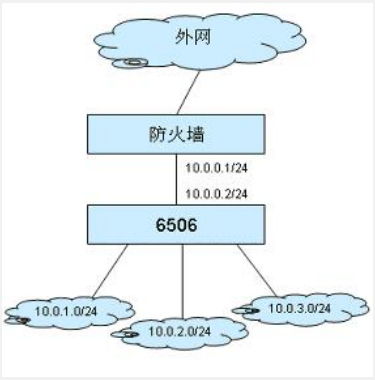
关于S6500系列交换机CPU高问题定位思路

祝江华 2006-08-03 发表

关于S6500系列交换机CPU高问题定位思路

一、组网



二、问题描述

如上图，某电信组网中经常出现ping不通网关的情况，但出现这种情况也时随机的，并且故障出现的时候使用display cpu时发现主控板的cpu利用率非常高。

三、过程分析

1．使用display cpu查看cpu利用率：

```
[6506]dis cpu
Board 0 CPU busy status:
  100% in last 5 seconds
  100% in last 1 minute
  100% in last 5 minutes
Board 1 CPU busy status:
   18% in last 5 seconds
   20% in last 1 minute
   16% in last 5 minutes
Board 2 CPU busy status:
   29% in last 5 seconds
   28% in last 1 minute
   28% in last 5 minutes
```

可以看到主控板的cpu利用率高达100%，那么就需要找出是什么任务让cpu如此之忙?

2．查看是什么任务让cpu忙。

使用<6506>_**dis task 0**发现SOCK任务高。

```
32      SOCK    Event Sem       44059       1381       12697   76.8%
```

确认为上CPU报文太多导致。

是什么报文上CPU了呢?

3．用 debug rxtx event slot 0 看收发包信息，发现主控板**收包**的计数增长很快。因此，可以肯定是上CPU报文太多，造成CPU处理非常繁忙。

```
/-------- the first chip NetID = 0 -------/
Packets received from chip(U=unicast M=Multicast B=BroadCast):
 port      bcm(U)     bcm(M)     bcm(B)
    2        45         0          0
    3        45         0          0
    8        1123123    0          0
    9        45         0          0
Packets transmit to chip(U=unicast M=Multicast B=BroadCast):
 port      bcm(U)     bcm(M)     bcm(B)
```

2 45 0 0
3 45 0 0
8 45 0 0
9 45 0 0

/-------- the second chip NetID = 1 -------/
Packets received from chip(U=unicast M=Multicast B=BroadCast):
port bcm(U) bcm(M) bcm(B)

很明显，第8端口存在大量的单播报文上CPU。
Packets transmit to chip(U=unicast M=Multicast B=BroadCast):
port bcm(U) bcm(M) bcm(B)
8 1123123 0 0

4．使用deb rxtx –c 100 packet slot 0命令。
分析打印出来的报文，发现有大量相同的数据包上送到CPU。这些数据包都是来自vlan103的报文：

Rcv from NetID:0,Port:4,Length:60,reason 0,copy 0,BPDU 0
*0.85967996-RXTX-8-S0-pkt:
---
00 e0 fc 0b e5 a9 00 e0 fc 0b 0c 62 81 00 00 67
08 00 45 00 00 30 4b f4 40 00 07 06 f5 21 50 0d
55 61 dc c0 b0 83 10 a0 00 50 6e 45 90 2f 00 00
00 00 70 02 40 00 01 08 00 00 02 04 05 b4 01 01

各项数据为：
00 e0 fc 0b e5 a9---》目的MAC
00 e0 fc 0b 0c 62---》原MAC
81 00 00 67--------》TAG报文，VLAN为67（HEX）= 103
08 00----------》IP报文
50 0d 55 61--------》 原IP地址：80.13.85.97
通过网络工具来不难分析出这些十六进制数据，可以得到数据报文的VLAN、SIP、DIP、SMAC、DMAC以及协议号等信息，从VLAN来看这些数据是来自上行设备的数据，分析网络配置发现在vlan103存在路由环，这样数据在上行设备和S6506之间来回转发，直到TTL为0。

大量TTL等于0报文送到CPU，这种情况基本是由路由环路导致。进一步分析组网和配置信息。
防火墙上对于内网设置静态路由：
ip route-static 10.0.0.0 255.0.0.0 10.0.0.2 preference 60，
对于内网的流量指向6506。

6506上配置缺省路由：
ip route-static 0.0.0.0 0.0.0.0 10.0.0.1 preference 60，
指向防火墙。

这样的配置当6506下挂的子网10.0.1.0/24与6506的连接断开时，该三层接口Down，导致6506上该子网的直连路由10.0.1.0/24 DIRECT 0 0 10.0.1.1 Vlan-interface2 消失。这样当其他子网的主机访问该网段主机时就会产生路由环。
例如10.0.2.0/24网段的主机10.0.2.123 和 子网10.0.1.0/24的主机10.0.1.123通信。该IP报文在6506上会匹配到默认路由ip route-static 0.0.0.0 0.0.0.0 10.0.0.1 preference 60，从而转发到防火墙同时TTL减1；而防火墙发现该报文是内网的，就会匹配静态路由ip route-static 10.0.0.0 255.0.0.0 10.0.0.2 preference 60 ，将该报文再次转发给6506同时TTL减1。 这样该报文就会在6506和防火墙之间来回转发，直到TTL等于0 ；

四、结论与解决方法：
该路由环的产生的根本原因是由于防火墙上的路由网段设置过大，当6506上的直连路由消失后防火墙上的路由没有刷新，导致路由环。
解决的方法是在6506上配置黑洞路由：
ip route-static 10.0.0.0 255.0.0.0 NULL 0 blackhole

此时显示6506上的路由表为：
Destination/Mask Protocol Pre Cost Nexthop Interface
0.0.0.0/0 STATIC 60 0 10.0.0.1 Vlan-interface1
10.0.0.0/8 STATIC 60 0 0.0.0.0 NULL0
10.0.0.0/24 DIRECT 0 0 10.0.0.1 Vlan-interface1

10.0.0.1/32 DIRECT 0 0 127.0.0.1 InLoopBack0
10.0.1.0/24 DIRECT 0 0 10.0.1.1 Vlan-interface2
10.0.1.1/32 DIRECT 0 0 127.0.0.1 InLoopBack0
10.0.2.0/24 DIRECT 0 0 10.0.2.1 Vlan-interface3
10.0.2.1/32 DIRECT 0 0 127.0.0.1 InLoopBack0
10.0.3.0/24 DIRECT 0 0 10.0.3.1 Vlan-interface4
10.0.3.1/32 DIRECT 0 0 127.0.0.1 InLoopBack0

当6506下挂的子网10.0.1.0/24与6506的连接断开时，路由表变为：
Destination/Mask Protocol Pre Cost Nexthop Interface
0.0.0.0/0 STATIC 60 0 10.0.0.1 Vlan-interface1
10.0.0.0/8 STATIC 60 0 0.0.0.0 NULL0
10.0.0.0/24 DIRECT 0 0 10.0.0.1 Vlan-interface1
10.0.0.1/32 DIRECT 0 0 127.0.0.1 InLoopBack0
10.0.2.0/24 DIRECT 0 0 10.0.2.1 Vlan-interface3
10.0.2.1/32 DIRECT 0 0 127.0.0.1 InLoopBack0
10.0.3.0/24 DIRECT 0 0 10.0.3.1 Vlan-interface4 10.0.3.1/32 DIRECT 0
0 127.0.0.1 InLoopBack0

相比原来少了两条路由：
10.0.1.0/24 DIRECT 0 0 10.0.1.1 Vlan-interface2
10.0.1.1/32 DIRECT 0 0 127.0.0.1 InLoopBack0

这样，当10.0.2.0/24网段的主机10.0.2.123 和 子网10.0.1.0/24的主机10.0.1.123通信时，报文会在6506上先匹配到黑洞路由（**而不是缺省路由**）
10.0.0.0/8 STATIC 60 0 0.0.0.0 NULL0

这些报文被设备丢弃，避免路由环路的产生。