空间数据的网格化存储技术

KONGJIAN SHUJU DE WANGGEHUA CUNCHU JISHU

高劲松 著

湖北科学技术出版社

空间数据的网格化存储技术

KONGJIAN SHUJU DE WANGGEHUA CUNCHU JISHU

ISBN 978-7-5352-3877-1

9 787535 238771 >

定价：30.00元

空间数据的网格化存储技术

KONGJIAN SHUJU DE WANGGEHUA CUNCHU JISHU

高劲松 著

湖北科学技术出版社

图书在版编目(CIP)数据

空间数据的网格化存储技术/高劲松著. —武汉：湖北科学技术出版社，2007.10
ISBN 978-7-5352-3877-1

Ⅰ.空… Ⅱ.高… Ⅲ.计算机网络-信息存贮-应用-地理信息系统：数据库系统 Ⅳ.P208 TP311.13

中国版本图书馆CIP数据核字(2007)第153120号

空间数据的网格化存储技术 © 高劲松 著

责任编辑：高城毅 陈 琦 封面设计：喻杨

出版发行：湖北科学技术出版社 电话：87679468

地 址：武汉市雄楚大街268号
湖北出版文化城B座12-13层 邮编：430070

印 刷：武汉中远印务有限公司 邮编：430034

850毫米×1168毫米 32开 6.75印张 157千字
2007年10月第1版 2007年10月第1次印刷

定价：30.00元

内容简介

本书对网格技术，特别是网格存储及其在空间信息领域的应用前景和发展趋势以及当前的主要研究方向进行了分析，重点研究了网格环境下的海量空间数据的存储和管理技术。

全书分为七章，主要阐述了网格 GIS 的基本概念和技术、空间数据网格化存储模型、基于虚拟 SAN 的空间数据网格存储技术、网格环境下的空间数据副本定位、创建与替换策略、空间数据的网格存储技术应用等。可作为计算机技术、地理信息系统、电子商务、信息管理与信息系统等相关专业技术人员的参考资料，也可供从事空间信息技术研究和对数据存储与管理感兴趣的研究生阅读。

前　言

近年来，随着GIS技术在各行业的广泛应用和对地观测技术的长足进展，空间数据量以几何级数快速增长，数据的应用形式不断多元化，应用范围不断拓宽。空间数据的应用在很大程度上存在着地域相关性，具有广域分散、局部集中的特点。各个应用组织或单位由于技术和经济实力的差异，存储设备和数据格式异构性普遍存在，使GIS正面临着空间数据增长速度过快和现有存储管理技术相对落后所带来的难题。网格计算和网格存储技术的出现对GIS的发展既是机遇也是挑战，网格GIS的兴起和发展为解决海量空间数据的存储管理提供了新的思路。

众所周知，网格以广域资源的全面共享为目标，能够充分吸纳网格中的各种计算资源、存储资源甚至技术和人力等资源，可以为广域范围的资源共享和协作提供理论依据和技术支持。在网格这种复杂的应用环境下，海量空间数据的存储管理如何与网格技术所催生的新的数据共享和应用模式相适应，如何有效提高海量空间数据的存取性能，不仅需要从存储系统结构上加以研究，而且对空间数据的存取模式也提出了新的更高的要求。

本书以空间数据的网格化存储为主要研究目标，以网格存储的基础设施与现有空间数据存储管理模式的结合为切入点，借助于存储虚拟化技术从存储模式的改进和访问性能提升两个角度来解决海量空间数据在广域应用环境下的存储管理问题。

存储虚拟化是被广泛认可的实现网格存储环境整合的重要技术手段，但在具体实现方案上，针对不同的网格应用存在着不同的解决方式。在存储模式问题上，本书立足于空间数据的异构性和海量特征，研究利用网格存储基础设施有效消除 GIS 信息孤岛，实现广域环境下的空间数据资源的透明访问和无缝共享。

通过对空间信息行业存储管理现状的分析，提出了一种基于虚拟 SAN 技术的空间数据网格存储体系结构。利用带外虚拟化技术和存储区域网络技术对海量空间数据的存取环境进行整合，结合网格环境下多用户并发访问机制优化数据通道的设计，使得存储系统能够较好地满足空间数据应用需求。在此基础上，借助于数据网格中普遍采用的副本技术，研究了空间数据网格存储的副本定位和优化技术。数据副本技术是网格环境中广泛应用的旨在提高数据访问和传输性能的主要手段，一般应用于远程海量数据访问上，它能根据用户的访问特征动态地向用户端扩展数据副本，还能有效地改善负载平衡，提高数据可靠性，是解决分布式存储效率的有效途径。

通过对空间数据资源和用户节点分布的分析，本书还提出了利用通信密度较高的多个节点建立空间数据存储域，并以空间数据存储域为基础建立双重副本管理模型的新思路，重点研究了空间数据副本定位、副本创建和副本替换等。在空间数据副本定位上，充分利用各个空间数据存储域之间的对等协同关系，借鉴 P2P 技术解决域间副本定位问题，通过为每个域建立基于 XML 的分布式副本元数据目录实现域内副本定位。在副本创建上，结合了经济模型中的拍卖思想和马尔可夫决策过程来选择最优的副本创建策略。根据空间数据的访问特点，在副本替换上采用 Zipf 定律和淘汰代价函数对副本状态进行动态控

制。与此同时，利用网格仿真平台对副本创建和替换算法的存取性能进行了有效性检验。最后以两个例子进一步探讨了空间数据网格存储的应用及在应用中需要加以解决的若干其他问题。

本书可作为计算机技术、地理信息系统、电子商务、信息管理与信息系统等相关专业技术人员的参考资料，也可供从事空间信息技术研究和对数据存储与管理感兴趣的研究生阅读。

本书的出版得到了湖北科学技术出版社的大力支持，在此深表谢意。

由于水平有限，加之网格技术和空间信息技术发展迅速，新技术、新理论、新观点、新方法不断涌现，书中不足和错误之处恳请读者批评指正。

作　者

2007 年于武昌

目　录

第一章 绪 论

1.1 研究背景及意义

GIS是一门汇集了地球科学、计算机技术、数据库技术、网络技术、智能处理技术、空间信息科学等多门学科和技术的边缘学科，在生态、能源、交通、水利、规划、土地等多个相关行业和领域发挥着日益重要的作用。随着计算机技术、网络技术等相关理论和技术的进步，空间信息应用的广度和深度都在不断扩大和深化，空间信息网络化和有效共享成为GIS发展的重要趋势，也使得GIS的应用迅速普及到人们的日常工作和生活中。但这种应用与普及却受到传统数据管理模式的极大约束，使得用户并不能得到高效的空间信息服务。导致这一现象的主要原因是数据处理和管理手段无法跟上数据量快速增长的步伐。目前在Internet上，全球有超过10亿用户不停地在创建各种数据，并且这些数据每天都在增加。Internet业已成为图像、视频和音频、各种测量数据的巨大存储库。手机、便携式计算机和PDA不断增多，它们通过专用网络、无线网络和Internet等多种途径传输和访问数据，所有这一切都加快了数据的增长速度。

有研究表明，全球数据存储量从1999年的18万TB(1TB=1 024GB)增加到了2003年的200万TB，而且还在以年均80%的速度持续增长。各行业所呈现的数据量增长趋势也是十分明

显的。科学领域的数据采集、存储、处理和传播的数量与日俱增，人类社会所积累的科学数据量已经超过了过去5000年的总和。全世界排名前2000位的巨型、大型企业所存储的数据量每年增长两倍，网络公司存储的数据量3个月内即增长两倍。数据量的大幅度增加促进存储行业迅速发展。据国内权威IT门户网站所做的预测分析报告表明，2010年中国网络存储市场总量规模将是2006年的3倍左右，增长态势如图1-1所示。

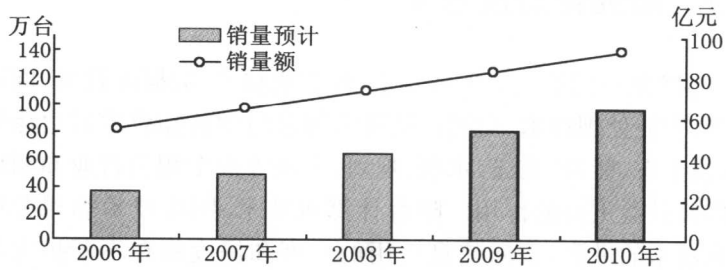

图1-1 2006—2010年中国网络存储市场总量和规模预测

存储市场的快速增长对存储技术本身将有越来越高的要求。随着相关技术的进展，数据存储技术的发展日新月异，先后产生了许多存储技术标准，如SCSI、ATA、SATA、SAS、iSCSI、FC、Ethernet、iFCP、FCIP、SMI-S等。根据存储介质的特点，存储的形式有在线存储、近线存储、离线存储三种，并且在不同的领域得到具体应用。借助于网络技术的进步，存储结构经历了直联存储、附网存储和存储网络三个阶段。当前的存储基本上以网络存储为特征，得到了大量应用，这也使得存储技术逐渐成为各行业的关键支撑技术之一。

有统计表明，全球所存储和使用的数据中，80%左右直接或间接地与空间位置有关，这表明海量空间数据的存储和管理问

题已成为迫切需要解决的问题之一，而传统的存储系统和管理模式却不能满足网络时代的需要。如果说20世纪70年代面临计算速度的挑战(CPU的速度和效率)、80年代面临网络性能挑战(网络通信和数据传输速度)、90年代面临共享的挑战(共享主机和信息)，那么进入新世纪以来，人们更加关注的是数据和信息服务及服务质量的提高。作为信息化建设的三个核心建设内容(计算、传输和存储)之一，数据存储技术在数据服务方面的基础地位是无庸质疑的。

在空间信息领域，从20世纪90年代中期开始，存储就成为了GIS产业中拖后腿的部分，海量空间数据的存储一直是GIS发展的瓶颈。庞大的空间数据如何为用户所获得和有效使用成为新的难题，主要体现在异构的空间数据在网络传输、处理、共享与互操作方面还存在理论和技术上的问题，大量的空间数据被搁置，没有得到合理地使用，造成资源的极大浪费。

从存储系统目前的发展趋势来看，空间数据存储系统的焦点集中在如何以合理的存储管理成本实现智能化存储，它包括四方面内容：第一，了解所存储的数据对象并能对其加以处理；第二，自动实现空间数据的存取，空间数据存储系统应具备主动学习新任务的能力；第三，能与其他智能存储系统协同工作；第四，空间数据存储系统升级性能比现有的存储系统要好得多。对于GIS来说，重要的是空间数据存储系统能进行智能的空间数据处理，也即能够屏蔽不同存储系统之间的异构特性，消除空间信息孤岛和知识孤岛，实现空间数据跨平台的共享和互操作。由于GIS数据中涉及大量的地理空间数据，数据量庞大，类型复杂，各种类型的数据存在着不同的用途，数据的使用频度也各不相同。因此，和一般数据相比，空间数据要求有更高的安全性、更及时的备份和更加灵活的扩充功能(周星，2002)。空间数据

对数据可靠性、存储空间和成本的要求也相对更高一些，这也使得空间数据的存储管理更加困难。

空间数据与人们的日常生活息息相关，它除了数据量庞大外，还具有数据异构性、非结构化等特点，这些都造成了空间数据在存储管理和处理上的复杂性，表现最为明显的是资源共享和互操作能力相对较弱。从网络 GIS 和 WebGIS 的现有影响来看，网络环境下的资源共享和互操作特性为 GIS 的应用提供了舞台，推动了 GIS 的大众化。网络环境为 GIS 达到较理想的应用目标提供了很好的平台。而同样基于网络环境的网格技术，在资源整合和互操作的实现上具有更加强大的优势，它的发展对于目前各种各样的 GIS 必将产生深刻的影响。许多 GIS 学者都认识到网格技术可为空间数据的共享和一体化管理带来巨大契机，网格 GIS 的概念随之产生。事实上，网格 GIS 的研究已经在世界范围内展开。

一般认为，网格 GIS 是一种网格计算环境下的地理信息系统，它可以汇集和共享地理上分布的各种海量空间数据，具有强大的空间数据存储、管理和处理能力，能实现空间数据的一体化组织与管理，是一种能够为用户提供按需服务能力的空间数据基础设施（孟令奎，史文中，2005）。

当前，建设完善的空间数据基础设施以为公众提供便捷的空间数据服务已经成为众多国家政府和学者的共识，许多专家和组织都针对全球海量空间数据的获取、处理、分析和发布展开了研究（Nebert D，2004）。研究的重点多集中在空间元数据服务技术、空间数据资源的调度技术以及空间数据共享技术方面。超大规模空间数据的存储管理技术一直未能有大的突破。网格环境下空间数据服务的目的是为广域范围内的各种用户提供透明的数据共享服务，空间数据存储管理是实现数据共享的技术

基础。但由于空间数据的格式多样、结构复杂、存储模式各异、存储位置分散、数据量增长迅速，导致空间数据存储管理问题高度复杂化。同时，空间数据产业的迅猛发展也使得空间数据的存储管理面临新的挑战。李德仁院士在 2003 年提出的空间信息多级网格框架(SIMG)，使得空间数据具有了更多的内涵，包括网格层次数据和网格内部数据。在空间信息多级网格框架下，如何解决数据存储问题，是实现 SIMG 的关键技术之一。也有人论及在地理信息网格构建中，存储技术是其中的重要支撑技术，包括数据备份技术、网格存储技术以及存储虚拟化技术等多个方面(龚强，2005)。

为了适应空间数据快速膨胀的形势和满足各种用户对空间数据高效服务的迫切要求，采用单纯的集中式存储方式是难以为继的。并且传统的存储管理技术还存在着存储容量有限、扩展不便的缺陷，不利于数据的统一管理和调用，需要采用新的技术手段来解决。在网格环境下解决空间数据的存储问题的基本思路是存储虚拟化和数据副本技术。

存储虚拟化是网格技术对物理存储设备进行整合的一种关键技术，数据副本技术则是网格环境中广泛应用的提高数据访问和传输性能的重要技术手段。存储虚拟化的物理基础设施是存储区域网络(SAN)(谢长生，2003)。在具体实现方案上，随着网络存储技术的不断成熟，存在着许多值得探讨和研究的研究方向，合理地对物理存储设备进行虚拟化统一管理，能够有效地提高数据的存储效率和数据的利用率，也是改善存储系统性能的重要方面。另一方面，数据副本技术可以根据用户的访问特点动态地向用户端扩展数据副本，能够有效地改善网络中的负载平衡，提高数据的可获得性和系统的可靠性及可用性，是解决分布式网络环境下数据存储效率的有效途径(Radic B，2005；邹

鹏，2004）。

数据副本技术主要应用于远程海量数据访问上，根据其固有的动态扩展用户端数据副本特征，可以方便地使用户就近访问数据，减少数据在节点间的移动，从而降低数据存取的代价，提高用户请求的响应速度，还能减少存取延迟，降低带宽消耗。对存储虚拟化的实现和副本技术进行针对空间数据存储管理特性的研究，将为网格 GIS 的数据存储管理奠定良好的技术基础，也有利于推动两者的有机结合，有助于在网格松耦合环境下构建大型 GIS 应用系统，实现海量空间数据的共享和互操作。以此为基础构造一种适合网格环境的空间数据存储管理技术或实用方法不仅能在不增大网络负载的情况下充分扩大空间数据的应用范围，有效地降低空间数据的访问代价，而且还将直接促进数据存储技术和网格技术的进展，因此具有重要的研究意义和实际应用价值。

1.2 研究问题的提出

高性能、高可靠性和高透明度的空间数据存取是实现空间数据共享和互操作的关键，但是由于现有网络带宽和传输速率以及并发用户访问量等的限制，GIS 用户难以及时高效地获得较大地理范围内的空间数据。就当前而言，空间数据存储与管理存在着以下几个主要问题：

（1）两种地理分布所导致的问题

第一种是 GIS 的用户往往位于不同的地理位置，体现为跨城市、跨区域乃至跨国界的地理分布。第二种是空间数据在地理位置上是分布的，为了完成用户的一项任务，可能涉及到多种不同来源和类型的空间数据，而这些数据可能分属于不同的地