Grid Workflow Modeling for Remote Sensing Retrieval Service with Tight Coupling

Jianwen Ai[1,3,4], Yong Xue[1,2], Jie Guang[1,4], Yingjie Li[1,4], Ying Wang[1,4], and Linyan Bai[1,4]

[1] 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, P.O. Box 9718, Beijing 100101, China
[2] Department of Computing, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK
[3] College of Resources and Environmental Sciences, Northeast Agricultural University, Harbin, 150030, China
[4] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
neau_ajw@hotmail.com, y.xue@londonmet.ac.uk

Abstract. Grid computing technology is a new way for remotely sensed data processing. Tight-coupling remote sensing algorithms can't be scheduled by grid platform directly. Therefore, we need a interactive graphical tool to present the executing relationships of algorithms and to generate automatically the corresponding submitted description files for grid platform. In this paper we mainly discusses some application cases based on Grid computing for Geosciences and the application limit of Grid in remote sensing, and gives the method of Grid Workflow modeling for remote sensing. Then based on the modeling, we design a concrete example.

1 Introduction

Remote sensing data is characterized by largeness and instantaneousness. The analysis and sharing of these huge amounts of data is a big challenge for the remote sensing community (Hu *et al.* 2005). The tremendous computing requirement of the algorithms and the high costs of high-performance supercomputers drive us to hunt for share of computing resources. The emerging computational grid technologies are expected to make feasible the creation of a computational environment handling many PetaBytes of distributed data, tens of thousands of heterogeneous computing resources, and thousands of simultaneous users from multiple research institutions (Giovanni et al. 2003).

Fortunately, within the spatial information field, there are successful application cases based on Grid computing. Work Package (WP) 9 of the DataGrid aims to demonstrate the use of Grid technology for remote sensing applications and earth observation (Giovanni *et al. 2003*). The Information Power Grid (IPG) (http:// www. ipg.nasa.gov) is NASA's high-performance computational Grid. Computational Grids are persistent networked environments that integrate geographically distributed

G. Allen et al. (Eds.): ICCS 2009, Part II, LNCS 5545, pp. 349–356, 2009.
© Springer-Verlag Berlin Heidelberg 2009

supercomputers, large databases, and high-end instruments. These resources are managed by diverse organizations in widespread locations and shared by researchers from many different institutions (http://www.ipg.nasa.gov). The IPG is a collaborative effort between NASA Ames, NASA Glenn, and NASA Langley Research Centers, and the NSF PACI programs at SDSC and NCSA (http://www.ipg.nasa.gov). GENIE (Grid ENabled Integrated Earth System Model) is a new Grid-enabled modeling framework that can compose an extensive range of Earth System Models (ESMs) for simulation over multi-millennial timescales, to study ice age cycles and long-term human-induced global change (Andrew *et al. 2005*). The scientific focus of GENIE is on long-term and paleo-climate change, especially through the last glacial maximum (~21kyr BP) to the present interglacial, and the future long-term response of the Earth system to human activities (http://www.genie.ac.uk/about/overview.htm. The goal of GENIE is to integrate models of the atmosphere, ocean, sea-ice, marine sediments, land surface, vegetation and soil, ice sheets and the energy, biogeochemical and hydrological cycling within and between components (http://www.genie.ac.uk/about/modelling.htm). The Earth System Grid II (ESG) (http://www.earthsystemgrid.org/about/ overviewPage.doc) is a new research project sponsored by the U.S. DOE Office of Science under the auspices of the Scientific Discovery through Advanced Computing program (SciDAC). The primary goal of ESG is to address the formidable challenges associated with enabling analysis of and knowledge development from global Earth System models (http://www.earthsystemgrid.org/about/overviewPage.doc). GEON (GEOsciences Network) is the cyberinfrastructure project that is bringing together information technology and geoscience researchers from multiple institutions in a large-scale collaboration (http://www.geongrid.org). The aim of GEON is to build data-sharing frameworks, identify best practices, and develop useful capabilities and tools to enable dramatic advances in how geoscience is done (Young *et al. 2005*). TeraGrid is a collaboration of partners providing a high-performance, nationally distributed capability infrastructure for computational science (Charles 2005). The Geographic Information Science Gateway (GISolve) is a TeraGrid Science Gateway project based at the National Center for Supercomputing Applications (NCSA) (http://kb.iu.edu/data/awod.html). Its focus is on geographic information science, an interdisciplinary field involving geography and other social sciences, computer science, geodesy, and information sciences for the study of generic issues in the development and use of computationally intensive geographic information systems (GIS) technologies (http://kb.iu.edu/data/awod.html).

2 The Application Limit of Grid in Remote Sensing

Remote sensing quantitative retrieval is a complex computing process due to the terabytes or petabytes of data processed and the tight-coupling remote sensing algorithms. The tight-coupling feature makes that remote sensing algorithm modules need to be processed by computer according to the logic order. The transfer among sensing algorithm modules not only includes processed data files, but also includes control files. Real largeness remote sensing data movement between remote sensing algorithm modules scheduled must be either via an interaction with a data movement service, or through specialized binary-level data channel running directly

between the tasks involved (Fox and Gannon 2006). The feature makes that the existing Grid platform cannot satisfy with our requirements. When we use Grid platform schedule sensing algorithm modules directly, we find that we cannot get the expectable result. Grid platform can only schedule irrelevant job. Therefore, we need to design a tool to control the order of remote sensing algorithms scheduled in Grid system. Besides, although much of Grid software technology addresses the issues of resource scheduling, quality of service, fault tolerance, decentralized control and security and so on, which enable the Grid to be perceived as a single virtual platform by the user, grid computing is not yet mature (Berman *et al. 2003*). There are many open issues to be addressed and missing functionality to be developed, and more will emerge as uses of computing Grids proliferate (Berman *et al. 2003*). Grid platform is not a special platform for remote sensing. It can also not make validity check of condition for remote sensing algorithm modules to be scheduled with short of special knowledge on remote sensing.

To solve the above problem, we must let the remotely sensed data processing module be scheduled in the Grid environment. We need design an interactive GUI interface to represent the processed steps of remote sensing algorithm modules and their relations including concurrence/synchronism and executed order. We need a tool which can use interactive graphical editors to present the executing relationships of algorithms on human-friendly diagrams and to generate automatically the corresponding submitted description files of grid platform. Fortunately, workflow technologies make it possible. Using workflow technology, we can construct a remote sensing information processing environment to integrate the distributed data and computational resources. It is not a new idea to apply workflow technologies to Grid platform. Actually, people try to integrate the workflow technologies under Grid platforms in many projects, such as Triana (Majithia et al. 2004), Unicore (Riedel et al. 2006), Kepler (Zhang 2006), ICENI (McGough et al. 2006), Taverna (Turi et al. 2007), GridFlow (Cao et al. 2002, 2003), Askalon (Fahringer et al. 2005), Karajan (http://www.cogkit.org), etc. These Grid workflows give a host of useful workflow composition tools with graph-based modeling or language-based modeling. About Language-based modeling, Yu and Buyya (2005) consider that language-based modeling may be convenient for skilled users, but they require users to enumerate a lot of language specific syntax; in addition, it is impossible for users to express a complex and large workflow by scripting workflow components manually; and workflow languages are more appropriate for sharing and manipulation, whereas the graphical representations are intuitive but they require to be converted into other forms for manipulation. So most Grid systems, workflow languages are designed to bridge the gap between the graphical clients and the Grid workflow execution engine (Guan et al. 2004). By analyzing, we find that these existing Grid workflow platforms can not satisfy with our requirement owing to the features of remote sensing, such tight-coupling remote sensing algorithm modules, largeness and instantaneousness remote sensing data, etc. Therefore, it is necessary to design a workflow composition tool using graph-based modeling for remote sensing services.

3 Grid Workflow Modeling for Remote Sensing Retrieval Service

Grid workflow modeling for remote sensing retrieval service includes that its task definition, structure definition and mapping relation of specific Grid resources for task execution. Task definition includes that all information to execute a task in grid environment, such as function descriptions of task, previous task, support environment requirements, the size of memory, minimum space of hard disk, etc. In general, a workflow structure can be represented as a Directed Acyclic Graph (DAG) or a non-DAG (Sakellariou and Zhao 2004). Non-DAG workflow includes the iteration structure, which isn't suitable for modeling for remote sensing algorithm modules. In DAG-based workflow, Yu and Buyya (2005) consider that workflow structure can be classified as sequence, parallelism, and choice; Sequence is defined as an ordered series of tasks, with one task starting after a previous task has completed; Parallelism represents tasks which are performed concurrently, rather than serially; and in choice control pattern, a task is selected to execute at runtime when its associated conditions are true. Mapping relation of specific Grid resources for task execution binds workflow tasks to specific grid resources. Aiming at largeness and instantaneousness of remote sensing data, we adopt tasks acting as data movement or computing code movement according to the schedule arithmetic of engine. Besides, grid workflow modeling for remote sensing retrieval service, unlike the object-orient design where interaction of the objects are driven by method calls, the states of our model can control their own actions and react to parameters which are the control-flow elements such as branching or an expression of value provided by their users. It also provides a mechanism for concurrency or sequential computation through tokens that are fired by the transition function. When there is a transition of the state, the model get the parameter of tokens from a FIFO (first in, first out) queue with capacity equal to one, executing the concurrency or sequential computation correspondingly. The Grid workflow modeling for remote sensing retrieval service can be represented as an 8-tuples GWP = (K, D, R, P, s, d, F, T), where:

- K is a finite set of states, where each black-box is regarded as a state. Any a element of K expresses a remote sensing algorithm module or a task of remote sensing data movement or remote sensing algorithm code movement.
- D is a set of data, which includes the initial data file of remote sensing information, the results of data disposed. D is the condition of remote sensing algorithm modules executed.
- R is a subset of binary relation K x K. It is a set of arcs, where each element represents the order relation among the executed remote sensing algorithm modules.
- P is a finite set of mapping parameters, parameters provided by users and control tokens. The mapping relation of specific Grid resources for remote sensing algorithm module execution, which binds workflow tasks to specific grid resources. The parameters provided by users include function descriptions of task, previous task, support environment requirements, the size of memory, minimum space of hard disk, etc. The control tokens expresses that remote sensing algorithm module gets the condition scheduled.
- $s \in K$ and $d \in D$ are the initial remote sensing state and initial remote sensing data file.

- F $\subseteq$ K is the set of final states, and
- T is a transition function from (K-F) x (P $\cup$ { Φ }) x D to K x D.
- We now formalize the operation of the model.

When the Workflow is started, triple (s, Φ ,d) (K $\subseteq$ F) x (P $\cup$ { Φ }) is its initial transition state of T; for all q$\in$ (K-F) and p$\in$ K, if (q, p)$\in$ R, the workflow scans its set of data and its set of parameters, getting relevant d1$\in$ D and p1$\in$ P; then the transition is fired, according to the semantic analysis of the parameters, and changing the state q to the state p and producing the result d2 from d1, until it finds a state pi$\in$ F; and it halts. The data file di is the result we need.

4 Implementation

We have implemented Grid workflow modeling for remote sensing retrieval service and Grid workflow composition tool using graph-based modeling for remote sensing services (see Figure 1). We can use XML to record the user's describe information of workflow (see Figure2). We have accomplished the scheduled of algorithms of remote sensing according workflow mechanism. The next work is to transform the XML format of workflows into the corresponding submitted description files of grid platform according to the criterions of Grid platform.

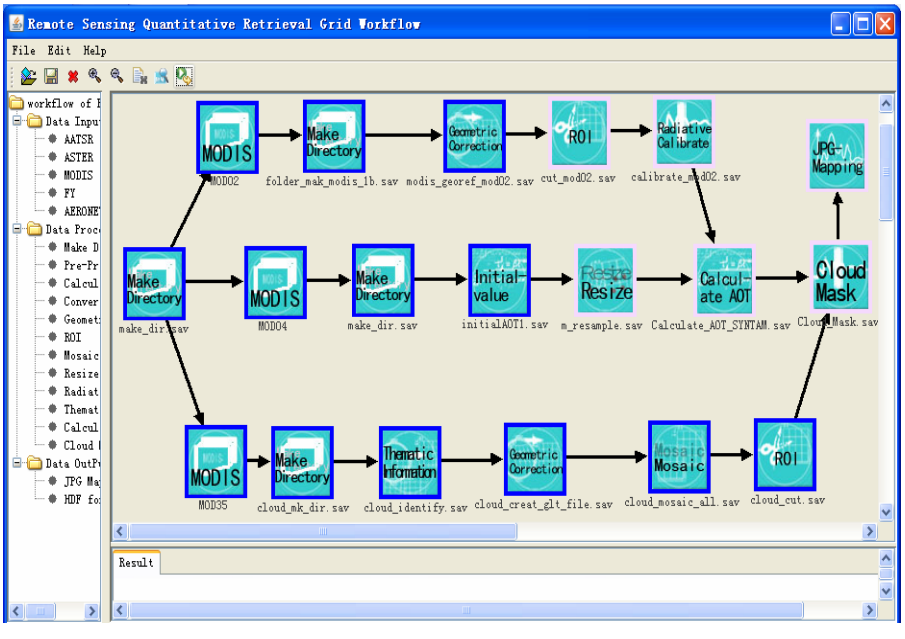

Fig. 1. A case of workflow scheduled

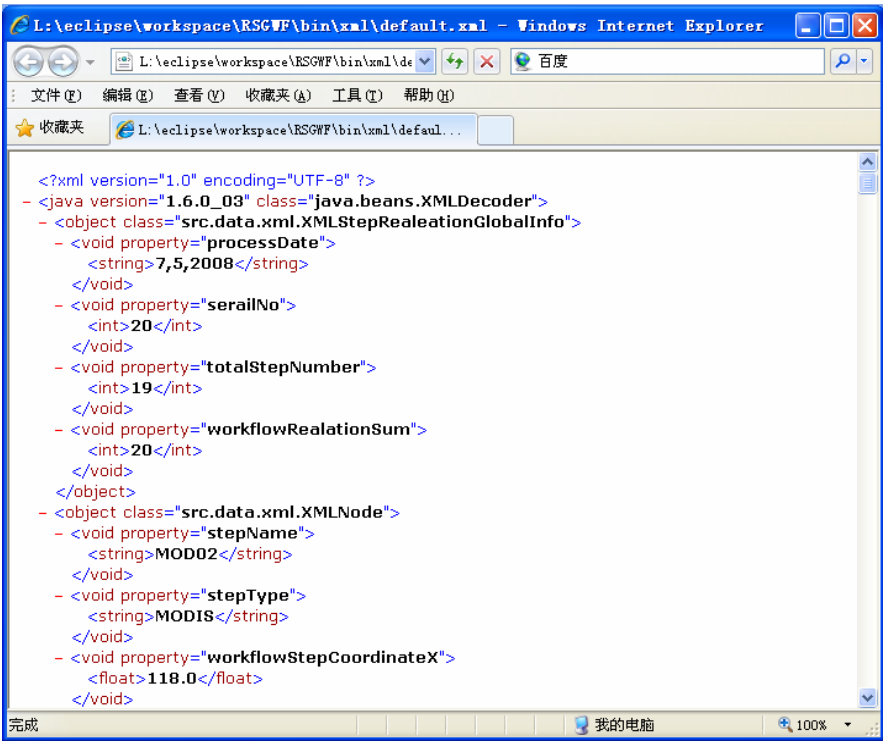

Fig. 2. A case of workflow xml format

5 Conclusion

Grid workflow composition tool using graph-based modeling can express the tight-coupling features of among remote sensing algorithms conveniently. Through it the existing remotely sensed data processing modules can be scheduled orderly by Grid platform. Through it, the existing remotely sensed data processing modules and algorithms resource can be shared. It seems like the common service in Grid environment shared. It much enhances the resource utility rate. In order to design a Grid workflow modeling for remote sensing retrieval service with tight coupling, the paper analyses the features of remote sensing data and algorithms and introduces the successful application cases based on Grid computing for Geo-sciences. By analyzing the application limit of Grid in remote sensing, we give Grid Workflow modeling for remote sensing retrieval Service. Combing the rule of modeling, we design Grid workflow composition tool using graph-based modeling for remote sensing services.

Acknowledgement

This work was supported in part by National Science Foundation of China (NSFC) under Grant No. 40671142, by the Ministry of Science and Technology (MOST),

China under Grant No. 2008AA12Z109 and Grant No. 2007CB714407, by Chinese Academy of Sciences (CAS) under Grant No. KZCX2-YW-313, by the NSFC under Grant No. 40471091.

References

[1] Xue, Y., Wang, J.Q., Wu, C.L., Hu, Y.C., Guo, J.P., Zheng, L., Wan, W., Cai, G.Y., Luo, Y., Zhong, S.B.: Information Registry of Remotely Sensed Meta-modeling Grid Environment. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 1–8. Springer, Heidelberg (2006)

[2] Fran, B., Hey Anthony, J.G., Fox Geoffrey, C.: Grid Computing Making the Global Infrastructure a Reality. John Wiley & Sons Ltd., UK (2003)

[3] Giovanni, N.A., Luigi, F.B., Linford, J.: Grid technology for the storage and processing of remote sensing data: description of an application. In: Proceedings of the society of photo-optical instrumentation engineers (SPIE), vol. 4881, pp. 677–685 (2003)

[4] Price, A., Lenton, T., Cox, S., Valdes, P., Shepherd, J., GENIE team: GENIE: Grid ENabled Integrated Earth System Model. ERCIM News 61, 15–16 (2005)

[5] Youn, C., Baru, C., Bhatia, K., Chandra, S., Lin, K., Memon, A., Memon, G., Seber, D.: GEONGrid Portal: Design and Implementations. In: GCE 2005 Workshop on Grid Computing based on SC 2005, November 2005, Seattle, WA (2005)

[6] Catlett, C.E.: TeraGrid: A Foundation for US Cyberinfrastructure. In: Jin, H., Reed, D., Jiang, W. (eds.) NPC 2005. LNCS, vol. 3779, p. 1. Springer, Heidelberg (2005)

[7] Yincui, H., Yong, X., Jiakui, T., Shaobo, Z., Guoyin, C.: Data-parallel Georeference of MODIS Level 1B Data Using Grid Computing. In: Sunderam, V.S., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2005. LNCS, vol. 3516, pp. 883–886. Springer, Heidelberg (2005)

[8] Majithia, S., Shields, M., Taylor, I., Wang, I.: Triana: a graphical Web service composition and execution toolkit. In: Proceedings of Web Services 2004, pp. 514–521 (2004)

[9] Riedel, M., Menday, R., Streit, A., Bala, P.: A DRMAA-based target system interface framework for UNICORE. In: 12th International Conference on ICPADS 2006, vol. 2 (2006) CD-ROM

[10] Zhang, J.: Ontology-Driven Composition and Validation of Scientific Grid Workflows in Kepler: a Case Study of Hyperspectral Image Processing. In: GCCW 2006. Fifth International Conference on Grid and Cooperative Computing Workshops, pp. 282–289 (2006)

[11] McGough, A.S., Lee, W., Darlington, J.: ICENI II. In: First International Conference on Comsware 2006, pp. 1–4 (2006)

[12] Turi, D., Missier, P., Goble, C., De Roure, D., Oinn, T.: Taverna Workflows: Syntax and Semantics. In: IEEE International Conference on e-Science and Grid Computing, Bangalore, pp. 441–448 (2007)

[13] Cao, J., Jarvis, S.A., Saini, S., Kerbyson, D.J., Nudd, G.R.: ARMS: an Agent-based Resource Management System for Grid Computing. Scientific Programming, Special Issue on Grid Computing 10(2), 135–148 (2002)

[14] Fahringer, T., Prodan, R., Duan, R., Nerieri, F., Podlipnig, S., Qin, J., Siddiqui, M., Truong, H.L., Villazon, A., ieczorek, M.: ASKALON: a Grid application development and computing environment. In: The 6th IEEE/ACM International Workshop on Grid Computing (2005) CD-ROM

[15] Fox, G., Gannon, D.: Workflow in Grid Systems. In: Concurrency and Computation: Practice & Experience, vol. 18, pp. 1009–1019. John Wiley and Sons Ltd., UK (2006)

[16] Yu, J., Buyya, R.: A Taxonomy of Workflow Management Systems for Grid Computing. Journal of Grid Computing 3, 171–200 (2005)
[17] Guan, Z., Hernandez, F., Bangalore, P., Gray, J., Skjellum, A., Velusamy, V., Liu Y.: "Grid-Flow": A Grid-Enabled Scientific Workflow System with a Petri Net-based Interface,Technical Report (December 2004), http://www.cis.uab.edu/gray/Pubs/grid-flow.pdf
[18] Sakellariou, R., Zhao, H.: A Low-Cost Rescheduling Policy for Efficient Mapping of Workflows on Grid Systems. Scientific Programming 12(4), 253–262 (2004)