Analysis and the CMS Computing Model

LPC JTerm IV
Eric Vaandering
4 August 2009

Introduction

Oliver and Dave gave very nice overviews of the CMS computing model yesterday

This talk covers some of the same ground (not too much I hope) but with a different emphasis

I want to go a bit deeper into the parts of the computing model that are most relevant to ordinary users and physics analysis

- Also cover some tools and concepts specific to your needs

Outline

- Computing Model Review
- Role of Tier2s in Analysis
- CMS Remote Analysis Builder (CRAB)
- Data Transfer and Bookkeeping (PhedEx/DBS)
- Analysis Datasets
- Resources and Facilities for Users
- Resources and Facilities for Physics Groups
- StoreResults service

Roles and Responsibilities

Tier-0

- Primary reconstruction
- 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
- User Analysis and Storage
- Physics Group Data Storage

Data Driven Baseline

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

- Tier-0 and Tier-1 are CMS experiment resources and activities are nearly entirely specified

Tier-2 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

Tier-2 computing centers represent the bulk of the analysis computing resources for the experiments

- In the early years of the experiment serious analysis may require frequent access back to the raw data samples
  - Making selections and moving the data to Tier-2s for detailed analysis

US Sites

The US has a Tier-1 Center at FNAL

- FNAL is the largest Tier-1 center in CMS
  - One of 7 Tier-1s
  - The only Tier-1 center in the Americas
    - The center will reach 6MSI2k, 2PB of disk, 4.7PB tape
    - WAN network is 20Gb/s

The US has 7 Tier-2 computing facilities

- Caltech, Florida, MIT, Nebraska, Purdue, UCSD, Wisconsin
- All have 300 TB (200TB nominal) of disk
- All US sites now have 10Gb/s network links
- Dedicated 50% CPU to Users, 50% to Monte Carlo production

In addition FNAL has an analysis farm called the LPC CAF

- Much like a Tier-2 in terms of analysis

Tier2 Data Storage

In CMS jobs go to the data.

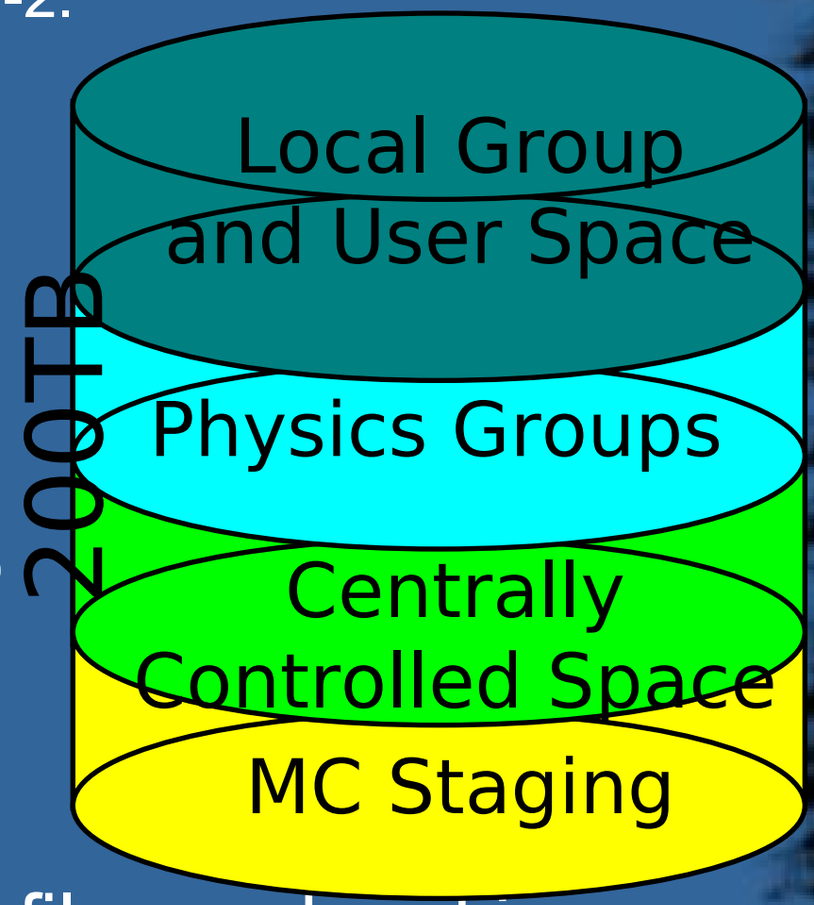
- The challenging part is making sure the right data is distributed broadly
  - There are 200TB of disk space at a nominal Tier-2.
  - CMS attempts to share the management of the space across groups

30 TB of space at each site is for DataOps

- We expect to be able to host most of the RECO data used in the first year

30TB is identified for use by the local group

- Local community controlled space



20TB is identified for storing user produced files and making them grid accessible

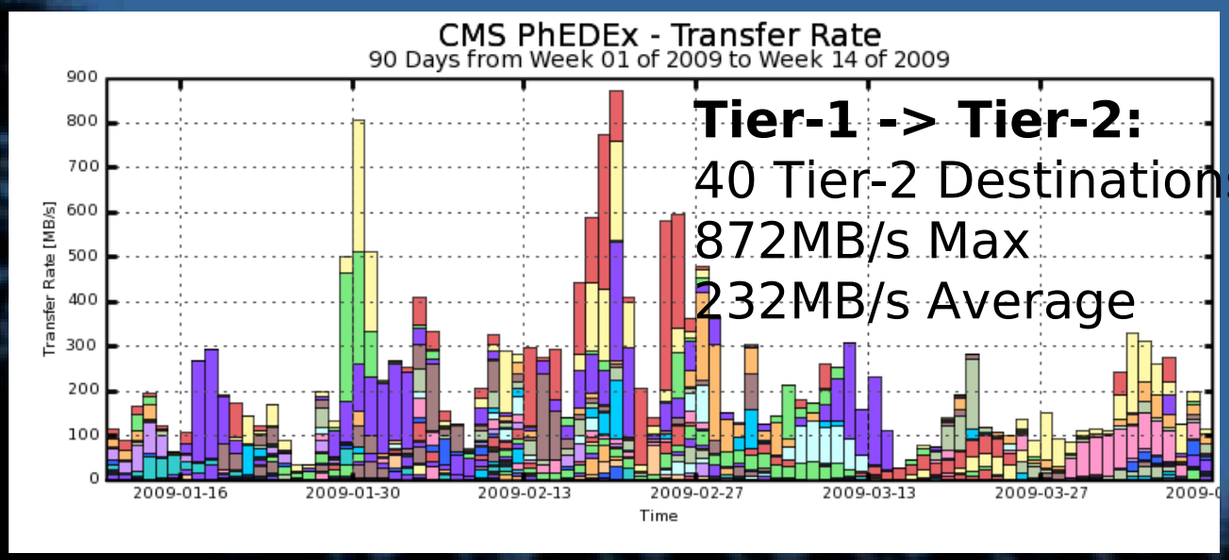
Tier-3s

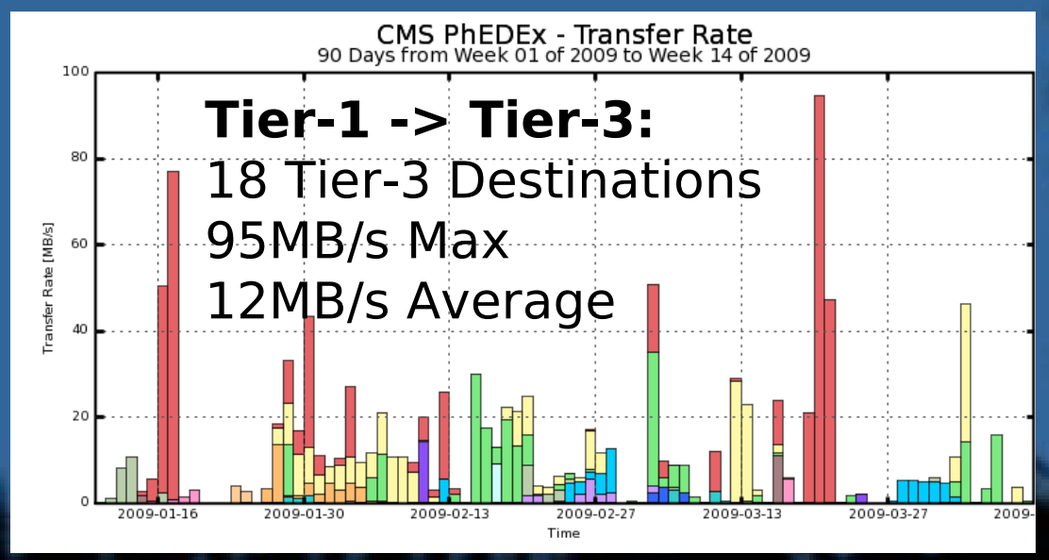
The Tier-3s are an important analysis resource

- Entirely controlled by the community that provides them
- Minimum to be a registered Tier-3 is a PhEDEx instance for data transfers
  - Currently 38 Tier-3s registered in PhEDEx, 17 in the US
  - A number of Tier-3s have a full grid interface and accept analysis jobs
    - 10 of the US Tier-3 sites on the Open Science Grid

Currently the Tier-3s can receive data from anywhere

- There are large variations in sizes and configurations for Tier-3s, though some standards are emerging





CRAB and the GRID

- The GRID
  - Interconnects world-wide computing resources (batch farms, storage systems, ...) in a transparent way
  - Provides every user with access to these resources
    - Authorization, job submission, ...
- CMS uses the GRID to provide data access for all collaborators.
- CRAB (CMS Remote Analysis Builder) is the CMS user front end to the GRID

Glossary

UI	User Interface: GRID tools on your local PC
RB	Resource Broker: central job submission point
CE	Compute Element: GRID door to compute farm
SE	Storage Element: GRID door to mass storage
WN	Worker Node: Batch slot in a farm accessed through CE
EDG	European Data Grid
EGEE	European Grid to Enable E-Science
OSG	Open Science Grid
SAM	Site Availability Monitoring: continuous tests of GRID infrastructure status

GRID certificates

- CMS uses GRID certificates and a dedicated Virtual Organization (VO) management to have better access control for specific tasks/groups
- You need this for making Phedex requests, submitting jobs, and can use to to access Indico and the TWiki
- Your GRID certificate is important, follow all rules, don’t let it expire!
- Can be a pain the first time, but then it's over

UI

- The User Interface is the GRID specific software for authentication, job submission and all other GRID interactions
- There are UI's from EGEE and from OSG
- FNAL: pre-initialized on cmslpc.fnal.gov
- CERN initialization: `source /afs/cern.ch/cms/LCG/LCG-2/UI/cms_ui_env.(c)sh`
- You won't have to use it directly, CRAB uses it for you

CRAB

- CRAB enables the user to submit CMSSW jobs to all CMS datasets within the data-location driven CMS computing structure
- The aim of CRAB is to hide as much of the complexity of the GRID as possible from the end user
- CRAB provides a user front-end to
  - Split user jobs into manageable pieces
  - Transport user analysis code to the data location for execution (compiled on submitting node)
  - Execute user jobs, check status and retrieve output

Expected analysis flow

User runs interactively on small samples in the local environment to develop the analysis code and test it
(Get a file with the file mover)

Once ready the user selects a large (whole) sample to submit the very same code to analyze many more events

The results are made available to the user to be analyzed interactively to produce final plots

How to run CRAB

The basic CRAB workflow is organized in 4 steps:

- Job Creation
- Job Submission
- Job Check status
- Output retrieval

Job Creation: **crab -create**

At this level CRAB interacts with the DBS system, organizes the jobs of the task according to the user's job splitting parameters, packs the users specific code(/lib /module /data), prepares the script to configure the remote environment, and (using BOSS) prepares the jdl file to communicate with the RB

It also creates the working directory which is organized in 4 subdirectories named:

/job : CRAB specific stuff
/log : CRAB log file location
/res : default results destination
/share : CRAB and BOSS specific stuff

How to run CRAB (2)

Job submission: **crab –submit**

The submission uses the previously created CRAB project to submit the jobs.
Before the real submission, CRAB always checks for available resources preventing the submission of unmatched jobs. By default all created jobs are submitted.

Job status: **crab –status**

This command checks the status of all jobs in the CRAB project.
For each job CRAB prints on the screen the job id, scheduler status, site hosting the jobs, cmssw exit code, & job exit code. The output gives also a summary with a list of job IDs sorted by status categories.By default the status of all jobs is checked.

Job output : **crab –getoutput**

This command retrieves the output of all jobs of a CRAB project which have status “Done”. By default the retrieved output files are copied in the “res” sub-dir of the CRAB workingdir. Included are the standard output and error of the jobs (CMSSW stdout and stderr) and the output files specified in crab.cfg.

Even if your job fails, run **crab -getoutput**. Otherwise your output clogs up a server.

CRAB Tutorials

- We no longer regularly give a basic (new user) CRAB tutorial at Jterm
  - Walkthroughs on the web are usually good enough
  - Only usual measure of success is if you followed all the instructions to get your certificate, registered in the CMS VO, etc.
  - We're generally pressed for time with CMSSW, FWLite, PAT, etc.
- Instead we have a tutorial for users who've been basic users of CRAB for a while and want to learn something new
- Latest walkthrough specific to FNAL users:

https://twiki.cern.ch/twiki/bin/view/Main/EricVaanderingCRABTutorialAug2009

How to get CRAB support

Best source for user support is the CRAB feedback hypernews:
https://hypernews.cern.ch/HyperNews/CMS/get/crabFeedback

All CRAB questions and suggestions can be posted to this forum, CRAB developers try to solve the problems and give solutions. User suggestions can help improve the tool.

A troubleshooting guide is at
https://twiki.cern.ch/twiki/bin/view/CMS/WorkBookGridJobDiagnosisTemplate

Questions not directly related to CRAB (GRID related problems, CMSSW specific problems, etc...) should be referred to other hypernews forums

Additional Documentation:

Wiki: https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideCrab

Data Management

CMS data is divided into

- Datasets - A group of events that can be accessed together
- Data Blocks - A group of files that is tracked by the data transfer system
  - A dataset is a group of blocks
- Logical file name (LFN) – A single file in a block, uniform name
- Physical file name (PFN) – Actual file name on a site

The Dataset Discovery Page is at

- https://cmsweb.cern.ch/dbs_discovery/
  - Wildcard searches
  - Select on software releases

Data is also divided by Event Content (Tier)

- RAW, RECO, AOD
  - Datasets may be skimmed and reduced

Data Management

CMS data is div

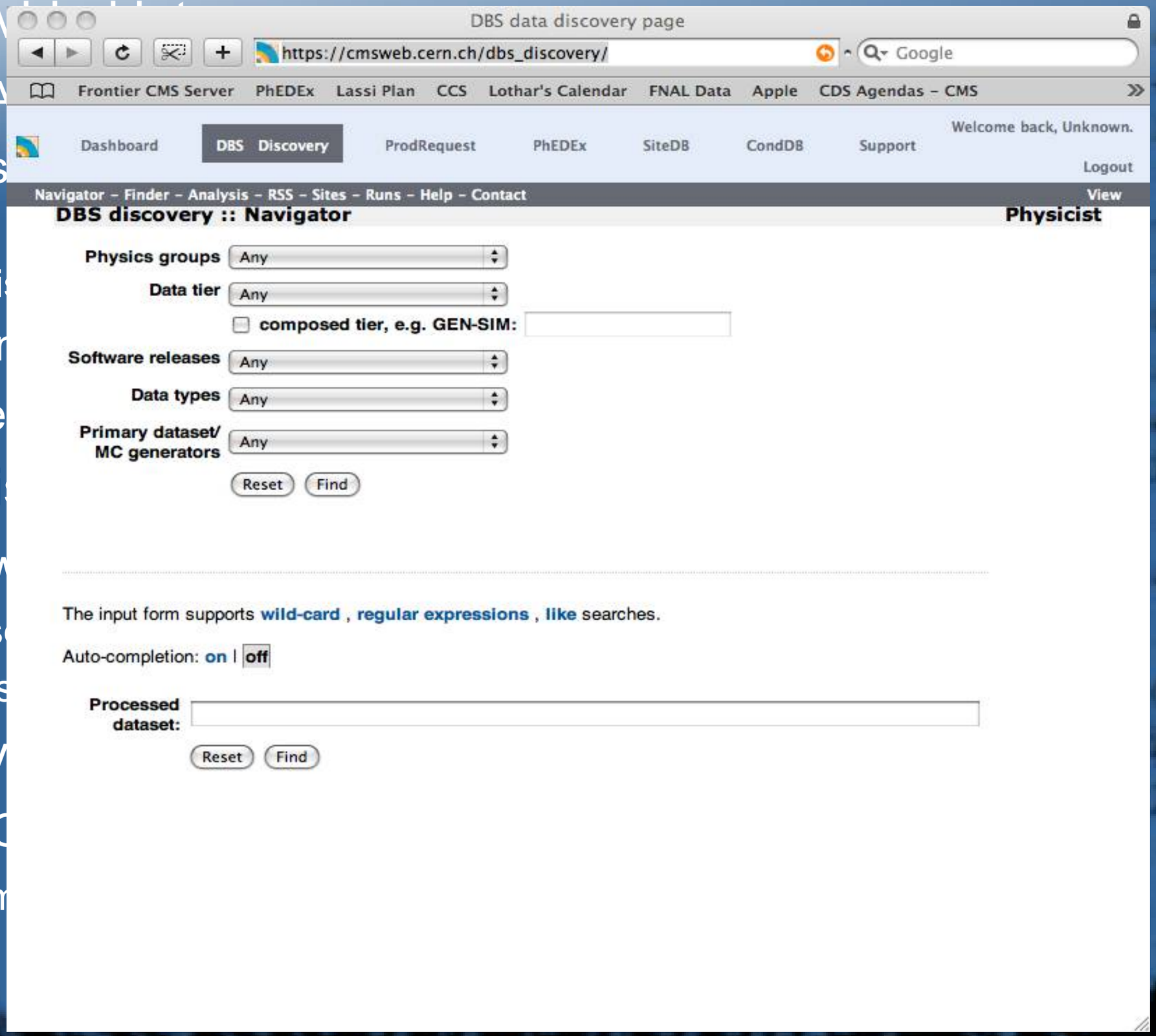
- Datasets - A
- Data Blocks
  system
  - A dataset i
- Logical file r
- Physical file

The Dataset Di

- https://cmsw
  - Wildcard s
  - Select on s

Data is also div

- RAW, RECO
  - Datasets m



Datasets

- Dave Mason covered this yesterday
- Primary dataset: name that describes the physics channel
  - Based on triggers (for data) or generator settings
- Processed dataset: describes the kind of processing applied
  - Includes era and version
- Data Tier: describes the kind of event information stored from each step in the simulation and/or reconstruction chain.
  - e.g. RAW, RECO, and for MC, GEN, SIM and DIGI.
  - May consist of multiple tiers, e.g., GEN-SIM-DIGI-RECO
- Example:
  - /Cosmics/CRAFT09-PromptReco-v1/RECO

Finding Datasets

- Probably the best way is to talk to others in your physics group to find out what datasets are relevant to your analysis
- DBS Discovery is a powerful tool, but can be a little daunting at first
  - Often searching on some text plus the wildcard '*' can give you the results you want
  - Remember, not every sample needed for a top analysis is listed under the 'Top' physics group
- Good documentation on the DBS query language (with examples!): https://twiki.cern.ch/twiki/bin/view/CMS/QL

Analysis Datasets

- A way to pick out the data *you* want for your analysis
- Consists of a processed dataset plus a mask
  - Can mask by run, luminosity block, data quality flags, etc
    - Quality can be things like “Muon detector working” or “Magnet on”
  - Criteria differ from analysis to analysis
- Support for this feature is not complete, but is progressing
  - Fully supported in CMSSW as of 3_1_0
  - Will be fully supported in next version of CRAB (2.6.2 or 2.7.0)

Data Movement

Data Movement in CMS is handled by a tool called PhEDEx

- PhEDEx replicates individual files and updates the data management system when complete blocks have been transferred

Data Movement to Tier-1 centers is a decision of central CMS

Data Movement to Tier-2 centers is intended to be driven by needs or users and groups

- There can be open data subscriptions that grow as data is available
- The majority of the storage at the Tier-2s is intended for serving experiment data
- You can make a request, but make sure you coordinate with the person(s) in your physics group responsible for managing the group's space
  - Or have them make the request

Data Transfers

The PhEDEx page is available at

- http://cmsdoc.cern.ch/cms/aprom/phedex/prod/Activity::RatePlots?view=global

Any user can make a transfer request.

- Only a site data manager can approve a request
- Data managers make sure the site is not over subscribed

Data on Tier-2 sites is not intended to be kept forever

Data Transfers

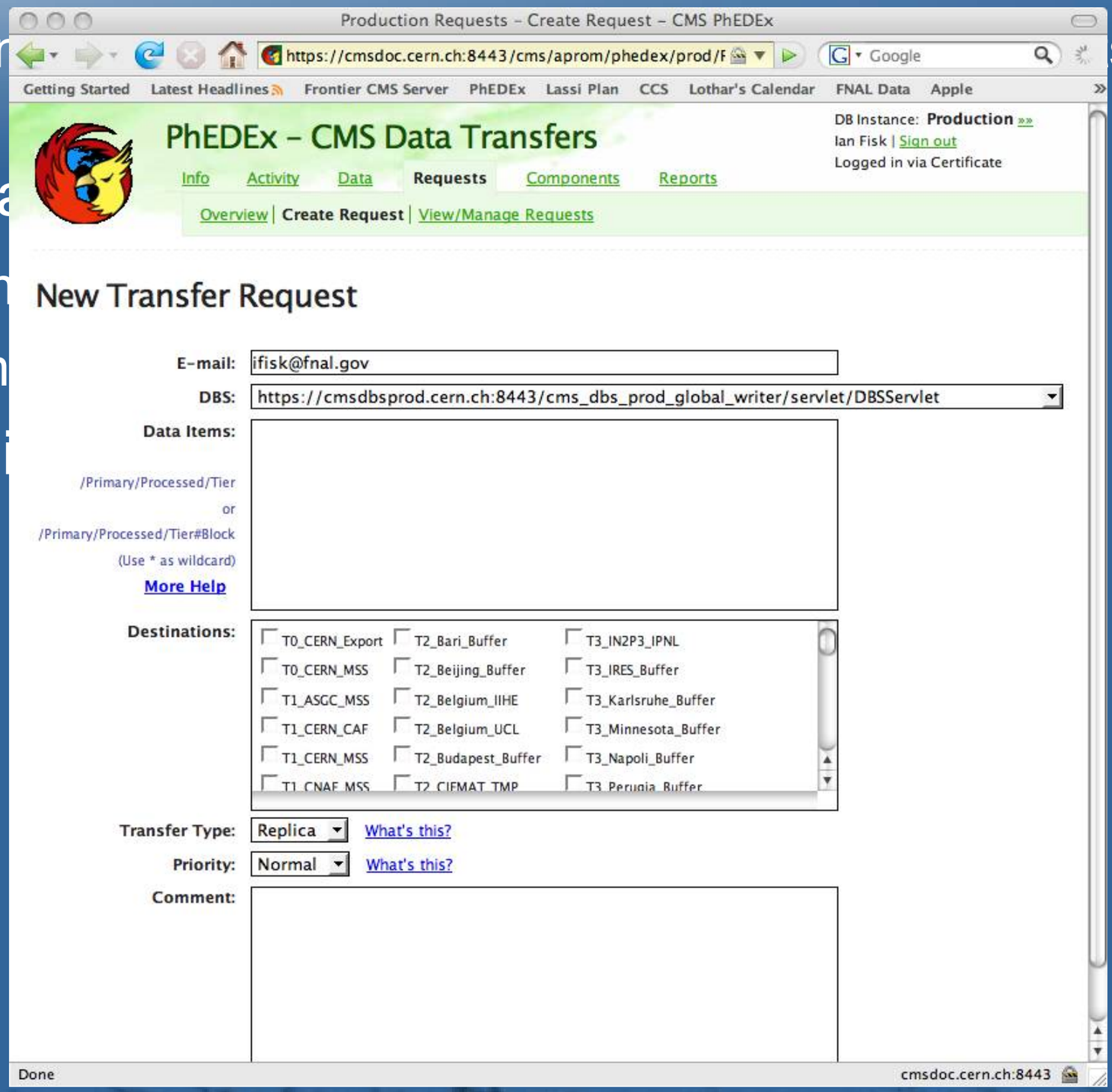
The PhEDEx page is available at

- http://cmsdoc.cern… s? view=global

Any user can make a…

- Only a site data m…
- Data managers m…

Data on Tier-2 sites i…



Facilities for Users

- As a user you are granted some resources at a Tier2
  - Each US institution is associated to one of the US Tier2s or FNAL
  - Generally split geographically (e.g. Big 12 @ UNL, Midwest @ Wisconsin)
  - About 1 TB of space for your own use in /store/user/<hnname>
  - Some may allow interactive log on as well: Check with them
  - Listed at http://www.uscms.org/uscms_at_work/software_computing/tier2/store_user.shtml
- As a USCMS member you are also allowed an interactive account at FNAL
  - Access to LPC CAF with >2000 CPUs
  - CAF has access to all FNAL Tier1 data

LPC CAF

The LPC CAF has a lot in common with other Tier-2-like analysis centers

- The disk requirements were specified with the expectation that the analysis center had access to the data stored on the Tier-1 disks
  - Data not expected to be served from FNAL T1 must be subscribed to the LPC dedicated storage and approved by a site data manager
- Unlike most of the Tier-2s the LPC CAF only excepts local job submissions
  - You need to log in and submit
  - Reasonably well utilized cluster
- CRAB knows how to submit to CAF just like the Grid

Accounts can be applied for using instructions on the web page

- http://www.uscms.org/SoftwareComputing/UserComputing/GetAccount.html

Facilities for Physics Groups

- Physics groups are also supported by Tier2 sites
  - Each physics group supported by al least two Tier2 sites
  - Up to the physics groups how they want to manage/allocate this space
  - Most have a site in the US
  - Tens of TB in storage at each site
  - List of associations:
    http://indico.cern.ch/materialDisplay.py?contribId=28&sessionId=22&materialId=slides&confId=41026
    Take a look at pages 10 & 11 for a list of who is where
- Also space for POGs (Physics Object Groups)

StoreResults

- Need a way to promote a single user's data to “official” status
- User creates some dataset with CRAB
  - Written and published in /store/user
  - Published in Local DBS and file size probably not right for moving to tape
- A physics group realizes it is useful, would like to distribute more widely
- After relevant approvals are made, migration is made
  - Have to make sure files are appropriate size for tape storage

Store Results process

- User requests migration of a sample
- Physics Group approves migration
- Data in /store/user is audited for size
- Data is migrated to /store/results/group
- Data is injected into Global DBS
- Data is injected into PhEDEx and transferred (if required) to new Tier2
- Prototype is working now, will be generally available “soon”

Analysis Operations

- This is a new group that is just beginning it's work
- Various tasks to provide analysis support for users
  - Operate CRABServers
  - Operate StoreResults service
  - Monitor CRAB Hypernews and user support
  - Support physics groups with data placement and validation
  - Monitoring, and Evaluation of distributed analysis
- Still understaffed. There is room here for service work