Research Storage

Andrew Caird, Paul Killey

14 March 2013

Contents

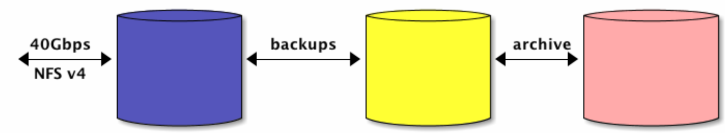

Figure 1: Storage stack described in this document

Research Storage Background

Types:

- Lustre / scratch
- NFS / working storage
- HDFS / Map-Reduce, Hadoop, HBase, columnnar storage
- Web Object Storage / HTTP semantics for large chunks of data

Characteristics:

- storage
- speed
- availability

Tiers:

- Lustre types: /nobackup on Nyx and /scratch on Flux
- NFS types: Value Storage and research storage
- HDFS: generic and highly-tuned
- WOS: local and distant

Implementation on campus	Avail.	Speed	Capacity	Semantics
/nobackup on Nyx	low			Lustre
/scratch on Flux	medium	100Gbps		Lustre
Value Storage	medium	3Gbps		NFSv3, POSIX
Research Working Storage	high	40Gbps		NFSv4, POSIX
HDFS	user-selected			M-R, HBase, etc

Strategy:

- multi-tier that is X, Y, and Z

Strategy Background

(This is from a Google document of similar name at `http://goo.gl/qHnIC`.)

As I understand demand, these are the minimal set of storage options required to address the scholarly and administrative data and management requirements at the U-M:

1. Lustre (parallel storage for HPC)

2. NFSv3 (file system based storage)

3. NFSv4 (file system based storage)

4. CIFS (file system based storage)

5. HDFS (distributed storage supporting "big data": map-reduce (Hadoop), column-oriented data (NoSQL), etc.)

6. Object Storage (something like Amazon S3, replacing or augmenting file system storage)

7. SQL (Oracle, MS SQL Server, MySQL, etc.)

8. Backups (relatively short time to recovery)

9. Cold storage (relatively long time to recovery)

These categories encode / encapsulate provisioning and configuration decisions regarding:

1. Media (SSD, spinning disk, tape, cloud)

2. On-disk formats and file systems

3. Network protocols and networking / fabric capacity

4. Media management (NAS, SAN, HFS, appliance, DAS, cloud, etc.)

These storage options have to be made available as accessible services at acceptable cost / capability tiers (where storage capability is traditionally expressed in terms of capacity, performance, and availability).

They all require implementation decisions.

These implementation decisions can be optimized along dimensions of interest: the number of platforms, vendors, campus providers, products, feature sets, etc., or otherwise expressed requirements from different communities.

Research Storage Strategy

The need for electronic storage of research data at U-M will be met with three broad classes of storage:

1. **very high-speed, temporary storage** that is tightly coupled with the high-performance computing environment

2. **high-speed, secure, and safe storage** that is broadly available on campus and appears via the network as local storage

3. **storage for distributed processing of large amounts of unstructured data** that is scalable in capacity and performance on premise or via a cloud-computing provider

The research working storage service is of the second type: data storage that provides speed of access, security of access, and safety of data—research working storage. Research working storage (RWS) and its associated business processes are well-suited to research data, although can be used for other types of data or as the basis for other services.

The intent of the storage service is that it will be useful to a large fraction[1] of researchers who need a service that will support a data lifecycle and also be useful as a basis for other services (such as data curation) or entitlements to staff, students, or faculty.

To be as useful as possible to as many researchers as possible, the Research Working Storage service will provide:

- tools to put as much of the control as possible over the life-cyle of data into the hands of the data owner
- an integrated data-archiving service that is pre-paid so that there are no on-going costs for archived data
- a for-fee subscription service appropriate for storing data associated with research that integrates active storage, backups, archives, business processes, and IT processes
- a service that is accessible to both the researchers and IT systems that need it
- a service that is presented securely to on-campus consumers in the broadest possible way and to the most possible clients
- a storage service that matches the performance of the data-generation and data-analysis systems available on campus and compares favorably with the performance of locally provisioned storage
- a service that includes back-ups and archiving of data while ensuring that the data owner has control over the back-ups, restores, and archives of their data
- a way that the storage consumers can manage the initiation, alteration and termination of active storage; the restore or deletion of backup copies of the data; and storage into and retrival from the archive location
- useful and actionable information to the consumer about the age and usage of the storage to which they have subscribed

[1] Once we have an understanding of the costs associated with the service, we will survey faculty members about their ability and interest in paying for the service and discuss cost-sharing options with Colleges.

- a cost structure based on a flexible operation that can adopt the best hardware-based or provider-based technology options without impacting the delivered service, allowing the service manager to optimize the operations for costs as the technologies and services change over time
- integration with U-M business practices so payment and billing is done in a familiar environment
- integration with U-M IT practices so it can be used in a familiar manner to other IT services
- professional IT and business operations and support

The RWS service follows the model established for computing by Flux, where there is a capital investment in the initial service and the unused capacity before there are enough subscribers to reach sustainability. Following the capital investment and aggregation of a subscriber base, the money recovered by the rate would fund the replacement hardware, and the amount of money recovered by the rate would inform the size of the next version.[2]

Aggregation and abstraction are the key components of this service from an administrative perspective, as they allow for cost management, economies of scale, and vendor optimization. At the same time, performance, security, and data protection are key components of this service from a researchers' perspective.

Implementation

There are three components to the RWS service, the integrated set of which includes procedures for acquisition, monitoring, and termination that are integrated into the U-M business processes and easily used by the research community we support.

High-speed Storage

The first of the three components is a high-speed storage service. This service is what is presented to the researcher and is a proxy for the back-up and archive services. The quantity of the storage for which the subscriber pays can be varied over time. The subscriptions can be funded by different sources with the storage presented to the subscriber as either an aggregated amount across funding sources or as separate amounts between funding sources, depending on the needs of the subscriber.

Technical Details

The technical details of the high-speed storage are:

[2] This could also be reflected in a section about costs.

- the storage system provides aggregate bandwidth on the order of high-performance computing interconnects[3]; today this is 40Gbps, but will increase as networking technology advances
- the storage system will provide snap-shots of data on disk for simple, user-directed recovery of data that is was changed or deleted and for which a previous version is needed
- the protocols of the storage will be directed toward the systems that are most likely to consume it; today this is NFS[4]
- the presentation of the storage will be to on-campus clients but will not be restricted to managed systems in order to support as many researchers and devices as possible; to ensure security with this broad presentation, the storage will initially be offered via Version 4 of the NFS protocol (NFSv4) and be integrated into the existing campus Kerberos infrastructure to provide strong authentication

Service Details
The service details of the storage are:

- the storage will be sold in an "allocation model" where each 50GB-6 month[5, 2] unit will have a start date, end date, and funding source. Units of storage can optionally be combined in projects so they are presented to the researcher as one aggregated pool of storage
- storage projects will be implemented by system quotas that will vary as the allocation units come and go; this will also support usage reporting to the researchers
- when an allocation unit expires, the disk quota will be set to the sum of the remaining active allocations; if this quota is less than the total data stored, no more data can be written, but data can be read
- if there are no additional allocations then the disk quota is set to zero. At this point no data can be written to the disk space owned by the project, but data can read; if no new allocations are made after two weeks[6], the data is removed from active storage but kept in the backups for the duration of the backup retention time. Backups can be restored to active storage if a new storage allocation is created or they can be archived to long-term storage if the project with which the data is associated has remaining archive credits.

[3] This is the current speed of the InfiniBand network on Flux.

[4] We are aware that Windows protocols may be useful, and we will try to choose a product that can support current Windows file system protocols (SMB, CIFS, etc.)

[5] This quantity and time combination is a policy decision that can be refined based on costs, market analysis, and specifics of the service.

[6] The expiration rule is a policy decision that can be refined based on the needs of the community of subscribers or other information and requirements.

Backups

All of the data stored is also backed up and a set of backups are kept for a reasonable period of time and can be restored, deleted, or archived by the owner of each project.

The management of the backups is via a web-based presentation of the data contained in the backups, command-line tools on some systems, and email-based support.

Technical Details The technical details of the backups are:

- the backups of the data are kept for a resonable time and a reasonable number of copies are kept; initially the backups will cover a time span of one year, with copies from the previous day; the previous one, two, three and four weeks; the previous one, two, three, four, five six, eight, ten, and twelve months (a total of fourteen copies)[7, 2].

Service Details The service details of the backups are:

- the backups are only of data from the active storage and are not offered as a general purpose backup system; there are already several on-campus options for that service
- the backup system integrated with the RWS service is unique to it because of the high performance required of it to back up the amount of data the system is designed to store and because there is only one client from which to back up data; other backup services on campus do not have the same performance requirement and must support many clients[8, 2]

Archives

The archive portion of the research working storage service will use a cloud-based data archive solution[9]. The process of depositing and withdrawing data from the archive is researcher-directed and the level of curation is at the discretion of the researcher. The style of archives in this service can be called “data graveyard” or “data dumping ground” in contrast with a curated archival solution that a data management group or library[10] might offer.

Technical Details The technical details of the archives are:

- the data archive will use a cloud-based data archive solution; a local abstraction layer will allow U-M to choose appropriate cloud service providers as the market changes over time[2]
- today the most likely cloud service provider of data archive services is Amazon and their Glacier product [11]

[7] The backup retention policy will be determined by requirements and costs; the example here is just one option.

[8] The cost and performance profiles for a general purpose back-up system differs from those of a single purpose back-up system; as we understand costs we will evaluate them against the costs of the available options.

[9] There are no on-campus archive options, so a cloud-based option is likely the best option.

[10] The U-M library could use this service as the technical component or back-end to an archiving service they offer.

[11] `http://aws.amazon.com/glacier` Amazon Glacier is an extremely low-cost storage service that provides secure and durable storage for data archiving and backup. In order to keep costs low, Amazon Glacier is optimized for data that is infrequently accessed and for which retrieval times of several hours are suitable. With Amazon Glacier, customers can reliably store large or small amounts of data for as little as $0.01 per gigabyte per month

  - archives will be initiated by the researcher via a web interface and the data to be archived will be from a backup set, not from the working set; this allows for stability of the data over the potentially long duration of the archiving process[12]
- restoring data from archive will be initiated by the researcher via a web interface and the data to be restored will be restored to active storage (the NFS tier); before the restore begins an allocation adequate to hold the restored data must be acquired. The restore will be stopped if the space is filled before the restore is complete

Service Details The business details of the archives are:

- as part of the storage allocation, each project will receive tokens for archiving and restoring data
- U-M will be able to assign archives (or parts of archives) to individuals outside of U-M or will have a federation model so each researcher has an identity at the cloud-based archive provider so they can restore their data outside of the U-M environment by paying the cloud service provider directly[12]

Possible Scenarios

Following are some scenarios that illustrate the concepts above integrated into contrived but hopefully representative examples.

Augmented base storage allocations

The College of Engineering has decided to make an allocation of 100GB an entitlement to all of the researchers in the College. There are many research groups in the College who require more than that to support their work, and they are expected to augment their base allocation to suit their needs with their own funds.

Dr. Smith has a research scientist and six graduate students and usually manages two or three grants, although sometimes there are additional projects.

Her base storage allocation of 100GB provided by the College is held in a project called `jmsmith_rs` ("Dr. J.M. Smith's research storage") which is accessible by her six graduate students, her research scientist, and herself, each with their own directories and some shared directories for collaboration. She and her graduate students connect to this storage from their laptops and workstations and it is also available on Flux and in computers in the U-M 3-D Lab.

Dr. Smith makes a request for an additional allocation for that project of 400GB for 4 years, bringing the total disk space available in

[12] This requirement informs either the selection process for a cloud/archive service or the level of staffing for a locally-developed solution.

`jmsmith_rs` to 1/2PB. She does this knowing that she can reduce the amount of space at any of the 6-month billing periods if she needs to, but by making it for 4 years, she avoids the risk of forgetting to renew it.

In addition to her main storage project, she is also working on a project called *NexMent* with her research scientist and two of her own graduate students and two graduate students from the Physics department in LSA. That research project has its own funding source and storage requirements, so she requests another project to provide storage to herself and the five other people involved and pays for it ith the grant money for that project. She now has another project, called `jmsmith1_rs` with a similar internal structure to her main project, but a different set of people who are using it.

When the *NexMent* project ends Dr. Smith initiates the process to archive the data from it and ends the storage allocation. The long-term preservation of the data, although slow to retrieve, fulfills a portion of the required data management plan associated with the grant that funded the project and there are no additional costs assigned to that grant.

As her career follows the typically successful arc of U-M faculty, her resource requirements wax and wane and she can control her costs and adjust her resources to follow that, all while knowing that her data is stored on professionally managed, high-quality infrastructure.

Researcher with intermediate funding

Dr. Jones' research in the ethnography of music is very data intensive—tape recorders in the field have been replaced with high-fidelity digital recording devices—but the funding for data intensive work in the world of music is sparse at best.

Dr. Jones has two main components to his research: field work collecting samples of music before the limitless arm of iTunes reaches everyone; and analysis, cataloging, and reporting on what he has collected. The funding for these two activities can come together or each one can be funded individually.

Dr. Jones uses the U-M Research Storage Environment for both aspects of his work.

After returning from the field with many gigabytes of audio on his digital recorders, he copies it to the working storage where he knows it is backed up and can be quickly accessed. He is able to search and play and work with his data directly from the working storage without having to hold it all on his laptop (which is good, because the 256GB SSD in his MacBook Air couldn't hold all of it). When the relatively short-term field-work grant ends, he archives all of the data from the most recent back-up of his working data and works on writing papers and more grants, knowing his data is safely archived and can be re-called when he needs (and can afford) to have it.

When Dr. Jones is funded to analyze a particular sub-genre of music that he has recorded on dozens of field trips over the years, he restores those from the archives to active storage and is able to look at them as a unit. For him, the ability to keep all of his data for a very long time allows for a kind of research that would be impossible if he had to decide what to keep as the end of a grant.

The ability to "warehouse" data at a low cost for long periods of time allows Dr. Jones to use storage as a service instead of risking his data on low-cost, low-performance, low-quality, or low-all-three hardware. The ability to work on data over a high-speed connection allows Dr. Jones to save time transferring data to his laptop and also allows him to work on much bigger data sets than he could on his laptop, while still offering excellent performance for his audio analysis tools.

Researcher who leaves U-M

Dr. Robbins has been at the University of Michigan for twelve years, but has decided to move to Minnesota to be closer to his parents. He has amassed data in support of his research into disease transmission over the years that he will need in his new job. Unfortunately, Minnesota State University does not have a Research Storage Environment like U-M's, but they will fund the purchase of several USB harddrives for him.

Before Dr. Robbins leaves U-M, he makes his last archive from his the backups of his active storage and prints the web page with the instructions on restoring archived data in a non-U-M environment.

When he arrives at Minnesota State University, he attaches his USB drives, and follows the instructions on restoring data from an archive, which include him paying the archive restoration and transfer fees from his funding in Minesota, so U-M does not incur any cost for this, although U-M will continue to maintain his data in the archive for 10 years after the last piece of data was added to the archive, so when Dr. Robbins' USB drives fail, he can pay for another restore and transfer.

Path to the future

This Research Storage Environment positions U-M well to be as efficient as possible in its support of research IT.

- By enabling researchers to use services for their research computing needs, U-M is positioned to either aggregate demand to one supply and enjoy economies of scale on campus or use off-campus alternatives at lower costs.

- Having an archive option of any sort, even if it is a "data graveyard", is an option that has not been available to researchers at U-M and has the potential to change what types of research can be done.

- Having an archive option will support the option of a curated archive, and because we would be using the same technology for both types of archive, the cost would be lowered for both.

- As more and more workload moves to off-site cloud providers, we can enable caching of data near the compute resources to ensure the data is available where it is needed[13], even if it is needed in two very distant locations at once; when the balance of the workload shifts to off-campus, we can start using cloud providers for the high-speed (in this example, NFSv4) storage, and put the smaller caches on campus for local access.

As mentioned, aggregation and abstraction are the key components of this service from an administrative perspective, as they allow for cost management, economies of scale, and vendor optimization. At the same time, performance, security, and data protection are key components of this service from a researchers' perspective.

Interaction with other on-campus storage services

The service proposed here is one of many different storage options available to researchers at the University of Michigan, and interaction with all of those is an important part of this service. In general, this is designed to be fast enough, large enough, and scalable to that it should present a reasonable interface to other options.

Scratch storage on Flux

Scratch storage on Flux is based on the Lustre parallel file system[14]. Lustre is tightly integrated with Flux and is not presented to hosts that are not managed by the Flux operators.

This level of integration is important to maintain the performance and security of the file system. In addition, Lustre is only supported on Linux—there are no Mac or Windows clients.

The research working storage service proposed here will provide a location for long-term storage of large inputs or outputs that are best

[13] Technically, this will be done with sophisticated NFS (or other storage protocol) caching appliances or software.

[14] `http://www.lustre.org` Lustre is a parallel distributed file system, generally used for large scale cluster computing. Lustre file systems are scalable and can support tens of thousands of client systems, tens of petabytes of storage, and hundreds of gigabytes per second of aggregate I/O throughput.

stored on Lustre while the related computational jobs are running or are staged to run.

Because the Lustre implementation on Flux is a very high-speed (40-80Gb/s) and very high-capacity (more than 600TB) filesystem, it has the ability to ingest, store, and output large quantities of data, so a long-term storage location for that data should be as fast as can be afforded, so that researchers don't spend any longer than necessary moving their data.

The research working storage service proposed here is a good complement to Flux's Lustre installation.

ITS Value storage

The NFS service offered by ITS called Value Storage[15] is based on NFSv3 and is available to anyone on campus. It was built as a low-cost, reliable NFS service. It was not built specifically for high speed. Value Storage includes an option to mirror data between two locations and the mirror is updated daily and there are snapshots of data on disk. Backups are not included but are offered via ITS' TSM service.

As we develop the components of the research working storage service, several may be suitable for integration with Value Storage.

For researchers who don't require the level of performance provided by the research working storage service proposed here, Value Storage offers good alternative.

Department or Lab storage

Many departments and research laboratories provision local storage and present that to clients via NFSv3 (for Linux or Mac clients) or CIFS (for Windows or Mac clients). Most of these storage services are small in capacity (less than 50TB) and low performance relative to the proposed research working storage service.

The advantage offered by local storage services is that they are usually a one-time cost that can be attributed to a grant as hardware. The disadvantages are that they are often not operated by people with operational experience in storage and that puts the data stored on these systems at some risk; these systems typically provide slow access to data because of their combination of networking (usually 1Gbps) and the number of disks in the system (usually less than 12); and these systems are often not expandable beyond a few tens of terabytes.

We expect that the combination of Value Storage, the proposed research working storage service and its backup and archival components, IT Rationalization with respect to staff, and the increasing

[15] `http://www.itcs.umich.edu/storage/value` Value storage is designed to provide a cost-effective way for University researchers (and others with large storage needs) to store large amounts of data in a centralized location. Disk space can be purchased in terabyte increments.

requirements for long-term data management will lead to fewer and fewer departments or research laboratories providing local storage.

Unstructured or Big Data storage

Much of the data at U-M that would fall under the new umbrella of "big data" or "unstructured data" (as opposed to relational data that is typically stored in relational database management systems like Oracle, MySQL, etc.) is currently stored where it is processed. In some cases this is in a Hadoop cluster, in other cases is it NoSQL systems and in other cases it is flat files.

The research working storage service will have the performance and capacity to ingest, store, and archive data from these systems as the current data is no longer needed but the space on the analysis platform is needed for the next research project.

As a data management support system, the research working storage service is an excellent complement to existing and future big data clusters.

ITS TSM product

The ITS TSM product[16] offers tape backups of data from many sources, and maintains two copies in separate geographic locations. This service has historically been viewed as expensive, which it is, and a bad value, which, for the right data, it is not. However, there is a class of data on campus for which ITS' TSM product is too richly featured and thus too expensive. The backups included in this research storage proposal are a very local, highly-integrated part of the service, and will not be offered as a generic backup service separate from the research working storage service. There will also be integration between the backups associated with the research working storage service and the archive, which is likely not appropriate for the TSM service.

In addition, we expect to make archive copies of data from backups, which is not supported in TSM today.

Web-based Data Sharing and Collaboration

In the College of Engineering researchers have expressed interest to us in a web-based method of sharing data and collaborating with other researchers (especially those from other institutions for whom getting U-M credentials is inconvenient). The characteristics of this web-based data sharing, as we understand them, are around all control of the service being held by the researcher, including hardware and software selection (Windows, Linux, or MacOS; a forum, a wiki,

[16] `http://www.itcs.umich.edu/tsm` The Tivoli Storage Manager (TSM) service provides networked backup of data on server-level machines (such as application and file servers, server-side databases, and research data collections).

a file upload/download service), maintainence of the access lists, data policies, and presentation.

The research working storage service described here would be suitable as the backing storage for a service like this:

- the performance of the storage would be sufficient to serve web-based requests
- snapshots and backups would offer some insurance against mistakes that would result in data loss were there only one copy
- the ability to archive the data at the end of the project without moving it would be nice
- the ability to have multiple, segregated storage areas (or "projects") will help with data management

While the option of a Mac Mini, a Drobo and a CrashPlan subscription is likely to be less expensive than a service like this, the features offered by this service may be worthwhile from the perspective of data security and external data management requirements.

Costs

For now, this is just the dumping ground of all of the places I mention costs[2] elsewhere in the document, other than those in the Scenarios section.

- the storage will be sold in units of Quantity per Time, where Quantity and Time will both vary as the technology, costs, and business operations change over time; today this will be 50GB of storage for 6 months
- there is no separate cost for the backups, they are integrated into the research working storage service
- These archives are intended to be a one-time cost for securely storing data to minimize the costs of active storage allocations. Using the archive service described here, the costs for active storage can be minimized to zero and there are no on-going costs for the data kept in the archive, only costs for storage and retreival.
- behind the scenes the web-based archiving tools will be a set of web services applications that will query the backup system and the archive system, presenting options and costs via a web page where the researcher (or other data manager) can initiate an archive, check on the progress of an in-progress archive, and view statistics about completed archives

- behind the scenes the web-based archive restore tools will be a set of web services applications that will query the archive system and active storage system, presenting options and costs via a web page where the researcher (or other data manager) can initiate a restore, check on the progress of an in-progress restore, and view statistics about completed restores

- if there are real cost differences between sending data to the archive and restoring data from the archive, that will be reflected in the number of tokens required for each action. Each 6 month, 50GB allocation[5, 2] will include enough tokens to archive 50GB two times and restore it once

- archives will be kept for 10 years at no cost to the researcher[2]

- While the costs aren't yet firm and we haven't surveyed potential subscribers to the service, if we don't think the service cannot be financially sustainable without subsidies we will investigate other options for storage appropriate for researchers.

- Because the backups associated with the research working storage service are so constrained (a single client, no campus-wide networking), they should be less expensive than TSM or any other option. In addition, we need some access to the backups to support the user-driven archives.