MPLS 协议原理与配置

MPLS （ Multi-Protocol Label Switching, 多协议标签交换 ）

LDP 的 hello 时间为 5s ，holdtime 15s , 发送的组播地址为 224.0.0.2 ，端口号为 UDP 646

57	42.078000	2.2.2.2	1.1.1.1	LDP	Keep Alive Message
58	42.250000	1.1.1.1	2.2.2.2	TCP	ldp > 63848 [ACK] Seq=157 Ack=175
59	42.250000	1.1.1.1	2.2.2.2	LDP	Keep Alive Message
60	42.297000	2.2.2.2	1.1.1.1	TCP	63848 > ldp [ACK] Seq=175 Ack=175
61	42.390000	192.168.12.2	224.0.0.5	OSPF	Hello Packet
62	42.859000	192.168.12.1	224.0.0.2	LDP	Hello Message
63	45.203000	192.168.12.2	224.0.0.2	LDP	Hello Message
64	47.906000	192.168.12.1	224.0.0.2	LDP	Hello Message
65	50.203000	192.168.12.2	224.0.0.2	LDP	Hello Message

9 6.563000 192.168.12.2 224.0.0.2 LDP Hello Message

```
Frame 9: 76 bytes on wire (608 bits), 76 bytes captured (608 bits)
Ethernet II, Src: HuaweiTe_a4:59:82 (54:89:98:a4:59:82), Dst: IPv4mcast_00:00:02 (01:
Internet Protocol, Src: 192.168.12.2 (192.168.12.2), Dst: 224.0.0.2 (224.0.0.2)
User Datagram Protocol, Src Port: ldp (646), Dst Port: ldp (646)
Label Distribution Protocol
    Version: 1
    PDU Length: 30
    LSR ID: 2.2.2.2 (2.2.2.2)
    Label Space ID: 0
  Hello Message
      0... .... = U bit: Unknown bit not set
      Message Type: Hello Message (0x100)
      Message Length: 20
      Message ID: 0x0000015c
    Common Hello Parameters TLV
        00.. .... = TLV Unknown bits: Known TLV, do not Forward (0x00)
        TLV Type: Common Hello Parameters TLV (0x400)
        TLV Length: 4
        Hold Time: 15
        0... .... .... .... = Targeted Hello: Link Hello
        .0.. .... .... .... = Hello Requested: Source does not request periodic hellos
        ..00 0000 0000 0000 = Reserved: 0x0000
    IPv4 Transport Address TLV
        00.. .... = TLV Unknown bits: Known TLV, do not Forward (0x00)
        TLV Type: IPv4 Transport Address TLV (0x401)
        TLV Length: 4
        IPv4 Transport Address: 2.2.2.2 (2.2.2.2)
```

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邻居的存在，如 Hello 消息。
会话（Session）消息：用于建立、维护和终止 LDP 对等体之间的会话，如 Initialization 消息、Keepalive 消息。
通告（Advertisement）消息：用于创建、改变和删除 FEC 的标签映射，如 Address 消息、Label Mapping 消息。
通知（Notification）消息：用于提供建议性的消息和差错通知。

0～15：特殊标签。如标签 3，称为隐式空标签，用于倒数第二跳弹出；
16～1023：静态 LSP
1024 及以上：LDP、MP-BGP

==========
标签管理

标签发布方式 ：DU（Downstream Unsolicited，下游自主方式）
Label Advertisement Mode
标签的分配控制方式：有序标签分配控制方式　　标签的保持方式：自由标签保持方式
Label Distribution Mode : Ordered
Label Retention Mode : Liberal

标签的发布方式：华为采用 DU 下游自主
DU（Downstream Unsolicited，下游自主方式）：
对于一个到达同一目地址报文的分组，LSR 无需从上游获得标签请求消息即可进行标签分配与分发。
DoD（Downstream on Demand，下游按需方式）：
对于一个到达同一目的地址报文的分组，LSR 获得标签请求

消息之后才进行标签分配与分发。

标签的分配控制方式：华为采用 Ordered 有序标签分配控制
Independent（独立标签分配控制方式）：本地 LSR 可以自主地分配一个标签绑定到某个 IP 分组，并通告给上游 LSR，而无需等待下游的标签。
Ordered（有序标签分配控制方式）：只有当该 LSR 已经具有此 IP 分组的下一跳的标签，或者该 LSR 就是该 IP 分组的出节点时，该 LSR 才可以向上游发送此 IP 分组的标签。

标签的保持方式：华为采用　Liberal 自由标签保持
Liberal（自由标签保持方式）：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。
Conservative（保守标签保持方式）：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。

0-15 保留标签的含义

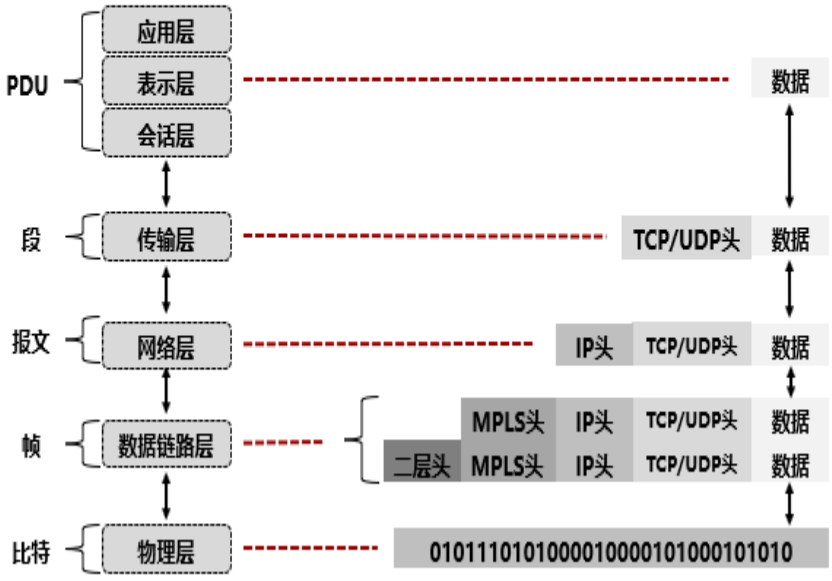
表1 特殊标签

标签值	含义	描述
0	IPv4 Explicit NULL Label	表示该标签必须被弹出（即标签被剥掉），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	-

前言

- 90年代初，互联网流量快速增长，而由于当时硬件技术的限制，路由器采用最长匹配算法逐跳转发数据包，成为网络数据转发的瓶颈。快速路由技术成为当时研究的一个热点。
- 在各种方案中，IETF确定MPLS协议作为标准的协议。MPLS采用短而定长的标签进行数据转发，大大提高了硬件限制下的转发能力；而且MPLS可以扩展到多种网络协议（如IPv6，IPX等）。

MPLS概述

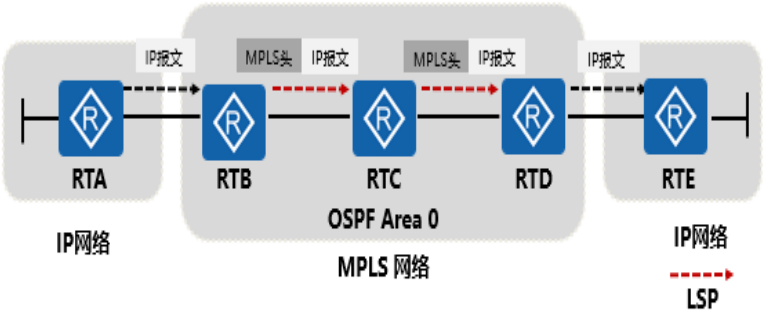


- MPLS 协议从各种链路层协议（如 PPP、ATM、帧中继、以太网等）得到链路层服务，又为网络层提供面向连接的服务。MPLS 能从 IP 路由协议和控制协议中得到支持，路由功能强

大、灵活，可以满足各种新应用对网络的要求。

- 我们将分为以下两部分讲解 MPLS 协议：
- MPLS 的基本结构；
- MPLS LSP 的建立过程和数据的转发过程。

MPLS基本网络结构



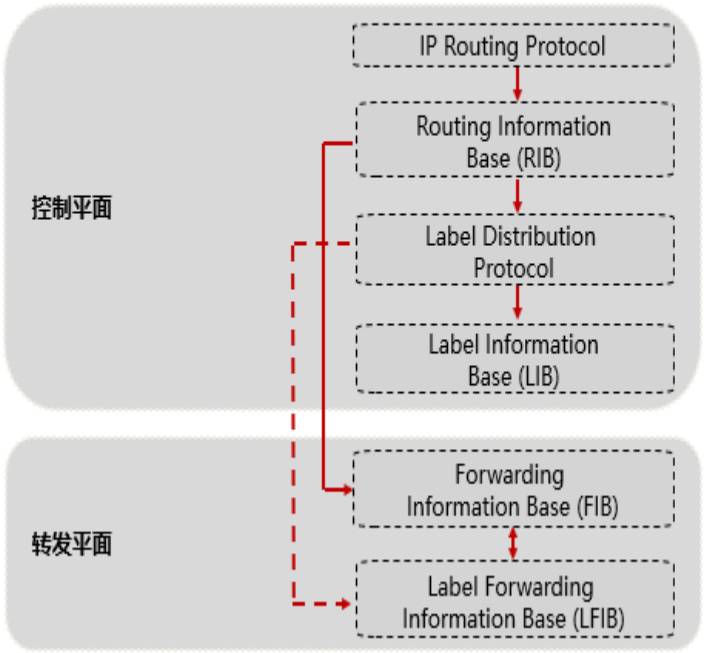
- LSP（Label Switched Path）：标签交换路径，即到达同一目的地址的报文在MPLS网络中经过的路径。
- FEC（Forwarding Equivalent Class）：一般指具有相同转发处理方式的报文。在MPLS网络中，到达同一目的地址的所有报文就是一个FEC。

- 在 MPLS 网络中，路由器的角色分为两种：
- LER（Label Edge Router）：在 MPLS 网络中，用于标签的压入或弹出，如上图中的 RTB，RTD。
- LSR（Label Switched Router）：在 MPLS 网络中，用于标签的交换，如图中的 RTC。
- 根据数据流的方向，LSP 的入口 LER 被称为入节点（Ingress）；位于 LSP 中间的 LSR 被称为中间节点（Transit）；LSP 的出口 LER 被称为出节点（Egress）。
- MPLS 报文由 Ingress 发往 Transit，则 Ingress 是 Transit 的上游节点，Transit 是 Ingress 的下游节点；同理，Transit

是 Egress 的上游节点，Egress 是 Transit 的下游节点。

- MPLS 作为一种分类转发技术，将具有相同转发处理方式的报文分为一类，称该类报文为一个 FEC（Forwarding Equivalent Class）。
- FEC 的划分方式非常灵活，可以是以源地址、目的地址、源端口、目的端口、协议类型或 VPN 等为划分依据的任意组合。

MPLS体系结构

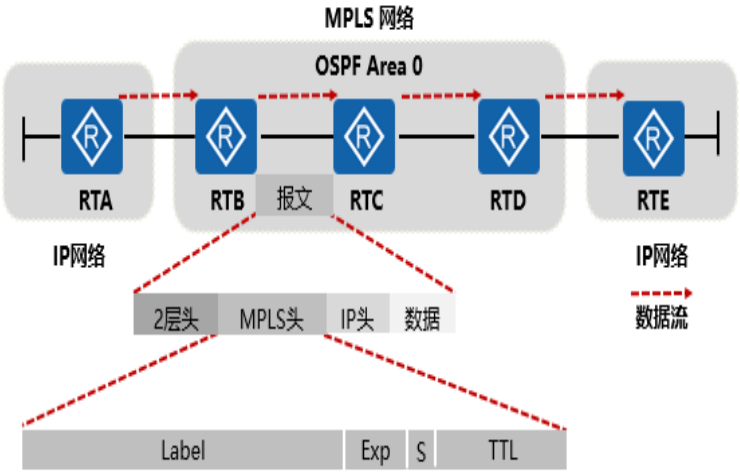


- 控制平面：负责产生和维护路由信息以及标签信息。
- 路由信息表 RIB（Routing Information Base）：由 IP 路由协议（IP Routing Protocol）生成，用于选择路由。
- 标签分发协议 LDP（Label Distribution Protocol）：负责标签的分配、标签转发信息表的建立、标签交换路径的建立、拆除等工作。
- 标签信息表 LIB（Label Information Base）：由标签分

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，由标签分发协议建立 LFIB，负责带 MPLS 标签报文的转发。
- MPLS 路由器上，报文的转发过程：
- 当收到普通 IP 报文时，查找 FIB 表，如果 Tunnel ID 为 0x0，则进行普通 IP 转发；如果查找 FIB 表，Tunnel ID 为非 0x0，则进行 MPLS 转发。
- 当收到带标签的报文时，查找 LFIB 表，如果对应的出标签是普通标签，则进行 MPLS 转发；查找 LFIB 表，如果对应的出标签是特殊标签，如标签 3，则将报文的标签去掉，进行 IP 转发。

MPLS数据报文结构



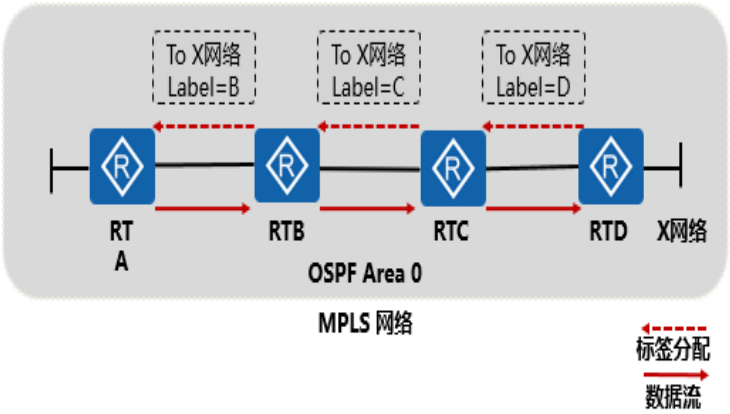
- 标签（Label）是一个短而定长的、只有本地意义的标识，用于唯一标识去往同一目的地址的报文分组。

- 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 标签的封装结构如图所示。
- MPLS 标签的长度为 4 个字节，共分 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），当设备发生阻塞时，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。
- 标签空间是指标签的取值范围。标签空间划分如下：
- 0～15：特殊标签。如标签 3，称为隐式空标签，用于倒数第二跳弹出；
- 16～1023：静态 LSP 和静态 CR-LSP（Constraint-base

d Routed Label Switched Path）共享的标签空间；

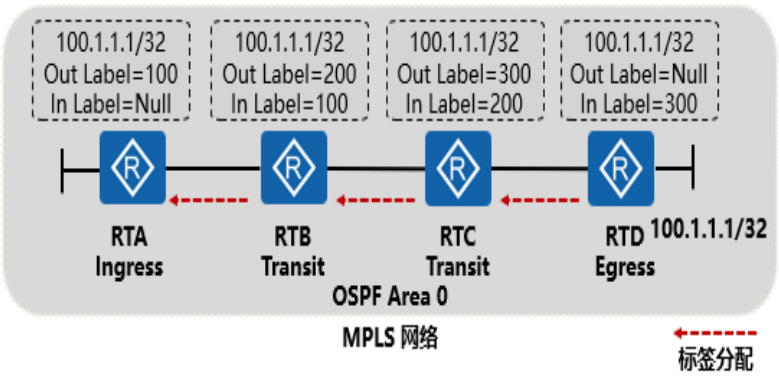
- 1024 及以上：LDP、RSVP-TE（Resource Reservation Protocol-Traffic Engineering）、MP-BGP（MultiProtocol Border Gateway Protocol）等动态信令协议的标签空间。

LSP建立方式



- 建立LSP的方式有两种：
  - 静态LSP：用户通过手工方式为各个转发等价类分配标签建立转发隧道；
  - 动态LSP：通过标签发布协议动态建立转发隧道。

静态LSP

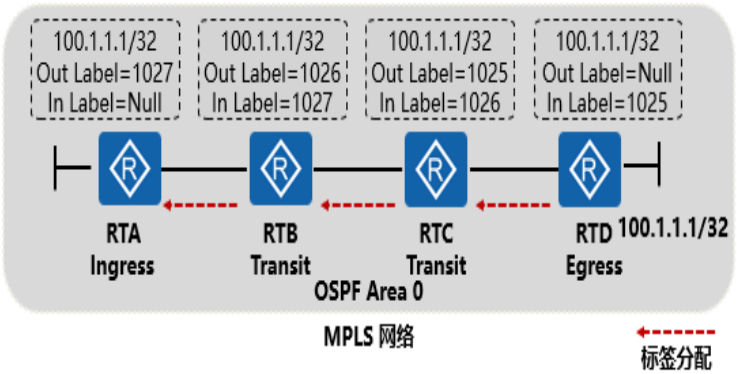


- 静态LSP的特点:
  - 不使用标签发布协议，不需要交互控制报文，资源消耗比较小;
  - 通过静态方式建立的LSP不能根据网络拓扑变化动态调整，需要管理员干预。
- 静态LSP适用于拓扑结构简单并且稳定的网络。

- 配置静态 LSP 时，管理员需要为各路由器手工分配标签，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。
- 如图所示拓扑，MPLS 网络中有一个 100.1.1.1/32 的用户，静态为该路由建立一条 LSP，配置过程如下：
- 配置 LSR ID 用来在网络中唯一标识一个 MPLS 路由器。缺省没有配置 LSR ID，必须手工配置。为了提高网络的可靠性，推荐使用 LSR 某个 Loopback 接口的地址作为 LSR ID。
- 配置命令：mpls lsr-id lsr-id
- 在 MPLS 域的所有节点与相应的接口上开启 MPLS 协议。
- 配置命令：system-view
- mpls
- interface interface-type interface-number

- mpls
- 在 Ingress 进行以下配置：
- static-lsp ingress lsp-name destination ip-address { mask-length | mask } { nexthop next-hop-address | outgoing-interface interface-type interface-number } * out-label out-label。

动态LSP

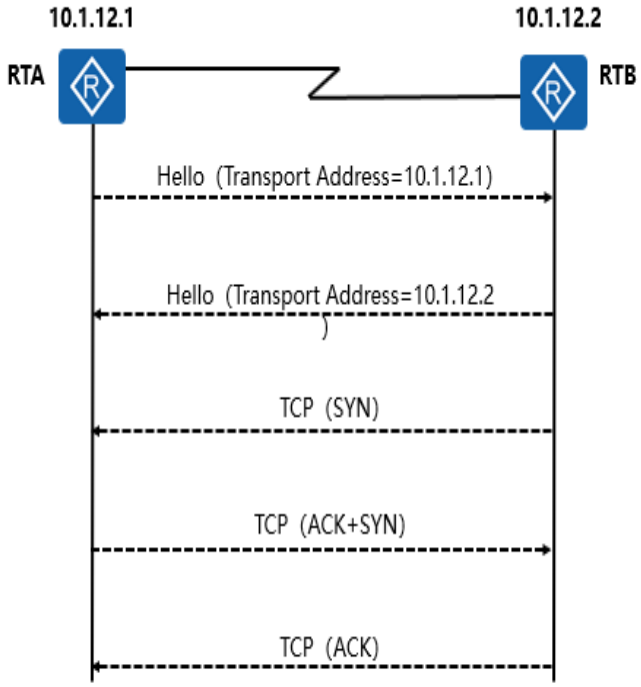


- 动态LSP通过LDP协议实现对FEC的分类、标签的分配及LSP的建立和维护等操作。
- 动态LSP的特点：
  - 组网配置简单，易于管理和维护；
  - 支持基于路由动态建立LSP，网络拓扑发生变化时，能及时反映网络状况。

- 如图所示拓扑：
- Egress 路由器 RTD 为本地存在的路由分配标签，并将路由和标签的绑定关系主动发送给上游邻居路由器 RTC；
- 路由器 RTC 收到下游邻居路由器 RTD 的路由和标签的绑定关系后，将其记录到 LIB 中，并将自己分配的标签和路由的绑定关系发送给上游邻居路由器 RTB；
- RTB 执行相同的动作将标签和路由的绑定关系发送给上游邻居路由器 RTA，RTA 为 Ingress 路由器，没有上游邻居，

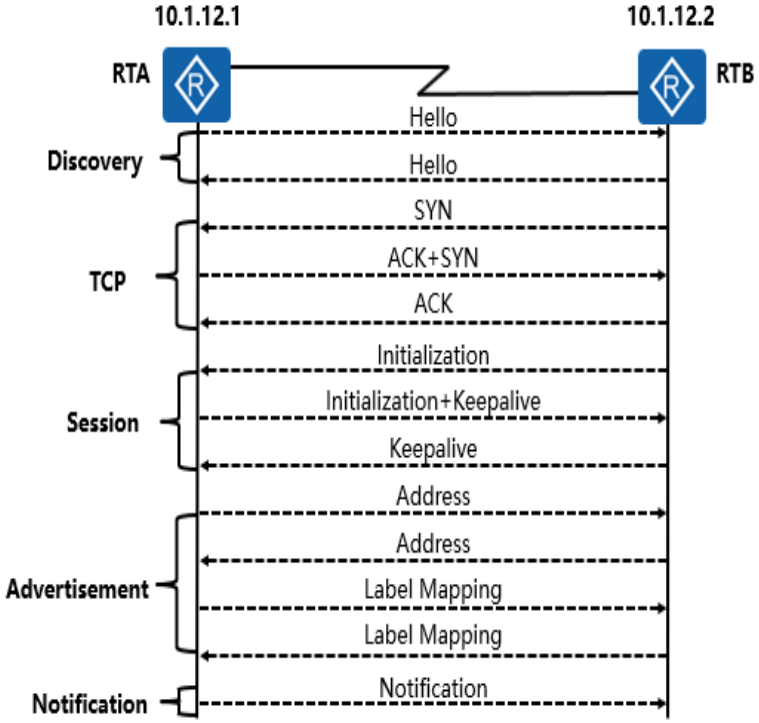
因此动态的 LSP 完成建立。

LDP邻居发现



- MPLS 路由器通过周期性地发送 LDP 链路 Hello 消息（LDP Link Hello），实现 LDP 邻居的发现，并建立本地 LDP 会话。
- 为了能使开启 LDP 协议的设备快速发现邻居，LDP 的 Hello 消息使用 UDP 封装。UDP 是无连接的协议，为了保证邻居的有效性和可靠性，Hello 消息周期发送，发送周期为 5s，使用组播 224.0.0.2 作为目的 IP 地址，意思是“发送给网络中的所有路由器”。
- LDP 的 Hello 消息中，携带有 Transport Address 字段，该字段与设备配置的 LSR ID 一致，表明与对端建立邻居关系时所使用的 IP 地址。如果该字段 IP 地址是直连接口 IP 地址，则直接建立邻居关系；如果该字段地址是 LoopBack 接口 IP 地址，保证该接口 IP 地址路由可达，才能建立邻居关系。

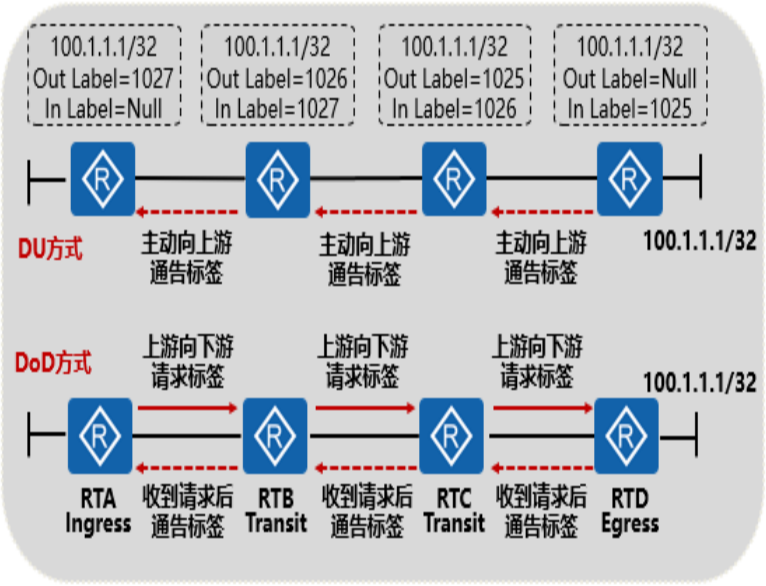
LDP邻居建立



- 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邻居的存在，如 Hello 消息。
- 会话（Session）消息：用于建立、维护和终止 LDP 对等体之间的会话，如 Initialization 消息、Keepalive 消息。
- 通告（Advertisement）消息：用于创建、改变和删除 FEC 的标签映射，如 Address 消息、Label Mapping 消息。
- 通知（Notification）消息：用于提供建议性的消息和差错通知。
- LDP 邻居建立过程如图所示：
- 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。
- 传输地址较大的一方作为主动方，发起 TCP 连接。

- 如图所示，RTB 作为主动方发起 TCP 连接，RTA 作为被动方等待对方发起连接。
- TCP 连接建立成功后，由主动方 RTB 发送初始化消息，协商建立 LDP 会话的相关参数。
- 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。

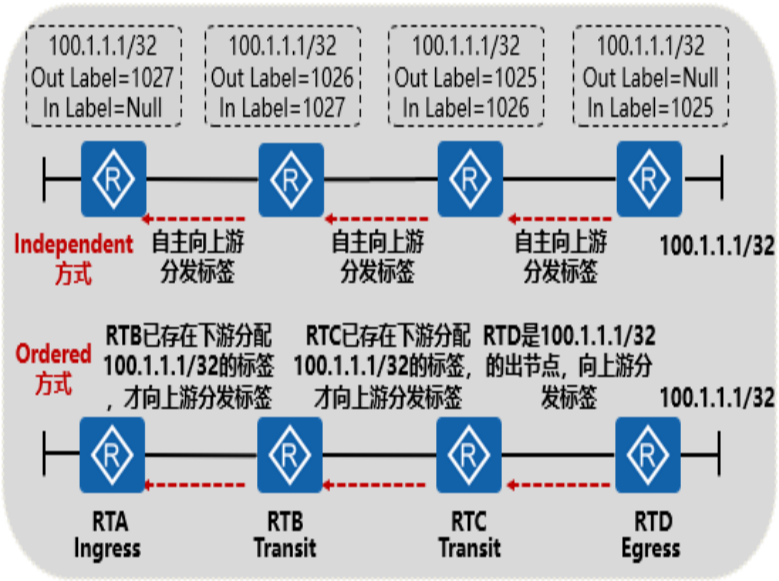
标签的发布方式



- 标签的发布方式：
- DU（Downstream Unsolicited，下游自主方式）：对于一个到达同一目地址报文的分组，LSR 无需从上游获得标签请求消息即可进行标签分配与分发。
- DoD（Downstream on Demand，下游按需方式）：对于一个到达同一目的地址报文的分组，LSR 获得标签请求消息之后才进行标签分配与分发。
- 如图所示拓扑：

- 采用 DU 方式分发标签，对于目的地址为 100.1.1.1/32 的分组，下游 RTD（Egress）通过标签映射消息主动向上游 RTC（Transit）通告自己主机路由 100.1.1.1/32 的标签。
- 采用 DoD 方式分发标签，对于目的地址为 100.1.1.1/32 的分组，上游 RTC（Transit）向下游发送标签请求消息，下游 RTD（Egress）收到标签请求消息后，才会向上游发送标签映射消息。
- 华为设备默认采用 DU 的方式发布标签。
- DU 无需等待上游的请求消息，可以直接向邻居分配标签。在网络拓扑发生变化时，采用 DU 方式可以快速反应为新的拓扑分发标签，收敛时间相对于 DoD 方式较短。

标签的分配控制方式

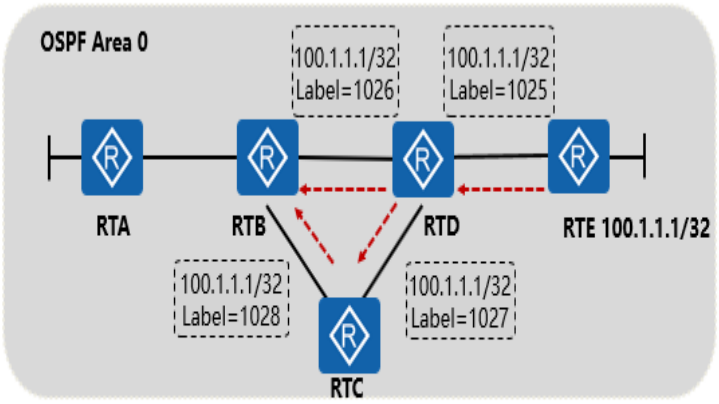


- 标签的分配控制方式：
- Independent（独立标签分配控制方式）：本地 LSR 可以自主地分配一个标签绑定到某个 IP 分组，并通告给上游 LS

R，而无需等待下游的标签。

- Ordered（有序标签分配控制方式）：只有当该 LSR 已经具有此 IP 分组的下一跳的标签，或者该 LSR 就是该 IP 分组的出节点时，该 LSR 才可以向上游发送此 IP 分组的标签。
- 如图所示拓扑：
- 采用 Independent 方式：
- 如果标签发布方式为 DU，且标签分配控制方式为 Independent，则 RTC（Transit）无需等待下游 RTD（Egress）的标签，就会直接向上游 RTB 分发标签。
- 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 RTB（Transit）的直连下游 RTC（Transit）会直接回应标签，而不必等待来自下游 RTD（Egress）的标签。
- 采用 Ordered 方式：
- 如果标签发布方式为 DU，且标签分配控制方式为 Ordered，则 RTC（Transit）只有收到下游 RTD（Egress）的标签，才会向上游 RTB 分发标签。
- 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 RTB（Transit）的直连下游 RTC（Transit）只有收到下游 RTD（Egress）的标签，才会向上游 RTB 分发标签。

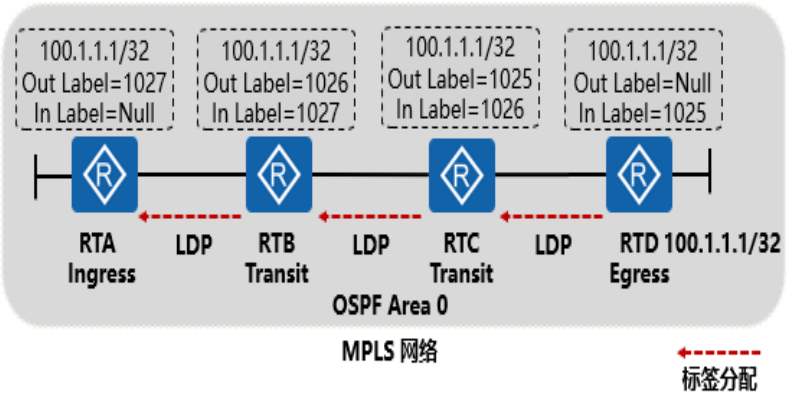
标签的保持方式



- 路由表中，RTB通过RTD到达100.1.1.1/32的路径最优，RTB从RTC收到分配给100.1.1.1/32的标签处理方式有以下两种:
- 一是RTB保留从RTC收到的标签信息，二是RTB不保留从RTC收到的标签信息，前者称为Liberal方式，后者称为Conservative方式。

- 标签的保持方式：
- Liberal（自由标签保持方式）：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。
- Conservative（保守标签保持方式）：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，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。
- 华为设备默认采用自由标签保持方式保存标签。

LDP建立LSP过程

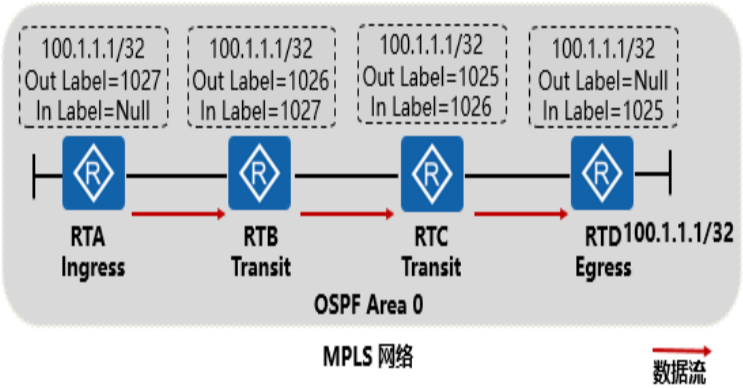


- IGP协议负责实现MPLS网络内路由可达，为FEC的分组提供路由；
- LDP协议负责实现对FEC的分类、标签的分配以及LSP的建立和维护等操作。

- 如图所示拓扑，LDP 动态建立 LSP 的过程如下：
- RTD 上存在 100.1.1.1/32 的主机路由，因为 RTD 是 Egress 节点，所以直接向自己上游邻居 RTC 发布 100.1.1.1/32 与标签的绑定关系；
- RTC 收到下游邻居 RTD 分配的 100.1.1.1/32 与标签的绑定关系后，将标签记录在自己的 LIB 表中，并向上游邻居 RTB 发布 100.1.1.1/32 与标签的绑定关系，同时 RTC 查看自己 IP 路由表中到达 100.1.1.1/32 的下一跳是否为 RTD，如果 IP 路由表中的下一跳为 RTD，则 RTC 使用 RTD 分配的标签封装到达 100.1.1.1/32 的数据；如果 IP 路由表中的下一跳不是 RTD，则 RTC 保留 RTD 分配的标签作为备用标签；
- RTB 收到下游邻居 RTC 分配的 100.1.1.1/32 与标签的绑定关系后，执行与 RTC 相同的动作；
- RTA 收到下游邻居 RTB 分配的 100.1.1.1/32 与标签的绑

定关系后，查看自己 IP 路由表中到达 100.1.1.1/32 的下一跳是否为 RTB，如果 IP 路由表中的下一跳为 RTB，则 RTA 使用 RTB 分配的标签封装到达 100.1.1.1/32 的数据；如果 IP 路由表中的下一跳不是 RTB，则 RTA 保留 RTB 分配的标签作为备用。因为 RTA 为 Ingress，最终到达 100.1.1.1/32 的 LSP 完成建立。

MPLS数据转发过程

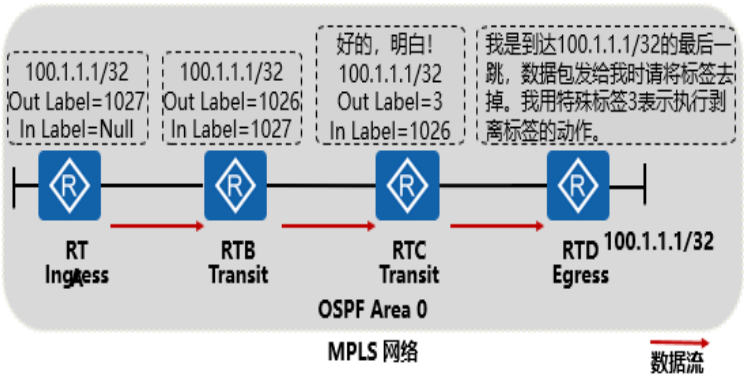


- 在MPLS网络中，数据包在每台路由器上根据已分配的标签进行标签的封装和转发；
- 分析数据包到达Egress节点RTD上怎么处理？如果MPLS网络中业务量很大，Egress节点的处理方式有何不妥？

- 如图所示拓扑，MPLS 数据转发过程如下：
- RTA 上收到访问 100.1.1.1/32 的数据包，如果数据包为普通的 IP 报文，则查找 FIB 表，因为 Tunnel ID 为非 0x0，封装已分配的标签 1027 进行 MPLS 转发；如果数据包为带标签的报文，查找 LFIB 表，封装已分配的标签 1027 进行 MPLS 转发；
- RTB 收到 RTA 发送的带标签 1027 的报文，查找 LFIB 表，封装已分配的出标签 1026 进行 MPLS 转发给 RTC；

- RTC 收到 RTB 发送的带标签 1026 的报文，查找 LFIB 表，封装已分配的出标签 1025 进行 MPLS 转发给 RTD；
- RTD 收到 RTC 发送的带标签 1025 的报文，查找 LFIB 表，出标签为 Null，表明数据包已经到达 Egress 节点，所以路由器将数据包的标签信息去掉，并对数据包进行三层处理，查找 IP 路由表发现 100.1.1.1/32 的路由为自己本地的路由，根据 IP 路由表中的出接口进行 IP 数据的封装并转发。
- 如果 MPLS 网络中的业务量很大，则每次数据包在 Egress 节点都要进行两次处理才能进行正确的路由转发，这样会导致 Egress 节点的处理压力增加，路由器的处理性能降低。我们希望在 Egress 节点上只处理一次就能将数据包正确转发，以提高 Egress 的转发性能，所以提出了 PHP 技术。

Penultimate Hop Popping



- PHP（Penultimate Hop Popping，倒数第二跳弹出），具体过程如下：
  - RTC收到RTB发送的带标签1026的报文，查找LFIB表，发现分配的出标签为隐式空标签3，于是执行弹出标签的动作，并将IP数据包转发给下游路由器RTD；
  - RTD收到RTC发送的IP报文，直接查找自己的FIB表，根据FIB表中的出接口进行IP数据的封装并转发。

思考题

1. MPLS的标签（Label）字段有多少位？（　）

   A. 10

   B. 20

   C. 30

   D. 40

2. 隐式空标签的标签值是多少？（　）

   A. 3

   B. 5

   C. 0

答案：B。
答案：A。