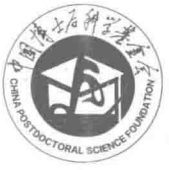
博士后文库
中国博士后科学基金资助出版

异构计算系统调度理论与方法

唐小勇 著

科学出版社



博士后文库
中国博士后科学基金资助出版

异构计算系统调度理论与方法

唐小勇 著

科学出版社
北京

内 容 简 介

随着信息技术的广泛应用和快速发展，以网络计算与分布式计算为基础的异构计算系统因其低成本与高性能而受到企业界和学术界的广泛关注。本书针对异构系统可靠性、安全性、任务计算量随机性、网络通信竞争和能耗等性能要素，从资源管理与任务调度角度提出一系列可行解决方案，以期提高其大规模计算应用性能。

本书作为异构计算系统资源管理与任务调度专著，主要面向从事高性能计算、并行分布式处理和云计算研究的科研人员，适合从事并行分布式应用的各类工程师和研发人员。同时，本书也可作为高等学校和计算机类科研机构高年级硕士研究生与博士生的教材。

图书在版编目 (CIP) 数据

异构计算系统调度理论与方法/唐小勇著. —北京：科学出版社，2017.11

（博士后文库）

ISBN 978-7-03-055071-2

Ⅰ. ①异… Ⅱ. ①唐… Ⅲ. ①并行算法②分布式算法 Ⅳ. ①TP301.6

中国版本图书馆 CIP 数据核字(2017)第 267489 号

责任编辑：陈 静 金 蓉 / 责任校对：郭瑞芝

责任印制：张克忠 / 封面设计：陈 敬

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2017 年 11 月第 一 版 开本：720×1 000 1/16

2017 年 11 月第一次印刷 印张：11

字数：202 000

定价：60.00 元

（如有印装质量问题，我社负责调换）

《博士后文库》编委会名单

《博士后文库》序言

1985年，在李政道先生的倡议和邓小平同志的亲自关怀下，我国建立了博士后制度，同时设立了博士后科学基金。30多年来，在党和国家的高度重视下，在社会各方面的关心和支持下，博士后制度为我国培养了一大批青年高层次创新人才。在这一过程中，博士后科学基金发挥了不可替代的独特作用。

博士后科学基金是中国特色博士后制度的重要组成部分，专门用于资助博士后研究人员开展创新探索。博士后科学基金的资助，对正处于独立科研生涯起步阶段的博士后研究人员来说，适逢其时，有利于培养他们独立的科研人格、在选题方面的竞争意识以及负责的精神，是他们独立从事科研工作的“第一桶金”。尽管博士后科学基金资助金额不大，但对博士后青年创新人才的培养和激励作用不可估量。四两拨千斤，博士后科学基金有效地推动了博士后研究人员迅速成长为高水平的研究人才，“小基金发挥了大作用”。

在博士后科学基金的资助下，博士后研究人员的优秀学术成果不断涌现。2013年，为提高博士后科学基金的资助效益，中国博士后科学基金会联合科学出版社开展了博士后优秀学术专著出版资助工作，通过专家评审遴选出优秀的博士后学术著作，收入《博士后文库》，由博士后科学基金资助、科学出版社出版。我们希望，借此打造专属于博士后学术创新的旗舰图书品牌，激励博士后研究人员潜心科研，扎实治学，提升博士后优秀学术成果的社会影响力。

2015年，国务院办公厅印发了《关于改革完善博士后制度的意见》（国办发〔2015〕87号），将“实施自然科学、人文社会科学优秀博士后论著出版支持计划”作为“十三五”期间博士后工作的重要内容和提升博士后研究人员培养质量的重要手段，这更加凸显了出版资助工作的意义。我相信，我们提供的这个出版资助平台将对博士后研究人员激发创新智慧、凝聚创新力量发挥独特的作用，促使博士后研究人员的创新成果更好地服务于创新驱动发展战略和创新型国家的建设。

祝愿广大博士后研究人员在博士后科学基金的资助下早日成长为栋梁之才，为实现中华民族伟大复兴的中国梦做出更大的贡献。

中国博士后科学基金会理事长

前　言

随着社会经济与信息技术的飞速发展，以网络计算和并行计算为基础的异构计算系统已成为 IT 的重要发展方向。随着系统规模持续扩大，系统平均无故障时间越来越低，任务执行行为可靠性已成为保证并行应用成功的关键；分布式系统为实现信息保密性、数据完整性和边界安全保护而面临大量安全威胁，这些威胁主要源于网络环境下资源与用户行为的不可控性和不确定性；并行应用实践表明任务实际执行时间往往受输入数据、系统环境和任务本身 If 分支的影响而具有随机性；对于任意处理机网络拓扑结构的异构计算系统，任务间通信竞争不可忽略，这种任务通信对整个并行应用程序执行具有重要影响；随着系统规模持续扩大，能耗也持续增长，大大增加了计算成本。面对由可靠性、安全性、任务计算量随机性、网络通信竞争和能耗等因素导致的系统性能降低，本书从资源管理与任务调度角度提出一系列可行的解决方案，以期提高异构系统应用性能。

针对上述问题，本书分 9 章进行阐述。第 1 章介绍高性能计算机发展历程、异构计算系统基本概念及典型异构计算技术。第 2 章主要介绍异构系统资源特征、资源管理与任务调度概念和基于任务优先约束的 DAG 调度模型，然后简述 DAG 模型调度方法、思想、分类和经典调度算法 DLS、MH、HEFT。第 3 章提出基于最短路径搜索技术的最早通信完成路径查找算法(EFCS)，采用插入式链路实现策略，以达到通信边的动态调度；对于任意处理机网络环境下的任务优先级计算问题，受 HEFT 算法启发提出递归优先权计算公式，并且按非升序排列获得各任务优先级。同时为了降低算法执行时间，采用 OpenMP 实现算法的并行化。第 4 章针对异构计算系统任务在不同处理机上执行时其计算成本不同的特点，克服以任务执行平均值、中间值、最好值或最坏值等方法给任务调度带来的困惑，提出异构计算系统计算能力异构因子，并依此实现任务优先级计算和处理机选择。在此基础上，提出基于可靠性驱动的分层任务调度算法(HEFD)。第 5 章利用 Weibull 分布分析任务在处理机上执行的可靠性、数据在通信网络上通信的可靠性及相互关系，建立应用程序任务执行行为可靠性模型。在此基础上提出可靠性驱动的最早完成时间任务复制调度算法(REFTD)。第 6 章针对大规模网格的异构性、动态性和广域性，提出可靠性驱动的层次调度体系。分析任务在处理机上执行的可靠性、数据在通信网络上通信的可靠性及相互关系，建立应用程序任务执行行为可靠性模型。在此基础上提出可靠性驱动的分层任务调度算法(HRDS)。其中，全局任务调度器负责把应用程序分配给虚拟节点，局部调度器则在虚拟节点内实现基于 DAG 模型的任务调度与分配。第 7

章针对分布式计算系统面临的安全威胁，受经济学品牌形象实践与心理学理论启发，在研究分布式信任表现形式与特点基础上提出基于博弈论微分对策技术的信任值动态量化计算方法。依据用户任务安全需求和分布式系统提供的安全信任保障，提出任务执行行为安全性开销计算方法和安全性风险评估技术。最后，提出考虑任务执行行为安全开销的调度模型和调度算法(SDS)。第 8 章首先证明基于 DAG 模型随机任务调度长度的下限是以任务期望构成的确定型任务调度长度，并利用 Clark 方程实现并行任务完成时间期望与方差的计算，在此基础上提出随机 sb-level 近似计算算法。受确定型表调度算法 DLS 启发，提出针对随机任务调度问题的随机动态级调度算法(SDLS)。第 9 章研究异构计算系统能耗与任务间相互依赖关系，建立任务执行时间随机性与能耗模型，提出时限能耗约束下的任务调度算法(ESTS)。

本书是作者多年来从事异构计算系统资源管理与任务调度研究的结晶。本书部分研究内容得到国家自然科学基金重点项目(61133005)、国家自然科学基金面上项目(61370098，61672219)、湖南省自然科学基金面上项目(2015JJ2078)和计算机软件新技术国家重点实验室开放课题(KFKT2016B02)等的资助，在此表示诚挚谢意。同时，本书在研究和编著过程中得到了湖南大学李肯立教授、李仁发教授、李克勤教授；湖南农业大学南方粮油作物协同创新中心廖桂平教授、方逵教授等的指导和帮助，在此向他们表示感谢。唐卓、吴帆、曾锃、Bharadwaj Veeravalli、秦啸、童钊、张龙信、李小春、中国博士后管理办公室等在本书的编著和出版过程中提出了宝贵意见，在此再次表示感谢。

由于本书涉及的专业知识面较广，作者水平有限，书中难免有不足之处，恳请广大读者批评指正。

作　者

2017 年 6 月

目　录

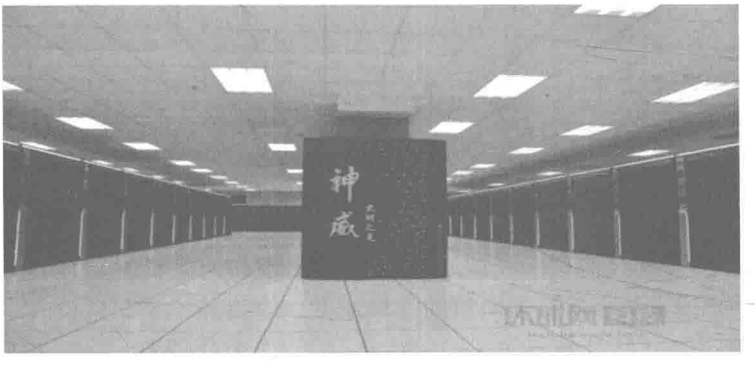
第1章 绪 论

1.1 高性能计算机发展历程

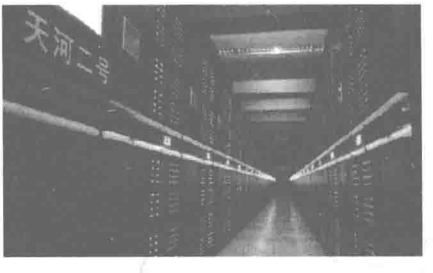
人类对高性能计算的需求一直是永无止境，在这种需求驱动下计算机的性能不断获得发展。20 世纪 60 年代初，Gray 公司制造了著名的 CDC6600 大型机，它使用非对称共享存储结构和流水线技术。1970 年出现的向量计算机则通过在计算机中加入流水部件来提高科学计算中向量的计算速度。1972 年，伊利诺伊大学和宝来公司联合研制的第一台并行计算机 ILLAC IV 包含了 32 个处理单元环形拓扑连接，称为单指令多数据流(single instruction multiple data，SIMD)类型机器。1976 年，Gray 公司推出了采用精简指令集计算机(reduced instruction set computer，RISC)的第一台向量计算机 Cray-1，其性能比当时标量系统高出一个数量级。在向量计算机方面，中国具有代表性的计算机是国防科学技术大学的“银河一号”和中国科学院计算技术研究所的 757 计算机。并行处理机于 20 世纪 80 年代初开始大量进入市场，多采用 SIMD、向量机和多指令流多数据流(multiple instruction multiple data，MIMD)结构。因 MIMD 计算机性价比高、可扩展性强，且人们一致认为传统向量处理机很难突破万亿次大关，MIMD 分布存储多处理机是唯一可以达到万亿次速度的技术。90 年代初，共享存储器结构的大规模并行计算机获得了新发展：IBM 将大量早期 RISC 微处理器通过蝶形网络联结起来。同一时期，基于消息传递机制的并行计算机也开始不断涌现：加州理工学院成功地将 64 个 i8086/i8087 处理器通过超立方体互联结构联结起来。Intel iPSC 系列、INMOS Transputer 系列、Intel Paragon、IBM SP2 等都是基于消息传递机制的并行计算机。从 20 世纪 90 年代开始，大规模并行处理(massively parallel processing，MPP)系统开始得到发展并逐渐成为主流。MPP 系统由多个微处理器通过高速互联网络构成，处理器之间通过消息传递方式进行通信和协调。典型的 MPP 系统如美国的 CM-5，采用超过 1000 个 SPARC 微处理器。1996 年，美国 ASIC RED 大规模并行计算机突破了万亿次计算性能。

20 世纪 90 年代后期，集群(cluster)技术逐渐成为高性能计算机主流。戴尔公司首先构建的集群系统 ThunderBird，系统 Linpack 测试性能达到 39Tflop/s。2004 年我国曙光 4000A 集群系统突破了 10 万亿次峰值运算能力，位列当年 Top500 第十名。随着科技的飞速发展，在科学研究和工程计算需求牵引下，超级计算机系统发展异常迅猛。在最新公布的全球计算机 Top500 排行榜上[1]，国家超级计算无锡中心的“神

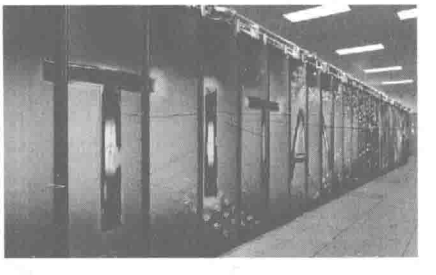
威·太湖之光”以 93014.6 Tflop/s Linpack 值和 125436 Tflop/s 峰值速度位列第一。国家超级计算广州中心的“天河二号”以 33862.7Tflop/s Linpack 值排名第二；美国能源部下属橡树岭国家实验室的 Titan 超级计算机以 17590Tflop/s 的运算速度位居第三。图 1.1 是这三台超级计算机的示意图。



(a) 神威·太湖之光



(b) 天河二号



(c) Titan

图 1.1　世界超级计算机 Top500 前三名计算机

1.2　异构计算系统概述

随着以信息技术变革为主的社会经济飞速发展，大型复杂科学工程计算和海量数据处理对高性能计算需求越来越高。由于硬件局限，单一计算机很难满足这些问题的海量计算需求。同时，高速互联网络与超大规模高性能集成电路技术迅猛发展，可移植高性能通信软件广泛应用，以网络为基础的高性能并行分布式计算成为可能。例如，1996 年，世界范围的 700 多台工作站和 PC(personal computer)协同计算发现了第 35 个素数；1997 年，分布在欧洲的 3500 台工作站协同工作破解了 48 位 RSA 密码，同年，使用约 78000 台计算机破解了 56 位数据加密标准(data encryption standard，DES)密码。这些工作都说明了基于网络的计算能力。

这种通过网络互联利用各类计算资源(包括超级计算机、PC、嵌入式系统、数

据源、仪器等)以低成本形成的高性能协同计算环境统称为分布式系统，用以解决许多中粗粒度的复杂科学计算、海量数据搜索处理问题，如高精度的中尺度数值天气预报、图形渲染、核爆炸模拟、新药筛选、核爆炸模拟、飞行器数字模拟、作战模拟和金融电子服务等[2]。并行分布式计算综合了网络计算与并行计算方面的主要技术，已成为高性能计算领域的一个重要发展方向。

并行分布式计算在异构计算系统上进行，通常称为异构计算。人们已从不同的角度对异构计算进行了定义，综合起来我们给出如下定义：异构计算是一种特殊形式的并行和分布式计算，它由能同时支持 SIMD 方式和 MIMD 方式的单个独立计算机，或是用由高速网络互联的一组独立计算机来完成计算任务。它能协调地使用性能、结构各异的机器以满足不同的计算需求，并使代码(或代码段)能以获取最大总体性能的方式来执行。概括来说，理想异构计算具有如下要素。

(1)它所使用的计算资源具有多种类型的计算能力，如 SIMD、MIMD、向量、标量、专用等。

(2)它需要识别计算任务中各子任务的并行性需求类型。

(3)它需要使具有不同计算类型的计算资源能相互协调运行。

(4)它既要开发应用问题中的并行性，还要开发应用问题中的异构性，即追求计算资源所具有的计算类型与它所执行的任务(或子任务)类型之间的匹配性。

(5)它追求的最终目标是使计算任务的执行具有最短时间。

因而异构计算是一种使并行分布式任务与计算处理资源有效协调计算的技术，具有应用广泛、可扩展性强和性价比高等特性。

近年来工业界和学术界陆续提出与实现了许多经典的异构计算系统，主要包括：P2P 计算(peer to peer computing)、集群系统、网格计算(grid computing)、多核 CPU(central processing unit)与众核协同计算以及云计算(cloud computing)等。在这些技术和项目的支持下，已开展了许多大型工程如预测地球气候(Climateprediction.net)、寻找外星文明(SETI@Home)、开发和利用粒子对撞机(LHC@home)、寻找引力波存在的证据(Einstein@Home)、清水计算(computing for clean water)、模拟蛋白质折叠(folding@Home)、寻找梅森素数 GIMPS (great internet mersenne prime search)(Prime95)等[3]。

1.3 典型异构计算

1.3.1 P2P 计算

P2P 计算也称对等计算，定义为通过直接交换共享分布式系统计算资源和服务。在 P2P 计算系统中，成千上万台彼此连接的计算节点都处于对等位置，这些

地位相等的节点可以互相进行资源利用和数据共享，不需要通过服务器来转接与通信，这样可以减少对服务器的依赖，也就降低了对服务器的性能要求(软件、硬件要求)。P2P 计算技术广泛应用于文件共享、协同服务计算、电子商务、深度搜索引擎等方面。

1.3.2　集群计算

集群系统由两台或多台计算处理资源(简称为节点)通过网络互联而成，节点可以是完整的 PC，也可以是对称多处理(symmetric multi processing，SMP)系统或者工作站等。每个节点都有单独的处理单元、存储器系统、I/O 接口及操作系统，从而可以单独地完成应用程序计算，也可以作为系统中的节点协同完成并行应用程序中的任务。对于用户而言，集群就是一个单一系统，拥有所有软硬件资源，不需考虑这些资源的位置与使用方式。目前集群系统采用的常见逻辑体系结构是由访问层、聚合层和核心层构成的三层胖树(three-tier fat tree)体系，如图 1.2 所示。

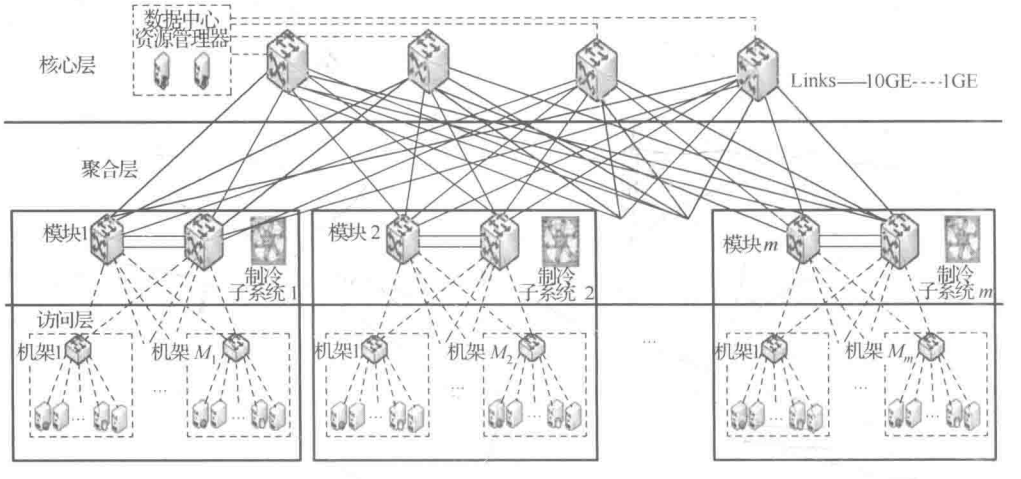


图 1.2　集群系统三层胖树体系结构示意图

依据组成集群计算系统节点(计算机)间体系结构的异同性，集群可分为结构相同的同构系统与体系各异的异构系统。如果集群系统按功能和结构分，则集群计算分为高性能计算集群(high-performance clusters)、负载均衡集群(load balancing clusters)和高可用性集群(high-availability clusters)等。

高性能计算集群主要将高性能应用计算任务分配到集群计算节点上并行执行来提高计算能力，因而在大型复杂科学计算领域里有广泛的应用。现在主流的高性能计算(high performance computing，HPC)都采用 Linux 操作系统和并行计算软件来完成大规模计算。这种集群配置通常称为 Beowulf 集群。这类应用程序一般运行特定的算法库，例如，专为科学计算设计的信息传递接口(message passing interface，MPI)库。

负载均衡集群运行时，一般通过一个或者多个前端负载均衡器，将工作负载迁移到后端服务器上，从而使整个系统具有高性能和高可用性。这种计算机集群有时也被称为服务器群，Linux 虚拟服务器(Linux Virtual Server，LVS)项目在 Linux 操作系统上提供了最常用的负载均衡功能。

高可用性集群一般是指当集群中有某个节点失效的情况下，其上任务会自动转移到其他正常运行节点上，从而能提高应用程序执行可靠性。这种集群还指能对某节点进行离线维护再上线的能力，该过程并不影响整个集群系统运行。这样的系统一般具有高可用性。该系统能够提供一种价格合理、高性能和高可用的并行分布式计算解决方案。

目前主流的超级计算机都采用集群计算体系来构成系统，如世界超级计算机排名 Top500 中的“神威·太湖之光”“天河二号”、Titan、K-computer 等。图 1.3 是国家超算长沙中心的“天河一号”超级计算机应用集群技术的具体体系结构图。

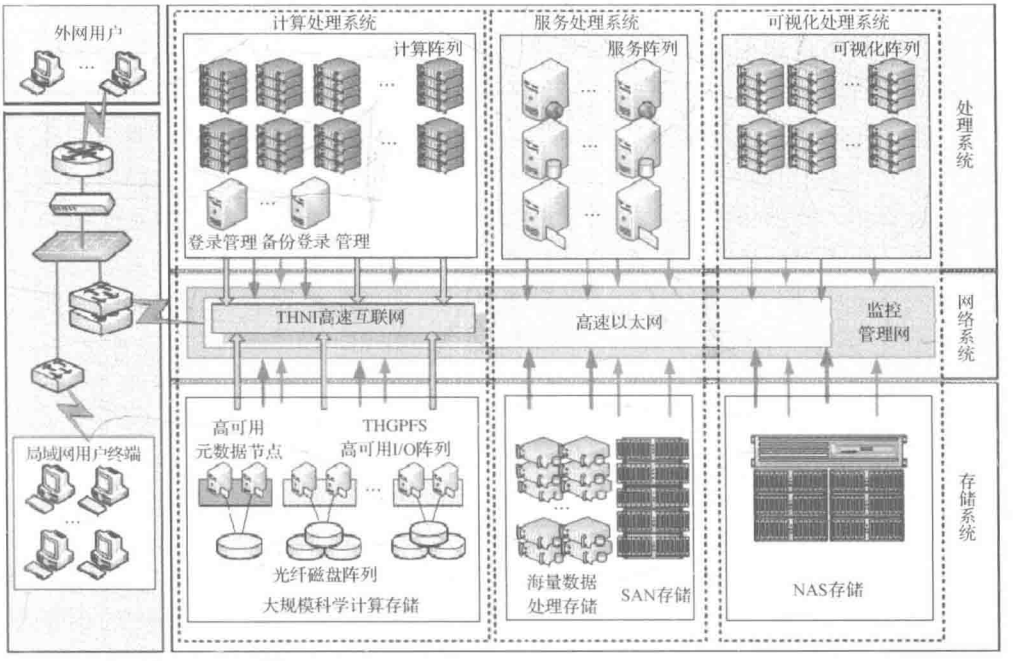


图 1.3 “天河一号”体系结构图

1.3.3 网格计算

网格计算通过高速互联网络把广域分布的高性能计算机、大型数据库、传感器、远程设备、信息资源等联结成一台巨大的“虚拟异构超级计算机”，虽然这些性质各异的网格资源分布在各自不同的计算机上，这些计算机可能有不同的操作系统、技