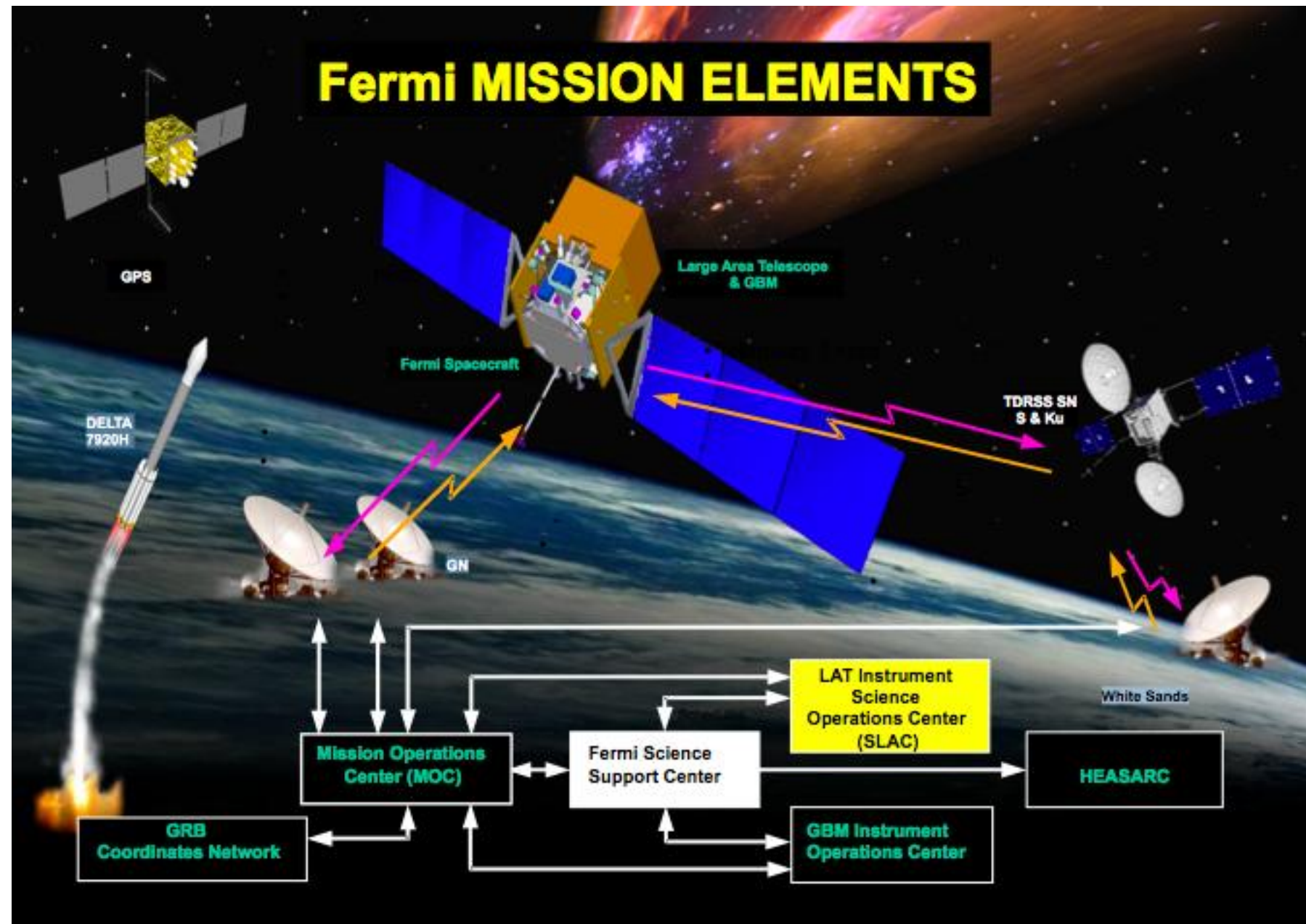


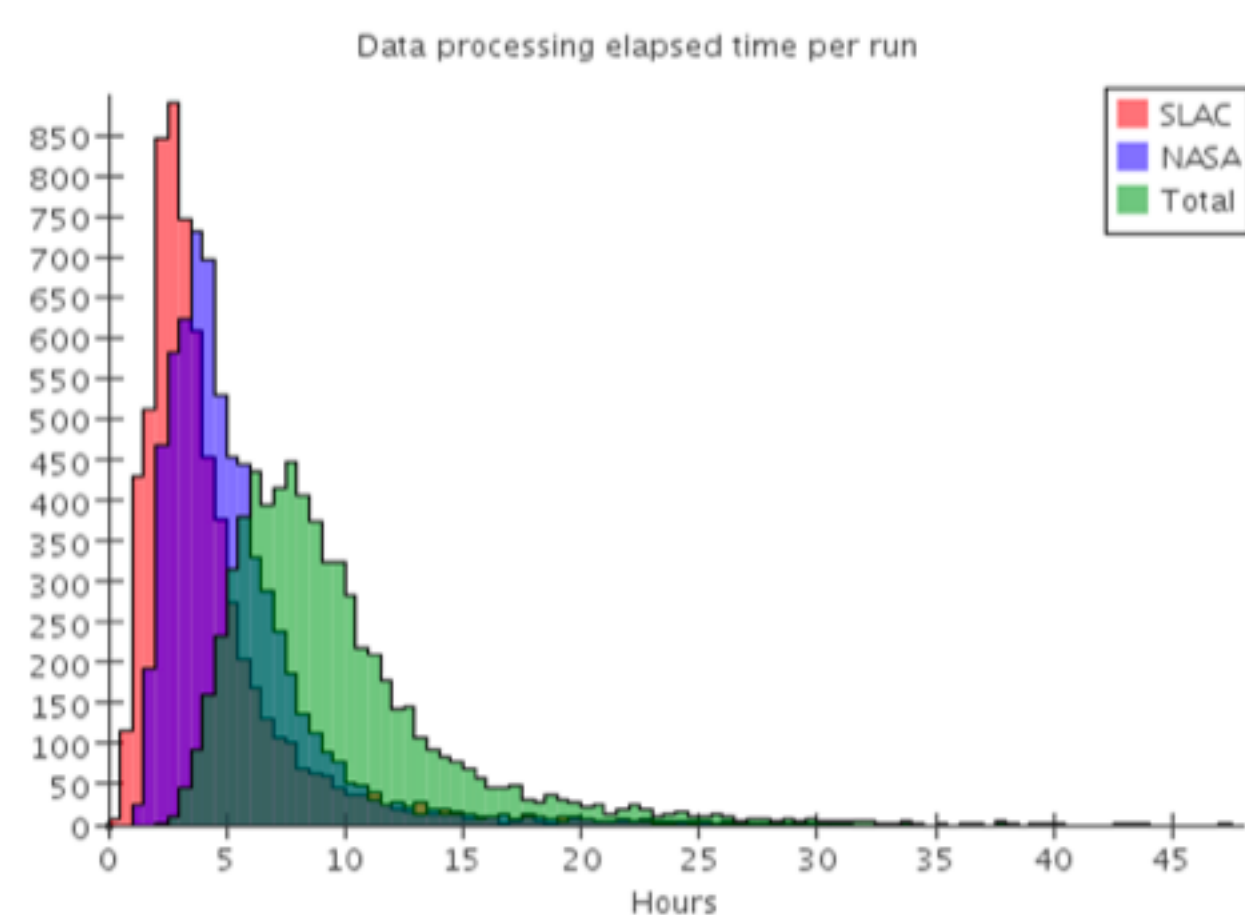
# Two Years Before the Mast: Fermi LAT Computing Two Years After Launch - Eight to Go!

Richard Dubois (SLAC)



## Data Flow from Space to Public

- ~13 GB/day downlinked via TDRSS to White Sands ~8 x/day
- Transferred to Mission Ops at Goddard to send to instrument teams
- Science data returned to Goddard Science within about 3 hours of receipt. Most probable time for public to get data is 8 hrs.



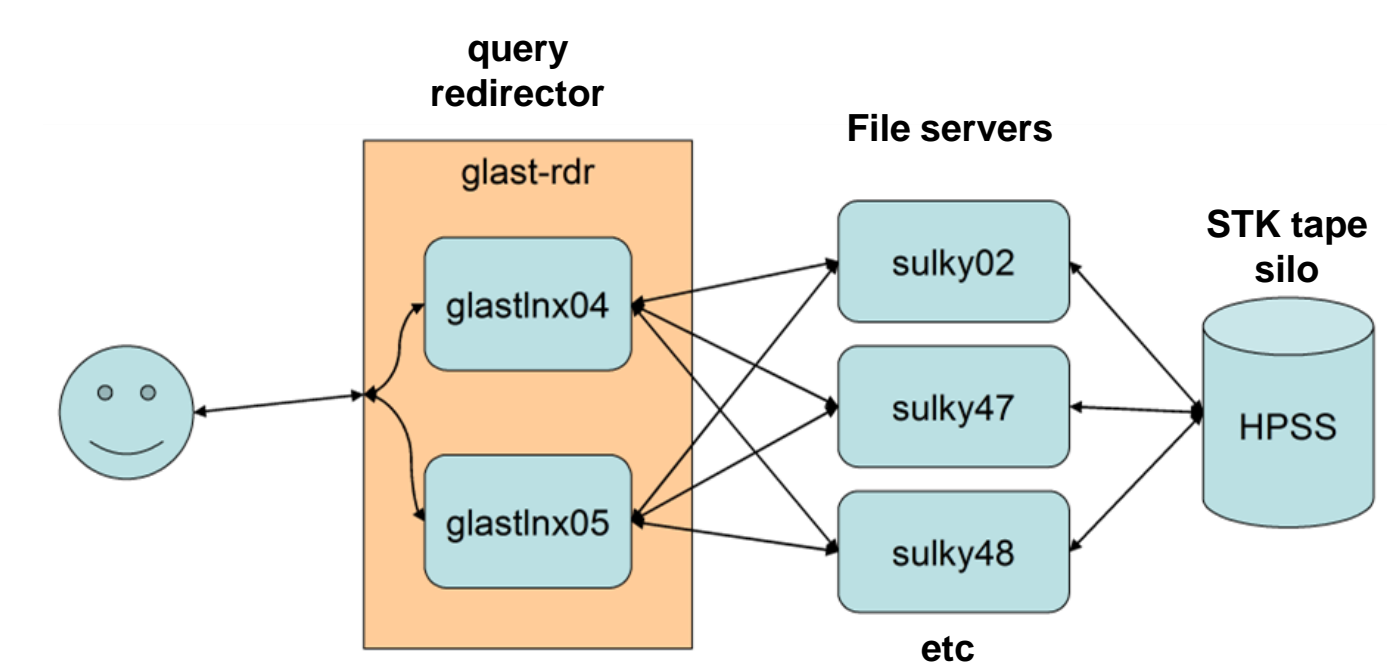
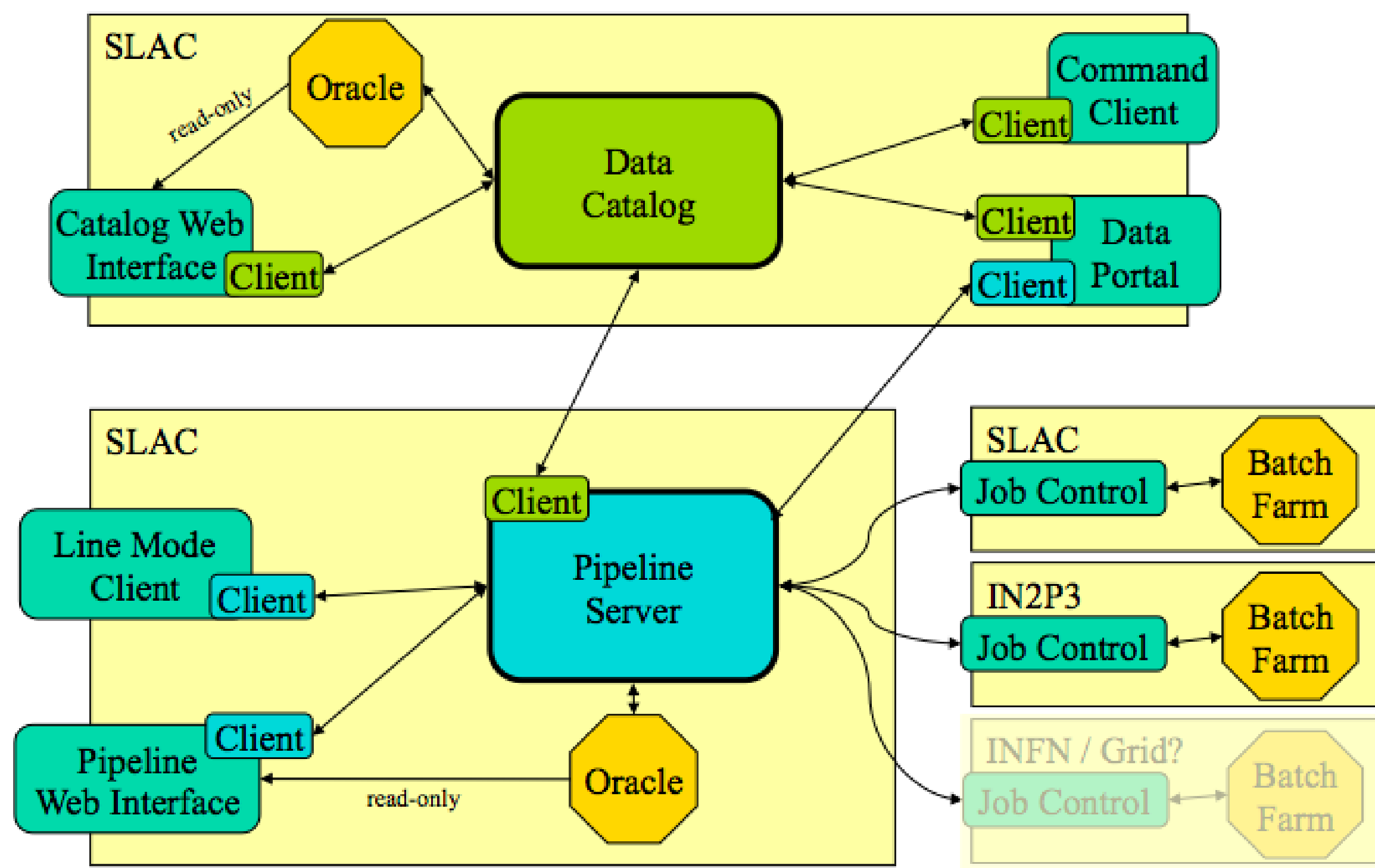
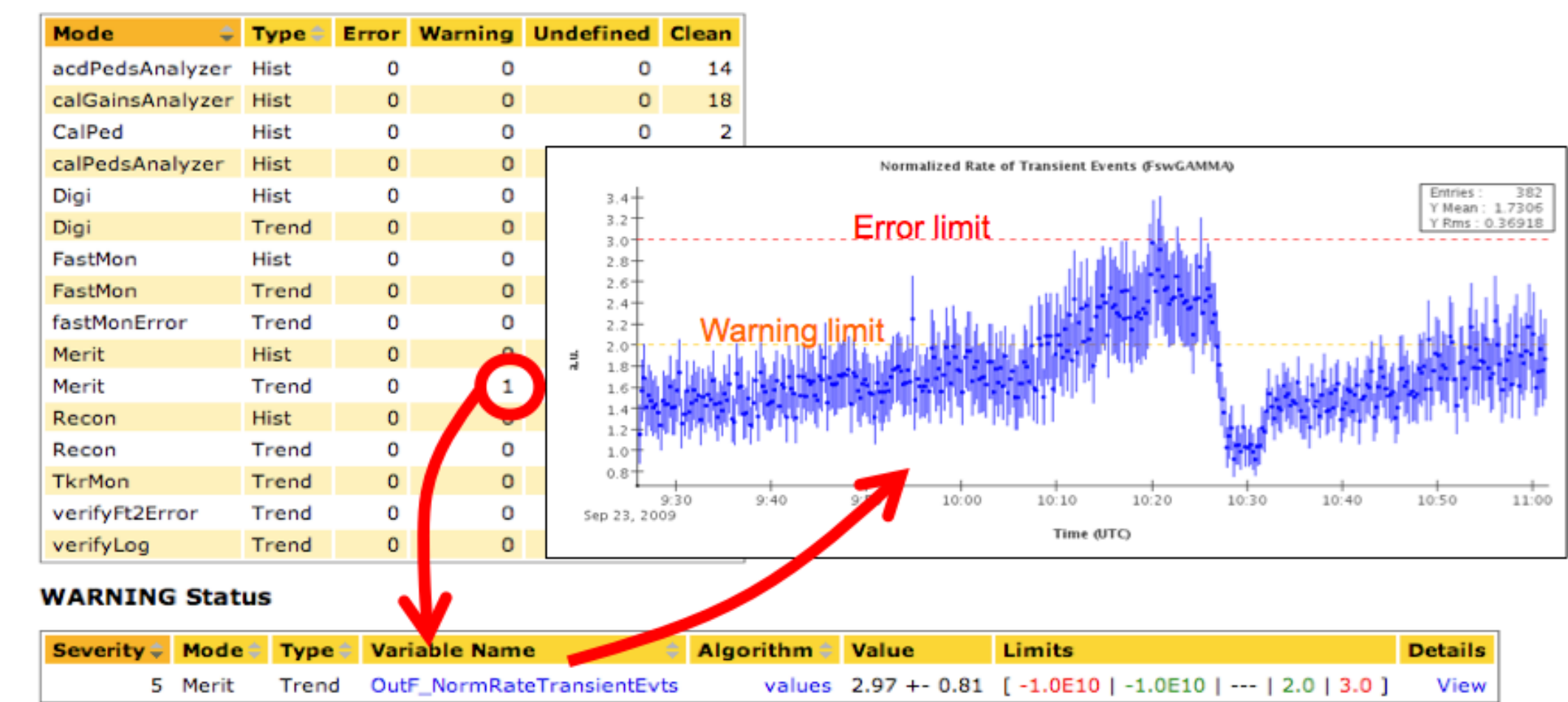
## Abstract

The Fermi Observatory was launched on June 11, 2008 and the Large Area Telescope (LAT) was activated on June 25. Some 13 GB of data is downlinked daily, transformed into 500 GB in the event reconstruction process, spread out over approximately 8 contacts per day. Each data run is farmed out to several hundred computing cores and results merged back together in our processing pipeline. The pipeline is designed to execute complex processing trees defined in xml and to handle multiple tasks simultaneously, including prompt data processing, simulations and data reprocessings. Our system has a pair of Oracle servers at its core to maintain all the state and dataset bookkeeping. Batch processing is centrally dispatched to the SLAC LSF and Lyon (France) BQS batch farms with more than 5000 shared cores. The xrootd cluster filesystem is used for high throughput and management of large disk pools. Nagios and Ganglia are used for problem alerts and tracking resource usage. The HEP-like instrument event reconstruction lives in a Root world, while high level science is done in FITS. LAT Collaboration users have access to the data via web query engines that slice and dice the data to their needs, also executing the queries in the processing pipeline. The data is now public, so we have the issues of new development vs stability for an outside user base. Two years later, we are dealing with the issues of long term support - how to keep a complex operation alive and vital for 10 years, and how to deal with dependencies on external packages whose support is out of our control.

## Keeping Watch

- LAT is very stable!
- Small team checks monitoring output via web
- Automated alarm system to warn if any of ~2000 distributions change outside limits

## Alarms for run 275390766



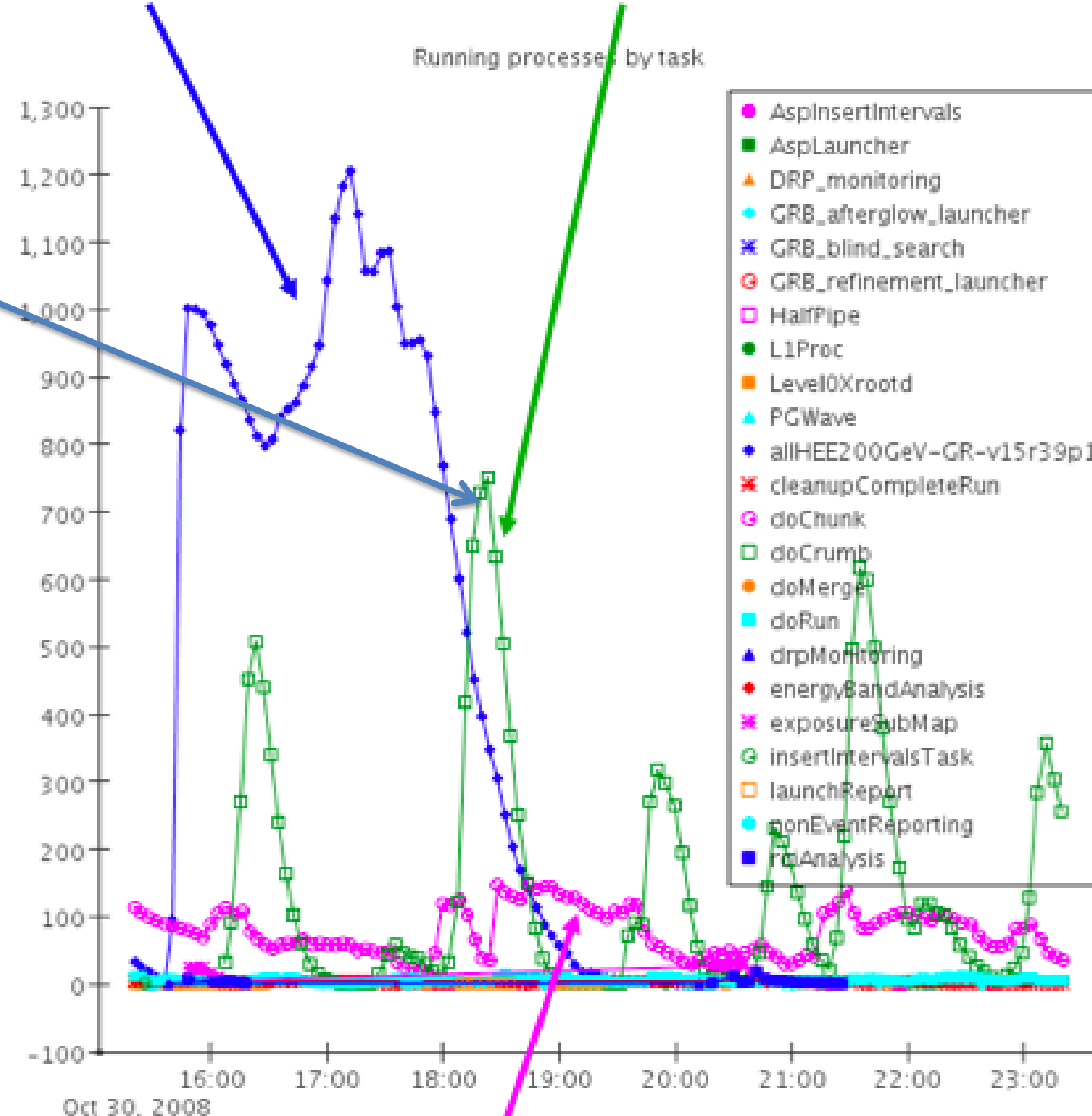
## xroot: Distributed Filesystem/Hierarchical Storage

- Maximizes throughput
- Minimizes manual disk management
- Automates archiving datasets to (and restoring from) tape
- Provides more reliability and scalability than NFS
- Remote access via proxies
- Supports access control based on Fermi collaborator list

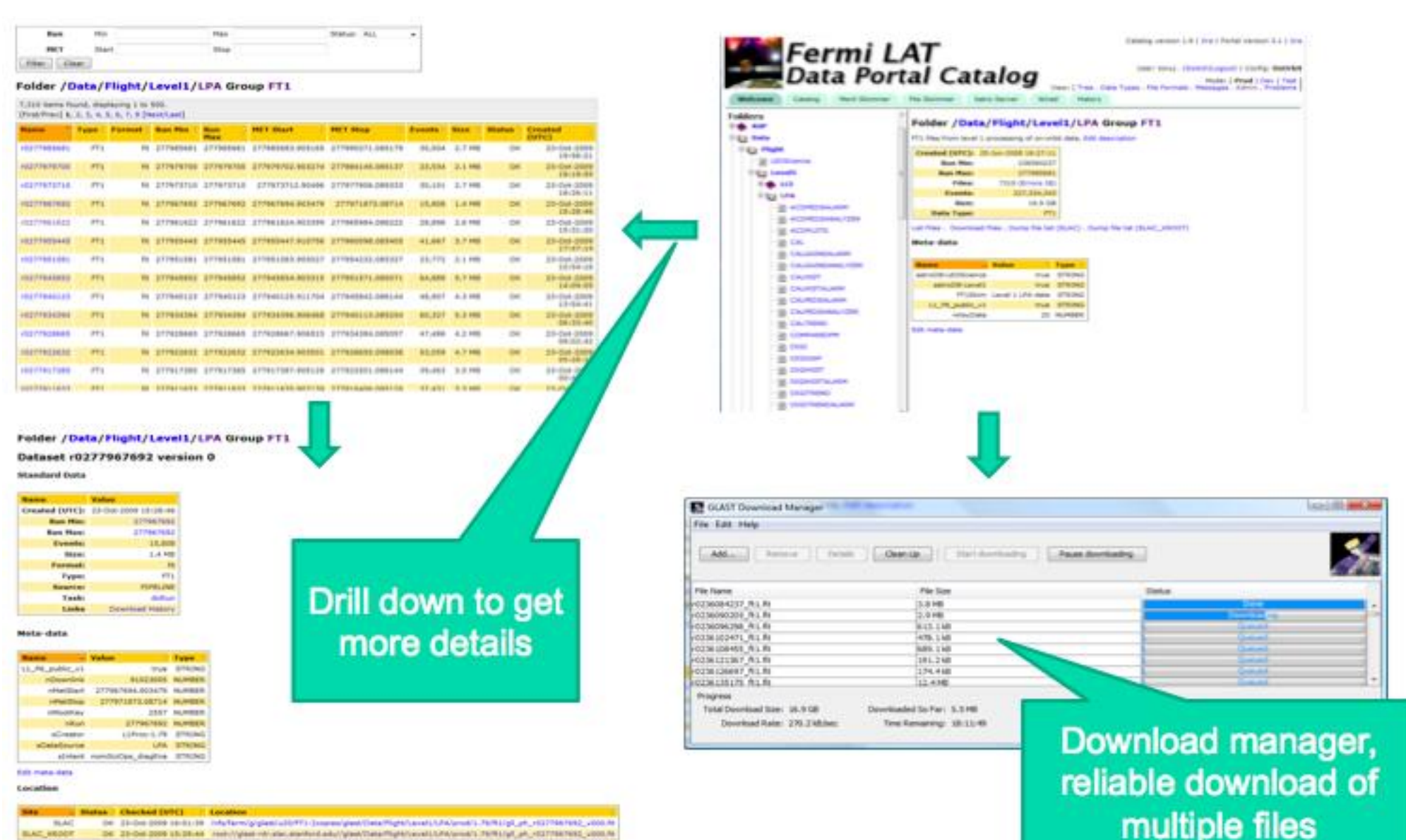
## Pipeline Processing

- Workflow engine to run independent "tasks" of processing
  - Have achieved > 40,000 jobs per day on >2000 cores
- Primary data processing run on SLAC batch farm
  - Up to 800 cores - can pre-empt other jobs to ensure quick turn-off when data arrives
  - Break up each run into many pieces to turn around in ~2 hours (before next downlink)
- Central dispatch sends jobs to Lyon computing center in France, primarily for simulations

## Simulation L1 Reconstruction



## L1 Digitization



Drill down to get more details

Download manager, reliable download of multiple files

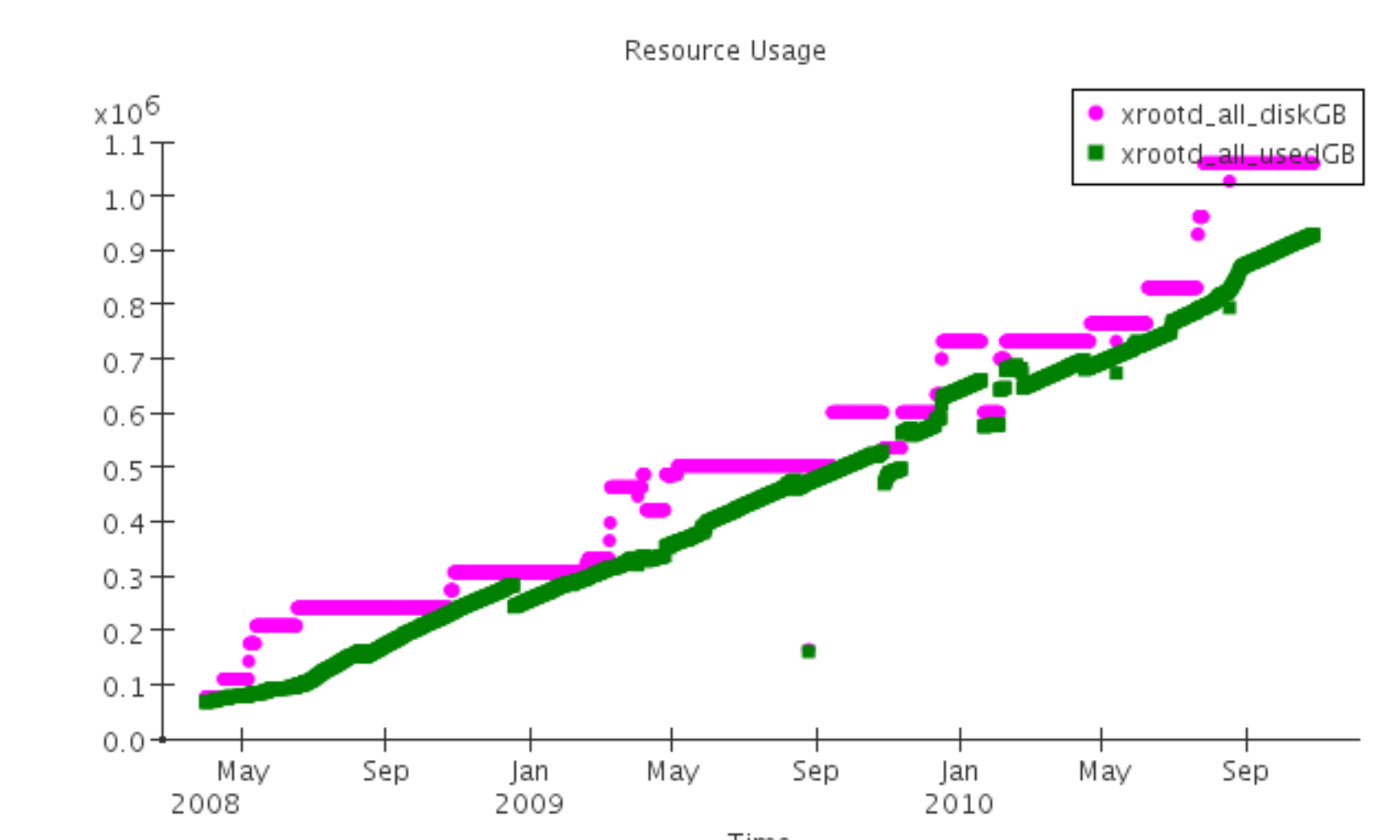
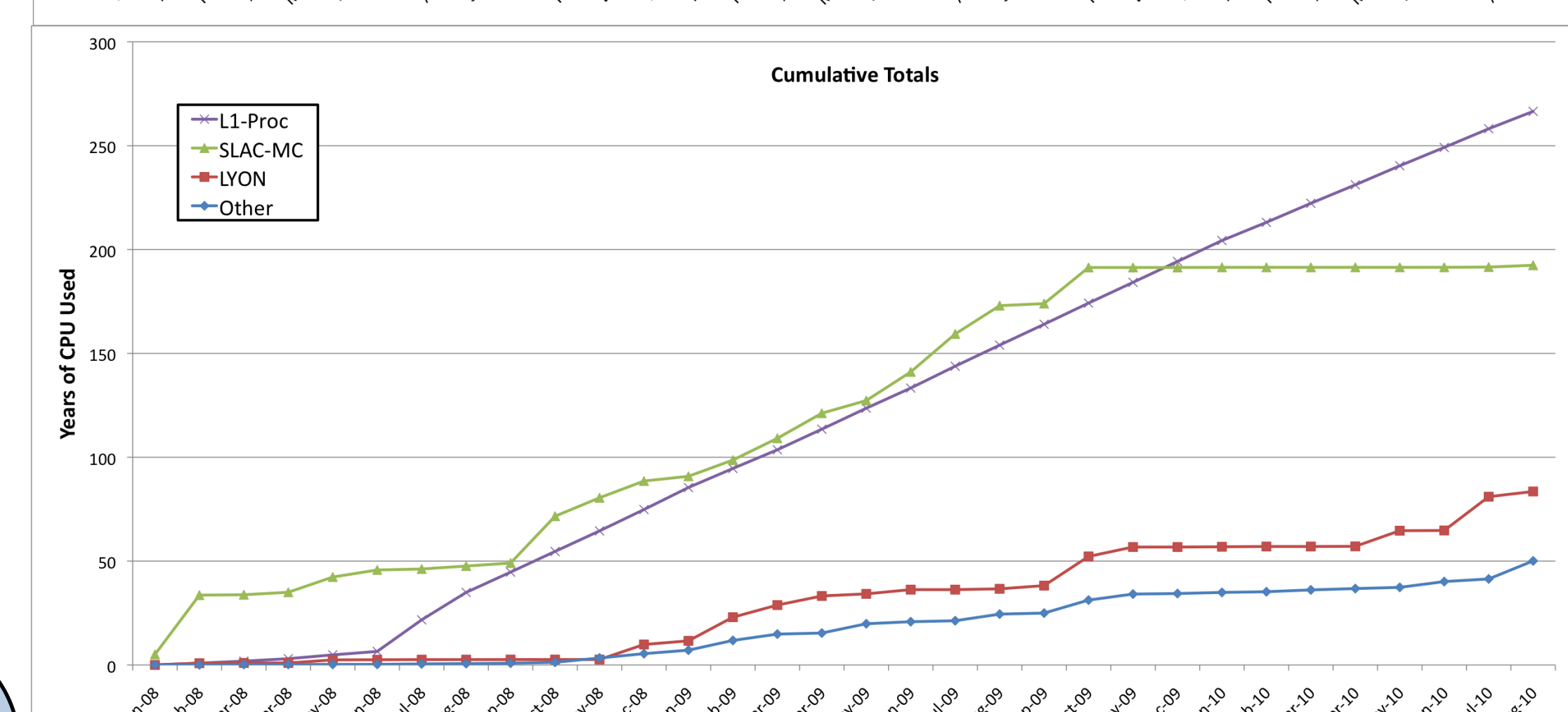
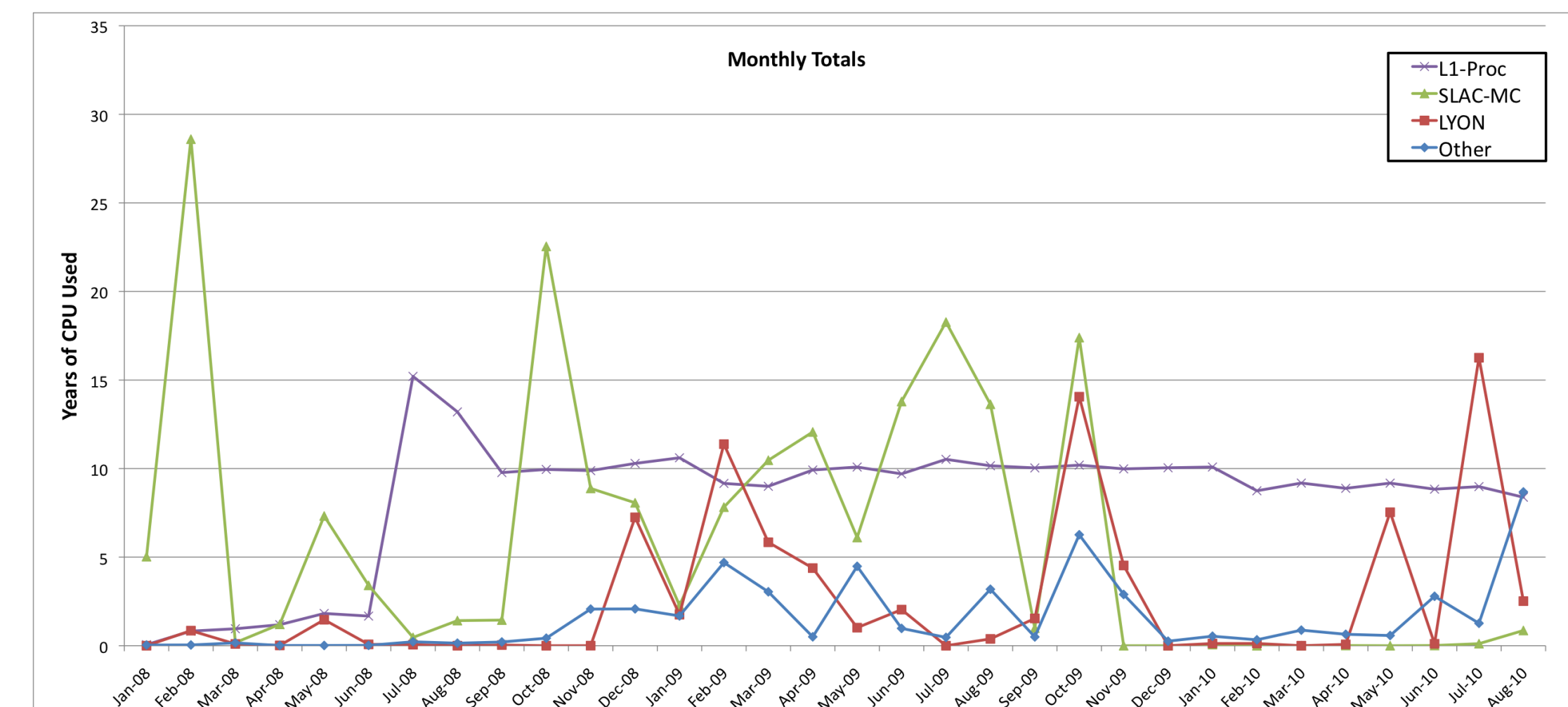
## Collaboration Access to Data

- All data files handled by the LAT are indexed in a data catalogue
- Java web-start gui app to download multiple streams of files
- Smart data servers to fetch data based on sky location and/or energy cuts

## Issues and Lessons Learned

No large system is constructed and works without issues. Here we highlight some of the hurdles we had to overcome and lessons learned along the way.

- Millions of Log Files** - overwhelming nfs file servers; only solution so far is to keep a pool of servers and switch when they fill up.
- Slow web pages** - some of the pipeline task monitoring pages have become very slow with the millions of jobs run. No fixes yet, but probably have to scale back on what is quickly seen. Getting painful.
- Monitoring the kitchen sink** - ~120k quantities being tracked (some different representations of the same detector quantities)! In canned histograms and ~30k stored in Oracle for dynamic trending
- Ever expanding definition of "usable photon"** - as understanding of LAT data deepens, more of it becomes useful. We are now storing 10x as many useable photons as planned. Made resource planning tricky, but expansion trays for the Oracle servers did the trick.
- PB of disk files** - storage model has been latest files all on disk - 0.75 TB/day. xrootd provides clustered file system with connection to hierarchical storage (tape). File load balancing and transparent HSM has required some effort on the client side.
- "Corporate Memory"** - maintaining experts' interest - how to keep a complex system humming in a stable experiment. We are leveraging the LAT tools for other experiments at SLAC to attempt keeping the tools under active development.



## Resource Usage

- Data processing steady at 10 CPU-yr and 25 TB per month
  - ~600,000 data files and counting
  - ~72 million log files!
- 600 CPU-yr of processing (data+sims) since 2008
- 1.1 PB of data on disk in xrootd cluster file system



richard@slac.stanford.edu