M-LAG典型案例及FAQ（第5期）

H3C交换机技术专题篇

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+VRRP服务器主备网卡接入流量泛洪
- Case2 M-LAG peer-link故障时主设备被MAD DOWN

FAQ：

- Q1 M-LAG组网需要修改mac老化时间吗？
- Q2 园区核心设备M-LAG组网，单挂接口是否只能配置为trunk接口？
- Q3 M-LAG设备健康值要检查哪些？
- Q4 M-LAG设备间可以同步静态mac表项吗？

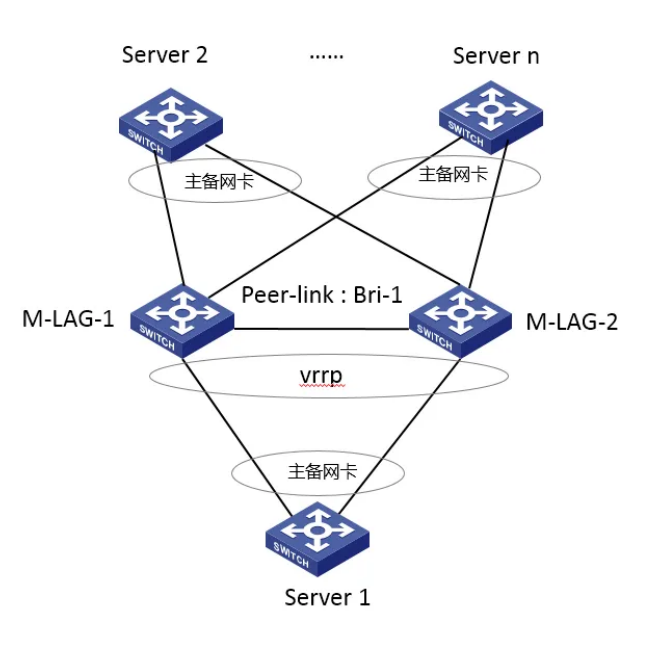
DRNI更名为M-LAG
原DR口>>M-LAG接口
原DR group>>M-LAG group
原IPL链路>>peer-link链路
原IPP口>>peer-link接口
本文涉及命令行暂不做修改

Case1 M-LAG+VRRP服务器主备网卡接入流量泛洪

组 / 网 / 说 / 明

设备及版本：S6850 R6635

组网：两台S6850配置M-LAG+VRRP作为服务器的网关，服务器采取主备网卡方式分别接在两台M-LAG设备上，M-LAG设备上没有配置M-LAG聚合口。



问 / 题 / 描 / 述

现场M-LAG+VRRP组网，两台S6850起M-LAG+VRRP作为服务器网关，服务器采取主备网卡方式分别连接两台M-LAG设备，M-LAG设备上不起M-LAG聚合口，相当于服务器单挂接入。现场开局时发现服务器使用主网卡访问时流量正常，使用备网卡访问时，Server1访问其他Server不通，且M-LAG从设备上其他还没上业务的接口出方向有大量泛洪流量。

查看从设备接口流量，发现其他接口WGE1/0/1-WGE1/0/17入方向几乎没有流量：

```
===========display counters rate inbound interface=========
Usage: Bandwidth utilization in percentage
Interface    Usage (%)  Total (pps)  Broadcast (pps)  Multicast (pps)
WGE1/0/1        0          0            --               --
WGE1/0/2        0          0            --               --
WGE1/0/3        0          0            --               --
WGE1/0/4        0          0            --               --
WGE1/0/5        0          0            --               --
WGE1/0/6        0          0            --               --
WGE1/0/7        0          0            --               --
WGE1/0/8        0          516          --               --
WGE1/0/9        0          0            --               --
WGE1/0/10       0          0            --               --
WGE1/0/17       0          0            --               --
```

但是这些接口的出方向流量都达到了21%：

```
==========display counters rate outbound interface==========
Usage: Bandwidth utilization in percentage
Interface    Usage (%)  Total (pps)  Broadcast (pps)  Multicast (pps)
WGE1/0/1        21       206896         --               --
WGE1/0/2        21       206903         --               --
WGE1/0/3        21       206854         --               --
WGE1/0/4        21       206876         --               --
WGE1/0/5        21       206959         --               --
WGE1/0/6        21       211652         --               --
WGE1/0/7        21       211650         --               --
WGE1/0/8        21       211636         --               --
WGE1/0/9        21       211653         --               --
WGE1/0/10       21       211653         --               --
WGE1/0/17       21       211656         --               --
```

在这些接口做outbound方向镜像抓包，发现报文均为Server1访问其他Server的报文，且报文目的mac均为网关VRRP的虚mac地址，说明Server1发出的报文从设备其他端口泛洪出去了。

过 / 程 / 分 / 析

查看从设备的VRRP状态，VRRP状态正常，在M-LAG+VRRP组网中，两台M-LAG设备底层都会下发VRRP虚mac，也就是说从设备收到目的mac为VRRP虚mac地址的报文时应该直接走三层转发，不应该泛洪：

```
==============display vrrp verbose==============
IPv4 Virtual Router Information:
 Running mode : Standard
Total number of virtual routers : 5
  Interface Vlan-interface100
    VRID           : 100            Adver Timer  : 100
    Admin Status   : Up             State        : Backup
    Config Pri     : 100            Running Pri  : 100
    Preempt Mode   : Yes            Delay Time   : 0
    Become Master  : 2940ms left
    Auth Type      : None
    Virtual IP     : 10.10.1.254
    Virtual MAC    : 0000-5e00-0001
    Master IP      : 10.10.1.1
```

那为什么从设备会将这部分流量泛洪呢？

我们在两台M-LAG设备上分别查看设备底层虚mac下发情况，发现M-LAG主设备上底层下发了VRRP虚mac地址，而M-LAG从设备的底层则没有该硬件表项：

```
M-LAG主设备表项：
[probe]debug l3intf-drv show virtualmac 0 slot 1
***********************************************************
- L3INTF Info by index  Slot 1
***********************************************************
- IntfId:      100
- VlanId:      100
- VmacStatus:  1
- RefCount:    1
- Vmac:        0000-5e00-0001
M-LAG从设备表项：
[probe]debug l3intf-drv show virtualmac 0 slot 1
***********************************************************
- L3INTF Info by index  Slot 1
***********************************************************
- IntfId:      0
- VlanId:      0
- VmacStatus:  0
- Vmac:        0x00-00-00-00-00-00
***********************************************************
```

从设备底层没有VRRP虚mac的硬件表项，当从设备收到Server1备网卡发过来的目的mac为VRRP虚mac的报文时，不认为目的mac是本机mac，因此没有走三层转发流程，而是走了二层转发；由于查不到该mac表项，就二层泛洪了。

那为什么从设备上底层没有下发VRRP虚mac表项呢？

现场服务器为主备网卡接入，相当于纯单挂接入，设备上也没有配置任何M-LAG聚合口。当前设备在纯单挂场景下存在如下限制：

服务器主备接入（bond1），并且配置VRRP网关的VLAN没有配置M-LAG接口时，需要创建一个M-LAG聚合组，并配置M-LAG聚合接口允许该VLAN通过。

例如：VLAN接口100配置了VRRP，但是没有任何M-LAG聚合接口加入VLAN 100，此时VLAN 100会出现预期外的未知单播流量泛洪，泛洪流量太大则可能导致丢包。

解 / 决 / 方 / 法

根据设备限制，创建一个M-LAG聚合接口，并配置聚合接口允许VLAN 100通过。后续我们计划开发一条专用命令来优化这个M-LAG纯单挂场景的限制。

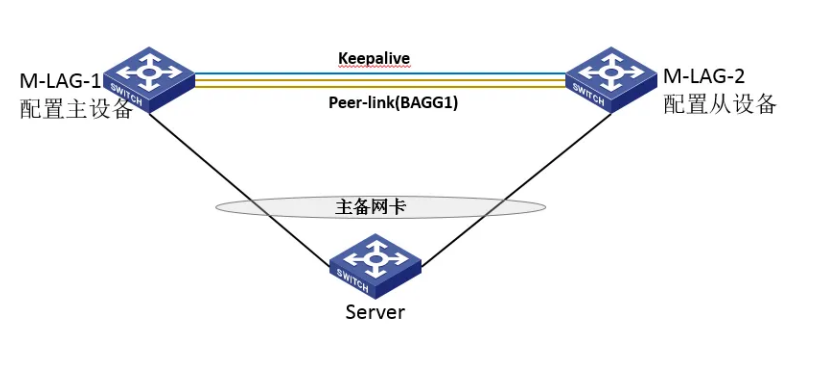
M-LAG+VRRP配置规范可参考：
交换那些事儿 | 技术专题篇 - DRNI+VRRP及DRNI+VLAN双活组网配置介绍

Case2 M-LAG peer-link故障时主设备被MAD DOWN

组 / 网 / 说 / 明

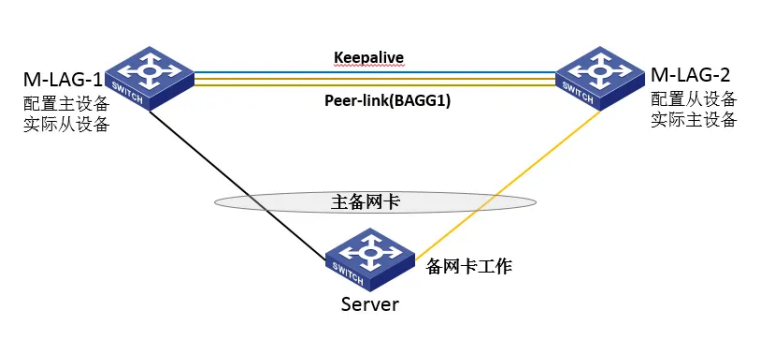
设备及版本：S12504G-AF R7624P15

组网：两台S12504G-AF起M-LAG，服务器主备网卡分别接M-LAG两台设备，没有配置M-LAG接口（和上文一样属于纯单挂场景）。下图为局部组网图：

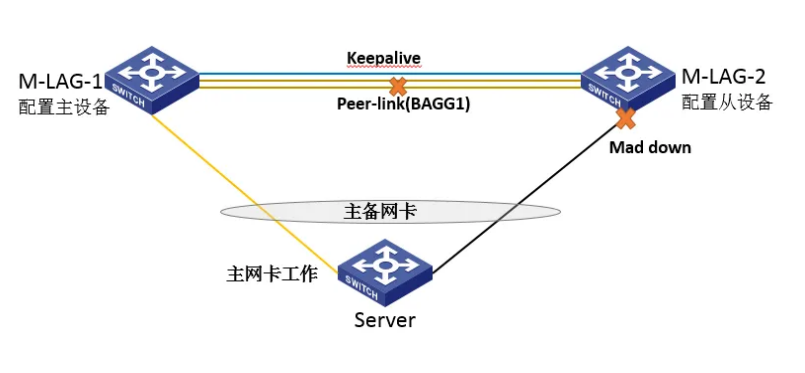


问 / 题 / 描 / 述

现场M-LAG组网，未起Vlan双活或者VRRP网关，服务器主备网卡分别接在M-LAG两台设备，相当于纯单挂接入场景。现场M-LAG-1为配置的主设备，M-LAG-2为配置从设备。但由于M-LAG-1之前重启过，发生过一次主备倒换，实际生效的角色M-LAG-2为主设备，M-LAG-1为从设备，因此实际流量到达服务器备网卡。



在这种场景下，现场peer-link链路因为链路故障down掉，keepalive链路正常，此时现场业务报障，一段时间后恢复。现场工程师查看业务状态，发现业务从备网卡切换到了主网卡，且发现原先生效的M-LAG主设备（即M-LAG-2）业务接口被MAD DOWN。



过 / 程 / 分 / 析

由于服务器主备网卡发生了倒换，与现场工程师确认，服务器主备倒换时自身需要对业务表项进行恢复，会导致业务中断。因此现场业务报障的直接原因是服务器主备网卡切换，主备网卡切换的原因又是因为M-LAG-2的业务口被MAD DOWN了。

此时有一个疑问，现场明明M-LAG-2是实际主设备，为什么peer-link故障时MAD DOWN的却是M-LAG-2设备?

这里我们再来详细研究一下M-LAG的故障处理机制，当peer-link故障而keepalive链路正常时，确实应该MAD DOWN从设备的接口，但这个从设备是当时计算出的从设备，而不是故障发生之前的从设备。

我们仔细去查阅官网配置指导，可以看到如下说明：

> **M-LAG角色计算触发条件包括：**
> - M-LAG设备在系统初始化时（包括新配置M-LAG或带M-LAG配置重启设备）。
> - peer-link链路UP时，设备角色通过peer-link计算。
> - peer-link故障，Keepalive正常工作，设备角色通过Keepalive链路计算。
> - peer-link和Keepalive链路均故障，根据本端M-LAG设备上M-LAG接口状态决定设备角色。

划重点了朋友们，peer-link故障，keepalive正常工作时，是会触发M-LAG角色计算的。也就是说，现场peer-link故障后，M-LAG-1和M-LAG-2设备的主从关系要通过keepalive链路重新计算，角色计算规则在配置手册中也有详细说明：

> 当通过peer-link或Keepalive链路交互报文计算设备角色时，依次比较如下因素：
>
> (1) 比较设备所有M-LAG接口的状态，有可工作M-LAG接口的一端为优；
>
> (2) 比较计算前角色，若有一端为Primary，另一端为None，则Primary端优；
>
> (3) 比较MAD DOWN状态，若一端存在处于MAD DOWN状态的接口，另一端不存在处于M-LAG MAD DOWN状态的接口，则不存在处于M-LAG MAD DOWN状态的接口的一端优；
>
> (4) 比较设备健康状况，健康值越小越优。设备的健康值可通过display system health命令查看，健康值越小设备越健康，设备无故障运行时，健康值为0；
>
> (5) 比较设备角色优先级，越高越优；
>
> (6) 比较设备桥MAC，越小越优。

根据这个计算规则，我们依次来看一下：

全单挂场景，两台设备均无可用M-LAG口，所以第1条pass；

在故障前不存在为none角色的设备和被MAD DOWN的设备，因此第2条、第3条规则也不满足；

由于故障前没有查看过设备健康度，所以第4条存疑，但故障前整个系统是正常运行的，大概率两台设备都没有问题，即健康度都为0；

M-LAG-1是配置的主设备，角色优先级配置得比M-LAG-2高，因此第5条M-LAG-1胜出；

由于第5条已经可以计算出两台设备的角色，第6条就没有比较的必要了。

从上述分析来看，现场peer-link故障后重新进行角色计算，很大可能M-LAG-1会被计算为主设备，那么M-LAG-2作为重新计算出的备设备被MAD DOWN就是正常的。

从以上信息可知，M-LAG的MAD DOWN实际会按照角色计算规则选举出一个从设备进行MAD DOWN，而非简单的将故障前实际生效的角色为从设备的接口给MAD DOWN。

解 / 决 / 方 / 法

M-LAG本身就是跨设备链路聚合，若业务侧切换时会存在恢复表象等导致业务中断的情况，建议及时修改组网为M-LAG聚合的形式；如果确有M-LAG全单挂的业务需求，建议尽可能保证业务主备与M-LAG配置的主从设备保持一致，同时注意前文所说的限制---创建一个M-LAG聚合接口，并配置聚合接口允许相应VLAN通过。

- FAQ -

Q1：M-LAG组网需要修改mac老化时间吗？
A1：

在M-LAG组网环境中，如果设备存在大量MAC地址表项，请通过mac-address timer aging命令增加MAC地址老化时间，建议配置MAC地址老化时间在20分钟以上。

Q2：园区核心设备M-LAG组网，单挂接口是否只能配置为trunk接口？
A2：

原本单挂接口需要与peer-link接口类型保持一致，peer-link一般为trunk口，因此单挂接口也只能配置为trunk接口。但当前园区核心设备M-LAG推荐的R7624Pxx及之后的版本，单挂接口可以使用access接口，无需与peer-link保持一致。

Q3：M-LAG设备健康值要检查哪些？
A3：

M-LAG角色计算时需要考虑设备健康值，健康值是指设备会定期执行一系列检查，根据检查项状态，判断设备的健康状态。健康值检查项包括：内存检测、进程异常检测、芯片堵死检测、CPU死循环检测、MMU表项错误检测、单板状态检测、单板间HG检测等等，可以通过display system health查看具体检查项：

表1-18 display system health命令显示信息描述表

字段	描述
Health: Normal(0)	系统的健康状态，取值为： • Normal(0)：表示系统健康 • Faulty(n)：表示系统不健康，n表示健康值，取值越大越不健康
Memory	内存检测
Communication	通信检测
DBM monitor	DBM进程异常检测
Chip jam	芯片堵死定时检测
CPU deadloop	CPU死循环检测
Forwarding channel	设备跨板业务检测
HG	单板间HG检测
External TCAM	转发芯片外部的TCAM检测
LSW CPU	LSW CPU故障检测
MMU	芯片表项MMU错误检测
LSW port	LSW端口检测
Board status	单板状态检测
Fan status	风扇状态检测
Temperature status	温度传感器状态检测

Q4：M-LAG设备间可以同步静态mac表项吗？
A4：

两端M-LAG设备之间不会同步静态MAC地址表项和黑洞MAC地址表项。

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

扫码关注我们哦