JLab Scientific Computing

Thomas Jefferson National Accelerator Facility

https://scicomp.jlab.org/docs

Sandy Philpott

HEPiX BNL

Oct 13, 2015

Updates since Oxford

- Computing
  - USQCD 2016 hardware acquisition planning
- Disk Storage
  - openZFS , Lustre 2.5 status
- Tape Storage
  - LTO-6 into production
  - TS3500 library frame relocation/reconfiguration
- Facilities
  - Data Center ongoing work through 2016
- Looking ahead …

Computing

USQCD – 3 sites: JLab, FNAL, BNL

FY16 procurement will be installed at Jlab; ~ $1M

Investigating several possibilities …

- Intel Xeon Phi / Knights Landing
  - Single socket, self hosting, largest on-package memory, >64 cores
- NVIDIA Pascal GPU, CUDA
- Intel Broadwell CPU server

Consideration factors

- hardware availability timeline
- high speed network – 100 Gbps price/performance
- reflective benchmarks
- available configurations

Optional FY17 upgrade will be included in the award; could include Experimental Physics combination purchase

Disk Storage

Lustre 2.5.3 upgrade (almost) complete

1.3 PB on 30 OSSs each with 30 * 2/3/4 TB disks; 8+2 RAID6 or RAID-Z2

- 8.1 GB / sec aggregate bandwidth, 100 MB/s – 1 GB/s single stream
- Mix of 8+2 RAID-6 and 8+2 RAID-Z2 OSTs
- retiring 14 2009 servers: 24 * 1 TB systems - repurpose elsewhere as large scratch

2014 disk hardware - in production

4 dual Xeon E5-2630v2 CPUs, 30*4TB and 4*500GB SATA Enterprise disk drives, LSI 9361-8I RAID Controller with backup, 2*QDR ConnectX3 ports

- With RAID-Z, don't need hardware RAID … JBOD …
- Ongoing issues with stability; 3 of 4 machines have had crashes since June 27 scheduled outage … likely bad RAM?

2015 hardware - just received

2 dual Xeon E5-2630v2 6 core 2.6GHz, 128GB RAM, SAS3, 40*8TB Hitachi Ultrastar, 6*400GB Seagate SSD, LSI 9300-8E HBA, QDR on motherboard, FDR add-on

- JBOD
- Fully redundant – 2 shelves connect 2 to hosts
- Into production in November

Storage Evolution

Dell MDS in producion

- 2 R720s, E5-2620 v2 2.1GHz 6C, 64 GB RDIMM, 2 * 500GB 7.2K SATA
- PowerVault MD3200 6G SAS, dual 2G Cache Controller, 6 * 600GB 10K disk

Upgrade from Lustre1.8 to 2.5 - almost complete; over 1PB

- 2 pools: fastest/newest, and older/slower – implementing with 6 SSDs in the new hardware
- Begin using striping, and all stripes will be fast (or all slow)
- Inactive projects moved from the main partition into the older, slower partition, freeing up highest performance disk space for active projects
- Use openZFS with JBOD

… and following CEPH developments

Mass Storage

- IBM TS3500 Tape Library
  - > 10 PB written; duplicates of all raw data, stored in tape vault
  - 6 LTO-6 drives into production for all writes
    - 6 LTO-6 drives
    - 8 LTO-5 drives
  - Replaced 8 440-slot frames with 3 1320 slot frames
    - From 14 to 9 frames; library supports 7 more
  - Relocated across the room within the Data Center

- Continue to increase capacity within the same library
- Will still need a second tape library, likely in the 2018 timeframe
- New write-through-to-tape filesystem to automatically move oldest files to tape goes into production in November with new hardware installation
  - Our poor man's HSM

Facilities Update

Computer Center Efficiency Upgrade and Consolidation

- Computer Center HVAC and power improvements in 2015 to allow consolidation of the Lab computer and data centers to assist in meeting DOE Computer Center power efficiency goal of 1.4 PUE
- Double cooling and power capacities
- Increase power density to 16-18 KW/rack

- Staged approach, to minimize downtime

Looking ahead

Computer Center Efficiency Upgrade and Consolidation continues…

Investigate / support CentOS 7

need Lustre 2.5 client

Deploy first cluster of LQCD-ext II, in the current location of our 2009/2010 clusters - reusing existing power and cooling infrastructure lowers installation costs

Install second tape library, as growth exceeds current library with 12GeV accelerator and experiments

Data management, mining, indexing for Physics discovery …