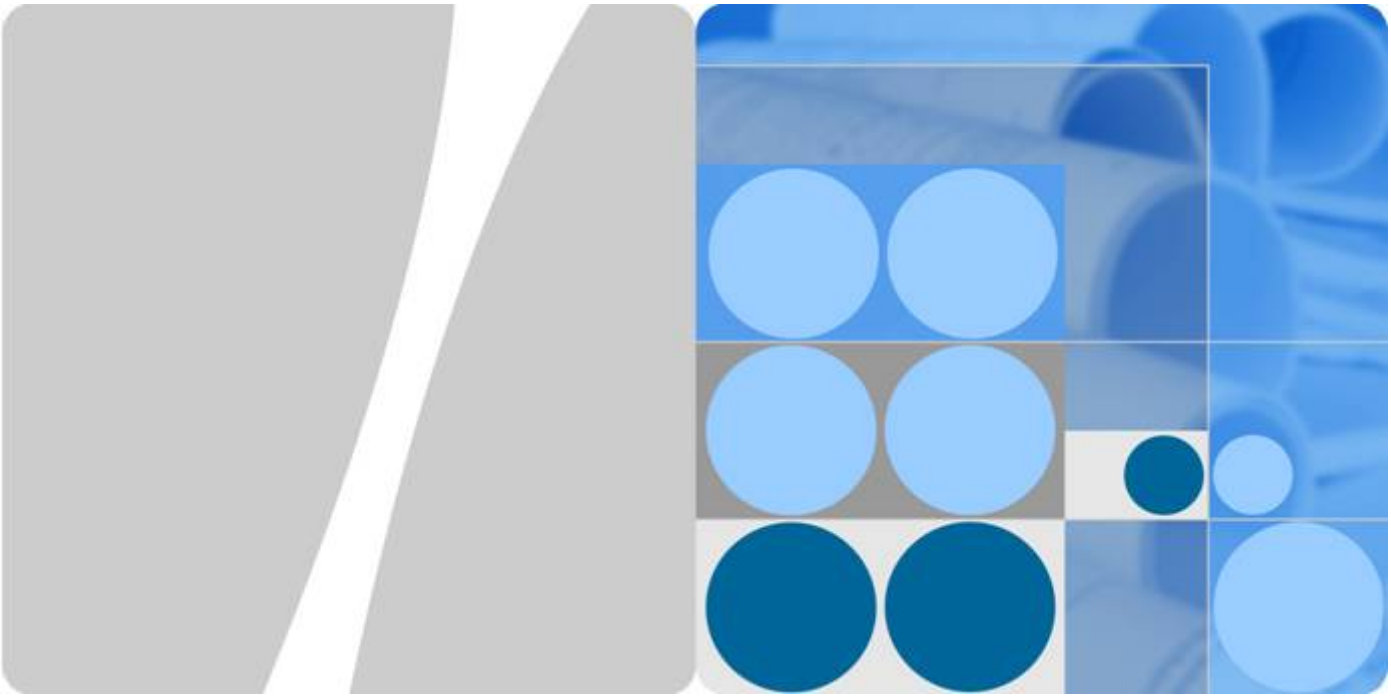


HUAWEI NetEngine80E/40E 路由器

V600R003C00

特性描述-可靠性

文档版本　02

发布日期　2011-09-10

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默示的声明或保证。

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： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

前 言

概述

本文档针对可靠性，从简介、原理描述和应用三个方面介绍了可靠性相关特性。

本文档与其它类型手册相结合，便于读者深入掌握特性的实现原理。

产品版本

与本文档相对应的产品版本如下所示。

产品名称	产品版本
HUAWEI NetEngine80E/40E	V600R003C00

读者对象

本文档主要适用于以下工程师：

- 网络规划工程师
- 调测工程师
- 数据配置工程师
- 系统维护工程师

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	以本标志开始的文本表示有高度潜在危险，如果不能避免，会导致人员死亡或严重伤害。

符号	说明
 警告	以本标志开始的文本表示有中度或低度潜在危险，如果不能避免，可能导致人员轻微或中等伤害。
 注意	以本标志开始的文本表示有潜在风险，如果忽视这些文本，可能导致设备损坏、数据丢失、设备性能降低或不可预知的结果。
窍门	以本标志开始的文本能帮助您解决某个问题或节省您的时间。
说明	以本标志开始的文本是正文的附加信息，是对正文的强调和补充。

修订记录

修改记录累积了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 02 (2011-09-10)

第二次正式发布，文档内容更新如下：

相对于上一版本无变更。

文档版本 01 (2011-05-30)

第一次正式归档。

目 录

1 VRRP

关于本章

1.1 介绍

定义

VRRP（Virtual Router Redundancy Protocol）虚拟路由冗余协议，是一种容错协议。该协议通过把几台路由设备联合组成一台虚拟的路由设备，使用一定的机制保证当主机的下一跳路由器出现故障时，及时将业务切换到其它路由器，从而保持通讯的连续性和可靠性。

以下是与 VRRP 协议相关的基本概念：

- VRRP 路由器（VRRP Router）：运行 VRRP 的设备，它可能属于一个或多个虚拟路由器。
- 虚拟路由器（Virtual Router）：由 VRRP 管理的抽象设备，又称为 VRRP 备份组，被当作一个共享局域网内主机的缺省网关。它包括了一个虚拟路由器标识符和一组虚拟 IP 地址。
- 虚拟 IP 地址(Virtual IP Address)：虚拟路由器的 IP 地址，一个虚拟路由器可以有一个或多个 IP 地址，由用户配置。
- IP 地址拥有者（IP Address Owner）：如果一个 VRRP 路由器将虚拟路由器的 IP 地址作为真实的接口地址，则该设备是 IP 地址拥有者。当这台设备正常工作时，它会响应目的地址是虚拟 IP 地址的报文，如 Ping、TCP 连接等。
- 虚拟 MAC 地址：是虚拟路由器根据虚拟路由器 ID 生成的 MAC 地址。一个虚拟路由器拥有一个虚拟 MAC 地址，格式为：00-00-5E-00-01-{VRID}(VRRP)；00-00-5E-00-02-{VRID}(VRRP6)。在 IPv4 网络中，当虚拟路由器回应 ARP 请求时，使用虚拟 MAC 地址，而不是接口的真实 MAC 地址。
- 主 IP 地址（Primary IP Address）：从接口的真实 IP 地址中选出来的一个主用 IP 地址，通常选择配置的第一个 IP 地址。VRRP 广播报文使用主 IP 地址作为 IP 报文的源地址。
- Master 路由器（Virtual Router Master）：是承担转发报文或者应答 ARP 请求的 VRRP 路由器，转发报文都是发送到虚拟 IP 地址的。如果 IP 地址拥有者是可用的，通常它将成为 Master。
- Backup 路由器（Virtual Router Backup）：一组没有承担转发任务的 VRRP 路由器，当 Master 设备出现故障时，它们将通过竞选成为新的 Master。
- 抢占模式：在抢占模式下，如果 Backup 路由器的优先级比当前 Master 路由器的优先级高，将主动将自己升级成 Master。

目的

随着 Internet 的发展，人们对网络的可靠性的要求越来越高。对于局域网用户来说，能够时刻与外部网络保持联系非常重要。

通常情况下，内部网络中的所有主机都设置一条相同的缺省路由，指向出口网关，实现主机与外部网络的通信。当出口网关发生故障时，主机与外部网络的通信就会中断。

配置多个出口网关是提高系统可靠性的常见方法，但局域网内的主机设备通常不支持动态路由协议，如何在多个出口网关之间进行选路是一个需要解决的问题。

VRRP 协议由因特网工程任务组 IETF（Internet Engineering Task Force）推出，旨在解决局域网主机访问外部网络的可靠性问题，包括如下应用特性：

- 主备备份：这是 VRRP 提供 IP 地址备份功能的基本方式。主备备份方式需要建立一个虚拟路由器，该虚拟路由器包括一个 Master 设备和若干 Backup 设备，这些路由器构成一个备份组。正常情况下，业务全部由 Master 承担。Master 出现故障时，Backup 接替工作。
- VRRP 负载分担：负载分担方式是指多台路由器同时承担业务，单个 VRRP 备份组是不具备负载分担功能的，只有在多台设备上建立两个或更多的备份组，所有备份组均匀分担 Master 状态，此时就每台设备只承担了部分的业务，从而达到负载分担的作用。
- VRRP 监视接口状态：每个 VRRP 备份组可以监视所有与此 VRRP 备份组绑定的接口的状态，从而当接口出现故障时，VRRP 通过改变优先级来重新选择主备关系。
- 虚拟 IP 地址 Ping 开关：提供了控制 Ping 通虚拟 IP 地址的开关命令。
- VRRP 的安全功能：对于安全程度不同的网络环境，可以在报头上设定不同的认证方式和认证字。
- VRRP 平滑倒换：CE 设备作为业务系统的网关，需要启用 VRRP 冗余备份功能，此时当设备处于进行主备倒换的过程中，本端和对端设备都不会出现 VRRP 状态切换，从而防止业务丢包。
- 管理 VRRP：
  - 一个管理 VRRP 备份组可以绑定多个业务 VRRP 备份组，但是它不可以作为业务 VRRP 和其他管理 VRRP 进行绑定。
  - 一个业务 VRRP 只能绑定到一个管理 VRRP。管理 VRRP 绑定多个业务 VRRP 后，通过只在管理 VRRP 上运行 VRRP 协议报文来决定其下面的所有业务 VRRP 的状态，这样极大降低 VRRP 协议报文数量，提高了协议报文处理效率。
  - 管理 VRRP 通过监视 Peer BFD 会话和 Link BFD 会话状态实现主备快速切换。VPLS 汇聚的方案中，为了提高可靠性，通常部署双 NPE，采用 VPLS 汇聚，VSI 之间的 PW、AC 接口监视管理 VRRP 的状态来决定主备 PW 和主备 AC 接口。除此功能外，管理 VRRP 的其它属性同普通 VRRP 备份组。
- VRRP 快速切换：VRRP 通过监视 BFD 会话状态实现主备快速切换，主备切换时间毫秒级。

1.2 参考标准和协议

本特性的参考资料清单如下：

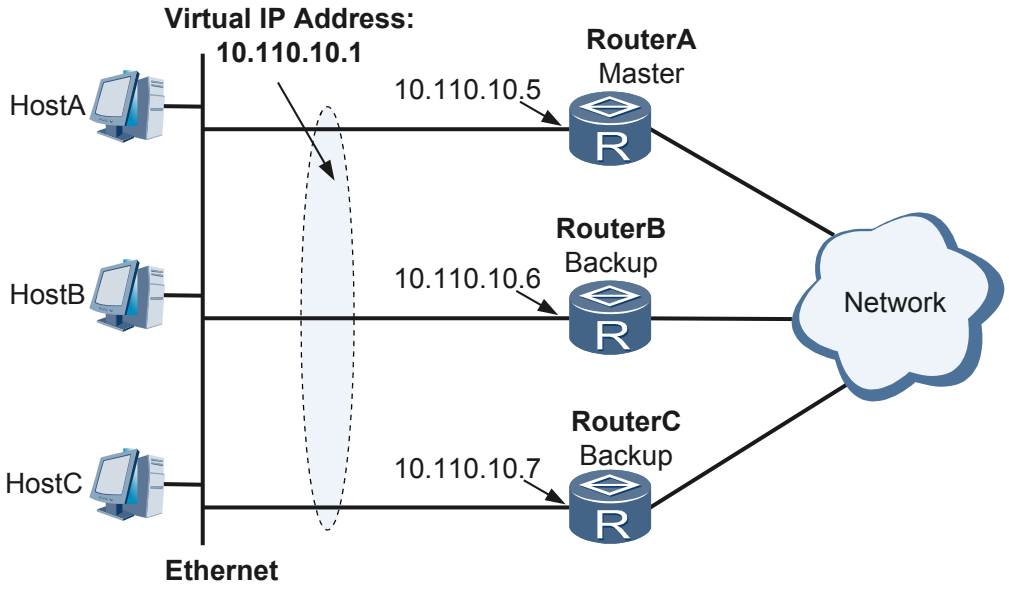
文档	描述	备注
RFC2281	Hot Standby Router Protocol（HSRP）	-
RFC2338	Virtual Router Redundancy Protocol（version number One1998）	-
RFC2787	Definitions of Managed Objects for the Virtual Router Redundancy Protocol	-
RFC3768	Virtual Router Redundancy Protocol（version number Two 2004）	-
RFC5798	Virtual Router Redundancy Protocol Version 3 for IPv4 and IPv6	-

1.3 原理描述

VRRP 将局域网的一组路由设备构成一个 VRRP 备份组，相当于一台虚拟路由器。局域网内的主机只需要知道这个虚拟路由器的 IP 地址，并不需知道具体某台设备的 IP 地址，将网络内主机的缺省网关设置为该虚拟路由器的IP 地址，主机就可以利用该虚拟网关与外部网络进行通信。

VRRP 将该虚拟路由器动态关联到承担传输业务的物理设备上，当该设备出现故障时，再次选择新设备来接替业务传输工作，整个过程对用户完全透明，实现了内部网络和外部网络不间断通信。

图 1-1 虚拟路由器示意图



如图 1-1 所示，虚拟路由器的实现原理如下：

- RouterA、RouterB 和 RouterC 属于同一个 VRRP 备份组，组成一个虚拟的路由器，这个虚拟路由器有自己的 IP 地址 10.110.10.1。虚拟 IP 地址可以直接指定，也可以借用该 VRRP 组所包含的设备上某接口地址。
- RouterA、RouterB 和 RouterC 的实际 IP 地址分别是 10.110.10.5、10.110.10.6 和 10.110.10.7。
- 局域网内的主机只需要将缺省路由设为 10.110.10.1 即可，无需知道具体设备上的接口地址。

主机利用该虚拟网关与外部网络通信。虚拟路由器工作机制如下：

- 根据优先级的大小挑选 Master 设备。Master 设备的选举有两种方法：
  - 比较优先级的大小，优先级高者当选为 Master 设备。

- 当两台优先级相同的设备，如果已经存在 Master，则 Backup 设备不进行抢占；如果同时竞争 Master，则比较接口 IP 地址大小，IP 地址较大的接口所在设备当选为 Master 设备。

- 其它设备作为备份设备，随时监听 Master 设备的状态。
  - 当主设备正常工作时，它会每隔一段时间（Advertisement_Interval）发送一个 VRRP 组播报文，以通知组内的备份设备，主设备处于正常工作状态。
  - 当组内的备份设备一段时间（Master_Down_Interval）内没有接收到来自主设备的报文，则将自己转为主设备。一个 VRRP 组里有多台备份设备时，短时间内可能产生多个 Master 设备，此时，设备将会将收到的 VRRP 报文中的优先级与本地优先级做比较。从而选取优先级高的设备做 Master。设备的状态变为 Master 之后，会立刻发送免费 ARP 来刷新交换机上的 Mac 表项，从而把用户的流量引到此台设备上来，整个过程对用户完全透明。

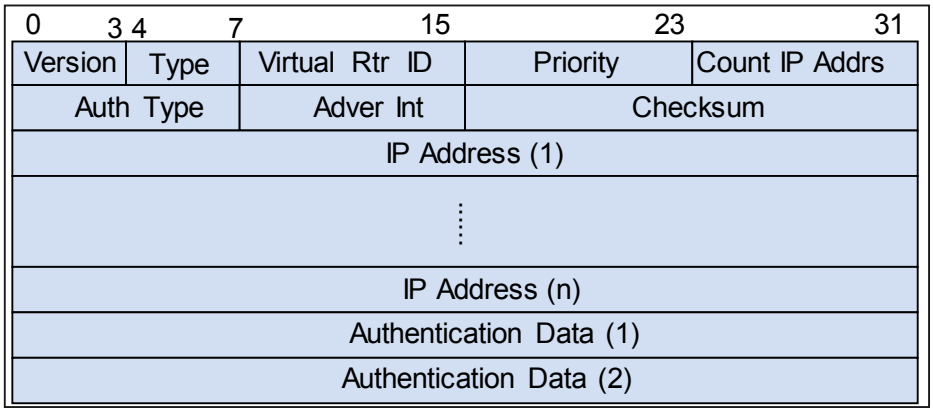
从上述分析可以看到，主机不需要增加额外工作，与外界的通信也不会因某台设备故障而受到影响。

VRRP 报文结构

VRRP 报文用来将 Master 设备的优先级和状态通告给同一虚拟路由器的所有 VRRP 路由器。

VRRP 报文封装在 IP 报文中，发送到分配给 VRRP 的 IP 组播地址。在 IP 报文头中，源地址为发送报文的主接口地址（不是虚拟地址或辅助地址），目的地址是 224.0.0.18，TTL 是 255，协议号是 112。VRRP 报文的结构如图 1-2 所示。

图 1-2 VRRP 报文结构



各字段的含义如下：

- Version：VRRP 协议版本号。此处取值为 2。
- Type：VRRP 通告报文的类型。只有一种取值 1，表示 Advertisement。
- Virtual Rtr ID（VRID）：虚拟路由器 ID，取值范围是 1 ～ 255。
- Priority：发送 VRRP 通告报文的设备在备份组中的优先级。取值范围是 0 ～ 255，但可用的范围是 1 ～ 254。0 表示设备停止参与 VRRP 备份组，用来使备份设备尽快成为 Master 设备，而不必等到计时器超时；255 则保留给 IP 地址拥有者。缺省值是 100。

- Count IP Addrs：VRRP 通告报文中包含的虚拟 IP 地址的个数。
- Authentication Type：VRRP 报文的认证类型。协议中指定了 3 种类型：
  - 0：Non Authentication
  - 1：Simple Text Password
  - 2：IP Authentication Header

  说明

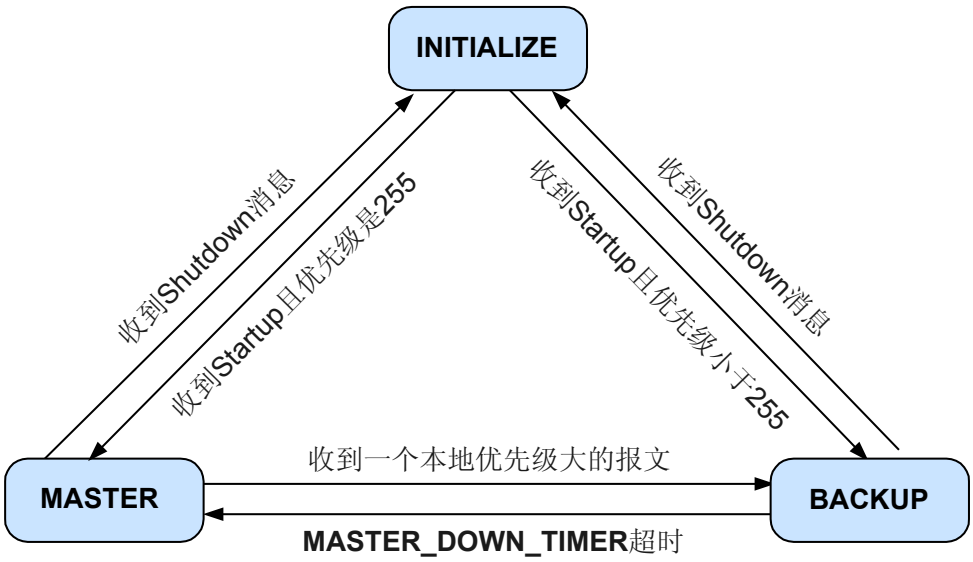
  目前，NE80E/40E 实现了
  - Simple Text Password：即明文认证方式。
  - IP Authentication Header：采用 MD5 认证方式。
- Advertisement Interval：发送通告报文的时间间隔。缺省值为 1 秒。
- Checksum：校验和。
- IP Address(es)：VRRP 备份组的虚拟 IP 地址。
- Authentication Data：认证字。目前只有明文认证和 MD5 认证才用到该部分，对于其它认证方式，一律填 0。

VRRP 的状态机

VRRP 协议中定义了三种状态机：初始状态（Initialize）、活动状态（Master）、备份状态（Backup）。其中，只有处于活动状态的设备才可以转发那些发送到虚拟 IP 地址的报文。

VRRP 状态转换如图 1-3 所示。

图 1-3 VRRP 状态机的转换



Initialize：设备启动时进入此状态，当收到接口 Startup 的消息，将转入 Backup 或 Master 状态（IP 地址拥有者的接口优先级为 255，直接转为 Master）。在此状态时，不会对 VRRP 通告报文做任何处理。

Master：当路由器处于 Master 状态时，它将会做下列工作：

- 定期发送 VRRP 通告报文。
- 以虚拟 MAC 地址响应对虚拟 IP 地址的 ARP 请求。
- 转发目的 MAC 地址为虚拟 MAC 地址的 IP 报文。
- 如果它是这个虚拟 IP 地址的拥有者，则接收目的 IP 地址为这个虚拟 IP 地址的 IP 报文。否则，丢弃这个 IP 报文。
- 如果收到比自己优先级大的报文则转为 Backup 状态。
- 当接收到接口的 Shutdown 事件时，转为 Initialize 状态。

Backup：当路由器处于 Backup 状态时，它将会做下列工作：

- 接收 Master 发送的 VRRP 通告报文，判断 Master 的状态是否正常。
- 对虚拟 IP 地址的 ARP 请求，不做响应。
- 丢弃目的 MAC 地址为虚拟 MAC 地址的 IP 报文。
- 丢弃目的 IP 地址为虚拟 IP 地址的 IP 报文。
- 如果收到比自己优先级小的报文时，丢弃报文，不重置定时器；如果收到优先级和自己相同的报文，则重置定时器，不进一步比较 IP 地址。
- 当接收到 MASTER_DOWN_TIMER 定时器超时的事件时，才会转为 Master 状态。
- 当接收到接口的 Shutdown 事件时，转为 Initialize 状态。

1.3.1 主备备份

这是 VRRP 提供 IP 地址备份功能的基本方式。主备备份方式需要建立一个虚拟路由器，该虚拟路由器包括一个 Master 和若干 Backup 设备。

- 正常情况下，业务全部由 Master 承担。
- Master 出现故障时，Backup 设备接替工作。

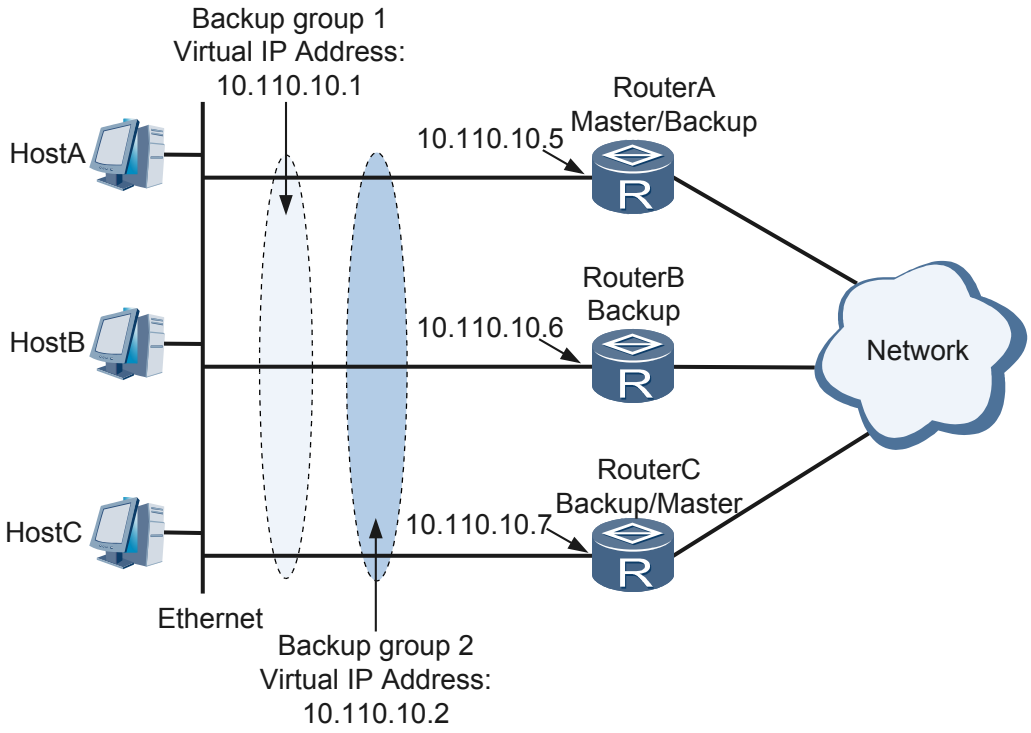
1.3.2 VRRP 负载分担

允许一台设备为多个 VRRP 备份组作备份。通过多个虚拟路由器可以实现负载分担。负载分担方式是指多台虚拟路由器同时承担业务，因此需要建立两个或更多的备份组。

负载分担方式具有以下特点：

- 每个备份组都包括一个 Master 设备和若干 Backup 设备。
- 各备份组的 Master 设备可以不同。
- 同一台设备可以加入多个备份组，在不同备份组中有不同的优先级。

图 1-4 VRRP 负载分担示意图



如图 1-4 所示，配置两个备份组：组 1 和组 2。

- RouterA 在备份组 1 中作为 Master，在备份组 2 中作为 Backup。
- RouterB 在备份组 1 和 2 中都作为 Backup。
- RouterC 在备份组 2 中作为 Master，在备份组 1 中作为 Backup。
- 一部分主机使用备份组 1 作网关，另一部分主机使用备份组 2 作为网关。

这样，可以达到分担数据流而又相互备份的目的。

1.3.3 VRRP 监视接口状态

VRRP 可以监视所有接口的状态。当被监视的接口 Down 或 Up 时，该设备的优先级会自动降低或升高一定的数值，使得备份组中各设备优先级高低顺序发生变化，VRRP 设备重新进行 Master 设备竞选。

VRRP 可以通过 Increased 方式和 Reduced 方式来监视接口（一个 VRRP 最多可以监视 8 个接口）。

- 如果 VRRP 以 Increased 方式监视一个接口，当被监视的接口状态变成 Down 后，VRRP 的优先级增加（增加值可以配置）。

  Increased 方式在 VRRP 状态为 Master 或 Backup 时都生效。
- 如果 VRRP 以 Reduced 方式监视一个接口，当被监视的接口状态变为 Down 后，VRRP 的优先级降低（降低值可以配置）。

  Reduced 方式在 VRRP 状态为 Master 或 Backup 时都生效。

具体的应用场景可以参考组网应用中的“VRRP 监视接口状态”。

1.3.4 VRRP 快速切换

双向转发检测 BFD（Bidirectional Forwarding Detection）机制能够快速检测、监控网络中链路或者 IP 路由的连通状况，VRRP 通过监视 BFD 会话状态实现主备快速切换，主备切换的时间控制在 1 秒以内。

对于以下情况，BFD 都能够将检测到的故障通知接口板，从而加快 VRRP 主备倒换的速度。

- 备份组包含的接口出现故障。
- Master 和 Backup 不直接相连。
- Master 和 Backup 直接相连，但在中间链路上存在传输设备。

BFD 对 Backup 和 Master 之间的实际地址通信情况进行检测，如果通信不正常，Backup 就认为 Master 已经不可用，升级成 Master。在以下情况下 Backup 转换为 Master：

- 当两台设备之间的背靠背连接全部断开时，Backup 主动升级成 Master，承载上行流量。
- 当两台设备之间的连接在如下情况下中断时，Backup 主动升级成 Master，承载上行流量。
  - 当 Master 重新启动
  - Master 与交换机之间的链路断开
  - 与 Master 相连的交换机重新启动

VRRP 快速切换的环境要求：

- 在 Backup 上，BFD Session 检测的接口必须和 Master 设备相连；
- 在 Master 不可用时，Backup 的优先级增加并大于原来 Master 的优先级，促使自己快速切换为 Master。

1.3.5 EFM for VRRP

在城域以太网解决方案 V1R3 中，主备 NPE 之间和 NPE-UPE 之间都是通过 BFD 来实现链路检测和保护的，但如果 UPE 设备不支持 BFD 时，则无法适用。若 UPE 支持 802.3ah，则 NPE-UPE 之间可使用 802.3ah 来替代 BFD。

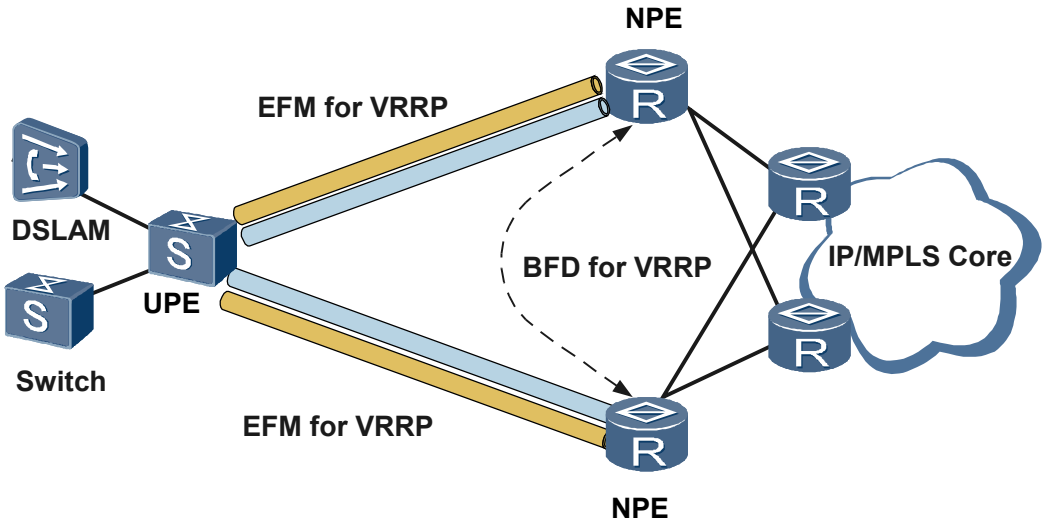
遵循 IEEE 802.3ah 协议的 EFM OAM（Ethernet in the First Mile OAM）属于链路级以太网 OAM 技术，针对两台直连设备之间的链路，提供链路连通性检测功能、链路故障监控功能、远端故障通知功能、远端环回功能。

VRRP（Virtual Router Redundancy Protocol）是 RFC2338/RFC3768 定义的一种容错协议，通过物理设备和逻辑设备的分离，实现在多个出口网关之间选路。同一备份组中有多个成员，只有一个作为主设备，负责流量转发，其他作为备份设备，当主设备故障时，在备份设备中再选择出一个主设备，保证通讯的连续性和可靠性。

802.3ah for VRRP 特性实现的是 VRRP 与 EFM 联动的功能，EFM 用于检测近端链路（本端 NPE 与 UPE 之间的链路）状态，Peer BFD 负责检测远端链路（本端 NPE 与对端 NPE 之间的链路）状态。VRRP 根据 EFM 和 Peer BFD 上报的链路状态，进行状态机计算，选择主备设备。

EFM for VRRP 组网图

图 1-5 EFM for VRRP 组网图

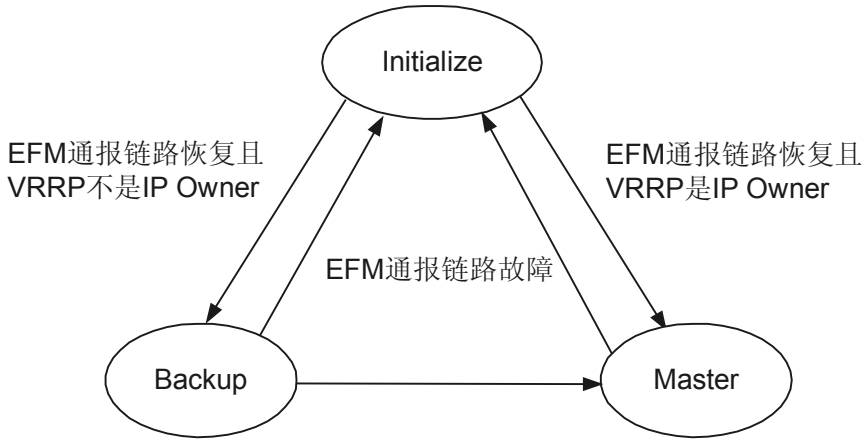


如图 1-5 所示，NPE 和 UPE 之间运行 802.3ah 故障检测协议，同时在 NPE 之间配置 BFD 故障检测协议，检测 NPE 之间是否出现故障，且要求 NPE 之间的故障检测协议报文通过 UPE 转发，通过这两种机制的保护，可以实现主备 NPE 的快速切换。

802.3ah 协议仅用于检测单跳链路的状态，无法检测跨设备的链路。

EFM for VRRP 状态机图

图 1-6 EFM for VRRP 状态机图



本特性相关的状态机如图 1-6 所示，VRRP 的基本状态机请查看 RFC2338/3768。

在 VRRP 所在接口 UP 的情况下，若 VRRP 处于 initialize 状态时，接收到 802.3ah 通报链路故障恢复，若该备份组是 IP Owner，则直接迁移至 master，否则迁移至 backup。若

VRRP 处于 backup 状态，接收到 peer BFD 通报链路故障，则迁移至 master。VRRP 处于非 initialize 状态时，若接收到 802.3ah 通报链路故障，则直接迁移至 initialize。

1.3.6 虚拟 IP 地址 Ping 开关

VRRP 备份组使用虚拟 IP 地址，能够 Ping 通虚拟 IP 地址可以比较方便的监控虚拟路由器的工作情况，但是带来可能遭到 ICMP 攻击的隐患。控制 Ping 通虚拟 IP 地址的开关命令供用户方便处理。

1.3.7 VRRP 安全

对于安全程度不同的网络环境，可以在报头上设定不同的认证方式和认证字。

在一个安全的网络中，可以采用缺省设置：设备对要发送的 VRRP 报文不进行任何认证处理，收到 VRRP 报文的设备也不进行任何认证，认为收到的都是真实的、合法的 VRRP 报文。这种情况下，不需要设置认证字。

在有可能受到安全威胁的网络中，VRRP 提供了简单字符（Simple）认证方式和 MD5 认证方式。对于简单认证字方式，可以设置长度为 1 ～ 8 的认证字；对于 MD5 认证方式，明文长度范围是 1 ～ 8，密文长度为 24。

1.3.8 VRRP 平滑倒换

当 Master 设备发生主备倒换后，从发生主备倒换到新主控板正常工作需要一段时间。该时间随不同设备和不同配置差别较大，结果可能导致 Master 设备不能正常处理 VRRP 协议报文，Backup 设备因为收不到广播报文而抢占到 Master 状态，并针对每一个虚拟路由器的虚拟 IP 地址发送免费 ARP，给相关绑定模块发送状态变化通知。在抢占方式下，如果原 Master 设备优先级高，则此设备倒换完成后又将抢占到 Master 状态，从而导致状态的两次抖动，影响业务流量。

因此需要启用了 VRRP 功能的设备支持 VRRP 的平滑倒换功能，避免因主备倒换影响业务流量。

在设备主用主控板和备用主控板状态都正常的情况下，VRRP 备份组中的 Master 设备会以 Advertisement_Interval 间隔定时发送 VRRP 广播报文，Backup 设备通过不断检测接收到的广播报文来判断 Master 设备状态是否正常。

在 VRRP 平滑倒换的过程中，Master 和 Backup 分工不同，相互配合，共同保证业务的平滑传输。

- 要进行 VRRP 整机平滑倒换处理，必须分别在 Master 和 Backup 上使能 VRRP 协议报文时间间隔学习功能。
  - 如果使能了 VRRP 协议报文时间间隔学习功能，Master 状态的 VRRP 不学习也不检查协议报文时间间隔的一致性。
  - 非 Master 状态的 VRRP 收到 Master 状态 VRRP 发来的协议报文后，会检查报文中的时间间隔值，如果和自己的不同，非 Master 状态的 VRRP 就会学习到报文中的时间间隔，并调整自己的协议报文时间间隔值，与报文中的值保持一致。
- RouterA 配置整机 VRRP 平滑倒换功能。设备主备倒换，新的主板启动后，VRRP 根据设备主备倒换前的状态判断，保存当前配置的 VRRP 协议报文时间间隔，并对 Master 状态的 VRRP 进行协议报文时间间隔调整，然后以当前配置的时间间隔发出 VRRP 平滑倒换报文，报文中携带着新的时间间隔发送到对端 RouterB。
- RouterB 收到的 VRRP 协议报文中携带的时间间隔和自己本地的间隔不一致，将对自己的运行时间间隔调整，并调整自己的定时器，与其保持一致。

- RouterA 平滑结束时将发出 VRRP 恢复报文，报文中携带着主备倒换前配置的时间间隔，此时 RouterB 上的 VRRP 会再进行一次时间间隔学习。

使用 VRRP 平滑倒换时的注意事项：

- 在平滑倒换的过程中，VRRP 的学习功能优先于抢占功能，即如果 Backup 状态的 VRRP 收到的协议报文里面的时间间隔和自己当前配置的不一致，并且报文中携带的优先级低于自己当前的配置优先级，这种情况 VRRP 首先考虑的是学习这个时间间隔并重置超时定时器，而后才会考虑是否抢占。
- VRRP 整机平滑倒换功能还依赖于系统本身，如果设备自身从主备倒换一开始系统便非常繁忙，无法调度 VRRP 模块运行的情况，VRRP 整机平滑倒换功能无效。

1.3.9 管理 VRRP(mVRRP)

mVRRP 是指管理 VRRP。管理 VRRP 备份组从本质上讲就是普通的 VRRP 备份组，它具备设备当前版本所实现的一切特征，唯一特殊之处在于：普通的 VRRP 备份组被配置为管理 VRRP 备份组之后，可以绑定其他的业务备份组，并根据绑定关系，决定相关业务备份组的状态。

普通 VRRP 备份组被加入管理 VRRP 备份组后，就不需要自己再发送 VRRP 协议报文来决定自己的状态。而管理 VRRP 通过发送 VRRP 协议报文来决定自己的状态，从而决定其绑定的所有业务 VRRP 的状态。这样就大大节约了 VRRP 协议报文占用带宽的问题。

一个管理 VRRP 备份组可以绑定多个业务 VRRP 备份组，但它不能作为业务备份组与其他管理备份组进行绑定。

在 VPLS 网络中，将 PW 或者业务口与管理 VRRP 进行绑定，可以实现 mVRRP 与 mVSI 联动应用。具体的应用场景可以参考后面组网应用中的“mVRRP”和“VRRP 在 ME 方案中的典型应用”。

1.3.10 VRRP6

VRRP6 是 VRRP for IPv6 的简称。它是一种容错协议，是对 VRRP 协议的扩展。它通过把几台路由设备联合组成一台虚拟的路由设备，并通过一定的机制保证当主机的下一跳设备出现故障时，及时将业务切换到其它设备，从而保持通讯的连续性和可靠性。

VRRP6 和 VRRP for IPv4 的区别

可以根据 VRRP6 和 VRRP for IPv4 所支持的功能，来查看它们的异同点。

VRRP6/VRRP for IPv4 支持的功能	VRRP6	VRRP for IPv4
主备备份 VRRP	是	是
负载分担 VRRP	是	是
监视接口状态	是	是
VRRP 快速切换	是	是
虚拟 IP 地址 Ping 开关	是	是
VRRP 的安全认证功能	否	是

VRRP6/VRRP for IPv4 支持的功能	VRRP6	VRRP for IPv4
VRRP 平滑倒换	否	是
管理 VRRP	是	是

基本原理

和 VRRP for IPv4 的实现原理相同，VRRP6 将局域网的一组设备构成一个备份组，相当于一个虚拟路由器。局域网内的主机只需要知道这个虚拟路由器的 IPv6 地址，并不需知道具体某台设备的 IPv6 地址，将网络内主机的缺省网关设置为该虚拟路由器的 IPv6 地址，主机就可以利用该虚拟网关与外部网络进行通信。

VRRP6 将该虚拟路由器动态关联到承担传输业务的物理设备上。当物理设备出现故障时，VRRP6 再次选择新设备来接替业务传输工作。整个过程对用户完全透明，实现了内部网络和外部网络不间断通信。

主机利用该虚拟 IPv6 地址与外部网络进行通信。设备工作机制如下：

1. 根据优先级的大小挑选 Master 设备。Master 设备的选举有两种方法：
   - 比较优先级的大小，优先级高者当选为 Master 设备。
   - 当两台优先级相同的设备同时竞争 Master 时，比较接口 IPv6 地址大小。接口 IPv6 地址大者当选为 Master 设备。
2. 其它设备作为备份设备，随时监听 Master 设备的状态。
   - 如果 Master 设备正常工作时，它会每隔一段时间发送一个 VRRP6 广播报文，以通知组内的备份设备，Master 设备处于正常工作状态。
   - 如果组内的备份设备一段时间内没有接收到来自 Master 设备发送来的 VRRP6 广播报文，则将自己转为主设备。一个 VRRP6 备份组里有多台备份设备时，短时间内可能产生多个 Master 设备。此时，设备将会将收到的 VRRP6 报文中的优先级与本地优先级做比较。从而选取优先级高的设备做 Master。

   说明
   - 对于 VRRP for IPv4，通过设置备份组内成员发送的 VRRP 通告报文时间间隔一致，可以避免同一备份组内出现多个 Master。
   - 对于 VRRP6，即使设置备份组内成员发送的 VRRP 通告报文的时间间隔不一致，同一备份组内也不会出现多个 Master。

主备备份 VRRP6 备份组

主备备份是 VRRP6 备份组工作的基本方式。它需要建立一个虚拟路由器，该虚拟路由器包含一台 Master 设备和若干台 Backup 设备。正常情况下，Master 设备承担全部业务；当 Master 设备出现故障时，Backup 设备接替 Master 设备工作，承担业务。

主备备份 VRRP6 备份组与主备备份 VRRP for IPv4 的处理机制相同。

负载分担 VRRP6 备份组

一台设备可以为多台设备作备份。负载分担方式是指建立两个或者更多个备份组，多台设备同时承担业务。

工作在负载分担下的 VRRP 备份组具有以下特点：

- 每个备份组都包括一个 Master 设备和若干 Backup 设备。
- 各个备份组的 Master 设备可以不同。
- 同一台设备可以加入多个备份组，在不同备份组中有不同的优先级。

负载分担 VRRP6 备份组与负载分担 VRRP for IPv4 的处理机制相同。

VRRP6 备份组监视接口状态

VRRP6 具备监视接口状态的功能。即不仅在备份组所在的接口出现故障时提供备份功能，而且在设备的某个其它接口出现故障时提供备份功能。

VRRP6 备份组监视接口状态与 VRRP for IPv4 备份组监视接口状态的处理机制相同。

VRRP6 备份组快速切换

BFD 会话能够快速检测网络中链路的故障。使用 VRRP6 监视 BFD 会话功能，在 BFD 会话状态改变后通知 VRRP 备份组，VRRP6 备份组将根据 BFD 会话的状态进行链路切换，从而实现快速切换的功能。

VRR6 备份组快速切换功能与 VRRP for IPv4 备份组快速切换功能的处理机制相同。

管理 VRRP6 备份组

管理 VRRP6 备份组从本质上讲就是普通的 VRRP6 备份组。它具有普通 VRRP6 备份组的一切特征。唯一的区别是：管理 VRRP6 备份组可以绑定其它的成员 VRRP6 备份组，并根据绑定关系，决定成员 VRRP6 备份组的状态。

一个管理 VRRP6 备份组可以绑定多个成员 VRRP6 备份组，但是它不可以作为成员 VRRP6 备份组和其它的管理 VRRP6 备份组进行绑定。

目前只支持成员 VRRP6 备份组与管理 VRRP6 备份组进行绑定。

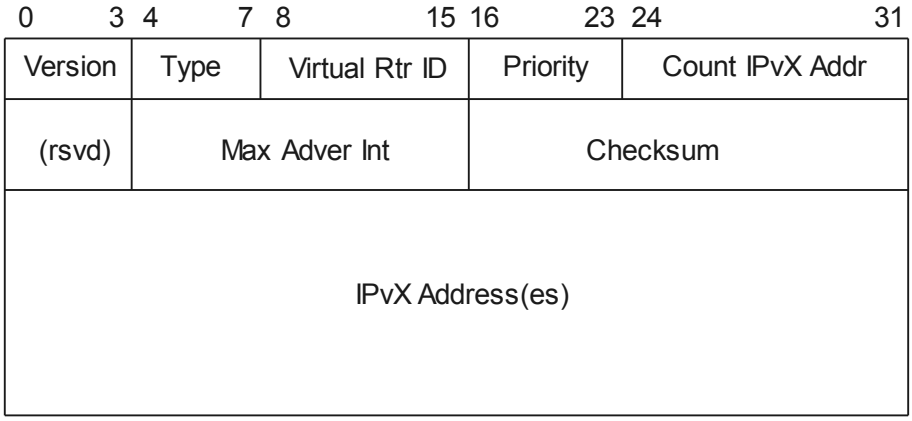
管理 VRRP6 备份组与管理 VRRP for IPv4 备份组的处理机制相同。

1.3.11 VRRPv3 的报文格式

目前，版本是 2 的 VRRP 协议不支持 IPv6 网路类型，而版本是 3 的 VRRP 协议支持 IPv4 和 IPv6 两种网络。

VRRP 协议既支持 VRRPv2 的报文，也支持 VRRPv3 的报文。VRRPv2 版本是由 RFC 3768 提出的，VRRPv3 版本是由 RFC5798 提出的。它们都是用来将 Master 设备的优先级和状态通告给同一备份组的其它设备。VRRPv3 报文的结构如图 1-7 所示。

图 1-7 VRRPv3 报文结构



各字段的含义如下：

- Version：VRRP 协议版本号。VRRPv3 报文只有一种取值：3。
- Type：VRRP 通告报文的类型。只有一种取值：1。
- Virtual Rtr ID：VRRP 备份组的 ID。取值范围是 1 ～ 255。
- Priority：发送 VRRP 通告报文的的设备在备份组中的优先级。取值范围是 0 ～ 255，但可用的范围是 1 ～ 255。0 表示设备停止参与 VRRP 备份组，用来使备份设备尽快成为 Master 设备，而不必等到计时器超时；255 则保留给 IP 地址拥有者。缺省值是 100。
- rsvd：保留字段，必须设置为 0。
- Count IP Addrs：VRRP 通告报文中包含的虚拟 IPv4 或虚拟 IPv6 地址的个数。
- Max Adver Int：发送 VRRP 通告报文的时间间隔。单位是厘秒。
- Checksum：校验和。
- IPvX Address(es)：VRRP 备份组的虚拟 IPv4 地址或者虚拟 IPv6 地址。

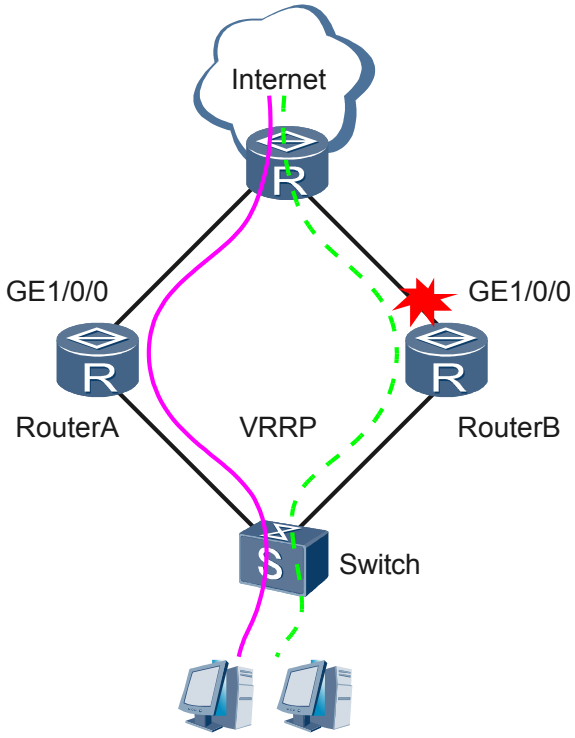
VRRPv2 版本和 VRRPv3 版本的主要区别：

- 认证功能不同。VRRPv3 不支持认证功能，而 VRRPv2 版本支持认证功能。
- 发送通告报文的时间间隔的单位不同。VRRPv3 支持的是厘秒级，而 VRRPv2 支持的是秒级。

1.4 应用

1.4.1 VRRP 监视接口状态

图 1-8 VRRP 监视接口的典型组网图



解决的问题：VRRP 无法感知非 VRRP 所在接口状态的变化，当上行链路出现故障时，VRRP 感知不到，从而导致业务中断。

配置说明如下：

- 通过配置 VRRP 监视指定的接口。
- VRRP 可以以 Increased 方式和 Reduced 方式来监视一个上下链路接口，一个 VRRP 最多可以监视 8 个接口。
- 当 VRRP 监视的接口的状态发生变化时，通知 VRRP，VRRP 根据接口的状态来增加或者是减少 VRRP 的优先级，从而达到指导 VRRP 状态切换的目的。

如图 1-8 所示，RouterA 和 RouterB 两台设备上面运行 VRRP 协议。并且 RouterB 的优先级比 RouterA 的优先级的高，RouterB 以 Reduced 方式监视接口。RouterB 为 Master 设备，用户侧的流量通过主用设备 RouterB 出去，如图 1-8 中虚线所示。现在 RouterB 连向 Internet 的出接口出现故障，由于 RouterB 上面 VRRP 以 Reduced 方式监视了这个接口，VRRP 的优先级降低，RouterA 抢占成为主用设备，以后用户侧的流量则通过 RouterA 出去。

1.4.2 VRRP 快速切换

VRRP 备份组发送广播报文的时间间隔为秒级或者毫秒级。缺省情况下，发送广播报文的时间间隔为秒级。

- 如果用户配置的发包间隔接近秒级或者为秒级，则需要通过监视 BFD 会话来实现 VRRP 的快速切换功能。
- 如果用户配置的发包间隔为毫秒级（不接近秒级，如 100 毫秒），则无需通过监视 BFD 会话来实现 VRRP 的快速切换功能，因为此 VRRP 备份组完全可以根据协议本身实现毫秒级的主备切换。

VRRP 监视 BFD

BFD 机制能够快速检测、监控网络中链路或者 IP 路由的连通状况，VRRP 通过监视 BFD 会话状态实现主备快速切换，主备切换的时间控制在 1 秒以内。

对于以下情况，BFD 都能够将检测到的故障通知接口板，从而加快 VRRP 主备倒换的速度。

- 备份组包含的接口出现故障。
- Master 和 Backup 不直接相连。
- Master 和 Backup 直接相连，但在中间链路上存在传输设备。

BFD 对 Backup 和 Master 之间的实际地址通信情况进行检测，如果通信不正常，Backup 就认为 Master 已经不可用，升级成 Master。在以下情况下 Backup 转换为 Master：

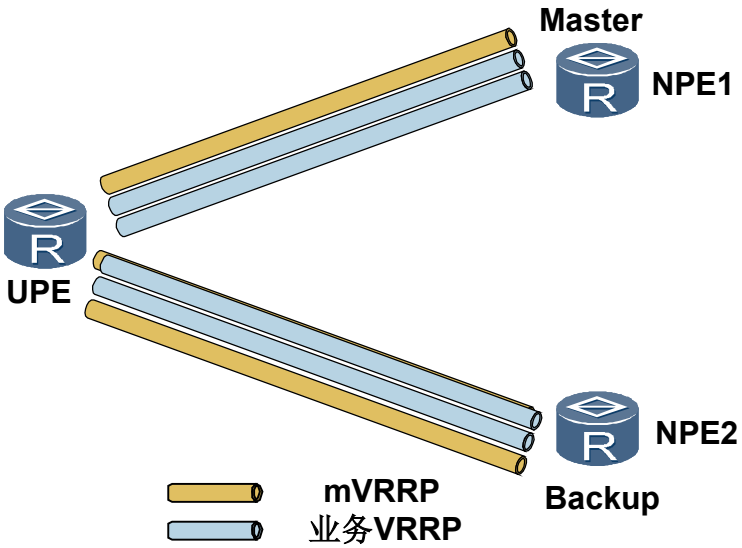
- 当两台设备之间的背靠背连接全部断开时，Backup 主动升级成 Master，承载上行流量。
- 当两台设备之间的连接在如下情况下中断时，Backup 主动升级成 Master，承载上行流量。
  - 当 Master 重新启动
  - Master 与交换机之间的链路断开
  - 与 Master 相连的交换机重新启动

VRRP 快速切换的环境要求：

- 在 Backup 上，BFD Session 检测的接口必须和 Master 设备相连；
- 在 Master 不可用时，Backup 的优先级增加并大于原来 Master 的优先级，促使自己快速切换为 Master。

1.4.3 mVRRP

图 1-9 mVRRP 的典型组网



解决的问题：大量 VRRP 协议报文浪费带宽，占用 CPU 的处理时间。

配置说明如下：

- 在 NPE1 和 NPE2 上配置管理 VRRP 备份组和普通 VRRP 备份组，然后将普通 VRRP 备份组与管理 VRRP 备份组绑定，此时被绑定的普通 VRRP 就称作业务 VRRP。
- UPE 不感知管理 VRRP 和业务 VRRP。

如图 1-9，当 NPE1 设备的一个 mVRRP 状态由 Master 变为 Backup 或者 Init 时，此 mVRRP 会通知它所绑定的所有业务 VRRP 与 mVRRP 的状态保持一致。而此时 NPE2 上对应的 mVRRP 的状态会由 Backup 变为 Master，此 mVRRP 同样会通知它所绑定的所有业务 VRRP 也把状态变为 Master。当 mVRRP 和业务 VRRP 的状态变为 Master 后，会发送免费 ARP 报文，从而将用户的流量引到新的 Master 端设备上来。

1.4.4 VRRP 在 ME 方案中的典型应用

图 1-10 UPE 双归到 NPE 组网图

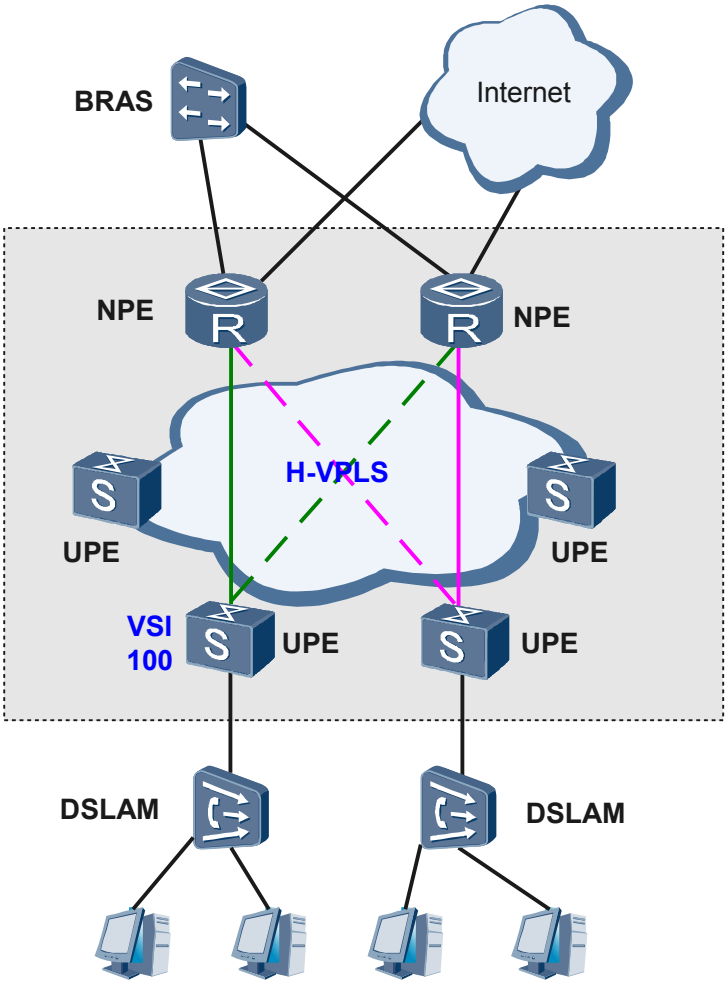
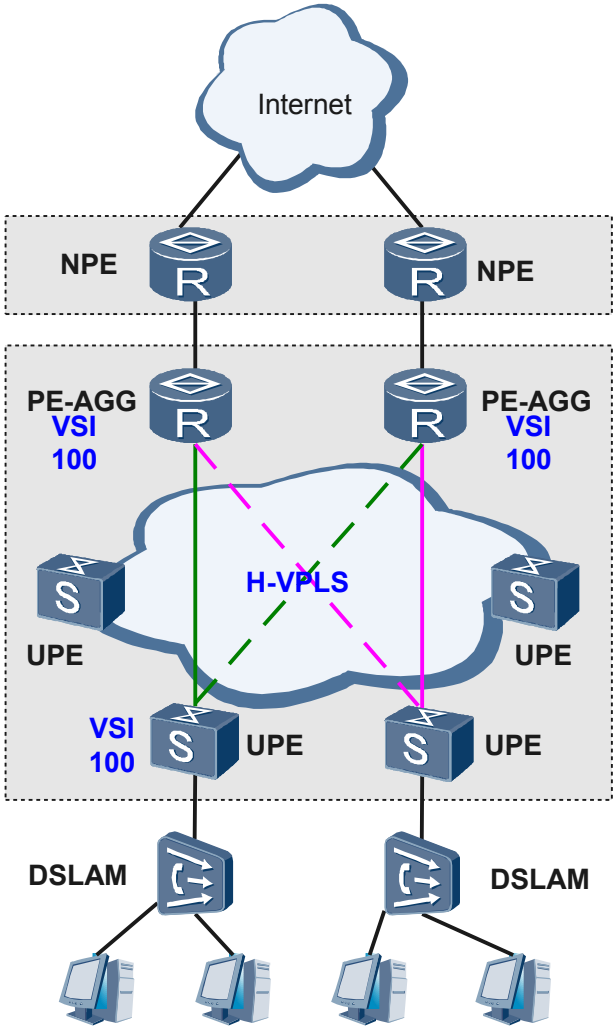


图 1-11 UPE 双归到 PE-AGG 组网图



问题一：如图 1-10 和图 1-11 所示，UPE 通过双归接入到 NPE（PE-AGG），如何实现 VRRP 的快速切换的问题。

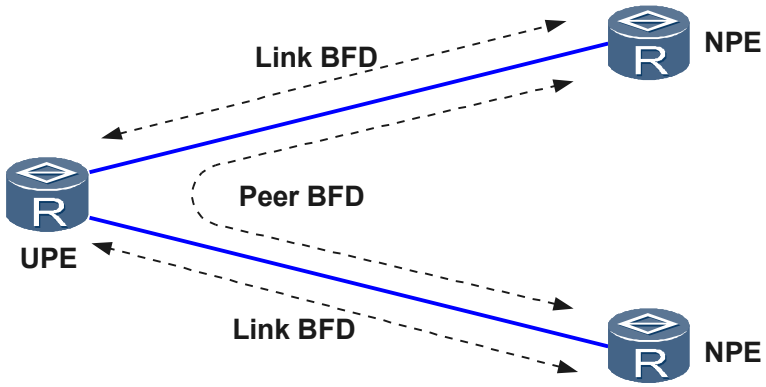
问题二：如图 1-10 所示，UPE 通过两条 PW 双归接入到 NPE，如何通过 VRRP 来保证两条 PW 的主备关系，以及如何达到主备两条 PW 的快速切换。

问题三：如图 1-11 所示，通过普通链路双归，如何实现普通链路的主备以及如何实现两条主备链路的快速切换（PE-AGG 组网环境）。

【解决问题一】：

通过配置 Peer BFD 会话和 Link BFD 会话来解决 UPE 通过双归接入到 NPE（PE-AGG），VRRP 的快速切换问题。

图 1-12 VRRP 监视 Peer BFD 和 Link BFD 典型应用



配置说明如下：

- 在两台 NPE 设备上配置 VRRP 备份组。
- 两台 NPE 之间配置 Peer BFD 会话。
- UPE 和两台 NPE 之间分别配置 Link BFD 会话。
- 配置 VRRP 同时监视 Link BFD 和 Peer BFD 会话。

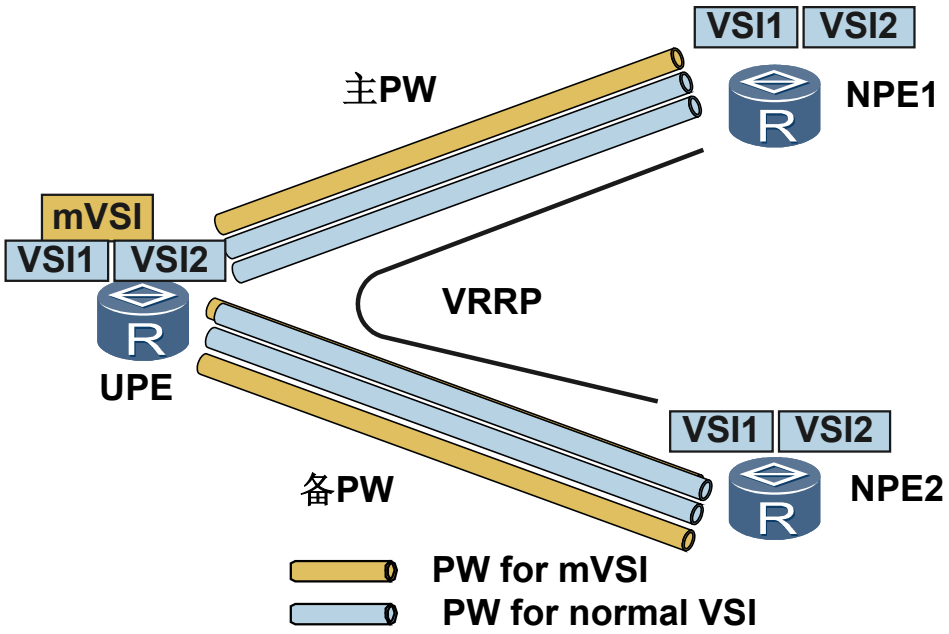
如图 1-12 所示，NPE 之间运行 VRRP 协议。NPE 之间运行的 BFD 叫做 Peer BFD，它检测 NPE 和 NPE 之间的链路和设备故障；NPE 与 UPE 之间运行的 BFD 叫做 Link BFD，它用来检测 NPE 和 UPE 之间的链路和设备故障。

Peer BFD 和 Link BFD 会话状态与普通的 BFD for VRRP 会话状态对 VRRP 备份组的影响不同：前两者直接影响备份组的状态，即直接设置备份组的状态；后者只是间接影响备份组的状态，即通过修改优先级来影响备份组的状态，但是优先级的修改并不一定会导致备份组状态的变化。mVRRP 通过监视 Peer BFD 和 Link BFD 的状态，可以更快的实现主备倒换，并感知故障发生的位置。

【解决问题二】：

通过配置 PW 绑定管理 VRRP，解决 UPE 通过两条 PW 双归接入到 NPE，可以通过 VRRP 来保证两条 PW 的主备关系，以及达到主备两条 PW 的快速切换（NPE 合一组网环境）。

图 1-13 PW 监视 VRRP 的典型应用



如图 1-13 所示，PW 与管理 VRRP 备份组绑定，如果 UPE 与 NPE 之间运行 VLL、PWE3 或者 VPLS，UPE 通过两条 PW 双归属到 NPE，此时可以配置 PW 与管理 VRRP 绑定来决定主用 PW 和备用 PW。

典型的应用：mVRRP over mVPLS，指 mVRRP 的报文通过 mVSI 以及 mPW 来交互。

配置说明如下：

- 在两台 NPE 设备上配置 VRRP，设置 VRRP 为管理 VRRP
- NPE 设备上的 PW 分别监视管理 VRRP
- UPE 上都配置 mVSI 和业务 VSI，并且 mVSI 与业务 VSI 绑定

NPE 与 UPE 之间运行 mVPLS，NPE 之间运行 mVRRP。mVRRP 的报文通过 NPE 与 UPE 之间的管理 PW 来传送，并通过 mVSI 转发。其他业务报文通过 UPE 和 NPE 之间的业务 PW 和业务 VSI 来传送。管理 VRRP 的报文和其他业务报文通过不同的 PW 传送，相互隔离。为了使 NPE 之间的管理 VRRP 能够快速切换，NPE 之间需要配置 Peer BFD。Peer BFD 报文也通过管理 PW 传送，并通过管理 VSI 交互。

当 NPE 的 VRRP 备份组发生主备切换时：

1. UPE 上的 mVSI 通过 NPE 与 UPE 之间的管理 PW 会收到 NPE 发送来的免费 ARP 报文
2. mVSI 判断本次收到的免费 ARP 报文和前一次收到的免费 ARP 报文是否有变化（是否来自同一个 PW，IP、入标签、入接口以及 MAC 是否相同）
   - 如果没有变化，则说明 NPE 之间的管理 VRRP 没有发生主备切换。
   - 如果有变化，则说明 NPE 之间的管理 VRRP 发生了主备切换。
3. UPE 根据配置的 mVSI 和业务 VSI 的关系，清除关联的所有业务 VSI 的 MAC

此外，由于 mVSI 用于传输和截获免费 ARP 报文和 BFD 报文，不同于普通 VSI，所以不允许用户进行 shutdown 操作。

PW 状态和管理 VRRP 状态的关联关系：

业务 VSI 的 MAC 清除后，当收到去往新的主 NPE 的报文，由于是未知帧，将进行广播，通过重新学习 MAC 切换到新的主用 NPE。

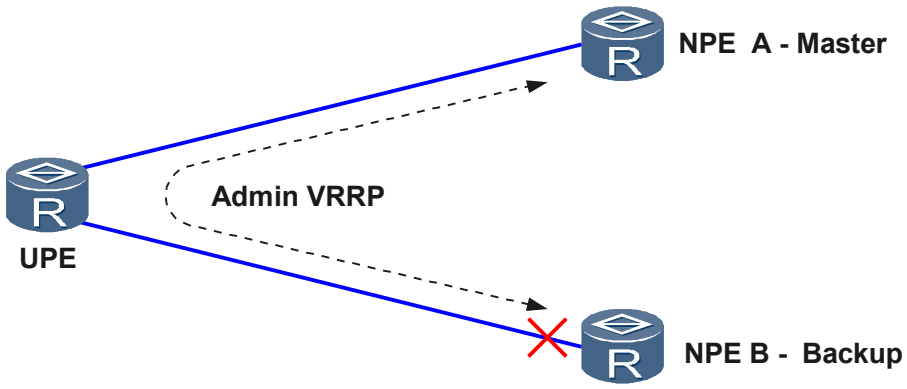
- 如果 PW 原先的状态为 Down，则 PW 与管理 VRRP 备份组绑定之后，PW 的状态仍为 Down。
- 如果 PW 原先的状态为 UP，则 PW 与管理 VRRP 备份组绑定之后，如果管理 VRRP 的状态为 Master，那么 PW 状态仍为 Up；如果管理 VRRP 的状态为 Backup 或 Initialize，那么 PW 的状态会变为 Backup。

此种场景下，即可以配置业务 VRRP 与 mVRRP 绑定，也可以不绑定，因为 VRRP 只是起网关的作用，属于三层业务，而 PW 跑的都是二层业务，它们之间没有必然联系。

【解决问题三】：

通过配置业务接口绑定管理 VRRP，实现普通链路的主备以及可以实现两条主备链路的快速切换（PE-AGG 组网环境）。

图 1-14 业务口 Track VRRP 的典型应用



如图 1-14 所示，业务接口与管理 VRRP 备份组绑定，如果 UPE 通过两条物理链路双归属到 NPE，可以配置业务接口与管理 VRRP 绑定来决定主用业务接口和备用业务接口。

配置说明如下：

- 在两台 NPE 设备上配置 VRRP，设置 VRRP 为管理 VRRP
- 将 NPE 设备上的两条主备链路分别 Track 管理 VRRP
- 根据 VRRP 的状态来决定链路是否转发流量，如果接口 Track 的管理 VRRP 为 Master，则转发流量，如果接口 Track 的管理 VRRP 为 Backup 或者 Initialize，则接口不转发流量
- VRRP 状态的切换引起接口状态的切换，实现链路主备的切换
- 倒换时清除 UPE 设备上的 MAC 地址，从而使流量可以广播到备用链路，减少流量丢失
- NPE 主备状态切换，包括回切，主用设备的接口 MAC 发生了变化，需要及时反映到用户设备的 ARP 表中，并且 UPE 上也需要及时生成 MAC 转发表项，减少广播流量

业务接口状态和管理 VRRP 状态的关联关系：

- 当业务接口所绑定的管理 VRRP 备份组状态变为 Master 状态：管理 VRRP 备份组通知所有关联的业务接口，如果接口运行三层业务，则将接口置为 UP 状态，生成网段路由，转发平面根据接口状态开启双向流量转发；如果接口运行二层业务，则直接将接口置为 UP 状态，开启双向流量转发。
- 当业务接口所绑定的管理 VRRP 备份组状态变为 Initialize 或者 Backup 状态：管理 VRRP 备份组通知所有关联的业务接口，如果接口运行的是三层业务，则将接口置为 Down 状态，取消网段路由，转发平面根据接口状态关闭双向流量转发；如果接口运行的是二层业务，则直接将接口置为 Down 状态，关闭双向流量转发。

业务口和普通接口没有区别，只是由于绑定了管理 VRRP，所以才叫业务口。

1.5 术语与缩略语

缩略语

缩略语	英文全称	中文全称
VRRP	Virtual Router Redundancy Protocol	虚拟路由冗余协议
ARP	Address Resolution Protocol	地址解析协议
BFD	Bidirectional Forwarding Detection	双向转发检测
L2VPN	Layer 2 virtual private network	二层虚拟专用网
PW	Pseudo Wire	虚电线
VSI	Virtual Switching Instance	虚拟交换实例
QinQ	802.1Q in 802.1Q	802.1Q 嵌套 802.1Q
ME	Metro Ethernet	城域以太
mVRRP	Manage Virtual Router Redundancy Protocol	管理 VRRP
mVPLS	Manage Virtual Private LAN Service	管理 VPLS
mVSI	Manage Virtual Switching Instance	管理 VSI

2 BFD

关于本章

2.1 介绍

定义

双向转发检测 BFD（Bidirectional Forwarding Detection）用于快速检测系统之间的通信故障，并在出现故障时通知上层应用。

目的

为了减小设备故障对业务的影响，提高网络的可用性，网络设备需要能够尽快检测到与相邻设备间的通信故障，以便及时采取措施，保证业务继续进行。

现有的故障检测方法主要包括：

- 硬件检测：例如通过 SDH（Synchronous Digital Hierarchy，同步数字体系）告警检测链路故障。硬件检测的优点是可以很快发现故障，但并不是所有介质都能提供硬件检测。
- 慢 Hello 机制：通常是指路由协议的 Hello 机制。这种机制检测到故障所需时间为秒级。对于高速数据传输，例如吉比特速率级，超过 1 秒的检测时间将导致大量数据丢失；对于时延敏感的业务，例如语音业务，超过 1 秒的延迟也是不能接受的。
- 其他检测机制：不同的协议或设备制造商有时会提供专用的检测机制，但在系统间互联互通时，这样的专用检测机制通常难以部署。

BFD 就是为解决现有检测机制的不足而产生的。

BFD 的目标如下：

- 对相邻转发引擎之间的通道提供轻负荷、快速故障检测。这些故障包括接口、数据链路，甚至有可能是转发引擎本身。
- 提供一种单一的机制，能够用来对任何媒介、任何协议层进行实时地检测，并且检测的时间与开销范围比较宽。

2.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC5880	Bidirectional Forwarding Detection	-
RFC5882	Generic Application of BFD	-
RFC5883	BFD for Multihop Paths	-
RFC5881	BFD for IPv4 and IPv6 (Single Hop)	-
RFC5884	BFD for mpls	-

2.3 原理描述

BFD 用于检测转发引擎之间的通信故障。具体来说，BFD 对系统间的、同一路径上的一种数据协议的连通性进行检测，这条路径可以是物理链路或逻辑链路，包括隧道。

可以把 BFD 看作是系统提供的一种服务：

- 上层应用向 BFD 提供检测地址、检测时间等参数。
- BFD 根据这些信息创建、删除或修改 BFD 会话，并把会话状态通告给上层应用。

BFD 具有以下特点：

- 对相邻转发引擎之间的路径提供轻负荷、短持续时间的检测。
- 采用单一机制对所有类型的介质、协议层进行检测，实现全网统一的检测机制。

下面从 BFD 检测机制、检测的链路类型、会话建立方式以及会话管理来介绍 BFD 的基本原理。

BFD 检测机制

BFD 的检测机制是两个系统建立 BFD 会话，并沿它们之间的路径周期性发送 BFD 控制报文，如果一方在既定的时间内没有收到 BFD 控制报文，则认为路径上发生了故障。

BFD 控制报文封装在 UDP 报文中传送。会话开始阶段，双方系统通过控制报文中携带的参数（会话标识符、期望的收发报文最小时间间隔、本端 BFD 会话状态等）进行协商。协商成功后，以协商的报文收发时间在彼此之间的路径上定时发送 BFD 控制报文。

为满足快速检测的需求，BFD 草案规定发送间隔和接收间隔单位是微秒。但限于目前的设备处理能力，大部分厂商的设备配置 BFD 时只能达到毫秒级，在进行内部处理时再转换到微秒。NE80E/40E 支持的最小检测时间为 30 毫秒。

BFD 提供两种检测模式：

- 异步模式：BFD 的主要操作模式称为异步模式。在这种模式下，系统之间相互周期性地发送 BFD 控制报文，如果某个系统连续几个报文都没有接收到，就认为此 BFD 会话的状态是 Down。
- 查询模式：BFD 的第二种操作模式称为查询模式。当一个系统中存在大量 BFD 会话时，为防止周期性发送 BFD 控制报文的开销影响到系统的正常运行，可以采用查询模式。在查询模式下，一旦 BFD 会话建立，系统就不再周期性发送 BFD 控制报文，而是通过其他与 BFD 无关的机制检测连通性（比如路由协议的 Hello 机制、硬件检测机制等），从而减少 BFD 会话带来的开销。

两种模式的一个辅助功能是回声功能。当回声功能激活时，一个 BFD 控制报文按照如下方式发送：本地发送一个 BFD 控制报文，远端系统通过它的转发通道将它们环回回来。如果连续几个回声包都没有接收到，会话状态就被宣布为“Down”。回声功能可以与异步模式或者查询模式一起使用。

目前系统只支持被动回声功能。

BFD 检测的链路类型

- IP 链路

  在 NE80E/40E 中，BFD 支持检测的 IP 链路如下，包括单跳检测和多跳检测。

- 三层物理接口
- 以太子接口（包括 Eth-Trunk 子接口）

对于一个物理以太网接口有多个子接口的情况，BFD 会话可以独立建立在各个子接口上和此物理以太网接口上。

- IP-Trunk
  - IP-Trunk 链路
  - IP-Trunk 成员链路

  检测 Trunk 成员口与检测 Trunk 口的 BFD 会话互相独立，可同时检测。
- Eth-Trunk
  - 二层 Eth-Trunk 链路
  - 二层 Eth-Trunk 成员链路
  - 三层 Eth-Trunk 链路
  - 三层 Eth-Trunk 成员链路

  检测 Trunk 成员口与检测 Trunk 口的 BFD 会话互相独立，可同时检测。
- VLANIF
  - VLAN 以太成员链路
  - VLAN 以太子接口
  - VLANIF 接口

  检测 VLANIF 与检测 vlan 成员口的 BFD 会话相互相独立，可同时检测。
- MPLS LSP

  检测 MPLS LSP 的连通性时，BFD 会话协商有两种方式：
  - 静态配置 BFD：通过手工配置 BFD 的本地标识符和远端标识符，由 BFD 本身的协商机制建立会话。
  - 动态创建 BFD 会话：通过在 LSP Ping 报文中携带 BFD Discriminator TLV 进行会话协商。

  静态配置方式下，BFD 能够检测的 LSP 类型有：
  - 静态 LSP
  - LDP LSP
  - TE：Tunnel、与 Tunnel 绑定的静态 CR-LSP 和 RSVP CR-LSP。

  BFD 能够检测信令协议为 CR-Static 和 RSVP-TE 的 TE 隧道，并且能够检测与 TE 隧道绑定的主用 LSP。

  动态配置方式下，BFD 检测的 LSP 类型有:
  - LDP LSP
  - TE：与 Tunnel 绑定的静态 CR-LSP 和 RSVP CR-LSP

  动态 BFD 不支持检测整条 TE 隧道。
- PW

  BFD 检测 PW 也分为静态方式（手工配置标识符）和动态方式。

  BFD 能够检测的 PW 类型有：
  - 单跳 PW
  - 多跳 PW
  - BGP PW

BFD 会话建立方式

BFD 会话的建立有两种方式，即静态配置 BFD 会话和动态建立 BFD 会话。

BFD 通过控制报文中的 My Discriminator 和 Your Discriminator 区分不同的会话。静态和动态创建 BFD 会话的主要区别在于 My Discriminator 和 Your Discriminator 的配置方式不同。

- 静态配置 BFD 会话

  静态配置 BFD 会话是指通过命令行手工配置 BFD 会话参数，包括了配置本地标识符和远端标识符等，然后手工下发 BFD 会话建立请求。

- 动态建立 BFD 会话

  动态建立 BFD 会话时，系统对本地标识符和远端标识符的处理方式如下：

  - 动态分配本地标识符

    当应用程序触发动态创建 BFD 会话时，系统分配属于动态会话标识符区域的值作为 BFD 会话的本地标识符。然后向对端发送 Your Discriminator 的值为 0 的 BFD 控制报文，进行会话协商。

  - 自学习远端标识符

    当 BFD 会话的一端收到 Your Discriminator 的值为 0 的 BFD 控制报文时，判断该报文是否与本地 BFD 会话匹配，如果匹配，则学习接收到的 BFD 报文中 My Discriminator 的值，获取远端标识符。

BFD 会话管理

BFD 会话有四种状态：Down、Init、Up 和 AdminDown。

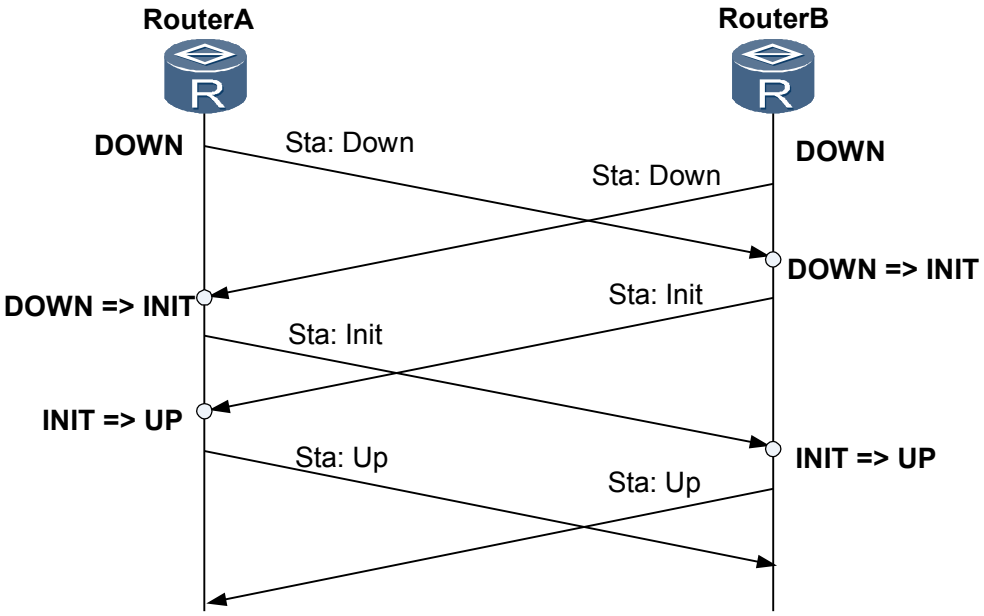
- Down：会话处于 Down 状态或刚刚创建。
- Init：已经能够与对端系统通信，本端希望使会话进入 Up 状态。
- Up：会话已经建立成功。
- AdminDown：会话处于管理性 Down 状态。

会话状态变化通过 BFD 报文的 State 字段传递，系统根据自己本地的会话状态和接收到的对端 BFD 报文驱动状态改变。

BFD 状态机的建立和拆除都采用三次握手机制，以确保两端系统都能知道状态的变化。

以 BFD 会话建立为例，简单介绍状态机的迁移过程。

图 2-1 BFD 会话连接建立



1. RouterA 和 RouterB 各自启动 BFD 状态机，初始状态为 Down，发送状态为 Down 的 BFD 报文。对于静态配置 BFD 会话，报文中的 Your Discriminator 的值是用户指定的；对于动态创建 BFD 会话，Your Discriminator 的值是 0。
2. RouterB 收到状态为 Down 的 BFD 报文后，状态切换至 Init，并发送状态为 Init 的 BFD 报文。
3. RouterB 本地 BFD 状态为 Init 后，不再处理接收到的状态为 Down 的报文。
4. RouterA 的 BFD 状态变化同 RouterB。
5. RouterB 收到状态为 Init 的 BFD 报文后，本地状态切换至 Up。
6. RouterA 的 BFD 状态变化同 RouterB。

2.3.1 BFD for IP

在 IP 链路上建立 BFD 会话，利用 BFD 检测机制快速检测故障。

BFD for IP 支持单跳检测和多跳检测：

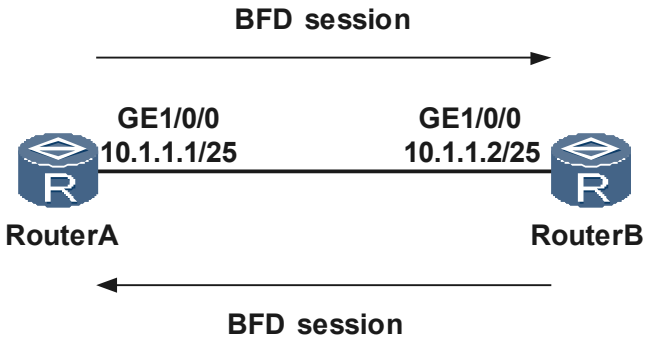
- BFD 单跳检测是指对两个直连系统进行 IP 连通性检测，这里所说的“单跳”是 IP 的一跳。在进行 BFD 单跳检测的两个系统中，对于一种给定的数据协议，在指定接口上只存在一个 BFD 会话。
- BFD 多跳检测是指 BFD 可以检测两个系统间的任意路径，这些路径可能跨越很多跳，也可能在某些部分发生重叠。

组网应用

典型应用一：

如图 2-2 所示，BFD 检测两台设备之间的单跳路径，BFD 会话绑定出接口。

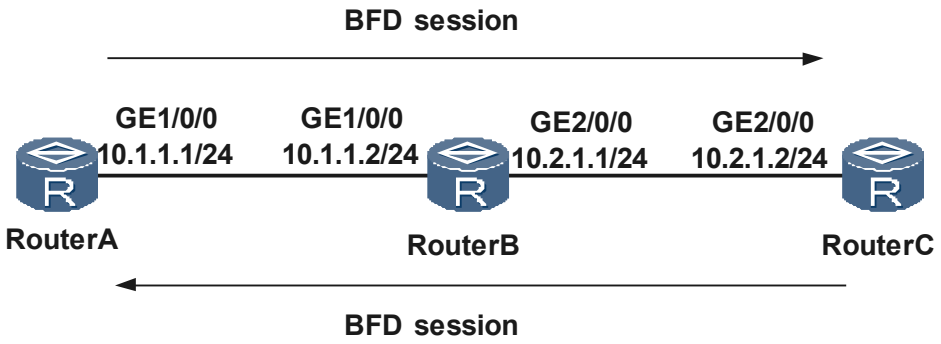
图 2-2 单跳 BFD for IP



典型应用二：

如图 2-3 所示，BFD 检测 RouterA 和 RouterC 之间的多跳路径，BFD 会话绑定对端 IP 但不绑定出接口。

图 2-3 多跳 BFD for IP



2.3.2 组播 BFD

组播 BFD 用于检测无 IP 地址等三层属性的接口之间的链路连通性，达到链路故障快速检测。

通过将检测报文通过 IP 层发送组播检测报文，在所需检测链路之间的路由器配置组播检测。本端发送组播报文，如果链路连通，则对端接口也可以收到这个组播报文，上送对端 BFD 应用，感知链路正常。对于二层 Trunk 链路，由于发送的是组播报文，IP 层转发不需要三层属性，直接下发链路层发送，快速检测链路的连通性。这里的 IP 是 BFD 模块配置的公认的组播地址 Default-IP，任何收到此 IP 的接口都将此报文上送 BFD 应用，完成 IP 转发。

组网应用

图 2-4 组播 BFD 组网示意图

如图 2-4 所示，组播 BFD 可以快速检测接口之间的链路连通性。在 RouterA、RouterB 上配置 BFD 会话，使用缺省组播地址对绑定 GE1/0/0 接口的单跳链路进行检测，这样就能快速检测接口之间的链路连通性。

2.3.3 BFD for PIS

BFD for PIS（Process interface status）提供一种简单的机制，使得 BFD 检测行为可以关联接口状态，提高了接口感应链路故障的灵敏度，减少了非直连链路故障导致的问题。

BFD 的 PIS 机制，对检测到链路故障的 BFD 会话，会立即上报 Down 消息到相应接口，使得接口进入一种特殊的 Down 状态：BFD Down 状态，该状态等效于链路协议 Down 状态，在该状态下只有 BFD 的报文可以正常处理，从而使接口也可以快速感知链路故障。

对于每个要配置接口联动的 BFD 会话，配置为组播检测并指定接口方式，从而避开对接口 IP 属性的依赖性。

组网应用

图 2-5 BFD for PIS 组网示意图

如图 2-5 所示，在 RouterA 和 RouterB 上配置 BFD 会话，使用缺省组播地址对绑定 GE1/0/0 接口的单跳链路进行检测，配置接口联动后，当 BFD 检测到链路出现故障，立即上报 Down 消息到相应接口，使接口进入 BFD Down 状态。

2.3.4 BFD for TTL

BFD 后续多跳草案（draft-ietf-bfd-multihop-04）规定，对于单跳会话继续使用 3784 作为目的端口号，新增加了关于多跳会话的端口号的使用限制，即多跳会话使用 4784 作为目的端口号。为了与最新的 BFD 多跳草案保持一致，则在 BFD for TTL 特性中增加了对于多跳 BFD 报文端口号的支持，使用 4784 作为多跳 BFD 报文的目的端口号；同时又保证与以前老版本设备互通时，对于以前的老版本的实现并不区分 BFD 单跳、多跳会话，统一使用 3784 作为 BFD 报文端口号，此时需要根据接收报文中携带的 TTL 值区分单跳会话和多跳会话。

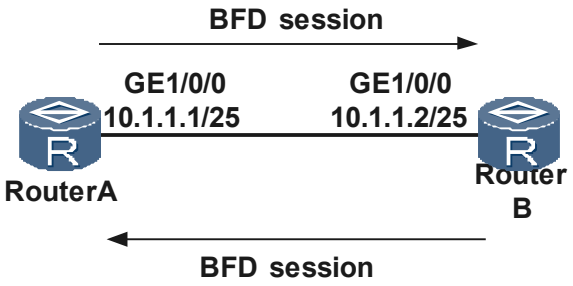
BFD 控制报文封装在 UDP 报文中传送，源端口号的取值范围是 49152 ～ 65535，目的端口号的取值范围是 3784 或 4784。BFD 草案规定，多跳 BFD 报文的目的端口号是 4784。

组网应用

典型应用一：

如图 2-6 所示，BFD 检测两台设备之间的单跳路径，BFD 会话绑定出接口。

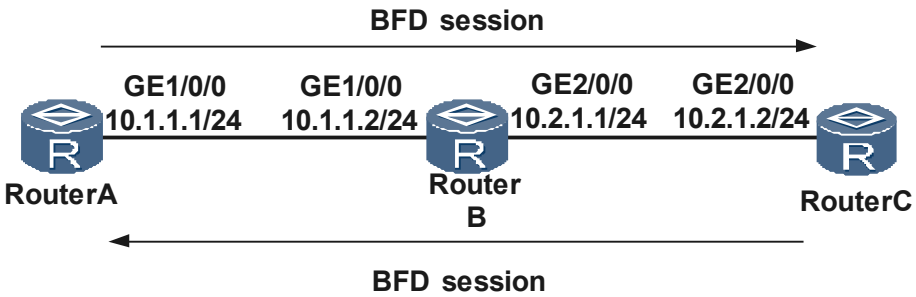
图 2-6 单跳 BFD for IP



典型应用二：

如图 2-7 所示，BFD 检测 RouterA 和 RouterC 之间的多跳路径，BFD 会话绑定对端 IP 但不绑定出接口。

图 2-7 多跳 BFD for IP



2.3.5 单臂 ECHO 功能

ECHO 功能是指通过 BFD 报文的环回操作检测转发链路的连通性。

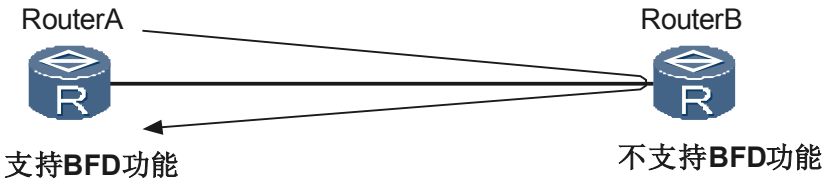
单臂 ECHO 功能：在两台直接相连的设备中，其中，一台设备支持 BFD 功能，另一台设备不支持 BFD 功能，只支持基本的网路层转发。为了能够快速的检测这两台设备之间的故障，可以在支持 BFD 功能的设备上创建单臂 ECHO 功能的 BFD 会话。支持 BFD 功能的设备主动发起 ECHO 请求功能，不支持 BFD 功能的设备接收到该报文后直接将其环回，从而实现转发链路的连通性检测功能。

说明

单臂 ECHO 功能只适用于单跳 BFD 会话中。

组网应用

图 2-8 单臂 ECHO 功能组网示意图



如图 2-8 所示，RouterA 支持 BFD 功能，RouterB 不支持 BFD 功能。在 RouterA 上配置单臂 ECHO 功能的 BFD 会话，检测 RouterA 到 RouterB 之间的单跳路径。RouterB 接收到 RouterA 发送的 BFD 报文后，直接在网络层将该报文环回，从而快速检测 RouterA 和 RouterB 之间的直连链路的连通性。

2.4 应用

2.4.1 BFD for USR

BFD for USR（Unicast Static Route）用于支持 IPv4 单播静态路由，支持 IPv4 单播静态路由绑定后快速感知链路状态。

与动态路由协议不同，单播静态路由自身没有检测机制，当网络发生故障的时候，需要管理员介入。BFD for USR 特性可为公网 IPv4 单播静态路由绑定 BFD 会话，利用 BFD 会话来检测单播静态路由所在链路的状态。

BFD for USR 可为每条 IPv4 单播静态路由绑定一个 BFD 会话，当这条 USR 上绑定的 BFD 会话检测到链路故障（由 Up 转为 Down）后，BFD 会将故障上报路由管理系统，由路由管理模块将这条路由设置为“非激活”状态（此条路由不可用，从 IP 路由表中删除）。

当这条 USR 上绑定的 BFD 会话成功建立或者从故障状态恢复后（由 Down 转为 Up），BFD 会上报路由管理模块，由路由管理模块将这条路由设置为“激活”状态（此路由可用，加入 IP 路由表）。

2.4.2 BFD for OSPF

网络上的链路故障或拓扑变化都会导致路由器重新进行路由计算，要提高网络的可用性，缩短路由协议的收敛时间非常重要。由于链路故障无法完全避免，因此，加快故障感知速度并将故障快速通告给路由协议是一种可行的方案。

BFD for OSPF 就是将 BFD 和 OSPF 协议关联起来，通过 BFD 对链路故障的快速感应进而通知 OSPF 协议，从而加快 OSPF 协议对于网络拓扑变化的响应。

说明

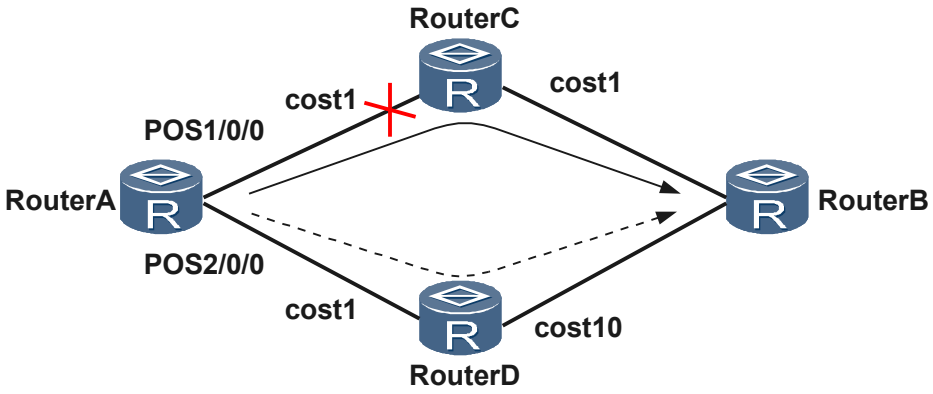
在与不同厂商设备互通的网络中，如果对端设备配置为 BFD 的单跳会话，建议用户在本端设备上也配置单跳会话。

表 2-1 显示了 OSPF 协议在有、无 BFD 协议下收敛速度的数据。

表 2-1 OSPF 协议收敛速度的数据

有无 BFD	链路故障检测机制	收敛速度
无 BFD	OSPF HELLO keepalive 定时器超时	秒级
有 BFD	BFD 会话 Down	毫秒级

图 2-9 BFD for OSPF 组网图



如图 2-9 所示，RouterA 分别与 RouterC 和 RouterD 建立 OSPF 邻居关系，RouterA 到 RouterB 的路由出接口为 POS1/0/0，经过 RouterC 到达 RouterB。邻居状态到达 FULL 状态时通知 BFD 建立 BFD 会话。

1. 当 RouterA 和 RouterC 之间链路出现故障，BFD 首先感知到并通知 RouterA。
2. RouterA 处理邻居 Down 事件，重新进行路由计算，新的路由出接口为 POS2/0/0，经过 RouterD 到达 RouterB。

2.4.3 BFD for IS-IS

通常情况下，ISIS 设定发送 Hello 报文的时间间隔为 10 秒钟，宣告邻居 Down 的时间即相邻设备失效的时间一般配置为 Hello 报文间隔的 3 倍。若在相邻设备失效时间内没有

收到邻居发来的 Hello 报文，将会删除邻居。设备能感知到邻居故障的时间最小也是秒级。在高速的网络环境中，这将导致报文大量丢失。

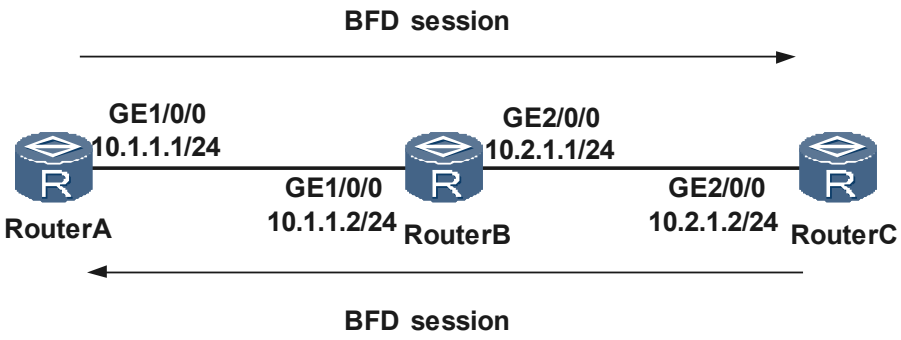
BFD for ISIS 是指 BFD 会话由 ISIS 协议动态创建，不再依靠手工配置，当 BFD 检测到故障时，通过路由管理通知 ISIS 协议，由协议进行相应邻居 Down 处理，快速更新 LSP 信息和进行增量路由计算，从而实现 ISIS 路由的快速收敛。

通过配置 BFD 可以设置毫秒级的时间检测间隔。使用 BFD 并不是代替 ISIS 协议本身的 Hello 机制，而是配合 ISIS 协议更快的发现邻接方面出现的故障，并及时通知 ISIS 重新计算相关路由以便正确指导报文的转发。

路由管理模块 RM（Routing Management Module）为 ISIS 提供与 BFD 模块交互的相关服务。ISIS 通过 RM 通知 BFD 来动态创建或删除 BFD 会话，同时 BFD 的事件消息也通过 RM 传递给 ISIS。

组网应用

图 2-10 BFD for ISIS 组网示意图



在 RouterA、RouterB 和 RouterC 上使能 BFD 后，当 RouterA 和 RouterB 之间的链路故障时，BFD 快速检测到故障并通过 RM 模块通告给 ISIS 协议，ISIS 将此接口关联的邻居的状态置为 Down，从而触发 ISIS 拓扑计算，同时更新 LSP 使得其他邻居（如 RouterB 的邻居 RouterC）及时收到 RouterB 的更新 LSP，最终实现了网络拓扑的快速收敛。

2.4.4 BFD for VRRP

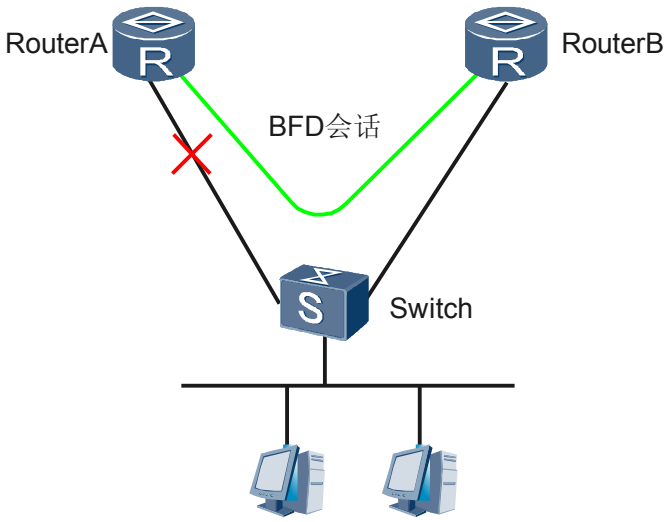
对于以下情况，BFD 都能够将检测到的故障通知接口板，从而加快 VRRP 主备倒换的速度。

- 备份组包含的接口出现故障
- Master 和 Backup 不直接相连
- Master 和 Backup 直接相连，但在中间链路上存在传输设备。根据中间链路的传输设备是否支持 BFD 特性，可从以下两方面进行介绍。
  - 如果此传输设备是交换机或者是其它支持 BFD 特性的设备，可配置两种类型的 BFD 会话来检测链路，一种是 Peer BFD 类型的会话，用于检测 Master 和 Backup 之间的链路；另一种是 Link BFD 类型的会话，用于检测传输设备和 Master 或者是 Backup 之间的链路。
  - 如果此传输设备不支持 BFD 特性：

- 可以在 Master 端和 Backup 端配置 Peer BFD 会话的会话来检测链路。当 BFD 会话的状态变为 Down 时，VRRP 备份组会将 Backup 上升为 Master，在此情况下，有可能会出现双 Master 情况。此时，双 Master 端会发送免费 ARP 报文，从而系统最终能够判断出哪条链路发生了故障。

BFD 对 Backup 和 Master 之间的实际地址通信情况进行检测，如果通信不正常，Backup 就认为 Master 已经不可用，升级成 Master。VRRP 通过监视 BFD 会话状态实现主备快速切换，切换时间控制在 50 毫秒以内。

图 2-11 VRRP Track BFD 典型组网



如图 2-11 所示，RouterA 和 RouterB 之间配置 VRRP 备份组建立主备关系，RouterA 为主用设备，RouterB 为备用设备，用户过来的流量从 RouterA 出去。在 RouterA 和 RouterB 之间建立 BFD 会话，VRRP 备份组监视该 BFD 会话，当 BFD 会话状态变化时，通过修改备份组优先级实现主备快速切换。

当 BFD 检测到 RouterA 和 Switch 之间的链路故障时，上报给 VRRP 一个 BFD 检测 Down 事件，RouterB 上 VRRP 备份组的优先级增加，增加后的优先级大于 RouterA 上的 VRRP 备份组的优先级，于是 RouterB 立刻升为 Master，后继的用户流量就会通过 RouterB 转发，从而实现 VRRP 的主备快速切换。

2.4.5 BFD for PIM

正常情况下，如果共享网段上的当前 DR（Designate Router）出现故障，其他 PIM 邻居会等到邻居关系超时才触发新一轮的 DR 竞选过程，组播数据传输中断的时间将会不小于邻居关系的超时时间，通常是秒级。

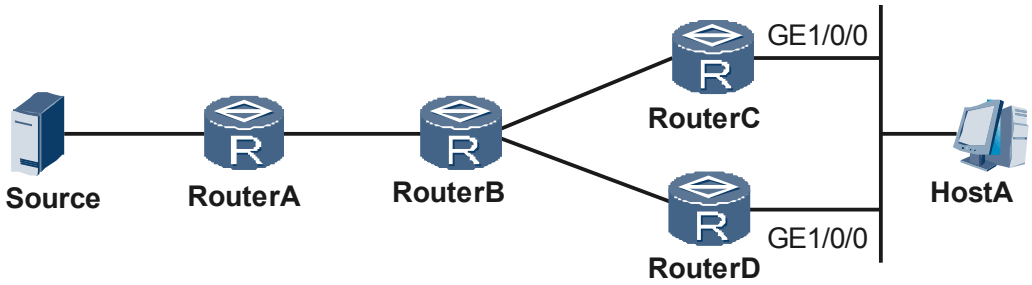
BFD for PIM 的特点是可以进行快速故障检测，能够实现在毫秒级别内通知 PIM 模块触发新一轮的 DR 竞选，而不是等到邻居关系超时。这在很大程度上缩短了组播数据传输的中断时间，提高了组播网络的可靠性。

BFD for PIM 同时也适用于共享网段上 Assert 竞选的过程，可以快速响应 Assert Winner 接口故障。

组网应用

BFD for PIM 适用于 NBMA 接口和 Broadcast 接口组成的共享网段上的链路故障快速检测。

图 2-12 BFD for PIM 组网示意图



如图 2-12 所示，主机 HostA 需要 Source 的数据流，在正确配置组播 PIM 协议后，数据流传输如下：

- 在 RouterC 和 RouterD 没有配置 DR 优先级的情况下，RouterC 将作为接收端 DR，数据流会从 Source 经 RouterA、RouterB、RouterC 到达 HostA。
- 分别在 RouterC 的和 RouterD 的 GE1/0/0 接口上使能 PIM BFD 功能，当 RouterC 上的 GE1/0/0 接口出现故障时，那么 RouterD 将会立即感知到链路对端发生故障，从而触发新一轮的 DR 竞选，使数据流迅速切换到从 Source 经 RouterA、RouterB、RouterD 到达 HostA，在很大程度上缩小组播数据传输的中断时间。（注：在没有配置 DR 优先级的默认情况下，IP 地址高的路由器担任 DR。）

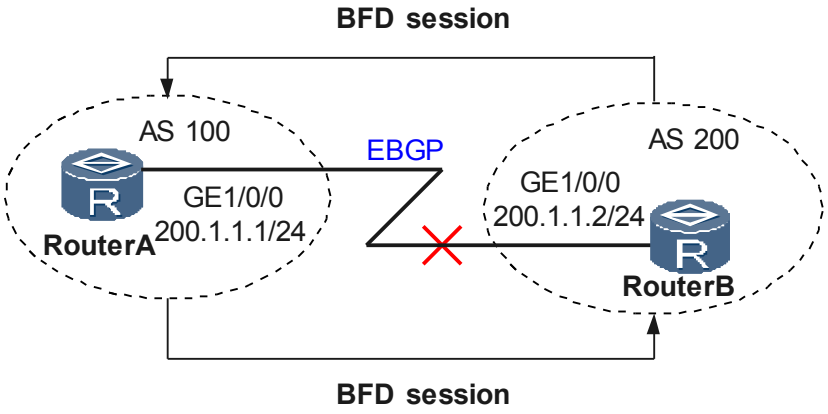
2.4.6 BFD for BGP

BGP 协议通过周期性的向对等体发送 Keepalive 报文来实现邻居检测机制。但这种机制检测到故障所需时间比较长，超过 1 秒钟。当数据达到吉比特速率级别时，将会导致大量的数据丢失，从而无法满足电信级网络高可靠性的需求。

因此，BGP 协议通过引入 BFD for BGP 特性，利用 BFD 的快速检测机制，迅速发现 BGP 对等体间链路的故障，并报告给 BGP 协议，从而实现 BGP 路由的快速收敛。

组网应用

图 2-13 BFD for BGP 组网图



如图 2-13 所示，RouterA 和 RouterB 分别属于 AS100 和 AS200，两台设备直接相连并建立 EBGP 连接。使用 BFD 检测 RouterA 和 RouterB 之间的 BGP 邻居关系，当 RouterA 和 RouterB 之间的链路发生故障时，BFD 能够快速检测到故障并通告给 BGP 协议。

2.4.7 BFD for LSP

在 LSP 链路上建立 BFD 会话，利用 BFD 检测机制快速检测 LSP 链路的故障，可以提供端到端的保护。

BFD 可以用来检测 MPLS LSP 转发路径上数据平面的故障，同时 BFD 的报文格式是固定的。使用 BFD 检测单向 LSP 路径时，反向链路可以是 IP 链路、LSP 或 TE 隧道。

检测 MPLS LSP 的连通性时，BFD 会话协商有两种方式：

- 静态配置 BFD：通过手工配置 BFD 的本地标识符和远端标识符，由 BFD 本身的协商机制建立会话。
- 动态创建 BFD 会话：通过在 LSP Ping 报文中携带 BFD Discriminator TLV 进行会话协商。

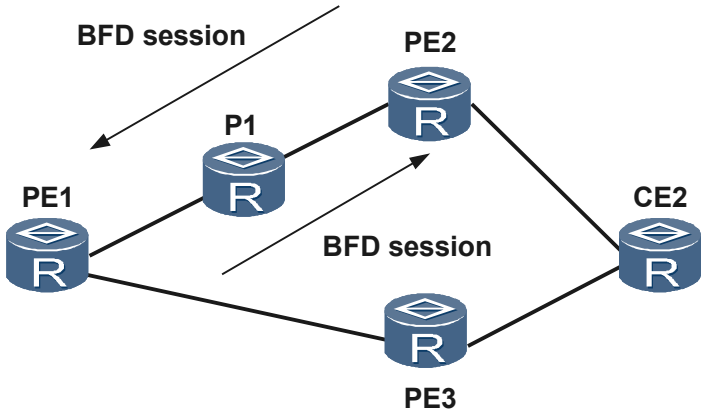
BFD 支持检测的 LSP 类型有：

- 静态 BFD for 静态 LSP
- 静态 BFD for LDP LSP
- 动态 BFD for LDP LSP

BFD 使用异步模式检测 LSP 的连通性，即 Ingress 和 Egress 之间相互周期性地发送 BFD 报文。如果任何一端在检测时间内没有收到对端发来的 BFD 报文，就认为 LSP 状态为 Down，并向 LSPM 上报 LSP Down 消息。

组网应用

图 2-14 BFD for LDP LSP 组网图



如图 2-14 所示，为简化起见，只考虑从 PE1 到 CE2 的流量。当 PE1-P1 之间的链路发生故障时，由于 PE1 可以通过接口感知到故障所以不配置 BFD for LDP LSP 也可以。但是如果 P1-PE2 之间的链路发生故障，则 PE1 无法通过接口感知，这时需要结合 BFD for LDP LSP 来进行快速故障检测。

在 PE1 上有到 PE2 的 LDP LSP，配置 BFD FOR LDP LSP 为这条 LDP LSP 建立 BFD 会话进行检测，同时在 PE1 上配置 VPN FRR 的相关策略，指定保护路径为 PE1-PE3。

当 PE1-P1 或者 P1-PE2 之间的链路发生故障时，PE1 上能迅速感知到 LSP 故障，并触发 VPN FRR 切换，使流量切换到 PE1-PE3-CE2，实现保护。

2.4.8 BFD for PST

当 BFD 检测到故障时，修改端口状态表 PST（Port State Table）中的接口状态，从而触发快速重路由。BFD 会话修改端口状态表功能只能用于绑定接口的 BFD 单跳会话。

BFD for PST 在很多类型的 FRR（快速重路由）中使用广泛。在绑定接口的 BFD 会话中使用 BFD for PST，会将该 BFD 会话与这个接口的 PST 表联动。在 BFD 会话检测链路 Down 后，将该接口的 PST 表对应比特位置 Down，从而立即进行 FRR 切换。

2.4.9 BFD for TE

BFD For TE 是 MPLS TE 中的一种端到端的快速检测机制，用于快速检测隧道所经过的链路中所发生的故障。

传统的检测机制依靠包括 RSVP Hello 或者 RSVP 刷新超时等检测，都具有检测速度缓慢的缺点。BFD 检测机制很好的克服了这些缺点，它采用快速收发报文的机制，完成这些隧道链路故障的快速检测，从而引导承载业务的快速切换，达到保护业务的目的。

BFD 支持的 TE 类型有：

- 静态 BFD for CR-LSP

  静态 BFD for CR-LSP 是指使用 BFD 检测 CR-LSP，做到快速发现 LSP 故障。BFD 会话需要手动配置。

- 静态 BFD for TE

  BFD for TE 使用 BFD 检测整条 TE 隧道，触发 VPN FRR、VLL FRR 等应用进行流量切换。

- 动态 BFD for CR-LSP

  动态 BFD for CR-LSP 的作用与静态 BFD for CR-LSP 相同。所不同的是建立 BFD 会话的方式，动态 BFD for CR-LSP 方式下 BFD 会话动态触发。

- 动态 BFD for RSVP

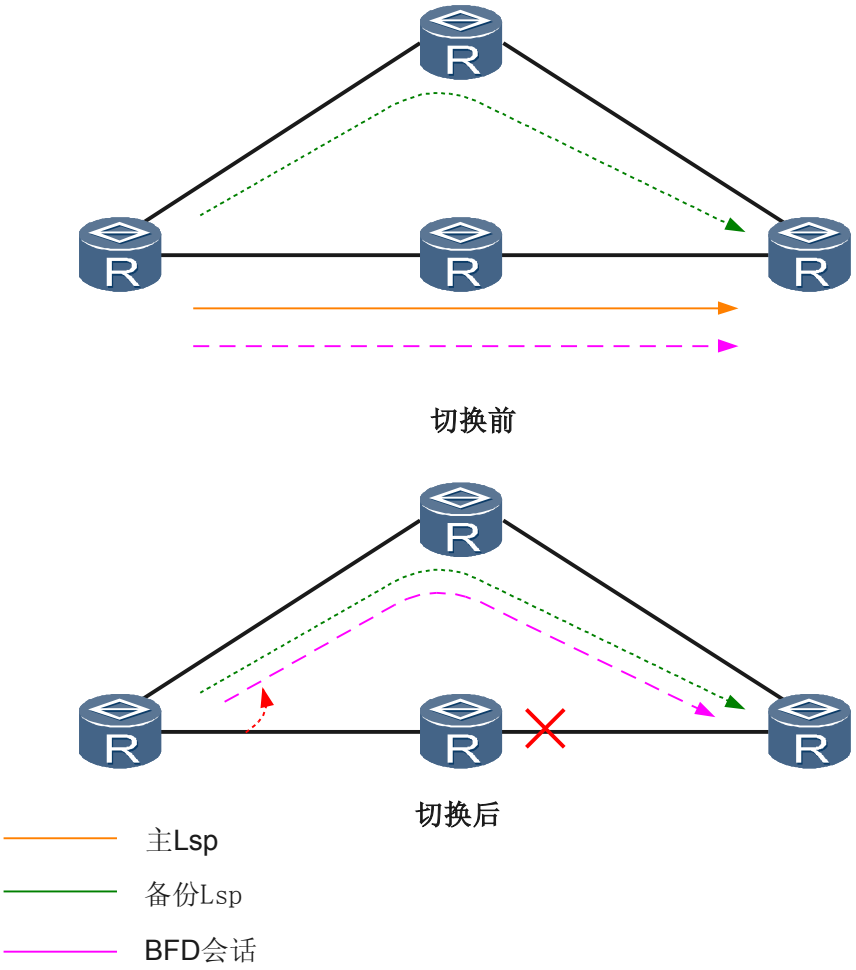
  BFD for RSVP 使用 BFD 检测 RSVP 邻居关系。当 RSVP 相邻节点之间存在二层设备时，这两个节点只能根据 Hello 机制感知链路故障，感知故障时间为秒级，这将导致数据大量丢失。BFD for RSVP 可实现毫秒级故障监测，并配合 RSVP 协议快速的发现 RSVP 邻接故障。BFD for RSVP 一般用在 TE FRR 中 PLR 节点与主路径的 RSVP 邻居之间存在二层设备的情况。

BFD for TE 与 BFD for CR-LSP 的区别是故障通告的对象不同。BFD for TE 是向 VPN 等应用通告故障，触发业务流在不同隧道接口上的切换；BFD for CR-LSP 是向 TE 隧道通告故障，触发业务流在同一 TE 隧道内的不同 CR-LSP 上的切换。

通过和 LSP 绑定，在 Ingress 和 Egress 之间建立 BFD 会话。BFD 检测报文从源端开始经过 MPLS 转发到达宿端，如果在一定时间内宿端没有没有收到检测保温，宿端会将本端 BFD 状态置 down，并通知源端，通过此方式在源端可以快速检测出 LSP 所经过链路的故障状态。

当检测出链路故障以后，BFD 将此信息上报给承载在该 LSP 上面的应用模块，应用再将流量切换到备份路径上，并在备份路径建立新的 BFD 会话。

图 2-15 BFD for LSP 组网图

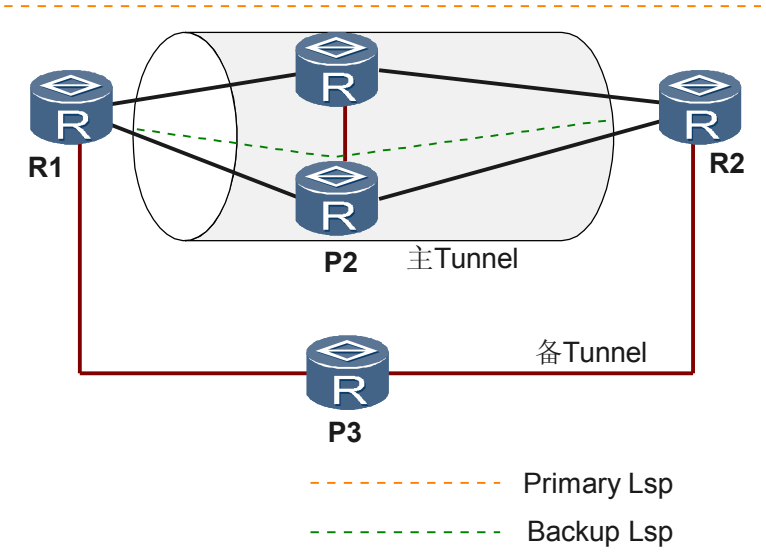


如图 2-15 所示，BFD 会话检测主 LSP 所经过的链路。当主 LSP 所经过的链路出现故障时，在源端 BFD 会立即报告该故障信息，Ingress 就会立即作出决定，将流量切换至备份 LSP（备份 LSP 上已经建立了 BFD 保护）。

组网应用

该组网图可同时满足 BFD for TE 的基本功能、hotstandby 组网和隧道保护组组网三种应用情况。

图 2-16 BFD for TE 组网应用



- 主 LSP 和热备份 LSP 之间的切换

  如图 2-16 所示，在 R1 和 R2 之间建立一条主 Tunnel，同时配置热备份 LSP。在 R1 上建立一个到 R2 的 BFD 会话，用于检测该 Tunnel 中的主 Tunnel。当主 Tunnel 链路出现故障时，BFD 会快速通知 R1。收到故障信息以后，R1 会立即将流量切换到热备份 LSP 上，从而保证流量的不中断性。

- 主 Tunnel 和备份 Tunnel 之间的切换

  如图 2-16 所示，在 R1->P2->R2 之间建立一条主 Tunnel，同时在 R1->P3->R2 之间建立一条备份 Tunnel。在路径 R1->P2->R2 上建立一个 BFD 会话，用于检测主 Tunnel 的路径。当主链路出现故障时，BFD 会快速通知 R1。收到故障信息以后，R1 会立即将流量切换到备份 Tunnel 上，从而保证了流量的不中断。

2.4.10 BFD for PW

BFD for PW 是一种对 L2VPN 网络进行故障检测的机制，并可以通告故障向 L2VPN 通告故障。

L2VPN 利用 BFD 完成隧道或 PW 故障的快速检测，从而引导所承载业务的快速切换，达到业务保护的目的。

BFD 能够对本地和远端的 PE 之间的 PW 链路进行快速故障检测，以支持 PW FRR，减少链路故障给业务带来的影响。

- 静态 BFD 检测 PW：支持 TTL 方式和非 TTL 方式。

  说明

  TTL 方式指通过自动计算或者手动配置的方式确定 TTL 的值；非 TTL 方式指 TTL 的值为 255，无需自动计算或者手动配置的方式确定 TTL 的值。

  - TTL 方式 BFD 检测 PW：TTL 方式的 BFD 检测是指 BFD 报文通过 PW 封装后在 PW 链路上传输。这种方式中，PW 支持控制字方式和非控制字方式。
    - 检测单跳 PW：指 BFD 会话检测端到端的单跳 PW。BFD 根据配置的目的地址和 TTL 值（单跳 PW 的 TTL 值为 1）进行 BFD 会话的协商和报文的收发，从而快速检测 PW 链路。

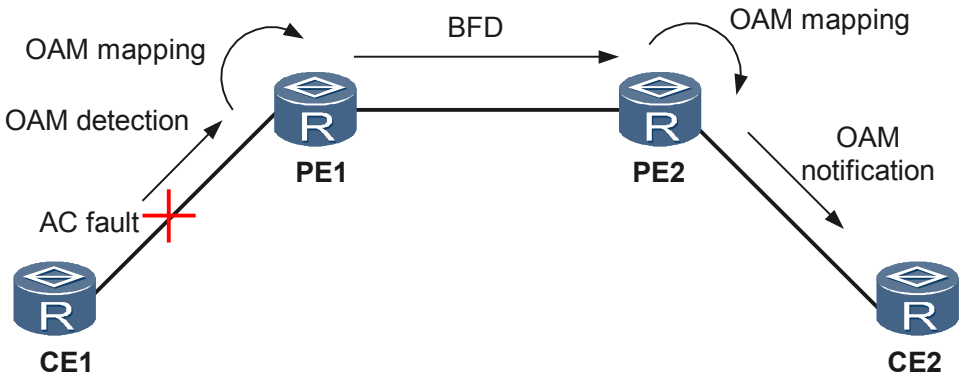
- 检测多跳 PW：需要指定待检测多跳 PW 的目的地址。无论此 PW 是控制字方式还是非控制字方式，BFD 报文都可以穿越多个 SPE 节点到达目的端。
- 检测单播 VPLS：BFD 能够检测单跳 PW 或者多跳 PW（VPLS 接入 VLL）。与 BFD 检测组播 VPLS 不同，BFD 检测单播 VPLS 能够实现对 VPLS 中任意一条 PW 链路进行检测。

- 非 TTL 方式 BFD 检测 PW：BFD 报文通过 PW 封装后在 PW 链路上传输。这种方式中，PW 只支持控制字方式，通过控制字来区分控制报文和数据报文。目前，BFD 支持端到端的单跳 PW 检测。

- 动态 BFD 检测 PW：仅支持非 TTL 方式。这种方式中，PW 只支持控制字方式。PW 的 Up 和 Down 会触发动态创建和删除 BFD 会话。当需要检测的 PW 的状态变为 Up 后，本端设备将邻居信息和检测参数通知 BFD 模块建立相应会话检测邻居之间的链路。

故障检测和传递机制

图 2-17 AC 故障检测和传递机制组网图



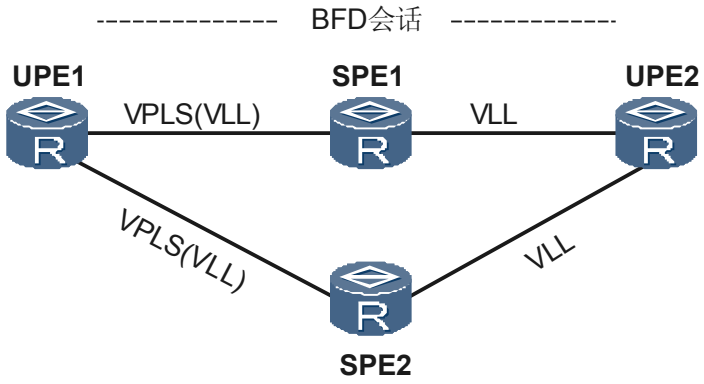
如图 2-17 所示，CE1 和 PE1 之间的 AC 发生故障后：

1. PE1 的 AC OAM 检测到 AC 故障。
2. PE1 的 OAM Mapping 根据 AC 映射出对应的 PW。
3. 通过 BFD 将 OAM 故障消息透明传输给 PE2。
4. PE2 收到 BFD 故障消息时，如果远端 PE1 有备份 PW，则进行流量切换，否则进行 OAM Mapping，映射出对应的 AC 后通告故障给 CE2。

应用

如图 2-18 所示，UPE1—>SPE1—>UPE2 之间的链路为主用 PW，UPE1—>SPE2—>UPE2 之间的链路为备用 PW。分别在 UPE1 和 UPE2 上配置 BFD 会话，检测 UPE1 和 UPE2 之间的多跳 PW。当 BFD 会话检测到 UPE1 和 UPE2 之间的主链路故障后，会及时通知 PW 将流量切换到备用 PW 上。

图 2-18 静态 BFD 会话检测多跳 PW 组网图



2.4.11 BFD6

BFD6 是 BFD for IPv6 的简称。它和 BFD for IPv4 的功能相同，用于检测系统之间的通信故障，并在故障出现时通知上层应用。

BFD6 和 BFD for IPv4 的异同点

可以根据 BFD6 和 BFD for IPv4 所支持的功能，来查看它们的异同点。

BFD6/BFD for IPv4 支持的特性	BFD6	BFD for IPv4
IP 链路	是	是
静态路由	是	是
OSPF	否	是
OSPFv3	是	否
BGP	是	是
ISIS	是	是
PIM	是	是
PST	是	是
PIS	否	是
PW	否	是
TE	否	是
LSP	否	是

BFD6 与路由协议关联

网络上的链路故障或拓扑变化都会导致路由器重新进行路由计算，要想提高网络的可用性，缩短路由协议的收敛时间是非常重要的。因此，加快故障感知速度并将故障快速通告给路由协议是一种可行的方案。

将 BFD6 与路由协议关联，路由协议在建立新邻居后动态创建 BFD6 会话来检测邻居之间的链路。当 BFD6 检测到链路故障时，能够将故障通告给路由协议，触发路由协议的快速收敛；当邻居关系为 Down 时，则动态删除 BFD6 会话。

目前，与 BFD6 关联的路由协议包括：

- OSPFv3
- BGP
- ISIS
- PIM

BFD6 for 静态路由

静态路由自身没有检测机制，当网络发生故障时需要管理员介入。

使用 BFD6 for 静态路由特性，可以利用 BFD6 会话检测对公网 IPv6 静态路由的状态。路由管理系统根据 BFD 会话的状态决定静态路由是否可用。

BFD6 检测的链路类型

BFD6 可以对 IPv6 链路进行检测。它和 BFD for IPv4 一样，也分为单跳会话和多跳会话。

BFD6 支持的接口类型如下：

- 三层物理接口
- 以太子接口（包括 Eth-Trunk 子接口）
- IP-Trunk
- Eth-Trunk
- VLANIF 接口

2.5 术语与缩略语

缩略语(Abbreviations)

缩略语	英文全称	中文全称
ISIS	Intermediate System-Intermediate System	中间系统到中间系统
BFD	Bidirectional Forwarding Detection	双向转发检测
VC	Virtual Circuit	虚电路
VLL	Virtual Leased Line	虚拟租用链路
AC	Attachment Circuit	接入电路

缩略语	英文全称	中文全称
PE	Provider Edge Router	边缘路由器
CE	Customer Edge Router	用户边缘路由器
OSPF	Open Shortest Path First	开放式最短路径优先协议
TE	Traffic Engineer	流量工程
CSPF	Constraint Shortest Path First	约束最短路径优先
VRRP	Virtual Router Redundancy Protocol	虚拟路由冗余协议
L2VPN	Layer 2 virtual private network	二层虚拟专用网
PW	Pseudo Wire	虚电线
MPLS	Multi Protocol Label Switching	多协议标签交换
BFD6	BFD For IPv6	BFD 支持 IPv6 特性
OSPFv3	Open Shortest Path First-v3	开放式最短路径优先协议

3 以太网 OAM

关于本章

3.1 介绍

定义

以太网 OAM 全称为以太网 Operations Administration and Maintenance，即针对以太网的操作管理和维护。

以太网 OAM 主要功能可分为以下两部分：

- 故障管理
  - 通过定时或手动发送检测报文来探测网络的连通性。
  - 提供类似 IP 网络中的 Ping（Packet Internet Groper）和 Traceroute 的功能，对以太网进行故障确认和故障定位。
  - 和保护倒换协议配合，当检测到连通性故障后触发保护倒换，以实现网络业务中断小于等于 50 毫秒的运营级可靠性目标。
- 性能管理

  性能管理主要是指对网络中的丢包、时延、抖动进行衡量，也包括对网络中各类流量进行统计。性能管理通常在用户接入点实施。有了性能管理工具，运营商可通过网管系统监测网络运行情况、定位故障，确认网络转发能力是否符合与用户签订的 SLA（Service Level Agreement）。

以太网 OAM 能够有效提高以太网的网络管理能力和维护能力，保障网络的稳定运行。

目的

以太网技术自诞生起，以其简单易用、价格低廉的特点逐步成为局域网的主导技术。近年来，随着千兆、万兆以太网技术的相继应用，以太网已经向城域网和广域网方向扩展。

以太网最初主要应用于局域网。相对于城域网和广域网，局域网对可靠性、稳定性要求都较低，因此以太网一直缺乏运行维护管理机制，这一点已成为以太网作为 ISP 网络的严重障碍。因此，在以太网上实现 OAM（Operations Administration and Maintenance）成为一个必然的发展趋势。

3.2 参考标准和协议

本特性的参考资料清单如下：

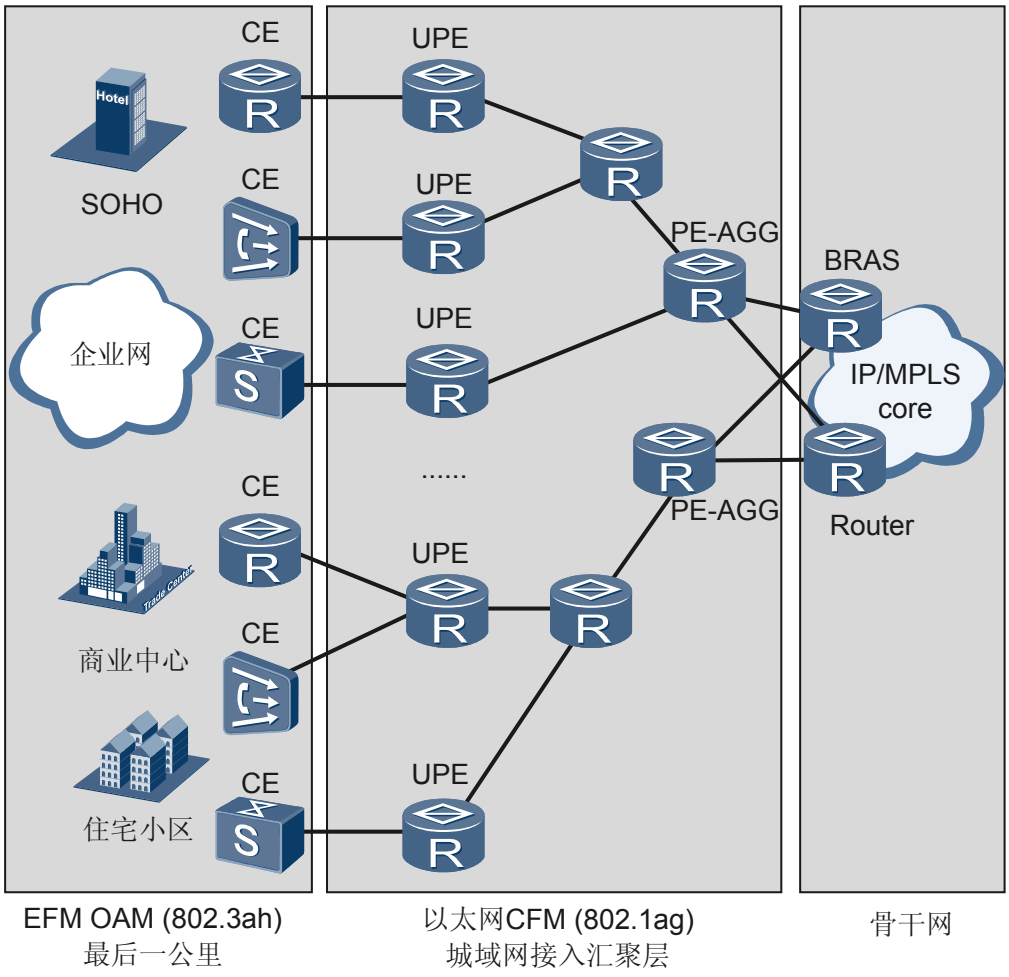
文档	描述	备注
IEEE Std 802.3ah-2004	Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks	

文档	描述	备注
IEEE Std 802.1ag-2007	IEEE Standard for Local and metropolitan area networks— Virtual Bridged Local Area Networks Amendment 5: Connectivity Fault Management	
IEEE 802.1ag/ Draft7.0	Virtual Bridged Local Area Networks— Amendment 5: Connectivity Fault Management	
MEF6	This document defines two generic service constructs called Ethernet Service types and specifies their associated service attributes and parameters used to create point-to-point and multipoint-to-multipoint Ethernet services. This document also defines the requirements for several Ethernet services that use these generic Ethernet Service types.	
Y.1731	Y.1731 is an OAM protocol organized by the ITU-T. It covers not only the contents defined by IEEE802.1ag but also combinations of OAM messages, including the Alarm Indication Signal (AIS), Remote Defect Indication (RDI), Locked Signal (LCK), Test Signal, Automatic Protection Switching (APS), Maintenance Communication Channel (MCC), Experimental (EXP), and Vendor Specific (VSP) for fault management and performance monitoring, such as frame loss measurement (LM) and delay measurement (DM).	

3.3 原理描述

以太网 OAM 技术分层实现，可分为链路级以太网 OAM 和网络级以太网 OAM。

图 3-1 城域网典型组网



链路级以太网 OAM

链路级以太网 OAM 技术，如遵循 IEEE 802.3ah 协议的 EFM OAM（Ethernet in the First Mile OAM），针对两台直连设备之间的链路，提供链路连通性检测功能、链路故障监控功能、远端故障通知功能、远端环回功能。如图 3-1 所示，在城域网中，链路级以太网 OAM 技术多应用在 CE（Customer Edge）设备和 PE（Provider Edge）设备之间，用于保证用户网络和运营商网络之间连接的可靠性和稳定性。

CE 设备是用户网络边缘设备。CE 设备用来连接用户网络和运营商网络。与 CE 设备不同，PE 设备是运营商网络边缘设备，它用来连接运营商网络和用户网络。

遵循 IEEE 802.3ah 协议的 EFM OAM 是 Ethernet in the First Mile OAM（以太最后一公里 OAM）的简称。它属于链路级以太网 OAM 技术，是一种较为简单的 OAM 协议，提供点到点的故障检测，主要用于两台直连设备之间链路的 OAM 检测。

网络级以太网 OAM

遵循 IEEE 802.1ag 协议规定的以太网 CFM（Connectivity Fault Management），属于网络级以太网 OAM 技术针对网络实现端到端的连通性故障检测、故障通知、故障确认和

故障定位功能，可用于监测整个网络的连通性，定位网络的连通性故障，并可与保护倒换技术相配合，提高网络的可靠性。

OAM 故障联动

OAM 故障联动用于在不同的检测协议之间传递故障信息，例如在同属于以太网 OAM 的 EFM OAM 和以太网 CFM 之间传递故障信息，在以太网 OAM 和 BFD 之间传递故障信息等。随着以太网 OAM、BFD、MPLS OAM 等检测协议的不断推广运用，OAM 故障联动模块将承担起更多的使用场景。

3.3.1 EFM OAM

EFM OAM 功能包括：对端发现、链路监控、故障通告和远端环回。

对端发现

EFM OAM 工作模式是具备 EFM OAM 功能的接口的一个属性，包括两种：主动模式和被动模式。接口的默认工作模式是主动模式。

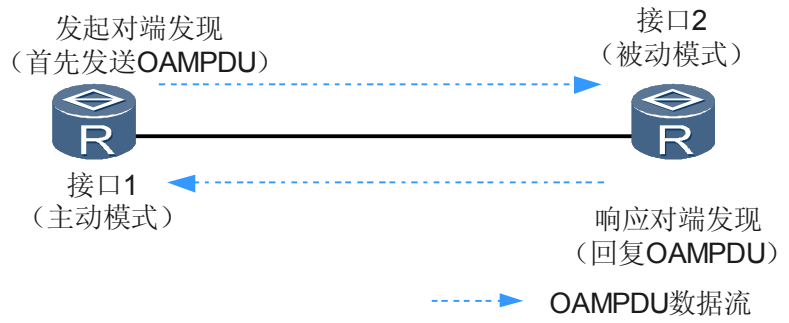
接口配置 EFM OAM 之前首先要对其工作模式进行配置：

- 如果配置为主动模式，则由该接口主动发起对端发现过程。
- 如果配置为被动模式，则不能由该接口主动发起对端发现过程。

  这样可以确保两个被动端不能进行会话协商。处于被动模式的端口不能主动发起远端环回和变量获取请求。

当某个接口使能 EFM OAM 功能时，如果该接口的 EFM OAM 工作模式是主动模式，则该接口发起对端发现过程，该接口和对端接口进入 EFM OAM 发现阶段。

图 3-2 对端发现示意图



如图 3-2 所示，假设接口 1 的 EFM OAM 工作模式是主动模式，接口 2 为被动模式。当使能接口 1 的 EFM OAM 功能时，对端发现过程如下：

1. 接口 1 发送 OAMPDU（OAM Protocol Data Unit），该 OAMPDU 中包含接口 1 的 EFM OAM 配置信息。
2. 接口 2 收到 OAM PDU 后，和自己的 EFM OAM 配置进行匹配比较，然后向接口 1 回复 OAM PDU。该 OAM PDU 除包含接口 1 和接口 2 的 EFM OAM 配置外，还包含接口 2 对接口 1 的 EFM OAM 配置是否满意的标志。

   报文格式请参看下图。

图 3-3 EFM OAM PDU 报文格式

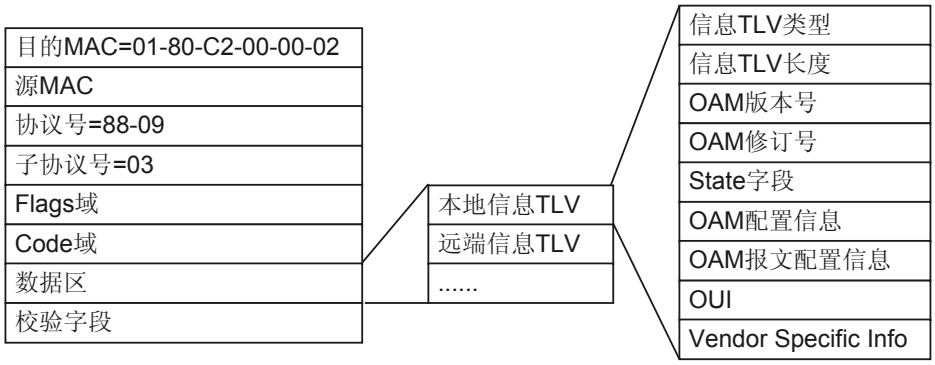


图 3-4 OAM 配置信息字段含义

值	名称	描述
OAM Configuration	OAM配置信息	7:5保留,在本地信息TLV中置0
		4 变量可达性 1=DTE支持发送变量响应OAMPDU 0=DTE不支持发送变量响应OAMPDU
		3 链路事件 1=DTE支持解析链路事件 0=DTE不支持解析链路事件
		2 OAM远端环回支持 1=DTE有OAM远端环回模式的能力 0=DTE没有OAM远端环回模式的能力
		1 单项支持 1=当接收链路不工作时DTE有能力发送OAMPDU 0=当接收链路不工作时DTE没有能力发送OAMPDU
		0 OAM模式 1=DTE配置为主动模式 0=DTE配置为被动模式

3. 接口 1 收到接口 2 发来的 OAMPDU 后，再来判断接口 2 的 EFM OAM 配置和本端的配置是否匹配。

通过以上过程，如果接口 1 和接口 2 的 EFM OAM 配置匹配，则双方进入 Detect 状态。在 Detect 状态，双方通过定时发送 OAMPDU 维持邻居关系。如果配置不成功，则仍然处于 Discovery 状态，继续进行发送信息报文进行协商，直到协商成功或是当前接口 EFM 去使能。OAMPDU 的发送时间间隔为 1 秒且不可调。

链路监控

在配置了 EFM OAM 链路监控功能以后，会查询接口管理模块的物理层的统计数据，检测当前接口所在链路的通信质量。在设定的观察时长内，如果接口检测到的误帧数量、

误码数量或误帧秒数量达到或超过设定的阈值，则表示该链路存在故障，在本地产生告警并通知网管，并通过向对端设备发送 OAMPDU 来通告故障。其中，误帧秒是指将设定的观察时长划分为若干个以 1 秒为单位的时间段，如果在某个时间段内检测到了至少一个误帧，则该时间段称为一个误帧秒。

故障通告

故障通告类型包括：协议报文超时、物理链路故障、OAM 管理模块传递故障。

- 对于协议报文超时和物理链路故障，将故障事件记入日志并上报给网管系统。
- 对于 OAM 管理模块传递的故障，将故障事件记入日志并上报给网管系统。

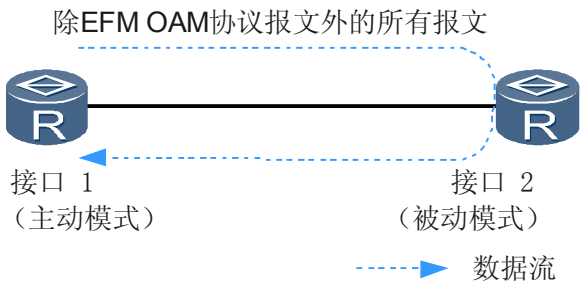
  当反向链路可达时，向对端发送 OAM PDU，通知对端发生故障。对端设备接到通知消息后，将消息中携带的故障事件记入日志，并上报给网管系统。
- 如果与 OAM 故障联动配合使用，配置了 EFM OAM 与其他模块的联动关系，如与 BFD 联动、与以太网 CFM 联动、与 MPLS OAM 联动，则 OAM 故障联动模块会将故障通告给后者以触发相应联动处理。

远端环回

如图 3-5 所示，远端环回是指当本端向对端发送除 EFM OAM 协议报文外的所有其它报文时，对端（远端）接收到报文后不按照报文的目的地址进行转发，而是将报文再发回给本端。

远端环回功能可用于定位链路故障和测试链路质量。进入环回模式后，可在本端往对端设备发送测试报文，根据发送的数据包数目和收到的数据包数目对比可计算当前链路的丢包率等通讯质量的参数。

图 3-5 远端环回示意图



处于主动模式的接口才有资格发起远端环回。当某个接口的 EFM OAM 工作模式为主动模式且它和对端接口处于 Detect 状态时，使能该接口的远端环回功能：

1. 该接口向对端发送环回请求消息并等待回复。
2. 对端接口接收到环回请求消息后，向本端回复环回响应消息并进入环回状态。
3. 如果本端在 2 秒内收到环回响应消息后并进入环回状态。如果本端在 2 秒内未到环回响应消息，会重新向对端发送环回请求消息。本端最多可以重发三次环回请求消息。

当本端需要停止远端环回时，本端向对端发送环回取消消息。对端接到环回取消消息后退出环回状态。

为避免由于用户忘记停止远端环回而造成链路长时间无法正常转发业务数据，EFM OAM远端环回具有超时自动取消功能。远端环回的超时时间可设置。远端环回超时后，本端自动向对端发送环回取消消息。

主备扩展功能

- 概念

  EFM OAM 主备扩展功能是指 EFM OAM 扩展了原有的 EFM 的检测功能和通告功能。

  路由器支持解析对端发送的的信息报文中携带的主备状态标志，根据状态标志来控制接入链路的状态。并通过与接口联动、静态路由联动来控制业务流量的走向。在主用链路出现故障时使流量自动切换到备用链路，保障网络的可靠性。

- 目的

  对于 IP 业务设备来说，为保证业务的可靠性，一般采用主备冗余方式接入到 IP 网络。当前主流设计中，IP 业务设备主要通过 VRRP 方式接入路由器，但这种方式当初是为 PC 终端设计，所以在可靠性上不能满足电信级的业务的需求。

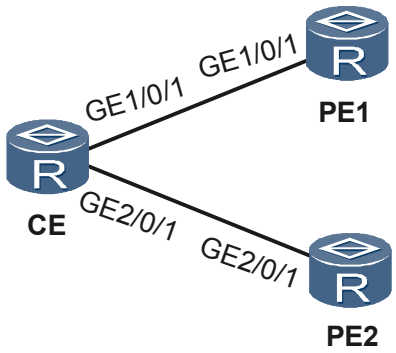
  EFM OAM 主备扩展功能，有如下优点：

  - 既可以接入三层网络（IP/MPLS），也可以接入二层网络（PBB-TE/T-MPLS/VPLS），目前被各设备厂商认可。
  - 主备链路切换时间短，可靠性高，满足电信级业务的需要。

- 基本原理

  接口使能 EFM OAM 功能后，直连链路的两端通过 EFM 信息报文，进行对端发现或链路监测。如图 3-6 所示，为了达到主备控制的目的，要求控制端设备 CE 能够发送携带主备状态标志的信息报文。在信息报文中加入带有主备信息的特定厂商 TLV。被控制端设备具有 EFM OAM 扩展功能，使路由器具有解析对端发送的主备扩展 TLV 的能力。其中主备标志的字段 0xAB 表示主用状态，0xBA 表示备用状态。路由器根据主备状态，改变端口检测状态或者改变路由有效状态，达到控制业务流量走向的目的，实现自动倒换。

  图 3-6 主备扩展功能示意图

  EFM OAM 的主备扩展功能包括以下三点：

  - 主备报文解析功能

    配置 EFM OAM 功能的接口不但能够解析普通的 802.3ah 信息报文的能力，同时能够正确识别信息报文中特殊厂商类型的 TLV 字段，解析出该 TLV 中的主备状态标志，从而能够正确处理信息报文。

- EFM OAM 扩展与接口联动功能

  当主备链路切换时，需要通知相关业务，切换业务流量走向。

  配置 EFM 与接口联动功能，EFM 可以感知接口主备切换后，通知接口进行链路状态切换。这样，其它业务根据接口链路状态进行相应操作（例如：重新发布路由等）。

  当接口主备状态由备变为主，EFM 通知接口，接口状态变为 Up。同时，通知其它业务，根据需要切换业务流量到该接口。

  当接口主备状态由主变为备，EFM 通知接口，接口状态变为 Down。同时，通知其它业务，根据需要将业务流量切换到其它 Up 状态的接口。

- EFM 与静态路由联动功能

  当主备链路切换，需要实时刷新相关静态路由。EFM 与静态路由联动功能，根据端口主备状态，实时通知路由管理模块，进行静态路由刷新

  当接口主备状态由备变为主，EFM 通知路由管理模块，生成相关静态路由。

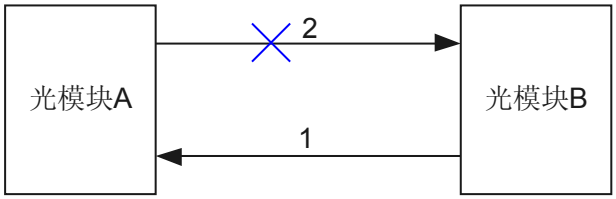
  当接口主备状态由主变为备，EFM 通知路由管理模块，删除相关静态路由。

单纤故障检测功能

目前的光接口具有全双工的特点。光接口如果能接收到报文就认为自身是物理 Up 的。但是存在一种情况导致端口实际工作状态和端口物理状态不一致。

如图 3-7 所示，光接口 A 和光接口 B 直连。若 2 号线发生故障，则接口 B 收不到报文，对应的端口物理状态为 Down。而接口 A 通过 1 号线仍然可以收到接口 B 发送的报文，则接口 A 的物理状态依然为 Up。但此时这个端口实际上已经不具有承载业务的能力，接口 A 所在设备的业务无法感知端口真正的工作状态，从而影响业务传输。

图 3-7 EFM OAM 单纤检测原理示意图



EFM OAM 的单纤检测功能可以很好的解决这个问题。

EFM 在检测到故障时，若配置 EFM 与接口管理联动则会改变端口的状态为 EFM Down。对于二、三层业务来说，EFM Down 相当于物理 Down，从而会让业务感知到真实的端口的工作状态。若故障消除，EFM OAM 重新协商成功后会将接口状态恢复正常，此时业务也可以恢复正常状态。

这种机制有效地解决了单纤故障导致端口状态和实际状态不一致的问题，保障了各业务特性感知的接口状态准确性和可靠性。

3.3.2 以太网 CFM

以太网 CFM 针对网络实现端到端的连通性故障检测、故障通知、故障确认和故障定位功能。可用于监测整个网络的连通性，定位网络的连通性故障，并可与保护倒换技术相配合，提高网络的可靠性。

现如今，802.1ag 协议分为 draft7 草案版本和 standard2007 标准版本，分别遵循 IEEE 802.1ag/Draft7.0 和 IEEE Std 802.1ag-2007。两者主要区别如**表 3-1** 所示。

表 3-1 802.1ag 协议 draft7 版本和 standard2007 版本的区别

特性	draft7	standard2007	备注
MD（Maintenance Domain）	支持	支持	标准版本和草案实现的功能、配置一样
默认 MD（Default Maintenance Domain）	不支持	支持	-
MA（Maintenance Association）	支持	支持	标准版本和草案实现的功能、配置一样
MEP（Maintenance association End Point）	支持	支持	标准版本和草案实现的功能、配置一样
RMEP（Remote Maintenance association End Point）	支持	支持	标准版本和草案实现的功能、配置一样
MIP（Maintenance association Intermediate Point）	支持	支持	标准版本和草案中 MIP 节点的生成规则是一样的，分为：default 型、explicit 型和 none 型。不同的是： ● 草案中，MIP 节点是基于端口创建。 ● 标准版本中，MIP 节点是基于 MD 或默认 MD 创建。
MP（Maintenance Point）	支持	支持	标准版本和草案实现的功能、配置一样

由于 draft7 草案版本和 standard2007 标准版本报文格式有差异，如果要在网络中使用以太网 CFM 功能，必须保证整网版本一致。

基本概念

- MD

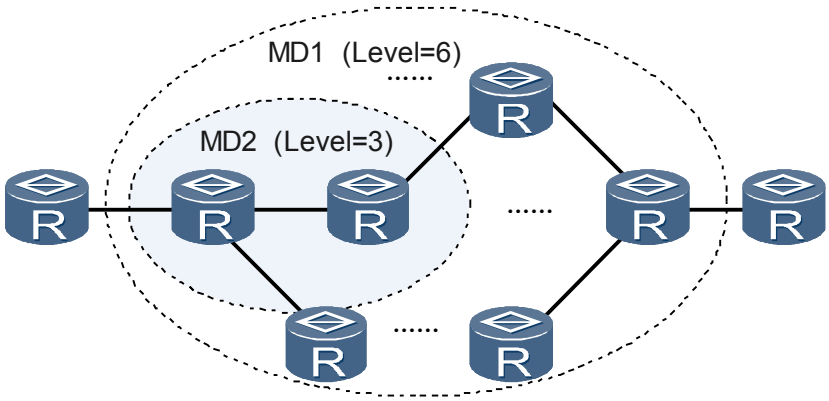
  维护域 MD（Maintenance Domain）是指对其实施以太网 CFM 管理的一个网络或一个网络的一部分，一个 MD 通常由一个统一的 ISP（Internet Service Provider）或者运营商进行管理。

  每个 MD 有一个级别，取值范围是 0 ～ 7，值越大 MD 的级别越高。低级别 MD 的 802.1ag 协议报文进入高级别的 MD 后被丢弃，高级别 MD 的 802.1ag 协议报文可以穿越低级别的 MD。

  在实际应用中，如果某个 MD 内还包含另一个范围较小的 MD，在对大的 MD 进行连通性检测时，802.1ag 协议报文需要穿越小的 MD，则可将大 MD 的级别配置得比小 MD 高，以实现 802.1ag 协议报文穿越。

  例如在图 3-8 所示网络中，MD2 包含在 MD1 中，MD1 的 802.1ag 协议报文需要穿越 MD2。因此，将 MD1 的级别配置为 6，MD2 的级别配置为 3，这样 MD1 内的 802.1ag 协议报文就可以穿越 MD2 实现整个 MD1 的连通性故障管理，而 MD2 的 802.1ag 协议报文不会扩散到 MD1 中。

  图 3-8 不同级别的 MD

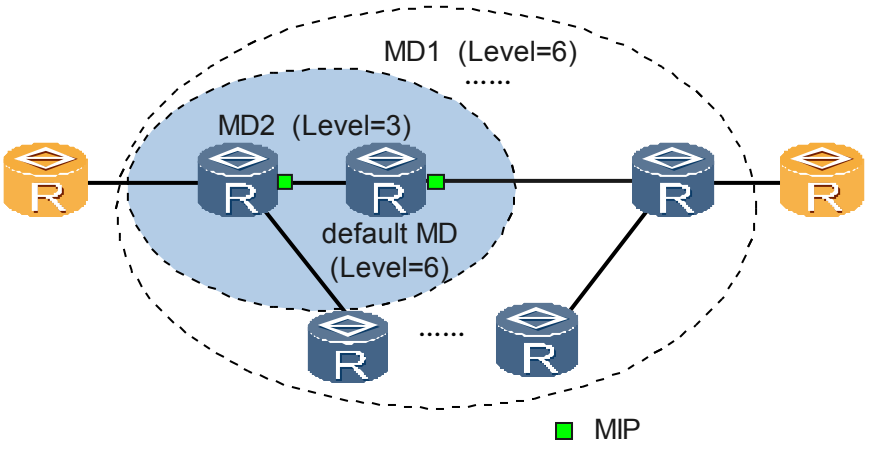
  

- 默认 MD

  standard2007 标准版本规定，每台设备上可配置一个默认 MD。默认 MD 可用于使高优先级 MD 感知低优先级 MD 内部拓扑。

  如图 3-9 所示，在 MD 嵌套场景下，高优先级 MD 中的设备可能是低优先级 MD 的边缘设备和中间设备。当高优先级 MD 内的 802.1ag 协议报文穿越低优先级 MD 时，报文会透传；在没有配置默认 MD 的情况下，如果需要感知低优先级 MD 内部拓扑，需要在低优先级 MD 内设备上的指定端口上创建指定优先级的 MIP 节点，用以回复 LBR 或 LTR 报文给高优先级 MD 中的设备。

图 3-9 默认 MD



如果在低级别 MD 中的设备上配置与高优先级 MD 相同级别的默认 MD，则设备基于默认 MD 按规则自动生成相应级别的 MIP 节点回复 LBR 或 LTR 报文给高优先级 MD 中的设备，这样高级别 MD 就能感知低级别 MD 中拓扑结构的变化，也实现了整个 MD1 的连通性故障管理。

默认 MD 的级别必须高于本设备配置 MEP 节点所有 MD 的级别，和高优先级 MD 的级别相等，用于高级别 CCM 报文穿越，创建 MIP 节点回复 LTR 报文。

IEEE802.1ag 标准版本规定，每台设备上可配置一个默认 MD，一个默认 MD 可以关联多个 VLAN。VLAN 内的端口将基于默认 MD 按规则自动创建 MIP 节点。

说明

在配置了默认 MD 的设备上，默认 MD 关联的 VLAN 一定不能关联 MA。

- MA

  维护联盟 MA（Maintenance Association）是 MD 的一部分。一个 MD 可以划分成一个或多个 MA。以太网 CFM 对每个 MA 分别进行连通性故障检测。

  在运营商的网络中，通常一个 VLAN 对应一个业务实例，通过划分 MA 可实现对传输某个业务实例的网络的连通性故障检测。

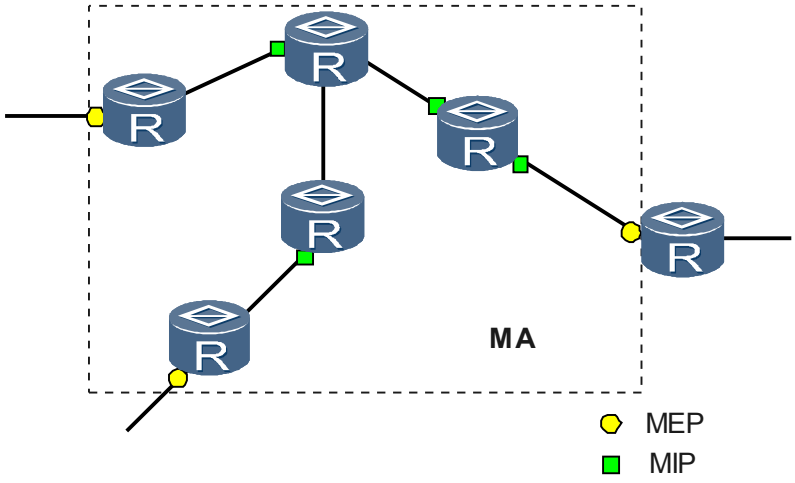
  MA 的级别等于它所在的 MD 的级别。

- MEP

  如图 3-10 所示，维护联盟边缘节点 MEP（Maintenance association End Point）是 MA 的边缘节点。

  MEP 位于设备的接口上，手工创建。MEP 的级别等于它所在的 MD 的级别。

图 3-10 MEP 和 MIP



对于运行以太网 CFM 的网络中的任意一台设备，该设备上的 MEP 称为本地 MEP，同一 MA 内其它设备上的 MEP 对本设备而言称为远端维护联盟边缘节点 RMEP（Remote Maintenance association End Point）。

MEP 分为 inward 型 MEP 和 outward 型 MEP。inward 型 MEP 发出的 802.1ag 协议报文通过当前 MA 关联 VLAN 内的所有接口（除 MEP 所在接口）向外发送，即在当前 MA 关联的 VLAN 内广播；outward 型 MEP 发出的协议报文直接通过该 MEP 所在的接口向外发送。

- MIP

  如图 3-10 所示，维护联盟内部节点 MIP（Maintenance association Intermediate Point）是 MA 的内部节点。

  MIP 位于设备的接口上，是按规则自动创建的，MIP 的创建规则有三种。在 draft7 草案版本和 standard2007 标准版本中，MIP 创建规则差异如表 3-2 所示。

表 3-2 draft7 草案版本和 standard2007 标准版本的区别

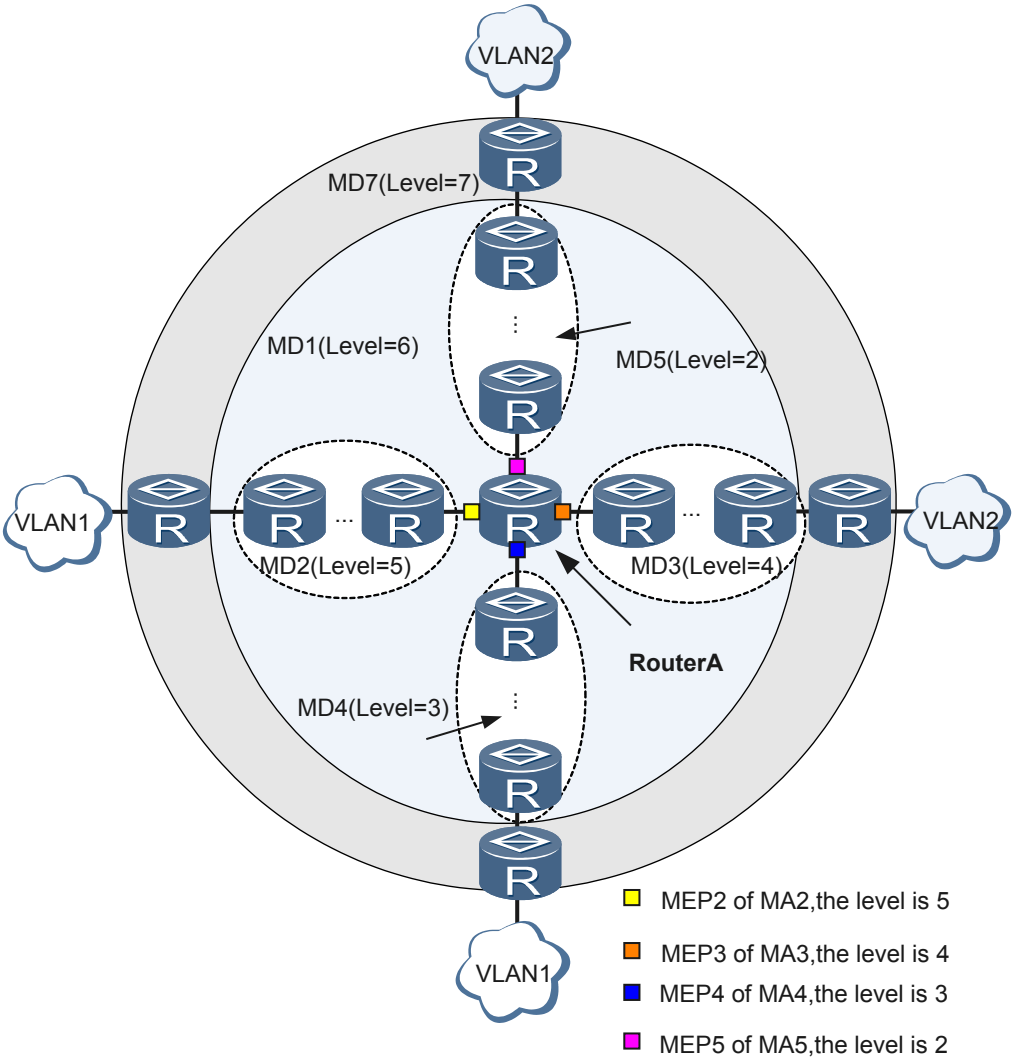
MIP 创建规则	draft7 草案版本	standard2007 标准版本
default 型	如果指定接口不存在更高级别的 MEP，并且不存在更低级别的 MIP，则可以在该接口上创建 MIP。	如果指定 MD 或默认 MD 所属的接口不存在更高级别的 MEP，并且不存在更低级别的 MIP，则可以在该接口上创建 MIP。
explicit 型	如果在指定接口下存在更低级的 MEP 但不存在更高级别的 MEP，而且不存在更低级别的 MIP，则可以创建 MIP。在本类型下，接口上只有已配置了更低级别的 MEP 才可能创建 MIP。	如果在指定 MD 或默认 MD 所属的接口存在更低级的 MEP 但不存在更高级别的 MEP，而且不存在更低级别的 MIP，则可以创建 MIP。在本类型下，接口上只有已配置了更低级别的 MEP 才可能创建 MIP。

MIP 创建规则	draft7 草案版本	standard2007 标准版本
none 型	不自动创建 MIP	不自动创建 MIP

MIP 节点的级别由生成规则以及生成其的M D 的级别决定。

如图 3-11 所示，MD1 ～ MD5 包含在 MD7 中，MD2 ～ MD5 包含在 MD1 中。MD7 的级别高于 MD1 ～ MD5 的级别，MD1 的级别高于 MD2 ～ MD5 的级别。MD1 中的设备 RouterA 上创建了多个 MEP，各个 MEP 属于不同级别的 MD。

图 3-11 standard2007 标准版本 MIP 节点创建示意图



如果，设备 RouterA 已经基于 MD1 配置了 MIP 节点的创建规则为 default，MIP 节点创建过程如下：

1. 比较各个 MEP，找出最高级别的 MEP，最高级别的 MEP 为 5。（MEP 的级别由自身所属的 MD 决定。）

2. 选择高于 5 级 MEP 的最小 MD，MD 的级别为 6。
3. 创建级别为 6 的 MIP 节点。

如果级别 6 或者大于级别 6 的 MD 都不存在，则 MIP 节点将无法创建。

如果 RouterA 上已经存在级别为 1 的 MIP 节点，则 MIP 节点按规则也无法创建。

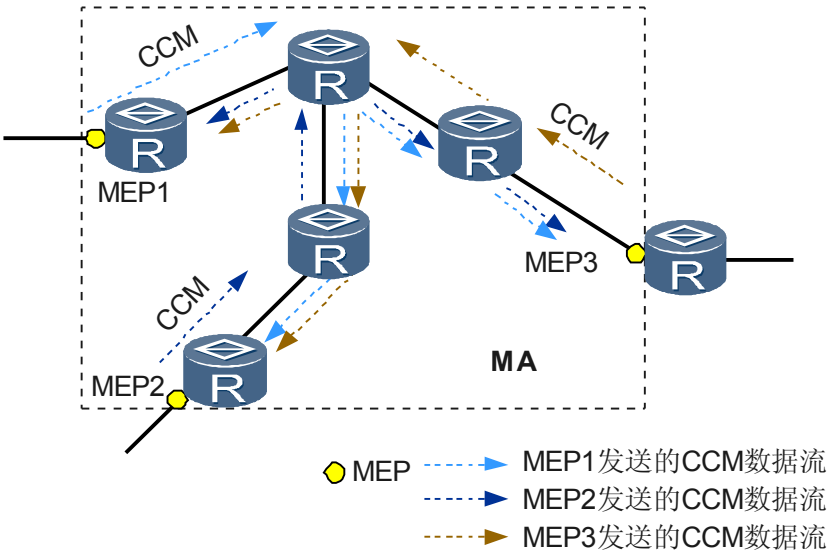
- MP

  MEP 和 MIP 统称为维护节点 MP（Maintenance Point）。

连通性检测

以太网 CFM 通过 MEP 之间定期互发 CCM（Continuity Check Message）来检测各 MEP 之间的连通性，这种检测称为连通性检测或 CC 检测。

图 3-12 CC 检测示意图



- CCM 的产生

  CCM 由 MEP 产生并发送。如图 3-12 所示，MEP1、MEP2 和 MEP3 是同一个 MA 的 3 个 MEP。当使能了 CCM 发送功能后，MEP1 定期以组播方式向 MEP2 和 MEP3 发送 CCM。同样，MEP2 以相同的周期向 MEP1 和 MEP3 发送 CCM，MEP3 也以相同的周期向 MEP1 和 MEP2 发送 CCM。

  CCM 中携带有该 CCM 的级别信息。CCM 的级别等于发送该 CCM 的 MEP 的级别。

- MEP 数据库

  每个启动了以太网 CFM 功能的设备上都有一个 MEP 数据库。MEP 数据库中记录着本设备上的 MEP（即本地 MEP）和同一 MA 内的其它设备上的 MEP（即远端 MEP）。本地 MEP 和远端 MEP 均由用户手工配置后由设备自动记入 MEP 数据库。

- 故障判定

  如果某个 MEP 连续 3 个 CCM 发送周期没有接收到另一个远端 MEP 发送的 CCM，则认为和该 MEP 之间发生了连通性故障。如果配置了 OAM 故障联动，OAM 管理模块会触发相应的联动或倒换处理。

- CCM 的终结

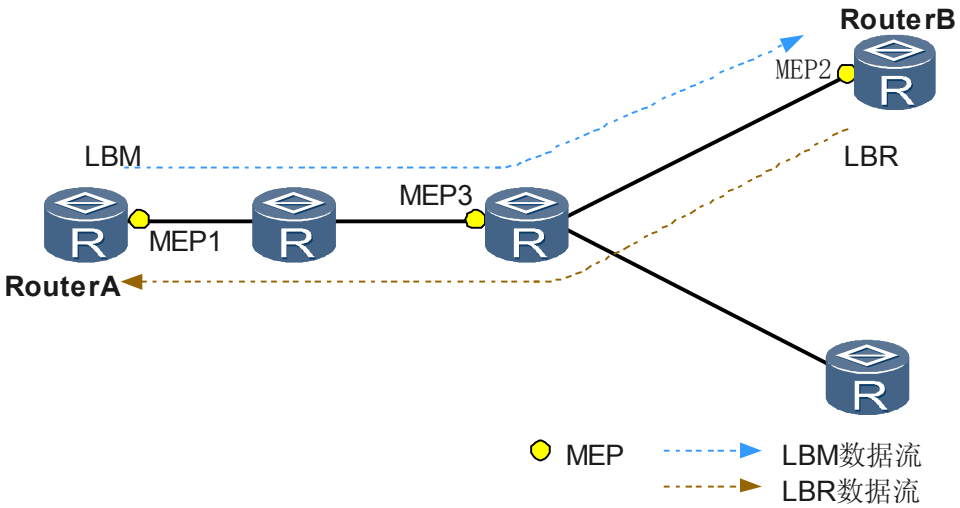
  CCM 由 MEP 产生也由 MEP 终结。当 MEP 接收到大于自身级别的 CCM 时，继续转发该 CCM；当 MEP 接收到小于或等于自身级别的 CCM 时，不再转发该 CCM，以确保低级别 MD 内的 CCM 不会扩散到高级别 MD 中。

802.1ag MAC Ping

802.1ag MAC Ping 与 Ping 类似，通过发送测试报文和接收应答报文来检测从本设备到目的设备是否可达。802.1ag MAC Ping 通过命令行触发，可用来确认本设备到目的设备的故障。

802.1ag MAC Ping 由 MEP 发起，目的节点可以是同一 MA 内的或不同 MA 内的，与发起节点级别相同的 MEP 或 MIP。

图 3-13 802.1ag MAC Ping 基本原理示意图



如图 3-13 所示，从 MEP1 向 MEP2 发起 802.1ag MAC Ping 探测，探测过程如下：

1. MEP1 向 MEP2 发送 LBM（Loopback Message）消息。消息内必须指定 MEP2 相关信息，可以是 MEP2 的主机 MAC 地址或 MEP2 的 MEP ID。
2. MEP2 接收到该 LBM 后，发送应答消息 LBR（Loopback Reply）。发起端会计算出 ping 操作的时间，用于分析网络性能。

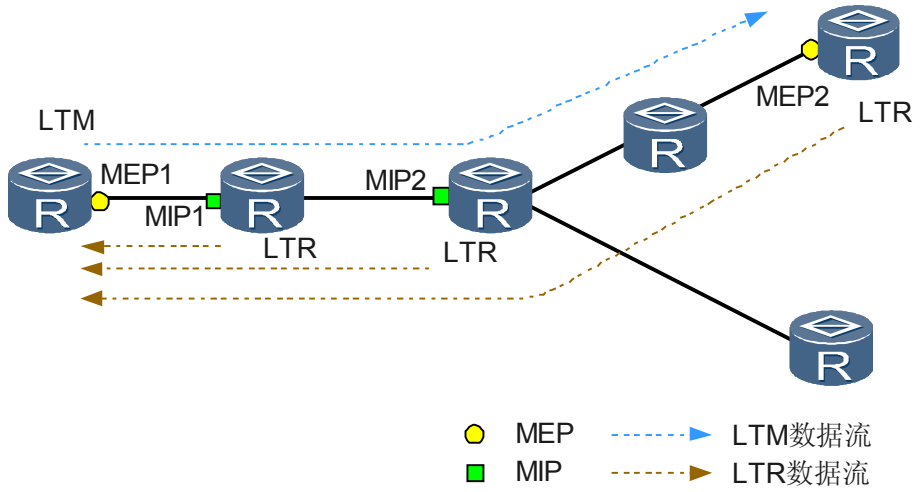
   在指定的超时时间内，如果 MEP1 没有收到 MEP2 回应的 LBR 消息，则认为 MEP1 和 MEP2 之间不可达；如果 MEP1 收到了 MEP2 回应的 LBR 消息，根据其中携带的时间戳信息计算出从 MEP1 到 MEP2 的网络时延。另外，可通过连续发送多个 LBM，观察 LBR 的返回情况，从而了解网络的丢包情况。

802.1ag MAC Trace

802.1ag MAC Trace 与 Traceroute（或 Tracert）类似，通过发送测试报文和接收应答报文来检测从本设备到目的设备的路径或定位故障点。

802.1ag MAC Trace 由 MEP 发起，目的节点可以是同一 MA 内的或不同 MA 内的，与发起节点级别相同的 MEP 或 MIP。

图 3-14 802.1ag MAC Trace 原理示意图



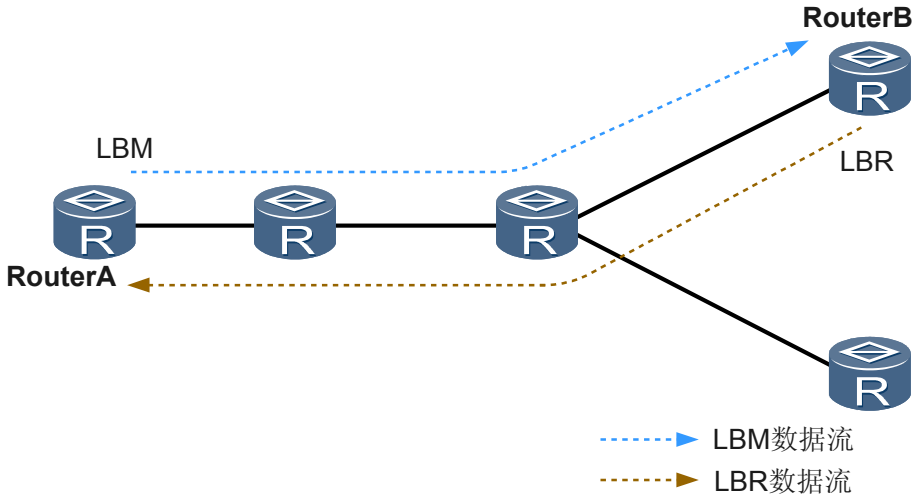
如图 3-14 所示，从 MEP1 向 MEP2 发起 8021.1ag MAC Trace 探测时，探测过程如下：

1. MEP1 向 MEP2 发送 LTM（Linktrace Message）消息。LTM 消息 中包含有 TTL（Time to Live）和目的节点 MEP2 的 MAC 地址。
2. 当 LTM 到达 MIP1 时，MIP1 将 LTM 中的 TTL 字段的值减 1，若此值为 0 不再转发，否则继续转发该 LTM。同时向 MEP1 回复 LTR（Linktrace Reply）。LTR 中还携带了分析报文路径的转发信息和收到的 LTM 报文的 TTL 字段。
3. MIP2 和 MEP2 收到 LTM 后，会做和 MIP1 相同的处理。但是，由于根据 LTM 中携带的目的节点 MAC 地址 MEP2 可以判断出自己是 LTM 的目的节点，因此 MEP2 不会再转发该 LTM。
4. MEP1 接收到 MIP1、MIP2、MEP2 回复的 LTR 后，根据 LTR 携带的信息即可得到从 MEP1 到 MEP2 的转发路径。

   如果 MEP1 到 MEP2 之间的路径有故障，则故障点下游的 MEP 或 MIP 将无法收到 LTM，也不会回复 LTR，可据此判定故障点的位置。例如当 MEP1 到 MIP2 之间的路径正常，而 MIP2 和 MEP2 之间的路径有故障时，MEP1 可以收到 MIP1、MIP2 回复的 LTR，但收不到 MEP2 回复的 LTR，于是可判定 MIP2 和 MEP2 之间的路径有故障。

通用 MAC Ping

图 3-15 通用 MAC Ping 原理示意图

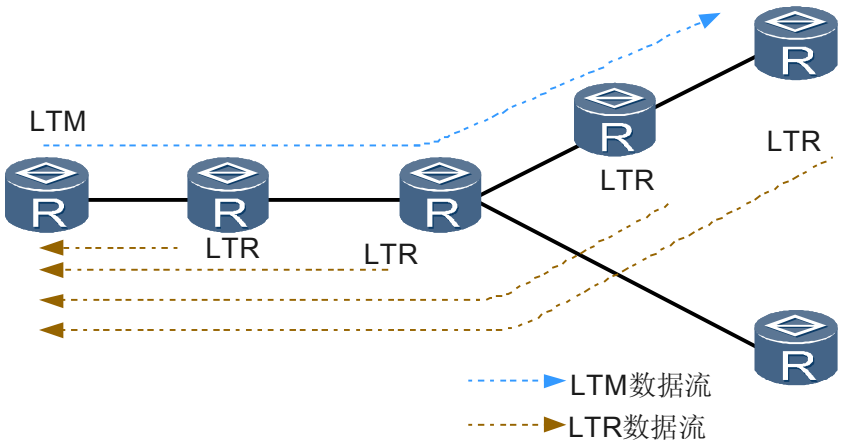


如图 3-15 所示，通用 MAC Ping 和 802.1ag MAC Ping 原理相似，但是通用 MAC Ping 不需要由 MEP 发起，目的节点也不需要是 MEP 或 MIP，即在源设备、中间设备和目的设备上均不需要配置 MD、MA 和 MEP 就可以进行通用 MAC Ping，只要中间设备上使能了 GMAC 功能。因此，在未配置 MD、MA 和 MEP 的网络（或网络的一部分）中，可直接使用通用 MAC Ping 进行连通性故障确认，而不需配置 MD、MA 和 MEP。

1. RouterA 向 RouterB 发送 LBM（Loopback Message）消息。消息内必须指定 RouterB 的主机 MAC 地址及该业务绑定的 VLAN ID 或 VSI ID。
2. RouterB 接收到该 LBM 后，发送应答消息 LBR（Loopback Reply）。发起端 RouterA 会计算出 ping 操作的时间，用于分析网络性能。

通用 MAC Trace

图 3-16 通用 MAC Ping 原理示意图



如图 3-16 所示，通用 MAC Trace 和 802.1ag MAC Trace 原理相似，但是通用 MAC Trace 不需要由 MEP 发起，中间节点和目的节点也不需要是 MEP 或 MIP，即在源设备、中间设备和目的设备上均不需要配置 MD、MA 和 MEP 就可以进行通用 MAC Trace，所有中间设备均回应 LTR 消息。

在未配置 MD、MA 和 MEP 的网络（或网络的一部分）中，可直接使用通用 MAC Trace 检测转发路径和定位故障，而不需配置 MD、MA 和 MEP。

1. RouterA 向 RouterB 发送 LTM（Linktrace Message）消息。LTM 中消息内必须指定 RouterB 的主机 MAC 地址及该业务绑定的 VLAN ID 或 VSI ID。
2. 当 LTM 到达 VLAN 或 VSI 内某节点，将 LTM 中的 TTL 字段的值减 1，若此值为 0 不再转发，否则继续转发该 LTM。同时向 RouterA 回复 LTR（Linktrace Reply）。当 LTM 到达 RouterB 时，回复 LTR 报文，终结 LTM 报文。
3. RouterA 接收到中间节点及 RouterB 回复的 LTR 后，根据 LTR 携带的信息即可得到从 RouterA 到 RouterB 的转发路径。

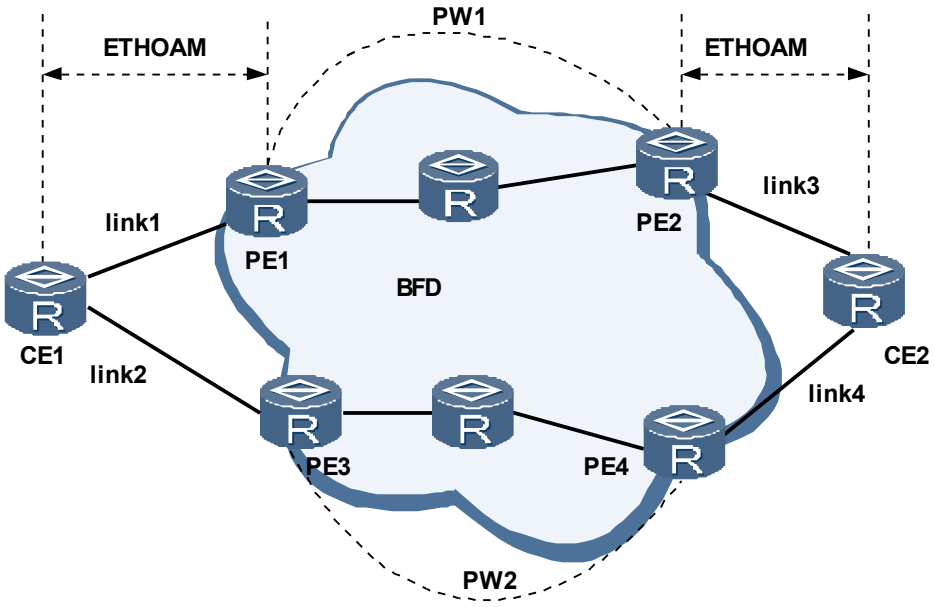
3.3.3 OAM 故障联动

以太网 OAM 与 BFD 间的故障联动

以太网 OAM（EFM OAM 或以太网 CFM）与 BFD 的故障联动是指以太网 OAM 模块检测到故障后，通过 PE 设备的 OAM 管理模块将故障消息传递给绑定的 BFD 会话实例，通过 BFD 再将故障消息传递到另一端的 PE；同时，BFD 检测到 PE 之间的故障也可以通过 PE 设备的 OAM 管理模块传递给以太网 OAM。

以太网 OAM 和 BFD 的故障联动主要用于 PE 间运行 BFD 检测，CE 设备和 PE 设备之间运行以太网 OAM 的场景。配置此功能后，两端的 CE 设备可以互相获知对端网络中 CE 设备和 PE 设备之间的连通性故障。

图 3-17 以太网 OAM 与 BFD 联动原理示意图



如图 3-17 所示：

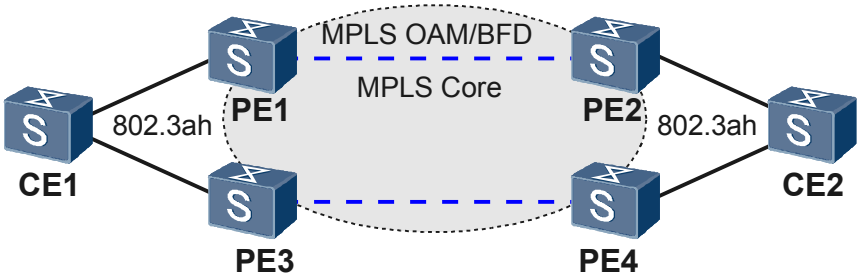
- PW1 和 PW2 相互独立，分别使用 BFD 进行故障检测。
- CE 使用双归方式接入到 PE1 和 PE2，PE 和 CE 间使用以太网 OAM 进行故障检测。

以太网 OAM 与 MPLS OAM 间的故障联动

EFM OAM 和 MPLS OAM 的故障联动是指 EFM OAM 模块检测到故障后，通过 PE 设备的 OAM 管理模块将故障消息传递给绑定的 MPLS OAM 会话实例，通过 MPLS OAM 再将故障消息传递到另一端的 PE。MPLS OAM 检测到 PE 之间的故障也可以通过 PE 设备的 OAM 管理模块传递给以太网 OAM。

EFM OAM 和 MPLS OAM 的故障联动主要用于 PE 间运行 MPLS OAM 检测，CE 设备和 PE 设备之间运行 EFM OAM 的场景。配置此功能后，两端的 CE 设备可以互相获知对端网络中 CE 设备和 PE 设备之间的连通性故障。

图 3-18 以太网 OAM 与 MPLS OAM 联动原理示意图



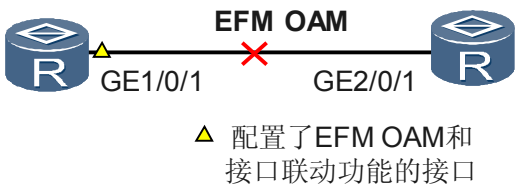
如图 3-18 所示：

- CE1 双归方式接入到 PE1 和 PE3，在 CE1 和 PE1、PE3 之间分别配置以太网 OAM
- PE 之间为 MPLS 网络，在 PE 之间的 MPLS LSP 上配置 MPLS OAM

EFM OAM 与接口联动

如图 3-19 所示，EFM OAM 与接口联动是指当某个运行 EFM OAM 的接口检测到和对端接口之间有连通性故障时，接口下除 EFM 协议报文外的所有其他报文都无法被转发，二三层业务将全部被阻塞。因此，一旦启用 EFM OAM 和当前接口联动功能，将可能对业务产生很大影响。如果当前接口通过 EFM OAM 检测到链路连通性故障恢复，则恢复该接口的所有转发报文，并通知该接口上的所有二三层业务作相应处理。

图 3-19 EFM OAM 和接口联动示意图



CFM OAM 与接口联动

以太网 CFM 与接口联动是指如果某个 MEP 检测到和同一 MA 内某个指定 RMEP 之间有连通性故障，OAM 管理模块对该 MEP 所在接口进行闪断处理，即先关闭该接口，然后再打开该接口，以使其它模块感知到该故障。

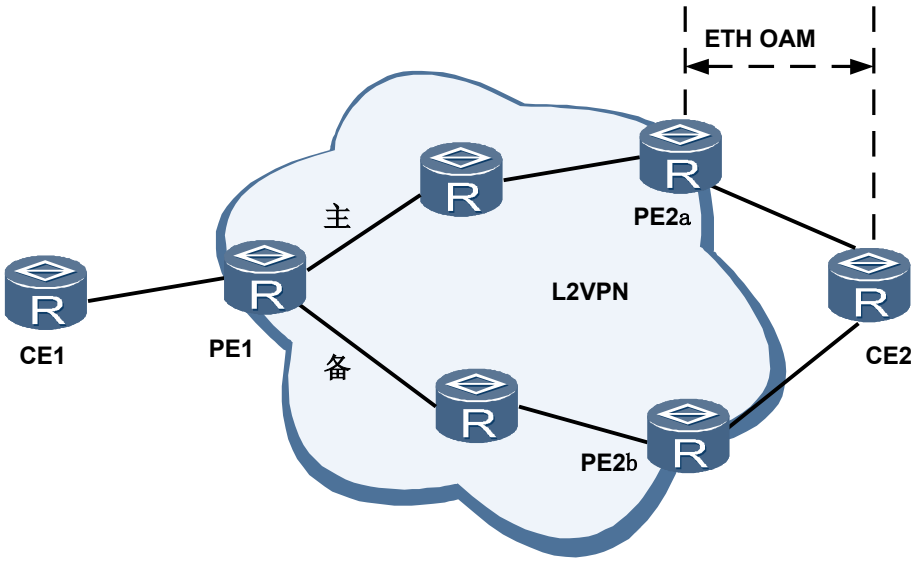
以太网 CFM 与接口联动功能主要用于 LACP 静态链路聚合组或手工 1：1 主备模式链路聚合的活动链路快速故障检测和触发保护倒换。

以太网 OAM 与清除 ARP 联动

CFM OAM 与清除 ARP 联动是指如果某个 MEP 检测到和同一 MA 内某个指定 RMEP 之间有连通性故障，OAM 管理模块清除该 MEP 所在接口的 ARP 相关表项，接口重新学习 ARP 表项。

EFM OAM 与清除 ARP 联动是指当某个运行 EFM OAM 的接口检测到和对端接口之间有连通性故障时，OAM 管理模块清除该接口与指定 VLAN 的 ARP 相关表项，接口重新学习 ARP 表项。

图 3-20 以太网 OAM 与清除 ARP 联动原理示意图



如图 3-20 所示：

- CE2 双归方式接入到 PE2a 和 PE2b，双归接口同时配置以太网 OAM 与清除 ARP 联动。
- PE 之间为 L2VPN 网络。

当以太网 OAM 检测到链路故障后会触发故障端口清除对应的 ARP 表项，使 CE2 设备从备用链路学习到新的 ARP 表项，并且将上行流量切换到备用路径。

3.3.4 OAM 安全

当网络不稳定时，如果以太网 CFM 已经启动故障连通性检测功能，在较短的时间内将有大量的告警和告警恢复产生。这些告警将会占用大量的系统资源，会影响到系统性能。因此，用户可以通过告警防抖动功能和告警抑制功能减少告警数量。

告警防抖动功能

说明

告警防抖动功能只支持在 standard2007 标准版本中使用。

- RMEP 激活时间

  RMEP 激活时间能够防止误告警的产生。它时间实际上是预留给用户对 RMEP 配置的时间。若本端设备上配置了 RMEP 的激活时间，当在本端设备上使能接收某个 RMEP 的 CCM 报文功能后，等待 RMEP 的激活时间到达后，本端才开始接收 CCM 报文。如果 RMEP 时间到达后，MEP 连续 3 个 CCM 周期没有收到 RMEP 发送的 CCM 报文，则认为发生了连通性故障，本端设备上将会产生连通性故障告警。
- 告警产生和告警恢复的防抖时间
  - 当 MEP 检测到连通性故障时：
    - 在告警产生防抖时间到后，上报告警。
    - 告警产生防抖时间内，如果告警恢复，则不会上报任何告警。
  - 当 MEP 检测到连通性故障并上报告警后：
    - 告警恢复的防抖时间内，如果 MEP 不再检测到连通性故障，则告警防抖时间到后上报告警恢复。
    - 告警恢复的防抖时间内，如果 MEP 再次检测到连通性故障，则不会上报告警恢复。

告警抑制功能

在网络中，多种故障可能同时产生，因此，多种告警也可能同时产生。CFM 的告警抑制功能的原理是：当多种告警同时产生时，只上报最高级别的告警。当最高级别的告警恢复后，如果低级别的告警仍然存在，则继续上报当前存在的最高级别的告警。如此反复，直至系统中不再有告警。

告警抑制的作用在于：

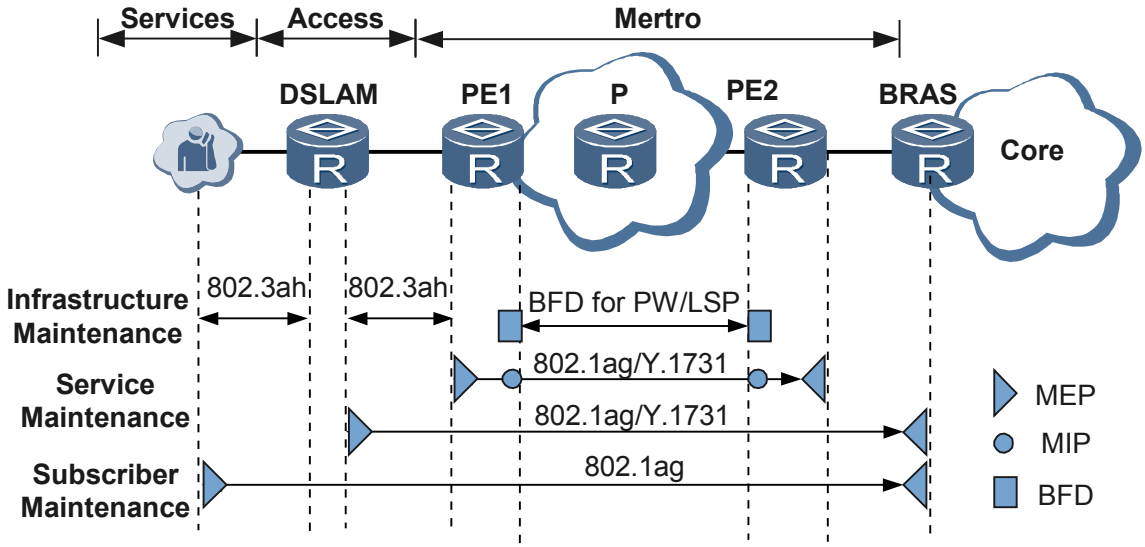
- 一般情况下，高级别的告警对应网络中比较严重的问题，需要用户优先解决。
- 同一个问题可能产生多种告警，这些告警的级别可能不相同。当最高级别的告警恢复后，当前系统中由相同问题产生的所有低级别的告警可能都会恢复。

3.3.5 Y.1731

Y.1731 是由 ITU-T 标准组织提出的 OAM 协议，它不仅包含 IEEE802.1ag 所规定的内容，而且又增加了更多的 OAM 消息组合，包括 AIS（Alarm Indication Signal），RDI（Remote Defect Indication），锁信号 LCK(Locked Signal)，测试信号，自动保护切换 APS（Automatic Protection Switching），维护通信渠道 MCC（Maintenance Communication Channel），试验 EXP（Experimental OAM），供应商特定的 VSP（Vendor Specific OAM）故障管理以及用于性能监视的丢包管理 LM（Loss Measurement）和延迟评估 DM（Delay Measurement）等。

如图 3-21 所示，Y.1731 一般部署于 PE 节点设备，用于端到端的业务故障快速检测和性能监视。Y.1731 支持与 802.1ag、802.3ah、BFD 等检测协议联动，能够将检测到的缺陷传递到远端设备或者接收远端设备传递来的缺陷通告，从而进行业务切换。

图 3-21 Y.1731 典型应用场景



当用户认为购买的以太网隧道业务出现了服务质量问题，或者运营商对自己的网络进行例行的 SLA 监测时，可以使用 Y.1731 所定义的性能测试工具进行检测。例如通过发送和接收 CCM 报文来计算丢包率统计，以及双向帧的延迟和抖动。

单端丢包统计功能

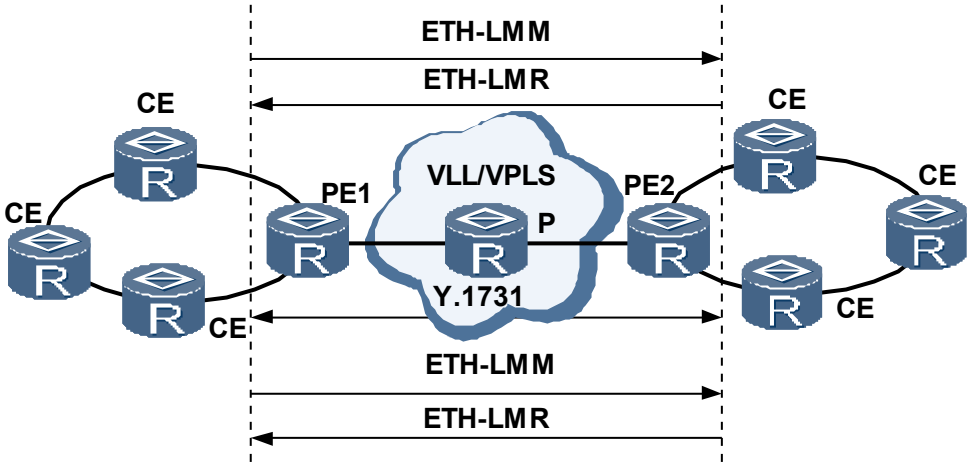
如果需要在一个指定的时间段内，对 VLL/VPLS 的丢包数量进行统计时，可以使用 VLL/VPLS 的单端丢包统计功能。目前 Y.1731 只支持基于 LDP 信令的 VLL/VPLS 进行丢包统计。

每个 MEP 统计结果包括：

- 近端丢包率：统计 Ingress 业务报文的丢失。
- 远端丢包率：统计 Egress 业务报文的丢失。

如图 3-22 所示，CE 设备分别通过 VLAN 方式接入 PE，PE1 和 PE2 之间建立使用 LDP 信令的 VLL/VPLS 连接。

图 3-22 基于 PE1 与 PE2 之间的指定 PW 单端丢包统计图



单端检测时，为进行丢包测量，PE1 设备的 MEP 向 PE2 的 MEP 发送带有 ETH-LM 请求信息的帧(ETH-LMM)，并从 PE2 MEP 接收带有 ETH-LM 回复信息的帧(ETH-LMR)。ETH-LMM 报文包含了报文发送时本地发送计数器 TxFCl 的值 TxFCf，PE2 收到后将回发 ETH-LMR 报文，包含如下信息：

- TxFCf：从 ETH-LMM 报文中复制的 TxFCf。
- RxFCf：ETH-LMM 报文接收时本地接收计数器 RxFCl 的数值。
- TxFCb：ETH-LMR 报文传输时本地计数器 TxFCl 的数值。

PE1 收到 ETH-LMR 帧后将使用如下数值进行近端和远端丢包测量：

- PE1 接收 ETH-LMR 帧的 TxFCf、RxFCf、TxFCb 的数值和该 LMR 帧接收时本地计数器 RxFCl 的数值分别被表示为 TxFCf[tc]、RxFCf[tc]、TxFCb[tc]和 RxFCl[tc]。

  这里的 tc 是当前 ETH-LMR 帧的接收时间。
- PE1 收到的前一个 LMR 帧的 TxFCf、RxFCf、TxFCb 的数值和该 LMR 帧接收时本地计数器 RxFCl 的数值分别被表示为 TxFCf[tp]、RxFCf[tp]、TxFCb[tp]和 RxFCl[tp]。

  这里的 tp 是前一个 ETH-LMR 帧的接收时间。

帧丢失(远端) = |TxFCf[tc] － TxFCf[tp]| － |RxFCf[tc] － RxFCf[tp]|

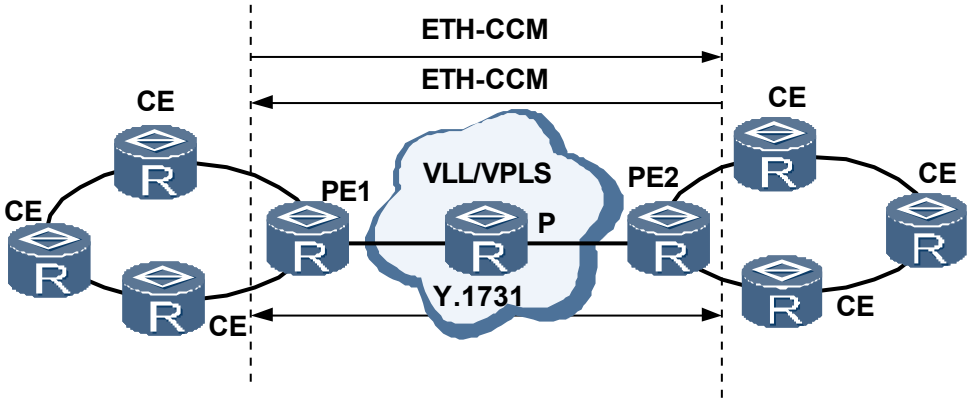
帧丢失(近端) = |TxFCb[tc] － TxFCb[tp]| － |RxFCl[tc] － RxFCl[tp]|

双端丢包统计功能

当需要对 VLL/VPLS 的丢包数进行在线不间断统计时，可以使用双端丢包统计功能。

如图 3-23 所示，CE 分别通过 VLAN 方式接入 PE，PE1 和 PE2 之间建立使用 LDP 信令的 VLL/VPLS 连接。

图 3-23 基于 PE1 与 PE2 之间的指定 PW 双端丢包统计图



双端检测时，MEP 向其远端 MEP 发送带有 ETH-LM 请求信息的双端的帧，以便于远端 MEP 进行丢包统计。每个 MEP 都终结对端发出的带有 ETH-LM 信息的双端的帧，进行近端和远端丢包统计。ETH-LM 信息的帧在这里就是 CCM 报文，报文中包含如下信息：

- TxFCf：在 CCM 帧传输时本地计数器 TxFCl 的数值。
- RxFCb：在从远端 MEP 接收到最后一个 CCM 帧时本地计数器 RxFCl 的数值。
- TxFCb：在从远端 MEP 接收到的最后一个 CCM 帧中的 TxFCf 的数值。

PE 收到带有 ETH-LM 信息的 CCM 帧后，将使用如下数值进行近端和远端丢包统计：

- 所接收 CCM 帧的 TxFCf、RxFCf、TxFCb 的数值和该 CCM 帧接收时本地计数器 RxFCl 的数值分别被表示为 TxFCf[tc]、RxFCf[tc]、TxFCb[tc]和 RxFCl[tc]。

  这里的 tc 是当前 CCM 的接收时间。
- 收到的前一个 CCM 帧的 TxFCf、RxFCf、TxFCb 的数值和该 CCM 帧接收时本地计数器 RxFCl 的数值分别被表示为 TxFCf[tp]、RxFCf[tp]、TxFCb[tp]和 RxFCl[tp]。

  这里的 tp 是前一个 CCM 帧的接收时间。

帧丢失(远端) = |TxFCb[tc] - TxFCb[tp]| - |RxFCb[tc] - RxFCb[tp]|

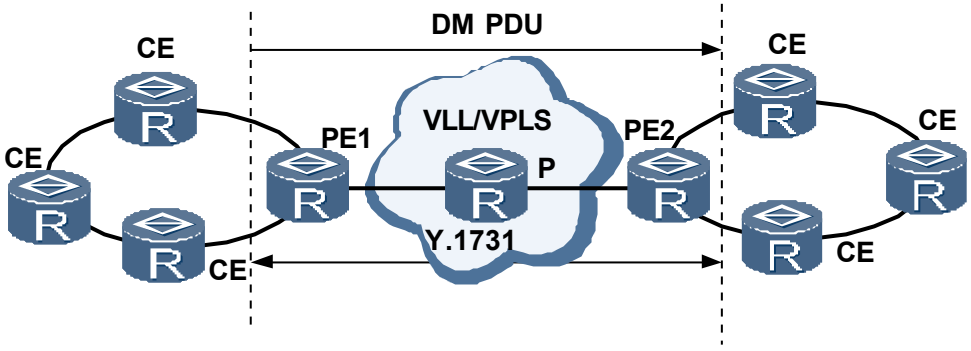
帧丢失(近端) = |TxFCf[tc] - TxFCf[tp]| - |RxFCl[tc] - RxFCl[tp]|

单向时延统计

如果需要在一个指定的时间段内，对 VLL/VPLS 的业务流量时延进行统计时，可以使用 VLL/VPLS 的单端时延统计功能。目前 Y.1731 只支持基于 LDP 信令的 VLL/VPLS 进行时延统计。

如图 3-24 所示，CE 分别通过 VLAN 方式接入 PE1 和 PE2，PE1 和 PE2 之间建立使用 LDP 信令的 VLL/VPLS 连接。

图 3-24 基于 PE1 与 PE2 之间的指定 PW 单端时延统计图



单向时延统计是在端到端 MEP 之间进行，通过发送/接收 DM 报文进行计算。当单向时延统计功能配置成功后，MEP 将周期性地发送带有 TxTimeStampf（DM 传输时的时间戳）数值的 DM 帧。RMEP 接收到 DM 报文后，解析出 DM 报文中 TxTimeStampf 值，与自身接收 DM 报文的时间 RxTimef 进行比较，得出单向时延值。计算公式如下：

帧时延 = RxTimef － TxTimeStampf

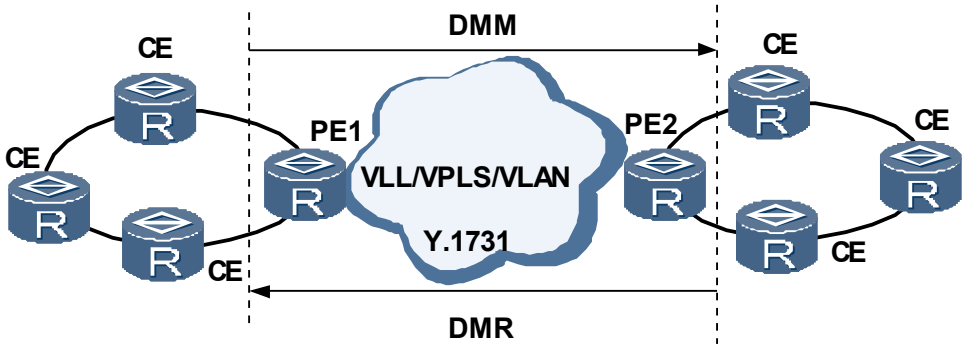
通过上面公式中的帧时延还可以计算出帧时延变化，即帧时延抖动。

时延抖动：是对一对业务帧之间帧时延变化的度量，本次时延与上次时延之间的绝对差值。

双向时延统计

如果需要在一个指定的时间段内，对双向的业务流量时延进行统计时，可以使用双端时延统计功能。

图 3-25 双向时延统计图



双向时延统计是在端到端 MEP 之间进行，通过接收 DMM 报文和发送 DMR 报文进行计算。当双向时延统计功能配置成功后，MEP 将周期性地发送带有 TxTimeStampf（DM 传输时的时间戳）数值的 DMM 帧。RMEP 接收到 DMM 报文后，填充 DMM 报文接收的时间戳值 RxTimeStampf，然后修改报文的类型为 DMR 报文，并交换报文的目的 MAC 和源 MAC 发送出去，同时带上发送时间戳 TxTimeStampb。当 DMM 发送端收到 DMR 时与接收到 DMR 报文的时间 RxTimeb 进行比较，得出双向时延值。计算公式如下：

帧时延 = (RxTimeb － TxTimeStampf) － (TxTimeStampb － RxTimeStampf)

通过上面公式中的帧时延还可以计算出帧时延变化，即帧时延抖动。

时延抖动：是对一对业务帧之间帧时延变化的度量，本次时延与上次时延之间的绝对差值。

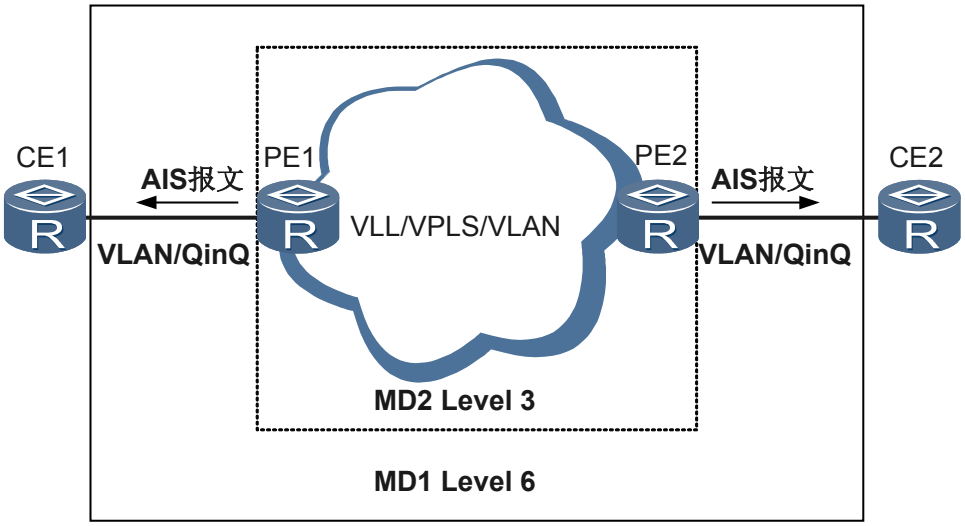
AIS 功能

AIS 是 Alarm Indication Signal 的简称。又被称为告警指示信号。它是一种用来传递故障信息的协议。

如图 3-26 所示，CE1 和 CE2 的接入端口配置的 MEP 在级别为 6 的 MD1 中，属于用户域网络，它对故障检测的时间要求相对较低。PE1 和 PE2 配置的 MEP 在级别为 3 的 MD2 中，属于运营商域，它对故障检测的时间要求相对较高。

- 当 PE 之间的 CFM 检测到连通性故障后，如果 PE 设备使能了 AIS 功能，则会向 CE 设备发送 AIS 报文。CE 设备接收到 AIS 报文后，可以抑制本设备的所有告警，从而减轻大量告警对网管设备的冲击。
- 当 PE 之间的链路恢复后，PE 设备将停止发送 AIS 报文。此时 CE 设备将不会再收到 AIS 报文，经过 3.5 倍 AIS 报文的发送周期后，本设备的告警抑制功能将会取消。

图 3-26 AIS 基本原理示意图



组播 Mac Ping 功能

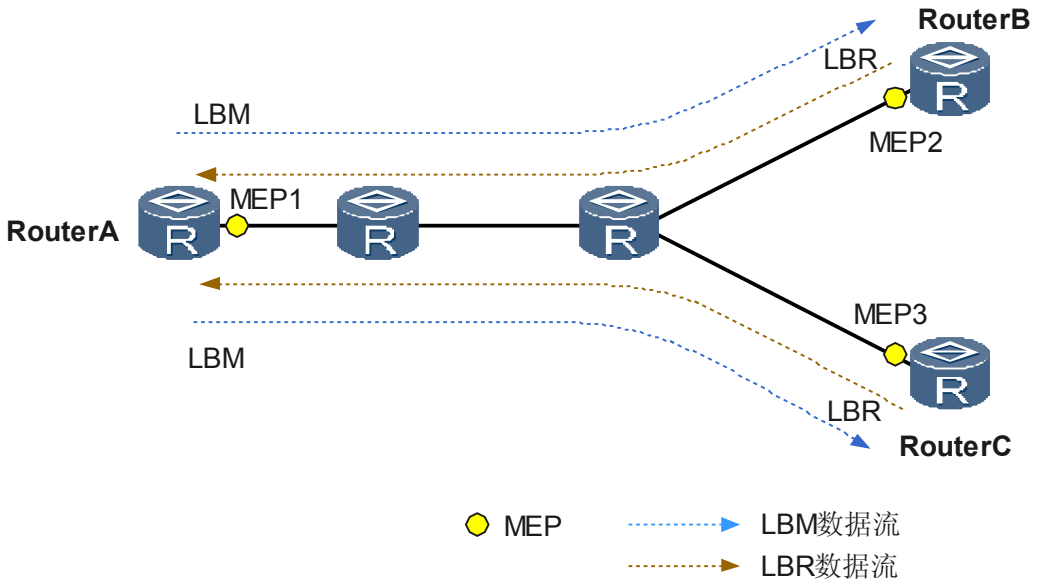
组播 Mac Ping 的目的 Mac 地址为组播 Mac 地址。组播 Mac Ping 主要有两个功能，一是故障确认功能，二是发现远端 MEP 功能。

- 功能一：故障确认功能

  组播 Mac Ping 提供与 802.1ag Mac Ping 类似的故障确认功能，通过发送测试报文来检测本设备和目的设备是否可达，但是组播 Mac Ping 能够一次确认多个 MEP 之间的连通性故障。

  一般情况下，组播 Mac Ping 功能应用在有多个 MEP 的场景下。它由本端 MEP 发起，由远端 MEP 回复。

图 3-27 组播 Mac Ping 基本原理示意图



如图 3-27 所示，RouterA 设备的 MEP1 存在两个远端 MEP：MEP2 和 MEP3。从 MEP1 发起组播 Mac Ping 探测，探测过程如下：

- 如果 MEP1 和远端 MEP2 之间或者 MEP1 和远端 MEP3 没有连通性故障:
  1. 本端设备的 MEP1 发送组播 LBM 报文。
  2. 远端 MEP2 和远端 MEP3 接收到该 LBM 报文后，会回复应答报文 LBR。
  3. 本端设备接收并显示应答报文 LBR 的内容。
- 如果 MEP1 和远端 MEP2 之间或者 MEP1 和远端 MEP3 存在连通性故障:
  1. 本端设备的 MEP1 发送组播 LBM 报文。
  2. 在指定的超时时间内，MEP1 没有收到远端 MEP 回复的应答报文，则认为 MEP1 和远端 MEP 之间出现了连通性故障。通过查看显示结果即可确定 MEP1 与远端 MEP2 或者是远端 MEP3 之间发生了连通性故障。

- 功能二：发现远端 MEP 功能

  如图 3-27 所示，用户需要在 RouterA 的本端 MEP1 上配置远端 MEP。由于在此情况下，用户可能不知道远端 MEP 的 ID 和 Mac 地址。此时可以配置 Mac Ping 功能，来发现远端 MEP。主要过程如下：

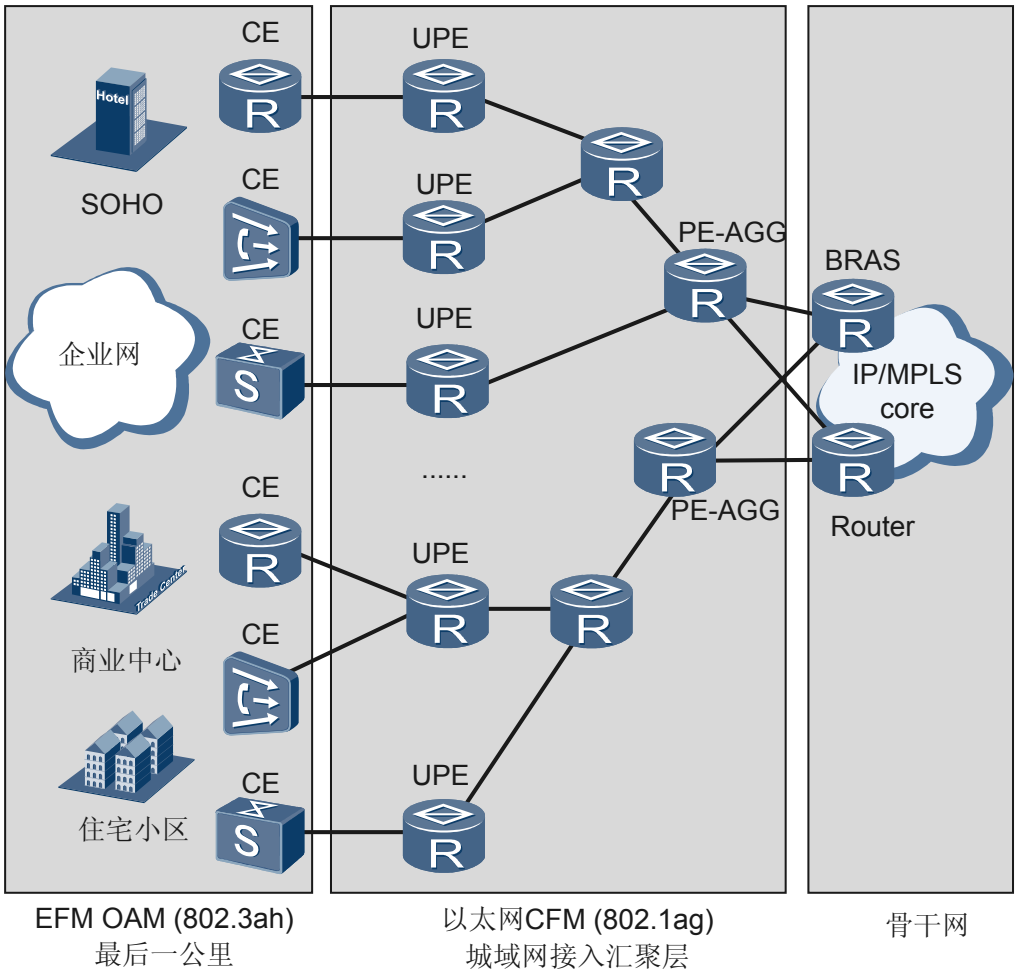
  1. 本端设备的 MEP1 发送组播 LBM 报文。
  2. 远端 MEP2 和远端 MEP3 接收到该 LBM 报文后，会回复应答报文 LBR。
  3. 本端设备接收并显示应答报文 LBR 的内容。该报文中包括远端 MEP 的 Mac 地址、MEP ID 和时延。

3.3.6 协议的比较

表 3-3 EFM OAM 与以太网 CFM 的比较

协议	特点
EFM OAM	EFM OAM（Ethernet in the First Mile OAM），针对两台直连设备之间的链路，提供链路连通性检测功能、链路故障监控功能、故障通知功能、远端环回功能。如图 3-28 所示，在城域网中，EFM OAM 技术多应用在 CE（Customer Edge）设备和 UPE（Underlayer Provider Edge）设备之间，用于保证用户网络和运营商网络之间连接的可靠性和稳定性。
以太网 CFM	以太网 CFM（Connectivity Fault Management），针对网络实现端到端的连通性故障检测、故障通知、故障确认和故障定位功能。如图 3-28 所示，在城域网中，CFM OAM 技术多应用在接入汇聚层网络中，用于监测整个网络的连通性，定位网络的连通性故障，并可与保护倒换技术相配合，提高网络的可靠性。

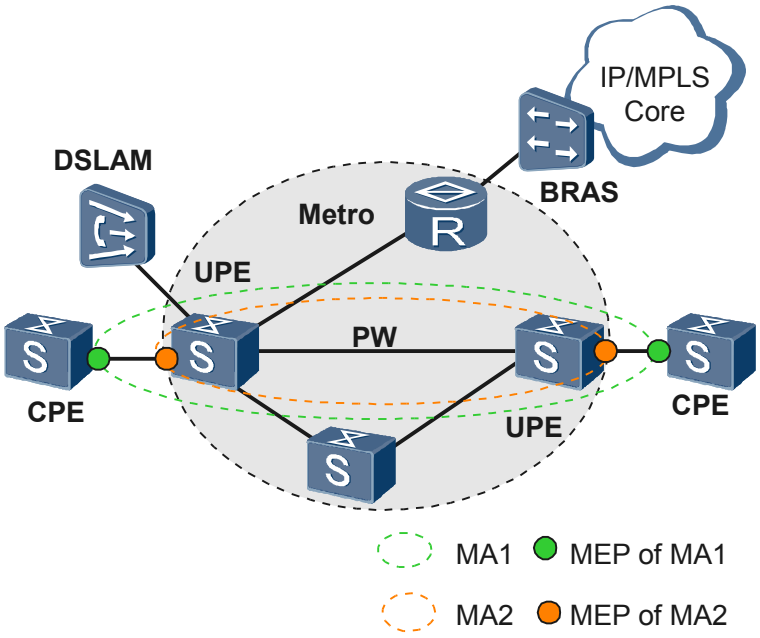
图 3-28 城域网典型组网



3.4 应用

E-LINE 业务的故障检测和性能检测

图 3-29 PW 隧道 E-LINE 业务



E-LINE 业务是由 MEF 6 标准定义的通用的业务类型，即点到点连接的称为以太网专线业务。

图 3-29 是利用 MPLS 技术建立起 PW 隧道。就整个业务通道考虑，隧道可视为一跳；就 Metro 网络考虑，利用 ETH-OAM，从一端 UPE 设备的 CPE 侧端口到远端 UPE 设备的 CPE 侧端口建立一个 MD，针对特定的用户业务来建立 MA，在 UPE 设备的 CPE 侧端口上建立 UP 型的 MEP，这样就可以从业务层面真正对 PW 隧道进行检测保护了（MA2）。

E-LAN 业务的故障检测和性能检测

图 3-30 多点到多点以太网

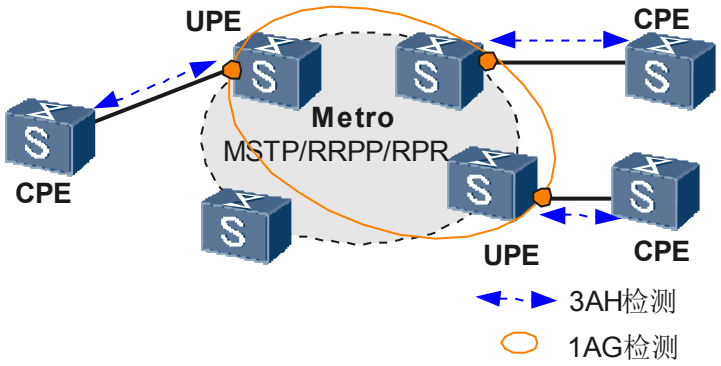
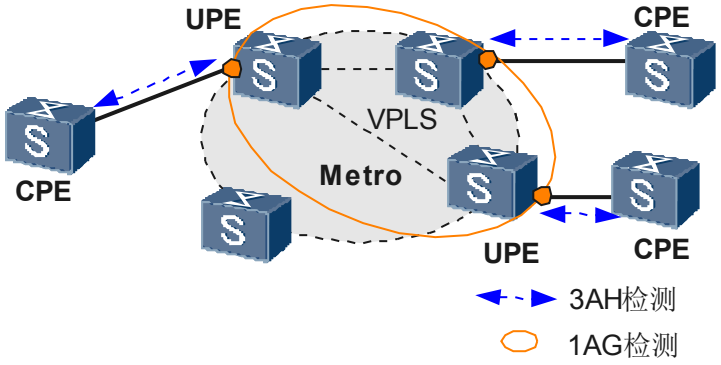


图 3-31 通过 VPLS 提供 E-LAN 业务



E-LAN 业务也是由 MEF 6 标准定义的通用业务类型，为用户提供多点到多点的连接，这种业务可以用纯以太技术、VPLS、QinQ 等技术实现，网络中各节点需要学习 MAC 地址。

UPE 与 UPE 多点网络的故障检测通过 802.1ag 在网络中发送组播 CC 报文实现，对于时延和抖动的检测只能是基于具体的两点之间。对于 CPE 与 UPE 之间，两点间的故障检测采用 802.3ah 协议，不存在多点对多点问题。

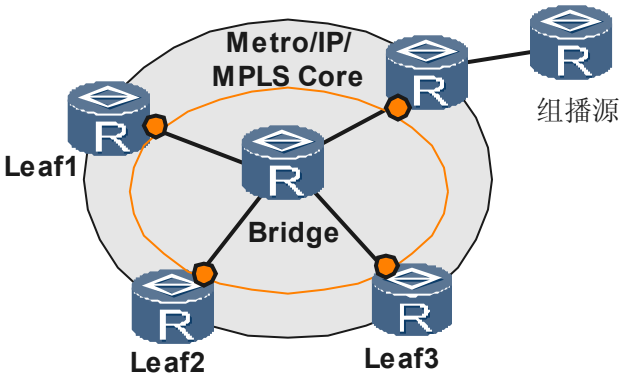
E-TREE 业务的故障检测

E-TREE 仍处于 MEF 草案状态，为用户提供点到多点的连接。E-TREE 业务从应用看多为组播业务，这种业务的特点是基本为单向流量，采用组播地址，可能存在跨 VLAN 复制操作。对于这样的业务，一般只要在叶子节点上对从根到叶子的链路进行故障检测，在叶子节点进行双归保护即可。

- 简单单向点到多点网络的故障检测：

如图 3-32 所示，网络中不存在跨 VLAN 复制的应用，各叶子节点只关心自己到 Root 节点的链路状态，不涉及其它叶子节点；同时 Root 节点不需要考虑到任何叶子的链路状态。因此可以只建立一个 MA，在所有叶子节点和根节点都设置 MEP，在根节点指明所有叶子为远端 MEP，使能 CC 发送，但不使能 CC 接收告警；在所有叶子节点只指明根节点为远端 MEP，不使能 CC 发送，只使能 CC 接收告警。这样通过简单配置即可完成该网络的 OAM 检测功能。

图 3-32 单向点到多点的 E-TREE 业务



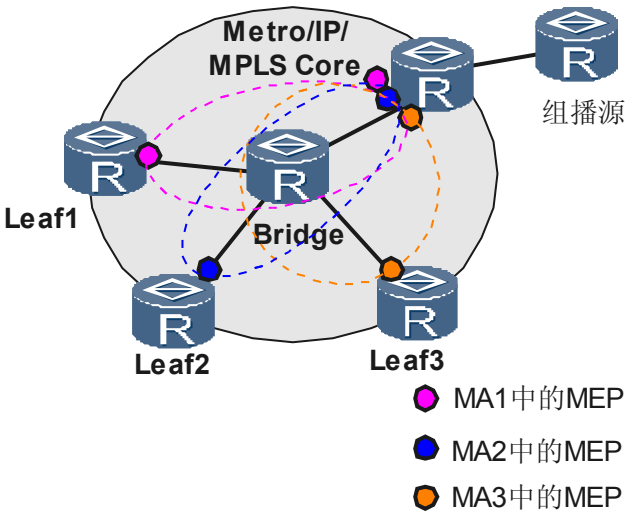
- 多 VLAN 单向点到多点网络的故障检测：

  中间桥实现跨 VLAN 复制。这时只需要建立基于 VLAN 组的 MA，其它处理与上一个场景相同；也可以简单的将其分段检测，并进行不同 MA 间的故障传递。

- 双向点到多点网络的故障检测：

  可能存在这样的组网，如图 3-33 所示，根节点与叶子节点之间存在通讯，叶子与叶子也存在通讯，但是叶子之间的通讯要通过根节点来进行中转。对此，可就根节点与不同叶子节点之间分别建立 MA，在根节点处将多个 MEP 关联起来，并进行告警传递即可。

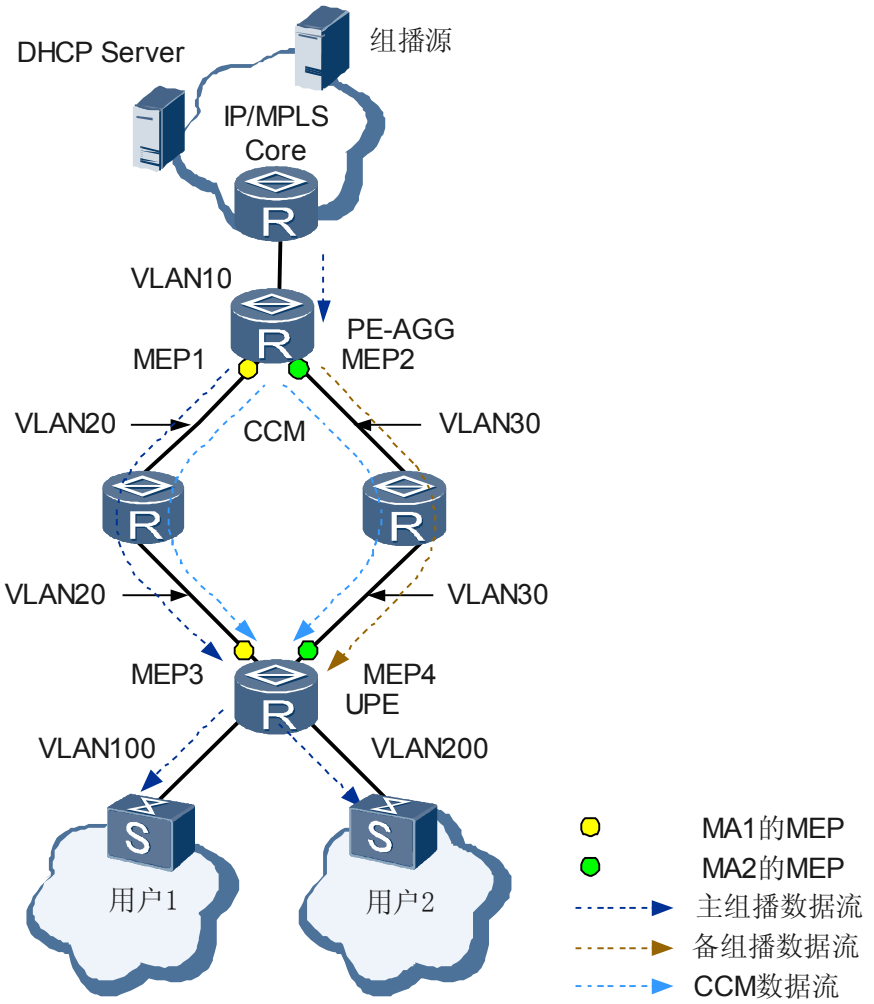
图 3-33 双向点到多点的 E-TREE 业务



组播 1+1 保护

以太网 CFM 可用于工作 VLAN 和保护 VLAN 的连通性故障检测，以及触发保护倒换，从而实现组播 1+1 保护功能。这里的组播指的是二层组播。

图 3-34 组播 1+1 保护



如图 3-34 所示，PE-AGG（PE-Aggregation）设备和 UPE 设备之间跑二层业务，PE-AGG 设备将组播流复制到两个 VLAN 中转发，这两个 VLAN 分别称为工作 VLAN（VLAN20）和保护 VLAN（VLAN30）。通常情况下，UPE（Underlayer PE）设备接收工作 VLAN 内的组播数据，丢弃保护 VLAN 内的组播数据。当工作 VLAN 出现连通性故障时，UPE 设备切换为接收保护 VLAN 内的组播数据。

工作 VLAN 和保护 VLAN 的连通性故障检测通过以太网 CFM 的 CC 检测来实现。在 PE-AGG 和 UPE 上各创建两个 MA，这两个 MA 分别对应工作 VLAN 和保护 VLAN。在每个 MA 内各创建一个 VLAN 型的 MEP，并使能 PE-AGG 上的两个 MEP（MEP1 和 MEP2）的 CCM 发送功能，使能 UPE 接收 MEP1 和 MEP2 发送的 CCM。

在指定时间内，如果 MEP3 接收不到 MEP1 发送的 CCM，则认为工作 VLAN 出现连通性故障。如果此时，MEP4 可以接收到 MEP2 发送的 CCM，则 UPE 切换为接收保护 VLAN 内的组播数据。

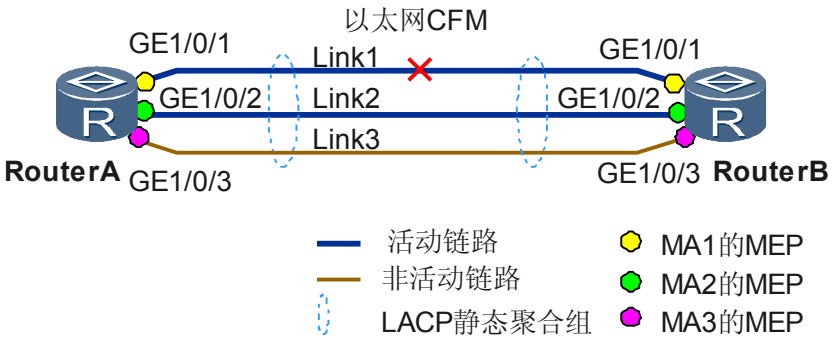
链路聚合故障检测和保护倒换

- 静态 LACP 聚合组故障检测和保护倒换

  由于 LACP 检测链路故障的时间最快为 3 秒，不能满足电信级网络保护倒换的速度小于等于 50 毫秒的要求。在可靠性要求较高的网络中，可在静态 LACP 聚合组成员链路两端的接口上配置以太网 CFM，同时配置以太网 CFM 与接口联动功能，借助以太网 CFM 的连通性故障快速检测功能和以太网 CFM 与接口联动功能，使静态 LACP 聚合组的保护倒换时间小于等于 50 毫秒，满足电信级网络的要求。

  如图 3-35 所示，RouterA 和 RouterB 配置了 LACP 静态链路聚合组。

图 3-35 静态 LACP 聚合组故障检测和保护倒换



在 RouterA 和 RouterB 上配置以太网 CFM 功能：在链路聚合组的所有成员接口上配置 MEP，并使同一条链路两端接口上的 MEP 位于同一个 MA 内，不同链路两端接口上的 MEP 位于不同 MA 内，所有接口上的 MEP 位于同一 MD 内。通过同一链路两端的 MEP 互发 CCM 消息检测链路的连通性。并配置以太网 CFM 和接口联动。

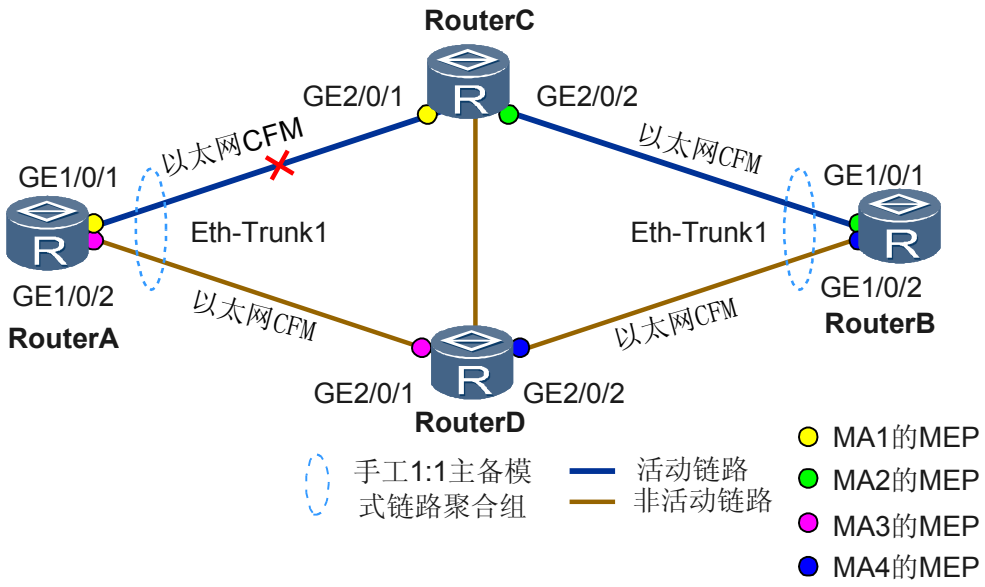
当 Link1 出现连通性故障时，RouterA 和 RouterB 的 OAM 管理模块分别对自己的 GE1/0/1 接口进行闪断。于是，LACP 模块就可以感知到 Link1 出现故障，并将通过该链路转发的业务数据切换到非活动链路 Link3 上转发。

- 手工 1：1 主备模式链路聚合组故障检测和保护倒换

  借助以太网 CFM 的 CC 检测功能以及以太网 CFM 和接口联动功能，可使手工 1：1 主备模式链路聚合组的保护倒换时间小于等于 50 毫秒，满足电信级网络的要求。

  如图 3-36 所示，RouterA 和 RouterB 配置了手工 1：1 模式主备链路聚合组。在 RouterA 和 RouterC、RouterA 和 RouterD、RouterB 和 RouterC、RouterB 和 RouterD 之间分别配置以太网 CFM，并在 RouterA 上配置以太网 CFM 和接口 GE1/0/1、GE1/0/2 联动，在 RouterB 上配置以太网 CFM 和接口 GE1/0/1、GE1/0/2 联动。

图 3-36 手工 1：1 主备模式链路聚合组故障检测和保护倒换

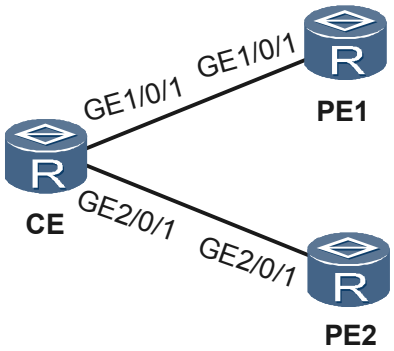


假设 RouterA 和 RouterC 之间的链路出现连通性故障，以太网 CFM 模块检测到该故障并通知 OAM 管理模块。OAM 管理模块对 GE1/0/1 接口进行闪断，以使链路聚合组感知到该故障并将业务数据切换到链路 A-D-C-B 上转发。其它链路的连通性故障的处理过程和以上 RouterA 和 RouterC 之间的链路故障处理过程类似。

EFM OAM 扩展功能

如图 3-37 所示，CE 为能够支持 EFM OAM 功能的 IP 接入设备，同时具备发送带有主备状态的 EFM 信息报文的能力。

图 3-37 EFM 扩展功能典型应用的组网图



在上述组网图中，采用了如下的方案配置：

- CE（IP 业务设备）双归接入 PE。
- CE 向 PE1 和 PE2 分别发送带有主和备状态标志的信息报文，控制链路的主用和备用关系，从而决定业务流量的走向。

- PE1 和 PE2 上同时配置了 EFM OAM 和接口联动或者 EFM OAM 和静态路由联动。

正常情况下，CE 向 PE1 发送主用状态标志的信息报文，且 PE1 上的接口状态为 Up，则静态路由存在，上下行业务流量通过 PE1 转发，PE2 上没有流量。

当 CE 和 PE1 之间的链路出现故障，CE 感知故障后，向 PE1 发送备用状态标志的信息报文，向 PE2 发送主用状态的信息报文。这样，PE1 与 CE 相连的接口状态变为 Down 或删除相关静态路由。PE2 与 CE 相连的接口状态变为 Up，生成相关静态路由，业务流量通过 PE2 转发。通过这种方式控制流量的走向，做到自动切换。

3.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
EFM	Ethernet in the First Mile	以太最后一公里
OAM	Operation Administration & Maintenance	运维管理
MPLS	Multiprotocol Label Switching	多协议标签交换
BFD	Bidirectional Forwarding Detection	双向转发检测
MD	Maintenance Domain	维护域
MA	Maintenance Association	维护联合
MEP	Maintenance association End Point	维护终结点
MIP	Maintenance association Intermediate Point	维护中间结点
CCM	Continuity Check Message	连续性检测报文
LBM	Loopback Message	环回报文
LBR	Loopback Reply	环回回应
LTM	Linktrace Message	链路跟踪报文
LTR	Linktrace Reply	链路跟踪回应
PBT	Provider Backbone Transport	运营商骨干传输

4 APS

关于本章

介绍 APS 特性的原理及应用。

4.1 介绍

4.2 参考标准和协议
介绍 APS 涉及到的参考标准和协议

4.3 原理描述

4.4 应用

4.5 术语与缩略语

4.1 介绍

定义

APS 是 Automatic Protection Switching 的简称，又被称为自动保护倒换(以下简称 APS)，是一种冗余保护的技术。在出现链路故障时，APS 将通过 APS 保护字节〔即位于保护通道的复用段开销字节（MSOH，Multiplex Section Overhead）中的 K1/K2 字节〕发出保护倒换请求，并由对端设备给予倒换桥接应答的保护倒换机制。

目的

APS 是 SDH 网络中的一个固有特性，在 SDH 网络上已经有很长的应用历史。在部署移动承载网络时，路由器需要和 SDH 设备 ADM 或无线设备 RNC 对接，这些设备都支持 APS 功能。路由器原有的保护特性不能很好的完成对路由器与 ADM 或 RNC 之间通信通道的保护，而 APS 特性很好的解决了这个问题。

受益

业务的高可靠性给运营商带来了明显的收益：

- 通过 APS 的自动倒换，减少了人为干预节省人力成本。
- 通过 APS 的快速倒换，减少了网络中断时间提高了网络传输的可靠性。
- 通过提高网络的可靠性，使的用户接入成功率大大提高。

4.2 参考标准和协议

介绍 APS 涉及到的参考标准和协议

文档	描述
ITU G.783	Characteristics of synchronous digital hierarchy (SDH) equipment functional blocks。
ITU G.841	Types and characteristics of SDH network protection architectures。

4.3 原理描述

4.3.1 APS 的基本原理

APS（Automatic Protection Switching）原本是 SDH 网络固有的一个特性，在 SDH 网络上已经有很长的应用历史，由 ITU G.783 和 G.841 给予定义。 APS 是一种冗余保护机制，为了保护某个通道的业务，需要有冗余的备份通道存在。 SDH 作为一个支撑网络，为数据通信网络提供了组建大型高速网络的可能。比如 A B 两地的数据网络通过光纤将业务复用到 SDH 的净负荷中，就可以实现 A B 两地的连通。由于 SDH 网络本身具备 APS 自动保护功能， 所以路由器上的 APS 特性可以用来保护路由器接入 SDH 网络的接入链路部分，即路由器到 SDH 的 ADM（Add/Drop Multiplex，分插复用器）设备

之间的通道的安全。当路由器之间直接连接在一起时，也能够实现路由器间的 APS 保护功能。

APS 倒换信息的传递

APS 协议信息利用复用段开销〔MSOH〕中的 K1、K2 字节来传递，具体含义如下所述。

K1 字节的含义

- K1(5-8)：通道号，0 表示保护通道，1 ～ 14 表示工作通道(对于 1+1 方式恒为 1)，15 表示额外业务通道(仅对 1:N 方式)，
- K1(1-4)：倒换请求码，具体含义见表 4-1。

表 4-1 K1 字节 1 ～ 4bit 含义

比特（4－1）	条件，状态或外部请求	优先级
1111	保护封闭	高
1110	强制倒换	
1101	信号失效高优先级(1+1 时不使用)	
1100	信号失效低优先级	
1011	信号劣化高优先级(1+1 时不使用)	
1010	信号劣化低优先级	
1001	未使用	
1000	人工倒换	
0111	未使用	
0110	等待恢复	
0101	未使用	
0100	练习	
0011	未使用	
0010	反向倒换请求（只双端倒换使用）	
0001	不倒回（只单端倒使用）	
0000	无请求	低

K2 字节的含义

- K2(5-8)：通道号，取值同 K1(5-8)
- K2(4)：保护方式，1 表示 1:N 方式，0 表示 1+1 方式
- K2(1-3)：指明了操作模式或码，具体含义见表 4-2。

表 4-2 K2 字节 1 ～ 3bit 含义

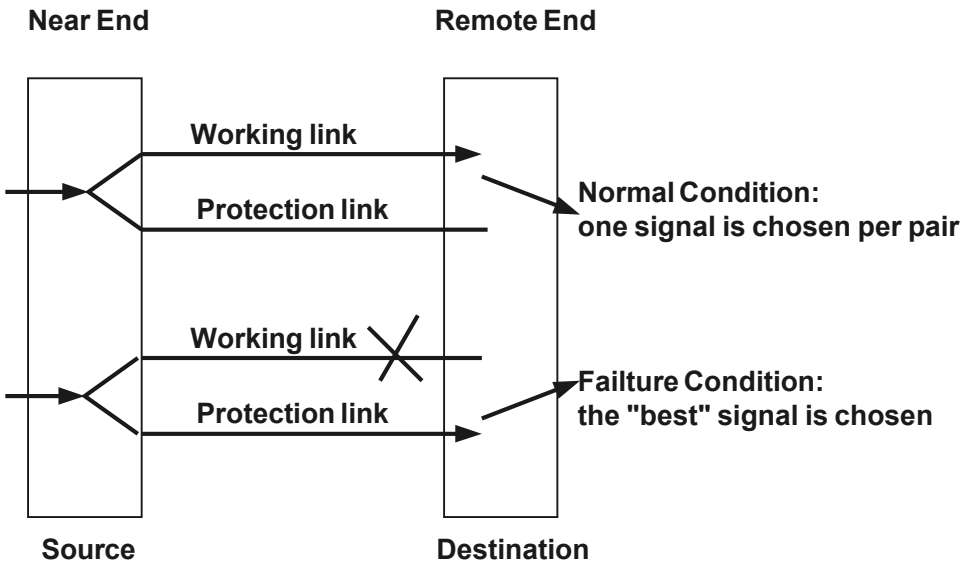
比特（3－1）	APS 操作模式，或其他线路信息
111	线路信号告警指示：Line AIS
110	线路远端错误指示：Line RDI
101	双端倒换模式
100	单端倒换模式
其他	保留

APS 模式的分类

- 根据保护结构分，可以分为 1+1 保护和 1:N (未实现 1:N，实现了 1:1)保护。
  - 1+1 保护是指每一条工作链路都有一条专有的保护链路为其提供备份。发送端在工作链路和保护链路上同时传输数据(此过程称为桥接：Bridge)，在正常情况下，接收端从工作链路上接收数据，当工作链路出现故障被接收端检测出来时，接收端将切换到保护接口上接收数据。一般情况下切换过程只在接收端进行动作，配合单端保护进行实现，不需要通过 K1K2 字节进行 APS 协商。

    它具有切换时间短，可靠性高等优点，但是缺点的是信道利用率低(50%)。具体过程如图 4-1 所示：

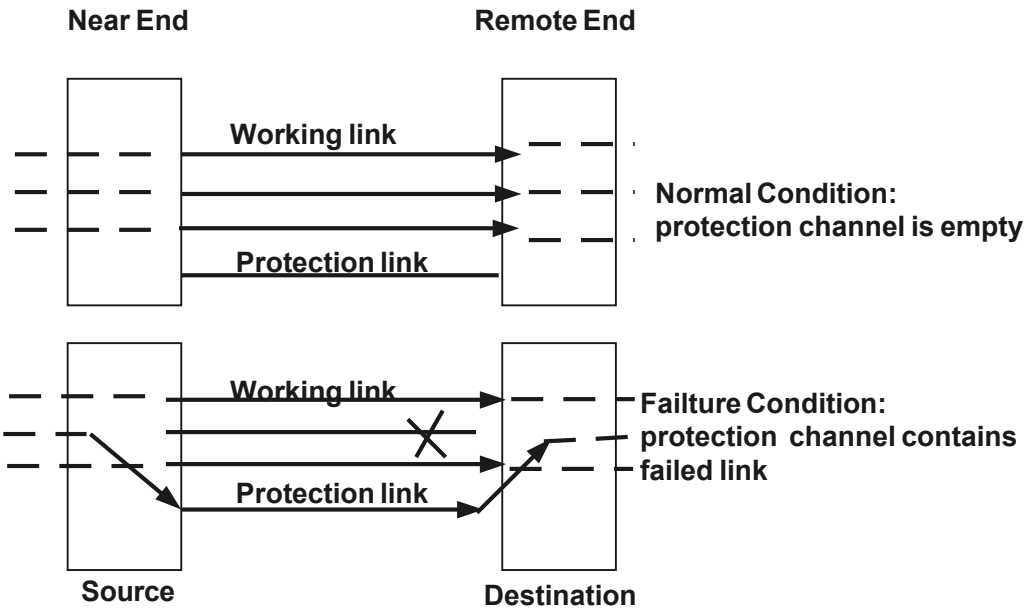
    图 4-1 1+1

  - 1:N 保护是指一条保护链路同时为 N 条工作链路提供保护(1<=N<=14)。 在正常情况下，发送端只在相应的工作链路上传输数据，保护链路上此时可以传输一些低优先级的数据(当然，也可以不进行数据的传输)。当工作链路出现故障时，发送端将要传输的数据桥接(Bridge)到保护链路上，接收端此时从保护链路上接收数据。如果保护链路上原本有低优先级数据在进行传输，此时，低优先级的

数据要让位于高优先级的被保护数据，也就不能进行传输了。实现过程如图 4-2 所示。

图 4-2 1:N



当同时有几条工作链路出现故障时，则根据工作链路的优先级，只有最高优先级链路上的数据可以倒换到保护链路上。其它出现故障的保护链路上的数据只能丢失。

特别的，当 N=1 时，就为我们所熟悉的 1:1 模式。

1:N 保护在保护过程中需要发送端和接收端同时进行切换，因此，这种保护需要通过 K1K2 字节进行协商。1:N 保护的优点是信道利用率高，但是可靠性不如 1+1 保护。

- 按照倒回模式分，可以分为倒回模式和非倒回模式。

  倒回模式是指工作链路恢复正常以后，过一段时间(这一段时间一般为几分钟到十几分钟)，待工作链路稳定后，在保护链路上的数据是否可以倒换回工作链路上。如果可以倒回，则称为这种保护为倒回模式；否则，则为非倒回模式。 1+1 保护默认为非倒回模式，也可以配置为倒回模式，1:1 保护只能配置为非倒回模式。

- 按照发生链路故障时，两端是否同时切换，可以分为单端倒换和双端倒换。
  - 单端倒换是指发生链路故障时，接收端检测故障发生切换，发送端未检测到故障不进行切换，只进行接收端的倒换桥接，切换的结果是 APS 连接的两端可能选择不同的链路接收流量。
  - 双端倒换是指发生链路故障时，接收端检测故障发生切换，发送端未检测到故障也需要通过 SDH 的 K 字节协商进行切换，切换的结果是 APS 连接的两端需要选择同一条链路进行发送接收。

  单端切换只能配合 1 ＋ 1 进行保护，双端切换即可以配合 1:1 也可以配合 1+1 进行保护。

- 按照配置 APS 的设备数目不同，可分为单机 APS 和双机 APS（即 E-APS）

4.3.2 APS 的实现方式

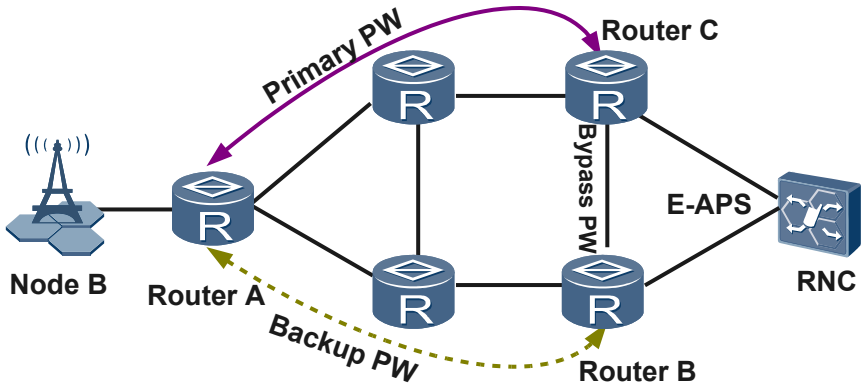
为了实现 APS 倒换对业务透明使物理层以上都不感知切换，APS 将通过逻辑接口（包括 Pos-Trunk、Cpos-Turnk（Trunk-Serial、Global-Mp-Group）、Atm-Trunk）来承载业务。通过配置物理接口加入 Trunk 接口，使用 Trunk 接口来承载业务，达到对物理层以上网络业务的透明，实现路由等模块不感知 APS 切换。但是，物理层属性还是要在物理接口上配置。

4.4 应用

在移动承载领域，RNC 通过两台路由器双规接入 IP 网络，路由器和 RNC 之间部署跨机框的 APS，即 E-APS。路由器通过部署主备两条 PW 穿越公网，通过 PW 透传 RNC 的数据到远端路由器。

当 RNC 和路由器间的主用通道发生故障时，业务会通过 APS 切换到 Router C 和 RNC 的通道，Router B 和 Router C 通过 Bypass PW 转接流量，给用户的关键业务提供高可靠性保障。详细组网如图 4-3 所示。

图 4-3 APS 与 PW 冗余联动



4.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称
APS	Automatic Protection Switching

缩略语	英文全称
SDH	Synchronous Digital Hierarchy

5 MPLS-TP OAM

关于本章

5.1 介绍

定义

MPLS-TP 是一种融合了 MPLS 包交换和传统传送网特性的传输技术，可以替代原有传送网作为未来的承载网。MPLS-TP OAM 可以有效检测、识别和定位 MPLS-TP 用户层面故障，在链路或节点出现缺陷或故障时迅速进行保护倒换。运维管理 OAM（Operation Administration & Maintenance）是降低网络维护成本的有效手段，MPLS-TP OAM 机制用于 MPLS-TP 层的运维管理。

目的

随着网络和业务的转型和融合，各种新兴的业务，例如三重播放、NGN、电信级以太网（Carrier Ethernet）、FTTx 等，都对单纯的分组传送网的投资成本、运维成本、QoS 保证、全业务接入、网络扩展性、网络可靠性和网络可管理性等提出了更高的要求。对比缺乏控制平面，不能适应这些新需求的传统传送网（MSTP，SDH，WDM）技术，MPLS-TP 具有传送网特性、又支持分组业务处理能力的下一代分组传送网来满足这些需求。

由于传统的传送网络（如 SDH/OTN）在可靠性和运维这两个方面树立了一个很高的基准，因此 MPLS-TP 需要提供完善的 OAM 能力。简单来说，MPLS-TP OAM 主要包括三个方面：

- 故障管理（Fault Management）
- 性能监控（Performance Monitoring）
- 保护倒换（Protection Switching）

5.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
Y.1731	This Recommendation provides mechanisms for user-plane OAM functionality in Ethernet networks according to the requirements and principles given in Recommendation Y.1730. This Recommendation is designed specifically to support point-to-point connections and multipoint connectivity in the ETH layer as identified in Recommendation G.8010.	
draft-bhh-mpls-tp-oam-y1731-04.txt	This document describes methods to leverage Y.1731 Protocol Data Units (PDU) and procedures (state machines) to provide a set of Operation, Administration, and Maintenance (OAM) mechanisms that meets the MPLS Transport Profile (MPLS-TP) OAM requirements. In particular, this document describes the MPLS-TP technology specific encapsulation mechanisms to carry these OAM PDUs within MPLS-TP packets to provide MPLS-TP OAM capabilities in MPLS-TP networks.	

文档	描述	备注
draft-ietf-mpls-tp-oam-framework-06.txt	This document describes a framework to support a comprehensive set of OAM procedures that fulfill the MPLS-TP OAM requirements.	

5.3 原理描述

5.3.1 MPLS-TP OAM 功能组件

MPLS-TP OAM 沿用[ITU-T Y.1731]所定义的一系列功能组件。这种定义方式已经为 IETF MPLS-TP 工作组所广泛采用，几乎在所有涉及 OAM 的 MPLS-TP 标准文稿中均使用下面的功能组件定义。

维护实体（Maintenance Entity）与维护实体组（Maintenance Entity Group）

MPLS-TP 的 OAM 操作均基于维护实体 ME（Maintenance Entity）进行。一个 ME 可以简单地理解为一条传送路径（transport path）的两个端点，也就是一对 MEP（Maintenance Entity Group End Points）以及中间经过的 MIP（Maintenance Entity Group Intermediate Points）。OAM 功能主要就运行在这两个 MEP 之间，而这里所指的传送路径可以是点到点的（包括 PW 和 P2P LSP），也可以点到多点的 P2MP LSP。

一个或者多个属于同一条传送路径的 ME 构成一个维护实体组 MEG（Maintenance Entity Group）。对于 MPLS-TP OAM 来说，一个 MEG 对应点到点单向路径时，就只包含一个 ME；对应点到点双向路径时，则包含正反两个方向的两个 ME（如果是双向共路，则只包含一个 ME）；对应一个点到多点的单向路径时，就是由从根到各个叶子节点的各个 ME 组成。

如图 5-1 所示的 P2P 场景，A、B、C 和 D 可以是 LSP 的 LER/LSR 或者 MS-PW 的 T-PE/S-PE。MEP 位于 A 和 D，而 MIP 位于 B 和 C 节点。中间的链路可以是物理链路，下一层的 LSP 或者下层网络提供的传输路径。

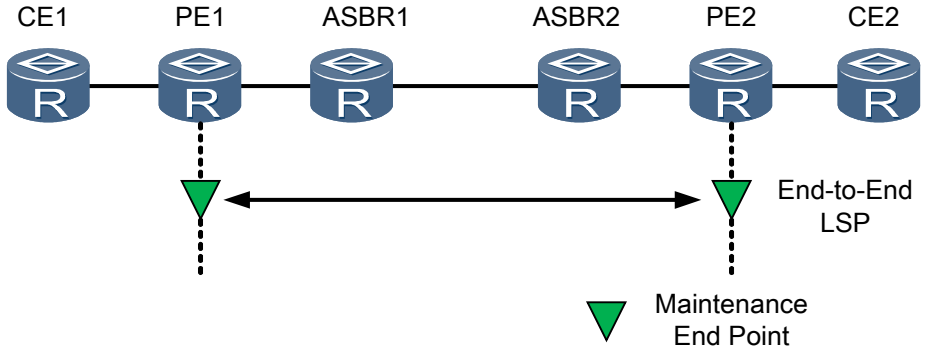
图 5-1 MEG 示意（P2P）

维护实体组端点（MEG End Point）

MEG End Point（MEP）是 MEG 的源、宿点。对于一个 MPLS-TP LSP 来说，只有 LER 才能作为 MEP；对于 MPLS-TP PW，只有 T-PE 可以作为 MEP。

例如，如图 5-2 所示，在端到端 LSP 上，只有 PE1 和 PE2 可以作为 MEP 节点。

图 5-2 MEP 概念示意



5.3.2 连通性（CC/CV）检测

CC（Continuity Check）和 CV（Connectivity Verification）实际上是两种不同的 MPLS-TP OAM 功能。CC 用于同一 MEG 的两个 MEP 之间的连续性缺陷检测 LOC（loss of continuity defect）。而 CV 用于检测两个 MEG 或者同一 MEG 的 MEP 之间错误的连通性缺陷。但是在实际应用中这两者通常是联系在一起，共同执行，所以把这两种检测功能放在一起。但是，CC 和 CV 在检测目的上存在很大区别。

CC 检测

双向共路 LSP 持续连通性检测（CC）作为一种主动性的 OAM，用于检测一个 MEG 中任何一对 MEP 间连续性的丢失。本端周期性的向目的端发送 CCM 报文，如果目的端在 3.5 个报文发送周期内没有收到 CCM 报文，则认为本端到目的端连通性有问题，上报告警，同时进入故障状态并触发 APS 倒换。当目的端重新收到本端发送的 CCM 报文，则恢复告警并且退出故障状态。

CV 检测

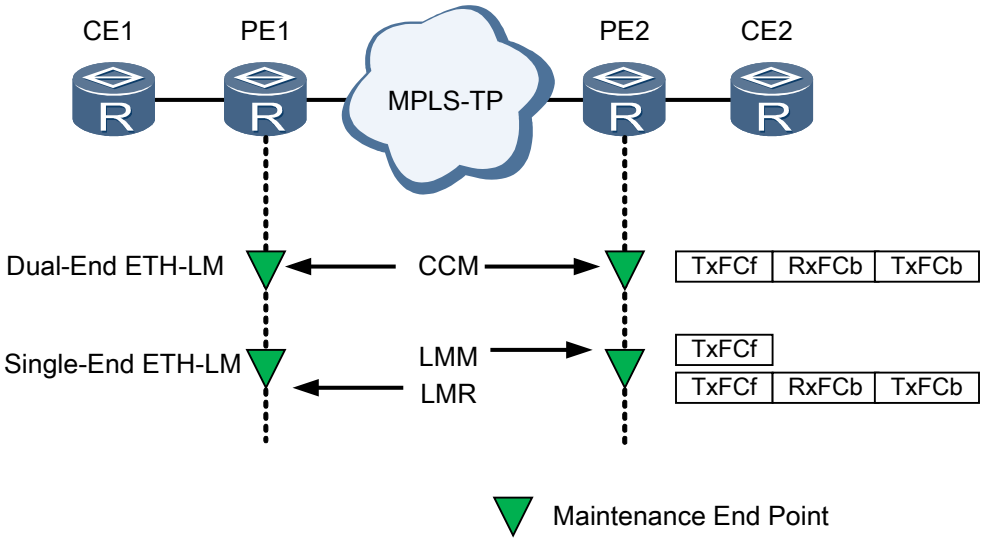
双向共路 LSP 转发故障检测（CV）同样是一种主动性的 OAM，MEP 在接收到非预期的报文时会上报告警。比如使能 CV 检测功能的设备从某条 LSP 收到报文，却发现该报文不是这条 LSP 应该承载的，此时会上报转发错误告警。

5.3.3 丢包率（LM）检测

丢包率度量 LM（Packet Loss Measurement）是 MPLS-TP 性能监控 PM（Performance Monitoring）的一项功能，在 PW、LSP 或 Section 的两个端点 MEP 进行。度量的数据包含两个方面，分别被称为近端丢包率和远端丢包率：近端丢包率表示本端 MEP 收方向（本端 MEP 接收从远端 MEP 发过来）的丢包统计，远端丢包率表示本端 MEP 发方向（远端 MEP 接收从本端 MEP 发过去）的丢包统计。为了计算各个方向的丢包统计，每个 MEP 本地均需维护两个基本的计数器：

- TxFCl：记录当前向对端 MEP 发送的报文数
- RxFCl：记录当前从对端 MEP 接收的报文数

图 5-3 Loss Measurement 功能示意



如图 5-3 所示，LM 有两种工作模式，双端（Dual-End）LM 和单端（Single-End）LM。双端 LM 只能工作于主动监控模式，两端 MEP 周期性地发送的 OAM 报文，携带如下信息：

- TxFCf：当前 OAM 报文发送时本地计数器 TxFCl 的值。
- RxFCb：收到上一个从对端 MEP 发来的 OAM 报文时本地计数器 RxFCl 的值。
- TxFCb：收到上一个从对端 MEP 发来的 OAM 报文时的 TxFCf 值（也就是当时本地计数器 TxFCI 的值）。

当两端 MEP 分别收到从对端发来的携带计数信息的 OAM 报文，通过比较如图 5-4 所示的两组数据，即可进行近端和远端丢包度量：

图 5-4 双端 LM 计算公式

$$\text{Frame Loss}_{\text{far-end}} = |\ TxFCb[t_c] - TxFCb[t_p]\ | - |\ RxFCb[t_c] - RxFCb[t_p]\ |$$

$$\text{Frame Loss}_{\text{near-end}} = |\ TxFCf[t_c] - TxFCf[t_p]\ | - |\ RxFCl[t_c] - RxFCl[t_p]\ |$$

TxFCf[tc]、RxFCb[tc]和 TxFCb[tc]分别表示当前收到对端发来的 OAM 报文中的 TxFCf、RxFCb 和 TxFCb 值，RxFCl[tc]表示收到该报文时本地 RxFCI 中的计数，tc 表示收到 OAM 的报文的当前时间。

TxFCf[tp]、RxFCb[tp]和 TxFCb[tp]分别表示收到上一个对端发来的 OAM 报文中的 TxFCf、RxFCb 和 TxFCb 值，RxFCl[tp]表示收到上一个报文时本地 RxFCI 中的计数，tp 表示收到上一个 OAM 的报文的时间。

单端 LM 一般用作按需监控的 OAM，在此模式下，只是本端 MEP 周期性地向对端 MEP 发送 LM 请求报文，对端 MEP 回应 LM 相应报文，携带如下信息：

- TxFCf：当前 LM 请求报文发送时本地计数器 TxFCl 的值

当对端 MEP 接收到一个正确的 LM 请求报文时需回应一个 LM 响应报文，该报文携带如下信息：

- TxFCf：从 LM 请求报文中拷贝的 TxFCf 值
- RxFCf：收到 LM 请求报文时远端 MEP 本地计数器 RxFCl 的值

- TxFCb：远端 MEP 发送 LM 响应报文时本地计数器 TxFCl 的值

于是当本端 MEP 收到对端 MEP 发来的 LM 响应报文，通过比较如图 5-5 所示的两组数据，即可进行近端和远端丢包度量：

图 5-5 单端 LM 计算公式

$$\text{Frame Loss}_{\text{far-end}} = | TxFCf[t_c] - TxFCf[t_p] | - | RxFCf[t_c] - RxFCf[t_p] |$$

$$\text{Frame Loss}_{\text{near-end}} = | TxFCb[t_c] - TxFCb[t_p] | - | RxFCl[t_c] - RxFCl[t_p] |$$

TxFCf[tc]、RxFCf[tc]和 TxFCb[tc]分别表示当前收到对端发来的 LM 响应报文中的 TxFCf、RxFCf 和 TxFCb 值，RxFCl[tc]表示收到该报文时本地 RxFCI 中的计数，tc 表示收到 LM 响应报文的当前时间。

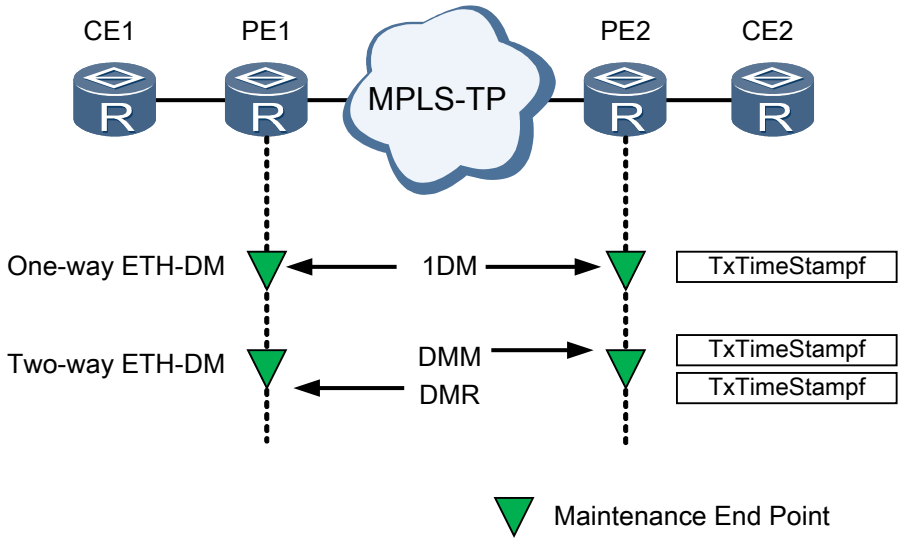
TxFCf[tp]、RxFCf[tp]和 TxFCb[tp]分别表示收到上一个对端发来的 LM 响应报文中的 TxFCf、RxFCf 和 TxFCb 值，RxFCl[tp]收到上一个报文时本地 RxFCI 中的计数，tp 表示收到上一个 LM 响应报文的时间。

5.3.4 时延（DM）检测

报文时延度量 DM（Packet Delay Measurement）同样也是 MPLS-TP 性能监控 PM（Performance Monitoring）中的一项功能。MPLS-TP OAM 首先应提供按需监控的 DM 功能，主动方式的 DM 可选。同时根据统计的时延信息，可监控出传送路径的报文时延抖动情况。

如图 5-6 所示，时延度量在两个端点 MEP 进行，包含单向 DM 和双向 DM：

图 5-6 Delay Measurement 功能示意



- 对于单向 DM，本端 MEP 周期性地发送 OAM 报文携带发送时的 TxTimeStampf（发送 DM 报文时的时间戳）。对端 MEP 收到 OAM 报文后，通过比对接收时间与 OAM 报文中的时间戳即可获得计算出报文时延：

  Frame Delay = RxTimef － TxTimeStampf （RxTimef 为接收到 OAM 报文的时刻）

  注意，执行 One-way DM 的两个节点间必须时间同步，否则计算出的时延不准，只能做时延抖动的度量。

- 对于双向 DM，本端 MEP 周期性地相对端 MEP 发送 DM 请求报文，该报文携带发送时的 TxTimeStampf。对端 MEP 收到 DM 请求报文后，本端 MEP 收到后，通过比对接收时间与 OAM 报文中的时间戳即可获得计算出报文时延：

  Frame Delay = RxTimeb － TxTimeStampf （RxTimeb 为接收到 OAM 报文的时刻）

  为了更精确地计算报文时延，排除掉对端节点处理 DM 报文所消耗的时间，回应报文也可引入另外两个值：RxTimeStampf，接收到 DM 请求报文的时刻；TxTimeStampb，发送 DM 回应报文的时刻 。这样当本端 MEP 接收到 DM 回应报文，做如下计算即可得出报文时延：

  Frame Delay = (RxTimeb－TxTimeStampf)－(TxTimeStampb－RxTimeStampf)

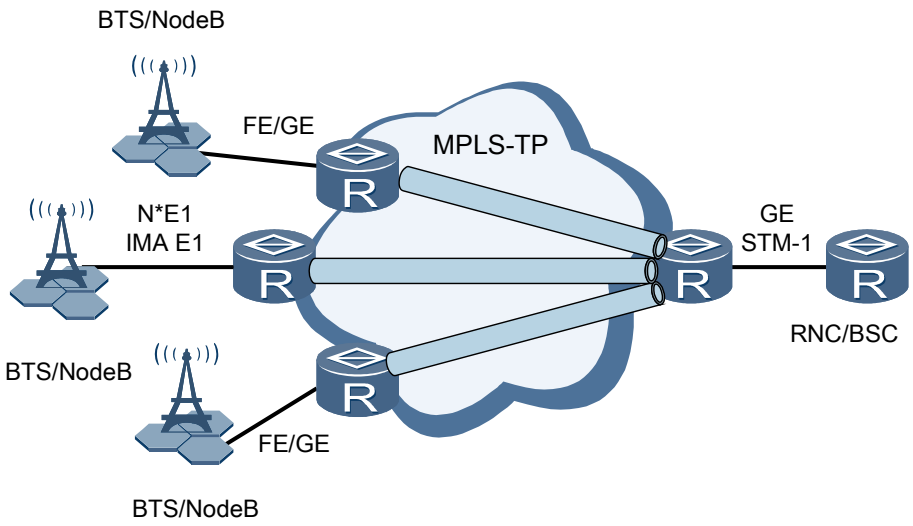
  Two-way DM 不需要两个节点间时间同步即可进行报文时延和抖动的度量。但是，这里计算出的时延是来、去两个方向上的报文时延之和。实际上，在两个节点时间同步的情况下，（RxTimeStampf -TxTimeStampf-）和（RxTimeb- TxTimeStampb）即分别为来、去两个方向上的报文时延（Two-way DM 就是两个方向上 One-way DM 之和）。

5.4 应用

5.4.1 IP RAN 二层到边缘场景 MPLS-TP OAM 检测描述

如图 5-7 所示，对于 IP RAN 中二层到边缘的场景，TDM/ATM/Ethernet 等各种制式的 BTS/NodeB、BSC/RNC 可直接连接到 MPLS-TP 组成的包交换传送网络中。利用成熟的 PWE3 技术承载 TDM/ATM/Ethernet 业务。

图 5-7 应用 MPLS-TP 构建 IP RAN 网络（二层到边缘场景）



MPLS-TP OAM 机制用于 MPLS-TP 层的运维管理。应用 MPLS-TP OAM 可以有效检测、识别和定位 MPLS-TP 用户层面故障，在链路或节点出现缺陷或故障时迅速进行保护倒换，并且可以降低网络维护成本。

5.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
AIS	Alarm Indication Signal	告警抑制
CC	Continuity Check	连续性检测
CV	Connectivity Verification	转发故障持续检测
DM	Delay Measurement	时延和时延抖动统计
LM	Loss Measurement	丢包率统计
MEP	Maintenance association End Point	维护终结点
MIP	Maintenance association Intermediate Point	维护中间结点
MPLS-TP	Multiprotocol Label Switching Transport Profile	多协议标签交换传输规范
OAM	Operation Administration & Maintenance	运维管理
RDI	Remote Defect Indication	远端缺陷通告