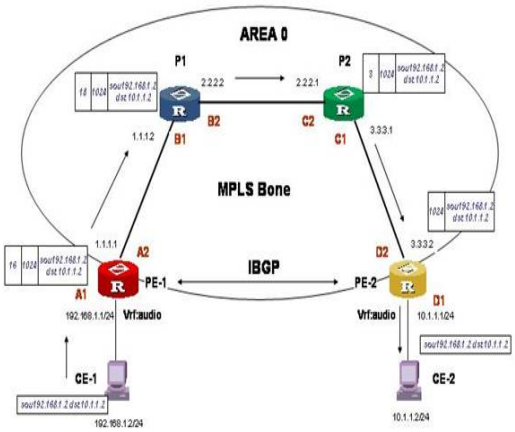
MPLS VPN中的Qos技术

郭晓翔 2006-03-24 发表

MPLS VPN中的Qos技术

网络模型



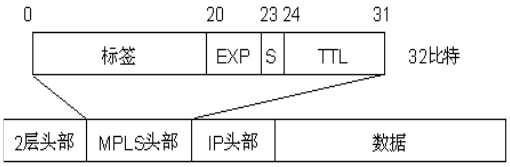
路由器 PE-1---P1或是P1-P2(假设)之间的广域网链路带宽是有限的，比如64K线路。在无QOS保证时，当A侧客户端从B侧服务器获取大量数据时，A侧电话听到B侧传来的声音出现明显断续，很长延迟，通话质量恶劣。

对MPLS/VPN和QOS必须了解的基本知识

MPLS/VPN

这里我们主要指的是基于BGP/MPLS VPN，也就是三层的VPN技术，网络组成由CE-PE-P-PE-CE这种方式构建。

MPLS是多协议标签交换协议的简称，多协议是指它能够支持多种三层协议；标签是一种短的，易于处理的，不包含拓扑信息，只具有局部意义的信息内容；MPLS的报文转发是基于标签的，在MPLS网络中，IP包在进入第一个MPLS设备时，MPLS边缘路由器分析IP包的内容并且为这些IP包选择合适的标签。以后所有MPLS网络中节点都是依据这个标签作为转发依据。当IP包最终离开MPLS网络时，标签被边缘路由器分离。



在MPLS中，一个标签标识了一个转发等价类（FEC——Forwording Equivalence Class）。一个转发等价类是在网络中遵循同样的转发路径的报文的集合，这些报文的目的地址甚至可以不同。

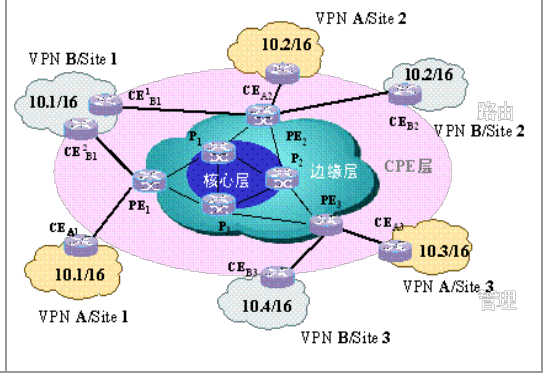
MPLS可以看做是一种面向连接的技术。可通过MPLS信令(如LDP，Label Distribute Protocol，标签分配协议)或手工配置的方法建立好MPLS标记交换连接(Label Switched Path，简称LSP）以后，数据转发过程中，在网络入口进行流分类，根据数据流所属的FEC选择相应的LSP，把需要通这条LSP的报文打上相应的标签，中间路由器在收到MPLS报文以后直接根据MPLS报头的标签进行转发，而不用再通过IP报文头的IP地址查找。在MPLS标记交换路径的出口（或倒数第二跳），弹出MPLS包头，还回原来的IP包（在VPN的时候可能是以太网报文或ATM报文等）。

由于FEC可以是按照目的地址划分的，这同传统的IP转发相同，也可以是基于源地址、目的地址、源端口、目的端口、协议类型、CoS、VPN等等信息的任意组合。而MPLS可以把任何流关联到一个FEC，然后把一个FEC映射到一个LSP，这个LSP可以是为了这种FEC而特殊构造的，这使得服务提供商可以非常精确地控制网络中的每个流。这种空前的控制能力使网络能够提供更加有效和可预测的服务。

根据扩展方式的不同MPLS VPN可以分为BGP扩展实现的MPLS VPN，和LDP扩展实现的VPN。根据PE（Provider Edge）设备是否参与VPN路由又细分为二层VPN和三层VPN。同依赖于IP Tunnel技术实现的传统IP VPN不同，MPLS VPN不依靠封装和加密技术，而是依靠转发表和数据包的标记来创建一个安全的VPN。

在L3 MPLS VPN（又称MPLS BGP VPN）的模型中，网络由运营商的骨干网与用户的各个Site组成，所谓VPN就是对site集合的划分，一个VPN就对应一个由若干site组成的集合。但是必须遵循如

下规则：两个Site之间只有至少同时属于一个VPN定义的Site集合，才具有IP连通性。MPLS BGP VPN的框架模型如图所示：

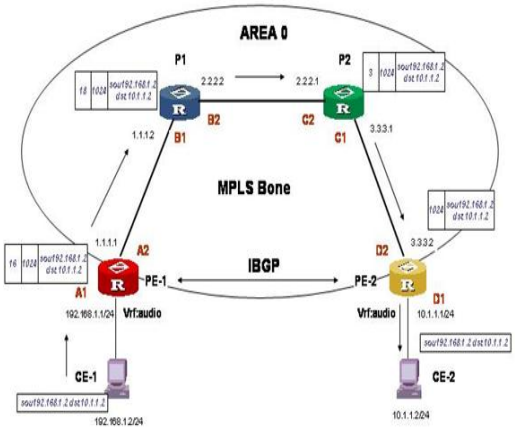


如图所示，基于BGP扩展实现的L3 MPLS VPN所包含的基本组件：
PE：Provider Edge Router，骨干网边缘路由器，存储VRF（Virtual Routing Forwarding Instance），处理VPN-IPv4路由，是MPLS三层VPN的主要实现者；
CE：Custom Edge Router，用户网边缘路由器，分布用户网络路由；
P router： Provider Router，骨干网核心路由器，负责MPLS转发；
VPN用户站点（site）：是VPN中的一个孤立的IP网络，一般来说，不通过骨干网不具有连通性，公司总部、分支机构都是site的具体例子。CE路由器通常是VPN Site中的一个路由器或交换设备，Site通过一个单独的物理端口或逻辑端口（通常是VLAN端口）连接到PE设备上；
用户接入MPLS VPN的方式是每个site提供一个或多个CE，同骨干网的PE连接。在PE上为这个site配置VRF，将连结PE-CE的物理接口、逻辑接口、甚至L2TP/IPSec隧道绑定的VRF上。
BGP扩展实现的MPLS-VPN扩展的了BGP NLRI中的IPv4地址，在其前增加了一个8字节的RD（Route Distinguisher）。RD时用来标识VPN的成员---即Site的。VPN的成员关系是通过路由所携带的route target属性来获得的，每个VRF配置了一些策略，规定一个VPN可以接收哪些Site来的路由信息，可以向外发布哪些Site的路由信息。 每个PE根据BGP扩展发布的信息进行路由计算，生成每个相关VPN的路由表。
PE-CE之间要交换路由信息一般是通过静态路由，也可以通过RIP、BGP等。PE-CE之间采用静态路由的好处是可以减少CE设备可能会因为管理不善等原因造成对骨干网BGP路由产生震荡，影响骨干网的稳定性；如果采用BGP可以实现动态的网络扩展，网络路由信息发生变化时，不必更改设备的配置信息。
PE与PE之间需要运行IBGP协议，存在可扩展性问题，但采用路由反射器RR可以显著地减少IBGP连接的数量。
MPLS/BGP VPN提供了灵活的地址管理。由于采用了单独的路由表，允许每个VPN使用单独的地址空间中，称为VPN-IPv4地址空间，RD加上IPv4地址就构成了VPN-IPv4地址。很多采用私有地址的用户不必再进行地址转换NAT。NAT只有在两个有冲突地址的用户需要建立Extranet进行通信时才需要。
在MPLS/BGP VPN中，属于同一的VPN的两个site之间转发报文使用两层标签来解决，在入口PE上为报文打上两层标签，第一层（外层）标签在骨干网内部进行交换，代表了从PE到对端PE的一条隧道，VPN报文打上这层标签，就可以沿着LSP到达对端PE，这时候就需要使用第二层（内层）标签，这层标签指示了报文应该到达哪个site，或者更具体一些，到达哪一个CE，这样，根据内层标签，就可以找到转发的接口。可以认为，内层标签代表了通过骨干网相连的两个CE之间的一个隧道。
L3 MPLS-VPN通过和Internet路由之间配置一些静态路由的方式，可以实现VPN的Internet上网服务，还可以为跨不同地域的、属于同一个AS但是没有自己的骨干网的运营商提供VPN互连，即提供"运营商的运营商"模式的VPN网络互连。
MPLS/MBGP VPN可以简化对用户端设备的需求和用户管理、维护Intranet/Extranet的复杂性，每个CE仅需要维持一个到PE的路由交换协议，CE间的路由交换、传输控制、路由策略由运营商根据VPN用户的需求来实施。由于BGP的策略控制能力很强，随之而来的是VPN用户路由策略控制的灵活性。

QOS
此处涉及的QOS，只是对QOS中的一种技术拥塞管理，也就是大家熟悉的PQ，CQ之类，主要用在当线路出现拥塞时，保证重要业务的带宽，而其它QOS策略，可依此类推。
因采用了differ-serv模型的QOS不是端到端的，不依靠信令来传送，所以整网全部链路拥塞着或者链路将要拥塞的参与IP流转发的路由器都必须参与QOS处理。
QOS对出方向的数据流量做处理是合理而必要的。
线路不拥塞时，QOS是不生效的。
提高重要数据的优先级，保证重要数据的带宽是我们要做的。
我们使用
CBQ对业务进行分流，同时保证带宽。
对支持differ-ser模型的设备，采用differ-ser模型来实现QOS。

第二章 实现思想

从最简单的组网分析



实现原理：对于MPLS VP的网络来说，CE-1（IP：192.168.1.2），发送一个报文，去访问CE-2(IP:10.1.1.2),报文格式：[sou192.168.1.2 dst10.1.1.2]

转发到PE时，PE根据自己的VRF定义，以及通过IBGP学习到的路由和标签表，给这样的报文进行了第一层私网mpls标签，假定报文格式为[1024 | sou192.168.1.2 dst10.1.1.2]。

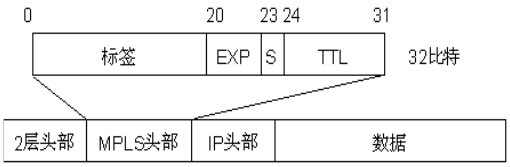
在从PE-1转发出去时，根据公网路由表和公网标签，发送出去的报文格式为：[16 | 1024 | sou192.168.1.2 dst10.1.1.2],这样报文形成了两层标签。

在从PE-1到P1就是这样的格式，当P1收到这样的报文，发现自己非倒数第二跳（通过标签表），转发报文，但替换掉公网标签，所以P1到P2的报文更改为[17 | 1024 | sou192.168.1.2 dst10.1.1.2]。

当报文到达P-2时，P-2发现自己是倒数第二跳，弹出外层标签，将带有一层私网标签的报文发送到PE-2上，报文格式为: [1024 | sou192.168.1.2 dst10.1.1.2]。

PE-2收到此报文，根据自己的私网标签表和路由表，查找到下一跳接口，将标签剥离，以纯IP报文方式发送给CE-2,所以报文格式为：[sou192.168.1.2 dst10.1.1.2]，CE-2就收到了它可以识别的IP报文。

所以从上述过程，我们可以看到，CE-PE是IP路由转发，而PE-P-PE 是标签转发，不查看IP内容，只查看标签，PE-CE任是IP转发，所以在IP网络中使用的标识IP报文优先级的TOS域在MPLS VPN中失去了它的存在价值，但是，我们再看一次MPLS的格式：



其中的EXP位占3个bit，和IP报文中的TOS域使用的ip precende值类似，所以我们把它定义为mpls报文的优先级。如何实现呢，做一个映射关系，IP报文中的TOS值是多少，在添加标签时将其拷贝一份到标签的EXP中，当有多层标签时，外层标签也复制一份。而在MPLS VPN的骨干域中，我们可以读取MPLS中的EXP位来实现标签转发中的报文优先和拥塞管理功能。

注意一点：对多层标签，完全可能出现不同的标签中的EXP位不一致的情况，缺省情况下，一般是都从IP报文TOS中拷贝，但BGP MPLS VPN中，P设备只查找外层标签，不考虑内层标签，在某些特定情况下，可能需要外层标签的EXP有不同的定义，所以这种策略是可以实现的，而且在应用中也增加了灵活性。

从配置看实现

从我们的产品实现来看，目前（截至到2004年8月）可支持做PE和P设备的产品为AR系列路由器，NE系列路由器，R26/36（VRP3版本）路由器，S8016。

NE40，NE80，S8016产品因支持differ serv模型，所以它们的实现和其他用软件实现CBQ方式的产品有不同，配置时注意方法。

而从CE发往PE的IP报文分两种，一种是本身IP的precedence或是DSCP值有置位，即在IP中就有优先级，这时PE可根据这个值直接转换为MPLS中的EXP值，另一种是IP报文没有优先级，在PE上要先对其做定义，根据报文的其他特征（如ACL的五元组）来定义EXP值，这种方式下对PE设备处理要求较高，不如第一种方式简单。

因我们配置MPLS时，版本差异较大，大家注意看手册。
我们先讨论第二种情况：IP报文没有定义优先级
1. 当PE-1用AR46这种软件实现的设备时：
定义从CE-1所在VPN的IP流，用ACL定义
```
acl number 2000 match-order auto
 rule 0 permit vpn-instance audio source 192.168.1.0 0.0.0.255
 rule 1 deny vpn-instance audio        (audio为VPN名称)
```
利用CBQ来定义EXP值：
```
traffic classifier test operator and
 if-match acl 2000
traffic behavior exp4
  remark mpls-exp 4
qos policy modi
 classifier test behavior exp4
```
在入接口上使能此CBQ，将IP报文封装MPLS标签后，定义EXP值为4。
```
interface Ethernet0/0/0
 ip binding vpn-instance audio
 ip address 192.168.1.1 255.255.255.0
 qos apply policy modi inbound
```

上述方法只是改变了EXP的值，而IP报文中的值并没改变，如果同时想改变IP报文中的DSCP值，可用如下方法：
```
acl number 2000 match-order auto
 rule 0 permit vpn-instance audio source 192.168.1.0 0.0.0.255
 rule 1 deny vpn-instance audio
acl number 2001 match-order auto
 rule 0 permit source 192.168.1.0 0.0.0.255
 rule 1 deny
traffic classifier test operator and
 if-match acl 2000
traffic classifier test-1 operator and
 if-match acl 2001
#
traffic behavior dscp11
  remark dscp af11
traffic behavior exp4
  remark mpls-exp 4
#
qos policy modi
 classifier test behavior exp4        (将符合ACL2000的报文修改它的EXP值为4)
 classifier test-1 behavior dscp11    (将符合ACL2001的报文的IP优先级的DSCP值置为af11)
```
应用到PE的私网接口的入方向上
```
interface Ethernet0/0/0
 ip binding vpn-instance audio
 ip address 192.168.1.1 255.255.255.0
 qos apply policy modi inbound
```

这时，报文的TOS和MPLS EXP值都改变了，再定义QOS中的CBQ时，可按以前对待IP方式一样处理了。
定义方法如下：

2． 当PE-1用NE40这种NP实现的设备时，它没有CBQ这种软件处理机制，它采用了一种differ serv的模型，这种模型要求它处理的IP报文中的DSCP值要有定义，根据这个值，相当于分类出了一种流，对这个流再使用相应的转发模型（自己定义的，如DSCP为AF43的流，在网络接口拥塞时，可保证2M带宽），缺省的转发模型是网络拥塞时，优先转发DSCP值大的报文（和交换机比较类似）。
实现方法：
定义流分类
```
rule-map intervlan audio ip 192.168.1.0 0.0.0.255 any
```
定义流动作
```
flow-action mark diffserv af43    (将流定义为af43)
```
关联流
```
eacl audio-vpn audio mark
```

在PE入接口上应用：

```
interface Ethernet2/0/15
 negotiation auto
 undo shutdown
 ip binding vpn-instance audio
 ip address 192.168.1.1 255.255.255.0
 access-group router eacl audio-vpn
```

而NE40/80它的智能在于，一旦定义了DSCP的值，它会自动生成一张DSCP - EXP的映射表，当然你也可以手工修改，表如下：

DSCP	EXP	DSCP	EXP
00 ~ 07	0	32 ~ 39	4
08 ~ 15	1	40 ~ 47	5
16 ~ 23	2	48 ~ 55	6
24 ~ 31	3	56 ~ 63	7

可用命令修改：

```
[Quidway-mpls] dscp-exp-map 08 4 （将DSCP值8对应EXP4）
```

进入MPLS骨干域时的处理

当带有EXP位的置位标识（表示此MPLS优先级）的报文进入骨干域，也就是从PE口转发到P设备在骨干域中时，多个VPN用户会共用同一骨干链路，在出接口上导致端口队列拥塞，这时我们就会像对待IP报文转发的处理一样，采用相应的队列调度机制，利用拥塞管理技术来实现对重要的VPN或是重要业务的带宽保证，常用的拥塞管理技术为PQ,CQ,WFQ,CBQ等，他们的实现原理和在IP网络中是一样的，不同点在于他们不是依据IP报文的特征来分类流，而是依据MPLS EXP位来分类流。

配置方法如下（在PE设备的上）：

```
traffic classifier ddd operator and
 if-match mpls-exp 4                          （定义流，这个流的MPLS EXP位为4）
traffic behavior dede
  queue af bandwidth pct 100                   （带宽100％可用，缺省只有带宽75％可用）
queue af bandwidth 2000                        （定义一个动作,保证带宽为2M）
qos policy change
 classifier ddd behavior dede                  （定义策略，将流定义和动作结合）
interface Ethernet0/0/1
 ip address 1.1.1.1 255.0.0.0
 mpls
 mpls ldp enable
 qos apply policy change outbound        (应用到PE - P的出接口上，这个接口可能拥塞)
```

中间的P设备同理可按此方式实现重要业务的带宽保证。

说明：带宽保证，多是应用在报文的出接口上，只有出接口队列在带宽不足时会出现拥塞，导致报文丢弃，而保证的带宽只有在出口拥塞时才会启用拥塞管理机制（即CBQ或是PQ等才会生效），在正常情况下，业务不会因这个带宽受限制（此QOS原理部分参考其他相关材料）。