文章编号:1001 - 2486(2005)05 - 0001 - 05

高速路由与交换技术现状及发展趋势*

卢锡城,管剑波
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:多个方向的可扩展特性是新一代互联网的重要特点,这对高性能路由器的性能、规模、功能和服务等能力提出了更高的要求,高速路由与交换技术是承载这种可扩展特性的关键技术。从网络处理器设计、路由查找技术以及高速交换技术三个方面总结了高速路由与交换技术的研究和发展现状,并展望了未来的发展趋势。

关键词:网络处理器;路由查找;高速交换;可扩展
中图分类号:TP393 文献标识码:A

A Survey on High Performance Routing and Switching Technology

LU Xi-cheng, GUAN Jian-bo
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Scalability is one of the most important characteristics of next generation Internet. And it makes more rigid demands for high performance routers. Since high performance routing and switching technology is important for routers to provide this characteristic, we make comprehensive a survey on three aspects of the network processor design, routing lookup and high speed switching. Related works are also summarized. The future directions of development are analyzed.

Key words: network processor; routing lookup; high speed switching; scalability

未来新一代互联网体系结构的一个显著特点就是要在规模、功能、性能、安全和服务等多个方向上具有良好的可扩展特性,以更好地适应互联网速度及应用的快速发展。然而现有网络体系结构难以满足网络功能复杂多样化趋势的要求。高性能路由器是互联网重要的网络单元基础设施,是承载规模可扩展、性能可扩展、功能可扩展和服务可扩展的主要平台[1],而高速路由与交换技术是高性能路由器中的重要关键技术。本文主要陈述高速路由与交换技术的研究工作,并展望了未来发展。

1 研究背景

当前高速路由与交换技术的研究主要集中于网络处理器设计、路由查找技术及高速交换技术三个方面,它们也是高性能路由器实现中的主要关键技术,它们在路由器中的作用如图 1 所示。

网络处理器(NP)负责对报文进行协议相关的处理,如路由转发,QoS 控制,安全加密等,它与路由器的性能和功能相关;路由查找(Lookup)为转发操作服务,它是进行报文分类和路由选择的关键技术,关系到路由器所能达到的性能;高速交换(Switching)是交换网络设计的关键技术,直接关系到路由器的性能和规模。

新一代互联网的流量快速增长,且对网络的功能和服务需求日趋多样化,下一代高性能路由器将向着规模、性能、功能和服务多维可扩展的网络核心平台发展,上述高速路由与交换技术的研究必然要适应这个发展的趋势。

* 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);国家自然科学基金项目(90104001)
作者简介:卢锡城(1946—),男,教授,博士生导师,中国工程院院士。

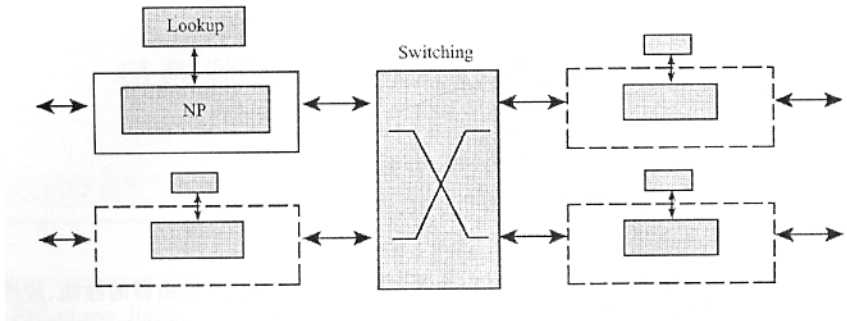

图1 高速路由与交换技术在路由器体系结构中的位置

Fig.1 Routing and switching technology in the architecture of routers

2 网络处理器设计

网络处理器是一种针对网络协议处理和报文转发进行了优化的特殊CPU,它兼具了ASIC转发引擎和通用CPU的优点,可同时获得较高的性能和灵活性,近年来在路由器中得到了广泛的应用。网络处理器技术已经成为推动下一代网络发展的核心技术。

2.1 网络处理器基本结构

从2000年起,很多公司推出了其网络处理器产品,例如Intel的IXP2800、IBM的NP4GS3、EZChip的NP-2以及AMCC的nP3700等。它们在体系结构上有共同之处,一般包含一组处理单元PE(Processing Element),多个协处理器CoP(Co-Processor)和多个硬件逻辑块HLB(Hardware Logic Block),如图2所示。其中,PE是一般功能的可编程处理单元,它执行网络处理器程序,对报文进行协议处理、路由转发等操作。一般网络处理器都使用多个PE,通过并行处理或流水线的方式提高性能。CoP完成报文处理中执行频度较高、处理较复杂的特定功能,如校验和计算、报文分类和路由查表等,一般在PE的指令下工作。HLB是一般的硬件逻辑模块,实现网络处理器的接口控制。

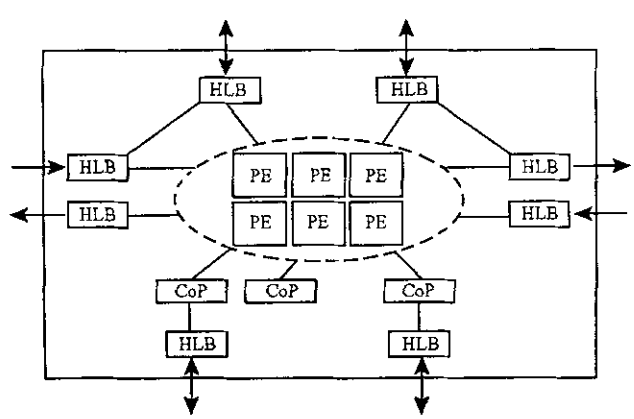

图2 网络处理器基本体系结构

Fig.2 Network processor architecture

2.2 网络处理器技术研究现状

近年来关于网络处理器技术的研究十分活跃,主要集中于以下几个方面。

2.2.1 网络处理器设计方法学

网络处理器设计方法学研究的目的是寻找更加有效、更加优化的一般性方法和规律,从而指导网络处理器的设计。Madhu[2]等人指出网络流量特征是指导网络处理器设计的重要依据,并分析了网络处理器中报文流的空间局部性、时间局部性和报文平均长度等特征。Crowley提出了网络处理器系统的模

型框架[3],力图基于模型的性能分析来评价实际系统的性能。Thiele 使用 Petri 网结构分析了网络处理器中的报文处理行为[4],提出网络处理器设计空间的概念。设计空间的开发是一个包含芯片面积、片上存储器需求、PE 能力及数量等多个问题参数的多对象优化问题。Thiele 还给出了一般性的设计空间优化策略。Wolf 等人从 SoC(System on Chip)设计的角度阐述了网络处理器设计空间开发问题[5],并且给出了处理器数目和 Cache 尺寸对工作负载的优化性能设计方法。

由于网络处理器是一种新生的技术,其设计方法学的研究还大都基于原来通用处理器的设计方法,因而对网络处理行为特征进行更进一步的研究和分析,以网络行为特征指导网络处理器设计将会达到更好的效果。

2.2.2 网络处理器体系结构

由于网络报文流中广泛存在的报文间相互独立的特性,使得并行结构成为网络处理器为提高处理能力而广泛采用的结构。Crowley[3] 的研究表明:网络处理特点决定了没有足够的指令级并行(ILP)空间,网络处理器获得高性能的关键是开发报文级并行(PLP)。因此面向 ILP 支持的处理器体系结构,例如乱序超标量(SS)结构或者细粒度多线程(FGMT)结构,都不能获得较高的性能,不适合用作网络处理器体系结构。而支持并发线程级并行的处理器体系结构,例如片上多处理器(CMP)和并发多线程(SMT)结构,更适合于开发报文处理中的 PLP 并行性,所以能够在处理报文时获得较好性能。这正是目前网络处理器体系使用嵌入式多处理器结构的发展趋势原因所在。

开发网络处理器并行性的另一方法是采用流水线体系结构。Intel 公司设计了基于流水线的网络处理器体系结构。在这种结构中,报文处理流水线跨越多个 PE,每个 PE 完成报文处理的某一个阶段,而流水线结构中任务阶段的划分成为至关重要的问题。

2.2.3 多协议支持技术

网络处理器要应对不断提高的链路速度,越来越多的高层协议也进入了它的处理范围,因此未来网络处理器需要具有支持多种协议的灵活性,尤其是对应用日益广泛的安全协议的支持。目前的通用网络处理器都没有针对安全协议的优化处理,设计安全协议专用网络处理器成为另一种研究的思路,它主要用于防火墙,VPN 网络等安全性要求很高的互联网接入级中。安全协议专用网络处理器的代表:Cavium 公司的 CN22 系列芯片支持 SSL 和 IPSec VPN,性能上支持 200Mbps 的大报文块加密和每秒 100 次的 IKE 或 SSL 握手;Broadcom 公司的 BCM58 系列支持 3DES/ESA 加解密功能。网络郸理器技术在面向核心应用不断发展的同时,也逐渐扩展到接入层的需要多协议支持的应用环境中,这也是网络处理器的一个发展趋势。

2.3 网络处理器技术发展趋势

网络处理器技术是支撑新一代互联网性能、功能和服务可扩展的关键技术,网络处理器技术的发展呈现如下趋势:

· 设计方法学方面将更加注重分析网络处理工作负载的特性,针对不同应用环境设计最佳的网络处理器体系结构;

· 在设计实践上将更多地利用多处理器片上系统(MP – SOC)技术、硬件多线程技术及片上通信技术提高处理性能;

· 研究更适合多协议支持的体系结构,应用将从核心向边缘扩展,范围更加广泛;

· 网络处理器接口将趋向标准化,支持互操作,便于系统集成。

3 路由查找技术

路由查找是路由器进行报文转发的关键技术,它通常由网络处理器中的专用 CoP 完成。路由查找的主要任务是根据报文所要到达的目的地址,在转发表中查找对应路由并按标示的路径转发。路由器采用的路由查找技术决定了其转发能力,对性能至关重要。

3.1 路由查找技术研究现状

对路由查找技术的评价标准包括性能、实现代价、更新复杂度和可扩展能力等,任何一种路由查找

技术都是在以上几个因素之间寻找平衡。路由查找技术的研究可分为以下三类：

3.1.1 基于软件的查找算法

基于软件的查找算法一般都是基于 Trie 树的查找，其思想是根据表项值(例如 IP 地址前缀)的二进制位构建 Trie 树，在检索时以查找关键字作为索引，在 Trie 树中进行遍历。Radix Trie[6] 算法、Patricia 算法[7] 都是基于二叉树的查找，在其上运用路径压缩技术[7] 或者 leaf pushing 技术[8] 可以获得时空效率的优化。为了加快搜索的收敛速度，LC – Trie 树算法[9] 和受控前缀扩展算法[8] 采用了多分支树结构来组织路由表。Srinivasan 提出了多分支 Trie 树的一般结构[8]，所有基于 Trie 树的算法都可以看做是该一般结构的特例或是变形。

所有基于 Trie 树的查找算法中，Trie 树的高度直接决定了查找过程访问存储器的次数，而存储器的速度远远低于 CPU 和 Cache 的速度，访存的次数就决定了算法的性能。因此，引入路径压缩、多分支、leaf pushing 以及前缀扩展技术，都是致力于减少访存次数、提高算法的性能，目前的优化技术往往使得实现代价和更新复杂度都有较明显的增加。

3.1.2 基于硬件的路由查找技术

近年来，随着硬件器件技术的发展以及对路由查找速度越来越高的要求，很多研究工作转向硬件支持的路由查找技术。24 – 8 DIR[10] 算法实际上是一种用硬件实现的多分支前缀扩展算法，它的 Tire 树只有两级，因此查找操作最多只需两次访存。这是一种用空间换取时间的查找技术，对存储器空间的要求很大。

另一种硬件查找技术是基于 TCAM(Ternary CAM)的查找。近年来 TCAM 器件在速度和容量上都有很大提高，能够满足大规模高速路由查找的需求。基于 TCAM 的查找速度快，但是实现代价较高，功耗较大，且表项的更新非常复杂。

3.1.3 利用 Cache 的路由查找算法

软件查找算法的访存次数决定了其性能，而利用 Cache 算法的目标是让更多的访存落到 Cache 中去，因此可以极大地减少搜索的时间，如 Lulea 算法[11] 以及 Degermark[12] 所提出的算法等都属于此类算法。这类算法的性能远高于软件查找算法，但是由于使用了树压缩技术，因此实现和更新都较为复杂。

3.2 路由查找技术发展趋势

未来的路由查找技术面临着三个因素的挑战：一是网络带宽需求的快速增长，使得对路由查找速度的要求越来越高；二是路由表的规模越来越大，路由查找技术必须适于大规模路由表的查找环境；三是 IPv4 向 IPv6 协议的过渡以及多维报文分类查找的广泛使用，使得路由查找关键字的位宽大大增加，未来有可能大于 300 位。为了应对上述挑战，未来路由查找技术的研究将会向以下几个方向发展：

- 与硬件技术相关的路由查找技术及算法将成为研究的主流；
- 路由查找技术的研究重点将从 IPv4 转移到 IPv6；
- 支持高层协议或专用协议的多维分类查找将是一个研究的重点；
- 路由查找技术的研究将和网络处理器技术的研究紧密结合，未来的查找技术和算法研究将以网络处理器为应用背景。

4 高速交换技术

交换网络是高性能路由器的核心，它关系到高性能路由器的规模和性能。未来高性能路由器需要具有更高的端口速率和更大的端口密度以及扩展灵活性，所以高速交换技术研究的目标就是交换网络要具备向更高性能和更大规模的扩展能力。

4.1 高速交换技术研究现状

近年来交换技术的研究主要体现为交换网络体系结构的发展和相关算法的研究。根据报文在交换网络中经历交换的次数，交换网络可以分为单级交换结构和多级交换结构两种。

4.1.1 单级交换

传统的路由器多采用单级交换结构,例如共享输出缓存结构(Shared Output Queuing)[13]和 Crossbar with VOQs(Virtual Output Queuing)[14]交换结构。共享输出缓存结构可以提供最理想的性能和 QoS 保证,但是它对存储器带宽的要求随端口数目或链路速率均呈线性增长,因而制约了它的扩展能力。Crossbar with VOQ 交换结构中必须有一个集中式的调度器来进行实时配置。调度算法对交换性能影响很大,代表性算法有 PIM[15],WFA[16]以及 iSLIP[17]等,这些调度算法的复杂度随端口数的增加而呈指数增长,而链路速率的增长提高了对调度器运行速度的要求,所以集中式调度器是阻碍 Crossbar with VOQ 结构扩展的瓶颈。传统的单级交换结构难以在性能和规模上进行扩展,采用新的交换结构是发展的趋势。

4.1.2 多级交换

相对于单级交换结构,多级交换结构的最大优点是具有良好的可扩展性,被认为是实现更大规模交换网络的必然选择。Dally 提出了 3D Torus 网络是一种适于实现交换结构的拓扑[18],并分析了该结构中报文路由、拥塞控制及虚通道等技术的应用。在工业界,Juniper 公司在其 T640 路由器中采用了 3 级 Clos 交换结构,而 Cisco 公司在 CRS-1 路由器中采用了 3 级 Benes 交换结构。总的来看,上述几种交换结构都是借鉴以往在高性能 MPP 计算机中广泛采用的拓扑结构,存在着实现复杂、内部互连代价高等缺点。Chang 提出了一种新颖的两级交换结构[19],并证明了这种交换结构可以在任何容许的流量下获得 100%的吞吐率,且实现代价低。Keslassy 扩展了这种两级交换结构[20],解决了报文乱序的问题,并描述了将这种结构应用于光互连路由器中的具体实现。两级交换结构被认为是比较有前途的一种交换结构。此外 Pappu 提出了一种被称为 Distributed Queueing 的分布式队列调度算法[21]。该算法在假设多级交换网络内部具有一定加速比的条件下,调度分配各输入端口的带宽,可以使整个交换网络获得近似 100%的吞吐率,为多级交换网络的实现提供了思路。

4.2 高速交换技术发展趋势

高速交换技术研究的主要目标是更高的性能和更大的规模,这是未来高性能路由器可扩展特性的重要基础。为了适应这一趋势,今后高速交换技术的研究将会有以下几个重要趋势:

- 可扩展多级交换结构将成为研究的热点;
- 多级交换结构内部一般存在着阻塞,如何减轻阻塞将是研究的一个重要方向;
- 研究多个基于标准开放接口的交换模块如何搭建更大规模的交换网络;
- 交换技术的研究将更多地注意如何有效利用 VLSI 技术的进步成果,例如由于 VLSI 技术的发展,内嵌缓冲区的 Crossbar 交换结构受到关注[22]。

5 结 论

网络流量的快速增长以及网络服务的日趋多样化对高性能路由器提出了新的要求,网络处理器技术、路由查找技术和高速交换技术作为高性能路由器实现的关键技术,如何适应路由器规模、功能、性能和服务可扩展性的要求,是未来研究发展的主要方向。

参 考 文 献:

[1] Braden R, Clark D, Shenker S, et al. Developing a Next-generation Internet Architecture[R]. White Paper, DARPA, http://www.isi.edu/newarch/documents/WhitePaper.ps, July, 2000.

[2] Madhu S, et al. Network Processors: Guiding Design through Analysis[R]. http://www.cs.wisc.edu/johnbent/Projects/net-proc.pdf.

[3] Crowley P, et al. Characterizing Processor Architectures for Programmable Network Interfaces[A]. International Conference on Supercomputing[C], Santa Fe, N.M., May, 2000.

[4] Thiele L, et al. Design Space Exploration of Network Processor Architectures[A]. First Workshop on Network Processors at the 8th International Symposium on High Performance Computer Architecture (HPCA8)[C], Cambridge MA, USA, February, 2002.

[5] Wolf T, et al. Design Tradeoffs for Embedded Network Processors[A]. International Conference on Architecture of Computing Systems (ARCS)[C], 1999: 149-164.

(下转第 11 页)

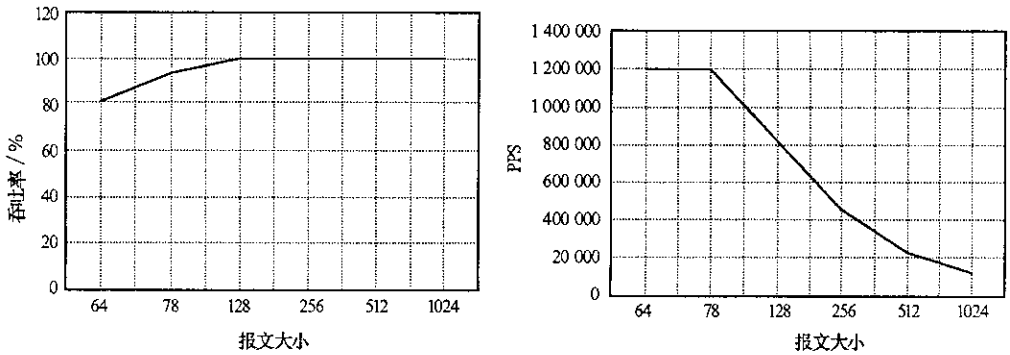

图 7 IPv6 核心路由器的性能测试结果

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.