

地学领域数据网格的构建与其应用案例分析

宋佳¹, 冯敏¹, 张金区^{1,2}, 尹芳¹

(1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101;

2. 华南师范大学, 广州 510631)

摘要: 网格是 20 世纪 90 年代中期发展起来的新技术, 它把互联网上分散的资源融为有机整体, 实现资源的全面共享和有机协作。本文简单介绍了信息领域网格的起源和发展过程, 并对几种网络计算技术进行了比较分析。进而将网格技术引入到地学领域, 对地学数据网格作了初步的需求分析, 阐述了构建地学数据网格的思路。最后, 以人地系统科学数据网格为例, 分析了人地系统科学数据网格数据层和中间件的实现方法, 并基于该网格通过 DEM 地表呈现模型给出了地学领域数据密集型网格计算的应用, 对改进地学研究方法, 提升应用效率, 推动地学科研信息化具有一定意义。

关键词: 网格; 云计算; 时空数据; 模型计算; 网格服务

DOI: 10.3724/SP.J.1047.2011.00338

1 引言

网格(Grid)的理念始于 1969 年 Kleinrock 表达“像使用电力和电话设施一样来使用计算机”的愿望^[1]。其原型是 1992 年美国国家超级计算应用中心提出的“元计算”^[2-3]概念, 即在网络上构筑虚拟计算机环境、执行大规模并行计算。主导 Globus 网格中间件的 Foster Ian 等将网格描述为:“提供可靠的、一致的、广泛的和廉价的访问高端计算能力的软硬件基础设施”^[4]及“网格计算是在动态变化的多个机构组成的虚拟组织中协调资源共享和求解问题”^[5]。随着计算机技术和网格技术的发展, 出现了许多与网格计算类似的计算形式, 如云计算、P2P 计算、集群计算、分布式计算等。这些计算形式很容易与网格计算混淆, 本文首先对其作简单的比较。

通常, 分布式计算是一个学术概念, 其研究焦点是如何分解单机很难完成的惊人计算量的任务, 即分布式计算面临的首先是数学问题。而网格计算是分布式计算的一种应用形式, 其面临的首要问题是各种计算资源和能力的整合问题。P2P 计算(对等网络计算)是分布式计算的一个子集, 突出的

是用户计算机既是客户机也是服务器。与网格计算相比, 大多 P2P 系统缺少安全机制, 并且连通的可靠性弱, 在资源上缺乏可调度性和平衡性^[6]。而集群是多台计算机服务器的部署方案, 集群中的处理器和资源的数量通常都是静态的, 集群因此被看作是一台私有的计算机。而网络连接的是不同类型及相互并不信任的计算机, 它的运作更像一个公共的计算设施, 网格上的资源可以根据需要添加到网格中, 或从网格中删除。云计算(Cloud Computing)是 2006 年 Google 首先提出的新名词。云计算实质上是网格计算的商业实现。虽然, 云计算在资源构成、应用类型等一些细节上与网格计算还是有区别的, 如表 1。但是, 网格计算与云计算都被看成是分布式计算所衍生出来的概念, 在理念上两者是殊途同归, 都是为了让用户透明访问各种信息资源, 为了更好地提高资源的使用率。

网格初期主要集中在高性能科学计算领域中。现在根据不同的侧重点进一步细分为计算网格、数据网格、信息网格、知识网格等^[7]。自上世纪 90 年代末以来, 国内外的网格研究和建设项目纷纷涌现(表 2, 表 3), 它们大多以计算网格和数据密集型的数据网格为主。

收稿日期: 2011-01-20; **修回日期:** 2011-05-22.

基金项目: 中国科学院“十一五”信息化专项“数据应用环境建设和服务”项目(INFO-115-C01)科学数据库子项目资助。

作者简介: 宋佳(1980-), 男, 博士, 助理研究员, 主要研究方向: 地学数据网格与地学 e-Science, 地理本体。

E-mail: songji@lreis.ac.cn

表 1 网格计算与云计算的比较

Tab. 1 The comparison between grid computing & cloud computing

| | 网格计算 | 云计算 |
|------|--------------------------|---|
| 资源构成 | 不同组织机构中的异构资源构成的资源池 | 同一机构同构资源构成的虚拟资源池 |
| 计算设施 | 高性能计算机为主 | 大量服务器/PC 集群 |
| 计算形式 | 元计算 | 简化的并行计算 (Map/Reduce) |
| 服务理念 | 面向科研问题的信息化服务 (e-Science) | 软件即服务 (SaaS); 平台即服务 (PaaS); 基础设施即服务 (IaaS); 管理服务供应商 (MSP); |
| 应用类型 | 科学计算与虚拟 | 密集型数据处理 |
| 应用领域 | 科学界 | 商业/产业社会 |

表 2 国外主要网格项目^[8-13]

Tab. 2 Some data grid projects abroad^[8-13]

| 名称 | 机构 | 说明 |
|-----------|---------------|---|
| OSG | 美国 NSF 和 DEOS | 构建在美国高校及研究机构间的研究网络,集成分布的资源,为数据深度研究提供计算设施和服务 |
| TeraGrid | 美国 NSF | 涉及 100 多个学科,50PB 的数据资源,应用于气候、生物、地震、高能物理等领域 |
| IPG | 美国 NASA | 高性能计算和数据网格 |
| ESG | 美国阿贡实验室等 | 围绕下一代气候研究,提供气候研究领域的数据及全球和区域模型 |
| GIG | 美国军方 | 致力于计算机、传感器、作战平台的融合及作战信息、知识的共享、决策 |
| DataGrid | 欧盟 | 提供广泛分布的科学团体间的海量数据库共享和深度计算分析 |
| e-Science | 英国 | 提供信息化的科学研究的环境和平台 |
| ITBL | 日本 | 构建日本的服务于科研的虚拟实验室 |

表 3 国内主要网格项目

Tab. 3 Domestic data grid projects

| 名称 | 机构 |
|----------------------|---------------|
| 国家网格 (CNGGrid) | 科技部 |
| 中国教育科研网格 (ChinaGrid) | 教育部 |
| E-Science 网络研究计划 | 国家自然科学基金委 |
| 上海网格 (ShanghaiGrid) | 上海市 |
| 科学数据网格 (SDG) | 中科院计算机网络信息中心等 |

我国地学领域从上世纪末逐步引入网格思想到地学应用中。相关研究主要集中在空间信息网格框架及基于 OGSA 的网格服务技术^[14-20]等方面。

本文主要从地学数据多元异构的特点出发,考虑网格的分布地学数据一站式共享,以及数据网络驱动地学模型的应用研究。

2 地学领域数据网络的构建思路

将网格技术应用到地学领域,有必要说明一下网格技术中的“网格”和地球空间信息多尺度网格中的“网格”之含义。其两者的英文都用“Grid”表达,但意义却不同,前者是信息科学中的一种基于网络和分布式计算的技术,后者应用于地球科学,表达地学空间信息在地球表面的空间划分,实际上后者应该用“格网”表达更准确,以与信息领域中的网格技术相区别。

2.1 地学数据网络的定位

本文所述的地学数据网络定位为地学 e-Science 的基础设施。这里,e-Science^[21-22],即科研信息化,1999 年由英国科学家提出。地学 e-Science^[23-24]是服务于地学研究,支撑地学科研信息化的战略,由地学数据子环境、地学计算子环境、地学可视化子环境三大子环境构成。地学数据网络是“地学数据子环境”的技术核心,如图 1。地学数据网络调取的数据资源,可直接与“地学可视化子环境”对接,以地图、统计图表形式渲染和呈现数据;另外,地学数据网络调取的数据也可作为地学模型计算和过程模拟的基础及初始数据资源,通过与“地学计算环境”对接,实现模型输入数据的自动化

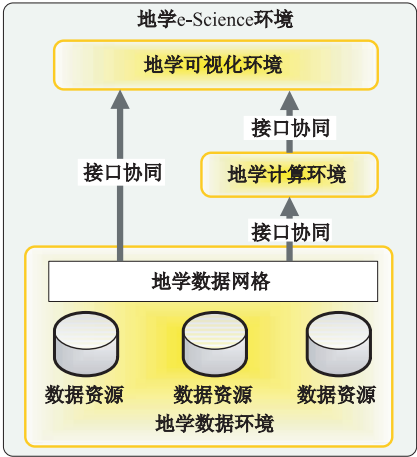


图 1 地学 e-Science 环境下的地学数据网络
Fig. 1 The data grid of geosciences based on e-GeoScience

调取。可以看出,地学数据网格驱动着地学 e-Science 三大环境的协同运转,是地学 e-Science 的基础设施。

2.2 地学数据网格的需求分析及构建思路

2.2.1 地学数据特点分析

地学数据资源具有强烈的时空特征和类型特征,这由地学的研究对象——地球决定的。地学研究中的时空不仅有范围,还有尺度,收集数据资料常常根据时间范围和空间区域筛选。但是,由于科学数据在数据内容层面缺少规范,不同来源的地学数据在时空表达及其存储结构上具有很大的随意性,这对地学数据网格广域共享资源并一站式地调取数据带来了难度。并且,地学数据具有属性、矢量、栅格三大数据类型。数据的“时间、空间、专题属性要素”在这三种类型中以不同的方式体现。在属性数据中,时空及专题属性均以表格字段的方式体现;矢量数据中,空间信息以地理几何坐标(点、线、面)表达,而时间、专题属性要素一般以表格字段表达;栅格数据中,以规则的格网阵列及空间参考体现空间信息,格网值反映某一专题要素的值,时间通过其他说明性的方式体现。

可以看出,“时间、空间、专题属性要素”是构成地学数据的基本三元结构。但在表达“时、空、属性”的数据结构上,则纷繁多样。这一方面是因为地学数据本身就存在属性、矢量、栅格三种不同类型的存储结构,更重要的因素在于缺乏一致的数据表达规范、特别是在时空表达上。为了方便从时空条件及要素属性角度进行一站式的数据调取,有必要研究适合于数据网格的“时间、空间、专题要素”三元组表达规范。

2.2.2 地学数据网格用户分析

地学数据网格宏观上面向两类数据用户:一类是传统数据共享平台中的数据用户,以获得原始的地学数据为目的,这里的“原始数据”并不是“纯”原始的、未经规范整理的数据,实际上是需要对异构多源的数据按照一定的规范整理后的数据,并且这类数据是要严格符合用户的时间、空间、专题要素条件。另一类是模型数据用户,即以运行模型、驱动模型计算为目的的数据用户。对这类用户,需要地学数据网格具备按时空范围、尺度、专题要素属性进行抽取和组装的能力,即需比元数据更细层面提供专题要素粒度的组织和服务模式。

2.2.3 地学数据网格服务模式

对一般数据用户和模型数据用户,应提供两种不同粒度的服务模式:元数据模式和操作粒度更细的数据要素模式(如图 2)。

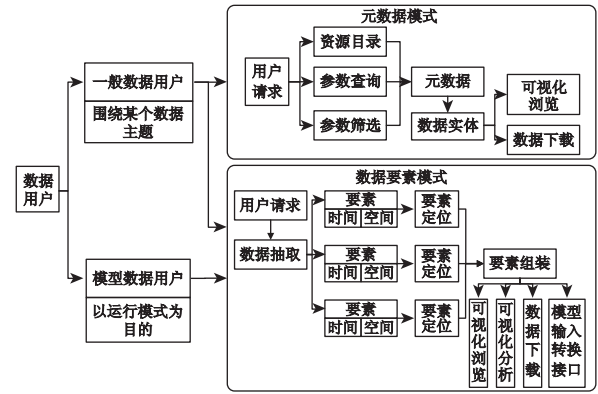


图 2 地学数据网格的用户及服务模式分析
Fig. 2 The analysis of users and service patterns for the data grid of geosciences

元数据模式中的元数据与数据实体是“一对一”对应的,元数据按照资源目录或者资源分类组织,用户可以通过定位资源目录/分类再定位元数据,进而访问到数据实体,也可通过时空和数据类型等参数查询元数据,进而访问到数据实体。采用元数据模式,用户得到的数据是一个完整的、原始的数据实体。

而数据要素模式中,需要从数据内容的层面遍历并定位要素所在的数据实体集,对其中的数据专题要素进行抽取,并按时空范围和尺度进行筛选,来自不同节点并经过抽取筛选后的数据通过组装后提供给用户下载、可视化浏览或驱动模型计算。也就是说,数据网格提供的数据不仅仅是经过了简单的分布式查询访问过程,更重要的是对数据按照用户的时空条件有一个时间范围的截取和空间区域的切割和拼接过程。如果和模型对接后,可通过模型输入接口直接输出到模型计算环境中,驱动模型计算。

3 人地系统科学数据网格的构建与应用分析

基于上述思路,在中国科学院“十一五”信息化专项“数据应用环境建设和服务”项目的支持下,以人口、资源、环境和发展(PRED)为核心初步构建了“人地系统科学数据网格”。该网格瞄准人地关系

研究主题,立足于将孤立的各类数据资源互联起来,通过综合、高效地利用各种计算、存储等资源,达到资源间协同处理和透明访问的目的。“人地系统科学数据网格”的构建及应用包括网格数据层、网格服务中间件和数据网格的模型计算应用三个层面。网格数据层重点研究资源的组织模式,资源存储模型与结构;网格服务中间件以网格服务的设计和技术实现为核心,研究基于 RESTful Web Service 方式的网格服务;在模型计算应用方面,主要研究了高精度 DEM 地表呈现模型在网格环境中的实现和应用。

3.1 人地系统科学数据网格数据层

人地系统科学数据网格中的资源主要包括网格节点资源、数据资源目录、元数据资源和数据实体资源四种类型。整个网格平台由若干自由的网格节点组成,每个网格节点通过一个或多个树型资源目录组织元数据,而元数据与数据实体是一一对应的。每个数据实体都有一对一的元数据对数据的主题、时空范围、类型等信息进行描述,如图 3 所示。

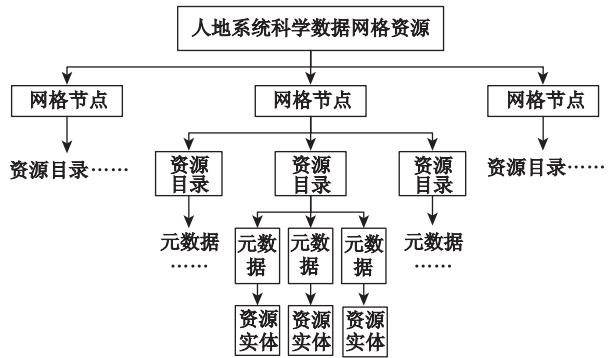


图 3 人地系统科学数据网格资源的层次结构
Fig. 3 The data architecture of the data grid for human - land system

另外,为了实现对人地系统数据按属性要素和时空条件的调取,提出了矢量、属性一体化的数据模型,突出了时间域、空间域、专题要素域的三元结构规范,如图 4 所示。其中,时间域设计了时间点和时间段两种表达模式,时间尺度包括年、月、日、时,并可扩展。空间域包含两个层次:一是用空间位置的名称表示空间信息,比如属性数据中字符串类型表达的各级行政单元、自然区域单元、观测站点等;二是用空间单元的地理坐标值来表示空间位置,即矢量数据的地理坐标部分。专题要素域指的

是地理数据的属性部分,包括属性的具体值和对属性的说明信息(属性元信息),即专题要素字段的标识、名称、单位量纲、描述等信息。

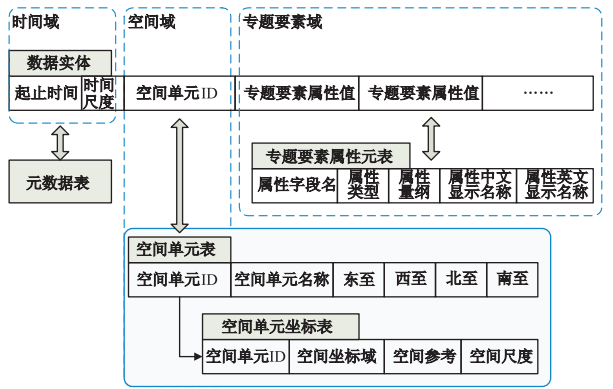


图 4 矢量、属性一体化的时间、空间、专题要素三元组结构
Fig. 4 The model for integrating vector data and attribute data

在人地系统科学数据网格的数据存储中采用了技术先进的开源数据库 PostgreSQL 及其 PostGIS 组件。PostGIS 支持所有 OGC 规范的“Simple Features”类型,同时在此基础上扩展了对 3DZ、3DM、4D 坐标的支持。利用 PostgreSQL + PostGIS,我们已实现了人地系统科学数据网格中全部属性和矢量数据的规范化入库管理。

3.2 人地系统科学数据网格中间件

人地系统科学数据网格中间件包括服务端组件和客户端组件两部分。服务端组件主要起发布和公开本节点数据资源的作用。也就是说,只有部署了服务端组件的节点才可以向其他网格节点提供数据目录、元数据,以及数据实体的查询、调取等服务。服务器端组件基于 RESTful Web Service 技术,以 URL 接口的方式提供网格服务。网格服务的结果可以根据请求的类型是“application/json”还是“application/xml”响应 JSON 和 XML 两种结果类型。对于空间数据部分实现了基于 GML 规范的结果响应。

设计客户端组件的主要作用有两个方面:一是对多个远程节点的数据抽取访问后执行组装过程,使用户可以一站式地获取到符合时空和属性条件的所有网格节点的数据;另一个方面是因为现在流行的 Ajax 调用方式并不能跨域调取其他网格节点的资源,所以,通过客户端组件与服务端组件通信,可以将对其他节点的远程调用转为客户端组件的

取访问服务是保证网格中数据资源广域共享的重要环节。其实现流程如图 7,涉及查询参数设定、执行查询、抽取访问、可视化浏览等过程。数据查询参数的设定包括时空范围和尺度、数据类型、主题词或选择专题要素字段,同时可以控制选择在哪些网格节点和网格节点中的哪些分类目录中进行查询和数据抽取。可用的网格节点和资源目录信息由网格节点监控模块和资源目录访问模块提供。跨节点的数据查询和抽取分别由网格中间件中的元数据查询模块和数据抽取访问模块实现。数据抽取操作根据用户设定的时空范围和属性要素自动执行。这样得到的数据是经过时间过滤和空间范围裁剪后的数据,更符合用户的意图。

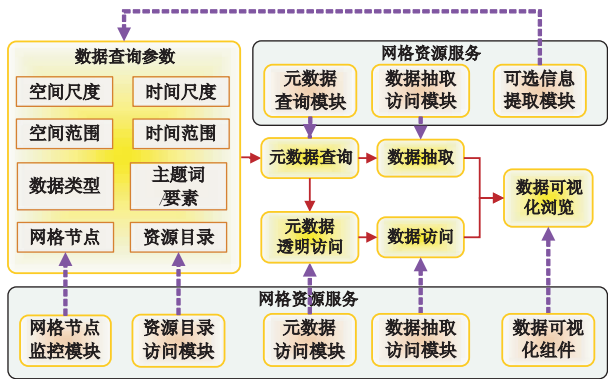


图 7 跨节点的元数据查询与数据抽取访问
Fig. 7 Metadata query on mutil-nodes and data extraction and access

3.3 基于网格的 DEM 的地表呈现模型

在入地系统科学数据网格中,不仅进行了网格中间件的研发,为用户提供了一站式的数据服务,并进行了地学模型的网格计算应用研究,即基于数据网格实现了高精度 DEM 的地表呈现计算模型。模型通过对地形高程数据 DEM 与土地利用数据进行融合计算,输出可视化较好的地理空间数据和地图,突破单独显示地形或土地利用数据带来的表现不足的问题。

在传统的科研环境下实现这种大数据量计算模型,首先,需要收集大量不同类型的相关数据,通过数据格式转换、重投影、数据切割、数据重组拼接等处理,解决数据格式和参考系等互操作问题;其次,基于 ESRI ArcGIS、ENVI 等地理信息系统和遥感软件实现模型,一方面依赖于这些商业软件;另一方面,由于这些软件提供的桌面交互计算环境适

合于处理小数据量,而在数据量较大的应用中往往导致计算效率低下,且难以直接应用并行计算环境。

而通过入地系统科学数据网格,首先,实现了模型输入端土地利用数据的动态调取,即模型通过数据网格的服务接口根据用户选择的地理空间范围处理原始数据,输出指定空间范围的土地利用数据,即基于数据网格实现了数据收集和处理过程的完全自动化。更重要的是,模型的计算过程是基于 Map/Reduce 模式,实现了地理空间数据处理分析过程在网格环境下的高效并行计算。通过构建分布式的地理空间数据存储和模型计算网格,实现了对大数据量地理空间数据的自动分块,以及计算任务的自动分配过程,计算效率得到了充分提升。如图 8 是模型参数输入和计算节点选择界面,包括太阳高度角、方位角、空间范围等模型参数。计算节点可选择“中科院计算机网络信息中心”或“中科院地理资源所”的网格计算节点群。输入参数提交任务作业后,网格开始自动调取数据,并提交数据到选定的计算节点中,然后进入作业监控页面。在作业监控页面中可以随时取消该计算任务。任务完成后即可向提交任务的用户提供运算结果的在线浏览和下载服务,如图 9。



图 8 基于 DEM 的地表呈现模型参数输入界面
Fig. 8 The interface of input parameters for the land surface appearance model based on DEM

4 结语

地学数据网格是地学科研信息化(地学 e-Science)的重要基础设施。本文对地学领域如何构建数据网格进行了探索性的分析和讨论,地学领域的

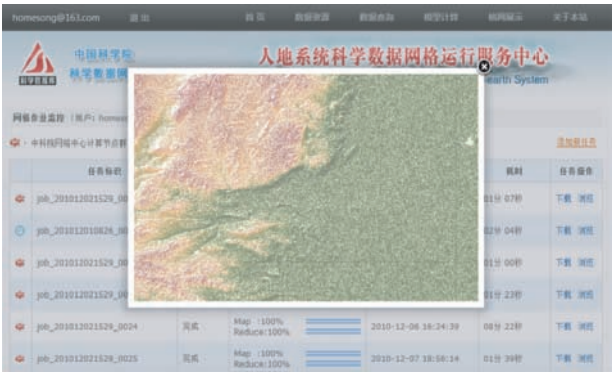


图 9 网格作业监控及模型计算结果

Fig. 9 Monitor of grid jobs and computation results of the model

置无关的一站式广域数据服务,这是地学数据网络的一个基本层次;基于它更高的一个层次是提供地学过程模型的分析计算服务。本文同时也以人地系统科学数据网络为例,阐述了地学数据网络数据层和中间件的设计思路及实现方法,给出了基于网络的模型计算应用案例。目前,人地系统科学数据网络已经在中国科学院地理科学与资源研究所等五家单位部署,在全国范围内初步形成了四个网格资源节点、两个网格计算节点群、一个数据网格门户的分布格局。

但是,人地系统科学数据网络作为地学领域数据网络的研究建设工作尚处于起步阶段,限于时间关系,在诸多方面还需进一步研究,如,在数据模型方面虽然已经提出实现了矢量、属性一体化的数据模型,但尚没有将栅格类型数据纳入到这个一体化体系中来。另外,在网格的安全与权限控制方面也需要今后进一步探索。

参考文献:

[1] Foster I. What is the Grid? A Three Point Checklist [DB/OL]. <http://dlib.cs.odu.edu/WhatIsTheGrid.pdf>, 2002-06.

[2] Smarr L, Catlett C. Metacomputing[J]. Communications of the ACM, 1992, 35(6).

[3] Smarr L, Catlett C. Overview of the I-WAY: Wide Area Visual Supercomputing[J]. International Journal of High Performance Computing Applications, 1996, 10(6).

[4] Foster I, Kesselman C. The Grid: Blueprint for a New Computing Infrastructure[M]. San Francisco, CA: Morgan Kaufmann Publishers, 1998.

[5] Foster I, Kesselman C, Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations[J]. International Journal of Supercomputer Applications, 2001, 15(3).

[6] 吴成义,王志喜. 综合 P2P 和网格计算模式的研究[J]. 西华大学学报·自然科学版, 2006, 25(2).

[7] 徐志伟,李伟. 织女星网络的体系结构研究[J]. 计算机研究与发展, 2002, 39(8).

[8] OSG. <http://www.opensciencegrid.org/About/>.

[9] TeraGrid. <https://www.teragrid.org/web/about/>.

[10] IPG. <http://www.ipg.nasa.gov/>.

[11] ESG. <http://www.earthsystemgrid.org/>.

[12] 苏长云. 网格技术与全球信息网格[J]. 情报指挥控制系统与仿真技术, 2005, 27(3).

[13] 宋琳琳. E—Science 发展情况简介[J]. 图书馆学研究, 2005, (7).

[14] 李德仁,易华蓉,汪志库. 论网格技术及其与空间信息技术的集成[J]. 武汉大学学报(信息科学版), 2005(9): 759 - 760.

[15] 李德仁. 论广义空间信息网格和狭义空间信息网格[J]. 遥感学报, 2005, 9(5): 513 - 520.

[16] 王家耀,孙庆辉,吴明光,等. 面向智能空间信息服务的网格 GIS 节点构建[J]. 武汉大学学报·信息科学版, 2009, 34(1): 1 - 6.

[17] 曾怡,刘定生,李国庆,等. 面向空间信息的数据网格研究[J]. 地理信息世界, 2007(5): 27 - 32.

[18] 王方雄,金宝轩,侯英姿,等. 空间信息网格的数据互操作方法[J]. 武汉理工大学学报(交通科学与工程版), 2008, 32(3): 569 - 572.

[19] 杜鹃,关泽群. 空间信息网格的框架体系和关键技术[J]. 地理空间信息, 2005, 3(2): 27 - 29.

[20] 张建兵. 基于网格的空间信息服务关键技术研究[D]. 中国科学院研究生院, 2006.

[21] 桂文庄. 什么是 e-Science[J]. 科研信息化技术与应用, 2008, (1): 1 - 7.

[22] 桂文庄. 再谈什么是 e-Science[J]. 科研信息化技术与应用, 2009, (2): 1 - 5.

[23] 诸云强,孙九林,宋佳,等. 地学信息化科研环境研究与应用示范[J]. 科研信息化技术与应用, 2009(4): 43 - 51.

[24] 诸云强,孙九林. 面向 e-GeoScience 的地学数据共享研究进展[J]. 地球科学进展, 2006, 21(3): 286 - 290.

The Construction & Application for Data Grid of Geosciences

SONG Jia¹, FENG Min¹, ZHANG Jinqu^{1,2}, YIN Fang¹

(1. *State Key Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China*; 2. *South China Normal University, Guangzhou 510631, China*)

Abstract: Grid is the new technology of IT (Information Technology) field and is developed from the 1990s. It aims to make distributed resources on internet become an organic whole and implement resources sharing and cooperation in an all-round way. The origin and development of grid is simply introduced in the beginning of the paper. And several computing technologies based on the internet are analyzed and compared with grid computing. Meanwhile, the construction of Chinese scientific data grid supported by Chinese Academy of Sciences dates from 2002 is summarized. Then the thought of constructing data grid of geosciences is illustrated in the paper. It is believed that the data grid of geosciences is the infrastructure of e-GeoScience. And the requirements for data grid of geosciences are discussed from several different sides including data characteristics, data users and service patterns in geosciences. Finally, taking scientific data grid for the human-land system as an example, the data layer and the overall framework of the middleware of that are given. And the key problems and its implementation ways in scientific data grid for the human-land system are presented. The scientific data grid for the human-land system has been deployed in four institutes of Chinese Academy Sciences and the distributed resources and unified services are formed. A data intensive application for the scientific data grid, the land surface appearance model based on DEM, is demonstrated in the paper. It is automatic for model when the model is deployed on the scientific data grid for the human-land system. It has positive effects for improving study efficiency of geographic science and promoting the development of e-GeoScience.

Key words: grid; cloud computing; spatio-temporal data; model computing; grid service