一种新的基于区域编码的标记交换及网络性能分析

邱智亮, 史 琰, 刘焕峰, 陈 鹏, 刘增基
(西安电子科技大学 综合业务网国家重点实验室, 陕西西安 710071)

摘 要: 提出一种新的基于区域编码的标记交换网实现方案, 描述了标记交换网的组成、工作原理、体系结构及主要特点. 该方案将整个互联网按地理区域进行划分, 并为每个区域分配唯一的长度固定的区域编码, 在链路层帧头和 IP 分组头之间插入相应的区域编码标记, 则分组在区域标记交换网中传输时, 标记交换机将定长的目的区域编码作为分组的转发标记完成分组的转发. 理论分析和仿真结果表明, 区域标记交换网通过采用分层区域编码结构, 极大地减少了骨干交换机的路由表项数, 降低了路由存储空间和路由查找复杂度, 同时也大幅度地降低了维护路由表项的处理开销和链路传输开销, 使网络具有良好的扩展性. 该方案能有效地克服现有互联网地址空间不足、路由表过大、不能提供良好的服务质量等缺点, 简化骨干网交换设备的实现复杂度, 提高网络的吞吐率, 并能提供良好的服务质量保证.

关键词: 标记交换; 路由查找; 区域编码; 聚合增益
中图分类号: TN915. 05 文献标识码: A 文章编号: 0372-2112 (2004) 12A-039-05

A New Regional Code Label Switching and Network Performance Analysis

QIU Zhi-liang, SHI Yan, LIU Huan-feng, CHEN Peng, LIU Zeng-ji
(*National Key Lab. of Integrated Service Networks, Xidian Univ., Xi' an, Shaanxi* 710071, *China*)

Abstract: A new label switching scheme based on regional code is proposed. The framework, operational principle and main characteristics of the system are described. The scheme divides the whole Internet into regions, and every region is assigned a unique fixed length regional code. The corresponding regional code label is inserted between data link frame header and IP packet header by accessing label switching when packet accesses label switching network. The core label switching identifies destination regional code of packet to follow the packet in label switching network only. It is showed by theoretical analysis and simulation that the network can greatly reduce the number of routing entries in routers, and largely lessen link bandwidth and processor overhead for maintaining these entries. The scheme can effectively overcome the Internet shortcoming of IPv4 address space exhausted, simplify the complexity of the implementation of the core switching, enhance the system throughput, and provide good QoS.

Key words: label switching; routing lookup; regional code; aggregation gain

1 引言

目前互联网已成为人们日常生活中不可缺少的一个组成部分, 但其地址空间的匮乏、骨干路由器表项过大和有限的服务质量保证, 严重地阻碍了互联网的应用和发展. 最初设计互联网时并没有预计到该网络的应用会如此普及、如此广泛, 所以当初的互联网设计并不一定适合现在的应用需求, 故对互联网进行合理的改进是与时俱进、适应社会需求和发展的必然结果. 本文提出一种新的基于区域编码的标记交换方案, 该方案既能够继承现有互联网无连接分组传输的特点, 又能够继承快速标记交换的优点, 简化骨干交换设备的实现复杂度, 提高网络的吞吐率, 提供区分服务和服务质量保证, 能够使现有互联网向区域编码标记交换网(RCLS, Regional Code Label Switching) 平滑过渡和升级.

2 RCLS 设计思想

RCLS 将整个互联网按地理区域或行政区域进行划分, 并为每个区域分配唯一的长度固定的区域编码, 则区域内的网络设备地址由该区域编码和原来的网络地址(IPv4 地址) 共同构成[1]. 通过采用分层区域编码结构不仅能够很好地体现互联网拓扑结构的分层特性, 而且能够对物理网络起到很好的聚合作用. 由于一个区域会覆盖多个物理网络, 而骨干网交换机仅关注如何将到达的分组转发到目的区域, 即骨干网只需将分组的目的区域编码作为转发标签, 就可以完成分组的转发工作, 所以骨干交换机只需存储定长的区域编码作为路由表项, 不必象 IPv4 互联网既要存储大量的网络地址又要进行

收稿日期: 2004-07-26; 修回日期: 2004-11-06
基金项目: 国家"863" 项目资助(No. 2002AA121061, No. 2002AA103062)

复杂的地址最大匹配. 这不仅可以极大地减小骨干交换机路由转发信息表的大小和分组的处理时延, 降低交换机的实现复杂度和分组转发的操作复杂度, 还可以提高分组的转发速率和交换机的吞吐能力.

RCLS 是利用在传统 IPv4 分组头与链路层帧头之间插入一个较短的长度固定的标记来实现分组的转发, 该标记不仅包含源区域编码和目的区域编码等路由信息, 还包含分组所需的服务等级、生存期等控制信息. 由于增加了区域编码(RC), 所以设备地址应该由本地 RC 和原 IP 地址共同组成, 所以该方案实现了对现有 IP 地址的扩展, 解决 IPv4 地址空间不足问题. 由于 RC 具有分层结构, 不同层次上的交换机可以按照本层次较短的编码进行路由, 既提高转发速度, 又节约了路由表的空间.

RCLS 是对现有互联网的改造和平滑升级, 使现有网络能够平滑过渡到区域标记交换网络. 网络的演进过程, 可以分为两个阶段. 第一阶段, 骨干网支持 RCLS, 即所有的核心标记交换机(CLS)和接入标记交换机(ALS)都支持区域编码, 而在接入网内部仍采用原有的 IPv4 体制, 这可以很好地保护用户和网络运营商的投资, 此时用户终端使用的 IPv4 地址和现在因特网相同, 应该是全球唯一的, RC 由 ALS 添加. 在第二阶段, 终端使用本地 RC 及原 IPv4 地址作为设备的唯一地址, 此时接入网既需要支持 RCLS, 也需要支持 IP 交换, 此时即可实现地址空间的扩展.

3 RCLS 的组成及协议体系结构

3.1 RCLS 组成

RCLS 由标记交换骨干网和接入网组成, 如图 1 所示. 骨干网由 ALS、CLS 和区域编码解析服务器(RCS)构成, CLS 的主要功能是根据分组的目的区域标记完成分组的转发和标记转发路由表的维护及更新; ALS 不仅要完成目的 IP 地址与目的区域编码的映射, 还要完成标记产生、成帧及对入网业务的流量控制等工作. 接入网仍保持现有的 IPv4 互联网接入, 由用户终端、二层交换机和传统路由器组成, 所以标记交换网能够很好地保护用户和运营商原有投资. RCS 主要完成 IPv4 地址与 RC 的映射, ALS 可以通过 RCS 根据 IP 地址查找到相应的区域编码.

3.2 RCLS 基本工作原理

用户设备和接入网仍按现有方式工作, 通过本地的二层交换机和路由器实现本地交换, 对于出本地网的分组, 由与本地网连接的 ALS 负责为其添加区域编码、服务级别等交换标记. 本地 ALS 肯定知道源 IP 地址对应的本地 RC, 而目的 IP 地址对应的 RC 可以通过本地 RC 高速缓存映射表或类似 ARP 的方式从上级 RCS 获得, 即根据目的地址(IPv4)查询 RCS 获得相应的 RC. ALS 还可以根据用户的业务类型为用户选择服务等级, 提供区分服务, 并根据目的区域编码将分组转发给下一个 CLS. CLS 收到带有交换标记的分组后, 只需将定长的 RC 作为分组的转发标记, 根据本地的 RC 路由转发表, 将分组转发到下一个 CLS. 当分组到达目的区域的 ALS 时, 目的 ALS 剥去分组交换标记, 将标记分组还原成现在的 IP 分组格式, 并根据目的 IP 地址在目的接入网内进行分组的转发. 在 ALS 和 CLS 之间进行分组转发时, 可以根据标记中给出的业务服务级别, 为分组提供区分服务.

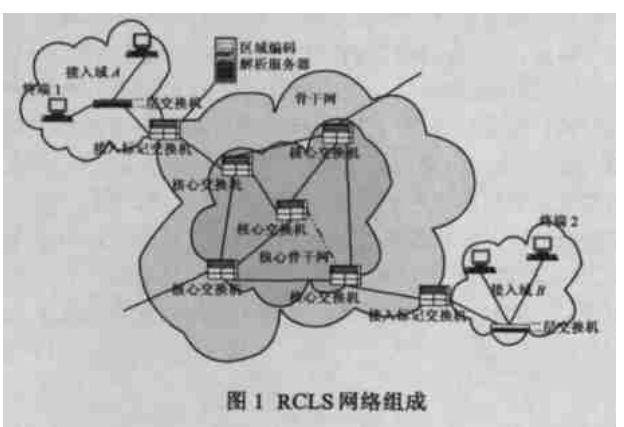

图 1 RCLS 网络组成

3.3 RCLS 协议体系结构

RCLS 的协议栈如图 2、图 3 所示, 图中给出了终端、ALS 和 CLS 的协议分层结构. 在第一阶段中 RC 标记由 ALS 和 CLS 处理, 终端的协议栈仍使用原来的协议栈. 在 ALS 设备中, 增加了 IPv4 到区域编码标记(RCL)映射处理模块; 而 CLS 设备的主要功能是标记快速转发, 因此不再处理到 IP 层, 只对 RCL 进行处理.

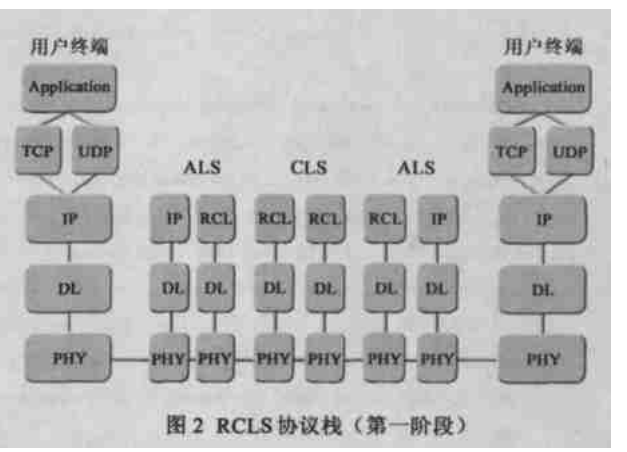

图 2 RCLS 协议栈(第一阶段)

图 3 RCLS 协议栈(第二阶段)

在第二阶段, RC 由终端自己产生并完成标记分组的成帧工作. ALS 主要根据标记中的控制域和地址域进行相应的处理和转发. CLS 的功能不变.

3.4 RCL 的组成

RCL 的设计主要考虑以下原则: (1) 能体现区域标记交换网的特点, 便于利用标记进行分组快捷转发; (2) 要能够支持特殊用户和不同业务的服务质量要求; (3) 便于扩展; (4) 支持 TCP/IP 协议族; (5) 支持拥塞控制、QoS 路由和组播等技术.

笔者设计的 RCL 由两部分组成: 路由域和控制域, 具体包括: 源区域编码、目的区域编码、版本号、TTL、服务质量编码及预留域, RCL 的长度初步定为 8 个字节, 如图 4 所示. 其中源区域编码和目的区域编码属于路由域, 用于分组选路; 其余属于控制域, 用于版本升级、保证分组的服务质量及防止分组的循环转发等控制功能.

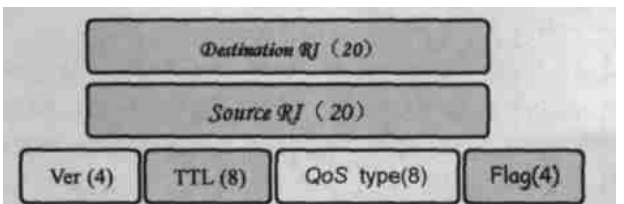

图 4 RCL 结构

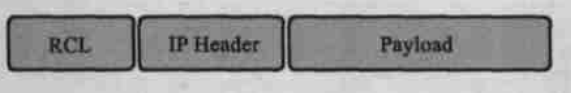

图 5 携带 RCL 的分组格式

4 区域编码对路由表规模的影响[2]

4.1 路由聚合增益的概念

网络经过区域编码后, 一个区域可以使用一条路由表项来表示到达另一区域内所有子网的路由, 因此 RC 可以减少区域间路由表项的数目. 其本质就是, RC 提供了很好的路由聚合的能力. 我们使用参数聚合增益(AG—Aggregation Gain)来恒量区域编码地址分配方法对路由表大小的影响.

$$AG=\frac{N_t-N_a}{N_t}\times 100\% \tag{1}$$

式(1)中, N_t 表示所有已经编码的区域数目, N_a 表示路由表稳定后的路由表项数目. 由于不同的区域其聚合增益不同, 对于一个网络, 有一个平均聚合增益, 为了描述方便, 在下文中就用 AG 表示网络的平均聚合增益.

CIDR[3-5] 技术也提供了某种程度的路由聚合增益, 根据有关科研机构的统计, 目前基于 CIDR 技术的 IP 地址分配的路由聚合增益约为 40%[6]. 下面, 对基于区域编码的互联网地址分配方法进行分析, 可以看出, 使用区域编码后, 网络的路由聚合增益有显著的提高.

4.2 基于区域编码的地址分配方法分析

条件: 区域编码的总长度 N bit, 区域编码的层次为 $M(1\leqslant M\leqslant N)$, 每层次编码的最小长度为 L bit, 每层次的编码长度分别为. 即存在一个样本空间,

$$S=\left\{s=(l_1,l_2,\cdots,l_m)\mid l_i\geqslant L,1\leqslant i\leqslant M,\text{且}\sum_{i=1}^{M}l_i=N\right\}$$

每个样本表明一次区域编码, 为了简化分析, 假设每个样本的取值概率是相同的.

在一次区域编码 $s=(l_1,l_2,\cdots,l_M)$ 下, 每个区域的域间路由表项数目最多为

$$N_{RE}(s)=\sum_{i=1}^{M}2^{l_i} \tag{2}$$

$N_{RE}(s)$ 是随机变量, 由上述条件, 在样本空间 S 下, 平均的路由表项的数目 $E(N_{RE})$ 为:

$$E(N_{RE})=\frac{1}{T(N,M,L)}\sum_{s\in S}N_{RE}(s) \tag{3}$$

其中, $T(N,M,L)$ 为样本空间 S 中的样本个数, 可以通过式(4)递推得到

$$\begin{cases}T(N,1,L)=1\\ T(N,M,L)=\sum_{i=l}^{N-(M-1)L}T(N-i,M-1,L),(M>1)\end{cases} \tag{4}$$

设 $T_{RE}(N,M,L)=\sum_{s\in S}N_{RE}(s)$, 其值通过式(5)递推得到

$$\begin{cases}T_{RE}(N,M,L)=\sum_{i=L}^{N-(M-1)L}2^iT(N-i,M-1,L)\\ \qquad\qquad +T_{RE}(N-i,M-1,L),(M>1)\\ T_{RE}(N,1,L)=2^N\end{cases} \tag{5}$$

在一个受限于 N,M,L 的样本空间 S 内, 其平均的路由表项的数目为

$$E(N_{Re})=\frac{T_{RE}(N,M,L)}{T(N,M,L)} \tag{6}$$

理论上, 区域编码的层次 M 可以很大(最大为 N), 但是在实际的区域划分中, 过大的 M 没有太大的意义, 表 1 中列出了 $M=1,2,3,4$ 时的 $T(N,M,L)$ 和 $E(N_{RE})$ 的值. 根据式(1)聚合增益的定义, 可以得到区域编码后的路由聚合增益为

$$AG(N,M,L)=1-\frac{E(N_{Re})}{2^N} \tag{7}$$

表 1 区域编码参数

M	$T(N,M,L)$	$E(N_{RE})$
1	1	2^N
2	$N-2L+1$	$\frac{2^{N-L+2}-2^{L+1}}{N-2L+1}$
3	$\frac{(N-3L+2)(N-3L+1)}{2}$	$\frac{2^{N-2L+4}+2^{N-2L+3}-(3N-9L+9)2^{L+1}}{(N-3L+2)(N-3L+1)}$
4	$\frac{(N-4L+1)(N-4L+2)(N-4L+3)}{6}$	$\frac{6[2^{N-3L+5}-(N^2-8LN+7N+16L^2-28L+14)2^{L+1}]}{(N-4L+1)(N-4L+2)(N-4L+3)}$

图6 表示出 $N=$ 20bit, $M=$ 2, 3, 4, $L=$ 1, …, N/M 时的平均路由表项的数目和平均聚合增益. 从图 6 可以看出, 在固定 N 的情况下, 随着 M 或 L 的增大, 平均的路由表项数目随之减少, 而路由的聚合增益可以得到大幅度的提升, 这也从理论上证明了基于区域编码的互联网地址分配方法的优点.

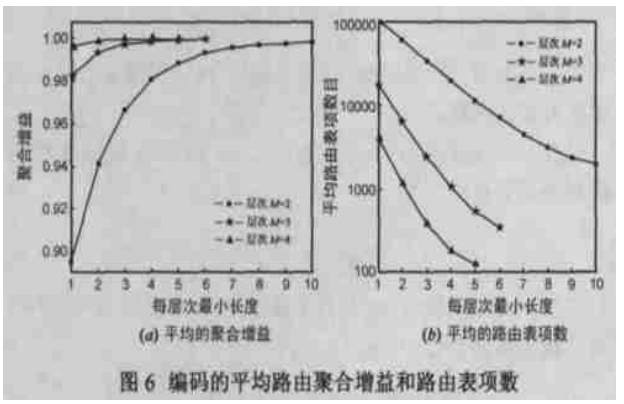

图 6 编码的平均路由聚合增益和路由表项数

4.3 RCLS 与 IP 网络路由查找性能比较[7~12]

RCLS 的骨干网络与传统 IP 网络就路由查找相比, 在操作复杂度、内存需求和更新复杂度等诸方面都具有明显优势. 由于路由查找算法的查找时间与实现复杂度往往取决于内存读取的次数, 因而一次查找中所需的内存读取次数成为衡量操作复杂度的标准. 表 2 列出了当路由表含有 40, 000 个表项时, 各种方案在所需的存储器类型、内存访问次数(包含最好和最坏两种情况) 和转发表大小等三个方面的比较. 表中 N 代表路由表中的前缀数, l 代表 RCLS 用来进行路由的那部分 RC 的比特长度.

表 2 不同方案的性能比较

算法	Nilsson	Gupta	Huang	TCAM	RCLS
存储器类型	SRAM	DRAM	SRAM	TCAM	SRAM 或 L1 Cache
内存访问次数	1/32	1/2	1/3	1/1	1/1
转发表大小	800KB	33MB	450~470KB	$O(N)$	2^l

通过比较可以看出大部分传统的 IP 查找算法要么在最坏情况下所需内存访问次数过多(Nilsson) , 难以采用硬件方法实现; 要么所需内存存储空间过大(Gupta) , 导致只能采用廉价的 DRAM. Huang 和基于 TCAM 的算法虽然在这两方面都较理想, 但 Huang 的算法在掩码长度超过 24、前缀数量超过 4000 的情况下转发表的大小会急剧上升[12] ; 而在基于 TCAM 的算法中, 转发表的大小也会随着路由表中前缀数目的增加而线性增长, 不利于扩展, 且 CAM 价格较高, 容量有限, 这都限制了该算法的应用.

相比之下, RCLS 骨干网交换机采用的是标记交换, 查找只需一次内存访问, 而不必进行复杂的最长前缀匹配, 这大大简化了交换机的操作, 提高了交换机的转发性能. 并且 RCLS 骨干网交换机所需要的存储空间小, 其所需的存储空间仅取决于区域标记的编码位数. 由于 RCLS 骨干网采用层次化的路由方式, 每次路由查找仅需要 20bit 区域编码中的一部分作为索引, 故不论哪一级交换机的转发表都可以做的非常的小(当 $M \geqslant 3$ 时, 最多仅需几十 KB 的内存空间, 远小于上述任何算法) . 这么小的转发表完全可以放在片内高速缓存(如 CPU 的一级缓存) 上, 以进一步加快转发速度.

5 路由协议性能仿真分析

基于 RC 的互联网地址分配方法可以带来极高的路由聚合增益, 并大大减少路由表项的数目. 这意味着降低交换机在传输、处理、查找和维护路由表的处理开销, 同时也减少了维护路由表的链路带宽. 本节通过计算机仿真, 对基于区域编码的路由协议[2] 与互联网中基于距离向量法的路由协议, 在路由收敛时间和路由协议引入的链路开销方面的性能进行比较. 所谓路由收敛时间是: 从网络拓扑发生变化的时刻起, 到这种拓扑变化被相关交换机获知为止, 所经历的时间. 因此, 希望路由收敛时间越短越好. 另外一个影响路由协议性能的因素是链路开销, 由于交换机之间需要周期性地交互路由信息来维持路由表项的有效性, 单位时间内传输的路由信息越多, 路由协议的链路开销越大.

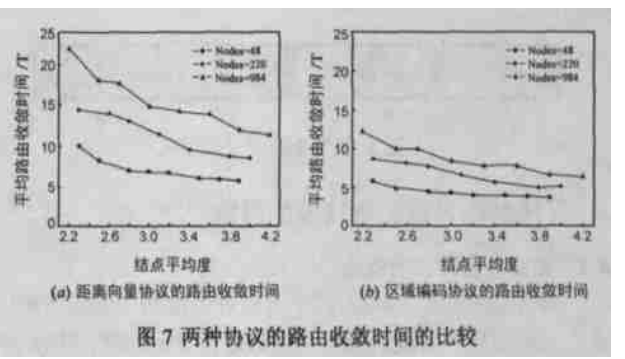

图 7 两种协议的路由收敛时间的比较

在图 7 中, 图 7(a) 表明基于距离向量的路由协议下网络的路由收敛时间, 图 7(b) 表明基于 RC 的路由协议下网络的路由收敛时间. 在图 7 中, 网络的路由收敛时间使用路由更新周期 T 的倍数表示. 从图 7 可以看出, 在相同网络拓扑下, 基于区域编码的路由协议比基于距离向量法的路由协议具有更短的路由收敛时间.

图 8 表明了三种网络规模下, 网络的平均链路开销随结点平均度变化的曲线. 图 8 中使用路由更新周期 T 内每条链路平均接收的路由表项数目来表示路由协议的链路开销.

从图 8 可以看出, 在相同的网络拓扑下, 基于区域编码的路由协议的链路开销明显小于基于距离向量法的路由协议的

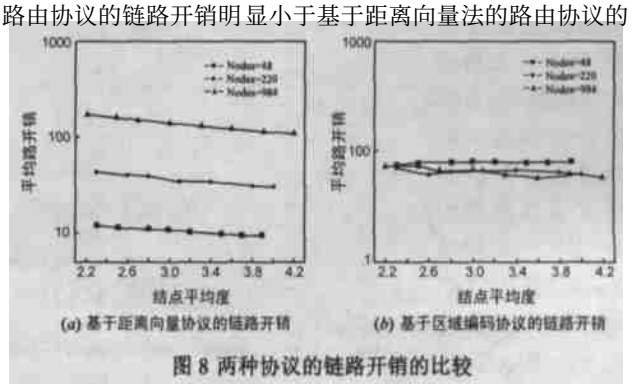

图 8 两种协议的链路开销的比较

链路开销,而且不随网络规模的变大而增加.

从图 7 和图 8 的仿真结果可以看出,相对于基于距离向量法的路由协议,基于区域编码的路由协议不仅具有较短的路由收敛时间,而且其协议的链路开销也较小,这与理论分析结果一致.

6 结束语

上述结果证明笔者提出的采用网络区域编码作为互联网分组交换标记的新思路,能够有效提高互联网性能,简化骨干交换设备的实现复杂度. 通过为不同的业务提供相应的区分服务,保证业务服务质量. 同时,解决 IPv4 地址空间严重不足的问题. 该方案是对现有因特网的维新和改良,是一个渐进改良过程,可以很好的保护现有的网络投资并具有良好的扩展性. 我们相信,通过对该技术方案全面、深入的研究,能够提出多项具有我国自主知识产权的网络技术协议和规范,如在新的标记交换思想指导下,提出新的路由算法、流控方法、区分服务方法及组播方法,争取形成若干国家标准,为我国提出和建设具有自主知识产权的新一代网络体系结构奠定基础.

参考文献:

[1] 邱智亮,游骅,刘焕峰,刘增基. 一种新的基于区域编码的标记交换技术[J]. 西安电子科技大学学报,2003,30(7):132-135.

[2] 史琰,刘增基,邱智亮. 基于区域编码的 Internet 地址分配方法[J]. 西安电子科技大学学报,2003,30(7):136-141.

[3] IETF Network Working Group. RFC1518, An Architecture for IP Address Allocation with CIDR[S]. 1993.

[4] IETF Network Working Group. RFC1519, Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy[S]. 1993.

[5] IETF Network Working Group. RFC1817, CIDR and Classful Routing[S]. 1995.

[6] Tony Bates, Philip Smith, Geoff Huston. CIDR REPORT[EB/OL]. http://www.cidr-report.org/, 2003.

[7] D R Morrison. Partricia-practical algorithm to retrieve information coded in alphanumeric[J], ACM, 1968, 15(4): 515-534.

[8] S Nilsson, G Karlsson. IP-Address lookup using LC-Tries[J]. in IEEE Journal on Selected Aeras in Communications, 1999, 17(6): 1083-1092.

[9] Lampson B, Srinivasan V, Varghese G. IP lookups using multiway and multicolumn search[J]. IEEE/ACM Transactions on Networking, 1999, 7(3): 324-334.

[10] P Gupta, S Lin, N McKeown. Routing Lookups in Hardware at Memory Access Speeds[C]. Proc IEEE Infocom, 1998, 3: 1240-1247.

[11] A Brodnik, S Carlsson, M Degermark, S Pink. Small Forwarding Table for Fast Routing Lookups[C]. Proc ACM SIGCOMM, 1997, 9: 3-14.

[12] Nerr Fu Huang, Shi-Ming Zhao. A Novel IP-Routing Lookup Scheme and Hardware Architecture for Multigigabit Switching Routers[J]. IEEE Journal on Selected Areas in Communications, 1999, 17(6): 1093-1104.

作者简介:

邱智亮　男,1965 年 5 月出生于吉林省长春市,博士,西安电子科技大学综合业务网国家重点实验室教授,主要研究领域为宽带综合业务网,ATM 接入与交换技术,高性能路由器交换技术.

史　琰　男,1975 年 1 月出生于河南洛阳,西安电子科技大学博士研究生,主要研究方向为现代通信网路由技术,QoS 保证,流量管理技术.

刘焕峰　男,1972 年 10 月出生于山东临朐,西安电子科技大学在职博士研究生,主要研究方向为宽带综合业务网流量拥塞控制.

陈　鹏　男,1978 年 8 月生于山东潍坊,2004 于西安电子科技大学通信工程学院获硕士学位,主要研究方向为宽带网络交换技术.

刘增基　男,1937 年 11 月出生于浙江丽水,西安电子科技大学教授博导,主要研究方向为宽带通信网,交换技术和光通信技术.