Storage Elements at BNL

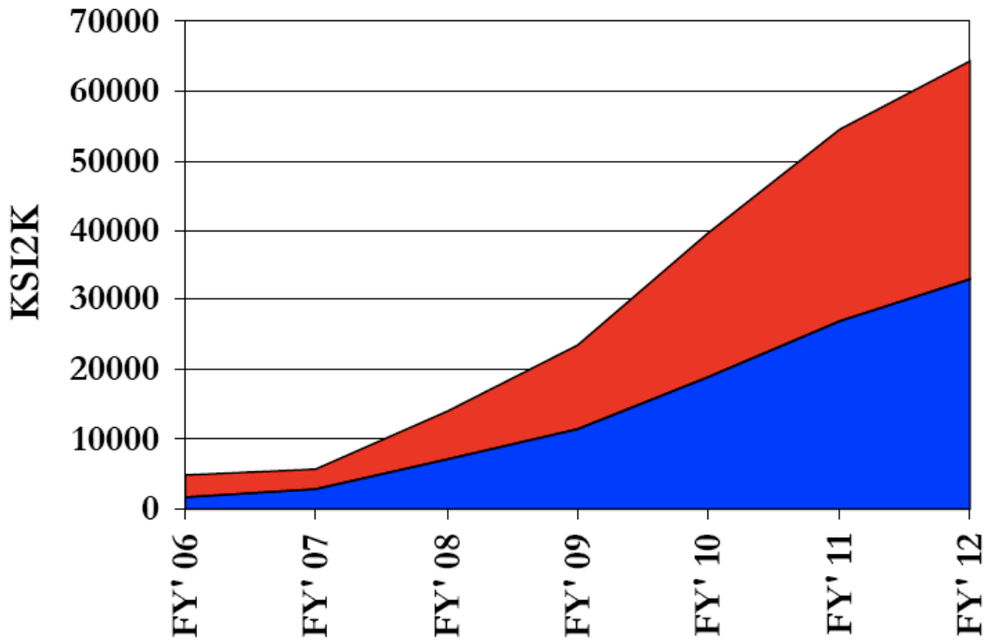
2008 HEPiX @ CERN | Robert Petkus | Brookhaven National Laboratory

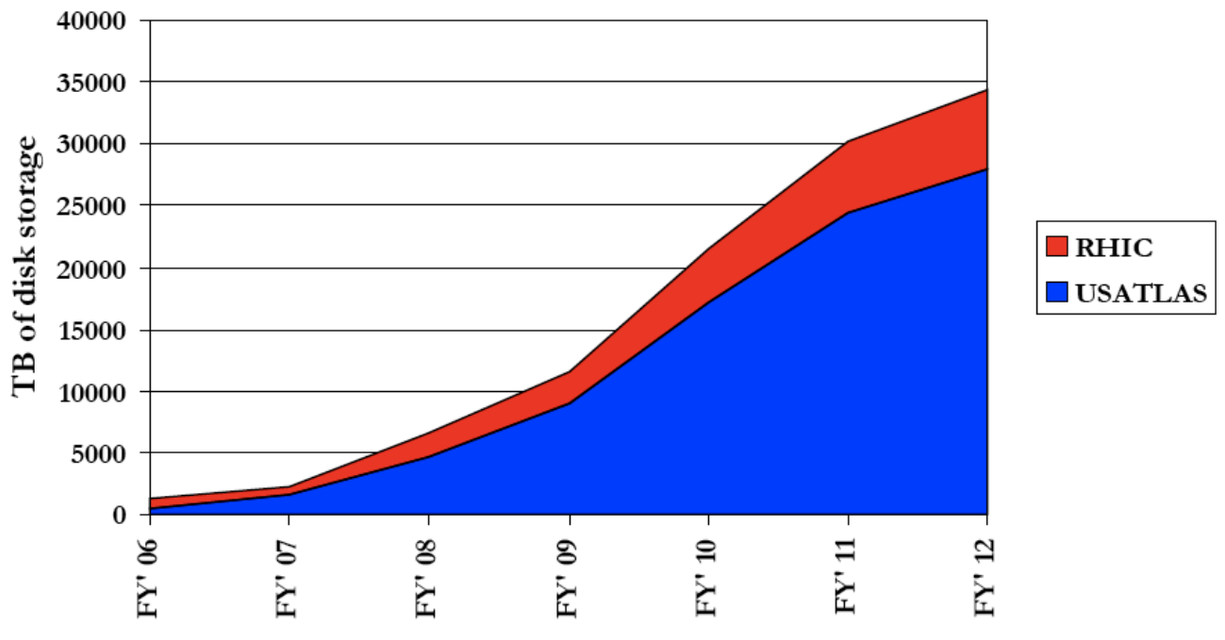
BlueArc

- (2) Node Titan 2200 cluster and stand-alone 2200 in use at RACF to meet all NFS central storage needs
- 222TB LSI fibre channel disks, RAID 5
- SATA looks like a likely incremental addition of storage but
  - Require RAID 6 with 1TB disks
  - Unclear which vendors offer the best SATA RAID 6 solution
  - Lower spindle count, 2/3 performance of FC
- Upgrade plans vacillate between (2) configurations
  - Add additional 2200 head to complete 2nd cluster
  - Create single large 3200 series cluster
  - Both configurations benefit from total 10GE upgrade
- Investigating DataMigrator for tiered storage classes
- Compelling company future roadmap

Distributed Storage @ BNL

- >2PB disk storage to be added in 2008, >4PB in 2009, etc.
- Storage demands for ATLAS outpace CPU requirements
  - Not the case for RHIC





Distributed Storage @ BNL

- Limited power, cooling, and space require the adoption of high-density disk arrays
  - Superior choices today have disks vertically mounted
  - Other designs imminent: side-by-side, back-to-back in rack (IBM)
  - We like to stripe across as many disks as possible
- SunFire x4500s seem to be a favorite now in the HEP community
  - Specs and performance characteristics are well documented
  - Delay of AMD Barcelona a real disappointment
- Other well-positioned competitors exist and will always be considered and re-evaluated as improvements are made
  - Forthcoming DDN S2A StorageScaler with (60) drives in 4U
  - 3PAR S800, (40) drives in 4U (10x4 disk magazines)
- Every major vendor will have jumped on the bandwagon within 2 years

SunFire X4500 + Solaris

- Solaris 10 is used, primarily for the benefit of ZFS and all its cool features, again well-documented
  - End-to-end checksumming, COW, high performance RAID60, no fsck
- Blazing fast ZFS fuse on Linux seems a ways off
- Will probably revisit Linux on X4500 in the future again but no real compelling reason for us in our environment
- Solaris and X4500 wish list (in no particular order)
  - IO module to daisy chain systems
    - Export ZFS pool if CPU module dies
  - Faster processors
    - Nehalem -> QuickPath
  - A real SNMP that works with fmadm
  - OS on compact flash
  - Clean-up SMF
  - Automated disk failure handling

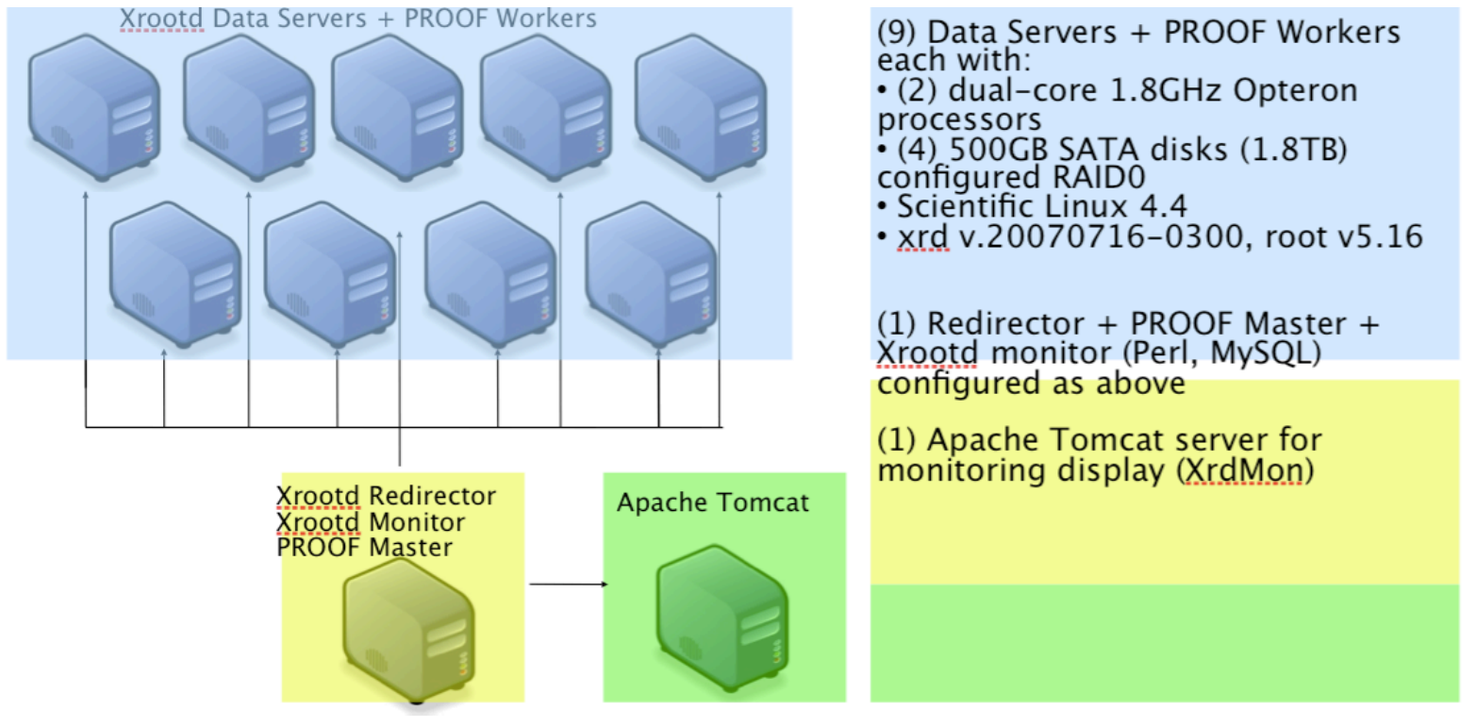
PROOF-Xrootd Testbed

Currently there is a (10) system Xrootd-PROOF testbed at BNL

- PROOF = Parallel ROOT Facility, also a plug-in for Xrootd
  - Developed as system for batch analysis of huge sets of ROOT data files on a cluster
  - PROOF uses Xrootd for data discovery and file serving

Main issue: match I/O supply and demand

- (1) I/O bound ROOT job = (1) core = ~10MB/sec
- (1) SATA HDD = ~20MB/sec = (2) jobs
- (8) core system + (1) HDD = bad



COURTESY SERGEY PANITKIN

PROOF-XrootD w/SSDs

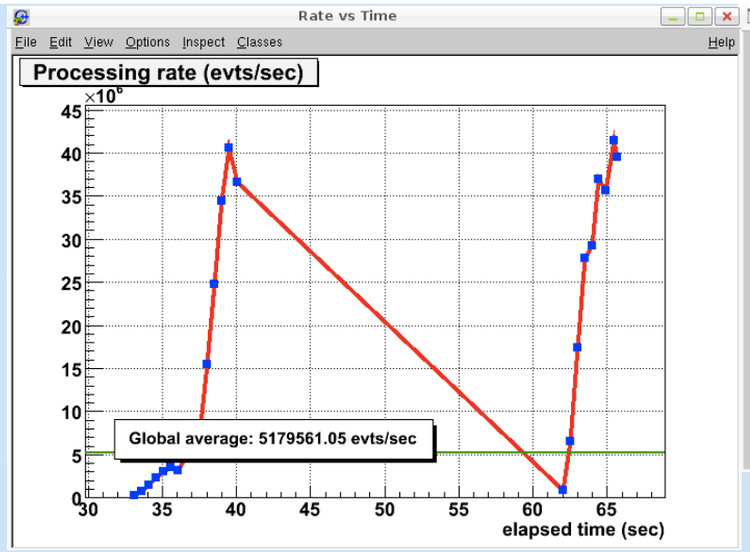
- Purchased (10) Mtron 3.5" SATAII SSDs, (1) per testbed system
  - 64GB, 120MB/sec read, 90MB/sec write sustained performance
  - Random access time = 0.1 ms (SATA HDD = ~10ms)
  - Write endurance >140 years @ 50GB/day
  - MTBF = 1 million hours
  - 7-bit error correction code

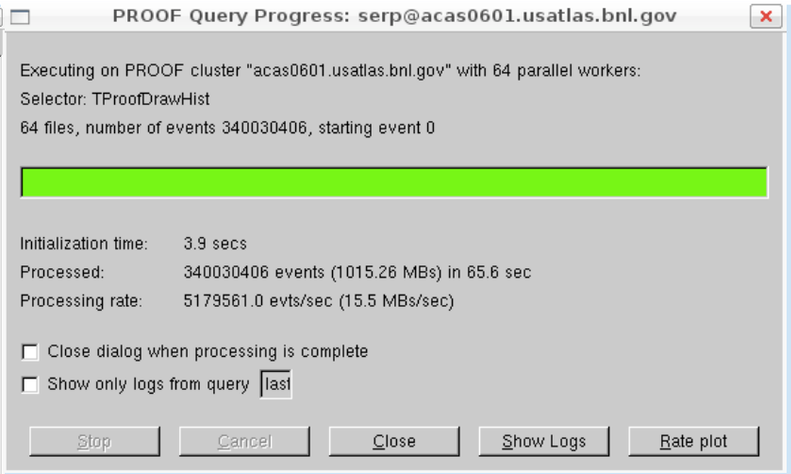
Access Type	IOPS operation unit size	
	512B	4KB
Sequential Read	78,000	11,200
Sequential Write	42,000	16,700
Random Read	18,000	12,000
Random Write	120	120

PROOF-Xrootd w/SSDs

```
root [2] .L make_tdset.C
root [3] TDSet *d = make_tdset("/data/test",1)
root [4] d->Draw("fTemperature","1")
Looking up for exact location of files: OK (64 files)
Validating files: OK (64 files)
Mst-0: grand total: sent 2 objects, size: 1032 bytes
<TCanvas::MakeDefCanvas>: created default TCanvas with name c1
(Long64_t)340030406
root [5]
```

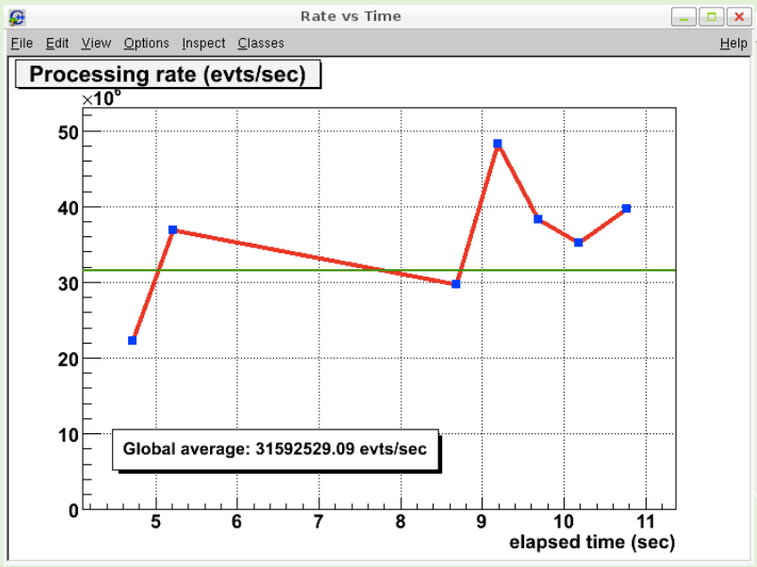
Test with SATA HDD

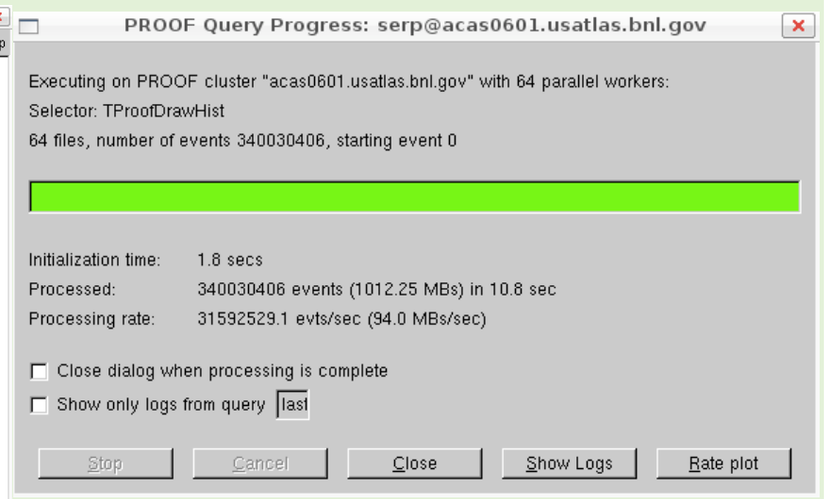




Test with Mtron SSD

```
root [3] .L make_tdset.C
root [4] TDSet *d = make_tdset("/ssd/test",1)
root [5] d->Draw("fTemperature","1")
Looking up for exact location of files: OK (64 files)
Validating files: OK (64 files)
Mst-0: grand total: sent 2 objects, size: 1032 bytes
<TCanvas::MakeDefCanvas>: created default TCanvas with name c1
(Long64_t)340030406
root [6]
```





TESTS COURTESY OF SERGEY PANITKIN

Future Directions

- Much larger and cheaper SSD presence to enter market accelerated by consumer demand (iPod, Smartphones, GPS devices)
- Fusion-io ioDRIVE (40-320GB)
  - ioMemory silicon-based storage architecture
  - 600MB/sec random writes, 700MB/sec random reads
  - 120k sustained random IOPs
  - <50 microseconds access time
  - Single / mulitple PCI-e x4
  - Wear leveling (24x7) writes = 8years (160 & 320GB models)
  - Strong ECC algorithms

Questions?