

M-LAG已逐步取代IRF走进千万张网
无论你是M-LAG小白
还是已经开局过M-LAG的大佬
都即将或者已经遇到各种各样的M-LAG问题
交换那些事儿M-LAG技术专题新系列
M-LAG典型案例及FAQ
每期推送2个案例+4个FAQ
让你在遇到类似问题时不迷茫不发愁
解决方案手到擒来~

本期内容

典型案例:

- Case1 M-LAG+EVPN组网leaf下终端互访丢包案例
- Case2 M-LAG+EVPN组网leaf下终端ARP学习异常案例

FAQ:

- Q1 M-LAG单挂场景下表项的学习和同步机制?
- Q2 二次故障时设备的处理机制?
- Q3 M-LAG系统MAC的推荐配置?
- Q4 M-LAG+VLAN双活网关ping下联主机不通?

DRNI更名为M-LAG

原DR口>>M-LAG接口
原DR group>>M-LAG group

原IPL链路>>peer-link链路
原IPP口>>peer-link接口

本文涉及命令行暂不做修改

Case1

M-LAG+EVPN组网leaf下终端互访丢包案例

组 / 网 / 说 / 明

设备及版本: S6805 F6631

组网: leaf-3和leaf-4起M-LAG作为一组leaf设备, 上行连接spine, 和远端leaf建立隧道, 下行M-LAG接口连接server。下图为局部组网图:



问 / 题 / 描 / 述

现场 M-LAG+EVPN 组网, leaf-3 和 leaf-4 起 M-LAG, 作为一组leaf设备, 发现远端leaf下挂终端二层访问本端leaf下终端存在丢包现象。现场工程师流统确认丢包时报文已经发送到了leaf-3/4上, 但leaf-3/4并未发给下联终端。

过 / 程 / 分 / 析

我们找到一条故障流进行详细分析, 该故障流的目的mac为fa16-3e9c-b63c, 是leaf-3/4通过M-LAG组BAGG2下联的终端。我们知道, 正常情况下, 这个mac表项在两台M-LAG设备上都应该学习在M-LAG接口上, 但是在现场设备上查看, leaf-3上学习正确, 但在leaf-4上mac却学习在了Tunnel2上:

```
<LEAF-03> display l2vpn mac-address | begin b63c
MAC Address : fa16-3e9c-b63c
VSI Name : SDN_VSI_100
State : Dynamic
Link ID/Name : Aging
BAGG2 Aging
< LEAF-04> display l2vpn mac-address | begin b63c
MAC Address : fa16-3e9c-b63c
VSI Name : SDN_VSI_100
State : EVPN
Link ID/Name : Aging
Tunnel2 NotAging
```

Tunnel2是个什么隧道呢? 我们在leaf-4上查看其详细信息, 发现它的源目IP分别为20.3.16.5和20.3.16.6:

```
Tunnel2
Current state: UP
Line protocol state: UP
Description: Tunnel2 Interface
...
Tunnel source 20.3.16.5, destination 20.3.16.6
Tunnel protocol/transport UDP_VXLAN/IP
```

Tunnel2的源IP为leaf-4上的虚拟VTEP地址, 目的地址为Leaf-3上配置的虚拟VTEP地址, 且Leaf-3上也存在对应的隧道Tunnel1:

```
leaf-4:
evpn drni group 20.3.16.5
Leaf-3:
evpn drni group 20.3.16.6
Tunnel1
Current state: UP
Line protocol state: UP
Description: Tunnel1 Interface
...
Tunnel source 20.3.16.6, destination 20.3.16.5
```

也就是说, 两台M-LAG设备leaf-3/4之间通过各自配置的虚拟VTEP地址建立了一条隧道。相信机智的小伙伴们已经发现了其中的不妥, 不急, 我们继续往下看。

排查过程中, 现场多次查看了mac地址表项, 发现该mac在不断漂移, 再次查看时leaf-4上学习正常, 而leaf-3上学到了Tunnel1上:

```
< LEAF-03> display l2vpn mac-address | begin b63c
MAC Address : fa16-3e9c-b63c
VSI Name : SDN_VSI_100
State : EVPN
Link ID/Name : Aging
Tunnel1 NotAging
< LEAF-04> display l2vpn mac-address | begin b63c
MAC Address : fa16-3e9c-b63c
VSI Name : SDN_VSI_100
State : Dynamic
Link ID/Name : Aging
BAGG2 Aging
```

排查到这里线索已经很明显, 我们基本就可以破案了。一组M-LAG的VTEP设备需要配置相同的VTEP虚地址以和远端设备建立隧道, 而现场在两台设备上配置了不同的虚拟VTEP地址, 导致这组M-LAG设备间建立了虚地址隧道。在正常情况下, 两台M-LAG设备上终端MAC都应该学习在AC口(即M-LAG接口), 而现场由于M-LAG设备间异常隧道的存在, 导致终端mac可以通过EVPN进行同步, 使得mac地址在AC口和隧道口之间漂移, 因此下行的流量在该组leaf上存在丢包现象。

解 / 决 / 方 / 法

按照规范配置, 将leaf-3、leaf-4上VTEP虚地址配置修正为相同地址后故障修复。

M-LAG+EVPN配置规范可参考:

交换那些事儿 | 技术专题篇 - DRNI+EVPN组网及配置介绍

交换那些事儿 | 技术专题篇 - DRNI+EVPN组网进阶介绍

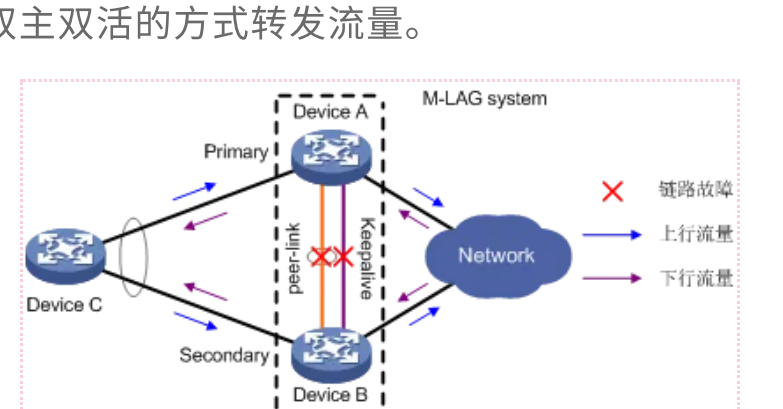
Case2

M-LAG+EVPN组网leaf下ARP学习异常案例

组 / 网 / 说 / 明

设备及版本: S6800 F2707+H0303

组网: 两台S6800起M-LAG作为leaf设备, 下联服务器网卡工作在主备模式, 即服务器单挂接入M-LAG。



问 / 题 / 描 / 述

服务器主备模式单挂接入M-LAG, 当服务器主备网卡倒换后, M-LAG设备上服务器的ARP没有正常迁移, 导致流量转发异常。现场服务器网卡迁移时不会主动发送ARP报文(即静默主机)。

过 / 程 / 分 / 析

问题处理过程中, 现场完整复现了一次故障现象。初始状态下, 服务器主网卡接在leaf-2设备的T1/0/14口, 此时两台M-LAG设备上表项如下, 分别学在peer-link接口和下行单挂口:

```
< leaf-1> display arp | inc 22.243.240.103
22.243.240.103 bc16-9524-f442 SDN_VSI_10702 BAGG98 -- R
< leaf-2> display arp | inc 22.243.240.103
22.243.240.103 bc16-9524-f442 SDN_VSI_10702 XGE1/0/14 1179 D
```

MAC表项也正常, 分别学习在peer-link接口和下行单挂口:

```
leaf-1:
bc16-9524-f442 EVPN SDN_VSI_10702 BAGG98 NotAging
leaf-2:
bc16-9524-f442 Dynamic SDN_VSI_10702 XGE1/0/14 Aging
```

随后现场将服务器主备倒换, 将主网卡迁移至leaf-1设备的下行口T1/0/14, 此时查看设备上MAC表项正常:

```
leaf-1:
bc16-9524-f442 Dynamic SDN_VSI_10702 XGE1/0/14 Aging
leaf-2:
bc16-9524-f442 EVPN SDN_VSI_10702 BAGG98 NotAging
```

但ARP就不对劲了, 两台M-LAG设备上的ARP都学习在了peer-link接口上:

```
< leaf-1> display arp | inc 22.243.240.103
22.243.240.103 bc16-9524-f442 SDN_VSI_10702 BAGG98 516 D
< leaf-2> display arp | inc 22.243.240.103
22.243.240.103 bc16-9524-f442 SDN_VSI_10702 BAGG98 783 R
```

服务器网卡迁移后, MAC表项迁移, 但ARP表项学习异常。

我们知道, 交换机是通过端口收到报文的源MAC来进行mac学习的, 服务器网卡迁移后, 只要主网卡在新端口发送报文, 即可触发MAC学习, 完成MAC迁移; 但现场服务器是静默主机, 迁移后不会主动发送ARP报文, 那么交换机就无法通过正常的ARP学习流程来进行ARP迁移。

这种情况下, MAC和ARP是可以联动迁移的, 当一个ARP表项的MAC迁移后, 也会同步刷新ARP表项。但当前版本M-LAG场景下ARP刷新流程存在异常, 导致ARP学习在两台M-LAG设备的peer-link接口。

解 / 决 / 方 / 法

临时规避方法: 配置服务器网卡迁移时主动发送ARP, 触发M-LAG设备正常ARP学习流程。

解决方案: 目前S6800设备已发布F2707H07补丁解决该问题, 取消了peer-link接口的MAC/ARP联动迁移。

- FAQ -

Q1: M-LAG单挂场景下表项的学习和同步机制?

A1:

ARP: 单挂终端ARP报文不会被Rlink报文封装同步给另一台M-LAG设备, M-LAG设备根据广播报文转发机制对单挂ARP报文进行正常泛洪, M-LAG设备的VLAN虚接口收到ARP报文后, 正常学习ARP表项到相应VLAN虚接口。
MAC: 单挂学习到的MAC表项会同步到另一台设备的IPP口, IPP口不主动学MAC。

Q2: 二次故障时设备的处理机制?

A2:

M-LAG二次故障是指在peer-link发生故障后, Keepalive链路也发生故障, 或者在Keepalive链路发生故障后, peer-link也发生故障。针对M-LAG设备上不同的配置情况, 当发生二次故障时, 处理方式不同。
缺省配置下, 若peer-link链路先发生故障, 此时两端M-LAG设备会根据Keepalive链路进行设备角色选举, 并依据MAD检测机制, 将从设备上除IRF保留接口、M-LAG保留接口外的所有接口置为M-LAG MAD DOWN状态。
此后, 若Keepalive链路也发生故障, 从设备也会升为主设备, 并解除设备上所有接口的M-LAG MAD DOWN状态, 以双主双活的方式转发流量。由于peer-link链路故障时, 无法同步表项, 可能导致流量转发错误。
若Keepalive链路先发生故障, peer-link链路后发生故障, 则M-LAG设备上的接口不会被置为M-LAG MAD DOWN状态, 而是直接以双主双活的方式转发流量。

Q3: M-LAG系统MAC的推荐配置?

A3:

通常情况下, M-LAG系统MAC配置没有限制, 只需要保证全网唯一, 不能和其它既有设备地址冲突; 一般推荐设置为M-LAG两台设备中较小的桥MAC。

Q4: M-LAG+VLAN双活网关ping下联主机不通?

A4:

M-LAG+VLAN双活场景下从网关去ping终端时, 可能出现不通, 这是正常现象, 原因是终端回复的报文可能会hash到另一台M-LAG设备, 由于VLAN双活场景下两台M-LAG设备的IP和MAC地址相同, 另一台M-LAG设备也能处理终端的reply报文。开局测试时, 建议在终端侧进行ping测试。

每多处理一个问题
对M-LAG的理解也更多一层
经验案例多看看
典型配置多学习
大家都可以成为大佬~
有问题欢迎留言哦~

END

扫码关注我們哦