Parallel Distributed Scientific Facility

- Rose from the ashes of the SSC
- Really came to life in 1997 when NERSC moved to LBNL
Many Years of Science

On track to deliver ~2500 CPU years in 2014
PDSF Today

- Used by Nuclear Physics and High Energy Physics groups
  - Simulation and analysis
  - Data mirrors
- Evergreen and condo
- 2600 cores with 20 – 60 GB RAM, 230 nodes
- Dedicated GPFS and XRootD storage

PDSF Shares 2014

- ALICE 40.6%
- STAR 34.5%
- ATLAS 12.0%
- Daya Bay 9.8%
- Majorana 2.3%
- IceCube 0.5%
PDSF System

- **PDSF Compute Cluster**
  - 158 Dell Servers (8, 12 cores, mostly 4GB/core)
  - 68 Mendel Servers (16 cores, 4GB/core, FDR IB – 30Gb/s link between PDSF core router and Mendel)
  - 3 hosts behind load balancer for interactive access
  - 4 backup interactive nodes used for special services and development
  - Auxiliary servers (mostly Dell)
    - 2 UGE servers (master, shadow) for reliability
    - 2 admin servers (managing deployment and configuration)

- **Data Storage**
  - 690 TB of GPFS local storage (in addition to NERSC global systems)
  - 2 data transfer nodes with 10Gb/s access to PDSF storage

- **Group Specific Hardware**
  - ALICE XRootD cluster: 10 Dell Servers with a total of 720 TB of storage and 10Gb/s ethernet
  - STAR XRootD cluster: Uses disks of compute nodes, 1.2 PB of storage
  - ALICE job submission: VO node with Condor-G and 2 CE gatekeepers with UGE job managers
  - ATLAS VO node
  - Daya Bay: Three interactive nodes where “heavy” processes can be run

- **Networking**
  - Access to NERSC global file systems
  - Combination of Dell, HP, and Cisco switches
  - Cisco core router
    - 2x10Gb/s connection to other NERSC systems/storage
    - 2x10Gb/s connection to the border router
Mendel System Backbone

- PDSF and several other HT clusters needed expansion
- NERSC elected to deploy a single new hardware platform ("Mendel") to handle:
  - Jobs from the "parent systems" (PDSF, Genepool, and Carver)
  - Support services (NX and MongoDB)
- Groups of Mendel nodes are assigned to a parent system
  - These nodes run a batch execution daemon that integrates with the parent batch system
  - Expansion experience is seamless to users
- PDSF is currently transitioning to full Mendel model
Bells and Whistles

- cvmfs installed on all nodes
- OSG software: Monitoring, job submission, usage reporting

Various databases
  - mysql
  - SPADE

Full time staff providing full user environment
Physics Highlights: STAR

• At RHIC in BNL, turned on in 2000
• Analysis and simulation done at PDSF, used 4.1 M hours in 2013
  – XRootD storage on compute
  – Dedicate STAR software stack
• Almost 600 journal publications, more than 11,000 citations
  – Elliptical flow from QGP
  – Jet quenching
  – Heaviest anti-matter particle ever
Physics Highlights: KamLAND

• Liquid scintillator $\bar{\nu}_e$ detector located in the Kamioka Observatory
  – Measured $\nu$’s from nearby nuclear reactors

• Took first data in 2002 at a rate of 200 GB / day
  – Stored on LTO tapes and driven to nearby Japanese university, copies flown to US

• HPSS read tapes on nights and weekends

• Data was processed on PDSF cluster
  – 400 core cluster, done in three months

• First measurement of terrestrial neutrino oscillation

Physics Highlights: ALICE

- Heavy Ion Collisions at LHC
- NERSC is their US Tier 2 facility
- PDSF provides dedicated 720 TB XRootD cluster
- Also 145 TB of local storage and NGF
- 9 million CPU hours in 2013
- System for running and reporting ALICE jobs submitted from all over the world
  - OSG reporting, debugging, and maintenance
Physics Highlights: IceCube

- Neutrino Detector in Antarctica
- Data analysis and simulation on PDSF and carver
- At its peak used ~1.1 M hours on PDSF
- In 2013 found first evidence of high energy astrophysics neutrinos

November 22nd, 2013
Towards the Future: Memory

• Most compute nodes have 16 job slots and 4 GB / slot
  – Jobs can use more, requested at submission

• Average memory request per job
  – 2010: 2.1 GB
  – 2014: 3.2 GB

• Frequently see the cluster “memory full”
Data and Lots of It

- **PDSF provides storage for groups**
  - ATLAS: 310 TB + 40 TB on NGF
  - ALICE: 27 TB, and 720 XRootD storage + 60 TB on NGF
  - STAR: 140 TB and 1.2 PB XRootD storage + 70 TB on NGF
  - Daya Bay: 144 TB + 750 TB on NGF
  - In addition each group heavily uses NERSC global file storage

- **Every new group that has joined PDSF in the last year has also brought a request for storage**
Towards the Future

• Several new groups interested in joining PDSF
  − Neutrinoless double beta decay
  − Dark matter

• Looking forward to 17 more years of science
Thank you.
Physics Highlights: Daya Bay

• Neutrino Oscillation experiment in China
• NERSC is their US Tier 1 facility
• Runs are copied to PDSF within ~15 – 20 minutes
• SPADE over sees transfer
  – Runs on three dedicated PDSF nodes
• Data is processed on PDSF
  – ~1.6 million CPU hours in 2013
• Processed data archived at HPSS
  – 125 TB / year at a rate of 350 – 400 GB / day
• First measurement of $\theta_{13}$ $\nu$ mixing angle
• Science magazine’s Top 10 Breakthrough of the year in 2012
NERSC Systems Today

Edison: 2.39PF, 333 TB RAM
- 5,192 nodes
- 125K Cores

Hopper: 1.3PF, 212 TB RAM
- 6,384 nodes
- 150K Cores

Production Clusters
- Carver, PDSF, JGI, KBASE, HEP
- 14x QDR

Vis & Analytics
- Data Transfer Nodes
- Adv. Arch. Testbeds
- Science Gateways

Ethernet & IB Fabric
- Science Friendly Security
- Production Monitoring
- Power Efficiency
- WAN

Global Scratch
- 3.6 PB
- 5 x SFA12KE

/project
- 5 PB
- DDN9900 & NexSAN

/home
- 250 TB
- NetApp 5460

HPSS
- 50 PB stored
- 240 PB capacity
- 20 years of community data

2 x 10 Gb
1 x 100 Gb

Software Defined Networking

WAN

16 x FDR IB
- 80 GB/s

16 x QDR IB
- 16 GB/s

12 GB/s
50 GB/s
5 GB/s