Overview of the Active Archive Facility (AAF) Service Offering

Fermilab provides an active (aka nearline) archival service for external customers to store their scientific data. This tape based storage system has a capacity of tens of Petabytes and an aggregate I/O bandwidth of multi GB/s. We assume tape storage to be permanent and provide custodial care for data on tape (see Appendix II for a description of what we do to maintain data integrity for custodial service).

Customers are charged yearly according to a cost model that is based on their projected and actual usage. The cost model includes a charge for the total amount of data on tape at the end of the year, a charge for the media needed for the year, and a charge for tape-drive hours used. We provide daily and monthly usage metrics for the customer through a web portal.

Customers can also opt for a dedicated disk cache that is integrated with the tape storage. This cache uses a Least Recently Used algorithm to decide on which files to remove for making room for new ones. Thus files in the cache have a finite lifetime. The disk cost is a yearly rate based on how many TB are dedicated to the customer. If this option is not chosen, then the customer will be added to a small, tape-backed, shared disk cache.

Remote access to files can be accomplished through several Wide Area Network protocols, including the Storage Resource Manager (SRM), GSI FTP, or WebDAV. Access via the xrootd protocol is also supported. SRM provides a means to manipulate files and their metadata (ls, rm, mv)

Each customer is identified by one or more "storage groups". Customer metrics and monitoring are provided on a per storage group basis.

Strategic Partnership Project Agreement

AAF service is provided through a Strategic Partnership Project (SPP) agreement between the customer and Fermilab. The SPP is a package that consists of a Statement Of Work (SOW) and Terms and Conditions. These are multi-year agreements and are generally written for a five-year period, with yearly adjustments/updates as necessary. The SOW contains the expected usage by the customer, estimated charges for each year, and agreed responsibilities of each party. Customers are charged each year in advance.

Customer Contact/Administrator

The customer must designate an administrator for its users. The responsibilities of this administrator are:

- To be the customer contact to AAF for reporting incidents, making requests and responding to any technical or usage issues.
- To receive notifications of scheduled and unscheduled outages and disseminate this information to its users in a timely fashion.
- To convey proper usage information to users, such as what kinds of data are permitted to be stored and what kind not. The customer administrator is responsible for assuring its users abide by the Computing Sector's policies on storage and that they follow the recommendations for proper use of the storage systems.
- To understand the total amount of space required by its users, the types and sizes of data they will store, and the rates they will store and retrieve the data.
- To define, in consultation with storage services, the directory hierarchy, including file families (see below), for its users and to identify any directories whose data is critical enough to be duplicated. To maintain the directory hierarchy and request updates as needed (new branches and file families).

The contact will also be given permission to send to a mailing list that will automatically generate a ticket in our incident tracking system. The address of this list is: aaf-incident@lfnal.gov . There is also a request mail that can be used for requests: aaf-request@fnal.gov.

In addition, the customer needs to designate a financial contact (which may be the same as the customer contact), to act as the financial intermediary for the purposes of budgeting and billing.

Laboratory and Computing Sector and Storage Service Policies

The scientific storage services at Fermilab are intended for the storage of only scientific data. Storage of personal information or protected Personally Identifiable Information (PII) as defined by the Fermilab PII Policy [1] are strictly prohibited.

It is up to the customer to assure that the data stored at Fermilab by their users:

- Does not support legally prohibited activities
- Is not offensive and will not result in public embarrassment to Fermilab
- Is not sexually explicit
- Does not violate a license or other computer related contract provisions
- Is not Personally Identifiable Information (according to the Fermilab PII policy)
- Fermilab has the right to remove any data not meeting these criteria.

Appendix I: For the Customer Administrator

A customer's files are not mixed up with other customer's files on tape. When a customer has a tape mounted for access, they have exclusive use of that tape drive from the time when the tape is mounted until when it is dismounted. Since tape drives are a limited resource, customer's are charged by the tape drive-hour for their use. Inefficient use of drives can have an overall system performance impact and a cost impact for the customer.

Latency and Rate

When a request to read or write is made to the AAF, a tape must be mounted and threaded to its load point, then a seek to the location of the file on tape is made. The mount time depends on the load of the system – the quickest time to load is about 45 seconds. A seek from the beginning of tape takes about 50 seconds on average and a seek from one location on tape to another random location is about 33 seconds. On average, you should expect the best time to access a single file on tape to be about 2 minutes. It is best to read or write as many files as possible while the tape is mounted. For writes, this means supply files at a high enough frequency that a number of them will be written per mount. For read accesses, files written about the same time will generally reside on the same tape, so accessing multiple files from around the same time will yield the best performance.

File size

It takes on the order of 5 seconds between writing a file mark and starting the next file write. File sizes should be large enough to keep a tape drive streaming for a significant amount of time. Our recommend file size is several Gigabytes, though smaller files down to 250MB are acceptable if the average is in the Gigabyte range.

We ask that experiments not write small files (according to the above guidelines). In order not to impact performance, resource availability, or drive life, if you have small files, you should tar them up into larger files if possible.

Storage Group and File Families

Each customer is assigned a storage group for writing to tape. This is usually a short experiment or organization name (e.g. "minos" "CMS" "CDF'. The storage group is used for accounting and also for operational functions such as assigning tapes

AAF also has the concept of file families. File families are "tags" that the customer can specify in the namespace used by AAF. Each tape volume being written has a volume family (storage group + file family) associated with it. Only files with the same volume family can be written to the tape. In addition, file families can have file family widths (with a default of 1) associated with them. If the file family width is N, then Enstore can write up to N tapes with that file family in parallel.

While file families are useful tools for grouping together similar files on tape, their use must be considered very carefully. If you have too many file families, many of the tapes may never be filled, wasting space and costing money.

A default file-family equal to the storage-group is assigned to each customer. If the customer has a need, or Fermilab detects a need for specific file families, they can be arranged through a consultation.

File Duplication

Files deemed to be critical can be automatically duplicated to separate tapes (and in separate libraries and in separate buildings). This is done on a per directory basis (inherited by sub-directories). Duplicated data costs twice as much as unduplicated data. When files are duplicated, one file is considered the primary and the other the secondary. User's can only access the primary copy. Storage Administrators can swap the role of primary and secondary copies if necessary.

Namespace

The storage system uses a namespace for identifying files that is hierarchical and file-system like. Customers are responsible for mapping their files to this namespace and cataloging the mapping. The namespace used by storage is not meant to be a catalog - for performance considerations, using it as such is not permitted. The namespace used by AAF is called Chimera (also known as pnfs).

Monitoring

Customer monitoring pages are provided. These include:

1. A complete file listing.
2. Daily accumulated TB on tape
3. Daily Tape drive-hour usage
4. Historical monthly usage summaries of the above

Any access goes through one or more intermediary disk caches. For this reason, files written from offsite may not immediately appear on tape. We provide the complete file listings and other tools for the customer to determine when their files are safely on tape.

Appendix II: Availability and Custodial Service for data on tapes

Availability

Storage systems require regular monthly maintenance. Maintenance is usually performed on the third Thursday of each month. Sometime it is necessary to have an off-schedule maintenance and we strive to give as much notice in advance as possible – a week at least if not an emergency. These downtimes are announced via a mailing list that includes the customer contact.

The expected availability of AAF is around 99% (this assumes only planned maintenance outages). This allows for 7 hours downtime per month, though the downtime are seldom this long.

The Storage Service Administrators are on call 24x7 for storage related issues. Customer Incident and Requests are 8x5.

Custodial Care

By default, the Scientific Computing Division assumes the retention period for data on tape to be permanent and that the owners of the data will bear the cost of maintaining the integrity of this data. The cost includes maintenance costs, operation costs and the costs associated with migration to newer and/or denser media over time.

AAF has numerous features built in to maintain the integrity of data on tape, and the Storage Services Administration group has procedures to maintain the integrity:

1. The tape storage software uses end-to-end checksuming to ensure the data of each file has remained unchanged. This checksum is generated when the users data lands on our disk cache (or at the customers end if using SRM), is calculated and verified before being written to tape, is stored in the file's metadata, and is calculated on every read from tape. This assures that the data written to tape is exactly the same as retrieved from tape.
2. When a tape gets filled, it will be temporarily ejected and have it's write protection tab physically locked to protect against accidental erasure. Jobs to write lock batches of filled tapes are generated automatically when enough filled tapes have accumulated, or two weeks has passed, whichever comes first.
3. Tapes are randomly sampled and the first, last and a random middle file on the selected tapes are read (and implicitly their checksum verified).
4. As files are written to tape, periodically once a file write is complete, the tape is backspaced to the file and it is read back. This verifies the checksum from

tape immediately after the file was written. We do this for about one file per tape.

5. The number of mounts for each tape is maintained in a database and alarmed if a threshold is crossed. If the threshold is crossed, the tape is "cloned"; the files are copied to a fresh new tape. This protects against data loss due to tape wear.
6. Data is refreshed on an approximately 5-6 year cycle by migrating to new denser media . (New generations of media that are about 2 times denser are rolled out around every 3 years, so we migrate to 3-4 times denser media every 6 years.)
7. Since tapes and tape drives are mechanical devices, (very) infrequently tape drives can damage tapes, a tape cartridge can fail mechanically, or a portion of the media can be damaged from wear. In these cases it usually requires vendor intervention to read past the bad spots on the tape to recover the data. If recovery seems possible, we send damaged tapes to data recovery vendors to recover what they can from the tape and restore the recovered data to new tape.
8. There are tape libraries at two computing centers at Fermilab that are separated by about a mile. Users can maintain two geographically separated copies of subsets of their files at these two centers. We can arrange for the customer to do this on their own, or the customer can use an AAF feature to do this automatically. This provides extra protection against tapes getting damaged by drives and against a disaster at one of the sites. It is highly recommended that two or more copies be made for with small amounts of data (such that loosing a tape constitutes a significant statistical loss), since, though very rarely, a tape drive can damage a tape or a tape cartridge will fail.
9. We monitor tapes at the level of soft errors (recovered reads or writes) and we use this information to proactively detect tape integrity issues.

References

[1] Fermilab Personally Identifiable Information Policy, http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=2134

[2] Fermilab Policy on Computing, http://security.fnal.gov/policies/cpolicy.html