文章编号:1001 - 2486(2005)05 - 0006 - 06

基于网络处理器的新型 IPv6 转发系统的设计与实现*

苏金树,时向泉,吴纯青
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:转发与控制分离结构的提出和网络处理器的发展对路由器的扩展性、灵活性、性能具有重要的影响,而 IPv6 作为下一代互联网协议的核心,是路由器研究的重要对象。简要阐述了基于转发与控制分离结构 ForCES 的 IPv6 路由器的系统结构,重点论述了基于网络处理器的 IPv6 路由器的转发结构、双栈转发系统的流程设计和隧道机制设计的实现,给出 IPv6 路由器原型系统的实际测试结果。

关键词:IPv6;转发与控制分离;网络处理器;双栈;隧道
中图分类号:TN393 文献标识码:A

The Design and Implementation of a New IPv6 Forwarding System Based on Network Processor

SU Jin-shu, SHI Xiang-quan, WU Chun-qing
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: ForCES (Forwarding and Control Element Separation) and network processor play very important roles in the router design and implementation. IPv6, the emerging standard for Internet, whose framework has been widely accepted and deployed by the Internet community. In this paper, we design an efficient framework for implementing the Ipv6 router. The approach uses the ForCES protocol to establish communication between the control plane and the data plane. To implement a high performance data forwarding system, the data plane adopts network processor to handle dual IPv4/IPv6 stacks packet forwording, tunneling. We tested the whole prototype system performance by AX4000, the result is excellent.

Key words: IPv6; forCES; network processor; daul stacks; performance

随着网络新应用、新协议的不断涌现,互联网逐渐暴露了 IP 地址不足、QoS 质量无法保障、对新协议的适应能力差以及安全可靠性不高等缺点,涌现出了许多新技术,例如 IPv6、IPQoS、MPLS 等,这些新的协议和应用需求,对互联网的核心路由器的体系结构、转发能力、扩展性和灵活性都提出了许多新的要求。为此,必须寻找一种新的路由器的实现方法,使之具有性能价格比高、可扩展性强的特点。

传统的路由器体系结构将控制和转发集成在一起,是一种紧耦合的体系结构。在这种体系结构下,控制层面和转发层面的任何变动都会牵一发而动全身,导致路由器的扩展性和软件的移植性较差。在该体系结构下,对现有系统引入新的协议和增加新的功能服务将会非常困难。而采用转发与控制相分离的体系结构,就可以较方便地对控制平面和转发平面的功能进行扩展,真正实现路由器的分离管理与控制,极大提高路由器的可靠性和可扩展性。

1 转发系统与控制系统分离的 IPv6 路由器体系结构

IPv6 作为下一代互联网协议,与 IPv4 相比,在很多方面都发生了革命性的变化,因此对于 IPv6 路由器,无论是协议栈还是路由协议都必须进行重新设计。转发与控制分离的体系结构思想使系统具有易扩展和重用的特点,使得系统软硬件的改动最小化,避免了大量软件移植的困境。采用基于转发与控制

* 收稿日期:2005 - 05 - 20
基金项目:国家重点基础研究发展计划项目(2003CB314802);国家 863 高技术研究发展计划基金项目(2003AA115130);国家自然科学基金项目(90104001)
作者简介:苏金树(1962—),男,教授,博士,博士生导师。

分离的 ForCES(Forwarding and Control Element Separation)体系结构,使得路由器能保证系统具有高性能的同时,还具有灵活、可扩展和快速开发等特性。ForCES 是 IETF 提出的标准的转发与控制分离协议[1]。图 1 是我们提出的一种基于转发与控制分离体系结构的双栈路由器体系结构。

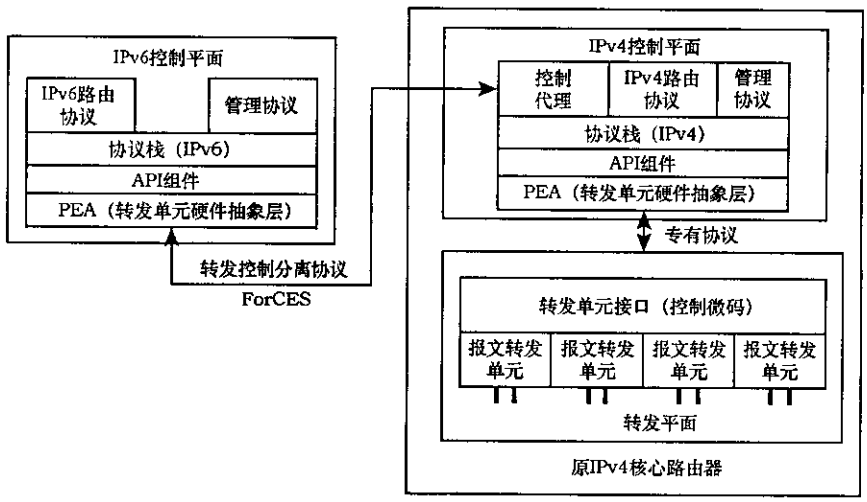

图 1 基于转发与控制分离体系结构的双栈路由器体系结构

Fig.1 The architecture of forCES based dual stack router

该结构可以使 IPv6 与 IPv4 得到有机的融合,又能充分利用原有的 IPv4 基础,所有与 IPv6 路由协议相关的处理全部放在外置的 IPv6 主控系统上处理。在外置 IPv6 服务器和原来的路由器的 CP 之间遵循标准的协议和开放的接口 ForCES,在这种结构下,可以把原来的 IPv4 路由器整体(包括主控和转发系统)看做一个转发单元,外置的 IPv6 主控系统和转发单元之间使用标准的 API 和消息格式进行交互。本结构具有很好的扩展性、灵活性。也可以采用一台主服务器对多台核心路由器进行控制,组成一个路由器集群。

2 基于网络处理器的高性能转发系统结构

网络处理器是典型的片上芯片系统,内部由若干个微处理器和若干硬件协处理器组成,网络处理器内部的多个微处理器并行处理,通过预先编制的微码来控制处理流程。而对于一些复杂的标准的操作(如内存操作、路由表查找算法、QoS 的拥塞控制算法、流量调度算法等)则采用硬件协处理器来进一步提高处理性能。由于网络处理器一般采用多线程和流水线的体系结构,这样可以对报文进行并行处理,具有很高的转发性能,此外由于网络处理器还具有可编程特性,这使网络处理器可以对报文进行深度的分析处理,以满足新的应用和协议的需求,例如对 IPv6、CPOS 以及 RPR 等协议软件进行快速的开发。

图 2 是采用网络处理器进行 IPv6 报文转发的路由器转发系统结构,每个转发板上都嵌入一个网络处理器,负责对 IPv6 和 IPv4 数据报文进行高速转发处理。输入报文经过网络处理器处理,查找转发表,确定下一跳报文的目的地址,然后通过交换开关转发到输出转发板进行输出处理。选用的网络处理器具有快速、功能强大、扩展性好的特点[2]。该处理器具有 16 个可编程微码处理器,提供 2128MIPS 处理能力,具有嵌入 PowerPC 微处理器核,同时可提供硬件加速器,用于 CRC 校验和生成、树搜索、资源控制等。其体系结构如图 3 所示。

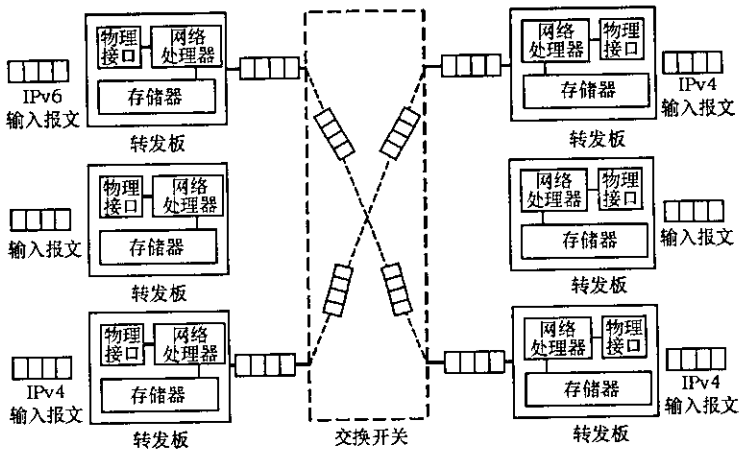

图2 基于网络处理器的IPv6路由器转发系统结构

Fig.2 The architecture of newtork processor based IPv6 forwarding system

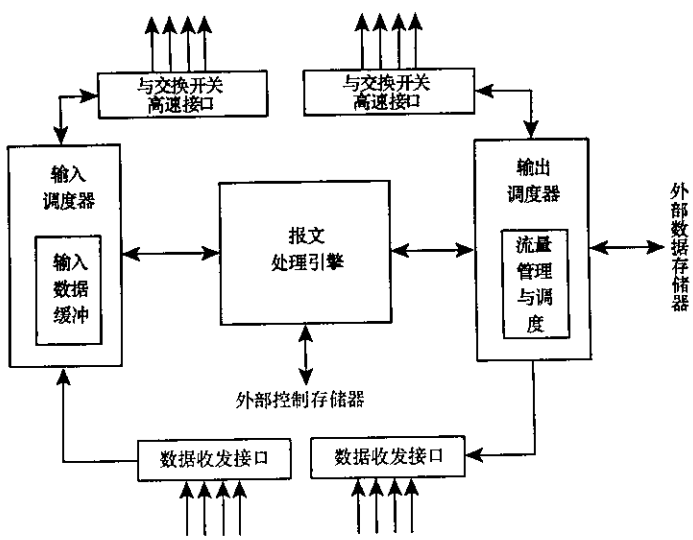

图3 选用的网络处理器体系结构

Fig.3 The architecture of network processor we used

3 基于网络处理器的转发系统实现

3.1 双协议栈的转发处理

IPv4/IPv6协议的报文由硬件分类器进行识别,由于选用的网络处理器对IPv6报文不能进行自动识别,需要对其硬件分类器进行扩展。通过协议域确定是否是IPv6报文。报文分类之后,进入各自的处理模块进行转发处理,如图4所示。此外还可能要对属于IPv4封装的IPv6报文进行解封装处理,以及对IPv6报文进行IPv4报文的封装处理。IPv6和IPv4的路由转发表相互独立,采用基于Patricia树算法进行管理与查找。具体执行动作由内置在网络处理器中的树搜索引擎实现。

3.2 IPv6报文转发处理流程

图5是一个典型的IPv6报文到IPv6报文的转发处理流程。

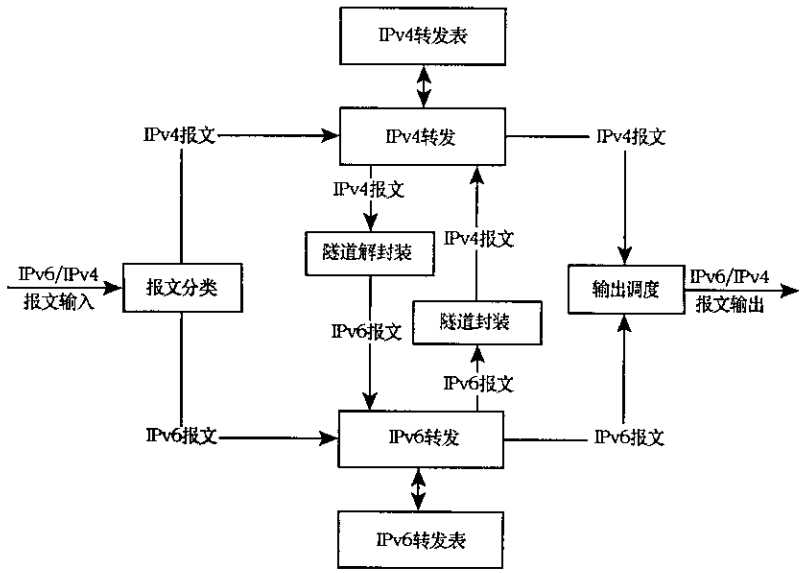

图 4 双栈报文转发处理流程

Fig.4 The data flow of dual stack forwarding system

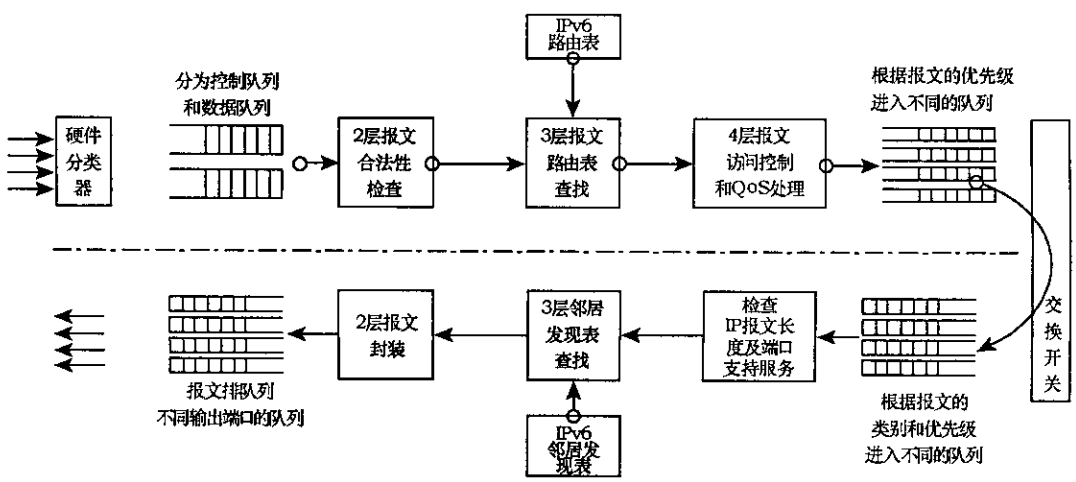

图 5 IPv6 报文到 IPv6 报文的转发处理

Fig.5 The pure IPv6 packet flow of forward system

一个以太网封装的 IPv6 报文经过千兆物理端口进入路由器,首先由网络处理器的硬件分类器进行报文类型的识别,同时报文的头部 64 字节被缓存到输入端的缓冲区进行进一步处理。首先进行二层封装的合法性检查,对于正常报文,抽取目的 IPv6 地址使用最长前缀匹配算法进行转发表的查找,根据查找结果确定目标转发板和物理输出端口号,进行 TTL 的修改和 QoS 以及一些访问控制处理,最后根据上述处理的结果确定报文的排队队列,通过交换开关转发到目标板进行输出处理。在目标板上首先对从交换开关传输过来的报文进行组装和检查,对不同的报文分派到不同的输入队列等待网络处理器的处理。首先对报文的长度和端口支持的服务进行检查,通过检查的报文根据下一跳的 IPv6 地址进行邻居发现表的查找,获得下一跳地址的 MAC 地址,然后进行二层报文的封装。对于 IPv6 报文,不需要路由器进行报文的分片和重组。最后报文被排到一个具体的物理端口的输出队列,等待硬件的调度输出[3,4,6]。

3.3 隧道机制实现

隧道处理过程包括三步 封装、解封装和隧道管理。两个隧道端节点通常是双栈节点,进行封装和解封装操作。隧道处理流程如图6所示。

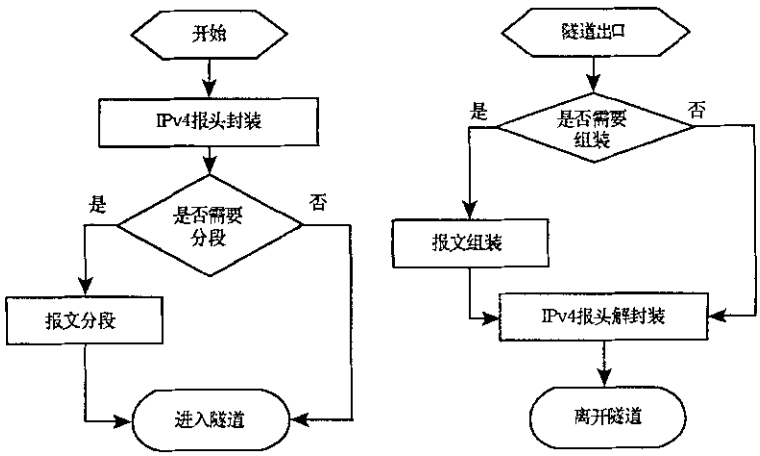

图6 隧道处理流程

Fig.6 The data flow of IPv6 tunnel

在隧道入口点,要求转发的IPv6报文到达隧道入口节点(通常是一个双栈路由器),网络处理器首先创建IPv4报头,进行IPv4报头封装,其中,IPv4报头的源地址和目标地址根据隧道配置确定。由于IPv6的路径MTU和IPv4的路径MTU存在差异,IPv6报文进入IPv4后可能要进行分段。算法判断是否需要分段,若需要则依据IPv4方法分段,分段报文进入隧道 否则封装直接进入隧道。

在隧道出口点,算法首先进行必要的IPv4报文组装,然后去除IPv4报头,获得IPv6报文。对IPv6报文进行必要的操作,例如跳限制域减1,然后按照正常的IPv6处理方法发送报文[5]。

4 原型系统实现结果

使用业内标准测试工具AX4000对我们实现的IPv6核心路由器原型系统的转发性能进行测试,测试接口为千兆以太网,测试方式为双端口全双工交叉测试。

表1和图7给出部分测试结果。

表1 测试结果

Tab.1 Test results

报文大小	吞吐率%	比特率 Kbit/s	帧/秒 packet/s
64	80	609 529.170	1 190 487
78	93.5	744 190.139	1 192 612
128.00	100.00	864 866.182	844 600.0
256.00	100.00	927 541 883.00	452 901.0
512.00	100.00	962 411 897.00	234 964.0
1024.00	100.00	980 849 004.00	119 733.0
1280.00	100.00	984 620 938.00	96 154.0
1518.00	100.00	985 727 952.00	81 170.0

测试结果表明利用我们前面提出的高性能路由器设计方法设计的路由器原型系统可以达到高性能路由器的要求。由于采用分离的体系结构,软件开发速度快,IPv4/IPv6两个路由系统软件相互独立,便于系统的升级和扩充。

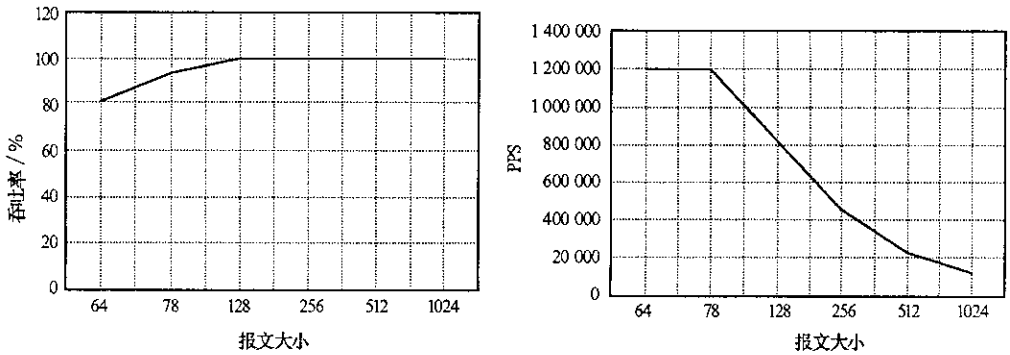

图 7 IPv6 核心路由器的性能测试结果

Fig.7 The test performance of IPv6 router

5 结 论

转发与控制分离结构和网络处理器技术对于路由器的实现具有重要的意义。从前文可以看出,使用上述技术实现的核心 IPv6 路由器在具有很好的扩展性、灵活性、可编程特性的同时,其性能也同样十分出色。我们下一步的工作就是深入研究更加适合 IPv6 的路由表查找算法,对原有 IPv4 主控系统和当前 IPv6 主控系统整合起来,对系统进行进一步的优化。

参 考 文 献:

[1] Yang L. Intel Corp. Dantu R, et al. Forwarding and Control Element Separation (ForCES) Framework [EB]. http://www.faqs.org/rfcs/rfc3746.html, 2004.

[2] IBM Corporation. IBM PowerNP. NP4GS3 Network Processor [EB]. http://www.ibm.com/chips, 2002.

[3] Deering S, Cisco, Hinden R, et al. Internet Protocol, Version 6 (IPv6) Specification [EB]. http://www.faqs.org/rfcs/rfc2460.html, 1998.

[4] Hinden R, Nokia, Deering S, et al. IP Version 6 Addressing Architecture [EB]. http://www.faqs.org/rfcs/rfc2373.html, 1998.

[5] Conta A, Lucent Technologies Inc., Deering S, et al. Generic Packet Tunneling in IPv6 Specification [EB]. http://www.faqs.org/rfcs/rfc2473.html, 1998.

[6] Crawford M, Fermilab. Transmission of IPv6 packets over Ethernet Networks [EB]. http://www.faqs.org/rfcs/rfc2464.html, 1998.

(上接第 5 页)

[6] Sklower K. A Tree-based Routing Table for Berkeley Unix [D]. University of California, Berkeley, 1993.

[7] Morrison D R. Patricia: Pratical Algorithm to Retrieve Information Coded in Alphanumeric [J]. Journal of ACM, 1968, 15(4): 514 – 534.

[8] Srinivasan V, Varghese G. Fast IP Lookups Using Controlled Prefix Expansion [J]. ACM Transactions on Computer Systems, 1999, 17(1): 1 – 40.

[9] Nilsson G. IP Address Lookup Using LC-Tries [J]. IEEE Journal on Selected Areas in Communications, 1999, 17(6): 1083 – 1092.

[10] Gupta P, et al. Routing Lookups in Hardware at Memory Access Speeds [A]. IEEE Infocom [C], 1998.

[11] Ruiz-Sanchez M A, et al. Survey and Taxonomy of IP Address Lookup Algorithms [J]. IEEE, Network, 15: 8 – 23, Mar./Apr, 2001.

[12] DegerMark M, et al. Small Forwarding Tables for Fast Routing Lookups [A]. ACM Sigcomm 97 [C], 1997: 3 – 14.

[13] Minkenberg C, et al. A Combined Input and Output Queued Packet-switched System Based on PRIZMA Switch-on-a-Chip Technology [J]. IEEE Communications Magazine, December 2000.

[14] McKeown N, et al. The Tiny Tera: A Packet Switch Core [J]. IEEE Micro Magazine, Jan. – Feb., 1997: 26 – 33.

[15] Anderson T E, et al. High Speed Switch Scheduling for Local Area Networks [R]. Digital Research Paper No. 99, Apr. 26, 1993.

[16] Tamir Y, et al. Symmetric Crossbar Arbiters for VLSI Communication Switches [J]. IEEE Transaction on Parallel and Distributed Systems, 1993, 4(1): 13 – 27.

[17] McKeown N. Fast Switched Backplane for a Gigabit Switched Router [R]. Cisco Systems, 2002.

[18] Dally W J. Scalable Switching Fabrics for Internet Routers [R]. Avici Systems Inc., White Paper, 1999.

[19] Chang C S, et al. Load Balanced Birkhoff-von Neumann Switches, Part I: One-stage Buffering [J]. Computer Comm., 2002, 25: 611 – 622.

[20] Keslassy I, et al. Scaling Internet Routers Using Optics [A]. ACM SIGCOMM [C], 2003: 189 – 200.

[21] Pappu P, et al. Distributed Queueing in Scalable High Performance Routers [A]. IEEE INFOCOM [C], 2003.

[22] Katevenis M, et al. Variable Packet Size Buffered Crossbar (CICQ) Switches [A]. IEEE ICC [C], 2004.