Tier-1 Summary from FNAL

Ian Fisk

LHCC-LCG Review

November 14-15, 2005

SC3 and Grid Deployment

SC3 for FNAL was a reasonably small evolution over previous activities

- SRM transfers were exercised as early as SC1
  - Failure rate in the SRM transfers was higher than before and increased operational load, but retries in the system kept this manageable
- Tape transfers were exercised in SC2
- Data hosting and publication for CMS and support for the analysis applications have gone on for several months
- Only a few SC3 services hadn't been tested before
  - OSG Analysis submission needed to be developed
  - Spent most of the SC3 effort helping to bring up the US Tier-2 centers

Grid Deployment

- Generally succeeding
- Scale increases every month and facility issues need to be found and fixed.

Scale of the Existing Facility

The Tier-1 Center at FNAL is completing the first year of a three year procurement cycle in preparation for the start of the experiment

FNAL used a system of about 10-20% of the complexity and capacity of the final system

- Currently we have 460 dual CPU Processor nodes in service
  - Grows by another 700 CPUs
- Around 40 server systems for facility and grid services
- 100TB of Normal dCache Space based on RAID5 devices
  - Grows by 2pB
- 75TB of resilient dCache space in the worker nodes themselves
- ~200TB of mass storage space
  - Grows by 4PB

Facility Services

Grid Interfaces: (LCG and OSG used in SC3)

- FNAL Supports both LCG-2 and the OSG-0.2 releases
  - Two doors into the same physical hardware

Processing: Condor for local batch queue (Execute Analysis Jobs)

- We switched all batch resources to condor, still learning the optimal configuration, but we have been happy with the setup
  - CDF experience at FNAL indicates we should be able to scale to goals
  - Priority scheduling works well, but we would like to implement hierarchical priority schedule this winter

Storage: dCache/Enstore for Mass storage (Data Replication and Analysis)

- The dCache system has performed well under heavy load
- New resilient instance is a nice feature. Performance good and users like it

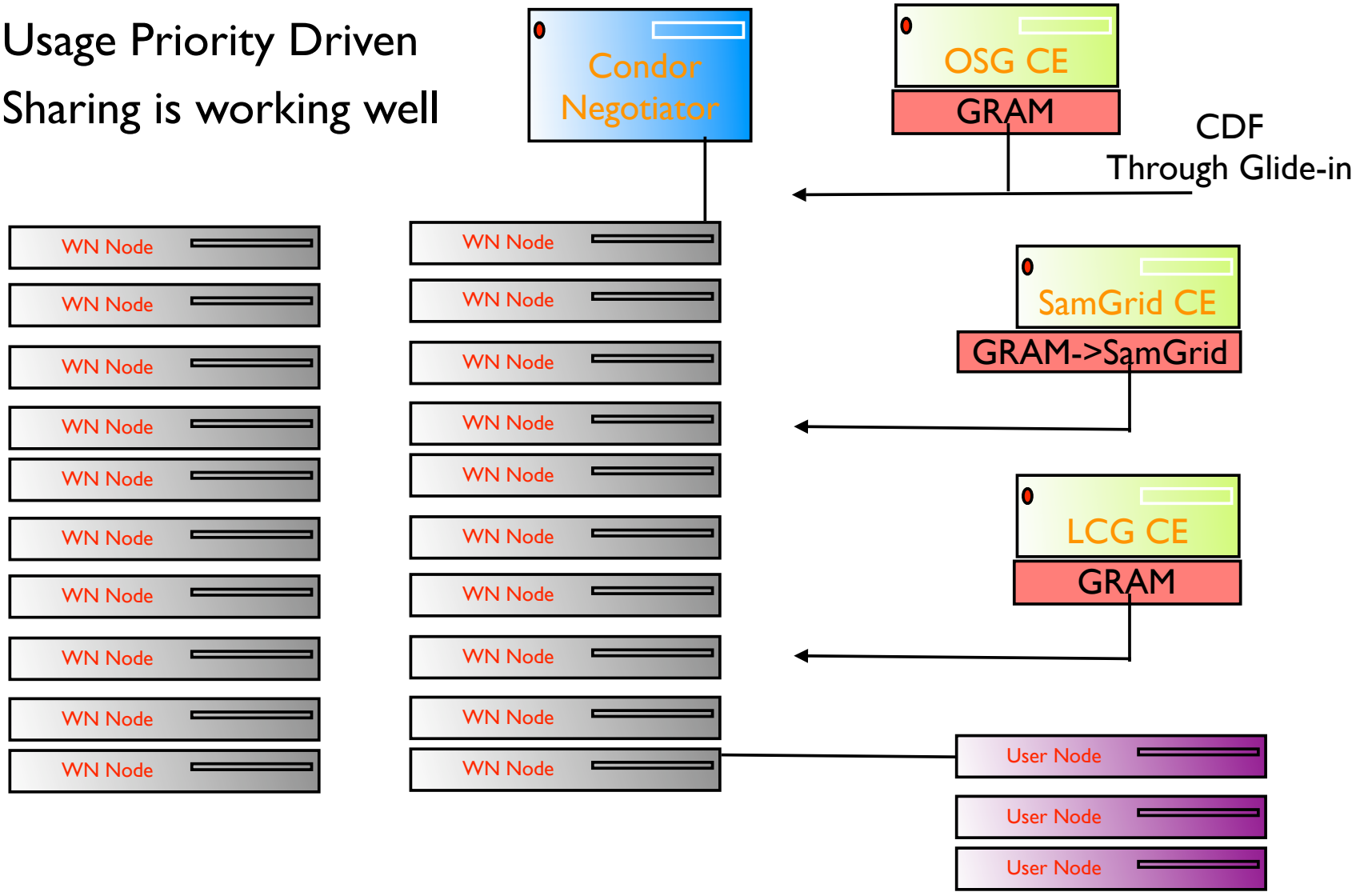
Networking: Current we have a 10Gb research link (Used in throughput)

- Progress toward a production link

Grid Deployment

LCG and OSG have individual gatekeepers

- Usage Priority Driven
- Sharing is working well



LCG Deployment Experience

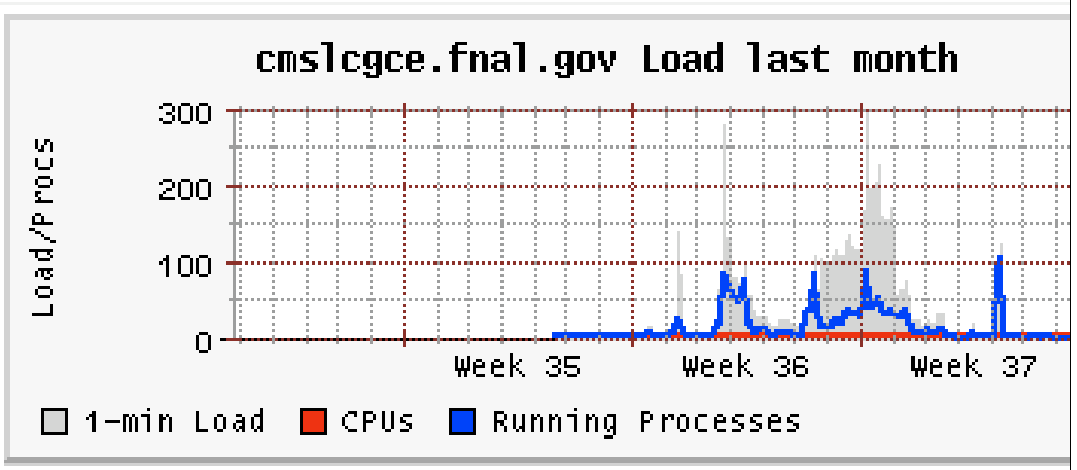
The deployment experience with LCG is generally good

- Packaging and distribution efforts are paying off

We have had thousands of job in the LCG queue and several hundred running jobs

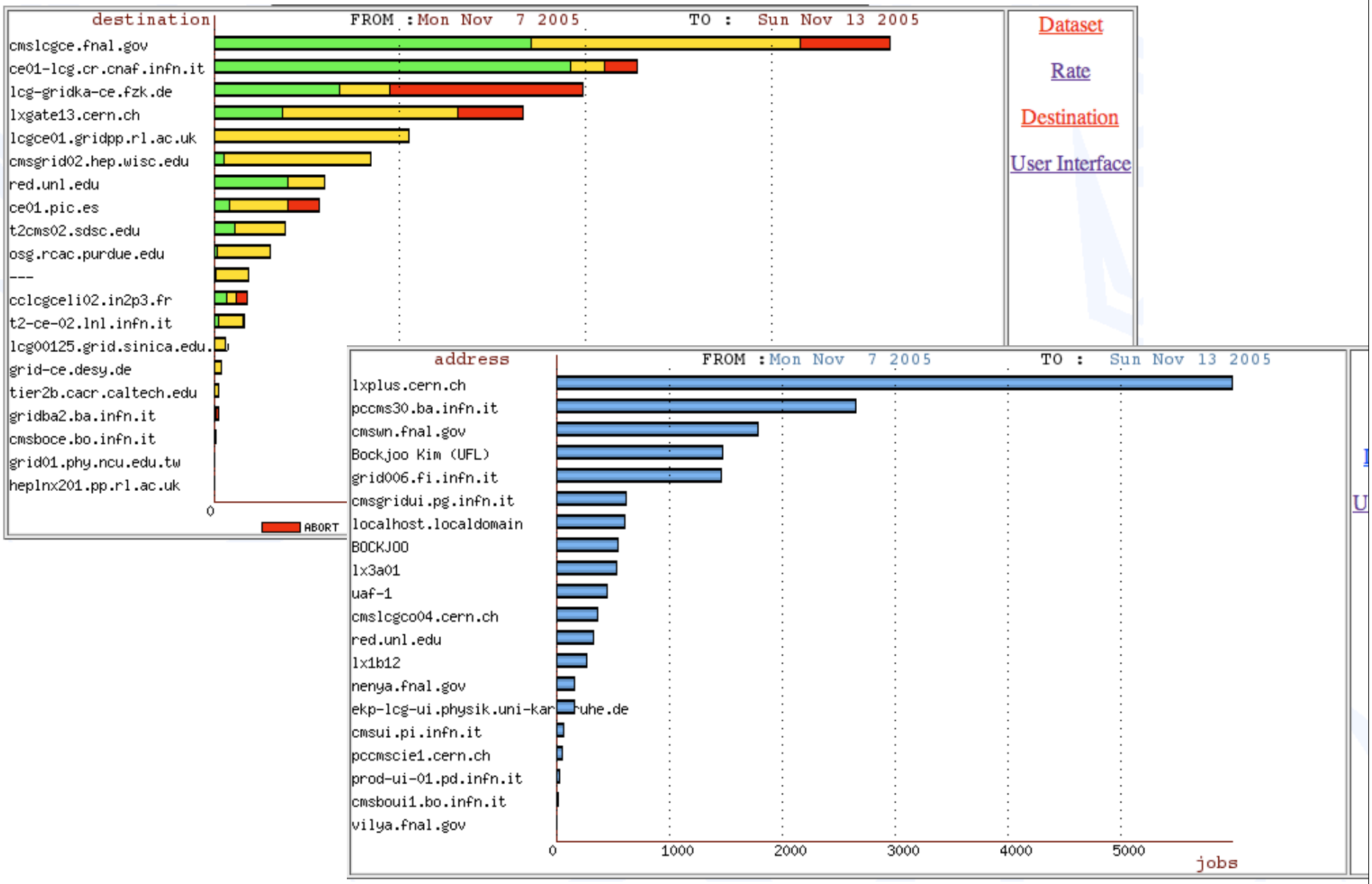
We support primarily user analysis jobs through CRAB on the LCG (CMS Remote Analysis Builder)

- 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.



- The execution site is final step in a chain. When problems happen upstream it is often hard to distinguish from a site problem

CRAB Plots Comb. of SC3 and CMS Users



OSG Experience

OSG Deployment and Operations Experience is also good

- Automated installation has become pretty reliable

OSG at FNAL is primarily used for CMS simulated event production

- OSG has implemented support for VOMS extended proxies with roles and groups
  - voms-proxy-init to define a production role
  - mapping callout assigns anyone with the role to a production user with a higher priority in the batch system
  - Many hundreds of jobs simultaneously

We also support opportunistic VO usage through the OSG interface

- Biology, astrophysics and gravitational communities. CDF uses glide-in
- Only give significant resources when the farm would be otherwise idle

Recently Tier-2 sites have supported CRAB jobs through the OSG interface

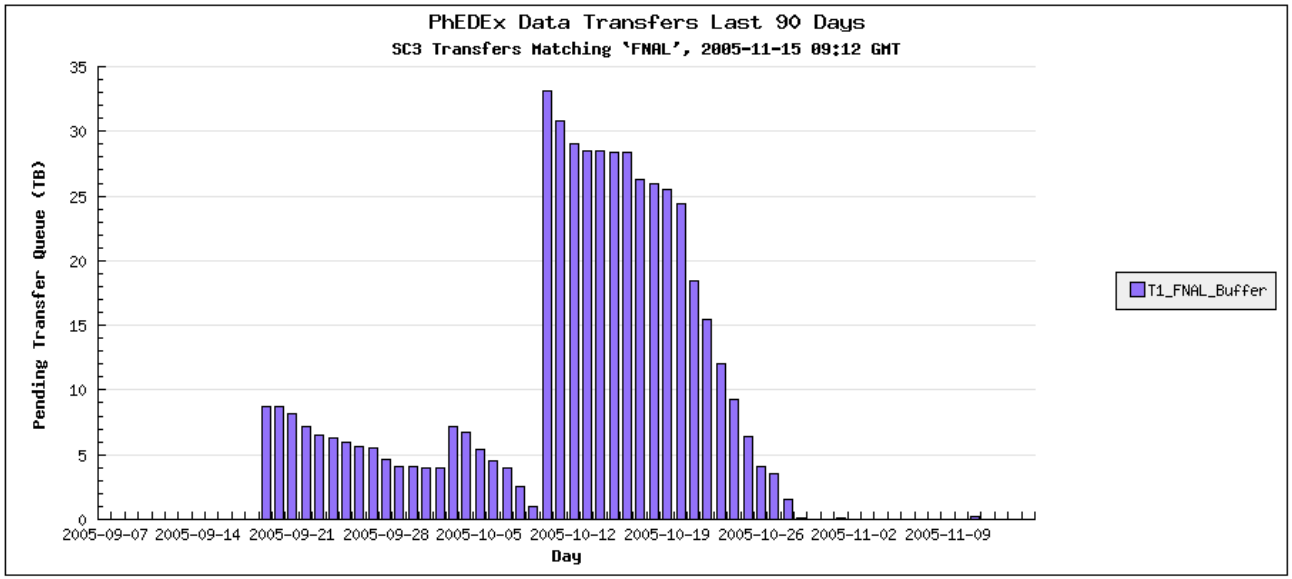
- Working nicely (5 of top 10 execution sites are US Tier-2s: SC3 jobs)

SC3 Transfers

SC3 Transfers to FNAL Buffer worked

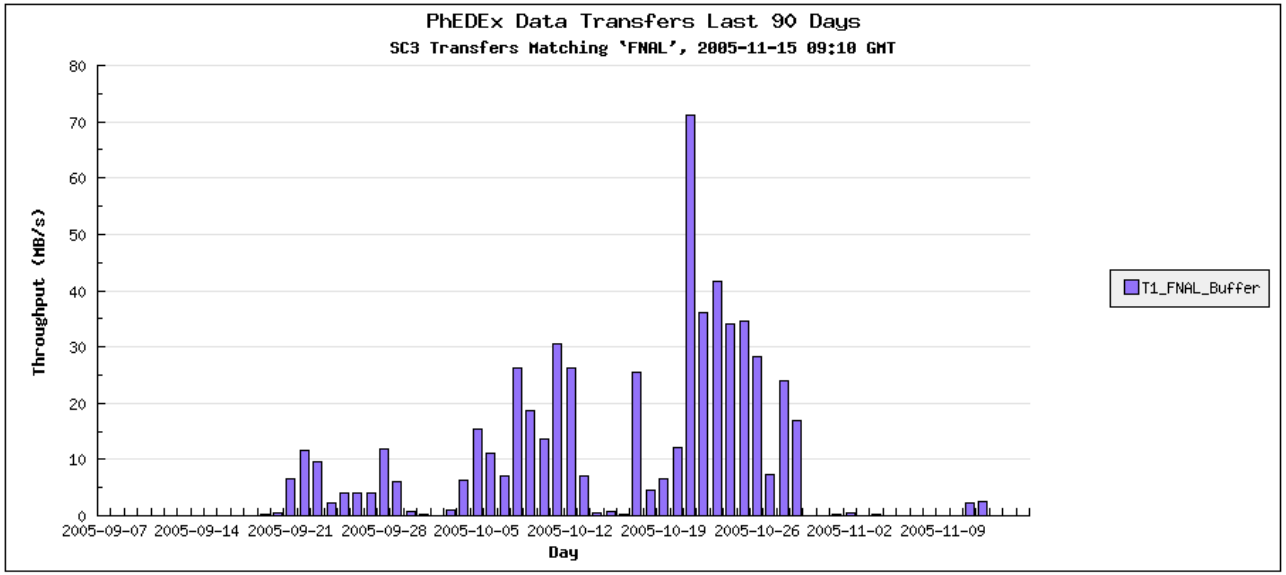
- 50TB transferred

Failure rate on SRM was high



Interesting interference effects

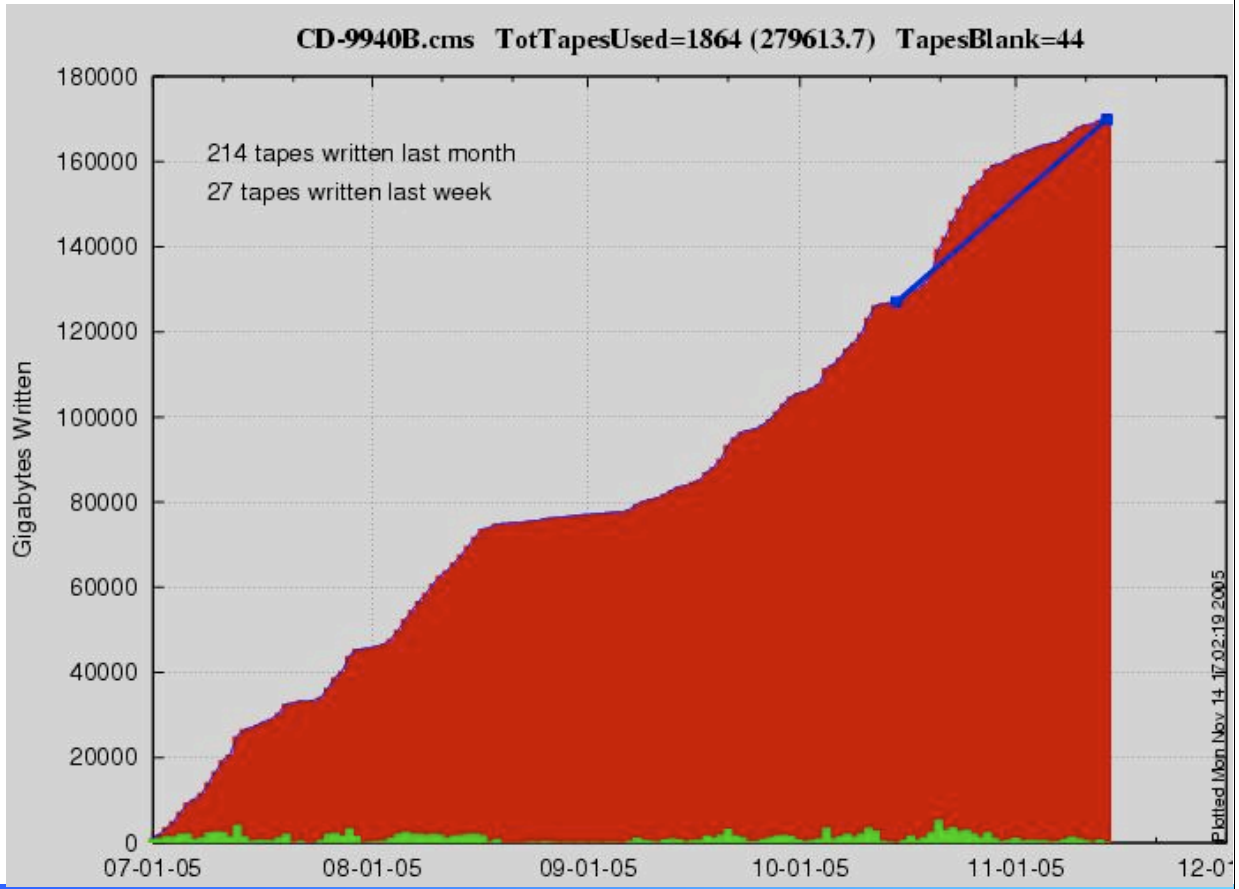
Deployment of priority queues in dCache made a big difference



Storage Tape

The Service Challenges have resulted in a lot of tapes written

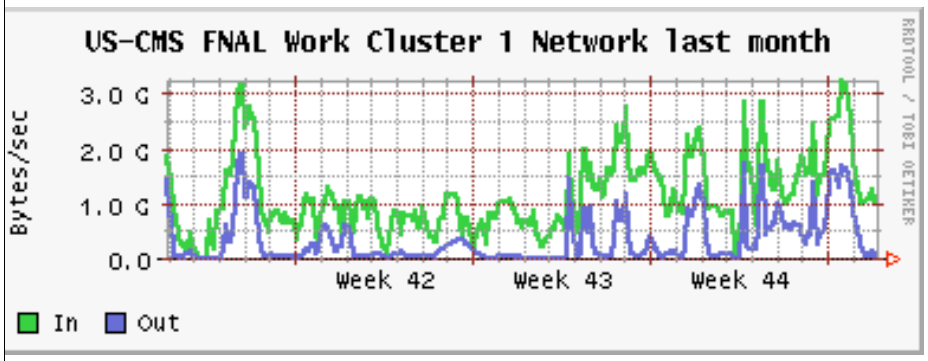
- Early challenges of just writing junk allowed for fast cleanup
- During SC3 data is read into mass storage and eventually to tape, though also accessed from disk
- Some data is samples we don't have and wish to keep for analysis community
- SRM transfers to FNAL had reasonable performance
- Failure rate is still high

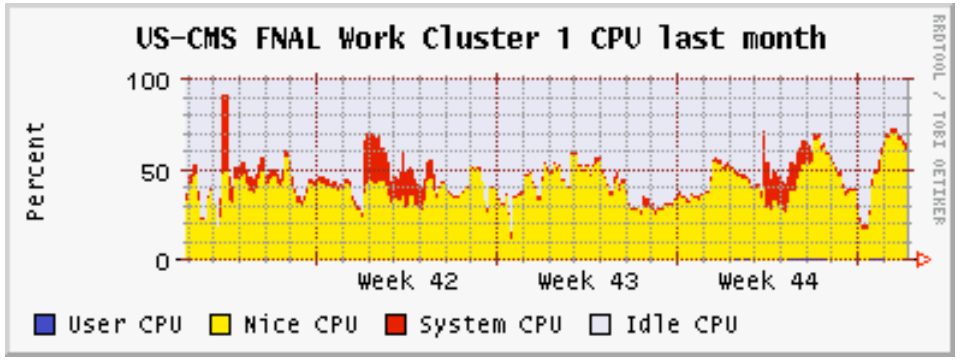


Storage dCache

dCache Storage in SC3

- The CMS Application is particularly hard on the dCache system
  - Uses buffer very inefficiently. This has been fixed in the newest release
- Over the last month we have averaged higher than 1Gigabyte per second
  - Sustained periods of 2 and 3 gigabytes per second. More than 200TB served in a day
    - Higher than expected rates in 2008. An excellent facility test
    - Even with the high rate we are seeing lower than expected CPU efficiency
    - Anxious to get this fixed
    - Repaired in new release, but a lot of older releases in use for PTDR





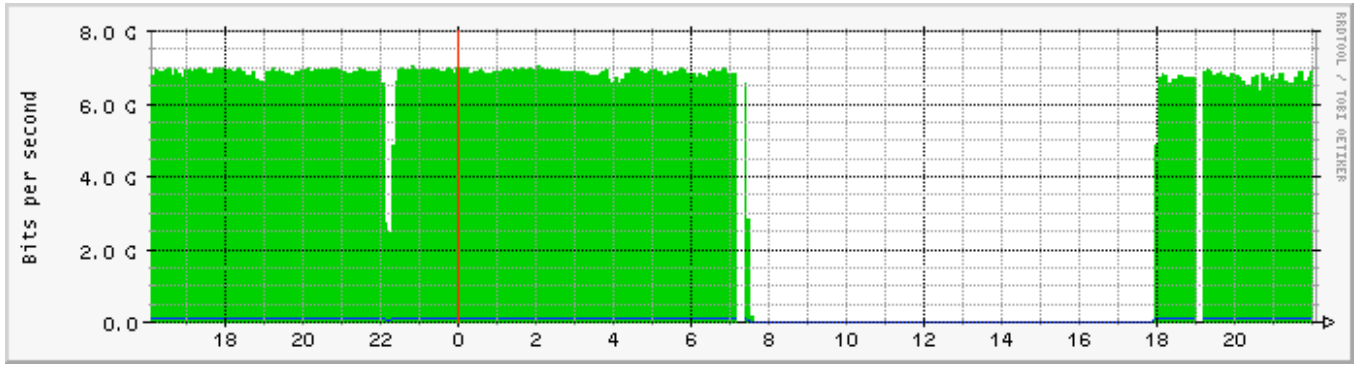
Networking

There are two sections of the network under stress.

- At FNAL there are two computing facilities: FCC the older center has generator backup and is home to the disk servers. Most of the worker nodes are located in the Grid Computing Facility
  - The two buildings are connected over a large bundle of fibers. US-CMS has 2 x 10Gb, which we recently hit 80% utilization for

The 10Gb research link to Starlight and then 10Gb LHCNet link to CERN are heavily utilized during the throughput phases

- Plot from SC2
- Rate lower in SC3 because more balanced



Fishing boat

- Promptly failed over to alternate path

Experiment Specific Services

The effort associated with the CMS specific services for SC3 was modest from the site standpoint (if site had experience)

- PhEDEx installation for data transfer
  - Working nicely as a relatively mature service
- PubDB for data publication and CMS Preparation scripts
  - Current CMS Data model is fragile and hard to support
- Recently started contributing to the Analysis Tool Development
  - SC3 was a good catalyst for development
  - CRAB deployed for local FNAL community
  - Supporting submission to OSG

Spent a good deal of effort helping to bring up Tier-2 sites

- More than half the US Tier-2 only started on May 1
- Reasonable progress in a short time

Outlook

From the standpoint of the FNAL Site all the Service Challenges have been generally positive experiences

- We had more debugging to do in previous challenges.

- Not all services scale will scale to the requirements
  - Will require development effort
  - Facility deployment will be a necessary tool to test service improvements
- Increasing local and grid supported community
  - Need to get a handle on user support

Looking forward to SC4