ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.6, June 2008, pp.1452−1464
DOI: 10.3724/SP.J.1001.2008.01452

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

可扩展路由器*

张小平+, 刘振华, 赵有健, 关洪涛

(清华大学 计算机科学与技术系,北京 100084)

Scalable Router

ZHANG Xiao-Ping+, LIU Zhen-Hua, ZHAO You-Jian, GUAN Hong-Tao

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: E-mail: zhxp@tsinghua.edu.cn

Zhang XP, Liu ZH, Zhao YJ, Guan HT. Scalable router. *Journal of Software*, 2008,19(6):1452−1464.
http://www.jos.org.cn/1000-9825/19/1452.htm

Abstract: This paper presents a survey on the study of the scalable router and proposes a hierarchical model based on the study of the architecture and models of the scalable router. The model splits the scalable router into six layers from bottom to top, which are interconnection network and data switching layer, routing lookup layer, standard interface layer, distributed operating system layer, distributed routing behavior layer and single image management layer. The paper provides surveys on the research of each layer, and finally concludes and analyses the difficulties of current development of the scalable router.

Key words: scalable router; routing node; control plane; data plane; interconnection network; routing lookup; distributed routing calculation; single image

摘 要: 对可扩展路由器的研究现状进行了综述,并在可扩展路由器体系结构和模型研究的基础上提出其分层模型,将可扩展路由器"自底向顶"地划分为互连结构和数据交换层,路由查找层、标准接口层、分布式操作系统层、分布式路由行为层和单映像管理层 6 层,并综述了每层的研究进展.最后进行了总结并分析了当前可扩展路由器发展的难点.

关键词: 可扩展路由器;路由节点;控制平面;数据平面;互连结构;路由查找;分布式路由计算;单映像

中图法分类号: TP393 **文献标识码**: A

从 1969 年 12 月 ARPANET 包含 4 个节点的实验网络开始运行至今,Internet 经历了历史上任何一种技术都未曾经历的飞速发展,在不到 40 年的时间内已经演变为一个全球性的巨大商业网络,并对人类社会的各个领域产生了深远的影响.近年来,随着各种新型应用在 Internet 上的大规模部署,尤其是基于 P2P 技术应用的流行,Internet 的流量快速增长;另一方面,超高速光通信技术、无线通信技术以及其他革命性技术的研究进展使得网络接入和传输方式日新月异,新的网络应用模式和网络协议不断涌现.在此情况下,用户对于 Internet 的性能

* Supported by the National Natural Science Foundation of China under Grant Nos.90604029, 60773150 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2077AA01Z219 (国家高技术研究发展计划(863))

Received 2007-08-02; Accepted 2008-01-24

提出了多维度的要求,其中速度仍为主要指标.

路由器作为 Internet 的核心设备,对于 Internet 的性能有着重要影响.历史上,路由器体系结构的变迁体现了由通用器件向专用器件,由串行处理向并行处理,从集中式到分布式的趋势.依此可将路由器体系结构划分为 4 代:单处理器集中式总线结构,多处理器分布式共享总线结构,多处理器分布式交换结构,多机互连的可扩展集群结构[1].第一代路由器中,单一的中央处理器和总线成为系统计算能力和通信能力的瓶颈.第二代~第四代路由器体系结构逐步解决了这方面的问题:第二代体系结构提高了线卡的报文转发能力,主要贡献为将中央处理器从报文转发中解放出来;第三代体系结构中采用了交换结构代替总线,通过构造"无阻塞"的交换网络来大幅度提高交换性能;第四代体系结构目前仍在不断发展的过程中,通过某种互连方式将多个路由节点有机地结合在一起的可扩展路由器体系结构将是充满希望的发展方向.

所谓"可扩展路由器",是由多个可独立运行的路由节点,通过某种互连结构连接而成性能、功能可扩展的单映像路由器.其可扩展性主要体现在以下 3 个方面:交换实体的分布性带来的规模可扩展性;路由实体的分布性带来的路由计算可扩展性;路由器操作系统的分布性带来的功能可扩展性.

总的来说,可扩展路由器研究主要分为数据平面可扩展和控制平面可扩展两个层次,而有关可扩展路由器的性能模型研究方面,一般也基于其中某个层面.对于可扩展路由器系统所包含的问题及其之间的关系,可参考图 1 的可扩展路由器系统层次模型.

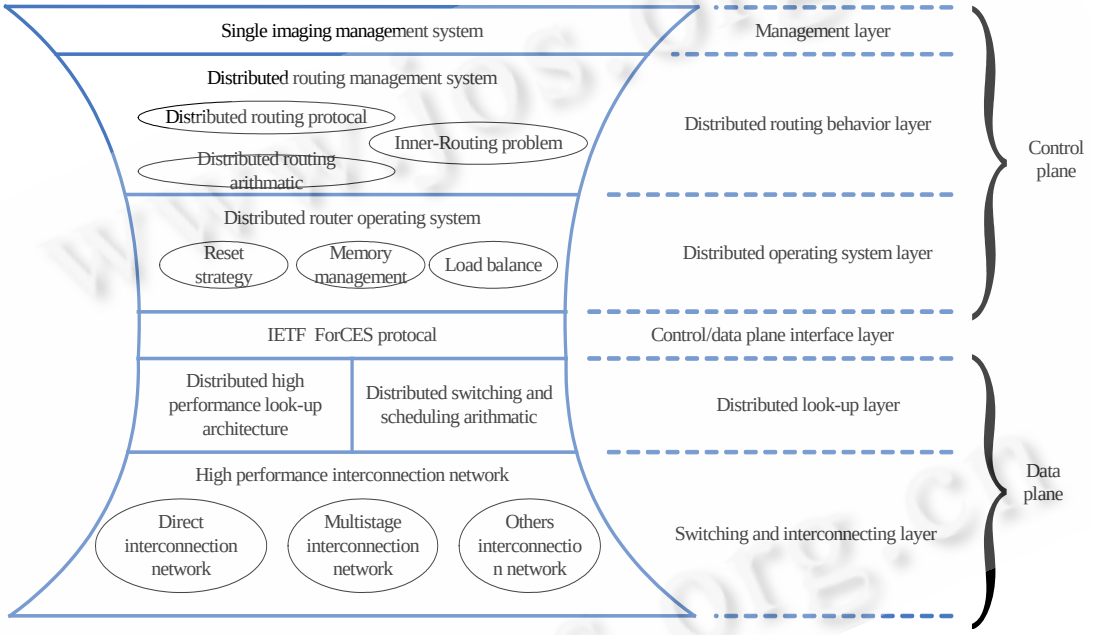

Fig.1 Hierarchical model of scalable router architecture

图 1 可扩展路由器体系结构层次模型

概括而言,可扩展路由器系统的具体研究内容如下:

(1) 数据平面可扩展性研究:数据平面可扩展性研究一般包括分布式高速路由查找结构、高性能可扩展互连结构和分布式交换及调度算法.在此方面,高性能可扩展互连结构的性能对于可扩展路由器的整体性能至关重要.目前,在数据平面可扩展性方面已经取得了部分研究成果[2−4].

(2) 控制平面可扩展性研究:控制平面可扩展性研究的主要问题有分布式路由、可扩展分布式操作系统和单映像管理等,控制平面只有具有与数据平面相匹配的可扩展能力才能满足整体的可扩展要求.

(3) 数据平面与控制平面接口标准研究:IETF 组织的 ForCES 工作组目前正在制定有关可扩展路由器中数据平面与控制平面分离的接口标准.

从工业界的发展情况来看,主要的路由器生产商也在努力增强其产品的可扩展性.JUNIPER 公司的路由矩阵(routing matrix)[5]由单机 T640 扩展而来.就其扩展性而言,数据平面根据 Clos 网络的特点,理论上可以实现无限扩展;控制平面虽然在物理上具有可扩展性,但由于仍采用集中式路由计算,因此从严格意义上讲并不具有可扩展性.相比之下,CISCO 路由器 CRS-1[6]交换结构为三级自路由 Benes 结构,实现了 1296×1296 缓存无阻塞交换;其控制平面的可扩展性主要体现在采用模块化,全分布式,基于微核结构的自诊断操作系统 Cisco IOS XR Software,总体来看,CRS-1 在数据平面和控制平面都具有强大的可扩展能力.AVICI 路由器 TSR[7]在数据平面,采用 3D torus 互连结构,具有高度的内部路径多样性;在控制平面上,软件采用 IPriori System Software,用于优化和控制交换及路由,支持 BGP-4,OSPF,IS-IS 和 PIM-Sparse 等协议,是分布式结构.TSR 最大的特点就是可以实现经济型扩展.

1 可扩展路由器模型及体系结构研究

目前,很多研究机构都在可扩展路由器模型研究方面得出了有价值的成果.Click[8]路由系统是由 MIT 的 Eddie Kohler 博士提出,并由 MIT 计算机技术系并行与分布式操作系统实验室开发完成.Click 是一种新的用来构建灵活和可配置路由器的软件体系结构.Click 的配置是模块化的,不仅有灵活的配置,而且有很高的性能.但是,Click 只是针对数据转发层的设计,并没有指出如何与控制层的应用相结合,而且不具备动态配置能力.Scout[9]是一个由 DARPA 和 NSF 支持的项目,它先由 Arizona 大学开发,目前由 Princeton 大学继续开发.作为一个针对网络应用的独立操作系统,Scout 有以下 3 个方面的特点:提出了一种称为路径(path)的面向通信的抽象;具有可配置性,Scout 的实例都是针对一种特殊的网络应用,由一系列的模块组成;包含有调度和资源分配机制.路径的抽象是基础;不同的路径依靠不同的模块配置实现其功能;调度和资源的分配策略也是基于路径的,不同的路径有不同的策略.Router Plugins[10]结构是在 Crossbow 项目下,由 ETH Zurich,Washington 大学,Ascom 这 3 个单位合作开发完成.Router Plugins 是在 NetBSD 操作系统内核中实现高性能、模块化、可扩展服务的路由器软件体系结构.这种结构允许在运行时动态地添加配置代码模块,即插件(plugins).

文献[11]对上述 3 种可扩展路由器结构进行了一个比较性研究,提出了一种一般化的路由器结构.文献认为在可扩展性方面,3 种结构都可以进行功能上的扩展,但是 Click 结构所付出的代价最小,拥有最好的可扩展性;在对报文的调度处理方面,Scout 可以有效地预测包的处理时间及所需要的资源等,调度最容易进行;在对流的隔离方面,Scout 结构中每条流均沿着自己的一条路径转发,不同的流之间没有交集,流隔离性最好.

VERA[12](virtual extensible router architecture)是一种虚拟可扩展路由器体系结构的概念.VERA 从平台的角度设计路由器的结构,并考虑到了硬件的多样性.它侧重于设计路由器抽象的可扩展结构及功能模块的分布,但忽视了对分布式结构的支持.

此外,加州大学伯克利分校为解决当前的网络研究中路由器的源代码不开放和 API 不开放的问题,提出了 XORP(eXtensible open router platform)[13]开放平台.其目标是提供功能全面、可扩展、有性能保证、稳定的科研工具和配置平台,使新的创意从实验转化为应用变得更加容易.总的来说,XORP 的可扩展性主要体现在路由器的上层协议和转发功能细节上.CLARA[14](a cluster-based active router architecture)是美国 NEC 公司和 Princeton 大学共同提出的一种配置路由和计算功能的体系结构,从而能够在网络中提供灵活的计算服务.CLARA 是 JOUNEY[15]主动网络中路由节点的原型.文献[16]提出的 Active Router Cluster 结构相对于 CLARA 引入了负载均衡策略,并提出了负载均衡中的乱序问题.

在高速路由器可扩展性研究过程中,很多模型重点研究数据层面的可扩展性.此类模型一般使用 ASIC 或 NP 以及高速交换结构来突破传统路由器所面对的性能瓶颈.就其结构而言,通常采用主控制器和子交换引擎的分布式结构[17].具有代表性的模型有 Washington 大学设计和实现的一个动态可扩展路由器 DER[18](dynamically extensible router),它可以自由地安装软、硬件的插件,主要的设计目的是为可编程网络、协议设计、路由器软硬件设计等研究提供一个实验平台.此外,Suez[19]是支持高速的尽力而为报文路由和可扩展 QoS 保证报文调度的高性能实时报文路由器方案,其硬件平台由通用 PC 集群通过 GB 级系统域网络(SAN)连接而成.项目旨在表

明 PC 集群体系可以像并行计算那样为高性能报文路由提供低成本平台.Suez 描述了一个可扩展路由器的功能分配和构建模式,并证实了 PC 集群体系是一种低成本的构建高性能路由器平台的途径,为进一步研究完整的分布路由器平台奠定了基础.Pluris[20]大规模并行路由器是由大量的处理节点通过线性可扩展高速数据互连而成.处理节点可以分为转发节点和路由节点两类:转发节点是通用 PC,负责转发 IP 报文;路由节点是具有更高处理性能的通用 PC,负责路由协议的执行及向其他转发节点广播报文转发表.为了提高性能,Pluris 路由器将转发节点通过低速线路连接到若干同步多路复用器上,从而将低速流聚合成骨干网上的高速流;为了提高可靠性,路由节点有多个冗余节点同时工作.

在国内的研究中,比较有代表性的是分布式路由器[21]和集群路由器[22,23]研究.分布式路由器也能实现路由器一定程度上的规模、性能的可扩展,但其研究均基于第三代路由器体系结构,因此严格意义上还不能称作是可扩展路由器.集群路由器研究方面,国防科学技术大学定义集群路由器为“由多个可独立运行的路由交换实体,通过某种互连方式组合而成的单映像路由器”[22].因此,集群路由器是真正意义上的可扩展路由器.在集群路由器体系结构研究方面,文献[23]提出了一个结构灵活、开放可扩展的通用路由器体系结构——OpenRouter 模型.OpenRouter 模型将路由器划分为转发实体(forwarding engine,简称 FE)和控制实体(control engine,简称 CE),其中控制实体又可以进一步垂直划分为控制服务层和操作服务层.模型的开放性特点来自于控制实体与转发实体之间、控制服务层与操作服务层之间、控制服务层与外部应用之间定义的 3 个层次的开放可编程接口;模型中对等层协同机制是实现可扩展性的关键.此外,文献[24]提出了软件集群路由器的两个参考模型:SCR-RM(softwarebased cluster router reference model)和 PCR-RM(parallel cluster routing reference model),并给出了这两个参考模型的具体描述.

1.1 分离控制模型及接口标准化工作

目前,国内外提出的分离控制模型有:IETF 的通用交换管理协议[25],网络处理器论坛(NPForum)的 NPF Software Mode[22],IEEE 基于电信模型的 P1520 分层参考模型[26]以及 IETF 的 ForCES 工作组的转发与控制单元分离的思想[27].其中,在可扩展路由器研究方面比较有影响力的是 ForCES 和网络处理器论坛.

2001 年,IETF 成立 ForCES 工作组,该工作组针对以下问题进行研究:IP 网络单元逻辑分离对控制和数据转发平面的机制需求;ForCES 模型和协议;定义组成 ForCES 网络单元的实体和识别实体间交互的体系结构框架;描述转发单元的功能模型.目前,ForCES 工作组已经提出了关于 ForCES 协议整体框架的 RFC3746 和 RFC3654 及大量的 Draft,涉及 FE Model,ForCES protocol,TML 和 neighbor discovery 等. ForCES 工作组提出的可扩展路由器模型如图 2 所示.

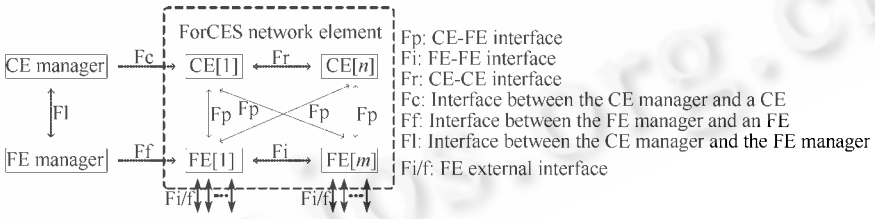

Fig.2 ForCES model

图 2 ForCES 模型

网络处理器论坛关注于定义用来构建功能块的标准接口.NPForum 与 ForCES 的不同主要体现在:使用的模型是静态的;控制与转发的分离是逻辑性的 API;而 ForCES 模型中的控制与转发的分离是协议性的互操作规范.NPForum 自 2002 年起也发布了一系列路由器内部接口规范,其中包括 NPSI[28],LA-1 等硬件接口规范.

1.2 本文的可扩展路由器模型

本文中使用的可扩展路由器体系结构遵循控制平面与数据平面分离的思想,其模型如图 3 所示,自底向顶

地将可扩展路由器分为数据交换和互连结构层、路由查找层、标准接口层、可扩展路由器操作系统层、分布式路由行为层和单映像管理层 6 层.其中,可扩展路由器操作系统层是协调软件系统和硬件系统的重要层,它应该对上屏蔽硬件的具体实现,对下兼容各种硬件结构和查找技术.控制平面和数据平面分别由独立的控制单元和转发单元通过互连结构连接而成.数据平面内部互连结构和控制平面内部互连结构功能不同,但在物理实现上可以共享.

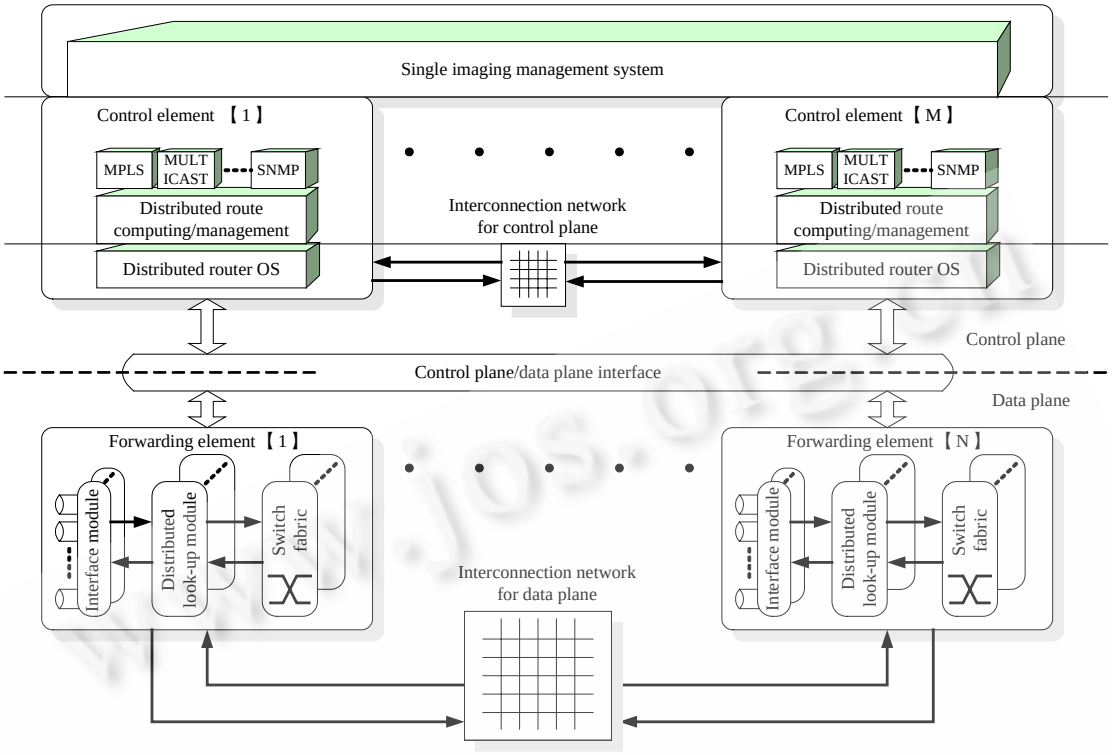

Fig.3 Model of scalable router

图 3 可扩展路由器模型

2 数据交换和互连结构层

2.1 分布式交换及调度算法

交换结构中调度算法的发展一直体现着从集中到分布、从复杂到简单的过程.早期基于 Crossbar 的调度算法多基于寻求二分图的某种匹配,如极大匹配算法等.它们都是集中式的,计算复杂度较高.于是,人们提出了一系列分布式实现的近似算法,如 PIM[29],iSLIP[30]等.这些算法降低了求匹配的计算复杂度,但依然受限于 Crossbar 的集中式结构.因此,基于 Buffered Crossbar 的交换结构[31]受到了重视,其中调度算法可以全分布式实现.近年来,随着多级交换结构的发展,两级负载均衡结构[32]备受关注,其最大优点在于具有 $O(1)$的在线调度复杂度.

2.2 高性能可扩展互连结构

2.2.1 直连结构

在直连结构中,每个节点除了负责分组处理的任务外,还需要承担分组转发的任务.各个节点间通过数据通道相连,相邻节点间可以直接通信.直连结构的拓扑特点决定了相邻节点之间的连接方式,主要有链状(chain)结构、环型(ring)结构、超立方体(hypercube)网络结构、完全图(complete graph)结构、蜂巢结构、蛛网结构、网

格(mesh,X-mesh,H-mesh)、带环网格(2D-torus,3D-torus,H-torus)等.目前,直连网络已经广泛应用于多处理器系统(multiprocessor)、多计算机(multi-computer)系统以及集群(cluster)系统中.

在以上几类典型的直连结构中,链表和环的拓扑结构简单,易于实现,但在路径多样性和吞吐率方面性能都比较差,超立方体结构拓扑结构复杂,不适合于扩展.完全图结构在路径多样性、时延和吞吐率方面都表现得不错,但是由于链路数目过多,导致成本开销大,在实际应用中用得很少.Mesh 结构的构建比较简单,而且相对于以上几种拓扑结构,平均性能不错.X-mesh 和 Torus 结构是对 Mesh 结构的改进,在路径多样性方面有所提高.文献[33]提出了一种基于带中心点正六边形拓扑的蜂巢结构可扩展路由器,并给出了其在单机架和多机架结构中的具体部署方案以及适于无限可扩展的蛛网式可扩展路由器结构.H-mesh 是蜂巢结构的一种变体,它是一种六边形网格结构,其编址方法和路由算法实现比较简单,适用于实际拓扑结构,而 H-torus 是对 H-mesh 结构的一种扩展,它的路由和广播算法实现简单,效率高,而且提高了路径多样性并降低了时延.

2.2.2 非直连结构

除了直连结构以外,其他互连结构都可以称为非直连结构.非直连结构又可以分为 3 种:总线结构,交叉开关网络(crossbar)和多级互连网络(multistage interconnecting network,简称 MIN).

总线结构的信道利用率较高,结构简单,价格相对便宜,但扩展能力受限于总线的带宽.交叉开关网络可为每端口的相同数据宽度和相等的连接端口提供更高的带宽并提供简单的无阻塞操作.为了减少冲突,可以在交叉开关网络的输入端口或输出端口增加缓冲,这样就构成了带缓冲的交叉开关网络(buffered crossbar).多级互连网络就是在输入端口和输出端口通过一级以上的开关相连.常见的多级互连网络有 Banyan 网络[34](包括Ω网络[35]、基准网络(baseline)[36]、Delta 网络[37]),Batcher 排序网络[38]、间接的二进制 *n*-立方体结构[39]、Clos 网络[40]、Benes 结构[41]等.

2.2.3 可扩展互连结构的实例

文献[42,43]提出了两种可扩展的互连结构:ICO 和 GMR.前者属于非直连结构,后者属于直连结构,它们都有分布式管理和硬件复杂度低的特点.

ICO(I-CubeOut)[42]是一种针对可扩展路由器的多级交换结构,这种结构以间接 *n*-立方体的结构为基础,使用报文再循环(packet recirculating)连接方式来提高吞吐率.为了解决间接 *n*-立方体结构中的阻塞问题,可以考虑以下两种方法:在结构中采用输出队列,但需要一个输出加速比;在原来一级的基础上增加一级拷贝(second copy),其中再循环连接方式可以是静态或动态的.

GMR(grid oriented multistage-connected RU's)[43]是一种针对 T 比特级路由器的可扩展交换结构.它由两种组件组合而成:路由单元(routing unit,简称 RU),用于决定报文的转发方式;连接组件(connecting component,简称 CC),从横向和纵向把 RU 以网格的方式连接起来.在实际应用中,只需 RU 在 *x*,*y* 方向上各连接 3 个 CC,并且配置适当的加速比,GMR 结构可以达到很好的性能(吞吐率为 1).

3 路由查找层

3.1 集中式路由查找

核心路由器中的路由查找就是根据报文中的目的地址字段查询目的端口,从而最终到达目的节点的过程.随着路由器端口速度的不断提高,快速路由查找算法成为研究热点,文献[44,45]分别从不同的角度综述了部分快速路由查找算法,目前形成的算法主要分为[17]:(1) 基于 Trie 的查找算法[3,46−48];(2) 基于 Hash 的查找算法[49.50];(3) 基于前缀的查找算法[51−53];(4) 基于硬件的查找算法[54−57]等.

3.2 分布式路由查找

对于可扩展路由器来说,报文的转发过程比较复杂,源节点到达目的节点的过程不是入端口到交换结构到出端口的传输过程,而是要跨越多个节点.在可扩展路由器中,分布式路由查找是未来的发展方向之一.在这方面,互连网络的路由选择和负载平衡问题已经有了一些比较成熟的理论[58].文献[59]中提出了基于 TCAM 的分

布式并行 IP 查找机制和性能分析.

3.3 对IPv6的支持

IPv6是下一代互联网的核心技术.在设计路由管理模型时,有必要对IPv6的扩展性进行考虑:IPv6中地址长度从32bit增大到128bit[60];路由表和转发表的结构都会发生变化;路由协议从RIPv2,OSPFv2,BGP4分别升级到RIPng,OSPFv3,BGP4+.Zebra 公司对此开发出一套协议[61]来规范路由协议和路由管理子系统之间的通信.这种协议规定了不同的通信数据类型,对IPv6有较好的扩展性.文献[62]提出了一种可扩展的IPv6查找方案.这种方案提出的主要原因是:IPv6中地址长度增加至128位,现有最快的IPv4路由查找方案也无法实现线速.文献[62]提出的方案是基于对于现实世界路由前缀的分布规律相关的 RFC 文档,结合了位图压缩和路径压缩,使用可变步幅(variable-stride)机制来最大化压缩比和最小化平均存储使用.他们的实验数据说明该方案可以支持10Gbps的线速转发.

4 操作系统层

操作系统是路由器软件体系结构的重要支撑[63].在传统的路由器体系结构中,操作系统需要完成系统资源管理,硬件抽象与屏蔽、任务调度与控制等操作.在可扩展路由器中,这些仍然是操作系统的功能基础.除此之外,可扩展的体系结构提出了新的需求.从根本上讲,可扩展路由器的操作系统是一种分布式操作系统[64],关于分布式操作系统的研究已经十分深入[65,66].

可扩展路由器软件体系结构支撑系统是在以经典分布式操作系统为基础、为路由器特有功能服务的一个软件平台,它的分布式支持包括对基本分布式特性的支持和对路由器特有应用的分布式支持.前者完全可以由现有的分布式操作系统来实现,而后者更是可扩展软件体系结构的支撑结构所要研究的根本问题.文献[67]提出了路由器的一种可扩展性支撑结构方案.但是,这种方案属于以功能可扩展为目标,不能解决规模可扩展的软件体系结构问题.随着路由器数据平面和控制平面分别进行的规模扩展,两个平面间的通信量大幅增加,连接拓扑更为复杂,传统信息流的传输模式会成为限制系统进一步扩展的潜在瓶颈.这个问题在通用分布式系统中没有遇到,在当前可扩展路由器中也没有解决.

此外,可扩展路由器操作系统需要支持任务迁移,传统的任务迁移多集中在进程状态的维护等基础问题上,这些都可以为路由器系统的任务迁移所用.但是,路由器中有一类任务,例如路由计算,它们的运算包含了大规模的有结构数据集,例如路由表.这类任务对路由器至关重要,它们的迁移既要保证数据的完整性正确性,还要保证迁移过程的实时性.传统的分布式操作系统与路由器理论都没有对这个问题提出解决的方案.这是路由器可扩展软件体系结构支撑平台的另一个关键性结构问题.

针对数据平面到控制平面的数据流瓶颈问题和对大数据集任务迁移的支持问题,文献[68]中给出了传输适配子层的解决方案,并对这个模型的性能与效率进行了分析;将其关键部分归结为有结构的大数据集在节点间的迁移问题,提出了一个存储管理方法 PS2 来解决这个问题,使大数据集的重建过程降低到线性的计算复杂度.文献[69]中讨论了其负载均衡问题.

5 分布式路由行为层

路由协议计算是路由器的核心应用之一,是路由器进行路由操作的功能基础.路由协议计算的复杂度高,计算量大,在路由器软件体系结构的各个功能子系统中是占用资源最多、最重要的任务.在对路由器软件体系结构进行可扩展化的设计中,路由协议计算的可扩展分布式模型是最重要的可扩展应用模型.OSPF 协议和 BGP 协议路由计算的分布式研究对可扩展路由器和Internet骨干网的发展具有重要意义.

分布式路由行为层的主要功能是路由计算,路由计算采用可扩展的分布式路由协议实现.分布式路由协议实现根据系统当前的运行状况,将路由计算的核心部分划分成若干个单元,分布到多个节点上并行地运行,并将结果整合,生成统一的路由决策表,下发到数据平面.分布式的路由算法实现与调度方法是这一部分的关键内

容.文献[70]中针对高性能分布式路由器中路由管理必须面对两个技术难点(实现高性能的路由查找算法和实现主从路由表同步)提出一种路由管理模型,给出了高性能路由查找算法和主从路由表同步的一种解决方案.文献[71]中提出了一种非对称的路由同步框架 AREF(asymmetrical routes electing framework)路由同步框架.

5.1 分布式OSPF

OSPF(open shortest path first)是一个广泛部署的复杂路由协议,它包含邻居关系维护、数据库维护、最短路径树(shortest path tree, 简称 SPT)计算等多个组成模块.对于路由器上运行的 OSPF 协议任务,其主要瓶颈在于 SPT 计算.尽管有研究表明,在快速检测网络拓扑变化的要求下,邻居关系维护也会产生很大的系统开销,不过,由于这种开销可以通过邻居关系模块向线卡上的功能转移而得到很好的改善[72],最终不会成为系统的决定性瓶颈.以并行 SPT 算法为基础的分布式 OSPF 协议模型是集群路由器体系结构的重要构成部分之一.

目前对 SPT 的并行计算有一定的研究,基于 BTH 结构的分布式 SPT 算法[73,74]尽管可以实现 $O(\log n)$的计算复杂度,但是否所有的图都可以转化为 BTH 结构尚没有相关文献论述.文献[75]提出了一种基于“点断集”的网络划分.对于单源最短路径树问题,该算法可以达到 $O(n)$的计算时间复杂度.但是,该算法的划分不能解决任意多个平面上的拓扑网络的划分,后者是一个 NP-H 问题,这是该算法最大的缺陷.文献[76]中也提出了一种多级划分的方法.这种划分是针对 mesh 结构的拓扑网络.该算法解决 SOSP 问题的计算时间复杂度是 $O\left(N^{\frac{L-1}{L}}\right)$,比一般的最短路径算法效率要高.但该算法适用性不广,它主要针对 mesh 拓扑结构的网络.文献[77]提出了 BPA 算法,可以针对任意拓扑进行 SPT 的分布式计算,但在算法复杂度方面没有本质改进.

5.2 分布式BGP

Internet 核心路由器控制平面的 BGP 协议性能也面临着新的挑战.传统单进程集中控制下,BGP 协议实现在可靠性、路由表容量、路由计算能力和支持的邻居规模上都无法满足未来需求,而路由器硬件平台的发展提供了分布式的计算资源与存储能力.如何充分利用可扩展路由器的特点,提高 BGP 协议实现的性能是亟待解决的一个重要问题.目前,Internet 骨干节点的 BGP 路由表容量呈现出线性增长与指数增长交替的趋势[78],尽管经过努力又回落到线性增长,但未来发展趋势仍很难预料.在大容量路由表条件下,路由器需要消耗更多的存储空间,还要为未来增长进行预留,造成路由更新处理变慢,增加 BGP 协议的计算开销.而 BGP 路由更新、抖动和边界路由器 BGP 邻居数量的快速增长进一步加剧了路由器控制平面的负载.文献[79]的研究表明,BGP 路由更新具有很强的突发性,突发情况下每秒需要处理的路由前缀数量可以达到几百个.文献[80]对 Internet 拓扑结构分析显示出 AS 间 BGP 连接数量呈现快速增长的趋势,1999 年,大 ISP 连接的 BGP 邻居数量为 1 418 个,2001 年为 2 376 个.到目前为止,AS701[81]的 BGP 邻居数量已经达到 3 024 个,每个边界路由器需要维护的 BGP 邻居数量达到几百个甚至近千个.

在骨干网络核心路由器上,BGP[82]是应用范围最广的路由协议[83].可扩展路由器体系结构需要可扩展的分布式 BGP 协议计算模型与实现方案来支持.随着网络规模的增大,BGP 协议的可扩展需求主要来自两个的方面,一是网络节点的数目以及网络复杂度的增加;另一个是核心路由器所连接的邻居数目的增加.这使得路由器要具备越来越高的路由协议处理速度和更大的内存容量.

对于传统的集中式 BGP 协议计算通用模型,已有比较成熟的研究结果[84].当前分布式 BGP 研究的主要内容是对路由计算过程的分布式模型研究.其中有代表性的[85],对基于路径向量的 BGP 协议提出了一种分布式的计算模型.这种算法借用 Agent 的概念对路由的前缀进行运算节点之间的划分.文献[86]中,对 BGP 协议的并行实现技术和集群路由器报文转发表一致性维护问题进行了系统的阐述,对 BGP 协议的行为、处理能力,对控制平面 CPU 资源的消耗提供了理论基础,并提出基于 BGP 实体集合及 C-BGP 协议的算法模型,充分利用路由器的多处理器资源来加速协议运行,提高 BGP 协议的处理性能和邻居规模的可扩展性.

6 单映像管理层

单映像操作管理是可扩展路由器软件体系结构中最重要的应用子系统之一.对路由器软件体系结构进行可扩展化的设计,要求其既在计算模型上是分布式的,又提供集中式操作管理接口.由于路由器的操作管理复杂度比通用计算机要高得多,并且路由器操作管理对网络的正常运行有着至关重要的意义,所以在一定程度上,可扩展路由器单映像操作管理子系统具有更重要的意义.

GENESIS 系统[87]在其支持并行计算的操作系统中提供了单映像的系统抽象模型与实现,它对整个集群计算机系统进行了多层次的单映像建模.作为一种支持并行计算的操作系统,它更侧重于在操作系统级别提供单映像的支持,它面向的主要对象是并行的编程模型.由于可扩展路由器的研究大多还处在以数据平面为重点的阶段,因而还没有相关的文献资料对可扩展路由器软件体系结构的单映像操作管理进行系统的模型分析.从网络管理软件演化而来的路由器管理软件提供了部分参考,Cisco 公司为路由器开发的应用产品 Craft Works Interface(CWI)[88]是其中典型的管理软件系统,可以通过单一的操作接口对复杂的多节点路由器系统进行配置和管理,但是这样的软件仍然停留在较高的软件抽象层次,而直接驻留在路由器系统中,在操作管理效率方面受到一定的限制.文献[22]提出了异构型集群路由器体系结构模型 HCR(heterogeneous cluster-based router),它支持以常规路由器作为节点,以标准接口和协议进行互连以构成单映像集群路由器.针对模型分析了异构交换单元聚合,路由计算/控制单元聚合和单映像保持等关键问题.

7 结论和下一步研究展望

本文主要讨论了可扩展路由器的研究现状.首先提出了可扩展路由器的分层结构,将可扩展路由器"自底向顶"地划分为互连结构和数据交换层、路由查找层、分布式操作系统层、分布式路由行为层和单映像管理层 5 层,进而对每层的发展现状进行综述和评价.总的来说,现在的可扩展路由器的研究尚处于起步阶段,还有以下难点需要进一步的研究:

(1) 高性能可扩展互连结构及内部路由算法设计

新型拓扑结构的选择.该结构应具有良好的扩展性和容错能力.直连式互连结构是研究方向之一.在研究直连式互连结构的性能时,等分带宽是一个非常重要的参数,直接关系到互连结构的吞吐性能,这对于路由器这类强调吞吐率的设备显得尤为重要.然而,求解任意拓扑结构的等分带宽属于 NP-complete 问题[89],因此,如何选择高性能可扩展的直连式互连结构是一个首要的难题.对此,作者认为自同构拓扑结构是理想选择.它具有节点对称或边对称、扩展粒度低、等分带宽大、网络直径小、节点的度适度等优良特性,拓扑结构应该具备良好的扩展性.

内部路由算法的设计.在直连式互连结构中,内部路由算法的选择将直接影响到整个系统的性能,例如:吞吐率、分组延时、路由容错等.而系统性能与互连结构的拓扑、流量模型等多方面的因素相关.因此,设计一个较通用的路由算法就显得更加困难.最终所采用的路由算法最好具备自适应特性,对各类允许类型流量均能提供较高的吞吐率、较低的延时和一定程度的路由容错功能,这样才能使得算法支持可扩展特性.

(2) 分布式路由计算

随着网络规模的扩大,路由计算问题的复杂度也随之增长.当前路由算法全部是一种集中式的计算模式.在可扩展路由器中,需要研究路由计算的分布式问题,其中主要研究内容为 BGP 协议和 OSPF 协议的分布式计算.将计算分布到各个节点后,计算复杂度问题、计算效率问题、分布式算法的平滑过渡问题等都是研究的难点.在 OSPF 分布式计算研究中,SPT 的分布式是难点,如何改进目前已有的 BPA 算法,使得计算任务可以动态迁移,最终降低计算复杂度是研究方向之一.对于 BGP 分布式计算研究,目前主要考虑控制平面与数据平面完全分离的模型假设,但数据平面中硬件平台物理细节和拓扑关系会影响节点间的通信方式,自然会对负载分配造成一定影响[68],如何将此约束加入 BGP 分布式计算模型中,是值得探讨的方向.

(3) 单映像一致性

为实现单映像操作管理,需要解决同步算法在开销和性能之间的折衷问题,数据库维护问题(集中式全局管理虽然没有由命令触发的管理信息,但是集中的数据集需要对各个节点的管理信息进行不间断的实时收集与整理.分布式数据库的维护会严重限制系统的规模扩展能力)以及可靠性问题(集中式模型在路由器体系结构中具有严重的单一失效节点问题,这是可扩展路由器要特别避免采用的结构).对于 HCR 中单映像一致性的研究,主要问题是缺乏标准接口和协议规范,由于不能对普通路由器作任何假设,因此,研究必须基于开放的标准接口和协议规范.

References:

[1] Xu K, Wu JP, Xu MW. Advanced Computer Networks: Architecture, Protocol Mechanism, Algorithm Design And Router Technology. Beijing: Mechanism Industry Press, 2005. 496−500 (in Chinese).

[2] Iyer S, McKeown NW. Analysis of the parallel packet switch architecture. IEEE/ACM Trans. on Networking, 2003,11(2): 314−324.

[3] Gupta P. Algorithms for routing lookups and packet classification [Ph.D. Thesis]. Stanford: Department of Computer Science, Stanford University, 2000.

[4] Keslassy I, Chuang ST, Yu K, Miller D, Horowitz M, Solgaard O, McKeown N. Scaling Internet routers using optics. In: Proc. of the 2003 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM Press, 2003. 189−200. http://tiny-tera.stanford.edu/~nickm/papers/sigcomm2003.pdf

[5] T640 Routing Node and TX Matrix™ Platform: Architecture. White Paper. http://www.juniper.net

[6] Cisco CRS-1 series carrier routing system getting started guide. http://www.cisco.com

[7] The Avici TSR: Cornerstone of the multi-terabit core. http://www.avici.com

[8] Kohler E. The Click modular router [Ph.D. Thesis]. Department of Electrical Engineering and Computer Science, MIT University, 2001.

[9] Montz AB, Mosberger D, O'Malley SW, Peterson LL, Proebsting TA. Scout: A communications-oriented operating system. In: Proc. of the 5th Workshop on Hot Topics in Operating Systems (HotOS-V). 1995. 58−61. http://lahtermaher.org/pub/plan/xkernel/Papers/scout_hotos.ps

[10] Decasper D, Dittia Z, Parulkar G, Plattner B. Router plugins: A software architecture for next generation routers. IEEE/ACM Trans. on Networking, 2000,8(1):2−15.

[11] Gottlieb Y, Peterson L. A comparative study of extensible routers. In: Proc. of the 2002 IEEE Open Architectures and Network Programming Conf. New York: IEEE Press, 2002. 51−62. http://www.cs.princeton.edu/nsg/papers/routerstudy_openarch_02/routerstudy-openarch.pdf

[12] Karlin S, Peterson L. VERA: An extensible router architecture. In: Proc. of the 4th Int'l Conf. on Open Architectures and Network Programming (OPENARCH). 2001. 3−14. http://www.cs.princeton.edu/nsg/papers/vera_cn_02/vera-cn.pdf

[13] Handley M, Hodson O, Kohler E. Xorp: An open platform for network research. ACM SIGCOMM Computer Communication Review, 2003,33(1):53−57.

[14] Welling G, Ott M, Mathur S. CLARA: A cluster-based active router architecture. IEEE Micro, 2001,21(1):16−25.

[15] Ott M, Welling G, Mathur S, Reininger D, Izmailov R. The JOURNEY active network model. IEEE Journal on Selected Areas in Communications, 2001,19(3):527−537.

[16] Guo JN, Chen F, Bhuyan L, Kumar R. A cluster-based active router architecture supporting video/audio stream trans-coding service. In: Proc. of the Int'l Parallel and Distributed Processing Symp. 2003. 8−15. http://www-lih.univ-lehavre.fr/Intranet/proceedings/IPDPS2003/DATA/11_02_059.PDF

[17] Yu X. Research on the key technologies of cluster based router [Ph.D. Thesis]. Wuhan: Huazhong University of Science and Technology. 2005 (in Chinese with English abstract).

[18] Kuhns F, DeHart J, Kantawala A, Keller R, Lockwood J, Pappu P, Richards D, Taylor D, Parwatikar J, Spitznagel E, Turner J, Wong K. Design of a high performance dynamically extensible router. In: Proc. of the DARPA Active Networks Conf. and Exposition (DANCE). 2002. http://www.arl.wustl.edu/projects/fpx/references/dance02.pdf

[19] Chiueh TC, Pradhan P. Suez: A cluster-based scalable real-time packet router. In: Proc. of the 20th Int'l Conf. on Distributed Computing Systems. 2000. 136−144. http://academic.csuohio.edu/yuc/perf-00/References/ICDCS00-suez.ps

[20] Pluris massively parallel routing. White Paper. http://www.kotovnik.com/~avg/pluris/wp/

[21] Fan XB, Lin C, Wu JP, Xu L. Performance model and analysis of a distributed router. Chinese Journal of Computers, 1999,22(11): 1223−1227 (in Chinese with English abstract).

[22] Guan JB. Research on the architecture and key technologies of cluster-based routers [Ph.D. Thesis]. Changsha: National University of Defense Technology, 2005 (in Chinese with English abstract).

[23] Wang BS. Research and implementation on a new router architecture [Ph.D. Thesis]. Changsha: National University of Defense Technology, 2005 (in Chinese with English abstract).

[24] Gong ZH, Fu B, Lu ZX. Research on the architecture of software-based cluster routers. Journal of National University of Defense Technology, 2006,8(3):40−43 (in Chinese with English abstract).
[25] Doria A, Hellstrand F, Sundell K, Worster T. General switch management protocol (GSMP) V3. RFC3292, 2002.
[26] Biswas J, Lazar AA, Huard JF, Lim K, Pau LF, Torstensson S, Wang WG, Weinstein S. The IEEE P1520 standards Initiative for programmable network interfaces. IEEE Communications Special Issue on Programmable Networks, 1998,36(10):64−70.
[27] Doria A. ForCES Protocol Specification. IETF draft. 2007.
[28] Bergen C, Lynch J. Streaming interface (NPSI) implementation agreement. White Paper, Network Processing Forum Hardware Working Group, 2002. http://www.npforum.org/techinfo/HWStreamingIA.pdf
[29] Anderson TE, Owicki SS, Saxes JB, Thacker CP. High speed switch scheduling for local area networks. ACM Trans. on Computer Systems, 1993,11(4):319−352.
[30] McKeown N. The *i*SLIP scheduling algorithm for input-queued switches. IEEE/ACM Trans. on Networking, 1999,7(2):188−201.
[31] Rojas-Cessa R, Oki E, Chao HJ. On the combined input-crosspoint buffered switch with round-robin arbitration. IEEE Trans. on Communications, 2005,53(11):1945−1951.
[32] Chang CS, Lee DS, Jou YS. Load balanced Birkhoff-von Neumann switches, part I: One-stage buffering. Computer Communications, 2002,25(6):611−622.
[33] Yue ZH, Zhao YJ, Wu JP, Zhang XP. Designing scalable routers with a new switching architecture. In: Proc. of Autonomic and Autonomous Systems and Int'l Conf. on Networking and Services. 2005.
[34] Narasimha MJ. The Batcher-Banyan self-routing network: universality and simplification. IEEE Trans. on Communications, 1988, 36(10):1175−1178.
[35] Lawrie DH. Access and alignment of data in an array processor. IEEE Trans. on Computers, 1975,24(12):175−189.
[36] Wu CL, Feng TY. On a class of multi-stage interconnection networks. IEEE Trans. on Computers, 1980,29(8):649−702.
[37] Patel JH. Performance of processor-memory interconnection for multiprocessors. IEEE Trans. on Computers, 1981,30(10):771−780.
[38] Batcher KE. Sorting networks and their applications. In: Proc. of the AFIPS Spring Joint Computer Conf. 1968. 307−314. http://www.cs.kent.edu/~potter/research/papers/sort.pdf
[39] M Pease. The indirect binary *n*-cube microprocessor array. IEEE Trans. on Computers, 1977,6(5):250−265.
[40] Jajszczyk A. Nonblocking, repackable, and rearrangeable Clos networks: Fifty years of the theory evolution. IEEE Communications Magazine, 2003,41(10):28−33.
[41] Sapountzis G, Katevenis M. Benes switching fabrics with $O(N)$-complexity internal backpressure. IEEE Communications Magazine, 2005,43(1):88−94.
[42] Tzeng NF. Multistage-Based switching fabrics for scalable routers. IEEE Trans. on Parallel and Distributed Systems, 2004,15: 304−318.
[43] Tzeng NF, Mandviwalla M. Cost-Effective switching fabrics with distributed control for scalable routers. In: Proc. of the ICDCS. 2002. 65−73.
[44] Fuller V, Li T, Yu J, Varadhan K. Classless inter-domain routing (CIDR): An address assignment and aggregation strategy. RFC1519, 1993.
[45] Xu K, Xu MW, Wu JP, Wu J. Survey on routing lookup algorithms. Journal of Software, 2002,13(1):42−50 (in Chinese with English abstarct). http://www.jos.org.cn/1000-9825/12/42.pdf
[46] Morrison DR. PATRICIA: Practical algorithm to retrieve information coded in alphanumeric. Journal of the ACM, 1968,15(14): 514−534.
[47] Srinivasan V, Varghese G. Fast address lookups using controlled prefix expansion. ACM Trans. on Computer Systems, 1999,17(1): 1−40.
[48] Nilsson S, Karlsson G. IP-Address lookup using LC-Tries. IEEE Journal on Selected Areas in Communications, 1999,17(6): 1083−1092.
[49] Waldvogel M, Varghese G, Turner J, Plattner B. Scalable high speed IP routing lookups. In: Proc. of the ACM SIGCOMM'97. 1997. 25−36. http://sigcomm.org/sigcomm97/papers/p182.pdf
[50] Degermark M, Brodnik A, Carlsson S, Pink S. Small forwarding tables for fast routing lookups. In: Proc. of the ACM SIGCOMM'97. 1997. 3−14. http://www.cs.ust.hk/~hamdi/Class/CSIT560/Reading/Lookup1.pdf
[51] Mehrotra P, Franzon PD. Binary search schemes for fast IP lookups. In: Proc. of the Global Telecommunications Conf. (GLOBECOM 2002), Vol.2. 2002. 2005−2009. http://ants.iis.sinica.edu.tw/presents/Binary%20search%20schemes%20for%20fast%20IP%20lookups.pdf
[52] Berger, M. IP lookup with low memory requirement and fast update. In: Proc. of the High Performance Switching and Routing, HPSR 2003. 2003. 287−291.
[53] Lim H, Lee B, Kim WJ. Binary searches on multiple small trees for IP address lookup. Communications Letters, 2005,9(1):75−77.
[54] Chiueh TC, Pradhan P. High-Performance IP routing table lookup using CPU caching. In: Proc. of the IEEE INFOCOMM'99. 1999. 1421−1428. http://citeseer.ist.psu.edu/131258.html

[55] Song YL, Aboelela E. A parallel IP-address forwarding approach based on partitioned lookup table techniques. In: Proc. of the 29th Annual IEEE Int'l Conf. on Local Computer Networks (LCN 2004). 2004. 425−426.
[56] Wang ZX, Wang HM, Sun YM, Zhang YX, Wu JX. High-Performance IPv4/IPv6 dual-stack routing lookup. In: Proc. of the 18th Int'l Conf. on Advanced Information Networking and Applications (AINA 2004). 2004. 476−481.
[57] Lim H, Jung Y. A parallel multiple hashing architecture for IP address lookup. In: Proc. of the Workshop on High Performance Switching and Routing (HPSR 2004). 2004. 91−95.
[58] Dally WJ. Performance analysis of *k*-ary *n*-cube Interconnection networks. IEEE Trans. on Computers, 1990,39(6):775−785.
[59] Zheng K, Hu CC, Lu HB, Liu B. A TCAM-based distributed parallel ip lookup scheme and performance analysis. IEEE/ACM Trans. on Networking, 2006,14(4):863−875.
[60] Deering S, Hinden R. Internet protocol version 6 (IPv6) specification. RFC1883, 1998.
[61] Zebra. Zebra Protocol. 2001. http://manticore.2y.net/doc/zebra
[62] Zheng K, Liu Z, Liu B. A scalable IPv6 lookup scheme via dynamic variable-stride bitmap compression. Computer Communications, 2006,29(16):3037−3050.
[63] Xu K, Wu JP, Yu ZC, Xu MW. HEROS: Router-Oriented distributed real-time operating system. Journal of Tsinghua University (Sci & Tech), 2002,42(1):52−55 (in Chinese with English abstract).
[64] Fan XB, Lin C, Wu JP, Xu K. Performance model and analysis of a distributed router. Chinese Journal of Computers, 1999,22(11): 1223−1227 (in Chinese with English abstract).
[65] Levy E, Silberschatz A. Distributed file systems: Concepts and examples. ACM Computing Surveys, 1990,22(4):321−374.
[66] Denys G, Piessens F, Matthijs F. A survey of customizability in operating systems research. 2002.
[67] Pradhan P, Chiueh TC. Operating systems support for programmable cluster-based internet routers. In: Proc. of the Workshop on Hot Topics in Operating Systems. 1999. 76−81. http://citeseer.ist.psu.edu/27538.html
[68] Wu K. Research on Software architecture for extensible router [Ph.D. Thesis]. Beijing: Tsinghua University, 2006 (in Chinese with English abstract).
[69] Chan HCB, Alnuweiri HM, Leung VCM. A framework for optimizing the cost and performance of next-generation IP routers. IEEE Journal on Selected Areas in Communications, 1999,17(6):1013−1029.
[70] Liang ZY, Xu K, Wu JP, Xu MW. Routing management model in distributed routers. Journal of Tsinghua University (Sci & Tech), 2003,43(4):503−506 (in Chinese with English abstract).
[71] Zhang XZ, Lu XC, Zhu PD, Peng W. A synchronization framework and critical algorithm maintaining single image of ip forwarding tables among cluster router's nodes. Journal of Software, 2006,17(3):445−453 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/445.htm
[72] Basu A, Riecke JG. Stability issues in OSPF routing. In: Proc. of the ACM SIGCOMM 2001. 2001. http://www.sigcomm.org/sigcomm2001/p18-basu.pdf
[73] Zhu S, Huang GM. A new parallel and distributed shortest path algorithm for hierarchically clustered data networks. IEEE Trans. on Parallel and Distributed Systems, 1998,9(9):841−855.
[74] Antonio JK, Huang GM, Tsai WK. A fast distributed shortest path algorithm for a class of hierarchically clustered data networks. IEEE Trans. on Computers, 1992,41(6):710−724.
[75] Cohen E. Efficient parallel shortest-paths in digraphs with a separator decomposition. In: Proc. of the 5th Annual ACM Symp. on Parallel Algorithms and Architectures. 1993. 57−67. http://akpublic.research.att.com/~edith/Papers/separator.ps.gz
[76] Romeijn HE, Smithy RL. Parallel algorithms for solving aggregated shortest path problems. Computers & Operations Research, 1999,26(10-11):941−953.
[77] Zhang XP, Wu JP, Zhang N, Zhao YJ. BPA-A parallel shortest path algorithm for cluster-router. In: Proc. of the PDCS 2006.
[78] Geoff Huston. Internet BGP Table. http://www.potaroo.net/
[79] Labovitz C, Malan GR, Jahanian F. Internet routing instability. In: Proc. of the ACM SIGCOMM'97. 1997. http://www.cdt.luth.se/net/courses/99-00/smd076/articles00/internet-routing-instability.pdf
[80] Ge ZH, Figueiredo DR, Jaiswal S, Gao LX. On the hierarchical structure of the logical Internet graph. In: Proc. of the SPIE ITCom. 2001. 208−222. http://routeviews.org/papers/hier.ps
[81] CIDR REPORT. http://www.cidr-report.org
[82] Rekhter Y, Li T, Hares S. A Border Gateway Protocol 4 (BGP-4). IETF RFC 4271, 2006.
[83] Huston G. Analyzing the Internet's BGP routing table. Cisco Internet Protocol Journal. 2001,4(1). http://www.potaroo.net/papers/ipj/2001-v4-n1-bgp/bgp.pdf
[84] Xu K, Wu JP. Design and implementation of border gateway protocol BGP-4 based on event-driven programming. Journal of Software, 2000,11(11):1516−1521 (in Chinese with English abstract).
[85] Zhang XZ, Zhu PD, Lu XC. Fully-Distributed and highly-parallelized implementation model of bgp4 based on clustered routers. In: Proc. of the 4th Int'l Conf. on Networking (ICN 2005). 2005. 433−441.

[86] Zhang XZ. Research on parallel processing of routing protocols [Ph.D. Thesis]. Changsha: National University of Defense Technology, 2005 (in Chinese with English abstract).
[87] Goscinski A, Hobbs M, Silcock J. GENESIS: An efficient, transparent and easy to use cluster operating system. Elsevier Science, 2002. 557−605.
[88] Cisco craft works interface configuration applications reference guide, release 3.2. Technical Report, Cisco Systems, Inc., 2005. http://www.cisco.com/
[89] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. Greeman, 1979.

附中文参考文献:

[1] 徐恪,吴建平,徐明伟.高等计算机网络:体系结构,协议机制,算法设计与路由器技术.北京:机械工业出版社,2005.
[17] 余鑫.集群路由器关键技术研究[博士学位论文].武汉:华中科技大学,2005.
[21] 范晓勃,林闯,吴建平,徐恪.分布式路由器的性能模型与分析.计算机学报,1999,22(11):1223−1227.
[22] 管剑波.集群路由器体系结构及其关键技术的研究[博士学位论文].长沙:国防科学技术大学,2005.
[23] 王宝生.一种新型路由器体系结构及其实现技术研究[博士学位论文].长沙:国防科学技术大学,2005.
[24] 龚正虎,傅彬,卢泽新.软件集群路由器体系结构的研究.国防科学技术大学学报,2006,8(3):40−43.
[45] 徐恪,徐明伟.路由查找算法研究综述.软件学报,2002,13(1):42−50. http://www.jos.org.cn/1000-9825/12/42.pdf
[63] 徐恪,吴建平,喻中超,徐明伟.面向路由器的分布式实时操作系统 HEROS.清华大学学报(自然科学版),2002,42(1):52−55.
[64] 范晓勃,林闯,吴建平,徐恪.分布式路由器的性能模型与分析.计算机学报,1999,22(11):1223−1227.
[68] 吴鲲.可扩展路由器软件体系结构研究[博士学位论文].北京:清华大学,2006.
[70] 梁志勇,徐恪,吴建平,徐明伟.分布式路由器中的路由管理模型.清华大学学报(自然科学版),2003,43(4):503−506.
[71] 张晓哲,卢锡城,朱培栋,彭伟.一种集群路由器转发表同步框架及关键算法.软件学报,2006,17(3):445−453.http://www.jos.org.cn/1000-9825/17/445.htm
[84] 徐恪,吴建平.基于事件驱动的边网关协议 BGP-4 的设计与实现.软件学报,2000,11(11):1516−1521.
[86] 张晓哲.路由协议并行处理技术研究[博士学位论文].长沙:国防科学技术大学,2005.

张小平(1975－),男,内蒙古包头人,博士生,助理研究员,CCF 会员,主要研究领域为路由器体系结构,分布式路由.

赵有健(1969－),男,博士,副研究员,主要研究领域为路由器体系结构,交换与调度算法.

刘振华(1983－),男,硕士生,主要研究领域为可扩展路由器体系结构.

关洪涛(1980－),男,博士生,主要研究领域为可扩展路由器体系结构.