



# US-CMS Storage Plans

May 17, 2005



# Tier-I Requirements



CMS has various requirements for storage depending on the application and the utilization of the facility

- ➔ The Tier-I offers custodial storage of CMS raw data. This involves tape
  - Tier-I Storage Solutions currently must support tape back-ends
  - Accepted and reliable wide area interface to the storage system
  - Input rate to tape is expect to be 70-100MB/s on average during running periods
    - Need head room to catch up from failures
- ➔ Tier-I centers are expected to serve data to analysis applications at a rate of 800MB/s for a nominal Tier-I, probably 1600MB/s for the US-CMS Tier-I
  - Data read by reprocessing applications
  - User analysis applications (bulk of serving is analysis)
- ➔ Expect about 2pB of disk storage



# Tier-2 Requirements



Tier-2s have ~200TB of disk storage

- ➔ No custodial storage
- ➔ Tier-2 storage is true cache

Read rate should be approximately 20% of Tier-1 rate

- ➔ ~200MB/s

Ingest rate should be able to keep up with a 2.5-10Gb/s WAN link

- ➔ 25-100MB/s

Interface needs to be standard, reliable, and with reasonable performance

Currently the assumption is the experiment will not track official CMS data on the file level

- ➔ Deal with blocks of data and complete datasets

For transfers between Tier-1 and Tier-2s resource reservation unnecessary

- ➔ Data Placement Service and allocated block of CMS Tier-2 storage

For user output and simulation output some ability to ensure output has a place to go



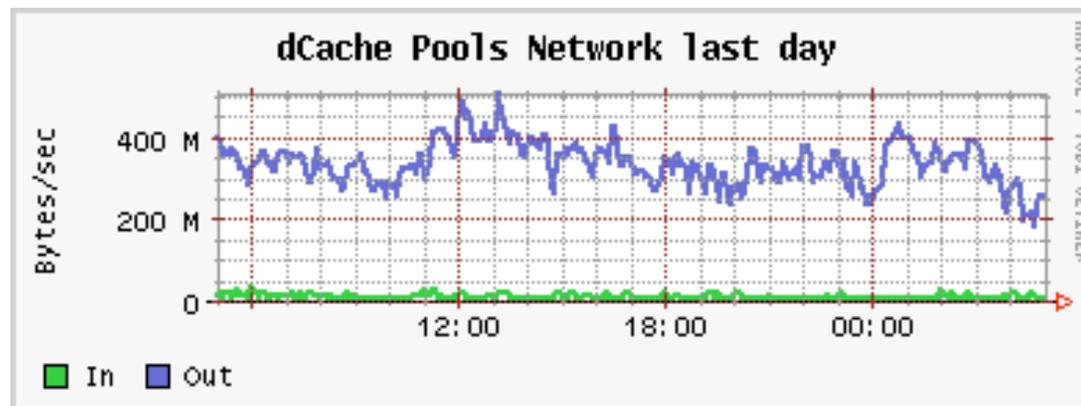
Primary Usage of Opportunistic computing is likely to remain simulation requests or specialized analysis applications that have large CPU requirements and reasonably small input samples

- ➔ Some form of resource reservation for simulation would end a current common failure for OSG
  - Currently directly transferring from the worker nodes would also solve the problem.
- ➔ For analysis applications the ability to reserve space for some period of time would facilitate utilization of resources with significant input datasets



US-CMS in collaboration with CCF is in the process of deploying and supporting SRM/dCache/Enstore at the Tier-I center

- ➔ Data Serving and ingestion rates are good
  - Demonstrated extended periods of 400MB/s of serving
  - Met SC2 goals for 50MB/s of ingest (limited by drives)



SRM performed well during service challenge

- ➔ Problems were found and fixed



## PNFS is a potential weak point in the system

- ➔ We have been learning how to mistreat the PNFS database
  - Some of the problems may be reduced by increasing the CMS file size and decreasing the number of entries
- ➔ For the final system we may need to partition the storage

## Operations load is significant

- ➔ We enjoy very good support from CCF, but it still involves hard work from both groups
- ➔ We expect the increase in scale to require development as we ramp up

## SRM works but we are struggling with adoption

- ➔ Seems to be contention in the user community it's hard to get running
- ➔ Aside from CERN and FNAL we don't have successful grid implementation
- ➔ Concern in grid community that we are trying to push too much into the protocol



# US-CMS Tier-2 Technology Choices



We are in the process of installing resilient dCache at the Tier-2s

- ➔ Of the existing prototypes we have 75% fully functional.
  - Working on the last one
- ➔ Of the new sites we have one basically working and one just starting.
- ➔ At all sites we are working on optimizing the SRM performance and experience with data serving operations

Baseline technology choice is dCache

- ➔ We will probably ask that at least one Tier-2 install and configure xrootd as well
  - Should be comparing performance, operational stability and operational difficulty
- ➔ Without standard wide area interface it doesn't meet the storage requirements
  - Might be useful for specialized applications and good to compare
  - It is not obviously a technical comparison that needs to be done here.



We need a viable model for supporting Tier-2 dCache operations

- ➔ Currently the model is to develop enough expertise to perform support functions locally on site and within US-CMS

Operational Support of dCache is labor intensive

- ➔ Partially inexperience and partially true challenge
- ➔ Installations generally proceed reasonably well. The questions usually come from operations issues

Optimizing the system is difficult

- ➔ A lot of effort went into getting the current local performance
- ➔ Jon recently got a factor of 150 improvement in performance between FNAL and UCSD
  - Non-trivial to find



# OSG Technology Choices



## Lots of possibilities in OSG

### ➔ DRM

- Not ready for wide deployment but SRM implementation on top of filesystem would improve reliability of current GridFTP door
- No built in virtualization

### ➔ dCache

- Need to understand how we constrain the local support load
- ATLAS Tier-2s are examining

### ➔ NEST

- Currently has GridFTP door, but does provide space reservation
- Does not appear to have a virtualization layer

### ➔ xrootd

- Good for serving data
- No SRM interface.



# OSG Storage Concerns



Open Science Grid is self organized and sites can choose what services they bring up.

- ➔ None of the production on the previous page is perfect and the optimization of what to choose will have difference answers depending the sites and the chooser
- ➔ CMS applications will have to be configured to make use of what they find.