SDSC Technical Report 2003-01

NARA Persistent Archives
NPACI Collaboration Project

Status Report

October 2002 to December 2002

Reagan W. Moore
Richard Marciano
San Diego Supercomputer Center
January 26, 2003

Sponsored by
National Archives and Records Administration

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Archives and Records Administration or the U.S. Government.

Table of Contents

1. Introduction..... 1
2. Hardware Systems Status..... 1
3. Software System Status..... 3
4. Collection Status 5
5. Theoretical Implications 6
6. Papers..... 8
9. Presentations, Workshops, and Meetings 9

1. Introduction

The National Partnership for Advanced Computational Infrastructure, an NSF sponsored program, is collaborating with the National Archives and Records Administration on a demonstration of prototype persistent archives. The NPACI partners formally involved in the collaboration include the research staff of the Data and Knowledge Systems group at the San Diego Supercomputer Center, the University of California at Berkeley, and the University of Maryland. The EVResearch Corporation is applying concept –based organization of material. Additional NPACI partners are informally involved, including Stanford, which is integrating the WebBase web crawling technology with the SDSC Storage Resource Broker data grid.

The goal of the development effort is to demonstrate support for automation of archival processes through use of data grid technology. Demonstrations are being done on representative collections provided by NARA. The collections include:

- Tape collections provided on 3480 tape cartridges
- Code of Federal Regulations
- Presidential web sites
    - President Eisenhower - http://www.eisenhower.utexas.edu/
    - President Ford - http://www.ford.utexas.edu/
    - President Johnson - http://www.lbjlib.utexas.edu/
    - President Reagan - http://www.reagan.utexas.edu/

This report will cover the status of the prototype hardware systems, the status of the persistent archive software development, and the status of the acquisition of the collections. A final section covers recent theory developments, and papers and presentations related to the topic of preservation and data grids.

2. Hardware Systems Status

At the planning meeting held on August 16, 2002 at the University of Maryland, an engineering document was developed for the design of grid bricks on which the persistent archives would be created. The systems included commercially available platforms (to be implemented at the University of Maryland and NARA), and commodity platforms that would be integrated as a cheaper cost solution at SDSC.

The platforms are described in Table 1, which shows the planned CPU, disk, and archival storage capacities. The key features are that each site is capable of standing alone, independently supporting the persistent archive. Data grid technology is planned for federating the independent collections. The systems are interconnected by high-performance networks, with the goal of sustaining data flows between sites at rates up to 4 MB/sec. Each site is planning on acquisition of at least 1 TB of disk cache on which the archival processes can be demonstrated.

The systems at the University of Maryland have been purchased and installed. This has been an effective demonstration of the efficacy of using commercial systems for integrating processing and storage.

The systems at SDSC are partially installed. SDSC has worked with three different integrators for the assembly of commodity-based systems. While the cost per Terabyte of storage for commodity-based systems is up to a factor of ten cheaper than the commercially integrated systems, the ability to get a robust platform is highly dependent on the skills of the integrator. The first system was developed in house, by Tom Guptil, and was used to identify problems in the Linux operating system, the preferred commodity products for RAID controller (3Ware), the disk cabling vendor, and the network interface. A grid brick was implemented by Tom and is now used in production, with a capacity of 1 TB. A second grid brick was ordered to the same specification from a local integrator, Datel, but was inoperable on arrival, with a failed disk. The system was returned, but came back with a failed RAID controller. The system is still in repair. Two additional systems were bought from a second integrator, at a higher cost of $4,500 per grid brick. Both systems are performing well, indicating that the original design could be integrated into a stable platform. We will try a third commercial integrator in order to explore whether the original price configuration can still be met.

Comparison of Grid Bricks

Site	NARA	U Md	SDSC	SDSC
System	Dell	Dell	Grid Brick (8)	Database brick
Name	DELL 1650	Dell 1650		
CPU	2-1.4 Ghz	2-1.4 Ghz	1-1.7Ghz	2-1.7Ghz
CPU type	Pentium 3	Pentium 3	Celeron	Pentium 3
Memory	4 GB	4 GB	1 GB	4 GB
Disk controller	FastT200	FastT200	3Ware	3Ware
Disk connect	2- fibrechannel	2- fibrechannel	IDE	IDE
Disk size	1.1 TB	1.1 TB	1.1 TB	0.64 TB
Disk drive	73 GB	73 GB	160 GB	80 GB
Disk RPM	10000	10000	5400	5400
Throughput	166 MB/s	166 MB/s	110 MB/s	110 MB/s
			25 MB/s write	25 MB/s write
Network	Gig-E	Gig-E	100/Gig-E	10/100/Gig-E
	10/100 baseT	10/100 baseT		
OS	Linux	Linux	Linux	Linux
Cabinet	Rack/3U	Rack/3U	Rack/4U	Rack/4U
Database	Oracle	Informix		DB2
Support	3-year	3-year		
Total Disk	1.1 TB	1.1 TB	8.8 TB	0.64 TB

Table 1. Hardware systems for the prototype persistent archive

SDSC has started an effort to understand the optimal database brick design. The goal is to tune the amount of memory to the size that can be processed effectively by the CPU for queries across all records. Multiple database bricks can be used to house the entire collection. A database that supports shared-nothing (partitioned data), such as DB2, will be used to federate multiple database bricks. The Storage Resource Broker data grid is used to federate the Grid Bricks. A paper on Grid Bricks has been accepted for publication in the proceedings on the 20th IEEE Symposium on Mass Storage Systems and Eleventh Goddard Conference on Mass Storage Systems and Technologies, to be held in San Diego in April 2003.

Both the University of Maryland and SDSC have additional archival storage systems that are accessible from the systems controlling the persistent archive disk caches. Data can be migrated between the disk caches and the archives under the control of the archival engineer.

The hardware at the NARA site is still in procurement. When the platforms are up and running, a trip will be scheduled to support installation of the persistent archive software.

A second hardware development effort has been completed at SDSC to prepare for the ingestion of collections. The 3480 tape drives originally planned for use at SDSC had been taken out of operation, since the current drive technology is based on 9940 tapes, which have a capacity of 200 GBs per cartridge. To support the ingestion of collections from 3480-tapes, a tape stacker was re-installed at SDSC with 3480 tape drives. The system is now being used to load the collections provided by NARA.

A third hardware effort still needs completion, that of establishing a network connection between NARA and the University of Maryland. This link is critical if data is going to be replicated between the three sites over high-speed networks.

3. Software System Status

The SDSC Storage Resource Broker data grid is being used to integrate the three persistent archive sites. The software at each site has three components:

- Operating system for the CPU
- Database instance for managing a metadata catalog
- SRB servers for managing a logical name space and interacting with storage repositories

We are using the Linux operating system on all of the platforms. SDSC does have a mod for the booting of file systems that are larger than 1 TB is size. Otherwise, standard operating systems releases of Linux are used.

Different database technologies are used at SDSC (Oracle) and the University of Maryland (Informix). At SDSC, multiple database instances are being used while the database grid brick is developed, which are in production use on Sun E15k platforms.

The Oracle database for the persistent archive is currently running on a Sun E15k. An instance will be installed in the future on the database grid brick.

The SRB data grid uses a metadata catalog (MCAT) to manage the logical name space into which digital entities are registered. The MCAT system is capable of running on Oracle, DB2, Sybase, and SQLServer relational databases. A port of MCAT to the Postgres database was completed at SDSC, to facilitate the creation of a version of MCAT suitable for the Informix operating system. The Postgres port was provided to the University of Maryland, which then created a version of the MCAT software that could be run on Informix.

The SRB server was ported at SDSC to the 64-bit version of the Linux operating system, in anticipation of the next generation of hardware platform. The SRB environment now runs on all of the platforms planned for the NARA persistent archive prototype.

A second major effort is still in progress for developing the software needed to automate archival processes. Many of the features that are desired, such as bulk metadata loading, bulk data registration, parallel I/O support, have been available as enhancements of release 1.1.8 of the Storage Resource Broker. These extensions are being integrated into a standard version 2.0 for release in the middle of February. The integration of the capabilities has been completed, and the new software release is in test. Some additional features include support for streaming of data in containers, and the generation of a list of the digital entities within the containers for remote processing. This will make it easier to replicate large collections between the sites. Once the testing is complete, version 2.0 of the SRB will be distributed to all of the sites involved in the persistent archive prototype.

A third major software effort is the generation of a robust system for the harvesting of web material, through the integration of web crawlers with data grids. A first effort has been completed as part of the NSDL project (NSF National Science Digital Library). The public domain Wget web crawling technology was integrated with the SRB data grid. In the NSDL project, the software was used to crawl 13,000 URLs from the NSDL repository, and deposit the retrieved digital entities into a SRB collection. Over 1.7 million digital entities were retrieved by crawling to a depth of two (through recursive processing of internal URLs within a digital entity to identify additional relevant material). The process demonstrated the need to parallelize the crawls and avoid the retrieval of duplicates. Of the 1.7 million entities, about 75% were duplicates.

In collaboration with Stanford University, the SRB data grid is being integrated with the Stanford WebBase technology. WebBase supports both parallel crawls and the identification of duplicates. The integration project has required the development of new utilities to stream containers. The collaboration with Stanford would be a profitable area for support in the next set of development activities with NARA.

A fourth major software development effort is in a planning stage. The federation of collections between SDSC, U Md, and NARA can be done at the application level, with a program controlling the explicit replication of data and metadata between the sites.

While this approach is feasible and straight forward to implement, the long-term goal is to support federation of collections directly within the data grid. This will enable the federation of collections across agencies, and can be used to simplify the accession process by automating data transfers from agencies into the NARA persistent archive. The design of the collection federation capabilities is being done in collaboration with the grid community (Carl Kesselman at ISI), the United Kingdom data grid (Kerstin Kleese at Daresbury Laboratory), the NIH Biomedical Informatics Research Network (Mark Ellisman at UCSD), and NARA. A first version is planned for the middle of calendar year 2003. The system will support registration of collections into a remote SRB metadata catalog, and simple controls for access and consistency between collections.

To support federation of collections that have different consistency constraints (such as access control lists, user authentication, collection organization, update policies), more powerful technology will be needed. SDSC is starting the design of a constraint-based collection management system, which will make it possible to support peer-to-peer federation of collections as digital entities. This will be used to integrate ontologies that describe the constraints that are imposed on each independent collection with the information management catalog. Examples include adding a concept space to impose an alternate organizational structure, adding control mechanisms for automating replication of digital entities between collections, and managing access controls between collections. The SRB technology makes it possible to associate access controls with the logical name space. This means that as data is moved between sites, the access controls are preserved. The related question is how access controls are preserved when digital entities are registered into an independent collection.

4. Collection Status

NARA has shipped a collection of two hundred 3480-tape cartridges to SDSC, which are in the process of being loaded. A first set of 30 tapes has been loaded over a weekend, and the remaining tapes will be loaded in February, 2003. The system works sufficiently well that an additional 300 3480-tapes can be easily handled.

The University of Maryland has loaded all of the collections that have been sent to them.

The Code of Federal Regulations (CFR) has been acquired by EVResearch Corp. and is being analyzed. The collection was obtained from the Office of the Law Revision Counsel and is 273.239 Mbytes in size. The analysis tools used by EVResearch Corp include a “break” module, an indexing module, a search module, and an “un-break” module. The “break” module uses expert rules to separate the CFR into finite elements and create categorical tags for the elements. The index module creates a searchable database of the finite elements across the contained strings (words and phrases). The search module supports queries on the database, and the “un-break” module presents a contiguous portion of the original document.

The collection contains 50 documents and about 25,000 elements. Current structures that are tagged include:

- CITE-
- EXPCITE-
- HEAD-
- MISC1-
- SECREF-
- STATUTE-
- SOURCE-

Iterations will be done over the collection to refine categorical tags and index the material. The time to process the collection is 73 seconds. A fully indexed version will be completed in March 2003.

We plan to crawl the presidential web sites using both the SDSC and Stanford technology. We will house the resulting collections on the grid bricks, and back up the digital entities into containers archived on HPSS. During the 1st quarter of calendar year 2003, the collections will be housed in the SRB instance at SDSC. An initial crawl of the Ronald Reagan web site resulted in the retrieval of 90 MBs of material. The results of the crawl are being analyzed to determine completeness, types of errors seen in the process, types of material that were retrieved, and the required depth of crawl.

The engineering tasks relative to the persistent archive collections will then be demonstrated, including multiple combinations of location of ingestion, management of metadata, and storage of data. We will demonstrate interoperability between the sites using the SRB data grid technology. The demonstrations will include the ability to vary the location of the accession, the metadata catalog, and the data storage.

Ingest Site	Metadata Catalog Site	Data Storage Site
U Md	U Md	U Md
SDSC	SDSC	SDSC
U Md	U Md	SDSC
U Md	SDSC	SDSC
NARA	NARA	NARA
NARA	U Md	SDSC

Table 2. Planned demonstrations of distributed data ingestion, data management, and data storage

5. Theoretical Implications

During the last quarter of calendar year 2002, multiple interactions were conducted with the archival community on the development of research agendas. The meetings included:

- EU(DELOS)-US Working Group on Digital Preservation and Archiving led by co-leaders: Margaret Hedstrom (US) & Seamus Ross (UK)
- NHPRC workshop on Electronic Records Research agenda, led by Robert Horton (Minnesota Historical Society)
- National Academy of Science Computer Science Technology Board assessment meeting on preservation research, held at SDSC
- Workshop on XML as a Preservation Language held in Urbino, Italy.
- InterPARES 2 workshops on Preserving Authentic Non-Textual and Interactive Records, held in Los Angeles and Rome in December.

The archivists who participated were firmly convinced of the need to preserve the information and knowledge content of their collections. The most succinct statements were made by the Canadian archivists in the InterPARES 2 workshop, who stated that their development agenda was predicated on the need to re-purpose collections to make them relevant at a future date. The goal was to develop the ability to examine prior practices based on archived material to improve future practices.

In a position paper generated for the Urbino, Italy workshop on XML as a preservation language, a new approach was defined that integrates aspects of migration and emulation into a common system. In both the data grid community and the persistent archive community, the need to automate the manipulation of digital entities is growing. Both groups need the equivalent of a digital ontology that will:

- Define the relationships that are present within a digital entity that are used to convert bit streams into structures, apply semantic tags to the structures, map the structures to coordinate systems, map units to semantic tags, and associate time stamps or location information with semantic tags
- Define the order of application of the relationships that are needed to manipulate the digital entity

The manipulation of the digital entity is then based upon the defined relationships. The application that presents the digital entity would:

- Define the set of operations that can be performed on the defined relationships
- Map the operations on the relationships to operations on the storage repository

With this approach the original bits of the digital entity are never changed. Only the digital ontology is migrated over time onto new encoding standards for the representation of the relationships that are present within the digital entity. Emulation, the migration of the presentation application to new operating systems, is then done based upon a set of defined operations on defined relationships. A generic presentation application can be created for a generic representation of the mapping from relationships to operations, and does not have to be re-done for every digital entity. Effectively, all the knowledge that is unique to a digital entity is captured within the digital ontology. This approach can be extended to encompass manipulations within data grids.

From the perspective of the data grid community, the persistent archive community is managing technology evolution, including the ability to manipulate digital entities when encoding formats become obsolete. Archival processes are applied to create a representation of the digital entity that characterizes the internal relationships, such as the creation of structures from the bit stream, the mapping of the structures to coordinate systems, the assignment of semantic labels to the structures, the assignment of physical units to semantic labels, the assignment of temporal relationships, etc. The mappings can be organized into a digital ontology that represents the procedural order that is needed to transform bits into semantically labeled components. The digital ontology is then used as the characterization of the digital entity. The digital ontology can be migrated onto new encoding standards for annotating relationships and semantic tags, while the original bit stream within the digital entity is preserved unchanged. The emulation program operates on the digital ontology, as a generic mapping from defined relationships to permitted operations. As new technology becomes available that allows additional operations on given relationships, the emulation environment becomes more sophisticated and richer.

The persistent archive community, therefore, needs to mange not just distributed state information across collections, but also the digital ontologies that characterize the information and knowledge content within digital entities. Grid ontologies are needed that characterize the collective properties of the digital ontologies. Operations on digital entities can be controlled by the set of relationships that are in common across the digital entities. Grid services would correspond to operations on these common relationships. The distributed state information managed by data grids then corresponds to mappings imposed on the data grid logical name space for the manipulation of the common relationships within the digital ontologies. Grid services can be expressible as a grid ontology, specifying which operations will be performed on the underlying digital ontologies.

The final major theoretical effort that continues is the definition of the minimal set of data grid capabilities that are needed to implement a persistent archive. A review of a draft document was held at the Global Grid Forum meeting in Chicago, on October 17, 2002. Corrections to the draft will be reviewed at the next Global Grid Forum meeting, to be held the week of March 3, 2003 in Tokyo, Japan.

6. Papers

1. R. Moore, “Common Consistency Requirements for Data Grids, Digital Libraries, and Persistent Archives”, submitted to 12th High Performance Distributed Computing conference, Seattle, Washington, Jun 2003
2. R. Moore, C. Baru, “Virtualization Services for Data Grids”, Book chapter in "Grid Computing: Making the Global Infrastructure a Reality", John Wiley & Sons Ltd, 2003.
3. Arcot Rajasekar, Michael Wan, Reagan Moore, George Kremenek, Tom Guptil, “Data Grids, Collections, and Grid Bricks”, Proceedings of the 20th IEEE Symposium

on Mass Storage Systems and Eleventh Goddard Conference on Mass Storage Systems and Technologies, San Diego, April 2003.
4. Michael Wan, Arcot Rajasekar, Reagan Moore, Phil Andrews, "A Simple Mass Storage System for the SRB Data Grid", Proceedings of the 20th IEEE Symposium on Mass Storage Systems and Eleventh Goddard Conference on Mass Storage Systems and Technologies, San Diego, April 2003.
5. R. Moore, "The San Diego Project: Persistent Objects", Proceedings of the Workshop on XML as a Preservation Language, Urbino, Italy, October 2002.
6. Edward A. Fox, Virginia Tech: Reagan W. Moore, San Diego Supercomputer Center; Ronald L. Larsen, University of Pittsburgh; Sung Hyon Myaeng, Chungnam National University; and Sung-Hyuk Kim, Sookmyung Women's University, Toward a Global Digital Library: Generalizing US-Korea Collaboration on Digital Libraries, D-Lib Magazine, October 2002, http://www.dlib.org/
7. Arcot Rajasekar, Reagan Moore, Bertram Ludäscher, Ilya Zaslavsky, " The Grid Adventures: SDSC's Storage Resource Broker and Web Services in Digital Library Applications:, 4th Russian Conference on Digital Libraries, Dubna, Russia, October, 2002.
8. R. Marciano, B. Ludäscher, R. Moore, I. Zaslavsky, K. Pezzoli, "Multi-Level Information Modeling and Preservation of eGOV Data", Intl. Conference on E-Government (EGOV), LNCS 2456, pp. 93-100, Aix-en-Provence, France, September 2002.

9. Presentations, Workshops, and Meetings

To promote the development of common systems between the digital library community, the data grid community, and the persistent archive community, presentations have been given at workshops, conferences, project coordination meetings, federal review meetings, and at standards meetings. In particular, persistent archives are being developed for the NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL). Federation of data resources through data grids is being developed for the NSF National Virtual Observatory (NVO), the NSF Grid Physics Network (GriPhyN), the NSF Southern California Earthquake Center (SCEC), the NASA Information Power Grid (IPG), the DOE Particle Physics Data Grid (PPDG) and the NLM Digital Embryo project. The goal is to promote common software solutions across all of these application areas, that will result is a standard implementation that can be used by other communities.

1.	Dec 19, 2002	DOE	PPDG all-hands meeting
2.	Dec 16, 2002	NSF	GriPhyN all-hands meeting
3.	Dec 13, 2002	LLNL	Presentation on Data Grids
4.	Dec 8, 2002	NHPRC	Workshop on NHPRC research directions
5.	Dec 2, 2002	NSDL	Launch of the NSDL repository
6.	Nov 28, 2002	SDSC	SDSC-UK data grid coordination meeting
7.	Nov 25, 2002	UCB	Seminar on NARA preservation projects
8.	Nov 25, 2002	NSF	Symposium on Knowledge Environments
9.	Nov 21, 2002	NVO	NVO advisory committee meeting

10.	Nov 17, 2002	SC-02	Tutorial on Storage Resource Broker
11.	Nov 11, 2002	NSF	EU-US preservation workshop
12.	Nov 8, 2002	NSF	DLI2 ADEPT coordination meeting
13.	Oct 28, 2002	Singapore	SRB tutorials
14.	Oct 25, 2002	Canada	Data grids for environmental data
15.	Oct 23, 2002	NIH	Program review committee
16.	Oct 21, 2002	CDL	Preservation of Intellectual Capital
17.	Oct 17, 2002	GGF6	Persistent Archive Research Group
18.	Oct 14, 2002	NSF	NPACI executive committee
19.	Oct 11, 2002	Italy	XML preservation workshop
20.	Oct 8, 2002	NASA	NASA data grid workshop
21.	Oct 4, 2002	NSF	SCEC coordination meeting
22.	Oct 3, 2002	NARA	NAS/CSTB preservation technology review

Richard Marciano gave a seminar at the University of California, Berkeley on Nov. 25, 200, that proposed evaluating the UCB document presentation mechanisms from a preservation context. The UCB presentation system is capable of processing multiple types of document formats, and in effect is a version of an emulation program that uses format-specific drivers to handle different document types. Given a specification of standard operations on a standard set of relationships, a driver can be written that will operate on a digital ontology. This would be a realization of the concept of differentiating the knowledge inherent in digital entities from the digital entity and the system used to display the digital entity.