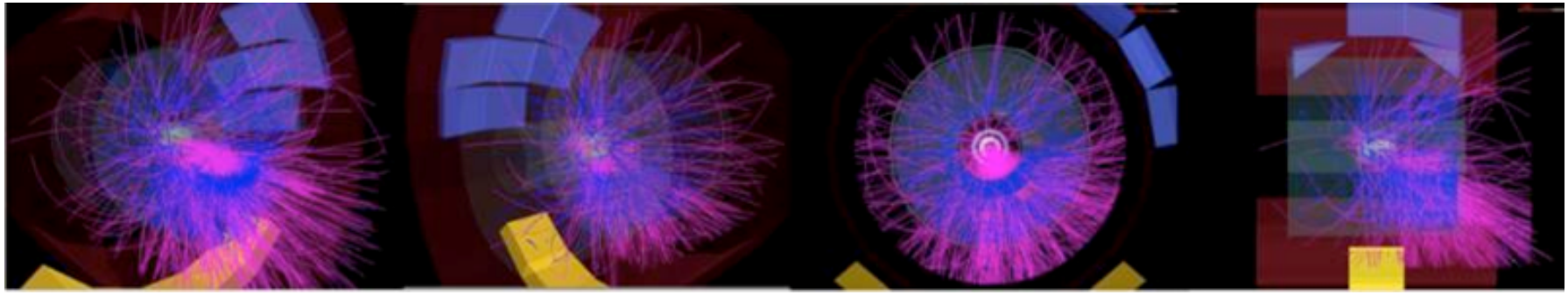


PDSF at NERSC

Site Report – HEPiX Spring 2012 Workshop

Eric Hjort, Larry Pezzaglia, Iwona Sakrejda

National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory

April 2012

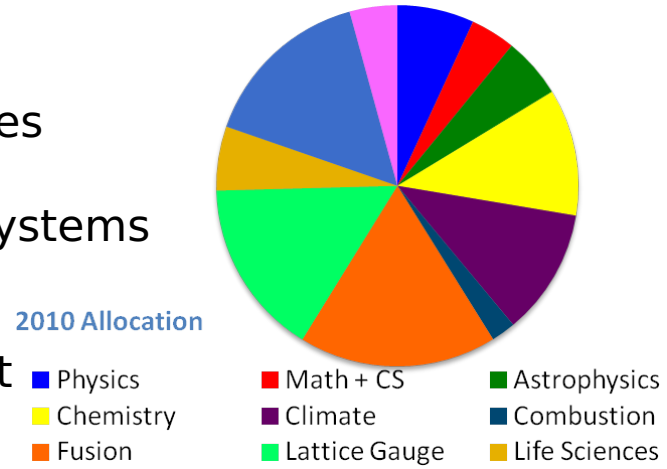
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

BERKELEY LAB Lawrence Berkeley National Laboratory

Snapshot of 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)

NERSC Update

- Franklin (Cray XT4) will be retired on 2012-04-30
- Ongoing integration of the Joint Genome Institute (JGI) computational systems and supporting infrastructure.
  - Workload is mostly serial, high-throughput jobs
- Upcoming mid-range procurement
- Transition to ServiceNow ticketing system complete

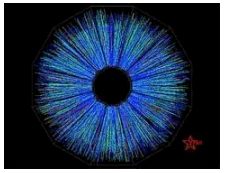
PDSF Overview

PDSF Overview

Parallel Distributed Systems Facility

- A commodity Linux cluster at NERSC serving HEP and NS projects
- 1GbE and 10GbE interconnect
- In continuous operation since 1996
- ~1500 compute cores on ~200 nodes
- Over 750 TB shared storage in 17 GPFS filesystems
- Over 650 TB of XRootD storage
- Supports SL5 and SL6 environments with CHOS
- Univa Grid Engine (formerly Sun Grid Engine) batch system

PDSF Workloads

- PDSF has a broad user base (including non-CERN and non-LHC projects)
- Workload consists of serial, high-throughput jobs
- Fair share scheduling mechanism with UGE
  - Projects “buy in” to PDSF and the UGE share tree is adjusted accordingly
- PDSF is Tier-1 for STAR, Tier-2 for ALICE (with LLNL), and Tier-3 for ATLAS

PDSF Data Flow

- Large quantities of experimental data are transferred into PDSF from other centers
- Processed data is shared with other centers.
- PDSF is well-situated to handle this model:
  - Excellent 10GbE connections to mass storage
  - Close proximity to ESnet
  - High-capacity and high-quality storage:
    - PDSF GPFS filesystems (“elizas”) (Over 750TB)
    - PDSF XRootD storage (Over 650TB)
    - 1.4PB “/project” NERSC Global Filesystem (mounted on all NERSC systems)
    - Local disks on compute nodes

PDSF Grid

- Grid services (OSG stack)
  - Data transfer nodes supporting BeStMan (SRM)
  - Gatekeeper interface to the UGE batch system
- Production “project” accounts (GSISSH based)

Highlighted PDSF Changes

PDSF Changes

- Staff Changes
- Dell Support Challenges
- Ongoing SL6 Deployment
- XRootD for STAR
- Improved node and image management with xCAT

Staff Changes

- Recap from October's update:
  - Jay Srinivasan became the Computational Systems Group (CSG) Lead
  - Iwona Sakrejda returned to PDSF as the PDSF System Lead
- Elizabeth Bautista became the Computer Operations and ESnet Support (CONS) Group Lead
- Larry Pezzaglia is still with PDSF
- Eric Hjort continues to provide user support

Dell Support

- We have a significant quantity of Dell equipment:
  - Servers: R410 and R710
  - Storage: MD3200/MD3000 and MD1200/MD1000
- Dell is not well equipped to support external storage **attached to Linux servers**
  - Communication barrier between Storage and Servers support groups
  - Storage support analysts are generally unfamiliar with Linux fundamentals and common UNIX/Linux tools (e.g., dd and ssh)
    - Most analysts assume a Microsoft shop using MD3xxx units to export LUNs via iSCSI for VMware VMs

Dell Support

- RHEL is a “certified” platform, but SL is not
- We understand that building a support infrastructure is not easy and we want to work with Dell to improve this situation

SL6 Deployment

- We have a production SL6 environment made available via CHOS.
- We are performing a rolling upgrade of node base OSes to SL6
  - ~50% of nodes have been converted
- EL6 challenges:
  - pam_tally vs pam_tally2 for tracking login failures
  - Ganglia 3.0 vs 3.1
  - Porting CHOS to EL6-based kernel
- Overall, SL6 is a solid release. Our thanks and congratulations to the SL team.

XRootD for STAR

- STAR is deploying XRootD on PDSF compute nodes in a similar manner as is done for STAR at RCF
- This is in addition to the existing ALICE XRootD storage
- This model will provide STAR with multiple benefits:
  - Leverages inexpensive disks on the compute nodes to serve read-only data to data-intensive tasks
  - Reduces reliance on large shared file systems that can be a single point of failure for a workflow
  - It also adds some costs:
    - XRootD configuration and deployment
    - Data maintenance (e.g., handling of disk failures)

xCAT Management

- xCAT (Extreme Cloud Administration Toolkit) is an infrastructure management software package
  - `http://xcat.sf.net`
  - We also use xCAT on Carver, our IBM iDataPlex system
- xCAT is an excellent and extensible tool
- Particularly helpful for PDSF have been:
  - Node discovery and auto-configuration
  - Node image build and management framework

xCAT Node Discovery

- New nodes are discovered, configured, and managed by xCAT.
  1. xCAT knows to which Ethernet switch and to which switch port every new node is connected
  2. When a new node boots, it is allocated a temporary IP address and requests “discovery”.
  3. xCAT queries Ethernet switches via SNMP to determine the switch port to which the new node is connected
     - With this information, xCAT now knows the identity of the new node.
     - xCAT configures the node’s IPMI BMC and boots it into the production base OS.

xCAT Images

- xCAT provides a framework for building and booting netboot node images.
  - These nodes are termed “diskless” in xCAT parlance.
- Our image build scripts:
  1. Modify the stock xCAT “genimage” utility
  2. Use “genimage” to create a minimal image
  3. Modify the image to add required PDSF packages (e.g., GPFS, CVMFS, Ganglia)
  4. Commit the changes to an SCM repository
  5. Perform a clean checkout from SCM
  6. Call the xCAT “packimage” utility to enable the image for production use.

Image versioning

- Every time the image is changed, it is re-assembled from scratch.
  - This ensures a reproducible and maintainable image.
- We use FSVS (“Fast System VerSioning”, pronounced *fisvis*) to version the image.
  - We can easily determine what changed between any two image revisions
  - We can easily revert to **any** previous image version
  - `http://fsvs.tigris.org`

PDSF History

History

2003

2012

History

2003

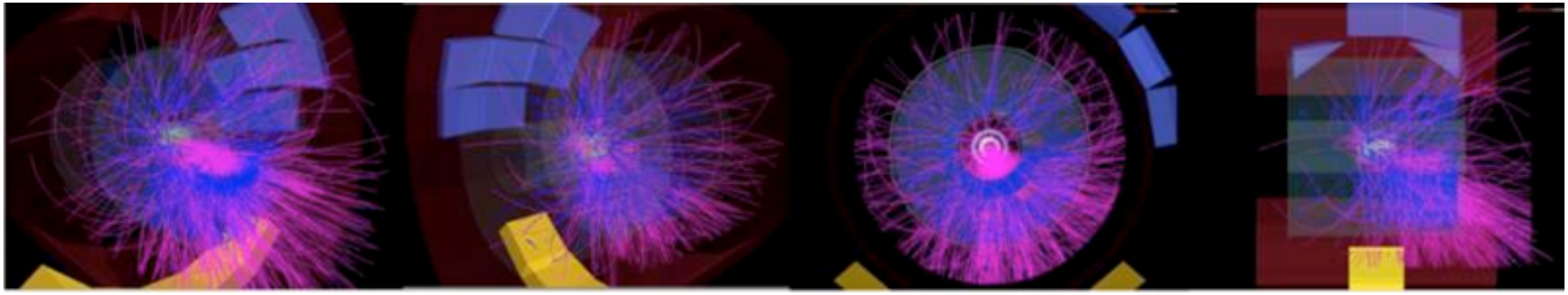
Shane Canon
Cary Whitney
Iwona Sakrejda
Tom Langley

2012

Iwona Sakrejda

Eric Hjort

Larry Pezzaglia



Questions?