Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

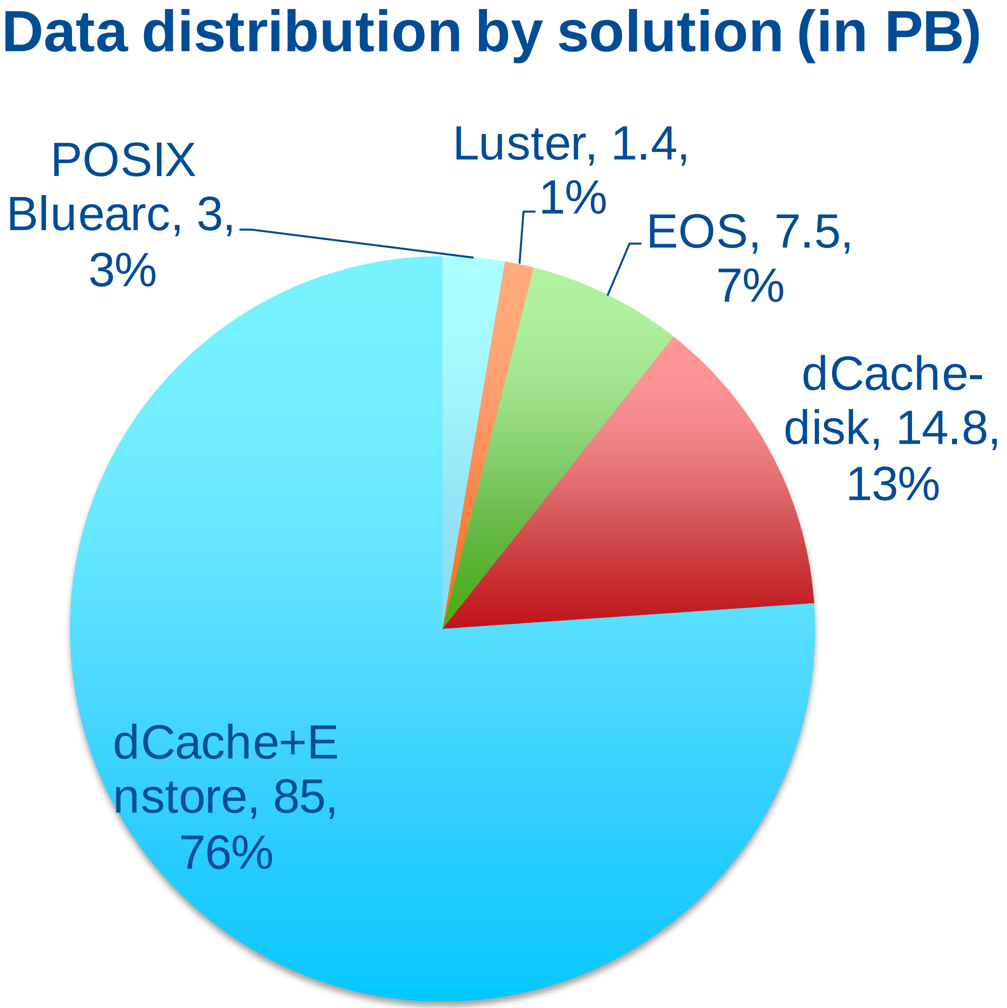
Scientific Storage at FNAL

Gerard Bernabeu Altayo
Dmitry Litvintsev
Gene Oleynik

14/10/2015

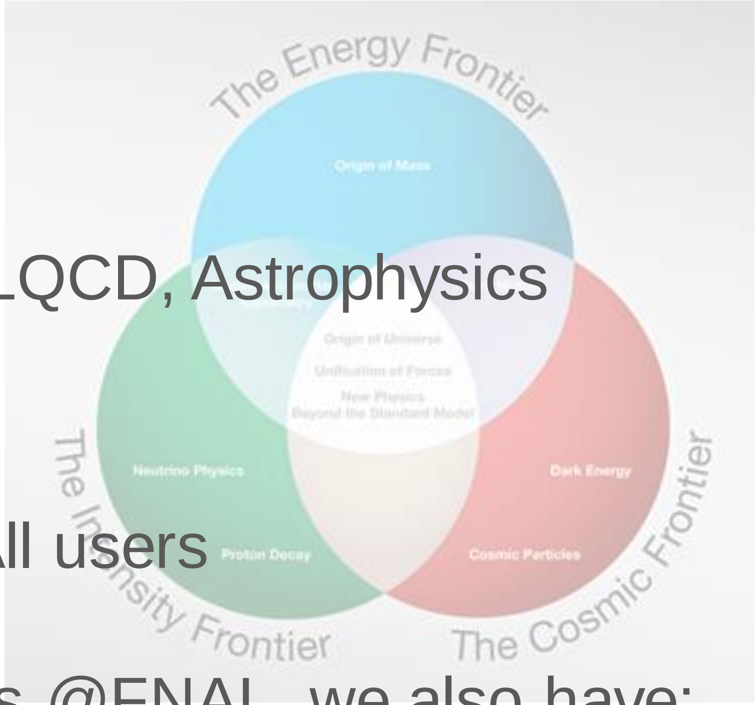
Index

- Storage use cases
- Bluearc
- Lustre
- EOS
- dCache disk only
- dCache+Enstore



Storage use cases

- Data intensive - Intensity Frontier (FIFE) & HEP experiments
  - DAQ long term storage
  - Offline computing (batch jobs)
- High performance computing - LQCD, Astrophysics
  - Low latency
  - Highly parallel
- Interactive full POSIX access - All users
- Not covering all storage solutions @FNAL, we also have:
  - DB storage (on Bluearc)
  - CVMFS (dedicated SAN)
  - Windows shares, etc.



Bluearc - Hitachi NAS Platform

Industry standard high performance NFS. This is the most standard storage system available for Scientific Storage.

- Use case: interactive and still some 'legacy' batch storage for both CMS (direct analysis users) and FIFE experiments. The biggest advantage is that this is a fully POSIX compliant solution.
- Storage size: 3PB total
- Theoretical performance: 2*10Gbps Throughput and 200K IOPs (on cached data)
- Access protocols: NFSv3, NFSv4 , CIFS/SMB, SRM/gridftp
- Users: ~ 1000 users

Solution outlook:

- Current solution has support until 2018.
- FIFE experiments’ data is being migrated to dCache.
  - This is crucial for efficient offsite usage, currently there are scalability issues.
- Users’ home areas will remain in Bluearc
- Some FIFE and CMS data will remain in a fully POSIX compliant solution.
  - Mounted from the Interactive Machines, where users compile and run test jobs locally.
  - Not mounted on the local WorkerNodes.

Lustre

Industry standard Distributed Filesystem. Used for High Performance and low latency access by local HPC computing resources.

- Use case: accessible from all the Fermilab USQCD clusters over QDR (Quad Data Rate, 40Gbps) Infiniband. One of the USQCD clusters is housed in a separate computer room as the Lustre file-system hardware and these clients access the Lustre storage using a pair of routers configured as Lustre routers. The Fermilab Theoretical Astrophysics Group has its own small instance (~130TB).
- Storage size: 1.4PB, ~156M files
- Theoretical performance: Observed rates to ~0.75GB/s for writes and ~2.75GB/s reads.
- Access protocols: POSIX-like Lustre mount (via Infiniband connection) and GridFTP via Globus Online for off-site transfers.
- Users: mounted on ~1200 clients, used by ~150 users

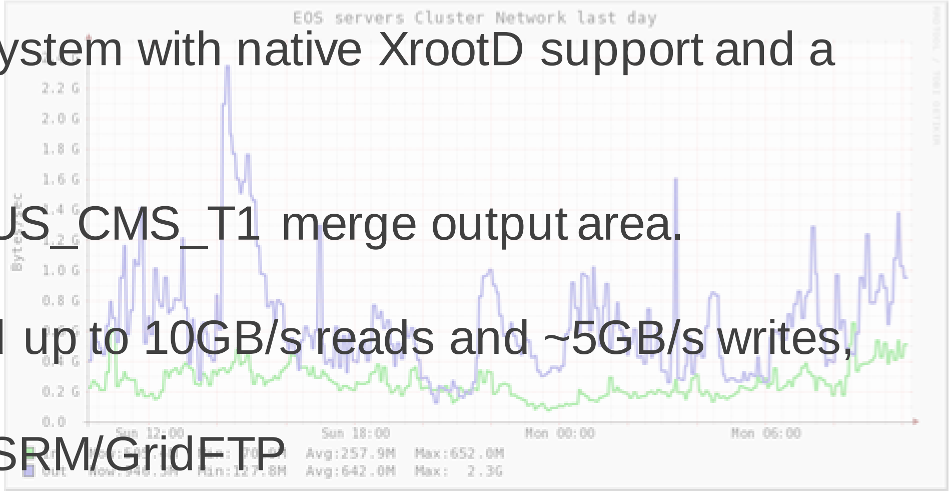
Solution outlook:

- In the process of upgrading to a current and supported Lustre version (1.8.9 -> 2.5.4).
- Lustre on top of ZFS on SL6, JBOD lower storage cost.
- Will keep running 2 separate Lustre instances.
- Support until 2019, future plans depend on LQCD project funding and support by the open source (OpenSFS) community.

EOS

CERN developed disk-based storage system with native XrootD support and a FUSE POSIX like interface.

- Use case: CMS LPC storage and US_CMS_T1 merge output area.
- Storage size: 7.5PB, 19.5M files
- Theoretical performance: observed up to 10GB/s reads and ~5GB/s writes, not hitting HW limits.
- Access protocols: FUSE, XrootD, SRM/GridFTP
- Users: ~1000

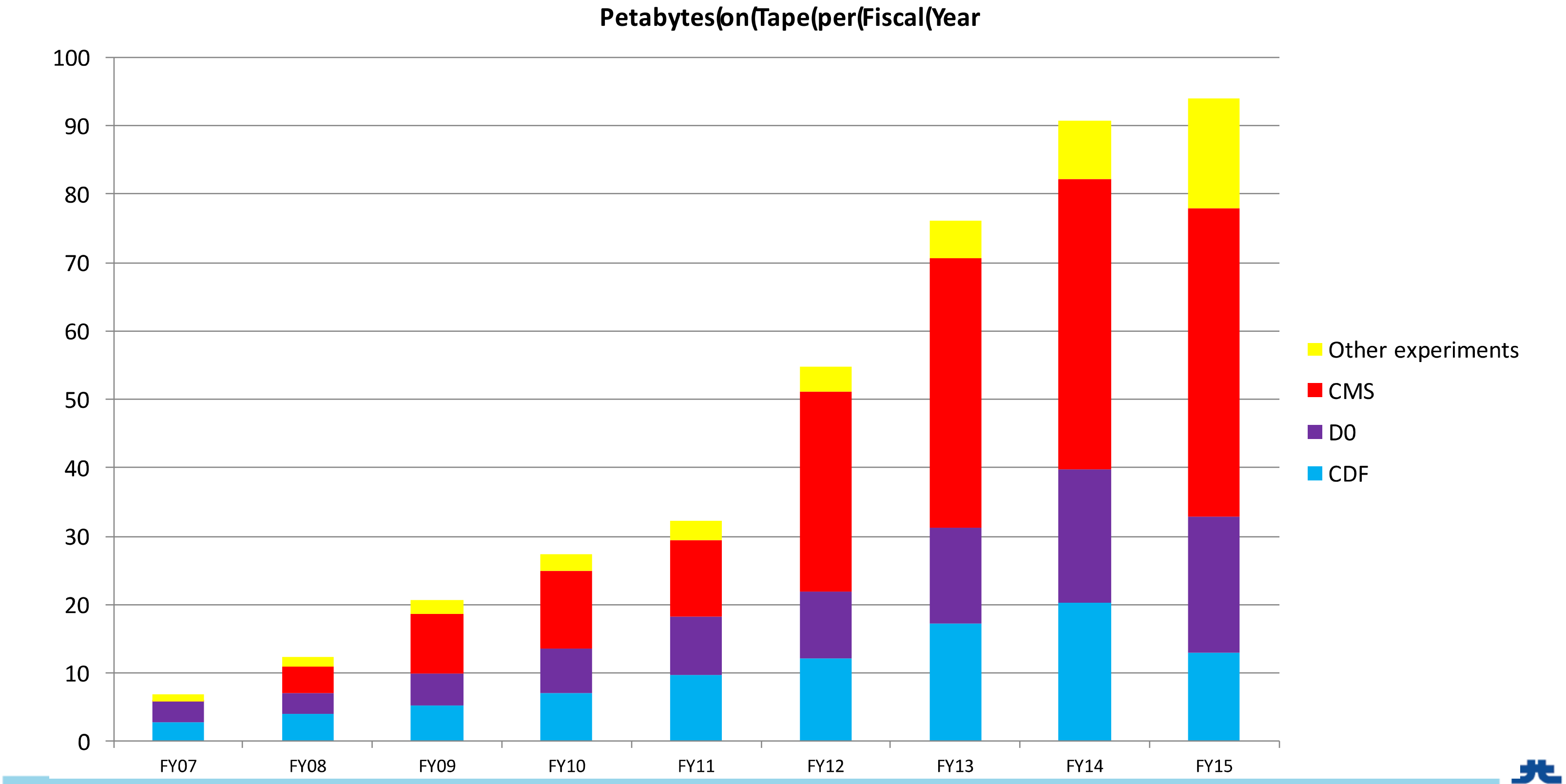


Solution outlook:

- Move most of the FNAL-LPC activities to EOS (out of Bluearc)
    - Interactive via FUSE and batch via xrootd (not FUSE mounted on WN)
- Move all Tier1 operations (xrootd) out of EOS
- Trial multiple copies on non-RAID JBOD (vs single copy on RAID protected disks)

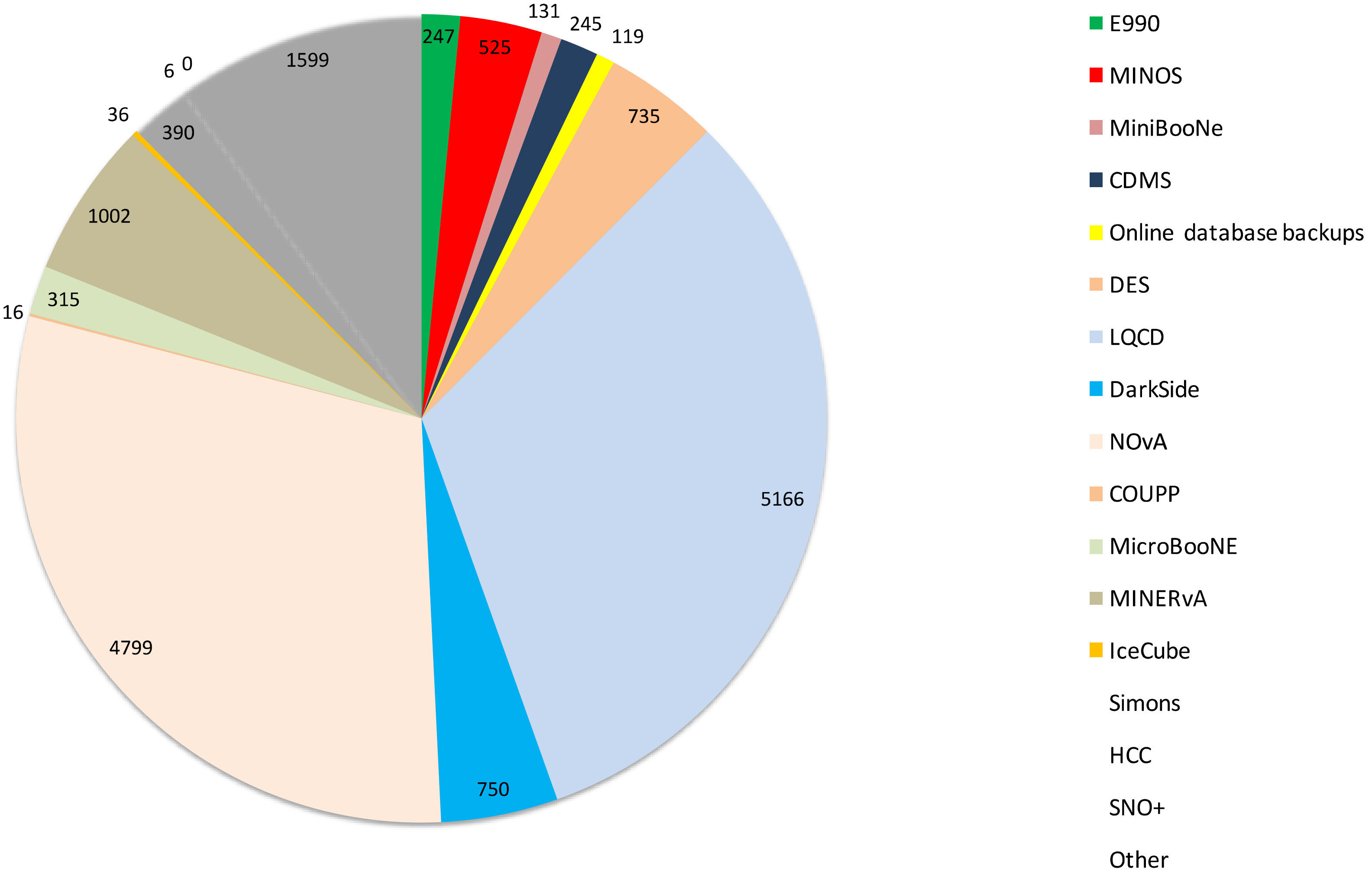
dCache + Enstore

- All Scientific Data at Fermilab are stored on tape managed by Enstore HSM
- Fermilab uses dCache as :
  - HSM front-end (3 instances)
  - Stand-alone disk-only system (1 instance)



dCache + Enstore

Detailed ‘Other experiments’ distribution in TB



CMS dCache disk only

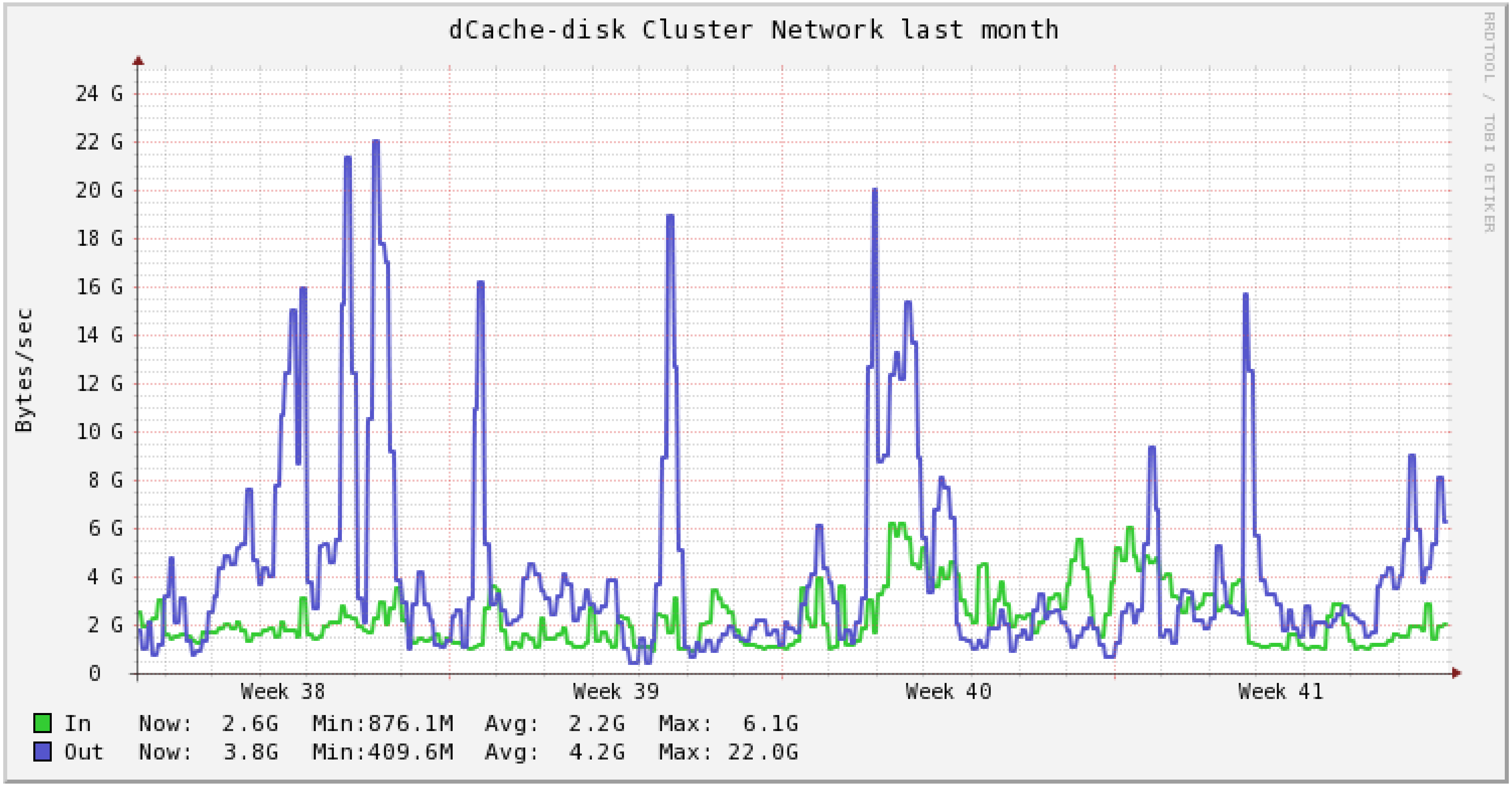
To completely separate disk and tape workflows CMS created a separate T1 dCache disk instance (published as its own Storage Element) which is directly accessed by all the batch jobs. This is the most performant SE at FNAL with ~200 dCache pools.

- Use case: Data is staged in/out by CMS data ops and CMS production jobs
- Storage size: 12.6 PB, 3.1M files
- Performance: topping off at ~20GB/s reads, up to 8GB/s writes observed without hitting HW limits.
- Access protocols: XrootD, SRM/gridftp, dcap, NFSv4
- Users: CMS T1 production and user analysis (AAA)

Solution outlook:

- Upgrade to a modern supported version (2.2 -> 2.13)
- Move merge jobs output from EOS to dCache (reducing dependencies for the jobs)
- Concentrate doors (which need certificates like gridftp) in a few systems
- Move Chimera DB to SSD disk
- Open up storage building block architecture (now server + FC SAN)
  - (please provide feedback on how you do your RFPs)
- Consolidate service management group

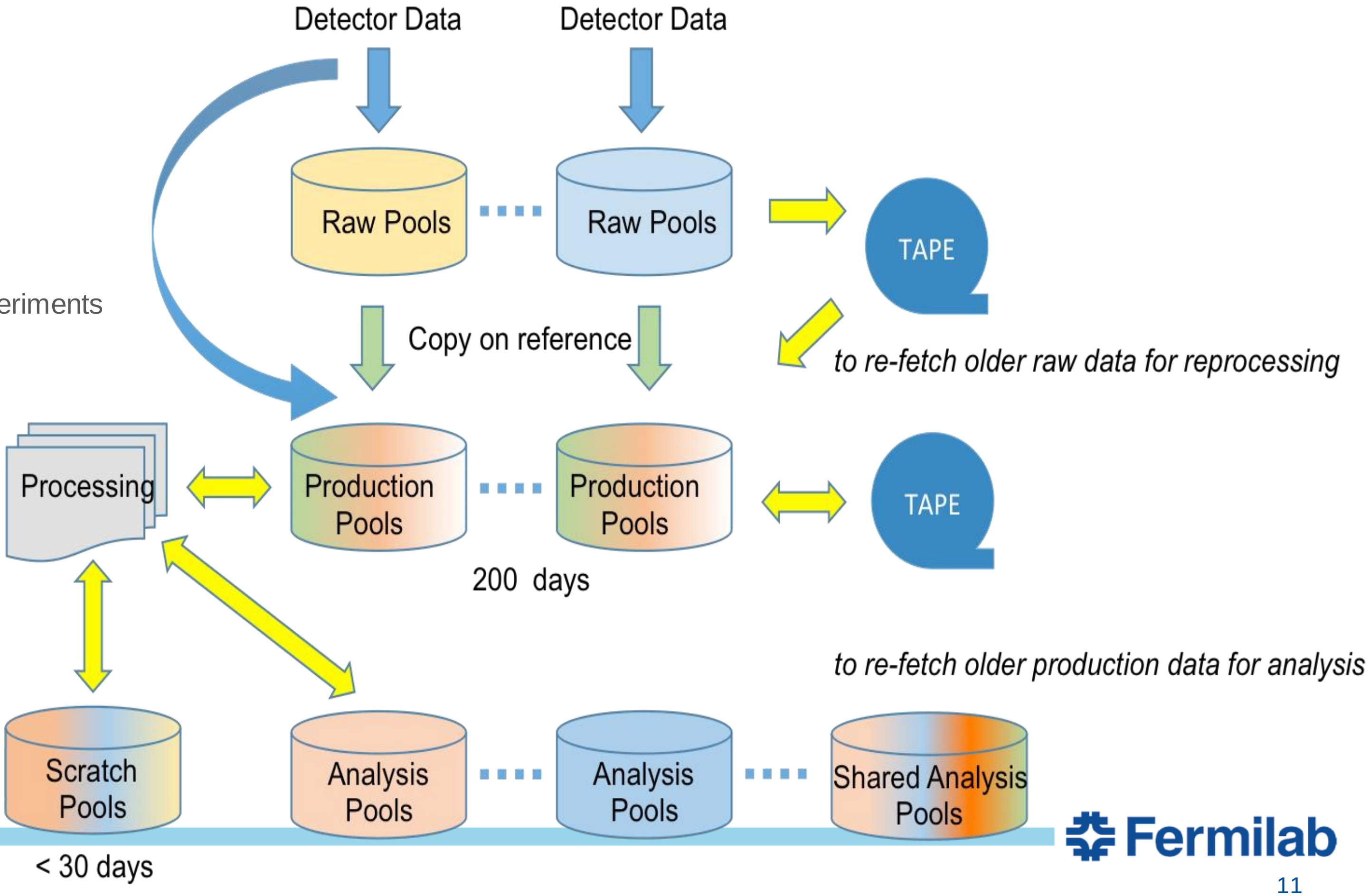
CMS dCache disk only



FIFE dCache + Enstore

- Tape backed cache – for production workflow
- Non tape backed cache – for production temporary scratch
- Persistent - disk only – for analysis workflow

Colors represent experiments.
Rainbow = shared between experiments



dCache + Enstore

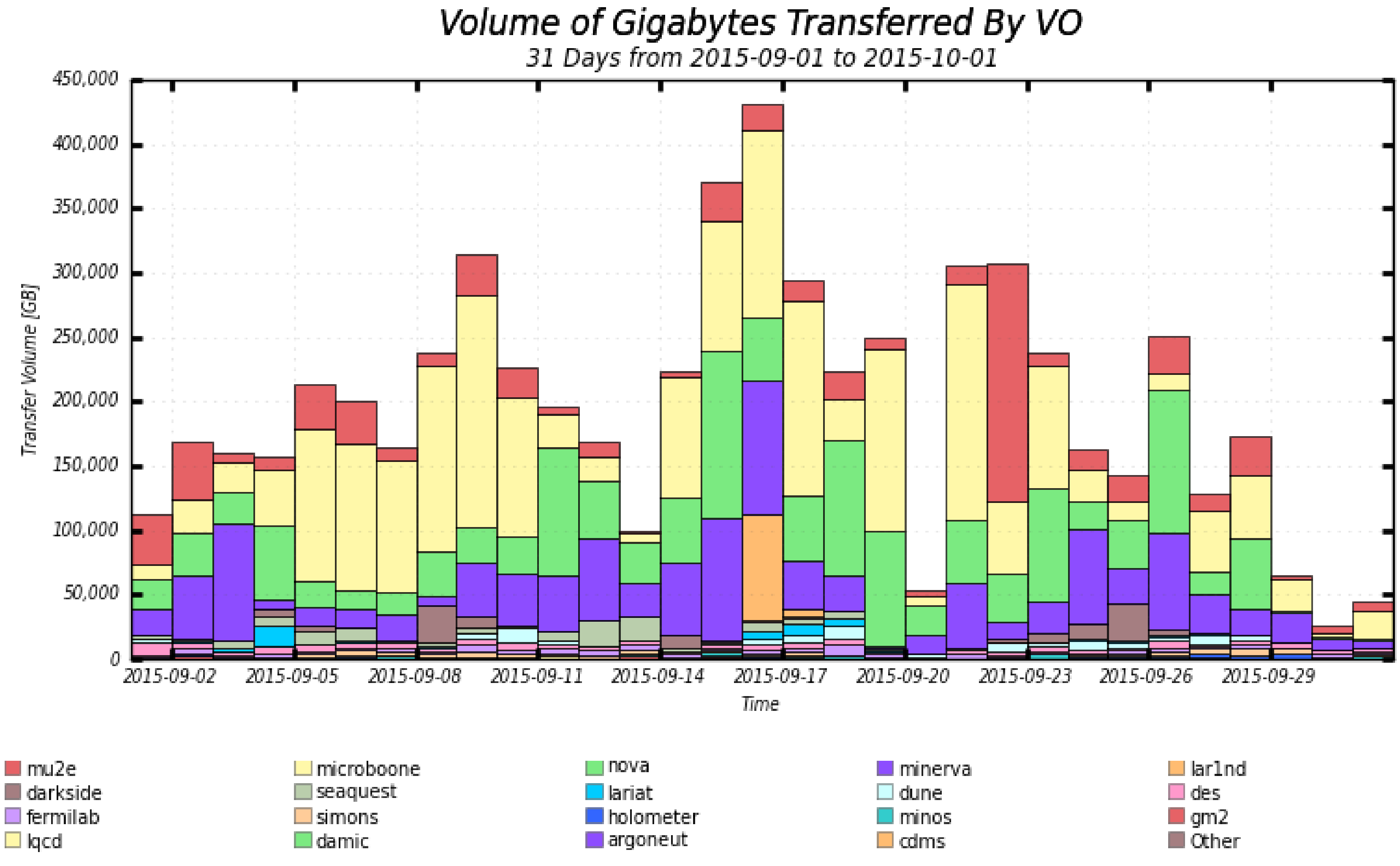
- Storage size: 4PiB of disk cache and 85 PiB tape space.
  - FIFE: 3 PiB disk cache, 16 PB data on tape, 100M files
    2.2 PiB disk only
  - CMS: 1 PiB disk cache, 22 PiB data on tape, 17.5M files
  - CDF: 1.4 PiB disk cache, 10 PiB data on tape, 8.9M files
  - D0: 100 TiB disk cache, 9.1 PiB data on tape, 15M files

- Access protocols: dcap, SRM/gridftp, WebDAV, XrootD, NFSv4.1 (and v3,4 for metadata only access)

Solution outlook:

- CMS: Upgrade all pools to 10GE on new HW, keep buffer at ~1PB (will have less spindles)
- CMS: Optimize data transfer rate to/from Tape
- CMS: move Chimera DB to SSD disk
- Consolidate service management group
- FIFE : keep scaling out disk capacity

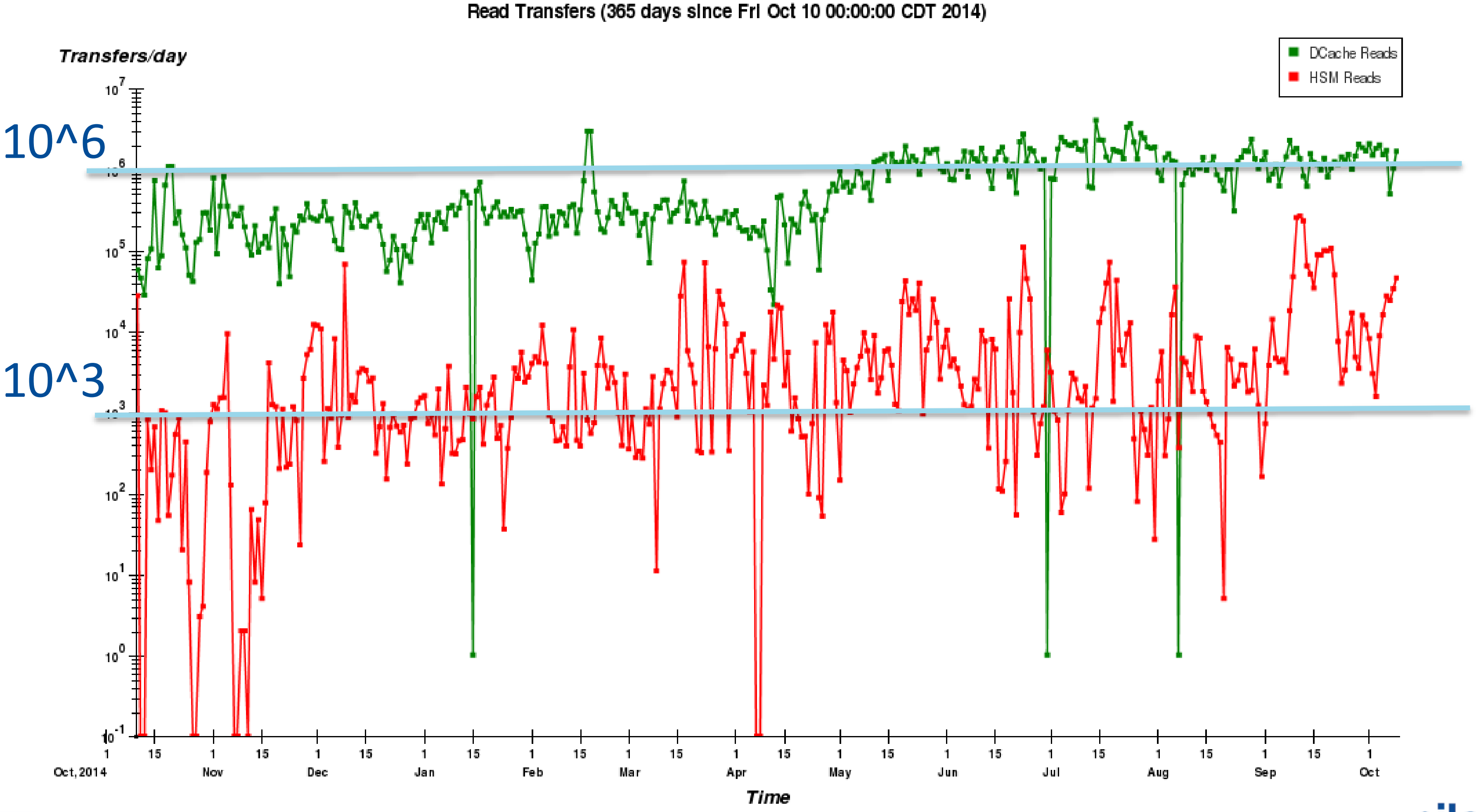
FIFE dCache



Maximum: 431,086 GB, Minimum: 25,781 GB, Average: 197,093 GB, Current: 44,602 GB

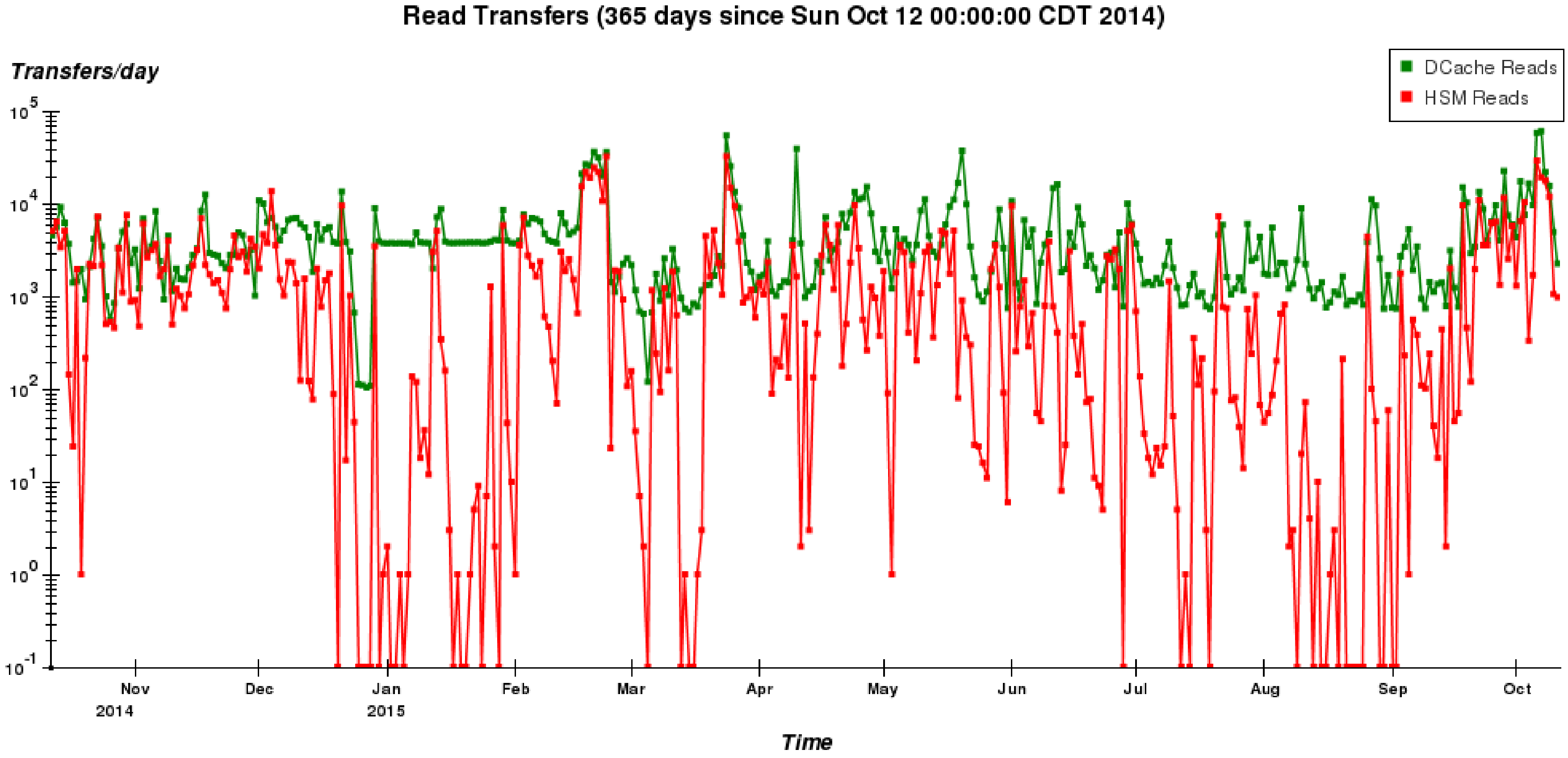
FIFE dCache

- For FIFE experiments we see >95% disk cache hit ratio



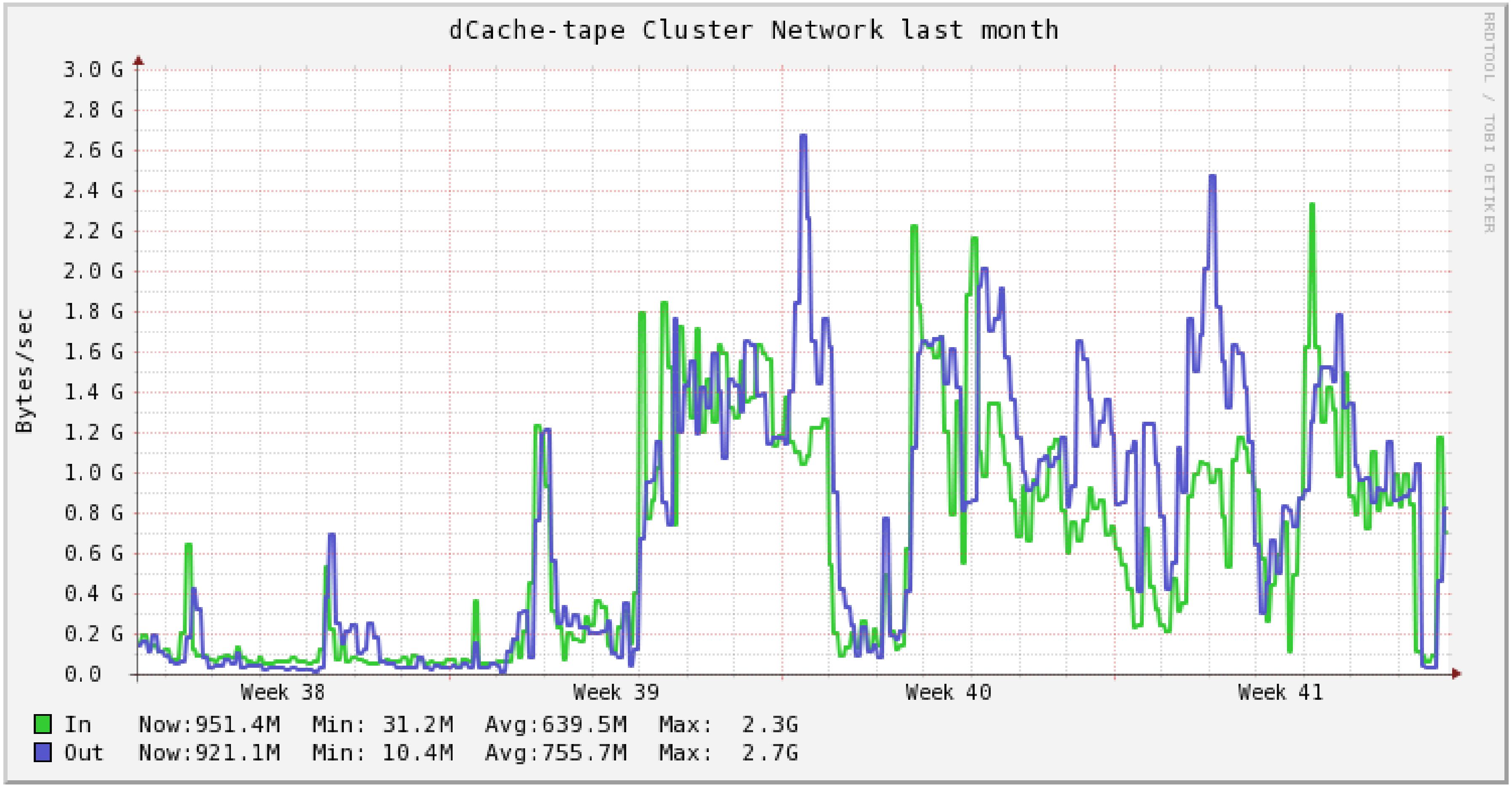
CMS Tape dCache

- For CMS tape the hit ratio is low. It is just a staging area and actual cache hits happen on the 'disk instance'.



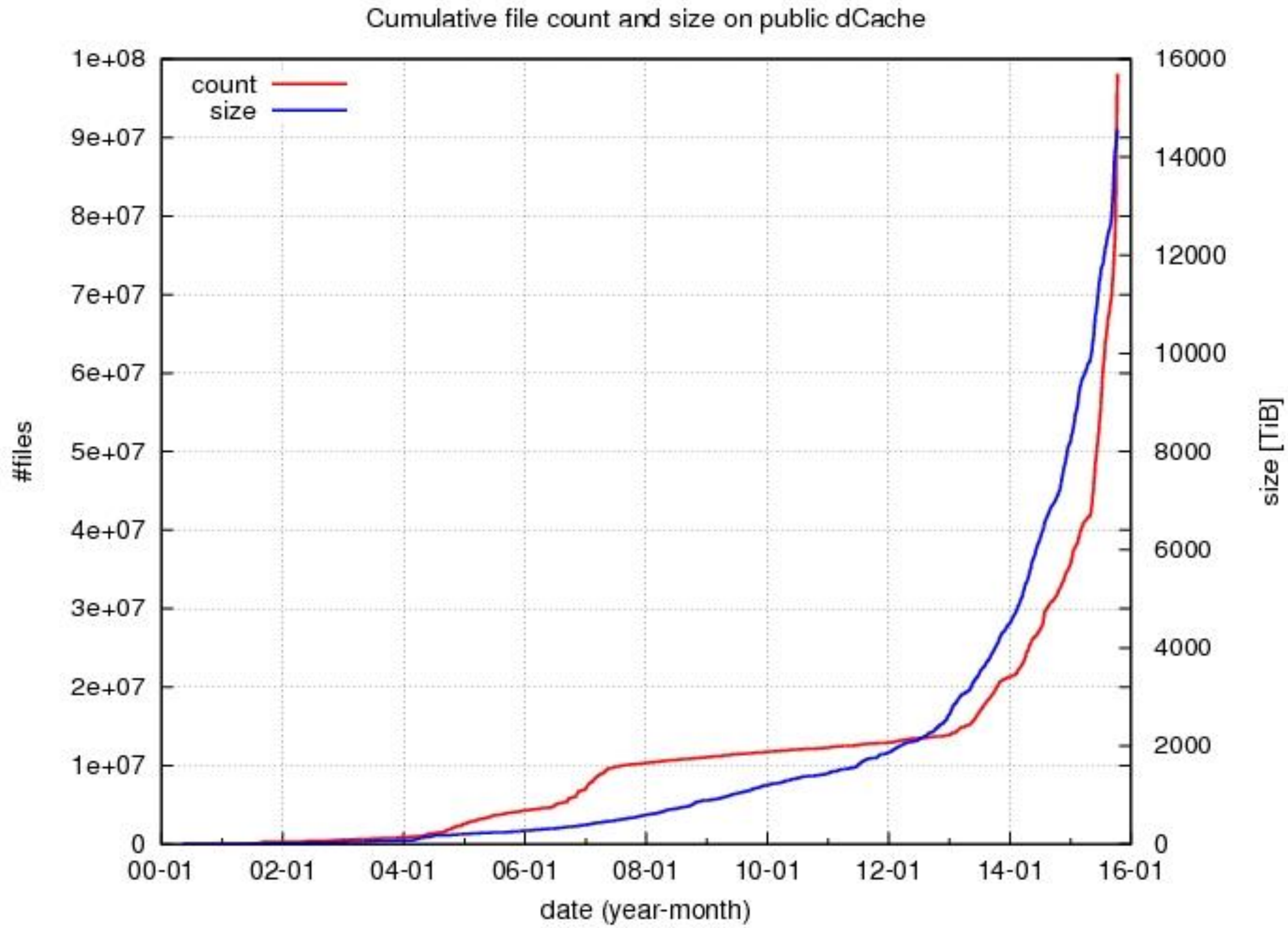
CMS Tape dCache

- For CMS tape the hit ratio is low. It is just a staging area and actual cache hits happen on the 'disk instance'.



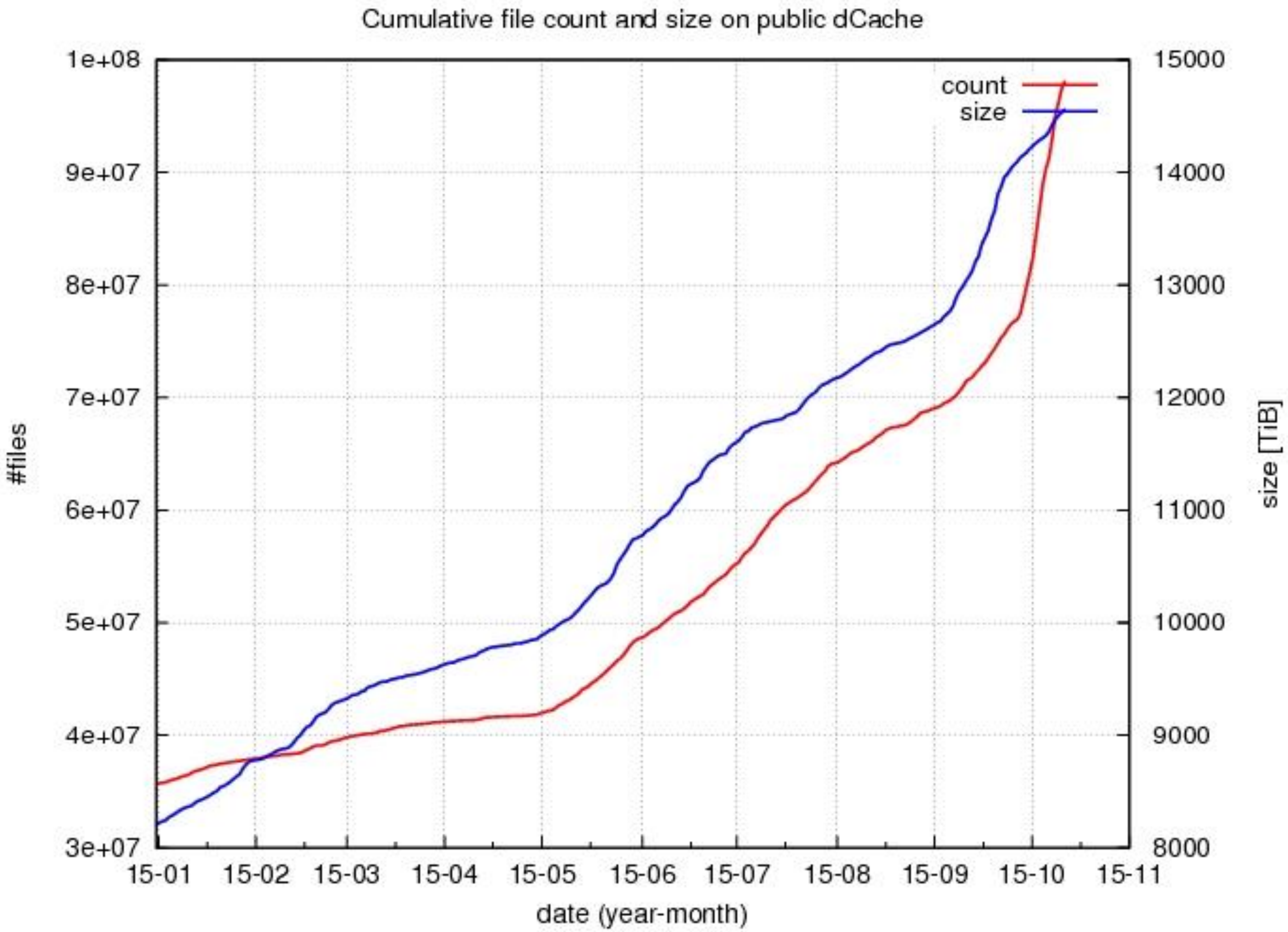
FIFE dCache files

- Explosion of data over the last few years



FIFE dCache files

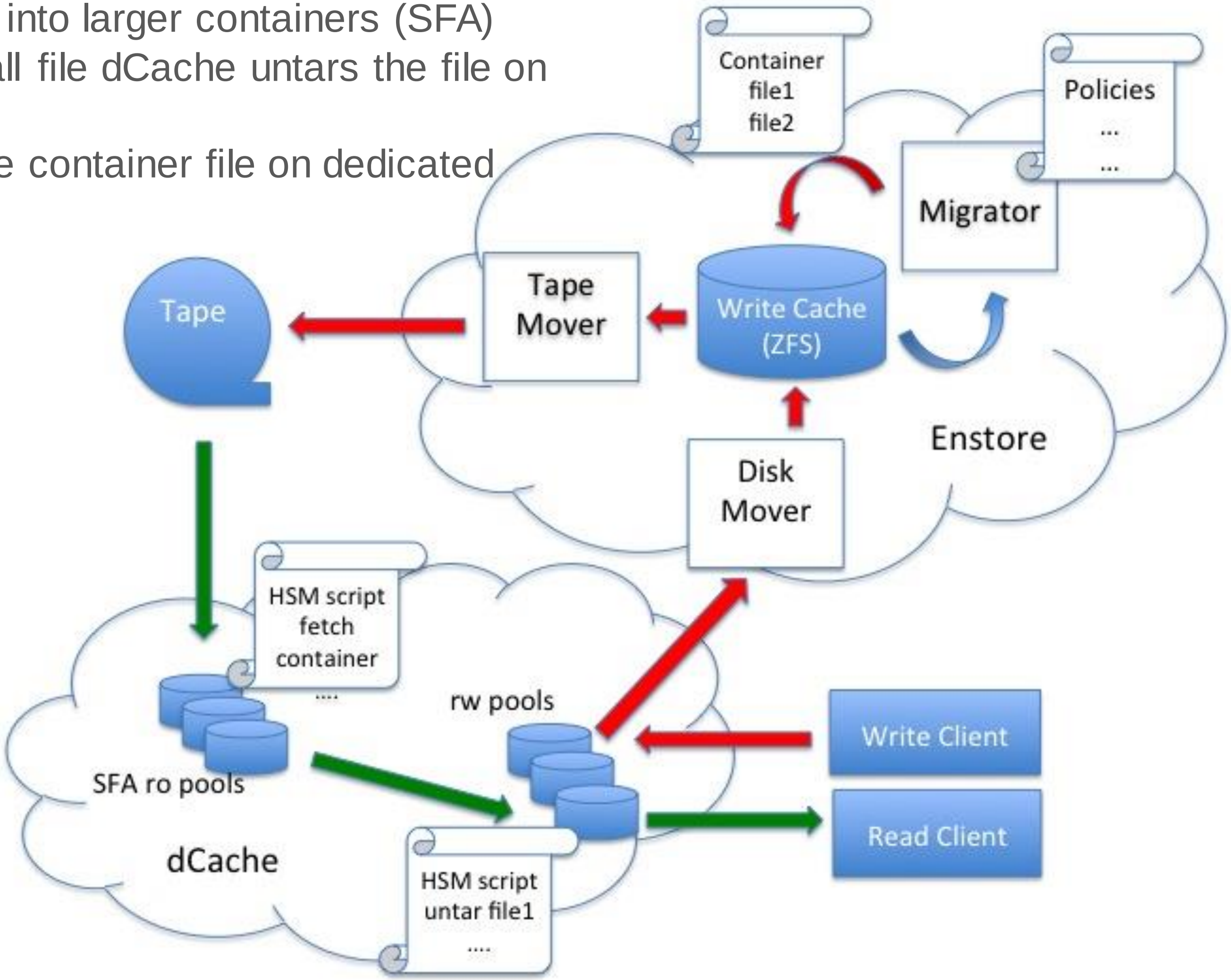
- Doubled the number of files in 6 months (40->80M)



Handling of Small Files

Enstore Small File Aggregator - Scalable handling of Small files:

- Enstore packs small files into larger containers (SFA)
- On request to read a small file dCache untars the file on a dCache read pool
- Untar triggers stage of the container file on dedicated set of dCache pools



Summary

- Consolidating Scientific Data Storage at FNAL across experiments
- FIFE and HEP experiments’ storage needs keep growing and challenging us
- Distributed Storage on top of HSM provides a cost effective and performant solution that fits our needs.