ALICE USA Computing Project

Operations Status & Plans

ALICE T1/T2 Workshop

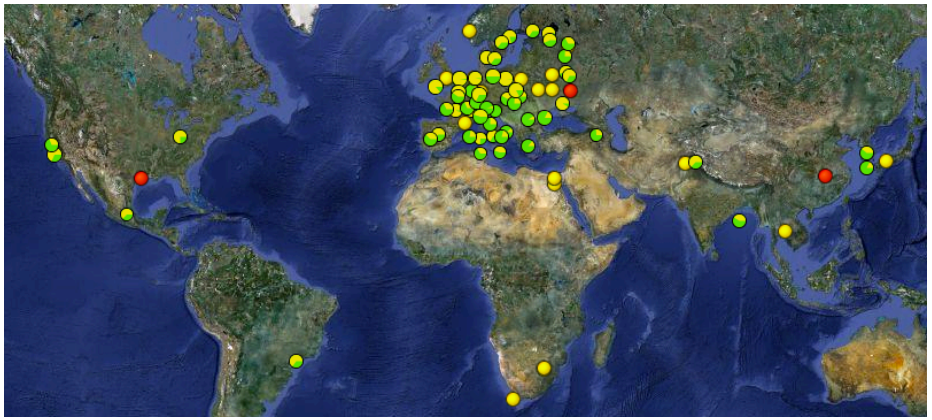
Outline

- US Computing project
- Facility snapshots & site configurations
- Resource utilization & performance
  - CPU utilization
  - Storage Elements
- Upcoming hardware tasks
- Upcoming operational tasks
- Longer term plans

Project Organization

- Goal: high-performance, cost-efficient ALICE computing facility
  - Fulfills MoU-based ALICE USA obligations for computing & storage resources to ALICE
  - ALICE USA participation is about 7-8% of ALICE

- 2009 Project Proposal
  - Operate facilities at two DOE labs
    - NERSC/PDSF at LBNL
    - Livermore Computing (LC) at LLNL
  - In operation since Summer 2010



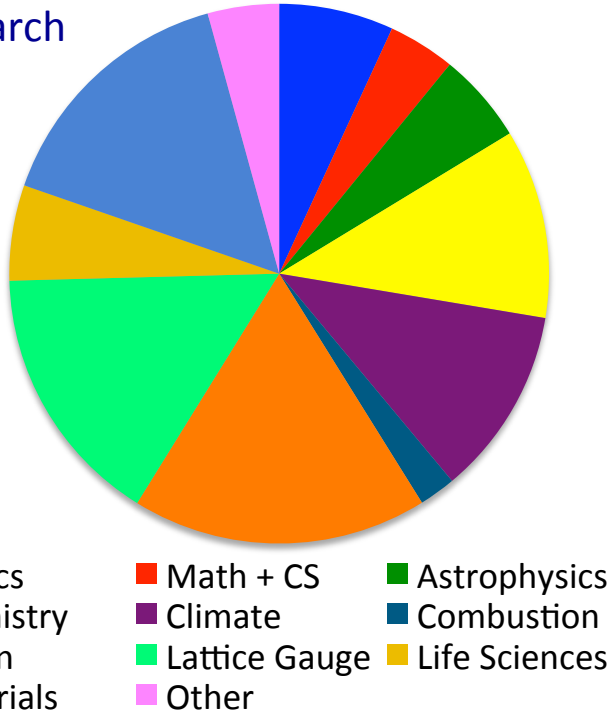
- Project personnel on steering/operations committee
  - Jeff Porter – project manager & ALICE Grid Manager for NERSC
  - Ron Soltz - Former Computing Coordinator & LLNL ALICE Rep.
  - Jeff Cunningham – LLNL System Admin and ALICE Grid Manager for LC-glcc
  - Iwona Sakrejda – PDSF project lead
  - Bjorn Nilsen – ALICE-USA contributor

Facility Snapshot: LLNL/LC

- Livermore Computing
  - Large & diverse institutional-based High Performance Computing Center
  - Supports Lab Science and Engineering activities
- Cost effective procurement and operations model
  - Able to buy into routine very large purchases of scalable units
  - In-house managed OS (CHAOS) & other software (e.g. SLURM)
- ALICE Deployment model @ LLNL/LC
  - Separate single-use facility
    - 100% ALICE
    - Grid only use → no user logins
  - Large HW purchase, refreshed every 3-4 years

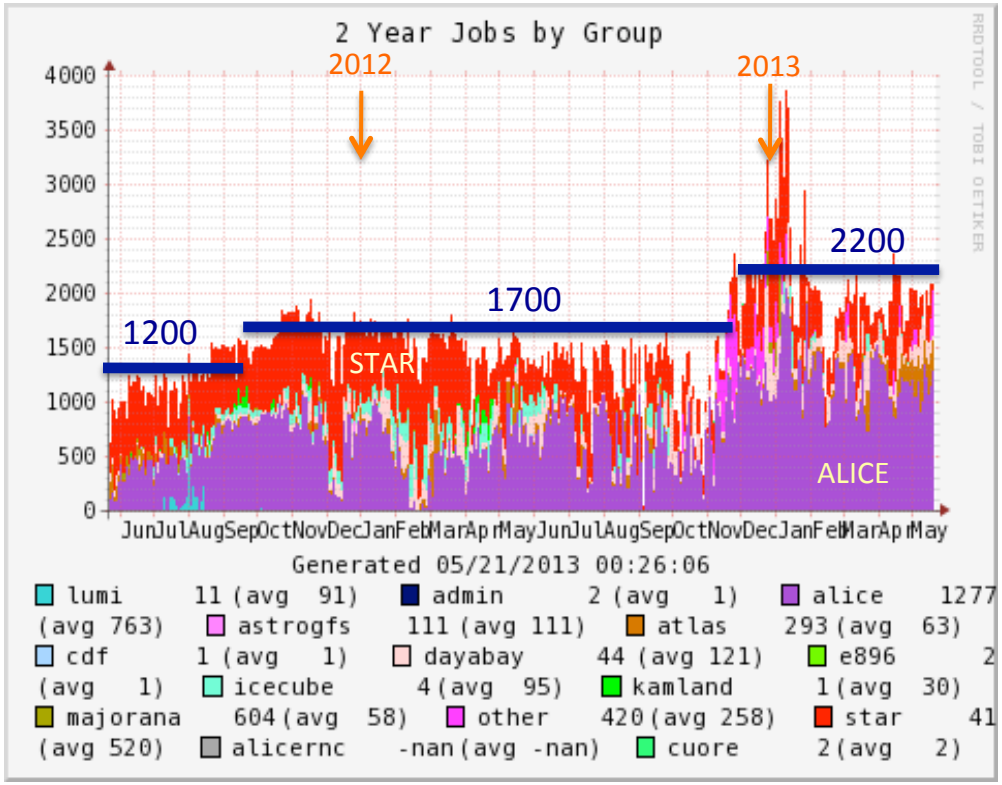
Facility Snapshot: NERSC

- NERSC: US Department of Energy (DOE) Office of Science Flagship High Performance Scientific Computing Center
  - Available to all DOE Office of Science sponsored research
- Computing for Scientific Research
  - Large HPC Systems (~200k cores)
  - Special Clusters: PDSF, Visualization,…
  - Large archival storage (HPSS)
  - Data Transfer, Gateway & OSG/Grid Services
  - Evaluation Systems: GPU & Cloud Services
- Extensive user support services
- ALICE Deployment Model @ NERSC
  - Project resources deployed on PDSF for ALICE Grid (see next slide)
  - Users can have login access with ALICE client tools available
  - Annual HW purchases to adjust to changing ALICE requirements



Facility Snapshot: PDSF

- Multi-group facility for Nuclear & High Energy Physics experiments
  - Allocations as "share" of resources
  - Fair share done in SGE (UGE)
- Share calculation includes
  - HW investment
  - FTE contribution
- Nuclear Science shares
  - ALICE 40%
  - STAR 30%
- Physics Div. shares
  - ATLAS T3 15%
  - Dayabay 10%

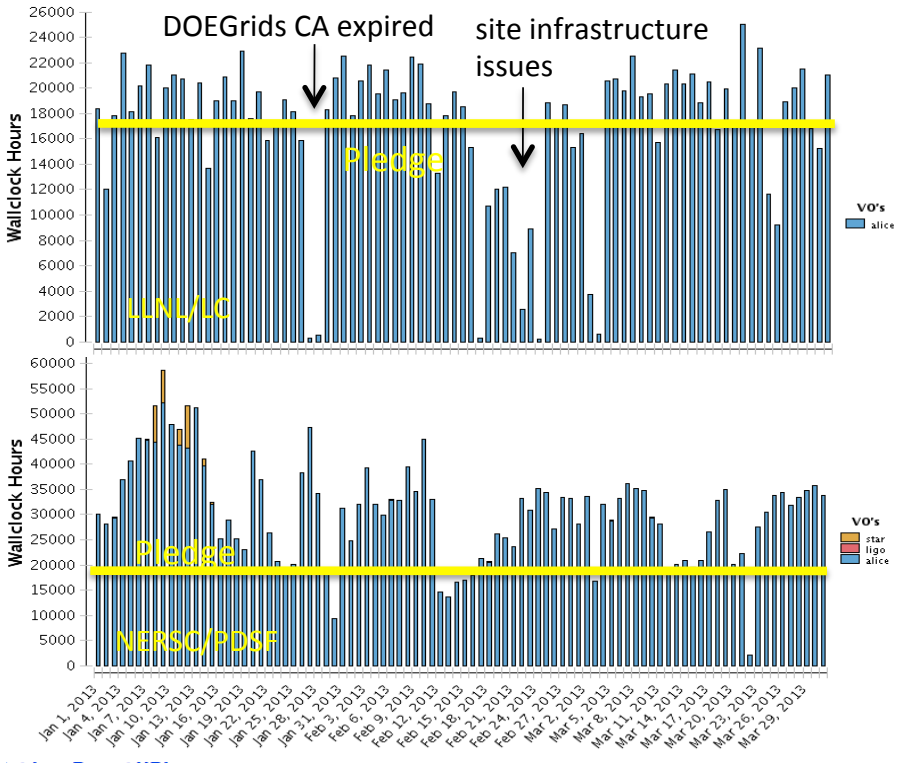


Running jobs

ALICE-USA & Open Science Grid

- ALICE-USA project leverages OSG capabilities
  - OSG Registration Authority
  - Resource reports sent to WLCG
    - Availability and Reliability Rep.
      - Critical services scans
    - Accounting Reports
      - Gratia site service
      - OSG central repository
  - Disposition of unused cpu @LLNL
    - OSG CE for DOE/NP funded groups

OSG Reports

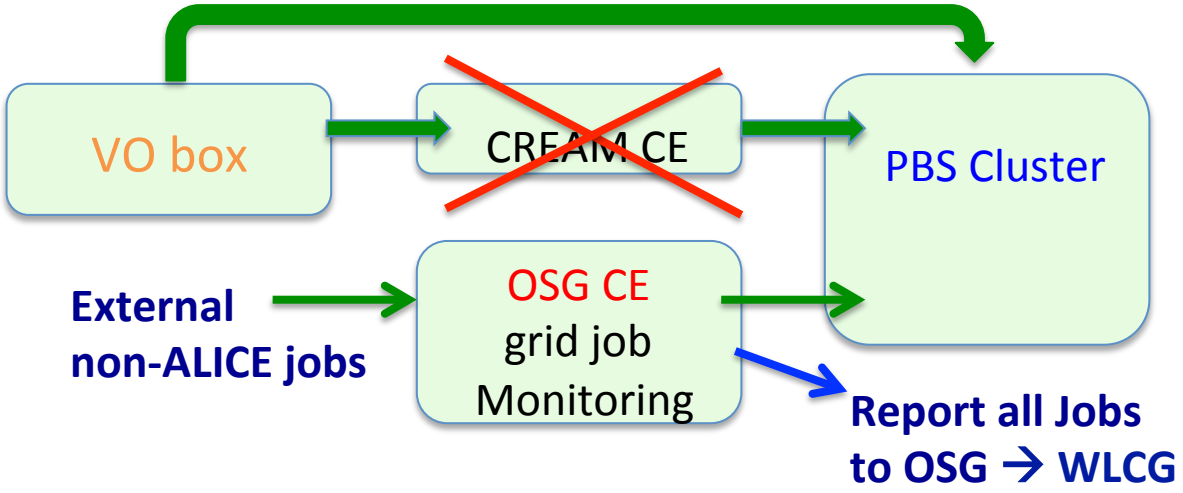


January 1 – March 31, 2013

Site Configurations

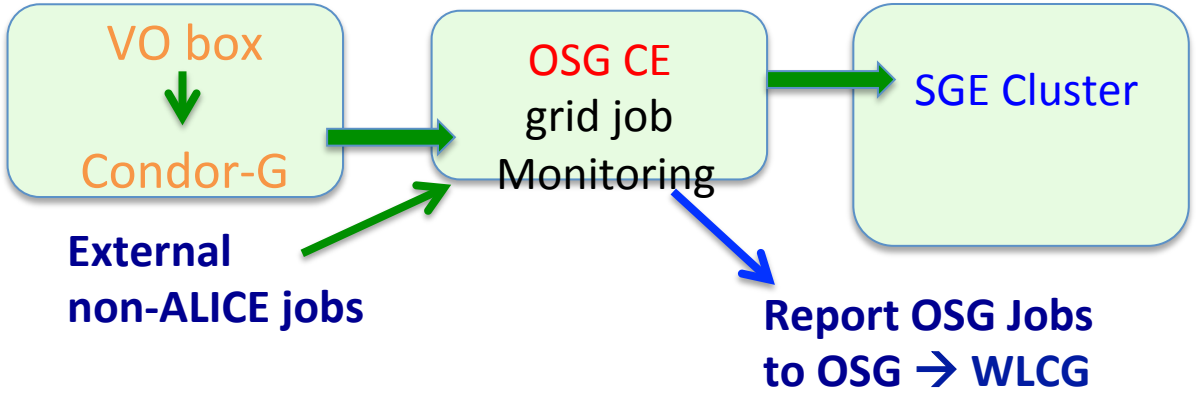
LLNL/LC VOBox

- Submits directly to ~~CREAM-CE~~ PBS
- OSG-CE: independent external interface
- OSG Gracia site service reports all jobs as grid jobs



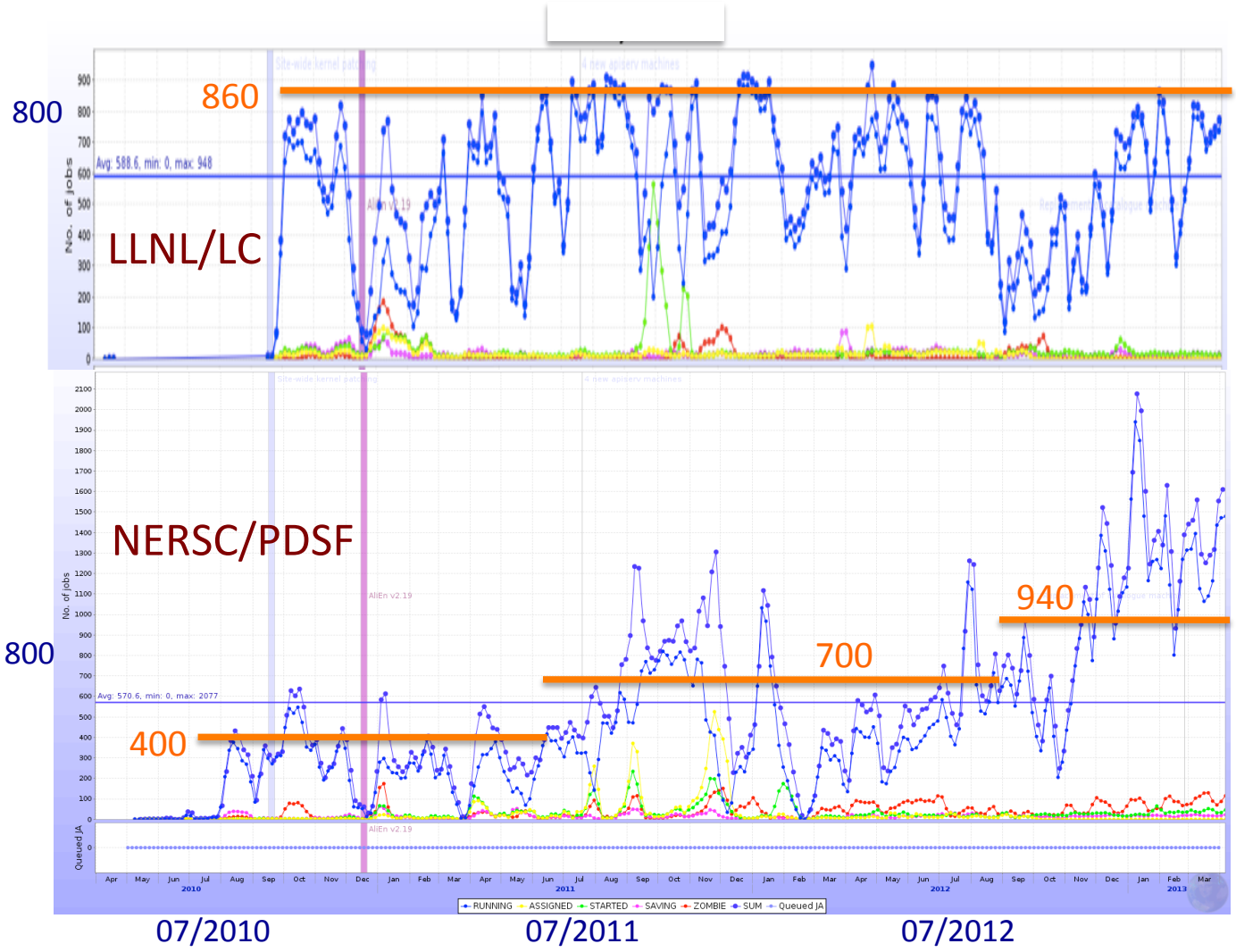
NERSC/PDSF VOBox

- Submits to Condor-G service on VO box
- Condor-G submits jto OSG-CE at PDSF
- OSG Gracia service reports OSG jobs



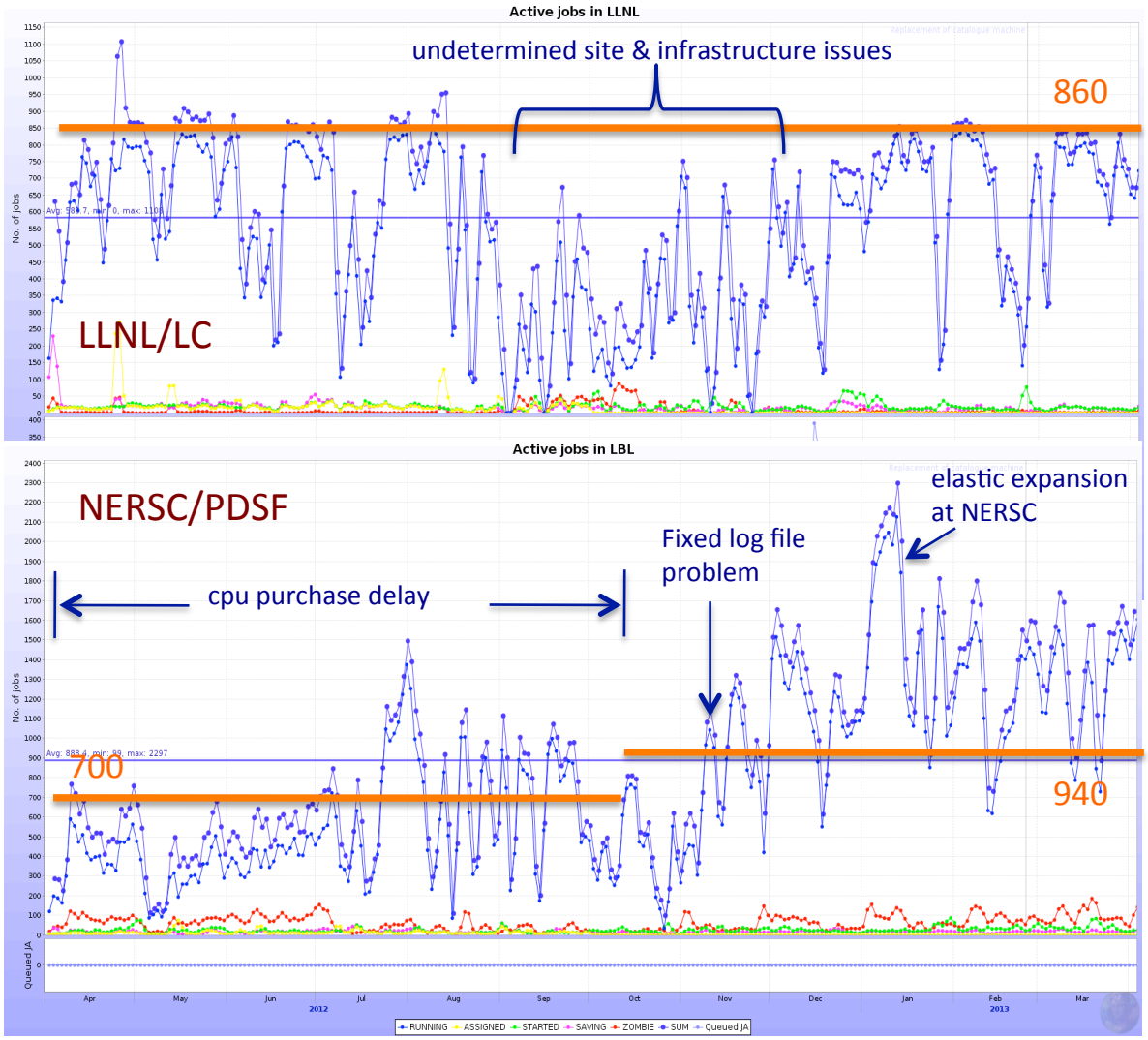
Both LLNL/LC and NERSC/PDSF Site & VO Box services are managed locally

Job & core count over project history



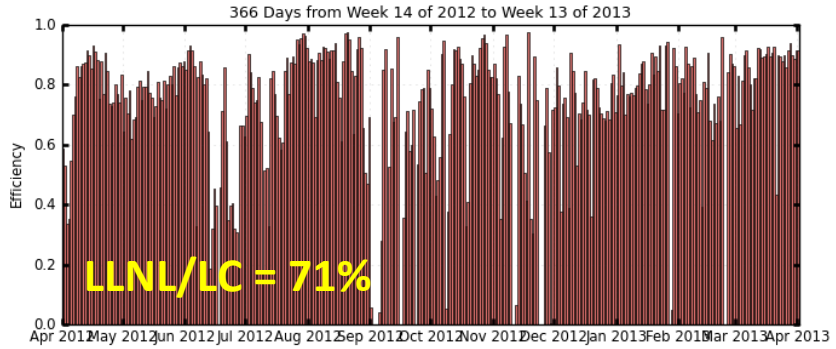
Job levels - 2012 RRB Year

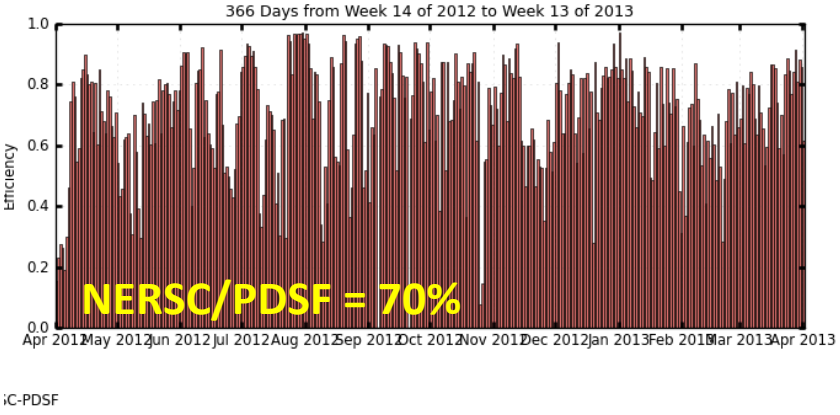
- Wall clock utilization:
  - LLNL/LC
    - <Jobs> = 590
    - 9.2 kHS ~ 80% pledge
  - PDSF
    - <Jobs> = 890
    - 13 kHS ~ 110% pledge
- Issues shown
  - Unaccounted usage gaps
    - Site services control?
  - cpu deployment delay
  - log issue suppressed use
    - AliEn vs Condor-G
  - Zombie grass on PDSF
    - V.Mem as a consumable
      - Jobs killed by SGE



Efficiency as cpu/wall

OSG Reports

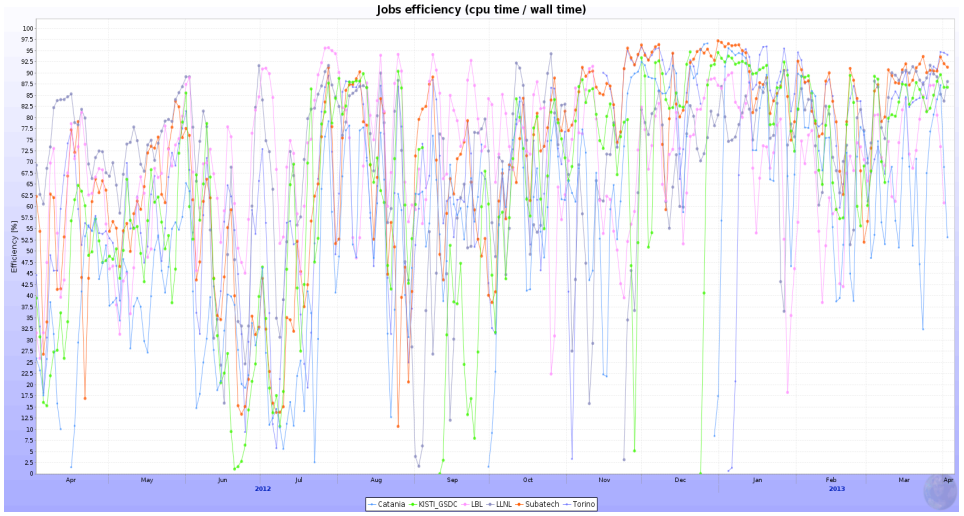




Efficiency = cpu/wall

Pilot + Payload over RRB 2012

MonaLisa Reports



Efficiency = cpu/wall

Payload over RRB 2012

Jobs efficiency (cpu time / wall time)

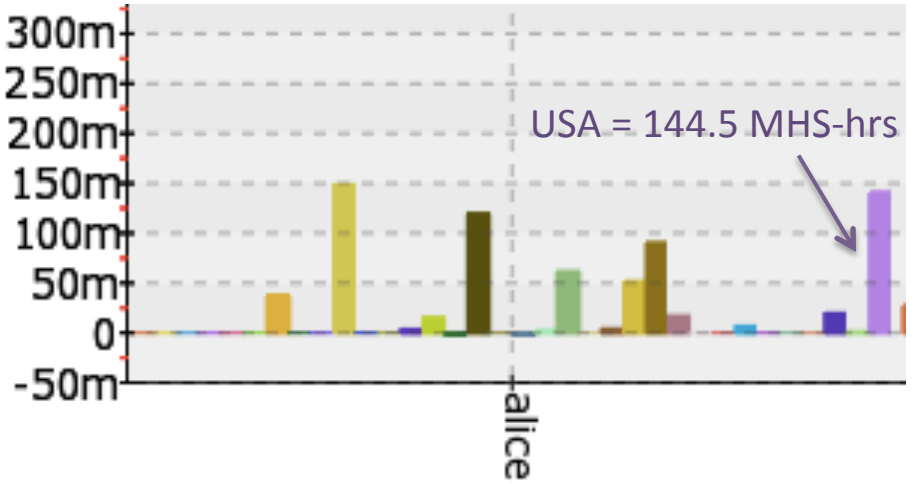
Series	Last value	Min	Avg	Max
1. LBL	76.94	0	69.74	100
2. LLNL	88.71	0	69.34	100
Total	82.83		69.54	

LLNL/LC = NERSC/PDSF ~ 70%

CPU Utilization 2012 RRB Year

- Wall Clock Utilization
  - LLNL/LC > 80% of pledge
  - NERSC/PDSF > 110% of pledge
    - Purchase delay offset by extra resources
  - Fluctuations →slow to diagnose / resolve
- CPU Utilization
  - LLNL/LC & NERSC/PDSF ~ 70% efficiency
    - Similar to other ALICE T2

EGI Portal Country T2

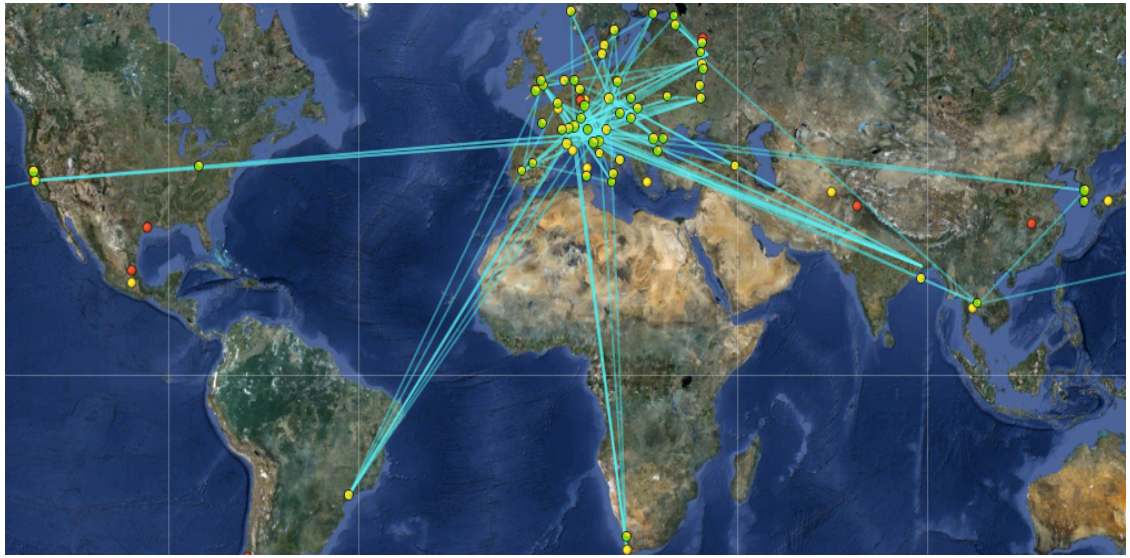


- Utilization relative to pledges
  - LLNL/LC
    - Pledge : 11,500 HS x 24 x 365 x 0.7 (allowed eff.) = 70.5 MHS-hrs
    - Delivered: 15.6 HS/core x 5.3 Mcore-hrs x 0.7 (measured) = 57.9 MHS-hrs } 82%
  - NERSC/PDSF
    - Pledge: 12,000 HS x 24 x 365 x 0.7 = 73.6 MHS-hrs
    - Delivered: (13.3 HS/c x 2.7Mcore-hrs + 15.5 HS/c x 4.7 Mcore-hrs)x0.7 = 86.9 MHS-hrs } 118%
  - Combined
    - Pledge = 144.1 MHS-hrs , Delivered = 144.5 MHS-hrs } 100%

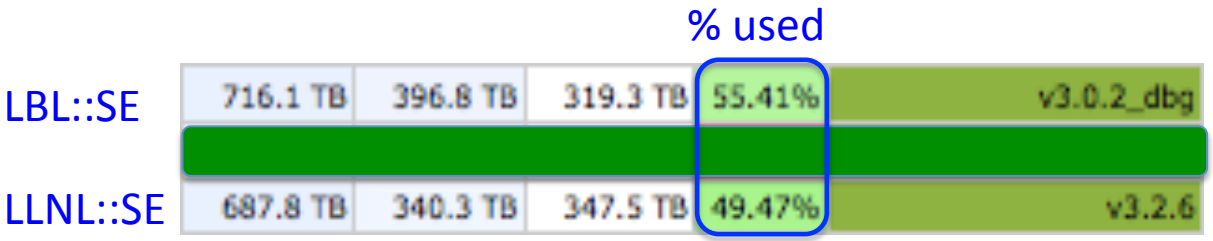
➢ Biggest concern seems to be fluctuations in wall clock utilization

Storage Element Deployments

- Pledged Obligations
  - LLNL/LC
    - 650 TB, 2011 - 2013
  - NERSC/PDSF
    - 740 TB in 2011
    - 1,020 TB in 2012
    - 1,200 TB in 2013



- Installed Capacity
  - LLNL/LC = 685 TB since Aug '10
  - PDSF = 720 TB since Oct '11
    - 144TB June 2010
    - 288TB Apr 2011
    - 288TB Oct 2011



SE Performance

- Reasonable bandwidth at sites
  - Average ~50MB/s writes
  - Average ~300MB/s read
- Bandwidth tests CERN:
  - 100 Mbps
  - 200 ms RTT
- Test history:
  - LBL::SE @ 96%
  - LLNL::SE @ 86%

CERN DNS / FW problem to LLNL

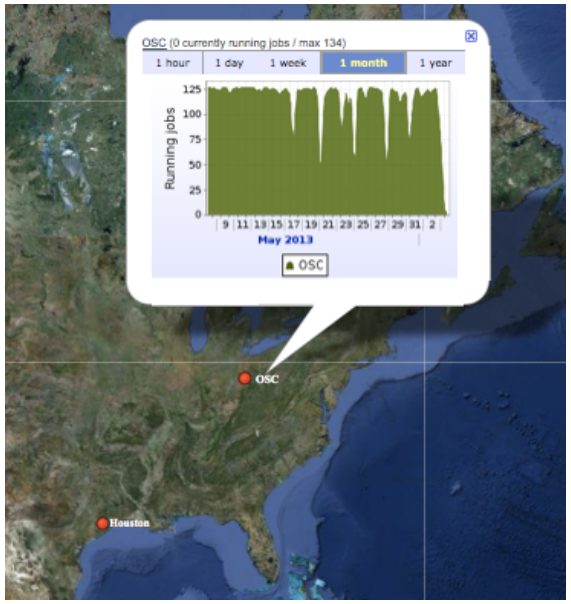
SE Test history

SE Utilization Summary

- Both LLNL:SE and LBL:SE have generally shown stable operations
  - LLNL:SE Add Test failures:
    - Difficult to ID & located at the tester
    - Likely reduced utilization for that period
- Overall utilization has been low relative to capacity
  - Bandwidth & rtt is about 2x worse for US relative to sites in Europe
  - Actual traffic at Storage Elements is comparable to other sites
    - dominated by local I/O
- Overall utilization per cpu-core doesn't appear low
  - US T2 have installed ~ 0.75 TB/core
  - My survey of other sites suggest 0.5 TB/core as a better target

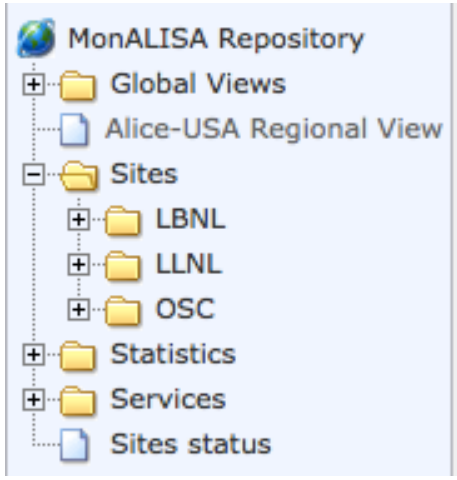
Additional US Sites & Activities

- Ohio Super Computing Center (OSC)
  - Different funding agency (NSF)
  - Stable resource
    - Not pledged
    - University grant allocation
  - No compatible storage



- University of Houston
  - An early important grid site in ALICE
  - out of service as DOE funded LBNL & LLNL
  - Possible solution for additional resources

- Regional Monitor via our own ML collector
  - Allows for targeted monitoring & site specific data
  - Prototype is deployed
    - Documentation on our local twiki --- will share



Upcoming Hardware Tasks

- Hardware installation plans
  - Installed CPU is sufficient for 2013 pledges, perhaps 2014 as well
  - Will install 260TB of disk on NERSC/PDSF as needed
- Disk and CPU start to go out of warranty this summer
  - Will operate CPU on a DNR model→ small but steady drop in capacity
  - Storage Elements:
    - Will purchase individual replacement disks (~5%) and swap in as needed
    - Will replace an individual server node if isolated problem, case by case basis
    - DNR for more significant problems

Upcoming Operational Tasks

- Upgrade resources to SL6.x
  - Upgrade underway at LLNL, path set for NERSC/PDSF ~ June/July
- New OSG CE software
  - Major change in installation model (rpms instead of pacman)
  - Upgrade underway at LLNL, awaiting UGE patch for NERSC/PDSF
- Upgrade XRootD at LBL::SE
  - Consider waiting until xrootd 4.x is released?
  - Will include SL6.x plus a change to our redirector identity at PDSF
    - LLNL::SE has been upgraded to 3.2.6 and RHEL6 (SL6)
  - Develop XRootD/HPSS interface
    - perhaps at next AliEn developers workshop
- LHCONE
  - Talks with ESNet & NERSC are underway, LLNL/LC to follow

Longer Term Plans

- Original Project plan called for LBNL to host an ALICE Tier 1 center
  - NERSC maintains a large HPSS tape storage facility
- Project directed to first establish & demonstrate T2 operations
  - While working on XROOTD/HPSS interface software
- Project External review in late 2011
  - Project was given a very favorable review, but
  - Recommended we consider T1 transition only beginning after LS1
- T1 option to be re-evaluated in Spring 2014 → We'll see