

Data Management in a Research Environment

Academic Systems and Technologies

April 2006

Although data storage is often treated as commodity class, especially when considering cost at the lower end of the spectrum, its status in a competitive cyberinfrastructure is actually quite different. This paper explores the role of data management in a research environment and the direction it is likely to take at UMDNJ.

The path taken by data from collection to ultimate retirement is influenced by many factors including:

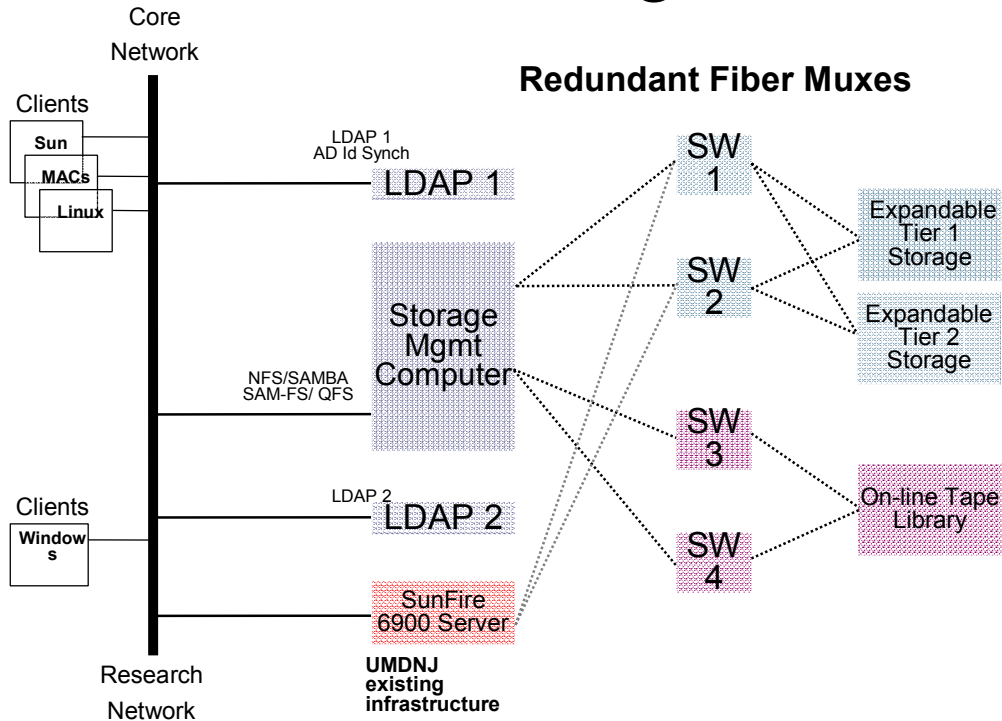
- Human or machine generated;
- Volume of data;
- Editing/vetting requirements;
- Availability;
- Analysis: Complexity of computation, size of datasets;
- Accessibility to multiple compute environments;
- Publishing and collaboration requirements;
- Security and data protection;
- Near-line and archival access;
- Disaster recovery.

The management of data from collection throughout its lifecycle of synthesis into information and finally knowledge cannot be taken for granted or left to chance. In fact, Federal granting agencies keenly recognize and assume that institutions already have or are building into their infrastructure the facility to manage data throughout its useful life. Rapidly evolving technologies such as Internet2 and LamdaRail are demanding far more synergistic approaches and connections between computers, storage, networks and people. Institutions that do not recognize these trends will be significantly disadvantaged in the funding arena for high end, collaborative research.

Information Services and Technology has always endeavored to provide data storage capabilities for the research and academic communities consistent with requirements for security, protection, access and performance. Technology and cost considerations have historically limited the extent to which storage solutions could be crafted to meet the most emergent research needs. Today, commercial solutions exist that address all of the factors articulated above. The next sections detail a vision of information life cycle management for the research community. It closely parallels commercially available solutions in this space.

The following diagram depicts a simplified view of the storage architecture that IST recommends for research data management and represents what is commonly referred to as a hierarchical storage environment.

Solution Block Diagram:



Each level on the right-hand side of the diagram from the tape library to tier 1 represents increasingly higher levels of performance and correspondingly higher cost per unit storage. In the environment envisioned by IST and other organizations at UMDNJ, data is introduced into tier 2 via network-attached clients, either at the desktop or through core facility or other laboratory devices and processes whereby data is collected and vetted prior to further analysis. Data arriving at tier 2 is quickly copied to the tape library for archive. Data in tier 2 can be manipulated by research network clients that see it as network-attached storage. Tier 2 storage is suitable for many storage and access purposes.

Tier 1 storage is considerably faster than tier 2 and consequently more costly. This tier is used by high performance computing equipment such as the Sunfire 6900 maintained by IST's division of Academic Systems and Technologies. HPC assets, which in the future are expected to also include high performance linux clusters along with more conventional systems such as high-end Sunfire class servers are directly connected to the storage environment, typically through fiber channel. In this way the I/O requirement of very fast computing equipment is able to take complete advantage of the very high performance data access and transfer available from tier 1. Data requested by HPC

systems are automatically migrated to tier 1 where they remain until policies cause the data to be migrated back to tier 2 and simultaneously copied to the tape library for archive.

Policies dictate how data courses through the environment and are set to insure optimum system performance. Considerations such as amount of storage available at each tier, access frequency, user quota in tiers 1 & 2 determine the migration of data in both directions in the environment. The user always sees the same view of his/her data regardless of where it resides in the system. Data is transparently moved to the tier at which it will be used. Some delay may be experienced when data is moved from archive to an active tier, but the process is always automatic and the user is presented the same view regardless of where the data initially resides. Automatic controls limit data migration rates to insure fair utilization of the environment for a large population of users. On-line library tapes are automatically replicated at intervals complying with local data center best practice for off-site storage for disaster recovery purposes.

A typical data management scenario might involve data collected and vetted by a core facilities, say flow cytometry, electron microscopy or DNA microarray analysis. Here, laboratory work is conducted at an investigator's request and results, often in the form of significant amounts of data, are made available for subsequent analysis. These data can be made available to the researcher in tier 2 storage and appear as network-attached storage. Similarly, the investigator can move data directly from the desktop to tier 2. Further analysis including web publication might be performed on this tier or on tier 1 depending on the nature of the computations. Another scenario might involve large scale statistical computation using SAS, where prototyping is done at the desktop or in tier 2 storage. Complex analysis using very large datasets could then be performed on tier 1. In either of these or any other scenario, the investigator is assured that data is readily accessible, secured and protected and available to whatever level of computational equipment/service is required.

Once research data is managed in a state-of-the-art environment as described herein, it becomes possible to consider emergent technologies designed to identify relationships among data from seemingly disparate disciplines and provide additional incentive for translational research. Technologies such as Shibboleth, which will facilitate identity management among cooperating institutions, are made easier to implement when research storage becomes an enterprise function and becomes an important building block for UMDNJ playing in a national cyberinfrastructure setting.

Initiatives such as this, targeted to the research community, not only insure that critical data is properly managed throughout its lifecycle, but helps the University to remain technologically competitive and well positioned to take advantage of funding opportunities that are increasingly focused on collaborative and translational research.