



# CMS Service Challenge Status

Ian Fisk  
June 27, 2006



# Operational Goals

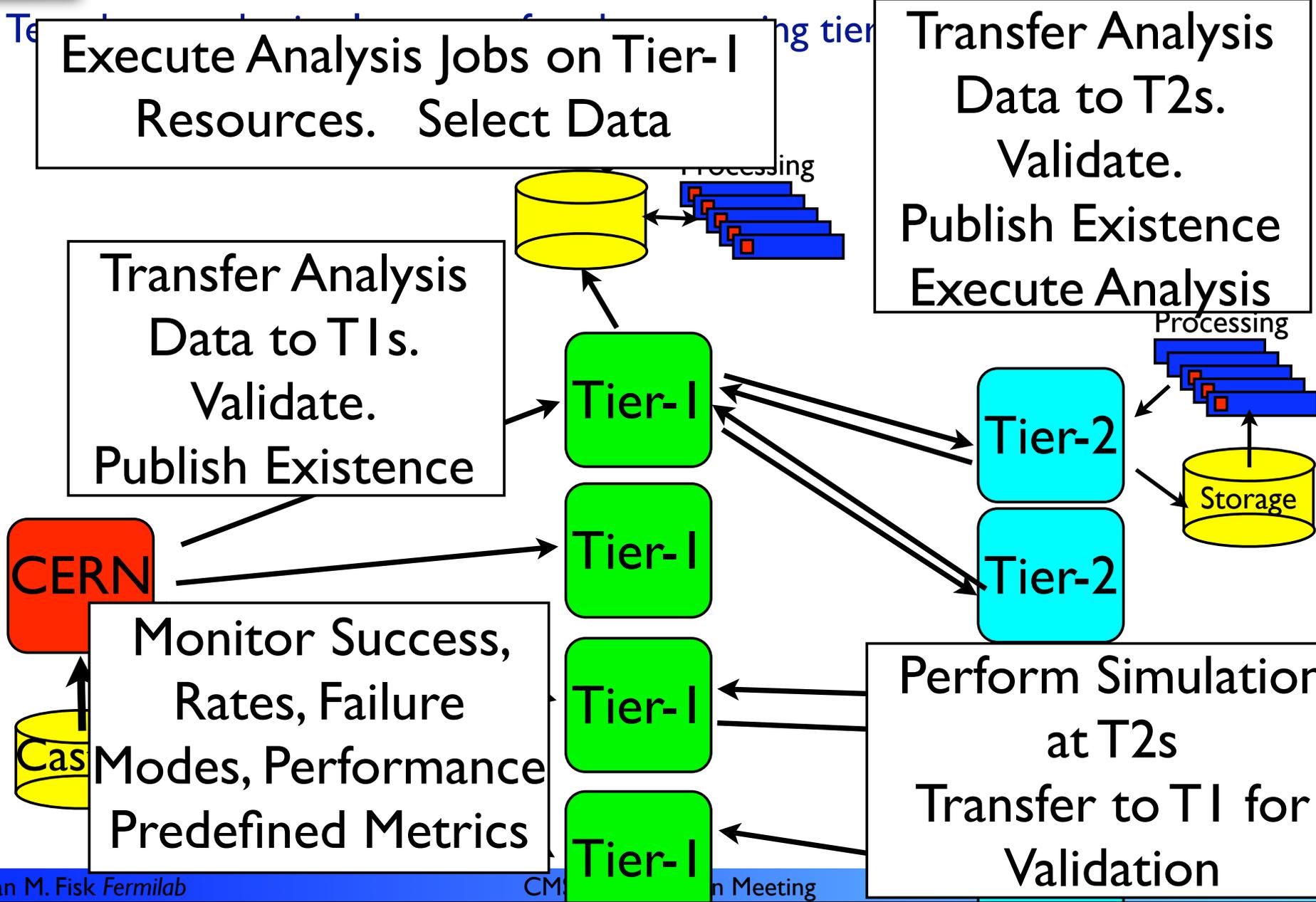
CMS needs to be at production scale services in 2008

- ➔ Assuming we cannot easily more than double the scale each year, we should be able to demonstrate 25% of the expected 2008 scale in this year and be able to reach 50% scale early in 2007

| Service                         | 2008 Goal              | 2006 Goal     | %   |
|---------------------------------|------------------------|---------------|-----|
| Network Transfers between T0-T1 | 600MB/s                | 150MB/s       | 25% |
| Network Transfers between T0-T1 | 50-500 MB/s            | 10-100 MB/s   | 20% |
| Job Submission to Tier-1s       | 50k jobs/d             | 12k jobs/d    | 25% |
| Job Submissions to Tier-2s      | 150k jobs/d            | 40k jobs/d    | 25% |
| MC Simulation                   | 1.5 $10^9$ events/year | 25M per month | 25% |



# SC4 Workflows





# Original Schedule Processing

Original schedule was to operate the first two weeks of June

- ➔ 25k jobs per day (50% analysis and 50% production)
  - Operate Job Robot on test simulation samples for analysis
  - Operate Prod\_Agent for production job
- ➔ 90% success rate to complete jobs
- ➔ We had a series of validation steps for the sites to meet
  - 25 Tier-2s signed up to participate
    - Pass the site functional tests, allow the CMS software to be installed, configure PhEDEx, download a test sample into the trivial file catalog namespace, demonstrate access with an analysis application
    - Demonstrate success with the production agent job
  - 19 Tier-2 sites and 5 Tier-1 sites completed the steps
- ➔ We spent a lot of the first two weeks commissioning sites and did not start large scale operations until last week



# Original Schedule Transfers

## Scaling Tape Rates by pledge aiming for 150MB/s

- ➔ ASGC: 10MB/s to tape
- ➔ CNAF: 25MB/s to tape
- ➔ FNAL: 50MB/s to tape
- ➔ GridKa: 20MB/s to tape
- ➔ IN2P3: 25MB/s to tape
- ➔ PIC: 20MB/s to tape
- ➔ RAL: 10MB/s to tape

## Networking provisioning should be at least twice this

- ➔ These goals are sufficiently modest that no center should struggle to sustain them



# Tier-2 Transfers

## Network Estimates for Tier-2 vary widely

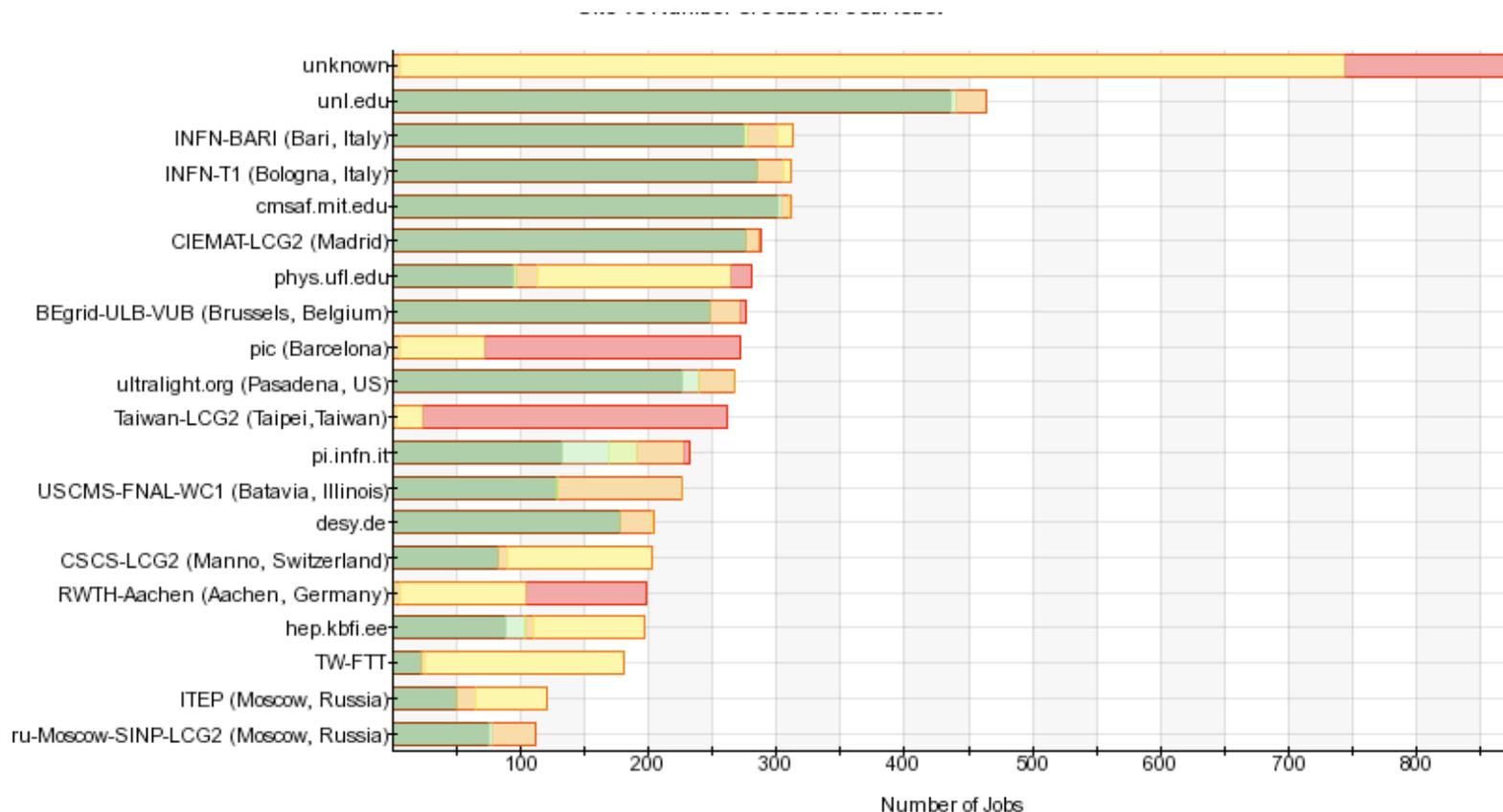
- ➔ The computing model defines the expected minimum in 2008 at 1 Gb/s
  - Naively taking 25% this would be 250Mb/s
- ➔ Given the number of Tier-2 centers already at 1 Gb/s to 10Gb/s and the difficulty using reasonable scale networking end-to-end it makes sense to try much larger scale tests at some Tier-2 centers
  - Try to sustain ingest rate to Tier-1 centers from all Tier-2s
  - Drive Tier-1 to Tier-2 rates at 10MB/s to 100MB/s



# Processing Status

Analysis processing is going OK, though we still have about a factor of 2 in scaling to meeting and some stability issues

- ➔ We have hit about 10k analysis submissions in a day
- ➔ We have hit 10k jobs a day in the production system on the OSG
- Lower on the LCG





# Worker Node configuration issues

We see issues where a few misconfigured worker nodes can significantly reduce the efficiency of a site

- ➔ Those nodes are preferentially available

We kill individual user jobs and need to resubmit them

- ➔ We can see 50% loss on sites with 1% badly configured nodes

This is an interesting use-case for pilot jobs

- ➔ Pilots determine the configuration and don't request work flows unless they are configured.
- ➔ We are finding even locally that we can have configuration issues that don't affect all VO's or only affect grid submissions
  - Diagnosis and debugging are challenging



# Transfer Status

## Transfers proceeding poorly

- ➔ Data rate out of CERN has struggled to reach 100MB/s and then with substantial numbers of errors and large structure
  - Periods with 10MB/s-20MB/s
  - Many errors
- ➔ Rate from Tier-1s to Tier-2s are not much better
  - Highest export rate is from FNAL that has not switched the Tier-2s to FTS channels

## We have a couple exceptions

- ➔ A number of Tier-2 have sustained 10MB/s from at least one Tier-1
- ➔ All the US Tier-2s have sustained 50MB/s for a 24 hour period from one T1
- ➔ 2 Tier-2 has sustained 100MB/s for a 24 hour period



# FNAL Transfers

All the transfer details are

- ➔ <http://pcardabg.cern.ch:8080/dashboard/phedex/>

CERN to FNAL is nearly meeting the modest 50MB/s goal, but there is a lot of structure

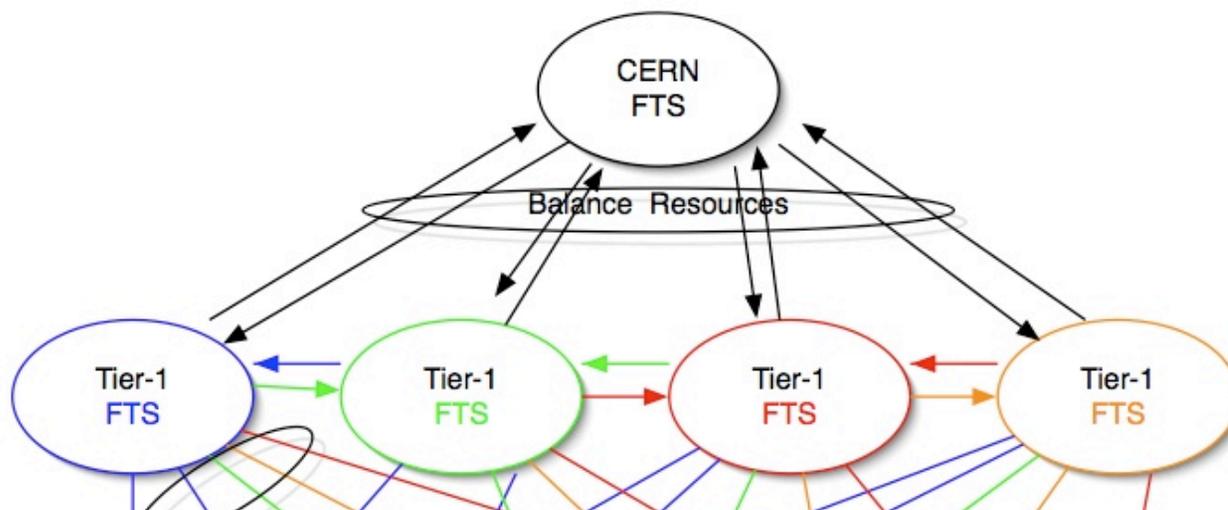
- ➔ We are increasing the size of the transfer file pool
- ➔ Investigating ways to increase the single stream performance to FNAL
- ➔ Should be able to do 4 times this without breathing hard

FNAL to Tier-2s

- ➔ Export rate is the best of any Tier-1,
  - Both in terms of total rate and number of centers connected
  - Trying to get FTS channels functional and increase the number of connected sites

File Transfer Service is a service implemented by EGEE

- ➔ It sits on top of srm (srmcp and get and put are possible)
- ➔ Database that tracks transfers
  - It's purpose is to act as a throttle
    - Defined VO share of the transfers between two end points
    - Defined maximum average bandwidth
  - There are some stability issues with the DB, requests fill and tables need to be cleaned out
    - Rapid development





# Preparation Activities

## Increasing the processing scale and including users

- ➔ Need to hit 50k jobs per day by CSA06

## Continuing to exercise the transfers

- ➔ Get Tier-0 to Tier-1 transfers under control and begin to have more successful Tier-1 to Tier-2

## Incorporating the calibration information into the workflow

- ➔ Deploy the LCG-3D infrastructure with Frontier for SQUID caches

## Improving the functionality and reliability of sites

- ➔ All participating sites should be able to complete the CMS workflow and metrics

## Prepare CSA06 Production Samples

- ➔ 25M Events per month in July and August



# Increasing the Processing

## Two improvements from SC4

- ➔ To grow from 25k jobs per day to 50k jobs per day we need to switch submission infrastructure
  - 25k jobs per day is already a strain on the resource broker infrastructure we have
- ➔ Job robots are good load generators, but they do not make mistakes and they are patient (flat load over a 24 hour period)
  - Users generate unexpected usage patterns and loads

## The switch to the gLite RB with bulk submission is needed for CSA06

- ➔ Deployment came later than we expected in May
  - Still not fully commissioned in CMS for SC4
  - We anticipate the latter portion of SC4 and continuing throughput the summer to work with the gLite submission infrastructure on LCG
  - Continue to improve the submission rate on OSG



# Improving Functionality and Reliability of Sites

Commissioning sites was a fair bit of work, but we are succeeding

- ➔ There are a number of new services and new sites

We have conducted and will continue open technical support sessions

- ➔ Opportunities for sites to call in and receive help with individual services

We have encouraged status reports

- ➔ A regular schedule of site reports
- ➔ Directed reports from Tier-I centers on SC4 progress

We will interact with sites through the LCG-MB

We are exercising the channels available to us, but there are still issues with site preparation and reliability

- ➔ The majority of sites are responsive, but there is a lot of work for this summer



# CSA06 Preparations

## The sample is 50M Events

- ➔ As a rough assumption let us assume 3 minutes per event on average
  - Based on the old GEANT4 production
- ➔  $50\text{M} \times 3 \text{ minutes} = 150\text{M} \text{ minutes}$
- ➔ July and August = 86k minutes
- ➔ If we were 100% efficient we would need 1750 CPUs
  - Assuming something more reasonable like 75% we need ~2400
- ➔ Given these assumptions, most of the non-CERN resources needed for CSA06 itself will also be for pre-challenge production
- ➔ Requirements go down if CERN resources are used for simulation as well
  - Hoping to reserve CERN resources for Tier-0 preparations



# Service Challenge Plans

From the standpoint of CMS computing, Service Challenge 4 is a dress rehearsal for CSA06

- ➔ A number of the elements in SC4 are lower scale exercises we expect to perform in CSA06
- ➔ There primary technical differences are
  - We add Tier-0 reconstruction (outside the scope of today)
  - We anticipate include user analysis jobs into Tier-1 and Tier-2 processing
    - Currently SC4 is entirely exercised with robots
    - Job submission rates for CSA06 are 50k jobs per day (double SC4)
  - Calibration and Alignment infrastructure plays a larger role
- ➔ SC4 is an integration activity we will continue running for as long as it continues to be productive
  - CSA06 is a data challenge with metrics that need to be met in a specified time



# Rough Schedule

## Now until the end of June

- ➔ Continue to try to improve transfer efficiency
- ➔ Attempt to hit 25k jobs per day and increase the number and reliability of sites performing 90% efficiency for job completion

## July

- ➔ Demonstrate CMS analysis submitter in bulk mode with the gLite RB

## July and August

- ➔ 25M events per month with the production systems

## Second half of July participate in multi-experiment FTS Tier-0 to Tier-1 transfers

- ➔ Continue through August with transfers

## Improve Tier-1 to Tier-2 transfers and the reliability of the FTS channels.