CloudEngine 数据中心交换机 M-LAG 最佳实践（V3 版本）

文档版本 01

发布日期 2023-12-01

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

前言

概述

本文档详细的描述了CloudEngine系列交换机（V3版本）M-LAG组网场景下推荐的基线方案和配置指导。

读者对象

读者对象本文档主要适用于项目规划设计和部署实施的操作人员。操作人员必须具备以下经验和技能：

- 熟悉华为数据中心网络CloudEngine交换机产品。
- 熟悉M-LAG特性的基本原理。

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
危险	表示如不避免则将会导致死亡或严重伤害的具有高等级风险的危害。
警告	表示如不避免则可能导致死亡或严重伤害的具有中等级风险的危害。
注意	表示如不避免则可能导致轻微或中度伤害的具有低等级风险的危害。
须知	用于传递设备或环境安全警示信息。如不避免则可能会导致设备损坏、数据丢失、设备性能降低或其它不可预知的结果。 “须知”不涉及人身伤害。
说明	对正文中重点信息的补充说明。 “说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。

修改记录

文档版本	发布日期	修改说明
02	2024-03-01	删除静态Bypass VXLAN隧道用到的VLAN下的**m-lag peer-link reserved**配置。
01	2023-12-01	第一次正式发布。

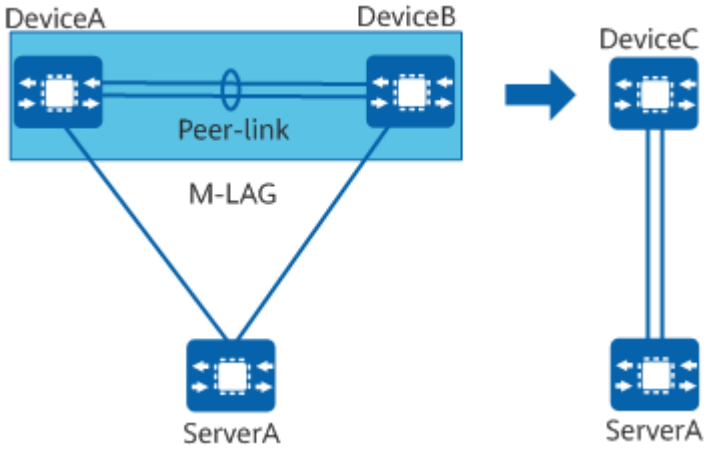
目 录

1 M-LAG 概述

1.1 M-LAG 简介

M-LAG（Multichassis Link Aggregation Group）即跨设备链路聚合组，是一种实现跨设备链路聚合的机制。如图1-1所示，M-LAG是将ServerA（可以是设备或主机）与另外两台设备DeviceA和DeviceB进行跨设备链路聚合，如同ServerA和一台设备建立了链路聚合关系，从而把链路可靠性从单板级提高到了设备级。

图 1-1 M-LAG 示意图



M-LAG主要应用于普通以太网络、VXLAN和IP网络的双归接入，可以起到负载分担或备份保护的作用。相较于另一种常见的可靠性接入技术——堆叠，M-LAG在可靠性、升级等方面有着显著的优势，两者的对比如表1 堆叠与M-LAG的对比所示。

表 1-1 堆叠与 M-LAG 的对比

对比维度	堆叠	M-LAG（推荐）
可靠性	**一般**：控制面集中，可能故障在成员设备上扩散	**更高**：控制面独立，故障域隔离
配置复杂度	**简单**：逻辑上是一台设备	**一般**：两台设备均需独立配置
成本	**一般**：需要部署堆叠线缆	**一般**：需要部署Peer-link连线

对比维度	堆叠	M-LAG（推荐）
性能	**一般：**Master控制面要控制所有堆叠成员的转发面，CPU载荷加重	**高：**成员设备独立转发，CPU载荷保持不变
升级中断时间	**相对较长：**通过堆叠快速升级，典型配置组网下，业务中断时间在20秒~1分钟左右；若采用堆叠整系统重启方式，升级时长可达数分钟	**短：**每台设备可独立升级，流量秒级中断
网络设计	**相对简单：**逻辑上单节点设计	**相对复杂：**逻辑上双节点设计
适用场景	• 对软件版本升级中断时间无要求 • 维护简单	• 对软件版本升级时业务中断时间要求较高 • 可靠性更高 • 可接受增加一定程度的维护复杂度

1.2 M-LAG 基本概念

M-LAG基本拓扑如图1-2所示，涉及的基本概念如表1-2所示。

图 1-2 M-LAG 基本拓扑

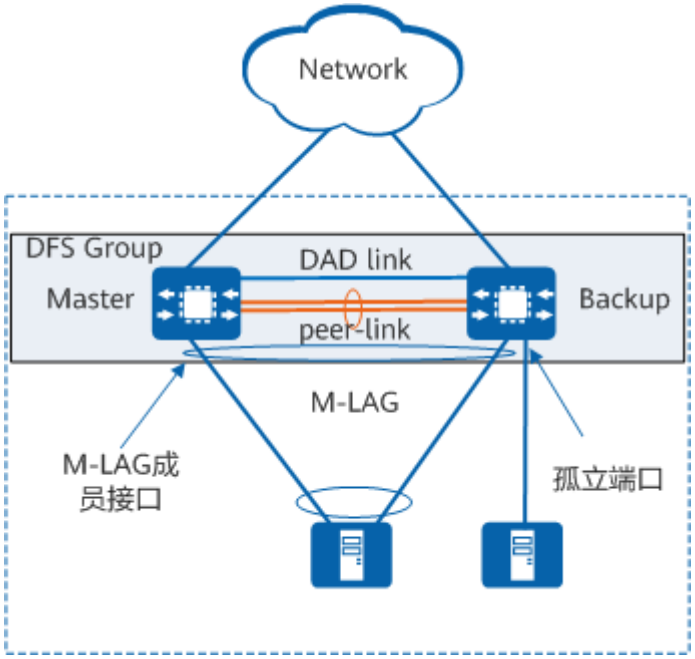


表 1-2 M-LAG 基本概念

概念	说明
DFS Group	动态交换服务组DFS Group（Dynamic Fabric Service Group），主要用于实现M-LAG设备之间的配对，M-LAG设备之间的接口状态、表项等信息同步需要依赖DFS Group协议进行同步。 DFS Group的角色区分主备，配对成功后，两台设备经过DFS Group协商，协商出DFS主设备（M-LAG主设备，Master）和DFS备设备（M-LAG备设备，Backup）。
peer-link	M-LAG设备之间的直连链路，链路两端直连的接口为peer-link接口，用于传输协议报文、表项同步报文，并转发部分流量。
DAD link	双主检测链路，又称为心跳链路，是一条三层互通链路，用于M-LAG设备之间发送双主检测报文。 正常情况下，双主检测链路不会参与M-LAG的任何转发行为，只在DFS Group配对失败或者peer-link故障场景下，用于检查是否出现双主的情况。
M-LAG成员接口	M-LAG主备设备上连接用户侧设备或主机的Eth-trunk接口，从而实现跨设备链路聚合的目的。
孤立端口	M-LAG设备上未加入任何M-LAG成员口的端口。
保留端口	当peer-link故障时，M-LAG分裂，配对的两台设备无法相互发送协议报文及同步报文，两台设备处于双主状态。为了避免流量转发异常，需要将一端M-LAG设备上的端口置为Error-Down，但在实际组网应用中，某些端口并不希望被置为Error-Down，这类peer-link故障时不被Error-Down的端口被称为保留端口（简称保留口）。 缺省情况下，设备上仅管理网口和peer-link接口为保留口，其他端口可以通过**m-lag unpaired-port reserved**命令配置为保留口。
工作模式	M-LAG分为双活模式和主备模式，两者在M-LAG主备成员口产生方式上有所不同，且主备模式仅适用于服务器主备网卡接入M-LAG场景，具体可参见**M-LAG建立过程**。

1.3 M-LAG 建立过程

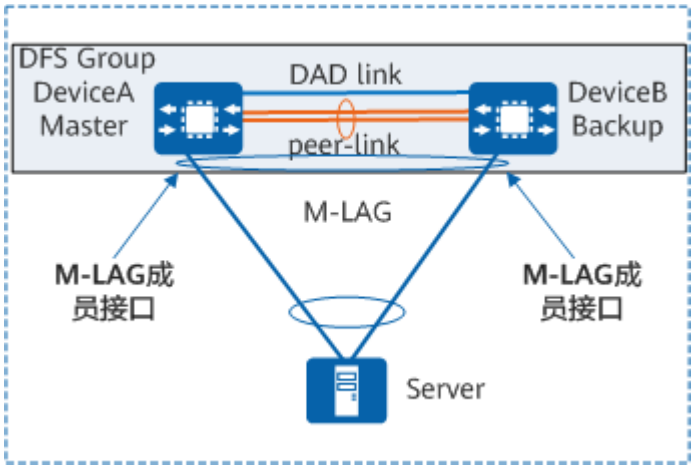
在M-LAG基本建立过程中，根据M-LAG主备成员口产生方式的不同，可将M-LAG工作模式分为两种：双活模式和主备模式。

双活模式

如**图1-3**所示，在两台独立的设备DeviceA和DeviceB上，建立一个跨设备的Eth-Trunk与用户侧设备Server的普通Eth-Trunk并通过M-LAG机制进行对接，共同组成一个M-LAG系统。双活模式下，DeviceA和DeviceB通过动态交换服务组DFS Group（Dynamic Fabric Service Group），进行M-LAG设备配对，协商出M-LAG主备设备和M-LAG成员口的主备状态，同时进行表项同步。DeviceA和DeviceB的直连链路为peer-link链路，用于交换协商报文及转发部分流量，peer-link口默认加入所有VLAN和BD。M-LAG成员口即DeviceA和DeviceB上连接用户侧Server的Eth-Trunk接口，用于流量接入。

双活模式下，两条链路可以起到负载分担流量转发的作用。

图 1-3 M-LAG 双活模式基本组网



M-LAG的建立过程分为如下几个步骤：

1. DFS Group配对

   当M-LAG两端设备完成配置后，设备首先通过peer-link链路发送DFS Group的Hello报文。当设备收到对端的Hello报文后，会判断报文中携带的DFS Group编号是否和本端相同，如果两台设备的DFS Group编号相同，则两台设备DFS Group配对成功。

2. DFS Group协商主备

   配对成功后，两台设备会向对端发送DFS Group的设备信息报文，设备根据报文中携带的DFS Group优先级以及系统MAC地址确定出M-LAG设备的主备状态。

   以DeviceB为例，当DeviceB收到DeviceA发送的报文时，DeviceB会查看并记录对端信息，然后比较DFS Group的优先级，如果DeviceA的DFS Group优先级高于本端的DFS Group优先级，则确定DeviceA为M-LAG主设备，DeviceB为M-LAG备设备。如果DeviceA和DeviceB的DFS Group优先级相同，比较两台设备的MAC地址，确定MAC地址小的一端为M-LAG主设备。

   说明

   正常情况下，M-LAG主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。

3. M-LAG成员接口协商主备

   在DFS Group协商出主备状态后，M-LAG的两台设备会通过peer-link链路发送M-LAG设备信息报文，报文中携带了M-LAG成员口的配置信息。在成员口信息同步完成后，确定M-LAG成员口的主备状态。

   初始状态，M-LAG成员口的主备状态与M-LAG设备的主备状态一致，此时M-LAG成员口的主备状态与M-LAG设备的主备状态一致，即：M-LAG主设备上的成员口为M-LAG主成员口，M-LAG备设备上的成员口为M-LAG备成员口。其他场景下，与对端同步成员口信息时，状态由Down先变为Up的M-LAG成员口成为M-LAG主成员口，对端对应的M-LAG成员口为M-LAG备成员口，且主备状态默认不回切，即：如果M-LAG成员口状态为主的设备故障，则对端对应的M-LAG成员口从备状态升级为主状态；当M-LAG成员口状态为主的设备故障恢复后，先前由备状态升级为主状态的接口仍保持主状态，恢复故障的M-LAG成员口状态为备。

4. 双主检测

协商出M-LAG主备设备后，两台设备之间会通过双主检测链路按照1s的周期发送M-LAG双主检测报文，当两台设备均能够收到对端发送的报文时，M-LAG系统即开始正常的工作。一旦设备感知peer-link故障，会在双主检测延时时间（缺省值为3s）后，按照200ms的周期发送三个双主检测链路报文进行加速检测，防止误触发双主加速检测，导致一端M-LAG设备端口被Error-Down。

正常情况下，双主检测链路不会参与M-LAG的任何转发行为，只在DFS Group配对失败或者peer-link故障场景下，用于检查是否出现双主的情况，所以即便双主检测失败也不会影响M-LAG正常工作。

5. M-LAG同步信息

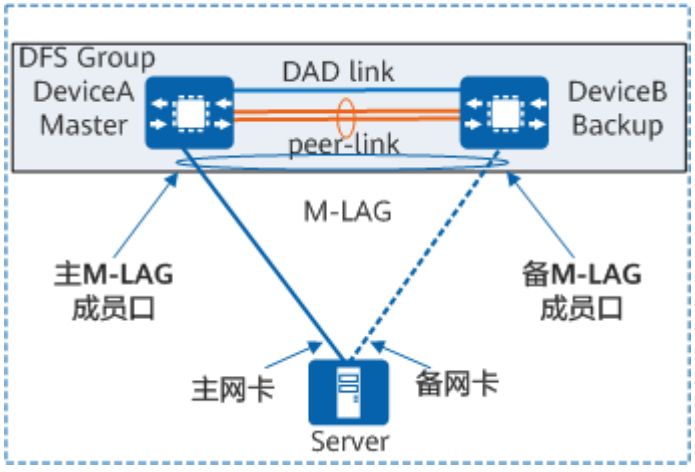
   正常工作后，两台设备之间会通过peer-link链路发送M-LAG同步报文实时同步对端的信息，M-LAG同步报文中包括MAC表项、ARP表项、ND表项等，发送M-LAG成员口的状态，这样任意一台设备故障都不会影响流量的转发，保证正常的业务不会中断。

主备模式

如图1-4所示，在两台独立的设备DeviceA和DeviceB上，建立一个跨设备的Eth-Trunk与用户侧服务器Server并通过M-LAG机制进行对接，共同组成一个M-LAG系统。主备模式下，DeviceA和DeviceB通过动态交换服务组DFS Group，进行M-LAG设备配对，协商出M-LAG主备设备和选举出M-LAG主备成员口，同时进行表项同步。DeviceA和DeviceB的直连链路为peer-link链路，用于交换协商报文及转发部分流量，peer-link口默认加入所有VLAN和BD。M-LAG成员口即DeviceA和DeviceB上连接用户侧Server的主备网卡，用于流量接入。

主备模式仅适用于服务器主备网卡接入M-LAG场景，两条链路可以起到备份保护的作用。正常情况下，仅主网卡对接的M-LAG设备收发流量，备网卡对接的M-LAG设备收到流量后，通过peer-link口绕行到主网卡对接的M-LAG设备；当服务器主网卡对接的M-LAG成员口状态变为Down时，流量从peer-link口引流到备网卡对接的M-LAG设备，流量通过M-LAG备成员口直接转发给服务器，完成快速切换，提高服务器主备网卡接入M-LAG场景下的故障切换丢包性能。

图 1-4 M-LAG 主备模式基本组网



与双活模式类似，主备模式下M-LAG的建立过程也分为如下几个步骤：

1. DFS Group配对
2. DFS Group协商主备

3. M-LAG成员接口选举主备
4. 双主检测
5. M-LAG同步信息

主备模式仅在M-LAG主备成员口的产生方式上与双活模式不同，M-LAG设备根据服务器主网卡发出的协议报文（即选主报文）进行M-LAG成员口的主备选举，与主网卡对接的M-LAG成员口成为M-LAG主成员口，对端对应的M-LAG成员口为M-LAG备成员口（即与备网卡对接的M-LAG成员口成为M-LAG备成员口）。

当服务器主网卡发生故障时，M-LAG备成员口将自动升主，实现故障场景下流量快速切换。另外，当服务器检测主网卡所在链路故障，也会将备网卡升主，升主后的备网卡重新发送选主报文，触发M-LAG成员口进行主备选举，与M-LAG备成员口对接的M-LAG设备收到选主报文后，M-LAG成员口备也会升主。

当服务器的主网卡故障恢复后，M-LAG成员口的主备状态默认不回切，主备状态是否回切，依赖于主网卡故障恢复后是否发送选主报文。如果故障恢复后的主网卡重新发送选主报文进行M-LAG成员口主备选举，则主备状态回切。

设备支持ARP、ND、IGMP和DHCP报文进行M-LAG成员口主备选举。从V300R023C00版本开始支持MLD报文进行选举。

说明

主备模式下，M-LAG成员口的主备选举依赖选主报文上送CPU。为了使IGMP/DHCP/MLD报文上送CPU，需要先在VLAN或BD下使能IGMP Snooping功能，才可以支持通过IGMP报文进行主备选举；或者在全局下使能DHCP功能，才可以支持通过DHCP报文进行主备选举；或者在VLAN或BD下使能MLD Snooping功能，才可以支持通过MLD报文进行主备选举。另外，当M-LAG设备作为二层透传设备或二层网关时，ARP/ND报文不会上送CPU，无法通过选主报文进行主备选举，这种情况下需要在VLAN下使能ARP二层代答功能，在BD下使能ARP广播抑制功能或NS组播抑制功能，才可以支持通过ARP/ND报文进行主备选举。

1.4 DAD 链路部署方案

双主检测（DAD）链路是一条三层互通链路，用于M-LAG设备之间发送双主检测报文。设备支持的双主检测链路部署方式具体如表1-3所示。

表 1-3 双主检测链路部署方式

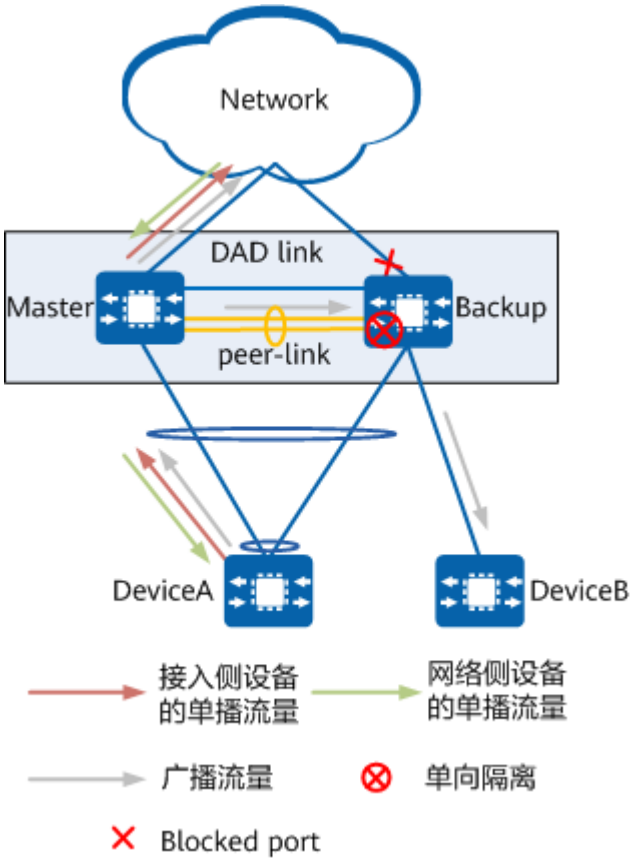
部署方式	说明
使用单独心跳线作为双主检测链路	推荐使用。使用单独心跳线作为双主检测链路时，建议使用三层主接口作为双主检测链路接口，如果使用VLANIF接口，则需要保证peer-link接口不允许通过该VLAN，否则会有环路或MAC漂移的现象。 另外，为避免peer-link故障场景下，双主检测链路接口被Error-Down，需要配置双主检测链路接口为保留口。
无单独心跳线，使用管理网口作为双主检测链路	不推荐使用。使用管理网口作为双主检测链路时，需要部署单独管理网。

部署方式	说明
无单独心跳线，使用业务接口作为双主检测链路	不推荐使用。使用业务口作为双主检测链路时，需要关闭二次故障增强功能（即不能配置双主检测的**peer** *ip-address*参数），否则在peer-link故障场景下，会因一端M-LAG设备的业务口被Error-Down，进而产生Error-Down震荡。

1.5 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1-5所示，从接入设备或网络侧到达M-LAG配对设备的单播流量，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链路广播到对端M-LAG设备，在peer-link链路与M-LAG成员口之间设置单方向的流量隔离，即从peer-link口进来的流量不会再从M-LAG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 1-5 M-LAG 接入二层网络流量转发示意图



单向隔离机制

机制生效前提

当M-LAG两台设备协商出M-LAG主备后，系统通过M-LAG同步报文判断接入设备是否双活接入：

- 若接入设备双活接入M-LAG系统，则M-LAG两台设备下发对应M-LAG成员口的单向隔离配置，来隔离由peer-link口发往M-LAG成员口的流量。

  说明

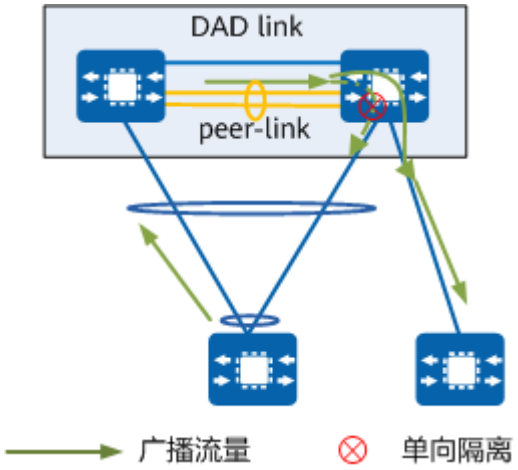
  M-LAG防环机制中的单向隔离对二层（包括单播、组播、广播）流量生效，三层组播流量生效，三层单播流量不生效。
- 若接入设备单归接入M-LAG系统，则M-LAG系统不会下发对应M-LAG成员口的单向隔离配置。
- 主备模式下，M-LAG系统不会下发对应M-LAG成员口的单向隔离配置。

单向隔离机制实现原理

如图1-6所示，在设备双活接入M-LAG场景下，设备通过端口隔离功能实现peer-link接口与M-LAG成员口之间的单向隔离，隔离由peer-link接口发往M-LAG成员口的广播等泛洪流量。

M-LAG两台设备会自动下发对应M-LAG成员口的单向隔离配置，无需用户手工配置，同一端口隔离组内的端口之间无法互通。当M-LAG设备感知到本端的M-LAG成员口状态为Down时，会通过peer-link发送M-LAG同步报文，通知对端设备自动将相应的M-LAG成员口从隔离组中清除。

图 1-6 M-LAG 单向隔离示意图



1.6 M-LAG 支持的破环协议

M-LAG组网中，为避免网络环路的发生，组建M-LAG的两台设备需要对外呈现为一台设备进行STP协议计算，设备当前支持两种破环协议：V-STP（Virtual Spanning Tree Protocol）和V-VBST（Virtual VLAN-Based Spanning Tree）。

V-STP

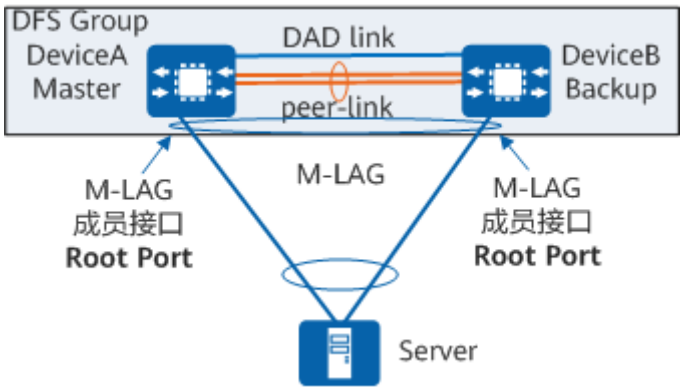
V-STP是二层拓扑管理特性，其核心思想是将两台设备的STP协议虚拟成一台设备的STP协议，对外呈现为一台设备进行STP协议计算。

M-LAG主备设备使能STP功能后，STP可以感知M-LAG主备协商状态。在M-LAG主备协商成功后，STP需要同步M-LAG主备设备的桥MAC信息和实例优先级信息，M-LAG备设备使用M-LAG主设备同步过来的桥MAC信息和实例优先级信息进行STP计算和收发报文，且peer-link链路不参与STP协议计算，保证M-LAG两台设备被虚拟化成一台设备进行端口角色计算和快速收敛计算。

当前，V-STP只能用于M-LAG组网，适用于与传统STP网络对接，可以解决多级M-LAG互联场景和组成M-LAG的设备作为非根桥场景的需求。

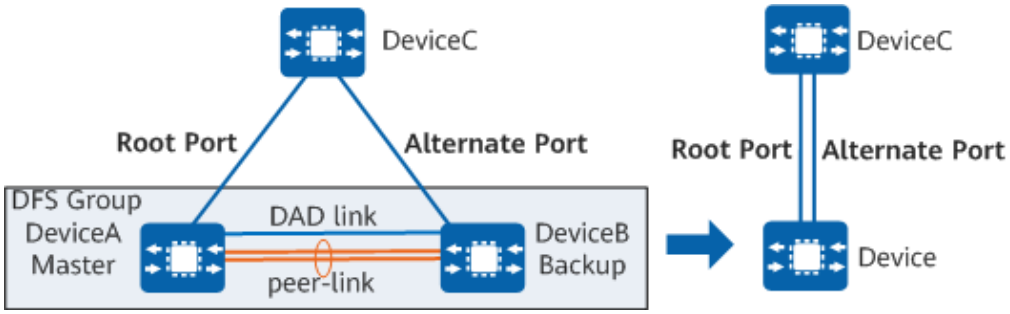
1. 如图1-7所示，将M-LAG成员接口协商成相同的角色。

图 1-7 M-LAG 成员接口协商成根端口（Root Port）



2. 如图1-8所示，将M-LAG两台设备虚拟成一台设备，两端M-LAG的端口分别为根端口和Alternate端口（Alternate Port，即根端口的备份端口）。

图 1-8 M-LAG 两台设备虚拟成一台设备



V-VBST

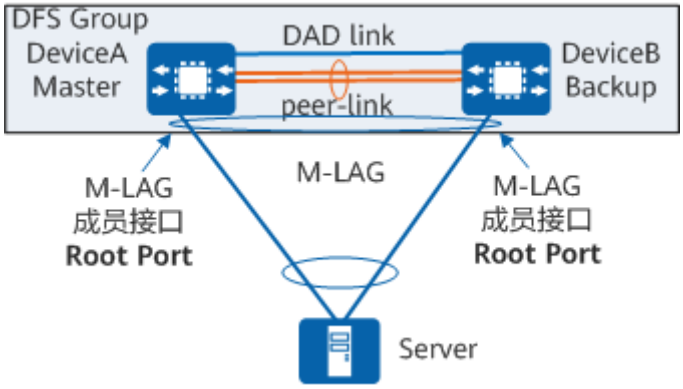
M-LAG支持VBST虚拟化计算，其核心思想是将两台设备的VBST协议虚拟成一台设备的VBST协议，对外呈现为一台设备进行VBST协议计算。

M-LAG主备设备使能VBST功能后，VBST可以感知M-LAG主备协商状态，在M-LAG主备协商成功后，VBST需要同步M-LAG主备的桥MAC信息和实例优先级信息。M-LAG主备协商成功后，M-LAG备设备使用M-LAG主设备同步过来的桥MAC信息和实例优先级信息进行VBST计算和收发报文，且peer-link链路不参与STP协议计算，保证两台设备被虚拟化成一台设备进行端口角色计算和快速收敛计算。

当前，V-VBST只能用于M-LAG组网，适用与其他网络对接，如PVST、VBST网络，需要基于VLAN计算，可以解决多级M-LAG互联场景和组成M-LAG的设备作为非根桥场景的需求。

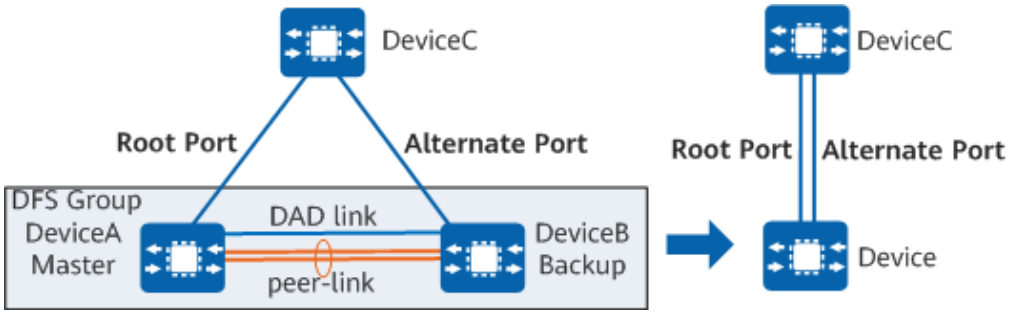
1. 如图1-9所示，将M-LAG成员接口协商成相同的角色。

图 1-9 M-LAG 成员接口协商成根端口（Root Port）



2. 如图1-10所示，将M-LAG两台设备虚拟成一台设备，两端M-LAG的端口分别为根端口和Alternate端口（Alternate Port，即根端口的备份端口）。

图 1-10 M-LAG 两台设备虚拟成一台设备



1.7 M-LAG 配置一致性检查

M-LAG是由两台设备组成的一个双活系统，是将两台设备在逻辑上虚拟成一台设备，形成一个统一的二层逻辑节点。这带来了逻辑拓扑的清晰高效，也决定了M-LAG两端设备的某些配置需要一致，某些配置需要不一致，否则可能会导致M-LAG无法正常工作或成环等问题。但M-LAG运用于网络中时，却面临一个突出的问题：部署网络时，通过手工配置、人工比对来保证每一个M-LAG系统两端设备的配置一致性，不仅处理效率低下，更多的是带来诸多潜在的误配置风险。

为了解决上述问题，华为公司提出了M-LAG配置一致性检查的解决方案。该解决方案中，通过M-LAG机制自带的配置一致性检查功能，去订阅M-LAG两端设备的各模块配置。我们可以通过检查功能返回的比对结果，及时地调整M-LAG两端设备的配置部署，防止组网成环或者数据丢包等问题发生。

M-LAG配置一致性检查将设备配置分为两类，常见的一致性检查项如表1-4所示，M-LAG支持的所有的一致性检查项请参见产品文档《配置指南-可靠性》中的“M-LAG配置”章节，分别为关键配置（Type 1）和一般配置（Type 2）。根据对关键配置检查不通过时的处理方式，M-LAG一致性又分为严格模式（strict）和松散模式（loose）。

1. 关键配置（Type 1）：如果M-LAG系统两端设备配置一致性检查不通过，会导致成环、状态正常但长时间丢包等问题。
   - 严格模式下，如果M-LAG两端设备存在Type 1配置一致性检查不通过，会导致M-LAG一端设备上成员口处于ERROR DOWN状态，且触发设备对Type 1类型配置检查不通过的告警。

- 松散模式下，如果M-LAG两端设备存在Type 1配置一致性检查不通过，则会触发设备对两种类型配置检查不通过的告警。

2. 一般配置（Type 2）：如果M-LAG系统两端设备配置一致性检查不通过，可能会导致M-LAG运行状态异常。与Type 1类型相比较而言，Type 2类型对组网环境的影响相对较小。

   无论处于何种模式，如果M-LAG两端设备存在表中Type 2配置一致性检查不通过，则会触发设备对两种类型配置检查不通过的告警。

此外，针对Type 2类型的配置，设备还支持基于模块开启或关闭M-LAG配置一致性检查功能。如果用户使能了M-LAG配置一致性检查功能，但又期望部分Type 2类型的配置不进行一致性检查，可以将指定模块的配置加入到M-LAG一致性检查白名单，白名单中的配置将不再进行一致性检查。

表 1-4 M-LAG 配置一致性检查配置分类列表

视图	配置	类型
全局	STP功能是否使能	Type 1
	V-STP功能是否使能	
	STP工作模式配置	
	BPDU保护功能是否使能	
	VLAN上的VBST功能是否使能	
M-LAG成员口	M-LAG主备模式配置	
	STP功能是否使能	
	STP端口的Root保护功能是否使能	
	M-LAG成员口的LACP模式配置	
	LACP M-LAG系统优先级和系统ID配置	
	STP边缘端口配置	
peer-link接口	STP功能是否使能	
	peer-link接口的LACP模式配置	
全局	参与主备选举的报文类型配置	Type2
	VLAN配置	

1.8 M-LAG 网关

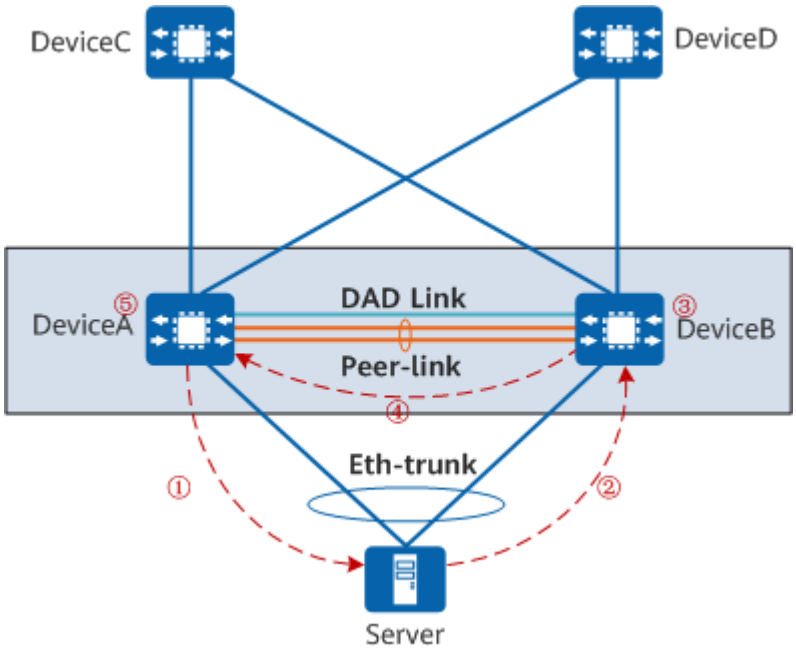
在M-LAG双归接入IP网络或VXLAN网络的场景中，M-LAG主备设备需要同时作为三层网关，必须保证M-LAG成员接口对应的VLANIF接口或VBDIF接口具有相同的IP地址和MAC地址。

如图1-11所示，DeviceA和DeviceB组建M-LAG并双归接入三层网络时，M-LAG双归设备成为二层网络和三层网络的分界点，即承担起网关的作用。由于是两台设备做网关，其对接入侧需要展示相同的网关IP地址和MAC地址。ARP/ND学习过程主要如下：

1. DeviceA向Server发送ARP/ND请求报文。
2. Server接收ARP/ND请求，学习到DeviceA端的IP地址和MAC地址，存入自己的ARP/ND表项中，并发送ARP/ND应答报文。应答报文在Eth-Trunk链路上随机HASH，被应答到DeviceB。
3. DeviceB接收ARP/ND应答报文，学习到Server端的IP地址和MAC地址，并存入自己的ARP/ND表项中。
4. DeviceB通过peer-link将ARP/ND表项同步到DeviceA。
5. DeviceA学习到Server端的IP地址和MAC地址，并存入自己的ARP/ND表项中。

经历以上5个阶段后，M-LAG两端设备的ARP表项保持一致，流量通过DeviceA或DeviceB都可以到达Server。

图 1-11 M-LAG 双活网关示意图



1.9 M-LAG 与组播协同工作机制

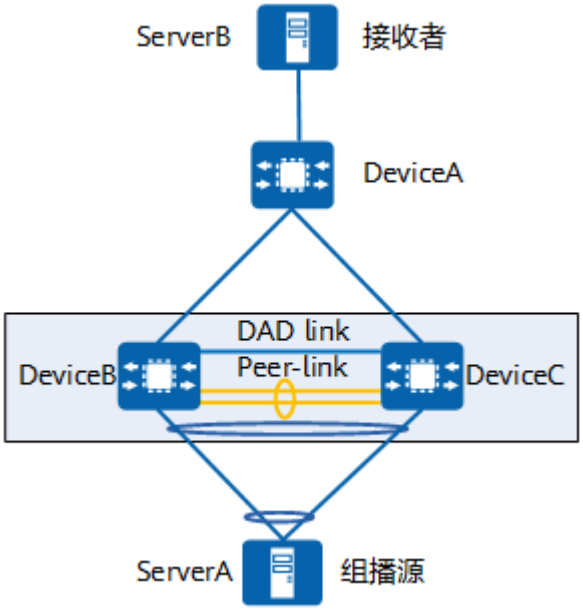
组播转发树的建立依赖于单播路由，当网络中的链路故障时，需要经历单播路由重新收敛，从而触发组播转发树的重新建立，但这个过程耗时较长，会导致组播流量丢失过多。

PIM FRR（Fast ReRoute）技术可以解决此问题。PIM FRR根据单播备份FRR路由，分别向组播源发送主备双加入，建立主备组播转发树，使主备链路交叉节点上分别从主备链路各收到一份组播流量。转发层面选收主链路的流量，丢弃备链路的流量，当主链路发生故障时，立刻选收备链路的流量，从而减少组播流量的丢失。

如图1-12所示，M-LAG与组播叠加场景，DeviceB和DeviceC为与组播源相连且负责转发该组播源发出的组播数据的第一跳PIM设备。DeviceA为与组播组成员相连且负责向该组成员转发组播数据的最后一跳PIM设备。在DeviceB和DeviceC间部署M-LAG，正

常工作时链路进行负载分担且任何一台设备故障对业务均没有影响。同时，在DeviceA、DeviceB和DeviceC上部署PIM FRR，设备选收主链路流量，当主链路故障时，流量立刻切换到备链路上，以实现组播链路故障保护的功能。

图 1-12 M-LAG 与组播叠加场景示意图

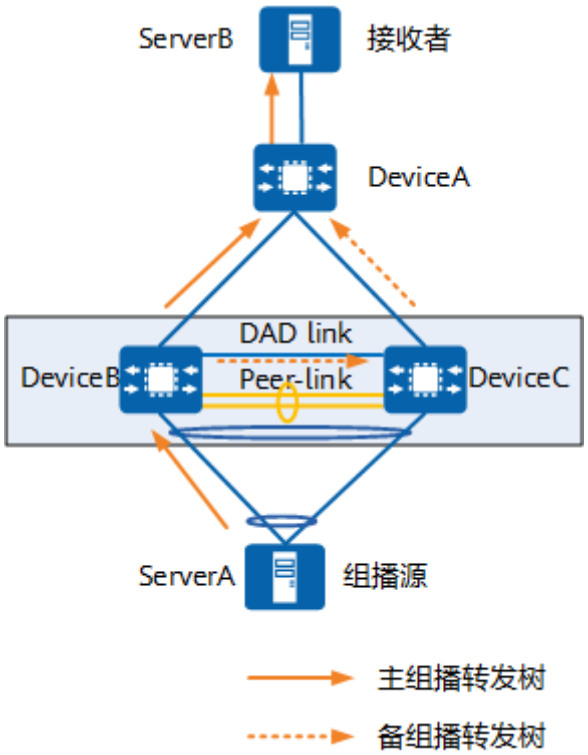


PIM FRR的实现过程包括下面三个步骤：

1. 主备组播转发树的建立

   PIM-SM中完成SPT切换的(S, G)表项或者PIM-SSM的(S, G)表项根据单播路由添加主入接口后，查找单播是否存在备份FRR路由，如果存在则添加备份入接口。同时向组播源发送主备双份PIM加入报文，建立主备组播转发树。主备组播转发树的建立过程如图1-13所示。

图 1-13 主备组播转发树的建立示意图

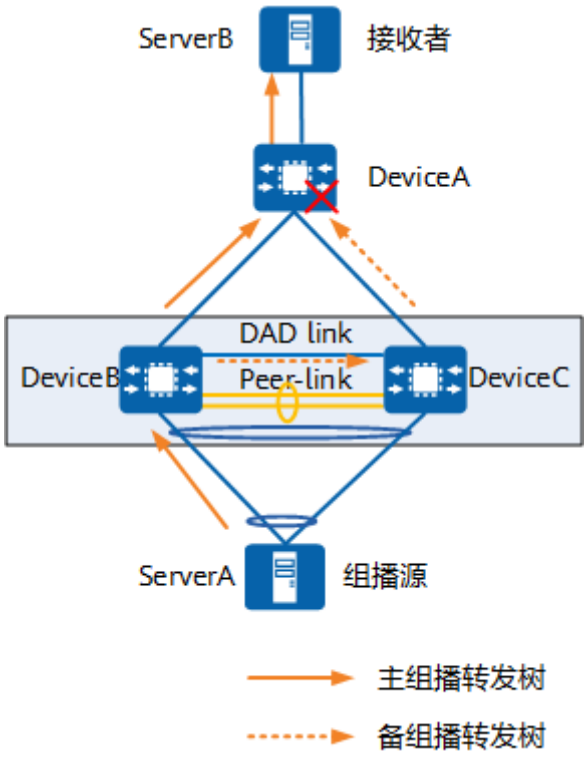


2. 故障检测及保护

   建立主备组播转发树之后，存在主备路由的设备能收到双份组播流量（如图1-13中的DeviceA）。转发层面选收主链路的流量，丢弃备链路的流量，当主链路发生故障时，通过感知主链路的链路状态Down，能够快速选收备链路流量。当链路是经过其他设备（如分波合波设备）实现直连，可能存在链路故障后链路状态仍为Up，PIM FRR无法通过感知链路状态快速切换的情况，此时可以通过部署BFD单跳检测来解决（如果是Eth-Trunk推荐配置BFD Link-bundle）。组播的入接口被BFD检测绑定后，PIM FRR可以通过感知BFD来实现快速切换。

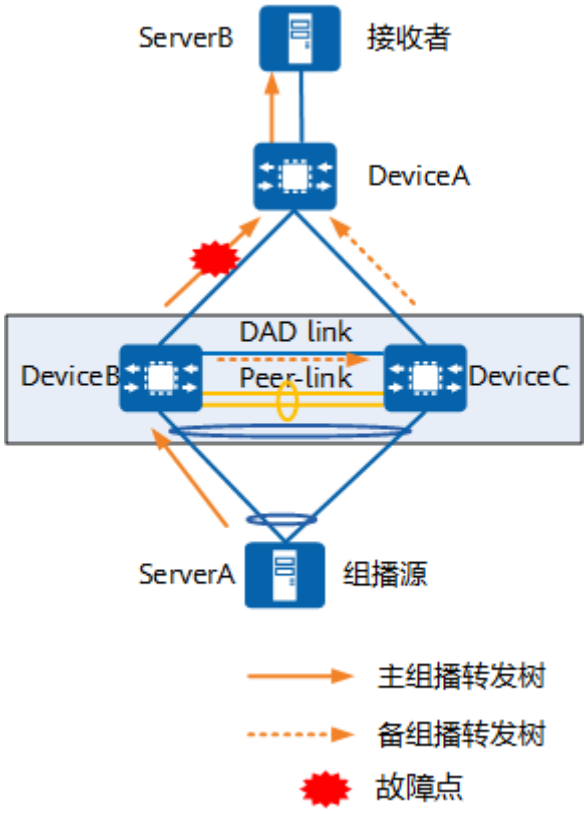
   如图1-14所示，DeviceA选收主链路的组播流量，丢弃备链路的组播流量。

图 1-14 PIM FRR 故障前示意图



如图1-15所示，当DeviceA所在的主链路发生故障时，则DeviceA立刻选收从DeviceC来的备份路径流量，即DeviceB -> DeviceC -> DeviceA。

图 1-15 PIM FRR 故障后示意图



3. 回切

   当链路故障恢复时，PIM协议层感知路由变化后，启动路由回切，而转发面为确保有充分时间进行转发表恢复，会进入WTR状态开始延迟回切，超过WTR时间后，会再平滑收敛到当前最优的主路径上。

   配置PIM FRR功能会生成具有两个入接口（主入接口和备入接口）的组播路由转发表。当主入接口发生故障（链路状态Down或BFD状态Down）时，为了快速恢复组播业务，对应的组播组会接收并转发来自备入接口的组播流。在组播路由转发表还未变化为单一入接口的情况下，如果主入接口故障状态恢复（链路状态Up或BFD状态Up），就需要等待一段时间再恢复（Wait to Restore，WTR）使用主入接口接收并转发组播流。WTR可以避免上游设备还未准备好的情况下恢复使用主入接口接收并转发组播流从而导致组播流中断。

   因为来自主入接口组播流量的恢复时长，与组播组的数量和CPU负载情况强相关，所以配置WTR的时长也需要结合组播组的数量和CPU负载情况调整，具体请参见**CloudEngine数据中心交换机 证券组播基线方案最佳实践**。

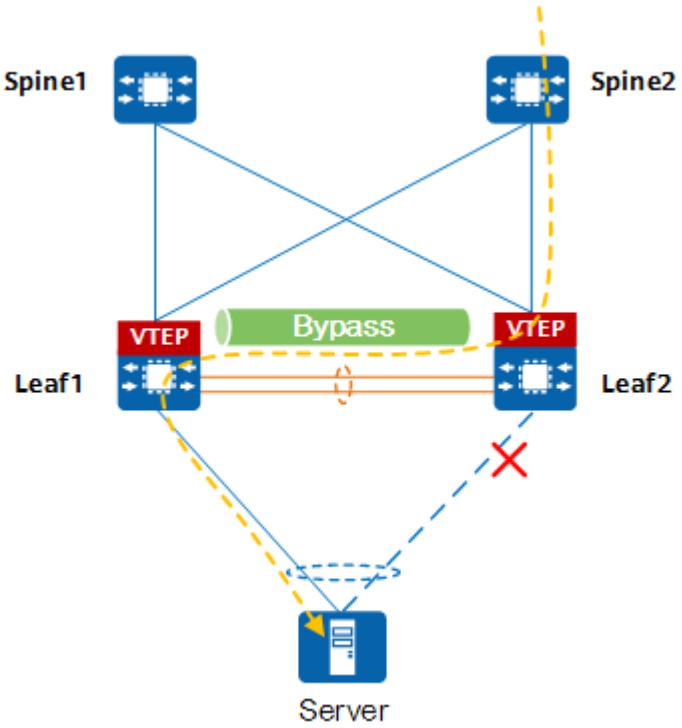
1.10 M-LAG 与 VXLAN 协同工作机制

在VXLAN网络中，组成M-LAG的设备作为VXLAN隧道的端点（VTEP）时，两台设备之间基于BD（Bridge Domain）进行ARP/ND/MAC表项的同步。在故障或者单边接入的场景，VXLAN Overlay流量可能需要经过peer-link绕行。而在缺省情况下，M-LAG两台设备之间并没有建立VXLAN隧道，无法实现VXLAN Overlay的互通。因此需要在两台设备之间基于peer-link链路部署Bypass VXLAN隧道，用于流量的绕行。Bypass VXLAN隧道的出接口为peer-link口，设备上所有BD默认加入Bypass VXLAN隧道中。

如**图1-16**所示，VXLAN Overlay流量从Spine通过ECMP Hash转发至Leaf2后，无故障情况时，Leaf2上通过查询设备ARP/ND/MAC表项将流量转发至服务器，出接口为下行

连接服务器的接口；在Leaf2的下行口故障时，ARP/ND/MAC表项的出接口切换至Bypass VXLAN隧道，Leaf2通过Bypass VXLAN隧道将流量转发Leaf1，再由Leaf1转发至服务器。

图 1-16 VXLAN 场景 M-LAG Bypass VXLAN 隧道示意图



1.11 M-LAG 流量转发机制（双活模式）

M-LAG建立成功后，M-LAG主备设备负载分担共同进行流量的转发。下面介绍在正常工作情况下M-LAG的流量转发机制。

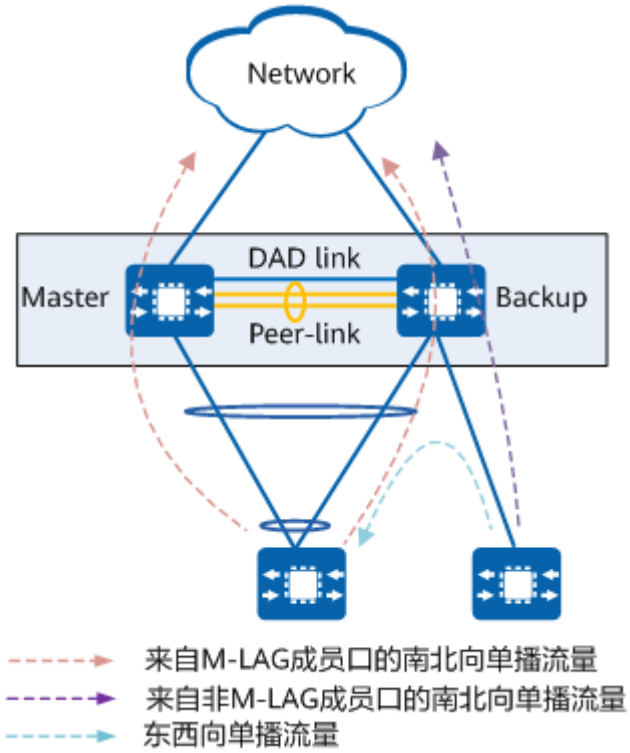
单播流量转发

如图1-17所示，M-LAG系统在设备双归接入场景下的已知单播流量转发：

对于南北向单播流量，在M-LAG接入侧，M-LAG的成员设备接收到接入设备通过链路捆绑负载分担发送的流量后，共同进行流量转发。流量到达M-LAG主备设备后，根据设备的路由表转发至网络侧。

对于东西向单播流量，在全部组建M-LAG，没有非M-LAG成员口的场景下，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主备设备转发至对应成员口。

图 1-17 M-LAG 已知单播流量转发示意图



组播流量转发

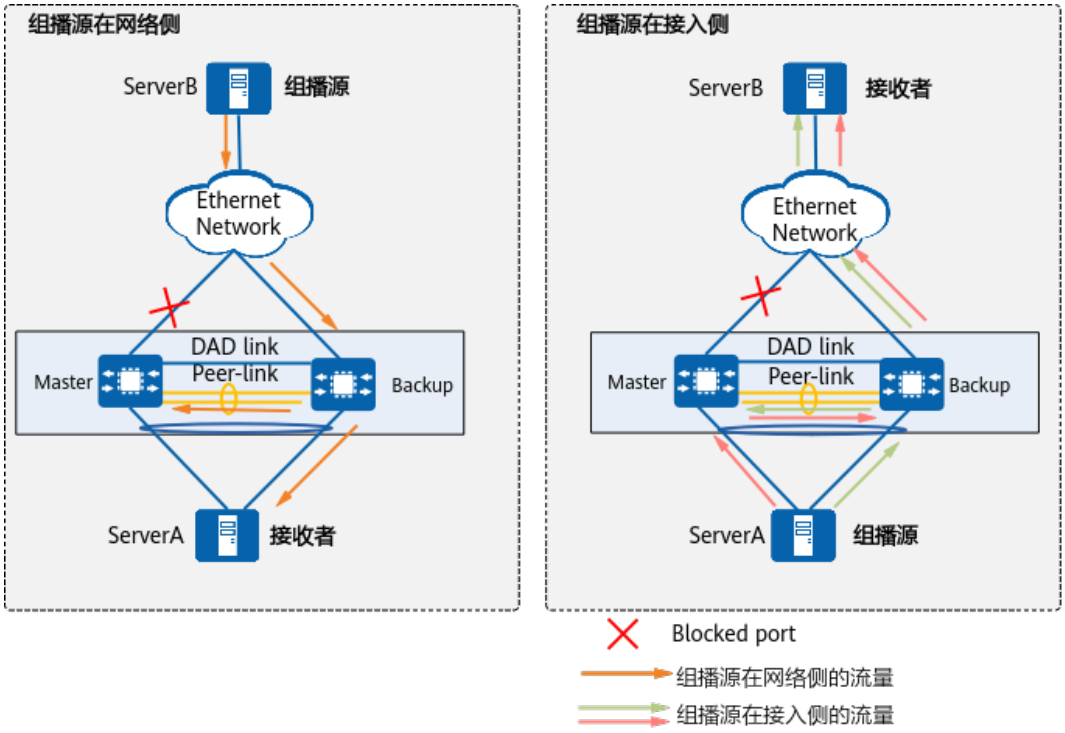
M-LAG接入二层网络

M-LAG接入二层网络，那么二层网络必须要保证发往M-LAG的流量只有一份，否则会有成环的风险。如图1-18所示，以M-LAG备设备的转发为例，假设左侧M-LAG主设备上行接口被STP协议阻塞。

组播源在网络侧的场景下，在ServerB作为组播源、ServerA作为组播组成员时，M-LAG成员口状态为主备的设备都可以转发组播流量，M-LAG备设备收到组播流量后向各个下一跳转发，网络侧流量会转发给ServerA、M-LAG主设备。当组播流量通过peer-link到达M-LAG主设备时，由于peer-link与M-LAG成员口存在单向隔离机制（即从peer-link口进来的流量不会再从M-LAG口转发出去），所以到达主设备的组播流量不会向ServerA转发。

组播源在接入侧的场景下，在ServerA作为组播源、ServerB作为组播组成员，且M-LAG设备无下挂其他组播组成员时，组播源发出的流量负载分担到M-LAG主备设备。收到流量后，上行接口被STP协议阻塞的主设备组播出接口指向peer-link链路，流量汇总到备设备端口，统一上送给ServerB。备设备的流量虽然也会通过peer-link转发到左侧主设备，但由于peer-link与M-LAG成员口存在单向隔离机制，主设备上行接口又被阻塞，因此这份流量无法转发到ServerA或ServerB。

图 1-18 M-LAG 接入二层网络组播流量转发示意图



M-LAG接入三层网络

M-LAG上行接入三层网络，M-LAG主备设备需要支持二三层组播混跑。如图1-19所示，M-LAG系统在接入设备双归接入场景下的组播流量转发：

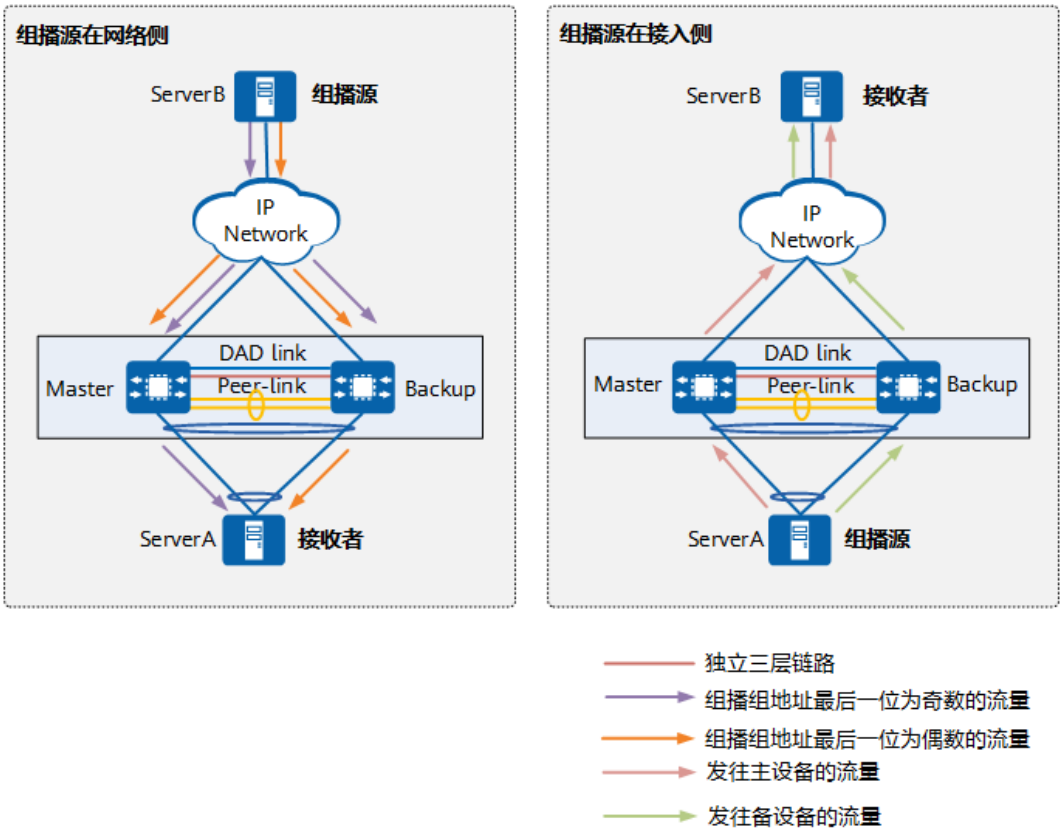
组播源在网络侧的场景下，在ServerB作为组播源、ServerA作为组播组成员时，M-LAG主备设备都从组播源引流，且按照以下规则由M-LAG主备设备在本地查找组播表后将流量负载分担转发至组播组成员：

- 若组播组地址最后一位为奇数（例如225.1.1.1），则由M-LAG成员口状态为主的设备转发至组播组成员；
- 若组播组地址最后一位为偶数（例如225.1.1.2），则由M-LAG成员口状态为备的设备转发至组播组成员；

组播源在接入侧的场景下，在ServerA作为组播源、ServerB作为组播组成员，且M-LAG系统无下挂其他组播组成员时，组播源发出的流量负载分担到M-LAG主备设备，收到流量后在本地查找组播表将报文发送出去。

区别于单播流量，由组播流量转发示意图可以看出，M-LAG系统在转发组播流量时需要在M-LAG两台设备间配置一条独立三层链路。因为在故障场景下，可能出现网络侧只有单链路上行，此时M-LAG主备设备间部署一条独立的单独三层链路可以用来传输组播报文。

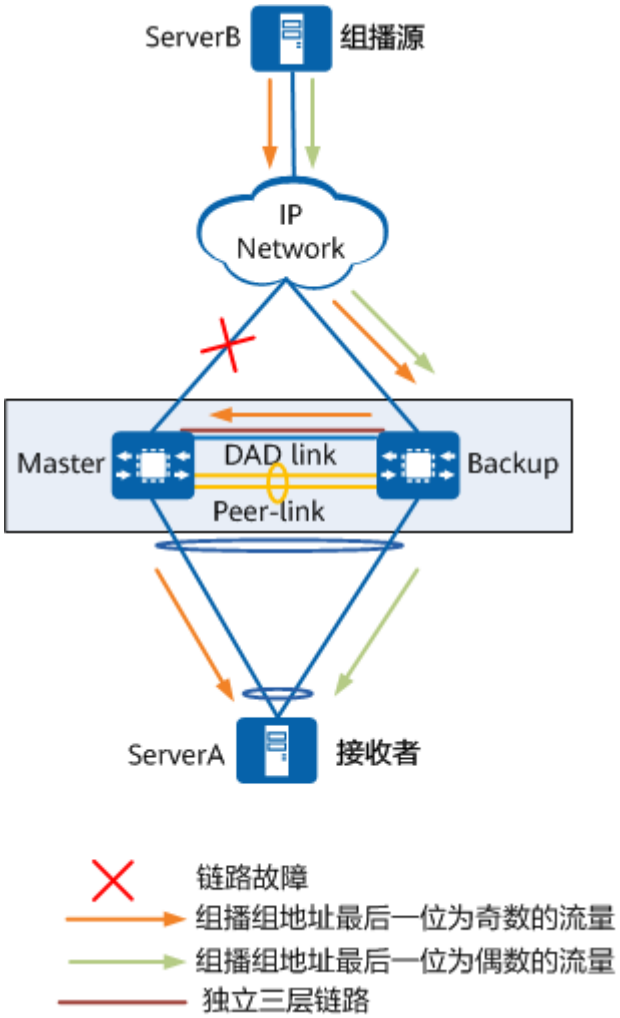
图 1-19 M-LAG 接入三层网络组播流量转发示意图



如图1-20所示，在网络侧设备连接到M-LAG备设备场景下，由peer-link接口转发的组播报文由于单向隔离无法转发至指定的M-LAG成员口，此时有两种方式将组播地址最后一位为奇数的组播报文转发至M-LAG成员口状态为主的设备：

- 组播报文通过独立三层链路转发至主设备。
- 配置VLAN及其对应的VLANIF接口，并且该VLAN只能包括peer-link接口，从而让组播报文通过peer-link链路转发至主设备。

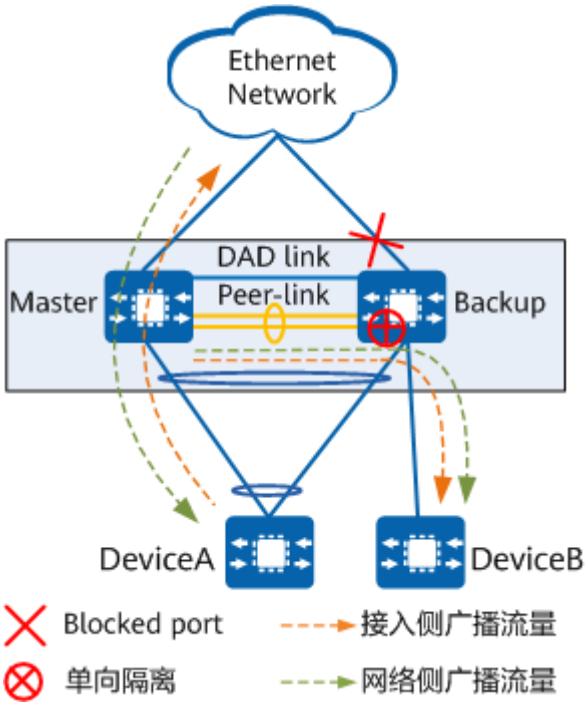
图 1-20 M-LAG 接入三层网络（单链路上行）组播流量转发示意图



广播流量转发

M-LAG上行接入二层网络，那么二层网络必须要保证从M-LAG跨设备的Eth-Trunk接口发出的流量只有一份，即必须保证一份流量只能通过一个M-LAG成员口发送给双归接入M-LAG的接入侧设备，否则会造成发送双份流量的问题。此处以M-LAG主设备的转发为例，如图1-21所示，假设右侧M-LAG上行接口被STP协议阻塞，M-LAG主设备收到广播流量后向各个下一跳转发，网络侧广播流量会转发给DeviceA、M-LAG备设备，接入侧广播流量会转发给网络侧设备、M-LAG备设备，当流量通过peer-link到达M-LAG备设备时，由于peer-link与M-LAG成员接口存在单向隔离机制（即从peer-link口进来的流量不会再从M-LAG口转发出去），所以到达M-LAG备设备的流量不会向DeviceA转发。

图 1-21 M-LAG 接入二层网络广播流量转发示意图



1.12 M-LAG 流量转发机制（主备模式）

M-LAG建立成功后，流量从M-LAG主成员口转发，M-LAG备成员口不转发流量。下面介绍在正常工作情况下M-LAG的流量转发机制。

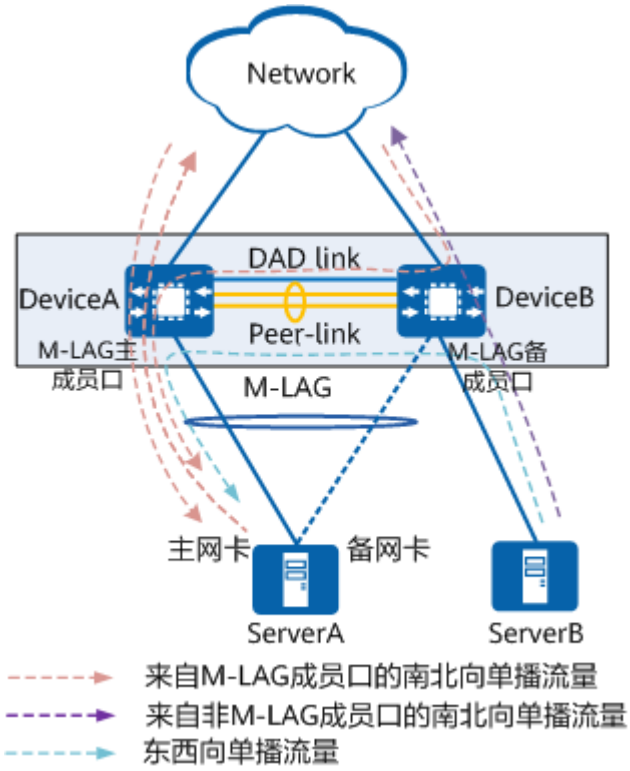
单播流量转发

如图1-22所示，M-LAG系统在服务器主备网卡接入场景下的已知单播流量转发：

对于南北向单播流量，在M-LAG接入侧，M-LAG设备DeviceA接收到服务器主网卡发送的流量并基于路由表转发至网络侧；在M-LAG网络侧，M-LAG设备接收到网络侧发送的流量，到达设备DeviceA的流量通过M-LAG主成员口转发至服务器，到达DeviceB的流量通过peer-link口绕行到设备DeviceA，再通过M-LAG主成员口转发至服务器。

对于东西向单播流量，在全部组建M-LAG，没有非M-LAG成员口的场景下，二三层流量均通过M-LAG主成员口转发。

图 1-22 M-LAG 已知单播流量转发示意图



组播流量转发

M-LAG接入二层网络

M-LAG上行接入二层网络，那么二层网络必须要保证发往M-LAG的流量只有一份，否则会有成环的风险。如图1-23所示，服务器ServerA的主网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接，假设M-LAG成员口状态为主的设备DeviceA对应的上行接口被STP协议阻塞。

组播源在网络侧的场景下，在ServerB作为组播源、ServerA作为组播组成员时，只有M-LAG成员口状态为主的设备可以转发组播流量，在网络侧只引流一份流量的情况下，接收到流量的DeviceB通过peer-link链路绕行到对端M-LAG设备DeviceA的M-LAG成员口。

如果设备DeviceA的M-LAG主成员口故障，M-LAG主成员口所在链路状态变为Down，DeviceA感知端口故障将M-LAG主成员口降备，并通知对端设备DeviceB，DeviceB收到消息将M-LAG备成员口升主；另外，当服务器的主网卡检测到链路故障，将备网卡升主并发送选主报文，DeviceB的M-LAG备成员口收到服务器发送的选主报文也会将M-LAG备成员口升主。此时，组播流量如图1-24所示，切换到M-LAG系统另一条设备DeviceB的成员口进行转发。

如果M-LAG成员口状态为备的设备DeviceB对应的上行接口被STP协议阻塞，接收到流量的设备DeviceA直接转发到本地的M-LAG成员口。

组播源在接入侧的场景下，在ServerA作为组播源、ServerB作为组播组成员时，组播源的流量发送至M-LAG状态为主的设备DeviceA，由于M-LAG设备DeviceA的上行接口被阻塞，那么设备DeviceA的组播出接口指向peer-link链路。如果设备DeviceA的M-LAG主成员口故障，此时组播流量切换到M-LAG系统另一条设备DeviceB的成员口进行转发。

图 1-23 M-LAG 接入二层网络组播流量转发示意图

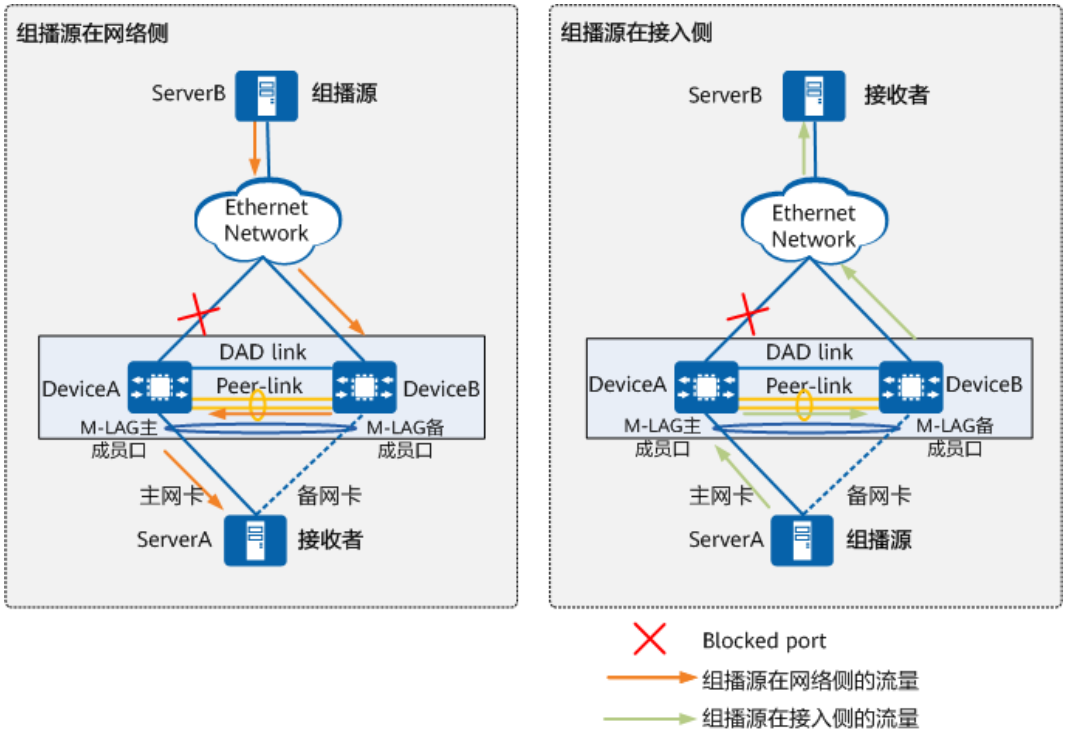
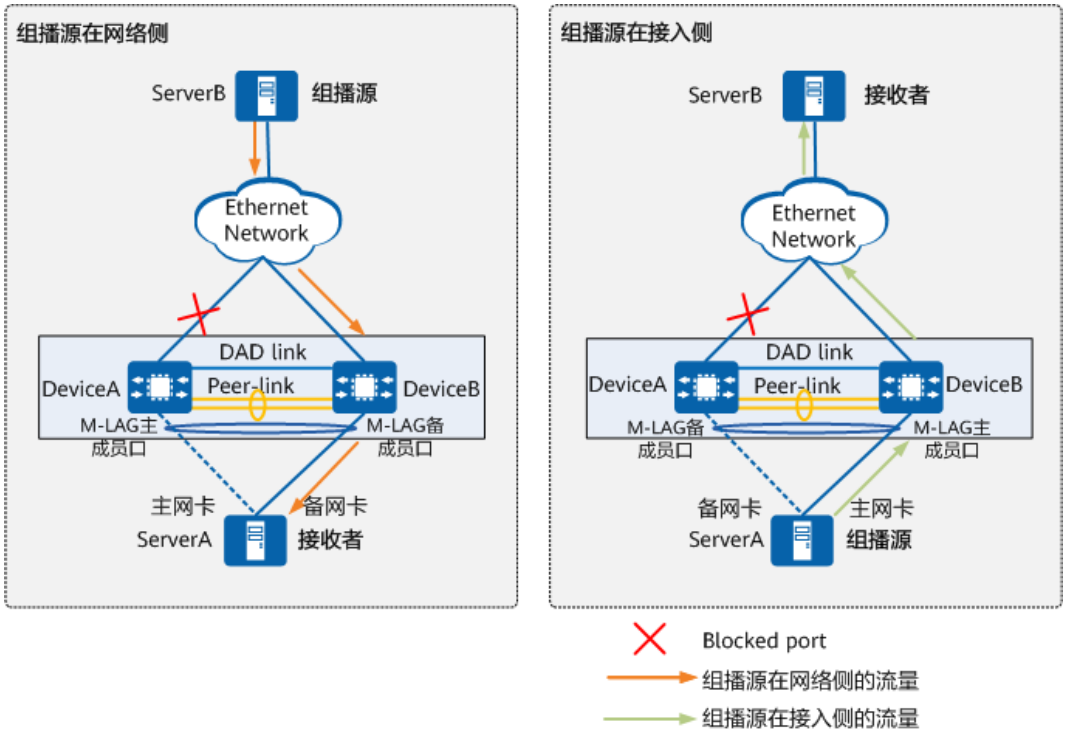


图 1-24 M-LAG 接入二层网络组播流量成员口故障转发示意图



M-LAG接入三层网络

M-LAG上行接入三层网络，M-LAG系统成员设备需要支持二三层组播混跑。如图1-25所示，M-LAG系统在服务器主备网卡接入场景下的组播流量转发，服务器ServerA的主

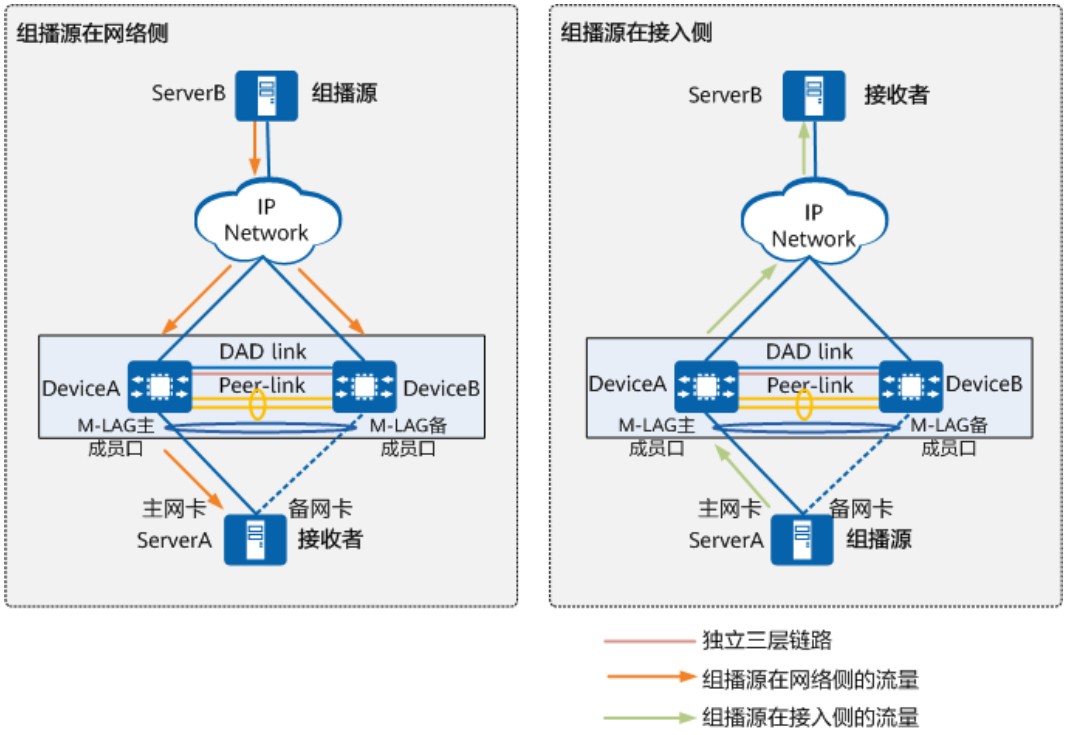
网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接：

组播源在网络侧的场景下，在ServerB作为组播源、ServerA作为组播组成员时，M-LAG主备设备都从组播源引流，且按照以下规则由M-LAG设备在本地查找组播表后将组播流量转发至组播组成员：

- M-LAG成员口状态为主的设备收到组播流量后，直接转发至组播组成员；
- M-LAG成员口状态为备的设备收到组播流量后，M-LAG备成员口出方向的组播流量被直接丢弃。

组播源在接入侧的场景下，在ServerA作为组播源、ServerB作为组播组成员，且M-LAG设备无下挂其他组播组成员时，组播源发送流量到M-LAG设备DeviceA，DeviceA收到流量后在本地查找组播表将报文发送出去。

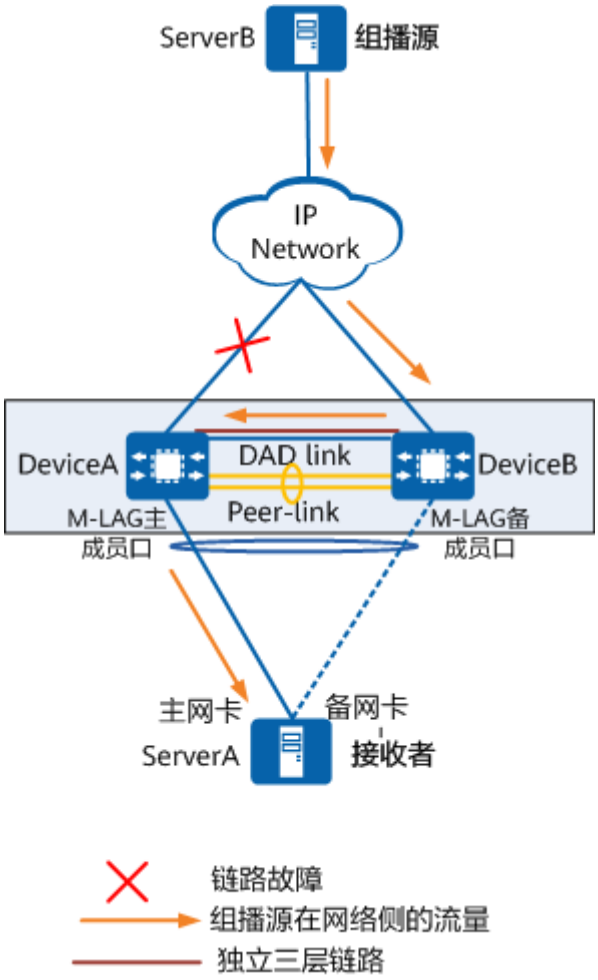
图 1-25 M-LAG 接入三层网络组播流量转发示意图



类似于双活模式，主备模式下M-LAG系统在转发组播流量时需要在M-LAG两台设备间配置一条独立三层链路。因为在故障场景下，可能出现网络侧只有单链路上行，此时M-LAG两台设备间部署一条独立的单独三层链路可以用来传输组播报文。如图1-26所示，在网络侧设备连接到M-LAG成员口状态为备的设备DeviceB场景下，由于M-LAG备成员口出方向的组播流量被直接丢弃，无法转发至对端设备的M-LAG成员口，此时有两种方式将组播报文转发至对端M-LAG成员口状态为主的设备：

- 组播报文通过独立三层链路转发至M-LAG成员口状态为主的设备。
- 配置VLAN及其对应的VLANIF接口，并且该VLAN只能包括peer-link接口，从而让组播报文通过peer-link链路转发至M-LAG成员口状态为主的设备。

图 1-26 M-LAG 接入三层网络（单链路上行）组播流量转发示意图

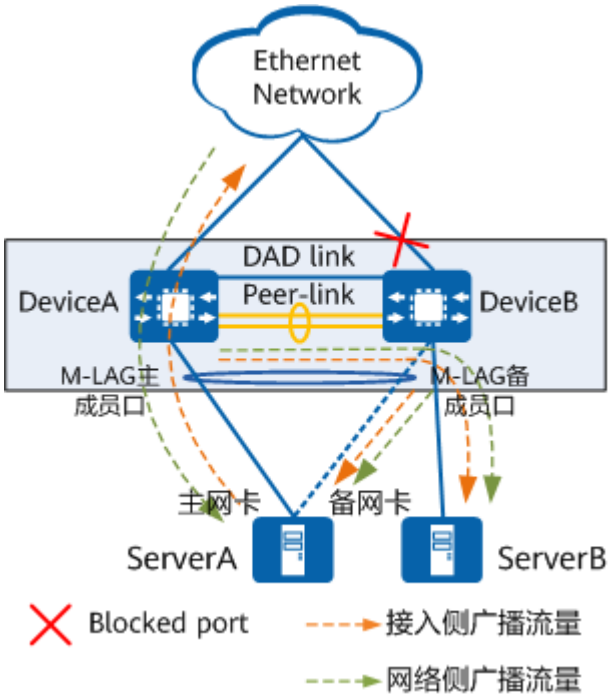


广播流量转发

如图1-27所示，服务器ServerA的主网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接。假设右侧M-LAG设备DeviceB上行接口被STP协议阻塞，M-LAG设备DeviceA收到广播流量后向各个下一跳转发，网络侧广播流量会转发给ServerA、DeviceB，接入侧广播流量会转发给网络侧设备、DeviceB。

当流量通过peer-link链路广播到对端M-LAG设备时，由于主备模式下M-LAG系统不会下发对应M-LAG成员口的单向隔离配置，流量将会从M-LAG备成员口转发至ServerA的备网卡，但ServerA的备网卡收到流量后不会进行处理，所以不会形成环路。

图 1-27 M-LAG 广播流量转发示意图

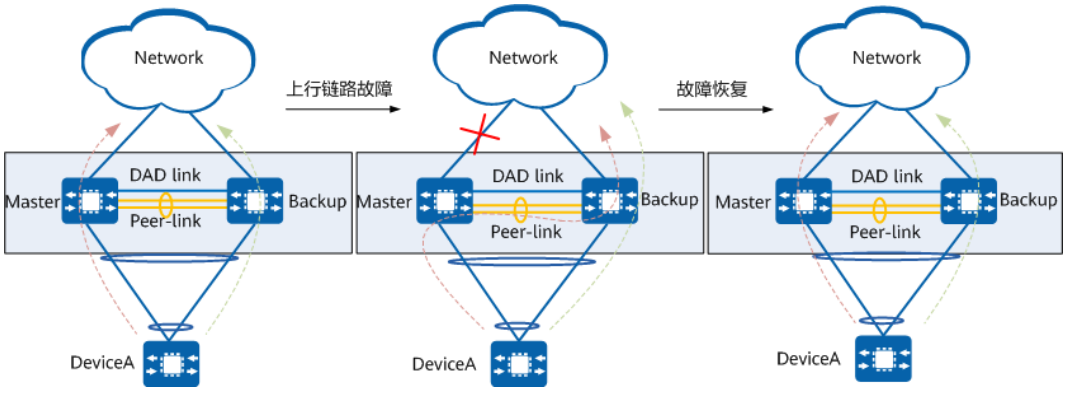


1.13 M-LAG 故障及恢复（双活模式）

M-LAG双活模式下的故障场景包括上行链路故障、下行链路故障、M-LAG设备故障、心跳故障、peer-link故障和二次故障。下面介绍流量在M-LAG故障和故障恢复场景下的转发情况和处理机制。

上行链路故障

图 1-28 上行链路故障组网示意图

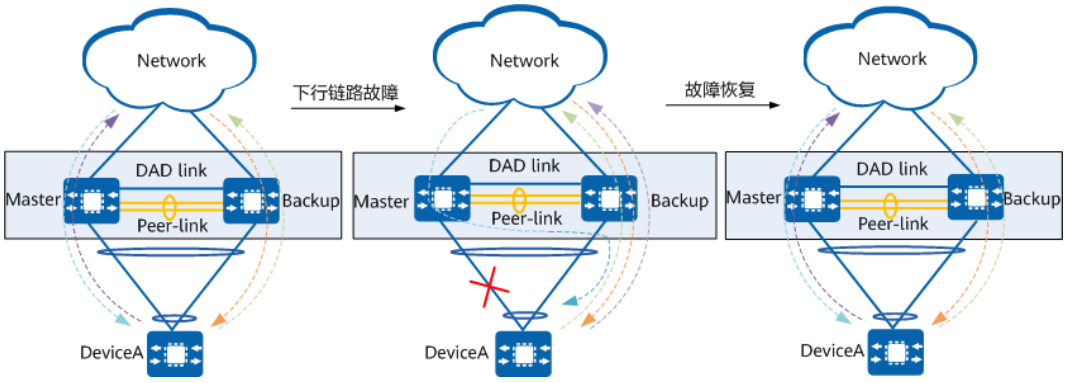


如图1-28所示，M-LAG接入普通以太网场景，M-LAG主设备的上行链路故障，通过M-LAG主设备的流量经过peer-link链路进行转发。M-LAG主设备上行链路故障恢复后，流量也恢复从主设备转发到网络侧。

三层场景下，需要在M-LAG主备设备之间配置逃生链路，否则到达Master设备的上行流量无法通过peer-link链路到达Backup设备。

下行链路故障

图 1-29 下行链路故障组网示意图



如图1-29所示，当M-LAG主成员口故障时，所在的链路状态变为Down，此时M-LAG备成员口状态由备升主，双归场景变为单归场景。在M-LAG主成员口故障的同时，主设备学习到的DeviceA侧MAC不会被清除，直接刷新MAC表的出端口指向peer-link口，实现流量快速切换，避免未知单播泛洪。

在故障M-LAG成员口恢复后，MAC表的出端口从peer-link指向M-LAG成员口，实现流量快速切换，避免未知单播泛洪。同时，为避免M-LAG成员口状态切换造成的某些协议振荡，M-LAG成员口主备状态不再回切，即由备升主的M-LAG成员口状态仍为主，原M-LAG主成员口在故障恢复后状态为备。

在M-LAG成员口故障，设备双归变单归场景下：

- 对于P系列单板、SAN系列单板、J系列单板、CE6866、CE6860-SAN、CE6860-HAM、CE6866K、CE8851-32CQ8DQ-P、CE8850-SAN、CE8850-HAM、CE8851K、CE8855、CE8851-32CQ4BQ、CE6855-48XS8CQ、CE6885、CE6885-T、CE6885-LL普通转发模式、CE6885-SAN、CE6885-LL低时延模式、CE6863E-48S8CQ：默认对报文出端口为M-LAG成员接口的所有ARP表项、ND表项、静态路由表项和动态路由表项申请备份的FRR资源，使得出接口指向peer-link口并形成主备路径下发，将表项的下一跳由M-LAG成员口切换为peer-link口，从而提升故障场景下的切换性能。
- 对于E系列单板、EK系列单板、CE6820H、CE6820H-K、CE6820S、CE6863H、CE6863H-K、CE6881H、CE6881H-K：在使能M-LAG三层转发增强功能后，将对报文出端口为M-LAG成员接口的所有ARP表项、ND表项、静态路由表项和动态路由表项申请备份的FRR资源，使得出接口指向peer-link口并形成主备路径下发，将表项的下一跳由M-LAG成员口切换为peer-link口，从而提升故障场景下的切换性能。

对于组播源在网络侧，组播成员在接入侧的组播流量，当M-LAG主设备的M-LAG成员口故障时，通过M-LAG同步报文通知对端设备进行组播表项刷新，M-LAG主备设备不再按照组播地址奇偶进行负载分担，而是所有组播流量都由端口状态Up的M-LAG备设备进行转发，反之亦然。

M-LAG 设备故障

图 1-30 M-LAG 主设备故障组网示意图

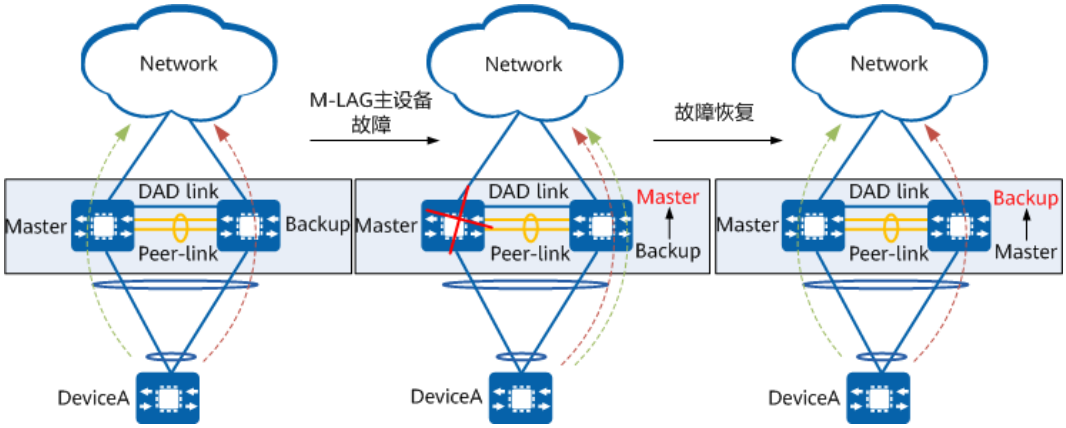
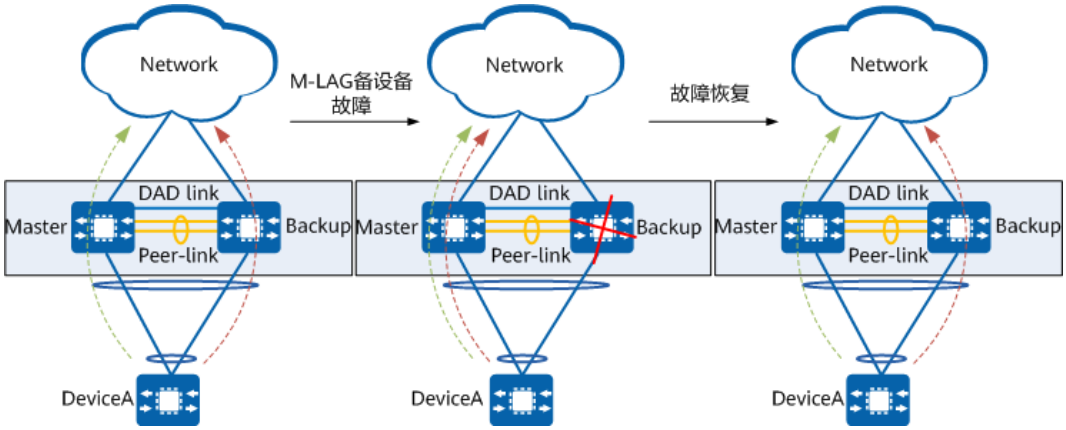


图 1-31 M-LAG 备设备故障组网示意图

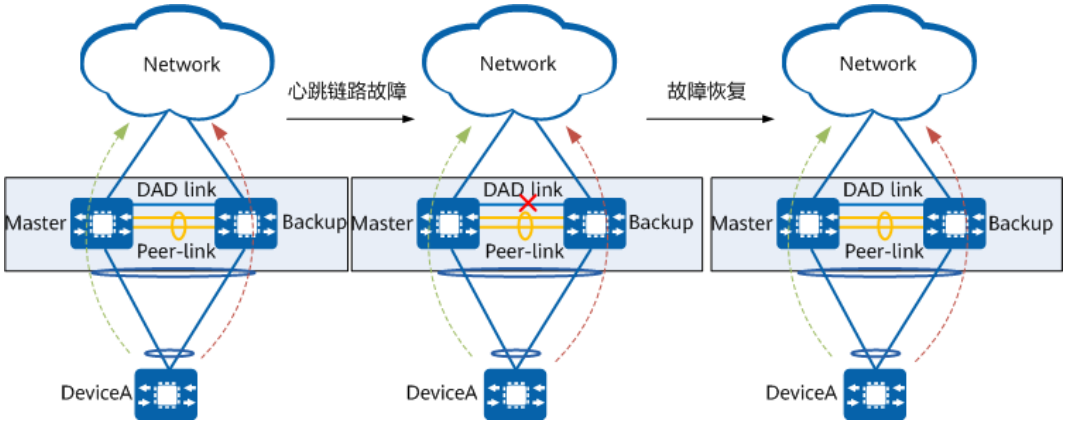


如图1-30所示，M-LAG主设备故障，M-LAG备设备将升级为主。原主设备侧M-LAG成员口链路状态变为Down，双归场景变为单归场景。如图1-31所示，M-LAG备设备故障，M-LAG的主备状态不会发生变化，M-LAG备设备侧成员口链路状态变为Down。M-LAG主设备侧成员口链路状态仍为Up，流量转发状态不变，双归场景变为单归场景。

M-LAG设备故障恢复时，peer-link先UP，DFS状态重新协商，M-LAG成员口恢复UP，流量恢复负载分担。M-LAG主设备恢复后设备状态仍然为主，M-LAG备设备恢复后设备状态仍然为备。

心跳链路故障

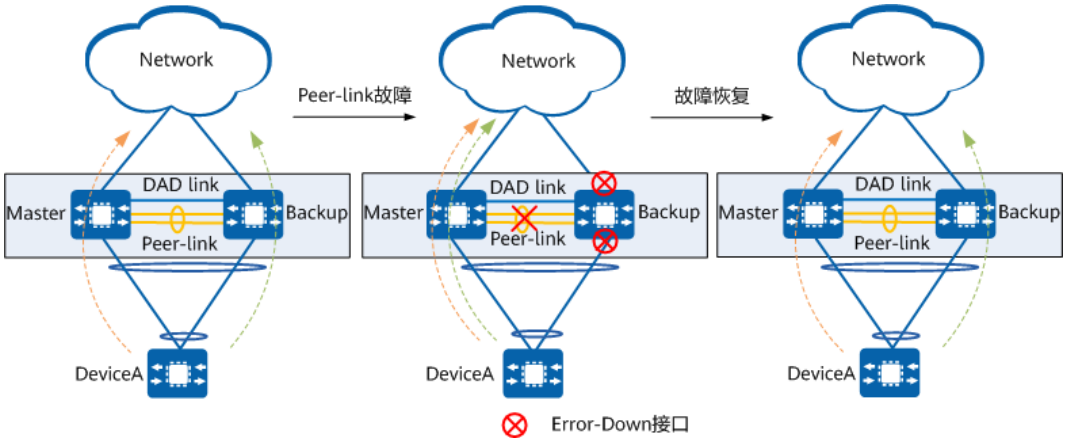
图 1-32 心跳链路故障组网示意图



心跳链路是用来处理peer-link故障时检测M-LAG系统是否是双主，但如果此时peer-link故障，则会因无法进行双主检测，出现双主，导致流量转发异常，如图1-32所示。两种情况都会产生心跳故障告警。心跳链路故障恢复后，产生心跳故障恢复告警。

peer-link 故障

图 1-33 peer-link 故障组网示意图



当peer-link故障但双主检测心跳状态正常时，在双主检测延时时间（缺省值为3s）后，会触发一端M-LAG设备上除逻辑端口、管理网口和peer-link接口以外的其他接口处于Error-Down状态，只保证另一端M-LAG设备正常流量转发。M-LAG系统按照如下先后顺序判断触发哪一端M-LAG设备的接口Error-Down：

1. 是否存在Up状态的上行口：若一端M-LAG设备的上行口全部为Down状态，且另一端M-LAG设备存在Up状态的上行口，则对上行口全部为Down状态的M-LAG设备触发端口Error-Down操作。
2. （仅框式设备涉及）peer-link接口所在接口板是否全部故障：若peer-link链路为直连聚合链路，一端M-LAG设备的peer-link接口所在接口板全部故障，且另一端M-LAG设备的peer-link接口所在接口板未全部故障，则对peer-link接口所在接口板全部故障的M-LAG设备触发端口Error-Down操作。

3. 带宽通量差值大小：若一端M-LAG设备计算出的带宽通量差值比另一端M-LAG设备的带宽通量差值更大，则对带宽通量差值更大的那一端M-LAG设备触发端口Error-Down操作。

   DFS配对成功后，M-LAG设备默认每间隔10s统计一次带宽通量；当触发双主检测时，会同时触发M-LAG设备统计此时的带宽通量。

   带宽通量差值计算公式：带宽通量差值=上一次统计到的带宽通量-触发双主检测时统计到的带宽通量，且每次统计带宽通量时，不包含peer-link接口。

   如果某一端M-LAG设备计算出的带宽通量差值为负值，则该M-LAG设备的带宽通量差值按照0处理。

4. 其他场景，如**图1-33**所示，则对M-LAG备设备触发端口Error-Down操作。

说明

建议将承载上行流量的接口配置为上行链路接口。如果设备没有配置上行链路接口，则按该设备上行口为Up状态处理。

对于框式设备，为了提升peer-link链路可靠性，需要跨版部署。

peer-link故障恢复时，处于Error Down状态的M-LAG成员口默认将在240s后自动恢复为Up状态，处于Error Down状态的其它接口将立即自动恢复为Up状态，流量恢复实现负载分担。

通过配置peer-link故障但双主检测心跳状态正常时触发Error-Down的端口包括逻辑端口，会触发M-LAG备设备上VLANIF接口、VBDIF接口、LoopBack接口以及M-LAG成员口处于ERROR DOWN状态。当peer-link故障恢复后，为保证大规格ARP同步正常，设备将在DFS Group配对成功后延迟6s恢复VLANIF接口、VBDIF接口、LoopBack接口为Up状态。此时，如果在接口下配置了接口三层协议状态延时Up时间，则VLANIF接口、VBDIF接口、LoopBack接口恢复Up状态的延迟时间为两者之和。

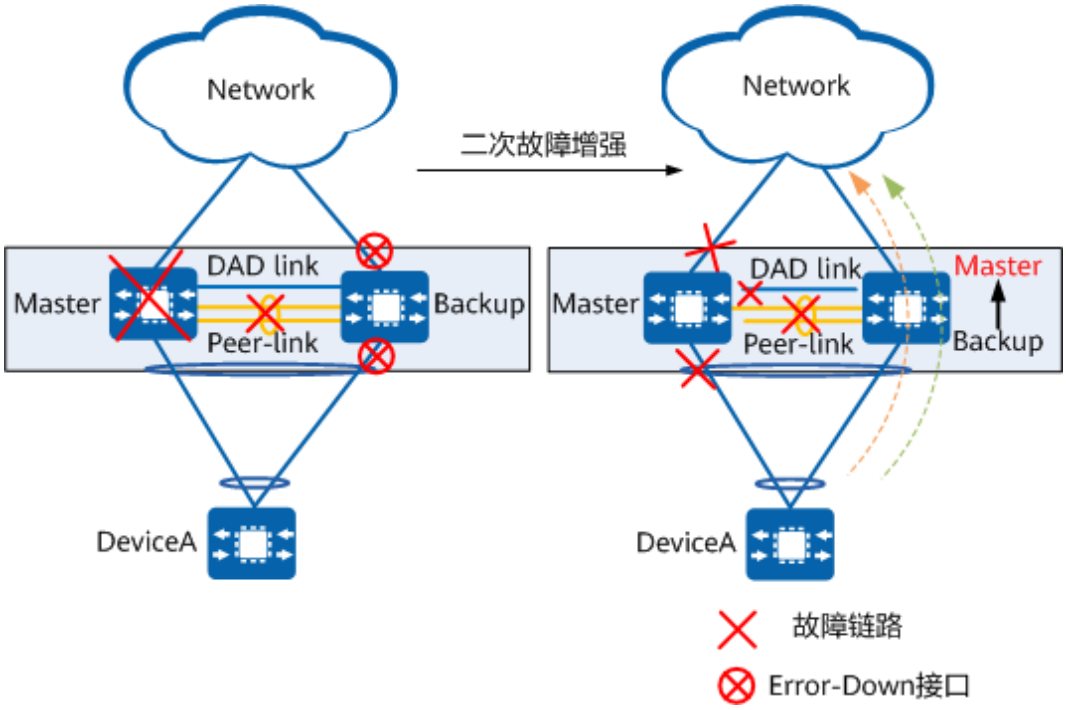
通过在端口下配置命令可以灵活配置某个端口在peer-link故障但双主检测心跳状态正常时是否将端口Error-Down。peer-link故障但双主检测正常场景下设备端口Error-Down对应情况如**表1-5**所示。

表 1-5 设备在 peer-link 故障但双主检测正常时接口 Error-Down 情况

设备配置情况	Error-Down接口类型
设备缺省情况	除逻辑端口、管理网口和peer-link接口以外的接口处于ERROR DOWN状态。
设备仅配置包括逻辑端口	VLANIF接口、VBDIF接口、LoopBack接口以及M-LAG成员口处于ERROR DOWN状态
设备仅配置suspend功能	仅M-LAG成员口以及配置该功能的接口处于ERROR DOWN状态。
设备仅配置reserved功能	除配置该功能的接口、逻辑端口、管理网口和peer-link接口以外的接口处于ERROR DOWN状态。
设备同时配置suspend功能和reserved功能	仅M-LAG成员口以及配置suspend功能的接口处于ERROR DOWN状态。

M-LAG 二次故障（peer-link 故障+M-LAG 设备故障）

图 1-34 M-LAG 二次故障组网示意图



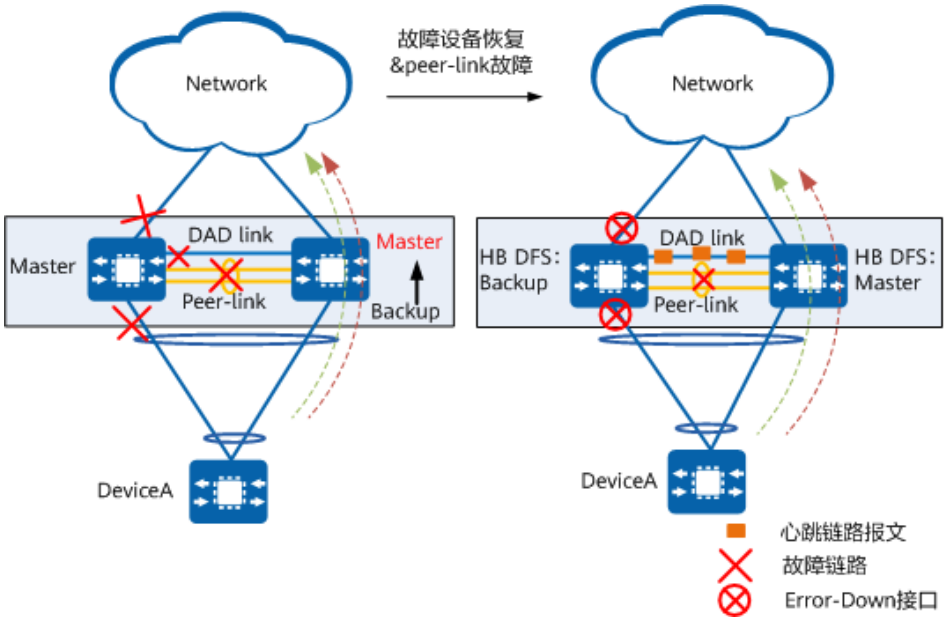
如图1-34所示，在M-LAG正常工作时，当peer-link故障但双主检测心跳状态正常且M-LAG主设备正常工作时（参见peer-link故障），M-LAG主设备又由于断电、主控板损坏、整机故障重启等其他故障导致主设备不能工作时，此时M-LAG主备设备皆不能正常转发流量。在该二次故障场景下，可以借助M-LAG二次故障增强来实现该故障场景下业务不中断的可靠性要求，下面通过M-LAG二次故障增强来说明不同的故障阶段和产生的行为：

1. 二次故障增强：在上述场景基础下，如果设备的二次故障增强功能已生效，则M-LAG备设备会借助M-LAG双主检测机制感知到M-LAG主设备故障（在一定周期内接收不到任何的M-LAG双主检测心跳报文）后，将升级为M-LAG主设备并恢复设备上处于ERROR DOWN状态的端口为Up状态，继续转发流量。

   若配置了双主检测的*peer-ip-address*参数，则二次故障增强功能生效；若未配置双主检测的*peer-ip-address*参数，则二次故障增强功能不生效。

2. 设备故障恢复：若原M-LAG主设备故障恢复后但peer-link链路仍故障
   - 若配置LACP M-LAG的系统ID在一定时间内切换为本设备的LACP系统ID，则在LACP协商时接入侧仅选择上行链路中的一条链路为活动链路，实际流量转发正常。
   - 若配置LACP M-LAG的系统ID为缺省情况，即系统ID不回切，M-LAG两台设备均使用同一系统ID来与接入侧设备协商，链路均能被选中成为活动链路。该场景下，由于peer-link链路仍然故障，M-LAG两端无法同步对端的优先级、系统MAC等信息，形成M-LAG两台设备双主的情况，可能导致流量异常。此时，如图1-35所示，可以借助心跳链路报文中携带必要的DFS Group协商主备的必要信息（如DFS Group优先级、系统MAC等）来协商M-LAG两台设备的HB（HeartBeat） DFS主备信息，触发HB DFS状态为备的设备上某些端口处于ERROR DOWN（端口Error-Down范围可以参见peer-link故障）状态，HB DFS状态为主的设备继续工作。

图 1-35 M-LAG 二次故障场景下设备故障恢复组网示意图



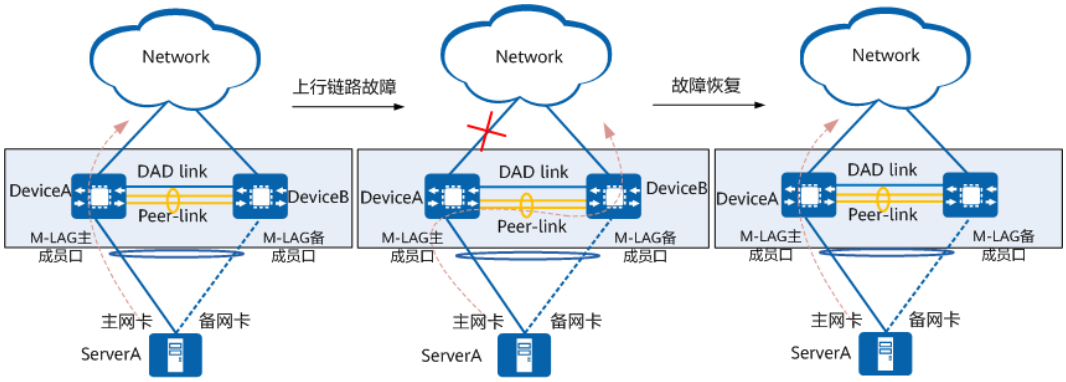
若在peer-link故障后，二次故障的设备为M-LAG备设备，则此时不会对流量转发行为产生影响，仍由M-LAG主设备进行流量转发。

1.14 M-LAG 故障及恢复（主备模式）

主备模式仅适用于服务器主备网卡接入M-LAG场景，故障场景包括上行链路故障、下行链路故障、M-LAG设备故障、心跳故障、peer-link故障和二次故障。下面介绍流量在M-LAG故障和故障恢复场景下的转发情况和处理机制。

上行链路故障

图 1-36 上行链路故障组网示意图



如图1-36所示，服务器ServerA的主网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接。M-LAG接入普通以太网场景，DeviceA上行链路故障，通过DeviceA的流量经过peer-link链路进行转发。DeviceA上行链路故障恢复后，流量也恢复从DeviceA转发到网络侧。

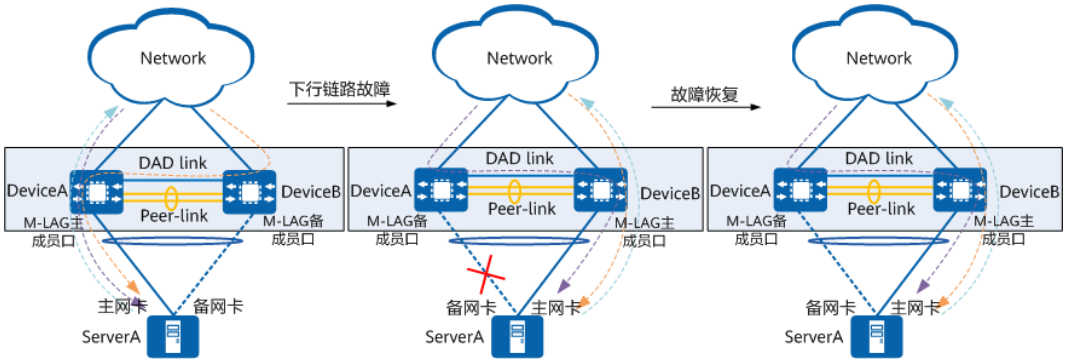
若DeviceB上行链路故障，由于DeviceB的M-LAG备成员口与ServerA的备网卡连接，不参与流量转发，所以对于M-LAG正常工作没有影响。

当故障的上行链路恰好为双主检测链路，此时对于M-LAG正常工作没有影响。一旦peer-link也发生故障，M-LAG出现双主冲突，双主检测又无法进行，则会出现丢包现象。

三层场景下，需要在M-LAG两台设备之间配置逃生链路，否则到达设备DeviceA的上行流量无法通过peer-link链路到达设备DeviceB。

下行链路故障

图 1-37 下行链路故障组网示意图



如图1-37所示，服务器ServerA的主网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接。

当M-LAG主成员口故障时，所在的链路状态变为Down，DeviceA感知端口故障将M-LAG主成员口降备，并通知对端设备DeviceB，DeviceB收到消息将M-LAG备成员口升主；另外，当服务器的主网卡检测到链路故障，将备网卡升主并发送选主报文，DeviceB的M-LAG备成员口收到服务器发送的选主报文也会将M-LAG备成员口升主。在M-LAG主成员口故障的同时，DeviceA学习到的Server侧MAC不会被清除，直接刷新MAC表的出端口指向peer-link口；DeviceB将M-LAG备成员口升主后，也会直接刷新MAC表的出接口由peer-link口指向M-LAG成员口，实现流量快速切换，避免未知单播泛洪。

在故障M-LAG成员口恢复后，为避免M-LAG主备成员状态切换造成的某些协议振荡，设备支持M-LAG成员口状态默认不主动回切。M-LAG成员口主备状态是否回切，依赖于原主网卡检查到链路故障恢复后是否升主。如果原主网卡恢复为主，重新发送选主报文进行M-LAG成员口主备选举，则主备状态回切。

若M-LAG备成员口故障，由于M-LAG备成员口与ServerA的备网卡连接，不参与流量转发，所以对于M-LAG正常工作没有影响。

在M-LAG成员口故障，设备双归变单归场景下，默认对于报文出端口为M-LAG成员口的所有ARP表项、ND表项、静态路由表项和动态路由表项申请备份的FRR资源，使得出接口指向peer-link口并形成主备路径下发，将表项的下一跳由M-LAG成员口切换为peer-link口，从而提升故障场景下的切换性能。

对于组播源在网络侧，组播成员在接入侧的组播流量，当M-LAG主成员口故障时，通过M-LAG同步报文通知对端设备（即M-LAG成员口状态为备的设备）进行组播表项刷新，实现组播流量快速切换到对端M-LAG设备转发。

M-LAG 设备故障

图 1-38 M-LAG 成员口状态为主的设备故障组网示意图

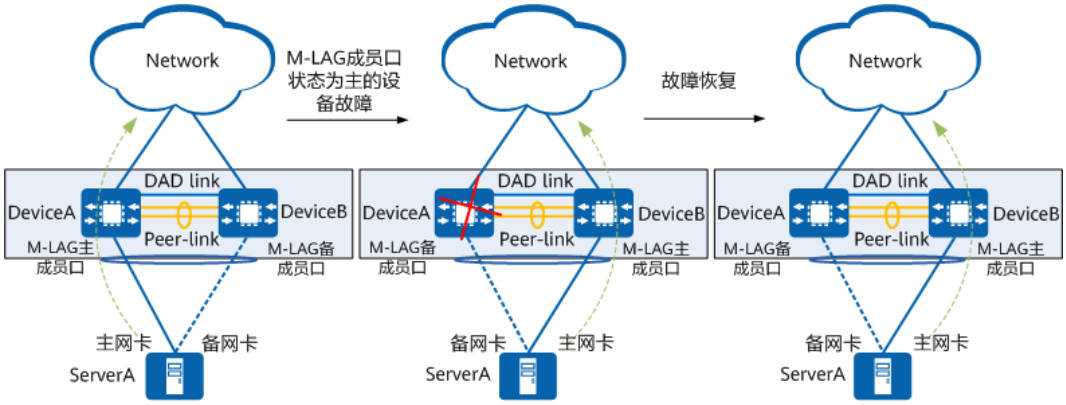
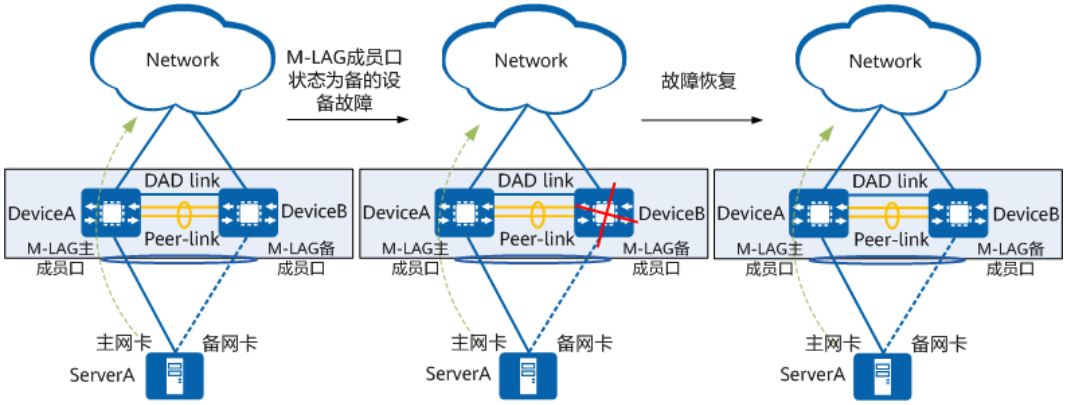


图 1-39 M-LAG 成员口状态为备的设备故障组网示意图



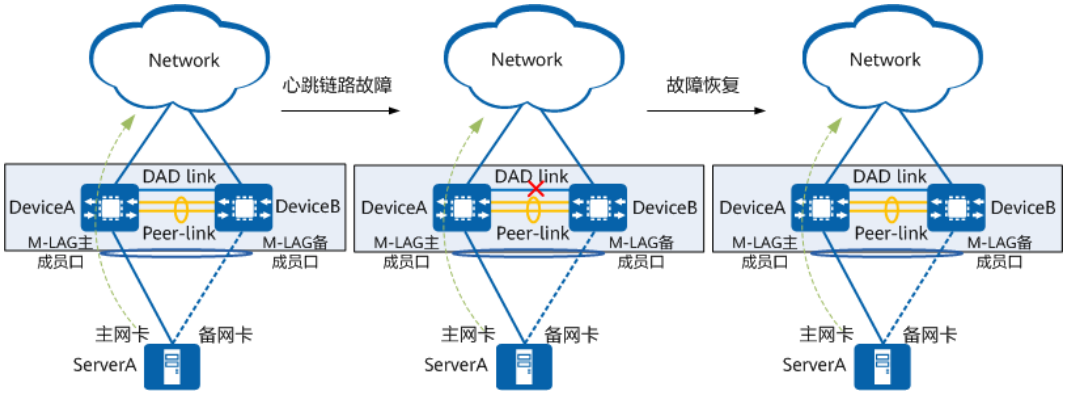
如图1-38所示，M-LAG成员口状态为主的设备DeviceA故障，DeviceA侧M-LAG成员口链路状态变为Down，对端设备DeviceB将升级为M-LAG主设备（如果DeviceB原来已经是M-LAG主设备，则M-LAG的主备状态不涉及变化），DeviceB检测到peer-link链路故障后，DeviceB侧M-LAG备成员口会升主；另外，当服务器的主网卡检测到链路故障，将备网卡升主并发送选主报文，DeviceB的M-LAG备成员口收到服务器发送的选主报文也会将M-LAG备成员口升主。

如图1-39所示，M-LAG成员口状态为备的设备DeviceB故障，DeviceB侧M-LAG成员口链路状态变为Down，对端设备DeviceA将升级为M-LAG主设备（如果DeviceA原来已经是M-LAG主设备，则M-LAG的主备状态不涉及变化），M-LAG成员口主备状态也不会发生变化。DeviceA侧M-LAG主成员口状态仍为UP，流量转发状态不变。

M-LAG设备故障恢复时，peer-link先UP，DFS状态重新协商，M-LAG成员口恢复UP。M-LAG主设备恢复后设备状态仍然为主，M-LAG备设备恢复后设备状态仍然为备。原M-LAG成员口状态为主的设备DeviceA故障恢复时，M-LAG成员口状态默认不主动回切。M-LAG成员口主备状态是否回切，依赖于原主网卡检查到链路故障恢复后是否升主。如果原主网卡恢复为主，重新发送选主报文进行M-LAG成员口主备选举，则主备状态回切。

心跳链路故障

图 1-40 心跳链路故障组网示意图



心跳链路是用来处理peer-link故障时检测M-LAG系统是否是双主，若心跳链路承载三层网络的业务，心跳故障对设备流量转发会有影响。若心跳链路承载二层业务或不承载三层业务，心跳故障对设备流量转发无影响；但如果此时peer-link故障，则会因无法进行双主检测，出现双主，导致流量转发异常，如图1-40所示，服务器ServerA的主网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接。两种情况都会产生心跳故障告警。心跳链路故障恢复后，产生心跳故障恢复告警。

peer-link 故障

图 1-41 peer-link 故障组网示意图（M-LAG 成员口状态为主的设备为 M-LAG 备设备）

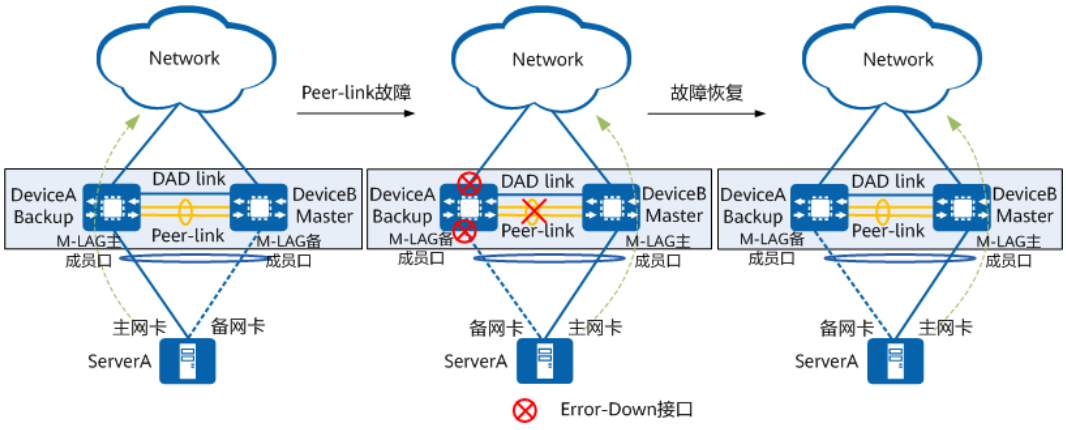
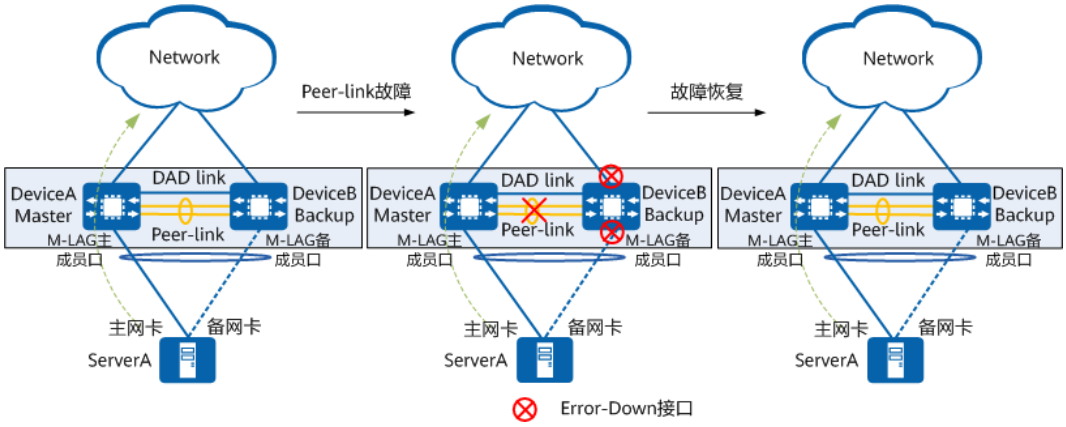


图 1-42 peer-link 故障组网示意图（M-LAG 成员口状态为主的设备为 M-LAG 主设备）



当peer-link故障但双主检测心跳状态正常时，在双主检测延时时间（缺省值为3s）后，会触发一端M-LAG设备上除逻辑端口、管理网口和peer-link接口以外的其他接口处于Error-Down状态，只保证另一端M-LAG设备正常流量转发。

1. 若一端M-LAG设备的上行口全部为Down状态，且另一端M-LAG设备存在Up状态的上行口，则对上行口全部为Down状态的M-LAG设备触发端口Error-Down操作。
2. （仅框式设备涉及）若两端M-LAG设备上行口状态相同，且peer-link链路为直连聚合链路，若一端M-LAG设备的peer-link接口所在接口板全部故障，且另一端M-LAG设备的peer-link接口所在接口板未全部故障，则对peer-link接口所在接口板全部故障的M-LAG设备触发端口Error-Down操作。
3. 带宽通量差值大小：若一端M-LAG设备计算出的带宽通量差值比另一端M-LAG设备的带宽通量差值更大，则对带宽通量差值更大的那一端M-LAG设备触发端口Error-Down操作。

   DFS配对成功后，M-LAG设备默认每间隔10s统计一次带宽通量；当触发双主检测时，会同时触发M-LAG设备统计此时的带宽通量。

   带宽通量差值计算公式：带宽通量差值=上一次统计到的带宽通量-触发双主检测时统计到的带宽通量，且每次统计带宽通量时，不包含peer-link接口。

   如果某一端M-LAG设备计算出的带宽通量差值为负值，则该M-LAG设备的带宽通量差值按照0处理。
4. 其他场景，则对M-LAG备设备触发端口Error-Down操作。
   - 如图1-41所示，服务器ServerA的主网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接，设备DeviceA为M-LAG备设备。当peer-link故障时，M-LAG的主备状态不会变化，但DeviceB检测到peer-link链路故障后，DeviceB侧M-LAG备成员口会升主；另外，当服务器的主网卡检测到链路故障，将备网卡升主并发送选主报文，DeviceB的M-LAG备成员口收到服务器发送的选主报文也会将M-LAG备成员口升主。
   - 如图1-42所示，服务器ServerA的主网卡与M-LAG设备DeviceA连接，服务器ServerA的备网卡与M-LAG设备DeviceB连接，设备DeviceB为M-LAG备设备。当peer-link故障时，M-LAG的主备状态不会变化，M-LAG成员口主备状态也不会发生变化。DeviceA侧M-LAG主成员口状态仍为UP，流量转发状态不变。

说明

建议将承载上行流量的接口配置为上行链路接口。如果设备没有配置上行链路接口，则按该设备上行口为Up状态处理。

对于框式设备，为了提升peer-link链路可靠性，需要跨版部署。

peer-link故障恢复时，处于Error Down状态的M-LAG成员口默认将在240s后自动恢复为Up状态，处于Error Down状态的其它接口将立即自动恢复为Up状态。Error Down状态的M-LAG成员口恢复成UP状态后，M-LAG成员口状态默认不主动回切。M-LAG成员口主备状态是否回切，依赖于原主网卡检查到链路故障恢复后是否升主。如果原主网卡恢复为主，重新发送选主报文进行M-LAG成员口主备选举，则主备状态回切。

通过配置peer-link故障但双主检测心跳状态正常时触发Error-Down的端口包括逻辑端口，会触发M-LAG备设备上VLANIF接口、VBDIF接口、LoopBack接口以及M-LAG成员口处于ERROR DOWN状态。当peer-link故障恢复后，为保证大规格ARP同步正常，设备将在DFS Group配对成功后延迟6s恢复VLANIF接口、VBDIF接口、LoopBack接口为Up状态。此时，如果在接口下配置了接口三层协议状态延时Up时间，则VLANIF接口、VBDIF接口、LoopBack接口恢复Up状态的延迟时间为两者之和。

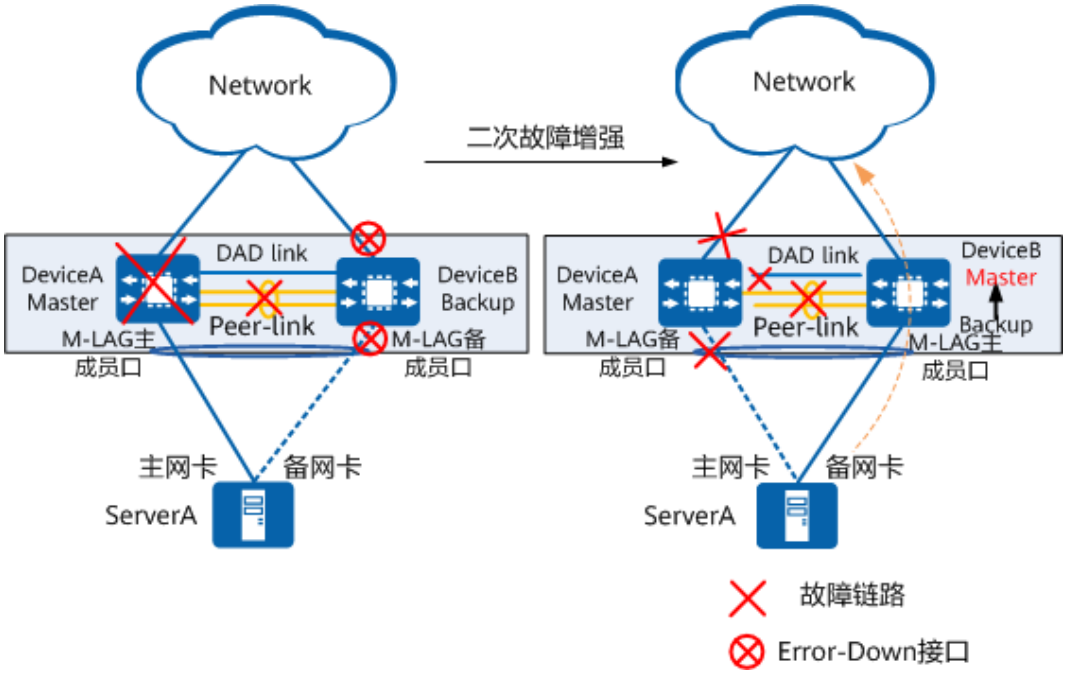
通过在端口下配置命令可以灵活配置某个端口在peer-link故障但双主检测心跳状态正常时是否将端口Error-Down。peer-link故障但双主检测正常场景下设备端口Error-Down对应情况如表1-6所示。

表 1-6 设备在 peer-link 故障但双主检测正常时接口 Error-Down 情况

设备配置情况	Error-Down接口类型
设备缺省情况	除逻辑端口、管理网口和peer-link接口以外的接口处于ERROR DOWN状态。
设备仅配置包括逻辑端口	VLANIF接口、VBDIF接口、LoopBack接口以及M-LAG成员口处于ERROR DOWN状态。
设备仅配置suspend功能	仅M-LAG成员口以及配置该功能的接口处于ERROR DOWN状态。
设备仅配置reserved功能	除配置该功能的接口、逻辑端口、管理网口和peer-link接口以外的接口处于ERROR DOWN状态。
设备同时配置suspend功能和reserved功能	仅M-LAG成员口以及配置suspend功能的接口处于ERROR DOWN状态。

M-LAG 二次故障（peer-link 故障+M-LAG 设备故障）

图 1-43 M-LAG 二次故障组网示意图



如图1-43所示，在M-LAG正常工作时，当peer-link故障但双主检测心跳状态正常且M-LAG主设备正常工作时（参见**peer-link故障**），M-LAG主设备又由于断电、主控板损坏、整机故障重启等其他故障导致主设备不能工作时，此时M-LAG主备设备皆不能正常转发流量。

在该二次故障场景下，可以借助M-LAG二次故障增强来实现该故障场景下业务不中断的可靠性要求，下面通过M-LAG二次故障增强来说明不同的故障阶段和产生的行为：

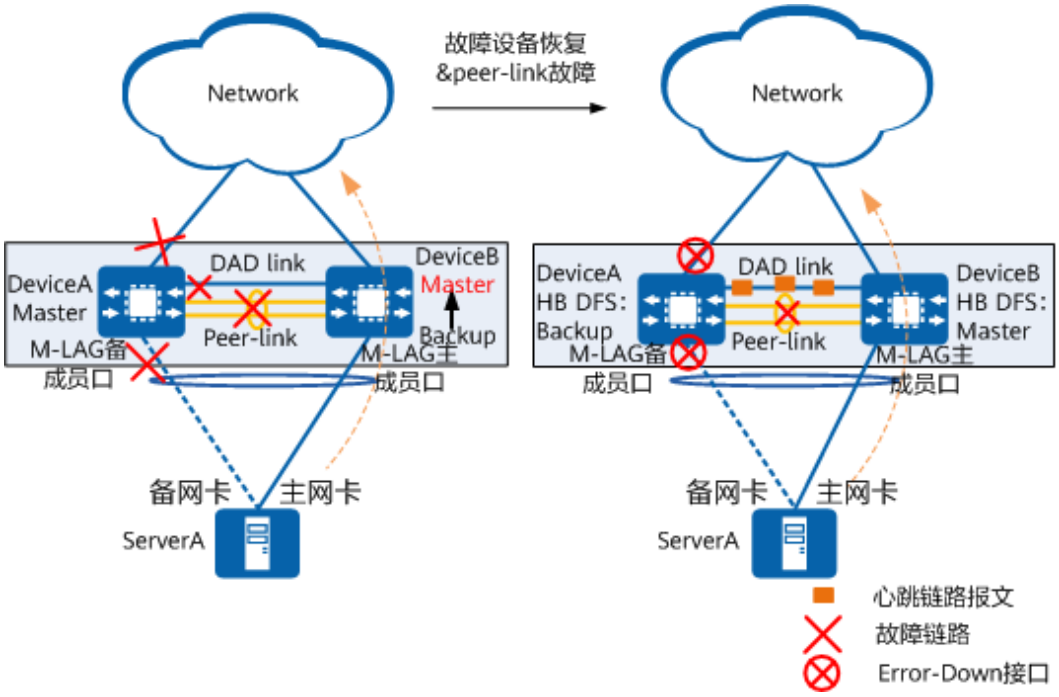
1. 二次故障增强：在上述场景基础下，如果设备的二次故障增强功能已生效，则M-LAG备设备会借助M-LAG双主检测机制感知到M-LAG主设备故障（在一定周期内接收不到任何的M-LAG双主检测心跳报文）后，将升级为M-LAG主设备并恢复设备上处于ERROR DOWN状态的端口为Up状态。升级为M-LAG主设备的M-LAG成员口恢复为UP状态后，检测到peer-link链路故障后将升主；另外，当服务器的主网卡检测到链路故障，将备网卡升主并发送选主报文，M-LAG备成员口收到服务器发送的选主报文也会将M-LAG备成员口升主。

   若配置了双主检测的*peer-ip-address*参数，则二次故障增强功能生效；若未配置双主检测的*peer-ip-address*参数，则二次故障增强功能不生效。

2. 设备故障恢复：若原M-LAG主设备故障恢复后但peer-link链路仍故障。

   原M-LAG主设备故障恢复时，由于peer-link链路仍然故障，M-LAG两端无法同步对端的优先级、系统MAC等信息，形成M-LAG两台设备双主的情况，可能导致流量异常。此时，如图1-44所示，由于M-LAG上行口已恢复为UP状态，心跳链路恢复正常，可以借助心跳链路报文中携带DFS Group协商主备的必要信息（如DFS Group优先级、系统MAC等）来协商M-LAG两台设备的HB（HeartBeat） DFS主备信息，触发HB DFS状态为备的设备上某些端口处于ERROR DOWN（端口Error-Down范围可以参见表1-6）状态，HB DFS状态为主的设备继续工作。

图 1-44 M-LAG 二次故障场景下设备故障恢复组网示意图



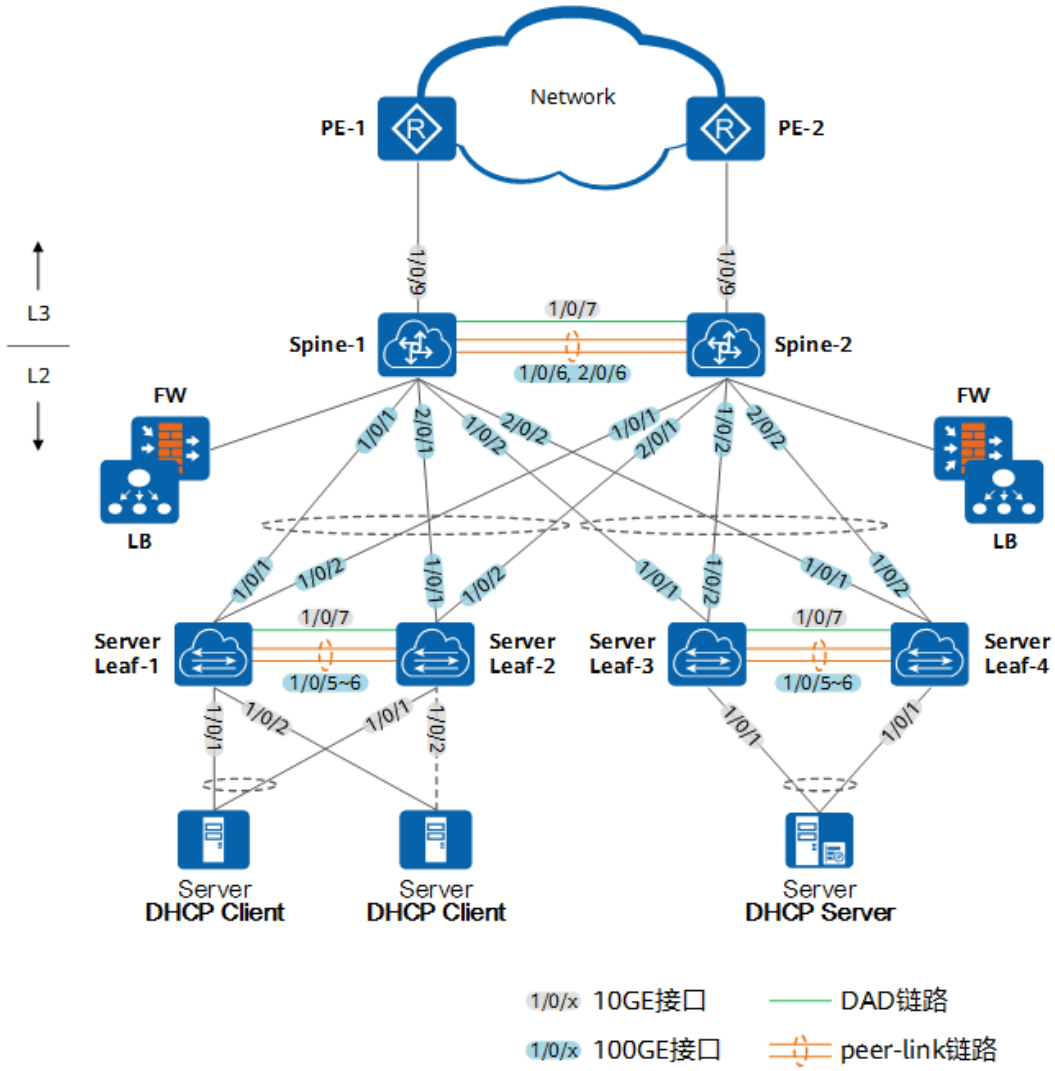
若在peer-link故障后，二次故障的设备为M-LAG备设备，则此时不会对流量转发行为产生影响，仍由M-LAG主设备进行流量转发。

2 典型配置案例

2.1 多级 M-LAG 互联

2.1.1 组网方案

图 2-1 多级 M-LAG 互联组网图



如图2-1所示，某网络采用Spine、Leaf之间M-LAG级联的方式互联，其中：

- Server Leaf部署M-LAG，用于服务器二层接入。
- Spine部署M-LAG，作为服务器的三层业务网关，与Sever Leaf之间使用M-LAG级联的方式互联。
- 防火墙（FW）及负载均衡器（LB）旁挂在Spine上。更多关于M-LAG设备组和防火墙、负载均衡器的对接描述请参考**2.5 M-LAG与防火墙对接**、**2.6 M-LAG与负载均衡器（LB）对接**。
- 服务器以负载分担或主备方式接入，服务器负载分担接入时，推荐M-LAG使用双活模式，服务器主备接入时，推荐M-LAG使用主备模式。
- 建议通过单独链路作为M-LAG设备之间双主检测链路，提升可靠性。
- 建议M-LAG设备通过静态LACP模式和服务器对接，同时需要配置LACP forced-up功能。当服务器重启或刚上线时，Eth-Trunk接口接受LACP协议报文超时后，Eth-Trunk成员口的状态会变为Down。这时可以通过配置成员口Force Up，使该接口继续转发业务流量，防止业务丢包。

- 两台设备组建M-LAG之后，可能产生环路，建议使用V-STP功能防止M-LAG成员口因为环路协议被堵塞。级联M-LAG场景下由于不支持根桥方式的M-LAG，必须使用V-STP方式配置M-LAG。如果M-LAG设备需要和友商PVST/PVST+协议对接，则推荐使用VBST协议和PVST/PVST+对接。
- 建议至少使用两条链路捆绑成Eth-Trunk链路作为M-LAG peer-link链路，如果是两台框式设备组建M-LAG，建议使用不同单板的接口捆绑成Eth-Trunk链路作为M-LAG peer-link链路。如果使用CE6881、CE6863、CE6881H和CE6863H设备组建M-LAG，建议选择（100GE 1/0/1~1/0/3和100GE 1/0/4~1/0/6）中的接口捆绑Eth-Trunk作为M-LAG peer-link链路。

2.1.2 配置 Server Leaf

配置概览

1. **配置M-LAG**
2. **配置Leaf与Spine互联链路**
3. **配置服务器接入**
4. **配置优化命令**

配置步骤

步骤1　配置M-LAG

ServerLeaf-1	ServerLeaf-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
stp bridge-address *1-1-2* # 	stp bridge-address *1-1-2* # 	配置当前设备参与生成树计算的桥MAC，两台M-LAG设备的桥MAC必须相同，建议选择其中一台设备的系统MAC作为共同桥MAC，不同M-LAG组里的设备桥MAC不同。
ip vpn-instance *DAD* ipv4-family route-distinguisher *3:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *4:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。

ServerLeaf-1	ServerLeaf-2	命令说明
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 #	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 #	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。 配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。

```
[~ServerLeaf-1] ping -vpn-instance DAD 192.168.10.2
  PING 192.168.10.2: 56  data bytes, press CTRL_C to break
    Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

  --- 192.168.10.2 ping statistics ---
    5 packet(s) transmitted
    5 packet(s) received
    0.00% packet loss
    round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~ServerLeaf-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW :
10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
    Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
    Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
    Input: 62343901 packets,93259180813 bytes
           42820517 unicast,132 broadcast,19523252 multicast
           0 errors,0 drops
    Output:48323130 packets,92501426032 bytes
           29723661 unicast,128 broadcast,18599341 multicast
           0 errors,0 drops
    Last 10 seconds input utility rate:  0.01%
    Last 10 seconds output utility rate: 0.01%
--------------------------------------------------------------
PortName                Status          Weight
--------------------------------------------------------------
100GE1/0/5              UP              1
100GE1/0/6              UP              1
--------------------------------------------------------------
```

```
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~ServerLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                   Working Mode: Static
Preempt Delay: Disabled         Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
--------------------------------------------------------------------------------
ActorPortName            Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/5(r)            Selected 100GE    32768   6      65      10111100  1
100GE1/0/6(r)            Selected 100GE    32768   6      65      10111100  1

Partner:
--------------------------------------------------------------------------------
ActorPortName            SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/5               32768    00fd-dffb-9a03  32768   6      65      10111100
100GE1/0/6               32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
[~ServerLeaf-1] display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance       : DAD
  State              : Backup
  Causation          : -
  System ID          : 00fd-fdfd-b703
  SysName            : ServerLeaf-2
  Version            : V300R022C00
  Device Type        : CE6800
```

步骤2 配置Leaf与Spine级联链路

ServerLeaf-1	ServerLeaf-2	命令说明
vlan batch *10 20*	vlan batch *10 20*	创建业务VLAN。

ServerLeaf-1	ServerLeaf-2	命令说明
interface *Eth-Trunk10* description *to-Spine* trunkport *100GE1/0/1* trunkport *100GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *10* #	interface *Eth-Trunk10* description *to-Spine* trunkport *100GE1/0/1* trunkport *100GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *10* #	配置级联链路。 按需放通VLAN。

步骤3 配置服务器接入

- 服务器负载分担方式接入时

ServerLeaf-1	ServerLeaf-2	命令说明
interface eth-trunk *1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *10GE1/0/1* storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* #	interface eth-trunk *1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *10GE1/0/1* storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* #	配置接入用的链路聚合组。 按需放通VLAN，不放通VLAN1，防止成环。 配置静态LACP模式。 配置边缘端口。 配置未知单播抑制，建议为端口带宽的5%。 配置组播报文抑制，建议为1 Mbit/s。 配置广播报文抑制，建议为1 Mbit/s。
interface *10GE1/0/1* lacp force-up #	-	（可选）在服务器需要进行PXE安装的场景，其中一台设备的成员口需要配置lacp force-up，该成员口为对接安装操作系统的网卡的接口。

- 服务器主备方式接入时

ServerLeaf-1	ServerLeaf-2	命令说明
interface *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *20* stp edged-port enable storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* #	interface *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *20* stp edged-port enable storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* #	配置服务器接入端口。 按需放通VLAN，不放通VLAN1，防止成环。 配置边缘端口。 配置未知单播抑制，建议为端口带宽的5%。 配置组播报文抑制，建议为1 Mbit/s。 配置广播报文抑制，建议为1 Mbit/s。

上述配置完成后，可以通过**display dfs-group 1 node 1 m-lag** [**brief**]命令查看M-LAG链路聚合组的状态。

```
[~ServerLeaf-1] display dfs-group 1 node 1 m-lag brief
* - Local node

M-Lag ID     Interface      Port State    Status              Consistency-check
      1     Eth-Trunk 1     Up           active(*)-active     success
      2     Eth-Trunk 2     Up           active(*)-inactive   success
     10     Eth-Trunk 10    Up           active(*)-active     success
```

可以通过**display mac-address** [**vlan** *vlan-id*]命令查看设备学习到的MAC地址信息。

```
[~ServerLeaf-1] display mac-address
Flags: * - Backup
       # - forwarding logical interface, operations cannot be performed based
           on the interface.
BD   : bridge-domain    Age : dynamic MAC learned time in seconds
--------------------------------------------------------------------------------
MAC Address    VLAN/VSI/BD   Learned-From        Type              Age
--------------------------------------------------------------------------------
9c7d-a378-3c8d 10/-/-        Eth-Trunk1          dynamic           319
cc64-a668-6814 20/-/-        Eth-Trunk2          dynamic           319
--------------------------------------------------------------------------------
```

可以通过**display stp** [**brief**]命令查看STP状态信息。

```
[~ServerLeaf-1] display stp brief
 MSTID  Port                  Role  STP State     Protection      Cost    Edged
    0   Eth-Trunk0            ROOT  forwarding    none            2000    disable
    0   Eth-Trunk1            DESI  forwarding    none            2000    enable
    0   Eth-Trunk2            DESI  forwarding    none            2000    enable
```

步骤4 配置优化命令

ServerLeaf-1	ServerLeaf-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.1.3 配置 Spine

配置概览

1. **配置M-LAG**
2. **配置Spine与Leaf级联链路**
3. **配置三层网关**
4. **配置出口网络**
5. **配置优化命令**

6. （可选）配置DHCP Relay

配置步骤

步骤1 配置M-LAG

Spine-1	Spine-1	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
stp root primary # 	stp root primary # 	配置Spine为生成树的根桥设备。
stp bridge-address *1-1-1* # 	stp bridge-address *1-1-1* # 	配置当前设备参与生成树计算的桥MAC，两台M-LAG设备的桥MAC必须相同，建议选择其中一台设备的系统MAC作为共同桥MAC，不同M-LAG组里的设备桥MAC不同。
ip vpn-instance *DAD* ipv4-family route-distinguisher *1:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *2:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。
interface *Eth-Trunk0* trunkport *100GE1/0/6* trunkport *100GE2/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	interface *Eth-Trunk0* trunkport *100GE1/0/6* trunkport *100GE2/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。

```
[~Spine-1] ping -vpn-instance DAD 192.168.10.2
  PING 192.168.10.2: 56  data bytes, press CTRL_C to break
    Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
```

```
  Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
  Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
  Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
  Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

 --- 192.168.10.2 ping statistics ---
  5 packet(s) transmitted
  5 packet(s) received
  0.00% packet loss
  round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~Spine-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW :
10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
   Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
   Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
   Input: 62343901 packets,93259180813 bytes
          42820517 unicast,132 broadcast,19523252 multicast
          0 errors,0 drops
   Output:48323130 packets,92501426032 bytes
          29723661 unicast,128 broadcast,18599341 multicast
          0 errors,0 drops
   Last 10 seconds input utility rate:  0.01%
   Last 10 seconds output utility rate: 0.01%
-------------------------------------------------------------
PortName                  Status            Weight
-------------------------------------------------------------
100GE1/0/6                UP                1
100GE2/0/6                UP                1
-------------------------------------------------------------
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~ServerLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                   Working Mode: Static
Preempt Delay: Disabled         Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
--------------------------------------------------------------------------------
ActorPortName             Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/6(r)             Selected 100GE    32768   6      65      10111100  1
100GE2/0/6(r)             Selected 100GE    32768   6      65      10111100  1

Partner:
--------------------------------------------------------------------------------
ActorPortName             SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/6                32768    00fd-dffb-9a03  32768   6      65      10111100
100GE2/0/6                32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
[~Spine-1] display dfs-group 1 m-lag
*                 : Local node
```

```
Heart beat state      : OK
Node 1 *
  Dfs-Group ID        : 1
  Priority            : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance        : DAD
  State               : Master
  Causation           : -
  System ID           : 00fd-dffb-9a03
  SysName             : Spine-1
  Version             : V300R022C00
  Device Type         : CE6800
Node 2
  Dfs-Group ID        : 1
  Priority            : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance        : DAD
  State               : Backup
  Causation           : -
  System ID           : 00fd-fdfd-b703
  SysName             : Spine-2
  Version             : V300R022C00
  Device Type         : CE6800
```

步骤2 配置Spine与Leaf级联链路

Spine-1	Spine-2	命令说明
vlan batch *10 20*	vlan batch *10 20*	创建业务VLAN。
interface *Eth-Trunk10* description *to-Leaf-Group-1* trunkport *100GE1/0/1* trunkport *100GE2/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *10* # interface *Eth-Trunk20* description *to-Leaf-Group-2* trunkport *100GE1/0/2* trunkport *100GE2/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *10* #	interface *Eth-Trunk10* description *to-Leaf-Group-1* trunkport *100GE1/0/1* trunkport *100GE2/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *10* # interface *Eth-Trunk20* description *to-Leaf-Group-2* trunkport *100GE1/0/2* trunkport *100GE2/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *10* #	配置级联链路。 按需放通VLAN。

上述配置完成后，可以通过**display dfs-group 1 node 1 m-lag** [**brief**]命令查看M-LAG链路聚合组的状态。

```
[~Spine-1] display dfs-group 1 node 1 m-lag brief
* - Local node

M-Lag ID     Interface      Port State    Status              Consistency-check
     10      Eth-Trunk 10   Up            active(*)-active    success
     20      Eth-Trunk 20   Up            active(*)-active    success
```

步骤3 配置三层网关

Spine-1	Spine-2	命令说明
interface vlanif *10* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd ra prefix *fc00::10:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0110* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean # interface vlanif *20* ip address *10.1.20.1 24* ipv6 enable ipv6 address *fc00::20:1 112* ipv6 nd ra prefix *fc00::20:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0120* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean #	interface vlanif *10* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd ra prefix *fc00::10:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0110* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean # interface vlanif *20* ip address *10.1.20.1 24* ipv6 enable ipv6 address *fc00::20:1 112* ipv6 nd ra prefix *fc00::20:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0120* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean #	配置三层网关VLANIF。 两台设备上配置的VLANIF接口的IP地址和MAC地址需要相同。

上述配置完成后，可以通过**display interface vlanif** *vlanif-id*命令查看双活网关的状态。

```
[~Spine-1] display interface vlanif 10
Vlanif11 current state : UP (ifindex: 5)
Line protocol current state : UP
Last line protocol up time : 2023-07-13 16:00:09+08:00
Description: "to FW"
Route Port,The Maximum Transmit Unit is 1500
Internet Address is 10.1.10.1/24
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 0000-5e00-0110
Physical is VLANIF
Current system time: 2023-07-13 16:04:39+08:00
   Input bandwidth utilization  : --
   Output bandwidth utilization : --
```

可以通过**display arp** [**interface** *interface-type interface-name*]命令查看设备学习到的ARP信息。

```
[~Spine-1] display arp
ARP Entry Types: D - Dynamic, S - Static, I - Interface, O - OpenFlow, RD - Redirect
EXP: Expire-time VLAN: VLAN or Bridge Domain

IP ADDRESS      MAC ADDRESS    EXP(M) TYPE/VLAN      INTERFACE       VPN-INSTANCE
------------------------------------------------------------------------------------
10.1.10.25   9c7d-a378-3c8d          I            Vlanif10        vpn1
10.1.20.21   cc64-a668-6814          I            Vlanif20        vpn1
------------------------------------------------------------------------------------
```

步骤4 配置出口网络

Spine-1	Spine-2	命令说明
ip vpn-instance *Internet* ipv4-family route-distinguisher *1:99* ipv6-family route-distinguisher *1:99* # 	ip vpn-instance *Internet* ipv4-family route-distinguisher *2:99* ipv6-family route-distinguisher *2:99* # 	创建公共出口VRF。
interface *Eth-Trunk99* description *to-PE-1* trunkport *10GE1/0/9* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.1 30* ipv6 enable ipv6 address *fc00::99:1 126* mode lacp-static # 	interface *Eth-Trunk99* description *to-PE-2* trunkport *10GE1/0/9* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.5 30* ipv6 enable ipv6 address *fc00::99:5 126* mode lacp-static # 	配置与PE互联接口，创建三层Eth-Trunk口，并加入对应物理成员口。
vlan *100* # interface vlanif *100* ip binding vpn-instance *Internet* ip address *10.1.99.9 30* ipv6 enable ipv6 address *fc00::99:9 126* # 	vlan *100* # interface vlanif *100* ip binding vpn-instance *Internet* ip address *10.1.99.10 30* ipv6 enable ipv6 address *fc00::99:10 126* # 	配置Spine间出口逃生路径，单台Spine与出口路由器互联的上行链路同时中断时生效。 本案例中逃生路径复用peer-link物理链路，通过peer-link上配置专用的VLANIF三层互联。
ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.2* ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:2* 	ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.6* ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:6* 	配置至出口PE的静态路由。
ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.10* preference 100 ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:10* preference 100 	ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.9* preference 100 ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:9* preference 100 	配置出口逃生路由，设置为低优先级。
ip route-static vpn-instance *Internet* *10.1.10.0 24* public ip route-static vpn-instance *Internet* *10.1.20.0 24* public ipv6 route-static vpn-instance *Internet* *fc00::10: 112* public ipv6 route-static vpn-instance *Internet* *fc00::20: 112* public 	ip route-static vpn-instance *Internet* *10.1.10.0 24* public ip route-static vpn-instance *Internet* *10.1.20.0 24* public ipv6 route-static vpn-instance *Internet* *fc00::10: 112* public ipv6 route-static vpn-instance *Internet* *fc00::20: 112* public 	配置至业务网段的静态路由，下一跳为Public VRF。
ip route-static 0.0.0.0 0.0.0.0 vpn-instance *Internet* ipv6 route-static :: 0 vpn-instance *Internet* 	ip route-static 0.0.0.0 0.0.0.0 vpn-instance *Internet* ipv6 route-static :: 0 vpn-instance *Internet* 	配置Public VRF的静态路由，下一跳为出口VRF。

上述配置完成后，可以检查服务器端和外部网络之间能否ping通。

步骤5　配置优化命令

Spine-1	Spine-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	关闭不使用的端口。

Spine-1	Spine-2	命令说明
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	配置VLAN 1的流量抑制功能，防止广播风暴。

步骤6 （可选）配置DHCP Relay

在服务器通过DHCP上线的场景，需要在M-LAG双活网关上配置DHCP Relay功能，使服务器正确获取IP地址。

- DHCP Client与DHCP Server属于同一VPN场景

Spine-1	Spine-2	命令说明
vlan 30 # interface vlanif *30* ip address *10.1.30.1 24* ipv6 enable ipv6 address *fc00::30:1 112* ipv6 nd na glean # 	vlan 30 # interface vlanif *30* ip address *10.1.30.2 24* ipv6 enable ipv6 address *fc00::30:2 112* ipv6 nd na glean # 	在作为DHCP Relay的Spine上配置DHCP专用VLANIF接口，作为发送DHCP Realy报文的源接口。 两台M-LAG设备需要配置不同的接口IP，避免DHCP报文来回路径不一致的问题。
dhcp enable # interface vlanif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* dhcp relay source-ip-address *10.1.30.1* dhcp relay information enable dhcp option82 vendor-specific format vendor-sub-option 10 source-ip-address *10.1.30.1* dhcpv6 relay destination *fc00::200:10* dhcpv6 relay source-ip-address *fc00::30:1* # 	dhcp enable # interface vlanif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* dhcp relay source-ip-address *10.1.30.2* dhcp relay information enable dhcp option82 vendor-specific format vendor-sub-option 10 source-ip-address *10.1.30.2* dhcpv6 relay destination *fc00::200:10* dhcpv6 relay source-ip-address *fc00::30:2* # 	在业务网关VLANIF下配置DHCP Relay，以VLANIF10配置为例。 *10.1.200.10*/*fc00::200:10*为DHCP Server的IPv4/IPv6地址，该地址需要和DHCP专用VLANIF接口的IPv4/IPv6地址三层互通。 DHCP Server和作为DHCP Client的业务服务器部署在同一个VPN中。部署DHCP Server的相关配置请参考上文业务VPN的配置，此处略。

- DHCP Client与DHCP Server跨VPN场景

ServerLeaf-1	ServerLeaf-2	命令说明
ip vpn-instance *DHCP* ipv4-family route-distinguisher *1:30* ipv6-family route-distinguisher *1:30* # 	ip vpn-instance *DHCP* ipv4-family route-distinguisher *2:30* ipv6-family route-distinguisher *2:30* # 	配置DHCP专用VPN。

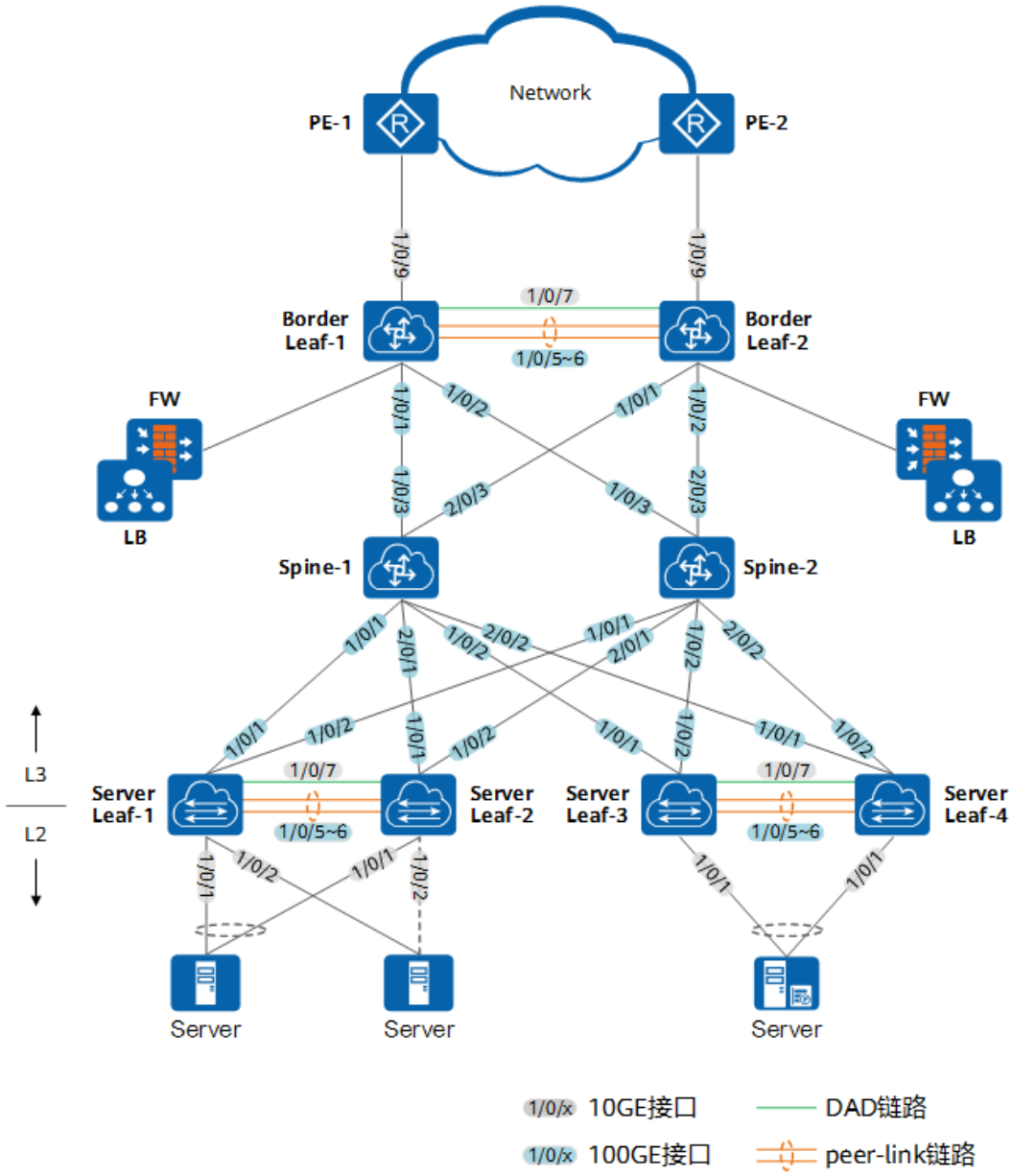
ServerLeaf-1	ServerLeaf-2	命令说明
vlan 30 # interface vlanif *30* ip binding vpn-instance *DHCP* ip address *10.1.30.1 24* ipv6 enable ipv6 address *fc00::30:1 112* ipv6 nd na glean # 	vlan 30 # interface vlanif *30* ip binding vpn-instance *DHCP* ip address *10.1.30.2 24* ipv6 enable ipv6 address *fc00::30:2 112* ipv6 nd na glean # 	在作为DHCP Relay的Spine上配置DHCP专用VLANIF接口，作为发送DHCP Realy报文的源接口。 两台M-LAG设备需要配置不同的接口IP，避免DHCP报文来回路径不一致的问题。
dhcp enable # interface vlanif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* vpn-instance *DHCP* dhcp relay giaddr source-interface *vlanif30* dhcp option82 link-selection insert enable dhcp option82 vss-control insert enable dhcp option82 server-id-override insert enable dhcpv6 relay destination *fc00::200:10* vpn-instance *DHCP* dhcpv6 relay option79 insert enable dhcpv6 relay vss-control insert enable dhcpv6 relay source-ip-address *fc00::30:1* # 	dhcp enable # interface vlanif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* vpn-instance *DHCP* dhcp relay giaddr source-interface *vlanif30* dhcp option82 link-selection insert enable dhcp option82 vss-control insert enable dhcp option82 server-id-override insert enable dhcpv6 relay destination *fc00::200:10* vpn-instance *DHCP* dhcpv6 relay option79 insert enable dhcpv6 relay vss-control insert enable dhcpv6 relay source-ip-address *fc00::30:2* # 	在业务网关VLANIF下配置DHCP Relay，以VLANIF10配置为例。 *10.1.200.10*/*fc00::200:10*为DHCP Server的IPv4/IPv6地址，该地址需要和DHCP专用VLANIF接口的IPv4/IPv6地址三层互通。 DHCP Server部署在DHCP专用VPN中。部署DHCP Server的相关配置请参考上文业务VPN的配置，此处略。

----结束

2.2 M-LAG 作为三层双活网关

2.2.1 组网方案

图 2-2 M-LAG 作为三层双活网关示意图



如图2-2所示，某新建数据中心网络采用Spine-Leaf架构，其中：

- Server Leaf部署M-LAG，作为三层双活网关与服务器对接。
- Spine独立部署，分别与Server Leaf、Border Leaf互联，运行路由协议保证路由三层可达。
- Border Leaf部署M-LAG，作为出口设备。防火墙（FW）及负载均衡器（LB）旁挂在Border Leaf上。更多关于M-LAG设备组和防火墙、负载均衡器的对接描述请参考**2.5 M-LAG与防火墙对接**、**2.6 M-LAG与负载均衡器（LB）对接**。
- 服务器以负载分担或主备方式接入，服务器负载分担接入时，推荐M-LAG使用双活模式，服务器主备接入时，推荐M-LAG使用主备模式。

- 建议通过单独链路作为M-LAG设备之间双主检测链路，提升可靠性。
- 建议M-LAG设备通过静态LACP模式和服务器对接，同时需要配置LACP forced-up功能。当服务器重启或刚上线时，Eth-Trunk接口接受LACP协议报文超时后，Eth-Trunk成员口的状态会变为Down。这时可以通过配置成员口Force Up，使该接口继续转发业务流量，防止业务丢包。
- 两台设备组建M-LAG之后，可能产生环路，建议使用V-STP功能防止M-LAG成员口因为环路协议被堵塞。级联M-LAG场景下由于不支持根桥方式的M-LAG，必须使用V-STP方式配置M-LAG。如果M-LAG设备需要和友商PVST/PVST+协议对接，则推荐使用VBST协议和PVST/PVST+对接。
- 建议至少使用两条链路捆绑成Eth-Trunk链路作为M-LAG peer-link链路，如果是两台框式设备组建M-LAG，建议使用不同单板的接口捆绑成Eth-Trunk链路作为M-LAG peer-link链路。如果使用CE6881、CE6863、CE6881H和CE6863H设备组建M-LAG，建议选择（100GE 1/0/1~1/0/3和100GE 1/0/4~1/0/6）中的接口捆绑Eth-Trunk作为M-LAG peer-link链路。
- 本例中Underlay协议使用OSPF路由协议，实际应用建议基于实际情况选择OSPF或者BGP协议。OSPF和BGP路由协议作为Underlay层面的路由协议时，两者之间的对比参见下表。

表 2-1 BGP 和 OSPF 优劣对比

对比维度	OSPF	BGP
收敛速度	收敛速度较快	收敛速度快 ● 可配置BGP关联接口或者与BFD联动优化收敛性能
协议部署	协议部署简单，但控制手段较少，依赖cost，需要全网调整	配置相对复杂，路由控制手段丰富
网络规模	适用中小型网络 ● OSPF计算消耗较大，扩展能力有限	适用大中型网络 ● BGP计算消耗低，可扩展性好
故障域	故障域较大	每个分区路由域独立，故障域可控
临时路由环路	网络拓扑变化时，OSPF可能会计算出短暂的临时路由环路	通过合理规划AS，网路拓扑变化时，BGP不会计算出临时路由环路
适用场景	● 中小规模DC使用，很少在大规模数据中心使用 ● 中小型网络单Area，大型网络三层架构多Area部署 ● 建议邻居数<200 ● 单POD邻居数<100，避免路由域过大影响网络性能	● 中大型网络，ISP大型DC网络广泛采用 ● 多POD多层Spine设计，POD间通过EBGP传递路由 ● 建议邻居数<500 ● 单POD邻居数<100，避免路由域过大影响网络性能

2.2.2 配置 Server Leaf

配置概览

1. **配置互联接口IP地址及Loopback接口**
2. **配置M-LAG**
3. **配置路由协议，实现三层互通**
4. **配置三层网关**
5. **配置服务器接入**
6. **配置优化命令**

配置步骤

步骤1 配置互联接口IP地址及Loopback接口

ServerLeaf-1	ServerLeaf-2	命令说明
interface *100GE1/0/1* description *to-Spine-1* undo portswitch ip address *192.168.1.2 30* # interface *100GE1/0/2* description *to-Spine-2* undo portswitch ip address *192.168.1.18 30* # 	interface *100GE1/0/1* description *to-Spine-1* undo portswitch ip address *192.168.1.6 30* # interface *100GE1/0/2* description *to-Spine-2* undo portswitch ip address *192.168.1.22 30* # 	配置与Spine互联接口。
interface LoopBack1 description *Router-id* ip address *10.1.8.5 32* # 	interface LoopBack1 description *Router-id* ip address *10.1.8.6 32* # 	配置Loopback1接口，用作Router-ID。

步骤2 配置M-LAG

ServerLeaf-1	ServerLeaf-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
ip vpn-instance *DAD* ipv4-family route-distinguisher *5:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *6:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。

ServerLeaf-1	ServerLeaf-2	命令说明
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 #	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 #	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。 配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。

```
[~ServerLeaf-1] ping -vpn-instance DAD 192.168.10.2
  PING 192.168.10.2: 56  data bytes, press CTRL_C to break
    Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

  --- 192.168.10.2 ping statistics ---
    5 packet(s) transmitted
    5 packet(s) received
    0.00% packet loss
    round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~ServerLeaf-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW : 10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
    Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
    Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
    Input: 62343901 packets,93259180813 bytes
          42820517 unicast,132 broadcast,19523252 multicast
          0 errors,0 drops
    Output:48323130 packets,92501426032 bytes
          29723661 unicast,128 broadcast,18599341 multicast
          0 errors,0 drops
    Last 10 seconds input utility rate:  0.01%
    Last 10 seconds output utility rate: 0.01%
-------------------------------------------------------------
PortName                Status          Weight
-------------------------------------------------------------
100GE1/0/5              UP              1
100GE1/0/6              UP              1
```

```
------------------------------------------------------------
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~ServerLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                   Working Mode: Static
Preempt Delay: Disabled         Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
--------------------------------------------------------------------------------
ActorPortName            Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/5(r)            Selected 100GE    32768   6      65      10111100  1
100GE1/0/6(r)            Selected 100GE    32768   6      65      10111100  1

Partner:
--------------------------------------------------------------------------------
ActorPortName            SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/5               32768    00fd-dffb-9a03  32768   6      65      10111100
100GE1/0/6               32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
[~ServerLeaf-1] display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance       : DAD
  State              : Backup
  Causation          : -
  System ID          : 00fd-fdfd-b703
  SysName            : ServerLeaf-2
  Version            : V300R022C00
  Device Type        : CE6800
```

步骤3 配置三层网关

ServerLeaf-1	ServerLeaf-2	命令说明
vlan batch *10 20* # interface vlanif *10* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd ra prefix *fc00::10:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0110* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean # interface vlanif *20* ip address *10.1.20.1 24* ipv6 enable ipv6 address *fc00::20:1 112* ipv6 nd ra prefix *fc00::20:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0120* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean #	vlan batch *10 20* # interface vlanif *10* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd ra prefix *fc00::10:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0110* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean # interface vlanif *20* ip address *10.1.20.1 24* ipv6 enable ipv6 address *fc00::20:1 112* ipv6 nd ra prefix *fc00::20:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0120* ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean #	配置三层网关VLANIF。 两台设备上配置的VLANIF接口的IP地址和MAC地址需要相同。

上述配置完成后，可以通过**display interface vlanif** *vlanif-id*命令查看双活网关的状态。

```
[~ServerLeaf-1] display interface vlanif 10
Vlanif11 current state : UP (ifindex: 5)
Line protocol current state : UP
Last line protocol up time : 2023-07-13 16:00:09+08:00
Description: "to FW"
Route Port,The Maximum Transmit Unit is 1500
Internet Address is 10.1.10.1/24
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 0000-5e00-0110
Physical is VLANIF
Current system time: 2023-07-13 16:04:39+08:00
    Input bandwidth utilization  : --
    Output bandwidth utilization : --
```

可以通过**display arp** [**interface** *interface-type interface-name*]命令查看设备学习到的ARP信息。

```
[~ServerLeaf-1] display arp
ARP Entry Types: D - Dynamic, S - Static, I - Interface, O - OpenFlow, RD - Redirect
EXP: Expire-time VLAN: VLAN or Bridge Domain

IP ADDRESS      MAC ADDRESS    EXP(M) TYPE/VLAN      INTERFACE       VPN-INSTANCE
------------------------------------------------------------------------------------
10.1.10.25   9c7d-a378-3c8d          I             Vlanif10
10.1.20.21   cc64-a668-6814          I             Vlanif20
------------------------------------------------------------------------------------
```

步骤4 配置路由协议，实现三层互通

本章节EBGP路由协议为例。

ServerLeaf-1	ServerLeaf-2	命令说明
bfd # bgp *65300* router-id *10.1.8.5* auto-frr advertise lowest-priority all-address-family peer-up delay 360	bfd # bgp *65300* router-id *10.1.8.5* auto-frr advertise lowest-priority all-address-family peer-up delay 360	全局使能BFD功能。 配置BGP AS号及相应Router-ID。 开启BGP Auto FRR功能，对于从不同对等体学到的相同前缀的路由，利用最优路由作为主链路进行转发，并自动将次优路由作为备份链路。 在邻居状态由Down到Up时将BGP路由的优先级调整为最低优先级，路由延时发布，解决回切场景丢包时间长问题。
group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.1* group *Group_Spine* peer *192.168.1.17* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 import-route direct maximum load-balancing 32	group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.5* group *Group_Spine* peer *192.168.1.21* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 import-route direct maximum load-balancing 32	创建一个用于与Spine对接对等体组，简化后续配置。 配置对等体组的BFD功能，并设置BFD参数。 引入路由，发布业务网段地址。
interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	将Leaf上与Spine互联的接口配置为M-LAG上行口，上行链路和Peer-link链路同时故障的设备的端口优先被Error-down。
vlan *100* # interface vlanif *100* ip address *10.1.100.5 30* # bgp *65300* peer *10.1.100.6* as-number *65300* peer *10.1.100.6* connect-interface *vlanif100* peer *10.1.100.6* next-hop-local #	vlan *100* # interface vlanif *100* ip address *10.1.100.6 30* # bgp *65300* peer *10.1.100.5* as-number *65300* peer *10.1.100.5* connect-interface *vlanif100* peer *10.1.100.5* next-hop-local #	配置三层逃生路径。通过peer-link上配置专用的VLANIF三层直连，加入到iBGP路由协议中。 当M-LAG组中一个Leaf与Spine互联的上行链路全部故障时，通过Leaf间的逃生路径，将流量转发至M-LAG组内另外一台Leaf上继续转发。

上述配置完成后，可以通过**display bgp peer**命令查看BGP邻居状态，或通过**display bgp routing-table**命令查看BGP路由信息。

步骤5 配置服务器接入

- 服务器负载分担方式接入时

ServerLeaf-1	ServerLeaf-2	命令说明
interface eth-trunk *1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *10GE1/0/1* storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* # 	interface eth-trunk *1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *10GE1/0/1* storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* # 	配置接入用的链路聚合组。 按需放通VLAN，不放通VLAN1，防止成环。 配置静态LACP模式。 配置边缘端口。 配置未知单播抑制，建议为端口带宽的5%。 配置组播报文抑制，建议为1 Mbit/s。 配置广播报文抑制，建议为1 Mbit/s。
interface *10GE1/0/1* lacp force-up # 	-	（可选）在服务器需要进行PXE安装的场景，其中一台设备的成员口需要配置lacp force-up，该成员口为对接安装操作系统的网卡的接口。

- 服务器主备方式接入时

ServerLeaf-1	ServerLeaf-2	命令说明
interface *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *20* stp edged-port enable storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* # 	interface *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *20* stp edged-port enable storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* # 	配置服务器接入端口。 按需放通VLAN，不放通VLAN1，防止成环。 配置边缘端口。 配置未知单播抑制，建议为端口带宽的5%。 配置组播报文抑制，建议为1 Mbit/s。 配置广播报文抑制，建议为1 Mbit/s。

上述配置完成后，可以通过**display dfs-group 1 node 1 m-lag** [**brief**]命令查看M-LAG链路聚合组的状态。

```
[~ServerLeaf-1] display dfs-group 1 node 1 m-lag brief
* - Local node

M-Lag ID     Interface      Port State    Status              Consistency-check
     1     Eth-Trunk 1     Up          active(*)-active      success
     2     Eth-Trunk 2     Up          active(*)-inactive    success
    10     Eth-Trunk 10    Up          active(*)-active      success
```

可以通过**display mac-address** [**vlan** *vlan-id*]命令查看设备学习到的MAC地址信息。

```
[~ServerLeaf-1] display mac-address
Flags: * - Backup
       # - forwarding logical interface, operations cannot be performed based
           on the interface.
```



```
BD   : bridge-domain   Age : dynamic MAC learned time in seconds
--------------------------------------------------------------------------------
MAC Address    VLAN/VSI/BD   Learned-From        Type                 Age
--------------------------------------------------------------------------------
9c7d-a378-3c8d 10/-/-        Eth-Trunk1          dynamic              319
cc64-a668-6814 20/-/-        Eth-Trunk2          dynamic              319
--------------------------------------------------------------------------------
```

可以通过**display stp** [**brief**]命令查看STP状态信息。

```
[~ServerLeaf-1] display stp brief
 MSTID  Port                        Role  STP State     Protection      Cost    Edged
    0   Eth-Trunk0                  ROOT  forwarding    none            2000    disable
    0   Eth-Trunk1                  DESI  forwarding    none            2000    enable
    0   Eth-Trunk2                  DESI  forwarding    none            2000    enable
```

步骤6　配置优化命令

ServerLeaf-1	ServerLeaf-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.2.3 配置 Spine

配置概览

1. **配置互联接口IP地址及Loopback接口**
2. **配置路由协议，实现三层互通**
3. **配置优化命令**

配置步骤

步骤1　配置互联接口IP地址及Loopback接口

Spine-1	Spine-2	命令说明
interface *100GE1/0/1* description *to-ServerLeaf-1* undo portswitch ip address *192.168.1.1 30* # interface *100GE2/0/1* description *to-ServerLeaf-2* undo portswitch ip address *192.168.1.5 30* # interface *100GE1/0/2* description *to-ServerLeaf-3* undo portswitch ip address *192.168.1.9 30* # interface *100GE2/0/2* description *to-ServerLeaf-4* undo portswitch ip address *192.168.1.13 30* # interface *100GE1/0/3* description *to-BorderLeaf-1* undo portswitch ip address *192.168.1.33 30* # interface *100GE2/0/3* description *to-BorderLeaf-2* undo portswitch ip address *192.168.1.37 30* #	interface *100GE1/0/1* description *to-ServerLeaf-1* undo portswitch ip address *192.168.1.17 30* # interface *100GE2/0/1* description *to-ServerLeaf-2* undo portswitch ip address *192.168.1.21 30* # interface *100GE1/0/2* description *to-ServerLeaf-3* undo portswitch ip address *192.168.1.25 30* # interface *100GE2/0/2* description *to-ServerLeaf-4* undo portswitch ip address *192.168.1.29 30* # interface *100GE1/0/3* description *to-BorderLeaf-1* undo portswitch ip address *192.168.1.41 30* # interface *100GE2/0/3* description *to-BorderLeaf-2* undo portswitch ip address *192.168.1.45 30* #	配置与Leaf互联接口。
interface LoopBack1 description *Router-id* ip address *10.1.8.3 32* #	interface LoopBack1 description *Router-id* ip address *10.1.8.4 32* #	配置Loopback1接口，用作Router-ID。

步骤2 配置路由协议，实现三层互通

本章节EBGP路由协议为例。

Spine-1	Spine-2	命令说明
bfd # bgp *65200* router-id *10.1.8.3* advertise lowest-priority all-address-family peer-up delay 360	bfd # bgp *65200* router-id *10.1.8.4* advertise lowest-priority all-address-family peer-up delay 360	全局使能BFD功能。 配置BGP AS号及相应Router-ID。 在邻居状态由Down到Up时将BGP路由的优先级调整为最低优先级，路由延时发布，解决回切场景丢包时间长问题。

Spine-1	Spine-2	命令说明
group *Group_ServerLeaf1* external peer *Group_ServerLeaf1* as-number *65300* peer *192.168.1.2* group *Group_ServerLeaf1* peer *192.168.1.6* group *Group_ServerLeaf1* group *Group_ServerLeaf2* external peer *Group_ServerLeaf2* as-number *65400* peer *192.168.1.10* group *Group_ServerLeaf2* peer *192.168.1.14* group *Group_ServerLeaf2* group *Group_BorderLeaf* external peer *Group_BorderLeaf* as-number *65100* peer *192.168.1.34* group *Group_BorderLeaf* peer *192.168.1.38* group*Group_BorderLeaf* peer *Group_ServerLeaf1* bfd enable peer *Group_ServerLeaf1* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_ServerLeaf2* bfd enable peer *Group_ServerLeaf2* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_BorderLeaf* bfd enable peer *Group_BorderLeaf* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 maximum load-balancing 32	group *Group_ServerLeaf1* external peer *Group_ServerLeaf1* as-number *65300* peer *192.168.1.18* group *Group_ServerLeaf1* peer *192.168.1.22* group *Group_ServerLeaf1* group *Group_ServerLeaf2* external peer *Group_ServerLeaf2* as-number *65400* peer *192.168.1.26* group *Group_ServerLeaf2* peer *192.168.1.30* group *Group_ServerLeaf2* group *Group_BorderLeaf* external peer *Group_BorderLeaf* as-number *65100* peer *192.168.1.42* group *Group_BorderLeaf* peer *192.168.1.46* group*Group_BorderLeaf* peer *Group_ServerLeaf1* bfd enable peer *Group_ServerLeaf1* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_ServerLeaf2* bfd enable peer *Group_ServerLeaf2* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_BorderLeaf* bfd enable peer *Group_BorderLeaf* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 maximum load-balancing 32	创建一个用于与ServerLeaf1、ServerLeaf2对接的对等体组，简化后续配置。 创建一个用于与ServerLeaf3、ServerLeaf4对接的对等体组，简化后续配置。 创建一个用于与BorderLeaf1、BorderLeaf2对接的对等体组，简化后续配置。 配置对等体组的BFD功能，并设置BFD参数。

上述配置完成后，可以通过**display bgp peer**命令查看BGP邻居状态，或通过**display bgp routing-table**命令查看BGP路由信息。

步骤3 配置优化命令

Spine-1	Spine-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.2.4 配置 Border Leaf

配置概览

1. **配置互联接口IP地址及Loopback接口**
2. **配置M-LAG**
3. **配置路由协议，实现三层互通**
4. **配置出口网络**
5. **配置优化命令**

配置步骤

步骤1 配置互联接口IP地址及Loopback接口

BorderLeaf-1	BorderLeaf-2	命令说明
interface *100GE1/0/1* description *to-Spine1* undo portswitch ip address *192.168.1.34 30* # interface *100GE1/0/2* description *to-Spine2* undo portswitch ip address *192.168.1.42 30* # 	interface *100GE1/0/1* description *to-Spine1* undo portswitch ip address *192.168.1.38 30* # interface *100GE1/0/2* description *to-Spine2* undo portswitch ip address *192.168.1.46 30* # 	配置与Spine互联接口。
interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.1 32* # 	interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.2 32* # 	配置Loopback1接口，用作Router-ID。

步骤2 配置M-LAG

BorderLeaf-1	BorderLeaf-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
ip vpn-instance *DAD* ipv4-family route-distinguisher *1:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *2:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。

BorderLeaf-1	BorderLeaf-2	命令说明
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。
interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。

```
[~BorderLeaf-1] ping -vpn-instance DAD 192.168.10.2
  PING 192.168.10.2: 56  data bytes, press CTRL_C to break
    Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

  --- 192.168.10.2 ping statistics ---
    5 packet(s) transmitted
    5 packet(s) received
    0.00% packet loss
    round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~BorderLeaf-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW : 10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
    Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
    Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
    Input: 62343901 packets,93259180813 bytes
          42820517 unicast,132 broadcast,19523252 multicast
          0 errors,0 drops
    Output:48323130 packets,92501426032 bytes
          29723661 unicast,128 broadcast,18599341 multicast
          0 errors,0 drops
    Last 10 seconds input utility rate:  0.01%
    Last 10 seconds output utility rate: 0.01%
--------------------------------------------------------------
PortName                  Status            Weight
--------------------------------------------------------------
100GE1/0/5                UP                1
100GE1/0/6                UP                1
```

```
------------------------------------------------------------
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~BorderLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                   Working Mode: Static
Preempt Delay: Disabled         Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
--------------------------------------------------------------------------------
ActorPortName            Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/5(r)            Selected 100GE    32768   6      65      10111100  1
100GE1/0/6(r)            Selected 100GE    32768   6      65      10111100  1

Partner:
--------------------------------------------------------------------------------
ActorPortName            SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/5               32768    00fd-dffb-9a03  32768   6      65      10111100
100GE1/0/6               32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
[~BorderLeaf-1] display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance       : DAD
  State              : Backup
  Causation          : -
  System ID          : 00fd-fdfd-b703
  SysName            : ServerLeaf-2
  Version            : V300R022C00
  Device Type        : CE6800
```

步骤3 配置路由协议，实现三层互通

本章节EBGP路由协议为例。

BorderLeaf-1	BorderLeaf-1	命令说明
bfd # bgp *65100* router-id *10.1.8.1* auto-frr advertise lowest-priority all-address-family peer-up delay 360	bfd # bgp *65100* router-id *10.1.8.2* auto-frr advertise lowest-priority all-address-family peer-up delay 360	全局使能BFD功能。 配置BGP AS号及相应Router-ID。 开启BGP Auto FRR功能，对于从不同对等体学到的相同前缀的路由，利用最优路由作为主链路进行转发，并自动将次优路由作为备份链路。 在邻居状态由Down到Up时将BGP路由的优先级调整为最低优先级，路由延时发布，解决回切场景丢包时间长问题。
group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.33* group *Group_Spine* peer *192.168.1.41* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 default-route imported import-route static maximum load-balancing 32	group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.37* group *Group_Spine* peer *192.168.1.45* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 default-route imported import-route static maximum load-balancing 32	创建一个用于与Spine对接对等体组，简化后续配置。 配置对等体组的BFD功能，并设置BFD参数。 引入路由。
interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	将Leaf上与Spine互联的接口配置为M-LAG上行口，上行链路和Peer-link链路同时故障的设备的端口优先被Error-down。
vlan *100* # interface vlanif *100* ip address *10.1.100.1 30* # bgp *65100* peer *10.1.100.2* as-number *65100* peer *10.1.100.2* connect-interface *vlanif100* peer *10.1.100.2* next-hop-local #	vlan *100* # interface vlanif *100* ip address *10.1.100.2 30* # bgp *65100* peer *10.1.100.1* as-number *65100* peer *10.1.100.1* connect-interface *vlanif100* peer *10.1.100.1* next-hop-local #	配置三层逃生路径。通过peer-link上配置专用的VLANIF三层直连，加入到iBGP路由协议中。 当M-LAG组中一个Leaf与Spine互联的上行链路全部故障时，通过Leaf间的逃生路径，将流量转发至M-LAG组内另外一台Leaf上继续转发。

上述配置完成后，可以通过**display bgp peer**命令查看BGP邻居状态，或通过**display bgp routing-table**命令查看BGP路由信息。

步骤4 配置出口网络

BorderLeaf-1	BorderLeaf-2	命令说明
ip vpn-instance *Internet* ipv4-family route-distinguisher *1:99* ipv6-family route-distinguisher *1:99* #	ip vpn-instance *Internet* ipv4-family route-distinguisher *2:99* ipv6-family route-distinguisher *2:99* #	创建出口VRF。
interface *Eth-Trunk99* description *to-PE-1* trunkport *10GE1/0/9* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.1 30* ipv6 enable ipv6 address *fc00::99:1 126* mode lacp-static # interface *10GE1/0/9* set up-delay 300 #	interface *Eth-Trunk99* description *to-PE-2* trunkport *10GE1/0/9* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.5 30* ipv6 enable ipv6 address *fc00::99:5 126* mode lacp-static # interface *10GE1/0/9* set up-delay 300 #	配置与PE互联接口，创建三层Eth-Trunk口，并加入对应物理成员口。 Border Leaf与PE互联接口配置延时UP，防止设备重启后，路由下发较慢或者下行隧道建立较慢导致业务流量回切时间较长。
interface *Eth-Trunk20* trunkport *10GE1/0/3* trunkport *10GE1/0/4* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.9 30* ipv6 enable ipv6 address *fc00::99:9 126* mode lacp-static m-lag unpaired-port reserved #	interface *Eth-Trunk20* trunkport *10GE1/0/3* trunkport *10GE1/0/4* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.10 30* ipv6 enable ipv6 address *fc00::99:10 126* mode lacp-static m-lag unpaired-port reserved #	配置Border Leaf间出口逃生路径，单台Border Leaf与出口路由器互联的上行链路同时中断时生效。 出于高可靠性考虑，逃生路径建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。
ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.2* ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:2*	ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.6* ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:6*	配置至出口PE的静态路由。
ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.10* preference 100 ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:10* preference 100	ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.9* preference 100 ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:9* preference 100	配置出口逃生路由，设置为低优先级。
ip route-static vpn-instance *Internet* *10.1.10.0 24* public ip route-static vpn-instance *Internet* *10.1.20.0 24* public ipv6 route-static vpn-instance *Internet* *fc00::10: 112* public ipv6 route-static vpn-instance *Internet* *fc00::20: 112* public	ip route-static vpn-instance *Internet* *10.1.10.0 24* public ip route-static vpn-instance *Internet* *10.1.20.0 24* public ipv6 route-static vpn-instance *Internet* *fc00::10: 112* public ipv6 route-static vpn-instance *Internet* *fc00::20: 112* public	配置至业务网段的静态路由，下一跳为Public VRF。
ip route-static 0.0.0.0 0.0.0.0 vpn-instance *Internet* ipv6 route-static :: 0 vpn-instance *Internet*	ip route-static 0.0.0.0 0.0.0.0 vpn-instance *Internet* ipv6 route-static :: 0 vpn-instance *Internet*	配置Public VRF的静态路由，下一跳为出口VRF。

上述配置完成后，可以检查服务器端和外部网络之间能否ping通。

步骤5 配置优化命令

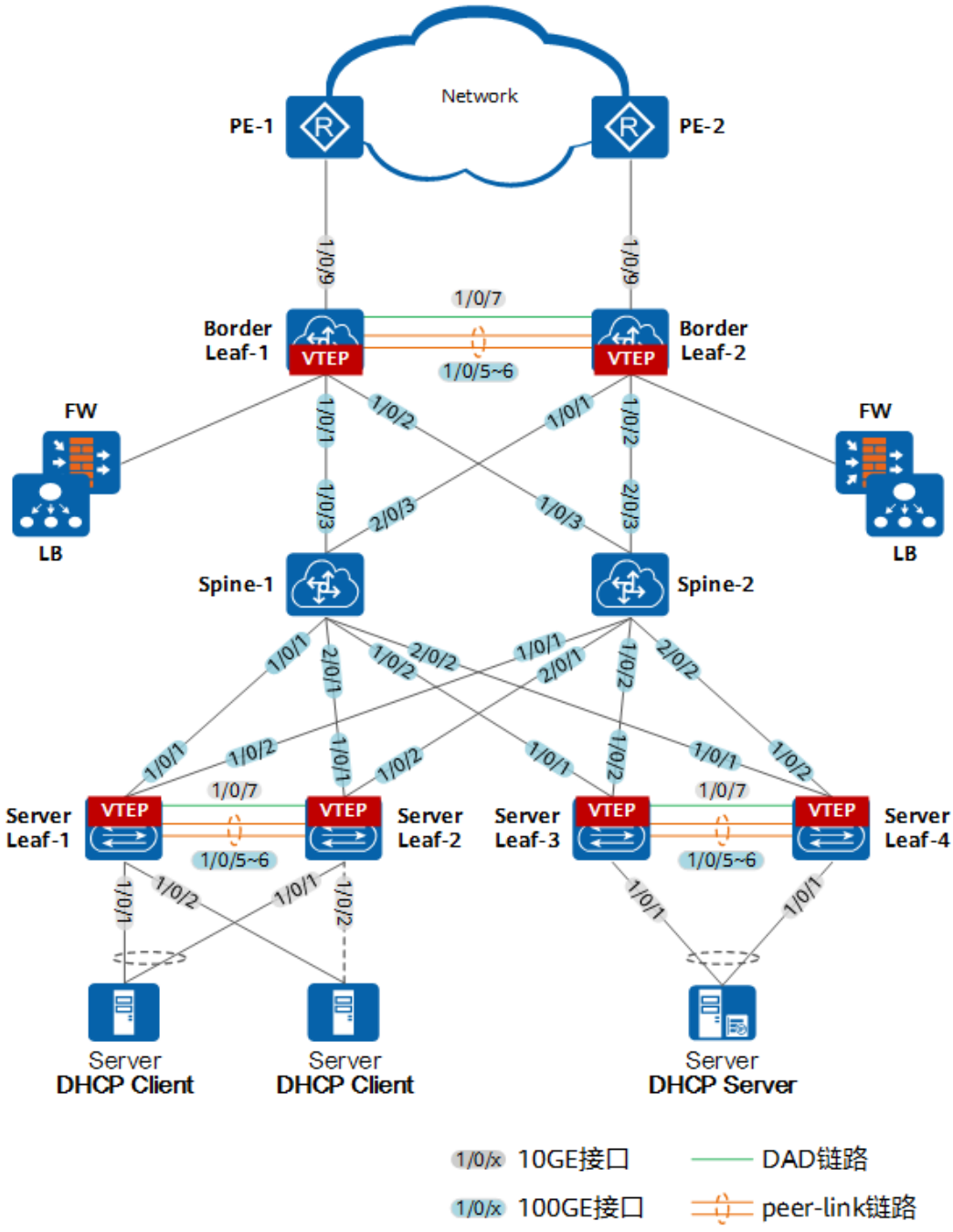
BorderLeaf-1	BorderLeaf-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.3 M-LAG 作为 VXLAN 分布式网关

2.3.1 组网方案

图 2-3 M-LAG 作为 VXLAN 分布式网关示意图



如图2-3所示，某新建数据中心网络采用VXLAN分布式网关部署方式，其中：

- Server Leaf部署M-LAG，作为VXLAN分布式网关与服务器对接。
- Spine独立部署，分别与Server Leaf、Border Leaf互联，运行路由协议保证Underlay三层可达。

- Border Leaf部署M-LAG，作为出口网关。防火墙（FW）及负载均衡器（LB）旁挂在Border Leaf上。更多关于M-LAG设备组和防火墙、负载均衡器的对接描述请参考**2.5 M-LAG与防火墙对接**、**2.6 M-LAG与负载均衡器（LB）对接**。
- 服务器以负载分担或主备方式接入，服务器负载分担接入时，推荐M-LAG使用双活模式，服务器主备接入时，推荐M-LAG使用主备模式。
- 建议通过单独链路作为M-LAG设备之间双主检测链路，提升可靠性。
- 建议M-LAG设备通过静态LACP模式和服务器对接，同时需要配置LACP forced-up功能。当服务器重启或刚上线时，Eth-Trunk接口接受LACP协议报文超时后，Eth-Trunk成员口的状态会变为Down。这时可以通过配置成员口Force Up，使该接口继续转发业务流量，防止业务丢包。
- 两台设备组建M-LAG之后，可能产生环路，建议使用V-STP功能防止M-LAG成员口因为环路协议被堵塞。级联M-LAG场景下由于不支持根桥方式的M-LAG，必须使用V-STP方式配置M-LAG。如果M-LAG设备需要和友商PVST/PVST+协议对接，则推荐使用VBST协议和PVST/PVST+对接。
- 建议至少使用两条链路捆绑成Eth-Trunk链路作为M-LAG peer-link链路，如果是两台框式设备组建M-LAG，建议使用不同单板的接口捆绑成Eth-Trunk链路作为M-LAG peer-link链路。如果使用CE6881、CE6863、CE6881H和CE6863H设备组建M-LAG，建议选择（100GE 1/0/1~1/0/3和100GE 1/0/4~1/0/6）中的接口捆绑Eth-Trunk作为M-LAG peer-link链路。
- 本例中Underlay协议使用OSPF路由协议，实际应用建议基于实际情况选择OSPF或者BGP协议。OSPF和BGP路由协议作为Underlay层面的路由协议时，两者之间的对比参见下表。

表 2-2 BGP 和 OSPF 优劣对比

对比维度	OSPF	BGP
收敛速度	收敛速度较快	收敛速度快 • 可配置BGP关联接口或者与BFD联动优化收敛性能
协议部署	协议部署简单，但控制手段较少，依赖cost，需要全网调整	配置相对复杂，路由控制手段丰富
网络规模	适用中小型网络 • OSPF计算消耗较大，扩展能力有限	适用大中型网络 • BGP计算消耗低，可扩展性好
故障域	故障域较大	每个分区路由域独立，故障域可控
临时路由环路	网络拓扑变化时，OSPF可能会计算出短暂的临时路由环路	通过合理规划AS，网路拓扑变化时，BGP不会计算出临时路由环路
适用场景	• 中小规模DC使用，很少在大规模数据中心使用 • 中小型网络单Area，大型网络三层架构多Area部署 • 建议邻居数<200 • 单POD邻居数<100，避免路由域过大影响网络性能	• 中大型网络，ISP大型DC网络广泛采用 • 多POD多层Spine设计，POD间通过EBGP传递路由 • 建议邻居数<500 • 单POD邻居数<100，避免路由域过大影响网络性能

2.3.2 配置 Server Leaf

配置概览

1. **配置互联接口IP地址及Loopback接口**
2. **配置M-LAG**
3. **配置Underlay路由协议，实现三层互通**
4. **配置BGP EVPN路由**
5. **配置VPN实例及EVPN实例**
6. **配置三层网关**
7. **配置服务器接入**
8. **配置优化命令**
9. **（可选）配置DHCP Relay**

配置步骤

步骤1 配置互联接口IP地址及Loopback接口

ServerLeaf-1	ServerLeaf-2	命令说明
interface *100GE1/0/1* description *to-Spine-1* undo portswitch ip address *192.168.1.2 30* qos phb marking dscp enable # interface *100GE1/0/2* description *to-Spine-2* undo portswitch ip address *192.168.1.18 30* qos phb marking dscp enable #	interface *100GE1/0/1* description *to-Spine-1* undo portswitch ip address *192.168.1.6 30* qos phb marking dscp enable # interface *100GE1/0/2* description *to-Spine-2* undo portswitch ip address *192.168.1.22 30* qos phb marking dscp enable #	配置与Spine互联接口。
interface LoopBack0 description *VTEP* ip address *10.1.7.2 32* # interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.5 32* # interface LoopBack2 description *Bypass-VXLAN* ip address *10.1.9.5 32* #	interface LoopBack0 description *VTEP* ip address *10.1.7.2 32* # interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.6 32* # interface LoopBack2 description *Bypass-VXLAN* ip address *10.1.9.6 32* #	配置Loopback接口。 Loopback0用作VTEP IP，两台Leaf的地址必须配置一样。 Loopback1用作Router-ID/建立BGP EVPN对等体时发送BGP报文的源接口。 Loopback2用作静态Bypass VXLAN隧道的源端IP地址。

步骤2 配置M-LAG

ServerLeaf-1	ServerLeaf-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
ip vpn-instance *DAD* ipv4-family route-distinguisher *5:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *6:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # mac-address m-lag notification evpn disable # 	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # mac-address m-lag notification evpn disable # 	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。
interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。
vlan batch *100* # interface vlanif *100* reserved for vxlan bypass ip address *10.1.100.5 30* # 	vlan batch *100* # interface vlanif *100* reserved for vxlan bypass ip address *10.1.100.6 30* # 	配置静态Bypass VXLAN隧道用到的VLAN及VLANIF的IP地址。该VLAN不能划分给其他业务使用。
ip route-static *10.1.9.6 32 10.1.100.6* preference 1 	ip route-static *10.1.9.5 32 10.1.100.5* preference 1 	配置静态路由，打通Bypass VXLAN隧道，该路由目的地址为对端Loopback2地址，下一跳为对端VLANIF接口地址。
interface nve 1 pip-source *10.1.9.5* peer *10.1.9.6* bypass # 	interface nve 1 pip-source *10.1.9.6* peer *10.1.9.5* bypass # 	创建静态Bypass VXLAN隧道，指定源端地址和对端地址。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。
  ```
  [~ServerLeaf-1] ping -vpn-instance DAD 192.168.10.2
    PING 192.168.10.2: 56  data bytes, press CTRL_C to break
  ```

```
  Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
  Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
  Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
  Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
  Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

 --- 192.168.10.2 ping statistics ---
  5 packet(s) transmitted
  5 packet(s) received
  0.00% packet loss
  round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~ServerLeaf-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW :
10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
   Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
   Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
   Input: 62343901 packets,93259180813 bytes
          42820517 unicast,132 broadcast,19523252 multicast
          0 errors,0 drops
   Output:48323130 packets,92501426032 bytes
          29723661 unicast,128 broadcast,18599341 multicast
          0 errors,0 drops
   Last 10 seconds input utility rate:  0.01%
   Last 10 seconds output utility rate: 0.01%
-------------------------------------------------------------
PortName                  Status            Weight
-------------------------------------------------------------
100GE1/0/5                UP                1
100GE1/0/6                UP                1
-------------------------------------------------------------
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~ServerLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                   Working Mode: Static
Preempt Delay: Disabled         Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
-------------------------------------------------------------------------------
ActorPortName             Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/5(r)             Selected 100GE    32768   6      65      10111100  1
100GE1/0/6(r)             Selected 100GE    32768   6      65      10111100  1

Partner:
-------------------------------------------------------------------------------
ActorPortName             SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/5                32768    00fd-dffb-9a03  32768   6      65      10111100
100GE1/0/6                32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。


```
[~ServerLeaf-1] display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance       : DAD
  State              : Backup
  Causation          : -
  System ID          : 00fd-fdfd-b703
  SysName            : ServerLeaf-2
  Version            : V300R022C00
  Device Type        : CE6800
```

步骤3 配置Underlay路由协议，实现三层互通

本章节EBGP路由协议为例。

ServerLeaf-1	ServerLeaf-2	命令说明
bfd # bgp *65300* router-id *10.1.8.5* auto-frr advertise lowest-priority all-address-family peer-up delay 360	bfd # bgp *65300* router-id *10.1.8.5* auto-frr advertise lowest-priority all-address-family peer-up delay 360	全局使能BFD功能。 配置BGP AS号及相应Router-ID。 开启BGP Auto FRR功能，对于从不同对等体学到的相同前缀的路由，利用最优路由作为主链路进行转发，并自动将次优路由作为备份链路。 在邻居状态由Down到Up时将BGP路由的优先级调整为最低优先级，路由延时发布，解决回切场景丢包时间长问题。
group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.1* group *Group_Spine* peer *192.168.1.17* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 network *10.1.7.2 255.255.255.255* network *10.1.8.5 255.255.255.255* maximum load-balancing 32	group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.5* group *Group_Spine* peer *192.168.1.21* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 network *10.1.7.2 255.255.255.255* network *10.1.8.6 255.255.255.255* maximum load-balancing 32	创建一个用于与Spine对接对等体组，简化后续配置。 配置对等体组的BFD功能，并设置BFD参数。 发布VTEP IP用于建立VXLAN隧道。 发布Router-ID。

ServerLeaf-1	ServerLeaf-2	命令说明
interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	将Leaf上与Spine互联的接口配置为M-LAG上行口，上行链路和Peer-link链路同时故障的设备的端口优先被Error-down。
vlan *100* # interface vlanif *100* ip address *10.1.100.5 30* # bgp *65300* peer *10.1.100.6* as-number *65300* peer *10.1.100.6* connect-interface *vlanif100* peer *10.1.100.6* next-hop-local #	vlan *100* # interface vlanif *100* ip address *10.1.100.6 30* # bgp *65300* peer *10.1.100.5* as-number *65300* peer *10.1.100.5* connect-interface *vlanif100* peer *10.1.100.5* next-hop-local #	配置三层逃生路径。通过peer-link上配置专用的VLANIF三层直连，加入到iBGP路由协议中。 当M-LAG组中一个Leaf与Spine互联的上行链路全部故障时，通过Leaf间的逃生路径，将流量转发至M-LAG组内另外一台Leaf上继续转发。

上述配置完成后，可以通过**display bgp peer**命令查看BGP邻居状态，或通过**display bgp routing-table**命令查看BGP路由信息。

步骤4 配置BGP EVPN路由

使用BGP EVPN作为VXLAN的控制平面，Leaf作为BGP路由反射器客户端，Spine作为BGP路由器反射器，在Leaf与Spine之间建立EVPN IBGP邻居。

ServerLeaf-1	ServerLeaf-2	命令说明
evpn-overlay enable	evpn-overlay enable	使能EVPN作为VXLAN的控制平面。
bgp 100 instance *evpn1* router-id *10.1.8.5* group *Spine* internal peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* connect-interface *LoopBack1* #	bgp 100 instance *evpn1* router-id *10.1.8.6* group *Spine* internal peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* connect-interface *LoopBack1* #	配置BGP EVPN。当Underlay路由为EBGP时，此处需单独构造一个AS号和BGP实例，用于EVPN IBGP对等体配置。 配置Spine的对等体组并加入相应对等体。 指定发送BGP报文的源接口。
l2vpn-family evpn policy vpn-target peer *Spine* enable peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* advertise irb peer *Spine* advertise irbv6 #	l2vpn-family evpn policy vpn-target peer *Spine* enable peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* advertise irb peer *Spine* advertise irbv6 #	使能并进入BGP EVPN地址族视图。 对接收的VPN路由或者标签块进行VPN-Target过滤。 配置BGP EVPN对等体组，并发布irb和irbv6路由。

上述配置完成后，可以通过**display bgp instance** *instance-name* **evpn peer**查看BGP EVPN邻居状态。

步骤5 配置VPN实例及EVPN实例

ServerLeaf-1	ServerLeaf-2	命令说明
ip vpn-instance *vpn1* ipv4-family route-distinguisher *5:5000* vpn-target *0:5000* evpn ipv6-family route-distinguisher *5:5000* vpn-target *0:5000* evpn vxlan vni *5000* # bridge-domain *10* vxlan vni *10* evpn route-distinguisher *5:10* vpn-target *0:10* vpn-target *0:5000* export-extcommunity # bridge-domain *20* vxlan vni *20* evpn route-distinguisher *5:20* vpn-target *0:20* vpn-target *0:5000 export-extcommunity* # bgp 100 ipv4-family vpn-instance *vpn1* import-route direct maximum load-balancing 32 advertise l2vpn evpn ipv6-family vpn-instance *vpn1* import-route static maximum load-balancing 32 advertise l2vpn evpn # interface nve 1 source *10.1.7.2* mac-address *0000-5e00-0102* vni *10* head-end peer-list protocol bgp vni *20* head-end peer-list protocol bgp #	ip vpn-instance *vpn1* ipv4-family route-distinguisher *6:5000* vpn-target *0:5000* evpn ipv6-family route-distinguisher *6:5000* vpn-target *0:5000* evpn vxlan vni *5000* # bridge-domain *10* vxlan vni *10* evpn route-distinguisher *6:10* vpn-target *0:10* vpn-target *0:5000* export-extcommunity # bridge-domain *20* vxlan vni *20* evpn route-distinguisher *6:20* vpn-target *0:20* vpn-target *0:5000 export-extcommunity* # bgp 100 ipv4-family vpn-instance *vpn1* import-route direct maximum load-balancing 32 advertise l2vpn evpn ipv6-family vpn-instance *vpn1* import-route static maximum load-balancing 32 advertise l2vpn evpn # interface nve 1 source *10.1.7.2* mac-address *0000-5e00-0102* vni *10* head-end peer-list protocol bgp vni *20* head-end peer-list protocol bgp #	配置VPN实例。 配置EVPN实例。 两台M-LAG设备上配置的Bridge Domain需要相同。 配置BGP引入直连路由。 配置NVE。两台设备上配置的NVE接口的IP地址和MAC地址需要相同。

上述配置完成后，可以通过**display ip routing-table vpn-instance** *vpn-name*、**display ipv6 routing-table vpn-instance** *vpn-name*查看VPN中通过本地引入或BGP EVPN发布的路由信息。

步骤6 配置三层网关

ServerLeaf-1	ServerLeaf-2	命令说明
interface vbdif *10* ip binding vpn-instance *vpn1* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd ra prefix *fc00::10:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0110* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean # interface vbdif *20* ip binding vpn-instance *vpn1* ip address *10.1.20.1 24* ipv6 enable ipv6 address *fc00::20:1 112* ipv6 nd ra prefix *fc00::20:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0120* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean #	interface vbdif *10* ip binding vpn-instance *vpn1* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd ra prefix *fc00::10:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0110* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean # interface vbdif *20* ip binding vpn-instance *vpn1* ip address *10.10.20.1 24* ipv6 enable ipv6 address *fc00::20:1 112* ipv6 nd ra prefix *fc00::20:0/112* 0 0 no-autoconfig off-link mac-address *0000-5e00-0120* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd ra halt disable ipv6 nd autoconfig managed-address-flag ipv6 nd autoconfig other-flag ipv6 nd na glean #	配置三层网关VBDIF。 两台设备上配置的VBDIF接口的IP地址和MAC地址需要相同

步骤7 配置服务器接入

- 服务器负载分担方式接入时

ServerLeaf-1	ServerLeaf-2	命令说明
interface eth-trunk *1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *10GE1/0/1* storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* #	interface eth-trunk *1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *10GE1/0/1* storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* #	配置接入用的链路聚合组。 按需放通VLAN，不放通VLAN1，防止成环。 配置静态LACP模式。 配置边缘端口。 配置未知单播抑制，建议为端口带宽的5%。 配置组播报文抑制，建议为1 Mbit/s。 配置广播报文抑制，建议为1 Mbit/s。

ServerLeaf-1	ServerLeaf-2	命令说明
interface *10GE1/0/1* lacp force-up # 	-	（可选）在服务器需要进行PXE安装的场景，其中一台设备的成员口需要配置lacp force-up，该成员口为对接安装操作系统的网卡的接口。
interface eth-trunk *1.10* mode l2 encapsulation dot1q vid *10* bridge-domain *10* # 	interface eth-trunk *1.10* mode l2 encapsulation dot1q vid *10* bridge-domain *10* # 	配置业务接入点。

- 服务器主备方式接入时

ServerLeaf-1	ServerLeaf-2	命令说明
interface *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *20* stp edged-port enable storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* # 	interface *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *20* stp edged-port enable storm suppression unknown-unicast *5* storm suppression multicast cir *1 mbps* storm suppression broadcast cir *1 mbps* # 	配置服务器接入端口。 按需放通VLAN，不放通VLAN1，防止成环。 配置边缘端口。 配置未知单播抑制，建议为端口带宽的5%。 配置组播报文抑制，建议为1 Mbit/s。 配置广播报文抑制，建议为1 Mbit/s。
interface *10GE1/0/2.20* mode l2 encapsulation dot1q vid *20* bridge-domain *20* # 	interface *10GE1/0/2.20* mode l2 encapsulation dot1q vid *20* bridge-domain *20* # 	配置业务接入点。

上述配置完成后，可以通过**display dfs-group 1 node 1 m-lag** [**brief**]命令查看M-LAG链路聚合组的状态。

```
[~ServerLeaf-1] display dfs-group 1 node 1 m-lag brief
* - Local node

M-Lag ID     Interface      Port State    Status              Consistency-check
     1     Eth-Trunk 1     Up          active(*)-active     success
     2     Eth-Trunk 2     Up          active(*)-inactive   success
    10     Eth-Trunk 10    Up          active(*)-active     success
```

可以通过**display mac-address** [**bridge-domain** *bridge-domain-id*]命令查看设备学习到的MAC地址信息。

```
[~ServerLeaf-1] display mac-address
Flags: * - Backup
       # - forwarding logical interface, operations cannot be performed based
           on the interface.
BD   : bridge-domain   Age : dynamic MAC learned time in seconds
--------------------------------------------------------------------------------
MAC Address    VLAN/VSI/BD   Learned-From        Type                Age
--------------------------------------------------------------------------------
9c7d-a378-3c8d -/-/10        Eth-Trunk1.10       dynamic             319
cc64-a668-6814 -/-/20        Eth-Trunk2.20       dynamic             319
--------------------------------------------------------------------------------
```

步骤8 配置优化命令

ServerLeaf-1	ServerLeaf-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	配置VLAN 1的流量抑制功能，防止广播风暴。

步骤9 （可选）配置DHCP Relay

- DHCP Client与DHCP Server属于同一VPN场景

ServerLeaf-1	ServerLeaf-2	命令说明
bridge-domain *30* vxlan vni *30* evpn route-distinguisher *5:30* vpn-target *0:30* vpn-target *0:5000* export-extcommunity # interface vbdif *30* ip binding vpn-instance *vpn1* ip address *10.1.30.5 24* ipv6 enable ipv6 address *fc00::30:5 112* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd na glean # interface nve 1 vni *30* head-end peer-list protocol bgp # 	bridge-domain *30* vxlan vni *30* evpn route-distinguisher *6:30* vpn-target *0:30* vpn-target *0:5000* export-extcommunity # interface vbdif *30* ip binding vpn-instance *vpn1* ip address *10.1.30.6 24* ipv6 enable ipv6 address *fc00::30:6 112* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd na glean # interface nve 1 vni *30* head-end peer-list protocol bgp # 	在作为DHCP Relay的Leaf上配置DHCP专用VBDIF接口，作为发送DHCP Realy报文的源接口。 两台M-LAG设备需要配置不同的接口IP，避免DHCP报文来回路径不一致的问题。
dhcp enable # interface vbdif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* dhcp relay source-ip-address *10.1.30.5* dhcp relay information enable dhcp option82 vendor-specific format vendor-sub-option 10 source-ip-address *10.1.30.5* dhcpv6 relay destination *fc00::200:10* dhcpv6 relay source-ip-address *fc00::30:5* # 	dhcp enable # interface vbdif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* dhcp relay source-ip-address *10.1.30.6* dhcp relay information enable dhcp option82 vendor-specific format vendor-sub-option 10 source-ip-address *10.1.30.6* dhcpv6 relay destination *fc00::200:10* dhcpv6 relay source-ip-address *fc00::30:6* # 	在业务网关VBDIF下配置DHCP Relay，以VBDIF10配置为例。 *10.1.200.10*/*fc00::200:10*为DHCP Server的IPv4/IPv6地址，该地址需要和DHCP专用VBDIF接口的IPv4/IPv6地址三层互通。 DHCP Server和作为DHCP Client的业务服务器部署在同一个VPN中。部署DHCP Server的相关配置请参考上文业务VPN的配置，此处略。

- DHCP Client与DHCP Server跨VPN场景

ServerLeaf-1	ServerLeaf-2	命令说明
ip vpn-instance *DHCP* ipv4-family route-distinguisher *5:6000* vpn-target *0:6000* evpn ipv6-family route-distinguisher *5:6000* vpn-target *0:6000* evpn vxlan vni *6000* # bridge-domain *30* vxlan vni *30* evpn route-distinguisher *5:30* vpn-target *0:30* vpn-target *0:6000* export-extcommunity # interface vbdif *30* ip binding vpn-instance *DHCP* ip address *10.1.30.5 24* ipv6 enable ipv6 address *fc00::30:5 112* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd na glean # interface nve 1 vni *30* head-end peer-list protocol bgp # bgp 100 ipv4-family vpn-instance *DHCP* import-route direct maximum load-balancing 32 advertise l2vpn evpn ipv6-family vpn-instance *DHCP* import-route static maximum load-balancing 32 advertise l2vpn evpn # dhcp enable # interface vbdif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* vpn-instance *DHCP* dhcp relay giaddr source-interface *vbdif30* dhcp option82 link-selection insert enable dhcp option82 vss-control insert enable dhcp option82 server-id-override insert enable dhcpv6 relay destination *fc00::200:10* vpn-instance *DHCP* dhcpv6 relay option79 insert enable dhcpv6 relay vss-control insert enable dhcpv6 relay source-ip-address *fc00::30:5* # ip vpn-instance *DHCP* ipv4-family route-distinguisher *5:6000* vpn-target *0:6000* evpn ipv6-family route-distinguisher *5:6000* vpn-target *0:6000* evpn vxlan vni *6000* # bridge-domain *30* vxlan vni *30* evpn route-distinguisher *5:30* vpn-target *0:30* vpn-target *0:6000* export-extcommunity # interface vbdif *30* ip binding vpn-instance *DHCP* ip address *10.1.30.5 24* ipv6 enable ipv6 address *fc00::30:5 112* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd na glean # interface nve 1 vni *30* head-end peer-list protocol bgp # bgp 100 ipv4-family vpn-instance *DHCP* import-route direct maximum load-balancing 32 advertise l2vpn evpn ipv6-family vpn-instance *DHCP* import-route static maximum load-balancing 32 advertise l2vpn evpn # dhcp enable # interface vbdif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* vpn-instance *DHCP* dhcp relay giaddr source-interface *vbdif30* dhcp option82 link-selection insert enable dhcp option82 vss-control insert enable dhcp option82 server-id-override insert enable dhcpv6 relay destination *fc00::200:10* vpn-instance *DHCP* dhcpv6 relay option79 insert enable dhcpv6 relay vss-control insert enable dhcpv6 relay source-ip-address *fc00::30:6* # ip vpn-instance *DHCP* ipv4-family route-distinguisher *6:6000* vpn-target *0:6000* evpn ipv6-family route-distinguisher *6:6000* vpn-target *0:6000* evpn vxlan vni *6000* # bridge-domain *200* vxlan vni *200* evpn route-distinguisher *6:30* vpn-target *0:30* vpn-target *0:6000* export-extcommunity # interface vbdif *30* ip binding vpn-instance *DHCP* ip address *10.1.30.6 24* ipv6 enable ipv6 address *fc00::30:6 112* vxlan anycast-gateway enable arp collect host enable arp broadcast-detect enable ipv6 nd collect host enable ipv6 nd na glean # interface nve 1 vni *30* head-end peer-list protocol bgp # bgp 100 ipv4-family vpn-instance *DHCP* import-route direct maximum load-balancing 32 advertise l2vpn evpn ipv6-family vpn-instance *DHCP* import-route static maximum load-balancing 32 advertise l2vpn evpn # dhcp enable # interface vbdif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* vpn-instance *DHCP* dhcp relay giaddr source-interface *vbdif30* dhcp option82 link-selection insert enable dhcp option82 vss-control insert enable dhcp option82 server-id-override insert enable dhcpv6 relay destination *fc00::200:10* vpn-instance *DHCP* dhcpv6 relay option79 insert enable dhcpv6 relay vss-control insert enable dhcpv6 relay source-ip-address *fc00::30:6* # dhcp enable # interface vbdif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* vpn-instance *DHCP* dhcp relay giaddr source-interface *vbdif30* dhcp option82 link-selection insert enable dhcp option82 vss-control insert enable dhcp option82 server-id-override insert enable dhcpv6 relay destination *fc00::200:10* vpn-instance *DHCP* dhcpv6 relay option79 insert enable dhcpv6 relay vss-control insert enable dhcpv6 relay source-ip-address *fc00::30:5* # dhcp enable # interface vbdif *10* dhcp select relay dhcp relay server-ip *10.1.200.10* vpn-instance *DHCP* dhcp relay giaddr source-interface *vbdif30* dhcp option82 link-selection insert enable dhcp option82 vss-control insert enable dhcp option82 server-id-override insert enable dhcpv6 relay destination *fc00::200:10* vpn-instance *DHCP* dhcpv6 relay option79 insert enable dhcpv6 relay vss-control insert enable dhcpv6 relay source-ip-address *fc00::30:6* #	配置DHCP专用VPN。 在作为DHCP Relay的Leaf上配置DHCP专用VBDIF接口，作为发送DHCP Realy报文的源接口。 两台M-LAG设备需要配置不同的接口IP，避免DHCP报文来回路径不一致的问题。 配置BGP引入直连路由。 在业务网关VBDIF下配置DHCP Relay，以VBDIF10配置为例。 *10.1.200.10*/*fc00::200:10*为DHCP Server的IPv4/IPv6地址，该地址需要和DHCP专用VBDIF接口的IPv4/IPv6地址三层互通。 DHCP Server即部署在DHCP专用VPN中。部署DHCP Server的相关配置请参考上文业务VPN的配置，此处略。

----结束

2.3.3 配置 Spine

配置概览

1. **配置互联接口IP地址及Loopback接口**
2. **配置Underlay路由协议，实现三层互通**
3. **配置BGP EVPN路由**
4. **配置优化命令**

配置步骤

步骤1 配置互联接口IP地址及Loopback接口

Spine-1	Spine-2	命令说明
interface *100GE1/0/1* description *to-ServerLeaf-1* undo portswitch ip address *192.168.1.1 30* # interface *100GE2/0/1* description *to-ServerLeaf-2* undo portswitch ip address *192.168.1.5 30* # interface *100GE1/0/2* description *to-ServerLeaf-3* undo portswitch ip address *192.168.1.9 30* # interface *100GE2/0/2* description *to-ServerLeaf-4* undo portswitch ip address *192.168.1.13 30* # interface *100GE1/0/3* description *to-BorderLeaf-1* undo portswitch ip address *192.168.1.33 30* # interface *100GE2/0/3* description *to-BorderLeaf-2* undo portswitch ip address *192.168.1.37 30* # 	interface *100GE1/0/1* description *to-ServerLeaf-1* undo portswitch ip address *192.168.1.17 30* # interface *100GE2/0/1* description *to-ServerLeaf-2* undo portswitch ip address *192.168.1.21 30* # interface *100GE1/0/2* description *to-ServerLeaf-3* undo portswitch ip address *192.168.1.25 30* # interface *100GE2/0/2* description *to-ServerLeaf-4* undo portswitch ip address *192.168.1.29 30* # interface *100GE1/0/3* description *to-BorderLeaf-1* undo portswitch ip address *192.168.1.41 30* # interface *100GE2/0/3* description *to-BorderLeaf-2* undo portswitch ip address *192.168.1.45 30* # 	配置与Leaf互联接口。
interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.3 32* # 	interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.4 32* # 	配置Loopback接口。 Loopback1用作Router-ID/建立BGP EVPN对等体时发送BGP报文的源接口。

步骤2 配置Underlay路由协议，实现三层互通

本章节EBGP路由协议为例。

Spine-1	Spine-2	命令说明
bfd # bgp *65200* router-id *10.1.8.3* advertise lowest-priority all-address-family peer-up delay 360	bfd # bgp *65200* router-id *10.1.8.4* advertise lowest-priority all-address-family peer-up delay 360	全局使能BFD功能。 配置BGP AS号及相应Router-ID。 在邻居状态由Down到Up时将BGP路由的优先级调整为最低优先级，路由延时发布，解决回切场景丢包时间长问题。
group *Group_ServerLeaf1* external peer *Group_ServerLeaf1* as-number *65300* peer *192.168.1.2* group *Group_ServerLeaf1* peer *192.168.1.6* group *Group_ServerLeaf1* group *Group_ServerLeaf2* external peer *Group_ServerLeaf2* as-number *65400* peer *192.168.1.10* group *Group_ServerLeaf2* peer *192.168.1.14* group *Group_ServerLeaf2* group *Group_BorderLeaf* external peer *Group_BorderLeaf* as-number *65100* peer *192.168.1.34* group *Group_BorderLeaf* peer *192.168.1.38* group*Group_BorderLeaf* peer *Group_ServerLeaf1* bfd enable peer *Group_ServerLeaf1* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_ServerLeaf2* bfd enable peer *Group_ServerLeaf2* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_BorderLeaf* bfd enable peer *Group_BorderLeaf* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 network *110.1.8.3 255.255.255.255* maximum load-balancing 32	group *Group_ServerLeaf1* external peer *Group_ServerLeaf1* as-number *65300* peer *192.168.1.18* group *Group_ServerLeaf1* peer *192.168.1.22* group *Group_ServerLeaf1* group *Group_ServerLeaf2* external peer *Group_ServerLeaf2* as-number *65400* peer *192.168.1.26* group *Group_ServerLeaf2* peer *192.168.1.30* group *Group_ServerLeaf2* group *Group_BorderLeaf* external peer *Group_BorderLeaf* as-number *65100* peer *192.168.1.42* group *Group_BorderLeaf* peer *192.168.1.46* group*Group_BorderLeaf* peer *Group_ServerLeaf1* bfd enable peer *Group_ServerLeaf1* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_ServerLeaf2* bfd enable peer *Group_ServerLeaf2* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 peer *Group_BorderLeaf* bfd enable peer *Group_BorderLeaf* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 network *10.1.8.4 255.255.255.255* maximum load-balancing 32	创建一个用于与ServerLeaf1、ServerLeaf2对接的对等体组，简化后续配置。 创建一个用于与ServerLeaf3、ServerLeaf4对接的对等体组，简化后续配置。 创建一个用于与BorderLeaf1、BorderLeaf2对接的对等体组，简化后续配置。 配置对等体组的BFD功能，并设置BFD参数。 发布Router-ID

上述配置完成后，可以通过**display bgp peer**命令查看BGP邻居状态，或通过**display bgp routing-table**命令查看BGP路由信息。

步骤3 配置BGP EVPN路由

使用BGP EVPN作为VXLAN的控制平面，Leaf作为BGP路由反射器客户端，Spine作为BGP路由器反射器，在Leaf与Spine之间建立EVPN IBGP邻居。

ServerLeaf-1	ServerLeaf-2	命令说明
evpn-overlay enable	evpn-overlay enable	使能EVPN作为VXLAN的控制平面。
bgp 100 instance *evpn1* router-id *10.1.8.3* group *BorderLeaf* internal peer *10.1.8.1* group *BorderLeaf* peer *10.1.8.2* group *BorderLeaf* peer *BorderLeaf* connect-interface *LoopBack1* group *ServerLeaf* internal peer *10.1.8.5* group *ServerLeaf* peer *10.1.8.6* group *ServerLeaf* peer *10.1.8.7* group *ServerLeaf* peer *10.1.8.8* group *ServerLeaf* peer *ServerLeaf* connect-interface *LoopBack1* # l2vpn-family evpn undo policy vpn-target peer *BorderLeaf* enable peer *10.1.8.1* group *BorderLeaf* peer *10.1.8.2* group *BorderLeaf* peer *BorderLeaf* advertise irb peer *BorderLeaf* advertise irbv6 peer *BorderLeaf* reflect-client peer *ServerLeaf* enable peer *10.1.8.5* group *ServerLeaf* peer *10.1.8.6* group *ServerLeaf* peer *10.1.8.7* group *ServerLeaf* peer *10.1.8.8* group *ServerLeaf* peer *ServerLeaf* advertise irb peer *ServerLeaf* advertise irbv6 peer *ServerLeaf* reflect-client # bgp 100 instance *evpn1* router-id *10.1.8.4* group *BorderLeaf* internal peer *10.1.8.1* group *BorderLeaf* peer *10.1.8.2* group *BorderLeaf* peer *BorderLeaf* connect-interface *LoopBack1* group *ServerLeaf* internal peer *10.1.8.5* group *ServerLeaf* peer *10.1.8.6* group *ServerLeaf* peer *10.1.9.7* group *ServerLeaf* peer *10.1.8.8* group *ServerLeaf* peer *ServerLeaf* connect-interface *LoopBack1* # l2vpn-family evpn undo policy vpn-target peer *BorderLeaf* enable peer *10.1.8.1* group *BorderLeaf* peer *10.1.8.2* group *BorderLeaf* peer *BorderLeaf* advertise irb peer *BorderLeaf* advertise irbv6 peer *BorderLeaf* reflect-client peer *ServerLeaf* enable peer *10.1.8.5* group *ServerLeaf* peer *10.1.8.6* group *ServerLeaf* peer *10.1.8.7* group *ServerLeaf* peer *10.1.8.8* group *ServerLeaf* peer *ServerLeaf* advertise irb peer *ServerLeaf* advertise irbv6 peer *ServerLeaf* reflect-client #	配置BGP EVPN。当Underlay路由为EBGP时，此处需单独构造一个AS号和BGP实例，用于EVPN IBGP对等体配置。 配置Spine的对等体组并加入相应对等体。 指定发送BGP报文的源接口。 使能并进入BGP EVPN地址族视图。 取消对接收的VPN路由或者标签块进行VPN-Target过滤。 配置BGP EVPN对等体组，并发布irb和irbv6路由。 配置将本机作为路由反射器，并配置对等体组设备作为其客户端。

上述配置完成后，可以通过**display bgp instance** *instance-name* **evpn peer**查看BGP EVPN邻居状态。

步骤4 配置优化命令

Spine-1	Spine-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.3.4 配置 Border Leaf

配置概览

1. **配置互联接口IP地址及Loopback接口**
2. **配置M-LAG**
3. **配置Underlay路由协议，实现三层互通**
4. **配置BGP EVPN路由**
5. **配置VPN实例及EVPN实例**
6. **配置出口网络**
7. **配置优化命令**

配置步骤

步骤1 配置互联接口IP地址及Loopback接口

BorderLeaf-1	BorderLeaf-2	命令说明
interface *100GE1/0/1* description *to-Spine1* undo portswitch ip address *192.168.1.34 30* qos phb marking dscp enable # interface *100GE1/0/2* description *to-Spine2* undo portswitch ip address *192.168.1.42 30* qos phb marking dscp enable #	interface *100GE1/0/1* description *to-Spine1* undo portswitch ip address *192.168.1.38 30* qos phb marking dscp enable # interface *100GE1/0/2* description *to-Spine2* undo portswitch ip address *192.168.1.46 30* qos phb marking dscp enable #	配置与Spine互联接口。
interface LoopBack0 description *VTEP* ip address *10.1.7.1 32* # interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.1 32* # interface LoopBack2 description *Bypass-VXLAN* ip address *10.1.9.1 32* #	interface LoopBack0 description *VTEP* ip address *10.1.7.1 32* # interface LoopBack1 description *Router-id/BGP* ip address *10.1.8.2 32* # interface LoopBack2 description *Bypass-VXLAN* ip address *10.1.9.2 32* #	配置Loopback接口。 Loopback0用作VTEP IP，两台Leaf的地址必须配置一样。 Loopback1用作Router-ID/建立BGP EVPN对等体时发送BGP报文的源接口。 Loopback2用作静态Bypass VXLAN隧道的源端IP地址。

步骤2 配置M-LAG

BorderLeaf-1	BorderLeaf-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection arp ip-conflict-detect enable #	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection arp ip-conflict-detect enable #	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。 使能设备的IP地址冲突检测的功能。
ip vpn-instance *DAD* ipv4-family route-distinguisher *1:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved #	ip vpn-instance *DAD* ipv4-family route-distinguisher *2:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved #	部署独立直连物理链路，用于M-LAG心跳检测。
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # mac-address m-lag notification evpn disable #	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # mac-address m-lag notification evpn disable #	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。
interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 #	interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 #	配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。
vlan batch *100* # interface vlanif *100* reserved for vxlan bypass ip address *10.1.100.1 30* #	vlan batch *100* # interface vlanif *100* reserved for vxlan bypass ip address *10.1.100.2 30* #	配置静态Bypass VXLAN隧道用到的VLAN及VLANIF的IP地址。该VLAN不能划分给其他业务使用。
ip route-static *10.1.9.2 32 10.1.100.2* preference 1	ip route-static *10.1.9.1 32 10.1.100.1* preference 1	配置静态路由，打通Bypass VXLAN隧道，该路由目的地址为对端Loopback2地址，下一跳为对端VLANIF接口地址。
interface nve 1 pip-source *10.1.9.1* peer *10.1.9.2* bypass #	interface nve 1 pip-source *10.1.9.2* peer *10.1.9.1* bypass #	创建静态Bypass VXLAN隧道，指定源端地址和对端地址。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。

```
[~BorderLeaf-1] ping -vpn-instance DAD 192.168.10.2
  PING 192.168.10.2: 56  data bytes, press CTRL_C to break
    Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

  --- 192.168.10.2 ping statistics ---
    5 packet(s) transmitted
    5 packet(s) received
    0.00% packet loss
    round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~BorderLeaf-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW :
10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
    Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
    Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
    Input: 62343901 packets,93259180813 bytes
           42820517 unicast,132 broadcast,19523252 multicast
           0 errors,0 drops
    Output:48323130 packets,92501426032 bytes
           29723661 unicast,128 broadcast,18599341 multicast
           0 errors,0 drops
    Last 10 seconds input utility rate:  0.01%
    Last 10 seconds output utility rate: 0.01%
--------------------------------------------------------------
PortName                  Status             Weight
--------------------------------------------------------------
100GE1/0/5                UP                 1
100GE1/0/6                UP                 1
--------------------------------------------------------------
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~BorderLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                     Working Mode: Static
Preempt Delay: Disabled         Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
--------------------------------------------------------------------------------
ActorPortName             Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/5(r)             Selected 100GE    32768   6      65      10111100  1
100GE1/0/6(r)             Selected 100GE    32768   6      65      10111100  1

Partner:
--------------------------------------------------------------------------------
ActorPortName             SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/5                32768    00fd-dffb-9a03  32768   6      65      10111100
100GE1/0/6                32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
[~BorderLeaf-1] display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance       : DAD
  State              : Backup
  Causation          : -
  System ID          : 00fd-fdfd-b703
  SysName            : ServerLeaf-2
  Version            : V300R022C00
  Device Type        : CE6800
```

步骤3 配置Underlay路由协议，实现三层互通

本章节EBGP路由协议为例。

BorderLeaf-1	BorderLeaf-1	命令说明
bfd # bgp *65100* router-id *10.1.8.1* auto-frr advertise lowest-priority all-address-family peer-up delay 360	bfd # bgp *65100* router-id *10.1.8.2* auto-frr advertise lowest-priority all-address-family peer-up delay 360	全局使能BFD功能。 配置BGP AS号及相应Router-ID。 开启BGP Auto FRR功能，对于从不同对等体学到的相同前缀的路由，利用最优路由作为主链路进行转发，并自动将次优路由作为备份链路。 在邻居状态由Down到Up时将BGP路由的优先级调整为最低优先级，路由延时发布，解决回切场景丢包时间长问题。
group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.33* group *Group_Spine* peer *192.168.1.41* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 network *10.1.7.2 255.255.255.255* network *10.1.8.5 255.255.255.255* maximum load-balancing 32	group *Group_Spine* external peer *Group_Spine* as-number *65200* peer *Group_Spine* allow-as-loop peer *192.168.1.37* group *Group_Spine* peer *192.168.1.45* group *Group_Spine* peer *Group_Spine* bfd enable peer *Group_Spine* bfd min-tx-interval 500 min-rx-interval 500 detect-multiplier 3 ipv4-family unicast preference 20 200 10 network *10.1.7.2 255.255.255.255* network *10.1.8.5 255.255.255.255* maximum load-balancing 32	创建一个用于与Spine对接对等体组，简化后续配置。 配置对等体组的BFD功能，并设置BFD参数。 发布VTEP IP用于建立VXLAN隧道。 发布Router-ID。

BorderLeaf-1	BorderLeaf-1	命令说明
interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	interface *100GE1/0/1* uplink-port enable # interface *100GE1/0/2* uplink-port enable #	将Leaf上与Spine互联的接口配置为M-LAG上行口，上行链路和Peer-link链路同时故障的设备的端口优先被Error-down。
vlan *100* # interface vlanif *100* ip address *10.1.100.1 30* # bgp *65100* peer *10.1.100.2* as-number *65100* peer *10.1.100.2* connect-interface *vlanif100* peer *10.1.100.2* next-hop-local #	vlan *100* # interface vlanif *100* ip address *10.1.100.2 30* # bgp *65100* peer *10.1.100.1* as-number *65100* peer *10.1.100.1* connect-interface *vlanif100* peer *10.1.100.1* next-hop-local #	配置三层逃生路径。通过peer-link上配置专用的VLANIF三层直连，加入到iBGP路由协议中。 当M-LAG组中一个Leaf与Spine互联的上行链路全部故障时，通过Leaf间的逃生路径，将流量转发至M-LAG组内另外一台Leaf上继续转发。

上述配置完成后，可以通过**display bgp peer**命令查看BGP邻居状态，或通过**display bgp routing-table**命令查看BGP路由信息。

步骤4 配置BGP EVPN路由

使用BGP EVPN作为VXLAN的控制平面，Leaf作为BGP路由反射器客户端，Spine作为BGP路由器反射器，在Leaf与Spine之间建立EVPN IBGP邻居。

ServerLeaf-1	ServerLeaf-2	命令说明
evpn-overlay enable	evpn-overlay enable	使能EVPN作为VXLAN的控制平面。
bgp 100 instance *vpn1* router-id *10.1.8.1* group *Spine* internal peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* connect-interface *LoopBack1* #	bgp 100 instance *vpn1* router-id *10.1.8.2* group *Spine* internal peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* connect-interface *LoopBack1* #	配置BGP EVPN。当Underlay路由为EBGP时，此处需单独构造一个AS号和BGP实例，用于EVPN IBGP对等体配置。 配置Spine的对等体组并加入相应对等体。 指定发送BGP报文的源接口。
l2vpn-family evpn policy vpn-target peer *Spine* enable peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* advertise irb peer *Spine* advertise irbv6 #	l2vpn-family evpn policy vpn-target peer *Spine* enable peer *10.1.8.3* group *Spine* peer *10.1.8.4* group *Spine* peer *Spine* advertise irb peer *Spine* advertise irbv6 #	使能并进入BGP EVPN地址族视图。 对接收的VPN路由或者标签块进行VPN-Target过滤。 配置BGP EVPN对等体组，并发布irb和irbv6路由。

上述配置完成后，可以通过**display bgp instance** *instance-name* **evpn peer**查看BGP EVPN邻居状态。

步骤5 配置VPN实例及EVPN实例

ServerLeaf-1	ServerLeaf-2	命令说明
ip vpn-instance *vpn1* ipv4-family route-distinguisher *1:5000* vpn-target *0:5000* evpn ipv6-family route-distinguisher *1:5000* vpn-target *0:5000* evpn vxlan vni *5000* # 	ip vpn-instance *vpn1* ipv4-family route-distinguisher *2:5000* vpn-target *0:5000* evpn ipv6-family route-distinguisher *2:5000* vpn-target *0:5000* evpn vxlan vni *5000* # 	配置VPN实例。
bgp 100 ipv4-family vpn-instance *vpn1* default-route imported import-route static maximum load-balancing 32 advertise l2vpn evpn ipv6-family vpn-instance *vpn1* default-route imported import-route static maximum load-balancing 32 advertise l2vpn evpn # 	bgp 100 ipv4-family vpn-instance *vpn1* default-route imported import-route static maximum load-balancing 32 advertise l2vpn evpn ipv6-family vpn-instance *vpn1* default-route imported import-route static maximum load-balancing 32 advertise l2vpn evpn # 	配置BGP引入静态路由。
interface nve 1 source *10.1.7.1* mac-address *0000-5e00-0101* # 	interface nve 1 source *10.1.7.1* mac-address *0000-5e00-0101* # 	配置NVE。两台设备上配置的NVE接口的IP地址和MAC地址需要相同。

上述配置完成后，可以通过**display ip routing-table vpn-instance** *vpn-name*、**display ipv6 routing-table vpn-instance** *vpn-name*查看VPN中通过本地引入或BGP EVPN发布的路由信息。

步骤6 配置出口网络

BorderLeaf-1	BorderLeaf-2	命令说明
ip vpn-instance *Internet* ipv4-family route-distinguisher *1:99* ipv6-family route-distinguisher *1:99* # 	ip vpn-instance *Internet* ipv4-family route-distinguisher *2:99* ipv6-family route-distinguisher *2:99* # 	创建出口VRF。
interface *Eth-Trunk99* description *to-PE-1* trunkport *10GE1/0/9* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.1 30* ipv6 enable ipv6 address *fc00::99:1 126* mode lacp-static # interface *10GE1/0/9* set up-delay 300 # 	interface *Eth-Trunk99* description *to-PE-2* trunkport *10GE1/0/9* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.5 30* ipv6 enable ipv6 address *fc00::99:5 126* mode lacp-static # interface *10GE1/0/9* set up-delay 300 # 	配置与PE互联接口，创建三层Eth-Trunk口，并加入对应物理成员口。 Border Leaf与PE互联接口配置延时UP，防止设备重启后，路由下发较慢或者下行隧道建立较慢导致业务流量回切时间较长。

BorderLeaf-1	BorderLeaf-2	命令说明
interface *Eth-Trunk20* trunkport *10GE1/0/3* trunkport *10GE1/0/4* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.9 30* ipv6 enable ipv6 address *fc00::99:9 126* mode lacp-static m-lag unpaired-port reserved # 	interface *Eth-Trunk20* trunkport *10GE1/0/3* trunkport *10GE1/0/4* undo portswitch ip binding vpn-instance *Internet* ip address *10.1.99.10 30* ipv6 enable ipv6 address *fc00::99:10 126* mode lacp-static m-lag unpaired-port reserved # 	配置Border Leaf间出口逃生路径，单台Border Leaf与出口路由器互联的上行链路同时中断时生效。 出于高可靠性考虑，逃生路径建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。
ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.2* ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:2* 	ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.6* ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:6* 	配置至出口PE的静态路由。
ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.10* preference 100 ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:10* preference 100 	ip route-static vpn-instance *Internet* 0.0.0.0 0.0.0.0 *10.1.99.9* preference 100 ipv6 route-static vpn-instance *Internet* :: 0 *fc00::99:9* preference 100 	配置出口逃生路由，设置为低优先级。
ip route-static vpn-instance *Internet* *10.1.10.0 24* vpn-instance *vpn1* ip route-static vpn-instance *Internet* *10.1.20.0 24* vpn-instance *vpn1* ipv6 route-static vpn-instance *Internet* *fc00::10: 112* vpn-instance *vpn1* ipv6 route-static vpn-instance *Internet* *fc00::20: 112* vpn-instance *vpn1* 	ip route-static vpn-instance *Internet* *10.1.10.0 24* vpn-instance *vpn1* ip route-static vpn-instance *Internet* *10.1.20.0 24* vpn-instance *vpn1* ipv6 route-static vpn-instance *Internet* *fc00::10: 112* vpn-instance *vpn1* ipv6 route-static vpn-instance *Internet* *fc00::20: 112* vpn-instance *vpn1* 	配置至业务网段的静态路由，下一跳为业务VRF。
ip route-static vpn-instance *vpn1* 0.0.0.0 0.0.0.0 vpn-instance *Internet* ipv6 route-static vpn-instance *vpn1* :: 0 vpn-instance *Internet* 	ip route-static vpn-instance *vpn1* 0.0.0.0 0.0.0.0 vpn-instance *Internet* ipv6 route-static vpn-instance *vpn1* :: 0 vpn-instance *Internet* 	配置业务VRF的静态路由，下一跳为出口VRF。

上述配置完成后，可以检查服务器端和外部网络之间能否ping通。

步骤7 配置优化命令

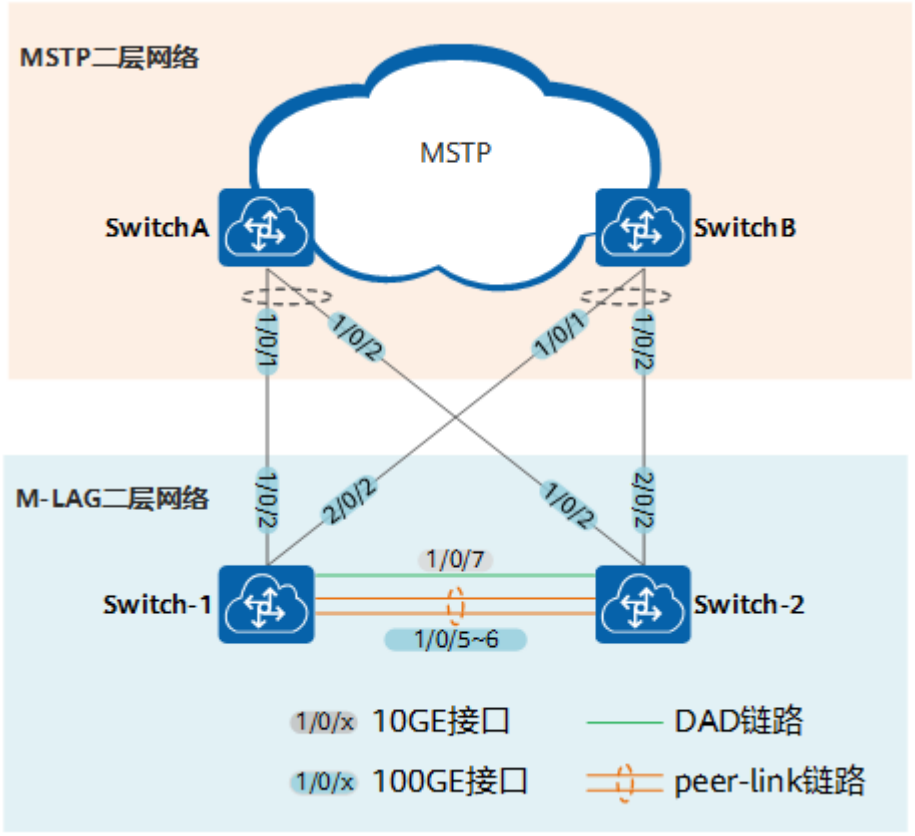
BorderLeaf-1	BorderLeaf-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown # 	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps # 	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.4 M-LAG 与 MSTP 二层网络对接

2.4.1 组网方案

图 2-4 M-LAG 与 MSTP 二层网络对接示意图



如图2-4所示，两台交换机组建M-LAG，与传统MSTP二层网络对接。

- MSTP设备需要配置TC抑制功能（stp tc-restriction enable），防止M-LAG设备抢根，防止M-LAG设备TC报文扩散，导致MSTP设备误认为整网拓扑变化，而清除自身MAC/ARP表项。
- 建议通过单独链路作为M-LAG设备之间双主检测链路，提升可靠性。
- 两台设备组建M-LAG之后，可能产生环路，建议使用V-STP功能防止M-LAG成员口因为环路协议被堵塞。
- 建议至少使用两条链路捆绑成Eth-Trunk链路作为M-LAG peer-link链路，如果是两台框式设备组建M-LAG，建议使用不同单板的接口捆绑成Eth-Trunk链路作为M-LAG peer-link链路。如果使用CE6881、CE6863、CE6881H和CE6863H设备组建M-LAG，建议选择（100GE 1/0/1~1/0/3和100GE 1/0/4~1/0/6）中的接口捆绑Eth-Trunk作为M-LAG peer-link链路。

2.4.2 配置 M-LAG 网络域交换机

配置概览

1. **配置M-LAG**
2. **配置M-LAG与MSTP网络对接**
3. **配置优化命令**

配置步骤

步骤1 配置M-LAG

Switch-1	Switch-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
stp bridge-address *1-1-1* # 	stp bridge-address *1-1-1* # 	配置当前设备参与生成树计算的桥MAC，两台M-LAG设备的桥MAC必须相同，建议选择其中一台设备的系统MAC作为共同桥MAC，不同M-LAG组里的设备桥MAC不同。
ip vpn-instance *DAD* ipv4-family route-distinguisher *1:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *2:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。
interface *Eth-Trunk0* trunkport *100GE1/0/6* trunkport *100GE2/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	interface *Eth-Trunk0* trunkport *100GE1/0/6* trunkport *100GE2/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。

步骤2 配置M-LAG与MSTP网络对接

Switch-1	Switch-2	命令说明
stp priority *32768* 	stp priority *32768* 	降低M-LAG设备的STP优先级，以保证MSTP网络域的根桥不受影响。

Switch-1	Switch-2	命令说明
interface *Eth-Trunk8* description *to-SwitchA* trunkport *100GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *8* #	interface *Eth-Trunk8* description *to-SwitchA* trunkport *100GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *8* #	配置与SwitchA互联链路 按需放通VLAN。
interface *Eth-Trunk9* description *to-SwitchB* trunkport *100GE2/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *9* #	interface *Eth-Trunk9* description *to-SwitchB* trunkport *100GE2/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static dfs-group 1 m-lag *9* #	配置与SwitchB互联链路 按需放通VLAN。

步骤3　配置优化命令

Switch-1	Switch-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.4.3 配置 MSTP 网络域交换机

说明

- 本节点仅以华为CloudEngine系列交换机为例，给出MSTP网络域中和M-LAG对接的SwitchA、SwitchB的配置，其他款型交换机配置请参考对应产品的指导文档。
- V-STP模式M-LAG与MSTP网络对接时，M-LAG设备在MSTP实例0中进行拓扑计算，其他MSTP实例不受影响。支持MSTP多进程。

配置概览

1. **配置MSTP**
2. **配置与Spine互联链路**
3. **配置优化命令**

配置步骤

步骤1 配置MSTP

SwitchA	SwitchB	命令说明
stp mode mstp # stp region-configuration region-name *RG1* instance 0 vlan 10 20 #	stp mode mstp # stp region-configuration region-name *RG1* instance 0 vlan 10 20 #	配置MSTP。

步骤2 配置与Spine互联链路

SwitchA	SwitchB	命令说明
vlan batch *10 20*	vlan batch *10 20*	创建业务VLAN。
interface *Eth-Trunk10* description *to-Spine* trunkport *100GE1/0/1* trunkport *100GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static stp tc-restriction enable #	interface *Eth-Trunk10* description *to-Spine* trunkport *100GE1/0/1* trunkport *100GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 20* mode lacp-static stp tc-restriction enable #	配置级联链路。 按需放通VLAN。 在与Spine互联接口上配置TC/TCN报文抑制功能，使接口收到TC/TCN报文后不刷新本地ARP和MAC表项，也不将TC/TCN报文扩散到本设备其他端口。

步骤3 配置优化命令

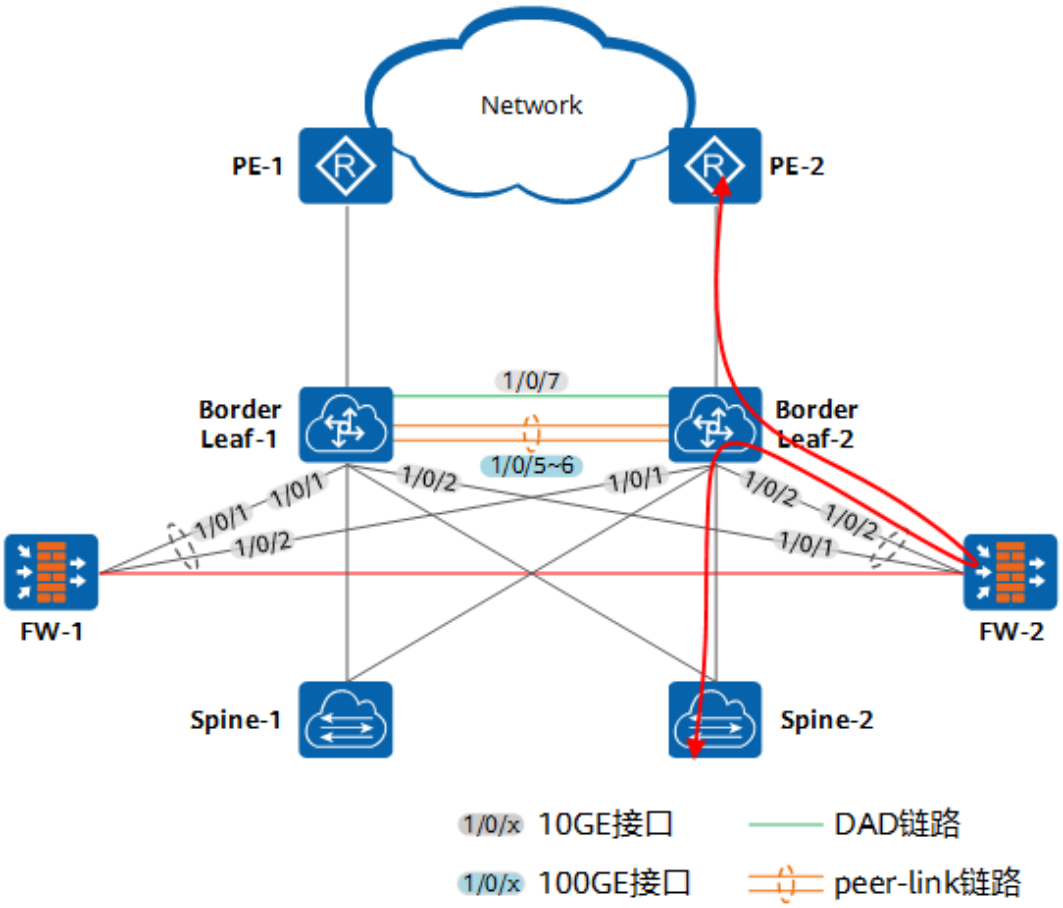
SwitchA	SwitchB	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.5 M-LAG 与防火墙对接

2.5.1 组网方案

图 2-5 M-LAG 对接防火墙组网示意图



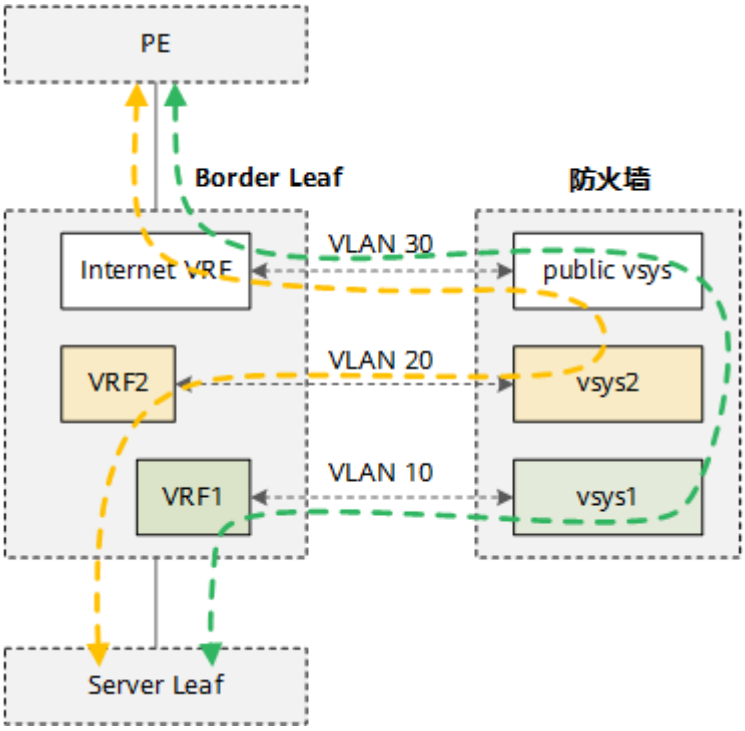
如图2-5所示，网络中两台Border Leaf作为业务出口网关，防火墙（FW）旁挂在Border Leaf上其中：

- Border Leaf部署M-LAG，作为业务网关，通过VLANIF与防火墙对接。
- 防火墙以主备镜像模式部署，旁挂在Border Leaf上。

流量模型如图2-6所示：

- Border Leaf上创建业务VRF及公共出口VRF，防火墙上也创建公共出口vsys及对应业务vsys（防火墙侧也可以使用安全域或VRF进行安全隔离，此处仅以vsys为例）。
- 业务流量进入Border Leaf后，由对应VRF发送至防火墙业务vsys，再经过防火墙公共出口vsys发送至Border Leaf的公共出口VRF，最后发往出口路由器。

图 2-6 过墙流量模型



- 建议通过单独链路作为M-LAG设备之间双主检测链路，提升可靠性。
- 两台设备组建M-LAG之后，可能产生环路，建议使用V-STP功能防止M-LAG成员口因为环路协议被堵塞。如果M-LAG设备需要和友商PVST/PVST+协议对接，则推荐使用VBST协议和PVST/PVST+对接。
- 建议至少使用两条链路捆绑成Eth-Trunk链路作为M-LAG peer-link链路，如果是两台框式设备组建M-LAG，建议使用不同单板的接口捆绑成Eth-Trunk链路作为M-LAG peer-link链路。如果使用CE6881、CE6863、CE6881H和CE6863H设备组建M-LAG，建议选择（100GE 1/0/1~1/0/3和100GE 1/0/4~1/0/6）中的接口捆绑Eth-Trunk作为M-LAG peer-link链路。

2.5.2 配置 Border Leaf

说明

本文仅给出交换机的配置，防火墙配置请参考对应产品的指导文档。

配置概览

1. **配置M-LAG**
2. **配置Leaf与防火墙互联链路**
3. **将过墙流量引入防火墙**
4. **配置优化命令**

配置步骤

步骤1　配置M-LAG

BorderLeaf-1	BorderLeaf-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
ip vpn-instance *DAD* ipv4-family route-distinguisher *1:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *2:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。
interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。

```
[~BorderLeaf-1] ping -vpn-instance DAD 192.168.10.2
  PING 192.168.10.2: 56  data bytes, press CTRL_C to break
    Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

  --- 192.168.10.2 ping statistics ---
    5 packet(s) transmitted
    5 packet(s) received
    0.00% packet loss
    round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~BorderLeaf-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW :
```

```
10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
    Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
    Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
    Input: 62343901 packets,93259180813 bytes
          42820517 unicast,132 broadcast,19523252 multicast
          0 errors,0 drops
    Output:48323130 packets,92501426032 bytes
          29723661 unicast,128 broadcast,18599341 multicast
          0 errors,0 drops
    Last 10 seconds input utility rate:  0.01%
    Last 10 seconds output utility rate: 0.01%
--------------------------------------------------------------
PortName                  Status             Weight
--------------------------------------------------------------
100GE1/0/5                UP                 1
100GE1/0/6                UP                 1
--------------------------------------------------------------
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~BorderLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                   Working Mode: Static
Preempt Delay: Disabled         Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
--------------------------------------------------------------------------------
ActorPortName             Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/5(r)             Selected 100GE    32768   6      65      10111100  1
100GE1/0/6(r)             Selected 100GE    32768   6      65      10111100  1

Partner:
--------------------------------------------------------------------------------
ActorPortName             SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/5                32768    00fd-dffb-9a03  32768   6      65      10111100
100GE1/0/6                32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
[~BorderLeaf-1] display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance       : DAD
  State              : Backup
  Causation          : -
```

```
System ID          : 00fd-fdfd-b703
SysName            : ServerLeaf-2
Version            : V300R022C00
Device Type        : CE6800
```

步骤2 配置Leaf与防火墙互联链路

BorderLeaf-1	BorderLeaf-2	命令说明
ip vpn-instance *vpn1* ipv4-family route-distinguisher *1:10* ipv6-family route-distinguisher *1:10* # ip vpn-instance *Internet* ipv4-family route-distinguisher *1:30* ipv6-family route-distinguisher *1:30* #	ip vpn-instance *vpn1* ipv4-family route-distinguisher *2:10* ipv6-family route-distinguisher *2:10* # ip vpn-instance *Internet* ipv4-family route-distinguisher *1:30* ipv6-family route-distinguisher *1:30* #	创建业务VRF及公共出口VRF。
vlan batch *10 30* # interface vlanif *10* ip binding vpn-instance *vpn1* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd na glean mac-address *0000-5e00-0110* # interface vlanif *30* ip binding vpn-instance *Internet* ip address *10.1.30.1 24* ipv6 enable ipv6 address *fc00::30:1 112* ipv6 nd na glean mac-address *0000-5e00-0130*	vlan batch *10 30* # interface vlanif *10* ip binding vpn-instance *vpn1* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00::10:1 112* ipv6 nd na glean mac-address *0000-5e00-0110* # interface vlanif *30* ip binding vpn-instance *Internet* ip address *10.1.30.1 24* ipv6 enable ipv6 address *fc00::30:1 112* ipv6 nd na glean mac-address *0000-5e00-0130*	创建与FW互联业务VLAN及VLANIF接口。
interface *Eth-Trunk1* description *to-FW-1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 30* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *Eth-Trunk2* description *to-FW-2* trunkport *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 30* mode lacp-static stp edged-port enable dfs-group 1 m-lag *2* #	interface *Eth-Trunk1* description *to-FW-1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 30* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* # interface *Eth-Trunk2* description *to-FW-2* trunkport *10GE1/0/2* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10 30* mode lacp-static stp edged-port enable dfs-group 1 m-lag *2* #	配置与防火墙对接端口。

步骤3 配置路由或策略路由（PBR），将过墙流量引入防火墙

- 通过路由方式，将过墙流量引入防火墙。

BorderLeaf-1	BorderLeaf-2	命令说明
ip route-static vpn-instance *vpn1* *10.1.99.0 24 10.1.10.2* ipv6 route-static vpn-instance *vpn1* *fc00::99:0 112 fc00::10:2*	ip route-static vpn-instance *vpn1* *10.1.99.0 24 10.1.10.2* ipv6 route-static vpn-instance *vpn1* *fc00::99:0 112 fc00::10:2*	配置业务VRF过墙流量的静态路由，下一跳为防火墙。 10.1.99.0/24、fc00::99:0/96为业务目的地址。
ip route-static vpn-instance *Internet* *10.1.1.0 24 10.1.30.2* ipv6 route-static vpn-instance *vpn1* *fc00::1:0 112 fc00:30::2*	ip route-static vpn-instance *Internet* *10.1.0.0 24 10.1.30.2* ipv6 route-static vpn-instance *vpn1* *fc00:1::0 112 fc00:30::2*	配置公共出口VRF至业务网段的静态路由，下一跳为防火墙。 10.1.1.0/24、fc00::1:0/96为业务网段地址。

- 通过策略路由方式，将过墙流量引入防火墙。

BorderLeaf-1	BorderLeaf-2	命令说明
acl number *3001* rule *5* permit ip source *10.1.1.0 0.0.0.255* destination *10.1.99.0 0.0.0.255* # acl ipv6 number *3002* rule *5* permit ipv6 source *fc00::1:0/112* destination *fc00::99:0/112* # traffic classifier *c1* if-match acl *3001* # traffic classifier *c2* if-match ipv6 acl *3002* # traffic behavior *b1* redirect vpn-instance *vpn1* nexthop *10.1.10.2* # traffic behavior *b2* redirect vpn-instance *vpn1* ipv6 nexthop *fc00:10::2* # traffic policy *p1* classifier *c1* behavior *b1* # traffic policy *p2* classifier *c2* behavior *b2* # interface 100GE1/0/1 traffic-policy *p1* inbound traffic-policy *p2* inbound # interface 100GE1/0/2 traffic-policy *p1* inbound traffic-policy *p2* inbound #	acl number *3001* rule *5* permit ip source *10.1.1.0 0.0.0.255* destination *10.1.99.0 0.0.0.255* # acl ipv6 number *3002* rule *5* permit ipv6 source *fc00::1:0/112* destination *fc00::99:0/112* # traffic classifier *c1* if-match acl *3001* # traffic classifier *c2* if-match ipv6 acl *3002* # traffic behavior *b1* redirect vpn-instance *vpn1* nexthop *10.1.10.2* # traffic behavior *b2* redirect vpn-instance *vpn1* ipv6 nexthop *fc00:10::2* # traffic policy *p1* classifier *c1* behavior *b1* # traffic policy *p2* classifier *c2* behavior *b2* # interface 100GE1/0/1 traffic-policy *p1* inbound traffic-policy *p2* inbound # interface 100GE1/0/2 traffic-policy *p1* inbound traffic-policy *p2* inbound #	配置策略路由，匹配业务VRF北向过墙流量，重定向至防火墙。 配置ACL，匹配源、目的地址。 配置流分类，匹配ACL规则。 配置流行为，重定向至防火墙。 配置流策略，匹配流分类及流行为。 在与Spine互联接口上应用流策略。

BorderLeaf-1	BorderLeaf-2	命令说明
acl number *3003* rule *5* permit ip source *10.1.99.0 0.0.0.255* destination *10.1.1.0 0.0.0.255* # acl ipv6 number *3004* rule *5* permit ipv6 source *fc00::99:0/112* destination *fc00::1:0/112* # traffic classifier *c3* if-match acl *3003* # traffic classifier *c4* if-match ipv6 acl *3004* # traffic behavior *b3* redirect vpn-instance *Internet* nexthop *10.1.30.2* # traffic behavior *b4* redirect vpn-instance *Internet* ipv6 nexthop *fc00:30::2* # traffic policy *p3* classifier *c3* behavior *b3* # traffic policy *p4* classifier *c4* behavior *b4* # interface *10GE1/0/9* traffic-policy *p3* inbound traffic-policy *p4* inbound #	acl number *3003* rule *5* permit ip source *10.1.99.0 0.0.0.255* destination *10.1.1.0 0.0.0.255* # acl ipv6 number *3004* rule *5* permit ipv6 source *fc00::99:0/112* destination *fc00::1:0/112* # traffic classifier *c3* if-match acl *3003* # traffic classifier *c4* if-match ipv6 acl *3004* # traffic behavior *b3* redirect vpn-instance *Internet* nexthop *10.1.30.2* # traffic behavior *b4* redirect vpn-instance *Internet* ipv6 nexthop *fc00:30::2* # traffic policy *p3* classifier *c3* behavior *b3* # traffic policy *p4* classifier *c4* behavior *b4* # interface *10GE1/0/9* traffic-policy *p3* inbound traffic-policy *p4* inbound #	配置策略路由，匹配公共出口VRF南向过墙流量，重定向至防火墙。 配置ACL，匹配源、目的地址。 配置流分类，匹配ACL规则。 配置流行为，重定向至防火墙。 配置流策略，匹配流分类及流行为。 在与PE互联接口上应用流策略。

步骤4 配置优化命令

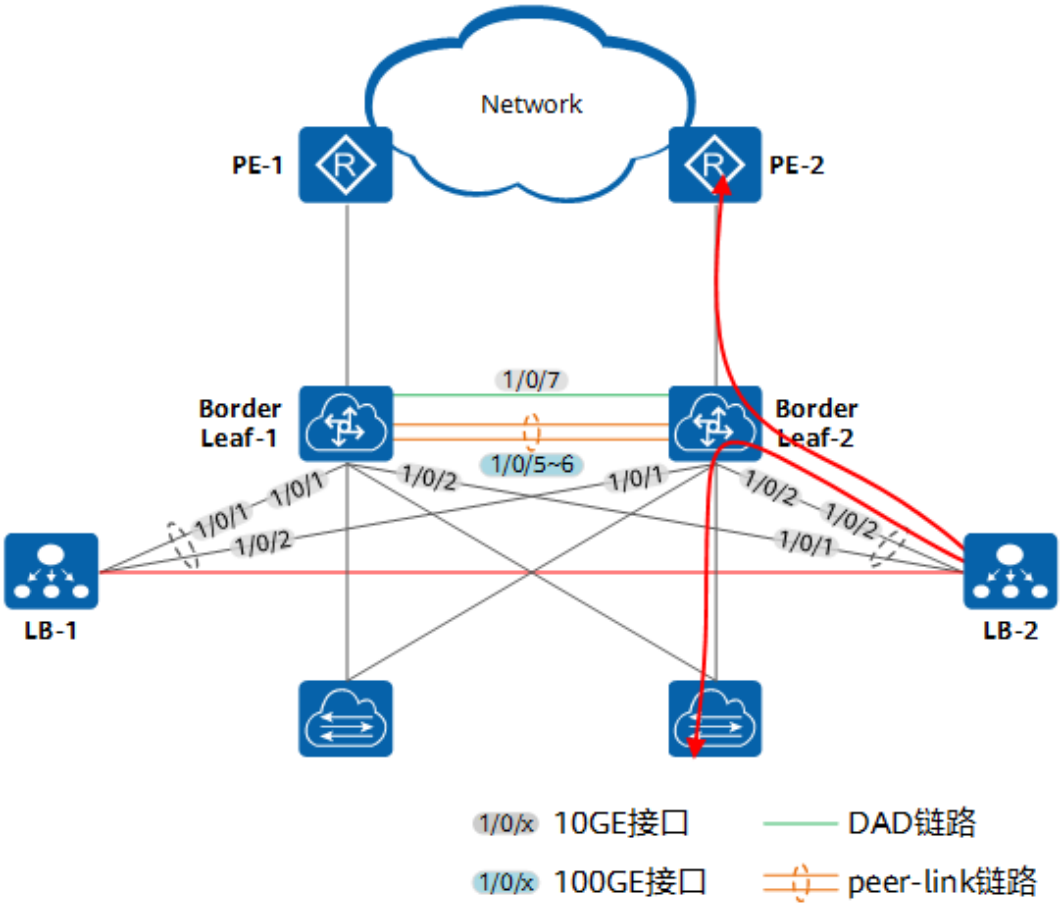
BorderLeaf-1	BorderLeaf-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

2.6 M-LAG 与负载均衡器（LB）对接

2.6.1 组网方案

图 2-7 M-LAG 对接 LB 组网示意图



如图2-7所示，负载均衡器（下文简称LB）旁挂在Border Leaf上，其中：

- Border Leaf部署M-LAG，通过VLANIF与LB对接。
- LB以主备模式部署，旁挂在Border Leaf上。

2.6.2 配置 Border Leaf

说明

本文仅给出交换机部分配置，LB配置请参考对应产品的指导文档。

配置概览

1. **配置M-LAG**
2. **配置Leaf与LB互联链路**
3. **配置路由，将流量引入LB**
4. **配置优化命令**

配置步骤

步骤1 配置M-LAG

BorderLeaf-1	BorderLeaf-2	命令说明
stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	stp mode rstp stp v-stp enable stp tc-protection stp bpdu-protection # 	配置V-STP方式M-LAG。 使能设备对TC类型BPDU报文的保护功能。 使能设备的BPDU保护功能。
ip vpn-instance *DAD* ipv4-family route-distinguisher *1:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.1 30* m-lag unpaired-port reserved # 	ip vpn-instance *DAD* ipv4-family route-distinguisher *2:1* # dfs-group 1 # interface *10GE1/0/7* undo portswitch ip binding vpn-instance *DAD* ip address *192.168.10.2 30* m-lag unpaired-port reserved # 	部署独立直连物理链路，用于M-LAG心跳检测。
dfs-group 1 priority 150 dual-active detection source ip *192.168.10.1* vpn-instance *DAD* peer *192.168.10.2* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	dfs-group 1 priority 100 dual-active detection source ip *192.168.10.2* vpn-instance *DAD* peer *192.168.10.1* authentication-mode hmac-sha256 password *dfs-group@M-LAG* # 	配置DFS Group。 组成M-LAG系统的两台设备的验证口令必须相同。
interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	interface *Eth-Trunk0* trunkport *100GE1/0/5* trunkport *100GE1/0/6* mode lacp-static peer-link 1 port vlan exclude 1 # 	配置M-LAG的peer-link口。 出于高可靠性考虑，peer-link建议多链路部署，多单板/多子卡场景需要跨单板/跨子卡部署。 配置peer-link接口不允许通过VLAN1。

上述配置完成后，可以：

- 通过ping检查两端心跳地址之间是否三层互通。

```
[~BorderLeaf-1] ping -vpn-instance DAD 192.168.10.2
  PING 192.168.10.2: 56  data bytes, press CTRL_C to break
    Reply from 192.168.10.2: bytes=56 Sequence=1 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=2 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=3 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=4 ttl=255 time=1 ms
    Reply from 192.168.10.2: bytes=56 Sequence=5 ttl=255 time=1 ms

  --- 192.168.10.2 ping statistics ---
    5 packet(s) transmitted
    5 packet(s) received
    0.00% packet loss
    round-trip min/avg/max = 1/1/1 ms
```

- 通过**display interface eth-trunk** *eth-trunk-id*或**display eth-trunk** *eth-trunk-id*命令查看peer-link口状态。

```
[~BorderLeaf-1] display interface eth-trunk 0
Eth-Trunk0 current state : UP (ifindex: 8)
```

```
Line protocol current state : UP
Last line protocol up time : 2023-06-30 11:00:17+08:00
Description:
Switch Port, PVID :    1, TPID : 8100(Hex), Hash Arithmetic : based on profile default, Maximal BW :
10Gbps, Current BW : 10Gbps, The Maximum Frame Length is 9216
Internet protocol processing : disabled
IP Sending Frames' Format is PKTFMT_ETHNT_2, Hardware address is 00fd-fdfd-b703
Current system time: 2023-07-13 14:41:01+08:00
Physical is ETH_TRUNK
   Last 10 seconds input rate 697186 bits/sec, 57 packets/sec
   Last 10 seconds output rate 687865 bits/sec, 29 packets/sec
   Input: 62343901 packets,93259180813 bytes
          42820517 unicast,132 broadcast,19523252 multicast
          0 errors,0 drops
   Output:48323130 packets,92501426032 bytes
          29723661 unicast,128 broadcast,18599341 multicast
          0 errors,0 drops
   Last 10 seconds input utility rate:  0.01%
   Last 10 seconds output utility rate: 0.01%
-------------------------------------------------------------
PortName                  Status             Weight
-------------------------------------------------------------
100GE1/0/5                UP                 1
100GE1/0/6                UP                 1
-------------------------------------------------------------
The Number of Ports in Trunk : 2
The Number of Up Ports in Trunk : 2

[~BorderLeaf-1] display eth-trunk 0
Eth-Trunk0's state information is:
(h): high priority
(r): reference port
Local:
LAG ID: 0                     Working Mode: Static
Preempt Delay: Disabled           Hash Arithmetic: based on profile default
System Priority: 32768          System ID: 00fd-fdfd-b703
Least Active-linknumber: 1      Max Active-linknumber: 256
Operating Status: up            Number Of Up Ports In Trunk: 2
Timeout Period: Slow
PortKeyMode: Auto
-------------------------------------------------------------------------------
ActorPortName             Status   PortType PortPri PortNo PortKey PortState Weight
100GE1/0/5(r)             Selected 100GE    32768   6      65      10111100  1
100GE1/0/6(r)             Selected 100GE    32768   6      65      10111100  1

Partner:
-------------------------------------------------------------------------------
ActorPortName             SysPri   SystemID        PortPri PortNo PortKey PortState
100GE1/0/5                32768    00fd-dffb-9a03  32768   6      65      10111100
100GE1/0/6                32768    00fd-dffb-9a03  32768   6      65      10111100
```

- 通过**display dfs-group 1 m-lag**命令查看M-LAG状态。正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
[~BorderLeaf-1] display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
```

```
Dual-active Address : 192.168.10.2
VPN-Instance        : DAD
State               : Backup
Causation           : -
System ID           : 00fd-fdfd-b703
SysName             : ServerLeaf-2
Version             : V300R022C00
Device Type         : CE6800
```

步骤2 配置Leaf与LB互联链路

BorderLeaf-1	BorderLeaf-2	命令说明
vlan *10* # interface vlanif *10* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00:10::1 64* ipv6 nd na glean mac-address *0000-5e00-0110* #	vlan *10* # interface vlanif *10* ip address *10.1.10.1 24* ipv6 enable ipv6 address *fc00:10::1 64* ipv6 nd na glean mac-address *0000-5e00-0110* #	创建与LB互联业务VLAN及VLANIF接口。
interface *Eth-Trunk1* description *to-LB-1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* #	interface *Eth-Trunk1* description *to-LB-1* trunkport *10GE1/0/1* port link-type trunk undo port trunk allow-pass vlan 1 port trunk allow-pass vlan *10* mode lacp-static stp edged-port enable dfs-group 1 m-lag *1* #	配置与LB对接端口。

步骤3 配置路由，将流量引入LB

BorderLeaf-1	BorderLeaf-2	命令说明
ip route-static *10.1.99.1 32 10.1.10.2* ipv6 route-static *fc00:99::1 128 fc00:10::2*	ip route-static *10.1.99.1 32 10.1.10.2* ipv6 route-static *fc00:99::1 128 fc00:10::2*	配置至LB虚拟服务地址的静态路由。

步骤4 配置优化命令

BorderLeaf-1	BorderLeaf-2	命令说明
port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	port-group group-member *10GE1/0/10 to 10GE1/0/48* shutdown #	关闭不使用的端口。
vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	vlan 1 storm suppression multicast cir 64 kbps storm suppression broadcast cir 64 kbps storm suppression unknown-unicast cir 64 kbps #	配置VLAN 1的流量抑制功能，防止广播风暴。

----结束

3 维护与故障处理

3.1 M-LAG 常用 display 命令

- **display dfs-group 1 m-lag**命令用来查看M-LAG设备组状态信息。

  正常情况下，会显示两台成员设备的状态，一台为“Master”，另一台为“Backup”。

```
<HUAWEI> display dfs-group 1 m-lag
*                : Local node
Heart beat state     : OK
Node 1 *
  Dfs-Group ID       : 1
  Priority           : 150
  Dual-active Address : 192.168.10.1
  VPN-Instance       : DAD
  State              : Master
  Causation          : -
  System ID          : 00fd-dffb-9a03
  SysName            : ServerLeaf-1
  Version            : V300R022C00
  Device Type        : CE6800
Node 2
  Dfs-Group ID       : 1
  Priority           : 100
  Dual-active Address : 192.168.10.2
  VPN-Instance       : DAD
  State              : Backup
  Causation          : -
  System ID          : 00fd-fdfd-b703
  SysName            : ServerLeaf-2
  Version            : V300R022C00
  Device Type        : CE6800
```

- **display dfs-group 1 peer-link**命令用来查看peer-link链路状态信息。

```
<HUAWEI> display dfs-group 1 peer-link
Peer-link information
------------------------------------------------------------------
Total Interface(s): 1
Peer-link Id     : 1
Port Name        : Eth-Trunk0
Port State       : Up
Link Type        : Physical link
------------------------------------------------------------------
```

- **display dfs-group 1 heartbeat**命令用来查看DAD心跳链路的状态信息。

```
<HUAWEI> display dfs-group 1 heartbeat
------------------------------------------------------------------------------
Heart beat status    : OK
Local:
```



```
  Dfs-Group ID       : 1
  Priority           : 150
  Udp port           : 1025
  Dual-active Address : 10.159.31.15
  VPN-Instance       : public net
  System ID          : 00fd-dffb-9a03
  Heart beat state   : --
Peer:
  Dfs-Group ID       : --
  Priority           : --
  Udp port           : --
  Dual-active Address : --
  VPN-Instance       : public net
  System ID          : --
  Heart beat state   : --
--------------------------------------------------------------------------------
```

- **display dfs-group 1 node** *node-id* **m-lag** [**brief**]命令用来查看M-LAG链路聚合组的状态。

```
<HUAWEI> display dfs-group 1 node 1 m-lag
* - Local node

M-Lag ID         : 1
Interface        : Eth-Trunk 1
Port State       : Up
Status           : active(*)-active
Member Port Role  : Master(*)-Backup

M-Lag ID         : 2
Interface        : Eth-Trunk 2
Port State       : Up
Status           : active(*)-active
Member Port Role  : Master(*)-Backup
```

- **display dfs-group consistency-check status**命令用来查看M-LAG配置一致性检查功能是否开启。

```
<HUAWEI> display dfs-group consistency-check status
--------------------------------------------------------------------------------
Local Status : Enable
Peer Status  : Enable
--------------------------------------------------------------------------------
Configuration          Send times       Failed times
--------------------------------------------------------------------------------
BD                       0                0
BDIF                     0                0
VLAN                     2                0
Port VLAN                3                0
VLAN instance            1                0
STP port                 3                0
LACP                     2                0
VLANIF                   1                0
MLAG member number       3                0
MAC aging                1                0
ARP aging                1                0
Static MAC               1                0
Static ARP               1                0
MLAG IP                  41               0
MLAG mode                0                0
V-STP Enable             1                0
LACP M-LAG System-Id     1                0
LACP M-LAG Priority      1                0
STP Edged-port           3                0
LACP Mode                2                0
Peer-link STP            2                0
Exclude VLAN             5                0
Election mode            1                0
STP Vlan                 0                0
--------------------------------------------------------------------------------
```

- **display dfs-group consistency-check** { **global** | **interface** { **m-lag** *m-lag-id* | **peer-link** *peer-linkid* } | **static-arp** | **static-mac** }命令用来查看M-LAG两台设备的相关配置是否一致性。

  如果两端相关配置不一致，则会显示不一致的内容：

  ```
  <HUAWEI> display dfs-group consistency-check global
  ------------------------------------------------------------------------------
  Configuration          Type   Local value              Peer value
  ------------------------------------------------------------------------------
  VLANIF(IPV4)              2   100(40.40.5.1:30)        100(40.40.5.2:30)
  ------------------------------------------------------------------------------
  ```

  如果两端相关配置一致，则会显示配置一致性检查通过：

  ```
  <HUAWEI> display dfs-group consistency-check interface m-lag 1
  Info: The device consistency check is passed.
  ```

- **display m-lag consistency-check whitelist status**命令用来查看M-LAG配置一致性检查的白名单。

  ```
  <HUAWEI> display m-lag consistency-check whitelist status
  ------------------------------------------------------------------------------
  module          service-type                  whitelist
  ------------------------------------------------------------------------------
  m-lag             m-lag-member-num                N
                  m-lag-id                      N
                  peer-link-exclude-vlan          N
                  m-lag-ipv4-address              N
                  m-lag-ipv6-address              N
                  m-lag-election-mode             N
  ------------------------------------------------------------------------------
  vlan              vlan-configuration              N
                  port-vlan-relation              N
  ------------------------------------------------------------------------------
  stp               stp-m-lag-priority              N
  ------------------------------------------------------------------------------
  mac               mac-aging-time                  N
                  static-mac                    N
  ------------------------------------------------------------------------------
  arp               arp-aging-time                  N
                  static-arp                    N
  ------------------------------------------------------------------------------
  vlanif            vlanif-configuration            N
                  ipv4-address                  N
                  ipv6-address                  N
                  vrrp4                         N
                  virtual-mac                   N
                  vlanif-status                 N
                  vlanif-bypass                 N
                  arp-timeout                   N
  ------------------------------------------------------------------------------
  vxlan             bd-configuration                N
                  vbdif-configuration             N
                  ipv4-address                  N
                  ipv6-address                  N
                  virtual-mac                   N
                  vbdif-status                  N
  ------------------------------------------------------------------------------
  ```

- **display m-lag troubleshooting** { **current** | **history** }命令用来查看M-LAG发生故障的原因。

  ```
  <HUAWEI> display m-lag troubleshooting current
  Total : 2
  ------------------------------------------------------------------------------
  Seq  Time                  Event Description
  ------------------------------------------------------------------------------
  1    2021-01-21 16:04:48      DFS pairing failed because the DFS group could not
                                receive Hello packets. Check the DFS configuratio
                                n of the local or remote switch.
  2    2021-01-21 16:04:37      DFS pairing failed because the peer-link was down.
  ```



```
                    Check the status of the peer-link interface.
------------------------------------------------------------------------------
```

- **display m-lag unpaired-port reserved**命令用来查看备设备上接口在peer-link故障但双主检测正常时不被Error-Down的接口。

```
<HUAWEI> display m-lag unpaired-port reserved
*    : errordown
Ports
------------------------------------------------------------------------------
100GE1/0/1      100GE1/0/2
```

- **display stp** [**brief**]命令用来查看STP状态信息。

```
<HUAWEI> display stp brief
 MSTID  Port                Role  STP State     Protection      Cost    Edged
    0   Eth-Trunk0          ROOT  forwarding    none            2000    disable
    0   Eth-Trunk1          DESI  forwarding    none            2000    enable
    0   Eth-Trunk2          DESI  forwarding    none            2000    enable
```

- **display mac-address** [**vlan** *vlan-id* | **bridge-domain** *bridge-domain-id*]命令用来查看设备学习到的MAC地址信息。

```
<HUAWEI> display mac-address
Flags: * - Backup
       # - forwarding logical interface, operations cannot be performed based
           on the interface.
BD   : bridge-domain    Age : dynamic MAC learned time in seconds
------------------------------------------------------------------------------
MAC Address    VLAN/VSI/BD   Learned-From        Type                Age
------------------------------------------------------------------------------
9c7d-a378-3c8d 11/-/-        Eth-Trunk13         dynamic             1058749
cc64-a668-6814 11/-/-        Eth-Trunk1          dynamic             1058745
00fd-fdfd-b706 100/-/-       Eth-Trunk0          dynamic             91258
------------------------------------------------------------------------------
```

- **display arp** [**interface** *interface-type interface-name*]命令用来查看设备学习到的ARP信息。

```
<HUAWEI> display arp
ARP Entry Types: D - Dynamic, S - Static, I - Interface, O - OpenFlow, RD - Redirect
EXP: Expire-time VLAN: VLAN or Bridge Domain

IP ADDRESS      MAC ADDRESS    EXP(M) TYPE/VLAN       INTERFACE        VPN-INSTANCE
------------------------------------------------------------------------------------------
10.159.31.254   0000-5e00-0222       I              Vlanif11
10.159.31.52    9c7d-a378-3c8d    8  D/11           Eth-Trunk13
10.40.5.1       00fd-dffb-9a06       I              Vlanif100
10.40.5.2       00fd-fdfd-b706   20  D/100          Eth-Trunk0
10.40.41.1      00fd-dffb-9a0a       I              Vlanif1000
10.40.41.2      00fd-fdfd-b70a   20  D/1000         Eth-Trunk0
10.255.9.2      00fd-dffb-9a04       I              25GE1/0/4
10.255.9.1      487b-6bee-0712   11  D              25GE1/0/4
------------------------------------------------------------------------------------------
```

3.2 M-LAG 组建失败故障处理

故障现象

两台设备组建M-LAG失败，执行**display dfs-group 1 m-lag**命令只显示一台设备的状态。

故障处理步骤

步骤1 用户可以使用**display dfs-group 1 m-lag**查看M-LAG组建失败的原因。

```
<HUAWEI> display dfs-group 1 m-lag
*                  : Local node        //标识为本地节点。
Heart beat state   : OK                //心跳状态。OK：表示正常在线状态；Lost：表示离线状态。
```



```
Node 1 *
  Dfs-Group ID       : 1              //DFS Group的编号。
  Priority           : 150            //DFS Group的优先级。
  Dual-active Address : 10.1.10.15    //DFS Group绑定的IP地址。
  VPN-Instance       : public net     //DFS Group绑定的VPN实例。
  State              : Master         //设备的主备状态。Master：主用状态；Backup：备用状态。
  Causation          : NOPEERLINK     //M-LAG组建失败的原因。
  System ID          : 00fd-dffb-9a03 //系统MAC地址。
  SysName            : ServerLeaf-1   //设备名称。
  Version            : V300R022C00    //设备当前的版本。
  Device Type        : CE6800         //设备对应的系列。
```

根据显示信息中**Causation**字段可以查看到M-LAG组建失败的原因，详细原因参见下表。

表 3-1 Causation 字段具体内容解释

Causation字段	失败原因	解决办法
NOPEERLINK	表示没有配置peer-link。	执行**display current-configuration | section include peer-link**命令，查看是否有Eth-trunk接口下配置了**peer-link 1**，如果没有，则需要在规划的Eth-trunk接口下配置该命令。
NOADDRESS	表示DFS Group没有绑定地址或绑定的IP地址不能互通。	进入DFS Group视图，执行**display this**查看是否有**dual-active detection source**命令。如果没有，则增加该配置；如果有，则检查两端配置的IP地址之间能否ping通，需要确保两端配置的IP地址能够互通。
SAMEMAC	表示本端和对端设备的系统MAC地址相同。	执行**display system mac-address**命令查看两台设备的系统MAC地址是否一致，如果是，则在诊断视图下执行**reset system-mac**命令刷新系统MAC。
TYPEMISMATCH	表示绑定到DFS Group的源地址类型不同。	进入DFS Group视图，执行**display this**查看**dual-active detection source**配置的IP地址类型是否一致，两端需要同时使用IPv4，或者同时使用IPv6。
TIMEOUT	表示接收协议报文超时，即没有收到协议报文。可能因为链路拥塞、CPU使用率高导致协议报文丢弃等。	请联系华为技术工程师进行处理。
PEERLINKDOWN	表示peer-link链路状态为Down。	检查被配置为peer-link的Eth-trunk接口配置及其物理成员接口状态，详细可参考**Eth-trunk故障定位指导**。

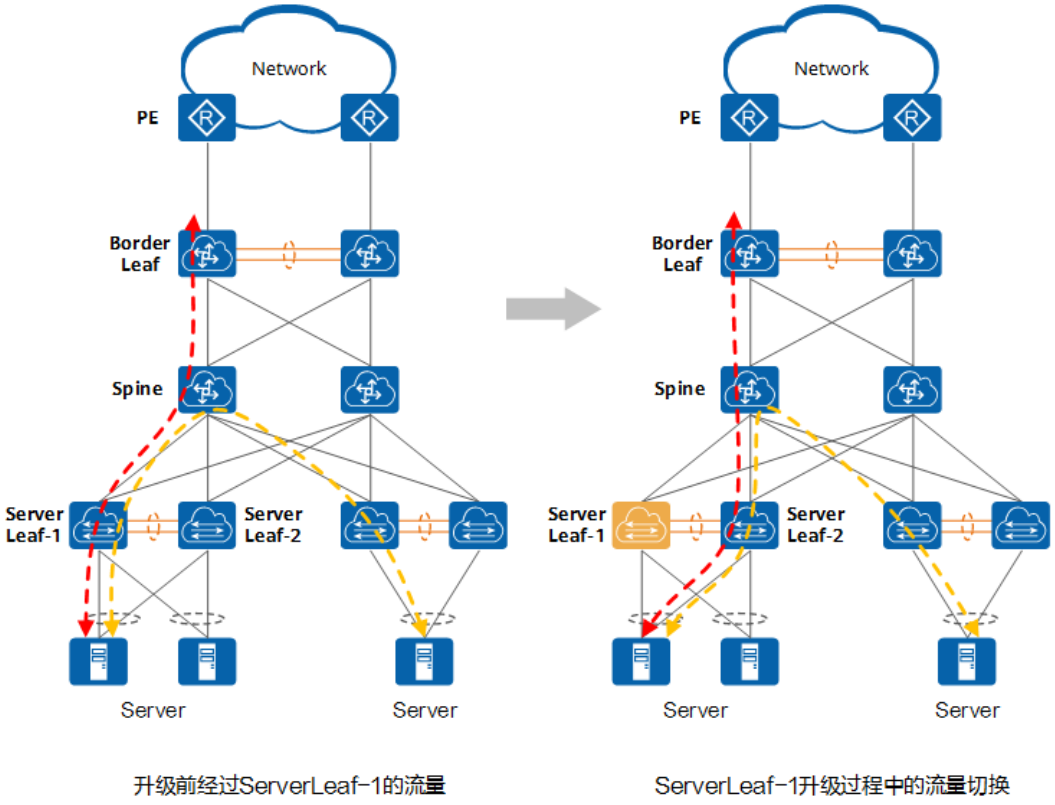
Causation字段	失败原因	解决办法
DETECT	表示能收到对端hello报文但收不到对端设备信息报文。可能因为链路拥塞、CPU使用率高导致协议报文丢弃等。	请联系华为技术工程师进行处理。
AUTHENTICATION FAILED	表示DFS Group配对的设备验证失败。	配置**authentication-mode hmac-sha256 password** *password*命令，确保两端配置的验证密码一致。
VERSIONMISMATCH	表示设备版本不匹配。	执行**display version**查看设备版本，确保两台设备版本一致。
NOAUTHENTICATION	表示未配置验证密码。	进入DFS Group视图，配置**authentication-mode hmac-sha256 password** *password*命令。
DEVICETYPEMISMATCH	表示设备类型不匹配。	执行**display device**查看设备类型，确保两台设备类型一致。
PEERLINKTYPEMISMATCH	表示peer-link类型不匹配。	检查peer-link配置，确保两台设备同时使用物理peer-link，或者同时使用virtual peer-link。

----结束

3.3 M-LAG 升级（维护模式）

组网需求

图 3-1 M-LAG 维护模式升级示意图



如图3-1所示，某网络采用Spine-Leaf架构，Server Leaf部署M-LAG，作为服务器的三层网关，服务器接入采用负载分担方式接入。Server Leaf、Spine、Border Leaf之间三层互联，运行路由协议保证三层可达。

为实现升级期间不影响部署在设备上的业务，可使用M-LAG在维护模式下的升级方式，提升升级过程中的可靠性，减少业务中断。

升级思路

采用如下思路进行M-LAG维护模式下升级，以组成M-LAG的ServerLeaf-1和ServerLeaf-2为例：

1. 进行网络基础配置，包括M-LAG、服务器接入、路由协议等，实现网络二三层互通。
2. 进行升级前检查。
3. 申请并加载License，M-LAG维护模式下的升级功能使用License控制。
4. ServerLeaf-1进入维护模式，进行调整路由Cost、调整路由发布优先级和设置接口Down等配置，将ServerLeaf-1的流量切换至ServerLeaf-2，然后进行ServerLeaf-1的升级操作。

5. ServerLeaf-1升级完成后，再次进入维护模式，进行恢复路由Cost、发布优先级、接口Up等配置，将流量恢复至ServerLeaf-1进行转发。
6. 重复ServerLeaf-1的流程，对ServerLeaf-2进行升级操作。

升级注意事项

- 如果M-LAG系统的备设备的MAC地址小于主设备，备设备升级重启后可能会触发STP根桥切换，收敛时间长，会导致重启丢包10s左右。所以升级前，请先在备设备上手工配置和主设备相同的STP根桥MAC地址。操作步骤如下：

  a. 在主设备上查询STP根桥MAC地址。

  ```
  <HUAWEI> display stp v-stp
  Bridge Information:
    V-STP Mode          :True
    Bridge Mac          :Config=00e0-fc11-1200 / Active=00e0-fc12-1234
    Peer-link Name      :Eth-Trunk1
  ```

  Active的取值“00e0-fc12-1234”即为STP根桥MAC地址。

  b. 在备设备上配置STP根桥MAC地址，与主设备一致。

  ```
  <HUAWEI> system-view
  [~HUAWEI] stp bridge-address 00e0-fc12-1234
  [*HUAWEI] commit
  ```

- 建议先升级备设备，备设备升级完成后，需要检查状态正常后，再升级主设备，同时主备设备升级间隔至少10分钟。

操作步骤

步骤1 进行网络基础配置，实现网络二三层互通。

M-LAG、服务器接入、路由协议等基础配置本案例不再介绍，详细描述可以参考前文**2 典型配置案例**。

步骤2 进行升级前检查。

检查网络运行状态及其可靠性，例如物理链路状态是否正常，物理链路是否存在备份，路由协议是否双平面备份等。详细步骤请参考升级指导书

步骤3 申请并加载License。

M-LAG维护模式下的升级功能使用License控制，缺省情况下，新购买的设备的M-LAG维护模式下的升级功能未打开。如果需要使用设备M-LAG维护模式下的升级功能，请联系设备经销商申请并购买License。申请及加载License的详细步骤请参考**License使用指南**。

可以通过**display license**命令查看设备是否已加载M-LAG维护模式下的升级功能License，其控制项名称为**CE-LIC-LU**。

```
<ServerLeaf-1> display license
......

Item name             Item type      Value      Description
-----------------------------------------------------------------
......

CE68-LIC-LU             --           1        CloudEngine 6800 lossless upgrade Function
  CE-LIC-LU           Function       YES        CE-LIC-LU

......
```

步骤4 升级ServerLeaf-1。

1. 进入维护模式，进行调整路由Cost、调整路由发布优先级和设置接口Down等配置，将ServerLeaf-1的流量切换至ServerLeaf-2。

   调整路由OSPF或OSPFv3的COST值为最大值，或降低路由BGP或BGP4+的路由发布优先级（用户可根据现网中使用的路由协议选择）。

   ```
   [~ServerLeaf-1] maintenance    //进入维护模式
   [~ServerLeaf-1-maintenance] ospf advertise max-cost    //将设备全局发布的OSPF LSA的COST值调整为最大值
   [*ServerLeaf-1-maintenance] ospfv3 advertise max-cost    //将设备全局发布的OSPFv3 LSA的COST值调整为最大值
   [*ServerLeaf-1-maintenance] advertise bgp ipv4-family unicast lowest-priority enable    //降低BGP IPv4单播地址族、VPN实例IPv4地址族路由发布优先级
   [*ServerLeaf-1-maintenance] advertise bgp ipv6-family unicast lowest-priority enable    //降低BGP IPv6单播地址族、VPN实例IPv6地址族路由发布优先级
   [*ServerLeaf-1-maintenance] advertise bgp l2vpn-family evpn lowest-priority enable    //降低EVPN地址族路由发布优先级
   [*ServerLeaf-1-maintenance] commit
   ```

   调整路由Cost、发布优先级后，先执行**display ip routing-table** [**vpn-instance** *vpn-name*]等命令检查路由是否收敛，及路由侧流量是否切换完成。待流量切换完成后，再设置加入M-LAG的Eth-Trunk成员口Down，引导服务器将上行流量切至备用链路。

   ```
   [~ServerLeaf-1-maintenance] lacp force-down    //设置加入M-LAG的Eth-Trunk成员口为Down，该配置对主备模式下的M-LAG接口不生效
   [*ServerLeaf-1-maintenance] commit
   ```

   执行完上述配置后，可以执行**display interface brief**等命令，查看流量是否成功切换至ServerLeaf-2。

2. 升级ServerLeaf-1软件版本，详细升级注意事项及操作步骤请参考升级指导书。

步骤5 待ServerLeaf-1升级完成后，再次进入维护模式，恢复路由Cost、路由发布优先级、接口Up等设置，将流量切换回ServerLeaf-1。

```
[~ServerLeaf-1] maintenance    //进入维护模式
[~ServerLeaf-1-maintenance] undo lacp force-down    //恢复加入M-LAG的Eth-Trunk成员口为Up
[*ServerLeaf-1-maintenance] commit
```

恢复接口Up后，先检查M-LAG成员口同步的ARP、ND和MAC表项是否恢复，待表项恢复完成后再执行路由流量的回切。

```
[~ServerLeaf-1-maintenance] undo ospf advertise max-cost    //恢复OSPF LSA的COST值
[*ServerLeaf-1-maintenance] undo ospfv3 advertise max-cost    //恢复OSPFv3 LSA的COST值
[*ServerLeaf-1-maintenance] undo advertise bgp ipv4-family unicast lowest-priority enable    //恢复BGP IPv4单播地址族、VPN实例IPv4地址族路由发布优先级
[*ServerLeaf-1-maintenance] undo advertise bgp ipv6-family unicast lowest-priority enable    //恢复BGP IPv6单播地址族、VPN实例IPv6地址族路由发布优先级
[*ServerLeaf-1-maintenance] undo advertise bgp l2vpn-family evpn lowest-priority enable    //恢复EVPN地址族路由发布优先级
[*ServerLeaf-1-maintenance] commit
```

执行完上述配置后，可以执行**display ip routing-table** [**vpn-instance** *vpn-name*]、**display interface brief**等命令，查看路由是否收敛，及流量是否成功切换回ServerLeaf-1。

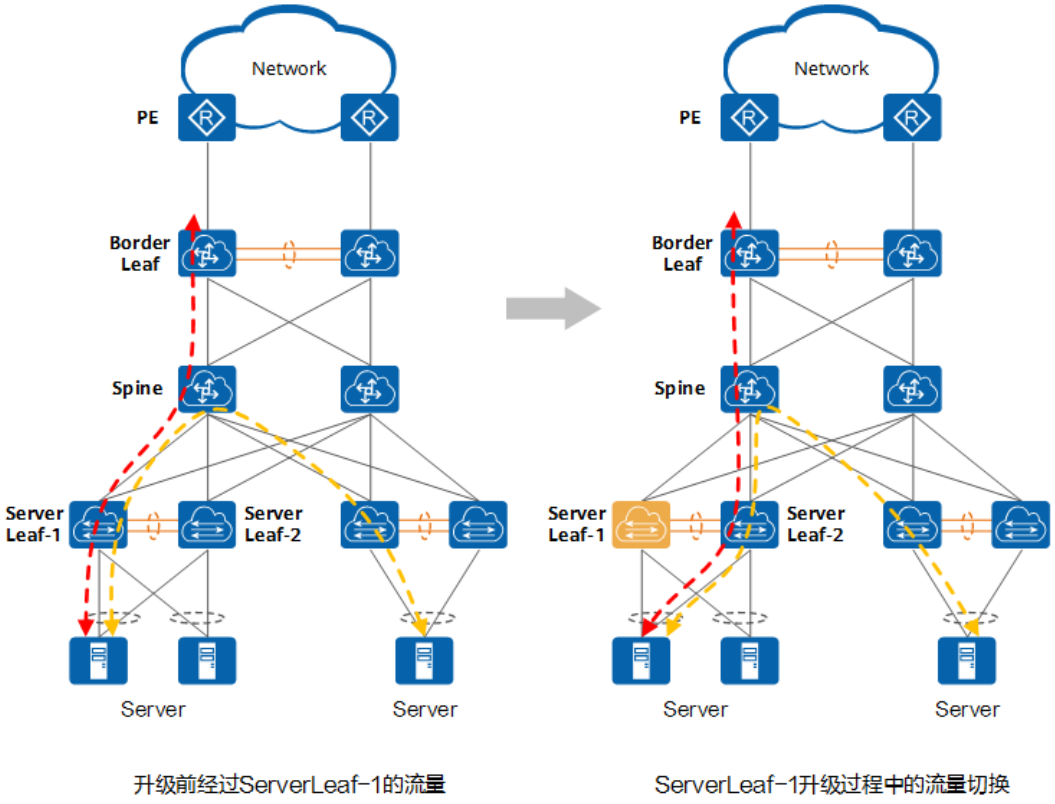
步骤6 参照上述流程升级ServerLeaf-2，最终完成整个M-LAG设备组的升级。

----结束

3.4 M-LAG 升级（非维护模式）

组网需求

图 3-2 M-LAG 非维护模式升级示意图



如图3-2所示，某网络采用Spine-Leaf架构，Server Leaf部署M-LAG，作为服务器的三层网关，服务器接入采用负载分担方式接入。Server Leaf、Spine、Border Leaf之间三层互联，运行路由协议保证三层可达。

升级思路

采用如下思路进行M-LAG非维护模式下升级，以组成M-LAG的ServerLeaf-1和ServerLeaf-2为例：

1. 进行网络基础配置，包括M-LAG、服务器接入、路由协议等，实现网络二三层互通。
2. 进行升级前检查。
3. 升级ServerLeaf-1。
4. ServerLeaf-1升级完成后，再对ServerLeaf-2进行升级操作。

升级注意事项

- 如果M-LAG系统的备设备的MAC地址小于主设备，备设备升级重启后可能会触发STP根桥切换，收敛时间长，会导致重启丢包10s左右。所以升级前，请先在备设备上手工配置和主设备相同的STP根桥MAC地址。操作步骤如下：

a. 在主设备上查询STP根桥MAC地址。

```
<HUAWEI> display stp v-stp
Bridge Information:
  V-STP Mode          :True
  Bridge Mac          :Config=00e0-fc11-1200 / Active=00e0-fc12-1234
  Peer-link Name      :Eth-Trunk1
```

Active的取值“00e0-fc12-1234”即为STP根桥MAC地址。

b. 在备设备上配置STP根桥MAC地址，与主设备一致。

```
<HUAWEI> system-view
[~HUAWEI] stp bridge-address 00e0-fc12-1234
[*HUAWEI] commit
```

- 建议先升级备设备，备设备升级完成后，需要检查状态正常后，再升级主设备，同时主备设备升级间隔至少10分钟。

操作步骤

步骤1 进行网络基础配置，实现网络二三层互通。

M-LAG、服务器接入、路由协议等基础配置本案例不再介绍，详细描述可以参考前文**2 典型配置案例**。

步骤2 进行升级前检查。

检查网络运行状态及其可靠性，例如物理链路状态是否正常，物理链路是否存在备份，路由协议是否双平面备份等。详细步骤请参考升级指导书

步骤3 升级ServerLeaf-1。

保存配置并执行reboot命令升级设备。详细操作步骤请参考升级指导书，下文仅为简单示例。

```
<ServerLeaf-1> startup system-software XXX.cc all   //设置下次启动使用的系统软件
<ServerLeaf-1> reboot
Warning: The current configuration will be saved to the next startup saved-configuration file. Continue?
[Y/N]: y    //如果使用新配置文件，则输入n，否则输入y
Warning:The system will reboot. Continue? [Y/N]: y
```

待ServerLeaf-1升级完成后，检查设备运行状态，及路由、流量是否恢复。

步骤4 参照**步骤3**升级ServerLeaf-2，最终完成整个M-LAG设备组的升级。

----结束