Feasibility Study of Geo-spatial Analysis Using Grid Computing

Yincui Hu[1], Yong Xue[1,2*], Jianqin Wang[1], Xiaosong Sun[1], Guoyin Cai[1], Jiakui Tang[1], Ying Luo[1], Shaobo Zhong[1], Yanguang Wang[1], and Aijun Zhang[1]

[1]Laboratory of Remote Sensing Information Sciences, Institute of Remote Sensing Applications, Chinese Academy of Sciences, P. Box 9718, Beijing 100101, China
[2]Department of Computing, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK
{huyincui@163.com, y.xue@londonmet.ac.uk}

Abstract. Spatial applications will gain high complexity as the volume of spatial data increases rapidly. A suitable data processing and computing infrastructure for spatial applications needs to be established. Over the past decade, grid has become a powerful computing environment for data intensive and computing intensive applications. In this paper, we tested and analyzed the feasibility of using Grid platform for spatial analysis functionalities in Geographic Information System (GIS). We found that spatial interpolation, buffers, and spatial query can be easily migrated to Grid platform. Polygon overlay and transformation could achieve better results on Grid platform. To do network analysis and spatial statistical analysis on Grid platform could be no significant improvement of performance. The most un-suitable spatial analysis on Grid platform is the spatial measurement.

1 Introduction

In numerous scientific disciplines, terabyte and petabyte-scale data collections are emerging as critical community resources. A new class of "data grid" infrastructure is required to support management, transport, distributed access to, and analysis of these datasets by potentially thousands of users. Researchers who face this challenge include the climate modeling community, which performs long-duration computations accompanied by frequent output of very large files that must be further analyzed. The number of applications that require parallel and high-performance computing techniques has diminished in recent years due to the continuing increase in power of PC, workstation and mono-processor systems. However, Geographic information systems (GIS) still provide a resource-hungry application domain that can make good use of parallel techniques. GIS applications are often run operationally as part of decision support systems with both a human interactive component as well as large scale batch or server-based components. Parallel computing technology embedded in a distributed system therefore provides an ideal and practical solution for multi-site organisations and especially government agencies who need to extract the best value from bulk geographic data.

* Corresponding author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 956–963, 2004.

Spatial applications will gain high complexity as the volume of spatial data increases rapidly. A suitable data processing and computing infrastructure for spatial applications needs to be established. Over the past decade, grid has become a powerful computing environment for data intensive and computing intensive applications.

Parallel and Distributed Knowledge Discovery (PDKD) is emerging as a possible killer application for clusters and grids of computers. The need to process large volumes of data and the availability of parallel data mining algorithms, makes it possible to exploit the increasing computational power of clusters at low costs. On the other side, grid computing is an emerging "standard" to develop and deploy distributed, high performance applications over geographic networks, in different domains, and in particular for data intensive applications. Cannataro (2000) proposed an approach to integrate cluster of computers within a grid infrastructure to use them, enriched by specific data mining services, as the deployment platform for high performance distributed data mining and knowledge discovery.

Integrating grid computing with spatial data processing technology, Pouchard *et al.* (2003) described the increasing role of ontologies in the, context of Grid Computing for obtaining, comparing and analyzing data. They presented ontology entities and a declarative model that provide the outline for an ontology of scientific information. Relationships between concepts were also given. The implementation of some concepts described in this ontology was discussed within the context of the Earth System Grid II (ESG).

In this paper, we tested and analyzed the feasibility of using Grid platform for spatial analysis functionalities in Geographic Information System (GIS). First, we listed the several basic spatial analysis functions used in GIS systems. Following the definitions of criteria and basic principles for spatial analysis middleware development for Grid platform, we analyzed the feasibilities of above basic spatial analysis functions for Grid platform and give the suggestions on how to develop the middleware for spatial analysis with Grid platform.

2 Spatial Analysis Functionalities Commonly Used in GIS

Spatial Analysis is a set of techniques whose results are dependent on the locations of the objects being analyzed and requiring access both to the locations of objects and also to their attributes (Goodchild 2001). GIS is designed to support a range of different kinds of analysis of geographic information: techniques to examine and explore data from a geographic perspective, to develop and test models, and to present data in ways that lead to greater insight and understanding. All of these techniques fall under the general umbrella of "spatial analysis". In general, it includes Query, Analyses which are simple in nature but difficult to execute manually, such as overlay (topological), map measurement, particularly area, and buffer zone generation, Browsing/plotting independently of map boundaries and with zoom/scale-change such as seamless database, need for automatic generalization and editing, and Complex modeling/analysis (based on the above and extensions). We will focus on the following spatial analysis functionalities: Query and reasoning: the identification

of objects and attributes either by their location or attribute query; Measurement: simple geometric measurements associated with objects; Transformation; Buffers; Spatial overlay; Spatial Interpolation; Network Analysis and Statistical Analysis.

3 Evaluation Criteria and Basic Principles

Two issues will be considered in use of Grid technology. One is the efficiency or high performance, i.e. to get the solution in a very short time period or to solve more complex problem in a same time period. Another is the high throughput computing to reduce the cost, etc. We will only evaluate the performance of spatial analysis algorithms on Grid platform and neglect the limitations of hardware and data storage.

To improve the efficiency of application algorithms is to enhance the parallel processing. Normally, there are two ways to deal with it. One is to parallel the processing algorithms and second is to parallel the data. We use 5 levels of criteria factors: I – worse, II – poor, III – good, IV – better, V – Best. For the parallel of algorithms, how many sub-jobs can a job be divided into? Table 1 shows the explanation of all five levels. For the parallel of data, how many sub-areas can the whole area be divided into? Table 2 explains the meaning of al five levels.

Table 1. The meanings of all five criteria levels of algorithm parallel.

Criteria Level	I	II	III	IV	V
Digital Score (A)	1	3	5	7	9
Description	There is few parallel.	Can be paralleled, but half of sub-jobs are inter-associated.	Can be paralleled, but between 1/2 and 1/3 of sub-jobs are inter-associated	Can be paralleled, less than 1/3 of sub-jobs are inter-associated	Can be paralleled, no sub-jobs are inter-associated

Table 2. The explanation of all five criteria levels for data parallel.

Criteria Level	I	II	III	IV	V
Digital Score (D)	1	3	5	7	9
Description	Data cannot be divided for processing.	Data can be divided with more than half of redundancy.	Data can be divided with half of redundancy.	Data can be divided with less than half of redundancy.	Data can be divided without redundancy.

It is easier to realize the data parallel processing. Because of the higher efficiency for data parallel processing than that for algorithm parallel processing, we define the different weight factors for data and algorithm parallel processing. The overall evaluation criteria score (E) will be

$$E = 0.6 * D + 0.4 * A \quad (1)$$

Where A is the score from Table 1 for algorithm parallel and D is the score from Table 2 for data parallel. The final overall evaluation criteria will be in five levels as shown in Table 3.

Table 3. The overall criteria levels for both data and algorithm parallel

Criteria Level	I	II	III	IV	V
Score (E)	$1 < E \leq 1.8$	$1.8 < E \leq 3.6$	$3.6 < E \leq 5.4$	$5.4 < E \leq 7.2$	$7.2 < E \leq 9$
Description	Worse	Poor	Good	Better	Best

4 Performance of Spatial Analysis on Grid Platform

4.1 Spatial Query

Database query is probably one of the most important and most commonly used application in Geographic Information Systems. Like any database, a GIS allows you to access information held in a data file in a variety of ways. Information can be grouped in categories, sorted, analyzed, printed, etc. The difference, once again, is that GIS deals with *spatially oriented data*. This means that when querying a database you cannot only see the data but its geographic location as well.

Database query simply asks to see already stored information. Basically there are two types of query most general GIS allow: viz., Query by attribute and Query by geometry. Map features can be retrieved on the basis of attributes. The attribute database, in general, is stored in a table (relational database mode.) with a unique code linked to the geometric data. This database can be searched with specific characteristics. However, more complex queries can be made with the help of SQL. GIS can carry out a number of geometric queries. The simplest application, for example, is to show the attributes of displayed objects by identifying them with a graphical cursor. There are five forms of primitive geometric query: viz., Query by point, Query by rectangle, Query by circle, Query by line, and Query by polygon. A more complex query still is one that uses both geometric and attributes search criteria together. Many GIS force the separation of the two different types of query. However, some GIS, using databases to store both geometric and attribute data, allow true hybrid spatial queries.

For spatial query, the database can be divided in to several smaller databases and searched in parallel. The score for criteria level of algorithm parallel is 1 and data parallel is 9.

4.2 Spatial Measurements

GIS makes spatial measurements easy to perform. Spatial measurements can be the distance between two points, the area of a polygon or the length of a line or boundary.

Calculations can be of a simple nature, such as measuring areas on one map, or more complex, such as measuring overlapping areas on two or more maps. Distance measurement refers to measuring straight-line distances between points or between points and their nearest points or lines.

As spatial measurements are always simple and the delay of data communication is much longer that processing, there is no need to use Grid computing technology. The overall criteria score is 1.

4.3 Transformation

The point transformation includes the algebra functions such as addition, subtraction, multiply and division and logical functions such as AND, OR, NOT, NOR, etc; Comparison functions such as GREAT, LESS etc; and other mathematical functions. Those are simple operations and the score for algorithm parallel is 1 and for data parallel is 9.

DEM analysis is one of the most popular transformation operations. Roros and Armstrong (1998) showed that three steps are needed: transect classification, cell classification, and feature topology construction. The score for algorithm parallel is 1, but the score for data parallel is 3 – 7.

4.4 Buffer Analysis

Buffer analysis is used for identifying areas surrounding geographic features. The process involves generating a buffer around existing geographic features and then identifying or selecting features based on whether they fall inside or outside the boundary of the buffer. This kind of analysis is also called proximity analysis. The buffer operation will generate polygon feature types irrespective of geographic features and delineates spatial proximity. The data parallel is scored 9 and algorithm parallel is scored 1.

4.5 Spatial Overlay

One basic way to create or identify spatial relationships is through the process of spatial overlay. Spatial overlay is accomplished by joining and viewing together separate data sets that share all or part of the same area. The result of this combination is a new data set that identifies the spatial relationships. This allows the user to view and analyze those portions of the various layers which cover the same place on the Earth.

Spatial overlay could be done in parallel. The data parallel has a score of 9 of criteria level and the algorithm parallel could be from 1 to 9 depending on the algebra of map overlay.

4.6 Network Analysis

Network analysis is used for identifying the most efficient routes or paths for allocation of services. This involves finding the shortest or least-cost manner in which to visit a location or a set of locations in a network. The "cost" in a network analysis is frequently distance or travel time. Network analysis can also be used to optimize the allocation of resources.

For the case of shortest distance analysis, the algorithm has to be serial and the criteria level's score is 1. Data could be divided into several parts. But the size of each part could be too small as the coordinating between each part takes much long time (Lanthier and Nussbaurm 2003). The data parallel scores 7.

4.7 Spatial Interpolation

A GIS can be used to estimate the characteristics of terrain or ecological conditions from a limited number of field measurements. Spatial interpolation is the procedure of estimating the value of properties at unsampled sites within the area covered by existing observations and in almost all cases the property must be interval or ratio scaled. Spatial interpolation can be thought of as the reverse of the process used to select the few points from a DEM which accurately represent the surface. Rationale behind spatial interpolation is the observation that points close together in space are more likely to have similar values than points far apart (Tobler's Law of Geography). Spatial interpolation is a very important feature of many GISs. Spatial interpolation may be used in GISs:

- to provide contours for displaying data graphically
- to calculate some property of the surface at a given point
- to change the unit of comparison when using different data structures in different layers
- frequently is used as an aid in the spatial decision making process both in physical and human geography and in related disciplines such as mineral prospecting and hydrocarbon exploration

Many of the techniques of spatial interpolation are two- dimensional developments of the one-dimensional methods originally developed for time series analysis. There are several different ways to classify spatial interpolation procedures: Point Interpolation/Areal Interpolation, Global/Local Interpolators, Exact/Approximate Interpolators, Stochastic/Deterministic Interpolators and Gradual/Abrupt Interpolators (Armstrong and Marciano, 1997, Wang and Armstrong 2003). In general, the data parallel criteria level is 9 and algorithm parallel is 5 - 7.

4.8 Spatial Statistical Analysis

All data have a more-or-less precise spatial and temporal label associated with them. Data that are close together in space (and time) are often more alike than those that are far apart. A spatial statistical model incorporates this spatial variation into the stochastic generating mechanism. Temporal information allows this mechanism to be

dynamic. Prediction of unobserveds from observeds and estimation of unknown model parameters are the principal forms of statistical inference. The search for well defined statistical criteria and a quantification of the variability inherent in the (optimal) predictor or estimator are instrinsic to a statistical approach.

It is almost always true that the classical, non-spatial model is a special case of a spatial model, and so the spatial model is more general (spatial-temporal models are even more general). Whether one chooses to model the spatial variation through the non-stochastic mean structure (sometimes called large-scale variation) or the stochastic-dependence structure (sometimes called small-scale variation) depends on the underlying scientific problem, and can be simply a trade-off between model fit and parsimony of the model description.

There are two different categories of spatial statistical analysis: spatial self-correlation and spatial self-regression analysis (Li 1996, Roros and Armstrong 1996). The score for data parallel is 5 and for algorithm is 3.

The summaries of above discussing are illustrated in Table 4. We found that spatial interpolation, buffers, and spatial query can be easily migrated to Grid platform. Polygon overlay and transformation could achieve better results on Grid platform. To do network analysis and spatial statistical analysis on Grid platform could be no significant improvement of performance. The most un-suitable spatial analysis on Grid platform is the spatial measurement.

Table 4. Feasibilities of spatial analysis on Grid platform

Spatial Analysis Functionalities	Algorithm Parallelity	Data Parallelity	Overall Score	Evaluation Level
Spatial Query	1	9	5.8	Better
Spatial Measurement	1	1	1	Worse
Transformation	1	3-9	2.2-5.8	Poor - Better
Buffers	1	9	5.8	Better
Overlay Analysis	1-9	7-9	4.6-9	Good – Best
Network Analysis	1	7	4.6	Good
Spatial Interpolation	5-7	9	7.4-8.2	Best
Spatial statistical Analysis	5	3	3.8	Good

5 Conclusions

Grid computing has emerged as an important new field in the distributed computing arena. It focuses on intensive resource sharing, innovative applications, and, in some cases, high-performance orientation. Grid technology is very effective method for spatial data analysis. It can give strong computing power in grid environment.

As our work was limited on the evaluation of performance of spatial analysis itself, in reality, many other factors such as hardware environment, distribution of data, etc. have to be considered. We are carrying on the research.

Acknowledgement. This publication is an output from the research projects "CAS Hundred Talents Program", "Digital Earth" (KZCX2-312) funded by Chinese Academy of Sciences and “Dynamic Monitoring of Beijing Olympic Environment Using Remote Sensing” (2002BA904B07-2) funded by the Ministry of Science and Technology, China.

References

1. Armstrong, M. P., and Marciano, R. J., 1997, Massively Parallel Strategies for Local Spatial Interpolation. Computers & Geosciences. Vol.23, No.8 , pp.859-867.
2. Cannataro, M., 2000, Clusters and grids for distributed and parallel knowledge discovery. Lecture Notes in Computer Science, Vol. 1823, 708-716, 2000.
3. Goodchild, M. F., 2001, *http://www.csiss.org/learning_resources/content/good_sa*.
4. Lanthier, M., and Nussbaurm, D., 2003, Parallel implementation of geometric shortest path algorithms. Parallel Computing, 29, 1445-1479.
5. Li, B., 1996, Implementing Spatial Statistics on Parallel Computers. In S. I. Arlinghaus, D. A. Griffith, W. C. Arlinghaus, W. D. Drake, & J. D. Nystuen (Eds.), Practical Handbook of Spatial Statistics, (New York:CRC Press) pp.107-148.
6. Pouchard, L; Cinquini, L; Drach, B; Middleton, D; Bernholdt, D; Chanchio, K; Foster, I; Nefedova, V; Brown, D; Fox, P; Garcia, J; Strand, G; Williams, D; Chervenak, A; Kesselman, C; Shoshani, A; Sim, A., 2003, An ontology for scientific information in a grid environment: The Earth system grid. In Proceeding of CCGRID 2003: 3rd IEEE/ACM International Symposium on Cluster Computing and the GRID held in Tokyo, Japan on May 12-15, 2003, pp626-632.
7. Roros, D. -K. D. and Armstrong, M. P., 1996, Using Linda to Compute Spatial Autocorrelation in Parallel. Computers & Geosciences.Vol.22, No.4, pp.425-432.
8. Roros, D. -K. D. and Armstrong, M. P., 1998, Experiments in the Identification of Terrain Features Using a PC-Based Parallel Computer. Photogrammetric Engineering & Remote Sensing.Vol.64, No.2, pp.135-142.
9. Wang, S. and Armstrong, M. P., 2003, A Quadtree Approach to Domain Decomposition for Spatial Interpolation in Grid Computing Environments. Parallel Computing. 29, 1481-1504.