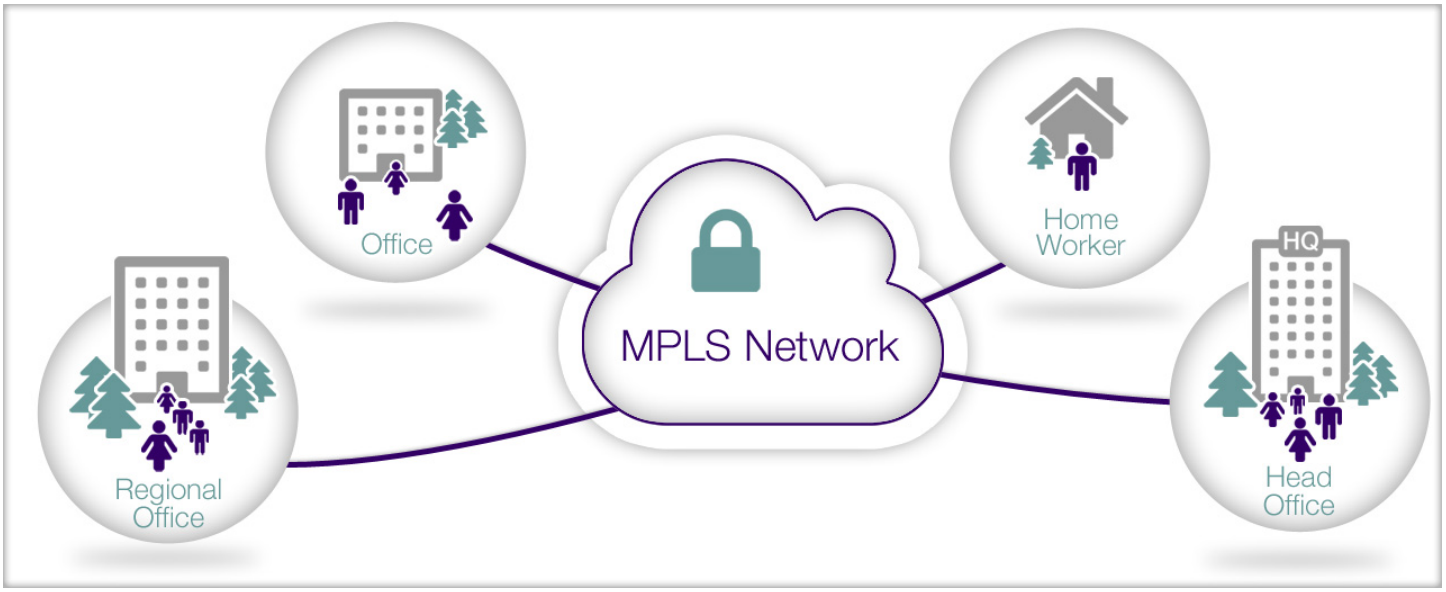


MPLS基础

肖宏辉 · 2 个月前

MPLS是什么？

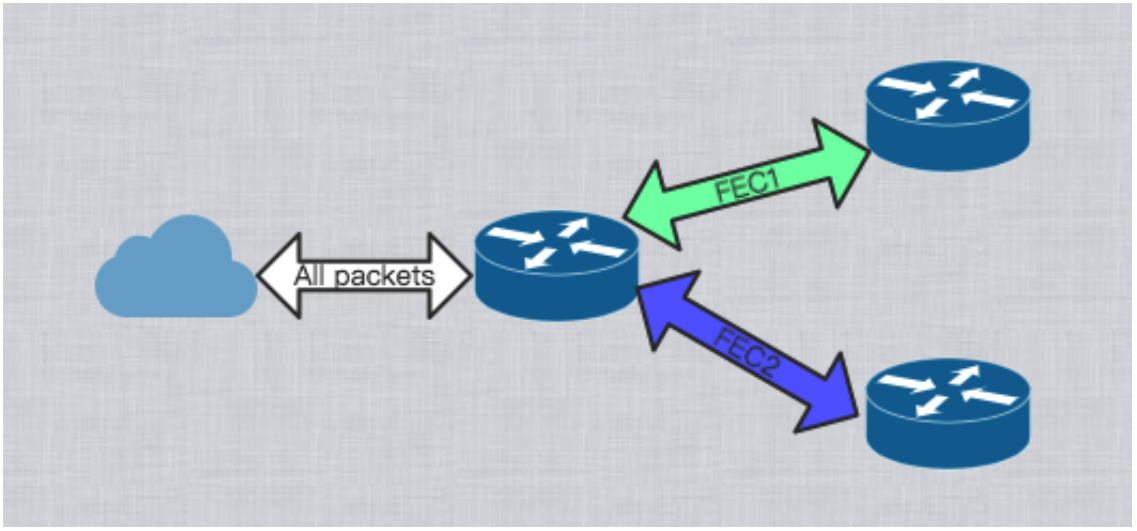
从发展历史来看，MPLS并非是一个非常新的技术，最早可以追溯在1997年，相应的IETF工作组就成立。在2001年，相应的RFC3031就发布了。从字面意思来看，MPLS全称是Multi-Protocol Label Switching，直译过来就是多协议标签交换技术。维基百科是这么定义MPLS的：一种在通讯网络上的高性能数据传输技术。信息太少，没法理解，不要着急，往下看。

MPLS解决了什么问题？

传统的路由网络里面，当一个（无状态的）网络层协议数据包（例如IP协议报文）在路由器之间游荡时，每个路由器都是独立的对这个数据包做出路由决策。**路由决策就是路由器决定数据包如何路由转发的过程。**路由决策在后面会多次提到，在这里指，每个路由器都需要分析包头，根据网络协议层的数据进行运算，再基于这些分析和运算，独立的为数据包选择下一跳（next hop），最后通过next hop将数据包发送出去。以IP协议报文为例，路由决策是

基于目的IP地址，路由器根据目的IP地址，选择路由条目，再做转发。路由决策可以认为是由两部分组成：

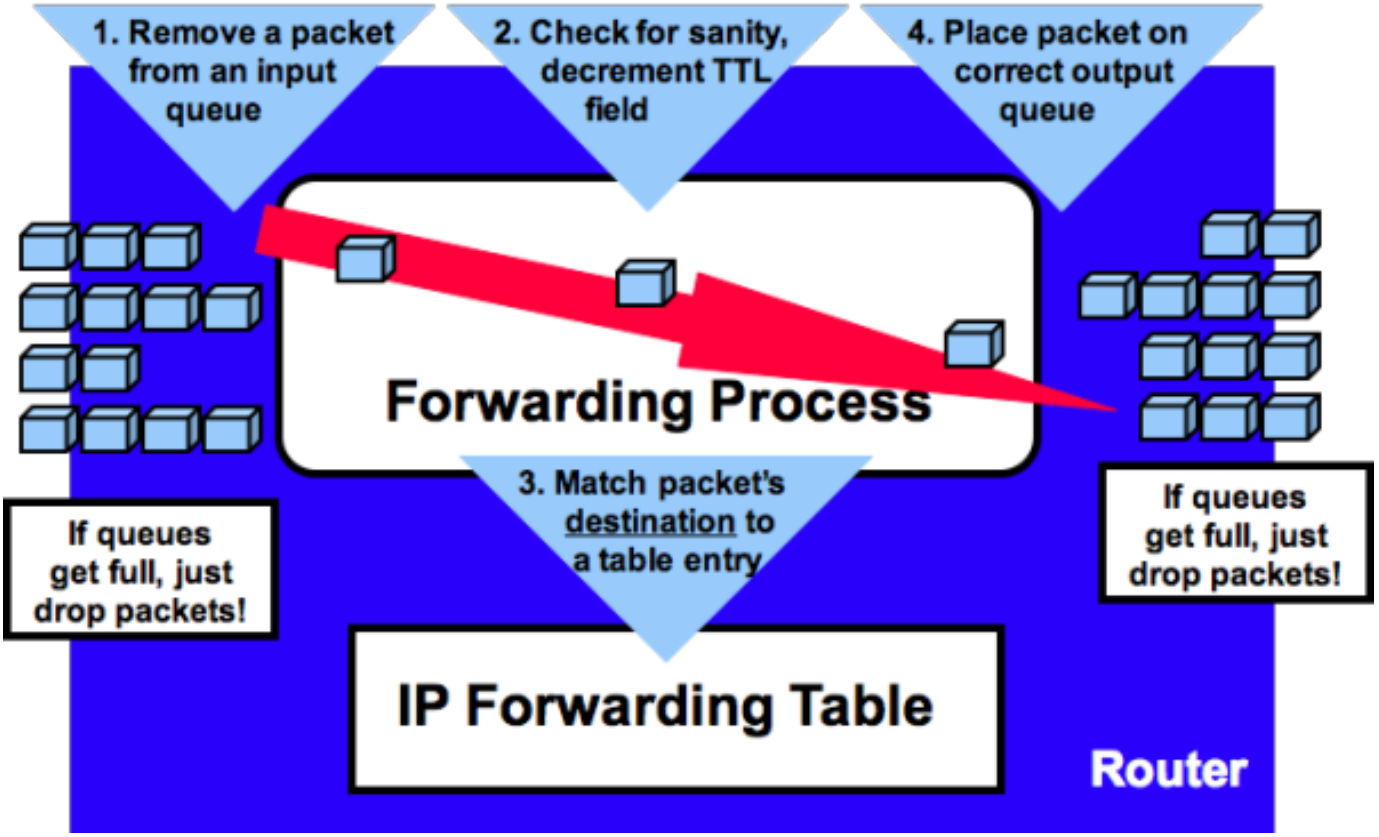
- 分类，将特定的数据包归属为一个等价转发类（Forwarding Equivalence Classes，FECs）
- 查找，查找FEC对应的next hop



对于同一个路由器来说，同一个FEC必然对应同一个next hop，那么属于同一个FEC的所有网络数据包必然会走同一条路径转发出去。（注，在多链路负载均衡的情况下，一个FEC也可能对应一组next hop，但是逻辑上还是能看成是一个next hop，因为殊途同归！）

具体到IP协议报文，当多个IP协议报文的目的地址都对应路由器的一条路由，且这条路由是所有路由里面最长匹配（longest match）的路由，那么对于这个路由器来说，就会认为这两个IP协议报文属于一个FEC。因此，这两个数据包就会走同一条路径出这个路由器。这就是我们最常见到的路由转发。

需要注意的是，这里的FEC是针对路由器的，而不是全局的。举个例子，目的地址为192.168.31.1和192.168.31.100的两个IP协议报文，第一个路由器具有192.168.31.0/24这条路由，那么在第一个路由器它们属于同一个FEC，都会被转发到第二个路由器。第二个路由器具有192.168.31.0/26和192.168.31.0/24两条路由，并且两条路由的next hop不一样。因为192.168.31.0/26能更精确的匹配192.168.31.1，所以192.168.31.1匹配第一条路由，而192.168.31.100匹配第二条路由，最终，这两个IP协议报文在第二个路由器被认为是不同的FEC，从不同的路径出去。这就是每个路由器都需要独立的做路由决策的原因之一。
路由器的工作原理如下图所示：



由于每个路由器都需要独立的路由决策（虽然会有这样那样的缓存机制加速决策），而路由器的收发队列一旦满了，就会丢包。所以在一个高流量，高容量的网络里面，无疑对每个路由器的要求都很高（否则就会丢包了！）

针对这个问题，MPLS提出了类似的，但是更简单的另外一种路由决策的方法。

传统的路由决策，路由器需要对网络数据包进行解包，再根据目的IP地址计算归属的FEC。而MPLS提出，当网络数据包进入MPLS网络时，对网络数据包进行解包，计算归属的FEC，生成标签（Label）。当网络数据包在MPLS网络中传输时，路由决策都是基于Label，路由器不再需要对网络数据包进行解包。并且Label是个整数，以整数作为key，可以达到O(1)的查找时间。大大减少了路由决策的时间。这里的Label就是MPLS里面的L。需要注意的是Label在MPLS网络里面，是作为网络数据包的一部分，随着网络数据包传输的。

也就是说，在MPLS网络里面，数据被封装在了盒子里，上面贴了标签，每个经手的人只需要读标签就知道盒子该送到哪。而传统的路由网络里面，每个经手的人都需要打开盒子，看看里面的内容，再决定送往哪。

这里提到了MPLS网络，这是一个由相连的，支持MPLS的设备组成的网络。打上MPLS标签的数据可以在这个网络里面传输。MPLS的核心就是，**一旦进入了MPLS网络，那么网络数据包的内容就不再重要，路由决策（包括FEC归属的计算，next hop的查找）都是基于Label来进行的。**

从目前看，MPLS带来的好处是，在MPLS网络里面，除了边界路由器，其他路由器可以由一些支持Label查找替换的低性能的交换机，或者路由器来完成。这一方面降低了组网的成本，另一方面提升了同样性能设备的转发效率。不过，随着路由器的发展，这方面的优势弱化了，而且，类似的问题，也不一定需要MPLS来解决。**MPLS的价值更多的在于其他方面。**不过初步理解MPLS，就先说这些。到目前为止，MPLS里面的M，P，L都介绍过，S其实也隐含的介绍过，S是Switching的意思，即基于Label做路由决策的意思，或者说标签交换里面的交换。相信大家也明白了为什么说MPLS是一种高效的数据传输的技术。

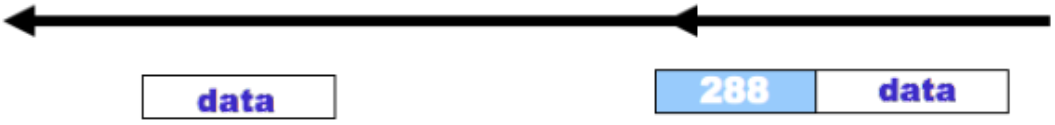
MPLS术语

相较于传统的路由交换技术，MPLS是一个全新的世界，因此有必要对MPLS中的一些术语和角色做一些解释。同时，为了表述简单，后面都用IP协议报文代替网络数据包来进行描述。

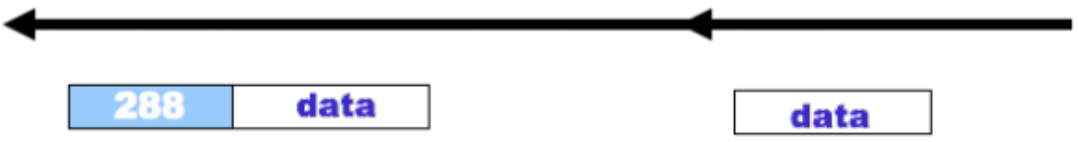
- FEC（Forwarding Equivalence Class）：交换等价类，前面描述过，同样的转发路径的网络数据包的集合。
- MPLS网络：由支持MPLS的，相连的设备的构成。

- LSH（Label Switching hop）：IP协议报文从一个MPLS设备发送到另一个MPLS设备，区别于传统的路由交换，LSH是基于Label的转发。
- NHLFE（Next Hop Label Forwarding Entry）：LSR中用来转发条目，相当于路由表之于路由器。包含了：
  - 下一跳：nexthop
  - 对数据包的当前label需要做的操作，包括了：
    - 替换（SWAP）

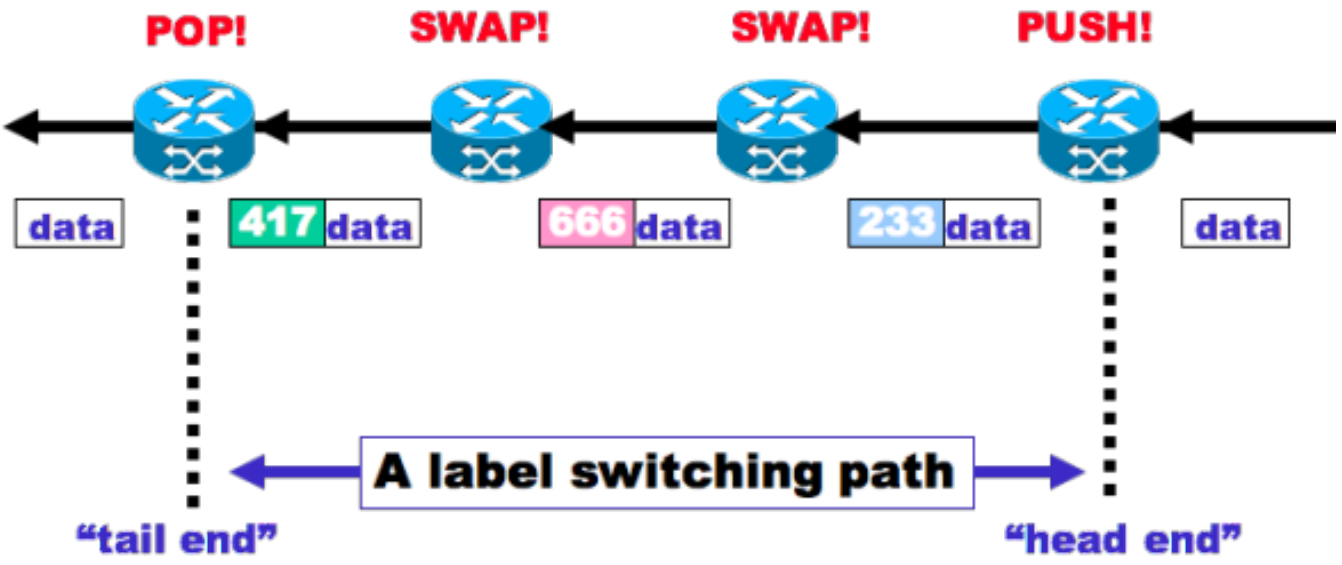
    - 删除（POP）



    - 添加（PUSH）



- LER（Label Edge Router）：有的地方也叫做 MPLS edge node。顾名思义，MPLS网络的边缘设备。
    - MPLS ingress node：进入MPLS网络的节点，也就是MPLS网络的入口路由器。该设备计算出IP协议报文归属的FEC，并把相应的Label放入IP协议报文。
    - MPLS egress node：出MPLS网络的节点，也就是MPLS的出口路由器。IP协议报文在这里回到传统的路由系统中。
- LSR（Label Switching Router）：支持MPLS转发的路由器。如果一个LSR有一个邻接的节点在MPLS网络之外，那么这个LSR就是LER。注意，这里的MPLS网络之外可以是：1.传统路由网络，2.另一个MPLS网络。
- LSP（Label Switching Path）：特定的FEC中的IP协议报文所经过的LSR的集合。LSP通常也被称为MPLS tunnel。



MPLS协议格式

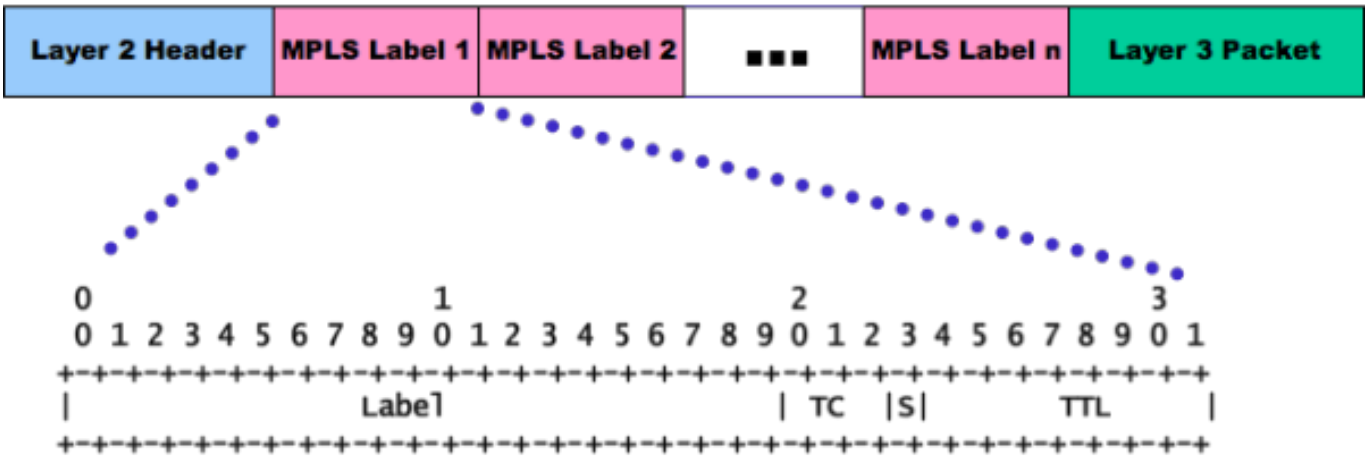
前面说了，MPLS把Label作为IP协议报文的一部分，存储在IP协议报文中。通常情况下，MPLS操作在OSI的2层（数据链路层）和3层（网络层）之间，因此也常常被认为是2.5层协议。这也就是MPLS能支持Multiprotocol的原因。Label不依赖于任何协议，直接定义在2-3层之间。当然，老司机们会说MPLS也可以在2层，例如MPLS-ATM和MPLS-FRMRLY。这种情况现在用的比较少，这里就不考虑。
MPLS的Label格式定义如下：

MPLS Label

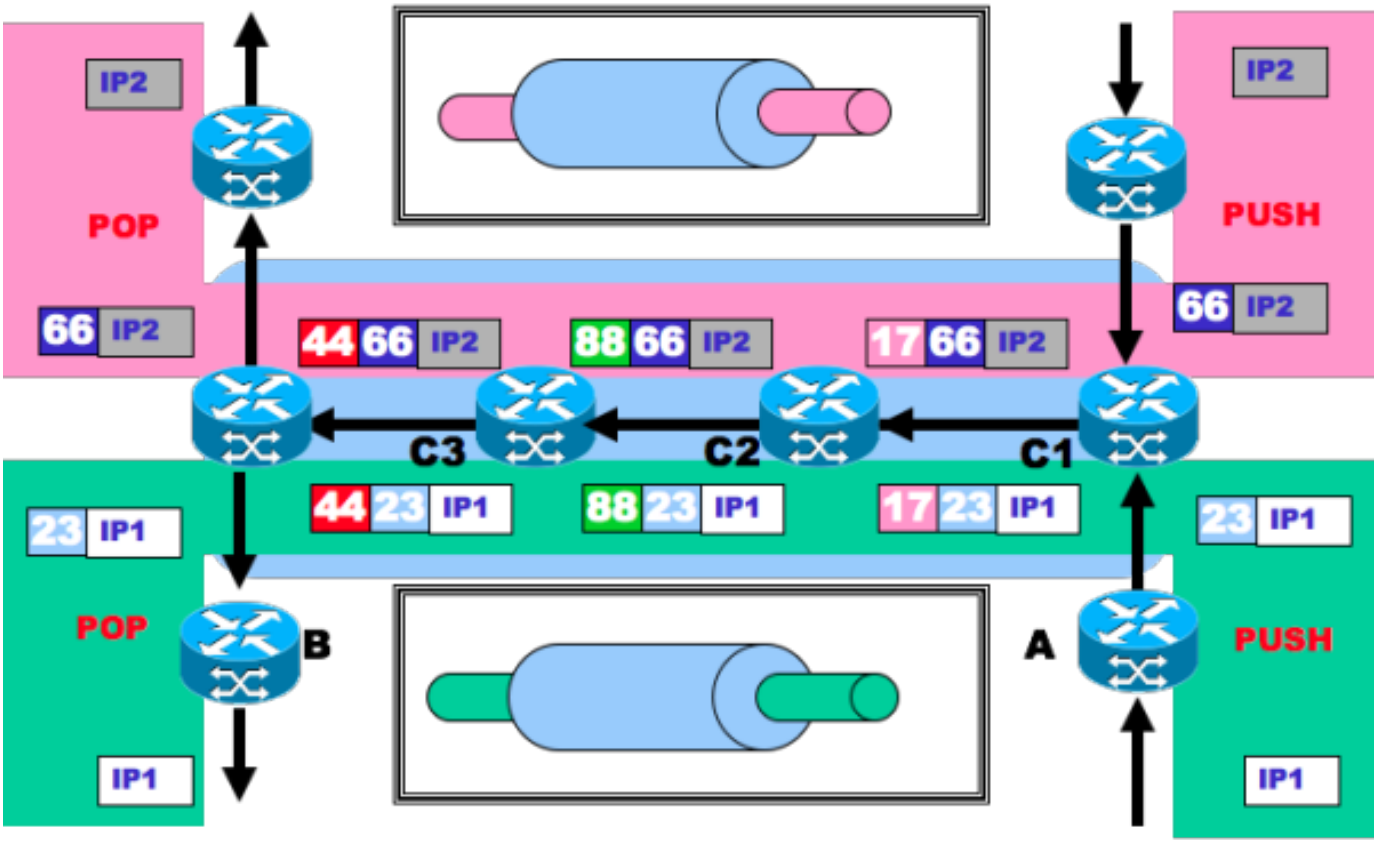
00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Label																				TC: Traffic Class (QoS and ECN)			S: Bottom-of-Stack	TTL: Time-to-Live							

- Label：前面提到过多次的Label，20bit的整数，也就是说Labe的容量是百万级的。虽然不是无限的，但是也很多了。
- TC：之前的EXP，改名成TC，由RFC5462定义。
- S：bottom of stack。什么是stack，一种常见的数据结构类型，特点是后进先出。S为1表明这已经是栈底了，即当前Label是IP协议报文最后一个MPLS标签。再执行一个POP操作，就能变成正常的IP协议报文了。
- TTL：TTL在IP协议里面的作用主要是防止环路和用于traceroute等工具。在之前的文章Traceroute里详细介绍过。MPLS里面的TTL作用是一样的。当数据包进入MPLS网络，网络层中的TTL会被拷贝至MPLS的TTL，每一次LSH，TTL减1，数据包出MPLS网络，MPLS中的TTL会拷贝至网络层。

刚刚提到了stack，MPLS中的Label不是指一个Label，而是由多个Label构成的Label Stack。



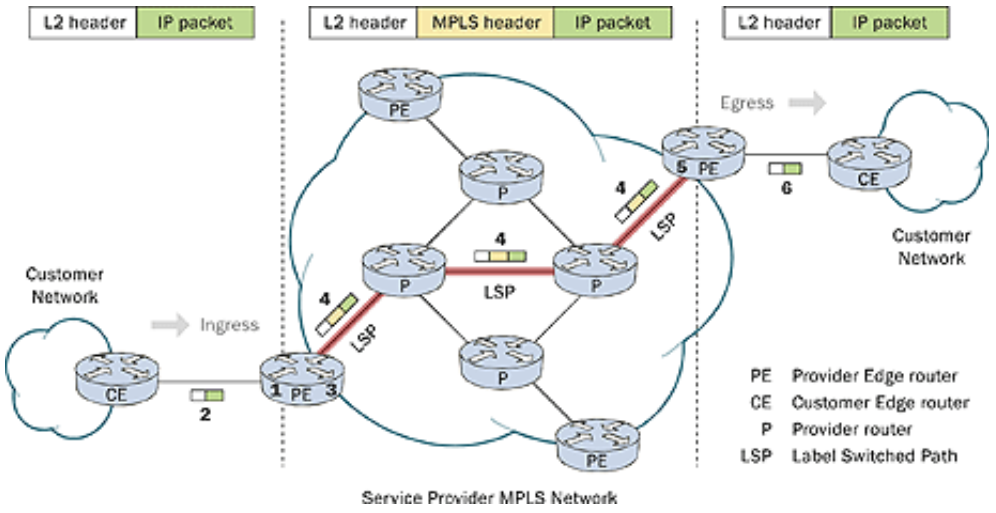
为什么要Label Stack？
从前面的描述看，MPLS似乎用一个Label就可以满足要求了。多Label的一个应用场景就是嵌套的LSP。直接看图吧：



对于IP1的报文来说，上层LSP就是由A->B组成。但是A并非直接转发给B的，而是通过另一个子LSP C1->C2->C3转发给的B。这么做，首先是因为A和B没有直接相连，没有办法直接转发。另一方面，因为IP1和IP2有一部分重合的路径，通过定义子LSP可以复用这部分路径。对于IP1和IP2来说，C1，C2，C3只需要存储一套NHLFE即可。

MPLS网络拓扑

看一个简单的MPLS网络：



这个图里面的术语跟之前描述的不一样，不过其实这个图里的术语更常见。可以简单的翻译

一下。

CE：前面没有介绍过，其实就是传统路由网络中与LER连接的路由器，可以理解成客户网络的边缘路由器。

PE：服务提供商的边缘路由器，对应LER。

P：服务提供商的路由器，对应LSR。

为什么会有这些术语上的不同，MPLS的提出本身是中立的，但是随着发展，现在应用最多的是电信网络，所以才有了customer provider这些概念，其实都是对应电信运营商网络中的设备。上图中间部分就是个MPLS网络，前面介绍过，MPLS网络中的IP报文都是带有MPLS标签的。下面来过一下工作过程：

1. 在所有的网络流量之前，PE路由器需要通过MPLS网络与远端PE路由器建立LSP。
2. 客户网络从CE发来的非MPLS 报文，发送到了ingress PE路由器，也就是MPLS ingress edge node。
3. ingress PE 路由器通过运算得出IP协议报文归属于哪个FEC，并把相应的Label加到了IP协议报文。
4. IP协议报文沿着LSP传输，每个P路由器都根据自身的NHLFE，替换Label，再把报文传给下一跳。
5. 在egress PE路由器，Label被从IP协议报文中删除，一个传统的IP协议报文又产生了。
6. IP协议报文被发送到了对端的CE路由器，最终进入了另一个客户网络。

到此为止，MPLS的data plane描述完了。MPLS是个很大话题，如果还有下次的话，应该会说说LSP的建立过程，也就是LDP（Label Distribution Protocol）。

计算机网络 Multiprotocol Label Switching(MPLS)