Theory HPC & Experimental Physics

Thomas Jefferson National Accelerator Facility

www.jlab.org

Sandy Philpott

HEPiX Oxford

March 23-27, 2015

Updates

Since our Nebraska meeting…

- Computing
  - Haswells into production
  - Continue core swaps for best-match load balancing
- Disk Storage
  - ZFS on Linux into production
  - Lustre 2.5 upgrade beginning
- Tape Storage
  - LTO-6 into production
  - TS3500 library frame consolidation
- Network
  - Redundant 10 / 40 Gigabit from Online DAQ to Offline MSS and farm
- Facilities
  - Data Center rework starts Oct 1
- Looking ahead

Computing

Latest procurement – 104 Experimental Physics nodes

- dual Intel E5-2670v3 Haswell 12 core, 2.3 GHz, 32 GB DDR4-2133 memory
- Adds 2500 cores to existing 1400 core farm; retire 2009 nodes
- In production at CentOS 6.5
- Systems are memory lean, so still working with users to
  - move to multi-threaded code
  - request only memory that job requires, so as not to block other jobs from running

Additionally, working to understand some high I/O requirements

Load sharing now automated between experimental and theoretical clusters

New workflow tool developed for better job and tape I/O management

Acquired NVIDIA K80 GPU system, for further USQCD code development

Disk Storage

Currently

- Lustre 1 PB on 30 OSSs each with 30 * 1/2/3 TB disks, 3 8+2 RAID6
  - 8.1 GB / sec aggregate bandwidth, 100 MB/s – 1 GB/s single stream
- ZFS servers 250 TB
  - **Move to ZFS on Linux** - retire 5 year old SunFire Thors, continue using our 2 year old Oracle 320 appliance

New 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 …

Storage Evolution

Lustre Upgrade and Partitioning

New Dell MDS installed

- 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 1.8 to 2.5, partition by performance

- 2 pools: fastest/newest, and older/slower
- Begin using striping, and all stripes will be fast (or all slow)
- By summer 2015, this will be in production, with “inactive” projects moved from the main partition into the older, slower partition, freeing up highest performance disk space for active projects
- Use ZFS on Linux, on top of JBOD

Planning to test 2 DDN Infinite Memory Engine servers (SSDs)

Interested in CEPH developments

Mass Storage

- IBM TS3500 Tape Library with 14 LTO drives, 14 frames
  - Currently 10 PB stored
  - Ready to move to production LTO-6
  - Replacing 8 low density frames with 3 high density frames
  - Continue to increase capacity within the same library
  - Will still need a second tape library in the 2017 timeframe

- New write-through-to-tape filesystem to automatically move oldest files to tape
  - Poor man's HSM

Network

Installed redundant gateways between online DAQ and offline

- Using 2 hosts, each with 10 gigabit Ethernet / 40 gigabit Infiniband interfaces
- Can install additional gateways as needed
- Makes Lustre staging area available via NFS

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

2015 – 2016

Computer Center Efficiency Upgrade and Consolidation

Operate current HPC resources: run the late Fall 2009 clusters through June 2015, and mid 2010 clusters through June 2016 -- longer than usual due to absence of hardware for 2015

Experimental Physics grows to match the size of LQCD, enabling efficient load balancing (with careful balance tracking)

Investigate / support CentOS 7

2016 – 2017

JLab will be the deployment site for the first cluster of LQCD-ext II. This resource will be installed in the current location of our 2009/2010 clusters (same power and cooling, thus lower installation costs)

Install second tape library in 2017

Continue to grow physics farm to meet 12GeV computing requirements