BFD 技术白皮书

目 录

1 概述 1
1.1 产生背景 1
1.2 技术优点 1
2 BFD 技术实现 1
2.1 概念介绍 1
2.2 原理简介 1
2.3 echo 报文方式的 BFD 会话 2
2.3.1 BFD echo 报文 2
2.3.2 BFD 会话的建立 3
2.3.3 BFD 会话的检测机制 3
2.4 控制报文方式的 BFD 会话 3
2.4.1 BFD 控制报文 3
2.4.2 BFD 会话建立方式 4
2.4.3 BFD 会话建立过程 4
2.4.4 定时器协商 6
2.4.5 BFD 会话的检测机制 8
2.4.6 BFD 回声功能 8
3 SBFD 技术实现 8
3.1 原理简介 8
3.2 运行机制 9
3.3 应用限制 9
4 SRv6 BFD 技术实现 9
4.1 原理简介 9
4.2 检测 SRv6 BE 9
4.2.1 功能简介 9
4.2.2 运行机制 10
4.3 检测 SRv6 TE Policy 10
4.3.1 功能简介 10
4.3.2 运行机制 10
5 Comware 实现的技术特色—硬件 BFD 技术 12
5.1 产生背景 12
5.2 运行机制 12

5.2.1 echo 报文方式的硬件 BFD 12
5.2.2 控制报文方式的硬件 BFD 12
5.3 应用限制 13
6 典型组网应用 13
6.1 路由协议与 BFD 联动典型组网应用 13
6.2 快速重路由与 BFD 联动典型组网应用 13
6.3 VRRP 与 BFD 联动典型组网应用 14
6.4 MPLS L3VPN over SRv6 快速重路由典型组网应用 15
6.5 SR-MPLS TE Policy 与 SBFD 联动典型组网应用 16
7 参考文献 17

1 概述

1.1 产生背景

网络设计时，通常使用冗余备份链路来保护关键应用。网络发生故障时，需要设备能够快速检测出故障，并将流量切换至备份链路以加快网络收敛速度。目前有些链路（如 POS）通过硬件检测机制来实现快速故障检测，但是某些链路（如以太网链路）不具备这样的检测机制。此时，应用就要依靠上层协议自身的机制来进行故障检测，然而上层协议的检测时间都在 1 秒以上，这样的故障检测时间对某些应用来说是不能容忍的。部分路由协议如 OSPF、IS-IS 虽然有 Fast Hello 功能来加快检测速度，但是检测时间也只能达到 1 秒的精度，而且 Fast Hello 功能只是针对本协议的，无法为其它协议提供快速故障检测。

1.2 技术优点

BFD 协议提供了一个通用的、标准化的、介质无关的、协议无关的快速故障检测机制。具有以下优点：

- 对网络设备间任意类型的双向转发路径进行故障检测，包括直连物理链路、SRv6 BE 转发路径、SRv6 TE Policy 转发路径、MPLS LSP、多跳路由路径以及单向链路等。
- 可以为不同的上层应用服务，提供一致的快速故障检测时间。
- 提供毫秒级的检测速度，从而加快网络收敛速度，减少应用中断时间，提高网络的可靠性。

2 BFD 技术实现

2.1 概念介绍

BFD 可以用来进行单跳和多跳检测：

- 单跳检测：是指对两个直连设备进行 IP 连通性检测，这里所说的“单跳”是 IP 的一跳。
- 多跳检测：BFD 可以检测两个设备间任意路径的链路情况，这些路径可能跨越很多跳。

2.2 原理简介

BFD 在两台网络设备上建立会话，用来检测网络设备间的双向转发路径，为上层应用服务。BFD 本身并没有发现机制，而是靠被服务的上层协议通知来建立会话。上层协议在建立新的邻居关系后，将邻居的参数及检测参数（包括目的地址和源地址等）通告给 BFD；BFD 根据收到的参数建立 BFD 会话。会话建立后，建立会话的双方会周期性地向彼此快速发送 BFD 报文。如果在检测时间内没有收到会话对端的 BFD 报文，则认为该双向转发路径发生了故障，并将故障信息通知给该会话所服务的上层应用，由上层应用采取相应的措施。下面以 OSPF 与 BFD 联动为例，简单介绍 BFD 的工作流程。

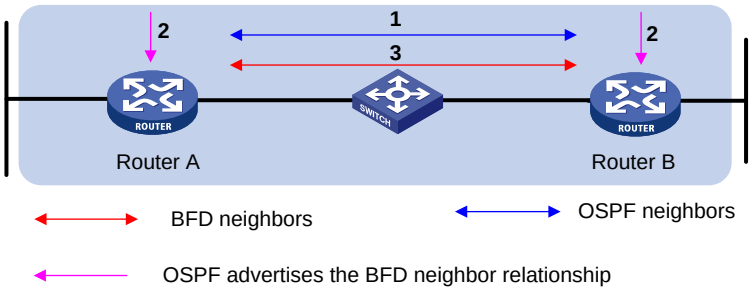
如(3)图 1 所示，上层应用与 BFD 联动触发建立会话的流程为：

(1) OSPF 通过自己的 Hello 机制发现邻居并建立连接；

(2) OSPF 在建立了新的邻居关系后，将邻居信息（包括目的地址和源地址等）通告给 BFD；

(3)　BFD 根据收到的邻居信息建立会话。

图1 BFD 会话建立流程图



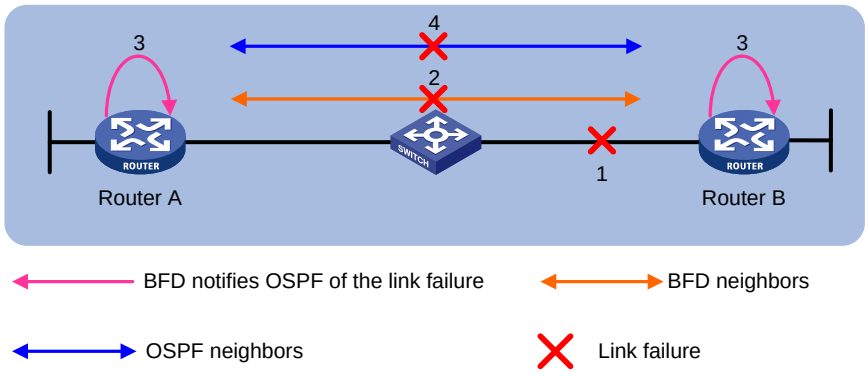
如图 2 所示，BFD 检测到链路故障通知上层应用的流程为：

(1)　BFD 检测到链路故障，BFD 会话状态变为 Down；

(2)　BFD 通知本地 OSPF 进程 BFD 邻居不可达；

(3)　本地 OSPF 进程中断 OSPF 邻居关系。

图2 BFD 故障发现处理流程图



根据检测过程采用的报文类型不同，BFD 会话分为两类：

- echo 报文方式的 BFD 会话。echo 报文方式的 BFD 会话不需要对端设备支持 BFD 功能，或者不需要对端配置 BFD。适用于仅一端设备需要故障检测的情况。
- 控制报文方式的 BFD 会话。控制报文方式需要两端设备均配置 BFD。适用于两端设备均需要故障检测的情况。

下面将分别对这两种方式进行介绍。

2.3 echo报文方式的BFD会话

2.3.1 BFD echo 报文

BFD echo 报文采用 UDP 封装，目的端口号为 3785，目的 IP 地址为发送接口的地址，源 IP 地址由配置产生（配置的源 IP 地址要避免产生 ICMP 重定向）。

BFD 协议并没有对 BFD echo 报文的格式进行定义，唯一的要求是发送方能够通过报文内容识别会话。

当 BFD 会话工作于 echo 报文方式时，仅在隧道应用中支持多跳检测，其他应用中仅支持单跳检测。

2.3.2 BFD 会话的建立

本端周期性发送 echo 报文建立 BFD 会话，对链路进行单向检测。对端不建立 BFD 会话。

2.3.3 BFD 会话的检测机制

本端发送 echo 报文，对端只需把收到的 echo 报文转发回本端。如果本端在检测时间内没有收到对端转发回的 echo 报文，则认为会话 DOWN。

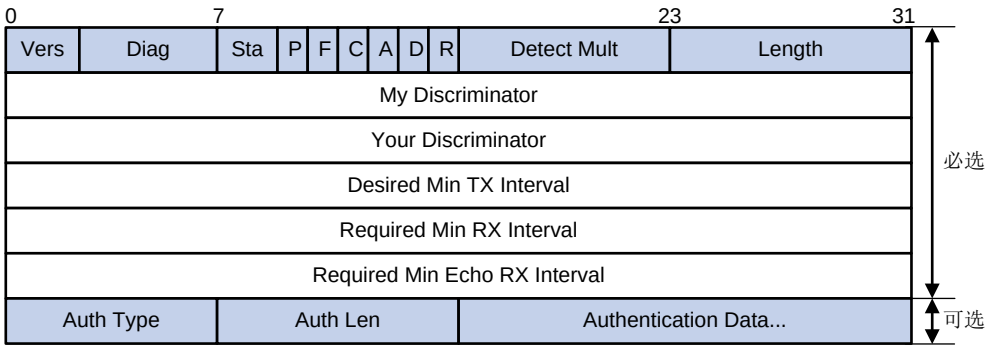
2.4 控制报文方式的BFD会话

2.4.1 BFD 控制报文

BFD 控制报文采用 UDP 封装，源端口号的范围为 49152 到 65535，对于单跳检测其 UDP 目的端口号为 3784，对于多跳检测其 UDP 目的端口号为 4784。

如图 3 所示，BFD 控制报文包括强制必选部分和可选认证部分。

图3 BFD 控制报文



BFD 控制报文各字段含义如表 1 所示。

表1 BFD 控制报文字段含义

字段	含义
Vers	BFD协议版本号，目前版本号为1
Diag	诊断码，表明发送方最近一次会话Down的原因
Sta	发送方BFD会话当前状态，取值为： • 0：代表 AdminDown • 1：代表 Down • 2：代表 Init • 3：代表 Up
P	会话参数变化时置位
F	如果收到的BFD控制报文P字段置位，则将下一个发送的BFD控制报文的F字段置位作为应答
C	该字段置位表明BFD的实现是独立于控制平面的

字段	含义
A	该字段置位表明报文包含认证部分，会话需要进行认证
D	该字段置位表明发送方希望以查询模式运行，不置位表明不希望以查询模式运行或不支持查询模式
R	保留位，发送时设为0，接收时忽略该字段
Detect Mult	检测时间倍数
Length	BFD控制报文长度，单位为字节
My Discriminator	发送方产生的一个唯一非0值，用来标识不同的BFD会话
Your Discriminator	如果已经收到会话邻居发送的BFD控制报文则该值为收到报文中的My Discriminator，否则为0
Desired Min TX Interval	发送方支持的最小BFD控制报文发送时间间隔，单位为微秒
Required Min RX Interval	发送方支持的最小BFD控制报文接收时间间隔，单位为微秒
Required Min Echo RX Interval	发送方支持的最小BFD Echo报文接收时间间隔，单位为微秒。为0表示不支持BFD Echo报文
Auth Type	认证类型
Auth Len	可选认证部分长度，包括Auth Type和Auth Len字段，单位为字节
Authentication Data	认证数据区

2.4.2 BFD 会话建立方式

建立控制报文方式的 BFD 会话有两种方式：

- 通过命令行静态创建 BFD 会话。
- 应用程序与 BFD 联动时动态建立 BFD 会话。

BFD 通过控制报文中的本地标识符（My Discriminator）和远端标识符（Your Discriminator）来区分不同的会话。静态创建 BFD 会话和动态建立 BFD 会话的主要区别在于本地标识符和远端标识符的获取方式不同：

- 静态 BFD 会话的本地标识符和远端标识符由用户手工配置。包括以下两种配置方式：
  - 手工创建静态 BFD 会话的同时，指定本地标识符和远端标识符。
  - 将应用程序与 BFD 联动时，手工指定本地标识符和远端标识符。
- 动态 BFD 会话的本端标识符由本端设备分配，远端标识符在 BFD 会话协商建立过程中获取。

2.4.3 BFD 会话建立过程

1. BFD 会话的状态机迁移机制

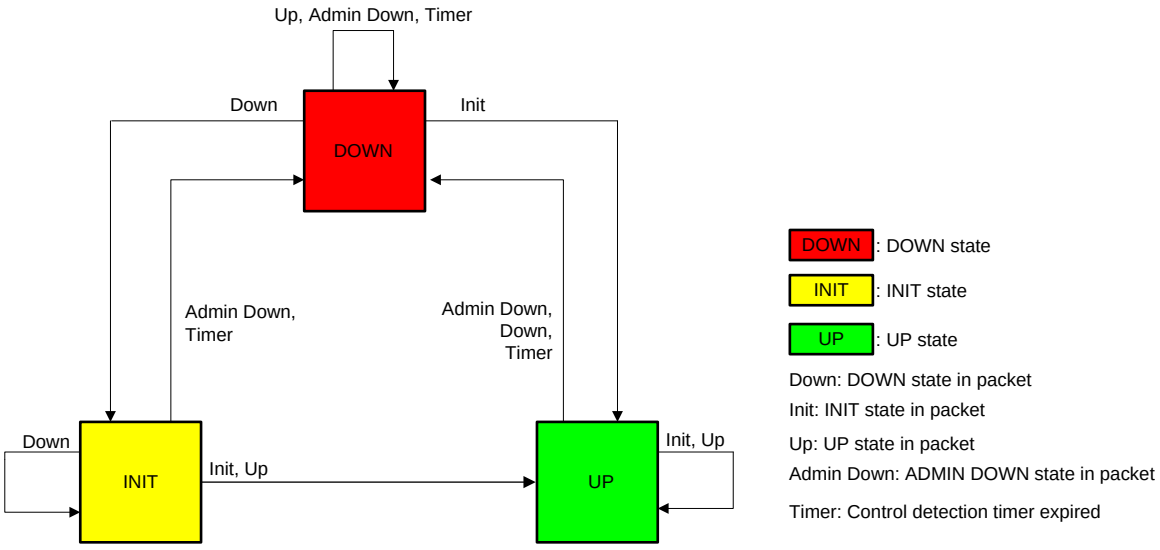
BFD 会话通常有三种状态，分别为：

- DOWN：本端会话已经关闭或刚刚创建。DOWN 状态表示转发路径不可用，与 BFD 会话联动的上层应用需要采取适当的措施，例如主备路径切换等。

- INIT：本端已经可以与对端通信，且本端希望会话进入 UP 状态。
- UP：本端会话已经建立成功。UP 状态表示转发路径可用。

另外，还有一个特殊状态：ADMIN DOWN，本端系统主动阻止建立 BFD 会话时，BFD 会话状态为 ADMIN DOWN。在状态机中 ADMIN DOWN 也是 DOWN 状态。状态机迁移机制如图 4 所示。

图4 BFD 会话状态机迁移机制



2. BFD 会话的建立

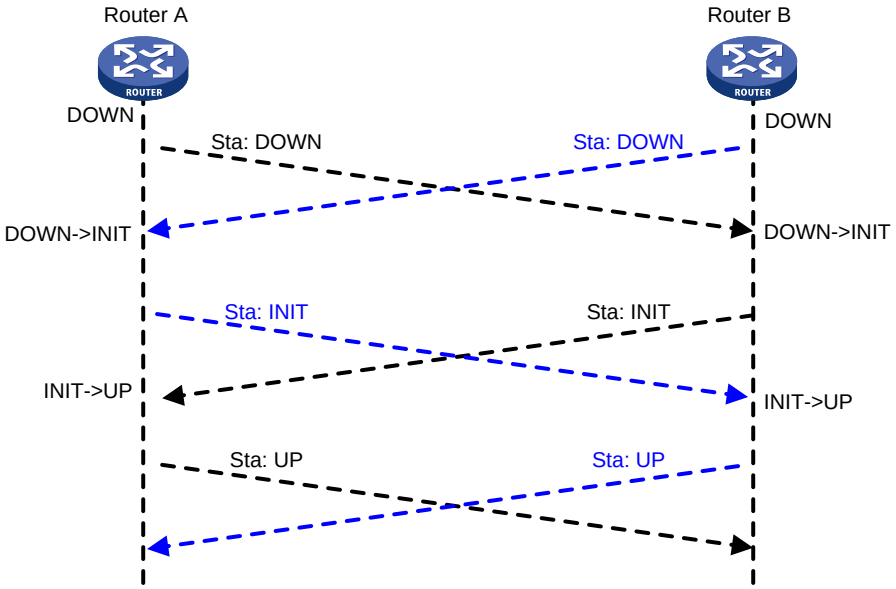
BFD 会话建立前，通过改变 BFD 会话的运行模式可以控制发送 BFD 控制报文的方式：

- 主动模式：在建立会话前不管是否收到对端发来的 BFD 控制报文，都会主动发送 BFD 控制报文；
- 被动模式：在建立会话前不会主动发送 BFD 控制报文，直到收到对端发送来的控制报文。

通信双方至少要有一方运行在主动模式才能成功建立起 BFD 会话。

BFD 使用三次握手的机制来建立会话，发送方在发送 BFD 控制报文时会在 Sta 字段填入本地当前的会话状态，接收方根据收到的 BFD 控制报文的 Sta 字段以及本地当前会话状态来进行状态机的迁移，建立会话。

图5 BFD 会话建立过程



如图 5 所示，以两端均为主动模式的会话建立过程为例，说明 BFD 如何通过报文交互和状态机的变化建立会话，一端主动模式一端被动模式的会话建立过程基本相同：

(1) Router A 和 Router B 的 BFD 收到上层应用的通知后，发送状态为 DOWN 的 BFD 控制报文。

(2) Router B 收到对端状态为 DOWN 的 BFD 控制报文后，本地会话状态由 DOWN 迁移到 INIT，随后将待发送的 BFD 控制报文中的 Sta 字段填为 2（表示会话状态为 INIT）。Router A 的 BFD 状态变化同 Router B。

(3) Router A 收到对端状态为 INIT 的 BFD 控制报文后，本地会话状态由 INIT 迁移到 UP，随后将待发送的 BFD 控制报文中的 Sta 字段填为 3（表示会话状态为 UP）。Router B 的 BFD 状态变化同 Router A。

(4) BFD 双方状态都为 UP 时，标志会话成功建立，两端开始检测链路状态。

(5) 如果本端 BFD 会话 DOWN，将会向对端发送 Sta 字段填为 1 的 BFD 控制报文，通知对端会话 DOWN，对端的 BFD 会话也迁移到 DOWN 状态。

2.4.4 定时器协商

1. 定时器简介

在建立 BFD 会话的过程中，需要通过报文交互协商如下定时器来控制 BFD 检测过程：

- BFD 控制报文的发送时间间隔：BFD 会话建立前，发送时间间隔至少为 1 秒，具体发送时间间隔与设备的型号有关。部分设备上会话数量越多发送的间隔越大，这样可以减小报文流量；在会话建立后，则以协商的时间间隔发送 BFD 控制报文，以实现快速检测。
- 检测时间：每当收到 BFD 控制报文时，就会重置检测时间定时器，保持会话 UP 状态。如果在检测时间内没有收到 BFD 控制报文，BFD 会话会迁移到 DOWN 状态，并通知该会话所服务的上层应用监测对象发生故障，由上层应用采取相应的措施。

2. 定时器协商机制

BFD 会话协商时，定时器值的选取原则为：

- BFD 控制报文发送时间间隔=MAX（本端 Desired Min TX Interval，对端 Required Min RX Interval），也就是说比较慢的一方决定了发送频率。
- 检测时间=对端 BFD 控制报文中的 Detect Mult×MAX（本端 Required Min RX Interval，对端 Desired Min TX Interval）。

不同方向的 BFD 会话定时器是各自独立协商的，双向定时器时间可以不同。

在 BFD 会话有效期间，控制报文发送时间间隔和检测时间可以随时协商修改而不影响会话状态。修改定时器参数会带来如下影响：

- 如果加大本端 Desired Min TX Interval，那么本端实际发送 BFD 控制报文的时间间隔必须要等收到对端 F 字段置位的报文后才能改变，这是为了确保在本端加大 BFD 控制报文发送时间间隔前对端已经加大了检测时间，否则可能导致对端检测定时器错误超时。
- 如果减小本端 Required Min RX Interval，那么本端检测时间必须要等收到对端 F 字段置位的报文后才能改变，这是为了确保在本端减小检测时间前对端已经减小了 BFD 控制报文发送间隔时间，否则可能导致本端检测定时器错误超时。
- 如果减小 Desired Min TX Interval，本端 BFD 控制报文发送时间间隔将会立即减小；加大 Required Min RX Interval，本端检测时间将会立即加大。

下面详细介绍参数改变后定时器的协商过程。如图 6 所示，Router A 与 Router B 建立 BFD 会话，双方的 Desired Min TX Interval 和 Required Min RX Interval（下面简称为 TX 和 RX）都为 100ms，Detect Mult 都为 3。根据定时器协商规则，Router A 的发送时间间隔为 Router A 的 TX 与 Router B 的 RX 中的最大值也就是 100ms，Router B 的发送时间间隔也是 100ms，双方的检测超时时间都为 300ms。

如果此时将 Router A 的 TX 和 RX 加大到 150 ms。

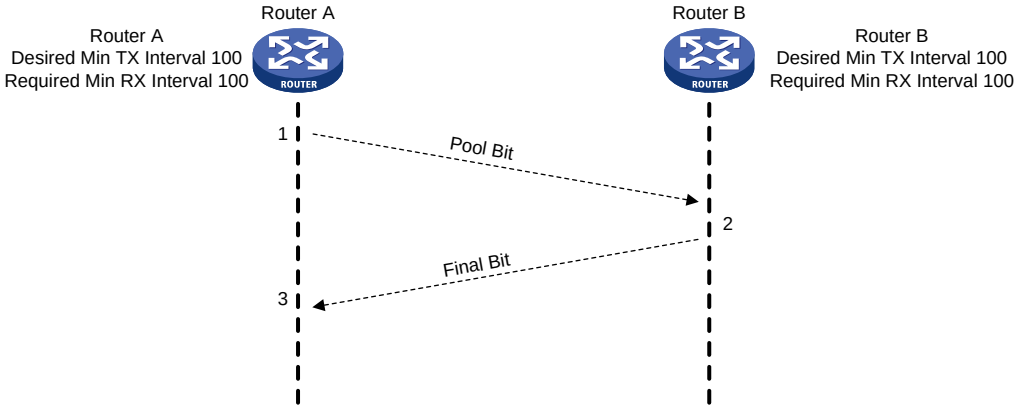
(1) Router A 比较本端的 RX（150ms）和 Router B 的 TX（100ms），然后将本端检测时间改为 450ms，同时向对端发送 P 字段置位的 BFD 控制报文（TX 和 RX 均为 150ms）。

(2) Router B 收到报文后，将收到报文中的 RX 与本端的 TX 进行比较，由于 RX 较大，故 Router B 的发送间隔改为 150ms。经过比较本端 RX 和对端的 TX，从而将检测时间也增大到 450ms。调整完成后给 Router A 回复 F 字段置位的 BFD 控制报文（TX 和 RX 均为 100ms）。

(3) Router A 收到对端发来 F 字段置位的控制报文，根据报文中的 RX 与本端的 TX 进行比较计算出新的时间间隔为 150ms。

(4) 定时器协商完成，双方的发送间隔和检测时间均分别为 150ms 和 450ms。

图6 BFD 检测时间协商



2.4.5 BFD 会话的检测机制

BFD 会话建立后，两端通过周期性发送控制报文对链路进行检测。

BFD 支持两种检测模式：

- 异步模式：设备周期性发送 BFD 控制报文，如果在检测时间内没有收到对端发送的 BFD 控制报文，则认为会话 DOWN。缺省情况下，BFD 会话为异步模式。
- 查询模式：当系统中的 BFD 会话数量较多时，采用查询模式可防止周期性发送 BFD 控制报文的开销对系统的正常运行造成影响。
  - 本端的 BFD 会话工作在查询模式时，本端设备会向对端发送 D 比特位置 1 的 BFD 控制报文，对端（缺省为异步模式）收到该报文后将停止周期性发送 BFD 控制报文。
  - 如果 BFD 会话两端都是查询模式，则双方在 BFD 会话建立后停止周期性发送 BFD 控制报文。仅当需要验证连通性的时候，设备会连续发送 P 比特位置 1 的 BFD 控制报文。如果在检测时间内没有收到对端回应的 F 比特位置 1 的报文，就认为会话 DOWN；如果在检测时间内收到对端回应的 F 比特位置 1 的报文，就认为链路连通，停止发送报文，等待下一次触发查询。

2.4.6 BFD 回声功能

使用异步模式的 BFD 会话检测直连网段的连通性时，可以使用 BFD 回声功能辅助进行故障检测。回声功能启动后，会话的一端周期性地发送 BFD echo 报文，同时自动降低控制报文的接收速率，减少对带宽资源的消耗。对端不对 BFD echo 报文进行处理，而只将此报文转发回发送端。如果发送端连续几个 echo 报文都没有接收到，会话状态将变为 DOWN。

3 SBFD 技术实现

3.1 原理简介

SBFD（Seamless BFD，无缝 BFD）是一种单向的故障检测机制，简化了 BFD 的状态机（SBFD 仅支持 UP、DOWN 两个状态），缩短了会话协商时间，其检测速度比 BFD 更快速。SBFD 适用于仅一端需要进行链路状态检测的情况。

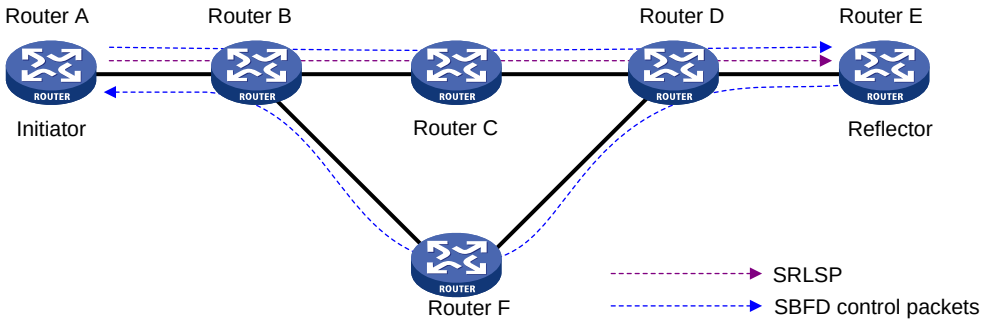
3.2 运行机制

SBFD 会话中，设备的角色分为发起端（Initiator）和响应端（Reflector）：

- Initiator：SBFD 会话的发起者，负责维护 SBFD 会话的状态。Initiator 周期性发送 SBFD 控制报文。
- Reflector：监听到达本地节点的 SBFD 控制报文，并判断是否需要生成 SBFD 响应报文。Reflector 无需维护 SBFD 会话状态。

以图 7 所示的组网为例，说明 SBFD 会话的运行机制。作为 Initiator 的 Router A 通过基于 SR（Segment Routing，段路由）建立的 MPLS TE 隧道，向作为 Reflector 的 Router E 发送 SBFD 控制报文。Router A 只要能够收到 Router E 发送的 SBFD 控制报文，即认为从 Router A 到 Router E 的 SRLSP 路径可达。

图7 SBFD 的 Initiator 和 Reflector



3.3 应用限制

Initiator 上指定的 SBFD 会话的远端标识符必须为 Reflector 上手工指定的 SBFD 会话本地标识符。否则，当 Reflector 收到 Initiator 发送的 SBFD 控制报文后，发现报文中携带的远端标识符不是自己的本地标识符时，不会发送响应报文给 Initiator。

4 SRv6 BFD 技术实现

4.1 原理简介

BFD 可以对 SRv6 BE 的转发路径和 SRv6 TE Policy 的转发路径进行快速故障检测，监测其连通状态。当故障发生时触发 SRv6 BE 或 SRv6 TE Policy 进行流量切换，提高整网可靠性。

4.2 检测SRv6 BE

4.2.1 功能简介

在 SRv6 BE 网络中，使用静态 BFD 会话检测 Locator 的可达性，能够提升故障切换性能。

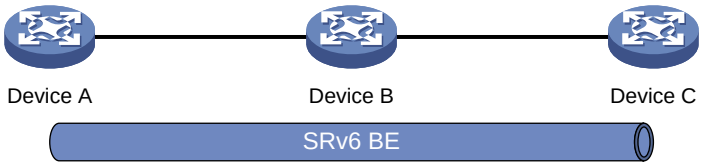
4.2.2 运行机制

如图 8 所示，Device A、Device B 和 Device C 为 SRv6 节点，建立 Device A 到 Device C 的 SRv6 BE 转发路径。Device A 的 Locator 前缀为 100:1::/120，Device B 的 Locator 前缀为 200:1::/120，Device C 的 Locator 前缀为 300:1::/120。分别在 Device A 和 Device C 上创建静态 BFD 会话，Device A 和 Device C 的静态 BFD 会话使用的源地址和目的地址分别如下：

- Device A 上的静态 BFD 会话使用的源地址是 100:1::0，目的地址是 300:1::0。
- Device C 上的静态 BFD 会话使用的源地址是 300:1::0，目的地址是 100:1::0。

完成静态 BFD 会话创建后，Device A 和 Device C 通过 IPv6 路由的最短路径周期性发送 BFD 控制报文。如果 Device A 和 Device C 在检测时间内收到对端发送的 BFD 控制报文，则认为 SRv6 BE 转发路径正常。否则，Device A 和 Device C 认为 SRv6 BE 转发路径故障。

图8 SRv6 BE 组网



4.3 检测SRv6 TE Policy

4.3.1 功能简介

SBFD 和 echo 报文方式的 BFD 均可以用来检测 SRv6 TE Policy 的连通性，为其提供毫秒级的故障检测速度，并实现快速的故障切换。

4.3.2 运行机制

1. SBFD 检测 SRv6 TE Policy

一个 SRv6 TE Policy 中，优先级最高的有效路径为主路径，优先级次高的有效路径为备份路径。SBFD 对 SRv6 TE Policy 的主、备路径进行检测。如果主、备路径中存在多个 SID 列表，SRv6 TE Policy 会建立多个 SBFD 会话分别用来检测每一个 SID 列表对应的转发路径。当 SBFD 检测到 SRv6 TE Policy 主路径的所有 SID 列表均无效时，SBFD 通知 SRv6 TE Policy 切换到备份路径。

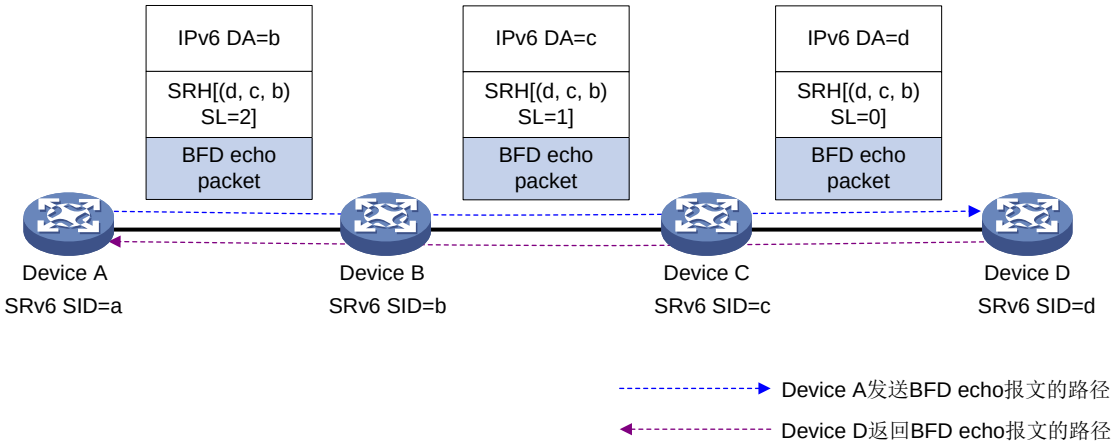
如(3)图 9 所示，在 Device A 上配置 SRv6 TE Policy，并使用 SBFD 检测该 SRv6 TE Policy。SBFD 检测 SRv6 TE Policy 的过程如下：

(1) 头节点作为 Initiator 发送 SBFD 报文，SBFD 报文封装 SRv6 TE Policy 对应的 SID 列表，分别对主、备路径进行检测。存在多个 SID 列表时，使用多个报文封装不同的 SID 列表。

(2) 作为 Reflector 的尾节点收到 SBFD 报文后，检查报文中携带的远端标识符是否与本地配置的标识符一致。如果一致，Reflector 将通过 IPv6 路由向 Initiator 发送 SBFD 响应报文。如果不一致，Reflector 将丢弃收到的 SBFD 报文。

(3) 如果头节点在检测时间超时前能够收到 SBFD 响应报文，则认为 SRv6 TE Policy 的 SID 列表正常。否则，头节点认为 SID 列表故障。如果主路径下的所有 SID 列表都发生故障，则 SBFD 触发主备路径切换。

图9 SBFD for SRv6 TE Policy 检测过程



2. echo 报文方式的 BFD 会话检测 SRv6 TE Policy

一个 SRv6 TE Policy 中，优先级最高的有效路径为主路径，优先级次高的有效路径为备份路径。echo 报文方式的 BFD 对 SRv6 TE Policy 的主、备路径进行检测。如果主、备路径中存在多个 SID 列表，SRv6 TE Policy 会建立多个 BFD 会话分别用来检测每一个 SID 列表对应的转发路径。当 BFD 检测到 SRv6 TE Policy 主路径的所有 SID 列表均无效时，BFD 通知 SRv6 TE Policy 切换到备份路径。

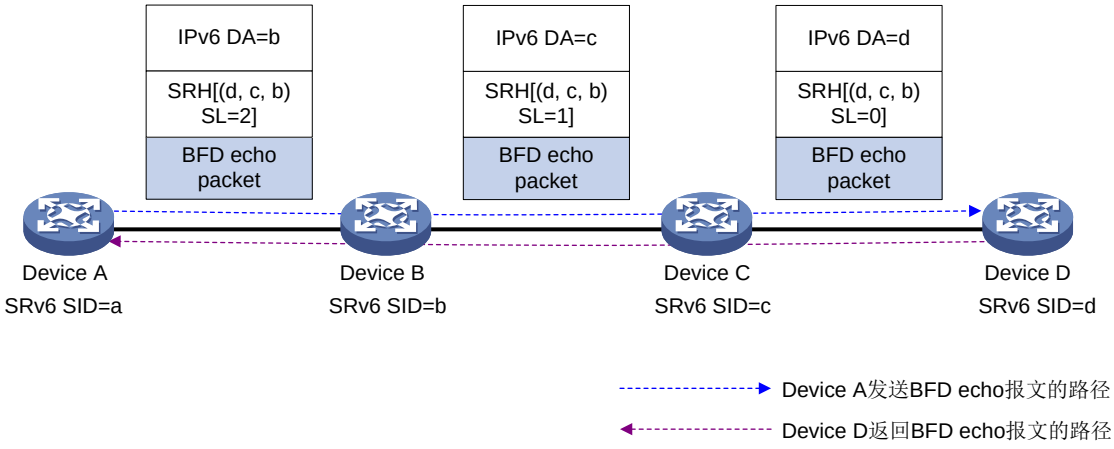
如(3)图 10 所示，在 Router A 上配置 SRv6 TE Policy，并使用 echo 报文方式的 BFD 检测该 SRv6 TE Policy，检测过程如下：

(1) 头节点发送 BFD echo 报文，报文封装 SRv6 TE Policy 对应的 SID 列表，分别对主、备路径进行检测。存在多个 SID 列表时，使用多个报文封装不同的 SID 列表。

(2) 尾节点收到 BFD echo 报文后，通过 IPv6 路由的最短路径将 BFD echo 报文转发回头节点。

(3) 如果头节点在检测时间超时前能够收到尾节点转发回的 BFD echo 报文，则认为 SRv6 TE Policy 的 SID 列表正常。否则，头节点认为 SID 列表故障。如果主路径下的所有 SID 列表都发生故障，则 BFD 触发主备路径切换。

图10 echo 报文方式的 BFD 会话检测 SRv6 TE Policy



5 Comware 实现的技术特色—硬件 BFD 技术

5.1 产生背景

软件 BFD 是指 BFD 检测过程中的报文收发、BFD 会话状态机的维护完全依赖 CPU 来处理。软件 BFD 会极大的消耗 CPU 能力。同时，受 CPU 性能影响，能够支持的 BFD 会话规格较小，无法用于大规格 BFD 会话需求的应用场景。

硬件 BFD 将发送报文、接收报文以及故障检测等消耗 CPU 性能的功能转移到硬件芯片上处理，从而提升 CPU 的利用率，以便支持大规格的 BFD 会话。

5.2 运行机制

5.2.1 echo 报文方式的硬件 BFD

对于 echo 报文方式的 BFD 会话，第一次收到转发回来的 echo 报文后，BFD 会话就会尝试将其转发到硬件芯片处理。具体处理机制如下：

- 如果检测到硬件芯片可以支持 BFD，系统会通知软件处理成功，软件不再维护 BFD 会话。
- 如果检测到硬件芯片不支持 BFD，系统会通知软件处理失败，仍然由软件维护 BFD 会话。

5.2.2 控制报文方式的硬件 BFD

控制报文方式的 BFD 会话状态需要通过控制报文进行协商，硬件芯片的功能比较简单，不能完成 BFD 会话协商功能。因此在会话状态 UP 之前，仍然需要通过 CPU 维护。会话 UP 之后，会尝试转移到硬件芯片处理。具体处理机制如下：

- 如果系统检测到硬件芯片可以支持 BFD，会通知软件处理成功，软件不再维护 BFD 会话。
- 如果系统检测到硬件芯片不支持 BFD，会通知软件处理失败，仍然由软件维护 BFD 会话。如果需要调整会话的各种参数，则由软件进行协商。

为了支持大规格 BFD 会话的并发协商能力，在协商定时器时，硬件 BFD 会话有一些特殊的处理：

(1) 本端在 DOWN 状态收到 INIT 报文，或者在 INIT 状态收到 UP 报文后，BFD 会话变成 UP 状态，并开始向对端发送 P 字段置位的报文。报文中携带的会话发送时间间隔、接收时间间隔和检测倍数都设置为设备支持的最大值。

(2) 对端收到 P 字段置位的报文时，回应 F 字段置位的报文。

(3) 当本端收到对端回应的 F 字段置位的报文，表明对端已经按照最大值调整好，此时开始尝试将发送时间间隔、接收时间间隔和检测倍数下调到配置的值，并向对端发送新的 P 字段置位的报文。报文中携带的发送时间间隔、接收时间间隔和检测倍数为配置的值。由于发送时间间隔变小，因此本端报文的发送时间间隔调整为 MAX（配置的发送时间间隔，上一次收到的对端发送时间间隔）；测时间调整为 MAX（最大接收时间间隔，对端接收时间间隔）×本端检测倍数。

(4) 对端收到新的 P 字段置位的报文时，回应 F 字段置位的报文。

(5) 当本端收到对端回应的新的 F 字段置位的报文，表明对端已经按照 P 字段置位的报文中携带的发送时间间隔、接收时间间隔和检测倍数调整好了实际发送时间间隔和检测时间。此时本端检测时间调整为 MAX（配置的接收时间间隔，对端接收时间间隔）×本端检测倍数。

(6) 如果本端未收到对端回应的 F 字段置位的报文，会持续向对端发送 P 字段置位的报文，在此期间定时器的值为发起协商前的值。直到收到对端回应的 F 字段置位的报文，协商过程才能结束，BFD 按照协商后的结果调整定时器的值。

5.3 应用限制

目前，硬件 BFD 存在如下限制：

- 硬件芯片对 BFD 会话的发送时间、接收时间和检测倍数有一定的限制，比如发送时间、接收时间可能要求必须是 10ms 的整数倍。
- 暂不支持对 BFD 报文进行认证。
- 硬件 BFD 的协商较普通 BFD 的协商流程复杂，因此耗时更久，如果在协商未完成之前链路发生故障，可能导致检测时间较长。

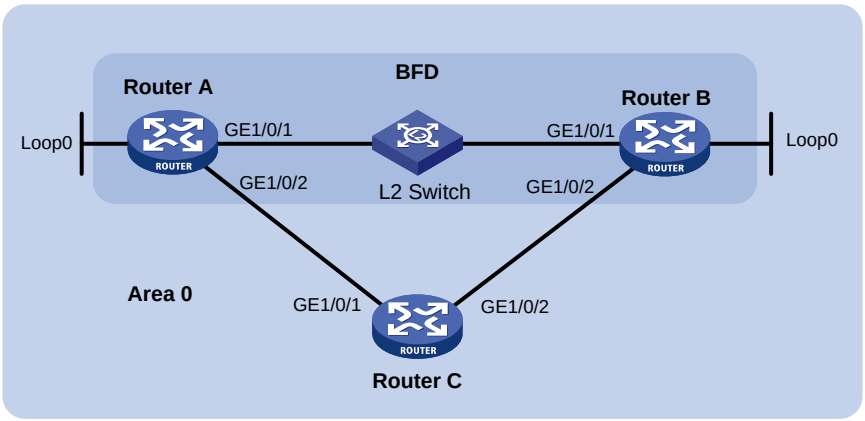
6 典型组网应用

6.1 路由协议与BFD联动典型组网应用

如图 11 所示，两台路由器 Router A、Router B 通过二层交换机互连，在设备上运行路由协议，网络层相互可达。

Router A 与 Router B 之间通过二层交换机通信的链路出现故障时，BFD 能够快速感知并通知 OSPF 协议，OSPF 协议收到 BFD 通知后尽快计算新的路由，从而缩短收敛时间。

图11 路由协议与 BFD 联动组网图



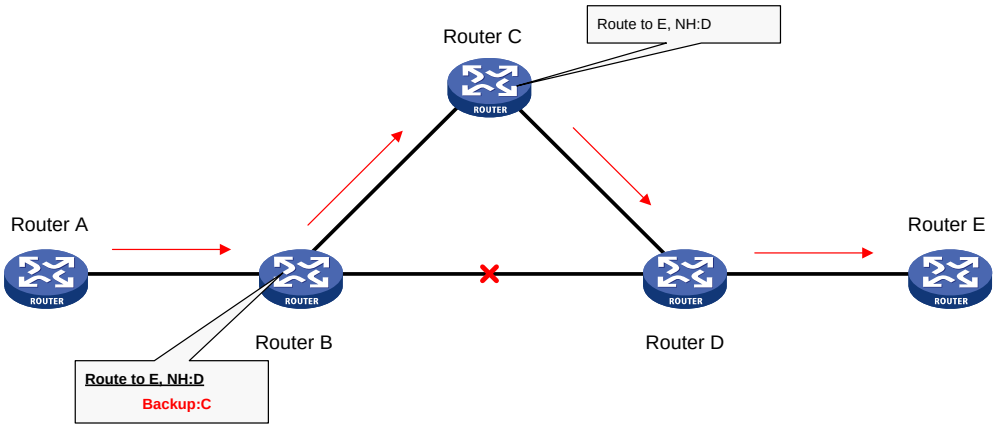
6.2 快速重路由与BFD联动典型组网应用

随着网络的快速发展，IP 网络越来越多的承载语音、视频等多种业务，这些业务对网络的高可靠性提出了更高的要求，从而运营商网络要求更快的收敛速度。

BFD 应用于路由协议以及路由协议快速收敛技术的使用虽然很大程度提高了收敛速度，但还是无法满足语音、视频等新业务对业务中断时间的要求。

而快速重路由和 BFD 联动技术可以很好的满足这种要求，如图 12 所示，通过提前计算备用路径，快速发现主用路径故障，并在主用路径故障时不依赖于控制平面的收敛而直接在转发平面切换至备用路径，极大的缩短了业务中断时间。

图12 快速重路由与 BFD 联动组网图

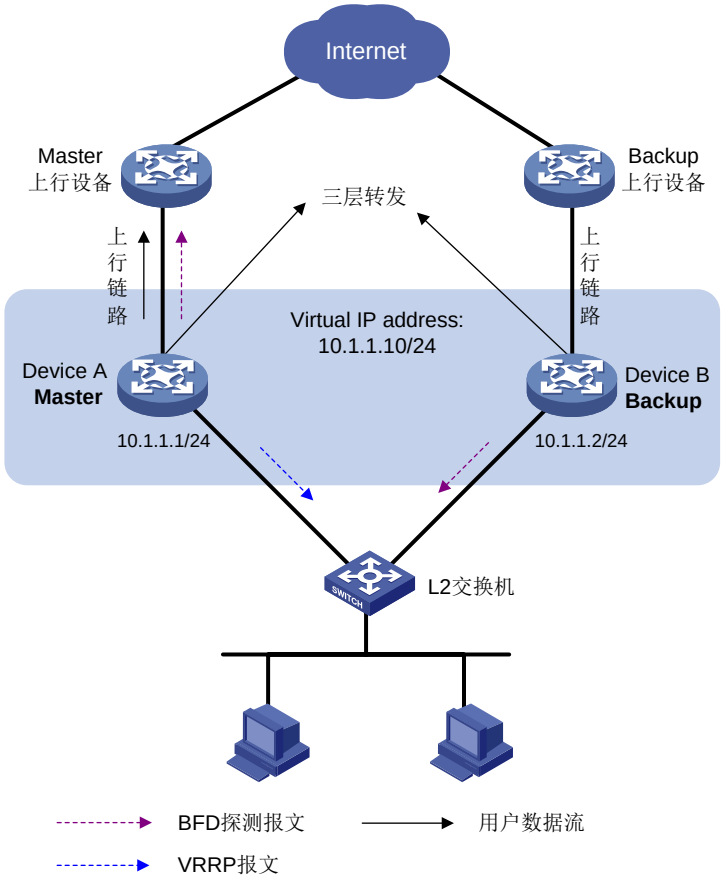


6.3 VRRP与BFD联动典型组网应用

VRRP 的协议关键点是当 Master 出现故障时，Backup 能够快速接替 Master 的转发工作，保证用户数据流的中断时间尽量短。当 Master 出现故障时，VRRP 依靠 Backup 设置的超时时间来判断是否应该抢占，切换速度在 1 秒以上。如图 13 所示，将 BFD 应用于 Backup 对 Master 的检测，可以实现对 Master 故障的快速检测，缩短用户流量中断时间。

VRRP 还会监视 Master 的上行链路能否正常工作，Master 即使正常工作，但是如果其上行链路出现了故障，用户报文实际上也是无法正常转发的。VRRP 是依靠监视接口状态来判断上行链路是否正常工作的，当被监视的接口 DOWN 掉时，Master 主动降低优先级，引起切换。这种监视接口的处理方式依赖于接口的协议状态，如果上行链路出现故障而接口不 DOWN 则无法感知到。将 BFD 应用于 VRRP 上行链路的检测可以有效解决问题。

图13 VRRP 与 BFD 联动组网图



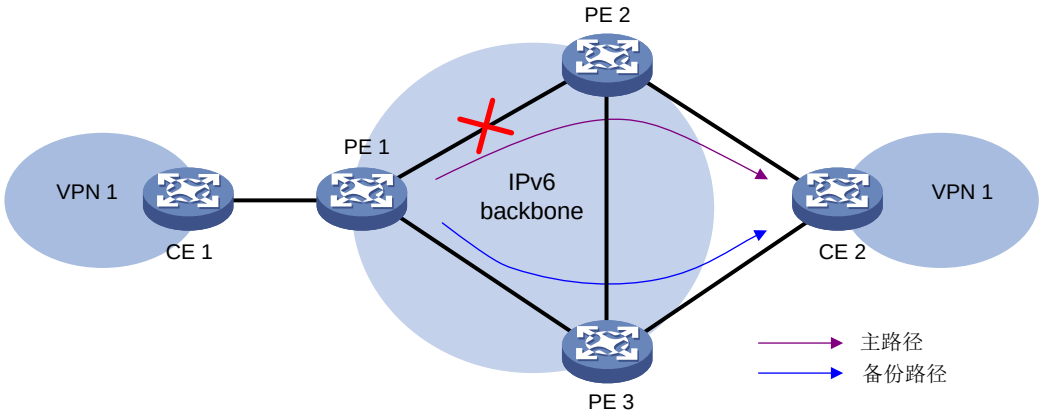
6.4 MPLS L3VPN over SRv6快速重路由典型组网应用

MPLS L3VPN over SRv6 快速重路由功能用来在 CE 双归属（即一个 CE 同时连接两个 PE）的组网环境下，通过为流量转发的主路径指定一条备份路径，并通过静态 BFD 检测主路径的状态，实现当主路径出现故障时，将流量迅速切换到备份路径，大大缩短了故障恢复时间。在使用备份路径转发报文的同时，会重新进行路由优选，优选完毕后，使用新的最优路由来转发报文。

以 VPNv4 路由备份 VPNv4 路由为例，如图 14 所示，在入节点 PE 1 上指定 VPN 1 的 FRR 备份下一跳为 PE 3，则 PE 1 接收到 PE 2 和 PE 3 发布的到达 CE 2 的 VPNv4 路由后，PE 1 会记录这两条 VPNv4 路由，并将 PE 2 发布的 VPNv4 路由当作主路径，PE 3 发布的 VPNv4 路由当作备份路径。

在 PE 1 上配置静态 BFD 检测，通过 BFD 检测 PE 1 到 PE 2 之间公网隧道的状态，由 PE 1 负责流量切换。当公网隧道正常工作时，CE 1 和 CE 2 通过主路径 CE 1—PE 1—PE 2—CE 2 通信。当 PE 1 检测到该公网隧道出现故障时，PE 1 将通过备份路径 CE 1—PE 1—PE 3—CE 2 转发 CE 1 访问 CE 2 的流量。

图14 MPLS L3VPN over SRv6 快速重路由组网图



6.5 SR-MPLS TE Policy与SBFD联动典型组网应用

SR-MPLS TE Policy 不会通过设备之间互相发送的消息来维持自身状态，需要借助 SBFD 检测 SR-MPLS TE Policy 路径故障。

如(3)图 15 所示，源节点 Router A 开启 SR-MPLS TE Policy 与 SBFD 联动功能后，将 End-point 地址作为 SBFD 的远端标识符。当 SR-MPLS TE Policy 中优先级最高的候选路径里存在多个 SID 列表时，会建立多个 SBFD 会话分别用来检测每一个 SID 列表对应的转发路径，所有 SBFD 会话的远端标识符均相同。

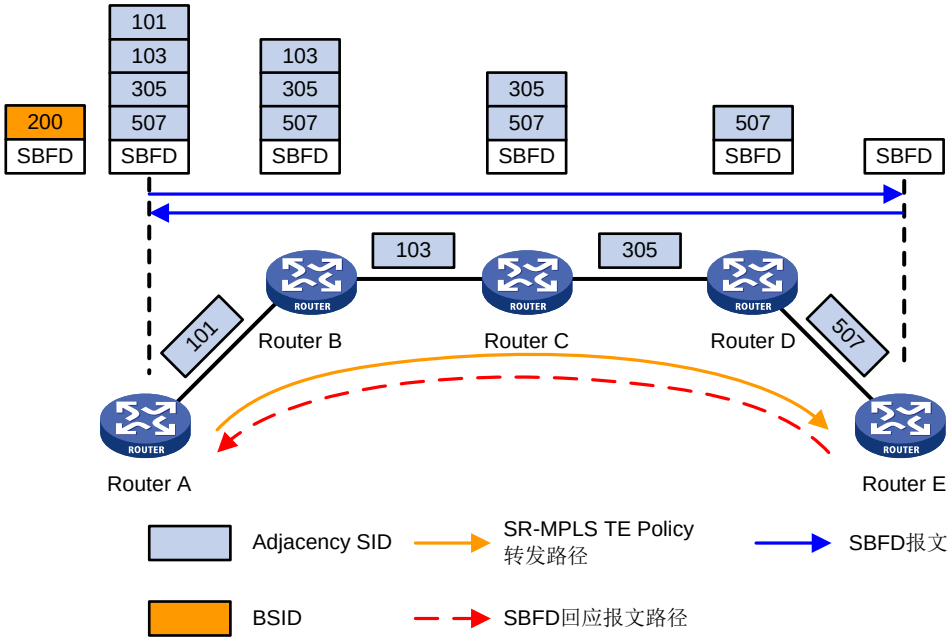
通过 SBFD 检测 SR-MPLS TE Policy 路径过程如下：

(1) 源节点 Router A 对外发送 SBFD 报文，SBFD 报文封装 SR-MPLS TE Policy 对应的 SID 列表。

(2) 尾节点 Router E 收到 SBFD 报文后，通过查找 IP 路由表按照最短路径发送回应报文。

(3) 源节点 Router A 如果收到 SBFD 回应报文，则认为该 SID 列表对应的转发路径正常；否则，会认为该 SID 列表对应转发路径故障。如果一个候选路径下所有 SID 列表对应的转发路径都发生故障，则 SBFD 触发候选路径切换。

图15 SR-MPLS TE Policy 与 SBFD 联动组网图



7 参考文献

- RFC 5880：Bidirectional Forwarding Detection (BFD)
- RFC 5881：Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)
- RFC 5882：Generic Application of Bidirectional Forwarding Detection (BFD)
- RFC 5883：Bidirectional Forwarding Detection (BFD) for Multihop Paths
- RFC 7880：Seamless Bidirectional Forwarding Detection (S-BFD)
- RFC 7881：Seamless Bidirectional Forwarding Detection (S-BFD) for IPv4, IPv6, and MPLS
- RFC 7882：Seamless Bidirectional Forwarding Detection (S-BFD) Use Cases