International Journal for Multidisciplinary Research
Plagiarism & Intellectual Property Rights in Digital Age (PIPRDA-2023)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Improvement of Accessibility of Digital Libraries Through Grid Technology

[1]Dr. Nutan Joshi, [2]Raghvendra Tripathi

[1,2]Librarian, Government KRG PG Autonomous College, Gwalior (M.P.)

Abstract:
The integration of grid and digital library technologies has resulted in software infrastructure that is perfectly suited to the generation and management of data. Data grid provides support for the organization, management, and application of processes in the digital library and in managing the digital entities in cyber environment. To build an effective digital library librarian need vast amount of storage space and have to invest large amount in buying storage oriented hardware's. The emerging trends on grid technology brought new solutions for improved storage and access of information in digital library environment.

Keyword: Data grid, Grid technology, Digital library.

1.1 Introduction:

Libraries have constantly been seeking newly emerging way (s) to promote their library services. Libraries all over the world are undergoing drastic transformation. As Internet rises and knowledge explodes, the concept of digital libraries has been extensively accepted in the developed and developing countries and traditional libraries are becoming hybrid libraries and new libraries that are being set up are increasingly of the digital kind. The growing size of different types of digital libraries and integration of them has various challenges. Some of them are: (i) resource discovery, (ii) standardization of interfaces, (iii) digital library administration, (iv) copyright and licensing, and (v) cost optimization. Access to global literature, books, and articles require efficient data management and querying techniques. Data mining, text mining and grid based tools within digital library architecture have been shown individually in the past to enhance existing content and services. The advancement of data and text mining techniques produced useful and accessible tools with which to implement or improve digital library services, including advanced document clustering and automated metadata extraction as a means to improve discovery. The natural language processing techniques used in text mining may also be applied to text and data sources in large scale digital libraries and repositories. The 21st century can be described as the era of innovative technology. The advancement in the field of ICT is unprecedented, but there are other issues just as important that libraries can identify - issues that will require libraries to review their management capabilities [1].

1.2 Grid Technology:

Grid technology has its roots in distributed computing which is developed in an effort to generate processing power for meeting workload challenges. In order to boost processing power, institutions aggregated computing resources across locations or across the entire institution. The idea was to match the supply of processing cycles with the demand created by applications. This concept is now a ubiquitous solution practiced by leading organizations around the world. It ensures continuous computing availability despite scheduled maintenance, power outages, and unexpected failures. The same idea of sharing resources has paved the way for grid computing - a term coined in the mid-1990s. The grid approach uses untapped processing cycles from across geographical boundaries, much like distributed computing but with a far wider scale and scope. Grid computing, in effect, provides a global reach to distributed computing. It promises lower total computing costs along with on-demand, reliable, and inexpensive access to the vast, available computing resources that would otherwise go unused. Grids and data grids are complementary technologies that together enable the creation and management of data. The integration of information management is one of the next steps in the evolution of grid technology[2].

Grids do not sit in isolation from rest of world. Grids need to interoperate with the data access and management technologies that are being developed by other communities. An example is the interchange between the Grid community and semantic web community on the meaning and application of web services. Multiple data management systems are being developed to organize digital entities for use by research communities. Examples include digital libraries, which focus on publication and discovery mechanisms, persistent archives, which focus on preservation and technology evolution management, and data grids, which focus on interoperability across distributed resources. Each system provides a logical name space for referencing digital entities. Each system provides services that operate on the digital entities (content). Each system maps distributed state information that results from the services (context) to the logical name space. Each system manages consistency requirements that define how the state information from different services should be updated.

Grid technology can be divided into three groups. The first is data grids, which are grid environments designed to process huge data samples. An example of this is a particle physics experiment that requires the analysis of millions of particle collisions. The second group is the computing grid where the focus is on the execution of parallel algorithms rather than on the distributed processing of huge amounts of data. Finally, the concept of an application grid addresses the set of services available on sites distributed throughout the grid. The original grid environments assumed that applications directly accessed remote data that was stored under the user's ID, that data would be pulled to the computation, that accesses could be based upon physical file names, and that the applications would access data through a library interface. Generalizations now exist for each of these functions, typically implemented as naming indirection mechanisms.[3] Digital libraries and persistent archives focus on the management of the data those results from the application of services. They define a context that includes the state information that results from all processes performed upon a digital entity, and organize the digital entities into a collection. If grid technology is going to be used to support end-to-end data management applications, grid technology will need to implement similar functions. The Grid provides workflow processing systems that specify each service that is applied to a digital entity. Control mechanisms are applied to the services to specify their completion status. A dataflow environment focuses on the digital entities, and applies control mechanisms to the digital entities that are processed by the services. An example of a dataflow environment is the execution of a query on a collection, then the processing of the result set. The processing status of each digital entity in the result set is maintained [4].

A workflow environment typically knows in advance the names of the digital entities and the processing steps that will be applied to each digital entity. The dataflow environment allows the digital entities to be identified as part of the dataflow and provides controls to allow looping, conditional execution, and branching based upon the results of each service. The output from the dataflow may be stored in a collection or consumed by another dataflow or a device (such as video streaming). The collection context includes the state information that is generated by the application of the services. The design of appropriate dataflow control mechanisms is an integral part of access to distributed data. Operations may be performed more efficiently at a remote storage system when the result is the movement of a smaller amount of data over the network. Processing mechanisms have been incorporated into data access by the database community.

Equivalent functionality will need to be supported by the grid community to improve performance. The grid will need to support application of processes at remote storage locations.

Data grids provide generic data management infrastructure that can be used to implement records preservation environments. Data grids excel at managing reproduction of records across multiple sites, essential for mitigating against risk of record loss. Federations of data grids provide the essential capabilities needed to implement high security environments for preservation of authentic copies. Data grids are being used today to manage aggregations that total hundreds of terabytes of data and tens of millions of files. The management of data has traditionally been supported by software systems that assume explicit control over local storage systems (file systems) or that assume local control over information records in databases. The

mechanisms provided by data grids to manage access to heterogeneous data resources can also be used to manage migration from old systems to new systems, and hence manage technology evolution [5].

1.3 Data Grids:

Data grids are software systems that provide generic software infrastructure for managing distributed records (and records metadata), and are used to support all types of data management environments. The capabilities provided by data grids are essential for automating preservation processes, mitigating risk of data loss through reproduction of digital components, assuring the permanent association of identity and integrity metadata with records, and supporting retrieval and access. At the same time, data grids are designed to manage digital entities stored in any type of storage system, while providing access through a very wide variety of access mechanisms. This ability to interact with multiple types of storage systems and access systems forms the core of data grid support for technology evolution.

Data grids manage changes in software and hardware systems and simplify the incorporation of new technology into a preservation environment. They make it possible for a trusted custodian to take advantage of advances in technology without risk to the authenticity of the records. In the software components of a data grid there are five principal layers. In the bottom layer are the storage systems where the digital components actually reside. Data grids provide a standard mechanism for interacting with the storage systems, that is, a standard set of operations that can be performed upon the digital component(s) registered into the logical file name space. Typical operations include the ability to read and write a file in the storage system, and to manage latency over wide area networks. When manipulating a thousand digital components, it is much faster to register and move them using bulk operations than it is to issue commands one at a time for each component. The bulk operations are also required for scalability, to ensure that the digital material that is preserved can grow to tens or hundreds of millions of digital components while providing good interactive response. Current data grid technology is capable of storing digital components in Unix file systems (Linux, Solaris, Mac OS), Windows file systems, binary large objects in databases, disk/tape-based storage systems etc. If a new storage technology becomes available, a new storage system access mechanism is written that understands how to interact with the storage system and map from the standard set of operations supported by the data grid to the operations provided by the storage system.

Data grids store the logical attributes in a database. To ensure that the data grid can take advantage of new database technology, a standard database control mechanism is used to define the set of operations required to manage sets of metadata in a database. The standard operations maintain intact each metadata set and preserve its link to the related record. Data grids manage ownership by storing all data under a Unix user ID that is created for the data grid itself. Only the data grid has the ability to authenticate access to the remote storage system. This means that all accesses to the digital components must be done through the data grid software. This in turn means that the data grid can track all operations that are performed, maintain audit trails of all accesses, and automatically update the storage metadata when a digital component is moved to another location.

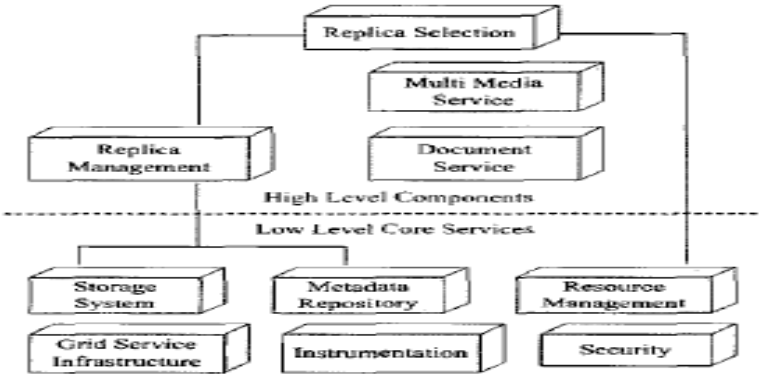

Figure 1 : Data Grid Architecture(source: Joshi and Jakharia)

Data grids support the concept of multiple access roles. A trusted custodian can be assigned access privileges that permit the execution of preservation processes, while the public is only given read permission

to selected digital components. The access permissions can be set separately for each digital component. Access permissions can also be set for each set of logical attributes. An implication is that the data grid must provide access controls for each digital component that is registered. Since the data grids manage the logical name spaces for users as well as digital components and logical attributes, the data grid can implement access controls that are permanent over time. Data grids effectively build a characterization / representation of the archival aggregation that is managed independently of the choice of storage technology.

1.4 Digital Library and Grid technology:

Nowadays digital libraries have become the source of information, sharing across the globe in the fields of education, research and knowledge. The full usage of digital libraries will be realized only when people can have access to the material from any location. The advantage of multimedia is that people of all ages can understand more clearly by seeing or hearing rather than reading. Considering the exponential growth in various technologies, developing a multimedia digital library in wireless is not an complicated task. Grid computing enables the virtualization data resources, process network bandwidth and storage capacities to create a single system image granting the user a seamless access to vast IT capabilities. By adopting peer-to-peer overlay networks, which are taking a central position in information systems, the storage space problem can be solved and by using grid computing the security can be maintained.

The digital data is stored in a cluster built of commodity components and users can access those data from anywhere, anytime securely. Advantages of this framework are: (i) existing capital investments are used for storing the multimedia files, (ii) increased access to data, (iii) balancing workloads among different systems connected to nodes, (iv) authenticated and secured transfer of files. Benchmarks used to test this framework are: (i) file size vs. download time, (ii) simultaneous connections, (iii) band-width utilization, (iv) security, (v) scalability, and (vi) robustness. The integration of data grids and digital libraries is forcing continued evolution of grid technology. Grid software focuses on services and execution of services. Digital libraries focus on management of results of services. Grids have been evolving through the addition of naming indirection mechanisms. The ability to manage information context will require further evolution of grid technology and the ability to characterize the assertions behind the application of the grid name spaces. The result will be the ability to manage the consistency of federated data collections while flowing information and data from digital libraries through grid services into preservation environments.

In the digital library community, the information is mapped onto a logical identifier that is associated with each digital entity, and organized as a collection of digital entities. The choice for which collection to assemble is independent of the set of services that was applied to the digital entities. The collections assert relationships between digital entities by providing support for metadata attributes. The digital library community constructs union catalogs to federate access across collections. An equivalent federation mechanism is needed to federate all of the logical name spaces managed by data grids, including name spaces for files, users, and resources. Federation extends the original naming indirection mechanisms developed for grids to support access across independently assembled collections of computational results.

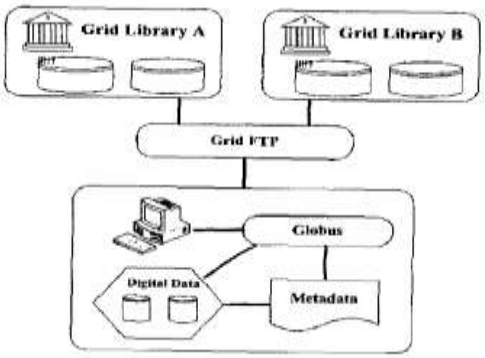

Figure 2: Architecture of Digital Library Grid (source: Joshi and Jakharia)

Digital libraries organize information in collections. The integration of data grid and digital library technology has resulted in software infrastructure that is suited to the generation and management of data.

Grids provide support for the organization, management, and application of processes. Data grids manage the resulting digital entities. Digital libraries provide support for the management of information associated with the digital entities. The grid environment and digital libraries have inter-related benefits and challenges [6]. The digital libraries have continued to grow and are now at the point where data grid-based storage solutions are becoming of great relevance. Data grids are able to provide seamless access to storage, geographically distributed to different locations, and stored in different types of repository. Files are identified by a logical identifier, rather than a file name; a system very familiar to digital libraries in terms of handles, digital object identifiers and so on. Preservation for digital records are successful when they can separate the digital record from any dependence on the original creating infrastructure. Data grid technology, which supports the management of distributed records, provides the software needed for infrastructure independence.

1.5 Grid technology and Knowledge management:

A major research issue in data grids and digital libraries is the integration of knowledge management systems with existing data and information management systems. Knowledge management is needed to support constraints that are applied in federation of data grids and in semantic cross-walks between digital libraries. Data grids can be viewed as systems that manage and manipulate consistency constraints on mappings of distributed information. Digital libraries add mappings to manage user-defined metadata to support discovery and browsing. Grid technology has the potential to improve the accessibility of digital libraries. The Integration of grid, data grid and digital library solves various issues related to the upcoming globalization of digital libraries. Joshi and Jakharia[7]proposed a Grid based digital library concept and examine the synergies between these data management systems, which would help in future evolution of digital libraries. In this Internet era accumulating knowledge has increased to a tremendous size. Accessing books on every field are continually increasing in large quantities.

Digital libraries are essential technology for the organization of data into collections. The integration of digital libraries with data grids provides the mechanisms needed to manage the massive amounts of distributed data that are now being generated. The federation of independent digital libraries is being driven by the desire to retain local control over collections, while supporting international access [8]. The result is the need to define management control and consistency update constraints that can be imposed between digital libraries. Life-cycle management (DSpace) and knowledge management (Fedora) technologies provide mechanisms to specify management controls and consistency constraints. The integration of these emerging digital library technologies with federated data grids promises to provide the data management infrastructure needed to support international collaborations [9].

1.6 Conclusion:

The information and communication technology has changed the complexion of today's libraries on a large scale. In developing a digital library, librarians will have a hard task to do. They will need the help of technologies to understand better the possibilities being created by digital technologies, and technologists will again need the help of librarians to make the process a successful one. The major component missing from grid technology is the ability to maintain consistency between content and context when multiple services are invoked. The challenge in managing information flow is that coordination is required between services. When result sets are manipulated instead of individual files, it may be appropriate to do processing at the remote storage location for some of the digital entities, but may be more efficient to move another member of the result set to a compute node for processing. Information flow imposes a generality of solution that is not available with current grid technology. A fundamental change for grids is the ability to define a context that can be managed independently of grid services.

Recently there is increasing interest in the re-application of methodologies and ideas from related areas of information access, organization, storage and retrieval within the digital library context. Today's information services offer lots of power and visibility to the libraries. It helps the user community without compromising on their quality and performance too. Developments in media related technologies like format migrations and resource intensive maintenance has added complexity and complications in building

effective digital library. A Data Grid that maps from access mechanisms provided by grids to management functions provided by digital libraries. Grid technology has evolved considerably over the last four years through the development of naming indirection mechanisms. The result has been the creation of an environment that manages web service execution independently of the underlying infrastructure components. An evolution of grid technology is needed to provide the capabilities required by digital libraries.

References :

1. Developing a Grid-Based Search and Categorization Tool", High Energy Physics Libraries Webzine, issue 8, October 2003. URL: http://library.cern.ch/HEPLW/8/papers/1/ (accessed on 01-03-2013)
2. A. Rajasekar, M. Wan, and R. Moore, "mySRB and SRB, components of a data grid," presented at the 11th High Performance Distributed Computing Conf., Edinburgh, U.K., 2002.
3. R. Moore, "Knowledge-based grids," presented at the 18th IEEE Symp. Mass Storage Systems and 9th Goddard Conf. Mass Storage Systems and Technologies, San Diego, CA, 2001.
4. R. Moore and A. Rajasekar. (2003) Common consistency requirements for data grids, digital libraries, and persistent archives. http://www.sdsc.edu/dice/Pubs/Moore-HPDC (accessed on 7-02-2013)
5. A. Rajasekar,M.Wan, R. Moore,W. Schroeder,G. Kremenek, A. Jagatheesan, C. Cowart, S.-Y. Chen, and R. Olaschanowsky, "Storage resource broker—Managing distributed data in a grid," J. Comput. Soc. India, vol. 33, no. 4, pp. 41–53, Oct.–Dec. 2003.
6. 6.Diligent Project, "A Digital Library Infrastructure on Grid Enabled Technology", http://www.diligentproject.org (accessed on 7-04-2013).
7. H Joshi, JC Jakharia, Digital library grid: a Roadmap to Next generation digital libraries using grid technologies. 4th International Convention CALIBER-2006, Gulbarga, INFLIBNET Centre, Ahmadabad (2006).
8. Larson, R. R., and Sanderson, R. "Grid-Based Digital Libraries: Cheshire3 and Distributed Retrieval" In Proc. 5th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2005), pages 112-113, Denver, USA, 2005.
9. Robert Sanderson and Paul WatryIntegrating Data and Text Mining Processes for Digital Library Applications JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.