

高性能路由器硬件抽象层的设计与实现^{*}

吴美娟 魏进武 陈庶樵 岳 俭

(国家数字交换系统工程技术研究中心 郑州450002)

摘 要 分布式体系结构已成为高性能路由器普遍采用的实现结构,然而迫切需要解决的问题是如何实现独立于上层软件及底层硬件的通用性支撑软件体系结构。本文针对高性能路由器体系结构及功能特点,提出了硬件抽象层(HAL)的设计方案并予以实现。该方案在主处理单元上模拟从处理单元线路接口行为,由系统管理模块维护从处理单元的正常工作的。测试结果表明,HAL在轻、重负载下,其效率及可靠性均能满足高性能路由器的需求,且其功能可满足软件体系结构的可扩展性需求。

关键词 网络,路由器,硬件抽象层,分布式结构

Design and Implementation of Hardware Abstraction Level of High-Performance Router

WU Mei-Juan WEI Jin-Wu CHEN Shu-Qiao YUE Jian

(National Digital Switching System Engineering & Technology R&D Center, Zhengzhou 450002)

Abstract High-performance router is normally designed on distributed architecture. However, how to implement a versatile support software architecture which is independent of its upper layer software is a problem that must be resolved. This paper presents a hardware abstraction level design method and implements it. The method simulates actions of the interface units of subcontrol unit, and uses system manage module to maintain the work of subcontrol unit. The results of testing show that the efficiency and security of this method can satisfy the need of high-performance router, and its function can meet extensible need.

Keywords Network, Router, Hardware abstraction level (HAL), Distributed architecture

1 引言

1.1 高性能路由器体系结构的特点

随着互联网技术的空前发展和网络带宽的不断提高,骨干网络核心路由器的体系结构也发生了一些变化。近年来,国际上高性能路由器体系结构的研究和大部分商用高端路由器的实现方案中,普遍采用了分布式的体系结构^[1,2]。

分布式体系结构基于路由与转发分离的思想,使用了两级处理器,由一个主处理单元和若干个从处理单元构成,主处理单元与从处理单元间的数据交换采用了网络交换的方式。主从处理单元之间的连接关系如图1所示,主处理单元负责实现设备管理、路由计算等功能,并负责将转发表下发到从处理单元。从处理单元上有多个接口单元(IU),可根据主处理单元下发的转发表独立进行路由转发。由于从处理单元的查表转发是由硬件实现的,因此可以达到线速转发的要求。

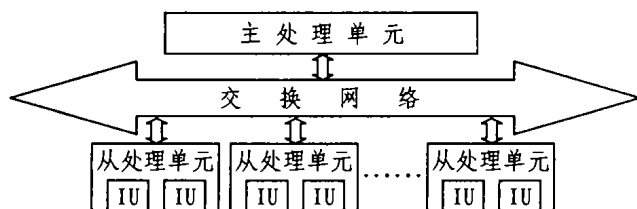


图1 高性能路由器中主从处理单元连接关系图

分布式结构为高性能路由器实现线速转发提供了必要条件,但同时也给路由器软件体系结构的设计与实现带来了迫切需要解决的一些问题。一是主、从处理单元之间必须实现可靠、高效的通信,否则将严重影响路由器上层软件的运行以及整机性能;二是如何设计一种可扩展的且不依赖于硬件具体实现方式的软件体系结构。因为,现代网络路由器除了提供基本的路由转发功能之外,诸如服务质量(QoS)保证、区分服务(DiffServ)、主动网络支持^[3]等功能需求也日益迫切。同时可用于实现路由器的硬件也日益广泛,例如网络处理机^[4,5]、可编程线卡、网络接口卡(NIC)等,且硬件体系结构的实现方式也不断更新。也就是说,路由器的上层软件和底层硬件的实现方式都呈现出多样化的趋势^[3]。如何提供一种简单高效的管理该体系结构的手段,使其对路由器软件的影响程度尽可能地减少,是高性能路由器设计中普遍面临的问题。

1.2 硬件抽象层的功能与意义

针对上述高性能路由器的实现所面临的问题,我们设计并实现了一种介于底层硬件与上层软件之间的硬件抽象层(HAL)。它包括虚拟驱动、系统管理和内部通信三大模块,在整个路由器系统中的位置如图2所示。

设计该硬件抽象层,主要目的是为了在高性能路由器中实现如下几点特性:

1. 适应分布式的路由器体系结构:将高性能路由器分布式转发和集中式处理的两大特点统一起来,对上层软件屏蔽

^{*} 基金项目:“十五”863计划信息技术领域重大项目(No. 2001AA121011)资助课题;河南省自然科学基金(No. 0111061300)资助课题。吴美娟 硕士生,研究方向为路由器软件体系结构。魏进武 博士生,研究方向为下一代网络体系结构及网络流量分析。陈庶樵 博士,研究方向为 Internet 服务质量保证及高端路由器体系结构。岳 俭 教授,研究方向为操作系统及软件体系结构。

了底层硬件的细节。既方便了主处理单元管理控制整个路由器系统,又给予从处理单元一定的独立转发空间。

2. 开放性好:硬件抽象层的出现,给上层软件提供了统一的接口,方便了协议软件等上层应用程序的升级和移植,对于提高路由器基础平台的开放性具有重要意义。

3. 可扩展性好:两级以上的多级处理单元也可采用类似的抽象层软件进行分级管理。

4. 符合标准化设计需求:对上层软件提供标准的编程接口。

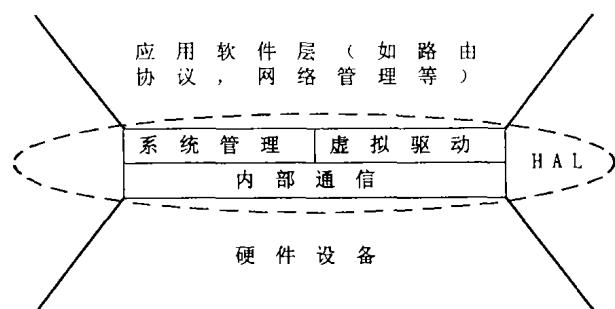


图2 硬件抽象层在路由器体系结构中的位置

本文针对高性能路由器体系结构的特点,提出了硬件抽象层的设计与实现方案。该方案在主处理单元上实现了与从处理单元线路接口对应的虚拟网络接口驱动程序,来完成对从处理单元线路接口行为的模拟;设置系统管理模块维护从

处理单元的正常运转,并且负责下发及实时更新转发表,保证从处理单元实现独立转发。

2 HAL 的总体结构及关键模块

HAL 的作用就是要将应用软件和底层硬件实现逻辑上的隔离,使应用软件通过控制 HAL 来达到操纵底层硬件的目的。在整个路由器体系结构中,HAL 介于主处理单元的应用软件和从处理单元(包括其上的线路接口单元)之间。路由器上运行的应用软件可分为两大类,一类为协议软件,一类为网络管理软件,因此 HAL 中相应地设置了虚拟驱动模块和系统管理模块,供上层应用软件使用,其中协议软件通过虚拟驱动间接管理线路接口单元的收发包动作,网络管理软件通过控制系统管理来对从处理单元及其线路接口进行管理配置及监控。此外,高性能路由器的分布式体系结构决定了主从处理单元间信息的交换必然需要内部通信模块。

基于上述考虑,HAL 分为内部通信、虚拟驱动和系统管理三大模块,其运行环境如图3所示。内部通信模块运行于路由器内部交换网络之上,负责主处理单元和从处理单元之间的通信。虚拟驱动模块位于路由器操作系统的底层,用于模拟从处理单元上线路接口单元的行为,即相对上层软件虚拟出与线路接口单元功能相同、数量一致的虚拟接口单元。系统管理模块统一管理所有从处理单元,统计从处理单元上报的各类系统信息,监控从处理单元的运行情况,并且定期下发或更新转发表。

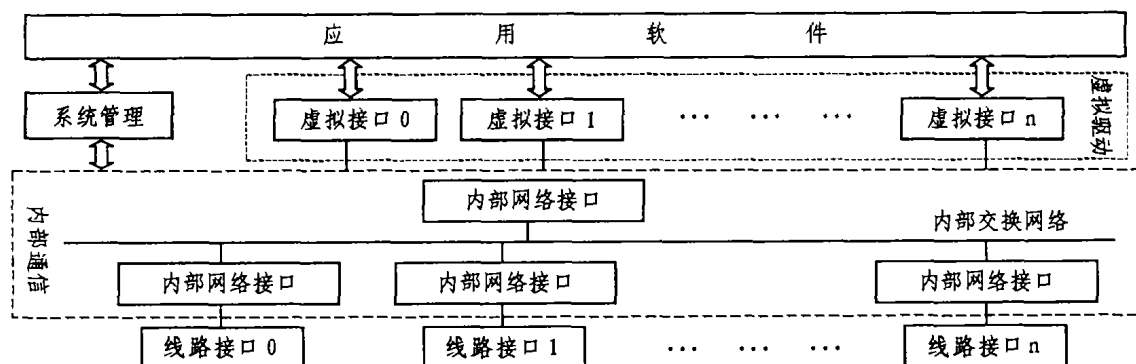


图3 硬件抽象层的运行环境

2.1 内部通信模块

内部通信模块负责主处理单元和从处理单元之间可靠的数据交换。

内部通信所传递的数据分为两大类,一类是主从处理单元间管理配置所需的各种信息,即系统管理模块所处理的数据。此类数据主要包括从处理单元向主处理单元上报的自身系统的运行情况以及主处理单元向从处理单元下达的配置命令及转发表等。

另一类数据是指需要经虚拟驱动模块处理后送协议层的信息。由于从处理单元不具备独立处理传输层协议报文的能力,因此所有需要协议处理的报文,都必须通过内部通信模块传到主处理单元进行处理。此类数据由线路接口单元进入,包括协议报文和从处理单元无法独立进行转发(如不存在该转发表项)的数据报文。

2.2 虚拟驱动模块

虚拟驱动模块的功能是模拟从处理单元上的线路接口单元。由于虚拟驱动模块的下层是内部通信模块,因此它的收发动作并不是对真实的网络接口卡的操作,而是与内部通信模

块之间的数据传递。

线路接口单元是路由器与外界网络交换数据的重要通道,与协议层紧密相关。虽然为了提高转发速率,将其设计在从处理单元上,但是它与主处理单元协议层的关系十分密切,需要在硬件抽象层中设置相应的软件模块。而对于线路接口而言,最通用的上层接口就是驱动程序;对于上层软件而言,驱动就代表了相应的设备。因此,协议层通过虚拟驱动控制线路接口单元的收发包动作,是可行的。

2.3 系统管理模块

系统管理模块也处于内部通信之上,它的功能是维护从处理单元的正常运转。它所管理的数据分为两部分:一是转发表,二是从处理单元及线路接口单元的配置管理信息。

系统管理模块在从处理单元初始化完毕后,将当前转发表下发。由于网络拓扑和路由信息是不断变化的,因此转发表需要进行相应的更新,当主处理单元的转发表发生变化时,系统管理模块要实时更新从处理单元上的转发表,使其与主处理单元的转发表保持一致。

从处理单元和线路接口单元的管理配置,也是由系统管

理模块来实现的。系统管理把从处理单元上报的信息存入相应的数据库中,随时可响应上层网络管理软件的查询。上层软件需要更改路由器的配置时,也可直接对系统管理下达命令,再由它控制从处理单元和线路接口单元。此外,系统管理需要协调多个从处理单元之间的关系,使它们正常高效地运转。

3 实现及性能测试

我们将上述硬件抽象层在 Linux 嵌入式实时操作系统下实现并对其进行了性能测试。测试环境是由两块 Motorola Lopec7410 PowerPC 板、一台 Adtech AX/4000 路由器测试仪以及一台以太网交换机组成,测试环境如图4所示。其中:用两块 Lopec7410 板模拟一台路由器,其中一块作为主处理单元,运行嵌入式实时 Hardhat Linux 操作系统,一块作为从处理单元,运行嵌入式实时 VxWorks 操作系统,分别由两个控制平台控制,它们之间均通过10M/100M 以太网交换机进行连

接;Adtech AX/4000 路由器测试仪的两个接口分别为 Generator 与 Analyzer,如图4中的 ADC00与 ADC01,分处于两个子网。在该环境中,HAL 中的虚拟驱动模块与系统管理模块运行于主处理单元,板间通信模块中的服务程序运行于主处理单元,客户程序运行于从处理单元。Adtech AX/4000测试仪通过 ADC00向路由器的 NIC00接口发包,经路由器处理后,由 NIC01输出,进入测试仪的接口 ADC01。

我们利用 Adtech AX/4000 路由器测试仪,在不同负载下,对 HAL 的效率及可靠性进行了测试,测试结果如图5所示。测试时的发包速率服从近似于网络流量特征的马尔可夫调制泊松过程(MMPP)^[6],其中图5(a)给出了重负载时包速率的变化情况,图5(b)为相应的 HAL 的性能测试结果;图5(c)给出了轻负载时包速率的变化情况,图5(d)为相应的 HAL 的性能测试结果。另外,测试仪的统计结果显示,在轻重两种负载情况下,丢包率都低于 10^{-8} 。

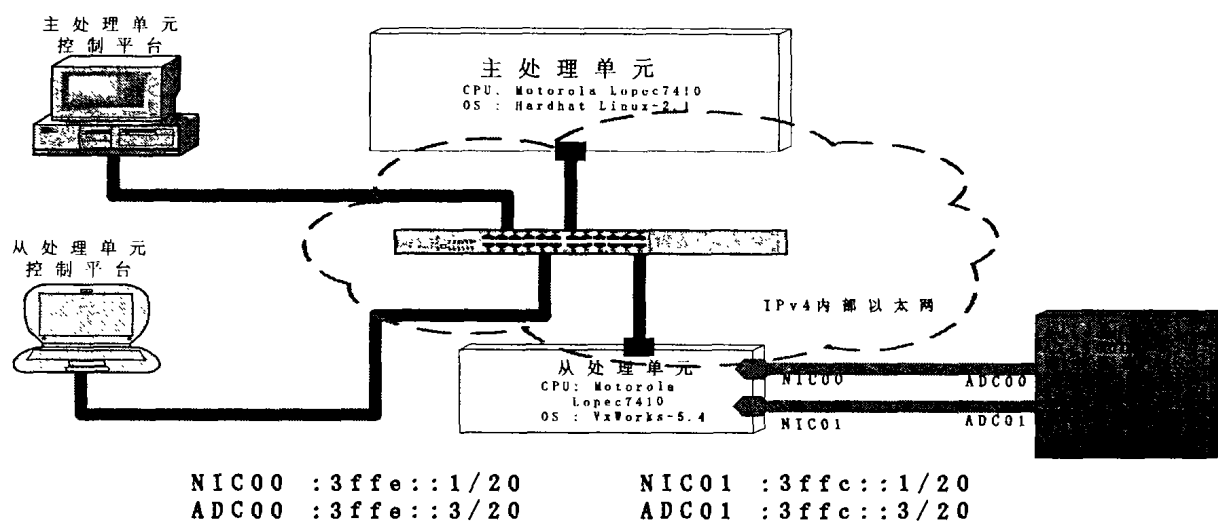


图4 测试环境组网示意图

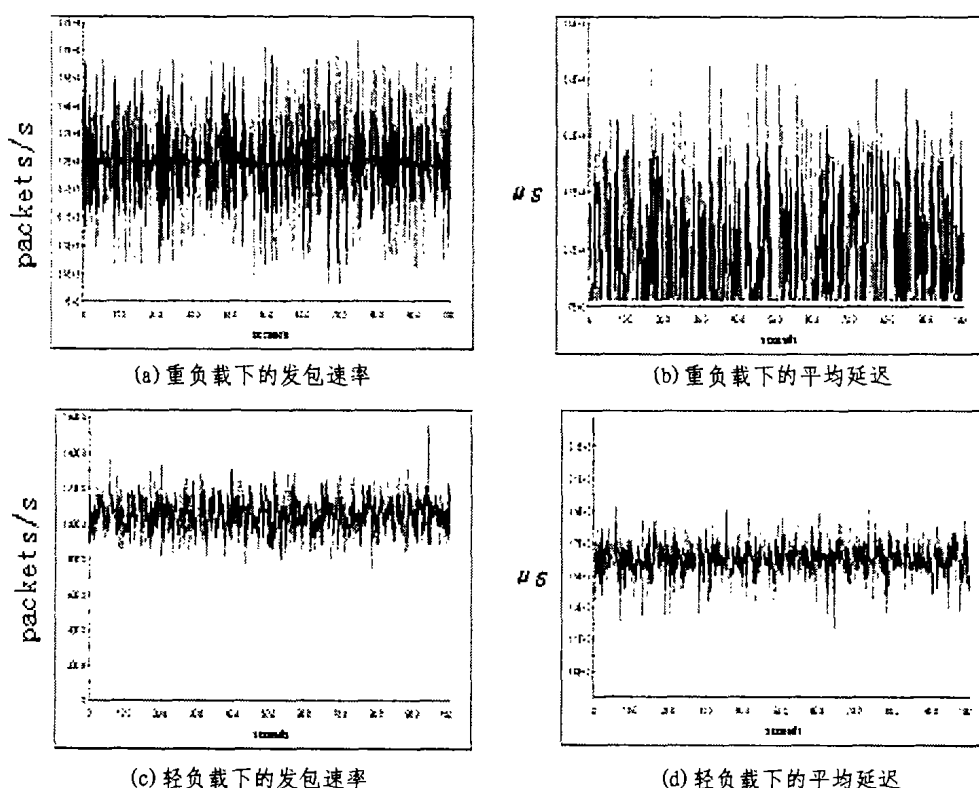


图5 不同负载下 HAL 的测试性能

(下转第191页)

choose-R3选择规则3

前提: $a: \forall (\forall (u_1, \dots, u_n) \exists (v_1, \dots, v_m). P). C \in A, \exists b \in O_A \exists x_1, \dots, x_n, y_1, \dots, y_m \in O_D: (fillerl_A(a, u_1) = x_1 \wedge \dots \wedge fillerl_A(a, u_n) = x_n \wedge fillerl_A(b, v_1) = y_1 \wedge \dots \wedge fillerl_A(b, v_m) = y_m \wedge (x_1, \dots, x_n, y_1, \dots, y_m): P \in A \wedge (x_1, \dots, x_n, y_1, \dots, y_m): \bar{P} \in A)$

结论: $A' = A \cup \{(x_1, \dots, x_n, y_1, \dots, y_m): P\}, A'' = A \cup \{(x_1, \dots, x_n, y_1, \dots, y_m): \bar{P}\}$

choose-R4选择规则4

前提: $a: \forall (\forall (u_1, \dots, u_n) \forall (v_1, \dots, v_m). P). C \in A, \exists b \in O_A, \exists x_1, \dots, x_n, y_1, \dots, y_m \in O_D: (fillerl_A(a, u_1) = x_1 \wedge \dots \wedge fillerl_A(a, u_n) = x_n \wedge fillerl_A(b, v_1) = y_1 \wedge \dots \wedge fillerl_A(b, v_m) = y_m \wedge (x_1, \dots, x_n, y_1, \dots, y_m): P \in A \wedge (x_1, \dots, x_n, y_1, \dots, y_m): \bar{P} \in A)$

结论: $A' = A \cup \{(x_1, \dots, x_n, y_1, \dots, y_m): P\}, A'' = A \cup \{(x_1, \dots, x_n, y_1, \dots, y_m): \bar{P}\}$

定义10 令 C 表示概念, R 表示任务, $P_i (i=1, \dots, k)$ 表示来自于 Φ_{DP} 的谓词名, a 和 b 表示来自于 O_A 的对象名, 令 f 是一个特征, 如果如下的冲突触发器中任意一个被触发, 则 $ABoxA$ 被称为矛盾的: ①基元冲突: $a: C \in A, a: \neg C \in A$; ②特征域冲突: $(a, x): f \in A, (a, b): f \in A$; ③所有域冲突: $(a, x): f \in A, a: \forall f. C \in A$; ④具体域冲突: $(x_1^{(1)}, \dots, x_1^{(1)}): P_1 \in A, \dots, (x_1^{(k)}, \dots, x_n^{(k)}): P_k \in A$ 并且相应的合取 $\bigwedge_{i=1}^k P_i(x^{(i)})$ 在 D 中是不可满足的。这是可以判定的, 因为 D 被要求是容许的。

DLSR(DP)的冲突触发器与 ALC(D)的冲突触发器是相同的。与 ALC(D)的相应规则相比, $\exists$ Pr-R1- $\exists$ Pr-R4 和 Choose-R1-Choose-R4 是新增加的规则。

规则 $\exists$ Pr-R1- $\exists$ Pr-R4 的意义显而易见。如果已知两个对象通过复合任务相联系, 那么我们可以立即建立具体对象作为用于定义复合任务的特征链的填充符, 用于定义在新建立的具体对象上成立的谓词也就知道了, 并且可以进一步添加适当的约束。

规则 Choose-R1-Choose-R4 的意义可以理解如下: 如果一个特征链集合用于定义复合任务 R , 并且对于这个特征链集合中的所有特征链来说, 存在两个对象 a 和 b 的适当的具具体填充符, 则 b 可能是 a 的任务 R 填充符。如果真是这种情况, 必须存在一个显式的约束来规定定义任务 R 的谓词 P 成立(或不成立)。

我们可以利用与文[9]类似的方法来证明算法的有效性和完备性。算法的有效性是从子代生成规则的局部有效性中得出的。算法的完备性是通过证明任意完备的 $ABox$ 是根据

定义初始 $ABoxA_0$ 模型的算法经计算得到的来证明的。

结束语 我们展示了如何利用描述逻辑 DLSR(DP)把空间和术语推理结合在 TBox 中, 关于描述逻辑的研究已经扩展到解决定性空间关系和定量空间数据, 主要思想之一是引进基于具体对象性质定义的任务概念。目前的检测有限个合取式的可满足性的工作是在目前的定性空间推理理论的基础上进行的。

参考文献

- 1 Baader F, Hanschke P. A scheme for integrating concrete domain into concept languages. In: Twelfth Intl. Joint Conf. Artificial Intelligence, Aug. 1991. 452~457
- 2 Lutz C. Reasoning with concrete domain. In: Proc. of the Sixteenth Intl. Joint Conf. on Artificial Intelligence IJCAI-99, Stockholm, Sweden, 1999
- 3 Lutz C. Reasoning with Complexity of terminological reasoning revisited. In: Proc. of the 6th Intl. Joint Conf. on Logic for Programming and Automated Reasoning (LPA'99), number 1705 in Lecture Notes in Artificial Intelligence, Springer-Verlag, Step. 1999. 181~200
- 4 Horrocks I, Sattler U, Tessaris S, Tobies S. Query Containment using a DLR ABox: [LTCS-Report 00-15]. LuFG Theoretical Computer Science, RWTH Aachen, Germany
- 5 Borgida A, Franconi E, Horrocks I, McGuinness D. Explaining ALC subsumption. In: Proc. of DL'99, 1999. 37~40
- 6 Artele A, Franconi E. A terporal description logic for reasoning about action and plans. In: J. Doyle, E. Sandewall, P. Torasso, eds. Fourth Intl. Conf. on Principles of Knowledge Representation, May 1994
- 7 Faltings B. Qualitative spatial reasoning using algebraic topology. In A U Frank and W Kuhn, editors, Spatial Information Theory: a theoretical basis for GIS, volume 988 of Lecture Notes in Computer Science, Springer-Verlag, 1995. 17~30
- 8 Randell D A, Cui Z, Cohn A G. A spatial logic based on regions and connection. In: B. Nebel, W. Swartout, C. Rich, eds. Principles of Knowledge Representation and Reasoning of the 3rd Intl. Conf. Cambridge MA, Morgan Kaufmann, 1992. 165~176
- 9 Baader F, Hanschke P. A scheme for integrating concrete domains into concept languages. In: Twelfth Intl. Joint Conf. Artificial Intelligence, Aug. 1991. 452~457
- 10 Nebel B. Computational properties of qualitative spatial reasoning: First results. In: I. Wachsmuth, C. -R. Rollinger, W. Brauer, eds. Proc. KI-95: Advances in Artificial Intelligence, 19th Annual German Conf. on Artificial Intelligence, Volume 981 of LNAI, Springer-Verlag, Sep. 1995. 233~244
- 11 Renz J, Nebel B. On the complexity of qualitative spatial reasoning: A maximal tractable fragment of the region connection calculus. In: Proc. of 15th Intl. Joint Conf. on Artificial Intelligence (IJCAI97), Aug. 1997. 522~527

(上接第158页)

图5中的测试结果表明, 在不同网络负载下, 包的平均传输延迟和丢包率, 都处于很低的水平。上述性能指标, 完全可以满足高性能路由器在效率、可靠性方面的需求。

结论 本文研究了分布式路由器软件体系结构通用化设计问题, 提出了高性能路由器系统中的硬件抽象层概念及其功能, 并就其功能及其关键模块在高性能 IPv6 路由器基础平台及实验系统中予以实现。通过所构建的测试环境, 对 HAL 进行了较系统的测试, 测试结果表明: HAL 达到较高的性能指标, 其运行效率与可靠性都能满足高性能路由器的需求。对 HAL 的功能分析结果表明, HAL 的提出较好地解决了路由器分布式体系结构所带来的软件系统设计的难题, 在上层软件实现中, 基本上可以不考虑底层硬件细节, 增强了路由器的开放性及其可扩展性, 具有较强的普适性。

参考文献

- 1 Keshav S, Sharma R. Issues and trends in router design. IEEE Communications Magazine, 1998, 36(5): 144~151
- 2 Basic Scalability and Performance Considerations for Evaluating Large-Scale Router Designs. <http://www.cisco.com/warp/public/cc/pd/rt/12000/tech/ruar-wp.htm>
- 3 Karlin S, Peterson L. VERA: An Extensible Router Architecture. In: Proc. of the 4th Intl. Conf. on Open Architectures and Network Programming, April, 2001
- 4 Intel Corporation. IXP1200 Network Processor Datasheet. Sep. 2000
- 5 Vitesse Semiconductor Corporation, Longmont, Colorado, IQ2000 Network Processor Product Brief, 2000
- 6 Andersen A T, Nielsen B F. A markovian approach for modeling packet traffic with long-range dependence. IEEE Journal on Selected Areas in Communications (JSAC), 1998, 16 (5): 719~732