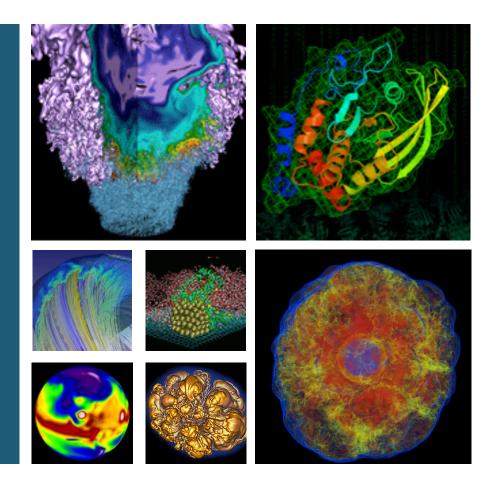
# PDSF at NERSC Site Report HEPiX Spring 2015





### James Botts, Lisa Gerhardt, Tony Quan, Iwona Sakrejda

March 20, 2015







**National Energy Research Scientific Computing Center** 





### **PDSF: Parallel Distributed Systems Facility**

**Continuous Operation Since 1996** 

















- serial, high throughput
- Univa Grid Engine batch system
- broad user base (including non-LHC)
- Fair share scheduling projects "buy in" to PDSF and the share tree is adjusted accordingly
- STAR Tier-1
- ALICE Tier-2
- ATLAS Tier-3



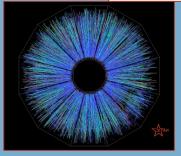


### **Snapshot of Science at PDSF**





ALICE, ATLAS, and STAR: 85% of usage Data analysis and storage as well as simulation





Daya Bay: Neutrino mixing angle experiment
Measures neutrinos from nuclear reactors in China
One of Science magazine's Top Ten
Breakthroughs of 2012



Majorana: Neutrinoless double beta decay experiment
Measuring the mass of the neutrino, is the neutrino its own antiparticle?

LUX: Large Underground Xenon dark matter experiment







### **PDSF Changes**



- retirement of Iwona Sakrejda in July 2015
- Moved users to global homes (GPFS) from old local storage
- 20 additional nodes deployed, increasing TFlops by 40%
- completed migration to SL6
- pilot migration from cfengine 3 to ansible for configuration management
- added two new local GPFS file systems
- retired 3 old file systems
- Upgraded GPFS from 3.4 to 3.5 to 4.1 (issues with mixed versions and deadlock tracing etc etc)





### Why bother with ansible?



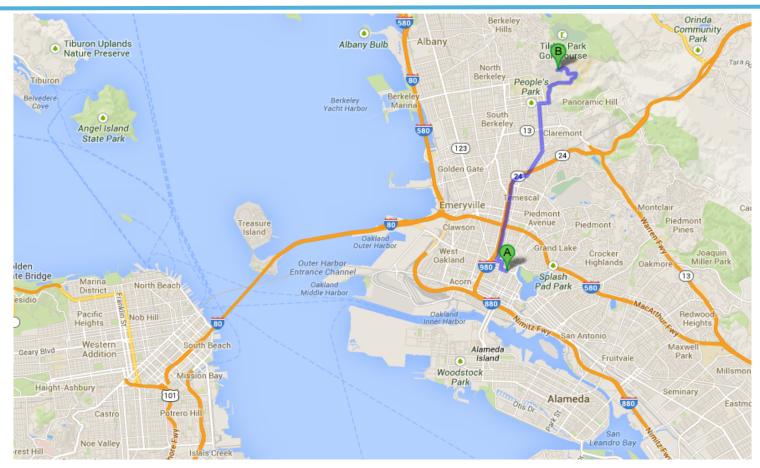
- incompatiblities between different versions of cfengine
   3 have made upgrading painful
- not a good experience with getting help from the cfengine community – only advice one gets is to upgrade
- Cray is probably going to use ansible for configuration and deployment management in their CLE 6/SMW 8 release (aka Rhine-Redwood)
- excuse to cut through decades of cruft and rewrite our configuration management
- easily supports push or pull model without deploying additional daemons or managing additional config files





#### **NERSC Move to CRT 2015**





#### ~6.4 miles from Downtown Oakland to LBNL





### **Computational Research and Theory Facility**









### Stage 0 – The current layout at OSF



#### **PDSF Section A - Legacy**

Legacy network 160 Computes

1540 cores

14.7 TFlops Theoretical

1 GbE

xCAT management server

4 MySQL database servers for STAR

2 Alice VO

3 XRootD redirectors (2 STAR, 1 ALICE)

4 Daya Bay login nodes

2 Batch System Servers

4 Grid Servers (OSG stack, data transfer nodes supporting BeStMan)

10 GPFS management and quorum nodes

3 Ganglia related servers

52 GPFS NSD servers and XRootD

managers

13 PDSF GPFS file systems 1.3 PB

#### PDSF Section B

68 Computes (20 added in 2014) 1088 cores 22.6 TFlops Theoretical 1 GbE, QDR InfiniBand

3 Login nodes for interactive use

#### PDSF Section C

18 Computes (added in 2014) 360 cores 7.2 TFlops Theoretical 1 GbE, QDR InfiniBand

2 Service nodes hosting VMs





### **Stage 1 – ~June 2015**



- PDSF Section C moves to CRT
- New PDSF
   Section D
   Shipped
   Directly to CRT
   and deployed
- UGE Batch server remains at OSF
- Grid Gateways, etc. stay at OSF
- The border router is at OSF
- Server functionality ported to VMs hosted at CRT

#### **PDSF Section A - Legacy**

Legacy network Computes - 14.7 TFlops

~80 Servers

13 PDSF GPFS file systems 1.3 PB

#### **PDSF Section B**

Computes - 22.6 TFlops

3 Login nodes

OSF

400 Gb/s

10.5 km

1 Gateway per 40-50 nodes

Each gateway: 2 x 40 Gb/s Bonded Ethernet

#### **PDSF Section C**

Computes - 7.2 TFlops

2 Service nodes hosting VMs

# New PDSF Section D

33 Computes 19.4 TFlops 1 GbE, QDR InfiniBand

#### New Storage

~ 2 PB xRootD storage for Alice

CRT

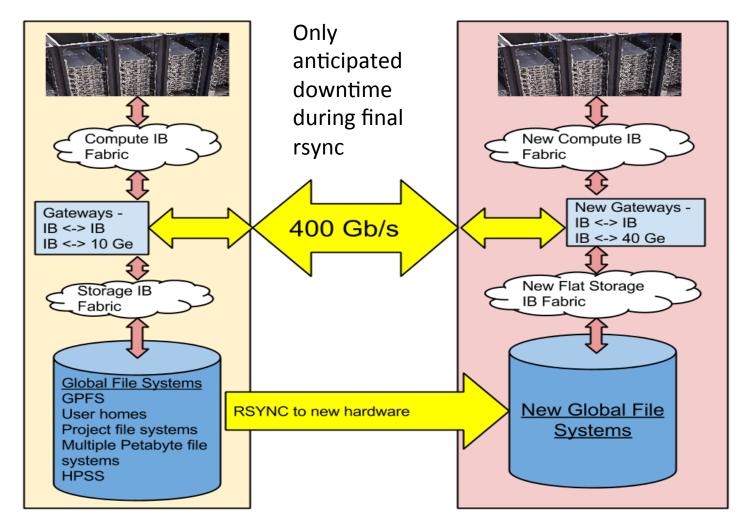




### **Global Storage Migration**



From June to October, **NERSC** global file systems will be migrated one by one to new hardware at the new facility.







## Stage 2 – July/August 2015



- All systems that will not be retired will be moved to the CRT
- New servers and storage will be deployed at the CRT to host the batch system and common files
- The batch system will be migrated from a legacy system at OSF to the new servers at CRT
- 3 PDSF GPFS file systems will be retired before the move
- The hardware of 3 PDSF GPFS file systems will move to the CRT, IB HCAs added to the NSDs

#### PDSF Section A - Legacy

Legacy network Computes - 14.7 TFlops

~80 Servers
Temporary resurrection
of 4 login nodes

8 PDSF GPFS file systems 470 TB

#### **OSF**

- All the systems at OSF will be retired shortly after the move
- The NERSC border router will move to CRT in October
- Complete IB FDR connectivity
- Gateways rebalanced

#### <u>PDSF</u>

Computes - 49.2 TFlops

3 Login nodes 2 VM Hosts

#### xRootD Storage

~2 PB, for Alice

3 PDSF GPFS file systems 680 TB

400 Gb/s

NEW 2 Servers
GPFS NSD server
UGE batch server
(HA)
with

6 TB directly attached storage

CRT







### Thank you.



