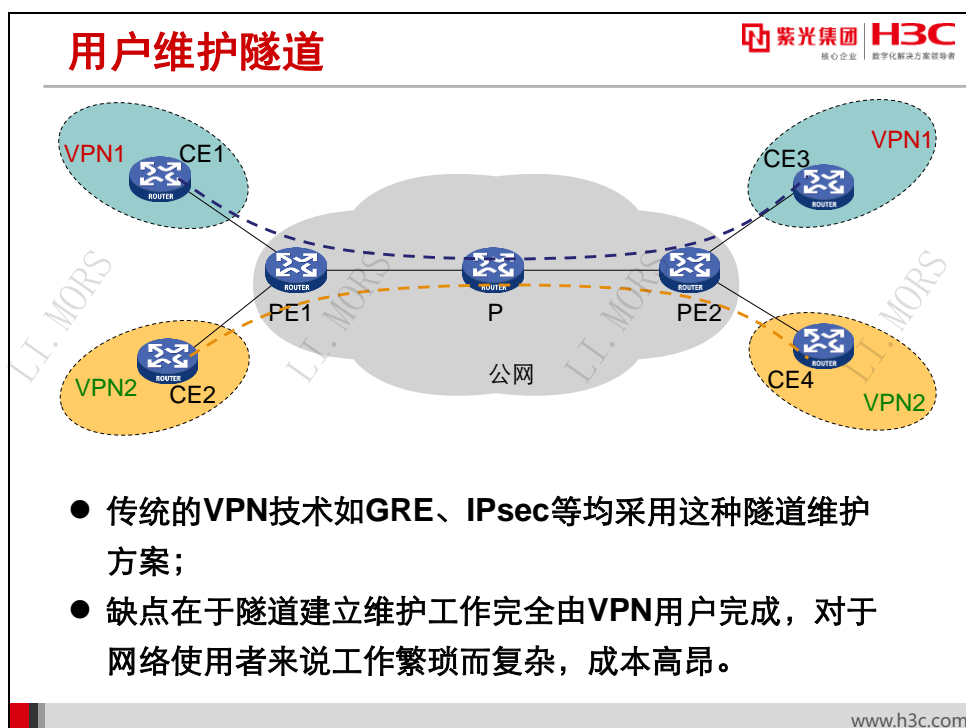
## 16.4 多VRF技术

### 16.4.1 优化 VPN 组网结构

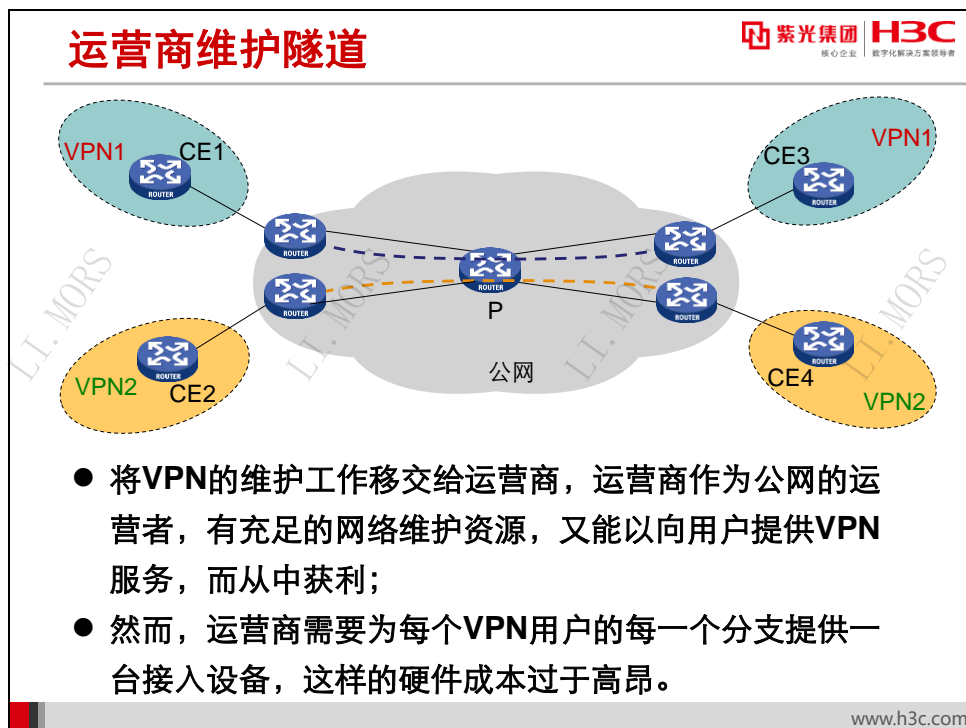


在 VPN 的组网结构中，我们将网络分成公网和私网两个部分。公网是指由服务提供商建设的公共网络，而私网是指用户自己建设的私有网络。同时将网络中的路由器按照在 VPN 网络中的位置区分成以下三种角色：

- **CE (Custom Edge Router, 用户边缘路由器)**：直接与公网相连的用户设备。
- **PE (Provider Edge Router, 服务商边缘路由器)**：指公网上的边缘路由器，与 CE 相连。
- **P (Provider Router, 服务商路由器)**：指公网上的核心路由器。



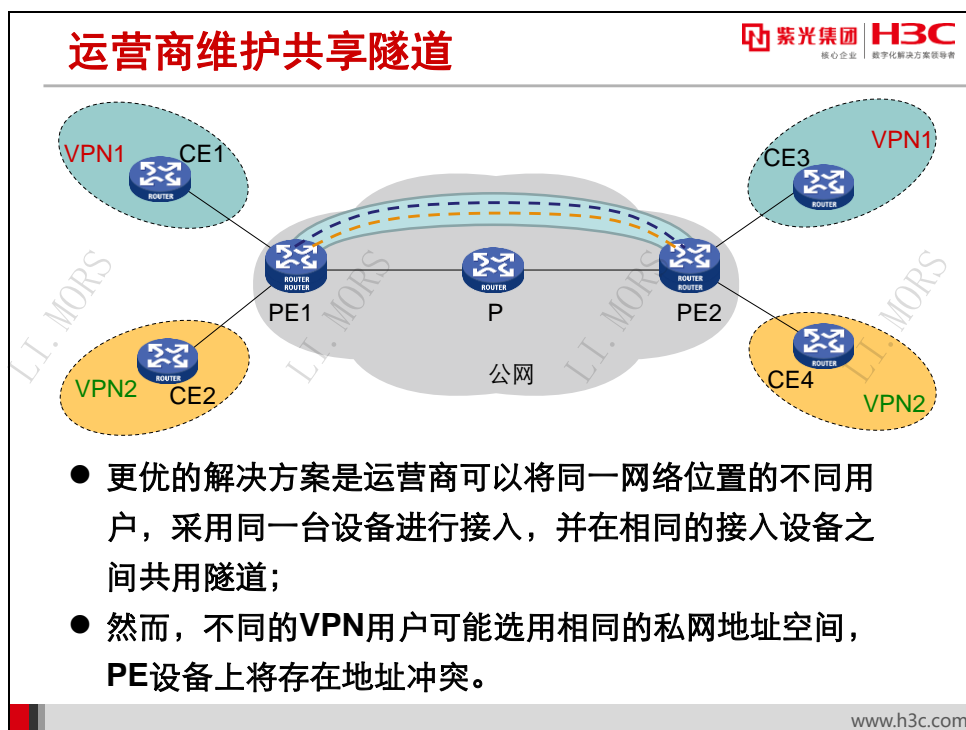
传统的 VPN 隧道建立在用户的 CE 设备之间，隧道的建立维护等工作完全需要由用户自己来完成，每个 VPN 用户的维护工作繁琐且分散，总体维护成本高昂。另一方面，作为专门提供网络服务的网络供应商，他们拥有充足的网络维护资源，但却完全无法感知用户的 VPN 应用，不能提供 VPN 相关的服务，从而无法获得相应的利润。传统 VPN 的这种结构限制了其大规模扩展，无论是用户还是运营商都希望能够将 VPN 隧道的建立维护等工作转移给运营商。



另外一种VPN方案是，隧道建立在PE设备之间，隧道的建立维护等工作由运营商来完成，如上图所示。这样，用户能够从繁琐的VPN维护工作中解脱出来，运营商也能够从VPN维护中获利。

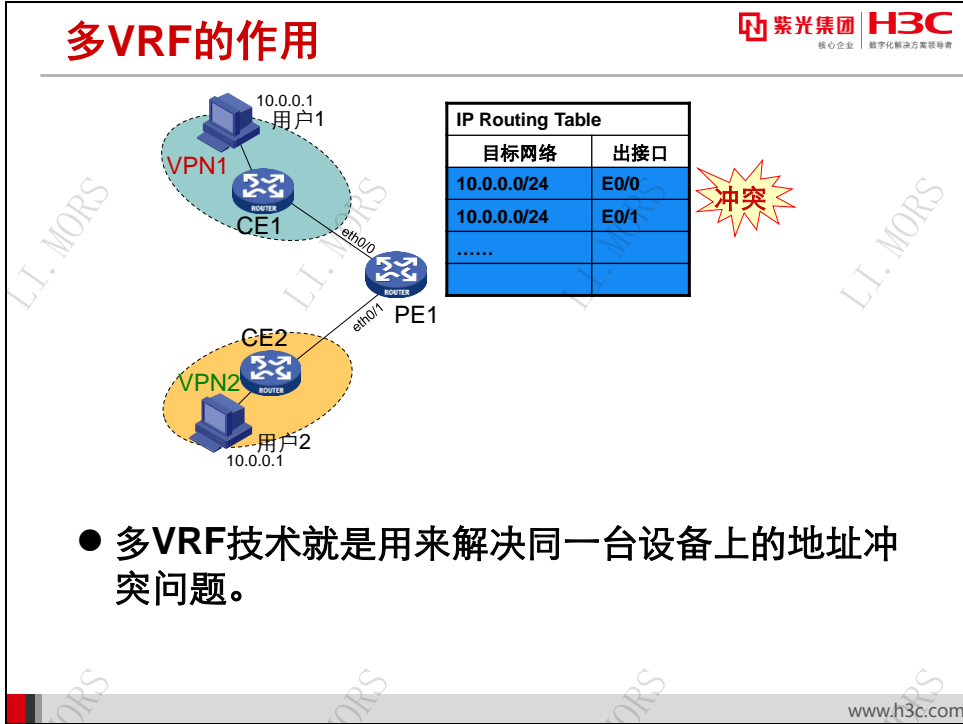
然而，在以上方案中因为不同的私网用户之间的地址空间可能完全重叠，所以运营商需要为每个VPN用户提供一台接入设备，哪怕这两个用户离得再近也不能共用。这种方案称之为专属PE的方式。

专属PE方式需要运营商为每个VPN用户提供专门的设备，硬件成本过于高昂，用户需要向运营商支付较高的费用，无法获得广泛认可。

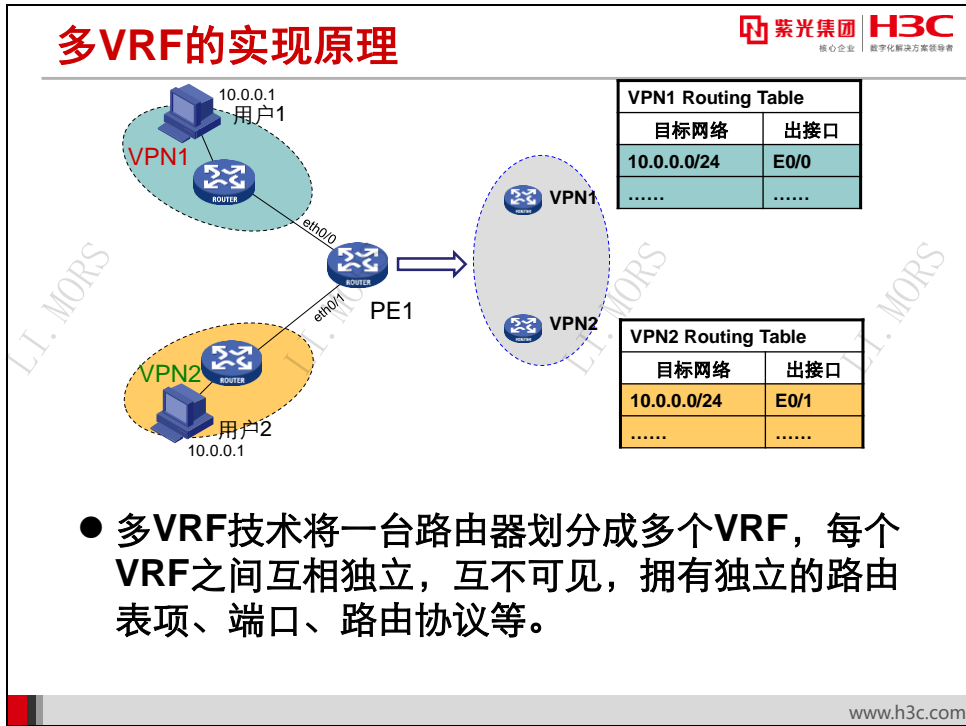


更优的 VPN 隧道方案是隧道建立在公网的 PE 之间，每一台公网 PE 设备还可以同时接入多个 VPN 用户，多个 VPN 用户能够共享一个隧道。这个方案无论从可维护性上，还是从成本上都满足了用户和运营商的需求。然而，由于不同的 VPN 用户可能选用了相同的地址空间，这样会造成 PE 设备无法区分这些用户的数据流。也就是说，此方案存在着同一台设备上地址冲突的问题。多 VRF（Virtual Routing and Forwarding，虚拟路由与转发）技术可以解决这个问题。

16.4.2 多 VRF 技术实现原理



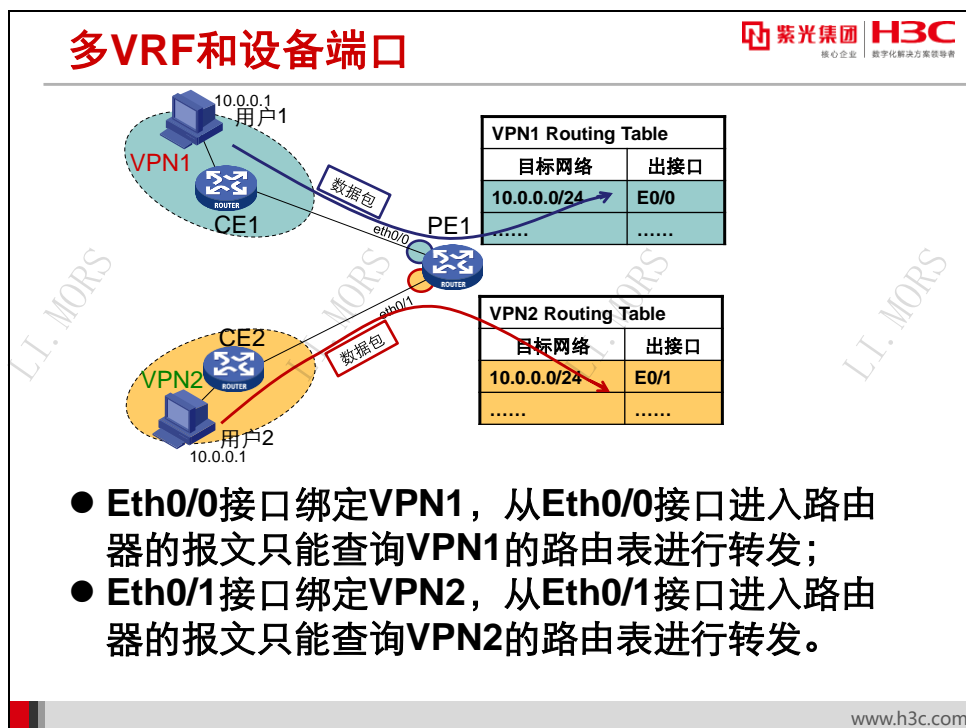
多 VRF 技术的目的就是要解决在同一台设备上的地址冲突问题。如上图所示，该 PE 路由器的两个不同接口，分别接入 VPN1 和 VPN2 用户，而 VPN1 用户和 VPN2 用户都选用了 10.0.0.1/24 这个网段的地址。在没有多 VRF 技术的情况下，这两条路由将会在 PE 上发生冲突，在 PE 的路由表里面只能保留一条 10.0.0.0/24 路由。



支持多 VRF 技术的路由器将一台路由器划分成多个 VRF，每个 VRF 之间互相独立，互不可见，各自拥有独立的路由表项、端口、路由协议等等，每一个 VRF 就类似一台虚拟的路由器。

如上图所示，PE1 为了能同时接入 VPN1 和 VPN2 这两个地址冲突的私网用户，启用两个 VRF，每个 VRF 与 VPN 相对应，在 VRF1 里面只看到与 VPN1 相连的接口，只学习 VPN1 的路由；VRF2 上也只看到与 VPN2 相连的接口，只学习 VPN2 的路由。这两个 VRF 各自拥有自己的路由表，VRF1 学习到的 10.0.0.0/24 路由，下一跳指向 Eth0/0；而 VRF2 学习到的 10.0.0.0/24 路由，下一跳指向 Eth0/1，这 2 条路由同时存在于 PE1 上，互不冲突，互不影响。

有了支持多 VRF 技术的 PE 设备，就可以实现一台 PE 接入多个 VPN 用户。

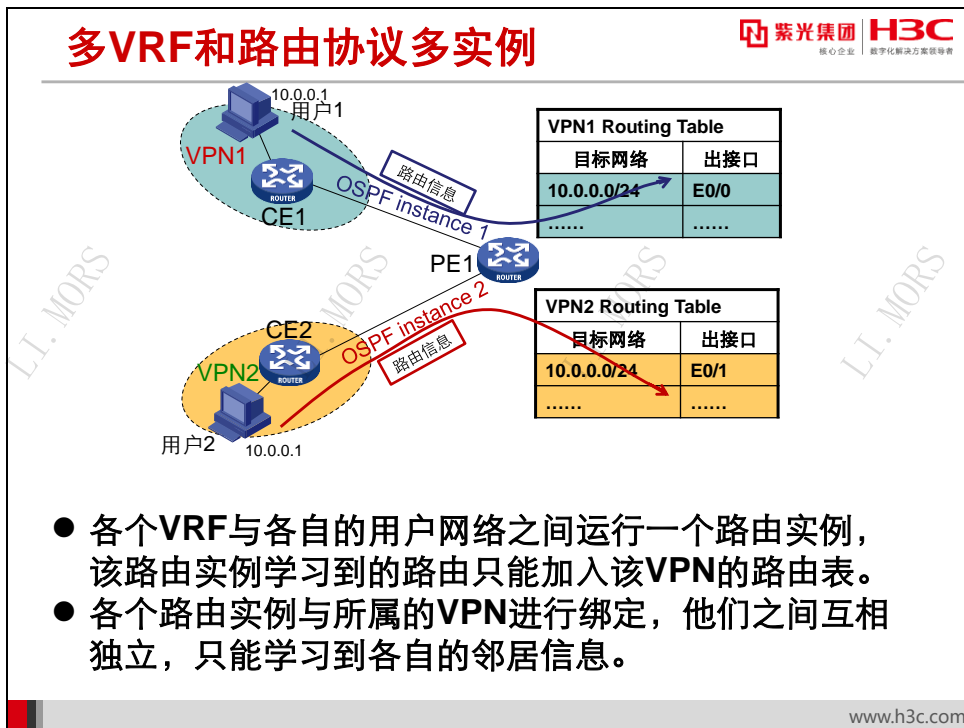


为了加深理解多 VRF 技术的实现，在此进一步说明多 VRF 与设备端口的关系。在支持多 VRF 的 PE 设备上，需要将和某一个 VPN 用户相连的接口与对应的 VRF 相绑定。如上图所示，PE1 会将 Eth0/0 与 VRF1 绑定，而将 Eth0/1 与 VRF2 绑定。与 VRF 绑定后的接口，只会出现在该 VRF 对应的路由表里面；并且，当报文从该接口进入路由器时，只能查询该 VRF 对应的路由表进行转发。

如上图所示，从 Eth0/0 接口进入路由器的报文，只能查询 VPN1 的路由表进行转发；从 Eth0/1 接口进入路由器的报文，也只能查询 VPN2 的路由表进行转发。

多 VRF 与设备端口的这种关系，确保了数据进入多 VRF 设备进行转发时，不同 VPN 用户的数据之间不会发生冲突。





在支持多 VRF 技术的路由器上，不同的 VRF 运行独立的路由协议，这些路由协议不会交互协议报文，且学习到的路由也放在各自 VRF 的路由表中，互不影响。实现这种效果需要将各个 VRF 运行的路由协议与该 VRF 进行绑定。

不同的 VRF 可以选择采用同一种路由协议与 VPN 用户之间交互路由，例如均采用 OSPF 路由协议；当然也可以选择不同的路由协议。在支持多 VRF 的路由器上需要运行多个路由协议进程，并将不同的进程与不同的 VRF 进行绑定，将路由协议的某个进程与某 VRF 进行绑定的做法称之为路由协议的多实例。

如上图所示，PE 路由器上运行两个 OSPF 实例。OSPF instance 1 和 VRF1 进行绑定，与私网 VPN1 用户的路由器 CE1 建立 OSPF 邻居；OSPF instance 2 和 VRF2 进行绑定，与私网 VPN2 用户的路由器 CE2 建立 OSPF 邻居。此时 OSPF instance 1 学习到的路由就只会收录到 VRF1 所对应的路由表中，而 OSPF instance 2 学习到的路由就只会收录到 VRF2 所对应的路由表中。

多 VRF 和路由协议多实例之间的绑定关系，确保了即使不同的 VPN 用户选用相同的地址空间，并与 PE 设备之间通过路由协议交互路由时，也不会在 PE 上出现路由冲突。


目前，绝大多数的路由器设备都能支持路由协议多实例功能，包括 OSPF、IS-IS、BGP、RIP、静态路由等。

## 16.5 MP-BGP 技术

### 16.5.1 MP-BGP 技术实现

### BGP 协议的特点

- 基于 TCP 链接，可以跨越多台设备建立路由邻居，传递路由；
- 基于 TLV 架构，可以扩展属性位，以便携带更多表明路由特征的信息；
- BGP 路由协议还有很多的特点，而上述的两个特点是他被选作 VPN 私网路由协议的主要原因。



紫光集团 H3C  
核心企业 数字化解决方案领导者

www.h3c.com

有了比传统 VPN 更加先进的隧道技术，也有了比传统 VPN 更加合理的 VPN 组网结构，最后只需要一种更加先进的路由协议，用于在 PE 设备之间交互私网路由信息。纵观各种路由协议，最合适充当这个角色的路由协议是 BGP 路由协议，因为它有如下两个非常关键的特质，适用于 VPN 技术中。

- BGP 路由协议是基于 TCP 连接建立路由邻居的，可以实现跨越多台路由器建立 BGP 邻居，直接交互路由信息。在 VPN 的组网中，就是需要两台 PE 设备直接交互私网路由，无需经过中间的 P 设备转达。因为作为隧道中间的 P 设备，它们无需了解 VPN 的信息；另外，不同 VPN 间地址空间会有重叠，如果这些 VPN 路由都经过 P 设备转达，P 设备上将无法区分。
- BGP 路由协议的协议报文是基于 TLV 结构的，具有扩展属性位，便于携带更多的表明路由特征的信息。也就是说，BGP 路由协议可以为 VPN 组网定义一些扩展属性，适应 PE 之间交互私网路由的需要。这一点对于 VPN 非常重要，因为有了多 VRF 技术以后，每一台 PE 设备需要接入多个 VPN 用户，一对 PE 设备之间交互的私网路由，需要设法携带上某种特征，才能够通知对端 PE 这些路由信息属于哪一个 VRF。

因为上述的两个特质，BGP 路由协议被 BGP MPLS VPN 技术选用作于穿越公网传递私网路由的路由协议。为了适应 VPN 技术的需要，BGP 路由协议进行了一定的扩展，扩展后的 BGP 路由协议叫着 MP-BGP (Multiprotocol BGP)，即多协议 BGP 路由协议。

## BGP路由更新



- BGP协议通过BGP更新（UPDATE）消息发布和删除路由，BGP更新消息结构如下：

Withdrawn Routes Length	Withdrawn Routes
Total Path Attribute Length	Path Attribute
Network layer Reachability Information (NLRI)	

- BGP协议更新（UPDATE）消息主要包含以下三个部分：
  - Withdrawn Routes：之前发布过，不再有效的路由；
  - Path Attributes：路由信息的属性（附加描述），是BGP用以进行路由控制和决策的信息；
  - NLRI：路由信息，有一个或多个IPv4地址/前缀长度组成。
- 由此可以看出普通的BGP路由更新消息只能发布或删除IPv4路由。

www.h3c.com

在普通的 BGP 路由协议里，BGP 协议通过 BGP 更新（UPDATE）消息发布和删除路由，格式如上图所示。

BGP 协议的更新消息组要包括以下三个部分：

- **Withdrawn Routes**：之前发布过的，现在不再有效的路由信息。
- **Path Attributes**：路由信息的属性（附加描述），是 BGP 用以进行路由控制和决策的信息，如 LP 属性、MED 属性等等；
- **NLRI**：路由信息，由一个或者多个 IPv4 地址前缀组成，这个就是需要发布给邻居的生效的路由信息。

普通的 BGP 路由协议就是采用这样的更新消息发布或者删除 BGP 路由信息，通过 BGP 更新消息的格式，可以看出，普通的 BGP 路由协议只能用于发布或删除 IPv4 路由。

## MP-BGP协议



- 普通BGP只能传递IPv4路由信息，为了能够承载多个协议的路由信息，RFC2858对BGP进行了扩展，扩展后的BGP协议称之为多协议BGP（MP-BGP）。
- MP-BGP新增了MP\_REACH\_NLRI和MP\_UNREACH\_NLRI两个属性
- RFC4360对团体（Communities）属性进行了扩展，新增扩展团体属性（Extended\_Communities）。
- MP-BGP路由协议可以传递BGP MPLS VPN、L2VPN、6PE等路由信息。

www.h3c.com

为了适应 VPN 技术的需要，让 BGP 路由协议能够承载更多形式的路由信息，比如说 VPN 的私网路由，RFC2858(更新的 RFC 为 4760)对 BGP 协议进行了扩展，扩展后的路由协议叫做 MP-BGP（Multiprotocol BGP，多协议 BGP）。

MP-BGP 新增了 MP\_REACH\_NLRI 和 MP\_UNREACH\_NLRI 两个属性。

RFC4360 对 BGP 协议原有的团体属性（Communities 属性）进行了扩展，新增了扩展团体（Extended Communities）属性。

MP-BGP 路由协议通过对 BGP 协议的扩展，不仅仅用于 BGP MPLS VPN 技术中传递私网路由，还可以用在 IPv6、6PE、L2VPN 等技术中，从而有了更广泛的应用。

## MP-BGP路由更新



- MP-BGP路由更新消息相对普通BGP路由更新消息作出如下改动：

- MP\_REACH\_NLRI属性代替原BGP更新消息里面的NLRI及Next-hop属性
- MP\_UNREACH\_NLRI属性代替原BGP更新消息里面的Withdrawn Routes
- 属性部分增加Extended\_Communities

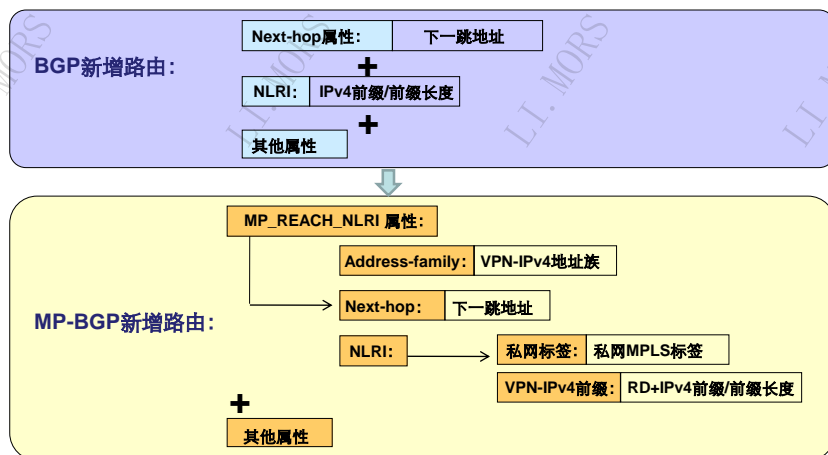
www.h3c.com

MP-BGP 协议相对 BGP 协议重点是对路由更新消息进行了改动，具体的改动包括以下的三个部分：

- 采用 MP\_REACH\_NLRI 属性代替了原 BGP 更新消息里的 NLRI 及 Next-hop 属性；
- 采用 MP\_UNREACH\_NLRI 属性代替了原 BGP 更新消息里面的 Withdrawn Routes；
- 在 BGP 属性部分增加的一种新的属性 Extended\_Communities 属性。

## MP\_REACH\_NLRI属性

- **MP\_REACH\_NLRI**是对原BGP更新消息中NLRI的扩展，增加了地址族的描述，以及私网Label和RD，并包含了原BGP更新消息中的Next-hop属性。



MP\_REACH\_NLRI 是对原 BGP 更新消息中的 NLRI 的扩展，增加了新的地址族的描述，以及私网标签和 RD，并且包含了原 BGP 更新消息中的 Next-hop 属性。如上图所示，MP\_REACH\_NLRI 属性中共同具体内容包括：

- **Address Family:** 说明下面的路由前缀将采用的地址类型，不仅仅包含原先的普通 IPv4 地址，可能是 IPv6 地址，或者是用于 BGP MPLS VPN 组网中传递私网路由所要用的 VPNv4 地址；
- **Next-hop:** 下一跳信息，与原 BGP 的下一跳属性的内容相同；
- **NLRI:** 如果在 Address Family 区域中指明采用的地址族是 VPNv4 地址族，那么该处的格式将包含两个部分，第一部分是私网标签，是一个 MPLS 标签值；第二部分是一个 VPNv4 地址，其格式是 RD+IPv4 地址。VPNv4 地址的这种格式是为了满足 BGP MPLS VPN 组网中传递私网路由的需要。



MP\_UNREACH\_NLRI 属性代替原 BGP 更新消息中的 Withdrawn Routes，格式如上图所示。它包括两个部分，Address Family 和 Withdrawn Routes，其中增加的 Address Family 与 MP\_REACH\_NLRI 属性中的 Address Family 形式完全相同，Withdrawn Routes 中的地址前缀的地址类型由 Address Family 的内容来决定。如在 BGP MPLS VPN 的应用中，Address Family 将指示地址族为 VPNv4 地址族，而 Withdrawn Routes 区域的地址就会是 VPNv4 地址，格式为 RD+IPv4 地址。

## 16.5.2 Route Target 属性

Route Target

紫光集团 核心企业 数字化转型方案领导者

- **BGP的扩展community属性：RT（Route Target）**
- **扩展的community有如下两种格式：其中 type 字段为 0x0002、0x0102 或者 0x0202 时表示 RT。**

TYPE(2字节)	Administrator Field	Assigned Number Field
0x0002	2字节AS号	4字节分配编号
0x0102	4字节IP地址	2字节分配编号
0x0202	4字节AS号	2字节分配编号

[www.h3c.com](http://www.h3c.com)

MP-BGP 协议的最后一个改动是增加了一个扩展团体属性（Extended\_Communities），这个属性的最主要内容就是 RT（Route Target）即路由目标，通常都被简称为 RT 属性。

扩展团体属性的格式如上图所示。当 BGP 属性类型域的值为 0X0002、0X0102 或者为 0X0202 时，标识为 RT 属性。这三种类型值代表三种不同的 RT 格式，通常由用户根据自己的使用习惯来决定使用那种格式。

0X0002 类型的 RT 格式为 2 字节的 AS 号加上 4 字节的用户自定义数字，如 100:1、200:1 等。其中 100、200 通常为 BGP 的 AS 号，当然不是必须的，用户可根据理解的方便进行配置；而后面的 1、2 通常是 VPN 编号，表示不同的 VPN。

0X0102 类型的 RT 格式为 4 字节的 IP 地址加上 2 字节的用户自定义数字，如 192.168.1.1:1、202.1.1.1:2 等。前面的 IP 地址通常为 PE 设备的 Router ID，当然也不是必须的，用户可根据理解的方便进行配置；后面的 1、2 通常是 VPN 编号，用于表示不同的 VPN。

0X0202 类型的 RT 格式为 4 字节的 AS 号加上 2 字节的用户自定义数字，如 65536:1、65537:2 等。其中的自治系统号最小值为 65536，用户可根据理解的方便进行配置；后面的 1、2 通常是 VPN 编号，用于表示不同的 VPN。



## Route Target本质



- RT的本质是每个VPN实例表达自己的路由取舍及喜好的方式。
- RT由Export Target与import Target两部分构成：
  - 在PE设备上，发送某一个VPN用户的私网路由给其BGP邻居时，需要在MP-BGP的扩展团体属性区域中增加该VPN的Export Target属性
  - 在PE设备上，需要将收到的MP-BGP路由的扩展团体属性中所携带的RT属性值，与本地每一个VPN的Import Target属性值相比较，当这两个值存在交集时，就需要将这条路由添加到该VPN的路由表中去

www.h3c.com

RT 属性的本质是为每个 VPN 实例表达自己的路由取舍及喜好，RT 属性分为两个部分：Export Target 属性和 Import Target 属性，在 PE 上定义某一个 VPN 时，需要给这个 VPN 设计并配置 RT 属性值。MP-BGP 在 PE 间交互私网路由时，需遵循如下的规则：

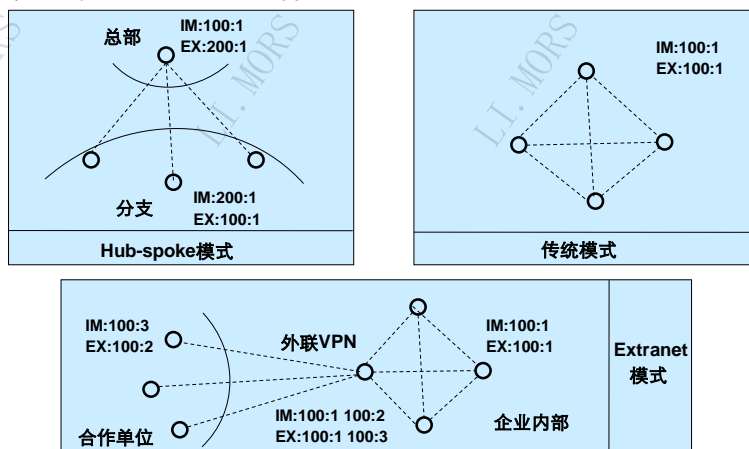
- 在 PE 设备上，发送某一个 VPN 用户的私网路由给其 BGP 邻居时，需要在 MP-BGP 的扩展团体属性区域中增加该 VPN 的 Export Target 属性；
- 在 PE 设备上，需要将收到的 MP-BGP 路由的扩展团体属性中所携带的 RT 属性值，与本地每一个 VPN 的 Import Target 属性值相比较，当这两个值存在交集时，就需要将这条路由添加到该 VPN 的路由表中去。

通过以上规则，能够实现同一个 VPN 用户的路由进行交互，而不同 VPN 的用户路由不能交互，也就控制了 VPN 用户之间的互访关系。

## RT的灵活应用

紫光集团 H3C  
核心企业 数字化转型先锋

- 由于每个RT Export Target与import Target都可以配置多个value，接收时是“或”操作，所以就可以实现非常灵活的VPN访问控制。



www.h3c.com

RT 的设计使得采用 BGP MPLS VPN 技术的 VPN 站点之间的互访关系的控制变得非常灵活。RT 实际上是一个属性列表，可以配置多个 RT Export Target 和 Import Target 属性值，在发送时，MP-BGP 会携带所有的 Export Target 属性值，而在接收时，MP-BGP 将接收到路由信息里携带的 RT Export Target 值与本地 VPN 的 Import Target 属性值进行比较。比较时采用的是“或”操作，也就是说，是只要这两个列表存在交集，该路由就可以被该 VPN 接收。通过以上的实现方式，可以使得 VPN 的互访关系非常的灵活多样。

依靠 RT 属性的这一特点，可以控制同一个 VPN 用户的不同站点之间不单单是简单的完全可互访关系，还可以是上图所示的 hub-spoke 模式，也可以使得不同的 VPN 用户在某些特殊的部分位置可以与其他 VPN 用户互通，如上图所示的 extranet 模式。下文以 hub-spoke 模式分析一下其具体实现过程。

hub-spoke 模式是一种 VPN 用户的特殊互访模式。这种互访关系要求，用户的总部可以与其每一个分部进行互通，而该 VPN 用户的每一个分部之间禁止互相访问。BGP MPLS VPN 可以通过 RT 的设计轻松满足用户的这一需要，而不需要像传统 VPN 那样通过大量的访问控制列表实现。

在上图所示的例子中，在用户总部接入的 PE 设备上，为某 VPN 配置 RT 的 Import Target 值为 100:2, Export Target 值为 100:1；而在每个分部接入的 PE 上，为该 VPN 配置 RT 的 Import Target 值为 100:1, Export Target 值为 100:2。以这样的 RT 设计，可以发现，当分部 PE 将该 VPN 的路由发送给总部 PE 时，会在 MP-BGP 路由信息中携带 RT 属性值 100:2，总部 PE 收到该路由后，发现与本地该 VPN 的 Import Target 属性值相同，于是将路由加入到该 VPN 的路由表中。而当总部 PE 将该 VPN 路由发送给各个分部 PE 时，会在 MP-BGP 路由信息中携带 RT 属性值 100:1，各个分部 PE 收到该路由后，发现与本地该 VPN 的 Import Target 的属性值

相同，于是将路由加入到该 VPN 的路由表中。这样，各个分部与总部之间可以完成路由交互，实现互通，总部可以访问各个分部，各个分部也均能访问总部。而如果一台分部 PE 将该 VPN 的路由发送给另一分部的 PE 时，也会在 MP-BGP 路由信息中携带 RT 属性值 100:2，另一分部 PE 路由器收到该路由后，对比发现与本地该 VPN 的 Import Target 的属性值 100:1 不同，则路由不能加入到该 VPN 的路由表中，因此每个分部 PE 之间无法完成路由交互，自然各个分部之间也就无法互通。

这种通过 RT 来控制路由的交互，从而控制 VPN 用户之间甚至 VPN 内部的互访关系的办法，非常的灵活，可以根据用户的需求进行规划。而在用户的互访关系发生变化时，也可以通过修改 RT 的配置进行灵活的变更，而不需要中断 VPN 用户的业务。相对传统 VPN 技术的采用访问控制列表控制互访的方法，通过 RT 通知的 BGP MPLS VPN 技术实现安全可靠且不会影响路由器转发的性能。VPN 的互访关系灵活可控也是 BGP MPLS VPN 技术相对传统 VPN 技术的优势之一。

### 16.5.3 RD 前缀

## RD (Route Distinguisher)

紫光集团 核心企业 数字化转型方案领导者

- **RD (Route Distinguisher) 路由区分**，在 BGP MPLS VPN 的网络中，私网路由的路由前缀的形式不再是普通的 IPv4 地址，而是 RD+IPv4 地址，这样可以在路由前缀中直接标识该路由的 VPN 信息。
- **RD 的格式**
  - 16 位自治系统号 : 32 位用户自定义数，例如：100:1
  - 32 位 IP 地址 : 16 位用户自定义数，例如：172.1.1.1:1
  - 32 位自治系统号:16 位用户自定义数字，例如：65536:1

TYPE(2字节)	Administrator Field	Assigned Number Field
0x0002	2字节AS号	4字节分配编号
0x0102	4字节IP地址	2字节分配编号
0x0202	4字节AS号	2字节分配编号

www.h3c.com

MP-BGP 协议为了能够传递 BGP MPLS VPN 私网路由，设计了一个 VPNv4 地址族，这个地址族的格式是 RD+IPv4 地址，也就是在普通的 IPv4 地址前面加了一个被称之为 RD 的前缀。

RD (Route Distinguisher) 是为了标识该路由信息所属的 VPN。RD 的格式如上图所示，它一共有 48 位。在定义一个 VPN 时，除了要为该 VPN 设计 RT 属性外，还要为该 VPN 设计一个 RD 值。每个 VPN 只能配置一个 RD 值，其格式有以下三种：

- 16 位自治系统号 (AS Number) 加上 32 位用户自定义数值。例如：100:1

- 32 位 IP 地址加上 16 位用户自定义数值。例如：172.1.1.1:1
  - 32 位自治系统号（AS Number）加上 16 位用户自定义数值。例如：65536:1
- 用户可以根据自己的使用习惯或表达需要来决定使用那一种 RD 格式。

## RD 的本质

紫光集团 H3C  
核心企业 | 数字化解决方案领导者

- RD 的作用是用于私网路由的撤销，因为按照 BGP 原理，在撤销路由时不会携带路由的属性值，也就不能携带 RT 属性，PE 在删除路由时无法判断是要撤销哪个 VPN 的路由。
- 理论上可以为每个 VPN 实例配置一个 RD。通常建议为每个 VPN 都配置相同的 RD，不同的 VPN 配置不同的 RD。但是实际上只要保证存在相同地址的两个 VPN 实例的 RD 不同即可。
- 如果两个 VPN 实例中存在相同的地址，则一定要配置不同的 RD，而且两个 VPN 实例一定不能互访，间接互访也不成。

www.h3c.com

在 MP-BGP 协议中，由 RT 来控制 VPN 之间的互访，当一条路由抵达 PE 时，PE 可根据该路由所携带的 RT 属性判断该路由应该被学习到本地哪一个 VPN 的路由表。而 RD 前缀的目的在于撤销路由时使用。在 BGP 协议中规定，BGP 在发布路由撤销消息时，将不会携带路由属性域，目的是为了减少路由撤销消息的报文大小，降低 BGP 路由协议占用的网络资源。在普通的 BGP 协议里，撤销消息不携带路由属性可以满足应用；但在 MP-BGP 协议中，发送路由撤销消息时，因无法携带 RT 属性，出现 PE 无法判断这条撤销消息想要删除的是哪个 VPN 的路由。RD 的目的就是用于解决 MP-BGP 撤销路由时遇到的问题。

在 MP-BGP 里，路由的前缀变成 RD+IPv4。当 PE 接收到 MP-BGP 路由时，路由表中的表项是 VPNv4 地址格式；当收到路由撤销消息时，撤销消息里面的路由前缀也是 VPNv4 的格式，也会带着 RD，这样，只要不同 VPN 用户设计成不同的 RD，就可以明确的区分出该路由撤销消息想要撤销的是哪一个 VPN 的路由，即使不同的 VPN 的 IPv4 地址是重叠的，也能够明确的区分开来。

进一步理解了 RD 的作用，可以发现，实际上只需要同一台 PE 上两个存在地址冲突的两个 VPN 的 RD 值配置得不相同就能解决路由撤销时的问题。当不存在地址冲突时，通过 IPv4 地址的不同就能分辨要删除的路由；而不同的 PE 设备上的 VPN 的 RD 值更是不存在关系，因为 PE 能判断撤销消息是由哪一个 PE 设备发来，也就只会删除从该 PE 设备学习到的路由。

不同 VPN 的所采用的地址空间都有可能存在重叠，所以通常情况下，PE 要为每一个 VPN 用户设计一个本地唯一的 RD 值，避免可能发生的地址冲突。

需要强调的是 RD 前缀的作用只是发生在路由撤销的时候，RD 是否相同并不决定路由的取舍，也不能控制 VPN 用户的互访关系，路由的取舍完全由 RT 属性来决定。

## VPNv4和IPv4 地址族

紫光集团 H3C  
核心企业 数字化解决方案领导者

- 在IPv4地址加上RD之后，就变成VPNv4地址族了。而原来的标准的地址族就称为IPv4。
- VPNv4地址族主要用于PE路由器之间传递VPN路由
- VPNv4地址只是存在于MP-BGP的路由信息和PE设备的私网路由表中，也就是只是出现在路由的发布学习过程中。
- 在VPN数据流量穿越供应商骨干时，包头中没有携带VPNv4地址。

VPNv4地址结构：

Route Distinguisher(8字节)	IPv4地址
--------------------------	--------


www.h3c.com

理解了 RD 前缀再来看 VPNv4 的地址就更为清晰了，VPNv4 的地址与普通的 IPv4 的地址相比，结构如上图所示，前面多 8 个字节的 RD 值。

需要说明的是，VPNv4 地址只是存在于 MP-BGP 的路由信息和 PE 设备的私网路由表中，也就是只是出现在路由的发布学习过程中。在用户业务的数据包中，并没有改变 IP 头的结构，将报文 IP 地址区域改成 VPNv4 地址，如果这样就需要更改路由器数据转发的实现了，因为需要去识别一种新的地址格式。

## 16.5.4 MPLS 私网 Label

## 私网Label



核心企业 数字化转型领导者

- RT属性和RD前缀顺利解决了私网路由的学习和撤销中存在的问题，然而因为VPN地址的冲突在数据转发过程也将遇到困难。
- 需要在数据报文中增加一个标识，以帮助PE判断该报文是去往本地的那个VPN。
- 由于MPLS支持多层标签的嵌套，这个标识可以定义成MPLS标签的格式，即私网Label。

www.h3c.com

RT 属性和 RD 前缀顺利解决了私网路由的学习和撤销中存在的问题，然而因为 VPN 地址的冲突在数据转发过程也将遇到困难。试想一个用户的私网报文通过 MPLS 隧道抵达了出口的 PE 设备，PE 需要查询报文的目的 IP 地址，根据本地私网路由表进行转发。但问题在于，在该 PE 上可能接入了多个 VPN，这些 VPN 的地址可能是重叠的，报文的目的地址可能在本地的多个 VPN 的私网路由表中都存在，此时，PE 将无法决定该按照本地的哪一个 VPN 私网路由表进行转发。

要解决上述问题，就需要在数据报文中增加一个标识，以帮助 PE 判断该报文是去往本地的那个 VPN。此时最快联想到的就是采用已经设计出来的 RT 或者 RD 来完成这个任务，然而如果在普通的 IP 报文中增加 RT 或者 RD 将创造出一种新的报文格式，PE 设备将需要升级该着支持这种格式的报文转发；此外 RD 有 8 个字节，RT 是一个更大的属性列表家在报文里面讲占用网络带宽资源。因此这个位置采用 RD 和 RT 并不是好的选择，相反 MPLS 的标签成为这个位置的最佳选择。

PE 设备将本端的私网路由发往对端 PE 时，为该路由分配个私网 MPLS 标签值，存放在 MP\_REACH\_NLRI 属性内发给对端，其格式上文已经说明。对端 PE 在接收到该路由以后，将路由中携带的 MPLS 私网标签值与该路由同时保存下来。当在对端 PE 上有数据包根据这条路由发往 PE 时，则在 IP 地址前先压上这个私网 MPLS 标签，再进入 MPLS 公网隧道。报文抵达 PE 以后，PE 检查收到的报文就会发现该 MPLS 标签值，根据该标签值，也就能清楚的知道该报文是依照本地哪个 VPN 的路由转发而来，从而决定按照哪个 VPN 的路由表将该数据转发给对应的 VPN 用户。

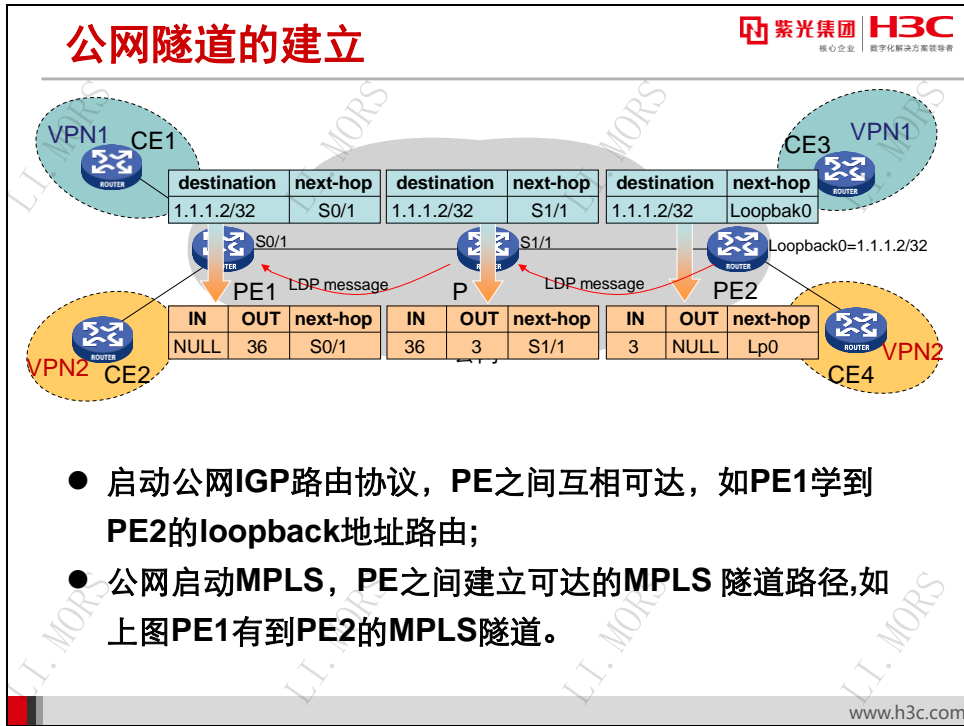
私网数据为了穿越公网，需要进入 MPLS 的公网隧道进行转发，也就是需要在报文的前面压上 MPLS 标签。而按照上文所述报文进入公网隧道前需要先加上的标识报文所属 VPN 的 MPLS 标签，因此就需要在报文前面有两层 MPLS 标签。为了在理解上加以区分，将实现公网隧道时所采用的 MPLS 标签称之为 MPLS 公网标签，而将实现标识报文所属 VPN 的 MPLS 标签称之为 MPLS 私网标签。MPLS 公网标签和私网标签的组成或外观等并没有什么区别，他们的区别在于在报文的转发过程中发挥不同的作用。实现在 IP 报文前压上两层 MPLS 标签依赖于 MPLS 技术本身就考虑支持了标签的嵌套，可见 MPLS 支持标签的嵌套是其可以应用于 VPN 技术的关键。

至此，MP-BGP 协议相对普通 BGP 协议的所有扩展及这些扩展的作用都已经说明，其中 RT 属性、RD 前缀和 MPLS 私网标签是最关键的三个扩展内容，它们分别在私网路由的学习、私网路由的撤销和私网数据的转发过程中发挥着重要的作用。



## 16.6 BGP MPLS VPN基本原理

### 16.6.1 公网隧道建立



BGP MPLS VPN 的实现分为以下四个步骤：

- 公网隧道的建立
- 本地 VPN 的建立
- 私网路由的学习
- 私网数据的转发

前面三个步骤是第四步私网数据转发的基础，首先第一步就是公网隧道的建立。在 BGP MPLS VPN 技术中，选用了 MPLS 技术来建立 VPN 用户的公网隧道。

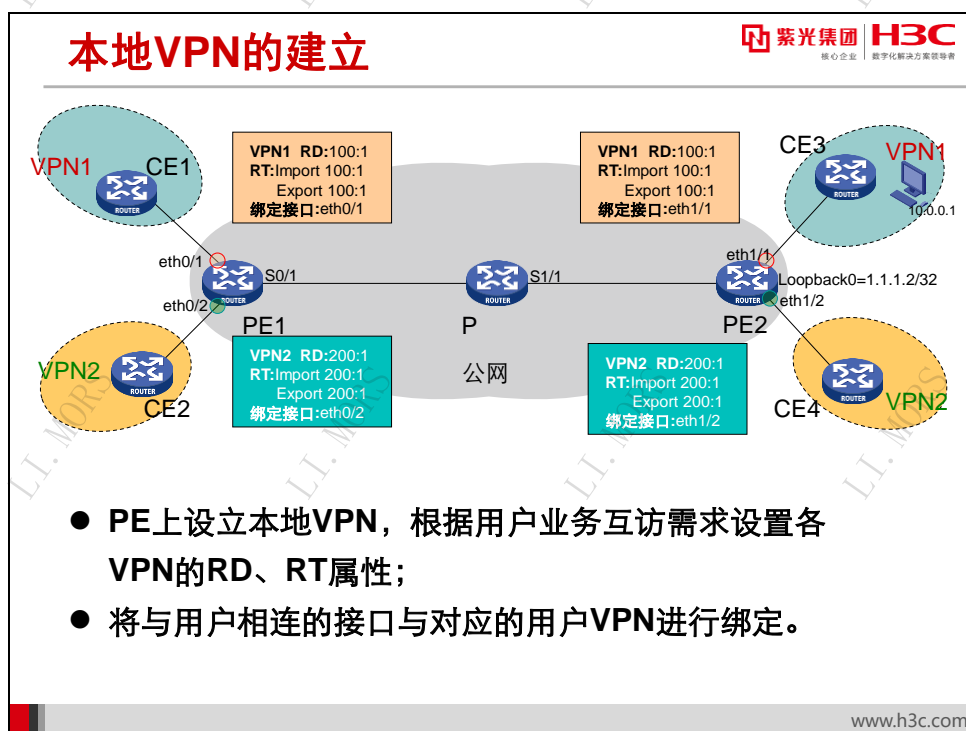
如上图所示，是 BGP MPLS VPN 的典型组网，VPN1 和 VPN2 是接入在公网上的两个 VPN 用户，他们各自拥有两个分部，分别接入在公网的 PE1 和 PE2 设备上。要求 VPN1 的各个分部之间能够互访，也就是上图的 CE1 和 CE3 可以互通，VPN2 的各个分部之间也能够互相访问，即 CE2 和 CE4 可以互通，相反 VPN1 和 VPN2 的用户之间不能互访，且 VPN1 和 VPN2 的用户私网地址空间可能重叠。

为了能让 VPN 的私网数据能够穿越公网进行互访，需要在他们的接入公网设备也就是 PE1 和 PE2 之间建立起 MPLS 隧道，如上图，以从 PE1 访问 PE2 的方向为例，需要建立起抵达 PE2 的 MPLS LSP。



按照 MPLS 技术的原理，首先需要公网上有抵达 PE2 的路由，如上图所示，P 和 PE1 均学习到了 PE2 的 loopback 地址 1.1.1.2/32 的路由。此时如果 PE1、P 和 PE2 均使能了 MPLS 功能和 MPLS 标签分配协议，就会在 PE1、P 和 PE2 上形成针对 1.1.1.2/32 这条路由的标签转发表。此时，如果有数据从 PE1 去访问 PE2 即访问 1.1.1.2，就会按照标签转发表进行转发，在公网中间的 P 设备上，不需要再检查报文 MPLS 标签内部的目的 IP 地址，就可以完成报文的转发，也就建立起了 PE1 访问 PE2 的 MPLS 隧道。当然，从 PE2 访问 PE1 的 MPLS 隧道的建立过程，与上述过程完全对称，这里就不再赘述。

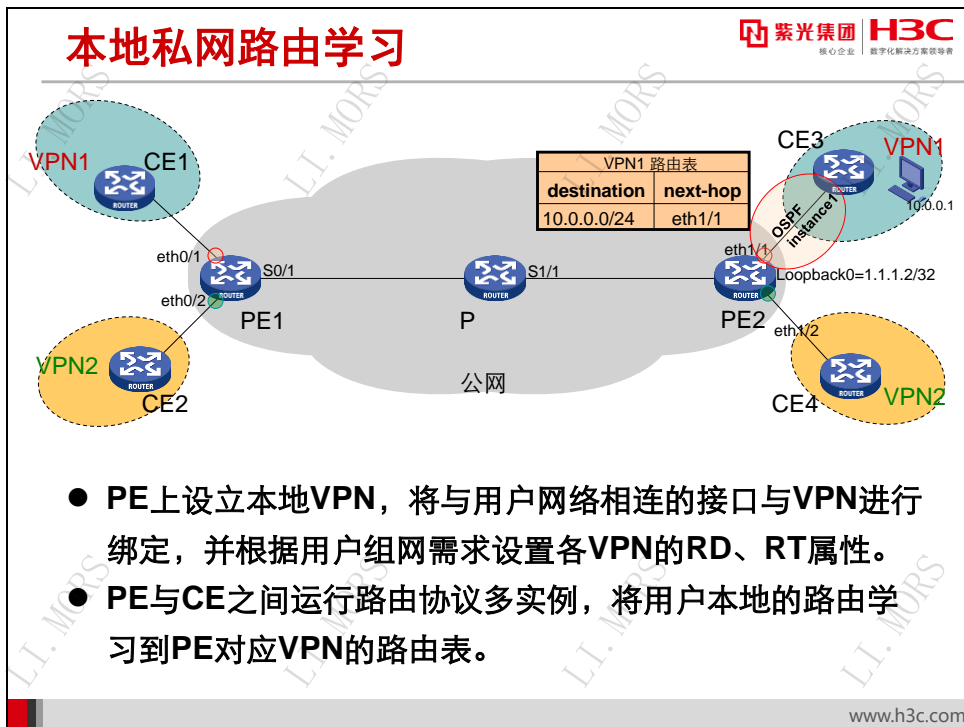
### 16.6.2 本地 VPN 的建立



实现 BGP MPLS VPN 技术的第二步是建立本地的 VPN，也就是在每一台 PE 设备上，根据接入的 VPN 用户的需求，为该 VPN 用户建立一个 VRF，并且完成 PE 设备与其本地的 VPN 用户网络的互通。

本地 VPN 的建立，是 BGP MPLS VPN 规划的关键，在 PE 上设立一个 VPN 需要为该 VPN 设计 RD 和 RT。RD 的设计比较简单，只要确保该 RD 值在本 PE 设备上唯一。而 RT 的设计通常会比较复杂，它决定了用户的互访关系。在上图的组网中，互访关系比较简单，要求 VPN1 用户内部互通，VPN2 用户内部互通，VPN1 和 VPN2 用户隔离。所以可以按照上图来设计这两个 VPN 的 RD 和 RT 值，即在 PE1 和 PE2 上均为 VPN1 用户设计 RD 值 100:1，RT 值 Import Target: 100:1，Export Target: 100:1，而为 VPN2 用户设计 RD 值 200:1，RT 值 Import Target: 200:1，Export Target: 200:1。根据 MP-BGP 的实现原理，可以发现 PE1 和 PE2 上 VPN1 用户的路由可以交互，VPN2 用户的路由也可以交互，相反 VPN1 用户无法接受 VPN2 用户的路由，且 VPN2 用户也无法接受 VPN1 用户的路由，满足用户的互访需求。

在 PE 建立起 VPN 以后, 需要将对应的 VPN 用户接入的接口与该 VPN 进行绑定, 完成绑定以后, 根据多 VRF 技术的原理, 从该接口进入 PE 的报文, 将只能访问对应的 VPN 的路由表进行转发, 而完全无法感知其他 VPN 的路由表, 也就不担心 VPN1 和 VPN2 的地址空间冲突了。



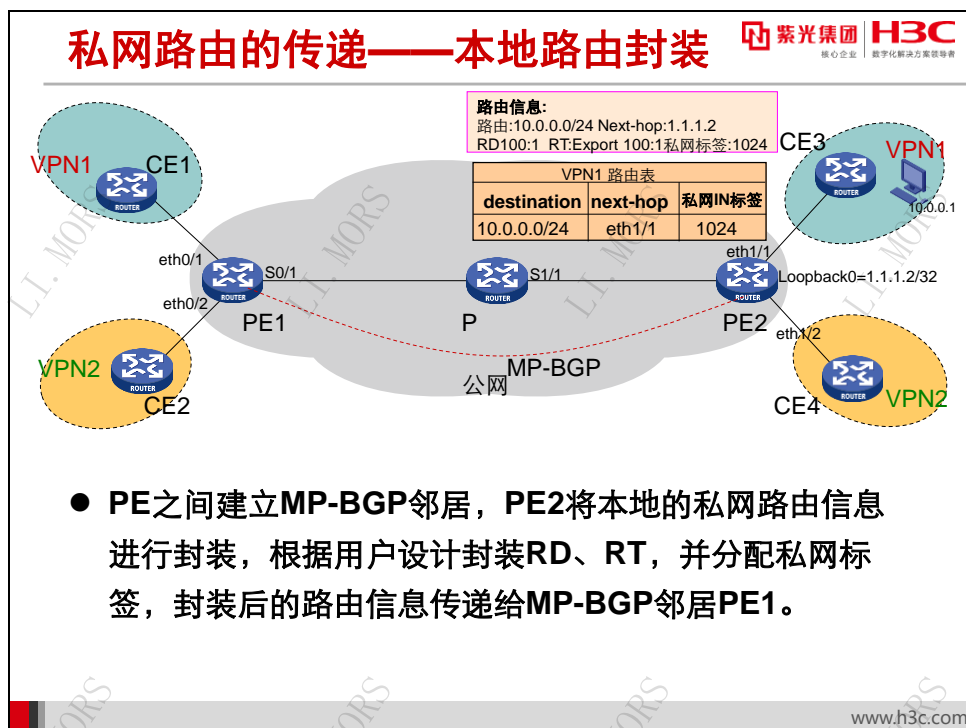
只是将接口与 VPN 进行绑定, 并没有最终完成本地 VPN 的建立, 还需要 PE 设备与本地的 VPN 用户完成私网路由的交互, 这样才能让接入到 PE 上的 VPN 用户知道 PE 是通往其他分部的出口。

如上图所示, PE2 设备需要与 CE3 之间进行路由交互, 路由交互可以采用各种路由协议, 这些路由协议在 PE2 上需要与 VPN1 进行绑定, 这样才能将学习到的路由信息存入 VPN1 的路由表, 不与其他 VPN 相混淆。上图的例子中选用 OSPF 路由协议, 其进程 1 在 PE2 上与 VPN1 进行绑定。当然需要说明的是路由协议的实例就跟进程一样, 只有本地意义, 作为 VPN 的内部用户设备, 如 CE3 并不需要感知到 VPN 的存在, 也不需要运行路由协议的多实例, 只是普通的路由协议就可以。

PE2 和 CE3 之间的路由协议邻居建立成功后, PE2 就可以学习到本地的 VPN1 用户的路由信息, 如上图所示, 在 PE2 的 VPN1 的路由表里面, 出现了 CE3 下连的某网段 10.0.0.1/24 这条路由。

当然所有的 PE 设备与接入它的所有 VPN 用户之间都要进行上述的路由交互的过程, 交互时可以选用各种路由协议, 如静态、OSPF、RIP、IS-IS 等等。

## 16.6.3 私网路由的学习



完成了本地 VPN 的建立，BGP MPLS VPN 实现开始了最关键的一个步骤，也就是私网路由的学习过程，这个过程的主角是 MP-BGP 协议，它的目的是将 PE 设备在本地 VPN 的建立过程中学习到的本地的私网路由信息通过 MP-BGP 协议传递给对端 PE 设备，并且根据用户的 VPN 互访关系的设计，将这些路由信息存放在对端 PE 设备的正确的 VPN 路由表中。

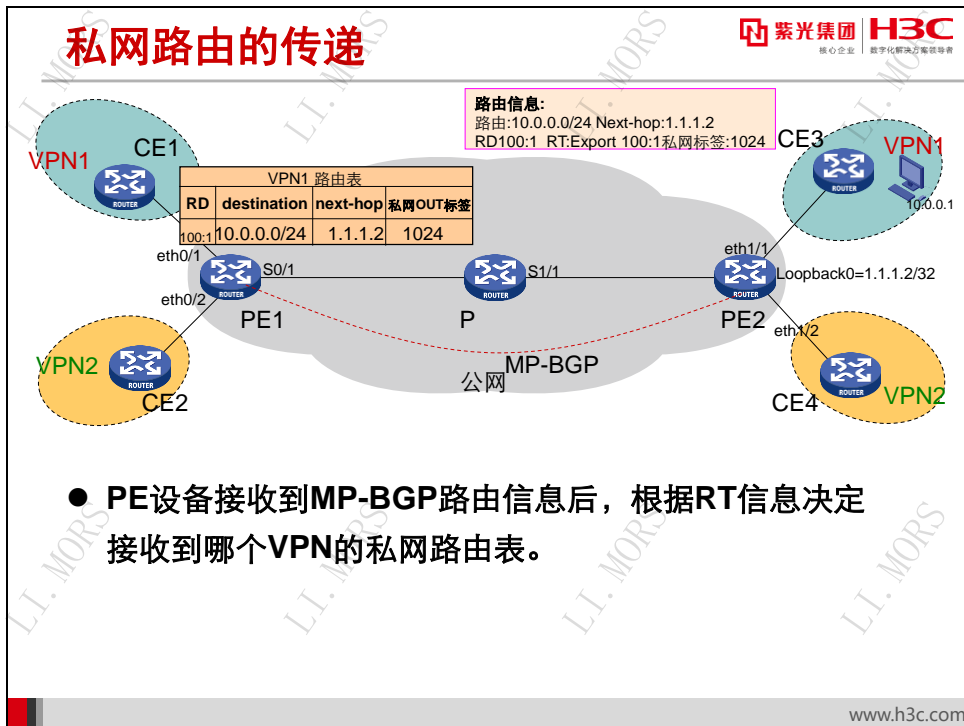
私网路由的学习可以分为两个部分来看，第一步是本地路由的封装，也就是按照 MP-BGP 协议的原理，在本端 PE 上对即将传递给对端的私网路由进行一系列的准备工作，最终封装成一个 MP-BGP 路由更新消息，发送给对端。

如图所示，PE2 将从 CE3 学习到的 10.0.0.1/24 这条路由，引入了 MP-BGP 要求将其发送给 BGP 邻居，这时按照 MP-BGP 原理，形成了一条 MP-BGP 路由更新消息，包含以下信息：

- **VPNv4 的路由前缀：**RD+IPv4 地址/掩码。根据在本地 VPN 建立过程中 RD 的设计，RD 值填上 100:1，路由前缀自然就是 10.0.0.1/24。
- **下一跳地址：**根据 BGP 的原理，当 BGP 发布一条本地的路由时，下一跳地址将填写与对端设备建立 BGP 邻居的地址，通常 IBGP 的邻居会选用本机的 loopback 地址与对端建立邻居，以确保 BGP 邻居的稳定性（loopback 地址不会像普通接口地址那样，因为链路故障，导致接口 down 而不可达），在 BGP MPLS VPN 的组网中，PE 之间的 MP-BGP 邻居都要求采用 loopback 地址来建立，所以这里的下一跳地址将填写 1.1.1.2 即 PE2 的 loopback 地址。
- **RT 属性：**根据 MP-BGP 协议原理，通过 MP-BGP 路由更新消息传送给对端的 RT 属性是本地 RT 设计的 Export Target 值，这里将填写 100:1。

- **私网标签：**根据 MP-BGP 协议原理，需要为发往 BGP 邻居的私网路由分配一个私网标签，这是一个随机值，只要不与为其他路由分配的 MPLS 标签相冲突即可，如上图所示，随机分配一个 MPLS 私网标签值为 1024。

以上这些信息将按照上文所介绍的 MP-BGP 更新消息的格式，组装成一个 MP-BGP 更新消息，这就完成的私网路由的本地封装过程。



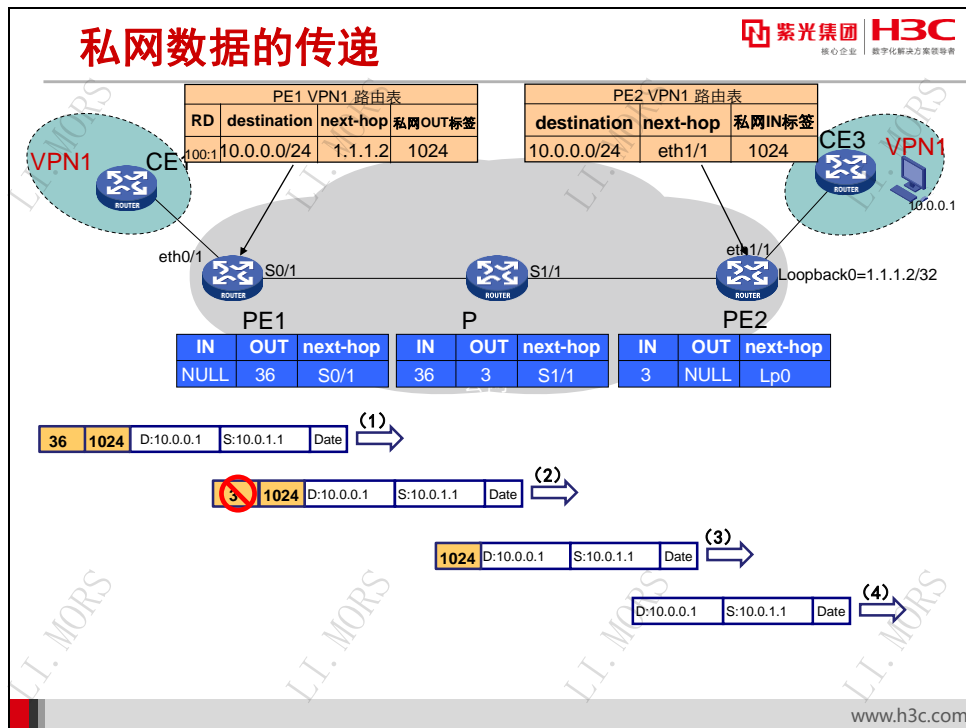
- **PE设备接收到MP-BGP路由信息后，根据RT信息决定接收到哪个VPN的私网路由表。**

完成的私网路由的本地封装后，当 PE 设备之间建立起 MP-BGP 邻居关系，这条路由信息就会发布给对端 PE 设备，如上图所示，10.0.0.1/24 这条路由从 PE2 传递给 PE1，PE1 收到这条路由信息后，将根据 MP-BGP 协议的原理，学习这条私网路由，学习过程将包括以下几个步骤：

- **比较 RT 属性值决定加入哪些 VPN 的路由表：**根据 MP-BGP 协议原理，PE 设备收到 MP-BGP 路由后，检查路由信息所携带 RT 值，与本地各个 VPN 的 RT 的 Import Target 值相比较，当存在交集时，则将该路由信息加入该 VPN 的路由表，如上图的例子，10.0.0.1/24 这条路由将被加入 PE1 的 VPN1 的路由表。
- **记录路由信息：**记录路由信息包括记录该路由的前缀（RD+IPv4 前缀），下一跳信息以及私网标签，如上图所示，10.0.0.1/24 这条路由被记录在 PE1 的 VPN1 的路由表中。
- **发布给本地 VPN：**通过 MP-BGP 学习到的路由信息，需要通过 PE 和本地接入 VPN 的 CE 设备之间的路由交互，发布给本地的 CE 设备，最终的用户才能知道其他分部的路由信息。如上图所示 10.0.0.1/24 这条路由信息将通过 PE1 和 CE1 之间的路由交互，发布给 CE1。

完成了公网隧道的建立，本地 VPN 的建立，私网路由的学习，BGP MPLS VPN 的网就已经搭建完成，此时私网 VPN 用户不同的分部之间就可以互相访问。

#### 16.6.4 私网数据的传递



BGP MPLS VPN 技术原理的最后一部分就是私网数据传递的过程，上图以 CE1 下面的用户 PC1 访问 CE3 下面的用户 PC3 为例，来体验这个过程。

首先，PC1 发出一个访问 10.0.0.1 即 PC3 的报文，该报文抵达 CE1 后，CE1 检查路由表发现下一跳指向 PE1。报文会从 PE1 的 eth0/1 接口进入 PE1 设备。因为 PE1 的 eth0/1 接口与 VPN1 相绑定，该报文将只能查询 VPN1 的私网路由表进行转发，根据报文的目地地址，将匹配到 10.0.0.1/24 这条私网路由，因此需要为该报文加上一个私网标签 1024，并转发给下一跳 1.1.1.2。下一跳 1.1.1.2 并不是一个直接的出口，PE1 需要进行路由的迭代，在 BGP MPLS VPN 的实现中，此处将自动迭代到公网路由表进行转发。因为 PE1 上使能了 MPLS 转发，此时 PE1 发现已经存在一个抵达 1.1.1.2 的标签转发表，该报文需要进行 MPLS 转发，于是按照该标签转发表项，为报文再压上标签 36，并将报文从 S0/1 接口发送出去。

如上图所示报文连续被压了两层 MPLS 标签，先压的标签是 MPLS 私网标签，也称为内层标签，如 1024；后压的标签是 MPLS 公网标签，也称为外层标签，如上图的 36。

报文从 PE1 的 S0/1 接口发出后，将抵达 P 设备，P 设备使能了 MPLS，会发现该报文是一个 MPLS 包，它将根据报文的 MPLS 标签值进行转发。P 设备首先看到的是报文的外层标签，根据外层标签检查标签转发表。P 设备找到对应的表项后，就可以根据表项进行转发，无需再检查报文外层标签内部的其他内容，即既不会去解读报文的私网标签，更不会去查看私网报文的目的 IP 地址。所以 P 设备也不需识别报文的私网标签和私网 IP 地址，这也就是隧道的效果。

上图的例子中只有一台 P 设备，实际组网中可能存在很多台 P 设备，所有 P 设备的数据转发过程都是相同的，都只需要检查报文的外层标签。

当 P 设备根据标签转发表进行报文的转发时，执行 MPLS 的 SWAP 动作，也就是将报文的外层标签值更换成对应的标签转发表的 OUT 标签值，如上图所示 P 设备将报文的外层标签值换成 3。根据 MPLS 倒数第二跳弹出的原理，标签值 3 代表一个特殊含义，也就是该 P 设备是该 LSP 的倒数第二跳设备，它需要将 3 标签弹出，再将报文根据标签转发表项从对应的出接口发出。如上图所示，P 设备将 3 标签弹出后，将报文从 S1/1 接口发出，此时的报文只剩下一层 MPLS 标签，也就是报文的私网标签 1024。

报文抵达 PE2 设备后，PE2 设备发现该报文是一个 MPLS 包，将检查该报文的 MPLS 标签值，发现是 1024，而 PE2 上记录着 1024 是 PE2 为 VPN1 的某一条私网路由分配的标签，此时 PE2 就会将该报文的私网标签值弹出，并根据报文的目的 IP 地址查询 VPN1 的私网路由表进行转发。检查 VPN1 的私网路由表发现，访问 10.0.0.1/24 需从 eth1/1 接口去往 CE3 方可抵达，于是将报文发往 CE3。需要注意的是，此时的报文已经不再携带任何 MPLS 标签，而是一个普通的 IP 包，CE3 只需要进行普通的 IP 转发，报文就可以顺利的抵达最终的目的地 PC3。

以上就是 VPN 用户的私网报文穿越公网进行互相访问的整个过程，当然如果是 PC3 访问 PC1，将是一个与上述过程完全对称的过程，此处不再赘述。

纵观整个 BGP MPLS VPN 的实现过程可以发现，该 VPN 技术有效减轻了 VPN 用户的负担，在整个 VPN 用户的私网内部包括 CE 设备，感知不到 VPN 的存在，也无需维护隧道，而其不同的分部之间就如存在实际的物理线路一样被相连起来。与此同时，这样的—个无需用户操心的 VPN 技术在运营商维护起来也极为方便，因为它的隧道的建立，私网路由的学习都是动态的，运营商只需要根据用户的互访需求为各个 VPN 设计 RD 和 RT 值就可以，而且通过 RT 值的设计，运营商可以满足用户丰富多彩的 VPN 互访关系需求。当然以上的这些都是传统的 VPN 技术所无法比拟的，这也正是 BGP MPLS VPN 技术得以盛行的原因。

## 16.7 本章总结

### 本章总结

- 了解BGP MPLS VPN技术背景
- 理解MPLS技术隧道应用
- 掌握多VRF技术和MP-BGP技术原理
- 理解BGP MPLS VPN实现过程

www.h3c.com

## 16.8 习题和解答

### 16.8.1 习题

1. BGP MPLS VPN 技术与传统 VPN 技术相比，下列描述正确的是（ ）  
A. BGP MPLS VPN 实现隧道的动态建立，无需手工创建 VPN 隧道  
B. BGP MPLS VPN 每增加一个 VPN 用户，需要增加对应的隧道配置  
C. BGP MPLS VPN 私网路由易于控制，可以支持更灵活的 VPN 互访关系  
D. BGP MPLS VPN 解决了本地地址冲突问题，多个 VPN 用户可共享接入设备
2. 下列哪些技术可以作为隧道使用？（ ）  
A. GRE    B. IPSec    C. MPLS    D. BGP
3. 使能多 VRF 技术的路由器，每一个 VRF 将（ ）  
A. 拥有独立的接口                      B. 拥有独立的路由表  
C. 拥有独立的路由协议                D. 拥有独立的 CPU
4. MP-BGP 路由协议相对 BGP 路由协议增加了下列哪些内容（ ）  
A. MP\_REACH\_NLRI 属性                B. MP\_UNREACH\_NLRI 属性  
C. Next-hop 属性                        D. Extended\_Communities 属性
5. 下列关于 RT 的描述正确的是（ ）  
A. RT 包含 Export Target 属性和 Import Target 属性  
B. RT 是配置使能 MP-BGP 路由协议后自动生成的  
C. MP-BGP 路由协议会将本地 VPN RT 的 Import Target 携带在路由信息里面发给 BGP 邻居  
D. 当收到的 MP-BGP 路由中携带的 RT 值与本地 VPN 的 Import Target 属性存在交集时，该路由就会被添加到该 VPN 的路由表
6. 下列关于 RD 的描述错误的是（ ）  
A. RD 在 BGP MPLS VPN 的实现原理中主要用于撤销路由时区分该撤销哪个 VPN 的路由  
B. 从原理上讲不同的 PE 设备在为不同的 VPN 配置 RD 值时是无需考虑是否相同的  
C. 从原理上讲某一台 PE 设备在为不同的 VPN 配置 RD 值时是无需考虑是否相同的



- ### 16.8.2 习题答案

- 416 -

## 第17章 BGP MPLS VPN 配置与故障排除

理解了 BGP MPLS VPN 的基本原理以后，掌握 BGP MPLS VPN 技术的配置将是一件非常容易的事，BGP MPLS VPN 技术的配置思路与其实现原理完全吻合，甚至其故障排查的思路也完全来源于其实现原理。

### 17.1 本章目标

#### 课程目标

● 学习完本课程，您应该能够：

- 了解BGP MPLS VPN主要配置
- 掌握BGP MPLS VPN配置思路与步骤
- 理解BGP MPLS VPN故障排查思路与步骤



www.h3c.com

## 17.2 BGP MPLS VPN的配置思路

### BGP MPLS VPN 配置思路

● BGP MPLS VPN的配置思路与对BGP MPLS VPN技术原理的理解一致，分为以下三个步骤：

- 配置公网隧道
- 配置本地VPN
- 配置MP-BGP

www.h3c.com


BGP MPLS VPN 的配置思路与 BGP MPLS VPN 的原理完全吻合，与 BGP MPLS VPN 的实现过程一样，配置也分为以下三个步骤：

- 1) 配置公网隧道，首先在公网上使能 MPLS，建立公网隧道；
- 2) 配置本地 VPN，其次就是要根据用户 VPN 的互访关系，设计本地 VPN；
- 3) 配置 MP-BGP，最后在 PE 之间建立其 MP-BGP 邻居，传递私网路由。

## 17.3 BGP MPLS VPN配置命令

### 17.3.1 配置公网隧道

### 配置公网隧道

 紫光集团 H3C  
核心企业 数字化解决方案领导者

- 配置本节点的LSR ID:  

```
mpls lsr-id lsr-id
```
- 系统模式下使能MPLS和MPLS LDP  

```
mpls ldp
```
- 接口模式使能MPLS和MPLS LDP  

```
interface interface-type interface-number  
mpls enable  
mpls ldp enable
```

www.h3c.com



在配置公网隧道时，有以下三个关键配置：

- 1) **配置该 LSR 设备的 LSR ID:** LSR ID 的格式是一个 IP 地址，它将用来在 MPLS 网络中标识这台 LSR 设备。所以要求 LSR 设备的 LSR ID 在 MPLS 网络中唯一，通常会选用 LSR 的 loopback 地址作为其 LSR ID。配置方法非常简单，如上图所示，在系统模式下配置。
- 2) **使能 LDP:** 这一步的配置是要将一台普通的路由器，转变成一台可以处理 MPLS 报文，可以进行 MPLS 标签分配的路由器。当然使能的标签分配协议由用户选用的标签分配协议来决定，不一定是 LDP 协议，此处以 LDP 为例，给出配置案例，该部分配置内容也在系统模式下。
- 3) **接口下使能 MPLS 和 LDP:** 具体使能某一接口的 MPLS 报文处理能力和 LDP 标签分配功能。系统模式下使能 MPLS 和 LDP 是接口使能对应功能的前提，只有具体接口使能了 MPLS 和 LDP，该接口才能处理 MPLS 报文。

完成上述配置后，如果网络中已经有对应的某一网段的路由，就能自动形成对应的标签转发表项，从而形成对应的隧道。

## 17.3.2 配置本地 VPN

## 配置本地VPN

 紫光集团  H3C  
核心企业 | 数字化转型领导者

- 创建VPN实例，进入VPN视图：  

**ip vpn-instance *vpn-instance-name***
- 配置RD和RT  

**route-distinguisher *route-distinguisher***  
**vpn-target *vpn-target*<1-8> [ both | export-extcommunity | import-extcommunity ]**
- 配置接口与VPN实例绑定  

**interface *interface-type interface-number***  
**ip binding vpn-instance *vpn-instance-name***
- 配置PE与CE之间的路由实例与VPN绑定，以OSPF为例：  

**ospf [ *process-id* | router-id *router-id* | vpn-instance *vpn-instance-name* ]**

www.h3c.com

配置 BGP MPLS VPN 的第二步是要建立本地 VPN，这部分的配置包含两个重要的部分，第一部分是根据用户 VPN 互访关系的要求，给对应的 VPN 设计 RT、RD 等参数，并在 PE 上配置该 VPN；第二部分是用户接入 PE 的接口与对应的 VPN 进行绑定，并启动路由协议多实例完成 PE 和 CE 之间的本地私网路由交互。


所以这部分的关键配置有如下几步：

- 1) **创建一个 VPN：**其中 VPN 的名称是一个任意字符串，可以根据用户的特点进行命名，建议尽可能考虑能从名称上识别是哪个用户，以方便维护。
- 2) **配置 RD 和 RT：**这部分首先要根据用户互访需求进行规划，规划完成后方可将规划的 RT 和 RD 值配置在该 VPN 视图下。
- 3) **配置接口与 VPN 绑定：**将用户接入的接口与对应的 VPN 进行绑定，方法只需要在对应的接口下配置上图所示的绑定命令。
- 4) **配置 PE 与 CE 之间的路由协议：**其中在 PE 一侧需要将对应的路由实例与对应的 VPN 进行绑定，上图是以 OSPF 在 PE 侧的配置为例，将 OSPF 某一进程与某一 VPN 进行绑定。

## 17.3.3 配置 MP-BGP

## 配置MP-BGP

紫光集团

H3C  
核心企业 | 数字化转型方案领导者

- 进入BGP-VPNv4子地址族视图

ipv4-family vpnv4

- 使能对等体交换BGP-VPNv4路由信息

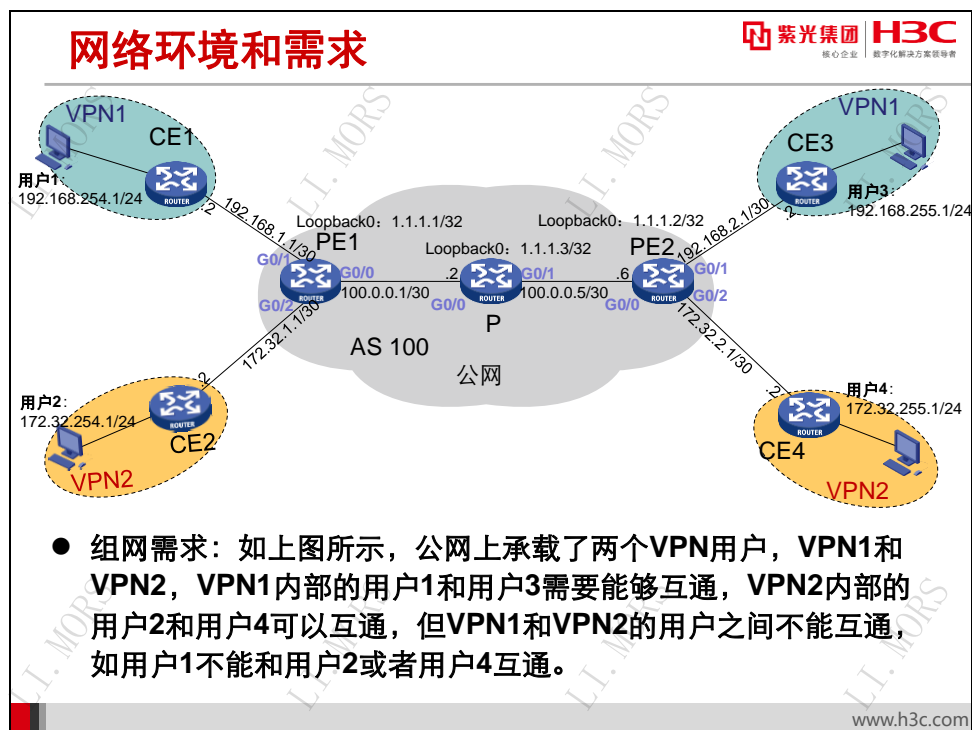
peer { group-name | ip-address } enable

www.h3c.com

配置 BGP MPLS VPN 的最后一步是要在 PE 之间建立器 MP-BGP 邻居，这部分的关键配置只是需要在 PE 之间建立普通 BGP 邻居的基础上，使能他们之间交互 VPNv4 路由的能力即可。配置方法如上图所示，在 BGP 视图下，进入 VPNv4 地址族，并在该地址族视图下，使能需要交互 VPNv4 路由的邻居。

## 17.4 BGP MPLS VPN配置示例

### 17.4.1 网络环境和需求




下文我们将以一个具体的组网案例为例，讲解 BGP MPLS VPN 的配置步骤。上图是 BGP MPLS VPN 的一个典型组网。事实组网中，公网通常存在大量的 P 和 PE 设备，本案例进行了简化，假设公网只有三台设备，PE1、P 和 PE2，其中 PE1 和 PE2 上分别接入了两个用户，用户 1、用户 2 和 用户 3、用户 4。根据各个用户的要求，用户 1 与用户 3 同属 VPN1 需要进行互通；用户 2 与用户 4 同属 VPN2 也需要进行互通；而 VPN1 和 VPN2 完全隔离，不能互访。

## 17.4.2 配置 BGP MPLS VPN 公网隧道

### 配置公网隧道的步骤

- 配置公网隧道分为两步：
  - 配置公网IGP路由协议
  - 配置使能MPLS及MPLS LDP

 紫光集团 | H3C  
核心企业 | 数字化转型方案领导者

[www.h3c.com](http://www.h3c.com)

按照 BGP MPLS VPN 的配置思路，首先在公网上配置 MPLS 公网隧道，该部分配置包括两个部分：

- 1) 首先在公网上使能某种 IGP 路由协议，使得公网设备之间 IP 互通。
- 2) 其次在公网的 P 和 PE 设备上使能 MPLS 和 MPLS LDP 协议，完成 PE 设备之间的公网隧道建立。



## 配置公网隧道步骤一

紫光集团 H3C  
核心企业 数字化转型方案领导者

### ● 配置公网IGP路由协议。

- 目标：使所有PE和P设备可以互相学到32位loopback地址路由。
- IGP路由协议可以选择OSPF、ISIS、甚至静态路由等等。
- 以OSPF为例：PE1、P和PE2设备之间建立OSPF邻居。

```
PE1: router id 1.1.1.1
      ospf 1
        area 0.0.0.0
          network 1.1.1.1 0.0.0.0
          network 100.0.0.1 0.0.0.3
```

```
P: router id 1.1.1.3
    ospf 1
      area 0.0.0.0
        network 1.1.1.3 0.0.0.0
        network 100.0.0.2 0.0.0.3
        network 100.0.0.5 0.0.0.3
```

```
PE2: router id 1.1.1.2
      ospf 1
        area 0.0.0.0
          network 1.1.1.2 0.0.0.0
          network 100.0.0.6 0.0.0.3
```

www.h3c.com

公网上的 IGP 路由协议，可以根据用户的实际需求进行选择，只要能达到以下这个目的，也就是公网上所有的设备可以学习到各个 PE 设备的 Loopback 地址的明细路由。需要注意的是，为了建立 PE 和 PE 之间端到端的隧道，不能将 PE 的 Loopback 地址路由加以聚合。

本文以 OSPF 为例，在公网设备间建立起 OSPF 邻居，并发布 PE 设备的 Loopback 地址的路由，相关配置步骤如上图所示。

## 配置公网隧道步骤二



### ● 配置使能MPLS及MPLS LDP。

→ 目标：所有PE和P设备之间建立LDPsession，建立起到达对端PE的LSP隧道

→ 需要在设备全局模式及公网接口上均使能MPLS及MPLS LDP

PE1:	系统模式:  mpls lsr-id 1.1.1.1 mpls ldp	接口模式:  interface GigabitEthernet0/0 mpls enable mpls ldp enable	
P:	系统模式:  mpls lsr-id 1.1.1.3 mpls ldp	接口模式:  interface GigabitEthernet0/0 mpls enable mpls ldp enable	接口模式:  interface GigabitEthernet0/1 mpls enable mpls ldp enable
PE2:	系统模式:  mpls lsr-id 1.1.1.2 mpls ldp	接口模式:  interface GigabitEthernet0/0 mpls enable mpls ldp enable	

www.h3c.com

建立公网隧道的第二步就是要在公网设备及其公网接口上使能 MPLS 及 MPLS LDP 协议。所以配置分为两个部分，一是在 P 和 PE 的系统模式下全局使能 MPLS LDP；二是在 P 和 PE 设备的公网接口上使能 MPLS 及 MPLS LDP。在系统模式下使能 MPLS LDP 之前，还需要为该公网设备配置 MPLS 的 LSR ID，如上图所示，为了确保 LSR ID 唯一，通常选用该公网设备的 Loopback 地址作为 LSR ID。

## 17.4.3 配置 BGP MPLS VPN 本地 VPN

## 配置本地VPN的步骤

● 配置本地VPN分为三步：

- 配置VPN，按照用户互访需求配置VPN的RD和RT
- 配置私网接口与VPN的绑定
- 配置PE和CE之间的路由协议

www.h3c.com

配置 BGP MPLS VPN 的第二步是要配置本地 VPN，该部分配置通常分为以下三个步骤来完成：

- 1) **创建 VPN：**按照用户互访需求创建 VPN 并配置该 VPN 的 RD 和 RT 值；
- 2) **配置私网接口与 VPN 的绑定：**即将 PE 上用户接入的私网接口与对应的 VPN 进行绑定；
- 3) **配置 PE 和 CE 之间的路由协议：**实现 PE 与本地 VPN 用户之间的路由交互。

## 配置本地VPN步骤一

### ● 配置VPN，按照用户互访需求配置VPN的RD和RT

→ 此步配置需要仔细分析用户互访需求，设计各PE上每个VPN的RD和RT属性；

→ 根据本案例的组网需求设计做如下配置：

```
PE1:
ip vpn-instance vpn1
 route-distinguisher 100:1
 vpn-target 100:1 export-extcommunity
 vpn-target 100:1 import-extcommunity
#
ip vpn-instance vpn2
 route-distinguisher 100:2
 vpn-target 100:2 export-extcommunity
 vpn-target 100:2 import-extcommunity
```

```
PE2:
ip vpn-instance vpn1
 route-distinguisher 100:1
 vpn-target 100:1 export-extcommunity
 vpn-target 100:1 import-extcommunity
#
ip vpn-instance vpn2
 route-distinguisher 100:2
 vpn-target 100:2 export-extcommunity
 vpn-target 100:2 import-extcommunity
```

首先是在 PE 上创建 VPN，该步骤是 BGP MPLS VPN 配置中最与用户业务相关的配置步骤，需要经过充分的用户需求调查并根据用户要求进行网络规划，才能够完成相应的配置。

按照本文给出的组网需要，在 PE1 和 PE2 上都需要创建两个 VPN，即 VPN1 和 VPN2。其中 PE1 上的 VPN1 需要和 PE2 上的 VPN1 互通，于是可以设计 PE1 和 PE2 上的 VPN1 的 RT 参数 import target 和 export target 属性值均为 100:1；同时 PE1 上的 VPN2 需要和 PE2 上的 VPN2 互通，于是可以设计 PE1 和 PE2 上的 VPN2 的 RT 参数 import target 和 export target 属性值均为 200:1。在这样的设计下，因为 RT 值的关系，PE1 上的 VPN1 和 PE2 上的 VPN1 路由可以互相学习，可以互通；同时 PE1 上的 VPN1 和 PE2 上的 VPN2 路由也可以交互，也可以互通；相反 VPN1 和 VPN2 路由将不能互相学习，VPN1 用户和 VPN2 用户不能互通，均满足用户的互访关系要求。

关于 RD 值的设计相对简单，只要同一台 PE 上不同的 VPN 的 RD 值不相同即可，这里为了设计更加的清晰，在 PE1 和 PE2 上 VPN1 的 RD 值均设计为 100:1，而 VPN2 的 RD 值均设计为 200:1。

根据如上的 VPN 规划，可以完成对应的 VPN 建立配置，具体配置步骤如上图所示。

## 配置本地VPN步骤二

紫光集团 H3C  
核心企业 数字化转型领导者

### ● 配置私网接口与VPN的绑定:

- 将与对应用户相连的接口与对应的VPN进行绑定;
- 对端CE设备无需感知VPN的存在, 仅需做普通的接口地址等配置。

```
PE1: interface GigabitEthernet0/1
      ip binding vpn-instance vpn1
      ip address 192.168.1.1 255.255.255.252
      #
      interface GigabitEthernet0/2
      ip binding vpn-instance vpn2
      ip address 172.32.1.1 255.255.255.252
```

```
PE2: interface GigabitEthernet0/1
      ip binding vpn-instance vpn1
      ip address 192.168.2.1 255.255.255.252
      #
      interface GigabitEthernet0/2
      ip binding vpn-instance vpn2
      ip address 172.32.2.1 255.255.255.252
```

www.h3c.com

配置本地VPN的第二步就是要将PE的私网接口与对应VPN进行绑定,如以本文的案例,PE1的eth0/1接口与VPN1的用户相连,则需要将该接口与VPN1进行绑定。如上图所示,绑定的方法就是在该接口下配置绑定VPN的命令。需要提醒的是,在配置将接口与VPN绑定时,接口下原有的IP地址等配置将会丢失,需要重新进行配置,正确的配置步骤应该是先进行接口与VPN的绑定,在完成接口的IP地址等配置。

接口与VPN绑定的配置只需要在PE设备上进行,与PE相连的CE设备无法感知到VPN的存在,也无需做任何配置,只是一个普通的接口。

## 配置本地VPN步骤三

紫光集团 H3C  
核心企业 数字化转型领导者

### ● 配置PE和CE之间的路由协议

- 目标：PE设备能学习到本端相连的用户路由，如PE1学习到用户1的路由，并能与用户1互通；
- PE和相连的CE设备之间运行路由协议，其中PE设备上的路由协议需要与对应的VPN进行绑定，也就是运行路由协议多实例；
- 路由协议可以选择OSPF、ISIS、EBGP、RIP、静态等；
- 本例中以OSPF为例，且以PE1和CE1及CE2之间的配置为例：

PE1:	<pre>ospf 11 vpn-instance vpn1 area 0.0.0.0 network 192.168.1.0 0.0.0.3 # ospf 12 vpn-instance vpn2 area 0.0.0.0 network 172.32.1.0 0.0.0.3</pre>
CE1:	<pre>ospf 11 area 0.0.0.0 network 192.168.1.0 0.0.0.3 network 192.168.254.1 0.0.0.0</pre>
CE2:	<pre>ospf 12 area 0.0.0.0 network 172.32.1.0 0.0.0.3 network 172.32.254.1 0.0.0.0</pre>

www.h3c.com

配置本地 VPN 的最后一个步骤，就是要在 PE 和 CE 之间运行某种路由协议，使得 PE 和 CE 之间可以交互路由信息，完成 PE 和本地 VPN 用户的路由交互。如以本文的案例，PE1 需要能够与其本地相连的用户 1 可以互通。

在 PE 和 CE 之间运行路由协议时，PE 侧相对特殊，其运行的路由协议需要与对应的 VPN 进行绑定，也就是要运行路由协议的多实例。如在 PE1 上，当他需要与 CE1 交互路由时，需要运行某种路由协议的一个实例，该实例需要与 VPN1 进行绑定，这样从 CE1 学习到的路由才会记录到 VPN1 的路由表。而在 CE1 上感知不到 VPN 的存在，只要运行普通的路由协议即可。

PE 和 CE 之间的路由协议可以任意选择，如 OSPF、ISIS、EBGP、RIP、甚至静态等等，本文是以 OSPF 为例。如上图所示，PE1 为了能分别与 CE1 和 CE2 交互路由，在 PE1 上运行了两个 OSPF 实例，分别与 VPN1 和 VPN2 进行绑定，而在 CE1 和 CE2 上只要运行普通的 OSPF，PE1 和 CE1 及 CE2 分别建立起 OSPF 邻居，交互路由。PE1 通过 OSPF11 学习到的路由，记录在 VPN1 的路由表，通过 OSPF12 学习到的路由则记录在 VPN2 的路由表。

在 PE2 上的配置与 PE1 上相当，对应的配置命令本文不再赘述。

## 17.4.4 配置 MP-BGP

## 配置MP-BGP的步骤

● 配置MP-BGP分为两步：

- 配置PE之间MP-BGP邻居
- 配置本地VPN路由与MP-BGP之间的路由引入引出

www.h3c.com

配置 BGP MPLS VPN 的最后一个步骤是配置 MP-BGP。这部分的目标是要能在 PE 之间建立起 MP-BGP 邻居，交互私网路由，一般分为以下两个步骤：

- 1) PE 之间配置建立 MP-BGP 邻居；
- 2) 配置本地 VPN 路由和 MP-BGP 之间的路由引入引出。

## 配置MP-BGP步骤一

紫光集团 H3C  
核心企业 数字化转型方案领导者

### ● 配置PE之间MP-BGP邻居

→ PE之间建立起MP-BGP传递BGP MPLS VPN的私网路由，也就是VPNv4路由

→ 建立MP-BGP邻居的方法是，进入BGP VPNv4 族，使能对应的BGP邻居

PE1:

```
bgp 100
peer 1.1.1.2 as-number 100
peer 1.1.1.2 connect-interface LoopBack0
address-family vpnv4
peer 1.1.1.2 enable
```

PE2:

```
bgp 100
peer 1.1.1.1 as-number 100
peer 1.1.1.1 connect-interface LoopBack0
address-family vpnv4
peer 1.1.1.1 enable
```

www.h3c.com

PE 之间需要使能 BGP 协议传递 VPNv4 路由的能力，也就是配置 PE 之间的 MP-BGP 邻居关系，这是配置 MP-BGP 的关键步骤。

建立 MP-BGP 邻居关系的方法是在 BGP 视图下进入 VPNv4 地址族，并使能该邻居。具体配置步骤如上图所示。



## 配置 MP-BGP 步骤二

紫光集团 H3C  
核心企业 数字化转型方案领导者

### ● 配置本地 VPN 路由与 MP-BGP 之间的路由相互引入

→ 本地的私网路由需要引入 BGP，才能通过 BGP 传给远端 CE；通过 BGP 学习到的远端私网路由需要引入本地 PE 与 CE 之间的路由协议，才能传递给本地的 CE。

→ 以 PE1 为例，PE2 与之对称不做重复：

```

PE1:
bgp 100
  ipv4-family vpn-instance vpn1
    address-family ipv4 unicast
    import-route ospf 11
  #
  ipv4-family vpn-instance vpn2
    address-family ipv4 unicast
    import-route ospf 12
  #
ospf 11 vpn-instance vpn1
  import-route bgp
#
ospf 12 vpn-instance vpn2
  import-route bgp
  
```

www.h3c.com

配置 MP-BGP 的最后一步尤为关键，也常常是配置 BGP MPLS VPN 时最容易被遗漏的，那就是配置本地 VPN 私网路由和 MP-BGP 路由的互相引入。

PE 和 CE 之间采用某种路由协议的多实例互相交互路由，这部分路由将学习到 PE 的对应的 VPN 路由表中，但这些路由并不会自动的被 MP-BGP 路由发布给对端 PE，需要在 MP-BGP 协议中加以引入才行。如上图所示，需要在 BGP 路由协议中，进入对应的 VPN 实例视图，在 ipv4 单播地址族视图下引入 PE 和 CE 之间运行的路由实例的路由，以本文为例就是对应的 OSPF 实例。

相反 MP-BGP 协议从远端 PE 学习到的私网路由，也会存放到对应的 VPN 的路由表，但这部分路由并不会自动的通过 PE 和 CE 之间运行的路由协议发布给 CE 设备，也需要在对应的路由协议多实例中引入 BGP 路由才行。如上图所示，各个路由协议多实例中都要配置引入 BGP 路由协议，以本文为例就是在 PE 和 CE 之间运行的 OSPF 实例中引入 BGP。

## 17.5 BGP MPLS VPN故障排查

### 17.5.1 BGP MPLS VPN 故障排查思路

**BGP MPLS VPN故障排查思路**

紫光集团 H3C  
核心企业 数字化转型领导者

- **BGP MPLS VPN的故障排查的思路与BGP MPLS VPN的原理也是统一的，当BGP MPLS VPN的网络出现两个私网用户之间无法互通，可以按照下面的思路进行排查：**
  - 排查公网隧道是否存在
  - 排查本地VPN建立是否符合要求
  - 排查MP-BGP私网路由传递是否正确

www.h3c.com

BGP MPLS VPN 的故障排查思路也是来源于 BGP MPLS VPN 技术的原理，当 BGP MPLS VPN 的网络出现故障，导致私网用户之间无法互访时，通常可以按照下面的思路来检查问题所在：

- 1) 首先检查公网隧道的情况，确认两个私网用户所接入的 PE 之间的公网隧道是否存在；
- 2) 排查 PE 设备上的本地 VPN 建立是否正确；
- 3) 最后排查两 PE 之间的 MP-BGP 邻居建立情况是否正常。

## 17.5.2 BGP MPLS VPN 故障排查步骤

### 公网隧道故障排查步骤

紫光集团 H3C  
核心企业 数字化转型解决方案领导者

- 排查公网隧道故障分为以下三步
  - 检查公网路由学习是否正确
  - 检查公网设备之间的MPLS LDP邻居关系是否正常
  - 检查到达对端PE的loopback地址的公网隧道是否存在

www.h3c.com

按照 BGP MPLS VPN 的故障排查思路，首先是要检查 PE 之间的公网隧道建立是否正确，该部分检查过程主要包括以下三个步骤：

- 1) 首先检查公网上的 IGP 路由学习是否正确；
- 2) 其次确认所有的公网设备都使能了 MPLS 和 MPLS LDP 协议，并排查公网设备之间的 MPLS LDP 邻居关系是否正常；
- 3) 最后是在 PE 设备上检查是否存在到达对端 PE Loopback 接口地址的 LSP，也就是 MPLS 公网隧道。

## 公网隧道故障排查步骤一

紫光集团 H3C  
核心企业 数字化转型领导者

### ● 检查公网路由学习是否正确

- 根据所选择的公网IGP路由协议，检查公网设备之间的IGP路由邻居关系是否正确；
- 在PE上检查是否存在到达对端PE的loopback地址的32位掩码的明细路由，以PE1为例：

PE1:

```
[PE1]dis ip routing-table
```

Destinations : 16 Routes : 16

Destination/Mask	Proto	Pre	Cost	NextHop	Interface
0.0.0.0/32	Direct	0	0	127.0.0.1	InLoop0
1.1.1.1/32	Direct	0	0	127.0.0.1	InLoop0
1.1.1.2/32	O_INTRA	10	2	100.0.0.2	GE0/0
1.1.1.3/32	O_INTRA	10	1	100.0.0.2	GE0/0

www.h3c.com

公网 IGP 路由学习正常是建立 MPLS 隧道的大前提，无论公网选择哪一种路由协议，对于 BGP MPLS VPN 的组网，只要在 PE 上可以学习到抵达对端 PE 的 Loopback 接口地址的明细路由。如上图所示，可以通过检查路由表的命令检查 PE1 上的路由，确认是否存在 1.1.1.2/32 这条路由，而这条路由正是 PE2 的 Loopback 接口地址的路由。当然相反在 PE2 上也要检查有没有到 PE1 的 Loopback 接口地址的明细路由，即 1.1.1.1/32 这条路由。

当互相都已经拥有对端的路由时，可以通过“ping”命令再次确认一下 PE1 和 PE2 之间是否互相可达。但是仍然需要强调说明的是，仅仅互相可达并不能达到 BGP MPLS VPN 组网的需求。在 BGP MPLS VPN 组网中，不能对 PE 的 Loopback 地址的路由进行聚合，因为一旦聚合，PE 之间的 LSP 就不连贯，也就无法建立 PE 之间端到端的隧道，而隧道一旦在公网中的 P 设备上中断，私网数据将暴露在公网 P 设备上，而 P 设备无法识别私网数据，报文将被丢弃。所以要求 PE 学习到的对端 PE 的 Loopback 接口地址路由一定没有经过聚合的 32 掩码的明细路由，如上图所示。

## 公网隧道故障排查步骤二

紫光集团 H3C  
核心企业 数字化转型领导者

### ● 检查公网设备之间的MPLS LDP邻居关系是否正常

- LDP邻居建立完成后，正确的状态应该处于Operational；
- 如果不能到达Operational状态，则应进一步确认LDP配置，或深入排查LDP邻居建立过程的故障所在。

```
PE1: [PE1]dis mpls ldp peer
Total number of peers: 1
Peer LDP ID      State      Role      GR      MD5      KA Sent/Rcvd
1.1.1.3:0        Operational Passive    Off      Off      420/425
```

www.h3c.com

确认 IGP 正确后，就需要检查所有公网设备的 MPLS 及 MPLS LDP 协议的使能情况，每台设备都需要在系统和公网接口模式下使能 MPLS 及 MPLS LDP 协议，使能的结果可以通过检查公网设备间的 LDP 邻居状况进行确认。

MPLS LDP 的邻居关系建立完成后，最终的状态应该是 Operational，如上图所示，可以通过“display mpls ldp peer”命令检查 LDP 邻居状态。公网上，每一台设备都应该与他相邻的公网设备建立 LDP 邻居，如上图所示 PE1 应该与 P 设备之间建立起 LDP 邻居。而如果检查 P 设备上的 LDP 邻居状况，应该看到它与 PE1 及 PE2 都建立了 LDP 邻居。

如果 LDP 邻居状态未能达到 Operational，故障原因就能确定在该公网设备或其相邻设备的 MPLS LDP 配置上，可进一步确认他们的配置是否正确，或者也可按照 MPLS LDP 协议邻居建立的步骤，深入排查不能建立邻居的原因，该部分本文不作介绍。

## 公网隧道故障排查步骤三



### ● 检查公网隧道是否存在

- 如果公网IGP和LDP均正常，PE之间的应建立起到达对方loopback地址的MPLS隧道；
- 隧道是单向的，需要在每台PE上分别查询

PE1:	[PE1]dis mpls ldp lsp			
	Status Flags: * - stale, L - liberal, B - backup			
	FECs: 3	Ingress: 2	Transit: 2	Egress: 1
	FEC	In/Out Label	NextHop	OutInterface
	1.1.1.1/32	3/-	-/1150(L)	
PE2:	[PE2]dis mpls ldp lsp			
	Status Flags: * - stale, L - liberal, B - backup			
	FECs: 3	Ingress: 2	Transit: 2	Egress: 1
	FEC	In/Out Label	NextHop	OutInterface
	1.1.1.1/32	-/1150	100.0.0.2	GE0/0

www.h3c.com

如果 LDP 邻居建立也正确，在公网隧道的排查中，还要做最后一步确认，那就是确认 PE 之间的 LSP 是否存在，也就是确认 PE 之间的公网隧道是否已经建成。MPLS 隧道是单向的，需要在每一台 PE 上确认是否有到达对端 PE 的隧道。如上图所示，分别在 PE1 和 PE2 上检查抵达对端 PE 的 LSP。对应的表项，就是抵达对端 PE 的标签转发表，可以看到标签转发表的组成部分：OUT 标签、IN 标签、出接口等等。

LSP 是从 PE 抵达对端 PE 的一个标签转发路径，可以根据该标签转发路径沿路检查各台公网设备上的 LSP 状况，确认对应的标签转发表的标签值，是否符合 MPLS 标签分配原理，当然这个一般很少出现错误，重点还是确认沿路所有的公网设备上都存在这该 LSP 对应的标签转发表项。

确认了公网隧道的存在，关于 BGP MPLS VPN 的公网隧道的故障排查就已经完成，如果私网用户之间仍然无法互通，需要按照下文进一步排查 BGP MPLS VPN 的本地 VPN 及 MP-BGP 的情况。

## 17.5.3 排查本地 VPN 故障的步骤

## 排查本地VPN故障

紫光集团 H3C  
核心企业 数字化转型方案领导者

● 排查本地VPN故障分为两步

- 检查确认对应VPN的私网路由邻居建立是否正常；
- 检查PE与本地CE之间的路由学习是否正确；

www.h3c.com

排查 BGP MPLS VPN 的第二个步骤是检查本地 VPN 的建立是否正确，该部分通常分为以下三个步骤。

- 1) 确认本地 VPN 的设计是否符合用户互访要求，RT 的规划是否正确；
- 2) 检查 PE 和 CE 之间运行的路由协议邻居状况是否正常；
- 3) 检查 PE 是否正确学习到了本地 VPN 用户的私网路由。

## 排查本地VPN故障(续)

紫光集团 H3C  
核心企业 数字化转型领导者

### ● 检查确认对应VPN的私网路由邻居建立是否正常；

- 首先在PE上查看与对应用户相连的接口状态应该处于UP并与对应的VPN实例绑定；
- 按照本例如果PE与CE之间采用OSPF路由协议，可查看对应的OSPF实例路由邻居是否建立成功：

```
PE1: [PE1]dis ospf 11 peer
```

OSPF Process 11 with Router ID 192.168.1.1  
Neighbor Brief Information

Area: 0.0.0.0	Router ID	Address	Pri	Dead-Time	State	Interface
	192.168.254.1	192.168.1.2	1	32	Full/BDR	GE0/1

www.h3c.com

确认 PE 上 VPN 的规划是否正确，只需要查看各个的 VPN 的 RT 配置，是否满足用户的互访关系描述，该部分是一个核实配置的过程，本文不再列出详细检查过程。

确认本地 VPN 设计完全正确后，下一步是需要确认 PE 和 CE 之间的路由协议邻居状况，当然首先要确认 PE 上与 VPN 用户相连的接口是否与对应的 VPN 绑定，且接口状态 UP，在这个前提下，再去查看 PE 和 CE 之间运行的路由协议邻居是否已经建立成功。

按照本文采用的配置案例，PE 和 CE 之间采用 OSPF 路由协议，在 PE 上就需要检查该 VPN 对应的 OSPF 路由实例邻居状况，如上图所示，检查 VPN1 所对应的 OSPF 实例 ospf 11 的邻居状况，可以看出 PE 与下连的 CE 设备 OSPF 邻居状态已经 FULL，表示邻居状况正常。



## 排查本地VPN故障(续)

紫光集团 H3C  
核心企业 数字化转型领航者

### ● 检查PE与本地CE之间的路由学习是否正确；

- 在PE上检查是否学习到本地CE及用户的路由信息
- 无论PE与CE之间采用何种路由协议，重点在于PE能学习到本地用户的路由信息，以PE1的VPN1为例，PE1可以学习到用户1的路由，如下：

PE1: [PE1]dis ip routing-table vpn-instance vpn1

Destinations : 14		Routes : 14			
Destination/Mask	Proto	Pre	Cost	NextHop	Interface
0.0.0.0/32	Direct	0	0	127.0.0.1	InLoop0
127.0.0.0/8	Direct	0	0	127.0.0.1	InLoop0
127.0.0.0/32	Direct	0	0	127.0.0.1	InLoop0
127.0.0.1/32	Direct	0	0	127.0.0.1	InLoop0
127.255.255.255/32	Direct	0	0	127.0.0.1	InLoop0
192.168.1.0/30	Direct	0	0	192.168.1.1	GE0/1
192.168.1.0/32	Direct	0	0	192.168.1.1	GE0/1
192.168.1.1/32	Direct	0	0	127.0.0.1	InLoop0
192.168.1.3/32	Direct	0	0	192.168.1.1	GE0/1
192.168.254.1/32	O INTRA	10	1	192.168.1.2	GE0/1
192.168.255.1/32	BGP	255	2	1.1.1.2	GE0/0

www.h3c.com

PE 和 CE 之间的邻居状况建立正常后，就需要确认 PE 是否已经学习到本地 VPN 用户的路由。也就是检查 PE 上对应 VPN 的私网路由表，确认是否有到本地 VPN 用户的路由。检查的命令如上图所示，与普通的检查路由表的方法相比，需要在后面加上 vpn-instance 的参数。

如果 PE 上已经学习到本地 VPN 用户的私网路由，表示 PE 的本地 VPN 的建立完全正确，如果私网用户之间仍然无法互通，需要按照下文检查 MP-BGP 的情况。

## 17.5.4 排查 MP-BGP 故障的步骤

## 排查MP-BGP私网路由传递故障

紫光集团 H3C  
核心企业 数字化转型方案领导者

- 排查MP-BGP私网路由传递故障分为以下三步：
  - 检查PE之间MP-BGP邻居是否建立成功
  - 检查PE是否学习到远端用户的私网路由
  - 检查CE是否学习到远端用户的私网路由

www.h3c.com

在 BGP MPLS VPN 组网中，如果 PE 之间公网隧道正常，且 PE 上本地 VPN 的状况也正常的，那么只剩下最后一个可能存在的故障点，那就是 PE 间 MP-BGP 的私网路由交互是否存在问题。

排查 MP-BGP 的故障主要分为以下三个步骤：

- 1) 检查 PE 之间 MP-BGP 的邻居是否建立成功；
- 2) 检查 PE 是否通过 MP-BGP 路由协议学习到了对端私网用户的路由；
- 3) 检查 CE 是否学习到了对端私网用户的路由。

## 排查MP-BGP私网路由传递故障(续)

紫光集团 H3C  
核心企业 数字化转型领导者

### ● 检查PE之间MP-BGP邻居是否建立成功

→ 如果MP-BGP邻居未能正常建立，检查对应的MP-BGP配置是否正确

PE1:

[PE1]dis bgp peer vpnv4

BGP local router ID: 1.1.1.1

Local AS number: 100

Total number of peers: 1

Peers in established state: 1

\* - Dynamically created peer

Peer	AS	MsgRcvd	MsgSent	OutQ	PrefRcv	Up/Down	State
1.1.1.2	100	140	135	0	2	01:50:59	Established

www.h3c.com

在 BGP MPLS VPN 组网中，PE 之间需要建立 MP-BGP 路由邻居。检查方法如上图所示，也就是查看 BGP 的 VPNv4 邻居状况，正常的状态应该是 Established。如果不能达到 Established，需要确认两 PE 上是否都将对端 PE 在 BGP VPNv4 视图下面使能。

## 排查MP-BGP私网路由传递故障(续)

紫光集团 H3C  
核心企业 数字化转型方案领导者

### ● 检查PE是否学习到远端用户的私网路由

- 在PE上检查是否学习到本地CE及用户的路由信息；
- PE之间建立起MP-BGP邻居后具备了互相传递私网路由的能力；
- 但此时需要将本地的私网路由引入BGP，BGP才能将这些引入的路由传递给对端，结果可以在PE上检查到对端用户的私网路由。

PE1: [PE1]dis ip routing-table vpn-instance vpn1

Destinations : 14      Routes : 14

Destination/Mask	Proto	Pre	Cost	NextHop	Interface
0.0.0.0/32	Direct	0	0	127.0.0.1	InLoop0
127.0.0.0/8	Direct	0	0	127.0.0.1	InLoop0
127.0.0.0/32	Direct	0	0	127.0.0.1	InLoop0
127.0.0.1/32	Direct	0	0	127.0.0.1	InLoop0
127.255.255.255/32	Direct	0	0	127.0.0.1	InLoop0
192.168.1.0/30	Direct	0	0	192.168.1.1	GE0/1
192.168.1.0/32	Direct	0	0	192.168.1.1	GE0/1
192.168.1.1/32	Direct	0	0	127.0.0.1	InLoop0
192.168.1.3/32	Direct	0	0	192.168.1.1	GE0/1
192.168.255.1/32	O_INTRA	10	1	192.168.1.2	GE0/1
192.168.255.1/32	BGP	255	2	1.1.1.2	GE0/0

www.h3c.com

前文已经说明过，BGP MPLS VPN 配置过程中最容易疏漏的点，就是将 PE 上本地 VPN 的私网路由引入到 BGP 中，这样 BGP 路由协议才会将该路由发布给它的 BGP 邻居。完成引入后，在 PE 上查看对应的私网路由表，可以发现通过 BGP 路由协议从远端 PE 学习过来的私网路由信息。上图是以 PE1 为例，其 VPN1 的私网路由表中已经通过 BGP 路由协议学习到了 PE2 侧的私网路由。同时还应该在 PE2 上做相应的检查，以确认 PE1 设备也做了正确的引入配置。

## 排查MP-BGP私网路由传递故障(续)

紫光集团 H3C  
核心企业 数字化转型领航者

### ● 检查CE是否学习到远端用户的私网路由

- PE通过BGP学习到远端用户的私网路由后，需要将该BGP路由引入到PE与CE之间的路由协议，CE才能学习到远端用户的私网路由；
- 以CE1为例，可以查看到用户3的路由信息：

CE1: [CE1]dis ip routing-table

Destinations : 17      Routes : 17

Destination/Mask	Proto	Pre	Cost	NextHop	Interface
192.168.1.0/30	Direct	0	0	192.168.1.2	GE0/0
192.168.1.0/32	Direct	0	0	192.168.1.2	GE0/0
192.168.1.2/32	Direct	0	0	127.0.0.1	InLoop0
192.168.1.3/32	Direct	0	0	192.168.1.2	GE0/0
192.168.254.0/24	Direct	0	0	192.168.254.1	Loop1
192.168.254.0/32	Direct	0	0	192.168.254.1	Loop1
192.168.254.1/32	Direct	0	0	127.0.0.1	InLoop0
192.168.254.255/32	Direct	0	0	192.168.254.1	Loop1
192.168.255.1/32	O INTER	10	3	192.168.1.1	GE0/0

www.h3c.com

在 PE 已经学习到远端 VPN 的私网路由后，需要完成另一个重要的配置步骤，那就是将通过 MP-BGP 路由协议学习来的远端 VPN 的私网路由引入到 PE 和 CE 之间运行的路由实例中，以通过该路由实例将路由发布给 PE 本地的 CE 设备，这样 CE 才能学习到远端 VPN 的路由。

完成这一步引入操作后，在 CE 设备上检查路由表应该可以看到远端 VPN 用户的路由。如上图是以 CE1 为例，在该设备上已经可以查看到远端 VPN 用户 PC2 的路由信息。

完成了公网隧道、本地 VPN、和 MP-BGP 的所有检查步骤，BGP MPLS VPN 的网络基本已经正常，这个就是 BGP MPLS VPN 的故障排查方法，可见仍然与 BGP MPLS VPN 的实现原理保持一致。

## 17.6 本章总结

### 本章总结

- 掌握BGP MPLS VPN的配置和故障排查思路
- 熟悉BGP MPLS VPN的配置步骤
- 了解BGP MPLS VPN的故障排查步骤

www.h3c.com

## 17.7 习题和解答

### 17.7.1 习题

1. 下列哪种路由协议不能作为 BGP MPLS VPN 的 IGP 路由协议？（ ）  
A. OSPF    B. ISIS    C. RIP    D. 静态路由    E. 以上都能
2. 下列哪些接口需要使能 MPLS 协议？（ ）  
A. PE 设备的公网接口    B. CE 设备的上联接口  
C. P 设备的公网接口    D. PE 设备的私网接口
3. 下列哪条命令可以检查 MPLS LDP 的 session 是否建立完成？（ ）  
A. display mpls ldp peer    B. display mpls ldp lsp  
C. display mpls ldp session    D. display mpls ldp
4. 在 PE 上创建一个 VPN 需要为该 VPN 配置下列哪些内容？（ ）  
A. VPN 的名称    B. RD 值    C. RT 列表    D. 私网标签
5. 应使用如下哪个命令查看两台 PE 设备之间的 MP-BGP 邻居是否已经建立成功？（ ）  
A. display bgp peer    B. display bgp vpnv4 peer  
C. display mp-bgp peer    D. display bgp peer vpnv4

### 17.7.2 习题答案

1. E
2. AC
3. A
4. ABC
5. D