FNAL T1 and the US region in CCRC'08/phase-2

Jon Bakken

Friday Facilities Ops Meeting

June 6, 2008

Furloughs

People, including myself, were on furlough during the month.

- Didn't pay full attention to tests because we were not at work
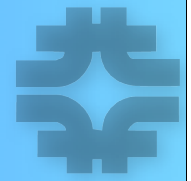
- Couldn't talk to necessary people.

It's great the furloughs are over!

Now we just have to deal with 'forced vacation' - use it or lose it.

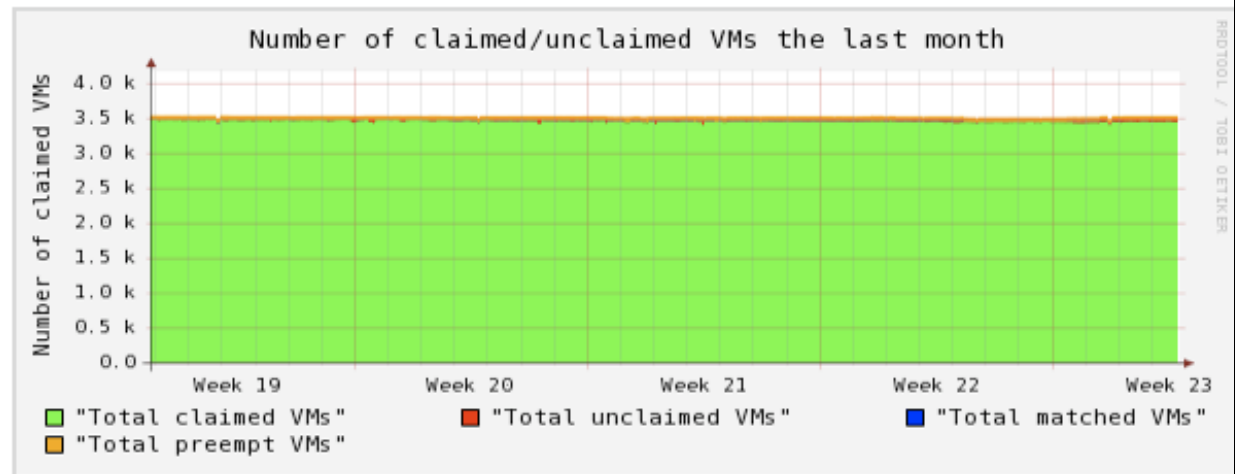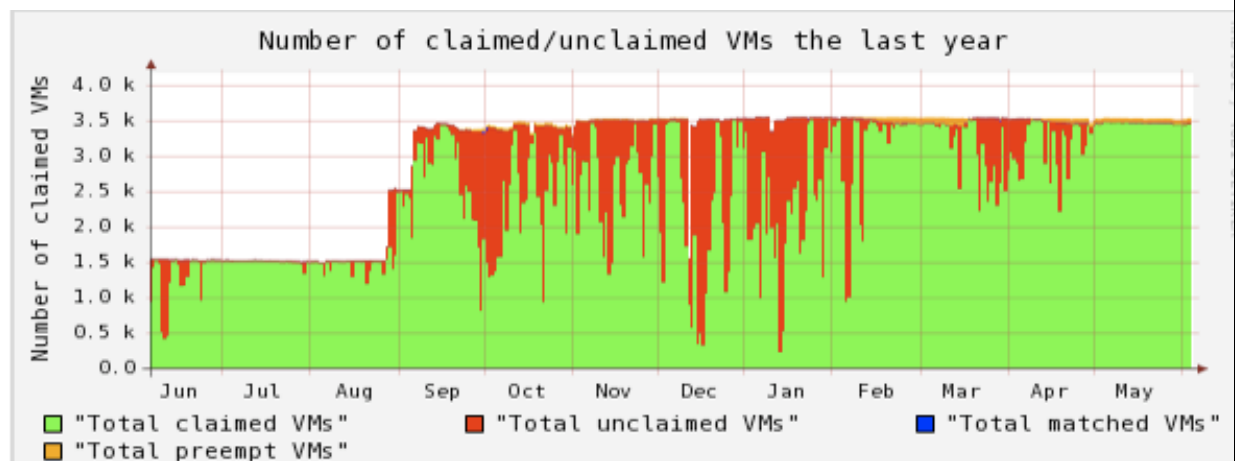- Lots of vacation coming up for Tier-1 staff in next few months.

Jobs

Running jobs on the FNAL cluster worked extremely well. Continuously running ~3500 jobs by production only, other users were starved out.

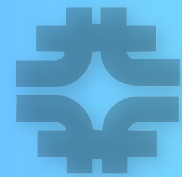Status last month: Entirely green means fully utilized

Status last year: red indicates some excess cpu.

Jobs

There was a comment during the month that jobs run faster at CERN than at FNAL. This is true, but it has to do with output IO, not cpu speed.

- Job Input   was read   via dcap - worked very well.
  Job Output was written via srm - added stress to overall system, relatively slow
       Note 3500 jobs writing output via srm = lots of srm stress.

  - SRM transfers - expensive, especially on site, due to authentication/ authorization for each transaction, and for 'detailed' pnfs queries. Also large overhead for SRM itself.

  - All our dcap doors at FNAL are read-only because we insist on using SRM functionality to authenticate and to direct files to root path location.

  - We just finished coding a restrictive dcap door - allows writing for specified users who know a simple secret. This will allow cmsprod jobs to write using dcap protocol, and deny the rest of the users.

  - This should dramatically increase IO write speed at FNAL, and reduce the stress on the SRM server.
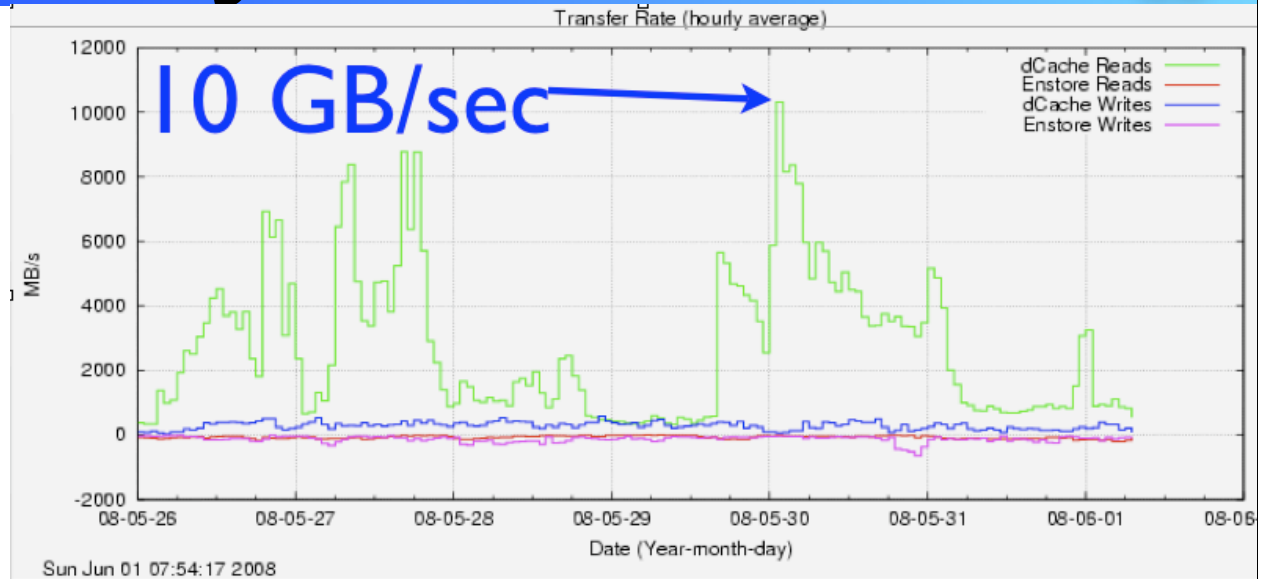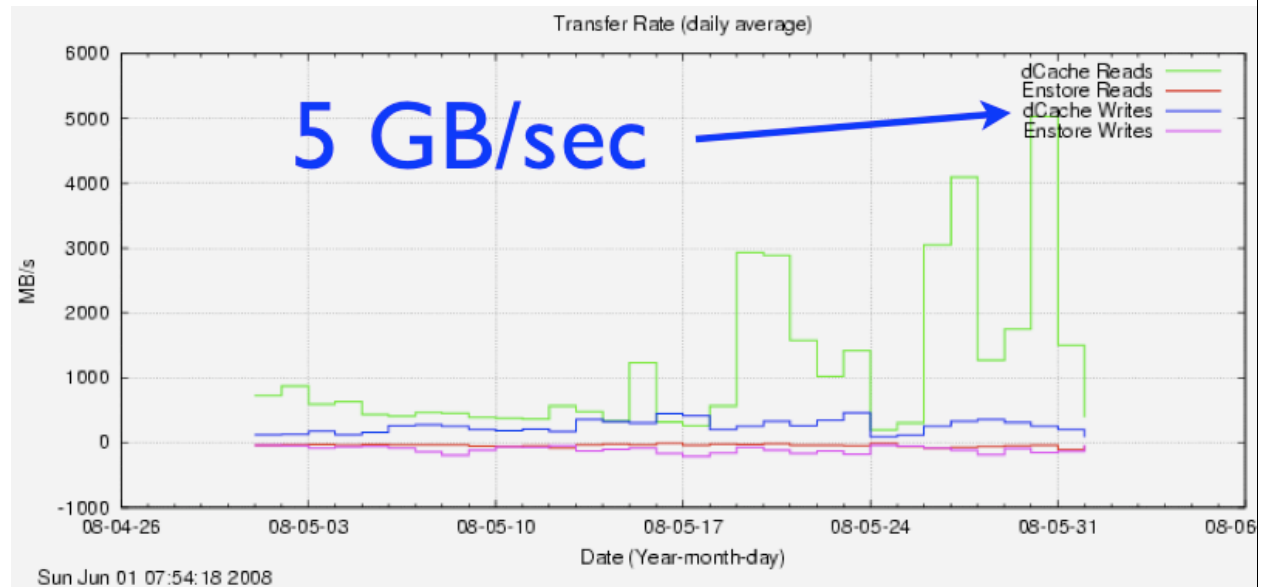
Jobs

IO rates
averaged over hour:

(X scale is 1 week)

averaged over day:
(X scale is 1 month)

Note - during the
infamous readahead
misconfiguration

10 GB/sec

5 GB/sec

File Families

USCMS tried a very fine-grained file family distribution this time.

- Directories were set up ahead of time, with correct file family and file family widths (number of active tapes).

- Permissions were set such that transfers could only write to these directories, and would fail on auto-creating directories.

Observations:

- During data taking - expect distribution of data files in the ratio that the file family widths were set up. But during test, all data from 1 file family was delivered at once - this caused tape backlog.
    FNAL can have a very large tape backlog and still be working properly!

- Target directories would change - for various reasons. This caused transfer failures due to permission errors on writing. We relaxed the restrictions and auto-created directories were written to catch-all file family. Happened several times.

We probably want to have courser grained file families next time.

Working on method for automatically specifying family during tape writing stage.

Phedex

We had trouble with agents stopping in the month.

We had 2 instances of phedex set up at FNAL - debug and production. During the month, the debug instance just stopped working. It was worked on for a few days, by several experts, but it never started working again. The reason is still unknown - we stopped trying to fix it.

- Debug instance was moved to production node, and it worked fine. We are now combining our debug/production scripts into a single instance with arguments instead of having 2 separate files.

PNFS access - we had previously removed all pnfs accesses in phedex since this caused much pnfs stress. We query our enstore databases whenever possible for the information we previously got from PNFS.

- Please do not add back in any pnfs accesses to the phedex scripts.

Transfers

We had a period when we were successfully transferring data, but at the end of the transfer, the transfers were all rejected because of CRC errors.

- Investigated - there were not any CRC errors. CERN wasn't recognizing request for CRC. This should not have been an error, but somehow it was.

  - Disabled all CRC checks - everything worked fine again

  - dCache team investigated - said it was impossible, but a week later has come back with a possible explanation. Working on code rewrite.

Castor upgrade
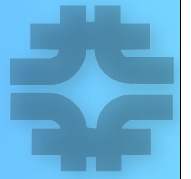
After the castor upgrade, we never got the rate we expected out of CERN.

- We twiddled many knobs at many levels, but we seemed to be source rate limited.

- Investigation is underway.

Monitoring

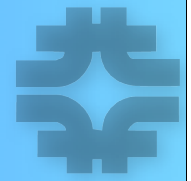The data transfer system is complex.

Excellent work on data monitoring - but - my opinion is that the info is mostly beneficial for management level presentations and not earnest debugging attempts.   (I mean this in a positive way.)

When trying to see what is going on after a parameter change, I think you need:

- clear button - reset stats to get affect of change, not old data

- number of files in flight at a site, number queued, number finished

- aggregate rate to the site, individual transfer rates

- success/attempt ratio

- And you need this all over a time period of 5 min, 10 min and 15 min.

- Finally, a list of actual globus-url-copy commands that are active, and are queued needs to be available to make progress in debugging transfer issues.

More comments

Paul Rossman has these comments, not in any particular order:

- Provide instructions whom to notify in case of irregularities

- We are now using swatch to monitor PhEDEx logs and page when necessary

- Problems with our custom PhEDEx scripts at FNAL after upgrade to 3.0 - these scripts needed code updates in order to work with new calls & libraries

- Lots of configuration options for new Phedex FTS download agent. what's the recommended config? lots of advice/emails with different parameters.

- All T1s have different restrictions on their FTS channel out of CERN - how does this affect FTS and Phedex

- Do we use CERN FTS server or our own?

- What is the status of the SL4 version of FTS?

- Do we use multiple phedex agents? We finally decided to increase our agent count to include a CERN agent and for a short period of time, we had a T2-US specific agent

- We had a few instances where we had to correct directory permissions in order for transfers to continue (see earlier slide). Some of this was caused by old SRM bug.