DIGITAL LIBRARY GRID : A ROADMAP TO NEXT GENERATION DIGITAL LIBRARIES USING GRID TECHNOLOGIES.

Hardik Joshi J C Jakharia

Abstract

In the current scenario, data is increasing rapidly. Chunks of data is being accumulated resulting in knowledge generation. Access to global literature, books, articles require efficient data management and querying techniques. The consequences associated are requirements of very large storage resources, complex queries, interoperability and scalability across global environment. The Integration of grid, data grid, digital library solves various issues related to the upcoming globalization of digital libraries. In this paper, we propose a Grid based digital library concept & examine the synergies between these data management systems, which would help in future evolution of digital libraries.

Keywords : Grid Computing, Data grid, Digital Library, Information management.

1. Introduction

Digital Library (DL) research is that field of research and development aiming to promote the theory and practice of processing, dissemination, storage, search and analysis of various digital data. Digital Libraries acting as knowledge depositories can be considered as complex information systems, development and use of which require solution of numerous scientific, technological, methodological, economic, legal and other issues. [1] Digital Library technologies are rapidly developing. Challenges in semantics, integration of information, and perceptions of presentation of various kinds of data call for significant innovations.

Digital libraries are more and more actively coming into use within scientific organizations and universities. At present, one can hardly imagine a western university having no well-developed digital library. Digital libraries created on the basis of licensing agreements with publishing houses frequently displace subscriptions to print journals. At the same time, most university digital libraries functionally resemble traditional libraries or are combined with them to become “hybrid libraries”. In perspective, digital libraries should become the depositories of knowledge.

Now-a-days, with the growing size of digital libraries and integration of digital libraries, there are various challenges in this filed, some of them are :

- Resource discovery
- Standardization of Interfaces
- Digital Library Administration
- Copyright and Licensing
- Cost optimization etc…

 © INFLIBNET Centre, Ahmedabad

Resource Discovery among various heterogeneous data sources is a critical issue. The emergence of grid technologies suggests it may be possible to integrate all independent digital libraries within the global digital library. Data grid technology is an example of the federation of completely independent digital libraries into a common name space with common access mechanisms. This strongly suggests that it will be possible to create global digital libraries, in which differing technical, legal, and policy issues are overcome.

2. Grid Technologies

Grid computing uses the resources of many separate computers connected by a network to solve large-scale computation problems. Grid computing involves sharing heterogeneous resources (based on different platforms, hardware/software architectures, and computer languages), located in different places belonging to different administrative domains over a network using open standards. In short, it involves virtualizing computing resources.

Functionally, one can classify grids into several types:

- Computational Grids (including CPU scavenging grids), which focuses primarily on computationally-intensive operations.
- Data grids, or the controlled sharing and management of large amounts of distributed data.
- Equipment Grids which have a primary piece of equipment e.g. a telescope, and where the surrounding Grid is used to control the equipment remotely and to analyze the data produced.

A data grid is a grid computing system that deals with data—the controlled sharing and management of large amounts of distributed data. These are often, but not always, combined with computational grid computing systems. [2]

Many scientific and engineering applications require access to large amounts of distributed data (terabytes or petabytes). The size and number of these data collections has been growing rapidly in recent years and will continue to grow as new experiments and sensors come on-line, the costs of computation and data storage decrease and performances increase, and new computational science applications are developed.

Here, we propose a concept for an integration of various digital libraries. Grid computing allows digital libraries to be linked with one another through a grid environment. It also supports in searching across various digital resources.

Through any node that is linked to the Library Grid, a user would be given access to search for information on the materials that the user is looking for. After the material of information is located, Replica selection in the data grid technology would be applied to locate the node with the fastest download rate. Then the digitized file would be transferred to the user. In this way, the library grid would help in global resource sharing and searching the metadata across various heterogeneous resources.

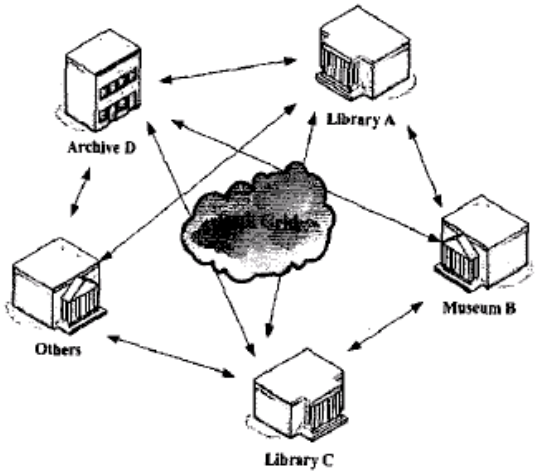

Figure 1 Conceptual View of a Global Digital Library

3. Requirements for a digital library system

The objective Is to create a software toolkit that could be used to set up a digital library according to the requirements of a given user community by instantiating the software appropriately and then explicitly submitting new documents or harvesting the content from existing sources.

1. There is a set of core DL functionalities, such as search, retrieval, access to information objects, that any DL should provide. The format in which each of these functionalities is presented to the user is usually different since it complies with the application specific vocabularies and rules. In addition to the core functionalities, each DL, usually, must provide other specific functionalities for serving the application-specific requirements.

2. New organizations may ask to participate to the DL during its lifetime and additional functionalities may be required to satisfy new needs. A DL mustbe able to dynamically evolve by adapting itself to these new situations.

3. The handling of a DL can be expensive in terms of financial, infrastructural and human resources. Many organizations are confident that this problem can be overcome by adopting a DL federated model. According to this model, multiple organizations can set up a DL by sharing their resources without losing, if required, control over their own resources. For example, they can store their information objects locally or host key services on their computers.

4. The users of a DL require a good quality of the service (QoS), i.e. an acceptable level of non-functional properties such as performance, reliability, availability and security.

5. Access to content and services is usually regulated by policies. These can specify, for example, that a collection of objects is only visible to a particular group of users, or that a set of services can only be used free of charge for a given time interval.

4. Grid Middleware & Data Grid

Grid computing enables the use of computation power, various resources and computational devices of isolated computer. When PCs or facilities are connected to the grid, other computers on the grid will search for components to begin processing and computing the related work. Hence apart from data repositories, computation power of idle computers can be harvested in Grid Libraries.

Data grid Architecture is meant to integrate the data storage devices and data management service into the grid environment. Data grid consists of scattered computing and storage resources, though located in different countries to remain accessible to users [3].

Various Grid middlewares are available to implement Grid environment like Globus, Legion, Unicore etc.. Globus being widely used and portable on open source systems, We have adopted the Globus Toolkit in our framework. It provides solutions such as security, resource management, data management and information service. [4]

Globus Data Grid comes in two layers. In the Low Level are the Data Grid Core Services, and the upper layer are High Level Components [5]. Figure 2 shows the Data Grid Architecture.

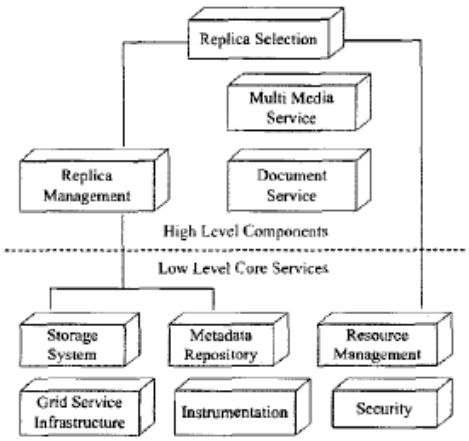

Figure 2 Data Grid Architecture

The storage system is a basic data grid component. Various file systems such as HPSS, DPSS can be considered [6]. Data access service is a set up of a mechanism for accessing, managing and transferring data in the storage system [7].

Metadata service is for managing and accessing Metadata, which contains Data Grid information. Metadata Application includes information describing the files and information on data environments. Replica Metadata is applied to manage replication of data objects.

Resource Management is responsible for storage system, networks and other data grid resources to assure end-to-end efficiency, technical assessment of the efficiency test, as well as crucial resources. Grid Security Infrastructure [8] provides environment authorization and certification mechanism to a large number of users.

Replica management [9] is important to the successful processing of large amounts of data in Data Grid. It is mainly to decide when and where to set up replica and it provides information on the location of Replica.

5. Proposed Framework

Service Oriented Architecture suits best to construct a federated DL. Grid Services and Web Services play a crucial role to implement interoperable digital libraries. The following listing shows some services, which must be provided by various components of a Library Grid.

Service name	Main performed tasks
Repository	Storage and dissemination of documents that conform to a documentmodel termed DoMDL [14]. These documents can be structured, multilingual and multimedia.
Multimedia Storage	Storage, streaming and downloading of video manifestation of a document, dissemination of videos either as whole documents or as aggregations of scenes, shots and frames.
Library Management	Submission, withdrawal and replacement of documents. It is configurable with respect to the metadata formats accepted.
Index	Document retrieval parametric w.r.t. the metadata format, the set of indexed fields, the result set format and the query terms language.
Query Mediator	Document retrieval by dispatching queries to the appropriate Indexservice instances and by merging the result sets, taking into account the peculiarities of available Index instances.
Browser	Construction and use of appropriate data structures, termed indexes,for browsing the library content, parametric w.r.t. the metadata formats, the set of browsable fields, and the result set format.
User Interface	Mediations, among human user and application services.

6. System Design & Implementation

In Grid Organization, Grid Portal serves as a communication with the grid members and its main function is to provide the user with a Web-based interface for the use of the services and resources on the grid [10]. A grid portal allows communication between the outside world and the grid itself.

There are several technologies in the constructions of the grid portal. GridSphere [11] has the Tomcat Application Server as its development platform and uses Java technology. Another grid portal technology is GridPort [12] which is a Perl-based API for Unix systems.

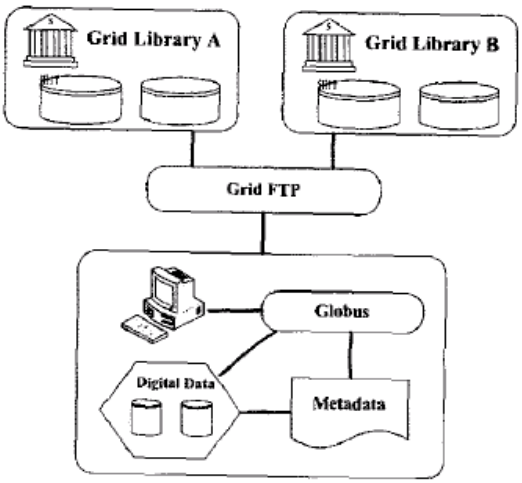

Figure 3 Architecture of Digital Library Grid

In our framework, Globus toolkit works as a Grid middleware whereas Grid Portals are used to disseminate the Grid services among the end users. This setup helps to solve the challenge of inconsistent softwares used among various digital libraries. Using Grid technologies like Globus and GridSphere we can bring all the DL to a unified schema. Although there are some assumptions residing in this problem, the following steps help to designing a unified approach.

1. Each member of the Grid (Library or museum) must set up a basic Grid environment, such as installing Globus Toolkit and software pack of the Data Grid.
2. Each member has to digitize the data to be shared in an acceptable format of the Grid.
3. The digitized data shall be stored in an individual storage system and Metadata file with the description of these data stored in Metadata Repository, before starting the Replica Management Mechanism. ISBN will serve as a key index, and the data framework would then be set up by using XML technology.
4. GridFTP [13] will be used as a protocol for data transfer. This will help in Linking the Grid.
5. Integration of the Grid Portal and Grid Organization(s) running Globus Toolkit.

Benefits :

Compared with ordinary digital libraries, the following benefits are available from Library Grid :

- All organizations which intent to become a member of the Library Grid Organization has to set up only the required Grid Middleware and Grid Portal services.
- As Digital Library Grid is built on Data Grid technology by using Data Grid's Replica management framework its users are allowed to locate the information they need efficiently and reduce the time of data transfer.
- Issues related to Security are handled by GSI, which need not to be considered.
- Load Balancing and Efficient resource utilization and Network management is carried out by Grid Middlewares, which assists in better administration.

7. Conclusion

In this paper we have proposed Digital Library Grid, which solves certain issues related to interoperation of various digital libraries. Apart from which, it also enhances the abilities of Library Grids to scale, efficient computation power harvesting of idle computers, efficient data storage and resource management, load balancing across various organizations in searches.

The proposed framework can be a future Roadmap to the Digital Libraries.

8. References

1 The Digital Library Concept, http://www.dinf.ne.jp/doc/english/index_e.html

2 Grid Computing, http://www.wikipedia.org

3 W. Hoschek, A. Samar, “Data Management in an International Data Grid Project”, IEEE/ACM International Workshop on Grid Computing, Dec. 2000.

4 Introduction to Grids and Globus Toolkit,http://www.globus.org/toolkit/about.html

5 Chervenak, I. Foster, “The Data Grid : Towards an Architecture for the Distributed Management and Analysis of Large Scientific datasets”, Journal of Network and Computer Applications, Volume 23: pp187-200, 2001.

6 Getting Started with the Globus Replica Catalog,http://www.globus.org/datagrid

7 S. Vazhkudai, I. Foster, “Replica Selection in the Globus Data Grid”, Cluster Computing and the Grid, 2001.

8 Globus : Grid Security Infrastructure (GSI),http://www.globus.org/security

9 Globus : Replica Management,http://www.globus.org/datagrid/replica-management.html

10 M.E. Pierce, G.C. Fox, “The Gateway computational Web Portal”, Concurrency and Computation : Practice and Experience, Vol 14, pp 1411-1426, 2002

11 GridSphere Portal,http://www.gridsphere.org/gridsphere

12 M.Thomas, S. Mock, “The GridPort toolkit : A system for building Grid portals”, Proceedings of the 10th IEEE International High Performance Distributed Computing Symposium pp216-227, Aug 2001

13 GridFTP : Universal Data Transfer for the Grid,http://www.globus.org/grid_software/data/gridftp.php

14 F. Berman, Geoffrey, “Grid Computing – Making the Global Infrastructure a Reality”, 2002 John Wiley & Sons Ltd, pp 9-49 ISBN : 0-470-85319-0.

About Authors

Mr. Hardik Joshi is working as a Lecturer with the Department of Computer Science, Gujarat University – Ahemedabad. He is MCA, M.Phil in Computer Science. His area of interests include Knowledge Management using Grid Technologies. He has presented 5 papers at National Level and a paper in Caliber 2004.
Email :

Mr. J. C. Jakharia is working as a Senior Lecturer with G.K. & C.K. Bosamia College Jetpur affiliated to Saurashtra University – Rajkot. He is M.Sc. (Statistics) and teaching since past 12 years. He has presented 3 papers at National Level.