Journal of Computational Science xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

# Large scale research data archiving: Training for an inconvenient technology

S. Patrick Calhoun, David Akin, Brett Zimmerman, Henry Neeman*

*The University of Oklahoma, Norman, OK, USA*

ARTICLE INFO

*Article history:*
Received 25 June 2015
Received in revised form 24 March 2016
Accepted 14 July 2016
Available online xxx

*Keywords:*
Research archive
Technology training
Large scale storage

ABSTRACT

At small scales, storage is straightforward to afford and to use, but at large scales – from several Terabytes (TB) to many Petabytes (PB) and soon Exabytes (EB) – tradeoffs must be made between cost and convenience, and training for use of such resources needs to take such inconveniences into account. A large scale, long term (over 10 year) institutional research data storage archive is described, focusing on both hardware and software. The technology choices give rise to inconveniences, which in turn not only lead to a crucial requirement for training on the proper use of the archive, but also inform the specifics of that training, as does each individual use case.



## 1. Introduction

How physical storage is structured, and how it is used, can vary substantially across scales, because of both pricing concerns and technological aspects. At the smallest scales – for example, handhelds such as mobile phones and tablets – pricing is affordable (typically under US$1 per GB, with maximum sizes typically well under 1 TB), and use mechanisms and administration are convenient and intuitive (for example, push a MicroSD card into a slot in the handheld, and the operating system automatically recognizes it and puts it into service). By contrast, at the largest scales (from several TB to many PB and soon EB), storage can either be reasonably convenient to use but expensive (for example, large scale enterprise-class disk systems, which can be comparable in purchase price per GB to small scale but are much more expensive to operate), or reasonably affordable but inconvenient to use (for example, magnetic tape).

At the same time, research datasets are increasingly being subject to requirements or needs not only to be retained over several to many years, but also to be made accessible to relevant communities external to the data owners, typically at no more than the incremental cost of creating and transferring a copy. For example, in 2013, the US Office of Science and Technology Policy released a memorandum [1] calling on every US federal research funding agency with a research funding budget over USD $100,000,000 to prepare a public access plan. In 2015, the US National Science Foundation (NSF) released its Public Access Plan [2], which stated:

> *NSF requires applicants for funding to prepare a [Data Management Plan] . . . [which] may address . . . [p]olicies for access and sharing . . .. All data resulting from [NSF-funded] research . . . should be deposited at the appropriate repository . . .. NSF's data-sharing policy states: "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data . . . created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing". . . . NSF requires applicants . . . to address archiving and preservation . . . Strategies for providing long-term storage and preservation will be a requirement for any future NSF-designated repository system whether for data or publication. . . .*

However, in an era of increasingly open access to massive data collections, some storage technologies and some extant business models, for large scale, long term (over 10 year) storage of "cold" data, including enterprise disk or tape systems and metered cloud providers, aren't universally viable under current research funding approaches. This typically is because (i) the cost of storage is too high to be practical, and/or (ii) the file owners are obligated to continue paying substantial recurring charges even after the relevant research funding has ended.

Among the key issues are: (1) the cost of storing large datasets (2) over the long term, while making the datasets (3) not only accessible to the owner (4) but also discoverable and accessible by third parties as appropriate, (5) and being able to use

* Corresponding author at: One Partners Place Suite 2600, 350 David L. Boren Blvd., Norman, OK, 73019, USA.
*E-mail address:* hneeman@ou.edu (H. Neeman).

http://dx.doi.org/10.1016/j.jocs.2016.07.005
1877-7503/

shorter term funding such as a 2–5 year research grant, (6) with minimal recurring costs, (7) encompassing multiple copies to improve resiliency (8) at minimal cost per TB per copy per year.

Under these constraints, the following storage strategies are extremely challenging: (a) funding a disk system refresh after end-of-life (5–7 years) is very difficult; (b) enterprise disk in general is too expensive per TB per year; (c) buying disk drives in a centrally-managed disk array gets too little lifetime for some disk drives, because the useable lifetime of the disk drives typically ends at the end-of-life of the disk array, so disk drive purchases late in the life of the disk array have even higher cost per TB per year; (d) metered cloud storage can be unsustainable beyond the lifetime of the relevant project, because it can be difficult to justify expending funds from later grants on irrelevant datasets from earlier grants; (e) collections of standalone disk drives (for example, USB disk drives) are undiscoverable, inaccessible, cumbersome to manage at scale (tens of TB to many PB), and don't last long enough; (f) buying a tape library per research team is impractical due to high fixed costs (5–8 figures per medium-to-large tape library[1]).

Large scale tape archives, by contrast, have the following advantages: (i) low incremental price per unit (other than fixed costs, tape costs substantially less per TB per year than even USB disk drives [3,4]); (ii) longevity (10 years or more); (iii) accessibility; (iv) discoverability (via metadata catalogs); (v) media (tape cartridges) can be paid entirely up front, with zero recurring costs for 10+ years.

Disadvantages of large scale tape archives include: (i) long latency (wait time) before any individual file can be read (30–120 s for tape, vs 1–10 milliseconds for disk), so tape is best for "cold" archiving of files that are expected to be accessed infrequently; (ii) high fixed costs, typically six or seven figures for a tape library with hundreds of tape cartridges. [5–7]

Thus, tape may be impractical at the research group scale, but can substantially reduce costs to researchers at institutional and national scales.

Note that discoverability – whether on a tape archive or a disk system – depends first on physical access (for example, via the Internet) to the contents of the storage system. Metadata and related information describing the contents of files on such a storage resource can be crucial for users who need to search for such content (as well as for provenance, reproducibility and other purposes), but only come into play once physical accessibility is resolved. (Issues relating to metadata are outside the scope of this article.)

At the University of Oklahoma (OU), the OU Supercomputing Center for Education and Research (OSCER), a division of OU Information Technology, has been using a very successful business model [8] that effectively addresses these concerns for an institutional-scale resource. This business model is based on three funding sources: (1) grant: an NSF Major Research Instrumentation (MRI) grant (OCI-1039829, "Acquisition of Extensible Petascale Storage for Data Intensive Research," USD $792,925, 10/1/2010 – 9/30/2014, PI H. Neeman) funds hardware, software and the first several years of warranty/maintenance/support; (2) institutional commitment by OU provides space, power, cooling and labor, as well as maintenance after the initial warranty period; (3) researchers buy their own media, typically but not exclusively via their own grants.

Thus, researchers' cost per TB per copy per year is significantly less than that of USB disk drives, because of both lower purchase costs (see above) and longer and more predictable media lifetimes [9,10].

Unfortunately, because of constraints of both budget and technology, the use of OU's storage archive is neither straightforward nor convenient. In particular, the technology choices (informed by budget constraints) compel inconvenient usage mechanisms, which in turn require targeted tailoring of user training. Effective training regarding proper use is crucial, and this training must be both brief and intuitive, in order to reduce violations of appropriate practices and policies, while minimizing the amount of time devoted to this training by both users and operations staff.

## 2. Technology

The Oklahoma PetaStore [11], OU's research data archive, consists of a tape library, a disk array, a set of servers, software packages for tape and disk, and a networking environment. Of these components, almost all are standard systems that are commonly used for purposes like this, but the tape library software is an unusual choice in this context. (See the Appendix for details of the hardware, software and operating environment.)

Similar institution-scale resources can be found at other US academic institutions; for example, the University of Washington's "lolo" archive [12] and the University of Colorado Boulder's PetaLibrary [13].

## 3. Inconvenience

Crucial to understanding OU's approach to PetaStore training (Section 4, below) is first understanding not only the extent to which use of the PetaStore is inconvenient, but also the specific nature of those inconveniences, and why the PetaStore has been designed in such an inconvenient manner. Ultimately, a key driver of inconvenience is the tradeoff between cost and convenience: as described above (Section 1), an enterprise-class disk system, or a metered third-party cloud resource, would be very convenient, but to match the PetaStore's capacity would cost far more – especially far more to the users, who on the PetaStore fund only the media, not the system – and would also need to be refreshed via institutional funding instead of user funding, which would be impractical.

### 3.1. Archive filesystem access mechanisms

#### 3.1.1. Cluster supercomputer data transfer nodes

On OSCER's cluster supercomputer, there are a pair of support nodes (servers), known as archive nodes, whose role is to execute large scale data transfers. These are the only nodes on OSCER's supercomputer that mount both (a) all of the supercomputer's globally accessible user filesystems and (b) the PetaStore filesystem. Thus, logging in to these archive nodes provides the simplest mechanism for transferring files from the supercomputer to the PetaStore: using the Unix cp (copy) command, as if the PetaStore filesystem were part of the supercomputer.

There are multiple reasons why the PetaStore filesystem isn't mounted on other supercomputer nodes. A key reason is that each node that mounts the PetaStore filesystem needs to run a General Parallel File System (GPFS, also known as IBM Spectrum Scale) client, and while such clients are very affordable on a modest number of servers, the cost would quickly become more than can be sustained if the GPFS client were deployed on hundreds of servers such as the compute nodes on OSCER's supercomputer.

In addition, mounting the PetaStore filesystem on (especially) the compute nodes (and to a lesser extent on other nodes such as

[1] For example, on February 16, 2016, an IBM TS4500 tape library with 730 tape cartridge slots and 2 tape drives, driven by 5 Lenovo x3650M5 servers, an IBM Storwize V3700 disk array, a pair of IBM SAN24B-4 Express Fibre Channel switches, IBM's General Parallel File System software and IBM's Linear Tape File System Enterprise Edition software, with only a single year of support, had a Manufacturer's Suggested Retail Price (MSRP, also known as list price) of over USD $450,000; the same configuration except with 12 tape drives and 9970 tape cartridge slots had an MSRP of over USD $1,000,000. [5–7].

Please cite this article in press as: S.P. Calhoun, et al., Large scale research data archiving: Training for an inconvenient technology, J. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.jocs.2016.07.005

login nodes) would present a very large, accessible filesystem – far larger than the aggregate of all other user-accessible filesystems combined – so there would be a high risk that some of the research applications running on the compute nodes would end up accessing the PetaStore filesystem and using it for writes and/or reads, as if the PetaStore were a live filesystem. This approach would consume substantial transaction capacity on both the PetaStore filesystem and the relevant campus backbone networks, risking significant congestion, and potentially slowing down both the PetaStore and those research applications. By not mounting the PetaStore filesystem on the compute nodes, this risk is avoided.

#### 3.1.2. External data transfer nodes

*3.1.2.1. Secure FTP/Secure copy.* The PetaStore filesystem is also accessible via a pair of servers external to OSCER's cluster supercomputer, via mechanisms such as Secure FTP (sftp) and Secure Copy (scp). For example, files can be transferred from a Windows laptop to the PetaStore via WinSCP. This mechanism tends to be slow, because most such connections are from desktop and laptop systems rather than from large scale, enterprise-class storage resources.

*3.1.2.2. Globus.* The PetaStore has been configured as a Globus [14–16] (formerly Globus Online) endpoint. Globus is both a software stack and a web-based service that facilitates data transfers between designated endpoints (each of which the Globus software stack has been installed on). The Globus software provides a Graphical User Interface (GUI) to files on endpoints, so that users of those endpoints can freely transfer files using high speed protocols such as gridftp. In addition, the Globus software monitors data transfers and automatically restarts those that are interrupted. User accounts are available for free at any time, and the use of Globus is free for those with accounts on Globus endpoints.

### 3.2. Duplication options

The PetaStore supports four duplication options:

disk_1copy_unsafe: This option is designed for files that are expected to be accessed regularly and/or rapidly, and that are duplicated on other repositories (for example, files downloaded from national centers).

disk_1copy_tape_1copy: This option is designed for files that are expected to be accessed regularly and/or rapidly, and that reside only on the PetaStore.

tape_1copy_unsafe: This option is designed for files that are "cold" (not expected to be accessed often and not needing rapid download), and that are duplicated on other repositories.

tape_2copies: This option is designed for files that are "cold," and that reside only on the PetaStore.

These duplication options are implemented as subdirectories under each user's top-level directory structure. For example:

/archive/username/disk_1copy_unsafe
/archive/username/disk_1copy_tape_1copy
/archive/username/tape_1copy_unsafe
/archive/username/tape_2copies

The mechanisms for implementing these duplication options are tied to these subdirectories, so that any content (files and/or subdirectories) inside one of these duplication option subdirectories will automatically undergo the correct procedure to implement the appropriate duplication option.

For the duplication options disk_1copy_unsafe and tape_1copy_unsafe, the duplication option subdirectory name intentionally includes the word "unsafe," in order to regularly reinforce that the choice to forgo a duplicate copy entails significant risks.

The duplication options are implemented via a variety of Tivoli Storage Manager (TSM, also known as IBM Spectrum Protect) features: disk_1copy_unsafe doesn't use TSM at all, disk_1copy_tape_1copy uses the Backup feature, tape_1copy_unsafe uses the Hierarchical Storage Management (HSM) feature, and tape_2copies uses a combination of TSM's HSM feature and Disaster Recovery feature.

However, underlying this implementation is a considerable degree of complexity, and this complexity results in constraints on system behavior such as, for example, the possibility of a lag of several hours between a file appearing in a user's tape_2copies subdirectory on the PetaStore disk staging area and the completion of the second copy on a separate tape cartridge from the first copy.

Offsite Duplication: For tape_2copies, secondary copies are regularly exported from the tape library and physically transported to an offsite disaster recovery location, several miles from the data center that houses the PetaStore. By this mechanism, secondary copies are well positioned to survive a natural disaster or a technological misadventure. This component of the data flow is designed to be visible to operators only, but the training does touch on this aspect.

The value of this approach was recently demonstrated when a failure of one of the TSM servers, combined with a TSM feature intended to benefit backup requirements instead of archive needs, led to a collection of user files being inadvertently deleted by the system, but 100% of offsite files were successfully recovered.

### 3.3. Minimum and maximum file sizes

Unlike disk, tape storage typically has very high latency per file, because a file retrieval typically requires preparation time of 30–120 s (typically 60) per file: (1) rewind the previous tape cartridge that is in the targeted tape drive to its beginning; (2) eject the previous tape cartridge; (3) carry the previous tape cartridge to its slot; (4) travel to the tape cartridge slot with the requested file; (5) select the requested tape cartridge; (6) carry the requested tape cartridge to the targeted tape drive; (7) insert the tape cartridge into the tape drive; (8) fast forward to the start of the requested file.

This preparation requirement imposes significant time costs and risks on retrieval. For example: (a) file retrieval times can be hugely increased due to the imbalance between preparation time and file read time; (b) very large numbers of files can lead to very long traversal times for the file catalog database, which can result in slowdowns of services; (c) having many small files stored on tape can lead to "shoeshining" wear and tear on the tape medium.

Because of these consequences, maintaining substantial numbers of small files isn't practical. On the other hand, excessively large files can lead to monopolization of limited resources (in particular the finite number of tape drives), so constraining the maximum permitted file size allows tape read requests to be serviced in a timely manner.

Therefore, on the PetaStore, the minimum permitted file size for each Linear Tape Open (LTO) version is the greater of 1 GB or 0.01% of tape cartridge capacity, the maximum permitted file size is 10% of tape cartridge capacity, and the recommended range is 10–100 times the minimum permitted file size. (See Table 1.)

As a result of these extrema, at the minimum permitted file size, preparation time dominates (4–20 times as long as file read time); at the minimum recommended file size, preparation time and file read time are balanced; at or above the maximum recommended file size, file read time dominates.

Note that none of these timings includes time that a request might spend waiting in the request queue until a tape drive becomes available to service that request. To date on the PetaStore, queue time hasn't been a major contributor to preparation time,

Please cite this article in press as: S.P. Calhoun, et al., Large scale research data archiving: Training for an inconvenient technology, J. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.jocs.2016.07.005

**Table 1**
Properties of LTO-5 and LTO-6 on the Oklahoma PetaStore.

| | LTO-5 | LTO-6 |
|---|---|---|
| Capacity, uncompressed | 1.5 TB | 2.5 TB |
| Peak read/write speed | 140 MB/sec | 160 MB/sec |
| Minimum permitted file size | 1 GB | 1 GB |
| Minimum permitted file size read time | 7 s | 6 s |
| Minimum recommended file size | 10 GB | 10 GB |
| Minimum recommended file size read time | 1.2 min | 1 min |
| Maximum recommended file size | 100 GB | 100 GB |
| Maximum recommended file size read time | 12 min | 10 min |
| Maximum permitted file size | 150 GB | 250 GB |
| Maximum permitted file size read time | 18 min | 26 min |

but there is an ever-increasing possibility that queue time could become significant or even endemic.

### 3.4. Compressing and aggregating

While some research workflows have individual files that are at or above the minimum permitted file size, some do not, and a substantial fraction have at least some files that are below (often well below) the minimum permitted file size. For a collection of files, some of which are below the minimum permitted file size, the minimum permitted file size can be achieved by creating ZIP or compressed tar files of the set of smaller files.

This approach accomplishes not only aggregation of small files but also reduction of data capacity footprint, because ZIP compresses by default, and tar can be induced to compress by adding a single character to the tar command. Thus, file footprints are reduced not only on tape but also on disk, which can have a significant impact on the usability of the PetaStore under heavy load, when the disk that serves as a staging area in preparation for migration to tape might otherwise approach or even reach full capacity. The ZIP/compressed tar mechanism also substantially reduces the contribution of drive-level compression (the default behavior), which can then be ignored as a factor in quota enforcement.

### 3.5. Store time vs retrieval time

For a given file, store time is substantially better than retrieval time. That is, the execution time of a user-level store command (for example, copying from supercomputer disk to the PetaStore) is only the time to transfer the data across the network and store it to PetaStore disk – the migration of a PetaStore disk file to tape (for every duplication option except disk_1copy_unsafe) occurs after the store command has completed (typically min to hours later). By contrast, the time to retrieve a file involves both the time to draw the file down from tape to disk (queue time, preparation time, file read time) and then the time to transfer the file from PetaStore disk to the destination storage platform. That is, total retrieval time is roughly equivalent to retrieval time from tape plus disk-to-disk store time, which is governed by the worst of (a) the performance of the PetaStore disk, (b) the performance of the destination disk and (c) the performance of the network.

### 3.6. Tape cartridge purchases

The hardware and software vendor specifies which tape cartridge brands and model numbers are permitted in the PetaStore, and which resellers these tape cartridge models can be purchased from, in order not to affect warranty coverage. The OSCER operations team works with research teams to assist in the purchase process, and to ensure that the correct models are purchased from the correct resellers. Tape cartridges can be shipped directly to OSCER, or, once an order of tape cartridges arrives, either the research team can transport them to OSCER staff, or OSCER staff can collect them from the research team.

### 3.7. Use agreement

The Oklahoma PetaStore is subject to a university-approved Use Agreement that constrains the use of the PetaStore as follows: (1) No files that are subject to the US federal Health Insurance Portability and Accountability Act (HIPAA) are permitted. (2) No files that are subject to the US federal Family Educational Rights and Privacy Act (FERPA) are permitted. (3) No files that are classified are permitted. (4) For any files that are subject to one or more Human Subjects Research agreements with any Institutional Review Board (IRB), including but not limited to OU's IRB, the user must take full responsibility for ensuring full compliance with any such agreement(s). (5) The PetaStore is only available for users who are at institutions within the US. (6) The PetaStore is subject to OU's Acceptable Use Policy. (7) If the use agreement signatory is a research team leader, it is the signatory's responsibility to ensure that any student members of their team comply with OU's Acceptable Use Policy. (8) OU disclaims any guarantee to continue providing the PetaStore. (9) If the PetaStore (or any successor) is withdrawn by OU, it is the signatory's responsibility to transfer any and all relevant files to other storage resources, and in a timely manner. (10) It is the signatory's responsibility to keep abreast of and comply with changes to any of the relevant laws, policies and circumstances described above.

## 4. Training

OU has extensive experience providing education, outreach and training to a broad variety of audiences [17–23], so development of the Oklahoma PetaStore training has fit straightforwardly into extant structures. Because of both obvious and subtle operational differences between the PetaStore and typical filesystems, and the risk that particular styles of use can detrimentally affect system performance and resiliency, one-on-one or one-on-few PetaStore training is required before access to the PetaStore is granted. The training session typically lasts 45 min, and consists of several core components.

The core competencies taught during training are as follows: (1) Log on to and navigate a remote Linux console via Secure Shell. (2) Transfer data to and from the PetaStore. (3) Meet the minimum file size requirements by utilizing tar or ZIP to group small files, while developing and employing a coherent schema for organizing such groups of files. (4) Make an educated decision on which duplication policy to use. (5) Interact with non-standard commands that are specific to the PetaStore, exposing the amount of media used and available, and the migration state (resident, premigrated, or migrated [24]) of a particular file. (6) Ensure faithful data transfers by calculating and comparing checksums at appropriate stages in a workflow. (7) Integrate these skills into a suitable workflow.

The approach to training is with the core competencies in mind, typically with a one-to-one or one-to-few trainer to attendee ratio, inviting questions at any time. Success of the PetaStore project as a whole hinges on adherence to the minimum/maximum file size policies and on system uptake, which is a function of successful integration of training knowledge into research workflows. Thus, much of the conversation is an engineering discussion toward that end – on a case-by-case basis, how can proper data archival practices be integrated into a research workflow to overcome the inconvenience of the technology?

To date, 93 users have received training. Over the production lifetime of the PetaStore, six administrators' interventions have been necessary, primarily to rectify infractions of the minimum file size policy.

Please cite this article in press as: S.P. Calhoun, et al., Large scale research data archiving: Training for an inconvenient technology, J. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.jocs.2016.07.005

### 4.1. Storage archive description

In order that proper workflows can be individually designed for each use case, a brief description of the PetaStore is needed. This description begins with the mix of storage hardware types and an overview of the internal data paths. Based on this description, workflows for the relevant use case(s) can be developed to more closely align with desired access patterns.

### 4.2. Tailoring the discussion

The relevant use case description(s) form the core driver of the training. In addition, the ownership of media (tape cartridges and/or disk drives) determines the choice of media type, and both ownership and data vulnerability inform duplication option choices (see Section 3.2).

Natural Dataset Size: Datasets typically have a natural aggregation size, and in the context of bulk data archiving, this size is typically in the desired 1–100 GB range. Establishing the aggregation size as a natural consequence of the research workflow leads to examining the issue of file size limits.

Data Proximity: Due to the high time cost of accessing datasets from remote repositories, it may be advantageous for a workflow to incorporate a local copy of the target files.

Data Vulnerability Assessment: The vulnerability of the data is assessed; for example, are these files already duplicated on a separate repository? This issue informs the choice of duplication option.

Data Sensitivity Assessment: The sensitivity of the data is assessed; for example, is this HIPAA data, FERPA data, classified data, human subjects data? This issue informs the appropriateness of the data to the PetaStore as well as possible adoption of encryption.

Additional Constraints: The technical environment to be used in the workflow can affect the delivery of training; for example, if the operating system on which the workflow is to be executed is Windows, which lacks a command line secure file transfer mechanism, then the training might incorporate FileZilla or WinSCP, whereas in a MacOS or Linux context, the scp command is more relevant.

### 4.3. System rules

File Sizes: Any file placed on the PetaStore must fall between the minimum and maximum permitted file sizes (see Section 3.3).

Approved Purchasing Channels: All media must be approved brands and models, purchased via approved resellers (see Section 3.6). This rule serves to ensure system consistency and compliance with the PetaStore's warranties and service agreements.

### 4.4. Duplication options

All nominally-available duplication options are presented, and their corresponding filesystem paths are explored. Even if a particular filesystem policy is inapplicable to the use case at hand, the function of that policy is examined, so that its availability informs possible future use cases.

Particular focus is given to the appropriate duplication level for the use case at hand, necessitating an ancillary discussion on how one should evaluate the duplication level options in order to determine the most appropriate option. In particular, best practices dictate that all archived data should either be stored with a minimum of two instances of the data (on any combination of storage systems), or be stored in one instance for cases where the data can be trivially reconstructed. In cases where a dataset can be readily acquired from external resources, such as from a public data collection (for example, at a regional, national or international data repository), it follows that only one instance needs to be stored on the PetaStore. In cases where data are a culmination of months of computationally- and/or labor-intensive research, it is likely that two instances (on two different physical media) ought to be stored on the PetaStore.

### 4.5. Compression and aggregation

The need to use compressed aggregation files (e.g., ZIP files, compressed tar files) on the PetaStore arises from its technological properties (see Sections 3.3, 3.4). Use cases often involve archiving thousands of small raw data files (much smaller than the minimum permitted 1 GB), dictating that those small files be bundled together as a single file prior to being transferred to the PetaStore. Additionally, the raw data files are often insufficiently or not at all compressed. In cases where the files to be archived reside on a Unix-like system (including but not limited to Linux), the natural choice of aggregation file format typically is compressed tar. Conversely, in cases where the data source is a Windows system (for example, a laptop), ZIP may be a more natural choice. Training includes hands-on exposure to an appropriate compressed aggregation file format, including at minimum, creation of the aggregation file, compression (if a separate step), and extraction of individual files from an aggregation file.

To facilitate the possibly unfamiliar and inconvenient steps of compression and aggregation in researchers' workflows, training includes suggestions on how data can be organized. For example, time series data may best be grouped by timestamp range (e.g., months for earth science data, millions of years for cosmological data), in order to meet a minimum file size while maintaining utility to a researcher. Alternatively, some hypothetical data may be best grouped first spatially (e.g., by geographical region), and then temporally, to be more useful for future analyses over a single region. The guidance in the training suggests that a coherent schema be developed by the researcher, in order to balance ease of archiving against ease of retrieving, the best solutions to which may or may not individually produce the same schema.

### 4.6. Interfaces

All available interfaces are discussed (supercomputer archive node, scp/sftp, gridftp, Globus). Even if a particular interface is inapplicable to the use case at hand, the function and merits of the interface are examined, so that its availability informs possible future use cases. Particular focus is given to the appropriate interface or interfaces for the use case at hand.

#### 4.6.1. Nonstandard commands

Due to the unusual nature of the PetaStore's implementation, a need arises for interaction with the system in ways that are not provided for on standard compute resources. To address such cases, the PetaStore's non-standard commands are discussed during the training.

The PetaStore's unusual combination of software technologies layers tape and disk on a single filesystem path for each file in either of the "tape_1copy_unsafe" or "tape_2copies" duplication option directories. Standard commands do not provide any exposure to which combination of disk and tape is currently in use for any given file, so a nonstandard analogue to the Unix "ls" command is provided, which reveals the migration state for each file (resident on disk only, pre-migated to tape but also still resident on disk, or migrated to tape only).

Additionally, proper management of one's allocated space on the system requires the ability to query the size of their allocated space, and to query how much of that space is used vs how much is unused. Standard commands are available to query such quota information for disk-based filesystems, but due to the layered

Please cite this article in press as: S.P. Calhoun, et al., Large scale research data archiving: Training for an inconvenient technology, J. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.jocs.2016.07.005

nature of disk and tape, no standard command accurately reflects the reality on the PetaStore. For this purpose, training includes instruction on the use of a nonstandard command that reports on total usage information for a researcher's media, by media type.

### 4.7. Hands-on exposure

As part of the training, actual data exemplars or, often, substitute files are archived. The exact method of archiving depends on the use case at hand, and involves accessing the PetaStore through one or more of the defined interfaces. In conjunction with this step, Unix commands are discussed that can be used as part of an archiving workflow. In particular, it is often the case that ssh, cd, ls, cp, mv, tar, zip, screen, and md5sum are discussed. Often during this portion of the training, the topic strays from technology that directly involves the PetaStore, to topics that otherwise address the requirements of the particular use case.

Discussion of md5sum entails a more rigorous examination of best practices for data validation and integrity testing throughout the workflow. With file sizes increasing in the range of tens to hundreds of gigabytes, transfer times increase in the range of minutes to hours, and the correlating risk of data stream interruption increases. In such an event, incomplete files can be written on any storage system, including the PetaStore. The training includes suggestions on when to validate the contents of a file (before archiving), and when to calculate and confirm checksums (initially at data creation/validation time, after transfer, and generally after migration to tape within the PetaStore).

## 5. Conclusion

Training for the Oklahoma PetaStore, a large scale, long term institution-scale research data archive, focuses substantially on the inconveniences associated with the technological choices, which were constrained by budget and therefore are inherently nonideal. This training continues to be a core activity in service of petascale storage capability at OU and across Oklahoma.

To date, the PetaStore has been serving 28 research groups, having incorporated over 50,000 files totaling almost 1 PB of content, for a mean of approximately 15 GB per file. Taking into account intentional file duplication, the total consumption of storage capacity is over 1.6 PB. There are approximately 1600 tape cartridges total (approximately 1200 in the PetaStore and approximately 400 at the offsite disaster recovery location), for an aggregate tape capacity of almost 2.5 PB so far. These are indicators of the value of the PetaStore to the institution and the state. Only 0.4% of data archived in the PetaStore is in files smaller than 1 GB (a mean of approximately 75 MB per such file), demonstrating the value of the training.

A final advantage of the PetaStore approach is that, although the PetaStore system will need to be refreshed in the near future, the PetaStore media (tape cartridges) will outlast the PetaStore system and can be transitioned physically to the PetaStore's successor, thereby minimizing user costs while keeping institutional and funding agency costs manageable.

## 6. Appendix

All hardware and software was purchased with 3 year warranties, funded by the NSF MRI grant. As noted above, after the initial warranties expired, OU IT began funding annual maintenance coverage.

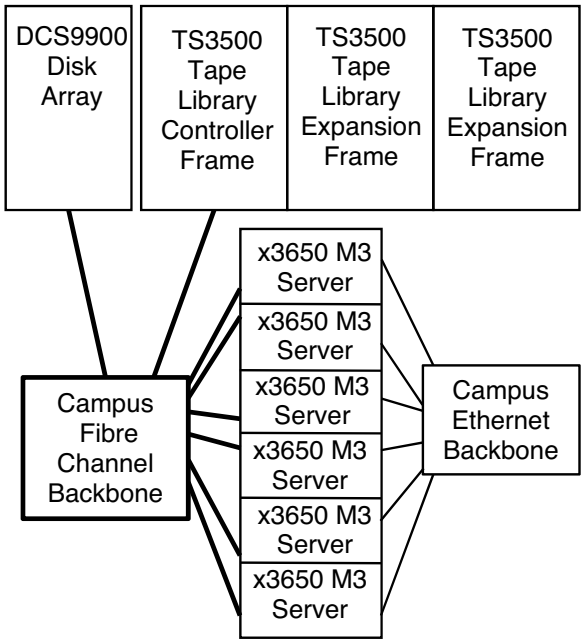



**Fig. 1.** Oklahoma PetaStore Hardware Conceptual Layout.

### 6.1. Hardware

Fig. 1 shows a conceptual diagram of the hardware layout, which is a common configuration for research data archives.

### 6.2. Disk hardware

The disk array is an IBM DCS9900 (rebranded DataDirect Networks S2A9900), which has dual controllers and was originally acquired with 1200 disk drive slots (20 enclosures of 60 disk drive slots each) and 300 2 TB SATA 7200 RPM disk drives (the minimum that could be acquired at initial purchase). After the initial purchase, another 230 2 TB SATA drives were acquired by individual research teams and were deployed within the DCS9900, for a total of 530 disk drives and approximately 830 TB of useable capacity. (However, because the DCS9900 model was discontinued and so was the ability to purchase additional disk drives without affecting warranty coverage, the disk drives were therefore consolidated into 600 disk drive slots, and ten of the twenty disk drive enclosures were decommissioned, to reduce annual maintenance costs. In addition, IBM provided a small DCS3700 disk array, which is a rebranded NetApp E5400, for additional slot and growth capacity.)

### 6.3. Tape hardware

The tape library is an IBM TS3500 with one L53 base frame with 12 tape drive slots and 219 tape cartridge slots, as well as two S54 high density cartridge-only expansion frames with 1320 tape cartridge slots each, for a total of 2859 tape cartridge slots, plus (originally) four LTO-5 tape drives and (purchased more recently) three LTO-6 tape drives.

The aggregate theoretical peak speed of the tape drives is just over 1 GB/s (each LTO-5 tape drive has a peak speed of 140 MB/s and each LTO-6 tape drive has a peak speed of 160 MB/s). Because LTO-6 drives can read and write LTO-5 tape cartridges, but only at LTO-5 speeds, the aggregate peak speed for LTO-5 tape cartridges is just under 1 GB/sec; on the other hand, LTO-5 tape drives cannot read nor write LTO-6 tape cartridges, so the aggregate peak speed for LTO-6 tape cartridges is 480 GB/sec, roughly half that of LTO-5.

Please cite this article in press as: S.P. Calhoun, et al., Large scale research data archiving: Training for an inconvenient technology, J. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.jocs.2016.07.005

Until recently, LTO-6 tape cartridges had cost significantly more per TB than LTO-5. As such, between higher cost and lower aggregate peak speed on the PetaStore, there was no incentive to pursue LTO-6 tape cartridges. However, once LTO-6 tape cartridge prices reached approximate parity with LTO-5 on price per TB, which occurred in mid-2015, LTO-6 became a recommended option.

### 6.4. Servers

The six servers that control the tape library and the disk array are IBM model x3650 M3, with dual Intel Xeon E5620 "Westmere" CPUs (quad core, 2.4 GHz, 1066 MHz memory speed), 24 GB RAM ($6 \times 4$ GB DDR3 1333 MHz), dual disk drives in RAID1 (mirrored) configuration for the operating system and software (each 300 GB, SAS, 10,000 RPM, 6 Gbps), QLogic Fibre Channel 8 Gbps dual-port host bus adapter, Chelsio S320E 10 Gbps Ethernet (10GE) dual-port adapter, and dual power supplies for redundancy. Four of the servers control the disk array, and the other two control the tape library.

### 6.5. Networks

#### 6.5.1. Intra-campus networks

Intra-campus Fibre Channel: The servers, disk array and tape library are all connected to OU's Fibre Channel backbone – dual connections per server for the four disk array servers and quad connections per server for the two tape library servers (16 connections total), some 4 Gbps and some 8 Gbps – for data traffic within the PetaStore (that is, among the PetaStore components), instead of having a set of Fibre Channel switches dedicated exclusively to the PetaStore. This approach was chosen as a cost savings measure, given that the PetaStore's Fibre Channel traffic is sufficiently modest that PetaStore data transfers don't have a significant negative impact on other Fibre Channel backbone uses (nor vice versa).

Intra-campus Ethernet: The servers are also connected to OU's campus Ethernet backbone network at a single 10 Gbps connection per server (six connections total). These connections facilitate data traffic into and out of the PetaStore, especially (but by no means exclusively) between the PetaStore and globally accessible user filesystems on OU's cluster supercomputer. Idealized benchmarks of the PetaStore disk array have shown the aggregate sustained disk bandwidth (capacity per unit time) to be approximately 4 GB/s (approximately 32 Gbps), but individual data transfers have achieved only a few hundred MB/s (a few Gbps).

#### 6.5.2. Inter-campus network

The PetaStore is part of a larger collection of resources distributed among multiple institutions in Oklahoma, specifically OU, Oklahoma State University (OSU), the Oklahoma Innovation Institute (a nonprofit corporation), Langston University (Oklahoma's only Historically Black University) and the University of Central Oklahoma (a non-PhD-granting university). These institutions form the service provider core of the OneOklahoma Cyberinfrastructure Initiative (OneOCII) [25], an informal but extremely active collaboration that to date has served over 100 institutions and organizations in Oklahoma, including over 50 academic institutions, with a broad variety of technological capabilities as well as education, outreach, training and workforce development activities.

### 6.6. Software

#### 6.6.1. Operating system

The servers run Red Hat Enterprise Linux version 6.6.

#### 6.6.2. Disk array software

The disk array uses IBM's General Parallel Filesystem (GPFS) [26], also known as IBM Spectrum Scale, specifically GPFS Server on the four disk array servers and GPFS Client on the two tape library servers. The version currently in production is 3.5.0.9. (A significant portion of the disk array is set aside for OU's High Energy Physics group. This portion runs the Lustre parallel filesystem [27–29], completely independently of the GPFS implementation, and is used as live storage for a working set of Large Hadron Collider [30] and similar datasets.)

#### 6.6.3. Tape library software

The tape library is controlled by IBM's Tivoli Storage Manager (TSM) [31], also known as IBM Spectrum Protect, specifically TSM Extended Edition and TSM Space Management, both of them version 7.1.1. TSM products are installed on the two tape library servers only, not on the four disk array servers. The use of TSM is unusual for a long term research data archive, because TSM was originally designed as a backup product, not as an archiving product, with Hierarchical Storage Management (HSM) features added in more recent versions.

In fact, IBM has a software product, High Performance Storage System (HPSS) [32], specifically designed for this kind of data archiving. HPSS is popular at, for example, national-scale academic and government supercomputing centers. Unfortunately, at the scale of an institutional archive such as the PetaStore, HPSS's price point is much higher than TSM's, because TSM and GPFS are charged per server (and the number of servers required for a tape archive is very modest), whereas HPSS is charged as a fixed cost for the entire archive, and the total pricing for HPSS (as quoted to OU by IBM) would have been near the total project budget; that is, HPSS is priced appropriately for a 7–8 figure national-scale storage archive, but not for a 6 figure institutional-scale storage archive.

## Acknowledgements

Portions of this material are based upon work supported by the National Science Foundation under the following grants: Grant No. OCI-1039829, "MRI: Acquisition of Extensible Petascale Storage for Data Intensive Research;" Grant no. EPS-1301789, "Adapting Socio-ecological Systems to Increased Climate Variability;" Grant no. ACI-1341028, "OneOklahoma Friction Free Network;" Grant no. ACI-1440783, "A Model for Advanced Cyberinfrastructure Research and Education Facilitators;" Grant no. EPS-1006919, "Oklahoma Optical Initiative." Additional support was provided by the University of Oklahoma.

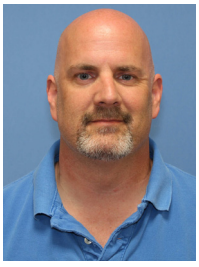
The authors are grateful to members of the OU Information Technology team for their contributions to designing and deploying the Oklahoma PetaStore, and to the OU MRI grant team for their contributions to obtaining and implementing the MRI grant. Some of the text appeared in somewhat modified form in various grant proposals.

## References

[1] White House Office of Science and Technology Policy, 2013. Increasing Access to the Results of Federally Funded Scientific Research. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013. pdf.
[2] National Science Foundation, Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation. 2015. https://www.nsf.gov/publications/pub_summ.jsp?ods_key = nsf15052.
[3] Pricewatch USB 6 T B USB 3.0 disk drive pricing webpage. http://www.pricewatch.com/price/hard_removable_drives/usb_6tb.
[4] Pricewatch LTO-7 media pricing webpage. http://www.pricewatch.com/price/media/lto-7.

[5] NASPO ValuePoint IBM Storage Hardware for Contract MNWNC-116: Update as of February 03, 2016. http://www-304.ibm.com/shop/americas/content/pdfs/naspo/NASPO_Storage_Hw_02032016_M-500k.pdf.
[6] NASPO ValuePoint IBM Storage Software for Contract MNWNC-116: Update as of February 10, 2016. http://www-304.ibm.com/shop/americas/content/pdfs/naspo/NASPO_Storage_Sw_02-10-2016_Master-500k.pdf.
[7] Lenovo System X x3650M5 server webpage. http://shop.lenovo.com/us/en/systems/servers/racks/systemx/x3650-m5/.
[8] S.P. Calhoun, D. Akin, J. Alexander, B. Zimmerman, F. Keller, B. George, H. Neeman, The Oklahoma PetaStore: A Business Model for Big Data on a Small Budget, Proc. XSEDE'14, article 48, 2014. DOI: 10.1145/2616498.2616548.
[9] Snowark, Decrypting hard-drive failures –MTBF and AFR. 2013. http://research.snowark.com/blog/2013/02/11/decrypting-hard-drive-failures-mtbf-and-afr/.
[10] B. Beach, Hard Drive Reliability Update – Sep 2014. 2014. https://www.backblaze.com/blog/hard-drive-reliability-update-september-2014/.
[11] Oklahoma PetaStore webpage. http://www.oscer.ou.edu/petastore/.
[12] University of Washington Shared Central File System for Research Archives (lolo Archive) webpage. https://itconnect.uw.edu/service/shared-central-file-system-for-research-archives-lolo-archive/.
[13] University of Colorado Boulder PetaLibrary webpage. https://www.rc.colorado.edu/resources/storage/petalibrary.
[14] I. Foster, Delivering a Campus Research Data Service with Globus Presentation at MAGIC Meeting, slide 53 http://www.slideshare.net/ianfoster/globus-status-and-publication-plans. 2014.
[15] I. Foster, Delivering a Campus Research Data Service with Globus Presentation at MAGIC Meeting, Slide 53, 2014 (http://www.slideshare.net/ianfoster/globus-status-and-publication-plans).
[16] Globus Online website. https://www.globus.org.
[17] C. Carley, B. McKinney, L. Sells, C. Zhao, H. Neeman, Using a shared, remote cluster for teaching HPC, Proc. IEEE Cluster (2013), http://dx.doi.org/10.1109/CLUSTER.2013.6702630.
[18] A. Fitz, P. Gibbon, D.A. Gray, T. Joiner, H. Murphy, R.M. Neeman, C. Panoff, S. Peck, Thompson, Teaching High Performance Computing to Undergraduate Faculty and Undergraduate Students, Proc. TeraGrid'10, article 7. 2010. DOI: 10.1145/1838574.1838581.
[19] H. Neeman, H. Severini, D. Wu, K. Kantardjieff, Teaching High Performance Computing via Videoconferencing, ACM Inroads, 1 (1), 67–71 (2010). DOI: 10.1145/1721933.1721954.
[20] H. Neeman, H. Severini, D. Wu, K. Kantardjieff, Teaching Supercomputing via Videoconferencing, Proc. TeraGrid. 2008.
[21] H. Neeman, H. Severini, D. Wu, Supercomputing in Plain English: Teaching Cyberinfrastructure to Computing Novices, inroads: SIGCSE Bulletin, 40 (2), 27–30. 2008. DOI: 10.1145/1383602.1383628.
[22] H. Neeman, L. Lee, J. Mullen, G. Newman, Analogies for Teaching Parallel Computing to Inexperienced Programmers, inroads: SIGCSE Bulletin, 38 (4), 64–67. 2006. DOI:10.1145/1189136.1189172.
[23] H. Neeman, J. Mullen, L. Lee, G. Newman, Supercomputing in Plain English: Teaching High Performance Computing to Inexperienced Programmers, Proc. 3rd LCI International Conference on Linux Clusters: The HPC Revolution 2002. DOI: 10.1.1.329.6450.
[24] IBM Knowledge Center: Tivoli Storage Manager 7.1.1 Documentation: Glossary. http://www.ibm.com/support/knowledgecenter/SSGSG7_7.1.1/com.ibm.itsm.ic.doc/glossary.html.
[25] H. Neeman, D. Brunson, J. Deaton, Z. Gray, E. Huebsch, D. Gentis, D. Horton, The Oklahoma Cyberinfrastructure Initiative, Proc. XSEDE'13, article 70. 2013. DOI: 10.1145/2484762.2484793.
[26] IBM General Parallel Filesystem (GPFS) webpage. http://www-03.ibm.com/software/products/en/software.
[27] Lustre website. http://lustre.org.
[28] OpenSFS Lustre webpage. http://opensfs.org/lustre/.
[29] Wikipedia Lustre webpage. https://en.wikipedia.org/wiki/Lustre_(file_system).
[30] Large Hadron Collider webpage. http://home.web.cern.ch/topics/large-hadron-collider.
[31] IBM Tivoli Storage Manager webpage. http://www-03.ibm.com/software/products/en/tivostormana.
[32] IBM High Performance Storage System website. http://www.hpss-collaboration.org.

**S. Patrick Calhoun** is a System Administrator for the OU Supercomputing Center for Education and Research, focusing on the Oklahoma PetaStore. He received his Bachelor of Science degree in Computer Engineering from the University of Oklahoma in 2006. He has worked for OU in his current position since 2011.



**David Akin** is a Senior System Administrator at the OU Supercomputing Center for Education and Research at the University of Oklahoma (OU). He received his Bachelor of Science in Computer Science from the University of Science and Arts of Oklahoma in 1996. He has worked for OU in his current position since 2005.

**Brett Zimmerman** has been a Unix and Linux systems administrator and systems programmer for 20 years, the last 9 at the OU Supercomputing Center for Education and Research at the University of Oklahoma (OU), working on the design and management of cluster supercomputers, user education, and optimization of scientific code for the university's computing environments.

**Dr. Henry Neeman** is the founding Director of the OU Supercomputing Center for Education and Research, Assistant Vice President for Information Technology - Research Strategy Advisor, Associate Professor of Engineering, and Adjunct Associate Professor of Computer Science (CS) at the University of Oklahoma (OU). He received his Bachelor of Science in CS and his Bachelor of Arts in Statistics from the University at Buffalo, State University of New York in 1987, his Masters of Science in CS from the University of Illinois at Urbana-Champaign (UIUC) in 1990 and his doctorate in CS from UIUC in 1996.

Please cite this article in press as: S.P. Calhoun, et al., Large scale research data archiving: Training for an inconvenient technology, J. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.jocs.2016.07.005