Research on the typical parallel processing algorithm of massive remote sensing data

Yang Jinlin [1*], Fan Dandan [2] and Xu Xiaoshen [2]

1 Beijing Aerospace Titan Polytron Technologies Inc，Beijing，China

2 Key Laboratory of Aerospace Broadband Network Technology，Beijing，China

Huairou，Beijing，box number 3380，156,101416

yangjun_801222@126.com

Keywords: Big data. parallel processing. OpenMP.

Abstract. With the rapid development of the earth observation technology, remote sensing image data is exponential growth, which has a huge challenge to the existing remote sensing processing system. Because of the limitation of single machine processing capability, the processing speed of single machine algorithm has been greatly improved. So the high performance parallel clustering technology is used. In this paper, the characteristics of remote sensing image geometric features, complex features and large data processing requirements of the algorithm are discussed. The parallel processing technology is studied. The parallel architecture and parallel computing models are discussed. The relationship between the number of threads and the number of cores is discussed. The experimental results are discussed. The running time of the system is influenced by the number of threads and the hardware.

Setting

With the rapid development of the high resolution observation technology, the data obtained by remote sensing satellite has a wide range of applications because of its fast update speed and large coverage. At the same time, the high resolution satellite launch directly causes the remote sensing image data quantity to grow, to the existing computer system will have the huge challenge [1]. In the current technology background, because of the limitation of the processing capability of the single machine, the processing speed of the single computer algorithm has been very difficult to improve the [2]. Therefore, with the development of national space and satellite applications, the main strategy of the use of high performance parallel clustering technology, for remote sensing image data and processing algorithm characteristics, the study of high performance parallel processing technology and the whole parallel cluster number of threads and hardware, the need for the remote sensing satellite image processing system, the scheme design and system construction is of great significance.

Research on parallel architecture

The production process of remote sensing products involves complicated operation model, complex structure, complex algorithm, complex and related events, multi processor and multi system cooperation. Parallel technology is divided into three parallel modes, such as data line, task parallel and parallel computing, and it uses [3-6].

Parallel data

Data parallelism is in the data processing operations, according to the characteristics of the processing algorithm in accordance with the strategy of automatic or manual data segmentation, the larger size of data segmentation into a number of small data blocks, and from multiple computing

servers to calculate, after the completion of the data in accordance with the principle of data segmentation, to complete the data form returned to the terminal [6].

There are four main ways of dividing data: by block segmentation (the conventional algorithm can be carried out by such segmentation), and the segmentation, the segmentation of the column, irregular segmentation (such as triangular mesh processing, adaptive enhancement, etc.).

Results the merging process is mainly based on the spatial information of the data block, and the results are mainly in two cases:

1. Regular Merger：After treatment, the space position does not change.
2. Irregular Merger：After processing, the space position changes.

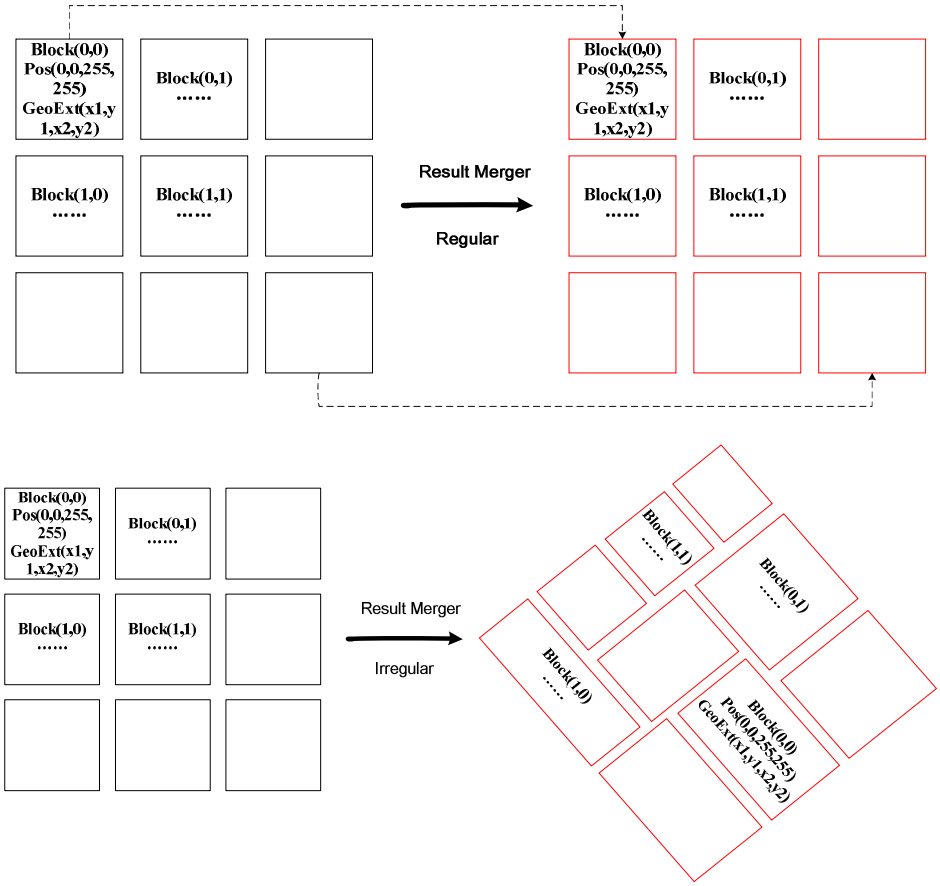

Fig. 1 Data merge sample graph

Task parallelization

Task parallelism in the task refers to the specific remote sensing computing tasks, in the background will be implemented to develop a good scheduling strategy, different computing tasks assigned to different computing tasks, and to build a queue of tasks. In this mode, the number of tasks by adjusting the background to control the size of the task parallel.

Computation parallelization

Computational parallelism refers to the remote sensing algorithm level, that is, an image data in the actual calculation process, will be allocated to multiple processors (including GPU and CPU) concurrent execution. Now the mainstream parallel architecture for GPU-CPU based parallel architecture, the calculation process of the parallel and data parallel.

The following figure is based on the GPU-CPU of the remote sensing algorithm for parallel processing.

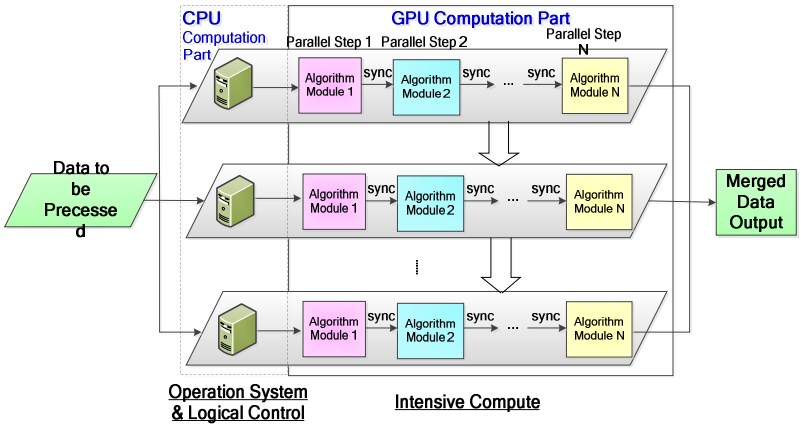

Fig. 2 The parallel processing of remote sensing algorithm based on GPU-CPU

Parallel model of the algorithm flow

In view of the characteristics of GPU and CPU, the parallel Parallel (Data) Raster Processing (RDP), Abstract Model (PAM), is studied.

Parallel mode of the algorithm flow

According to the characteristics of GPU and CPU, GPU computing is based on the computation intensive and a large number of data parallel computing, CPU computing is based on general-purpose computing, which is the logic control of the operating system and instruction, so the whole algorithm can be divided into CPU algorithm and GPU algorithm.

Parallel mode of CPU algorithm

The modern parallel computing platform is built on multi processor cluster platform. Considering the heterogeneous and node computing power, the impact of network congestion on the communication between nodes and memory (CMT) is realized. The task is to split the task into multiple threads. The thread can be divided into multiple threads, and the thread is created, which is based on the number of kernel level threads.

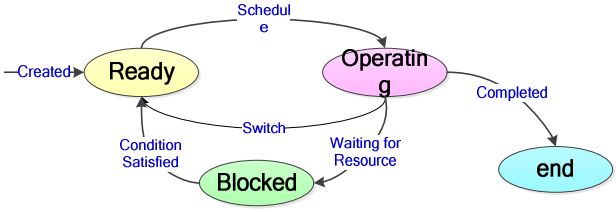

Fig. 3 Thread state transition diagram

GPU processing algorithm parallel mode

The core of the GPU processing algorithm of remote sensing image is the image of the sub block, that is the image of the slice, through the analysis of the parallel mechanism of CUDA, make the following points:

1 . The original image and its auxiliary data copied from the system memory to the memory device of global storage space;

2 . To correct the original image (if the image has been corrected, skip this step), and stored in the memory device of global storage space;

3 . The image of block, and were added to different thread blocks corresponding to the shared memory space, for block after image processing, and processing the data saved to the device memory of global storage space;

4 . Will the final data copied from the memory device to system memory, as the output results are preserved.

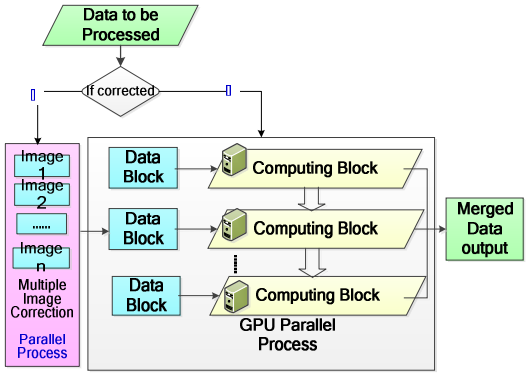

Fig. 4 GPU parallel processing flow chart

Abstract model of parallel processing

The abstract model for grid data parallel processing includes two modules: algorithm flow modeling and time estimation. The function of algorithm process modeling is to model the process of a given grid data processing algorithm, and the feature of the algorithm is based on the same model.

Algorithm process modeling

Refer to Heather Betel [7] (Cellular Cellspace) is defined as a Transition ($Tran_i$,Where i represents the sequence number), which is defined as a CS (Rule), which is defined by the, which is defined as a CA, which is defined as the evolution rule of a cell in the cellular space. The transition (Automata) is defined as a finite element:

$$CA = \{CS, \underset{i\in[1,n_t]}{Arr}\langle Tran_i\rangle, \underset{i\in[1,n_t]}{Arr}\langle R_i\rangle\}$$

Among them, Arr represents a sequence of one-dimensional linked list,transition is a list of elements, n_t Represents the total number of transition.

On the basis of the theory of cellular automaton, the high performance parallel cluster remote sensing element is made by using the remote sensing element of the cluster (Parallel Remote-Sensing Cellular Automata , PRSCA) . Compared to the original model, PRSCA made the following three points to change:

(1) Using CS to describe the image spatial data set, where each element is uniquely corresponding to a real place in a reference coordinate system of a geographic space.

(2) For PRSCA based on data parallel idea, the cellular space will be divided into a number of sub cellular spaces according to a data partitioning method. (Sub-Cellspace , Scs_j , Where j represents the only number of cells in each of the sub cells.)

(3) In order to realize the parallel evolution computation, all of them need to be allocated to a certain task scheduling strategy (Task Scheduling , $TS(n_p)$, which n_p represents the number of parallel computing units) to allocate at least one parallel computing unit. Each computing unit has at least one evolutionary computation task of Scsj, which is performed at the same time, until the end of the evolutionary computation tasks of all the computing units, which marks the end of CA.

According to the above analysis, the formal description of a PRSCA will increase three elements based on the original three tuple, become:

$$PGCA = \{CS, SRS, DP, TS(n_p), \underset{i\in[1,n_t]}{Arr}\langle Tran_i\rangle, \underset{i\in[1,n_t]}{Arr}\langle R_i\rangle\}$$

In which, the spatial information of the spatial data sets is described, and the essential data partition and task scheduling methods are described.

For most remote sensing image data processing algorithms, the process is to enter a remote sensing image data set, and then traverse the input data set of each point in the input data set $Step_i$.

The final output data set. The calculation process can be decomposed into a series of discrete calculation steps.

According to the above analysis, similar to CA, an image data processing algorithm flow can be represented as a three tuple:

$$RDP = \{RDs, Arr_{ie[1,n]} < Step_i >, Arr_{ie[1,n]} < Eq_i >\}$$

In which, Arr represents a sequence of one-dimensional chain, is a list of elements, n_t represents the length of the list, that is, the total number of steps after the decomposition of the algorithm, the calculation equation is used to describe how to calculate the next result values of each data point.

For a remote sensing image data processing algorithm in parallel cluster, its calculation process can be just a step, or it can be decomposed into a number of steps.

In the process of remote sensing image data parallel processing, the remote sensing image data set is divided into a number of sub data sets (Sub-Dataset，Sds_j，which j represents the unique number of each sub data set.) Then all sub data sets are assigned to a task scheduling strategy (Task Scheduling，$TS(n_p)$，which n_p represents the number of parallel computing units) to be allocated at most parallel computing units.

According to the above analysis, referring to the parallel remote sensing data of the automaton's derivative process, a raster data parallel processing algorithm flow is described in the form of the original serial algorithm based on the three elements of two elements, become:

$$GDP^2 = \{RDs, DP, TS, Arr_{ie[1,n]} < Step_i >, Arr_{ie[1,n]} < Eq_i >\}$$

Among them, the essential data partition and task scheduling in the parallel algorithms are described.

According to the above analysis, the process of a data parallel processing algorithm of remote sensing data can be described by a PRSCA, which can be modeled by the PRSCA based on the map of the spatial grid data parallel processing algorithm.

$$GDP^2 \rightarrow PGCA$$

$$via \begin{cases} GDs \rightarrow CS + SRS \\ Step \rightarrow Trans \\ Eq \rightarrow R \end{cases}$$

The role of algorithm process modeling is to describe the remote sensing image data parallel processing algorithm, which is based on PRSCA theory and model, and from another point of view, the design and implementation of model is based on CA model.

Time estimation

Another core of the parallel processing abstract model is the optimization of time, which reveals the serial / parallel time cost of the remote sensing image data processing algorithm.

Referring to the algorithm flow modeling method, the time cost (t_{seq}) estimation equation for the serial processing algorithm of the geo spatial grid data is presented:

$$t_{seq} = t_{ipt} + t_{evls} + t_{ept}$$

Where, t_{ipt} and t_{ept}, respectively, to import and export data to import and export CS time, t_{evls} expressed the total time for the implementation of the evolution of the CS within the cell.

Assuming that the proposed algorithm can be decomposed into a k step, corresponding to transition k, then the time cost equation can be further deformed:

$$t_{seq} = t_{ipt} + \sum_{i=1}^{k} t_{evl,i} + t_{ept}$$

Among them, $t_{evl,i}$ said the time used for the evolution of the CS implementation of the transition-i.

The essence of parallelization is to decompose a large amount of computing tasks into a number of independent sub tasks that can be performed simultaneously by multiple computing units, which are not affected and restricted by each other.

However, not all of the sub tasks of data generation can be implemented independently, and they have no effect on each other. In this case, the time cost of the parallel algorithm is also introduced to the CS, which is a kind of CS:

$$t_{para} = t_{pipt} + \sum_{i=1}^{k} (\max_{j\in[1,n_t]} t_{cmm,i,j} + \max_{j\in[1,n_p]} \sum_{g_n} t_{evls,i,j}) + t_{pept}$$

Among them, $t_{cmm,i,j}$ transition-i on the j implementation of evolutionary computation sub cellular space required communication time consumption, t_{pipt} and t_{pept} are respectively used for data import and export all sub cellular space time, n_p said parallel computing unit number, $t_{evl,i,j}$ represents a to j the chant cellular space execution transition-i evolutionary computation with the time, g_n said no. n calculation unit allocation of sub cellular space set, $\sum_{g_n} t_{evls,i,j}$ said no. n calculation unit for the distribution of all sub cellular space implementation of transition-i evolution which are used to calculate the time and.

Experimental results

In order to verify the effectiveness of the algorithm flow of the parallel model, based on the existing experimental platform, the design of the edge extraction of parallel algorithm experiments. Experimental data using real data from the national elevation data in 2015, the data size is 13000 x 250, the spatial resolution of 21000 meters, the data volume is 520MB.

Software and Hardware Environment

Data test is carried out based on OpenMP. The hardware and software environment are shown in Table 1.

Table 1 Experimental environment of time prediction model of parallel programming model

(a) Hardware Environment

Hardware Environment	Configuration Parameter
Manage/Login Node	x3650M3：2* Intel Xeon 4C X5540 2.4GHz 6* 4GB DDR3 LP RDIMM
I/O Node	x3550M3：2* Intel Xeon 4C X5570 2.80GHz 6* 2GB DDR3 LP RDIMM
Compute Node	Blade HS22：2* Intel Xeon 4C X5570 2.80GHz 6* 2GB DDR3 LP RDIMM
Shared File System	GPFS：2*1T RAID5 GPFS
Network	Infiniband：High-speed 40Gb QDR

(b) Software Environment

Software Environment	Configuration Parameter
Operation System	Red Hat Enterprise Linux 5 Server
Compiler	GCC 4.1.2
Software Base	OpenMPI 1.4.3 GDAL 1.8.0

Test Result

The result of this experiment is based on the edge extraction algorithm of OpenMP, for example, the edge extraction algorithm for OpenMP parallelization, change the number of threads (from 1 threads to 32 threads), test the effect of the number of threads on the performance of parallel, the experimental results are shown in the following figure.

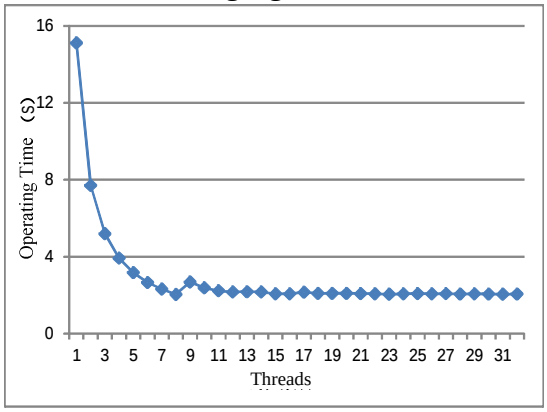

Fig. 5 The effect of thread number on the execution time of OpenMP parallel program

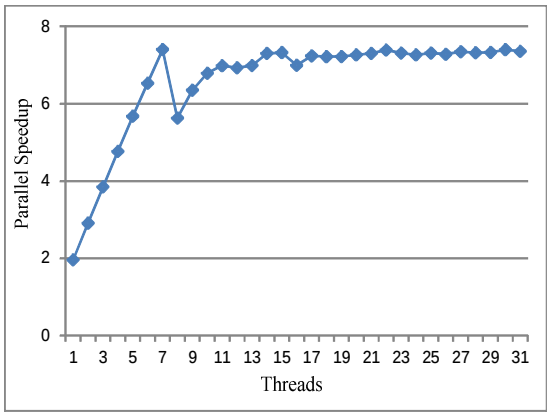

Fig. 6 Effect of speedup on execution time of OpenMP parallel program

Experimental result analysis:

A single node in the experiment is 8 core physical hardware environment, within the 8 threads running in parallel with the time increasing, the number of threads and less, this is because the thread load gradually by the unbalanced equilibrium. In this process, each thread allocation amount of data processed with threads is increased less, therefore, parallel the running time of each thread maximum operation time decreasing, the speedup can approximately achieve linear speedup when the number of threads; each thread reached the nuclear number, load balancing and load minimum, so the running time is minimized. When the thread number is greater than the number of nuclear, add the line number of passes, because the load of each thread and parallel operation time load imbalance. The maximum running time, with the increase in the number of threads, the load will slowly be equal, so the running time decreases, finally tends to be stable, the number of threads and the number of visible nuclear hardware together affect the parallel running time.

Conclusion

I In this paper, the characteristics of remote sensing image geometric features, complex features and large data processing requirements of the algorithm are discussed. The parallel processing technology is analyzed. The parallel computing model and the time evaluation method are discussed.

The parallel experiments are carried out. The experimental results confirm that the running time of the system can be used to construct the parallel processing system.

References

[1] Li Deren, Shao Zhenfeng. China to earth observation satellite and its application [J]. science. 2007 (06)

[2] Chen Chen, Tan Yihua, Li Haitao, Gu Haiyan. Fast parallel algorithm of remote sensing image mosaic [J]. microelectronics and computer (03) 2011.

[3] Chen Guoliang, Sun Ninghui. A popular high performance computer [J]. Journal of University of Science & Technology China, 2008 (07)

[4] Chen Guoliang, Sun Guangzhong, Xu Yun, Miao Kun, Zheng Qilong. The hierarchical parallel computing model [J]. Journal of University of Science & Technology China 2008 (07).

[5] Chen Guoliang,. A parallel programming model and language [J]. software. 2002 (01)

[6] Chen Guoliang, Sun Guangzhong, Xu Yun, Lv Min. Research methodology of parallel algorithms [J]. Chinese Journal. 2008 (09)

[7] Heather Betel,Paola Flocchini. On the Relationship Between Boolean and Fuzzy Cellular Automata[J]. Electronic Notes in Theoretical Computer Science . 2009