PanDA for LSST/DESC: ImSim on Demand

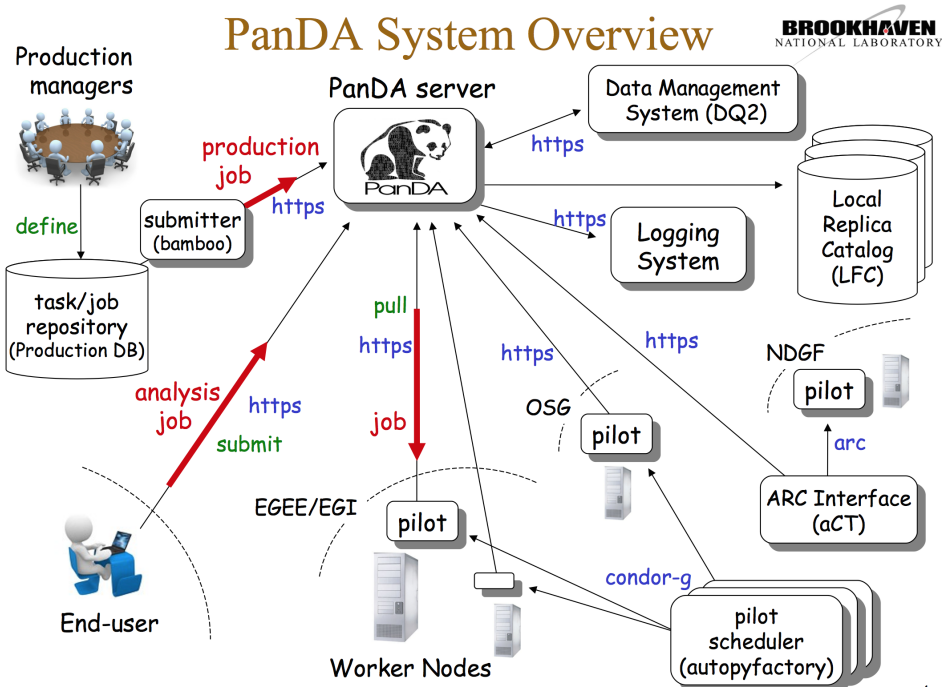
Torre Wenaus, Jarka Schovancová

BNL Astro Group Meeting

May 20, 2014

BROOKHAVEN NATIONAL LABORATORY

PanDA



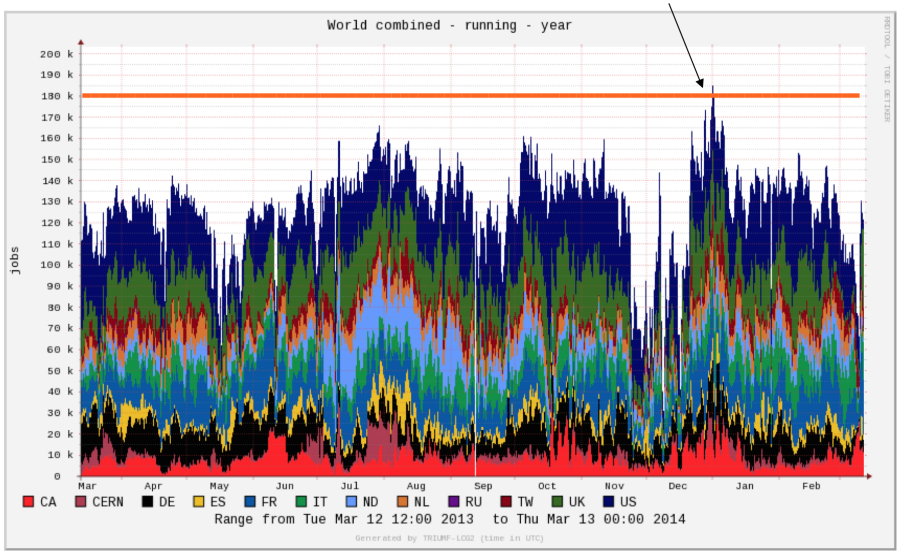
BNL's Physics Applications Software (PAS) group leads development of the PanDA workload management system with UT Arlington

PanDA manages processing and data workflows for large scale data intensive computing

- 2005: Initiated for US ATLAS
- 2008: Adopted ATLAS-wide
- 2009: First use beyond ATLAS
- 2012: ASCR funding for Exascale BigPanDA
- 2013-14: New PanDA based ATLAS prod system
- 2014: New Event Service fine grained processing
- 2014: PanDA community growing in HEP, NP, cosmology… US and international (supp. slides)

Global ATLAS PanDA operations:
180k concurrent jobs, 1.5M jobs/day peak
~1400 ATLAS physicist users
~140 sites around the world

~ 180 k running jobs at peak



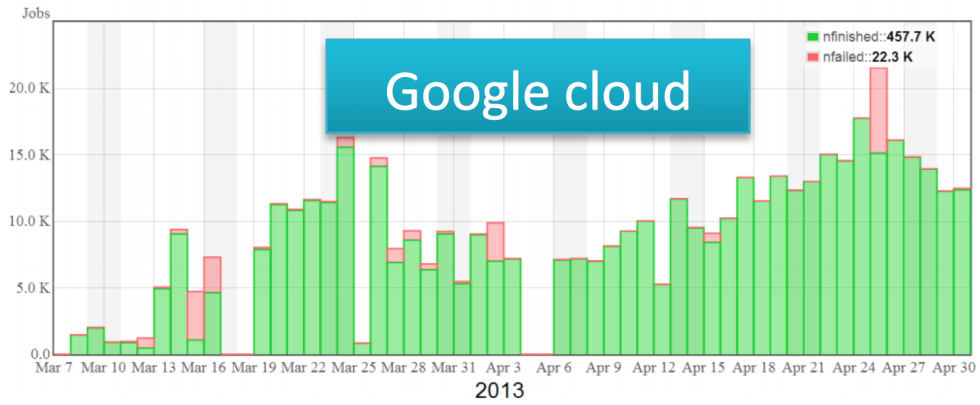
PanDA Beyond ATLAS

BigPanDA: Next Generation Workload Management and Analysis System for Big Data

- BNL PIs: Alexei Klimentov (lead), Sergey Panitkin, Torre Wenaus
  - ASCR + DOE HEP partnership funding a BNL/UT Arlington project
  - 3 FTEs funded FY12-FY14 (the work will carry over through FY15)
- Extending PanDA as a proven large-scale data-intensive workload management system to serve the exascale, big data communities
  - Lower the entry barrier for users to large scale processing
  - Close collaboration with ORNL on PanDA@Titan
  - **Supporting the collaboration with SLAC LSST DESC for PanDA@LSST/DESC**
- Turning PanDA into a generic product and service
  - PanDA refactoring, packaging for general use well advanced and ongoing
  - New generic PanDA instance in US, MySQL based, for OSG, others
    - Planning with OSG to evolve this into an OSG-hosted PanDA service
- Leveraging intelligent networking to make PanDA network-aware
  - Network awareness integrated into PanDA brokerage
- Work is well advanced, high marks from ASCR on progress in the latest briefing
- A challenge to keep up with growing interest

PanDA on Diverse Resources

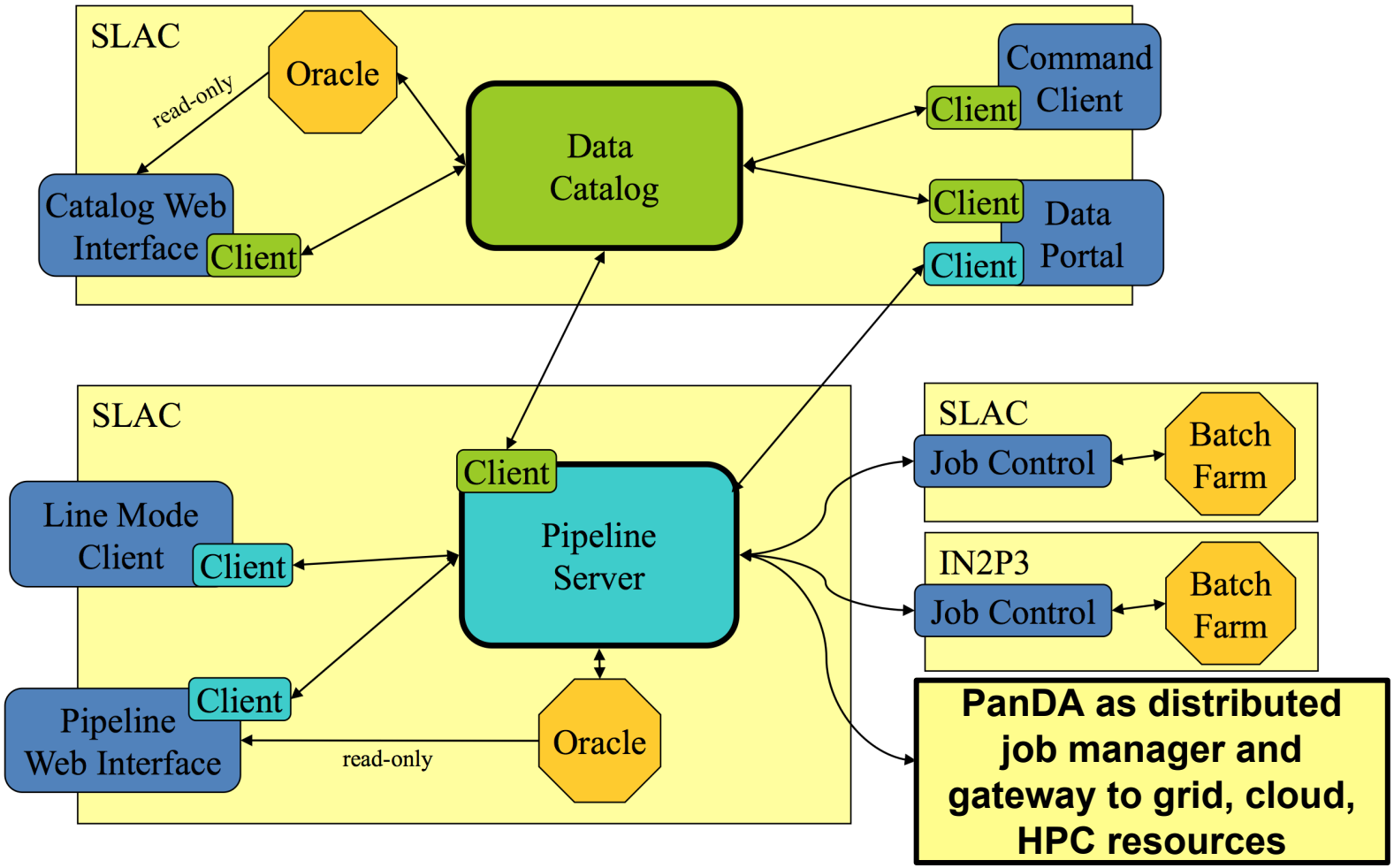
- Grid is the longstanding baseline
- Clouds in production: Amazon AWS, Google compute engine, research clouds (Openstack based)
  - Capability stands ready as cloud economics steadily improve
- Supercomputers have been/are being ported; many cycles available for crack-filling applications
  - ORNL Titan, ANL Mira, NERSC Hopper, XSEDE, EU machines
- Distributed storage via xrootd



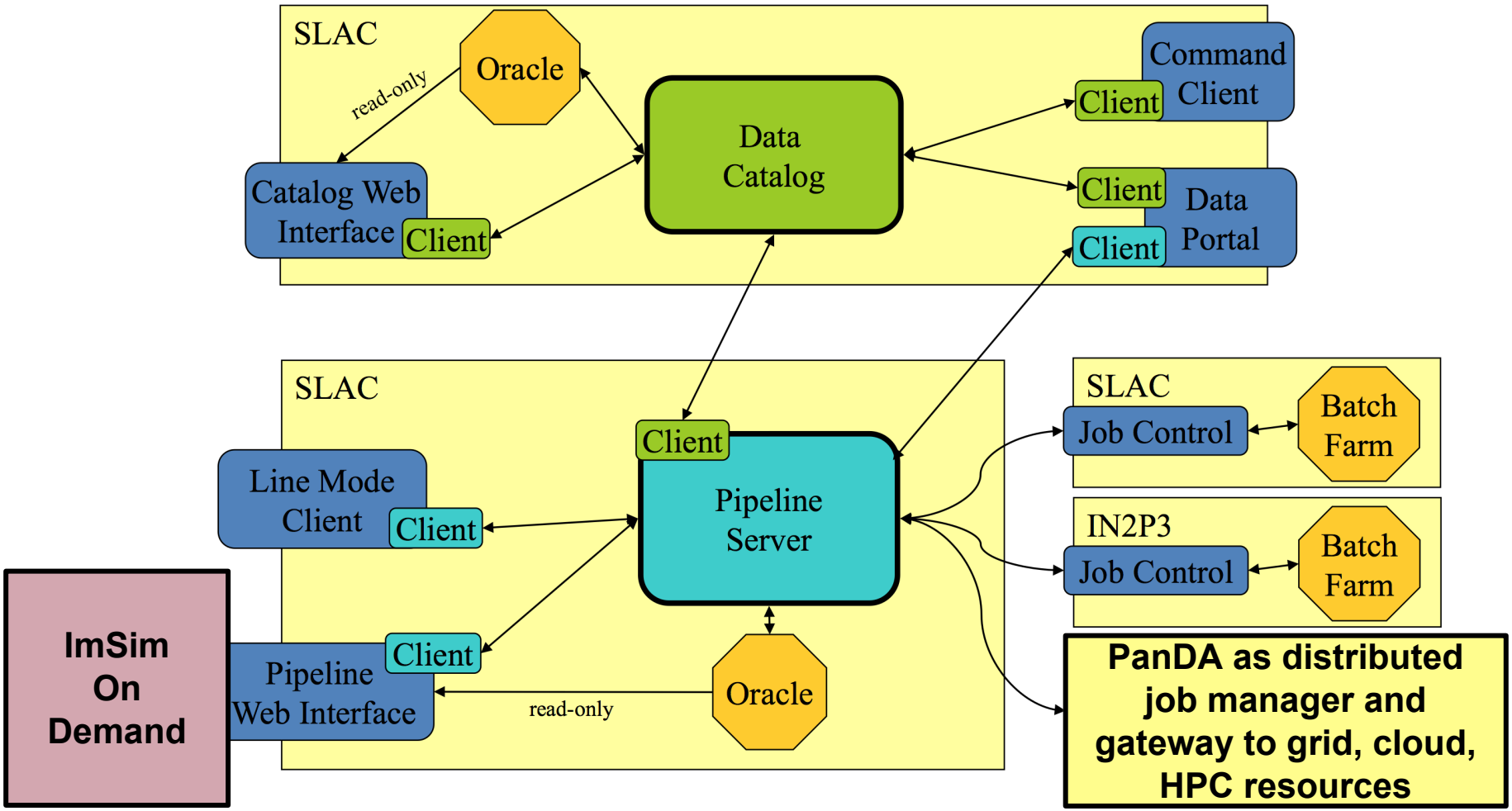
PanDA for LSST -- Background

- PanDA integrated with the SLAC DESC workflow pipeline in spring 2013, in a collaboration between BNL and SLAC (Tony Johnson the lead at SLAC), based on ATLAS PanDA
  - PanDA as gateway to diverse resources: initially BNL RACF and US ATLAS Tier 2 (s), later OSG-wide, HPCs, clouds, …
- Summer/Fall 2013: Generic PanDA service implemented in the Amazon cloud – reimplementation of PanDA using MySQL
  - Work begun to move the pipeline integration to this infrastructure
- Dec 2013: Pipeline + PanDA taken as the basis for ImSim On Demand, chosen by DESC as a 'highlight project' for spring/summer 2014; close collaboration with SLAC on development
  - Prospective further project: integrating SLAC data catalog with PanDA
- Winter 2013: progress temporarily derailed by an unexpected departure from BNL PanDA team
- Now: BNL+SLAC blitzing to have first version of ImSim on Demand working and demonstrated at the DESC collab meeting in June
  - Weekly meetings with SLAC, steady attention & progress at SLAC & BNL

SLAC DESC Workflow Pipeline + PanDA



SLAC DESC Workflow Pipeline + PanDA + ImSim On Demand



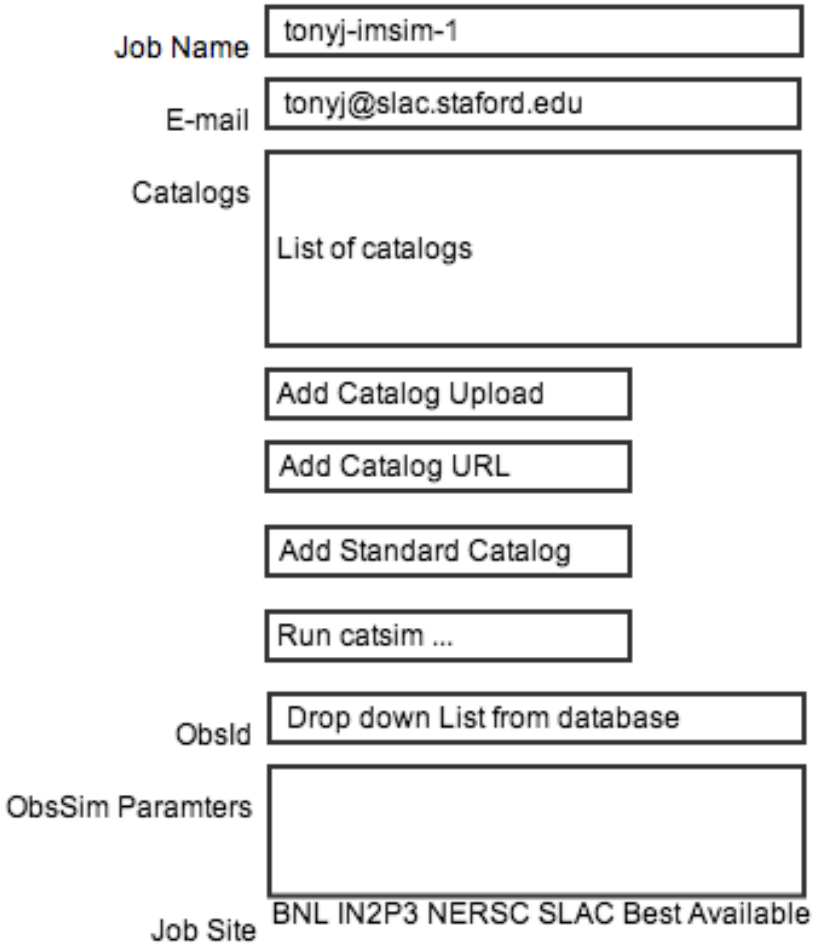
ImSim on Demand
(Richard Dubois @ Dec '13 DESC meeting, Pitt)

ImSim on Demand

- PhoSim is easy to set up and run on a small system
  - Much harder to manage production on thousands of cores and worse on big MPI machines
- Work towards a simple web interface for running ImSim on demand, supplying a catalogue and PhoSim parameters
  - Fire off jobs into the ether and notify the user when done (or if action is needed to rerun any failed jobs)
- Make use of existing workflow tools
  - Workflow engine to run graph of jobs
  - data catalogue to record where the data went
  - Interfaces to GRID, HPC, clouds etc to find the cycles

ImSim on Demand web interface mock-up

Seeking input from prospective users



Data Access: xrootd

- Distributed processing requires distributed data management
- PanDA is integrated with a HEP (and ATLAS) standard distributed data access tool called xrootd (SLAC developed)
  - Exposes data stores to distributed (as well as local) clients
  - Soon via http clients as well as via xrootd tools
- BNL astro's LSST disk space being made accessible via xrootd (Hiro Ito, RACF, in progress)
- Will allow PanDA to use this area as a data source/sink for LSST jobs, wherever they run
- Initially, data (file) cataloging is via PanDA's internal file catalog
- SLAC team working on an interface to their data catalog, integrated with the pipeline, that will allow PanDA to send file info to it

Software access: CERNVM File System (CVMFS)

- Critical for distributed processing is an easy means of distributing the software
- We now have one, thanks to the CERNVM (virtual machine) project: CVMFS
- Together with an OSG-developed layer on top, OASIS
- Allows one central installation of the software to be propagated to OSG sites, available from worker nodes
- Path: /cvmfs/oasis.opensciencegrid.org/lsst/
- First SW package installed: phosim 3.3.2

/cvmfs/oasis.opensciencegrid.org/lsst/test.lsst/bnl/phosim/3.3.2/phosim-3.3.2

/cvmfs/oasis.opensciencegrid.org/lsst/test.lsst/bnl/phosim/3.3.2/phosim-3.3.2-modif-oasis.tar.gz

phosim via PanDA

- Compile at BNL -> tarball in CVMFS -> access from a grid job
- Grid job test I
```
cd $PHOSIM_DIR
phosim ./examples/star -c ./examples/nobackground
```
  - Issue: memory consumption > 2 GB; Memory limit increased to 3 GB at BNL.
    - raytrace is the most expensive (RAM, CPU)
- In progress
  - Grid job test II

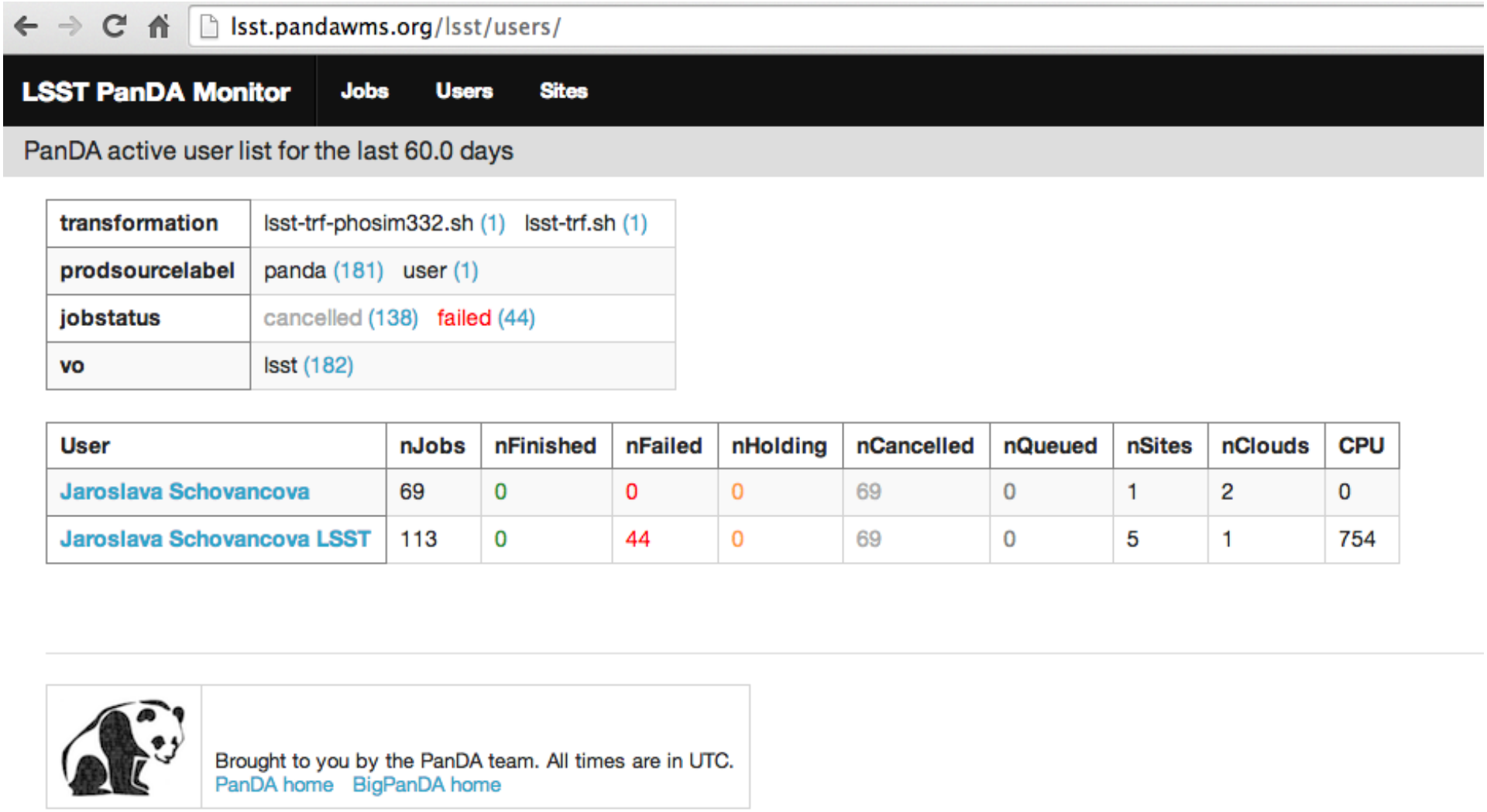
/nfs/slac/g/ki/ki06/lsst/djbard/LSST/IMSIM/1degrad-87393588/bdgals-fixed.dat

  - pipeline task with phosim
- Memory consumption
  - What is a realistic LSST phosim input?
  - What is usual memory consumption pattern?

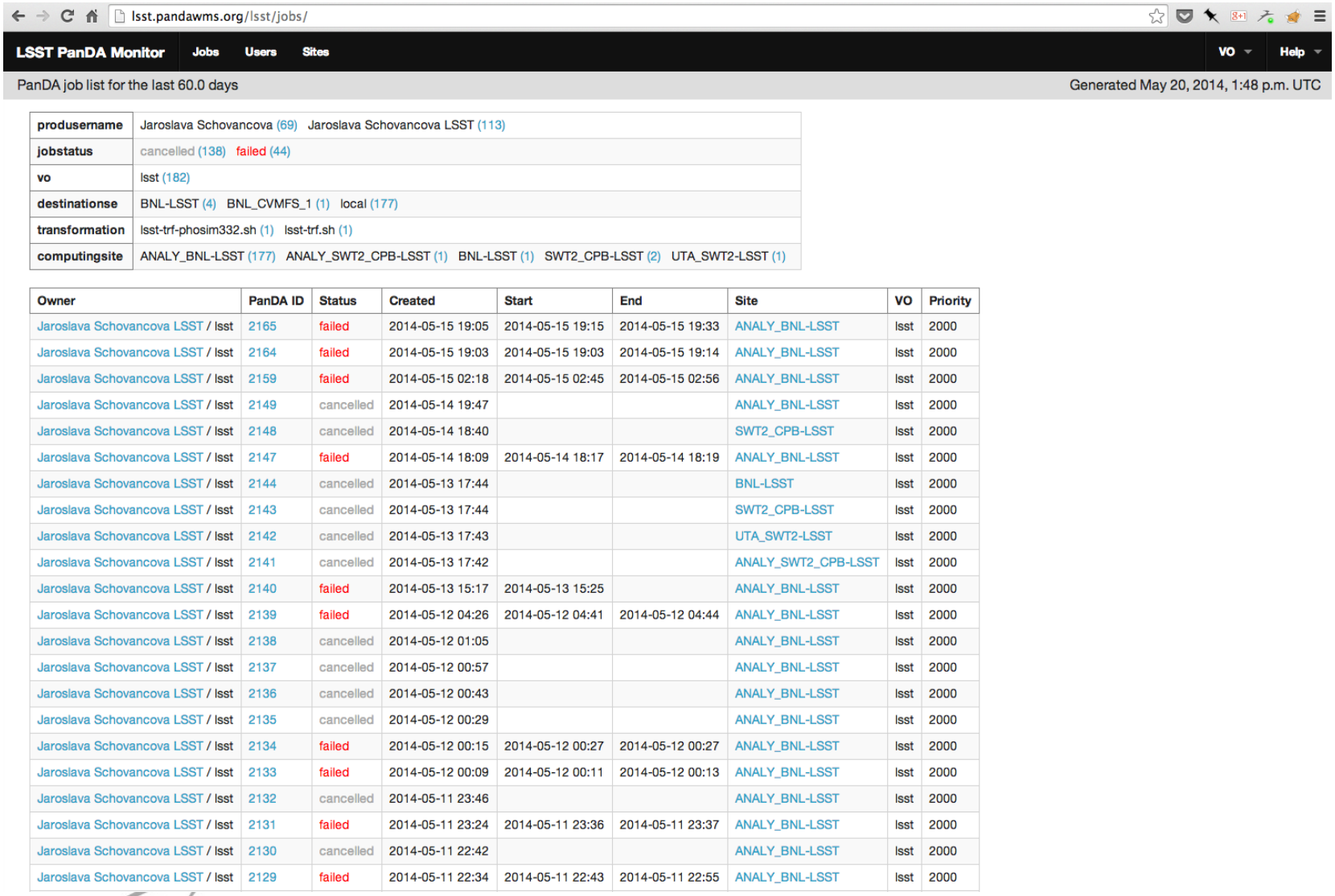
Processing Sites

- LSST jobs are managed as OSG (Open Science Grid) jobs, VO=lsst
- First available site: BNL RACF ATLAS resources
  - ~15k job slots; resources available to non-ATLAS fluctuates but can be substantial during quiet ATLAS periods (which in principle should not exist but they do)
- Second: US ATLAS Tier 2 Center(s), starting with UT Arlington
  - Demonstrator/debug site to validate that LSST processing is grid-enabled
  - When it is working, expand to other US ATLAS and OSG sites
- Third: IN2P3-CC plans to enable LSST PanDA processing
  - They are working on the pre-requisites (OASIS etc)
- And then… HPC (NERSC, …)

Monitor - Users



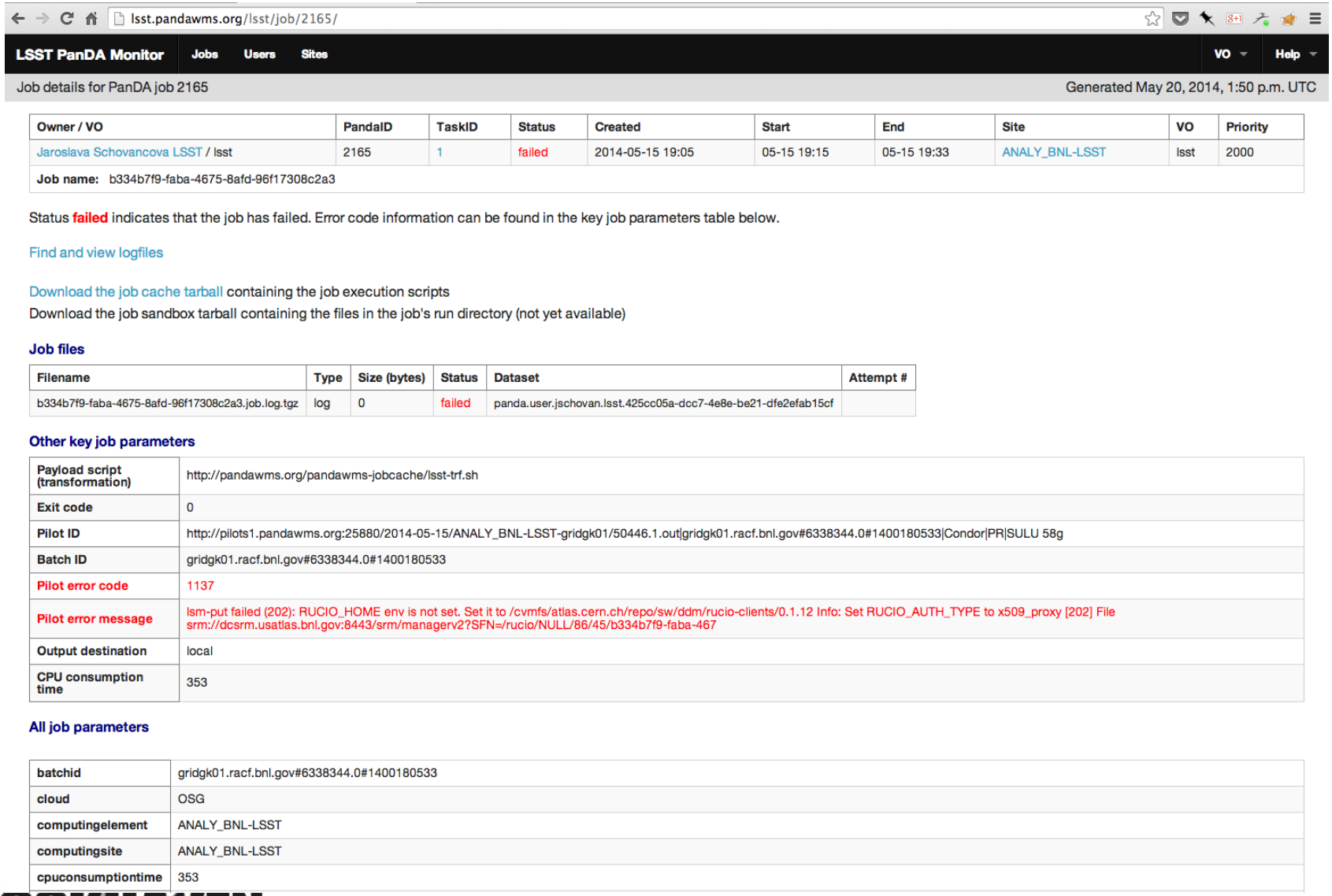
Monitor - Jobs



produsername	Jaroslava Schovancova (69) Jaroslava Schovancova LSST (113)
jobstatus	cancelled (138) failed (44)
vo	lsst (182)
destinationse	BNL-LSST (4) BNL_CVMFS_1 (1) local (177)
transformation	lsst-trf-phosim332.sh (1) lsst-trf.sh (1)
computingsite	ANALY_BNL-LSST (177) ANALY_SWT2_CPB-LSST (1) BNL-LSST (1) SWT2_CPB-LSST (2) UTA_SWT2-LSST (1)

Owner	PanDA ID	Status	Created	Start	End	Site	VO	Priority
Jaroslava Schovancova LSST / lsst	2165	failed	2014-05-15 19:05	2014-05-15 19:15	2014-05-15 19:33	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2164	failed	2014-05-15 19:03	2014-05-15 19:03	2014-05-15 19:14	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2159	failed	2014-05-15 02:18	2014-05-15 02:45	2014-05-15 02:56	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2149	cancelled	2014-05-14 19:47			ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2148	cancelled	2014-05-14 18:40			SWT2_CPB-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2147	failed	2014-05-14 18:09	2014-05-14 18:17	2014-05-14 18:19	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2144	cancelled	2014-05-13 17:44			BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2143	cancelled	2014-05-13 17:44			SWT2_CPB-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2142	cancelled	2014-05-13 17:43			UTA_SWT2-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2141	cancelled	2014-05-13 17:42			ANALY_SWT2_CPB-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2140	failed	2014-05-13 15:17	2014-05-13 15:25		ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2139	failed	2014-05-12 04:26	2014-05-12 04:41	2014-05-12 04:44	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2138	cancelled	2014-05-12 01:05			ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2137	cancelled	2014-05-12 00:57			ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2136	cancelled	2014-05-12 00:43			ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2135	cancelled	2014-05-12 00:29			ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2134	failed	2014-05-12 00:15	2014-05-12 00:27	2014-05-12 00:27	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2133	failed	2014-05-12 00:09	2014-05-12 00:11	2014-05-12 00:13	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2132	cancelled	2014-05-11 23:46			ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2131	failed	2014-05-11 23:24	2014-05-11 23:36	2014-05-11 23:37	ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2130	cancelled	2014-05-11 22:42			ANALY_BNL-LSST	lsst	2000
Jaroslava Schovancova LSST / lsst	2129	failed	2014-05-11 22:34	2014-05-11 22:43	2014-05-11 22:55	ANALY_BNL-LSST	lsst	2000

Monitor - Job details



lsst.pandawms.org/lsst/job/2165/

LSST PanDA Monitor | Jobs | Users | Sites | VO | Help

Job details for PanDA job 2165 — Generated May 20, 2014, 1:50 p.m. UTC

Owner / VO	PandaID	TaskID	Status	Created	Start	End	Site	VO	Priority
Jaroslava Schovancova LSST / lsst	2165	1	failed	2014-05-15 19:05	05-15 19:15	05-15 19:33	ANALY_BNL-LSST	lsst	2000
Job name: b334b7f9-faba-4675-8afd-96f17308c2a3									

Status **failed** indicates that the job has failed. Error code information can be found in the key job parameters table below.

Find and view logfiles

Download the job cache tarball containing the job execution scripts
Download the job sandbox tarball containing the files in the job's run directory (not yet available)

Job files

Filename	Type	Size (bytes)	Status	Dataset	Attempt #
b334b7f9-faba-4675-8afd-96f17308c2a3.job.log.tgz	log	0	failed	panda.user.jschovan.lsst.425cc05a-dcc7-4e8e-be21-dfe2efab15cf	

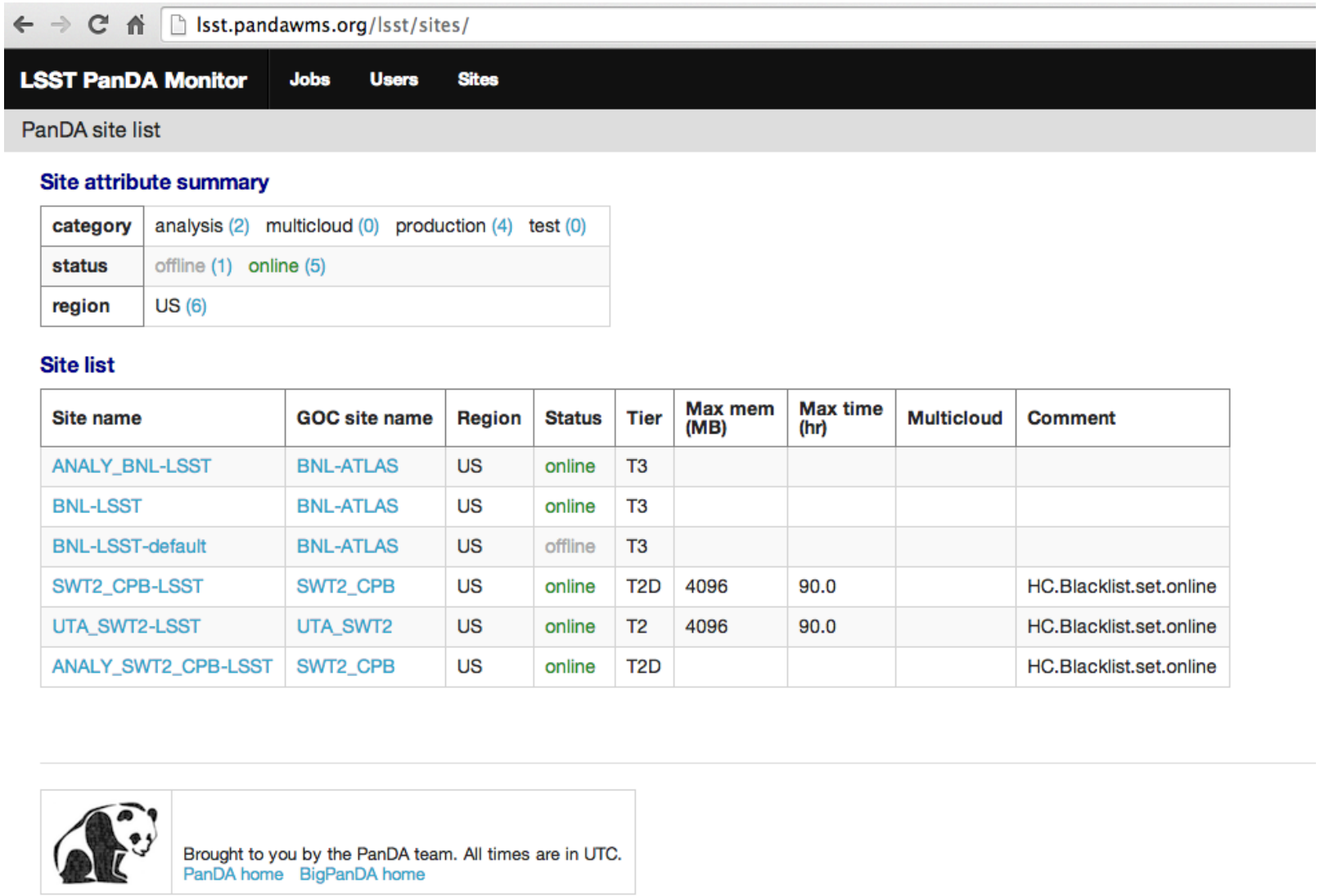
Other key job parameters

Payload script (transformation)	http://pandawms.org/pandawms-jobcache/lsst-trf.sh
Exit code	0
Pilot ID	http://pilots1.pandawms.org:25880/2014-05-15/ANALY_BNL-LSST-gridgk01/50446.1.out|gridgk01.racf.bnl.gov#6338344.0#1400180533|Condor|PR|SULU 58g
Batch ID	gridgk01.racf.bnl.gov#6338344.0#1400180533
Pilot error code	1137
Pilot error message	lsm-put failed (202): RUCIO_HOME env is not set. Set it to /cvmfs/atlas.cern.ch/repo/sw/ddm/rucio-clients/0.1.12 Info: Set RUCIO_AUTH_TYPE to x509_proxy [202] File srm://dcsrm.usatlas.bnl.gov:8443/srm/managerv2?SFN=/rucio/NULL/86/45/b334b7f9-faba-467
Output destination	local
CPU consumption time	353

All job parameters

batchid	gridgk01.racf.bnl.gov#6338344.0#1400180533
cloud	OSG
computingelement	ANALY_BNL-LSST
computingsite	ANALY_BNL-LSST
cpuconsumptiontime	353

Monitor - Sites



That's It! A work in progress.

- Your input on useful workloads to support would be most welcome
- It should be ready to try for any daring volunteers before the DESC collaboration meeting