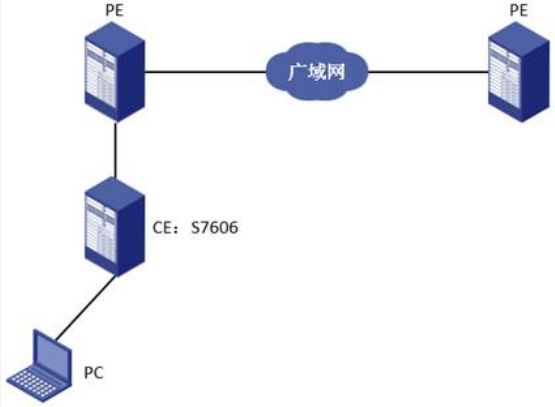
S7600迭代路由问题的分析和判断

静态路由 RIP OSPF BGP 秦军 2014-07-25 发表

S7600迭代路由问题的分析和判断

一、 组网：



二、 问题描述：

现场S7606交换机升级到R6708版本后，交换机二层转发的流量都正常，但部分三层流量转发不通。如果将设备重启，原本不通的流量转发有时可以恢复，但重启后本来能够正常转发的三层流量也可能会出现不通的情况。

三、 过程分析：

现场找到两个有问题的路由条目，通过S7600交换机访问10.43.38.13和10.43.38.14这两个地址时，在设备上可以查找到相应的Fib表项：

display fib

Destination count: 19339 FIB entry count: 19339

Flag:

U:Useable G:Gateway H:Host B:Blackhole D:Dynamic S:Static R:Relay

Destination/Mask Nexthop Flag OutInterface InnerLabel Token

10.43.38.0/24 10.120.1.5 UDG Vlan7 Null Invalid

检查芯片硬件表项，可以看到在3槽和5槽有此地址的表项，表项指向的出口索引是100006：

[S7606-diagnose]bcm 3 0 l3/l3table/show

Unit 0, free L3 table entries: 16052

Entry VRF IP address Mac Address INTF MOD PORT CLASS HIT

92167 0 10.43.38.0/24 00:00:00:00:00:00 100006 0 0 0 32 y

但是100006指向的是黑洞表项：

[S7606-diagnose]bcm 3 0 l3/egress/show

Entry Mac Vlan INTF PORT MOD MPLS_LABEL ToCpu Drop DisDA DisSA DisVLAN DisTTL

100006 00:00:00:00:00:00 4095 4095 63 14 0 no yes no no no no

这样报文匹配上此表项后无法转发到正确的出端口上，而是直接丢弃。

经检查设备的运行信息发现设备有大量的路由下发失败的记录，下发失败的原因是传入的参数出现了错误，路由下发失败就会出现上面转发指向了黑洞表项。这时，查看设备记录的local log信息，发现路由表项有大量的下发失败记录，而且类型均相同：

[S7606-diagnose]local logbuffer 3 display 360

Feb 03 2013 01:18:31:0865:

LINE:1379, File platform_bcm/drv/l3/utils/drv_ipv4_shim.c-TASK:FIB- FUNC::

Fail to add route entry!ulUnitID=0,vrf=0,ip=a2b0418,mask=fffffffc, IRv=-4

Feb 03 2013 01:18:31:0865:

LINE:961, File platform_bcm/drv/l3/utils/drv_ipv4_shim.c-TASK:FIB- FUNC::

Call drv_ipv4_shim_uc_ipv4_add_forward_route return 1,ip = 0xa2b0418, mask = 0xfffffffc

Feb 03 2013 01:18:31:0865:

LINE:410, File platform_bcm/drv/l3/utils/drv_ipv4_shim.c-TASK:FIB- FUNC::

Call drv_ipv4_shim_ipv4_addroute_hard error,Ret=1,ip = 0xa2b0418,mask = 0xfffffffc

除此以外，在设备上还有ECMP没有资源的告警记录：

Feb 03 2013 01:18:31:0877:

LINE:10702, File platform_bcm/drv/l3/ipv4/drv_ipv4_uc_intf.c-TASK:FIB- FUNC::

ERR_NO_ENOUGH_RESOURCE !ulType=0xcc010009

Feb 03 2013 01:18:31:0877:

LINE:858, File platform_bcm/drv/l3/ipv4/drv_ipv4_uc_intf.c-TASK:FIB- FUNC::

call DRV_L3UC_ECMP_Malloc error:No resource to add ECMP entry!

ECMP Group的规格是127条，目前设备上规格满了，所以会有此告警：

[S7606-diagnose]debug ipv4-drv show config slot 3

****

- IPv4 Config Slot 3

****

- ARP SIZE: 16384

- ArpCanNotSetToHW: NO

- IPV4 ROUTE SIZE: 131072

- ECMP SIZE: 16

- SDK ECMP SIZE: 16

- ECMP GROUP SIZE: 127 //ECMP Group规格

通过命令可以看出实际下发到硬件的条数显示，这时ECMP Group已经占满了：

[S7606-diagnose]bcm 3 0 l3/multipath/show

Multipath Egress Object 200000 //第1组

Ecmp Cnt: 67

  Interfaces: 100006

Multipath Egress Object 200001 //第2组

Ecmp Cnt: 1

  Interfaces: 100191

Multipath Egress Object 200002 //第3组

Ecmp Cnt: 175

  Interfaces: 100006

Multipath Egress Object 200003 //第4组

Ecmp Cnt: 57

  Interfaces: 100191

Multipath Egress Object 200004 //第5组

Ecmp Cnt: 18

  Interfaces: 100006

Multipath Egress Object 200005 //第6组

Ecmp Cnt: 1

Interfaces: 100006

... ...

Multipath Egress Object 200126 //第127组

Ecmp Cnt: 73

Interfaces: 100006

然而，10.43.38.0/24这条路由并不是等价路由，为什么也会受影响呢？经过进一步确认，ECMP Group资源不仅等价路由会占用，需要迭代查找出接口的路由也会占用，这个是在R6708及其以后版本上增加的F R R特性，即路由迭代功能，配置指导上说明如下：

路由迭代配置：

如果路由所携带的下一跳信息并不是直接可达的，需要找到到达下一跳的直连出接口，路由迭代的过程就是通过路由的下一跳信息来找到直连出接口的过程。

缺省情况下，路由运算允许进行迭代过程，这样可以保证路由快速收敛。

而在使用等价路由进行负载分担的场合，建议您不要使用迭代模式；在需要路由表项数目较多的场合，也可以配置为非迭代模式来保证路由规格。这时可以通过下面配置，配置为非迭代模式。

表1 配置路由迭代模式

操作	命令	说明
进入系统视图	**system-view**	-
配置路由迭代模式	**switch-mode route-iterative**	缺省情况下，为非迭代模式

由于平台下发的路由为FRR路由，路由指向VN表项(VN表项是平台软件下发的中间软件表项)，而VN表项需要占用ECMP Group资源，因此这种算法会导致消耗大量的ECMP资源。当ECMP Group资源耗尽时，如果出现了路由的切换就会导致部分普通路由表项的下发也出现错误，从而造成部分路由表项下发失败，现场10.43.38.0/24这条路由就是在这种情况下产生的。

四、 解决方法：

对于R6708版本，默认会进行迭代路由运算，从R67078P03版本开始，路由迭代模式命令缺省值发生变化，由缺省进行迭代模式路由运算改为非迭代模式路由运算。

此问题可以手工通过配置非迭代模式路由运算来解决，这样迭代路由不会再占用ECMP Group资源。注意：修改路由计算模式后，必须将设备整机重启才能生效。或者升级设备的软件版本，升级到R6708P03及其以后的版本，这样系统默认的路由运算模式为非迭代模式，不会出现类似问题。