BFD 技术实现分层 VPLS 系统①

Hierarchical VPLS System Realized by BFD Technique

黄小芳　刘敬彪　(杭州电子科技大学 电子信息学院 浙江 杭州 310018)

鄢　能　(杭州华三通信技术有限公司 浙江 杭州 310053)

摘　要： 针对 IP 网络在设计上无法在少于 1s 的时间内从故障中恢复，BFD 技术提供了一种简单的检测链路或系统转发传输流能力的方法，保证了小于 50ms 的故障检测，大大提高故障检测与恢复速度。为了简化网络管理和提高网络的扩展性，提出了分层 VPLS 的组网方式。利用 BFD 技术对分层 VPLS 系统进行链路故障的检测，指导主备 PW 的切换，大大减少了链路检测时间，减少报文丢失。

关键词： BFD 技术 分层 VPLS PW 链路备份 冗余保护

1　引言

VPLS（Virtual Private Lan Service，虚拟专用局域网业务）是公用网络中提供的一种点到多点的 L2VPN(Layer2 Virtual Private Network，二层虚拟专用网络)业务，它有效结合了以太网、VPN(Virtual Private Network，虚拟专用网络)和 MPLS(Multi-Protocol Label Switching，多协议标签交换)等多种技术的优势，使地域上隔离的用户站点能通过 MAN 或 WAN 相连，并且使各个站点间的连接效果象在一个 LAN 中，从而形成虚拟专用网络[1]。VPLS 的各站点对应的 PE(Provider Edge，服务商边缘路由器)设备之间逻辑全连接，它可以像 L3VPN(Layer3 Virtual Private Network，三层虚拟专用网络)一样提供多点到多点的连接服务，PE 设备能在多点之间进行 MAC 地址学习以及数据报文转发。

VPLS 要求 PE 之间全连接，因此一个 VPLS 实例的 PW（pseudo-wire，虚连接）的条数跟 PE 设备的个数之间的关系是：PW 的条数＝PE 的个数×(PE 的个数－1)/2。在 VPLS 网络规模比较大的情况下，PW 的数目非常庞大，PW 信令开销很大，网络的管理和扩展都将变得复杂。为了简化网络管理和提高网络的扩展性，引出了分层 VPLS 的组网方式。

2　分层VPLS模型

分层 VPLS 将 PE 划分为 UPE(User-side PE，面向用户的 PE 设备)和 NPE(Network-side PE，面向网络的 PE 设备)[2]。UPE 主要作为用户接入 VPN 的 MTU(Maximum Transmission Unit，最大传输单元)，用于连接 CE(Customer Edge，用户边缘设备)和服务商网络；NPE 处于 VPLS 网络的核心域边缘，提供用户报文在核心网上的透明传输服务。UPE 不需要与所有的 NPE 建立全连接，只需在 NPE 之间建立全连接。分层 VPLS 通过分级，减少了 PW 的数目和 PW 信令的负担。

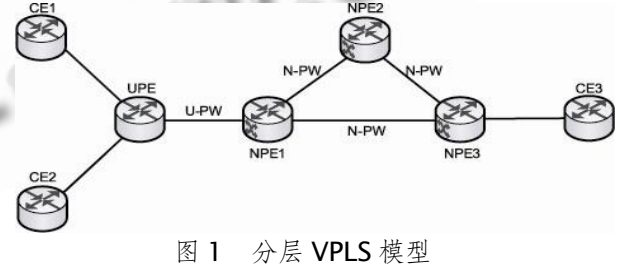

图 1　分层 VPLS 模型

如图 1 所示，分层的 VPLS 网络结构中，在 NPE 中逻辑全连接，UPE 只与最近的 NPE 建立虚连接，通过 NPE 与对端 VPN 站点进行报文交换，这样可以层次化网络拓扑和扩展接入范围。

分层 VPLS 模型对 NPE 性能要求高，因为这里 VPN 业务流量集中，而对 UPE 性能要求比较低，主要

① 收稿时间:2008-08-20

用于 VPN 业务接入。同时，UPE 与 NPE 之间可以链路备份，增强网络健壮性。

3　分层VPLS的链路备份实现方式

UPE 与 NPE 之间只有单条链路连接的方案具有明显的弱点，一旦该接入链路出现故障，汇聚设备连接的所有 VPN 都将丧失连通性。所以，分层 VPLS 的组网上，需要有冗余备份链路存在。在正常情况下，设备只使用一条链路（主链路）接入，一旦 VPLS 系统检测到接入链路失败，它将启用备用链路继续提供 VPN 业务。

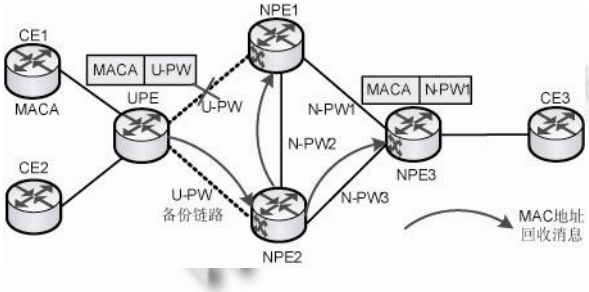

图 2　分层 VPLS 接入链路的备份

如图 2 所示，UPE(User-PW，面向用户的虚连接)检测到与 NPE1 之间的 U-PW 失败，它发起 PW 切换，激活另一条 U-PW 传输数据。

在 PW 失效后的一段时间内，其它站点的 NPE 设备仍然向此 PW 连接的 NPE1 转发流量。当流量到 NPE1 时，报文无法继续转发。假设 CE1 内有一个 MAC 地址为“A”的报文原来走主 PW 到达 CE3，由于 VPLS 的 MAC 学习机制，在 NPE1，NPE3 上都将 MACA 学习到了对应的虚接口。由于 NPE3 不知道对端发生了链路倒换，仍然保留了该 MAC 地址表项，很显然这是错误的。所以 UPE 在主备 PW 切换是，需要将相关的 MAC 地址回收。MAC 地址回收可以使用 LDP(Label Distribution Protocol，标签分发协议)的地址回收消息来实现。地址回收消息中携带 MAC TLV，收到这个消息的设备根据 TLV 中指定的参数进行 MAC 地址的删除或重新学习这些 MAC 地址。当 MAC 地址数目巨大的时候，为了加快收敛速度，可以发送一个空的 MAC 地址列表。收到地址回收消息后，NPE 将移除指定 VSI 中的所有 MAC 地址(从发送此消息的 PE 处学习到的 MAC 地址除外)。

MAC 地址回收消息的发送和处理如下：UPE 发送 MAC 地址回收消息给 NPE2，NPE2 处理该 MAC 地址回收消息，将 MACA 学习到备份 PW 上，再发送地址回收消息给其他对端口（NPE1，NPE3），其他对端进行回收消息处理，将 MACA 学习到对应的 PW 上。

4　BFD检测和冗余保护

为了实现冗余保护，分层 VPLS 组网中，UPE 与两个 NPE 之间建立主备 PW 连接。当主用 PW 发生故障时，快速切换到备用 PW，以保证通信的连续性。目前的网络中一般采用比较慢的协议报文 Hello 机制，尤其在路由协议中，没有硬件的帮助下，检测时间会很长。当数据达到吉比特速率级时，故障检测时间长代表着大量数据的丢失。并且对于不允许路由协议的节点没有办法检测链路状态。现有的 IP 网络中并不具备秒级以下的间歇性故障修复功能，而传统的路由架构对现实应用（如语音）进行准确故障检测方面能力有限。

BFD（Bidirectional Forwarding Detection，双向转发检测）是一套全网统一的检测机制，用于快速检测、监控网络中链路连通状况[3]。BFD 的目标是对相邻转发引擎之间通道故障提供轻负荷、持续时间短的检测。在 UPE 和 NPE 之间采用 BFD（报文周期最快可达 10ms）检测 UPE 和 NPE 之间的路由可达性。UPE 设备上，通过 BFD 检测到与对端 NPE 之间出现连通性故障时，UPE 设备发起 PW 的切换过程，将所有与 NPE 上的 VSI 连接的 PW，都切换到备用的 PW 上，以保证最小流量丢失。

4.1 BFD 工作机制

BFD 提供了一个通用的、标准化的、介质无关、协议无关的快速故障检测机制，可以为各上层协议如路由协议、MPLS 等统一地快速检测两台路由器间双向转发路径的故障。

BFD 在两台路由器上建立会话，用来监测两台路由器间的双向转发路径，为上层协议服务。BFD 本身并没有发现机制，而是靠被服务的上层协议通知其该与谁建立会话，会话建立后如果在检测时间内没有收到对端的 BFD 控制报文则认为发生故障，通知被服务的上层协议，上层协议进行相应的处理。

4.2 BFD 工作流程

BFD 建立过程：

步骤 1：上层协议发现邻居后并建立连接；

步骤 2：上层协议在建立了新的邻居关系时，将邻居的参数及检测参数都(包括目的地址和源地址等) 通告给 BFD；

步骤 3：BFD 根据收到的参数进行计算并建立邻居。

当网络出现故障时：

步骤 1：BFD 检测到链路/网络故障；

步骤 2：拆除 BFD 邻居会话；

步骤 3：BFD 通知本地上层协议进程 BFD 邻居不可达；

步骤 4：本地上层协议处理邻居不可达，如果有备份链路，则切换到备份链路继续运行[4]；

BFD 草案中没有规定检测的时间精度，目前支持 BFD 的设备大多数提供的是毫秒级检测。

5　BFD检测分层VPLS

5.1 利用 LDP 协议报文检测 PW 链路状况

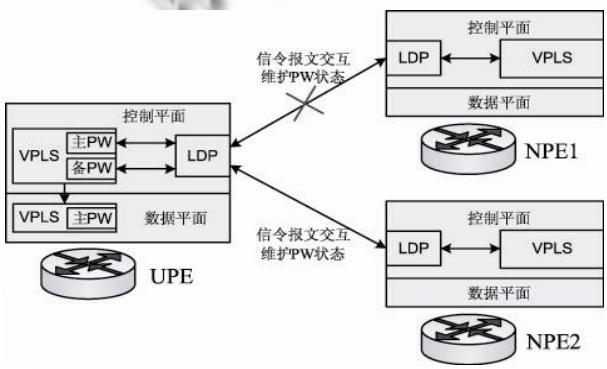

图 3　Hello 报文检测 PW 链路状况

如图 3 所示，由于 UPE 与 NPE 之间运行 LDP 会话，可以根据 LDP 会话的活动状态来判断主 PW 是否失效。UPE 与 NPE1、NPE2 之间的 LDP 会话 up 后，UPE 中的控制平面 VPLS 模块建立主备两条 PW 表项，将主 PW 表项下发数据平面，数据平面将表项下发硬件指导转发。LDP 会话由 LDP 的协议报文进行维护，当检测到连续 3 个 Hello 报文不可到达时，认为链路发生故障，LDP 会话 down。UPE 中的控制平面 VPLS 模块处理 LDP 会话 down 事件，发起 PW 切换，先通知数据平面删除主 PW 和硬件表项，后将备 PW 下发数据平面，数据平面再将表项下发硬件指导转发。协议报文周期最快可达 1s，由此可见，利用协议报文检测 PW 链路故障，指导 PW 切换的时间为 3s 以上。

5.2 利用 BFD 技术检测 PW 链路状况

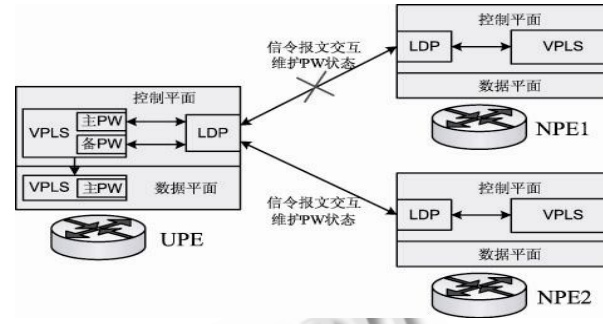

图 4　BFD 技术检测 PW 链路状况

如图 4 所示，BFD 维护 PW 状态的过程如下：

①UPE 与 NPE1、NPE2 之间的 LDP 会话 up 后，LDP 创建两个 BFD 会话，目的地址分别为 NPE1 和 NPE2 的 IP 地址；同时控制平面 VPLS 模块创建主备两条 PW，并将两条 PW 表项下发数据平面。

②数据平面的 VPLS 模块收到表项后，先将两条 PW 表项保存在数据平面中，并将主 PW 表项下发硬件用于进行流量转发；然后根据主备 PW 表项中携带的 NPE1 和 NPE2 的信息创建 BFD 会话。

③BFD 引擎根据会话信息开始收发报文，一般情况下，10ms 一个 BFD 报文，如果连续 3 个报文未收到，则认为链路发生故障，BFD 模块通知数据平面的 VPLS 模块链路异常事件，再通知控制平面的 LDP 模块链路异常。

④数据平面的 VPLS 模块处理链路异常，直接将硬件中的主 PW 表项修改为备 PW 表项，快速的进行转发切换，确保流量少量丢失；同时将 VPLS 模块中的链路切换到备 PW。

⑤控制平面的 LDP 模块处理 BFD 通知的链路异常，通知 VPLS 模块将控制平面 PW 的主备属性进行更换，以保持控制平面和数据平面的一致性。

由上述可知，利用 BFD 机制检测 PW 链路故障，指导 PW 切换的时间一般在 50ms 之内。

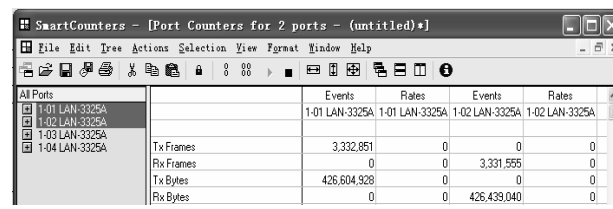

图 5　SmartBits 打流量

(下转第 120 页)

(上接第 145 页)

用 SmartBits 仪器打流量，收发报文结果如图 5 所示，流量为每秒钟 39987 帧报文，从 1 号口发出报文为 3332851 帧报文，从 2 号口收到报文为 3331555 帧，则丢失流量为(3332851 − 3331555) = 1296 帧，从而算出 BFD 检测到链路出异常和主备 PW 切换的总时间为 1296/39987 = 32.4ms。

6 总结

利用 BFD 技术实现分层 VPLS 的主备 PW 切换，大大减少了链路故障的检测时间，并且主备 PW 的切换是在数据平面完成，不需要通过控制平面下发备 PW 表项，减少了主备 PW 的切换时间。BFD 的定位更多的是绑定到数据平面，从而脱离具体的网络协议，使快速检测缺陷实现电信级倒换成为可能；加上 BFD 处理的低开销使得 BFD 具有很广的推广性和更广的适用性，BFD 必将成为 IP 网络电信化的一支重要的推动力量。

参考文献

1 吴昱,胡钧.VPLS—一二层虚拟专用网业务的最佳选择.通讯世界,2004,(10):35 − 37.

2 Stokes O, Kompella V, Heron G, Serbest Y. Testing Hierarchical Virtual Private LAN Services. 2004 − 10.

3 高鑫.双向转发检测（BFD）协议研究.北京:京邮电大学,2007.

4 魏功,张贞贞.基于 MPLS 的分布式快速重路由算法.计算机工程与设计, 2007,(2) .

5 李伦伊.VPLS 转发平台的实现方案.计算机与数字工程, 2005,(6):141 − 144.

6 陈利兵.BFD 技术在 IP 承载网中的应用.现代电信科技, 2008,(1):61 − 64.

7 W.Richard Stevens.TCP/IP 详解.北京:机械工业出版社,1999:103 − 255.