

组网及说明

接入m-lag-6850-G(ASW)和核心-S125G-AF(DSW)

告警信息

不涉及

问题描述

6850-G(ASW)和S125G-AF(DSW)上的日志，ASW侧BGP断开原因显示未收到keepalive报文超时断连，DSW侧收到远端发来的BGP断开消息状态切换。

ASW侧：

%Feb 23 15:57:01:273 2024 ASW-132-C03-1.AM11 BGP/5/BGP_STATE_CHANGED: BGP.: FD00:0:AC8:D038::AC8:D039 state has changed from ESTABLISHED to IDLE forhold timer expiration caused by peer device.

%Feb 23 15:57:01:273 2024 ASW-132-C03-1.AM11 BGP/5/BGP_STATE_CHANGED_REASON: BGP.: FD00:0:AC8:D038::AC8:D039 state has changed from ESTABLISHED to IDLE. (Reason: no keepalives or updates had been received from the peer when the hold timer expired, Error code: Send Notificationcode 4/0)

%Feb 23 15:57:59:273 2024 ASW-132-C03-1.AM11 BGP/5/BGP_STATE_CHANGED: BGP.: 10.200.208.49 state has changed from ESTABLISHED to IDLE for hold timer expiration caused by peer device.

%Feb 23 15:57:59:273 2024 ASW-132-C03-1.AM11 BGP/5/BGP_STATE_CHANGED_REASON: BGP.: 10.200.208.49 state has changed from ESTABLISHED to IDLE. (Reason:no keepalives or updates had been received from the peer when the hold timer expired, Error code: Send Notificationcode 4/0)

DSW侧：

%Feb 23 16:49:36:378 2024 DSW-VM-G1-P-1.SM132 BGP/5/BGP_STATE_CHANGED_REASON: BGP.: FD00:0:AC8:D054::AC8:D056 state has changed from ESTABLISHED to IDLE. (Reason: a notification was received from the peer, Error code: Receive Notificationcode 4/0)

%Feb 23 16:49:36:555 2024 DSW-VM-G1-P-1.SM132 BGP/5/BGP_STATE_CHANGED_REASON: BGP.: FD00:0:AC8:D098::AC8:D09A state has changed from ESTABLISHED to IDLE. (Reason: a notification was received from the peer, Error code: Receive Notificationcode 4/0)

%Feb 23 16:49:37:185 2024 DSW-VM-G1-P-1.SM132 BGP/5/BGP_STATE_CHANGED_REASON: BGP.: 10.200.208.186 state has changed from ESTABLISHED to IDLE. (Reason: a notification was received from the peer, Error code: Receive Notificationcode 4/0)

过程分析

查看ASW侧BGP断开log-info，基本都是收不到keepalive报文超时down，发送notification给对端，但是排查DSW诊断未发现异常，且其还和其他接入网关也有建立BGP邻居都是正常的，怀疑是ASW侧可能存在CPU处理BGP报文不及时问题。

<ASW-132-C03-1.AM11>dis bgp peer ipv4 10.200.209.125 log-info

Peer: 10.200.209.125

Date	Time	State	Notification
Error/SubError			
23-Feb-2024	18:59:48	Down	Send notification with error 4/0
Hold Timer Expired/ErrSubCode Unspecified			
Keepalive last received time : 18:59:12-2024.2.23			
Update last received time : 18:59:17-2024.2.23			
EPOLLIN last occurred time : 18:59:17-2024.2.23			
23-Feb-2024	18:57:15	Up	
23-Feb-2024	18:56:58	Down	Send notification with error 4/0
Hold Timer Expired/ErrSubCode Unspecified			
Keepalive last received time : 18:56:27-2024.2.23			
Update last received time : 18:56:27-2024.2.23			
EPOLLIN last occurred time : 18:56:27-2024.2.23			
23-Feb-2024	18:41:08	Up	
23-Feb-2024	18:40:45	Down	Send notification with error 4/0
Hold Timer Expired/ErrSubCode Unspecified			
Keepalive last received time : 18:40:14-2024.2.23			
Update last received time : 18:40:14-2024.2.23			
EPOLLIN last occurred time : 18:40:14-2024.2.23			

```
23-Feb-2024 18:38:07 Up
23-Feb-2024 18:37:44 Down Send notification with error 4/0
Hold Timer Expired/ErrSubCode Unspecified
Keepalive last received time : 18:37:13-2024.2.23
Update last received time : 18:37:13-2024.2.23
EPOLLIN last occurred time : 18:37:13-2024.2.23
23-Feb-2024 18:22:50 Up
23-Feb-2024 18:22:26 Down Send notification with error 4/0
Hold Timer Expired/ErrSubCode Unspecified
Keepalive last received time : 18:21:55-2024.2.23
Update last received time : 18:21:50-2024.2.23
EPOLLIN last occurred time : 18:21:55-2024.2.23
23-Feb-2024 18:19:10 Up
23-Feb-2024 18:18:53 Down Send notification with error 4/0
Hold Timer Expired/ErrSubCode Unspecified
Keepalive last received time : 18:18:17-2024.2.23
Update last received time : 18:18:22-2024.2.23
EPOLLIN last occurred time : 18:18:22-2024.2.23
```

查看ASW上有持续的上送cpu收包计数，分析是有异常报文冲击CPU，将上送cpu的报文打印出来，发现有很多TTL等于1的TCP报文，查看对应报文的目的ip在设备上表项，发现两台设备上该目的ip的arp都学到横联口，形成路由环路，该目的IP和现场确认是ASW下挂服务器地址。

10.200.198.3	0000-5e00-0101	9	BAGG100	20	D
10.200.198.4	0000-5e00-0101	9	BAGG100	18	D
10.200.198.10	0000-5e00-0101	9	BAGG100	44	D
10.200.198.21	0000-5e00-0101	9	BAGG100	20	D
10.200.198.22	0000-5e00-0101	9	BAGG100	16	D
10.200.198.28	0000-5e00-0101	9	BAGG100	44	D
10.200.198.34	0000-5e00-0101	9	BAGG100	84	D
10.200.198.42	0000-5e00-0101	9	BAGG100	30	D
10.200.198.44	0000-5e00-0101	9	BAGG100	90	D
10.200.198.46	0000-5e00-0101	9	BAGG100	90	D
10.200.198.48	0000-5e00-0101	9	BAGG100	90	D
10.200.198.50	0000-5e00-0101	9	BAGG100	90	D
10.200.198.51	0000-5e00-0101	9	BAGG100	18	D
10.200.198.53	0000-5e00-0101	9	BAGG100	16	D
10.200.198.54	0000-5e00-0101	9	BAGG100	16	D
10.200.198.55	0000-5e00-0101	9	BAGG100	88	D
10.200.198.59	0000-5e00-0101	9	BAGG100	86	D
10.200.198.65	0000-5e00-0101	9	BAGG100	30	D
10.200.198.66	0000-5e00-0101	9	BAGG100	28	D
10.200.198.68	0000-5e00-0101	9	BAGG100	28	D
10.200.198.73	0000-5e00-0101	9	BAGG100	38	D
10.200.198.74	0000-5e00-0101	9	BAGG100	16	D
10.200.198.75	0000-5e00-0101	9	BAGG100	64	D
10.200.198.76	0000-5e00-0101	9	BAGG100	40	D
10.200.198.77	0000-5e00-0101	9	BAGG100	68	D
10.200.198.78	0000-5e00-0101	9	BAGG100	60	D
10.200.198.80	0000-5e00-0101	9	BAGG100	90	D
10.200.198.86	0000-5e00-0101	9	BAGG100	84	D

```
#
interface Vlan-interface9 //arp的mac是网关mac
ip address 10.200.199.247 255.255.254.0
mac-address 0000-5e00-0101
local-proxy-arp enable
arp route-direct advertise
arp timer aging second 90
#
interface Bridge-Aggregation100 //聚合100是横联口
port link-type trunk
undo port trunk permit vlan 1
port trunk permit vlan 2 to 4094
link-aggregation mode dynamic
port m-lag peer-link 1
undo mac-address static source-check enable
#
```

经过实验室按照现场的配置复现打流测试，发现是因为现场配置的双活网关下配置了arp代理（local-proxy-arp enable），在该场景下，当上行DSW往ASW发送流量时，此时ASW设备如果没有学到下行服务器的arp，上行流量下来后因为网关配了arp代理，此时会往横联peer-link广播arp请求，m-lag对端网关也配了本地arp代理，这样对端又会再往回发送一份arp请求，这样目的ip的arp就会学在m-lag两边的peer-link口上，导致路由环路。路由环路会导致报文最终ttl减到1后上

送cpu处理，又由于当前设备BGP keepalive报文也是走的ttl=1报文上送cpu队列，如果异常ttl等于1报文过大会挤掉keepalive报文，导致keepalive收包超时，BGP断开。

解决方法

- 1、M-LAG双活网关组网环境下，M-LAG设备的下行VLAN接口不要配置本地代理ARP/ND功能，否则会触发流量环路。
- 2、升级软件版本，后续版本优化了keepalive上送cpu队列，保障BGP高优处理，建议升级到推荐版本R8307P08。