USCMS Plans for SC2

Jon Bakken
FNAL

SC1 Summary

Many independent specific feature tests, such as network rate using iperf servers or disk rates using filesystem exercisers, have allowed CMS to make data system component choices.

What was previously missing was an integrated system test where the robustness of the data transfers was the primary goal.

- SC1 Goal was to find and fix bugs, find optimal transfer tuning parameters, and thereby provide robust data transfers within the existing production environment and framework at the USCMS Tier 1 site at Fermilab.
- Secondary goal was to sustain transfers at a high rate, up to 500 MB/sec, for an extended period.

SC1 Summary

SC1 was an extremely valuable system integration exercise

It showed that the currently deployed system has a high degree of usability.

- SC1 demonstrated a 10x higher throughput (25 TB/day WAN) than prior use in fairly realistic deployment [c.f. CDF LAN rate is ~30 TB/day]

Many problems exposed due to high number of transfers and high rate. Problems were fixed, or new features added or system redesigned, before proceeding with tests

Rate results:

- Using a fully integrated system, rate - 300 MB/sec, CERN to FNAL, disk to disk
- Using dcache-java-class gridftp script - 500 MB/sec, no disk at FNAL
- Using dcache-java-class gridftp script - 400 MB/sec, to dCache disks
- 20 parallel streams per transfer, each with 2 MB buffers is optimal tune

SC2 Goals

Continue the goal of making transfers robust and reliable

- Understand and fix problems rather than just restart

Only use SRM managed transfers to FNAL dCache pools

- No Point-to-point contrived gridftp transfers or scripts
- No gridftp transfers to gridftp doors, only 3rd party transfers to dCache pools (might turning off gridftp doors to Tier 1)
- 20 streams per transfers, 2 MB buffers, standard TCP stack

Continue to use the CMS Production environment

- Real data transfers have to coexist with users, service challenge should do so as well. Solutions deployed right away.
- CMS has scheduled downtimes on Thursday mornings to update or install new services

Tier1 Prod Environ

dCache Pools - total of 54 TB deployed

- 26 TB Read Pools on 11 nodes (oldest unused data flushed)
- 12 TB Write Pools on 6 nodes (can read from these, too)
- 3 TB Volatile Pools on 17 nodes (no lifetime guarantee)
- 13 TB Resilient Pools on 174 nodes (copies = 5)

Enstore MSS

- CMS has first priority on 8 drives, and shares 3 more
- 162 TB on tape, around 200 blanks available now
- Powderhorn 9310 Silo

SC2 Goals

Rate Goals

- Transfer user requested data from Castorgrid first
  - Expect to get 10-15 MB/sec of user data from Castorgrid
  - 10 MB/s is about 4.2 9940B tapes/day
  - Expect to use PhEDEx for these user data transfers,
    - PhEDEx has been shown to work for this many times
- Transfer fake data from oplapro nodes - what is available?
  - 40 MB/s is about 16.8 9940B tapes/day. Will write this data to tape to exercise full system. Recycle tapes after one day.
  - Just released version of PhEDEx allows one to specify test files on the oplapro disks to be known to the database. This allows us to use PhEDEx to drive the fake data transfers, too. Need list of files + node names, if different than SC1

SC2 Goals

PhEDEx will therefore be used as the driving agent for both user-requested data from Castor and fake data from oplapro nodes

- Bug fixes or development in PhEDEx may be necessary and are part of the USCMS SC2 goals.
- Potential PhEDEx bottlenecks, previously present, due to pool filecatalog accesses may be present today, these will need to be addressed during SC2. Exercising PhEDEx is a goal of SC2

We are willing to participate in short bursts of 500 MB/s SRM managed transfers to CMS's resilient or volatile pools.

SC2 Goals

USCMS Tier 1 to Tier 2 Complement of SC2

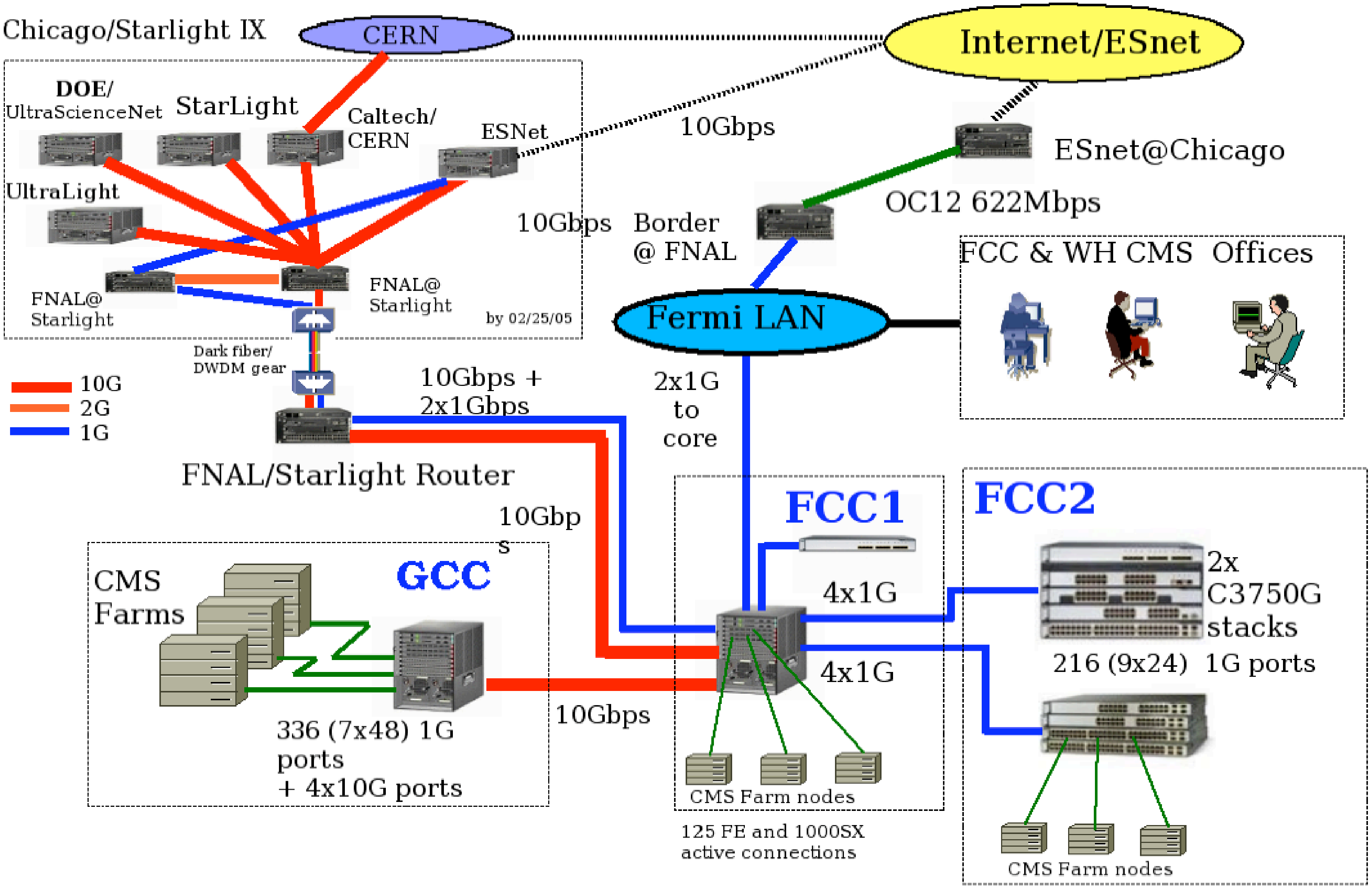
- US CMS Tier 2 sites want some of the data we are transferring. I plan on delivering these data sets to them as part of this challenge
- Initially expect low rate to each, to 3-4 Tier 2 sites
  - End-to-end functionality test of many components
  - Rate can grow as Tier 2 can accept data
- Will use PhEDEx to drive these transfers as well
  - Coordinating with Tier 2 sites now and getting PhEDEx daemons working at each site. Each site has resilient dCache + SRM deployed, or is deploying it now
  - Underlying mechanism is srm managed transfers to dCache pools using gridftp protocol

Other USCMS Activities

FNAL and Caltech are investigating Terabyte transactions

- Move a few TB in an hour
- This will be a demonstration of LambdaStation integrated with SRM/dCache project
- This work is concurrent with SC2 but not part of it

CMS Network @ Fermilab (02/23/2005 AB)



Answers to Questions

What is your data transfer cluster?

- It is the CMS production system: 54 TB total, 12 TB in write pools, 26 TB in read pools, 3 TB volatile pools and 13 TB in resilient pools. We have 8 cms-owned 9940B drives, and access to 3 more in a powderhorn silo.
  - We have ordered 26 TB more dCache pools, and expect to be deploying them in about month or two. Another 26 TB will be ordered late spring.
  - Adding 300 worker nodes, each with 250 GB disk ~early summer. This gives another 75 TB of resilient Pools.
  -

What is your network, how much bandwidth can be used when?

- FNAL shares a 2x10 GE, 1 GE and 622 MB link. The 10 GE and 1 GE links will be used for these tests.

Answers to Questions

What is your software: OS, kernel version, globus version?

- OS is LTS303 (Redhat SL3) with a few 7.31 nodes
- All have Redhat 2.4.21 kernels with security patches
- Java 1.5
- dCache and SRM code are 'head' of repository, I cut a new versions daily and deploy it, after testing, every Thursday.
- SRMCP is v1.11 Available from FNAL KITS.
- We do not use globus servers for data transfers.
- The dCache + SRM use these modules (all part of DESY CVS)
    - GLUE from "The Mind Electric", version 3.2.
    - Axis 1.2 Release candidate 2
    - Java cog kit from Globus, Head of jglobus cvs repository on 12/9/4

Answers to Questions

When do you need to be alone for perf testing ?

- I expect CERN and the Tier2s to be involved with the transfers.
- PhEDEx transfers scripts and databases being developed now
- Gridftp transfers optimized at 20 streams each with 2 MB buffers gives best aggregate throughput
- On track, so I am not sure what standalone performance testing is needed for FNAL right now.

Answers to Questions

What software will be tested and when ?

- End-to-end data srm managed transfers between CERN and FNAL and US Tier2s driven by PhEDEx. The underlying protocol is 3rd party gridftp to the dCache pool nodes.

What is your SRM implementation / timeline ?

- We use the SRM that FNAL wrote in collaboration with DESY. The SRM code is 'head of the repository', updated weekly. SRMCP is version v1.1. It is working now.

What are your performance milestones

- Sustained 50 MB/s to tape, 10 MB/s from Castorgrid,
- Sustained 2 MB/s to each participating Tier 2 (3-4 sites)