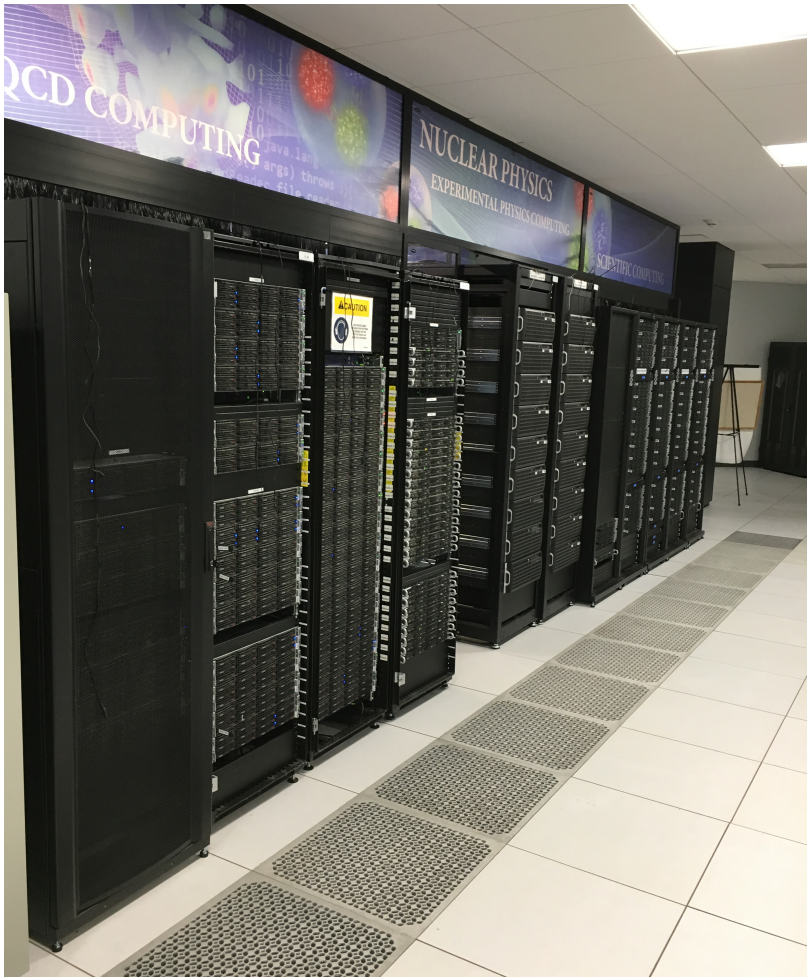
Scientific and High Performance Computing



Thomas Jefferson National
Accelerator Facility

https://scicomp.jlab.org/docs

Sandy Philpott

HEPiX PIC

October 8, 2018

Highlights since Fall16 @ LBNL

- CEBAF 12GeV Upgrade complete – Beam to all 4 Halls
  - New SSD gateway, from online DAQ to offline
  - Next computing and software readiness review in November
- Computing
  - Hardware installations
    - Additional KNL cluster for USQCD
    - Skylake compute nodes for ExpPhy
  - Resource sharing update for Theory, Experimental Physics
  - Access to offsite resources
- Disk Storage
  - 2 NFS ZFS servers; 4 Lustre OSS installations
  - LNet routing IB->OPA
  - Lustre 2.5 -> 2.11 upgrade planning
- Tape Storage
  - LTO-8 support, finally
- Facilities
  - Data Center modernization work completed
- A look ahead…

Computing

2018 hardware additions (after none in 2017):

- 18p: 180 KNL 7250 nodes, with Omni-Path
- farm18: 88 Skylake nodes, with Infiniband

complement the existing Theory and ExpPhy resources:

- 16p: 264 KNL 7230 nodes, OPA
- 12k: 40 Kepler nodes, FDR IB
- farm12-16: Sandy Bridge through Broadwell, IB

And, a new 19g (GPU) cluster procurement is expected soon, with the start of our new fiscal year …

We are in the process of switching to Slurm,

- from PBS/Torque/Maui

Offsite Computing Resources

- Open Science Grid
  - We have added an OSG submit host for GlueX; CLAS12 may also begin OSG activities
  - CHEP poster: “Limits of the HTCondor transfer system”
    - https://indico.cern.ch/event/587955/contributions/2937378/
- NERSC
  - We have adapted our SWIF workflow tool to ease submissions and Globus data transfers for users, and integrate with our local compute farm
- Cloud Services, as they are becoming more cost effective and we need bursts

- We’ve also set up a HepSim server at Jlab

Disks and Filesystems

Lustre and ZFS/NFS, 3 petabytes

- 2019: replace Lustre MDS, upgrade from 2.5 to 2.11
- 2018: adding 4 Lustre OSS
  - 60 disk shelves were slow to arrive …
  - Also adding SSDs into them, for a fast I/O DAQ gateway
    - At some point … (the first we received weren't supported!)
- 2017: added 2 NFS servers, running ZFS
- 2016: 4 Lustre OSS
- 2015: 2 Lustre OSS
- 2014: 4 Lustre OSS, MDS – retiring in 2019

Mass Storage

- IBM TS3500 Tape Library
  - 30 petabytes in 11 frames, room for 5 more
  - 8 LTO-8, 4 LTO-7, 8 LTO-6, 4 LTO-5 drives

When we relocated the library for the 2nd time within the Data Center in 2017 to its final location, IBM required that we move the frames only after unloading its 10,000 tapes – but we still had similar issues as the first time to resolve

Added LTO-7 drives in 2017, but after SC17, we bought LTO-8 drives. We skipped LTO-7 media and ordered LTO-M8, but never got it and had to cancel the order … Now we are using LTO-8 media since the spring, and are ironing out a few small problems with corrupt files (13 files on 4 tapes, out of several hundred tapes…)

Migration off of all LTO-4 media is almost complete

Facilities Update

To meet DOE goal of PUE of 1.4, power and cooling were refurbished - completed in 2017

- 800 KW UPS
- 3 200 KW air handlers + refurbished 180
- All file, interactive, infrastructure servers moved to dual fed power, one side of which is generator backed (99.99% uptime)
- Rolling cluster outages relocated and re-racked to 18-20 KW/rack as opposed to 10-12 KW
- Downtime <2% for the year

Looking ahead…

SSDs into production in the ZFS and Lustre fileservers

Shared Slurm environment, for resource sharing between Theory and Experimental Physics computing

Automate suppression of users with “bad I/O impacts”

Lustre 2.5 to 2.11 upgrade

More offsite resource usage, at NERSC, OSG, and also at cloud providers as it is becoming more cost effective

Procuring a new USQCD cluster, likely GPUs

Joint WLCG, OSG, HEP Software Foundation meeting @ JLab, March 18-22, 2019