LSST Simulations on OSG

Overview

- OSG Engagement of LSST
- LSST Simulation Workflow Requirements
- System Architecture
- Resource Utilization
- Operational Experience
- Results Validation

September 22, 2010

Parag Mhashilkar, Gabriele Garzoglio for the OSG Task Force on LSST
Computing Division, Fermilab

Introduction

- LSST at Purdue (Ian Shipsey) and OSG are collaborating to explore the use of the OSG to run LSST computations
- Integrated the current LSST image simulation with OSG
- In September, Bo Xin (LSST Purdue) has…
  - produced 150 image pairs and is working on 350 more
  - validated the OSG production against one official reference (PT1) image

OSG “engagement” of LSST

- Goal of the effort: OSG to empower LSST to use OSG resources independently

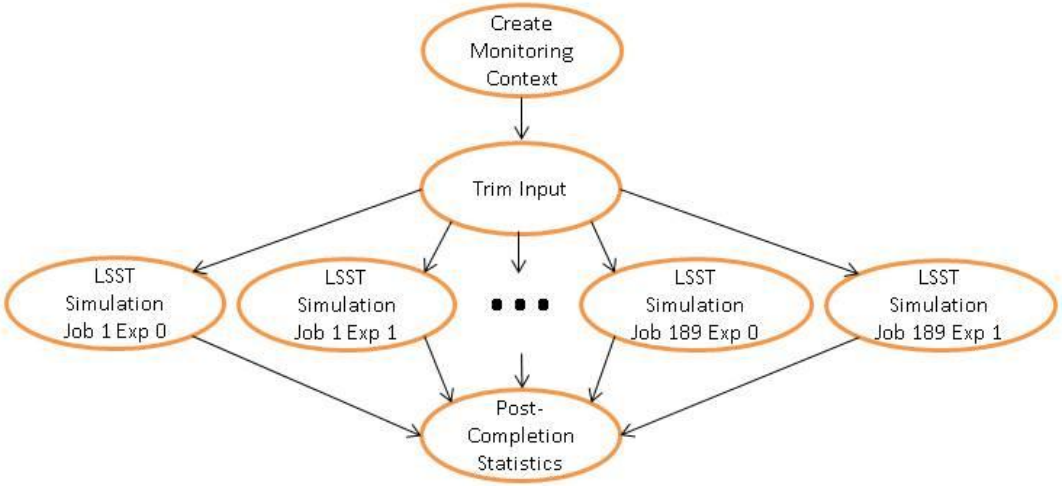
- Engagement phase: OSG provides experts for a limited time to help with
  - commissioning a submission system (software & resources) to run LSST workflows on OSG resources
  - supporting the integration of LSST applications with OSG
  - supporting the initial LSST operations

History

- Feb 2010 – Proof of principle of 1 LSST image simulated on OSG
- Jun 2010 – OSG EB forms an OSG task force for LSST
  - Goal of the project is to simulate one night of image taking (500 image pairs)
- Jul 2010 – Commissioning of the LSST submission system on OSG
  - Using an old application release (svn-11853), 1 person produced 183 times the same image in 1 day
- Aug 2010 – Integrated “current” LSST release (svn-16264) with the system
- Sep 2010 – LSST operator (Bo Xin) ran operations to produce 529 pairs.

Workflow Requirements

- LSST simulation of 1 image: 189 trivially parallel jobs for the 189 chips
- Input to the workflow:
  - SED catalog files: 15 GB uncompressed, pre-installed at all sites
  - Catalog files (SED files + wind speed, etc.): 500 MB compressed per image pair
- Workflow:
  - Trim catalog file into 189 chip-specific files
  - Submit 2 x 189 jobs: 1 image pair (same image w/ 2 exposures)
- Output: 2 x 189 FITS files, 10 MB each compressed

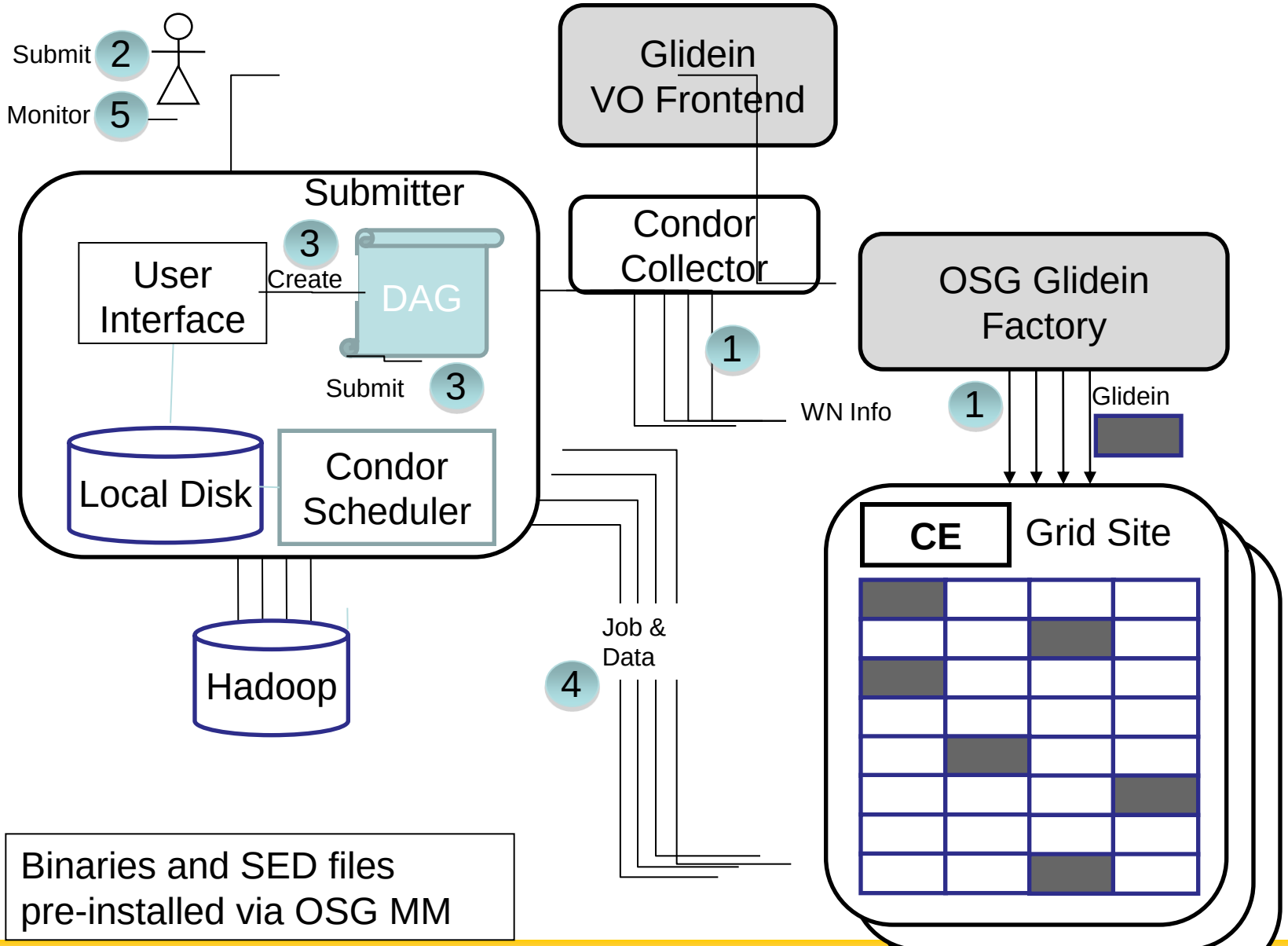


Production by Numbers

- Goal: simulate 1 night of LSST data collection: 500 pairs
- 200k simulation jobs (1 chip at a time) + 500 trim jobs
- Assume 4 hours / job for trim and simulation (over-est.)
  → 800,000 CPU hours
- Assume 2000 jobs DC → ~50,000 CPU hours / day

- 17 days to complete (w/o counting failures)
- 12,000 jobs / day i.e. 31 image pairs / day
- 50 GB / day of input files moved (different for every job)
- 300 GB / day of output
- Total number of files = 400,000 (50% input - 50% output)
- Total output compressed = 5.0 TB (25 MB per job)

Architecture



Operations

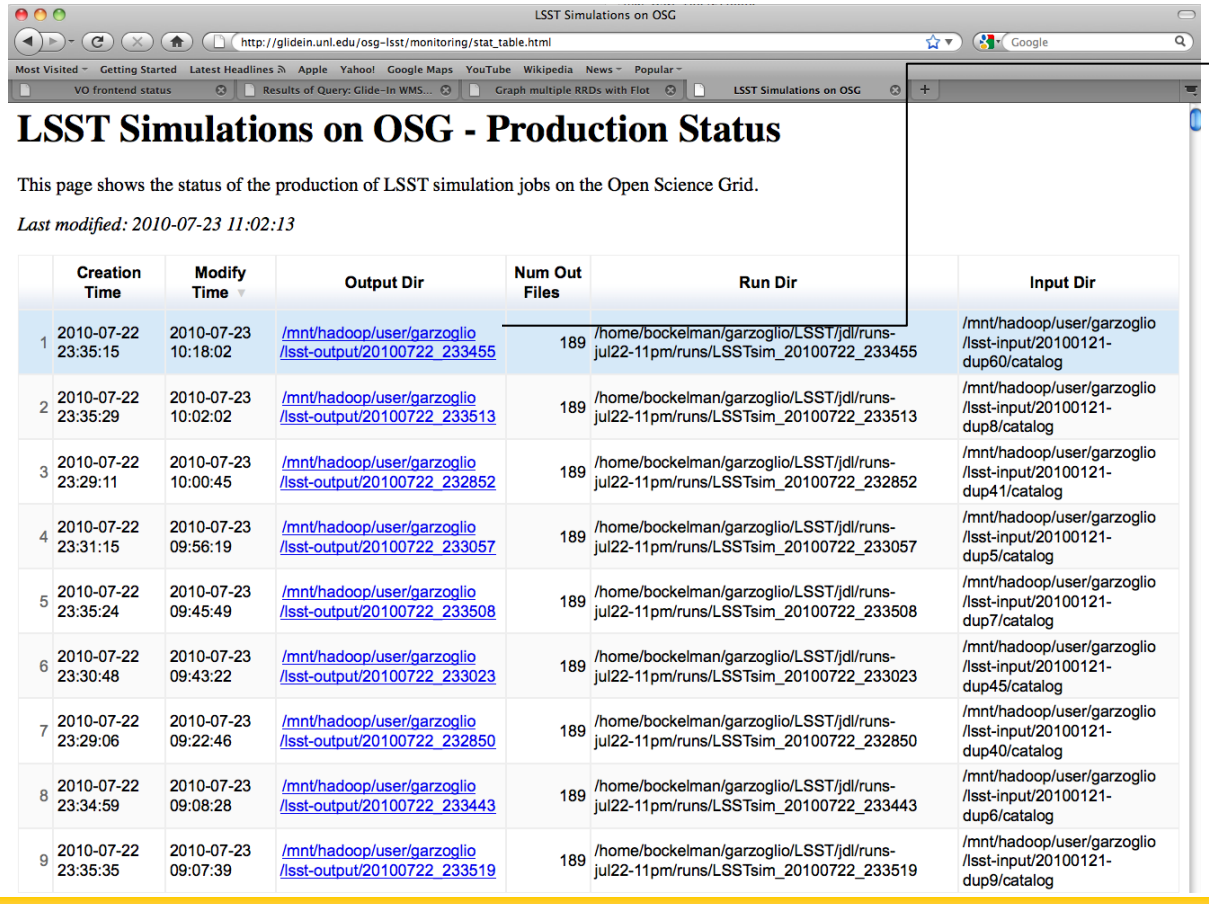
- Current setup: Central machine hosts
  - glideinWMS VO Frontend
  - User Pool (collector), Condor Scheduler
  - Unmanaged dedicated storage

- If LSST became a stand-alone VO
  - it could benefit from managed public storage
  - upload input data to OSG and download output data from OSG public storage
  - data maintenance could be done through metadata (e.g. remove trim and catalog input files for all workflow that have an output)
  - data processing would not depend on the availability of a single resource

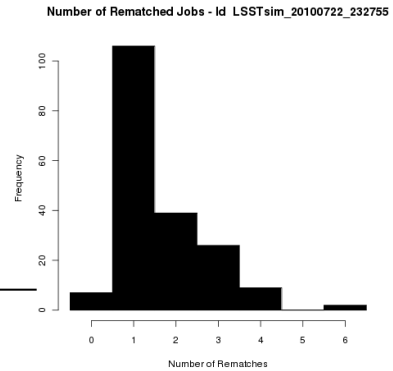
Monitoring Operations

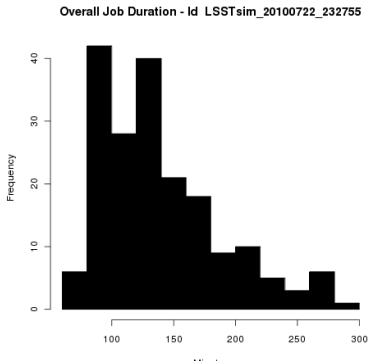
- If job exits due to system failure → automatically resubmit indefinitely
- If job exits due to app. failure → resubmit 5 times

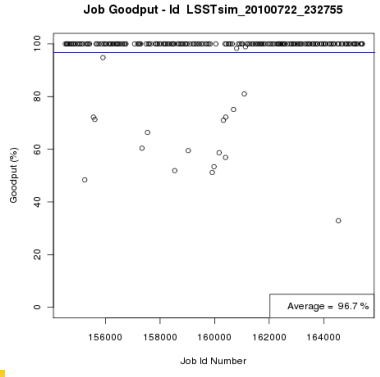
Periodically look at the output dir



	Creation Time	Modify Time	Output Dir	Num Out Files	Run Dir	Input Dir
1	2010-07-22 23:35:15	2010-07-23 10:18:02	/mnt/hadoop/user/garzoglio /lsst-output/20100722_233455	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_233455	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup60/catalog
2	2010-07-22 23:35:29	2010-07-23 10:02:02	/mnt/hadoop/user/garzoglio /lsst-output/20100722_233513	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_233513	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup8/catalog
3	2010-07-22 23:29:11	2010-07-23 10:00:45	/mnt/hadoop/user/garzoglio /lsst-output/20100722_232852	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_232852	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup41/catalog
4	2010-07-22 23:31:15	2010-07-23 09:56:19	/mnt/hadoop/user/garzoglio /lsst-output/20100722_233057	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_233057	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup5/catalog
5	2010-07-22 23:35:24	2010-07-23 09:45:49	/mnt/hadoop/user/garzoglio /lsst-output/20100722_233508	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_233508	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup7/catalog
6	2010-07-22 23:30:48	2010-07-23 09:43:22	/mnt/hadoop/user/garzoglio /lsst-output/20100722_233023	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_233023	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup45/catalog
7	2010-07-22 23:29:06	2010-07-23 09:22:46	/mnt/hadoop/user/garzoglio /lsst-output/20100722_232850	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_232850	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup40/catalog
8	2010-07-22 23:34:59	2010-07-23 09:08:28	/mnt/hadoop/user/garzoglio /lsst-output/20100722_233443	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_233443	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup6/catalog
9	2010-07-22 23:35:35	2010-07-23 09:07:39	/mnt/hadoop/user/garzoglio /lsst-output/20100722_233519	189	/home/bockelman/garzoglio/LSST/jdl/runs-jul22-11pm/runs/LSSTsim_20100722_233519	/mnt/hadoop/user/garzoglio /lsst-input/20100121-dup9/catalog

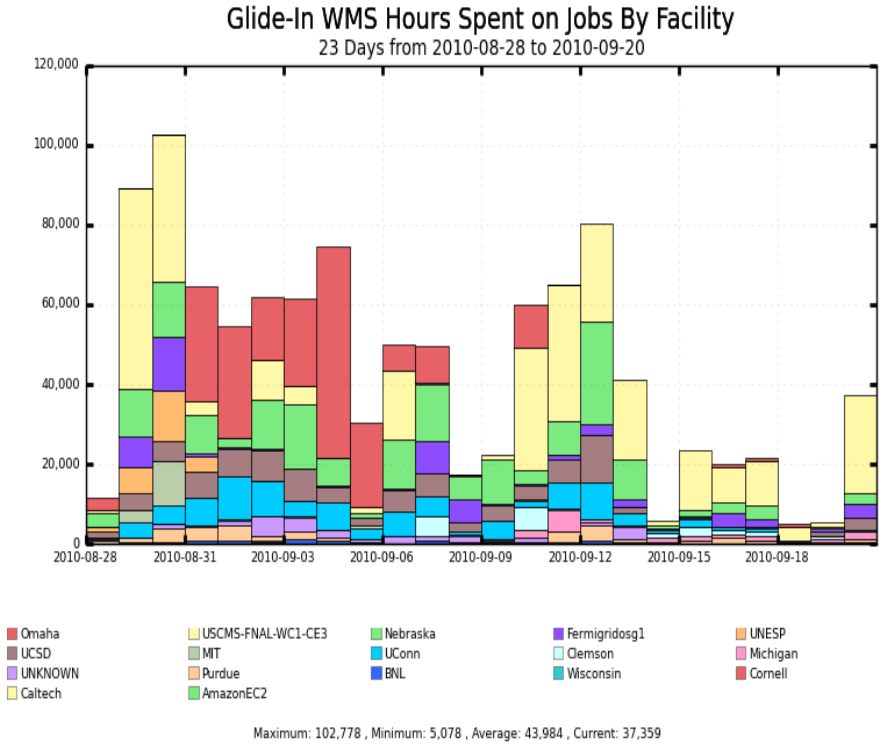




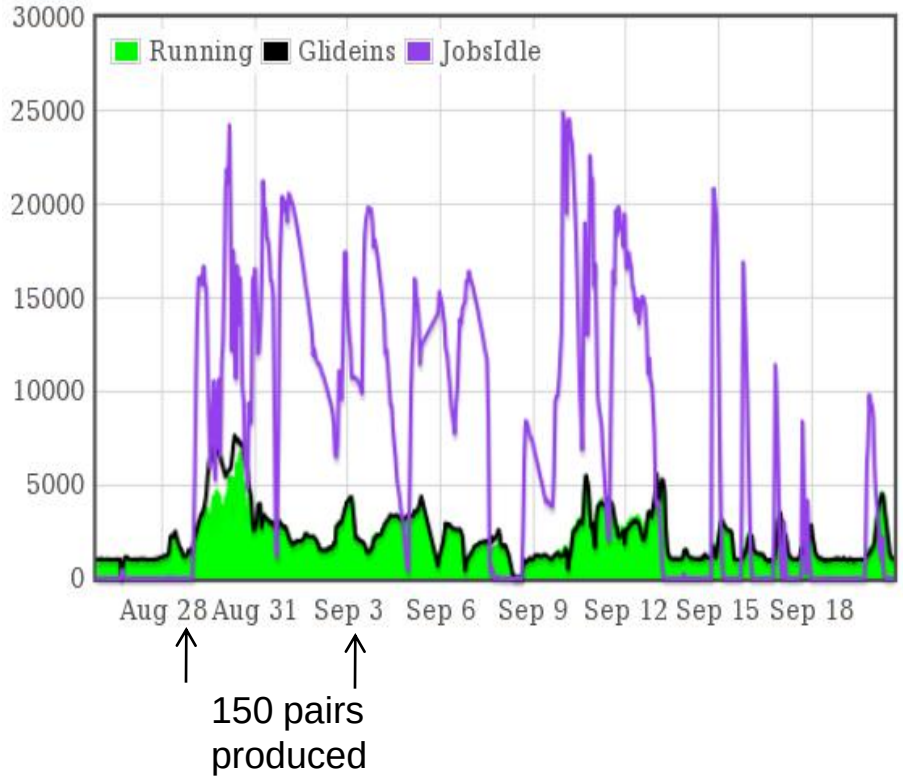


Resource Utilization

- By September 3, produced 150 pairs in 5 days using 13 sites.
- Now 400 / 529 pairs are produced (some chips job may require recovery)

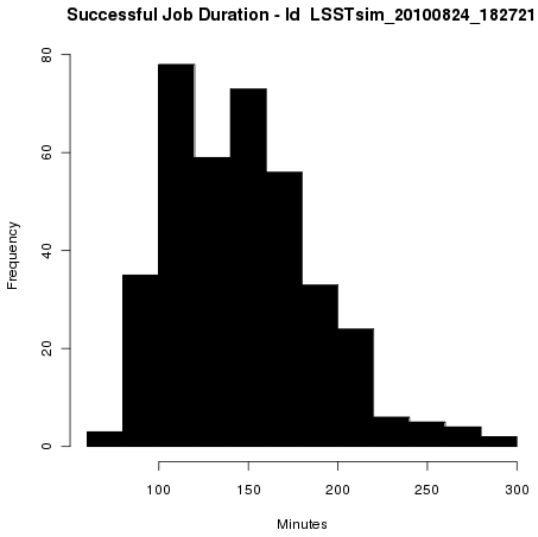


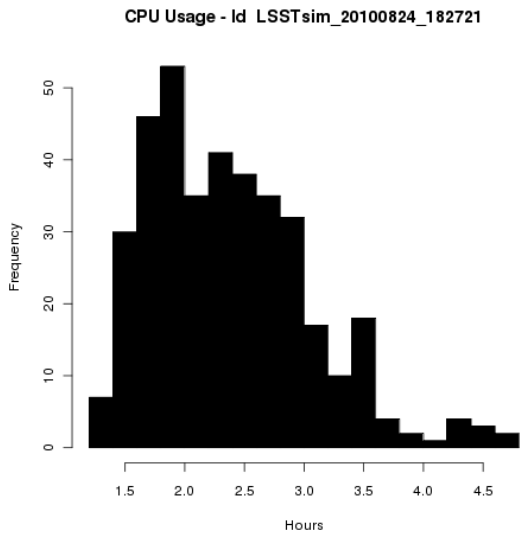
Gratia Resource Utilization plots

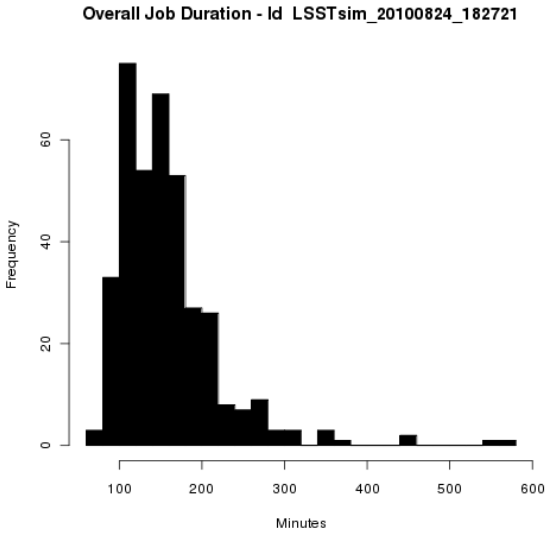


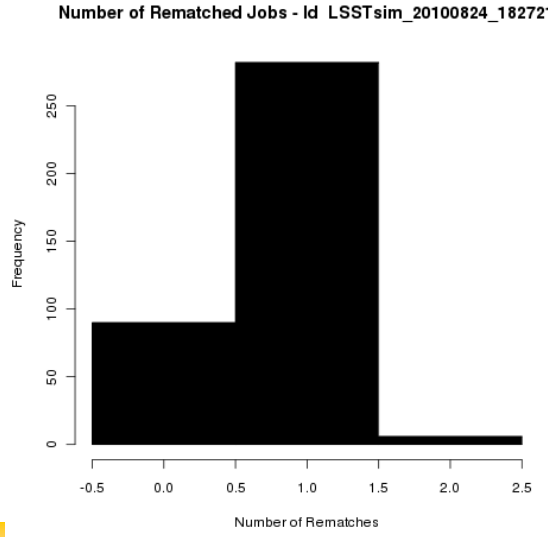
Frontend Status: Jobs & Glideins

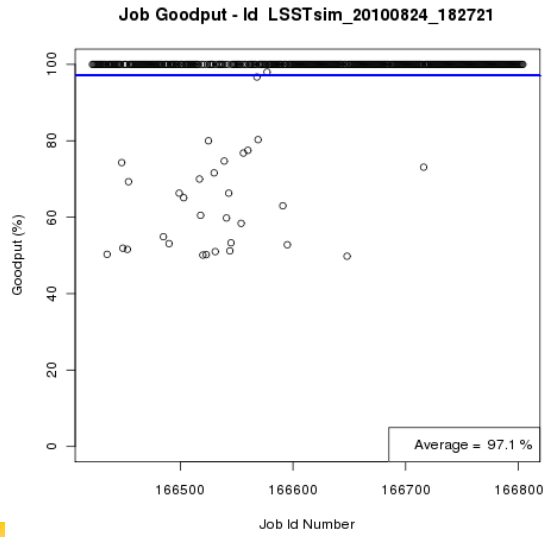
Typical Workflow Statistics









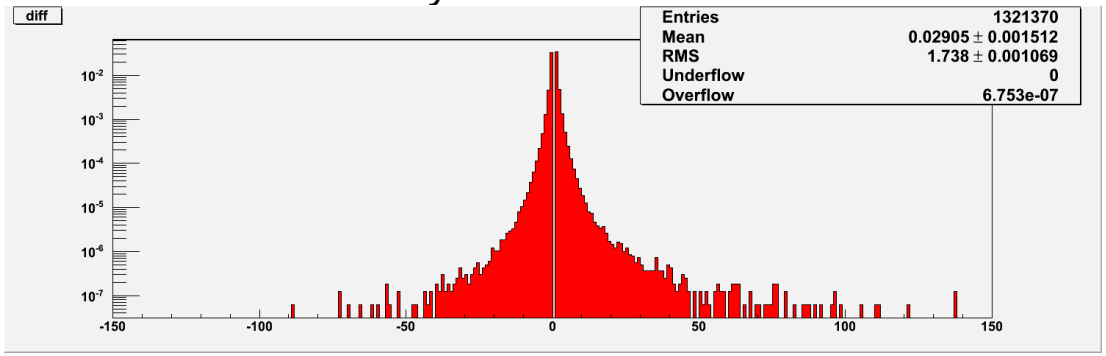


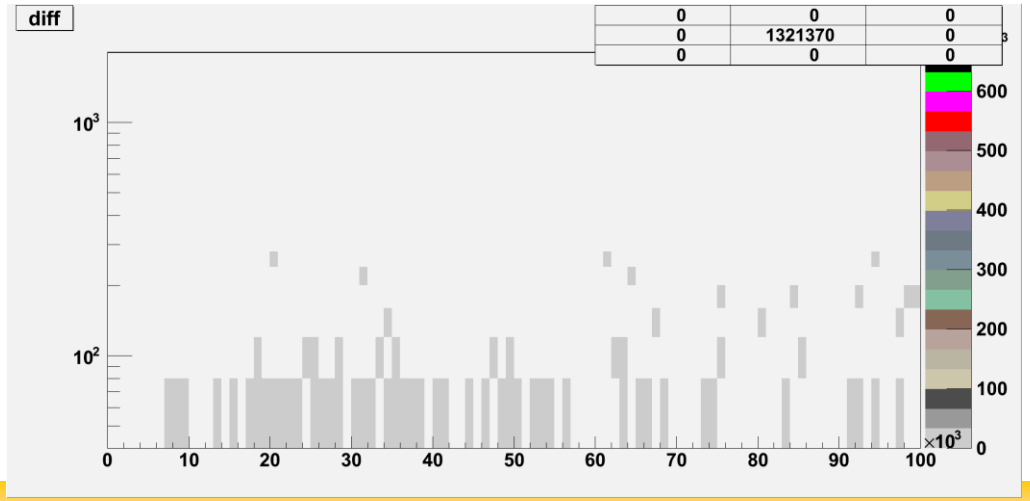
Operational Challenges

- LSST binaries were not Grid-ready
  - Application assumed writable software distribution
  - Application assumed path-lengths too short for the Grid
  - Orchestration script did not exit with failure upon error (required manual recovery until fixed)
- Typical failures at sites:
  - Job required more memory than the batch system allotted
  - Storage unavailable due to maintenance at some of the most productive sites
- Limited disk quota on the submission machine
- After the production of 150 pairs, the operator was mostly traveling and had limited time to dedicate to the operations

Validation: comparison with PT1

- **Validation Mechanism:**
  - Subtract an image pixel by pixel from a reference image.
  - Produced 2 different images twice on OSG, each on a different mix of resources.
  - In both cases, the pixel-subtraction is consistently 0.
- We compared 1 image pair (378 chips) with an "official" reference (PT1).
- For 374 chips the pixel-subtraction is consistently 0.
- For 4 chips it is negligible*.





More investigations are under way. Study done by Bo Xin.

Conclusions

- This project has demonstrated that OSG is a valuable platform to simulate LSST images
- 1 LSST person, Bo Xin, with expert support has produced 150 image pairs in 5 days, using in average 50,000 CPU h / day.
- Bo is currently finishing the production of 529 pairs (400 done with possible needs for recovery)
- Image simulation is the first application integrated with OSG. The integration of data-intensive applications might follow
- If LSST became a stand-alone VO, it could benefit from managed public storage supported by OSG
- Thank you to the OSG Task Force
  - Especially to Brian Bockelman and Derek Weitzel