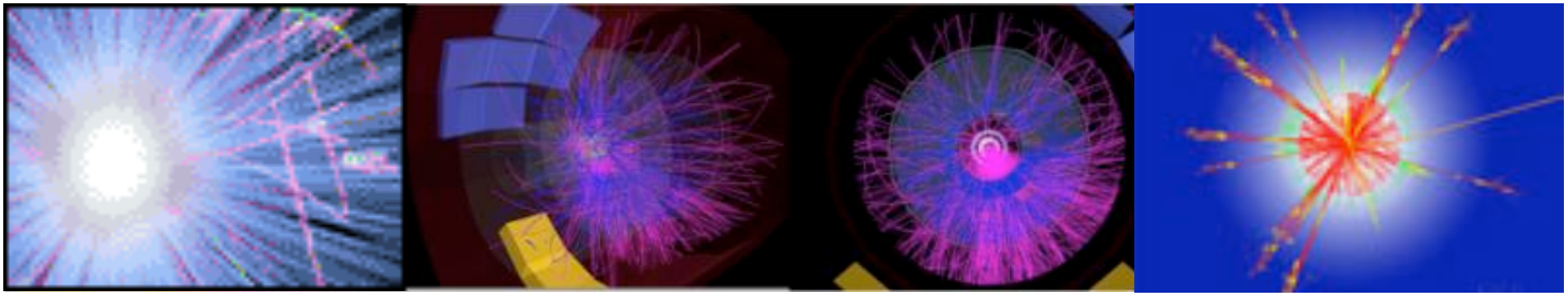


PDSF@NERSC

Site Report

HEPiX Fall 2011 Workshop
October 24-28 2011

Elizabeth Bautista, Eric Hjort, Larry Pezzaglia, Iwona Sakrejda, Jay Srinivasan

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

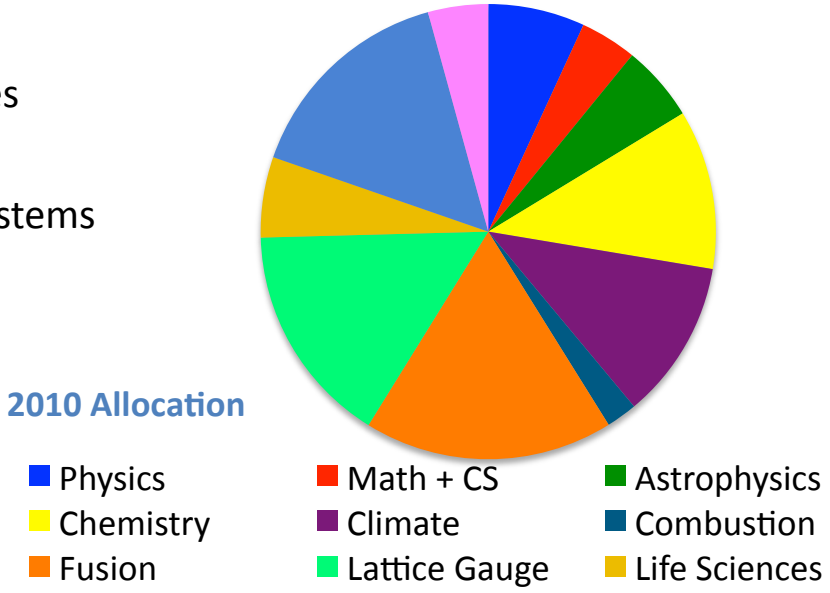
BERKELEY LAB Lawrence Berkeley National Laboratory

PDSF Overview

- PDSF@NERSC
- PDSF in numbers
- Changing user base
- Changes in the compute systems hardware
- Storage expansion
- Batch system (from SGE to UGE)
- System Monitoring (Benchmarks,Nagios and Ganglia)
- User Support
- A little bit of history

Why PDSF@NERSC

- Located at LBNL NERSC is the primary computing center for the US DOE Office of Science
- NERSC serves a large population of ~4000 users, 400 projects and 500 codes
- Focus is on "unique" resources
  - Expert computing and other services
  - 24x7 monitoring
  - High-end computing and storage systems
- NERSC is known for
  - Excellent services and user support
  - Diverse workload



Snapshot of NERSC

Large-Scale Computing Systems

Franklin (NERSC-5): Cray XT4

- 9,532 compute nodes; 38,128 cores
- ~25 Tflop/s on applications; 356 Tflop/s peak

Hopper (NERSC-6): Cray XE6

- 6,384 compute nodes, 153,216 cores
- 120 Tflop/s on applications; 1.3 Pflop/s peak

Clusters

140 Tflops total

Carver

- IBM iDataplex cluster

Magellan Cloud testbed

- IBM iDataplex cluster

Magellan+Carver =~10k cores total

GenePool (JGI)

- ~5K core throughput cluster

PDSF (HEP/NP)

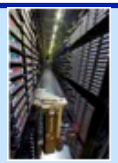
- ~1.5K core cluster

NERSC Global Filesystem (NGF)

Uses IBM's GPFS

- 1.5 PB capacity
- 5.5 GB/s of bandwidth

HPSS Archival Storage



- 59 PB capacity
- 4 Tape libraries
- 150 TB disk cache

Analytics

Euclid

(512 GB shared memory)

Dirac GPU testbed (48 nodes)

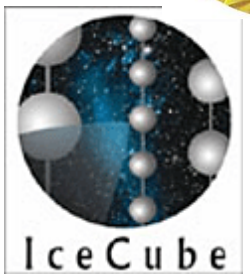
PDSF Quick Overview

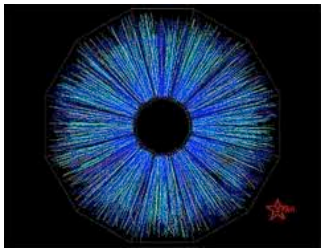
- Parallel Distributed Systems Facility
  - A Computing Facility at NERSC primarily used now by the NS and HEP community
  - Commodity Linux Cluster
  - Interconnected with 10GE, Gigabit Ethernet and partially with Infiniband.
  - In continuous operation since 1996 (15 years!).
  - Over 1500 compute cores (quad-core and six-core nodes)
  - Over 1PB of globally accessible storage (GPFS and xRootD based)
  - Tier-1 for STAR, Tier-2 for ALICE (w/LLNL), and Tier-3 for ATLAS

PDSF Workloads

- PDSF is an essential resource for a number of groups such as STAR, KamLAND, ATLAS, ALICE, DayaBay, IceCube, etc.
- Groups such as CUORE, Planck, Majorana also use PDSF
- Groups such as SNO, SNFactory, CDF, BaBaR, have used PDSF in the past.

ATLAS





STAR

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

PDSF Workloads and Services

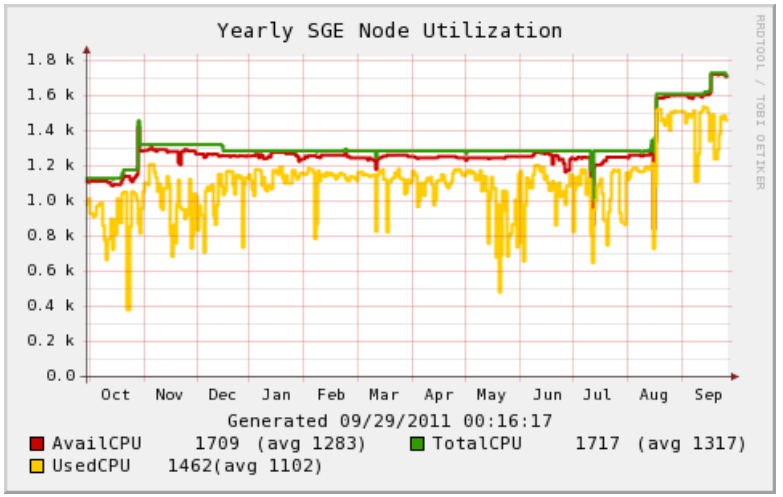
- Computing is serial
  - High throughput of jobs
  - Fair share scheduling mechanism
  - Virtualization of software environment (CHOS)
- Large quantities of experimental data transferred in and processed data shared with other centers
  - Excellent 10GigE (multiple) connection to mass storage
  - Proximity to ESNet
  - Large and good quality storage (GPFS, scratch local to compute systems and XRootD)
- Grid services (OSG stack)
  - Data transfer nodes supporting bestman (srm)
  - Gatekeeper interface to the SGE batch system
- Production accounts (gsissh based)

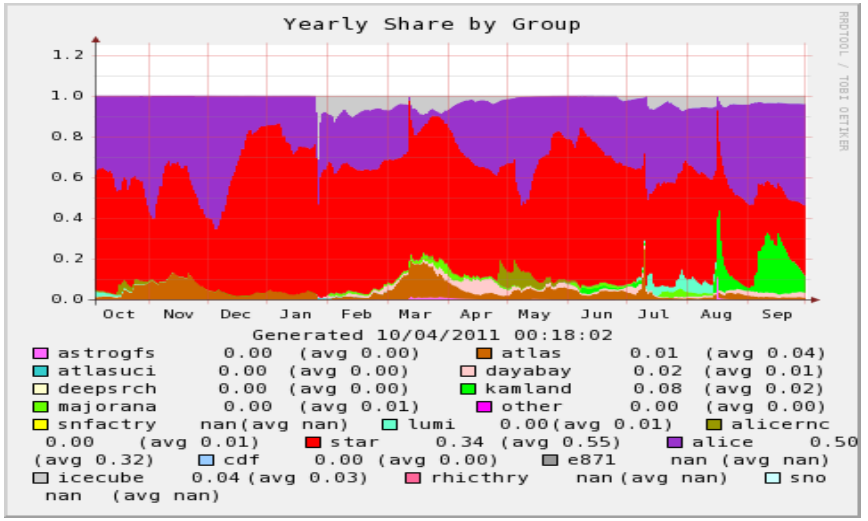
Hardware Procurement

- Hardware is kept in full service until warranty (usually 3 years) expires.
  - Hardware is retired once warranty has expired (usually after a period of 3-6 months) and only once replacement hardware is available
- Try to bunch up procurements
- ALICE, ATLAS (and other LHC experiments) get special pricing from Dell for compute and storage servers and networking equipment
- Typical compute node is
  - 2.6 GHz, Intel Westmere, 2-socket, 6-core/socket, 4GB RAM/core,
  - 4 disk slots available (for node-local storage)
- Storage is ~$350/TB, we usually purchase in units of ~20TB
- To the extent possible, we share resources amongst groups
  - More efficient during procurement
  - Better performance for users

Cluster Utilization

- PDSF Compute and Storage resources are used by a number of groups in the Nuclear Science and Physics Divisions
- We utilize the Grid Engine Batch System and Scheduler to control usage of the compute resources
  - Fair Share scheduler to control the distribution of jobs amongst the various groups
  - But scheduling of jobs also depends on resources jobs request (memory, I/O, HPSS, disk space)
- Groups get their fair share (a function of the contribution to shared resources) over a period of time
- Unused cycles are available to other groups as a function of that groups fair share
- Cluster is well utilized as a result





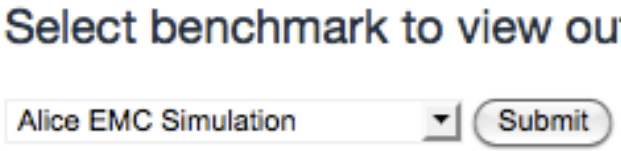
Batch System

- Evaluating torque/MOAB – NERSC is considering a site license
  - Global resources
  - Reservations
  - Fair Share scheduling
- Purchased support for Grid Engine from UNIVA
  - we are doing our first “supported” upgrade in November

Computing Hardware

- PDSF consists of compute, storage, and ancillary systems
- Compute nodes are primarily 2 socket 4 and 6 core Intel and AMD based 1U servers, with 3 or 4 GB RAM/core
- We have ~1500 compute cores on ~200 nodes
- The cluster installs are managed with xCAT
  - Moving towards diskless xCAT installs
  - nodes are discovered, configured, and managed by xCAT. When in production, the nodes netboot into a prepared OS image.
  - We aim to keep the image lean and maintainable. Our build scripts wrap around xCAT's image tools and version all changes with FSVS.
  - The image is built mostly from unmodified Scientific Linux RPMs and is easily kept under 400 MB.
  - image assembly completes in under 15 minutes.
  - CHOS provide flexibility for jobs without bloating the base OS image.
- Configuration files stored in cvs (transitioning to svn) and managed by cfengine.

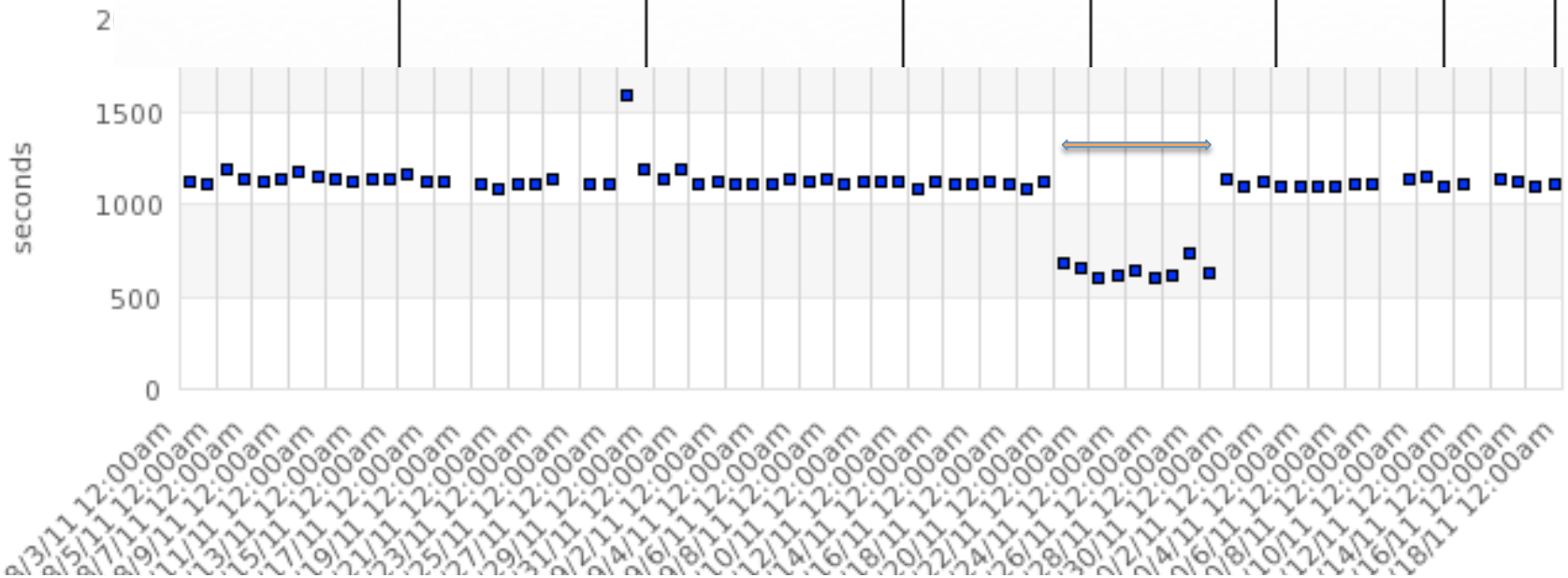
Job Benchmarks

- Actual applications were selected and are run on PDSF daily.
- Results stored in a DB and accessible to users via a web interface: 

- At this point the following applications are used :
  - ATLAS Fragmentation
  - ALICE EMC
  - Daya Bay Analysis
  - STAR pp500 Reconstruction
  - STAR AuAu200 Reconstruction

Job Benchmarks

Completion Date	Benchmark	Wall Time (seconds)	Metric Value	Metric Units	Job ID	AP ID	Notes
October 17, 2011 1:13:19pm	AuAu200-1	1,108	1,108.00	seconds	8055489		cputime 1040.512 maxvmem 983.750



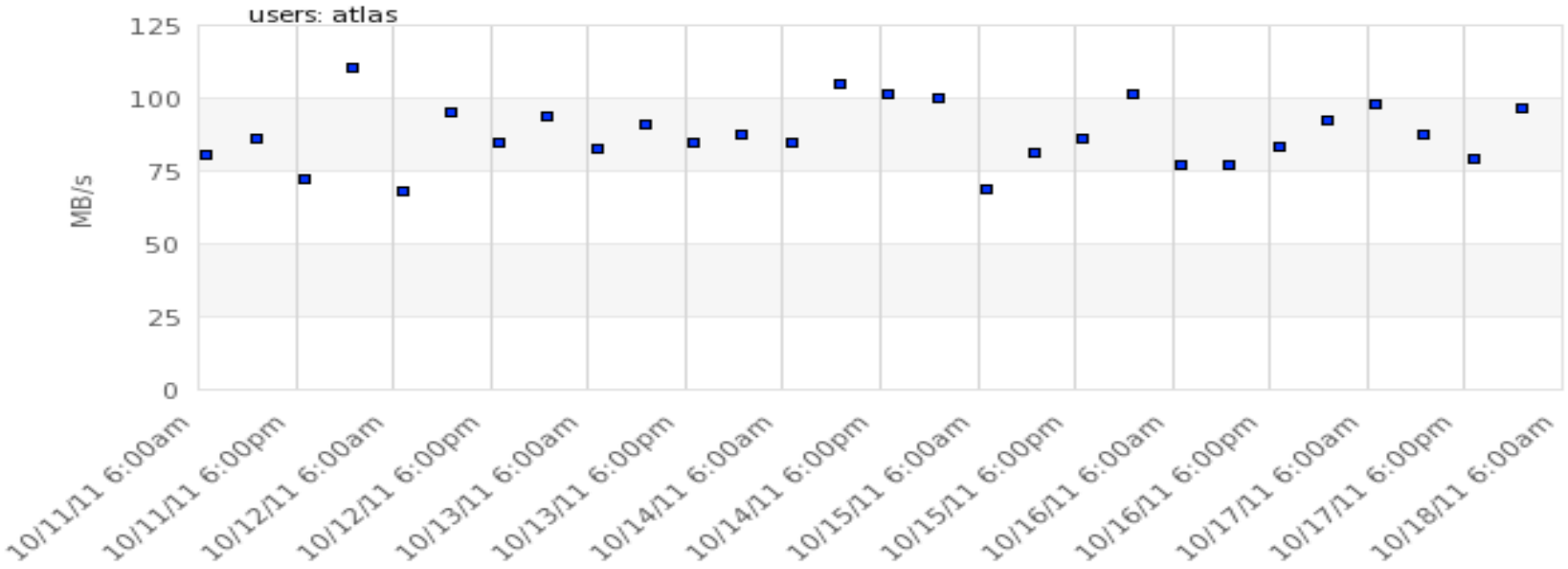
Benchmark	# Successful Runs	Avg Wall Time (seconds)	Avg Metric Value - seconds	COV (%)
AuAu200-1	72	1,066.03	1,066.03	16.14

Storage Hardware

- Storage systems are either F/C-SATA or SAS/SAS-2 based systems
- Storage is either 1GE or 10GE attached storage (backend is either 3/6 Gbps SAS, or 4/8 Gbps FC)
- We have 650TB in shared, globally accessible filesystems using GPFS ("elizas" 1-10,14-18)
- Users also have access to the 1.6PB "Project" NERSC Global Filesystem, which is shared all across NERSC

IO Benchmarks

- All the elizas are monitored regularly
- Transfer rates measured 4 times per day for the past week on all of the eliza file.
- Benchmark description available at I/O Benchmarking Details.
- Click on the plot to get details for that eliza and click on the ganglia link to get the corresponding weekly ganglia plot.



User Support

Ticketing System

- Last year a decision was made to move away from Right Now Web.
- Selection process identified 2 candidates – "Service Now" one of them.
- ESNet and LBNL already using
- NERSC decided thus to go with it too.
- Navigis selected as an implementation partner
  - Define requirements
  - Assist with move of tickets to the new system
- "Service Now" will provide on-going support once move is completed.
- NERSC will consider using "Service Now" in other areas of operation.

NERSC Web Pages

- SilverStriper Content Management Software selected

- Quinn Interactive assisting in training and transitioning. They customized the SilverStripe for us, including:
  - New logo
  - Site design
  - Design implementation (skinning the site with Silverstripe templates)
  - Additional functionality (such as the publications database)

A little bit of history….

- Personell changes 2011:
  - Jay Srinivasan became Head of NERSC Computational Systems Group
  - Iwona Sakrejda came back as the PDSF lead
  - Larry Pezzaglia joined the CSG with focus on PDSF
  - Elizabeth Bautista (CSG) is still with us
  - Eric Hjort continues to provide user support

2011

2000 – Moving into OSF

2003 – Shane, Cary, Iwona, Tom, PDSF

Vintage Hardware

2003