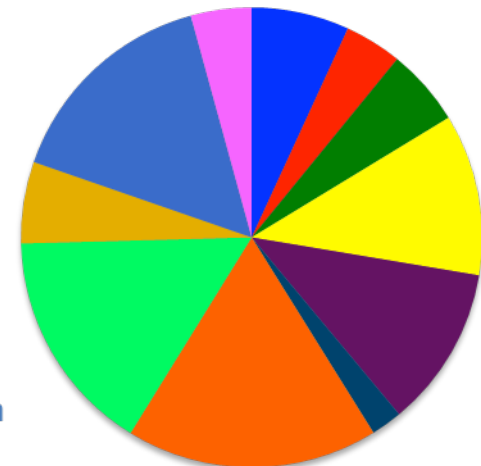


- **Background – motivation behind switching new hardware vendor.**
- **Description of the new acquisition**
- **Deployment and configuration**
  - Racking – challenges of integrating into existing setup
  - Configuration – for fail-over (redundancy)
- **Integration with the GPFS cluster**
  - Replacing the old file system live and preserving name
- **Support Model**

# Snapshot of NERSC



- Located at the Oakland Scientific Facility (until 2015), NERSC is the primary computing facility for the US DOE Office of Science
- Division of LBNL
- over 5000 users
- over 400 projects
- 40<sup>th</sup> Anniversary in 2014



2010 Allocation

- Physics
- Chemistry
- Fusion
- Math + CS
- Climate
- Lattice Gauge
- Astrophysics
- Combustion
- Life Sciences

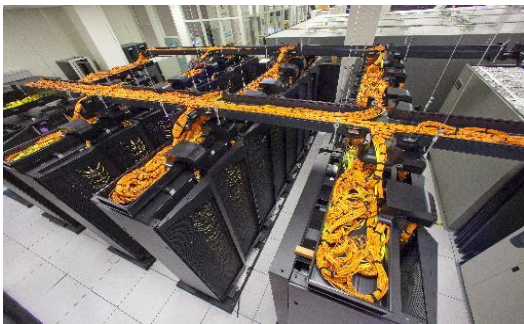
# Systems at NERSC (Except PDSF!)



NERSC-7 Cray XC30  
5200 Nodes  
124 800 cores  
2.4 PFlops Theoretical

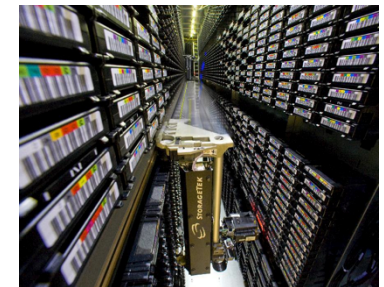


NERSC-6 Cray XE6  
6384 Nodes  
153 216 cores  
1.3 PFlops Theoretical

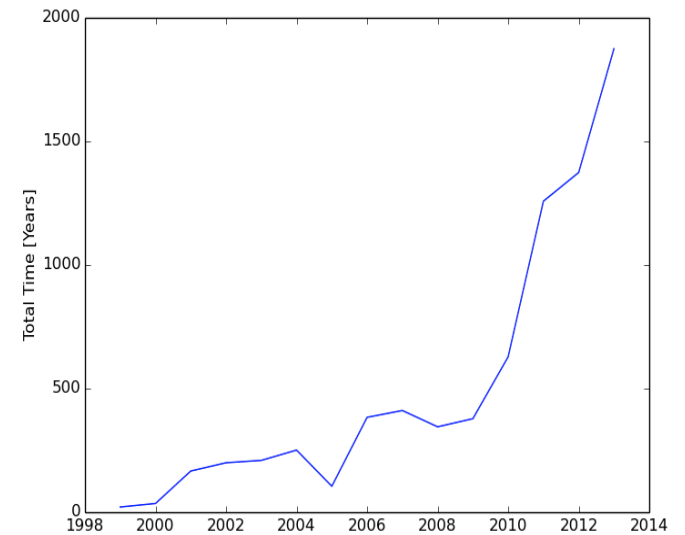
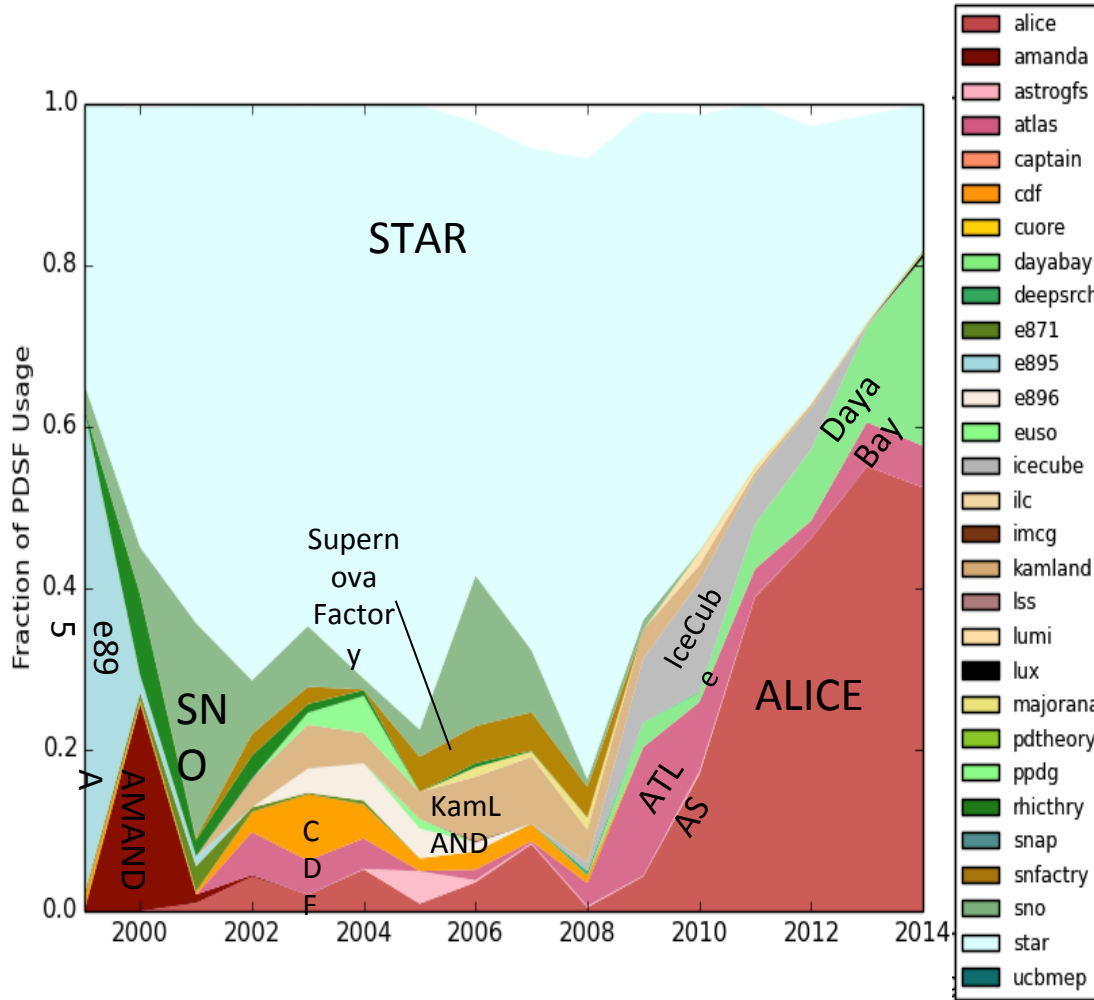


Carver  
IBM iDataplex  
1202 compute nodes  
9984 cores  
106.5 TFlops  
Theoretical

Global Filesystems and HPSS Data Storage  
Archive



# PDSF Overview



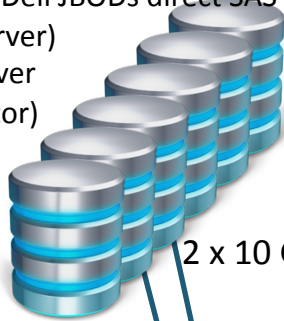
On track to deliver ~2500 CPU years in 2014

# PDSF Cluster Layout



## XRootD Storage Cluster:

10 R710 servers  
MD1200 Dell JBODs direct SAS attached (4 per server)  
R410 server (redirector)



2 x 10 Gb/s

## PDSF local storage (GPFS):

Added 1070TB total  
New formatted: 525TB (Half of PetaByte)



## Auxiliary servers (mostly Dell R410):

2 VO boxes with Condor-G  
2 CE gatekeepers with SGE job managers  
2 UGE servers (master, shadow) for reliability  
2 admin servers (managing deployment and configuration)  
4 backup interactive nodes used for special services and development  
6 DB nodes

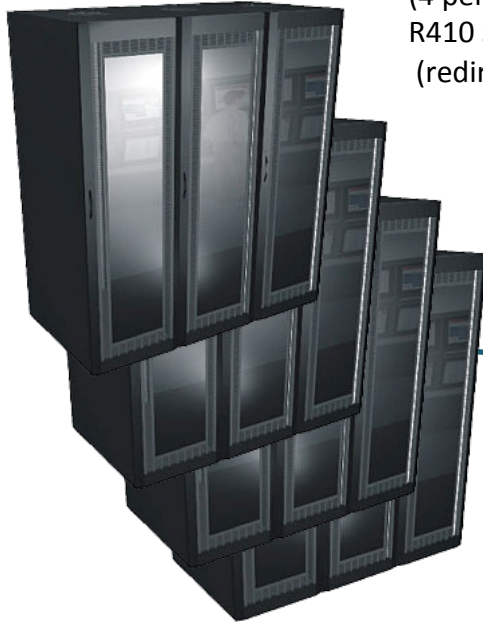
## Network:

Dell, HP and Cisco switches  
Cisco core router

30 Gb/s

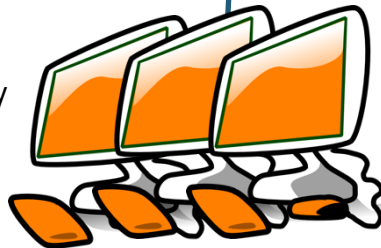
2 x 10 Gb/s

2 x 10 Gb/s



## PDSF Compute Cluster:

200 Dell R410 Servers (8, 12 cores, memory mostly 4GB/core)  
68 Mendel Servers (16 cores, memory 4GB/core, FDR IB)



3 interactive hosts behind load balancer



NERSC Global File System and HPSS

# Background of PDSF: Parallel Distributed Systems Facility



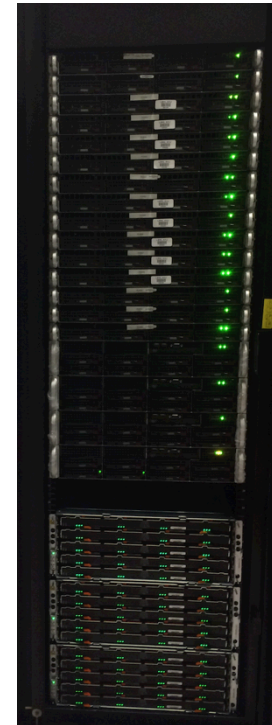
- PDSF have 15 GPFS file systems local to cluster, > 1000 TB for users
- Current GPFS file systems built with Dell hardware platform MD3200 and MD1200
- For a uniform hardware profile with NERSC global file system(NGF), PDSF procured the NetApp storage systems to add to the existing GPFS infrastructure.
  - Pre-established vendor relationship
  - Options on a larger contract for future acquisition
  - Very familiar to our storage group

# Complications



- **Racking**

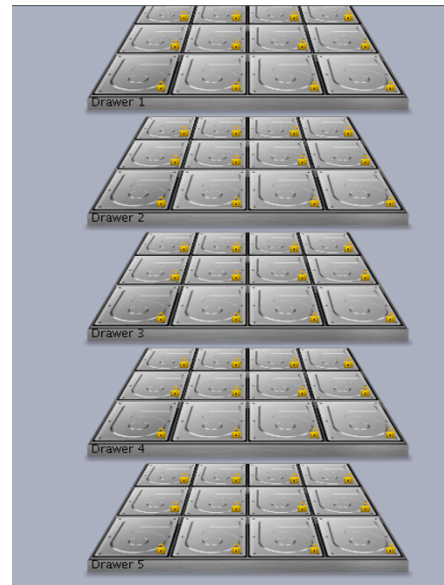
- Challenges encountered the APC cabinet rails needed to expand to accommodate the NetApp E5500 storage units
- Shutting down nodes within the rack and getting them back online.



# Configurations



- **Configurations – design to have a high availability system system with a fail over**
  - Two NetApp controllers are configured with out-of-band management. One of the enclosure with two controllers is SAS directly attached to the E5500 expansion system.
  - 8+2 RAID6 striping across 10 trays (over two enclosures with 120 drives). First enclosure is configured as Raid6 array with 2 disks per drawer expanding vertically to all 10 drawers with 5 drawers from the controllers and the attached expansion system.
  - The vertical expansion of the raid 6 array configuration will allow us with up to two drawers failing.

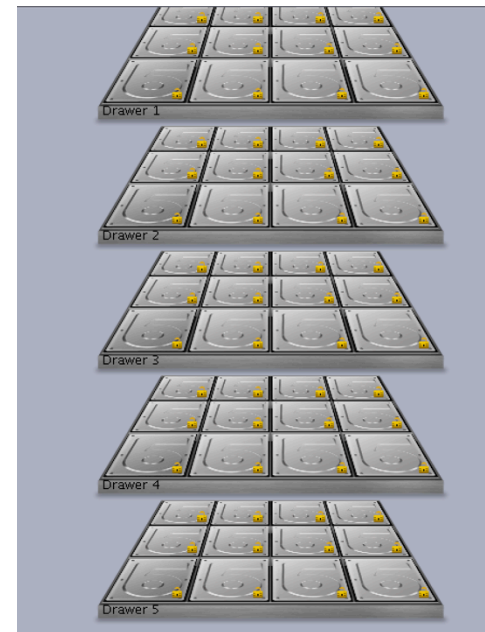




# Deployment and Configuration

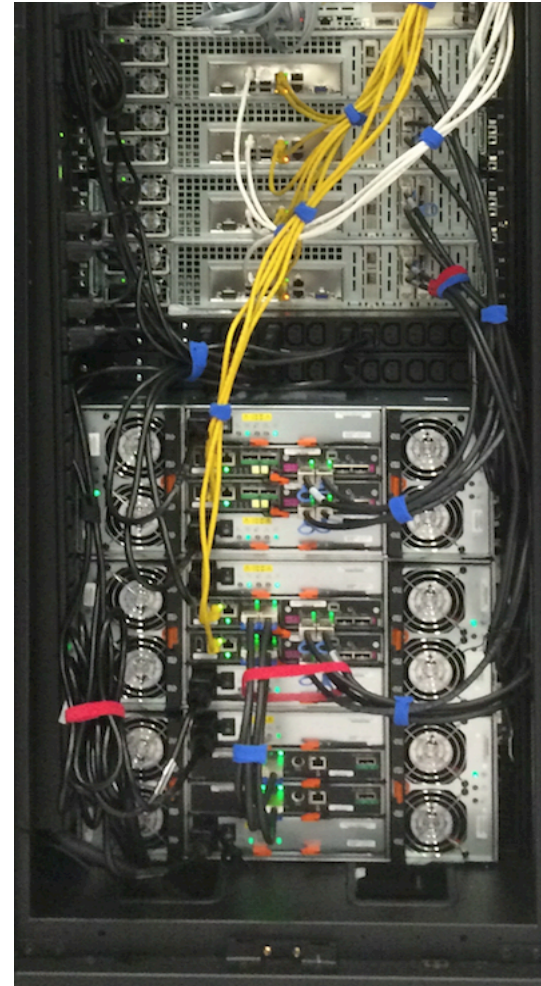


- **Configurations on the second enclosure**
  - 8+2 Raid 6 over 5 trays (on a single enclosure with 60 drives).
  - 2 disks per drawer expanding vertically to all 5 drawers.



# Deployment and Configuration

- Each of the NetApp controllers is directly attach to two Supermicro servers for redundancy.



# Deployment and Configuration



- Logical volume overview:

The screenshot shows the SANtricity Enterprise Management interface. The left pane displays a tree view under 'pd2015.nersc.gov' with 'Discovered Storage Arrays (10)'. The right pane shows a table of these arrays.

Name	T...	Status	Management Connections
pd2005		Optimal	Out-of-Band( <a href="#">details</a> )
Storage Array pd15 11			
Storage Array pd2009			
Storage Array pd15 05			
Storage Array pd2431			
Storage Array pf23 19			
Storage Array pd23 17			
Storage Array pf16 14			
Storage Array pd2414			
Storage Array pd2407			

# Deployment and Configuration



- Host mappings on the NetApp enclosure with a SAS directly attached expansion system.

The screenshot shows the NetApp SANtricity management interface for storage array pd2005. The 'Host Mappings' tab is active, displaying a table of defined mappings. The table has columns for Volume Name, Accessible By, LUN, Volume Capacity, and Type. The mappings are as follows:

Volume Name	Accessible By	LUN	Volume Capacity	Type
Access	Default Group	7		Access
pd2 005 av1	Host Group eliza4	0	29.000 TB	Standard
pd2 005 bv1	Host Group eliza4	1	29.105 TB	Standard
pd2 005 cv1	Host Group eliza4	2	29.105 TB	Standard
pd2 005 fv1	Host Group eliza4	3	29.105 TB	Standard
pd2 005 gv1	Host Group eliza4	4	29.105 TB	Standard
pd2 005 cv1	Host Group eliza4	5	29.105 TB	Standard
pd2 005 ev1	Host Group eliza4	6	29.105 TB	Standard
Access	Host Group eliza4	7		Access
pd2 005 hv1	Host Group eliza4	8	29.105 TB	Standard
pd2 005 iv1	Host Group eliza4	9	29.105 TB	Standard
pd2 005 jv1	Host Group eliza4	10	29.105 TB	Standard
pd2 005 kv1	Host Group eliza4	11	29.105 TB	Standard
pd2 005 lv1	Host Group eliza4	12	29.105 TB	Standard

# Deployment and Configuration



- **Host Mappings on the NetApp standalone enclosure:**

The screenshot shows the NetApp SANtricity Array Management interface for storage array 'pd2009'. The 'Host Mappings' tab is selected, displaying a table of defined mappings. The left sidebar shows a tree view with 'Storage Array pd2009' expanded to show 'Undefined Mappings', 'Default Group', 'Host pd2019', 'Host pd2021', and 'Unassociated Host Port Identifiers'. The 'Defined Mappings' table lists the following:

Volume Name	Accessible By	LUN	Volume Capacity	Type
pd2009av1	Default Group	0	29.105 TB	Standard
pd2009bv1	Default Group	1	29.105 TB	Standard
pd2009cv1	Default Group	2	29.105 TB	Standard
pd2009dv1	Default Group	3	29.105 TB	Standard
pd2009ev1	Default Group	4	29.105 TB	Standard
pd2009fv1	Default Group	5	29.105 TB	Standard
Access	Default Group	7		Access

# Integration with the GPFS cluster



- **Replaced the old file system**
- **Migrating data**
  - 130TB of data to transfer
  - Number of inodes ~14M
  - Parallel copy old to new filesystem when old filesystem is in use
- **Naming – preserving names on ATLAS file catalog**
- **Rsync to bring the new file system up to date**
  - Take old offline – first evaluate the down time by performing live rsync to estimate how long it takes to walk the tree
  - Scheduled downtime to run final rsync for 12 hours for 14 million inodes.

- **Santricity software was configured to monitor**
- **Configure to notify vendor support about failures**
- **NetApp four hour turn around support**
- **Send email notification to operations (24x7 Alarm monitoring by NERSC OTG)**
  - Operators staff trained to resolve simple issues
  - 24x7 escalation to CSG Systems Engineer on call
- **Nagios working on plugins to monitor disk array health status.**

# Computational Research and Theory (CRT) Building



- **Four story, 140,000 GSF building on the main LBL campus**
  - 300 offices in collaborative office setting (2 20Ksf floors)
  - 20K -> 29Ksf HPC floor
  - Mechanical floor
- **Energy efficient**
  - Year-round free air and water cooling
  - PUE < 1.1
  - LEED Gold design
- **42MW to building**
  - 12.5MW provisioned
- **Occupancy Early 2015**

Construction Webcam







# National Energy Research Scientific Computing Center