Real Experiences with Data Grids - Case studies in using the SRB

Arcot Rajasekar, Michael Wan, Reagan Moore, Arun Jagatheesan, George Kremenek

San Diego Supercomputer Center (SDSC), University of California at San Diego

{sekar, mwan, moore, arun, kremenek}@sdsc.edu

ABSTRACT

In scientific communities, Data Grids are becoming increasingly important for sharing large data collections in collaborative environments. In this paper we describe the use of the SRB, a data grid infrastructure, in building collaborative environments for large-scale data. We describe our experiences with generating Digital Simulation Videos for the Hayden Planetarium, Distributed Data Management for the Visible Embryo project, sorting and storing large number of sky images for the 2MASS Digital Sky project and the experience with building the Biomedical Informatics Research Network.

Keywords

Data Grids, Grid Computing, High Performance Storage and Data Management, Data intensive computing

INTRODUCTION

Data Grids are becoming increasing useful for managing large-scale distributed data across multiple heterogeneous resources. Examples of data grids can be found in the physics community [1,2,4], biomedical applications [3] and for ecological sciences [7, astronomy [8], geography, and earthquake and plate tectonic systems [9]. The SDSC Storage Resource Broker (SRB) is a client-server based middleware that provides a facility for collection-building, managing, querying and accessing, and preserving data in a distributed data grid framework [10,11]. In a nutshell, the SRB provides the following data grid functionalities (see [6] for details).

- Manage metadata for creating data, including system-metadata, domain-specific and user-defined metadata.
- Handle heterogeneity of platforms, storage and data.
- Provide seamless authorization and authentication to data and information stored in distributed sites.
- Provide transparent access to resources and attribute-based access to data and collections
- Provide virtual organization structure for data and information based on a digital library framework.
- Handle and manage replication of data,
- Handle caching, archiving and data placement,
- Handle access control and provide auditing facilities,
- Provide remote operations for data sub-setting, metadata extraction, indexing, data movement, etc.

The SRB has been in use for the past four years in support of multiple projects. The following list provides a glimpse at the breadth of projects that currently using the SRB:

1. **Astronomy:** 2Micron All Sky Survey (2MASS), Hayden Planetarium project, National Virtual Observatory
2. **Earth-systems and Environmental Sciences:** HyperLTER Project, Land Data Assimilation System (LDAS), CEED: Caveat Emptor Ecological Data Repository, ROADNet
3. **Medical Sciences:** Visible Embryo Project
4. **Molecular Sciences:** SSRL Data Repository, Alliance for Cellular Signaling (AfCS)
5. **Neurosciences:** NPACI Brain Data Archiving Project, Biomedical Informatics Research Network
6. **Physics and Chemistry:** GriPhyN Project, GAMESS Portal, Babar Project
7. **Digital Libraries and Archives:** National Science Digital Library (NSDL), National Archives and Library of Congress, Univ. of Michigan Digital Library Archive
8. **Data Grids:** NASA Information Power Grid, NPACI Grid Portal Project, UK eScience Grid.

More information about these and other projects can be found at http://www.npaci.edu/dice/srb/Projects/main.html. In this paper, we describe our experience in using the SRB in four diverse applications. We picked these projects for discussion not only for their large-scale data integration needs but also for their diversity in highlighting the various functionalities of the SRB system. In particular they show the large data handling capability, data aggregation and movement features, parallel data transfer functionality, collaborative strengths and metadata management facilities inherent in the SRB.

HAYDEN PLANETARIUM PROJECT

Data Sizes: 7 TeraBytes

Number of Files: more than 10,000

Sites: NCSA, AMNH, SDSC, CalTech

Resources: 512 & 64 procs. Origin2000, SP-2

Data Storage: UniTree, HPSS, GPFS, Unix FS

Project Duration: 3 months (project concluded)

This data-intensive project for the Hayden Planetarium (http://www.haydenplanetarium.org/) at the American Museum of Natural History (AMNH) involved the

generation of a movie showing the formation of an emission nebula leading to the creation of our solar system.

The simulation was performed at NCSA, AMNH and SDSC with the largest being done at NCSA using 512 processors of an SGI Origin2000. The simulation at AMNH used 64 processors of an SGI Origin2000 and that at SDSC used the Blue Horizon, a 1000 processor IBM SP2. The final simulation produced more than 10,000 files with more than 2.5 TBytes of data. This data at NCSA was initially stored in the UniTree tape system, and then transferred to SDSC using the SRB over a period of 9 days on Internet2 network using parallel data streams.

The data at SDSC was first staged on a 800GB cache system and also partially replicated on the IBM SP2's GPFS file system. The data was also replicated onto the HPSS archive at SDSC. Since the data size was around 2.5TB and the cache was only 800GB, the SRB was used to optimally place the data at multiple caches, including the HPSS system at CalTech and disks belonging to other projects. The SRB made this data shuffling and data placements seamless to the application since the applications see only the logical path names and not the changing physical path names of the files. The rendering of the movie was done at SDSC on the IBM Blue Horizon using the SDSC 3D Volume Renderer. The intermediate steps in the cleaning and rendering process generated more than 5TBytes of data of which 3TBytes were needed during the rendering process. The rendering resulted in a movie run on 7 video projectors used in full-dome projections at the Hayden Planetarium. During rendering the SRB was used to move data in and out of the IBM GPFS from the various sites where data sets were replicated. Data needed for rendering runs were placed on the GPFS and removed when it was done with new data taking its place.

During the whole process, the interactive dialogue between the various players at AMNH, NCSA and SDSC was intense to make the simulation as close to physical reality as possible. SRB was not only used for data movement and data placement but also as a collaboratory for communicating data pieces across the sites. Figure 1. shows the data movement architecture.

2MASS DIGITAL SKY PROJECT

Data Sizes: 10 TeraBytes

Number of Files: more than 5,000,000

Sites: SDSC, IPAC at CalTech, CACR at CalTech

Resources: Sun E10K and IBM SP2

Data Storage: Tape systems, HPSS, and Unix FS

Project Duration: 1.5 years ingestion, continuing access

The 2Micron All Sky Survey [5] is a star catalog with images of stars taken using fixed telescopes at the 2 micron wavelength. The 2MASS survey used two highly-automated 1.3-m telescopes, one at Mt. Hopkins, AZ, USA and one at CTIO, Chile. The raw files from the telescopes were stored in the FITS format on tapes. The aim of the project was to make the files that were on off-line tapes into a near-line system where astronomers can access the raw images through the web. The tapes were read at CalTech and moved over to SDSC using the CalRen2 network. The data was transferred in two streams continuously. The ingestion routine was written at IPAC, and made to be 'fault-tolerant' by an SRB expert at SDSC. The data movement was bottle-necked by several factors at various times during the process: the amount of space available at CalTech for reading tapes, cache space limitations at SDSC, limited tape drive availablility and HPSS transfer problems. In spite of these shortcomings and network failures, the data transfer was done smoothly by the SRB.

Before ingesting into the HPSS storage system at SDSC, we had to deal with another complication. Archives such as HPSS do not handle small files very well. HPSS gets choked up when small files are written and ingesting 5 million of them would have overwhelmed the system. SRB has a feature called 'containers' that addresses this problem [6]. We also used the container to sort the data in a more useful manner. The data on tapes at IPAC are stored temporally, i.e., in the order in which they were taken. Since the telescopes sweep the sky nightly each tape contained images from large swaths of the sky. But the astronomers generally use the data by locality, i.e., they look at specific regions of the sky and access objects in a region. To make this data access optimal, we decided to store the data in containers by the spatial locality instead of storing as they come from the tape. This temporal-to-spatial sorting implied that the images from a tape would go into as many as 3000 containers per tape. Due to the lack of cache space, during the later periods of the ingestion, the SRB was moving containers in and out of HPSS to accommodate filling them from the tapes. This led to some trashing, but was unavoidable because of the nature of the sorting that we were performing. The whole data ingestion process took around 18 months.

The images in the SRB are currently being used by the astronomy community accessing them through the web: we have more than 1000 hits per day for these files which were unavailable online just two years back. Moreover, there are at least two groups that are planning to perform full-sky mosaicing using all the images in the Survey. Figure 2 provides a sketch of the of the 2MASS project w.r.t. SRB.

VISIBLE EMBRYO PROJECT

Data Sizes: 1 TByte pilot. Can grow to 10 PetaBytes

Number of Files: several 100,000

Sites: GMU, LLNL, SDSC, AFIP, UICMC, OHSU, JHU

Resources: Sun E10K, SP2

Data Storage: HPSS, NTFS, Unix FS

Project Duration: ongoing ingestion and access

The Visible Embryo project (http://netlab.gmu.edu/visembryo.htm) is part of the Next

Generation Internet Initiative and is funded by the National Library of Medicine. The purpose of the project is to demonstrate applications of leading-edge information technologies in computation, visualization, collaboration, and networking in order to enable newer capabilities in science and medicine for developmental studies, clinical work and teaching. The project has three aims: 1) deploy high-grade workstations for collaboration, 2) digitize the Carnegie Collection and place it in a digital library setting using the SRB, and 3) demonstrate the system in annotation and modeling, embryology education and clinical management planning.

The National Museum of Health and Medicine (AFIP/NMHM) is in charge of digitization and data acquisition. SDSC is involved with data storage, replication and volume rendering of the images. The other sites are involved in using the data collections for scientific and medical purposes. The SRB is used for data management and metadata management through out the project. The digitized images are converted to multiple resolutions and are moved on a continuous basis to be stored in the HPSS archive at SDSC. The images go through a staging process at NMHM were some metadata are extracted. The metadata is stored as files in the SRB and in a pilot project extracted and placed into an Oracle database that was under the control of the SRB. Access to the metadata for querying and access to associated images are provided over the web. Figure 3. provides a sketch of the data movement and processes involved in the project.

BIOMEDICAL INFORMATICS RESEARCH NETWORK

Data Sizes: several TBytes

Number of Files: several million files

Sites: SDSC, UCSD, Duke, UCLA, CalTech, Harvard

Resources: Linux BIRN Racks

Data Storage: HPSS, NTFS, Unix FS

Project Duration: ongoing ingestion and usage

The BIRN project [3] is an NCRR/NIH initiative aimed at creating a testbed to address the needs of biomedical researchers to access and analyze data at a variety of levels of aggregation located at diverse sites throughout the country. Many data grid issues such as user authentication and auditing, data integrity, security, and data ownership will also be addressed as part of the BIRN project.

The BIRN Coordinating Center is located at the University of California, San Diego, and provides the integration needed for the BIRN network which is currently funded for two activities: 1) The Mouse BIRN Project includes activities at Duke University, UCLA, Caltech, and UCSD and 2) The Brain Morphology BIRN project includes two research groups at Harvard Medical School, one at Duke University, and two groups at UCSD.

A main important part of the BIRN project is to deploy custom hardware that is extensible but administered from a central site. A linux-based system, called the BIRN RACK [12], is deployed at each site and is an essential grid component that makes it very easy to manage data grid resources in a wide area network. The BIRN Racks are customized to run the SRB data grid, and a Coordinating Center manages the SRB administration aspect of the grid. Data is replicated across sites under SRB control. The SRB is being used for accessing data through several visualization programs that are currently under use by the various BIRN partners. In a sense this uniformity of access will allow one to apply different techniques to the same data and see the relative advantages and disadvantages. The main aim of the BIRN project is to overcome the challenges and problems in accessing large datasets across sites while keeping the strict compliance needed by the medical data sharing regulations. The BIRN project is in progress and several data integration and collaboration efforts are under development. Figure 4 shows the role played by the SRB.

ACKNOWLEDGEMENT

This research has been sponsored by National Institute of Health (Award No. NIH-8P41RR08605) and the National Science Foundation (Award No. ASC 96-19020).

REFERENCES

[1] "The Particle Physics Data Grid", (http://www.ppdg.net/,http://www.cacr.caltech.edu/ppdg/).

[2] "The Grid Physics Network", (http://www.griphyn.org/proj-desc1.0.html).

[3] "BIRN: Biomedical Informatics Research Network", (http://www.nbirn.net).

[4] "Network for Earthquake Engineering Simulation", (http://www.eng.nsf.gov/nees/).

[5] 2MASS, http://www.ipac.caltech.edu/2mass/.

[6] Rajasekar, A., M. Wan, and R. Moore, "MySRB & SRB - Components of a Data Grid," *The 11th International Symposium on High Performance Distributed Computing (HPDC-11)* Edinburgh, Scotland, July 24-26, 2002.

[7] "The Knowledge Network for Biocomplexity", (http://knb.ecoinformatics.org/).

[8] "National Virtual Observatory", (http://www.srl.caltech.edu/nvo/).

[9] "EarthScope", (http://www.earthscope.org/).

[10] "Storage Resource Broker, Version 1.1.8", SDSC (http://www.npaci.edu/dice/srb).

[11] Moore R., and A. Rajasekar, "Data and Metadata Collections for Scientific Applications", High Performance Computing and Networking, Amsterdam, NL, June 2001.

[12] Rajasekar, A., and M. Wan, "SRB & SRBRack - Components of a Virtual Data Grid Architecture**",** *Advanced Simulation Technologies Conference (ASTC02)* San Diego, April 15-17, 2002

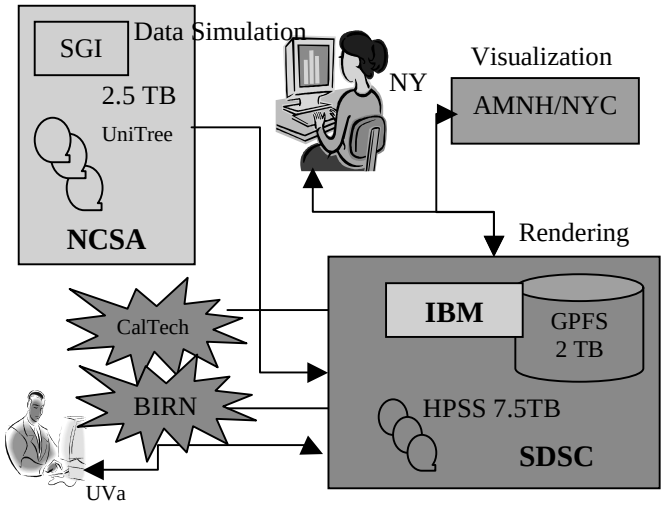

Figure 1: Hayden Planetarium Project involved simulation at NCSA (Urbana-Champaign, IL), rendering at SDSC (San Diego, CA) and final visualization at the Hayden Planetarium (New York, NY). 2.5 TB was moved using the Internet2 Network using as much as 19 streams in parallel. The rendering produced 5 TB of data stored at multiple location to juggle space for near-time usage. The visualization is done using 7 projectors at the Hayden Planetarium. Intensive collaboration during the project used the SRB as a central black board for quick turnaround.

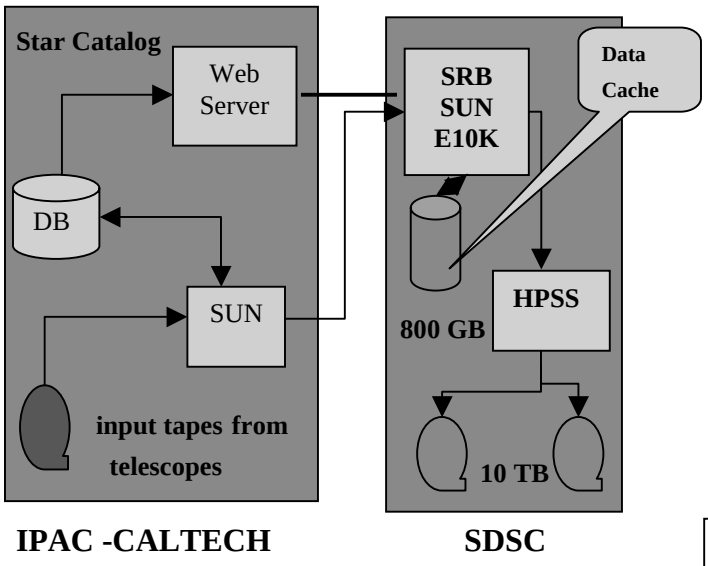

Figure 2: 2MASS Digital Sky Project involved IPAC at Caltech and SDSC. Tapes were read at IPAC and data transferred over CalRen2 to SDSC where it was spatially sorted into containers in data cache and finally archived into HPSS. The data transferred was limited by tape reading capacity at CalTech. 10 TBytes of data was transferred over a year in two parallel streams. The temporal-to-spatial sorting involved data movement between cache and archive.

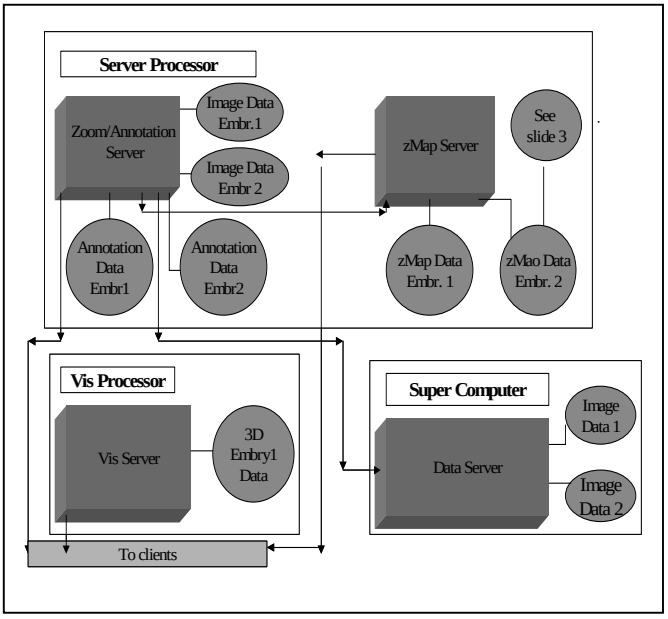

Figure 3: Visible Embryo Project: The diagram shows the distributed players and the various activities that are performed including annotation, zMap search services, visualization and data serving.

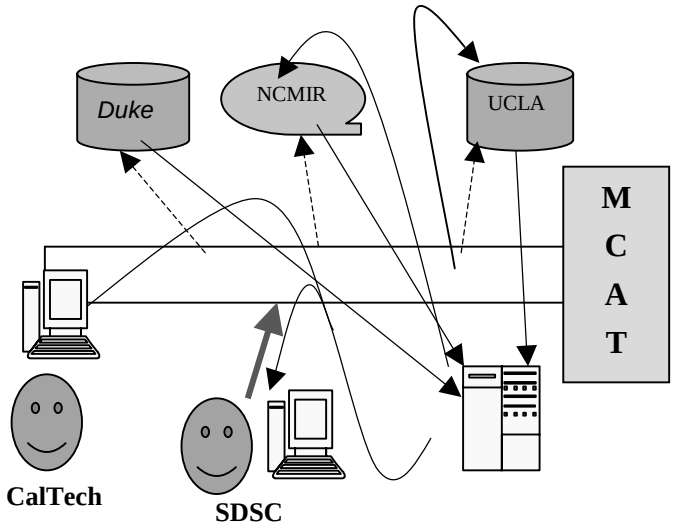

Figure 4: BIRN Data Grid: Shows possible interaction between various sites. Data residing at Duke, NCMIR and UCLA is copied to a high-powered compute platform for rendering and visualized concurrently at SDSC and CalTech. The MCAT mediates the process, helping in data discovery and method discovery and in checking for authentication and authorization.