

# Hidra: 一个分级域间路由架构

王 娜<sup>1),2)</sup> 马海龙<sup>2)</sup> 程东年<sup>2)</sup> 汪斌强<sup>2)</sup>

<sup>1)</sup>(解放军信息工程大学电子技术学院 郑州 450004)

<sup>2)</sup>(解放军信息工程大学信息工程学院 郑州 450002)

**摘 要** Internet 域间路由系统的扩展性面临着严峻挑战. 主要表现在全球路由表膨胀和路由更新频繁. 分析发现造成全球路由表膨胀的根本原因是标识自治系统位置的 IP 前缀数目不可控, 造成路由更新频繁的根本原因是扁平的域间路由结构. 基于此, 文中提出了一个分级域间路由架构 Hidra(Hierarchical inter-domain routing architecture). Hidra 的核心思想是隔离网络边界与核心: “相对稳定”的核心网络位于高阶路由层, 运行高阶域间路由协议, 以维持核心网络的可达性; 在“变化相对剧烈”的边界, 引入一个低阶映射层和相应的映射服务, 以维持边界网络与核心网络之间的可达性. 因为与不稳定的边界网络隔离, 核心网络路由的稳定性增强. Hidra 引入一个标识传送自治系统位置的域间路由标识(Routing IDentity, RID). 它由传送自治系统及其提供商自治系统唯一确定, 显著降低了全球路由表的规模.

**关键词** 域间路由; BGP; 扩展性; 映射服务

**中图法分类号** TP393

**DOI 号:** 10.3724/SP.J.1016.2009.00377

## Hidra: A Hierarchical Inter-Domain Routing Architecture

WANG Na<sup>1),2)</sup> MA Hai-Long<sup>2)</sup> CHENG Dong-Nian<sup>2)</sup> WANG Bin-Qiang<sup>2)</sup>

<sup>1)</sup>(College of Electronic Technology, PLA Information Engineering University, Zhengzhou 450004)

<sup>2)</sup>(College of Information Engineering, PLA Information Engineering University, Zhengzhou 450002)

**Abstract** The Internet inter-domain routing system is facing serious scalability challenge, which represents super-linear growth of the global routing table and increasingly frequent routing updates. The root cause that contributes to the rapid routing table growth is that number of IP prefix identifying location of autonomous system is uncontrollable; the root reason that leads to increasingly frequent routing updates is the flat inter-domain routing architecture. To address these limitations, this paper proposes a Hierarchical Inter-domain Routing Architecture (Hidra). The basic idea of Hidra is to separate edge networks from the core: “relatively stable” core networks lie in a high-rank routing layer, and a high-rank inter-domain routing protocol is operated among routers in core networks to maintain reachability to all other core networks; in the “relatively queasy” edge, a low-rank mapping layer along with a corresponding mapping service is introduced to keep the reachability between the edge and core. Because being separated from unstable edge networks, core networks have the enhanced routing stability. Hidra introduces an inter-domain routing identity RID (Routing Idengity) to identify the location of transit autonomous system, which is only defined by transit autonomous system and its provider autonomous system. As a result, scale of global routing table is observably reduced.

**Keywords** inter-domain routing; BGP; scalability; mapping service

收稿日期: 2008-09-20; 最终修改稿收到日期: 2009-02-01. 本课题得到国家“九七三”重点基础研究发展规划项目基金(2007CB307102)、国家“八六三”高技术研究发展计划项目基金(2007AA01Z405)资助. 王 娜, 女, 1980 年生, 博士研究生, 主要研究兴趣为网络路由与安全. E-mail: tinatwf@gmail.com. 马海龙, 男, 1980 年生, 博士研究生, 主要研究兴趣为网络路由技术. 程东年, 男, 1957 年生, 教授, 研究兴趣包括网络体系结构、网络安全协议及网络性能分析技术. 汪斌强, 男, 1963 年生, 教授, 博士生导师, 研究兴趣为宽带信息网络.

# 1 引 言

诸多研究表明 Internet 域间路由系统的扩展性面临着严峻挑战<sup>[1-3]</sup>. 主要表现在以下两个方面:

(1) 全球路由表膨胀. 通过分析 Oregon 大学 RouteViews 项目<sup>[4]</sup>提供的 Internet 全球路由表“实

时快照”,从 2004 年 3 月到 2008 年 3 月,全球路由表表项数目从 154K 增加到 254K,增长了 65%.

(2) 路由更新频繁. 根据 BGP 稳定性报告<sup>①</sup>,在 2008 年 5 月 4 日到 6 月 3 日的 31 天内,全球路由更新报文数目达到 5M,每秒平均值是 1.9,峰值达到 10K. 图 1 和图 2 显示了 31 天内每秒前缀更新速率的每小时平均值和峰值.

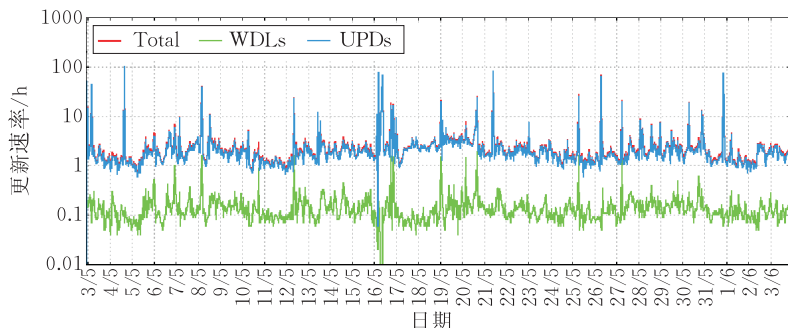


图 1 每秒前缀更新速率的每小时平均值

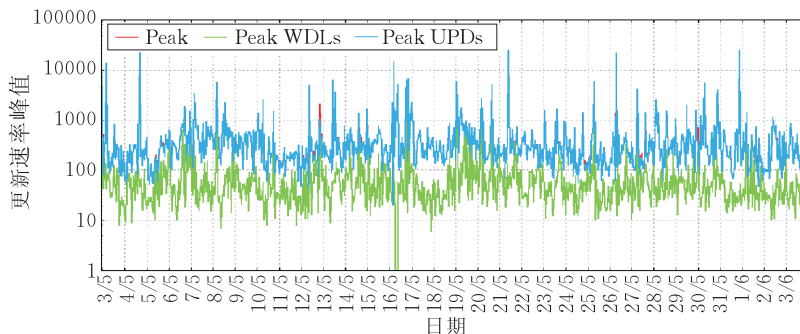


图 2 每秒前缀更新速率的峰值

全球路由表膨胀和频繁的路由更新会导致以下结果: (1) 路由器内存开销显著增加<sup>[5]</sup>. 当路由器内存容量难以满足需求时, BGP 路由器崩溃, 停止接收更新报文, 或者进入某个中间状态, 导致端到端通信的严重破坏<sup>[6]</sup>; (2) 假设路由器的处理开销与路由表规模、路由表项稳定性直接相关<sup>②</sup>, 全球路由表膨胀和频繁的路由更新将导致路由器处理负担加重, BGP 路由处理占用大量的处理器周期<sup>[7]</sup>; (3) 路由器处理更新报文时, 路由表查询时间变长, 导致路由收敛时间延长; (4) 路由器转发表表项数目显著增加, 导致数据包转发时间延长, 端到端通信延时增长; (5) 经济开销增加, 有研究指出一条 BGP 路由更新的经济影响是每年大约 8000 美元, 不管该路由更新的类型和其中包含 IP 前缀的数目<sup>③</sup>.

通过分析全球路由表及路由更新速率的变化规律, APNIC (Asia-Pacific Network Information Center, 亚洲太平洋地区互联网络信息中心) 的著名学者 Huston 预测到 2011 年全球路由表会拥有表项

370K, 每天前缀更新量 2.8M, 前缀撤销量 1.6M, 路由器相关内存开销达到 550M, 处理路由更新会占用 1.5GHz 处理器的 120%<sup>[8-9]</sup>. 在考虑了 IPv6 网络后, 思科学者 Fuller 预测到 2011 年全球 IPv4/v6 路由表表项数目会达到 1049K<sup>④</sup>.

可见, Internet 域间路由系统面临着日益严峻的扩展挑战. 并且, 随着 IPv6 技术的广泛部署应用, 问题会更加严重.

本文分析发现导致全球路由表膨胀的根本原因是标识自治系统位置的 IP 前缀数目不可控, 造成路由更新频繁的根本原因是扁平的域间路由结构. 基

① The BGP instability report. <http://bgpupdates.potaroo.net/instability/bgpupd.html>

② Troubleshooting high CPU caused by the bgp scanner or bgp router process. <http://www.cisco.com/warp/public/459/highcpu-bgp.pdf>

③ What does a BGP route cost? <http://bill.herrin.us/network/bgpcost.html>. 2008

④ Fuller V. Scaling of Internet routing and addressing: Past view, present reality, and possible futures. <http://www.vaf.net/~vaf/apricot-workshop.pdf>. 2007

于此,本文提出了一个新型域间路由架构——分级域间路由架构 Hydra (Hierarchical inter-domain routing architecture). Hydra 的核心思想是隔离“变化相对剧烈”的网络边界和“相对稳定”的网络核心,仅在“相对稳定”的核心网络中运行动态路由协议,在“变化相对剧烈”的网络边界引入映射服务.具体地,Hydra 将域间路由分为低阶映射层和高阶路由层,低阶映射层被用于维持位于网络边界的端(stub)自治系统与位于网络核心的传送(transit)自治系统之间的可达性,高阶路由层被用于维持核心网络中传送自治系统之间的可达性.另外,Hydra 新定义一个域间路由标识 RID(Routing Identity),以标识核心网络中传送自治系统的位置.它由自治系统号码和位置符组成,位置符由传送自治系统及其提供商自治系统唯一确定.

因为向网络核心隔离了网络边界路由的不稳定,评估结果显示 Hydra 网络中高阶域间路由由更新报文数目显著减少,核心网络路由的稳定性增强.域间路由标识格式使标识传送自治系统位置的域间路由标识数目与该自治系统的提供商数目相关.评估结果显示 Hydra 网络全球路由表的规模得到显著降低.

Hydra 还具有以下特点:支持端自治系统和传送自治系统的多宿主、流量工程技术;支持客户网络使用 PI(Provider Independent,提供商无关的)前缀,避免了更改提供商时的重新编号(renumbering)问题;使数据包的源自治系统和源传送自治系统具有一定的选择数据包传输路径的能力;提供了一定的数据包冗余传输能力;增加了核心网络的安全性.

本文第 2 节概述当前研究现状,指出与本文思想相近的 eFIT 存在的不足;第 3 节分析造成全球路由表膨胀和路由更新频繁的根本原因;基于第 3 节的分析,第 4 节给出 Hydra 架构;第 5 节详细说明了 Hydra 网络中数据包传输过程;第 6 节讨论 Hydra 具有的一些特点;第 7 节进行评估;第 8 节比较 Hydra 与 eFIT;最后,总结全文并指出下一步的研究方向.

## 2 研究现状

Internet 域间路由系统面临的扩展挑战引起了全球互联网专家的注意和重视. IRTF 专门成立了工作组 RRG(Routing Research Group,路由研究

组)对该问题展开研究<sup>①</sup>. 大量的解决方案被提出. 根据解决问题的着眼点和方法不同,这些方案可分为位置/身份分离方案、新型路由方案和短期方案 3 类. 下面详细说明.

### (1) 位置/身份分离方案

位置/身份分离方案(见表 1)的提出者们认为造成 Internet 域间路由系统扩展性差的根本原因是 IP 地址语义过载. IP 地址既标识终端身份又标识路由位置. 当 IP 地址标识终端身份时,它是根据 ISP(Internet Service Provider, Internet 服务提供商)组织结构而非拓扑结构分配,导致目前 Internet 唯一有效的路由缩减技术——拓扑聚合技术——失效. 位置/身份分离方案的基本思想是分离身份标识和路由位置标识,路由位置标识可拓扑聚合. 需要说明的是位置/身份分离方案没有修改 Internet 域间路由基本机制和算法.

表 1 典型的位置/身份分离方案

		位置/身份分离方案
基于主机的方案		Shim6 <sup>[10]</sup>
基于网络的方案	地址重写	GSE <sup>[11]</sup>
	Map-and-Encap	LISP <sup>[12]</sup> , IPvLX <sup>[13]</sup> , Ivip <sup>[14]</sup> , CRIO <sup>[15]</sup> , tldr <sup>[16]</sup> , trrp <sup>②</sup>
基于主机+网络的方案		Six/one <sup>[17]</sup>

### (2) 新型路由方案

新型路由方案包括有基于 AS 的新型路由机制,如 HLP<sup>[18]</sup>、HRA<sup>[19]</sup>, CAIDA 的 compact inter-domain routing<sup>[20]</sup>、atomized routing<sup>③</sup>, GIRO(Geographically Informed inter-domain Routing)<sup>[21]</sup>, ROFL(Routing On Flat Labels)<sup>[22]</sup>和 eFIT(enable Future Internet innovation through Transit wire)<sup>[23]</sup>等.

基于 AS 的新型路由机制能够有效缩减全球路由表的规模. 但是,路由粒度过大,ISP 不能够灵活地实现流量控制. Compact inter-domain routing 只是分析了 compact 路由算法在类 Internet 拓扑结构上的一些静态属性,而没有开发或分析实现这些算法的动态路由协议. Atomized routing 面临着以下难题:如何直接发起基于原子(atom)的路由通告和建立原子与原子化前缀(atomized prefix)之间的映

① IRTF Routing Research Group (RRG). [http://www.irtf.org/charter?\\_gtype=rg&.group=rrg](http://www.irtf.org/charter?_gtype=rg&.group=rrg)  
② Tunneling Route Reduction Protocol (TRRP). <http://bill.herrin.us/network/trrp.html>  
③ Atoms-atomised routing. <http://www.caida.org/projects/routing/atoms/>

射关系? GIRO 的基本思想是在地址中增加地理信息,以改善 Internet 域间路由系统的扩展性和性能.但是,实际中许多 ISP 出于安全或保密考虑,可能不愿意暴露自己的详细网络结构(例如,自治系统含有多少路由器,这些路由器详细的地理位置信息等);并且,在网络拓扑图中,两个地理位置邻近的网络可能距离最远. ROFL 提出直接基于标识主机身份的 Flat 标签路由的新思想.

eFIT 的基本思想是隔离网络边界与核心,与本文所提 Hydra 的基本思想相似.但是,与 GIRO 相同,eFIT 核心网络地址(提供商地址)含有所标识路由器的详细地理信息(如经、纬度).并且,eFIT 定义的数据包传送过程可能导致数据包传送失败.根据 eFIT 定义,数据包的源客户网络首先选择数据包进入传送核心网络的入口边界路由器,并发送数据包到该入口边界路由器;入口边界路由器收到数据包后,根据目的客户地址,查询映射表,获得目的提供商地址;最后,封装数据包,并向外发送.在这个过程中,数据包的源客户网络首先确定入口边界路由器;并且,入口边界路由器在查询映射表,获得目的提供商地址时,不考虑本路由器是否可达目的提供商地址.因此,所选择的数据包入口边界路由器与目的提供商地址标识的出口边界路由器之间可能不存在可达路径,数据包传送失败.以图 3 所示拓扑为例说明.假设提供商网络  $F$  与  $A$ 、 $C$  不可达,客户网络  $S$  和  $D$  之间只存在一条可达路径: $S-A-B-D$ . 如果  $S$  选择路由器  $P_2$  作为入口边界路由器,因为  $P_2$  没有到达目的提供商地址( $P_3$  或  $P_4$ )的路由信息,该数据包将会被丢弃,数据包传送失败.即使  $S$  选择  $P_1$  作为入口边界路由器,如果  $P_1$  查询映射表,选择的提供商地址是  $P_4$ ,该数据包的传送仍然失败.

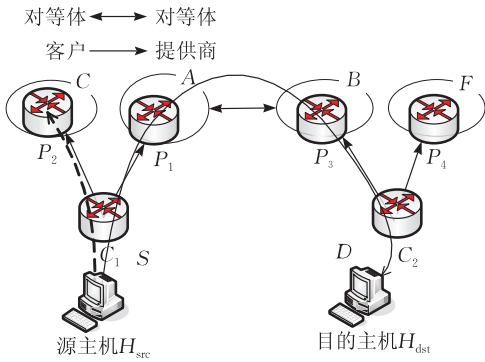


图 3 eFIT 数据包传送失败示意

(3) 短期方案

Forgetful routing<sup>[5]</sup> 的出发点是减少路由表占用的内存空间,方法是压缩替代路由. Forgetful

routing 不要求修改 BGP 协议,可增量式部署且不影响路由收敛.它能够有效缓解路由器因路由表膨胀而带来的存储压力,是一个很好的短期解决方案.

3 问题分析

我们认为造成 Internet 全球路由表膨胀的根本原因是标识自治系统位置的 IP 前缀数目不可控.将 Internet 视为一个无向图  $G=(V,E)$ ,其中节点集合  $V$  表示自治系统集合,边集合  $E$  表示自治系统之间连接集合.域间路由协议 BGP 可理解为在图  $G$  中寻找任两点之间满足一定约束条件的可达路径. IP 前缀被用于标识图  $G$  中节点(自治系统)的位置.它分为拥有前缀和 more specific 前缀两类.拥有前缀指自治系统拥有的前缀,可能是 PI 前缀和(或)PA (Provider Assigned,提供商分配的)前缀.因为多宿主自治系统可任意地通告它所拥有 PA 前缀的 more specific 前缀,且 more specific 前缀和 PI 前缀无法被聚合,所以标识自治系统位置的 IP 前缀数目不可控,导致全球路由表膨胀.

造成域间路由更新频繁的根本原因是 Internet 扁平的域间路由结构.因为扁平的域间路由结构和 PI、more specific 不可聚合前缀的出现,网络边界路由的不稳定会直接影响核心网络路由的稳定性;并且,核心网络连接的丰富性会进一步地放大网络边界路由的不稳定性.例如,不稳定端自治系统引发的路由抖动会被洪泛到整个 Internet<sup>[24]</sup>;端自治系统的一个配置错误会造成网络大规模的损害<sup>[25]</sup>等.大量研究<sup>[26-28]</sup>发现 Internet 中少数不稳定的边界网络贡献了大量的更新报文,其中大部分更新报文被证明所含路由信息无效,并不是实际网络拓扑变化的反应.

4 Hydra

基于上述分析,下面提出一个新型域间路由架构 Hydra (Hierarchical inter-domain routing architecture,分级域间路由架构). Hydra 的核心思想是分级路由,即在“相对稳定”的核心网络中运行动态路由协议,在“变化相对剧烈”的网络边界引入映射服务.具体地,Hydra 将域间路由分为低阶映射层和高阶路由层,低阶映射层被用于维持位于网络边界的端自治系统与位于网络核心的传送自治系统之间的可达性,高阶路由层被用于维持核心网络中传送

自治系统之间的可达性. 另外, Hidra 新定义一个域间路由标识(RID), 以标识传送自治系统的位置.

Hidra 由映射表建立、管理协议、路由表分发协议、高阶域间路由协议和路径维持协议组成. 其中, 映射表建立、管理协议和路由表分发协议用于维持端自治系统与传送自治系统之间的可达性; 高阶域间路由协议用于维持传送自治系统之间的可达性; 路径维持协议用于维持两个直接相连的端自治系统之间和端自治系统与直接相连的传送自治系统之间的可达性. Hidra 的结构如图 4 所示.

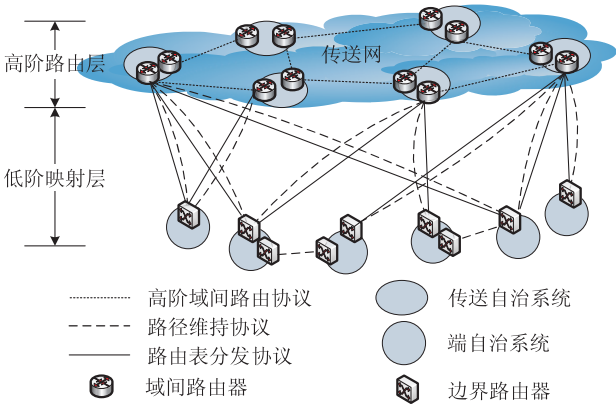


图 4 Hidra 架构

在进一步说明之前, 解释以下名词.

- (1) Hidra 网络: 采用 Hidra 作为域间路由架构的网络.
- (2) 传送网: 由所有传送自治系统构成的网络.
- (3) Transit-edge 自治系统: 传送网中发起(originate)路由通告的自治系统.
- (4) 域间路由器: 传送自治系统内运行高阶域间路由协议、路径维持协议和路由表分发协议的路由器.
- (5) 边界路由器: 端自治系统内运行路径维持协议和路由表分发协议的路由器.

4.1 域间路由标识

本小节将讨论域间路由标识的格式. 在数据平面, 域间路由标识(RID)标识数据包进入和离开传送自治系统的位置.

将传送网视为一个无向图  $G=(V,E)$ , 其中节点集合  $V$  表示传送自治系统集合, 边集合  $E$  表示传送自治系统之间连接集合. 假设自治系统  $AS_i$  和  $AS_j$  直接相连,  $AS_i$  和  $AS_j$  之间的连接表示边  $(AS_i, AS_j)$ . 在传送网无向图  $G$  中, 数据包进入和离开某个传送自治系统的位置可由与该传送自治系统直接相连的自治系统确定. 例如, 在图 5 中, 数

据包进入和离开自治系统  $AS_3$  的位置可表示为  $Loc(AS_3, AS_2)$ 、 $Loc(AS_3, AS_4)$ 、 $Loc(AS_3, AS_5)$  和  $Loc(AS_3, AS_6)$ .

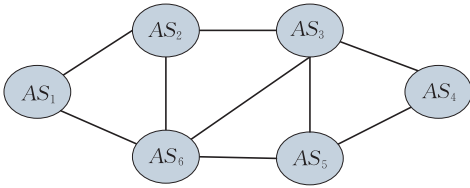


图 5 传送网无向图

考虑自治系统之间的商业关系. 假设自治系统  $AS_i$  和  $AS_j$  直接相连. 当  $AS_i$  和  $AS_j$  之间是对等体-对等体关系(或客户-提供商关系)时,  $AS_i$  以及  $AS_j$  所有客户自治系统发起的到达  $AS_j$  或者  $AS_j$  任一客户自治系统的数据包都被转发, 经由  $AS_i$  和  $AS_j$  之间的连接到达  $AS_j$ .  $AS_j$  不能控制从对等体自治系统(或客户自治系统)进入的数据包; 反之, 亦然. 但是, 当  $AS_j$  拥有多个提供商自治系统时, 它能够控制数据包经由哪个提供商自治系统进入或离开到哪个提供商自治系统. 因此, 对某个自治系统, 数据包从它的对等体/客户自治系统进入或离开到对等体/客户自治系统的位置可仅由该自治系统确定, 数据包从提供商自治系统进入或离开到提供商自治系统的位置可由该自治系统及其提供商自治系统共同确定. 例如, 在图 6 中, 对  $AS_3$ , 数据包从  $AS_2/AS_4$  进入(或数据包离开到  $AS_2/AS_4$ )的位置可表示为  $Loc(AS_3)$ , 数据包从  $AS_5/AS_6$  进入(或数据包离开到  $AS_5/AS_6$ )的位置可表示为  $Loc(AS_3, AS_5)/Loc(AS_3, AS_6)$ .

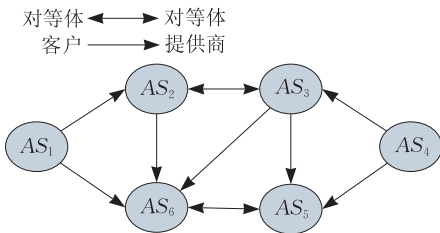


图 6 传送网

基于上述分析, 定义域间路由标识由自治系统号码(ASN)和位置符(Locator)组成. 称自治系统号码为域间路由标识前缀(RID 前缀). 位置符由传送自治系统及其提供商自治系统唯一确定, 可通过下式计算获得:

$$Locator = Loc_{provider}(AS_{transit} \parallel AS_{provider}) \quad (1)$$

其中,  $Loc_{provider}$  表示位置符生成算法,  $AS_{transit}$  表示传



送自治系统号码,  $AS_{\text{provider}}$  表示传送自治系统的提供商自治系统号码.

要求算法  $Loc_{\text{provider}}$  至少满足以下条件: (1) 能够验证该位置符是由自治系统  $AS_{\text{provider}}$  生成; (2) 位置符唯一, 且不可伪造.

定义域间路由标识分为默认域间路由标识 ( $dRID$ ) 和特殊域间路由标识 ( $pRID$ ) 两类. 默认域间路由标识 ( $dRID$ ) 的位置符为空.

- 基于 RID 的域间路由通告要满足以下要求:
- (1) 自治系统向对等体和客户自治系统发起关于自己  $dRID$  的路由通告, 向提供商自治系统发起关于含有由它生成位置符的  $pRID$  的路由通告;
  - (2) 当  $dRID$  被通告到对等体或客户自治系统后, 它们只向自己的客户自治系统继续通告此  $dRID$ ;
  - (3) 当  $pRID$  被通告到提供商自治系统后, 该提供商自治系统验证  $pRID$  中位置符是否由自己生成; 若是, 继续向外通告此  $pRID$ ; 反之, 丢弃该路由通告.

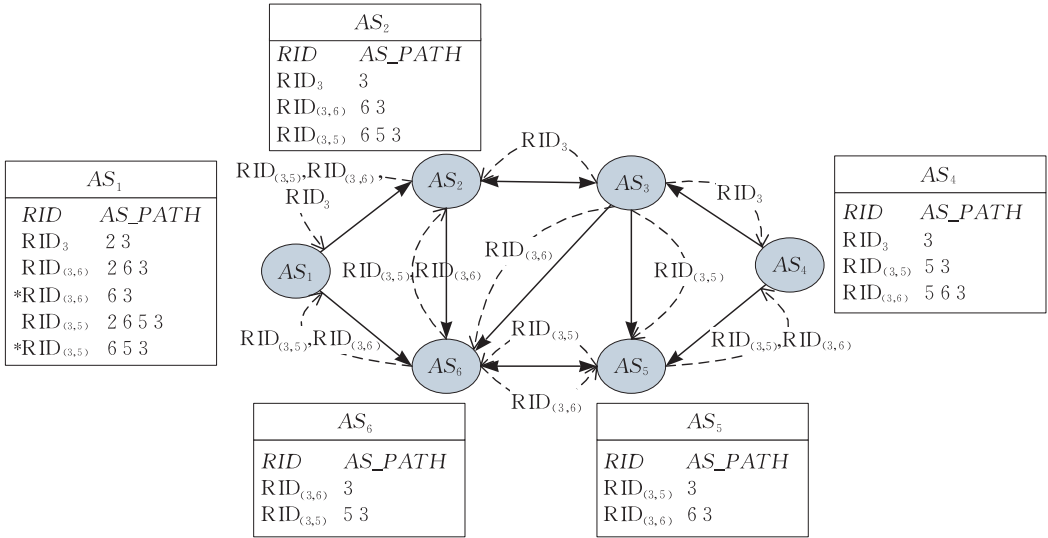


图 7 路由计算及路由表示例

4.3 映射表建立、管理协议

映射表建立、管理协议用于建立、管理 IP 前缀与 RID 之间的映射关系. IP 前缀与 RID 是一对多的映射关系.

- 根据 RID 的不同类型, 映射表项分为两类:
- (1) 默认映射表项: (IP 前缀,  $dRID$ );
  - (2) 特殊映射表项: (IP 前缀,  $pRID, weight$ ).

给定一个 IP 地址, 当映射表中存在多个与该 IP 地址相匹配的特殊映射表项时,  $weight$  项用于确定选择哪个特殊映射表项. “与 IP 地址相匹配的映射表项”指表项中 IP 前缀包含该 IP 地址的映射表

4.2 高阶域间路由协议

传送网内传送自治系统之间运行高阶域间路由协议, 以建立基于 RID 的域间路由表, 维持传送自治系统之间的可达性. BGP 可用作高阶域间路由协议, 只是基于域间路由标识, 满足 RID 的路由通告要求.

举例说明基于 RID 的 BGP 路由计算和路由表. 在图 7 中,  $AS_3$  拥有一个默认域间路由标识  $RID_3$  和两个特殊域间路由标识  $RID_{(3,5)}$ 、 $RID_{(3,6)}$ . 以  $AS_1$  为例. 根据基于 RID 的域间路由通告要求,  $AS_3$  向  $AS_2$  发起关于  $RID_3$  的路由通告, 向  $AS_6$  发起关于  $RID_{(3,6)}$  的路由通告, 向  $AS_5$  发起关于  $RID_{(3,5)}$  的路由通告;  $AS_5$  向  $AS_6$  继续通告关于  $RID_{(3,5)}$  路由;  $AS_6$  向  $AS_2$  和  $AS_1$  继续通告关于  $RID_{(3,5)}$  和  $RID_{(3,6)}$  的路由;  $AS_2$  向  $AS_1$  继续通告关于  $RID_3$ 、 $RID_{(3,5)}$  和  $RID_{(3,6)}$  的路由; 最后, 根据最短路径优先原则,  $AS_1$  将分别选择从  $AS_6$  获得的路由作为到达  $RID_{(3,5)}$  和  $RID_{(3,6)}$  的最优路由 (图中用 \* 标出).

项. 对拥有相同 IP 前缀  $prefix$  的特殊映射表项集合  $SM = \{M_1, \dots, M_k\}$ , 其中  $M_i = (prefix, pRID_i, weight_i)$ ,  $1 \leq i \leq k$ ,  $k \geq 1$ , 满足以下条件:

$$weight_i \leq 1 \text{ 且 } \sum_{i=1}^k weight_i = 1 \tag{2}$$

即映射表中所有 IP 前缀相同的特殊映射表项的  $weight$  值之和等于 1.

当端和传送自治系统多宿主时, 它们可通过设置 IP 前缀与不同  $pRID$  的映射关系以及不同的  $weight$  值来实现输入流量控制. 以图 8 所示拓扑为例说明. 假设端自治系统 X 拥有前缀 63.63.62.0/23;

$RID_{(A,C)}$  表示传送自治系统  $A$  拥有的与提供商自治系统  $C$  相关的  $pRID$ ;  $RID_{(B,D)}$ 、 $RID_{(B,E)}$  分别表示  $B$  拥有的与提供商自治系统  $D$ 、 $E$  相关的  $pRID$ . 如果  $X$  希望所有到达  $63.63.62.0/23$  的输入流量都只经过  $B$ ,  $B$  希望到达  $63.63.62.0/23$  的流量输入均衡, 可通过设置映射表项:  $(63.63.62.0/23, RID_{(B,D)}, 0.5)$ ,  $(63.63.62.0/23, RID_{(B,E)}, 0.5)$  实现; 如果  $X$  希望到达  $63.63.62.0/23$  的流量输入均衡,  $B$  希望到达  $63.63.62.0/23$  流量只经过  $E$ , 那么相应的映射表项是  $(63.63.62.0/23, RID_{(A,C)}, 0.5)$ ,  $(63.63.62.0/23, RID_{(B,E)}, 0.5)$ .

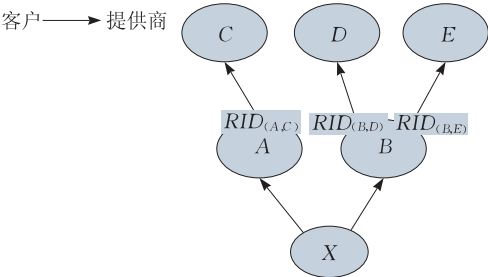


图 8 网络拓扑结构

建议由专门的设备(如映射表服务器)完成映射表的建立和管理. 对于映射表的管理和查询, 可借鉴目前已经提出的一些方法, 如 LISP-NERD<sup>[29]</sup>, APT<sup>[30]</sup>等, 本文不再展开讨论.

4.4 路由表分发协议

传送自治系统运行路由表分发协议, 向直接相连的端自治系统分发它的域间路由表. 路由表分发协议类似于单向 BGP 协议: 端自治系统只接收传送自治系统发出的基于 RID 的域间路由通告, 它并不进一步地传播该路由通告, 也不会和无法向传送自治系统发出基于 RID 的域间路由通告. 根据直接相连传送自治系统提供的域间路由表, 端自治系统选择数据包的目的域间路由标识和源传送自治系统<sup>①</sup>, 以确保所选择源传送自治系统可达目的域间路由标识. 详见第 5 节的讨论.

4.5 路径维持协议

在直接相连的两个端自治系统之间或直接相连的端自治系统和传送自治系统之间传输的数据包不穿越传送网. 因此, 这些系统运行路径维持协议, 建立基于 IP 地址的本地路由表, 以维持彼此可达. 路径维持协议只在两个直接相连的端自治系统之间或直接相连的端自治系统和传送自治系统之间运行. BGP 可用作路径维持协议, 只是所建立路径只有 1 跳距离.

5 数据包传送

数据包在 Hidra 网络中的传送分为基于本地路由表的传送和基于域间路由表的传送两类. 因篇幅所限, 本文只详细描述基于域间路由表的数据包传送过程.

在端自治系统与不直接相连的端自治系统和传送自治系统之间, 或传送自治系统之间传输的数据包穿越传送网, 基于域间路由表.

基于域间路由表的数据包传送基本过程如下:

- 1. 源主机发出源地址是  $IP_{src}$ , 目的地址是  $IP_{dst}$  的数据包;
- 2. 因目的地址  $IP_{dst}$  不在本域内, 该数据包被路由到源主机所在自治系统的边界(域间)路由器;
- 3. 边界(域间)路由器收到数据包后, 根据  $IP_{dst}$  查询本地路由表; 若本地路由表中不存在相关路由信息, 根据源自治系统的不同类型, 分别执行以下操作:
  - 3.1. 若源自治系统是传送自治系统, 域间路由器
    - 3.1.1. 查询映射表, 获得与  $IP_{dst}$  相匹配的映射表项; 对映射表项中的 RID, 查询域间路由表, 判断是否可达, 获得可达映射表项<sup>②</sup>; 若存在多个可达映射表项, 根据一定的策略和原则<sup>③</sup>, 确定目的域间路由标识  $RID_{dst}$ ;
    - 3.1.2. 判断自己通过哪些域间路由标识可达  $RID_{dst}$ ; 若存在多个可达域间路由标识, 根据域间路由表和一定的路径选择策略<sup>④</sup>, 选择其中一个作为源域间路由标识  $RID_{src}$ ;
    - 3.1.3. 使用  $RID_{src}$  和  $RID_{dst}$  封装数据包; 转发封装数据包进入传送网;
  - 3.2. 若源自治系统是端自治系统, 边界路由器
    - 3.2.1. 查询映射表, 获得与  $IP_{dst}$  相匹配的映射表项; 对映射表项中的 RID, 分别查询所拥有的域间路由表, 判断是否可达, 获得可达映射表项, 并记录每个可达映射表项是由哪个域间路由表查询获得(即通过哪个传送自治系统可达, 称该传送自治系统为“可达传送自治系统”); 若存在多个可达映射表项, 根据一定的策略和原则, 确定目的域间路由标识  $RID_{dst}$ ;
    - 3.2.2. 对  $RID_{dst}$ , 若存在多个可达传送自治系统, 根据

① 源传送自治系统指数据包进入传送网的入口自治系统.  
② 可达映射表项指域间路由表拥有到达该映射表项中 RID 路由的映射表项.  
③ 如果可达映射表项都是特殊映射表项, 可根据表项中的 *weight* 值, 确定目的域间路由标识; 如果可达映射表项包括特殊映射表项和默认映射表项, 自治系统可根据一定的策略和原则, 如特殊映射表项优于默认映射表项, 确定目的域间路由标识.  
④ Hidra 支持的路径选择策略可分为本地策略(如输出流量控制策略)和远程策略(如数据包的传输不经过某个传送自治系统).

这些可达传送自治系统提供的关于  $RID_{dst}$  的路由信息及一定的路径选择策略<sup>①</sup>, 选择一个可达传送自治系统作为源传送自治系统;

3.2.3. 转发数据包和所选择的  $RID_{dst}$  至源传送自治系统;

3.2.4. 源传送自治系统收到数据包后, 判断自己通过哪些域间路由标识可达  $RID_{dst}$ ; 若存在多个可达域间路由

标识, 根据域间路由表和一定的路径选择策略, 选择其中一个作为源域间路由标识  $RID_{src}$ ; 使用  $RID_{src}$  和  $RID_{dst}$  封装数据包; 转发封装数据包进入传送网;

4. 传送网中域间路由器根据  $RID_{dst}$  转发封装数据包到目的传送自治系统<sup>②</sup>;

5. 目的传送自治系统收到该数据包后, 解封装包, 根据  $IP_{dst}$  查询本地路由表, 转发数据包至目的主机。

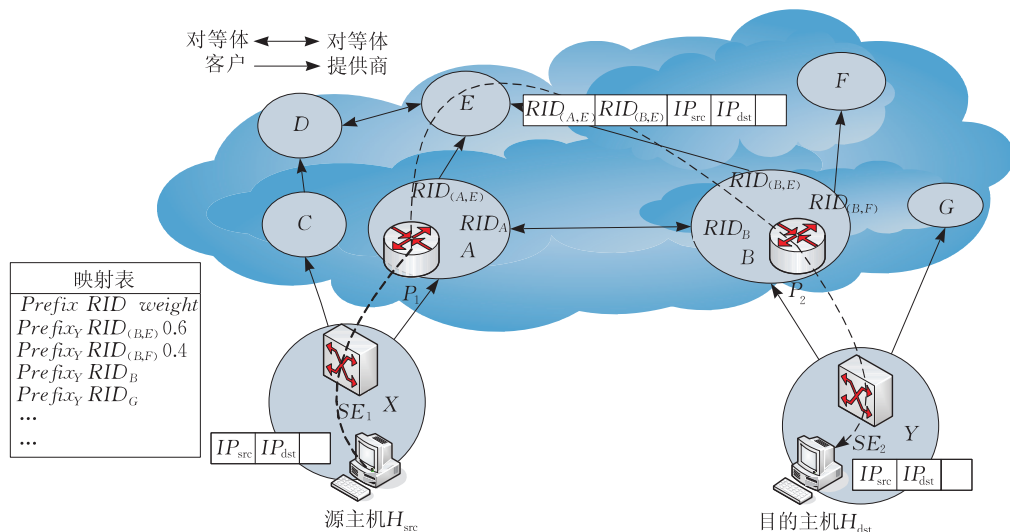


图 9 基于域间路由表的数据包传送

以图 9 所示拓扑为例说明. 假设传送自治系统  $G, F$  与  $C, A$  不可达, 端自治系统  $Y$  拥有前缀  $Prefix_Y$ , 源主机  $H_{src}$  欲发送到达目的主机  $H_{dst}$  的数据包:

1.  $H_{src}$  发出源地址是  $IP_{src}$ , 目的地址是  $IP_{dst}$  的数据包;

2. 因目的地址  $IP_{dst}$  不在本域内, 该数据包被路由到边界路由器  $SE_1$ ;

3.  $SE_1$  收到数据包后, 根据  $IP_{dst}$  查询本地路由表; 本地路由表中不存在相关路由信息,

3.1.  $SE_1$  查询映射表, 获得与  $IP_{dst}$  相匹配的映射表项:  $(Prefix_Y, RID_{(B,E)}, 0.6)$ ,  $(Prefix_Y, RID_{(B,F)}, 0.4)$ ,  $(Prefix_Y, RID_B)$  和  $(Prefix_Y, RID_G)$ ; 对每个表项中的  $RID$ , 分别查询所拥有的域间路由表, 因为  $G, F$  与  $C, A$  不可达, 获得以下可达映射表项:  $(Prefix_Y, RID_{(B,E)}, 0.6)$  和  $(Prefix_Y, RID_B)$ ; 假设  $X$  选择  $RID_{(B,E)}$  作为目的域间路由标识;

3.2. 因为  $A$  和  $C$  都拥有  $RID_{(B,E)}$  的路由信息, 根据输出流量控制策略,  $X$  选择  $A$  作为源传送自治系统, 并转发数据包和  $RID_{(B,E)}$  至  $A$  中的域间路由器  $P_1$ ;

4.  $P_1$  收到数据包后, 判断自己通过域间路由标识  $RID_{(A,E)}$  可达  $RID_{(B,E)}$ ; 使用  $RID_{(A,E)}$  和  $RID_{(B,E)}$  封装数据包; 并向外转发封装后数据包;

5. 传送网中域间路由器根据  $RID_{(B,E)}$  转发封装数据包到目的传送自治系统  $B$  中的域间路由器  $P_2$ ;

6.  $P_2$  收到该数据包后, 解封装, 根据  $IP_{dst}$  查询本地路由表, 转发数据包至  $SE_2$ ;

7.  $SE_2$  收到数据包后, 根据  $IP_{dst}$  转发数据包到目的主机  $H_{dst}$ .

下面将证明 Hydra 所定义的基于域间路由表的数据包传送过程正确. 本文认为只要根据所定义数据包传送过程获得的数据包源和目的域间路由标识之间存在可达路径, 该数据包传送过程就正确.

**定理 1.** Hydra 所定义的基于域间路由表的数据包传送过程正确.

证明.

在数据包的源和目的自治系统都是端自治系统的情况下, 根据 Hydra 所定义的基于域间路由表的数据包传送过程, 源自治系统首先查询映射表和直接相连传送自治系统提供的域间路由表, 获得可达映射表项和可达传送自治系统, 并分别从中选择目的域间路由标识和源传送自治系统. 此步骤保证了源传送自治系统可达目的域间路由标识. 在数据包和目的域间路由标识被发送到源传送自治系统后, 源传送自治系统将选择一个可达目标域间路由标识的域间路由标识作为源域间路由标识, 以保证数据包的源和目的域间路由标识之间存在可达路径.

① 同前页脚注④. 如果支持远程策略, 需要存在一个机制, 将源自治系统的远程策略告知源传送自治系统, 以避免源和源传送自治系统的路径选择策略发生冲突.

② 目的传送自治系统指数据包出传送网的出口自治系统.



对源和(或)目的自治系统是传送自治系统的情况,证明过程相似,本文不再赘述。

因此,Hidra 定义的基于域间路由表的数据包传送过程保证了数据包的源和目的域间路由标识之间存在可达路径,该数据包传送过程正确。 证毕。

6 一些讨论

与 Internet 域间路由系统不同,Hidra 允许源和源传送自治系统根据一定的路径选择策略,选择数据包的传输路径。它与源路由协议<sup>[31]</sup>的不同之处是 Hidra 中源和源传送自治系统只能从其它自治系统通告的最优路径中选择数据包的传输路径,而源路由协议是源自治系统自己发现和选择到达目的自治系统的路径。

当不存在与数据包目的 RID 完全匹配的转发表项时,Hidra 允许路由器根据与目的 RID 前缀相同的其它 RID( $dRID$  或  $pRID$ )转发该数据包,以使数据包不被丢弃。例如,在图 10 拓扑中,正常情况下,传送自治系统 A 的域间路由器  $R_A$  拥有到达  $RID_{(E,C)}$ 、 $RID_{(E,B)}$  和  $RID_E$  的转发表项。假设 A 中主机  $H_{src}$  欲发送数据包至 E 中主机  $H_{dst}$ ,且当  $H_{src}$  发出数据包被转发至  $R_A$  时, $R_A$  确定该数据包的域的域间路由标识为  $RID_{(E,C)}$ 。如果在  $H_{src}$  与  $H_{dst}$  通信过程中,E 和 C 之间链路发生故障,彼此不要达。C 检测出故障后,向 A 发出  $RID_{(E,C)}$  路由撤销报文;收到此报文后,A 撤销到达  $RID_{(E,C)}$  的路由,同时删除转发表中相应表项。在这种情况下, $R_A$  在转发到达  $RID_{(E,C)}$  的数据包时,可根据与  $RID_{(E,C)}$  拥有相同 RID 前缀的其它转发表项( $RID_{(E,B)}$  或  $RID_E$ )进行转发。假设它选择转发表项  $RID_E$ ,将该数据包转发至 D。因为没有到达  $RID_{(E,C)}$  的转发表项,D 依然根据  $RID_E$  转发该数据包至 E。这个过程将一直持续到 E 与 C 之间再次彼此可达。

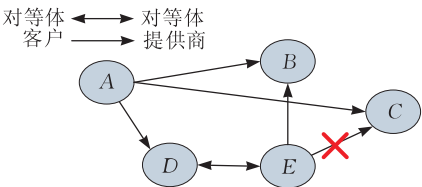


图 10 域间路由器可根据与目的 RID 拥有相同 RID 前缀的转发表项转发数据包

在建立 IP 前缀与 RID 映射关系时,如果不建立 IP 前缀与源自治系统的绑定关系,由源自治系统自由地(不需授权地)要求建立某个 IP 前缀与 RID 的映射关系,Hidra 网络也会受到 IP 前缀劫持攻

击。以端自治系统为例说明。当某个端自治系统要求建立一个非本自治系统内 IP 前缀与其提供商自治系统 RID 的映射关系,导致网络中以该前缀为目的地的全部/部分数据包被路由到该提供商自治系统时,针对 Hidra 网络的前缀劫持攻击发生。我们称该前缀为被劫持前缀,称以被劫持前缀为目的地址的被路由到提供商自治系统的数据包为被劫持数据包。如果该提供商自治系统不能够验证被劫持前缀是否由这个端自治系统拥有,当该端自治系统根据路径维持协议,发出关于被劫持前缀的路由通告到此提供商自治系统时,提供商自治系统的本地路由表中将存在关于被劫持前缀的路由表项。从而,它将根据该表项转发被劫持数据包到该端自治系统。如果该提供商自治系统能够验证被劫持前缀不被这个端自治系统拥有,提供商自治系统的本地路由表中将不存在关于被劫持前缀的路由表项,它将直接丢弃这些被劫持数据包。但是,不论提供商自治系统能否验证被劫持前缀是否由这个端自治系统拥有,前缀劫持攻击都已经发生,对网络都已经造成危害。因此,为了抵抗 IP 前缀劫持攻击,正确地建立 IP 前缀与 RIDs 的映射关系,建立 IP 前缀与源自治系统的绑定关系至关重要。可采用文献[32]所提方法建立 IP 前缀与源自治系统的绑定关系。

Hidra 网络可抵抗域间路由标识劫持攻击。域间路由标识劫持攻击是指恶意传送自治系统发起关于其它传送自治系统域间路由标识的路由通告,使以该域间路由标识为目的地的全部/部分数据包被转发到恶意传送自治系统。根据域间路由标识通告要求:提供商自治系统只进一步地通告含有自己生成位置符的  $pRID$ ,恶意传送自治系统不能向它的提供商自治系统发起关于其它传送自治系统 RID( $dRID$  和  $pRID$ )的路由通告。对对等体和客户自治系统,可通过一定的检测机制,即传送自治系统只接收对等体和提供商自治系统发起的关于  $dRID$  的路由通告,且通过验证该  $dRID$  前缀与发起对等体或提供商自治系统号码是否一致,来避免恶意传送自治系统向对等体和客户自治系统发起其它传送自治系统 RID( $dRID$  和  $pRID$ )的路由通告。

7 评 估

7.1 全球路由表

传送网中只有 transit-edge 自治系统发出关于 RID 的域间路由通告。若 transit-edge 自治系统单宿主,它只需通告两个 RID( $pRID$  和  $dRID$ );若 transit-edge 自治系统多宿主,拥有  $n$  个提供商自治

系统( $n>1$ ),它需要通告  $n+1$  个 RID.

为便于说明,定义以下符号: $N_{te}$  表示 transit-edge 自治系统数目; $Pe_{mte}$  表示多宿主 transit-edge 自治系统所占比例; $\overline{NP}_{mte}$  表示多宿主 transit-edge 自治系统的平均提供商数目; $N_{Hidra}$  表示传送网域间路由表表项数目.

获得下式:

$$N_{Hidra}=N_{te} \cdot Pe_{mte} \cdot (\overline{NP}_{mte}+1)+N_{te} \cdot (1-Pe_{mte}) \cdot 2$$

(3)

基于 2008 年 3 月的 RouteViews 数据分析结果<sup>①</sup>, Internet 全球路由表有表项 253750,且  $N_{te}=4277$ ,  $Pe_{mte}=0.76268$ ,  $\overline{NP}_{mte}=3.297$ . 基于式(3),

$N_{Hidra}=16049$ . 可见,与 Internet 全球路由表相比, Hidra 全球路由表缩减了一个数量级.

根据 2004 年 3 月至 2008 年 3 月 RouteViews 数据分析结果,  $N_{te}$ 、 $Pe_{mte}$  和  $\overline{NP}_{mte}$  的变化曲线如图 11 所示. 基于式(3),图 12 给出了  $N_{Hidra}$  的变化曲线. 显然,  $N_{Hidra}$  呈线性增长,且可拟合成线性函数(如图 13 所示):

$$f_1(x)=138.2 \times x+8901$$

(4)

其中  $x$  表示距离 2004 年 3 月的月份数.

假设时间表示为  $Y$  年  $M$  月,那么,

$$x=(Y-2004) \times 12+M-2$$

(5)

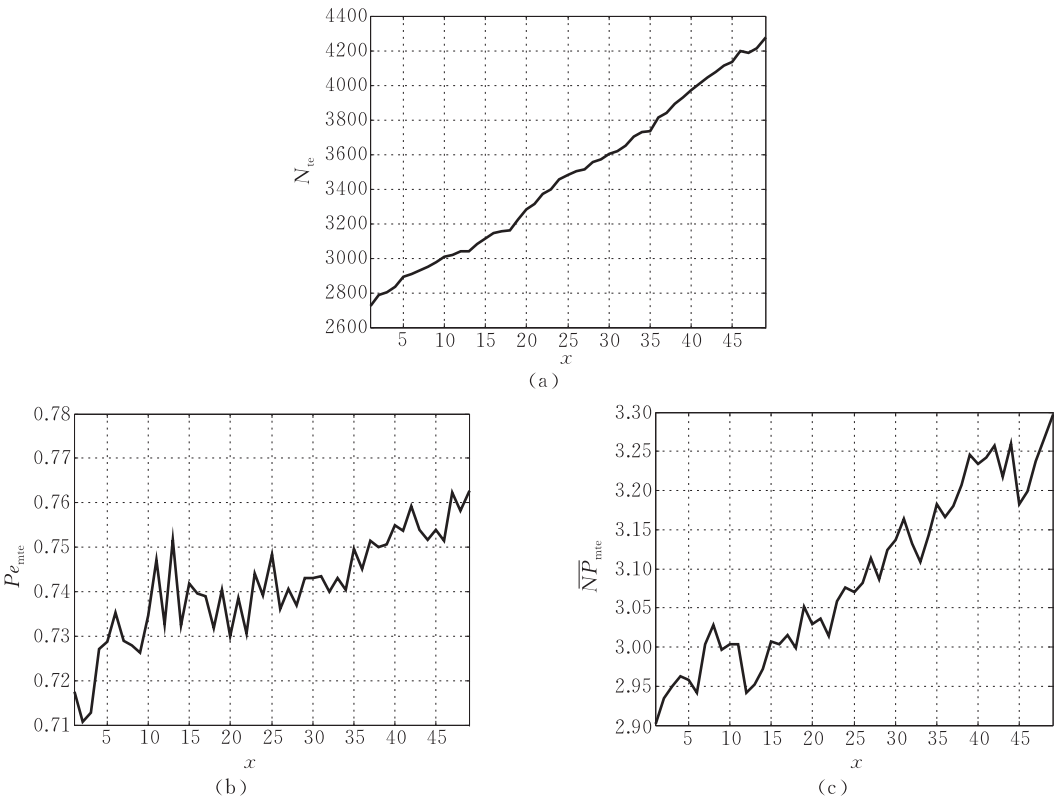


图 11  $N_{te}$ 、 $Pe_{mte}$  及  $\overline{NP}_{mte}$  的变化曲线

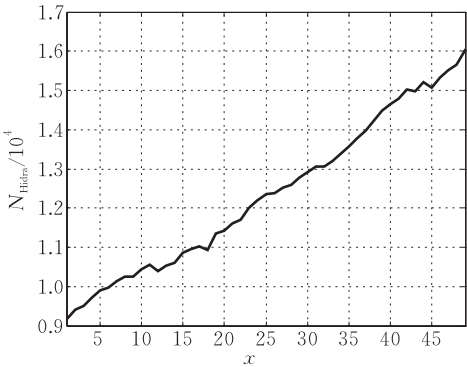


图 12  $N_{Hidra}$  的变化曲线

根据式(4)和(5),当  $f_1(x)=253750$  时,

$$(Y-2004)=147$$

(6)

可见,按照目前  $N_{te}$ 、 $Pe_{mte}$  和  $\overline{NP}_{mte}$  的增长趋势,到 2151 年左右, Hidra 全球路由表才会达到目前 Internet 全球路由表的规模.

① Transit-edge 自治系统由 Internet 中存在客户是端自治系统的传送自治系统(称此传送自治系统为 stub transit 自治系统)和发起 BGP 路由更新报文的传送自治系统(称此传送自治系统为 origin transit 自治系统)组成. 实际中 stub transit 自治系统集合与 origin transit 自治系统集合不完全相交,即存在 stub transit 自治系统不是 origin transit 自治系统而是 transit-only 自治系统的情况.

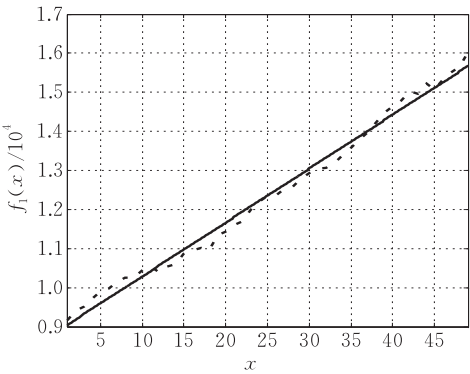


图 13  $N_{\text{Hydra}}$  及其拟合曲线

7.2 路由更新报文

根据 Internet 中发起路由更新报文的源自治系统类型,更新报文可分为 stub 更新报文和 transit 更新报文. 其中,stub 更新报文指由端自治系统发起的路由更新报文,transit 更新报文指由传送自治系统发起的路由更新报文.

假设在一定时间段内, $NU_{\text{Internet}}$  表示 Internet 中路由更新报文数目, $NU_{\text{stub}}$  表示 Internet 中 stub 更新报文数目, $NU_{\text{transit}}$  表示 Internet 中 transit 更新报文数目, $NU_{\text{Hydra}}$  表示 Hydra 高阶域间路由更新报文数目. 以 Internet 中 transit 更新报文数目度量 Hydra 高阶域间路由更新报文数目. 获得以下公式:

$$NU_{\text{Internet}} = NU_{\text{stub}} + NU_{\text{transit}} \tag{7}$$

$$NU_{\text{Hydra}} = NU_{\text{transit}} \tag{8}$$

基于文献[33]中对 2005 年 Internet 路由更新报文数目的分析结果(如图 14 所示),Hydra 高阶域间路由更新报文数目与 Internet 相比至少缩减了 60%.

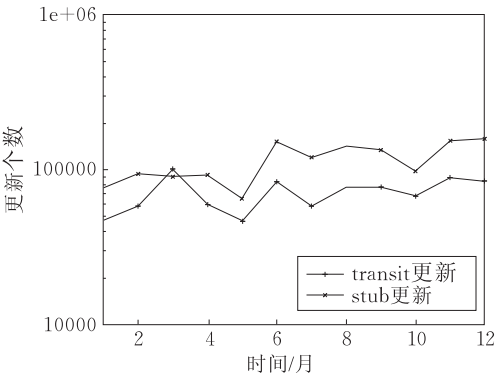


图 14 Internet 路由更新报文个数<sup>[33]</sup>

8 与 eFIT 的比较

Hydra 与 eFIT 的基本思想相似,都是建议隔离

网络边界与核心;核心网路由协议都基于一个新的地址(提供商地址或者域间路由标识),该地址空间与边界网络地址空间隔离;都引入映射服务,以建立、维持和管理边界网络地址与核心网络地址之间的映射关系;数据包在核心网络中都采用隧道传输等.

但是,Hydra 与 eFIT 存在以下不同:

(1) 核心网络地址的语义和格式不同.

Hydra 核心网络地址(域间路由标识)被用于标识传送自治系统的位置,它由传送自治系统及其提供商自治系统唯一确定,由自治系统号码和位置符两部分组成. eFIT 核心网络地址(提供商地址)被用于标识提供商网络中边界路由器的位置. 与 GIRO 地址相似,它由提供商 ID(如自治系统号码)、地理 ID 等部分组成. 其中,地理 ID 标识了边界路由器所在位置的详细地理信息(如经、纬度).

(2) eFIT 的设计缺陷可能导致数据包传送失败,而 Hydra 能够保证数据包在网络中被成功传送<sup>①</sup>.

如第 2 节所讨论,eFIT 确定数据包源和目的提供商地址顺序的错误以及在确定目的提供商地址时没有考虑该地址的可达性,可能导致数据包在网络中传送失败. 与 eFIT 不同,Hydra 首先确定数据包的目的域间路由标识,并在确定目的域间路由标识时,通过查询相关路由信息,保证目的域间路由标识可达. Hydra 保证了所选择的源和目的域间路由标识之间存在可达路径,数据包能够被成功传送.

(3) Hydra 支持端、传送自治系统的多宿主和流量工程技术,eFIT 只支持客户网络(端自治系统)的多宿主和流量工程技术.

Hydra 和 eFIT 都采用映射表实现多宿主与流量工程技术. 但是,eFIT 映射表保存客户网络地址到提供商地址的映射关系,而提供商地址只标识提供商网络中边界路由器的地理位置. 因此,eFIT 只能够支持客户网络的多宿主和流量工程技术. Hydra 映射表保存 IP 前缀到域间路由标识的映射关系,而域间路由标识由传送自治系统及其提供商自治系统唯一确定. 所以,Hydra 支持端和传送自治系统的多宿主、流量工程技术.

Hydra 与 eFIT 的详细比较如表 2 所示.

① 本文默认为数据包在穿越传送网时,如果它的源和目的域间路由标识之间存在可达路径,该数据包就能够被成功地传送.

表 2 Hydra 与 eFIT 的比较

		Hydra	eFIT
基本思想		隔离网络边界与核心:分级路由	隔离网络边界与核心:在寻址和路由层面隔离客户网络和传送提供商网络
可扩展性	全球路由表	基于域间路由标识,规模显著降低	基于提供商地址,规模显著降低
	路由更新报文	由 transit-edge 自治系统发起;与 Internet 相比,数目显著减少,传送网稳定性增强	由提供商网络发起;与 Internet 相比,数目显著减少,核心网络稳定性增强
核心网络地址		域间路由标识	提供商地址
是否引入映射服务		是	是
数据包是否隧道穿越核心网络		是	是
数据包是否能够被成功传送		是	不一定,存在数据包传送失败的可能
是否支持端自治系统(客户网络)使用 IP 前缀,避免更改提供商时的重新编码问题		是	是
多宿主、流量工程技术		端、传送自治系统	客户网络
安全性		边界网络不能够发起针对核心网络路由基础设施的攻击;易于部署安全路由方案;易于检测和诊断网络问题;抵抗前缀劫持和域间路由标识劫持攻击	边界网络不能够发起针对核心网络路由基础设施的攻击;易于部署安全路由方案;易于检测和诊断网络问题
数据包的源自治系统(及源传送自治系统)是否具有一定的选择数据包传输路径的能力		是	否

9 总结与下一步研究

本文提出了一个新型域间路由架构 Hydra. 它通过隔离网络边界与核心和新定义一个域间路由标识,增强了核心网络路由的稳定性,显著降低了全球路由表规模.

与相似的 eFIT 方案相比,Hydra 保证了数据包在网络中能够被成功传送,具有支持端和传送自治系统多宿主、流量工程技术,使数据包的源和源传送自治系统具有一定的选择数据包传输路径的能力等特点.

任何事物都有利有弊. 本文提出的新型域间路由架构 Hydra 增强了域间路由系统的可扩展性. 但是,它引入了一个映射表管理系统. 因此,需要展开对映射表管理系统中相关问题的研究. 例如,映射表建立机制,映射表项更新、撤销机制,映射表查询方法,映射表服务器的部署、管理等,以避免映射表管理系统成为网络新的发展瓶颈. 需要说明的是虽然目前解决 Internet 域间路由系统扩展难题的出发点和方法不同,但是引入映射表已成为学术界共识<sup>[23]</sup>. 这意味着 Hydra 面临的与映射表有关的研究问题,其它解决方案也都存在. 从而,Hydra 可借鉴其它解决方案在该方面的一些成熟的研究成果.

参 考 文 献

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

[2] Narten T. Routing and addressing problem statement. draft-narten-radir-problem-statement-02. txt. April, 2008

[3] Fuller V. Scaling issues with routing + multihoming//Proceedings of the Asia Pacific Regional Internet Conference on Operational Technologies. Bali, Indonesia, 2007

[4] Zheng C, Ji L, Pei D, Wang J, Francis P. A light-weight distributed scheme for detecting IP prefix hijacks in real-time. SIGCOMM Computer Communication Review, 2007, 37(4): 277-288

[5] Elliott K, Jennifer R. Using forgetful routing to control BGP table size//Proceedings of the 2006 ACM CoNEXT Conference. Lisboa, Portugal, 2006; 1-12

[6] Di-Fa C, Ramesh G, John H. An empirical study of router response to large BGP routing table load//Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment. Marseille, France, 2002; 203-208

[7] Agarwal S, Chuah C-N, Bhattacharyya S, Diot C. Impact of BGP Dynamics on Router CPU Utilization//Proceedings of the Passive & Active Measurement Workshop. Antibes Juan-les-Pins, France, 2004; 278-288

[8] Huston G. Auto-detecting hijacked prefixes?//Proceedings of the RIPE 50 meeting. Hanoi, Vietnam, 2005

[9] Huston G, Armitage G. Projecting future IPv4 router requirements from trends in dynamic BGP behaviour//Proceedings of the Australian Telecommunication Networks and Applications Conference (ATNAC). Melbourne, Australia, 2006; 70-75

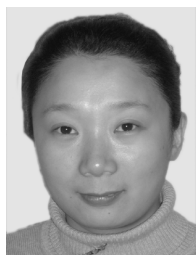
[10] Nordmark E, Bagnulo M. Shim6: Level 3 multihoming shim protocol for IPv6. draft-ietf-shim6-proto-10. txt. February, 2008

[11] O’Dell M. GSE—An alternate addressing architecture for IPv6. draft-ietf-ipngwg-gseaddr-00. txt. February, 1997

[12] Fuller D F V, Oran D, Meyer D. Locator/ID separation protocol(LISP). draft-farinacci-lisp-07. txt. 2008



- [13] Templin F. The IPvLX Architecture. draft-templin-ipvlx-08.txt. 2007
- [14] Whittle R. Ivip (Internet Vastly Improved Plumbing Architecture). draft-whittle-ivip-arch-01.txt. 2008
- [15] Zhang X, Francis P, Wang J, Yoshida K. Scaling IP routing with the core router-integrated overlay//Proceedings of the 2006 IEEE International Conference on Network Protocols. 2006; 147-156
- [16] Adan J J. Tunneled Inter-domain Routing (TIDR). draft-adan-idr-tidr-01.txt. 2006
- [17] Vogt C. Six/One: A solution for routing and addressing in IPv6. draft-vogt-rrg-six-one-01.txt. 2007
- [18] Lakshminarayanan S, Matthew C, Cheng Tien E, Mark H, Morley M, Scott S, Ion S. HLP: A next generation inter-domain routing protocol//Proceedings of the ACM SIGCOMM. Philadelphia, Pennsylvania, USA, 2005; 13-24
- [19] Xu X, Guo D. Hierarchical routing architecture (HRA). draft-xu-rrg-hra-00.txt. February, 2008
- [20] Krioukov D, Fall K, Yang X. Compact routing on Internet-like graphs//Proceedings of the IEEE INFOCOM. 2004; 219-229
- [21] Oliveira R, Lad M, Zhang B, Zhang L. Geographically informed inter-domain routing//Proceedings of the IEEE International Conference on Network Protocols 2007 (ICNP2007). Beijing, China, 2007; 103-112
- [22] Caesar M, Condie T, Kannan J, Lakshminarayanan K, Stoica I, Shenker S. ROFL: Routing on flat labels//Proceedings of the ACM SIGCOMM 2006. Pisa, Italy, 2006; 363-374.
- [23] Massey D, Wang L, Memphis U, Zhang B, Arizona U, Zhang L. A proposal for scalable Internet routing & addressing. draft-wang-ietf-efit-00.txt. 2007
- [24] Huston G. 2005—A BGP Year in Review//Proceeding of the APNIC 21 meeting. Perth, Australia, 2006
- [25] Ratul M, David W, Tom A. Understanding BGP misconfiguration//Proceedings of the ACM SIGCOMM. Pittsburgh, Pennsylvania, USA, 2002; 3-16
- [26] Lad M, Zhao X, Zhang B, Massey D, Zhang L. Analysis of BGP update surge during slammer worm attack//Proceedings of the Distributed Computing—IWDC 2003. 2003; 833-835
- [27] Oliveira R V, Lzhak-Ratzin R, Zhang B, Zhang L. Measurement of highly active prefixes in BGP//Proceedings of the IEEE GLOBECOM. 2005; 5-9
- [28] Lan W, Xiaoliang Z, Dan P, Randy B, Daniel M, Allison M, Wu S F, Lixia Z. Observation and analysis of BGP behavior under stress//Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement. Marseille, France, 2002; 183-195
- [29] Lear E. NERD: A not-so-novel EID to RLOC database. draft-lear-lisp-nerd-04.txt. 2008
- [30] Jen D, Meisel M, Wang M D L, Zhang B, Zhang L. APT: A practical transit mapping service. draft-jen-apt-01.txt. 2007
- [31] Yang X. NIRA: A new Internet routing architecture//Proceedings of the ACM SIGCOMM FDNA Workshop. Karlsruhe, Germany, 2003; 301-312
- [32] Wang N, Zhi Y, Wang B. AT: An origin verification mechanism based on assignment track for securing BGP//Proceedings of the IEEE International Conference on Communications. Beijing, China, 2008; 5739-5745
- [33] Zhang B, Kambhampati V, Massey D, Oliveira R, Pei D, Wang L, Zhang L. A secure and scalable Internet routing architecture (SIRA)//ACM SIGCOMM 2006 Poster Session. Pisa, Italy, 2006



**WANG Na**, born in 1980, Ph. D. candidate. Her research interests focus on network routing and security.

search interests focus on network routing.

**CHENG Dong-Nian**, born in 1957, professor. His research interests include network architecture, network secure protocol and network performance analysis.

**WANG Bin-Qiang**, born in 1963, professor, Ph. D. supervisor. His research interests focus on broadband information network.

**MA Hai-Long**, born in 1980, Ph. D candidate. His re-

## Background

Many studies reveal that Internet inter-domain routing system is facing a scaling challenge. The current global routing table has been growing at an alarming rate over the recent years. With the wide IPv6 deployment, the routing table size in the default free zone (DFZ) is growing dramatically. This routing scalability concern is exacerbated by an increase in users' requests for provider-independent addresses. At the

same time, the need for effective traffic engineering at both edge networks and ISPs also add potential scaling challenges to the global routing table by techniques such as announcing more specific prefixes. Overall, the Internet community is presented with a challenge to keep the global routing table scalable in face of an expected growth in the address space, an increased allocation of provider-independent address pref-

xes, and a demand for effective traffic engineering. Another major factor regarding routing scalability is the amount of update messages routers must process in real time. Because of the flat nature of the Internet routing, a routing flap to any destination can trigger routing updates to be propagated through the entire Internet. Research results have shown that the overwhelming majority of BGP updates are generated by a very small number of sources, most of them being small edge networks. Several efforts have point out that the routing scaling problem necessitates architectural changes and now is an important crossroad when changes can and must be made.

The main contribution of the paper is to propose a new inter-domain routing architecture, Hydra (Hierarchical inter-domain routing architecture). The paper found that the root cause that contributes to the rapid routing table growth is that number of IP prefix identifying location of autonomous system is uncontrollable; the root reason that leads to increasingly frequent routing updates is the flat inter-domain routing architecture. As a result, Hydra divides the inter-domain routing into two layers: a low-rank mapping layer and a

high-rank routing layer. The low-rank mapping layer is used to maintain the reachability between stub and transit autonomous systems; the high-rank routing layer is used to maintain the reachability between transit autonomous systems. Hydra consists of mapping table establishment and management protocol, routing table distribution protocol, high-rank inter-domain routing protocol and path maintenance protocol. For separating network edge and core, evaluation results indicate that in Hydra network, high-rank inter-domain routing updates are observably decreased, the stability of core network routing is enhanced. Hydra introduces an inter-domain routing identifier — RID (Routing Identifier) — to identify transit autonomous system's location. RID comprises autonomous system number and a locator. The locator is uniquely determined by transit autonomous system and its provider autonomous system. The RID address format brings about number of RID correlating with number of the transit autonomous system's providers. Evaluation results indicate that in Hydra network, global routing table is observably reduced, grows linearly and controllably.