MPLS 基础

MPLS 简介

多协议标签交换 MPLS（Multiprotocol Label Switching）是一种 IP（Internet Protocol）骨干网技术。MPLS 在无连接的 IP 网络上引入面向连接的标签交换概念，将第三层路由技术和第二层交换技术相结合，充分发挥了 IP 路由的灵活性和二层交换的简捷性。

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）和 CLNP（Connectionless Network Protocol）等。MPLS 中的“Multiprotocol”指的就是支持多种网络协议。

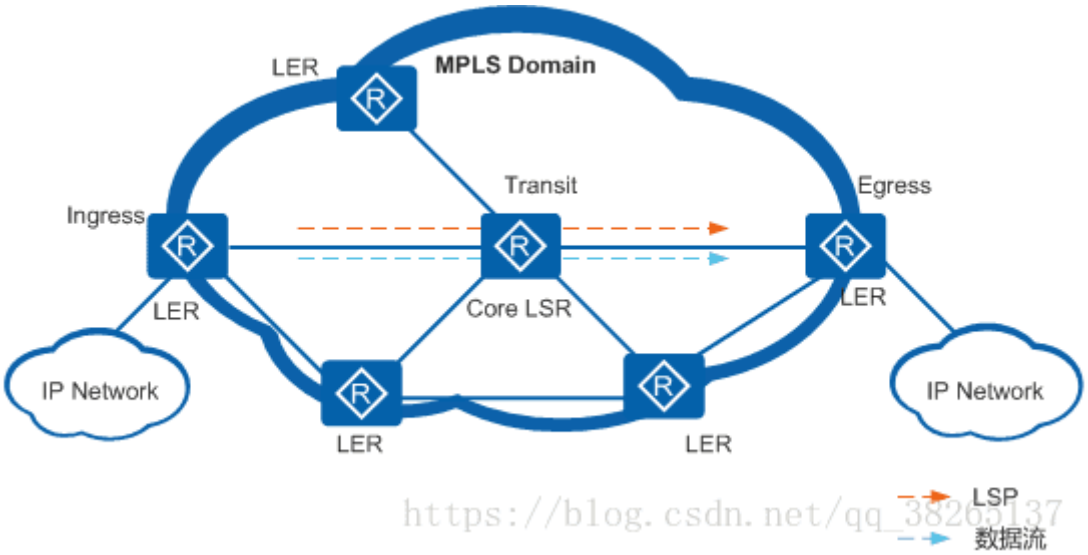
由此可见，MPLS 并不是一种业务或者应用，它实际上是一种隧道技术。这种技术不仅支持多种高层协议与业务，而且在一定程度上可以保证信息传输的安全性。

MPLS 基本结构

网络结构

MPLS 网络的典型结构如下图所示。MPLS 基于标签进行转发，下图中进行 MPLS 标签交换和报文转发的网络设备称为标签交换路由器 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、连接其他网络的 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）。



图：MPLS 网络结构图

IP 报文进入 MPLS 网络时，MPLS 入口的 LER 分析 IP 报文的内容并且为这些 IP 报文添加合适的标签，所有 MPLS 网络中的 LSR 根据标签转发数据。当该 IP 报文离开 MPLS 网络时，标签由出口 LER 弹出。

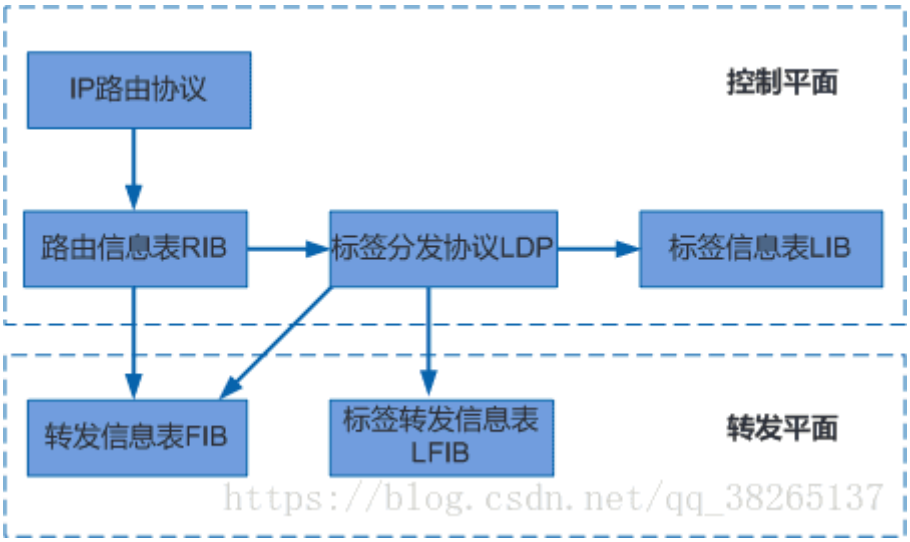
IP 报文在 MPLS 网络中经过的路径称为标签交换路径 LSP（Label Switched Path）。LSP 是一个单向路径，与数据流的方向一致。

如上图，LSP 的入口 LER 称为入节点（Ingress）；位于 LSP 中间的 LSR 称为中间节点（Transit）；LSP 的出口 LER 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一个出节点。

根据 LSP 的方向，MPLS 报文由 Ingress 发往 Egress，则 Ingress 是 Transit 的上游节点，Transit 是 Ingress 的下游节点。同理，Transit 是 Egress 上游节点，Egress 是 Transit 的下游节点。

体系结构：

MPLS 的体系结构如下图所示，它由控制平面（Control Plane）和转发平面（Forwarding Plane）组成。



图：MPLS 体系结构图

- 控制平面：负责产生和维护路由信息以及标签信息。
    - 路由信息表 RIB（Routing Information Base）：由 IP 路由协议（IP Routing Protocol）生成，用于选择路由。
    - 标签分发协议 LDP（Label Distribution Protocol）：负责标签的分配、标签转发信息表的建立、标签交换路径的建立、拆除等工作。
    - 标签信息表 LIB（Label Information Base）：由标签分发协议生成，用于管理标签信息。
- 转发平面：即数据平面（Data Plane），负责普通 IP 报文的转发以及带 MPLS 标签报文的转发。

- 转发信息表 FIB（Forwarding Information Base）：从 RIB 提取必要的路由信息生成，负责普通 IP 报文的转发。
- 标签转发信息表 LFIB（Label Forwarding Information Base）：简称标签转发表，由标签分发协议在 LSR 上建立 LFIB，负责带 MPLS 标签报文的转发。

MPLS 标签

转发等价类：

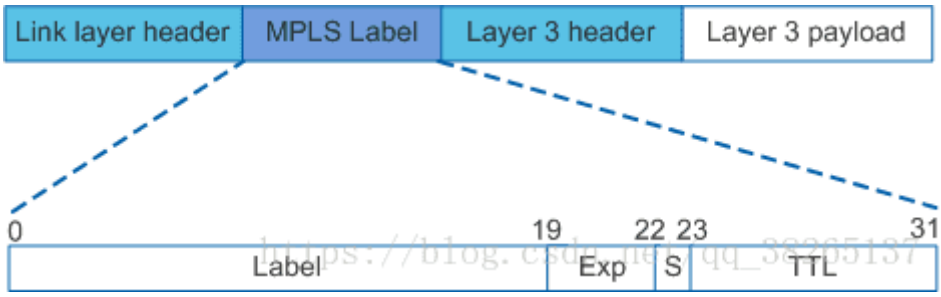
MPLS 将具有相同特征的报文归为一类，称为转发等价类 FEC（Forwarding Equivalence Class）。属于相同 FEC 的报文在转发过程中被 LSR 以相同方式处理。

FEC 可以根据源地址、目的地址、源端口、目的端口、VPN 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，但是一台设备上，一个标签只能代表一个 FEC。

MPLS 报文与普通的 IP 报文相比增加了 MPLS 标签信息，MPLS 标签的长度为 4 个字节。MPLS 标签封装在链路层和网络层之间，可以支持任意的链路层协议。MPLS 标签的封装结构如下图所示：



图：MPLS 标签封装结构

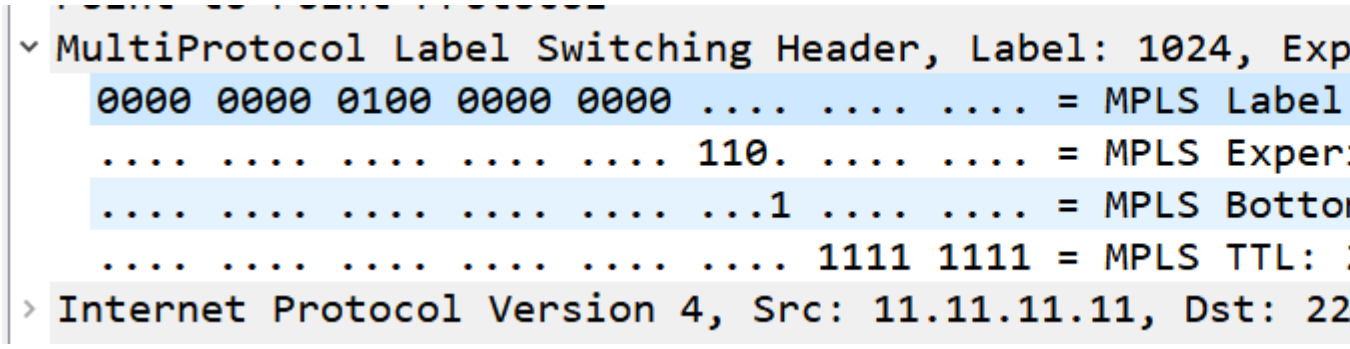
标签共有 4 个字段：

- Label：20bit，标签值域。
- Exp：3bit，用于扩展。现在通常用做 CoS（Class of Service），当设备阻塞时，优先发送优先级高的报文。
- 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 报文中的 TTL（Time To Live）意义相同。
- TTL：8bit，和 IP 报文中的 TTL（Time To Live）意义相同。

标签栈（Label Stack）是指标签的排序集合。靠近二层首部的标签称为栈顶 MPLS 标签或外层 MPLS 标签（Outer MPLS label）；靠近 IP 首部的标签称为栈底 MPLS 标签或内层 MPLS 标签（Inner MPLS label）。理论上，MPLS 标签可以无限嵌套。目前 MPLS 标签嵌套主要应用在 MPLS VPN、TE FRR（Traffic Engineering Fast ReRoute）中。

标签栈按后进先出方式组织标签，从栈顶开始处理标签。

MPLS 报文抓包示例：



图：MPLS 报文抓包示例

标签空间：

标签空间就是指标签的取值范围。标签空间划分如下：

- 0～15：特殊标签。
- 16～1023：静态 LSP 和静态 CR-LSP（Constraint-based Routed Label Switched Path）共享的标签空间。
- 1024 及以上：LDP、RSVP-TE（Resource Reservation Protocol-Traffic Engineering）、MP-BGP（MultiProtocol Border Gateway Protocol）等动态信令协议的标签空间。

特殊标签表：

标签值	含义	描述
0	IPv4 Explicit NULL Label	表示该标签必须被弹出（即标签被剥掉），且报文的转发必须基于 IPv4。如果出倒数第二跳 LSR 需要将值为 0 的标签正常压入报文标签值顶部，转发给最后一跳标签弹出。
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

标签值	含义	描述
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，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 OAM 报文对于 Transit LSR 和倒数第二跳 LSR（penultima
15	保留	-

LSP 的建立

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。LSP 分为静态 LSP 和动态 LSP 两种。

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。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。

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。

配置静态 LSP 时，管理员需要为各 LSR 手工分配标签，需要遵循的原则是：前一节点出标签的值等于下一个节点入标签的值。

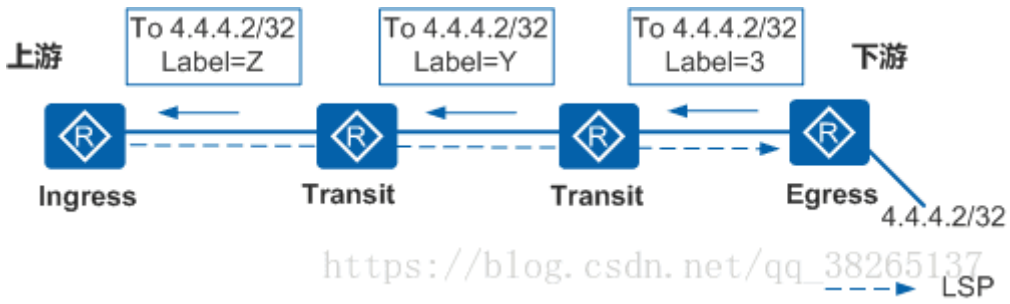
动态 LSP 的建立：

动态 LSP 的标签发布协议

动态 LSP 通过标签发布协议动态建立。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。

动态 LSP 的基本建立过程

标签由下游 LSR 分配，按从下游到上游的方向分发。如下图，由下游 LSR 在 IP 路由表的基础上进行 FEC 的划分，并根据 FEC 分配标签，通告给上游的 LSR，以便建立标签转发表和 LSP。



图：动态 LSP 的基本建立过程

MPLS 转发

MPLS 基本转发过程：

基本概念

在 MPLS 基本转发过程中涉及的相关概念如下：

标签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。

- Push：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）。

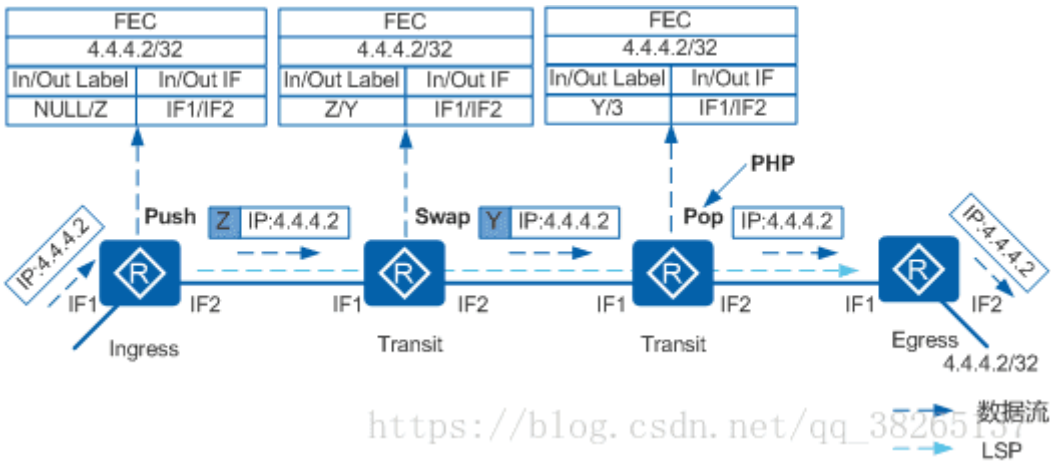
- Swap：当报文在 MPLS 域内转发时，根据标签转发表，用下一跳分配的标签，替换 MPLS 报文的栈顶标签。
- Pop：当报文离开 MPLS 域时，将 MPLS 报文的标签剥掉。

  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。

  默认情况下，设备支持 PHP 特性，支持 PHP 的 Egress 节点分配给倒数第二跳节点的标签值为 3。

基本转发过程：

以支持 PHP 的 LSP 为例，说明 MPLS 基本转发过程。



图：MPLS 基本转发过程

如上图所示，MPLS 标签已分发完成，建立了一条 LSP，其目的地址为 4.4.4.2/32。则 MPLS 基本转发过程如下：

1. Ingress 节点收到目的地址为 4.4.4.2 的 IP 报文，压入标签 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 换成标签 Y。
3. 倒数第二跳 Transit 节点收到带标签 Y 的报文。因为 Egress 分给它的标签值为 3，所以进行 PHP 操作，弹出标签 Y 并转发报文。从倒数第二跳转发给 Egress 的报文以 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 4.4.4.2/32。

MPLS 详细转发过程：

基本概念：

在 MPLS 详细转发过程中涉及的相关概念如下：

- 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，系统自动为隧道分配了一个 ID，也称为 Tunnel ID。该 Tunnel ID 的长度为 32 比特，只是本地有效。

- 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。

FEC 到一组 NHLFE 的映射称为 FTN（FEC-to-NHLFE）。通过查看 FIB 表中 Tunnel ID 值不为 0x0 的表项，能够获得 FTN 的详细信息。FTN 只在 Ingress 存在。

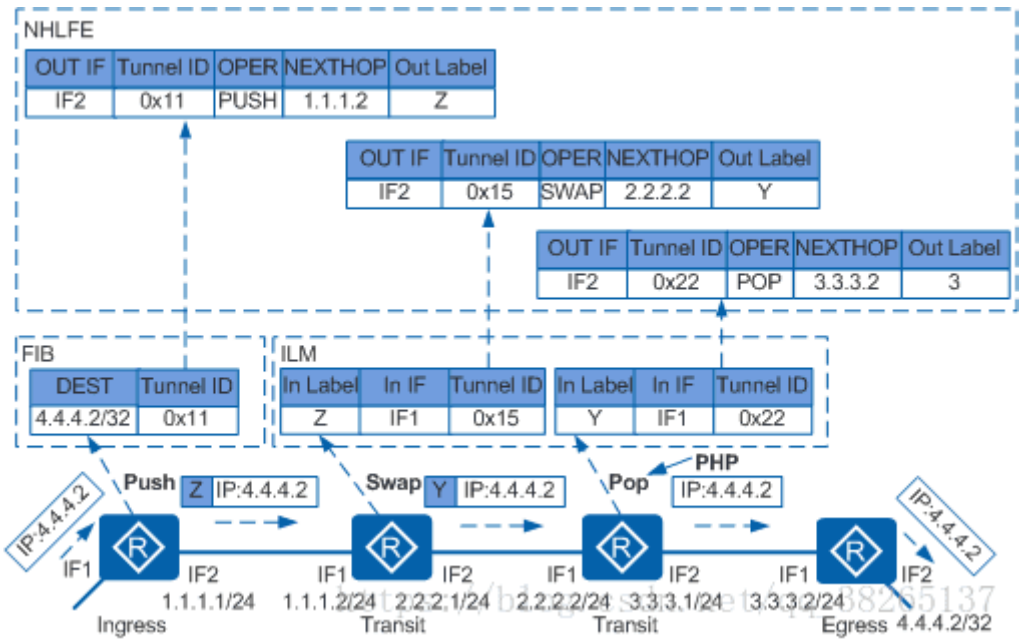
- ILM

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。

详细转发过程：



图：MPLS 详细转发过程

MPLS 的详细转发过程如上图所示：

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。

在 MPLS 转发过程中，FIB、ILM 和 NHLFE 表项是通过 Tunnel ID 关联的。

- Ingress 的处理：通过查询 FIB 表和 NHLFE 表指导报文的转发。
    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。
    4. 在 IP 报文中压入出标签，同时处理 TTL，然后将封装好的 MPLS 报文发送给下一跳。
- Transit 的处理：通过查询 ILM 表和 NHLFE 表指导 MPLS 报文的转发。
    1. 根据 MPLS 的标签值查看对应的 ILM 表，可以得到 Tunnel ID。
    2. 根据 ILM 表的 Tunnel ID 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。

1. 如果标签值> = 16，则用新标签替换 MPLS 报文中的旧标签，同时处理 TTL，然后将替换完标签的 MPLS 报文发送给下一跳。
2. 如果标签值为 3，则直接弹出标签，同时处理 TTL，然后进行 IP 转发或下一层标签转发。

- Egress 的处理：通过查询 ILM 表指导 MPLS 报文的转发或查询路由表指导 IP 报文转发。

  - 如果 Egress 收到 IP 报文，则查看路由表，进行 IP 转发。
  - 如果 Egress 收到 MPLS 报文，则查看 ILM 表获得标签操作类型，同时处理 TTL。
  - 如果标签中的栈底标识 S=1，表明该标签是栈底标签，直接进行 IP 转发。
  - 如果标签中的栈底标识 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 对 TTL 的处理包括 MPLS 对 TTL 的处理模式和 ICMP 响应报文这两个方面。

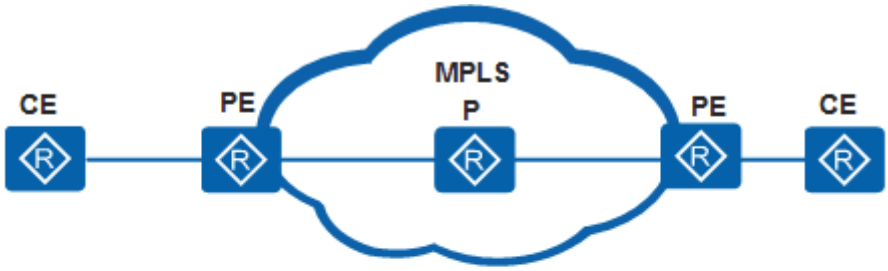
MPLS 对 TTL 的处理模式：

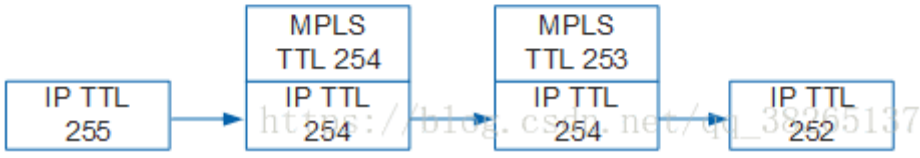
MPLS 标签中包含一个 8 比特的 TTL 字段，其含义与 IP 头中的 TTL 域相同。MPLS 对 TTL 的处理除了用于防止产生路由环路外，也用于实现 Traceroute 功能。

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 Uniform。

- **Uniform 模式**

  IP 报文经过 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。如下图所示：



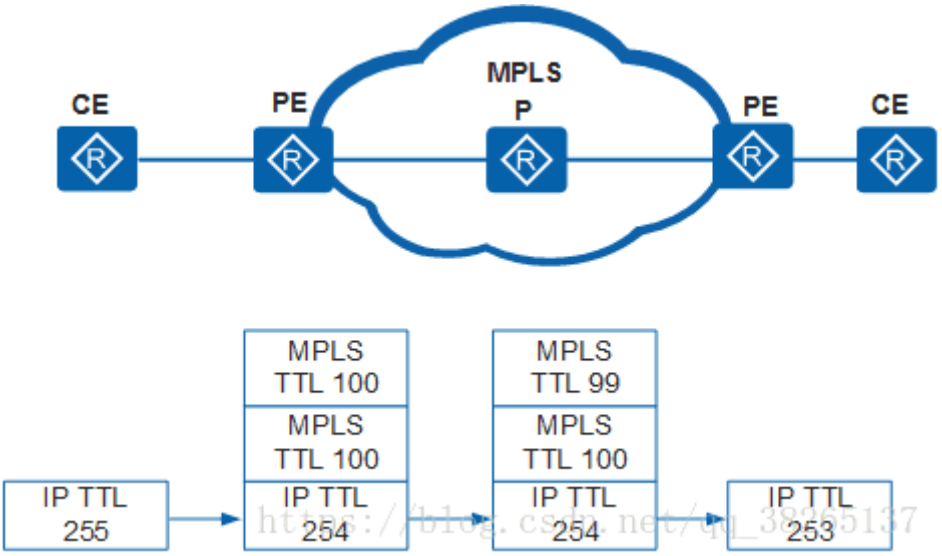


图：Uniform 模式下入方向 TTL 的处理

- **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。如下图所示：



图：Pipe 模式下入方向 TTL 的处理

- 在 MPLS VPN 应用中，出于网络安全的考虑，需要隐藏 MPLS 骨干网络的结构，这种情况下，对于私网报文，Ingress 上使用 Pipe 模式。

ICMP 响应报文：

在 MPLS 网络中，当 LSR 收到 TTL 为 1 的含有标签的 MPLS 报文时，LSR 生成 ICMP 的 TTL 超时消息。

- 如果 LSR 上存在到达报文发送者的路由，则可以通过 IP 路由，直接向发送者回应 TTL 超时消息。
- 如果 LSR 上不存在到达报文发送者的路由，则 ICMP 响应报文将按照 LSP 继续传送，到达 LSP 出节点后，由 Egress 节点将该消息返回给发送者。

通常情况下，收到的 MPLS 报文只带一层标签时，LSR 可以采用第一种方式回应 TTL 超时消息；收到的 MPLS 报文包含多层标签时，LSR 采用第二种方式回应 TTL 超时消息。

但是，在 MPLS VPN 中，ASBR（Autonomous System Boundary Router，自治系统边界路由器）和 HoVPN 组网应用中的 SPE（Superstratum PE or Sevice Provider-end PE，上层 PE 或运营商侧 PE），接收到的承载 VPN 报文的 MPLS 报文可能只有一层标签，此时，这些设备上并不存在到达报文发送者的路由，则采用第二种方法回应 TTL 超时消息。

LSP 连通性检测

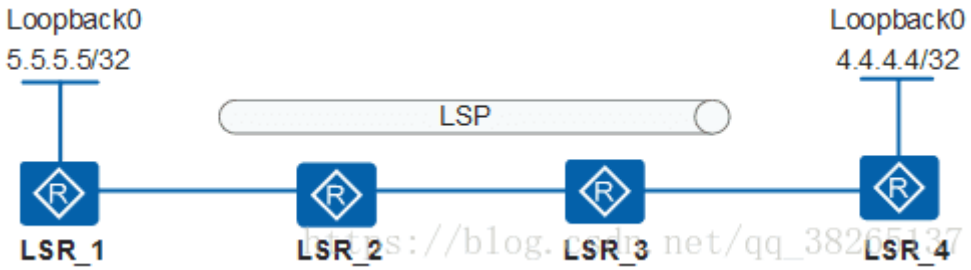
在 MPLS 网络中，如果通过 LSP 转发数据失败，负责建立 LSP 的 MPLS 控制平面将无法检测到这种错误，加大了网络维护的难度。MPLS Ping/MPLS Tracert 为用户提供了发现 LSP 错误、并及时定位失效节点的机制。

MPLS Ping 主要用于检查 LSP 的连通性。MPLS Tracert 在检查 LSP 的连通性的同时，还可以分析网络什么地方发生了故障。类似于普通 IP 的 Ping/Tracert，MPLS Ping/MPLS Tracert 使用 MPLS 回显请求（Echo Request）报文和 MPLS 回显应答（Echo Reply）报文检测 LSP 的可用性。这两种消息都以 UDP 报文格式发送，其中 Echo Request 的 UDP 端口号为 3503，该端口号只有使能 MPLS 功能的设备才能识别。

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 报文通过

MPLS 转发给目的端，而 MPLS Echo Reply 报文则通过 IP 转发给源端。另外为了防止 LSP 断路时 Echo Request 进行 IP 转发，从而保证 LSP 的连通性测试，将 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 值为 1。

MPLS Ping：



图：MPLS 网络

如上图，LSR_1 上建立了一条目的地为 LSR_4 的 LSP。从 LSR_1 对该 LSP 进行 MPLS Ping 时的处理如下：

1. LSR_1 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. LSR_1 构造 MPLS Echo Request 报文，IP 头中的目的地址为 127.0.0.1/8，IP 头中的 TTL 值为 1，同时将 4.4.4.4 填入 Echo Request 报文中的目的 FEC 中。然后查找相应的 LSP，压入 LSP 的标签，将报文发送给 LSR_2。
3. 中间节点 LSR_2 和 LSR_3 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 MPLS Echo Reply 报文。

4. 当 MPLS 转发路径无故障，则 MPLS Echo Request 报文到达 LSP 的出节点 LSR_4。然后检查目的 FEC 中包含的目的地址 4.4.4.4 是否为自己的 Loopback 接口地址，以此来确认 LSR_4 是该 FEC 的真正出口后，返回正确的 MPLS Echo Reply 报文。至此整个 MPLS Ping 过程结束。

MPLS Tracert:

从 LSR_1 对 4.4.4.4/32 进行 MPLS Tracert 时的处理如下：

1. LSR_1 检查 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 Tracert，否则继续进行如下处理。
2. LSR_1 构造 MPLS Echo Request 报文，IP 头中的目的地址为 127.0.0.1/8，同时将 4.4.4.4 填入 MPLS Echo Request 报文中的目的 FEC 中，然后查找相应的 LSP，压入 LSP 的标签并且将 MPLS TTL 设置为 1，将报文发送给 LSR_2。此 MPLS Echo Request 报文中包含 Downstream Mapping TLV（用来携带 LSP 在当前节点的下游信息，主要包括下一跳地址、出标签等）。
3. LSR_2 收到 LSR_1 发送来的报文后，将 MPLS Echo Request 中 MPLS TTL 减 1 为 0 后发现 TTL 超时，然后 LSR_2 需要检查是否存在该 LSP，同时检查报文中 Downstream Mapping TLV 的下一跳地址、出标签是否正确，如果两项检查都正确，返回正确的 MPLS Echo Reply 报文，并且报文中必须携带 LSR_2 本身的包含下一跳和出标签的 Downstream

Mapping TLV 给 LSR_1。如果检查有不正确，则返回错误的 MPLS Echo Reply 报文。

4. LSR_1 收到正确的 MPLS Echo Reply 报文后再次发送 MPLS Echo Request 报文，报文的封装方式跟步骤 2 类似，只是将 LSP 标签的 MPLS TTL 设置为 2，此时的 MPLS Echo Request 报文中的 Downstream Mapping TLV 是从 MPLS Echo Reply 报文中复制过来的。然后 LSR_2 收到该报文后进行普通 MPLS 转发。LSR_3 收到此报文，标签的 TTL 超时，跟步骤 3 同样的处理方式后返回 MPLS Echo Reply 报文。
5. LSR_1 收到正确的 MPLS Echo Reply 报文后重复步骤 4 把 LSP 标签的 MPLS TTL 设置为 3，复制 Downstream Mapping TLV 后发送 MPLS Echo Request 报文。LSR_2 和 LSR_3 对该报文进行普通 MPLS 转发。LSR_4 收到此报文，重复步骤 3 处理方式对报文进行处理，同时检查目的 FEC 中包含的目的 IP 4.4.4.4 为自己的 Loopback 接口地址，以此来发现已经是该 LSP 的出节点，因此返回不带下游信息的 MPLS Echo Reply 报文，至此整个 MPLS Tracert 过程结束。

通过上述步骤中返回携带下游信息的 MPLS Echo Reply 报文，在 LSR_1 上就获取了该 LSP 沿途每一个节点信息。