Thomas Jefferson
National Accelerator Facility

JLab TOP500

Sandy Philpott

www.jlab.org/hpc

HEPiX - UM Atlas Tier 2 – Oct 28, 2013

Jefferson Lab Computing

Clusters

- HPC Accelerated – GPUs, plus MICs for R&D
  - **TOP500! #364**
- HPC Infiniband – FDR, QDR
- Physics Data Analysis – DDR, SDR IB (recycled)

Storage

- Disk – Lustre & ZFS over IB, NFS over Ethernet
- Tape – IBM library, LTO drives

Upgraded CEBAF 6->12GeV !

Clusters

In 2012, JLab was awarded the computing hardware for US Lattice QCD.

The hardware was divided between traditional IB, & GPU accelerated nodes.

12s “2012 Sandy Bridge” – Atipa Technologies

- 276 SuperMicro dual 8 core 2.0GHz nodes
- 4 nodes in 2u, 32 GB RAM, 500GB disk
- QDR, full bi-sectional bandwidth, leaf and spine
  - Mellanox onboard hosts; Qlogic switches
- CentOS 6.2
- Power upgraded to 30 amp 5 wire 3 phase

- Short of TOP500 in 11/2012 – barely
  - Needed to include 32 more 12s nodes that were DDR retrofitted
    - But HPL code didn’t run on them (?)
  - Speed Step/Turbo Mods finally tripped the 12s power - Game Over.

- 12s Qlogic ←→ Mellanox core: problem: fiber links degraded
  - 4x 2.5Gbps, not 4x 10Gpbs – remains unresolved
    - Using a “bandaid” Mellanox in between, over copper

Clusters (cont)

12k: “2012 Kepler”: Seneca Data

TOP500 06/2013! **#364**

- 42 nodes, 4 NVIDIA K20 each, FDR IB, openmpi 1.6
  - **117 TFlop/s, 2652 cores**

12m: “2012 MIC”: Seneca Data

- 18 nodes, 4 Intel Xeon Phi each, FDR IB
- research & development cluster

Physics data analysis cluster – batch farm

- Save procurement overhead when purchased together with HPC
- 32 **farm12** nodes identical (interchangeable!) with **12s** compute nodes except
  - 2 disks rather than 1
  - DDR Infiniband recycling, from old 2007 cluster

  - Consider moving from Torque/Maui to SLURM… ?

Storage - Disk / Filesystem

- **Lustre** 1.8.8wc-1 – AMAX, & ICC

    1PB+ on 30 OSSs, 2 or 3 OSTs each, 30 * 1 to 3 TB disks, LSI controllers – Amax & ICC

    Still unresolved, but we are decommissioning - slow writes on 14 old 3ware 24 * 1TB disk systems; Happened at upgrade to 1.6.6-wc1 (?)

  Migrate MDS from Dell MD3000 RAID 10 to Dell MD1220 with 3 SSDs

    (Dell will no longer support the 3-yr-old MD3000 (!)

  Add backups

  Investigate Lustre 2.x (2.4, 2.5?) …

- **ZFS**

    Oracle SunFire X4540 Thors – still 5 of them – run another year or two…

    Oracle 7320 appliance added

      2 head units for redundancy

      2 shelves, 20 and 24 3TB disks, 1 with write accelerator SSD

      /home over Infiniband would hang; still unresolved; serve over Ethernet instead

  Interested in OpenZFS release! …

Storage - MSS

Tape & Mass Storage System

IBM TS3500 Library

- Installed 14th of 16 possible frames, 9400 slots
  14 LTO drives: **2 new LTO6**, 10 LTO5, 2 LTO4
- All new writes to LTO5 for now
- Migrate data from LTO4s in background
  - frees slots, almost 1 LTO5 slot for 2 LTO4s
  - exchange blanked LTO4s for new LTO5/6 cartridges

JASMine, local JLab software, used for management

Data Preservation

http://scicomp.jlab.org/scicomp/#/static/data-management-plan

6->12GeV Accelerator Upgrade

- Accelerator returning to operation after 18 month upgrade
  - One additional Experimental Hall D
  - Double the current data rates in existing Halls A,B,C
  - 15PB yearly at full operation; use LTO-7, …
- External IT/Computing reviews
  - Ensure readiness of data acquisition and analysis on day 1
  - Data challenges
- Workflow tools under development for processing large data sets
- Starting to auto-rebuild compute nodes between HPC and the batch farm on demand
  - Newest Ivy Bridge installed last week are working in both clusters
- Globus Online – users love it!
  - Gateway offsite data transfer node updated to 10GigE / QDR IB