中低端交换机（V5）网络丢包问题排查经验案例

以太网接口 张磊 2014-07-28 发表

中低端交换机（V5）网络丢包问题排查经验案例

一、 组网：

无

二、 问题描述：

无

三、 过程分析：

常见的几种网络丢弃因素有以下几种：

1）网络中线路存在质量问题，存在线路丢包的情况。

2）指导业务转发的相关表项没有及时学习建立，如MAC、ARP、路由表项。

3）网络存在二层环路，mac及arp表项出端口学习错误，导致交换机将报文转发给错误的出端口。

4)设备路由、ARP超规格相关表项无法正常下发到设备硬件导致报文转发失败或者相关表项底层下发错误。

5）网络流量比较大，网络设备存在拥塞丢包。

首先在PC 上ping 沿途各个节点，初步判断丢包位置，然后查看沿途端口上是否存在错误报文，特别注意的是IRF跨设备转发的报文，irf端口也需要查看是否存在错误报文。通过该排查方法基本上可以排除掉线路质量导致的网络丢包。

若是设备之间通过链路聚合进行互联则需关注下聚合两端的成员端口是否都处于选中状态，避免因为一段选中一段未选中造成的链路丢包。

通过在沿途设备上进行流量统计准确确认报文丢在何处。流统时需要注意以下注意事项。

1） 设备上用于匹配流量的ACL 一定要精确匹配，配置的acl 匹配报文类型ICMP,源目的IP 的反掩码配置为0。平时在处理的问题的过程中一定要严格匹配流量特征，如TCP 、UDP 报文的端口号等特征。

2） 在终端PC PING测试前一定要首先把QOS 策略下发在各个交换机的端口上，即在PING 测试前查看各个交换机的流统计结果一定要是0。

3） 查看统计的结果前一定要停止ping 测试。

4） 部分交换机配置流量统计的动作为accounting ，部分交换机为accounting packet，各个交换机的具体配置方法请参看相关设备及相关版本的操作手册。

5） QOS 必须下发到物理端口上且聚合组内的所有成员端口都需要下发。

具体的流量统计方法可以参考KMS – 23894。

确认报文丢弃位置后，查看该设备的MAC地址表、ARP 表，路由表中的相关表项学习是否正常，关注表项出端口是否正确。

查看特定mac地址表项的命令：

```
[H3C]display  mac-address  60eb-6926-aca4

MAC ADDR      VLAN ID  STATE      PORT INDEX            AGING TIME(s)

60eb-6926-aca4 1       Learned   GigabitEthernet1/0/1     AGING

---  1 mac address(es) found  ---
```

若是MAC地址经常在不同的端口上切换学习到，说明网络中肯定存在环路。对于不该学习到对应MAC 的端口下面的网络可能存在环路，需要重点排查。也可以根据下文介绍的查看mac地址漂移的方法来确定环路的位置。

查看特定ARP表项的命令：

```
display  arp 172.16.0.121

                Type: S-Static    D-Dynamic    M-Multiport

IP Address      MAC Address     VLAN ID  Interface      Aging Type
```

172.16.0.121 60eb-6926-aca4 1 GE1/0/1 20 D

由上述正常的表项可知，同一个终端的mac表项及arp表项中的vlan出端口这些信息是一致的，若是不一致则网络肯定存在问题，需要排查网络是否存在环路。

若是ARP 表项无法正常建立，则可以尝试测试下静态ARP 表项的情况下网络是否正常，已确认网络丢包或者不通是否就是由arp 模块导致的。若是ARP 无法建立则需要排查网络中是否存在较多的arp 报文对设备造成攻击，通过抓包等手段抓取arp 报文的交互过程同时在设备debug arp packet 观察报文交互过程。

查看特定路由表项信息：

[H3C]display ip routing-table 192.168.1.100

Routing Table : Public

Summary Count : 1

Destination/Mask Proto Pre Cost NextHop Interface

192.168.1.0/24 Static 60 0 172.16.0.121 Vlan1

通过查看下一跳地址的ARP信息可以确认报文的出接口。

上述命令是查看软件层面相关表项学习是否正常，若是软件层的路由表项都无法建立则需要先排查相关协议是否正常。

网络中存在二层环路的一般都伴随着广播风暴，mac学习在错误的端口上。一般input流量特别大的端口及学习到其他端口下终端mac的端口存在环路，需要重点排查。也可以通过查看芯片上记录的mac地址漂移记录来分析环路位置。

1、 通过命令查看MAC地址漂移的方法：

[H3C-diagnose]debug l2 1 0 mac/move_rec/show（bcm l2 slotid chipid mac/move_rec/show）

==========L2MACMOVEMODULE INFO==========

L2MacMoveModule Enabled

L2MacMoveDebug Switch Off

==========L2MACMOVE Record INFO==========

MacAddress Vlan Agg Mod Port ->Agg Mod Port Cnt LatestTime

f8:bc:12:31:8c:9 301 0 5 19 ->0 5 18 689 2014/1/17 01:41:20

f8:bc:12:31:8c:e9 301 1 0 0 ->0 5 19 690 2014/1/17 01:37:46

f8:bc:12:31:85:f1 301 0 5 23 ->0 5 22 964 2014/1/17 01:39:20

c8:1f:66:d7:a6:a7 301 0 5 21 ->0 5 20 023 2014/1/17 01:38:18

2、 诊断信息中搜素“mac/move_rec/show”也可找到对应槽位存在的mac地址漂移记录。

以下针对上面红色的字体部分的记录进行分析，排查漂移所在的物理端口号。

从信息中可以看到f8:bc:12:31:8c:9(补全为f8bc-1231-8c09)这个mac在vlan 301 内，聚合组0（Agg 为零表明物理端口非聚合口），从Mod 5 的 Port 19 漂移到了Mod 5 的Port 18，漂移次数为689 最后一次漂移的时间为2014/1/17 01:41:20。以下为查看Mod 5 的 Port 19 和Mod 5 的Port 18的具体方法：

1、 在诊断视图下输入以下命令可以显示端口的对应关系。

[H3C-diagnose]debug port mapping 1（debug port mapping slotid）

==========

[Interface] [Unit][Port][Name][Combo?][Active?][IfIndex] [MID][Link]

==========

GE1/0/1 0 3 ge1 no no 0x900000 4 up

GE1/0/2 0 2 ge0 no no 0x900001 4 up

……

GE1/0/43 1 18 ge32 no no 0x90002a 5 up

GE1/0/44 1 19 ge33 no no 0x90002b 5 up

2、 在诊断视图下搜素“debug port mapping ”也可以查看对应槽位的映射信息。

以上信息可以知道之前从Mod 5 的 Port 19 漂移到了Mod 5 的Port 18实际上是从物理口GE1/0/44迁移到GE1/0/43，存在大量mac地址漂移一般多为环路导致，需要排查涉及物

理端口的下的设备是否存在环路。

蓝色字体部分的漂移记录Agg标记位为1 说明后面的Mod Port表示的实际物理端口为聚合口，查看Agg 1 Mod 0 Port 0 对应聚合口的方法如下。
在诊断视图下敲入以下命令：

```
[H3C-diagnose]debug port  trunk-global 1  (debug port  trunk-global slotid)
=============================================================
If=9530000
TG=0
Agg=1
ANum=2
HKey=0x6
Port: 0x01900000
-------------------------------
```

从以上信息可知Agg 1 Mod 0 Port 0对应的聚合组号为聚合组1.

也可以在diag 信息中搜索“trunk show”来找到对应槽位的链路聚合底层信息。

==========trunk show==========

Device supports 136 trunk groups:

128 front panel trunks (0..127), 16 ports/trunk

8 fabric trunks (128..135), 16 ports/trunk

trunk 0: (front panel, 2 ports)=ge4,ge5 dlf=any mc=any ipmc=any psc=portflow (0x9)

trunk 0: egress ports=cpu,ge,xe0-xe1,xe4-xe11,hg

由上述信息可知trunk 0 包含ge4 、ge5两个端口，从对应槽位的port mapping 信息可以这两个端口对应的物理端口，自然知道其对应的聚合组号。

==========

==========debug port mapping 1==========

==========

[Interface] [Unit][Port][Name][Combo?][Active?][IfIndex] [MID][Link] [Attr]

==========

GE1/0/1 0 3 ge1 no no 0x900000 4 up

GE1/0/2 0 2 ge0 no no 0x900001 4 up

GE1/0/3 0 5 ge3 no no 0x900002 4 up

GE1/0/4 0 4 ge2 no no 0x900003 4 up

GE1/0/5 0 7 ge5 no no 0x900004 4 up

GE1/0/6 0 6 ge4 no no 0x900005 4 up

GE1/0/7 0 9 ge7 no no 0x900006 4 up

GE1/0/8 0 8 ge6 no no 0x900007 4 up

GE1/0/9 0 11 ge9 no no 0x900008 4 up Bridge下面介绍BCM 芯片查看底层芯片硬件上表项的方法，排查软硬件表项不一致导致的转发失败。

通过查看系统路由表可知设备配置去往192.168.1.0/24的路由。同时，配置去网172.16.1.0/24的两条静态路由，下一跳分别为10.0.0.2和10.0.0.6。设备上的路由表如下所示：

display ip routing-table

Routing Tables: Public

Destinations : 10 Routes : 11

Destination/Mask Proto Pre Cost NextHop Interface

10.0.0.0/30 Direct 0 0 10.0.0.1 Vlan2

10.0.0.1/32 Direct 0 0 127.0.0.1 InLoop0

10.0.0.4/30 Direct 0 0 10.0.0.5 Vlan3

10.0.0.5/32 Direct 0 0 127.0.0.1 InLoop0

172.16.0.0/24 Direct 0 0 172.16.0.1 Vlan1

172.16.0.1/32 Direct 0 0 127.0.0.1 InLoop0

172.16.2.0/24 Static 60 0 10.0.0.2 Vlan2

Static 60 0 10.0.0.6 Vlan3

192.168.1.0/24 Static 60 0 172.16.0.121 Vlan1

下面我们分三种情况分析报文的转发流程：

(1)三层报文的目的ip为设备的直连路由

假设报文的目的ip为172.16.0.121。首先查host表，得到下一跳表的索引号为100004。

[H3C-diagnose]bcm 1 0 l3/l3table/show （diag 中直接搜索“l3table show”）

Unit 0, free L3 table entries: 8189

Entry VRF IP address Mac Address INTF MOD PORT CLASS HIT

1 0 10.0.0.2 00:00:00:00:00:00 100002 0 0 0 n

2 0 172.16.0.121 00:00:00:00:00:00 100004 0 0 0 y

3 0 10.0.0.6 00:00:00:00:00:00 100003 0 0 0 n

根据下一跳索引号查找下一跳表，得到下一跳的MAC、出VLAN、三层接口索引号、mod和port。如下所示：

[H3C-diagnose]bcm 1 0 l3/egress/show （diag中直接搜索“l3 egress show”）

Entry Mac Vlan INTF PORT MOD MPLS_LABEL

100000 ff:ff:ff:ff

:ff:ff 0 4095 28 0 0

100002 00:0f:e2:74:93:18 2 2 1 0 0

100003 00:0f:e2:74:93:18 3 3 2 0 0

100004 60:eb:69:26:ac:a4 1 1 3 0 0

说明：如果host表中没有172.16.0.121的表项，设备会首先发出arp请求报文，学习到arp后，再按照上述流程进行转发。

根据mod 和 port 信息确定物理出接口,如下所示：

[H3C-diagnose]debug port mapping 1

[Interface] [Unit][Port][Name][Combo?][Active?][IfIndex] [MID][Link] [Attr]

==========

GE1/0/1 0 3 ge1 no no 0x900000 4 up

GE1/0/2 0 2 ge0 no no 0x900001 4 up

(2)三层报文的目的ip为的非直连路由且不存在等价路由

假设报文的目的ip为192.168.1.100。查找LPM表，得到下一跳索引号为100004：

[H3C-diagnose]bcm 1 0 l3/defip/show （diag中直接搜索“l3 defip show”）

Unit 0, Total Number of DEFIP entries: 12289

VRF Net addr Next Hop Mac INTF MODID PORT PRIO CLASS HIT VLAN

0 0 192.168.1.0/24 00:00:00:00:00:00 100004 0 0 0 0 n

得到下一跳索引号后，接着查找下一跳表，得到下一跳MAC、出vlan、三层接口索引等信息，处理流程同第一种情况。

(3)三层报文的目的ip为设备的非直连路由且存在等价路由

假设报文的目的ip为172.16.2.100。查找LPM表，发现去往该目的有多条路由存在，并得到等价路由表的索引号200000；查找等价路由表，得到两个下一跳表项的索引号100002和100003：

[H3C-diagnose]bcm 1 0 l3/defip/show

VRF Net addr Next Hop Mac INTF MODID PORT PRIO CLASS HIT VLAN

0 0 172.16.2.0/24 00:00:00:00:00:00 200000 0 0 0 32 n (ECMP)

//ECMP表

[H3C-diagnose]bcm 1 0 l3/multipath/show

Entry Mac Vlan INTF PORT MOD MPLS_LABEL

Multipath Egress Object 200000

Interfaces: 100002 100003

//host表

[H3C-diagnose]bcm 1 0 l3/l3table/show （diag 中直接搜索“l3table show”）

Unit 0, free L3 table entries: 8189

Entry VRF IP address Mac Address INTF MOD PORT CLASS HIT

1 0 10.0.0.2 00:00:00:00:00:00 100002 0 0 0 n

2 0 172.16.0.121 00:00:00:00:00:00 100004 0 0 0 y

3 0 10.0.0.6 00:00:00:00:00:00 100003 0 0 0 n

//下一跳表

[H3C-diagnose]bcm 1 0 l3/egress/show

Entry Mac Vlan INTF PORT MOD MPLS_LABEL

100002 00:0f:e2:00:19:21 2 5 2 4 0

100003 00:0f:e2:2d:6a:f8 3 3 5 4 0

100009 00:0a:eb:f2:51:7f 20 5 8 8 0

最后根据port mapping表确认出端口及mac 地址：

==========debug port mapping 1==========

==========

[Interface] [Unit][Port][Name][Combo?][Active?][IfIndex] [MID][Link] [Attr]

==========

GE1/0/1 0 3 ge1 no no 0x900000 4 up

GE1/0/2 0 2 ge0 no no 0x900001 4 up

GE1/0/3 0 5 ge3 no no 0x900002 4 up

GE1/0/4 0 4 ge2 no no 0x900003 4 up

经过上述方法可以根据软件上学习的表项查看底层芯片下发的转发表项是否正确。

若是底层表项下发错误会出现查看软件表项正常但是转发失败的情况，表现是完全不通。

网络中常见的一种故障是丢包（间隔性丢弃个别报文），造成该故障的一个常见原因是拥塞。

常见的拥塞原因有：

1、 网络设备上存在10M,half的端口。目前大多数交换机的缓存都是基于芯片共享，低速半双工的端口极容易造成拥塞而占满整个芯片的缓存。

2、 网络架构存在高速打低速的情况，且网络流量突发性大。如上行速率为10G 或40G，下行速率为100M 或者10M，此时由于上下行速率存在较大的速率差，在下行存在突发流量的情况下很容易造成下行端口拥塞，造成瞬间丢包。

3、 网络流量模型存在多打一的情况，且流量存在较大突发。例如设备上存在5个GE端口连接了台server，但是业务模型上4台server流量都向其中一台server打流量，且流较大或者存在较大突发。

4、 框式设备个别单板非线速单板，在流量较大情况下单板内部由于带宽限制存在拥塞丢包。

xp32

上图为12500上的32端口万兆单板的简易图，由此可知4个端口共用一个芯片，两个芯片连接到一个PP芯片上，PP芯片与下面的芯片间仅仅有10G的带宽，PP芯片上行有20G带宽。由此可知该单板的收敛比为4 ：1 。若用户需要使用8个10G,则应该在每个芯片上选择一个端口，这样就可以限速转发了。

针对上述几种常见原因，下面介绍一下排查优化方法：

1、 首先查看各个端口的双工速率是否正常，排查低速半双工的端口。

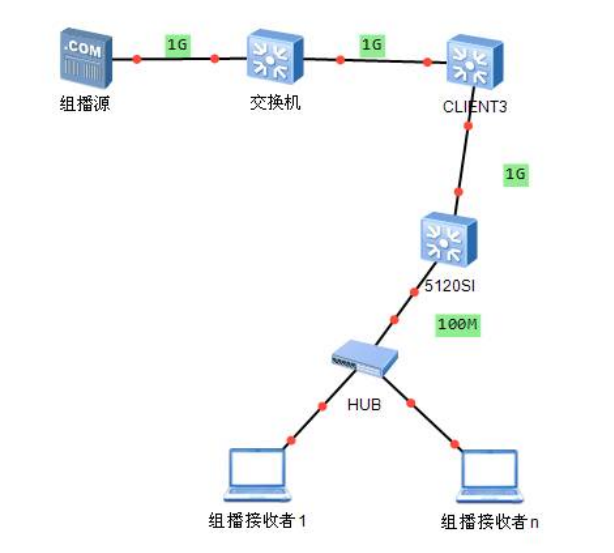
2、 了解组网中是否存在高速打低速，多打一的情况。若存在可以在设备上开启burst-mode enable （部分设备不支持）来优化缓存分配方式以适应流量突发的模型来观察。

3、 针对多打一，高速打低速的情况也可用通过链路聚合增加出口带宽的方式来规避，配置链路聚合的时候请注意成员端口的流量hash是否均匀，若是不均匀及时调整hash算法。

4、 对于非限速单板可以通过调整使用的端口或者更换单板来解决。单板缓存比较小的可以考虑更换大buffer的单板来观察解决。

5、 调整报文源端口的速率有时候也可以解决问题。如下面的这个问题就是通过调整组播源服务器的速率为100M解决的。

拓扑：



用户网络运行组播业务，server上承载着n多路组播数据。5120SI 设备上行1G下行100G ，存在高速打低速的情况。开启burst-mode enable 后有好转但是仍存在故障。

经排查现场server 发送的组播数据存在突发情况，流量波动很大。在无法通过链路聚合增加下行带宽，也无法调整下行端口的工作速率的情况下，最后将server的网卡强制工作在100M的模式后，全网观察组播效果很好。

下面介绍几种在设备上查看拥塞丢包的方法：

对于S10500 、S75E 、S5800、S5830、S5500HI 、S5500EI、S5500SI 、S5120EI等bcm 芯片的设备可以通过如下底层命令查看芯片丢包计数。该命令记录的是设备自启动以来记录的历史累积记录，因此需要读取多次，查看后一次与前一次的计数差值来比较。

75E 55设备一般是下面的记录：

[H3C]en_diag

[H3C-diagnose]bcm 1 0 show/c（bcm slotid chipid show/c）

RIPC4.ge0 ： 2,178,933,490 +638,721,773 5,235/s

RDISC.ge0 ： 1,069,666 +301,634

RUC.ge0 ： 3,446,748,216 +1,033,593,076 7,549/s

RDBGC0.ge0 : 1,070,049 +301,678

RDBGC1.ge0 : 60,240 +17,424

RDBGC3.ge0 : 1,070,049 +301,678

RDBGC4.ge0 : 1,069,666 +301,634

HOLD.ge0 ： 196 +57

TDBGC3.ge0 : 10,578 +2,447

以上信息说明ge0端口存在拥塞记录，ge0代表的具体物理端口号请参考上文介绍的port mapping 部分。

105 、5830、5800 设备上一般是下面的记录：

[H3C]en_diag

[H3C-diagnose]bcm 1 0 show/c（bcm slotid chipid show/c）

RUC.ge0 : 11,551,819,325 +11,551,819,325 5,391/s

GR64.ge0 : 14,904,837 +14,904,837 4/s

GT511.ge0 : 11,247,276 +11,247,276 2/s

……

PERQ_BYTE(7).ge0 ： 27,815,256 +27,815,256

PERQ_DROP_PKT(1).ge0: 1,705 +1,705

PERQ_DROP_BYTE(1).ge0: 2,030,874 +2,030,874

上述红色的字体计数代表有拥塞的历史记录。Ge0代表的具体端口号请参考前方介绍的port mapping的查找方式来确认。

对于5120SI 设备可以在隐藏视图下（_h进入）采集_display drv walkthrough命令来查看芯片是否存在拥塞的历史记录（在diag中直接搜索“walkthrough”）。

以下信息可Tail Dropped为零，说明没有丢包计数。由于该命令是读取设备

自启动以来的累积计数，因此一般在丢包前后各采集一次进行对比分析。

==========

==========_display drv walkthrough==========

==========

Counter Offset Chip0 Chip1

---------- ---------- ---------- ----------

Global security Breach filter 2040104 0 0

Port/VLAN Security Breach Drop 2040108 0 0

Bridge Filter Counter 2040150 0 0

---------- ---------- ---------- ----------

Set VLAN Ingress Filter

Set0 20400e4 0 0

Set1 20400f8 0 0

---------- ---------- ---------- ----------

---------- ---------- ---------- ----------

Set Security Filter

Set0 20400e8 0 0

Set1 20400fc 0 0

---------- ---------- ---------- ----------

---------- ---------- ---------- ----------

Set Bridge Filter

```
 Set0                          20400ec          0         0
 Set1                          2040100          0         0
-------------------------------- ---------- ---------- ----------
-------------------------------- ---------- ---------- ----------
Set Incoming Packet
 Set0                          20400e0    1155068    1487041
 Set1                          20400f4    1155061    1487030
-------------------------------- ---------- ---------- ----------
Ingress Drop                     b000040          0       2661
-------------------------------- ---------- ---------- ----------
Receive SDMA Packet
 Set0                          2820       2552       315
 Set1                          2824       0          1
 Set2                          2828       0          0
 Set3                          282c       0          0
 Set4                          2830       0          2
 Set5                          2834       0          0
 Set6                          2838       0          0
 Set7                          283c       0          0
-------------------------------- ---------- ---------- ----------
-------------------------------- ---------- ---------- ----------
Receive SDMA Byte
 Set0                          2840     188922     38652
 Set1                          2844       0        74
 Set2                          2848       0          0
 Set3                          284c       0          0
 Set4                          2850       0       1208
 Set5                          2854       0          0
 Set6                          2858       0          0
 Set7                          285c       0          0
-------------------------------- ---------- ---------- ----------
-------------------------------- ---------- ---------- ----------
Receive SDMA Resource Error
 Set0                          2860       0          0
 Set1                          2864       0          0
-------------------------------- ---------- ---------- ----------
-------------------------------- ---------- ---------- ----------
Set Outgoing Unicast
 Set0                          1b40144    2160017    2620425
 Set1                          1b40164    2160017    2620420
-------------------------------- ---------- ---------- ----------
-------------------------------- ---------- ---------- ----------
Set Outgoing Multicast
 Set0                          1b40148      67269     115967
 Set1                          1b40168      67269     115967
-------------------------------- ---------- ---------- ----------
-------------------------------- ---------- ---------- ----------
```

Set Outgoing Broadcast

Set0	1b4014c	33622	54354
Set1	1b4016c	33623	54355

---------- ---------- ---------- ----------

---------- ---------- ---------- ----------

Set Bridge Egress Filter

Set0	1b40150	1	0
Set1	1b40170	1	0

---------- ---------- ---------- ----------

---------- ---------- ---------- ----------

Set Tail Dropped

Set0	1b40154	0	0
Set1	1b40174	0	0

---------- ---------- ---------- ----------

---------- ---------- ---------- ----------

Set Control Packet

Set0	1b40158	13131	5293
Set1	1b40178	13131	5293

---------- ---------- ---------- ----------

---------- ---------- ---------- ----------

Set Egress Forward Restrict

Set0	1b4015c	0	0
Set1	1b4017c	0	0

---------- ---------- ---------- ----------

---------- ---------- ---------- ----------

Set Multicast FIFO Dropped

Set0	1b40180	175	1186
Set1	1b40184	175	1186

---------- ---------- ---------- ----------

四、 解决方法：

具体请参考过程分析部分。