SUB-WORKFLOW PARALLEL IMPLEMENTATION OF AEROSOL OPTICAL DEPTH RETRIEVAL FROM MODIS DATA CASE ON A GRID PLATFORM

Jia Liu[1,3], Yong Xue[1,2], Vassil Vassilev[2], Xingwei He[1,3]*

[1] Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China
[2] Faculty of Life Sciences and Computing, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK
[3] University of Chinese Academy of Sciences, Beijing 100049, China
{Email: libraspring@163.com; y.xue@londonmet.ac.uk}

ABSTRACT

Aerosol Optical Depth (AOD) is an significant parameter of aerosol optical properties. Operational production of AOD datasets over long time series, large-scale coverage puts on a severe challenge to computing technologies due to both the complexity of retrieval algorithm and the huge data amounts. The Grid computing solution-Remote Sensing Service Node (RSSN) was constructed as a high-throughput platform for remote sensing applications. Taking the sub-workflow level characteristics of some remote sensing retrieval applications into consideration, a sub-workflow parallel implementation for the Synergetic Retrieval of Aerosol Properties (SRAP) algorithm from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor data was taken on the RSSN, and an initial experiment result proved that the sub-workflow parallel could further reduce the runtime of data parallel solutions commonly used.

Index Terms— Aerosol Optical Depth, sub-workflow parallel, Grid computing

1. INTRODUCTION

Global aerosol characterization from satellite remote sensing arouses increasing interest due to the importance of aerosol radiative forcing of climate, its effect on cloud microphysics and albedo, satellite imagery atmospheric correction, etc. [1] Research on multiple retrieval algorithms of Aerosol Optical Depth (AOD), a significant parameter of aerosol optical properties, has been conducted for recent years. However, operational long time series AOD dataset retrieval covering large spatial scales with high accuracy and spatial resolution is still a time-consuming issue. For instance, the daily AOD retrieval procedure including the preprocessing and retrieval based on the Synergetic Retrieval of Aerosol Properties (SRAP) algorithm [2] which was implemented with the IDL language on a dual 3.0 GHz Intel Core (TM) 2 PC, with 3GB of memory costs over 50 hours in general. Effective computing solutions should be further investigated for long time series AOD dataset production.

Grid system could provide uniform and location independent access to geographically and organizationally dispersed heterogeneous resources that are persistent and supported including computing environment, data, and instrument systems, etc. [3], and is still a more common production environment in the scientific field compared with the latest emerging Cloud environment. Endeavours towards remote sensing applications in Grid computing such as the parallel algorithm design and implementation of geometric correction by Zhou et al. [4], and the TARIES.NET platform by Shen et al. [5] have proved it a feasible way to accelerate the data-intensive remote sensing applications. Hence, a high throughput Grid environment especially for geoscientists – Remote Sensing Service Node (RSSN) was put forward [6]. The RSSN was constructed based on the HTCondor, an open source workload management system for compute-intensive jobs, and could effectively harness wasted CPU power from otherwise idle desktop computing resources (http://research.cs.wisc.edu/htcondor/description.html).
Generally remote sensing applications could be assigned as Grid computing tasks and distributed to the computing nodes on the RSSN in a data parallel pattern, for instance, taking the preprocessing or retrieval of a satellite image scene as a task. The data parallel strategy could be easily implemented and adapted to many quantitative remote sensing retrieval applications due to the spatial independent characteristics of pixels. Xue et al. also took the SRAP algorithm from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor data as a study case and proved the runtime of SRAP-MODIS retrieval procedure deployed on the RSSN could be

reduced to an average of 8.36 hours from 50.65 hours based on a single PC in [6].

Except for the favorable data parallel nature of remote sensing applications, some workflows could also be divided into several sub-workflows. The sub-workflows either deal with different input datasets whose processing results might be corresponding to different parameters of retrieval models or involve different processing modules. In these cases, the sub-workflows are independent with each other, and could be made parallel, thus might further reduce the runtime. An example is the thermal inertia retrieval procedure by Xue and Cracknell [2]. Three sub-workflows to obtain the a) temperature difference between day and night, b) albedo, and c) other parameters could be divided before the thermal inertia calculation. Sub-workflow level relation and parallel has been paid attention. Liang et al. proposed a hierarchy workflow model to support cross-enterprise collaboration and its application [7], and in the work of distributed workflow execution algorithms by Schuster, sub-workflow distribution proved to be a well-suited application-independent and generally applicable distributed execution model [8]. The sub-workflow level issue could be further investigated in remote sensing application implementations.

In this paper the sub-workflow parallel of retrieving AOD procedure from the MODIS satellite data based on the SRAP-MODIS algorithm on the RSSN was implemented. In Section 2 we addressed the workflow analysis of SRAP-MODIS retrieval procedure, the RSSN platform and the sub-workflow parallel implementation. The AOD retrieval result and the sub-workflow parallel implementation performance were presented in Section 3. At last the work was concluded and several future aspects were raised.

2. DESIGN AND IMPLEMENTATION

5.1. SRAP-MODIS Algorithm Workflow Analysis

The detail information of SRAP-MODIS algorithm could be found in [1] and [2], while the workflow of SRAP-MODIS procedure is shown in Fig. 1 including the preprocessing and algorithm retrieval. From Fig. 1 it could be seen that in the preprocessing stage, three sub-workflows could be extracted, each of which is designated to a MODIS input dataset, i.e. MOD_02/03/04 data (MOD_02 for the top of atmosphere reflectance, MOD_03 for the longitude, latitude, sensor zenith angle and solar zenith angle, and MOD_04 for the MODIS AOD dataset products). The preprocessing includes the data resize, geometric correction, cloud mask, image cut, image mosaic, etc. and converges on the algorithm retrieval module after all the preprocessing results are prepared. Most processes in the three sub-workflows are independent from each other.

The preprocessing sub-workflow corresponding to the MOD_02 and MOD_03 are much more time-consuming than the MOD_04 (MOD_04 sub-workflow runtime takes up about 13% of that of MOD_03 sub-workflow and 4% of that of MOD_02 sub-workflow) through analyzing the historical runtime logs of the SRAP-MODIS retrieval procedure, which mainly due to the differences in the data amount to be transferred and processed. As a result, in the parallel design we took the strategy of making MOD_04 sub-workflow run in a local PC rather than be submitted to the HTCondor computing pool, thus leaving more computing resources to the MOD_02 and MOD_03 sub-workflows.

5.2. RSSN Grid Computing Platform

The RSSN Grid computing platform was constructed based on the middleware HTCondor and at present computing resources available of RSSN are over 100 computing cores in total. Some computing nodes will join in the HTCondor computing pool or quit.

Only one machine was taken as the central management node on the RSSN, and any other machine could be taken as the submit machine or work machine. When the server received a job and parsed, a corresponding script which met the demands of the HTCondor middleware was generated and then submitted to the computing pool. After the Grid tasks finished, the results would be sent back to the submit machine to gather and mosaic.

5.3. SRAP-MODIS Sub-workflow Implementation

To implement the sub-workflow parallel, a middle agent layer was added to the original layer structure of RSSN, shown in Fig.2. In the middle layer each agent was responsible for a sub-workflow and played the role of submit machine in the HTCondor pool, and would also receive the request from the users' task parser.

There were four roles in the logical structure in this situation: a) outside users, b) several submit machines as agents, c) a manager server, and d) execution machines. The roles b), c) and d) are deployed in an HTCondor pool.

For the sub-workflow parallel implementation of SRAP-MODIS retrieval procedure specifically, three agents called Fiji, Brazil and Georgia were deployed. Since the MOD_04 preprocessing sub-workflow didn't take up much time, once the server received and parsed the job submitted by an outside user, the main script including the MOD_04 sub-workflow process and the complete workflow would be generated and submitted to the agent "Fiji", which was a submit machine in the HTCondor computing pool. The other two agents dealt with the MOD_02 and MOD_03 sub-workflow scripts. After preprocessing results were generated, the MOD_02 relevant image files would be transferred back to the agent "Brazil", and MOD_03 relevant image files back to the agent "Georgia". Agent Fiji, Brazil and Georgia would then mosaic and partition the MOD_04/02/03 data results in separate. When the partitioning MOD_02/03/04 data were prepared, the main agent "Fiji" generated retrieval task executable files and submitted to the HTCondor pool. The computing nodes would fetch temporary data in need from agents. When the retrieval was finished, the final result data were transferred back to the “Fiji” and restored, and the “Fiji” would then send a job finished message to the server.

3. RETRIEVAL RESULTS AND PERFORMANCE

An initial experiment including six-day data samples was conducted. Fig. 3 shows an example of retrieved AOD from MODIS at 1km spatial resolution based on the RSSN.

The runtime comparison was made and presented in Fig. 4 between the original workflow implementation with only the data parallel strategy on the RSSN and its sub-workflow parallel combined strategy version.

The (a), (b) and (c) in Fig. 4 shows the runtime of the pre-processing, mosaic and partition, and total workflow execution respectively. The runtime of pre-processing in Fig. 4 (a) reduced to less than 4 hours compared with 5-7 hours in original strategy. On one hand the sub-workflow parallel mitigated the data transferation bottleneck including a) the executable files from the submit machines to the work machines, b) data to be processed from the data node to the work machines, and c) data results from work machines to three agents rather than one submit machine. On the other hand, the modules corresponding to the same spatial data area were distributed to different computing nodes rather than serially executed on a work machine. The runtime of mosaic and partition stage in Fig .4 (b) reduced 33.6% on average mainly because this stage was completed in three agents rather than one in the original solution. There was a severe cut down in the total runtime shown in Fig. 4 (c).

4. CONCLUSION AND DISCUSSION

Grid computing has been a common computing solution to many scientific issues. The attempts of construction of RSSN have provided an effective solution specifically for the remote sensing applications. The RSSN generally adopts the data parallel implementation due to the data independent nature of most remote sensing applications. From an initial design, implementation and experiment results in this paper, parallel considering the sub-workflow level could further help accelerate the image data preprocessing and retrieval, thus providing an operational computing solution for long-time series, large spatial scale AOD datasets production.

5. ACKNOWLEDGEMENT

This work was partly supported by the National Natural Science Foundation of China (NSFC) under grant no. 41101323, the Ministry of Science and Technology, China, under grant no. 2013CB733403 and the Major International Cooperation and Exchange Project of NSFC (grant no. 41120114001). Many thanks are due to the Centre for High Throughput Computing at UW-Madison for the open source HTCondor software used in this paper.

REFERENCES

[1] J. Tang, Y. Xue, T. Yu, and Y. Guan, "Aerosol optical thickness determination by exploiting the synergy of TERRA and AQUA MODIS," *Remote Sensing of Environment,* vol. 94, pp. 327-334, 2005.

[2] Y. Xue and A. Cracknell, "Operational bi-angle approach to retrieve the Earth surface albedo from AVHRR data in the visible band," *International Journal of Remote Sensing,* vol. 16, pp. 417-429, 1995.

[3] W. E. Johnston, "Using computing and data grids for large-scale science and engineering," *International Journal of High Performance Computing Applications,* vol. 15, pp. 223-242, 2001.

[4] H. Zhou, X. Yang, H. Liu, and Y. Tang, "GPGC: a Grid‐enabled parallel algorithm of geometric correction for remote‐sensing applications," *Concurrency and Computation: Practice and Experience,* vol. 18, pp. 1775-1785, 2006.

[5] Z. Shen, J. Luo, G. Huang, D. Ming, W. Ma, and H. Sheng, "Distributed computing model for processing remotely sensed images based on grid computing," *Information Sciences,* vol. 177, pp. 504-518, 2007.

[6] Y. Xue, Z. Chen, H. Xu, J. Ai, S. Jiang, Y. Li, Y. Wang, J. Guang, L. Mei, and X. Jiao, "A high throughput geocomputing system for remote sensing quantitative retrieval and a case study," *International Journal of Applied Earth Observation and Geoinformation,* vol. 13, pp. 902-911, 2011.

[7] Z. Liang, Z. Jun, L. X. Shen, X. J. Yang, and X. L. Xie, *Hierarchy workflow model supporting cross-enterprise collaboration and its application*. Marietta: American Scholars Press, 2007.

[8] H. Schuster, "Pros and cons of distributed workflow execution algorithms," in *Data Management in a Connected World: Essays Dedicated to Hartmut Wedekind on the Occasion of His 70th Birthday*. vol. 3551, T. Harder and W. Lehner, Eds., ed Berlin: Springer-Verlag Berlin, 2005, pp. 215-234.

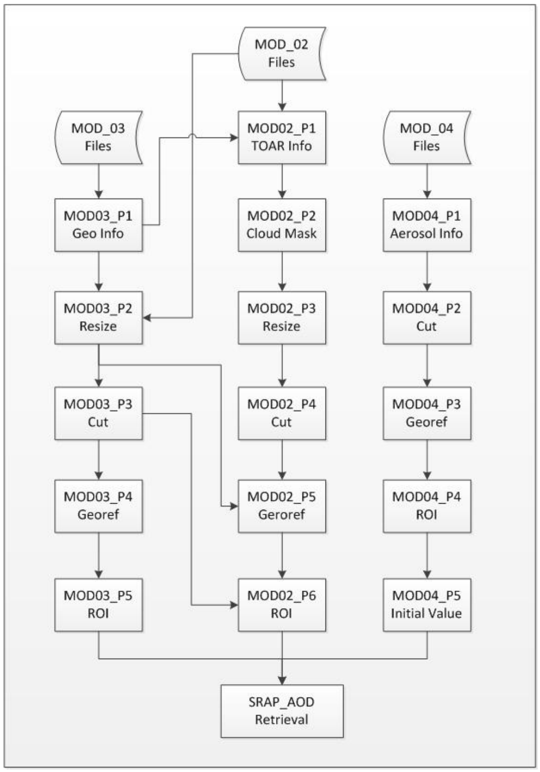


Figure 1. The Workflow of SRAP-MODIS Algorithm.

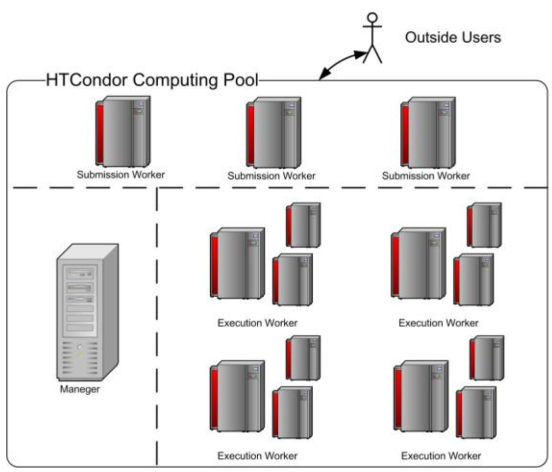


Figure 2. The Logical Structure of RSSN Computing Pool.

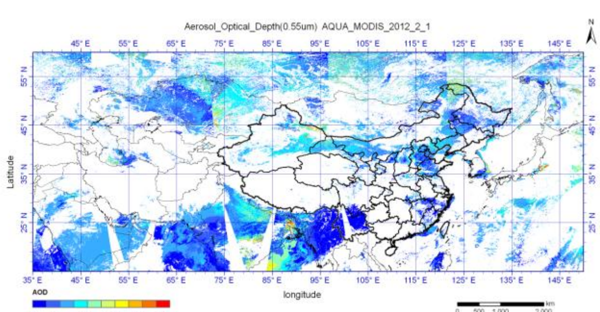
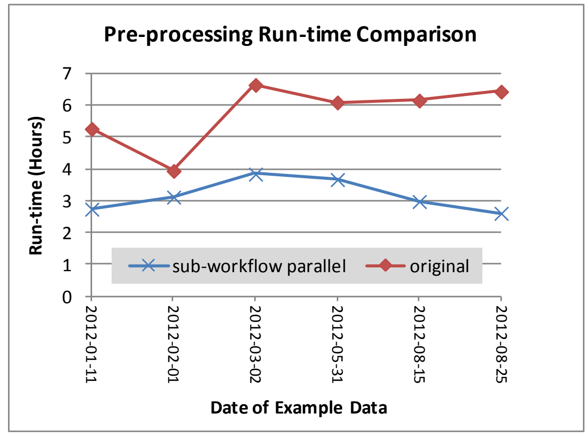
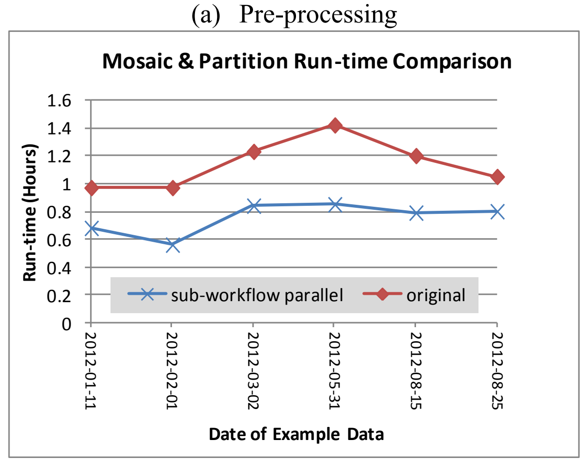


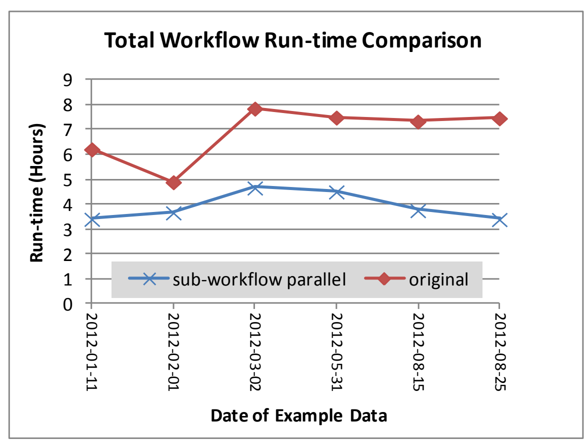
Figure 3. An Example of Retrieved AOD at 1km from SRAP-MODIS Algorithm on the RSSN.



(a) Pre-processing



(b) Mosaic and Partition



(c) Total Workflow

Figure 4. Run-time Comparison of Original Workflows and Sub-workflow Parallel Implementation.