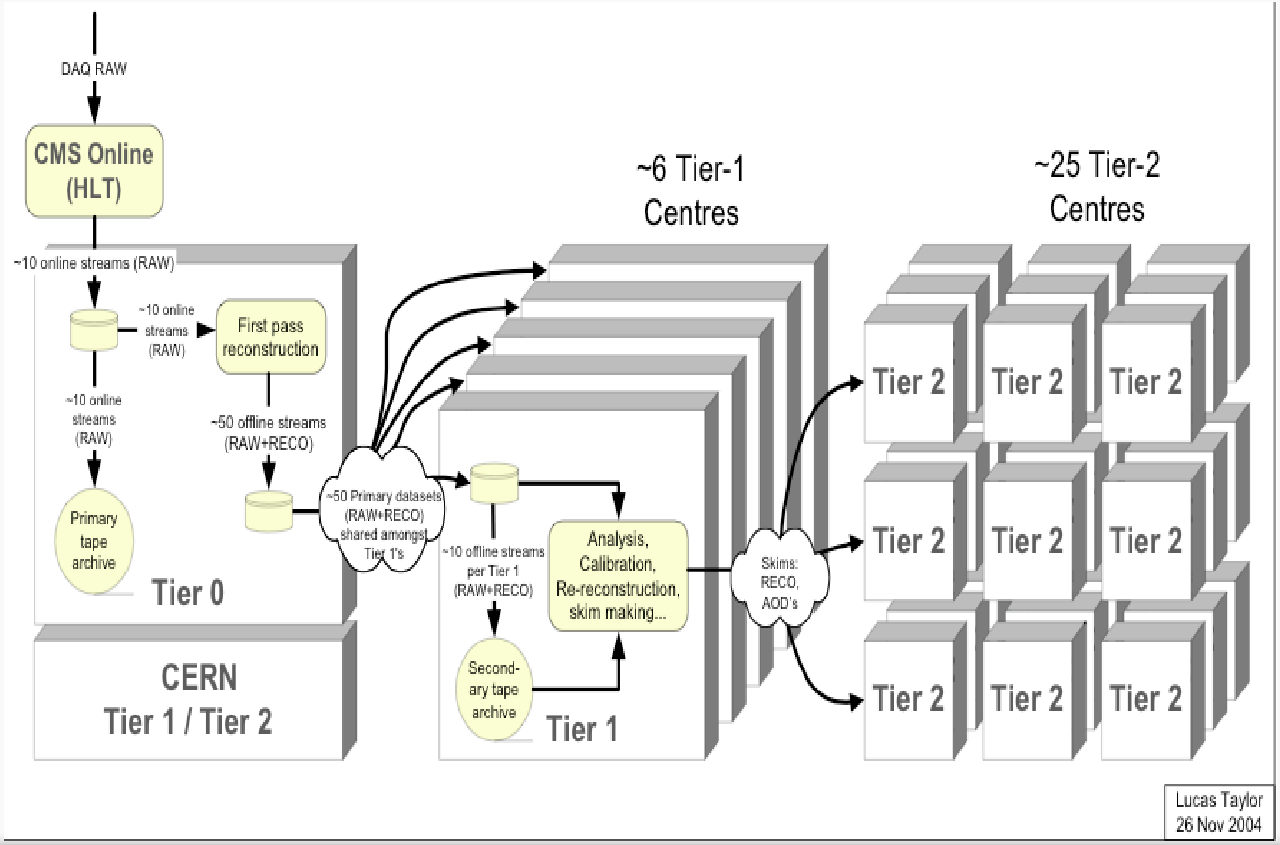
Development of the Tier-1 at FNAL

Jon Bakken, David Fagan, Ian Fisk, Lisa Giachetti, Oliver Gutsche, Joseph Kaiser, Tim Messer, Gary Stiehr, Hans Wenzel

The CMS Computing Model

CMS has proposed a computing model where the site activities and functionality is largely predictable

- Activities are driven by data location
- The majority of the computing capacity is located away from the experiment



Responsibilities of the Tier-1

Tier-1 Centers serve as an extension of the experiment on-line computing

- ➨ Share of raw data for custodial storage
  - Second copy of the raw data is distributed to Tier-1 centers
- ➨ Data Reprocessing
  - CMS anticipates 2 reprocessing runs per year

They are entrusted with serving the data entrusted to them

- ➨ Selecting and Skimming data for User Analysis and Calibration Tasks
- ➨ Data Serving to Tier-2 centers for analysis

Tier-1 centers also support Tier-2 centers

- ➨ Some specific operational support responsibilities
  - FNAL supports 7 US Tier-2 centers
- ➨ Archival Storage for Simulation and Important Analysis products from Tier-2 centers
  - Tier-2 centers typically do not have tape-based mass storage

The Tier-1 Center at FNAL

- FNAL is a dedicated Tier-1 Facility for CMS
  - Meeting the obligations of the U.S. to CMS Computing
    - Supporting the local community
  - The only Tier-1 center in the Americas
- By head count US-CMS is about 30% of the CMS collaboration
  - FNAL is about two nominal Tier-1 centers by the computing model numbers
    - The single largest Tier-1 center for CMS

FNAL Tier-1 2008	CPU	4.3MSI2k	1000 dual CPU nodes
	Disk	2PB	200 Servers (1600MB/s IO)
	Network	15Gb/s	CERN to FNAL
	People	30FTE	Includes Developers and Ops

- FNAL is completing the first year of a three year procurement ramp in preparation for the start of the experiment

Facility Services

Grid Interfaces:

- ➡ FNAL Supports both LCG-2.6 and the OSG-0.4 releases
  - Two doors into the same physical hardware
    - Cluster utilization is roughly half grid submission and half local jobs

Processing:

- ➡ All resources were switched to a Condor based system in 2005
  - Cluster is scaling well. Priority scheduling allows reasonable allocation of resources.
    - Currently 1000 batch slots. Experience through CDF with several times more

Storage:

- ➡ dCache/Enstore deployed for Mass storage
  - The dCache system has performed well under heavy load
    - Over 200TB delivered to applications in a single day

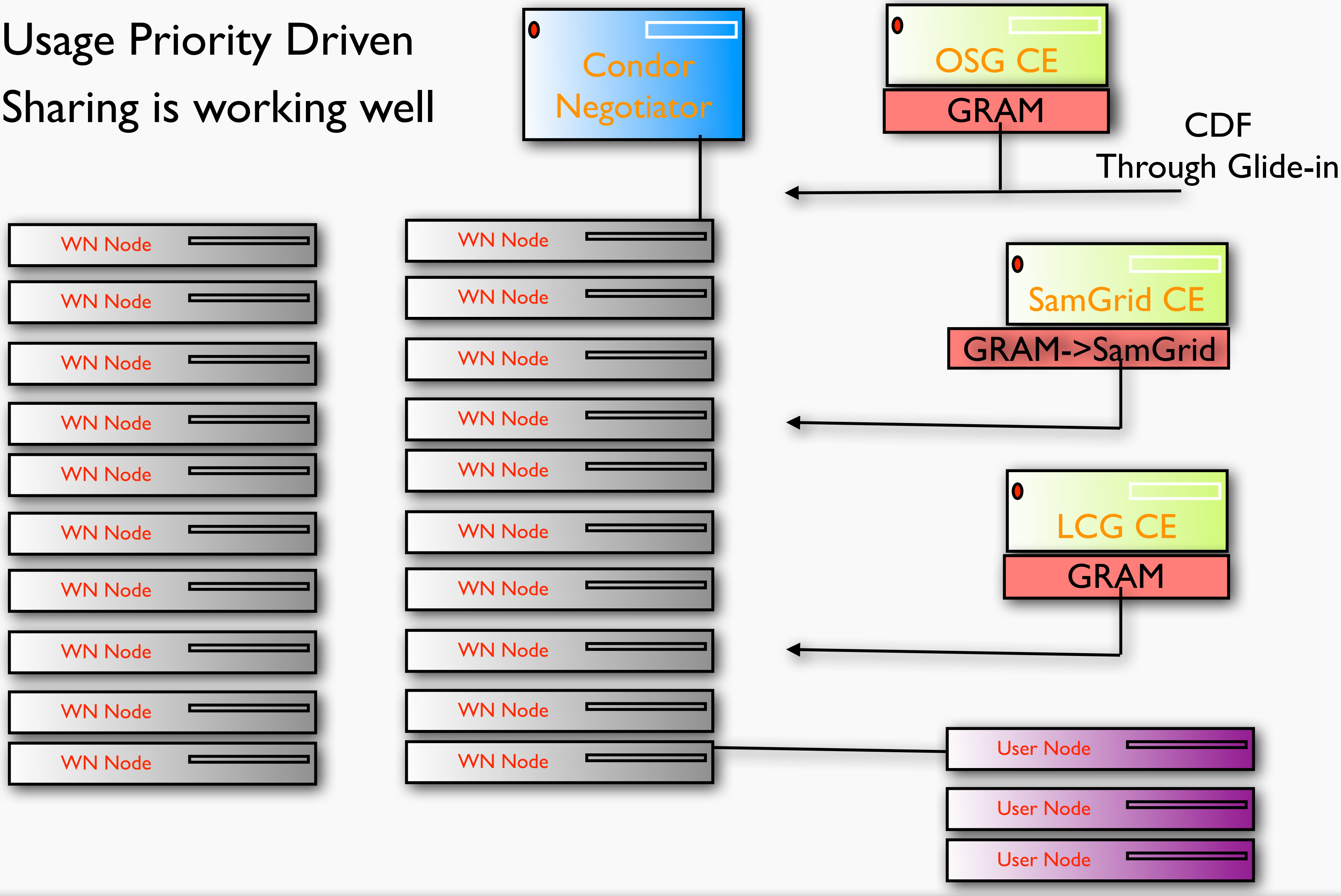
Networking:

- ➡ Current we have a 10Gb research link

Grid Deployment

LCG and OSG have individual gatekeepers

- Usage Priority Driven
- Sharing is working well



Grid Exerience

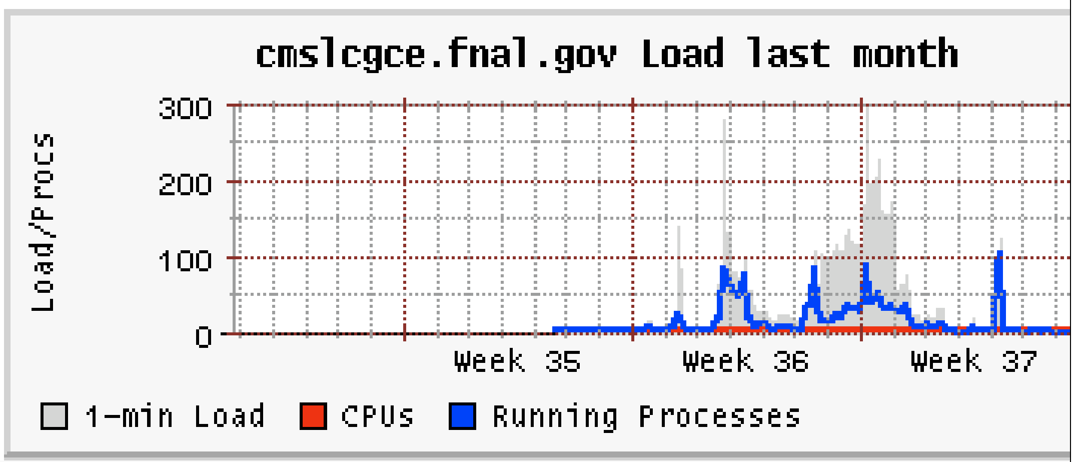
The deployment experience with LCG and OSG is generally good

- Packaging and distribution efforts are paying off

We have had thousands of job in the LCG queue and several hundred running jobs. In the OSG job queue we have balanced opportunistic users

We support primarily user analysis jobs through CRAB on the LCG (CMS Remote Analysis Builder)

- User Support load
- New failure modes
- Discovered the distributed file system could not keep up with the process tracking and locking mechanism deployed in the LCG. Worked around.



We currently supported mainly simulation production through the OSG interface

- In addition we have opportunistic use by other sciences

Processing Resources

FNAL is currently at ~500 dual CPU nodes (1000 CPUs)

- Slowest are 2.4GHz Xeons and the Fastest are single core Opteron 268s
- Facility is ~1000kSI2k (25% of the expected capacity in 2008)

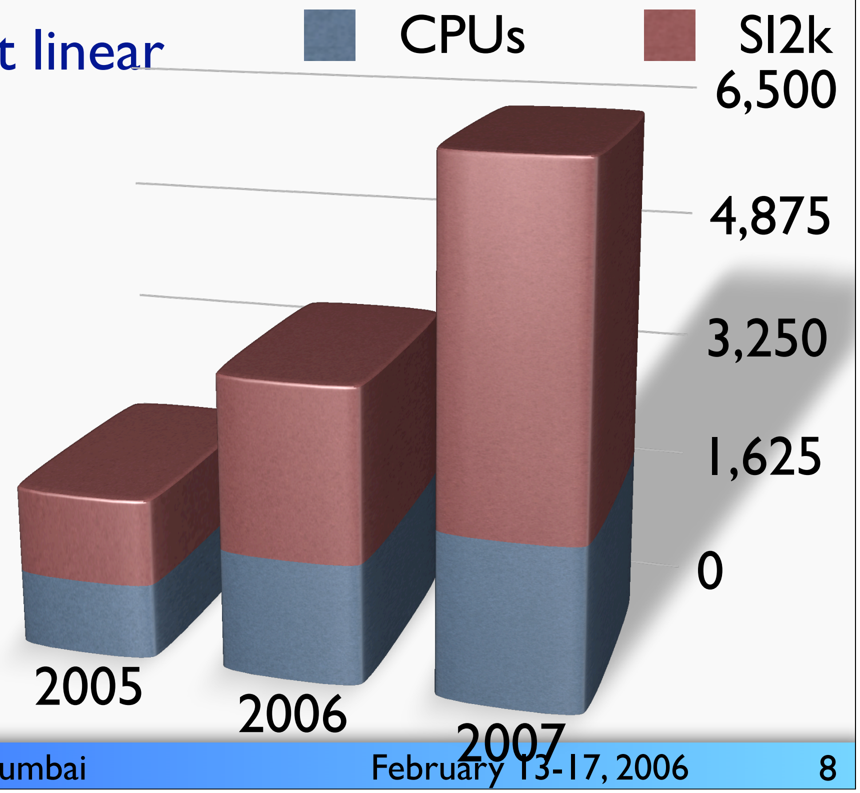
The operational ramp to the start of the experiment is manageable

- Experience at FNAL configuring and running farms this size

The increase in number of nodes is almost linear

- Performance increase is a fairly conservative improvement estimate
- Dual cores may improve the situation

Estimate is that all worker systems will be gigabit connected



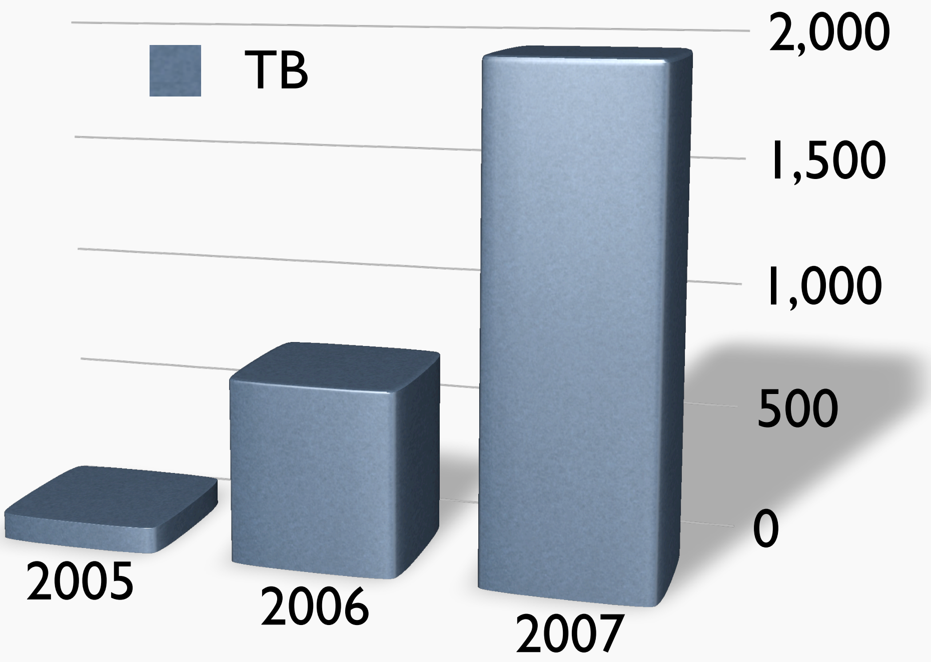
Storage Resources

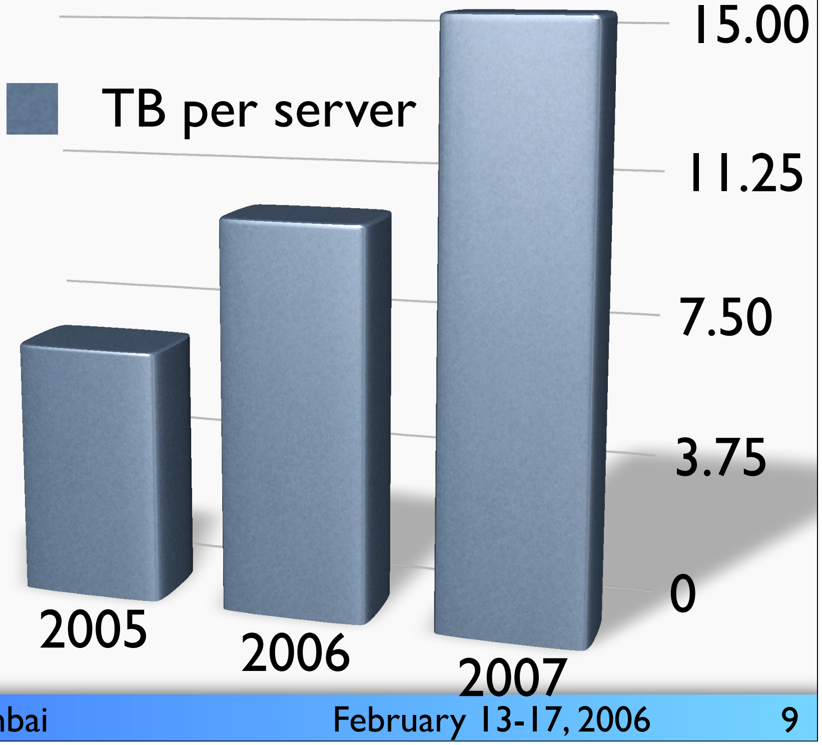
Currently the FNAL Tier-1 has ~100TB of dCache storage

- Roughly 5% of the capacity expected in 2008, but 20% of the number of servers

Very steep operations ramp in disk storage before the experiment start

- Procuring, deploying and commissioning at a large scale

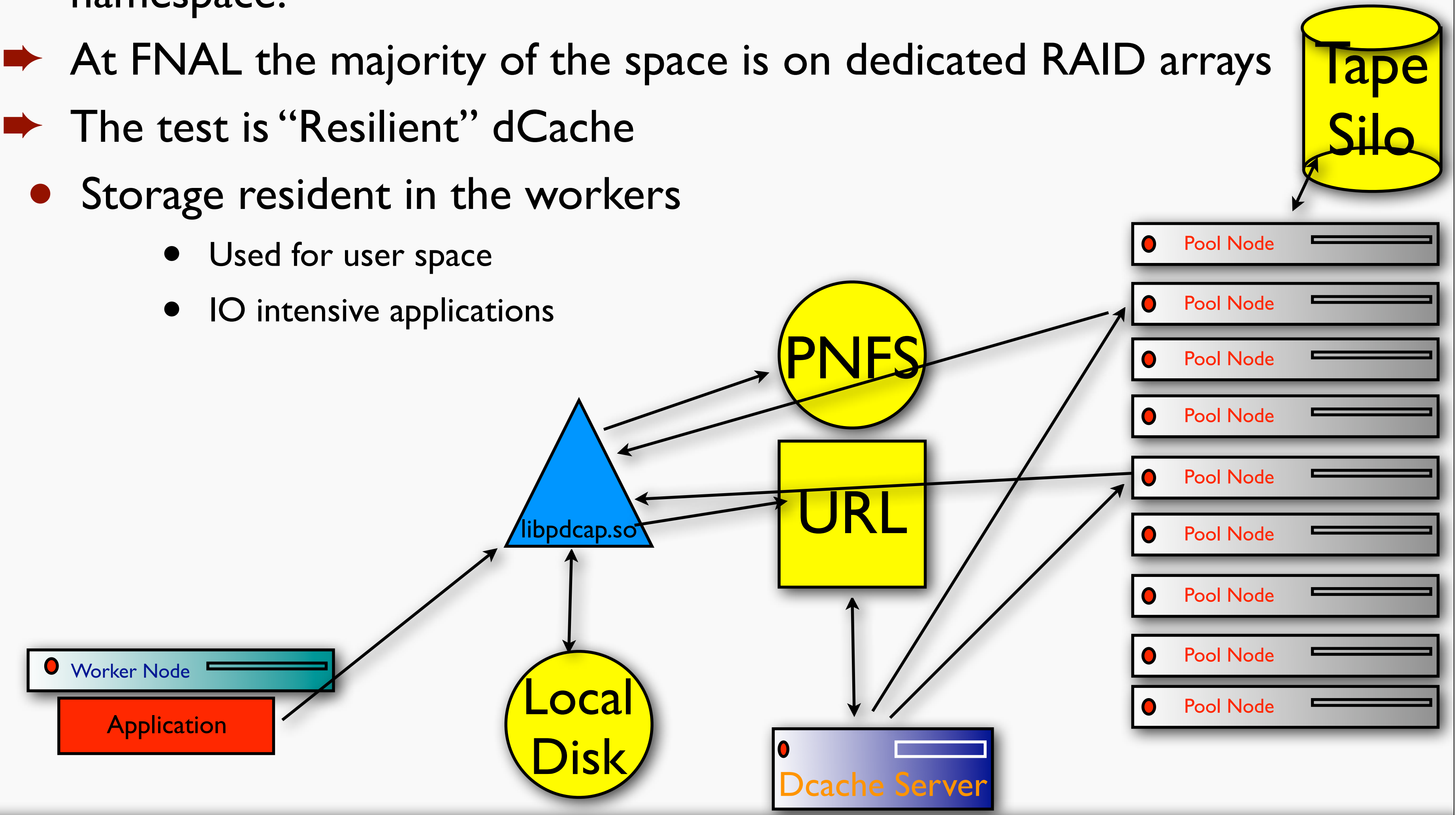




dCache Storage

dCache was jointly developed by DESY and FNAL

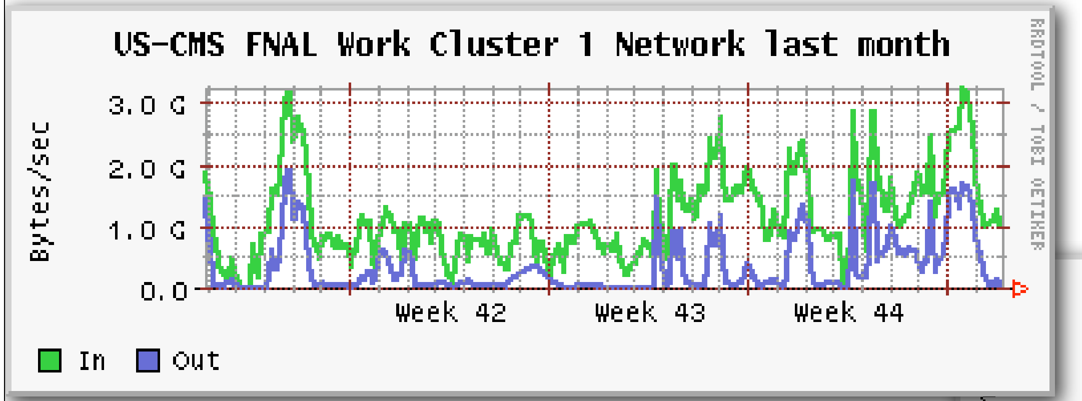
- Allows a group of physics disk resources to have a consistent namespace.
- At FNAL the majority of the space is on dedicated RAID arrays
- The test is “Resilient” dCache
  - Storage resident in the workers
    - Used for user space
    - IO intensive applications

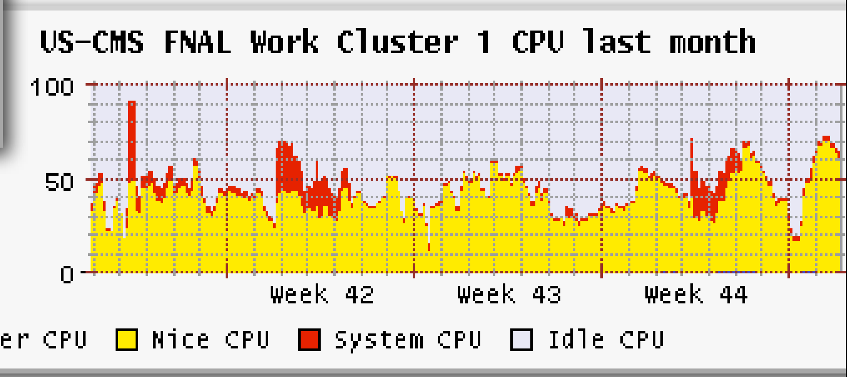


Stress Testing dCache

dCache Storage

- The CMS Application had a feature that caused the buffer to be inefficiently used in dCache. It has since been fixed.
- It was an excellent performance test of the system
  - Sustained periods of 2 and 3 gigabytes per second. More than 200TB served in a day
    - Higher than expected rates in 2008.
    - Even with the high rate we are seeing lower than expected CPU efficiency



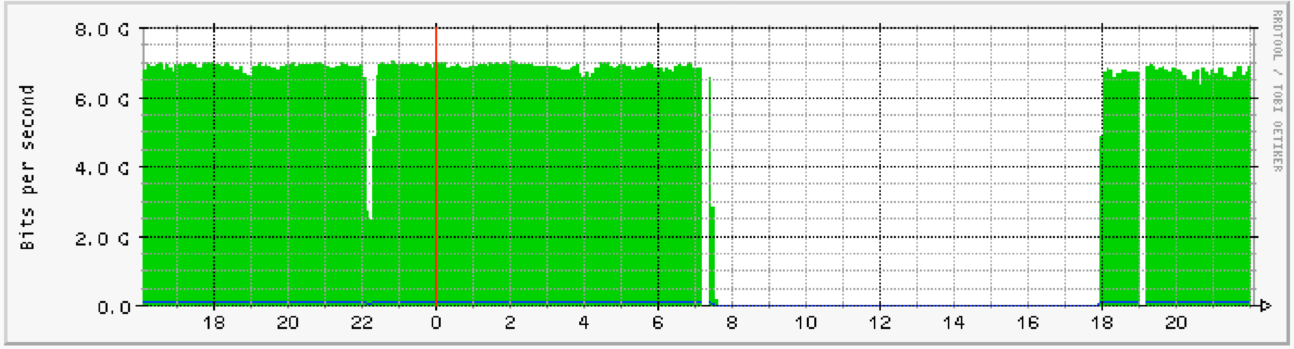


Network Resources

The WAN networking for CMS is provided by a 622Mb/s production link and a 10Gb/s research link (second 10Gb/s link is available to light).

- Most of the CMS data traffic goes over the research link

Traffic between CERN and FNAL during a Service Challenge



The LAN networking is provided by 10Gb/s links between large switches

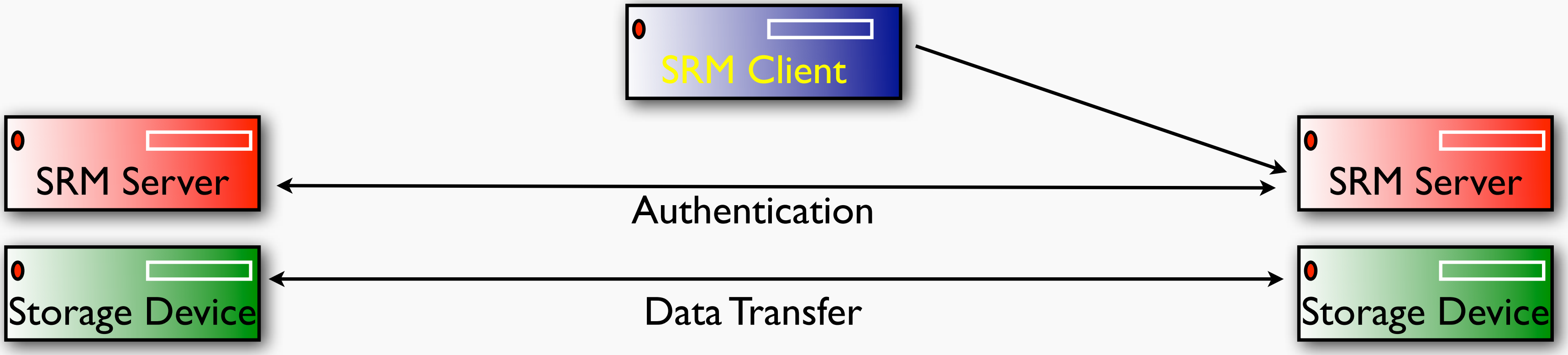
The dCache and mass storage systems are in one building while the worker nodes are in a dedicated building

- The buildings are connected for CMS with two 10Gb links
  - We recently hit 80% utilization of 20Gb/s and will shortly add a link
- The CMS estimates for data delivering in 2008 is 1600MB/s
  - We estimate needing to grow to 4 10Gb/s

Interface to Storage

We have SRM version 1.1 deployed for the dCache storage element.

- Allows load balancing and traffic shaping
- Scalable solution to enable multiple servers to send and receive gridFTP streams
- Migrating to SRM version 2 in the first quarter of 2006



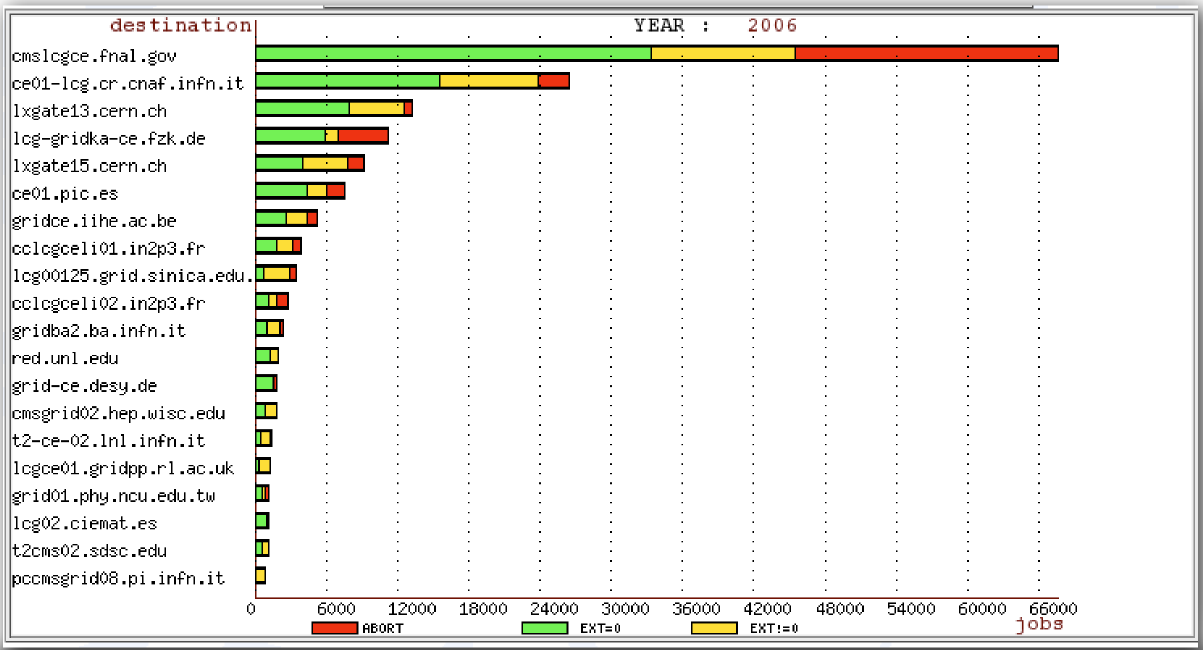
Growing User Load

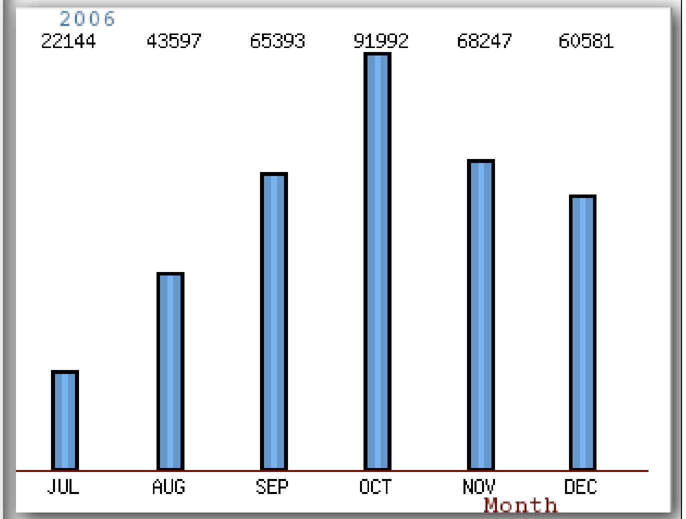
There are over 450 individuals signed up for interactive access to the CMS farm at FNAL

- Somewhat ahead of our projected ramp for users.
- At any given time 10-15% of those are actively running jobs.

There is 10TB of user controlled disk space on a distributed filesystem

- Some quota controlled space, some physics group managed space
- Heavily accessed





Outlook

CMS Tier-1 at FNAL is Growing

- ➡ Procurement ramp for the final two years is steep, but we expect to be ready with the appropriate sized center at the beginning of 2008

Grid, Storage and Processing Services are coming on line and becoming more reliable

- ➡ Operations experience still needs to improve and some of the services need to become more robost
    - ● Grid submission failures and failures of CMS services are still too high
- ➡ Facility scale is roughly one forth of the final capacity
- ➡ Services are increasing in performance and capability, but development is needed

User access and subsequent support load is increasing

- ➡ Grid user is ramping up