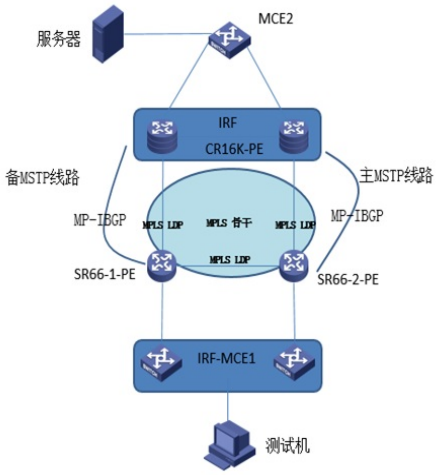
SR66路由器MPLS VPN ping 大包不通的经验案例

李长剑 2015-01-04 发表

SR66路由器MPLS VPN ping 大包不通的经验案例

一、 **组网**

SR66-1-PE路由器、SR66-2-PE路由器和CR16K-PE路由器属于AS号为65500，SR66-1-PE、SR66-2-PE和CR16K-PE之间建立MP-IBGP的邻居关系，通过OSPF来传递公网路由，三台设备之间分别跑MPLS LDP，建立LDP隧道，通过LDP分配公网标签。SR66-1-PE路由器和SR66-2-PE路由器分别通过子接口的方式和IRF-MCE1相连，之间运行OSPF多实例。通过控制IRF-MCE1下连测试机接口上的OSPF Cost控制测试机访问服务器的路径，测试机和服务器属于相同的VPN。SR66-1-PE路由器和SR66-2-PE路由器使用的线卡为FIP-110，子卡为MIM-2GE板卡，SR66-1-PE路由器和SR66-2-PE路由器之间互联使用的是MIM-2GE子卡，SR66-1-PE路由器和SR66-2-PE路由器与CR16K-PE路由器之间互联使用的是F11O自带的千兆接口，SR66-1-PE路由器和SR66-2-PE路由器与MCE1之间互联使用的是MIM-2GE子卡。测试机访问服务器走主MSTP线路，当主MSTP线路down掉后，业务切换到备份的MSTP线路。



二、 问题描述

当SR66-2-PE路由器和CR16K-PE路由器之间的主用链路故障后，测试机通过备份链路访问服务器的时候大部分业务中断，现场使用测试机ping 服务器1468字节数据包能通，但是测试机ping服务器1469字节数据包不通。

三、 **过程分析**

通过对故障现象的分析，怀疑是MTU的导致的，问题分析过程如下：

1、首先

判断可能是报文超出了SR66路由器MIM卡接口最大帧长度，因此将SR66-1-PE路由器和SR66-2-PE路由器之间互联接口改为FIP 110自带的接口，然而调整接口后，仍然ping 1469字节大包不通。

2、其次

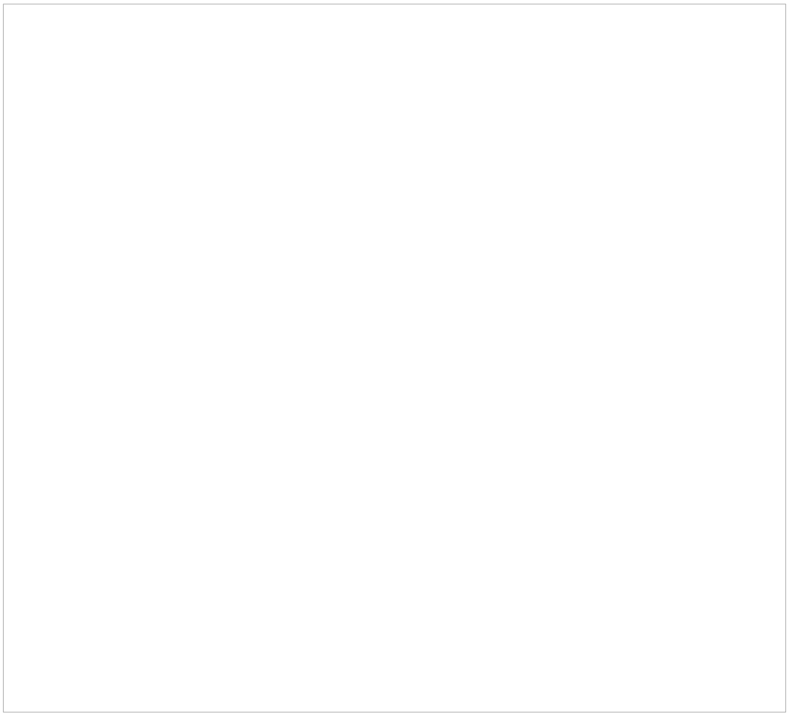
怀疑SR66-1-PE路由器和SR66-1-PE路由器之间MTU影响报文转发，尝试修改SR66-1-PE路由器和SR66-2-PE路由器之间的MPLS MTU 1400解决问题，但是修改后故障依旧，仍然ping 1469字节大包不通。

通过更换、修改了SR66-1-PE和SR66-2-PE之间的接口、MTU值后，故障依旧，那么超过1469字节的报文到底丢在哪里了呢？且看下问分解。

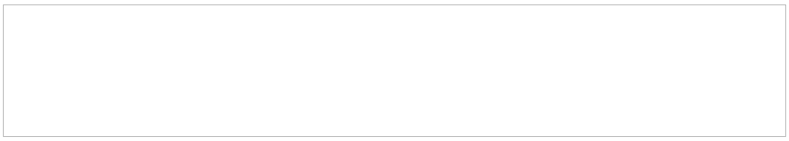
3、最后

详细的分析ICMP的报文的路径。

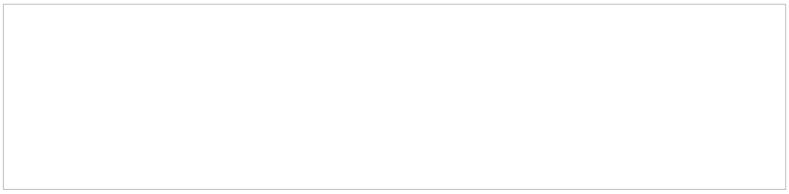
1）正常业务的路径图：



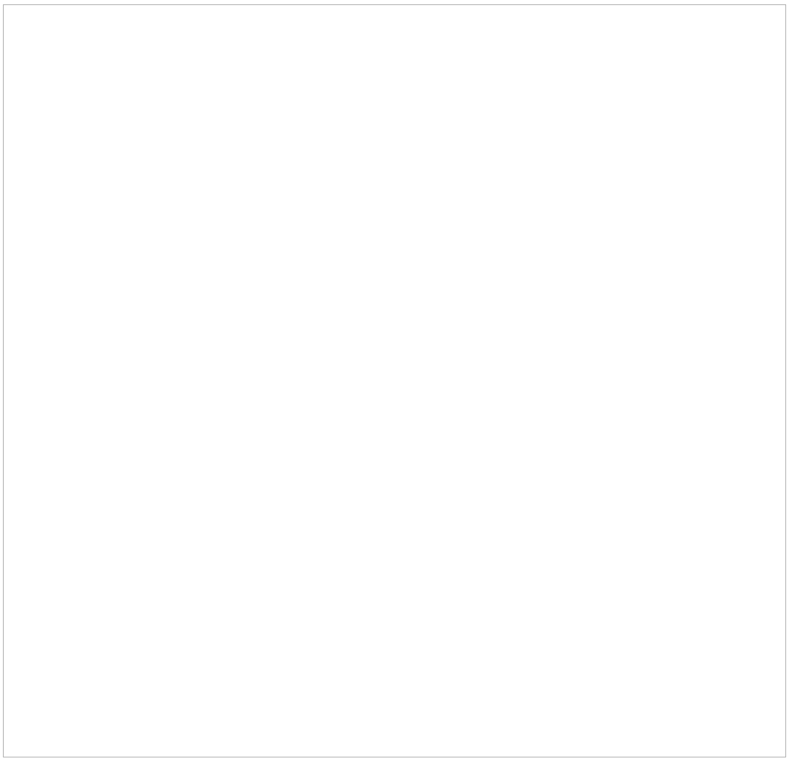
测试机访问服务器的报文路径：



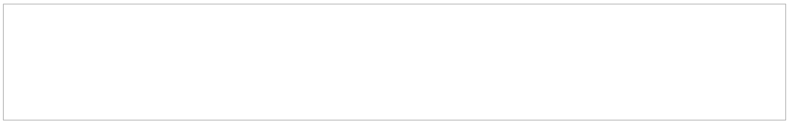
服务器回复测试机的报文路径：



2）当主的MSTP线路出现故障，备份线路的报文路径图：



测试机访问服务器的报文路径：

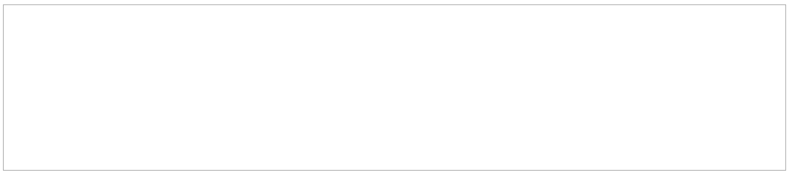


服务器回复测试机的报文路径：

3）变换路径测试报文路径图：

现场为了测试是否为单边设备、单边线路出现了问题，变换路径测试。
断开备份线路和MCE1与SR66-2-PE设备之间线路，让流量走反方向路径，发现测试结果：测试机ping服务器1468字节数据包能通，但是测试机ping服务器1469字节数据包不通。
测试机访问服务器的报文路径：

服务器回复测试机的报文路径：



4）分析帧大小

1、线路正常的时候SR66-2-PE路由器和CR16K-PE路由器之间数据报文分析如下：
主传输线路ICMP 的request请求报文长度1519字节：
1519=1469（数据载荷）+8(ICMP包头)+20(IP包头)+4（MP-BGP分配的私网标签）+18（以太网帧头）
主传输线路ICMP的reply应答报文长度1519字节：

1519=1469（数据载荷）+8(ICMP包头)+20(IP包头)+4（MP-BGP分配的私网标签）+18（以太网帧头）

此时，线路正常时，PE之间没有P设备，SR66-2-PE路由器作为PE设备PHP弹出公网标签，所以在MSTP传输线路上跑的数据包，只有一层私网标签。

2、SR66-2-PE路由器上行链路故障时，CR16K-PE路由器和SR66-1-PE路由器之间链路数据报文的大小分析如下：

备份传输线路上的ICMP 的request请求报文长度1519字节：

1519=1469（数据载荷）+8(ICMP包头)+20(IP包头)+4（MP-BGP分配的私网标签）+18（以太网帧头）

此时，SR66-1-PE路由器做为P设备，PHP弹出公网标签，所以在MSTP传输线路上跑的数据包，只有一层私网标签。

备份传输线路上的ICMP的reply应答报文长度1523字节：

1523=1469（数据载荷）+8(ICMP包头)+20(IP包头)+4（MP-BGP分配的私网标签）+4（LDP分配的公网标签）+18（以太网帧头）

此时，SR66-1-PE路由器做为P设备，CR16K-PE路由器作为PE设备发送过来的报文，在MSTP线路上跑的数据包携带了两层标签，公网标签和私网标签。

主MSTP线路故障的时候，CR16K-PE路由器返回ICMP的reply数据包比SR66-1-PE发送的ICMP的request数据报文多了4个字节的LDP公网标签。在SR66-1-PE上debug physical packet all interface GigabitEthernet 2/1/1（GigabitEthernet 2/1/1为和CR16K-PE路由器互联的接口），信息如下：

*Aug 26 07:48:07:975 2014 SR66-1-PE DPMPLS/7/MPLSFW PACKET: -Slot=2;

MPLS Output: Sending the packet to GE2/1/1, PktLen = 1505, Label(s) = 1146,9143, EXP = 0, TTL = 255!

*Aug 26 07:48:07:975 2014 SR66-1-PE DRVDBG/7/debugging: -Slot=2;

(GigabitEthernet2/1/1)PHY/PKT:

Packet Output, Packet Len =1519,Partial data as follows:

5C DD 70 4A 04 19 5C DD 70 4A 05 13 88 47 00 47

A0 FF 02 3B 71 FF 45 00 05 D9 1A 09 00 00 7E 01

BD FC **AC 10 01 01 C0 A8 01 01** 08 00 AF D7 00 01

07 5C 61 62 63 64 65 66 67 68 69 6A 6B 6C 6D 6E

SR66-1-PE路由器已经将ICMP的request报文发送出去，这点可以证明并非SR66-1-PE路由器导致的ping大包不通的故障现象。SR66-1-PE上面没有debug到服务器返回ICMP的reply的数据包；与此同时，在CR16K-PE路由器将和SR66-1-PE互联接口镜像抓包，能够看到ICMP的reply报文已经从接口发送出了，但是为什么SR66-1-PE没有收到呢?

经过上述的分析，到这里问题已经显而易见了，当主MSTP线路故障后，业务切换后，SR66-1-PE路由器的角色变为了P设备，那么CR16K-PE路由器返回的ICMP的reply的报文比链路正常的时候多增加了4个字节LDP公网标签，数据帧长度为1523字节，超过了MSTP传输线路最大接收帧长度，导致传输线路将服务器返回ICMP的reply的报文丢弃了。

四、 解决方法

通过调大传输线路接收最大帧长度解决问题。