

High Energy Physics at Fermilab: Current Practices and Future Directions in Data Storage and Movement

Donald Petravick

Fermi National Accelerator Laboratory

P.O. Box 500, Batavia, IL, 60510

Phone: + 01-630-840-3935 FAX: + 01-630-840-2783

E-mail: Petravick@fnal.gov

Presented at the THIC Meeting at the Embassy Suites Hotel
Denver South

Englewood CO 80112

on June 27-28, 2000

The Premier Advanced Recording Technology Forum

THIC Inc.

Basic approach seen in HEP

- A Statistical Science - Very, Very Large number of outcomes of a repeated experiment
 - Each outcome ~1MB
 - Some data loss is tolerable
 - Amount of data which can be stored and processed constrains the design of experiments

Beer Bottle Metaphor

- We have a large number of small things, like bottles of beer in a brewery
- Some breakage can be traded off to gain system level economies.
- For handling, you want the chunk size to be vary:
 - Bottle, case, hand-truck, delivery-truck, boxcar
- Contrary case : traditional supercomputing

Main design desiderata

- Low cost per byte stored
- Many independent streams at one time, the data rate on any one of them is not high
- High ensemble throughput of data to and from tape 24x7
- Interchange amongst labs and universities.
- Volume capacious enough justify ATL

Other Characteristics

- Since the data sets primarily reside on tape, high level software is structured around the idea of iterating over tapes, getting the data in bunches.
 - (attempts to hide the tape completely have typically failed)
- Many of our problems are “embarrassingly parallel”, we are used to exploiting this by building parallel systems from lowest-price components.

Past experiments



- Exabyte 8200->Eliant
- \$1.00 GB.
- Hand-mounted tapes, 24x7 operators.
- Few 10**6 volumes.
- Some DLT as well.
- Drives attached locally
- Platform independent I/O library

Past Experiments

Some use of STK equipment

IBM 3494/3590 with HPSS software as a demonstration,

Deployed as a tape based cache

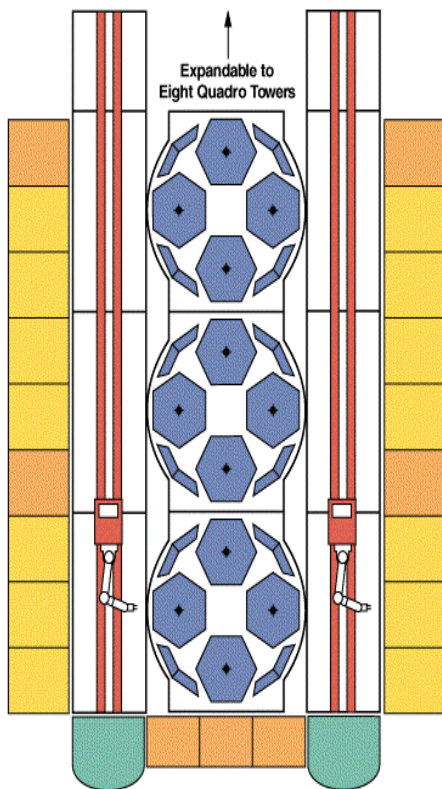
30 TB in total

Other HEP labs automated earlier than FNAL

Run II: petabyte/year

- Strategy was to flexible w.r.t media, how for low cost/byte stored
- Acceptable quality
- Low cost tape drive
 - Easy to interchange with universities
 - Easy to adjust tape bandwidth as experiments take data
- Automate tape mounting

Procure flexible media robot, make drive choice late

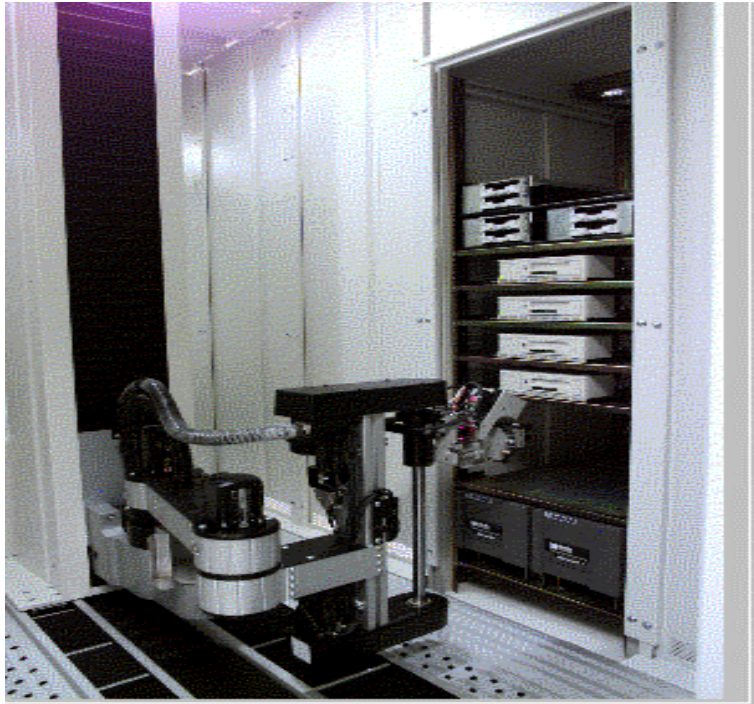


- Experiments procured ADIC (nee EMASS) AML/2 robots.
- Shelved out for 8mm and some DLT.
- Approx 35,000 tape slots capacity

Serial Media Evaluation

- Lifetime testing of drives
 - new tape each day
 - run for months
 - watch for degraded capacity
- ATL tests
 - many mounts
 - space, skip, write, read
- Early production use

The ATL experience



- Mount testing
 - DLT 7000
 - Sony AIT-1
 - Exabyte Mammoth-1
- Production experience
 - Mammoth-1
 - Few 10**6 mounts
 - 13 drives
- Less exp w/ others

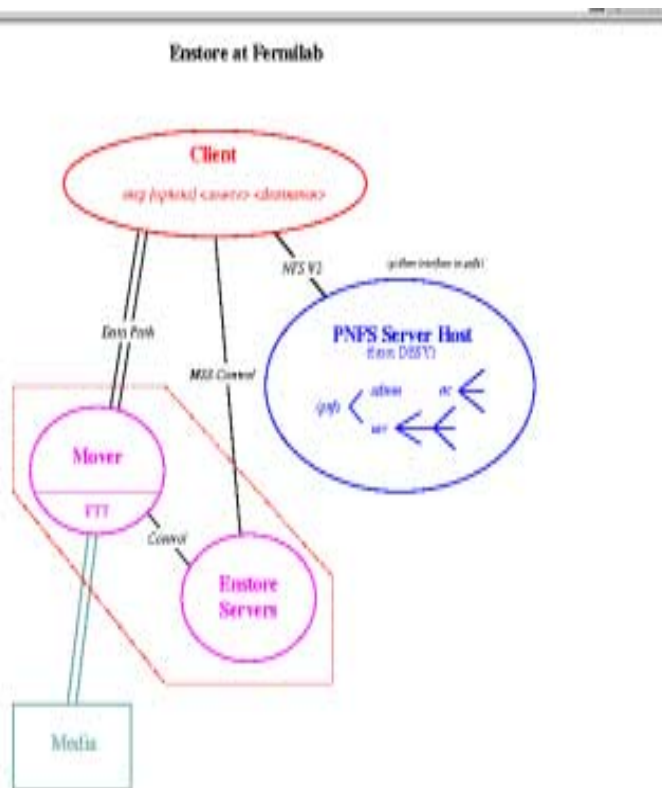
Status of selection

- Interest in both the second generation AIT and Mammoth drives.
- Experiments software will work with either.
- Evaluations continue, preparing to choose.
- Expect:
 - Well under \$2.00 /GB
 - Affordable to expand tape plant

Higher level system issues

- D0 Experiment: sustain 150 MB/sec to/from tape 24x7
- Attach a pool, tape drives via standard, 1500 byte MTU Ethernet LAN
 - To farms of Linux PC's (10 MB/sec)
 - To large SGI boxes (100 MB/sec)
 - To TRU64 boxes (30 MB/sec)
 - And to whatever else needs to be provided.

Enstore software for distributed data movement



- NFS transport for name space
- Scalable “data movers”
- Provide the reliability of an ensemble.
- Provides volume and location hints
- Data transport via

Enstore – hardware system



- UNIX
 - This deployment is Linux
- Intel based computers.
 - Very low cost computer system
 - System is portable
- In production

After Fermilab Run II

- Some large future experiments in the field will be truly global.
- Data storage and movement will be international in scope.
- The Large Hadron Collider (LHC) experiments will be based at CERN, near Geneva, Switzerland.
- Fermilab will be the U.S. “Tier 1” Data Center for one such experiment, “Compact Muon Solenoid”

Global Experiments

- Global Distribution means that storage systems will need more features.
- Effort to add these features/understand how they work in applications are part of “The Grid”
- Types of features
 - Authentication/Authorization (privacy matter less to us)
 - Wide are network transfers.
 - File or object catalogs.
 - Description of capabilities. (meta computing)
- Multi national effort

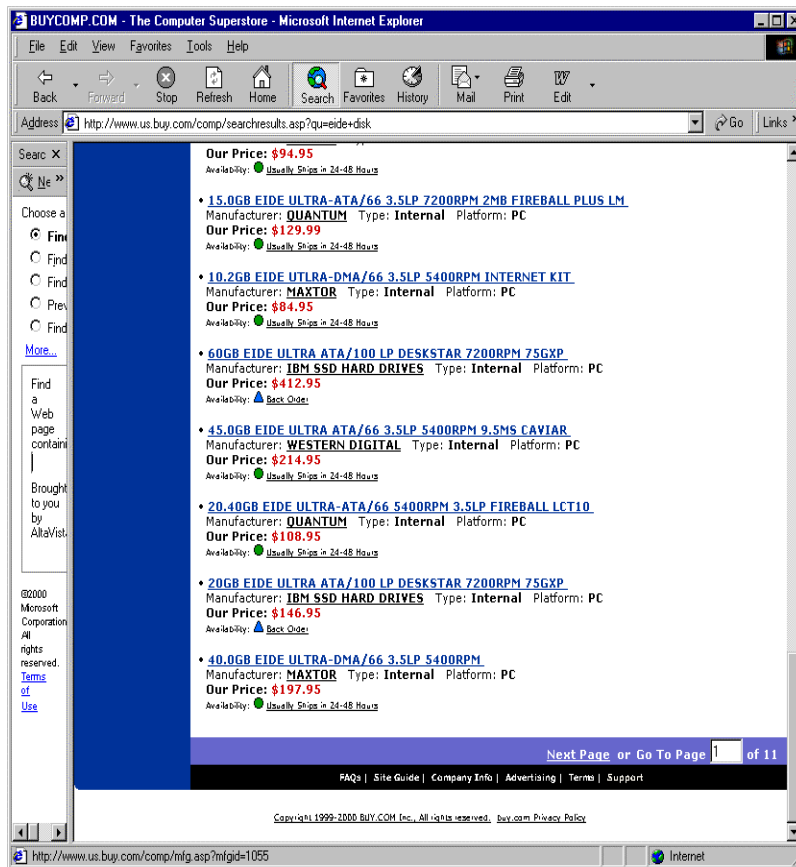
BTeV at Fermilab

- BTeV is a large-data-volume proposed future experiment at Fermilab
- Its draft budget notes the large amount of storage which will come for “free”
 - Projects a few thousand commodity P.C.’s will required.
 - Notes the increases in disk size.
 - Its budget recognizes that the data volume is budget limited.

Future (2005) as seen by LHC

- CERN has issues “PASTA” report.
- Worries about the market for tape based storage.
- Specifically worries that the order-of-magnitude price advantage of tape will not be sustained.
- LHC era data will be replicated (since it is global), so less reliance on archival properties of tape.
- The area is a worry.

Whither tape?



- Easy to find disk at \$5/GB or less @ quantity one.
- Some in the field are paying \$3/GB for tape.
- Have integrated 24 disks/PC.
- Tremendous fixed, inflexible, investment in ATL's people, software.

IDE investigations

- The field has integrated large amounts of IDE disks into one P.C.
- “farms” of these have vast capacity and immense throughput.
- Clearly useable in storage systems today!
- The field knows how to write software, integrate systems for its own storage systems.
- We will investigate adding archival quality features to these cheap disk systems.

Summary

- HEP is constrained by storage costs.
- The Field is interested in novel solutions
 - MUST be cost effective
 - Some latitude for loss and low data rates
- World-wide data distribution framework will be constructed for next generation experiments.
- Capable and experienced people are working on systems.