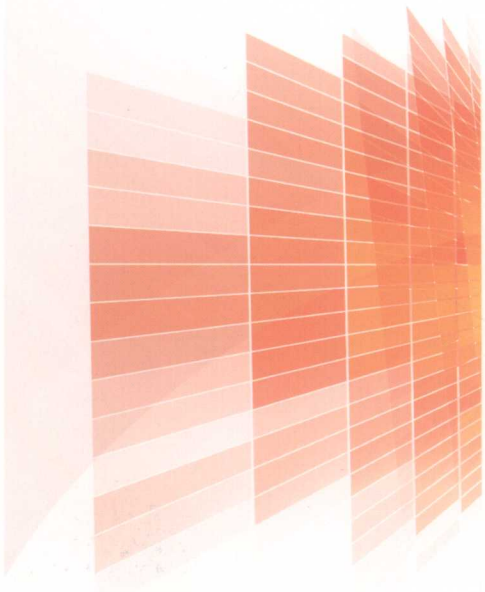
大规模分布式系统的任务分配算法

罗香玉◎著



西北工业大学出版社

大规模分布式系统的任务分配算法

罗香玉 著

西北工业大学出版社
西安

【内容简介】 本书分14章，分别介绍分布式系统分配问题概述、分布式系统任务分配研究现状、分布式储存系统概述、文件放置算法研究现状、一种基于非一致副本数的文件放置算法、一种不依赖文件访问热度信息的放置算法、云计算系统概述 、云计算虚拟机分配算法研究现状、虚拟机分配性能影响因素分析、一种基于资源需求分布特征的虚拟机分配法、大图分布式处理概述、大图划分算法研究现状、大图结构特征对划分效果的影响和一种基于结构特征的大图划分算法等。

本书主要适用于对分布式计算平台内部算法研究感兴趣的读者，也可供相关技术人员阅读参考。

图书在版编目(CIP)数据

大规模分布式系统的任务分配算法/罗香玉著. —
西安:西北工业大学出版社,2018.9
ISBN 978-7-5612-6245-0

Ⅰ.①大… Ⅱ.①罗… Ⅲ.①分布式操作系统—研究
Ⅳ.①TP316.4

中国版本图书馆CIP数据核字(2018)第209090号

策划编辑:季 强
责任编辑:王 静

出版发行:西北工业大学出版社
通信地址:西安市友谊西路127号 邮编:710072
电 话:(029)88493844 88491757
网 址:www.nwpup.com
印 刷 者:西安真色彩设计印务有限公司
开 本:710mm×1000mm 1/16
印 张:12
字 数:211千字
版 次:2019年5月第1版 2019年5月第1次印刷
定 价:48.00元

前　言

分布式计算是一个历久而弥新的领域，分布式计算为IT其他领域的发展铺平了道路，其他领域的发展又促进了分布式计算领域的发展。历史表明，分布式计算总能与最新的概念和思想相遇，并诞生实实在在的产品，促进新概念新思想在工程领域的大发展。20世纪90年代，分布式计算与面向对象相遇，诞生了CORBA；21世纪的前10年，分布式计算与云计算相遇，诞生了Hadoop；紧接着，分布式计算与大数据相遇，诞生了Spark；分布式计算与深度学习相遇，诞生了TensorFlow。

自2006年开始，笔者一直从事分布式计算相关的研究。在东南大学分布式计算实验室攻读博士学位期间，有幸参与多项大型分布式系统相关的科研项目，完成了CORBA中间件的移植工作、发布/订阅中间件容错机制的实现工作以及大规模分布式存储统一实验平台的设计和开发工作。进入西安科技大学计算机学院工作以来，笔者先后主持了陕西省教育厅专项研究计划项目“访问特征驱动的副本存储策略自动生成”、陕西省自然科学基金项目“面向云计算的虚拟机部署策略自动生成机制研究”和国家自然科学基金项目“基于结构感知的大规模动态图划分算法研究”等课题。这些课题针对的对象都是分布式系统，分别是分布式文件存储系统、云计算系统和分布式大图处理系统。它们处理的任务种类虽然不同，但都涉及分布式计算的一个基本问题，即任务如何在分布式系统各节点之间进行分配的问题。在认真总结研究成果的基础上，形成了本书，旨在为分布式系统任务分配算法的研究提供借鉴。

本书主要适用于对分布式计算平台内部算法研究感兴趣的读者，也可以为相关技术人员提供参考。

写作本书曾参阅了相关文献、资料，在此，谨向其作者深表谢意。

由于水平所限，书中疏漏不妥之处敬请读者批评指正。

著　者

2018年5月

目　录

第一篇　分布式系统任务分配基础

第二篇　分布式存储系统文件放置

第三篇 云计算系统的虚拟机分配

第四篇 大图分布式划分

第一篇

分布式系统任务分配基础

▶第 1 章

分布式系统任务分配问题概述

1.1 分布式系统概述

分布式系统自提出以来，因其良好的可扩展性和鲁棒性广受追捧。在一个分布式系统中，一组独立的计算机展现给用户的是一个统一的整体，就好像是一台虚拟的超强计算机。分布式系统拥有多种通用的物理和逻辑资源，可以动态地分配任务，分散的物理和逻辑资源通过计算机网络实现信息交换。当分布式系统不能满足应用需求时，可以通过引入更多的计算机来实现计算能力和存储能力的无限扩展。当分布式系统中的部分计算机故障时，可以通过其他计算机上的数据备份或者任务执行的备份来保证服务不中断。

在云计算和大数据时代，分布式系统表现出更加强大的生命力，成为云计算和大数据应用必不可少的支撑平台。无论是数据的存储还是计算，都与分布式系统密不可分。

1.2 分布式系统任务分配目标

在分布式系统中，无论是数据存储任务还是分析计算任务，都需要将之合理地分配至多台计算机。若分布式系统的计算机数量给定，则任务分配的目标之一是实现负载平衡，即各台计算机尽可能分担相同负载量的任务，以实现系统整体的服务性能最佳。反之，若分布式系统的性能需求给定，则任务分配的目标之一是尽可能减少所使用计算机的数量，以降低系统资源及能源开销。后者一般可以抽象为装箱问题，每一个任务看作一个物体，而每一台计算机看作一个箱子。

1.3 分布式系统任务分配实例

分布式文件存储系统是一个典型的分布式系统。它能够实现存储空间容量和I/O吞吐量的灵活扩展，是大数据和云计算的基础构件。分布式文件存储系统中的一个基本问题是如何对文件存储任务进行分配，即文件放置问题。一个良好的文件放置算法能够在给定存储节点情况下，尽可能提升整体的I/O访问性能；在给定I/O访问性能条件下，尽可能减少存储节点的数量，从而降低资源及能源成本。

云计算平台是一个典型的分布式系统。它以虚拟机方式为云租户提供服务。云计算系统中的一个基本问题是如何将虚拟机部署至合适的物理机。一个良好的虚拟机部署算法能够在给定物理机情况下，尽可能满足更多虚拟机的资源需求；在虚拟机给定条件下，尽可能减少物理机的使用数量。

图数据是大数据中较难处理的一类。随着图数据规模的不断扩大，图数据处理已步入分布式处理阶段。分布式图数据处理的一个基本前提是实现图数据的合理划分。一个良好的图划分算法一方面能够保证各子图规模相近，从而保证负载均衡，另一方面能够减少交叉边数量，以减少图数据处理阶段的通信开销。

▶第 2 章

分布式系统任务分配研究现状

根据任务分配目标的不同，分布式系统任务分配相关研究工作可分成两大类：一类研究负载平衡问题，另一类研究装箱问题。

2.1 负载平衡研究现状

分布式系统负载平衡是当今计算机领域研究热点之一。由于各分布式节点的系统资源利用率常常不同，这将会影响系统的性能。而负载平衡问题要考虑各节点的计算性能、通信性能等参数，从而提高系统的处理能力和整体性能[1]。因此研究负载平衡问题相当重要。

负载平衡属于 NP 问题，这说明无法获得最优的负载平衡，所以只能通过不断优化以接近最优解。目前对负载平衡的研究在任务调度模型、算法复杂程度、信息获取、数据传输代价、调整策略等方面都还存在一些问题[2]。

而负载平衡分为静态和动态两类[4]。静态负载平衡算法忽略系统当前的负载状态，将新任务依次分配给各节点，但并不能根据当前的负载动态变化，可用性不高；而动态负载平衡算法能根据当前实际情况，将新任务分配给各节点。动态负载平衡算法的关键任务是选择空闲节点，并将新任务转发给它。目前，常见的选择空闲节点的负载平衡策略[5][6]有最快响应法、最少连接法、最低缺失法和随机法。但这些方法仅将一种性能指标作为负载评价指标，这并不能反映节点的负载状况。

静态负载均衡又称为状态无关均衡，动态负载均衡又称为状态相关平衡。负载均衡算法可以分为局部和全局、静态和动态（在全局类中）、最优和次优（在静态和动态两种类型中）、近似和启发式（在次优模型中）、集中式和分散式（在动态类型中）以及协作和非协作的（对分散式）的六大类。Lewis[7]提供了另外一种负载均衡算法的分类，分为单个和多个应用程序、非抢占式的和抢占式的以及非自适应的和自适应的三大类，这是对前一种分类方法的补充。用户可以从有效性、稳定性、可靠性、用户透明性和通用性五方面评价一个负载均衡算法的好坏。

由于通过对静态任务划分的研究[7]，间接地实现了静态负载平衡，故负载均衡算法的研究主要集中在动态负载均衡算法的研究上。Watts J，Taylor S 提出基于减少通信开销的负载平衡策略[9]，VDS[10]和 Millipede[11]通过计算线程的散步和迁移实现了自动动态负载均衡。2002 年李登[12]提出了信息中心调度策略。2008 年周美娜[13]根据大规模分布式系统通信开销时变的特点，提出一种基于随机延迟论的层次结构负载平衡策略，该策略具有应用通信优化的层次结构减小超大规模机群的负载平衡开销、考虑到节点计算速率以及通信介质的随机延迟性和通过广义神经网络理论建模进行延迟预测，从而优化任务通信延迟及迁移延迟的三大特点。2014 年曲全民[14]通过分析目前负载平衡存在的一些问题，涉及了任务调度模型，在此模型基础得出一种高效的负载平衡算法。

负载指标[7,15]可以分为节点指标和网络通信指标。国内的董立岩等人，针对节点计算能力及带宽等方面的异构性导致的系统负载不均衡问题，提出了一种改进的分布式系统负载平衡策略。将模糊综合评判理论运用到节点性能评价中，选出性能最佳的节点，使负载平衡，提高了分布式资源的利用率。实验结果表明，运用此策

略可以准确地选出最佳节点。

虽然负载均衡研究了数十年，但由于负载均衡策略是 NP 完全问题，真正令人满意的实现系统并不多，对它的研究依然是一个难题。

2.2　装箱问题研究现状

装箱问题（Bin Packing Problem，BPP）也称为组合优化问题（Container Loading Problem，CLP），它是求解小物品装入大容器的整体布局方案，并使之在某些约束条件下达到某种优化目标，如消耗的容器数量最少，或装入的物品总价值最高等。它的解的形式是一系列的箱子布局图，布局分为两大类：分层布局和非层布局。它是研究最早、应用最广泛的 NP 难题问题之一[17]。它的研究主要关注输出（价值）最大化与输入（价值）最小化两项指标，以达到最佳装箱效率。输出最大化问题中包含相同物品装箱问题、单大容器放置问题、单背包问题、多个相同大容器的装箱问题、多个相同背包问题、多个异构大容器的装箱问题和多个异构背包问题。输入最小化问题包含了单大物体下料问题、单箱尺寸的装箱问题、多库存尺寸下料问题、多箱尺寸的装箱问题、剩余下料问题、剩余装箱问题和开维问题。研究装箱问题的意义是通过节约箱子、减少装箱工作量和化简装箱流程等手段，最终达到降低付出成本的目的[18]。它在多处理器调度、资源分配和日常生活中的计划、包装、调度等各领域有着广泛的应用背景。装箱问题的某些应用可能要求的附加约束有剪切和定向、重量和重心、叠放次序、物品种类限制、敞口和封闭、在线和离线等，它可以分为优化目标的分类、物品的分类、箱子的分类这三个大的类别。

解决装箱问题的经典算法有在线算法、离线算法、半在线算法和并行算法。装箱问题根据维数主要分为一维、二维、三维三大类。一维问题因其直观性、应用广泛性和约束条件相对较少等原因得到了学者们大量的关注和研究，研究成果众多；三维问题因其复杂性、建模难度大、可行解集合庞大等原因研究成果偏少；二维问题的研究规模介于前两者之间，相关研究成果较多。

对于一维问题，2004 年孙春玲，陈智斌和李建平[19]提出了一个新的近似算法：交叉装填算法，该算法获得了问题的最优近似值 3/2，它的近似值和复杂性从总体上也都得到了较大的提高，但是并不适用于一维经典装箱问题所衍生出来的其他装箱问题。2013 年邵飞牛[20]针对一维装箱问题提出了一种带缓冲箱的启发式算法，并分析了带缓冲箱的启发式算法的平均和最坏情况性能，指出它的平均性能比是，最坏性能比优于最佳适应算法，并通过实验指出参数 K 对算法性能比的影响。

对于二维问题，2006 年 Salto 和 Molina[21]提出了一个三阶段的二维下料方案，结合了顺序算法和并行遗传算法，取得了比普通下料算法更优越的实验效果。2011 年 Charalambous 和 Fleszar[22]针对带剪切约束的 2BP 提出了一个基于偏差和后处理机制的构造型启发式算法，该算法有效避免了因单纯追求箱子的面积利用率而导致小物品的过度使用。2016 年姚怡[23]针对带剪切约束的单箱尺寸二维装箱问题提出了三种不同的算法：VCH－RI，VCH－SP 和 HHBP，这三个算法解的质量和时间性能都在可接受的范围内，理论上采用并行算法进行求解是可行的，但是遇到了将装箱问题分解成若干个尽量相互独立的子问题，如何使用多台计算机同时求解它的问题。

对于三维装箱问题，2007 年张德富，魏丽军，陈青山，陈火旺[24]通过组合拟人启发式和模拟退火算法，提出了三维装箱问题的组合启发式算法。2009 年张德富，彭煜，朱文兴，陈火旺[9]提出了一个高效求解三维装箱问题的混合模拟退火算法。

装箱问题的应用领域广泛，很难有一种装箱算法能够对所有的装箱问题都能适用，故对装箱问题的研究仍然是一个难题。

刘胜等人[25]提出了一种启发式二叉树搜索算法，首先将所有箱子组合成多个优条，每个优条中的箱子沿容器高度方向排成一列；接着开始构建二叉树，其根节点表示空的装箱方案，每个树节点沿长度方向增加一排优条形成左子树节点，沿宽度方向增加一排优条形成右子树节点，二叉树必须扩展到所有叶子节点都无法再放入任何剩余的箱子为止，所有叶子节点中填充率最高的装箱方案即为最终结果。该算法满足三维装箱的 3 个著名的约束条件，分别是方向性约束、稳定性约束、完全切割约束。

崔会芬等人[26]提出一种基于改进遗传算法的人工智能算法，用来实现所建立的优化模型。结合实际装箱问题，分析装箱问题的约束条件，建立数学优化模型，通过将目标函数作为适应度函数和遗传操作中采用排序选择法、部分匹配交叉来实现对传统遗传算法的改进。

李孙寸等人[27]提出多元优化算法 MOA（Multi – variant Optimization Algorithm）。算法通过随机放置和局部调整从而逐步逼近最优解。随机放置是将随机选择的几个箱子装入容器中；局部调整是根据目标函数值对随机放置容器的箱子序列作局部调整优化；通过递推的随机放置和局部调整优化，目标函数值逐步逼近最优值，从而获得一个较为理想的三维装箱方案。