GI_Forum 2021, Issue 1
Page: 112 - 119
Best Practice Paper
Corresponding Author:
yx9@hotmail.com
DOI: 10.1553/giscience2021_01_s112

Dynamic Workflow Engine of Atmospheric Big Remote Sensing Data Processing Powered by Heterogenous Platform for High Performance Computing

Sheng Zhang[1], Yong Xue[1] and Xiran Zhou[1]

[1]China University of Mining and Technology

Abstract

The development of big remote sensing data related technologies and applications poses a big challenge that massive computing capability is needed to support big data processing. In order to solve this challenge, this paper proposes an architecture of heterogeneous platform of high performance computing, which employs the computer hardware resources to improve the efficiency of big remote sensing data processing by optimizing scheduling strategies and designing high-performance algorithms. Furthermore, the proposed platform can dynamically incorporated with a workflow engine regarding big remote sensing data processing. These algorithms are modular to meet the flexible combination of different processes.

Keywords: high performance computing (HPC), heterogeneous computing platform, workflow engine, atmospheric big remote sensing data

1 Introduction

Currently, the ability to generate remote sensing data has achieved an unprecedented level. We have entered an era of big remote sensing data. Big remote sensing data are attracting more and more attentions from government officers, commercial investment planers, academic researchers, et at.(Liu et al., 2018).

Based on the requirements of big data technology and application, the methodological framework for multi-scale, long-term, and multi-source atmospheric remote sensing data processing is pressing needed. These framework generally include data preprocessing, spatial processing, denoising, fusion, inversion, classification, interpretation and so on (Ma et al., 2014). These complex, heterogeneous and massive computing tasks could generate a huge cost of time consuming and computing load. The traditional systems for remote sensing data processing can't meet the needs of efficient processing on atmospheric big remote sensing data (Xu et al., 2020).

With the development of parallel computing, distributed computing, grid computing, cluster computing and cloud computing technology, a number of high performance computing

(HPC) for remote sensing data big have emerged, which can significantly improve the efficiency of remote sensing data processing. However, there are still some challenges, including how to make rational using of heterogeneous computing resources, and how to access data and optimize task scheduling to reach the full utilization of computing power (Chi et al., 2016).

In this paper, the distributed computing resources and storage resources in atmospheric big remote sensing data computing are optimized by our proposed dynamic workflow engine, which can improve the logic dependent relation of data in different processing processes. In addition, for the specific needs of atmospheric big remote sensing data processing, we developed a parallel processing algorithm suitable for different GPU hardware. Thirdly, we build an extensible model library for atmospheric big remote sensing data processing.

2 Related Works

For big remote sensing data processing, many researchers proposed high-performance computing method based on GPU, and improved resource scheduling strategy based on workflow.

Jia Liu et al. (2015) proposed two high-performance computing architectures for aerosol optical depth (AOD) retrieval: one is multi-core processor architecture, and the other is GPU architecture, they are all based on OpenMP and CUDA Programming environment. According to the characteristics of orthorectification algorithm of remote sensing image, DAI. Chenguang et al. (2011) proposed a fast GPU-CPU collaborative processing algorithm based on CUDA, which realized the parallel processing of image resampling based on single GPU and multi-GPU. Ma Yan et al. (2015) proposed a parallel processing model for remote sensing image based on GPU, and established a set of parallel programming template which provides a simpler and more effective method for programmers to write parallel remote sensing image processing algorithm. Yang Xue et al. (2018) proposed a general, fast and effective denoising method, which combines Huber function and GPU adaptive partition technology, after analyzing the Markov random field prior model method. This method significantly improves the computational efficiency of processing massive remote sensing images. Wang Z et al. (2011) proposed a method to manage MODIS sensor data processing based on workflow engine, which can configure high-performance computing resources. It reduces the execution cost by using the existing program modules and distributed resources, and finally helps users manage and process a large number of remote sensing data through workflow. Based on Web services and Activiti 5.0 workflow engine, Fang Huang et al. (2020) built a high-performance computing service platform, which reduced the platform discrepancy between different high-performance exchange systems. This platform simplifies the operation of complex geospatial information processing applications in the field of high-performance geographic computing and realizes the efficient processing of massive data.

Overall, big remote sensing data computing and processing has accumulated some research results. However, in these achievements, the high-performance computing for atmospheric big remote sensing data processing is mainly to solve a specific problem, and there are still

deficiencies in modularization, integration ability, distributed scheduling, process optimization, etc., which cannot make full use of high-performance computing resources.

3 System Framework Design

The dynamic workflow customization technology proposed in this paper is based on a big data platform with 5-layer architecture (Fig. 1), which is composed of scheduling layer, hardware layer, data layer, algorithm layer, application layer.

Based on the workflow engine, the scheduling layer can realizes the dynamic customization of remote sensing data processing process, the allocation of data resources, computing resources and storage resources, and solves the problem of dependence between different process data through the communication mechanism between different processing processes, so as to optimize the scheduling strategy. Therefore, the whole gas big remote sensing data processing is driven by the scheduling layer. The scheduling layer includes two key modules: task launcher and task scheduler. According to the workflow in terms of data processing designed by users, we design the task launcher to process each sub-task in parallel or serial order. In additional, we design the task scheduler for deciding the partition scheme of the whole task, to allocate every sub-task to a distributed computing node, and receive processing results.

The hardware layer provides storage and computing resources for the platform, and can dynamically add new computing nodes and new storage devices. Conversely, new storage and computing hardware resources need to be registered through the platform which dynamically monitors the computing resources and storage resources.

The data layer includes two kinds of data resources: remote sensing data and thematic data. Remote sensing data includes original data, process data and result data. Thematic data is used to assist remote sensing data processing, such as administrative boundaries, coordinate transformation parameters, digital high-range model data needed for orthorectification, etc. The data layer can be updated dynamically, and the first addition of data needs to be registered by providing metadata information of scope, type and time.

The algorithm layer is the program library related to atmospheric remote sensing data processing, including reading and writing, format conversion, projection, fusion, correction, splicing, cutting, inversion and other programs. The flexible customization of processing flow is realized through modular program, and other program tools are added dynamically through registration to expand the platform functions. According to the characteristics of data processing, algorithm layer tools support different GPU hardware.

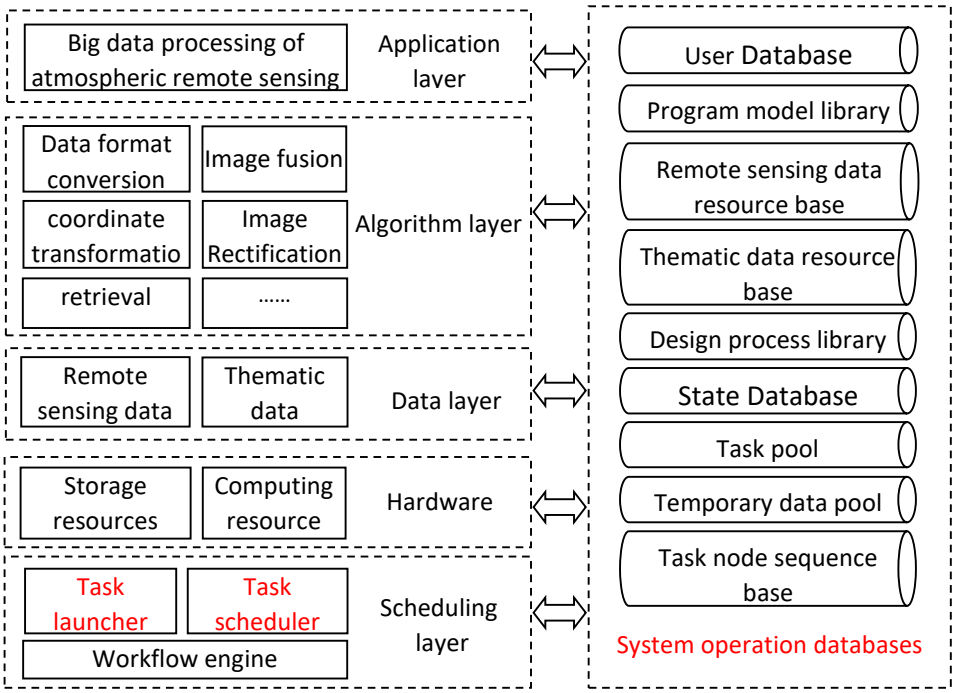

Figure 1: Architecture of the system framework

The application layer is oriented to professional users, customizes the remote sensing data processing flow through the visual window, configures the required data resources, computing resources and storage resources, makes full use of the distributed computing and storage resources, and optimizes the operation strategy by designing the processing flow and setting the interdependence between the processing flows.

The layers in the system mutually transfers information and exchanges data through nine databases. Users can customize and submit their data processing flow through the process designer. Then, the task scheduler automatically completes the task partition strategy according to the computing and storage resources, and distributes the partitined sub-task to each computing node. For each computing node, we design the task initiator to process its assigned task based on users' cunstermized processing flow, and transform the result to the master node. Finally, the task scheduler combines the results from each distributed computing nodes into the final one. Specifically, if the processing breaks unexpectedly due to an abnormal calculation node or unreasonable process design, we create the breakpoint continuation module to ensure the processing would still continue to reaching the completed point.

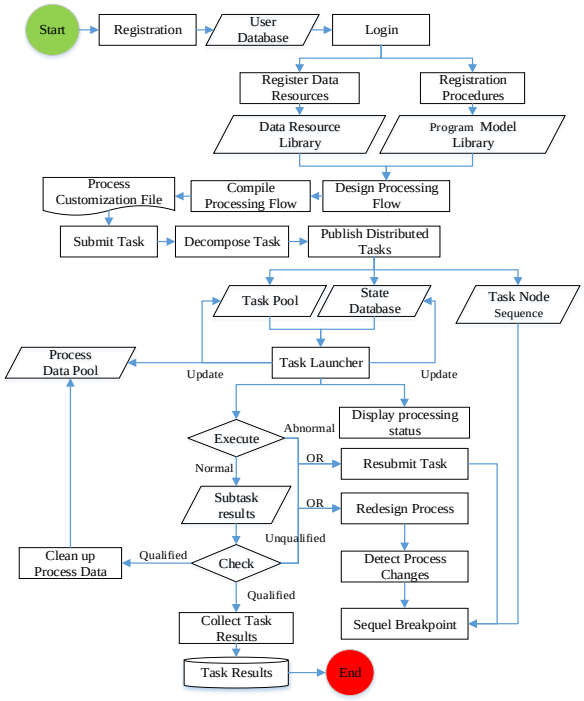

Figure 2: Task workflow procedure

4 Experiments

The experimental environment is shown in Figure 3, which is developed by six personal including one master node (management node) and five general computing nodes.

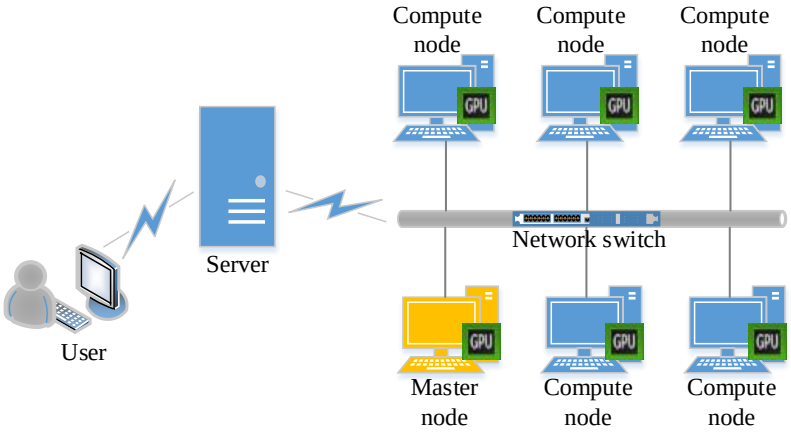

Figure 3: Supporting environment

Moreover, the hardware details of these six nodes are shown in Table 1.

Table 1: Configuration of the testing machines

Hardware category	Master node	Compute node
CPU Processor	Intel Core i7-10700F(2.90GHz ,16 CPUs)	Intel Core i7-10700F(2.90GHz ,16 CPUs)
RAM	16GB	8GB
GPU Processor	NVIDIA GeForce GTX 1660 SUPER(6GB)	NVIDIA GeForce GTX 1660 SUPER(6GB)
Hard disk	4TB(HDD)+1TB(SSD)	1TB(SSD)

The experiment regarding data processing is desgined for the retrieval of satellite-based aerosol optical depth (AOD) data product. The data range is: 35°E -150°E (longitude) and 0°N -60°N (latitude). The satellite data is MODIS, and the data phase is April 9, 2017. We select the SRAP algorithm proposed by Yong Xue et al. (2014) as the AOD retrieval algorithm. The spatial resolution of the AOD retrieval result is 1KM.

Figure 4 shows the workflow designed for AOD retrieval with SRAP algorithm. The input data includes M*D02, M*D03 and M*D04_L2 (* stands for O or Y, which respectively refers to the Terra and Aqua satellite sensor).

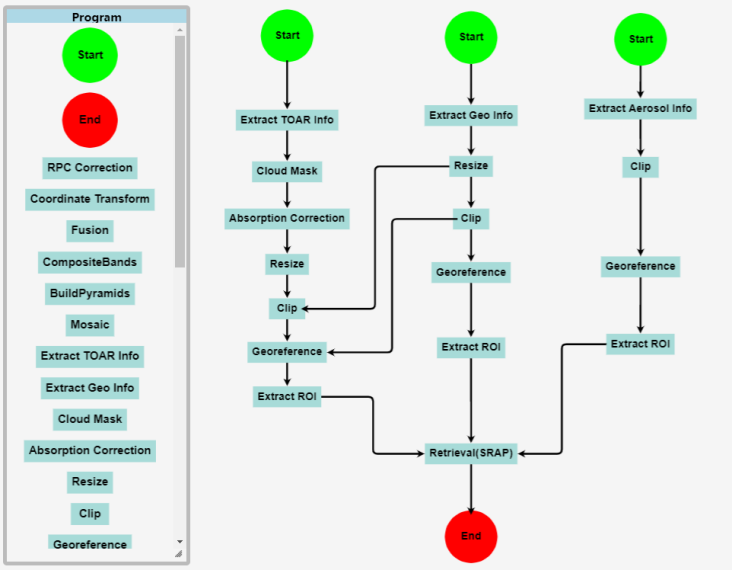

Figure 4: Workflow design of AOD retrieval

Table 2 shows the statistics of processing efficiency, hich compares the method by a single machine processing and the proposed method.

Table 2: Statistics of processing efficiency

Satellite data	Ground resolution	Number	Data size	Processing time of Single machine			Processing time of this paper		
				Prepr-ocessing	Retrieval	Total	Data interaction	Calculation	Total
M*D021KM	1km	61	8.7GB	153min	381 min	576 min	36 min	72 min	108 min
M*D03	1km	61	2GB	5min					
M*D04_L2	10km	61	156MB	37min					

Compared with the traditional atmospheric remote sensing big data computing method, the previous test work proves that the platform proposed in this paper has advantages in the following aspects: Firstly, in the aspect of data processing systematicness, different processing tasks can be customized through this platform, which makes the integration of functions more convenient, systematic and processing content more flexible; Secondly, in terms of task computing efficiency, the overall efficiency is improved about 5 times.

5 Conclusions

In this paper, the computational efficiency of atmospheric big remote sensing data processing was improved by optimizing the scheduling strategy by the dynamic workflow and the parallel algorithm based on GPU. A variety of data processing modules were integrated into a platform to decrease the workload of big data platform, and facilitate the collaboration and communication among researchers in different disciplines. We focused on designing the workflow-driven modular processing and dynamic scaling to allow researchers for customizing their own developed program in the high performance computing.

In the future, the platform could be improved in terms of quality control, fault tolerance and so on. Due to the heterogeneous resources and poor quality of atmospheric remote sensing data, some data processing links might have an issue, resulting in reducing the operating effects of the overall data processing. Therefore, we believe that the improvement of fault tolerance and robustness, and the building of the front-end and process quality monitoring function are significance of developing the platform. We hope our work can facilitate the research regarding heterogeneous computing such as resource expansion, functional expansion, process expansion, etc.

References

Liu, P., Di, L., Du, Q., & Wang, L. (2018). Remote sensing big data: Theory, methods and applications. Remote Sensing, 10(5). doi:10.3390/rs10050711.

Ma, Y., Chen, L., Liu, P., & Lu, K. (2014). Parallel programing templates for remote sensing image processing on gpu architectures: Design and implementation. Computing, 98(1-2), 7-33. doi:10.1007/s00607-014-0392-y.

Xu, C., Du, X., Yan, Z., & Fan, X. (2020). Scienceearth: A big data platform for remote sensing data processing. Remote Sensing, 12(4). doi:10.3390/rs12040607.

Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J., & Zhu, Y. (2016). Big data for remote sensing: Challenges and opportunities. Proceedings of the IEEE, 104(11), 2207-2219. doi:10.1109/jproc.2016.2598228.

Liu, J., Feld, D., Xue, Y., Garcke, J., & Soddemann, T. (2015). Multicore processors and graphics processing unit accelerators for parallel retrieval of aerosol optical depth from satellite data: Implementation, performance, and energy efficiency. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(5), 2306-2317. doi:10.1109/jstars.2015.2438893.

DAI C., & YANG J. (2011). Research on Orthorectification of Remote Sensing Images Using GPU-CPU Cooperative Processing.

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems, 51, 47-60. doi:10.1016/j.future.2014.10.029.

Yang, X., Li, F., Xin, L., Wang, C., Wang, XY., &Chang, X. (2018). Destriping methods for high resolution satellite multispectral remote sensing image based on GPU adaptive partitioning technology. Conference on Remote Sensing for Agriculture, Ecosystems, and Hydrology XX. doi:10.1117/12.2325311.

Wang, Z., King, E., Smith, G., Bellgard, M., Broomhall, M., Chedzey, H., Fearns, P., Garcia, R., Hunter, A., Lynch, M., & Shibeci, D. (2011). RS-YABI: A workflow system for Remote Sensing Processing in AusCover.MSSANZ 19th Biennial Congress on Modelling and Simulation (MODSIM), 1167-1173.

Huang, F., Yang, H., Tao, J., & Zhu, Q. (2020). Universal workflow-based high performance geo-computation service chain platform. Big Earth Data, 4(4), 409-434. doi:10.1080/20964471.2020.1776201.

Levin, N., Ali, S., Crandall, D., & Kark, S. (2019). World heritage in danger: Big data and remote sensing can help protect sites in conflict zones. Global Environmental Change, 55, 97-104. doi:10.1016/j.gloenvcha.2019.02.001.

Xue, Y., He, X., Xu, H., Guang, J., Guo, J., & Mei, L. (2014). China collection 2.0: The aerosol optical depth dataset from the synergetic retrieval of aerosol properties algorithm. Atmospheric Environment, 95, 45-58. doi:10.1016/j.atmosenv.2014.06.019.