JLab Scientific Computing

Thomas Jefferson National Accelerator Facility

https://scicomp.jlab.org/docs

Sandy Philpott

HEPiX DESY

April 19, 2016

Updates since BNL

- Computing
  - USQCD 2016 hardware acquisition purchase request is in
  - Software workshop at JLab in March
- Disk Storage
  - openZFS , Lustre 2.5 status
- Tape Storage
  - LTO-6 production
  - TS3500 library frame relocation/reconfiguration, round 2
- 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; OmniPath or IB
- reflective benchmarks
- available configurations

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

JLab Workshop

“Future Trends in Nuclear Physics Computing”

held at JLab in March

https://www.jlab.org/conferences/trends2016/

Goal: discuss trends in scientific computing,collect ideas on how to improve analysis workflows at existing nuclear physics programs, work towards analysis techniques and tools for future projects like the Electron-Ion Collider

One outcome – GlueX collaboration (JLab Hall D) will revisit using OSG resources for computing cycles

Disk Storage

Lustre 2.5.3

1.6 PB on 21 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
- retired 14 2009/2010 servers: 24 * 1 TB systems - repurpose elsewhere as large scratch

2014 disk hardware

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 in production, except SSDs

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; currently installed non-HA

2016 hardware purchase – require 12 MB/s/TB

- Purchased Intel Enterprise Edition; scheduled for June install

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
- ldiskfs

Lustre1.8 to 2.5 upgrade completed last fall - over 1PB

- Intend to have 3 pools:
  - Fast - will implement with 6 SSDs in the new hardware
  - Production – all RAID 6 or RAID z2
  - Work – smaller files, 3 mirrored sets of 4 disks
- 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 we still follow CEPH developments with interest

Storage Woes

Remember last HEPiX, I mentioned we were having issues with the 4 2014 fileservers?

- CATERRs (Catastrophic Errors) – systems hang
  - oss1401 - First 2 CATERRs in March! 03/12/16,03/27/16
  - oss1402 - up to CATERR 38 on 01/30/16; first 08/24/15; failed/replaced disk controller 01/18/16 lost 30 TB
  - oss1403 - up to CATERR 49 on 03/12/16; first 10/08/15
  - oss1404 - into production 03/31/16; RMAd in December-new motherboard; in testbed Jan-Mar for benchmarking configurations, work pool, HBA
- Updated motherboard BIOS, disk controller firmware, and zfs_arc_max
- Now working directly with Supermicro
  - Newer motherboard BIOS just released on Thursday
  - Requires update to IPMI v3
  - Firewall opening for UDP port with their diagnostic server

Mass Storage

- IBM TS3500 Tape Library
  - 12 PB written; duplicates of all raw data, stored in tape vault
  - 6 LTO-6 drives into production for all writes
    - 8 LTO-6 drives, 6 LTO-5 drives
  - Replaced 8 440-slot frames with 3 1320 slot frames
    - From 14 to 10 frames; library supports 6 more
  - Relocated across the room within the Data Center
  - Still missing a few tapes?!
  - Continue to increase capacity within the same library
    - Will need a second tape library, likely in the 2018 timeframe
- New write-through-to-tape filesystem automatically moves oldest files to tape; our poor man's HSM

When we relocate the library for the 2nd time within the Data Center, IBM will require that we move the frames with the 10,000 tapes unloaded

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
- Ready for new 2016 hardware procurement to arrive this summer

Looking ahead

Computer Center Efficiency Upgrade and Consolidation continues…

Investigate / support CentOS 7, including Lustre 2.x 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 - ~2018 timeframe?

Data management, mining, indexing for Physics discovery …