Jefferson Lab Site Report

HEPiX

October 26, 2009

Sandy Philpott

Sandy.Philpott@jlab.org

Jlab News

- Celebrating 25th Anniversary: 1984-2009
  - Visit from U.S. Secretary of Energy Dr. Steven Chu

- 12GeV Project
  - Continuous Electron Beam Accelerator doubles energy
    - 6GeV -> 12GeV
  - New Experimental Hall D, in addition to A,B,C upgrades
  - Increases raw experimental data, therefore scientific computing
  - Online in 2014

- $5M ARRA money for Lattice QCD
  - ARRA 2009: American Recovery and Reinvestment Act
  - Spend immediately
  - Increase from 3TF to 20TF
    - Maybe up to 60TF with GPUs
  - Increase data to mass storage system; how much?

LQCD Theory Clusters

- 6n – “2006 Infiniband”
  - 280 3.0GHz single cpu, dual core Pentiums D
  - SDR Infiniband (10Gb/s)
  - 70 nodes since reallocated to analysis farm
  - Next summer becomes all farm when 9q is operational
- 7n – “2007 Infiniband”
  - 396 1.9GHz dual cpu, quad core Opterons
  - DDR IB (20Gb/s)

LQCD Theory Clusters

- 9q – “2009 Quad Infiniband - ARRA Cluster”

  Storage – Whitebox – 14 AMAX servers
  - OpenSolaris or Solaris x86 w/ ZFS
  - Lustre

  Compute nodes _ Dell PowerEdge R410

  Nehalems – dual socket, quad core 2.4GHz

  QDR Infiniband (40 Gb/s)

  24GB RAM – DDR3-1333

  GPU nodes – KOI SuperMicro, 4*2GB GPUs
  - 15% of nodes, 80% of compute power
  - $0.20 / MF on std cluster
  - $0.01/MF on GPUs

  - Powerful and cheap, BUT difficult to program; disruptive technology
  - When C++ is supported, programming may become easier

Investigated Skyld – diskless, contained cluster – Not for us.

Auger, Analysis Cluster

- No spending on compute nodes since 2006, now purchasing again
- Nehalems – Intel X5530 dual cpu, quad core, 24MB RAM, 500GB SATA disk
  - Runs 12-14 jobs typically
  - Local disk is now a bottleneck
- Operating System switch: Fedora 8 32 bit → CentOS 5.3 64 bit
  - RedHat EL5.3 compatibility, for Users' desktops
  - RPC timeouts an unresolved issue
- Physics Applications supported by Jlab Physicists, Users
  - Cernlib - status on 64 bit?
  - Root
  - Geant4

- Benchmarking Jlab Physics analysis codes…
- Still lab-centric computing model, vs. grid

JASMine, Mass Storage

JASMine system upgraded this year, to accommodate new hardware:

IBM TS3500 tape library installed; 8 LTO-4 drives, 4400 slots

- Fully operational, >4PB; 10% capacity left
- Data set size increased recently; investigating better data management
  - Ex: 30->120TB raw + 100->500TB analysis data
    - Eject, or expand?
    - Upgrade to LTO-5 when available in 2010
      - Use data movers with SDR IB?

StorageTek Powderhorn silos replaced

- 20 9940B drives and 12000 tapes
- All >2PB data migrated
- Physically removed 2 silos this summer

Sun Fire X4500, X4540s for Cache - Disk copy of data on tape

- 60TB
- farm, general cache automanaged
- experiment-managed DST
  - Investigating DST cache usage and practicality
- Migrate to Lustre? Fast, recoverable

Facilities

- 2 physical rooms, same building, different floors:
  - 1990 – Central Computing Facility
  - 2006 – Data Storage Facility

- Power capacity
- Cooling capacity

- Requirements for ARRA
  - 380KW in 2 phases, 190KW each Nov/Feb
    - Add power; adhere to Procurement regulations
  - Have to move 100KW downstairs to upstairs now

Virtualization

80 production VMs. Plan 30-50 more over the next year.

- 5 ESX hosts:
  - Three Dell PowerEdge 2950III, Dual 2.8Ghz quad core Xeons, 64 GB RAM each
  - Two Dell PowerEdge R710, Dual 2.6Ghz quad core Nehalems, 96 GB RAM each
  - VMware ESX 3.5 (plan to move to vSphere 4 next year)
  - VMware VCenter Management server

Backend Storage

Two Lefhand iSCSI arrays. Each volume is 2 way replicated. Can take one array down for maintenance without impacting ESX at all. SAN based snapshots can be used for recovery of single VM (or single file with some work) or entire volume.

Notes

- Memory is the bottleneck for us. We were using 80% of a ESX host's memory, but less than 20% of it's cpu. (before upgrade from 32GB to 64GB)
  - - When using iSCSI for storage, VM backups were not as straight forward as when using NFS for storage. Settled on vRanger from Vizioncore.

Desktops

- In-house support; recommend hardware and configuration
  - RedHat Linux, Windows Vista, XP/Pro
  - Central applications, including MS-Office, anti-virus pushed
  - Windows – firewall off onsite, on offsite
- Security patches pushed monthly, critical ones immediately
  - Dedicated-use systems - SysAdmin pulls during downtime
- Macs – separate network; not centrally managed
  - Some apps provided, including anti-virus
- Guest networks
  - wired and wireless; appear outside of Jlab network

Other Computing News

- Network Asset Management - NAM
  - NetReg - registration
  - NetMan - management

  Auto-VLANed immediately
  - no time between plug-in and discovery

CyberSecurity annual self-assessment almost complete

Summary

ARRA cluster installation

Lustre, openSolaris or Solaris x86 w/ zfs

Issue is data integrity – check on read

- In hardware on some controllers, like DDN
- Not on SATABeast
- Lustre doesn't
- Zfs does

Other GPU applications?

Recycle systems HPC -> farm after 3+ years

Recycle SDR IB from HPC into batch farm and data movers for "free"?

Sun Fires on Infiniband?

ITIL?

12GeV era computing, by 2014