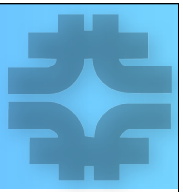


Introduction to CMS

Ian Fisk

October 23, 2006

Introduction

CMS has had a distributed computing model from early in on. Motivated by a variety of factors

- The large quantity of data and computing required encouraged distributed resources from a facility infrastructure point of view
- Ability to leverage resources at labs and university
  - Hardware, expertise, infrastructure
- Benefits of providing local control of some resources
- Ability to secure local funding sources

~20% of the resources are located at CERN, 40% at T1s, and 40% T2s

Can only be successful with sufficient networking between facilities

- Availability of high performance networks has made the distributed model feasible

Also relies on the development and success of Grid services and interfaces

- Efficient distributed computing services

Input Parameters For Computing Model

Event Sizes

- Current estimate of raw data event size is 1.5MB (1-2MB)
- Size of Reconstructed Event is 0.25MB
- Analysis Object data is 0.05MB per event

CMS best estimate is about 150Hz for the DAQ target Event rate

- ~ 250MB/s
- CMS is looking at first year scenarios with larger trigger rates

During normal CMS running we expect to log about 2PB of data per year of raw data

- About 30%-50% of that comes directly to FNAL for archiving and serving
  - 30% of raw, a larger fraction of reconstructed, and a full AOD copy
- During the first several years of the experiment the analysis will have to access more raw data
- Leads to larger data sets for analysis and larger selected datasets

CMS Computing Model

The CMS computing model is not the MONARC model circa 1998

- The strict hierarchies of access do not exist
  - Tier-2 and Tier-3 centers have to be able to connect to any Tier-1 center
  - Tier-1 centers communicate with each other

The CMS model is also not a pure grid computing cloud model

- Activities running at each tier are predictable and prescribed
  - Opportunistic computing is reserved for a very limited set of functionality
- The data location drives the activities at a site

Data is divided into on-line trigger streams and assigned to Tier-1 centers

- Approximately 10
- Sub-divided into off-line trigger streams
  - Approximately 50

Data Driven Baseline

Data placement drives activity at the Tier-0 and Tier-1 centers in the CMS baseline model.

- Data is partitioned by the experiment as a whole
- Tier-0 and Tier-1 are resources for the whole experiment
- Leads to very structured usage of Tier-0 and Tier-1
  - Tier-0 and Tier-1 centers are CMS experiment resources and activities are nearly entirely specified
    - Primary reconstruction, Re-reconstruction, Data and Simulation Archiving, Data and Simulation Serving, and Data Skimming

Tier-2 and Tier-3 Centers are the place where more flexible, user driven activities can occur

- Portion of resources are controlled by the local community
- More chaotic analysis activities
- Very significant computing resources in need of good access to data

Roles and Responsibilities

Tier-0

- Primary reconstruction / Partial Reprocessing
- First archive copy of the raw data

Tier-1s

- Share of raw data for custodial storage
- Data Reprocessing
- Data Selection
- Data Serving to Tier-2 centers for analysis
- Archive Simulation From Tier-2

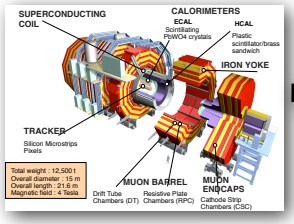
Tier-2s

- Monte Carlo Production
- Analysis

Tier-3

- Local Analysis and Opportunistic Computing

Data Flows



Tier-0

Tier-0 to Tier-1 Flow
Predictable
High Priority(300MB/s
From CERN)

Tier-1 to Tier-1
Burst with Rereco
Load Balancing
(100MB/s)

Tier-1

Tier-1

Tier-1

Tier-1 to Tier-2
Burst with User
Needs (50-500MB/s)

Tier-2

Tier-2 to Tier-1
Predictable
Simulation Data
(10MB/s)

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

Computing Center Specifications

Tier-0 Center

		Running Year 2007	2008	2009	2010	
Conditions		Pilot	2E33+HI	2E33+HI	E34+HI	
Tier-0	CPU	2.3	4.6	6.9	11.5	MSi2k
	Disk	0.1	0.4	0.4	0.6	PB
	Tape	1.1	4.9	9	12	PB
	WAN	3	5	8	12	Gb/s

Tier-1 Centers

A Tier-1	CPU	1.3	2.5	3.5	6.8	MSi2k
	Disk	0.3	1.2	1.7	2.6	PB
	Tape	0.6	2.8	4.9	7.0	PB
	WAN	3.6	7.2	10.7	16.1	Gb/s

- 1/6
- US-CMS is roughly twice as large

Tier-2 and Tier-3 Centers

A Tier-2 center in CMS is approximately 1MSI2k of computing

- Tier-3 centers belong to university groups and can be of comparable size

A Tier-2 center in CMS ~200TB of disk

- Currently procuring and managing this volume of storage is expensive and operationally challenging
  - Requires a reasonably virtualization layer

A Tier-2 center has between 1Gb/s and 10Gb/s of connectivity

- This is similar between Tier-2 and Tier-3 centers

In the US planning a Tier-2 supports 40 Physicists performing analysis

- This is a primary difference between a Tier-2 and a Tier-3

Tier-2 centers are a funded effort of the experiment

- The central project has expectations of them

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

CMS Data Concepts

Data File

- A file you can access with an application
  - Currently CMS opens one file

Data Block

- A group of files large enough for the data transfer system to worry about and the data publishing system to publish

Dataset

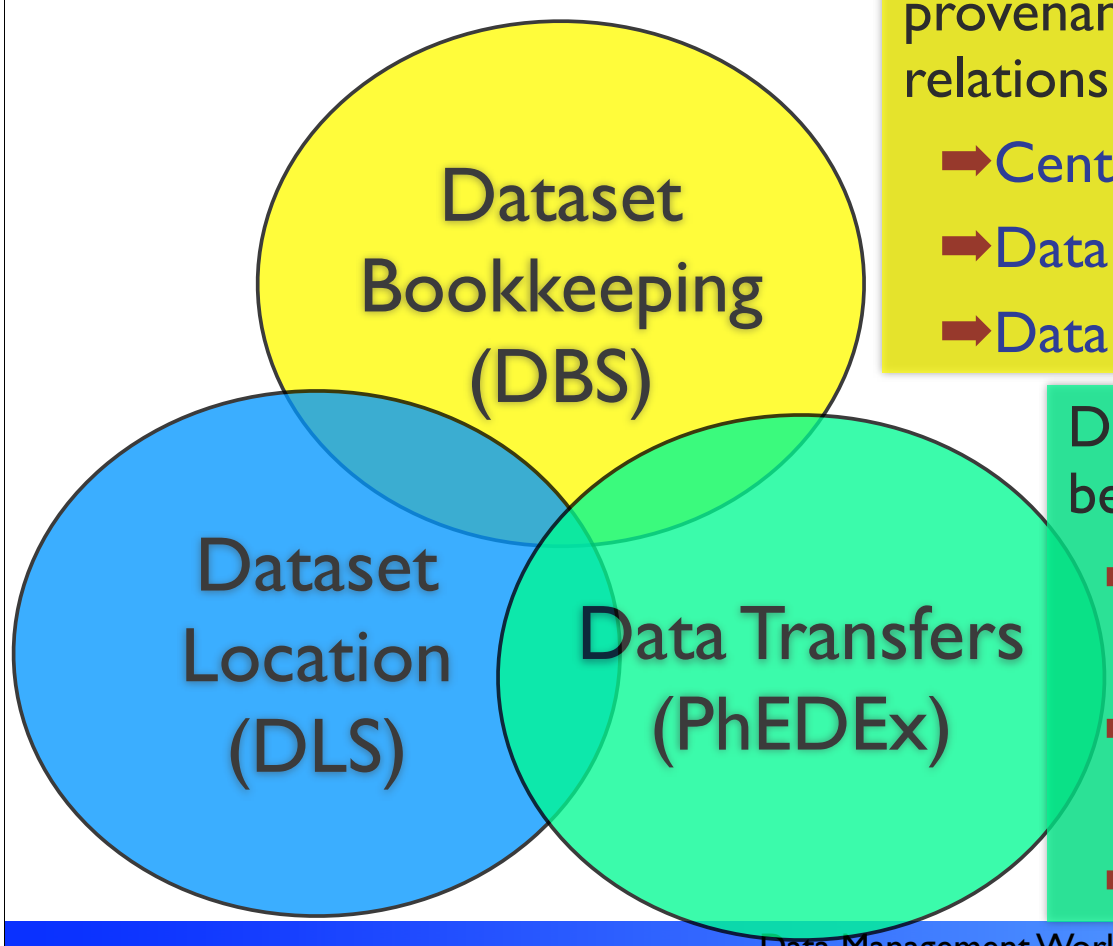
- A group of data blocks associated with a production or an analysis
  - Published in the DBS

The name space

- A consistent namespace used for data the experiment tracks
  - Allows resolution of logical to physical file name without an external catalog

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 to send and receive official experiment data is PhEDEx

- For Tier-3s the best way to receive datasets 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

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.

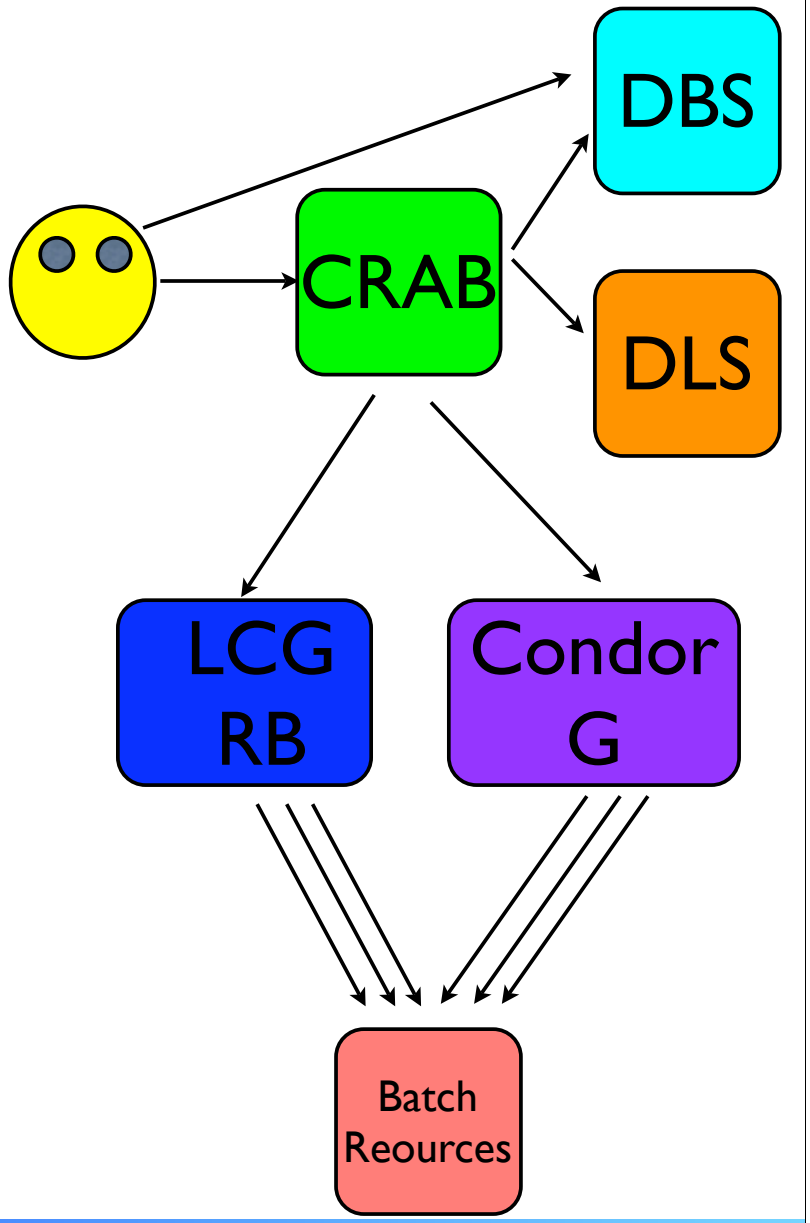
The identified dataset is defined by a number of data blocks

- ➡ 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



ProdAgent

Simulation is handled by the ProdAgent infrastructure

- Jobs come from central teams
  - Output is written to local SE
  - Moved out by PhEDEx
- Expectations on the sites are low