基于 BFD 检测的 IP 快速重路由机制①

胡建萍, 陈瑞森

(杭州电子科技大学 电子信息学院, 杭州 310018)

摘 要: IP 网络目前被广泛应用在互联网中, 更进一步被应用在各种电信业务的承载中. 但传统 IP 网络的设计原则是提供“尽力而为”的服务, 已很难满足现今 IP 电话、视频会议等实时性业务对网络高可靠性的要求. 介绍一种结合 BFD 快速检测机制的 IP 快速重路由技术,能够实现对网络故障的快速修复, 减少丢包, 详细阐述 BFD 检测技术及 IP 快速重路由的实现原理, 并通过实验进行测试验证.

关键词: 快速重路由; IP 网络; BFD

Implementation of IP Fast-Reroute Based on BFD

HU Jian-Ping, CHEN Rui-Sen

(School of Electronic Information, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: IP network is widely applied in internet nowadays, Further be applied in a variety of telecommunications service bearing. but the design principles of the traditional IP network is to provide the service of best effort, and can't meet the high reliability requirements of real-time business such as IP telephone and video conference. This paper will introduce the IP FRR of BFD rapid detection mechanism, which is presented to repair failure rapidly, reduce the packet dropout. And also the detection technology of BFD and the principle of IP FRR will be introduced in detail. At last, the result of experimentation is provided.

Key words: fast reroute; IP network; BFD

传统的 IP 网络通过路由协议的收敛来避开网络的故障链路和节点, 这种修复方式往往要耗费数秒的时间, 这种级别的收敛时间已很难满足现代运营商级 IP 承载网的业务质量的要求, 且随着语音、视频等新网络业务的出现, 现有的 IP 网络在快速收敛方面的不足显得越发明显[1]. 目前对网络故障的快恢复技术, 国内外进行了大量的设计研究, 提出了多种的新技术[2]. 例如通过各种定时器的退避算法以及 Fast Flood 特性缩短路由信息重新扩散时间、通过 ISPF(Incremental Shortest Path First, 增量最短路径优先)以及 RPC(Partial Route Calculation, 路由局部计算)缩短路由计算时间等. 在故障检测方面, 提出了 SDH(Synchronous Digital Hierarchy, 同步数字体系)硬件检测技术、Keepliving 以及 Fast Hello 机制[3]. 通过应用这些技术使路由收敛速度得到大幅度的提升, 但收敛速度还是只能达到亚秒级, 这种级别的故障修复速度还是无法满足现今的网络业务 50 毫秒内修复速度的需求.

为实现这种严格的要求, 本文采用一种通过提前计算备份路由的方法来实现对故障网络的快速修复. 即当故障发生时, 通过 BFD(Bidirectional Forwarding Detection, 双向转发检测)检测机制快速地检测出网络设备间的故障, 利用事先计算好的备份路由替换失效路由先在本地直接修复故障, 在整个新路由完成重新收敛期间, 一直使用事前确定的备份路由指导转发, 这样可将流量中断时间缩短到 50ms 以内, 确保业务的可靠性运行.

1 BFD检测技术

1.1 BFD 简介

故障检测速度是决定网络故障修复速度的一个重

① 收稿时间:2012-07-14;收到修改稿时间:2012-09-14

要环节, 使用传统的“Hello”机制进行故障检测, 其检测时间都在 1s 以上, 会导致大量报文的丢失. 因此, BFD 检测技术应运而生, 其提供了一种通用的、标准化的、介质无关、协议无关的快速故障检测机制, 可以为上层协议实现毫秒级的快速故障的检测[4,5].

1.2 BFD 会话建立流程

BFD 协议本身没有邻居发现机制, 而是靠被服务的上层协议通知其与谁建立会话[6,7]. BFD 会话建立模式有主动和被动两种. 主动模式即先主动向对端的设备发送控制报文即使在未收到对端的报文时. 被动模式即必须先收到对端的控制报文后才会作出相应的报文回应. 所以要建立一条 BFD 会话连接, 前提必须是至少有一端的设备是处于主动模式. 下面以一端处于主动模式另一端处于被动模式来介绍 BFD 会话建立的过程, 如图 1 是会话建立的握手过程.

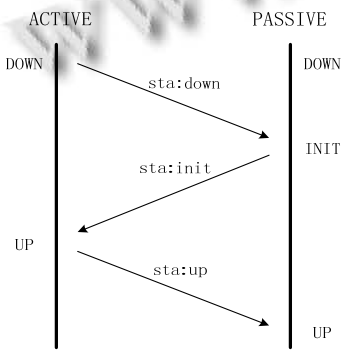

图 1 会话建立过程

① 开始时, 主动端和被动端都处于 down 状态, 主动方这时收到上层的通知后, 向被动端发送状态为 down 的 BFD 控制报文.

② 被动端收到对端发来的 down 的控制报文后, 将自身的状态从 down 转换成 init, 之后会向主动方发送状态为 init 的报文.

③ 主动方收到 init 的报文后, 随即将本端的会话状态由 down 转为 up, 接着又向被动方发送状态为 up 的报文.

④ 被动方接收到 up 的报文后, 将本端状态从 init 转为 up.

⑤ 至此, 两端的状态都为 up, BFD 会话正式建立并开始进行链路状态的检测.

1.3 BFD 定时器协商

BFD 会话建立前, 报文发送间隔不小于 1000ms, 在会话建立后则会以两端协商好后的间隔发送报文来进行链路的检测. 两端的控制报文的发送时间间隔及检测时间在会话建立后可随时进行协商修改, 以达到最优的检测效果. BFD 控制报文发送时间间隔是本端 Desired Min TX Interval 与对端 Required Min RX Interval 两者的最大值, 本端检测时间为对端 BFD 报文中的 Detect Mult 乘以协商好后的对端报文的发送时间间隔. 参数协商过程的报文交互流程如图 2 所示.

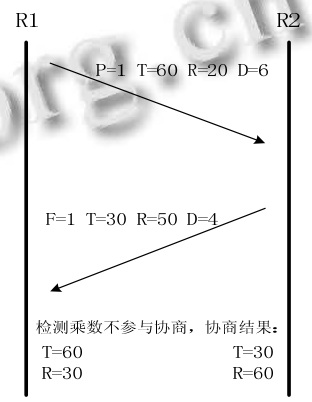

图 2 参数协商过程

① 源端 R1 修改配置参数, 发起协商请求, 修改报文发送间隔为 60ms, 接收间隔为 20ms, 检测乘数为 6. 在发送的协商报文中将 P 置位, 并携带新参数信息.

② 对端 R2 收到 P 置位的报文, 比较自身当前配置, 发送 F 置位的回应报文并携带自身可以接收的检测参数: 发送间隔为 30ms, 接收间隔为 50ms, 检测参数为 4.

③ 源端接收到 F 置位报文, 根据其所携带的信息, 保证满足最低能力来确定自身的参数.

④ 最终确定的协商结果 R1 端发送间隔为 60ms, 接收间隔为 30ms, R2 端发送间隔为 30ms, 接收间隔为 60ms, 检测乘数不参与协商. R1 端检测时间=R2 端发送间隔 30ms*R2 端检测乘数 4=120ms, R2 端检测时间=R1 端发送间隔 60ms*R1 端检测乘数 6=360ms.

1.4 BFD 故障检测

BFD 会话建立及定时器协商好后, 双方将根据协商好的时间间隔互发控制报文, 当收到对端的报文后将重置本端的检测定时器, 并保持报文会话状态. 如果在检测时间内没有收到对端的报文, 将视为有故障发生, 将会话状态由 up 转换为 down 状态, 同时通知上层协议模块进行相应的故障处理.

2 IP FRR

2.1 IP FRR 基本原理

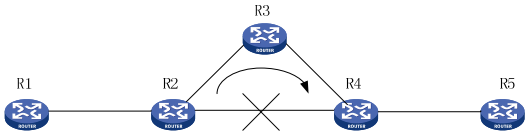

图 3　IP FRR 基本原理图

IP FRR(IP Fast Reroute, IP 快速重路由)基本原理就是在 IP 网络中为你想保护的链路建立一条备份路由, 主链路发生故障, 则将流量快速切换到备用链路上来.

举例说明, 如上图 3 所示, 在正常情况下, 源路由器 R1 到目的路由器 R5 所走的路由路径是 R1-R2-R4-R5. R2 的下一跳是 R4, 同时 R2 的路由表中同时安装一条备份的路径. 当路由器 R2 探测到 R2 和 R4 间的链路发生故障时, 将目的地为 R5 的报文通过备份的路径 R1-R2-R3-R4-R5 进行转发.

2.2 IP FRR 实现方案

目前 IP FRR 有多种不同的技术方案, 各有不同的优缺点, 而其中 Loop Free Alternate 方案对现有的路由协议的更改最少且相对成熟, 是目前应用较广泛的一种 IP FRR 技术方案[8].

Loop Free Alternate 技术其实就是为每条路由的下一跳找出全网无环的备份下一跳, 备份下一跳的计算方法就是这种技术的核心, 且得到的备份下一跳需满足无环的条件.

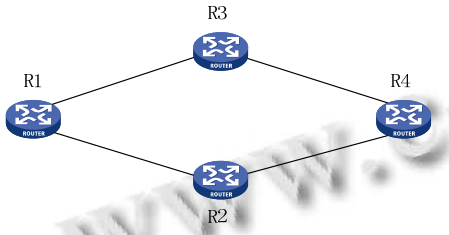

图 4　Loop Free Alternate 组网图

如上图 4 所示[9], R1 表示源, R4 表示目的, 假设 R1 到 R4 的路由主一跳为 R2, 当 R1 到 R2 的链路发生故障时, 可以将 R1 到 R2 的流量切换到 R3 上, 但是如果 R3-R4 的 cost 小于 R3-R1 的 cost 和 R1 到 R4 的 cost 之和的话, 这时由于 R1 到 R2 的链路故障而切换到 R3 上进行转发的流量将重新转发给 R1, 从而形成环路.

Loop Free Alternate 技术就是通过 SPF(shortest path first, 最短路径优先)算法去区分上述的情况, 去找出不会产生环路而真正能保持流量不中断的备份下一跳. 其选取出的下一跳需满足如下的不等式:

Distance_opt(R3,R4)<Distance_opt(R3,R1)+Distance_opt(R1,R4)

说明: Distance_opt(R1,R4)表示通过正常 SPF 计算得到的 R1 到 R4 的最短路径 cost.

一般正常的 SPF 计算的是以本地为源到其他设备的最短路径 cost, 而如何在 R1 上得到 Distance_opt(R3,R1)和 Distance_opt(R3,R4), 则需要特殊的计算, 目前我们的实现方法是在 R1 上以 R3 为 root 进行 SPF 的计算来得到 R3 到其他节点设备的最短路径 cost.

2.3 路由快速切换方法

传统的路由转发表项的组织结构: FIB(Forwarding Information Base, 转发信息库)通过下一跳及出接口信息关联 ARP(Address Resolution Protocol, 地址解析协议)表. 在 IP 报文转发流程中, 先查询 FIB, 并根据 FIB 表项中出接口、下一跳信息获取 ARP, 再根据 ARP 中链路层信息封装并发送. 当 FIB 相关联的下一跳发生改变时, 需要反刷 FIB 表. 如图 5 所示, 当大量的 FIB 表关联同一条 ARP 表时, 这时候下一跳发生变化, 需逐条更新 FIB 表, 这需要大量的时间, 无法实现快速的切换.

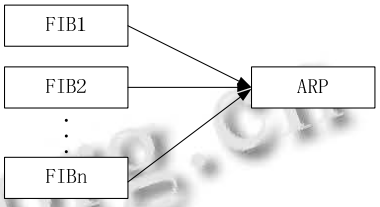

图 5　传统表项组织结构

为实现路由的快速切换, 我们在传统的路由表项结构的基础上增加一条 VN(Virtual Nexthop, 虚拟下一跳)表, 如图 6 所示. 针对 IP FRR 这种情况, 为 FRR 的主备下一跳分配一条 VN 表项, 当主链路发生故障时, FIB 表关联的下一跳由原先的主下一跳切换为备下一跳, 只修改 VN 和 ARP 的关联关系, 不必逐条刷大量的 FIB 表, 即可实现路由表项的快速切换, 节省了刷大量 FIB 表所耗的时间.

2.4 IP FRR 与 BFD 联动

当 BFD 检测到链路故障时, 需要通知 FRR 进行主备链路的切换. 在实现中, 通过在 FIB 模块中将 IP FRR 主下一跳 IP 与 BFD 进行关联, 在主路径上进行链路检测. 主路径的状态发生改变时, BFD 通知 FIB 对

IP FRR 路由的主备路径进行切换, 再通知路由模块重新计算路由, 收敛完成前采用备路径进行流量转发.

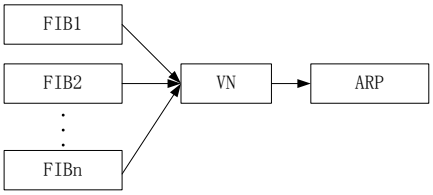

图 6 改进表项组织结构

3 结果验证

测试验证组网如图 7 所示, 将源 R1 和目的 R5 用思博伦的 TestCenter 测试仪模拟进行打流和接收流量. 在 R2、R3 和 R4 上使能 OSPF 协议, 相互学习路由. R1 上构造目的为 R5 的报文进行打流, 同时在 R5 记录流量接收情况, 计算故障切换时间, 切换时间(ms)=丢包数(Byte)*8/发包速率(Mbps)*0.001.

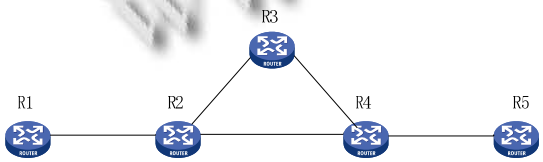

图 7 测试组网图

未部署 FRR 和 BFD 时, 测试仪以 10Mbps 发包速率打入流量, 流量稳定后 shutdown 掉 R4 与 R2 间的链路, 观察链路切换情况, 结果如表 1 所示. 这种情况下设备是通过传统协议重路由来重新学习路由进而来修复故障网络, 其故障链路切换时间在 1s 左右, 丢包数比较大.

部署 FRR 和 BFD 后, 测试仪以相同发包速率打入流量, 同样通过 shutdown 掉 R4 与 R2 间的链路来模拟链路故障, 测试结果如表 2 所示. 这种情况下与未部署 FRR 所不同的是设备会同时学到主备两条路由, 在故障发生时, 通过 BFD 感知故障, 并快速将流量切换到备路由上, 结果显示其切换时间为 30ms 左右, 较未部署时有了大幅提高, 并达到了 50ms 以内的预期结果.

通过以上两项数据对比, 说明本文介绍的快速重路由机制是能够起到快速修复网络故障、提高网络可靠性作用的.

表 1 未使能 FRR 测试结果

测试次数	发送报文(Byte)	接收报文(Byte)	丢失报文(Byte)	切换时间(ms)
1	880485120	879122497	1362623	1090.1
2	992574361	991338489	1235872	988.7
3	986587955	985187209	1400746	1120.6

表 2 使能 FRR 测试结果

测试次数	发送报文(Byte)	接收报文(Byte)	丢失报文(Byte)	切换时间(ms)
1	874192042	874155414	36628	29.3
2	914454450	914414333	40117	32.1
3	804113273	804077359	35914	28.7

4 结语

本文提出基于 BFD 快速检测的 IP FRR 机制能够有效缩短 IP 网络故障的修复速度, 实现 50ms 内快速切换的要求, 解决了链路失效但新路由尚未完全收敛这段时间报文大量丢失的问题, 相较于传统的单纯依靠上层路由协议的重收敛实现故障修复的方法, 其修复速度有了大幅度的提升. 这项研究大大提高了 IP 网络的可靠性, 对视频、语音等对网络有高可靠性要求的实时业务的应用具有非常重要的意义.

参考文献

1 宋哲.IP 快速重路由技术的研究[硕士学位论文].成都:电子科技大学,2010.
2 商荣亮.基于 OSPF 路由协议的 IP 快速路径切换技术的研究与实现[硕士学位论文]. 长沙:国防科技大学, 2011.
3 周跃文,张新菊,曾玉林.BFD 协议分布式实现方案的剖析. 网络与多媒体,2011,35(7):71−74.
4 陈利兵.BFD 技术在 IP 承载网中的应用.现代电信科技, 2008,1(1):61−64.
5 徐俊,秦艺力,唐淼,汤楹.通信 IP 网 BFD 应用的研究.电子设计工程,2012,20(3):39−43.
6 Katz D, Ward D. Bidirectional Forwarding Detection. IETF RFC5880, 2010.
7 Katz D, Ward D. Generic Application of Bidirectional Forwarding Detection. IETF RFC5882, 2010.
8 商荣亮,张晓哲,鄢苏丹.面向 IP 快速路径切换的 OSPF 冗余路径算法.计算机技术与发展,2011,21(6):1−3.
9 Atlas A, Zinin A. Basic Specification for IP Fast-Reroute: Loop-free Alternates. IETF RFC5286, 2008.