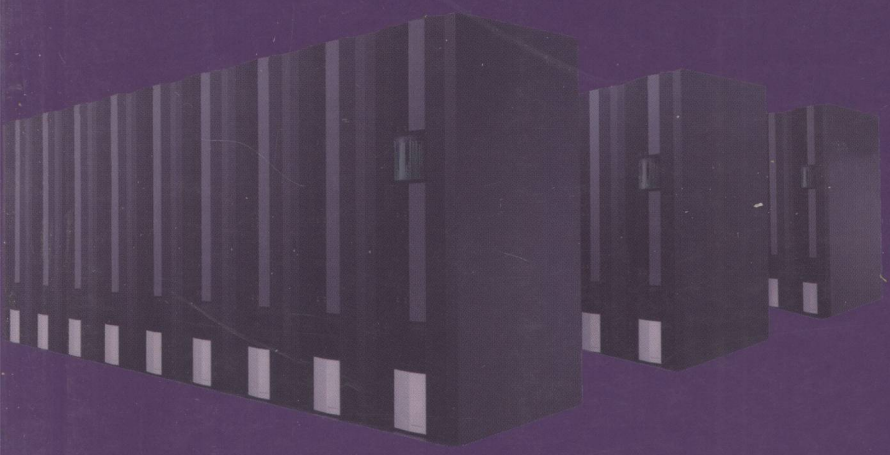
国家科学技术学术著作出版基金资助

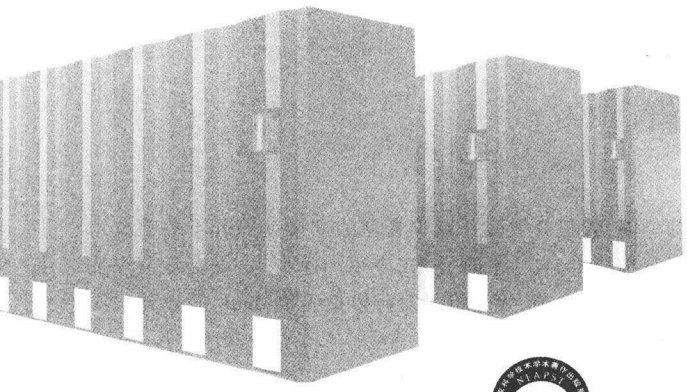


高效能计算机系统

——若干关键技术分析

曾 宇 等著

高等教育出版社



国家科学技术学术著作出版基金资助

高效能计算机系统

Gaoxiaoneng Jisuanji Xitong

——若干关键技术分析

曾 宇 许建卫 孙国忠 王 勇 王 洁 著

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

内容简介

本书对大规模高性能计算机如何降低能耗、提升效能并减少管理的复杂度等若干关键技术进行了阐述，主要内容包括：自适应功耗管理技术、自主管理技术、应用加速技术、高密度节点技术、网络内存技术、事件流应用技术、并行模拟技术等，并介绍了并行模拟引擎 SimK 的设计与实现，最后给出了高效能计算机曙光 5000A 的相对效能评价结果。

本书可供从事高性能计算机研究及应用的科研人员使用，也可作为高年级本科生及研究生学习高性能计算相关课程的参考书。

图书在版编目（CIP）数据

高效能计算机系统：若干关键技术分析 / 曾宇等著
--北京：高等教育出版社，2010.1
ISBN 978-7-04-028409-6

Ⅰ.①高… Ⅱ.①曾… Ⅲ.①计算机系统-研究
Ⅳ.①TP30

中国版本图书馆 CIP 数据核字（2010）第 001839号

策划编辑 刘 英 **责任编辑** 刘 英 **封面设计** 张 楠 **责任印制** 陈伟光

出版发行	高等教育出版社	**购书热线**	010-58581118
社　　址	北京市西城区德外大街 4 号	**免费咨询**	400-810-0598
邮政编码	100120	**网　　址**	http://www.hep.edu.cn
总　　机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经　　销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印　　刷	涿州市星河印刷有限公司	**畅想教育**	http://www.widedu.com
开　　本	787×1092　1/16	**版　　次**	2010 年 1 月第 1 版
印　　张	18	**印　　次**	2010 年 1 月第 1 次印刷
字　　数	370 000	**定　　价**	36.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

物料号 28409-00

前　言

在过去10年，高性能计算越来越得到产业界认可，应用领域正从科学计算领域向商业计算、信息化建设领域拓展，应用的种类不断增多，普及程度逐渐深入，使用者也从专家型研究人员向普通用户拓展。当前，内存墙、I/O 墙、功耗墙、复杂性墙、编程墙、扩展性墙是高性能计算机发展及应用的主要瓶颈。以高性能、可编程、可移植、稳定性为特征的高效能技术是高性能计算机研究的新方向。在高效能、多核、虚拟化等诸多技术的推动下，由超级计算拓展到广域计算机基础设施的科学研究正处于鼎盛阶段，一场在全球范围内的高性能计算平民化运动正在拉开帷幕，我们也称之为“泛高性能计算时代”。

本书对大规模高性能计算机如何降低能耗、提升效能以及减少管理的复杂度等关键技术进行了阐述。第1章为引言。第2章介绍了一种基于遗传算法的功耗管理方法，该方法采用作业队列的能效比作为调度因素，与面向资源效率的传统作业调度算法相比，在确保提升资源利用率、减少资源碎片、提升作业吞吐率、减少饥饿作业的前提下，大幅提升了系统的能效比。实际应用测试表明，该方法能有效降低系统能耗。第3章设计了一种分布式层次化的自主管理机制，该机制中采用的触发式 Bully 选举算法具有较高的执行效率，并能够适应节点故障、链路故障和节点变化等情况，具有一定的容错能力和动态特性。该机制支持在不停机条件下新增设备、修改特征参数，支持引入新的规则和进行启发式推理，从而提高了管理效率和故障定位的准确性。第4章阐述了一种基于 CBF 哈希过滤的五元组 IP 包分类算法，在网络安全应用加速卡中进行了有效实现，加速了 TCP/IP 协议栈的处理；同时，还介绍了 DGEMM 应用加速器的设计与实现，基于2颗 Xilinx Virtex V350 MHz FPGA 可实现 44.8 GFLOPS 的浮点计算性能。第5章阐述了一种高密度可扩展的计算节点结构，包括四路 SMP 的刀片设计，可实现多功能 I/O 扩展的 PCI-E 扩展模块的设计，以及管理模块、互连网络模块、机箱结构的设计，并应用于曙光5000A 高性能计算机。第6章提出并实现了网络内存服务系统 NMS（Network Memory Service），验证了网络内存技术在资源分配和加速应用方面的有效性；此外重点研究了网络内存优化技术，包括主动内存技术、M-PPM 预取技术以及二级缓存管理算法 LIRS-A。第7章以一个网络安全监控应用为基础，给出了一种从系统运行的 trace 中提取负载特征的方法，设计并实现了一个可扩展的并行查询引擎来提高单个查询的性能，提出了一种通过有效利用多种资源来提高系统吞吐率的并发查询调度方法，以上工作在一个事件流系统 DBroker 中进行了实现和实验。第8

章介绍了并行高性能计算机模拟的关键技术，包括基于阻塞/唤醒的同步机制、锁避免的调度技术、高性能通信技术、超步执行技术和多线程缓冲区优化技术等；除了上述内容外，还介绍了影响并行模拟性能的另一大要素——负载平衡，并提出了面向毫秒级的协作负载迁移机制的实现。第9章介绍了并行模拟引擎SimK的设计和实现，并给出了SimK在目前流行的高性能计算机SMP集群上的优化措施。SimK的设计目标是为了加速并行模拟器的开发，第9章后半部分基于SimK给出了千万亿次高性能计算机体系结构HPP模拟器实例——节点模拟器HppSim和互连模拟器HppNetSim的实现。第10章通过一个相对效能评价指标RPI，综合考虑了系统采购成本、运营维护成本、关键应用性能、代码编程难易度、管理复杂度等诸多因素，有效地解决了各参数直接测量值量纲不同的问题，更具合理性和测量的简洁性，并给出了高效能计算机曙光5000A的相对效能评价结果。

在高等教育出版社刘英编辑的支持和帮助下，此书从概念变成了现实，同时刘英编辑的建议也使本书的写作受益匪浅，在此表示衷心的感谢。在本书的撰写过程中，中国科学院计算技术研究所（以下简称中科院计算所）孙凝晖研究员从篇章结构、内容乃至行文方面都提出了许多建设性的指导和修改意见，在此表示衷心的感谢。常莉女士做了大量文字整理工作，特此致谢。曙光公司研发中心方信我、刘新春、刘朝晖等同志在应用加速技术的研究方面，温鑫等同志在自主管理技术研究方面，李麟等同志在自适应节能、高效能评价等方面，沙超群、郑臣明等同志在高密度计算节点研究方面都做出了贡献，在此一并表示感谢。

高效能计算机技术的发展是一个长期的实践过程，当前虽然学术界、工业界基于可重构计算、混合异构结构、高效能编程模型及编程语言、体系结构创新等相关技术有效提升了高性能计算机系统的效能，减少了机房面积、电能消耗、系统管理的一系列压力，更好地满足了应用需求，但必须看到，上述技术的突破与实现和高效能计算的总目标相比仍然有很大差距，未来高效能计算机技术仍将高速发展。本书对高效能计算机若干关键技术进行了原理性论述和必要的定量分析，目的是与同行交流观点和方法，为高等院校相关专业的教师、研究生和高性能计算机研制者提供一些技术参考。

本书反映了作者多年来在高性能计算机领域的研发体会、经验和成果，但技术发展快，涉及面广，个人水平有限，书中难免存在不妥之处，敬请指正。全书共分10章，第1、2、4、5、10章由曾宇撰写，第3章由王洁撰写，第6章由孙国忠撰写，第7章由王勇撰写，第8、9章由许建卫撰写，全书由曾宇统稿。

作 者

2009年9月9日

目　录

第1章 引 言

1.1 高性能计算机的应用

高性能计算机系统不可替代的优势，是为工业、商业、政府决策支持等领域的计算密集型应用和数据密集型应用提供快速、精确、海量数据的处理平台。当前，高性能计算机系统应用广泛，如情报分析、信号和图像处理及自动目标识别、武器系统集成模拟和测试等国防应用领域，气象气候、地球物理、空间物理、天体物理、高能物理、凝聚态物理、加速器物理、生命科学、材料设计和模拟、系统科学、人工智能、医学等基础科学研究领域，航空航天、舰船、能源、电子、电气等领域中的计算流体力学，有限元分析、电磁分析、多场耦合分析、系统性能分析等工程产品设计领域，金融数值模拟、商业数据挖掘、类物流系统规划等商业应用领域，传染病扩散、社会动力学、宏观经济学等社会学研究领域，环境保护与生态整治对策分析、城乡建设与居住环境规划、城市电磁污染分析、减灾防灾决策支持、人口普查、人口预测、人口规划、城市产业经济发展规律及机制等的模拟、城市交通规划与优化、城市零售等服务网点规划、城市不同时间及空间尺度上的形态演变等政府决策支持应用等领域。

从全球范围看，高性能计算机系统在主要发达国家和部分发展中国家都有部署，其中美国数量最多。在美国主要是政府机构支持公共计算平台的建设，如能源部(DOE)、国防部(DOD)、美国自然科学基金(NSF)，很多州政府也直接为国家、州或大学各级别公共计算平台提供资助，包括美国最知名的超级计算中心圣地亚哥超级计算中心(SDSC)、美国国家超级计算应用中心(NCSA)、匹兹堡超级计算中心(PSC)、美国劳伦斯·利弗莫尔国家实验室(LLNL)、美国阿贡国家实验室(ANL)、美国橡树岭国家实验室(ORNL)等。欧盟在每一期欧盟研究与技术开发框架项目中，均对高性能计算投入巨资。英国是欧洲最大的超级计算使用者，主要有爱丁堡并行计算中心(EPCC)和曼彻斯特大学学术研究计算服务设施中心(CSAR)。德国在TOP500拥有的装机数量基本与英国相当，3个国家级超级计算中心分别是斯图加特的HLRS、Julich的NIC以及慕尼黑的LRZ。法国紧随德国和英国之后，最大的超级计算机由法国原子能委员会运行，另外还有两家学术性超级计算中心CINES和IDRIS。欧洲其他国家由于国家规模小，超级计算中心的数目较少。总体来说，欧洲的超级计算中心从设施、运营模式、客户支持到应用领域均有诸多特色。

日本较大的超级计算中心有地球模拟器中心、日本物理和化学研究所(RIKEN)、日本国立先进工业科学和技术研究所、日本国家宇航实验室(JAXA)等。

可以说,高性能计算在国家安全和科学发现中已经扮演了十分关键的角色,各发达国家对高性能计算相关领域的科研、产业投入力度巨大,而且投入具有持续、时间跨度长等特点,因此其在高性能计算科研与技术方面具有良好的基础,积累了丰富的经验,储备了丰富的人才,高性能计算对其经济建设的贡献度不断提高,从而形成了公共计算发展的良性循环。

高性能计算机作为现代服务业的创新载体之一,也必将为企业自主创新的发展推波助澜。互联网、电信网、广电网融合是一种必然,移动互联网日益增强的媒体化应用趋势、视频化应用趋势使3G成为三网融合的重要切入点。网络的融合和发展,将有力促进新计算模式、新服务模式的形成和发展。按照互联网领域的长尾定律,20%的需求有较大的经济规模,80%的需求应用多样、单个应用资源需求规模较小、单个应用需求并不一定复杂但用户量大、个性化程度高,在SOA、Web Services、网格、云计算等相关技术的驱动下,为每个需求打造虚拟应用,开创SaaS、PaaS、IaaS等服务新模式是一种必然。

30年来,高性能计算机技术经历了一个从向量并行向大规模并行的发展时期,从20世纪70年代的向量并行发展到紧耦合共享存储多处理机并行处理结构,再发展到20世纪90年代的基于商用微处理器的大规模并行处理系统。1996年,美国ASIC RED大规模并行计算机首次突破了万亿次计算性能。2003年,日本采用定制向量处理器技术研制出并行向量计算机"地球模拟器",标志着定制技术路线在最高端高性能计算机中占主要地位。2004年,曙光4000A机群系统突破10万亿次峰值运算能力,位列当年TOP500第10名,表明机群成为市场主流。同年,美国IBM公司采用定制通用微处理器研制的BlueGene/L,突破百万亿次计算性能。2008年6月,人类历史上首台Linpack测试达到千万亿次的超级计算机IBM Roadrunner诞生,性能最高达到1026 TFLOPS,标志着混合异构结构成为实现千万亿次量级计算的可行途径之一。

从全球高性能计算机TOP500排行榜的历史来看,TOP500中第1名和第500名保持6~8年的时间距离,也就是说当今的第1名在6~8年后将排名第500名,而8~10年后微处理器芯片就可达到当今第500名的性能。2008年,Roadrunner和随后Jaguar的出现,使我们能够期待其后8年,也就是2016年,千万亿次超级计算时代的真正到来,即千万亿次将是全球TOP500排行榜的入门槛。

1.2 高效能计算机的技术挑战

当前高性能计算机应用正从科学计算领域向商业计算、信息化建设领域拓展,应用的种类不断增多,普及程度逐渐深入,使用者也从专家型研究人员向普通用户拓展。未来应用对计算能力的需求进一步增加,例如,在生物医学领域蛋白质电子

态的计算、药物发明中的遴选过程、蛋白质折叠等，航空航天制造领域的发动机燃烧模拟、机翼设计模拟、气象领域的短期天气预报、局部突发性灾难预报（如洪水、海啸）、核能领域的完全等离子分析、纳米技术领域的复合材料结构分析和功能预测、新材料的发明、天体物理学领域的超新星三维模拟等，这些应用都需要持续100 TFLOPS以上的计算性能，有些应用甚至需要1PFLOPS的持续应用性能。

高性能计算机将面临扩展性、可靠性、功耗、均衡性、可编程性、管理复杂性等诸多挑战[1]：

① 内存和I/O墙（Memory and I/O Wall）：系统结构的失衡问题，存储器性能与处理器性能差距越来越大，本地带宽及延迟和全局带宽及延迟发展不一致所造成的差距形成了阻碍性能提升的“内存墙”；系统I/O能力欠缺，让系统吃得进，吐不出，从而形成“I/O墙”。

② 功耗墙（Power Consumption Wall）：当前功耗已经成为制约计算机系统发展的主要因素之一，未来千万亿次计算时代高性能计算机系统，其每瓦GFLOPS性能应在1.0 GFLOPS/W以上，现在采取的各种应用加速技术，不能从根本上解决能耗问题。

③ 编程墙（Programming Wall）：在编程方面，用户为完成一个具体的并行应用在建模、编码、调试、优化、运行、维护和故障处理上所遇到的各种困难交错形成了“编程墙”，怎样利用为数众多的处理器海及怎样面对数十万并发线程，是面临的严峻挑战。

④ 复杂性墙（Complexity Wall）：在管理方面，高性能计算机软/硬件部件数高速增长，管理的复杂性随之成倍增长，形成了新的“复杂性墙”。

⑤ 可靠性墙（Availability Wall）：对于高性能计算机系统来说，可靠性也是其挑战之一，当其扩展到上万颗CPU以及几百Terabytes内存时，硬件系统的可靠性很难保证；同时，在这样的大规模系统中，软件错误也很难避免。

⑥ 扩展性墙（Scalability Wall）：当系统规模扩展到数万个以上处理器时，延时将变成一个非常重要的问题。同时，目前基本上没有应用软件能有效地扩展到上万个处理器的规模，需要重新设计软件模型，以适应系统的大规模扩展和求解问题的大规模扩展。

虽然多核处理器已经成为构建高性能计算机的基础，但多核处理器也对传统的系统结构提出了新的挑战，例如：如何对芯片级、板级、系统级三级并行结构进行均衡设计和并行编程，如何将通信延伸到多核内，发挥由几十万个处理器核构成的大规模并行系统的计算能力，如何增加应用的可移植性，如何减少多核带来的存储器壁垒的加剧，这些都是高性能计算机面临的技术挑战。

1.3 高效能计算机

内存墙、I/O 墙、功耗墙、复杂性墙、编程墙、扩展性墙等成为高性能计算机发展的主要瓶颈。美国国防部于 2002 年制定了“高效能计算系统”(High Productivity Computing Systems,HPCS)研究计划,提出了以高效能作为新一代高性能计算机研制的目标,IBM PERCS、Cray Cascade、SUN Hero 成为首批入选计划,高效能代表了高性能计算机研究的新方向,它包含高性能、可编程性、可移植性、稳定性等多个方面的要求:

高性能(Performance):在国家安全应用方面,由上千个节点组成的系统的计算效能将提高 10 到 40 倍。

可编程性(Programmability):减少应用方案的开发时间,降低系统运行及维护成本,提高系统的使用效率。

可移植性(Portability):将高效能应用软件与系统平台分离开。

健壮性(Robustness):针对外界攻击、硬件故障及软件错误,开发相应的保护技术,为用户提供增强的可靠性,减少恶意行为的风险。

高效能计算机研究包括平衡处理器、存储器、网络带宽、系统软件与开发语言的均衡系统结构、健壮性策略、新的度量准则和基准测试程序、系统自适应性、节能以及简化管理复杂性等诸多内容。当前 HPCS 项目已完成最后阶段的研究,2010 年前基于 IBM PERCS 系统和 Cray Cascade 系统完成两台千万亿次高效能计算机系统的研制。

IBM PERCS 系统基于 Power7 微处理器、AIX 操作系统、通用并行文件系统(GPFS)、IBM 并行计算环境、互联和存储子系统进行开发。当前 IBM PERCS 项目已经公布的几个研究方向有:片上多处理器(CMP);智能内存,在每个 DIMM 内存条上增加一个智能 Hub 芯片,实现预取、Scatter/Gather、重排序、缓存等功能;全局名字空间支持;混合型 DSM,通过 X10 编程语言实现,支持 OpenMP 程序;异步 SMP,简化目前 SMP 硬件一致性协议,使之更接近软件的应用模式;片上 FIFO,将同步和数据传送结合,减少应用中的延迟;基于目录的 Cache 一致性协议等。IBM 还计划开发高效用软件和开发工具以提高开发人员的生产率。

Cray 公司联合了 Stanford、Caltech/JPL 和 Notre Dame 学院的研究人员共同进行 Cascade 系统开发。它具有独特的处理器设计,节点处理器有机地结合了向量处理器、流处理器、多线程处理器设计,轻量级处理器采用了 PIM 技术和多线程技术。存储系统采用 UMA + NUMA 共享内存方式,并提供了灵活的地址变换和分布。Cray Cascade 系统本质上是一个能在单系统中提供包括标量、FPGA 和混合矢量/超级多线程(MMT)处理的 MPP 系统,该系统将采用统一的高带宽光互连网络,提供分布式共享内存、多层次多线程执行模型、硬件支持的分析和调试功能[2]。Cray 基于 Cascade 系统开发的代号为“Jaguar”的千万亿次超级计算机系统采用

37 538颗四核 AMD Bacelona 处理器,基于 Cray 专用 Sea Star 3D Torus 互连网络和 AMD HyperTransport 总线实现互连。

SUN 公司虽然没有获得继续支持,但是 SUN HERO 计划采用的 Sea of Memory 技术、Proximity Interconnect 技术、Guarded Pointer 技术和 Interval Arithmetic(IA)技术,将得到继续发展。

当前除了上述 IBM、Cray、SUN 等企业以外,学术界和企业界也已经开展了高效能计算机相关技术研究,包括应用加速、混合异构结构、芯片级、系统级以及基础架构级节能等技术。

当前高效能计算机技术已经取得了很多进展:

① 采用多核芯片构建高性能计算机系统,为提高计算机的效能和可扩展性带来直接的好处。IBM Power、Intel Xeon、AMD Opteron、Sun UltraSparc 都有了 2 ~ 8 核的产品,采用多核后可以不追求单核的高主频,多个较低主频的核构成的 CPU 功耗会减小,也较容易增加片内的聚合计算能力,降低系统空间占用。

② 在应用加速方面,由于 FPGA 可以根据不同的应用实现可重构计算,适应不同计算模型,同时 FPGA 在内存带宽、并行处理和低功耗方面有突出的优势,因此与主处理器配合,可实现提高特定应用性能和降低系统功耗的双重目标,应用前景广阔,是实现高效能计算的有效途径之一;在提高存储器性能方面,基于多层次 Cache、加大处理器和存储器之间的带宽、多线程、预取、PIM 等技术以消除内存墙。

③ 在系统可靠性研究方面,国内外研究主要在硬件可靠性、操作系统可靠性和应用可靠性 3 个方面。硬件可靠性主要沿用 20 世纪 60 年代大型机系统发展起来的一系列基础技术框架,如 N 模冗余、专用组件或模块等,通过硬件冗余达到提高系统硬件可靠性的目的。操作系统可靠性研究主要有操作系统隔离技术、故障忽略技术等,通过将故障忽略或隔离从而减少对应用的影响。针对科学计算应用,多采用检查点技术对应用运行的阶段性结果进行保存,以备在出错时进行恢复。

④ 在解决管理的复杂性方面,包括自动监控全局资源,对监控数据进行深度挖掘、关联分析预测系统行为,根据应用特征动态构造虚拟计算环境,实现应用间性能隔离和安全隔离,对应用软件的可靠性和扩展性提供支持,提供验证、模拟、评价工具,并在保证系统性能的情况下,实施自适应功耗管理及可靠性管理。

⑤ 在解决功耗方面,除采用液体冷却、低功耗专用芯片、芯片级冷却等技术以外,一些系统级节能技术也有望解决高效能计算机能耗问题,包括:基于负载情况动态调整系统状态、实施部分节点或部件的休眠;根据各进程能耗的不同对 CPU 任务队列进行调整。

⑥ 在海量存储和文件系统方面,包括聚合一组存储设备的容量和 I/O 带宽,采用面向对象存储技术开发高性能并行文件系统和支持大规模共享文件系统的存储系统。

⑦ 在并行编程模型方面,近年来分割全局地址空间模型(Partitioned Global Address Space Model,PGAS)广受关注。PGAS 既有共享内存编程模型的易编程性,

又能让程序员控制数据的分布,以达到和消息传递编程模型媲美的性能。

虽然当前学术界和工业界在高效能计算机研究领域已取得长足的进展,但与高效能计算的目标相比仍然有很大差距,研制高效能计算机是一个长期的实践过程。

曙光5000A是由中科院计算所和曙光公司联合研制的以高效能为目标的百万亿次量级高性能计算机,该系统的研制得到了国家"八六三"高技术研究发展计划的支持,并作为公共计算平台部署在上海超级计算中心。本书的部分研究成果已应用于曙光5000A高效能计算机系统。

1.4 本书的组织

本书从高效能计算机若干关键技术的研究与实现出发,第1章概述了高性能计算机的发展历程及趋势,阐述了当前基于机群架构实现千万亿次扩展面临的主要技术瓶颈,指出高效能是未来高性能计算机的发展方向。第2章描述了自适应功耗管理技术,详细阐述了基于遗传算法的功耗调度策略。在第3章中,基于高效能计算机管理的复杂性问题,分析研究了层次化自主管理技术。在第4章中着重研究应用加速器技术,基于网络安全应用,分别从CBF HASH算法、数据预处理、五元组过滤等几个方面详细介绍了网络应用加速器的设计及实现技术,并同时介绍了BLAS加速的设计,对其Linpack加速效果做了分析。第5章介绍了基于刀片结构的曙光5000A高效能计算节点的设计及实现。第6章通过对计算机系统内存资源进行虚拟化和网格化,实现了一个网络内存系统,并通过多个应用对其使用效果进行了验证。第7章着重介绍了事件流相关技术研究及事件流应用技术,基于网络监控系统对事件流的查询特性及查询调度进行分析,并设计实现了一个并行查询引擎。第8章着重介绍了并行模拟及并行模拟器的相关技术,提出了基于阻塞/唤醒机制的同步策略、无锁调度和通信机制,以及线程安全的缓冲区管理等方法。第9章着重介绍了并行模拟引擎SimK的设计与实现。第10章概述了曙光5000A系统的实现及评价,包括总体架构、硬件及软件架构,提出了高性能计算机相对效能评价指标RPI(Relative Productivity Indicator)。

高效能计算机技术的发展是一个长期的实践过程,本书虽然在自适应功耗管理、应用加速、高密度计算节点、自主管理、网络内存、事件流应用、并行模拟等方面做了有益的探索,但必须看到,上述技术的研究与实现和高效能计算的总体目标相比仍然有很大差距,未来高效能计算机技术仍将高速发展。