网格环境下银河系化学演化研究*

罗 泽[1+], 崔辰州[2], 南 凯[1], 阎保平[1]

[1](中国科学院 计算机网络信息中心,北京 100080)
[2](中国科学院 国家天文台,北京 100012)

Chemical Evolving Research of the Galaxy Under Grid Environment

LUO Ze[1+], CUI Chen-Zhou[2], NAN Kai[1], YAN Bao-Ping[1]

[1](Computer Network Information Center, The Chinese Academy of Sciences, Beijing 100080, China)
[2](National Astronomical Observatories, The Chinese Academy of Sciences, Beijing 100012, China)
+ Corresponding author: Phn: +86-10-58812350, E-mail: luoze@cnic.cn, http://www.cnic.ac.cn
Received 2003-12-04; Accepted 2005-01-07

Luo Z, Cui CZ, Nan K, Yan BP. Chemical evolving research of the galaxy under grid environment. ***Journal of Software*****, 2005,16(8):1465−1473.** DOI: 10.1360/jos161465

Abstract: Many computing or data intensive scientific applications require a mechanism of sharing and coordinated use of diverse resource as part of their sophisticated problem solving processes. This paper describes the design and implementation of the chemical evolving research of the Galaxy under Grid environment, which is an important part of China Virtual Observatory application system. Through this exemplification, this paper shows a mechanism of the sharing and coordinated use of resourced that support e-Science applications by integrating Grid rvices and Grid service workflow into Grid environment. Such a mechanism could be beneficial for modeling and managing scientific processes of experimental investigation, evidence accumulation and result assimilation.
Key words: grid service; workflow; scientific data grid; virtual observatory; chemical evolution of the galaxy

摘 要: 许多计算或数据密集型的科学应用要求一种共享和协同使用分布异构资源的机制,作为其复杂的问题求解过程的一部分.网格环境下银河系化学演化研究是中国虚拟天文台应用系统的一个重要部分.详细描述了该演化研究的设计和实现.通过这个范例,提出一种共享和协同使用资源的机制.该机制通过在网格环境中集成网格服务和网格服务工作流,能够有效地支持 e-Science 应用.该机制对于建模和管理实验研究、数据积累和科研成果的消化吸收这样的科研过程是有益的.

---

* Supported by the National Natural Science Foundation of China under Grant No.90412011 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2002AA104240 (国家高技术研究发展计划(863)); the Important Informationization Constructing Project of the 'Tenth Five-Year-Plan' of the Chinese Academy of Sciences, the Scientific Database and its Application System under Grant No.INF105-SDB (中国科学院"十五"信息化建设重大项目"科学数据库及其应用系统")

LUO Ze was born in 1976. He is a research associate at Computer Network Information Center, the Chinese Academy of Sciences. His current research areas are grid computing and application. **CUI Chen-Zhou** was born in 1976. He is a research associate at National Astronomical Observatory, the Chinese Academy of Sciences. His research areas are virtual observatory and grid computing. **NAN Kai** was born in 1974. He is an associate professor at Computer Network Information Center, the Chinese Academy of Sciences. His research areas are grid computing, computer network and database application. **YAN Bao-Ping** was born in 1950. She is a professor at Computer Network Information Center, the Chinese Academy of Sciences. Her research areas are grid computing and next generation Internet.

关键词: 网格服务;工作流;科学数据网格;虚拟天文台;银河系化学演化
中图法分类号: TP393 文献标识码: A

1 Introduction

Grid computing involves the agglomeration of diverse resources in dynamic, distributed virtual organizations[1]. Grid technologies[2] are infrastructures that support the sharing and coordinated use of such diverse resources. Grid technologies are currently seeing widespread adoption in the scientific computing community. Apart from the problems that are inherent in dealing with resources (i.e., algorithms and problem-solving techniques, resource management, security, instrumentation and performance analysis, network infrastructure, etc.), Grid technologies have to solve problems similar to those addressed by Web Services, namely description, discovery, communication, remote invocation, and so forth. Grid community has recognized this fact, resulting in the development of the Open Grid Services Architecture (OGSA)[3]. OGSA is the result of lessons learned while developing the Globus Toolkits, which has become the de facto standard for Grid computing. OGSA uses the Web Services Description Language (WSDL) to achieve self-describing and discoverable services. It defines a set of standard interfaces that a Grid Service may export and that enable features such as discovery, service lookup, lifetime management, notification, and credential management.

Many computing and data intensive scientific applications, such as climate modeling, astrophysics, high energy physics, structural biology and chemistry, require a mechanism of the sharing and coordinated use of diverse resources as part of their sophisticated problem solving processes. Although elementary Grid Services exist that enable scientific application developers to authenticate, access, manage, and discovery remote resources, developing scientific applications for these large-scale collaborative Grid infrastructure is a difficult task. In a similar vein to Web Services technologies, Grid services can realize their full potential only if there is a method to dynamically compose new services out of the existing ones. By using the Grid computing approach, a collaborative Grid service workflow can be provided in order to enable the coordinated use of numerous distributed and heterogeneous resources.

Scientific Data Grid is one part of the China National Grid. Supported by the National High Technology Research and Development Program of China (863), Scientific Data Grid project is an application Grid of "High Performance Computer and its Key Software", an important special project of the 863 in the 10th Five-Year-Plan. Scientific Data Grid is also one part of the "Integration Experimental Platform of Network Computing Environment", an important project supported by National Natural Science Foundation of China (NSFC). At the same time, Scientific Data Grid is an important part of the "Scientific Database and its Application System", one of the important informationization constructing projects of the Chinese Academy of Sciences (CAS). One of the primary goals of the Scientific Data Grid is to develop and run scientific applications based on Grid technologies, as an illustration of e-Science enabled by Grid technologies. In the Scientific Data Grid, Computer Network Information Center (CNIC) of CAS collaborates with National Astronomical Observatories of CAS to develop China Virtual Observatory as one of the scientific application systems.

As an exemplification of deep-going analysis and processing based on the physical data resources, chemical evolving research of the Galaxy under Grid environment is an important part of the China Virtual Observatory application system. This paper describes the design and implementation of the exemplification in detail. Through this exemplification, we show a mechanism of the sharing and coordinated use of diverse resources that support e-Science applications by integrating Grid services and Grid service workflow into any OGSA compliant environment. Such a mechanism could be beneficial for modeling and managing scientific processes of experimental investigation, evidence accumulation and result assimilation. Scientific processes themselves can then be reused, shared, adapted and annotated.

2 Overview

Chemical evolving research of the Galaxy is one of the important research directions in astronomical field. Through direct observation and simple data processing, astronomer can obtain data like stars' position, velocity and element abundances. Astronomers have provided several mass distribution models of the Galaxy, from which we can induce potential field function of the Galaxy. The galactic potential field function can be used for calculating star's acceleration at a given position in the Galaxy. Taking star's position and velocity as input, according to potential field function, astronomer can calculate stellar orbit around the Galaxy during the given time by the way of numerical integration. Combined with stellar element abundances and orbital parameters, such as mean orbital radius, astronomer can calculate element distribution function in the Galaxy and provide constraints for chemical evolving research of the Galaxy.

There are two distinguished characteristics in the chemical evolving research of the Galaxy: One is the access for numerous, distributed and heterogenous data resources. To perform numerical integration for stellar orbit calculating and analysis of element abundances, astronomer should firstly obtain sample stars' information about position, velocity and element abundances. However, generally, these data are geographically distributed, stored with different formats and managed by different database management systems. The other is the requirement for high performance computation resources. In order to obtain sample start's orbit parameters, we calculate stellar orbit by the way of numerical integration, which is based on sample stars' original kinematic parameters and potential field function induced from mass distribution model of the Galaxy. The whole procedure is complicated and requires high performance computation resources.

In chemical evolving research of the Galaxy under Grid environment, based on Globus Toolkits 3 (GT3), which is a java-based implementation of the current Grid services specification using Apache Tomcat and Apache Axis, we implement several Grid services including registry service, data access service, stellar orbit calculating service and statistical analysis service. According to different resources, these Grid services distribute on different hosting environments. Based on Index Service of GT3, registry service implements registry and discovery of Grid service. Data access service supports unified access and integration on distributed and heterogenous databases. Stellar orbit calculating service calculates sample stars' orbits around the Galaxy. Statistical analysis service processes and analyzes the data produced by stellar orbit calculating service, and returns visualized results to astronomer.

The Grid services described above are the necessary components of chemical evolving research of the Galaxy. A Grid service workflow is required to compose these Grid services, which can implement chemical evolving research as a whole. We take Grid Workflow Execution Language (GWEL)[4] as a specification of the workflow descriptions for Grid service in the OGSA framework. GWEL is an XML based language that reuses ideas from BPEL4WS. Similar to the BPEL4WS process model, GWEL's process model represents a peer-to-peer interaction between services described in WSDL. Using GT3, we implement a Grid service workflow execution engine, for processing chemical evolving research workflows specified in GWEL document. The Grid service workflow execution engine itself is a high-level Grid service, hence automatically Grid aware, that can be used within any GT3 environment.

3 Design and Implementation of Individual Grid Service

Standardization and commercialization of Grid promotes combination between Grid technologies and Web

Services technologies and produces service-based Grid architecture. Grid services encapsulate different resources including computing resources, storage resources, network resources, application software and databases under distributed, heterogenous and dynamic environment. Grid services abstract and unify different resources, data and information. This abstraction facilitates flexible, unified and dynamic sharing mechanism and makes Grid system possess standard interfaces and behaviors. Under service-based Grid environment, the interaction model between user and Grid system is shown in Fig.1.

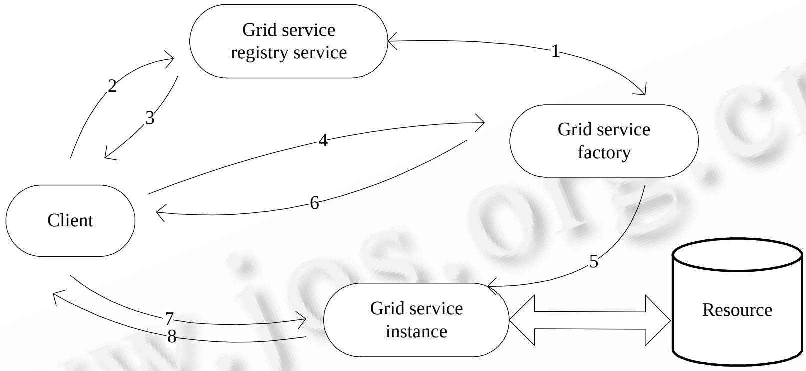

Fig.1 Interaction model between user and Grid system

The interaction between user and Grid system includes following steps:

(a) Grid Service Factory (GSF) registers with Grid Service Registry Service (GSRS), publishes Grid service's introspection, interface and other information, and exposes Grid service interfaces and information about the resources that Grid service encapsulates to user (step 1).

(b) User queries GSRS according to user's requirements. GSRS returns user GSF Handle which satisfies user's requirements (steps 2, 3).

(c) User sends request to GSF according to GSF Handle, asks for functions that Grid service provides (step 4).

(d) GSF creates Grid Service Instance (GSI), and returns GSI Handle to user (steps 5, 6).

(e) It is GSI's responsibility to keep session with user. User sends requests to GSI. GSI interacts with underlying physical resources and returns results to user (steps 7, 8).

It makes sense to destroy the instance as soon as we are done with it, so we can free system resource.

3.1 Data access Grid service

Data access Grid service supports unified access and integration on distributed, heterogenous data resources and conceals distributed, heterogenous characteristics of the underlying physical data resources by a set of common service interfaces. The goals of data access Grid service are:

- To provide controlled exposure of the physical data resources to the Grid.
- To support access to the heterogenous physical data resources through a common interface style while employing the underlying query mechanisms.
- To provide base services that allow higher-level data integration services to be constructed, e.g. distributed query processing and data federation services.
- To leverage emerging Grid infrastructure for security, management, accounting etc.

The data sources of chemical evolving research are two heterogenous databases. One is a Postgres database, which holds sample stars' element abundances information. The other is a MySql database, which stores information about sample stars' original position and velocity in the Galaxy. These two databases distributed on two different computers. We encapsulate these two databases by data access Grid service so they can provide greater transparency. Data access Grid service follows interaction model shown in Fig.1. Data providers statically configure data access Grid service factory and data access Grid service instance. Configuration information consists of meta-data information of physical data resources and behaviors supported by data access Grid service. Meta-data information of physical data resources includes data resources' inherent structure, associated content and capability meta-data, properties and access mechanisms. Behaviors supported by data access Grid service expose operations that physical data resources accept. Data access Grid service factory and data access Grid service instance expose configuration information to the Grid. Data movement to/from the physical data resources takes place as a stream, primarily of XML structured data.

3.2 Stellar orbit calculating Grid service

Supposed that the potential field of the Galaxy has not changed greatly during its evolution lasting billions of years, kinematic characteristics of the stellar orbits will be good indicators of the galactic formation and evolution. Using galactic mass distribution model, we can calculate the stellar orbital parameters in the Galaxy and use kinematic parameters, such as Galactocentric Distance (DG), disk radial distance (R) and disk normal distance (Z), to analyze element abundances gradients. This can reflect the process of galactic evolution factually.

Given the original kinematic data of a star (*X*, *Y*, *Z*, *U*, *V*, *W*. where *X*, *Y* and *Z* are position coordinates in the galactocentric right-angle coordinate system; *U*, *V* and *W* are velocities along the directions of *X*, *Y* and *Z* respectively), we can calculate its orbit in the galactic potential field. In chemical evolving research of Galaxy, we use the Runge-Kutta-Fehlberg algorithm[5] with automatic step length control. The orbits of stars are numerically integrated backward 1.6×10^{10} with the maximum energy variation $\delta(E)/E<10^{-6}$ and the maximum momentum variation $\delta(H)/H<10^{-6}$ at the end of integration.

Currently, we can only obtain stellar right ascensions, declinations, proper motions and parallaxes, but no kinematic parameters. We have to calculate the stellar initial position parameters from these data. We use an algorithm provided by Johnson et al. with which we could calculate the galactocentric space-velocity components *U*, *V* and *W* for a star if its proper motion, parallax and radial velocity are given[6].

The computation process is complicated, and then requires high performance computing resources. Through Grid service, Grid can provide high performance computing resources where high performance computing resources are available, like Supercomputing Center of CNIC, CAS.

3.3 Statistical analysis Grid service

Statistical analysis Grid service matches and statistically analyzes sample stars' element abundances data and kinematic parameters, and then returns the visualized analytic results to astronomers. This visualized experimental results show the sample stars' element abundances distribution. In order to analyze abundances gradients, we took Linear Regression algorithm to fit sample data linearly. With the *X* data, the independent variable, and *Y* data, the dependent variable, the estimated linear regression model is stated as follows:

$$Y_i = A + BX_i\,,\quad A = EY - B\times EX\,,\quad B = \frac{\sum_i^N (X_i - EX)(Y_i - EY)}{\sum_i^N (X_i - EX)^2}.$$

The parameters, *A* and *B*, are estimated by the method of least squares. X_i means disk radial distance *R* or disk normal distance *Z* of the sample star *i*, and Y_i means element abundances of the sample star *i*. *EX* denotes arithmetical average of all sample stars' *R* or *Z*. *EY* denotes arithmetical average of all sample stars' element abundances. *B* is the slope and is also the value of abundances gradient.

4 Orchestration of Grid Services

Currently, workflow systems for Grid services are evoking a high degree of interest. However, it is an area that is currently not defined explicitly by the WS-Resource Framework specification. Furthermore, we are not aware of any tool that makes fully use of the features provided in GT3. In chemical evolving research of the Galaxy under Grid environment, we take GWEL as specification of workflow descriptions for Grid service in the OGSA framework. Using GT3, we implement a Grid service workflow execution engine, for processing chemical evolving research workflows specified in GWEL document. We try to makes fully use of the features provided by GT3 to implement the orchestration of Grid services.

The most difference between Web Services and Grid services is that Grid service is dynamic and stateful. In Grid computing it is necessary to utilize resources efficiently and therefore not occupying them longer than necessary. Therefore instances of Grid services taking part in the workflow should be created by the workflow engine from the specified factory and destroyed when they are not needed any more.

Web Services technologies typically define workflow in a way that all the data that are exchanged between the services involved is transferred through the workflow engine. This is because they usually deal with a moderate level of data transmission across the Web Services[7]. Since e-Science applications typically involve large amount of data, it is essential for workflow engine of Grid services to just reference to data rather than transfer data itself. Data are transferred equivalently between participating Grid services.

Based on GT3, we implement a Grid service workflow engine. A workflow file parser is responsible for recursively parsing GWEL documents and then dispatching the elements to the appropriate instruction classes. In chemical evolving research of Galaxy, we choose SAX to parse the document. Grid service workflow engine creates virtual ports and services that map to processes internal to the workflow. These ports do not physically exist on the Grid service workflow engine and client can invoke methods on these ports, as they get mapped to a set of other calls with respect to the workflow engine. One of the most important parts of the workflow engine is the implementation of the life cycle instructions in the GWEL document, which handle the dynamic creation and destruction of Grid service instance. Creation of an instance is done through importing the stubs for the instance that has to be created. The stubs for an instance could be generated by GWSDL document of a Grid service. Dynamic destruction of instance is achieved through the subscription method, which is available through GT3. Control flow implementation is another important part of Grid service workflow engine. This is achieved by dispatching the control flow instructions in GWEL document to the appropriate class for sequential or parallel execution. Once parser acquires an operation, Java reflection is used to execute the appropriate invocation. In chemical evolving research of the Galaxy under Grid Environment, the workflow process is visualized in Fig.2.

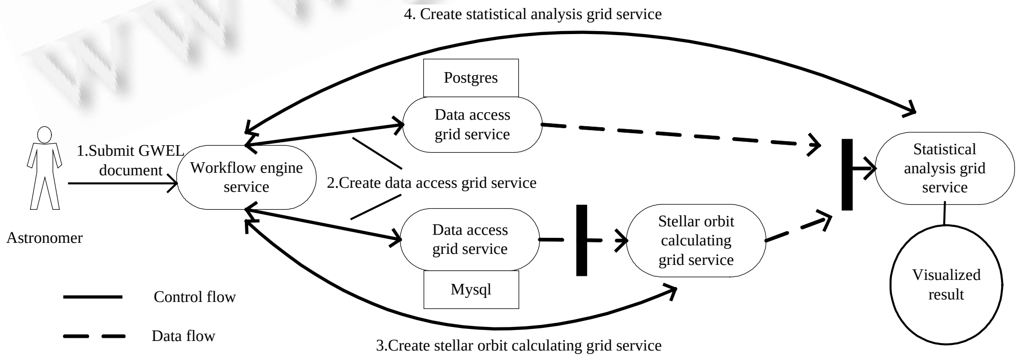

Fig.2 Grid services workflow in chemical evolving research of the Galaxy under Grid environment

5 Experimental Results

We calculate 89 sample stars' orbits around the Galaxy and obtain their kinematic parameters. Partial experimental results are given in Table 1. HD denotes a star's identifier. MaxDG and MinDG indicate respectively the maximum and minimum galactocentric distance of a star. MaxR and MinR mean respectively the maximum and minimum disk radial distance and MaxZ and MinZ denote respectively the maximum and minimum disk normal distance of a star.

Table 1 Partial experimental results of stellar orbit calculating gird service

HD	MaxDG	MinDG	MaxR	MinR	MaxZ	MinZ
400	9.417724979	7.359853302	9.417712117	7.359839485	0.02315704131	−0.02315864969
3454	12.90799519	8.180644823	12.9057315	8.175776174	0.4946127570	−0.4945968310
6834	10.63136987	8.425455728	10.62976865	8.422575452	0.3284933682	−0.3285850230
6840	8.882434716	5.668196142	8.881220688	5.666863401	0.2367993813	−0.2367939522
10307	8.758476619	6.520187255	8.758358689	6.520068895	0.06737815629	−0.06737717094
11007	11.45421300	8.14144423	11.44715744	8.123503068	0.8571721439	−0.8571155595
11592	10.10630223	6.366070751	10.10429491	6.363433894	0.341395952	−0.3415754221
19373A	10.25461737	6.657543681	10.25176439	6.653210030	0.4364442012	−0.4365446729

The distribution of Fe element abundances along disk radial direction, disk normal direction and results of linear fitting are shown in Fig.3.

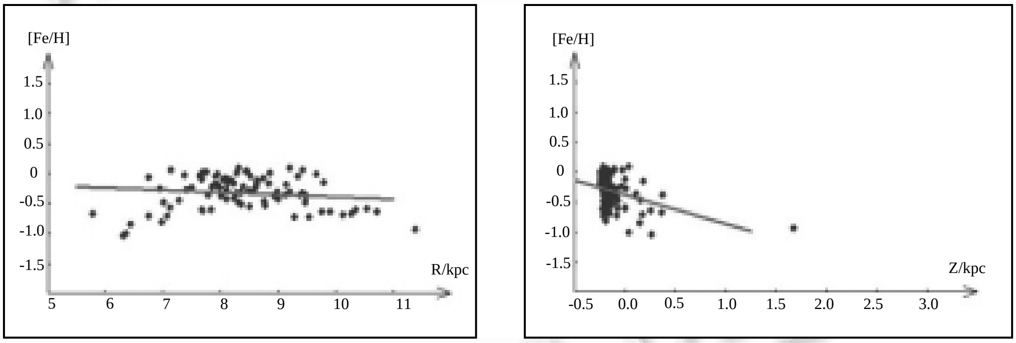

Fig.3 Distribution alone disk radial direction, disk normal direction and linear fitting

With the correlation analyses between element abundances and orbital parameters, we obtain that the Fe element gradient is −0.083dex/kpc along disk radial direction and −0.403dex/kpc along disk normal direction. We get the result that the vertical gradient is steeper than the radial gradient. Our work suggests the existence of the galactic gradient and supports those chemical evolution models, which show that the halo is formed before the disk at the early stage of the Galaxy.

6 Related Work

Like the STScI Digitalized Sky Survey, most of the astronomical software and tools that support online data access and computing are currently based on traditional technologies of web application, such as CGI and JSP. With the development of Web Services and Grid technologies, more and more astronomical applications developed under service-oriented architecture are emerging. Based on Web Services technologies, the Sloan Digital Sky Survey Data Release 1 is now available. In chemical evolving research of the Galaxy, we choose Grid services as common components, which can be composed into higher-level services. Compared with Web Services, the most distinct difference is the dynamic feature of Grid services, which makes Grid services utilize resource more efficiently, and the integration of Grid services is easier.

There have been a number of efforts within the Grid community to develop general-purpose workflow management solutions. One of the most well known projects that address workflow management in a Grid environment is presented in Ref.[7]. The project mainly focuses on the definition of the Grid Services Flow Language (GSFL). However, the project is based on GT version 2.4 and therefore does not consider any WS-Resource Framework features. In Ref.[8], the authors propose a distributed software component architecture, XCAT, for Grid computing that is compatible with the Common Component Architecture specification. CCA components have two types of ports. One type of port, which is called provides port, is identical to a Web Service port. The other type, called uses port, is an external reference from one component to a provided port on another component that can be bound at runtime. The result is a set of Web Services that could be composed as an application. GridAnt[9] uses the Ant workflow processing engine. GridAnt has predefined tasks for authentication, file transfer and job execution, while reusing the XML-based workflow specification implicitly included in ant. The implementations of XCAT and GridAnt are not based on Globus Toolkits. They cannot be applied in OGSA environment directly. Our implementation of Grid service workflow engine completely conforms to Open Grid Service Infrastructure (OGSI) specification[10]. We try to makes fully use of the features provided by GT3 to implement the orchestration of Grid services. The workflow engine itself is a high level Grid service, hence automatically Grid aware, that can be used within any GT3 environment.

7 Conclusions

In chemical evolving research of the Galaxy under Grid environment, based on Globus Toolkits 3, we implement several Grid services including registry service, data access service, stellar orbit calculating service and statistical analysis service. Grid services abstract and unify different resources, data and information. This abstraction facilitates flexible, unified and dynamic sharing mechanism and makes Grid system possess standard interfaces and behaviors. We implement a Grid service workflow execution engine in the OGSA framework, for processing chemical evolving research workflows specified in GWEL document. The Grid service workflow execution engine itself is a high-level Grid service, hence automatically Grid aware, that can be used within any GT3 environment. Through this exemplification, we show that the sharing and collaboration mechanism provided by Grid services and Grid service workflow could be beneficial for modeling and managing scientific processes of experimental investigation, evidence accumulation and result assimilation. Scientific processes themselves can then be reused, shared, adapted and annotated.

Acknowledgement We thank the researchers in Network Technology & Applications Research Laboratory, CNIC of CAS for their valuable discussions and comments.

References:

[1] Foster I, Kesselman C, Tuecke S. The anatomy of the grid: Enabling scalable virtual organizations. Int'l Journal of Supercomputing Applications, 2002,15(3):200−222.

[2] Foster I, Kesselman C. The Grid: Blueprint for a Future Computing Infrastructure. Morgan Kaufmann Publishers, 1998.

[3] Foster I, Kesselman C, NickJ, Tuecke S. The physiology of the Grid: An open Grid services architecture for distributed systems integration. DRAFT document, 2002. http://www.globus.org/research/paper/ogsa.pdf

[4] Cybok D. A grid workflow infrastructure. In: Proc. of the Workflow in Grid Systems Workshop in GGF10. Germany, 2004. http://www.extreme.indiana.edu/groc/ggf10-ww/grid_workflow_infrastructure__dieter_cybok/GWI.pdf

[5] Press WH, Teukolsky SA, Vetterling WT. Numerical Recipes in C, The Art of Scientific Computing. 2nd ed., Cambridge: Cambridge University Press, 1993.

[6] Johnson DRH, Soderblom DR. Calculating galactic space velocities and their uncertainties, with an application to the Ursa Major group. Astronomical Journal, 1987,93(4):864−867.

[7] Krishnan S, Wagstrom P, von Laszewski G. GSFL: A workflow framework for Grid services. In: Proc. of the SCC2002. Baltimore, 2002. http://www-unix.globus.org/cog/papers/gsfl-paper.pdf

[8] Govindaraju M, Krishnan S, Chiu K, Slominski A, Gannon D, Bramley R. Merging the CCA component model with the OGSI framework. In: Proc. of the 3rd Int'l Symp. on Cluster Computing and the Grid (CCGrid2003). 2003. http://www.extreme.indiana.edu/xcat/publications/ccgrid03_xcatogsi.pdf

[9] van Laszewski G, Amin K, Hategan M, Zaluzec NJ, Hampton S, Rossi A. GridAnt: A client-controllable Grid workflow system. In: Proc. of the 37th Hawaii Int'l Conf. on System Sciences (HICSS-37). Hilton Wailkoloa Village, Island of Hawaii, 2004. http://www.mcs.anl.gov/~gregor/papers/vonLaszewski--gridant-hics.pdf

[10] Tuecke S, Czajkowski K, Foster I, Frey J, Graham S, Kesselaman C, Maguire T, Sandholm T, Vanderbilt P, Snelling D. Open grid services infrastructure (OGSI) version 1.0. Technical Report, Open Grid Services Infrastructure WG, Global Grid Forum, 2002.