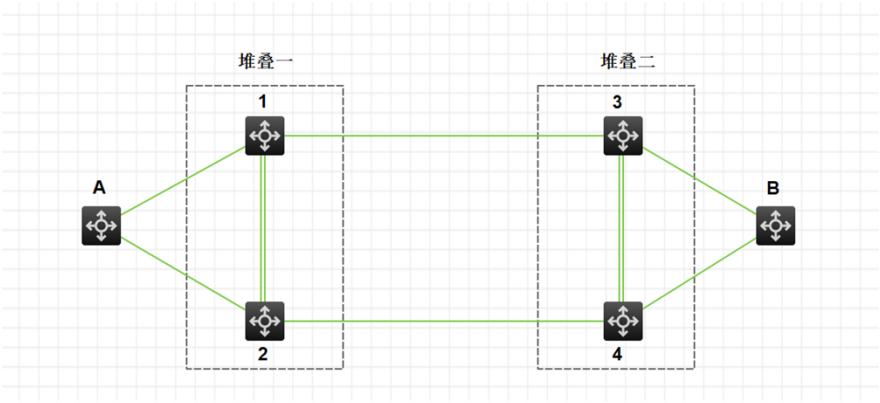
某局点S6800交换机EVPN组网堆叠切换丢包时间长

OSPF BFD IRF2 EVPN 苏亚东 2021-11-30 发表

组网及说明

某局点两组堆叠中间通过各个设备的48口相连，跑的是运营商的波分线路，用于连接两个数据中心，堆叠设备的48口做了三层动态聚合。并且两组堆叠设备间建立了用于数据中心间互联的VXLAN隧道，该隧道由两组堆叠设备间建立的EVPN自动创建。两组设备间underlay采用OSPF，并建立BGP EVPN邻居，借此建立起tunnel 0用于传输数据中心间流量。OSPF进程上配置了BFD和GR，以保证堆叠切换时能够减少丢包数量。两组设备下联分别用二层聚合接口连接A、B两台数据中心汇聚设备，并且在二层聚合口上起AC，用于数据中心间流量的加解封装。A、B设备上起三层聚合口用于对接DCI设备，并切在A、B三层聚合口上起同网段的地址互联。



问题描述

按照上述组网，A、B之间通信正常， 现场当时通过依次重启1、2、3、4设备的方式进行堆叠切换测试（即每次重启的均是堆叠主设备），测试方式为A、B之间通过同网段地址互ping。1、2、3设备的重启切换均正常，切换时间一般都在5-6秒左右。但是在进行4设备的重启时，发生以下现象：刚重启时A、B互ping丢包同样在5秒左右，然后恢复，但是过了10秒左右，又发生丢包，并持续20多秒，然后恢复。

过程分析

1、日志分析

%Oct 20 22:51:38:815 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to down.

%Oct 20 22:51:38:816 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to down. //开始丢包

%Oct 20 22:51:46:874 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to up.

%Oct 20 22:51:46:875 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to up. //恢复

%Oct 20 22:51:52:570 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to down.

%Oct 20 22:51:52:570 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to down. //重新丢包

%Oct 20 22:52:16:057 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to up.

%Oct 20 22:52:16:059 2021 XX-B01-N08-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to up. //重新恢复

从日志能看出该问题的原因是tunnel 0在整个过程中down/up了两次，并且第二次的时间有20多秒，与丢包的时长也吻合；而1、2、3设备重启时，均只有一次tunnel 0的down/up。

2、异常丢包分析

从反馈的信息来看，现场1-2堆叠，3-4堆叠，当设备3重启完成后再重启设备4。从日志中查看，故障时设备1-2上在22:51:46时OSPF邻居恢复full，tunnel 0就up起来了；但是过了大约5S后（22:51:52）OSPF邻居又变成了exstart，因而导致underlay网络中断、tunnel 0 down。又过了2S（22:51:54）OSPF邻居恢复，同时伴随着BFD会话由down→up，而直至22:52:11 BGP邻居重新建立，tunnel 0重新up后网络通信可达。查看设备3上对应时间点的日志信息可以看到，OSPF邻居down的同时也有bfd会话down的信息。

%Oct 20 22:51:38:696 2021 XX-B01-N08-DCI-ZDS-6800 BFD/5/BFD_CHANGE_FSM: Sess[10.130.254.1/10.130.254.6, LD/RD:2006/2004, Interface:RAGG100, SessType:Ctrl, LinkType:INET], Ver:1, Sta: UP->DOWN, Diag: 1 (Control Detection Time Expired)

%Oct 20 22:51:38:698 2021 XX-B01-N08-DCI-ZDS-6800 OSPF/6/OSPF_LAST_NBR_DOWN: OSPF 1 Last neighbor down event: Router ID: 10.130.253.101 Local address: 10.130.254.1 Remote address: 10.130.254.6 Reason: BFD session down.

初步怀疑OSPF邻居中断应该是因为BFD检测出现问题，导致BFD会话down了，因此将OSPF邻居给down了。

```
#
interface Route-Aggregation100
ospf 1 area 0.0.0.0
ospf bfd enable
link-aggregation mode dynamic
bfd min-transmit-interval 1000
bfd min-receive-interval 1000
bfd detect-multiplier 3
#
```

```
==============display bfd session verbose===============
Total Session Num: 2     Up Session Num: 1     Init Mode: Active

          Local Discr: 2006              Remote Discr: 2006
           Source IP: 1.1.1.1       Destination IP: 1.1.1.2
         Session State: Up                Interface: Route-Aggregation100
          Min Tx Inter: 1000ms            Act Tx Inter: 1000ms
          Min Rx Inter: 1000ms            Detect Inter: 3000ms
            Rx Count: 785                 Tx Count: 782                 //这里bfd报文收发数不一致，有可能是这个导致bfd down
         Connect Type: Direct          Running Up for: 00:11:18
           Hold Time: 2436ms              Auth mode: None
         Detect Mode: Async                 Slot: 1
           Protocol: OSPF
            Version: 1
          Diag Info: No Diagnostic
```

3、 复现分析

由于设备配置了bfd，且现场设备并未设置irf链路down延迟上报时间。根据手册中的说明，在存在bfd、GR等功能时，建议将irf link-delay设置为0，避免不必要的切换中断。

解决方法

1.2. 配置限制和指导

在两侧的堆叠设备上配置irf link-delay 0可以解决该问题。

如果某些协议配置的超时时间小于延迟上报时间（例如CFD、OSPF等），该协议将超时。此时请适当调整IRF链路down的延迟上报时间或者该协议的超时时间，使IRF链路down的延迟上报时间小于协议超时时间，保证协议状态不会发生不必要的切换。

下列情况下，建议将IRF链路down延迟上报时间配置为0：

- 对主备倒换速度和IRF链路切换速度要求较高时
- 在IRF环境中使用RRPP、BFD或GR功能时
- 在执行关闭IRF物理端口或重启IRF成员设备的操作之前，请首先将IRF链路down延迟上报时间配置为0，待操作完成后再将其恢复为之前的值

发现该问题后，现场已不具备继续测试的条件，于是实验室搭建环境进行复现，结果如下：

（1）未配置irf link-delay 0复现问题

经过几次主备切换，在1-2堆叠主设备重启过程中，在3-4堆叠打印如下：

Tunnel0 恢复后，又经过down,up

<QSH-NET06-DCI-ZDS-6800>%Nov 16 15:09:30:369 2021 QSH-NET06-DCI-ZDS-6800 LAGG/6/LAGG_INACTIVE_CONFIGURATION: Member port FGE1/0/49 of aggregation group RAGG100 changed to the inactive state, because the aggregation configuration of the port is incorrect.

%Nov 16 15:09:30:383 2021 QSH-NET06-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface FortyGigE1/0/49 changed to down.

%Nov 16 15:09:34:532 2021 QSH-NET06-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface FortyGigE1/0/49 changed to down.

%Nov 16 15:09:37:502 2021 QSH-NET06-DCI-ZDS-6800 BFD/5/BFD_CHANGE_FSM: Sess[10.130.254.6/10.130.254.1, LD/RD:2002/2002, Interface:RAGG100, SessType:Ctrl, LinkType:INET], Ver:1, Sta: UP->DOWN, Diag: 1 (Control Detection Time Expired)

%Nov 16 15:09:37:505 2021 QSH-NET06-DCI-ZDS-6800 OSPF/5/OSPF_NBR_CHG: OSPF 1 Neighbor 10.130.254.1(Route-Aggregation100) changed from FULL to DOWN.

%Nov 16 15:09:37:597 2021 QSH-NET06-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to down.

%Nov 16 15:09:37:598 2021 QSH-NET06-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to down.

%Nov 16 15:09:44:840 2021 QSH-NET06-DCI-ZDS-6800 OSPF/5/OSPF_NBR_CHG: OSPF 1 Neighbor 10.130.254.1(Route-Aggregation100) changed from LOADING to FULL.

%Nov 16 15:09:45:212 2021 QSH-NET06-DCI-ZDS-6800 BGP/5/BGP_STATE_CHANGED: BGP.: 10.130.253.1 state has changed from ESTABLISHED to IDLE for two connections exist and MD5 authentication is configured for the neighbor.

%Nov 16 15:09:45:531 2021 QSH-NET06-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to up.

%Nov 16 15:09:45:532 2021 QSH-NET06-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to up.

%Nov 16 15:09:48:946 2021 QSH-NET06-DCI-ZDS-6800 OSPF/5/OSPF_NBR_CHG: OSPF 1 Neighbor 10.130.254.1(Route-Aggregation100) changed from FULL to EXSTART.

%Nov 16 15:09:48:956 2021 QSH-NET06-DCI-ZDS-6800 OSPF/5/OSPF_NBR_CHG: OSPF 1 Neighbor 10.130.254.1(Route-Aggregation100) changed from LOADING to FULL.

%Nov 16 15:09:48:959 2021 QSH-NET06-DCI-ZDS-6800 BFD/5/BFD_CHANGE_FSM: Sess[10.130.254.6/10.130.254.1, LD/RD:2002/2004, Interface:RAGG100, SessType:Ctrl, LinkType:INET], Ver:1, Sta: DOWN->INIT, Diag: 0 (No Diagnostic)

%Nov 16 15:09:48:959 2021 QSH-NET06-DCI-ZDS-6800 BFD/5/BFD_CHANGE_FSM: Sess[10.130.254.6/10.130.254.1, LD/RD:2002/2004, Interface:RAGG100, SessType:Ctrl, LinkType:INET], Ver:1, Sta: INIT->UP, Diag: 0 (No Diagnostic)

%Nov 16 15:09:49:460 2021 QSH-NET06-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to down.

%Nov 16 15:09:49:460 2021 QSH-NET06-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to down.

%Nov 16 15:10:10:214 2021 QSH-NET06-DCI-ZDS-6800 BGP/5/BGP_STATE_CHANGED: BGP.: 10.130.253.1 state has changed from OPENCONFIRM to ESTABLISHED.

%Nov 16 15:10:14:386 2021 QSH-NET06-DCI-ZDS-6800 IFNET/3/PHY_UPDOWN: Physical state on the interface Tunnel0 changed to up.

%Nov 16 15:10:14:387 2021 QSH-NET06-DCI-ZDS-6800 IFNET/5/LINK_UPDOWN: Line protocol state on the interface Tunnel0 changed to up.

%Nov 16 15:11:32:002 2021 QSH-NET06-DCI-ZDS-6800 LLDP/5/LLDP_NEIGHBOR_AGE_OUT: Nearest bridge agent neighbor aged out on port FortyGigE1/0/49 (IfIndex 49), neighbor's chassis ID is 000f-0000-0002, port ID is FortyGigE1/0/49.

（2）在1-2，3-4堆叠配置irf link-delay 0后

多次主备切换过程，都没有打印tunnel0 down ,up的现象，故障消除
%Nov 17 10:04:47:689 2021 QSH-NET06-DCI-ZDS-6800 LAGG/6/LAGG_INACTIVE_CO