



CMS Analysis Use- Cases at Tier-2 Centers

Ian Fisk
Tier-2 Workshop
June 13, 2006



Introduction



Tier-2 centers are the largest resource for user specified analysis

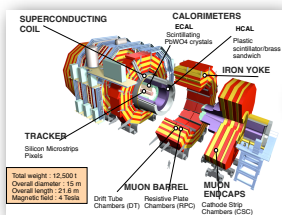
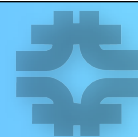
- ➔ Tier-1 centers have processing capacity, but it is largely specified
 - Roughly half the resources are intended for event reprocessing
 - The other half is intended for skimming and data selection, but the usage will need to be structured.
 - CPU intensive processing will be performed at analysis computing at Tier-2 centers

Data placement drives the usage of computing resources

- ➔ Jobs are sent to locations where the data exist
- ➔ At some point in the future CMS may have automation to balance the access load
 - At the start the data placement and replication will be governed by policy set by the experiment

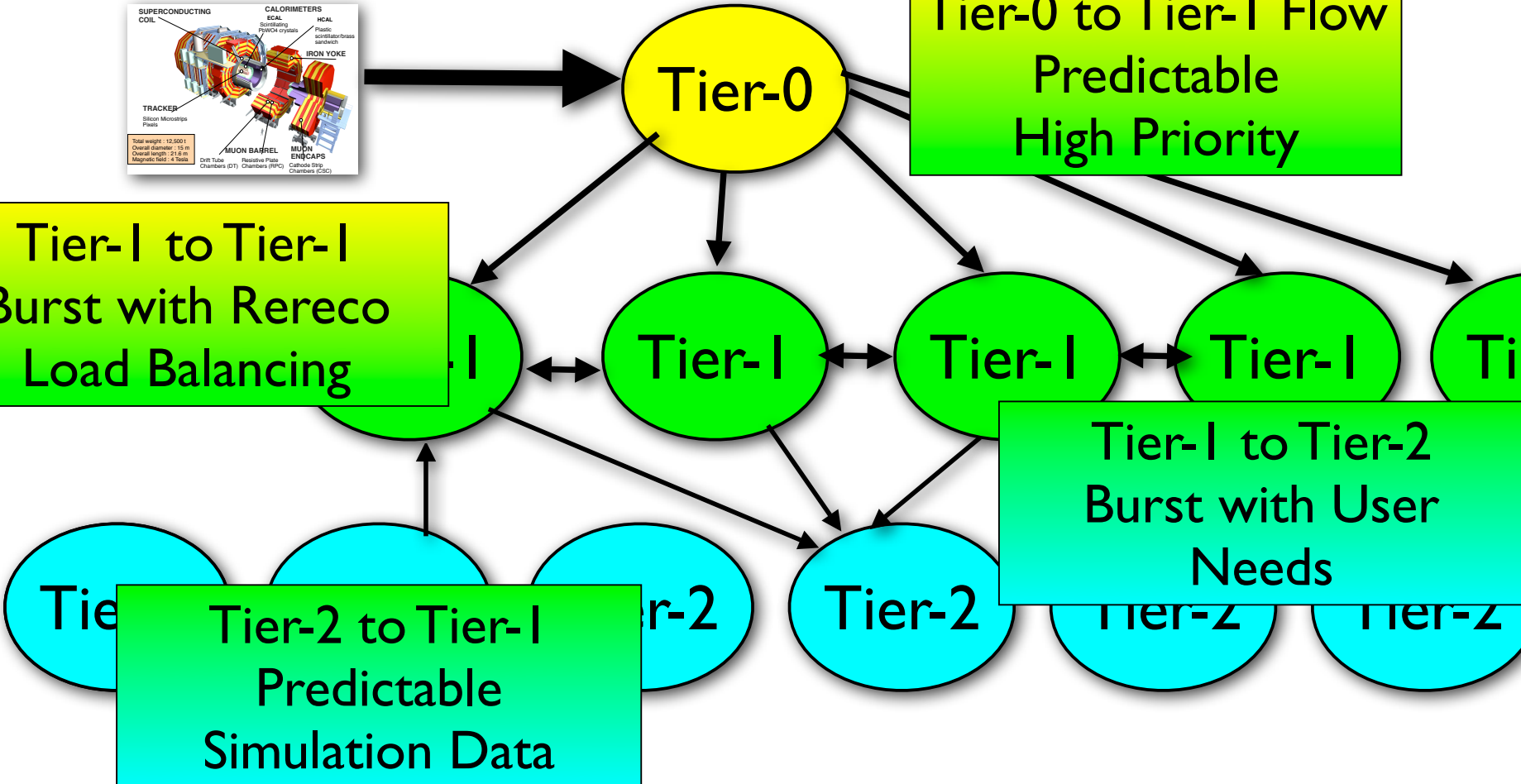
Data is entrusted to Tier-1 centers to archive and to serve

- ➔ Each Tier-2 center has the possibility to connect to any Tier-1 center to access data. It is not the purely hierarchical MONARC model



Tier-0 to Tier-1 Flow
Predictable
High Priority

Tier-1 to Tier-1
Burst with Rereco
Load Balancing



Tier-1 to Tier-2
Burst with User
Needs

Tier-2 to Tier-1
Predictable
Simulation Data

Tier-2 centers may have relationships with Tier-1 centers for management, support, and operations

➔ Data access may come from a variety of Tier-1 centers



Surviving the first years



The computing for CMS is hardest as the detector is being understood

- ➔ The analysis object data for CMS is estimated at 0.05MB
 - An entire year's data and reconstruction are only 300TB
- ➔ Data is divided into ~10 trigger streams and ~50 offline streams
 - A physics analysis should rely on 1 trigger stream
 - A Tier-2 could potentially maintain all the analysis objects for the majority of the analysis streams

Unfortunately, until the detector and reconstruction are completely understood the AOD is not useful for most analysis and access to the raw data will be more frequent

- ➔ The full raw data is 35 times bigger
- ➔ Given the CDF experience, we should expect about 3 years to stabilize
- ➔ People working at Tier2 centers can make substantial, but bursty requirements of the data transfers



Analysis Selections



When going back to the raw data, analysis selections on a complete trigger streams

- ➔ 1% selection would be 2TB, 10% selection would be 20TB
 - Smaller by factor of 5 if only the offline stream can be used
- ➔ There are an estimated 40 people working at a Tier-2
 - If people perform these sort of selections at the level of once a week
 - Rapidly get to the 2.5Gb/s - 10Gb/s in the initial Tier-2 estimates

Size of selections, number of active people and frequency of selections all have significant impact on the total network requirements

Network estimates for the Tier-2 centers were made with the expectation of treating the disk systems as a dynamic cache

- ➔ With 200TB of disk how long would be reasonable to wait to flush the cache
 - 2.5Gbs/s is week 10Gb/s is about 2 days



CMS Components Involved in Analysis



Data Management

- ➔ How does data get to a site?

Data Discovery

- ➔ How does a user or application determine what data is desired?

Data Location

- ➔ Where is the desired data?

Job Specification

- ➔ Configuring the jobs to run

Job Submission

Job Execution

Job Monitoring

Returning results



Data Management



CMS System for Data Management involves the following components

- ➔ Dataset Bookkeeping Service
 - Answers the question what data exists
 - Central service of the experiment
- ➔ Dataset Location Service
 - Identifies data location
 - Central service of the experiment
- ➔ Data Placement Service (PhEDEx)
 - Moves datasets between sites
 - Can monitor the progress of data to and from tape
 - Checks data completeness and data integrity
- ➔ Local File Catalog
 - Tool to resolve the logical file name, which is tracked by CMS, to the physical file name, which is tracked by the site.



The DBS and DLS are centrally run experiment services, PhEDEx and the local file catalog are issues for the sites including Tier-2 sites

➔ PhEDEx uses a database called the transfer management database to communicate between agents

- Agents at sites

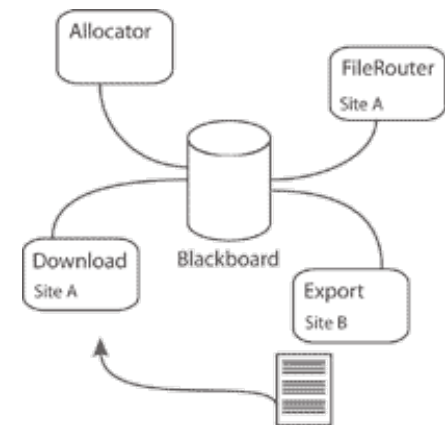
- Migrate files to mass storage,
- manage local mass storage stager pool
- stage in files efficiently based on transfer demand,
- calculate file checksums when necessary before transfers.

- Management Agents

- in particular the allocator agent, assigns files to destinations based on site data subscriptions, and agents to maintains file transfer topology routing information.

➔ PhEDEx transfer agents can call FTS to initiate transfers or use srmcp or even gridFTP directly

- It is the layer above individual file transfers and deals with replication datasets and managing the locations of the experiment data





Local File Catalog



The second service at every site is the local file catalog

How using the Logical File name does a a service resolve the physical file location?

- ➔ Variety of ways to do this
- ➔ CMS is working with an implementation called the trivial file catalog
 - Basic idea is that all you don't need an additional catalog to resolve logical to physical file names
 - Trivial File Catalog resolves using a consistent name space
 - Also could implement the use of the underlying storage catalog to resolve the physical file names
- ➔ The implementation CMS is using has a site local configuration file
 - discoverable from a consistent path
 - File gives the head in the directory structure
 - Logical file names look like paths beneath the head
 - `dcap:/pnfs/cms/WAX/11 /store/preprod/2006/05/15/mysamples/filenames.root`

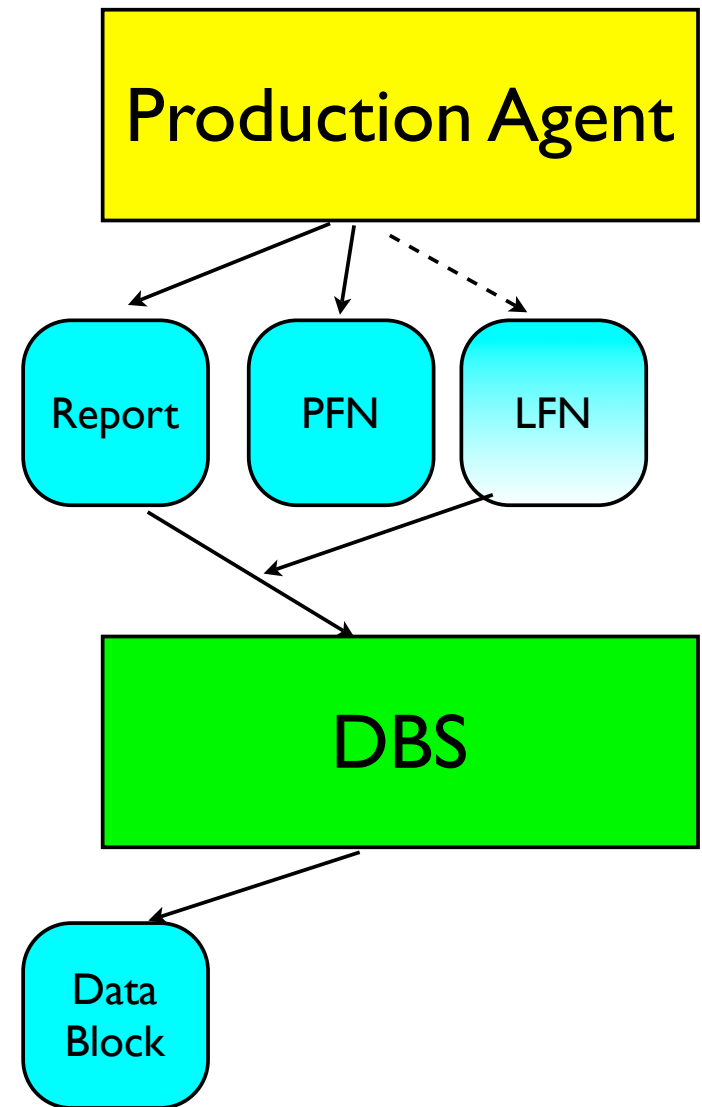


Simulated Data is produced by the Production_Agent

- ➔ A new dataset is registered into the Dataset bookkeeping service
- ➔ The output of individual jobs is reported by the Framework Job Report which is used to enter the file entries into the DBS

The DBS registers the files as a file block

- ➔ A file block is the smallest item the CMS data management system will manage
 - Data blocks can be open to enable subscription to increasing datasets
 - Or closed which are then managed as logical units
 - Simplifies the cataloging

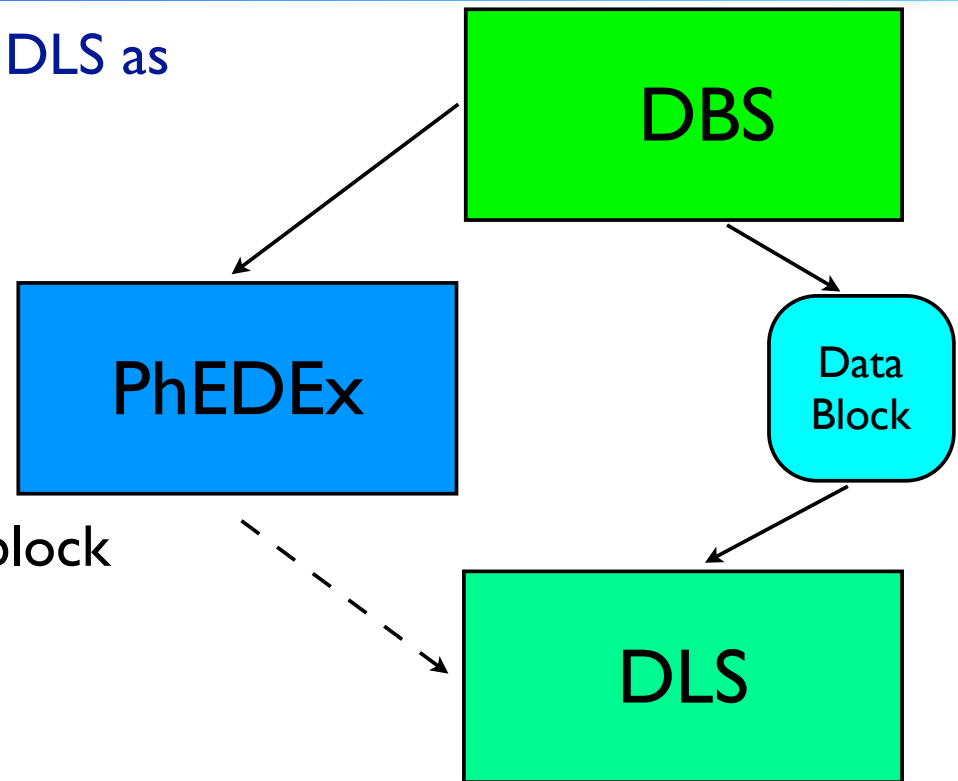




The newly formblock is registered in DLS as existing at a site

PhEDEx uses DBS to determine the files associated with a data set

- ➔ Transfers the files between sites
- ➔ Updates the DLS with the new block replica





A user can query the DBS to determine dataset

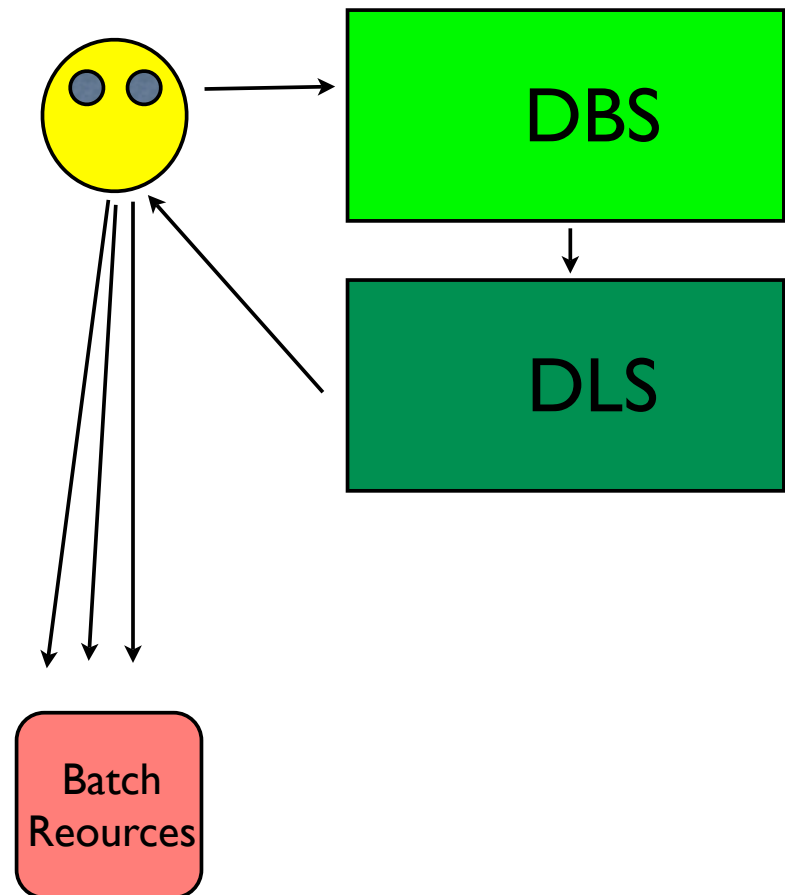
- Current query capabilities are fairly primitive, but will improve.

The identified dataset is defined by a number of data blocks

- ➔ Block location is provided by DLS
- ➔ Job can be sent to any site with the published set of blocks

A File list from DBS allows job splitting

- ➔ Logical file name splitting
- ➔ Trivial file catalog to resolve entries





Analysis Specification and Submission



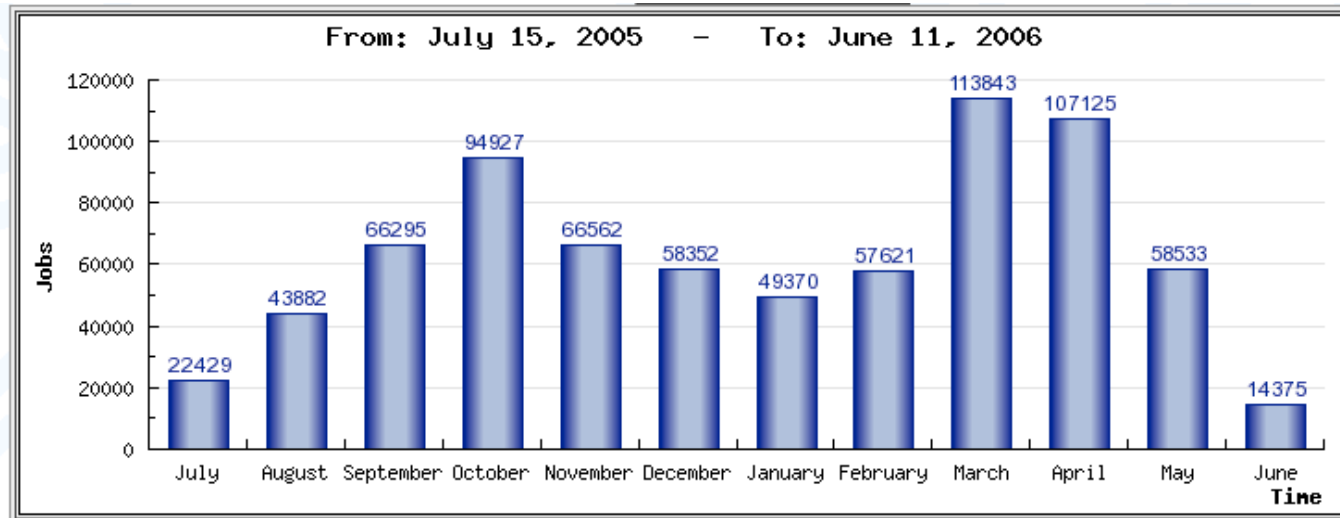
In July of 2005 CMS introduced the CMS Remote Analysis Builder (CRAB)

- ➔ A system in which a user could specify the data set desired, the application and input parameters to run, and the the number of events to process per job
- ➔ CRAB would handle the data discovery
 - With a query to DBS and DLS for
- ➔ The job preparation
 - Tarring up the user application and parameters, while making the appropriate number of jobs for the events needed to process
 - Assigning jobs to files as dictated by the entries in DBS
- ➔ Submitting the application
 - Submitting jobs through the LCG RB to sites where the appropriate software environments exist.
 - Submission location is identified by sites identified by block location in the DLS



CRAB submission has reached more than 100k jobs per month

- ➔ Tends to peak before Physics TDR submissions



While the majority of the access has so far has been to Tier-1 centers

- ➔ A number of Tier-2 centers have hosted data samples and accepted analysis jobs

There will be a more detailed presentation on CRAB in the CMS tutorial session on Thursday



CMS Dashboard



CMS analysis and service challenge jobs report to the ARDA developed Dashboard

➔ <http://arda-dashboard.cern.ch/cms/>

Allows to plot users, activities, and job success rates

➔ Very useful for diagnosis and debugging



ASAP

Arda Support for cms Analysis Process

User

Site

Job Submission Tool

Input Collection

Application

Activity

Grid Flavour

sort by

Start Date (d/m/y HH:MM:SS):

End Date (d/m/y HH:MM:SS):



Enabling Priorities



CMS Expects to support finer grained authorization at Tier-2 centers for roles and groups through VOMS

- ➔ Enable the experiment to control the prioritization on the fraction of resources pledged to CMS

For Analysis

- ➔ We anticipate 20 analysis groups
 - Each with one analysis coordinator role for common tasks
 - Pool of group members for analysis activities
- ➔ Storage authorization for analysis coordinators for common data and software
- ➔ User authorization for individual storage space.



Site Status



CMS is in the process of transition of Event Data Models and Software Framework

- ➔ We are commissioning sites with the new software, new data management infrastructure, and new trivial catalog logical file name resolution
- ➔ For Service Challenge 4
 - We currently have 18 of 25 Tier-1 sites validated
 - A small sample of the new data had been transferred down and configured into the trivial file catalog name space
 - The CMSSW software framework is installed and accessible from the worker nodes
 - A CRAB analysis job has been submitted through the RB to the site and run successfully against the test sample

Larger validation and commissioning data samples will be distributed shortly.



Outlook



CMS analysis activity including Tier-2 centers is ramping both in terms of number of users and total number of jobs

- ➔ We are still at a small fraction of the resources we expect to use in 2008
- ➔ The PTDR structure has been an interesting taste of the conference season experience where requests jump

Analysis resources at Tier-2 centers are a sizable fraction of the experiment as a whole, so our efficiency for using them needs to increase greatly over the next year

- ➔ Bringing up sites and services is challenging and sites are encouraged to join the CMS integration effort as early as they can.