ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.10, October 2010, pp.2524−2541
doi: 10.3724/SP.J.1001.2010.03911

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

互联网可扩展路由*

唐明董[1,2,3+], 张国清[1], 杨 景[1,4], 张国强[1]

[1](中国科学院 计算技术研究所,北京 100190)

[2](湖南科技大学 知识处理与网络化制造湖南省普通高校重点实验室,湖南 湘潭 411201)

[3](中国科学院 研究生院,北京 100049)

[4](中国移动通信研究院,北京 100053)

Scalable Routing for the Internet

TANG Ming-Dong[1,2,3+], ZHANG Guo-Qing[1], YANG Jing[1,4], ZHANG Guo-Qiang[1]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Laboratory of Knowledge Processing & Networked Manufacturing, Hu'nan University of Science and Technology, Xiangtan 411201, China)

[3](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

[4](China Mobile Research Institute, Beijing 100053, China)

+ Corresponding author: E-mail: mdtang@126.com

Tang MD, Zhang GQ, Yang J, Zhang GQ. Scalable routing for the Internet. *Journal of Software*, 2010,21(10): 2524−2541. http://www.jos.org.cn/1000-9825/3911.htm

Abstract: The Internet routing system is facing a serious scaling challenge due to the rapid growth of the global routing table. For the purpose of reducing routing table size, many studies have developed a lot of new routing solutions. After the paper introduces the background of the Internet routing system, a classification of new routing solutions is presented. Then, a typical scalable routing algorithms and architectures become the focus, and their basic ideas and characteristics are deeply analyzed and compared. Finally, some key issues and ideas for future research are discussed.

Key words: inter-domain routing; scalability; routing algorithm; routing architecture

摘 要: 全球路由表的高速膨胀,使互联网路由系统的可扩展性面临着严峻的挑战.为了缩减路由表,很多研究提出了新的路由解决方案.在介绍了互联网路由系统现状之后,从较高层次上将存在的解决方案分为短期方案、路由架构和可扩展路由算法 3 部分.着重介绍了路由算法和路由架构这两类工作,对经典的可扩展路由算法和路由架构进行了深入的分析和比较.最后讨论了有待解决的关键问题和未来的研究方向.

关键词: 域间路由;可扩展性;路由算法;路由架构

中图法分类号: TP393 **文献标识码**: A

如今,互联网路由系统的扩展性正面临着十分严峻的挑战[1−4].据统计,基于 IPv4 的全球路由表(global

* Supported by the National Natural Science Foundation of China under Grant Nos.60673168, 90818004 (国家自然科学基金)
Received 2009-12-20; Accepted 2010-06-28

routing table)表项数目前已在 30 万以上,且继续呈现指数级增长[5].庞大的路由表显著增加了路由器的内存和处理器开销,导致通信延时的增长和路由收敛属性的恶化.

为了应对路由表的膨胀,网络提供商(Internet service provider,简称 ISP)采取了以下措施:1) 升级路由器硬件;2) 压缩路由表数据结构;3) 过滤 IP 前缀.但是,升级路由器硬件提高了 ISP 的经营成本,降低了网络的性价比;并且高端路由器的性能发展能否跟上路由表的膨胀速度还是一个问题[1].压缩路由表的数据结构会带来更多的计算代价,不利于路由器的快速反应.过滤 IP 前缀将导致一些站点不可达.这些措施都没有触及根本问题.许多专家认为,为了从根本上解决路由扩展问题,有必要建立新的路由架构甚至替换现有的路由协议[1].为此,近年来,针对互联网路由扩展问题的研究掀起了一个热潮.这些研究从不同角度进行了探索,提出了很多解决方案.

互联网可扩展路由研究最近已得到国内学术界的关注.涂睿等人[6]和侯婕等人[7]对基于位置与标识分离(locator/ID split)的路由架构研究进行了综述.但是,他们的工作仅覆盖了可扩展路由研究中的一部分,没有涉及其他类型路由架构和可扩展路由算法方面的研究成果.本文对互联网可扩展路由研究方面的最新工作进行了比较全面的介绍,深入分析和比较了它们的基本思想和特点,并指出了有待解决的问题和未来的研究方向.

1 背景:互联网路由系统现状

互联网是由许多自治系统(autonomous system,简称 AS)连接而成的.一个 AS 可以自主决定在内部如何选择路由.网络运营商通常对 AS 内部的链路分配代价,然后沿链路代价之和最小的路径转发流量.这类路由选择协议如 OSPF(open shortest path first),ISIS(intermediate system to intermediate system)等.对于较大的 AS,它的网络通常被分为若干个路由区域,以便降低路由复杂性和提高路由扩展性.

在 AS 之间唯一使用的路由协议是 BGP(border gateway protocol)协议,每个 BGP 路由器会告诉它的邻居:哪些目的地址前缀标识的站点通过它的网络可达以及需要穿越的 AS 路径.因此,BGP 协议是基于路径向量的.在互联网的边缘网络中, BGP 路由器维护的路由表项数相对较少,对目的地未知的包使用缺省路由发送.然而在互联网的核心区域,BGP 路由器并不存在缺省路由,因此,该区域又被称为互联网的 DFZ(default-free zone).DFZ 路由器常常需要为互联网的每个可达的 IP 前缀安装一条路由,其结果是导致路由表随着全球 IP 前缀数量的增加而膨胀.

早期的互联网由较少的 AS 和客户网络连接而成,比较稀疏.每个 AS 被分配一块 *A* 类、*B* 类或 *C* 类地址.域间路由系统也相应地使用这几类前缀.然而,由于 *A* 类和 *B* 类地址的块过大,很快导致 IP 地址空间的耗尽,尽管很多地址并没有使用;而 *C* 类地址的块过小,往往需要多个 *C* 类地址块才能满足一个客户网络的需要,过多的 *C* 类地址前缀使得路由表不堪重负.因此,IETF(Internet Engineering Task Force)在 20 世纪 90 年代引入了无分类域间路由(classless inter-domain routing,简称 CIDR)[8]来解决这个问题.CIDR 使用一种更灵活的地址分配方案,一个 *A* 类和 *B* 类地址可以划分为多个地址块分配给不同客户,而多个连续的 *C* 类地址块可以聚合,用一个前缀来代表.CIDR 通过地址聚合能够减少 DFZ 路由表的 IP 前缀数量,因为 ISP 可以将多个连续的长前缀用一个短的前缀来发布.

CIDR 一度有效地降低了全球路由表的膨胀速度.然而,近年来各种反聚合因素的增长,使得 CIDR 的路由聚合作用逐渐失效,IPv4 前缀数量迅速增加,DFZ 的路由表再度呈现爆炸式增长.根据互联网架构委员会(IAB)在 2007 年的报告[1],这些因素主要包括:

(1) 提供商独立的地址

客户网络倾向于使用提供商独立的(provider-independent,简称 PI)地址,这样在改变提供商时可以避免重编号(renumbering)——人工地重新分配 IP 地址.因为现实中重编号往往要花费很高的代价.PI 前缀由于不能被上级提供商聚合,因此必须登记到 DFZ 的路由表中.增加的地址前缀不需要客户付费,然而 DFZ 路由表也正因此而膨胀.

(2) 多宿主

多宿主(multi-homing)是指一个站点从多个提供商那里获得服务.多宿主得到广泛应用的原因在于提供了备用路由,能够增加连接到互联网的可靠性.一个多宿主的站点可以使用 PI 地址或 PA(provider-aggregatable)地址.如果使用 PI 地址,那么 PI 前缀必须出现在它的所有提供商的路由表中;如果使用 PA 地址,那么 PA 前缀仅能够被分配该地址的提供商聚合,但是不能被其他提供商聚合.实际上,由于最长前缀匹配规则的存在,为了保证 PA 前缀可以经过它的提供商可达,往往该提供商也需要单独发布该前缀.因此,无论哪种情况都将导致前缀聚合失效.

(3) 流量工程

流量工程(traffic engineering,简称 TE)的目的是让某些互联网流量避免使用特定的网络路径.使用流量工程既包括提供商网络也包括客户网络,具体原因有负载平衡、降低费用和安全需求等.在 BGP 级,如果要对某块地址实施流量工程,那么网络运营商就必须将该地址的前缀从原来的较短前缀中分裂出来,单独发布到全球路由表中.

上述因素使得 IPv4 前缀不断分裂,DFZ 路由表中的前缀粒度越来越细,而数量越来越大.尽管 DFZ 路由表的规模受到 IPv4 的地址空间的约束,但是这并不意味着路由表的膨胀速度会减缓.随着 IPv6 的广泛部署和应用,由于 IPv6 能够提供庞大的地址空间,可能导致路由表项数以更快的速度增长[1−3].基于以上因素,很多专家认为,提高互联网路由系统的扩展性已经迫在眉睫[1−3].

2 互联网可扩展路由研究分类

为了降低路由表规模和解决互联网路由的扩展问题,目前已经提出了许多解决方案,根据着眼点的不同,这些解决方案从较高层次上可以分为 3 类.

(1) 短期方案

短期方案大多是对 BGP 协议提出增量式(incremental)的修改,并且以改善 BGP 路由的收敛属性和降低时延为主.Forgetful routing[9]可以降低路由表所占用的内存空间.它的基本思想是,选择性地丢弃路由表中的部分替代路由,只有在必要时才从邻接路由器那里获取.它不需要改变 BGP 协议,且基本不影响路由的收敛属性.但是,Forgetful routing 并没有减少 IP 前缀数量和路由表的增长速度,因此是一个短期方案.考虑到短期方案并不能真正提高路由扩展性,本文不作具体讨论.

(2) 路由架构

从中长期来解决路由扩展性的目标出发,很多研究人员提出了新的路由架构,其中许多是 IRTF(Internet Research Task Force)[10]的提案.绝大多数路由架构都是考虑在现有编址系统上增加一个间接层(indirection layer).间接机制一般将 IP 地址空间分离为主机标识和路由标识两部分,并在它们之间建立映射;或者更激进地,可能引入一种新的名字空间作为路由标识空间,IP 地址仅作为主机标识来使用.主机标识可以不包含语义,独立于提供商分配,而路由标识则隐含了网络的位置.报文传递过程会在某个阶段将主机标识用路由标识替换来穿越互联网核心.路由标识要么是能够聚合的,要么具有较大的粒度,使得路由标识的数量是可控的,从而缩减路由表.

基于路由标识的类型和来源,可以将新的路由架构分成核心/边缘分离、位置符/标识符分离、基于 AS 号的路由、基于虚拟聚合的路由等.根据路由架构以改变主机为主还是以改变网络为主,可以分为基于主机的和基于网络的.而根据报文传递过程中使用路由标识封装报文还是用路由标识改写主机标识,可以将新路由架构分为地址重写的(address rewriting)和映射&封装的(map & encapsulation).

(3) 可扩展路由算法

这类研究旨在设计路由算法降低网络节点维护的状态数量和控制开销,立足于从根本上解决路由扩展问题.针对互联网的可扩展路由算法的研究一直没有停止过.最早是 Kleinrock 和 Kamoun[11]提出的基于分区的层次化路由算法.它在互联网上得到了广泛应用,如划分自治系统以及 OSPF,ISIS 等支持分区的域内路由协议.而 BGP 协议采用的路由算法是基于路径向量的,本质上不具备良好的扩展性.到目前为止,已存在很多种可扩展路

由算法,如层次化路由、地理路由、紧凑路由、DHT(distributed hash table)路由等等.这些路由算法有的被互联网路由系统所接受,如基于提供商的层次化路由,更多的还停留在理论层面上,目的是为未来的互联网提供候选的可扩展路由方法.

Rekhter 有一句名言:“要么编址服从拓扑,要么拓扑服从编址,二者必居其一(addressing can follow topology or topology can follow addressing, choose one)”,这一针对路由扩展性的基本假设被称为 Rekhter 定律.针对互联网的可扩展路由研究大都是试图重新将互联网路由扳回到遵循该定律的轨道上来.

3 可扩展路由研究成果分析

3.1 可扩展路由算法

传统的最短路径路由算法,如距离向量、链路状态和路径向量等算法,在每个节点上要维护到所有节点的路由信息,因此路由表项数为$\Omega(n)$,其中,n 是节点总数.此时,路由表占用的内存和维护路由表的控制开销随网络规模都增长得很快,因此扩展性不好.可扩展路由算法旨在降低路由表的内存开销和控制消息开销.下面分类介绍已知的经典可扩展路由算法,并进行比较和分析.

3.1.1 基于分区的层次化路由

基于分区的层次化路由(area hierarchical routing)的基本思想是:对网络嵌套地划分区域并隔离不同区域的拓扑更新,每个节点维护较近区域的详细路由信息;而对其他较远区域只维护简略的路由信息.基于分区的层次化路由是由 Kleinrock 和 Kamoun[11]最早提出来的,也称为基于分簇的层次化路由(cluster hierarchical routing).它可以使节点的路由表项数缩减至 $kn^{1/k}$,其中,n 是节点总数,k 是分区的级数.下面简单介绍它的基本设计.

图 1 给出了一个具有 3 级的层次化网络,展示了如何对一个网络划分多级区域.原网络自顶向下首先划分为 1,2,3 这 3 个区域(第 1 级),然后对区域 1,2,3 继续划分子区域(第 2 级),接着划分第 3 级区域,这里就是节点本身.一个节点在每一级属于且仅属于一个区域.对于每个节点,根据它所属的区域使用层次化的编址:第 1 个字段对应节点所属的第 1 级区域的标识,第 2 个字段对应节点所属的第 2 级区域的标识,以此类推.

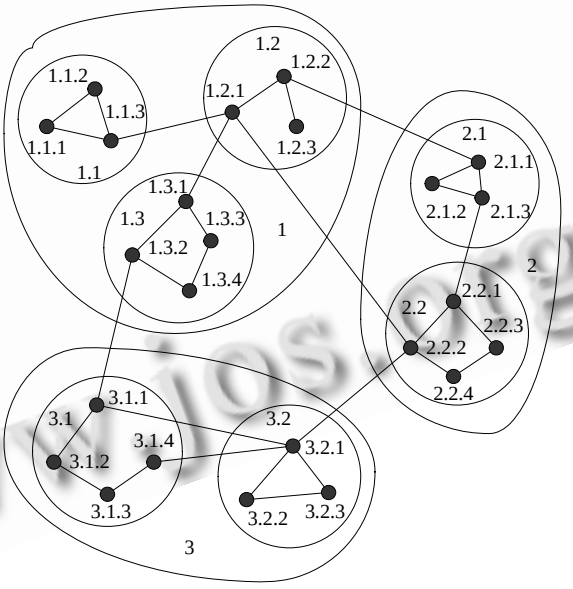


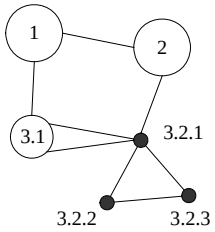
Fig.1 A 3-level area-hierarchy network

图 1 基于分区的 3 级层次化网络

在 k 级层次化网络中,一个节点 v 的路由表设置如下:如果节点自身看成是第 k 级区域,那么 v 所在的第 k–1

级区域中的每个节点在路由表中占一项;与 v 所在的第 k–2 级区域并列的每个兄弟区域占一项;以此类推,直到第 1 级区域.以图 1 中的层次化网络为例,节点 3.2.1 所能见到的网络视图和路由表设置如图 2 所示.节点 3.2.1 的路由表项数总共是 6 项(而如果不划分区域,需要 24 项,等于网络节点总数).

转发过程如下:当前节点在路由表中查找与目标地址具有最长相同前缀的项,该项代表了当前节点已知的离目标最近的区域,然后向通向该区域的下一跳转发消息;如此递归下去,直到消息到达目的节点为止.以图 1 和图 2 为例,从节点 3.2.1 路由一个消息到节点 1.2.1,使用的路径为 3.2.1-3.1.1-1.3.2-1.3.1-1.3.3-1.2.1,跳数等于 5.注意,该路径并不是最短路径,因为存在跳数等于 2 的路径 3.2.1-2.2.2-1.2.1.



Detination	Next noder	Hop number
1	3.1.1	2
2	2.2.2	1
3.1	3.1.1	1
3.2.1	–	0
3.2.2	3.2.2	1
3.2.3	3.2.3	1

Fig.2 View and routing table entries of node 3.2.1

图 2 节点 3.2.1 所具有的网络视图和路由表项

上述例子说明,基于分区的层次化路由可能导致路径长度的拉伸.从源节点 s 到目标节点 t 的路径的拉伸度(stretch)通常定义为该路径的长度与 s 到 t 的最短路径长度的比值,它总是大于或等于 1.考虑到基于分区的层次化路由存在边界效应(boundary effect)——邻近的两个节点如果划分到不同的区域中,路由时可能使用很长的路径.因此,在一般的网络上不能保证拉伸度具有较低上限.最坏情况下的路径拉伸度可能达到 $O(n)$,其中,n 是网络的节点总数.

基于分区的层次化路由思想在互联网上得到了广泛的研究和应用.表现在:互联网划分为不同的自治系统(AS),使用域间-域内两级的路由结构;并且在域内,OSPF 和 ISIS 协议也是分区的.针对域间路由,一些研究也提出了分区的层次化路由架构[12,13],具体内容将在第 3.2 节加以介绍.

3.1.2 基于地标的层次化路由

基于地标的层次化路由(landmark hierarchical routing)是对基于分区的层次化路由的改编,以使层次结构更易于动态管理.它最早是由 Tsuchiya[14]提出来的,基本思想是,迭代地选择网络中的节点作为地标并构建层次结构,每个节点的路由表只存放到本地节点和若干地标的路由信息:路由时对目标不在路由表中的消息向离目标最近的地标发送,到达地标后由地标转发给目标节点.下面介绍它的基本设计.

这类路由的一个重要概念就是地标.一个地标是从网络中选择的一个节点,设该地标的半径为 r,那么可以将与它的距离不超过 r 的那些节点的集合看成是它的邻域,邻域中的每个节点都要维护到该地标的路由信息.值得注意的是,地标可以不需要维护到邻域中节点的路由信息.

地标的层次化结构按如下方式构建:设 $LM_i[id]$代表一个第 i 级地标,id 是它的标识.网络中所有的节点可以看成是第 0 级地标,设第 0 级地标的半径是 r_0.对于所有 i,第 i+1 级地标是从第 i 级坐标中选择,第 i+1 级地标的半径 r_{i+1} 必须大于 r_i,以便对于任意 $LM_i[id]$,在它的 r_i 跳范围内至少存在一个第 i+1 级地标.最顶级地标的半径必须足够大,以便对所有节点都是可见的,即所有节点都要维护到全部顶级地标的路由信息.每个节点的路由表只维护到它可见的地标的路由信息,也就是说,对于任意 $LM_i[id]$,如果一个节点离 $LM_i[id]$的跳数不超过 r_i,那么在该节点中的路由表中安装一条到 $LM_i[id]$的路由信息.

在具有 k 级地标的层次化网络中,每个节点的地址由一系列地标标识构成,表示为$(LM_{k-1}(id_{k-1}),\ldots,LM_0(id_0))$,地址中每个地标必须是在它的上一级地标的半径范围内.图 3 展示了一个基于地标的 3 级层次化网络,并给出了每个节点的地址.以节点 g 为例来观察它的路由表项目.由于 r_0=2,所以第 0 级地标 f,k,e 分别在路由表中占据一项.类似地,第 1 级地标 i 和第 2 级地标 d 分别在路由表中各占据一项,因此,g 的路由表总共含有

5 项.

路由时先检查目标地址($LM_{k-1}(id_{k-1}),\ldots,LM_0(id_0)$),在路由表中查找标识包含在目标地址中且级别最低的地标,然后向该地标转发消息.如果消息到达了目标节点能见的某个地标,设为 $LM_i[id]$,因为 $LM_i[id]$的路由表至少含有到 $LM_{i-1}(id_{i-1})$的路由信息,所以 $LM_i[id]$一定能够将消息转发给离目标更近的地标.如此重复下去,消息将最终转发到目的地.以图 3 中的网络为例(所有节点都是第 0 级地标,用两个环绘制的节点是第 1 级地标,用 3 个环绘制的节点是第 2 级地标.r_0=2;r_1=4;r_2=8),考虑从节点 g 发送一个消息到节点 t,将使用路径 g-f-e-d-u-t,总跳数等于 5.该路径并不是最短路径,因为 g 到 t 的最短路径是 g-k-i-u-t.

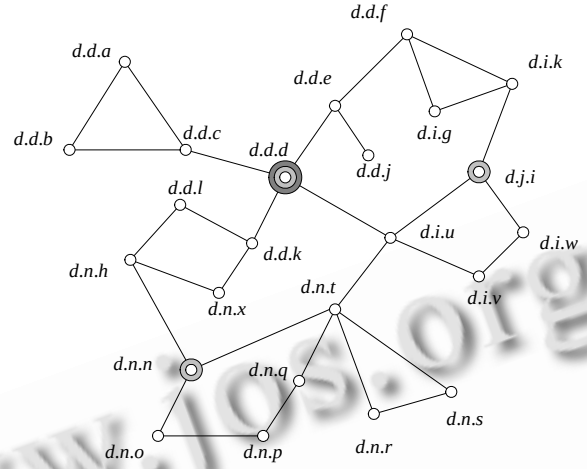


Fig.3 A 3-level landmark-hierarchy network

图 3 基于地标的 3 级层次化网络

基于地标的层次化路由可以大幅度地缩减节点的路由表,典型情况可以缩减至 $O(\sqrt{n})$[14].基于地标的层次化路由可以解决基于分区的层次化路由存在的边界效应问题,邻近节点之间的路径质量得到了提高.由于具有相对较好的动态管理特性,基于地标的层次化路由比较适用于一些拓扑变化较快的网络,如无线自组织网络.缺点是,由于许多流量都要经过地标转发,地标节点可能成为网络的瓶颈,且地标层次越高越容易成为瓶颈.

3.1.3 基于提供商的层次化路由

基于提供商的层次化路由(provider hierarchical routing)是目前互联网域间唯一使用的可扩展路由机制,CIDR 的地址聚合就是以该方法为基础的.基于提供商的层次化路由的基本思想是,利用域间存在的提供商-客户(provider-customer)层次结构,对客户网络分配可由提供商聚合的 IP 地址,即 PA 地址.

图 4 展示了互联网的基于提供商的层次结构(带箭头的实线代表客户-提供商连接,虚线代表对等连接).Stub 域必须连接到一个提供商域来获取转发(transit)服务,而本地级或地区级提供商必须连接到骨干提供商来获取转发服务.互联网的层次结构并不是树状的.一个域可以连接到多个提供商,这种情况称为多宿主.并且两个域之间还可以建立对等(peer-peer)关系,以互相提供穿越服务,如 Regional 2 和 Regional 3.一个域甚至可以直接与不同层次的提供商相连,如 Local 1.因此,一个域可能具有由不同的提供商分配的多个地址前缀.如果一个客户网络改变了它的提供商,那么它应当从它的新提供商那里获得 IP 地址并重新配置,该过程称为重编号.

网络重编号需要很高的代价,而且使用 PA 地址也将阻碍多宿主、主机移动性和流量工程的应用.因此,客户网络越来越倾向于使用 PI 地址,而不是 PA 地址.这些使得基于提供商的层次化路由在互联网上失效,从而导致路由表膨胀.一种解决方法是将 IP 地址同时扮演的位置符和标识符角色分离,对客户网络分配 PI 地址来标识身份.但是使用 PA 地址来标识位置不需要重编号,并且支持多宿主和流量工程.这类工作将在第 3.2 节加以介绍.

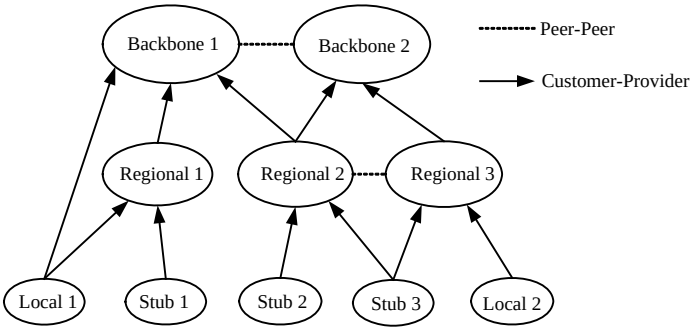


Fig.4 Hierarchical structure of the Internet

图 4 互联网的层次结构

3.1.4 DHT 路由

DHT 技术在网络上构建了一种键-值(key-value)查询服务,节点和键共用一个特定的标识空间.假定键对应的值存放在标识与键相等或相近的节点上,从一个节点上根据键查找值的过程等同于从该节点到相应目的节点的路由过程.到目前为止,DHT 主要用于在网络层之上构建逻辑网络,前提是存在网络层路由协议.这种逻辑网络是根据节点的标识度量空间来构建的,节点之间的逻辑距离可以由它们的标识推算出来,每个节点将收到的消息转发给离目的节点“距离”最近的邻居,从而逐步逼近目的节点.DHT 网络通常具有比较规则的拓扑结构,如环状[15]、树状[16]等.

Caesar 等人[17]考虑了在物理网络上直接构建 DHT,提出了一种虚拟环路由(virtual ring routing,简称 VRR)方法.VRR 使用不包含语义的扁平标识路由,只要保证标识唯一即可.基本思想是,将所有节点根据它们的标识大小顺序组织成一个虚拟环.假定标识使用{0,1,2,…,n−1}中的数字,节点 i 在路由表中维护如下的路由信息:1) 到标识等于 i−1 和 i+1 的节点的路由信息,因此形成虚拟环;2) 到物理邻居的路由信息;3) 设 P 是一对标识相邻的节点之间的最短路径,如果 P 经过节点 i,那么存放到 P 的两个端点的路由信息.路由时,当前节点从路由表中选择标识与目标标识最接近的节点转发消息.VRR 保证在任意一对连通的节点之间总是可达的.平均路由表项数一般在 $O(n^{1/2})$,其中,n 是网络的节点总数.但是,VRR 不能保证拉伸系数的大小,最坏情况下,拉伸系数可能达到 $O(n)$.VRR 路由方法在多种类型网络上的应用都得到了研究,如无线自组织网络[17]、企业局域网[18]和互联网[19]等.

ROFL(routing on flat labels)[19]是一种基于 VRR 的面向互联网的可扩展路由算法.ROFL 在网络层消除了位置符,完全使用与位置无关的标识进行路由.该方法继承了位置符/标识符分离策略的诸多优点,如主机移动性、多宿主和标识稳定等,但是不需要引入映射系统,编址和路由显得更加简单.ROFL 的基本原理是,在网络层将主机和路由器标识组织成层次化的 DHT,能够支持基本的域间路由策略.

ROFL 能够大幅度地缩减路由表规模,仿真结果表明,在具有 6×10^8 个节点的网络中,平均情况下,路由表规模仅为 10^2 的数量级.但是,ROFL 使用的路径与最短路径相比,长度显著增加,表明在路径质量上还不太理想.

3.1.5 紧凑路由

紧凑路由(compact routing)是一类在理论上保证同时具有较小路由表和较低拉伸度的路由算法.紧凑路由使每个节点的路由表规模为 $O(n)$,即随节点总数 n 呈亚线性增长,因此保证了路由表的可扩展性.它的基本思想是,平衡路由表大小和路径长度,允许有限的路径拉伸以大幅缩减路由表.

紧凑路由算法根据节点的地址是拓扑有关的还是拓扑无关的而分为命名约束的(name-dependent)和命名独立的(name-independent)两类.根据网络拓扑类型是任意的还是特定的,可以分为通用的(universal)和专用的(specialized)两类.命名约束的算法对节点分配包含拓扑信息的地址,代表工作有 Cowen 算法[20]、TZ 算法[21].Cowen 算法是第 1 个最大拉伸度不超过 3 且路由表大小为 $O(n)$的通用路由算法,每个节点的路由表大小

最多为 $\tilde{O}(n^{2/3})$ 比特($\tilde{O}(f(n))$ 是 $O(f(n)\log^c n)$的缩写,其中,c 取常数).TZ 算法是 Cowen 算法的改进,在最大拉伸度不超过 3 的情况下,路由表大小不超过 $\tilde{O}(n^{1/2})$ 比特.命名独立的算法允许节点使用任意的名字路由,比命名约束的算法相对更加复杂,代表工作有 Arias 等人提出的算法[22]和 Abraham 等人提出的算法[23].Arias 算法的最大拉伸度等于 5,路由表大小等于 $\tilde{O}(n^{1/2})$ 比特.Abraham 算法与 TZ 算法在拉伸度和路由表大小方面取得了近乎一致的理论结果,即拉伸度和路由表大小的上限分别为 3 和 $\tilde{O}(n^{1/2})$ 比特.Gavoille 和 Gengler[24]证明不存在通用路由算法满足拉伸度小于 3 且每个节点的路由表大小为 $O(n)$.因此,TZ 算法和 Abraham 算法很大程度上都是最优的.

进入 21 世纪以来,专用的紧凑路由算法成为研究热点.真实的网络往往呈现特定的拓扑结构特征.如互联网 AS 级拓扑具有幂律[25]、强聚集[26]、富人俱乐部[27]、自相似[28]和核心稳定[29]等许多特征.因此,通用的紧凑路由算法在实际网络上可能并不是最优的.研究人员常常用特定结构的图来建模真实的网络,如随机图(random graph)、幂律图(power-law graph)、增长受限的图(growth-bounded graph)、平面图、UDG 图(unit disk graph)等.关于这些特定图上的紧凑路由的研究如文献[30–37]等.

将紧凑路由用于互联网的设想始于 Krioukov 等人[30].Krioukov 等人使用 TZ 算法在幂律图和真实的互联网 AS 图上进行了仿真发现,平均的路由表项数很小,而平均拉伸度约为 1.1.此后,针对类互联网拓扑结构的图的紧凑路由开始得到更多关注.Brady 和 Cowen[31]提出了一种具有增量型拉伸系数(1,d)的紧凑路由算法(简称 BC 算法),路由表大小具有 $O(e\log^2 n)$比特的上限.通过仿真证明,在无标度网络上,d 和 e 都可以取较小的值,平均拉伸度比使用 TZ 算法更低.文献[32]同时利用了幂律特征和强聚集特征,在类互联网的图上可以获得比 BC 算法更优的平均路由性能.从实验角度分析幂律网络紧凑路由性能的类似工作还有文献[33,34]等.最近的工作[38,39]从数学上证明了随机幂律图上的紧凑路由可以取得比一般图上更优的性能.文献[40]对紧凑路由和在互联网上的应用问题进行了综述.

然而,为了保证路由表规模和拉伸度具有较低上限,紧凑路由算法需要精心地为节点分配标记和构造路由表,这样有可能引起这些数据结构维护代价的上升,增加了复杂度.

3.1.6 地理路由

地理路由(geographical routing)是指基于节点的地理位置使用贪心路由.一个节点通常只需要维护自己和邻居的地理位置,总是选择在地理位置上离目的节点最近的邻居作为消息的下一跳.地理路由有可能陷入局部最小点(local minimum),即找不到比自己离目的节点更近的邻居,因此贪心路由将失效.解决该问题的常用方法是,使用基于拓扑的替代路由方案,如面路由[41]等.一旦贪心路由遇到局部最小点,就启用替代路由方案来跳出局部最小点,然后再恢复贪心路由.

地理路由在每个节点上使用很小的路由表(与节点度成正比),几乎不需要维护网络的拓扑信息,因此具有十分理想的可扩展性.地理路由最早由 Finn 提出[42],在互联网和无线网络领域都得到了广泛研究.文献[43]较早提出了在互联网上如何使用地理信息编址和路由的方案,并与基于提供商的路由方案进行了比较;文献[44]针对 IPv6 提出了使用基于地理位置的编址方案;文献[45]则考虑了在包首部中携带地理信息来辅助互联网路由.

为了保证较高的路由成功率和路径质量,地理路由要求网络连接密度尽可能地高,即要求地理位置靠近的节点在拓扑上也是邻近的,最好是直接互连的.然而,真实网络可能难以满足上述条件.对互联网来说,AS 之间的连接关系是由它们之间的商业利益决定的,相邻地区的 AS 并不一定是互连的.而且,互联网的路由选择要受到策略(policy)的限制.因此,地理路由并不能直接用于当前的互联网.

3.1.7 图嵌入路由

图嵌入(graph embedding)的基本思想是,对网络中的每个节点分配虚拟坐标,将节点映射到虚拟几何空间或特定度量空间中的点,简化节点之间的“距离”计算.通过图嵌入,路由算法也可以使用贪心路由,每个节点只需要知道邻居的坐标,在转发消息时总是选择与目的节点“距离”最近的邻居转发,这一点与地理路由相似.但是与地理路由方法不同,虚拟坐标的构造不需要感知节点的物理位置,而通常是基于网络的拓扑信息(也许还包括网络的其他信息)来构造.这样做的优点是,虚拟坐标能够反映网络的连通信息,能够提高路由的效率.缺点是当网络的拓扑发生改变时,为保证一致性,至少一部分节点的虚拟坐标也要随之更新,因此动态属性不如地理路由.

基于图嵌入的路由适用于局部性较好的网络,因此这方面的研究大多针对无线网络.文献[46]提出将网络节点嵌入到多维欧几里德空间,使用欧几里德坐标来标识节点的位置;文献[47]则使用了一种类似于赋范空间(norm space)ℓ_1的实数向量空间R^d,节点的坐标是由该节点在网络中到d个地标的距离所构成.图嵌入也可能存在局部最小点问题.如果分配的虚拟坐标与网络连通信息不一致,贪心路由可能失败,这一点类似于地理路由.为了解决这个问题,文献[48]提出了贪心嵌入(greedy embedding)的概念.贪心嵌入是指保证在任意连通的节点之间用贪心路由可达的嵌入.贪心嵌入研究是最近的一个研究热点.典型的有:文献[49]提出的贪心嵌入是从网络的一个生成树构造一个双曲空间,将网络节点嵌入到该双曲空间中;文献[37]针对UDG图,使用图的多个生成树将图贪心嵌入到多维的向量空间,路由算法在每个节点上产生$O(\log^3 n)$比特的路由表,拉伸系数等于3.类似的工作还有文献[50].

研究发现,互联网AS级拓扑也存在较强的局部聚集特征[51,52].Krioukov等人[53]认为,基于复杂网络的隐藏度量空间的贪心路由有可能为互联网提供一种理想的路由方法.文献[53]提出了从一种双曲空间构造类互联网拓扑的无标度网络的模型,发现基于该双曲空间的贪心路由比较理想的性能.文献[54]提出了将复杂网络嵌入到由它的骨架导出的度量空间的方法,分析表明,基于该度量空间的贪心路由具有高扩展性.

3.1.8 比 较

不同种类的可扩展路由算法适用的拓扑类型和场景可能不一样,因此很难比较优劣.表1比较了它们的基本思想、适宜的网络拓扑类型、路由表大小、拉伸度和应用于域间路由的可行性.

Table 1 Comparison of various types of scalable routing algorithms

表1 不同类型可扩展路由算法的比较

Routing algorithms	Basic idea	Preferred topological types	Routing table size (entries)	Stretch	Applicability to inter-domain routing
Area hierarchical routing	Dividing network into clusters and each cluster into sub-clusters nestedly	Sparse or modular	$kn^{1/k}$ in the optimal case	$O(n)$ for arbitrary networks	Not so far
Landmark hierarchical routing	Selecting landmarks iteratively to build a landmark hierarchy	Arbitrary	Typically in $O(n^{1/2})$	Typically better than area hierarchical routing	Not so far
Provider-Based hierarchical routing	Addressing according to the provider-client hierarchical structure of the Internet	Tree-Like	$O(k)$ for any node with degree k	One for tree-like networks but $O(n)$ for arbitrary networks	Yes
Geographical routing	Greedy routing over geographic coordinates	With high link density	$O(k)$ for any node with degree k	Low in average; but routing fails sometimes	Not so far
Graph embedding based routing	Greedy routing over virtual coordinates via embedding	With good locality	$O(k)$ for any node with degree k	Depends on the quality of embedding; routing fails sometimes	Not so far
Compact routing	Trade-Offing routing table size by relaxing stretch	Arbitrary	$\tilde{O}(n^{1/2})$ for arbitrary networks	$O(1)$ for arbitrary networks	Not so far
Virtual ring routing (VRR)	Routing on flat labels by constructing DHT directly on the network layer	Arbitrary	$O(n^{1/2})$ in average	Typically high compared to other scalable routing algorithms	Not so far

可以总结,为了缩减路由表,可扩展路由算法大都是在两个方面进行平衡:一方面是平衡路由表大小和路径长度,允许使用非最优路径来降低维护的路由状态数;另一方面是平衡路由表大小和节点地址的复杂度.除了地理路由和VRR,上述其他可扩展路由方法都是通过对节点分配依赖于网络拓扑的地址来缩减路由表,同时有利于提高路径质量.但是,这样不可避免地会引起节点地址复杂度的增加.

3.2 路由架构

为了克服现有域间路由系统的不足,很多研究提出新的域间路由架构.这些路由架构根据其基本特点可分为下面几类.值得注意的是,该分类并非严格的划分,一种路由架构可能兼有多种特点,因此也可属于不同类别.

3.2.1 基于 AS 号的路由

由于 AS 具有比 IP 前缀更大的粒度,因此一些研究提出在域间基于 AS 号而不是 IP 前缀来路由.这样,核心路由表的表项数最多等于 AS 的总数.由于 AS 总数目前比 IP 前缀总数少一个数量级,因此可以大幅度地缩减路由表规模.这类工作的一个代表是 HLP(hybrid link-state path-vector routing protocol)[55].HLP 还利用了划分区域的分级路由思想,根据提供商-客户的层次关系对域间网络划分树状区域,对不同的区域进行隔离,将路由更新和故障限制在本地区域内.HLP 使用了链路状态和路径向量两种路由协议来提高域间路由的收敛属性.但是,域间拓扑由于多宿主和对等连接的广泛使用已经远不是树状的,划分树状区域在效果上并不理想.并且,HLP 要求 ISP 公开它们之间的连接关系,这在目前也是不太可行的.

Atomized routing[56]引入了一种称为“原子”的对象,聚合那些拥有相同 AS 路径的长 IP 前缀,在一定程度上也可以看成是基于 AS 号路由的.原子路由被设计用于边缘的客户网络,即被通告的原子来自于客户网络.在核心网络中,根据原子标识或 IP 前缀路由.通过用原子标识聚集长 IP 前缀,核心路由表规模得到了缩减.

Nimrod[57]也可以归于此类路由架构.它引入一种 map 分发机制来帮助客户网络发现拓扑和选择路由.在互联网上,由于 AS 级拓扑目前相对较小,因此路由表规模可以在一定程度上得到控制.然而,Nimrod 并没有考虑域间丰富的路由策略,因而目前基本上不具备可部署特性.

基于 AS 号的路由必须构建和维护 IP 前缀到 AS 号的映射表,并提供查询.从长期来看,这类路由架构的路由扩展性仍然受到较强的限制.这是因为近年来分配的 AS 号的增长速度比 IP 前缀数量增长得更快,从 2000 年的 10 000 个左右迅速增加到目前的近 50 000 个[7],AS 号目前已从 16 位升级到 32 位.照此趋势,全球路由表的增长速度仍得不到有效控制.

3.2.2 基于虚拟聚合的路由

这类路由架构也是基于比 IP 前缀更大的拓扑粒度来路由.基本思想是,分离网络的拓扑和编址,使用虚拟聚集对象来聚合较长的 IP 前缀,在核心网络中使用虚拟聚合对象的标识来路由.这类工作主要有 CRIO(core router-integrated overlay)[58],ISLAY(见:http://ietfreport.isoc.org/idref/draft-irtf-routing-islay/)[59]等.

CRIO 使用一种虚拟前缀对 IP 前缀进行聚合.如图 5 所示(携带长 IP 前缀的包先用虚拟前缀发送至相应的聚合代理,然后由代理隧道至目的网络),虚拟前缀与核心区域的路由器关联,这些路由器称为聚合代理(aggregation proxy).聚合代理用较短的虚拟前缀能够聚合很多较长的 IP 前缀,虚拟前缀的分配可以与网络拓扑无关.聚合代理向其他核心路由器发布自己的虚拟前缀,并负责将接收到的包用隧道方式路由到与实际的 IP 前缀对应的路由器.由于核心区域使用了较短的虚拟前缀,因此 CRIO 可以使核心路由表缩小两个数量级.

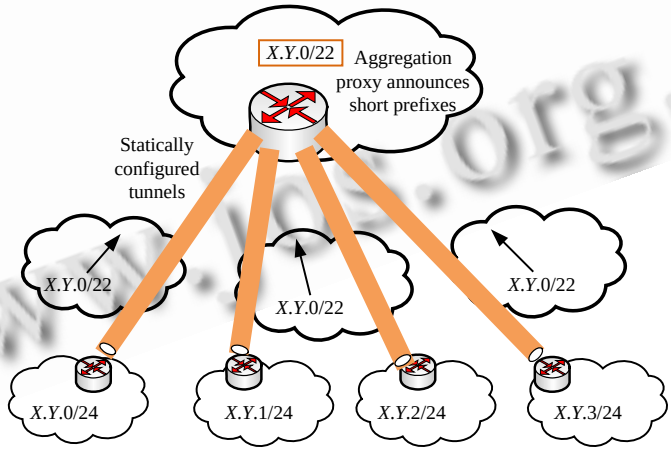


Fig.5 Virtual aggregation based architecture

图 5 基于虚拟聚合的架构

虚拟聚合的一大不足之处是,使用的路径长度比原来的路径有所拉伸.实际上,虚拟前缀的粒度越大,聚合的子前缀越多,路径的拉伸度越高.因此,这里存在一个平衡问题.此外,聚合代理由于集中了所有流向被聚合的子前缀的流量,很容易成为网络的瓶颈.

3.2.3 核心与边缘分离

核心与边缘分离(core-edge separation)是指将位于核心的提供商网络(即 transit 网络)与位于边缘的客户网络(即 stub 网络)分离,即分离它们的地址和路由空间.这种架构的提出者认为,互联网路由扩展问题的根源在于核心网络和边缘网络使用同一个地址和路由空间.核心网络的地址聚合要求与客户网络的流量工程、多宿主等反聚合因素存在矛盾.并且,核心网络无论是在 IP 前缀数量上,还是在拓扑上都相对比较稳定,而边缘网络的 IP 前缀数量和引发的路由更新都多得多,并且增长得更快.通过分离核心网络和边缘网络,在核心路由器中只维护核心网络通告的 IP 前缀,可以大幅度地缩小核心路由表,并且能够隔离客户网络引起的路由更新,减少路由更新数量以及提高路由收敛属性.这方面的工作有 eFit[12],Hidra[13]等.

eFit 提出将 IP 地址空间划分为两种,即提供商地址和用户地址.前者在核心网络中使用,而后者针对边缘网络,不能出现在核心路由表中,但必须是全局可达的.图 6 给出了 eFit 架构中端到端的路由过程(包在进入核心网络时,使用核心地址封装隧道至出口边界路由器),描述如下:源自客户站点的 IP 包首先使用目的用户地址路由到提供商网络的入口边界路由器 *P*1;然后,*P*1 根据用户地址通过映射系统查询对应的目的提供商地址,即目的提供商网络的边界路由器地址,设为 *P*2;在 *P*1 处将 *P*2 作为首部对 IP 包进行封装,将包隧道至 *P*2,*P*2 对收到的包进行拆封,再使用目的用户地址路由至目的主机.eFit 对核心路由表的缩减程度依赖于提供商地址的聚合程度.一般地,核心路由表规模与提供商网络数量呈线性比例.

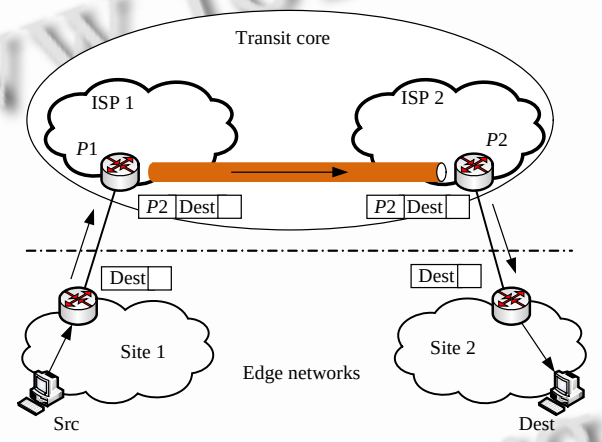


Fig.6 Core-Edge separation architecture

图 6 核心网络与边缘网络隔离的路由架构

Hidra 在核心网络中使用一种域间路由标识作为地址.一个域间路由标识由中介网络的 AS 号以及它的提供商网络的 AS 号唯一确定.因此,Hidra 在核心网络中使用了比 IP 前缀更大的路由粒度.Hidra 能够大幅度地降低核心路由表规模和路由更新数目.但是,使用新的路由标识要求对路由协议做出更多的改变.

核心-边缘分离的思想实际上在多数的路由架构和 IRTF 提案中都得到了不同程度的应用.不同的是,其他的路由架构,如位置与标识分离等,并不严格要求隔离边缘和核心的路由空间,而是允许路由协议穿越边界.核心-边缘分离方案也需要一个映射系统维护用户地址空间到提供商地址空间的映射,并提供查询.

3.2.4 位置符与标识符分离

位置符/标识符分离(locator/ID split)的提出者认为,现有互联网路由系统扩展性差的一个因素是 IP 地址语义过载,即既作为位置符又作为身份标识符.当 IP 地址标识终端身份时,它是根据 ISP 组织结构而非拓扑结构来分配,导致目前互联网唯一有效的路由缩减技术——拓扑聚合的失效.解决的方法是对网络节点引入位置符与

标识符两种地址空间.节点的位置符可以按照基于提供商聚合的方式分配,因此路由表规模能够得到有效控制.而节点的标识符可以使用PI地址,在应用层和传输层使用.拓扑变化时标识符不需要改变,只需要改变相关的位置符.位置符/标识符分离给多宿主、流量工程、主机移动性等提供了较好的支持.最近几年提出的这类路由架构和IRTF提案如LISP[60],Ivip[61],TIDR[62],TRRP[63],APT[64],IPvLx[65],GSE[66],Six/One[67],Six/One Router[68],HIP[69],Shim6[70]等.Cisco公司已经实现了LISP的一个原型系统.

根据位置符/标识符分离的位置,这类路由可以分为基于主机的、基于网络的和基于主机+网络的.基于主机的方案,如HIP,Shim6等.这类方案采用主机可见的新名字空间,在主机协议栈网络层之上增加了新的标识层,由主机完成对ID/Locator的解析和转换,一般采用地址重写的处理方式.基于主机的方案需要修改主机的应用层和传输层,部署起来比较困难;并且由于位置符必须使用PA地址,缺乏对站点多宿主和流量工程的支持.基于网络的方案,如GSE,LISP,Ivip,TIDR,TRRP,APT,IPvLx等.这类方案对主机不需要做出改变,在网络边缘路由器上进行ID/Locator转换,一般采用映射&封装的处理方式.以LISP为例,位置符分配给边缘网络路由器,而标识符分配给端点设备.在客户网络,本地主机通常可以使用标识符通信,但是穿越核心网络使用隧道方式.即在隧道入口,路由器(ITR)通过映射系统查询目的位置符,然后使用目的位置符在核心网络中路由,到达隧道出口路由器(ETR),然后再使用目的标识符路由.ITR(ingress tunnel router)和ETR(egress tunnel router)可以放置在客户网络中,也可以放置在提供商网络中.基于网络的方案可部署性较好,但是提供商网络的流量工程仍然受到很大的限制.基于主机+网络的一个典型方案是Six/One.该方案依赖于主机和网络共同完成对ID/Locator的解析,兼有基于主机和基于网络的方案的优点,但是由于需要改变主机和网络设备,部署开销仍然较大.

位置符/标识符分离架构中的一个核心部件是位置符/标识符映射系统,不同的方案在映射系统的设计上可能有很大的区别.根据映射信息的存储和访问模式,映射系统可以分为push模型、pull模型以及push+pull混合模型.push模型是指每个边缘路由器ITR拥有一份完整的映射表,ID/Locator查询在本地就可以完成,查询时间快,但是状态规模很大.pull模型中映射系统可以分布存放,ITR通过访问映射服务器得到映射信息并缓存到本地,ITR维护的状态规模小,但是会增加延迟.push+pull混合模型则平衡了push模型和pull模型的优缺点.push模型的典型代表如LISP-NERD[71].pull模型的典型代表包括TRRP,TIDR以及LISP-DHT[72,73]等.基于push+pull混合模型的则包括APT,LISP-CONS[74],LISP-ALT[75]等.GSE,Six/One,HIP,Shim6等没有引入特别的映射系统,而是借助于DNS(domain name system)系统来实现映射信息的分发和查询.

3.2.5 支持源路由的架构

这类路由架构的特点主要是,在解决路由扩展性问题的同时支持用户源路由选择和多路径路由.代表工作有NIRA[76]和Pathlet Routing[77].

NIRA(new Internet routing architecture)根据用户站点到互联网核心(通常由tier-1 AS构成)或对等连接点的上行路径来对用户站点编址,一条路径用一个地址来编码,这种编址方式是基于提供商聚合的.NIRA主要是针对IPv6的,但是也可以改编用于IPv4.此时,一条上行路径可以用一串IPv4地址来编码.假设使用IPv6地址,那么从源主机到目的主机的一条路径可以通过它们的地址来联合推出.如图7所示(用户可以根据自己的地址和目的地

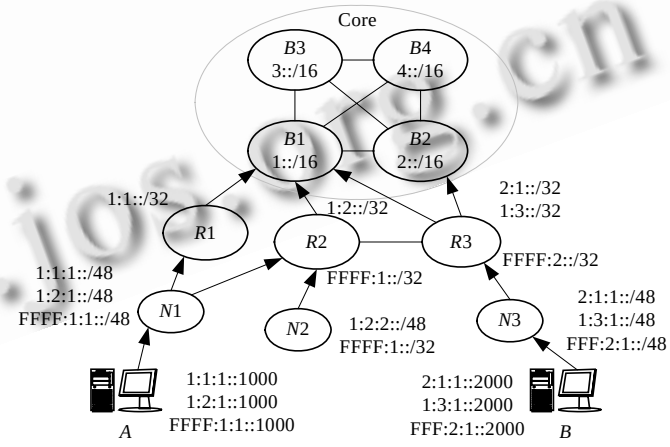


Fig.7 Example of the strict provider-rooted hierarchical addressing in NIRA

图7 NIRA使用严格的基于提供商的层次化地址分配方案

址选择一条路径),网络 *N*1 的用户 *A* 具有从核心网络分配来的地址 1:1:1::1000 和 1:2:1::1000,以及从对等连接点分配来的地址 FFFF:1:1::1000.网络 *N*3 的用户 *B* 具有从核心网络分配来的地址 2:1:1::2000 和 1:3:1::2000,以及从对等连接点分配来的地址 FFFF:2:1::2000.用户可以使用源地址和目的地址来计算一条合法的路径.如用户 *A* 可以使用源地址 1:1:1::1000 和 *B* 的地址 1:3:1::2000 来得到一条路径 *N*1→*R*1→*B*1→*R*3→*N*3,或者用源地址 FFFF:1:1::1000 和 *B* 的地址 FFFF:2:1::2000 得到路径 *N*1→*R*2→*R*3→*N*3.可见,NIRA 不仅允许用户选择路径,还支持多路径路由.

类似于 HLP,NIRA 也根据域间提供商-客户层次关系对全部 AS 划分树状区域,隔离不同区域的故障和路由更新.基于层次化的地址分配和区域划分,NIRA 可以获得较好的路由扩展性.但是,由于对等连接和多宿主等因素的增长,这种地址分配方案可能导致一个用户站点具有太多的地址.并且,用户站点的地址在改编提供商时必须重新分配.此外,NIRA 也不能支持复杂的路由策略.

Pathlet Routing 的基本思想是,将每个 AS 中的路由器聚合成一些抽象的虚拟节点 vnode,AS 提供的路由线路就是由一串 vnode 构成.这样的 vnode 序列称为 pathlet,AS 可以向互联网发布自己提供的 pathlets.源站点可以将连接源和目的地的一串首尾相连的 pathlets 作为一条端到端的源路由.这个框架提供了比较灵活的路由选择,因为 vnodes 能够以任意的粒度来抽象网络,pathlets 能够表达丰富的路由策略,同时提供了大量可选的路径,从而使得源站点能够从指数级的组合方式中选择并缝合 pathlets 形成端到端路由.Pathlet Routing 结合了 BGP 和源路由这两种路由策略的优点,其代价是更复杂的控制平面和更多的控制消息开销.

3.2.6 比 较

表 2 对上面重点讨论的路由架构进行了比较.比较的项目包括:基本思想、针对 IPv4 还是 IPv6、是否需要引入映射系统、可扩展性、是否支持流量工程、是否支持多宿主、是否支持主机移动性、是否易于部署等.

Table 2 Comparison of characteristics of routing architectures

表 2 路由架构的基本特点比较

Solution	Basic idea	IPv4/ IPv6	Mapping system	Scalability	Traffic engineering	Multi-Homing	Mobility	Deployability
HLP	AS-Based routing	IPv4/ IPv6	Yes	Medium	Yes-Limited	Yes-Limited	No	Poor
Atomized routing	AS-Based routing	IPv4/ IPv6	Yes	Medium	Yes	Yes	No	Poor
CRIO	Virtual aggregation	IPv4	Yes	Good	Yes	Yes	No	Medium
ISLAY	Virtual aggregation	IPv4/ IPv6	Yes	Good	Yes	Yes	Yes	Poor
eFit	Core-Edge separation	IPv4/ IPv6	Yes	Good	Yes	Yes	Yes	Good
Hidra	Core-Edge separation	IPv4/ IPv6	Yes	Good	Yes	Yes	Yes	Medium
Six/One	Locator/ID split; network+host based;	IPv6	No-Using DNS	Medium	Yes-Limited	Yes-Limited	Yes	Poor
Six/One router	Locator/ID split; network-based;	IPv6	Yes	Good	Yes	Yes	Yes	Good
TIDR	Locator/ID split; network-based;	IPv4/ IPv6	Yes	Good	Yes	Yes	Yes	Poor
LISP	Locator/ID split; network-based;	IPv4/ IPv6	Yes	Good	Yes	Yes	Yes	Good
Ivip	Locator/ID split; network-based;	IPv4/ IPv6	Yes	Good	Yes	Yes	Yes	Good
TRRP	Locator/ID split; network-based;	IPv4/ IPv6	Yes	Good	Yes	Yes	Yes	Good
HIP	Locator/ID split; host-based;	IPv6	No-Using DNS	Medium	Yes-Limited	Yes-Limited	Yes	Poor
Shim6	Locator/ID split; host-based;	IPv6	No-Using DNS	Medium	Yes-Limited	Yes-Limited	Yes	Poor
GSE	Locator/ID split; network-based;	IPv6	No-Using DNS	Medium	Yes	Yes	No	Poor
NIRA	Source-Based routing	IPv4/ IPv6	Yes	Medium	Yes-Limited	Yes-Limited	No	Poor

4 总结与展望

互联网可扩展路由方面的研究已经取得了很多成果,主要包含路由算法和路由架构两个方面的工作:前者着眼于从算法角度提出优于现有互联网路由算法和协议的解决方案;后者着眼于从架构的角度提出解决方案,强调可部署性,尽可能地不改变现有的路由协议.然而迄今为止,多数解决方案还停留在理论上,还没有一种方案得到应用和部署,针对未来互联网路由的解决方案还在不断形成.鉴于互联网的重要性和复杂性,对互联网路由系统做出改变或者找到理想的路由解决方案都不是一件容易的事情,有很多问题需要考虑和解决.下面从路由算法和路由架构两个层面来对其中的关键问题和未来的研究方向进行探讨.

4.1 路由算法

这方面的工作目前聚焦在寻求比地址聚合更好的可扩展路由机制和比 BGP 协议更具可扩展性的路由算法上.地址聚合是当前的互联网唯一使用的可扩展路由机制,然而如今的互联网自治系统级拓扑结构已经逐渐从树状演化成 mesh 状,呈现无标度特征.一些专家认为,地址聚合在本质上越来越与互联网的拓扑结构相背离,故聚合的效果将受到限制[40,53].另一方面,BGP 协议是以最优化路由选择为基础的,如同传统的最短路径路由算法一样,从根本上来说并不具备高扩展性,因为每个节点都需要维护到所有其他节点的路由状态.目前,针对互联网虽然已提出了许多种可扩展路由算法,但是以下问题仍然值得研究:

(1) 大多数可扩展路由算法,如紧凑路由、贪心路由等,仍然停留在静态性能的水平上,将互联网抽象成一个静态的图进行分析,很少考虑互联网上的实际因素,如拓扑变化、节点移动、路由策略约束等.在更加实际的互联网环境下,对这些算法的改进和评估值得进一步研究;

(2) 利用互联网的拓扑结构特征和演化机理来提高路由扩展性是当前的一个热点.然而,目前人们在互联网拓扑结构和演化机理方面的知识还比较有限.随着这方面研究的深入,针对互联网的可扩展路由预期将得到进一步优化.

4.2 路由架构

位置符/标识符分离是目前受到最多关注的一类路由架构,现阶段 IRTF 的路由架构提案几乎都属于该类.除了可以解决路由扩展问题以外,这种架构还带来了许多额外的好处,能够更好地满足当前路由架构无法提供可扩展性支持的应用需求,例如多宿主、流量工程、终端移动性等.但是,也仍然存在以下的关键问题尚待进一步研究:

(1) 报文处理方式的评估问题

位置符/标识符分离架构中有两种报文处理方式,即映射&封装方式和地址重写方式.映射&封装方式不需要引起主机和核心路由设施的改变,然而将增加报文的开销,可能加大网络延迟,有时甚至会因为传送数据过大而丢失数据.地址重写方式不会增加报文开销,但却需要对主机和网络做出更多的改变,并且可能丢失防火墙或其他安全系统所需要的信息.对未来的互联网来说,映射&封装方式是否真正优于地址重写方式还需要研究和评估.

(2) 映射系统的构建问题

引入映射系统已经成为许多路由架构的共识.映射系统的设计涉及到映射表的建立机制、映射表项的更新/撤销机制、映射表查询方法、映射表服务器的部署/管理等.映射系统应具有高扩展、安全性和低延迟等性质.映射系统的可扩展性是指映射表的规模不能太大且增长速度必须是可控的,映射数据的更新频率也必须是可控的.映射系统的安全性是指映射系统在受到攻击或出现故障的情况下,映射数据仍然真实可用,而路由系统仍然可以正常运行.映射系统的低延迟是指映射信息的查询时间应尽可能地短,以降低报文传送的延迟.如何设计和构建满足上述要求的映射系统极具研究价值,对提出的映射系统的性能评估也很重要.

(3) 可扩展性和流量工程之间的平衡问题

位置符/标识符分离架构依赖于位置符的积极聚合来缩减核心路由表和减少路由更新,从而达到提高路由可扩展性的目的.然而,流量工程以及路由策略要求使用更具体的地址前缀,过分的聚合可能使得网络无法有效

地开展流量工程,因此有必要对它们之间的平衡进行研究和评估.

最后,对新路由架构引入的利益关系,即谁付费、谁受益,还需要有更深入的理解.比如映射系统谁来部署、是否需要向用户收费等.这也是路由架构研究中必须考虑的因素.

References:

[1] Meyer D, Zhang L, Fall K. Report from the IAB workshop on routing and addressing. RFC 4984, 2007.

[2] Menth M, Hartmann M, Tran-Gia P, Klein D. Future Internet routing-motivation and design issues. Information Technology, Special Issue on Next Generation Internet, 2008,50(6):1−8.

[3] Wu JP, Wu Q, Xu K. Research and exploration of next-generation Internet architecture. Chinese Journal of Computers, 2008,31(9): 1537−1548 (in Chinese with English abstract).

[4] Narten T. Routing and addressing problem statement. Draft-narten-radir-problem-statement-02.txt, 2008.

[5] Huston G. BGP routing table analysis reports. 2009. http://bgp.potaroo.net/

[6] Tu R, Su JS, Peng W. Survey of naming and addressing architecture based on locator/identifier split. Journal of Computer Research and Development, 2009,46(11):1777−1786 (in Chinese with English abstract).

[7] Hou J, Liu YP, Gong ZH. Key techniques of identifier-based routing. Journal of Software, 2010,21(6):1326−1340 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3797.htm [doi: 10.3724/SP.J.1001.2010.03797]

[8] Fuller V, Li T, Yu J, Varadhan K. Classless inter-domain routing (CIDR): An address assignment and aggregation strategy. RFC 1519, 1993.

[9] Karpilovsky E, Rexford J. Using forgetful routing to control BGP table size. In: Proc. of the CoNEXT. Lisboa: ACM Press, 2006. http://www.cs.princeton.edu/~jrex/papers/forgetfulrouter.pdf

[10] Internet research task force. http://www.irtf.org/

[11] Kleinrock L, Kamoun F. Hierarchical routing for large networks: Performance evaluation and optimization. Computer Networks, 1977,1(3):155−174.

[12] Massey D, Wang L, Zhang B, Zhang L. A proposal for scalable Internet routing and addressing. Internet Draft, draft-wang-ietf-efit-00.txt, 2007.

[13] Wang N, Ma HL, Cheng DN, Wang BQ. Hidra: A hierarchical inter-domain routing architecture. Chinese Journal of Computers, 2009,32(3):377−390 (in Chinese with English abstract).

[14] Tsuchiya PF. The landmark hierarchy: A new hierarchy for routing in very large networks. ACM SIGCOMM Computer Communication Review, 1988,18(4):35−42. [doi: 10.1145/52325.52329]

[15] Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. In: Proc. of the ACM SIGCOMM. San Diego: ACM Press, 2001. 149−160. http://pdos.csail.mit.edu/papers/chord:sigcomm01/chord_sigcomm.pdf

[16] Plaxton CG, Rajaraman R, Richa AW. Accessing nearby copies of replicated objects in a distributed environment. In: Proc. of the 9th Annual ACM Symp. on Parallel Algorithms and Architectures. Newport: ACM Press, 1997. 311−320. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.1850&rep=rep1&type=pdf

[17] Caesar M, Castro M, Nightingale E, O'Shea G, Rowstron A. Virtual ring routing: network routing inspired by DHTs. In: Proc. of the ACM SIGCOMM. Pisa: ACM Press, 2006. 351−362. http://research.microsoft.com/en-us/um/people/antr/MS/vrr_sigcomm.pdf

[18] Kim C, Caesar M, Rexford J. Floodless in SEATTLE: A scalable Ethernet architecture for large enterprises. In: Proc. of the ACM SIGCOMM. Seattle: ACM Press, 2008. 3−14. http://www.cs.princeton.edu/~chkim/Research/SEATTLE/seattle.pdf

[19] Caesar M, Condie T, Kannan J, Lakshminarayanan K, Stoica I, Shenker S. ROFL: Routing on flat labels. In: Proc. of the ACM SIGCOMM. Pisa: ACM Press, 2006. 363−374. http://www.cs.uiuc.edu/~caesar/papers/rofl.pdf

[20] Cowen L. Compact routing with minimum stretch. Journal of Algorithms, 2001,38(1):170−183. [doi: 10.1006/jagm.2000.1134]

[21] Thorup M, Zwick U. Compact routing schemes. In: Proc. of the ACM Symp. on Parallel Algorithms and Architecture (SPAA). Heraklion: ACM Press, 2001. 1−10. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.4139&rep=rep1&type=pdf

[22] Arias M, Cowen L, Laing KA, Rajaraman R, Taka O. Compact routing with name independence. In: Proc. of the ACM Symp. on Parallel Algorithms and Architecture (SPAA). San Diego: ACM Press, 2003. 184−192. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.524&rep=rep1&type=pdf

[23] Abraham I, Gavoille C, Malkhi D, Nisan N, Thorup M. Compact name-independent routing with minimum stretch. In: Proc. of the ACM Symp. on Parallel Algorithms and Architecture (SPAA). Barcelona: ACM Press, 2004. 20−24. http://www.cs.huji.ac.il/~ittaia/papers/AGMNT04.pdf

[24] Gavoille C, Gengler M. Space-Efficiency for routing schemes of stretch Factor three. Journal of Parallel and Distributed Computing, 2001,61(5):679−687. [doi: 10.1006/jpdc.2000.1705]

[25] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology. ACM SIGCOMM Computer Communications Review, 1999,29(4):251−262. [doi: 10.1145/316194.316229]

[26] Dorogovtsev SN. Clustering of correlated networks. Physical Review E, 2004,69(027104).

[27] Zhou S, Mondragon RJ. Accurately modeling the Internet topology. Physical Review E, 2004,70(066108).

[28] Zhou S, Zhang GQ, Zhang GQ. Chinese Internet AS-level topology. IET Communications, 2007,1(2):209−214. [doi: 10.1049/iet-com:20060518]

[29] Zhang GQ, Zhang GQ, Yang QF, Cheng SQ, Zhou T. Evolution of the Internet and its cores. New Journal of Physics, 2008, 10(123027).

[30] Krioukov D, Fall K, Yang X. Compact routing on Internet-like graphs. In: Proc. of the IEEE INFOCOM 2004. Hong Kong: IEEE Press, 2004. 209−219. http://www.ieee-infocom.org/2004/Papers/05_4.PDF

[31] Brady A, Cowen L. Compact routing on power-law graphs with additive stretch. In: Proc. of the 8th Workshop on Algorithm Engineering and Experiments. SIAM, 2006. 119−128. http://www.siam.org/meetings/alenex06/

[32] Tang MD, Zhang GQ, Yang J, Zhang GQ. A compact routing scheme for scale-free networks, Journal of Software, 2010,21(7): 1732−1743 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3582.htm [doi: 10.3724/SP.J.1001.2010.03582]

[33] Enachescu M, Wang M, Goel A. Reducing maximum stretch in compact routing. In: Proc. of the IEEE INFOCOM. Phoenix: IEEE Press, 2008. 977−985. http://www-cs-students.stanford.edu/~mihaela/pdfs/cr_infocom.pdf

[34] Carmi S, Cohen R, Dolev D. Searching complex networks efficiently with minimal information. Europhysics Letters, 2006,74(6): 1102−1108. [doi: 10.1209/epl/i2006-10049-1]

[35] Abraham I, Malkhi D. Name independent routing for growth bounded networks. In: Proc. of the 17th Annual ACM Symp. on Parallelism in Algorithms and Architectures. 2005. 49−55. http://www.cs.huji.ac.il/~ittaia/papers/AM-SPAA05.pdf

[36] Lu H. Improved compact routing tables for planar networks via orderly spanning trees. In: Proc. of the 8th Int'l Conf. on Computing and Combinatorics. LNCS 2387, Berlin: Springer-Verlag, 2002. 57−66. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.1903&rep=rep1&type=pdf

[37] Flury R, Pemmaraju SV, Wattenhofer R. Greedy routing with bounded stretch. In: Proc. of the IEEE INFOCOM. Rio de Janerio: IEEE Press, 2009. 1737−1745. http://www.disco.ethz.ch/publications/infocom09routing.pdf

[38] Tang MD, Yang J, Zhang GQ. Compact routing on random power law graphs. In: Proc. of the IEEE Int'l Conf. on Dependable, Autonomic, and Secure Computing. Chengdu: IEEE Press, 2009. 575−578. http://doi.ieeecomputersociety.org/10.1109/DASC.2009.133

[39] Chen W, Sommer C, Teng SH, Wang Y. Compact routing in power-law graphs. In: Proc. of the DISC. Elche: Springer-Verlag, 2009. 379−391. http://research.microsoft.com/en-us/people/yajunw/disc2009sp.pdf

[40] Krioukov D, Claffy KC, Fall K, Brady A. On compact routing for the Internet. ACM SIGCOMM Computer Communication Review, 2007,37(7):43−52. [doi: 10.1145/1198255.1198262]

[41] Karp B, Kung HT. Gpsr: Greedy perimeter stateless routing for wireless networks. In: Proc. of the 6th Annual MOBICOM. Boston: ACM Press, 2000. 243−254. http://www.eecs.harvard.edu/~htk/publication/2000-mobi-karp-kung.pdf

[42] Finn G. Routing and addressing problems in large metropolitan-scale Internet-works. Technical Report, ISI/RR-87-180, ISI, University of Southern California, 1987.

[43] Francis P. Comparison of geographical and provider-rooted Internet addressing. Computer Networks and ISDN Systems, 1994, 27(3):437−448. [doi: 10.1016/0169-7552(94)90118-X]

[44] Hain T. An IPv6 provider-independent global unicast address format. Internet Draft, draft-hain-ipv6-pi-addr-10, 2006.

[45] Oliveira R, Lad M, Zhang B, Zhang L. Geographically informed inter-Domain routing. In: Proc. of the ICNP. Beijing: IEEE Press, 2007. 103−112. http://www.cs.ucla.edu/~lixia/papers/07SIGPoster_giro.pdf

[46] Cao Q, Abdelzaher T. A scalable logical coordinates framework for routing in wireless sensor networks. ACM Trans. on Sensor Networks, 2006,2(4):557−593. [doi: 10.1145/1218556.1218561]

[47] Fonseca R, Ratnasamy S, Zhao J, Ee CT, Culler D, Shenker S, Stoica I. Beacon vector routing: Scalable point to point routing in wireless sensornets. In: Proc. of the 2nd USENIX/ACM Symp. on Networked Systems Design and Implementation (NSDI). Boston: ACM Press, 2005. 329−342. http://berkeley.intel-research.net/sylvia/bvr.pdf

[48] Papadimitriou C, Ratajczak D. On a conjecture related to geometric routing. Theoretical Computer Science, 2005,344(1):3−14. [doi: 10.1016/j.tcs.2005.06.022]

[49] Kleinberg R. Geographic routing using hyperbolic space. In: Proc. of the IEEE INFOCOM. Anchorage: IEEE Press, 2007. 1902−1909. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.2848&rep=rep1&type=pdf

[50] Westphal C, Pei G. Scalable routing via greedy embedding. In: Proc. of the IEEE INFOCOM. Rio de Janerio: IEEE Press, 2009. 2826−2830. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.6328&rep=rep1&type=pdf

[51] Zhang GQ, Zhang GQ. Research on Internet correlation. Journal of Software, 2006,17(3):490−497 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/490.htm. [doi: 10.1360/jos170490]

[52] Zhang GQ, Zhang GQ. Exploring the local connectivity preference of the Internet AS-level topology. In: Proc. of the IEEE Int'l Conf. on Communications (ICC). Glasgow: IEEE Press, 2007. 6439−6445. http://sourcedb.ict.cas.cn/cn/ictthesis/200907/P020090722624457401017.pdf

[53] Krioukov D, Papadopoulos F, Boguna M, Vahdat A. Greedy forwarding in scale-free networks embedded in hyperbolic metric spaces. ACM SIGMETRICS Performance Evaluation Review, 2009,37(2):15−17. [doi: 10.1145/1639562.1639568]

[54] Tang MD, Zhang GQ, Yang J. Graph embedding based scalable routing in large networks. Journal of Computer Research and Development, 2010,47(7):1225−1233 (in Chinese with English abstract).

[55] Subramanian L, Caesar M, Ee CT, Handley M, Mao M, Shenker S, Stoica I. HLP: A next generation inter-domain routing protocol. SIGCOMM Computer Communication Review, 2005,35(4):13−24. [doi: 10.1145/1090191.1080095]

[56] Verkaik P, Broido A, Claffy KC, Gao R, Hyun Y, van der PR. Beyond CIDR aggregation. Technical Report, TR-2004-1, CAIDA, 2004.

[57] Castineyra I, Chiappa N, Steenstrup M. The nimrod routing architecture. RFC 1992, IETF, 1996.

[58] Zhang X, Francis P, Wang J, Yoshida K. Scaling IP routing with the core router-integrated overlay. In: Proc. of the IEEE Int'l Conf. on Network Protocols (ICNP). Santa Barbara: IEEE Press, 2006. 147−156. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.9844&rep=rep1&type=pdf

[59] Kastenholz F. ISLAY: A new routing and addressing architecture. Internet Draft, irtf-routing-islay-00.txt, 2002.

[60] Farinacci D, Fuller V, Oran D, Meyer D. Locator/ID separation protocol (LISP). Internet Draft, draft-farinacci-lisp-05.txt, 2007.

[61] Whittle R. Internet vastly improved plumbing architecture (Ivip). Internet Draft, draft-whittle-ivip-arch-01.txt, 2008.

[62] Adan JJ. Tunneled Inter-Domain Routing (TIDR). Internet Draft, draft-adan-idr-tidr-01.txt, 2006.

[63] Herrin W. Tunneling route reduction protocol (TRRP). 2008. http://bill.herrin.us/network/trrp.html

[64] Jen D, Meisel M, Massey D, Wang L, Zhang B, Zhang L. APT: A practical transit mapping service. Internet Draft, Draft-jen-apt-01.txt, 2007.

[65] Templin F. The IPvLX architecture. Internet Draft, draft-templin-ipvlx-08.txt, 2007.

[66] O'Dell M. GSE—An alternate addressing architecture for IPv6. Internet Draft, draft-ietf-ipngwg-gseaddr-00, 1997.

[67] Vogt C. Six/One: A solution for routing and addressing in IPv6. Internet Draft, draft-vogt-rrg-six-one-02, 2007.

[68] Vogt C. Six/One router: A scalable and backwards-compatible solution for provider-Independent addressing. In: Proc. of the ACM SIGCOMM MobiArch Workshop. Seattle: ACM Press, 2008. 13−18. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.729&rep=rep1&type=pdf

[69] Moskowitz R, Nikander P. Host identity protocol (HIP) architecture. RFC 4423, 2006.

[70] Nordmark E, Bagnulo M. Shim6: Level 3 multihoming shim protocol for IPv6. Internet Draft, draft-ietf-shim6-proto-09, 2007.

[71] Lear E. NERD: A not-so-novel EID to RLOC database. Internet Draft, draft-lear-lisp-nerd-04, 2008.

[72] Mathy L, Lancaster U. LISP-DHT: Towards a DHT to map identifiers onto locators. In: Proc. of the Re-Architecting the Internet Conf (ReArch 2008). Madrid: ACM Press, 2008. 7−12. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.139.735&rep=rep1&type=pdf

[73] Luo H, Qin Y, Zhang H. A DHT-based identifier-to-locator mapping approach for a scalable Internet. IEEE Trans. on Parallel and Distributed Systems, 2009,20(10):1790−1802. [doi: 10.1109/TPDS.2009.30]

[74] Brim S, Chiappa N, Farinacci D, Fuller V, Lewis D, Meyer D. LISP-CONS: A content distribution overlay network service for LISP. Internet Draft, draft-meyer-lisp-cons-04, 2008.

[75] Farinacci D, Fuller V, Meyer D. LISP alternative topology (LISP-ALT). Internet Draft, draft-fuller-lisp-alt-02, 2008.

[76] Yang X, Clark D, Berger AW. NIRA: A new inter-domain routing architecture. IEEE/ACM Trans. on Networking, 2007,15(4):775−788. [doi: 10.1109/TNET.2007.893888]

[77] Godfrey PB, Ganichev I, Shenker S, Stoica I. Pathlet routing. In: Proc. of the ACM SIGCOMM. Barcelona: ACM Press, 2009. 111−122. http://conferences.sigcomm.org/hotnets/2008/papers/17.pdf

附中文参考文献:

[3] 吴建平,吴茜,徐恪.下一代互联网体系结构基础研究及探索.计算机学报,2008,31(9):1537−1548.

[6] 涂睿,苏金树,彭伟.位置与标识分离的命名和寻址体系结构研究综述.计算机研究与发展,2009,46(11):1777−1786.

[7] 侯婕,刘亚萍,龚正虎.标识路由关键技术.软件学报,2010,21(6):1326−1340. http://www.jos.org.cn/1000-9825/3797.htm [doi: 10.3724/SP.J.1001.2010.03797]

[13] 王娜,马海龙,程东年,汪斌强.Hidra:一个分级域间路由架构.计算机学报,2009,32(3):377−390.

[32] 唐明董,张国清,杨景,张国强.针对无标度网络的紧凑路由方法.软件学报,2010,21(7):1732−1743. http://www.jos.org.cn/1000-9825/3582.htm [doi: 10.3724/SP.J.1001.2010.03582]

[51] 张国强,张国清.Internet 网络的关联性研究.软件学报,2006,17(3):490−497. http://www.jos.org.cn/1000-9825/17/490.htm [doi: 10.1360/jos170490]

[54] 唐明董,张国清,杨景.大规模网络上基于图嵌入的可扩展路由方法研究.计算机研究与发展,2010,47(7):1225−1233.

唐明董(1978－),男,湖南祁阳人,博士,CCF 高级会员,主要研究领域为计算机网络,路由.

杨景(1952－),男,研究员,博士生导师,主要研究领域为网络业务控制和集成体系结构.

张国清(1965－),男,博士,研究员,CCF 高级会员,主要研究领域为计算机网络,网络科学,网络融合.

张国强(1980－),男,博士,助理研究员,CCF 会员,主要研究领域为计算机网络.