IP 网络的快速故障恢复

张民贵, 刘 斌
(清华大学计算机科学与技术系, 北京 100084)

摘 要: 随着互联网的迅速发展, 人类通信对其依赖性日益增强, 而 IP 层故障恢复能力低下, 阻碍了互联网性能的提高. 近些年来, 国际学术界对 IP 网络快速故障恢复的方案研究异常活跃, 提出了加快故障恢复速度的三条途径: (1) 加快 IP 路由收敛; (2) 使用主动式故障恢复; (3) 提高故障检测的速度与准确性. 针对已有解决方案的不足, 本文得出, 要推动 IP 网络的快速故障恢复方案的实现, 必须做好: (1) 故障后的通信负载均衡; (2) 互操作测试及路由器体系结构的重新设计.

关键词: IP 网络; 故障恢复; 路由收敛; 备份路径; 负载均衡
中图分类号: TP393. 05 文献标识码: A 文章编号: 0372-2112 (2008) 08-1595-08

Fast Failure Recovery of IP Networks

ZHANG Min-gui, LIU Bin
(*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*)

Abstract: Internet has become an essential infrastructure of communication. Limited survivability in the face of failure impairs the performance of Internet. Recent years, fast failure recovery of IP network has drawn much attention. As the state-of-art, there three ways to realize fast failure recovery of IP Networks: (1) to accelerate the convergence of IP routing; (2) to employ the proactive recovery strategy; (3) to increase the speed and accuracy of failure detection. This paper analyzes the defects of the existing solutions. In order to realize fast failure recovery of IP networks, we should: (1) balance the traffic load after recovery; (2) do interoperability test and redesign the routers' architecture.

Key words: IP network; failure recovery; routing convergence; backup path; load balance

1 引言

1983 年, ARPANET 的所有主机更改成 TCP/IP 协议栈, 并在网络中加入路由器, 从而形成了互联网的雏形[1]. 最初, 互联网主要用于传输非实时业务, 如收发电子邮件, 浏览网页等. 互联网的路由收敛需要数秒的时间[2], 对于非实时业务来说, 这个时间是可以忍受的. 而在接下来的 20 多年里, 互联网已经发展成为一个全球性的通信网络, 过去许多在电信网和有线电视网中传输的业务也开始转向互联网. 随着规模的不断扩大, 互联网呈现出许多特点: (1) 大量实时业务开始在互联网上传输, 例如 VoIP、在线聊天、视频点播、多用户在线游戏等, 这些业务要求毫秒级的故障恢复时间; (2) 业务复用程度越来越高, 尤其是密集波分复用 (dense wave-length division multiplexing, DWDM) 技术的采用使单根光纤拥有 Tbits/s 数量级的传输能力[3,4]. 这样, 单根链路故障造成的后果非常严重[5]; (3) 大量关键性 (mission-critical) 业务, 如电子商务, 在互联网上传输, 这些业务对网络可用性要求很高. 上述新特征对传统互联网的故障恢复能力提出了挑战. 与此同时, 互联网是一个拓扑结构不断变化的动态网络, 这是因为: (1) 互联网是一个即联即用的网络, 不断的有新的设备加入互联网或损坏的设备离开互联网, 使其拓扑结构不断变化[6]; (2) 自然灾害 (如地震等)、设备断电、自然老化等导致节点或链路出现硬件故障; (3) 人为原因造成配置错误或软件漏洞, 使网络设备运行异常; (4) 对网络进行日常维护需要关闭某些设备; (5) 网络攻击频繁发生, 恶劣的网络攻击能够短时间内造成大量网络设备瘫痪[7,8]. 这些原因使互联网拓扑

收稿日期: 2007-07-18; 修回日期: 2008-03-12
基金项目: 国家自然科学基金 (No. 60373007, No. 60573121); 教育部科技创新工程重大项目培育资金 (No. 705003); 教育部博士点基金 (No. 20040003048, No. 20060003058); 清华大学基础研究基金 (No. JCpy2005054); 国家 973 重点基础研究发展规划 (No. 2007CB310701); 国家 863 高技术研究发展计划 (No. 2007AA01Z216)

结构频繁变化, 迫切需要通过故障恢复来保证其可靠性. 鉴于互联网的规模, 无法在域间实现满足实时业务要求的快速故障恢复[9], 本文讲述的是互联网域内故障的快速恢复.

节点或链路产生故障时, 互联网通过 IP 层的路由查找算法确定替代路径. 链路状态路由协议是互联网中主流的域内路由协议, 在链路状态路由协议中, 路由器通过链路状态通告(link-state advertisements, LSAs) 来描述本地的拓扑结构, 即该节点和其他节点的连接关系[1]. 若网络中产生故障, 会影响到与故障节点或链路相连的路由器, 导致其网络拓扑结构发生变化, 而检测到故障的路由器通过产生新的 LSAs 来描述网络拓扑结构的变化. 通过洪泛, 新的 LSAs 被散布到网络中所有的路由器上, 路由器通过收集和更新 LSAs 来维护本地的链路状态数据库(link-state database) , 并通过查找链路状态数据库来建立路由. 由于故障链路或节点不出现在更新过的链路状态数据库中, 新建立的路由将顺利地绕过故障节点或链路. 新的 LSAs 在网络中进行洪泛, 路由器更新自己的链路状态数据库和路由器重新计算路由的过程被称为路由收敛(convergence) . 路由收敛行为使互联网有一定的健壮性, 传统互联网正是依赖自身的这种健壮性来实现故障恢复的. 然路由收敛耗时长, 通常需要数秒的时间才能完成, 无法适应互联网的发展需求.

按照备份路径是否在故障发生前确定这一特征, 可以将 IP 网络故障恢复方案分为两类: 主动式故障恢复和被动式故障恢复. 在网络故障发生之前, 为工作路径确定好备份路径并预留资源, 这就是主动式故障恢复[10]. 主动式故障恢复能够保证故障的快速恢复, 但是资源利用率却不高. 在故障发生后, 根据当时的网络状态, 利用空闲资源建立备份路径, 这就是被动式故障恢复. 被动故障恢复提高了资源的利用率, 但是故障恢复时间较长. 可以看出: 缩短备份路径的计算时间与提高资源利用率是相互矛盾的. 作为一种折衷方案, IP 层使用的不预留资源的主动式故障恢复成为近几年学术界研究的热点.

加快 IP 网络故障恢复速度主要有三条途径: (1) 加快 IP 路由收敛过程, 提高故障恢复的速度. IP 路由收敛过程可以分为故障检测时间、LSAs 的传播时间和最短路由的重新计算时间等多个阶段. 为缩短上述时间阶段, 需要加快 HELLO 包的传输, 减少收敛过程中为保证网络稳定而人为设置的计时器(timer) 延迟, 但同时需要维护网络的稳定性; (2) 使用主动式故障恢复, 在故障产生前就计算好备份路径, 以加快故障恢复的速度. 使用主动式故障恢复, 能够在路由收敛之前迅速确定备份路径, 尤其适合解决网络中频繁发生的、持续时间又短的故障[11]; (3) 加快故障检测, 同时兼顾故障检测的准确性, 实现故障的快速准确隔离. 故障的检测速度和准确性是一对矛盾, 为满足 IP 网络快速检测故障的目标, 需要抑制 IP 路由收敛过程, 由此带来的故障检测不准确的问题需要通过故障隔离来解决. 使用互联网工程任务组(Internet Engineering Task Force, IETF) 草案中提出的"双向故障检测协议" (Bidirectional Forwarding Detection, BFD)[12]有望解决故障检测速度与准确性的矛盾.

为推进 IP 网络的快速故障恢复的实现, 需要做好两个方面的工作: (1) 通过流量工程对故障恢复后的通信负载进行均衡; (2) 对快速故障恢复方案进行互操作协议测试, 重新设计路由器的体系结构.

2 IP 网络快速故障恢复的解决方案

IP 网络的故障恢复速度缓慢, 破坏了互联网的性能. 如何提高故障恢复速度, 成为近几年学术界研究的热点问题. 这些研究主要从三个方面入手, 下面分别予以介绍.

2.1 加快 IP 路由收敛

这是故障恢复时间过长问题直观的解决方案. 在文献[16~ 18] 中认为 IP 路由的收敛时间主要由故障检测时间、LSAs 的传播时间和最短路由的重新计算时间三部分构成, 而文献[2] 中又将路由的收敛时间细分为 7 部分: (1) 检测到端口状态发生变化所需的时间; (2) 将链路状态变化通知给路由协议栈之前的计时器延迟; (3) 产生新的 LSAs 之前的计时器延迟; (4) LSAs 在网络中洪泛的时间; (5) 从收到 LSAs 到开始运行 SPF 算法之前的计时器延迟; (6) SPF 路由计算和路由表的更新时间; (7) 在路由器线卡的转发表中插入新的路由入口的时间.

加快 IP 路由收敛, 应当分别从缩短这 7 部分时间入手. 针对第 1 部分时间, 如果底层故障检测信号对 IP 层不可用, 那么检测时间主要由 IP 层的 HELLO 间隔决定. 传统 IP 网络中的 HELLO 间隔是 10s, 对于特定端口, 连续四个 HELLO 间隔内都没有收到邻居的 HELLO 包就认为该端口失效, 这样故障的检测时间需要 40s[1]. 为了使得总的故障恢复时间控制在毫秒级, 必须缩短 HELLO 间隔. 虽然从产生 HELLO 包所造成的处理器费用和 HELLO 包占用的带宽来看, 加快 HELLO 包的产生速度不会给网络造成太大的负担[18, 19], 但是 HELLO 间隔缩小后, 网络中路由振荡次数急剧增加, 导致网络不稳定, 在文献[20] 中, 仿真得出, 当前互联网中最优的 HELLO 间隔可以缩短到 275ms; 第 2 部分时间用于过滤掉端口状态的振荡式变化, 在文献[2] 的实验中, 端口状态下跳变和上跳变使用的延迟分别为 2s 和 12s; 第 3 部分时间受 LSA 的最小产生间隔(MinLSInterval) 的约束, 这里, MinLSInterval 是为了防止网络设备状态频繁变化

导致过多的控制开销而给出的时间间隔,如果路由器刚刚更新过 LSA,那么下一次该 LSA 的更新至少要等待 MinLSInterval 的时间间隔. LSA 的更新间隔不能太短,否则会导致网络不稳定,例如 OSPF 中规定 MinLSInterval 为 5s[1]; 第 4 部分时间和网络直径有关系,LSAs 在网络中洪泛时,每前进一跳需要的时间在 10ms 到 100ms 之间[15]; 第 5 部分时间将受 SPF 算法运行间隔的约束,因为 SPF 算法的运行过程开销较大,为了防止 SPF 算法的运行过于频繁,商用路由器通常会对 SPF 算法的运行间隔加以限制,有的使用固定值 5s,有的使用可变的间隔,但是这个间隔通常不宜小于 1s[18]; 第 6 部分时间中 SPF 算法的运行时间和网络的规模有关,因为 SPF 算法通常使用 Dijkstra 算法,该算法的复杂度可以达到 $O(l * \log(n))$,这里 l 是网络中的链路数,n 是目的路由器的数目[1]. 在文献[2]的实验中,这部分时间为 100ms 到 400ms 之间,可以采用增量 SPF 计算的方法来减少开销[21]. 而第 6 部分时间中路由表更新的时间可以这样得到: 当前,路由器中更新 20 个路由入口的时间约为 1ms,照此计算,对于影响上千个路由入口的常见故障来说,路由表更新的时间需要几百毫秒的时间; 第 7 部分时间为 1.5s[2].

可见,加快 IP 路由收敛的主要限制因素来自用于确保网络稳定性的延迟. 随着技术的进步,路由器生产厂商已经能够将这些延迟缩小到毫秒级且不影响网络稳定性,但是这些路由器在目前的互联网中并没有广泛部署[21]. 因此,在当前的互联网中,在维持网络稳定性的前提下使 IP 路由的收敛时间低于 1s 还很难做到[20].

2.2　使用主动式故障恢复

在 IP 路由收敛完成之前,数据包会因为目的节点不可达或路由产生环路等原因而被丢弃; 为了满足网络稳定性要求,短暂性的、频繁发作的故障会被过滤掉而不被处理,然而,这些故障正是网络中最常见的故障[2]. 使用主动式故障恢复就是在故障产生之前就计算好替代路径,当故障发生时直接利用替代路径进行通信. 和传统的主动式故障恢复不同,这里的主动式故障恢复不必为替代路径预留网络资源. 使用主动式故障恢复能够在毫秒数量级的时间内完成故障恢复,而且这类方案特别适合解决短暂性的、频繁发作的故障,可以作为 IP 路由收敛完成之前进行故障恢复的一道防线. 因此该类恢复方案的研究得到了长足的发展.

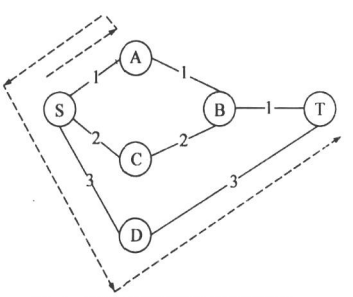

图 1　FIR 中的基于端口转发

文献[13]提出了"故障非敏感路由(Failure Insensitive Routing, FIR)"的方法. FIR 使用基于端口转发的方法进行故障恢复. 如图 1 所示,网络没有故障时 S 到 T 的通信路径是 S-A-B-T. 当链路 A-B 发生故障时,从节点 S 到 T 的包会从 A 返回 S, S 通过检查包的进入端口,可以推断链路 A-B 和 B-T 发生了故障,否则 A 不会将去往 T 的包发到端口 $A \rightarrow S$. 这样节点 S 不必等待故障通知,就可以选择避开故障的通信路径 S-D-T. FIR 特别适合解决短暂性的、频繁发作的单链路故障,由于这类故障是网络中的多发故障,因此 FIR 可以大幅度地提高 IP 网络的故障恢复速度. 在故障发生时,FIR 抑制了故障引发的 IP 路由收敛过程,使用上面提到的基于端口转发的方法确定备份路径,这个确定备份路径的过程可以在故障发生前完成,故 FIR 是一种主动式故障恢复方案.

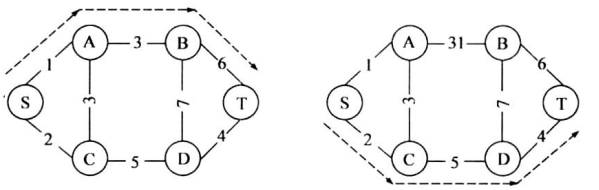

图 2　MRC 中的正常配置和备份配置

文献[16]提出了"多配置路由(Multiple Routing Configurations, MRC)"的方法. MRC 将路由层的节点和链路看成有多个配置,并用 IP 包中的 DSCP 字段来标识这些配置. 如图 2 所示,对于相同的拓扑结构,通过对链路赋予不同的权值,得到两个不同的配置: 正常配置 a 和备份配置 b. 假设链路 A-B 故障,在正常配置 a 中,从节点 S 到 T 的通信路径 S-A-B-T 将受影响,而在备份配置 b 中,将 A-B 的权值设为较大权值(如该配置中所有链路的权值之和). 这样,当 S 运行 SPF 算法计算到 T 的路径时将成功避开链路 A-B,选择通信路径 S-C-D-T. 如果在某个备份配置中将与节点相连的所有链路的权值设为较大值,那么在该备份配置中运行 SPF 算法时,该节点将被避开,故 MRC 还能用于解决节点故障. 在 MRC 的备份配置中运行 SPF 算法确定备份路径的过程可以在故障发生前完成,因此 MRC 也是主动式故障恢复方案.

文献[19]提出了"二出度(Outdegree 2, O2)路由"的方案. O2 路由要求从某个节点出发,对于任何目的节点都至少有两个互不重合的下一跳可以到达. 如图 3 所示,对于目的节点 T,网络中的所有节点都满足 O2 路由的要求. 链路 A-B,在文献[19]中被定义为"百搭链路",只有 A-T 或 B-T 两者中的一条产生故障时才能使用. 在 O2 路由中,

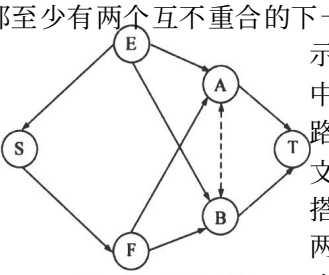

图 3　O2 路由示例

当去往目的节点的某一条路径出现故障时, 可以迅速地将通信切换到另外一条路径上进行. 产生 O2 路由的 O2 算法和传统 SPF 算法中常用的 Dijkstra 算法不同, 可以将 O2 路由看成是一种新的路由协议. 和传统的 IP 路由协议相比, 除了故障恢复的速度快以外, O2 路由的优势在于: 网络的负载更加均衡; 故障发生后在本地完成故障恢复, 而不必在全网内洪泛故障信息, 网络更加稳定. O2 路由可以在故障产生前完成, 因此 O2 路由也是主动式故障恢复方案.

标准组织 IETF 也提出了大量主动式故障恢复方案, 例如: 文献[16]和文献[23]类似, 也为网络配置多个拓扑(Multi-Topology, MT), 通过在备份配置中运行 SPF 算法来建立备份路径; 文献[24]将节点、链路和 SRLG 作为故障单元, 描述了对每个目的节点都给出其备份路径的方案; 文献[25]在检测到故障后, 将数据包封装在"不经过地址"中, 以避开故障单元; 文献[20]使用隧道技术建立备份路径; 在文献[26]中, 由检测到故障的路由器的上游邻居路由器负责建立绕过故障的备份路径.

2.3 提高检测故障的速度与准确性

故障的快速与准确的检测是其能够被及时有效恢复的基础, 而使用 IP 层原有的 HELLO 协议检测故障必须在两个目标之间进行折衷: 一方面, 为了满足故障检测的准确性, 获得全网详细的故障信息, 必须进行 LSAs 的洪泛, 这需要借助于 IP 层的路由收敛过程, 但是这个过程需要的时间较长; 另一方面, 为了加快故障检测速度, 以保证对故障做出及时响应, 就必须抑制 IP 层自身的路由收敛过程, 这必然又会降低故障检测的准确性. 对 IP 网络的快速故障恢复来说, 故障检测速度比准确性更重要, 下面讨论如何在确保检测速度的基础上解决检测不准确带来的问题.

在抑制路由收敛过程之后, 路由器无法从其他路由器接收 LSAs, 这样在路由器的某个端口失效时, 将无法区分究竟是由节点故障还是由链路故障导致的. 如图 4 所示, 源节点 S 向目的节点 T 发送数据包, 无论是节点 A 失效还是链路 S-A 失效, 都会导致 S 收不到 A 的 HELLO 包, 这时 S 不能确定究竟是 A 出了故障还是链路 S-A 出了故障. 这样, 在故障产生时, 存在两种解决方案: (1) 统一假设链路出现故障; (2) 统一假设节点出现故障. 在故障假设发生错误时, 就会出现问题. 如果采取第一种解决方案, 如图 4(*a*)所示, A 失效导致 S 收不到来自 A 的 HELLO 包, S 假设链路 S-A 失效. 因为 A 是 S 去往 T 的关键节点, S 到 T 的路径已经不存在了, 但 S 仍然会建立到 T 的无效路径 S-B-A-T. 替代路径无效是因为故障节点 A 出现在该路径中. 接下来, B 收到来自 S 的数据包后, 会再次假设链路 B-A 失效, 然后继续查找去往 T 的替代路径, 这相当于将单个节点故障替代成多个链路故障分别进行处理, 这种解决方案牺牲了故障恢复的速度, 在文献[16]中采用了这种解决方案; 如果采取第二种方案, 如图 4(*b*)所示, 链路 S-A 失效导致 S 收不到来自 A 的 HELLO 包, S 假设节点 A 失效. 虽然 S 去往 T 的路径还存在, 例如 S-B-A-T 和 S-C-A-T, 但是 S 无法建立去往 T 的替代路径, 这种解决方案夸大了故障, 但是加快了故障检测和恢复的速度, 在文献[27]中采用了这种解决方案.

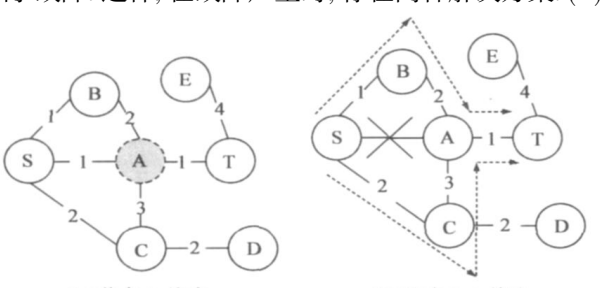

图 4 节点 A 故障或链路 S-A 故障导致端口 S→A 失效

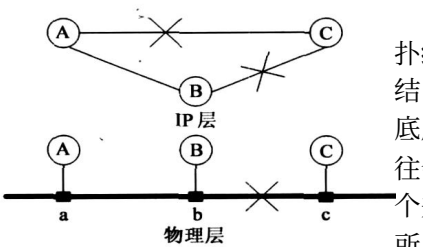

图 5 IP 层和底层拓扑结构不一致对故障检测的影响

由于 IP 层的拓扑结构和底层拓扑结构不完全对应, 底层的单个故障往往会导致 IP 层的多个并发故障. 如图 5 所示, 物理层的链路 b-c 出现故障时, 将同时导致 IP 层的链路 A-C 和 B-C 产生故障. 通过在 IP 层进行 LSAs 的洪泛, 最终 A 能够知道 B-C 也产生了故障, 但为了提高故障检测的速度, 需要抑制这种故障信息在全网的洪泛操作. 这样, 节点 A 通过 HELLO 只能判断 A 无法到达 C, 而无法知道 B-C 也不能相互通信. 同样, A 在选择替代路径时无法把 B-C 排除在外, 就会建立无效的替代路径. 如果将这些并发故障作为多个故障分别进行处理, 可以解决替代路径无效的问题, 但是必然降低故障检测和恢复的效率. 因此需要将这种并发故障作为单个故障单元进行处理. 如果在 IP 层建立的替代路径和底层故障单元是不相交(disjoint)的, 就能够保证替代路径的有效性, 这要求我们做到故障的有效隔离. 在文献[25]中, 利用共风险链路组(Shared Risk Link Group, SRLG)来隔离故障. 网络中的单个设备, 如线卡、管道(duct)等, 产生了故障会导致多条 IP 层的链路失效, 将这些链路划分到同一个集合中就得到了 SRLG. SRLG 中只要有一条链路出现故障, 就认为该 SRLG 中的所有链路都出现了故障. 例如, 图 5 中物理层的链路 b-c 产生故障时, 会导致 IP 层的链路 A-C 和 B-C 失效, 即 SRLG(b-c) = {A-C, B-C}; 再如物理层的节点 b 产生故障会导致 IP 层的链

路 A-B 和 B-C 同时失效, 即 SRLG(b) = {A-B, B-C}. 文献[25]的思路是: 找出和故障单元可能相交的所有链路的集合, 在建立备份路径时避开这些集合. 有的链路同时属于多个 SRLG, 例如链路 B-C 同时属于 SRLG(b-c) 和 SRLG(b), 如果 B-C 产生故障, 需要假定两个集合中的所有链路出现了故障. 这种做法虽然在一定程度上夸大了故障的影响范围, 但是能保证建立有效的替代路径, 同时也使故障检测的速度得以提高. 将 SRLG 作为故障单元能够解决部分多条链路同时发生故障的情况, 这是因为发生故障的链路之间存在联系. 然而, 对于相互之间不存在联系的多条链路同时发生故障的情况, 这种方案就无能为力了.

在 FIR 中使用了一种特殊的故障隔离方式, 即在链路产生故障时, FIR 假设故障链路下游的链路也产生了故障. 如图 1 所示, 链路 A-B 发生故障后, FIR 认为链路 B-T 也产生了故障. 这种方案和使用 SRLG 作为故障单元的做法类似, 也夸大了故障的影响范围. 从图 1 中可看出, 建立的备份路径 S-G-T 不是 S 到 T 的最短路径 S-C-B-T. 显然, 这种做法会降低故障下游链路的利用率.

由此得出: 故障检测速度慢, 故障检测不准确和无法知道底层故障的细节都是路由层 HELLO 协议的局限所造成的. 如果 IP 层能够直接利用底层的故障检测信号, 例如 SONET 层的"警告信号(alarm signal)", 那么故障的检测速度将大大加快[1, 12, 14], 而且知道底层故障的细节, 故障检测的准确性也提高了. 然而, 有的介质并不支持底层的故障检测信号, 例如, 以太网不支持 SONET 的"警告信号", 在这种情况下, 只能依赖于路由层的 HELLO 协议来检测故障[12, 14]. 为了克服 HELLO 协议的局限, 在文献[12]中设计了一种独立于路由层的"双向故障检测协议(BFD)". BFD 的基本原理和简单的 HELLO 协议类似, 在建立连接的两个通信实体之间周期性的互发 BFD 包, 如果一方连续地收不到另一方的 BFD 包, 就认为连接中断. 但是, 和路由层的 HELLO 协议相比, BFD 拥有很多优势. 首先, BFD 适应于各种传输介质; 其次, BFD 可以工作在网络的各个层面, 这为了解底层故障细节提供了途径; 另外, BFD 根据应用的需求建立会话, 并可以实时地协商、调整发包周期, 通过缩短发包周期可以缩短故障检测的时间.

3　推动 IP 网络快速故障恢复方案的实现

IP 网络快速故障恢复的研究取得了很大进展, 但也存在不少局限. 所提出的方案中, 有的适合解决链路故障, 而不擅长解决节点故障; 有的适合解决单链路故障, 而不擅长解决多链路同时发生故障的情况, 等等. 当前, 要推动 IP 网络快速故障恢复方案的实现进程, 需要在三个方面做出努力: (1) 对故障恢复后的通信负载进行均衡, 从而有效地利用网络资源, 防止拥塞的发生; (2) 研究对备份路径进行高效更新的算法; (3) 对快速故障恢复方案进行互操作测试, 设计路由器的体系结构以支持快速故障恢复方案的实现. 下面对这三个方面进行论述.

3.1　故障后的负载均衡

在主动式故障恢复方案中, 主要关注如何缩短故障的恢复时间, 而对于恢复后的通信负载如何在网络中进行均衡分配则考虑的不多. 他们或者没有考虑故障恢复后通信负载的均衡问题(如 FIR), 或者为了便于分析, 假设链路的容量足够大(如 MRC). 在网络产生故障时, 这些方案简单地将故障路径上的通信转移到备份路径上传送, 这种转移容易造成备份路径的拥塞. 为了避免这种转移造成通信负载的不均衡分配, 导致网络吞吐能力的下降, 需要对故障后的网络实施流量工程[28, 29]. 通过 LSP 配置和资源预留, 在 MPLS 中可以方便地进行流量工程, 避免通信负载失衡[40], 但是在纯 IP 网络中情况将有所不同, 流量工程需要通过合理地设置链路权值来实现[30~33].

用有向图 $G(N, A)$ 来表示网络, 链路 a 的容量为 $c(a)$, 通信负载为 $l(a)$, 链路的利用率为 $u(a) = l(a)/c(a)$. 直观地讲, 流量工程的目标是对于任意的 $a \in A$, 维持 $u(a) \leqslant 1$ 成立, 但这个目标过于笼统. 在文献[33]中定义了描述链路费用的函数:

$$\Phi(l(a)) = \begin{cases} l(a), & 0 \leq u(a) < 1/3 \\ 3l(a) - 2c(a)/3, & 1/3 \leq u(a) < 2/3 \\ 10l(a) - 16c(a)/3, & 2/3 \leq u(a) < 9/10 \\ 70l(a) - 178c(a)/3, & 9/10 \leq u(a) < 1 \\ 500l(a) - 1468c(a)/3, & 1 \leq u(a) < 11/10 \\ 5000l(a) - 16318c(a)/3, & 11/10 \leq u(a) < \infty \end{cases}$$

其中, $a \in A$, $\Phi(0) = 0$. 显然, 随着链路利用率的增加, 费用也在增加, 而且增加的速度不断加快, 尤其当链路利用率超过 1(即链路发生拥塞)时. 将 $\sum_{a \in A} \Phi(l(a))$ 作为目标函数, 并寻求其最小值, 可将流量工程转化为线性规划问题来解决. 链路权值的设置决定了 SPF 算法的运行结果, 进而决定通信负载在网络中的分配, 最终决定了目标函数的取值. 寻求最优的权值设置已被证明是 NP 难问题[33], 需要借助启发式算法来设置链路权值.

文献[27]和[28]中分别提出在备份配置中通过权值的设置进行流量工程的设想. 在文献[28]中, 首先确定备份配置中费用较大的关键链路集合 L_C, 然后将备份配置中每条链路的权值设置为某个随机数, 通过多次试探来寻求减小目标函数 $\sum_{a \in A} \Phi(l(a))$ 的权值设置; 文献[27]在确定 L_C 之后试探性地增加 L_C 中链路的权值, 以转移其通信负载. 和文献[33]相比, 针对可能产生的故障, 在相应的备份配置中进行独立的权值设置, 更有利

于解决由于故障路径向备份路径转移通信流量所造成的通信负载失衡的问题.

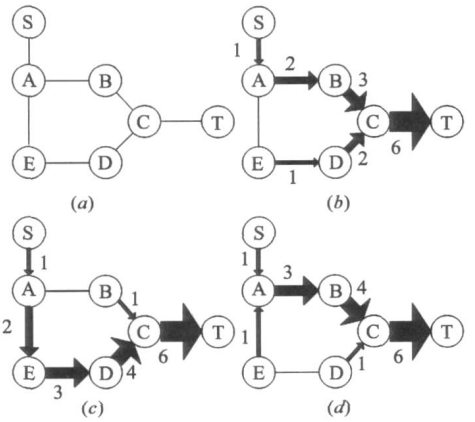

图 6 路由计算与链路利用率互反馈导致网络进入振荡状态
(图中用箭头粗细表示通信负载的大小)

需要指出, 在上述链路权值调整的过程中, 路由计算依赖于链路利用率, 同时又影响链路利用率, 二者之间的互反馈容易导致网络进入振荡状态. 根据文献[27, 28, 31~33]中权值设置的原则, 链路利用率越大, 应当设置的权值也越大, 以便于其转移通信负载. 为简化分析, 我们假设每条链路的容量相同, 而链路权值设置为:

$$w(a) = k * l(a) + d$$

其中, k 和 d 为常数. 造成网络振荡的例子不难构造. 如图 6 所示, 假设某网络处于状态 a 时, 节点 S, A, B, C, D 都要向目的节点 T 进行负载量为 1 的持续通信. 根据 SPF 算法建立的通信路径如状态 b 所示; 对处于状态 b 下的网络进行链路权值的调整, 链路 S-A, A-B, B-C, C-T, C-D, D-E, A-E 的权值分别设置为 $k+d$, $2k+d$, $3k+d$, $6k+d$, $2k+d$, $k+d$, d, 再根据SPF 算法建立的通信路径如状态 c 所示; 同理, 由状态 c 得到状态 d. 但是在对状态 d 下的网络进行权值调整后, 网络又返回状态 c. 这样网络将在状态 c 和状态 d 之间来回振荡. 实际上, 只要链路权值随着链路利用率单调增加都可能导致网络进入振荡状态[34]. 在实际的网络中, 可能产生的振荡情况更加复杂, 需要采取措施避免振荡的发生.

3.2 对 IP 网络快速故障恢复方案的支持

实现 IP 网络快速故障恢复方案不可避免地要对现有 IP 协议及路由器的体系结构做出改动. 加快 IP 路由收敛的方案需要缩短 HELLO 协议中包的发送间隔, 并缩短 LSAs 产生、更新以及运行 SPF 算法的计时器延迟. 在主动式故障恢复方案中, 所作的改动有: FIR 需要识别包的进入端口, 而当前在 IP 层进行 SPF 路由查找时不考虑包的进入端口; O2 路由算法中使用了一种非 SPF 的路由查找算法; MRC 和 MT 需要使用 IP 包的业务区分字段(Differentiated Services Field[35], DS 字段)来标识不同的拓扑配置, 但是 DS 路由在互联网中一直很少被采用[1]. 加快故障检测速度的方案需要抑制 IP 层的路由收敛过程. BFD 使用独立于 IP 路由协议的故障检测协议来保证故障快速准确检测.

优秀的协议应当满足对该协议进行独立地实现时, 各个不同的实现版本之间能够协调工作[1], 为此, 需要支持 IP 网络快速故障恢复方案的设备生产商之间进行互操作测试, 通过在大量路由器构成的拥有复杂拓扑的实际网络中运行快速故障恢复方案来不断发现并改进协议中存在的不足[36]. 设备生产商可以定期地举办会议, 也可以借助专门的平台来进行互操作测试, 例如新罕布什尔大学互操作性实验室[35]. IP 网络快速故障恢复方案的逐渐成熟并在互联网中广泛采用需经历一个长期过程.

从路由器角度来看, 为支持快速故障恢复方案的实现, 需要提高路由器的性能, 改进路由器的体系结构. 为支持 IP 路由的快速收敛, 以及使用 BFD 来实现故障的快速检测, 必须提高路由器的处理器速度. 为支持主动式故障恢复, 需要在路由器的转发表中存储备份路径的下一跳地址信息, 这将占用路由器额外的存储资源. 主动式故障恢复方案需要计算备份路径. 在 FIR 中, 路由

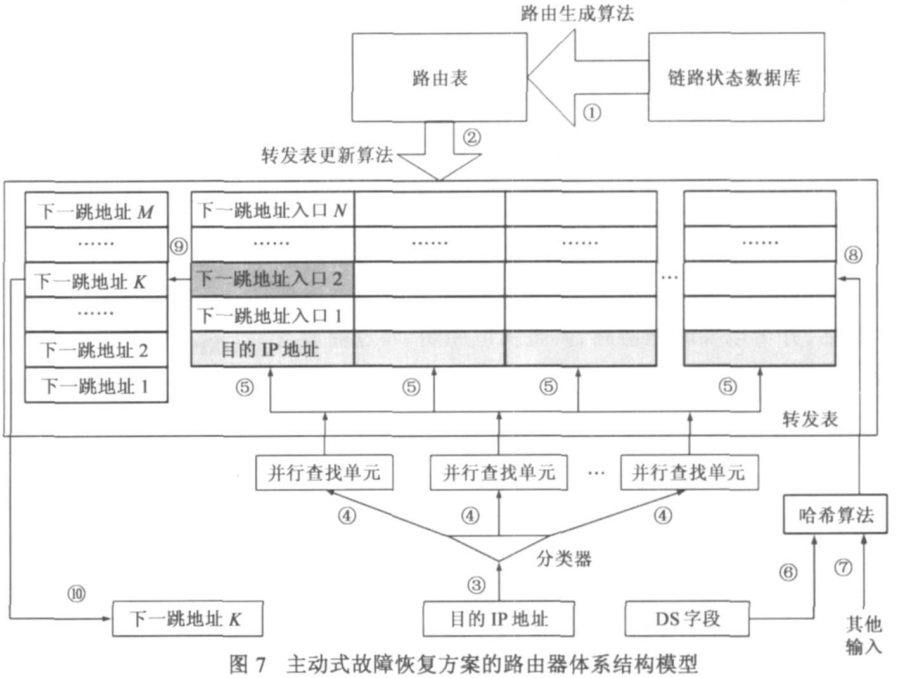

图 7 主动式故障恢复方案的路由器体系结构模型

器需要判断到达包的进入端口, 在 MRC 和 MT 中, 路由器需要读取和修改包的 DS 字段, 这也将占用路由器额外的计算资源. 在实现主动式恢复方案时, 为节省路由器有限的存储资源和计算资源, 同时保证其高吞吐率, 需要设计新的路由器体系结构. 参考文献[27]和[38], 我们在图 7 给出了主动式故障恢复方案的路由器体系结构模型. 其中: (1) 路由器运行"路由生成算法"从"链路状态数据库"生成"路由表"."路由表"中既存储了工作路径的下一跳地址, 也存储了备份路径的下一跳地址. (2)"路由表"中的信息经过压缩后, 通过"转发表更新算法"插入到路由器线卡的转发表中. 如何组织网络中的路由器按序更新自己的转发表, 防止更新过程中产生路由环路, 目前还是一个新的研究方向[39]. 对于特定目的地址, 在转发表中存在多个下一跳地址可以到达. 因为下一跳地址需要维护 MAC 地址等信息, 重复存储下一跳地址会占用转发表额外的存储空间[27], 所以模型中使用它在下一跳地址表中的入口来代替, 如图 7 中的"下一跳地址入口 1"到"下一跳地址入口 N". 在文献[27]中就采用了这种方法. (3) 路由器通过查找"目的 IP 地址"来确定下一跳地址入口. (4) 为了提高查找速度, 通过"分类器"将查找任务发射到"并行查找单元"中[39]. (5)"并行查找单元"根据"目的 IP 地址"查找转发表, 得到多个下一跳地址入口. (6) 在 MRC 和 MT 中哈希算法的输入是 IP 包的 DS 字段. (7) 作为其他输入, 在 FIR 中, 哈希算法的输入是到达包的进入端口的信息; 在 $O2$ 中, 输入是工作路径产生故障与否的信息, 等等. (8) 为了在"下一跳地址入口 1"到"下一跳地址入口 N"中做出快速选择, 需要使用哈希算法的输出. (9) 根据哈希算法的输出选出下一跳地址入口, 并访问下一跳地址表. (10) 得到真正的下一跳地址.

4　总结

通过本文的论述可以得到以下结论:

(1) 为提高互联网的性能, 必须增强 IP 网络故障恢复能力. IP 层的路由收敛使互联网拥有一定的健壮性, 但是, 因为要在全网洪泛故障信息, 路由收敛过程需要的时间过长, 低下的故障恢复速度无法满足实时业务、复用程度高的业务以及关键性业务的需求. 互联网拓扑结构变化频繁, 要保证其可靠性, 必须加快其故障恢复速度.

(2) 提高 IP 网络的故障恢复速度主要有三条途径: (a) 加快 IP 路由收敛; (b) 使用主动式故障恢复; (c) 统一假设链路、节点、SRLG 或下游链路产生故障, 或者使用独立的故障检测协议, 如 BFD 等, 以提高故障检测的速度与准确性.

(3) 为推动 IP 网络快速故障恢复方案的实现, 需要做好以下工作: (a) 通过调整链路权值来解决 IP 网络中故障恢复后通信负载失衡的问题, 但同时要防止网络振荡的发生; (b) 通过互操作测试促进快速故障恢复方案协议簇的成熟, 重新设计路由器的体系结构以支持快速故障恢复方案的实现.

参考文献:

[1] John T M. OSPF: Anatomy of an Internet Routing Protocol [M]. Massachusetts, USA: Addison Wesley Longman, Inc, 1998.

[2] Iannaccone G, Chuah G-N, Mortier R, et al. Analysis of link failures in an IP backbone[A]. Proceedings of ACM SIGCOMM Internet Measurement Workshop, , Marseille, France, 2002[C]. New York, USA: ACM Press, 2002. 237- 242.

[3] Sahasrabuddhe L, Ramamurthy S, Mukherjee B. Fault management in IP-over-WDM networks: WDM protection versus IP restoration[J]. IEEE Journal on Selected Areas In Communications, 2002, 20(1): 21- 33.

[4] Zhou DY, Subramaniam S. Survivability in optical networks [J]. IEEE Network, 2000, 14(6): 16- 23.

[5] Park J T. Resilience in GMPLS path management: model and mechanism[J]. IEEE Communications Magazine, 2004, 42(7): 128- 135.

[6] Awduche D, et al. Overview and principles of Internet traffic engineering[S]. Internet RFC 3272, May 2002.

[7] Moore D, Shannon C, Claffy K. Code-Red: a case study on the spread and victims of an Internet worm[A]. Proceedings of ACM SIGCOMM Internet Measurement Workshop, Marseille, France, 2002[C]. New York, USA: ACM Press, 2002. 273- 284.

[8] Moore D, Shannon C, Voelker G, et al. Internet quarantine: requirements for containing self-propagating code[A]. Proceedings of INFOCOM[C]. San Francisco, CA, USA: IEEE Press, 2003. 1901- 1910.

[9] Nate Kushman, Srikanth Kandula, Dina Katab, et al. R-BGP: Staying connected in a connected world[A]. NSDI 2007[C]. Cambridge, Massachusetts, USA, Proceedings. USENIX, 2007.

[10] 石兵, 周明天. 宽带网络中预先恢复路由配置的研究[J]. 电子学报, 2004, 32(7): 1209- 1211.
Shi Bing, Zhou Ming tian. Study of using pre-configured alternative route in broadband network [J]. Acta Electronica Sinica, 2004, 32(7): 1209- 1211. (in Chinese)

[11] S Bryant, C Filsfils et al. IP fast reroute using tunnels[Z]. IETF Internet Draft (work in progress), Apr 2005.

[12] Katz D, Ward D. Bidirectional forwarding detection[Z]. IETF Internet Draft (work in progress), June 2006.

[13] Lee S, Yu YZ, Nelakuditi S, et al. Proactive vs reactive approaches to failure resilient routing[A]. Proceedings of INFOCOM[C]. Hong Kong: IEEE Press, 2004. 7- 11.

[14] Ken O, Vishal S, Mathew O. Network survivability considera-

tions for traffic engineered IP networks[Z]. IETF Internet Draft (work in progress), May 2002.
[15] Shand M. IP fast reroute framework[Z]. IETF Internet Draft (work in progress), June 2005.
[16] Kvalbein A, Audun Fosselie Hansen, Cicic T, et al. Fast IP network recovery using multiple routing configurations[A]. Proceedings of INFOCOM[C]. Barcelona, Spain, IEEE Press, 2006. 1- 11.
[17] Pierre F, Clarence F, John E, et al. Achieving sub-second IGP convergence in large IP networks[J]. ACM SIGCOMM Computer Communication Review, 2005, 35(2): 5- 44.
[18] Alattinoglu C, Jacobson V, Yu H. Towards milli-second IGP convergence[Z]. IETF Internet Draft (work in progress), Nov 2000.
[19] Schollmeier G, Charzinski J, Kirstdter A, et al. Improving the resilience in IP networks[A]. Proceedings of High Performance Switching and Routing, Torino, Italy, 2003[C]. USA: IEEE press, 2003. 91- 96.
[20] Basu A, Riecke J. Stability issues in OSPF routing[A]. Proceedings of ACM Special Interest Group on Data Communication, San Diego, California, United States, 2001[C]. New York, USA: ACM Press, 2001. 225- 236.
[21] Iannaccone G, Chuah C-N, et al. Feasibility of IP restoration in a tier 1 backbone[J]. IEEE Network, 2004, 18(2): 13- 19.
[22] Markopoulou A, Iannaccone G, et al. Characterization of failures in an IP backbone[J]. IEEE Network, 2004, 18(2): 13- 19.
[23] Psenak P, Mirtorabi S, Roy A. Multi-Topology (MT) routing in SPF[Z]. IETF Internet Draft (work in progress), February 2006.
[24] Atlas A. Basic Specification for IP fast-reroute: loop-free alternates[Z]. IETF Internet Draft (work in progress), February 2006.
[25] Bryant S, Shand M, Previdi S. IP fast reroute using not-via addresses[Z]. IETF Internet Draft (work in progress), Oct 2005.
[26] Atlas A. U-turn alternates for IP/LDP fast-reroute[Z]. IETF Internet Draft (work in progress), Feb 2005.
[27] Apostolopolous G. Using Multiple Topologies for IP-only Protection Against Network Failures: A Routing Performance Perspective[R]. Crete, Greece: ICS-FORTH, 2006.
[28] Kvalbein A, Cicic T, Gjessing S. Post-failure routing performance with multiple routing configurations[A]. Proceedings of INFOCOM[C]. Anchorage, Alaska, USA: IEEE Press, 2007. 98- 106.
[29] Awduche D, et al. Overview and principles of Internet traffic engineering[Z]. IETF Internet Draft (work in progress), 2001.
[30] Fortz B, Thorup M. Traffic engineering with traditional IP routing protocols[J]. IEEE Communications Magazine, 2002, 40(10): 118- 124.
[31] Fortz B, Thorup M. Optimizing OSPF/IS-IS weights in a changing world[J]. IEEE Journal on Selected Areas in Communications, 2002, 20(4): 756- 767.
[32] Nucci A, Schroeder B, Bhattacharyya S, et al. IGP link weight assignment for transient link failures[R]. Burlingame, CA 94010, USA: Sprint Advanced Technology Labs, Technical report: TR02-ATL-071000, 2003.
[33] Fortz B, Thorup M. Internet traffic engineering by optimizing OSPF weights[A]. Proceedings of INFOCOM[C]. Tel Aviv, Israel: IEEE Press, 2000. 519- 528.
[34] Bertsekas D, Gallager R. Data Networks[M]. 2nd ed. Englewood Cliffs, New Jersey: Prentice Hall 1992.
[35] Nichols K, et al. Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers[S]. Internet RFC 2474, December 1998.
[36] Hinden R. Internet Routing protocol standardization criteria [S]. Internet RFC 1264, October, 1991.
[37] University of new Hampshire InterOperability lab[DB/OL]. http://www.iol.unh.edu/, 2007-07-01/2007-07-10.
[38] Zheng K, Hu C C, Lu H B, et al. An ultra high throughput and power efficient TCAM based IP lookup engine[A]. Proceedings of INFOCOM[C]. Hong Kong: IEEE Press, 2004. 1984-1994.
[39] Francois P, Bonaventure O, Shand M, et al. Loop-free convergence using oFIB[Z]. IETF Internet Draft (work in progress), Oct 2006.
[40] 倪 华, 唐宝民. MPLS 网络多路径动态流量分配的研究[J]. 电子学报, 2005, 33(4): 718- 720.
Ni Sur hua, Tang Bao-min. Dynamic traffic partitioning on MultiPath in MPLS systems[J]. Acta Electronica Sinica, 2005, 33(4): 718-720. (in Chinese)

作者简介:

张民贵 男, 1980 年生于山东青岛. 清华大学计算机科学与技术系博士研究生, 研究方向为网络故障恢复、网络安全.
Email: Zmg06@mails.tsinghua.edu.cn

刘 斌 男, 1964 年生于山东临朐, 清华大学计算机系教授, 博士生导师. 主要研究领域为高性能路由器、ATM 交换结构与理论、ISDN 交换技术、网络处理器和网络安全等.