中端交换机ND资源不足分析

IPv6 ARP **郭威** 2014-04-25 发表

中端交换机ND资源不足分析

一、组网：

本文主要讨论的是V5及V7平台版本的S7500E/S10500关于ND资源的分配规则及超规格后的排查方法。以V5平台的S10500为例，作为汇聚层开启IPV4/IPV6功能，下挂的终端运行了IPV4/IPV6双栈，网关均在S10500上。在S10500上配了防火墙板卡、无线板卡及交换板卡。

二、问题描述：

客户反馈IPV4/IPV6网络不稳定，出现如下现象：部分终端出现可以PING通百度，但是无法正常访问百度网页的情况；正常上网的用户丢包情况严重，根据流量统计可以确认报文丢在S10500上。

三、过程分析：

收集故障发生时问题设备的诊断信息，有如下现象可供判断。
1、CPU利用率过高
查看设备CPU占有率，发现单板CPU居高不下。
[H3C]display cpu
Slot 2 CPU usage:
97% in last 5 seconds
100% in last 1 minute
100% in last 5 minutes
进入隐含模式下查看该单板进程，发现有3个进程明显偏高
[H3C-hidecmd]display cpu-usage task slot 2
bRX1 36% 0/5ffc92c8 // 报文收发任务
SOCK 24% 0/3f35a441 // 平台处理报文任务
ARP 21% 0/3863aa4a // arp任务
由于ARP或ND表项没有能够下发到硬件，交换机便无法完成硬件转发，报文会上CPU进行软件转发，从而造成CPU高的情况，同时有可能大量丢包。该现象能从侧面证明ARP或ND表项超规格。

2、ND表项
Mar 10 2014 18:09:57:0964:
LINE:351, File platform_bcm/drv/l3/ipv6/drv_ipv6_uc.c-TASK:ND-FUNC::
ND table is full, no resource to add nd entry!ulNdFlag=512, ulFound=263
这条信息说明ND表项已经满了，进一步核实时间是否对得上问题发生的时间。如果出现多条，则可确认ND资源不足。

3、NH表项
Mar 11 2014 08:54:03:0089:
LINE:3041, File platform_bcm/drv/l3/ipv4/drv_ipv4_uc.c-TASK:ARP-FUNC::
add_arp_nexthop():No resource to add arp entry!
ulDrvIPv4ArpCount=6902,ArpNHCount=6902,IPAddr=739dd2d1,vrf=0,MAC=00-00-00-00-00-00,ArpStatus=0x2,ArpFlag=0x0.
这条信息说明ARP的下一跳表项（即NH表项）已经满了。由于默认情况下，NH表项的规格与ARP规格相同，因此通常都是先出现ARP表项满了之后才会出现NH表项超规格。另外当更改了单板模式为标准IPV6之后，也有可能单独出现该告警信息。

4、查看故障主机
查看上网异常的主机的表项是否已经下发到硬件上。
H3C-diagnose]debug ipv6-drv show nd XX::XX slot 1
[H3C-diagnose]debug ipv4-drv show arp X.X.X.X slot 1
****
- IPv4 ARP Information Slot 1
****
--- UNIT: 0 ---
- Entry not found //此处显示表项未能成功下发到硬件上

****

5、查看具体下发情况
如果能够在问题复现时登陆设备，则可以通过以下方法进行判断：
进入系统视图，输入display ipv6 neighbors all，统计出在软件层面上下发的ND数目。输入display arp count，统计出在软件层面上下发的ARP数目。
进入诊断视图，输入命令debug l3intf-drv show statistics slot<>中，可以通过查看标红的ND COUNT及ARP COUNT看到下发到底层槽位上的数目。

```
**************************************************************
- L3INTF Statistics     Slot 2
**************************************************************
- NH:                        8192  //下一跳地址，与ARP规格相同
- ARP
      SPECIFICATION:        8192   //ARP规格
      COUNT:                3734   //ARP下发到SLOT2上的数目
      NHCOUNT:              3734   //占用的下一条资源
- IPV4 ROUTE
      SPECIFICATION:        12288
      COUNT:                1050
- ND
      SPECIFICATION:        4096   //ND规格，是ARP规格的一半
      COUNT:                1139   //ND下发到SLOT2上的数目
      NHCOUNT:              1141   //占用的下一条资源
```

如果看到下发到软件层面的ND表项及ARP表项数目要远大于下发到底层槽位硬件上的。同时进入诊断视图下，输入local logbuffer slot <> display all，也能够看到ND table is full的报错信息。由于local logbuffer较小，很快就会被刷新掉，因此需要在问题复现时收集查看。由此可以进一步确认问题产生的原因是ND资源不足。

四、 解决方法：

1、当S10500上有SC单板，且下挂的终端数量小于(ARP表项/2)=4K时，可以通过更改交换机业务单板的工作模式来解决。
在系统视图下输入
Switch-mode standard-ipv6 chassis <> slot <>
可将业务单板从普通模式转为标准IPV6模式，注意所有业务板均需要修改（带有业务槽位的主控板也同样需要修改），修改完后保存，重启业务板生效。
以SC类单板为例，其原理如下：
(a) 普通模式：
普通模式中ND与ARP共同占用一个iphost表项，即宣称的ARP的8K表项。其中每一个ARP占用一条表项，一个ND占用两条表项。
最大规格计算方式即 ARP + ND * 2 = 8K
对于学校的IPV6用户，一台PC终端，最多会对应3个IPV6地址，分别是FE80开头的本地链路地址和2001或2002开头的隧道地址。因此实际上一个IPV6终端占用了2条ND表项，即4条表项。
最大终端规格计算方式： IPV4终端数目 + IPV6终端数目 * 4 = 8K
注：需要注意的是IPV6终端通常也是IPV4终端。
(b) 标准IPV6模式：
标准ipv6模式对上面进行了优化，将arp放在路由表中与路由表共享12k路由表项。而nd独占原来arp+nd的8k表项，因此ND表项规格为8K/2 = 4K（一个ND占用两条表项）.
由于该模式下进行了优化，本地链路地址不再下发到硬件上，因此实际上一个IPV6终端占用了一条ND表项。即可用的IPV6终端数最大为4K。
ARP表项虽然放在了路由表中，但其规格仍然是最大8K。
值得注意的是由于NH即下一跳表项（如果没有外扩ARP，则与普通模式ARP相同，此处为8K）的限制，ARP+ND的数目仍然不能超过8K，即ARP + ND = 8K。

2、当S10500下挂的终端数量大于4K小于8K时，则可以通过将SC单板更换更高性能的板卡，如SE/EA/EB系列；如果下挂的终端数量大于8K时，则只能通过调整组网，将一部分网关放在其他设备上。

3、可以通过修改ND老化时间进行优化
在之前的版本实现中，ND表项的老化时间为24小时，且不能调整。为了适应网络变化的需要，新版本对此进行了优化，可以通过如下命令调整：
Ipv6 neighbor stale-aging aging-time ,
aging-time取值为1~24小时。用户可以根据实际情况选择老化时间，以对网络进行优化，释放资源。

注：下发ND表项的时候需要占用入方向的ACL资源。如果16K全部下发为ND表项，那么最多可能占用2个slice，即512条。

五、规格总结：

由于框式设备均是是分布式转发，一个端口学习到的ARP或者ND需要全局同步到所有单板，因此整机中以能力最小的单板规格作为整机规格，要尽量避免不同规格的单板混搭的情况。否则高规格单板发挥不了优势，特别需要注意的是无线及安全插卡的后插卡同样有规格限制。
例如在10500上配置有SC及SE单板。虽然10500的整体ARP规格为16K，SE单板也为16K，但是由于SC单板的ARP规格为8K，那么整机的规格均按照最低的SC单板计算，即整机ARP规格为8K。

如下是常见的业务板卡的ARP及ND规格。
1、关于交换业务板卡：
S7500E上的规格：

	SA系列	SC系列	SD系列	EA系列	EB系列
ARP规格	2K	8K	16K	16K	16K
ND规格	不支持	最大4K	最大8K	不支持	最大8K

S10500上的规格：

	SC系列	SE系列	EA系列	EB系列
ARP规格	8K	16K	16K	16K
ND规格	最大4K	最大8K	最大8K	最大8K

注：
1、S7600可以参考S7500E；S7600-X可以参考S10500；
2、S7500E主控板上有业务槽位的，其规格可参考SC类单板。
3、S10500上的SF类板卡，不支持修改代理模式。

2、关于安全插卡：
防火墙板卡，在交换侧对应的规格可以从其命名确认。
例如LSQ1FWBSC0单板，从其开头三位字母LSQ可知是S7500E的板卡，从其最后两位字母SC可知对应规格为SC类单板。因此查表可知其ARP规格为8K，ND规格为4K。

3、关于无线插卡：
在75E及105上主要有两种无线插卡，分别是LSQM1WCMB0、LSUM3WCMD0。前者的ARP规格为8K，ND规格为4K；后者的ARP规格是16K，ND规格为8K。

4、关于V7平台的S10500和S7500E：
在V7平台上，在普通模式下本地链路地址默认不下发到底层，即不占用表项资源。同时修改为标准IPV6模式后，也会同V5平台一样将ARP&ND表项全部拿给ND用。

5、关于12500-X：
修改为标准IPV6模式对于12500-X没有意义，因为支持该命令需要芯片支持，现在暂时不能支持。