Cluster Comput (2016) 19:139–152
DOI 10.1007/s10586-015-0512-2

Geographical information system parallelization for spatial big data processing: a review

Lingjun Zhao[1,2] · Lajiao Chen[1] · Rajiv Ranjan[3] · Kim-Kwang Raymond Choo[4] · Jijun He[5]

Received: 17 October 2015 / Revised: 14 November 2015 / Accepted: 15 November 2015 / Published online: 26 November 2015

Abstract With the increasing interest in large-scale, high-resolution and real-time geographic information system (GIS) applications and spatial big data processing, traditional GIS is not efficient enough to handle the required loads due to limited computational capabilities.Various attempts have been made to adopt high performance computation techniques from different applications, such as designs of advanced architectures, strategies of data partition and direct parallelization method of spatial analysis algorithm, to address such challenges. This paper surveys the current state of parallel GIS with respect to parallel GIS architectures, parallel processing strategies, and relevant topics. We present the general evolution of the GIS architecture which includes main two parallel GIS architectures based on high performance computing cluster and Hadoop cluster. Then we summarize the current spatial data partition strategies, key methods to realize parallel GIS in the view of data decomposition and progress of the special parallel GIS algorithms. We use the parallel processing of GRASS as a case study. We also identify key problems and future potential research directions of parallel GIS.

Keywords Spatial big data · Parallel GIS · Parallel spatial analysis algorithm

1 Introduction

With the rapid advancements in information communications and technologies (ICT), such as grid computing [43,72], cloud computing [70,71,73], green computing, data centre computing [9,75], high performance computing [37], and earth observation technologies [74,80], geographic information system (GIS) has developed as a useful tool and technology for geospatial data process and data applications [4]. GIS is widely used in various fields, such as resource survey, environmental assessment, disaster prediction, land management, urban planning [27].

In recent years, with the increasing of spatial data acquisition methods, the rapidly growing geospatial data has become an important part of the big data stream, Meanwhile, what GIS processes has developed from single mapping data to spatial big data. Big data, simply refers to the data whose size is too huge to be processed and interpreted by the usual methods. While, a large proportion of big data is likely to be geographically referenced and may be real time, which is called big geo data [24,28,30] or spatial big data [13,63]. Apart from the feature of geographic location, spatial big data is also characterized by three Vs: volume (or bigness), velocity (or timeliness), variety (or the disparate nature of its sources). Due to the explosive growth of big spatial data, complex and large-scale algorithms, demanding real-time response requirement, etc, traditional GIS is limited by the capability of spatial computing technologies, which become

✉ Lajiao Chen
chenlj@radi.ac.cn

✉ Jijun He
hejiun_200018@163.com

Lingjun Zhao
ljzhao@ceode.ac.cn

[1] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100096, People's Republic of China

[2] University of Chinese Academy of Sciences, Beijing, People's Republic of China

[3] University of Newcastle, Newcastle upon Tyne, UK

[4] University of South Australia, Adelaide, Australia

[5] Capital Normal University, Beijing, People's Republic of China

a major bottleneck and challenge to storage and processing in GIS field.

To address the challenges due to the demands in computational requirement to process complicated algorithms and data storage, high performance computing is generally considered to be an effective solution [50,51,89,90]. Parallel GIS [33] has emerged as a promising solution to increase computation capacities for massive spatial data processing and large-scale applications. The construction of a parallel GIS consists of three steps: Firstly, one needs to design the advanced parallel architecture and parallel strategies from the system level; Secondly, one needs to design data partition strategies and efficient parallel database to realize the parallel processing of vector and raster data typically used in sophisticated offline analysis; Thirdly, one has to adopt numerical or task parallel methods to realize spatial analysis algorithms parallelization to improve the operation efficiency. High performance computation techniques, such as designs of advanced architectures [1,84], strategies of data partition [46,92], and direct parallelization methods of spatial analysis algorithm [33], have been applied to address the limitations of existing GIS architecture. The importance of the area, parallel GIS, is evident in the increasing focus of research funding in recent years. For example, national high technology research and development program of China (863 program) launched the project of high performance GIS for new architecture platform [49]. The project focuses on the construction of GIS architecture, parallel strategies, parallelization of GIS algorithm, and other related areas. Despite the increasing importance of this topic, there is a lack of a comprehensive literature review. Therefore, in this paper, we seek to contribute to this knowledge gap by presenting an overview of the state-of-the-art advances in parallel GIS.

The rest of the paper is organized as follows. In Sect. 2, we outline the evolution of the GIS architecture. In Sect. 3, we summarize the current spatial data partition strategies, which are an important method to realize parallel GIS. Next, briefly describe the parallel processing algorithms for GIS in Sect. 4, as well as presenting a case study involving CRASS (an open source GIS software), in Sects. 4 and 5, respectively. Finally, key challenges and potential research directions of parallel GIS are presented in Sect. 6. The last section concludes this paper.

2 Architecture evolution of geographical information systems

At the beginning of 1960s, Tomlinson proposed the concept of GIS and built the first GIS—Canadian geographical information system (CGIS) in the world [68]. GIS comprises multiple techniques, including geography, cartography, computer technology, database technology, communication network technology and spatial information technology. With the development of computing mode, the system architecture of GIS evolved from the host-based centralized computing GIS structure, distributed computing GIS structures based on client/server (C/S) or browser/server (B/S), to service-oriented distributed GIS architecture and parallel GIS architecture (see Fig. 1).

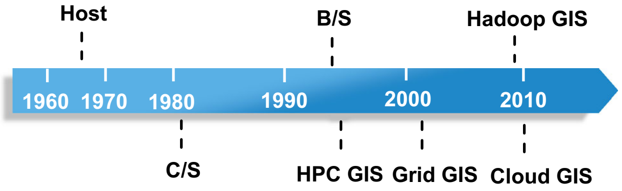

Fig. 1 Architecture evolution of GIS

2.1 Earlier architecture

GIS became popular in the 1970s, and mainly consisted of host and peripheral devices. Host-based centralized GIS [42] is simple but lacks sharing and computing capabilities.

To overcome the inherent limitations of the host-based centralized GIS architecture, the distributed GIS architecture based on the C/S [7,16] model became popular in the 1980s. The earlier distributed model consists of a local area network, which includes a multi-user shared server, and user client devices (e.g. computers). The C/S model is a semi-closed system structure, where spatial data can be shared. This is, perhaps, the most mature and known GIS architecture to date.

As web-based computing became more popular in the 1990s, GIS architecture evolved to the three-layers B/S architecture, also known as WebGIS [3]. WebGIS is a change in GIS system architecture, which makes the GIS enter into the spatial data services from the spatial data management, so that GIS applies from the level of department, enterprise into the socialization.

2.2 The overview of service-oriented distributed GIS architecture

With the increase of spatial data and the growing demand for the large-scale processing, GIS system architecture developed towards two diversification. On one hand, with the rise of new types of distributed service-oriented computing model of grid computing and cloud computing, grid GIS and cloud GIS began to develop to eliminate the resource and information island of traditional GIS. On the other hand, with the development of parallel computing, parallel GIS was born, dedicating to provide fast computing power for GIS.

2.2.1 Grid GIS

Grid computing is the collection of computer resources from multiple locations to reach a common goal. Just like the grid computing, grid GIS [58,64] can integrate multiple distributed and heterogeneous computers, spatial data server, large search storage systems, GIS, virtual reality system with high-speed network to form a super processing environment with virtual spatial information resources for the same username. Comparing with the traditional WebGIS that emphasizes to realize data sharing by the existing network, grid GIS stresses the common resource sharing, achieving real independence with the platform. The features of grid GIS are huge amounts of data distributed storage, high sharing of the spatial data and GIS service, self-regulating of each grid node and autonomy, high performance computing ability.

Generally, grid GIS adopts three-layer structure, consisting of application layer, grid service layer and data resources layer. The typical grid GIS is CyberGIS [56,76,77], which can access to the national science foundation terragrid and the open science grid, and can also process big spatial data.

2.2.2 Cloud GIS

Cloud computing is a kind of internet-based computing, where shared resources and information are provided to computers and other devices on-demand, which is an effective method to deal with big data. With the high computing capability and convenient operation management provided by Laas and Paas, cloud GIS [8] encapsulates the processing of vast amounts of geospatial data, corresponding query retrieve and geographic information processing functions as standard web services. Therefore, cloud GIS can provide the GIS application for all types of users, which is convenient, quick, stable, reliable, elastic, on-demand deployment under the support of cloud computing technology [83].

Currently, the main two could GIS platforms include ArcGIS online and SuperMap iServer 7C, both adopt four-layer cloud architecture, which are mainly similar. For example, the ArcGIS could can be divided into services portal layer, service management layer, resource pool layer, infrastructure layer [47]. Besides, Another typical cloud platform, MapGIS 10 adopts the unique T–C–V three-layer software structure, which incudes the terminal application layer (T layer), cloud computing layer (C layer) and the virtual equipment layer (V layer) [81].

2.3 The overview of parallel GIS architecture

With the invalidation of Moore's law, computing technology begins to develop towards the direction of multi-core and multi-CPU, which makes parallel method necessary to solve massive geospatial data processing. The concept of parallel GIS arises at such historic moment, dedicating to provide computing ability for vast spatial data processing and large-scale GIS applications.

Generally, the parallel GIS system architecture refers to the one based on cluster. Cluster is a method of using multi-machine high performance platform and high-speed switch to connect the cluster nodes. At present, the most common cluster includes HPC cluster and Hadoop cluster. The corresponding parallel GIS system architecture is mainly divided into parallel GIS architecture based on HPC cluster and based on Hadoop. In addition, there are other parallel GIS architectures which are few studied, such as CudaGIS [88].

2.3.1 Parallel GIS based on HPC cluster

HPC cluster is large-scale clustering system with high-speed switch, high-speed storage and high-speed processing nodes, adopting multi-core, multi-CPU and GPU technology to speedup computation. With large memory capacity and high-speed disk array system for data storage, cluster's software is fully compatible with a variety of Linux systems, most of which can support the message passing interface (MPI) and is acted as the cluster scheduling system.

In 1998, Healey put forward a ideal parallel GIS system structure based on the MPI, whose structure is composed of three layers [33], including MPI, parallel utility library (PUL) and GIS algorithms library. It is completely schemed on the MPI, with which the bottom can realize the data exchange and control information between the parallel nodes. The middle PUL, which is the core part of parallel GIS, mainly provides a unified general parallel processing interface for parallel GIS algorithms standing in the descend layer. The upper GIS library fully realizes parallel GIS algorithms Library. Healey's architecture adopted layered idea, each layer only focusing on their own functions, without interference. The structure has been gradually becoming the typical parallel GIS architecture in follow-up studies and showed as Fig. 2.

With the application of parallel file system and job scheduling, the architecture of parallel GIS system based on

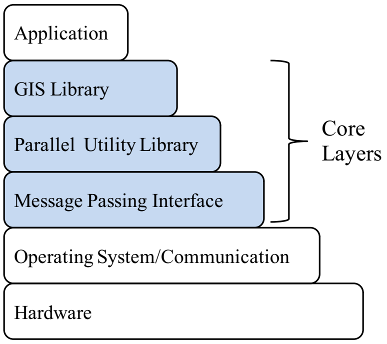

Fig. 2 Healey's ideal architecture of parallel GIS

cluster is gradually improved. At early time, Jin [39] presented a parallel GIS system architecture of four layers based on the Rocks cluster,which takes parallel task decomposition, parallel file system and task scheduling into consideration. Recently, to meet the requirement of multi-users concurrent access on WebGIS and large data updating on internet, Guo putted forward a high performance computing WebGIS model (HPCWM) [31], which is a five-layer structure including distributed massive spatial database layer, GIS server layer, cluster scheduling management layer, GIS web service layer and application layer. Both of them can perfect well in the area of massive spatial data fast network publishing, concurrent performance optimization, and visualization, etc.

2.3.2 *Parallel GIS based on Hadoop cluster*

Hadoop cluster is cluster system based on the ordinary commercial computer and network equipment, with Hadoop distributed file system (HDFS) and parallel computing architecture named MapReduce. It is relatively in low cost and is widely used in the data analysis in the internet. Hadoop has grown rapidly in recent years, thus high performance GIS based on it also gained rapid development in these years. And parallel GIS based on Hadoop is the main solution for big spatial data processing.

Hadoop-GIS, developed by Aji [1], is an extensible high-performance spatial data warehouse, with the platform of Hadoop, which can support huge amounts of data query. Hadoop-GIS consists of the five parts: data partition, data storage, QL^{sp} query language, query transformation, query engine. Hadoop-GIS realized the fast query of massive spatial data by using various technologies, such as data pre-split technology, real time spatial query engine (RESQUE), implicit parallel query algorithm based on the Map-Reduce, boundary special processing strategy. The architecture of Hadoop-GIS is showed in Fig. 3.

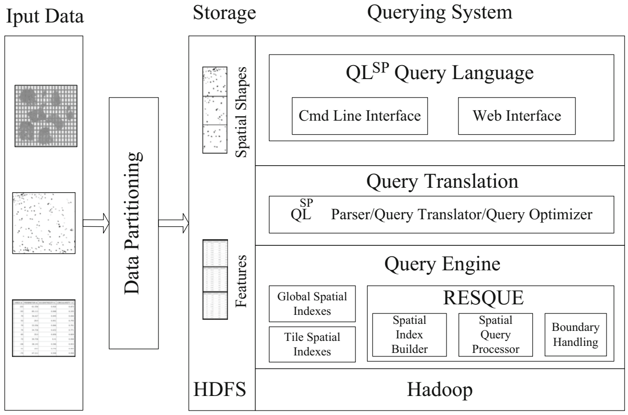

Fig. 3 Architecture of Hadoop-GIS

Zhong designed vegaGIS spatio-temporal data engine (VegaSTDE) [92] based on the cloud computing platform and the Hadoop platform, which used cluster technology, distributed storage and parallel processing technology. VegaSTDE includes VegaStore (spatiotemporal data distributed storage middleware), VegaIndexer (distributed spatiotemporal index middleware), VegaProcessor (parallel spatiotemporal query processing middleware) and VegaCache (spatiotemporal data objects distributed cache middleware). VegaSTDE effectively solves the performance bottleneck problems of storage access and query processing of the existing data management systems.

2.4 The summary of GIS architectrue development

With the development of computation patterns, GIS system has experienced the traditional architecture to the new ones. The GIS architecture is mainly based on hierarchical structure. The service-oriented distributed GIS provides a comparatively perfect spatial information service function from different angles, and improves the GIS parallel computing ability on a macro-scale. While, parallel GIS focuses on solving the problem of huge amounts of spatial data as an independent computing resources node on micro scale, providing computational support for grid GIS, cloud GIS as the extension node and realizing specific GIS application as an independent system. The summary of the new GIS architectures is showed as Table 1.

3 The spatial data parallel partition strategy

Algorithm parallelization usually adopts two ways to partition the problems as the sub-problem, including data decomposition and functional decomposition [21]. Data decomposition refers to partition the data to different nodes and each node processes its data block using the same action. The operation of the most of GIS algorithms are relatively simple whereas the amount of data is huge, therefore data partition is usually preferred.

3.1 The decomposition strategy of raster data

The raster data has a lot of dividing methods. For each dividing method, the divided sub-data should be combined in the original order. Usually the raster data is averagely divided and each sub-data needs the same computing resources and at the same time, minimizing the communication of sub-data among processors.

The raster data mainly has four partition ways: regular decomposition, irregular decomposition, scattered spatial decomposition and task farm. The regular decomposition aims at the balance problem, namely the processing time only

Table 1 Features of the new GIS architecture model

Architecture model	Typical software	Advantages	Shortages	Application scene
Grid GIS	CyberGIS, ASU	Taking advantage of distributive resource and data from internet to solve the interoperability between heterogeneous systems and share the spatial data GIS services	Need to develop complex scheduling model for developer; Could not make full use of network resources	Scientific research
Cloud GIS	ArcGIS Online, ESRI	In the way of network service, to provide the ability of storage and computing, also can create new applications; To reduce the business demand of the user and the workload of the developer; High resource utilization and low network burden	Lack of data security and privacy; Higher requirements for the network for online operation	The location and spatial service to the public
Parallel GIS based on HPC	HiGIS, NUDT, China	Strong computing power, suitable for complex operations, can achieve a variety of characteristics of the GIS computing tasks	High requirements for developers; High cost; Poor scalability	Scientific research
Parallel GIS based on Hadoop	HadoopGIS, Emory Univ.	To achieve high performance architecture at a lower cost; High scalability, high reliability; Low requirements for developers	Only suitable for data intensive computing and non real time operation	Common solution in big data processing, especially in spatial data storage and management

relies on the size of every sub-raster block (SRB) and not its contents, including horizontal stripe, vertical stripe, rectangular block. The regular decomposition consists of two ways, PUL-RD decomposition which is a trade-off between dividing the SRB into rectangular and minimizing the perimeter, and heuristic decomposition [46] which is is a trade-off between dividing the SRB into rectangular and minimizing the perimeter.

The rest of other three methods aim at the unbalanced problem. Irregular decomposition retains the way of each process processes the SRB in the regular decomposition, and Fig. 4 shows irregular decomposition compared with regular decomposition. while the scattered spatial decomposition and the task farm achieve dynamic balances by dividing multiple SRBs to processes. The feature of the four dividing method as shown in Table 2.

3.2 The decomposition strategy of vector data

Vector data has not yet gained acceptable rules for data partition, because of its complex structure and compact topological relationship between objects. Under the support of GIS-PAL, Sloan, Healey and others adopt three steps including separately the sorting, merging, entity attributes partition to do the vector data partition [65]. Up to now, the data partition methods can be summarized into three groups, they are vector partition strategies based on geometry object, strip and grid.

3.2.1 The vector partition strategy based on geometry object

Geometry object partition is the simplest strategy for vector data, which divides the same number of objects into the individual computing nodes. It does not always keep the balance on each nodes and serval experts want to improve it. Yang [84] proposed a data partition method of vector equilibrium target set, considering the task balance and load balancing needs. The vector data are evenly divided into p parts according to the number of the process p target set (number of n), and each process gets the n/p vector of target set. The vector target set is based on the geometry entity, as a result, the partition strategy is a kind of partition way based on the object. Such partition strategy which considers load balance, however, it is not sensitive to the space and is not suitable for popularization.

3.2.2 The partition strategy based on strip

In the parallelization of the GIS algorithm and the partition strategy of the vector data, the parallel architectures Laboratory for GIS in University of Edinburgh is the most typical and the most successful. Healey et al. proposed a segmentation approach based on banding using region partition to divide the vector data into strips, which are orthogonal or parallel to sweep direction. Each strip is then sent to the each compute node to participate in the parallel computing. This approach has been in the frontier of parallel GIS.

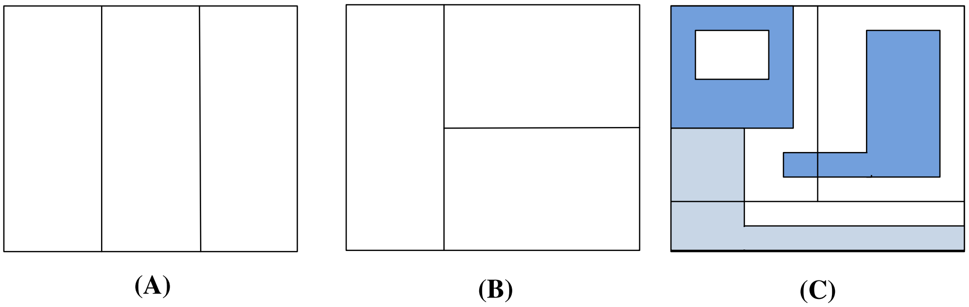

Fig. 4 Regular and irregular decomposition of raster data **a** PUL-RD partitioning, **b** heuristic partitioning, **C** irregular decmoposition

Table 2 Decomposition features of raster data

Decomposition mode	Decomposition feature	Advantages	Shortages
Regular decomposition	Each process receive one SRB that contain the same amount of data	Easy to achieve; maintain the data volume balance better	Not suitable the algorithm whose processing time is sensitive to the contents of pixels
Irregular decomposition	According to the complexity of the data, each SRB has the same amount of calculation but different amount of data size	The overheads in defined-extent, iterative algorithms would be lower than in the alternatives ones	The preliminary analysis is necessary but difficult, and can not change the distribution
Scattered spatial decomposition	Entails decomposition into a large number of SRBs of the same size. Each process has the same number of SRBS	The SRBs needing more intensive processing are distributed evenly	The additional fragmentation may lead to additional messaging
Task farm	Decomposes the data set into a number of SRBs with different size. In this case, each Worker process holds one SRB at a time, and requests its next SRB on completion of processing of the previous SRB	Little analysis is required by the source process, load-balancing is achieved	The additional fragmentation may lead to additional messaging

3.2.3 The vector partition strategy based on grid

Map-Reduce of Hadoop platform requires similar structure of each data block, so, it is especially strict with the vector partition strategy and usually divides the spatial data into grids. Zhong [92] designed a vector-based spatio-temporal data block structure (VBS) based on the space grid division. Therefore, a spatial data set is expressed as a data file composed of several VBS blocks. The Hadoop-GIS [1] also build its own sub block structure based on the division strategy of grid. Additionally, The partition of Hadoop-GIS considers the objects across regions. Comparatively speaking, grid is a two-dimensional partition, and it is more easily reached the spatial balance than the one-dimensional partition such as geometry object and strip.

3.3 The data partition strategy based on the parallel database

Parallel database system is the new generation of database system with high performance, based on massively parallel processing (MPP) and cluster parallel computing environment.The partition strategy of the parallel database is put forward, in order to solve data skew phenomenon (that is, the data block is distributed uneven in the disk), which is an effective method of improving the query efficiency of parallel spatial database.

Oracle spatial database provides a data partition strategy based on space position, similar to the process of building a grid index on the data set, and adopting the rotation method to map each grid to each node. Oracle's strategy is simple but result in storage capacity imbalance. Aimed to the imbalance problem, Zhou [93] put forward a spatial data partition method based on Hilbert space filling curve hierarchical decomposition, Wang [78] proposed another massive spatial data partition strategy, based on the Hilbert space permutation code from the perspective of spatial clustering, while Jia [38] raised a spatial data partition method according to the K-average clustering algorithm.

The parallel database partition strategies mainly belong to the static partition and do the division based on well-distributed data amount. They are easy to cause space assembling entity hard to segmented, thereby can reduce the entity aggregation after group. Most of the partition strategies deal with point instead of the surface in gather, which have some shortcoming when applied in regional scale.

3.4 The summary of spatial data parallel partition strategy

Among the four partition ways of raster data, regular decomposition is simple and clear, and it also has a direct mapping relationship with the row or line of image. So, regular decomposition, especially horizontal strip, is the most widely used one.

Among the three partition methods of vector data, although the strategy based on geometry object is simple, but its data volume balance is poor, and does not consider the spatial relationship of each object, so its partition effect is weak. The strategy based on strip is put forward earlier, but the spatial relationship between the strips is maintained well, and it is widely used in the parallelization based on MPI. While, the strategy based on grid, is further improved in the spatial relation, and is mainly adopted in the spatial big data storage and parallel spatial query.

Between the two partition methods of parallel database, the strategy based on curve is worse than the strategy based on cluster in terms of balance, while is better than the other in terms of computing time.

4 Parallel spatial analysis algorithms in GIS

As early as 1991, Richard and Steve led the study of the GIS parallel computing model and parallel computing algorithms. The most direct way to parallel the spatial analysis algorithms is analyzing the principle of the algorithms and adopting the suitable method, no matter data decomposition or functional decomposition are adopted. Data and data structure is the foundation and core of GIS. Vector and raster are the most basic data structure of GIS, which have their own advantages and disadvantage. A library of GIS serial algorithms of vector and raster has emerged in recent years, providing support for the foundation of GIS software and the various application.

4.1 Parallel algorithms of vector data

Research on parallel algorithms of vector data has been about 20 years, however, only the simple and typical spatial analysis algorithm has been parallelized, such as the shortest path analysis, superposition analysis and network analysis. In view of application, the vector algorithms can be divided into spatial index construction, spatial relationship computation and query analysis, buffer analysis, overlay analysis, network analysis.

The spatial index is a directory of the spatial dataset, aim to improve the efficiency of locating a geometry object in the collection. The performance of spatial index algorithm is very important, which directly determines the efficiency of spatial data concurrent visit, and is the key factor to solve the problem of spatial big data retrieval, query and visit. Based on the R tree, a variety of improved R tree is proposed, such as MXR-R tree [41], MCR-tree [62], GPR-tree, DPR-tree [86], Parallel R tree index [69], Upgraded Parallel R-tree [45], AP-n*mD R-tree [10] and HCMPR-tree [91].

The spatial relations mainly describe the geometry and topo relations between the spatial objects, such as containing, convered, crosses, disjoint etc, and provide the basic theory and methods for the analysis of GIS. The serial algorithm performance has been basically achieved perfection [18,22,48], but studies of spatial relations parallel algorithms rarely seen in the literature. The parallel methods mianly focus on the data partition strategy, such as the partition strategy based on geomtry [84].

Buffer analysis is one of the important spatial analysis functions of GISwhich automatically bulid some polygons of a certain width around the specified geometric objects. The current parallel studies mainly focus on data parallel. In order to avoid the low computational efficiency of complex data structure and huge data amount, Yao [85] adopted the division method of layers or data sets, dividing each data set to different computer nodes, and the result of combined calculation is the buffer zone. This parallel method only fit the application that multiple data sets work together. Pang [57] adopted the division method of geometry object, dividing different objects to each computing node, but the corresponding overlap operation was not mentioned. Fan [19] also used the idea of data parallel and MPI model to improved the parallel efficiency of buffer analysis in each steps.

Spatial overlap analysis refer to multiple layers create a new layer by geometrical logic computing. The parallel strategy of vector data mainly include three ways: pipeline stack, data parallel superposition and block type superimposed [79], but all the ways are insufficient in load balancing and frequent communication. Theoharis used two ways which are control parallel method and novel data parallel method to implement the Sutherland-Hodgman polygon clipping algorithm with multiple instruction multiple data stream (MIMD) processor [67]. Qatawneh [59] did a parallel implementation of Liang-Barsky clipping algorithm on a pipeline network configuration of four transputers. While Franklin [23] used uniform grid technique and single instruction, multiple data stream (SIMD) to parallel map overlay with good success.

Spatial network analysis calculates and analyzes the performance characteristics of the network on the basis of network topology relationship, such as the connected relation between node and node, node and line, line and line. Spatial network analysis mainly includes the shortest path, center location and vehicle routing problem (VRP) algorithm. The shortest path adopts parallel strategy of combining data segmentation and accelerate methods [15,40,66]; p-median positioning parallel algorithm mainly includes parallel scatter search algorithm [25,82] and parallel GRASP algorithm

Table 3 Characteristics of vector parallel algorithms

Functional category	Characteristics of parallel algorithms
Spatial Index	Mainly based on R tree search to generate multiple path plan, and through the R tree is decomposed into each sub tree or maintenance of R tree copies and mapped to each computing node, in order to reduce the retrieval system dependence on the root node and improve the efficiency of cable lead; Balancing of tasks and the index of load are unable to simultaneous
Calculation and analysis of spatial relations	Spatial relations computing easier to realize the parallel
Overlay analysis	Has good parallelism, but the accuracy of the algorithm is slightly less, although the realization of parallel superposition analysis, but in the parallel strategy, load balancing scheduling, node communication management is deficient
Buffer analysis	Has good parallelism, but did not give specific method for data or task decomposition, the accuracy and nuclear distribution also were not considered in the calculation of nodes within the computing task, and can not fully use mufti-core and mixed cluster hardware resources
Network analysis	Network parallel algorithm performance has not been tested by geography application, the parallel efficiency is still needed further determination

[20]. Parallel algorithm for VRP mainly includes parallel iterative abut search [12] and parallel local search mathematical programming algorithm [29].

The current characteristics of vector parallel algorithms researches are summarized in Table 3 according to the classifications mentioned above.

4.2 Parallel algorithms of raster data

Research on parallel algorithms of raster data has been carried out earlier than vector parallel algorithm. Guan and Clark achieved a more generic raster geographic parallel algorithm for computing programming tool library based on MPI which has been applied to the simulate cellular automate. In view of application, the raster algorithms can be divided into basic raster analysis, terrain analysis, cluster analysis, remote sensing analysis and so on.

Basic raster analysis computes and analogizes the raster cell, and easy to parallel. Bader [5] completed the parallel algorithms for image histogram and connected components, while Boukerram [11] made the mathematical morphology parallel algorithms at the pixel level and sub image level.

Digital terrain analysis (DTA) is a data information processing technology of terrain attribute calculation and feature extraction based on digital elevation model (DEM). In view of DTA, Bader [87] studied the parallel connected domain algorithm, Gong [26] made a parallel extraction of drainage networks from large terrain data sets using high throughput computing, while Ortega completed the parallel drainage network computation on CUDA.

Spatial clustering analysis divides the similar geometry objects into the same group. Since the Computational cost is larger, the are many parallel studies about it. The parallel researches mainly include parallel minimum spanning tree clustering [60], pPop algorithm [14], the adaptive clustering [54], the fuzzy c-means clustering algorithm based on data decomposition [44], the parallel fuzzy c-means cluster analysis [52,53] and so on.

Remote sensing digital processing carries out a series of computer operations for remote sensing digital images to obtain the expected results. With the development of high resolution technology, the amount of remote sensing image becomes more huge, and the parallel researches becomes more popular, such as parallel classification algorithm departmentalization [32], P-PCA [34] and NIPALS-PCA, GS-PCA parallel algorithm based on GPU, parallel FFT transformation [32].

Besides the academic researches, there are some other real parallel applications in business GIS software, such as parallel map tiling in SuperMap [61].

The current characteristics of raster parallel algorithms researches are summarized in Table 4 according to the categories mentioned above.

4.3 The summary of parallel spatial analysis algorithms in GIS

From the above analysis, the current studies on the parallel of the spatial analysis algorithms are presented as follows: (1) The parallel algorithms and methods are varied, but there is not a simple or public parallel way, so that the parallelization of serial algorithms is not popular enough. (2) Although there are a lot of parallel algorithms, they can not be integrated into one system for there is no standardized interface. (3) Most of the existing parallel algorithms are implemented for a special parallel environment, which are universal. (4) The mainstream commercial GIS, such as ArcGIS and SuperMap, and high performance GIS, such as CyberGIS, both only

Table 4 Characteristics of vector parallel algorithms

Functional category	Characteristics of parallel algorithms
Basic raster analysis	Most of them are realized by singal pixel or sub neighborhood window, easier to implement efficient departmentalization
Terrain analysis	In relation to global compute, difficult to parallel
Clustering analysis	Computational cost is larger, involving global operations and inter process communication
Remote sensing anlasis	Large calculation, but easier to parallelize, usually divided the image into sub image

provide a few parallel algorithms, and do not provide large-scale parallel computing functions or algorithms library. So, the research on the algorithm of parallel spatial analysis is very extensive. As far as a single algorithm is concerned, it is very mature, but as a whole, spatial parallel algorithms of tool has not yet reached the level of available algorithms library.

5 Parallel processing in GRASS: a case study

geographic resources analysis support system (GRASS GIS) is a free, open-source GIS, which is widely used in processing raster, topological vector, image and chart data, etc. Also, they are server parallel researches on it. So, we take parallel processing in GRASS as a case.

5.1 GRASS GIS introduction

The development team of GRASS is a multinational group consisting of the developers from many countries and areas, and GRASS is one of the original eight software projects of open source Geo-Spatial foundation. GRASS GIS has a clear Hierarchy architecture which are conducive to code reuse and easy to parallelize, and various algorithms and functions are encapsulated into the upper module, such as raster module, vector module, image module, etc. The software architecture of GRASS is showed in Fig. 5.

5.2 Current state of parallel GRASS

So far, many researches of the GRASS has been conducted in many aspects, such as parallel system architecture, data partition strategy, and separate algorithms, etc.

In the aspect of architecture, according to the multi-user characteristics of the GRASS GIS, Huang [35] encapsulated the GRASS modules and put forward multi-user parallel model (MUPM), making multiple GRASS users which are distributed in different nodes, carry on the subtask processing in parallel.

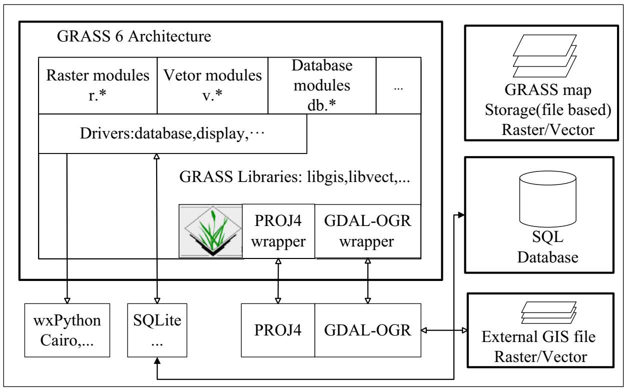

Fig. 5 Architecture of GRASS

With respect to data partition strategy, the raster data adopts partition ways in lines or blocks, whereas for vector data, there is no parallel strategy has been proposed.

As to single GRASS algorithm parallelization, Huang [35,36] presented the parallelization of inverse distance weighted (IDW) interpolation algorithm and the shortest path algorithm. Khter [2] used MPI, openMP, Ninf-G programming to compare the remote sensing algorithm efficiency based on HPC, such as vegetation indices. Aim to develop a parallel radio-coverage prediction tool, Lucas [6] presented a new approach that adopted an external database to improve the communication overlap, and to reduce the idle times for the worker processes, which exhibits better scalability than the traditional master-worker approach. Osterman [55] implement the r.cuda.los module by using parallel computation on the NVIDIA CUDA graphic cards.

GRASS research organization also begins to introduce the openMP to the GRASS module for large scale data manipulation, and parallelizes the GRASS partial differential equations library (GPDE) and gmath library.

5.3 Our study on the parallel GRASS

Our team also have done some researches in GRASS parallelization. We have accomplished a more ideal parallel GRASS architecture [17], which adds MPI layer, parallel interface layer, parallel scheduling management layer to the original structure. With our architecture, scale of parallel grass modules can be implemented without difficulty. Based on algorithm feature, We put forward entity partition and results partition. We have also parallelized nearly twenty GRASS algorithms, which include buffer analysis, overlay analysis, interpolation, hydrological analysis and so on. In this way, we have built a simulative space location application with parallel buffer and overlap algorithms with a New York vector data set which is over 2 GB, while the the original GRASS could not process such big data.

5.4 The summary of parallel GRASS

From the above studies, we can see that the parallelization of GRASS mainly focuses on raster algorithms, and little attention paid to the parallelization of vector algorithms, system architecture and data partition strategy. The parallelization study of GRASS although is rarebut it is gradually entering the stage of development. In future, the parallel GRASS facing on spatial big data will be a new research field.

6 Problems and perspective on future work

Despite there are many existing studies on parallel GIS for large scale, many technical limitations and challenges remain for their implementation.

6.1 Problems and future work in architecture

So far, although several parallel GIS architectures have been designed based on the HPC cluster or Hadoop cluster, most of these architectures are still in the stage of design and experiment. A recognized and practical application of parallel GIS system architecture has not been formed yet. In the era of big data, the future work in terms of the parallel GIS architecture may include efficient distributed storage and management strategy for big spatial data, convenient integrated strategy for large scale parallel spatial analysis algorithms, simple expansion strategy for geoprocessing flow and system architecture design of software and hardware integration. Parallel GIS can be the extension node to provide computational support for grid GIS, cloud GIS, and can also be used as an independent system, implementing GIS application's efficiency.

6.2 Problems and future work in data partition strategy

Raster data structure is simple and it's partition strategies are well developed, which have taken into account the load balance and spatial aggregation of data. With respect to vector data, due to the complex structure and compact topological relationships between objects, there didn't form a common partition strategy for the vector data yet. Many problems remain in the currently strategies. Firstly, it is difficult to reconcile the balance of the amount and spatial aggregation of data. Secondly, the features of the vector object's characteristics of cross-regional coverage has rarely been considered. Thirdly, spatial association between data sets with a large of data to calculate simultaneously. In the future work, more attention should been paid to a strategy which is compatible with both spatial aggregation and calculation balance, entity integrity, and databases spatial association, in order to improve GIS algorithm parallel efficiency on the data decomposition.

6.3 Problems and future work in parallel processing algorithms

The present situation of the algorithm can be summarized into two aspects. On one hand, the parallelization of raster-based spatial analysis algorithm is relatively mature while parallelization of vector-based analysis algorithms is still in the early stage. Most of the vector-based research is for specific algorithm which is not systematic and far from actual application. So, great efforts should been paid to complete the parallel transformation especially for each vector-based spatial analysis algorithms and to add a large number of experiments and debugging of parallelized algorithms for better parallel efficiency. On the other hand, the object of the parallel algorithms is mainly the traditional static, formative spatial data, and can not meet the application requirements of large spatial data. In future, there will be at least two studies we should do. One is that, for the most direct parallel of spatial analysis algorithms combined with data partition strategies, we should present a relatively easy way for parallel implementation on the system level so as to reduce the difficulty in the development of the GIS algorithm, the other is that more and more parallel spatial analysis algorithms, parallel spatial data mining algorithms should be developed for big spatial data that with the feature of volume, velocity, variety, veracity and value.

7 Conclusions

This paper provides a perspective on the current state of parallel GIS for spatial big data. We firstly summarize the progress of the GIS architecture which are based on HPC cluster and Hadoop cluster. Then we summarize the current spatial data partition strategies and progress of the special parallel GIS algorithms. Next, parallel processing of GRASS, an open GIS software, is illustrated as a case study of parallel GIS. In the end, The key problems and future potential research directions are addressed, such as a parallel GIS architecture incorporated with efficient storage and computing, dynamical business work flow combination, a multiplicity balanced data partition strategy and a relatively easy parallel way in the system level.

Acknowledgments This study is supported by National Natural Science Foundation of China (41301028).

References

1. Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J.: Hadoop gis: a high performance spatial data warehousing system over mapreduce. Proc. VLDB Endow. **6**(11), 1009–1020 (2013)
2. Akhter, S., Aida, K., Chemin, Y.: Grass gis on high performance computing with mpi, openmp and ninf-g programming framework. In: Proceeding of ISPRS 2010 (2010)

3. Alesheikh, A., Helali, H., Behroz, H.: Web gis: technologies and its applications. In: Symposium on Geospatial Theory, Processing and Applications, vol. 15 (2002)
4. Aronoff, S.: Geographic Information Systems: A Management Perspective. Taylor & Francis, London (1989)
5. Bader, D.A., JáJá, J.: Parallel algorithms for image histogramming and connected components with an experimental study (1998)
6. Benedičič, L., Cruz, F.A., Hamada, T., Korošec, P.: A grass gis parallel module for radio-propagation predictions. Int. J. Geogr. Inf. Sci. **28**(4), 799–823 (2014)
7. Berson, A.: Client-Server Architecture. IEEE-802. McGraw-Hill, New York (1992)
8. Bhat, M.A., Shah, R.M., Ahmad, B.: Cloud computing: a solution to geographical information systems(gis). Int. J. Comput. Sci. Eng. **3**(2), 594–600 (2011)
9. Bilal, K., Khan, S.U., Zhang, L., Li, H., Hayat, K., Madani, S.A., Min-Allah, N., Wang, L., Chen, D., Iqbal, M.I., Xu, C.Z., Zomaya, A.Y.: Quantitative comparisons of the state-of-the-art data center architectures. Concurr. Comput. Pract Exp. **25**(12), 1771–1783 (2013). doi:10.1002/cpe.2963
10. Bok, K., Seo, D., Song, S., Kim, M., Yoo, J.: An index structure for parallel processing of multidimensional data. In: Advances in Web-Age Information Management, pp. 589–600. Springer, New York (2005)
11. Boukerram, A., Azzou, S.A.K.: Parallelisation of algorithms of mathematical morphology. J. Comput. Sci. **2**(8), 615–618 (2006)
12. Cordeau, J.F., Maischberger, M.: A parallel iterated tabu search heuristic for vehicle routing problems. Comput. Oper. Res. **39**(9), 2033–2050 (2012)
13. Dalton, C.M., Thatcher, J.: Inflated Granularity: Spatial Big Dataand Geodemographics. Available at SSRN 2544638 (2015)
14. Dash, M., Petrutiu, S., Scheuermann, P.: ppop: fast yet accurate parallel hierarchical clustering using partitioning. Data Knowl. Eng. **61**(3), 563–578 (2007)
15. Delling, D., Katz, B., Pajor, T.: Parallel computation of best connections in public transportation networks. J. Exp. Algorithmics **17**, 4–4 (2012)
16. Dewitt, D.J., Kabra, N., Luo, J., Patel, J.M., Yu, J.B.: Client-server paradise. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 558–569 (2001)
17. Dong, W., Liu, D., Zhao, L.: A new mpi-based grass technology for parallel processing and its architecture[j]. Remote Sens. Inf. **28**(01), 102–109 (2013)
18. Egenhofer, M.J.: Reasoning about binary topological relations. In: Advances in Spatial Databases, pp. 141–160. Springer, New York (1991)
19. Fan, J., Ji, M., Gu, G., Sun, Y.: Optimization approaches to mpi and area merging-based parallel buffer algorithm. Boletim de Ciências Geodésicas **20**(2), 237–256 (2014)
20. Festa, P., Resende, M.G.: Hybridizations of grasp with path-relinking. In: Hybrid Metaheuristics, pp. 135–155. Springer, New York (2013)
21. Foster, I.: Designing and Building Parallel Programs. Addison Wesley Publishing Company, Reading (1995)
22. Frank, A.U.: Qualitative spatial reasoning: cardinal directions as an example. Int. J. Geogr. Inf. Sci. **10**(3), 269–290 (1996)
23. Franklin, W.R., Narayanaswami, C., Kankanhalli, M., Sun, D., Zhou, M.C., Wu, P.Y.: Uniform grids: a technique for intersection detection on serial and parallel machines. In: Proceedings of Auto Carto 9: Ninth International Symposium on Computer-Assisted Cartography, pp. 100–109 (1989)
24. Gao, S., Li, L., Li, W., Janowicz, K., Zhang, Y.: Constructing gazetteers from volunteered big geo-data based on hadoop. Comput. Environ. Urban Syst. (2014). doi:10.1016/j.compenvurbsys.2014.02.004
25. Garcıa-López, F., Melián-Batista, B., Moreno-Pérez, J.A., Moreno-Vega, J.M.: Parallelization of the scatter search for the p-median problem. Parallel Comput. **29**(5), 575–589 (2003)
26. Gong, J., Xie, J.: Extraction of drainage networks from large terrain datasets using high throughput computing. Comput. Geosci. **35**(2), 337–346 (2009)
27. Goodchild, M.F.: Geographical information science. Int. J. Geogr. Inf. Syst. **6**(1), 31–45 (1992)
28. Goodchild, M.F.: The quality of big (geo) data. Dialogues Human Geogr. **3**(3), 280–284 (2013)
29. Groër, C., Golden, B., Wasil, E.: A parallel algorithm for the vehicle routing problem. INFORMS J. Comput. **23**(2), 315–330 (2011)
30. Guo, H., Wang, L., Chen, F., Liang, D.: Scientific big data and digital earth. Chin. Sci. Bull. **59**(35), 5066–5073 (2014). doi:10.1007/s11434-014-0645-3
31. Guo, M.: Research on the key technologies of high performance computing webgis model. Ph.D. thesis, China University of Geosciences, Wuhan (2012)
32. Hawick, K.A., Coddington, P.D., James, H.A.: Distributed frameworks and parallel algorithms for processing large-scale geographic data. Parallel Comput. **29**(10), 1297–1333 (2003)
33. Healey, R., Dowers, S., Gittings, B., Mineter, M.J.: Parallel Processing Algorithms for GIS. CRC Press, Basingstoke (1997)
34. Hu, B., Wang, H.F., Wang, P.F., Liu, H.Z.: A parallel algorithm of pca image fusion in remote sensing and its implementation. Microelectron. Comput. **23**(10), 153–157 (2006)
35. Huang, F., Liu, D., Liu, P., Wang, S., Zeng, Y., Li, G., Yu, W., Wang, J., Zhao, L., Pang, L.: Research on cluster-based parallel gis with the example of parallelization on grass gis. In: Sixth International Conference on Grid and Cooperative Computing, 2007. GCC 2007, pp. 642–649. IEEE (2007)
36. Huang, F., Liu, D., Tan, X., Wang, J., Chen, Y., He, B.: Explorations of the implementation of a parallel idw interpolation algorithm in a linux cluster-based parallel gis. Comput. Geosci. **37**(4), 426–434 (2011)
37. Hussain, H., Malik, S.U.R., Hameed, A., Khan, S.U., Bickler, G., Min-Allah, N., Qureshi, M.B., Zhang, L., Wang, Y., Ghani, N., Kolodziej, J., Zomaya, A.Y., Xu, C.Z., Balaji, P., Vishnu, A., Pinel, F., Pecero, J.E., Kliazovich, D., Bouvry, P., Li, H., Wang, L., Chen, D., Rayes, A.: A survey on resource allocation in high performance distributed computing systems. Parallel Comput. **39**(11), 709–736 (2013)
38. Jia, T., Wei, Z., Tang, S., Kim, J.H.: New spatial data partition approach for spatial data query. Comput. Sci. **37**(8), 198–200 (2013)
39. Jin, H., Meng, L., Wang, X.: Cluster-based architecture design of parallel gis [j]. Geospat. Inf. **5**, 015 (2005)
40. Kalpana, R., Thambidurai, P.: Optimizing shortest path queries with parallelized arc flags. In: International Conference on Recent Trends in Information Technology (ICRTIT), 2011, pp. 601–606. IEEE (2011)
41. Kamel, I., Faloutsos, C.: Parallel R-Trees, vol. 21. In: ACM (1992)
42. Katz, R.H.: High-performance network and channel-based storage. Proc. IEEE **80**(8), 1238–1261 (1992)
43. Kolodziej, J., Khan, S.U., Wang, L., Byrski, A., Min-Allah, N., Madani, S.A.: Hierarchical genetic-based grid scheduling with energy optimization. Clust. Comput. **16**(3), 591–609 (2013). doi:10.1007/s10586-012-0226-7
44. Kwok, T., Smith, K., Lozano, S., Taniar, D.: Parallel fuzzy c-means clustering for large data sets. In: Euro-Par 2002 Parallel Processing, pp. 365–374. Springer, New York (2002)
45. Lai, S., Zhu, F., Sun, Y.: A design of parallel r-tree on cluster of workstations. In: Databases in Networked Information Systems, pp. 119–133. Springer, New York (2000)
46. Lee, C.K., Hamdi, M.: Parallel image processing applications on a network of workstations. Parallel Comput. **21**(1), 137–160 (1995)

47. Lin, D., Liang, Q.: Research progress and connotation of cloud gis [j]. Prog. Geogr. **11**, 013 (2012)
48. Liu, D., Liu, Y.: A review on spatial reasoning and geographic information system. J. Softw. **11**(12), 1598–1606 (2000)
49. Liu, L., Yang, A., Chen, L., Xiong, W., Wu, Q., Jing, N.: Higis-when gis meets hpc. In: 12th International Conference on GeoComputation, Wuhan (2013)
50. Liu, P., Yuan, T., Ma, Y., Wang, L., Liu, D., Yue, S., Kolodziej, J.: Parallel processing of massive remote sensing images in a gpu architecture. Comput. Inf. **33**(1), 197–217 (2014)
51. Ma, Y., Wang, L., Liu, D., Yuan, T., Liu, P., Zhang, W.: Distributed data structure templates for data-intensive remote sensing applications. Concurr. Comput. Pract. Exp. **25**(12), 1784–1797 (2013). doi:10.1002/cpe.2965
52. Modenesi, M.V., Costa, M.C., Evsukoff, A.G., Ebecken, N.F.: Parallel fuzzy c-means cluster analysis. In: High Performance Computing for Computational Science-VECPAR 2006, pp. 52–65. Springer, New York (2007)
53. Modenesi, M.V., Evsukoff, A.G., Costa, M.C.: A load balancing knapsack algorithm for parallel fuzzy c-means cluster analysis. In: High Performance Computing for Computational Science-VECPAR 2008, pp. 269–279. Springer, New York (2008)
54. Nagesh, H., Goil, S., Choudhary, A.: Parallel algorithms for clustering high-dimensional large-scale datasets. In: Data Mining for Scientific and Engineering Applications, pp. 335–356. Springer, New York (2001)
55. Osterman, A.: Implementation of the r. cuda. los module in the open source grass gis by using parallel computation on the nvidia cuda graphic cards. ELEKTROTEHNIËĞSKI VESTNIK **79**(1–2), 19–24 (2012)
56. Padmanabhan, A., Wang, S., Navarro, J.P.: A cybergis gateway approach to interoperable access to the national science foundation teragrid and the open science grid. In: Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery, p. 42. ACM (2011)
57. Pang, L., Li, G., Yan, Y., Ma, Y.: Research on parallel buffer analysis with grided based hpc technology. In: IEEE International Geoscience and Remote Sensing Symposium, 2009, IGARSS 2009, vol. 4, pp. IV–200. IEEE (2009)
58. Paulsen, J., Körner, C.: Gis-analysis of tree-line elevation in the swiss alps suggests no exposure effect. J. Veg. Sci. **12**(6), 817–824 (2001)
59. Qatawneh, M., Sleit, A., Almobaideen, W.: Parallel implementation of polygon clipping using transputer. Am. J. Appl. Sci. **6**(2), 214 (2009)
60. Rajasekaran, S.: Efficient parallel hierarchical clustering algorithms. IEEE Trans. Parallel Distrib. Syst. **6**, 497–502 (2005)
61. Rao, Q., Ding, J., Su, L., Gu, Y., Xia, L., Hu, Z.: The design and implementation of distributed map tiling service based on cloud computing. Geomat. Spat. Inf. Technol. **36**, 29–35 (2013)
62. Schnitzer, B., Leutenegger, S.T.: Master-client r-trees: a new parallel r-tree architecture. In: Eleventh International Conference on Scientific and Statistical Database Management, 1999, pp. 68–77. IEEE (1999)
63. Shekhar, S., Gunturi, V., Evans, M.R., Yang, K.: Spatial big-data challenges intersecting mobility and cloud computing. In: Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access, pp. 1–6. ACM (2012)
64. Shen, Z., Luo, J., Zhou, C., Cai, S., Zheng, J., Chen, Q., Ming, D., Sun, Q.: Architecture design of grid gis and its applications on image processing based on lan. Inf. Sci. **166**(1), 1–17 (2004)
65. Sloan, T.M., Mineter, M.J., Dowers, S., Mulholland, C., Darling, G., Gittings, B.M.: Partitioning of vector-topological data for parallel gis operations: Assessment and performance analysis. In: Euro-Par'99 Parallel Processing, pp. 691–694. Springer, New York (1999)
66. Sun, W., Tan, Z., Wang, J., Zhou, C., He, J.: An analysis of parallelizing shortest path algorithm. Geogr. GeoInf. Sci. **4**, 005 (2013)
67. Theoharis, T., Page, I.: Two parallel methods for polygon clipping. In: Computer Graphics Forum, vol. 8, pp. 107–114. Wiley Online Library (1989)
68. Tomlinson, R.F., Calkins, H.W., Marble, D.F.: Computer Handling of Geographical Data. UNESCO Press, Paris (1976)
69. Wang, B., Horinokuchi, H., Kaneko, K., Makinouchi, A.: Parallel r-tree search algorithm on dsvm. In: Proceedings of the 6th International Conference on Database Systems for Advanced Applications, 1999, pp. 237–244. IEEE (1999)
70. Wang, L., Chen, D., Hu, Y., Ma, Y., Wang, J.: Towards enabling cyberinfrastructure as a service in clouds. Comput. Electr. Eng. **39**(1), 3–14 (2013)
71. Wang, L., Kunze, M., Tao, J., von Laszewski, G.: Towards building a cloud for scientific applications. Adv. Eng. Softw. **42**(9), 714–722 (2011)
72. Wang, L., von Laszewski, G., Kunze, M., Tao, J., Dayal, J.: Provide virtual distributed environments for grid computing on demand. Adv. Eng. Softw. **41**(2), 213–219 (2010)
73. Wang, L., von Laszewski, G., Younge, A.J., He, X., Kunze, M., Tao, J., Fu, C.: Cloud computing: a perspective study. New Gener. Comput. **28**(2), 137–146 (2010)
74. Wang, L., Lu, K., Liu, P.: Compressed sensing of a remote sensing image based on the priors of the reference image. IEEE Geosci. Remote Sens. Lett. **12**(4), 736–740 (2015)
75. Wang, L., Tao, J., Ma, Y., Khan, S.U., Kolodziej, J., Chen, D.: Software design and implementation for mapreduce across distributed data centers. Int. J. Appl. Math. Inf. Sci. **7**(1), 85–90 (2013)
76. Wang, S.: A cybergis framework for the synthesis of cyberinfrastructure, gis, and spatial analysis. Ann. Assoc. Am. Geogr. **100**(3), 535–557 (2010)
77. Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M.F., Liu, Y., Nyerges, T.L.: Cybergis software: a synthetic review and integration roadmap. Int. J. Geogr. Inf. Sci. **27**(11), 2122–2145 (2013)
78. Wang, Y., Meng, L., Zhao, C.: The research of massive spatial data partitioning algorithm, based on the hilbert space permutation code. Geomat. Inf. Sci. Wuhan Univ. **32**(7), 650–653 (2007)
79. Wilson, G.: Assessing the usability of parallel programming systems: The cowichan problems. In: Proceedings of the IFIP Working Conference on Programming Environments for Massively Parallel Distributed Systems, pp. 183–193 (1994)
80. Wu, X., Huang, B., Wang, L., Lu, K., Zhang, J.: Gpu-based parallel design of the hyperspectral signal subspace identification by minimum error (hysime). IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. Accepted (2015)
81. Wu, X., Xu, S., Wan, B., Wu, L.: Next generation software architecture t-c-v. Earth Sci. J. China Univ. Geosci. **39**(2), 221–226 (2014)
82. Yan, Z., Sun, W., Zhou, C., Xiong, T., Wang, J.: A parallel scatter search algorithm for the p-median problem. Geogr. GeoInf. Sci. **4**, 011 (2013)
83. Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., Bambacus, M., Fay, D.: Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? Int. J. Digit. Earth **4**(4), 305–329 (2011)
84. Yang, Y., Lixin, W.: A vector data partitioning method for realizing efficient parallel computing of topological relations. Geogr. GeoInf. Sci. **29**(7), 25–29 (2013)
85. Yao, Y., Gao, J., Meng, L., Deng, S.: Parallel computing of buffer analysis based on grid computing [j]. Geospat. Inf. **1**, 035 (2007)
86. Yu, B., Hao, Z.: Research of distributed and parallel spatial index mechanism based on dpr-tree [j]. Comput. Technol. Dev. **6**, 012 (2010)

87. Zhang, J., Xu, M.: Design and implementation of connected component labeling parallel algorithm with multi-core processor. Comput. Syst. Appl. **19**(4), 140–143 (2010)
88. Zhang, J., You, S.: Cudagis: report on the design and realization of a massive data parallel gis on gpus. In: Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming, pp. 101–108. ACM (2012)
89. Zhang, W., Wang, L., Liu, D., Song, W., Ma, Y., Liu, P., Chen, D.: Towards building a multi-datacenter infrastructure for massive remote sensing image processing. Concurr. Comput. Pract. Exp. **25**(12), 1798–1812 (2013)
90. Zhang, W., Wang, L., Ma, Y., Liu, D.: Design and implementation of task scheduling strategies for massive remote sensing data processing across multiple data centers. Software: Practice and Experience 44(7), 873–886 (2014)
91. Zhao, Y., Li, C.: Research on the distributed parallel spatial indexing schema based on r-tree. Geogr. GeoInf. Sci. **6**, 009 (2007)
92. Zhong, Y.: Towards distributed management scheme for big spatiotemporal data. Ph.D. thesis, Institute of Computing Technology, Chinese Academy of Sciences, Beijing (2013)
93. Zhou, Y., Zhu, Q., Yeting, Z.: The spatial data partitioning method, based on the hilbert curve hierarchical decomposition. Geogr. GeoInf. Sci. **23**(4), 13–17 (2007)

Lingjun Zhao is a Senior Engineer in institute for remote sensing & digital earth, Chinese Academy of Sciences. He works in the research fields of satellite data processing and assessment along with parallel GIS. He graduated with a bachelor's degree and a master's degree one after another from Jilin University, currently he is studying for a Ph.D. in University of Chinese Academy of Sciences. He is now the engineer in charge of projects from National Development and Reform Commission.

Lajiao Chen received the Bachelor of Geography and the Master of Physical Geography degrees from Zhejiang Normal University, Jinhua, China, and the Doctor of Geographic Information System degree from the Chinese Academy of Sciences (CAS), Beijing, China. She is currently an Assistant Professor with the Institute of Remote Sensing and Digital Earth, CAS. Her research is focused on geocomputing.

Rajiv Ranjan is an Associate Professor (Reader) in Computing Science at Newcastle University, United Kingdom. At Newcastle University he is working on projects related to emerging areas in parallel and distributed systems (Cloud Computing, Internet of Things, and Big Data). Previously, he was Julius Fellow (2013–2015), Senior Research Scientist (equivalent to Senior Lecturer in Australian/UK University Grading System) and Project Leader in the Digital Productivity and Services Flagship of Commonwealth Scientific and Industrial Research Organization (CSIRO — Australian Government's Pllremier Research Agency). Prior to that he was a Senior Research Associate (Lecturer level B) in the School of Computer Science and Engineering, University of New South Wales (UNSW). He has a Ph.D. (2009) in Computer Science and Software Engineering from the University of Melbourne. He is broadly interested in the emerging areas of distributed systems. The main goal of his current research is to advance the fundamental understanding and state of the art of provisioning and delivery of application services (web, big data analytics, content delivery networks, and scientific workflows) in large, heterogeneous, uncertain, and emerging distributed systems.

Kim-Kwang Raymond Choo is a Fulbright Scholar and Senior Researcher in Cyber Security and Cloud Forensics at University of South Australia. He is currently a Visiting Scholar at INTERPOL Global Complex for Innovation (IGCI; Aug 2015–Feb 2016). He has co-edited "Cloud Security Ecosystem" (Elsevier 2015), and (co)authored two books (Springer 2008; Elsevier 2014—Forewords written by Australia's Chief Defence Scientist and Chair of the Electronic Evidence Specialist Advisory Group), seven Australian Government refereed monographs, 156 refereed book chapters, journal and conference articles. He has been a Keynote/Plenary Speaker at conferences organized by Infocomm Development Authority of Singapore (2015), CSO Australia and Trend Micro (2015), Anti-Phishing Working Group (2014), National Taiwan University of Science and Technology (2014), Asia Pacific University of Technology & Innovation (2014), Nanyang Technological University (2011), and National Chiayi University (2010); and more recently in 2015, an Invited Expert at events organized by UNAFEI, INTERPOL, Taiwan Ministry of justice Investigation Bureau, and at the World Internet Conference (Wuzhen Summit) in 2014 and 2015, jointly organized by the Cyberspace Administration of China and the People's Government of Zhejiang Province. He has also examined theses from University of Ballarat (Australia), University of Wollongong

(Australia), Melbourne University (Australia), University of Waikato (New Zealand), Macquarie University (Australia), Queensland University of Technology (Australia), University of Pretoria (South Africa), and Victoria University of Wellington (New Zealand).

Jijun He got Ph.D. from Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (CAS) on July 2009. Dr. He currently is an Associate Professor at the College of Resource Environment and Tourism in Capital Normal University. His research interests include soil erosion, environmental assessment, and remote sensing application.