

# MySQL 集群与 MPI 的并行空间分析系统设计与实验

周玉科<sup>1,2</sup>, 马 廷<sup>1</sup>, 周成虎<sup>1</sup>, 高锡章<sup>1</sup>, 范俊甫<sup>1,2</sup>

(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. 中国科学院大学, 北京 100049)

**摘要:** GIS 应用正面对空间数据规模日益增加和空间分析算法复杂度逐渐提高的挑战, 本文提出一种基于 MySQL 空间数据库集群与 MPI 的并行计算库分布式空间分析框架的解决方案。该框架使用 MySQL 空间数据库集群解决大量空间数据存储与管理问题, 利用 MySQL Spatial 的 Replication 机制加强空间数据的冗余备份和并发访问控制, 同时使用 MPI 负责分布式计算节点间的通信减少人工控制通信的开发成本。并行框架的任务管理与调度系统采用优先队列式管理, 通过 Master 节点监控集群状态, 合理分发计算任务实现负载均衡和容错。最后, 以多边形 Overlay 算法为例, 研究其在该并行空间分析系统下的并行策略, 采用数据并行的管道流水线作业方式在框架中运行测试, 结果表明, 该并行框架相比串行算法可以得到可靠的加速比。

**关键词:** MySQL 集群; 并行 GIS; 并行空间数据库; MPI; 叠加分析

**DOI:** 10.3724/SP.J.1047.2012.00448

## 1 引言

随着空间数据规模的膨胀和模型的复杂化, 交互式的串行方法解决 GIS 问题容易出现瓶颈。并行计算依靠其强大的计算能力为解决海量空间数据分析提供了一种解决方案。21 世纪初空间信息领域专家提出高性能地学计算的概念<sup>[1]</sup>, 旨在将空间信息科学领域的理论与并行计算技术相结合, 解决 GIS 中海量地理空间数据的并行存储、查询、检索、分析等关键技术问题, 提高 GIS 的空间数据处理与管理能力, 为地学计算密集型和数据密集型的各类空间操作提供强大、高性能的并行处理能力。

集群技术具有高度可伸缩、高可用、易管理和高性价比的特性, 英国爱丁堡大学 Recharad Healey 团队研究并开发了基于分布式集群的并行 GIS 算法库和基于 SIMD(单指令多数据)并行计算体系的空间数据处理分析原型系统<sup>[2]</sup>, 但是由于当时计算机软硬件技术的限制, 该研究只是实验性探索并未大规模应用。国内遥感领域利用网格计算进行高性能地学计算, 代表项目有 ChinaGrid 和空间信息网格 SIG<sup>[3]</sup>, 目的是通过网络汇集空间上分布的海

量空间信息资源, 形成虚拟化组织实现存储和计算资源的共享。高性能空间分析算法的研究仍然以数据分解的并行方式为主<sup>[4]</sup>。本文框架采用数据与计算本地化策略, 可以改善分布式计算松散结构造成的性能下降。

本文给出了 MySQL 和 MPI(message parallel interface)的并行 GIS 框架总体设计框架, 旨在实现小范围局域网集群内的高性能空间计算平台, 提供强大的数据存储和空间分析服务; 同时介绍了框架的数据存储、任务调度等关键技术与解决方案; 并以 Overlay 算法的并行化运算对该框架进行测试。

## 2 系统的架构设计

从层次方面划分, 系统框架可以分为以下 5 层(图 1)。

(1) 数据存储层: 矢量空间数据的主要载体为基于 Linux 文件系统(LinuxFS)的 MySQL Spatial 集群, 计算中的临时结果可以存储在 Linux 的文件系统中, 也可以在 MySQL 中创建临时表存储。

(2) 通信层: 利用底层 socket 管道进行快速文

收稿日期: 2012-03-17; 修回日期: 2012-07-26.

基金项目: 国家科技支撑计划(2011BAH06B03、2011BAH24B10); 国家自然科学基金项目(40830529、41171307)。

作者简介: 周玉科(1984-), 男, 博士研究生, 从事高性能空间分析。E-mail: zyk@lreis.ac.cn

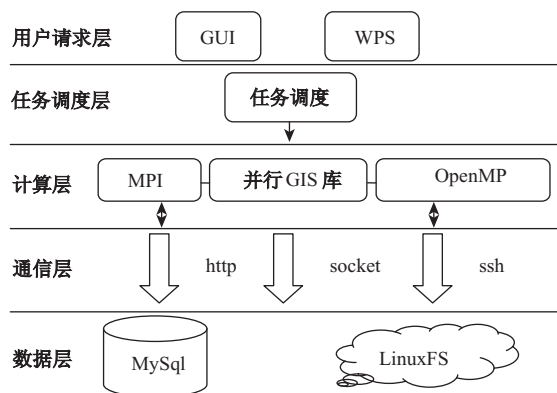


图 1 并行空间分析架构

Fig. 1 Architecture of parallel spatial analysis

件传输,http 协议支持广域网访问,ssh 进行集群间的远程控制和任务启动。

(3)计算层:主要由并行 GIS 空间分析模块组成。各个子节点均部署一份并行 GIS 算法包。该部分是工具集中的核心,从底层进行研发,充分利用 MPI 多进程技术提升原始算法性能。模块内容包括:空间几何分析(buffer, overlay),空间关系分析(nearest, disjoint, touch)、空间插值分析等,本文将选取典型 Overlay 算法进行并行化封装,实现高性能目标。

(4)任务调度层:是整个并行框架的核心部分,负责计算节点的负载均衡。并行空间分析算法与传统的串行算法最大的区别,就是必须考虑各个子节点运行的状态和任务、数据在这些节点中的分配和工作流的整合。用户请求服务必须通过任务调度器进行排队,等待群资源状态达到基本运行条件才启动该服务。

(5)用户请求层:该层的具体表现为一个跨平台的客户端应用程序,可以运行在 Windows 和 Linux 下,更高层次的应用可以将客户端封装为基于 Web 的 WPS(Web Process Service)。

### 3 系统关键技术与实现

与数值并行计算架构类似,并行空间分析系统中的关键技术也会涉及到数据存储、数据分发、任务调度与协同等多种技术,不同的是矢量空间数据具有多尺度和分布不规则等空间特性,不能简单地按照数值并行计算方法进行组织,其系统实现的关键技术包括:

#### 3.1 并行空间数据的管理

除具有一般并行计算框架的存储管理、资源监控和发现等功能外,并行空间分析系统的资源管理更侧重空间数据库、GIS 算法库和模型库的管理<sup>[5]</sup>。空间分析数据通常具有数据量大、数据分布广等特点,在并行计算体系下如何组织管理这些空间数据,直接关系到并行 GIS 算法效率的提升。空间数据划分问题已被证明为 NP(非确定多项式)问题<sup>[6]</sup>,因此须采取合理方案组织管理数据。本架构中对待处理数据采取冗余存储机制,每个子节点均保存一份,减少算法输入数据的传输。这种空间数据组织方法可以实现本地负载均衡,每个计算子节点上的数据负载尽可能的接近,同时保证了计算的本地化,计算节点只需要读取本地数据库。

为提高分布式集群的数据并发访问效率和容错能力,数据组织形式采用 MySQL 集群的复制机制(Replication)<sup>[7]</sup>。MySQL 数据库存储引擎有 MyISAM 和 InnoDB 两种<sup>[8]</sup>,该并行空间数据库采用的 MyISAM 引擎可支持空间数据结构和 R-tree 索引。矢量数据导入到各节点数据库中并建立空间索引,加快数据抽取速度。在并行空间分析操作中,数据读操作远大于写操作,因此,瓶颈在于如何快速抽取数据。MySQL Replication 可以实现 READ/WRITE 操作的分离,这个功能在大规模读写操作中会非常实用,而且通过同步复制策略可以提升集群扩展时的性能和负载均衡(图 2)。

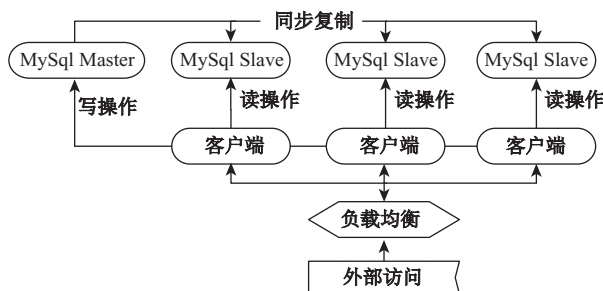


图 2 读写分离的并行 MySQL 空间数据库

Fig. 2 Parallel MySQL spatial database with standalone Read/Write node

MySQL 空间数据库的复制策略用到一个 Master 服务器和多个 Slave 服务器,Master 是需要复制与同步的操作和数据的源,而 Slave 则是这些

源材料的接受者。MySQL 集群本身依赖于 NDB 存储引擎, Master 利用二进制日志(Binlog)系统设置需要备份的空间数据库, Slave 节点通过 I/O 线程→Binlog→Sql 线程的顺序将 Master 节点 Binlog 指定的源进行复制,其中 Binlog 过程存在不同程度的延迟,可用于读压力比较大的应用的数据库端廉价扩展。面对海量空间数据库,可以指定备份特定的数据表,而无需备份整个数据库。集群性能的加速性能通过各节点协同计算获得,单一节点承担某一计算片段,这些片段计算性能会高于集群整体评价因此可以获得加速比。

### 3.2 任务监控与调度

并行 GIS 集群系统中任务调度通常由资源管理器和任务调度器组成,负责计算任务的提交、排队分发和状态监控。资源管理器负责实时获取每个子节点心跳信息(Heartbeat),解析该节点的健康状态(内存容量、CPU 占有率、硬盘资源等),为任务队列的优先级排序提供参考信息,以合理分配资源保证集群运行效率。并行 GIS 空间分析系统使用 xpbsMon 可视化工具监控集群状态,它不仅可通过 ssh 隧道获取每个阶段的心跳信息来监控节点的存活,如图 3,绿色代表可用节点,红色代表宕机节点,还可以通过集群高级命令查看节点情况的进程信息(进程 ID,用户,进程命令等)。

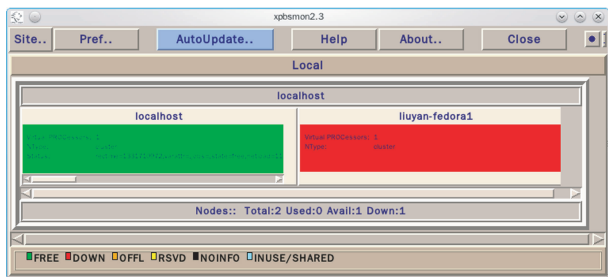


图 3 集群状态监控

Fig. 3 Monitor of cluster status

本系统的任务装载管理系统由 Torque 和 Maui 组成。Torque 作为一个 pbs\_server 负责接收用户作业提交信息和收集任务队列的状态<sup>[9]</sup>。Torque 默认包含一个队列维护控件,但是功能比较薄弱,Maui 可以很好地支持队列的优先级维护,用户先提交的作业位于队首,已经完成的作业将从 Maui 队列中弹出。并行 GIS 中的空间分析算法可以分为批处理作业和交互式作业两种<sup>[10]</sup>。以并行

Overlay 为例,其中的求交操作(intersect)只需 2 个图层的一次叠加便可获得结果图层,可以批处理的方式进行任务分发,而 Overlay 的联合操作(union)需要在局部合并以后与其他节点结果进行进一步的合并。批处理作业调度只需要将任务提交到队列,没有上下文依赖关系,便于任务队列的维护,而交互式作业需要关注作业前后的依赖关系,往往需要多个队列维护一个操作。图 4 展示了 Torque 和 Maui 作业排队情况,还可以报告 JobID 和可用 CPU 等情况,并能够手动改变作业的优先级,让依赖少的任务优先执行。

Job id	Name	User	PEs	CpuUse	WaitUse	S	Queue
11.localhost	pbs	kekezhou	1	0	0	Q	batch@kekezhou.fedora kekezhou.fedora 11.localhost
13.localhost	showxclock	kekezhou	1	0	0	Q	batch@kekezhou.fedora kekezhou.fedora 13.localhost
15.localhost	tor	kekezhou	1	0	0	Q	batch@kekezhou.fedora kekezhou.fedora 15.localhost
16.localhost	showxclock	kekezhou	1	0	0	Q	batch@kekezhou.fedora kekezhou.fedora 16.localhost
17.localhost	pbs	kekezhou	1	0	0	Q	batch@kekezhou.fedora kekezhou.fedora 17.localhost
18.localhost	STDIN	kekezhou	1	0	0	Q	batch@kekezhou.fedora kekezhou.fedora 18.localhost
19.localhost	map_overlay	kekezhou	1	0	0	Q	batch@kekezhou.fedora kekezhou.fedora 19.localhost

图 4 并行任务队列

Fig. 4 Parallel task queue

### 3.3 节点通信

并行 GIS 空间分析系统节点间的通信建立在 MPI 基础之上,有 4 种通信模式:标准通信模式、缓冲通信模式、同步通信模式和就绪通信模式<sup>[11]</sup>。MPI 消息传递的最大通信量不超过 2M,并且数据类型只有浮点型、整型和字符型,适合传递算法中函数的简单输入参数或小型矩阵数组。本系统中 MPI 主要负责启动各子节点的 GIS 算法包。利用 MPI 的扩展数据类型,可以用来传递 GIS 简单要素对象中的点对象。对于集群中较大容量的数据传输则直接利用 socket 套接字进行通讯,避免 MPI 传递带来的信道阻塞。

## 4 并行 Overlay 算法策略与实验分析

地图叠加分析(Overlay)是 GIS 空间分析中最基础使用最频繁的一种操作,指同一地域范围的两个或多个地图图层在同样空间参考系下进行叠加,获取具有新属性的空间区域,其结果是生成一个新的图层,图层要素属性由叠加运算符决定。叠加分析是一系列计算几何布尔操作和属性传递过程的集合,其中几何对象之间的操作已经被证明为时间复杂度最少为  $O(n \log n)$  的操作<sup>[12]</sup>,因此考虑地图

全幅的要素叠加分析属于计算密集型操作。对此,首先从方法论探索 Overlay 的可并行特征,然后,结合本并行框架进行可行性和加速比测试实验。

#### 4.1 并行策略

空间叠加分析实质是图层间的 boolean 运算,因此,其须从算法本身的并行特征出发,才能使并行算法的效率更理想。与传统的并行计算模式一样,并行 Overlay 算法也可以利用功能划分(Function—Decomposition)和数据分解(Data—Decomposition)来实现。功能划分是从任务级别对并行算法进行切分,对于 Overlay 算法来讲,由于其具有代数运算的性质,按功能并行必须与分布式计算系统和具体数据结构紧密耦合,导致算法通用性和可移植性较差。因此,Overlay 算法比较适合以数据并行的方式进行并行化。

本文并行 Overlay 以求交操作为例,采用数据并行策略并以管道方式进行叠加分析。A 地图中每一个  $N_a$  多边形都将配送到  $p$  个节点中的某一个,  $1 < p < N_a$ , 而 B 地图中的多边形将依次对其进行空间叠加运算。运算结果将同时产生并存储至结果收集器中,或进入下次的管道线叠加操作中。管道式并行 Overlay 过程如图 5,其中第一步相交操作无结果,因此直接进入下一次求交。

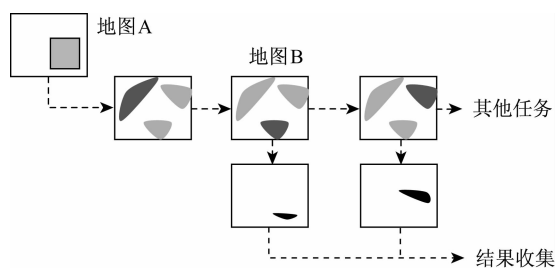


图 5 管道线式 Overlay 方法

Fig. 5 Pipeline styled Overlay algorithm

在并行 GIS 框架中,并不是发送地图 A 中多边形的真实数据到计算模块中,只需要 Master 任务控制器发送需要叠加的地图 A 中的多边形索引号,计算模块会在就近(几乎都是本地)的 MySQL 空间数据库中抽取该行数据。为提高效率,使用按图层 FID 等步长划分的方式,每次将一个步长内的多边形索引发送到一个计算节点,剩余多边形按此规则依次轮询发送,该方法基本能够保持计算任务的负载均衡。

#### 4.2 算法的实现与应用

实验中算法的实现过程如下:

(1)启动计算环境。通过系统数据管理工具将 Shapefile 格式矢量数据导入到 MySQL Master 节点,其 Replication 机制会自动将数据同步到各 Slave 节点<sup>[13]</sup>;开启管理节点任务监控和调度服务;将 Overlay 算法包分发到各子节点。

(2)用户向管理节点提交作业。任务调度将该作业压入队列并分配唯一的 JobID。调度器根据集群的状态和队列优先顺序启动作业。

(3)Overlay 运算启动。先根据分配的 FID 去元数据表中寻找待处理图层,抽取数据后各自计算,求交结果存储到本地临时表中。

(4)Master 节点收集各节点的临时结果形成最后结果,任务结束,将该 Job 弹出队列。

#### 4.3 实验结果与分析

为测试并行空间分析框架中 MySQL 空间集群和 MPI 通信的计算能力,利用多边形图层的求交分析进行实验。

实验环境:操作系统为 Fedora15\_x86\_64 (Linux2.6.38),编程环境为 gcc 编译器和标准 c/c++ 开发语言,网络环境为 100M 局域网,并行库采用 MPICH2-1.3.2,MySQL 空间数据库为 5.5 集群版,PC 集群硬件设施为 Dell (Optiplex 980),1T 硬盘,4 核 CPU。程序运行过程中,MPI 根据配置脚本文件为多核 CPU 分配计算任务而无需手动干预,缺省情况下任务会平均分配到各处理器核。

实验数据:底层矢量数据为程序生成矩形有洞多边形(Polygon),要素个数分 6 万、20 万、100 万多个级别,叠加图层为世界行政区图层 224 个多边形要素。操作过程为用世界行政区图层对矩形多边形图层进行求交,部分图层进行 intersect 运算,效果如图 6。并行效率统计情况如表 1 所示。

表 1 并行 Overlay 操作性能

节点数	图层 1	图层 2	串行时间(S)	并行时间(S)	加速比
4	6 万	251	508	61	1.08
4	20 万	251	732	157	1.15
4	100 万	251	>5 小时	<7500	>1.20

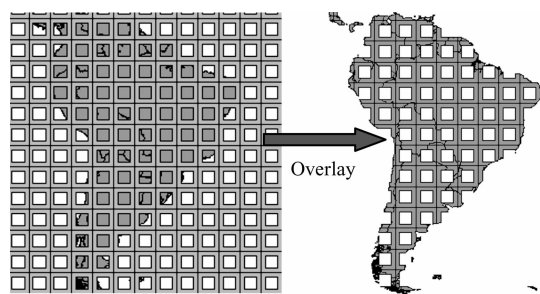


图6 并行叠加分析效果

Fig. 6 Result of parallel overlay

从实验统计结果可以看出,并行 Overlay 运行时间肯定小于串行方法。随着数据量的增大,节点间通信开销和数据频繁读写增加,导致加速比有所下降,但是总体效率仍有优势。在数据量特别大的情况下(百万级别要素),串行方法甚至无法完成运算。总体说明并行空间分析框架能够加速复杂 GIS 算法的效率,系统稳定性和扩展性需要进一步加强。

## 5 结语

空间信息服务和地学问题求解往往要求即时响应,而空间数据的海量化膨胀增加了信息分析和挖掘的难度,高性能并行计算成为快速高效解决问题的一种经典方案。传统高性能地学计算的应用通常以遥感影像处理分析为主,本文提出的并行空间分析框架采用 MySQL 与 MPI 结合的方式进行矢量数据并行化存储与处理分析的方法。使用 MySQL 空间数据库集群弥补矢量空间数据文件格式存储的不足,实现高效的数据同步和冗余备份。采用 MPI 消息传递机制负责算法参数的传送,GIS 开发人员无需过多关心节点通信而专注于并行 GIS 算法设计。任务调度是并行计算系统的核心,本框架采用的优先队列任务调度方可以选择性能最佳

的节点进行任务分发。在该框架的基础上,利用数据并行的策略实现了并行 Overlay 运算,实验证明该并行算法可以提高叠加分析的效率,但是在多个算法步骤形成工作流时的复杂情况需要进一步优化任务调度策略,以致数据划分方法有所创新。

### 参考文献:

- [1] 赵春宇. 高性能并行 GIS 中矢量空间数据存取与处理关键技术研究[D]. 武汉:武汉大学,2006.
- [2] Mineter M J, Dowers S and Gittings B M. Towards a HPC framework for integrated processing of geographical data: Encapsulating the complexity of parallel algorithms[J]. Transactions in GIS, 2000(4): 245 - 261.
- [3] 薛勇,万伟,艾建文. 高性能地学计算进展[J]. 世界科技研究与发展,2008(3):314 - 319.
- [4] 王结臣,王豹,胡玮,等. 并行空间分析算法研究进展及评述[J]. 地理与地理信息科学,2011(6):1 - 5.
- [5] 罗英伟,汪小林. 空间信息合作与并行处理[J]. 计算机辅助设计与图形学学报,2003,15(10):1307 - 1314.
- [6] 方裕,邬伦,谢昆青,等. 分布式协同计算的 GIS 技术研究[J]. 地理与地理信息科学,2006,22(3):9 - 12,54.
- [7] MySQL Replication. <http://dev.mysql.com/doc/refman/5.5/en/replication.html>
- [8] 朱江,张立立. 海量影像数据的发布集群系统与应用[J]. 地球信息科学,2006,8(2):101 - 105.
- [9] Torque. <http://www.clusterresources.com/torque-docs21>.
- [10] 吴亮,谢忠,陈占龙,等. 分布式空间分析运算关键技术[J]. 地球科学(中国地质大学学报),2010(3).
- [11] 陈国良,孙广中,徐云,等. 并行算法研究方法学[J]. 计算机学报,2008,12(9):1493 - 1502.
- [12] Bentley J L, Ottmann T A. Algorithms for reporting and counting geometric intersections[J]. IEEE Trans. Comput., 1979, C-28:643 - 647.
- [13] 王璟,张云泉,李玉成. 基于 MPI 和 MySQL 的并行数据库系统搭建[J]. 计算机科学,2003,31(10):418 - 421.

## Design and Implement of Parallel Spatial Analysis System Based on MySQL & MPI

ZHOU Yuke<sup>1,2</sup>, MA Ting<sup>1</sup>, ZHOU Chenghu<sup>1</sup>, GAO Xizhang<sup>1</sup> and FAN Junfu<sup>1,2</sup>

- (1. *State Key Laboratory of Resources and Environmental Information System (LREIS), Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;*  
2. *University of Chinese Academy of Sciences, Beijing 100049, China*)

**Abstract:** With the rapid development of space survey technology, GIS is facing a challenge of fast growing size on spatial data and complexity of spatial analysis algorithm. Traditional serial spatial analysis method isn't able to deal with this condition well. High performance computer and new computing methods provide an innovative way for spatial data processing and analysing problem. Remote sensing data processing is data-intensive and an ideal domain to use parallel computing, but vector data operation is computing-intensive which needs more computing ability. In this paper, a distributed spatial analysis framework based on MySQL spatial and MPI is described. Parallel spatial vector data mean is explored in kind of cluster way. This framework uses MySQL spatial cluster to store and manage GIS data which can resolve the problem about fault-tolerant and concurrent access for the same data block. MPI is good at passing messages in distributed network nodes, so it's not necessary to control telecom between nodes manually. Task management and distribution use prior queue to achieve load balance and fault-tolerant through monitoring the status of cluster. Finally, a parallel polygon overlay operation is experimented on this distributed system to test the performance of the cluster. The strategy of parallel Overlay operation is in a pipeline way, each node gets a part set of the polygons in the overlaid layers. And this method got relative better speedup than the serial overlay operation.

**Key words:** MySQL cluster; parallel GIS; distributed spatial database; MPI; Overlay