



**HUAWEI NetEngine20E-X6 高端业务路由器
V600R003C00**

特性描述-MPLS

文档版本 01

发布日期 2011-05-15

版权所有 © 华为技术有限公司 2011。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、复制本档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本档内容会不定期进行更新。除非另有约定，本档仅作为使用指导，本档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编： 518129

网址： <http://www.huawei.com>

客户服务邮箱： support@huawei.com

客户服务电话： 0755-28560000 4008302118

客户服务传真： 0755-28560111

前言

读者对象

本文档针对 MPLS 特性，从简介、原理描述和应用三个方面介绍了分流和引流特性。

本文档与其它类型手册相结合，便于读者深入掌握 MPLS 特性的实现原理。

本文档主要适用于以下工程师：

- 网络规划工程师
- 调测工程师
- 数据配置工程师
- 系统维护工程师

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
 危险	以本标志开始的文本表示有高度潜在危险，如果不能避免，会导致人员死亡或严重伤害。
 警告	以本标志开始的文本表示有中度或低度潜在危险，如果不能避免，可能导致人员轻微或中等伤害。
 注意	以本标志开始的文本表示有潜在风险，如果忽视这些文本，可能导致设备损坏、数据丢失、设备性能降低或不可预知的结果。
 窍门	以本标志开始的文本能帮助您解决某个问题或节省您的时间。
 说明	以本标志开始的文本是正文的附加信息，是对正文的强调和补充。

修订记录

修改记录累积了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 01 (2011-05-15)

第一次正式发布。

目录

前言.....	iii
1 MPLS 基础.....	1-1
1.1 介绍.....	1-2
1.2 参考标准和协议.....	1-2
1.3 原理描述.....	1-3
1.3.1 MPLS 基本概念.....	1-3
1.3.2 LSP 的建立.....	1-9
1.3.3 MPLS 转发.....	1-11
1.3.4 MPLS Ping/Traceroute.....	1-16
1.4 应用.....	1-17
1.4.1 基于 MPLS 的 VPN.....	1-18
1.4.2 策略路由到 LSP.....	1-19
1.5 术语与缩略语.....	1-19
2 MPLS LDP.....	2-1
2.1 介绍.....	2-2
2.2 参考标准和协议.....	2-2
2.3 原理描述.....	2-2
2.3.1 LDP 基本概念.....	2-3
2.3.2 LDP 会话.....	2-4
2.3.3 标签的发布和管理.....	2-6
2.3.4 LDP LSP 的建立.....	2-8
2.3.5 LDP 跨域扩展.....	2-8
2.3.6 LDP Outbound 策略和 Inbound 策略.....	2-10
2.3.7 LDP-IGP 联动.....	2-10
2.3.8 LDP 和静态路由同步.....	2-12
2.3.9 LDP GR.....	2-13
2.3.10 LDP NSR.....	2-14
2.3.11 LDP FRR.....	2-14
2.3.12 LDP MTU.....	2-16
2.3.13 LDP MD5.....	2-16
2.3.14 LDP 认证.....	2-17
2.3.15 LDP 为 BGP 分标签.....	2-18

2.3.16 LDP over TE.....	2-20
2.3.17 LDP over GRE.....	2-21
2.3.18 LDP GTSM.....	2-21
2.3.19 LDP 本远端会话共存.....	2-22
2.3.20 LDP 为所有 Peer 分标签.....	2-23
2.4 术语与缩略语.....	2-24
3 MPLS TE.....	3-1
3.1 介绍.....	3-2
3.2 参考标准和协议.....	3-5
3.3 原理描述.....	3-7
3.3.1 RSVP-TE 协议.....	3-7
3.3.2 Make-before-break	3-9
3.3.3 P2MP RSVP-TE.....	3-10
3.3.4 自动带宽调整.....	3-12
3.3.5 重优化.....	3-13
3.3.6 TE FRR.....	3-14
3.3.7 SRLG.....	3-17
3.3.8 CR-LSP 备份.....	3-18
3.3.9 DS-TE.....	3-23
3.3.10 TE 隧道保护组.....	3-33
3.3.11 BFD For TE CR-LSP.....	3-35
3.3.12 BFD for TE Tunnel.....	3-37
3.3.13 RSVP 认证.....	3-37
3.3.14 RSVP GR.....	3-38
3.3.15 RSVP 摘要刷新.....	3-39
3.3.16 RSVP Hello.....	3-40
3.3.17 BFD for RSVP.....	3-40
3.3.18 TE-LSP 配置模板.....	3-41
3.3.19 MPLS LSR-ID 的多区域发布.....	3-42
3.4 术语与缩略语.....	3-44
4 MPLS OAM.....	4-1
4.1 介绍.....	4-2
4.2 参考标准和协议.....	4-2
4.3 原理描述.....	4-3
4.3.1 MPLS OAM 检测.....	4-3
4.3.2 反向通道.....	4-5
4.3.3 MPLS OAM 自动协议.....	4-5
4.3.4 保护倒换.....	4-5
4.4 术语与缩略语.....	4-6

插图目录

图 1-1 MPLS 网络结构.....	1-4
图 1-2 MPLS LSP.....	1-4
图 1-3 MPLS 报文首部结构.....	1-5
图 1-4 标签在分组中的封装位置.....	1-5
图 1-5 标签栈.....	1-7
图 1-6 上游和下游概念.....	1-8
图 1-7 标签分发示意图.....	1-8
图 1-8 MPLS 体系结构示意图.....	1-9
图 1-9 LSP 建立.....	1-10
图 1-10 Tunnel ID 的结构.....	1-11
图 1-11 MPLS 标签分发和报文转发.....	1-13
图 1-12 MPLS 转发流程.....	1-14
图 1-13 Uniform 模式下 TTL 的处理.....	1-15
图 1-14 Pipe 模式下 TTL 的处理.....	1-16
图 1-15 MPLS 网络.....	1-17
图 1-16 基于 MPLS 的 VPN.....	1-18
图 1-17 策略路由到 LSP 的应用.....	1-19
图 2-1 LDP Session 建立过程.....	2-5
图 2-2 DU 方式.....	2-6
图 2-3 DoD 方式.....	2-7
图 2-4 LDP 跨域扩展组网拓扑.....	2-9
图 2-5 LDP-IGP 联动需要解决的回切问题.....	2-11
图 2-6 LDP 和静态路由同步解决的 LSP 切换场景.....	2-12
图 2-7 LDP GR 实现原理.....	2-14
图 2-8 LDP FRR 典型应用场景-三角型拓扑.....	2-15
图 2-9 LDP FRR 典型应用场景-口字型拓扑.....	2-16
图 2-10 L3VPN 跨域 OptionC 组网拓扑.....	2-18
图 2-11 运营商的运营商组网拓扑.....	2-19
图 2-12 LDP over TE 组网拓扑.....	2-20
图 2-13 LDP over GRE 拓扑.....	2-21
图 2-14 LDP GTSM 组网拓扑.....	2-22
图 2-15 LDP 本远端会话共存组网拓扑.....	2-23
图 2-16 LDP 为所有 Peer 分标签组网拓扑.....	2-24

图 3-1 传统路由的困难.....	3-2
图 3-2 严格显式路径.....	3-4
图 3-3 松散显式路径.....	3-4
图 3-4 LSP 建立过程图.....	3-8
图 3-5 Make-before-break 基本原理.....	3-9
图 3-6 P2MP RSVP-TE 隧道建立示意图.....	3-11
图 3-7 P2MP RSVP-TE 隧道 FRR 保护示意图.....	3-12
图 3-8 链路保护示意图.....	3-14
图 3-9 节点保护示意图.....	3-15
图 3-10 TE FRR 示意图.....	3-16
图 3-11 SRLG 示意图.....	3-18
图 3-12 逃生路径示意图.....	3-19
图 3-13 MAM 模型示意图.....	3-26
图 3-14 RDM 模型示意图.....	3-26
图 3-15 同一 VPN 不同业务接入同一条 LSP 组网.....	3-29
图 3-16 VPN 与非 VPN 业务接入同一条隧道组网.....	3-30
图 3-17 VPN 与非 VPN 业务接入分别接入一条隧道组网.....	3-30
图 3-18 隧道备份与快速重路由结合使用应用场景.....	3-31
图 3-19 HQoS 调度.....	3-32
图 3-20 N:1 保护模式图.....	3-34
图 3-21 隧道保护组的原理图.....	3-34
图 3-22 BFD 检测示意图.....	3-35
图 3-23 BFD 会话切换前后示意图.....	3-36
图 3-24 BFD for TE 部署.....	3-36
图 3-25 通过 GR Path 消息和 RecoveryPath 消息重建示意图.....	3-39
图 3-26 Hello 握手机制.....	3-40
图 3-27 BFD for RSVP 示意图.....	3-41
图 3-28 ABR 同时作为两个区域中的 Tunnel 接口终点的示意图.....	3-43
图 3-29 ABR 同时向两个区域发布 MPLS LSR-ID 的区域内部路由的示意图.....	3-43
图 4-1 MPLS OAM 连通性示意图.....	4-4
图 4-2 MPLS OAM 隧道保护示意图.....	4-6

表格目录

表 1-1 特殊标签.....	1-6
表 1-2 Tunnel ID 各字段含义描述.....	1-11
表 2-1 LSRA 的路由表项信息和 FEC 携带的路由信息.....	2-9
表 3-1 热备份与普通备份的区别.....	3-19
表 3-2 CR-LSP 备份与隧道保护组的区别.....	3-22
表 3-3 IETF 模式和 Non-IETF 模式的区别.....	3-27
表 3-4 DS-TE 模式切换.....	3-27
表 3-5 缺省的 TE-Class 映射表.....	3-28
表 3-6 隧道保护中的 DS-TE.....	3-31

1 MPLS 基础

关于本章

- 1.1 介绍
- 1.2 参考标准和协议
- 1.3 原理描述
- 1.4 应用
- 1.5 术语与缩略语

1.1 介绍

MPLS 的起源

90 年代中期，基于 IP 技术的 Internet 快速普及。但由于硬件技术存在限制，基于最长匹配算法的 IP 技术必须使用软件查找路由，转发性能低下，因此 IP 技术的转发性能成为当时限制网络发展的瓶颈。

为了适应网络的发展，ATM（Asynchronous Transfer Mode）技术应运而生。ATM 采用定长标签（即信元），并且只需要维护比路由表规模小得多的标签表，能够提供比 IP 路由方式高得多的转发性能。然而，ATM 协议相对复杂，且 ATM 网络部署成本高，这使得 ATM 技术很难普及。

传统的 IP 技术简单，且部署成本低。如何结合 IP 与 ATM 的优点成为当时热门话题。多协议标签交换技术 MPLS（Multiprotocol Label Switching）就是在这种背景下产生的。

MPLS 最初是为了提高路由器的转发速度而提出的。与传统 IP 路由方式相比，它在数据转发时，只在网络边缘分析 IP 报文头，而不用在每一跳都分析 IP 报文头，节约了处理时间。

随着 ASIC（Application Specific Integrated Circuit）技术的发展，路由查找速度已经不是阻碍网络发展的瓶颈。这使得 MPLS 在提高转发速度方面不再具备明显的优势。但是 MPLS 支持多层标签和转发平面面向连接的特性，使其在 VPN（Virtual Private Network）、流量工程、QoS（Quality of Service）等方面得到广泛应用。

MPLS 概述

MPLS 位于 TCP/IP 协议栈中的链路层和网络层之间，用于向 IP 层提供连接服务，同时又从链路层得到服务。MPLS 以标签交换替代 IP 转发。标签是一个短而定长的、只具有本地意义的连接标识符，与 ATM 的 VPI/VCI 以及 Frame Relay 的 DLCI 类似。标签封装在链路层和网络层之间。

MPLS 不局限于任何特定的链路层协议，能够使用任意二层介质传输网络分组。

MPLS 起源于 IPv4（Internet Protocol version 4），其核心技术可扩展到多种网络协议，包括 IPv6（Internet Protocol version 6）、IPX（Internet Packet Exchange）、Appletalk、DECnet、CLNP（Connectionless Network Protocol）等。MPLS 中的“Multiprotocol”指的就是支持多种网络协议。

由此可见，MPLS 并不是一种业务或者应用，它实际上是一种隧道技术。这种技术不仅支持多种高层协议与业务，而且在一定程度上可以保证信息传输的安全性。

1.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC 3031	Multiprotocol Label Switching Architecture	-
RFC 3036	LDP Specification	-

文档	描述	备注
RFC 3032	MPLS Label Stack Encoding	-
RFC 3443	Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks	-
RFC 3034	Use of Label Switching on Frame Relay Networks Specification	-
RFC 2702	Requirements for Traffic Engineering Over MPLS	-
RFC 3209	RSVP-TE: Extensions to RSVP for LSP Tunnels	-
RFC 2547	BGP/MPLS VPNs	-

1.3 原理描述

1.3.1 MPLS 基本概念

1.3.2 LSP 的建立

1.3.3 MPLS 转发

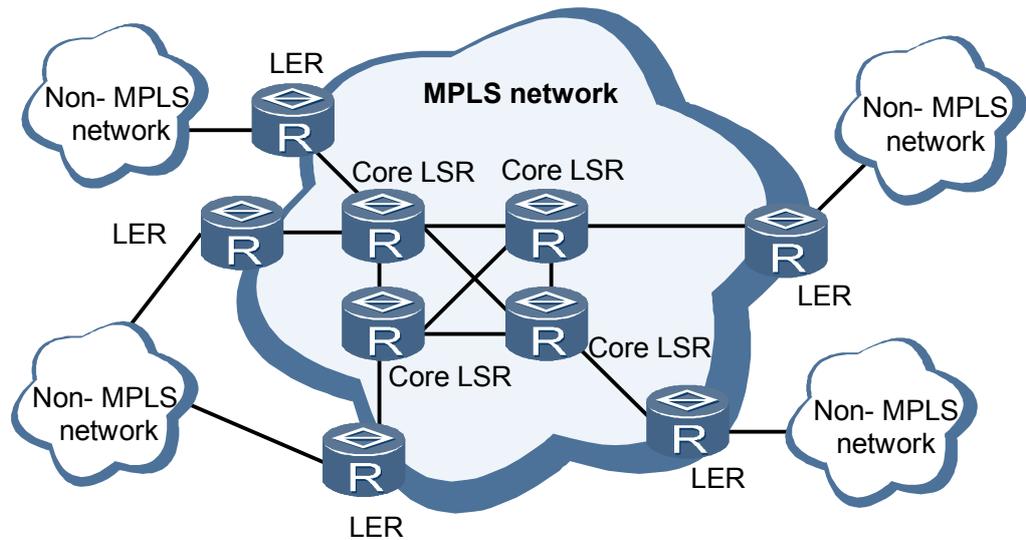
1.3.4 MPLS Ping/Traceroute

1.3.1 MPLS 基本概念

MPLS 网络结构

MPLS 网络的典型结构如图 1-1，MPLS 网络的基本组成单元是标签交换路由器 LSR (Label Switching Router)，由 LSR 构成的网络区域称为 MPLS 域 (MPLS Domain)。位于 MPLS 域边缘、连接其它网络的 LSR 称为边缘路由器 LER (Label Edge Router)，区域内部的 LSR 称为核心 LSR (Core LSR)。如果一个 LSR 有一个或多个不运行 MPLS 的相邻节点，那么该 LSR 就是 LER。如果一个 LSR 的相邻节点都运行 MPLS，则该 LSR 就是核心 LSR。

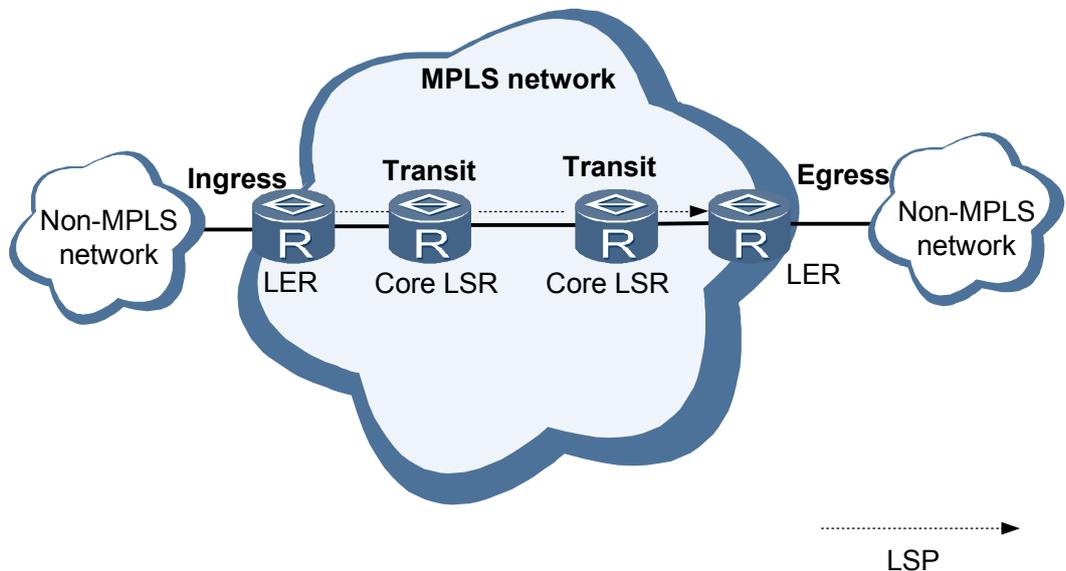
图 1-1 MPLS 网络结构



MPLS 基于标签进行转发。IP 包进入 MPLS 网络时，MPLS 入口的 LER 分析 IP 包的内容并且为这些 IP 包添加合适的标签，所有 MPLS 网络中的节点都是依据标签来转发数据的。当该 IP 包离开 MPLS 网络时，标签由出口边缘路由器删除。

IP 包在 MPLS 网络中经过的路径称为标签交换路径 LSP (Label Switched Path)。LSP 是一个单向路径，与数据流的方向一致。

图 1-2 MPLS LSP



LSP 的起始节点称为入节点 (Ingress)；位于 LSP 中间的节点称为中间节点 (Transit)；LSP 的末节点称为出节点 (Egress)。一条 LSP 可以有 0 个、1 个或多个中间节点，但有且只有一个入节点和出节点。

转发等价类

转发等价类 FEC (Forwarding Equivalence Class) 是一组具有某些共性的数据流的集合。这些数据流在转发过程中被 LSR 以相同方式处理。

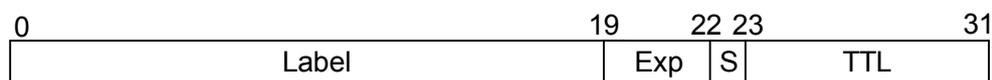
FEC 可以根据地址、业务类型、QoS 等要素进行划分。例如，在传统的采用最长匹配算法的 IP 转发中，到同一条路由的所有报文就是一个转发等价类。

标签

标签 (Label) 是一个短而定长的、只具有本地意义的标识符，用于唯一标识一个分组所属的 FEC。在某些情况下，例如要进行负载分担，对应一个 FEC 可能会有多个入标签，但是一台路由器上，一个标签只能代表一个 FEC。标签与 ATM 的 VPI/VCI 以及 Frame Relay 的 DLCI 类似，是一种连接标识符。

标签长度为 4 个字节，封装结构如 [图 1-3](#) 所示。

图 1-3 MPLS 报文首部结构



标签共有 4 个域：

- Label: 20bit, 标签值域。
- Exp: 3bit, 用于扩展。现在通常用做 CoS (Class of Service), 其作用与 Ethernet802.1p 的作用类似。
- S: 1bit, 栈底标识。MPLS 支持多层标签, 即标签嵌套。S 值为 1 时表明为最底层标签。
- TTL: 8bit, 和 IP 分组中的 TTL (Time To Live) 意义相同。

标签封装在链路层和网络层之间。这样, 标签能够被任意的链路层所支持。标签在分组中的封装位置如 [图 1-4](#) 所示。

图 1-4 标签在分组中的封装位置



标签空间

标签空间就是指标签的取值范围。NE20E-X6 实现中, 标签空间划分如下：

- 0 ~ 15: 特殊标签。特殊标签的详细介绍请参见 [表 1-1](#)。
- 16 ~ 1023: 静态 LSP 和静态 CR-LSP 共享的标签空间。
- 1024 以上: LDP、RSVP-TE、MP-BGP 等动态信令协议的标签空间。

动态信令协议的标签空间不是共享的，而是独立且连续的，互不影响。各个动态信令协议的标签空间由设备的许可证文件定义。如果要修改标签空间，需要修改许可证文件。

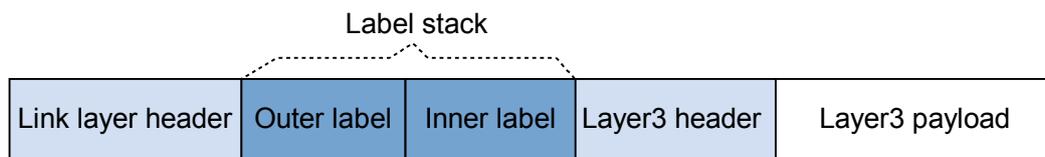
表 1-1 特殊标签

标签值	含义	描述
0	IPv4 Explicit NULL Label	表示该标签必须被弹出，且报文的转发必须基于 IPv4。如果出节点分配给倒数第二跳节点的标签值为 0，则倒数第二跳 LSR 需要将值为 0 的标签正常压入报文标签值顶部，转发给最后一跳。最后一跳发现报文携带的标签值为 0，则将标签弹出。0 标签只有出现在栈底时才有效。
1	Router Alert Label	只有出现在非栈底时才有效。类似于 IP 报文的“Router Alert Option”字段，节点收到 Router Alert Label 时，需要将其送往本地软件模块进一步处理。实际报文转发由下一层标签决定。如果报文需要继续转发，则节点需要将 Router Alert Label 压回标签栈顶。
2	IPv6 Explicit NULL Label	表示该标签必须被弹出，且报文的转发必须基于 IPv6。如果出节点分配给倒数第二跳节点的标签值为 2，则倒数第二跳节点需要将值为 2 的标签正常压入报文标签值顶部，转发给最后一跳。最后一跳发现报文携带的标签值为 2，则直接将标签弹出。2 标签只有出现在栈底时才有效。
3	Implicit NULL Label	倒数第二跳 LSR 进行标签交换时，如果发现交换后的标签值为 3，则将标签弹出，并将报文发给下最后一跳。最后一跳收到该报文直接进行 IP 转发或下一层标签转发。
4 ~ 13	保留	-
14	OAM Router Alert Label	MPLS OAM (Operation Administration & Maintenance) 通过发送 OAM 报文检测和通告 LSP 故障。OAM 报文使用 MPLS 承载。OAM 报文对于 Transit LSR 和倒数第二跳 LSR (penultimate LSR) 是透明的。
15	保留	-

标签栈

标签栈 (Label stack) 是指标签的排序集合。MPLS 报文支持同时携带多个标签，靠近二层首部的标签称为栈顶标签或外层标签；靠近 IP 首部的标签称为栈底标签，或内层标签。理论上，MPLS 标签可以无限嵌套。

图 1-5 标签栈



标签栈按后进先出（Last In First Out）方式组织标签，从栈顶开始处理标签。

标签操作类型

标签的操作类型包括标签压入（Push）、标签交换（Swap）和标签弹出（Pop），它们是标签转发的基本动作，是标签转发信息表的组成部分。

- **Push:** 指当 IP 报文进入 MPLS 域时，MPLS 边界设备在报文二层首部和 IP 首部之间插入一个新标签；或者 MPLS 中间设备根据需要，在标签栈顶增加一个新的标签（即标签嵌套封装）。
- **Swap:** 当报文在 MPLS 域内转发时，根据标签转发表，用下一跳分配的标签，替换 MPLS 报文的栈顶标签。
- **Pop:** 当报文离开 MPLS 域时，将 MPLS 报文的标签去掉；或者 MPLS 倒数第二跳节点处去掉栈顶标签，减少标签栈中的标签数目。

倒数第二跳弹出

在最后一跳节点，标签已经没有使用价值。这种情况下，可以利用倒数第二跳弹出特性 PHP（Penultimate Hop Popping），在倒数第二跳节点处将标签弹出，减少最后一跳的负担。最后一跳节点直接进行 IP 转发或者下一层标签转发。

PHP 在 Egress 节点上配置。支持 PHP 的 Egress 节点分配给倒数第二跳节点的标签只有一种：

标签值 3：表示隐式空标签（implicit-null），这个值不会出现在标签栈中。当一个 LSR 发现自己被分配了隐式空标签时，它并不用这个值替代栈顶原来的标签，而是直接执行 Pop 操作。Egress 节点直接进行 IP 转发或下一层标签转发。

标签交换路由器

标签交换路由器 LSR（Label Switching Router）是指可以进行 MPLS 标签交换和报文转发的网络设备，也称为 MPLS 节点。LSR 是 MPLS 网络中的基本元素，所有 LSR 都支持 MPLS 协议。

LER

位于 MPLS 域边缘的 LSR 称为 LER（Label Edge Router）。如果一个 LSR 有一个不运行 MPLS 的相邻节点，那么该 LSR 就是 LER。

LER 负责对进入 MPLS 域的报文划分 FEC，并为这些 FEC 压入标签，进行 MPLS 转发。当报文离开 MPLS 域时弹出标签，恢复成原来的报文，再进行相应的转发。

标签交换路径

一个转发等价类在 MPLS 网络中经过的路径称为标签交换路径 LSP (Label Switched Path)。

LSP 在功能上与 ATM 和 Frame Relay 的虚电路相同，是从入口到出口的一个单向路径。

入节点、中间节点和出节点

标签交换路径 LSP 是一个单向路径，LSP 中的 LSR 可以分为：

- 入节点 (Ingress)：LSP 的起始节点，一条 LSP 只能有一个 Ingress。
Ingress 的主要功能是给报文压入一个新的标签，封装成 MPLS 报文进行转发。
- 中间节点 (Transit)：LSP 的中间节点，一条 LSP 可能有多个 Transit。
Transit 的主要功能是查找标签转发信息表，通过标签交换完成 MPLS 报文的转发。
- 出节点 (Egress)：LSP 的末节点，一条 LSP 只能有一个 Egress。
Egress 的主要功能是弹出标签，恢复成原来的报文进行相应的转发。

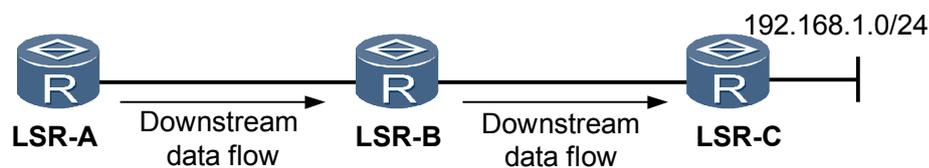
上游和下游

根据数据传送的方向，LSR 可以分为上游和下游。

- 上游：以指定的 LSR 为视角，根据数据传送的方向，所有往本 LSR 发送 MPLS 报文的 LSR 都可以称为上游 LSR。
- 下游：以指定的 LSR 为视角，根据数据传送的方向，本 LSR 将 MPLS 报文发送到的所有下一跳 LSR 都可以称为下游 LSR。

如图 1-6 所示，对于发往 192.168.1.0/24 的数据流来说，LSR-A 是 LSR-B 的上游节点，LSR-B 是 LSR-A 的下游节点。同理，LSR-B 是 LSR-C 上游节点。LSR-C 是 LSR-B 的下游节点。

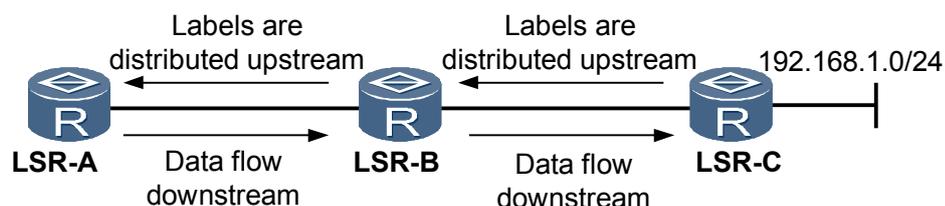
图 1-6 上游和下游概念



标签分发

将目的地址相同的分组报文划分为一个 FEC，然后从 MPLS 标签资源池中取出一个标签，分配给这个 FEC。LSR 记录该标签和 FEC 的对应关系，并将该对应关系封装成消息报文，通告给上游的 LSR，这个过程称为标签分发。

图 1-7 标签分发示意图



如图 1-7，LSR-B 和 LSR-C 对去往 192.168.1.0/24 的报文划分为一个 FEC，然后为该 FEC 分配标签，并向上游通告。因此标签是由下游分配的。

标签发布协议

标签发布协议是 MPLS 的控制协议（也可称为信令协议），负责 FEC 的分类、标签的分发以及 LSP 的建立和维护等一系列操作。

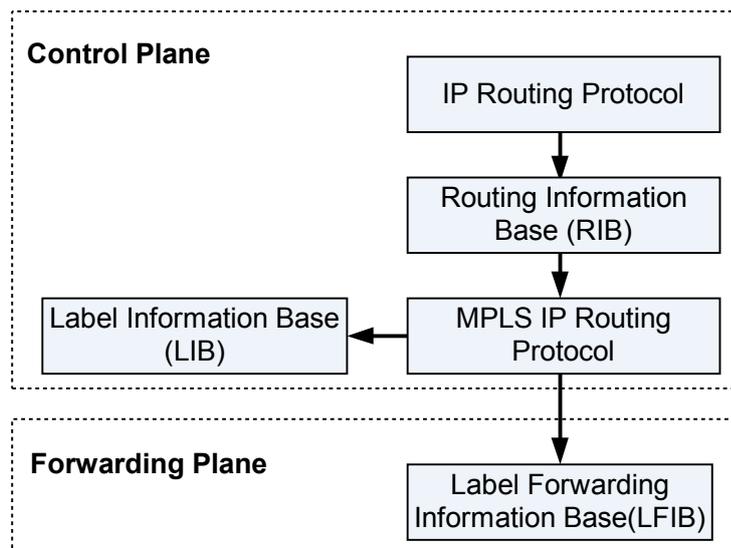
MPLS 可以使用多种标签发布协议，例如 LDP（Label Distribution Protocol）、RSVP-TE（Resource Reservation Protocol Traffic Engineering）和 MP-BGP（Multiprotocol Border Gateway Protocol）。

MPLS 的体系结构

MPLS 的体系结构由控制平面（Control Plane）和转发平面（Forwarding Plane）组成。

MPLS 体系结构如图 1-8。

图 1-8 MPLS 体系结构示意图



- 控制平面是无连接的，主要功能是负责标签的分配、标签转发表的建立、标签交换路径的建立、拆除等工作。
- 转发平面也称为数据平面（Data Plane），是面向连接的，可以使用 ATM、帧中继、Ethernet 等二层网络。转发平面的主要功能是对 IP 包进行标签的添加和删除，同时依据标签转发表对收到的分组进行转发。

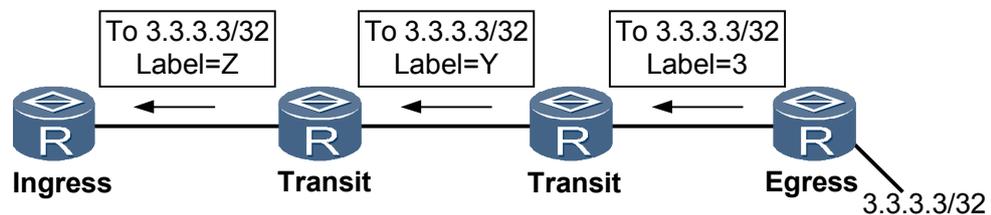
1.3.2 LSP 的建立

LSP 的基本建立过程

MPLS 需要为报文事先分配好标签，建立一条 LSP，才能进行报文转发。

标签由下游分配，按从下游到上游的方向分发。如图 1-9，由下游 LSR 在 IP 路由表的基础上进行 FEC 的划分，并将标签分配给特定 FEC，再通过标签发布协议通知上游 LSR，以便建立标签转发表和 LSP。

图 1-9 LSP 建立



LSP 分为静态 LSP 和动态 LSP 两种。静态 LSP 由管理员手工配置，动态 LSP 则利用路由协议和标签发布协议动态建立。

静态 LSP 的建立

静态 LSP 是用户通过手工为各个转发等价类分配标签而建立的。手工分配标签需要遵循的原则是：上游节点出标签的值就是下游节点入标签的值。

由于静态 LSP 各节点上不能相互感知到整个 LSP 的情况，因此静态 LSP 是一个本地的概念。

- 配置静态 LSP 的入节点，且在出接口使能了 MPLS，如果路由可达，则静态 LSP 就为 UP 状态，无论是否有 Transit 节点或出节点。这里的“路由可达”的含义是本地路由表中存在与指定目的 IP 地址精确匹配的路由项（包括目的地址和下一跳都匹配）。
- 配置静态 LSP 的 Transit 节点，且在入接口和出接口使能了 MPLS，如果入接口和出接口的物理层及协议层状态为 Up，则该静态 LSP 就为 UP 状态，无论是否有入节点、出节点或其他 Transit 节点。
- 配置静态 LSP 的出节点，且入接口使能了 MPLS，如果入接口的物理层及协议层状态为 Up，则该静态 LSP 就为 UP 状态，无论是否有入节点或 Transit 节点。

说明

只有入节点的静态 LSP 的建立依赖于路由，中间节点和出节点无此限制。

静态 LSP 不使用标签发布协议，不需要交互控制报文，因此消耗资源比较小，适用于拓扑结构简单并且稳定的小型网络。但通过静态方式分配标签建立的 LSP 不能根据网络拓扑变化动态调整，需要管理员干预。

动态 LSP 的建立

动态 LSP 通过标签发布协议动态建立。MPLS 可以使用多种标签发布协议，例如：

- LDP

LDP 是专为标签发布而制定的协议。LDP 通过逐跳方式建立 LSP 时，利用沿途各 LSR 路由转发表中的信息来确定下一跳，而路由转发表中的信息一般是通过 IGP、BGP 等路由协议收集的。LDP 并不直接和各种路由协议关联，只是间接使用路由信息。

虽然 LDP 是专门用来实现标签分发的协议，但 LDP 并不是唯一的标签分发协议。通过对 BGP、RSVP 等已有协议进行扩展，也可以支持 MPLS 标签的分发。

- RSVP-TE

资源预留协议 RSVP 是为 Integrated Service 模型而设计的，用于在一条路径的各节点上进行资源预留。RSVP 工作在传输层，但不参与应用数据的传送，是一种网络上的控制协议，类似于 ICMP。

为了能够建立基于约束的 LSP，对 RSVP 协议进行了扩展。扩展后的 RSVP 信令协议称为 RSVP-TE 信令协议，主要用于建立 TE 隧道。它拥有普通 LDP LSP 没有的功能，如发布带宽预留请求、带宽约束、链路颜色和显式路径等。

- MP-BGP

MP-BGP 是在 BGP 协议基础上扩展的协议。MP-BGP 引入 Community 属性，支持为 MPLS VPN 业务中私网路由和跨域 VPN 的标签路由分配标签。

1.3.3 MPLS 转发

基本概念

- Tunnel ID

为了给使用隧道的上层应用（如 VPN、路由管理）提供统一的接口，系统自动为各种隧道分配了一个 ID，也称为 Tunnel ID。该 Tunnel ID 只是本地有效。

Tunnel ID 的长度为 32 比特，不同类型的隧道，各字段的长度可能不同。Tunnel ID 的结构如图 1-10 所示。

图 1-10 Tunnel ID 的结构



各字段含义如下：

表 1-2 Tunnel ID 各字段含义描述

字段	含义
Token	用于在 MPLS 转发表中查找指定的 MPLS 转发信息。Token 只是一个查找转发信息的索引号。
Sequence Number	隧道 ID 的序列号。
Slot Number	出接口槽位号，指定了实际发送报文的槽位。

字段	含义
Tunnel Type	<p>Tunnel 的类型。对于 MPLS，Tunnel Type 包括：</p> <ul style="list-style-type: none"> ● LSP：通过 LDP 协议动态建立的不带约束条件的 LSP 隧道。 ● CRLSP：通过 CR-LDP 或 RSVP-TE 协议动态建立的带约束条件的 LSP 隧道。 ● MPLS Local IFNET：OptionB 和 OptionC 方式跨域 VPN 中，ASBR 通告给 BGP 对等体的 VPN 路由信息（包括 L2VPN 标签块信息）必须包含隧道信息。但是 ASBR 之间没有配置隧道。因此，为了将域外的路由信息发送给域内的对等体，ASBR 的 EBGP 为 ASBR 之间的 MPLS 接口生成一个 MPLS Local IFNET 类型的隧道。
Allocation Method	<p>Token 的分配方式，包括：</p> <ul style="list-style-type: none"> ● Global：所有隧道共用一个公共的全局空间，不可能存在两个相同的 token 值。 ● Global with reserved tokens：与 Global 方式基本相同。区别在于它会预留某些 Token 值不被隧道使用，即隧道的 Token 值起始于一个指定的数值。 ● Per slot：每一个槽位都有独立的 Token 空间。同一个槽位的 Token 值一定不同，不同槽位的 Token 值可能相同。 ● Per slot with reserved Tokens：与 Per slot 方式基本相同。区别在于它会预留某些 Token 值不被隧道使用，即隧道的 Token 值起始于一个指定的数值。 ● Per slot with different avail value：与 Per slot 方式基本相同。区别在于每个槽位的 Token 的取值范围不同。 ● Mixed：全局空间和每个槽位的空间都会被创建，然后根据出接口类型选取其中一种方式。如果是 VLANIF 或者是主干接口，则使用 Global 方式；否则使用 Per slot 方式。 ● Mixed with 2 global space：全局空间 1、全局空间 2 和每个槽位的空间都会被创建。 ● 2 global space：全局空间 1 和全局空间 2 都会被创建。 <p>说明 设备实际采用哪种 Allocation Method 取决于许可证文件。</p>

- NHLFE

下一跳标签转发表项 NHLFE（Next Hop Label Forwarding Entry）用于指导 MPLS 报文的转发。

NHLFE 包括：Tunnel ID、出接口、下一跳、出标签、标签操作类型等信息。

- ILM

入标签到一组下一跳标签转发表项的映射称为入标签映射 ILM (Incoming Label Map)。

ILM 包括: Tunnel ID、入标签、入接口、标签操作类型等信息。

ILM 在 Transit 节点的作用是将标签和 NHLFE 绑定。通过标签索引 ILM 表, 就相当于使用目的 IP 地址查询 FIB, 能够得到所有的标签转发信息。

- FTN

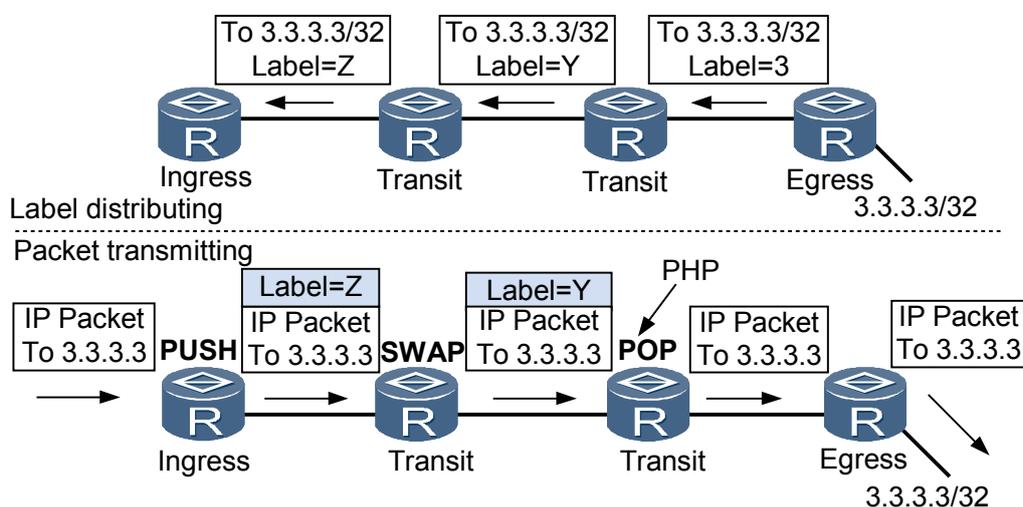
FEC 到一组 NHLFE 的映射称为 FTN (FEC-to-NHLFE)。

通过查看 FIB 表中 Token 值不为 0x0 的表项, 能够获得 FTN 的详细信息。FTN 只在 Ingress 存在。

MPLS 报文的基本转发过程

以支持 PHP 的 LSP 为例, 说明 MPLS 报文的基本转发过程。

图 1-11 MPLS 标签分发和报文转发



如图 1-11, MPLS 建立了一条 LSP, 其目的地址为 3.3.3.3/32。则 MPLS 报文基本转发过程如下:

1. Ingress 节点收到目的地址为 3.3.3.3/32 的 IP 报文, 添加标签 Z 并转发。
2. Transit 节点收到该标签报文, 进行标签交换, 将标签 Z 弹出, 换成标签 Y。
3. 倒数第二跳 Transit 节点收到带标签 Y 的报文。因 Egress 分给它的标签值为 3, 进行 PHP 操作, 弹出标签 Y 并转发报文。从倒数第二跳到 Egress 之间报文以 IP 报文形式传输。
4. Egress 节点收到该 IP 报文, 将其转发给目的地 3.3.3.3/32。

MPLS 转发流程

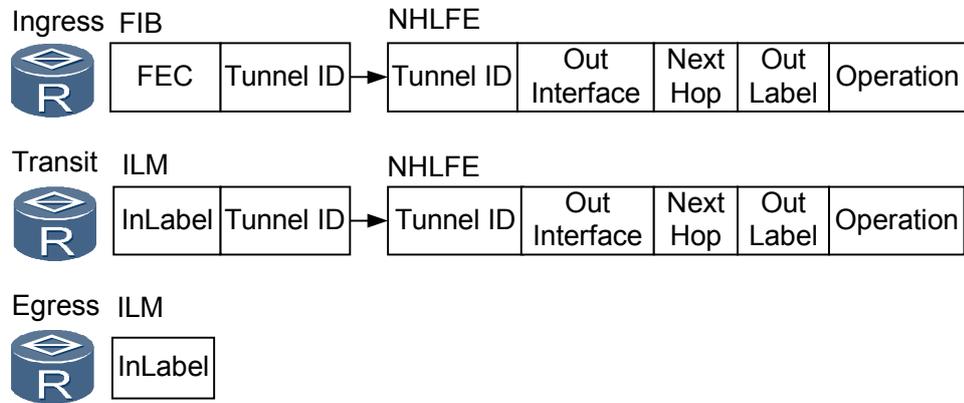
当 IP 报文进入 MPLS 域时, 首先查看 FIB 表, 检查目的 IP 地址对应的 Tunnel ID 值是否为 0x0。

- 如果 Tunnel ID 值为 0x0, 则进入正常的 IP 转发流程。

- 如果 Tunnel ID 值不为 0x0，则进入 MPLS 转发流程。

MPLS 转发流程如图 1-12 所示。

图 1-12 MPLS 转发流程



在报文转发过程中：

1. 在 Ingress，通过查询 FIB 表和 NHLFE 表指导报文的转发。
2. 在 Transit，通过查询 ILM 表和 NHLFE 表指导 MPLS 报文的转发。
3. 在 Egress，通过查询 ILM 表指导 MPLS 报文的转发。

在 MPLS 转发过程中，FIB、ILM 和 NHLFE 表项实际上是通过 Tunnel ID 中的 Token 字段关联的。

- Ingress 的处理

Ingress 节点的处理如下：

1. 查看 FIB 表，根据目的 IP 地址找到对应的 Tunnel ID。
2. 根据 FIB 表的 Tunnel ID 找到对应的 NHLFE 表项，将 FIB 表项和 NHLFE 表项关联起来。
3. 查看 NHLFE 表项，可以得到出接口、下一跳、出标签和标签操作类型，标签操作类型为 Push。
4. 在 IP 分组报文中压入获得的标签，并根据 QoS 策略处理 EXP，同时处理 TTL，然后将封装好的 MPLS 分组报文发送给下一跳。

- Transit 的处理

Transit 节点收到 MPLS 报文后的处理：

1. 根据 MPLS 的标签值查看对应的 ILM 表，可以得到 Token。
2. 根据 ILM 表的 Token 找到对应的 NHLFE 表项。
3. 查看 NHLFE 表项，可以得到出接口、下一跳、出标签和标签操作类型。
4. MPLS 报文的处理方式根据不同的标签值而不同。
 - 如果标签值 ≥ 16 ，则用新标签替换 MPLS 分组报文中的旧标签，同时处理 EXP 和 TTL，然后将替换完标签的 MPLS 分组报文发送给下一跳。

- 如果标签值为 3，则直接弹出标签，同时处理 EXP 和 TTL，然后进行 IP 转发或下一层标签转发。
- Egress 的处理
Egress 节点收到 MPLS 报文后，查看 ILM 表获得标签操作类型，同时处理 EXP 和 TTL。
 - 如果标签中的 S=1，表明该标签是栈底标签，直接进行 IP 转发。
 - 如果标签中的 S=0，表明还有下一层标签，继续进行下一层标签转发。

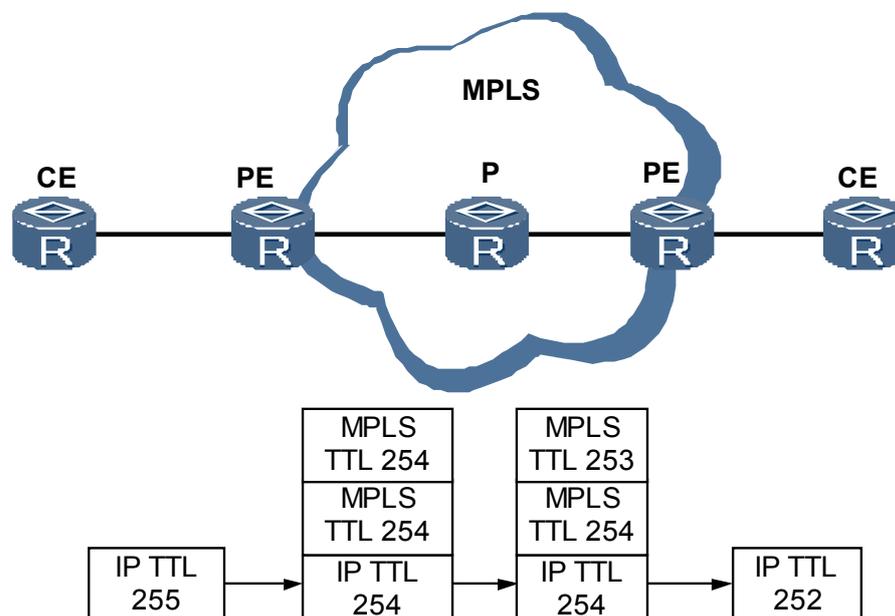
MPLS 对 TTL 的处理

MPLS 标签中包含一个 8 位的 TTL 域，其含义与 IP 头中的 TTL 域相同。MPLS 对 TTL 的处理除了用于防止产生路由环路外，也用于实现 Traceroute 功能。

RFC3443 中定义了两种 MPLS 对 TTL 的处理模式：Uniform 和 Pipe。缺省情况下，MPLS 对 TTL 的处理模式为 Pipe。

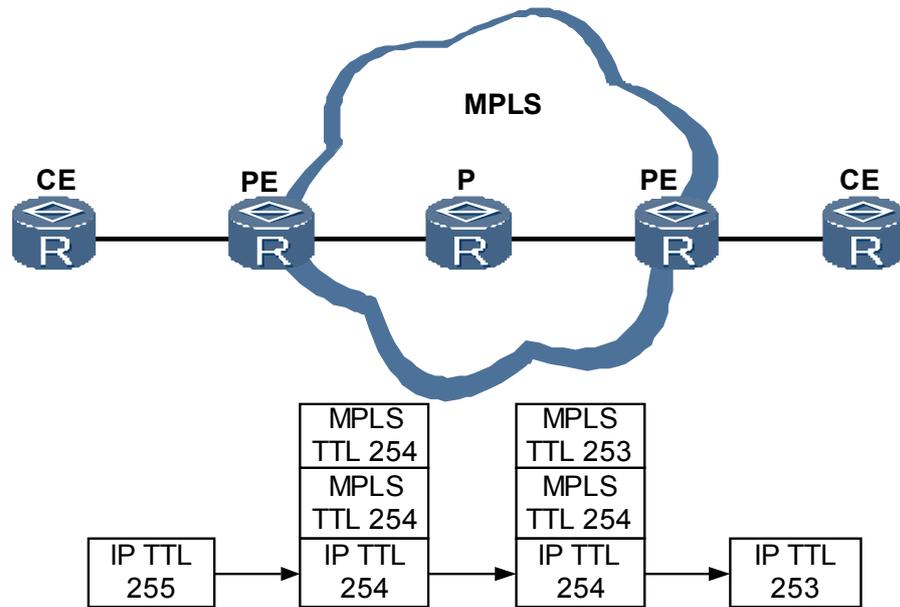
- Uniform 模式
IP 分组经过 MPLS 网络时，在入节点，IP TTL 减 1 映射到 MPLS TTL 字段，此后报文在 MPLS 网络中按照标准的 TTL 处理方式处理。在出节点将 MPLS TTL 减 1 后映射到 IP TTL 字段。如图 1-13 所示。

图 1-13 Uniform 模式下 TTL 的处理



- Pipe 模式
在入节点，IP TTL 值减 1，MPLS TTL 字段为固定值，此后报文在 MPLS 网络中按照标准的 TTL 处理方式处理。在出节点会将 IP TTL 字段的值减 1。即 IP 分组经过 MPLS 网络时，无论经过多少跳，IP TTL 只在入节点和出节点分别减 1。如图 1-14 所示。

图 1-14 Pipe 模式下 TTL 的处理



1.3.4 MPLS Ping/Traceroute

概述

在 MPLS 中，如果 LSP 转发数据失败，负责建立 LSP 的 MPLS 控制平面将无法检测到这种错误，这会给网络维护带来困难。

MPLS Ping/Traceroute 为用户提供了发现 LSP 错误、并及时定位失效节点的机制。

MPLS Ping 主要用于检查网络连接及主机是否可达。MPLS Traceroute 在检查网络连接是否可达的同时，还可以分析网络什么地方发生了故障。

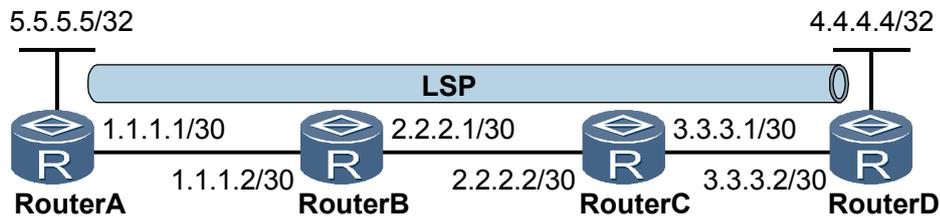
类似于普通 IP 的 Ping/Traceroute，MPLS Ping/Traceroute 使用 MPLS Echo Request 和 MPLS Echo Reply 检测 LSP 的可用性。这两种消息以 UDP 报文格式发送，端口号为 3503。接收端通过 UDP 端口号识别出 MPLS Echo Request 和 MPLS Echo Reply 报文。

MPLS Echo Request 中携带需要检测的 FEC 信息，和其他属于此 FEC 的报文一样沿 LSP 发送，从而实现对 LSP 的检测。MPLS Echo Request 通过 MPLS 转发给目的端，而 MPLS Echo Reply 则通过 IP 转发给源端。

为了防止消息到达 Egress 节点后又转发给其他节点，Echo Request 消息的 IP 头中目的地址设置为 127.0.0.1/8（本机环回地址），IP 头中的 TTL 值 = 1。

MPLS Ping

图 1-15 MPLS 网络



如图 1-15，RouterA 上建立了一条目的地为 RouterD 的 LSP。从 RouterA 对该 LSP 进行 MPLS Ping 时的处理如下：

1. RouterA 查找该 LSP 是否存在（对于 TE 隧道，查找 Tunnel 接口是否存在且 CR-LSP 是否建立成功）。如果不存在，返回错误信息，停止 Ping。如果存在，则继续进行以下操作。
2. RouterA 构造 MPLS Echo Request 报文，IP 首部目的地址为 127.0.0.1/8，IP TTL = 1。查找相应的 LSP，压入 LSP 的标签（标签的 TTL = 255），将报文发送给 RouterB。
3. 中间节点 RouterB 和 RouterC 对 MPLS Echo Request 报文进行普通 MPLS 转发。如果中间节点 MPLS 转发失败，则中间节点返回带有错误码的应答消息。
4. 当 MPLS 转发路径无故障，则 MPLS Echo Request 报文到达 LSP 的出节点 RouterD。RouterD 处理后返回 MPLS Echo Reply 报文。

MPLS Traceroute

如图 1-15，从 RouterA 对 4.4.4.4/32 进行 MPLS traceroute 时的处理如下：

1. RouterA 检查 LSP 是否存在，如果不存在，返回错误消息，停止 Traceroute，否则继续进行如下处理。
2. RouterA 构造 MPLS Echo Request 报文，IP 首部目的地址为 127.0.0.1/8，IP TTL = 1。查找相应的 LSP，压入 LSP 的标签（标签的 TTL = 1），将报文发送给 RouterB。RouterB 收到此报文，标签的 TTL 超时，因此返回 MPLS Echo Reply 消息。MPLS Echo Reply 消息的目的 UDP 端口和目的 IP 地址就是 MPLS Echo Request 报文的源 UDP 端口和源 IP 地址，IP TTL = 255。
3. RouterA 收到 MPLS Echo Reply 消息后发送 MPLS Echo Request 报文，其中标签的 TTL = 2。RouterB 对该报文进行普通 MPLS 转发。RouterC 收到此报文，标签的 TTL 超时，返回 MPLS Echo Reply 消息。
如果中间节点 MPLS 转发失败，则没有 MPLS Echo Reply 消息返回。
4. RouterA 收到 MPLS Echo Reply 消息后发送 MPLS Echo Request 报文，其中标签的 TTL = 3。RouterB 和 RouterC 对该报文进行普通 MPLS 转发。RouterD 收到此报文，发现报文目的地址为本机回环地址，返回 MPLS Echo Reply 消息。

1.4 应用

1.4.1 基于 MPLS 的 VPN

1.4.2 策略路由到 LSP

1.4.1 基于 MPLS 的 VPN

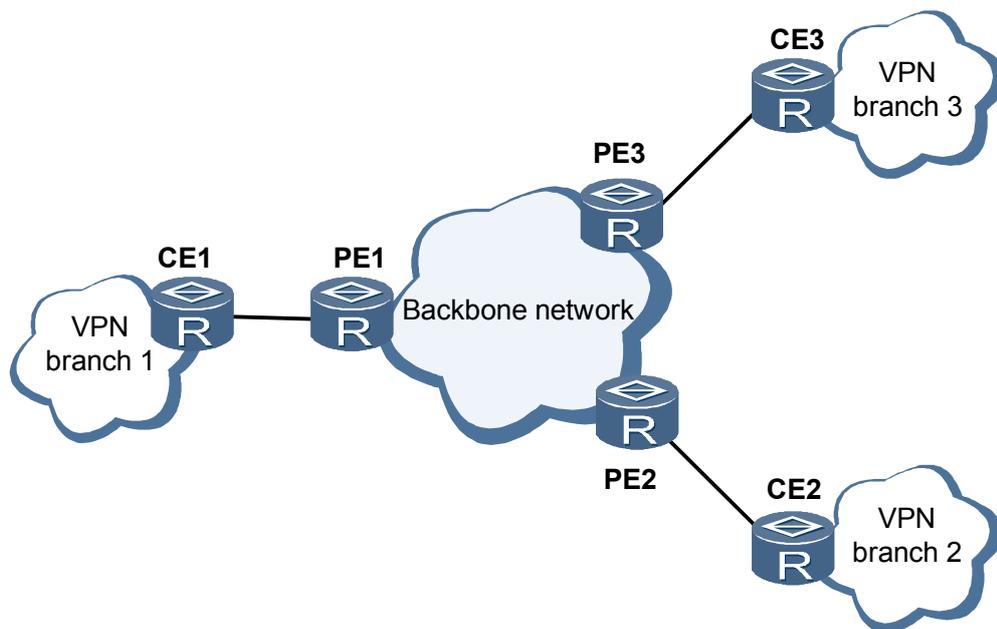
传统 VPN 一般是通过 GRE（Generic Routing Encapsulation）、L2TP（Layer 2 Tunneling Protocol）、PPTP（Point to Point Tunneling Protocol）等隧道协议来实现私有网络间数据在公网上的传送。

基于 MPLS 的 VPN 可以创建一个同 FR 网络具备的安全性很相似的专用网。用户设备一般不需要使用 IPSec 等安全技术，也无需为 VPN 配置 GRE、L2TP 等隧道，网络时延被降到最低，因为数据包不再经过封装或者加密。

基于 MPLS 的 VPN 通过 LSP 将私有网络的不同分支联结起来，形成一个统一的网络，如图 1-16 所示。基于 MPLS 的 VPN 还支持对不同 VPN 间的互通控制。图 1-16 中：

- CE（Customer Edge）是用户边缘设备，可以是路由器，也可以是交换机或主机；
- PE（Provider Edge）是服务商边缘节点，位于骨干网络。
- P（Provider）是服务提供商网络中的骨干设备，不与 CE 直接相连。P 设备只需要具备基本 MPLS 转发能力，不维护 VPN 信息。

图 1-16 基于 MPLS 的 VPN



基于 MPLS 的 VPN 具有以下特点：

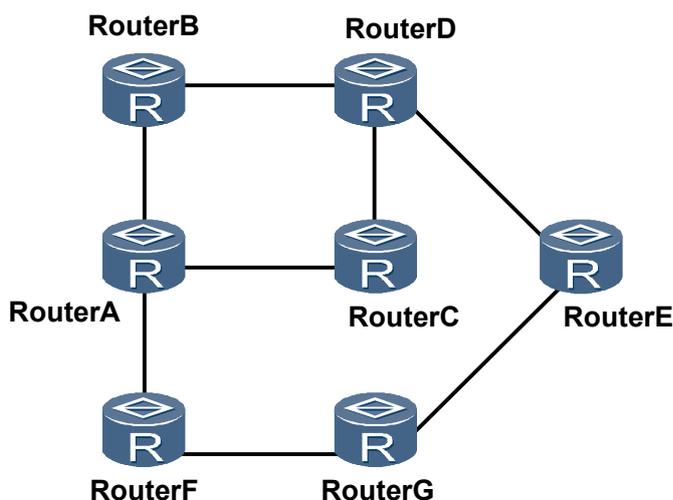
- PE 负责对 VPN 用户进行管理、建立各 PE 间 LSP 连接、同一 VPN 用户各分支间路由分派。
- PE 间的路由分派通常是用 LDP 或 MBGP 协议实现。
- 支持不同分支间 IP 地址复用和不同 VPN 间互通。

1.4.2 策略路由到 LSP

策略路由是指按照用户制定的策略进行路由选择，可应用于安全、负载分担等目的。NE20E-X6 支持策略路由到 LSP，在 MPLS 网络中，可以使符合过滤条件的 IP 报文通过指定的 LSP 路径转发。

在图 1-17 中，原网络有 RouterA、RouterB、RouterC、RouterD、RouterE 这几台设备，为提供新业务增加了 RouterF 和 RouterG 两台设备。新的业务通过 RouterF、RouterG 转发，老的业务通过原有网络。

图 1-17 策略路由到 LSP 的应用



如果希望一部分新业务通过老的网络传输，可以在 RouterA 上配置策略路由到 LSP，使符合特定条件的流量使用老的网络的 LSP 转发。

如果将策略路由到 LSP 与 LDP FRR 相结合，可以将流量指定到备份 LSP 转发。因为备份 LSP 上可能相对空闲，可以分担一部分流量。

1.5 术语与缩略语

术语

术语	解释
标签空间	分配标签的数值范围称为标签空间（Label Space）。
ILM	入标签到一组下一跳标签转发表项的映射称为入标签映射 ILM（Incoming Label Map）。ILM 包括 Tunnel ID、入标签、入接口等信息。

术语	解释
LDP 对等体	LDP 对等体是指相互之间存在 LDP 会话、使用 LDP 来交换标签/FEC 映射关系的两个 LSR。
LDP 标识符	LDP 标识符 (LDP Identifier) 用于标识特定 LSR 的标签空间范围。
NHLFE	下一跳标签转发表项 NHLFE (Next Hop Label Forwarding Entry) 用于指导 MPLS 报文的转发。NHLFE 包括: Tunnel ID、出接口、下一跳、出标签、标签操作类型等信息。
PHB	用来描述拥有相同 DSCP 值的报文的下一步转发动作。一般 PHB 包括延时、丢包率等流量特性。
控制平面	无连接的, 主要功能是负责标签的分配、标签转发表的建立、标签交换路径的建立、拆除等工作。
转发平面	也称为数据平面 (Data Plane), 是面向连接的, 可以使用 ATM、帧中继、Ethernet 等二层网络。转发平面的主要功能是对 IP 包进行标签的添加和删除, 同时依据标签转发表对收到的分组进行转发。

缩略语

缩略语	英文全称	中文全称
DoD	Downstream-on-Demand	下游按需方式
DU	Downstream Unsolicited	下游自主方式
LSP	Label switched path	标签转发路径
FEC	Forwarding Equivalence Class	转发等价类
ILM	Incoming Label Map	入标签映射
LAM	Label Advertisement Mode	标签发布模式
LDP	Label Distribution Protocol	标记分发协议/标签分发协议
LER	Label Edge Router	标记边缘路由器
LFIB	Label Forward Information Base	转发信息库
LSP	Label Switched Path	标签交换路径
LSR	Label Switching Router	标签交换路由器

缩略语	英文全称	中文全称
MPLS	Multiprotocol Label Switching	多协议标签交换
NHLFE	Next Hop Label Forwarding Entry	下一跳标签转发项
PHP	Penultimate Hop Popping	倒数第二跳弹出

2 MPLS LDP

关于本章

- 2.1 介绍
- 2.2 参考标准和协议
- 2.3 原理描述
- 2.4 术语与缩略语

2.1 介绍

定义

标签分发协议 LDP (Label Distribution Protocol) 是多协议标签交换 MPLS (Multi-Protocol Label Switching) 的一种控制协议, 相当于传统网络中的信令协议, 负责转发等价类 FEC (Forwarding Equivalence Class) 的分类、标签的分配以及标签交换路径 LSP (Label Switched Path) 的建立和维护等操作。LDP 规定了标签分发过程中的各种消息以及相关处理过程。

目的

MPLS 支持多层标签, 并且转发平面面向连接, 故具有良好的扩展性, 使在统一的 MPLS/IP 基础网络架构上为客户提供各类服务成为可能。通过 LDP 协议, 标签交换路由器 LSR (Label Switched Router) 可以把网络层的路由信息直接映射到数据链路层的交换路径上, 建立起网络层的 LSP。目前, LDP 广泛地应用于 VPN 服务, 具有组网和配置简单、支持路由拓扑驱动建立 LSP、支持大容量 LSP 等优点。

2.2 参考标准和协议

本特性的参考资料清单如下:

文档	描述	备注
RFC5036	LDP Specification	不支持环路检测。
RFC3215	LDP State Machine	-
RFC5443	LDP IGP Synchronization	不支持 end-of-lib 消息, 其它支持。
RFC3478	Graceful Restart Mechanism for Label Distribution Protocol	-
RFC1321	The MD5 Message-Digest Algorithm	-
RFC3037	LDP Applicability	-
RFC3988	Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol	-

2.3 原理描述

2.3.1 LDP 基本概念

2.3.2 LDP 会话

- 2.3.3 标签的发布和管理
- 2.3.4 LDP LSP 的建立
- 2.3.5 LDP 跨域扩展
- 2.3.6 LDP Outbound 策略和 Inbound 策略
- 2.3.7 LDP-IGP 联动
- 2.3.8 LDP 和静态路由同步
- 2.3.9 LDP GR
- 2.3.10 LDP NSR
- 2.3.11 LDP FRR
- 2.3.12 LDP MTU
- 2.3.13 LDP MD5
- 2.3.14 LDP 认证
- 2.3.15 LDP 为 BGP 分标签
- 2.3.16 LDP over TE
- 2.3.17 LDP over GRE
- 2.3.18 LDP GTSM
- 2.3.19 LDP 本远端会话共存
- 2.3.20 LDP 为所有 Peer 分标签

2.3.1 LDP 基本概念

MPLS 体系有多种标签发布协议，LDP（Label Distribution Protocol）是其中使用较广的一种。

LDP（Label Distribution Protocol）规定了标签分发过程中的各种消息以及相关的处理过程。LSR 之间将依据本地转发表中对应于一个特定 FEC 的入标签、下一跳节点、出标签等信息联系在一起，从而形成标签交换路径 LSP。

关于 LDP 的详细介绍可以参考 RFC5036（LDP Specification）。

LDP 邻接体

当一台 LSR 接收到对端发送过来的 hello 消息，意味着可能存在 LDP 对等体，此时 LSR 会建立维护对端存在的 LDP 邻接体。LDP 邻接体存在两种类型：本地邻接体（Local Adjacency）和远端邻接体（Remote Adjacency）。

LDP 对等体

LDP 对等体是指相互之间存在 LDP 会话、使用 LDP 来交换标签消息的两个 LSR。

LDP 对等体通过它们之间的 LDP 会话获得对方的标签。

LDP 会话

LDP 会话用于 LSR 之间交换标签映射、释放等消息。LDP 会话分为两种类型：

- 本地 LDP 会话（Local LDP Session）：建立会话的两个 LSR 之间是直连的；
- 远端 LDP 会话（Remote LDP Session）：建立会话的两个 LSR 之间可以是直连的，也可以是非直连的。

本地 LDP 会话和远端 LDP 会话可以共存。

LDP 动态能力通告功能

LDP 动态能力通告功能可以确保在不中断会话的情况下，动态的使能或者去使能 LDP 新的扩展能力，保证 LSP 的稳定性。

LDP 邻接体/对等体/会话之间的关系

LDP 通过邻接体来维护对等体的存在，对等体的类型取决于维护它的邻接体的类型。一个对等体可以由多个邻接体来维护，可以由本地邻接体和远端邻接体两者来维护，则对等体类型为本远共存对等体。只有存在对等体才能建立 LDP 会话。

LDP 消息类型

LDP 协议主要使用四类消息：

- 发现（Discovery）消息：用于通告和维护网络中 LSR 的存在。
- 会话（Session）消息：用于建立、维护和终止 LDP 对等体之间的会话。
- 通告（Advertisement）消息：用于创建、改变和删除 FEC 的标签映射。
- 通知（Notification）消息：用于提供建议性的消息和差错通知。

为保证 LDP 消息的可靠发送，除了 Discovery 消息使用 UDP 外，LDP 的 Session 消息、Advertisement 消息和 Notification 消息都使用 TCP 传输。

标签空间与 LDP 标识符

- 标签空间

LDP 对等体之间分配标签的数值范围称为标签空间（Label Space）。可以分为：

- 全局标签空间（Per-Platform Label Space）：整个 LSR 使用一个标签空间。
- 接口标签空间（Per-Interface Label Space）：为 LSR 的每个接口指定一个标签空间。

- LDP 标识符

LDP 标识符（LDP Identifier）用于标识特定 LSR 的标签空间范围。LDP 标识符的格式为<LSR ID>: <Label space ID>，长度为六字节，其中：

- LSR ID：表示 LSR 标识符，占四字节。
- Label space ID：表示标签空间标识符，占两字节。

2.3.2 LDP 会话

LDP 发现机制

LDP 发现机制用于 LSR 发现潜在的 LDP 对等体。LDP 有两种发现机制：

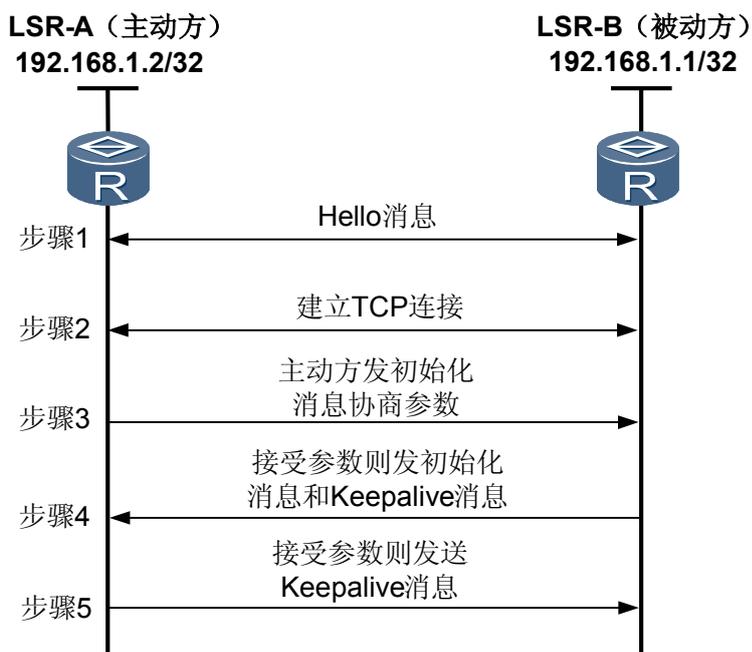
- 基本发现机制：用于发现链路上直连的 LSR。
LSR 通过周期性的发送 LDP Hello 报文，实现 LDP 基本发现机制，建立本地 LDP 会话。
Hello 报文中携带 LDP Identifier 及一些其他信息（例如 hold time、transport address）。如果 LSR 在特定接口接收到 LDP Hello 消息，表明该接口存在 LDP 对等体。
- 扩展发现机制：用于发现链路上非直连 LSR。
LSR 周期性的发送 Targeted Hello 消息到指定地址，实现 LDP 扩展发现机制，建立远端 LDP 会话。
Targeted Hello 消息使用 UDP 报文，目的地址是指定地址，目的端口是 LDP 端口（646）。Targeted Hello 消息同样携带 LDP Identifier 及一些其他信息（例如：transport address、hold time）。如果 LSR 在特定接口接收到 Targeted hello 消息，表明该接口存在 LDP 对等体。

LDP Session 建立过程

两台 LSR 之间交换 Hello 消息触发 LDP session 的建立。

LDP Session 的建立过程如图 2-1 所示：

图 2-1 LDP Session 建立过程



1. 两个 LSR 之间互相发送 Hello 消息。Hello 消息中携带传输地址，双方使用传输地址建立 LDP 会话。首先选择传输地址较大的一方作为主动方，发起建立 TCP 连接。如图 2-1 所示，LSRA 作为主动方发起建立 TCP 连接，LSRB 作为被动方等待对方发起连接。
2. TCP 连接建立成功后，由主动方 LSRA 发送 Initialization 消息，协商建立 LDP 会话的相关参数，包括 LDP 协议版本、标签分发方式、Keepalive 保持定时器的值、最大 PDU 长度和标签空间等。

3. 被动方 LSRB 收到 Initialization 消息后，如果不能接受相关参数，则发送 Notification 消息终止 LDP 会话的建立；如果被动的 LSRB 能够接受相关参数，则发送 Initialization 消息，同时发送 Keepalive 消息给主动方 LSRA。
4. 主动方 LSRA 收到 Initialization 消息后，如果不能接受相关参数，则发送 Notification 消息给被动的 LSRB 终止 LDP 会话的建立；如果能够接受相关参数，则发送 Keepalive 消息给被动的 LSRB。

当双方都收到对端的 Keepalive 消息后，LDP 会话建立成功。

2.3.3 标签的发布和管理

LDP 会话建立后，LDP 协议开始交换标签映射等消息用于建立 LSP。RFC5036 分别定义了标签发布方式、标签分配控制方式、标签保持方式来决定 LSR 如何发布和管理标签。

NE20E-X6 支持如下组合方式：

- 下游自主方式（DU）+有序标签控制方式（Ordered）+自由标签保持方式（Liberal）。
- 下游按需方式（DoD）+有序标签控制方式（Ordered）+保守标签保持方式（Conservative）。

标签发布方式

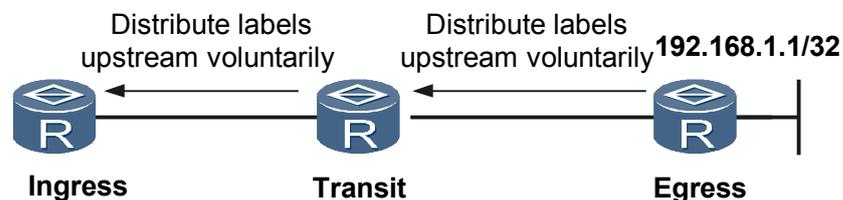
在 MPLS 体系中，由下游 LSR 决定将标签分配给特定 FEC，再通知上游 LSR。即标签由下游指定，标签的分配按从下游到上游的方向分发。

标签发布方式（Label Advertisement Mode）可以分为以下两种：

- 下游自主方式
下游自主方式 DU（Downstream Unsolicited）是指对于一个特定的 FEC，LSR 无须从上游获得标签请求消息即进行标签分配与分发。

如图 2-2 所示，对于目的地址为 192.168.1.1/32 的 FEC，根据主机方式触发，下游（Egress）通过标签映射消息主动向上游（Transit）通告自己的主机路由 192.168.1.1/32 的标签。

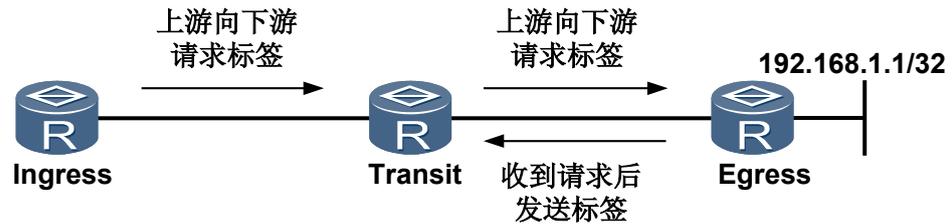
图 2-2 DU 方式



- 下游按需方式
下游按需方式 DoD（Downstream on Demand）是指对于一个特定的 FEC，LSR 获得标签请求消息之后才进行标签分配与分发。

如图 2-3 所示，对于目的地址为 192.168.1.1/32 的 FEC，根据主机方式触发，上游（Ingress）向下游发送标签请求消息，下游（Egress）收到标签请求消息后，才会向上游发送标签映射消息。

图 2-3 DoD 方式



具有标签分发邻接关系的上游 LSR 和下游 LSR 必须对使用的标签发布方式达成一致。

标签分配控制方式

标签分配控制方式（Label Distribution Control Mode）是指在 LSP 的建立过程中，LSR 分配标签时采用的处理方式。

标签分配控制方式可以分为以下两种：

- 独立标签分配控制

独立标签分配控制（Independent）是指本地 LSR 可以自主地分配一个标签绑定到某个 FEC，并通告给上游 LSR，而无需等待下游的标签。

- 如图 2-2 所示，如果标签发布方式为 DU，且标签分配控制方式为 Independent，则 LSR（Transit）无需等待下游（Egress）的标签，就会直接向上游（Ingress）分发标签。
- 如图 2-3 所示，如果标签发布方式为 DoD，且标签分配控制方式为 Independent，则发送标签请求的 LSR（Ingress）的直连下游（Transit）会直接回应标签，而不必等待来自最终下游（Egress）的标签。

- 有序标签分配控制

有序标签分配控制（Ordered）是指对于 LSR 上某个 FEC 的标签映射，只有当该 LSR 已经具有此 FEC 下一跳的标签映射消息、或者该 LSR 就是此 FEC 的出节点时，该 LSR 才可以向上游发送此 FEC 的标签映射。

- 如图 2-2 所示，如果标签发布方式为 DU，且标签分配控制方式为 Ordered，则 LSR（Transit）只有收到下游（Egress）的标签映射消息，才会向上游（Ingress）分发标签。
- 如图 2-3 所示，如果标签发布方式为 DoD，且标签分配控制方式为 Ordered，则发送标签请求的 LSR（Ingress）的直连下游（Transit）只有收到最终下游（Egress）的标签映射消息，才会向上游（Ingress）分发标签。

标签保持方式

标签保持方式（Label Retention Mode）是指 LSR 对收到的、但目前暂时不需要的标签映射的处理方式。

LSR 收到的标签映射可能来自下一跳，也可能来自非下一跳。

标签保持方式可以分为以下两种：

- 自由标签保持方式

自由标签保持方式（Liberal）是指对于从邻居 LSR 收到的标签映射，无论邻居 LSR 是不是自己的下一跳都保留。

- 保守标签保持方式

保守标签保持方式（Conservative）是指对于从邻居 LSR 收到的标签映射，只有当邻居 LSR 是自己的下一跳时才保留。

当网络拓扑变化引起下一跳邻居改变时：

- 使用自由标签保持方式，LSR 可以直接利用原来非下一跳邻居发来的标签，迅速重建 LSP（关于 LDP LSP 的建立，请参见 [LDP LSP 的建立](#)），但需要更多的内存和标签空间。

- 使用保守标签保持方式，LSR 只保留来自下一跳邻居的标签，节省了内存和标签空间，但 LSP 的重建会比较慢。

保守标签保持方式通常与 DoD 方式一起，用于标签空间有限的 LSR。

已经被分配标签，但是没有建立成功的 LSP 叫做 Liberal LSP。

2.3.4 LDP LSP 的建立

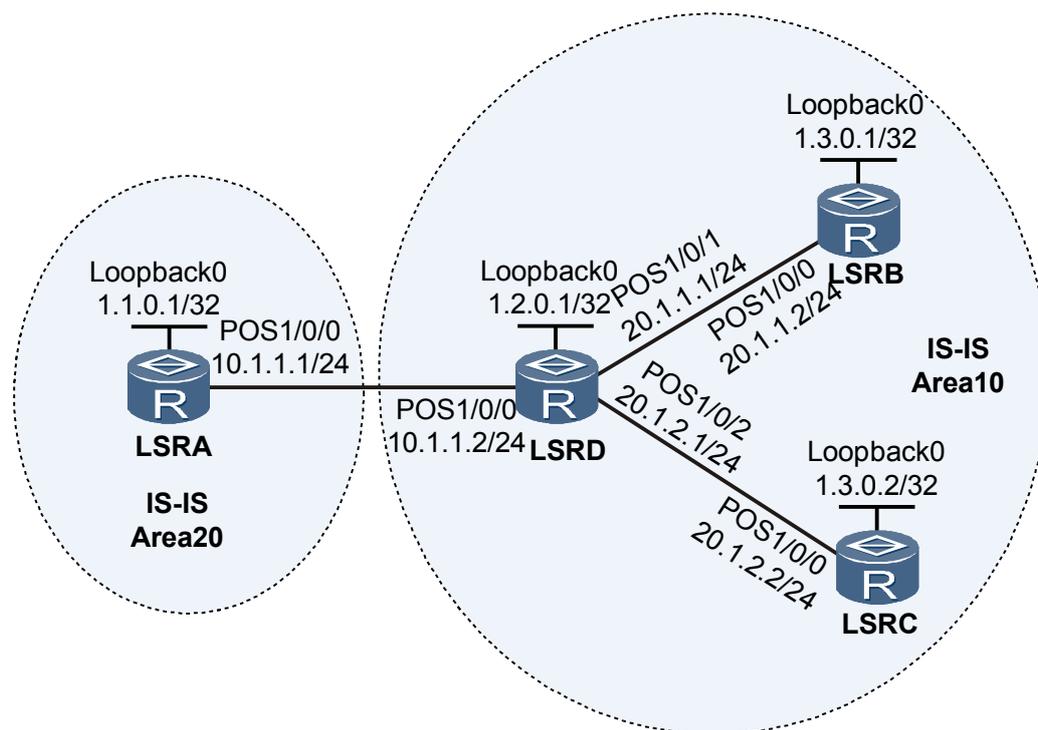
LSP 的建立过程实际就是将 FEC 和标签进行绑定，并将这种绑定通告 LSP 上的相邻 LSR。下面结合下游自主标签发布方式和有序标签控制方式来说明其主要步骤：

1. 当网络的路由改变时，如果有一个边缘节点发现自己的路由表中出现了新的目的地址，并且这一地址不属于任何现有的 FEC，则该边缘节点需要为这一目的地址建立一个新的 FEC。
2. 如果 MPLS 网络的出节点有可供分配的标签，则为 FEC 分配标签，并主动向上游发出标签映射消息，标签映射消息中包含分配的标签和绑定的 FEC 等信息。
3. 收到标签映射消息的 LSR 在其标签转发表中增加相应的条目，然后主动向上游 LSR 发送对于指定 FEC 的标签映射消息。
4. 当入节点 LSR 收到标签映射消息时，它也需要在标签转发表中增加相应的条目。这时，就完成了 LSP 的建立，接下来就可以对该 FEC 对应的数据分组进行标签转发。

2.3.5 LDP 跨域扩展

本特性可以使能 LDP 建立跨越多个 IGP 区域的 LDP LSP，提供穿越公网的隧道。

图 2-4 LDP 跨域扩展组网拓扑



如图 2-4 所示，存在 Area 10 和 Area 20 两个 IGP 区域。

在 Area 10 区域边缘的 LSRD 的路由表中，存在到 LSRB 和 LSRC 的两条主机路由。通常，为了避免路由数量多而引起的对资源的过多占用，在 LSRD 上可以通过 ISIS 路由协议将这两条路由聚合为 1.3.0.0/24 发送到 Area 20 区域，那么 LSRA 的路由表中只有这条聚合后的路由，而没有 32 位的主机路由。然而，缺省情况下，LDP 在建立 LSP 的时候，会在路由表中查找与收到的标签映射消息中携带的 FEC 精确匹配的路由，对于图 2-4 中的情况，LSRA 的路由表项信息和 FEC 携带的路由信息如表 2-1 所示。

表 2-1 LSRA 的路由表项信息和 FEC 携带的路由信息

LSRA 路由表项信息	FEC
1.3.0.0/24	1.3.0.1/32
	1.3.0.2/32

对于聚合路由，LDP 只能建立 Liberal LSP，无法建立跨越 IGP 区域的 LDP LSP，以至于无法给 VPN 业务提供必要的骨干网隧道。

因此，在图 2-4 的应用中，需要使能 LDP 按照最长匹配方式查找路由建立 LSP。在 LSRA 的路由表中，已经存在聚合路由 1.3.0.0/24。当 LSRA 收到 Area 10 区域的标签映射消息时（例如携带的 FEC 为 1.3.0.1/32），按照 RFC5283 使用最长匹配的查找方式，LSRA 能够找到聚合路由 1.3.0.0/24 的信息，将该路由的出接口和下一跳作为到 FEC 1.3.0.1/32 的出接口和下一跳。这样，LDP 就可以建立跨越 IGP 区域的 LDP LSP。

2.3.6 LDP Outbound 策略和 Inbound 策略

一般情况下，LSR 对于标签映射消息的接收和发送没有限制，导致大量的 LSP 的建立。这对于性能较低的设备，会造成资源的大量消耗，设备将无法承受。配置 LDP Outbound 策略和 Inbound 策略，可以限制标签映射消息的发送和接收，减少 LSP 的数量，节省内存。

LDP Outbound 策略

LDP Outbound 策略可以用于过滤向对等体发送的标签映射消息，只过滤路由 FEC 的标签映射消息，不会过滤 L2VPN 的标签映射消息，支持对 BGP 标签路由和非 BGP 路由分别指定 FEC 范围。

如果一组对等体或者全部对等体发送标签映射消息时，对 FEC 的限制范围是相同的，则可以对这一组或全部对等体应用相同的 Outbound 策略。

LDP Outbound 策略还支持水平分割策略，即只向上游 LDP 对等体分配标签。

LSR 在向对等体发送路由 FEC 的标签映射消息前会查看本地是否配置了对应路由类型（BGP 标签路由或非 BGP 路由）的 Outbound 策略。

- 如果本地没有配置 Outbound 策略，则发送标签映射消息。
- 如果本地配置了 Outbound 策略，则根据策略中指定的路由 FEC 的范围来确定是否发送该路由 FEC 的标签映射消息。

如果路由 FEC 没有通过任何 Outbound 策略，则不允许建立 Transit LSP 或 Egress LSP。

LDP Inbound 策略

LDP Inbound 策略可以用于过滤从对等体接收的标签映射消息，只过滤路由 FEC 的标签映射消息，不会过滤 L2VPN 的标签映射消息，支持对非 BGP 路由指定 FEC 范围。

如果一组对等体或者全部对等体接收标签映射消息时，对 FEC 的限制范围是相同的，则可以对这一组或全部对等体应用相同的 Inbound 策略。

LSR 在接收对等体发送的的路由 FEC 的标签映射消息前会查看本地是否配置了 Inbound 策略。

- 如果本地没有配置 Inbound 策略，则接收标签映射消息。
- 如果本地配置了 Inbound 策略，则根据策略中指定的路由 FEC 的范围来确定是否接收该路由 FEC 的标签映射消息。

如果路由 FEC 没有通过本地任何 Inbound 策略，则不接收标签映射消息。

在不接收标签映射消息的情况下：

- 如果本地和邻居建立的是 DU 会话，则建立 Liberal LSP，且此 Liberal LSP 不可以被 LDP FRR 选作备份 LSP。
- 如果本地和邻居建立的是 DoD 会话，则发送 Release 消息拆除标签绑定。

2.3.7 LDP-IGP 联动

LDP-IGP 联动是一种通过改变 IGP 发布路由的方式，保证 IGP 和 LDP 流量一致的技术。

在存在主备链路的组网中有如下问题：

- 当主链路发生故障时，IGP 路由和 LSP 均切换到备份链路上。但当主链路从故障中恢复时，由于 IGP 路由比 LDP 收敛速度快，IGP 会先于 LDP 切换回主链路，因此

造成 LSP 流量丢失。在 LDP Over TE 场景下，LSP 首先切到物理链路上，紧接着又切到备份链路上，导致 LSP 震荡。

- 当主链路正常，但主链路节点间的 LDP 会话发生故障时，IGP 路由仍然使用主链路，而主链路的 LSP 被删除。同时，由于备份链路不存在 IGP 路由，故 LSP 无法在备份链路建立，导致 LSP 流量丢失。在 LDP Over TE 场景下，路由可能首先切到联动的 Tunnel 上，接下来 Tunnel 被设为最大开销值后流量切到备份路径，导致 LSP 震荡。
- 当发生主备倒换时，LDP 会话的建立可能晚于 IGP 的 GR 结束，从而 IGP 发布链路的最大开销值，导致路由振荡。

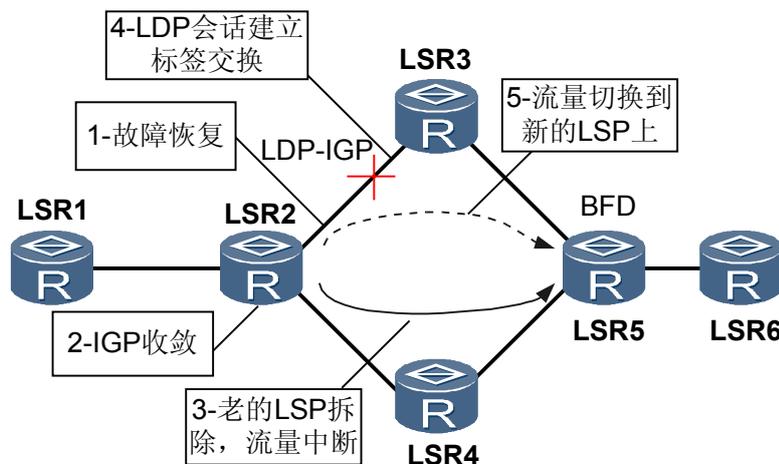
LDP-IGP 联动的基本原理是：通过抑制 IGP 发布正常路由，来保证 LDP 完成收敛。也就是在 LSP 发生故障的时候，IGP 改变发布路由的方式，以便 LDP 和 IGP 达到收敛一致。

LDP-IGP 联动包括四个定时器：

- Hold-down timer
- Hold-max-cost timer
- Delay timer
- GR Delay timer

LDP-IGP 联动主要应用于以下场景：

图 2-5 LDP-IGP 联动需要解决的回切问题



- 如图 2-5 所示，在存在主备链路的组网中，当主链路故障恢复后，流量从备份链路切换到主链路。但是如果当 IGP 收敛之后，原来的 LSP 路径无法再被使用时，新的 LSP 还没有建立，则在这个时间差内，流量被丢弃。在这种情况下，可以通过配置 LDP-IGP 联动，令 IGP 推迟路由的回切，直至 LDP 完成收敛。即在新的 LSP 没有收敛时，保持原来的 LSP，让流量继续从原来的 LSP 路径转发，直至新的 LSP 建立成功，再删除原来的 LSP。具体过程如下：

1. 链路故障恢复；
2. LSR2 与 LSR3 之间建立 LDP 会话，同时 IGP 抑制邻居关系的建立；
3. 流量仍然会按照原来的 LSP 转发；
4. LDP 会话建立，交换标签消息，通告 IGP 启动同步；

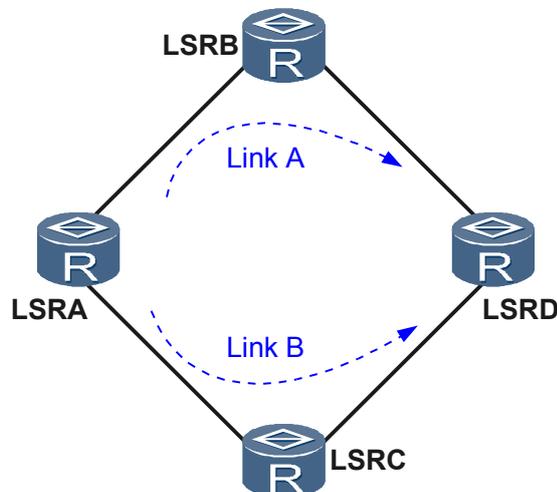
5. IGP 启动邻居关系建立，IGP 收敛到主链路上，LSP 重新建立并收敛到主链路上（一般在毫秒级）。
- 当主链路节点间的 LDP 会话发生故障时，主链路上的 LSP 被删除，但是 IGP 仍然使用主链路，导致 LSP 流量不能切换到备份链路，流量持续丢失。在这种情况下，可以配置 LDP-IGP 联动。在 LDP 会话发生故障时，LDP 向 IGP 通告 LDP 会话故障，这样 IGP 就会在该链路上发布最大开销值，实现路由切换至备份链路，从而 LSP 也切换至备份链路，具体过程如下：
 1. 主链路节点间 LDP 会话故障；
 2. LDP 通告 IGP 主链路会话故障，IGP 在主链路发布最大开销值；
 3. IGP 路由切换至备份链路；
 4. LSP 在备份链路重新建立并下发转发表项。
 为防止 LDP 会话一直不能重新建立，可通过配置 Hold-max-cost 定时器为永久发布最大开销值，使流量在主链路的 LDP 会话重新建立之前，一直都使用备份链路。
 - 当系统发生主备倒换时，LDP-IGP 联动的具体过程如下：
 1. Restarter 端的 IGP 会先发布正常开销值，并启动内部定时器 GR Delay timer 等待 LDP 会话建立，然后再结束 GR。
 2. 如果 GR Delay timer 超时，LDP 会话依然没有建立，IGP 启动 Hold-max-cost 定时器，并发布接口链路的最大开销值。
 3. 当 LDP 会话重新建立或 Hold-max-cost 定时器超时时，IGP 恢复本地链路的正常开销值，更新路由。
 4. Helper 端一直保留路由和 LSP 不会被删除，所以当 LDP 会话进入 Down 状态时，LDP 不会通知 IGP Session Down 事件，IGP 仍然发布链路的正常开销值，保证流量和 LSP 不发生切换。

2.3.8 LDP 和静态路由同步

使用 LDP 和静态路由同步技术，可以通过抑制静态路由活跃来迫使正切场景下流量切换到备份链路，回切场景下推迟流量切换到主链路，从而静态路由和 LDP 保持同步。

LDP 和静态路由同步主要是解决 LSR 之间存在备份路径组网中正切和回切过程的丢包问题。在 LSP 存在主备链路的组网中，如图 2-6 所示，LSRA 和 LSRD 之间通过静态路由连通，LDP 在两端基于静态路由建立 LSP，正常情况下优选 LinkA。

图 2-6 LDP 和静态路由同步解决的 LSP 切换场景



- 正切场景

LDP 会话正常建立的情况下，MPLS 流量走 LinkA。如果 LSRB 上的 LDP 被去使能或 LDP 出现故障，导致 LSRA 和 LSRB 之间的 LDP 会话发生中断，但 LSRA 和 LSRB 之间的链路没有问题，静态路由是活跃的，而 LSP 切换到 LinkB，导致 LSRA 和 LSRD 之间的 MPLS 流量中断。

如果在 LSRA 上使能 LDP 和静态路由同步后，当 LDP 会话 Down 掉时，静态路由自动切换到 LinkB，使 LSP 切换到备份 LSP，保证流量不中断。在同步期间，调整 LinkA 的静态路由优先级，使流量切换到备份链路。

- 回切场景

当 LSRA 和 LSRB 之间的链路失效时，LSP 正切到 LinkB。随后 LSRA 和 LSRB 之间的链路恢复时，LSP 随静态路由回切。原来的 LSP 路径无法使用，新的 LSP 还没有建立时，在这个时间差内，LSRA 和 LSRD 之间的 MPLS 流量中断。

如果在 LSRA 上使能 LDP 和静态路由同步后，当 LDP 会话 Up 时，静态路由才开始活跃，保证流量不中断。

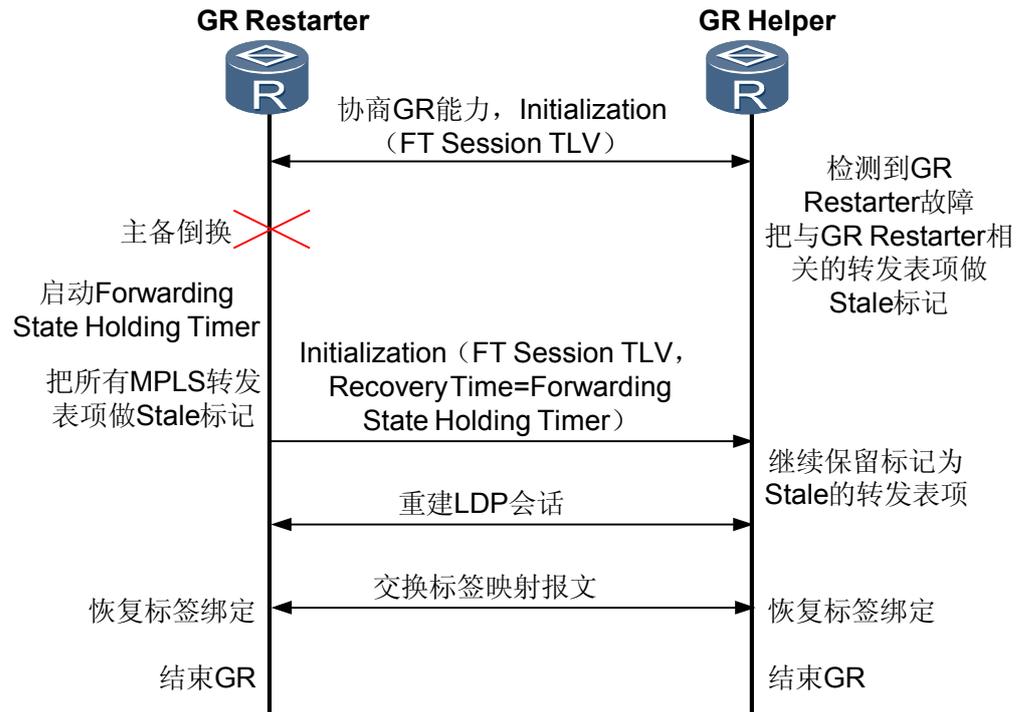
2.3.9 LDP GR

LDP GR (Graceful Restart) 是借助邻居设备 (Helper) 的帮助，实现主备倒换的设备 (Restarter) 转发不中断。

在没有 GR 能力的情况下，发生主备倒换时，邻居会因为会话进入 Down 状态而删除 LSP，产生流量短时间中断、业务短时间中断。对于以上情况，如果配置了 LDP GR 能力，可以保证意外主备倒换前后的标签保持一致，即保持了 MPLS 转发不中断。具体过程如图 2-7 所示：

1. 主备倒换前，LDP 邻居在建立 LDP 会话时进行 GR 能力协商。
2. Helper 感知到 Restarter 进行主备倒换后，启动 GR Reconnect 定时器，在该定时器超时前保留 Restarter 相关的转发表项。其中转发不中断的前提是 Restarter 保留了 MPLS 转发表项。
3. 如果在 Helper 的 GR Reconnect 定时器超时前 Restarter 和 Helper 之间的 LDP 会话重建完成，则 Helper 会删除 GR Reconnect 定时器，并启动 GR Recovery 定时器。
4. 在 Helper 的 GR Recovery 定时器超时前，Helper 会协助 Restarter 恢复转发表项，Restarter 也会协助 Helper 恢复转发表项，该定时器超时后，Helper 会删除所有未恢复的与 GR Restarter 相关的转发表项。
5. 与此同时，Restarter 进行主备倒换后，会启动 Forwarding State Holding 定时器。该定时器超时前，Restarter 保留重启前的转发表项，并在 Helper 的协助下进行转发表项的恢复工作。该定时器超时后，Restarter 将删除所有未恢复的转发表项。

图 2-7 LDP GR 实现原理



2.3.10 LDP NSR

基本原理

NSR(Non-Stop Routing)技术和 NSF (Non-Stop Forwarding) 技术有本质的区别，是 NSF 技术的革新。在系统控制平面发生故障（软件故障或硬件故障）时，NSR 技术除了能够保证转发不中断，还能够保证控制平面的连接不中断，且邻居控制平面不感知。

LDP 通过主备板同步来实现 NSR。首先，备板启动时，通过批量备份，达到主备板数据阶段性一致。然后，通过实时备份和主备板同时接收报文，达到主备板的实时数据同步。这样，NSR 就能够保证主备倒换后备板快速接管原主控板的业务，使邻居不感知本路由器故障。

LDP 实现了如下关键数据的主备实时同步：

- LSP 转发表项
- 标签、XC 等关键资源
- LDP 协议控制块

2.3.11 LDP FRR

LDP FRR (Fast Re-Route) 为 MPLS 网络提供快速重路由功能，实现了本地的端口备份。

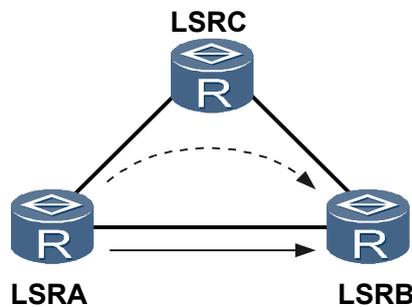
LDP FRR 原理是通过 LDP 信令的 Liberal 标签保持方式，先获取 Liberal Label，为该标签申请转发表项资源，并将转发信息下发到转发平面作为主 LSP 的备用转发表项。当接口故障（接口自己感知或者结合 BFD 检测）或者主 LSP 不通（结合 BFD 检测）时，可

以快速的将流量切换至备份路径，从而实现了对主 LSP 的保护。因为 LDP FRR 生成的备用转发表项依附于主 LSP，所以 LDP FRR 没有回切的概念。

- 手动配置的 LDP FRR 需要使用命令来指定建立的备份 LSP 的出接口和下一跳。当 Liberal Label 的来源匹配指定的出接口和下一跳的时候，就能够建立备份 LSP 并下发转发表项。
- LDP Auto FRR 依赖 IP FRR 的实现。只有 Liberal Label 的来源匹配存在的备份路由，即保留的 Liberal Label 来自备份路由出接口和下一跳，并且满足备份 LSP 触发策略，同时没有根据该备份路由手工配置的备份 LSP 存在的时候，才能够为之建立备份 LSP 并下发转发表项。LDP Auto FRR 策略默认是 32 位的备份路由触发 LDP 建立备份 LSP。在手工配置的 LDP FRR 和 LDP Auto FRR 同时满足创建条件的时候，优先建立手工配置的 LDP FRR。

应用场景

图 2-8 LDP FRR 典型应用场景-三角型拓扑



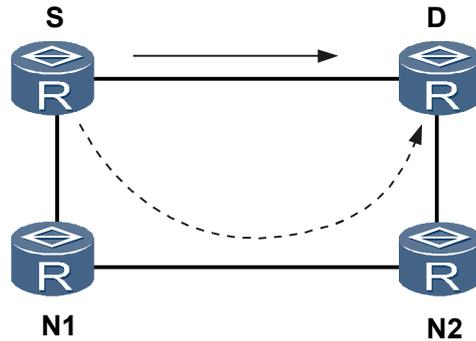
如图 2-8 所示，为 LDP FRR 的典型应用场景。LSRA 到 LSRB 的优选路由为 LSRA-LSRB，次优路由为 LSRA-LSRC-LSRB，因此 LSRA 上会建立一条 LSRA-LSRB 的主 LSP，并用 LSRA-LSRC-LSRB 作为备份 LSP 保护该主 LSP。当 LSRA 收到 LSRC 发来的标签后，会和路由比较，因为 LSRA 到 LSRB 的路由下一跳不是 LSRC，所以 LSRA 会把这个标签存为 Liberal Label，如果存在下列情况中的任意一种：

- 对于手动配置的 LDP FRR，该 Liberal Label 的来源对应着指定的出接口和下一跳
- 对于 LDP Auto FRR，该 Liberal Label 的来源对应的备份路由存在，且该路由的目的地址满足 LDP 创建备份 LSP 的策略，同时没有根据该备份路由手工配置的备份 LSP 存在

就可以为该 Liberal Label 申请一个转发表项资源，创建备份 LSP 作为主 LSP 的备用转发表项，和主 LSP 一起下发到转发平面，这样主 LSP 就和这条备份 LSP 关联起来了。

接口自己感知接口故障、BFD 感知接口故障、或者 BFD 感知主 LSP 不通等，都能触发 LDP FRR 切换。当 LDP FRR 切换后，流量根据备用转发表项切换到备份路径上，至此 LDP FRR 生效。之后的变化过程是路由从 LSRA-LSRB 收敛到 LSRA-LSRC-LSRB，在新的路径（原来的备份路径）上根据路由新建 LSP，再把原来的主 LSP 删除，流量按照 LSRA-LSRC-LSRB 上新建的 LSP 进行转发。

图 2-9 LDP FRR 典型应用场景-口字型拓扑



LDP FRR 对图 2-8 的三角形拓扑支持情况较好，但对口字型拓扑不一定能够完全支持。如图 2-9 所示，如果 N1 到 D 的最优路由是 N1-N2-D（不可负载分担），S 就会收到来自 N1 的 Liberal 标签，并绑定 LDP FRR。S-D 之间的链路故障时，流量会切换到 S-N1-N2-D，不会形成环路。

但如果 N1 到 D 的最优路由是 N1-N2-D 和 N1-S-D 负载分担，S 作为 N1 的下游邻居，不一定会收到来自 N1 的 Liberal 标签。并且，即使 S 有了该 Liberal 标签（LDP 为每 peer 分标签项目），绑定了 LDP FRR，发生切换后流量走到 N1 后还很有可能会走回给 S，从而形成环路，直至 N1 到 D 的路由收敛为 N1-N2-D。

2.3.12 LDP MTU

MTU 称为最大传输单元（Maximum Transmission Unit）。当同一个网络上的两台设备进行互通时，该网络的 MTU 是非常重要的。MTU 的大小决定了发送端一次能够发送报文的最大字节数，如果 MTU 设置的超过了接收端所能够承受的最大值，或者是超过了发送路径上途经的某台设备所能够承受的最大值，这样就会造成报文分片甚至丢弃，加重网络传输的负担。所以设备在进行通信之前必须要把 MTU 计算明确，才能保证每次发送的报文都能够畅通无阻的到达接收端，确保报文发送一次成功。

基本原理

LDP LSP 的转发和普通的 IP 转发虽然在实现方式上存在很大不同，但是 LDP LSP MTU 和普通的 IP 转发涉及的 MTU 在原理上有很多类似的地方。它们都是使得发送出去的报文能够畅通无阻的途经每一台中转设备，不需要报文重组就能直接到达接收端。

MPLS MTU 和接口 MTU 一样，有系统默认值，也可以通过命令行配置。LDP MTU 的计算方法是，LSR 针对某个 FEC，把所有下游设备通告的 MTU 和本机出接口 MTU 做比较，取出最小值通告给上游。通告方式为，把计算出来的 MTU 值放在 Label Mapping 消息的 MTU TLV 里面，然后把 Label Mapping 消息发送给上游。如果 MTU 发生变动，如本机出接口改变或者配置变更，那么 LSR 就应该再次通过 Label Mapping 消息，把重新计算过的 MTU 通告给它的所有上游。

2.3.13 LDP MD5

MD5 称为 Message-Digest Algorithm 5，是 RFC 1321 定义的国际标准摘要密码算法。MD5 的典型应用是针对一段信息（Message）计算出对应的信息摘要（Message-Digest），从而防止信息被篡改。MD5 信息摘要是通过不可逆的字符串变换算法产生的，结果唯一。因此，不管信息内容在传输过程中发生任何形式的改变，只要重新计算就会产生不同的信息摘要，接收端就可以由此判定收到的是一个不正确的报文。

基本原理

LDP MD5 就是应用其对同一信息段产生唯一摘要信息的特点来实现 LDP 报文防篡改校验，比一般意义上 TCP 校验和更为严格。

LDP MD5 验证是在 TCP 发出去之前进行的：LDP 消息在经 TCP 发出前，会在 TCP 头后面填充一个唯一的信息摘要再发出。而这个信息摘要就是把 TCP 头、LDP 消息、以及用户设置的 Password 一起作为原始信息，通过 MD5 算法计算出的。

当接收端收到这个 TCP 报文时，首先会取得报文的 TCP 头、信息摘要、LDP 消息，并结合 TCP 头、LDP 消息以及本地保存的 Password，利用 MD5 计算出信息摘要，然后与报文携带的信息摘要进行比较，从而检验报文是否被篡改过。

在用户设置 Password 时有明文和密文两种形式选择，这里的明文密文是对用户设置的 Password 在配置文件中的记录形式而言的。明文就是直接记录用户设置的字符串，密文就是记录经过特殊算法加密后的字符串。

但无论用户选择 Password 记录形态是明文还是密文形式，参与摘要计算时都是直接使用用户输入的字符串，也就是说私有加密算法计算出的 Password 并不会参与 MD5 摘要计算。由于明文和密文的转化算法各友商私有，此种实现做到了私有算法对友商透明。

2.3.14 LDP 认证

LDP MD5

MD5 称为 Message-Digest Algorithm 5，是 RFC 1321 定义的国际标准摘要密码算法。MD5 的典型应用是针对一段信息计算出对应的信息摘要，从而防止信息被篡改。MD5 信息摘要是通过不可逆的字符串变换算法产生的，结果唯一。因此，不管信息内容在传输过程中发生任何形式的改变，只要重新计算就会产生不同的信息摘要，接收端就可以由此判定收到的是一个不正确的报文。

LDP MD5 应用其对同一信息段产生唯一摘要信息的特点来实现 LDP 报文防篡改校验，比一般意义上 TCP 校验和更为严格。

LDP MD5 验证是在 TCP 发出去之前进行的：LDP 消息在经 TCP 发出前，会在 TCP 头后面填充一个唯一的信息摘要再发出。而这个信息摘要就是把 TCP 头、LDP 消息、以及用户设置的密码一起作为原始信息，通过 MD5 算法计算出的。

当接收端收到这个 TCP 报文时，首先会取得报文的 TCP 头、信息摘要、LDP 消息，并结合 TCP 头、LDP 消息以及本地保存的密码，利用 MD5 计算出信息摘要，然后与报文携带的信息摘要进行比较，从而检验报文是否被篡改过。

在用户设置密码时有明文和密文两种形式选择，这里的明文密文是对用户设置的密码在配置文件中的记录形式而言的。明文就是直接在配置文件中记录用户设置的字符串，密文就是在配置文件中记录经过特殊算法加密后的字符串。

但无论用户选择密码记录形态是明文还是密文形式，参与摘要计算时都是直接使用用户输入的字符串，也就是说私有加密算法计算出的密码并不会参与 MD5 摘要计算。由于明文和密文的转化算法各友商私有，此种实现做到了私有算法对友商透明。

LDP Keychain

Keychain 是一种增强型加密算法，类似于 MD5，Keychain 也是针对同一段信息计算出对应的信息摘要，实现 LDP 报文防篡改校验。

Keychain 允许用户定义一组密码，形成一个密码串，并且分别为每个密码指定加解密算法（包括 MD5，SHA-1 等）及密码使用的有效时间。在收发报文时，系统会按照用户

的配置选出一个当前有效的密码，并按照与此密码相匹配的加密解密算法以及密码的有效时间，进行发送时加密和接收时解密报文。此外，系统可以依据密码使用的有效时间，自动完成有效密码的切换，避免了长时间不更改密码导致的密码易破解问题。

Keychain 的密码、所使用的加解密算法以及密码使用的有效时间可以单独配置，形成一个 Keychain 配置节点，每个 Keychain 配置节点至少需要配置一个密码，并指定加解密算法。

Keychain 节点配置完成后，在全局 MPLS LDP 视图下指定需要引用 Keychain 节点的对等体和 Keychain 节点名称，对 Keychain 进行引用，即可实现对 LDP 会话的加密。不同的对等体可以引用同一个 Keychain 配置节点。

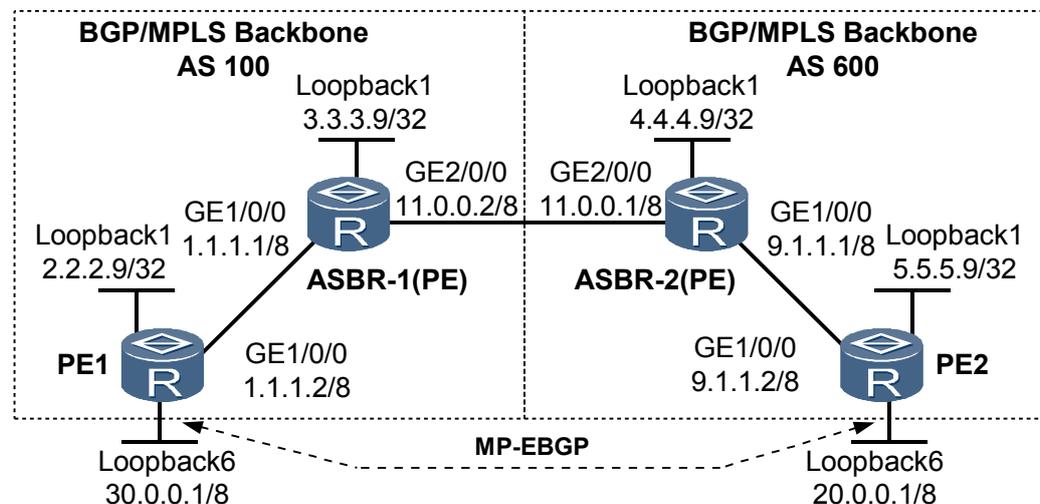
2.3.15 LDP 为 BGP 分标签

LDP 为 BGP 分标签是指：LDP 为带标签的公网 32 位掩码的 BGP 路由建立 LDP Egress LSP，以及 LDP 为公网 BGP 路由建立 Transit LSP。

LDP 为 BGP 路由分标签主要应用在 L3VPN 跨域 Option C 或者运营商的运营商场景，实现 LDP 为 BGP 路由分标签之后，跨域 Option C 场景中 PE 和 ASBR 之间不用再建立 full-mesh 的 IBGP 邻居关系（对应运营商的运营商场景，二级运营商的 PE 之间不用再建立 full-mesh 的 IBGP 邻居关系），易于扩展。

L3VPN 跨域 Option C

图 2-10 L3VPN 跨域 OptionC 组网拓扑



- 配置部署：
与普通 L3VPN 跨域 OptionC 的配置不同的是，PE1 和 ASBR1 之间，以及 PE2 和 ASBR2 之间不需要建立 IBGP 邻居，而是在两个 ASBR 上分别配置 LDP 为 BGP 分标签策略，并将 BGP 路由引入 IGP 协议发布，最终使 PE 上建立目的地址是到对端 PE 的 LDP ingress LSP 作为 L3VPN 的公网隧道。
- 路由发布过程：
以 PE1 的 Loopback 地址为例。PE1 将 2.2.2.9 / 32 通过 IGP 协议发布给 ASBR-1；ASBR-1 学到 2.2.2.9 / 32 的路由后，分配一个 BGP 标签，并通过 BGP 协议发布到 EBGP 邻居 ASBR-2；ASBR-2 学到 2.2.2.9 / 32 的带标签的公网 BGP 路由后，LDP

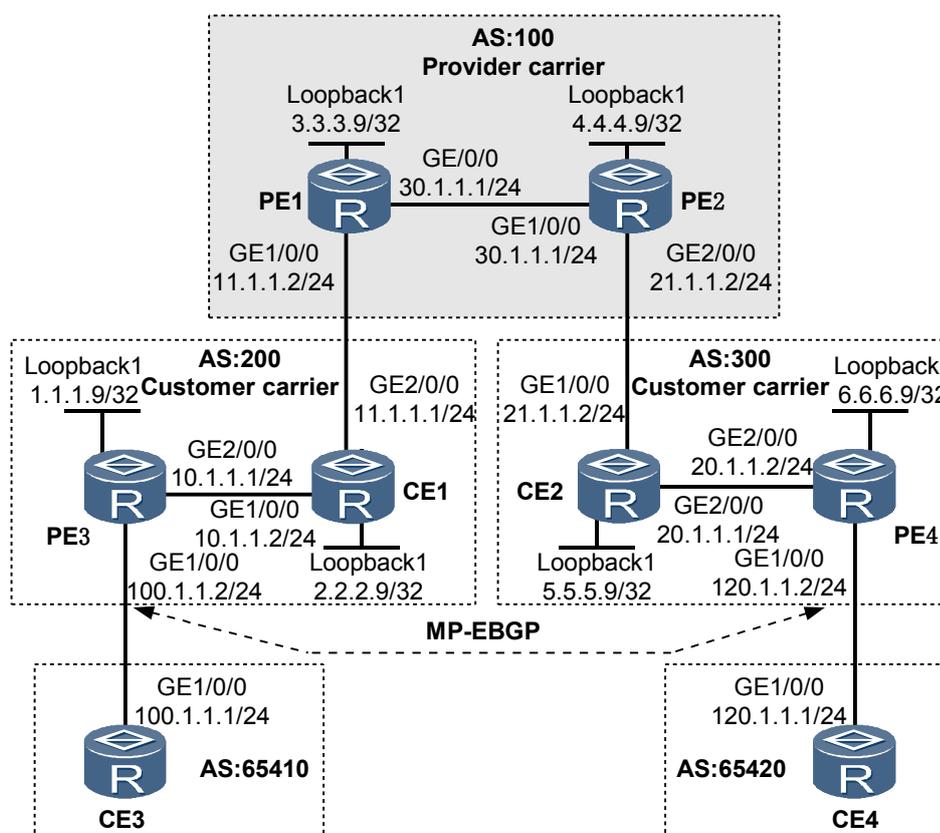
会为 2.2.2.9 / 32 建立 LDP Egress LSP，并将 2.2.2.9 / 32 通过 IGP 协议（此时已将 BGP 路由由 2.2.2.9 引入 IGP 协议）发布给 PE2；PE2 学到 2.2.2.9 / 32（此时是 IGP 路由），会为 2.2.2.9 建立 LDP Ingress LSP。

- 标签操作类型：

以私网路由 30.0.0.1 / 8 发出的报文为例，在 PE1 上压入一层 BGP 标签，再压入一层 LDP 标签；在 ASBR-1 上弹出外层 LDP 的标签，进入到 BGP LSP，并压入一层 BGP 标签；在 ASBR-2 上弹出外层的 BGP 标签，进入到 LDP LSP，并压入一层 LDP 标签；在 PE2 上弹出外层的 LDP 标签，弹出内层的 BGP 标签，走 IP 转发到 20.0.0.1 / 8。

运营商的运营商

图 2-11 运营商的运营商组网拓扑



- 配置部署：

如图 2-11 所示，PE3、CE1、CE2、PE4 是二级运营商的 PE 设备，PE1、PE2 是一级运营商的 PE 设备，CE3、CE4 为私网 CE 设备。与普通运营商的运营商配置不同的是，在 PE3 和 CE1 之间，以及 CE2 和 PE4 之间不需要建立 IBGP 邻居，而是分别在 CE1 和 CE2 上配置 LDP 为 BGP 分标签策略，并将 BGP 路由引入 IGP 协议发布，最终使 PE3 和 PE4 上分别建立到对端 PE 的 LDP Ingress LSP 作为二级运营商的公网隧道。

- 路由发布过程：

以 PE4 的 Loopback 地址为例。PE4 将 6.6.6.9 / 32 通过 IGP 协议发布给 CE2；CE2 学到 6.6.6.9 / 32 的路由后，分配一个 BGP 标签，并通过 BGP 协议发布到 EBGP 邻

居 PE2；PE2 学到 6.6.6.9 / 32 的私网 BGP 路由后，通过 BGP 协议发布到 IBGP 邻居 PE1；PE1 学到 6.6.6.9 / 32 的私网 BGP 路由后，通过 BGP 协议发布到 EBGP 邻居 CE1；CE1 学到 6.6.6.9 / 32 路由时，6.6.6.9 / 32 此时是带标签的公网 BGP 路由，LDP 会为 6.6.6.9 / 32 建立 LDP Egress LSP，并将 6.6.6.9 / 32 通过 IGP 协议（此时已将 BGP 路由 6.6.6.9 / 32 引入 IGP 协议）发布给 PE3；PE3 学到 6.6.6.9 / 32（此时是 IGP 路由），会为 6.6.6.9 / 32 建立 LDP Ingress LSP。

- 标签操作类型：

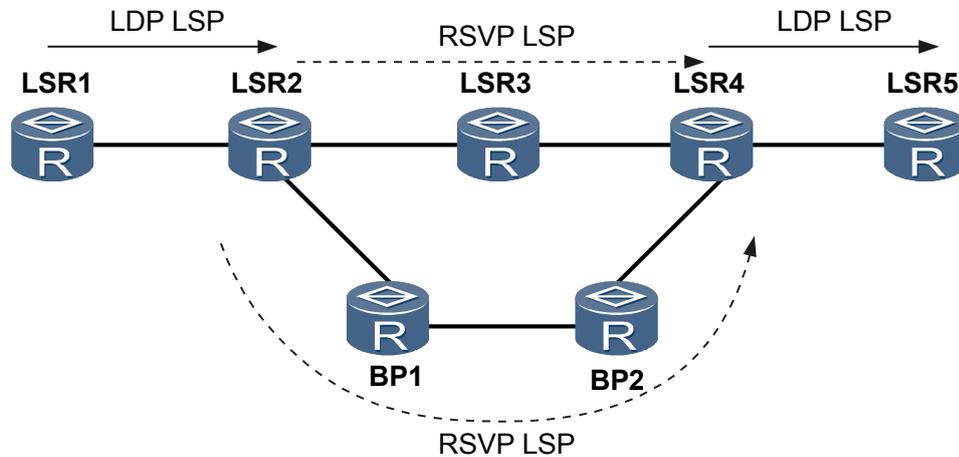
以私网路由 100.1.1.1 / 24 发出的报文为例，在 PE3 上压入一层 BGP 标签，再压入一层 LDP 标签；在 CE1 上弹出外层的 LDP 标签，进入到 BGP LSP，并压入一层 BGP 标签；在 CE2 上弹出外层 BGP 标签，进入到 LDP LSP，并压入一层 LDP LSP；在 PE4 上弹出外层的 LDP 标签，弹出内层的 BGP 标签，通过 IP 转发到达地址 120.1.1.1 / 24。

2.3.16 LDP over TE

RSVP TE 是一种 MPLS 隧道技术，生成的 LSP 可以提供给其他协议作为隧道，用来透传数据报文。LDP 是另一种 MPLS 隧道技术，用于生成 LDP LSP。LDP Over TE 是一个用于让 LDP 的 LSP 跨过 RSVP TE 区域，提供给 VPN 服务器使用的技术。从 VPN 应用的现状来看，为实现 MPLS 流量工程，运营商在整个网络布置 TE 的难度比较大，因此运营商可以规划一个核心的 TE 区域，在此区域里部署 TE，然后在此区域外运行 LDP。

在 RSVP TE 隧道建立完毕后，IGP（如 OSPF、ISIS）可以通过本地的计算，或者通过发布 LSA，让路由的出接口选中 TE 的隧道。可以认为起始设备和 TE 隧道的目的设备通过 TE 的隧道接口（逻辑接口）直接相连，报文实际转发时，是通过 TE 的隧道透传。

图 2-12 LDP over TE 组网拓扑



如图 2-12 所示，整个网络为 MPLS VPN 网络，信令协议为 LDP，提供普通的 VPN 业务，LSR1 和 LSR5 为 PE，在接入大量客户后，发现 LSR1 和 LSR5 之间的流量都要经过此链路，因此 LSR2 到 LSR3 的链路拥塞。而 LSR2 与 BP1 之间的链路非常空闲，但由于 IGP 的 cost 值的问题，LSP 不能用到 LSR2 和 BP1 之间的链路。

在这种情况下，可以通过部署 LDP Over TE，在 LSR2 和 LSR4 之间建立一条穿过 BP1、BP2 的 TE 隧道，通过调整 IGP 的 cost 值，让路由在 LSR2 上有两个负载分担：

- LSR2 和 LSR3 的物理接口

- LSR2 到 LSR4 的 TE 隧道接口

这样 LDP 就可以建立起负载分担的 LSP，使部分流量从空闲的链路通过。

此方法的优点：可以任意建立 TE 隧道，可以是一条，也可以是多条，而且只要发现有空闲的路径，都可以适用。与调整 IGP 的 cost 值有巨大的优势，也是目前流量工程比较广泛的应用场景。

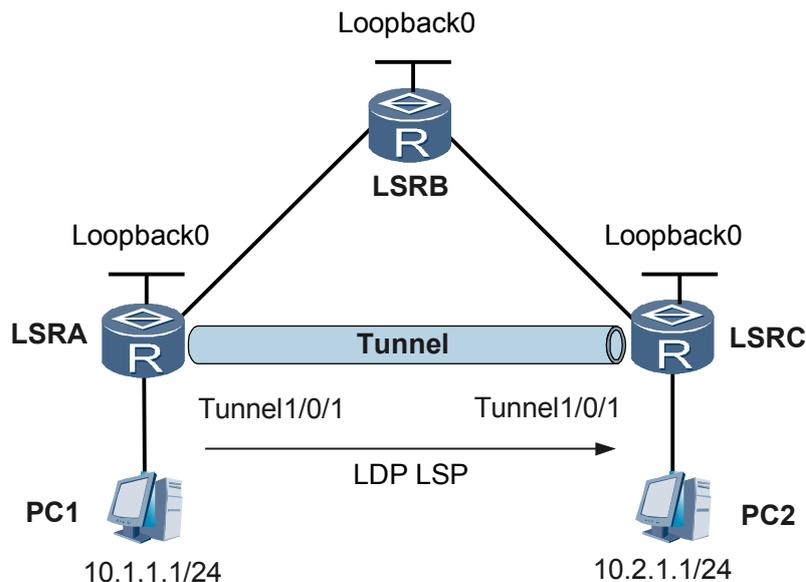
2.3.17 LDP over GRE

GRE（Generic Routing Encapsulation）是通用路由封装协议，提供了将一种协议的报文封装在另一种协议报文中的机制，使报文能够在异种网路中传输，而异种报文传输的通道称为隧道（Tunnel）。例如将某些网路层协议（如 IPX、ATM、IPV6、AppleTalk 等）的数据报进行封装，使这些被封装的数据报能够在 IPv4 网络中传输。

LDP over GRE 即通过在 GRE Tunnel 接口上配置 MPLS LDP，使 LDP 隧道建立在 GRE Tunnel 接口上，跨越 GRE 隧道承载 L2VPN 业务。

如图 2-13，LSRA、LSRB、LSRC 属于 VPN 骨干网。LSRA 和 LSRC 之间使用三层隧道协议 GRE，实现 PC1 和 PC2 互联。为了给 L2VPN 提供服务，需要在 LSRA 和 LSRC 之间建立一条跨越 GRE 隧道的 LDP LSP，实现 LDP over GRE。

图 2-13 LDP over GRE 拓扑



2.3.18 LDP GTSM

通用 TTL 安全保护机制 GTSM（Generalized TTL Security Mechanism）是一种通过检查 IP 报文头中的 TTL 值是否在一个预先定义好的范围内来实现对 IP 业务进行保护的机制。

使用 GTSM 的两个前提：

- 设备之间正常报文的 TTL 值确定
- 报文的 TTL 值很难被修改

GTSM 的应用范围包括 BGP（含 IPV6）、OSPF、LDP、RSVP、OSPFv3 等。

基本原理

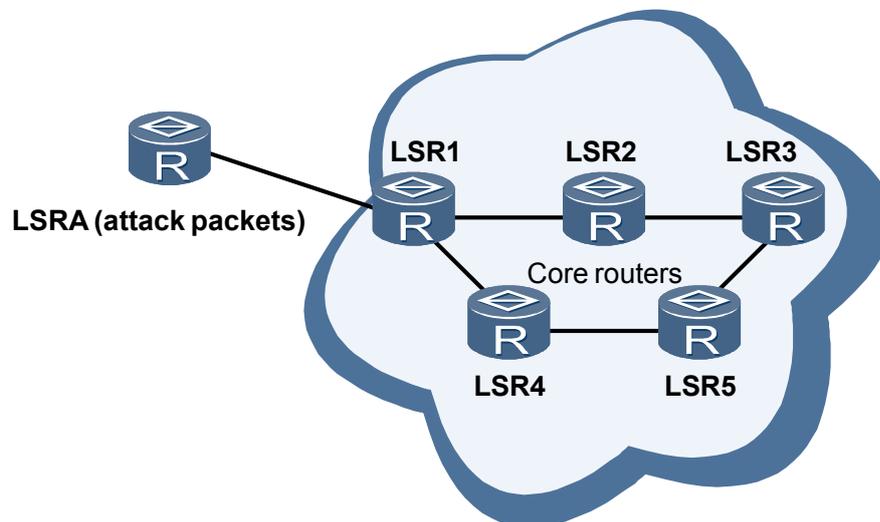
LDP GTSM 是 GTSM 在 LDP 方面的具体应用。

GTSM 通过判定报文的 TTL 值，确定报文是否有效，从而保护设备免受攻击。GTSM For LDP 即是对相邻或相近（基于只要跳数确定的原则）设备间的 LDP 消息报文应用此种机制，预先在各路由上设定好针对其他设备报文的有效范围，使能 GTSM，这样当相应设备之间应用 LDP 时，如果 LDP 消息报文的 TTL 不符合之前设置的范围要求，就认为此报文为非法攻击报文予以丢弃，进而实现对上层协议的保护。

应用场景

GTSM 在实际应用中，主要服务于保护建立在 TCP/IP 层上的控制层面免受 CPU-utilization（CPU overload）类型的攻击。应用于 LDP，就是将 GTSM 应用于 LDP 各种消息报文，以免 LDP 协议在收到大量伪装报文时，因处理报文导致 CPU-utilization 等情况的攻击。

图 2-14 LDP GTSM 组网拓扑



如图 2-14 所示，LSR1-LSR5 为骨干网中核心路由器，当 LSRA 通过其他设备与核心路由器间接相连时，LSRA 可能会伪造 LSR1 ~ LSR5 之间的 LDP 相关报文来达到攻击的目的。

当 LSRA 经过其他设备接入后，伪造报文中携带的 TTL 值被认为是不可伪造的，这就是 GTSM 的前提。

在 LSR1 ~ LSR5 上分别配置到各个可能邻居的 GTSM 策略：如在 LSR5 上配置 LSR2 发送的报文有效跳数为 1 ~ 2，有效 TTL 值为 254 ~ 255。当 LSRA 发送的伪造报文到达 LSR5 时，由于经过多跳中间设备，导致 TTL 比预先设置的有效范围小，因此 LSR5 可以直接把伪造报文丢弃，避免了攻击。

2.3.19 LDP 本远端会话共存

基本原理

本特性主要应用于 L2VPN 业务。

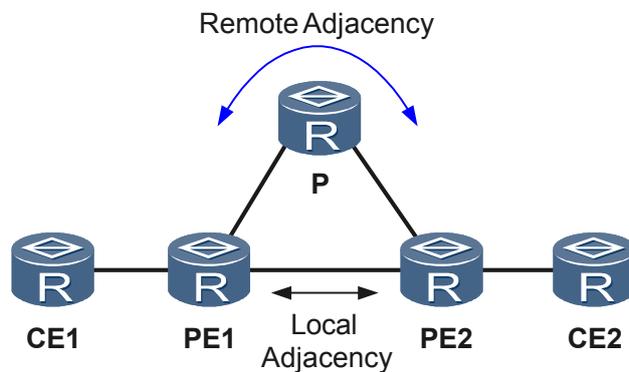
LDP 本远端会话共存，其本质是 LDP 的本地邻接体和远端邻接体可以同时绑定到同一个对等体上，也就是说该对等体同时由本地、远端两种邻接体来维持。

如图 2-15 所示，当和本地邻接体相关的链路发生故障，导致该邻接体被删除时，只可能会改变该对等体的类型（对等体类型由维持对等体的邻接体的类型来决定，可以在本地、远端、本远共存之间进行切换），不会影响到对等体的存在和状态。

在链路发生故障或者故障恢复的过程中，对等体类型可能发生改变，和该对等体对应的会话类型也随之发生变化，但是，在整个过程中会话都不会被删除，也不会被置 Down，始终能够起作用。

应用场景

图 2-15 LDP 本远端会话共存组网拓扑



本方案的典型应用场景是 L2VPN。如图 2-15PE1 和 PE2 之间部署 L2VPN 业务，当 PE1 和 PE2 之间的直连链路断开又恢复时的处理过程如下：

1. 在直连的两台设备上，建立本远共存会话，包括本地和远端两个邻接体。L2VPN 信令消息通过此会话传递。
2. PE1 和 PE2 之间的物理链路进入 Down 状态，导致 LDP 对等体的本地邻接体进入 Down 状态。PE1 和 PE2 之间的路由通过 P 可达，即远端邻接体依然处于 Up 状态，会话类型发生变化，成为远端会话，但会话依然处于 Up 状态，L2VPN 感知不到会话状态变化，不会主动拆除，这就避免了 L2VPN 断开邻居并重新恢复的过程，减少业务中断时间。
3. 随后故障恢复，PE1 和 PE2 之间的链路进入 Up 状态，随后本地邻接体随之进入 Up 状态。会话类型发生变化，恢复为本远共存会话，会话依然处于 Up 状态，L2VPN 感知不到会话状态变化，不会主动拆除，这也减少了业务中断时间。

2.3.20 LDP 为所有 Peer 分标签

本子特性主要是为了解决目字形组网发生链路故障收敛比较慢的问题而提出的。

在只给上游 Peer 分标签的情况下，发送标签 Mapping 消息的时候，要根据路由信息对会话的上下游关系进行确认。对于某一条路由，上游节点不向下游节点发送标签 Mapping 消息；如果发生路由变化，上下游关系倒换，新的下游需要重新给上游节点发送标签 Mapping 消息，收敛比较慢。

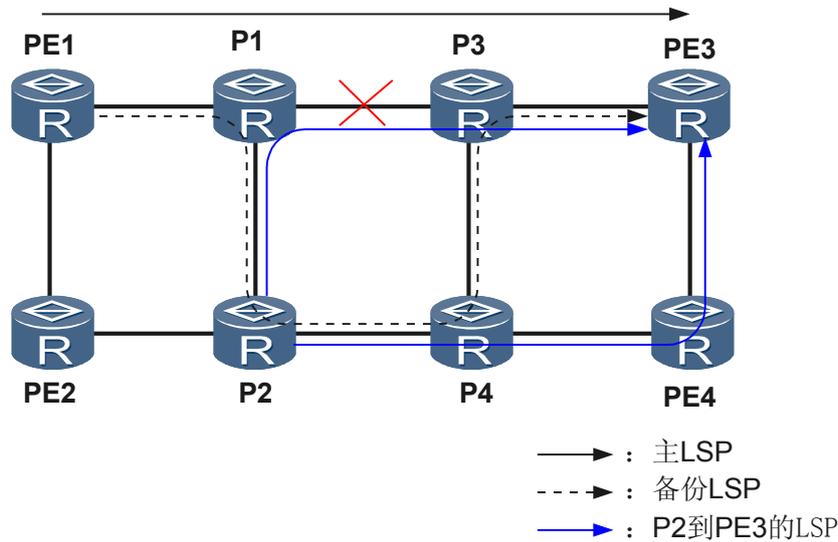
在本子特性完成后，对所有符合条件的对等体都要发送标签 Mapping 消息，不再区分上下游关系。即每个节点都可以向所有的对等体发送标签 Mapping 消息。

如图 2-16 所示，P2 原来到 PE3 的路径为 P2→P1→P3→PE3 和 P2→P4→PE4→PE3，对于 PE3 上的 Loopback 接口路由，P1 是 P2 的路由的下一跳。当 P2 从 P1 上接收到标签 Mapping 消息之后，只向上游分标签的情况下，P2 不会向 P1 发送关于这个路由的标签 Mapping 消息，这样当 P1P3 之间发生链路故障，从 PE1 到 PE3 的路径由 PE1→P1→P3→PE3 切换为 PE1→P1→P2→P4→P3→PE3，P2 切换为 P1 的下游，但是由于 P2 没有向 P1 发送标签 Mapping 消息，必须等待重新发送标签 Mapping 消息后才能使 LSP 重新收敛，速度比较慢。

在 LDP 为所有 Peer 分标签的情况下，P2 从 P1 上接收到标签 Mapping 消息之后，会直接向 P1 发送关于这个路由的标签 Mapping 消息，并在 P1 上生成一个 Liberal LSP。这样当 P1P3 之间发生链路故障，从 PE1 到 PE3 的路径由 PE1→P1→P3→PE3 切换为 PE1→P1→P2→P4→P3→PE3，P2 切换为 P1 的下游，Liberal LSP 直接变化为 Normal LSP，LSP 收敛速度加快。

另外，通过配置水平分割命令可以用来决定为哪些下游对等体发送标签 Mapping 消息，对哪些下游对等体不能发送标签 Mapping 消息。

图 2-16 LDP 为所有 Peer 分标签组网拓扑



2.4 术语与缩略语

术语

术语	解释
GR Restarter	GR 重启节点，指由管理员触发或故障触发重启的节点，必须具备 GR 能力。
GR Helper	Restarter 的邻居，必须具备 GR 能力。

缩略语

缩略语	英文全称	中文全称
LDP	Label distribution protocol	标签分发协议
LSP	Label switched path	标签转发路径
FEC	Forwarding Equivalence Class	转发等价类
GTSM	Generalized TTL Security Mechanism	通用 TTL 安全保护机制
TTL	Time to Live	报文生命周期
GR	Graceful Restart	优雅重启
FRR	Fast Re-Route	快速重路由
MTU	Maximum Transmission Unit	最大传输单元

3 MPLS TE

关于本章

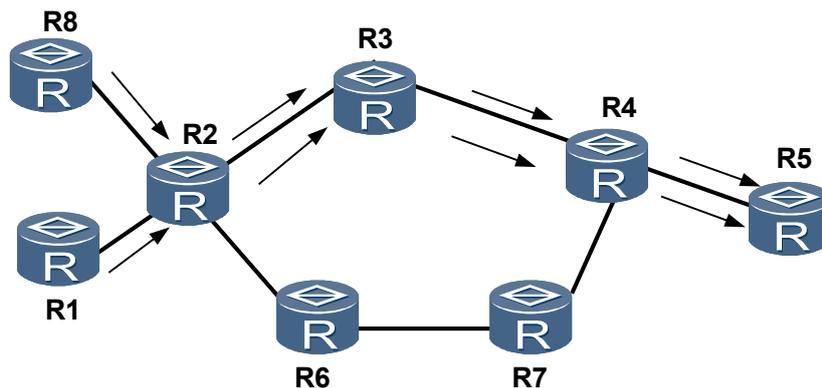
- 3.1 介绍
- 3.2 参考标准和协议
- 3.3 原理描述
- 3.4 术语与缩略语

3.1 介绍

目的

传统的 IP 网络中，节点选择最短的路径作为路由，不考虑带宽等因素。这样，容易出现流量集中于最短路径而导致拥塞。其他可选的路径则较为空闲。

图 3-1 传统路由的困难



如图 3-1 所示，假设每个链路的 metric 值相同。R1（R8）到 R5 的最短路径为 R1（R8）→R2→R3→R4→R5，尽管存在其他到达 R5 的路径，数据转发也走 R1（R8）→R2→R3→R4→R5 这条最短路径。这样，就可能出现一条链路 R1（R8）→R2→R3→R4→R5 过载而出现拥塞，而另外一条链路 R1（R8）→R2→R6→R7→R4→R5 空闲的情况。

网络拥塞是影响骨干网络性能的主要问题。拥塞的原因可能是网络资源不足，也可能是网络资源负载不均衡，导致局部拥塞。流量工程解决的是由于负载不均衡导致的拥塞。为解决网络拥塞，传统的流量工程通常采取如下方法：

- 通过调整路径 Metric 而控制网络流量：这种解决方法能够解决某些链路上的拥塞，但是可能会引起另外的链路拥塞。另外，在拓扑结构复杂的网络上，Metric 值的调整比较困难，往往一条链路的改动会影响多条路由，难以把握和权衡。
- 采用叠加的网络模型，建立虚连接引导部分流量：现有的 IGP 协议都是拓扑驱动，如果只考虑网络的连接情况，不能灵活反映带宽和流量特性这类动态状况。

解决 IGP 缺点的一种方法是使用重叠模型（Overlay），如 IP over ATM、IP over FR 等。重叠模型在网络的物理拓扑结构上提供了一个虚拟拓扑结构，从而容易实现流量的合理调配和良好的 QOS 功能。然而，重叠模型的额外开销大，可扩展性差。

为了在大型骨干网络中部署流量工程，必须采用一种可扩展性好、简单的解决方案。MPLS（Multiprotocol Label Switching）作为一种叠加模型，可以方便地在物理的网络拓扑上建立一个虚拟的拓扑，然后将流量映射到这个拓扑上。因此，MPLS 与流量工程相结合的技术——MPLS TE（Traffic Engineering）应运而生。

定义

MPLS TE（MPLS Traffic Engineer），即 MPLS 流量工程。可以通过 RSVP-TE 协议建立基于约束的 LSP（CR-LSP）隧道，这些约束可以包括：带宽、优先级、亲和属性、显

式路径、CSPF 仲裁策略、metric 类型、跳数限制、SRLG 等；也可以通过静态配置 LSP 来建立 CR-LSP，我们称之为静态 CR-LSP。

MPLS TE 的特征：MPLS TE 结合了 MPLS 技术与流量工程，通过建立经过指定路径的 LSP 进行资源预留，使网络流量绕开拥塞节点，达到平衡网络流量的目的。

MPLS 有实现流量工程所需要的特征，例如：

- 支持建立显式 LSP，可以对路径进行控制；
- 通过信令建立 LSP，配置起来比较简单，容易维护；
- 建立 LSP 的负荷小，不会影响网络的其它业务；
- 网络流量可以方便地映射到某条 LSP 上；
- LSP 有优先级、抢占等多种属性，可以方便地控制 LSP 的行为；
- 管理组和亲和属性可以控制 LSP 经由的路径；
- MPLS 允许流量聚合和非聚合两种方式，比 IP 灵活；
- MPLS 容易集成约束路由。

通过 MPLS TE，可以解决一部分路径过载而另一部分路径空闲的问题，充分利用现有的带宽资源。同时，MPLS TE 在建立 LSP 的过程中，可以通过预留资源来保证服务质量。

为了保证服务的连续性，MPLS TE 还引入路径备份和快速重路由 FRR（Fast Reroute）的机制，在链路故障时可及时进行切换。

MPLS TE 主要包括如下相关属性。

- CR-LSP

基于一定约束条件建立的 LSP 称为 CR-LSP（Constraint-based Routed Label Switched Path）。与普通 LSP 不同，CR-LSP 的建立不仅依赖路由信息，还需要满足其他一些条件，比如指定的带宽、颜色、建立优先级、保持优先级、显式路径和 QoS 参数。

CR-LSP 可分为两类：

- 静态 CR-LSP：通过手工配置转发信息和资源信息，不涉及信令协议和路径计算。由于不需要交互 MPLS 相关控制报文，消耗资源比较小，但静态 CR-LSP 不能根据网络的变化动态调整，实际应用有限。
- 动态 CR-LSP：通过信令机制建立和维护，需要进行路径计算。

- RSVP-TE

为了能够建立 CR-LSP，对 RSVP 协议进行了扩展。扩展后的 RSVP（Resource Reservation Protocol）信令协议称为 RSVP-TE 信令协议。

- 带宽

带宽是 MPLS TE 的基本属性之一，在建立 CR-LSP 时，可以指定其带宽，根据带宽要求确定合适的路径。RSVP-TE 信令可以携带带宽信息，在沿路径的各个节点实施带宽预留。CR-LSP 建立成功后，能够为对应的业务提供所要求的带宽保证。

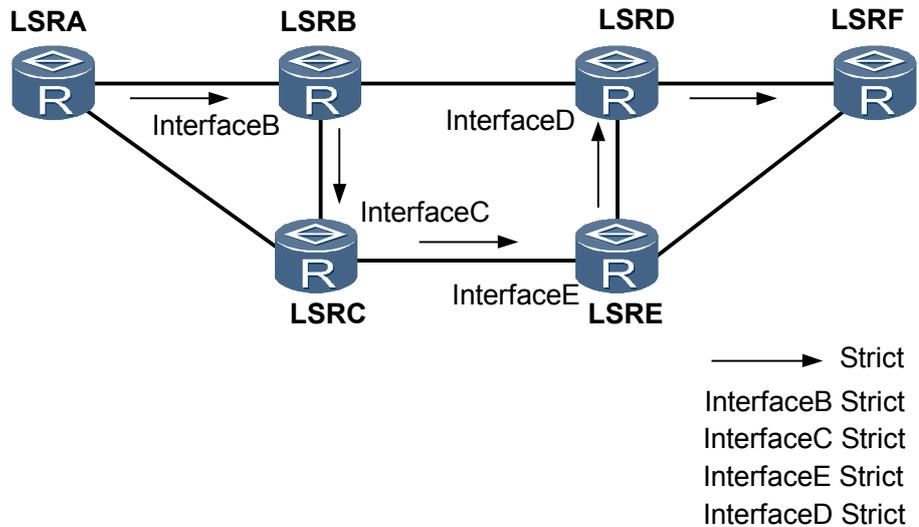
- 显式路径

CR-LSP 能够按照指定的路径建立，该路径称为显式路径。显式路径可以分为如下两种：

- 严格显式路径

所谓的严格显式路径，就是路径上每个节点都至少有一个接口的 IP 地址被指定。通过严格显式路径，可以最精确地控制 LSP 所经过的路径。

图 3-2 严格显式路径

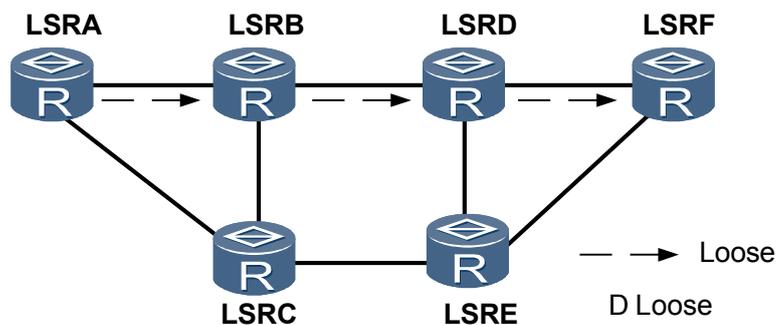


在图 3-2 中，LSRA 作为 LSP 的入节点，LSRF 作为出节点，从 LSRA 到 LSRF 用严格显式路径建立一条 LSP。“InterfaceB strict”表示该 LSP 必须经过 LSRB 的 InterfaceB 接口，并且 LSRB 的前一跳是 LSRA，“InterfaceC strict”表示该 LSP 必须经过 LSRC 的 InterfaceC 接口，并且 LSRC 的前一跳是 LSRB，依此类推，就可以精确控制该条 LSP 所经过的路径。

- 松散显式路径

松散方式可以指定路径上必须经过哪些节点，而不严格指定路径。

图 3-3 松散显式路径



在图 3-3 中，从入节点 LSRA 到出节点 LSRF 用松散显式路径建立一条 LSP。“D loose”表示该 LSP 必须经过 LSRD，但是 LSRD 与 LSRA 之间可以存在其他 LSR 节点，不必直接相连。

MPLS TE 信令能够携带显式路径及其严格或松散属性，按照指定的路径建立 CR-LSP。

ERO (Explicit Route Object) 描述 LSP 经过的路径信息，可以为严格显式路径也可以是松散显式路径。Path 消息沿 ERO 指定的路径转发，不受 IGP 最短路径约束。

● CSPF 仲裁策略

CSPF 在计算路径的过程中，如果遇到多条权值相同的路径，将根据策略选择其中的一条。这个过程称为仲裁（tie-breaking）。可用的仲裁策略有：

- **Most-fill**: 选择已用带宽和最大可预留带宽的比值最大的链路，使链路带宽资源高效使用；
- **Least-fill**: 选择已用带宽和最大可预留带宽的比值最小的链路，使各条链路的带宽资源均匀使用；
- **Random**: 随机选取，使每条链路上的 LSP 数量均匀分布，不考虑带宽因素。

在比率相同的情况下，比如没有利用保留带宽，或者利用的份额都是一样，此时不管配置的是 least-fill 还是 most-fill，选择的是首先发现的链路。

- **优先级和抢占**

如果在建立 CR-LSP 的过程中，无法找到满足所需带宽要求的路径，则拆除另外一条优先级较低的 CR-LSP，占用为它分配的带宽资源，这种处理方式称为抢占（Preemption）。

CR-LSP 使用两个优先级属性来决定是否可以进行抢占：建立优先级（Setup Priority）、保持优先级（Holding Priority）。优先级的取值为从 0 到 7 的整数。值越小，优先级越高。7 为最低优先级。

- **链路管理组和亲和属性**

链路管理组是一个表示链路属性的 32 位向量。

亲和属性是一个 32 位向量，表示隧道链路的颜色。为隧道配置亲和属性后，隧道选择链路时，将亲和属性和链路管理组属性进行比较，决定选择还是避开特定属性的链路。

在隧道入节点配置隧道亲和属性后，亲和属性将通过 RSVP-TE 信令协议传递给 CR-LSP 途经各节点。

- **跳数限制**

跳数限制值作为 CR-LSP 建立时的选路条件之一，就像管理组和亲和属性一样，可以限制一条 CR-LSP 允许选择的路径跳数不超过限制值。

- **SRLG**

共享风险链路组 SRLG（Shared Risk Link Group）是一组共享一个公共的物理资源（或是共享一根光纤）的链路。同一个 SRLG 的链路具有相同的风险等级，即如果 SRLG 中的一条链路失效，组内的其他链路也失效。

SRLG 主要用在 CR-LSP 热备份和 TE FRR 组网中增强 TE 隧道的可靠性。共享相同物理资源的多个链路具有相同的风险。例如，当主接口 Down 时，其子接口也通常变为 Down，所以子接口和主接口通常具有相同的风险。如果主隧道所在的链路和其备份隧道所在链路具有相同的风险，则主隧道变为 Down 时，备份隧道也很可能变为 Down。所以 SRLG 应该作为备份隧道路径计算的限制条件。

3.2 参考标准和协议

本特性的参考资料清单如下：

文档编号	描述	备注
RFC 2205	Resource ReSerVation Protocol	-
RFC 2209	Resource ReSerVation Protocol (RSVP) -- Version 1 Message Processing Rules	-

文档编号	描述	备注
RFC 2370	The OSPF Opaque LSA Option	-
RFC 2547	BGP/MPLS VPNs	-
RFC 2702	Requirements for Traffic Engineering Over MPLS	-
RFC 2747	RSVP Cryptographic Authentication	-
RFC 2961	RSVP Refresh Overhead Reduction Extensions	-
RFC 3031	Multiprotocol Label Switching Architecture	-
RFC 3032	MPLS Label Stack Encoding	-
RFC 3034	Use of Label Switching on Frame Relay Networks Specification	-
RFC 3209	RSVP-TE: Extensions to RSVP for LSP Tunnels	-
RFC 3210	Applicability Statement for Extensions to RSVP for LSP-Tunnels	-
RFC 3473	Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions	-
RFC 3630	Traffic Engineering (TE) Extensions to OSPF Version 2	-
RFC 3784	Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)	-
RFC 4124	Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering	-
RFC 4127	Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering	-
RFC4128	Bandwidth Constraints Models for Differentiated Services (Diffserv)-aware MPLS Traffic Engineering: Performance Evaluation	-
RFC 4139	Requirements for Generalized MPLS (GMPLS) Signaling Usage and Extensions for Automatically Switched Optical Network (ASON)	-
draft-ietf-mpls-rsvp-lsp-fastreroute-02	Fast Reroute Extensions to RSVP-TE for LSP Tunnels	-
draft-ietf-mpls-nodeid-subobject-01	Definition of an RRO node-id subobject	-
draft-ietf-tewg-diff-te-02	Protocol extensions for support of Diff-Serv-aware MPLS Traffic Engineering	-

文档编号	描述	备注
draft-ietf-mpls-diff-te-reqts-00	Requirements for support of Diff-Serv-aware MPLS Traffic Engineering	-
draft-ietf-mpls-diff-ext-07	MPLS Support of Differentiated Services	-

3.3 原理描述

MPLS TE 通过建立一条保证带宽的 TE 隧道来进行标签转发及流量保证。

3.3.1 RSVP-TE 协议

3.3.2 Make-before-break

3.3.3 P2MP RSVP-TE

3.3.4 自动带宽调整

3.3.5 重优化

3.3.6 TE FRR

3.3.7 SRLG

3.3.8 CR-LSP 备份

3.3.9 DS-TE

3.3.10 TE 隧道保护组

3.3.11 BFD For TE CR-LSP

3.3.12 BFD for TE Tunnel

3.3.13 RSVP 认证

3.3.14 RSVP GR

3.3.15 RSVP 摘要刷新

3.3.16 RSVP Hello

3.3.17 BFD for RSVP

3.3.18 TE-LSP 配置模板

3.3.19 MPLS LSR-ID 的多区域发布

3.3.1 RSVP-TE 协议

资源预留协议 RSVP (Resource Reservation Protocol) 是为 Integrated Service 模型而设计的, 用于在一条 LSP 的各节点上进行资源预留。RSVP 工作在传输层, 但不参与应用数据的传送, 是一种网络上的控制协议。RSVP-TE 是对 RSVP 的扩展, 通过扩展对象支持 TE (Traffic Engineering) 的相关属性, 用于基于约束的 LSP 的建立和删除。

简单来说，RSVP 具有以下几个主要特点：

- 单向；
- 面向接收者，由接收者发起对资源预留的请求，并维护资源预留信息；
- 使用“软状态”（soft state）机制维护资源预留信息。

RSVP-TE 对 RSVP 扩展的内容主要有：

- 在 RSVP 的 PATH 消息中引入 Label Request 对象，支持发起标签请求；在 RSVP Resv 消息中引入 Label 对象支持标签分配。这样就可以建立 CR-LSP 了。
- 扩展的消息除了可以携带标签绑定信息外，还可以携带限制信息，从而支持 CR-LSP 约束路由的功能。
- 此外，RSVP-TE 通过扩展对象支持 MPLS-TE 相关的属性，使其具有资源预留功能。

RSVP-TE 主要消息

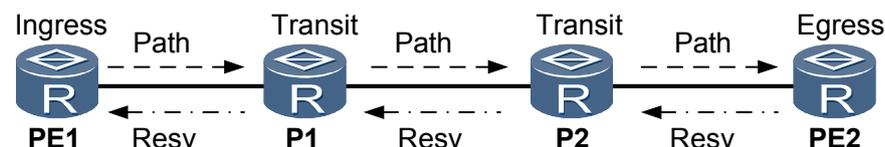
RSVP-TE 消息有：Path 消息、Resv 消息、PathErr 消息、ResvErr 消息、PathTear 消息、ResvTear 消息、ResvConfirm 消息、Hello 消息、Ack 消息、Srefresh 消息、GRPath 消息和 Recovery Path 消息。

RSVP-TE 主要包括如下消息：

- Path 消息：Path 消息用于发送者请求下游节点为此路径分配标签。途经每一个节点时记录路径信息，并且建立路径状态块 PSB（Path State Block）。
- Resv 消息：Resv 消息用于在各个节点预留资源。Resv 消息携带了发送者申请的资源预留信息，沿着数据流的反方向发送，在沿途节点创建预留状态块 RSB（Reserved State Block），记录分配的标签信息。
- PathErr 消息：RSVP 节点在处理 Path 消息的时候，如果发生错误，就向上游发送 PathErr 消息。中间节点收到 PathErr 消息后，继续向上游转发，直至入节点。
- ResvErr 消息：RSVP 节点在处理 Resv 消息时，如果发生错误，就向下游发送 ResvErr 消息；中间节点收到 ResvErr 消息后，继续向下游转发消息，直至出节点。
- PathTear 消息：由入节点向下游发送，用于删除各个节点创建好的本地状态。
- ResvTear 消息：由出节点向上游发送，用于删除对应的本地资源等。入节点收到 ResvTear 消息后，向下游发送 PathTear 消息。

LSP 建立过程

图 3-4 LSP 建立过程图



如图 3-4 所示，LSP 的建立过程如下：

1. Ingress 节点收到建立 LSP 的消息后创建 PSB，向下游发送 Path 消息。

2. Transit 节点处理并转发 Path 消息，在各个节点根据 Path 消息创建 PSB。
3. Egress 节点收到 Path 消息后创建 PSB，根据 Path 消息生成 Resv 消息，同时创建 RSB 等状态块，并且向上游发送 Resv 消息。
4. Transit 节点处理并转发 Resv 消息，创建 RSB 等状态块。
5. Ingress 节点收到 Resv 消息后，创建 RSB 等状态块，确认资源预留成功。

至此，一条 LSP 建立成功。

软状态

RSVP 节点周期性的发送 RSVP 刷新消息，用于在 RSVP 邻居节点进行状态（包括 PSB 和 RSB）同步，或恢复丢失的 RSVP 消息，这就是 RSVP 的“软状态”机制。对于某个状态，如果在指定刷新周期内没有收到刷新消息，这个状态将被删除。

当有状态需要刷新时，节点会创建对应的刷新消息，并发送给它的后续节点。当路由发生变化时，如果使能隧道重优化，下一个 Path 消息会基于新的路由初始化路径状态，之后的 Resv 消息将在新路径上建立预留状态。不再使用的路径状态将会超时删除。

预留风格

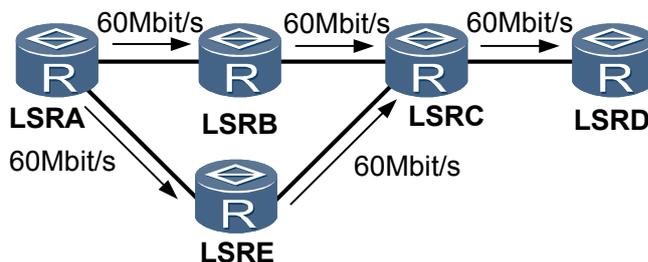
资源预留风格是指 RSVP 节点处理上游节点的资源预留请求时的资源预留方式。当前华为支持的资源预留风格包括：

- FF (Fixed Filter)：即固定过滤风格，为每个发送者创建固定的预留，每个发送者的预留都是相互独立的，并且分配唯一的标签。
- SE (Shared Explicit)：即共享显式风格，为接收者显式地指定一个预留中的发送者，这些发送者共享同一个预留，分配不同的标签。

3.3.2 Make-before-break

Make-before-break 是一种高可靠性的 CR-LSP 切换技术。在创建新的 CR-LSP 时，先不删除原有的 CR-LSP。当新的 CR-LSP 建立成功后，先将流量切换到新的 CR-LSP 上，再删除原来的 CR-LSP，保证业务流量不中断。

图 3-5 Make-before-break 基本原理



Make-before-break 机制常用的方式如下：

- 修改路径
如 [图 3-5](#)，假设所有链路最大可预留带宽为 60Mbit/s。新建一条 LSRA 到 LSRD 的 CR-LSP，路径是 LSRA→LSRB→LSRC→LSRD，带宽为 40Mbit/s。现在希望将路径改为 LSRA→LSRE→LSRC→LSRD，通过负载较轻的 LSRE 进行数据转发。但是

LSRC→LSRD 上剩余的可预留带宽只有 20Mbit/s，不足 40Mbit/s。这时可以通过 Make-before-break 机制来解决。

通过 Make-before-break，新建立的路径 LSRA→LSRE→LSRC→LSRD 在 LSRC→LSRD 上进行资源预留时采用原路径使用的带宽。新隧道建立成功后，流量转到新路径上后拆除原路径。

- 增加隧道的带宽

只要共用链路的可预留带宽满足增量要求，新的 CR-LSP 就可以建立成功。如图 3-5，假设所有链路最大可预留带宽为 60Mbit/s。新建一条 LSRA 到 LSRD 的 CR-LSP，带宽为 30Mbit/s，路径是 LSRA→LSRB→LSRC→LSRD。现在希望将路径改为 LSRA→LSRE→LSRC→LSRD，通过负载较轻的 LSRE 进行数据转发，并将带宽增大为 40Mbit/s。但是 LSRC→LSRD 上剩余的可预留带宽只有 30Mbit/s，不足 40Mbit/s。这时可以通过 Make-before-break 机制来解决。

通过 Make-before-break，新建立的 CR-LSP 的路径 LSRA→LSRE→LSRC→LSRD 在 LSRC→LSRD 上进行资源预留时采用原路径使用的带宽，并追加增量带宽。新的 CR-LSP 建立成功后，流量转到新路径上后拆除原路径。

3.3.3 P2MP RSVP-TE

目前多播业务存在以下问题：

- 带宽浪费：传统的核心网和骨干网通常使用 IP/MPLS 进行数据报文的转发。这种部署对于转发单播报文而言，具有很高的灵活性，并且提供了充分的可靠性和流量工程的能力。但是对于转发多播报文而言，需要对每个接收者进行报文的复制，这样导致了带宽的浪费。
- 业务中断：传统的基于 PIM 协议的组播路由协议在网络中根据路由逐跳进行通告和建立组播转发树，无法避开拥塞节点。此外，在发生网络链路故障时，收敛时间比较长，从而导致组播业务中断。

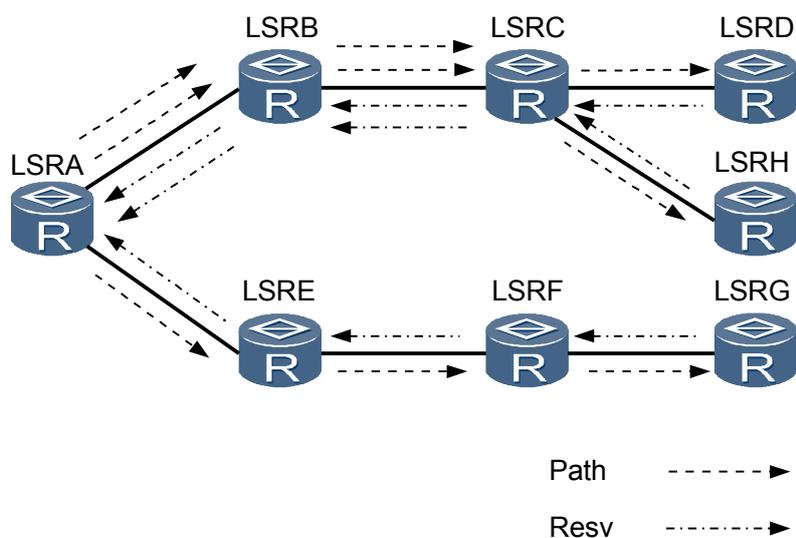
P2MP RSVP-TE 可以较好的解决上述问题，它对 MPLS 协议进行扩展以满足建立 P2MP 路径的需求。同时 MPLS 可以提供有效的可靠性和 QoS 保证，提供有效的流量工程能力。在现有的 IP/MPLS 网络上，通过简单升级即可获得多播的能力，可有效的节省升级和维护成本，符合网络融合的趋势。

P2MP RSVP-TE 隧道的建立

P2MP RSVP-TE 隧道的建立过程如下：

- 在首节点指定所有叶子。从首节点到达叶子的路径可以通过规划显式路径或者动态计算两种方法得到。
 - 首节点发送 RSVP-TE Path 消息沿着已经确定的路径到达各叶子节点，然后各个叶子节点向首节点发送 RSVP-TE RESV 消息。在 RSVP-TE 信令的接收和发送过程中预留带宽。
 - 首节点收到叶子返回的 RSVP-TE RESV 信令消息，CR-LSP 建立成功。
- 此时，P2MP RSVP-TE 隧道建立成功。

图 3-6 P2MP RSVP-TE 隧道建立示意图



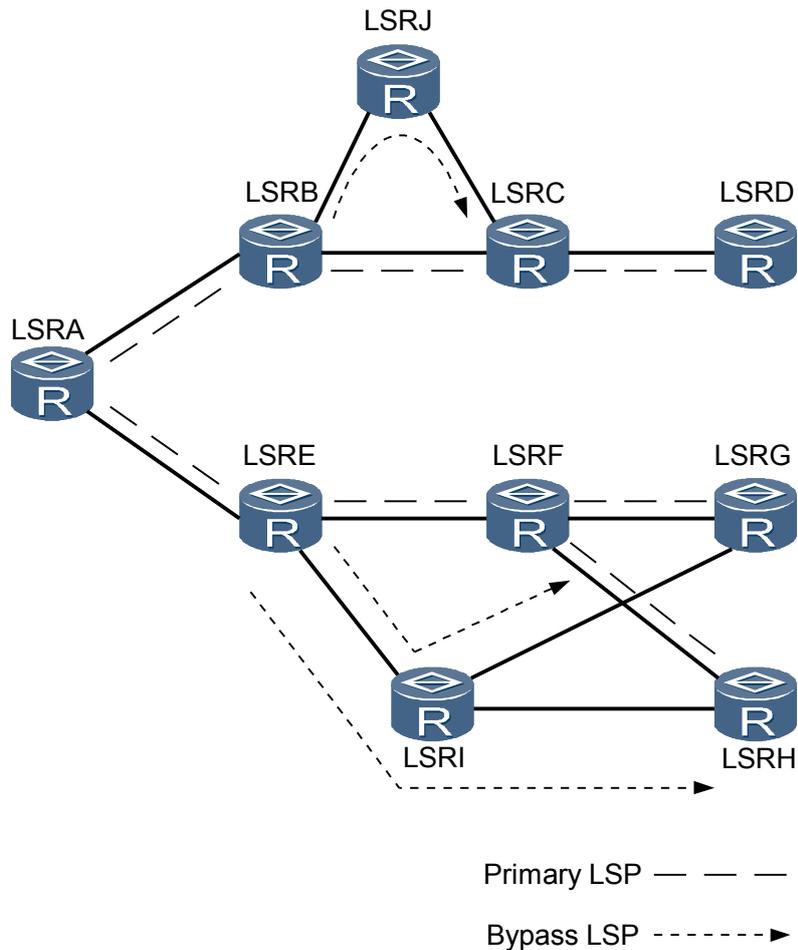
如图 3-6 所示，P2MP RSVP-TE 隧道的建立过程如下：

1. 用户使用 RSVP-TE 建立从 LSRA 到叶子 LSRD、LSRG、LSRH 的 P2MP 隧道。其路径为：LSRA->LSRB->LSRH->LSRD、LSRA->LSRB->LSRH->LSRG 和 LSRA->LSRE->LSRF->LSRG。用户可以提前规划好显式路径，也可以在 LSRA 上通过动态计算得到。
2. LSRA 分别为 LSRD、LSRG、LSRH 构造各自的 RSVP-TE PATH 消息，沿着已经确定的路径向下游发送。在叶子节点，收到 RSVP-TE PATH 消息后沿着反方向回复 RSVP-TE RESV 消息，RSVP-TE RESV 消息中包含标签。
3. LSRA 收到 LSRD、LSRG 或 LSRH 返回的 RESV 消息。
此时，P2MP RSVP-TE 隧道建立成功。

P2MP RSVP-TE 隧道 FRR 保护

与 RSVP-TE P2P 隧道类似，P2MP RSVP-TE 隧道提供快速的 FRR 保护措施。当链路或节点故障时，可以保证流量在 50ms 内切换到备份隧道。

图 3-7 P2MP RSVP-TE 隧道 FRR 保护示意图



如图 3-7 所示，P2MP RSVP-TE 隧道支持链路保护和节点保护。

- 链路保护：在 LSRB 节点，LSRB->LSRC 链路被路径为 LSRB-LSRJ-LSRC 的 P2PLSP 保护。当 LSRB->LSRC 链路故障时，流量切换到 Bypass LSP，即 LSRB-LSRJ-LSRC 链路。
- 节点保护：在 LSRE 节点，为了保护 LSRF 节点和 LSRE->LSRF 之间的链路，可以采用以下两种方式。
 - 使用两条 P2P LSP 进行保护，路径分别为 LSRE->LSRI->LSRG 和 LSRE->LSRI->LSRH。
 - 使用一条 P2MP LSP 进行保护，路径为 LSRE->LSRI->LSRG 和 LSRE->LSRI->LSRH。

当 LSRF 出现故障时，流量被切换到两条 Bypass LSP。即分别通过 LSRE->LSRI->LSRG 和 LSRE->LSRI->LSRH 链路到达 LSRG 和 LSRH。

目前 NE20E-X6 仅支持链路保护。

3.3.4 自动带宽调整

自动带宽调整，即可以动态检测流量，并根据检测结果按照实际的带宽需求重新建立隧道。

一条隧道的带宽可以通过手工配置，但是通常对于一条隧道来说，流量并不是固定不变的。为了保证流量的不间断，可以在手工配置带宽时配置最大流量所需的带宽作为预留带宽。这样虽然能够满足流量要求，但是却造成了带宽资源的浪费。在通信网络中，带宽资源是宝贵的。为了节省带宽，而产生了 TE 自动带宽调整。

通过定时的采样，可以获得通过该隧道在一个采样周期的平均带宽。通过在一段时间的多次采样后，根据平均的采样值来计算新带宽，并按照该带宽新建一条隧道。建立成功后，将流量切换到新的 LSP 上，同时拆除原来的 LSP；如果建立不成功，则流量继续从原来的 LSP 通过，在下一个采样周期结束后再做下一次调节。

此外，用户可以配置带宽的最小值和最大值，所调节的带宽必须在这个范围之内。

说明

- 缺省情况下，隧道不使用自动带宽调整。
- 带宽为多 CT (Class Type: 带宽类型) 的情形，目前不支持自动带宽调整。
- 如果使用自动带宽调整，则不能同时使用 CR-LSP 重优化。

自动带宽调整就是在一定周期内定时触发调整隧道的带宽。

采样周期对所有 MPLS TE 隧道有效。MPLS TE 隧道的出口速率每经过一个采样周期就会被记录下来，这样可以获得一条 MPLS TE 隧道在一个采样周期内的实际平均带宽。

隧道启用自动带宽调整后，当调整频率定时器 (mpls te auto-bandwidth adjustment frequency) 超时时，系统将将该定时器周期内多次采样的最大值作为带宽约束，发起建立一条新 LSP 的请求。建立成功后，将流量切换到新的 LSP 上，同时拆除原来的 LSP；如果建立不成功，则流量继续走原来的 LSP。

在配置自动带宽调整时，我们可以配置 adjustment 模式和 collect-bw 模式。

- adjustment 模式：系统会自动收集流量信息，并自动调整隧道带宽。
- collect-bw 模式：系统只会自动收集流量信息，而不会自动调整隧道带宽。

3.3.5 重优化

重优化是重新计算隧道需要的路径，完成对隧道路径的优化。

流量工程一个主要的目标就是优化网络上流量的分布。隧道建立之后，可以根据网络上的带宽变化、流量变化、管理策略变化等对已经建立的 CR-LSP 进行优化。当某条隧道发现更优的路径时，可以对其进行优化。所谓更优的路径，即满足 metric 值更小的路径。

隧道重优化有两种方式：

- 自动重优化：根据自动重优化机制，用户可以指定路径优化的时间间隔。在时间到达时，CR-LSP 会主动发现更优的路径来进行优化。
- 手工重优化：除了自动重优化，用户也可以通过配置方式直接触发 CR-LSP 进行优化，重新请求计算更优路径来建立 CR-LSP。

在优化时，用户的业务流不中断是非常重要的。即新的 CR-LSP 必须先建立，业务在旧的 CR-LSP 被拆除前切换到新的 CR-LSP 上。在新旧 CR-LSP 共享的链路上，由于旧的 CR-LSP 使用的资源不能在新的 CR-LSP 建立前释放，共享链路上资源是否被计算两次非常关键。这是因为可能会由于资源的缺乏而造成新的 CR-LSP 无法建立。

RSVP-TE 信令的 SE 预留风格能够非常好的解决这个问题，SE 预留风格允许新旧 CR-LSP 共享资源，使新的 CR-LSP 不会因为链路资源缺乏而必须等到旧的 CR-LSP 拆除才能完成。

 说明

- 缺省情况下，不进行重优化。定时重优化的缺省时间间隔是 3600 秒。
- 如果使用自动重优化，则不能同时使用自动带宽调整。
- 如果隧道的资源预留为固定过滤风格 FF，则不能配置 CR-LSP 重优化。

隧道重优化的实现如下：

- 对于自动重优化，是根据用户在隧道下配置的时间，重新触发 CSPF 计算该隧道的路径。如果 CSPF 计算出来的路径比现有路径的 Metric 值更小，则以新的路径创建 CR-LSP。若建立成功则通知转发层面进行流量切换，删除原 CR-LSP，重优化完成。若建立不成功，则流量还按照原路径转发。
- 对于手动重优化，是在用户视图下输入重优化命令，触发进行重优化。

3.3.6 TE FRR

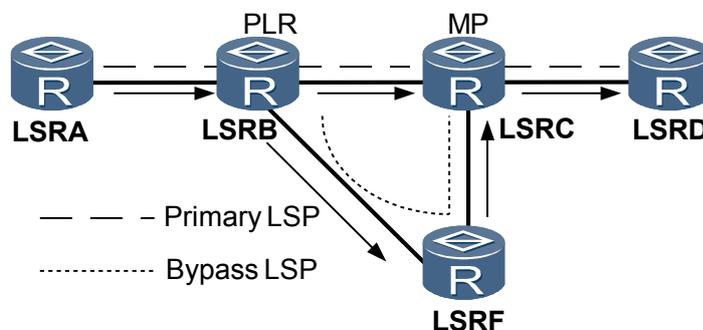
TE FRR 是 MPLS TE 中的一种局部性保护机制，用于保护 CR-LSP 的链路和节点故障，而且可以对主 LSP 配置带宽保护或非带宽保护。

TE FRR 通过预先建立绕过故障的链路或者节点的 Bypass 隧道达到保护主 CR-LSP 的目的。当 CR-LSP 链路或节点故障时，允许流量继续从旁路隧道传输，同时头节点可以在数据传输不受影响的同时继续发起主路径的重建。

FRR 中涉及的概念：

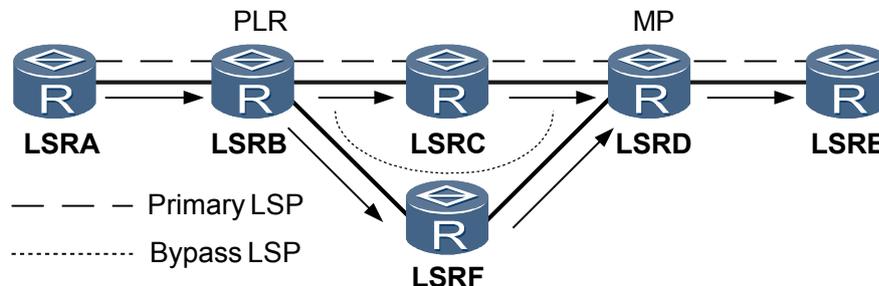
- 主 CR-LSP：被保护的 CR-LSP。
- Bypass CR-LSP：保护主 CR-LSP 的 CR-LSP。Bypass CR-LSP 的入节点是 PLR，出节点是 MP。Bypass CR-LSP 一般处于空闲状态，不承载业务。如果需要使用 Bypass CR-LSP 保护主 CR-LSP 的同时承载业务数据的转发，需要为 Bypass CR-LSP 分配足够的带宽。
- PLR（Point of Local Repair）：本地修复节点。Bypass CR-LSP 的入节点，必须在主 CR-LSP 的路径上，可以是主 CR-LSP 的入节点，但不能是主 CR-LSP 的出节点。
- MP（Merge Point）：汇聚点。Bypass CR-LSP 的出节点，必须在主 CR-LSP 的路径上，并且不能是主 CR-LSP 的入节点。
- 链路保护：如 [图 3-8](#)，PLR 和 MP 之间有直连链路（LSRB→LSRC）连接，主 CR-LSP 经过这条链路。当这条链路失效时，流量可以切换到 Bypass CR-LSP（LSRB→LSRF→LSRC）上。

图 3-8 链路保护示意图



- 节点保护：如图 3-9，PLR 和 MP 之间存在一台 LSR（LSRB→LSRC→LSRD），主 CR-LSP 经过该节点（LSRC）。当该节点失效时，流量可以切换到 Bypass CR-LSP（LSRB→LSRF→LSRD）上。

图 3-9 节点保护示意图



- 主 CR-LSP 的建立

主 CR-LSP 的建立过程与普通 CR-LSP 的建立过程一致，不同之处主要是：建立主 CR-LSP 过程中，隧道入节点会在 Path 消息中添加局部保护标记、记录路由标记和 SE 风格标记。如果需要进行带宽保护，则在该对象中添加带宽保护标记。

- 绑定旁路隧道

为主 CR-LSP 寻找合适的 Bypass CR-LSP 的过程称为绑定。只有具有局部保护标记的主 CR-LSP 才会触发绑定策略，绑定是在隧道切换之前完成的。绑定过程中，节点需要计算出 Bypass CR-LSP 的出接口、NHLFE、MP 的 LSR ID、MP 分配的标签及保护的类型等信息。

对于入节点和中间节点而言，主 LSP 的下一跳（NHOP）或下一跳（NNHOP）是已知的。如果 Bypass LSP 的 Egress LSR ID 与 NHOP 的 LSR ID 相等，就可以形成链路保护；如果 Bypass CR-LSP 的 Egress LSR ID 与 NNHOP 的 LSR ID 相等，就可形成节点保护。

当有多条 Bypass CR-LSP 可以保护同一条 CR-LSP 时，首先选择满足用户带宽配置需求的 Bypass CR-LSP，如果同时满足用户带宽配置需求，则节点保护优于链路保护，手工保护优于自动保护。

如果旁路隧道绑定成功，主 CR-LSP 的下一跳标签转发项 NHLFE（Next Hop Label Forwarding Entry）表项中记录 Bypass CR-LSP 的 NHLFE 表项索引以及 MP 为上一个节点分配的标签，即内层标签。内层标签用于 FRR 切换时的流量转发。

- 故障检测

链路保护直接使用链路层协议实现故障检测和通告，链路层发现故障的速度与链路类型直接相关。节点保护则使用链路协议检测链路故障，在链路没有故障的情况下，可以通过配置 BFD 机制检测被保护节点的故障。

对于节点保护，只保护被保护节点及其与 PLR 之间的链路。对于被保护节点和 MP 之间的链路故障，PLR 无法感知。

无论是检测到链路故障还是节点故障，最终都会导致 PLR 上的出接口被置为老化状态。出接口老化就会触发 FRR 的流量切换。

- 切换

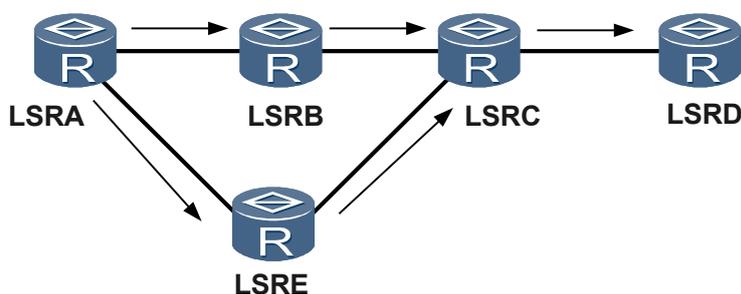
切换是指主 CR-LSP 故障后，业务流量和 RSVP 消息从主 CR-LSP 切换到 Bypass CR-LSP 上，并向上游通告切换已经发生。在切换的时候，主 CR-LSP 的 NHLFE 表项会被置上切换标志。

- 回切

切换后，PLR 会向上游发送携带切换标记的 PathError 消息。Ingress 收到该消息，试图重建主 CR-LSP，但不拆除主 CR-LSP，而是借用 Bypass CR-LSP 作为主 CR-LSP 继续转发数据，直到主 CR-LSP 重建成功。尝试重建的 CR-LSP 称为 Modified CR-LSP。

当主 CR-LSP 重建成功后，业务流量和 RSVP 消息需要从 Bypass CR-LSP 回切到主 CR-LSP 上。此过程，TE FRR（包括 Auto FRR）采用 Make-before-break 机制，即只有 Modified CR-LSP 建立成功后，原来的 Primary CR-LSP 才能被删除。

图 3-10 TE FRR 示意图



例如图 3-10，TE FRR 的主路径为 LSRA→LSRB→LSRC→LSRD；旁路隧道的路径为 LSRA→LSRE→LSRC；LSRB 为被保护节点。

当主路径故障时，借用 Bypass CR-LSP 作为主 CR-LSP 继续转发数据 LSRA→LSRE→LSRC→LSRD，并尝试建立 Modified CR-LSP。Modified CR-LSP 使用原路径建立成功后，流量转到 Modified CR-LSP 上，重新切回主路径 LSRA→LSRB→LSRC→LSRD，原来的主 CR-LSP 拆除。

Auto TE FRR

TE FRR 要求在配置被保护隧道的时候，需要手工配置一条旁路隧道来与之绑定，当出现链路或节点故障时，可以将数据切换到旁路隧道。

这种 FRR 保护需要手工配置旁路隧道，如果没有配置，或者旁路隧道配置不符合要求，将无法对被保护隧道进行保护。为了简化配置并保持 FRR 特性，Auto TE FRR 在这样的情形下应运而生。

在使能了 Auto TE FRR 的节点上建立一条要求 FRR 保护的主隧道，并且可以为主隧道配置带宽保护或非带宽保护，系统将自动尝试建立一条符合条件并最优的 Auto Bypass CR-LSP。在建立成功后，所建立的 Auto Bypass CR-LSP 也能对主 CR-LSP 进行保护。

Auto TE FRR 实现原理与手工配置的 TE FRR 一样，区别仅在于手工 TE FRR 需要手工配置 Bypass Tunnel，而 Auto FRR 的 Bypass Tunnel 是由系统根据策略自动计算生成的，此处不再赘述。

TE FRR 几种保护的优先顺序

FRR 可分为：

- 带宽与非带宽保护
- 节点保护与链路保护
- 手工保护与自动（Auto）保护

其中，带宽保护与非带宽保护是根据用户配置，没有优先级之分。如果需要选择，则首先选择满足用户带宽配置需求的 Bypass CR-LSP，如果同时满足用户带宽配置需求，则节点保护优于链路保护，手工保护优于自动保护。

拔板保护

拔板保护是指当 PLR 上的主隧道出接口所在接口板被拔出时，将 MPLS TE 流量快速切换到旁路隧道上。当被拔出的接口板再次被插入时，如果主隧道的出接口仍处于可用状态，可进行 MPLS TE 流量的回切。拔板保护主要是用来保护 PLR 节点上的主隧道出接口。

一般情况下，如果配置了隧道接口的接口板被拔出，隧道信息就丢失了。因此要实现 TE FRR 拔板保护，PLR 的主隧道接口、旁路隧道接口及旁路隧道的出接口都不能在被拔出的接口板上。一般情况，建议将 PLR 上的 TE tunnel 接口配置在主控板上。

将 PLR 上的 TE tunnel 接口配置在主控板上后，当主隧道物理出接口所在接口板被拔出或故障时，该接口就被置为 Stale 状态，经过该接口的有 FRR 保护的主 CR-LSP 保持不删除。当接口板重新插入时，恢复接口为可用状态，并开始重建主 CR-LSP。

部署 TE FRR

TE FRR 是 MPLS TE 中的一种局部性保护机制。在配置 Bypass CR-LSP 时，应该规划好它所保护的链路或节点，并确保该 Bypass CR-LSP 不会经过它所保护的链路或节点，否则不能真正起到保护作用。

FRR 不支持多点故障。即，如果发生了 FRR 切换，数据从主 CR-LSP 切换到 Bypass CR-LSP，在数据通过 Bypass CR-LSP 转发期间，Bypass CR-LSP 的状态必须始终保持 UP。一旦 Bypass CR-LSP 在此期间出现故障，被保护的数据将不能通过 MPLS 转发，从而可能出现流量中断，FRR 功能失效。即使 Bypass CR-LSP 重新建立，流量也无法转发，只能等待主 CR-LSP 恢复或重新创建后，流量才恢复转发。

TE FRR 支持 N:1 的保护模式，即一条 Bypass 隧道可以保护多条主隧道。

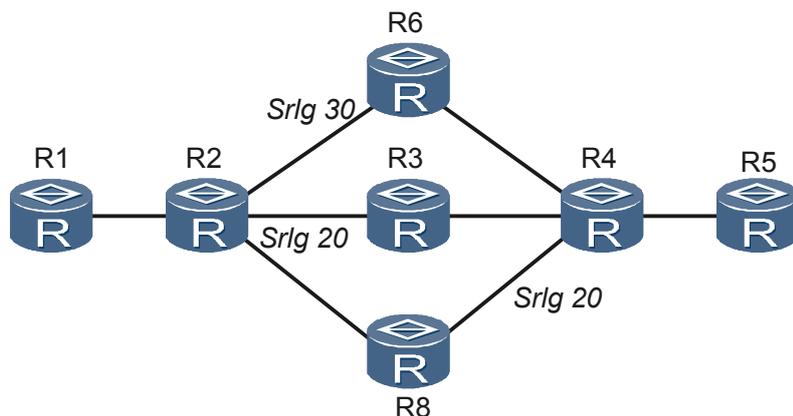
另外，FRR 的 Bypass 隧道需要预先建立，这需要占用额外的带宽。在网络带宽余量不多的情况下，只能对关键的链路或节点进行 FRR 保护。

3.3.7 SRLG

SRLG（Shared Risk Link Group，共享风险链路组）是具有相同故障风险的一组链路集合。即如果其中一条链路失效，那么组内的其他链路也可能失效，这种情况下使用组内的其他链路作为失效链路的热备份 CR-LSP 或 FRR Bypass 隧道，将起不到保护的作用。

SRLG 是链路属性，用数值表示，数值相同的链路表示属于同一个共享风险链路组。部署 SRLG 可以约束热备份 CR-LSP 和 FRR Bypass 隧道路径的选择，避免了主、备隧道建立在具有相同故障风险的链路上。

图 3-11 SRLG 示意图



如图 3-11 所示，以 TE FRR 为例，主隧道路径为 R1-R2-R3-R4-R5，其中链路 R2-R3-R4 需要 FRR Bypass 隧道保护。而在 R2 上满足要求的 Bypass 隧道共有两条，路径分别是 R2-R6-R4 和 R2-R8-R4。

在部署的时候发现，当链路 R2-R3 发生故障时，链路 R8-R4 通常也会发生故障，即两条链路存在相同的故障风险。如果选择 R2-R8-R4 作为 Bypass 隧道，当链路 R2-R3 出现故障时，链路 R8-R4 也可能出现故障，这样就失去了保护的意义。

因此在配置隧道之前，建议先收集并配置链路的 SRLG 信息。将 R2-R3 和 R8-R4 之间的链路配置相同的 SRLG 值，即两条链路处于相同的共享风险链路组，这样就可以保证主隧道优先选择与其不在相同共享风险链路组的 Bypass 隧道。

实现原理

SRLG 是通过 IGP 协议对 TE 属性的扩展来发布的。根据 CSPF 计算路径时，不仅会考虑带宽等约束条件，同时将 SRLG 作为约束条件来计算。

MPLS TE SRLG 具有两种操作模式：

- 严格 SRLG 模式（strict mode）：对热备份 LSP 和 FRR Bypass 隧道路径的计算，必须考虑 SRLG 的约束条件；
- 首选 SRLG 模式（preferred mode）：对热备份 LSP 和 FRR Bypass 隧道计算路径时候可以不考虑 SRLG 的约束条件。例如，当 Tunnel 配置了 hot-standby 时，如果第一次 CSPF 考虑 SRLG 约束计算备份路径失败，那么第二次进行算路的时候将不会考虑 SRLG 的约束条件。

SRLG 的操作模式及接口的 SRLG 信息都是在本地配置的。但是链路的 SRLG 信息可以通过 IGP 路由通告给整个域，即相同域中的节点能够获取域内所有链路的 SRLG 信息。CSPF 在计算路径的时候，根据当前节点配置的操作模式及 SRLG 约束信息进行路径计算。

3.3.8 CR-LSP 备份

同一条隧道下对主 CR-LSP 进行路径备份的 CR-LSP 称为备份 CR-LSP。

备份 CR-LSP 用于实现对重要 LSP 的流量保护。作为流量保护的一个重要组成部分，在主 CR-LSP 失败后，流量需要及时被切换到备份隧道上。

当入节点感知到主 CR-LSP 不可用时，将流量切换到备份路径上，当主 CR-LSP 路径恢复后再将流量切换回来，以实现主 CR-LSP 路径的备份保护。

CR-LSP 的备份有热备份和普通备份两种方式。

- 热备份 (Hot-standby)：创建主 CR-LSP 后随即创建备份 CR-LSP。主 CR-LSP 失效时，将业务切换至备份 CR-LSP。当主 CR-LSP 恢复时，将业务切回到主 CR-LSP。热备份 CR-LSP 可以支持逃生路径。
- 普通备份：指主 CR-LSP 失效后创建备份 CR-LSP。主 CR-LSP 失效时，将业务切换至备份 CR-LSP。当主 CR-LSP 恢复时，将业务切回到主 CR-LSP。
- 热备份和普通备份方法的区别请参见表 3-1。

表 3-1 热备份与普通备份的区别

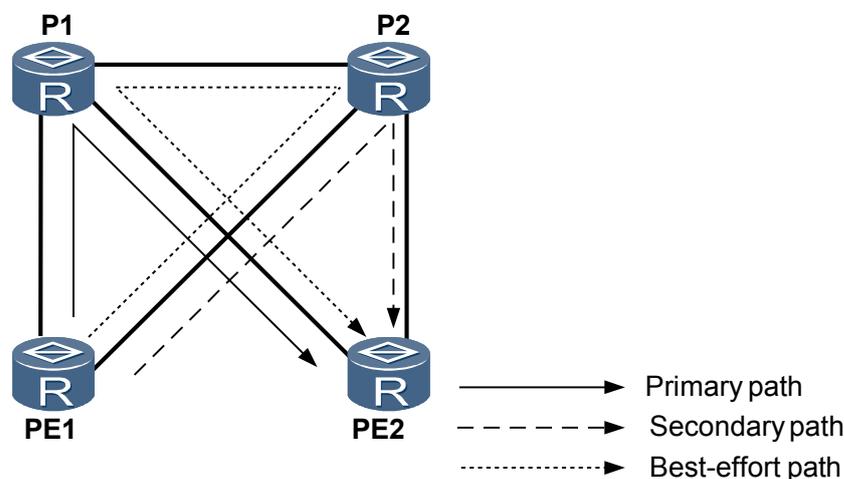
比较内容	热备份	普通备份
备份 LSP 创建时间	创建主 CR-LSP 成功后即创建备份 CR-LSP	主 CR-LSP 失效后才创建备份 CR-LSP
主、备显式路径	可以通过配置指明是否允许主备路径重合。当备份 CR-LSP 使能显式路径建立时，以显式路径作为约束条件建立备份 CR-LSP	无论备份 CR-LSP 是否用显式路径建立，备份 CR-LSP 的创建路径都可以与主 CR-LSP 的路径部分或完全重合
是否可以与逃生路径同时配置	是	使用属性模板配置的普通备份 CR-LSP 可以与逃生路径同时配置

说明

备份 CR-LSP 不能同时配置自动带宽调整与重优化功能。

- 逃生路径
热备份方式支持逃生路径。在主、备 CR-LSP 都故障时，可触发建立一条临时的 CR-LSP，称为逃生路径，流量将切换到逃生路径上。
如图 3-12，主 CR-LSP 路径为 PE1→P1→PE2；备份 CR-LSP 路径为 PE1→P2→PE2。当主备 CR-LSP 都故障时，PE1 触发建立逃生路径 PE1→P2→P1→PE2。

图 3-12 逃生路径示意图





说明

逃生路径没有带宽保证，可以根据需要配置逃生路径的亲属性和跳数限制。

Make-Before-Break、删除延迟与切换延迟

CR-LSP 备份链路支持 Make-Before-Break、删除延迟与切换延迟。当用户修改了备份 CR-LSP 的 Modify 属性时，Tunnel 会触发系统重建备份 CR-LSP。备份 CR-LSP 重建的过程采用了 Make-before-break 机制。当携带新属性的 CR-LSP 完全建立成功以后，如果此时备份 CR-LSP 已经承载了流量，Tunnel 会将流量切换到新的 CR-LSP 上，然后删除原来的备份 CR-LSP。

流量回切分为如下三种情况：

- 流量从逃生路径回切到主 CR-LSP 或备份 CR-LSP。其回切的优先级顺序是主 CR-LSP 有最高优先级，其次是热备份 CR-LSP，普通备份 CR-LSP 优先级最低。
- 当流量承载在普通备份 CR-LSP 上时，如果热备份 CR-LSP 先于主 CR-LSP 建立成功，由于热备份 CR-LSP 优先级比普通备份 CR-LSP 优先级高，流量会从普通备份 CR-LSP 切换到热备份 CR-LSP 上。
- 在备份 CR-LSP 承载流量的期间，如果主 CR-LSP 重新可以承载流量，则 Tunnel 会触发流量从备份 CR-LSP 回切到主 CR-LSP。

流量的切换过程可以使用删除延迟与切换延迟保证流量不丢失。以流量从普通备份 CR-LSP 回切到主 CR-LSP 为例，需要经过如下 4 个步骤：

1. 主 CR-LSP 的状态重新变成 Up 后，系统将设置一个缓冲时间，为流量回切做好准备。在缓冲时间内流量将不从普通备份 CR-LSP 回切到主 CR-LSP。
2. 在缓冲时间内，主 CR-LSP 链路在各个节点创建新 CR-LSP 的相关表项。
3. 缓冲时间结束后，流量正式回切。
4. 回切成功后，头结点将再次设置一个缓冲时间，为删除普通备份 CR-LSP 作准备。当缓冲时间超时后，Tunnel 将正式删除普通备份 CR-LSP。

备份 CR-LSP 轮流创建

同一条隧道下可能存在多个建立隧道的选项。在创建 CR-LSP 时，为了使隧道能够尽可能快速的承载业务，系统会轮流创建各个选项，直到隧道建立成功。

CR-LSP 的轮流创建规则如下：

1. 当新提交一条隧道或者隧道状态变为 Down 时，系统将按一定的优先级顺序轮流尝试创建主 CR-LSP、热备份 CR-LSP、普通备份 CR-LSP、逃生路径，直到隧道建立成功。
2. 在创建热备份或者普通备份类型的 CR-LSP 时，用户可以在这两种 CR-LSP 下面最多配置三个 CR-LSP 配置模板，这些配置模板之间存在一定的优先级别，系统会按照优先级别从高到底轮流尝试创建各个配置模板，直到该 CR-LSP 建立成功。
3. 如果使用的是低优先级的 CR-LSP 配置模板创建隧道成功，当链路状况发生改变时，系统将被触发去尝试向高优先级的 CR-LSP 配置模板升级。在升级过程中，系统将以 Make-before-break 机制创建新的 CR-LSP 保证流量不会中断。

备份 CR-LSP 属性模板锁定功能

在前面的备份 CR-LSP 轮流创建中提到，系统可以为热备份或者普通备份类型的 CR-LSP 提供三个属性模板，且各个属性模板之间存在优先级。如果当前的备份 CR-LSP 的是由

低优先级的属性模板创建的，那么系统会不断去尝试使用更高优先级的属性模板创建新的 LSP。

如果备份 CR-LSP 下的三个属性模板都可以创建出稳定性较好的 CR-LSP，则用户可以通过配置备份 CR-LSP 属性模板锁定功能来取消模板的自动升级。当配置了流量锁定功能以后，即使当前的备份 CR-LSP 使用的是一个低优先级属性模板，系统也不会去尝试更高优先级属性模板。这样可以减少很多不必要的流量切换，降低系统开销。

当用户去使能备份链路流量锁定功能以后，系统将自动恢复升级功能。

CR-LSP 热备份 dynamic-bandwidth 保护机制

热备份 CR-LSP 可以配置 dynamic-bandwidth 保护，即动态带宽保护功能。在该功能下，在主 CR-LSP 出现故障之前，热备份 CR-LSP 并不会额外占用网络中的带宽资源，只有当热备份 CR-LSP 真正承载流量后才会占用网络的带宽资源。这样可以更大幅度的节省网络资源，缩减网络开销。

使能动态带宽保护功能后，主 CR-LSP 承载流量，而热备份 CR-LSP 将不占用链路带宽。其具体过程为：

1. 当主 CR-LSP 出现故障后，流量立即切换到带宽值为 0 的热备份 CR-LSP。同时在 Ingress 节点，Tunnel 立即触发热备份 CR-LSP 采用 Make-Before-Break 机制进行重建。
2. 当新的热备份 CR-LSP 建立成功后，流量就切换到新的 LSP 上，并将带宽值为 0 的热备份 CR-LSP 删除。
3. 当主 CR-LSP 故障恢复后，流量会重新回切到主 CR-LSP。此时热备份 CR-LSP 会释放已占用的带宽，重新采用 0 带宽建立热备份 CR-LSP。

与其他特性的比较

1. CR-LSP 备份与 TE FRR 的区别

- CR-LSP 备份是一种端到端的路径保护（Path Protection, End-to-End Protection），提供整条 LSP 的保护。
- FRR 则是一种局部保护措施，只能保护 LSP 中的某条链路和某个节点，并且，FRR 是一种快速响应的临时性保护措施，对于切换时间有严格要求。

2. 隧道备份常常和快速重路由联合部署

在普通的联合部署网络中：

- CR-LSP 热备份与 TE FRR 结合使用：快速重路由可以及时响应链路故障，将流量在最短的时间内切换到旁路隧道上。链路故障信息通过信令带到入口节点，然后将流量切换到备份隧道上。
- CR-LSP 普通备份与 TE FRR 结合使用：快速重路由可以及时响应链路故障，将流量在最短的时间内切换到旁路隧道上。当 TE FRR 的主隧道和旁路隧道都出现故障之后，才会建立备份隧道并将流量切换到备份隧道上。

说明

如果主隧道使能了旁路保护隧道与备份 CR-LSP 同步功能，当主 CR-LSP 出现故障时，流量会切换到 TE FRR 的旁路保护隧道并尝试恢复主 CR-LSP 的同时，也会尝试创建备份 CR-LSP。当备份 CR-LSP 创建成功，并且主 CR-LSP 未恢复时，流量会切换到备份 CR-LSP。

3. CR-LSP 备份与 TE FRR 的同步联用

当用户同时部署 TE FRR 的局部保护和备份 CR-LSP 的端到端保护之后，由于后者的可靠性高于前者，因此为了增加隧道的安全性，系统支持单独配置增强的旁路保

护隧道与备份 CR-LSP 同步联用的功能。当使能了 CR-LSP 备份同 TE-FRR 同步联用之后：

- 如果配置的是普通备份：

当被保护链路或节点出现故障时，流量切换到 TE FRR 的旁路保护隧道并尝试恢复主 CR-LSP 的，同时也会轮流尝试创建备份 CR-LSP。

当备份 CR-LSP 创建成功，并且主 CR-LSP 未恢复时，流量会切换到备份 CR-LSP。

当主 CR-LSP 恢复成功，则不管当前流量是处于旁路保护隧道还是备份 CR-LSP，流量都会切换到新的主 CR-LSP 上。

当备份 CR-LSP 创建不成功，且主 CR-LSP 恢复也不成功，流量仍然从旁路保护隧道通过。

- 如果配置的是热备份：

如果备份 CR-LSP 的状态为 Up，且被保护链路或节点出现故障时，则流量先切换到旁路保护隧道后，立即切换到备份 CR-LSP 上，同时尝试恢复主 CR-LSP。

如果备份 CR-LSP 的状态为 Down，则处理方式与普通备份相同。

当主 CR-LSP 处于 Up 状态的同时，热备份 CR-LSP 也会一直尝试创建，当创建成功，就会占用额外的带宽。而普通备份 CR-LSP 在主 CR-LSP 处于 Frr-in-use 状态时才开始创建，在主 CR-LSP 没有出现故障时，不需要占用额外的带宽资源。因此推荐选择普通备份的方式同 TE FRR 同步联用。

4. CR-LSP 备份与隧道保护组的区别

CR-LSP 备份和隧道保护组都是 MPLS TE 中的端到端的保护机制。这两种保护机制的主要区别请参见表 3-2。

表 3-2 CR-LSP 备份与隧道保护组的区别

比较内容	CR-LSP 备份	隧道保护组
保护的對象	在同一条隧道中创建主、备两条 CR-LSP。备份 CR-LSP 保护主 CR-LSP	保护组是用一个隧道保护另一个隧道
TE FRR	只有主 CR-LSP 支持用 TE FRR 保护，备份 CR-LSP 不支持用 TE FRR 保护	主隧道支持用 TE FRR 保护，备份隧道不支持用 TE FRR 保护
LSP 属性	主、备 CR-LSP 的属性（例如带宽，建立优先级，保持优先级）一致，除了 TE FRR 属性	保护组内的各隧道的属性没有联系。例如，带宽为 0 的备份隧道可以保护有带宽保护要求的主隧道
保护模式	支持 1:1 保护模式，即每条主 CR-LSP 都有自己的备份 CR-LSP	支持 N:1 保护模式，即将一条隧道作为多条主用隧道的备份隧道，当任何一条主用隧道故障时，都将数据倒换到共享的备份隧道上

3.3.9 DS-TE

定义

MPLS DS-TE 是将 MPLS TE 和差分服务 (Diff-Serv) 相结合的一种提供有效的 QoS 保证的技术。

MPLS DS-TE 通过 MPLS TE 建立标签交换路径, 精确控制流量流经的路径, 保证带宽资源, 同时提供多种保护机制来提高业务可靠性; 通过 Diff-Serv 模型对业务进行服务等级 (CoS) 的分类, 根据 CoS 细粒度来预留资源, 设置控制策略和保障机制, 提供有效的 QoS 保证。

目的

随着网络技术不断发展, 一方面基于包交换的 IP 网络逐渐取代传统的基于电路交换的 ATM 及 FR 网络, 另一方面网络运营商提供的业务类型由单纯的数据或者语音业务演变成视频、语音和数据等业务共存的局面, 随着商务行为与网络共同的日益结合, 网络运营商承载的业务量和业务类型都在不断增长, 客户从自身的利益考虑, 对运营商能否提供服务质量 (QoS, Quality of Service) 保证变得日益重视。

Diff-Serv 模型, 对将客户的业务划分到不同的服务类型, 然后针对不同的服务类型对流量进行转发, 是一种可以满足多种 QoS 需求的服务模型。Diff-Serv 模型可以根据业务的不同服务等级, 有差别地进行流量的控制和转发; 将多种业务的数据流映射到有限的几个类型, 所需维护的信息量正比于数据流的类型, 而不是数据流的数量, 因此 Diff-Serv 可扩展性较好。

但是 Diff-Serv 模型只能在单个节点上预留资源, 无法在整个路径上保证服务质量。

MPLS TE 技术利用可用资源沿链路建立标签交换路径 (LSP), 从而确保始终为特定数据流提供有保证的带宽, 以避免在稳定或故障情况下出现拥塞。由于 LSP 只是在资源可用的条件下建立, 提高了现有设备的带宽利用率。此外, MPLS TE 可以精确控制流量流经的路径, 从而使现有的带宽资源被充分利用。

MPLS TE 的不足之处在于: 无法基于每个流量类型提供有区别的 QoS 保证。例如, 网络中有两类流量, 分别为语音数据流和视频数据流。视频数据帧可能会出现发生长时间的重复, 因此视频数据可能要求比语音数据具有更高的丢弃优先级。MPLS TE 并不区分流量类型, 而是将语音和视频流量都聚合在同一个丢弃优先级上。

针对以上两种技术的优缺点, MPLS DS-TE 两种技术结合, TE 技术可以保证带宽资源, 通过链路保护和快速重新路由等机制实现故障发生时的快速恢复, 提高业务的可靠性, 而通过 Diff-Serv 模型对业务进行服务等级 (CoS) 的分类, 根据 CoS 细粒度来预留资源, 每个 CoS 级别提供 MPLS 容错机制, 设置速率限制和访问许可控制等机制, 以确保将流量控制在资源预留时设置的限制范围内, 提供 QoS 保证。

受益

首先, DS-TE 通过利用可用资源沿链路建立标签交换路径 (LSP), 从而确保始终为特定流提供有保证带宽, 以避免在稳定或故障情况下出现拥塞。由于 LSP 只是在资源可用的条件下建立, 因此不需要轻载, 提高了现有设备的带宽利用率。

其次, DS-TE 还可以通过链路保护和快速重新路由等机制实现故障发生时的快速恢复, 提高业务的可靠性。

最后, DS-TE 通过对业务进行服务等级 (CoS) 的分类, 根据 CoS 细粒度来预留资源, 并在每个 CoS 级别提供 MPLS 容错机制, 设置速率限制和访问许可控制等机制, 以确

保将流量控制在资源预留时设置的限制范围内，提供 QoS 保证以满足严格的 SLA，如话音、ATM 和帧中继等。

MPLS DS-TE 的基本概念

- DS 字段
为了实现 Diff-Serv，RFC2474 中对 IPv4 报文头的 ToS 字段进行了重新定义，称为 DS (Differentiated Services) 字段。DS 字段的高 2 位是预留位，低 6 位是区分服务代码点 DSCP (DS CodePoint)。
- PHB
PHB (Per Hop Behavior) 用来描述拥有相同 DSCP 值的报文的下一步转发动作。一般 PHB 包括时延、丢包率等流量特性。
目前，IETF 定义了三种标准的 PHB：加速转发 EF (Expedited Forwarding)、确保转发 AF (Assured Forwarding) 和尽力而为 BE (Best-Effort)，BE 是缺省的 PHB。
- CT
为了提供差分服务，DS-TE 将 LSP 的带宽划分为 1 ~ 8 个部分，每部分被赋予不同的服务等级。这样一条或一组 LSP 的相同服务等级的带宽集合称为一个 CT(Class Type: 服务类型)。一个 CT 只能承载一种业务类型的流量。
IETF 规定 DS-TE 最多支持 8 个 CT，可以记为 CT_i，其中 i 的取值范围是 0 ~ 7。
- 单 CT 和多 CT
单 CT 是指一条 LSP 只允许承载一个 CT 的流量。
多 CT 是指一条 LSP 允许同时承载多个 CT 的流量。多 CT 中，必须所有 CT 带宽都满足时，资源预留、LSP 建立或带宽抢占才能成功。
华为实现的 DS-TE 支持这两种类型的 LSP。
- TE-Class
TE-Class 是指 CT 和优先级的组合 <CT, 优先级>。
这里的优先级是指 CR-LSP 的抢占优先级，而不是 MPLS 首部的 EXP 值。抢占优先级的取值范围是 0 ~ 7，数值越小，优先级越高。只有当一条 CR-LSP 的 CT 和建立优先级的组合 <CT, setup-priority> 以及 CT 和保持优先级的组合 <CT, hold-priority> 同时存在于 TE-Class 映射表中，该 CR-LSP 才能建立成功。例如，假设某节点的 TE-Class 映射表中仅有 TE-Class[0]=<CT0, 6> 和 TE-Class[1]=<CT0, 7>，则只有以下三类 CR-LSP 能建立成功：
 - Class-Type = CT0, setup-priority = 6, hold-priority = 6;
 - Class-Type = CT0, setup-priority = 7, hold-priority = 6;
 - Class-Type = CT0, setup-priority = 7, hold-priority = 7。CT 和优先级可以任意组合，因此理论上 TE-Class 共有 64 个。VRP 最多支持 8 个 TE-Class。这 8 个 TE-Class 通过用户配置指定。
- DS-TE
DS-TE 有以下两种模式：
 - IETF 模式：IETF 定义的模式，实现 8 个 CT 和 8 个优先级的组合，共 64 种 TE-Class。NE20E-X6 中，可配置最多 8 个 TE-Class。
 - Non-IETF 模式：非 IETF 定义的模式，实现两个 CT 和 8 个优先级的组合，共 16 种 TE-Class。
- TE-Class 映射表
TE-Class 映射表 (Mapping) 由一组 TE-Class 组成。NE20E-X6 实现中，TE-Class 映射表最多由 8 个 TE-Class 组成。对于有带宽的 LSP 来说，在头结点配置时需要

考虑满足 TE-Class 映射表，其在沿途各节点预留资源时同样需要考虑满足 TE-Class 映射表，否则就不能建立成功，建议 MPLS 网络中所有的 LSR 都配置相同的 TE-Class 映射表。

带宽约束模型

带宽约束模型用来定义 BC (Bandwidth Constraints) 的最大数目和每个 BC 的带宽可被哪些 CT 使用，以及 CT 如何使用 BC 带宽。

MPLS TE 和 Diff-Serv 的结合原理

Diff-Serv 的边界节点将流量分为少量的几个类，并将类信息直接标记在报文中的 DSCP 字段中。内部节点在调度转发报文时根据 DSCP 选择相应的 PHB。

MPLS 首部的 EXP 域可以用于承载 Diff-Serv 相关信息。如何将 DSCP 值（最多有 64 个值）映射到 EXP 域（最多 8 个值）是 DS-TE 实现的关键。RFC3270 定义了如下两种方案：

- L-LSP (Label-Only-Inferred-PSC LSP)：将丢弃优先级放在 EXP 字段中，PHB 调度类型由标签决定，在转发期间，标签决定数据包转发路径并为其分配调度行为。
- E-LSP (EXP-Inferred-PSC LSP)：将调度类型和丢包优先级都由 MPLS 标签的 EXP 字段表示。转发期间，标签决定数据包的转发路径，EXP 决定 PHB。E-LSP 适用于支持不多于 8 个 PHB 的网络。

VRP 实现了 E-LSP。DSCP 到 EXP 的映射遵从 RFC3270 的定义。EXP 到 PHB 的映射则通过手工配置指定。

为了基于流量类型分配资源，DS-TE 引入了 CT 的概念。DS-TE 将具有相同 PHB 的流量映射到同一个 CT 上，并为每个 CT 单独分配资源。因此，DS-TE LSP 是基于 CT 建立的。即 DS-TE 在路径计算过程中，需要将 CT 及每个 CT 可获得的带宽作为约束条件；在进行资源预留过程中，也需要考虑 CT 及其带宽需求。

IGP 和 RSVP 的相关扩展

为支持 DS-TE，RFC4124 对 IGP 进行了扩展，引入了新的可选子 TLV (Bandwidth Constraints sub-TLV)，并重新定义了原有子 TLV (Unreserved bandwidth sub-TLV) 的含义，用于通告和收集链路上各优先级的每个 CT 的可预留带宽，详细描述请参见 RFC4124。

同时 IETF 对 RSVP 进行了扩展：RFC4124 为 Path 消息定义了 CLASSTYPE 对象；IETF 草案 (draft-minei-diffserv-te-multi-class) 定义了 EXTENDED_CLASSTYPE 对象。关于 CLASSTYPE 和 EXTENDED_CLASSTYPE 对象的详细描述请参考 RFC4124 和 draft-minei-diffserv-te-multi-class。

当沿途的 LSR 收到带有 CT 信息的 RSVP Path 消息，如果有足够的资源则接受建立新的 LSP；LSP 建立成功后，该节点重新计算各优先级的每 CT 的可预留带宽。这些预留信息会反馈给 IGP，向网络中的其它节点通告。

带宽约束模型

目前，IETF 定义了如下几种带宽约束模型：

- 最大分配模型 MAM (Maximum Allocation Model)：将一个 BC 映射到一个 CT，CT 间不共享带宽。MAM 的 BC Mode ID 为 1。

图 3-13 MAM 模型示意图



MAM 模型中，CT_i 所有 LSP 的带宽总和不超过 BC_i ($0 \leq i \leq 7$)；所有 CT 的所有 LSP 带宽总和不超过链路最大可预留带宽。

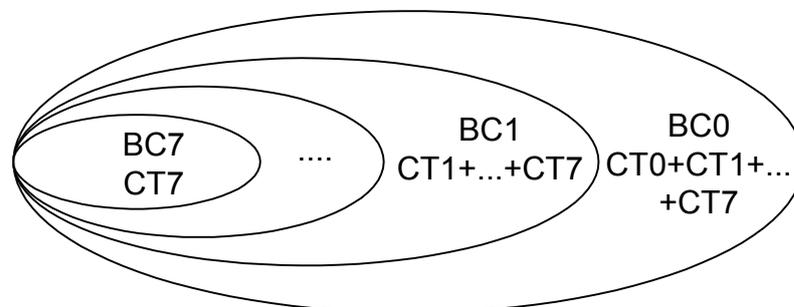
例如，假设某链路的带宽是 100Mbit/s，带宽模型为 MAM，且支持 3 个 CT（CT₀、CT₁ 和 CT₂）。BC₀ 为 20Mbit/s，用于承载 CT₀（假设为 BE 流）；BC₁ 为 50Mbit/s，用于承载 CT₁（假设为 AF 流）；BC₂ 为 30Mbit/s，用于承载 CT₂（假设为 EF 流）。则承载 BE 流的所有 LSP 的带宽总和不能够超过 20Mbit/s；承载 AF 流的所有 LSP 的带宽总和不能够超过 50Mbit/s；承载 EF 流的所有 LSP 的带宽总和不能够超过 30Mbit/s。

MAM 模型优点是不存在 CT 间的带宽抢占；缺点是可能存在带宽浪费。

- 俄罗斯套娃模型 RDM（Russian Dolls Model）：允许 CT 间共享带宽。RDM 的 BC Mode ID 为 0。

$BC_0 \leq$ 链路的最大可预留带宽。各 BC 之间存在不同的包含关系，如图 2：BC₇ 的带宽固定不变；BC₆ 包含 BC₇ 的带宽；……；BC₀ 包含所有 BC 的带宽。这种模型与俄罗斯玩偶类似：一个大娃娃套一个小娃娃，小娃娃又套一个更小的娃娃，依次类推。

图 3-14 RDM 模型示意图



例如，假设某链路的带宽是 100Mbit/s，带宽模型为 RDM，且支持 3 个 CT（CT₀、CT₁ 和 CT₂）。其中，CT₀ 用于承载 BE 流；CT₁ 用于承载 AF 流，CT₂ 用于承载 EF 流。BC₀ 为 100Mbit/s；BC₁ 为 50Mbit/s；BC₂ 为 20Mbit/s。那么承载 EF 流的所有 LSP 带宽总和不超过 20Mbit/s；承载 EF 流和 AF 流的所有 LSP 带宽总和能够超过 50Mbit/s；所有 LSP 的带宽总和不超过 100Mbit/s。

RDM 模型允许存在 CT 间的带宽抢占。各 CT 的抢占关系是：如果 $0 \leq m < n \leq 7$ ， $0 \leq i < j \leq 7$ ，则优先级为 m 的 CT_i 可以抢占优先级为 n 的 CT_i 的带宽和优先级为 n 的 CT_j 的带宽。但是 CT_i 的所有 LSP 带宽总和不超过 BC_i 的带宽值。

RDM 模型的优点是有效利用带宽。

- 扩展的最大分配模型 Extended-MAM：为了支持 E-LSP 而提出的一种带宽分配模式。Extended-MAM 的 BC Mode ID 为 254。

Extended-MAM 与 MAM 的区别在于，Extended-MAM 比 MAM 模型多支持了 8 个隐式 TE-Class（CT0 和 8 个优先级的组合）。这 8 个隐式 CT 使用 IGP 的 Unreserved BW TLV 进行泛洪。

IETF 模式与 Non-IETF 模式的区别

在按照 IETF 标准实现 DS-TE（IETF DS-TE）之前，华为已经实现了自己的 DS-TE (Non-IETF DS-TE)，在实际网络部署或者设备版本升级中，可能出现非 DS-TE 设备与 DS-TE 的设备共存，因此在 DS-TE 设计中包含了这两种模式，配置不同模式 DS-TE 的设备之间可以实现互通，以下是两种模式之间的区别：

表 3-3 IETF 模式和 Non-IETF 模式的区别

DS-TE 模式	Non-IETF 模式	IETF 模式
带宽模型	支持 MAM 和 RDM	支持 RDM、MAM 和 Extended MAM
CT 类型	支持 CT0 和 CT1，CT 配置为单 CT，允许的其中一种 CT	支持 CT0 ~ CT7，CT 配置为多 CT，支持八种 CT 类型同时配置
BC 类型	支持 BC0 和 BC1	支持 BC0 ~ BC7
TE-Class 映射表	可以配置 TE-Class 映射表，但不生效	支持配置和使用 TE-Class 映射表
IGP 消息	由 Unreserved bandwidth sub-TLV 携带基于优先级的可预留带宽	同时由 Unreserved bandwidth sub-TLV 和 Bandwidth Constraints sub-TLV 携带 CT 信息
RSVP 消息	由 ADSPEC 对象携带 CT 信息	单 CT： 由 CLASSTYPE 对象携带 CT 信息； 多 CT： 由 EXTENDED_CLASSTYPE 对象携带 CT 信息

MPLS DS-TE 模式切换

VRP 支持 Non-IETF 模式向 IETF 模式的切换，也支持 IETF 模式向 Non-IETF 模式的切换。模式切换的处理如下表所示。

表 3-4 DS-TE 模式切换

项目	Non-IETF 模式-->IETF 模式	IETF 模式-->Non-IETF 模式
带宽模型变化情况	带宽模型不变	Extended-MAM-->MAM;

项目	Non-IETF 模式-->IETF 模式	IETF 模式-->Non-IETF 模式
TE-Class 映射表变化情况	如果已配置 TE-Class 映射表，则使用配置的 TE-Class 映射表，否则采用缺省的 TE-Class 映射表。缺省的 TE-Class 映射表请参见表 3-5。	不使用 TE-Class 映射表。 如果用户配置了 TE-Class 映射表，则不删除映射表； 如果用户没有配置 TE-Class 映射表，则删除缺省的 TE-Class 映射表。
LSP 删除情况	在 Ingress 节点和 Transit 节点删除<CT,set-priority>组合和<CT,hold-priority>组合不在 TE-Class 映射表中的 LSP。	在 Ingress 节点和 Transit 节点删除以下类型的 LSP： 多 CT LSP； CT2 ~ CT7 的单 CT LSP。

表 3-5 缺省的 TE-Class 映射表

TE-Class	CT	优先级
TE-Class[0]	0	0
TE-Class[1]	1	0
TE-Class[2]	2	0
TE-Class[3]	3	0
TE-Class[4]	0	7
TE-Class[5]	1	7
TE-Class[6]	2	7
TE-Class[7]	3	7

典型组网应用

- DS-TE 应用场景 1：同一 VPN 不同业务接入

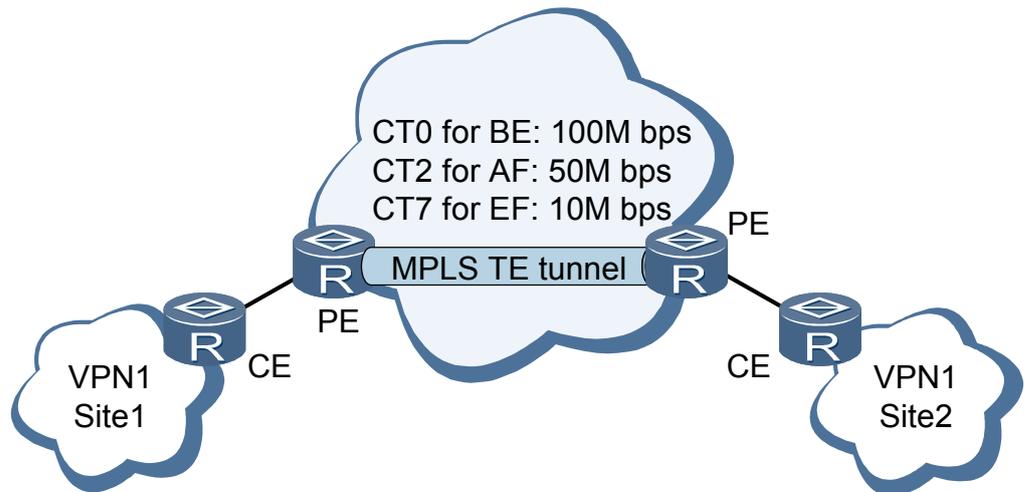
在使用 MPLS TE 作为隧道的 VPN 环境中，同一个 VPN 可能同时承载 EF（Expedited Forwarding）、AF（Assured Forwarding）及 BE（Best-Effort）业务。这意味着同一条隧道需要同时承载不同类型的业务。

为了防止同一条隧道中的不同业务之间相互干扰，可以创建多个 VPN 和 TE 隧道，不同的隧道承载不同的业务。但如果网络上存在多个 VPN 同时承载多种业务，则需要创建大量的 VPN 和隧道，造成资源浪费。

另一种方案是部署 DS-TE，使用一条多 CT 类型的 LSP 承载一个 VPN 的业务。一条多 CT LSP 可以预留多达 8 个 CT，每个 CT 对应一个 VPN 的一种业务，且业务之间互不干扰。

例如图 3-15，VPN1 同时接入三种业务：BE、AF 和 EF。只需要创建一条 TE 隧道，该隧道上配置 CT0（带宽 100M）、CT1（带宽 50M）、和 CT2（带宽 10M），然后在 Ingress 节点上将 VPN1 与隧道绑定，VPN1 的所有流量经过分类后进入对应 CT 的队列进行处理。

图 3-15 同一 VPN 不同业务接入同一条 LSP 组网



● DS-TE 应用场景 2: 不同 VPN 接入

在使用 MPLS TE 作为隧道的 VPN 环境中，可能存在多个 VPN 共用一条 TE 隧道的情况。这些 VPN 有着不同的 QoS 需求，会导致 VPN 之间竞争资源，使得 VPN 各个业务的 QoS 要求得不到保证。上述场景分为三种情况，下面对每种情况描述其解决方案：

- 不同 VPN，业务类型完全不同

这种情况下，只要所有 VPN 的业务数的总和不超过 8，则可以使用一条 TE 隧道承载。

例如 VPN1 和 VPN2 同时接入。VPN 的业务有 EF、BE 类型，VPN2 的业务为 AF 类型，则只需建立一条 TE 隧道，为每个 VPN 的不同业务配置不同 CT。CT 的个数为 VPN1 的业务数量 + VPN2 的业务数量 = 3。

- 不同 VPN，业务类型完全相同

这种情况下，所需 TE 的隧道数目为 VPN 的数目。每个隧道上所需的 CT 数等于对应 VPN 的业务数。

例如 VPN1 和 VPN2 同时接入。VPN 的业务有 EF、BE 类型，VPN2 的业务也是 EF、BE 类型，则需要建立两条 TE 隧道，不同 VPN 使用不同隧道。每个隧道上再为不同业务类型配置不同 CT。

- 不同 VPN，业务类型部分相同

这种情况下，也需要为每个 VPN 建立一个隧道。每个隧道上所需的 CT 数等于对应 VPN 的业务数。

● DS-TE 应用场景 3: VPN 与非 VPN 流量同时接入

VPN 网络中同时存在 VPN 与非 VPN 流量，且各自有着不同的 QoS 需求。如果使用同一条 TE 隧道承载，则可能导致 VPN 与非 VPN 流量之间竞争资源，使得各类业务的 QoS 要求得不到保证。

上述场景分为三种情况，下面对每种情况描述其解决方案：

- VPN 与非 VPN 业务类型完全不同

这种情况下，可以使用一条 TE 隧道承载。为 VPN 和非 VPN 的各类业务配置不同 CT。CT 的个数为 VPN 的业务数量 + 非 VPN 的业务数量，如 [图 3-16](#)。

- VPN 与非 VPN 业务类型完全相同

这种情况下，需要建立两条隧道，VPN 和非 VPN 业务分别使用一条隧道。每个隧道上为不同业务类型配置不同 CT，如 图 3-17。

- VPN 与非 VPN 业务类型部分相同

这种情况下，需要建立两条隧道，VPN 和非 VPN 业务分别使用一条隧道。每个隧道上为不同业务类型配置不同 CT，如 图 3-17。

图 3-16 VPN 与非 VPN 业务接入同一条隧道组网

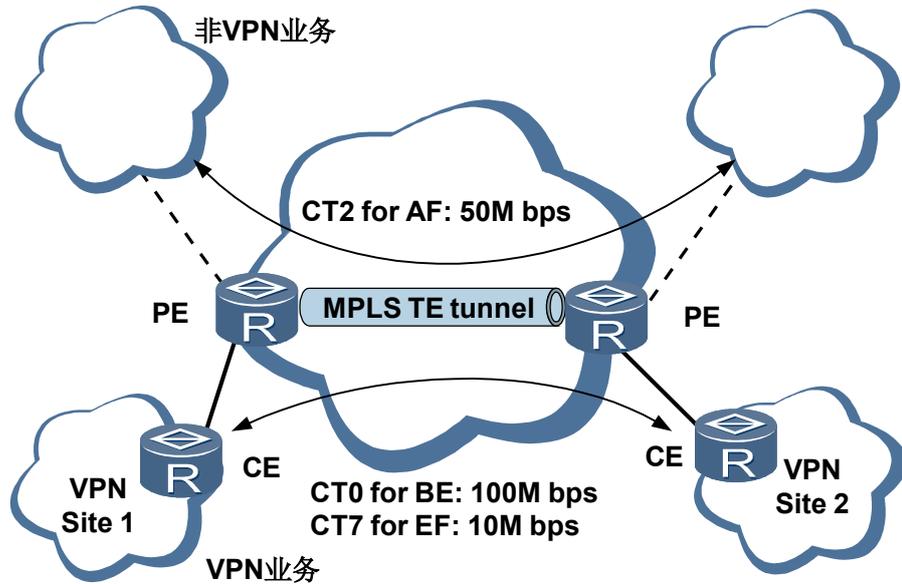
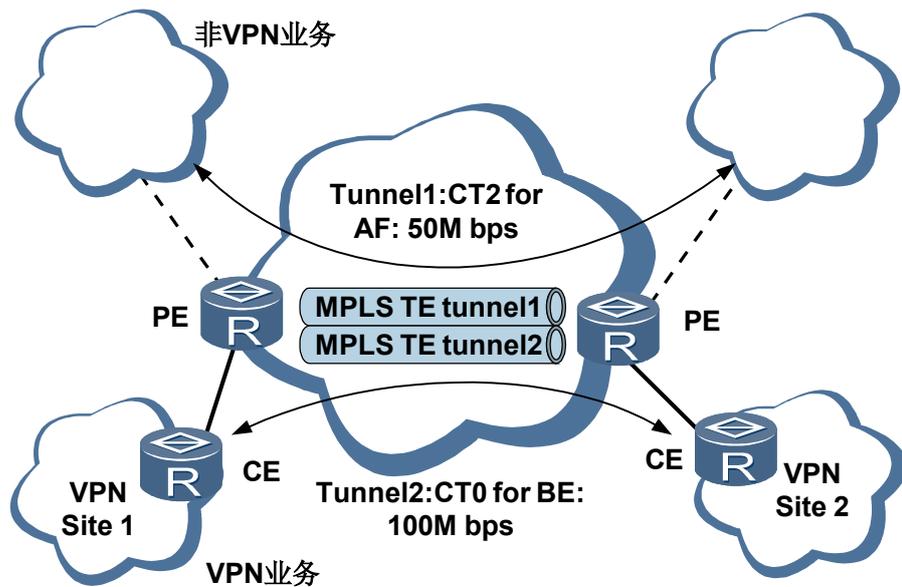


图 3-17 VPN 与非 VPN 业务接入分别接入一条隧道组网



- DS-TE 应用场景 4: 隧道保护中的 DS-TE

VPN 网络中同时存在 VPN 与非 VPN 流量，且各自有着不同的 QoS 需求。如果使用同一条 TE 隧道承载，则可能导致 VPN 与非 VPN 流量之间竞争资源，使得各类业务的 QoS 要求得不到保证。

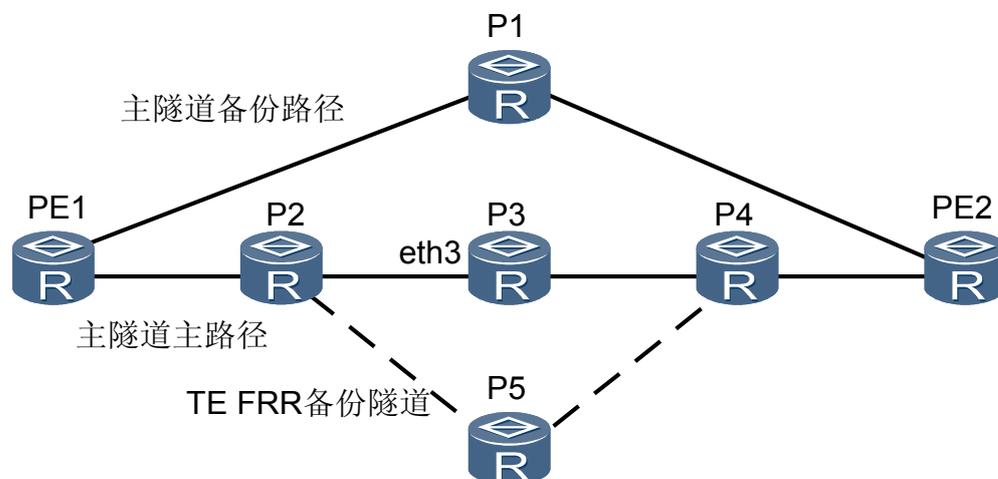
上述场景分为三种情况，下面对每种情况描述其解决方案：

表 3-6 隧道保护中的 DS-TE

保护方式	DS-TE 特性
TE FRR	分为两种情况： <ul style="list-style-type: none"> ● 需要带宽保护：手工 FRR 通过手工配置 Bypass 隧道的 CT 及其带宽来保证业务的 QoS，支持 1:1 保护和 N:1 保护；Auto FRR 则通过 Bypass 隧道继承主隧道的 CT 及带宽来保证业务的 QoS，只支持 1:1 保护，不支持 N:1 保护。 ● 不需要带宽保护：不考虑 Bypass 隧道的 CT 及其带宽，不论手工 FRR 还是 Auto FRR，都支持 1:1 保护和 N:1 保护。
CR-LSP 备份	备份隧道从主隧道继承 CT 类型及其带宽。逃生路径不保证 QoS，因此不继承主隧道的 CT 类型及其带宽。
隧道保护组	隧道保护组将两条独立配置的隧道进行绑定，使两条隧道形成主备关系。因此，备份隧道的 DS-TE 特性由配置决定。为保证业务的 QoS，备份隧道的 CT 类型及其带宽应该与主隧道相同。 另外，MPLS OAM 的检测报文通过 TE 隧道优先级最高的队列发送。

不同的保护可以为隧道提供保护，也可以将多种方式结合使用，如下图隧道备份和快速重路由的结合使用：

图 3-18 隧道备份与快速重路由结合使用应用场景



● DS-TE 应用场景 5：基于不同 DS-TE 模式的设备互通

在实际网络部署或者设备版本升级中，可能出现非 DS-TE 设备与 DS-TE 的设备共存，部署 Non-IETF 模式的设备与部署 IETF 模式的设备共存的现象。

对于普通的 TE 隧道（非 MPLS DS-TE 隧道），VRP 按照 CT0 方式，映射到 AF 类型。因此，VRP 支持如下情况的设备互通：

- DS-TE 设备与非 DS-TE 设备互通
 - 支持从非 DS-TE 设备建立普通 TE 隧道至 DS-TE 设备；
 - 支持从 DS-TE 设备建立普通 TE 隧道至非 DS-TE 设备。
- 与其他厂商的不支持 CLASSTYPE 对象的 DS-TE 设备互通
 - 可以解析携带如下 CT 信息的 Path 消息：
 - 由 EXTENDED_CLASSTYPE 对象携带 L-LSP 的 CT 信息；
 - 由 EXTENDED_CLASSTYPE 对象携带 CT0 信息。

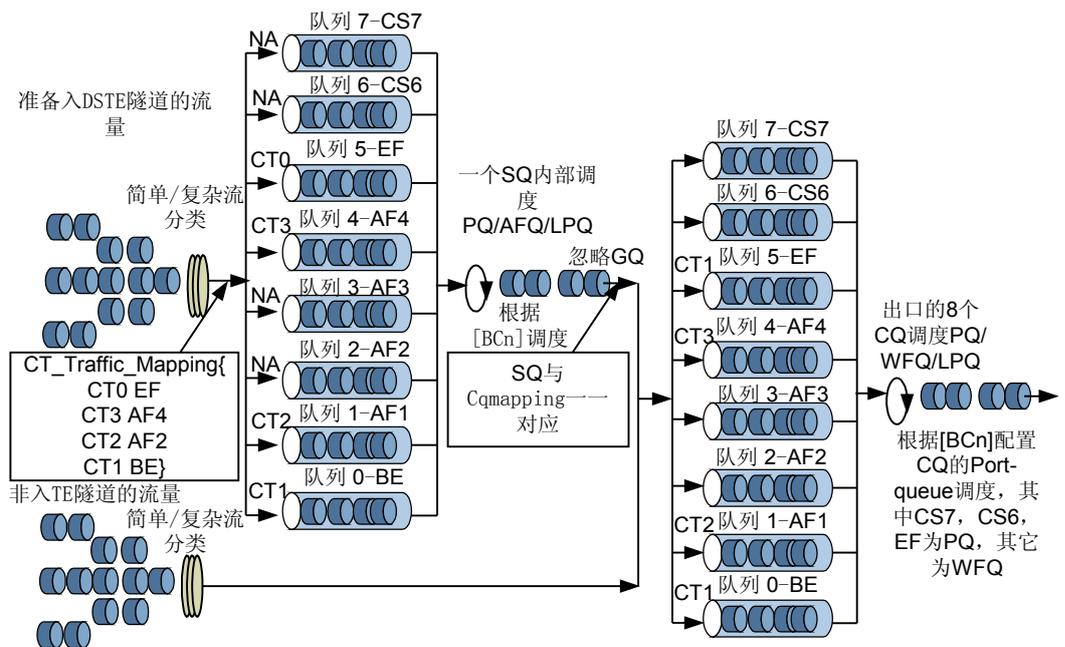
产品 DS-TE 的调度处理

R3 版本支持单 CT 的 DS-TE，支持两种业务类型 CT1 和 CT0，而且实现并非标准，但是目前产品对于 TE 的保护机制，LDP over TE 以及 RRVPN 项目中实现的 TE 带宽保证都已经支持，并且在支持 HQoS 的硬件上具备了支持最多 8CT 的 DS-TE 能力。

因此设计实现支持 8CT 的标准模式的 DS-TE，在原有单 CT 的 DS-TE 流量带宽保证实现 8CT 的带宽保证，8CT 的优先级调度，以及 CT 和 Cos 的灵活映射。

DS-TE 产品侧，Ingress PE 上行完成流分类，区分出报文的优先级。下行对不同优先级的报文进行差分优先级调度，通过调用 5 级层次化 QoS 的处理，在 FQ 侧完成流量的优先级区分，在 SQ 级进行总带宽的限制，如图 3-19 所示：

图 3-19 HQoS 调度



DS-TE 在 GQ 这里没有进行处理，而其中接口一级的层次化 QoS 由各个接口模块自己实现，DS-TE 不改变已有的接口级的层次化 QoS 功能。

用户无需详细掌握 HqoS 的内部处理，只需在全局配置 CT 与 FQ 队列的映射关系即可。其中每一个 CT 和 FQ 的一条流队列是一一对应的，这条流队列的调度行为决定了这条 CT 的调度行为。

用户全局可以配置 8 个 CT 到 FQ 队列的映射模板（包括一个默认的模板）。

默认的 CT 和 FQ 队列的映射关系如下：

CT 类型	FQ 队列	调度方式
CT7	CS7	PQ
CT6	CS6	PQ
CT5	EF	PQ
CT4	AF4	WFQ
CT3	AF3	WFQ
CT2	AF2	WFQ
CT1	AF1	WFQ
CT0	BE	LPQ

用户可以全局配置 8 个这样的映射模板，产品根据 TE 出接口应用的模板和各个 CT 的带宽就能够计算出一个 FQ 模板。通过灵活的映射关系，就可以满足 DS-TE 的各种模型的需要。

3.3.10 TE 隧道保护组

TE 隧道保护组主要用于在对网络性能要求较高的情况下，预留保护通道，从而在主隧道发生故障时通过备份隧道较快的恢复数据流的传输。

TE 隧道保护组通过预先配置一个主隧道和一个备份隧道，并将备份隧道绑定到主隧道上，则主隧道和备份隧道共同形成一个隧道保护组。

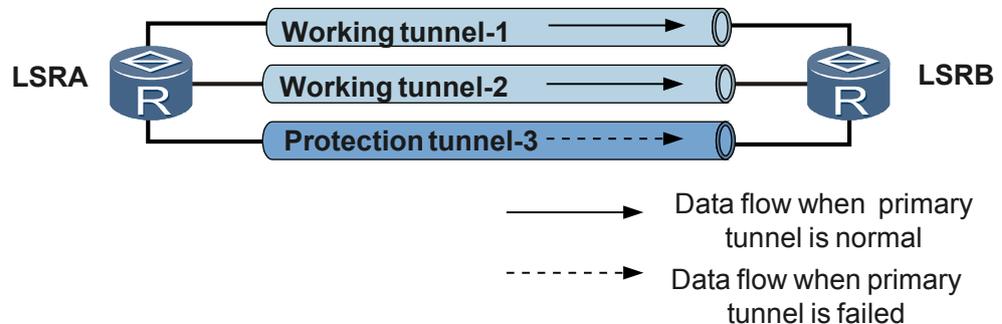
配置备份隧道时，可以单独配置自己的属性，这样更加利于用户规划网络。

隧道保护组中涉及的概念：

- 主隧道：被保护的隧道。
- 保护隧道：用来保护主隧道的隧道。
- 保护倒换：在隧道保护组中，当主隧道发生故障时，数据流迅速切换到备份隧道，从而提高网络可靠性。
- 1:1 保护：在隧道的入节点和出节点间提供主备两条隧道。正常情况下，数据在主隧道传输。当入节点通过检测机制发现主隧道故障时，进行保护倒换，将数据切换到备份隧道上继续传输。

- N:1 保护：将一条隧道作为多条主隧道的备份隧道，当任何一条主隧道故障时，都将数据倒换到共享的备份隧道上，这种模式主要是为了在网状拓扑结构的网络中节省带宽（如图 3-20）。

图 3-20 N:1 保护模式图

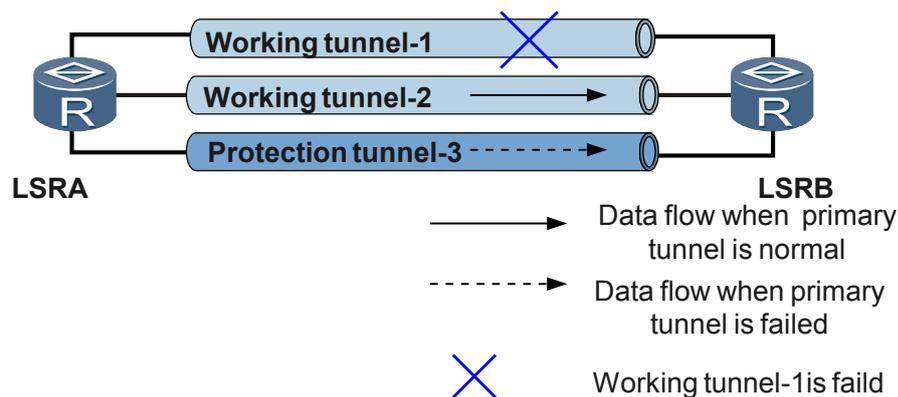


TE 隧道保护组的实现比较简单，主要操作都在入节点完成。用户在入节点通过配置为主隧道指定保护隧道，当主隧道发现故障(可以通过配置 OAM 或者 BFD 检测)之后，就会检测是否配置有保护隧道，及其保护隧道状态是否可用。如果有可用的保护隧道，则在入节点将流量切换到保护隧道。

如图 3-21，在 LSRA 上有主隧道 tunnel-1 和 tunnel-2，以及备份隧道 tunnel-3。

在入节点 LSRA 上指定主隧道 tunnel-1 和 tunnel-2 的保护隧道为 tunnel-3。根据配置的故障检测机制，当在入节点检测到 tunnel-1 隧道故障之后，则可以将流量切换到保护隧道 tunnel-3 上。而 tunnel-1 会一直尝试重建，如果建立成功，流量会根据配置的策略决定是否回切到主隧道上来。

图 3-21 隧道保护组的原理图



部署隧道保护组

TE 隧道保护组主要是通过预先的规划来提高主隧道的可靠性，所以在部署 TE 隧道保护组之前，需要对网络进行规划。为了能让保护隧道起到更好的保护作用，需要让保护隧道尽量不经过主隧道经过的链路和设备。

在配置隧道保护组完成之后，在控制平面可以使用多种切换模式供用户进行流量切换，包括手动和自动。而且用户可以配置相关切换时间。

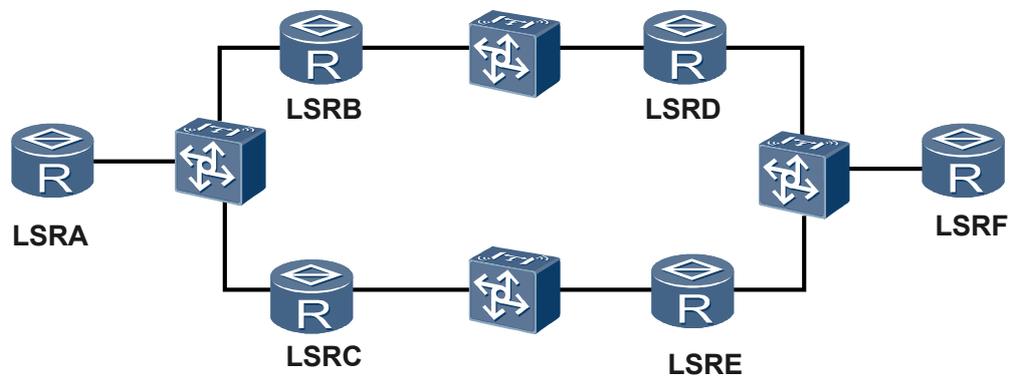
 说明

在部署过程中，保护隧道不能被其它保护隧道嵌套保护，而且保护隧道也不能使能 TE FRR。

3.3.11 BFD For TE CR-LSP

传统的检测机制，依靠包括 RSVP Hello 检测或者依靠 RSVP 刷新超时等检测，都具有检测速度缓慢的缺点。BFD 检测机制很好的克服了这些缺点，BFD 采用快速收发报文的机制，完成对隧道链路故障的快速检测，从而引导隧道上承载业务的快速切换，达到业务保护的目。

图 3-22 BFD 检测示意图



如图 3-22 所示，如果没有应用 BFD 检测，在 LSRE 发生故障时，由于二层交换机的出现，导致 LSRA 和 LSRF 无法立刻感知到故障发生；转而由 Hello 协议来检测，但会出现检测时间长的问题。

如果 LSRA、LSRB、LSRC、LSRD、LSRE 和 LSRF 全部应用 BFD，当 LSRE 发生故障时，LSRA 和 LSRF 会在很短的时间内检测到故障发生，并使数据流切换到 LSRA→LSRB→LSRD→LSRF。

BFD for TE CR-LSP 是对 CR-LSP 的检测，能够快速检测到 CR-LSP 的故障，并及时通知转发层面，从而保证流量的快速切换。BFD for TE CR-LSP 通常与 hot-standby CR-LSP 或者隧道保护组配合使用。

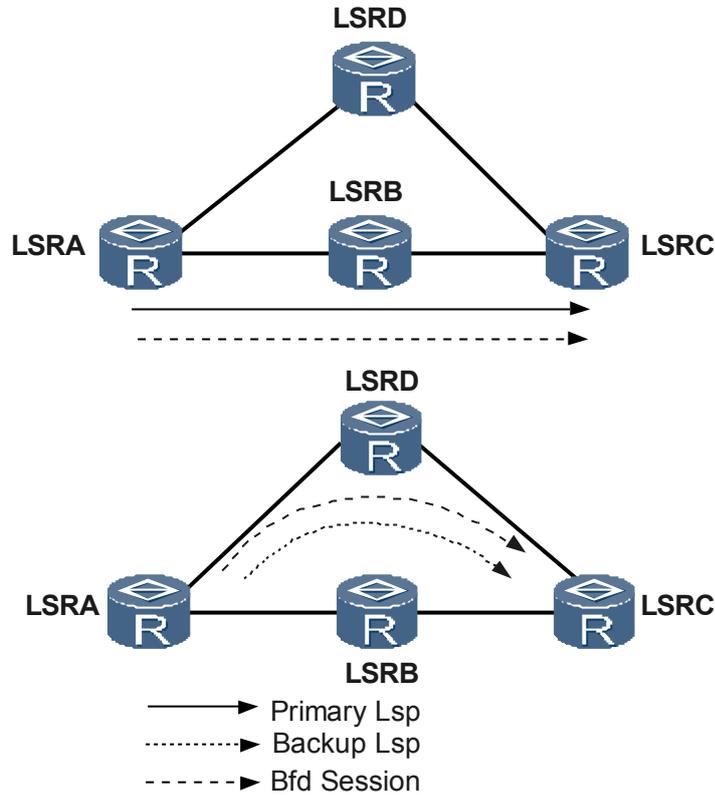
BFD 中涉及的基本概念：

- 静态 BFD 会话：本端标识符和对端标识符都需要手工指定，两端标识符必须匹配，否则会话无法建立。会话建立后，发送和接收时间间隔参数可以修改。
- 动态 BFD 会话：不需要指定本端标识符和对端标识符。路由协议邻居关系建立后，RM 会下发参数通告 BFD 建立会话，会话通过协商来确定本端标识符、对端标识符、发送和接收时间间隔参数。
- 检测周期：用于检测 BFD 会话是否正常的时间间隔。如果在检测周期内没有收到远端系统发送来的报文，则认为会话 down。

BFD 会话与 CR-LSP 绑定，即在入节点和出节点之间建立 BFD 会话。BFD 报文从源端开始经过 CR-LSP 转发到达宿端；宿端再对该 BFD 报文进行回应，通过此方式在源端可以快速检测出 CR-LSP 所经过链路的状态。

当检测出链路故障以后，BFD 将此信息上报给产品转发模块。转发模块查找备份 LSP，然后将业务流量切换到备份 CR-LSP 上。然后产品转发模块再将故障信息上报给控制层面，如果采用的是动态 BFD for TE CR-LSP，控制层面会主动去创建备份 CR-LSP 的 BFD 会话。如果采用的是静态 BFD for TE CR-LSP 时且需要对备份 CR-LSP 进行检测，则可以为其配置 BFD 检测。

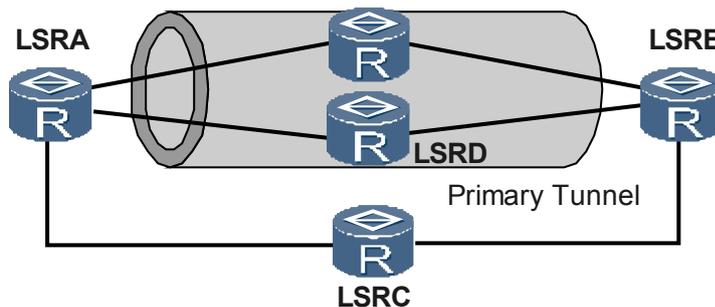
图 3-23 BFD 会话切换前后示意图



如图 3-23 所示，BFD 建立会话检测主 CR-LSP 所经过的链路。当主 CR-LSP 所经过的链路出现故障以后，在源端 BFD 会立即通告该故障信息。然后入节点将流量切换至备份 CR-LSP，同时在备份 CR-LSP 所经过的路径上建立新的 BFD 会话，用于检测备份 CR-LSP 所经过的链路状态。

BFD for TE 的部署

图 3-24 BFD for TE 部署



1. 主 CR-LSP 和热备份 CR-LSP 之间的切换

如图 3-24，在 LSRA→LSRB 之间建立一条主隧道，同时配置热备份 CR-LSP。在 LSRA 上建立一个到 LSRB 的 BFD 会话，用于检测该隧道中的主 CR-LSP。当主 CR-LSP 链路出现故障时，BFD 会快速通知 LSRA。LSRA 收到故障信息以后，立即将流量切换到热备份 CR-LSP 上，从而保证流量不中断。

2. 主隧道和备份隧道之间的切换

如图 3-24，在 LSRA→LSRD→LSRB 之间建立一条主隧道，同时在 LSRA→LSRC→LSRB 之间建立一条备份隧道。在路径 LSRA→LSRD→LSRB 上建立一个 BFD 会话，用于检测主隧道的路径。当主链路出现故障时，BFD 会快速通知 LSRA。LSRA 收到故障信息以后，立即将流量切换到备份隧道上，从而保证流量不中断。

3.3.12 BFD for TE Tunnel

双向转发检测 BFD（Bidirectional Forwarding Detection）主要用于检测转发引擎之间的通信故障。具体来说，BFD 是对两个系统间的、同一路径上的一种数据协议（data protocol）的连通性进行检测。这条路径可以是物理链路或逻辑链路，其中就包括 TE 隧道。

BFD for TE Tunnel 是使用 BFD 检测整条 TE 隧道，可触发 VPN FRR 应用在主路径故障时进行快速流量切换，以减少对业务的影响。

VPN FRR 的技术场景下，在 PE1 和 PE2 之间部署隧道，且对该隧道进行 BFD 检测。如果 BFD 检测出该隧道故障，则实现 VPN FRR 毫秒级的快速倒换。

TE 隧道在部署 BFD for TE Tunnel 业务后，无法再部署 MPLS OAM 检测业务。

3.3.13 RSVP 认证

RSVP 使用 RawIP 传递协议报文，而 RawIP 本身不提供安全性，报文容易被篡改，设备容易受到攻击。RSVP 消息通过验证摘要信息的正确性，来防止消息被篡改或伪造的恶意攻击，增强网络的可靠性和安全性。

基本原理

在需要认证的两个节点上，用户需要配置相同的密钥。在发送报文时，节点使用密钥为报文计算得到一个摘要（通过 HMAC-MD5 算法），摘要信息作为报文的一个对象（Integrity 对象），随着报文一起发送到对端节点。对端节点使用相同的密钥、相同的算法重新计算报文摘要，然后比较两个摘要是否相同，如果相同则接受，否则丢弃。

RSVP 认证不能防止回放攻击，也无法解决因 RSVP 报文的失序导致邻居之间认证关系终止的问题。为了解决该问题，引入了 RSVP 认证增强功能。RSVP 认证增强是在基于原有认证的基础上增加了认证生存时间、握手和消息滑窗等特性，这样使得 RSVP 自身的安全性得到很大的提高，并大大加强了在网络阻塞等恶劣网络环境时对用户进行身份验证的能力。

RSVP 密钥管理方式

RSVP 密钥管理包括以下两种方式：

- MD5 密钥

用户可以在 RSVP 接口、邻居下，以明文或者密文的方式输入密钥，密钥算法为 MD5。这种密钥管理方式的特点是：

- 每个协议特性都需要配置自己的密钥，密钥不能共享。
- 每个接口、邻居只能配置一个密钥，要更换密钥必须重新配置。

- **Keychain 密钥**

Keychain 是一种增强型加密算法，允许用户定义一组密码，形成一个密码串，并且分别为每个密码指定加解密算法及密码使用的有效时间。在收发报文时，系统会按照用户的配置选出一个当前有效的密码，并按照与此密码相匹配的加密解密算法，进行发送时加密和接收时解密报文。此外，系统可以依据密码使用的有效时间，自动完成有效密码的切换，避免了长时间不更改密码导致的密码易破解问题。

这种密钥管理方式的特点是：

- Keychain 的密码、所使用的加解密算法以及密码使用的有效时间可以单独配置，形成一个 Keychain 配置节点，每个 Keychain 配置节点至少需要配置一个密码，并指定加解密算法。
- Keychain 可以被各个协议特性引用，实现密钥集中管理、多特性共享。

RSVP 支持在接口、邻居下引用 Keychain，并仅支持 HMAC-MD5 算法。

RSVP 的认证级别

RSVP 的认证级别分为两种：

- **面向邻居的认证**

该级别的认证是指用户可以根据不同的邻居地址配置认证密钥等信息，RSVP 会针对每个邻居进行单独的认证。

有两种配置方式：

- 以邻居设备的某接口的 IP 地址作为邻居地址进行配置。
- 以邻居设备的 LSR ID 作为邻居地址进行配置。

- **面向接口的认证**

用户在接口上配置认证，RSVP 会根据消息的入接口进行认证处理。

这两个认证级别的优先级顺序由高到低依次为：面向邻居的认证、面向接口的认证。只有当高优先级没有使能认证的情况下才会进行低优先级的认证处理，一旦高优先级认证没有通过，则丢弃该报文。

3.3.14 RSVP GR

RSVP GR (Graceful Restart) 是 RSVP-TE 的一种状态恢复机制。

RSVP GR 功能基于无间断转发 NSF (Non-Stop Forwarding) 思想设计。当节点控制层面发生故障时，通过上下游邻居发送消息对节点的 RSVP 软状态进行恢复，而转发层面则不感知故障也不受故障影响，从而保证了流量的稳定性和可靠性。

RSVP GR 通过 Hello 特性扩展来检测邻居的 GR 状态，关于 Hello 特性的相关描述请参见 [RSVP Hello](#)。

RSVP GR 的实现机制如下：

如图 3-25，重启节点故障重启后，将停止向邻居节点发送 Hello 消息。使能了 GR 的邻居节点在连续 3 次未收到 Hello 消息后，认为邻居在做 GR，所有的转发信息都将继续保持。同时接口板继续传输业务，并等待重启节点进行 GR 恢复。

在重启节点启动后，如果收到邻居的 Hello 消息，会向邻居节点发送 Hello 消息，处于隧道上游和下游节点的处理方式是不一样的：

- 上游的支持节点收到该消息后，向重启节点（自己的下游）发出 GR Path 消息。
- 下游的支持节点收到该消息后，向重启节点（自己的上游）发出 Recovery Path 消息。

图 3-25 通过 GR Path 消息和 RecoveryPath 消息重建示意图



当重启节点收到 GR Path 消息和 Recovery Path 消息时，根据这两个消息重新建立 LSP 相关的状态信息，这样本地控制层面信息就恢复了。

若下游支持节点不能发送 Recovery Path 消息，则本地状态仅通过 GR Path 消息来重建。

3.3.15 RSVP 摘要刷新

在不用传输整个 RSVP 刷新消息的情况下，仅传输刷新消息对应的很小的摘要来维护 RSVP 软状态，同时响应 RSVP 状态变化。

每个 RSVP 会话都需要在每个刷新周期内产生、传送、接收和处理 RSVP 的 Path 消息和 Resv 消息。随着 RSVP 会话的不断增加，会产生大量的刷新消息来维持 RSVP 软状态。当一个非 RSVP 刷新消息在传输时丢失，就会产生可靠性和延迟问题。摘要刷新机制的引入，可以解决上述这两个问题。在减少大量刷新消息的同时，提高了 RSVP 消息传输的可靠性，并提高资源的利用率。

摘要刷新 Srefresh (Summary Refresh) 可以不传送标准的 Path 或 Resv 消息，而仍能实现对 RSVP 状态的刷新。使用摘要刷新的好处主要是它减少了维持 RSVP 状态所需传输及处理的信息量。使用摘要刷新消息更新 RSVP 状态时，常规的刷新消息就被抑制了。

摘要刷新消息承载了一系列 Message_ID 对象，用于识别需要被刷新的 Path 及 Resv 状态。摘要刷新需要与 Message_ID 扩展配合使用。只有那些已经包含 Message_ID 的 Path 和 Resv 消息发布过的状态才能使用摘要刷新机制刷新。

当节点接收到一条摘要刷新消息时，与本地状态块 (PSB 或 RSB) 进行匹配。

- 如果匹配，就更新本地状态，就象接收到一个标准的 RSVP 刷新消息一样。
- 如果不匹配，节点将发送一个 NACK 消息来通知摘要刷新消息的发送者。并根据 Path 或 Resv 消息刷新相应的 PSB 或 RSB，同时更新 Message_ID。

Message_ID 对象中包含了 Message_ID 序列号。当 LSP 发生变化时，相应的 Message_ID 序列号增大。节点收到 Path 消息时，将其中的 Message_ID 序列号与本地状态块中保存的 Message_ID 序列号比较：

- 如果相等，则保持状态不变。
- 如果大于，则表示状态已更新。

RSVP 目前支持全局使能摘要刷新和接口使能摘要刷新。如果使能了全局摘要刷新，则整套设备支持 RSVP 摘要刷新能力，如果只能使能接口摘要刷新能力，则该接口对应的链路上支持 RSVP 摘要刷新能力。

3.3.16 RSVP Hello

RSVP 节点给邻居发送 Hello 消息来检查邻居是否可达。

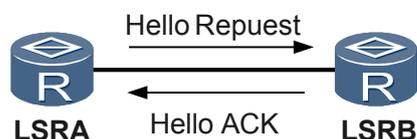
Hello 机制提供了节点到节点的缺陷检测方式。当发现节点间链路故障时，它的处理类似于链路层缺陷的处理。当链路层缺陷的通告不可用，或链路层提供的这个缺陷检测机制对节点缺陷检测不充分时，就可以使用 Hello 检测机制。

- 当使用快速重路由功能时，若 Hello 检测到邻居不可达时，会触发流量切换到 Bypass 路径上，防止因为下一跳不可达而造成流量丢失。
- 当使用 RSVP GR 功能时，Hello 还可以检测邻居节点是否重启，只有检测到邻居节点重启才能支持邻居节点恢复 RSVP 状态。

RSVP Hello 的实现原理如下：

1. Hello 握手机制

图 3-26 Hello 握手机制



如图 3-26，LSRA、LSRB 之间有链路直接相连。

- 当 LSRA 接口下使能了 RSVP Hello 时，LSRA 会向 LSRB 发送 Hello Request 消息。
 - 若 LSRB 收到了 Hello 消息，并且 LSRB 也使能了 RSVP Hello，就会给 LSRA 节点回复 Hello ACK 消息。
 - LSRA 收到 LSRB 的 Hello ACK 消息后，就确认 LSRA 的邻居 LSRB 是可达的。
- #### 2. 检测邻居丢失

在 LSRA 向 LSRB 发送 Hello Request 握手成功后，LSRA 与 LSRB 就开始互通 Hello 消息。当 LSRA 连续三次向 LSRB 发送 Hello Request 消息后，LSRB 仍然没有给 LSRA 回 Hello ACK 消息，此时就认为 LSRB 邻居丢失，重新初始化 RSVP Hello。

当 LSRA 和 LSRB 之间存在 LSP 时，

- 如果没有使能 GR 功能，但使能了快速重路由功能，则 Hello 检测到邻居丢失时，会触发流量切到 Bypass 路径上，保证流量不中断。
 - 如果使能了 GR 功能，则优先按照 GR 方式处理。
- #### 3. 检测邻居重启
- 当 LSRA 和 LSRB 都使能 RSVP GR 功能时，在 Hello 检查到邻居 B 丢失后，LSRA 就等待 LSRB 发送有 GR 扩展的 Hello Request 消息，此时 LSRA 开始支持 B 恢复 RSVP 状态。LSRB 收到 LSRA 回复的 Hello ACK 消息后，知道 LSRA 开始支持 LSRB 做 GR。然后 LSRA 和 LSRB 互通 Hello 消息，维持 GR 恢复状态。

3.3.17 BFD for RSVP

BFD for RSVP 使用 BFD 检测 RSVP 邻居关系。当 RSVP 相邻节点之间存在二层设备比如 HUB 时，这两个节点只能根据 Hello 机制感知链路故障，故障时间为秒级，这将导

致数据大量丢失。BFD for RSVP 可实现毫秒级故障监测时间，并配合 RSVP 协议快速的发现 RSVP 邻接故障。BFD for RSVP 一般用在 TE FRR 中 PLR 节点与主路径的 RSVP 邻居之间存在二层设备的情况。

图 3-27 BFD for RSVP 示意图



如图 3-27，BFD for RSVP 主要是在 RSVP 邻居之间建立 BFD 会话，用以检测这两个邻居之间的链路状态，让 RSVP 模块快速感知到链路失效。

BFD for RSVP 的检测对象与 BFD for CR-LSP 和 BFD for TE 不同。BFD for RSVP 是对 IP 层的检测，在 RSVP 邻居之间只能建立单跳 BFD 会话。在应用场景上，BFD for RSVP 也与 BFD for Tunnel 和 BFD for TE 不同，BFD for Tunnel 和 BFD for TE 主要是端到端的检测，而 BFD for RSVP 则是检查邻居节点间的链路状态。

BFD for RSVP 可以与 BFD for OSPF、BFD for ISIS 和 BFD for BGP 共享会话。如果本地的 BFD for RSVP 与其他协议共享 BFD 会话时，则本地节点分别选择所有共享 BFD 会话的协议的发送时间间隔、接收时间间隔、本地检测倍数的最小值做为本地的 BFD 会话参数。

3.3.18 TE-LSP 配置模板

为了简化 TE Tunnel 接口下配置的命令，将与 LSP 路径相关的配置信息提炼到 TE-LSP 属性模板下配置，这样的 TE-LSP 属性模板又称为 TE-LSP 配置模板。

配置模板下可配置的属性包括：带宽、显式路径、颜色、优先级、跳数限制、记录路由（标签）标志、快速重路由标志、bypass 隧道约束属性等与 LSP 建立相关的属性。

全局视图下配置了 TE-LSP 属性模板之后，在隧道接口下可以指定主 CR-LSP、热备份 CR-LSP 或普通备份 CR-LSP 使用的属性模板，引用某个属性模板的 LSP 建立时使用属性模板中指定的属性。

在隧道接口下配置属性模板，有以下几个优点：

- 当多条隧道使用相同的属性时，使用属性模板可以简化 TE 隧道接口下的配置；
- 使用属性模板配置热备份/普通备份 CR-LSP 时，可以为热备份/普通备份 CR-LSP 提供更加丰富的属性选项；
- 在隧道接口下可以为热备份/普通备份 CR-LSP 指定多个属性模板，从而为 TE 隧道的 LSP 建立提供多种不同的路径保护选项；
- 使用属性模板后，隧道接口下可以同时提供热备份、普通备份和逃生路径的保护方式；
- 通过修改 TE-LSP 配置模板的配置，使引用属性模板的隧道的 LSP 所使用的属性得到更新，从而增加了隧道配置的灵活性。

实现机制

- 属性模板配置

在全局视图下配置属性模板，属性模板中可以配置的属性包括：带宽、显式路径、颜色、优先级、跳数限制、记录路由（标签）标志、快速重路由标志、bypass 隧道约束属性。

- 隧道接口下配置和使用属性模板

隧道接口下可以指定主 CR-LSP、热备份 CR-LSP 或普通备份 CR-LSP 所使用的属性模板，并且热备份/普通备份 CR-LSP 的属性模板可以有多个。这些属性模板采用不同的配置，为隧道接口下 CR-LSP 的建立提供多种不同的路径选项。

隧道接口下配置的主 CR-LSP 属性模板中的属性用于创建主 CR-LSP，隧道接口下配置的热备份 CR-LSP 属性模板中的属性用于创建热备份 CR-LSP，隧道接口下配置的普通备份 CR-LSP 属性模板中的属性用于创建普通备份 CR-LSP。配置了属性模板之后，还可以在隧道接口下配置逃生路径。

当主 CR-LSP 使用属性模板创建成功后，可以使用热备份 CR-LSP 属性模板创建热备份 CR-LSP。由于热备份 CR-LSP 模板有多个，因此在创建热备份 CR-LSP 时依次尝试各个模板，直到使用其中某个模板创建热备份 CR-LSP 成功。

主 CR-LSP 无法创建成功或主 CR-LSP 失效并且无热备份 CR-LSP 时，可以创建普通方式的备份 CR-LSP。使用属性模板方式创建普通备份 CR-LSP 时，同样可以使用多个不同的属性模板依次进行尝试。

如果主 CR-LSP、热备份 CR-LSP 和普通备份 CR-LSP 都无法创建成功，则可以尝试创建使用逃生路径。

- 修改被隧道引用的属性模板的配置

当属性模板被隧道引用后，可以通过修改属性模板的配置改变所有引用该属性模板的隧道的 CR-LSP 所使用的属性。

根据所修改的属性模板的属性的不同，对隧道下 CR-LSP 建立的影响也不同。如果修改了属性模板的优先级或带宽类型，会引起隧道的 break-before-make；如果修改了属性模板中的其他 make-before-break 属性，会进行隧道的 make-before-break。

- 隧道接口下属性模板配置与已配置的命令的共存

隧道接口下配置属性模板后，还可以单独配置属性模板中已配置过的属性。

当隧道接口下既配置了属性模板，又使用已配置的命令配置了属性模板中配置过的属性时，按照隧道接口下配置优先的原则选取属性进行 CR-LSP 的创建。

3.3.19 MPLS LSR-ID 的多区域发布

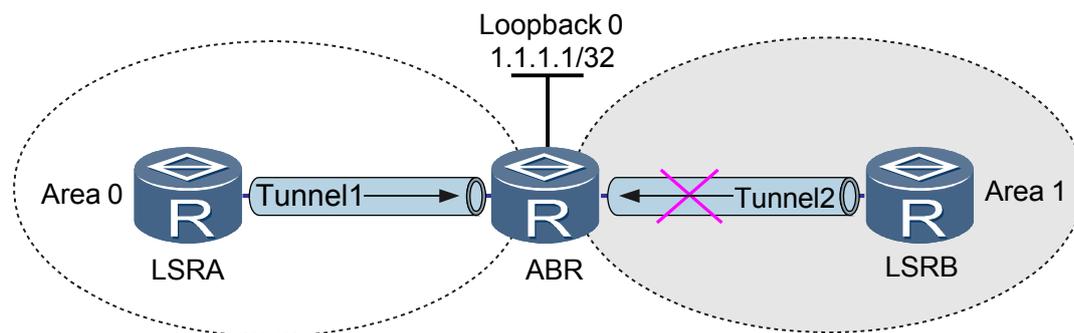
定义

OSPF 向本设备连接的所有区域发布地址为 MPLS LSR-ID 的区域内路由。

目的

当一台 ABR 同时作为两个区域中 Tunnel 接口的终点时，OSPF 只认为其中一个 Tunnel 是有效的。

图 3-28 ABR 同时作为两个区域中的 Tunnel 接口终点的示意图



对于 OSPF 而言，Tunnel 生效的必要条件是有到达目的地址的区域内路由。而在 OSPF 协议中一个接口只能属于一个区域，即 Loopback 接口的 IP 地址对应的路由只能在一个区域中成为区域内路由，在其它区域中都是区域间路由。

为了解决这个问题，OSPF 向本设备连接的所有区域发布地址为 MPLS LSR-ID 的区域内路由。

原理

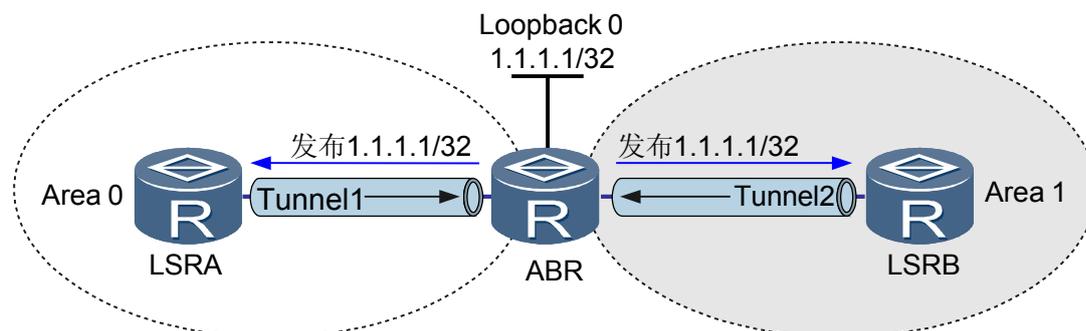
OSPF 进程会为每个区域生成一个 Router-LSA。在 Router-LSA 中添加一条 MPLS LSR-ID 的 stub link 后，每个区域在进行区域内路由计算时，就可以计算出以 MPLS LSR-ID 为目的地址的区域内路由。

说明

如果同时满足以下两个条件，则一个 OSPF 区域的 Router-LSA 中会有两条相同的 stub link。

- 在 OSPF 的某个区域下与 MPLS LSR-ID 对应的 Loopback 接口使能了 OSPF。
- 在 OSPF 视图下配置 `advertise mpls-lsr-id` 命令且此命令生效。

图 3-29 ABR 同时向两个区域发布 MPLS LSR-ID 的区域内路由的示意图





说明

多区域发布的条件:

- 存在 IP 为 MPLS LSR-ID 的接口。
- OSPF 进程下配置 **advertise mpls-lsr-id** 命令。
- 使能了 **mpls** 和 **mpls te** 能力。

3.4 术语与缩略语

缩略语

缩略语	英文全名	中文解释
RSVP	Resource Reservation Protocol	资源预留协议
FRR	Fast Re-Route	快速重路由
CSPF	Constrained Shortest Path First	基于约束的路径最短优先
TE	Traffic Engineering	流量工程
MP	Merge Point	聚合点
PLR	Point of Local Repair	本地修复节点
CT	Class Type	分类, 这里是指带宽类型
PSB	Path State Block	路径状态块
RSB	Reserved State Block	预留状态块

4 MPLS OAM

关于本章

- 4.1 介绍
- 4.2 参考标准和协议
- 4.3 原理描述
- 4.4 术语与缩略语

4.1 介绍

定义

OAM (Operation Administration & Maintenance) 是降低网络维护成本的重要手段, MPLS OAM 机制用于 MPLS 层的运维管理。

MPLS 支持多种三层和二层协议, 如 IP、FR、ATM、Ethernet 等。MPLS 提供一个完全不依赖于任何上层或下层的 OAM 机制, 在 MPLS 的用户平面实现以下功能:

- 检测 LSP 的连通性
- 衡量网络的利用率以及度量网络的性能
- 在链路出现缺陷或故障时迅速进行保护倒换, 以便能根据与客户签订的 SLA (Service Level Agreements) 提供业务

目的

MPLS 作为可扩展的下一代网络的关键承载技术, 提供具有 QoS 保障的多业务能力, 并且, MPLS 引入了一个独特网络层次, 会存在由这个新的网络层引起的故障。因此, MPLS 网络需要具备 OAM 能力。

承载 MPLS 的服务层 (server-layer), 例如 SONET/SDH, 以及利用 MPLS 的客户层 (client-layer), 例如 IP、FR、ATM, 都有各自的 OAM 机制。但 MPLS 网络层本身的故障不能完全通过其他层的 OAM 机制解决。并且, 网络技术的分层要求, 也需要 MPLS 具有自己独立的 OAM 机制, 从而减少各层之间的依赖关系。

使用 MPLS OAM 机制, 可以有效地检测、确认并定位出 MPLS 层内部的缺陷; 报告缺陷并做出相应的处理; 在出现故障的时候, 能够提供保护倒换的触发机制。

4.2 参考标准和协议

本特性的参考资料清单如下:

MPLS OAM 基于 ITU-T 协议设计和实现, RFC 只做参考。

文档	描述	备注
ITU-T Recommendation Y.1710	Requirements for Operation & Maintenance functionality for MPLS networks	基于此实现
ITU-T Recommendation Y.1711	Operation & Maintenance mechanism for MPLS networks	基于此实现
ITU-T Recommendation Y.1720	Protection switching for MPLS networks	基于此实现

文档	描述	备注
RFC3429	Assignment of the 'OAM Alert Label' for Multiprotocol Label Switching Architecture (MPLS) Operation and Maintenance (OAM) Functions	对 RFC 只做参考
RFC4377	Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks	对 RFC 只做参考
RFC4378	A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)	对 RFC 只做参考

4.3 原理描述

4.3.1 MPLS OAM 检测

4.3.2 反向通道

4.3.3 MPLS OAM 自动协议

4.3.4 保护倒换

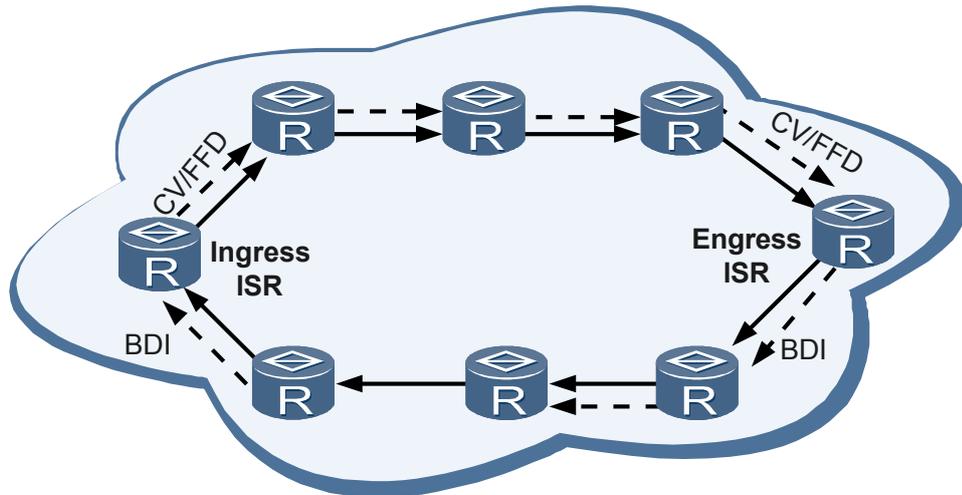
4.3.1 MPLS OAM 检测

MPLS OAM 使用的报文分为三类：

- 连通性检测：包括两种类型的探测报文
 - FFD (Fast Failure Detection)
 - CV (Connectivity Verification)
- 前向缺陷通告 FDI (Forward Defect Indication)
- 后向缺陷通告 BDI (Backward Defect Indication)

MPLS OAM 检测功能是指对 TE LSP 的连通性检测。MPLS OAM 通过在被检测的 TE LSP 上周期性发送检测报文 CV 或 FFD 实现。

图 4-1 MPLS OAM 连通性示意图



如图 4-1，MPLS OAM 的工作过程如下：

1. 入节点发送 CV/FFD 检测报文，报文通过被检测的 LSP 到达出节点。
2. 出节点把接收到的报文类型、频率、TTSI 等信息与本地记录的应该收到的值相比较，判断报文是否正确，并统计检测周期内收到的正确报文与错误报文的数量，从而对 LSP 的连通性进行监控。
3. CV 报文的检测频率为固定值，FFD 报文的检测周期为检测频率的三倍。
4. 当出节点检测到 LSP 缺陷后，分析缺陷类型，通过反向通道将携带缺陷信息的 BDI 报文发送到入节点，从而使入节点及时获知缺陷状态。如果正确配置了保护组，还会触发相应的保护倒换。

检测到的缺陷分为三类：非 MPLS 层缺陷、MPLS 层缺陷、其他缺陷。

- 非 MPLS 层缺陷：

- dServer：服务层缺陷。任何来自 MPLS 网络下层的非 MPLS 层技术的的服务层缺陷。

承载 MPLS 的下层网络可能会有其自身的保护及缺陷检测机制，当一条 LSP 上某处出现底层缺陷后，距离该故障最近的 LSR（下游方向）应能将该缺陷通告给 Egress 节点。对于发生的底层缺陷不应触发倒换动作，仅上报网管，但可以通过适当的方式通知 Ingress 节点（BDI 报文）。

- dPeerME：对等实体缺陷。任何来自 MPLS 子网外对等维护实体的非 MPLS 层技术的的服务缺陷。

- MPLS 层缺陷：

- dLOCV：连通性校验丢失缺陷。

在任意 3 个连续的 CV/FFD 发送周期内没有接收到相应的 CV/FFD 报文即认为发生该缺陷。

- dTTSI_Mismatch：TTSI 失配缺陷。

在任意 3 个连续的 CV/FFD 发送周期内没有接收到带有正确 TTSI 的 CV/FFD 报文即认为发生该缺陷。

- dTTSI_Mismerge：TTSI 错误合并缺陷。

在任意 3 个连续的 CV/FFD 发送周期内接收到既带有正确 TTSI 又带有错误 TTSI 的 CV/FFD 报文，即认为发生该缺陷。

- dExcess: 连通性检测报文超速缺陷。

在任意 3 个连续的 CV/FFD 发送周期内接收到超过（包括）5 个正确的 CV/FFD 报文，即认为发生该缺陷。

- 其他缺陷:

dUnknown: 在 MPLS 网络中出现未知缺陷。

这种缺陷可以自行定义用，比如 Egress 检测到在同一条 LSP 中既存在 CV 报文又存在 FFD 报文，类似这种协议没有规定的特殊缺陷可以用 dUnknown 来标识。

4.3.2 反向通道

在配置 OAM 基本检测功能时，需要为被检测的 LSP 绑定一条反向通道，传送 BDI 报文。

反向通道可以是与被检测 LSP 具有相反入节点和出节点的 LSP，也可以是连接被检测 LSP 的入节点和出节点的非 MPLS 路径。

具体来说，承载 BDI 报文的反向通道包括以下三种类型：

- 专用反向 LSP
- 共享反向 LSP
- 非 MPLS 返回路径

 说明

目前的实现中，华为设备只支持使用 TE 隧道作为反向通道。

4.3.3 MPLS OAM 自动协议

ITU-T Recommendation Y.1710 和 Y.1711 协议中存在一些不完善的地方，例如：

- 如果 LSP 入节点晚于出节点开启 OAM 功能，或出节点开启而入节点不开启 OAM 功能，出节点将产生连通性检测丢失 dLOCV（Loss of Connectivity Verification defect）告警。
- 停止 OAM 功能时存在同样的问题。当需要修改检测报文的类型以及刷新频率时，必须先分别在出节点和入节点停止 OAM 功能。
- 入节点和出节点的 OAM 参数需要分别配置，这容易导致人为的检测报文类型及发送频率配置不一致。

华为设备实现 OAM 自动协议解决如上的问题。

OAM 自动协议在出节点配置，提供首包触发和动态启停功能。

MPLS OAM 自动协议为华为专利。

4.3.4 保护倒换

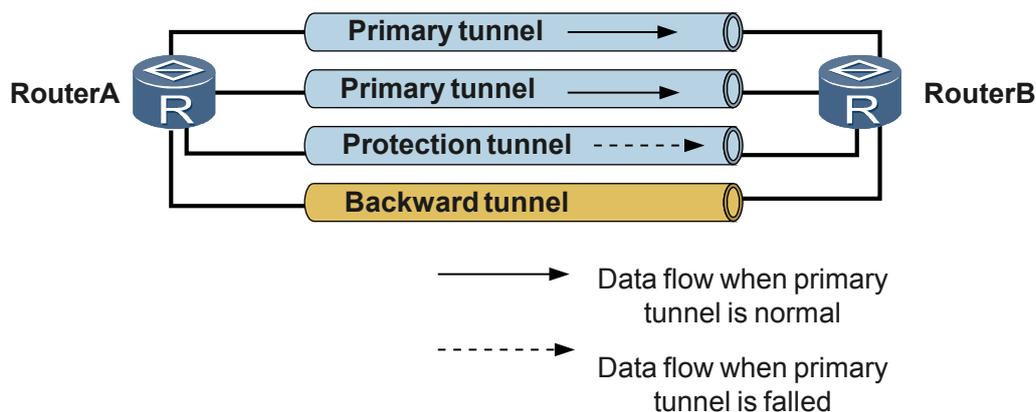
保护倒换 PS（Protection Switching）是为主 Tunnel 预先建立相应的保护 Tunnel（备用 Tunnel）并为其分配带宽，主 Tunnel 和备用 Tunnel 构成一对保护组。当主 Tunnel 发生缺陷时，数据流迅速倒换到备用 Tunnel，减少由于 LSP 失效造成的丢包或时延等问题，从而提高网络可靠性。保护倒换是端到端的保护。

配合 MPLS OAM 的快速缺陷检测，可以使保护倒换达到毫秒级切换。

华为设备实现了 1:1、N:1 两种保护倒换模式：

- 1:1 模式是在 Tunnel 的入节点和出节点间提供主备两条 Tunnel
 - 正常情况下，数据在主 Tunnel 传输。
 - 当入节点通过检测机制发现主 Tunnel 故障时，进行保护倒换，将数据切换到备用 Tunnel 上继续传输。
- N:1 模式是将一条 Tunnel 作为多条主用 Tunnel 的备用 Tunnel，当任何一条主用 Tunnel 故障时，都将数据倒换到共享的备用 Tunnel 上。这种模式主要是为了在网状拓扑结构的网络中节省带宽。

图 4-2 MPLS OAM 隧道保护示意图



4.4 术语与缩略语

术语

术语	解释
后向	被检测 LSP 上流量的反方向。
前向	被检测 LSP 上流量的方向。
Path Merge LSR	MPLS OAM 保护倒换功能中，接收保护路径流量的 LSR。如果该 LSR 不是流量目的地，它将流量合并到工作路径。如果该 LSR 是流量目的地，它将流量交给上层协议处理。
Path Switch LSR	MPLS OAM 保护倒换功能中，在主 LSP 和备 LSP 之间交换或复制流量的 LSR。
用户平面	流量经过的一系列转发组件的集合。OAM CV 或 FFD 报文周期性地插入到转发的流量中，监测转发组件的工作状态。在 IETF 的文档中，用户平面被称为数据平面（data-plane）。
源端	前向 LSP 的 Ingress LSR，也是反向 LSP 的 Egress LSR。
宿端	前向 LSP 的 Egress LSR，也是反向 LSP 的 Ingress LSR。

缩略语

缩略语	英文全称	中文全称
MPLS	Multi-Protocol Label Switching	多协议标记交换
CV	Connectivity Verification	连通性检测
FFD	Fast Failure Detection	快速缺陷检测
FDI	Forward Defect Indication	前向缺陷通告
BDI	Backward Defect Indication	后向缺陷通告
TTSI	Trail Termination Source Identifier	LSP 源宿端连接标识