

- 吉林财经大学资助出版图书 -

Research on Key Technology in Data Grid

数据网格几个关键技术的研究

姜建华◎著



科学出版社

吉林财经大学资助出版图书

数据网格几个关键技术的研究

姜建华 著

科学出版社

北 京

内 容 简 介

数据网格是网格计算领域的一个重要分支,也是当前大数据处理领域的研究热点。在数据网格中,数据密集型作业是其中最重要的一类作业类型。因此,如何提高这类作业的处理性能是数据网格中的一个重要研究内容。本书主要探讨如下内容:①数据副本管理问题;②数据密集型作业调度问题;③空间数据网格中的即插即用协议体系结构和设计问题;④空间数据网格中的按需动态扩展协议体系结构和设计问题。

本书适合全国高等院校从事大数据领域研究的研究生、教师、科研工作者,科研机构相关研究人员,高科技企业从事大数据应用开发与管理的专业技术人员,以及大数据相关领域爱好者等参考和阅读。

图书在版编目(CIP)数据

数据网格几个关键技术的研究/姜建华著. —北京:科学出版社, 2017.9

ISBN 978-7-03-054521-3

I. ①数… II. ①姜… III. ①网格计算—研究 IV. ①TP393.028

中国版本图书馆CIP数据核字(2017)第231024号

责任编辑:王喜军/责任校对:樊雅琼

责任印制:吴兆东/封面设计:壹选文化

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

北京厚诚则铭印刷科技有限公司印刷

科学出版社发行 各地新华书店经销

*

2017年9月第 一 版 开本:720×1000 1/16

2017年9月第一次印刷 印张:9 1/4

字数:160 000

定价:68.00 元

(如有印装质量问题,我社负责调换)

前 言

目前,市场上大量存在关于云计算、大数据等方面的通识性书籍,缺乏对其深度介绍和关键问题的认识。这就导致大部分读者难以深入理解当前流行的计算模式的本质内涵,进而难以有效深入领会和探索。本书试图打破这种僵局,对数据网格的几个关键问题的分析和求解,对读者了解数据网格的运作机理、数据密集型作业的计算模式、数据副本管理、存储设备管理等具有关键性作用。

在本书的写作过程中,作者首先考虑的是数据网格中有哪些是读者比较关心的重点难点问题,并以此为出发点,研究基本解决方案,并给出其具体效果。

本书的主要内容如下。

(1) 详细研究副本替换问题。在副本管理中,主要涉及副本创建,而副本创建的过程中不可避免地导致副本替换。然而,有效的副本替换将提升数据密集型作业的处理效率。因此,副本替换成为当前数据网格中的一个研究热点。首先分析当前副本替换所采用的策略,认为当前的副本替换问题的研究主要存在两个问题:①当前副本替换算法仅考虑单个数据文件,而忽视了其关联特性;②当前副本替换算法更多地考虑单个数据文件的访问情况,而忽视了其在数据网格全局中的访问特性。针对第一个问题,本书提出一个基于 Apriori 的关联数据文件替换算法,该算法首先通过各个存储节点的数据文件使用情况挖掘出数据文件的关联规则,然后根据这些关联规则产生副本替换决策,最后与其他副本替换算法的对比实验中表明本书所提出的算法具有较好的效果。针对第二个问题,本书提出一个 LFU-Min 算法,该算法将针对单个存储节点上数据文件的访问情况扩展到数据文件在全局数据网格上的访问情况,认为那些在全局网格上不频繁使用的数据文件具有优先被替换的权利,最终在 OptorSim 数据网格模拟器中的实验结果表明本书所提出的 LFU-Min 算法具有较好的效果。

(2) 详细研究数据密集型作业调度问题。在数据网格中, 针对数据密集型作业的调度好坏将直接影响数据网格的处理效率。在分析当前的数据密集型作业的特点的基础上, 简要阐述数据密集型作业调度的研究情况, 并认为当前的研究存在两个缺陷: ①缺乏有效的 Gfarm 数据网格中的数据密集型作业调度和管理手段; ②在以访问代价为目标函数的数据网格的作业调度策略中, 访问代价的影响因素应考虑处理节点上作业等待队列中的作业行为的影响。针对第一个问题, 本书在分析 Gfarm 和 LSF 之后, 认为可以通过对 LSF 中的插件机制的研究来设计和实现一个基于 Data-aware 的批数据密集型作业调度算法, 并从批作业的大小来选择两种不同的处理策略。针对第二个问题, 本书首先针对由于副本频繁替换导致在调度与运动时的数据分布产生变化, 进而使得访问代价计算产生偏差, 然后认为这种偏差是由于处理节点上的作业等待队列中作业的潜在行为导致估计结果的偏差, 最后设计和实现一个考虑作业等待队列中作业的潜在行为的副本替换算法。为了保证该副本替换算法的高效性, 本书引入一个简单的副本情况非集中式的反馈机制。与传统基于访问代价的副本替换算法进行对比模拟实验分析, 本书所提出的副本替换算法具有较好的效果。

(3) 研究空间数据网格中的即插即用的协议体系结构和设计。首先阐述空间数据网格, 然后对当前即插即用协议进行综述, 最后提出一个空间数据网格中的即插即用协议的体系结构和设计。本书所提出的即插即用协议主要包括设备动态上/下线协议、设备访问控制协议、数据资源动态上/下线协议和数据资源融合协议。相比于其他的设备即插即用协议, 本书所提出的数据资源动态上/下线协议和数据资源融合协议是其有益的补充, 特别是数据资源融合协议的设计为空间数据网格性能的提升提供了保证。

(4) 研究空间数据网格中的按需动态扩展的协议体系结构和设计。首先对当前的按需动态扩展协议进行调研, 并在调研的基础上认为当前的按需动态扩展协议应该考虑分析者行为需求, 提出一个针对空间数据网格的按需动态扩展协议。整个协议分为信息采集层、决策制定层和决策执行层。为了实现这三个层的功能, 分别设计信息采集协议、决策制定策略和决策执行协议。在决策制定策略中, 采

用数据挖掘的分类思想，以原有的决策条件和决策规则为训练数据集，通过挖掘分类规则，找到新数据集的所属类别，最终根据其所属的类别来映射出对应的决策执行方案。

感谢默默付出的王欢女士，感谢同事的帮助，是你们的无私关怀和帮助、最大限度的宽容和付出才使本书能够顺利出版。

本书得到了吉林财经大学的资助，在此一并表示感谢。

由于作者水平与时间的限制，书中难免会存在疏漏之处。如果读者在阅读过程中发现了一些问题，请 E-mail: jjh@jlufe.edu.cn，作者会及时给予回复。

姜建华

2016 年 12 月

于吉林财经大学

目 录

前言

第 1 章 绪论	1
1.1 网格计算研究背景	1
1.2 数据网格研究的理论意义及价值	5
1.3 数据网格研究现状	5
1.4 本书的主要研究方法	9
1.5 本书的结构安排	11
第 2 章 网格与数据网格	13
2.1 网格系统	13
2.2 数据网格系统	14
2.3 数据网格的主要项目	19
2.4 本章小结	22
第 3 章 数据网格中副本替换算法研究	23
3.1 数据网格副本管理	23
3.2 数据网格副本替换综述	24
3.3 ARRA: 基于 Apriori 的关联数据副本替换算法	26
3.4 LFU-Min: 基于 LFU 的全局最少使用副本替换算法	36
3.5 本章小结	42
第 4 章 数据网格中数据密集型作业的调度算法研究	44
4.1 数据密集型计算	44
4.2 研究概述	46
4.3 数据密集型作业在 LSF 和 Gfarm 中的 Data-aware 调度算法	49

4.4	ACPB: 基于访问代价并结合作业潜在行为的作业调度算法	61
4.5	本章小结	75
第 5 章	空间数据网格即插即用协议研究	77
5.1	空间数据网格概念	77
5.2	即插即用机制	78
5.3	空间数据网格即插即用协议	79
5.4	本章小结	102
第 6 章	空间数据网格中数据资源的按需动态扩展协议研究	104
6.1	研究概述	104
6.2	数据资源按需动态扩展协议	105
6.3	本章小结	124
第 7 章	结论与展望	126
参考文献		129
附录		136

第1章 绪 论

1.1 网格计算研究背景

技术的发展是为了解决人们的需求问题，而满足这些需求的同时也就拓展了人的能力，所以需求是创新之母。在 19 世纪 70 年代，贝尔和格雷发明了电话，从此人类就具备了“千里耳”的能力；自研究人员于 1969 年开始网络技术的研究和发展以来，人类就具备了“一网知天下”的能力。而今，人们的需求不仅局限于通过网络来获取信息，更多地将网络视为计算机、系统、软件、服务等，因此人们对网络的要求已经从计算机之间的简单互联到资源的充分共享。自 20 世纪 90 年代中期以来，网格计算技术的发展就在于解决各种资源的动态共享问题。为了更好地让大家认识网格，网格技术往往以电网做比喻，研究人员希望未来的网格服务像人们用电一样去使用网格服务。也就是说，网格技术解决了人们“按需服务”的要求，极大地拓展了人的“服务能力”。网格技术试图使得互联网上的所有资源全面连通，包括计算资源、存储资源、数据资源、通信资源、软件资源以及知识资源等，其目标就是实现整个网络虚拟环境下的资源全面共享和协同工作，消除资源孤岛和信息孤岛。

1.1.1 网格的概念和内涵

网格之父 Foster 将网格计算认为是在动态、多机构的虚拟组织中通过协作等方式解决资源共享问题^[1]。资源共享不仅仅限于文件交换，更包括对计算机、软件、数据和其他资源的直接访问。虚拟组织是指一些个体或者机构具有这种资源共享机制的集合。在 1999 年，Foster 在其著作《网格：21 世纪信息技术基础设施的蓝图》中将网格定义为：网格是构筑于互联网之上的一组新兴技术，其目的是

将高速互联网、高性能计算机、大型数据库、传感器、远程设备等融为一体，为科技人员及普通用户提供更多可用的资源、功能和交互性^[2]。

从网格之父 Foster 的定义中可以看出，网格计算所需要解决的是如何在动态、异构的不同虚拟组织间实现资源的充分共享和协同解决某一个特定问题。因此，若所需要解决的问题有所不同，则对网格的理解会有所区别。

网格可定义为一种在广域范围里的动态地组织、共享和使用各种资源（包括计算力、存储器、数据库、共享设备、通信设施等）的机制，这种机制借助于飞速发展的高速网络进行各种资源的共享和集成，并最终提供了一种超级的、协同的服务能力。其目的是要将广域网络上的各种资源通过网格所建立的机制进行集成和全面共享，并最终成为一台超级计算机，具有无所不能的服务能力。

1.1.2 网格技术的发展趋势

网格技术的发展主要看其体系结构的变化及其应用领域的变化。早期的网格体系结构是 Foster 等提出的五层沙漏结构，如图 1.1 所示。

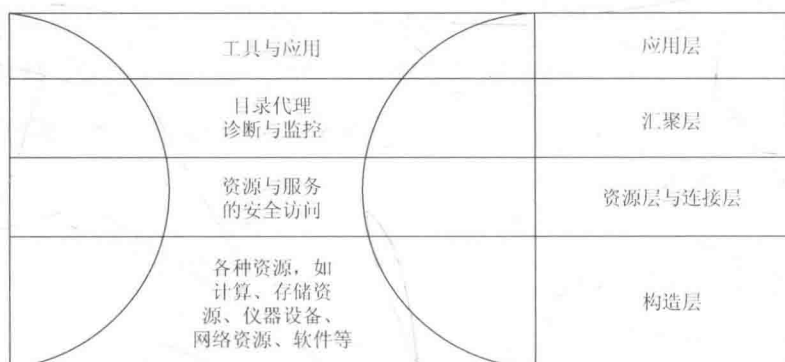


图 1.1 五层沙漏结构体系图

五层沙漏结构参考了 TCP/IP 的设计思想，将网格体系设计为以“协议”为中心，其中资源层与连接层实现网格计算中最核心的协议。核心协议要求数量不多，但要求所有网格应用都能够得到广泛支持。TCP/IP 的成功是层次性协议设计的典

型代表,但这并不能保证借鉴 TCP/IP 体系的网格计算协议体系的成功。由于不同的网格应用的需求有较大的区别,因此很难对各个层次的协议进行标准化。同时,由于相关技术的飞速发展,特别是 Web Service 技术的发展,对网格计算的体系结构的演化起到了积极的作用。

在 2002 年全球网格论坛的论文中, Foster 等^[3]提出了一个全新的网格体系结构,即开放网格服务体系结构(open grid service architecture, OGSA)。这种体系结构是在当时 Web Service 技术的飞速发展的背景下提出的。它将 Globus 和 Web Service 技术进行融合,统一以网格服务的方式对外界提供服务。XML、SOAP、WSDL、UDDI 等技术与平台无关的特性造就了 Web Service 的成功。这些技术使得访问异构平台的各种服务能够顺利进行,大大提高异构系统上的资源可用性和便利性。

近年起步的云计算技术由于得到一些大型公司,如 IBM、Microsoft、Google 等大力推进而发展迅速。维基百科认为云计算是一个能够将动态伸缩的虚拟化资源通过网络以 Web 服务的方式提供给用户的计算模式,对用户而言无需了解如何管理那些支持云计算的基础设施^[4]。简单地说,云计算是一种通过 Web 来获得软件和服务的计算模式。相比于网格计算,云计算的成功主要在于其合理的商业模式,而从计算处理的性能的角度考虑,网格计算更侧重于通过对不同组织的各种资源的有效管理来提升其作业的处理性能。从趋势上看,网格计算的发展为云计算的发展提供了良好的基础设施。

1.1.3 网格计算的主要项目

计算网格和数据网格是网格计算的两个重要分支。下面主要对这两种类型的网格项目进行介绍。

计算网格的主要项目有 Globus^[5]、Nimrod/G^[6]、NetSolve^[7]、AppLeS^[8]、Gfarm^[9]等。

Globus: Globus 项目是网格计算领域最知名的项目。Foster 等^[5]讨论了原有分

布式元计算环境中难以处理异构和动态资源获取的特点,提出了 Globus 解决方案。当前的 Globus Toolkits 已经发展到第 5 版,其已经成为一个工具集合,主要包括安全、信息基础设施、资源管理、数据管理、通信等。自 1996 年启动 Globus 项目以来,已经得到世界众多知名高校和科研机构的合作,同时与 Microsoft 和 IBM 等知名企业进行了战略合作。

Nimrod/G: Nimrod/G^[10]是 Nimrod 系统^[6]的新发展。Nimrod 系统对于一系列静态计算资源具有很好的解决方案,而对于大范围的动态计算资源则缺乏有效的解决途径。因为这些资源大范围分散地分布于不同的管理域,而这些管理域具有各自的用户管理政策,拥有各自的作业管理策略,以及不同的访问代价和计算能力。Nimrod/G 采用 Globus 中间件服务来解决网格计算所需要的在计算网格环境中的动态资源发现和作业分配问题。因此 Nimrod/G 是 Nimrod 和 Globus 的结合系统。

NetSolve: Casanova 和 Dongma^[7]介绍了 NetSolve 项目。NetSolve 是一个客户/服务器系统,采用远程方式解决复杂的科学问题。该系统允许用户访问分布于网络中的各种硬件和软件计算资源。NetSolve 通过搜索网络上的计算资源找到最合适的计算处理节点,同时整个作业处理过程要确保负载平衡。与此同时,容错性是 NetSolve 的另一大特点。

AppLeS: Berman 等^[11]在 1996 年 IEEE 超级计算国际会议上提出了一个应用层次上的调度方案用于解决异构系统网上的作业处理。AppLeS 是一种有效的元计算应用调度,在这种调度中,每个应用都与定制的调度器相结合。通过特定的应用性能模型和从 AppLeS 系统中动态地获取各种资源的信息,AppLeS 调度器进行最佳处理节点的决策^[8]。这种调度的依据就是处理节点可能的负载情况和处理性能的预测。

数据网格是网格计算的另一个重要分支。这种网格类型主要处理数据密集型作业类型。当前主要的数据网格项目包括 Gfarm^[9]、EU Data Grid^[12]、Globus Data Grid^[13]、SRB^[14]和 GriPhyN^[15]等。空间数据网格是一类应用于地理信息领域的的数据网格技术。在这种类型的数据网格中,数据网格所处理的数据是一类海量、多

维、多粒度、异构并具有时间和空间特征的数据。第2章将详细介绍数据网格，细节请参考第2章。

1.2 数据网格研究的理论意义及价值

网格技术至今仍然不够成熟，特别在数据网格领域的研究仍然需要进一步的探索。数据密集型作业是数据网格之上的主要的作业类型。针对数据密集型作业的调度是提升数据网格处理性能的关键点之一，因此有必要去探索 Data-aware 机制的作业调度算法。这种机制研究的意义在于为数据网格中的数据密集型作业调度问题找到一个合适的解决方案。

与此同时，在数据网格中，副本管理的目的是对数据文件及其副本进行有效分布和管理以提升数据密集型作业的处理性能，并减少网络带宽的使用。其中，影响作业处理性能和网络使用的关键在于数据副本在合适的节点创建以及各个网格存储节点上数据文件或副本的替换策略。本书主要针对副本替换策略进行研究，其意义在于找到一个有效的副本替换策略以提升数据密集型作业的处理性能和减少网络带宽在处理时的占用率。

在空间数据网格领域，如何进行副本管理和作业调度仍然是一个研究的热点，同时，其中的存储设备的即插即用机制以及存储空间的按需动态扩展机制也是两个关键点。即插即用机制实现了存储设备在空间数据网格中的动态加入和移除，这种处理方案的意义在于保证了空间数据网格的稳定和可用性。空间数据网格的存储空间按需动态扩展机制为海量数据的存储空间申请、使用和回收等一系列管理问题找到一个有效的处理途径。这种按需动态扩展机制能够智能、有效地进行可用存储空间的申请、使用和回收。

1.3 数据网格研究现状

对于数据网格的研究，大量学者对其各个方面进行了广泛的研究。本书就其中的几个关键技术进行深入研究，主要的研究方面包括作业调度、副本复制策略、

空间数据网格中的即插即用机制和按需动态扩展机制。

1.3.1 数据网格作业调度

在数据网格系统上运行的作业可以根据不同的标准进行分类。若从所提交的作业是否在运行过程中与用户进行交互的角度划分,可以将作业分为批作业和互动作业;若以作业在处理过程中是以计算为中心还是以数据为中心划分,又可以简单分为计算密集型作业和数据密集型作业;若从作业处理的时间要求的角度考虑,又可以将作业分为实时作业和非实时作业;当然,还可以根据作业不同的优先级进行分类。本书首先进行以上分类是由于当前的作业调度的研究往往针对不同类型的作业。

作业调度是网格计算的关键的核心技术之一,这是因为在网格计算的环境中,资源是广域松散耦合且异构的,那么对于某一个计算任务,对其进行合理而有效的调度将是这个作业能否得到高效处理的关键。就数据网格上的作业调度而言,现有的作业调度算法主要分为三类:以系统为中心(system-centric)、基于经济(economy-based)和以应用为中心(application-centric)。以系统为中心的调度算法关注整个数据网格系统上所有作业的整体处理性能;基于经济的调度算法强调采用市场经济理论的思想将数据网格中的资源分配和作业处理进行优化;以应用为中心的调度算法则关注单个作业的处理性能达到最优。Condor^[16]、Condor-G^[17]是以系统为中心的调度策略的主要代表;Nimrod/G^[10]实现了数据网格中的生产者和消费者的关系机制,其为基于经济的调度策略的典型代表;AppLeS^[18]则是一个以应用为中心的典型代表。

基于访问代价的调度模型分为两种,若仅针对单个作业的访问代价最小的作业调度策略,则为以应用为中心的调度策略,反之,若针对所有作业的访问代价最小,则为以系统为中心的调度策略。文献[19]~[21]采用了统计方法来估计作业在各个处理节点的预测所需完成时间,而忽视了影响作业完成时间的因素。文献[22]、[23]则认为数据密集型作业所需处理时间与其访问代价有关,将访问代价的

计算作为选择合适作业处理节点的依据。

1.3.2 副本替换策略

对于数据网格中的数据密集型作业,将数据密集型作业调度到具有该作业所需的绝大多数数据文件的节点是一个有效的解决途径。然而,必然存在处理节点上有些数据文件并不存在的现象,此时就需要将其他节点的数据文件复制到合适的处理节点。当前副本复制和替换策略主要有 Cache Replacement 模型、Economic 模型、Value 模型、Popularity 模型、Prediction 和 Cost 模型等。对于副本复制,其中一个重要环节就是要进行副本替换。作业处理节点的物理存储空间并不总是富余,若空闲的存储空间不够,就需要进行副本替换。草率的副本替换往往导致数据文件频繁地替换,进而造成数据网格上的数据文件的颠簸。

(1) Cache Replacement 模型: Cache Replacement 模型是传统的副本替换策略,主要有 Least-Recently-Used (LRU)^[24]、Least-Frequency-Used (LFU)^[25]、Greedy Dual-Size^[26]和 LCB-K^[27, 28]等。LRU 是指存储节点在最近一段时间内最长时间内不被使用的数据文件优先被替换,而 LFU 是指存储节点在最近一段时间内最不经常被使用的文件首先被替换。Greedy Dual-Size 是指每个置入 Cache 中的 Web 页面都赋予一个 H 值,若某个页面被替换则其 H 值减少,若该页面被访问则 H 值增加,那些具有最小 H 值的页面被优先替换^[26]。LCB-K 是指根据计算代价的角度进行选择候选替换数据文件,因为相比于内存替换、网页替换等算法,数据网格中的数据文件的替换的代价更大^[27, 28]。

(2) Economic 模型: Carman 等于 2001 年提出了一个基于 Economic 模型的文件访问和副本复制策略^[29]。Buyya 等^[30]也提出了在 P2P 和 Grid 环境中的基于 Economic 模型的副本复制策略。在文献[29]和[30]中,数据文件被描述为市场中的商品。对于运行中的作业,数据文件可以被计算单元 (computing element) 所购买,也可以被存储单元 (storage element) 来进行投资。对于存储单元,其侧重于如何管理自己的数据文件从而达到其利润的最大化;而对于计算单元,其侧重

于如何优化以使得其购买成本最低化,并提出存储单元主要通过 Auction 协议方式来出售数据文件给计算单元来获得最佳收益。文献[31]对 Economic 模型在 OptorSim^[22]数据网格模拟器中进行评估。

(3) Value 模型:在 Economic 模型的基础上,Yan 等提出了在移动网格环境中的 Value 模型^[32]。在这个模型中,作者考虑了副本的价值,并对替换所产生的价值进行了计算。作者所设定的函数包括 Value income function, Value payout function 和 Value variation function。在网格中的模拟结果表明其具有较好的效果。

(4) Popularity 模型:Philippe Cudre-Mauroux, Karl Aberer 在文献[33]中提出了一个基于 Popularity 模型的副本替换策略。作者将 Popularity 定义为 Zipf 分布函数,通过数学统计的方法对副本的 Popularity 以及其存储空间的 Popularity 进行了对应处理,并在 P-Grid 系统^[34]中进行了验证,取得了较好的效果。

(5) Prediction 和 Cost 模型:Ma 和 Luo 提出了一个基于 Predication 和 Cost 的副本替换算法^[35]。作者认为通过数据网格节点上的数据文件的历史访问累计统计情况进行预测,同时考虑网络延迟、带宽、副本大小以及系统可靠性等因素所导致的副本复制代价,来综合得出副本替换的策略^[35]。这种副本替换算法在平均作业处理时间、网络带宽的使用率、访问延迟等方面在数据网格的模拟中均具有较好的表现。

1.3.3 即插即用

空间数据网格是当前数据网格的一个重要的应用领域。当前,空间数据网格往往与地理信息系统进行耦合构成空间信息网格(spatial information grid, SIG)。在空间数据网格中,其空间数据呈现出海量、多维、多粒度等特点。如何对这类数据进行存储、管理和使用是空间数据网格的主要研究内容。其中,存储设备及其数据资源的即插即用也是一个重要的研究热点。

即插即用机制是指设备在网络系统中的热插拔机制。UPnP 聚焦于家庭中各种设备在无需其他驱动程序的条件下能够自动加入家庭网络,从而实现各种设备的

通用即插即用^[37]。那么这种即插即用机制应用于集群系统中,可以快速实现各种计算、存储设备的自动加入和移出。当前,PVFS^[38]、DCFS^[39]和 LUSTRE^[40]集群文件系统在原有静态存储设备扩展的基础上开始探索存储设备的动态加入,但其动态移除机制尚需要进一步的研究。Chord^[41]实现了广域网上的各种计算节点和存储节点的即插即用。但是,由于空间数据网格的数据具有海量的特点,因此存储设备的即插即用事件主要发生在各个子网中,也就是集群中。为了实现集群文件系统中的存储设备的即插即用,Handy 集群文件系统采用了元数据服务器环的机制实现了存储设备的动态加入和移出^[42]。然而,Handy 集群文件系统忽视了存储设备上数据资源的即插即用,同时忽视了存储设备上数据资源的无缝融合。

1.3.4 按需动态扩展

由于空间数据网格的数据具有海量的特点,运行于其中的数据密集型作业所需的数据量也是巨大的。如何提升这种数据密集型作业的处理性能是一个关键点。相关领域的研究,如内存管理问题,文献[43]系统总结了计算机中的内存管理问题,并认为核心的问题在于解决时间开销和碎片问题。文献[44]则对分布式共享内存系统进行了全面阐述,认为有效地利用各个分布节点上的多个副本是在分布式计算环境中提高作业处理能力的关键。数据网格中的数据文件管理文献^[45,46]则在分析了数据网格中副本管理的基础上,强调副本管理优化的重要性,认为针对数据文件访问历史记录、副本初始创建和副本选择这三个方面进行副本管理优化是提升数据密集型作业的处理性能的关键。然而,在空间数据网格领域,由于数据海量,研究者的分析行为需要得到应急响应,因此进一步要求空间数据网格中的数据管理能够分析行为者的需求,做到快速响应。

1.4 本书的主要研究方法

当前,对数据网格的研究开始进入全面研究的状态。其中关键技术的研究尤