

Tbit 路由器双协议栈的设计与实现

王向东^{1,3}, 刘 强^{2,3}

(1 中国科学院 研究生院, 北京 100049; 2 国家数字交换系统工程技术研究中心, 河南 郑州 450002;

3 总后勤部档案馆, 北京 100842)

摘 要: 由于技术的进步和网络的日益复杂, 网络路由技术发展迅速. 自主研发开发国产高性能路由器在经济上、技术上以及安全上都具有重大的理论和现实意义. 以实际的工作为背景, 对于高性能路由器双协议栈进行了系统的分析, 并对于双协议栈的实现和协议之间的信息交换进行了详尽的阐述, 最后对高性能路由器进行了功能测试, 满足项目的需要.

关键词: Tbit 路由器; 双协议栈; 设计; 实现

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2010)03-0085-06

Design and Implementation of Dual Protocol Stack on Tbit Router

WANG Xiang-dong^{1,3}, LIU Qiang^{2,3}

(1 School of Graduate, The Chinese Academy of Sciences, Beijing 100039, China;

2 National Digital Switching System Research Center, Zhengzhou 450002, China;

3 The Archives Bureau of General Logistics of PLA, Beijing 100842, China)

Abstract: Because of the implement of the technology and the increasing complexity of the network, network routing technology development rapidly. Independent research and development of domestic high-performance router are great theoretical and practical significance in the economic, technical and security. Based on the actual work, this paper analyzed the dual protocol stack of the high-performance of the router in system, this paper described the implementation of the dual protocol stack and the exchange of information between the protocol. At last, the function tests of the high-performance router proved the design and implementation meet the needs of the project.

Key words: high-performance router; dual protocol stack; design; implementation

1 引言

随着技术的进步和网络的日益复杂, 网络路由技术发展变化迅速. 作为不同网络之间互相连接的枢纽, 路由器系统构成了基于 TCP/IP 的 Internet 主体脉络. 路由器的处理速度是网络通信的主要瓶颈之一, 其可靠性则直接影响着网络互联的质量. 因此, 在整个 Internet 研究领域中, 路由器技术始终处于核心地位. 对于 IPv6 技术, IPv6 路由器的大规模实用化将是互联网上 IPv6 时代的起点. 因此, IPv6 路由器的研究与应用水平将决定 IPv6 网络发展的进程^[1].

为此科技部启动了高性能路由器研制项目“可

扩展到 T 比特的高性能 IPv4/v6 路由器^[2]基础平台及实验系统”, 目的是研制具有自主知识产权的国产高端 IPv6 和可扩展到 T 比特的 IPv4/IPv6 双栈路由器, 并作为中国下一代因特网 CNGI 中的关键网络设备. 文中结合该课题, 主要研究 Tbit 路由器软件子系统的双协议栈的实现技术.

2 T 比特路由器结构特点

路由器技术发展主要是在包转发技术方面的进步. 对于低中端路由器, 交换路由信息、路由表查找以及转发数据报文都是由软件实现的, 所有的 IP 包都要经过中央 CPU 进行转发处理, 吞吐率比较低.

而高端路由器使用基于硬件的报文转发.包转发和路由表查找都是由 ASIC 芯片完成,软件只完成路由协议报文处理、路由表计算以及将计算好的路由表下发到转发板上等功能^[3-4].

中央处理机采用多处理器结构并行处理各类协议及设备管理维护信息,并提供系统级全局唯一的路由转发表的维护命令,集中控制各路由单元的操作和维护,确保各路由单元操作和维护的实时性和一致性,它通过系统级内部通信对所有各路由单元进行控制、管理与维护.图 1 是 T 比特路由器主控软件分层结构图,由图可见主控软件由六层组成,它们分别是硬件抽象层,网络驱动层,双协议栈 IP 层, TCP/UDP 层,路由协议,SNMP Agent、Telnet shell、Web sever 等.

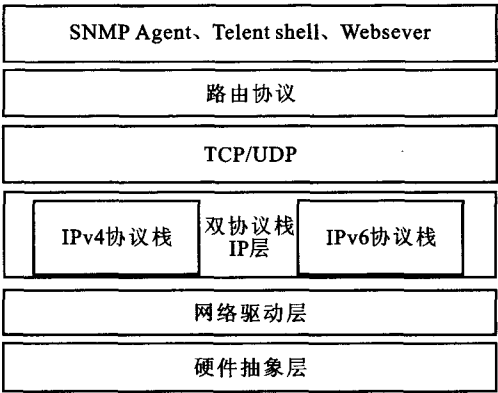


图 1 T 比特路由器主控软件分层结构图

3 双协议栈的简介

作为面向新一代互联网的核心网络路由器必须在 IP 协议的支持上实现 IPv4、IPv6 的双协议栈配置,以支持 IPv4 和 IPv6 的网络互联.因此 T 比特路由器是具有 IPv4/IPv6 双协议栈的路由器,它支持

IPv4、IPv6 以及与之相关的协议.

双协议栈工作的算法简单描述如下:

如果应用程序使用的目的地址是 IPv4 地址,那么将使用 IPv4 协议栈;如果应用程序使用的目的地址是兼容 IPv4 的 IPv6 地址,那么 IPv6 封装到 IPv4 中;如果目的地址是其他类型的 IPv6 地址,那么就使用 IPv6 地址.

双协议栈的缺省 IP 包发送算法如下所示.

- (1) 如果 IP 包的目的地址是 IPv4 地址:
目的站点在可达链路上,直接发送;
目的站点不可达,要么送往在线路由器,要么不可达.
- (2) 如果 IP 包的目的地址是 IPv4 兼容的 IPv6 地址:

目的站点在可达链路上,直接发送 IPv6 包;
目的站点处于 off-link,①有可达 IPv4 路由器,则封装在 IPv4 包中发往 IPv4 路由器;②有可达 IPv6 路由器,则不封装,直接发送;③没有可达路由器,则不可达.

- (3) 如果 IP 包的目的地址是纯 IPv6 地址:
目的站点在可达链路上,直接发送 IPv6 包;
目的站点处于 off-link,①有可达 IPv6 路由器,则直接发送到路由器;②如果目的地通过手动隧道可达,并且链路上有可达 IPv4 路由器,则封装成 IPv4 包,目的 IP 地址为隧道终点地址,链路地址为可达路由器的链路地址;③否则为不可达.

在 T 比特路由器的设计实现过程中,双协议栈 IP 层的功能需求包括三部分:IPv4 协议与 IPv6 协议部分,ICMPv4 协议与 ICMPv6 协议部分和 ARP 协议与 ND 协议部分,三个部分具体的功能需求如表 1 所示^[5].

表 1 双协议栈 IP 层的功能需求

IPv4 协议与 IPv6 协议	ICMPv4 协议与 ICMPv6 协议	ARP 协议与 ND 协议
对从外部网络进来的数据包进行合法性验证	处理接收的 ICMPv4、ICMPv6 数据包	地址解析
数据包的接收和发送	为 IP 层和高层协议产生和发送 ICMPv4、ICMPv6 差错数据包	邻居不可达性检测(仅对 ND 协议)
对 IPv4 报头中选项字段的处理(对 IPv6 扩展头的处理,包括接收数据时对扩展头的分析处理以及发送数据时扩展头的生成)	审查 ICMPv4、ICMPv6 差错数据包并提交给高层协议进行处理,例如目的不可达数据包、包过大数据包、超时数据包以及参数出错数据包等	重复地址检测(仅对 ND 协议)
数据包的分段与重组	对差错数据包的输出速率与带宽限制	路由器、前缀及参数发现(仅对 ND 协议)
路由查表	状态统计	重定向(仅对 ND 协议)
差错处理与状态统计		

4 双协议栈 IP 层的实现方案

T 比特路由器主控软件系统中的操作系统必须是嵌入式实时操作系统,以此来保证协议包处理的实时性,并将处理结果实时提供给硬件处理.因此采用 Linux 作为开放源代码操作系统做为路由器协议栈的基础.

4.1 Linux 操作系统的网络系统分层结构

基于 Linux 的双协议栈仍然采用传统的分层技术来实现.其网络系统分层结构如图 2 所示.

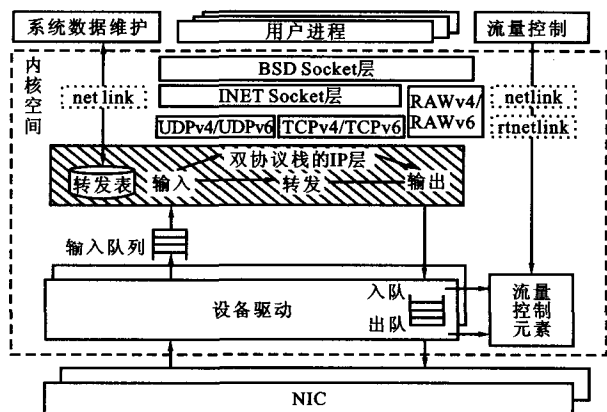


图2 Linux的协议栈网络系统分层结构图

其中,BSD socket 层提供了标准的 UNIX 套接字 API,因此一般使用普通 UNIX 中 BSD socket 通用接口编程的程序都可移植到 Linux 系统中;INET socket 层管理基于 IPv4/IPv6 协议的 TCPv4/TCPv6 与 UDPv4/UDPv6 协议通信端点,它比 IP 协议层次高,实现对 IP 分组排序、控制网络系统效率等功能;设备驱动负责从物理介质接收和发送数据.图中阴影部分——双协议栈 IP 层的实现技术即为本课题研究的主要内容.

4.2 双协议栈网络层数据传输流程

4.2.1 数据接收流程

接收数据时,由虚拟驱动程序负责接收数据包,分析协议类型并去 MAC 头,然后由软中断(softirq)处理程序唤醒数据接收任务,将接收到的数据包交给 IP 协议模块进行处理,取出 IP 包的头部信息,并根据数据包的类型交给相应的协议模块(TCP、UDP、RAW)处理,数据接收流程图如图 3 所示.

4.2.2 数据发送流程

应用程序通过 socket 编程接口使用 TCP/IP 协议栈提供的数据通讯功能.发送数据时,应用程序通过 TCP 或 UDP 协议模块将用户数据交给 IP 协议模块.IP 模块在 TCP 报文段或 UDP 数据报上添加

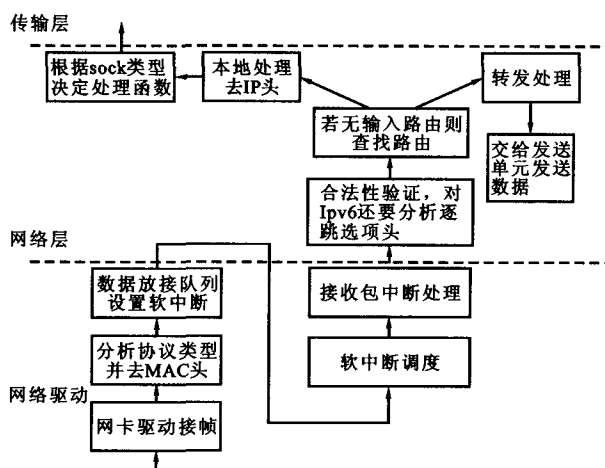


图3 数据接收流程

IP 头部(对 IPv6 还有扩展头部),封装成 IP 数据包,然后查找路由表确定路由.若找不到相应路由,则由 ICMP 模块发送相应的出错报文;若找到了路由且发送接口为以太网接口则需要启动地址解析功能找到目的 IP 地址对应的链路层地址,然后封装成以太网帧,并交给虚拟驱动程序,最后通过整个系统的外部接口发送出去,数据发送流程图如图 4 所示.

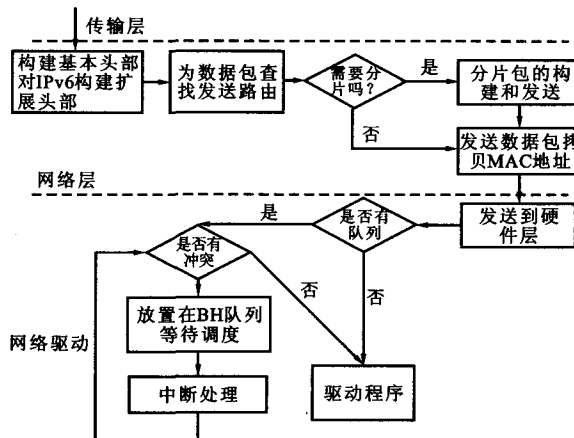


图4 数据发送流程

4.2.3 网络层总体数据流图

双协议栈的数据接收流程和数据发送流程中详细描述了从 IP 层到底层驱动之间的传输过程,下面的网络层总体数据流图简化了网络层以外的传输过程,详细描述了网络层的整个数据接收和发送的过程,如图 5 所示.

4.3 基本协议实现方案

IP 层最重要的任务是将数据从信源传送到信宿.因此,IP 协议是网络层的重要协议,为传输层提供尽最大努力的、不可靠的、无连接的服务.除此之外还应为系统间的调整提供必要的功能.所谓系统

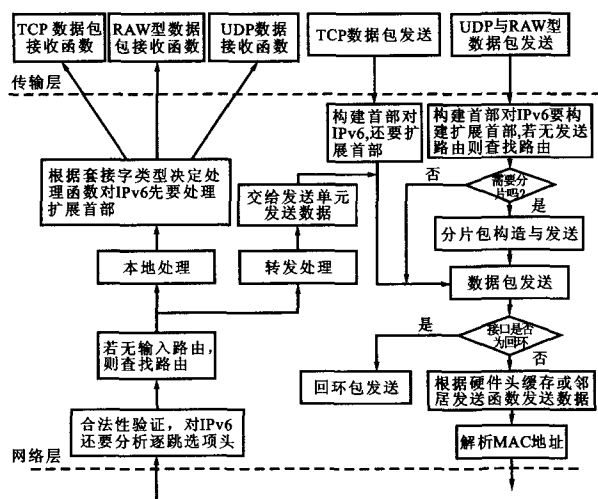


图5 总体数据流图

间的调整是指邻机系统发现、地址分配的控制和成员管理等功能,同时也提供差错报告与诊断功能。这些功能分别对应于双协议栈体系中的地址解析协议(ARP)、邻居发现协议(ND)、地址自动配置协议、组播管理协议以及控制报文协议(ICMPv4/ICMPv6)。根据系统需求,本设计仅考虑IPv4协议栈部分中的IPv4、ICMPv4与ARP和IPv6协议栈部分中的IPv6、ICMPv6与ND协议。

4.3.1 双协议栈IP协议实现方案

双协议栈IP协议完成的主要功能是接收或转发处理从网络获得的协议包和发送本地产生的协议包,对IPv6还包括接收或转发处理带有扩展头的数据包。

发送数据时,上层的TCP、UDP、RAW模块以及同层的ICMP模块的发送函数最终都将所发数据交给ip_output()/ip6_output()函数。此时,上层发送的数据是本路由器产生的各种协议包。在此过程中,调用ip_route_output()/ip6_route_output()函数查找发送路由。若找到路由则调用发送函数dev_queue_xmit()实现发送;若未找到相应路由,则丢弃需要发送的包。

接收数据时,网络驱动中的接收函数netif_rx()接收并处理由虚拟驱动程序递交的数据包,并将IP包交给IP层的入口函数ip_rcv()/ip6_rcv()进行处理。若该IP包无输入路由则需要调用ip_route_input()/ip6_route_input()函数在内核路由表中查找路由,决定需要转发(ip_forward()/ip6_forward())还是直接交给本地处理(ip_local_deliver()/ip6_input()),对于发往本地的包需要去掉IP的头部,对发往本地的IPv6数据包来说,还需要处

理它所带的扩展头。最后根据其所属的套接字类型查找协议处理链表,调用上层模块(TCP、UDP以及RAW模块)的相应接收处理函数对该数据包进行高层协议的处理。

IP协议的函数关系图如图6所示,网络层所示部分为双协议栈IP协议实现。

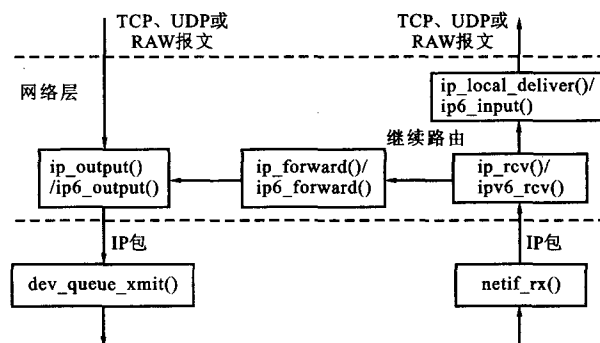


图6 双协议栈IP协议的函数关系

4.3.2 双协议栈ICMPv4/ICMPv6协议实现方案

双协议栈ICMP协议的基本功能是为IP层和高层协议产生和发送ICMP差错报文;处理接收到的各类ICMP报文;审查ICMP差错数据包并提交给高层协议进行处理。ICMP包是封装在IP包中进行传送的。

其实现比较简单,主要由两个函数组成:icmp_rcv()/icmpv6_rcv()和icmp_send()/icmpv6_send()。其中,icmp_rcv()/icmpv6_rcv()函数负责解释、处理收到的ICMPv6报文;icmp_send()/icmpv6_send()函数则负责发送ICMP报文。函数关系如图7所示,虚线的右半部分为ICMPv6协议实现。

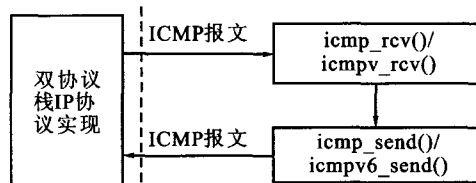


图7 双协议栈ICMP协议的函数关系

4.3.3 双协议栈ARP协议实现方案

双协议栈中的ARP协议是IPv4协议栈部分重要的协议,它为IP地址到对应的硬件地址之间提供动态映射。ARP工作时,送出一个含有所希望的IP地址的以太网广播数据包。目的地主机,或另一个代表该主机的系统,以一个含有IP和以太网地址对的数据包作为应答。ARP的功能主要包括:地址解析、邻居表的管理与维护等。

ARP协议的实现可分为三部分:ARP报文接收处理部分、ARP报文发送部分和邻居表管理部分。

ARP 报文接收处理部分处理来自网络的 ARP 分组,进行合法性检验后,根据类型作出相应的处理;ARP 报文发送部分,定义了 ARP 报文的发送函数,供 ARP 模块自身或其他协议调用;邻居表管理部分负责邻居表项的更新与维护,并定义了邻居表的各种操作函数,如邻居表项的查找、更新、增加、删除以及针对该邻居的各种形式的发送函数等。

双协议栈 ARP 协议的函数关系如图 8 所示,实线框内为 ARP 协议实现。其中 `arp_rcv()` 是接收 ARP 报文的函数,收到若为 ARP 请求报文,则发送 ARP 宣告报文并更新相应的邻居表(`arp_send()`/`neigh_update()`);若为 ARP 宣告报文,则根据 ARP 宣告报文的选项更新邻居表并将所有缓存等待的数据包发送出去(`neigh_update()`)。当 IP 层有数据需要向下层发送时,若此时没有相应邻居节点的链路层地址,则发送 ARP 请求报文(`arp_solicit()`),直至收到 ARP 回应报文获得相应的 MAC 地址信息后再将所有需要发送的数据包发送出去。

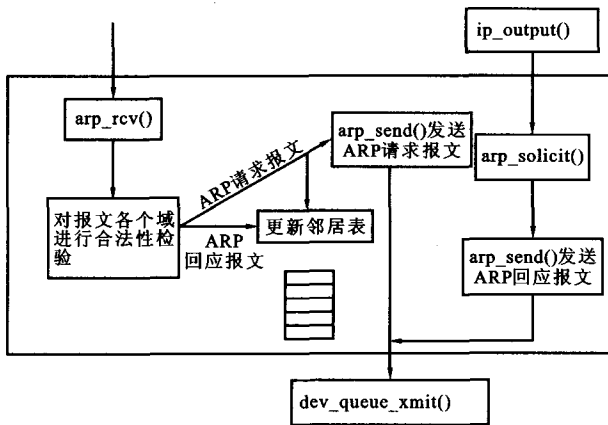


图 8 双协议栈 ARP 协议的函数关系

4.3.4 双协议栈 ND 协议实现方案

双协议栈中的 ND 协议是 IPv6 协议栈部分重要的协议,是 IPv6 所独有的协议。它使用一系列 IPv6 控制信息报文 (ICMPv6) 来实现相邻节点 (同一链路上的节点) 的交互管理。ND 协议以高效的组播和单播 ND 报文替代了以往基于广播的地址解析协议 ARP、ICMPv4 路由器发现和 ICMPv4 重定向报文,其功能主要包括:地址解析、邻居不可达性探测、路由器前缀及参数发现、重定向以及邻居表的管理与维护等。

ND 协议的实现可分为以下几个部分:ND 报文接收处理部分、ND 报文发送部分以及邻居表管理部分。其中 ND 报文接收处理部分负责从 ICMPv6

协议模块接收各种 ND 报文,进行合法性检验后,根据类型作出相应的处理;ND 报文发送部分定义了各种 ND 报文的发送函数,供 ND 模块自身或其它协议 (如 IP 和无状态地址自动配置) 调用;邻居表管理部分与 ARP 邻居表管理部分功能类似。

双协议栈 ND 协议的函数关系如图 9 所示,实线框内为 ND 协议实现。

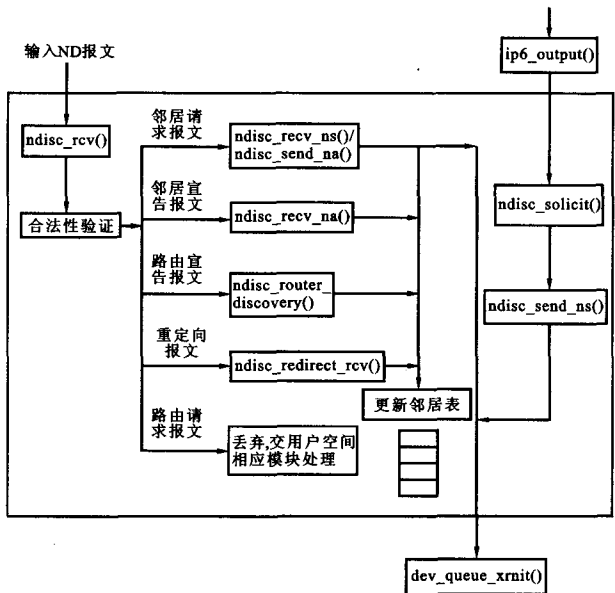


图 9 双协议栈 ND 协议的函数关系

其中 `ndisc_rcv()` 是接收 ND 报文的函数,收到若为邻居请求报文,则发送邻居宣告报文并更新相应的邻居表 (`ndisc_rcv_ns()`);若为邻居宣告报文,则根据宣告报文的选项更新邻居表并将所有缓存等待的数据包发送出去 (`ndisc_rcv_na()`);若为路由器宣告报文,则实现主机的路由器、前缀及参数发现功能 (`ndisc_router_discovery()`);若为重定向报文,则实现主机的重定向功能 (`ndisc_redirect_rcv()`);若为路由器请求报文,则进行合法性检查后丢弃,由用户空间相关模块处理。当 IP 层有数据需要向下层发送时,若此时没有相应邻居节点的链路层地址,则发送邻居请求报文 (`ndisc_solicit()`),直至收到邻居宣告报文获得相应的 MAC 地址信息后再将所有需要发送的数据包发送出去。

5 一致性测试

以上所述的双协议栈 IP 层基本协议的实现方案完成了协议栈网络层的基本功能,该部分通过了信息产业部的测试,具体的协议一致性测试结果如表 2 与表 3 所示。

表 2 IPv4 部分的协议一致性测试结果

测试项目	应测项数	实测项数	允许不支持项数	通过项数
IPv4 部分协议测试	37	36	1	36
ARP 协议的测试	7	7	0	7
IP 协议的测试	20	19	1	19
ICMP 协议的测试	10	10	0	10

表 3 IPv6 部分的协议一致性测试结果

测试项目	应测项数	实测项数	允许不支持项数	通过项数
IPv6 部分协议测试	164	162	2	162
基本协议测试	46	46	0	46
邻居发现协议测试	88	86	2	86
Path_ MTU 发现测试	4	4	0	4
ICMPv6 测试	22	22	0	22
IPv6 无状态地址自动配置测试	4	4	0	4

6 结束语

以科技部“863”课题“可扩展到 T 比特的高性能 IPv4/v6 路由器基础平台及实验系统”为背景,主要介绍了高性能路由器双协议栈的实现,通过一致性测试证明该协议栈基本满足 IETF 中规定的基本功能.不过本协议栈还有许多需要改进和增强的地方,MIPv6 等功能的扩展是此协议栈下一步的研究重点.

参考文献:

[1] 朱培栋. 高性能路由器[M]. 北京:人民邮电出版社, 2005.
[2] 李文正,郭巧. 移动 Ipv6 管理策略与仿真的研究[J]. 微

电子学与计算机,2006,23(9):173-175.
[3] Deering S, Hinden R. Internet protocol version 6 (IPv6) specification[S]. RFC 2460, 1998.
[4] Conta A, Deering S. Internet control message protocol (ICMPv6) for the internet protocol version 6 (IPv6) specification[S]. RFC 2463, 1998.
[5] Narten T, Nordmark E, Simpson W. Neighbor discovery for IP version 6 (IPv6)[S]. RFC 2461, 1998.

作者简介:

王向东 男,(1977-),硕士研究生.研究方向为智能信息处理.
刘 强 男,(1982-),博士研究生.研究方向为宽带信息网络.