



# **Tianlai survey and Fermilab Scientific Computing Division (SCD)**

9/27/2016

Stu Fuess, Margaret Votava

Fermilab / SCD

# What we know about your survey (1/2)

- Programmatic background
  - It is our understanding that this has been presented to the Fermilab PAC ([1/20/2016](#), [6/20/2016](#), [6/21/2016](#)) as a component of the Theory strategic plan
    - The recommendations ([1/2016](#), [6/2016](#)) of the PAC did not address lab support of this effort
      - LDRD proposal was not funded
      - We conclude that there are no direct lab support funds
  - We understand that there is a 3-year NSF [award](#) that could potentially provide funding
- It needs to be clear that the SCD cannot provide resources or effort utilizing base program funds; the SCD thus can...
  - direct you to available tools
  - provide resources chargeable to a supplied budget code
  - provide consulting services chargeable to a supplied budget code

# What we know about your survey (2/2)

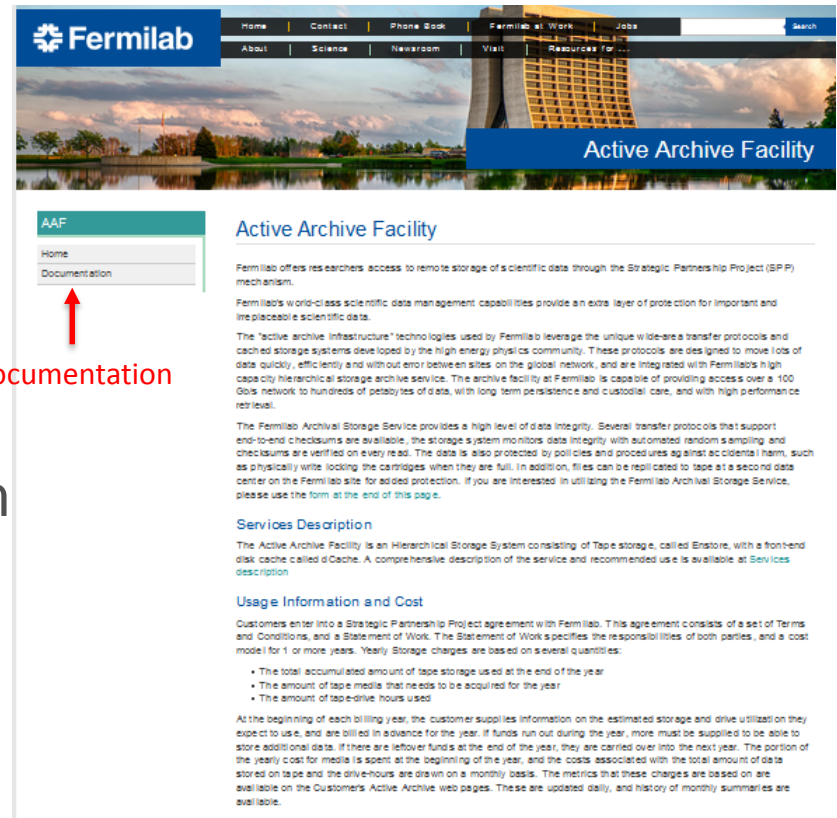
- Technical background
  - The SCD had a presentation from Albert Stebbins on 11/5/2015
    - See this [link](#) for the talk
    - See this [link](#) for meeting minutes
  - and has also had updates in advance of this meeting
- Data production and analysis
  - Roughly 100 MByte/s of correlation streams (TOD)
    - written to disk at site (eg 4TB disk fills in ~11 hours)
  - Sets of disks (how many?) shipped to US (Fermilab or other?)
  - Disks read, data imported to Fermilab disk cache and tape
    - Estimate 1.6 PByte/year total import (130 TB/month max rate)
    - Equivalent to average rate of 50 MByte/s
  - Expect to utilize opportunistic processing (eg OSG) for analysis

# Numbers

- 1.6 PByte in 4 TByte disks -> 400 disk imports
  - Equivalent of ½ year of 100 MByte/s data acquisition
- Noted that TOD to ASD step is embarrassingly parallel
  - but expect will inject a complete TOD file for production on a grid worker node, which for OSG opportunistic is a single core
  - parallelism may be best exercised by processing multiple files
- Data types:
  - TOD 1.6 PByte/yr (e.g. 400x 4 TB disks)
  - ASD 4 TB
  - Maps 1 TB

# Storage Resources

- Fermilab uses dCache disk as a cacheing layer in front of enstore tape storage
  - We would suggest that Tianlai purchase resources within the [Active Archive Facility](#) (AAF)
    - gridftp, xrootd, NFS, etc access methods
    - Ingest to disk from system(s) that mount the data disks
    - Automatically goes to tape
    - Cache provides buffer to/from tape
      - I/O and cache file lifetime needs determine cache size



**Fermilab**

Home | Contact | Phone Book | Fermilab at Work | Jobs

About | Science | Newsroom | Visit | Resources For

## Active Archive Facility

**AAF**

- Home
- Documentation

**Documentation**

### Active Archive Facility

Fermilab offers researchers access to remote storage of scientific data through the Strategic Partnership Project (SPP) mechanism.

Fermilab's world-class scientific data management capabilities provide an extra layer of protection for important and irreplaceable scientific data.

The "active archive" infrastructure technologies used by Fermilab leverage the unique wide-area transfer protocols and cached storage systems developed by the high energy physics community. These protocols are designed to move lots of data quickly, efficiently and without error between sites on the global network, and are integrated with Fermilab's high capacity hierarchical storage archive service. The archive facility at Fermilab is capable of providing access over a 100 Gbit/s network to hundreds of petabytes of data, with long term persistence and custodial care, and with high performance retrieval.

The Fermilab Archival Storage Service provides a high level of data integrity. Several transfer protocols that support end-to-end checksums are available, the storage system monitors data integrity with automated random sampling, and checksums are verified on every read. The data is also protected by policies and procedures against accidental harm, such as physically write locking the cartridges when they are full. In addition, files can be replicated to tape at a second data center on the Fermilab site for added protection. If you are interested in utilizing the Fermilab Archival Storage Service, please use the form at the end of this page.

#### Services Description

The Active Archive Facility is a Hierarchical Storage System consisting of Tape storage, called Enstore, with a front-end disk cache called dCache. A comprehensive description of the service and recommended use is available at [Services description](#)

#### Usage Information and Cost

Customers enter into a Strategic Partnership Project agreement with Fermilab. This agreement consists of a set of Terms and Conditions, and a Statement of Work. The Statement of Work specifies the responsibilities of both parties, and a cost model for 1 or more years. Yearly Storage charges are based on several quantities:

- The total accumulated amount of tape storage used at the end of the year
- The amount of tape media that needs to be acquired for the year
- The amount of tape-drive hours used

At the beginning of each billing year, the customer supplies information on the estimated storage and drive utilization on they expect to use, and are billed in advance for the year. If funds run out during the year, more must be supplied to be able to store additional data. If there are leftover funds at the end of the year, they are carried over into the next year. The portion of the yearly cost for media is spent at the beginning of the year, and the costs associated with the total amount of data stored on tape and the drive-hours are drawn on a monthly basis. The metrics that these charges are based on are available on the Customers' Active Archive web pages. These are updated daily, and history of monthly summaries are available.

# Storage Costs (take a deep breath...)

- With the assumptions:
  - 1.6 PBytes per year
  - Equates to an average of <50 MByte/s> purely for data ingest
- Then AAF costs are estimated to be:
  - \$32/TB, including overheads, for media
    - 1.6 PB → \$51.2K
  - \$13/TB/year, including overheads, labor, maint., ...
    - 1.6 PB → \$20.8K for year 1, 2x that for year 2 if another 1.6 PB, etc
  - \$149/TB/year for disk cache, including overheads, labor, maint., ...
    - To get 30-day lifetime with <50 MB/s> → 130 TB → \$19.4K/year
  - \$0.96/drive-hour
    - To ingest 1.6 PB at 50MB/s per drive → ~9K hours → \$8.5K/year
    - Add appropriately for reads from tape (hopefully small)
- **Net disk/tape cost for 1.6 PB is ~ \$100K per year**



# Processing Resources

- Without explicit funding to purchase resources or contribute to shared resources, only option is to use opportunistic resources
  - Available within GP Grid or OSG
    - Location choice may depend on I/O needs
      - or more explicitly, ratio of I/O to processing
- Be aware of the default grid job limitations:
  - single CPU core/thread
  - 2 GBytes (2000 MBytes) memory
  - ~40 GBytes local disk
- The job defaults can be overridden, but...
  - “Effective” job slot usage is 2x, 3x, etc
  - Harder to acquire “fill in the holes” opportunistic resources
- Effectively no associated costs beyond “consulting”
  - see next pages...

# Available services

- The sector provides a catalog of services in SNOW (the service desk software interface).
  - Complete list is [here](#)
    - Email lists
    - Backup services
    - Database hosting
    - etc
  - Scientific only list is [here](#).
    - Data catalog tools
    - Batch job submission wrappers/monitoring
    - Source code repositories
    - Electronic log book.
    - etc





# Services of potential interest to you

- Scientific Computing Systems / Interactive Server Facility
  - *to get an interactive node (GPCF, and/or to configure "disk ingestion" machine)*
- Distributed Computing / Batch Job Management
  - / Community On-Boarding (consulting)
  - / User Jobs Monitoring
    - *for submitting/monitoring grid jobs*
- Scientific Data Management / IFDHC
  - *tools for moving data around*
- Scientific Data Storage and Access
  - / Active Archive Facility
  - / dCache Disk Cache Storage
  - / Enstore Tape Storage

# Other services/tools of interest...

- Scientific Data Management / FTS (File Transfer Service)  
/ SAM (Sequential Access via Metadata)

Depending on the number of files that the survey will manage, consider a data management system

- The FTS service manages file transfers; this is a possible tool for use on the ingest from the raw data disks
- The SAM service associates the files with metadata, lists all replica locations, and allows for dataset definitions via metadata queries

# Cost of services

- Setup cost – consulting hours by service management
  - Accrued on an hourly basis
  - \$150/hour (fully burdened) for highly experienced staff
  - Charged against a billable task code.
- Maintenance cost
  - A small annual cost [tiny fraction of FTE]
  - Depends on particular service(s). Can discuss if you are interested in pursuing.
- Experiment needs to provide a point of contact to receive computing related announcements.

# Conclusion

- The relationship with the SCD will ultimately hinge on funding. We have no headroom to support anything outside of the funded CMS, DES, and Intensity Frontier programs (and even that fenced funding is insufficient).
- We have tools, expertise and can consult on resources - but cannot devote any effort unless reimbursed.
- We can continue to help describe these and give cost estimates.
- In many cases it is hoped that you can find the effort within your collaboration that, given modest guidance (at hopefully modest cost), can provide most of the needed functionality.
- Hardware resources, and the effort to configure such, will require funding.