

# ***JLab Scientific Computing***

***Experimental Physics***

***Lattice QCD***



***Sandy Philpott***

[sandy.philpott@jlab.org](mailto:sandy.philpott@jlab.org)

*October 8, 2012*

*EIC Detector R&D Simulation Workshop*

*Brookhaven National Lab*

# Overview

- Scientific Computing at Jefferson Lab
- Software and Hardware Systems
  - Mass Storage
  - Clusters
  - Filesystems
  - Network
  - Infrastructure
  - Physics Software
- Management
- Staffing
- 12GeV Upgrade Requirements
  - Cores
  - Disk
  - Tape
- Summary

# Scientific Computing

## Compute Clusters

- Experimental Physics: 2 racks
  - Data Analysis “Farm”: 1200 cores
- Theoretical Physics
  - Lattice Quantum Chromo Dynamics, part of the USQCD national collaboration
  - High Performance Computing: 37 racks
    - Infiniband: 11,000 cores
    - GPUs – disruptive technology – 530 gpus, 200,000 cores
      - » Adding 6 racks this fall (VNIDIA GPU and Intel MIC)

## Mass Storage

- Tape Library
  - Raw experimental physics data, processed results
  - Other lab-supported efforts, mostly LQCD

# Data Analysis Cluster

- **Auger** - Jlab written software for user interface, job management
- **Maui/PBS/Torque** open source resource manager/scheduler
- **Mysql** open source database
  
- 80 compute nodes, 700 cores – **CentOS 5.3** 64-bit
  - mostly dual quad-core Nehalem, 24GB RAM
  - also 6 quad eight-core Opteron, 64GB RAM
    - support move to multithreaded codes
- **New!** 32 compute nodes, 512 cores – **CentOS 6.2** 64-bit
  - 16-core Sandy Bridge, 32GB RAM

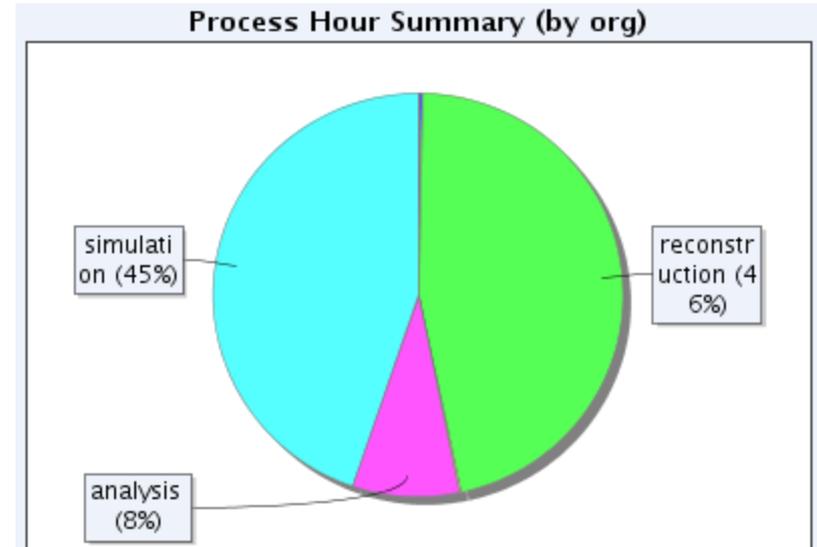
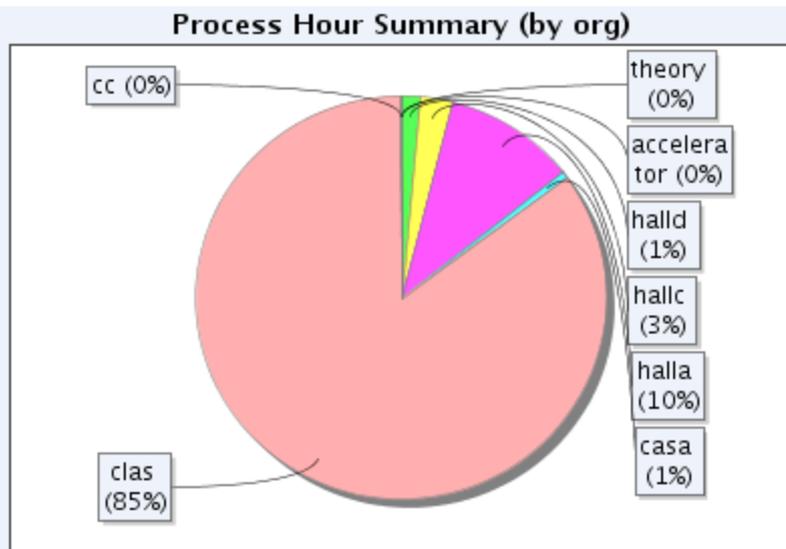
Moore's Law holds true; replacing oldest  $\frac{1}{4}$  compute nodes every 4 years keeps pace with increasing computing requirements.

Also, reallocate 4-year-old HPC compute nodes to the farm, after nodes don't function well in large parallel system but are fine individually.

# LQCD Clusters

- **Infiniband clusters**
  - **12s – 250 16-core Sandy Bridge**
  - **10q – 224 8-core Westmere**
  - **9q – 320 8-core Nehalem**
- **GPU clusters – nodes with 4 cards, professional or gamer**
  - **12k, 12m coming this fall**
    - **NVIDIA K20 and Intel MIC**
  - **11g – 8 nodes, dual IB, C2050 and M2050**
  - **10g – 50 nodes, C2050 and GTX 480,530**
  - **9g – 65 nodes, GTX 285, 480, 580**

# Data Analysis 2012



7.5M process hours delivered in FY12.

For comparison, LQCD delivered 81M process hours in FY12.

# Mass Storage System

**Tape Library-** IBM TS3500, LTO tape drives, >1GB/s aggregate bandwidth

- 13 frames, 8000 slots; Capacity ~12PB
  - expandable to 16 frames, 1300 slots/frame
- 12 LTO5 drives – all writes, copying LTO4s and duplicate raw data tapes
  - 1.5 TB/tape, 140MB/s
- 4 LTO4 drives – can be replaced
  - 1TB/tape, 120MB/s

**JASMine software** - JLab written; modular, distributed architecture; scalable

- manages data on tape plus 50TB disk cache

**12GeV projections include 2<sup>nd</sup> tape library procurement in 2014**

- expecting ~15PB yearly storage requirements in full 12GeV production
- LTO6 drives and media expected FY14, in advance of 12GeV turn-on
  - 3.2 TB /tape, 210MB/s
- 15-frame library with 1300 slots each = 19,500 slots x 3.2TB ea = ~60PB
- Upgrade to LTO7 expected in FY18
  - 6.4TB/tape, 300MB/s = ~120PB capacity

Scale-out solution – add more datamovers if more bandwidth needed

# Filesystems

**zfs** filesystems exported via **NFS** (Network File System) over trunked gigabit Ethernet

- Continue to use for project work data; migrating to Infiniband
- 230TB production total
  - Mostly Experimental Physics
- Space managed by individual projects

**Lustre** filesystem over Infiniband

- Volatile scratch space, automanaged
- Cache – copy of data on tape in mass storage system
- 24 servers, scale-out solution
- 800TB production total
  - 200TB for Experimental Physics
- Improves scalability and I/O performance.
- Require a specific kernel and Lustre client.

# Network

## Local

- Compute nodes
  - Oldest are 1Gbps Ethernet
  - Latest are 10Gbps Infiniband for Experimental Physics, 40Gbps for LQCD
- Fileservers
  - Trunked gigabit Ethernet filesystems, 4Gbps
  - Infiniband, 10 – 40Gbps interfaces
- Data acquisition systems to Computing Center
  - 10Gbps Ethernet
    - Plan to test 40Gbps Infiniband next year
      - 4x bandwidth, lower CPU load

## Wide area

- ESNNet: DoE's Energy Sciences Network via E-LITE
- 10 gigabit shared; can be assigned our own lambda when needed
- Used for remote logins, offsite data transfers

# Infrastructure

- Space
  - 7400 square feet today; 50% used
  - 2000 square foot conference room carved out; could be reclaimed if necessary
- Power
  - ~650KW, including PDU loss
  - can add another large UPS
  - Expect to increase power density 1.5x
- Cooling
  - Today: ~470KW; Fall: ~550KW with a new air handler
    - limited by 8” pipe
  - Fall 2014: >650KW with upgraded 12” pipe

# Physics Software

- Experimental Physics - supported by Physicists at Jlab
  - Geant4
  - Root
  - Cernlib
- LQCD codes – USQCD collaboration
  - Chroma
  - QMP

## CUDA on GPUs

- Software must be heavily optimized; USQCD codes are
- 10x performance of investment (\$800K GPU expenditure yields performance of \$8M CPU core investment)

# Management

## Information Technology

- The Scientific Computing Group is in JLab's Information Technology Division.
- IT/SciComp manages computing, disk and tape storage systems for Experimental Physics as a **Service Provider (SP)**.
- **Funding** for staffing, hardware, and storage media are in the **Physics budget**.
- JLab's annual work plans (**AWP**) define labor, funding, and requirements.

## Physics

- Physics computing requirements are gathered and conveyed by the **Physics Computing Coordinator**, Graham Heyes, to the **Head of Scientific Computing**.
- **Offline Computing Coordinators** for each hall manage their requirements.
  - Physics experiments/run groups communicates with their hall offline computing coordinator for resource allocation.
- **Regular meetings** occur with Physics and IT -- SciComp & CNI -- to ensure requirements are met and issues are addressed.

## JLab

- **Users Group Board of Directors** includes a Computing coordinator, now Mark Ito.
- **JLab IT Steering Committee** includes Physics Computing Coordinator, the UGBoD Computing contact, and several physicists.

# Staffing

## 10 FTE:

1 Manager, 1 Theory Physicist, 4 Development, 4 Operations (1 open)

### Operations today:

CPU servers: 1.5

Disk servers: 0.5

Robots: 0.5

Services: 0.5

User Support: 0.5

### 12 GeV:

+ 0.5 computing

< 0.5 all else

# 12GeV Requirements

## Cores

- Based on 2011 Intel Westmere 2.533 core, \$200

Hall\Year	2014	2015	2016
A	20	40	100
B	550	4400	11,000
C	20	40	100
D	500	5000	10,000
Total	1090	9480	21,200
Cost/core	\$80	\$50	\$35
#cores/node	32	32	64
#nodes/rack	64	64	64

Base spending plan includes \$50K/yr on computing.

At 12GeV operations in 2016, 7 racks.

# 12GeV Requirements

Storage – Tape (TB)/year

	2014	2015	2016
A	170	410	660
B	150	5200	9500
C	600	250	1750
D	200	1800	3600
Total	1120	7660	15510
Cost/tape	\$ 40+30 (LTO5)	\$60+30 (LTO6)	\$45+30 (LTO6)

# 12GeV Requirements

Storage – Disk (TB)/year

\$300/TB in 2012

	2014	2015	2016
A	20	50	110
B	90	1300	2700
C	60	125	175
D	50	750	2000
Total	220	2225	4985
Cost/TB	\$160	\$120	\$90

# Summary

JLab scientific computing systems can support 4096-core jobs, likely useful for EIC detector simulation.

Resources are available for science groups to run test and benchmarking jobs on batch compute farm, Infiniband clusters, and accelerated gpu clusters.

Future system growth is requirements driven, based on information provided to IT from all projects.

Ample space, cooling, and power are available for 12GeV era computing and storage needs in CEBAF Center F112, the IT Data Center.

Websites:            <http://www.jlab.org/hpc>  
                          <http://lqcd.jlab.org>  
                          <http://scicomp.jlab.org>