DATA MANAGEMENT SERVICES FOR ENGINEERING

Professor SJ Cox, Dr Z Jiao, Dr JL Wason
School of Engineering Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.

Key words to describe the work: Grid data management, Grid services, ontology applications, XML Schema, data access and visualization, design search, optimisation, Geodise.

Key Objectives: The objective of the data management in Geodise is to provide: data storage and replication services, metadata management service, secure and transparent data access services, data resources for visualization and data mining services.

Motivation for the work: To manage the large volume of data created by engineering design and optimisation processes running in the Grid environment, and to provide applications with data storage, access, and visualization services.

Abstract

Geodise [1] is a Grid-enabled, intelligent engineering design and optimisation search environment that provides engineers with a central point of access to a suite of tools that will enable them to improve designs, keep track of running and past jobs, and obtain advice. The service-based architecture consists of a number of Grid services providing design optimisation and search tools, CFD analysis applications, computation, knowledge and data resources.

Engineering design and optimisation is a computationally intensive process and data may be generated at different locations with different characteristics and must be stored for further analysis and post-processing. Databases therefore play an essential role in our architecture where it is important to capture the process of how results are obtained in addition to storing the results themselves, and both types of data need to be accessible to a large community.

Whilst other Grid projects have some similar components and requirements [2][3][4], our focus is on providing data management in an engineering environment where all the data generated by optimisations and simulations are traditionally stored in flat files with little descriptive metadata provided by the file system. When there are a large number of files it becomes difficult to find, compare and share the data. The Geodise project leverages existing database tools which are not commonly used in engineering and makes them accessible to users of the system. Various services are provided for data storage, access, query and transfer, and schema generation, allowing engineers to concentrate on their application problems without needing to be aware of specific low-level data management mechanisms. When result data is generated and analysed some of it will be retained for a short time and then discarded, whereas process information about how the data was generated may be retained for a long time and may be reused by others or used to regenerate the data at a later date.

Effective management of metadata can help make large quantities of engineering data more accessible and easier to locate. Although existing Database Management Systems do not yet offer Grid integration 'out-of-the-box', they provide a wide range of important functionality for Grid applications. We require relational and XML databases for managing different types of data (e.g. standard metadata and complex, changing engineering metadata). We therefore require a set of services that allow us to access and interrogate both types of data storage in a standard way.

We adopt a service approach for database integration into a Grid environment, providing other Grid applications with a well-defined interface for accessing and archiving data. The Data Access and Integration Services Working Group of the GGF [5], in which we are participating, are developing requirements, functionalities and standards for Grid Database Services in the Open Grid Services Architecture [6][7]. Each of our Geodise specific data management services will communicate with the underlying databases through such services. Other projects are specifically tackling relational databases [8] and XML repositories [9] and we will follow these closely and use implementations that follow the proposed standards.

We now further describe data management services currently provided by Geodise shown in Figure 1.

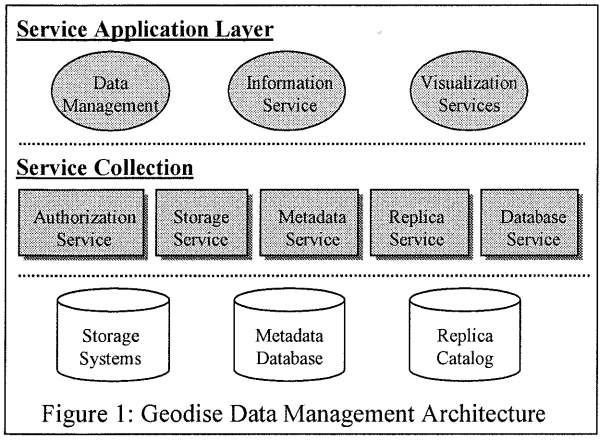

Figure 1: Geodise Data Management Architecture

- Storage service - Allows applications to archive data (e.g. result sets, log files, images) sent over GridFTP in file systems curated by Geodise for benefits of accessibility by a larger community (via authorization), storage capacity, and a uniform query interface.

- Metadata service - The data can be stored and retrieved based on additional descriptive information detailing technical characteristics (e.g. location, format), ownership, versioning, context [10] and application domain specific metadata. For the latter, applications first provide an XML Schema which is used to construct a storage schema, validate data before insertion, and build generic query interfaces.

- Authorisation service - Access rights to data can be granted to an authenticated user based on affiliation, roles, and organisational hierarchy, all of which are described by the Geodise user profile ontology. A user may grant and revoke access privileges on their data, for example specifying that it may only be viewed by them or members of a particular group, or used in aggregate queries.

An example of how these services are used during a typical Geodise session is now described. An engineer logs on to the Geodise portal and is authenticated. They set up a CFD design optimisation problem and set it running. Data is created and stored at various points: the portal logs problem setup information, the user sends requests to save data in the Geodise data repository, intermediate information created inside and passed between Grid services may also be stored for further analysis and provenance purposes. These requirements can be met by our storage, metadata and database services, and by plugging in our logging tool which transparently records messages exchanged between services into databases based on WSDL [11] definitions. The optimisation may take days to complete, and the user can log in and check its progress and retrieve the final results. The metadata service is used to locate data of interest, providing permission is granted by the authorisation service. Queries can be expressed through the query facility provided by the database services, through a simple interface where the Grid application provides a list of restrictions of name-value pairs, or using a web based query interface and visualization service.

We plan to provide a replica service to copy files between sites on demand, making use of GridFTP, where information about logical file names and the physical locations is stored in a replica catalog. We will also investigate further the application of ontologies in engineering data management.

References

[1] Geodise project. http://www.geodise.org

[2] Storage Resource Broker (SRB), San Diego Supercomputer Center. http://www.npaci.edu/DICE/SRB/

[3] W. Allcock, A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, S. Tuecke. *The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets.* Journal of Network and Computer Applications, 23:187-200, 2001.

[4] European Data Grid Project Work Package 2: *Architecture and Design Document.*
http://grid-data-management.web.cern.ch/grid-datamanagement/docs/DataGrid-02-D2.2-0103-1_2.pdf

[5] Global Grid Forum http://www.gridforum.org/

[6] M.P. Atkinson, V. Dialani, L. Guy, I. Narang, N.W. Paton, D. Pearson, T. Storey and P. Watson. *Grid Database Access and Integration: Requirements and Functionalities.* 4 July 2002. http://www.cs.man.ac.uk/grid-db/papers/dairf.pdf

[7] I. Foster, C. Kesselman, J. Nick, and S. Tuecke. *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration.* 2002. http://www.globus.org/research/papers/ogsa.pdf

[8] B. Collins, A. Borley, N. Hardman, A. Knox, S. Laws, J. Magowan, M. Oevers, E. Zaluska, *Grid Data Services - Relational Database Management Systems (Version 1).* 5 July 2002. http://www.cs.man.ac.uk/grid-db/papers/grdb.pdf

[9] A. Krause, K. Smyllie and R. Baxter, *Grid Data Service Specification for XML Databases* 19 June 2002. http://www.cs.man.ac.uk/grid-db/papers/GXDS-spec-1.0.pdf

[10] D. Pearson, *Data Requirements for the Grid: Scoping Study Report*, 2002. http://www.cs.man.ac.uk/grid-db/papers/Requirements.pdf

[11] World Wide Web Consortium (W3C). 2001. *Web Services Description Language (WSDL) 1.1* W3C Note http://www.w3.org/TR/wsdl/