软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(1):97−107 [doi: 10.3724/SP.J.1001.2012.04066]

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

双层 IP 地址空间体系结构*

钱华林[1], 鄂跃鹏[1,2+], 葛敬国[1], 任勇毛[1], 游军玲[1]

[1](中国科学院 计算机网络信息中心,北京 100190)

[2](中国科学院 研究生院,北京 100190)

Dual IP Address Spaces Architecture

QIAN Hua-Lin[1], E Yue-Peng[1,2+], GE Jing-Guo[1], REN Yong-Mao[1], YOU Jun-Ling[1]

[1](Computer Network Information Center, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100190, China)

+ Corresponding author: E-mail: eyp@cstnet.cn, http://www.cnic.cn

Qian HL, E YP, Ge JG, Ren YM, You JL. Dual IP address spaces architecture. *Journal of Software*, 2012, 23(1):97−107. http://www.jos.org.cn/1000-9825/4066.htm

Abstract: One of the challenges the current Internet faces is the scalability of routing system. The rapid growth of routing table and the more and more frequent updates of BGP are bringing heavy pressure on the performance, complexity, energy consumption and cost of core routers. In recent years, many researchers have been seeking solutions for these issues. One of the key research directions is splitting the current IP address into separate identifiers and locators. This paper proposes a novel ID/locator split solution, which forms dual IP address spaces architecture. It overcomes the difficulty of implementation and deployment. Also, it not only mitigates the routing scalability issue, but also solves the IPv4 address exhaustion problem. Besides simple modification on DNS and installation of a gateway for each user network, the original backbone network and user networks do not need any modification.

Key words: ID/locator split; architecture; address space; routing scalability

摘 要: 互联网面临的挑战之一就是路由系统的可扩展性.路由表的快速增长以及越来越频繁的 BGP 更新,对核心路由器的性能、复杂性、能耗和成本产生了越来越大的压力.近年来,大量网络研究人员正在针对这些问题寻找解决方案.将现有的 IP 地址分解为标识和位置的思想,是重要的研究方向之一.提出一种新的标识与位置分离方案,形成双层地址空间体系结构,克服了可实现性和可部署性的困难,在缓解路由系统扩展性难题的同时,解决了 IPv4 地址耗尽的问题.除了对 DNS 作简单的修改并增设一种网关设备外,原有的骨干网和用户网不作任何改动.

关键词: 标识与位置分离;体系结构;地址空间;路由扩展性

中图法分类号: TP393 **文献标识码**: A

互联网路由系统的可扩展性问题已被关注了很多年,但却一直没有好的解决方案.2006 年 10 月,互联网体

* 基金项目: 国家科技部支撑计划(2012BAH01B03)
收稿时间: 2011-02-24; 定稿时间: 2011-05-25; jos 在线出版时间: 2011-10-17
CNKI 网络优先出版: 2011-10-17 16:37, http://www.cnki.net/kcms/detail/11.2560.TP.20111017.1637.001.html

系结构理事会(Internet Architecture Board,简称 IAB)为此专门在荷兰阿姆斯特丹召开了一个由世界著名网络专家和网络公司参加的会议,讨论路由系统的可扩展性问题,随后又形成了一个标准文件(RFC 4984)[1].与会者一致认为,该问题已经是刻不容缓的了,一些可能的解决问题的方向也在会上提了出来.参会者的一个结论是,目前担负位置和标识功能的 IP 地址语义严重过载,这是造成路由系统可扩展问题的原因之一.因而把两者分离似乎是必须的,但如何从体系结构的角度实现位置与标识的分离,并没有满意的方案.

其实,现有的 IP 体系结构以及由此而引入的全局路由系统,不仅造成了路由表的急速增长,使得路由信息的交换、路由表的更新、路由表查询等难以满足高速数据包转发的要求,更为严重的是,路由系统的全局性使得网络中的每一个哪怕是微小的变动,例如信道的失效与恢复、设备或设备端口的失效与恢复、网络管理人员的一个误操作、一个不恰当的网络系统配置等,都要向全世界的网络通报并修改其路由信息库.随着网络规模的不断扩大,BGP 路由更新的频率越来越高,其处理强度几乎需要超级计算机才能胜任.

现有的 IP 地址体系结构由于标识和位置功能混为一体,不仅造成核心网络中路由表呈线性增长、IPv4 地址快速耗尽,而且还影响到网络或主机的移动、用户网络向上游 ISP 的多宿连接等问题的解决.

近年来,已经有大量的研究工作试图从标识和位置分离的角度来解决上述问题,但由于方案复杂,又引入了新的可扩展性问题,因而难以实现,更难以在现有的互联网上部署,从而引起了很大的争议.本文提出一种新的位置与标识分离方案,除了对域名解析系统(DNS)作少量修改并增设一种网关设备外,保持现有的骨干网络和用户网络的协议不作任何改动,从而获得很好的可部署性.

本文第 1 节讨论现有 IP 地址系统的缺陷.第 2 节介绍以往已有的相关工作,简单介绍它们的工作原理,分析其不易实现和部署的问题所在.第 3 节介绍新提出的双层 IP 地址空间体系结构及其工作原理.第 4 节分析该体系结构的特点和问题.第 5 节讨论网络和主机的移动性以及多宿连接.第 6 节研究新体系结构的部署和过渡.第 7 节总结并指出未来的工作方向.

1 IP 地址系统的缺陷

当前,互联网中 IP 地址系统最大的缺陷是其语义承载不当和承载过度.无论是 IP 信息空间还是人类社会空间,表达一个客观物体应包含 3 种信息:标识、位置、路由.标识表明是“谁”,位置表示“在哪里”,路由指示“怎么去”.在有结构的系统中,位置往往隐含了路由.例如,张三住在 *X* 国 *Y* 城市 *Z* 街 *W* 号,由于我们的社会是有结构的系统,*X* 国 *Y* 城市 *Z* 街 *W* 号不仅表达了张三在哪里,而且邮递员可以按照位置信息中隐含的路由信息逐步找到张三的住处.我们的电话网络也是一个有结构的系统,电话号码不仅表示了位置,而且信令系统可以利用电话号码结构中的国家号、地区号、区内终端号接通电话.电话号码的有结构,依赖于电话网络的有结构.在互联网中,网络拓扑是无结构的(各网络节点之间可以任意连接,不必遵守任何规则),无结构的网络拓扑必然导致位置信息(IP 地址)也是无结构的,而无结构的位置信息无法隐含路由信息.因此,对有结构的系统,只要有了标识和位置,寻址系统就是完备的,因为位置隐含了路由.对无结构系统,如互联网,除了标识和位置信息以外,还必须有一套单独的路由系统,实现路由信息的交换、更新、路由表查询、选路决策等功能,才能完成 IP 数据包的投递.目前的 IPv4 或 IPv6,其位置信息都没有隐含路由信息,从而导致其承载不当.

我们说目前的 IP 地址承载过度,是因为 IP 地址肩负了标识和位置双重功能.表面上看,主机或服务器可以有自己的域名,起到了标识的作用.但这是一种伪标识:它只是因为人不便记忆一串数字构成的 IP 地址,才专门设计了域名.计算机开始通信之前,首先将域名解析成了 IP 地址,随后的通信过程就完全与域名不再相关.撇开人的因素,就计算机而言,IP 地址兼任了位置与标识的功能.

一方面,IP 地址没有(也无法)隐含路由信息;另一方面,IP 地址代替了主机的标识.前一个缺陷导致互联网离不开一套复杂的、依赖全局网络行为的路由系统,带来的问题是缺乏可扩展性(scalability).随着网络规模的不断扩大,路由系统产生了一系列致命性的问题,例如,路由表越来越大、BGP 的收敛速度越来越慢、数据包路径不确定导致 QoS 无法实现、网络行为不确定导致网络监控、管理和网络安全难以做好、路由表处理和查询速度慢导致路由器设备复杂昂贵并且耗电量大.后一个缺陷相当于用“*X* 国 *Y* 城市 *Z* 街 *W* 号”来标识“张三”这个人,

导致这个人一旦搬家,就没有人认识他了.带来的问题是缺乏可移动性(mobility)和多宿连接(multi-homing)的能力.显而易见,互联网的可扩展性、可移动性和多宿连接等问题无法解决,根本原因就是 IP 地址承载不当和语义过载.

RFC 4984 指出,“IP 地址语义的位置与标识过载,是造成路由系统可扩展性问题的原因之一.因而,把它们分离对解决可扩展性问题似乎是必须的”.

2 相关工作

试图将 IP 地址分解为标识与位置两个部分的工作已经进行了多年,早在 1997 年,O’Dell 等人提出了一种新的编制方案[2],将网络划分成两部分:骨干网和用户网.骨干网是可穿越网(transit),把来自一个用户网的数据包送到另一个用户网;用户网是不提供穿越的末端网络.相应地,把 IPv6 地址也分成两部分:高 8 字节用于路由(称为 Routing Goop,简称 RG),表达骨干网连接用户网的边缘处的地址(locator);低 8 字节表示用户网络中主机的标识(identifier).该方案被称为 8+8 方案.其核心思想是,让 IP 地址的路由部分用于网络层协议,而标识部分用于传输层及以上的协议.由于高层协议只使用相对固定的标识而与接入点位置(由 RG 表达)无关,用户网从一个骨干网接入点迁移到另一个接入点时不必重新编号,当用户网主机发出数据包时,把源地址的 RG 部分置成空,数据包到达骨干网之前,由用户网的出口网关添上 RG.这样,用户网设备不必关心上连的 RG 地址,就有了可移动的能力.为了在骨干网中进一步增加汇聚能力从而减少路由表项,该方案还建议将骨干网划分成多个结构和子结构,构成一个层次型的网络.它的核心思想包括诸如把网络按骨干网和用户网划分、把 Locator 与 Identifier 分离、把骨干网结构化等,给后来的研究提供了很多有益的思路.

LISP(locator identifier separation protocol)协议[3]也是利用位置与标识分离来缓解骨干网中路由系统的可扩展性困难.骨干网与用户网的连接处,其设备端口的 IPv4 地址称为 Locator,用户网络内部的 IPv4 地址称为 Identifier.当用户主机发起通信时,通过 DNS 查得对方的 Identifier.如果通信目的地不在本用户网,则被缺省地送到骨干网,骨干网边缘设备用一对 Locator IPv4 地址对数据包进行隧道封装后,经骨干网送到目的地边缘设备,拆除封装后,交给目的地用户网络,利用目的地的标识 IPv4 地址将数据包送达目的地主机.问题的关键在于,源端的骨干网边缘设备如何获得封装信息.封装包头的源 Locator 地址就在本地,容易得到,但目的地的 Locator 就无法简单地获得了.这时,要进行一次从目的 Identifier 到目的 Locator 的映射.如下两个问题使这种 LISP 方法难以简单地实现和部署:一个困难是当骨干网源边缘设备发起映射查询时,从源主机发来的后续数据包可能源源不断地进入该边缘设备,需要提供足够的缓冲,甚至有丢失数据包的可能;另一个困难是到哪里去查询映射数据库.如果映射数据库集中在骨干网的某处或某几处,则肯定会造成瓶颈.如果把一个汇总的数据库分布地存放在所有的骨干网边缘路由器,则不但存储负担很重,且当某个映射关系变更时,还要通知到所有边缘路由器进行更新.任何局部的映射关系变更都需要全局性的更新,其负载不比目前的 BGP 协议轻.为此,很多文献对这个映射系统(mapping system)进行了研究,提出了各种不同的解决方案[4−7].其中,较受重视的一种是另建一个逻辑拓扑,以覆盖网的方式架构在现有的骨干网上,利用 BGP 协议传递和汇聚标识前缀的可达性信息以及执行映射的几个特殊控制包[3].但这种方法不仅使得映射系统复杂化,还产生易遭受攻击的安全问题[8].

文献[9]提出了一种新的也是基于标识和位置分离的体系结构 ILNP(identifier-locator network protocol),其主要目的是解决移动性和多宿连接等问题.该方案把地址结构分为 L:I 两部分,I 是与传输层及以上层有关的、通常不大变动的标识(identifier),L 是只与网络层路由有关的、可以在需要时变动的位置地址(locator).Locator 定位了一个子网也就是用户网,I 在同一个用户网络中唯一.同时,在 DNS 中增加了 L 记录和 I 记录,当一台或一批正在通信的主机移动到另一个接入点时:一方面利用 DNS 动态更新技术[10]更新 L 记录,新来的通信采用新的 L 地址;另一方面,移动的主机可以发一个更新 L 信息的控制包(ICMP 包),将移动后获知的新的 L 信息告诉通信对方,确保通信不被移动所中断.虽然该方法不要求骨干网的协议有任何改动,但对用户网的改动要求很多:要改动用户主机的操作系统内核、改变 socket 的调用方式、增加向通信对方通知新 Locator 值的控制报文及相应的协议、增加对 DNS 中 L 记录的动态更新、增加向边缘设备获知新 Locator 值的功能等.这些改变都要求每

台用户计算机的操作系统、网络协议等作大量的改动,因而并不容易部署.

此外,文献[11]提出了一种利用公开密钥体制中 Public Key 衍生出来的两个定长 HASH 值来标识网络主机的主机标识协议(host identity protocol,简称 HIP),利用大的密钥空间和 HASH 空间使标识全局唯一,其主要目的是解决主机身份认证和移动性.但要求对使用这种协议的主机操作系统和网络协议层次进行大量修改,与文献[9]要求的修改几乎没有区别.为了报告移动后新的位置信息,还要建立一个集中登记和查询的机制(rendezvous mechanism).尽管如此,要保证移动过程中不中断通信,还可能出现因处理时过大的延迟而得不到满足的情况.

3 双层 IP 地址空间体系结构

双层 IP 地址空间体系结构(dual IP space architecture,简称 DIPSA)沿用了 LISP 等方案提出的划分骨干网和用户网,并在边缘设备中进行隧道封装等概念,结合 DNS 和映射系统,解决骨干网路由系统扩展性和 IPv4 地址耗尽等问题.

虽然本方案既适用于全 IPv4,也适用于全 IPv6,甚至是用户网 IPv4/骨干网 IPv6 或用户网 IPv6/骨干网 IPv4 等不同情况,但考虑到从 IPv4 向 IPv6 过渡的艰难性,我们主要针对 IPv4 环境来考虑问题.如果能在缓解路由系统扩展性问题的同时还能较为容易地解决 IPv4 地址耗尽问题,那么,就可以缓解向 IPv6 过渡的困难.

目前的过渡方案可分为两类:隧道类,如 6rd,ds-lite,A+P;翻译类,如 NAT64,IVI 和 PNAT 等.[12]在 IPv4 应用占主导地位的情况下,隧道类方案仍然需要为用户提供 IPv4 地址,如果运营商没有足够的 IPv4 地址,那么他只能为用户提供私有地址使用 NAT(网络地址翻译)与公网地址通信;翻译类方案则可以只使用 IPv6 地址,与 IPv4 的通信通过翻译进行.在翻译类方案中,也存在 NAT 的应用层网关问题.本文描述的方法能够提供比 IPv4 更大的地址空间,同时能够实现比 NAT 更简单的双向访问,可用于替换过渡方案中使用 NAT 的 IPv4 部分.

3.1 基本概念

虽然本方案利用了很多相关工作中提出的概念,但存在很大差别.为便于理解,在图 1 中给出了一个骨干网和两个用户网的示例,骨干网的两个边缘端口的 IPv4 地址分别用 Loc-A 和 Loc-B 表示,两用户网中各设立一个网关 RMT-A 和 RMT-B,用户网的主机 IP 地址分别用 EID-A 和 EID-B 来标识.

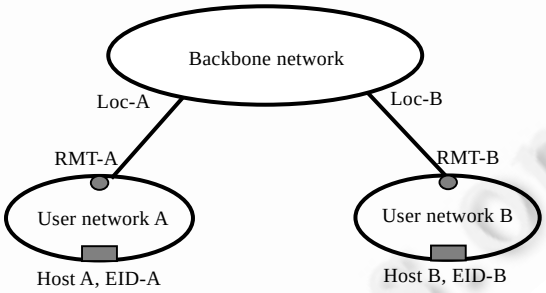

Fig.1 Network architecture illustration

图 1 网络结构示意图

结合图 1,我们对本方案中用到的一些基本概念作如下说明:

(1) 在骨干网和用户网的连接处,设立一个特殊的网关设备.该网关设备执行 3 个主要功能:1) 域名解析器(即 DNS 系统中的递归服务器);2) 映射系统;3) 隧道封装与拆封.为了方便起见,称这个特殊的网关为 RMT 网关,3 个字母分别为 3 项功能(resolver,mapping,tunneling)的英文字头.与 LISP 不同,这个网关属于用户网络而不是骨干网络.

(2) 在 DNS 的 zone file 中,除了现有的 A 记录外,增加一类新的资源记录(假设叫做 B 记录).B 记录与 A 记录一样,属于 IP 地址类型.A 记录表达用户网络中通信设备的标识(即图中的 EID).为了简单地获得标识在用户网络中的可路由性,并且不改变现有用户网络的任何协议,与 LISP 一样,我们用 IP 地址作为标识.因此,A 记录就

是现有 DNS 中的 A 记录.B 记录表达骨干网中相应的 Locator IP 地址.解析域名时,权威服务器向递归服务器同时返回 A 记录和 B 记录.在一个 DNS 注册项中,可以有一个或多个 B 记录,与用户网向上游 ISP 的一个或多个连接(multi-homing)相对应.

(3) 用户网络设备配置的递归域名服务器,指向本用户网络的 RMT 网关.发起通信的用户主机首先要解析通信对方的域名,由 RMT 负责递归解析,并将结果返回给用户主机.一个重要的特点是,当 RMT 向用户主机返回解析结果时,扣留 B 记录.用户主机得到的解析结果与现有的 IPv4 网络没有任何差别,这样,用户网络里的任何设备既不知道 RMT 的存在,也不知道 DNS 系统发生了什么变化.这个特性保证了用户网络在新的体系结构中无需作任何改动,也为用户网络的移动和多宿连接提供了基础.

(4) RMT 网关执行映射功能,将用户主机请求解析后获得的解析结果中的目的地 A 记录、目的地 B 记录以及发起域名查询的源主机的 IPv4 地址等信息,保存到一张映射表(映射数据库)中.RMT 负责这个映射表的动态生成、查询以及更新,形成一个映射系统.与 LISP 的映射系统不同:首先,发起映射的动作、获得映射关系的数据是由作为端设备的用户主机启动的,不会产生 LISP 中由中间设备(骨干网边缘路由器)发起映射造成后续数据包堆积的问题;其次,映射数据库的内容是局部性的(只与本用户网络中正在通信的设备有关),而不像 LISP 那样是全局性的,这使得该映射数据库具有局部性、规模小、易维护、系统结构简单、查询速度快等特点.映射表项如图 2 所示.

(5) 每个用户网络独立地占用整个 IPv4 地址空间,其中:少部分地址作为私有地址空间(private IP,简称 PIP),不能分配给网络设备;其余地址均可分配.PIP 可以占用当前尚未分配的地址,例如,可在 IANA 或 IETF 保留的空间中划出一个/8 就足够了.

(6) 送出用户网络的 IPv4 数据包,经过 RMT 时封装一个 IPv4 包头.其目的地址是映射系统中保存的、对方的 B 记录包含的 IPv4 地址,源地址是自己上连的骨干网的边缘端口的 IPv4 地址.这两个地址都处于骨干网的 IPv4 地址空间,骨干网可以独立地占有一个完整的 IPv4 地址空间.为了让 RMT 网关更简便地判定这个封装包头,将包头中协议类型字段的值置成一个 IANA 分配的数值,称为 DIPSA 标志.

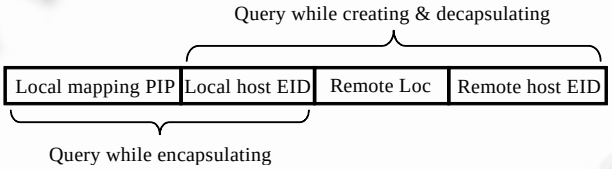

Fig.2 Mapping item

图 2 映射表项

3.2 通信过程

设用户网 A 中主机 A 发起对用户网 B 中主机 B 的通信,通信过程如图 3 所示,具体描述如下:

1. 主机 A 经 RMT-A 查询主机 B 的域名;
2. 递归服务器 RMT-A 获得主机 B 的 A 记录值 EID-B 和 B 记录值 Loc-B.RMT-A 进行如下操作:
   (1) 比较 Loc-B 和 Loc-A,若 Loc-B=Loc-A,则表明通信双方处于同一用户网,RMT-A 直接将 EID-B 返回给主机 A,双方在本域中通信,不涉及骨干网,其通信过程不再描述;
   (2) 若 Loc-B≠Loc-A,则表明通信双方不在同一用户网,需经骨干网转送数据包.RMT-A 从 PIP 地址池中取得一个私有地址 PIP-A,将 PIP-A 作为目的地址返回给主机 A;
   (3) 同时将 Loc-B,EID-B,PIP-A 和 EID-A 等参数存入映射数据库;
3. 主机 A 发出正常的 IPv4 数据包(称为原始数据包),其目的地址为 PIP-A,源地址为 EID-A.由于数据包目的地址 PIP-A 不在本用户网可路由地址范围内,数据包被缺省地向上游骨干网方向传送,必定经过 RMT-A;

4. RMT-A 收到原始数据包后,做如下操作:
   (1) 查询映射数据库(通常在 Cache 中有),获得 EID-B 和 Loc-B;
   (2) 用 EID-B 替换原始包头的目的地址 PIP-A;
   (3) 用 Loc-B 和 Loc-A 分别作为封装包头的目的地址和源地址进行隧道封装.封装后的数据包送交骨干网边缘路由器;
5. 骨干网利用封装包头的 Loc-B 和 Loc-A,将数据包送达 RMT-B;
6. RMT-B 收到数据包,进行如下操作:
   (1) 向私有地址池申请一个地址 PIP-B;
   (2) 将 PIP-B,EID-B,EID-A,Loc-A 等参数保存到映射库及 Cache 中;
   (3) 拆去封装包头,用 PIP-B 替换原始包头的 EID-A.将原始数据包传递给主机 B.

从主机 B 返回主机 A 的数据包,将收到数据包中的源地址和目的地址交换,即用 PIP-B 和 EID-B 分别作为目的地址和源地址组装数据包,其余过程同上.

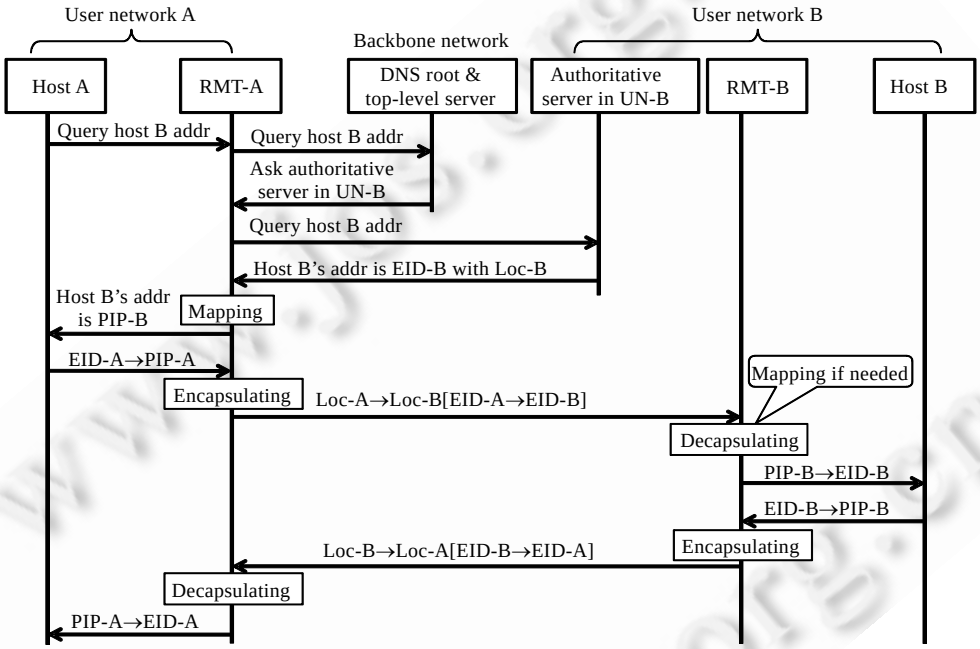

Fig.3 Time sequence of DIPSA

图 3 通信过程时序图

4 系统特点与问题分析

DIPSA 有一些吸引人的特点:

首先,原有网络路由系统的扩展性问题主要发生在非缺省区(default free zone,简称 DFZ).DIPSA 骨干网中用一个 Locator 地址代替了一个用户网络的地址前缀,而一个用户网络可能有多个地址前缀,这样就能减少骨干网路由表的规模.DIPSA 的骨干网如果进一步采用层次式交换技术替代路由技术[13,14],则能彻底消除路由可扩展性问题和因路由系统而带来的一系列重大问题.

其次,该方案使得每个用户网络拥有几乎全部 IPv4 地址空间,这就获得了地址空间无限扩展的能力.例如,当一个用户网地址空间不够用时(可能性极小),可以再开一个用户网.未来互联网向物联网、传感器网、家庭网、个人网等延伸时,地址空间的需求都在向下层扩展,DIPSA 让最下层的用户网络拥有丰富地址空间的思想,比

IPv6 的地址安排更合理、地址也更丰富.骨干网的地址空间需求增长相对缓慢,即使一个完整的 IPv4 空间仍然不够用,也可以很容易地转为 IPv6.向 IPv6 过渡的难度主要在用户网,因为所有的应用都在这里,而仅将骨干网过渡到 IPv6,现在的大部分核心设备已经就绪了.NAT 最多可以使一个用户网拥有一个 A 类地址(10.0.0.0/8),但对于从用户网外访问用户网内主机的情况,必须借助 UPnP 或 STUN 等机制才能实现,需要应用程序予以支持,这种情况下 NAT 不是透明的.

第三,一个新方案能否可行,最主要的问题是对现有系统的改动要求尽量少.DIPSA 对骨干网和用户网都不要求改动,这就使得新系统对现有的网络协议和网络应用完全透明,所有的功能都由一个新增加的网关 RMT 和在 DNS 中增加一个新记录来实现.

第四,由于用户网络的地址空间不再依赖上层 ISP,并且不知道上层接入骨干网的 Locator 信息,与现有网络相比,用户网络改变接入地点或改换上层 ISP 非常容易,只要改动 DNS 的注册信息就可以了,避免了地址重新编号的麻烦.如果一个用户网络里的一个或一部分设备改接到另一个用户网中去,由于原先两个用户网的地址空间是相互独立的,因此有可能在目的用户网中发生新来设备与原有设备产生地址冲突的可能.这种情况发生的可能性是极小的,因为属于不同单位的设备永久地转给另一个单位,是不大可能发生的.如果是临时地移动到一个新的用户网去(例如出差、开会),则可通过无线的 WiFi 利用 DHCP 等方法临时获得 IP 地址.总之,非实时的移动是没有问题的;对在线的实时性移动,将在第 5.1 节中加以讨论.

第五,前面在描述通信过程时,设立了私有 PIP 地址池,为每个 session 申请一个唯一的 PIP 地址(例如 PIP-A, PIP-B 等).实际上,PIP 地址的主要功能是让发出数据包的用户网络感到数据包目的地不在本网,让接收数据包的用户网络感到数据包并不来自本网.另外,PIP 地址的一个附加功能是可用作查询映射表的索引(index).显然,其主要功能并不要求每个 session 的 PIP 地址是唯一的.当对发出的包进行封装时,需要使用源 EID 和 PIP 地址的联合索引,如图 2 所示,即同一用户网中不同的主机通信时可以采用相同的 PIP 地址,同一主机与多个远程主机通信时,要采用不同的 PIP 地址.这时,PIP 地址池的大小并不取决于一个用户网络中有多少同时对外通信的主机,而是与一台服务器同时能建立的 session 数量有关.

下面我们来分析 DIPSA 可能遇到的问题,这些分析不仅可以评估 DIPSA 的可行性,还可以为其工程实现提供参考:

首先,若对 DNS 中权威服务器的改动只是增加一个记录,则不会有什么问题.但对 RMT 网关中的递归解析器要作一些改动:递归解析得到结果后,要进行目的 Locator 与本地 Locator 的比较;还可能需要扣留对方 A 记录的内容,而将一个私有地址交给用户主机;要将解析得到的信息传递给映射系统处理.需要增加的新操作并不复杂,但毕竟要对源代码作少量的改动.由于这些改动只在网关 RMT 中执行,因而容易实现.

其次,除 DNS 外,唯一新的东西就是 RMT 网关.该网关的功能有点像当前广泛应用的 NAT 设备[15].与 NAT 设备相同的是它们都只为一个用户网络服务,因而负担不会太重;与 NAT 设备不同的是,NAT 的穿透能力较弱[16],当外界网络用户发起对私有地址空间用户的通信时,需要另外的代理服务器协助才有可能,并且只适合某些应用,而 DIPSA 在两个通信方向上都是简单而通畅的.尽管在两个用户网边界的 RMT 分别对原始包头进行了目的地址和源地址的重写,并把 PIP-A 和 PIP-B 称为“私有地址”,而在本质上,两个用户网只是把这些地址看作缺省路由用的“外网”,与 NAT 的地址翻译是完全不同的.NAT 的公用地址池占用一批公用地址,DIPSA 只占用一个或几个骨干网的 Locator 地址.DIPSA 的封装功能操作简单,并不占用 RMT 网关多少 CPU 和存储资源.

第三,DIPSA 的隧道封装增加了 20 个字节,与 LISP 相比,超过 MTU 规定的概率较小,对少数长度超过 MTU 的包,出口 RMT 负责将其分片,与普通的路由器对于长度超过 MTU 的包的操作相同.

第四,DIPSA 的映射功能是唯一可能出现瓶颈因而须特别关注的地方.DIPSA 映射功能分为 3 个环节:映射表项入库、映射表项查询、映射表项删除.映射表项的生成和入库,每次通信过程只有 1 次,且发生在域名解析阶段,与域名解析相比,其处理延迟几乎可以忽略,对 RMT 资源的占用也不多.映射表的查询速度也不会出现问题,因为整个映射数据库仅涉及一个用户网络中正在对外界通信的各 session,其数量有限,且其建立映射和解封装时的索引长度为 96bit(如图 2 所示),与对称 NAPT(网络地址端口翻译)的索引长度近似,可以利用哈希表等方

法加快查询速度.映射表项的删除针对于已经结束的通信,用来释放映射表项和所占用的私有地址.如何判定一个映射表项已经不再活动因而可以删除了,可以参考 NAT 等使用的方法.

第五,最容易出现问题的地方是由于 IP 地址的改写,使得 TCP 和 UDP 头部的伪校验和发生不一致.虽然接收方的RMT对原始包实施源地址重写后可以重新计算并改写TCP/UDP的头部伪校验和,使得数据包可被正确地接收,但当采用 IPsec 等协议时,RMT 无法改写 TCP/UDP 头部校验和.这是所有采用地址改写技术的方案(包括 IPv4 NAT 以及向 IPv6 过渡中的 IPv4/IPv6 翻译)面对的共同问题,只能从 IPsec 等协议方面加以变通才可能解决.另外,与 NAT 类似,对 Payload 中含有 IP 地址信息的应用,必须配备相应的应用层网关.由于域名解析时获得的 IP 地址与 PIP 地址无关,NAT 中必不可少的 ALG-DNS 网关在这里是不需要的.

第六,对于某些直接使用 IP 地址进行通信的应用,由于不通过 RMT 解析域名,RMT 没有映射信息,就可能产生问题.分几种情形:用户网内部相互通信,没有问题;对外界网络用户通信时,RMT 因查询映射表失败而将数据包丢弃;对外界众所周知地址的访问(例如对 DNS 根服务器的访问),只有通过 RMT 才能进行,其他访问行为将被 RMT 阻止.

5 网络与主机的移动性和多宿连接

DIPSA 的主要作用是解决核心网络中的路由系统扩展性和 IPv4 地址空间耗尽两大难题.由于用户网络不知道自己接入骨干网的 Locator 位置地址信息,因而为用户网络的迁移(mobility)并避免重新编号(renumbering)提供了可能性.

5.1 网络与主机的移动性

对于非实时的静态移动,只涉及重编号的问题.前面已经讨论过,由于用户网络中的设备并不知道自己上连的 Locator 地址信息,用户网络静态地从一个 ISP 接入点改接到一个新的 ISP 接入点,仍然占有一个独立的 IPv4 地址空间,不存在重编号问题.对一个用户网络与另一个用户网络合并成一个更大用户网的情形,如果这两部分网络原本属于同一个用户单位,则合并前的两个用户网通常不会分配相互冲突的地址,需要重编号的可能性较小;如果两个用户网原先属于不同的管理单位,则合并后有可能发生地址冲突,需要把地址有冲突的设备重编号.这种重编号对任何使用 IP 地址作为主机标识的网络体系结构都是无可避免的,即使不用 IP 地址作为标识,如果允许不同用户网中独立地自选标识,也可能有标识冲突而要求合并后的部分设备重新选用标识,这与 IP 地址的重编号本质上是一样的.显然,要求设计的标识系统具有严格的全局唯一性,并非易事.

对于计算机的实时移动,要求无线连接的设备移到新的用户网时,原先正在进行的通信不被中断,难度就很大.文献[9]给出了一种解决方案,由于将 MAC 地址用作标识,并且假设标识是全局性的,移动到新域后标识不变,因而 TCP 层用同样的标识计算伪包头校验和结果不变,端到端的连接得以继续.但仍然要求动态修改 DNS 记录,以便新的通信呼叫使用新的 Locator 地址;提供移动计算机查询新 Locator 的协议;新设若干 ICMP 控制报文,将自己新的 Locator 通知对方,以便双方在 IP 报文中使用新的 Locator 地址.所有这些操作都应当足够快,以保证端到端的通信连接不会因超时而断开.这些对现有系统的改动,对实际部署是十分不利的.在 DIPSA 方案中,如果移动到新域后,主机的 IP 地址可以不变,以保证 TCP 层的伪包头校验结果不变,则可以借用文献[9]的方法.当然,同样存在改动大、部署困难的问题.可以说,实时移动是至今仍未解决的问题.

5.2 多宿连接

为了增加上连的可靠性或者获得更好的 ISP 服务,用户网络需要上连到不同的 ISP 或同一 ISP 的不同接入点.在 DIPSA 系统中,每一个上连接入点对应一个 RMT 网关.虽然允许每一个用户网络设置的网关数量没有限制,但过多的网关意义不大而且管理和操作变得复杂.我们以两个网关(RMT-1 和 RMT-2)为例加以研究.如果让用户网络中的一部分主机使用 RMT-1,另一部分主机使用 RMT-2,则通信过程与单个网关的情形没有差别.当一个网关(如 RMT-1)对应的上连信道出现故障时,受故障影响的那部分主机将自己的 DNS 解析器配置改为另一个网关(RMT-2),虽然不能做到实时切换,但也只需几分钟时间就可以保持网络的连通性了.

如果要求实时地切换,就不能让用户参与修改配置的工作.这就要求两个网关之间保持联络,上连信道失败的 RMT-1,继续为自己的用户主机服务,只是把接到的用户请求转交给 RMT-2,由 RMT-2 完成域名解析、地址映射和隧道封装等功能.这种上连接入点的动态、实时切换对用户网络中的设备是透明的,但由于网络中有两个出口点,且用户网络中的缺省出口不唯一,因而要在路由配置中设定.

如果用户网络只用一个 RMT 网关设备向上连到多个骨干网边缘端口,同样也能获得多宿连接的能力,但操作却简单得多,不仅避免了两个网关之间的通信联络,还由于用户网络中只有一个网关而使得出口的缺省路由选择变得非常简单.两条上连信道上的负载分配、负载平衡或故障时的负载切换都在这个网关内部进行,无论是负载分配策略的设定,还是负载切换的操作以及对网络管理员的技能要求,都很简单.而要防止单个网关本身的单点失效,可采用双机热备份的任何一种已有技术来解决.

为了简化多宿信道上均衡数据流的操作,可对多个不同连接分配不同的优先数值.例如,希望在两条带宽相同的上连信道上均衡分配负载,可把两个连接的优先数都置成 50.如果一条信道的容量为另一条的两倍,则分别置成 67 和 33.在保证每个用户数据流走相同上连信道的前提下,由 RMT 网关将不同的通信 session 分配到不同的上连信道,可以获得大致的平衡.

6 部署与过渡

任何一种新的结构或功能,其部署过程对现有系统的影响决定了它是否可行.部署 DIPSA 的过程就是从部分 DIPSA 向全部 DIPSA 过渡的过程,这就要求 DIPSA 用户网络能够顺畅地与非 DIPSA(为叙述方便,把现有的非 DIPSA 网络称为 SIPSA(single IP space architecture)网络)用户网络通信.为此,要研究过渡过程中对 DIPSA 用户网络的某些限制,以便适应 SIPSA 用户网络.

首先要考虑的是地址冲突问题.SIPSA 网络中只有一个 IPv4 地址空间,具有全局唯一性,如果 DIPSA 网络独立地使用全部 IPv4 地址空间(私有地址 PIP 除外),则由于 SIPSA 网络缺少 RMT 网关的隔离而造成地址冲突.为此,要求过渡过程中 DIPSA 网络仍然使用现有的全局唯一分配的地址.

其次是封装问题.DIPSA 采用了包头封装,而 SIPSA 没有封装包头.SIPSA 网络中没有地方能够识别封装并拆除封装,也就不能正确处理包头信息.为此,要求 DIPSA 网络与 SIPSA 网络通信时不做封装,因此也就不使用 B 记录、映射等功能.

通信关系有 3 种:DIPSA 与 DIPSA 通信;SIPSA 与 SIPSA 通信;DIPSA 与 SIPSA 通信.其中:DIPSA 与 DIPSA 通信的过程前面已经讨论过;而 SIPSA 与 SIPSA 通信是现有网络的通信,也无需讨论;因而只需讨论 DIPSA 与 SIPSA 之间如何通信的问题.为了保证现有网络不作任何变动,所有的限制性要求都应加在 DIPSA 网络一侧的 RMT 中.

DIPSA 与 SIPSA 通信又分两种情形:一种是 DIPSA 网络用户发起的对 SIPSA 用户的通信,另一种是 SIPSA 用户对 DIPSA 用户发起的通信.为讨论方便,我们假设主机 EID-A 处在 DIPSA 用户网络中,主机 EID-B 处于 SIPSA 用户网络中.

6.1 DIPSA对SIPSA发起通信

在 DIPSA 网络用户 EID-A 发起通信时,第 1 步工作是经 RMT 网关发起 DNS 域名解析,RMT 网关发现对方的域名记录中不包含 B 记录,就可断定对方是 SIPSA 用户.这时,DIPSA 的 RMT 网关不执行映射和封装等功能,只是简单地将对方的 A 记录如实地返回给自己的用户,并在映射表中简单地记录双方的 IPv4 地址(EID-A,EID-B).B 记录(Loc-B)和本地 PIP(PIP-A)部分设置成空(null).对来自 EID-A 的外发数据包,照样查询映射表,发现 Null 项后,将数据包原封不动地送给骨干网.对来自骨干网的到达数据包,也是在 Null 信息的指引下原封不动地传给本地用户 EID-A.其实现思想是把 DIPSA 降格为 SIPSA 来运行,RMT 网关对此的操作也很简单.

6.2 SIPSA对DIPSA发起通信

SIPSA 用户发起对 DIPSA 用户的通信,因为 EID-B 主机并没有 RMT 网关帮助并判断对方的 B 记录,当

EID-B 使用的解析器将对方的 A,B 记录都返回给 EID-B 时,它不认识 B 记录,不作任何处理,只用对方的 A 记录来构造发出的数据包.当该数据包到达 DIPSA 网络的 RMT 时,RMT 必须判断发起通信的对方是 DIPSA 还是 SIPSA 网络.

最简单的判断方法是对第 1 个到达包(映射表中没有相应表项,表明是对方发起的一次新的通信)进行包头分析,如果发现收到包的 IPv4 包头的协议类型字段为 DIPSA 标志,就按 DIPSA-DIPSA 方式通信;否则,就按 DIPSA-SIPSA 方式通信(在映射表中置对方的 B 记录和本地的 PIP 为 null,实现与第 6.1 节一样的 DIPSA 与 SIPSA 通信).另一种判断方法是,RMT 网关利用对方的 IP 地址在 IN-ADDR.ARPR 顶级域名中进行反向 DNS 解析,获得对方的域名,然后再作一次正向域名解析,看对方是否有 B 记录.进行两次解析会引入延迟,并且很多主机并不注册域名,所以前一种判断方法更简单、可靠.

6.3 域名服务器与IP地址直接访问

前面介绍的通信过程都要通过域名解析判断通信对方是否存在 B 记录以及获取通信对方 B 记录值等信息,由于每一级服务器都在它的直接上级服务器中注册,各级域名服务器的地址可以在递归解析过程中由上级服务器给出.根服务器是最高一级,它不能在更高一级的服务器中注册,其地址不能通过解析系统获得,所以只能使用一些固定的地址.这些地址被称为众所周知(well-known)的地址.由于这些众所周知的地址没有考虑到 B 记录,根服务器只能占用骨干网地址空间,对其访问时,直接用 IP 地址.其他各级域名服务器都在上级服务器注册,能标明 B 记录信息,过渡阶段中,它们可以处于骨干网地址空间中,但从长远看,应该都放在用户地址空间中.对处于骨干网地址空间的根以外的所有域名服务器的访问,因为不存在 B 记录,因此只能执行 DIPSA-SIPSA 通信方式.如果域名服务器处于受 RMT 隔离的用户空间,上级服务器给出的地址资源记录中有 B 记录,则可用 DIPSA-DIPSA 的方式通信.要求服务器在引用下级服务器时,将 B 记录(如果有的话)与 A 记录一起返回给递归服务器.

任何用户网络中的设备,过渡阶段中不论是否已安装了 RMT 网关,由于地址是全球唯一的,都可以不经域名解析直接用 IP 地址进行通信.完成过渡后,用户网独占 IPv4 地址空间,必须先通过域名解析获得 IP 地址后才能通信.对目前的大量“客户-服务器”通信方式,总是从客户向服务器发起呼叫,先解析域名是很自然的事;但对 Peer-to-Peer 方式的通信,例如网络电话、互动的电子游戏等,双方都是普通的 PC 机,不一定能通过 DNS 获得 IP 地址.一种解决办法是要求普通个人通信终端都注册域名;另一种办法是双方通过应用服务器作为通信代理来建立连接,这种方法也正被广泛地应用着.

对根域名服务器访问时,采用 SIPSA 方式通信.根服务器占用的特定地址禁止在用户网络中使用:除 RMT 网关外,用户网络中的任何设备都不能直接访问根服务器,也不能将根服务器占用的地址分配给用户网络中的任何设备.这些限制条件对当前的互联网也是必须的,若 RMT 网关检测到用户设备发出的数据包使用这些特定地址作为目的地址或源地址时,则按非法访问来处理.

7 总结与未来工作

本研究工作的主要目的是解决骨干网路由扩展性和 IPv4 地址耗尽两个重要问题,顺便对移动性和多宿连接也带来了一些好处.除在 DNS 系统中增加一个地址记录外,主要工作都在一个新的网关中执行,对现有的骨干网和用户网、用户主机等没有任何修改要求.网关的处理也相对简单,映射表只与本地通信 session 有关,不存在扩展性问题.用户所感到的变化仅仅是需要通过域名来访问其他主机.未来的工作是对该网关系统进行实验,从实际环境中验证 DIPSA 想法的可行性.通过工程优化方法,能够实现优于 NAT 的性能.另外,让 DIPSA 用户网络在过渡阶段中也能使用全部 IPv4 地址空间的研究工作,已经取得了进展,受篇幅所限,将另文发表.

References:

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

[2] O'Dell M. GSE—An alternate addressing architecture for IPv6. Internet-Draft, IETF, 1997. http://tools.ietf.org/html/draft-ietf-ipngwg-gseaddr-00

[3] Meyer D. The locator identifier separation protocol (LISP). The Internet Protocol Journal, 2008,11(1):23−36.

[4] Lear E. NERD: A not-so-novel EID to RLoc database. Internet-Draft, IETF, 2010. http://tools.ietf.org/html/draft-lear-lisp-nerd-08

[5] Brim S, Farinaci D, Meyer D. EID mappings multicast across cooperating systems for LISP. Internet-Draft, IETF, 2007. http://tools.ietf.org/html/draft-curran-lisp-emacs-00

[6] Brim S, Farinacci D, Fuller V, Lewis D, Meyer D. LISP-CONS: A content distribution overlay network service for LISP. Internet-Draft, IETF, 2007. http://tools.ietf.org/html/draft-meyer-lisp-cons-03

[7] Jen D, Meisel M, Massey D, Wang L, Zhang BC, Zhang LX. APT: A PracticalTransit mapping service. Internet-Draft, IETF, 2007. http://tools.ietf.org/html/draft-jen-apt-01

[8] Bagnulo M. Preliminary LISP threat analysis. Internet-Draft, IETF, 2007. http://tools.ietf.org/html/draft-bagnulo-lisp-threat-01

[9] Atkinson R, Bhatti S, Hailes S. ILNP: Mobility, multi-homing, localised addressing and security through naming. Telecommunication Systems, 2009,42(3):273−291. [doi: 10.1007/s11235-009-9186-5]

[10] Vixie P, Thomson S, Rekhter Y, Bound J. Dynamic updates in the domain name system (DNS UPDATE). RFC 2136, 1997.

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

[12] Tong YP, Li ZQ, Wei B. Analysis of IPv6 transition technologies. Telecommunications Science, 2011,27(1):52−60 (in Chinese with English abstract).

[13] Qian HL, Ge JJ, Li J. Hierarchical Switching Network Architecture. Beijing: Tsinghua University Press, 2008 (in Chinese).

[14] Qian HL, E YP. Hierarchically switched networks. ZTE Communications, 2010,16(2):1−5 (in Chinese with English abstract).

[15] Egevang K, Francis P. The IP network address translator (NAT). RFC 1631, IETF, 1994.

[16] Holdrege M, Srisuresh P. Protocol complications with the IP network address translator. RFC 3027, IETF, 2001.

附中文参考文献:

[12] 仝亚鹏,李振强,魏冰.IPv6 过渡技术分析.电信科学,2011,27(1):52−60.

[13] 钱华林,葛敬国,李俊.层次交换网络体系结构.北京:清华大学出版社,2008.

[14] 钱华林,鄂跃鹏.层次式交换网络.中兴通讯技术,2010,16(2):1−5.

钱华林(1940－),男,上海人,研究员,博士生导师,主要研究领域为网络体系结构,路由系统,网络安全.

任勇毛(1981－),男,博士,助理研究员,CCF 会员,主要研究领域为高速网络,传输协议.

鄂跃鹏(1976－),男,博士生,主要研究领域为网络体系结构,传输层拥塞控制.

游军玲(1981－),女,工程师,主要研究领域为网络体系结构.

葛敬国(1973－),男,博士,副研究员,主要研究领域为计算机网络体系结构.