RSSN: A High Throughput Geocomputing Platform for Remote Sensing Quantitative Retrieval and it's Task Scheduling Sub-System

Ziqiang Chen[1, 3], Yong Xue[1*], Jing Dong[2, 3], Jia Liu[2, 3], Longli Liu[1,3]

[1]Center for Earth Observation and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China
[2]State Key Laboratory of Remote Sensing Science, Jointly Sponsored by the Institute of Remote Sensing Applications of Chinese Academy of Sciences and Beijing Normal University, Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101, China
Telephone (86-10-64889540)
Email: yxue@irsa.ac.cn
[3]University of the Chinese Academy of Sciences, Beijing 100049, China

1. Introduction

In recent years, the constantly launched satellite sensors bring us vast amounts of Earth observation data. The quality and accuracy of remote sensing instruments have been improved significantly. But it is a rather urgent problem how to handle and analyze the large-scale remote sensing data and to extract valuable information from them timely and efficiently.

Since parallel computers and supercomputers cost too much, and also need for professional maintenance, Grid Computing platform built on the basis of network, which is a scalable virtual unified platform with unlimited computing power and storage capacity, has been one of the most effective solutions to overcome these shortcomings. (Foster and Kesselman 2004, Xue *et al.* 2011a)

This paper introduces our experience in building the RSSN (Remote Sensing Information Service Grid Node) platform, our recent researches in Grid tasks scheduling are also expounded.

2. The system architecture of the RSSN platform

The RSSN platform is a PC cluster for remote sensing quantitative retrieval. It was deployed in the Institute of Remote Sensing Applications, Chinese Academy of Sciences, and used for producing aerosol optical depth (AOD) production covered the land area of Asia. (Xue *et al.* 2011c, 2011d)

In the study of global climate change and other relative fields, large numbers of historical data need to be processed, so we are concerned with how many floating point operations per week or per month we can get, rather than the number of such operations the environment can provide us per second or minute. HTCondor (http://research.cs.wisc.edu/htcondor/description.html) is an open source Grid middle ware, which provides mechanisms to harness wasted CPU power from idle desktop workstations effectively and provide a HTC (High Throughput Computing) environment. In the RSSN platform, we use HTCondor as a resource manager to integrate the idle computing powers which are cheap and easy to obtain. Most of the computing nodes are commodity desktop PC used for our daily work, some of them are even outdated and

* Corresponding author

about to be scrapped. The platform consists of 40 PCs with more than 100 CPU cores totally, as a result, it provide us a considerable computing power for satellite data processing (Chen *et al.* 2012).

When we built the Grid platform, the computers were classified into four levels according to their hardware features, the nodes belong to different level play different roles. The physical architecture is shown in Figure 1. Above the infrastructure layer, there is a B/S-based Grid workflow composition system. Users use workflow composer to submit their tasks through browsers.

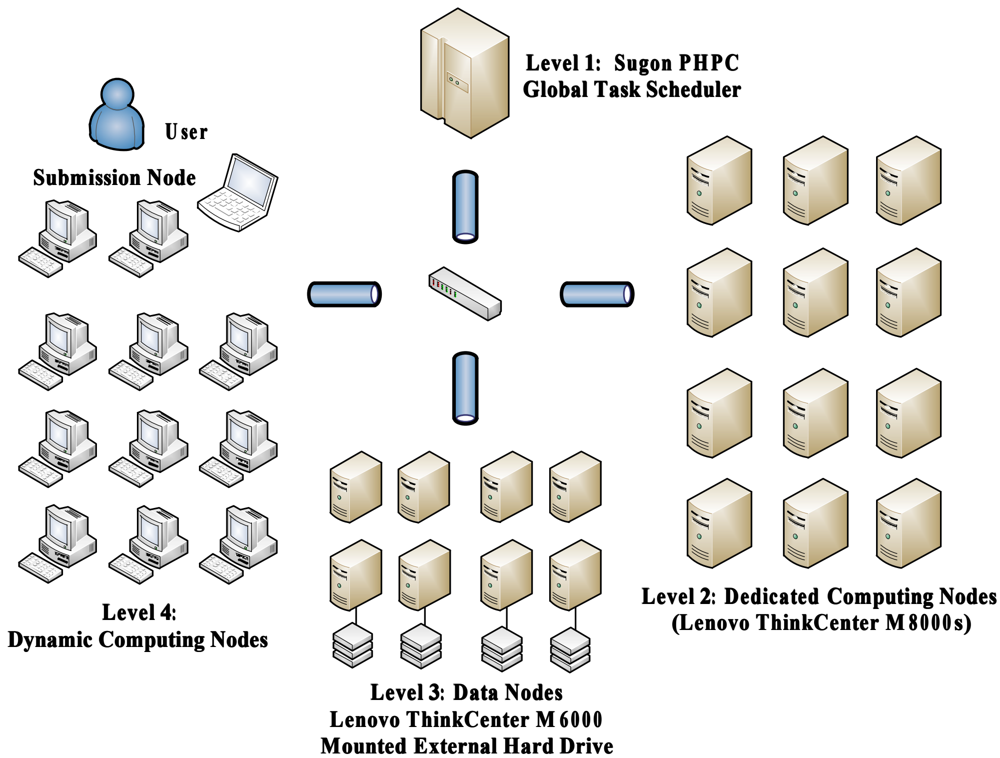

Figure 1. The physical architecture of the RSSN platform.

3. The task scheduling sub-system of the RSSN platform

The RSSN platform is used to generate AOD production over the land area in east of Asia. It gives a perspective on the potential of improving the efficiency of the remote sensing quantitative retrieval process by collecting the computing resources from the unoccupied desktop PCs. In our extensive experiments, we found that the speedup is limited not only by the sequential fraction of programs, but also the data transmission time. Meanwhile, the stability of the system is affected by the computing nodes which join in or quit out from the cluster dynamically. So we designed a new task scheduling sub-system for the RSSN platform to enhance the robustness, fault tolerance and throughput.

3.1 The system architecture of the scheduling system.

We designed the task scheduling sub-system based on the RSSN's three-level framework. The system architecture is shown in Figure 2. The top-layer is remote sensing application layer, there is a workflow composer GUI tool. The middle -layer is remote sensing Grid core components layer. This layer consists of task scheduling engine, workflow parser, user QoS (Quantity of Services) quantify, resource information database, task information database, task dispatcher and so on. The bottom-layer is Grid infrastructure layer. Resource monitor, task monitor and local task manager are running in each computing node.

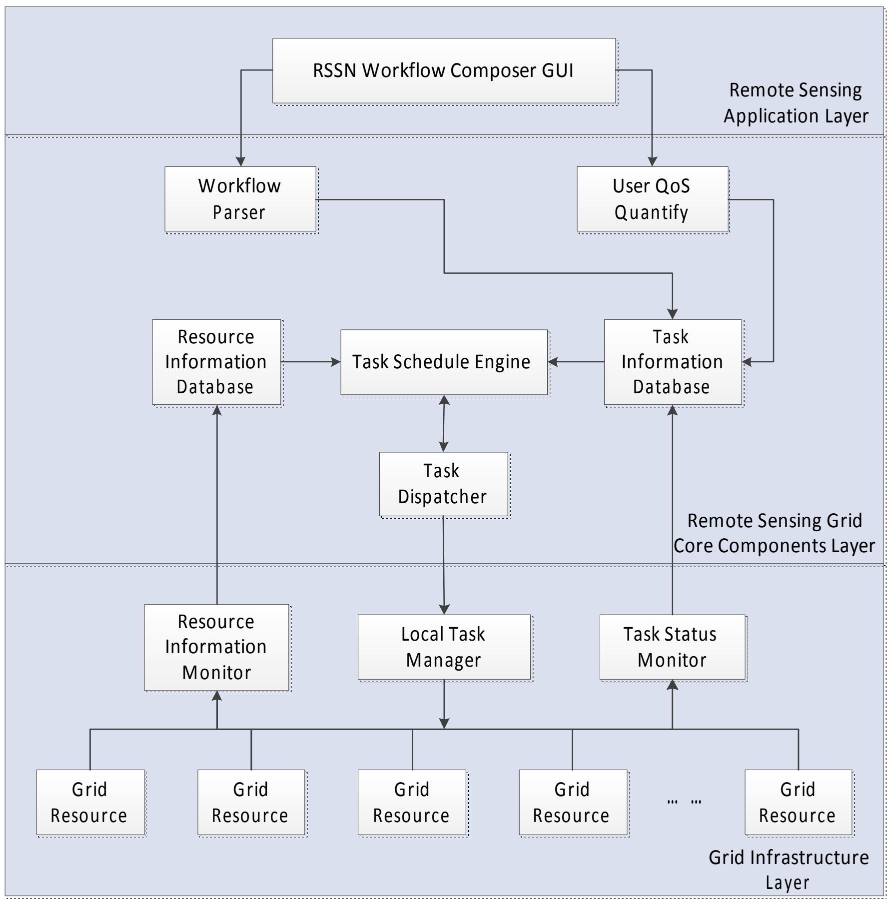

Figure 2. The architecture of the task scheduling sub-system for RSSN platform.

3.2 The scheduling strategy for the Remote Sensing application.

The task scheduling of remote sensing quantitative retrieval included two dimensions: first one is the scheduling between workflow steps which have temporal dependencies, the other one is parallel scheduling of the tasks inside a workflow step which are independent (Wan 2008, Xue *et al*. 2010). But most of the existing scheduling algorithm are not applicable to the dynamic PC clusters which loosely-coupled, or are not applicable to the massive satellite data transfer applications.

In our research, we separate data transfer as independent job just like the computing task. The time-consuming of data transfer is treated as an important factor in the scheduling algorithm. Meanwhile, we take the tasks success ratio and the computing nodes stability into account.

4. Case study and conclusion

Experiments from a realistic application were examined. We use SRAP (Synergetic Retrieval of Aerosol Properties) algorithm (Xue *et al.* 2011b) and the data over the land area of Asia with the resolution at 1km collected by MODIS (or Moderate Resolution Imaging Spectroradiometer) to produce the AOD production. With repeated experiments and analyze, we found the RSSN platform and its scheduling sub-system can be stably operated, the overall performance is also improved in some extend.

5. Acknowledgements

This work was partly supported by the National Natural Science Foundation of China under Grant No. 41271371and the Major International Cooperation and Exchange Project of National Natural Science Foundation of China (Grant No. 41120114001).

6. References

Ian Foster, Carl Kesselman, 2004, *The Grid 2: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Wei Wan, 2008, Study on Grid Enabled Aerosol Quantitative Remote Sensing Retrieval System. *Ph.D. dissertation*, The Graduate University of Chinese Academy of Sciences, Beijing, China.

Yong Xue, Dominic Palmer-Brown, Huadong Guo, 2011a, The use of high-performance and high-throughput computing for the fertilization of digital earth and global change studies. *International Journal of Digital Earth*, 4(3), 185-210

Yong Xue, Hui Xu, Yingjie Li, Leiku Yang, Linlu Mei, Jie Guang, Tingting Hou, Xingwei He, Jing Dong, Ziqing Chen, and Yang Qi, 2011b, Long-term aerosol optical depth datasets over China retrieved from satellite data. *Atmospheric Measurement Techniques Discussions*, 4(6), 6643-6678.

Yong Xue, Jianwen Ai, Wei Wan, Huadong Guo, Yingjie Li, Ying Wang, Jie Guang, Linlu Mei and Hui Xu, 2011c, Grid-enabled high-performance quantitative aerosol retrieval from remotely sensed data. *Computers & Geosciences*, 37(2): 202-206.

Yong Xue, Jianwen Ai, Wei Wan, Yingjie Li, Ying Wang, Jie Guang, Linlu Mei, Hui Xu, Qiang Li and Linyan Bai, 2010, Workload and task management of grid-enabled quantitative aerosol retrieval from remotely-sensed data. *Future Generation Computer Systems*, 26, 590-598.

Yong Xue, Ziqiang Chen, Hui Xu, Jianwen Ai, Shuzheng Jiang, Yingjie Li, Ying Wang, Jie Guang, Linlu Mei, Xijuan Jiao, Xingwei He and Tingting Hou, 2011d, A high throughput geocomputing system for remote sensing quantitative retrieval and a case study. *International Journal of Applied Earth Observation and Geoinformation*, 13(6): 902-911.

Ziqiang Chen, Yong Xue, Jing Dong, Jia Liu, Yingjie Li, 2012, The task scheduling for Remote Sensing Quantitative Retrieval based on hierarchical grid computing platform. In: *proceedings of the 32nd International Geoscience and Remote Sensing Symposium (IGARSS 2012)*, held in Munich, on July 22-27 2012, 487-490.