MANAGING THE GLOBAL DATASET IN BLOCK-INDEX BASED ON ESSG

YANG Yizhou[a], WU Lixin[a,b,c*], YU Jieqing[b]

a Institute of Geoinformatics and Digital Mine Research, Northeastern University, Shenyang, China - yangyizhou1987@163.com

b School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou, China - awulixin@263.net

c Academy of Disaster Reduction and Emergency Management, Beijing Normal University, Beijing, China

Commission VI, WG VI/4

KEY WORDS: SDOG, ESSG, Global Scale, Raster Data

ABSTRACT:

Earth observation system, which can realize global coverage in three dimensions, has provided massive data from earth observations at global scale for Earth System Science (ESS) and global change researches. It is named SDOG-based ESSG that the Spheroid Degenerated Octree Grid (SDOG) was chosen as an initial grid to develop an Earth System Spatial Grid (ESSG), which provides storage strategy and management framework based on triple (C,T,A) and SDOG grid coding for massive data. The objection of this paper is to provide an effective spatial data organization and indexing method, which organizes triple units into block and encodes for each triples block and converts multi-source global scale datasets into triples block so as to manage massive raster data at global scale.

1. INTRODUCTION

With the continuous development of earth observation technologies, spatial data is available at a global scale[1]. How to organize and manage the global-scale spatial data effectively has been a hotspot in ESS. As a type of space frame, grid is a frequently-used tool for organizing, analyzing, expressing, simulating, sharing and visualizing spatial data in many fields such as geography, geology and meteorology. According to different spatial divisions, there are two types of grid, one is Planar Spatial Grid (PSG)[2] and the other is Global Spatial Grid (GSG)[3,4]. While PSG is divided in Euclidean space including 2-D PSG and 3-D PSG, GSG is divided in sphere (spheroid) space including sphere GSG and spheroid GSG. Due to the disunity of map projection, traditional methods of displaying spatial data have been limited to a local scale, which are not suitable for global-scale data. SDOG[5] is a newly developed GSG in the intrinsic space of the Earth (i.e., spheroid) without using any map projections, and is able to represent global-scale spatial data. SDOG is now a good implementation of the ESSG[6] which has been adopted as the reference framework of global dataset in GEO working plan in 2012-2015[7] . It has also been applied to the global-scale lithosphere modelling[8] and meteorological modelling[9]. SDOG has great advantages in a) latitude-longitude consistent, orthogonal, approximate grid size and non-overlapping; b) multi-hierarchy, multi-resolution and unique code; c) all data related with any a grid can be considered as a triple (C, T, A) [10]. However, the number of triples for a global dataset will be raised dramatically as the grid resolution increasing which will beyond the ability of data management in present software and hardware environments, especially for the data retrieval and the I/O. How to enhance the efficiency of data accessing is a key issue for global spatial data management by SDOG-based ESSG. Dividing the dataset into blocks with numerical index is found out to be an efficient way to manage massive dataset [10, 11]. This paper is intent to manage the global-scale dataset using SDOG triples model which contains blocks division and hierarchical index.

2. METHODS

2.1 GLOBAL-SCALE DATA REPRESENTATION

The ESSG has excellent characteristics[6,10] in spherical structure, continuous coverage, hierarchical structure, approximate size, definite frame, unique code, multiple granularities, and geographical consistency, which make it to be a common global GSG. It helps to promote interdisciplinary cooperation between the earth system exchange and data sharing.

SDOG based on ESSG (Figure 1a) is a 3-D GSG which takes many advantages in multi-hierarchy, multi-resolution, orthogonality, latitude-longitude consistency, approximate size and non-overlapping. However, it also takes feature in unique code with a method of degenerated-Z curve (DZ-curve) filling (Figure 1b)[14]. Since each grid of SDOG is uniquely coded (Figure 1c) and each code can be transformed to actual coordinates freely by encoding and decoding algorithms, the code can be used as the indication of spatial information.

In order to get a historical record of the dynamic change of the attribute, a triple (C, T, A) is proposed as a new data structure to organize global scale dataset with ESSG[10,11]. It takes the code *(C)* of a grid as the spatial identifier, the time-stamp *(T)* as a time slice, and the attribute values *(A)* related with the grid. Any data sets collected with a time-stamp ineverywhere of the planet Earth, can be interpolated as a set of *C-T-A* values and input to the ESSG at corresponding granularity. If a triple model (*C, T, A*) is used to indicate each grid, any global-scale dataset can be represented by a set of triples (Figure 2).

3. MANAGE GLOBAL DATASET WITH ESSG

The global 3-D coverage of Earth Observation Systems (EOS) provides plenty of earth system data for the ESS and global change researches. How to share these datasets effectively is vital to SDOG-based ESSG. Surface data is one of the important parts of earth system data. This paper is mainly aimed at how to convert the global raster dataset into SDOG-based ESSG data blocks. Because of the extensive source channels of global-scale raster data and the diversity of data format and

doi:10.5194/isprsarchives-XL-4-W2-63-2013

huge amounts of data, how to convert the data into a unified data model is the research emphasis and difficulty. Therefore, the huge data organization and management based on ESSG mainly includes the following two parts: 1) to realize the management of block dataset based on ESSG coding; and 2) to convert the multi-source heterogeneous data into ESSG triples data block (Figure 3).

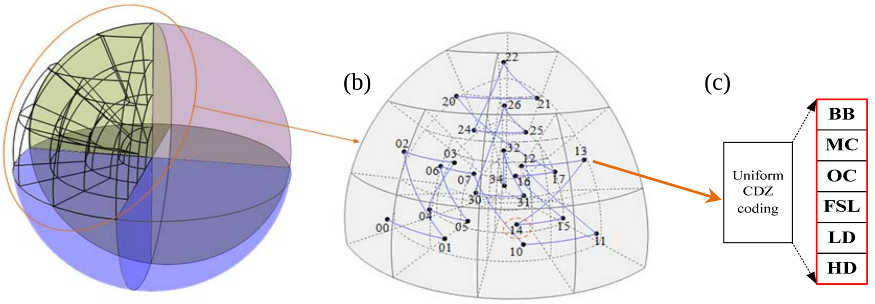

Figure 1 SDOG and its coding scheme

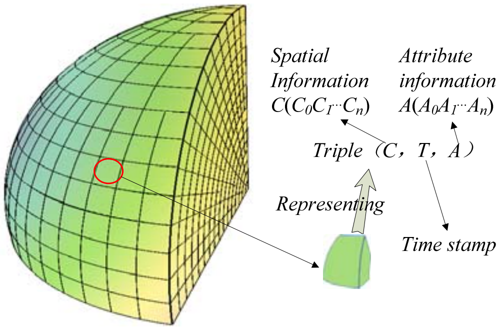
Spatial Information $C(C_0C_1 \cdots C_n)$
Attribute information $A(A_0A_1 \cdots A_n)$
Triple（C，T，A）
Representing
Time stamp

Figure 2 Representing global datasets in SDOG-based triple model

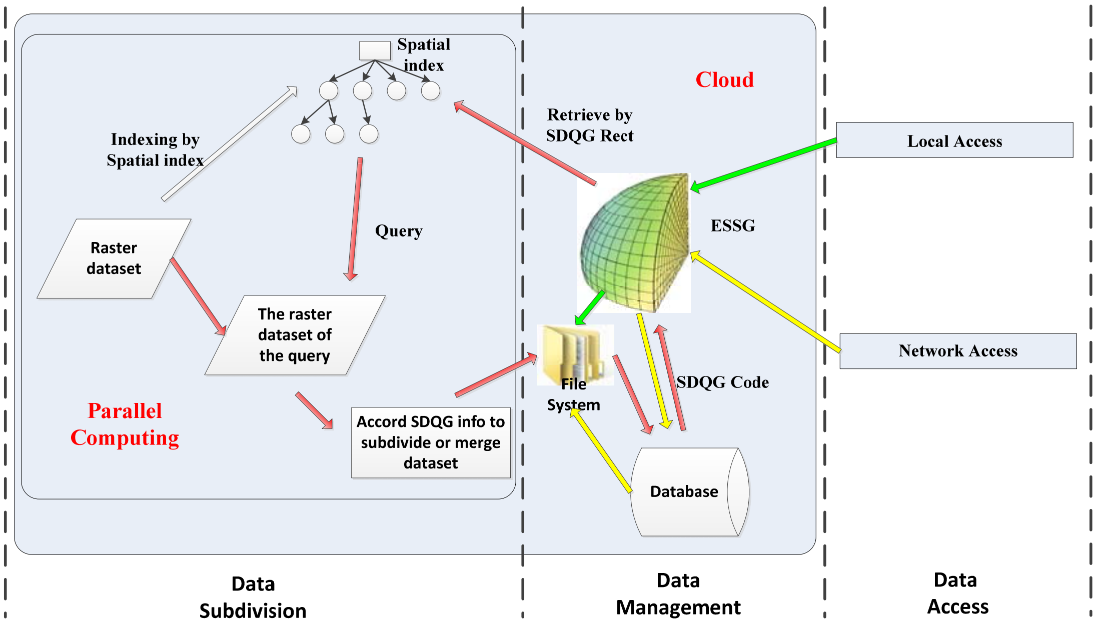

Figure 3 Data management with SDOG-based ESSG

3.1 GLOBAL-SCALE DATA MANAGEMENT

SDOG-based ESSG divides the global earth system space (including the interior, surface and exterior) into 8 octants and subdivided each octant based on SDOG method. If we begin the binary subdivision recursively along the radial dimension based on SDOG until the radial size of grid meets the requirements, then we call the produced ESSG as SDOG-R ESSG[10]. Otherwise, if we replaced the radial direction as the spherical dimension, then we call the produced ESSG as SDOG-S ESSG[10]. Together they constitute the SDOG–based ESSG family(Figure4).

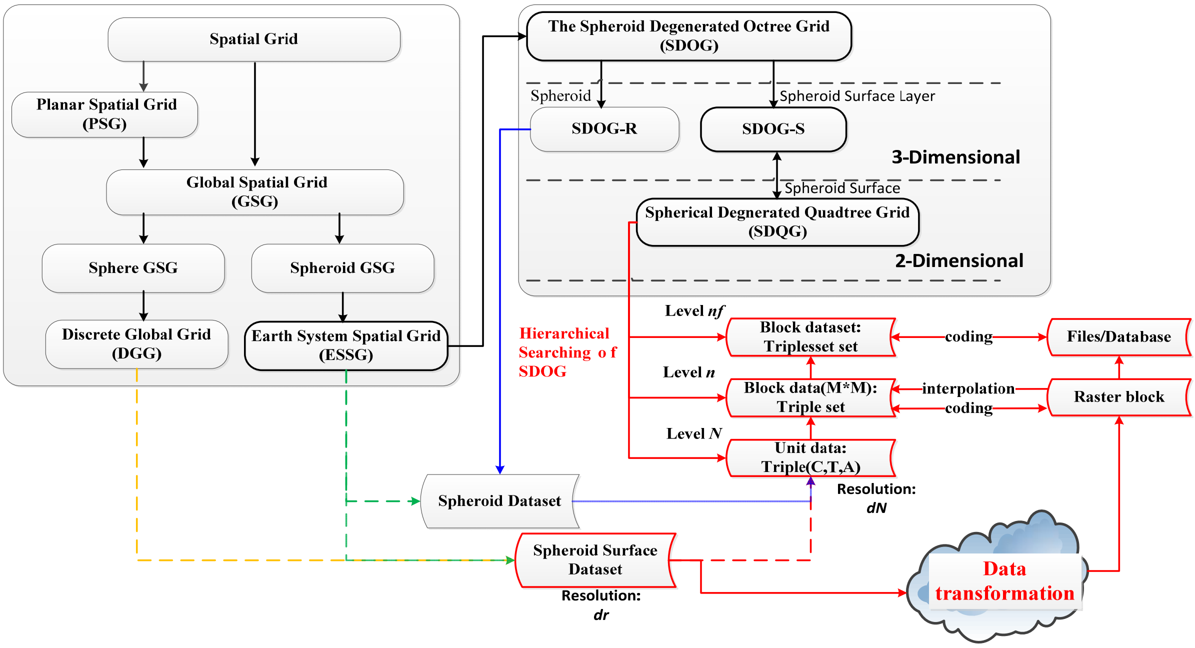

Fig 4 Data organization based on SDOG

We organize global datasets in SDOG-based ESSG with a triple (C, T, A). There are two parts for SDOG-based ESSG management for global spatial datasets (Figure 4): 1) SDOG-R ESSG manages the interior and exterior space dataset, and 2) SDOG-S ESSG manages the surface dataset. This paper is focus on the organizing and management of global scale surface datasets (Figure 4 red part).

A global dataset (e.g., global remote sensing image, global terrain, global air temperature field) with 10m resolution will take 1,466,015,503,704 triples (SDOG-S) at least. It obviously will be far beyond the ability of existing software when quickly accessing such a global dataset if no valid measurement is taken. As known from the grid subdivision, SDOG is a highly hierarchical tree, where a grid at *SL*=*i*-1 will completely cover the sub-grids at *SL*=*i*. Thus, we can take the full advantage of this characteristic, resample the global data at a lower SL of grids, and then to group these grids of representation into blocks, which are in fact the girds at a higher *SL* (Figure 5). For example, we could choose an SDOG at *SL*=*n* whose resolution is closed to the dataset as the resample grid to build the triples, and then choose an SDOG at *SL*=*m* where *n*<*m* as the grid to build the index. Consequently, each index block will contain $2^{n-m}\times2^{n-m}\times2^{n-m}$ triples at most when the third dimension is considered, and no more than $2^{n-m}\times2^{n-m}$ triples if the third dimension is ignored. To make the index more simple, we adopted a Spherical Degenerated Quadtree Grid (SDQG)[15], which is the surface part of SDOG-S. SDQG is actually a 2-D version of SDOG-S. When the grids of SDOG were projected to the surface of spheroid, they turn out to be the grids of SDQG (Figure 5).

With regard to determining the subdivision level-*n* of grid data, 3-D subdivision method for SDOG can be converted to 2-D method, and sub-blocks dataset which under SDQG subdivided are one-to-one correspondence to the cells in grid of level-n^{th} subdivision by grid code.

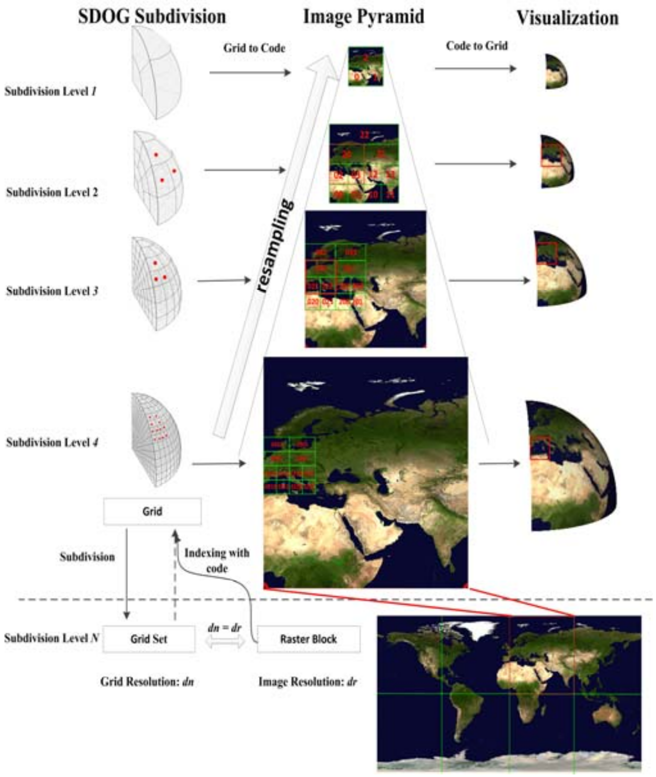

Figure 5 Multi-level global image organization with ESSG

Key issues of managing the global dataset based on SDQG with blocks-index includes: (1) how to divide the large scale raster data; (2) how to define the rule for naming the sub-block of raster data; and (3) how to connect the sub-block to the grid coding.

1) With the increasing of subdivision levels, the spatial resolution of related grid unit will get much higher. According to the resolution of the raster data, we calculate the grid subdivision level firstly, and then to decide the subdivision level of the raster data based on the grid's level (Figure 6A).

2) On account of determining the subdivision level of grid data, 3-D subdivision method for SDOG-S can be converted to 2-D method (SDQG) and sub-blocks which under subdivided are one-to-one correspondence to the cells in grid of subdivision level- *n*.

3) Because of the code uniqueness of SDQG-based ESSG and the ability of reciprocal transformation between the code and space coordinates, we name the sub-block data by using SDQG code of the level-n^{th}. Efficient retrieval of raster data block (Figure 6C) can be implemented by the transformation of code and coordinates (Figure 6B).

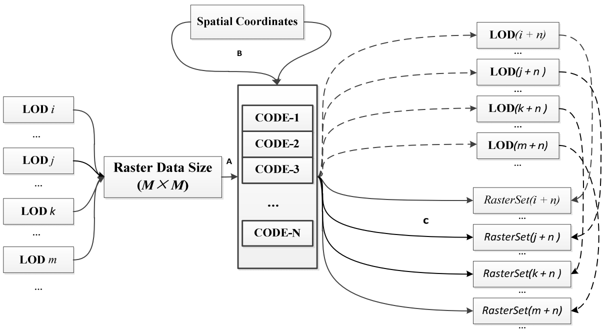

Figure 6 The management of raster data

3.1.1 Calculation of subdivision level (child level)

It is necessary to calculate the subdivision level of SDQG-based ESSG for dividing the source dataset into blocks. There is one problem that the spatial resolution of source dataset is usually inconsistent with SDQG grids' resolution. As a result, we calculate the subdivision level of the grid firstly according to the resolution of raster data, then to decide the subdivision level of the raster data based on the grid's level (Figure 6A). The details are illustrated as follows:

1) The relationship between the division level on the surface of sphere and the raster data resolution is shown as equation 1.

$$d_N = \frac{2\pi\times R}{2^N} \quad (1)$$

Where the d_N is spherical resolution of grid at subdivision level *N*, *R* stands for the radius of sphere .

2) The resolution of the division raster data is d_r,and choose the division level based on the rules: if $d_{N-1} > d_r > d_N$ then choose the division level *N* for re-gridding the raster data.

SDQG-based ESSG is a highly hierarchical grid system, which can act as the foundation of multi-lever indexing, multi-scale modelling and visualization simultaneously. Therefore, the father grid contains completely its son grid. It is feasible to use the size of a grid block (*M*×*M*) to calculate the subdivision level (*n*) of father grid, i.e., every grid in level-*n* is correspond to a *M*×*M* sized data block.

1) Assuming that the size of the block is *M*×*M*. if the grid of SDQG at subdivision level-*n* will be subdivided into *M*×*M* son grids at subdivision level-*N*, then dividing the high resolution raster data into blocks at subdivision level-*n*. The calculation of *n* is shown as equation 3. The equation 2 is to calculate the number of spherical grid.

$$NGs_i = 8\ \times \left(1 + \frac{\left(4^{(i-1)}-1\right)\times 2}{4-1}\right) (i > 0,\ i \text{ is subdivision level}) \quad (2)$$

$$\mathrm{n}\ \approx\ \frac{NGs_N}{M\times M} \quad (3)$$

2) Confirming the division level as *n* through the equation 2 and 3, where the *n* is smaller than *N*.

3.1.2 Calculation of subdivision level (father level)

The volume of high spatial resolution data will certainly be very huge. However, the Input/Output (IO) access is limited. These will lead to a low efficiency of data querying and updating in consideration of unified storage and management with SDQG-based ESSG. Hence, it is necessary to store the datasets separately to improve the efficiency of data accessing and updating. Since the highly hierarchical characteristic of SDQG-based ESSG, it is available to traverse all the children grids directly based on the grid code by using a search algorithm, as well as the father grids. In a word, storing datasets separately can be achieved by subdivision dataset based on the hierarchical character of SDQG-based ESSG. The procedures are as follows:

1) Calculating the subdivision level *n* of data block according to equation 2 and 3.

2) If the number of dataset contains data blocks is no more than *sn*， then calculate the dataset at level-*nf.*

3) Calculating each grid C_i at the level *nf* contains data blocks set$C_i\{C_{ni}, C_{ni+1}, \ldots\}$ at level-*n*

4) Storing the setC_i marked by the code of grid C_i.

The first level above is to realize the data storage separately to improve the dada management efficiency based on SDQG-based ESSG. The second level is to realize the dataset division to further improve the efficiency of data IO access and data updating.

3.1.3 Data retrieval

Each grid code of SDOG-based ESSG has the corresponding space scope, and the level index of the grid code is mainly based on its spatial encoding mechanism. The interaction index between father and children grid can be implemented by SDQG-based ESSG hierarchy search, i.e, by using the father grid coding (space scope) to retrieve the children grid coding (space scope), and vice versa.

Hierarchical searching based on hierarchical SDOG-based ESSG can be expressed by using the general form: $\mathrm{SDOG}_{\mathrm{ESSG}(p+x,q+y)} = \mathrm{g}_{\mathrm{x,y}}\left(\mathrm{SDOG}_{\mathrm{ESSG(p,q)}}\right)$. SDOG-S-ESSG*(p, f+1)* can obtained from the grid of SDOG-S-ESSG*(p,f)* by 1 time degraded quadtree subdivision. Consequently, spherical hierarchical searching of SDOD-S ESSG(*p*, *f*) only needs to change the *LD* and *FSL* of grid code. Set C (*p*, *f*) as a random grid code of SDOD-S ESSG, thus $g_{0,1}(\mathrm{C}(p,f))$ will has 3 or 4 grid code by all. The calculation method is as follows:

1) Judge the type of grid unit according to $\mathrm{C}(\mathrm{p},\mathrm{f})$ and work out FSL, LD and the others inversely. The Types of grid[10] unit are NG(Normal grid), LG(Latitude degenerated Grid), and SG(Sphere degenerated grid).

2) If grid unit is NG, then make FSL_1=FSL_2=FSL_3=FSL_4=FSL+1. Furthermore, respectively add the binary of 0, 1, 2, and 3 in the end of binary code of LD in

order to get LD$_1$, LD$_2$, LD$_3$ and LD$_4$. Then build the grid code respectively based on the combination of FSL$_i$ and LD$_i$ to get the 4 grid code of $g_{0,1}(\mathrm{C}(p,f))$.

3) If grid unit is LG or SG, then make FSL_1=FSL_2=FSL_3=FSL+1. and respectively add the binary of 0, 1and 2 in the end of binary code of LD in order to get LD$_1$, LD$_2$ and LD$_3$. Then build the grid code respectively based on the combination of FSL_i and LD_i to get the 3 grid code of $g_{0,1}(\mathrm{C}(p,f))$.

Since the level-*n* data block and the level-*nf* datasets of SDQG-based ESSG are all named after the SDQG code, coupled with the SDQG subdivision level mentioned above, we can come to a conclusion that source data index can be achieved by the firstly calculation the level-*n* of SDQG and the dataset can be retrieval by the level-*nf*. The main procedure includes:

1) Input the region of retrieval rater data which named as *Rect* (space coordinates are required);

2) Transform the space coordinates into code of the level- *nf* grid of SDQG;

3) Calculate *Rect* contains Codes *setC*$_{\mathrm{nf}}${*C1*,*C2*, ...} at level-*nf*;

4) Get all code *C*{ c_1, c_2, ...} at level-*n* within the range of *set*C$_{nf}$ calculated;

5) Get data blocks from database/file system when name of data block in *C*.

6) Look up for the correspondent raster data block *R*{r_{c1}, r_{c2} , ...} according to the code.

3.2 GLOBAL SCALE DATA TRANSFORMATION

There are a lot of methods to divide the global raster data, such as the latitude/longitude method. The most effective way for searching the global raster data rapidly is to use the coordinate information of these datasets to establish spatial index. Spatial index are used by spatial databases to optimize spatial queries. Usual spatial index methods include: Quadtree, Octree, R tree, etc. It is the code information recorded in the triple block model of SDOG-based ESSG, and the code can be converted into spatial range. Hence, the SDQG-based ESSG retrieval pattern also belongs to the space retrieval. Therefore, it is a conversion between different spatial indexes and a process of cutting and stitching the corresponding raster data of spatial index unit (Figure 7). The procedures to convert raster data into SDQG-based ESSG triple blocks are as follows:

1) According to the coordinate information from data blocks of the source data to build a spatial index *SI*;

2) Get the SDQG subdivision level-*n* according to the equation 3;

3) Calculate the corresponding space range of all the codes in level-*n*, the range sets is *setRec*$_n${rc_{n1}, rc_{n2}, rc_{n3},...};

4) Retrieval each scope rc_{ni} of *setRec*$_n$ in *SI* and the intersecting data block set *setRas*$_i$ {*Ras*$_1$, *Ras*$_2$, ...};

5) Cutting (suturing) the data set *setRas*$_i$ to raster data block *cr*$_i$;

6) Resampling (interpolation)the raster data *cr*$_i$ to a triple block *Tbk*$_i$;

7) Store *cr*$_i$ and *Tbk*$_i$;

8) Repeat procedure 4 to 7 until the end.

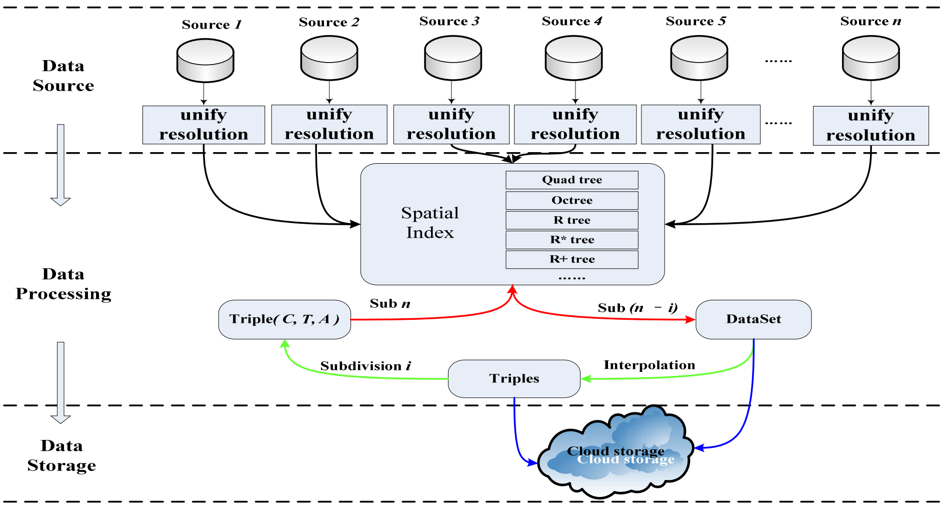

Fig 7 Convert Multi-source Global Data into ESSG-based triple(C,T,A)

The serial computation mention above can converte source data into triples block set. However, it can not satisfy the requirements on global scale and high performance. In order to solve this problem, we neede to develop a parallel algorithm, which can make full use of the computing resources of new hardware architectures. Each code corresponding to space range in the same SDQG subdivision level is independent, which is not influenced by some other codes corresponding to space range in processing by algorithms. So, SDOG-based ESSG is suitable for parallel computing. This paper used Master-Slave parallel strategy to design parallel algorithms. The parallel strategy to convert raster data into SDQG-based ESSG triple blocks are as follows:

1) Master-process *mp* use the coordinate information of the source dataset to build a spatial index *SI*;
2) *mp* get the SDQG subdivision level-*n* with the equation 3;

3) *mp* calculate the corresponding space range of all the codes in level-*n*, the range set is $setRec_n\{rc_{n1}, rc_{n2}, rc_{n3,...}\}$;

4) *mp* average distribute $setRec_n$ into Slave-processes, the range set of in each Slave-process $sp_j(0 < j \leqslant m,\ m$ is number of process) is $subRec_i\{ rc_{n1}, ...\}(0 < i \leqslant m)$;

5) sp_j retrieves each scope rc_{ni} of $subRec_i$ in *SI* and gets intersecting data block set $setRas_i$ $\{Ras_1, Ras_2, ...\}$;

6) sp_j cuts (sutures) the data set $setRas_i$ to raster data block cr_i;

7) sp_j calculates the raster data cr_i into a triple block Tbk_i by resampling (interpolation);

8) sp_j stores cr_i and Tbk_i;

9) Each *sp* repeats procedure 5 to 8 until the end.

4. EXPERIMENTS

All tested algorithms were executed on a server with an intel 2.0GHz CPU (each CPU has 16 processors) and 48GB memory.

An experiment on global dataset of terrain was carried on to test the performance of the management of block dataset based on ESSG coding. In the experiment, SDOG-S(20) (about 1,466,015,503,704 triples) was used to represent the global-scale dataset. Every $2^8 \times 2^8$ sub-grids were then united to one block, which turns to a grid in SDOG-S(11). Thus, SDOG-S(11) (about 5,592,408 blocks) was used to build the index for blocks. Results show that it costs nearly 175 seconds for the construction of all triples and index, and 1.3 seconds is consumed for retrieving 10 blocks and write into files.

Another experiment was to test the performance of converting the source data into ESSG triples data block. we transferred global 1 kilometre DEM (Digital Elevation Model) dataset into SDQG-based ESSG data blocks (19,804 blocks, 1.73GB) by using parallel algorithm is conducted. The experiment shows that 1 process consumed 1233s, and 8 processes consumed 197s. The speedup ratio can reach to 6.3 (Figure 8) when the process number is 8, and the parallel efficiency stable at more than 75% (Figure 9).

5. CONCLUSIONS

This paper employed multi-blocks and hierarchical index strategies to manage the global-scale dataset with SDOG-based ESSG triple model. The methods of blocks coding and index building was introduced, and an experiment was carried out to test the performance of data management based on SDOG based ESSG. Result shows that it has greatly improved the performance of data management and data transformation by the way used in this paper. It offers a reference for the exploring research on parallel computing based on SDOG-ESSG by the parallel technology.

SDOG-based ESSG can integrate all data sets on planet Earth. In this paper, we just have organized and managed the data related to the surface of the earth. Other papers will focus on the 3-D data organization and management, such as global crust data and atmosphere temperature field data.

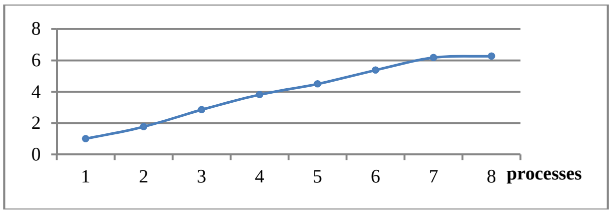

Figure 8 The speedup of the algorithm

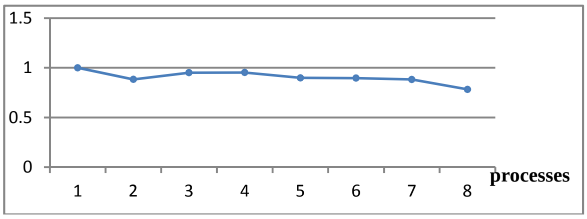

Figure 9 The efficiency of the parallel algorithm

Acknowledgement

This research is jointly sponsored by the National Natural Science Foundation (No. 41001228) and the National 863 High-Tech Program of China (No. 2011AA120302).

REFERENCE

1. LI Deren, GONG Jianya, SHAO Zhenfeng, 2010. From Digital EarthtoSmart Earth, Geomatics and Information Science of Wuhan University, 2, pp. 127-132.

2. Kiester, A. R., K. Sahr, 2008. Planar and spherical hierarchical, multi-resolution cellular automata. Computers, Environment and Urban Systems, 32(3), pp. 204-213.

3. GoodChild, M. F., 2000. Discrete Global Grids for Digital Earth. Internal conference on Discrete Global Grid, Santa babara.

4. Sahr, K., D., 2003. White,A. J. Kimerling. Geodesic Discrete Global Grid Systems. Cartography and Geographic Information Science, 30(2), pp. 121-134.

5. Lixin Wu, Jieqing Yu. 2009. A New Digital Earth Reference Model: Spheroid-based 3D Grid for Earth System (3DGES). The 6-th International Symposium on Digital Earth.September,9-12, Beijing, China.

6. WU Lixin, YU Jieqing, 2012. Earth System Spatial Grid and Its Application Modes, Geography and Geo-Infomation Science, 28(1), pp.14-18.

7. http://www.earthobservations.org/geoss_imp.php

8. YU Jieqing, WU Lixin, ZI GuoJie, et al., 2012. SDOG-based multi-scale 3D modeling and visualization on global lithosphere, Science China Earth Science, 55(6), pp. 1012-1020.

9. LI Zhifeng, WU Lixin, BO Haiguang, et al., 2012. VisIt-based Parallel Visualization of Global Scientific Data: Atmosphere Temperature Field Being a Case, Geography and Geo-Infomation Science, 28(1), pp. 24-28.

10. Yu J.Q., 2012. SDOG-based Earth System Spatial Grid and Its Application and Three Dimensional Modeling. Doctoral dissertation of Beijing Normal University, pp. 103.

11. Yu J.Q., Wu L.X., Li, Z.F., et al., 2012. A SDOG-based intrinsic method for three-dimensional modelling of large-scale spatial objects. Annals of GIS, 18(4), pp. 267-278

12. Beck, A., 2006. Google Earth and World Wind: remote sensing for the masses. Antiquity, 80(308). http://antiquity.ac.uk/ProjGall/beck/index.html.

13. Yoon S E, Salomon B, Gayle R, et al. Quick-VDR: Out-of-core view-dependent rendering of gigantic models[J]. Visualization and Computer Graphics, IEEE Transactions on, 2005, 11(4), pp. 369-382.

14. YU Jie-qing, WU Li-xin. On Coding and Decoding for Sphere Degenerated-Octree Grid. Geography and Geo-information Science.2009, 25(1), pp. 5-9.

15. CUI Ma-jun, ZHAO Xue-sheng. Tessellation and Distortion Analysis Based on Spherical DQG. Geography and Geo-information Science. 2007, 23(6), pp. 23-25.