面向地址空间分离网络的地址映射模型:AMIA

陈文龙[1)] 徐明伟[2)]

[1)](首都师范大学信息工程学院 北京 100048)

[2)](清华大学计算机科学与技术系 北京 100084)

摘 要 地址空间分离是解决互联网路由可扩展问题的有效方法,其关键技术是边缘网地址到核心网地址的映射机制.现有典型地址映射模型基于缓存映射项机制实施,其映射信息交互协议复杂,路由器对映射信息缓存的维护开销很大.而且,缓存机制中映射项查询延迟较长,明显影响到端系统用户的网络体验.文中设计了一种新型的面向地址空间分离网络的域间地址映射模型:AMIA,通过BGP协议扩展完成映射信息交互,映射项存储不带有任何缓存机制,方便实施.文中还为AMIA模型设计研制了集成PE和KMS功能的多功能路由系统,并在CERNET2中搭建实验网进行实验验证.理论分析及实验结果证明AMIA模型具有高性能、可行性及易实施等特点.

关键词 地址映射;封装;PE

中图法分类号 TP393 **DOI号**:10.3724/SP.J.1016.2012.00076

AMIA: Address Mapping Model Facing the Network with Separated Address Space

CHEN Wen-Long[1)] XU Ming-Wei[2)]

[1)](*Information Engineering College, Capital Normal University, Beijing* 100048)

[2)](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract Address space separation is an effective solution to the scalability of Internet routing, key technique of which is the address mapping mechanism from edge network to core network. Currently, all typical address mapping models are designed based on caching of mapping items, which require complex control protocol and bring great cost to PEs. Moreover, delay of address mapping lookup is so long as to severely impact the experience of network users. In this paper, we propose a novel model, AMIA (Address Mapping for Inter-AS), in which mapping messages transferring are fulfilled through BGP extension without any caching mechanism. Therefore, AMIA is very easy for practical deployment. Besides, we also develop the novel router integrating functions of PE&KMS and construct an experimental network on CERNET2 to validate AMIA model. Both theory analysis and experimental results show that AMIA is of high performance, feasibility and easy implementation, etc.

Keywords address mapping; encapsulation; PE(Provider Edge)

1 引 言

随着互联网规模的快速发展,互联网路由可扩展问题[1]被广为关注.Multi-homing和流量工程的广泛实施,导致互联网核心路由表容量快速增加,BGP核心路由表已超过300 K[2].其带来的负面影响就是:互联网核心路由器需要为此消耗大量路由计算

收稿日期:2011-08-24;最终修改稿收到日期:2011-09-22.本课题得到国家"九七三"重点基础研究发展规划项目基金(2009CB320502)、国家自然科学基金(61073166,61170209)和国家"八六三"高技术研究发展计划项目基金(2008AA01A323,2009AA01A334)资助.
陈文龙,男,1976年生,博士,主要研究方向为网络体系结构、网络协议.E-mail:chenwenlong2008@gmail.com.**徐明伟**,男,1971年生,博士,教授,主要研究领域为网络体系结构、高速路由器体系结构和协议测试.

开销和 FIB 表存储开销. 上述问题的根源就在于整个互联网一体化的地址及路由机制，这导致大量边缘网路由进入到核心网，使得核心网设备不堪重负. 目前，学术界解决路由可扩展问题的主流思路是将互联网地址分为两个空间：边缘网和核心网，从而减少核心网路由容量[3]. 根据文献[3]的分析，将边缘网地址前缀从核心网剥离后，整个互联网的路由表容量及路由更新频率将降低一个数量级. 地址空间分离的思想最早由文献[4]提出，实施的关键是建立好边缘网地址与核心网地址的映射模型，以实现端到端数据在核心网的封装传输. 当前，典型的地址空间分离模型，如 LISP[5] 和 APT[6]，都是基于核心网边界路由器(Provider Edge，PE)缓存映射项机制完成. 缓存机制最大的问题在于：端系统用户网络体验会受到较大影响，缓存机制的控制协议开销和 PE 维护开销较大，实际部署难度较大. 同时，对于无缓存的地址映射机制，如 Softwire[7] 中，PE 需要存储全网所有映射信息，难以承受.

本文设计了一种新型的面向地址空间分离网络的域间地址映射模型：AMIA(Address Mapping for Inter-AS)，实现全网的边缘网地址与核心网地址空间分离. AMIA 模型在每个自治域设置一台核心映射服务器(Kernel Mapping Server，KMS)，PE 只与本自治域的 KMS 通过 IBGP 扩展[8] 进行本自治域映射信息交互. 互联网端到端数据在核心网的传输过程通过报文封装完成. AMIA 模型信息交互通过对 BGP 协议扩展完成，映射项存储不带有任何缓存机制，简单易实施. 而且，通过对映射表存储容量的分析，论证了 AMIA 模型较好的存储性能. 本文还为 AMIA 模型设计实现了集成 PE 和 KMS 功能的多功能路由系统 MFRT(Multi-Function Router)，并在 CERNET2 中搭建网络进行实验，进一步验证了 AMIA 模型高性能、可行性及易实施等特点.

本文主要贡献包括：(1) 设计了一种无缓存机制的域间地址映射模型，克服了现有典型地址空间分离模型缓存机制对于用户网络体验影响较大的缺点；(2) 研制了集成 PE 和 KMS 功能的多功能路由系统，并测试验证了其高带宽、低转发时延等特点；(3) 基于 VegaNet[9] 完成 AMIA 模型在 CERNET2 上的真实部署并正常运行.

本文第 2 节介绍地址映射的相关研究工作；第 3 节描述 AMIA 模型具体思想和映射项存储分析以及 AMIA 模型的优劣评价；第 4 节给出 AMIA 模型的具体实施方法；第 5 节介绍 AMIA 模型在 CERNET2 上的真实部署实验及相关性能分析；第 6 节对全文工作进行总结.

2　相关研究

针对互联网的 Multi-homing 和流量工程导致的核心路由表的快速增长问题，LISP[5] 协议提出了地址空间分离模型. 它将互联网 IP 地址分为端系统标识域(EIDs)和路由标识域(RLOCs). LISP 无需对主机协议栈和除 PE 外的核心网路由器做任何修改. 数据层面，LISP 是一个简单的隧道转发机制. 核心网中，源端系统所属的 PE 路由器将封装报文发至目的端系统所在的 PE 路由器. 边缘网和核心网的数据转发过程和传统转发机制一样，关键点只是 PE 路由器的报文封装和解封装. 控制层面，主要是指映射信息的学习及存储，LISP 设定了一种分布式映射系统，并提供了 PUSH 和 PUSH/PULL 等几种不同的映射信息学习机制. 而且，LISP 定义了基于 UDP 的新型协议报文来交互映射信息. 然而，LISP 模型中映射项平均查询时间高达 3.6 s[10]，这些额外增加的访问延时对网络用户影响太大. LISP 协议较为复杂，PE 维护映射信息缓存需要较大开销，在映射信息系统方面也涉及新型设备部署，真实部署难度较大.

DAN[6] 希望将 APT 设计成一种能真正在互联网部署的地址空间分离模型. APT 模型的映射信息获取是一种 PULL/PUSH 混杂模式. 数据层面，APT 也是通过报文封装完成数据在核心网的传输. 不过，APT 定义了默认映射器 DM 来存储全部映射信息. PE 只缓存部分最近使用的映射信息，对于在缓存中没有找到映射项的报文，PE 会将其封装发送给 DM，由 DM 负责转发到对端 PE. APT 模型中，PE 和 DM 对映射信息缓存的维护代价较高，而且 DM 根据转发数据向 PE 发送映射项的机制有较大的安全问题，容易受到攻击.

清华大学提出的 Softwire[7] 架构，虽然是一种 IPv4 到 IPv6 的过渡机制，但其处理过程及实施结果却和地址空间分离模型极为相似，是地址映射模型研究的很好借鉴. Softwire 模型中，PE 通过扩展的 BGP 协议两两建立邻居关系，并相互传递 4～6 映射信息. 而且，IPv4 报文在 IPv6 网络中的传输也是通过封装机制完成. Softwire 模型的不足是 PE 需要建立太多的 BGP 邻居，而且尚不能解决全网跨域的映射信息交互.

3 AMIA 模型

方便描述，定义边缘网地址前缀为 E_PFX，核心网地址为 C_Addr. AMIA 模型关于映射信息存储及信息交互有如下基本思想：

(1) PE 只存储本自治域的映射信息；

(2) KMS 存储全网所有的映射信息；

(3) PE 只与本自治域的 KMS 交互映射信息；

(4) KMS 之间通过邻居洪泛模式学习映射信息.

定义 1(地址映射项). 地址映射项用于描述边缘网地址前缀到核心网地址的映射关系，记作“$E\_PFX \rightarrow C\_Addr$”.

定义 2(通配映射项). 任何边缘网地址都可匹配的映射项被称为通配映射项，记作“$E\_PFX_{\text{any}} \rightarrow C\_Addr$”. 假设边缘网是 IPv4 地址空间，则 E_PFX_{any} 就是一个 0 位掩码的 IPv4 地址前缀，即 0.0.0.0/0.

AMIA 模型的映射信息传播机制遵循如下原则：

(1) PE 与本自治域的 KMS 之间建立 IBGP 邻居关系，并通过扩展的 IBGP 协议进行映射信息交互；

(2) KMS 向所属自治域的 PE 通告该域所有的映射项，即映射项的 E_PFX 和 C_Addr 都属于该域；

(3) KMS 向所属自治域的每个 PE 通告一条通配映射项，通配映射项的 C_Addr 为 KMS 的地址；

(4) PE 之间不进行映射信息交互；

(5) 相邻自治域的 KMS 之间会建立 EBGP 邻居关系，并通过扩展的 EBGP 协议交互映射信息；

(6) KMS 之间相互传递各自拥有的全部映射信息，保证映射信息在不同自治域的 KMS 间逐个传递.

根据图 1 所示拓扑，通过 $PE_1 \sim PE_3$ 和 $KMS_1 \sim KMS_2$ 分析AMIA模型的映射信息传播过程. 令 PE_1、PE_2、PE_3 核心网地址分别为 C_Addr_1、C_Addr_2、C_Addr_3，它们连接了边缘网主机 H_1、H_2、H_3，各主机所属的地址前缀分别为 E_PFX_1、E_PFX_2、E_PFX_3. 令 KMS_1 和 KMS_2 的核心网地址为 C_Addr_4 和 C_Addr_5. PE 通过边缘网路由协议获得该边缘网路由信息，并针对每个路由前缀生成一条“边缘网地址前缀到核心网地址”的本地映射项. 其中，PE_1 会生成主机 H_1 所属边缘网地址前缀的映射项：$E\_PFX_1 \rightarrow C\_Addr_1$. 同理，$PE_2$ 和 PE_3 也会生成本地映射项：$E\_PFX_2 \rightarrow C\_Addr_2$ 和 $E\_PFX_3 \rightarrow C\_Addr_3$.

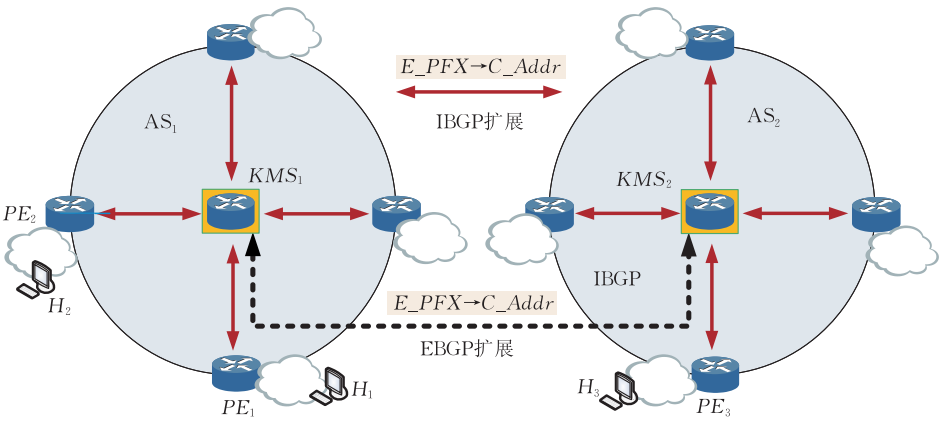

图 1 AMIA 模型网络拓扑示例

以映射项“$E\_PFX_1 \rightarrow C\_Addr_1$”为例解释上述映射信息传播机制，$PE_1$ 会把“$E\_PFX_1 \rightarrow C\_Addr_1$”通告给 KMS_1，KMS_1 会继续将该映射项转告给 PE_2 和 KMS_2. 不过，KMS_2 却不会将这种域外学习到的映射项通告给 PE_3. 映射信息传播后，各路由器所拥有映射信息如表 1 所示. 显然，映射信息传播机制达到了既定目标：“PE 只存储本域映射信息，KMS 存储全网映射信息”. 需要说明，KMS 生成的通配映射项专为 PE 服务，无需在自身存储. 对于跨域数据通信，PE 查询映射项必定匹配通配映射项，并将报文转交给 KMS 处理.

表 1 PE 及 KMS 的映射信息

角色	映射项
PE_1	$E\_PFX_1 \rightarrow C\_Addr_1$、$E\_PFX_2 \rightarrow C\_Addr_2$、$E\_PFX_{\text{any}} \rightarrow C\_Addr_4$
PE_2	$E\_PFX_1 \rightarrow C\_Addr_1$、$E\_PFX_2 \rightarrow C\_Addr_2$、$E\_PFX_{\text{any}} \rightarrow C\_Addr_4$
PE_3	$E\_PFX_3 \rightarrow C\_Addr_3$、$E\_PFX_{\text{any}} \rightarrow C\_Addr_5$
KMS_1	$E\_PFX_1 \rightarrow C\_Addr_1$、$E\_PFX_2 \rightarrow C\_Addr_2$、$E\_PFX_3 \rightarrow C\_Addr_3$
KMS_2	$E\_PFX_1 \rightarrow C\_Addr_1$、$E\_PFX_2 \rightarrow C\_Addr_2$、$E\_PFX_3 \rightarrow C\_Addr_3$

AMIA 模型中，报文转发过程分为两种情况：域内通信和域间通信. 域内通信是指通信两端在同一自治域的不同边缘网，转发过程中发送端所属 PE 路由器完成报文的封装，接收端所属 PE 路由器进行解封装. 域间通信是指通信两端分属不同自治域，由于 PE 路由器没有全网的地址映射信息，需要 KMS 路由器作转发跳板，转发过程会出现两次报文封装和解封装. 发送端主机所属 PE 路由器完成报文的第一次封装，报文送到源端所在自治域的 KMS 路由器进行第一次解封装；接着，仍是这台 KMS 路由器进行第二次封装，报文送到目的端所属 PE 路由器进行第二次解封装. PE 或 KMS 进行报文封装时，根据报文边缘网目的地址进行映射表匹配，当有多条映射项的 E_PFX 与该地址匹配时，选择 E_PFX 前缀长度最长的映射项作为匹配结果.

继续以图 1 所示拓扑及表 1 所述映射表为依据介绍地址空间分离网络中报文转发过程. 首先分析 H_1 到 H_2 的域内通信过程，转发步骤如下：

1. H_1 发送原始报文，源地址为 H_1，目的地址为 H_2；

2. 报文到达 PE_1（路由所致，PE_1 会向边缘网发布默认路由），PE_1 根据报文目的地址 H_2 查询映射表，匹配映射项“$E\_PFX_2 \rightarrow C\_Addr_2$”；

3. PE_1 对原始报文进行封装，封装报头源地址为 C_Addr_1，目的地址为 C_Addr_2，PE_1 发出封装报文；

4. PE_2 收到封装报文并进行解封装，得到原始报文，PE_2 将原始报文送达目的主机 H_2.

H_2 到 H_1 的反向通信过程与上述步骤类似，是它的逆过程. H_1 和 H_2 的双向域内通信过程如图 2 所示.

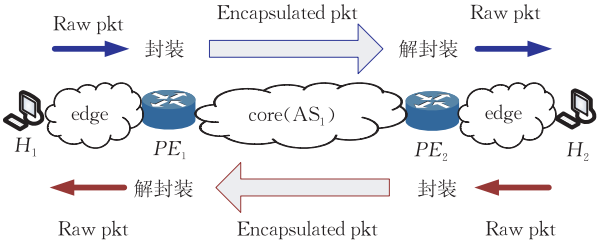

图 2　域内转发过程

接着，分析 H_1 到 H_3 的域间通信过程，转发步骤如下：

1. H_1 发送原始报文，源地址为 H_1，目的地址为 H_3；

2. 报文到达 PE_1，PE_1 根据报文目的地址 H_3 查询映射表，匹配通配映射项“$E\_PFX_{any} \rightarrow C\_Addr_4$”；

3. PE_1 对原始报文进行封装，封装报头源地址为 C_Addr_1，目的地址为 C_Addr_4，PE_1 发出封装报文；

4. KMS_1 收到封装报文并进行解封装，得到原始报文；

KMS_1 继续根据报文目的地址 H_3 查询映射表，匹配通配映射项“$E\_PFX_3 \rightarrow C\_Addr_3$”；

5. KMS_1 对原始报文进行封装，封装报头源地址为 C_Addr_4，目的地址为 C_Addr_3，KMS_1 发出封装报文；

6. PE_3 收到封装报文并进行解封装，得到原始报文，PE_3 将原始报文送达目的主机 H_3.

需要注意，H_3 到 H_1 的反向通信过程中，封装/解封装的实施点并不同于上述步骤. 其中，PE_3 进行第一次封装，KMS_2 进行第一次解封装，并继续第二次封装发送给 PE_1，由 PE_1 完成第二次解封装. H_1 和 H_3 的域间通信过程如图 3 所示.

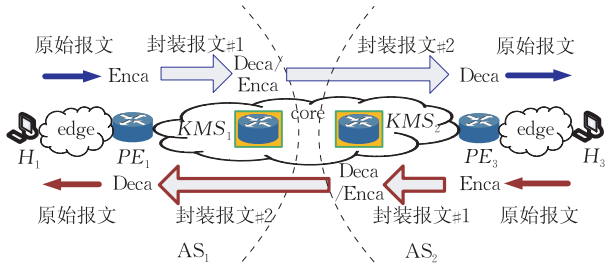

图 3　域间转发过程

AMIA 模型中，每一个 KMS 设备需要负责所属域端系统的所有域间流量的转发，负载较重，需要部署处理能力较强的高性能网关设备充当 KMS. 当然，我们还可以通过在一个域中部署多个 KMS 进行负载分摊. 负载策略可简单根据地址前若干比特位不同实施流量分摊，本文不做详细阐述.

3.1　映射项存储分析

映射项存储分析有表 2 定义的评价参数. 高端路由器的存储资源紧缺，主要针对数据层转发表项的存储，涉及 TCAM 和 SRAM 等芯片. 地址空间分离模型的映射表项也需占用上述资源，所以本节分析 AMIA 模型在数据层的硬件存储消耗. 现有非缓存方式的映射地址模型主要有 Softwire[7] 涉及的 PE 全存储模型 CSPE(Complete Storage on PE). 全网的映射表项数为所有自治域映射项总和：$\sum_{i=1}^{s}\sum_{j=1}^{E_i} P_i(j)$. 在 CSPE 模型中，所有 PE 路由器转发模块都要存储其它 PE 产生的映射项，本地产生的映射项无需在自身数据层存储. 全网的 PE 路由器数量为 $\sum_{i=1}^{s} E_i$，所以 CSPE 模型全网路由器存储的映射项总和 T_{CSPE} 满足式(1).

$$T_{\mathrm{CSPE}} = \left(\left(\sum_{i=1}^{s} E_i\right) - 1\right) \times \left(\sum_{i=1}^{s}\sum_{j=1}^{E_i} P_i(j)\right) \quad (1)$$

表 2 参数及变量定义

参数/变量	描述
s	全网自治域数量
E_i	第 i 个自治域中边缘网数量(PE 个数)
E_a	平均每个自治域中边缘网数量(PE 个数)
$P_i(j)$	第 i 个自治域中第 j 个边缘网生成的前缀路由数
P_a	平均每个边缘网生成的前缀路由数
$PE_i(j)$	第 i 个自治域中第 j 个 PE
T_{CSPE}	CSPE 模型全网存储的映射项数
T_{AMIA}	AMIA 模型全网存储的映射项数

便于分析,对自治域中边缘网数和边缘网路由数取平均值分析,则 T_{CSPE} 满足式(2).

$$T_{\mathrm{CSPE}}=s^2\times E_a^2\times P_a-s\times E_a\times P_a \tag{2}$$

AMIA 模型中,全网映射表项数仍然是 $\sum_{i=1}^{s}\sum_{j=1}^{E_i}P_i(j)$,但只有 KMS 存储全网映射项. 所以,$s$ 个自治域的 s 个 KMS 存储的映射项总数为 $s\times\sum_{i=1}^{s}\sum_{j=1}^{E_i}P_i(j)$. 第 i 个自治域的映射表项数为 $\sum_{j=1}^{E_i}P_i(j)$,每个 PE 的数据层只存储本域其它 PE 产生的映射项以及一条由 KMS 发布的通配映射项. 即 $PE_i(k)$存储的映射项:

$$\sum_{j=1}^{E_i}P_i(j)-P_i(k)+1 \tag{3}$$

所以,第 i 个自治域中E_i个 PE总共存储的映射表项数为$(E_i-1)\times\sum_{j=1}^{E_i}P_i(j)+E_i$. 同时,全网所有 PE 存储的映射表项数为$\sum_{i=1}^{s}\left((E_i-1)\times\sum_{j=1}^{E_i}P_i(j)+E_i\right)$. 所以,AMIA 模型全网所有路由器存储的映射项总和 T_{AMIA} 满足式(4),包括 PE 和 KMS 的存储映射项.

$$T_{\mathrm{AMIA}}=\sum_{i=1}^{s}\left((E_i-1)\times\sum_{j=1}^{E_i}P_i(j)+E_i\right)+s\times\sum_{i=1}^{s}\sum_{j=1}^{E_i}P_i(j) \tag{4}$$

便于分析,对自治域中边缘网数和边缘网路由数取平均值分析,则 AMIA 模型存储的映射项总和 T_{AMIA} 满足式(5).

$$\begin{aligned}T_{\mathrm{AMIA}}&=s\times E_a\times(E_a-1)\times P_a+s\times E_a+s^2\times E_a\times P_a\\&=s\times E_a^2\times P_a+s^2\times E_a\times P_a+s\times E_a-s\times E_a\times P_a\end{aligned} \tag{5}$$

继而分析 AMIA 模型和 CSPE 模型的存储映射项的差值,见式(6).

$$\begin{aligned}T_{\mathrm{CSPE}}-T_{\mathrm{AMIA}}&=(s^2\times E_a^2\times P_a-s\times E_a\times P_a)-(s\times E_a^2\times P_a+s^2\times E_a\times P_a+s\times E_a-s\times E_a\times P_a)\\&=(s\times E_a-s-E_a)\times(s\times E_a\times P_a)-s\times E_a\\&=((s\times E_a-s-E_a)\times P_a-1)\times(s\times E_a)\end{aligned} \tag{6}$$

从式(6)可以看出,只要 s、E_a两个参数大于 2,式(6)就为正值. 而且,式中 3 个参数的值越大,式(6)差值越大. 显然,真实互联网中 s 和 E_a都远大于 2,所以 AMIA 较之 CSPE 模型在映射项存储数量上有很大的减少.

由于一个边缘网产生的路由数一般为 1 或 2,图 4 给出了针对平均边缘网路由数为 1 和 2 以及自治域数分别为 10 和 100 时,两种模型在不同边缘网数量情况下需要存储映射项数的分析. 显然,无论针对哪种情形,AMIA 模型全网需要存储的映射项数都要远远少于 CSPE 模型,有较好的存储性能.

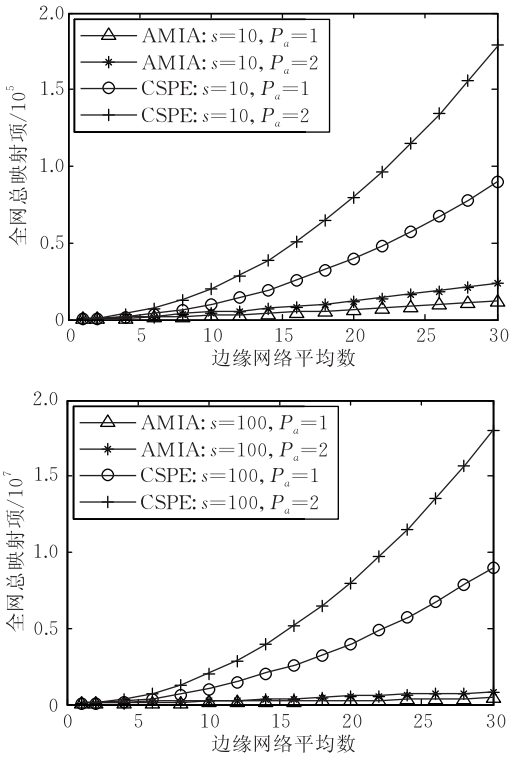

图 4 不同参数下全网映射项存储数量分析

3.2 AMIA 模型性能评价

相对现有其它方案,AMIA 模型在映射项存储性能、控制协议、表项维护、端系统用户体验、网络安全等方面都有了较大的改善,并容易在真实网络实施,其付出的代价只是少量的冗余转发路径. 下面从不同角度对 AMIA 模型与其它方案进行具体比较分析,并有表 3 所述总结.

(1)映射项存储性能. 由于一个边缘网一天的目的地址访问数一般高达 10^6 级别[10],所以缓存映射项的模型(如 LISP 和 APT)若想达到较好的转发效果,在PE端需要维护一个容量较大的映射项缓

表 3　不同模型关键性能比较

模型	存储性能	控制协议及表项维护	线速转发	通信时延	路径冗余	网络安全
APT	一般	复杂	达到	小	有(少量)	较差
LISP	一般	非常复杂	无法达到	每个连接初始流量非常大，后续较小	有(少量)	一般
CSPE	差	简单	达到	小	无	最好
AMIA	好	简单	达到	小	有	较好

存. 对于 AMIA 模型，由于一个边缘网产生的路由数通常为 1～2 条，而平均一个自治域的边缘网数不超过 100. 所以，根据式(3)可得 AMIA 模型 PE 的映射项存储量不超过 200. 由于 LISP 或 APT 都会为每个自治域部署至少一个存储全网映射项的映射服务器，等效于 AMIA 模型中的 KMS 映射项存储，所以 AMIA 模型在映射项存储总量上也少于 LISP 或 APT. 结合 3.1 节，可知 CSPE 的映射项存储性能较差. 所以，AMIA 模型在映射项全网存储或平均每个 PE 存储等方面，都具有最好的存储性能.

(2) 控制协议及表项维护. AMIA 模型和 CSPE 模型都是通过对 BGP 协议扩展完成映射信息的传递，功能简单易实施. APT 模型中 DM 间的信息交互通过一种新型的类似 OSPF 的洪泛协议完成，而且 DM 向 PE 通告映射信息过程需要有抑制机制，控制协议较为复杂. LISP 为映射信息交互设计了一种全新的控制协议，而且 LISP 有多种映射信息存储模式和获取机制，协议极为复杂，难以被广泛接受. 对于数据层映射项维护，AMIA 模型和 CSPE 模型与现有单播转发表维护类似，完全根据控制层协议的指示进行表项增、删、改. 相对而言，LISP 和 APT 由于采用映射项缓存机制，需要根据数据转发过程中映射项命中频率等参数对缓存进行维护，较为复杂.

(3) 线速转发. LISP 模型中，当 PE 缓存没有待转发报文所需的映射项时，需要数据层通告控制层发送映射信息请求并等待回应. 也就是说，LISP 模型中 PE 数据层无法为所有数据提供线速转发支持. AMIA、CSPE 和 APT 等模型都是通过查询映射项进行报文封装来完成数据在核心网的传输，根据本文第 5 节的实验，说明都能达到线速转发. 然而，APT 模型中 DM 数据层在对封装报文重新封装并转发的同时，需要向控制层提交相关信息，使其向报文源端 PE 发送映射信息. 这种数据层对控制层的反向影响，需要设备在硬件数据层付出较大的逻辑处理开销.

(4) 通信时延. 数据封装及解封装会给转发增加 μs 级的设备转发时延，但这对端到端路径转发时延来说可忽略不计. CSPE 没有任何转发路径冗余. AMIA 和 APT 可能出现转发路径冗余，但只会发生在一个自治域内部，发生在域内的转发路径冗余带来的通信时延较小，端用户不易感知. 而且，可以通过对 KMS 和 DM 优化部署来减少或完全规避这种额外时延. 然而，LISP 模型中类似 DNS 的映射信息查询，在每个连接初始流量的数据转发时，会给端用户带来明显的时延.

(5) 路径冗余. 路径冗余是指地址空间分离网络中端到端的转发路径与传统路由转发路径相比，是否在转发跳数上有所增加. LISP 和 APT 模型在 PE 没有缓存所需的映射项时，都会产生转发路径冗余，反之则不会有路径冗余. AMIA 模型中，对于自治域内部端到端通信没有任何转发路径冗余，对域间通信可能产生转发路径冗余. 不过，可以将 KMS 部署在自治域出口来减小甚至消除 AMIA 模型的转发路径冗余. CSPE 模型针对所有通信都不会产生转发路径冗余，但这是以 PE 存储全网映射项作为代价.

(6) 网络安全. 本文主要考虑用户通过发送目的地址无法命中映射表的攻击报文给网络带来的破坏性结果，如带宽消耗、产生无效映射信息交互等. CSPE 模型防御攻击能力最好，PE 拥有全部映射信息，一旦 PE 查询映射表失败则立刻丢弃报文. AMIA 模型和 APT 模型中，每个 PE 都有通配映射项，攻击报文在 PE 中会匹配通配映射项转发到 KMS，浪费了网络带宽. 当然，AMIA 模型的攻击报文最终会在 KMS 中因映射表匹配失败而丢弃. 不过，APT 中 DM 会根据报文回送映射信息的 REPLY 报文，使得 PE 缓存无效表项，造成无效存储. LISP 防范能力最差，收到攻击报文会使 PE 与映射系统通信并缓存无效表项，大量攻击报文易导致 PE 拒绝服务.

4　AMIA 模型实现设计

本文以边缘网为 IPv4 地址空间，核心网为 IPv6 地址空间为例，研究 AMIA 模型的实现机制. AMIA 模型需要在每个自治域设置一台高性能路由器作为 KMS，并让 KMS 和本域 PE 之间以及相邻域的 KMS 之间建立带有扩展能力的 IPv6 BGP 邻居关系. 传统 IPv6 BGP 连接用来传播 IPv6 地址前缀和 IPv6 下一跳信息，而 AMIA 模型需要对 BGP 进行扩展，使其能够携带“$E\_PFX \rightarrow C\_Addr$”映射信息. 本文实例中，E_PFX 是 IPv4 地址前缀，C_Addr 是 IPv6 地址. BGP 扩展主要包括两方面：

(1)扩展能力协商. AMIA 模型利用 OPEN 报文的可选参数“Optional Parameters”描述 BGP 实例的映射信息交互能力,双方都有扩展能力的邻居才能收发映射信息;(2)映射信息通告. 利用 UPDATE 报文的路径属性“Path Attributes”描述通告或撤消通告的 IPv4 地址前缀信息[8].

除了通配映射项由 KMS 直接发布,普通映射信息总是在 PE 路由器上产生,并通告给本域的 KMS,由 KMS 向外发布. PE 中映射信息是由 IPv4 路由管理将 IPv4 前缀通告给 IPv6 BGP 模块生成. 考虑实施的高效性,映射信息集成在 IPv4 转发表中存储. 地址映射项区别于普通 IPv4 转发项之处包括:(1)映射项转发标志位“M”;(2)下一跳是 IPv6 地址. 集成 PE 和 KMS 功能的多功能路由系统 MFRT 功能结构如图 5 所示,下面根据重要事件处理过程描述该设备的功能特征.

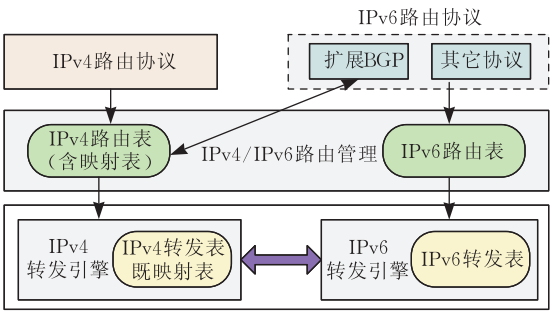

图 5 多功能路由系统 MFRT 功能结构

(1)学习到 IPv4 路由. 该事件发生在 PE,系统学习到 IPv4 路由. 一方面,按传统模式处理,生成 IPv4 转发项. 另一方面,路由管理将 IPv4 路由前缀通告 IPv6 BGP 模块生成映射信息,继而向它的 KMS 邻居发送.

(2)学习到 IPv6 路由. 该事件发生在 PE 或 KMS 设备上,完全按照传统模式处理,生成 IPv6 转发表项.

(3)学习到 4～6 映射信息. 4～6 映射信息是通过扩展 BGP 学习,该事件可能发生在 PE 或 KMS. ①若 PE 收到 4～6 映射信息,对端 BGP 邻居肯定是本域的 KMS,PE 将映射信息存储于 IPv4 转发表并设置标志位“M”,不再向外通告. ②若是 KMS 收到 4～6 映射信息,首先将映射信存储于 IPv4 转发表并设置标志位“M”,并向其它 KMS 邻居通告. 另外,KMS 还需确定是否将收到的映射信息向本域 PE 广播. 分析对端邻居角色,如果是 PE 发来的本域映射信息,则需向本域其它 PE 邻居通告. 相反,如果是 KMS 邻居发来的其它域映射信息,则无需向本域的 PE 邻居发送.

(4)收到 IPv4 报文. 一般发生在 PE. 根据最长前缀匹配原则查询 IPv4 转发表,若查得普通 IPv4 转发项则按传统模式转发报文. 若查得转发项具有封装标志“M”,则说明匹配的是一条映射项,需要进行 IPv6 封装,封装后进行 IPv6 转发表查询,并按 IPv6 查询结果发送报文. 具体过程参见图 6.

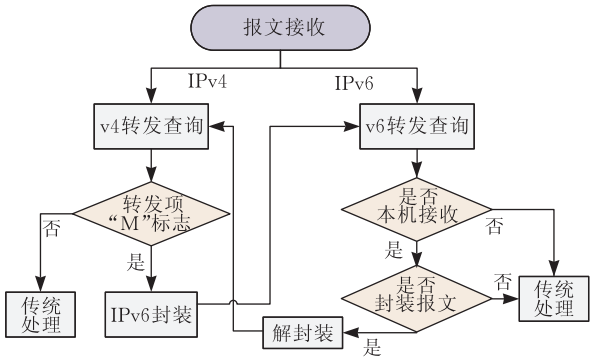

图 6 MFRT 数据处理流程

(5)收到 IPv6 报文. 事件可能发生在 PE 或 KMS. 查询 IPv6 转发表,若报文目的地址不是本地地址则正常转发. 否则,分析其是否封装报文. 对于非封装报文,则按传统模式本机接收. 对于封装报文,解封装得到内层 IPv4 报文,并进行 IPv4 转发处理. 后续 IPv4 报文处理过程和上述功能特征(4)一样,具体参见图 6.

5 实验及性能分析

本文基于 VegaNet 虚拟路由器[9],按照第 4 节所述机制,设计实现了 MFRT 原型系统. VegaNet 是一种为网络研究提供真实实验环境以及对核心网络进行模拟分析的高性能虚拟网络,它基于 CERNET2 的实施能提供一个接近于真实网络状况的网络实验环境,并能灵活支持对核心网络的模拟分析. VegaNet 的核心设备——虚拟路由器,基于真正的商业路由平台实现,支持高带宽的虚拟网络流量.

首先,在实验室通过测试仪对 PE 及 KMS 转发性能进行测试. IXIA 测试仪的两个千兆接口分别连接 MFRT 的千兆接口. 测试仪从一个接口发包,通过 MFRT 系统转发,从另一接口接收报文. MFRT 被测试的转发模式包括:模式 1,对 IPv4 报文进行 IPv6 封装再转发;模式 2,对 IPv4 in IPv6 封装报文解封装再转发;模式 3,对 IPv4 in IPv6 封装报文解封装再二次封装并转发. 测试仪发送 64 Bytes～1478 Bytes 随机大小报文,测试结果如图 7 所示. 统计 3 种转发模式平均转发能力分别为 798 Mbps、910 Mbps、895 Mbps. 由于转发性能是针

对设备接收报文统计，而模式 1 是对原始 IPv4 报文进行封装后转发，转发过程增加了报文大小. 考虑任何地址空间分离模型都要付出的封装代价，说明支持 AMIA 模型的 MFRT 系统各种转发模式基本都能达到线速转发能力.

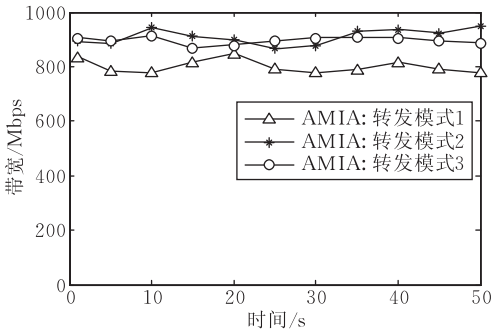

图 7　MFRT 不同转发模式的转发性能

接着，继续基于上述测试环境分析单台 MFRT 设备的转发时延. IXIA 测试仪通过一个接口发送不同字节大小的报文，并在另一接口收包以获取 MFRT 的报文转发时延，测试分别针对普通 IPv4 转发、普通 IPv6 转发和 AMIA 模型封装转发进行. 针对不同大小报文的 MFRT 转发时延如图 8 所示，相对于普通的 IPv4 和 IPv6 转发，AMIA 模型的数据转发时延会略有增加，但最高约为 80 μs. 微秒级的设备转发时延对于端到端通信至少毫秒级的路径传输时延可以忽略，所以 MFRT 的报文处理不会增加网络应用的额外延迟.

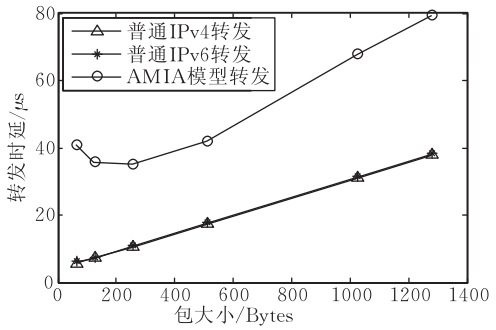

图 8　不同转发模式的设备转发时延

其次，本文利用现有 VegaNet 完成了 AMIA 模型基于 CERNET2 的真实部署，通过 8 个虚拟路由器在 CERNET2 构建了如图 9 所示的实验拓扑. 实验网由两个 AS 组成，每个 AS 包括 4 台虚拟路由器，D、E 分别充当两个 AS 的 KMS 路由器，其它 6 台设备的角色为 PE. 另外，有 4 台主机分别下连在 4 个 PE 的边缘网中. 图 9 中椭圆范围内接口配置 IPv6 地址，椭圆外部接口配置 IPv4 地址，即主机和与之相连路由器的接口为 IPv4 地址，其它均为 IPv6 地址. 所有虚拟路由器都按照本文的 AMIA 模型实现，并按图 9 中连线建立 BGP 邻居关系. 实验中，PE 上配置 IPv4 静态路由生成本地映射表项并向外通告. 分析 PE 上配置不同数量的静态路由时，全网存储的映射项数量和 PE 平均存储的映射项数量. 接着，按照图 10 所示将 KMS 撤除，其它 6 个 PE 进行 BGP 全连接，按 CSPE 模型进行映射信息交互，同样分析相关映射项存储. 实验结果如图 11 所示，无论 PE 平均存储映射项数还是全网映射项存储数，AMIA 模型相对 CSPE 模型均有减少. 结合 3.2 节分析可知，这种差异随网络规模增大越为明显. 最后，在图 9 中主机间进行了多种 IPv4 应用测试：PING、FTP、WEB 等常用网络应用程序，各项测试均能正常完成，说明了 AMIA 模型的功能可用性.

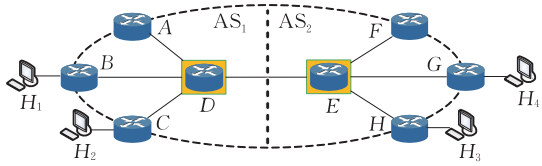

图 9　AMIA 模型实验拓扑

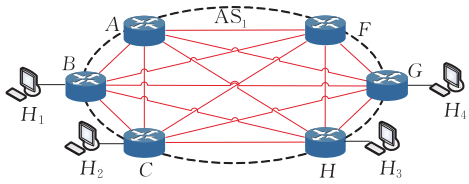

图 10　CSPE 模型实验拓扑

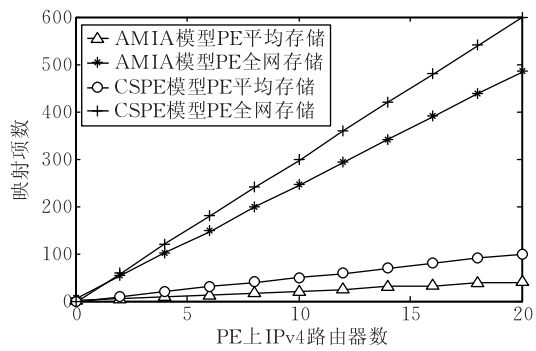

图 11　PE 映射项存储分析

6　总　结

现有地址空间分离模型尚存许多不足：控制协议复杂、映射项缓存机制需要较大的缓存维护开销、影响端系统网络用户的网络体验等. 本文设计了一种新型的面向地址空间分离网络的域间地址映射模型：AMIA，它以少量冗余转发路径为代价实现互联网边缘网地址和核心网地址的空间分离. AMIA 通

过 BGP 协议扩展完成映射信息的交互,通过隧道封装完成边缘网分组在核心网的传输. AMIA 模型无需任何缓存机制,并且具有很好的映射项存储性能.它的映射信息交互关系简单易实施,具有很好的网络攻击防范能力. AMIA 模型区别于其它相关模型的另一贡献还在于它不会影响端到端用户的网络体验. 本文基于 VegaNet 虚拟路由器,设计并研制了集成 AMIA 模型多种角色功能的多功能路由系统 MFRT. 实验表明,多功能路由系统在实施 AMIA 模型数据转发时能达到线速处理及最高 80 μs 的转发时延. AMIA 模型基于 CERNET2 进行了实验网部署,通过实验网的运行及相关实验,进一步验证了 AMIA 模型的高效存储性能及功能可用性.

致 谢 本论文工作在清华大学完成,感谢清华大学网络研究所的支持!

参 考 文 献

[1] Meyer D, Zhang L, Fall K. Report from the IAB Workshop on Routing and Addressing. RFC 4984, September 2007

[2] CIDR Report. http://www.cidr-report.org/as2.0/

[3] Massey D, Wang L, Zhang B, Zhang L. A scalable routing system design for future Internet//Proceedings of the ACM SIGCOMM Workshop on IPv6 and the Future of the Internet. 2007

[4] Deering S. The map & encap scheme for scalable IPv4 routing with portable site prefixes. Presentation, Xerox PARC, 1996

[5] Farinacci D, Fuller V, Meyer D, Lewis D. Locator/ID Separation Protocol (LISP) draft-ietf-lisp-13, 2011

[6] Jen D, Meisel M, Massey D et al. APT: A Practical Transit Mapping Service. draft-jen-apt-00.txt, 2007

[7] Wu J, Cui Y, Li X, Xu M, Metz C. 4over6 Transit Solution Using IP Encapsulation and MP-BGP Extensions. RFC 5747, 2010

[8] Bates T, Rekhter Y, Chandra R, Katz D. Multiprotocol Extensions for BGP-4. RFC 2858, 2000

[9] Chen Wen-Long, Xu Ming-Wei, Yang Yang, Li Qi, Ma Dong-Chao. Virtual network with high performance: VegaNet. Chinese Journal of Computers, 2010, 33(1): 63-73(in Chinese)
(陈文龙, 徐明伟, 杨扬, 李琦, 马东超. 高性能虚拟网络 VegaNet. 计算机学报, 2010, 33(1): 63-73)

[10] Sun Le-Tong. Study on the Mapping Mechanism in Separation Scheme for Scalable Routing. Beijing: Tsinghua University, 2011(in Chinese)
(孙乐童. 可扩展路由分离方案下映射机制的研究. 北京: 清华大学, 2011)

CHEN Wen-Long, born in 1976, Ph. D.. His research interests include network protocol and network architecture.

XU Ming-Wei, born in 1971, Ph. D., professor. His research interests include network architecture, high-performance router architecture and protocol test.

Background

This research is supported by the National Basic Research Program of China (973 Program) of China under Grant No. 2009CB320502; the National Natural Science Foundation of China under Grant Nos. 61073166, 61170209; and the National High Technology Research and Development Program (863 Program) of China under Grant Nos. 2008AA01A323, 2009AA01A334.

Along with rapid development and scale increasing of Internet, some defects of network architecture of current Internet are more obvious. Prevalent implementation of Multi-homing and traffic engineering bring on rapid increasing of kernel routing table of Internet. It takes kernel routers great overhead, including routing computing and FIB storage. Therefore, the routing system of Internet is facing severe scalability problem. The root reason of the above problem is the concentrated address and routing system. When lots of custom networks are connected to Internet through different ISPs, a mass of edge prefixes would be brought to core network of Internet. Currently, the primary solution to address the problem is to isolate the address space of edge network from core network. The main challenge of address separating scheme is the address mapping mechanism from edge network to core network.

Many researches are focused on address separating, such as LISP and APT. LISP addresses the Internet routing scaling problem by separating the current addressing into endpoint identifiers and routing locators. For LISP, there needn't change for either host protocol stacks or the core network of Internet. APT presents a tunneling architecture and new routing mechanism for the rapid growth of Internet routing table. APT partitions the Internet address space into one for transit network and one for edge networks. In this way, edge network prefixes can be removed from the routing table of the transit core. However, the address mapping models of above schemes are designed based on caching of mapping items, which require complex control protocol and bring great cost to PE routers.