此文章已于 14:06:41 2016/6/16 发布到 MADHEX

BGP/MPLS VPN的江湖恩仇录

MPLS

一、MPLS物种起源

•        IP的危机

在90年代中期，当时路由器技术的发展远远滞后于网络的发展速度与规模，主要表现在转发效率低下、无法提供QOS保证。原因是：当时路由查找算法使用最长匹配原则，必须使用软件查找；而IP的本质就是"只关心过程，不注重结果"的"尽力而为"。当时江湖上流行一种论调：过于简单的IP技术无法承载网络的未来，基于IP技术的因特网必将在几年之后崩溃。

•        ATM的野心

此时ATM跳了出来，欲收编所有帮派，一统武林。不幸的是：信奉唯美主义的ATM走向了另一个极端，过于复杂的心法与招式导致没有任何厂商能够完全修练成功，而且无法与IP很好的融合。在与IP的大决战中最终落败，ATM只能寄人篱下，沦落到作为IP链路层的地步。

ATM技术虽然没有成功，但其中的几点心法口诀，却属创新：

•    屏弃了繁琐的路由查找，改为简单快速的标签交换

•    将具有全局意义的路由表改为只有本地意义的标签表

这些都可以大大提高一台路由器的转发功力。

MPLS的创始人"label大师"充分吸取了ATM的精华，但也同时认识到IP为江湖第一大帮派，无法取而代之。遂主动与之修好，甘当IP的承载层，但为了与一般的链路层小帮有所区别，将自己定位在第2. 5层的位置。"label大师"本属于八面玲珑之人，为了不得罪其他帮派，宣称本帮是"multiprotocol"，来者不拒，也可以承载其他帮派的报文。在经过一年多的招兵买马、上下打点之后，于1997年的武林大会上，正式宣布本帮成立，并命名为MPLS（MultiProtocol label Switch）

二、MPLS包头结构

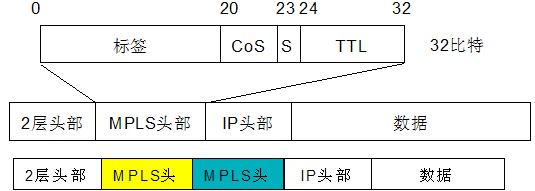
通常，MPLS包头有32Bit，其中有：

20Bit用作标签（Label）

3个Bit的EXP, 协议中没有明确，通常用作COS

1个Bit的S,用于标识是否是栈底，表明MPLS的标签可以嵌套。

8个Bit的TTL

[](http://photo.blog.sina.com.cn/showpic.html)  
理论上，标记栈可以无限嵌套，从而提供无限的业务支持能力。这是MPLS技术最大的魅力所在。

三、MPLS术语

标签（Label）：是一个比较短的，定长的，通常只具有局部意义的标识，这些标签通常位于数据链路层的数据链路层封装头和三层数据包之间，标签通过绑定过程同FEC相映射。

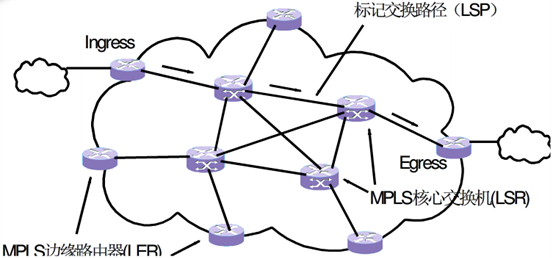
FEC：Forwarding Equivalence Class，FEC（转发等价类），是在转发过程中以等价的方式处理的一组数据分组， MPLS创始人在秘笈本来规定：可以通过地址、隧道、COS等来标识创建FEC，只可惜后辈弟子大多资质愚钝，不能理解其中的精妙之处，所以我们现在看到的MPLS中只是一条路由对应一个FEC。通常在一台设备上，对一个FEC分配相同的标签。

LSP：标签交换通道。一个FEC的数据流，在不同的节点被赋予确定的标签，数据转发按照这些标签进行。数据流所走的路径就是LSP。

LSR：Label Switching Router，LSR是MPLS的网络的核心交换机，它提供标签交换和标签分发功能。

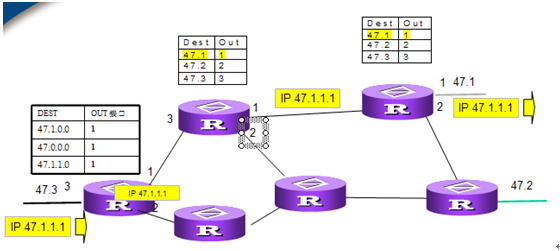
LER：Label Switching Edge Router,在MPLS的网络边缘，进入到MPLS网络的流量由LER分为不同的FEC，并为这些FEC请求相应的标签。它提供流量分类和标签的映射、标签的移除功能。

四、MPLS北斗七星阵法图

[](http://photo.blog.sina.com.cn/showpic.html)

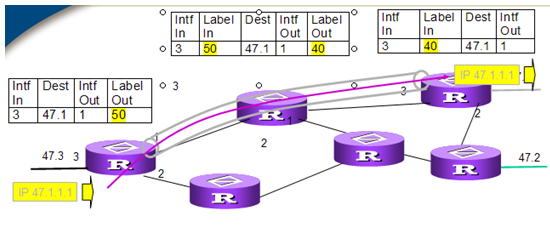
    该阵法分为内外两层，外层由功力高强的弟子担纲（至少是个堂主（LER），在IP报文冲阵时负责接收IP报文，查找标签转发表，给IP报文打标签操作（PUSH）在IP报文出阵时对标签报文进行弹出操作（POP），按IP路由进行转发。内层由功力较低的入门弟子组成，负责对标签报文进行快速的标签交换操作（SWAP）

五、IP的hop-by-hop逐跳转发

[](http://photo.blog.sina.com.cn/showpic.html)

    IP的逐跳转发，在经过的每一跳处，必须进行路由表的最长匹配查找（可能多次），速度缓慢。

六、Label Switched Path (LSP)

[](http://photo.blog.sina.com.cn/showpic.html)  
MPLS的标签转发，通过事先分配好的标签，为报文建立了一条标签转发通道（LSP），在通道经过的每一台设备处，只需要进行快速的标签交换即可（一次查找）。

FEC的精妙之处：不同目的地址（属于相同的网段）的IP报文，在ingress处被划分为相同的FEC，具有相同的标签，这样在LSR处，只需根据标签做快速的交换即可。而对于传统的IP路由，在每一跳处实际上都是一次重新划分FEC的过程。如果一台路由器对于ip路由和标签交换同样使用了cache功能，由于对于路由来说，在cache中只能记录主机路由，条目将十分有限，而标签对应的是FEC，可能是网段，可以做到很少的条目匹配大量的报文。

FEC的致命缺陷：对于一条FEC来说，沿途所有的设备都必须具有相同的路由（前缀和掩码必须完全相同）才可以建成一条LSP。换句话说，使用MPLS转发的所有沿途设备上，对于要使用标签转发的路由，都不能做路由聚合的操作。

七、上下打点

当一个链路层协议收到一个MPLS报文后，她是如何判断这是一个MPLS报文，应该送给MPLS处理，而不是象普通的IP报文那样，直接送给IP层处理？

回答：还记得MPLS的创始人"label大师"曾用了一年的时间来"上下打点"吗？当时主要的工作就是取得各个链路层帮派的通行证。

例如：

在以太网中：使用值是0x8847(单播)和0x8848（组播）来表示承载的是MPLS报文（0800是IP报文）

在PPP中：增加了一种新的NCP：MPLSCP，使用0x8281来标识

八、LDP

有了标签，转发是很简单的事，但是如何生成标签，却是MPLS中最难修练的部分。在MPLS秘笈中，这部分被称为LDP（Label Distribution Protocol），是一个动态的生成标签的协议。

其实LDP与IP帮派中的动态路由协议（例如RIP）十分相像，都具备如下的几大要素：

•    报文（或者叫消息）

•    邻居的自动发现和维护机制

•    一套算法，用来根据搜集到的信息计算最终结果。

只不过前者计算的结果是标签，后者是路由罢了。

九、LDP消息

在LDP协议中，存在4种LDP消息：

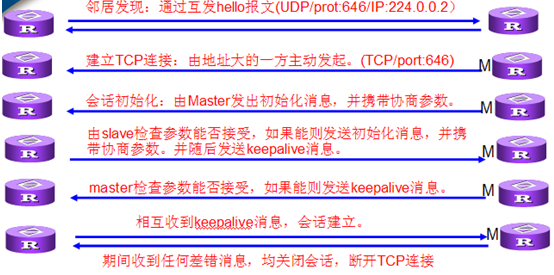
•       发现（Discovery）消息：用于通告和维护网络中LSR的存在。

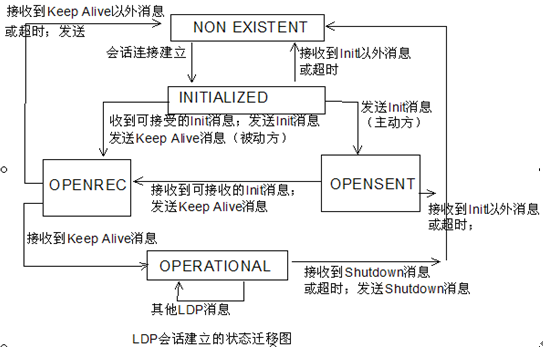
•       会话（Session）消息：用于建立，维护和结束LDP对等实体之间的会话连接。

•       通告（Advertisement）消息：用于创建、改变和删除特定FEC-标签绑定。

•       通知（Notification）消息：用于提供消息通告和差错通知。

十、LDP会话的建立和维护

[](http://photo.blog.sina.com.cn/showpic.html)  
十一、LDP邻居状态机

[](http://photo.blog.sina.com.cn/showpic.html)  
十二、标签的分配和管理

标记分发方式：

DOD（Downstream On Demand）下游按需标记分发

DU（Downstream Unsolicited）下游自主标记分发

标记控制方式：

有序方式（Odered）标记控制

独立方式（Independent）标记控制

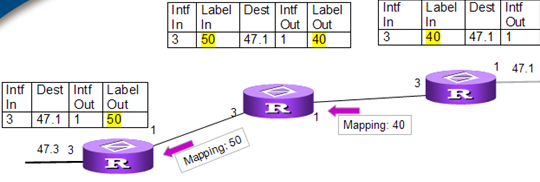
标签保留方式

保守方式

自由方式

上游与下游：在一条LSP上，沿数据包传送的方向，相邻的LSR分别叫上游LSR(upstream LSR )和下游LSR（downstream LSR）。下游是路由的始发者。

十三、LDP标签分配方式（DU）

[](http://photo.blog.sina.com.cn/showpic.html)

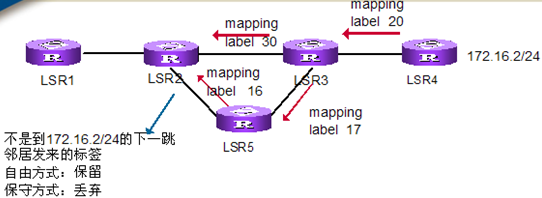
下游主动向上游发出标记映射消息。

标签分配方式中同样存在水平分割，即：对我已经选中的出口标签，就不再为下一跳分配出标签。

标签是设备随机自动生成的，16以下为系统保留。

还有一种DOD方式（由上游向下游请求），修练的人较少。

十四、LDP标签保留方式

[](http://photo.blog.sina.com.cn/showpic.html)

自由方式（Liberal retention mode）：保留来自邻居的所有发送来的标签

优点：当IP路由收敛、下一跳改变时减少了lsp收敛时间

缺点：需要更多的内存和标签空间。

保守方式（Conservative retention mode）：只保留来自下一跳邻居的标签，丢弃所有非下一跳邻居发来的标签。

优点：节省内存和标签空间。

缺点：当IP路由收敛、下一跳改变时lsp收敛慢

比较流行的是自由方式。

十五、LDP标签控制方式

有序方式（Odered）标记控制：除非LSR是路由的始发节点，否则LSR必须等收到下一跳的标记映射才能向上游发出标记映射。

独立方式（Independent）标记控制：LSR可以向上游发出标记映射，而不必等待来自LSR下一跳的标记映射消息。

比较流行的是有序方式。

十六、LDP标签分配

如果采用（DU+自由＋有序）的标签分配及控制方式：

•       发现自己有直连接口路由时会发送标签；

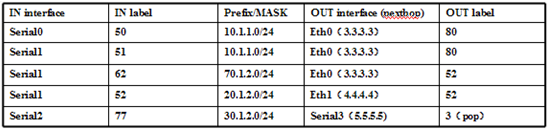
•       收到下游到某条路由的标签并且该路由生效（也就是说，在本地已经存在该条路由，并且路由的下一跳和标签的下一跳相同）时会发送标签。

•       标签表中会存在大量的非选中的标签。

下面的说法正确吗：如果某个网络中只有部分设备运行MPLS（MPLS域嵌在IP域中），则只会对运行MPLS的设备（MPLS域）的直连路由生成标签，对于其他设备（IP域）始发的路由则不会生成标签。

如果没有标签，那对于通过MPLS域的目的地址在IP域的报文如何转发呢？

十七、标签转发表心法口诀

[](http://photo.blog.sina.com.cn/showpic.html)

 标签转发表中的IN和OUT，是相对于标签转发而言，不是相对于标签分配的IN和OUT：

心法口诀：入标签是我分给别人的，出标签是别人分给我的。

             的标签是给别人用的，我不会添加到报文中。

对于一台设备的标签转发表（全局标签空间）来说：

• 所有的入标签( )

• 对于相同的路由（下一跳也相同），出标签( )

• 对于不同的路由（但下一跳相同），出标签( )

• 对于不同的路由（下一跳也不同），出标签( )

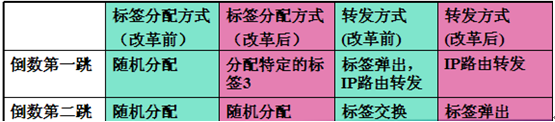
• 对于同一条路由，入标签和出标签( )

A 一定不同  B 一定相同 C 可能相同

十八、倒数第二跳弹出（P H P）

话说MPLS传到了第二代，由PHP接任掌门。PHP天资聪颖且富有创新精神。他经过对MPLS北斗七星阵法的深入研究，发现本帮的这门绝学虽然暗合天数、精妙无比，但并非没有可改进之处：在阵法的出口处，Egress LSR本应变MPLS转发为IP路由查找，但是他收到的仍旧是含有标签的MPLS报文，按照常规，这个报文应该送交MPLS模块处理，而此时MPLS模块不需要标签转发，能做的只是去掉标签，然后送交IP层。其实对于Egress LSR，处理MPLS报文是没有意义的。最好能够保证他直接收到的就是IP报文。这就需要在ELSR的上游（倒数第二跳）就把标签给弹出来。但关键问题是：上游设备如何知道自己是倒数第二跳呢？其实很简单，在倒数第一跳为其分配标签时做一下特殊说明即可（分配一个特殊的标签3）。

经过几次实战检验，效果很好，遂正式以自己的名字命名为：PHP（Penultimate Hop Popping),倒数第二跳弹出。

[](http://photo.blog.sina.com.cn/showpic.html)

十九、路由环路的预防与检测

路由环路的预防：任何涉及到转发或者是路由的阵法，都容易发生"路由环路"这样的走火入魔的事件。MPLS也不例外。创始人"label大师"深知武功中"借力打力"的原理，既然LSP的建立是依赖IP路由的，那么环路的预防也应该交给IP来做。自己无需处理了。

路由环路的检测：把自己的身家性命完全交给他人，毕竟不妥，万一IP没有把持住，后果不堪设想。所以虽然可以不作预防，但是必要的检测手段还是必需的，使用武林中通行的做法TTL即可。每经过一次MPLS转发，TTL减一。

在标签转发过程中，MPLS报文头中的TTL减一，那么ip报文头中的TTL是否还减一？

二十、MPLS的衰落……

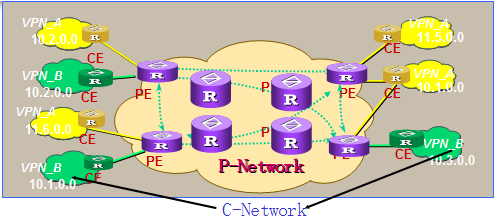
虽然MPLS的历任掌门都致力于本帮的发扬光大，但是要想整个武林都重新学习一门新功夫谈何容易。更为致命的是：MPLS标称的"身手敏捷"、"让一台IP路由器快速完成转发"也遇到了极大的挑战。由于社会进步，武林界已经告别以提升内力为主的冷兵器时代（软件转发），快速步入火器时代（硬件转发）。各种自动（ASIC）、半自动（NP）的武器价格低廉、江湖上几乎人手一把。当第二任掌门PHP发现凭借自己多年的修行，竟然连一个手持AK47的入门马仔（L3）都对付不了时，不禁仰头长叹，意识到日后再无人会苦练内力、提高身手了。联想到这几年的帮派斗争，自己早已心力交瘁，又感到十分愧对自己的恩师"label大师"，无法担当掌门的重任，遂弃掌门职位不坐，浪迹江湖，过起了隐居生活，当然，他没忘了在街边买上一把左轮手枪防身……

VPN

一、隐身术

江湖中除了IP、ATM等几个传统大派别之外，武林中还有一部分人醉心于修练一种"隐身术"，他们的领地通常四处分散，中间必须经过其他帮派（主要是IP）的地盘，为了免交养路费，在江湖中行走时如果经过IP的领地，便打扮成IP帮的弟子模样，到了本帮的领地，再去掉伪装，恢复本来面目。这些人自称为VPN,掌门为"虚通道长",手下的两个堂主分别是："Overlay VPN"和"Peer-to-Peer VPN"

二、VPN中的角色

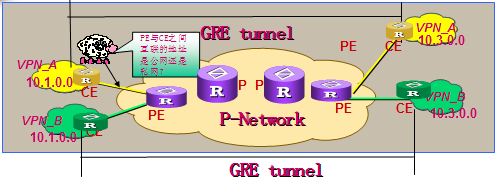
[](http://photo.blog.sina.com.cn/showpic.html)  
CE（Custom Edge）：直接与服务提供商相连的用户设备。

PE（Provider Edge Router）：指骨干网上的边缘路由器，与CE相连，主要负责VPN业务的接入。

P （Provider Router）：指骨干网上的核心路由器，主要完成路由和快速转发功能。

由于网络规模不同，网络中可能不存在P路由器。PE路由器也可能同时是P路由器。

三、Overlay VPN－隧道建立在CE上

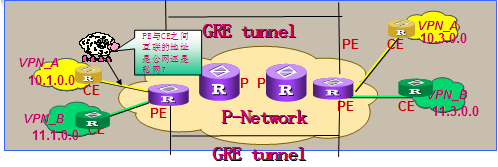
[](http://photo.blog.sina.com.cn/showpic.html)

特点：在CE与CE之间建立隧道，并直接传递路由信息，路由协议数据总是在客户设备之间交换，服务商对客户网络结构一无所知。典型代表是GRE、IPSec

优点：不同的客户地址空间可以重叠，保密性、安全性非常好。

缺点：需要客户自己创建并维护VPN。通常客户不愿意，也没有这个能力。

四、Overlay VPN－隧道建立在PE上

[](http://photo.blog.sina.com.cn/showpic.html)  
特点：在PE上为每一个VPN用户建立相应的GRE隧道，路由信息在PE与PE之间传递，公网中的P设备不知道私网的路由信息。

优点：客户把VPN的创建及维护完全交给服务商，保密性、安全性比较好。

缺点：不同的VPN用户不能共享相同的地址空间，即使可以共享，则PE与CE之间的地址、tunnel之间的地址一定不能相同，并且必须使用大量的ACL和策略路由。在实际中不具备可行性。

五、Overlay VPN的本质

Overlay VPN的本质是一种"静态"VPN，这好比是静态路由，所以他具有类似静态路由的全部缺陷：

1.     所有的配置与部署都需要手工完成，而且具有N^2问题：如果某个客户的VPN中新增了一个结点，则需要完成如下工作

•       在这个新增结点上建立与所有已存在的N个结点的隧道及相关的路由。

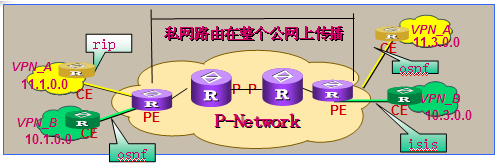
•       对于已存在的N个结点，需要在每个结点上都建立一个与新增结点之间的隧道及相关的路由。

2.     由于是"静态"VPN，则无法反应网络的实时变化。而且，如果隧道建立在CE上，则必须由用户维护，如果建立在PE上，则又无法解决地址冲突问题。

六、Peer-to-Peer VPN

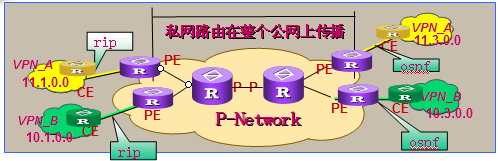
如同静态路由一样，所有具有"静态"性质的东西都不太适合大规模的应用和部署，难以担当重任。所以，首先要解决的问题就是将VPN的部署及路由发布变为动态性。Peer－to－Peer VPN的产生就是源于这种思想。这里的 Peer－to－Peer是指CE－to－PE，也就是要在CE与PE之间交换私网路由信息，然后由PE将这些私网路由在P－Network中传播（P-Network上肯定是运行了一种动态路由协议），这样这些私网路由会自动的传播到其他的PE上。这种VPN由于私网路由会泄露到公网上，所以必须严格的通过路由来控制，即：要确保同一个VPN的CE路由器上只能有本VPN的路由。所以，通常CE与PE之间运行的路由协议，与P-Network上运行的路由协议是不同的，即使相同，也要有很好的路由过滤和选择的机制。

七、Peer-to-Peer VPN——共享PE方式

[](http://photo.blog.sina.com.cn/showpic.html)  
所有VPN用户的CE都连到同一台PE上，PE与不同的CE之间运行不同的路由协议（或者是相同路由协议的不同进程，比如OSPF）。由路由始发PE将这些路由发布到公网上，在接收端的PE上将这些路由过滤后再发给相应的CE设备。

缺点：为了防止连接在同一台PE上的不同CE之间互通，必须在PE上配置大量的ACL。

八、Peer-to-Peer VPN——专用PE方式

[](http://photo.blog.sina.com.cn/showpic.html)  
为每一个VPN单独准备一台PE路由器，PE和CE之间可以运行任意的路由协议，与其他VPN无关。PE与P之间运行BGP，并使用路由属性进行过滤。

优点：无需配置任何的ACL了。

缺点：每一个VPN用户都有新增一台专用的PE，代价过于昂贵了。

九、Peer-to-Peer VPN的本质

Peer-to-Peer VPN虽然很好的解决了"静态的问题"，但是仍旧有很多局限性：

•       由于没有使用隧道技术，导致私网路由泄露到公网上，安全性很差。

•       VPN的"私有"特性完全靠路由来保证，导致在CE设备上无法配置缺省路由。（why？）

•       仍旧存在所有的设备无法共享相同的地址空间问题。

如果要确保安全性，则必须使用隧道技术，虽然本帮并不缺少隧道，但如GRE、IPSec都已被证实由于其"静态性"无法委以重任。而地址冲突的问题根本就不是本帮的势力范围，更是无法解决。

至此VPN帮已经黔驴技穷，好在掌门"虚通道长"是个留洋多年的"海龟"，思想很开放，觉得这个问题的解决需要整个武林一起出力。于是贴出了一张"招贤榜"——

十、招贤榜

为了尊重报文的隐私，提高我华夏的人权水准，大力推动网络私有化的进程，特向各位武林高手招贤纳士。如有能解决如下问题的好汉，无论出身、派别，皆可得千金重赏，并与本帮结为友好邻邦，共举VPN大业。

•        可以提供一种动态建立的隧道技术。

•        可以解决不同VPN共享相同地址空间的问题。

                     VPN 掌门

虚通道长敬上

十一、重赏之下，必有勇夫

话说招贤榜一贴出来，立刻轰动了整个武林。一日，众多武林中人正围着一张榜议论纷纷，忽然人群中一个腰挂左轮手枪，状如乞丐者抚掌大笑，口中念道，"嗌~~~~~~,中了！中了！"，言迄休克倒地。众人急忙将其救醒。此人醒来之后，揭下招贤榜，发足狂奔，喊道"兴邦有望！兴邦有望！……"

没错，当然是MPLS隐居的掌门PHP了，"可以提供一种动态建立的隧道技术"，MPLS中的LSP正是一种天然的隧道，而且这种隧道的建立是基于LDP协议，又恰恰是一种动态的标签生成协议。舍我其谁！！

自从PHP揭了第一张招贤榜之后，江湖上纷纷猜测会由谁来搞掂第二个问题，大家普遍认为最佳人选应该在几个路由协议中产生。

MP-BGP

一、为什么是BGP

如果要解决地址冲突问题，必须对现有的协议进行大规模的修改，这就要求一个协议具有良好的可扩展性。而具备条件的协议一定是基于TLV元素的。符合标准的只有EIGRP、BGP、ISIS。

•       ISIS本不是中土人士，前年刚刚从OSI逃荒过来，帮中弟兄都不会说IP语，而说NSAP语，目前连户口还没解决，估计无暇他顾了。

•       EIGRP向来闭关锁帮，夜郎自大的认为本帮的功夫天下一流，从不与别人切磋，也不参加武林大会。而且看见别人的武功与自己有几分相似，便跳出来要与他打官司。在江湖上名声臭极。

•       而BGP看来是十分合适的人选：

1.     网络中VPN路由数目可能非常大，BGP是唯一支持大量路由的路由协议；

2.     BGP是基于TCP来建立连接，可以在不直接相连的路由器间交换信息，这使得P路由器中无须包含VPN路由信息；

3.     BGP可以运载附加在路由后的任何信息，作为可选的BGP属性，任何不了解这些属性的BGP路由器都将透明的转发它们，这使在PE路由器间传播路由非常简单。

BGP的掌门叫——无为长老，是位得道的高僧。

二、无为长老

话说当年IP的掌门人——"尽力而为"，自知年事已高，便想在帮中选择下一任接班人。在众多弟子，唯有两个最为得意——OSPF和BGP。一日，掌门将二人叫到面前，让他们说一下这些年的修行心得。

OSPF念道："身如路由器，心似转发表，报文何其多，日夜勤查找。"

BGP念道：  "路由本非器，转发何需表？报文虽然多，自有他人找。"

"尽力而为"抚掌大笑曰："BGP得吾真传也！"，于是将衣钵传给了他。

OSPF很不服气，说："弟子日夜辛劳，编撰的OSPF心法一共300多页，构思精妙，算法复杂，堪称武林绝学。而师弟BGP终日游山玩水，草草写了一本70多页的心得就交差了。"掌门笑问他"那你的心法一共可以管理多少台路由器，多少条路由呢？"OSPF答道："设备百台，路由千条。"掌门又问BGP："那么你呢？"BGP道："整个internet百万台路由器，十余万路由皆由弟子一人掌管。"掌门笑道："BGP虽然表面看来无所作为，其实他只是不想过分的拘泥于细节，实际上是"无为而治"啊。我当年就是靠的这一点才将来势凶猛的ATM斩于马下。"

从此BGP便在武林中得到了"无为长老"的雅号。

三、冥思苦想

无为长老虽然觉得此事责无旁贷，但确实非常麻烦，要想解决地址冲突的问题，至少有如下三个难题需攻克：

1.     本地路由冲突问题，即：在同一台PE上如何区分不同VPN的相同路由。

2.     路由在网络中的传播问题，两条相同的路由，都在网络中传播，对于接收者如何分辨彼此？

3.     报文的转发问题，即使成功的解决了路由表的冲突，但是当PE接收到一个IP报文时，他又如何能够知道该发给那个VPN？因为IP报文头中唯一可用的信息就是目的地址。而很多VPN中都可能存在这个地址。

四、计上心来

无为闭关修练了数月，冥思苦想了很久，渐渐有了些思路：

1.     本地路由冲突问题，可以通过在同一台路由器上创建不同的路由表解决，而不同的接口可以分属不同的路由表中，这就相当于将一台共享PE模拟成多台专用PE。

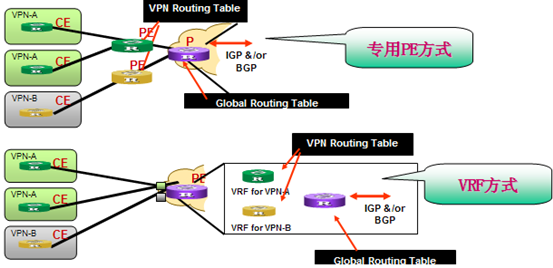
2.     可以在路由传递的过程中为这条路由再添加一个标识，用以区别不同的VPN。

3.     由于IP报文的格式不可更改，估计指望不上他了，但可以在IP头之外加上一些信息，由始发的VPN打上标记，这样PE在接收报文时可以根据这个标记进行转发。

虽然大致的解决方案已有，但要做到可以具体实施，却还有很多工作要做。

五、理论突破——VRF

其实解决地址冲突的问题，也存在一些方法：使用ACL、IP unnumber、NAT。但这些办法都是基于"打补丁"的思想，没能从本质上解决问题。要想彻底解决，必须在理论上有所突破。可以从专用PE上得到启示。专用路由器方式分工明确，每个PE只保留自己VPN的路由。P只保留公网路由。而现在的思路是：将这些所有设备的功能，和在一台PE上完成。

[](http://photo.blog.sina.com.cn/showpic.html)  
六、VRF

VRF---VPN路由转发实例（VPN Routing & Forwarding Instance）：每一个VRF可以看作虚拟的路由器，好像是一台专用的PE设备。该虚拟路由器包括如下元素：

•       一张独立的路由表，当然也包括了独立的地址空间。

•       一组归属于这个VRF的接口的集合。

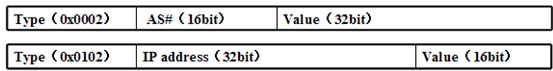
•       一组只用于本VRF的路由协议。

对于每个PE，可以维护一个或多个VRF，同时维护一个公网的路由表（也叫全局路由表），多个VRF实例相互分离独立。其实实现VRF并不困难，关键在于如何在PE上使用特定的策略规则来协调各VRF和全局路由表之间的关系。

七、RT

我们回忆一下，其实在专用PE的方式中，已经很好的解决了这个问题。当时使用了BGP的community属性。这次仍旧使用这个思路，只不过"旧瓶装新酒"把community扩展了一下，并且起了一个新名字：RT（Route Target）。

扩展的community有如下两种格式：其中type字段为0x0002或者0x0102时表示RT。

[](http://photo.blog.sina.com.cn/showpic.html)  
八、RT的本质

RT的本质是每个VRF表达自己的路由取舍及喜好的方式。可以分为两部分：Export Target与import Target；前者表示了我发出的路由的属性，而后者表示了我对那些路由感兴趣。例如：

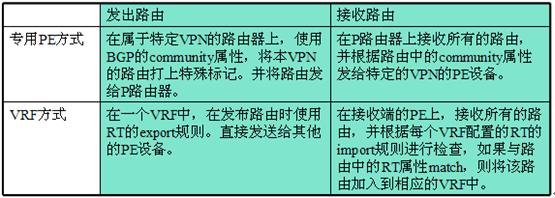
SITE-A：我发的路由是红色的，我也只接收红色的路由。

SITE-B：我发的路由是红色的，我也只接收红色的路由。

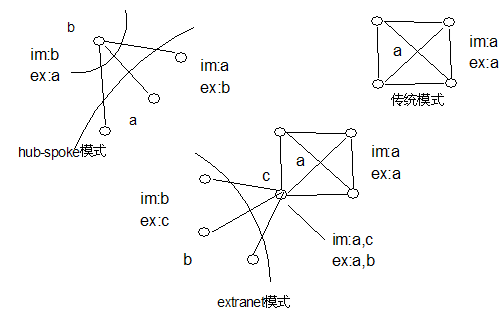
SITE-C：我发的路由是黑色的，我也只接收黑色的路由。

SITE-D：我发的路由是黑色的，我也只接收黑色的路由。

这样，SITE－A与SITE-B中就只有自己和对方的路由，两者实现了互访。同理SITE－C与SITE-D也一样。这时我们就可以把SITE-A与SITE－B称为VPN-A，而把SITE-C与SITE-D称为VPN-B。

[](http://photo.blog.sina.com.cn/showpic.html)  
九、RT的灵活应用

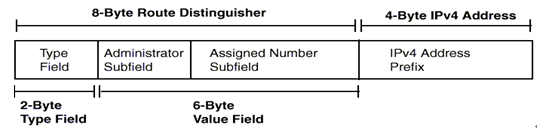
由于每个RT Export Target与import Target都可以配置多个属性，例如：我对红色或者蓝色的路由都感兴趣。接收时是"或"操作，红色的、蓝色的以及同时具备两种颜色的路由都会被接受。所以就可以实现非常灵活的VPN访问控制。

[](http://photo.blog.sina.com.cn/showpic.html)  
十、RD(Route Distinguisher)

在成功的解决了本地路由冲突的问题之后，路由在网络中传递时的冲突问题就迎刃而解了。只要在发布路由时加上一个标识即可。

既然路由发布时已经携带了RT，可否就使用RT作为标识呢？

理论上讲，肯定是可以的。但RT不是一个简单的数字，通常是一个列表，而且他是一种路由属性，不是与IP前缀放在一起的，这样在比较的时候不好操作。特别是：BGP的Route withdraw报文不携带属性，这样在这种情况下收到的路由就没有RT了。所以还是另外定义一个东西比较好，这个东东就叫做 RD。他的格式与RT基本上一样。

[](http://photo.blog.sina.com.cn/showpic.html)  
十一、RD的本质

在IPv4地址加上RD之后，就变成VPN-IPv4地址族了。

理论上可以为每个VRF配置一个RD。通常建议为每个VPN都配置相同的RD，不同的VPN配置不同的RD。但是实际上只要保证存在相同地址的两个VRF的RD不同即可，不同的VPN可以配置相同的RD，相同的VPN也可以配置不同的RD。如果两个VRF中存在相同的地址，则一定要配置不同的RD，而且两个VRF一定不能互访，间接互访也不成。

同一台PE上的不同VRF不能配置相同的RD（why？）。

RD并不会影响不同VRF之间的路由选择以及VPN的形成，这些事情由RT搞定。

PE从CE接收的标准的路由是IPv4路由，如果需要发布给其他的PE路由器，此时需要为这条路由附加一个RD。

VPN-IPv4地址仅用于服务供应商网络内部。在PE发布路由时添加，在PE接收路由后放在本地路由表中，用来与后来接收到的路由进行比较。CE不知道使用的是VPN-IPv4地址。

在其穿越供应商骨干时，在VPN数据流量的包头中没有携带VPN-IPv4地址。

十二、革命尚未成功

至此，前两个问题：在PE本地的路由冲突和网络传播过程的冲突都已解决。但是如果一个PE的两个本地VRF同时存在10.0.0.0/24的路由，当他接收到一个目的地址为10.0.0.1的报文时，他如何知道该把这个报文发给与哪个VRF相连的CE？肯定还需要在被转发的报文中增加一些信息。

既然路由发布时已经携带了RD，可否就使用RD作为标识呢？

理论上讲肯定是可以的。但是RD一共有64个bit，太大了。这会导致转发效率的降低。所以只需要一个短小、定长的标记即可。由于公网的隧道已经由MPLS来提供，而且MPLS支持多层标签的嵌套，这个标记定义成MPLS标签的格式。这个私网的标签就由MP-BGP来分配，与私网的路由一同发布出去。

十三、概念总结

VRF：在一台PE上虚拟出来的一个路由器，包括一些特定的接口，一张路由表，一个路由协议，一个RD和一组RT规则。

RT：表明了一个VRF的路由喜好，通过他可以实现不同VRF之间的路由互通。他的本质就是BGP的community属性。

RD：为了防止一台PE接收到远端PE发来的不同VRF的相同路由时不知所措，而加在路由前面的特殊信息。在PE发布路由时加上，在远端PE接收到路由后放在本地路由表中，用来与后来接收到的路由进行比较。

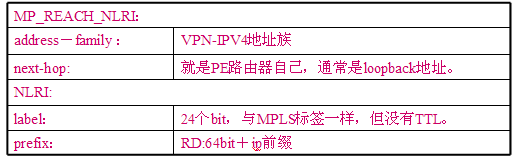
Label：为了防止一台PE接收到远端PE发给本地不同VRF的相同地址的主机时不知所措，而加在报文前面的特殊信息。由本地PE在发布路由时加上，远端PE接收到保存在相应的VRF中。

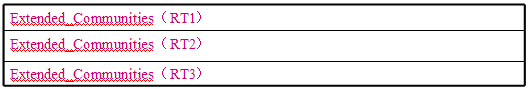
SITE：一个VRF加上与其相连的所有的CE的集合。

VPN：是一些SITE的集合，这些SITE由于共享了相同的路由信息可以互通。

十四、BGP发布路由时需要携带的信息

一个扩展之后的NLRI（Network Layer Reachability Information），增加了地址族的描述，以及私网label和RD。

[](http://photo.blog.sina.com.cn/showpic.html)  
跟随之后的是RT的列表

[](http://photo.blog.sina.com.cn/showpic.html)  
对于使用了扩展属性MP\_REACH\_NLRI的BGP，我们称之为MP-BGP。

BPG/MPLS VPN

一、宴桃园豪杰三结义

MPLS掌门PHP与BGP掌门无为一起来到VPN掌门虚通道长处商议结盟之事，宾主谈笑甚欢。三人皆是性情中人，胸怀坦荡，知无不言，感觉甚是投机。于是效仿古人，结拜兄弟。祭拜天地之后，序了长幼，无为年长，做大哥，PHP次之，虚通道长年幼，做了小弟。

三个掌门在本帮中精选得力弟子，成立JV公司，各占三分之一的股份。新帮派命名为：BGP/MPLS VPN。并且详细规定了新帮派的各项规章制度，并昭示天下。一时在江湖中传为佳话。

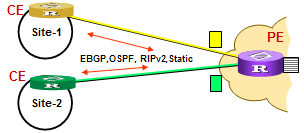
虽然成立了新帮派，但是三个掌门并没闲着。

MPLS继续潜心钻研QOS和流量工程；

BGP准备和IPv6以及组播成立新公司；

而虚通道长则致力于扩大VPN家族的势力范围。

二、CE与PE之间如何交换路由

[](http://photo.blog.sina.com.cn/showpic.html)  
VRF在PE上配置。

PE 维护独立的路由表，包括公网和私网(VRF)路由表

•       公网路由表：包含全部PE和P 路由器之间的路由，由骨干网IGP产生。

•       私网路由表：包含本VPN用户可达信息的路由和转发表。

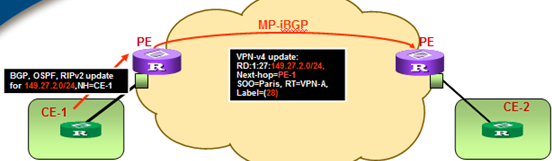
PE 和 CE 通过标准的EBGP、OSPF、RIP或者静态路由交换路由信息。

•       静态路由、RIP都是标准的协议，但是每个VRF运行不同的实例。相互之间没有干扰。与PE的MP-iBGP之间只是的redistribute操作。

•       EBGP也是普通的EBGP，而不是MP-EBGP，只交换经过PE过滤后的本VPN路由。

•       OSPF则做了很多修改，可以将本site的LSA放在bgp的扩展community属性中携带，与远端VPN中的ospf之间交换LSA。每个site中的OSPF都可以存在area 0，而骨干网则可以看作是super area 0。此时的OSPF由两极拓扑（骨干区域＋非骨干区域）变为3级拓扑（超级骨干区域＋骨干区域＋非骨干区域）

三、VRF路由注入到MP-iBGP

[](http://photo.blog.sina.com.cn/showpic.html)  
PE路由器需要对一条路由进行如下操作：

•       加上RD（RD为手工配置），变为一条VPN-IPV4路由。

•       更改下一跳属性为自己（通常是自己的loopback地址）

•       加上私网标签（随机自动生成，无需配置）

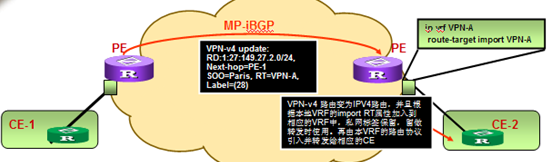
•       加上RT属性（RT需手工配置）

发给所有的PE邻居。

为何要更改下一跳属性？

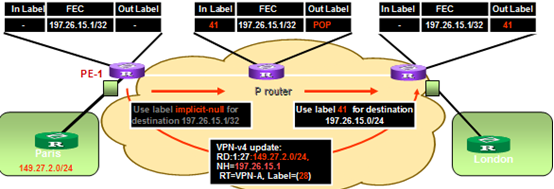
携带RT的export还是import属性？

四、MP-iBGP路由注入到VRF

[](http://photo.blog.sina.com.cn/showpic.html)  
VPN-v4 路由变为IPV4路由，并且根据本地VRF的import RT属性加入到相应的VRF中，私网标签保留，留做转发时使用。再由本VRF的路由协议引入并转发给相应的CE。

这条VPN路由的下一跳是谁？

五、公网标签分配过程

[](http://photo.blog.sina.com.cn/showpic.html)

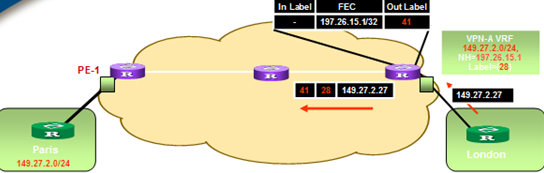
    •       PE和P路由器通过骨干网IGP学习到BGP邻居下一跳的地址。

•       通过运行LDP协议，分配标签，建立LSP通道。

•       标签栈用于报文转发，外层标签用来指示如何到达BGP下一跳 ，内层标签表示报文的出接口或者属于哪个VRF（属于哪个VPN）。

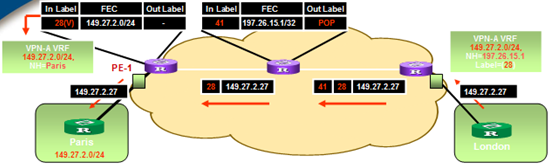
•       MPLS 节点转发是基于外层标签，而不管内层标签是多少。

六、报文转发——从CE到Ingress PE

[](http://photo.blog.sina.com.cn/showpic.html)  
•       CE将报文发给与其相连的VRF接口，PE在本VRF的路由表中进行查找，得到了该路由的公网下一跳地址（即：对端PE的loopback地址）和私网标签。

•       在把该报文封装一层私网标签后，在公网的标签转发表中查找下一跳地址，再封装一层公网标签后，交与MPLS转发。

七、Ingress PE－>Egress PE－>CE

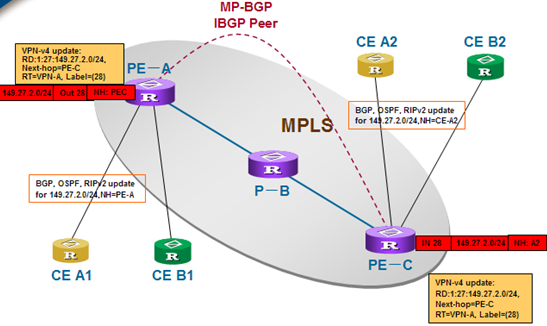
[](http://photo.blog.sina.com.cn/showpic.html)  
•       该报文在公网上沿着LSP转发，并根据途径的每一台设备的标签转发表进行标签交换。

•       在倒数第二跳处，将外层的公网标签弹出，交给目的PE设备

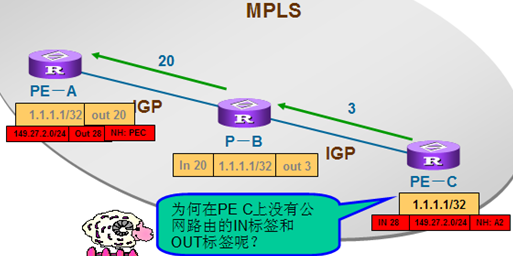
•       PE设备根据内层的私网标签判断该报文的出接口和下一跳。

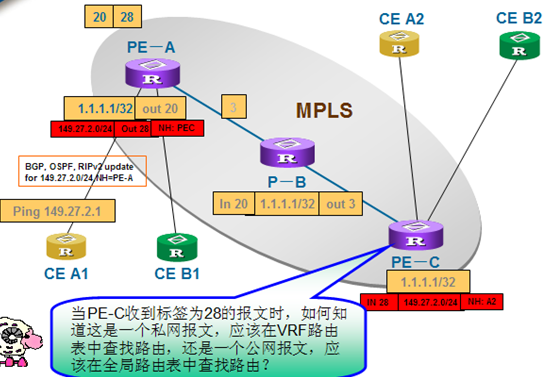
•       去掉私网标签后，请报文转发给相应的VRF中的CE。

八、MPLS VPN控制流程－"私网"路由及标签传递

[](http://photo.blog.sina.com.cn/showpic.html)  
九、MPLS VPN控制流程－"公网"LSP的建立

PE C的loopback地址为1.1.1.1

[](http://photo.blog.sina.com.cn/showpic.html)  
十、MPLS VPN数据流程－私网数据包的转发

[](http://photo.blog.sina.com.cn/showpic.html)  
十一、MPLS/VPN招募新弟子入门考试试题

1.     在MPLS/VPN中公网标签是由（）分配的，私网标签是由（）分配的。

2.     在CE上需要运行（），在PE上需要运行（），在P上需要运行（）

       A. 普通的路由协议            B.MP-BGP    C.MPLS

1.     RD是（），RT是（），私网标签是（），公网标签是（）                           A.手工配置的                     B.随机生成的

2.     虽然运行MPLS协议后，路由器会自动为公网路由表中的所有路由分配标签，但实际上，只需要为所有PE的loopback地址分配标签即可，不必为其他的任何公网路由分配标签。（）T or F

3.     虽然建议为不同的VPN配置不同的RD，相同的VPN配置相同的RD。但根本就别理他，只要保证存在相同地址的两个VRF的RD不同即可，不同的VPN可以配置相同的RD，相同的VPN也可以配置不同的RD。（）T or F

4.     既然已经定义了RD，就不可能存在两条相同的路由同时在网络中传播。（）

5.     对于一台PE，可能会出现接收到的不同目的地址的报文具有相同的私网标签，不可能会出现发送的不同目的地址的报文具有相同的私网标签（）

6.     一台运行MPLS的路由器如何知道自己相对于每个LSP是倒数第二跳，又如何知道自己是倒数第一跳？

思考题：如果是BGP/GRE VPN，他的运行模式又是如何的？

答对5题者，可入我帮，答对思考题者，可直接升为堂主。

配置

1.1、MPLS的配置

全局模式下：

Lsr的ID，可以配置成与router id相同。

mpls lsr id *10.5.80.250*

!

在全局模式启动LDP协议

mpls ldp

!

在接口上启动LDP Session

interface Ethernet4/1/0

ip address 10.5.3.93 255.255.255.252

mpls ldp enable

1.2、查看MPLS的邻居状态

PE3\_NE16#show mpls ldp session

Showing information about all sessions:

   Peer LDP Ident: 192.168.255.38:0; Local LDP Ident: 220.163.42.126:3

   Tcp connection:192.168.255.38 - 220.163.42.66

   Session State: Operational

   Session Role: Active   Hello packets sent/received: 72121/82424

   KeepAlive packets sent/received: 15018/20607

   Negotiated Keepalive Timer Value: 60   Peer PV Limit: 0

   LDP discovery source:GigabitEthernet4/1/0.1

1.3、查看MPLS的标签分配情况

NCC-R#    show mpls lsp brief

ID     I/O-Label       In-Interface         Prefix/Mask        Next-Hop

22     382/264         VT20                 10.5.61.250/32    10.5.3.94

23     388/266         VT20                 10.5.37.250/32    10.5.3.94

24     408/274         VT20                 10.5.32.250/32    10.5.3.94

25     ---/24          ----------           10.5.22.250/32    10.5.3.10

26     132/24          VT49                 10.5.22.250/32     10.5.3.10

27     153/24          Eth4/1/0             10.5.22.250/32    10.5.3.10

28     155/24          Eth10/2/0            10.5.22.250/32    10.5.3.10

29     ---/20          ----------           10.5.23.250/32    10.5.3.10

30     186/20          VT49                 10.5.23.250/32     10.5.3.10

31     229/20          Eth4/1/0             10.5.23.250/32    10.5.3.10

2、MP-BGP配置

2.1、PE上的配置

VRF配置：

ip vrf *VPN-HW*

创建一个VRF并命名。同时进入vrf配置模式

RD配置：

在VRF模式下，每个VRF配置一个RD，建议相同的VPN配置相同的RD。

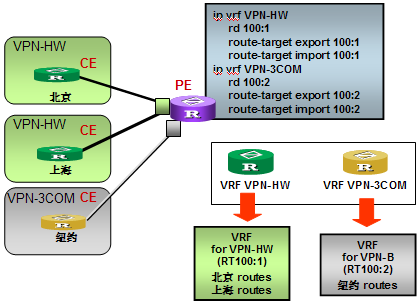
 rd *100:1*

RT配置：

在VRF模式下，每个VRF配置不同的RT列表，如果只要一个RT,建议与RD配成相同。

 route-target import *100:1*

 route-target export *100:1*

[](http://photo.blog.sina.com.cn/showpic.html)  
2.2、将VRF与接口关连

在与某个VPN相连的接口下配置如下命令：

ip vrf forwarding *VPN-HW*

interface Serial3/5

 ip vrf forwarding *VPN-HW*

 ip address 10.168.61.6 255.255.255.252

 encapsulation ppp

2.3、PE与CE之间的路由协议

目前支持：RIPv2、BGP、Static，每个协议都被改造成多实例的了，换句话说，就是"VRF化"了。

Static：

ip route vrf *VPN-HW 10.10.1.0 255.255.255.0 10.10.1.2*

ip route vrf *VPN-3COM 10.10.1.0 255.255.255.0 10.10.1.2*

RIPv2:

       router rip

      !

       address-family ipv4 vrf *VPN-HW*

       no  auto-summary

      network 10.0.0.0

       !

       address-family ipv4 vrf *VPN-3COM*

       no  auto-summary

      network 10.0.0.0

BGP：

router bgp 109

!

 address-family ipv4 vrf VPN-HW

 neighbor 10.168.62.5 remote-as 65503

 exit-address-family

 !

 address-family ipv4 vrf VPN-3C0M

 redistribute static

 redistribute connected

 redistribute rip

 exit-address-family

!

2.4、MP-BGP的配置

router bgp 30000

 no synchronization

 neighbor 10.5.80.240 remote-as 30000

 neighbor 10.5.80.240 update-source LoopBack0

address-family ipv4 vrf *VPN-HW*

 redistribute connected

 redistribute static

 no synchronization

 exit-address-family

 address-family vpnv4

 neighbor 10.5.80.240 activate

 exit-address-family

调试命令

一、查看VPN的路由

PE3\_NE16#show ip route vrf VPN-HW

  VPN-HW Route Information

 Routing Table:  VPN-HW    RD:   65400:1

Destination/Mask   Proto  Pre Metric      Nexthop         Interface

        1.1.1.1/32 BGP    170 0           220.163.42.62   LoopBack0

   192.168.20.0/29 BGP    170 0           220.163.42.62   LoopBack0

   192.168.20.0/30 BGP    170 0           220.163.42.62   LoopBack0

  192.168.20.65/32 DIRECT 0   0           127.0.0.1       InLoopBack0

  192.168.20.96/29 DIRECT 0   0           192.168.20.101  GE4/1/0.2

 192.168.20.101/32 DIRECT 0   0           127.0.0.1       InLoopBack0

对于路由表中的BGP路由，下一跳地址是对端PE的loopback地址，出接口则是自己的loopback接口。

二、查看BGP的VPN路由

PE3\_NE16#show ip bgp  vpnv4 all

BGP local router ID is 220.163.42.126

Status codes: s suppressed, d damped, h history, \* valid, > best, i internal

Origin codes: i - IGP, e - EGP, ? - incomplete

    Network            Next Hop          Label(I/O)  Metric    LocPrf     Path

 Route Distinguisher:65400:1   (default for vrf vpna)

\*>i 1.1.1.1/32         220.163.42.62        0/17               100         ?

\*>i 192.168.20.0/29    220.163.42.62        0/17                100         ?

\*>i 192.168.20.0/30    220.163.42.62        0/16                100         ?

\*>  192.168.20.65/32   0.0.0.0              19/0                            ?

\*>  192.168.20.96/29   0.0.0.0              18/0                            ?

此命令用来查看BGP学习到的VPNv4路由的具体信息，以及私网标签的分配情况。特别是本地始发的路由（next hop 0.0.0.0）的标签分配情况，只能通过本命令查看。

三、查看私网标签命令

PE3\_NE16#show  mpls lsp vrf brief

ID     I/O-Label       In-Interface         Prefix/Mask        Next-Hop

1      ---/141|17      ----------           1.1.1.1/32        220.163.42.62

2      ---/141|17      ----------           192.168.20.0/29   220.163.42.62

3      ---/141|16      ----------           192.168.20.0/30   220.163.42.62

     3 Record(s) Found

此命令只可以查看学习到的BGP路由的私网标签情况，对于本地始发的路由无法查看，必须通过命令 show ip bgp  vpnv4 all查看。

四、Ping&Telnet&tracert

由于现在一台PE上存在多张路由表了，所有针对VPN路由的ping、telnet、tracert等常用命令，必须加上vrf参数，而且最好加上－a参数，指明源地址。

PE3\_NE16#ping -vrf vpna -a 192.168.20.65 1.1.1.1

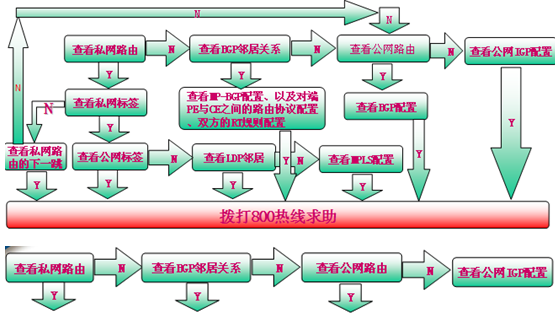
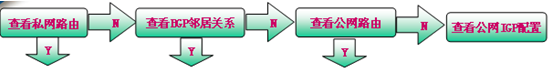
PE3\_NE16#tracert -vrf vpna -a 192.168.20.65 1.1.1.1

PE3\_NE16#telnet vrf vpna 1.1.1.1

Trouble shooting

一、MPLS/VPN的trouble-shooting

由于MPLS/VPN的报文转发是基于LSP，而LSP是依附于路由的。所以定位故障的思路是：先查路由、再查标签；先查私网、再查公网。

[](http://photo.blog.sina.com.cn/showpic.html)  
[](http://photo.blog.sina.com.cn/showpic.html)  
查看私网路由：

分别查看两端PE路由器的VRF中是否存在对端PE的VRF路由

命令： show ip route vrf

查看BGP邻居关系：

邻居状态机是否达到Established状态

命令：show ip bgp summary

Neighbor        V   AS     MsgRcvd    MsgSent   OutQ  Up/Down   State

220.163.42.62   4 65400       6818       6895     0    14h48m Established

查看公网路由：

是否在公网LSP途径的所有设备上都存在对端PE的loopback地址的精确路由？（必须是32位mask）

查看公网IGP配置：

是否通过IGP将PE的loopback地址的路由发布出去

为什么沿途设备上都必须是精确路由？

[http://www.madhex.com/wp-content/uploads/2016/05/053116_1211_BGPMPLSVPN35.png](http://photo.blog.sina.com.cn/showpic.html)  
查看私网标签：

查看本端PE路由器的私网标签是否为对端PE所分配，相关命令：

•                               本端    show  mpls lsp vrf brief

ID     I/O-Label       In-Interface         Prefix/Mask        Next-Hop

•                               ---/141|17      ----------           1.1.1.1/32        220.163.42.62

•                               对端 show ip bgp  vpnv4 all

Network            Next Hop          Label(I/O)  Metric     LocPrf     Path

\*> 1.1.1.1/32         0.0.0.0        17/0                   ?

查看MP-BGP配置、以及对端PE与CE之间的路由协议配置、双方的RT规则配置：

address-family ipv4 vrf *VPN-HW*

 redistribute connected

exit-address-family

 address-family vpnv4

 neighbor 10.5.80.240 activate

查看BGP配置：

查看普通BGP的配置，是否正确的配置了BGP邻居。

[http://www.madhex.com/wp-content/uploads/2016/05/053116_1211_BGPMPLSVPN36.png](http://photo.blog.sina.com.cn/showpic.html)  
查看私网路由的下一跳：

查看本地的公网路由表中是否存在私网路由的下一跳（即：对端PE的loopback地

址）的精确路由？（必须是32位mask）

查看公网标签：

查看整个LSP上的所有设备是否已经为两个PE的loopback地址正确的分配了

公网标签，相关命令：show mpls lsp brief 每台设备的入标签是否为其下

一跳的出标签。

ID     I/O-Label       In-Interface    Prefix/Mask        Next-Hop

22     382/264         VT20            10.5.61.250/32     10.5.3.94

查看LDP邻居：

查看两台相邻的PE或P路由器之间是否正确建立了LDPsession

相关命令：show mpls ldp session

   Session State: Operational

查看MPLS配置：

查看该设备是否在全局使能了MPLS，以及在相应的接口上使能了LDP。

全局命令：

mpls lsr id 10.5.80.250!

mpls ldp

在接口上启动LDP Session

interface Ethernet4/1/0

mpls ldp enable

<http://down.51cto.com/data/97614>