JLab Scientific Computing: Theory HPC & Experimental Physics

Thomas Jefferson National Accelerator Facility
Newport News, VA
www.jlab.org

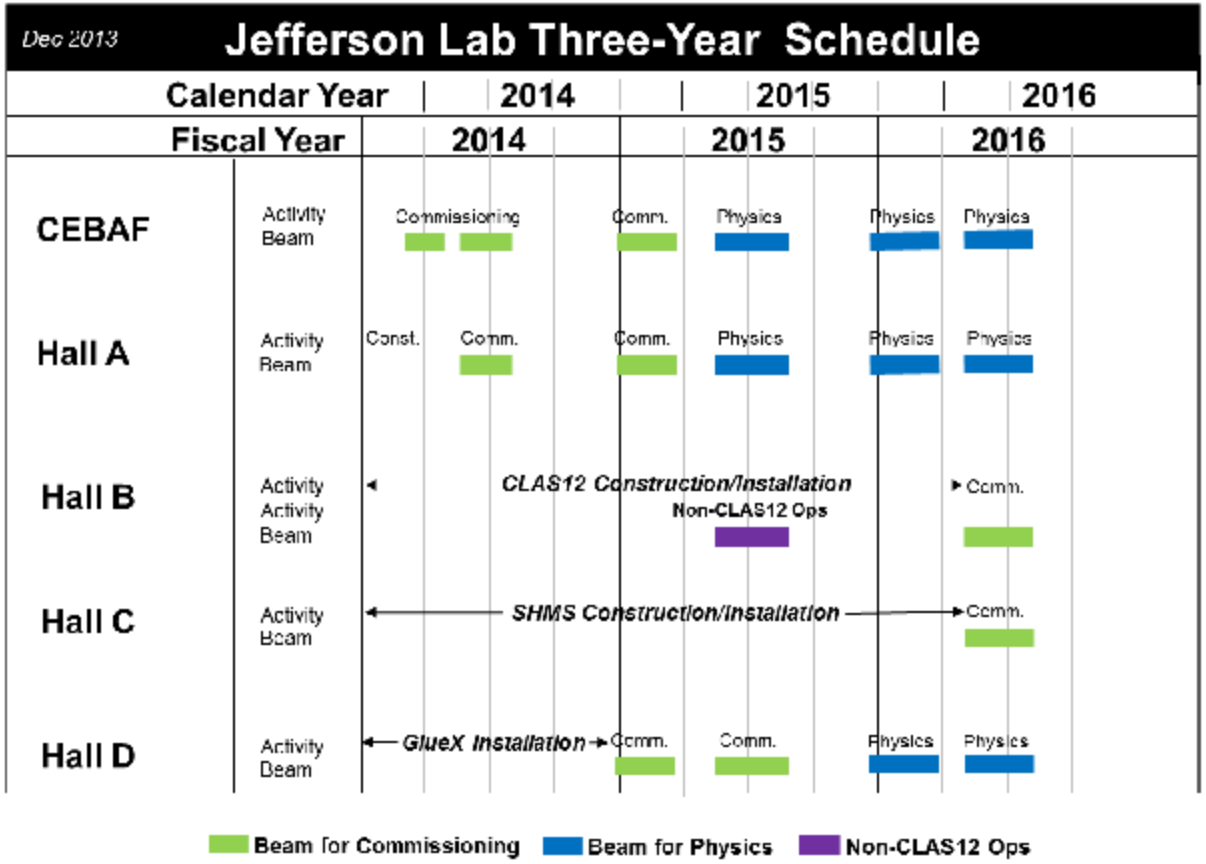
Sandy Philpott
HEPiX Nebraska a– October 13, 2014

Updates

Since our Annecy meeting…

- 12 GeV Accelerator status
- Computing
  - Haswells ordered
  - Continue core swaps for best-match load balancing
- Disk Storage
  - Newest storage servers
  - openZFS
  - New Lustre MDS system
  - Path to Lustre 2.5
- Facilities update
- Looking ahead

12 GeV Accelerator Status



Computing

Latest procurement – 104 Experimental Physics nodes

- dual Intel E5-2670v3 Haswell 12 core, 2.3 GHz, 32 GB DDR4-2133 memory
- GlueX (Hall D) testing indicates these cores are 50% faster than the Sandy Bridge / Ivy Bridge cores already in the farm
  - measured by events/sec/core on an 18 core part also at 2.3 GHz; system scaled linearly through all cores, so no memory bandwidth bottlenecks at a mere 12 cores per CPU

New trick helps USQCD , neutral for Experimental Physics:

- USQCD queue for 16 core nodes is always full; queue for 8 core nodes often sags, so give two 8 core nodes to experimental physics, take in return one 16 core node
- similar to our core exchange approach described in the spring talk, but now takes into account the type of load
- currently manual, soon (we hope) automatic
  –

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 procured

- 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

- Plan 2 pools: fastest/newest, and older/slower
- Begin using striping, and all stripes will be fast (or all slow)
- By the end of 2014, 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 openZFS, rather than ext4 / ldiskfs ?

Other sites’ Lustre migration plans and experience?

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 goals.

- Staged approach, to minimize downtime

Looking ahead

Rest of 2014 …

Increase the farm by ~ double as Halls come online. Upgrade Lustre; deploy 4 new OSSs (30 * 4 TB RAID6); move to ZFS on Linux. Begin using Puppet? Deploy workflow tool for farm. Continue to automate core sharing / load balancing. Hire a 3rd SysAdmin.

2015 – 2016

Computer Center Efficiency Upgrade and Consolidation

Operate current HPC resources (minus oldest gaming cards): 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).

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 the 9q / 10q clusters (same power and cooling, thus lower installation costs).

Continue to grow physics farm to meet 12GeV computing requirements. Final configuration will use ~20,000 cores.