M-LAG 技术白皮书

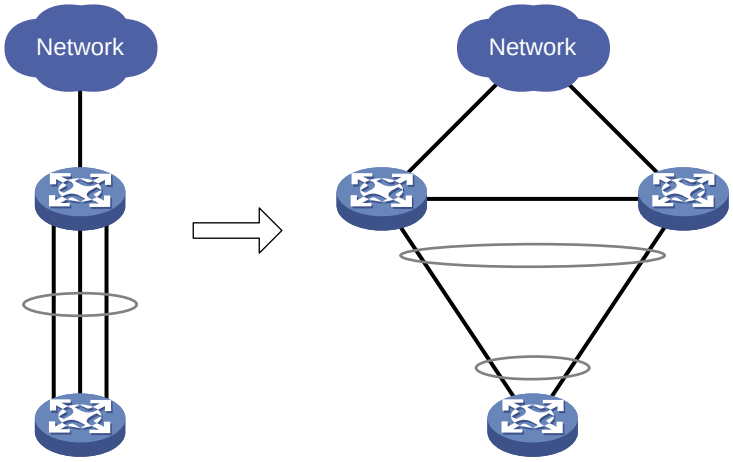
目 录

1 概述

1.1 产生背景

如图 1 所示，普通聚合的链路只能够在一台设备上，只能提供链路级的保护，当设备故障以后，普通聚合将无法工作，所以需要设备级保护的技术。M-LAG（Multichassis link aggregation，跨设备链路聚合）是基于 IEEE P802.1AX 协议的跨设备链路聚合技术。M-LAG 将两台物理设备虚拟成一台设备来实现跨设备链路聚合，从而提供设备级冗余保护和流量负载分担。

图1 链路聚合和 M-LAG 对比示意图



1.2 IRF和M-LAG对比

表 1 为 IRF 和 M-LAG 对比，组网可靠性要求高，升级过程要求业务中断时间短的场景推荐使用 M-LAG。在同一组网环境中，不能同时部署 IRF 和 M-LAG。

表1 IRF 和 M-LAG 对比

项目	IRF	M-LAG
控制面	• 所有成员设备控制面统一，集中管理 • 所有成员设备需要同步所有表项	• 两台独立设备，控制平面解耦 • 主要同步 MAC 表项/ARP 表项/ND 表项
设备面	紧耦合 • 硬件要求：芯片架构相同，一般要求同系列 • 软件要求：必须相同版本	松耦合 • 硬件要求：支持不同型号 • 软件要求：支持不同版本（由于 M-LAG 的特性支持情况还在快速发展阶段，现阶段部分产品要求相同版本）
版本升级	• 需要成员设备同步升级，或者主设备、从设备分开升级但操作较复杂 • 升级时业务中断时间 2 秒左右	可独立升级，升级时业务中断时间小于1s 对于支持GIR（Graceful Insertion and Removal，平滑插入和移除）的版本，可以做到不中断
配置管理	统一配置，统一管理，操作简单	独立配置，M-LAG系统会进行配置一致性检查，具

项目	IRF	M-LAG
	耦合度高，和控制器配合存在单点故障可能	体业务配置需要手工保证 独立管理，耦合度低，和控制器配合使用不存在单点故障，可靠性更高

说明

GIR 提供了一种设备隔离方案，适用于设备进行维护或升级的场景。通过 GIR 模式切换功能，可以一次下发多个业务模块的隔离命令，各业务协议模块会先将流量切换至冗余路径，再将设备置于维护模式，此时处于维护模式下的设备与其他设备之间网络隔离。当完成维护或者升级操作之后，将设备切换到普通模式，恢复流量的正常转发和处理。

1.3 技术优点

M-LAG 作为一种跨设备链路聚合的技术，除了具备增加带宽、提高链路可靠性、负载分担的优势外，还具备以下优势：

- 无环拓扑

  M-LAG 提供无环拓扑，即使在 M-LAG 组网中部署 STP，M-LAG 组网中的接口也不会被 STP 阻塞。
- 更高的可靠性

  把链路可靠性从单板级提高到了设备级。
- 双归接入

  允许设备双归接入，将两台设备的链路进行聚合，实现流量负载分担。
- 用户流量不中断

  M-LAG 组网中的接口、链路或者设备发生故障时，可将用户流量快速切换到正常设备/链路转发，确保用户业务不中断。
- 简化组网及配置

  提供了一个没有环路的二层拓扑，同时实现冗余备份，不再需要繁琐的防环协议配置，极大地简化了组网及配置。
- 独立升级

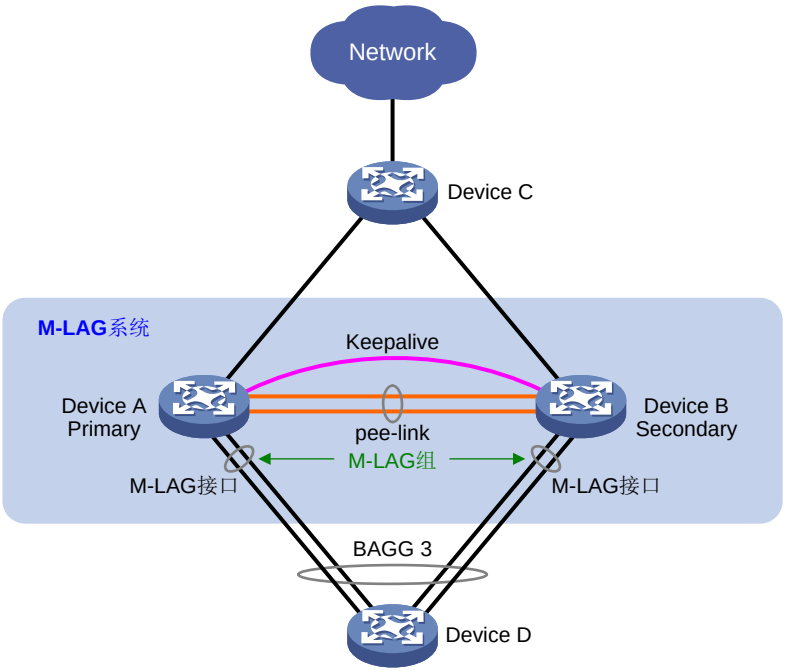
  两台设备可以分别进行升级，保证有一台设备正常工作即可，对正在运行的业务几乎没有影响。

2 M-LAG 技术实现

2.1 M-LAG基本概念

如图 2 所示，Device D 接入到 Device A 和 Device B 组成的 M-LAG 系统，通过 Device A 和 Device B 共同进行流量转发，保证网络的可靠性。

图2 M-LAG 组网示意图



M-LAG 涉及的相关概念如下：

- M-LAG 主设备：部署 M-LAG 且状态为 Primary 的设备。
- M-LAG 备设备：部署 M-LAG 且状态为 Secondary 的设备。
- peer-link 链路：M-LAG 设备间的交互 M-LAG 协议报文及传输数据流量的链路。peer-link 可以是聚合链路，也可以是 Tunnel 隧道，管理员需要根据不同组网环境选择 peer-link 链路。当采用聚合链路作为 peer-link 链路时，建议将多条链路进行聚合。一个 M-LAG 系统只有一条 peer-link 链路。
- peer-link 接口：peer-link 链路对应的接口，可以是聚合接口，也可以是 Tunnel 接口。每台 M-LAG 设备只有一个 peer-link 接口。
- Keepalive 链路：M-LAG 主备设备间的一条三层互通链路，用于 M-LAG 主备设备间检测邻居状态，即通过交互 Keepalive 报文来进行 peer-link 链路故障时的双主检测。
- M-LAG 组：用于部署 M-LAG 设备之间的配对，M-LAG 设备上相同编号的 M-LAG 接口属于同一 M-LAG 组。
- M-LAG 接口：M-LAG 主备设备与外部设备相连的二层聚合接口。为了提高可靠性，需要使用动态聚合。M-LAG 设备上相同编号的 M-LAG 接口属于同一 M-LAG 组。M-LAG 组 ID 为 M-LAG 接口编号。

说明

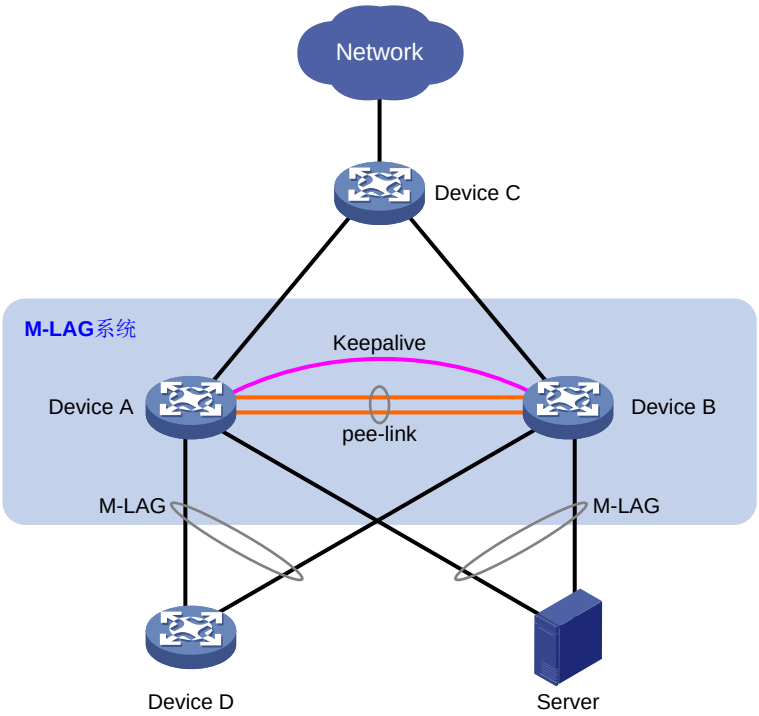
M-LAG 的角色区分为主和备，正常情况下，主设备和备设备同时进行业务流量的转发，转发行为没有区别，仅在故障场景下，主备设备的行为会有差别。

2.2 M-LAG网络模型

2.2.1 双归接入 M-LAG 网络模型

如图 3 所示，Device D 设备和 Server 分别双归接入 Device A 与 Device B 组成的 M-LAG 系统。Device A 与 Device B 形成负载分担，共同进行流量转发，当其中一台设备发生故障时，流量可以快速切换到另一台设备，保证业务的正常运行。

图3 双归接入 M-LAG 网络模型示意图



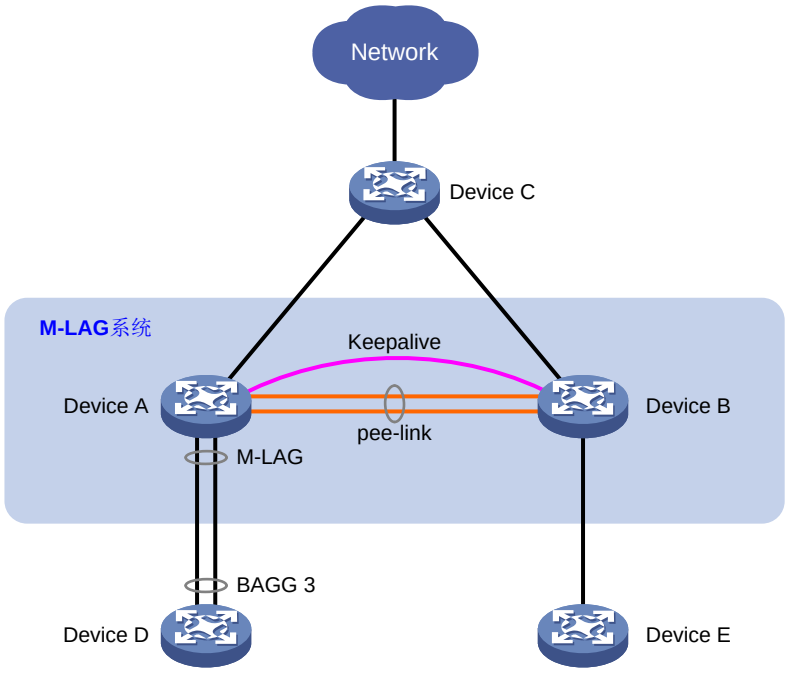
2.2.2 单归接入 M-LAG 网络模型

单归接入是指一台外部设备仅接入 M-LAG 系统的其中一台 M-LAG 设备。该外部设备称为单挂设备。根据接入接口的不同，单归接入分为：

- M-LAG 单归接入：通过 M-LAG 接口接入 M-LAG 系统的其中一台 M-LAG 设备。
- 非 M-LAG 单归接入：通过非 M-LAG 接口接入 M-LAG 系统的其中一台 M-LAG 设备。

如图 4 所示，Device D 设备以 M-LAG 单归接入方式接入 M-LAG 系统，Device E 设备以非 M-LAG 单归接入方式接入 M-LAG 系统。Device D 和 Device E 以单归方式接入 M-LAG 系统，Device D 和 Device E 的 MAC 地址、ARP/ND 等表项会 M-LAG 系统间进行备份，为南北向流量提供备份路径，提高可靠性。

图4 单归接入 M-LAG 网络模型示意图



2.3 M-LAG系统建立和维护

M-LAG 设备间通过交互 DRCP（Distributed Relay Control Protocol，分布式聚合控制协议）报文和 Keepalive 报文建立和维护 M-LAG 系统。在 M-LAG 系统正常工作时，M-LAG 系统的主备设备负载分担共同进行流量转发。如果 M-LAG 系统中出现故障（无论是接口故障、链路故障还是设备故障），M-LAG 系统都可以保证正常的业务不受影响。

2.3.1 M-LAG 系统建立及工作过程

如图 5 所示，Device A 和 Device B 之间 M-LAG 系统建立及工作过程如下：

(1) DRCP 协商

当 M-LAG 设备完成 M-LAG 系统参数配置后，两端设备通过 peer-link 链路定期发送 DRCP 报文。

(2) M-LAG 配对

当本端收到对端的 DRCP 协商报文后，会判断 DRCP 协商报文中的 M-LAG 系统配置是否和本端相同。如果两端的 M-LAG 系统配置相同，则这两台设备组成 M-LAG 系统。

(3) 主备协商

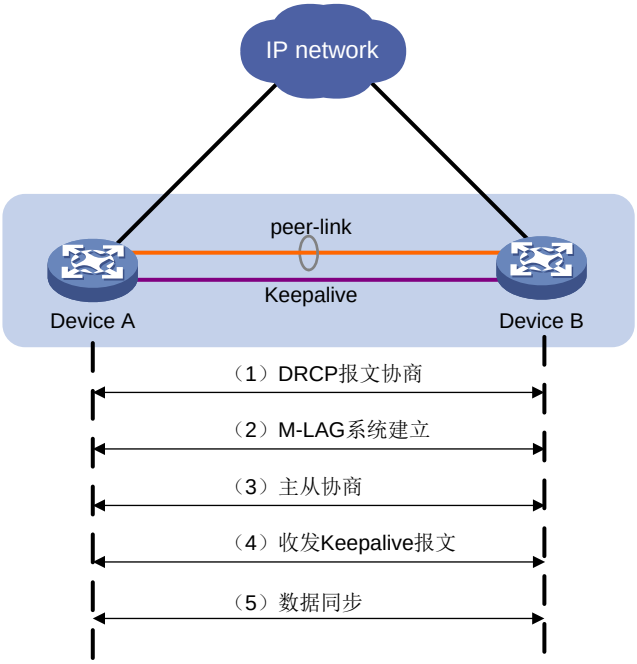
配对成功后，两端设备会确定出主备状态。依次比较两端 M-LAG 设备的初始角色、M-LAG MAD DOWN 状态、设备的健康值、角色优先级、设备桥 MAC，比较结果更优的一端为主设备。主备协商后，M-LAG 设备间会进行配置一致性检查。有关一致性检查的详细介绍，请参见“2.3.8 配置一致性检查功能”。

(4) 双主检测

当主备角色确定后，两端设备通过 Keepalive 链路周期性地发送 Keepalive 报文进行双主检测。

(5) M-LAG 系统开始工作后，两端设备之间会通过 peer-link 链路实时同步对端的信息，例如 MAC 地址表项、ARP 表项，从而确保任意一台设备故障都不会影响流量的转发，保证业务不会中断。

图5 M-LAG 建立及工作过程示意图



2.3.2 DRCP 协议

M-LAG 通过在 peer-link 链路上运行 DRCP 来交互分布式聚合的相关信息，以确定两台设备是否可以组成 M-LAG 系统。运行该协议的设备之间通过互发 DRCPDU（Distributed Relay Control Protocol Data Unit，分布式聚合控制协议数据单元）来交互分布式聚合的相关信息。

1. DRCPDU 的交互

两端 M-LAG 设备通过 peer-link 链路定期交互 DRCP 报文。当本端 M-LAG 设备收到对端 M-LAG 设备的 DRCP 协商报文后，会判断 DRCP 协商报文中的 M-LAG 系统配置是否和本端相同。如果两端的 M-LAG 系统配置均相同，则这两台设备可以组成 M-LAG 系统。

2. DRCP 超时时间

DRCP 超时时间是指 peer-link 接口等待接收 DRCPDU 的超时时间。在 DRCP 超时时间之前，如果本端 peer-link 未收到来自对端 M-LAG 设备的 DRCPDU，则认为对端 M-LAG 设备 peer-link 接口已经失效。

DRCP 超时时间同时也决定了对端 M-LAG 设备发送 DRCPDU 的速率。DRCP 超时有短超时（3 秒）和长超时（90 秒）两种：

- 若本端 DRCP 超时时间为短超时，则对端 M-LAG 设备将快速发送 DRCPDU（每 1 秒发送 1 个 DRCPDU）。
- 若本端 DRCP 超时时间为长超时，则对端 M-LAG 设备将慢速发送 DRCPDU（每 30 秒发送 1 个 DRCPDU）。

2.3.3 Keepalive 机制

M-LAG 设备间通过 Keepalive 链路检测邻居状态，即通过交互 Keepalive 报文来进行 peer-link 链路故障时的双主检测。

1. Keepalive 定时器

缺省情况下，Keepalive 各个定时器如表 2 所示。

表2 Keepalive 定时器描述表

Keepalive 定时器类型	定时器中文涵义	定时器缺省值
Keepalive interval	Keepalive报文发送的时间间隔	1秒
Keepalive hold timeout	peer-link链路down后等待检测故障原因的时间	3秒
Keepalive timeout	Keepalive报文超时时间间隔	5秒

2. Keepalive 实现机制

如果在 Keepalive timeout 时间内，本端 M-LAG 设备收到对端 M-LAG 设备发送的 Keepalive 报文：

- 如果 peer-link 链路状态为 down，则认为 peer-link 故障，启动 Keepalive hold timeout 定时器：
  - 在该定时器超时前收到 DRCP 报文，则 peer-link 链路状态恢复 up，M-LAG 系统正常工作。
  - 在该定时器超时前未收到 DRCP 报文，则本端和对端 M-LAG 设备根据收到的 Keepalive 报文选举主备设备，保证 M-LAG 系统中仅一台 M-LAG 设备转发流量，避免两台 M-LAG 设备均升级为主设备。
- 如果 peer-link 链路状态为 up，则 M-LAG 系统正常工作。

如果在 Keepalive timeout 时间内，本端 M-LAG 设备未收到对端 M-LAG 设备发送的 Keepalive 报文：

- 如果 peer-link 链路状态为 down，则认为对端 M-LAG 设备状态为 down，启动 Keepalive hold timeout 定时器，在该定时器超时后：
  - 本端设备为主设备时，如果本端设备上存在处于 up 状态的 M-LAG 口，则本端仍为主设备；否则，本端设备角色变为 None 角色。
  - 本端设备为备设备时，则升级为主设备。此后，只要本端设备上存在处于 up 状态的 M-LAG 口，则保持为主设备，否则本端设备角色变为 None 角色。

    当设备为 None 角色时，设备不能收发 Keepalive 报文，Keepalive 链路处于 down 状态。
- 如果 peer-link 链路状态为 up，则认为 Keepalive 链路状态为 down。此时主备设备正常工作，同时设备打印日志信息，提醒用户检查 Keepalive 链路。

2.3.4 MAD 机制

设备上接口在 M-LAG 系统分裂后有以下状态：

- M-LAG 系统分裂后接口处于 M-LAG MAD DOWN 状态。
- M-LAG 系统分裂后接口保持原状态不变。

peer-link 链路故障后，为了防止备设备继续转发流量，M-LAG 提供 MAD（Multi-Active Detection，多 Active 检测）机制，即在 M-LAG 系统分裂时将设备上部分接口置为 M-LAG MAD DOWN 状态，仅允许 M-LAG 口、peer-link 接口等接口转发流量，避免流量错误转发，尽量减少对业务影响。如果希望 M-LAG 系统中有特殊用途的接口（比如 Keepalive 接口）保持 up 状态，可以将其指定为 M-LAG 保留接口。

M-LAG 系统分裂时，设备上以下接口不被置为 M-LAG MAD DOWN 状态：

- M-LAG 保留接口（包括用户配置的和系统保留的）。
- 配置了强制端口 up 功能的接口。

M-LAG 保留接口包括系统保留接口和用户配置的保留接口。系统保留接口包括：

- peer-link 接口
- peer-link 接口所对应的二层聚合接口的成员接口
- M-LAG 接口
- 管理以太网接口

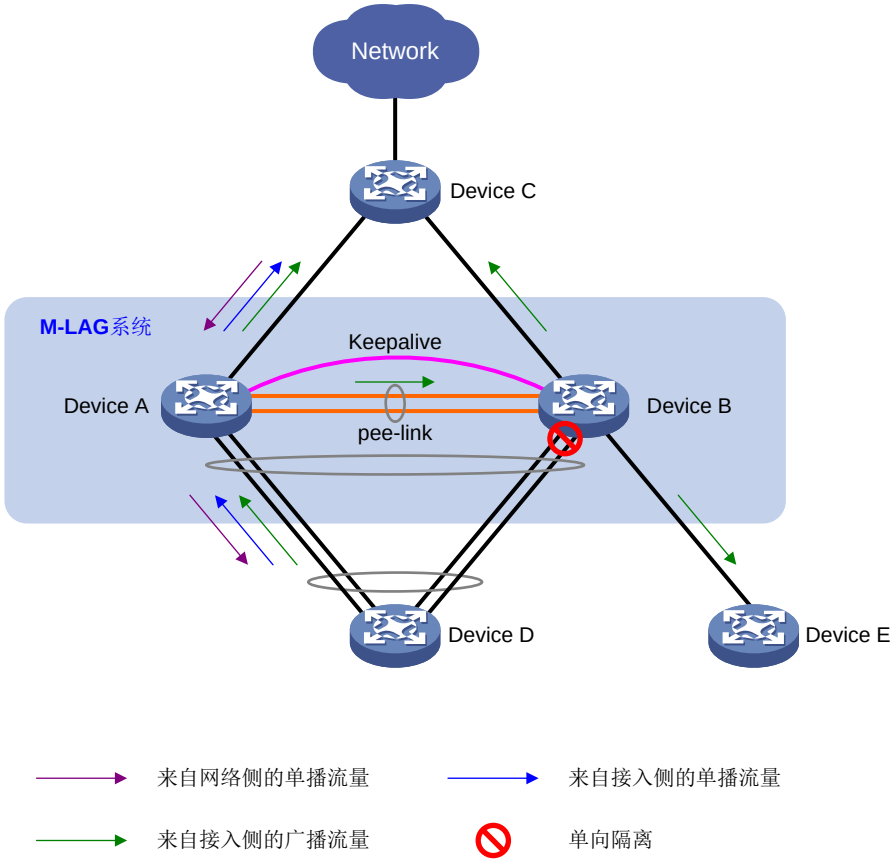
当 peer-link 链路故障恢复后，为了防止丢包，备设备尽可能在延迟恢复时间内完成表项（MAC 地址表、ARP 表等）同步，其后该设备上处于 M-LAG MAD DOWN 状态的接口将恢复为 up 状态。

在 EVPN VXLAN 支持 M-LAG 组网环境中，当使用 VXLAN 隧道作为 peer-link 链路时，为了保证 M-LAG 系统分裂后 M-LAG 设备能够正常工作，需要将大量逻辑接口（例如 Tunnel 接口或 LoopBack 接口）配置为保留接口。此时，为了减少配置工作量，可以指定部分无关接口在 M-LAG 系统分裂后处于 M-LAG MAD DOWN 状态，指定其他接口在 M-LAG 系统分裂后保持原状态不变。

2.3.5 M-LAG 防环机制

M-LAG 本身具有防环机制，可以构造出一个无环网络。如图 6 所示，从接入设备或网络侧到达 M-LAG 设备的单播流量，会优先从本地转发出去，peer-link 链路一般情况下不用来转发数据流量。当流量通过 peer-link 链路转发到对端 M-LAG 设备，在 peer-link 链路与 M-LAG 接口之间设置单方向的流量隔离，即从 peer-link 接口进来的流量不会再从 M-LAG 接口转发出去，所以不会形成环路，这就是 M-LAG 单向隔离机制。

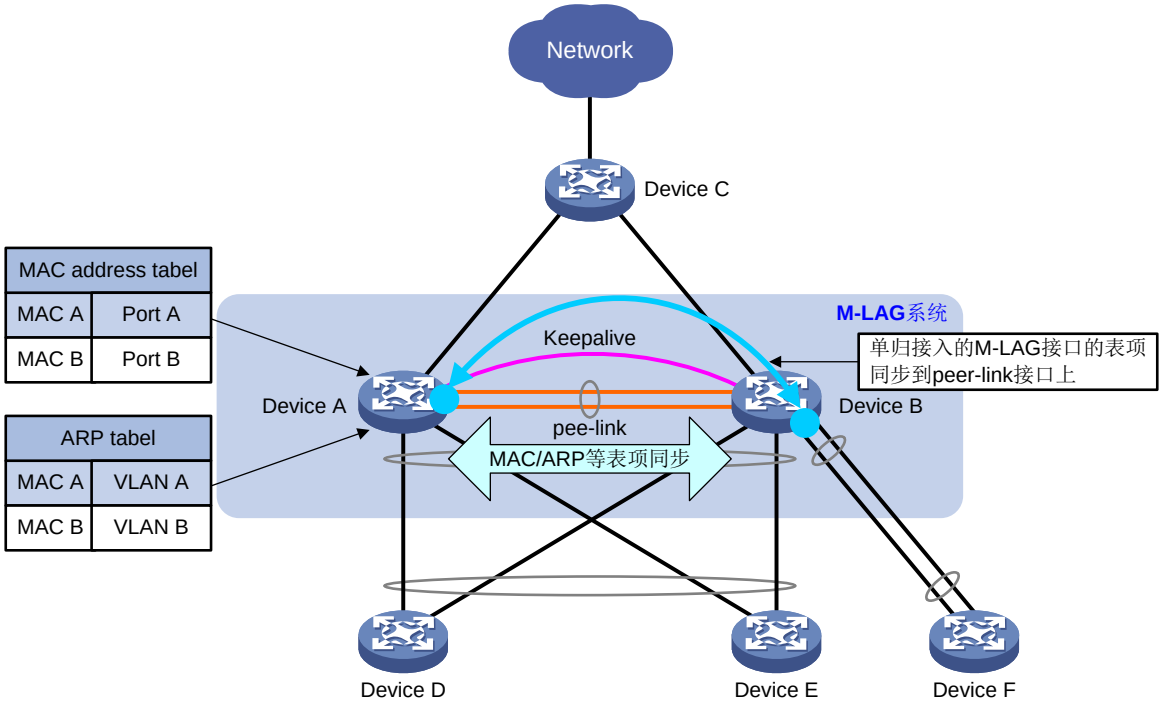
图6 M-LAG 防环机制示意图



2.3.6 M-LAG 表项同步机制

如图 7 所示，M-LAG 主备设备之间会实时同步 MAC 地址表项、ARP 表项、DHCP 表项、ND 表项等表项。单归接入场景中，单归接入的 M-LAG 接口的表项将同步到对端设备的 peer-link 接口上，以便下行流量绕行 peer-link 链路转发到 Device F。

图7 M-LAG 表项同步机制示意图



2.3.7 M-LAG 设备工作模式

M-LAG 设备工作模式分为以下两种：

- M-LAG 系统工作模式：作为 M-LAG 系统成员设备参与报文转发。接入设备与 M-LAG 设备交互的 LACPDU（Link Aggregation Control Protocol Data Unit，链路聚合控制协议数据单元）中，LACDU 携带的 LACP System ID 由 M-LAG 系统 MAC 地址和 M-LAG 系统优先级组成。
- 独立工作模式：脱离 M-LAG 系统独立工作，独自转发报文。接入设备与 M-LAG 设备交互的 LACPDU（Link Aggregation Control Protocol Data Unit，链路聚合控制协议数据单元）中，LACDU 携带的 LACP System ID 由 LACP 系统 MAC 地址和 LACP 系统优先级组成。

当 M-LAG 系统分裂时，为了避免 M-LAG 系统中的两台设备都作为主设备转发流量的情况，需要 M-LAG 设备独立工作。在 peer-link 链路和 Keepalive 链路均处于 DOWN 状态时，备设备会立即或经过一段时间切换到独立运行模式。

M-LAG 设备切换到独立运行模式后，聚合接口发送的 LACPDU 中携带的 M-LAG 系统参数还原为聚合接口的 LACP 系统 MAC 地址和 LACP 系统优先级，使同一 M-LAG 组中的两个聚合接口的 LACP 系统 MAC 地址和 LACP 系统优先级不一致。这样只有一边聚合接口的成员端口可以被选中，通过被选中的设备转发业务流量，避免流量转发异常。

2.3.8 配置一致性检查功能

M-LAG 系统建立过程中会进行配置一致性检查，以确保两端 M-LAG 设备配置匹配，不影响 M-LAG 设备转发报文。M-LAG 设备通过 peer-link 链路交换各自的配置信息，检查配置是否匹配。目前 M-LAG 支持对两种类型的配置进行一致性检查：

- 关键配置：Type 1 类型配置，即影响 M-LAG 系统转发的配置。如果 Type 1 类型配置不匹配，则将备设备上 M-LAG 接口置为 down 状态，将导致链路状态正常但是长时间丢包等问题。
- 一般配置：Type 2 类型配置，即仅影响业务模块的配置。如果 Type 2 类型配置不匹配，备设备上 M-LAG 接口依然为 up 状态，不影响 M-LAG 系统正常工作。与 Type 1 类型配置相比而言，Type 2 类型配置对网络环境影响较小。Type 2 类型配置仅影响其对应的业务模块功能。

2.3.9 M-LAG 双活网关

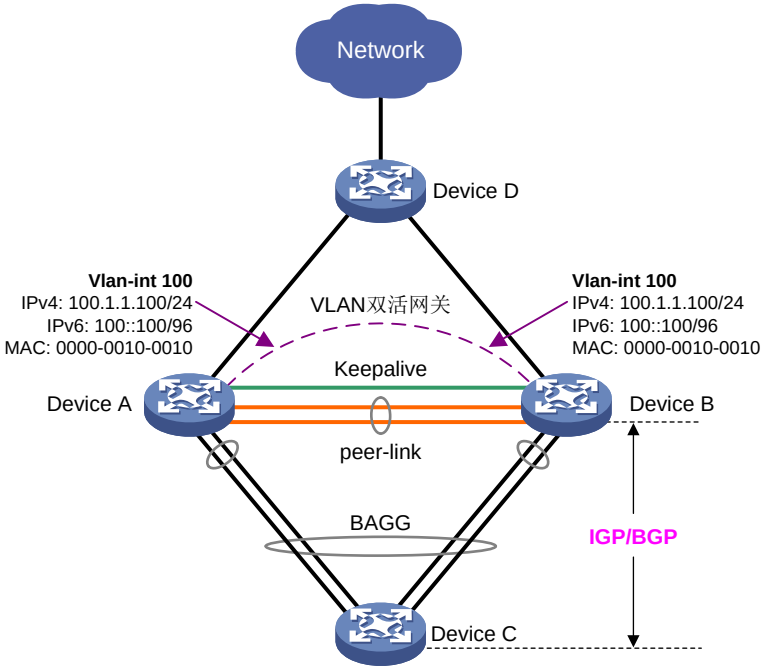
在 M-LAG 双归接入三层网络的场景中，两台 M-LAG 设备需要同时作为三层网关，必须保证 M-LAG 设备上存在相同的 IP 地址和 MAC 地址的逻辑接口，以便实现：

- 当一条接入链路发生故障时，流量可以快速切换到另一条链路，保证可靠性。
- 两条接入链路可以同时处理用户流量，以提高带宽利用率，使流量在两条接入链路上负载分担。

M-LAG 双活网关主要用于接入侧设备通过动态路由接入 M-LAG 的组网环境中。如图 8 所示，用户可以在 Device C 上部署静态路由通过三层路由方式接入到 M-LAG 系统，但部署静态路由将带来运维难度的上升和缺乏灵活快速部署能力，无法满足快速增长的业务需要。为解决静态路由表带来的问题，需要在 M-LAG 系统与用户侧设备之间建立动态路由协议邻居：

- M-LAG 设备 Device A 和 Device B 上各创建一个相同编号的 VLAN 接口（例如 Vlan-interface100），该接口作为网关接口，具有相同的 IPv4 地址、IPv6 地址和 MAC 地址，且 M-LAG 接口允许该 VLAN 通过。
- 在该 VLAN 接口下配置 IP 地址不同的 M-LAG 虚拟地址，用于 IGP/BGP 动态路由协议邻居建立，使得 M-LAG 设备和 Device C 之间建立 IGP/BGP 邻居，M-LAG 设备之间也会建立 IGP/BGP 邻居。

图8 动态路由接入 M-LAG 示意图



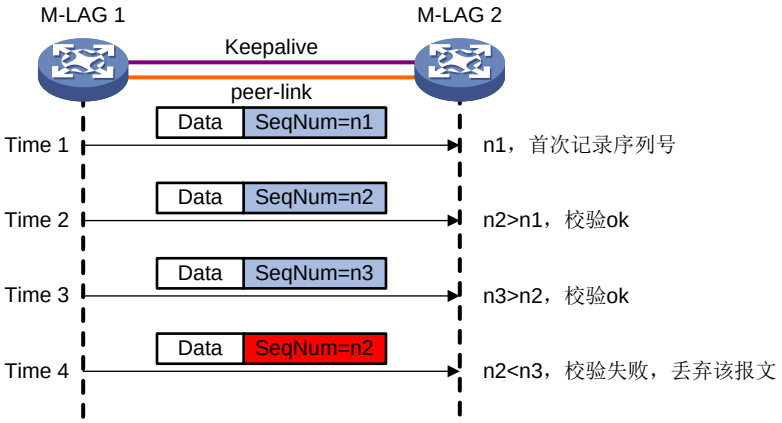
2.3.10 M-LAG 序列号校验

当网络入侵者从网络上截取 M-LAG 1 发送给 M-LAG 2 的 DRCP 报文/Keepalive 报文，并将这些报文发送给 M-LAG 2，使 M-LAG 2 误以为入侵者就是 M-LAG 1，然后 M-LAG 2 向伪装成 M-LAG 1 的入侵者发送应当发送给 M-LAG 1 的报文。M-LAG 1 接收不到 M-LAG 2 发送的 DRCP 报文/Keepalive 报文，导致 M-LAG 系统分裂，出现双主双活问题。

为了防止重放攻击，保证流量正常转发，M-LAG 支持序列号校验，以识别非法攻击报文。

如图 9 所示，如果 M-LAG 设备本次收到的 DRCP 报文/Keepalive 报文的序列号与已经收到的 DRCP 报文/Keepalive 报文的序列号相同，或小于上次收到的 DRCP 报文/Keepalive 报文的序列号，则认为发生重放攻击。M-LAG 设备会丢弃序列号校验失败的 DRCP 报文/Keepalive 报文。

图9 M-LAG 序列号校验示意图

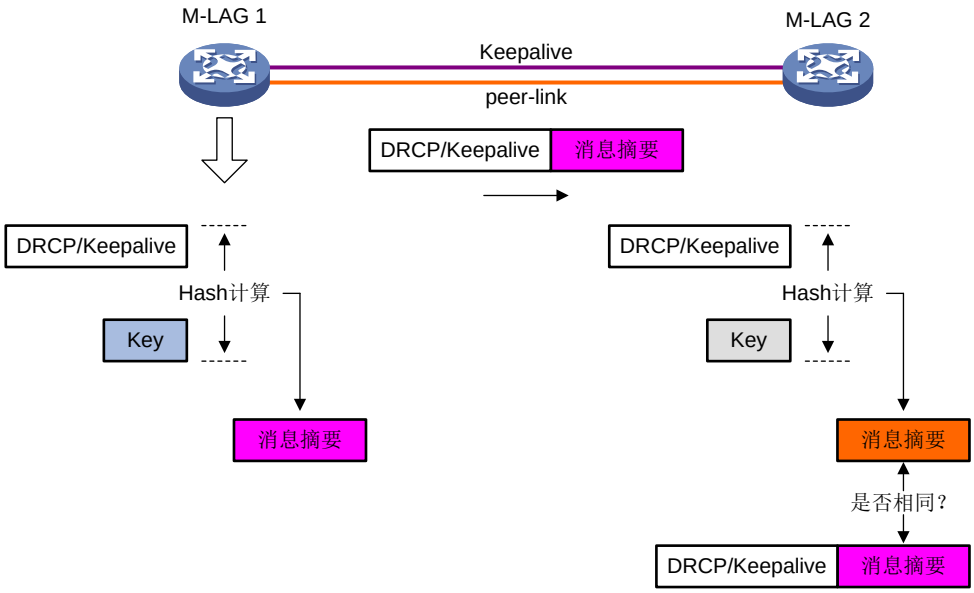


2.3.11 M-LAG 报文认证

为防止攻击者篡改 DRCP 报文/Keepalive 报文内容，M-LAG 提供报文认证功能，提高安全性。

如图 10 所示，M-LAG 设备发送的 DRCP 报文/Keepalive 报文中会携带一个消息摘要，该消息摘要是对报文内容经 Hash 计算得到。对端 M-LAG 设备收到该报文时，会与自己计算的该报文的消息摘要进行比对，如果一致，则认为其合法。

图10 M-LAG 报文认证



2.4 流量转发

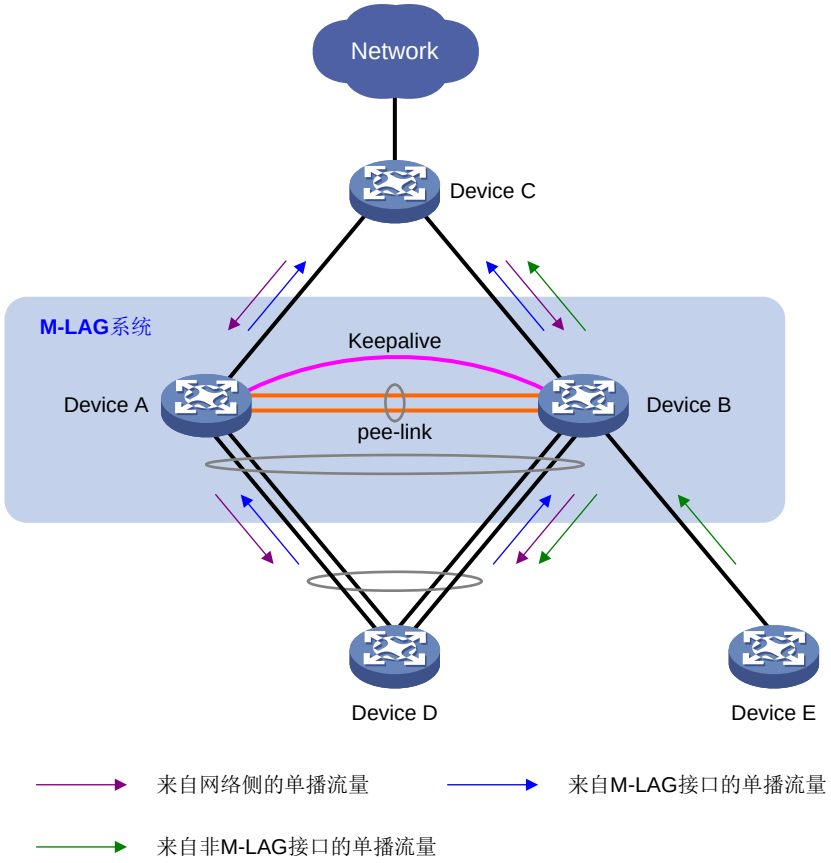
M-LAG 系统建立成功后即进入正常的工作，M-LAG 主备设备负载分担共同进行流量的转发，转发行为没有区别。

2.4.1 单播流量转发

如图 11 所示，Device D 设备接入 M-LAG 系统，已知单播流量转发机制如下：

- 对于南北向的单播流量，在 M-LAG 接入侧，M-LAG 设备接收到接入设备通过聚合链路负载分担发送的流量后，按本地转发优先原则，共同进行流量转发。发往 Network 侧的流量到达 M-LAG 设备后将根据路由表转发流量。
- 对于东西向的单播流量，二层流量通过 M-LAG 本地优先转发，三层流量通过双活网关转发，都不经过 peer-link 链路，直接由 M-LAG 设备转发。

图11 单播流量转发示意图

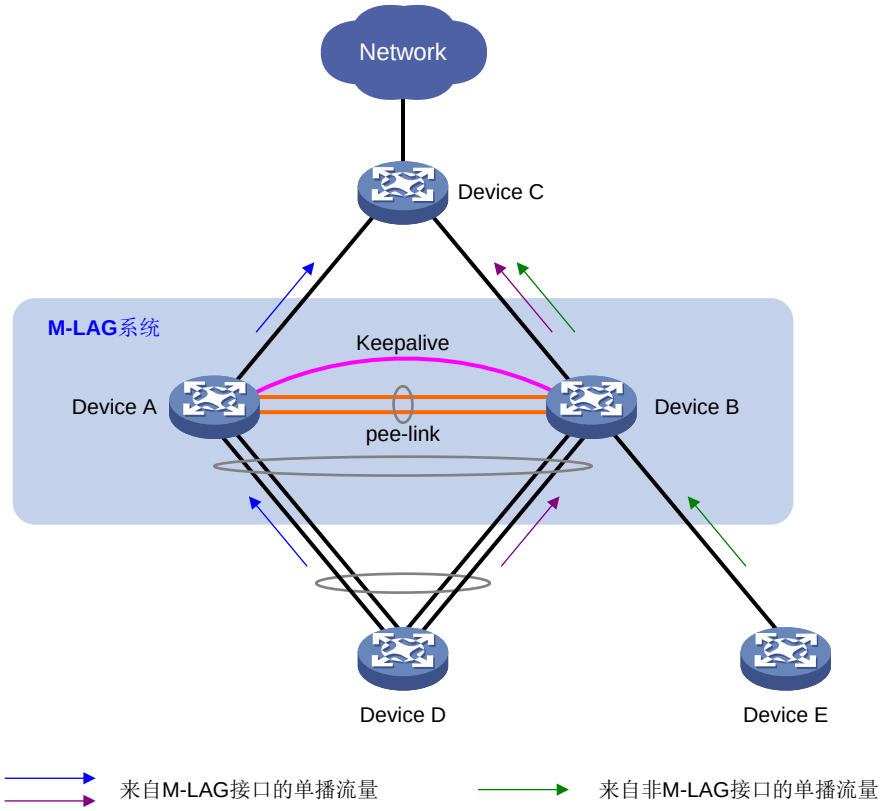


1. 南北向流量转发

如图 12 所示，上行流量转发机制如下：

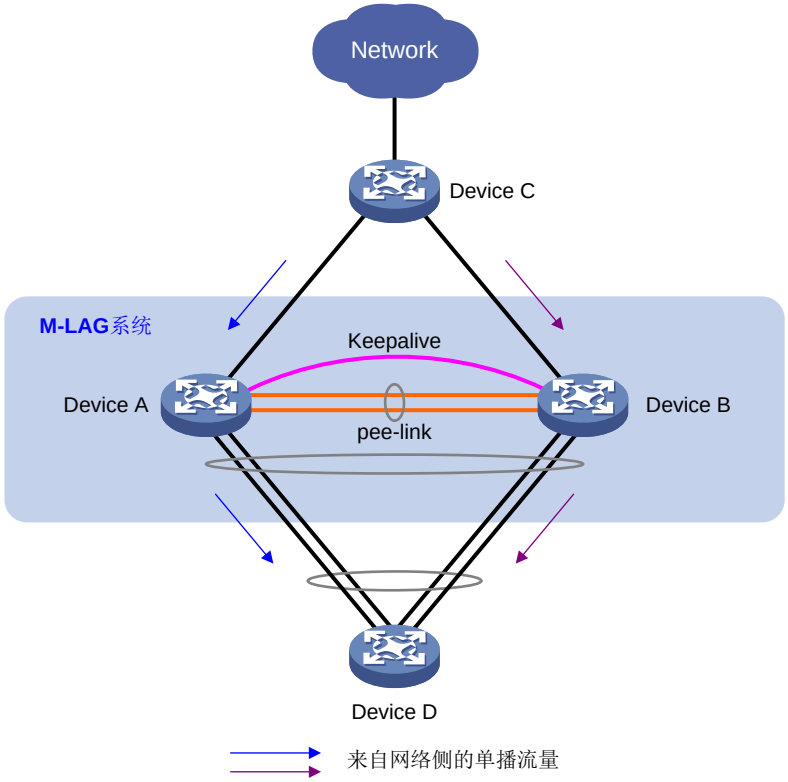
- 对于来自 M-LAG 接口的单播流量，Device A 与 Device B 形成负载分担，共同对来自 M-LAG 接口端口的单播流量进行转发。单播流量按本地转发优先原则，避免对 peer-link 链路造成压力。
- 对于来自非 M-LAG 接口的单播流量，Device B 按本地转发优先原则直接转发，不向 peer-link 链路转发。

图12 上行流量转发示意图



如图 13所示，对于来自网络侧的单播流量，根据本地转发优先原则，直接本地转发。

图13 下行流量转发示意图

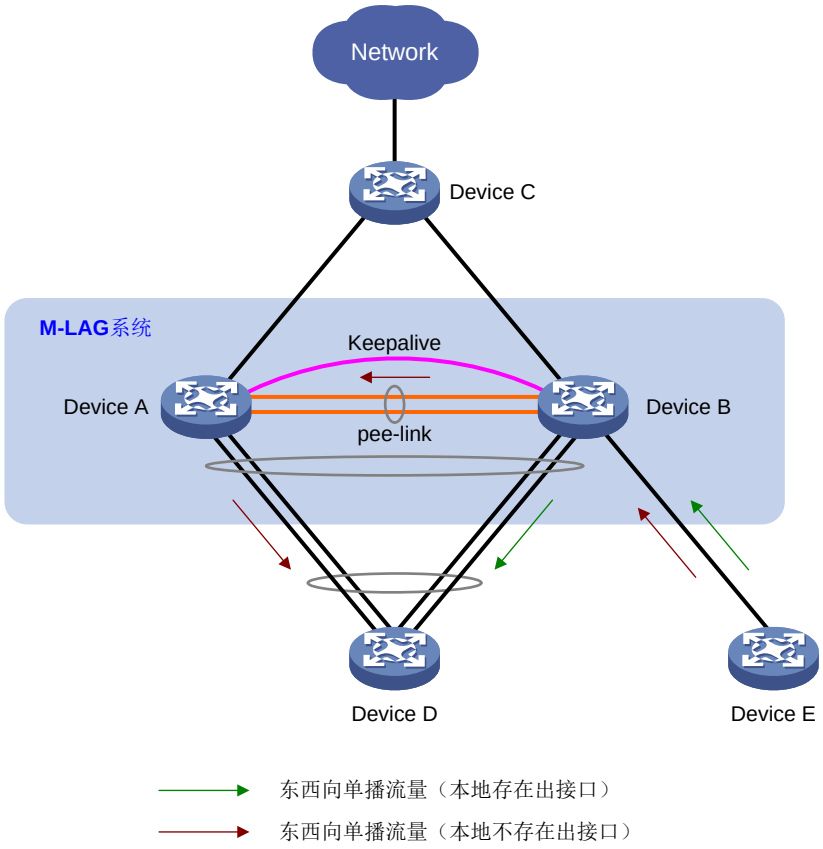


2. 东西向流量转发

如图 14所示，东西向流量转发机制如下：

- 如果本地存在出接口，则按照本地转发优先原则，本地直接转发。
- 如果本地不存在出接口，则流量绕行 peer-link 链路，通过对端 M-LAG 设备转发。

图14 东西向流量转发示意图

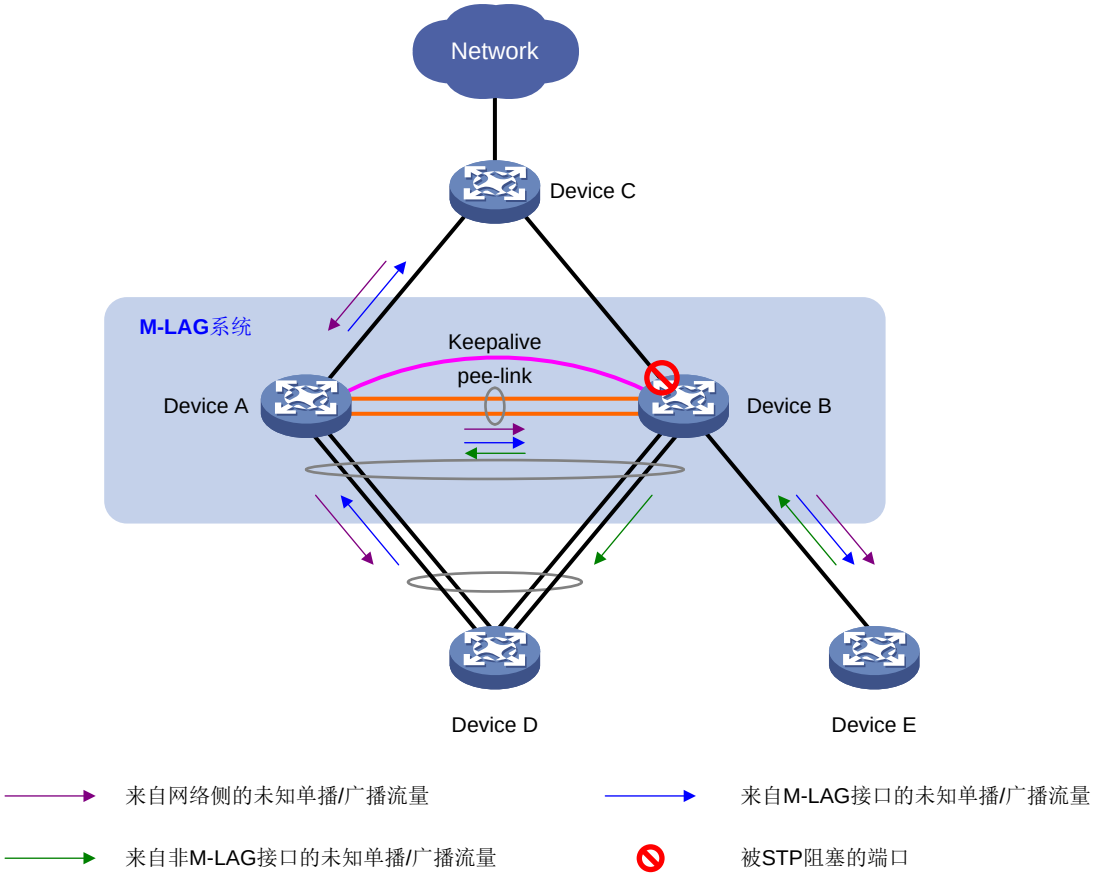


2.4.2 未知单播/广播流量转发

如图 15 所示，Device D 设备接入 M-LAG 系统，未知单播/广播流量转发机制如下：

- Device B 收到来自非 M-LAG 接口的未知单播/广播流量后，将向相连设备的同一广播域转发。当流量到达 Device A 时，由于 peer-link 接口与 M-LAG 接口存在单向隔离机制，到达 Device A 的流量不会向 Device D 转发。
- Device D 发送的未知单播/广播流量到达 M-LAG 设备（以 Device A 为例）后，将向相连设备的同一广播域转发。当流量到达 Device B 时，由于 peer-link 接口与 M-LAG 接口存在单向隔离机制，到达 Device B 的流量不会向 Device D 转发。

图15 广播流量转发示意图



2.4.3 组播流量转发

M-LAG 组网环境中组播流量的转发过程，具体请参见“8 二层组播支持 M-LAG”和“9 三层组播支持 M-LAG”。

2.5 M-LAG故障处理机制

2.5.1 M-LAG 接口故障处理机制

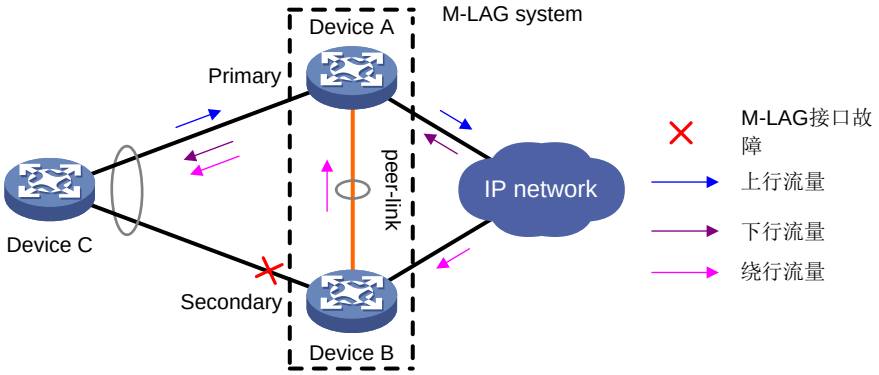
如图 16 所示，某 M-LAG 接口故障，来自外网侧的流量会通过 peer-link 链路发送给另外一台设备，所有流量均由另外一台 M-LAG 设备转发，具体过程如下：

(1) Device B 的某 M-LAG 接口故障，外网侧不感知，流量依然会发送给所有 M-LAG 设备。

(2) Device A 的相同 M-LAG 接口正常，则 Device B 收到外网侧访问 Device C 的流量后，通过 peer-link 链路将流量交给 Device A 后转发给 Device C。

(3) 故障恢复后，Device B 的该 M-LAG 接口 up，流量正常转发。

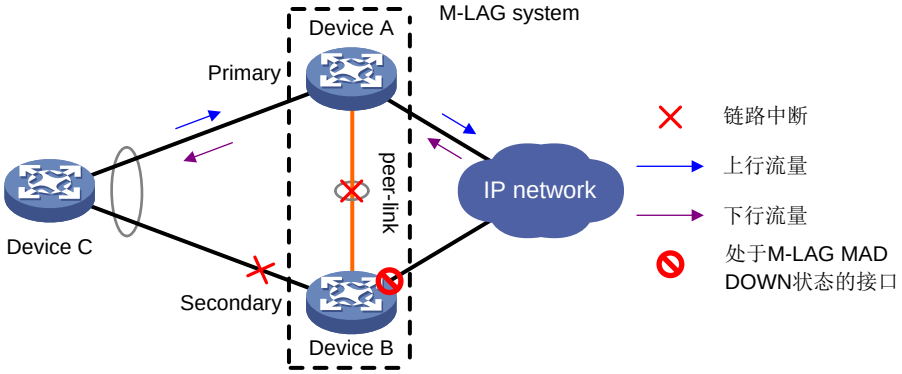
图16 M-LAG 接口故障处理机制示意图



2.5.2 peer-link 链路故障处理机制

如图 17 所示，peer-link 链路故障但 Keepalive 链路正常会导致备设备上除 M-LAG 保留接口以外的接口处于 M-LAG MAD DOWN 状态。主设备上的 M-LAG 接口所在聚合链路状态仍为 up，备设备上的 M-LAG 接口所在聚合链路状态变为 down，从而保证所有流量都通过主设备转发。一旦 peer-link 链路故障恢复，处于 M-LAG MAD DOWN 状态的接口经过延迟恢复时间自动恢复为 up 状态。

图17 peer-link 链路故障处理机制示意图

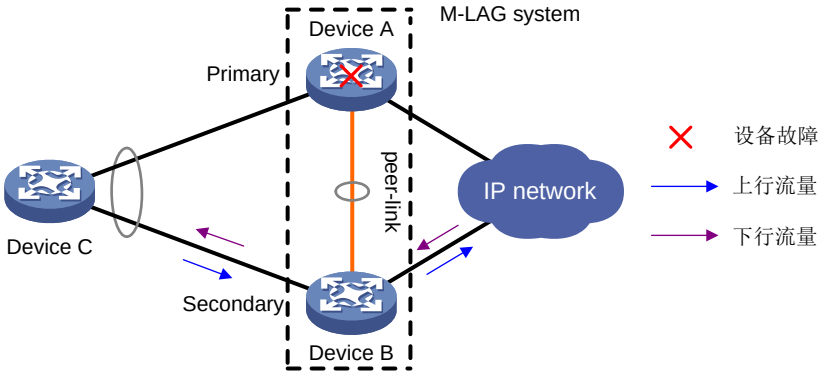


2.5.3 设备故障处理机制

如图 18 所示，Device A 为主设备，Device B 为备设备。当主设备故障后，主设备上的聚合链路状态变为 down，不再转发流量。备设备将升级为主设备，该设备上的聚合链路状态为 up，流量转发状态不变，继续转发流量。主设备故障恢复后，M-LAG 系统中由从状态升级为主状态的设备仍保持主状态，故障恢复后的设备成为 M-LAG 系统的备设备。

如果是备设备发生故障，M-LAG 系统的主备状态不会发生变化，备设备上的聚合链路状态变为 down。主设备上的聚合链路状态为 up，流量转发状态不变，继续转发流量。

图18 设备故障处理机制示意图

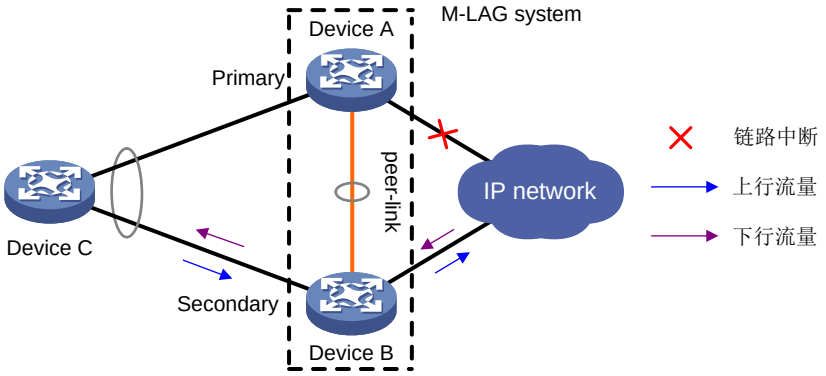


2.5.4 上行链路故障处理机制

上行链路故障并不会影响 M-LAG 系统的转发。如图 19 所示，Device A 上行链路虽然故障，但是外网侧的转发相关表项由 Device B 通过 peer-link 链路同步给 Device A，Device A 会将访问外网侧的流量发送给 Device B 进行转发。而外网侧发送给 Device C 的流量由于接口故障，自然也不会发送给 Device A 处理。

上行链路故障时，如果通过 Device A 将访问外网侧的流量发送给 Device B 进行转发，会降低转发效率。此时用户可以配置 Monitor Link 功能，将 M-LAG 组成员端口和上行端口关联起来，一旦上行链路故障了，会联动 M-LAG 组成员端口状态，将其状态变为 down，提高转发效率。

图19 上行链路故障处理机制示意图



2.5.5 M-LAG 二次故障处理机制

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。

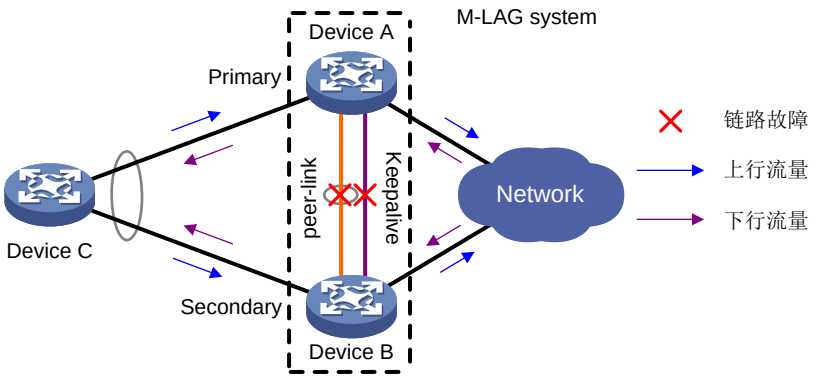
1. 缺省配置场景

如图 20 所示，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 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，可能导致流量转发错误。

图20 缺省配置场景下二次故障处理机制示意图



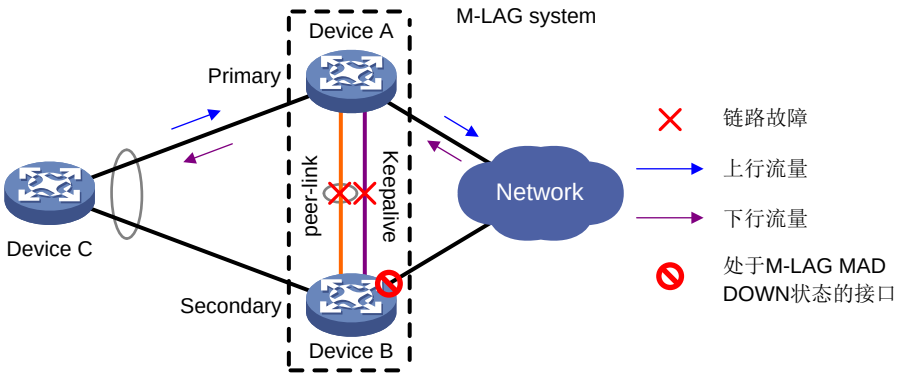
2. 开启 M-LAG MAD DOWN 状态保持功能场景

如图 21 所示，若peer-link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，但由于M-LAG设备已开启M-LAG MAD DOWN状态保持功能，将不会解除设备上所有接口的M-LAG MAD DOWN状态，继续只从原来的主设备转发流量。这样将不会出现双主双活的情况，避免流量转发异常。

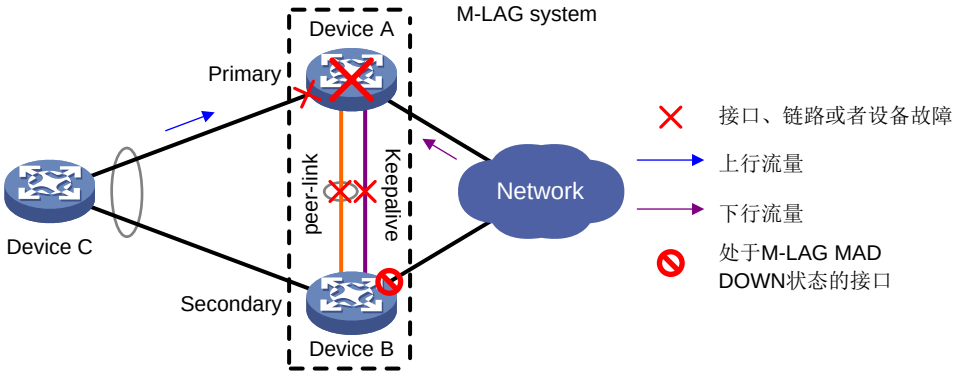
若Keepalive链路先发生故障，peer-link链路后发生故障，则M-LAG设备上的接口不会被置为M-LAG MAD DOWN状态，而是直接以双主双活的方式转发流量。M-LAG MAD DOWN状态保持功能不能解决Keepalive链路先故障，peer-link后故障导致的双主双活问题。

图21 开启 M-LAG MAD DOWN 状态保持功能场景下二次故障处理机制示意图（一）



如图 22 所示，如果主设备故障或者主设备上M-LAG接口故障，则无法转发流量。为了避免这种情况可以关闭M-LAG MAD DOWN状态保持功能，解除从设备上所有接口的M-LAG MAD DOWN状态，使从设备升级为主设备，以保证流量正常转发，减少流量中断时间。

图22 开启 M-LAG MAD DOWN 状态保持功能场景下二次故障处理机制示意图（二）



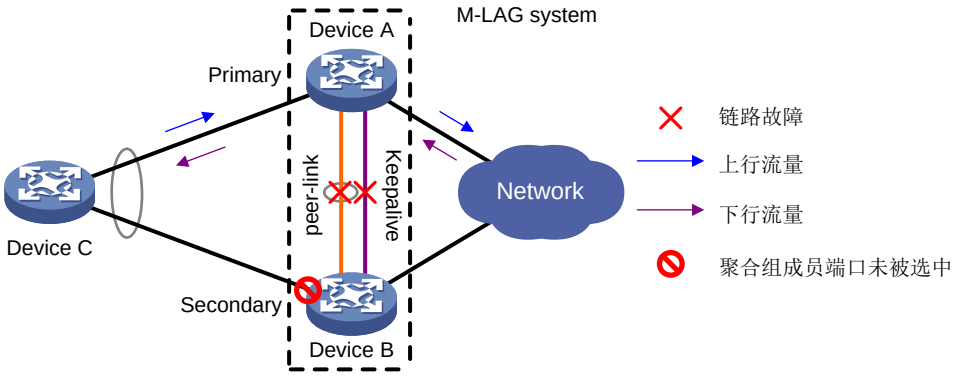
3. 开启设备独立工作功能场景

如图 23 所示，若 peer-link 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 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，解除所有接口的 M-LAG MAD DOWN 状态。但由于已开启立即或延迟切换到设备独立工作状态功能，两台 M-LAG 设备将切换到独立工作状态，切换后 M-LAG 接口对应的聚合接口发送的 LACP 报文中携带的 M-LAG 系统参数还原为聚合接口的 LACP 系统 MAC 地址和 LACP 系统优先级，使同一 M-LAG 组中的两个聚合接口的 LACP 系统 MAC 地址和 LACP 系统优先级不一致。这样 M-LAG 设备中只有一台设备的聚合接口的成员端口可以被选中（接入设备上仅一个成员端口可以被选中），通过被选中的设备转发业务流量，避免流量转发异常。成员端口的选中与 LACP 系统优先级和系统 MAC 地址相关，与 M-LAG 设备角色无关。LACP 系统优先级和系统 MAC 地址越小，则优先被选中。若选中的成员端口也发生故障，则将选中另外一台设备上聚合接口的成员端口，通过该聚合接口继续转发流量。

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，将立即或延迟一段时间切换到设备独立工作模式。

图23 开启设备独立工作功能场景下二次故障处理机制示意图



3 M-LAG 网络中运行 STP

3.1 STP应用场景

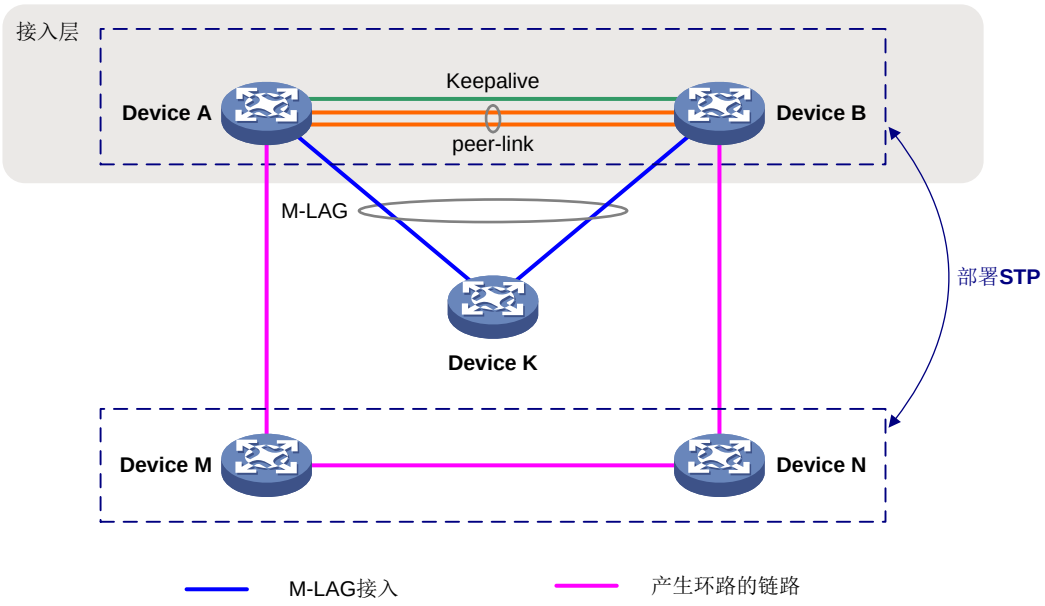
M-LAG 本身具有环路避免机制，正常情况下，M-LAG 组网中不会产生环路。多级 M-LAG 组网中，网络搭建错误、初始化 M-LAG 配置或设备空配置重启时，网络中可能会产生环路，需要部署 STP 来避免环路。需要部署 STP 的典型场景包括：

- 下行接入设备产生环路。
- 多级 M-LAG 设备间接入设备产生环路。
- 初始化 M-LAG 配置产生环路。
- 设备空配置重启产生环路。

3.1.1 下行接入设备产生环路

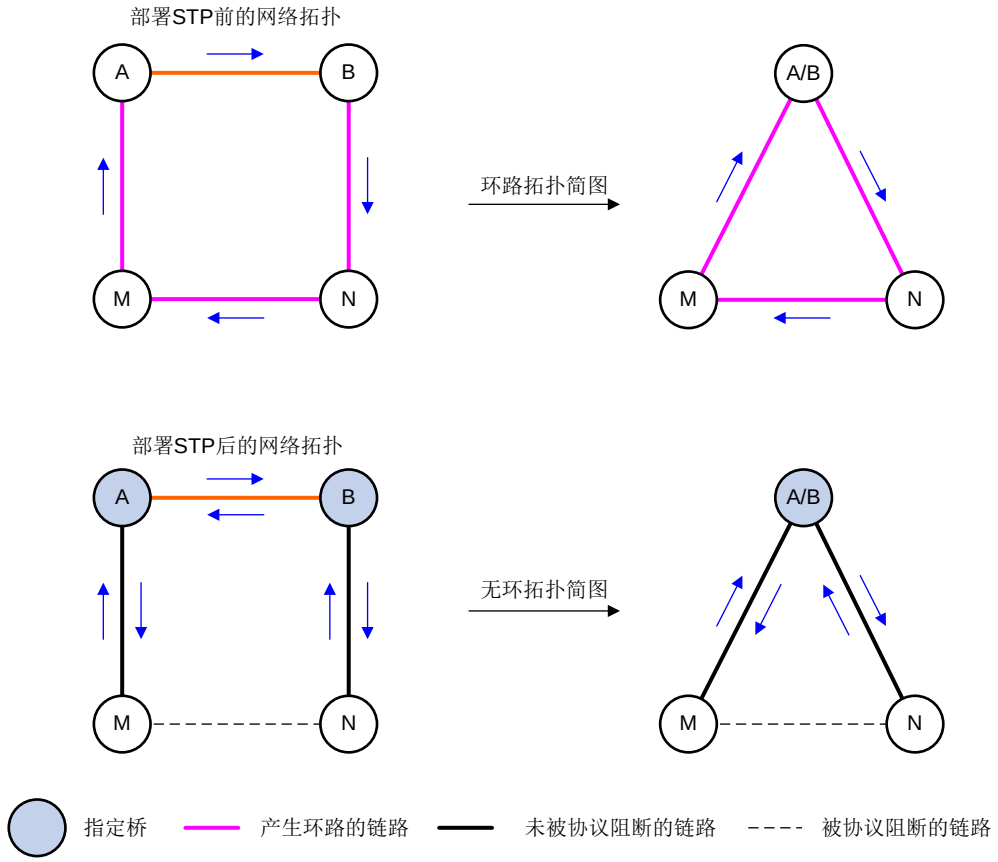
如图 24 所示，下行设备 Device M 和 Device N 通过非 M-LAG 接口接入 Device A 和 Device B，且 Device M 和 Device N 互连。

图24 下行设备接入 M-LAG 系统产生环路示意图



如图 25 所示，Device M 和 Device N 间的流量会经过 Device A 和 Device B 间的 peer-link 链路绕行，形成环路。为了避免环路，需要在 Device A 和 Device B、Device M 和 Device N 上部署 STP，阻塞 Device M 和 Device N 之间链路。其中，Device A 和 Device B 作为指定桥，Device M 和 Device N 间流量将通过 Device A 和 Device B 转发，避免了环路。

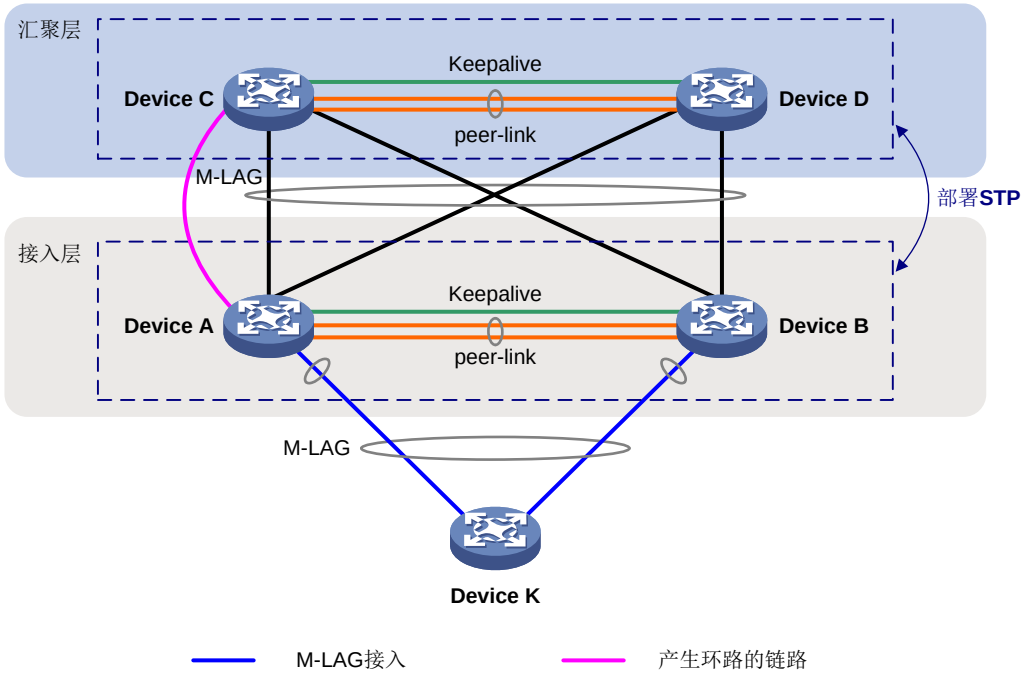
图25 下行设备接入 M-LAG 系统产生环路拓扑图



3.1.2 多级 M-LAG 设备间产生的环路

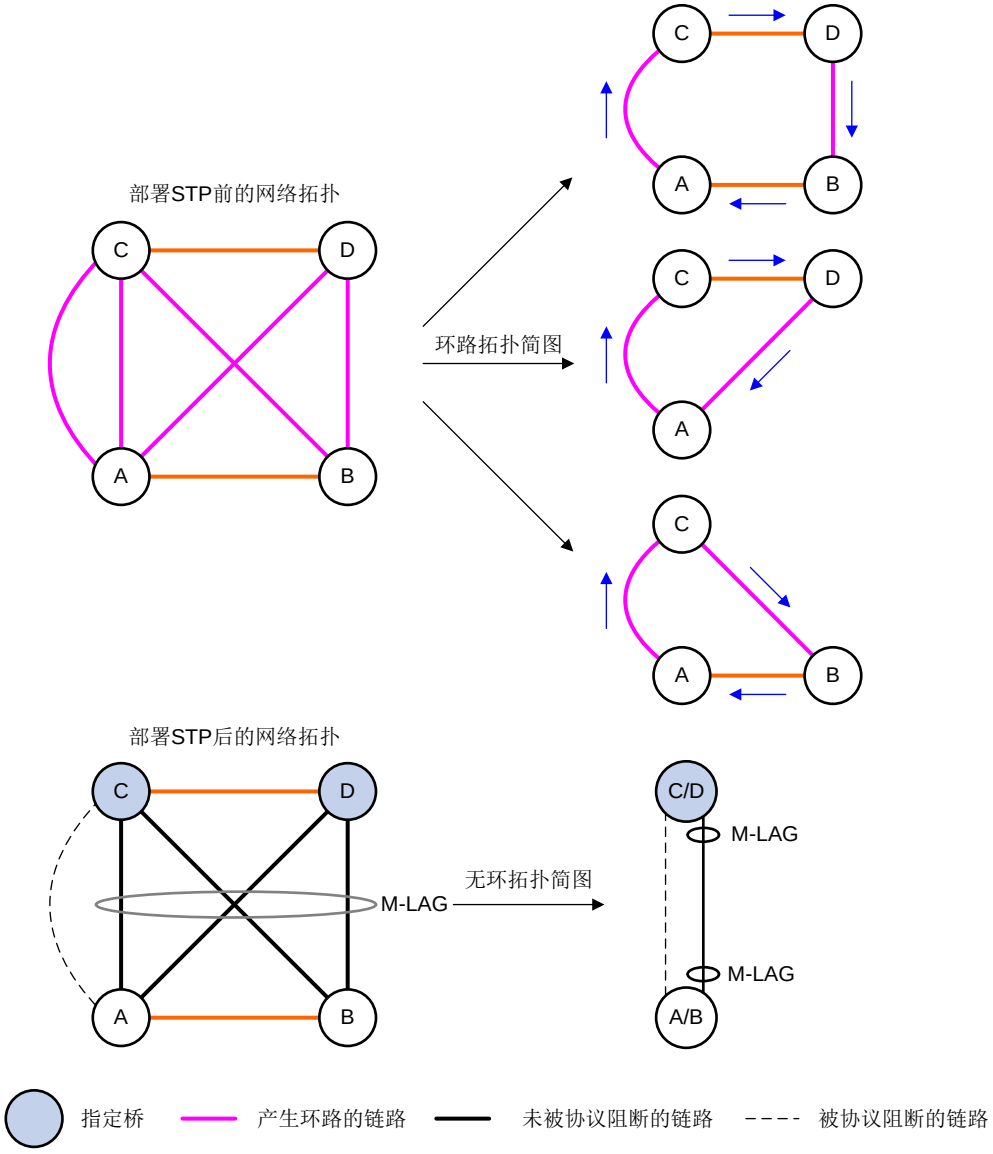
如图 26 所示，多级 M-LAG 组网中，Device A 和 Device C 之间通过非 M-LAG 接口误接线。

图26 多级 M-LAG 设备间环路产生示意图



如图 27 所示，Device A 和 Device C 间的流量会经过 peer-link 链路绕行，形成环路。为了避免环路，需要在 Device A 和 Device B、Device C 和 Device D 上部署 STP，阻塞 Device A 和 Device C 之间的误连接。其中，Device C 和 Device D 作为指定桥，Device A 和 Device C 间流量将通过多级 M-LAG 中的 M-LAG 接口转发，避免了环路。

图27 多级 M-LAG 设备间环路产生拓扑图

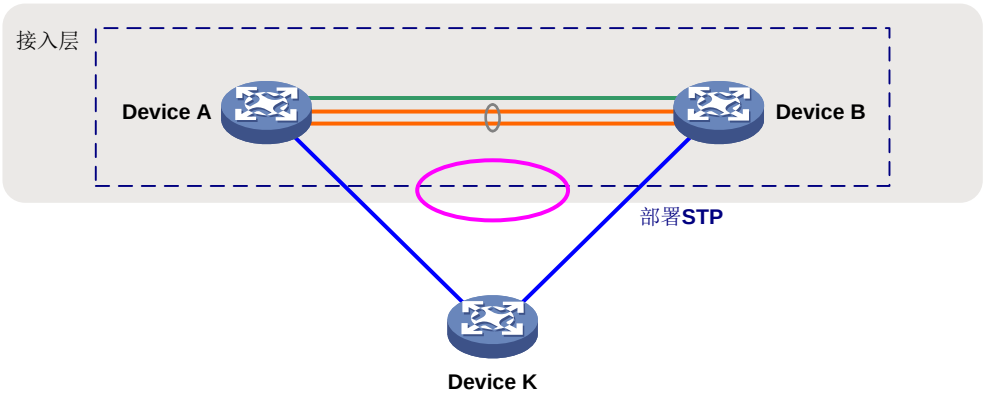


3.1.3 初始化 M-LAG 配置产生环路

如图 28 所示，按照 M-LAG 组网要求完成设备间的线路连接，并在 M-LAG 设备上执行 M-LAG 相关配置后，在 M-LAG 系统建立前，网络中存在短暂的环路。此时，通过部署 STP，可以阻塞端口，避免流量转发环路。

建议 M-LAG 系统建立完成，并验证通过后，再将 M-LAG 系统接入到现网中，以避免初始化 M-LAG 配置产生环路。

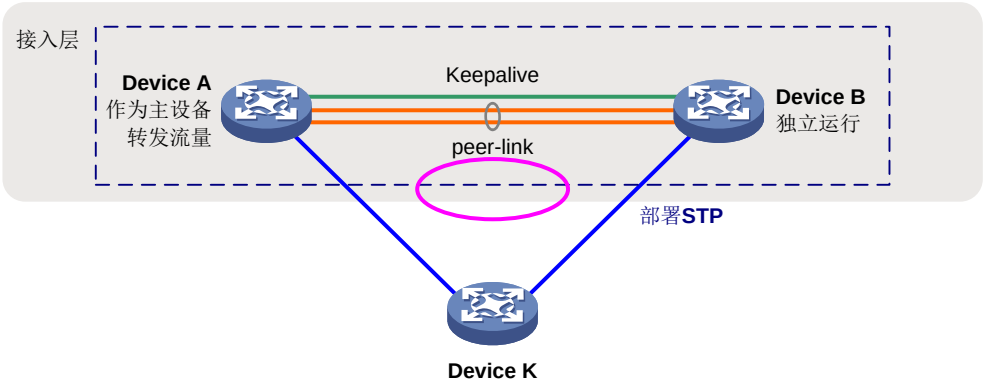
图28 初始化 M-LAG 配置环路产生示意图



3.1.4 设备空配置重启产生环路

如图 29 所示，两台 M-LAG 设备组成 M-LAG 系统后，如果其中一台 M-LAG 设备进行空配置重启，则该设备重启后不会加入 M-LAG 系统，作为独立的物理设备运行，可以转发流量。另一台 M-LAG 设备认为对端 M-LAG 设备故障，承担流量转发工作。从而，导致网络中存在环路。通过部署 STP，可以用来避免环路。

图29 设备空配置重启环路产生示意图



3.2 STP在M-LAG中的工作机制

在 M-LAG 组网中，由于组成 M-LAG 系统的两台 M-LAG 设备虚拟为一台设备，为了确保 STP 在 M-LAG 组网中的正常运行，M-LAG 设备上的 STP 运行机制需要进行如下调整：

- STP 协议由主设备控制。无论指定端口位于哪台 M-LAG 设备，都是由主设备生成 STP 的 BPDU 报文，并在指定的端口上发送 BPDU 报文。端口的 STP 状态也由主设备决定。
- 备设备不生成 BPDU 报文，也无法决定端口的 STP 状态。备设备接收到 BPDU 报文后，通过 peer-link 链路将其转发给主设备。
- 两台 M-LAG 设备上 M-LAG 接口的 STP 端口状态始终保持一致。
- peer-link 链路上不运行 STP 协议。

说明

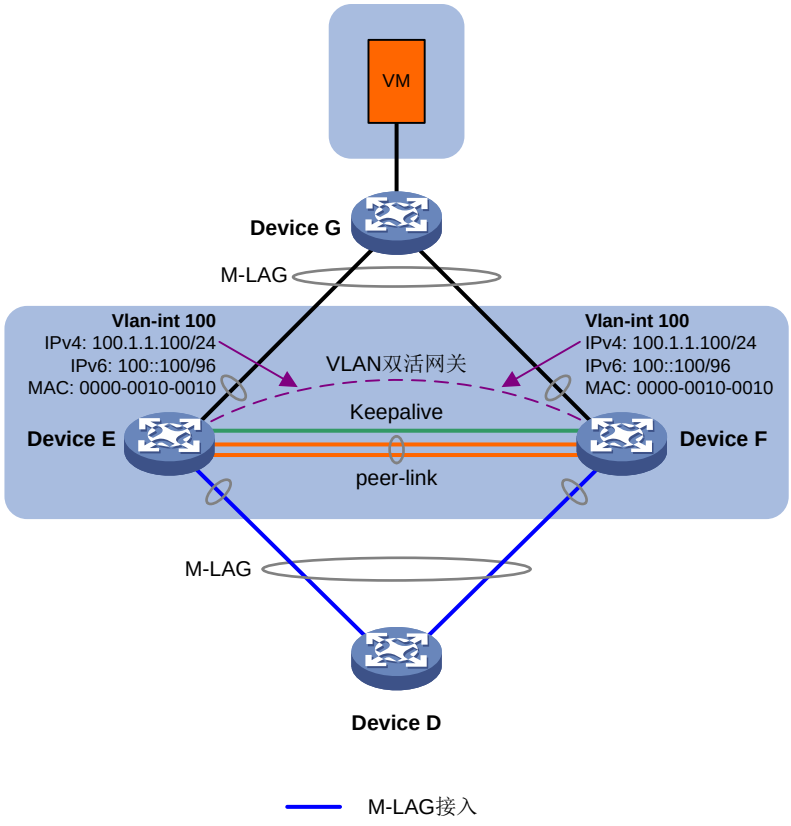
组成 M-LAG 系统的两台 M-LAG 设备具有相同的虚 MAC 地址（M-LAG 系统的 MAC 地址）。M-LAG 设备基于该虚 MAC 地址运行 STP 协议，因此，两台 M-LAG 设备可以同时作为 STP 的根。

4 M-LAG 三层网关

4.1 VLAN双活网关

如图 30 所示，VLAN 双活网关是指组成 M-LAG 系统的两台 M-LAG 设备均作为用户侧的网关，回应用户侧的 ARP/ND 请求并转发用户侧的报文，以提高网关的可靠性。

图30 M-LAG VLAN 双活网关部署方案



M-LAG VLAN 双活网关的部署方案为：在同一 M-LAG 系统的两台 M-LAG 设备上各创建一个相同编号的 VLAN 接口（例如 Vlan-interface100），并为其配置相同的 IPv4 地址、IPv6 地址和 MAC 地址。该接口的 IPv4 地址和 IPv6 地址作为网关地址，以便 IPv4 和 IPv6 用户均可通过该网关访问外部网络。

M-LAG VLAN 双活网关的工作机制为：

- M-LAG 设备采取本地优先转发原则，设备收到报文后直接转发，无需绕行 peer-link 链路到对端 M-LAG 设备转发。例如，M-LAG 设备 Device E 收到 VM 侧发送的 ARP 请求，Device E 直接向 VM 侧发送 ARP 应答报文，无需转发到 M-LAG 设备 Device F 处理。
- 当一条接入链路发生故障时，流量可以快速切换到另一条链路，保证可靠性。例如，Device E 和 Device G 之间链路故障，则流量处理方式为：
  - 访问 Device D 的下行流量快速切换到 Device F 处理，不再转发到 Device E。
  - 访问 VM 的上行流量，转发到 Device F 时，Device F 处理完成后直接向 VM 侧转发；转发到 Device E 时，流量将通过 peer-link 链路绕行到 Device F 处理，然后向 VM 侧转发。
- 两条接入链路可以同时处理用户流量，以提高带宽利用率，使流量在两条接入链路上负载分担。

在 M-LAG VLAN 双活网关场景中，M-LAG 成员设备作为网关进行三层转发。由于作为网关的 VLAN 接口具有相同的 IP 地址和 MAC 地址，M-LAG 成员设备无法用该 IP 地址与用户侧设备之间建立路由邻居关系。当 VLAN 双活网关需要与 Device B 建立路由邻居关系时，可以在作为网关的 VLAN 接口上配置 M-LAG 虚拟 IP 地址，并部署路由协议，使用虚拟 IP 地址与下行设备 Device B 建立邻居关系。具体部署方式请参见图 31 和表 3。

图31 M-LAG VLAN 双活网关场景网关接口配置 M-LAG 虚拟 IP 地址建立路由邻居

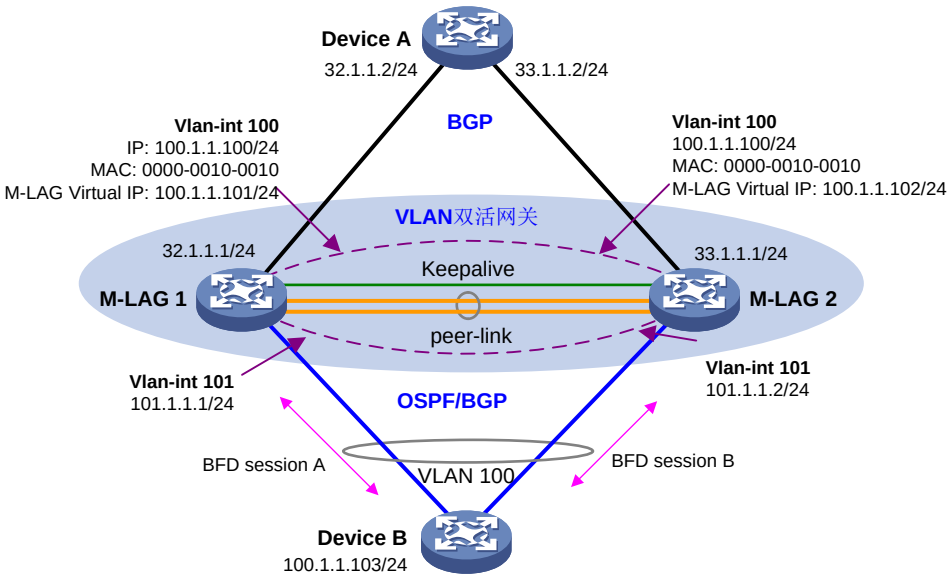


表3 M-LAG VLAN 双活网关场景网关接口配置 M-LAG 虚拟 IP 地址建立路由邻居

应用场景	部署方案	流量模型
下行设备Device B与M-LAG设备部署动态路由	• 在同一 M-LAG 系统的两台 M-LAG 设备上各创建一个相同编号的 VLAN 接口（例如 VLAN 100）作为 IPv4 和 IPv6 双活网关，在两台 M-LAG 设备上为该 VLAN 接口配置相同的 IP 地址和 MAC 地址作为网关地址。Device B 通过 M-LAG 接口接入到 M-LAG 设备，且 IPv4 和 IPv6 流量均可通过网关地址访问外部网络 • 在同一 M-LAG 系统的两台 M-LAG 设备上，作为网关的 VLAN 接口下分别配置同一网段不同的 M-LAG 虚拟 IP 地址，使用该虚拟 IP 地址与 Device B 建立三层连接，通过 BGP 或 OSPF 实现三层互通	• Device B 发出的二层流量，查找 MAC 地址表找到出接口为聚合接口，将流量负载分担到 M-LAG 设备上。M-LAG 设备根据本地优先转发原则，根据 MAC 地址表进行二层转发 • Device B 发出的三层流量，根据配置的动态路由生成的路由表找到

应用场景	部署方案	流量模型
	• 在同一 M-LAG 系统的两台 M-LAG 设备上各自再创建一个相同编号的 VLAN 接口（例如 VLAN 101），将 peer-link 链路聚合接口加入该 VLAN。两台 M-LAG 设备上分别为该 VLAN 接口配置同一网段的不同 IP 地址，以实现两台 M-LAG 设备的三层互通。如果 M-LAG 1 或 M-LAG 2 与上行设备 Device A 的链路故障，报文可以通过路由绕行到对端 M-LAG 设备处理 • M-LAG 设备与上行设备 Device A 间通过三层接口部署等价路由进行负载分担	出接口为 VLAN 100，通过 VLAN 100 加入的聚合接口转发，将流量负载分担到 M-LAG 设备上，M-LAG 设备根据 FIB 表对流量进行三层转发 • 外部网络访问 Device B 的流量根据 ECMP 路由，将流量负载分担转到 M-LAG 设备。M-LAG 设备根据本地路由信息将流量转发到 Device B
BFD快速检测（如有需要）	两台M-LAG设备分别使用M-LAG虚拟IP地址与下行设备的VLAN接口100的从IP地址建立BFD会话	-

4.2 VRRP网关

在 M-LAG 设备上部署 VRRP，可以实现为下行接入设备提供冗余备份的网关。M-LAG+VRRP 的三层转发部署方案请参见图 32 和表 4。

图32 M-LAG+VRRP 的三层转发方案

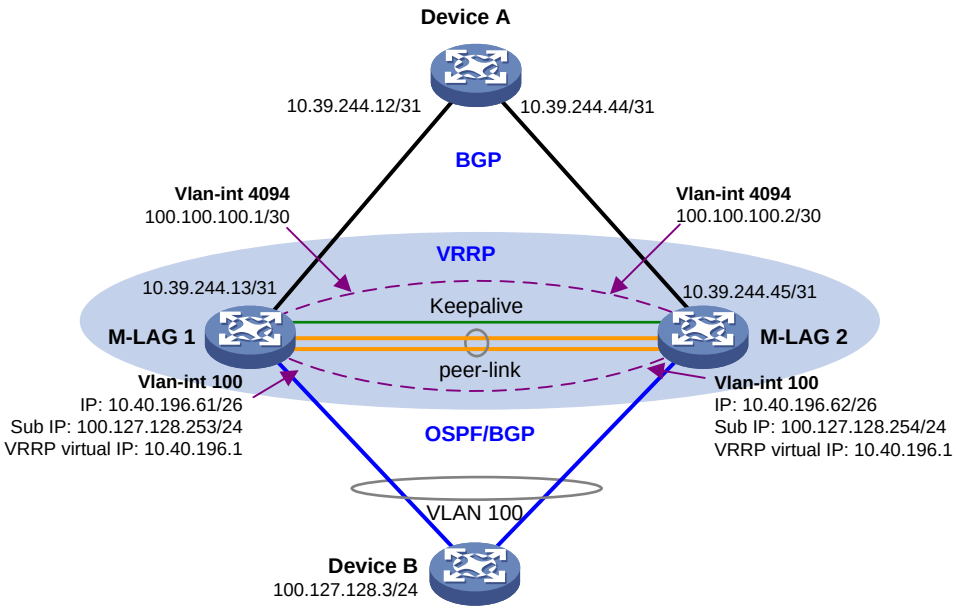


表4 M-LAG+VRRP 的三层转发方案说明

部署方案	流量模型
• M-LAG 设备部署 VRRP，VRRP 虚拟 IP 地址作为 Device B 的网关地址，Device B 通过 M-LAG 接口双归接入到 VRRP	• Device B 发往其它网段的报文，通过 M-LAG 接口负载分担到两台 M-LAG 设备，两台 M-LAG 设备均可以作为 VRRP

部署方案	流量模型
网关 • M-LAG 口所属 VLAN 创建 VLAN 接口，两台 M-LAG 设备的 VLAN 接口分别配置同网段内不同的 IP 地址作主 IP 地址，再配置另一网段内同网段不同的 IP 地址作从 IP 地址 • 两台 M-LAG 设备通过 peer-link 链路建立的三层接口建立路由邻居作为三层链路备份，如果 M-LAG1 或 M-LAG2 与上行设备 Device A 的链路故障，报文可以通过路由绕行到对端 M-LAG 设备处理 • M-LAG 设备与上行设备 Device A 间通过三层接口部署等价路由进行负载分担	虚拟路由器对报文进行转发 • Device B 发出的三层流量，根据 Device B 与 M-LAG 设备 VLAN 接口从 IP 建立的路由信息转发 • 外部网络访问 Device B 的流量根据 ECMP 路由，将流量负载分担转发到 M-LAG 设备上。M-LAG 设备根据本地路由信息将流量转发到 Device B

VRRP 网关接收到外网发送给 Device B 的、目的 MAC 地址为对端 M-LAG 设备实 MAC 地址的报文后，如果按照正常转发流程，本地 M-LAG 设备通过 peer-link 链路将该报文发送给目的 M-LAG 设备，则由于 M-LAG 的防环机制，目的 M-LAG 设备不会将从 peer-link 链路上接收的报文通过 M-LAG 接口转发给 Device B，从而导致报文被丢弃。为了避免报文丢失，在 M-LAG+VRRP 组网中，M-LAG 设备间需要同步各自的实 MAC 地址，使得 M-LAG 设备可以对目的 MAC 地址为对端 M-LAG 设备实 MAC 地址的报文进行本地三层转发。实 MAC 地址同步功能始终处于开启状态，无需手工配置。

5 M-LAG 网络中运行环路检测

5.1 功能简介

在 M-LAG 网络中运行环路检测时，如果设备在 M-LAG 接口上检测到环路，则形成 M-LAG 的两台设备作为一台虚拟设备，在同一编号的 M-LAG 接口上对环路作出相同的响应；如果设备在单归接入设备的接口上检测到环路，则 M-LAG 的两台设备各自作为独立的一台设备对环路作出单独的响应。

5.2 工作机制

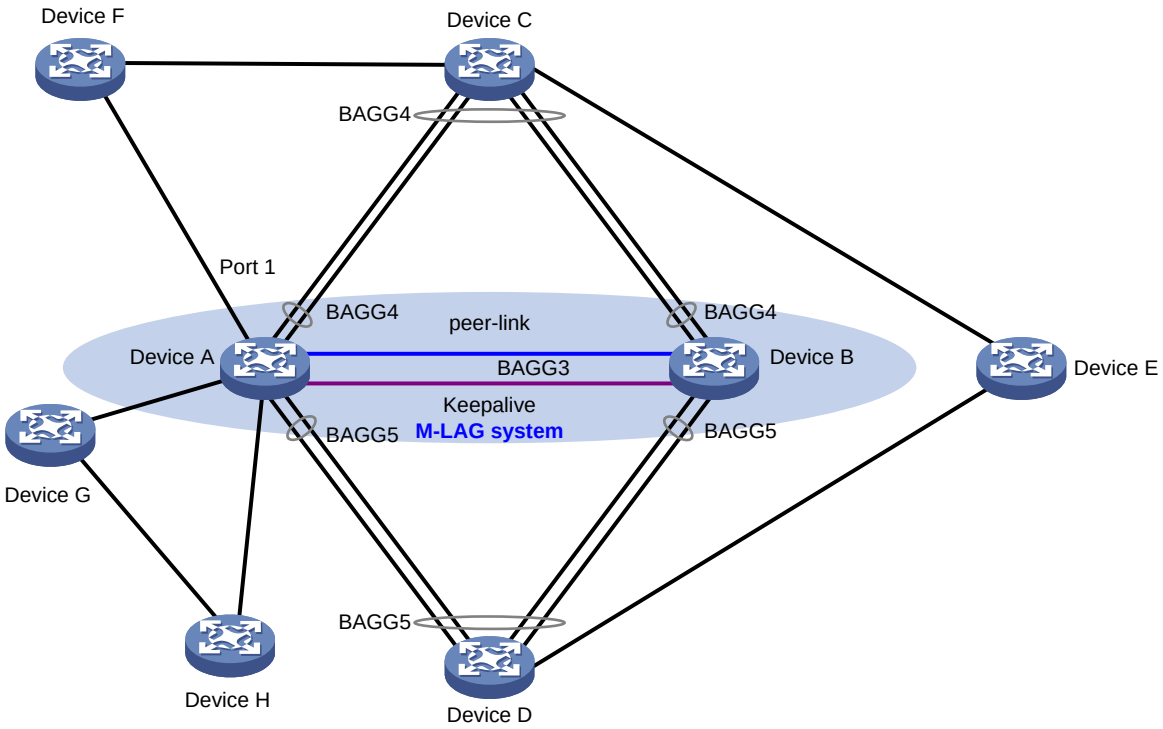
5.2.1 普通组网下的 M-LAG 设备环路检测工作机制

1. 功能简介

如图 33 所示，Device A 和 Device B 组成 M-LAG 系统，作为一台虚拟设备接入在网络中，以提高网络的可靠性。Device C 和 Device D 双归接入 M-LAG 系统，Device F、Device G 和 Device H 单归接入 Device A，本章将通过以下几种情形，分别对环路检测的工作机制进行介绍：

- 双归接入情形：例如 M-LAG 系统、Device C、Device F 以及 Device D 形成的环路。
- 单归接入情形一：例如 Device A、Device B、Device C 以及 Device F 形成的环路。
- 单归接入情形二：例如 Device A、Device G 以及 Device H 形成的环路。

图33 普通组网运行环路检测组网图



2. 环路检测功能的生效机制

M-LAG 设备只有在开启了环路检测功能后，才能检测到环路以及对发生环路的端口进行处理。所以根据不同的情形，推荐的环路检测配置如下：

- 对于双归接入情形，用户需要在 M-LAG 的两台设备上均开启环路检测功能，否则将只有一台设备对端口环路进行处理，无法起到消除环路的作用。
- 对于单归接入情形一，如果用户仅在 Device A 上开启环路检测功能，建议在 BAGG4 接口上开启环路检测功能，或全局开启环路检测功能，这样 Device A 会在 Port 1 上检测到环路，能够有效地消除环路。如果 Device A 仅在 Port 1 上开启环路检测功能，则 Device A 将在 BAGG4 接口上检测到环路，而仅阻塞或者关闭 Device A 的 BAGG4 接口不足以消除环路，此时建议在 Device B 上也开启环路检测功能，以确保能够消除环路。
- 对于单归接入情形二，用户只需要在单归接入的 Device A 上开启环路检测功能，Device A 将作为一台独立的设备检测环路。

3. 环路检测报文发送机制

对于双归接入的情形：

- 开启环路检测功能后，M-LAG 的两台设备都会在 M-LAG 接口上发送环路检测报文，报文的源 MAC 地址相同，均为 M-LAG 系统 MAC 地址。
- M-LAG 设备从 M-LAG 接口收到环路检测报文后，会通过 peer-link 链路同步给另一台 M-LAG 设备，以避免单点故障导致另一台 M-LAG 设备无法收到环路检测报文。

对于单归接入一的情形：

- 开启环路检测功能后，M-LAG 设备即在开启了环路检测功能的接口上发送环路检测报文，报文的源 MAC 地址为 M-LAG 系统 MAC 地址。
- M-LAG 设备从 M-LAG 接口收到环路检测报文后，会通过 peer-link 链路同步给另一台 M-LAG 设备；从非 M-LAG 接口收到环路检测报文后，不会通过 peer-link 链路同步。

对于单归接入二的情形：

- 开启环路检测功能后，M-LAG 设备即在开启了环路检测功能的接口上发送环路检测报文，报文的源 MAC 地址为 M-LAG 系统 MAC 地址。
- M-LAG 设备从非 M-LAG 接口收到环路检测报文后，不会通过 peer-link 链路同步给另一台 M-LAG 设备。

4. 产生环路的判断机制

开启环路检测功能后，M-LAG 设备从 peer-link 链路以外的任意端口收到 M-LAG 系统发送的环路检测报文时，均会判断该端口存在环路。

M-LAG 设备在开启了环路检测功能并从 peer-link 链路接收到同步的环路检测报文后，本端设备会判断与对端接收到环路检测报文的 M-LAG 接口属于同一 M-LAG 组的 M-LAG 接口也产生了环路。例如，Device A 从 BAGG4 接口收到环路检测报文并通过 peer-link 链路同步给 Device B 后，即使 Device B 没有从 BAGG4 接口收到环路检测报文，也会认为 BAGG4 接口上产生了环路，并对其进行相应的处理。

5. 检测到发生环路的处理机制

M-LAG 设备接收到本机发送或通过 peer-link 链路同步的环路检测报文，并判断端口出现环路后，会根据设备的配置对出现环路的端口执行关闭、阻塞或禁止 MAC 地址学习等操作。

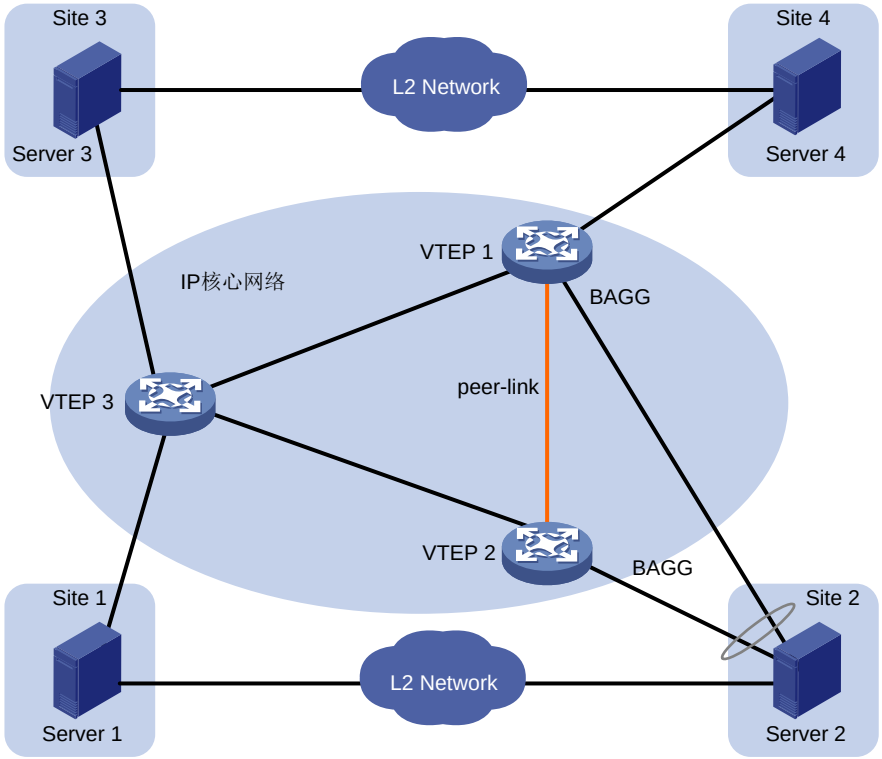
5.2.2 VXLAN 组网下的 M-LAG 设备环路检测工作机制

1. 功能简介

如图 34 所示，VXLAN 利用 M-LAG 功能将两台物理设备连接起来虚拟成一台设备，使用该虚拟设备作为 VTEP，用以避免 VTEP 单点故障对网络造成影响，从而提高 VXLAN 网络的可靠性。本章将通过以下几种情形，分别对环路检测的工作机制进行介绍：

- 双归接入情形：Server 2 双归接入 VTEP 经过 VXLAN 网络与 Server 1 实现互通，同时还通过直连的二层网络与 Server 1 互通，形成了环路。
- 单归接入情形：Server 4 单归接入 VTEP 经过 VXLAN 网络与 Server 3 实现互通，同时还通过直连的二层网络与 Server 3 互通，形成了环路。

图34 VXLAN 支持 M-LAG 网络运行环路检测组网图



2. 环路检测功能的生效机制

M-LAG 设备只有在开启了环路检测功能后，才能检测到环路以及对发生环路的 AC 进行处理。所以根据不同的情形，推荐的环路检测配置如下：

- 对于双归接入情形，用户需要在 M-LAG 的两台设备上均开启环路检测功能，否则将只有一台设备对环路进行处理，无法起到消除环路的作用。
- 对于单归接入情形，用户只需要在终端接入的 VTEP 上开启环路检测功能即可。

3. 环路检测报文发送机制

对于双归接入的情形：

- 开启环路检测功能后，M-LAG 的两台设备都会在 M-LAG 接口的 AC 上上发送环路检测报文，报文的源 MAC 地址为 M-LAG 系统 MAC 地址。
- M-LAG 设备从 M-LAG 接口收到环路检测报文后，会通过 peer-link 链路同步给另一台 M-LAG 设备，以避免单点故障导致另一台 M-LAG 设备无法收到环路检测报文。

对于单归接入的情形：

- 开启环路检测功能后，M-LAG 设备在终端接入的 AC 上发送环路检测报文，报文的源 MAC 地址为 M-LAG 系统 MAC 地址。
- M-LAG 设备从 M-LAG 接口收到环路检测报文后，会通过 peer-link 链路同步给另一台 M-LAG 设备；从非 M-LAG 接口收到环路检测报文后，不会通过 peer-link 链路同步。

4. 产生环路的判断机制

开启环路检测功能后，M-LAG 设备从 peer-link 链路以外的 AC 收到环路检测报文、且该环路检测报文携带的 VLAN Tag 与 AC 发送的环路检测报文相同时，会判断该 AC 存在环路。

M-LAG 设备在开启了环路检测功能并从 peer-link 链路接收到同步的环路检测报文后，本端设备会判断本端同一 M-LAG 组中，与环路检测报文属于相同 VXLAN 的 AC 也产生了环路。例如，VTEP 1 从 BAGG 接口上的 AC 收到环路检测报文并通过 peer-link 链路同步给 VTEP 2 后，即使 VTEP 2 没有从 BAGG 接口上的 AC 收到环路检测报文，也会认为本机上与 VTEP 1 上收到环路检测报文 AC 属于同一 VXLAN 的 AC 产生了环路。

5. 检测到发生环路的处理机制

M-LAG 设备从 AC 或者 peer-link 链路接收环路检测报文，并判断 AC 出现环路后，会根据收到的环路检测报文的优先级进行判断：

- 如果收到的环路检测报文的优先级更高，则 M-LAG 设备会根据配置对出现环路的 AC 进行阻塞等操作。
- 如果收到的环路检测报文的优先级更低，则 M-LAG 设备不会对出现环路的 AC 触发环路的处理动作。

6 DHCP/DHCPv6 支持 M-LAG

说明

- 当前 DHCP 支持 M-LAG 主要应用场景为 DHCP Snooping，DHCP 中继支持 M-LAG 的工作机制与 DHCP Snooping 基本一致，本章主要介绍 DHCP Snooping 支持 M-LAG 的工作机制。
- 本章内容中涉及的 DHCP Snooping 包括了 DHCPv4 Snooping 和 DHCPv6 Snooping。

6.1 功能简介

在网络中，为了提高 DHCP Snooping 设备的可靠性，可以配置 M-LAG 组网下的 DHCP 功能。将两台 DHCP Snooping 设备在聚合层面虚拟成一台设备来实现跨设备链路聚合，从而提供设备级冗余保护和流量负载分担。

6.2 工作机制

6.2.1 M-LAG 系统建立

DHCP Snooping 作为 M-LAG 设备，M-LAG 系统的建立及工作过程参见“2.3 M-LAG 系统建立和维护”。

6.2.2 用户表项同步机制

- 实时备份

DHCP 用户触发上线、续约或者下线时，由一端设备上生成、更新或删除用户表项，并将数据通过 peer-link 链路实时备份到对端设备上。

- 批量备份

  peer-link 接口状态由 DOWN 变 UP 时会触发批量备份事件，由 M-LAG 主设备将本端保存的用户表项发送给 M-LAG 备设备备份，备设备比主设备多余的用户表项也会发送给主设备备份，两端最终保存的表项是双方表项的并集。处于批备阶段时，实备消息会延迟发送。

6.3 流量转发

DHCP Snooping 支持 M-LAG 主要涉及如下几种应用场景。

6.3.1 M-LAG 设备双边聚合场景

图35 M-LAG 设备双边聚合场景（本端设备 M-LAG 上行口 UP）

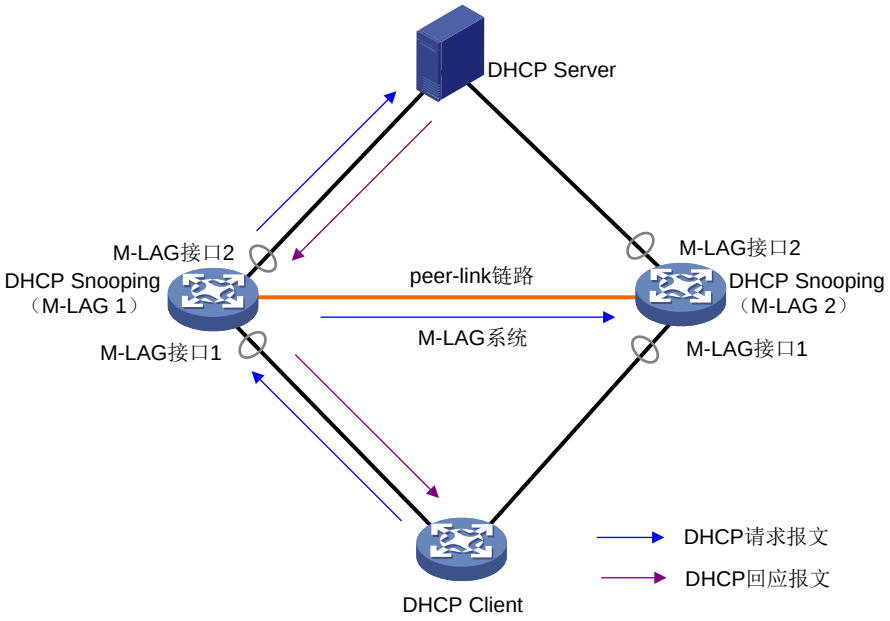
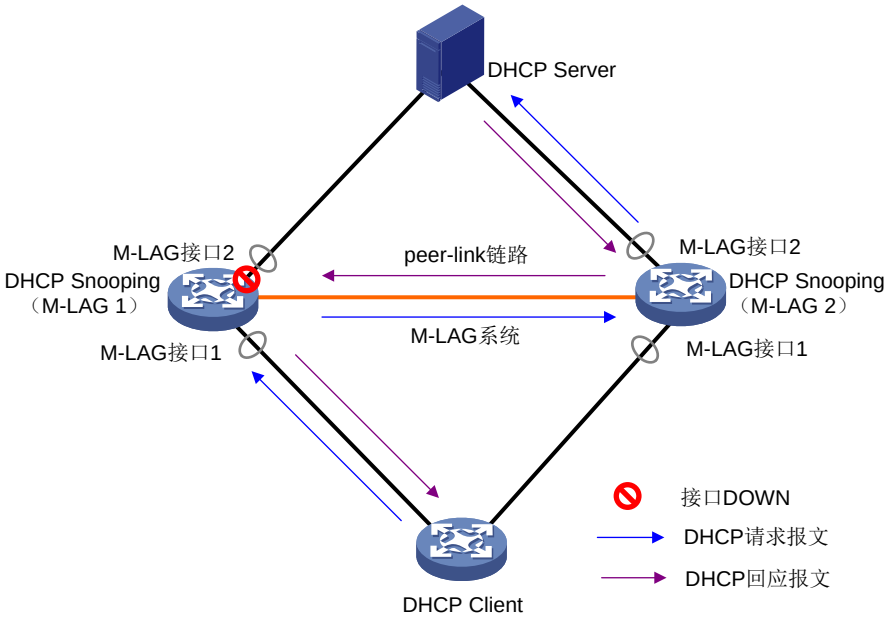


图36 M-LAG 设备双边聚合场景（本端设备 M-LAG 上行口 DOWN）



如图 35、图 36 所示，以 DHCP 请求报文发送给了 M-LAG 1（以下称为本端设备）为例：

(1) 本端设备收到 DHCP 客户端的请求报文后，生成用户的 MAC 地址临时表项（上线接口为 M-LAG 接口）。

(2) 本端设备通过 peer-link 链路将请求报文同步给对端设备，对端设备生成用户的 MAC 地址临时表项（上线接口为 peer-link 接口）。

(3) 后续请求报文转发与设备上行口状态有关：

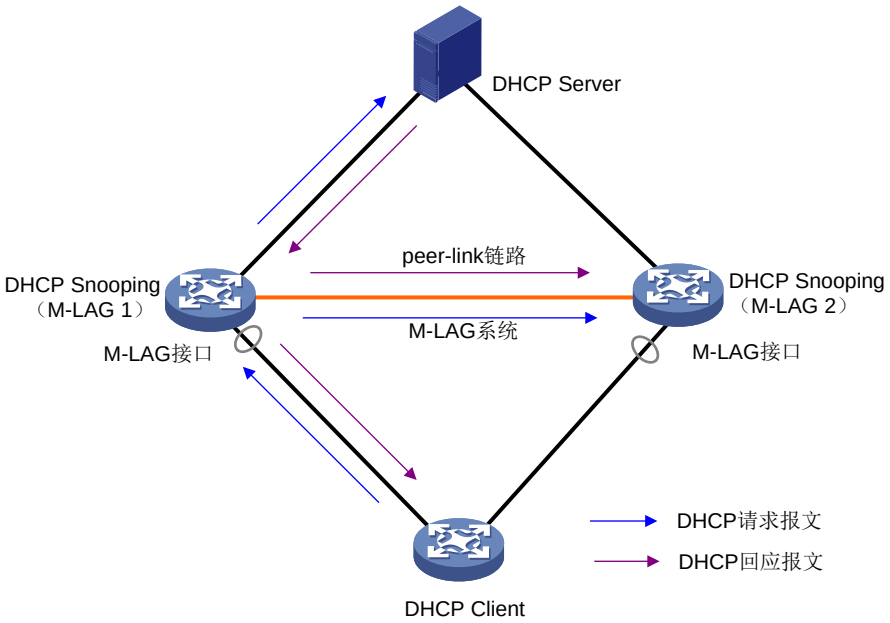
- 如图 35 所示，如果本端设备 M-LAG 上行口 UP，则本端设备会将请求报文上送给 DHCP 服务器处理。对端设备通过 peer-link 链路感知到本端 M-LAG 口状态为 UP，则直接丢弃请求报文。
- 如图 36 所示，如果本端设备 M-LAG 上行口故障，对端设备通过 peer-link 链路感知到本端设备上行 M-LAG 口 DOWN，则由对端设备将请求报文上送给 DHCP 服务器。

(4) DHCP 服务器收到请求报文处理后后，发送回应报文：

- 如果是本端设备收到回应报文，查找 MAC 地址转发表项后发现上线口为 M-LAG 口，则直接转发给 DHCP 客户端，不再同步给对端设备，同时在本端生成用户的 DHCP Snooping 绑定表项，并通过 peer-link 链路实时备份到对端设备。
- 如果是对端设备收到回应报文，会通过 peer-link 链路将回应报文同步给本端设备。对端设备查找 MAC 地址转发表项后发现上线口为 peer-link 接口，收包口是 M-LAG 口，直接丢弃此报文，由本端设备处理回应报文并转发给 DHCP 客户端。同样，本端会生成用户的 DHCP Snooping 绑定表项，并通过 peer-link 链路实时备份到对端设备。

6.3.2 M-LAG 设备下行口单边聚合场景

图37 M-LAG 设备下行口单边聚合场景



如图 37 所示，M-LAG 设备下行口单边聚合与双边聚合流量转发的主要差异在于：

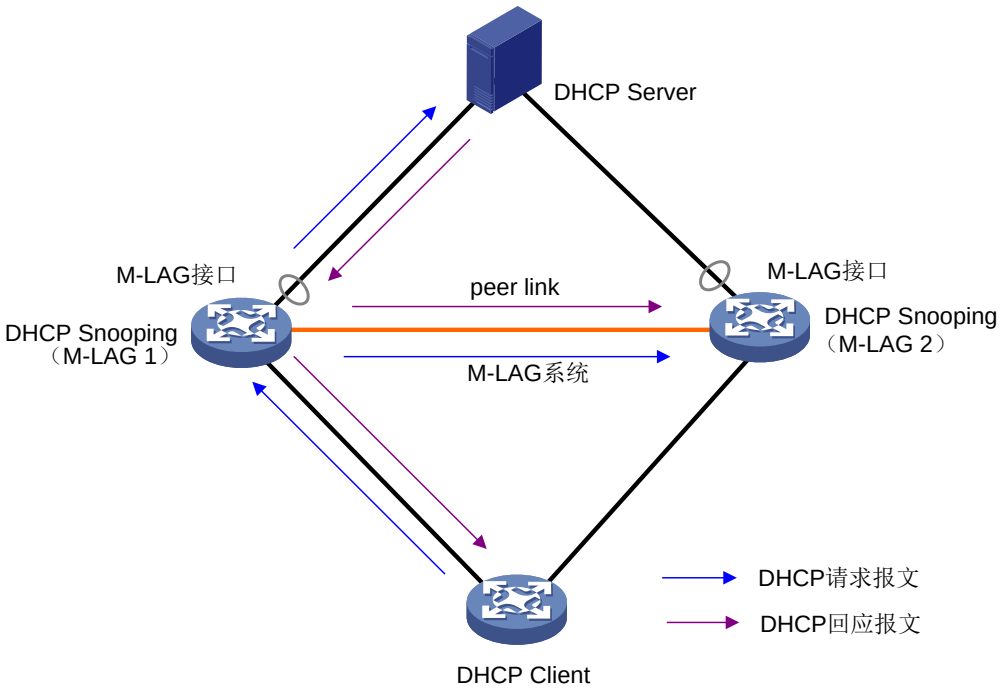
- 双边聚合场景下，M-LAG 设备的上行口均为 M-LAG 口；下行口单边聚合场景下，M-LAG 设备的上行口均为普通口。
- 双边聚合时两端 M-LAG 口均可以同时处于 UP 状态，而上行口非聚合情况下，由于生成树协议作用，为了避免回环，会选择阻塞一端的上行口。
- M-LAG 设备收到回应报文后，不是通过 peer-link 链路将回应报文转发给对端，而是通过 DHCP Snooping 模块转发给对端。

对于流量转发：

- 如果生成树协议选择阻塞对端设备上行口，则流量转发情况同 6.3.1 图 35。
- 如果生成树协议选择阻塞本端设备上行口，则流量转发路线同 6.3.1 图 36。不同的是，此时收包口为普通口，对端设备收到回应报文后，不会直接丢弃报文，会处理回应报文，但由于上线口为 peer-link 口，仍无法将报文转发给 DHCP 客户端，同样由本端设备处理同步过来的回应报文并转发给 DHCP 客户端。

6.3.3 M-LAG 设备上行口单边聚合场景

图38 M-LAG 设备上行口单边聚合场景



如图 38，M-LAG 设备上行口单边聚合与双边聚合的主要差异在于：

- 双边聚合场景下，M-LAG 设备上行口均为 M-LAG 口；上行口单边聚合场景下，M-LAG 设备的下行口均为普通口。
- 上行口单边聚合场景下，本端设备收到 DHCP 客户端的请求报文后，生成的 MAC 地址临时表项中用户的上线接口为普通口。
- 双边聚合时下行口均为 M-LAG 口，DHCP 客户端的请求报文随机发送。而下行口非聚合情况下，由于生成树协议作用，所有 DHCP 客户端的请求报文只会发送给一端设备。
- M-LAG 设备收到请求报文后，不是通过 peer-link 链路将请求报文转发给对端，而是由 DHCP Snooping 模块转发给对端。

上行口单边聚合流量转发过程同双边聚合场景。

7 安全机制支持 M-LAG

7.1 功能简介

设备正常运行时，用户可通过 M-LAG 接口发起认证。对于同一用户，认证、计费动作只在其中一台 M-LAG 设备执行，授权动作在两台 M-LAG 设备都会执行。M-LAG 接口上授权成功的用户，可以通过任意 M-LAG 设备上的 M-LAG 接口访问网络资源。当其中一台 M-LAG 设备发生故障，原来在该设备进行认证、计费的用户将在另一台 M-LAG 设备上接替运行，继续与服务器交互计费报文等信息，保持在线状态，可以继续访问网络资源。因此，M-LAG 场景下可以实现端口安全、Portal 业务的设备级负载分担和冗余备份。

本文提到的“端口安全用户”是 802.1X 用户、MAC 地址认证用户、Web 认证用户、静态用户的统称。

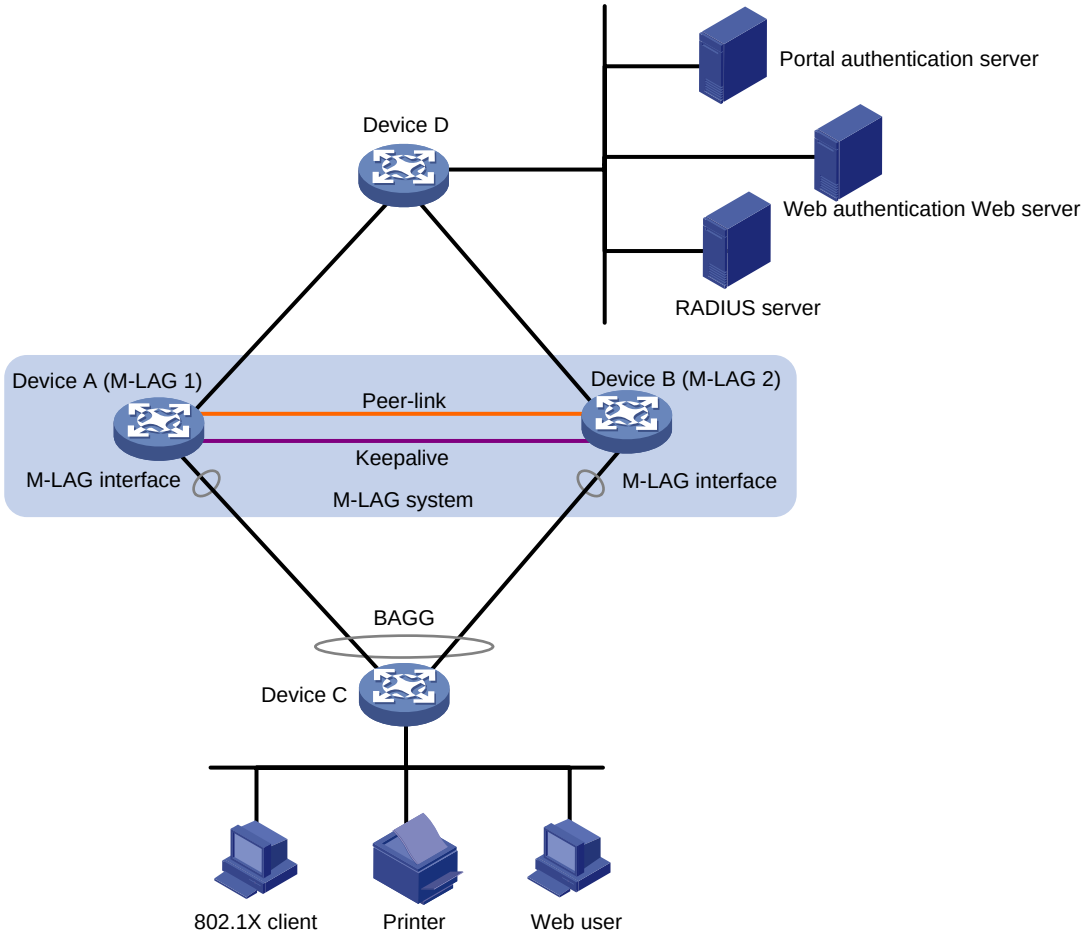
7.2 工作机制

7.2.1 端口安全支持 M-LAG

1. 典型组网

有线 802.1X 用户、MAC 地址认证用户、Web 认证用户、静态用户支持 M-LAG 的典型组网如下所示。

图39 端口安全支持 M-LAG 组网示意图



2. 配置一致性检查

M-LAG 系统会对端口安全业务进行 Type 2 类型的配置一致性检查：

- 如果配置一致性检查不通过，则丢弃触发认证的用户报文。
- 如果端口安全的配置从一致改变为不一致，则已经在线用户保持在线，新用户不允许上线。
- 如果 peer-link 链路故障恢复，M-LAG 系统发现端口安全的配置不一致，则会强制已在线用户下线。

- 允许 M-LAG 接口单边接入的情况下，不会对该 M-LAG 接口的端口安全业务进行配置一致性检查。当从 M-LAG 接口单边接入变为两台 M-LAG 设备上的 M-LAG 接口都正常工作时，M-LAG 系统发现端口安全的配置不一致，会强制之前单边接入的 M-LAG 接口上的已在线用户下线。

说明

M-LAG 接口单边接入是指，仅当前设备上配置了 M-LAG 接口，对端未配置 M-LAG 接口。

3. 用户报文的上送与分发

用户发送的报文，经由接入设备转发时，将根据接入设备的聚合口上配置的负载分担模式决定最终转发到哪一台 M-LAG 设备的 M-LAG 接口上进行处理。

M-LAG 设备收到用户的未知源 MAC 报文和 802.1X EAPOL 协议报文后，在本地创建用户表项，并根据端口安全模块配置的 M-LAG 接口上用户认证的负载分担模式，决定报文是直接在本设备处理，还是通知对端 M-LAG 设备处理。

- 集中式处理模式下：
  - 如果主设备收到用户报文，那么就在主设备上直接处理。
  - 如果备设备收到用户报文，先进行必要的报文解析，然后通知主设备进行后续处理，并由主设备主动与 AAA 服务器、客户端进行认证相关报文的交互。

  该模式下，配置较为简单，RADIUS 服务器上仅需管理一个接入设备 IP，两台 M-LAG 设备上仅需配置一个相同的 RADIUS 报文源 IP 地址，但用户报文的分发效率较低，适用于接入用户量较小的场景。
- 分布式处理模式下：
  - 本地模式：本端 M-LAG 接口上送的用户报文就在本端 M-LAG 设备处理。
  - 奇偶模式：当主机收到用户报文时，解析报文中的用户源 MAC，根据端口安全配置决定，奇 MAC 在一台 M-LAG 设备上处理，偶 MAC 在另外一台 M-LAG 设备上处理。如果需要对端处理，就透传到对端处理。

  该模式下，用户报文的分发效率较高，但是 RADIUS 服务器上需管理两个接入设备 IP，两台 M-LAG 设备上需同时配置本地设备以及对端设备使用的 RADIUS 报文源 IP 地址，适用于集中上线用户量较大的场景。其中，本地模式的分发效率最高，奇偶模式次之。

说明

当允许 M-LAG 接口单边接入时：

- 集中式处理模式下，M-LAG 接口会将用户的认证报文丢弃。
- 分布式处理模式下，无论配置了本地模式还是奇偶模式，均按照本地模式来处理。

如果任何一台 M-LAG 设备上修改了 M-LAG 接口上用户认证的负载分担模式，则所有 M-LAG 接口上的用户都会被强制下线。

peer-link 接口故障时，M-LAG 备设备上的所有 M-LAG 接口将处于 MAD DOWN 状态，且接口上的用户表项会被删除，此时将仅由 M-LAG 主设备处理用户报文。

4. 用户认证、授权、计费

(1) 假设用户报文分发给 M-LAG 1 设备处理，那么 M-LAG 1 上该用户表项就会置为 Active 状态，并由 M-LAG 1 与 RADIUS 服务器交互用户的认证、授权、计费报文。

(2) M-LAG 1 上用户认证通过后，执行本地授权，同时同步包含授权信息的用户数据到对端 M-LAG 设备。对端 M-LAG 设备同步此用户数据之后，也会进行本地授权，使得用户无论是通过 M-LAG 1 还是通过 M-LAG 2 上的 M-LAG 接口，都可以访问授权的网络资源。

(3) 两端 M-LAG 设备授权成功后，M-LAG 1 需要向客户端发送认证通过报文，向服务器发送计费开始报文。由于 M-LAG 2 上的该用户作为备份用户，不需要与服务器交互。

5. 用户表项同步

用户表项的创建及实时备份过程如下：

(1) 本端用户认证成功后，本端 M-LAG 设备将向对端 M-LAG 设备实时同步用户信息（包括源 MAC 地址、VLAN、授权信息等）。

(2) 对端 M-LAG 设备根据同步信息建立用户表项，下发授权信息。

说明

对于 Web 认证，两台设备上都要下发相同的重定向 URL 规则，以保证用户的 HTTP 报文从任意一台设备上来时，均可以重定向到 Portal Web 服务器。

用户表项的删除过程如下：

(1) 本端用户下线时，本端 M-LAG 设备通知对端 M-LAG 设备同步对用户进行下线处理，包括删除用户表项、取消用户授权、统计最终流量，并随此下线通知消息交互最终的用户流量。

(2) 由用户处于 Active 状态的 M-LAG 设备向 RADIUS 服务器发送计费停止报文，报文中携带的用户流量数据为两台 M-LAG 设备上叠加的最终用户流量数据。

(3) 对于 802.1X 用户，用户状态为 Active 的 M-LAG 设备需要通知 802.1X 客户端下线。

说明

两端 M-LAG 设备上均开启了相应认证类型用户的下线检测功能的情况下，当两端 M-LAG 设备均未检测到用户流量时，才会触发用户下线，并删除用户表项。

当 peer-link 链路 UP、端口安全进程重启或者单台 M-LAG 主备倒换时，两台 M-LAG 设备将会互相批量同步自身的用户数据，收到对端数据后都以激活状态的用户数据为准保存用户表项。

6. 流量统计及流量数据同步

由于两台 M-LAG 设备上都存在相同的用户表项，而且同一个用户的流量可能分布在两台设备上，所以同一个用户在两台设备上的流量统计信息需要互相同步：

- 两台 M-LAG 设备都会在用户授权成功后，开启统计流量功能，并将用户的流量统计值周期性（缺省 60 秒）地发往对端，并在收到对端的数据后叠加出用户的总流量。
- 如果当前 M-LAG 设备上的用户处于 Active 状态，则由该 M-LAG 设备使用叠加出的总流量，向 RADIUS 服务器发送计费更新报文，并在下线时发送计费停止报文。

7. 端口迁移

端口安全支持如下情况下，两台 M-LAG 设备之间或者单台 M-LAG 设备上发生的端口间用户迁移：

- M-LAG 接口间发生的用户迁移。
- M-LAG 口与非 M-LAG 口之间的用户迁移。
- 非 M-LAG 口之间的用户迁移。

说明

如果需要两台 M-LAG 设备上的非 M-LAG 接口之间进行用户迁移，需要在两台 M-LAG 设备均配置允许同步的远端 MAC 表项覆盖本端的原 MAC 表项功能。

8. 用户状态切换

表5 用户状态切换

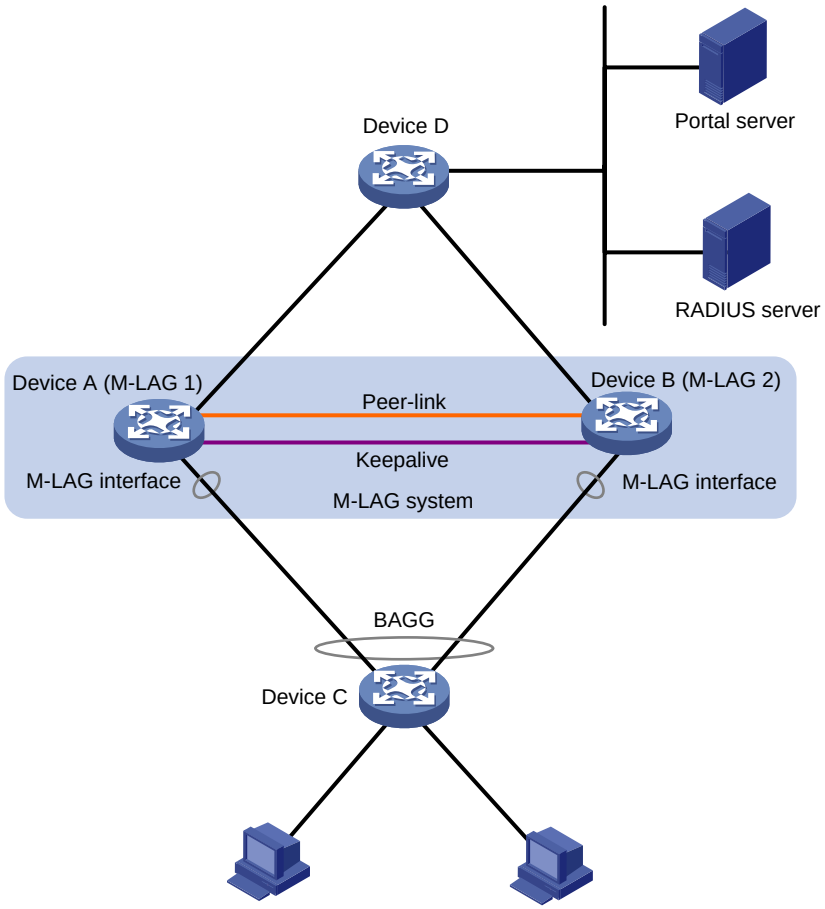
系统状态	用户认证的负载分担模式	用户状态
系统正常运行	集中式处理模式	• 主设备上的用户全部为 Active 状态 • 备设备上的用户全部为 Inactive 状态
	分布式处理模式	• 本设备上线的用户为 Active 状态 • 对端备份过来的用户为 Inactive 状态
peer-link接口故障	集中式处理模式	• 主设备上的用户全部为 Active 状态 • 备设备上的用户全部为 Inactive 状态
	分布式处理模式	• 主设备上的用户全部为 Active 状态 • 备设备上的用户表项被删除

7.2.2 Portal 支持 M-LAG 功能

1. 典型组网

在该组网中，若 Portal 用户在本端 M-LAG 设备上通过认证，本端 M-LAG 设备会将用户数据同步发送到对端 M-LAG 设备进行备份。当一端 M-LAG 设备发生故障时，对端 M-LAG 设备可以使用备份的用户数据接替处理业务，从而保证用户业务的正常运行。

图40 Portal 支持 M-LAG 功能组网图



2. 配置一致性检查

Portal 业务启动后，会对 M-LAG 接口所属的相同 VLAN 接口上的 Portal 配置进行一致性检查：

- 如果配置一致性检查不通过，则丢弃触发认证的用户报文。
- 如果 Portal 的配置从一致转变为不一致，则已经在线用户保持在线，新用户不允许上线。
- 如果 peer-link 链路故障恢复后，M-LAG 系统发现 Portal 配置不一致，则已经在线用户保持在线，新用户不允许上线。

说明

在单归接入的设备上不允许 Portal 用户上线。

---

3. 用户报文的上送与分发

用户发送的报文，经由接入设备转发时，将根据接入设备的聚合口上配置的负载分担模式决定最终转发到哪一台 M-LAG 设备的 M-LAG 接口上进行处理。

4. HTTP 重定向报文的处理

M-LAG 设备按照如下原则对用户的 HTTP/HTTPS 报文进行重定向处理：

(1) 如果主机或者备机收到首个 HTTP/HTTPS 报文，则记录一下报文的五元组信息。

(2) 下一个 HTTP/HTTPS 报文来到主机或者备机的时候，则检查本机是否有首个 HTTP/HTTPS 报文的记录。如果有，就在本端进行重定向处理；如果没有，就丢弃该 HTTP/HTTPS 报文。如果收到了非分片报文，或者最后一个 HTTP 分片报文，则也会尝试进行重定向处理。

说明

如果用户的同一条 HTTP/HTTPS 报文流（五元组一致）被上送到不同的 M-LAG 设备上，那么该用户的 HTTP 重定向处理可能会失败。

5. Portal 协议报文的处理

M-LAG 设备完成用户 HTTP 报文的重定向处理后，由 Portal 服务器向 M-LAG 设备发起认证请求。

- 集中式处理模式下：
  - 如果主设备收到 Portal 协议报文，那么就在主设备上直接处理，并创建用户表项。
  - 如果备设备收到 Portal 协议报文，先进行报文的必要解析，然后将报文透传给主设备进行后续的 Portal 业务处理，并由主设备上回应 Portal 服务器。

  该模式下，配置较为简单，RADIUS 服务器上仅需管理一个接入设备 IP，两台 M-LAG 设备上仅需配置一个相同的 RADIUS 报文源 IP 地址，但用户报文的分发效率较低，适用于接入用户量较小的场景。
- 分布式处理模式下：
  - M-LAG 设备会按照 Portal 协议报中携带的用户 IP 地址信息做负载分担。
  - 当 M-LAG 设备收到一个 Portal 协议报文时，解析报文中的用户源 IP，根据设备上的 Portal 配置决定，奇 IP 在一台 M-LAG 设备上处理，偶 IP 在另外一台 M-LAG 设备上处理。如果需要对端处理，就透传到对端进行 Portal 业务处理，并由对端设备回应 Portal 服务器，否则在本端进行 Portal 业务处理，并创建用户表项。

  该模式下，用户报文的分发效率较高，但是 RADIUS 服务器上需管理两个接入设备 IP，两台 M-LAG 设备上需同时配置本地设备以及对端设备使用的 RADIUS 报文源 IP 地址，适用于集中上线用户量较大的场景。

6. 用户认证、授权、计费

(1) 假设用户的 Portal 协议报文由 M-LAG 1 设备处理，那么就由 M-LAG 1 与 RADIUS 服务器交互用户的认证、授权、计费报文。

(2) M-LAG 1 上的用户认证通过后，该用户表项就会置为激活状态，M-LAG 1 将向 Portal 服务器发送认证通过报文。

(3) M-LAG 1 使用 RADIUS 服务器下发的授权信息在本地对用户进行授权，并同时将包含授权信息的用户数据同步给对端 M-LAG 设备。对端 M-LAG 设备同步此用户数据之后，也会进行本地授权，使得用户无论是通过 M-LAG 1 还是通过 M-LAG 2 上的 M-LAG 接口，都可以访问授权的网络资源。

(4) 两端 M-LAG 设备授权成功后，M-LAG 1 向 RADIUS 服务器发送计费开始报文。由于 M-LAG 2 上的该用户只作为备份用户，不需要与 RADIUS 服务器交互。

7. 用户表项同步

用户表项的创建及实时备份过程如下：

(1) 本端用户认证成功后，本端 M-LAG 设备将向对端 M-LAG 设备实时同步用户信息（包括用户 IP 地址、源 MAC 地址、VLAN、授权信息等）。

(2) 对端 M-LAG 根据同步信息建立用户表项，下发授权信息。

用户表项的删除过程如下：

(1) 如果用户主动下线，Portal 服务器会通知 M-LAG 设备对用户进行下线处理。M-LAG 系统将按照配置的 Portal 业务的 M-LAG 处理模式，对 Portal 服务器的下线请求报文进行分发。最终，由用户状态为 Active 的 M-LAG 设备对用户进行下线处理，并通知对端 M-LAG 设备同步处理，包括删除用户表项、取消用户授权、统计最终流量。

(2) 如果在 M-LAG 设备上强制用户下线，则由用户状态为 Active 的 M-LAG 设备向 Portal 服务器发送下线通知报文，并通知对端 M-LAG 设备同步处理，包括删除用户表项、取消用户授权、统计最终流量。

(3) 由用户处于 Active 状态的 M-LAG 设备向 RADIUS 服务器发送计费停止报文，报文中携带的用户流量数据为两台 M-LAG 设备上叠加的最终用户流量数据。

当 peer-link 链路 UP、端口安全进程重启、单台 M-LAG 主备倒换或者 Portal 配置从不一致转变为一致时，两台 M-LAG 设备将会互相批量同步自身的用户数据，收到对端数据后都以激活状态的用户数据为准保存用户表项。

8. 流量统计及流量数据数据同步

由于两台 M-LAG 设备上都下发了相同的 Portal 规则，而且同一个用户的流量可能分布在两台设备上，所以同一个用户在两台设备上的流量统计信息需要互相同步（包括 ITA 用户的流量）：

- 当 M-LAG 设备收到的 Portal 用户流量数值与上次备份的流量数值差值达到设定的阈值，或者 Portal 用户流量数值达到指定的流量备份周期，将触发本端设备将用户流量备份给对端 M-LAG 设备。
- 如果当前 M-LAG 设备上的用户处于 Active 状态，则由该 M-LAG 设备使用叠加出的总流量，向 RADIUS 服务器发送计费更新报文，并在下线时发送计费停止报文。

9. 用户状态切换

表6 用户状态切换

系统状态	用户认证的负载分担模式	用户状态
系统正常运行	集中式处理模式	• 主设备上的用户全部为 Active 状态 • 备设备上的用户全部为 Inactive 状态
	分布式处理模式	• 本设备上线的用户为 Active 状态 • 对端备份过来的用户为 Inactive 状态
peer-link接口故障	集中式处理模式	• 主设备上的用户全部为 Active 状态 • 备设备上的用户全部为 Inactive 状态
	分布式处理模式	• 主设备上的用户全部为 Active 状态 • 备设备上的用户表项保持现状

7.2.3 RADIUS 协议报文处理

认证服务器需要根据 NAS-IP 识别用户身份，为保证交互不受影响，处理备份用户业务时，需要使用对端 M-LAG 的 IP 地址作为源 IP 地址。在两台 M-LAG 设备上分别部署 M-LAG 虚拟 IP 地址后，设备正常工作时，使用本端 M-LAG 虚拟 IP 地址与 RADIUS 服务器交互本地用户信息，一台设备故障时，另一台设备就要使用对端的 M-LAG 虚拟 IP 地址和 RADIUS 服务器交互对端备份用户信息。

1. 认证报文、授权报文、计费报文（设备作为 RADIUS 客户端）

总处理原则：用户上线后，与 RADIUS 服务器交互时采用的 RADIUS 源 IP 地址保持不变。

- 集中处理模式下，需要在两台 M-LAG 设备上均配置相同的 Local 源 IP 地址，所有的认证、授权、计费报文都在主设备上进行处理。该源 IP 地址必须为同一个 M-LAG 虚拟 IP 地址。
  - 主设备正常的情况下，由主设备使用 Local 源 IP 地址与服务器交互 RADIUS 协议报文。
  - 主设备发生故障时，由备设备使用 Local 源 IP 地址与服务器交互 RADIUS 协议报文。
- 分布处理模式下，需要在两台 M-LAG 设备上均配置一个 Local 源 IP 地址和一个 Peer 源 IP 地址。即，两台设备上分别配置两个不同的 M-LAG 虚拟 IP 地址，且这两对地址彼此相反。

  假设，M-LAG 1 设备上的 Local 源 IP 地址为 A、Peer 源 IP 地址为 B，M-LAG 2 设备上的 Local 源 IP 地址为 B、Peer 源 IP 地址为 A，则：
  - 两台 M-LAG 设备均正常的情况下，M-LAG 1 设备使用 A 地址与服务器交互 RADIUS 协议报文，M-LAG 2 设备使用 B 地址与服务器交互 RADIUS 协议报文。
  - 当一台 M-LAG 设备发生故障时，原故障设备上处理的用户在另外一台 M-LAG 设备上使用配置的 Peer 源 IP 地址发送 RADIUS 报文。

说明

M-LAG 组网环境中，M-LAG 设备上指定的发送 RADIUS 报文使用的源 IP 地址必须为 M-LAG 虚拟 IP 地址。两台 M-LAG 设备的 Loopback 接口配置不同的虚拟 IP 地址，且均配置为 active 状态。

2. COA 报文（设备作为 RADIUS 服务端）

对于端口安全业务，COA 报文处理机制如下：

- 如果主机收到 COA 协议报文，那么就在主机上直接处理。
- 如果备机收到 COA 协议报文，那么就在备机上直接处理。

对于 Portal 业务，COA 报文处理机制如下：

- 如果收到 COA 协议报文的 M-LAG 设备上该用户状态为 Active，那么就在本机上直接处理。
- 如果收到 COA 协议报文的 M-LAG 设备上该用户状态为 Inactive，那么就透传到对端处理。

8 二层组播支持 M-LAG

8.1 功能简介

二层组播利用 M-LAG 功能将两台物理设备连接起来形成 M-LAG 系统，该 M-LAG 系统作为一台虚拟二层组播设备使用。使用该虚拟设备连接组播源或组播接收者，可避免单点故障对组播网络造成影响，提高组播网络可靠性。

8.2 工作机制

说明

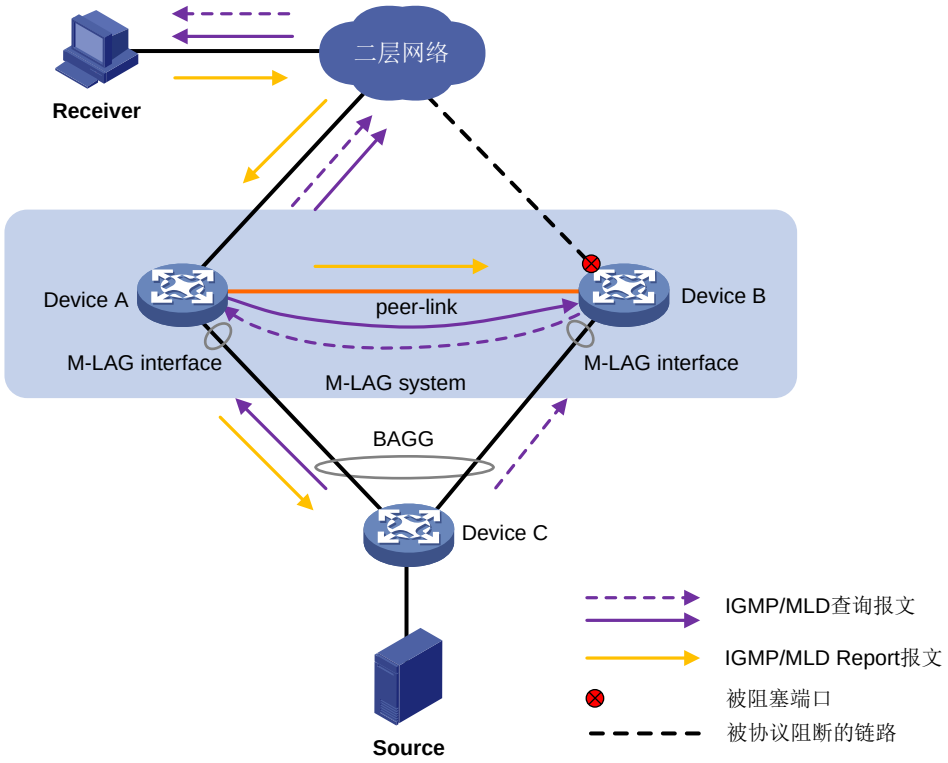
二层组播支持 M-LAG 组网中，IPP 口作为静态路由器端口，承载所有的组播流量。

8.2.1 组播源连接 M-LAG 系统

1. 二层组播转发表项建立过程

如图 41 所示，作为 M-LAG 设备的 Device A 和 Device B 通过 peer-link 链路连接，Device C 与组播源相连。由于 M-LAG 系统上行接入二层网络，因此需要配置 STP 协议，选择性地阻塞某些端口来消除二层环路，确保组播源发送到 M-LAG 系统的组播数据流量，只能被 Device A 或 Device B 的其中一台转发给组播接收者。此处，以阻断 Device B 与二层网络相连的链路为例。

图41 二层组播转发表项建立过程（组播源连接 M-LAG 系统）



组播源连接 M-LAG 时，二层组播表项的建立过程如下：

(1) Device C 发送的 IGMP/MLD 查询报文，通过 Device C 的聚合口进行负载分担，分别到达 Device A 和 Device B 的 M-LAG 接口。

(2) Device A 和 Device B 分别将各自的 M-LAG 接口添加到路由器端口列表中，并通过 peer-link 链路互相发送给对端设备。Device A 和 Device B 从 peer-link 接口收到的对端发送的查询报文，不会再向各自的 M-LAG 接口转发。

(3) Device A 收到来自下游接收者的 IGMP/MLD Report 报文后，从 Report 报文中解析出主机要加入的组播组地址 G，生成二层组播转发表项（*，G），将接收端口作为成员端口添加到出端口列表。同时，将 IGMP/MLD Report 报文通过 peer-link 链路发送给 Device B。

(4) Device B 收到 Report 报文后，同样生成二层组播转发表项（*，G），同时将 peer-link 接口作为成员端口添加到出端口列表中。但是，Device B 并不会将该报文从自己的路由器端口（M-LAG 接口）转发出去。

通过上述机制使得 Device A 和 Device B 上生成相同的二层组播转发表项，形成设备级别的备份，当一台成员设备发生故障（设备故障、上下行链路故障等）时，组播流量可以由另一台成员设备进行转发，从而避免组播流量转发中断。

2. 正常情况下组播流量转发过程

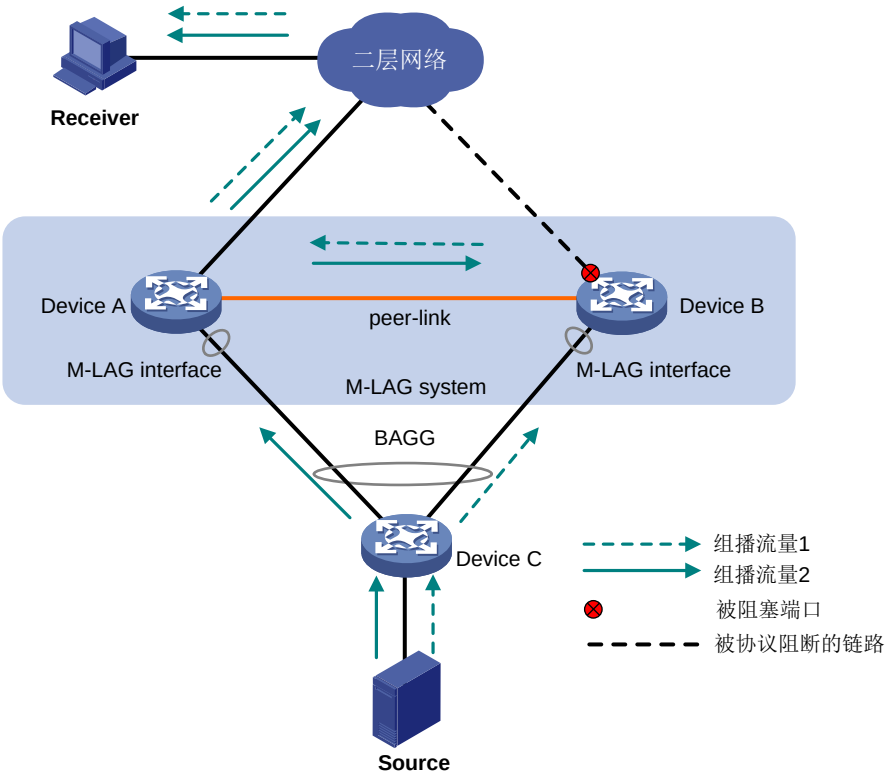
M-LAG 设备正常工作时，组播流量转发过程如图 42 所示。

(1) 组播源发送的组播流量，通过负载分担方式到达 Device A 和 Device B。

(2) Device A 和 Device B，通过 peer-link 链路互相发送各自接收到的组播数据流量给对方。这样，保证 Device A 和 Device B 上都能收到完整的组播数据流量。

(3) Device A 将完整的组播数据流量发送给下游接收者。

图42 组播流量转发过程（组播源连接 M-LAG 系统）

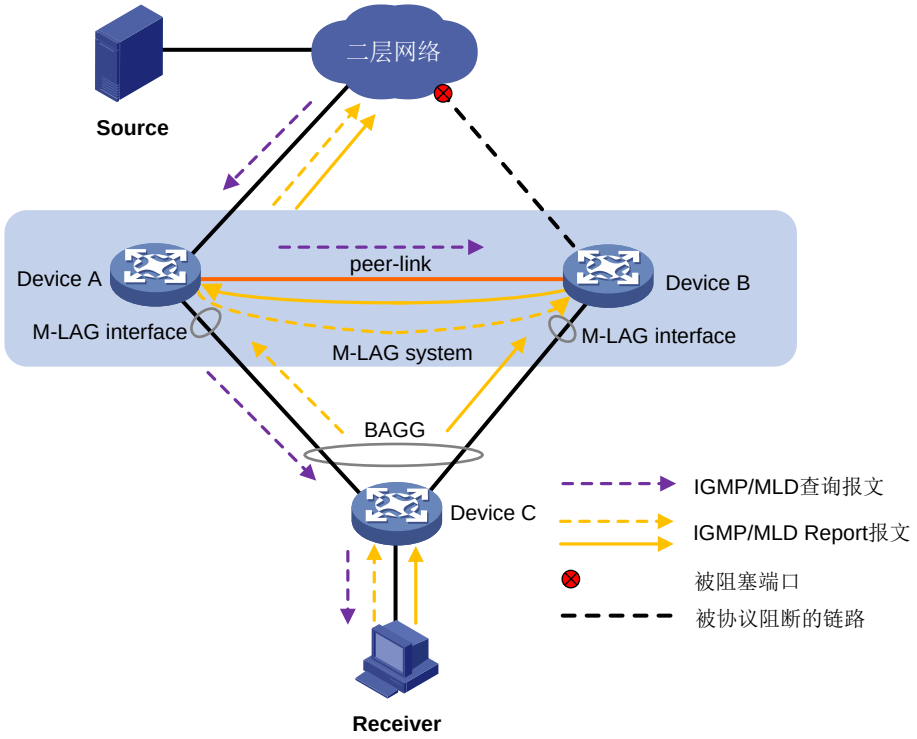


8.2.2 组播接收者连接 M-LAG 系统

如图 43 所示，作为 M-LAG 设备的 Device A 和 Device B 通过 peer-link 链路连接，Device C 与组播接收者相连。由于 M-LAG 系统上行接入二层网络，因此需要配置 STP 协议，选择性地阻塞某些

端口来消除二层环路，确保组播源发送到 M-LAG 系统的组播数据流量，只能被 Device A 或 Device B 的其中一台转发给组播接收者。此处，以阻断 Device B 与二层网络相连的链路为例。

图43 二层组播转发表项建立过程（组播接收者连接 M-LAG 系统）



组播接收者连接 M-LAG 时，二层组播表项的建立过程如下：

(1) Device A 收到上游二层网络发送的 IGMP/MLD 查询报文后，将接收报文的端口添加到动态路由器端口列表中，并通过 peer-link 链路发送给 Device B。

(2) Device B 收到查询报文后，将 peer-link 接口添加到路由器端口列表中。

(3) Device A 或者 Device B 中的任意一台设备，收到下游设备 Device C 发送的 IGMP/MLD Report 报文后，从 Report 报文中解析出主机要加入的组播组地址 G1 和 G2，分别在各自设备上生成二层组播转发表项（*，G1）和（*，G2），出端口分别为各自的 M-LAG 接口。

说明

假设 Device A 收到的为加入组播组 G1 的 Report 报文，Device B 收到的为加入组播组 G2 的 Report 报文。

(4) Device A 和 Device B 分别将 Report 报文通过 peer-link 链路透传给对端设备。以 Device A 为例，Device A 收到 Device B 发送的 Report 报文后，在生成二层组播转发表项（*，G2），同时将 Device A 上的 M-LAG 接口添加到出端口列表中。Device B 的处理过程类似，将在本地生成（*，G1）的二层组播转发表项。

通过上述机制使得 Device A 和 Device B 上的生成相同的二层组播转发表项，形成设备级别的备份，当一台成员设备发生故障（设备故障、上下行链路故障等）时，组播流量可以由另一台成员设备进行转发，从而避免组播流量转发中断。

2. 正常情况下组播流量转发过程

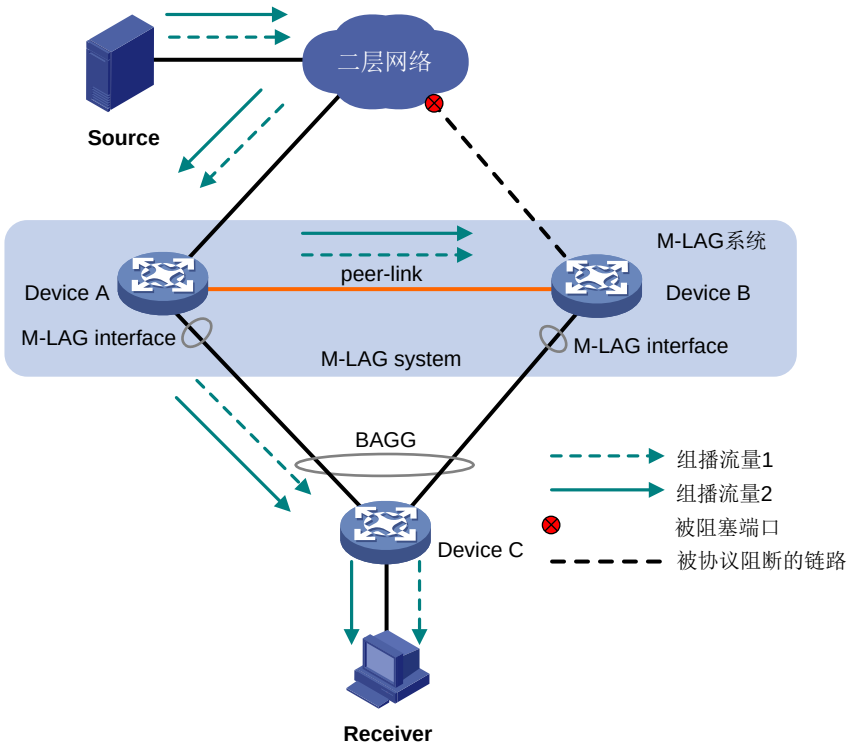
M-LAG 设备正常工作时，组播流量转发过程如图 44 所示。

(1) 组播源发送的组播流量，通过二层网络到达 Device A。

(2) Device A 通过 peer-link 链路将收到的组播数据流量发送给 Device B。此时，只有 Device A 会向下游 Device C 转发组播流量，而 Device B 上虽然生成了二层组播转发表项，但不会向下游 Device C 转发组播流量。

(3) Device C 收到后，将流量转发给接收者。

图44 组播流量转发过程（组播接收者连接 M-LAG 系统）



9 三层组播支持 M-LAG

9.1 功能简介

三层组播利用 M-LAG 功能将两台物理设备连接起来虚拟成一台设备，使用该虚拟设备连接组播源或组播接收者，可避免单点故障对组播网络造成影响，提高组播网络可靠性。

9.2 工作机制

说明

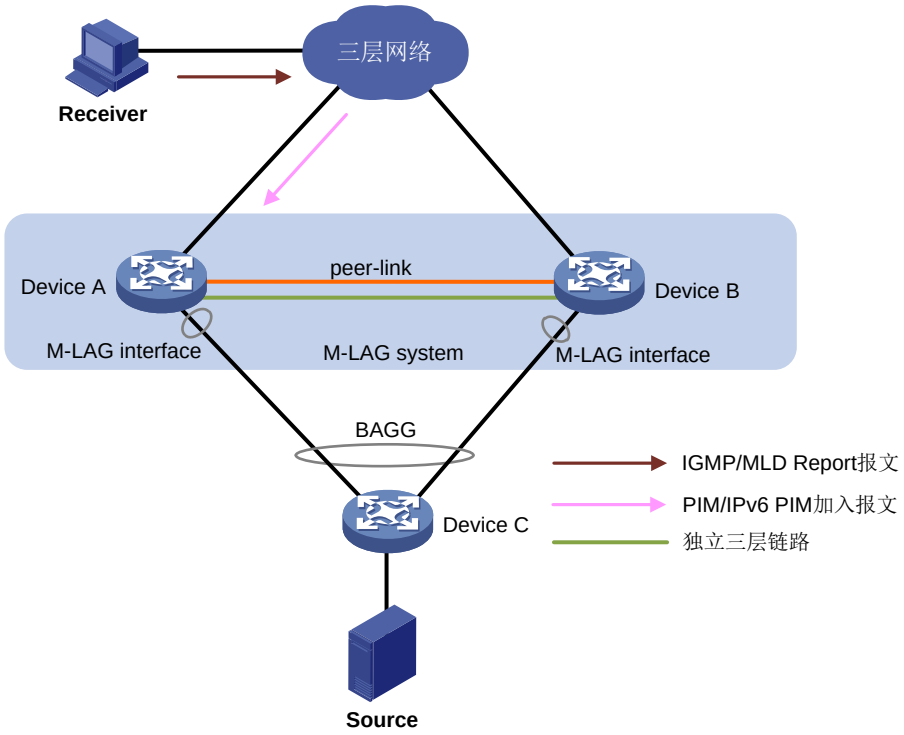
三层组播支持 M-LAG 场景中，需要在 M-LAG 设备之间配置一条独立的三层链路建立 PIM 邻居，并且作为 M-LAG 系统的 Keepalive 链路使用。

9.2.1 组播源连接 M-LAG 系统

1. 三层组播转发表项建立过程

如图 45 所示，作为 M-LAG 设备的 Device A 和 Device B 通过 peer-link 链路连接，Device C 与组播源相连。其中，peer-link 链路与 M-LAG 接口属于同一个 VLAN。接收者位于三层网络侧，会在三层网络上选择一条到组播源的最优路径，此处以选择上游设备 Device A 为例。

图45 三层组播转发表项建立过程（组播源连接 M-LAG 系统）



组播源连接 M-LAG 时，三层组播表项的建立过程如下：

(1) Device A 收到 PIM/IPv6 PIM 加入报文后，不会将报文通过 peer-link 链路同步给 Device B，而是在获取了组播组的接收者信息后，生成了（*，G）表项。

(2) 组播源发送给接收者的组播数据报文，会通过负载分担的方式分别到达 Device A 和 Device B 的 M-LAG 接口。Device A 和 Device B 会通过 peer-link 链路将各自接收到的报文发送给对端，从而使两台设备上都能收到完整的组播流量。Device A 上根据收到的组播数据流量建立（S，G）表项。

2. 正常情况下组播流量转发过程

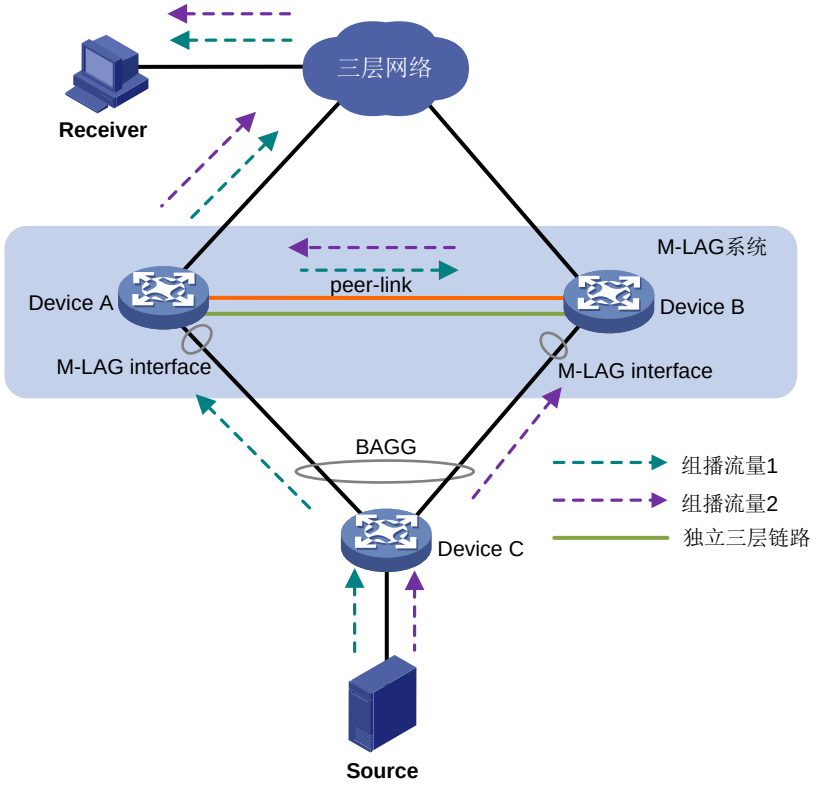
M-LAG 设备正常工作时，组播流量转发过程如图 46 所示。

(1) 组播源发送的组播流量，通过负载分担方式到达 Device A 和 Device B。

(2) Device A 和 Device B，通过 peer-link 链路互相发送接收到的组播数据流量。这样，保证 Device A 和 Device B 上都能收到完整的组播数据流量。

(3) Device A 将收到的完整的组播数据流量发送给下游接收者。

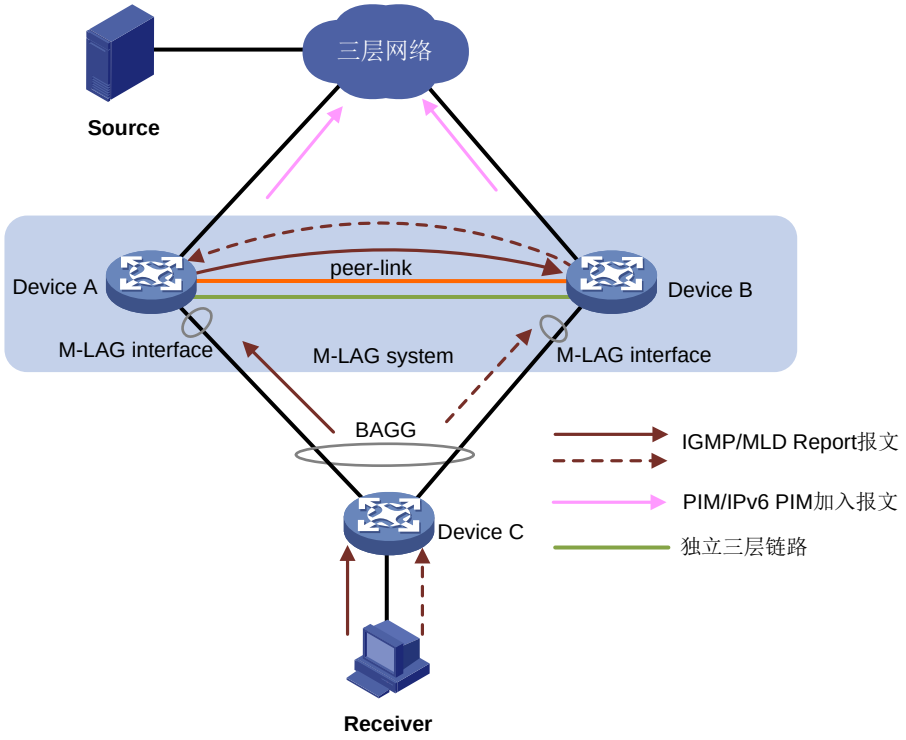
图46 组播流量转发过程（组播源连接 M-LAG 系统）



9.2.2 组播接收者连接 M-LAG 系统

如图 47 所示，作为 M-LAG 设备的 Device A 和 Device B 通过 peer-link 链路连接，Device C 与组播接收者相连。Device A 和 Device B 上与 Device C 相连的 M-LAG 接口上，均需要配置 PIM/IPv6 PIM 消极模式，保证 Device A 和 Device B 均能收到组播源发送的所有的组播数据流量。

图47 三层组播转发表项建立过程（组播接收者连接 M-LAG 系统）



(1) Device A 或者 Device B 中的任意一台设备，收到下游设备 Device C 发送的 IGMP/MLD Report 报文。Device A 或者 Device B 从 Report 报文中解析出主机要加入的组播组地址，分别生成（*，G1）和（*，G2）三层组播转发表项，将 M-LAG 接口添加到出端口列表中。

(2) Device A 和 Device B 分别将 Report 报文通过 peer-link 链路透传给对端设备。以 Device A 为例，Device A 收到 Device B 发送的 Report 报文后，在生成三层组播转发表项（*，G2），同时将 Device A 上的 M-LAG 接口添加到出端口列表中。Device B 的处理过程类似，将在本地生成（*，G1）的组播转发表项。

(3) Device A 和 Device B 上的组播组信息将保持同步，生成相同的（*，G1）和（*，G2）组播转发表项。

(4) Device A 和 Device B 根据网络中配置 PIM 模式工作机制，最终在设备上生成相同的 PIM 路由表项。此处，以网络配置的 PIM 模式为 PIM-SM 示例。

通过上述机制使得 Device A 和 Device B 上生成的 PIM 路由表项保持一致，形成设备级别的备份，当一台成员设备发生故障（设备故障、上下行链路故障等）时，组播流量可以由另一台成员设备进行转发，从而避免组播流量转发中断。

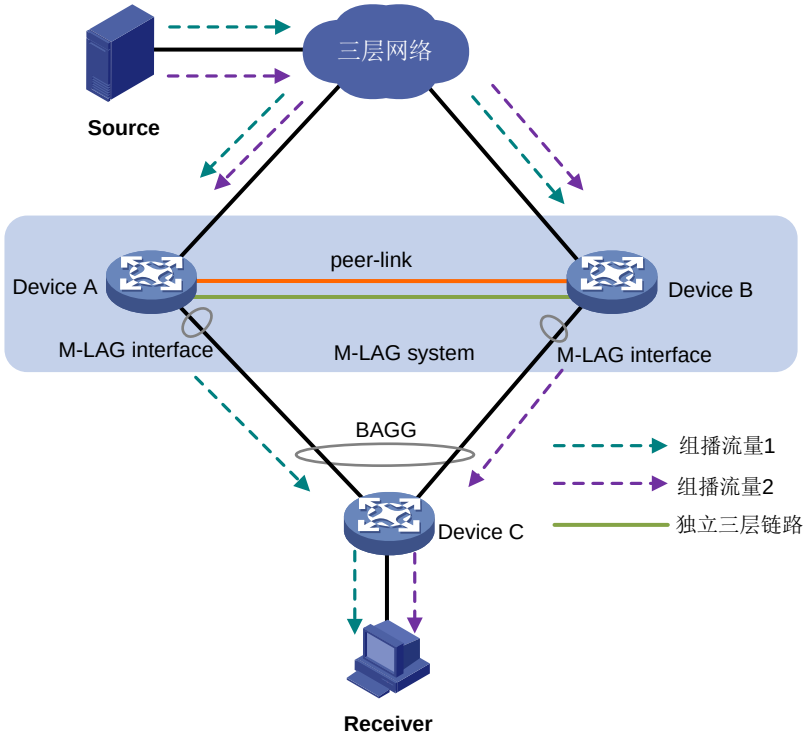
2. 正常情况下组播流量转发过程

M-LAG 设备正常工作时，组播流量转发过程如图 48 所示。

(1) 组播源发送的组播流量，通过三层网络分别到达 Device A 和 Device B。

(2) Device A 和 Device B 向下游转发组播流量时，采用奇偶原则对组播流量进行负载分担。M-LAG 系统编号为奇数的成员设备转发组播组地址为奇数的流量，M-LAG 系统编号为偶数的成员设备转发组播组地址为偶数的流量。

图48 组播流量转发过程（组播接收者连接 M-LAG 系统）



10 EVPN VXLAN 支持 M-LAG

10.1 功能简介

EVPN（Ethernet Virtual Private Network，以太网虚拟专用网络）VXLAN 采用 M-LAG 技术将两台物理设备连接起来虚拟成一台设备，使用该虚拟设备作为 VTEP（既可以是仅用于二层转发的 VTEP，也可以是 EVPN 网关），可以避免 VTEP 单点故障对网络造成影响，从而提高 EVPN 网络的可靠性。

提示

目前，本功能仅支持站点网络和 Underlay 网络同为 IPv4 网络，或站点网络和 Underlay 网络同为 IPv6 网络。

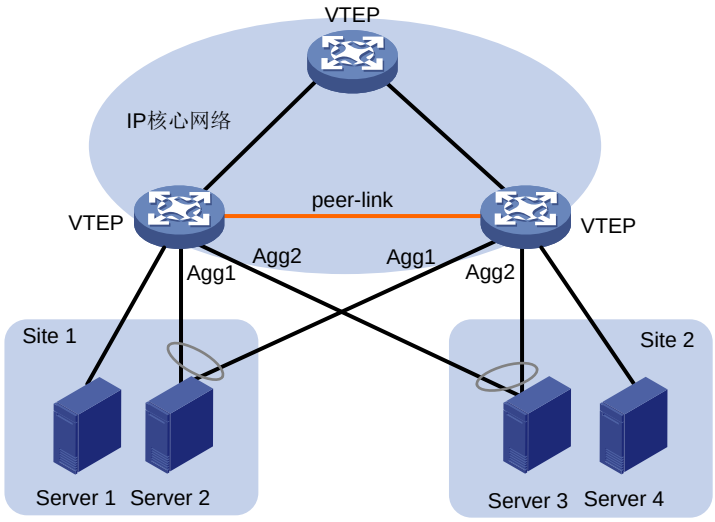
10.2 典型组网

EVPN VXLAN 支持 M-LAG 的典型组网如图 49 所示。在该组网中：

- 两台 VTEP 组成 M-LAG 系统，它们具有相同的虚拟 VTEP 地址，对外表现为一台虚拟 VTEP 设备。
- M-LAG 设备间的 peer-link 连接既可以是以太网聚合链路，也可以是 VXLAN 隧道。以太网链路作为 peer-link 连接的组网，称为有 peer link 组网；VXLAN 隧道作为 peer-link 连接的组网，称为无 peer link 组网，作为 peer-link 的 VXLAN 隧道自动与设备上的所有 VXLAN 关联。

- Server 2 和 Server 3 通过 M-LAG 方式接入 VTEP。
- Server 1 和 Server 4 单挂接入 VTEP。

图49 EVPN VXLAN 支持 M-LAG 典型组网图



10.3 同步MAC地址和ARP/ND信息

作为 M-LAG 设备的两台 VTEP 通过 peer-link 连接，在 peer-link 上同步 MAC 地址、ARP/ND 表项和 ARP/ND 泛洪抑制表项等，以确保两台 VTEP 上的 MAC 地址和 ARP/ND 信息保持一致。当一台 VTEP 故障时，另一台 VTEP 可以快速接替其工作，转发流量。

M-LAG 接口的表项通过 peer-link 同步到对端 M-LAG 设备上后，对端 M-LAG 设备将该表项添加到本地对应的 M-LAG 接口；单挂接口的表项通过 peer-link 同步到对端 M-LAG 设备上后，对端 M-LAG 设备将该表项添加到 peer-link 接口。

10.4 VXLAN隧道建立

在 EVPN VXLAN 组网中，VTEP 之间可以根据 BGP EVPN 的 IMET（Inclusive Multicast Ethernet Tag Route，包含性组播以太网标签路由）路由、MAC/IP 发布路由和 IP 前缀路由建立 VXLAN 隧道。

10.5 备份用户侧链路

在用户侧，两台 VTEP 均通过以太网链路接入同一台虚拟机，跨设备在两条链路间建立二层聚合接口，将该聚合接口配置为 AC（在聚合接口上创建以太网服务实例、配置报文匹配规则并关联以太网服务实例与 VSI），从而避免单条以太网链路故障导致虚拟机无法访问网络。

10.5.1 peer-link 为以太网聚合链路时的用户侧链路备份机制

peer-link 为以太网聚合链路时，VTEP 通过在 peer-link 上自动创建 AC 或自动创建 VXLAN 隧道来实现用户侧链路备份。

- 在 peer-link 上自动创建 AC

  在 peer-link 上，M-LAG 设备会根据用户侧 AC 或用户所属的 VXLAN ID 自动创建 AC。

  通过自动创建的 AC 实现用户侧链路备份的过程为：当一台 VTEP 上的 AC 故障后，从 VXLAN 隧道上接收到的、发送给该 AC 的报文将通过 peer-link 转发到另一台 VTEP，该 VTEP 根据 peer-link 上配置的 AC 判断报文所属 VSI，并转发该报文，从而保证转发不中断。

- 自动创建 VXLAN 隧道

  作为 M-LAG 设备的两台 VTEP 之间自动建立 VXLAN 隧道，并将该 VXLAN 隧道自动与所有 VXLAN 关联。

  通过自动创建的 VXLAN 隧道实现用户侧链路备份的过程为：如果一台 VTEP 上的 AC 故障，则该 VTEP 从 VXLAN 隧道上接收到远端 VTEP（非 M-LAG 设备）发送给故障 AC 的报文后，为报文添加 VXLAN 封装，封装的 VXLAN ID 为故障 AC 所属 VSI 对应的 VXLAN ID，并通过自动创建的 VXLAN 隧道将其转发到另一台 VTEP（M-LAG 设备）。该 VTEP 根据 VXLAN ID 判断报文所属的 VSI，并转发该报文。

10.5.2 peer-link 为 VXLAN 隧道时的用户侧链路备份机制

peer-link 为 VXLAN 隧道时，用户侧链路备份机制为：如果一台 VTEP 上的 AC 故障，则该 VTEP 从 VXLAN 隧道上接收到发送给故障 AC 的报文后，为报文添加 VXLAN 封装，封装的 VXLAN ID 为故障 AC 所属 VSI 对应的 VXLAN ID，并通过作为 peer-link 的 VXLAN 隧道将其转发到另一台 VTEP。该 VTEP 根据 VXLAN ID 判断报文所属的 VSI，并转发该报文。

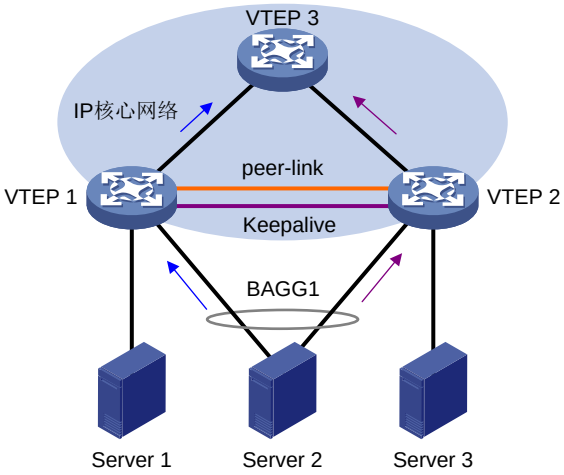
10.6 流量转发

10.6.1 M-LAG 接口的单播流量转发

1. M-LAG 接口的上行单播流量

如图 50 所示，接入服务器（Server 2）通过聚合接口采用 M-LAG 方式接入 VTEP 时，利用聚合接口的负载分担功能，接入服务器将发送到外网的上行单播流量负载分担到多台 M-LAG 设备（VTEP 1 和 VTEP 2）。M-LAG 设备接收到单播流量后，按照本地转发优先原则，通过本地的 VXLAN 隧道将单播流量转发给远端 VTEP。

图50 M-LAG 接口的上行单播流量转发

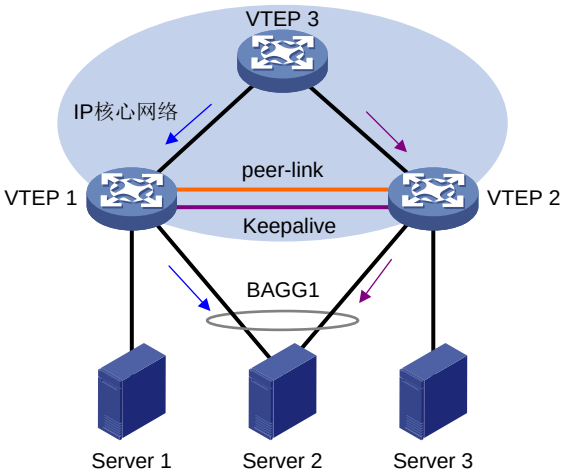


2. M-LAG 接口的下行单播流量

如图 51 所示，接入服务器（Server 2）通过聚合接口采用 M-LAG 方式接入 VTEP。外网向 Server 2 发送单播流量时，该流量通过 VXLAN 隧道转发给 M-LAG 设备（VTEP 1 和 VTEP 2）。Underlay 网络中，M-LAG 设备均发布到达虚拟 VTEP 地址的路由，以便在 VTEP 3 上形成等价路由。从而，使得外网发送给 Server 2 的单播流量负载分担到多台 M-LAG 设备。

M-LAG 设备接收到下行的单播流量后，按照本地转发优先原则，通过本地 AC 将单播流量转发给接入服务器（Server 2）。

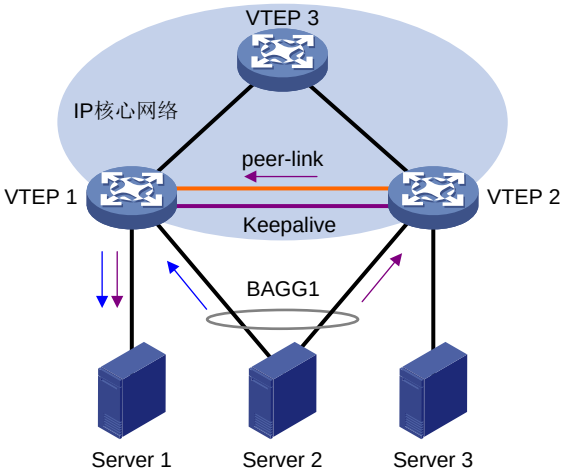
图51 M-LAG 接口的下行单播流量转发



3. M-LAG 接口发往单挂接口的单播流量

如图 52 所示，接入服务器（Server 2）通过聚合接口采用 M-LAG 方式接入 VTEP 时，利用聚合接口的负载分担功能，Server 2 将发往单挂接口的单播流量负载分担到多台 M-LAG 设备（VTEP 1 和 VTEP 2）。单挂接口所在的 M-LAG 设备（VTEP 1）通过查找本地表项，将单播流量转发到单挂接口；其他 M-LAG 设备（VTEP 2）接收到单播流量后，通过 peer-link 将流量转发给 VTEP 1，再由 VTEP 1 转发到单挂接口。

图52 M-LAG 接口发往单挂接口的单播流量转发

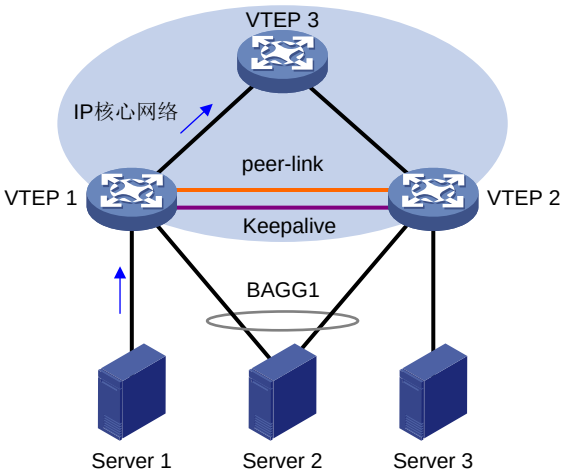


10.6.2 单挂接口的单播流量转发

1. 单挂接口的上行单播流量

如图 53 所示，单挂接入服务器（Server 1）发送的单播流量到达 VTEP 1 后，VTEP 1 通过本地的 VXLAN 隧道将单播流量转发给远端 VTEP。

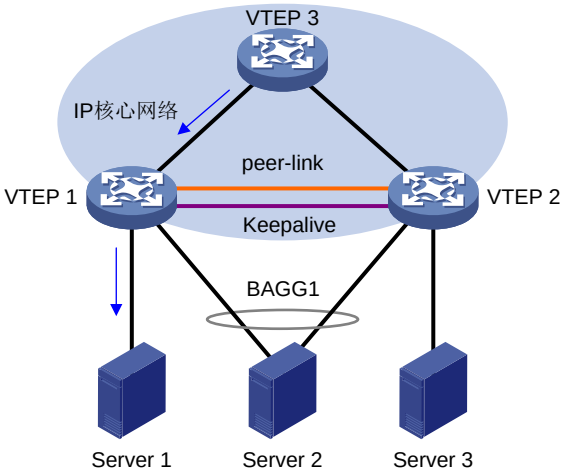
图53 单挂接口的上行单播流量转发



2. 单挂接口的下行单播流量

如图 54 所示，外网向单挂接入的服务器（Server 1）发送单播流量时，该流量会通过 VXLAN 隧道转发给 VTEP 1，VTEP 1 接收到流量后，直接将其转发到单挂接口。该流量不会发送给 VTEP 2，从而避免流量绕行 peer-link 的问题。

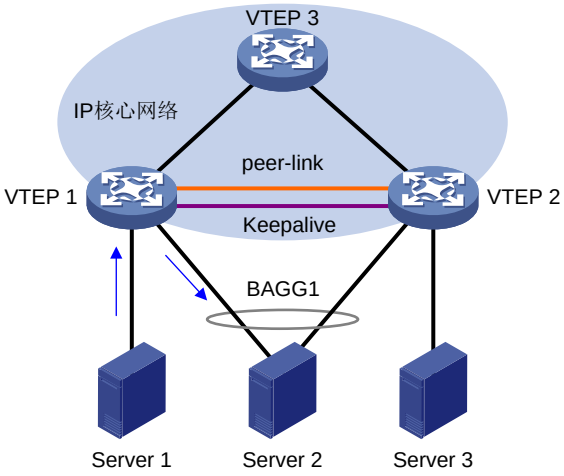
图54 单挂接口的下行单播流量转发



3. 单挂接口发往 M-LAG 接口的单播流量

如图 55 所示，单挂接入服务器（Server 1）发送给 M-LAG 接入服务器（Server 2）的单播流量到达 VTEP 1 后，VTEP 1 按照本地转发优先原则，通过本地 AC 将单播流量转发给接入服务器（Server 2）。

图55 单挂接口发往 M-LAG 接口的单播流量转发



4. 单挂接口互通的单播流量

当互通的单挂接口连接到同一台 M-LAG 设备时，流量转发过程与 EVPN VXLAN 的流量转发过程相同。

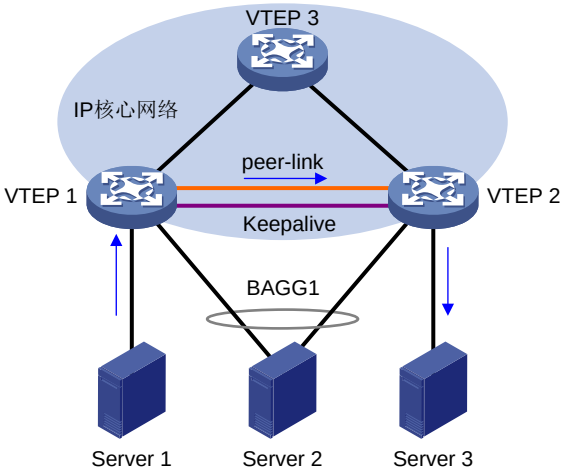
如图 56 所示，当互通的单挂接口连接到不同的 M-LAG 设备时，通过 peer-link 实现单挂接口的互通：

- peer-link 为以太网聚合链路时，单挂接口互通机制为：在单挂接口上创建单挂 AC 后，M-LAG 设备会在 peer-link 上自动创建对应的 AC，并将其与 VSI 关联。当一台 M-LAG 设备从单挂

AC 上收到报文后，报文将通过 peer-link 转发到另一台 M-LAG 设备，另一台 M-LAG 设备根据 peer-link 上自动创建的 AC 判断报文所属 VSI，并转发该报文。

- peer-link 为 VXLAN 隧道时，单挂接口互通机制为：当一台 M-LAG 设备从单挂 AC 上收到报文后，为报文添加 VXLAN 封装，封装的 VXLAN ID 为单挂 AC 所属 VSI 对应的 VXLAN ID，并通过作为 peer-link 的 VXLAN 隧道将其转发到另一台 M-LAG 设备。另一台 M-LAG 设备根据 VXLAN ID 判断报文所属的 VSI，并转发该报文。

图56 单挂接口互通的单播流量转发

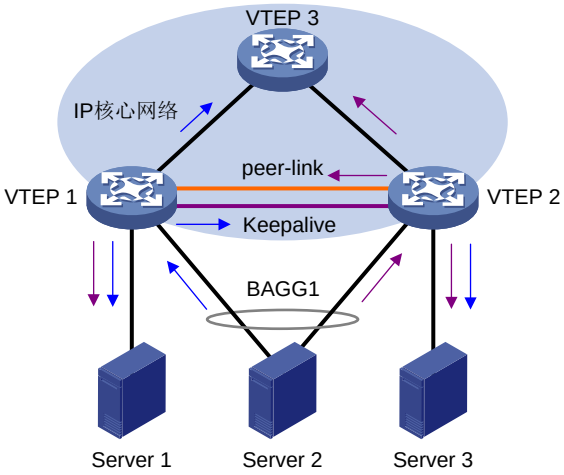


10.6.3 BUM 流量转发

1. M-LAG 接口的上行 BUM 流量

如图 57 所示，接入服务器（Server 2）通过聚合接口采用 M-LAG 方式接入 VTEP 时，利用聚合接口的负载分担功能，接入服务器将发送到外网的上行 BUM（Broadcast/Unknown unicast/Unknown Multicast，广播/未知单播/未知组播）流量负载分担到多台 M-LAG 设备（VTEP 1 和 VTEP 2）。M-LAG 设备接收到 BUM 流量后，判断 BUM 流量所属的 VSI，通过该 VSI 内除接收 AC 外的所有本地 AC（包括单挂 AC）、VXLAN 隧道和 peer-link 转发该流量。M-LAG 设备（即 VTEP）从 peer-link 上接收到 BUM 流量后，仅将该流量转发给本地的单挂 AC。

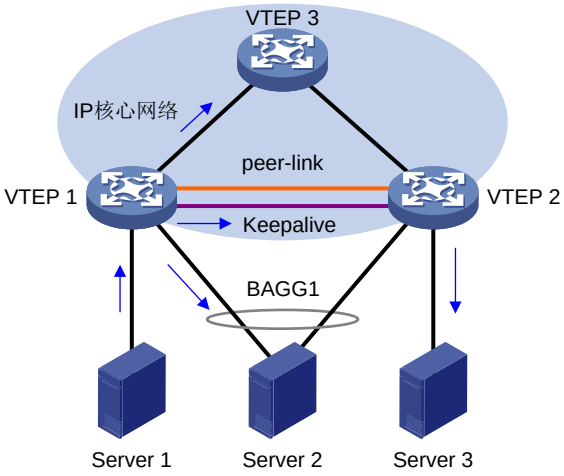
图57 M-LAG 接口的上行 BUM 流量转发



2. 单挂接口的上行 BUM 流量

如图 58 所示，单挂接入服务器（Server 1）发送的 BUM 流量到达 VTEP 1 后，VTEP 1 判断流量所属的 VSI，通过该 VSI 内除接收 AC 外的所有本地 AC、VXLAN 隧道和 peer-link 转发该流量。VTEP 2 从 peer-link 收到该流量后，只将其转发给本地单挂 AC。

图58 单挂接口的上行 BUM 流量转发

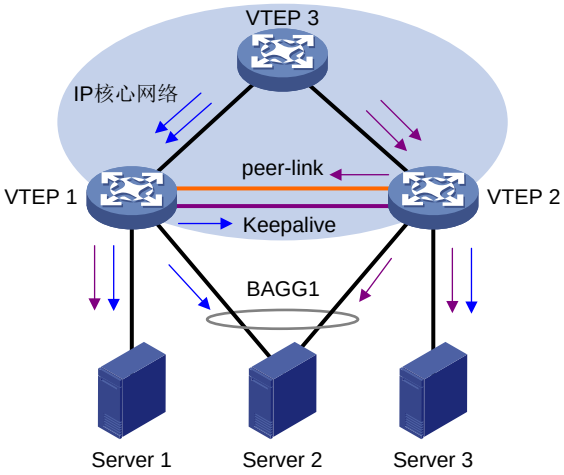


3. 网络侧的下行 BUM 流量

如图 59 所示，接入服务器（Server 2）通过聚合接口采用 M-LAG 方式接入 VTEP，Server 1 和 Server 3 分别单挂接入 VTEP 1 和 VTEP 2。外网向 Server 所在的内网发送 BUM 流量时，VTEP 3 通过 VXLAN 隧道将流量转发给 VTEP 1 和 VTEP 2。

M-LAG 设备接收到 BUM 流量后，判断流量所属的 VSI，并在该 VSI 内的所有 AC（包括 M-LAG 接口对应的 AC 和单挂 AC）、peer-link 上转发该流量。M-LAG 设备从 peer-link 上接收到 BUM 流量后，仅将该流量转发给本地的单挂 AC。

图59 M-LAG 接口的下行 BUM 流量转发



10.7 故障处理机制

在 EVPN VXLAN 支持 M-LAG 组网中，peer-link 故障、M-LAG 设备故障时的故障处理机制与 M-LAG 组网中的故障处理机制相同，详细介绍请参见“2.5 M-LAG 故障处理机制”。本节仅介绍下行链路故障、上行链路故障、peer-link 和 Keepalive 链路同时故障时的故障处理机制。

10.7.1 下行链路故障处理机制

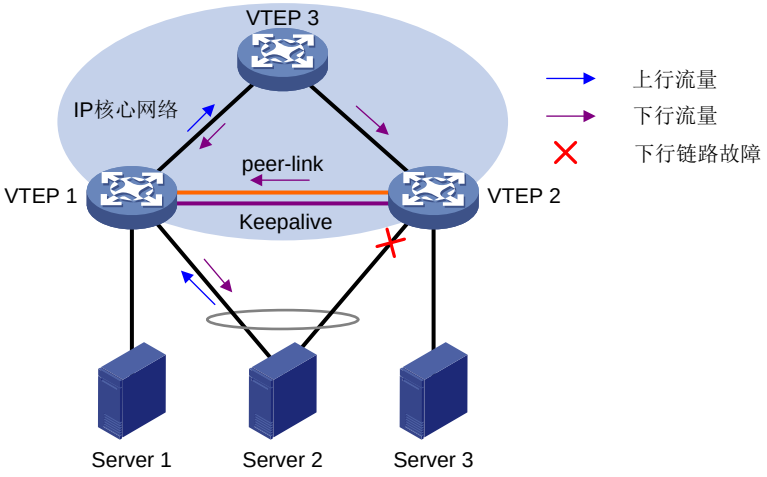
1. peer-link 为以太网聚合链路

如图 60 所示，peer-link 为以太网聚合链路的 EVPN VXLAN 组网中，某台 M-LAG 设备（VTEP 2）的下行链路故障时，上下行流量的处理方式分别为：

- 上行流量会通过未故障的链路发送给另一台 M-LAG 设备（VTEP 1），上行流量均通过 VTEP 1 转发。
- 下行流量转发过程为：
  a. 由于网络侧感知不到下行链路故障，流量依然会发送给所有 M-LAG 设备。
  b. VTEP 2 收到网络侧访问 Server 2 的流量后，会通过 peer-link 上自动创建的 AC 转发到 VTEP 1，VTEP 1 根据 peer-link 上自动创建的 AC 判断报文所属 VSI，并将该报文转发给 Server 2。

故障恢复后，VTEP 2 上的 AC up，流量正常转发。

图60 下行链路故障处理机制（peer-link 为以太网聚合链路）



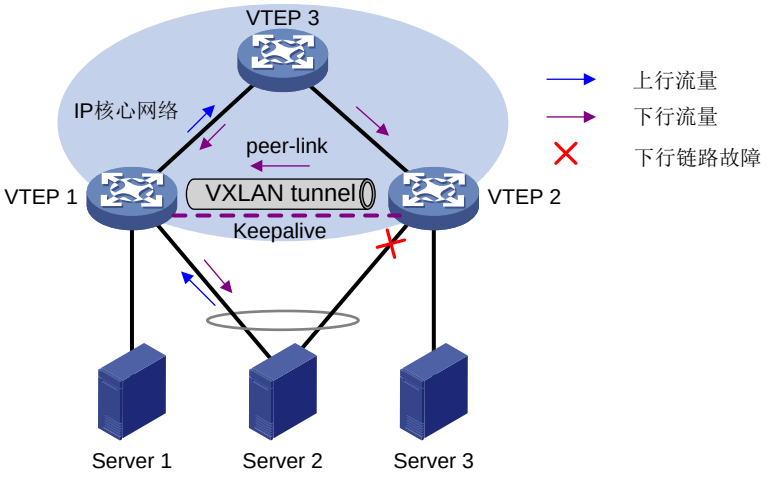
2. peer-link 为 VXLAN 隧道

如图 61 所示，peer-link 为 VXLAN 隧道的 EVPN VXLAN 组网中，某台 M-LAG 设备（VTEP 2）的下行链路故障时，上下行流量的处理方式分别为：

- 上行流量会通过未故障的链路发送给另一台 M-LAG 设备（VTEP 1），上行流量均通过 VTEP 1 转发。
- 下行流量转发过程为：
  a. 由于网络侧感知不到下行链路故障，流量依然会发送给所有 M-LAG 设备。
  b. VTEP 2 收到网络侧访问 Server 2 的流量后，为报文添加 VXLAN 封装（封装的 VXLAN ID 为故障 AC 所属 VSI 对应的 VXLAN ID），然后通过作为 peer-link 的 VXLAN 隧道将其转发到 VTEP 1。VTEP 1 根据接收到的报文中携带的 VXLAN ID 字段判断报文所属 VSI，并将该报文转发给 Server 2。

故障恢复后，VTEP 2 的 AC up，流量正常转发。

图61 下行链路故障处理机制（peer-link 为 VXLAN 隧道）



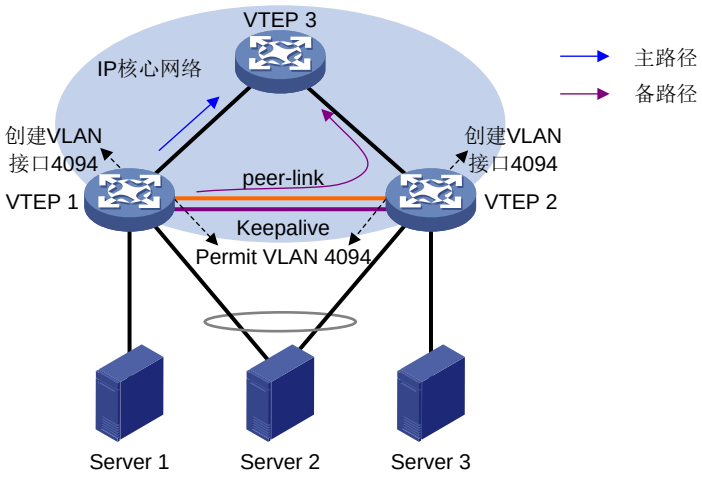
10.7.2 上行链路故障处理机制

1. peer-link 为以太网聚合链路

peer-link 为以太网聚合链路的 EVPN VXLAN 组网中，建议将 peer-link 部署为逃生链路。部署逃生链路是指允许 peer-link 转发三层流量，并在 peer-link 上运行路由协议，使得 M-LAG 设备可以通过 peer-link 与远端 VTEP 三层互通。在 underlay 网络上，peer-link 所在的路径作为 M-LAG 设备与远端 VTEP 之间 VXLAN 隧道的备份路径。当 M-LAG 设备的上行链路故障，导致 VXLAN 隧道的 underlay 主路径故障时，VXLAN 隧道保持 up 状态，通过 peer-link 所在的 underlay 备份路径转发流量。

如图 62 所示，部署逃生链路的方式为在 peer-link 上允许某个 VLAN 通过，在 M-LAG 设备上创建该 VLAN 对应的 VLAN 接口，并在 VLAN 接口上运行路由协议，使得该接口与远端 VTEP 三层互通。部署逃生链路后，VTEP 1 和 VTEP 3 之间的 VXLAN 隧道在 underlay 网络上具有主备两条路径，经由 peer-link 的路径作为备份路径。推荐使用 VLAN 4094 来部署逃生链路。

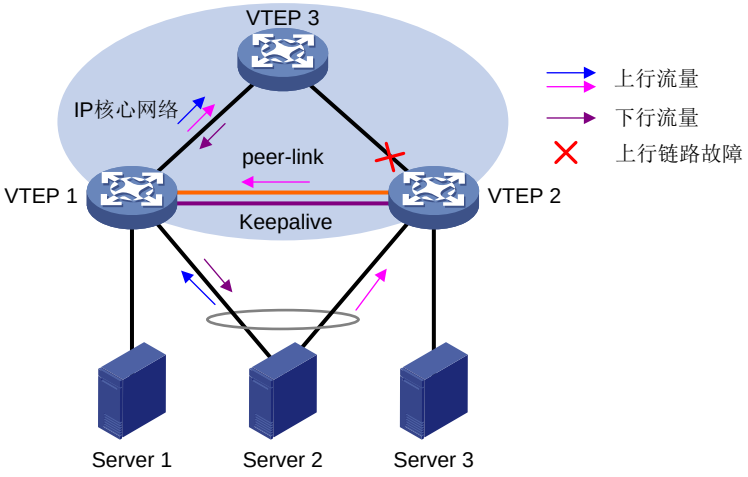
图62 逃生链路部署示意图



如图 63 所示，当某台 M-LAG 设备（VTEP 2）的上行链路故障时，M-LAG 接口上下行流量的处理方式分别为：

- 上行流量：接入服务器（Server 2）通过聚合接口采用 M-LAG 方式接入 VTEP 时，利用聚合接口的负载分担功能，接入服务器将发送到外网的上行单播流量负载分担到多台 M-LAG 设备（VTEP 1 和 VTEP 2）。
  - VTEP 1 接收到上行流量后，正常转发该流量。
  - 由于部署了逃生链路，上行链路故障的 M-LAG 设备（VTEP 2）与远端 VTEP 之间的 VXLAN 隧道仍然 up。该 VXLAN 隧道对应的 underlay 路径为经由 peer-link 的路径。因此，VTEP 2 接收到上行流量后，对其进行 VXLAN 封装，并通过 peer-link 绕行另一台 M-LAG 设备 VTEP 1，将流量发送给远端 VTEP。
- 下行流量：远端 VTEP 通过 VXLAN 隧道将下行流量发送给 M-LAG 设备。由于 VTEP 2 的上行链路故障，下行流量只会发送给 VTEP 1，再由 VTEP 1 将流量转发给 Server 2。

图63 上行链路故障处理机制（peer-link 为以太网聚合链路）



当某台 M-LAG 设备的上行链路故障时，该 M-LAG 设备上单挂接口的上下行流量将通过 peer-link 发送给另一台 M-LAG 设备，由另一台 M-LAG 设备进行转发。

2. peer-link 为 VXLAN 隧道

peer-link 为 VXLAN 隧道的 EVPN VXLAN 组网中，某台 M-LAG 设备的上行链路故障时，作为 peer-link 的 VXLAN 隧道也会 down，此时的故障处理机制为：

- 如果 M-LAG 设备之间没有 Keepalive 链路，则 M-LAG 系统会分裂，两台 M-LAG 设备均使用实际 IP 地址与远端 VTEP 建立 VXLAN 隧道，两台 M-LAG 设备均可以转发流量。
- 如果 M-LAG 设备之间存在 Keepalive 链路，则备设备上的接口会 MAD down，仅主设备使用虚拟 VTEP 地址与远端 VTEP 建立 VXLAN 隧道，仅主设备可以转发流量。

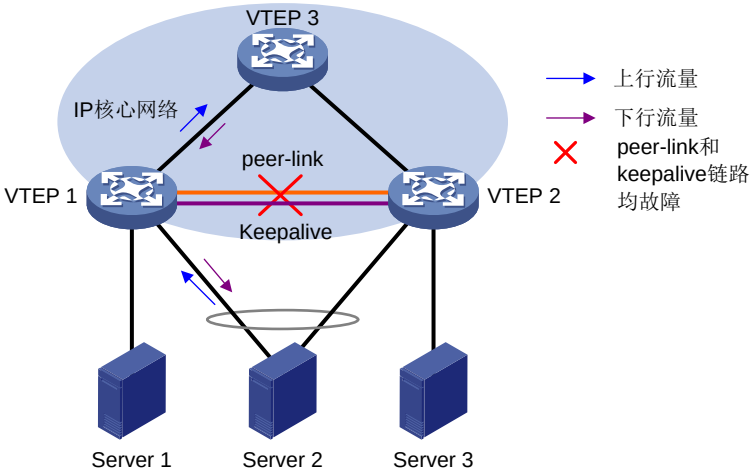
10.7.3 peer-link 和 Keepalive 链路同时故障

1. peer-link 为以太网聚合链路

peer-link 为以太网聚合链路的 EVPN VXLAN 组网中，peer-link 和 Keepalive 链路同时故障时，M-LAG 系统会分裂，两台 M-LAG 设备均使用实际 IP 地址与远端 VTEP 建立 VXLAN 隧道，两台 M-LAG 设备均可以转发流量。

如图 64 所示，peer-link 和 Keepalive 链路同时故障时，M-LAG 接口的上下行流量转发方式为：Server 2 根据 LACP 优先级选择将上行流量发送给一台 M-LAG 设备（如 VTEP 1），下行流量也通过目的 IP 为实际 IP 地址的 VXLAN 隧道发送给 VTEP 1。即，M-LAG 接口的上下行流量均通过 VTEP 1 转发。

图64 peer-link 和 Keepalive 链路同时故障（peer-link 为以太网聚合链路）



peer-link 和 Keepalive 链路同时故障时，单挂接口的流量转发不受影响。

2. peer-link 为 VXLAN 隧道

peer-link 为 VXLAN 隧道的 EVPN VXLAN 组网中，peer-link 和 Keepalive 链路同时故障时的故障处理机制与 peer-link 为以太网聚合链路的 EVPN VXLAN 组网相同。

11 VXLAN 支持 M-LAG

VXLAN（Virtual eXtensible LAN，可扩展虚拟局域网络）采用 M-LAG 技术将两台物理设备连接起来虚拟成一台设备，使用该虚拟设备作为 VTEP（既可以是仅用于二层转发的 VTEP，也可以是 VXLAN IP 网关），可以避免 VTEP 单点故障对网络造成影响，从而提高 VXLAN 网络的可靠性。

提示

- 目前，本功能仅支持站点网络和 Underlay 网络同为 IPv4 网络，或站点网络和 Underlay 网络同为 IPv6 网络。
- 集中式 VXLAN IP 网关保护组不支持 M-LAG 功能。

VXLAN 支持 M-LAG 的工作机制与 EVPN VXLAN 支持 M-LAG 的工作机制基本相同，此章节不再赘述，仅介绍两种工作机制的差异点。

11.1 同步MAC地址和ARP/ND信息

VXLAN 支持 M-LAG 组网中 M-LAG 接口和单挂接口的表项同步与 EVPN VXLAN 支持 M-LAG 组网的中表项同步工作机制相同。

对于 VXLAN 隧道上通过动态/静态方式学习到的 MAC 地址、ARP/ND 表项和 ARP/ND 泛洪抑制表项，需要通过 peer-link 同步到对端 M-LAG 设备上，对端 M-LAG 设备根据同步的表项信息中的 VXLAN ID、隧道的源端地址和目的地址，在本地相同 VXLAN ID 下查找是否存在相同源端地址和

目的地址的 VXLAN 隧道。若存在相同的 VXLAN 隧道，则将表项添加到该 VXLAN 隧道接口；否则，不添加表项。

11.2 VXLAN隧道建立

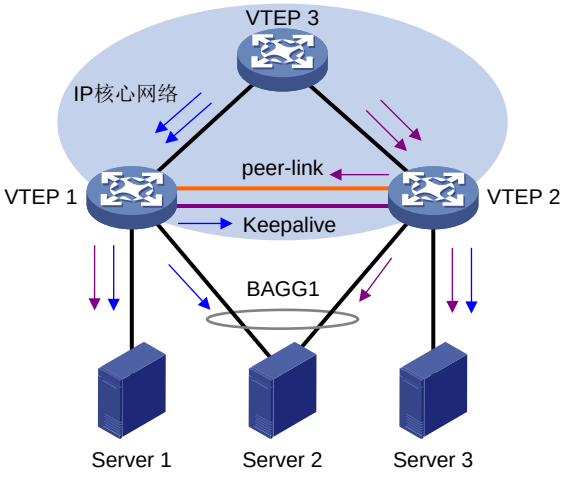
在 VXLAN 组网中，VTEP 之间通过手工方式创建 VXLAN 隧道，即手工指定隧道的源端地址和目的端地址需要分别手工指定为本地和远端 VTEP 的接口地址。组成 M-LAG 系统的两台 VTEP 设备上，需要配置一个相同的 IP 地址作为虚拟 VTEP 的地址，并采用虚拟 VTEP 地址作为本地 VXLAN 隧道的源端地址与远端 VTEP 建立隧道。

在 peer-link 为 VXLAN 隧道的 VXLAN 组网中，还需要在组成 M-LAG 系统的两台 VTEP 设备之间手工创建一条 VXLAN 隧道。

11.3 网络侧的下行BUM流量

如图 59 所示，接入服务器（Server 2）通过聚合接口采用 M-LAG 方式接入 VTEP，Server 1 和 Server 3 分别单挂接入 VTEP 1 和 VTEP 2。外网向 Server 所在的内网发送 BUM 流量时，VTEP 3 将通过目的 IP 地址为虚拟 VTEP 地址的 VXLAN 隧道将流量转发给 M-LAG 设备（VTEP 1 和 VTEP 2）。由于 VTEP 1 和 VTEP 2 共用虚拟 VTEP 地址，因此通过该 VXLAN 隧道转发的流量会负载分担到 VTEP 1 和 VTEP 2。VTEP 1 和 VTEP 2 判断流量所属的 VSI，并在该 VSI 内的所有 AC（包括 M-LAG 接口对应的 AC 和单挂 AC）、peer-link 上转发该流量。M-LAG 设备从 peer-link 上接收到 BUM 流量后，仅将该流量转发给本地的单挂 AC。

图65 M-LAG 接口的下行 BUM 流量转发



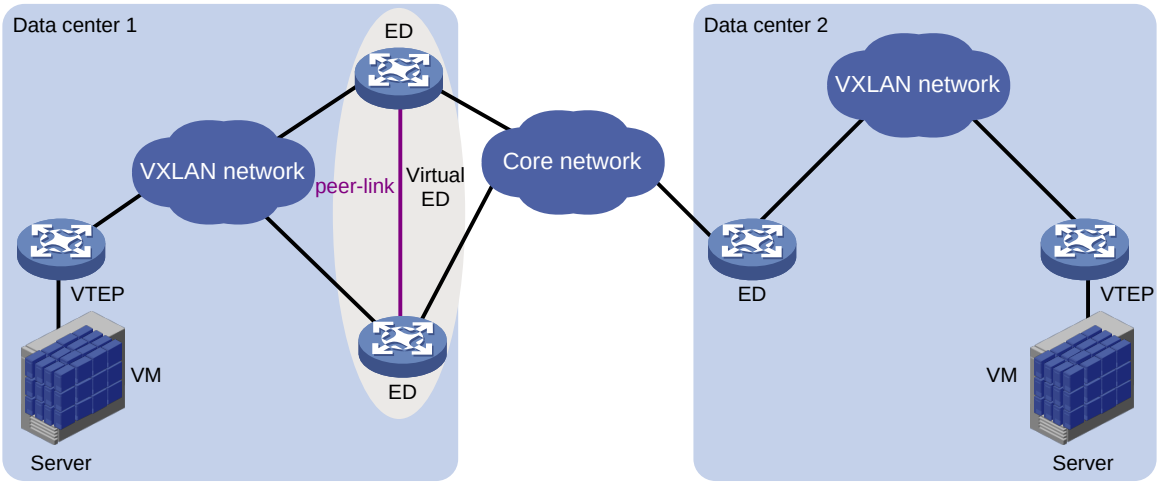
12 EVPN 数据中心互联支持 M-LAG

说明

目前，本功能仅支持站点网络和 Underlay 网络同为 IPv4 网络，或站点网络和 Underlay 网络同为 IPv6 网络。

如图 66 所示，EVPN 数据中心互联组网中，利用 M-LAG 将两台物理设备连接起来虚拟成一台设备，使用该虚拟设备作为 ED，可以避免 ED 单点故障对网络造成影响，从而提高 EVPN 网络的可靠性。

图66 EVPN 数据中心互联支持 M-LAG 示意图



EVPN 数据中心互联支持 M-LAG 的工作机制与 EVPN VXLAN 支持 M-LAG 相同，详细介绍请参见“10 EVPN VXLAN 支持 M-LAG”。

13 组播 VXLAN 支持 M-LAG

说明

仅 EVPN VXLAN 组网中的 MDT 模式组播 VXLAN 支持 M-LAG。VXLAN 组网中的入方向复制模式组播 VXLAN 不支持 M-LAG。

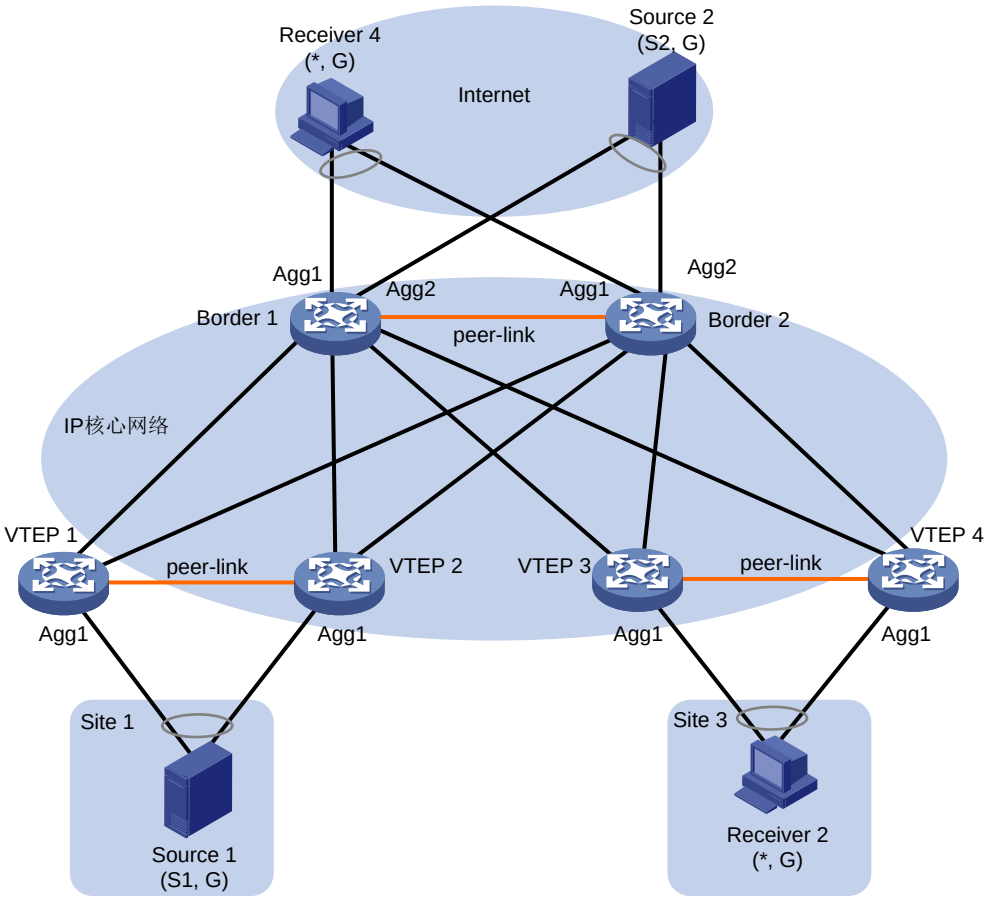
13.1 组播VXLAN支持M-LAG工作机制概述

13.1.1 典型组网

组播 VXLAN 利用 M-LAG 将两台物理设备连接起来虚拟成一台设备，避免设备单点故障对网络造成影响，从而提高组播 VXLAN 网络的可靠性。在组播 VXLAN 组网中，VTEP 和 Border 设备均支持 M-LAG，且作为 M-LAG 设备的 VTEP 和 Border 均可以连接组播源和组播接收者。

EVPN 组播支持 M-LAG 的典型组网如图 67 所示。VTEP 1 和 VTEP 2 组成 M-LAG 系统，VTEP 3 和 VTEP 4 组成 M-LAG 系统，Border 1 和 Border 2 组成 M-LAG 系统。组成 M-LAG 系统的两台 VTEP/Border 具有相同的虚拟地址，对外表现为一台虚拟设备。其他 VTEP/Border 使用该地址与这台虚拟设备自动建立单播 VXLAN 隧道。组播 VXLAN 隧道的源地址也使用该虚拟 VTEP 地址。由于存在 Underlay 网络的组播 RPF 检查机制，设备只会加入到 M-LAG 系统中一台设备的组播 VXLAN 隧道。例如，VTEP 3 和 VTEP 4 在加入组播 VXLAN 隧道时，只会加入到 VTEP 1 和 VTEP 2 中一台 VTEP 的组播 VXLAN 隧道，不会同时加入 VTEP 1 和 VTEP 2 的组播 VXLAN 隧道。

图67 组播 VXLAN 支持 M-LAG 组网



13.1.2 用户侧备份机制

组播 VXLAN 支持 M-LAG 功能通过 peer-link 在组成 M-LAG 系统的成员设备间同步组播流量和组播接收者加入请求（IGMP 成员关系报告报文或者 PIM 加入报文），使成员设备上的组播源和组播接收者信息保持一致，形成设备级备份。当一台成员设备发生故障（设备故障、上下行链路故障等）时，组播流量可以由另一台成员设备进行转发，从而避免组播流量转发中断。

在图 67 所示的组网中，用户侧备份机制为：

- 组播源侧备份：组播源 Source 1 通过 M-LAG 接入后，Source 1 的组播流量会发送到 VTEP 1 和 VTEP 2 中的一台设备。接收到组播流量的 VTEP 通过 peer-link，将组播流量同步到另外一台 VTEP，从而实现 VTEP 1 和 VTEP 2 上都存在组播流量。

- 组播接收者侧备份：组播接收者通过 M-LAG 接入后，组播接收者的加入请求会发送到 VTEP 3 和 VTEP 4 中的一台设备。接收到加入请求的 VTEP 通过 peer-link，将加入请求同步到另外一台 VTEP，从而实现在 VTEP 3 和 VTEP 4 上都建立组播转发表项，表项的出接口为 M-LAG 接口。

13.1.3 组播流量分担

组播接收者侧的 M-LAG 设备接收到组播流量后，采用奇偶原则对组播流量进行负载分担，M-LAG 系统编号为奇数的成员设备转发组播组地址为奇数的流量，M-LAG 系统编号为偶数的成员设备转发组播组地址为偶数的流量。当一台设备发生故障时，另一台设备可以接替其工作，避免流量转发中断。

说明

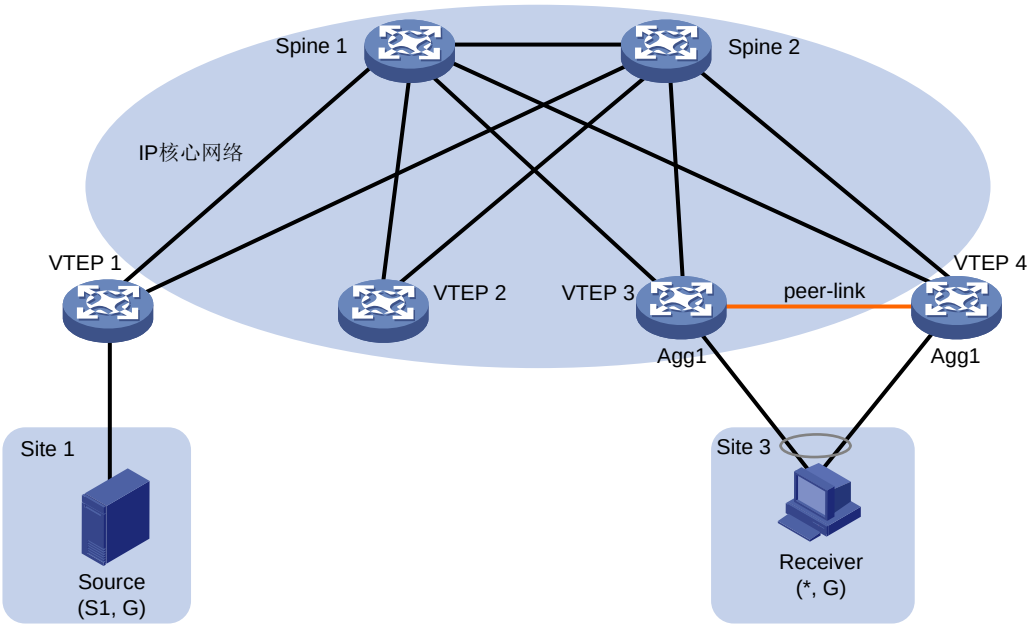
M-LAG 设备上的组播流量奇偶负载分担原则仅针对三层组播转发生效，对二层组播转发不生效。

13.2 二层组播支持M-LAG

二层组播是指组播源和组播接收者位于同一个 VXLAN 网络，组播流量在同一个 VXLAN 网络内根据二层组播转发表项（IGMP snooping、PIM snooping 表项等）进行转发。

如图 68 所示，二层组播支持 M-LAG 组网中，仅支持组播接收者通过 M-LAG 方式接入，不支持组播源通过 M-LAG 方式接入，且流量转发仅支持头端复制方式。

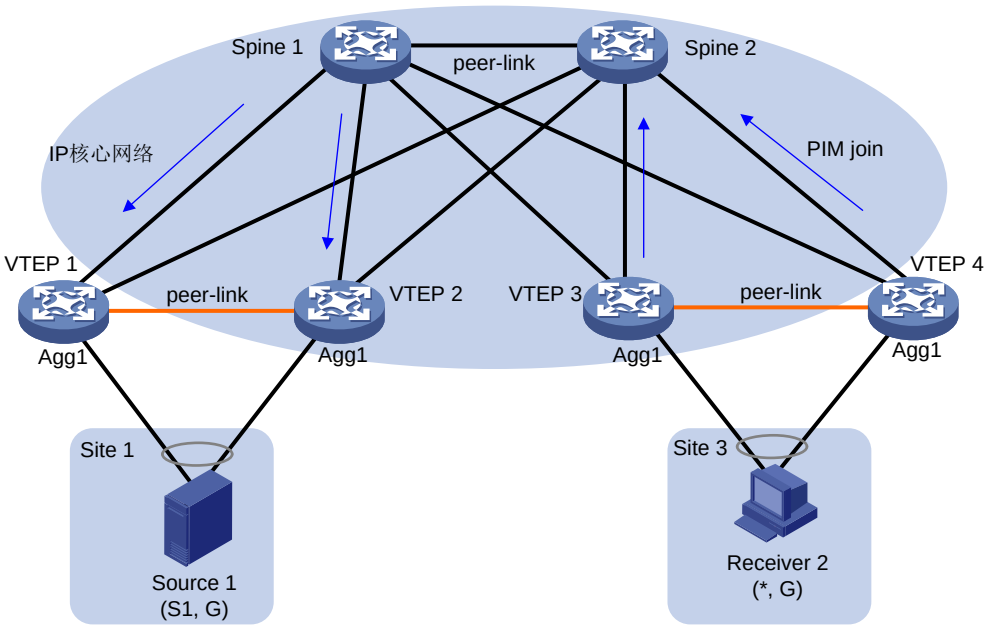
图68 二层组播支持 M-LAG 典型组网



13.3 三层组播支持M-LAG

三层组播是指组播源和组播接收者位于不同的 VXLAN 网络、相同的 VPN，组播流量在同一个 VPN 内跨越 VXLAN 网络，根据三层组播转发表项（IGMP、PIM 表项等）进行转发。如图 69 所示，EVPN VXLAN 三层组播场景中，支持通过 M-LAG 提高网络的可靠性。

图69 三层组播支持 M-LAG 典型组网



13.4 三层组播数据中心互联支持M-LAG

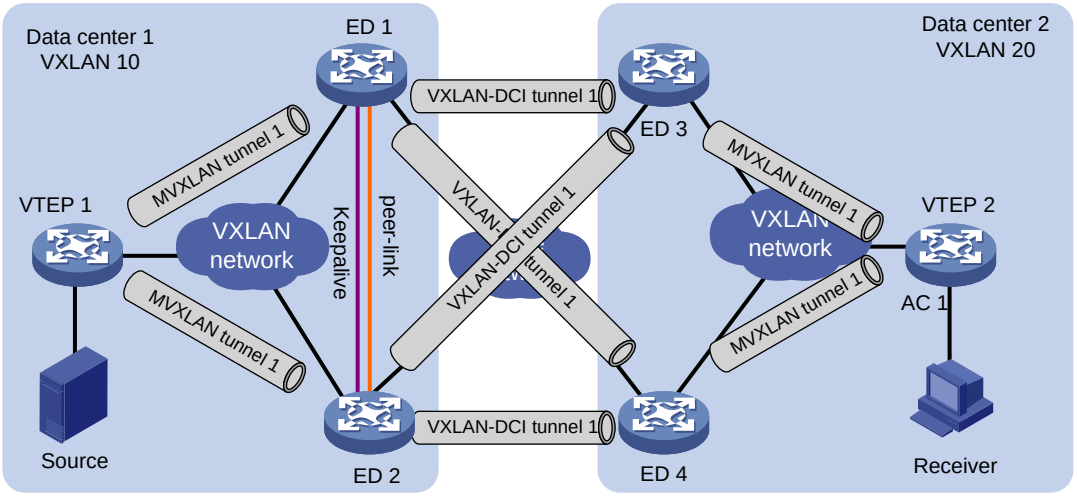
在三层组播 VXLAN 跨数据中心互联场景中，为了提高 ED 的可靠性，避免单点故障，在数据中心的边缘可以部署两台 ED 设备与其他数据中心互联。这两台 ED 设备组成 M-LAG 系统，通过 M-LAG 机制为 ED 提供冗余保护。

如图 70 所示，在组播数据中心互联支持 M-LAG 组网中，DC 内需要建立组播 VXLAN 隧道，ED 之间建立单播 VXLAN-DCI 隧道。组成 M-LAG 系统的 ED 设备具有相同的虚拟 ED 地址，虚拟成一台 ED 设备，使用虚拟 ED 地址与 VTEP、远端 ED 建立隧道，以实现冗余保护和负载分担。

说明

组播数据中心互联支持 M-LAG 组网中，ED 上不存在 M-LAG 接口。

图70 三层组播数据中心互联支持 M-LAG 典型组网

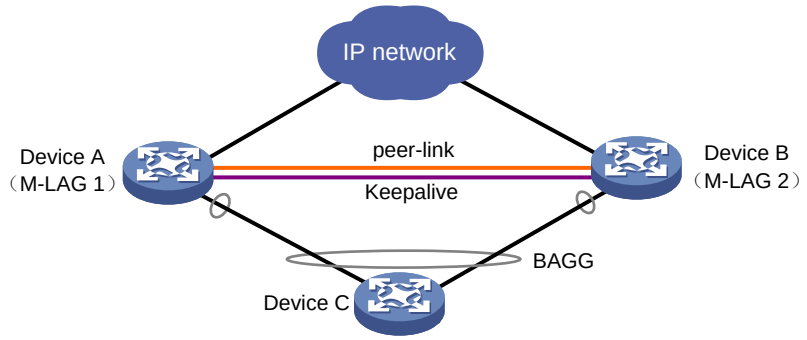


14 典型组网应用

14.1 单级M-LAG场景

如图 71 所示，为了保证可靠性，Device C 在接入网络时需要考虑链路的冗余备份，虽然可以采用部署 MSTP 等环路保护协议的方式，但是这种方式下链路的利用率很低，浪费大量的带宽资源。为了实现冗余备份同时提高链路的利用率，在 Device A 与 Device B 之间部署 M-LAG，实现设备的双归属接入。这样 Device A 与 Device B 形成负载分担，共同进行流量转发，当其中一台设备发生故障时，流量可以快速切换到另一台设备，保证业务的正常运行。

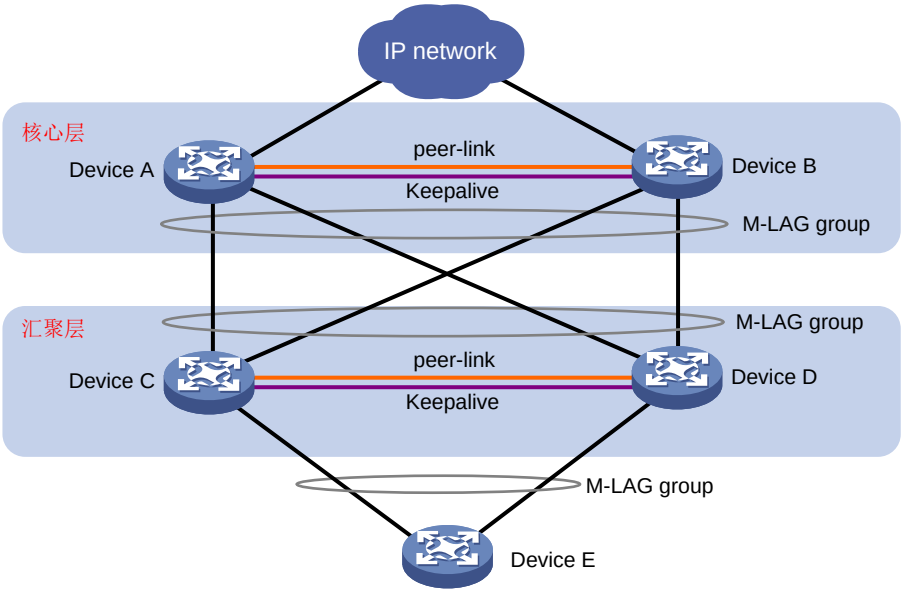
图71 单级 M-LAG 场景组网图



14.2 多级M-LAG互联场景

如图 72 所示，多级 M-LAG 互联可以在保证可靠性、提供链路利用率的同时扩展双归属接入的网络规模，可以在大二层数据中心网络设备数量比较多时提供稳定网络环境。同时汇聚层设备作为双活网关，核心层设备和汇聚层设备之间采用 M-LAG 组成聚合链路，保证设备级可靠性。

图72 多级 M-LAG 互联组网图

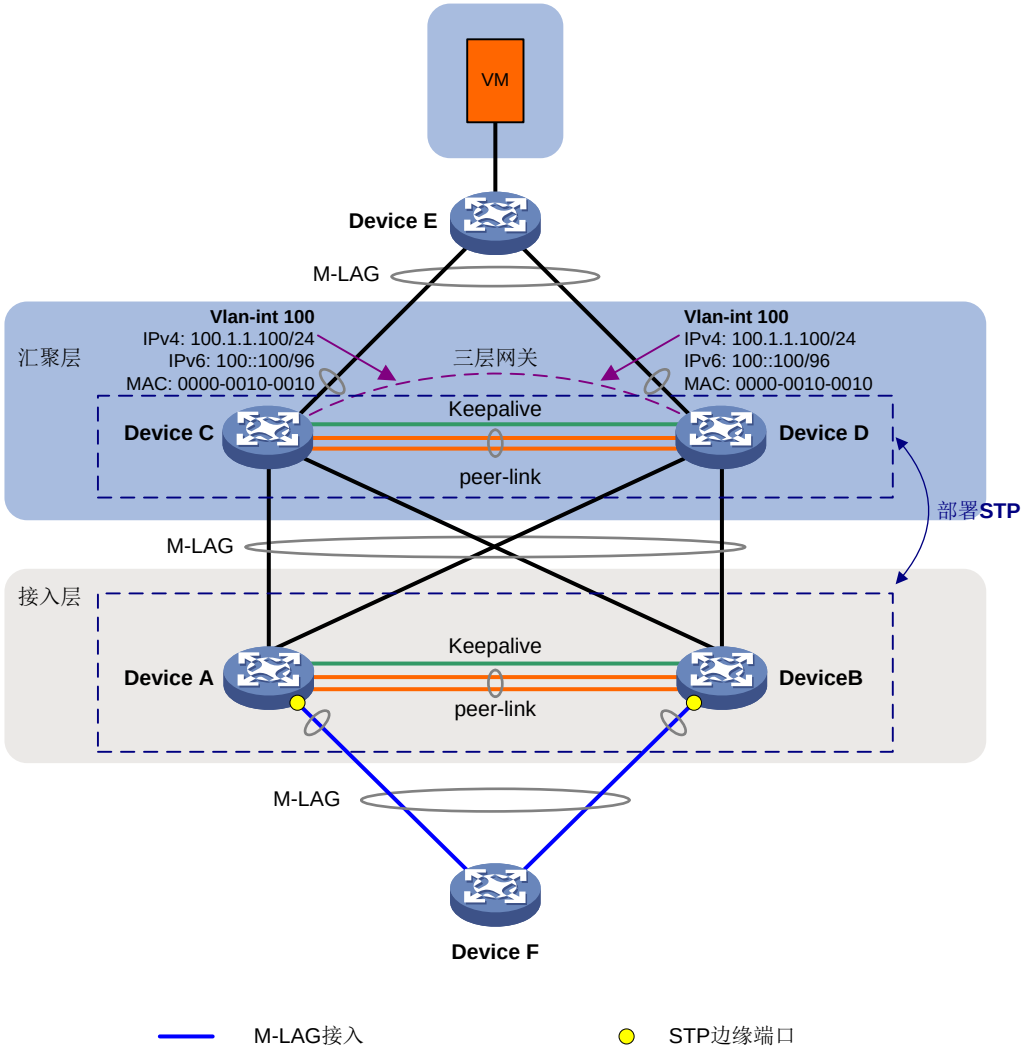


14.3 M-LAG与STP结合应用场景

如图 73所示，M-LAG 与 STP 结合应用场景的部署方案如下：

- 接入层的 Device A 和 Device B、汇聚层的 Device C 和 Device D 分别组成 M-LAG 系统，以避免单点故障造成流量转发中断，提高网络的可靠性。
- Device F 和 VM 通过 M-LAG 方式双归接入到 M-LAG 系统，以提高上行流量和下行流量的可靠性。其中，Device F 双归接入到 Device A 和 Device B 组成的 M-LAG 系统；VM 通过 Device G 双归接入到 evice C 和 Device D 组成的 M-LAG 系统。
- 多级 M-LAG 组网中，汇聚层的 Device C 和 Device D 作为三层网关，为 Device F 提供网关和路由接入服务。M-LAG 支持 VLAN 双活网关和 VRRP 网关两种网关部署方案。
- 在 Device A～Device D 上部署 STP，并指定 Device C 和 Device D 作为根桥，以消除 M-LAG 系统之间的环路。

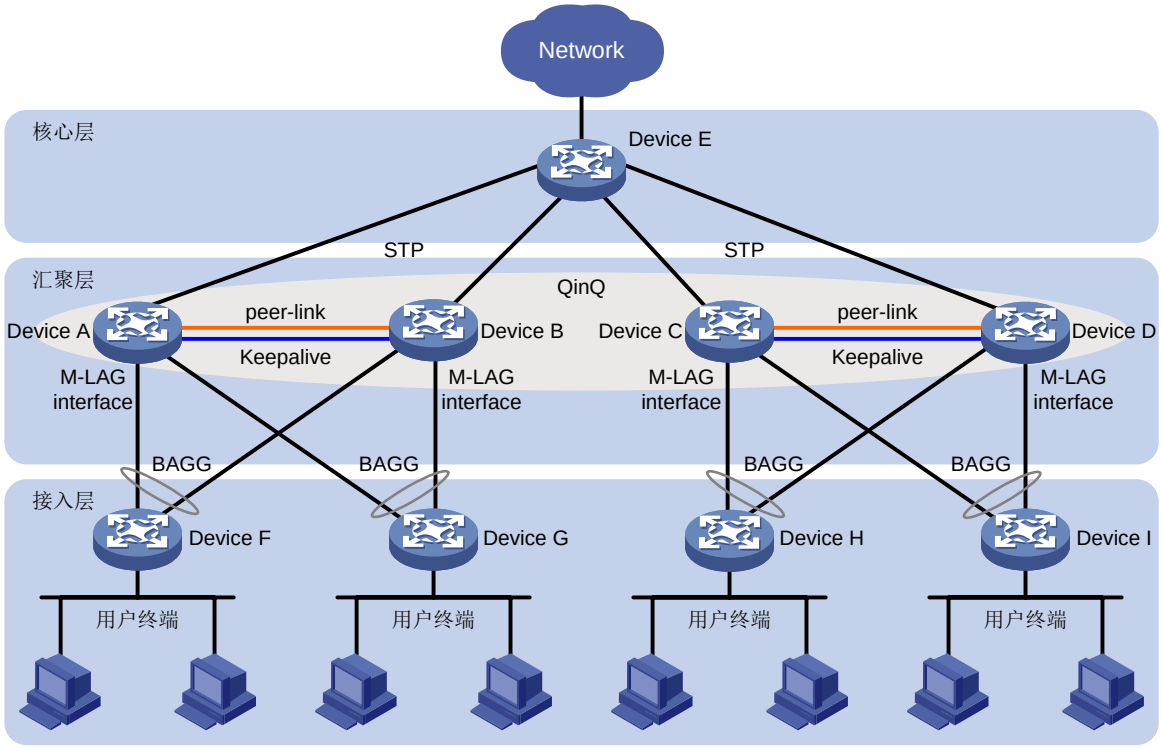
图73 二级 M-LAG+STP 组网示意图



14.4 M-LAG与QinQ结合应用场景

如图 74所示，在汇聚层设备上部署 M-LAG 和 QinQ，为 QinQ 服务提供设备高可靠性和负载分担；在汇聚层与核心层间部署生成树，以避免产生环路。

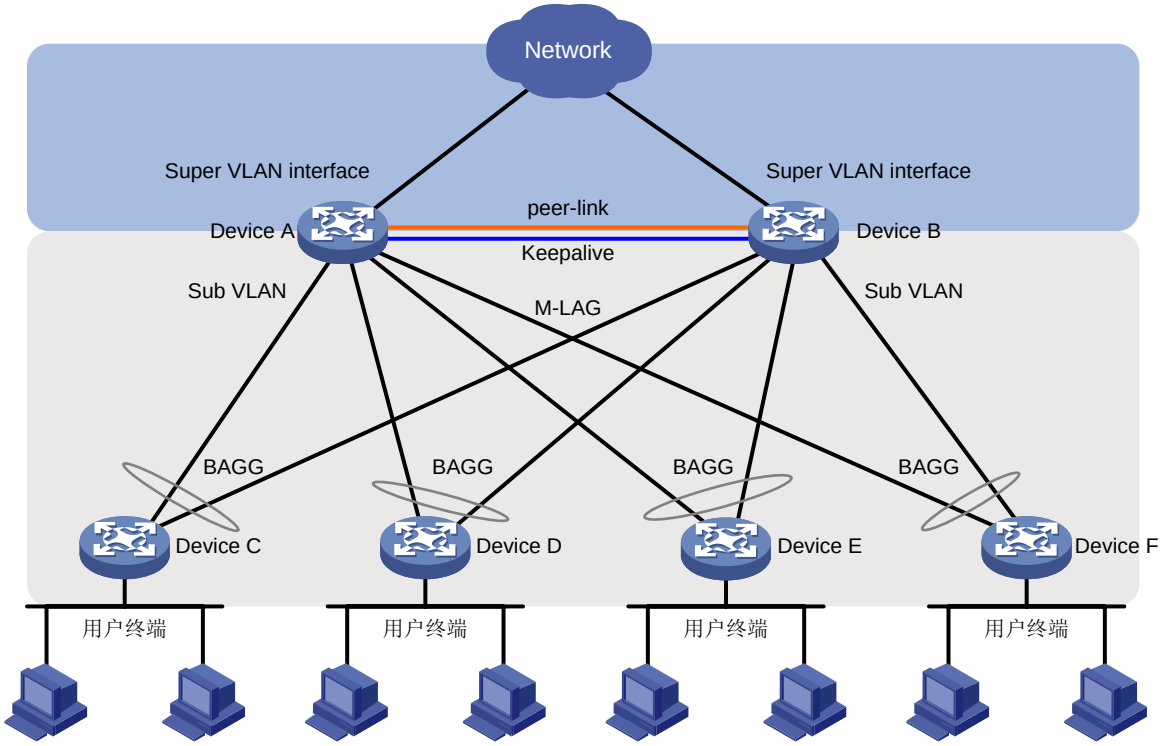
图74 M-LAG 与 QinQ 结合组网图



14.5 M-LAG与Super VLAN结合应用场景

如图 75 所示，为用户终端提供接入的设备均双归接入到与外界网络进行三层通信的 M-LAG 系统网关，M-LAG 设备运行 Super VLAN 服务以节约 IP 地址资源，同时通过 M-LAG 系统提供设备的高可靠性和链路的负载分担。M-LAG 系统的两台设备与外部网络间的上行链路可以借助路由协议实现负载分担。

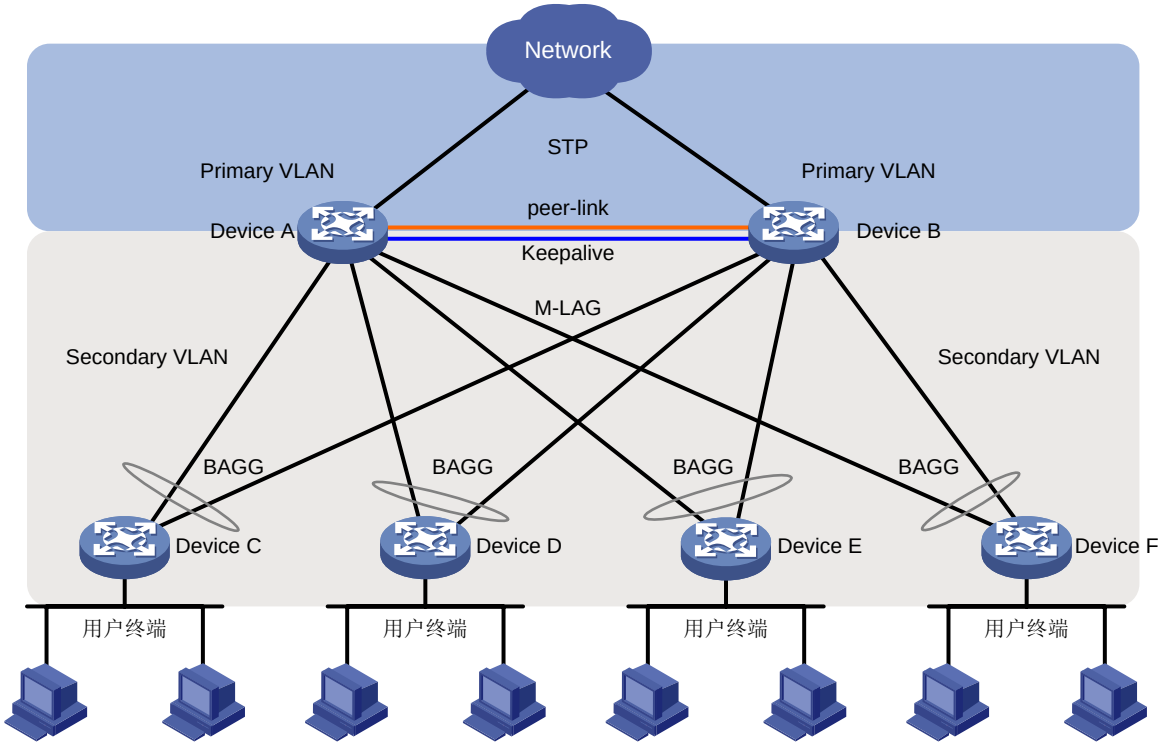
图75 M-LAG 与 Super VLAN 结合组网图



14.6 M-LAG与Private VLAN结合应用场景

如图 75 所示，为用户终端提供接入的设备均双归接入到与外界网络进行二三层通信的 M-LAG 系统网关，M-LAG 设备运行 Private VLAN 服务以节约 VLAN 资源，同时通过 M-LAG 系统提供设备的高可靠性和链路的负载分担。M-LAG 系统的两台设备与外部网络间的上行链路可以借助路由协议实现三层流量的负载分担，如果 M-LAG 系统的两台设备需要与外部网络进行二层通信，需要在上行链路配置生成树协议，以避免产生环路。

图76 M-LAG 与 Private VLAN 结合组网图



14.7 EVPN VXLAN支持M-LAG

EVPN VXLAN 支持 M-LAG 组网中，两台 VTEP 虚拟为一台 VTEP，在 VTEP 之间通过 peer-link 同步 MAC 地址和 ARP 信息，以确保两台 VTEP 上的 MAC 地址和 ARP 信息保持一致。peer-link 连接既可以是以太网聚合链路，如图 77 所示，也可以是 VXLAN 隧道，如图 78 所示。

在下行方向，跨 VTEP 设备形成链路聚合，实现用户侧链路的备份，从而避免单条以太网链路故障导致虚拟机无法访问网络。

图77 EVPN VXLAN 支持 M-LAG 组网图（以太网聚合链路作为 peer-link）

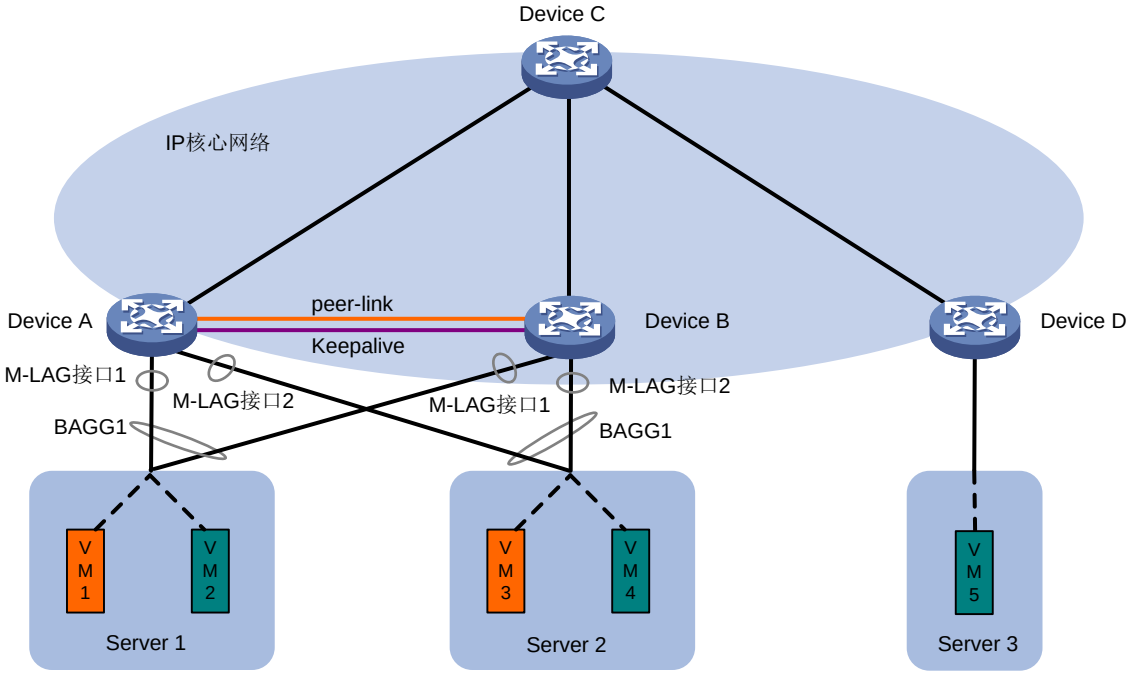
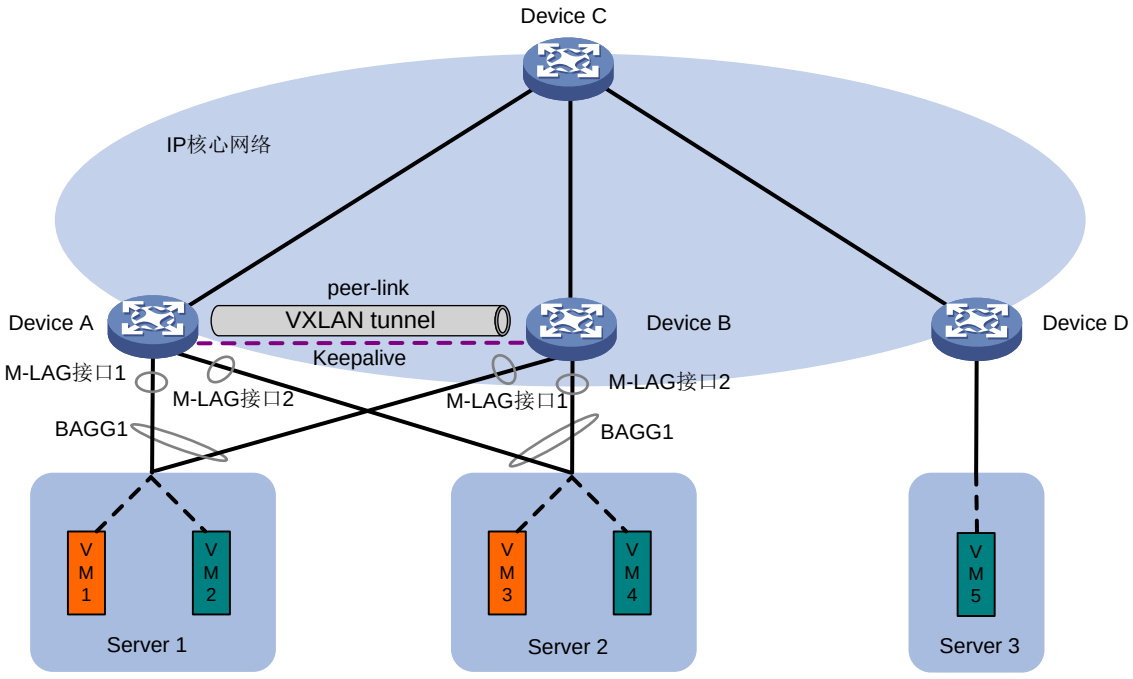


图78 EVPN VXLAN 支持 M-LAG 组网图（VXLAN 隧道作为 peer-link）



15 参考文献

- IEEE P802.1AX-REV™/D4.4c：Draft Standard for Local and Metropolitan Area Networks

- RFC 7432：BGP MPLS-Based Ethernet VPN