某局点S12500X-AF交换机转发业务延迟大问题

直通转发 张文宁 2021-12-14 发表

组网及说明

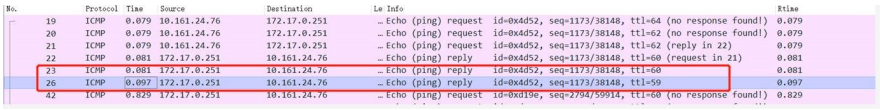
不涉及

问题描述

现场反馈新上的业务中，部分存在10ms+的转发时延，未影响业务。

过程分析

根据现场多次的抓包和测试，最终明确了确实是经过设备2框slot 0槽位的时候，就会产生10ms+的时延：



No.	Protocol	Time	Source	Destination	Le	Info	Rtime
19	ICMP	0.079	10.161.24.76	172.17.0.251	…	Echo (ping) request id=0x4d52, seq=1173/38148, ttl=64 (no response found!)	0.079
20	ICMP	0.079	10.161.24.76	172.17.0.251	…	Echo (ping) request id=0x4d52, seq=1173/38148, ttl=62 (no response found!)	0.079
21	ICMP	0.079	10.161.24.76	172.17.0.251	…	Echo (ping) request id=0x4d52, seq=1173/38148, ttl=62 (reply in 22)	0.079
22	ICMP	0.081	172.17.0.251	10.161.24.76	…	Echo (ping) reply id=0x4d52, seq=1173/38148, ttl=60 (request in 21)	0.081
23	ICMP	0.081	172.17.0.251	10.161.24.76	…	Echo (ping) reply id=0x4d52, seq=1173/38148, ttl=60	0.081
26	ICMP	0.097	172.17.0.251	10.161.24.76	…	Echo (ping) reply id=0x4d52, seq=1173/38148, ttl=59	0.097
42	ICMP	0.829	172.17.0.251	10.161.24.76	…	Echo (ping) request id=0xd19e, seq=2794/59914, ttl=60 (no response found!)	0.829

在上联设备将9台服务器下一跳都指定为S12504备框slot0板卡的下一跳，服务器出现时延问题率为100%：

```
NHZ07_M05_7368_FN_CSW2(s1)(config)#sh run | in route
ip route 10.161.24.74/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.75/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.76/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.82/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.83/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.84/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.85/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.86/32 Ethernet4/4/1 10.239.170.86
ip route 10.161.24.87/32 Ethernet4/4/1 10.239.170.86
route-map rp_bgpadv_to_br permit 10
route-map rp_bgprec from br permit 10
```

然后现场就保留其它3个ecmp，只删除2框slot 0这个下一跳后，问题消失：

```
NHZ07_M05_7368_FN_CSW3(s1)#
NHZ07_M05_7368_FN_CSW3(s1)#sh ip route 10.161.24.76

VRF: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B - BGP, B I - iBGP, B E - eBGP,
       R - RIP, I L1 - IS-IS level 1, I L2 - IS-IS level 2,
       O3 - OSPFv3, A B - BGP Aggregate, A O - OSPF Summary,
       NG - Nexthop Group Static Route, V - VXLAN Control Service,
       DH - DHCP client installed default route, M - Martian,
       DP - Dynamic Policy Route, L - VRF Leaked,
       RC - Route Cache Route

 S        10.161.24.76/32 [1/0] via 10.239.170.90, Ethernet4/4/1

NHZ07_M05_7368_FN_CSW3(s1)#configure
NHZ07_M05_7368_FN_CSW3(s1)(config)#no ip route 10.161.24.76/32 Ethernet4/4/1 10.239.170.90
NHZ07_M05_7368_FN_CSW3(s1)(config)#
NHZ07_M05_7368_FN_CSW3(s1)(config)#sh ip route 10.161.24.76

VRF: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B - BGP, B I - iBGP, B E - eBGP,
       R - RIP, I L1 - IS-IS level 1, I L2 - IS-IS level 2,
       O3 - OSPFv3, A B - BGP Aggregate, A O - OSPF Summary,
       NG - Nexthop Group Static Route, V - VXLAN Control Service,
       DH - DHCP client installed default route, M - Martian,
       DP - Dynamic Policy Route, L - VRF Leaked,
       RC - Route Cache Route

 B E      10.161.24.0/21 [20/0] via 10.239.170.26, Ethernet2/4/1
                                via 10.239.170.42, Ethernet3/3/1
                                via 10.239.170.90, Ethernet4/4/1
                                via 10.239.170.106, Ethernet5/3/1

NHZ07_M05_7368_FN_CSW3(s1)(config)#
```

上联设备删除下一跳路由后，路径不再走4/4/1这个端口

进一步查看设备信息，发现2框slot 0 上送cpu的报文超了线速：

====debug rxtx softcar show chassis 2 slot 0====

ID Type RcvPps PpsMax Rcv_All DisPkt_All Pps Dyn Swi Hash ACLmax

126 ROUTE_TO_CPU_MASK 606 1002 185242450 20800232 200 S On SMAC 8

The last discarded packet of ROUTE_TO_CPU_MASK :

0000 d4 61 fe 69 c6 01 d4 af f7 39 26 c3 81 00 0f ff

0010 08 00 45 00 00 3c 00 00 40 00 3d 06 f7 09 0a c2

0020 21 01 0a a1 10 4f 1a 0a c0 06 f4 91 c6 f4 57 a1

0030 58 76 a0 12 71 20 c3 de 00 00 02 04 05 b4 04 02

按理设备硬件转发，不应该这么多上送cpu的报文，怀疑是老版本parity error导致底层寄存器跳变软转了，进一步查看2框slot 0底层寄存器值，发现跳变为0了，确认为老版本已知问题：

[FAB_H3C_S12508X-probe]bcm ch 2 s 0 c 0

g/IHP_MACT_MANAGEMENT_UNIT_CONFIGURATION_REGISTER

IHP_MACT_MANAGEMENT_UNIT_CONFIGURATION_REGISTER.IHP0[0x2a3]=0x3800: <MACT_MNGMNT_UNIT_ENABLE=0,MACT_MNGMNT_UNIT_ACTIVE=0,FIELD_8_13=0x38>

路由黑洞走软转，是R1152H08解决的已知问题：

201912090347

问题现象：业务板部分端口不学习MAC。

问题产生条件：业务板问题端口所在交换芯片的LEM硬件表项产生parity error。

说明：无。

可以通过手工命令修复（该修复不影响业务，不会比错误状态下有更坏的影响）：

[probe]bcm ch 2 s 0 c 0 m/IHP_MACT_MANAGEMENT_UNIT_CONFIGURATION_REGISTER/MACT

解决方法

综上，老版本已知问题导致报文上送cpu，产生了转发时延。
1.规避措施：手工命令行修复

2.手工修复后，打上R1152H08补丁彻底解决此类问题。