软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(1):65−81 [doi: 10.3724/SP.J.1001.2012.04119]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

域间多路径路由协议*

苏金树+, 戴 斌, 刘宇靖, 彭 伟

(国防科学技术大学 计算机学院,湖南 长沙 410073)

Inter-Domain Multipath Routing Protocols

SU Jin-Shu+, DAI Bin, LIU Yu-Jing, PENG Wei

(School of Computer, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: E-mail: sjs@nudt.edu.cn, http://www.nudt.edu.cn

Su JS, Dai B, Liu YJ, Peng W. Inter-Domain multipath routing protocols. *Journal of Software*, 2012,23(1): 65−81. http://www.jos.org.cn/1000-9825/4119.htm

Abstract: BGP (border gateway protocol) is widely known for some problems in terms of poor reliability, suboptimal path use, and insufficient support for load balancing because it is a single-path routing protocol. Inter-Domain multipath routing explores the underlying network AS-level path diversity to improve the Internet's reliability, performance, and resource utilization. Thus, inter-domain multipath routing is considered a useful and necessary method to address the problems faced by BGP. This paper surveys current proposals on inter-domain multipath routing protocols and classifies these protocols into three categories: Protocols on a single announcement and multipath forwarding, protocols on multiple announcements and multipath forwarding, and new Internet routing architecture based protocols. They are compared under some different features of path diversity, control message overhead, loop-freeness property, etc. In addition to a review of existing protocols, the challenges in designing new inter-domain multipath routing protocols that could be taken as the future research direction are pointed out.

Key words: BGP (border gateway protocol); inter-domain multipath routing; network performance; network reliability

摘 要: 边界网关协议(border gateway protocol,简称BGP)是当前互联网的核心协议,但是由于BGP是一种单路径路由协议,所以仍存在可靠性差、无法有效使用次优路径以及负载均衡支持较弱等问题.域间多路径路由可以通过发挥底层网络的 AS 级路径多样性,提高域间路由的可靠性、报文分组转发的总体性能和整个网络资源的利用率.因此,域间多路径路由是解决上述 BGP 问题的一种有效手段,符合互联网应用不断深入、促进路由技术发展的需求.主要综述域间多路径协议,并将其分为 3 类:单径通告多路转发协议、多径通告多路转发协议和新型域间多路径路由体系结构.提出路径多样性、控制平面和数据平面开销、无环路特性等 8 项主要路由系统性能指标,并比较、分析了域间多路径路由协议.最后,指出域间多路径路由协议面临的主要挑战和未来的研究方向.

关键词: 边界网关协议(BGP);域间多路径路由;网络性能;网络可靠性

中图法分类号: TP393 **文献标识码**: A

* 基金项目: 国家自然科学基金(61070199, 61003301); 国家重点基础研究发展计划(973)(2009CB320503)
收稿时间: 2011-02-16; 定稿时间: 2011-08-31; jos 在线出版时间: 2011-10-11
CNKI 网络优先出版: 2011-10-11 14:12, http://www.cnki.net/kcms/detail/11.2560.TP.20111011.1412.001.html

互联网逐渐成为以全球信息化为广泛背景的人类社会发展和繁荣的关键基础设施.随着互联网应用种类的快速增长,网络用户对路由协议的可靠性和健壮性提出了更高要求,而现有的 BGP(border gateway protocol)[1]协议无法满足日益增长的可靠性和性能的需求.大量研究表明,BGP 可靠性差直接导致互联网可靠性差.例如,Labovitz 等人[2]指出:两分钟内的一个路由变化可产生大约 30%的报文丢失;许多 IP 地址由于 BGP 的动态性而在短时间内不可访问[3];Kushman 等人[4]通过实验得出,有多半 VoIP 故障发生在 BGP 更新报文出现后的 15 分钟内.

互联网的可靠性很大程度上取决于路由协议在链路(或节点)故障发生后,重新获得可用路径所需的反应时间.出于可扩展性和稳定性的考虑,当前互联网路由协议通常只选择一条“最佳”路径到达目的地,如域内协议 OSPF[5]和 ISIS[6]、域间协议 BGP.单路径路由协议,在链路(或节点)发生故障时,不具备故障瞬时恢复的能力.在链路(或节点)故障等引发网络故障后,单路径路由协议需要一定延迟才能完成路由重建,恢复正常的数据通信.而此类延迟可能会导致瞬时路由环路或者瞬时路由失效等路由故障[2,7].为减少延迟,人们提出了一些减少路由协议收敛时间的方法[8−10],但只依靠降低收敛时间减少延迟,在实际应用时是不可行的[11].多路径路由在主路由出现故障后,能够快速提供备用路由,保证网络通信会话不被中断,从而成为提高网络可靠性的一种有效方法.

与单路径路由相比,域间多路径路由具备下列优点[12]:

(1) 多路径路由能够提高网络可靠性,当主路径失效时,备用路径可以立即启用,而无需等待路由协议的重新收敛,从而避免了瞬时的数据丢失.

(2) 边缘 AS(autonomous system,自治系统)可以通过多条路径进行数据传输,从而可以采取更加灵活的负载均衡策略.

(3) 多路径路由可以增强数据传输的安全性[13].例如,一个 AS 为防止恶意数据侦听,可将数据分成多个部分,并通过不同的路径进行传输.

本文的主要贡献是:提出了 8 个性能指标,构成一组衡量域间多路径路由协议的指标体系;提出了域间多路径路由协议的多路径发现、路径选择和报文转发这 3 个核心问题;提出了域间多路径路由协议分类方法,即划分为单径通告多路转发协议、多径通告多路转发协议和新型域间多路径路由体系结构;综合分析了属于这 3 类协议的主流相关算法;最后,指出域间多路径路由协议面临的主要挑战和未来的研究方向.

1 域间多路径路由协议指标体系与核心问题

尽管研究者普遍认为多路径路由是一种提高可靠性、安全性以及网络性能的解决方法,但是多路径路由,尤其是域间多路径路由,与完全部署之间仍然存在一定的距离.设计域间多路径路由协议,必须研究多路径路由的性能指标.综合当前研究,本文主要采用以下 8 个性能指标进行分析比较:

(1) 路径多样性(path diversity).路径多样性是指网络层路径多样性.网络层发掘链路层路径多样性的突出优势是,在链路/节点故障、策略改变等情况下,只要网络链路层拓扑存在连接就能保证端到端的连续正常通信.

(2) 控制平面开销(control-plane overhead).控制平面开销分为路由消息开销和路由平面开销:路由消息开销是指相邻 AS 之间交换路由更新的数目.具体地,路由消息开销分为路径建立消息开销和网络故障引起的路由更新消息开销;路由平面开销是指存储路由的内存开销和更新路由表的计算开销.路由消息的交换会占用链路带宽,同时,计算和存储更多路由需要占用更多的路由器计算资源和存储资源.

(3) 数据平面开销(data-plane overhead).数据平面开销分为两部分:报文携带信息(报文头大小)和转发表信息(转发表大小).与数据平面开销密切相关的是转发操作方式(forwarding operation),为了使同源同目的地址的报文通过不同的路径,要求路由器能够将报文发往不同的下一跳地址,就需要改变或扩展路由器转发表.

(4) 是无环路特性(loop-freeness).无环路特性是指当网络动态发生变化时,路由协议检测路由环路的能力.转发环路会导致网络资源的浪费甚至导致报文的不可达,因此,路由协议的设计需要保证无环路特性.BGP 中使用 AS 路径通告来保证路由不包含环路.域间多路径路由在多路径转发的条件下确保无环路特性的能力,是衡量路由协议正确性和性能的重要指标.

(5) 增量部署能力(incremental deployment).增量部署是衡量一个协议能否快速部署的重要指标,主要表现在与现有 BGP 协议的兼容性上.

(6) 策略表达(policy expressiveness).策略表达体现了 AS 对路由的控制能力.AS 期望协议支持丰富、灵活的策略.策略表达能力可以通过协议所允许的端到端的路由多样性来表示.

(7) 可扩展性(scalability).域间路由系统复杂而庞大,因此,域间路由协议的可扩展性是重点考量指标.

(8) 用户对路由的控制能力(route control for users).单路径路由的一个缺陷是用户对路由没有选择权.多路径路由协议提供用户多条路由,因此,能否支持用户根据需要进行路由选择是非常重要的功能.然而,用户对路由的控制还需要充分考虑 ISP 流量工程的目标,否则会引起流量震荡.

需要说明的是,上述 8 项性能指标并不是衡量域间多路径路由协议的全部指标,而是我们所认为的最能体现各种路由协议之间差异的评价标准.另外,一些没有纳入指标体系的指标包括:路径构造延迟(latency of constructing paths)、协议收敛时间(convergence time)和安全性(security).路径构造延迟是指多路径尤其是备用路径建立的时间;协议收敛时间是指当网络状态发生变化后,网络节点到达统一转发视图需要的时间;安全性能是指网络协议检测网络中恶意节点或者攻击的能力.此外,各项指标之间存在一定的联系,比如,控制平面和数据平面的开销与协议的可扩展性密切相关,协议开销越小,可扩展性越强.

在实现多路径路由方法的内部机制时,需要综合考虑多路径发现、路径选择和报文转发这 3 个核心问题:

(1) 多路径发现问题.如何让边缘 AS 掌握网络中存在的到达同一个网络前缀的多条 AS 路径是域间多路径路由协议需要解决的首要问题.如图 1 所示,多路径发现方法可以分为 3 种:第 1 种方法是基于链路状态协议.链路状态协议的网络拓扑广播功能使得每个网络节点都可获取整个网络的拓扑.因此,采用此种方法的多路径协议可以获得最大程度的路径多样性.然而,由于域间路由系统的规模庞大,采用链路状态协议作为多路径发现的主要问题是由大量链路通告导致的可扩展性比较差;第 2 种方法是基于路径向量协议.BGP 是典型的路径向量协议,它只允许路由器向邻居通告一条“最好”的路径.因此,一种最简单的多路径发现方法是允许路由器通告给邻居多条路径.但是,通告多条路径容易导致更多的消息及存储开销.因此,一些多路径协议选择控制路径通告数目的方式或者在通告单一路径的基础上采用其他方法进行路径发现;第 3 种方法是基于主动多路径探测技术.这种多路径发现过程通常由边缘路由器发起,主动进行多路径的探测和发现.采用这种方法的协议通过对等路由器(peering router)的显式请求方式交换多路径信息.

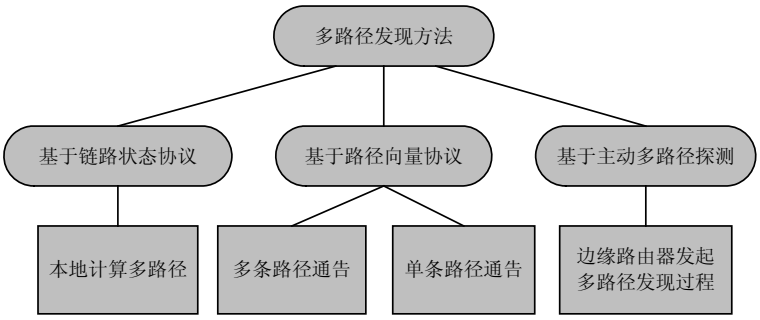

Fig.1 Methods of discovering multiple paths

图 1 多路径发现的方法

(2) 路径选择问题.域间多路径协议必须解决如何从多条路径中选择主路径和备用路径的问题,以保证达到较大的路径多样性和最低的消息开销.主路径选择方法一般采用 BGP 的“最好”路径选择方法.备用路径选择一般基于最大不相交节点或者最大不相交边的路径,也就是与默认路径边或者节点最不同的路径.也可以基于其他一些路径特征,比如像延迟、带宽以及报文丢失等,文献[14]提出了“路径可用性”作为路径选择的依据.

(3) 报文转发问题.由于存在到达同一网络前缀的多条路径,因此传统的基于目的地的逐跳转发方法不能

有效地工作.在多路径背景下,我们可以采用如图 2 所示的几种方法进行报文转发.源路由中发送者仅需将路径信息置于报文头中,途经的路由器通过读取的路径信息就可以进行报文分组的转发.基于目的地的路由则可以采用路径标识的方法,路径标识用来区分到达同一目的地的不同路径,根据设计者的选择,可以是局部的,也可以是全局的.基于目的地的路由除了采用路径标识以外,也可以通过 IP 隧道、MPLS 和虚拟接口进行有效的报文分组转发.

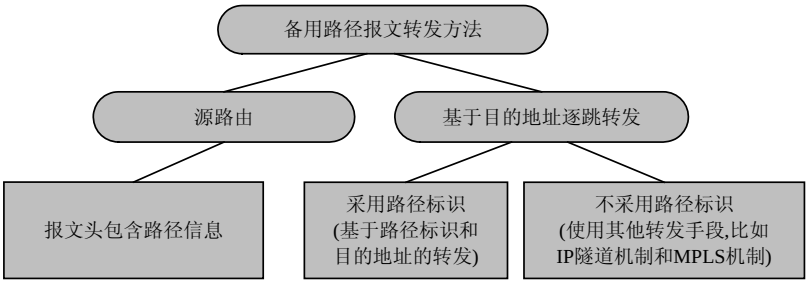

Fig.2 Methods of packet fowarding

图 2 报文转发方法

2 域间多路径路由协议分类

目前,域间多路径路由虽然是学术界研究的热点问题,但至今还没有一个被广泛接受的定义.我们认为,如果一个域间路由协议针对同一个网络前缀的报文可以进行多路径转发,则该路由协议就被称作是域间多路径路由协议.多路径转发要求协议在转发操作时对同一个网络前缀有多个不同的下一跳地址,并需要有协议在控制平面上的支持.围绕域间多路径路由协议的 3 个核心内部机制问题,根据协议如何获取多路径以及在控制平面的不同操作手段对域间多路径路由协议进行分类是比较直观和科学的,即将域间多路径路由协议分为 3 类:单径通告多路转发协议、多径通告多路转发协议和新型域间多路径路由体系结构,如图 3 所示.其中:单径通告多路转发协议是指针对某个特定网络前缀,AS 只允许选取一条 AS 路径并使用一个 BGP 路由更新通告给邻居;多径通告多路转发协议是指针对某个特定网络前缀,AS 允许选取多条 AS 路径并采用多个 BGP 路由更新通告给邻居.这两类域间路由协议采用的都是 BGP 路由更新报文,因此可以被认为是 BGP 基础上的扩展协议.最后一类协议类型是基于所提出的新型路由体系结构来实现多路径路由.在后续章节中,本文选取重要的、具有典型意义的域间多路径路由协议进行分析论述.

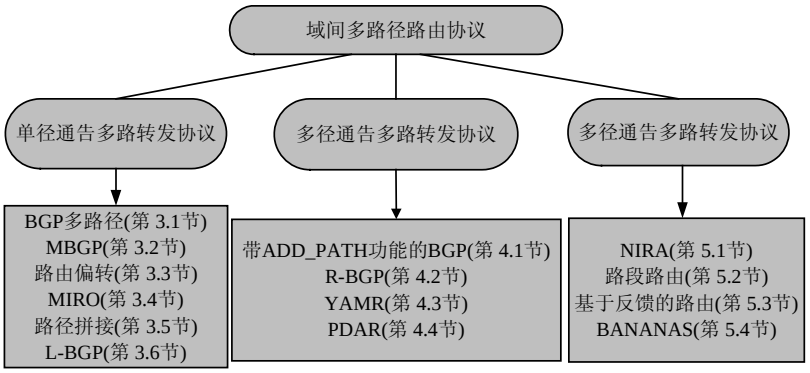

Fig.3 Classification of inter-domain multipath routing protocols

图 3 域间多路径路由协议的分类

3 单径通告多路转发协议

本节主要介绍单径通告多路转发协议.这类协议的特点是每个 AS 针对每个网络前缀,只向邻居通告一条路径,但是报文则可以通过多条路径进行传送.单径通告可以降低路由器CPU的处理负荷,具有较低的网络通信开销,但路径多样性会受到一定的限制.

3.1 BGP多路径

路由器主要生产商之一 Cisco 提出 BGP 多路径[15]的思想,允许路由器针对同一个网络前缀在路由表和转发表中安装满足条件的多条路由.Cisco 路由器对路径选取有非常严格的条件,多路径的候选路径必须与最佳路径具有相同的通告源、本地优先权值、MED 值以及 AS 路径长度等.BGP 多路径的配置比较简单,只需要使用控制命令 maximum-paths 限定多路径的最大允许数量.然后,满足相应条件的多条路径就被安装到路由表和转发表中.当负载均衡功能开启后,网络流量就可以通过多条路径进行转发.另一个路由器生产厂商 Juniper 也提出类似方案[16].

BGP 多路径的突出优点是简单,并且得到大量实际部署路由器的支持.然而,严格的多路径选取条件减少了可利用的多路径数量,从而限制了路径的多样性.在多路径发现上,BGP 多路径利用 BGP 所发掘的路径信息;在路径选择上,BGP 多路径采用严格的路径比较;在报文转发上,BGP 多路径采用的是流量分发.这种方法是最简单的利用多路径的方法.

3.2 MBGP(multipath BGP)

MBGP[17]的主要思想是,基于 BGP 协议建立的路径采用基于源的主动探测方法发现路径多样性.MBGP 定义了 3 类路由器:源 MBGP 路由器、宿 MBGP 路由器以及 MBGP 路由器.源 MBGP 路由器是指用户所在的 AS 中距离用户最近的运行 MBGP 的路由器;宿 MBGP 路由器是指在与源 MBGP 路由器进行通信的目标用户所在的 AS 中,距离目标用户最近的运行 MBGP 的路由器;MBGP 路由器是指除源 MBGP 路由器和宿 MBGP 路由器以外,所有运行 MBGP 协议的路由器.由源 MBGP 路由器发起多路径发现操作,通过其他 MBGP 路由器的配合,实现多路径的发现与安装.MBGP 的多路径发现建立在 BGP 的最佳路径选择完成的基础上,因此与 BGP 是兼容的.下面使用如图 4 所示的网络拓扑详细阐述 MBGP 的多路径发现机制.

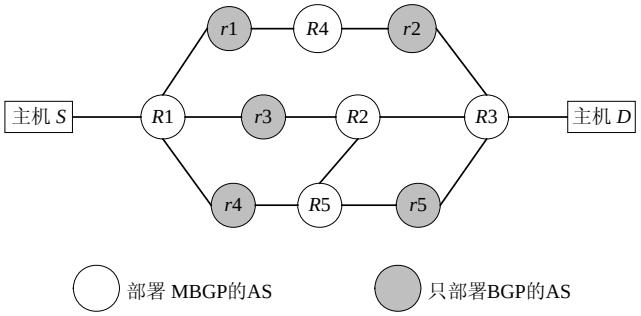

Fig.4 An example of MBGP

图 4 MBGP 示例

在图 4 中,用大写字母表示 MBGP 路由器,小写字母表示 BGP 路由器.假设主机 *S* 是源,主机 *D* 是目标.*R*1 是源 MBGP 路由器,*R*3 是宿 MBGP 路由器.假设从主机 *S* 到主机 *D* 的主 BGP 路径是 *R*1-*r*3-*R*2-*R*3.源 MBGP 路由器 *R*1 发送一个路径发现探测报文给宿 MBGP 路由器 *R*3.为发现所有的路径,*R*1 向所有的邻居路由器广播路径发现探测报文.若探测报文接收者是 MBGP 路由器,则探测报文将记录经过的路由器信息;否则,探测报文被当作普通 IP 报文进行转发.当多个探测报文到达 *R*3 时,*R*3 根据探测报文中记录的路由器信息就可以知道从 *R*1 出发的多条路径.然后,*R*3 将这些路径信息通过特定报文发回给 *R*1.值得注意的是,当 *R*1 收到这些路径时,并不

能使用这些路径,因为这些路径中包含的路由器可能并没有将有些路径安装到转发表中.由此,MBGP 定义了一个特殊报文,称作转发表设置消息.*R*1 需要发送转发表设置消息给那些 MBGP 路由器,使它们将 *R*1 需要使用的多条路径安装到转发表中.例如,当 *R*5 收到从 *R*1 发送的转发表设置消息后,它就会将 *R*2 和 *r*5 都设置成到达主机 *D* 的下一跳.MBGP 不修改报文头,数据流量只是在多条路径进行简单的分发.

MBGP 的特点是由边缘路由器发起多路径发现,并支持增量部署.多路径的发现过程将不可避免地产生路径使用延迟,MBGP 在该过程进行时使用 BGP 路径传输数据,从而避免了多路径建立延迟对数据传输的影响.通过设计一种基于源的路径发现机制,MBGP 有效避免了可能的环路问题.虽然与现有 BGP 兼容,但 MBGP 的设计相对比较复杂,同时,运行 MBGP 也会引入额外的计算开销.另外,MBGP 的作者提出通过设定可发现路径数量的最大值来控制消息开销,但未给出路径数量和消息开销的实际量化关系,因此难以具体操作.

3.3 路由偏转(route deflection)

路由偏转[18]的主要思想是,网络中的路由器通过路由偏转规则,将报文在不引起路由环路的情况下,发往非最短路径的路径上.为实现端用户具有路径选择的功能,文献[18]的作者提出了一种标签体系结构,端用户在报文中设置不同的标签,使途经的路由器选择不同的路径进行报文分组转发.

为实现域间多路径路由,路由偏转的主要思想是,使报文在一个 AS 内选择不同的出口路由器.在一个 AS 内,报文分组通过不同的出口路由器转发,就可以开发域间路由的差异性.也就是说,报文分组由域内路由器按照路由偏转规则进行偏转,就可以实现域间多路径.3 条偏转规则能够保证在每个路由器上都能独立进行报文分组转发,而不出现路由环路问题.

这 3 条偏转规则是:

规则 1. 一个节点 n_i 的偏转集合中的节点 n_{i+1} 应该满足条件 $cost(n_{i+1})<cost(n_i)$.

规则 2. 一个节点 n_i 的偏转集合中的节点 n_{i+1} 应该满足下列条件之一:$cost(n_{i+1})<cost(n_i)$或者 $cost(n_{i+1})<cost(n_{i-1})$.

规则 3. 一个节点 n_i 的偏转集合中的节点 n_{i+1} 应该满足下列条件之一:$cost(G/l_{i+1},n_{i+1})<cost(G/l_i,n_i)$或者 $cost(G/l_{i+1},n_{i+1})<cost(G,n_{i-1})$.

其中,n_i 表示当前需要转发报文分组的路由器,n_{i+1} 表示当前路由器可以安全进行转发的下一跳路由器集合,$cost(n_i)$($cost(G,n_i)$)表示当前路由器 n_i(在网络拓扑 G 中)与目的地址的最短距离,G 表示整个网络的拓扑,G/l 表示除去链路 l 后的整个网络拓扑.

这 3 条路由规则不是并列的,而是逐渐加强的.第 1 条规则提供的可偏转邻居集合最小,但是最容易实现,计算开销最小;第 3 条规则提供的可偏转邻居集合最大,但是实现难度最大,涉及的计算开销也非常大;第 2 条规则的实现难度和计算开销介于上述两者之间.

标签体系结构使得端用户可以选择不同的路径进行报文分组转发.路由器通过提取报文中的标签信息进行路径映射,从而将报文发往不同的路径上.需要指出的是,标签信息不具有全局意义,不同的路由器根据自身的配置,同一标签信息在不同的路由器中具有不同的路径映射.

路由偏转的一个显著优点是便于增量部署.部署路由偏转的路由器可以通过提取标签信息进行报文转发,而其他路由器则将携带标签信息的报文发往主路由上.路由偏转的端用户不了解路径信息,并不知道报文分组是经过哪些 AS(路由器)进行传输的.

3.4 MIRO

MIRO[19]的主要思想是,当一个 AS(被称作请求 AS)需要特定路由时,就发送路由请求给邻居 AS(被称作应答 AS)进行索取.MIRO 的多路径发现是通过主动请求模式,而不是被动接受模式来实现的.如图 5 所示,假设 *A* 是源 AS,*F* 是目的 AS,节点旁罗列的是节点通过 BGP 所得的到达目的 AS 的路由,AS *A* 到 AS *F* 的主路径由粗线显示.这里,我们假设 AS *A* 由于某种理由不希望它的数据报文通过 AS *D*.显然,当前 AS *A* 所具有的路由不能满足这个要求,而其邻居 AS *C* 却有路由 *C-E-F* 满足要求.在 MIRO 中,AS *A* 可以向 AS *C* 提出请求:请求不包括

AS *D* 的路由.AS *C* 根据请求将满足条件的路由发送给 AS *A*.

MIRO 采用隧道机制区分那些需要在备用路径上传输的报文,即从隧道接收到的数据都发往备用路径.在具体实现上,MIRO 需要修改路由器的转发平面和控制平面,控制平面主要负责两个路由器之间的路径协商和隧道建立.如图 5 所示,AS *A* 需要和 AS *C* 进行协商建立隧道,AS *C* 会通告 AS *A* 路由 *A-C-E-F* 的隧道标识,AS *A* 希望通过路由 *A-C-E-F* 发送的报文的报文头中能够写入隧道标识,AS *C* 根据报文头的隧道标识则可知该报文是否通过主路由发送.一般情况下,一个 AS 可能会与多个 AS 建立隧道,因此,MIRO 建立一个类似于基于目的地转发表的基于隧道标识的转发表.此外,MIRO 为请求 AS 和应答 AS 分别提供了相应的策略,以保证备用路由遵循服务提供商的策略配置.

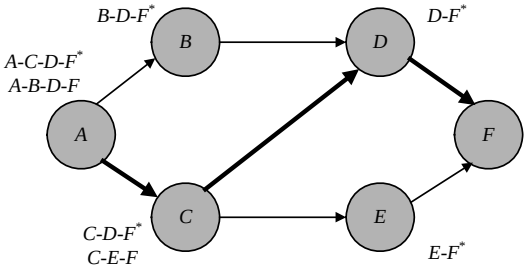

Fig.5 An example of MIRO

图 5 MIRO 示例

MIRO 的优点是:

(1) 消息开销低,互联网上大部分 AS 对于 BGP 提供的主路由是比较满意的,因此,只有当 AS 需要额外路由时才产生额外的消息开销;

(2) 途经的 AS 可以对备用路径进行策略设置,从而意味着端用户使用的备用路径不会违反途经 AS 的策略配置;

(3) 能够开发路径多样性的优势,因为它允许跨 AS 进行协商,如果邻居 AS 没有满足请求 AS 条件的路径,则请求 AS 可以向邻居的邻居进行请求.以此类推,直到有 AS 具有满足条件的路径.然而,MIRO 中多路径的建立需要引入延迟,难以满足需要故障快速恢复的场合.

3.5 域间拼接(inter-domain splicing)

域间拼接是路径拼接(path splicing)[20]在域间路由中的应用.路径拼接的主要思想是,在一个网络拓扑上,根据预设的不同链路权值,路由器可以计算出不同的报文转发树;端用户可以通过对报文头的特定设置,允许途经路由器将报文在不同的报文转发树中进行发送.部署域间拼接的路由器在控制平面选出 k 条最优路由并将其装入转发表;在数据平面,根据报文头中包含的拼接位(splicing bit)使用相应路径.域间拼接所基于的假设是:针对每个目的地址,BGP 路由器已经具有到达目的地址的多条路由.并且文献[20]的作者认为,利用这些路由进行域间多路径路由已经足够,从而无需增加 BGP 的路由消息开销和改变 BGP 更新报文格式.报文中的拼接位既可以由端用户设置(实现端用户对路径的控制),也可以由路由器设置(实现本地故障的快速恢复).

数据发送者将域间拼接位写入报文头中,路由器根据拼接位选取下一个邻居 AS.报文头中增设策略位,保证报文的“无谷底(valley free)”[21]路由.

域间拼接机制的最大缺陷是不能有效地避免转发环路.另外,随着域间路由策略的复杂化以及 AS 之间关系的多样化,路径拼接中只使用一位来区别 AS 间的关系显然是不够的.

3.6 L-BGP

Beijnum 等人[22]认为,严格遵守 Cisco 的多路径选择方法虽然可以避免路由环路的发生,但却极大地限制了路径多样性.L-BGP 的主要思想是,通过放宽 Cisco 提出的 BGP 多路径中备用路径和主路径必须“一致”的条件,

从而使各 AS 可以使用更多的备用路径.同时,L-BGP 使本地路由器通告一条特定的具备较长 AS 路径长度的路径,并安装所有 AS 路径长度比通告路径长度短的路径.L-BGP 在只允许 AS 通告一条路径的前提下使用多条路径,如何避免路由环路是 L-BGP 需要解决的首要问题.L-BGP 的提出者指出,只要满足根据 BGP 修改的 LFI 条件[23]就可以避免路由环路的发生.

L-BGP 根据 BGP 语义将 LFI 条件修改如下:

$$cp(p_r)<cp_r(p_r),$$

$$P=\{p|cp(p)\leqslant cp(p_r)\wedge p\in\pi\},$$

其中,p_r 是路由器广播给邻居的路径,$cp(p_r)$是 p_r 在本地路由器到目的地的开销(对于 eBGP 而言,这个值是 AS 路径长度;对于 iBGP 而言,这个值是由域内协议通告给 BGP 的内部开销),$cp_r(p_r)$是本地路由器通告给邻居的 p_r 的开销,P 是路由器考虑使用的路径集合,π表示邻居通告给本地路由器的路径集合.

在多路径发现上,L-BGP 利用了 BGP 所发掘的路径信息;在路径选择上,L-BGP 采用安装所有 AS 路径长度比通告路径短的路径.L-BGP 虽然从理论上保证了报文转发的无环路特性,但对采用通告非最佳路由而有可能引起的其他问题未作讨论,比如控制平面和数据平面的一致性问题.

3.7 小 结

本节主要介绍了单径通告多路转发协议,表 1 列出了此类协议的特点.在路径多样性中,Low 表示利用的仅为 BGP 协议提供的路径多样性,High 表示能够发掘链路层所有的路径多样性,Medium 表示能够发掘一定程度的路径多样性.在控制平面开销上,Low 表示协议产生的消息开销与 BGP 产生的消息开销无异,Medium 表示协议产生的开销比 BGP 产生的消息要多,High 表示协议产生的开销比 Medium 的程度要深.Yes 表示协议具有该特性,No 表示协议不具有该特性,“–”表示该协议对此指标未作考虑.

Table 1 Comparison of protocols on single announcement and multipath forwarding

表 1 单径通告多路转发协议的特点比较

协议名称	路径多样性	控制平面开销		数据平面开销		无环路特性	可扩展性	端用户路由控制
		路由消息开销	路由平面开销	报文携带信息	转发表信息			
BGP 多路径	Low	Low	Low	No	Multiple next-hop	Yes	Yes	No
MBGP	High	High	High	No	Multiple next-hop	Yes	No	No
路由偏转	Low	Low	Medium	Tag	Multiple next-hop	Yes	Yes	Yes
MIRO	High	Medium	High	Tunnel ID	Tunnel ID based forwarding table	Yes	Yes	Yes
域间拼接	Low	Low	Low	Splicing bits	Multiple forwarding table	No	Yes	Yes
L-BGP	Low	Low	Medium	None	Multiple next-hop	Yes	Yes	No

单径通告协议由于采用单 BGP 路径通告,因此路由开销相对于多径通告协议要低.另外,此类协议基于 BGP,因此易于增量部署.但是,此类协议的一个共同问题是,如何在仅由单路径通告信息的基础上开发路径的多样性.MBGP 和 MIRO 是通过主动多路径发现的手段,该方式基于大部分网络用户满足于现有 BGP 提供的单路径路由的前提来保证较低的路由开销,同时引入了多路径建立的延迟,不利于需要立即使用备用路径的场景.除 MBGP 和 MIRO 以外,其他 4 个协议都基于现有 BGP 提供的路径多样性,因此路径多样性相对较低.

4 多径通告多路转发协议

本节主要介绍多径通告多路转发路由协议,其特点是每个 AS 向其邻居通告多条 AS 路径.多径通告可以明显增加 AS 的路径多样性,如图 6 所示,假设 AS 5 是目标 AS,在单径通告中,AS 1 只知道 1 条 AS 路径,因为 AS 2 只允许通告 1 条路径给 AS 1.在多径通告的情况下,AS 2 可以通告多条路径给 AS 1,因此 AS 1 可以掌握多条 AS 路径信息.显然,多径通告增加了路由消息开销.因此,多径通告多路转发协议需要尽量控制每个 AS 通告给邻居的路径数目.

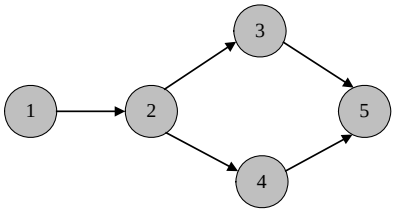

Fig.6 A simple topology

图 6 一个简单拓扑

4.1 带ADD_PATH功能的BGP

带 ADD_PATH 功能的 BGP[24]的主要思想是,利用 RFC 3392 提出的“功能通告”定义“ADD_PATH”功能,当通信双方都支持 ADD_PATH 功能时,相互间可以进行多路径通告.“路径标识符”用来区分针对同一个网络前缀的多条路径,“路径标识符”由本地路由器指定并通告给邻居.为了支持上述操作,带 ADD_PATH 功能的 BGP 对 NLRI 编码方式进行了扩展,并给出了 ADD_PATH 功能的具体格式.带 ADD_PATH 功能的 BGP 从实现层面提出了一种解决方案,但未考虑域间多路径路由协议设计的许多其他因素,比如控制平面开销、环路特性等.

4.2 R-BGP

为了减少控制平面开销尤其是路由消息开销,R-BGP[11]只通告 1 条备用路径给指定的邻居节点.R-BGP 提出了 3 种备用路径选择方案:第 1 种是与主路径交叉节点差异最大并忽略路由策略的路径,第 2 种是与主路径交叉节点差异最大但考虑路由策略的路径,第 3 种是由 BGP 路径选择过程得到的次优路径.方案 1 虽然提供了最大程度的可靠性,但其忽略了路由策略,因而在商业互联网上难以实施;方案 2 是一个比较好的折衷,放弃一部分路径多样性但遵循路由策略,更容易被互联网服务提供商所采用;方案 3 最易实现,但不能提供足够强的路径多样性.备用路径只通告给在主路径上的下一跳 AS(称为“下一跳机制”),因为每个 AS 都是自私的,他们首先得保证自己的报文不丢失,所以通常愿意给主路径的下一跳 AS 提供备用路径.R-BGP 在消息报文中携带中断链路标识信息(root cause information),避免了在收敛过程中出现的转发环路,并加速了协议收敛过程.

由于备用路径通告目标的选择性(主路径的下一个 AS),有些 AS 可能就不具备备用路径.R-BGP 的提出者指出“使用原主路径机制”,即链路中断发生后,只有中断链路的上游邻接 AS 将后续报文通过备用路径发送,其他 AS 仍然使用他们的原有主路径.同时,提出“确保收敛机制”来控制 AS 使用备用路径的时间,即当一个 AS 从所有邻居中收到撤销消息时,这个 AS 停止转发内部生成的流量到备用路径上;一个 AS 延迟发送路径撤销消息给邻居,直到它确定在协议收敛时它不能提供这个邻居一条“无谷底”路径.通过上述一系列的机制——“下一跳机制”、“使用原主路径机制”和“确保收敛机制”,R-BGP 被理论上证明了能够确保在任何情况下(链路中断、策略改变等),只要两个 AS 间存在一条“无谷底”的与路由策略兼容的路径,那么他们之间将不会出现通信中断的情况.

R-BGP 对于区分备用路径和主路径上的报文给出多种解决方案,分别是“虚拟”链接、MPLS 或者 IP 隧道.R-BGP 主要针对提高域间路由的可靠性,并不支持用户对路由的选择.

4.3 YAMR

YAMR[25]选择通告一定数目的路径,以达到足够程度的路径多样性.除选取主路径以外,YAMR 针对主路径上每条链路选取备用路径,也就是说,使选取的每条路径可以避免使用主路径中的 1 条链路.YAMR 路径选择的主要目的是,保证在主路径上 1 条链路发生故障后,AS 总存在 1 条可用(如果存在)路径.通告多条备用路径会大大增加 YAMR 的控制平面和数据平面开销,YAMR 提出了“路由信息隐藏”技术,即当链路故障发生时,如果知晓该故障发生的 AS 拥有可以绕开该故障的可用路径,那么它将不向网络中其他节点通告此故障的发生,从而大大降低了通告多路径而引起的消息开销和协议收敛时间.

YAMR 的每条路径都通过标签加以区分:主路径用统一的标签标识,比如[0,0];而备用路径用其所规避的链

路进行标识(此链路链接的两个 AS 号).报文转发基于目的地地址和路由标签来进行,YAMR 设计了环路检测令牌和断连检测令牌以保证“路由信息隐藏”技术不引起路由环路以及 AS 断连问题.另外,YAMR 使用故障根源信息(root cause notification,简称 RCN)广播故障链路信息,用来提高网络协议的收敛速度.

通过图 7 我们来简述 YAMR 的运行流程:假设 *A* 是目的 AS,当 YAMR 协议收敛后,各 AS 的路径如图 7 中节点旁标注所示.每个 AS 的主路径标签是[0,0],*E* 的主路径是 *E-C-A*.根据 YAMR 的路径建立规则,*E* 针对主路径的每条链路选取备用路径并以此链路为标签,即以[*E*,*C*]为标签的路径 *E-D-A* 和以[*C*,*A*]为标签的路径 *E-C-B-A*.在协议运行过程中,若出现链路 *A-C* 中断,*E* 就可以使用路径 *E-C-B-A*;若出现链路 *C-E* 中断,*E* 就可以使用路径 *E-D-A*.若链路 *A-C* 中断,由于 *C* 仍然可以通过路径 *C-B-A* 到达 *A*,则根据“路由信息隐藏”,*C* 可以不通告链路 *C-A* 中断的消息给 *E*.因此,*E* 继续将到达 *A* 的报文发送给 *C*,而 *C* 则将报文发送给 *B* 并到达 *A*.

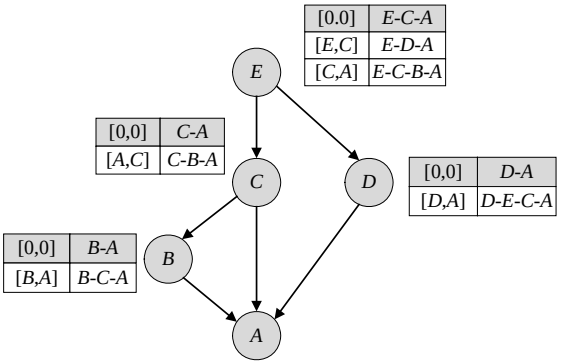

Fig.7 An example of YAMR

图 7 YAMR 示例

在多路径建立期间,由多路径通告引起的高路由开销(至少 3 倍以上)是 YAMR 实现真正部署的主要障碍.另外,YAMR 虽然通过模拟实验论证了其方法的有效性和正确性,但其主要实验采用的网络拓扑规模仍远小于现实的互联网规模,因此难以确定 YAMR 在互联网规模下的正确性和有效性.

4.4 PDAR

PDAR[26]允许 AS 在通告主路径的同时再通告 1 条与主路径节点(或者边)具有极小相似度的备用路径,PDAR 采用选择性备用路由通告策略有效降低了路由消息开销.选择性备用路由通告策略主要包括两项:

(1) 如果邻居已经拥有差异度很大的路由,则无需再向其发送备用路由,只需发送主路由;

(2) 向邻居发送的备用路由必须满足能够要增大邻居路由差异程度的条件.

PDAR 使用故障根源信息广播故障链路信息,用来提高网络协议的收敛速度.PDAR 的提出者认为,与多路径路由技术结合的故障根源信息广播可以进一步加快收敛速度,提高网络健壮性.基于极大差异路径通告、选择性宣告备用路由策略和 RCN 技术的 PDAR 协议称为 D-BGP.

但故障根源信息的广播会泄露 AS 内部私有信息,如策略配置、链路设置等.同时,在使用选择性备用路由通告策略后的消息开销仍比较大.为了解决上述两个问题,PDAR 的提出者指出采用 Bloom Filter,称为 B-BGP.将故障链路信息和路径信息通过 Bloom Filter 进行编码,保证了一些敏感信息不至于泄露,并有效降低了消息开销.

若发生链路故障,则在一个 AS 开始使用其备用路径后,PDAR 为避免转发环路,会采用 IP 封装或者 MPLS 技术进行备用路径报文的转发.

4.5 小 结

本节主要综述了多径通告多路转发协议,表 2 列出了此类协议的特点比较(相关说明与表 1 相同).显然,通告

所有已知路径是让边缘 AS 掌握所有路径的最简单方式,然而必定引起控制平面和数据平面开销的指数级膨胀,从而使得协议不具有可扩展性.因此,多径通告多路转发协议选择通告一部分路径,比如,R-BGP 和 PDAR 只通告 1 条备用路径,而 YAMR 通告与主路径链路数相同的备用路径.另外,此类协议与 BGP 兼容,具有增量部署的特性.带有 ADD_PATH 功能的 BGP 和 YAMR 都设计了路径标识,用于区分到达同一目的前缀的多条路径.R-BGP 和 PDAR 则采用 MPLS 或者 IP 隧道来进行备用路径的数据转发.我们认为,路径标识的引入虽然需要改变现有的基于地址的转发表结构,但却是将来多路径路由设计必须采用的方法.特别需要指出的是,YAMR 的“路由信息隐藏”技术对于减少路由消息开销和加快协议收敛时间可以起到非常重要的作用.此外,路由消息报文携带中断链路标识信息的方式可以帮助域间多路径路由协议加速收敛,因此是未来协议设计的不可缺少的部分.

Table 2 Comparison of protocols on multiple announcements and multipath forwarding

表 2 多径通告多路转发协议的特点比较

协议名称	路径多样性	控制平面开销		数据平面开销		无环路特性	可扩展性	端用户路由控制
		路由消息开销	路由平面开销	报文携带信息	转发表信息			
BGP with ADD-path capability	High	High	–	Local path ID	Destination and path ID based forwarding	–	–	–
R-BGP	Medium	Low	High	No	依赖于具体转发技术	No	Yes	No
YAMR	High	High*	High	Path ID	Destination and path ID based forwarding	No	Yes	Yes
PDAR	Medium	Low	High	No	依赖于具体转发技术	No	Yes	No

*YAMR 在多路径创建过程中路由消息开销较大,在网络故障发生后,由于采用了“路由信息隐藏”技术,路由消息开销降低.

5 新型域间多路径路由体系结构

本节主要介绍具有新型域间多路径路由体系结构特征的路由协议,主要是 NIRA 和路段路由(pathlet routing).与前两类协议不同,此类协议是域间路由系统的长期解决方案,长期方案通常不考虑兼容性而是构建一个全新的域间多路径路由体系结构.出于完整性考虑,我们将基于反馈的路由(feedback based routing)和 BANANAS 也归于此类,虽然它们并没有提出完整的路由体系结构,但是包含了一些未来域间多路径路由设计可能需要考虑的因素.显然,长期方案在短时间内实现是不可能的,但却能够给设计短期方案带来一定的启示.

5.1 NIRA

NIRA[27]提出了一种新的路由体系结构并支持用户选择路由,NIRA 提出拓扑信息传播协议(topology information propagation protocol,简称 TIPP).TIPP 协议的主要功能是使端用户发现可用路由.NIRA 中的一个重要概念是 up-graph,up-graph 是由一个端用户的服务提供者及其提供者的提供者组成的网络.通过 TIPP 协议中的路径向量协议,一个端用户容易构建自己的 up-graph.TIPP 协议中的链路状态协议在进行链路消息通告时主要就是在某个 up-graph 中进行,从而增加了协议的可扩展性.NIRA 定义由顶级 AS(tier-1 AS)组成的网络称为互联网核(core).NIRA 只告知端用户一部分由端到互联网核的路由,因此,端到端的路由是通过路由拼接完成的,即由发送者到核的路由和核到接收者的路由组成.这里,发送者需要通过名字路由查找服务(name-to-route lookup service,简称 NRLS)来获取核到接收者的路由.在 NIRA 中,顶级 AS(tier-1 AS)将分配所得网络地址进行细分授权给下一级 AS,然后这一级 AS 将所分配得到的网络地址同样进行细分授权给下一级 AS.依此类推,直到所有的末端 AS 都有网络地址.基于上层服务提供商的层次化地址分配,NIRA 允许端用户可以拥有多个网络地址.层次化地址分配的直接好处是,端用户的网络地址可以用来进行路由编码.NIRA 中的多路径主要是通过源地址和目的地址的不同组装实现的,在 NIRA 中,每个路由器都有 3 个路由表,分别是 uphill 路由表、downhill 路由表和 bridge 路由表.在转发过程中,路由器按照一定的顺序进行路由表查找.路由器首先根据目的地址在 downhill 表中查找;若未找到,就根据源地址在 uphill 表中查找;若仍未找到,就在 bridge 路由表中进行查找.若当端用户的 up-graph 中链路出现故障时,TIPP 协议将故障信息通告给端用户,端用户可以重新组装可用路由.若当

目的用户的 up-graph 中链路出现故障时,由于 TIPP 协议只在 up-graph 中传播故障,因此发送者需要采用比如超时等手段来判断路径故障.

下面,我们根据图 8 来了解 NIRA 的运行过程.图 8 中,*A*1,*A*2,*A*3 和 *A*4 是处于互联网核心的路由器,分别有统一分配的地址 1::/16,2::/16,3::/16 和 4::/16.对于对等关系的 AS,NIRA 提出给他们单独分配一类地址,如 *R*2 和 *R*3 分别有对等地址 FFFF:1::/32 和 FFFF:2::/32.此时,*R*2 具有两个地址.然后,他们分别向自己的客户进行地址分配,比如,*R*1 和 *R*2 分别获得由 *A*3 分配的地址 3:1::/32 和 3:2::/32,*R*1 和 *R*2 再向其客户(如 *N*1,*N*2)继续分配地址.客户 Bob 获得由 *N*1 分配的 3 个地址.同理,客户 Alice 获得由 *N*3 分配的 3 个地址.这种分配的地址实际上包含了路由信息.比如,Bob 的地址 3:2:1:1000 包含了路由 *N*1-*R*2-*A*3.当 Bob 需要和 Alice 进行通信时,Bob 首先通过名字路由查找服务获得从核到 Alice 的 3 条路由,而自己具有 3 条到达核的路由.通过路径拼接,Bob 可以获得 5 条到达 Alice 的路由路径.Bob 将选择的路由中对应的源地址和目标地址写入报文头,途经的路由器通过相应的路由表查找顺序进行查找和报文转发.

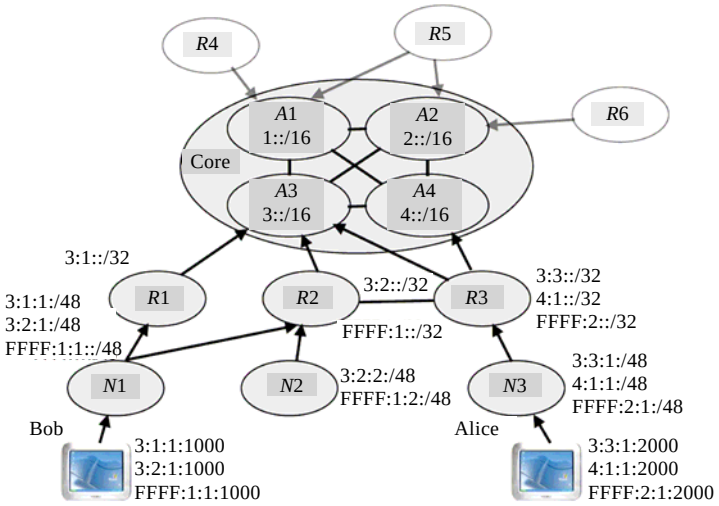

Fig.8 An example of NIRA

图 8 NIRA 的具体实例

5.2 路段路由(pathlet routing)

路段路由[28]的核心概念之一是路段(pathlet),路段是指路由的一部分.与传统的 BGP 通告整条路由不同,路段路由只通告路段.路段路由的主要思想是,多个路段可以进行拼接,组成有效的整条路由.路段路由的另一个核心概念是虚拟节点(vnode),虚拟节点的主要作用是作为路段的组成部分,也就是说,路段是由一系列虚拟节点组成的.从上层来看,路段路由可以简单看成是在一个以虚拟节点为节点和以路段为链路组成的网络上的路由.路段通告(类似于 BGP 中的路由通告)采用的是路径向量协议,路由计算则类似于链路状态协议.路径向量协议只用于进行路段信息的传播,而不对路段进行选择(与 NIRA 相似).每条路段都携带额外信息,比如路段的转发标识符和它所包含的虚拟节点序列,路段的转发标识符用于路段的查找.边缘路由器从控制平面获得路段信息,将这些信息构成一个由路段和虚拟节点组成的网络拓扑,并在此拓扑上进行最短路由算法以获得路由.发送者在发送数据时,需要将所途经路由的路段的转发标识符放入报文头中.路由器根据报文中的转发标识符进行转发表查找,若查找到对应标识符,则进行相应处理,包括转发标识符重写以及将报文发至下一跳.

策略表达能力是域间多路径协议的本质属性之一,路段路由的提出者指出,策略表达能力可以作为横向比较多种路由协议的指标,并揭示多种协议之间的内部联系.首先,我们介绍与讨论策略表达能力相关的术语和定义.

我们定义协议配置为网络中每个路由器在一个路由协议下运行的转发表状态集合.

在同一网络拓扑下,给定协议 P 的配置 c_1 和协议 Q 的配置 c_2,c_1 蕴含 c_2 需满足下列条件:

a) c_1 的可达(不可达)路由在 c_2 也可达(不可达);

b) 对于网络的每一个路由器 i,$|c_1(i)|=O(|c_2(i)|)$,其中,$c_j(i)$表示在协议配置 j 中路由器 i 的转发状态数目.

我们定义协议 P 的策略表达能力强于 Q,当对于 Q 中的任何配置 c_2,P 中都存在一个配置 c_1,使得 c_1 蕴含 c_2.

如图 9 所示,同一方框内的协议表示策略表达能力基本相同,$P\rightarrow Q$ 表示 P 策略表达能力强于 Q.从图 9 中可以看出,路段路由的策略表达能力强于大部分多路径路由协议.后面将介绍的基于反馈的路由不仅完全独立于 BGP,并且自身包含链路的过滤机制,因此不在我们的策略表达比较范围内,这里不再赘述.

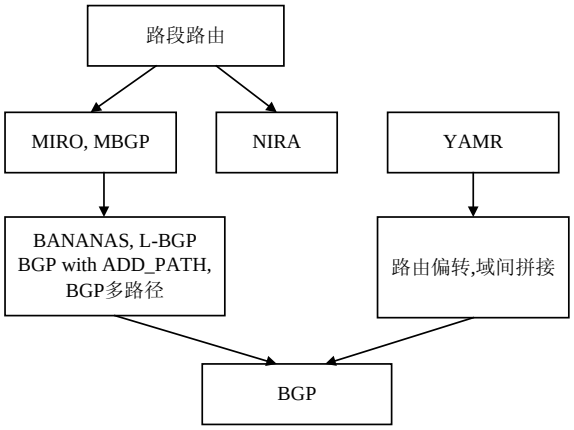

Fig.9 Comparison of policy expressiveness

图 9 路由协议的策略表达能力比较

5.3 基于反馈的路由(feedback based routing,简称FBR)

FBR[29]的核心思想是,在 AS 级作源路由,FBR 将域间路由器分为边缘路由器和过渡路由器,将路由信息分为结构路由信息和动态路由信息.结构路由信息是指网络的链路存在信息,而动态路由信息是指链路状态变化的信息.在 FBR 中,过渡路由器负责结构路由信息的传播,边缘路由器负责动态路由信息的检测.发送者在报文头中写入报文所途经的自治系统,过渡路由器只需根据报文中的路径信息就可以进行报文转发而无需计算路由,计算路由的工作由边缘路由器完成.通过过渡路由器传播的结构路由信息,边界路由器可以得到全网的网络拓扑结构,并由此进行路由计算.FBR 中,每个边缘路由器针对每个网络(前缀)计算两条路由:一条称为主路由,另一条称为备份路由.当主路由出现故障时,报文就通过备份路由进行发送.边缘路由器周期性地对路由进行实时测量,一旦发现路由出现问题,就将问题路由所包含的链路排除并重新进行路由计算.通过只传播结构路由信息和源路由机制,FBR 的路由系统具有很高的可扩展性,同时降低了过渡路由器的复杂性.

下面,我们根据图 10 来介绍 FBR 的运行过程.假设 A,G 是边缘路由器,并且 A 是发送者,G 是接收者.当 A 得到网络拓扑信息后进行路由计算,获得两条路由,分别是 A-C-F-G 和 A-B-D-E-G.A 通过测量获得两条路由的 RTT 分别是 20ms 和 200ms,并将 20ms 的路由设为主路由.当 A 发送报文时,报文头写入主路由信息 A-C-F-G,途经的路由器根据报文头中的路由信息进行转发.同时,A 持续地对两条路由进行检测,若发现问题,比如主路由传输延迟突然增大,则使用备用路由进行传输.

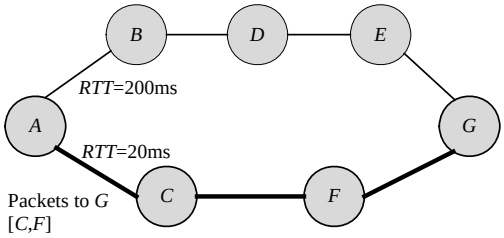

Fig.10 An example of FBR

图 10 FBR 示例

5.4 BANANAS

BANANAS[30]提出一种显式多路径路由方式,我们着重讨论其域间路由多路径机制.在 BANANAS 中,每条 AS 路由都被哈希成一个全局可认的标识,称为 e-PathID.为了使每个 AS 能够获得到达每一个目的地的多条路由, BANANAS 要求 AS 向其邻居广播多条路由信息,并且发送者需要在报文头中写入 e-PathID.BANANAS 提出两个核心功能:显式出口转发和显式 AS 路由转发.显式出口转发是指,在域内,为了增加路由多样性,路由器可以将到达同一个网络的报文分发到不同的出口路由器上;显式 AS 路由转发要求在支持 BANANAS 的路由器转发表中的每一个表项包含 4 部分[目的地址,入口 e-PathID,出口接口号,出口 e-PathID].当一个 BANANAS 路由器收到报文后,首先根据报文的目的地址和 e-PathID 查找转发表,若找到相应表项,则将报文中的 e-PathID 域改写成出口 e-PathID,然后发送到对应的出口接口上;否则,报文将被丢弃.

5.5 小 结

本节主要介绍了新型域间多路径路由体系结构协议,其特点比较见表 3.

Table 3 Comparison of new Internet routing architecture based protocols

表 3 新型域间多路径路由体系结构协议

协议名称	路径多样性	控制平面开销		数据平面开销		无环路特性	可扩展性	端用户路由控制
		路由消息开销	路由平面开销	报文携带信息	转发表信息			
NIRA	High	Medium	Low	No	Multiple forwarding table	Yes	Yes	Yes
路段路由	Medium	High	High	Forwarding ID list	Forwarding ID based	Yes	Yes	Yes
基于反馈的路由	High	Low*	Low*	AS path	Source routing	Yes	No	Yes
BANANAS	Medium	Medium	High	e-PathID	Destination and e-PathID based forwarding	Yes	Yes	No

*FBR 中路由的计算和存储主要由边缘路由器完成,这里比较的是核心路由器的控制平面开销.

从路径数量上来看,FBR 只给边缘路由器提供了针对每个网络前缀的两条路径,而 NIRA、路段路由和 BANANAS 则提供多条路由,其中,BANANAS 提供的路由数目基于 AS 间通告路由的数目.通观新型域间多路径路由体系结构协议我们可以发现,路由技术组合成为设计未来路由协议的趋势,如图 11 所示.具体地,NIRA 和路段路由都采用了链路状态协议、路径向量协议和源路由作为内部运行机制,分别负责不同的功能.FBR 是基于链路状态协议计算路由,而基于源路由进行报文转发.BANANAS 则是基于路径向量协议计算路由,而采用与源路由类似的机制进行报文转发.

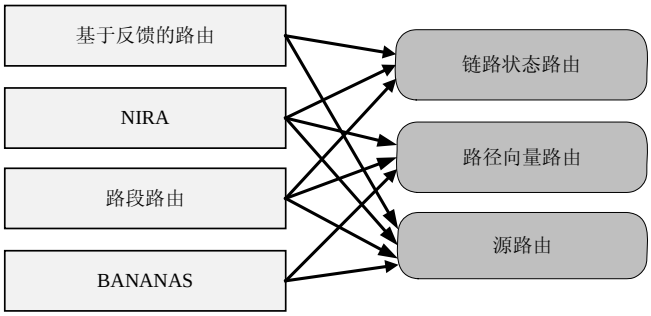

Fig.11 Combination of routing techniques

图 11 路由技术组合

6 结论和未来展望

当前广泛使用的 BGP 协议存在如可靠性差、次优路由使用、对负载均衡支持较差等问题,域间多路径路由被认为是解决这些问题的有效方法之一.本文集中阐述和分析了主要域间多路径路由协议,并将这些协议分为 3 类:单径通告多路转发协议、多径通告多路转发协议和新型域间多路径路由体系结构.我们认为,未来研究中需要进一步解决的问题主要包括:

(1) 路由稳定性问题.互联网上的每个 AS 根据自己的需求独立配置路由策略.研究者发现,独立的路由策略配置和 AS 的自私性会导致持久的路由震荡,从而导致报文丢失或者出现路由环路[31].多路径功能在域间多路径的扩展是否可能加剧(或者减轻)路由振荡,是值得研究的问题.研究表明[32],通过 AS 通告多条路由,可以减轻甚至避免 BGP 中的路由振荡问题.然而,域间多路径路由是否本身存在路由振荡也是值得研究的问题.

(2) 域间安全性问题.域间多路径路由作为域间路由未来的发展方向,必须能够有效解决近年来日益突出的 BGP 安全问题[33].一方面,目前的研究工作没有综合考虑多路径特性与安全问题的结合,比如利用多路径路由的特性进行前缀劫持的检测;另一方面,有些域间多路径协议本身的设计就存在安全隐患,比如在 MIRO 中,恶意 AS 可以充当具有满足条件路由的应答 AS,从而达到监控流量的目的.基于这两个方面,未来可能的工作是:

(a) 用域间多路径路由的自身特点,比如多路径通告、多路径转发等,检测并防范前缀劫持等网络攻击行为;

(b) 设计一种安全的域间路由协议,协议机制本身能够对一些网络路由攻击进行检测和防御.

(3) 全网多路径路由.互联网路由体系结构是层次性的,分为域间路由和域内路由,因此,互联网多路径路由可分为域间多路径路由和域内多路径路由.目前,域内多路径路由的研究相对比较成熟,域间多路径路由协议的研究也已经充分展开.然而,目前仍没有关于如何将两者结合起来实现整个互联网多路径路由的更细致的研究.实际上,路径拼接已经尝试综合考虑域内和域间多路径路由,但没有作为其研究的重点.未来的域间多路径路由设计必须考虑其与域内多路径的交互,唯此才能最终实现高效的全网多路径路由.

(4) 基于路由代数的域间多路径路由的建模分析.近年来,应用路由代数验证网络协议的可收敛性是研究的热点[34,35].如何使用路由代数对域间多路径路由进行建模来分析研究域间多路径路由的收敛性,具有非常重要的意义.

References:

[1] Rekhter Y, Li T, Hares S. A border gateway protocol 4 (BGP-4). RFC 4271, 2006.

[2] Labovitz C, Ahuja A, Bose A, Jahanian F. Delayed Internet routing convergence. ACM SIGCOMM Computer Communication Review, 2000,30(4):175−187. [doi: 10.1145/347057.347428]

[3] Rexford J, Wang J, Xiao Z, Zhang Y. BGP routing stability of popular destinations. In: Proc. of the 2nd ACM SIGCOMM Workshop on Internet Measurement (IMW 2002). New York: ACM Press, 2002. 197−202. [doi: 10.1145/637201.637232]

[4] Kushman N, Kandula S, Katabi D. Can you hear me now?! It must be BGP. ACM SIGCOMM Computer Communication Review, 2007,37(2):75−84. [doi: 10.1145/1232919.1232927]

[5] Moy J. RFC2328: OSPF Version 2. 1998.

[6] Callon R. RFC1195: Use of OSI IS-IS for routing in TCP/IP and dual environments. 1990.

[7] Wang F, Mao ZM, Wang J, Gao LX, Bush R. A measurement study on the impact of routing events on end-to-end Internet path performance. ACM SIGCOMM Computer Communication Review, 2006,36(4):375−386. [doi: 10.1145/1159913.1159956]

[8] Bremler-Barr A, Afek Y, Schwarz S. Improved BGP convergence via ghost flushing. In: Proc. of the 22nd IEEE INFOCOM Annual Joint Conf. of the IEEE Computer and Communications Societies, Vol.2. 2003. 927−937. [doi: 10.1109/INFCOM.2003.1208930]

[9] Luo JB, Xie JQ, Hao RZ, Li X. An approach to accelerate convergence for path vector protocol. In: Proc. of the IEEE Global Telecommunications Conf., Vol.3. 2002. 2390−2394. [doi: 10.1109/GLOCOM.2002.1189059]

[10] Pei D, Zhao XL, Wang L, Massey D, Mankin A, Su SF, Zhang LX. Improving BGP convergence through consistency assertions. In: Proc. of the 21st IEEE INFOCOM Annual Joint Conf. of the IEEE Computer and Communications Societies, Vol.2. 2002. 902−911. [doi: 10.1109/INFCOM.2002.1019337]

[11] Kushman N, Kandula S, Katabi D, Maggs BM. R-BGP: Staying connected in a connected world. In: Proc. of the 4th USENIX Symp. on Networked Systems Design and Implementation. 2007. 341−354.

[12] He JY, Rexford J. Toward Internet-wide multipath routing. IEEE Network Magazine, 2008,22(2):16−21. [doi: 10.1109/MNET.2008.4476066]

[13] Wendlandt D, Avramopoulos I, Andersen DG, Rexford J. Don't secure routing protocols, secure data delivery. In: Proc. of the ACM HotNets. 2006. 7−12.

[14] Zhang X, Perrig A, Zhang H. Availability-Oriented path selection in multi-path routing. Technical Report, CMU-CyLab-07-012, Carnegie Mellon University, 2007.

[15] Cisco Inc. BGP best path selection algorithm. 2006. http://www.cisco.com/image/gif/paws/13753/25.pdf

[16] JuniperNetworks. Configuring BGP to select multiple BGP pathshttp://www.juniper.net/techpubs/software/junos/junos90/swconfig-routing/configuring-bgp-to-select-multiple-bgp-paths.html#id-13280349

[17] Fujinoki H. Multi-Path BGP (MBGP): A solution for improving network bandwidth utilization and defense against link failures in inter-domain routing. In: Proc. of the IEEE Int'l Conf. on Networks. 2008. 1−6. [doi: 10.1109/ICON.2008.4772612]

[18] Yang XW, Wetherall D. Source selectable path diversity via routing deflections. In: Proc. of the 2006 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM Press, 2006. 159−170. [doi: 10.1145/1159913.1159933]

[19] Xu W, Rexford J. MIRO: Multi-Path interdomain routing. In: Proc. of the 2006 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM Press, 2006. 171−182. [doi: 10.1145/1159913.1159934]

[20] Motiwala M, Elmore M, Feamster N, Vempala S. Path splicing. ACM SIGCOMM Computer Communication Review, 2008,38(4):27−38. [doi: 10.1145/1402958.1402963]

[21] Gao LX, Rexford J. Stable Internet routing without global coordination. IEEE/ACM Trans. on Networking, 2001,9(6):681−692. [doi: 10.1109/90.974523]

[22] van Beijnum I, Crowcroft J, Valera F, Bagnulo M. Loop-Freeness in multipath BGP through propagating the longest path. In: Proc. of the Int'l Workshop on the Network of the Future (Fut-Net 2009). 2009. 1−6. [doi: 10.1109/ICCW.2009.5207968]

[23] Vutukury S, Garcia-Luna-Aceves JJ. A simple approximation to minimum-delay routing. ACM SIGCOMM Computer Communication Review, 1999,29(4):227−238. [doi: 10.1145/316188.316227]

[24] Walton D, Retana A, Chen E, Scudder J. Advertisement of multiple paths in BGP. Internet Draft, 2009.

[25] Ganichev I, Dai B, Godfrey PB, Shenker S. Yamr: Yet another multipath routing protocol. ACM SIGCOMM Computer Communication Review, 2010,40(5):13−19. [doi: 10.1145/1880153.1880156]

[26] Wang F, Gao LX. Path diversity aware interdomain routing. In: Proc. of the INFOCOM. 2009. 307−315. [doi: 10.1109/INFCOM.2009.5061934]

[27] Yang XW. NIRA: A new Internet routing architecture. In: Proc. of the ACM SIGCOMM Workshop on Future Directions in Network Architecture. New York: ACM Press, 2003. 301−312. [doi: 10.1145/944759.944768]

[28] Godfrey PB, Ganichev I, Shenker S, Stoica I. Pathlet routing. ACM SIGCOMM Computer Communication Review, 2009,39(4): 111−122. [doi: 10.1145/1592568.1592583]

[29] Zhu DP, Gritter M, Cheriton DR. Feedback based routing. ACM SIGCOMM Computer Communication Review, 2003,33(1):71−76. [doi: 10.1145/774763.774774]

[30] Kaur HT, Kalyanaraman S, Weiss A, Kanwar S, Gandhi A. BANANAS: An evolutionary framework for explicit and multipath routing in the Internet. In: Proc. of the ACM SIGCOMM Workshop on Future Directions in Network Architecture. New York: ACM Press, 2003. 277−288. [doi: 10.1145/944759.944766]

[31] Griffin TG, Shepherd FB, Wilfong G. The stable paths problem and interdomain routing. IEEE/ACM Trans. on Networking (TON), 2002,10(2):232−243. [doi: 10.1109/90.993304]

[32] Agarwal R, Jalaparti V, Caesar M, Godfrey PB. Guaranteeing BGP stability with a few extra paths. In: Proc. of the 2010 Int'l Conf. on Distributed Computing Systems. 2010. 221−230. [doi: 10.1109/ICDCS.2010.85]

[33] Butler K, Farley TR, McDaniel P, Rexford J. A survey of BGP security issues and solutions. Proc. of the IEEE, 2010,98(1): 100−122.

[34] Sobrinho JL. Network routing with path vector protocols: Theory and applications. In: Proc. of the 2003 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. 2003. 49−60. [doi: 10.1145/863955.863963]

[35] Griffin TG, Sobrinho JL. Metarouting. In: Proc. of the 2005 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. 2005. 1−12. [doi: 10.1145/1080091.1080094]

苏金树(1962－),男,福建莆田人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络,信息安全.

刘宇靖(1985－),女,博士生,主要研究领域为域间路由安全.

戴斌(1982－),男,博士生,主要研究领域为域间多路径路由技术,域间路由安全.

彭伟(1973－),男,博士,副教授,CCF 会员,主要研究领域为域间路由,路由算法.