MPLS MTU问题

赵刚 2007-09-25 发表

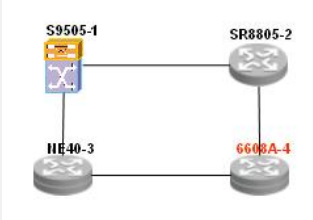
MPLS MTU问题

MPLS MTU每次想写的时候都觉得很麻烦，因为问题的原因只有一个，但是现象千差万别，各个产品实现还不一样，即使同一个产品不同板卡也有区别。

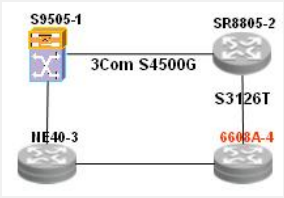
【问题描述】

问题一：

现象：如下网络结构，四个设备PE设备间，互为IBGP邻居。相互间使用loopback口建立BGP peer。S9505与SR6608A之间的BGP邻居不断UP/Down



经排查，发现更改IGP路由优先级，使S9505与SR6608A间数据流优选NE40就不会出现问题，如果数据流优选SR8805就会出问题。查看线路发现S8805的线路上还经过了两个交换机。如下图：S9505和SR8805之间有一台3com S4500G全千兆交换机。S8805和SR6608A之间有一台S3126T交换机。并且中间的报文为二层互通，带VLAN tag。



问题二：

现象：宁波江北政务网为MPLS L3 VPN环境，AR46作为纯P设备，下挂两台S85作为PE。

拓扑如下，一台S85下用户业务正常，另一台则邮件附件不能上传，部分业务不能使用。两台S85配置基本相同

AR46

| |

| |

| |

S85B--------S85A（业务不正常）

经排查，发现业务正常S85接AR46主控板以太口，业务不正常S85接AR46 2GE单板中的一个千兆口。

AR46

ERPU口 2GE口

| |

| |

| |

S85B--------S85A

问题三：

现象：SR6602A和C7509两个PE设备间IBGP邻居不断UP/Down。

经排查，发现SR6602A和C7509发现设备之间使用的使PPP MP进行的互连。

问题四：

现象：三个PE设备AR46，AR46B下私网用户访问和AR46A，AR28互访正常。AR46

A和AR28之间的用户可以ping通，但无法互访。

经排查，发现AR46A到AR46B的数据，因为倒数第二跳的原因就一层标签。而AR46A到AR28的数据有两层标签。

总结MPLS网络中这些不通的现象有一个共同点：能ping通，部分业务不通。其实这些业务的共同点就是业务中几乎都是大报文。
如果ping小包都不通，那么去查线路吧，问题就和本文无关了。J

【问题分析】
第一步：定位办法ping报文。这种情况一般都是ping小包通，ping –s 1500的大包不通。逐步改小ping包的大小，找到临界值。
第二步：如果找到临界值，那么百分百肯定是线路上某一个节点的MTU的问题，那么下一步就是确认是哪一个点MTU有问题。

问题一的分析过程：
ping1468的报文可以ping通，1469的报文无法ping通。
那么在线路上实际传输报文的长度。
目的mac＋源mac＋TAG＋类型＋ ＋MPLS＋IP头＋ICMP头＋报文
6 6 4 2 4 20 8 1468＝1518

目的mac＋源mac＋TAG＋类型＋ ＋MPLS＋IP头＋ICMP头＋报文
6 6 4 2 4 20 8 1469＝1519

而S4500的最大接收报文长度就是1518，所以导致1519的报文无法接收而丢弃。

但是设备间不是有MTU协商吗，为什么S4500还会收到超过接收能力的报文呢?
那么就要看什么是MTU。MTU最大传输单元，这个最大传输单元实际上和链路层协议有着密切的关系，让我们先仔细回忆一下EthernetII帧的结构DMAC+SMAC+Type+Data+CRC
由于以太网传输电气方面的限制，每个以太网帧都有最小的大小64bytes最大不能超过1518bytes，对于小于或者大于这个限制的以太网帧我们都可以视之为错误的数据帧，一般的以太网转发设备会丢弃这些数据帧。（注：小于64Bytes的数据帧一般是由于以太网冲突产生的“碎片”或者线路干扰或者坏的以太网接口产生的，对于大于1518Bytes的数据帧我们一般把它叫做Giant帧，这种一般是由于线路干扰或者坏的以太网口产生）
由于以太网EthernetII最大的数据帧是1518Bytes这样，刨去以太网帧的帧头（DMAC目的MAC地址48bit=6Bytes+SMAC源MAC地址48bit=6Bytes+Type域2bytes）14Bytes和帧尾CRC校验部分4Bytes（这个部门有时候大家也把它叫做FCS），那么剩下承载上层协议的地方也就是Data域最大就只能有1500Bytes这个值我们就把它称之为MTU。这个就是网络层协议非常关心的地方，因为网络层协议比如IP协议会根据这个值来决定是否把上层传下来的数据进行分片。

那么说我S4500的1518的最大接收报文长度 -（目的mac＋源mac＋TAG＋类型）＝1500
那么对端发送大于1500的报文就要分片，为什么我还收到了1519大小的报文?
那就是因为对端发送的时候有MPLS标签。MPLS是一个二层三层之间的概念。路由器在发送报文的时候，原始报文1469＋（IP头＋ICMP头）＝1497，不到1500没有分片。这样在封装MPLS头和以太头后，就超过了对端的接收范围。

[解决办法]
1．V3上可以在MPLS视图添加mtu label-including把标签长度统计进来，目前V5尚没有这个命令，目前正在开发。

2．NE上有更改MPLS MTU的命令规避该问题，这个实现和Cisco是一样的。
[NE40-3-GigabitEthernet3/0/0]mpls mtu ?
<328-8000> MTU value

3．当然也可以从S4500上下手，在S4500上开启jumboframe enable，让S4500可以接收大报文，也能解决这个问题。

问题二的分析过程：

也是一样ping1468字节可以通，1469字节不能通。
因S85发送报文MTU计算的时候是不计算MPLS头的。
S85A发出的1468字节的ping报文封装（ip+icmp）后8+20+1468=1496 因小于MTU=1500，发送。发送报文的大小为1496+4+4+14（两个标签+以太头）=1518；1469ping报文，最后发出包大小为1519。

AR46上2GE使用Intel GE8254x芯片，最大接收报文长度为1518。所以1468的ping报文能够通过。并且在AR46和S85A链接的端口上有大量超大帧错包。
GigabitEthernet1/0/0 current state :UP
Last 300 seconds input rate 0.00 bytes/sec, 0 bits/sec, 0.00 packets/sec
Last 300 seconds output rate 0.00 bytes/sec, 0 bits/sec, 0.00 packets/sec
Input: 167736958 packets, 724939975 bytes, 167737023 buffers
3 broadcasts, 1222013 multicasts, 0 pauses
298998 errors, 0 runts, 298978 giants
0 crc, 0 align errors, 0 overruns
0 dribbles, 0 drops, 0 no buffers
但是和S85B连接的ERPU使用BCM1250芯片，最大接收报文长度为1700，这样S85发出的报文因MTU为1500，1500+4+4+14<1700 所以能通过。

[解决办法]
1．参照上面的解决办法，因为S85使用的V5平台的版本，不支持mtu label-including，那么在发送上只能将S85出口的MTU改为1496，让S85不发出大于1518的大包。

2．在接收上，考虑到不同接口的接收能力不同，可以将S85都连在主控板的接口上，幸好AR46 ERPU上自带3个combo口。

问题三的分析过程：
如果SR6602A发大包目的是SR6608A的环回口，会走MPLS转发，下一跳是VT接口。当大包时，IP按照MTU分片为1500并加4个字节MPLS标签送MP转发。MP将1504的报文分两片处理。CISCO收到这个报文后会丢弃。分析原因是我们与CISCO做LCP协商时会把本地的MRU=1500发给CISCO，而CISCO收到报文时会与该MRU做比较，如果大于该MRU，认为报文非法，而丢弃。

用PING报文也可以验证这个问题，如果PING包为1469，+20字节IP头+8字节ICMP头=1497。这时候走MPLS转发时为1501。VT口为两个通道，分为2片。可以看到CISCO收到了2个分片：
01:32:32: Se4/1/5:0 MLP: I frag 80000A1E size 759 encsize 4
01:32:32: Se4/1/5:0 MLP: I data FF03 003D 8000 0A1E 0281 0019 11FF 4500
01:32:32: Se4/1/6:0 MLP: I frag 40000A1F size 760 encsize 4
01:32:32: Se4/1/6:0 MLP: I data FF03 003D 4000 0A1F C5C6 C7C8 C9CA CBCC

对分片进行重组后发现大于接口MRU=1500，CISCO将包丢弃：
00:50:11: Mu1 MLP: I pkt len 1503 > MRRU 1500 in 'r*-HG-', discarding
//CISCO错误的多统计了0281这两个标志字节
00:50:11: Mu1 MLP: Unable to add to reassembled pkt in 'r*-HG-'
00:50:11: Mu1 MLP: Discard reassembled packet
C7507-8(config-if)#

这个问题几个因素导致，一个是CISCO判断过严，我们的MP其实在1600字节以下都是可以接收的。二是SR66目前没有MPLS MTU的概念，IP分片后又格外增加了标签，导致超过接口的MTU。
另：与CISCO的MP只有一个通道时，可以正常转发。看来CISCO只是在重组分片后才做该检查。
01:27:44: Se4/1/6:0 MLP: I frag C0000976 size 1511 encsize 4
01:27:44: Se4/1/6:0 MLP: I data FF03 003D C000 0976 0281 0019 11FF 4500

[解决办法]
1． 参照上面的解决办法，因为S66使用的V5平台的版本，不支持mtu label-including。而且这个问题更改MTU也没有用，那么在发送端没有任何办法。
2． 接收端Cisco是因为双方协商为1500，但是收到1504的IP报文而任务是错误报文导致。可以配置Cisco允许接收大于协商的MTU来解决。
C7507(config-if)#ppp multilink mrru local ?
<128-16384> Our MRRU value

C7507(config-if)#ppp multilink mrru local 1504

问题四的分析过程：
问题四的分析过程和前面的类似，只不过要考虑一下相邻PE间有PHP问题，只有一层标签。读者可以自己分析一下。

归结上面的问题的原因就是链路上有的端口的最大接收报文长度为1518或1522。这样一个1500的报文＋14个以太头＋一两层MPLS头或者QinQ头就很容易超过接收能力。目前常见的MIM 2GE和下面四款FIC接口卡，因为最大接收长度为1518（不含CRC），都存在这个问题。

0231A58G	RT-FIC-1GBE-M3
0231A59U	RT-FIC-2GBE-M3
0231A59B	RT-FIC-1GEF-M3
0231A59C	RT-FIC-2GEF-M3

【解决办法总结】
1． 在接收端，增强接收能力，开启jumboframe的能力，cisco上增大MRRU。或者更换接收能力强的端口。
2． 在发送端，如果在接收端上面的操作都无法实施的时候，那么考虑发送端更改MTU的计算方法mtu label-including；或者把发送端端口的MTU值改小；对于能设置MPLS MTU的设备，更小这个设置。
3． 最终用户端，当然有的时候发送端不是H3C设备，或不允许更改的时候，只有最后一个最土的办法，更改CE下面用户PC机的MTU。 J

8062支持MIM 2GE，导致某些特别应用受限，如GRE多层封装，MPLS多标签等等，需要配置对端为小MTU才可以。
V5平台正在开发支持MPLS MTU功能。2007年9月10日提交的需求，单号：200709100010