

# An Introduction to Scientific Data Grid

LUO Ze, NAN Kai, YAN Bao-Ping

Computer Network Information Centre/Chinese Academy of Sciences

## Abstract

Scientific Data Grid is an application Grid which aims at scientific data resources sharing and collaboration, and supporting scientific application based on Grid technologies. This paper introduces its background, system platform, middleware and some scientific applications supported by this application Grid.

## 1. Background

As China's natural science research centre, Chinese Academy of Sciences (CAS) has produced and accumulated a great store of scientific data and materials in its long history of scientific research and practice.

In 1982, Chinese Academy of Sciences proposed the program of "The Scientific Database and Information System", which was intended to integrate the scattered databases of different specialties for sharing through utilizing the ever-developing computer, database and network technologies.

Through two decades continuous development, the Scientific Database (SDB) has already become the most characterized scientific database resource on China Science and Technology Network (CSTNET). It provides scientific data service to scientific research, national macro decision-making, as well as to the public.

Scientific Data Grid (SDG) is an application grid which aims at scientific data resources sharing and collaboration. It integrates different resources in informatization environment of scientific research, i.e. scientific data and computing capacity for data analysis and process, connect more than 40 institutes under Chinese Academy of Sciences via data resources in SDB, realize effective sharing of distributed and heterogeneous data resources by applying Grid technology, especially data Grid technology, and develop some application systems that have practical importance for scientific research.. We want to resolve following key problems through the research of SDG:

1. How to access large scale, distributed and heterogeneous scientific data uniformly, promote convenient sharing of scientific data resource and enhance efficiency and utility of sharing data resource.

2. How to integrate heterogeneous databases

- by metadata technology, implement sharing and service of relative information by Grid information service. Further, how to make advanced application systems based on Grid thinking and technology possible by way of combining metadata and information of data resource.

3. Via some application systems of special domain, provide Grid application framework of science research fields, explore main technical difficulties and problems in spreading Grid application of science research field and create elementarily a Grid application standard in some fields.

## 2. System Platform

The system platform for SDG consists of scientific data resources, network storage resources and computing resources.

By the end of October 2004, the SDB has established 388 databases of different specialties, and increased its gross data volume to 13TB, 7.7TB are available on the Internet, and 45 websites of different domains now provides on-line service with most of the data.

Storage resource includes 20TB network storage and 50TB tape system. SDG provides more than 1TFLOPS computing capability. Storage and computing resources are mainly provided by 59 nodes of the super data server, DeepComp 6800, situating at the data centre of Computer Network Information Centre under Chinese Academy of Sciences, as shown in figure 1.



Figure 1 DeepComp 6800

### 3. SDG Middleware

SDG middleware is composed of two parts, core services and application-oriented services. The architecture is shown in Figure 2.

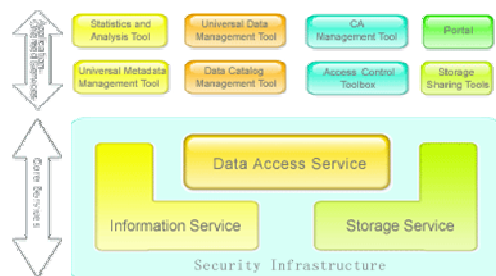


Figure 2 Architecture of SDG middleware

Information Service, Data Access Service, Security Infrastructure and Storage Service comprise the core services.

**Information Service.** On the basis of metadata built for Scientific Databases, the Information Service, including Information and Metadata Service (IMS) and SDG Finder, a resource finding tool, supplies information service for SDG and advanced application systems.

**Data Access Service (DAS).** DAS is designed to realize uniform access to massive, distributed, heterogeneous and autonomous databases. At present, we can access, via DAS, a wide range of relation databases, such as Oracle, Microsoft SQL Server and MySQL, and file systems. Through the interface provided by DAS, client can acquire metadata of data resource and execute query. The DAS is implemented by OGSA compliant grid services.

**Security Infrastructure.** Security Infrastructure implements primary functions of Certificate administration and access control. We implement software for constructing a Certificate Authority (CA) in a simple manner. CA is an entity in Public Key Infrastructure (PKI), which is responsible for establishing and vouching the authenticity of public keys.

**Storage Service.** Storage Service are made up of file storage service, database service and Internet publishing service, provides a series of storage service tools with the utilities of data transfer, storage management and quota assignment.

Application-oriented services include Statistics and Analysis Tool, Universal Metadata Management Tool, CA Management Tool, Access Control Toolbox, Storage Sharing Tools and Portal.

**Statistics and Analysis Tool.** Statistics and Analysis Tool is installed and deployed in data

centre and Institute that participated in SDG. According to the Interface provided by Statistics and Analysis Tool, we can get dynamically data volume information about data resource provided by particular organization. Data volume information could be processed and visualized to demonstrate the state of data resource. This tool is implemented by OGSA compliant grid service.

#### Universal Metadata Management Tool.

This tool is used for integrating metadata provided by different field. We adopt XML to exchange information among different modules of SDG middleware. This tool implements some management function for metadata, including add, remove and modify operation, and of course, supporting metadata query.

**CA Management Tool.** We provide a client-side tool, called CertUtility. This tool simplifies the integration and interaction between application and security infrastructure.

**Access Control Toolbox.** By using Access Control Toolbox, user can configure flexibly access right for given user, customize the mapping between account and role. The toolbox provides a way to control the user's access in a fine granularity manner. Currently, this toolbox supports RDBMS like Oracle, MySQL, etc.

**Storage Sharing Tools.** Based on open source software JFtp, we developed Storage Sharing Tools with two important enhancements. First, we enforce the security function and make data transport reliable. Second, these tools support quota assignment.

**Portal.** In our SDG Portal, we integrated grid service in the portlets. Every portlet service is compounded by one or more grid service. Portal has a few portlets which can provide service to users. Currently, the basic portlets have been developed.

After 3 years research and development, SDG middleware gained some important achievements. We released SDG middleware version 1.0 by the end of 2003, and released version 2.0 by the end of 2004. The software package was installed and deployed on Institutes that participated in SDG project after special annually training. One step of installation wizard of SDG middleware version 2.0 is shown in figure 3. The prototype of SDG now comes into being.



Figure 3 One step of installation wizard

#### 4. Applications

One of the primary goals of SDG is to develop and run scientific application based on Grid technologies, as an illustration of e-Science enabled by Grid technologies. In SDG, we currently support three domain applications: China Virtual Observatory; International Cosmic Ray Data Pre-processing Centre; and Chinese Herbal Medicine Virtual Academe.

**China Virtual Observatory.** In SDG, Computer Network Information Centre of CAS collaborates with National Astronomical Observatories of CAS to develop China Virtual Observatory as one of scientific application systems. Currently, services, including Statistical Analysis of Fe Abundances Gradients in the Galaxy, The Decoding Grid Service and Query Grid Service for some catalogue, DSS image retrieval grid service, and Basic Astronomical Computing Service, have been set up. Based on layered GRID infrastructure, China Virtual Observatory mainly addresses following three tasks: (1) astronomical data interoperation; (2) spectrum auto-process; (3) VO-enabled LAMOST. LAMOST, means Large Sky Area Multi-Object Fibre Spectroscopic Telescope, is a meridian reflecting Schmidt telescope, using active optics technique to control its reflecting corrector makes it a unique astronomical instrument in combining large aperture with wide field of view. LAMOST is shown in figure 4.



Figure 4 LAMOST

**International Cosmic Ray Data Pre-processing Centre.** YBJ International Cosmic Ray Observatory is located at 90°26'E and 30°13'N in Yangbajing (YBJ) valley of Tibetan highland. The ARGO -YBJ Project is a Sino-Italian cooperation started its detector installation in 2000. It aims at the research of the origin of high energy cosmic rays. It explores the approximately 100 GeV uncultivated land and measuring the antiproton/proton ratio by cosmic ray moon shadow. The ARGO-YBJ project will be full operational in 2007 and will generate more than 200TB of raw data each year. The raw data will be transferred from Tibet to Beijing Institute of High Energy Physics and processed in to reconstructed data. The physicists will work on the reconstructed data for physics researches. For this purpose a grid based computing system will be built with about 400 CPUs, mass storage system and broad band network links among Tibet, Beijing and institutes in Italy. Cosmic ray air-shower array detectors installed on YBJ International Cosmic Ray Observatory are shown in figure 5.

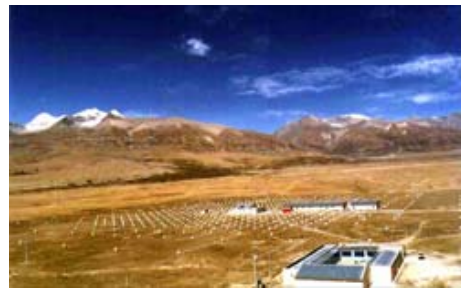


Figure 5 Cosmic ray air-shower array detectors

**Chinese Herbal Medicine Virtual Academe.** Based on databases of Chinese herbal medicine information distributed around China, Chinese Herbal Medicine Virtual Academe constructs a Chinese herbal medicine application grid, which implements interconnection and interoperability of Chinese herbal medicine information databases and high degree sharing of Chinese herbal medicine resources, supports the scientific research of Chinese herbal medicine and pushes the process of Chinese herbal medicine modernization.



Figure 6 Chinese traditional medicines

## **5. Conclusion**

This paper is a brief introduction to SDG. We describe its background, system platform, middleware and some scientific applications supported by this application Grid. After 3 years research and development, SDG gained some important achievements. The system platform is almost accomplished. We released SDG middleware version 2.0 by the end of 2004. The research and development of scientific applications on SDG is ongoing. The prototype of SDG now comes into being. We are now working hard for next version of SDG to make it stronger, more stable and more user-friendly.