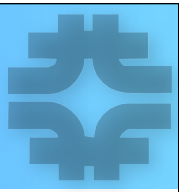


Resource Usage of CMS

Ian Fisk

March 20, 2007

CMS Data Driven Baseline

CMS expects the data location to drive the activities at sites

- Analysis Groups will submit selection jobs to Tier-1 centers to form analysis group datasets
  - New data samples can be replicated to Tier-2s
- Analysis Users will be submitted to Tier-2 centers from remote locations to resident data
- Portions of data can be transferred from Tier-1 and Tier-2 centers to local clusters for more detailed analysis

The replication of data between centers to improve efficiency or to utilize unused resources will be a deliberate act by an operator, at least at the beginning

- Activities are well specified

Tier-3

In the CMS model there are a lot of similarities between the Tier-2 and Tier-3 functionality

- Tier-3s do not have necessarily the same priority access to other centers for data transfer
  - But they have complete control of what they do
  - The number of active physicist supported at a Tier-3 center is potentially much smaller than a Tier-2
    - 4-8 people
    - This leads to smaller sustained network use
      - but similar requirements to T2s to enable similar turn-around times/latencies for physics datasets copied to T3 sites for analysis

CMS would like to have access to opportunistic cycles at the Tier-3 centers through the OSG interface

- A number of the normal CMS services have expectations of common grid infrastructure

Data Concepts

CMS data is divided into on-line and off-line streams

- Roughly 10 online trigger streams. Around 50 off-line analysis streams
- Analyses only require access to one stream and there can be overlap in the events in each stream

Data streams are hosted by at least one Tier-1 center, served to any Tier-2

- Raw data is expected to be 1.5MB/event
- 0.25MB of reconstruction
- 0.05MB of analysis object data
- Tier-1s also archive and host the relevant MC samples for the streams entrusted to them

CMS Data is written as files, Files are bound into logical quantities called blocks, blocks for datasets.

- Users typically worry about datasets, the transfer system deals with blocks, and an individual applications opens a file

Minimum Needed for Analysis at Tier-3

There is a big range in the capabilities and capacity of Tier-3 centers in CMS

- We will have Tier-3 centers which are shared university facilities which we hope will be made useful for opportunistic computing through OSG
- We will have Tier-3 centers that are desktop clusters used by local folks

At a minimum you need

- A user interface machine (UI) to submit a grid job. CMS remote analysis builder (CRAB) client
- A local CMS software installed, pretty trivially installed with apt
- In this model everything runs remotely and essentially the histograms are returned

If you have some storage and processing resources

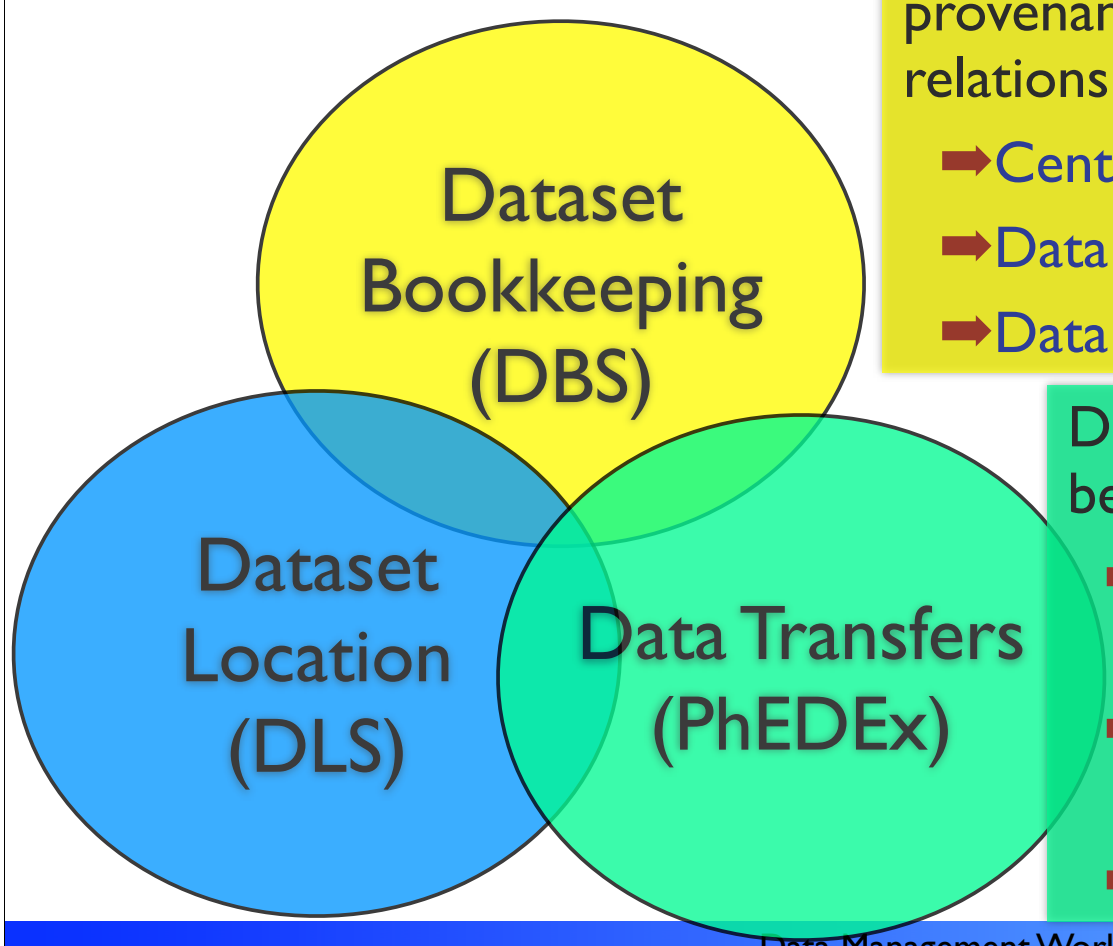
- Install the transfer system and bring datasets down and access locally

If you have a lot of resources

- Install OSG and make your resources available for opportunistic simulated event production

CMS Data Management Services

DLS Identifies the location of the blocks



Dataset bookkeeping tracks data provenance, meta data, and data relationships

- ➡Central database with server interface
- ➡Data Attributes
- ➡Data Discovery

Data Transfer moved data between sites

- ➡Ensures consistency and data integrity
- ➡Can enforce priority, load balance, and traffic shape
- ➡Database, agent architecture

PhEDEx

The way for sites, including Tier-3 centers, to send and receive official experiment data is PhEDEx

- PhEDEx makes subscriptions in a central Oracle DB at CERN
- Series of agents execute transfer requests
- PhEDEx is configurable and can handle a number of end-point configurations
  - Most common is SRM to SRM with either FTS or srmcp
  - Possible to use gsiftp as the end-point of even local file output

A Local data administrator can make data subscriptions using a web interface and valid grid credentials

Specifying and Submitting Applications

Once the data blocks have been located at a site the analysis jobs must be submitted

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

- CRAB was originally developed by INFN, though has grown into a global effort with contributions from the US and the UK
- 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 handles the data discovery
    - Query the DBS to determine the blocks required to complete the request and then the DLS to determine the clusters that can satisfy the request
  - The job preparation
    - Tarring up the user application and parameters, while making the appropriate number of jobs for the events needed to process
  - Submitting the application
    - Submitting jobs through the appropriate grid infrastructure

CRAB Submission

A user can query the DBS to determine dataset parameters

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

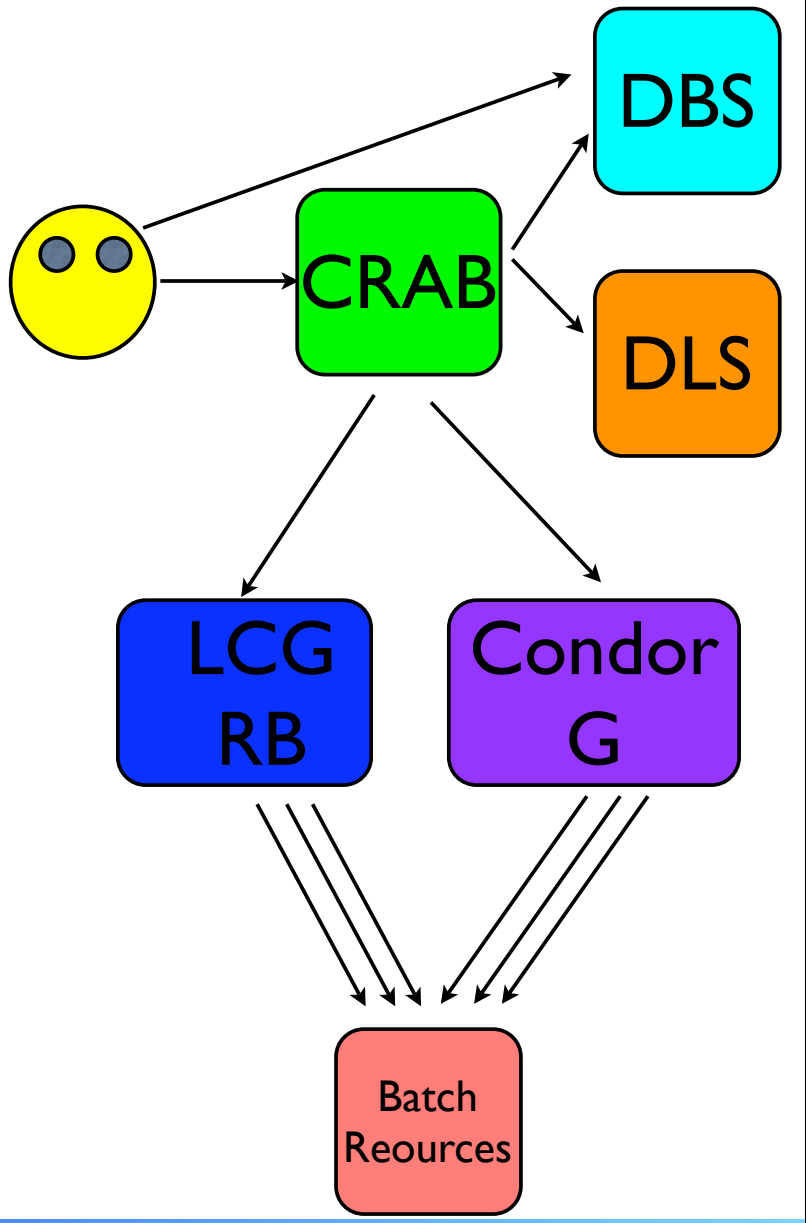
The identified dataset is defined by a number of data blocks

- ➡ Job can be sent to any site with the published set of blocks

A File list from DBS allows job splitting

Specified jobs are sent either to the LCG resource broker for the EGEE resources or Condor-G for the OSG resources

- ➡ RB has more functionality, while Condor-G is faster



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. A Tier-3 could host for local users

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 experience of the previous generation of detectors, we should expect about 3 years to stabilize
- People working at Tier2 and Tier-3 centers can make substantial, but bursty requirements of the data transfers

Analysis Selections

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

- 1% selection on data and MC would be 4TB, 10% selection would be 40TB
  - Smaller by factor of 5 if only the offline stream can be used
- There are an estimated 4-8 people working at a Tier-3 center
  - If half the people perform the small selections at the level of twice a month
    - This is already 5MB/s on average and everyone is working asynchronously
      - Conceivably you don't want to wait for 2 weeks to get a dataset
    - The original analysis estimates were once a week
    - 10% selections will happen. Easily get to 50-100MB/s of burst connectivity for T3

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

Getting Calibration

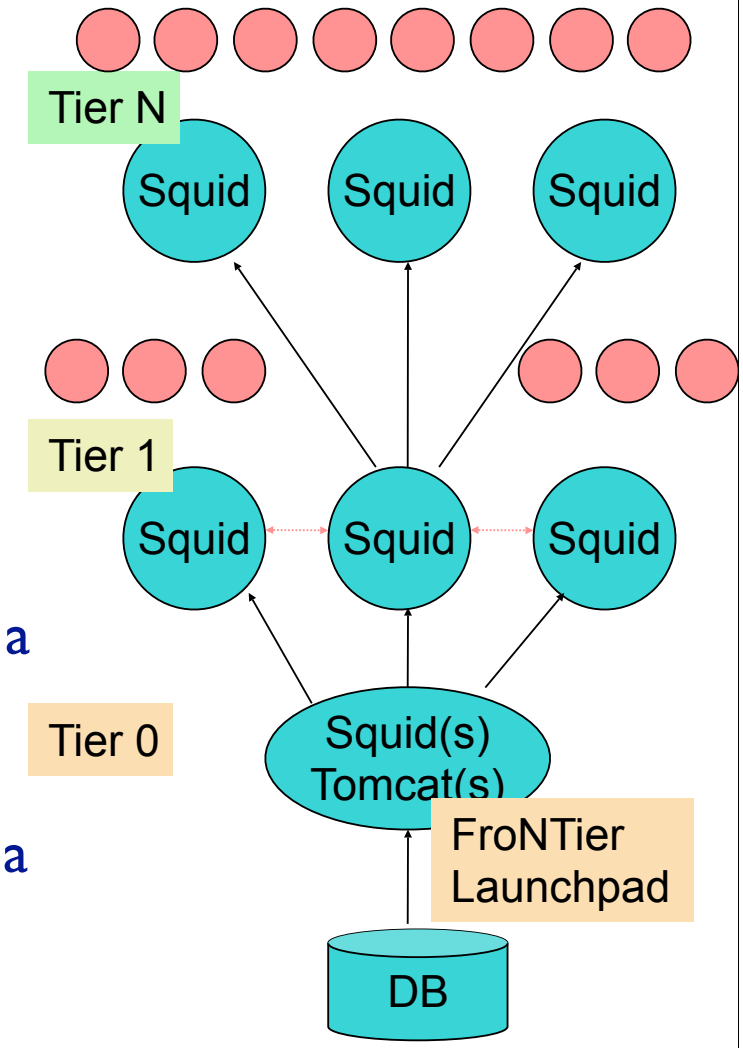
Especially, when dealing with raw or reconstructed data access to reliable and current calibration information is important

A traditional difficult operation with distributed computing is access to databases

- For remotely processed applications to be trusted they need up-to-date access to calibration constants and conditions data
- Database synchronization is difficult and it requires remote sites to have licenses and/ or database administration effort

CMS is attempting to solve this problem with a solution that caches database queries in web caches (SQUIDS)

Tier-3 centers could have a SQUID or access a remote centr



Getting Simulation

The goal in CMS is to make the official request system for simulation perform with a sufficiently low latency that users won't feel compelled to make their own Monte Carlo

- The Monte Carlo request system was recently rolled out.
  - It's a web site where an individual can request dataset to be created
- Eventually it will be tied to the Production infrastructure with some management components
  - Guide priorities and keep the production system busy
- Goal is to be able to deliver small samples in a couple of days
- Request system will eventually also put the finished data in a defined spot for analysis

Outlook

CMS Services are intended to make as few constraints as possible on the analysis users

- Data discovery, analysis job submission, simulation production are largely external services. This allows local analysis submission machines to be pretty easily supported

If a cluster wants to make itself available for opportunistic computing for simulation this is also intended to be light on the cluster

- An OSG gatekeeper is needed
  - The software environment is centrally installed, jobs are centrally submitted

Using the local cluster for local access to local data requires some services

- PhEDEx installation is needed to bring data down to the site
- Local storage.