CMS Grid Computing at TAMU
Performance, Monitoring and Current Status of the Brazos Cluster

Vaikunth Thukral

Department of Physics and Astronomy

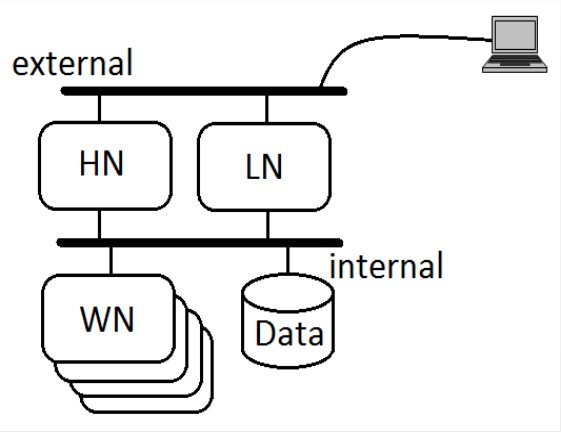
Texas A&M University

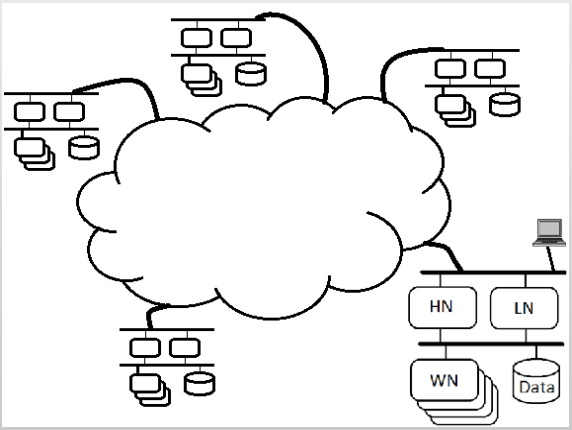
Outline

- **Grid Computing with CMS: PhEDEx and CRAB**
- **Our Local Computing center: Brazos/T3_US_TAMU**
- **Performance and Monitoring**
  - **Data Transfers**
  - **Data Storage**
  - **Jobs**
- **Summary**

Introduction to Grid Computing

- Cluster
  - Multiple computers in a Local Network
- The Grid
  - Many clusters connected by a Wide Area Network
  - Resources expanded for thousands of users as they have more access to distributed computing and disk
- CMS Grid: Tiered Structure (Mostly about size & location)
  - Tier 0: CERN
  - Tier 1: A few National Labs
  - Tier 2: Bigger University Installations for national use
  - Tier 3: For local use (Our type of center)





Next: Define PhEDEx and CRAB which are CMS ways of managing data and running "jobs" on the grid

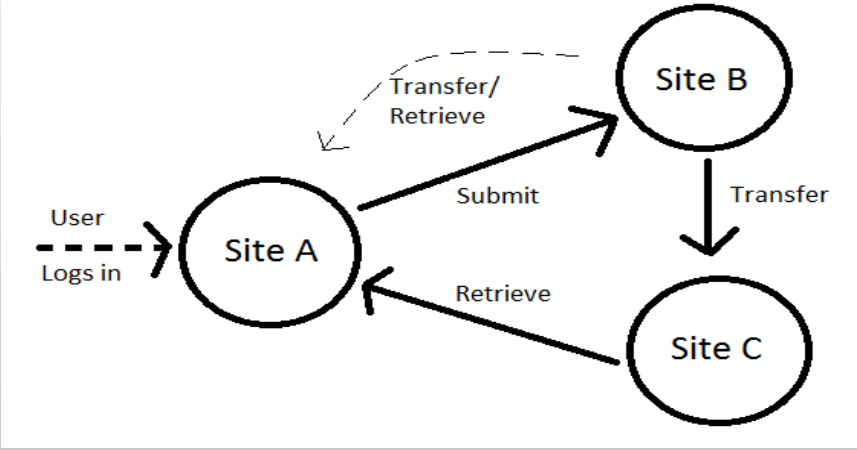
- "Jobs" - Breaking up the data analysis into lots of parallel pieces

PhEDEx

- **Ph**ysics **E**xperiment **D**ata **Ex**port
  - Data is spread around the world
  - Transport tens of Terabytes of data to A&M per month

CRAB

- **C**MS **R**emote **A**nalysis **B**uilder
  - Jobs are submitted to "the grid" using CRAB
  - CRAB decides how and where these tasks will run
  - Same tasks can run anywhere the data is located
  - Output can be sent anywhere you have permissions



Advantages of Having a CMS Tier 3 Computing Center at TAMU

- Don't have to compete for resources
  - CPU priority - Even though we only bought a small amount of CPUs, can periodically run on many more CPUs at the cluster at once
  - Disk space - Can control what data is here
- With a "standardized" Tier 3 on a cluster, can run same here as everywhere else
- Physicists don't do System Administration

T3_US_TAMU as part of Brazos

- Brazos cluster already established at Texas A&M
- Added our own CMS Grid Computing Center within the cluster
- Named T3_US_TAMU as per CMS conventions

T3_US_TAMU added CPU and Disk to Brazos as our way of joining

- Disk
  - Brazos has a total of ~150TB of storage space
  - ~30TB is assigned to our group
  - Space is shared amongst group members
    - N.B. Another 20TB in the works
- CPU
  - Brazos has a total of 307 compute nodes/2656 cores
  - 32 nodes/256 cores added by T3_US_TAMU
    - Since we can run 1 job on each core → 256 jobs at any one time, more when cluster is underutilized, or by prior agreement
  - 184,320 (256 x 24 x 30) dedicated CPU hours/Month

Grid Computing at Brazos Summary

- Tier 3 is fully functional on the cluster
- Instructions on how to use it can be found at:

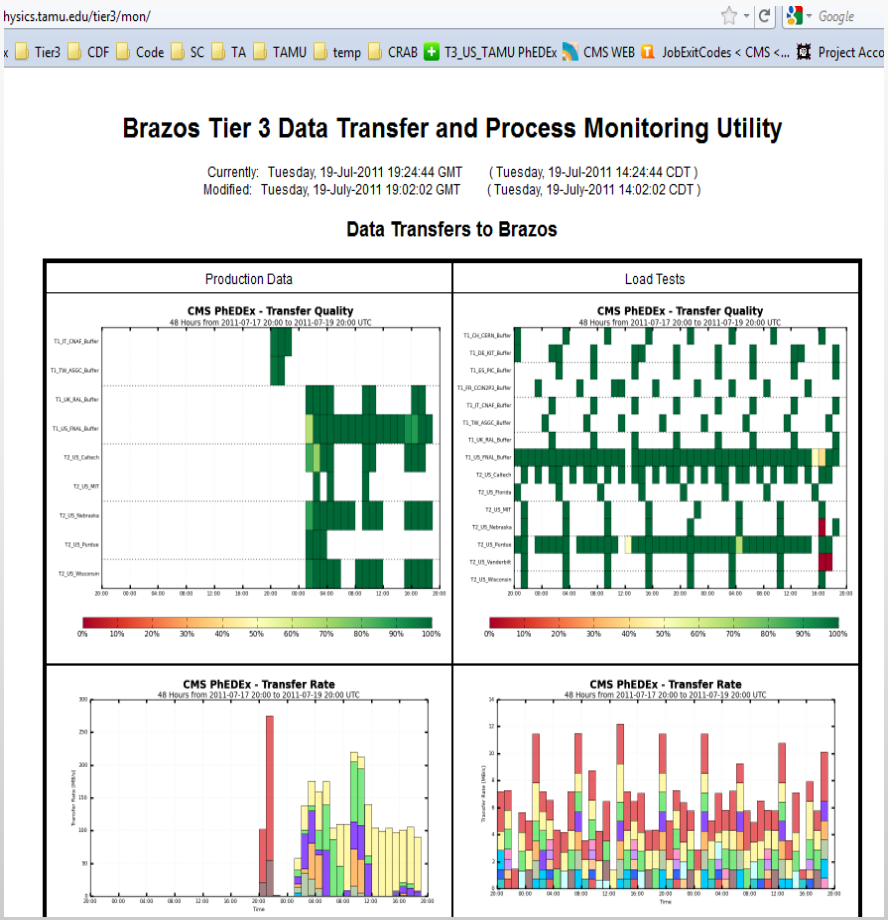
  http://collider.physics.tamu.edu/tier3/
- Updating our "Best Practices" on how to bring over data and run Jobs :

  http://collider.physics.tamu.edu/tier3/best_practice

Grid Computing at Brazos Summary

- Next: Move to describe how well the system is working by showing results from our Online Monitoring Pages(*)

- (*) Thanks to Dr. Joel Walker for leading this effort



Three Main Topics

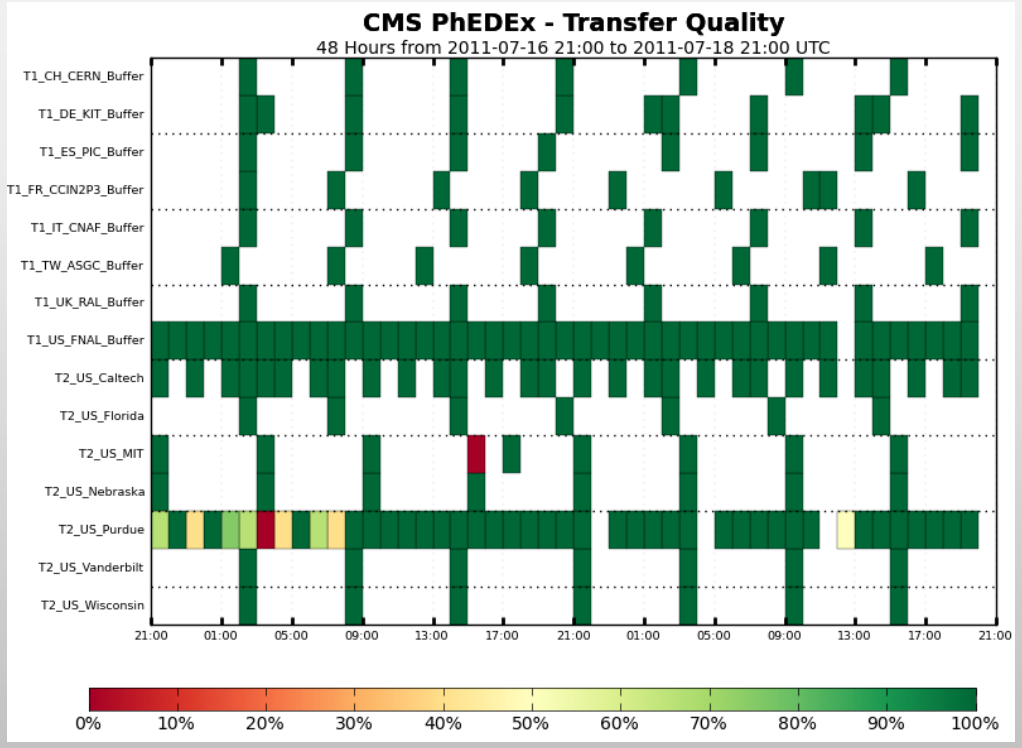
- Tier 3 Functionality
  - Data Transfers (PhEDEx)
  - Data Storage: PhEDEx Dataset + Local User Storage
  - Running Jobs (CRAB)
- Need to test and monitor all of these
  - CMS provides some monitoring tools
  - We have designed additional Brazos-specific/ custom monitoring tools

PhEDEx at Brazos

- PhEDEx performance is continually tested in different ways:
  - LoadTests
  - Transfer Quality
  - Transfer Rate

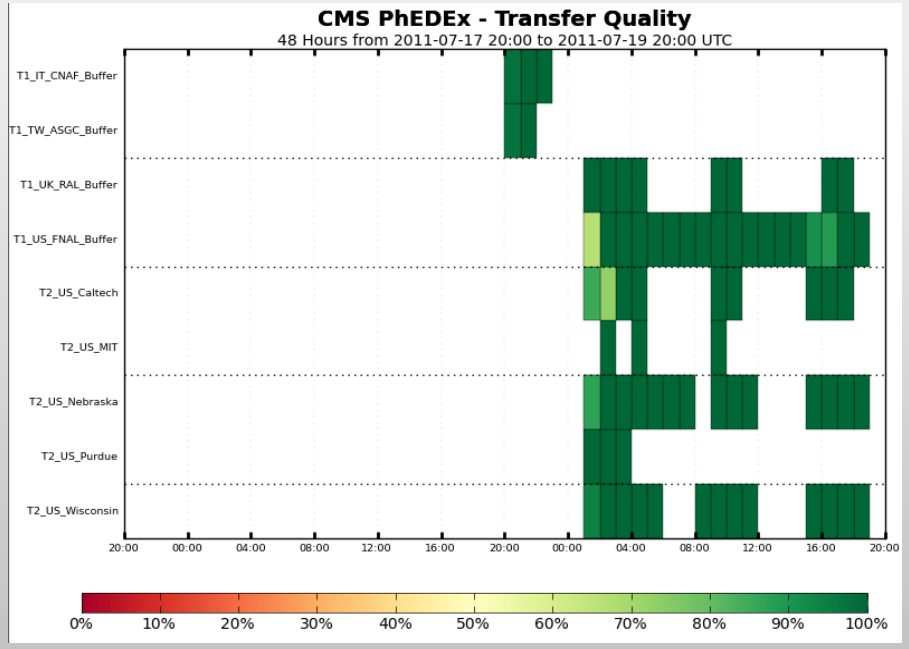
PhEDEx at Brazos (cont.)

- LoadTest
  - Acts as a test of the "handshake" between TAMU and linked sites in Taiwan, Europe, US etc.



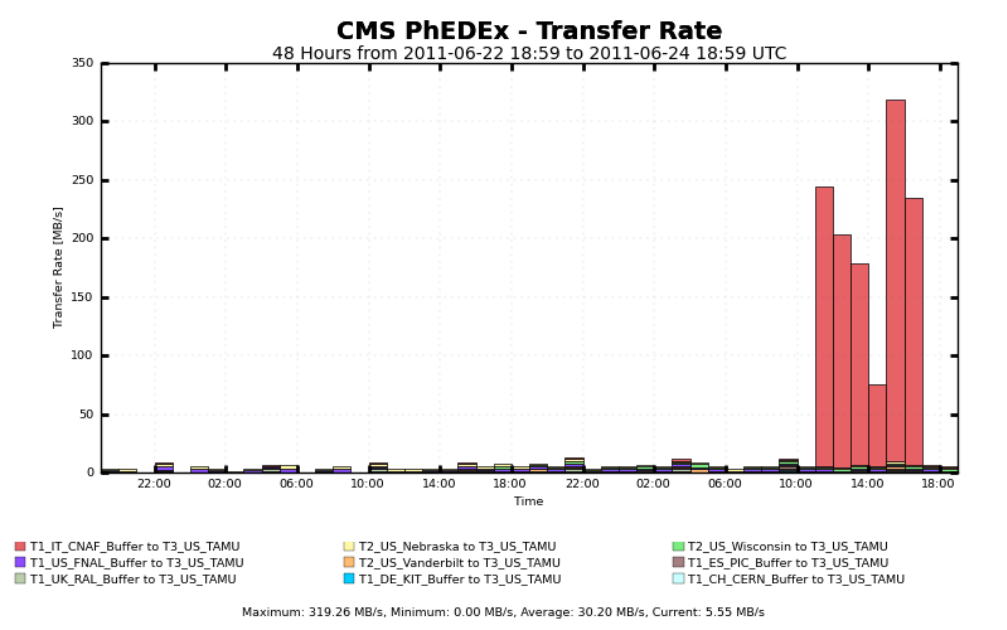
PhEDEx at Brazos (cont.)

- Transfer Quality
  - Monitors whether the transfers we have requested are actually coming across successfully
  - Transfers from Italy, Taiwan, UK etc.



PhEDEx Transfers

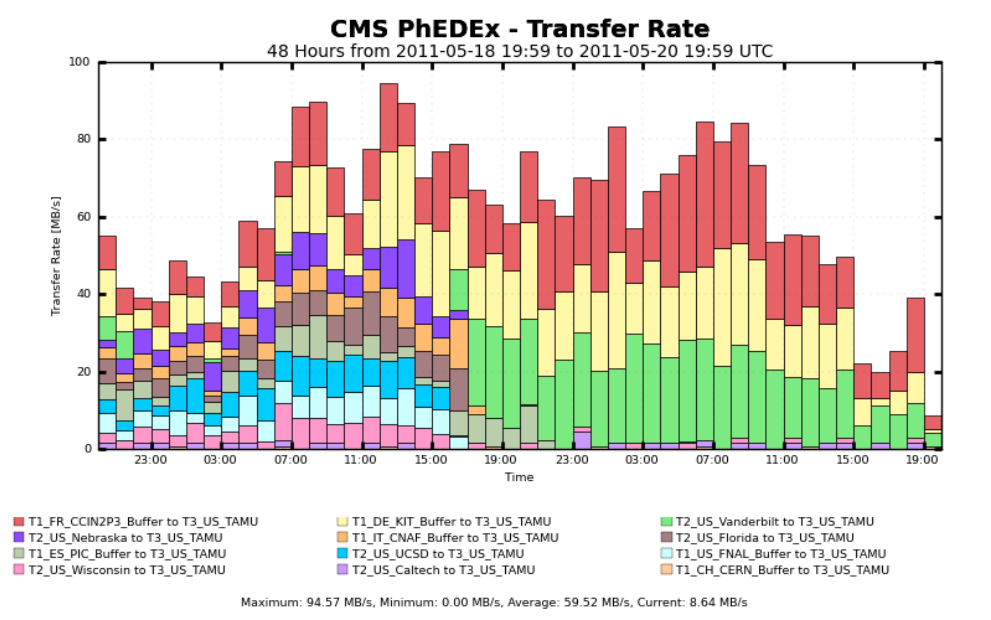
- PhEDEx Data Transfer Performance (Peak at 320 MB/s)



  - Network and client settings optimized
  - 20-fold increase in average transfer speeds from January to June
    - ~10 MB/s → ~ 200 MB/s
  - Other T3 sites average between 50-100 MB/s

PhEDEx Transfers (cont.)

- Can download from multiple locations at once and for extended periods of time



  - Capable of transferring large volumes consecutively
  - In principle, could download up to 25 TB in one day!
    - Have done 10TB just yesterday
  - Last month we brought over ~45TB

Data Storage and Monitoring

- Monitor PhEDEx and User files
  - HEPX User Output Files →
  - PhEDEx Dataset Usage ↓

Data on Brazos

	Subscribed			Resident			
Group Name	Items	Files	Bytes	Items	Files	Bytes	Percent
FacOps	1	8	10.1 GiB	1	8	10.1 GiB	100.0 %
deprecated-undefined	2	142	203.3 GiB	2	142	203.3 GiB	100.0 %
exotica	8	5667	8.7 TiB	8	5667	8.7 TiB	100.0 %
higgs	1	10	34.4 GiB	1	10	34.4 GiB	100.0 %
qcd	26	2411	4.5 TiB	16	2348	4.4 TiB	97.4 %
susy	5	6145	9.8 TiB	5	6145	9.7 TiB	100.0 %
Total	43	14383	23.2 TiB	33	14320	23 TiB	99.2 %

HEPX Disk Store Usage	
Directory	Bytes
User Output	2.6 TiB
Tai Sakuma	1.1 TiB
Indara Suarez	1.0 TiB
Roy Montalvo	436.7 GiB
Alfredo Gurrola	45.4 GiB
Jieun Kim	10.2 GiB
Vaikunth Thukral	20.0 KiB
PhEDEx Monte Carlo	20.3 GiB
PhEDEx Data	20.0 KiB
PhEDEx Load Tests	10.1 MiB
Miscellaneous	404.0 KiB
Total	2.6 TiB
↑ Click to Expand or Collapse Table	

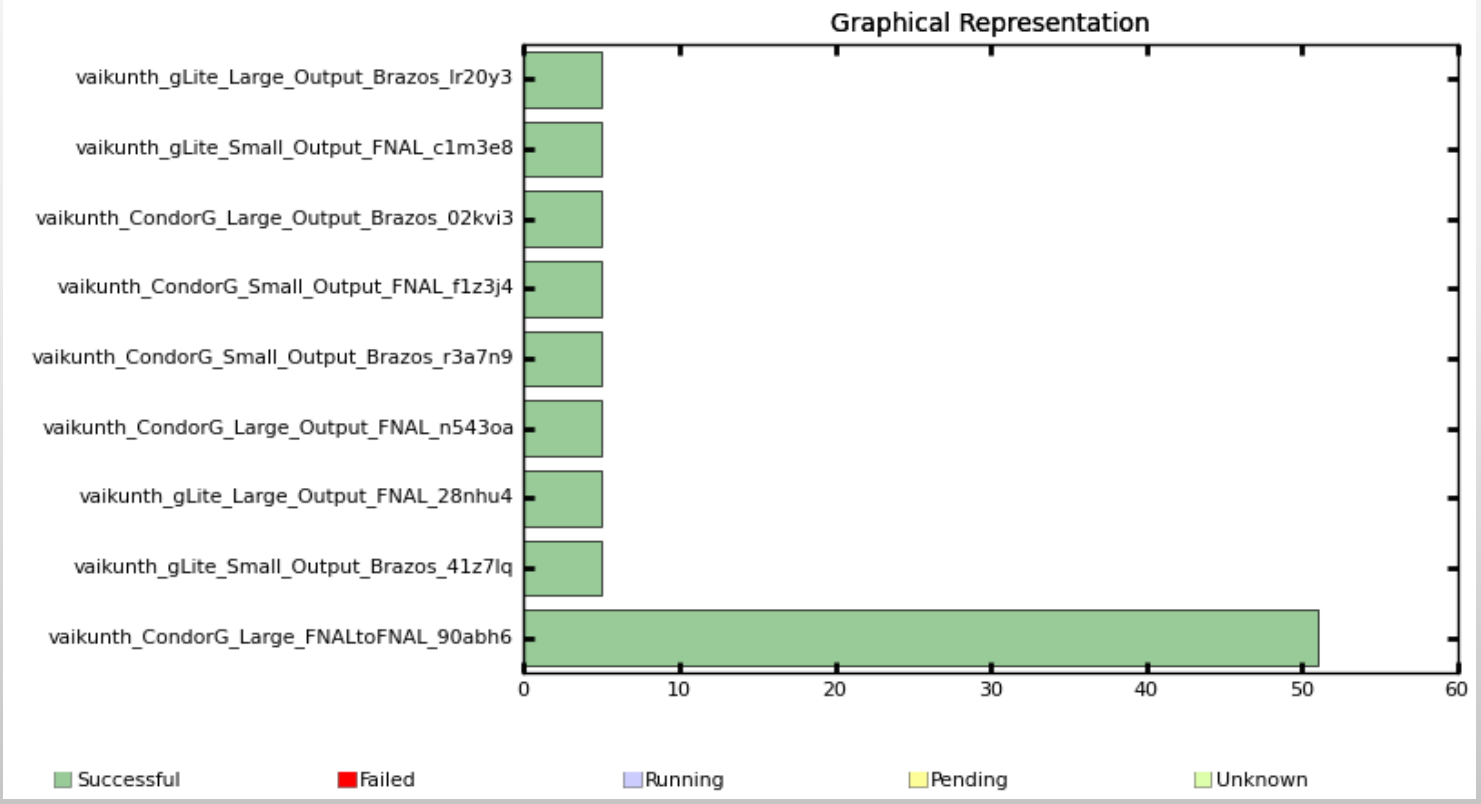
Note that this is important for self-imposed quotas. Need to know if we are keeping below our 30TB allocation. Will expand to 50TB soon. Will eventually be sending email if we get near our limit.

Running CRAB jobs on BRAZOS

- Have set up two fully functional ways for the convenience of users
  - condor_g – run jobs locally
  - gLite - can submit from anywhere in the world!
- More as needed
  - PBS - in the process of making this work
- Have created standard test jobs
  - **C**RAB **A**dmin **T**est **S**uite (CATS?) - These test both condor_g and gLite, output to FNAL and Brazos, big and small outputs, as well as large numbers of jobs.

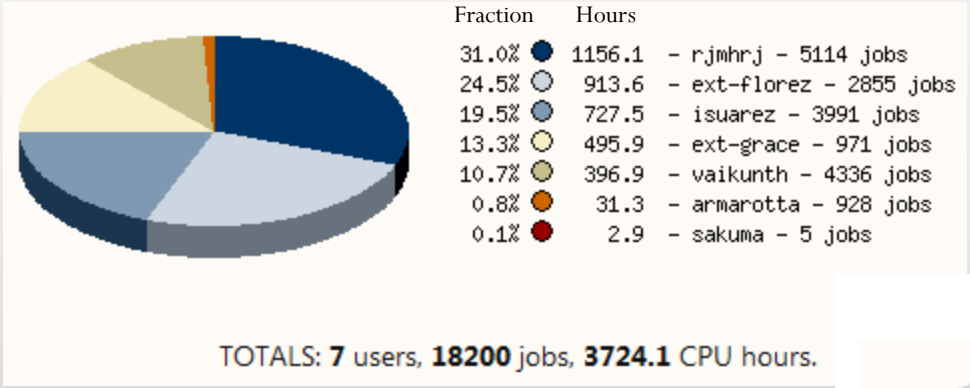
Current Status of CATS

- Validation test jobs (CATS) – All work
- Working on automating these to run periodically → Then add to monitoring site

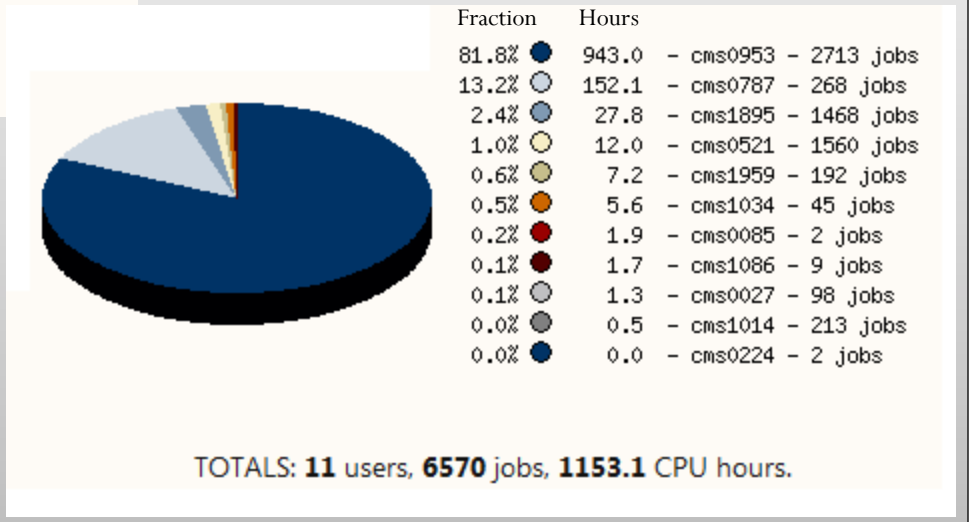


Already running LOTS of CRAB Jobs

- CRAB Usage by different groups
  - The A&M group has used a large amount of CPU



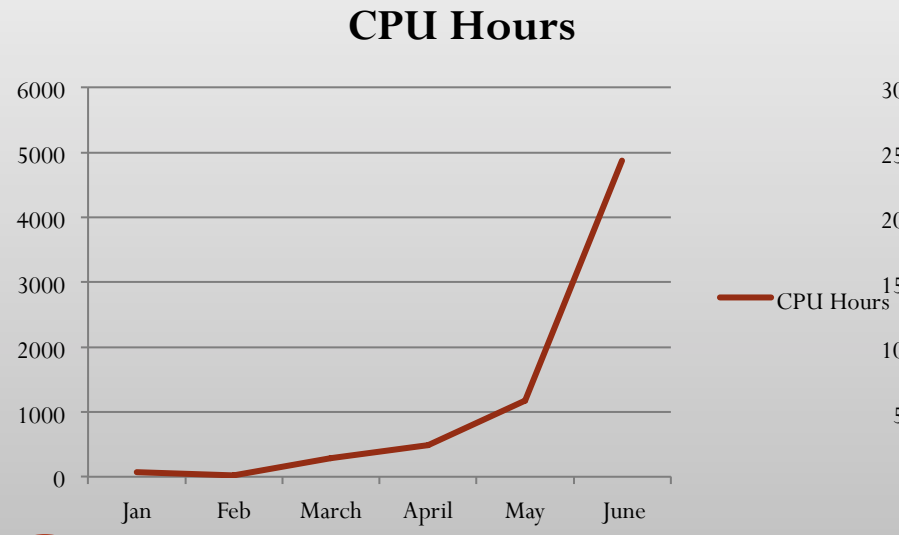
← HEPX group usage in June



Not all CPU available to use has been used by members of the Aggie Family →
We have provided a lot of CPU to outside users on CMS

More on CPU used and Jobs Run

- CRAB Performance
  - Job success rates have improved
  - Larger number of use-cases accommodates more jobs
  - Current trend indicates exponential usage, and we can still use MANY more CPU hours



Future Plans and Upgrades for Brazos/ Monitoring

- Add ability to run jobs via PBS
- Add more monitoring of jobs and CPU
- Add more disk
- Automate the running of CATS regularly, report results on Monitoring page
- Automate the checking of the monitoring to send mail on a failure or critical condition (disk space nearly fully, jobs failing, PhEDEx transfers failing, etc.)

Summary

- Grid Computing is a central part of CMS Analysis around the world
- Our own Grid Computing Center gives us high priority access to disk space and CPU which is an important competitive advantage in the search for Supersymmetry and the Higgs
- T3_US_TAMU at Brazos is fully functional and has already provided useful resources to the group
- We are constantly working to improve the monitoring of the system
- More resources will be added as we max out current ones

BACKUP SLIDES

Resources

- Future expansion
  - Adding 20 TB more to current disk space
  - Can continue to get more as needs increase
  - Possibly upgrade to a Tier 2 site
- The Brazos Team

Job:	Principal Investigators	CMS Admin	Grid/OSG Admin	Brazos SysAdmin
Who:	Dave Toback	Vaikunth Thukral	Steve Johnson	Many people
	Guy Almes	Daniel Cruz		brazos-help@listserve.tamu.edu
Responsibility:	Project Leadership	CMSSW/CRAB	Compute Element	Security
	Future Planning	PhEDEx	Storage Element	Software/TORQUE
	Funding	Monitoring		Maintenance
From:	Physics/Brazos	Physics	Brazos	

Operation Procedures

- Submitting jobs at T3_US_TAMU
  - Many use cases tested extensively
  - Some cases work better than others
  - Best way to run tasks on Brazos: http://collider.physics.tamu.edu/tier3/best_practice
- An Example

```
[USER]
return_data             = 0
copy_data               = 1
storage_element         = T3_US_TAMU
ui_working_dir          = CondorG_Small_Output_Brazos
user_remote_dir         = CondorG_Small_Output
#storage_element        = cmssrm.fnal.gov
#storage_path           = /srm/managerv2?SFN=/11
#user_remote_dir        =/store/user/vaikunth/

publish_data            = 0
#publish_data_name      = TestSkimTest

#dbs_url_for_publication = https://cmsdbsprod.cern.ch:8443/cms_dbs_ph_analysis_01_writer/servlet/DBSServlet

[CRAB]
scheduler               = condor_g
use_server              = 0
jobtype                 = cmssw

[CONDORG]
globus_rsl              = (queue=hepx)


[GRID]
rb                      = CERN
proxy_server            = fg-myproxy.fnal.gov
virtual_organization    = cms
se_white_list           = T3_US_TAMU
```

Operation Procedures

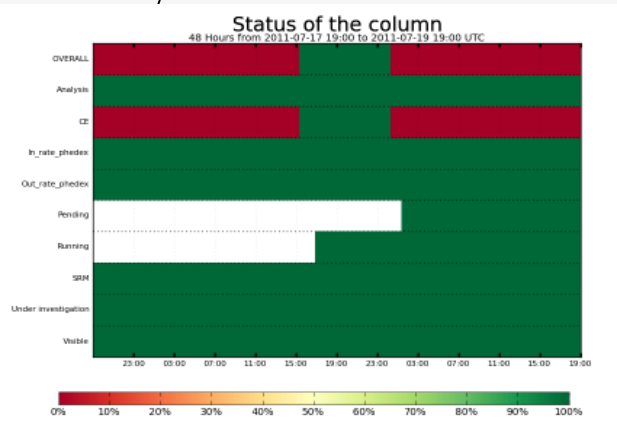
- Important Configuration Parameters
  - Schedulers
  - Datasets
  - White Lists
- Scheduler Options at T3_US_TAMU
  - Condor_g – Quick, suited for local submissions
  - gLite – Slower, suited for grid submissions
  - PBS – Currently being tested

Optimization

- Testing CRAB
  - Construct test jobs
  - These jobs test every aspect of the process
  - Currently run 8 test jobs
  - Particular cases tested:
    - Scheduler type
    - Output file size (Small/Large)
    - Local output destination
    - Remote output destination
    - Number of jobs (Small/Large)

Monitoring

- Tools Provided by the Grid
  - SAM tests



  - PhEDEx webpage

110 recent errors on links to T3_US_TAMU from .*

Please note that PhEDEx only stores the last 100 link errors to the database for a limited period of time. More errors may have occurred that are not visible here.

Page 1 of 11: 1 2 3 4 5 6 [Next] [Last]

Error 1

To Node:	T3_US_TAMU	**From Node:**	T1_DE_KIT_Buffer
Time Assigned:	2011-07-08 19:46:00 UTC (0m00 since assigned) (-41m39 from now)	**Time Transfer Start:**	2011-07-08 20:03:25 UTC (17m25 since assigned)
Time Exported:	2011-07-08 19:50:29 UTC (4m29 since assigned) (-37m10 from now)	**Time Transfer Done:**	2011-07-08 20:03:42 UTC (17m42 since assigned)
Time Pumped:	2011-07-08 20:03:04 UTC (17m04 since assigned) (-24m35 from now)	**Time Transfer Expires:**	2011-07-09 02:25:21 UTC (6h39 since assigned) (
Report Code:	2	**Transfer Code:**	1