

Research on data storage system of geographical spatial information based on grid

Chunhua Zhang^{1, a} Yueguang Li^{2, b}

¹Gansu Normal University for Nationalities, Hezuo, gansu, 747000, China

²Gansu Normal University for Nationalities, Hezuo, gansu, 747000, China

^aYueguangli7@163.com, ^bYueguangli7@sohu.com

Keywords: Grid; Geographic information; spatial data; storage system

Abstract: This paper proposes and constructs a network storage engine of spatial data based on grid. Realize storage of spatial data that is located in different places and effective management using grid, server cluster, spatial data division and data distribution to the storage devices linked by high-speed network.

1 Introduction

In recent years, with the popularization of information technology, data storage volume increases by 80% every year. About the rate of growth of information, Gray proposed an empirical law: the amount of data generated every 18 months under network environment is equal to the amount of data in history^[1]. For the spatial data of scientific computation, current, although the reading rate of the single disk has generally reached more than 100m/s, but as the number of storage space of a single disk and concurrent constraints, so that it can not solve the massive spatial data management and multi users high speed concurrent access^[2]. Parallel computing is an effective way to improve storage efficiency of spatial information, and the grid is a powerful weapon to provide high performance parallel computing.

This paper uses the grid ideas and server cluster, the spatial data uses the method of block, the data block is distributed to the storage device of a high-speed network, so when the users request data, storage devices can read or write the spatial data in a parallel manner, although the I/O performance of individual disks is low, but the network can take the low rate of disk I/O parallel the high rate data network stream. Therefore, in this paper, several typical network parallel storage structure based on the comparison and analysis, construct a reasonable, efficient spatial data storage system based on grid environment.

At present, there are mainly typical parallel data storage systems: 1. Direct attached storage (DAS); 2. Storage Area Network (SAN); 3. Network Attached storage (NAS); 4. With the file system SAN; 5. Object based storage device (OSD). Table 1 is the comparison table of parallel data storage system^[3, 4].

Table 1 The comparison table of parallel data storage system

System structure	Performance	Scalability	Cross platform	Stability	Ease of use	Sharing and security.	Performance
DAS	Bad	Bad	Bad	Bad	Good	Good	Good
SAN	General	General	Bad	General	Good	General	Bad
NAS	Good	General	Bad	General	General	General	Good
the file system SAN	Good	Good	Good	General	General	Good	Bad
OSD	Good	Good	Good	Good	Good	Good	Good

It can see from table 1, storage system OSD is high performance, security and scalability, it can basically solve the problems of spatial data network storage. It is an ideal system structure of geographic spatial network data storage on grid environment.

2 Construction of the spatial data storage system based on distributed parallel

2.1 Construction of the spatial data storage server

Spatial data storage server is responsible for the actual spatial data storage and handling the interaction between all clients and physical storage. Spatial information data files are stored in the object storage servers, or storage server is responsible for the actual spatial information file management, various types of spatial information such as create, delete, read, write and be responsible for reporting their management to metadata server, such as the spatial information of name, position, size, modification date, format, and receiving query from the metadata server.

In order to improve the parallel speed of spatial data access operations, the design thought of the paper is: a spatial data file is divided into data bars according to certain block algorithm, and then the data bars will be divided in multiple spatial data storage server on the basis of RAID algorithm, it can be seen in figure 1. If a data file is stored in the multiple spatial data storage server, so the user can send read / write requests to multiple servers simultaneously, it can greatly improve system performance.

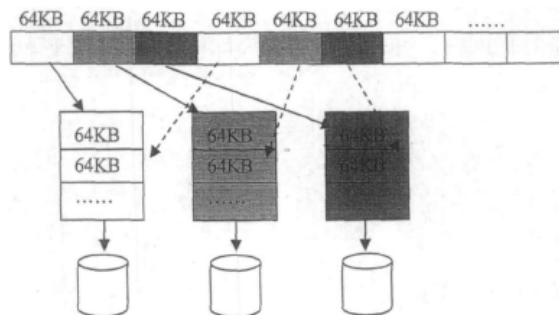


Fig.1 Flow chart of data storage space

2.2 Spatial data index

Spatial index is a kind of spatial data structure, it refers to certain spatial relationship between spatial position and shape of spatial entity or entities, a data structure are arranged according to a certain order [5].

The grid index idea is simple, easy to understand and easy to realize, and it has advantage on graphic display and zoom, meet the habit of people on the map subdivision and the convenience of multi scale spatial data nested operation, but the grid index not suitable for the requirements of current map. In the multi-stage mixed show rule map and non rule map, the traditional grid index is incapable of action. Aiming at this problem, this paper designs an adaptive double grid index. The specific algorithm is as follows:

The establishment of adaptive double grid index model step is: First let the computer determine the map is regular map or irregular map, if it is non rules map, then use the grid tree index to establish the image index, if it is the rule map, then use the grid index to establish map index. Code the grid tree index and the grid index. The establishment of the grid index of the geographic feature in each layer.

For regular map, it is uses the grid index, not uses grid tree index. The grid tree index is used a layer or map of the minimum bounding rectangle to represent the object. The grid tree is made up of the root node, leaf nodes and non leaf nodes. The node is made up of many records (index, P) forms. For a leaf node, index is the minimum bounding rectangle of a data object, P is the object identifier. For the non leaf node, MBR is the minimum bounding rectangle of all spatial data object of the grid tree node by the P.

Figure 2 shows a schematic diagram of a grid tree structure, solid line rectangle in the figure represents the circumscribed rectangle of spatial entity, in this paper refers to the circumscribed rectangle of the map. The grid tree of order is M, the layer or map number is n, the sector size is fan.

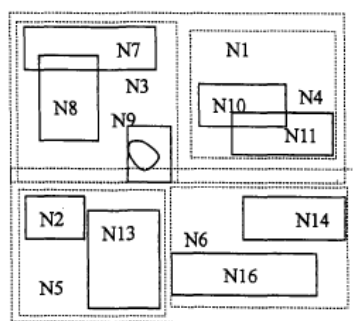


Fig.2 The grid tree structure chart

The second level index builds grid index of the spatial elements of layer. As shown in Figure 3, in the grid index technology, rectangular geographic scope of a layer are average divided into M rows and N columns by $M \times N$ a small rectangular grids, we put each grid area as an index, the index entries record the entity number and external rectangular geometry of entity, and then operate on the grid sequence, so as to accelerate the operation speed of space target. The establishment of the grid index step is:

1. Establish a unified serial number for each grid.
2. Calculate the offset of each element to file header and data block.
3. Traversal the spatial data files, calculate the grid number of the geographical elements.
4. Each grid index will be written into the file according to the serial number.

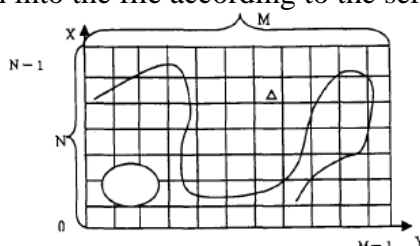


Fig.3 The grid index chart

3 The system performance testing

Run environment of algorithm is MATLAB (R2010b) under Win7 operating system. Project testing based on the environment of campus network (outlet bandwidth is 1G), 18 terminals are concurrent tested at the same time, each terminal has 5 threads simultaneously request data to the cluster. Table 2 is the test schedule of a terminal. Figure4 is cluster response line chart.

Table 2The efficiency table of the cluster

The number of operations	Cluster response time / ms	Total time/ ms
1	160	600
2	40	156
3	470	763
4	16	390
5	0	130

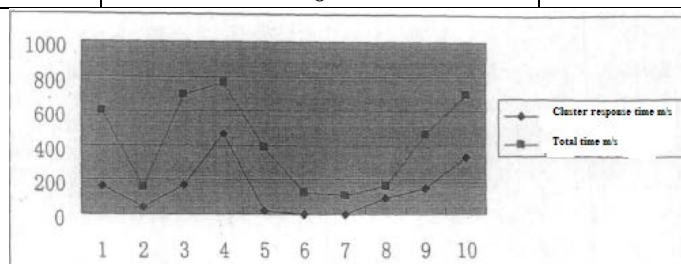


Fig. 4 The cluster response line chart

It can be seen from Figure 4, the cluster corresponding to each client time no more than 0.5 seconds, so the ability of cluster provides data service is far greater than 18*10 users access. The

fifth operating time is 0. It is due to the client cache and cluster lead to the overall performance of the system upgrade.

4 Conclusions

This paper firstly analyzes and compares the advantages and disadvantages of the several storage systems, then according to the characteristics of spatial data services in grid environment, the choice of a storage system: storage system based on object. On this basis, this paper proposes the construction rules of spatial data storage system under grid environment, and according to this rule, establish the spatial data storage system formed by spatial data storage server, metadata directory server and server load balancing, the related algorithms are described in detail, finally, test and verify the system.

Acknowledgment

This work is supported by Gansu Provincial College graduate tutor of scientific research project (1112-09) and Dean Fund of Gansu Normal College for Nationalities (09-07).

References

- [1] Guo Yudong, Yin Qing. Network storage based on object [M]. Electronics industry press. 2007.
- [2] Zhang Yongsheng, Cao Qihua. Structure analysis of the spatial data server based on distributed parallel [J]. Journal of Surveying and Mapping Science and technology. 2007 (05):67-70.
- [3] Lustre.<http://www.Lustre.Org>.
- [4] Panasas. <http://www.Panasas.com>.
- [5] Guttman A.R-Trees: A dynamic index structure for spatial searching. Proc, ACM SIGMOD int. Conf on Management of Data. Boston, MA. 2008(07):47~57.