MC-LAG 技术白皮书

文档版本 V1.0

发布日期 2022-12-16

商标说明

Inspur 浪潮、Inspur、浪潮、Inspur NOS 是浪潮集团有限公司的注册商标。

本手册中提及的其他所有商标或注册商标，由各自的所有人拥有。

技术支持

技术服务电话：400-860-0011

地　　址：中国济南市浪潮路 1036 号

浪潮电子信息产业股份有限公司

邮　　箱：lckf@inspur.com

邮　　编：250101

前言

变更记录

版本	时间	变更内容
V1.0	2022-12-16	首版发布

目 录

1 概述

1.1 背景

无论是传统的企业网，还是方兴未艾的数据中心，其组网拓扑都面临一个共同的需求：高可靠性。换而言之，存在一个共同的需要解决的问题：潜在的网络单点故障。

为了防止出现单点故障，业界经常采用的网络拓扑冗余方式有以下几种：

- 链路捆绑

在两台交换机之间连接多条链路，将多条链路捆绑成一个 portchannel，交换机之间运行 LACP 进行协商，或者直接进行静态捆绑。这种方法能够防止单条链路故障，当某条链路故障时，流量可以从其它链路进行转发。链路捆绑无法消除因对端交换机整机故障而导致的流量转发错误。

- 堆叠

堆叠是将多台交换机设备组合在一起，从逻辑上组合成一台交换机。堆叠支持成员交换机之间的冗余备份，支持跨设备的链路聚合，实现了跨设备的链路备份。但是堆叠的缺点也比较明显，即其控制平面完全耦合，控制面故障可能导致整台交换机故障，另外，堆叠升级版本时，需要所有成员同时升级，可能导致业务长时间中断。

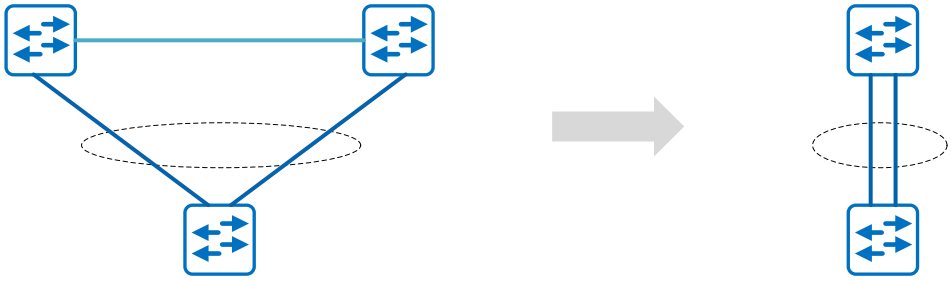
- 双上行

双上行即上连两台交换机，当其中一台交换机发生故障，业务可以切换到另外一台交换机。设备冗余会带来潜在的网络环路风险。为了防止可能出现的环路，需要配置 xSTP 等类似的防环协议，而 xSTP 的配置往往非常繁琐。

1.2 定义

MC-LAG（MultiChassis Link Aggregation Group，跨设备链路聚合组），是一种跨设备链路聚合的机制，如图 1-1 所示，将两台交换机与被接入设备进行链路聚合协商，从被接入设备来看，对端的两台交换机被虚拟成了一台交换机，从而把可靠性从链路级别提高到了设备级别。

图 1-1MC-LAG 示意图



1.3 优点

MC-LAG 作为一种跨设备进行链路聚合的技术，具有如下优点：

- 更大的带宽

通过将链路进行捆绑，可以将带宽成倍增加。

- 更高的可靠性

把可靠性从链路级别提高到了设备级别，一台设备出故障，另外一台设备可以正常工作。

- 组网及配置简单

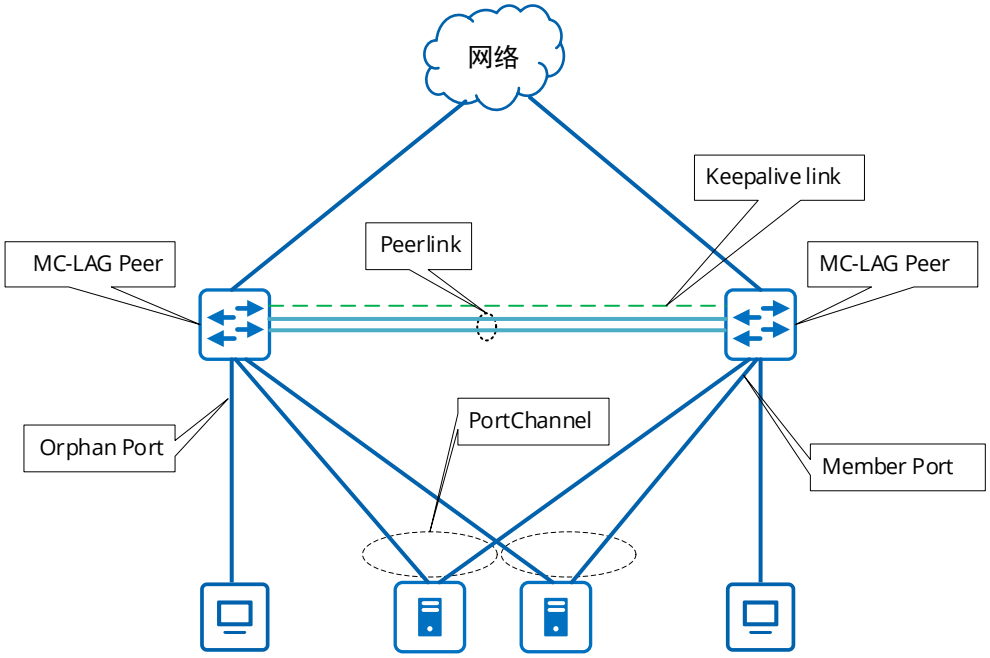
MC-LAG 类似于一种虚拟化技术，将双归设备虚拟成了一台设备，在防止环路的同时实现了设备冗余备份，不再需要进行繁琐的生成树协议配置，极大地简化了组网及配置。

- 独立升级

在对一台设备进行升级时，另外一台设备可以正常工作，对业务基本没有影响。

2 缩写和术语

图 2-1MC-LAG 典型组网



如图 2-1 所示：

缩写和术语	解释
MC-LAG	MultiChassis Link Aggregation Group，跨设备链路聚合组
ICCP	Inter-Chassis Communication Protocol，机架间通信协议，对应RFC7275
MC-LAG peer	MC-LAG交换机，两台交换机互为邻居关系，目前只支持两台交换机建立邻居关系
Member port	MC-LAG成员端口，跨设备形成MC-LAG，必须是Portchannel端口
Peerlink	两个peer之间作为数据备份链路的连线，可以是物理端口或者portchannel；在二层流量转发的场景，peerlink必须配置，当MC-LAG下联的member port down掉的时候，相关流量改由peerlink转发；在三层流量转发的场景，流量如何转发是通过查询路由决定的，peerlink是可选配置
Keepalive link	用于建立peer邻居关系的链路，一般是配置了ipv4地址的三层链路；当邻居关系建立以后，该链路用于同步MC-LAG状态、周期发送心跳报文。Keepalive link可以和peerlink合二为一，也可以不同；

	Keepalive link无需配置，只要对端peer用于建立邻居关系的ipv4地址是路由可达的，即可认为keepalive link是UP的
Orphan port	非MC-LAG member port，即本端口在对端设备上没有配置对应的跨设备聚合端口
BUM	Broadcast、Unknown unicast、Multicast，广播、未知单播、组播报文
FDB	Forwarding Data Base，二层MAC地址转发表

3 技术介绍

3.1 技术概述

MC-LAG 在完成配置以后，会通过比较 ipv4 地址大小确定主备关系，地址大的为备设备，小的为主设备。主设备会主动使用 TCP 与备设备建立邻居关系。在两台设备建立起邻居关系以后，备设备会把本设备上所有 MC-LAG 成员端口的 MAC 地址修改为与主设备的 system MAC 相同，这样在进行 LACP 协商时，主、备设备发送的 LACP PDU 中包含相同的 system id，使对端设备感觉是在与同一台设备进行协商，从而达到跨设备进行链路聚合的目的。邻居关系建好以后，设备之间会进行 FDB MAC、ARP 等信息的同步。通过控制 FDB、ARP 表项下发的端口，可以控制单播报文的转发通道，即在正常情况下，peerlink 只作为备份链路不转发流量，而当出现故障时，流量可以从 peerlink 转发，从而实现流量不会因为故障而中断。对于 BUM 报文，MC-LAG 设计了单向隔离机制，以保证同一台设备不会收到 BUM 报文的多份拷贝。

3.2 主备选举

MC-LAG 互为 peer 的两台交换机，其用于建立邻居关系的本端 ipv4 地址（local ip）、对端 ipv4 地址（peer ip）都是手工配置指定的。系统对配置的 ipv4 地址进行比较，地址大的为 standby，地址小的为 active。所谓的 active、standby，都是控制层面的概念，跟数据转发无关（参考“ 邻居关系建立”一节）。在数据平面，互为 peer 关系的两台交换机独自决定自身的数据转发路径，跟主备角色无关。

3.3 邻居关系建立

RFC 7275 定义了 ICCP 协议，该协议用于 MC-LAG 控制平面，即建立邻居关系、及时同步信息等。

ICCP 使用 TCP 建立连接关系，TCP 端口号为 8888。Active 交换机为 TCP client，standby 为 TCP server，client 主动向 server 发起 TCP 连接请求。

TCP 连接建立起来以后，两台交换机会每隔 1s 钟向对方发送 heartbeat 心跳报文。如果超过 15s 未收到对方发送过来的心跳报文，则认为邻居关系断开。

3.4 信息同步

MC-LAG 建立起邻居关系且工作正常以后，两台设备之间会通过 keepalive link 发送消息实时同步对端的信息，包括所有 FDB 表项、MC-LAG 成员端口学习到的 ARP 以及 ND 表项，并且发送 MC-LAG 成员端口状态变化信息，这样任意一台设备故障都不会影响流量转发，从而保证业务不会中断。

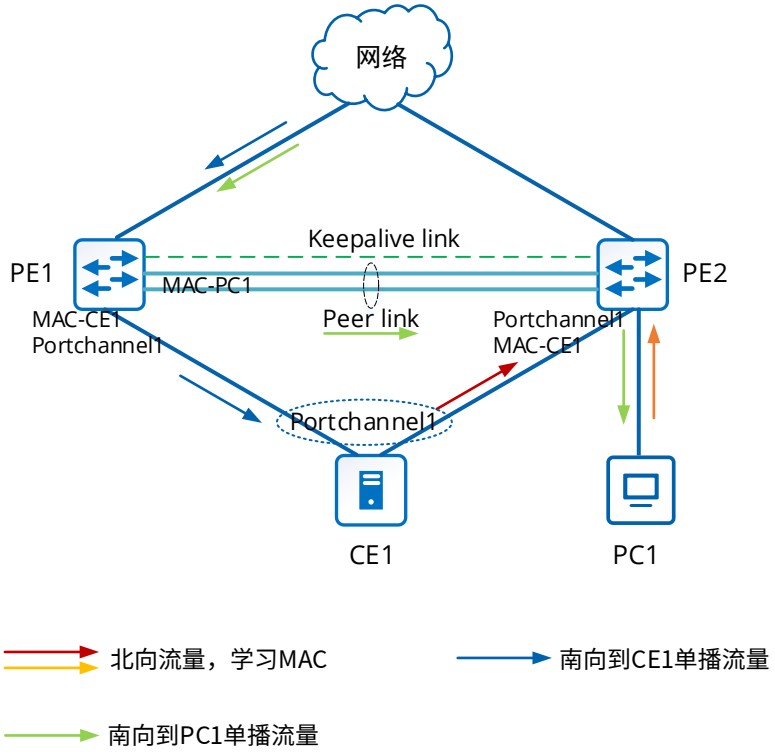
在邻居关系建立时，两台设备之间还会给对方同步自身的系统 MAC 地址信息，也即 LACP 协议所定义的 system ID。当 standby 从 active 收到系统 MAC 地址信息，会把自己的 system ID 修改为对端的系统 MAC。

3.5 MAC 学习及同步

默认在 MC-LAG 交换机上会关闭 peerlink 相关端口的 MAC 学习功能。这样在二层转发的场景中，正常情况下，单播流量不会查找到指向 peerlink 的 FDB MAC 表项，从而不会经 peerlink 转发。

MC-LAG 设备学习到的 FDB MAC 表项可以分成两类：MC-LAG 成员端口学习到的 MAC 和 orphan port 学习到的 MAC。MC-LAG 会把所有学习到的 MAC 表项同步给对端邻居，邻居收到同步的 MAC 表项以后，会根据表项类别进行不同的处理：对于 MC-LAG 成员端口学习到的 MAC，会下发到本端相同 name 的成员端口上；而 orphan port 学习到的 MAC，则会下发到 peerlink 端口上。

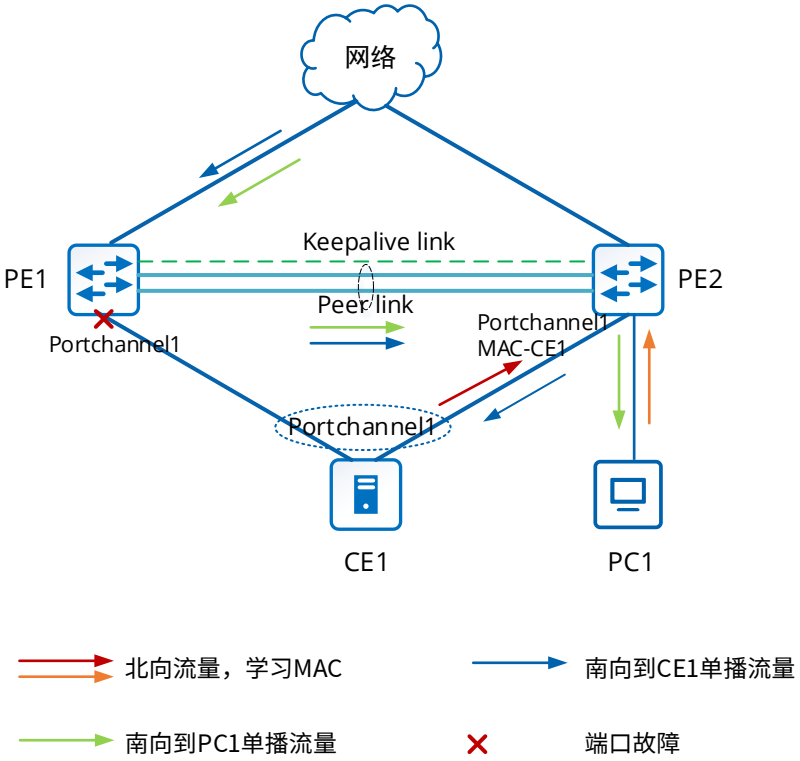
图 3-1MAC 学习和同步



如图 3-1 所示，CE1、PC1 发送流量到 PE2，PE2 从 MC-LAG 成员端口 portchannel1 学习到 MAC-CE1、从 orphan 端口学习到 MAC-PC1，并且将学习到的 MAC 信息同步给 PE1。PE1 收到同步信息，会将 MAC-CE1 下发到成员端口 portchannel1 上，将 MAC-PC1 下发到 peerlink 端口。

PE1 收到目的为 CE1 的南向单播流量，会直接从本端的成员端口转发，而收到目的为 PC1 的南向单播流量，则会从 peerlink 转发。

图 3-2 端口故障 MAC 地址重定向



如图 3-2 所示，如果 PE1 的成员端口 portchannel1 状态迁移到 down，则 MAC-CE1 会重定向下发到 peerlink 上。此时如果 PE1 收到目的为 CE1、PC1 的南向单播流量，都会从 peerlink 转发。

3.6 防环机制

MC-LAG 将两台设备虚拟成一台设备，构造了一个无环网络。

对于单播流量，正常情况下，通过关闭 peerlink 端口 MAC 地址学习功能，可以控制二层单播流量不通过 peerlink 转发；通过控制路由的优先级，可以控制三层单播流量不通过 peerlink 转发。

图 3-3MC-LAG 防环机制

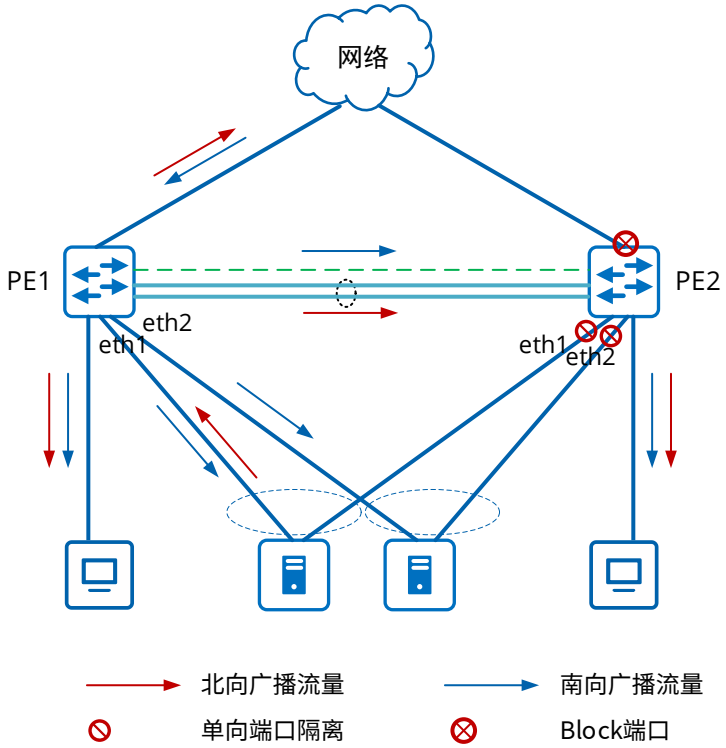


图 3-3 图示了从交换机 PE1 转发 BUM 流量的示意图（从另一侧交换机 PE2 转发时原理相同）。PE1 收到 BUM 报文（北向或者南向），会从除入端口外的同一 VLAN 内的其余端口 flood，包括 peerlink 对应端口。PE2 交换机从 peerlink 端口收到 BUM 报文，会从除入端口和 MC-LAG 成员端口外的同一 VLAN 内的其余端口 flood，也即 peerlink 端口和 MC-LAG 成员端口（图中为端口 eth1 和 eth2）进行了单向隔离。换而言之，当从 peerlink 收到数据报文（包括单播、BUM 报文）时，都不会从 MC-LAG 成员端口进行转发。当 MC-LAG 设备感知到本端某个 MC-LAG 成员端口状态为 down 时，会及时给 peer 发送状态同步信息，通知对端设备取消 peerlink 端口和相关 MC-LAG 成员端口的单向隔离。以图 3-3 为例，当 PE1 监控到端口 eth1 状态迁移为 down，会把此消息通知 PE2；PE2 收到消息，会取消 peerlink 端口和 eth1 端口之间的单向隔离。当 PE1 监控到端口 eth1 状态迁移为 up，也会把此消息通知 PE2；PE2 收到消息，会重新对 peerlink 端口和 eth1 端口进行单向隔离。

如果 PE1、PE2 的北向链路也是二层转发，且是双归属，则需要运行 xSTP 协议来防止环路。图 3-3 中，xSTP 协议计算结果是 block 掉 PE2 的北向端口。

3.7 跨设备链路聚合

如“ 防环机制”一节图示，PE1、PE2 要实现跨设备与 CE1 链路聚合，从 CE1 的角度看起来，CE1 是在与同一台设备的两个端口进行 LACP 协商。聚合协商成功的前提是 PE1、PE2

从 eth1 端口发送的 LACP PDU 报文中源 MAC 地址是相同的，这样在 CE1 看起来是从同一台设备收到的报文。

在两台设备建立起邻居关系以后，standby 会把本设备上所有 MC-LAG 成员端口的 MAC 地址修改为与 active 设备的 system MAC 相同。以上图为例，假设 PE1 为 active，PE2 为 standby。当邻居关系建立以后，PE2 会把自己设备上的 MC-LAG 成员端口 eth1、eth2 端口对应的 MAC 地址修改为 PE1 的系统 MAC。这样从 PE2 成员端口发送的 LACP PDU 源 MAC 地址和 PE1 相同，在 CE1 看起来是从同一台设备收到的报文。

3.8 流量转发

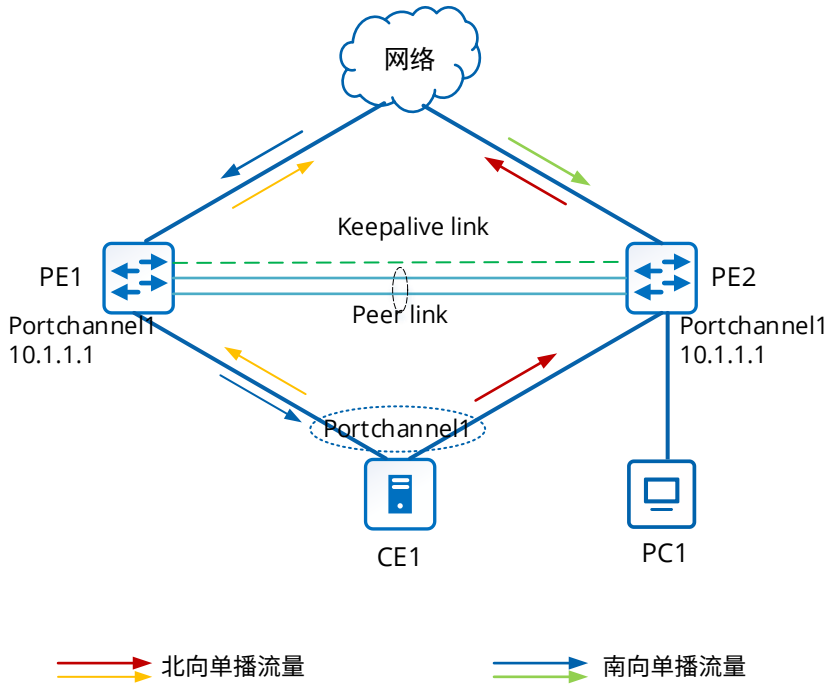
3.8.1 正常工作场景流量转发

所谓的正常工作场景，是指两台设备都处于正常工作状态，包括邻居正常可达、peerlink 相关端口处于 up 状态、MC-LAG 所有成员端口都处于 up 状态等。

1. L3 场景

- L3 单播流量正常转发

图 3-4 三层单播正常转发



在三层转发的场景中，PE1、PE2 一般会扮演网关的角色，其下联 CE 的 portchannel 接口需要配置相同的 ip 地址（如图中 10.1.1.1），也即所谓的 anycast ip。

如图 3-4 所示，南向流量到达 PE 以后，会匹配直连路由从 portchannel1 发送给 CE1，北向流量到达 PE 以后，也会匹配路由向 network 转发。无论是南向还是北向流量，都是在 PE 上直接匹配路由转发，不需要通过 peerlink 从另外一台设备转发。在三层场景中，只需要两台 PE 之间路由可达即可，peerlink 是可选的。

图 3-4 中没有东西向流量转发的图示，如 CE1、PC1 之间的流量，这些流量在 PE 上或者直接匹配到直连路由进行转发，或者进行二层转发（如果在同一 vlan 内）。

- L3 组播流量转发

目前暂不支持 L3 组播流量转发。

2. L2 场景

- L2 单播流量正常转发

图 3-5 二层单播正常转发

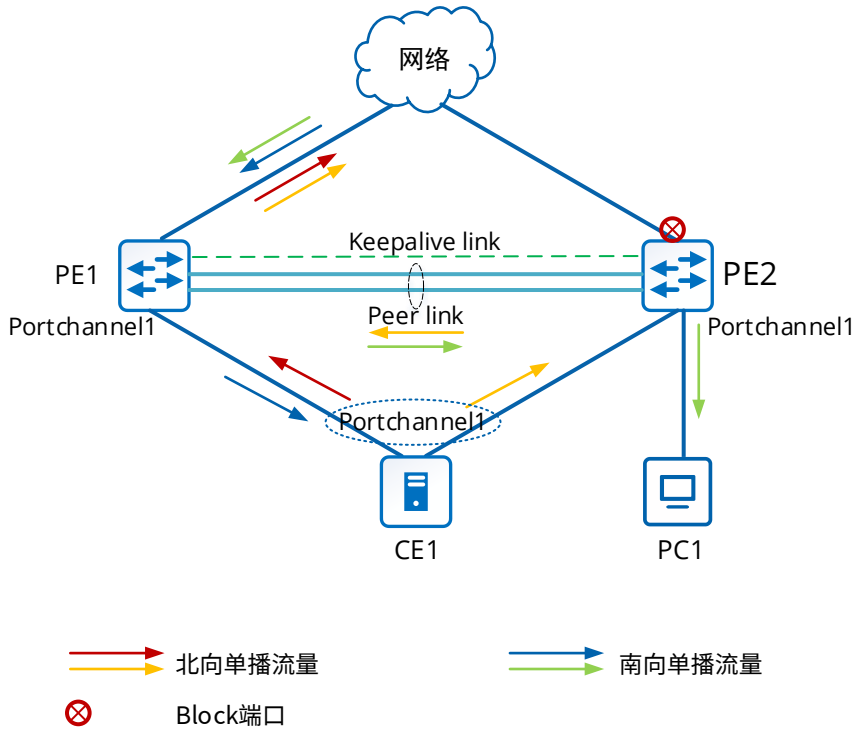


图 3-5 所示的二层单播转发场景中，PE 北向链路也是二层转发，且是双归属。此时需要运行 xSTP 协议来防止环路。图 3-5 中，xSTP 协议计算结果是 block 掉 PE2 的北向端口，也即 PE2 北向链路不会收发流量。

PE1 北向端口是 orphan 端口，其学习到的 MAC 表项同步到 PE2 以后，出端口会指向 peerlink 端口（参考“ MAC 学习及同步”一节）。此时 PE2 收到从 CE1、PC1 发送过来的北向流量都会从 peerlink 单播发送给 PE1，PE1 再查询 FDB 表项向 network 转发。

PE1 收到从 network 发送过来的到 CE1 的南向流量，直接在 PE1 上查询到匹配的 MAC 表项从 portchannel1 转发，无需通过 peerlink 绕到 PE2 转发。而到 PC1 的南向流量，会经 peerlink 到 PE2，由 PE2 转发。

图 3-5 中没有东西向流量转发的图示，如 CE1、PC1 之间的流量，这些流量在 PE 上都可以直接匹配到 FDB 表项进行转发。

图 3-6 上联网关二层单播正常转发

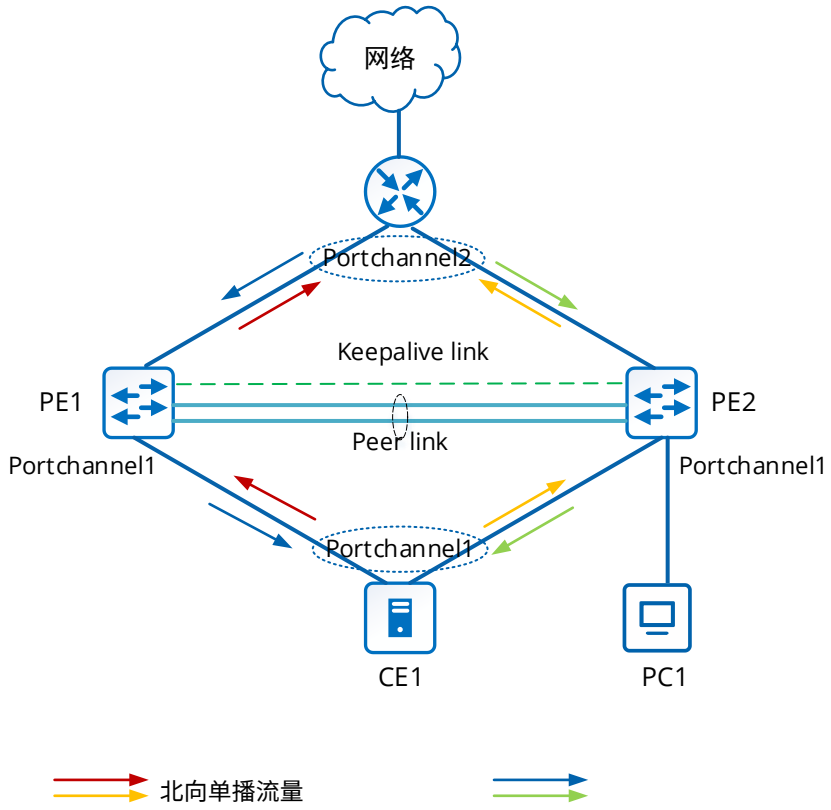


图 3-6 图示了第二种拓扑设计。PE1、PE2 北向接 GW 网关，网关与 PE1、PE2 也形成 MC-LAG 端口聚合（PE 可以与双备份网关连接，图 3-6 只示意了一个网关），这样 PE 设备南北向链路都不会形成环路，也不需要运行 xSTP 协议。在此种拓扑中，peerlink 上除了有从 peer 邻居 orphan port 上学习到的 MAC 表项外，没有从 MC-LAG 成员端口学习到的 MAC 表项，南向、北向流量在 PE 上都能够直接查找到对应的 MAC 表项，而无需经 peerlink 转发。

- L2 BUM 流量正常转发

图 3-7 二层 BUM 正常转发

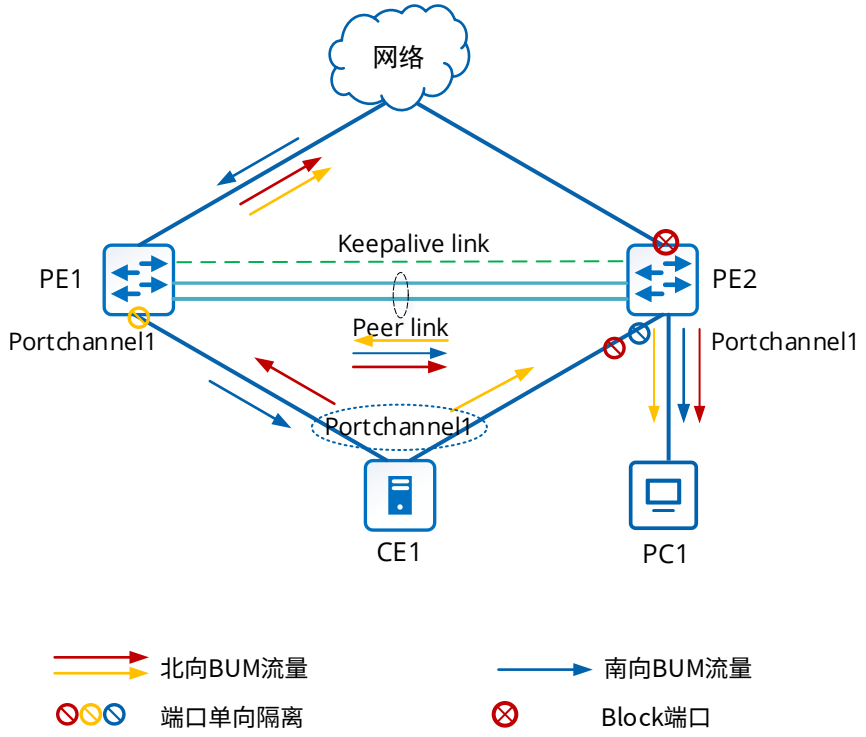


图 3-7 所示的二层 BUM 流量转发场景中，PE 北向链路也是二层转发，且是双归属。此时需要运行 xSTP 协议来防止环路。图 3-7 中，xSTP 协议计算结果是 block 掉 PE2 的北向端口，也即 PE2 北向链路不会收发流量。

PE1 收到从 network 发送过来的南向 BUM 流量，会从 portchannel1、peerlink 端口 flood。PE2 接收到 BUM 流量，会转发给 PC1，但是因为 peerlink 到 portchannel1 单向隔离的原因，不会转发给 CE1，CE1 不会收到两份相同的报文。

PE1 收到从 CE1 发送过来的北向 BUM 流量，会从北向端口、peerlink 端口 flood。PE2 接收到 BUM 流量，会转发给 PC1，但是因为 peerlink 到 portchannel1 单向隔离的原因，不会转发回 CE1。

PE2 收到从 CE1 发送过来的北向 BUM 流量，会从 orphan 端口、peerlink 端口 flood。PE1 接收到 BUM 流量，会从北向端口向 network 转发，但是因为 peerlink 到 portchannel1 单向隔离的原因，不会转发回 CE1。

图 3-8 上联网关二层 BUM 正常转发

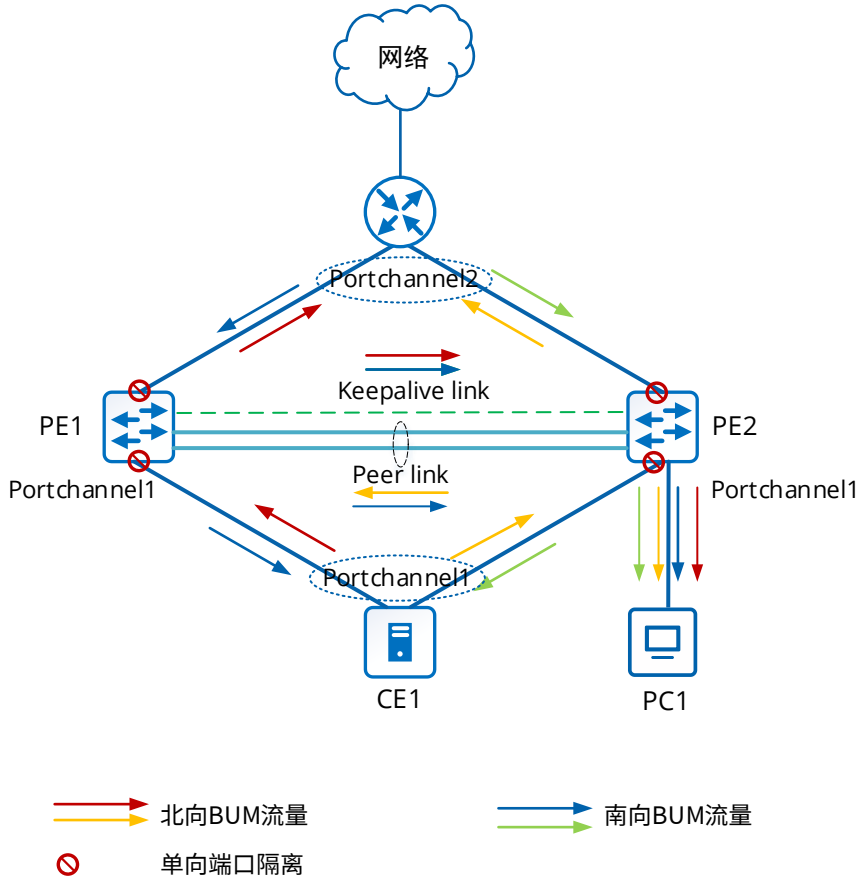


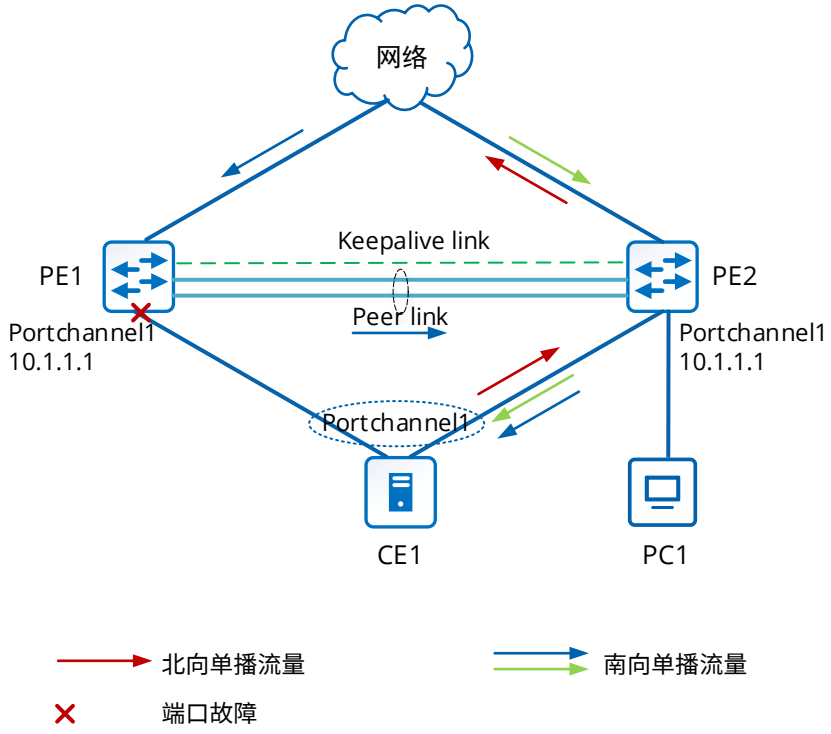
图 3-8 图示了第二种拓扑设计。PE1、PE2 北向接 GW 网关，网关与 PE1、PE2 也形成 MC-LAG 端口聚合。在此种拓扑中，从 peerlink 上 flood 转发的流量到达 peer 邻居以后，都会因为单向隔离而不会从 MC-LAG 成员端口转发，从而保证了同一设备不会收到同一 BUM 流量的多份拷贝。

3.8.2 故障场景流量转发

1. MC-LAG 成员端口 down

- L3 场景

图 3-9 三层场景 MC-LAG 成员端口 down



在三层转发的场景中，PE1、PE2 一般会扮演网关的角色，其下联 CE 的 portchannel 接口需要配置相同的 ip 地址（如图 10.1.1.1），也即所谓的 anycast ip。

为了提供到 CE1 的备份链路，在配置网络时，需要设计在 PE 上生成到 CE1 网段的备份路由（通过路由协议生成或者配置静态路由，如图 3-9 的网段 10.1.1.0/24），备份路由的下一跳为邻居 PE。在 L3 场景中，peer 邻居之间路由可达即可，peerlink 可选。为了图示方便，图 3-9 中配置了 peerlink 作为备份链路，备份路由的下一跳通过 peerlink 可达。正常情况下，直连路由的优先级高于备份路由。

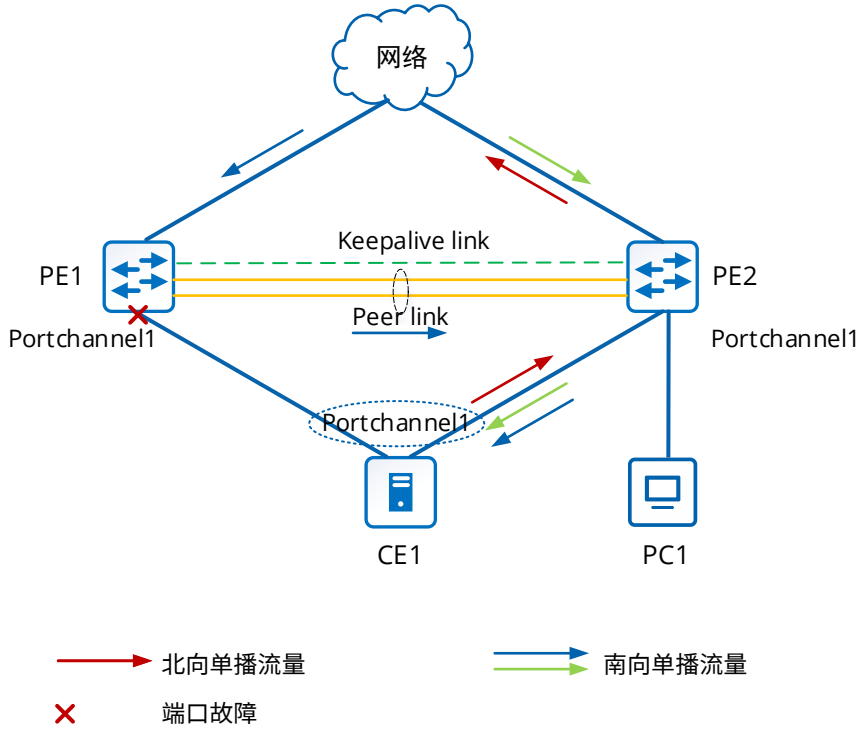
如图 3-9，PE1 监控到 MC-LAG 成员端口 portchannel1 迁移到 down，PE1 马上将此信息同步给 PE2。PE2 收到信息以后，会取消 peerlink 端口到 MC-LAG 成员端口 portchannel1 的单向隔离。PE1 上 MC-LAG 成员端口 portchannel1 迁移到 down，其对应的直连路由会被删除，此时备份路由会成为最优路由。

PE1 收到从 network 发送过来的南向单播流量，匹配备份路由，从 peerlink 相关接口发送给 PE2。PE2 接收到流量，匹配直连路由，从 portchannel1 接口转发给 CE1。

从 CE1 发送的北向流量转发跟正常场景类似。

- L2 场景

图 3-10 二层场景 MC-LAG 成员端口 down



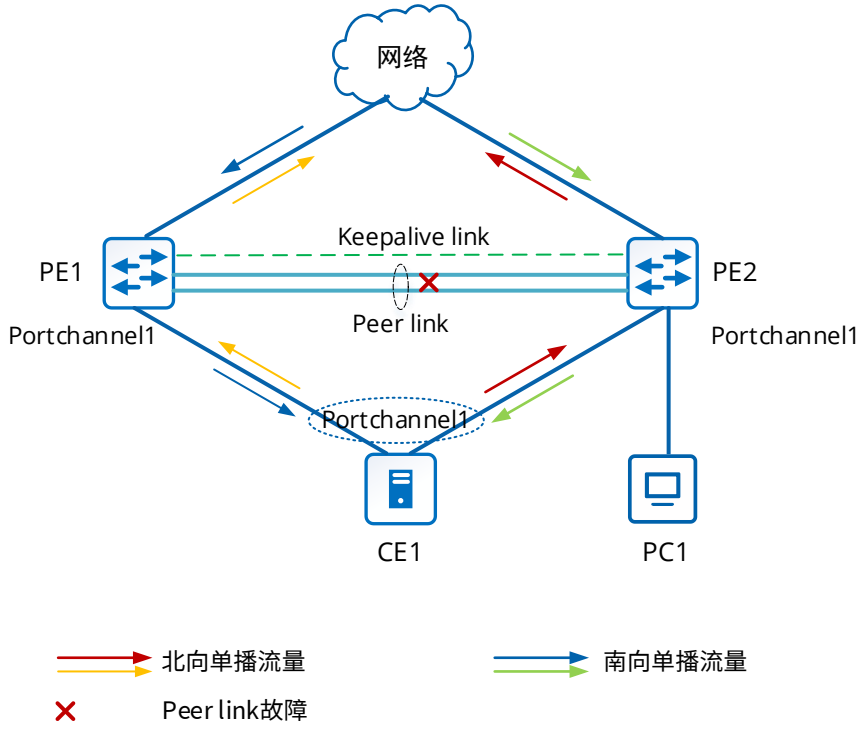
如图 3-10，PE1 监控到 MC-LAG 成员端口 portchannel1 迁移到 down，PE1 马上将此信息同步给 PE2。PE2 收到信息以后，会取消 peerlink 端口到 MC-LAG 成员端口 portchannel1 的单向隔离。PE1 同时会将原指向 MC-LAG 成员端口 portchannel1 的 FDB MAC 表项重定向指向 peerlink 端口。

PE1 收到从 network 发送过来的南向单播流量，查找 FDB MAC 表项，从 peerlink 相关端口发送给 PE2。PE2 接收到流量，再次二层查找，从 portchannel1 接口转发给 CE1。

从 CE1 发送的北向流量转发跟正常场景类似。

2. Peerlink down

图 3-11Peerlink down 故障场景



要让 MC-LAG 邻居双方能同时感知到 peerlink down 事件，一种方法是邻居双方链路直连，另一种方法使用工具例如 BFD 来监控探测链路状态。

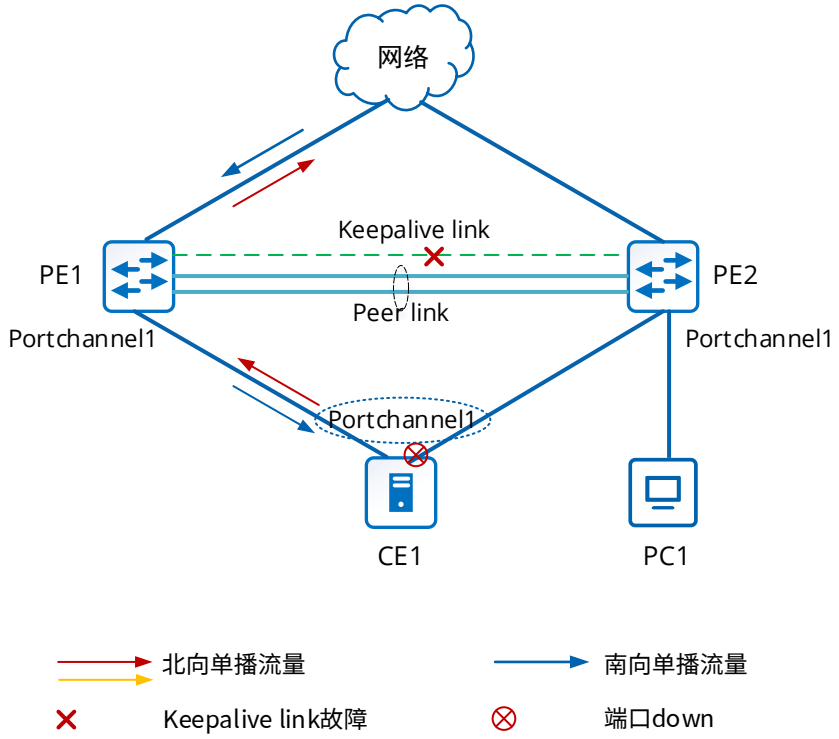
如果 peerlink 同时也是 keepalive link，则 peerlink down 会导致邻居关系断开，其处理请参考下面的“keepalive link down”一节的描述。

当 PE 感知到 peerlink down，邻居双方都会把指向 peerlink 的 MAC 表项删除。因为 peerlink 承担了备份链路的功能，此时经 MC-LAG 成员端口的流量转发不会收到影响，而需要经 peerlink 转发的 orphan port 相关流量则会中断，例如图 3-11 中经 PE1 到 PC1 的流量不可达了。此时如果同时有 MC-LAG 成员端口迁移到 down，则相关流量转发也会收到影响。

3. 邻居断开

- Keepalive link down

图 3-12Keepalive link down 故障场景



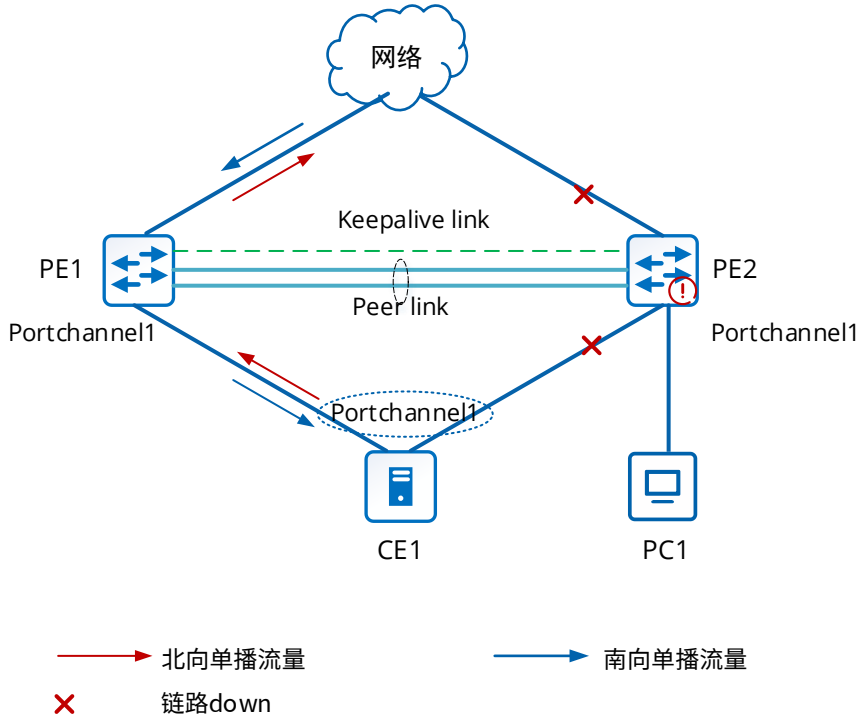
正常情况下，MC-LAG peer 之间路由可达，双方周期性地向对方发送心跳报文。当 keepalive link 出现故障，会导致心跳超时，邻居关系断开。此时 active、standby 都还活着，但彼此之间的联系已经中断了，信息也无法进行同步了。这种故障场景称为脑裂（split-brain scenario）。

如图 3-12 所示，PE1 为 active，PE2 为 standby。当作为 standby 的 PE2 发现邻居关系断开，会把自己成员端口的 system id 修改回自己原始的 MAC 地址。CE1 发现两个聚合端口收到的 LACP PDU 中 system id 不一致，无法再进行聚合操作，只能选择一个端口协商成功。此时 CE1 会 down 掉连接 standby 的链路，而维持与 active 的正常协商，结果是 CE1 收发的流量都需要流经 active 进行转发。

在此故障场景下，从 network 到 CE1 的南向流量会避免流经 PE2。在二层转发场景，此时 PE2 学习不到 CE1 的 MAC 地址表项，因此二层单播流量不会导向 PE2。在三层转发场景，此时 PE2 上的 portchannel1 接口会 down 掉，直连路由也会随之删除，因此三层流量也不会导向 PE2。

- 邻居设备 down

图 3-13 邻居设备 down 故障



当某台设备出现故障，会导致另外一台设备感知到心跳超时，邻居关系断开。

当 active 设备 down 时，连接 CE 设备的成员端口也会随之 down 掉，对端的 CE 马上能够感知到链路 down 的事件。Standby 发现邻居关系断开，会把自己成员端口的 system id 修改回自己原始的 MAC 地址。CE1 会选择跟 standby 进行聚合协商，结果是 CE1 收发的流量都需要流经 standby 进行转发。

类似的原理，当 standby 设备 down 时，连接 CE 设备的成员端口也会随之 down 掉，对端的 CE 马上能够感知到链路 down 的事件。CE1 会维持与 active 的正常协商，结果是 CE1 收发的流量都需要流经 active 进行转发。

4 主要特性

- 目前只支持两台设备建立邻居关系。
- 在一台设备中支持配置大约 520 个 MC-LAG 成员 portchannel 端口。
- Peerlink 两端所连接的端口类型必须相同，例如都为 ethernet 端口或者都为 portchannel。
- 互为 MC-LAG 邻居两台设备上，连接到同一台 CE 设备的 MC-LAG 成员端口名称必须相同，例如设备 1 连接到 CE1 的成员端口名称为 portchannel1，则设备 2 上连接到 CE1 的成员端口名称也必须为 portchannel1。

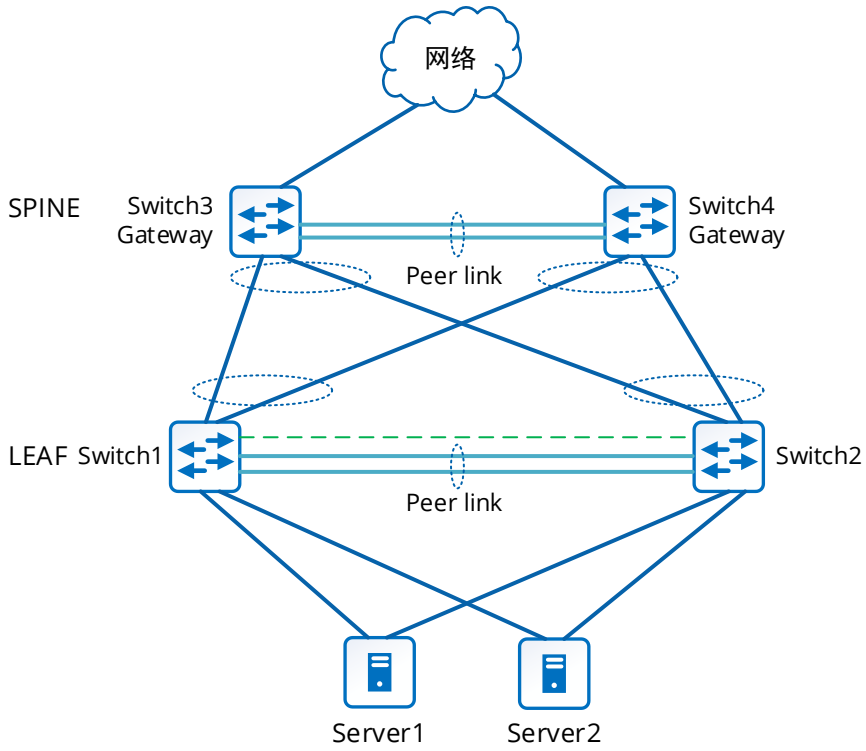
5 典型应用指南

5.1 典型组网方案

图 5-1 是数据中心常用的典型拓扑，采用了两级 MC-LAG 配置。

- Server1、server2 双归接入到 LEAF 交换机，每台 server 的两个上行接口绑定到一个 portchannel 中，可以实现流量的负载分担。
- 在 LEAF 层，Switch1、Switch2 组成 MC-LAG 邻居。Peerlink 接口配置为 portchannel，既可增加带宽，也可以提供链路备份。在两级 MC-LAG 拓扑中，每台 LEAF 设备上的南向、北向端口都需要配置为 MC-LAG 成员端口。Switch1、Switch2 的 portchannel101（包含端口 ethernet1）、portchannel102（包含端口 ethernet2）、portchannel103（包含端口 ethernet20、ethernet21）都需要配置为 MC-LAG 成员端口。
- 在 SPINE 层，Switch3、Switch4 为 gateway，配置相同的 anycast ip 作为网关 ip。Switch3、Switch4 组成 MC-LAG 邻居，peerlink 接口配置为 portchannel，其南向端口都配置为 MC-LAG 成员端口。
- SPINE、LEAF 的 MC-LAG 进行级联，在提供冗余链路的同时消除了环路风险。

图 5-1 数据中心典型拓扑



5.2 MC-LAG 主要配置命令

1. mclag <domain-id>

命令模式：

config 配置模式。

参数说明：

<domain-id>取值范围为 1 到 4095。

命令说明：

配置 MC-LAG domain 域，并且进入 MC-LAG domain 域配置模式。目前系统只支持配置一个 MC-LAG domain。

2. local-ip <ipv4-address> peer-ip <ipv4-address> [peer-link <port-name>]

命令模式：

MC-LAG domain 域配置模式。

参数说明：

<ipv4-address>为本端和对端用于建立邻居关系的 ipv4 地址。

<port-name>为 peerlink 对应的端口名称，目前只能配置为 ethernet 端口或者 portchannel 端口，在三层转发时为可选参数，二层转发时为必选参数。

命令说明：

配置 MC-LAG 本端、对端用于建立连接的 ipv4 地址，配置 peerlink 对应的端口。

3. interface <portchannel-list>

命令模式：

MC-LAG domain 域配置模式。

参数说明：

<portchannel-list>为 MC-LAG 成员端口名称列表，只能配置为 portchannel 端口，在配置多个 portchannel 名称时，名称之间用“,”进行分隔，例如“portchannel101,portchannel102”。

命令说明：

配置 MC-LAG 成员端口名称列表，必须为 portchannel。

4. MC-LAG config-commit

命令模式：

config 配置模式。

参数说明：

无。

命令说明：

提交 MC-LAG 配置，使配置生效。

5.3 具体配置

1. 配置三层交换机 Switch 1

1.1. 端口在加入 portchannel 时，需要配置为 no swtichport 模式

```
Switch1# configure terminal
Switch1(config)# interface ethernet 1-2,20-21,49-50
Switch1(config-if-ethernet1-8,20-21,49-50)# no switchport
Switch1(config-if-ethernet1-8,20-21,49-50)# exit
```

1.2. 配置 peerlink 相关端口 ethernet49、ethernet50 加入 portchannel4096，portchannel4096 加入 vlan2

```
Switch1(config)# vlan 2
Switch1(config-vlan2)# exit
Switch1(config)# interface port-channel 4096
Switch1(config-if-port-channel4096)# set timeout short
Switch1(config-if-port-channel4096)# switchport mode trunk
Switch1(config-if-port-channel4096)# switchport trunk allowed vlan add 2
Switch1(config-if-port-channel4096)# exit
Switch1(config)# interface ethernet 49-50
Switch1(config-if-ethernet49-50)# channel-group 4096
Switch1(config-if-ethernet49-50)# exit
```

1.3. 配置下联 server 的相关端口 ethernet1 加入 portchannel101、ethernet2 加入 portchannel102、ethernet20 和 ethernet21 加入 portchannel103，然后 portchannel101、

portchannel102、portchannel103 也加入 vlan2

```
Switch1(config)# interface port-channel 101-103
```

#配置 portchannel 协商时采用 fast 模式，可以快速感知链路 down 事件

```
Switch1(config-if-port-channel101-103)# set timeout short
Switch1(config-if-port-channel101-103)# switchport mode access
Switch1(config-if-port-channel101-103)# switchport access vlan 2
Switch1(config-if-port-channel101-103)# exit
Switch1(config)# interface ethernet 1
Switch1(config-if-ethernet1)# channel-group 101
Switch1(config-if-ethernet1)# exit
Switch1(config)# interface ethernet 2
Switch1(config-if-ethernet2)# channel-group 102
Switch1(config-if-ethernet2)# exit
Switch1(config)# interface ethernet 20-21
Switch1(config-if-ethernet20-21)# channel-group 103
Switch1(config-if-ethernet1)# exit
```

1.4. 配置 vlan 三层接口，使 MC-LAG 邻居 ipv4 地址可达，可以建立连接及同步信息

```
Switch1(config)# vlan 4000
Switch1(config-vlan4000)# exit
Switch1(config)# interface vlan 4000
Switch1(config-if-vlan4000)# ip address 112.0.0.1/24
Switch1(config-if-vlan4000)# exit
Switch1(config)# interface port-channel 4096
Switch1(config-if-port-channel4096)# switchport trunk native vlan 4000
Switch1(config-if-port-channel4096)# exit
```

1.5. 配置 MC-LAG

#配置 mclag domain，进入相关配置模式

```
Switch1(config)# mclag 1
```

#配置 MC-LAG 本端、对端用于建立连接的 ipv4 地址，配置 peerlink 对应的端口

```
Switch1(config-mclag-1)# local-ip 112.0.0.1 peer-ip 112.0.0.2 peer-link PortChannel4096
```

#配置 MC-LAG 成员端口（必须为 portchannel）

```
Switch1(config-mclag-1)# interface portChannel101,portChannel102,portChannel103
Switch1(config-mclag-1)# exit
```

#提交 MC-LAG 配置，使配置生效

```
Switch1(config)# mclag config-commit
```

2. Switch2 的配置除了配置 MC-LAG 时，将 local-ip、peer-ip 地址进行互换外，其余配置跟 Switch1 相同

```
Switch1(config)# mclag 1
Switch1(config-mclag-1)# local-ip 112.0.0.2 peer-ip 112.0.0.1 peer-link PortChannel4096
Switch1(config-mclag-1)# interface portChannel101,portChannel102,portChannel103
Switch1(config-mclag-1)# exit
Switch1(config)# mclag config-commit
```

3. SPINE 层交换机 Switch3、Switch4 的 MC-LAG 配置与 LEAF 交换机 Switch1、Switch2 基本类似，此处不赘述

4. Server 上行接口可以运行 LACP，也可以通过 bond 进行静态绑定，具体配置跟 server 安装的系统相关，此处也不赘述

6 维护

下面以“典型应用指南”一章所举例子介绍如何监控 MC-LAG 模块运行状态及进行相关的故障排查。

1. 查看 MC-LAG 邻居关系状态

在 Switch1 上查看：

```
Switch1# show running-config mclag
The MCLAG's keepalive is: OK
Domain id: 1
Local Ip: 112.0.0.1
Peer Ip: 112.0.0.2
Peer Link Interface: PortChannel4096
Role: Active
MC-LAG Interface: PortChannel102,PortChannel101
Loglevel: NOTICE
```

在 Switch2 上查看：

```
Switch2# show running-config mclag
The MCLAG's keepalive is: OK
Domain id: 1
Local Ip: 112.0.0.2
Peer Ip: 112.0.0.1
Peer Link Interface: PortChannel4096
Role: Standby
MC-LAG Interface: PortChannel102,PortChannel101
Loglevel: NOTICE
```

“The MCLAG's keepalive is: OK”表示邻居关系建立正常；如果显示“The MCLAG's keepalive is: ERROR”则表示邻居断开，造成此现象的原因一般是邻居 down 掉、连线错误、邻居 ipv4 地址配置错误、路由不可达等。需要保证建立邻居的 ipv4 地址相互可达，可以使用 ping 命令来验证对端地址是否可达。

2. 查看 MC-LAG 成员端口、peerlink 端口状态

在 Switch1、Switch2 上查看相关端口状态，示例中都是 portchannel 端口

```
Switch1# show port-channel summary
Flags: A - active, I - inactive, Up - up, Dw - Down, N/A - not available,
                S - selected, D - deselected, * - not synced
Load Balance: default
  No.       Team Dev              Protocol            Ports                                  MTU
Fallback      Time Out          Min Links
----    ------------    -----------    -------------------    -----    -------    ---------    -----------
1     PortChannel101      LACP(A)(Up)     Ethernet1(S)                          9100     false
short                     1
2     PortChannel102      LACP(A)(Up)     Ethernet2(S)                          9100     false
short                     1
3     PortChannel103      LACP(A)(Up)     Ethernet49(S)                         9100     false
short                     1
```

正常情况下，MC-LAG 成员端口、peerlink 端口都需要处于 Up 状态，否则需要追查没有 up 的原因。