JLab Scientific Computing Status and Perspectives

CLAS Collaboration Meeting
March 7, 2018

Sandy Philpott

http://scicomp.jlab.org

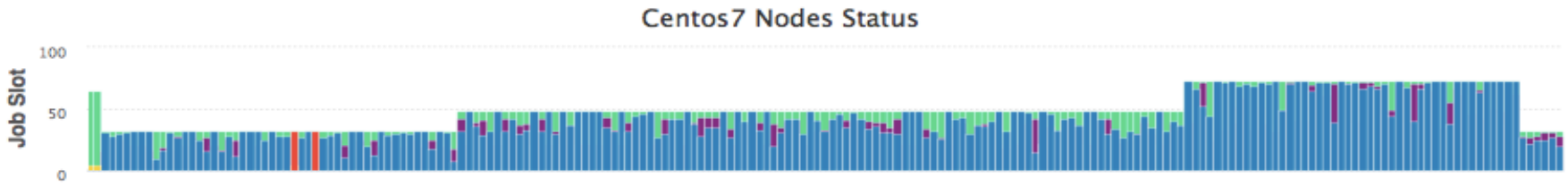
JLab Scientific Computing

- Resources
  - ENP Data Analysis Cluster (aka “The Farm”)
  - HPC Clusters (for reference, not part of this talk)
  - Mass Storage System
  - Human
- Challenges
- Outlook

The Farm – Compute Hardware

- 200 Batch Compute nodes, 5000 cores
  - 16, 24, or 36 core nodes – hyperthreaded
  - 32, 48, or 72 jobslots per node = # hyperthreaded cores
  - Using only physical cores yields ~85% performance of the node
  - 32 or 64 GB of memory,;128 GB virtual memory
- 2 Interactive User nodes: ifarm.jlab.org
- Data Gateway nodes
  - 6 DAQ,:10 GigE to DAQ, QDR IB to tape /stage disks
  - 2 Offsite: 10 GigE
  - 1 Desktop: 1 GigE
- 40 Gbps Infiniband network

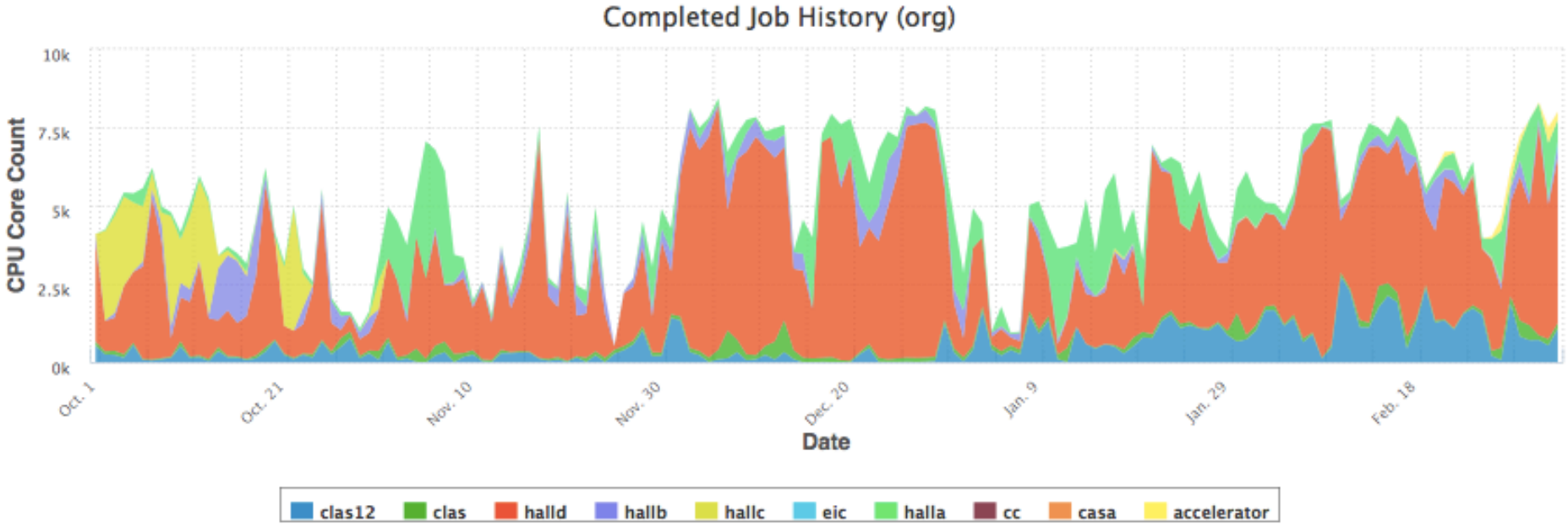
Farm Cluster Node Status



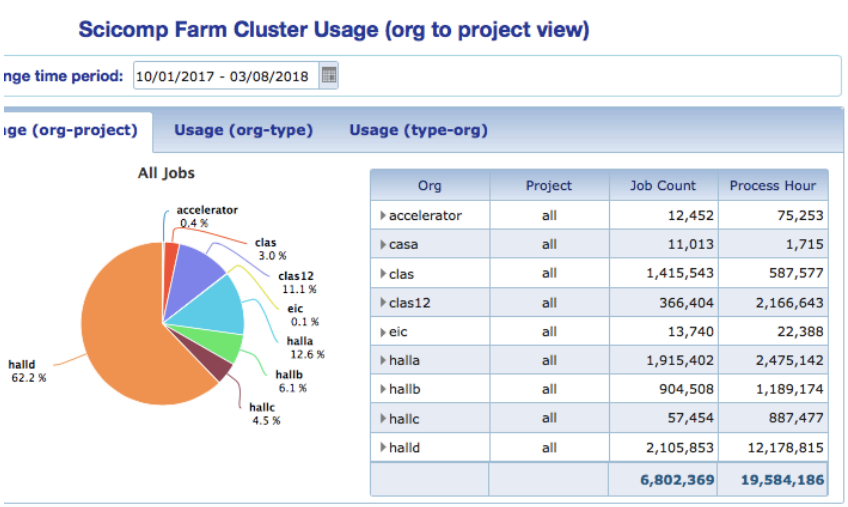
Scicomp Farm Cluster Job History

10/01/2017 - 03/06/2018

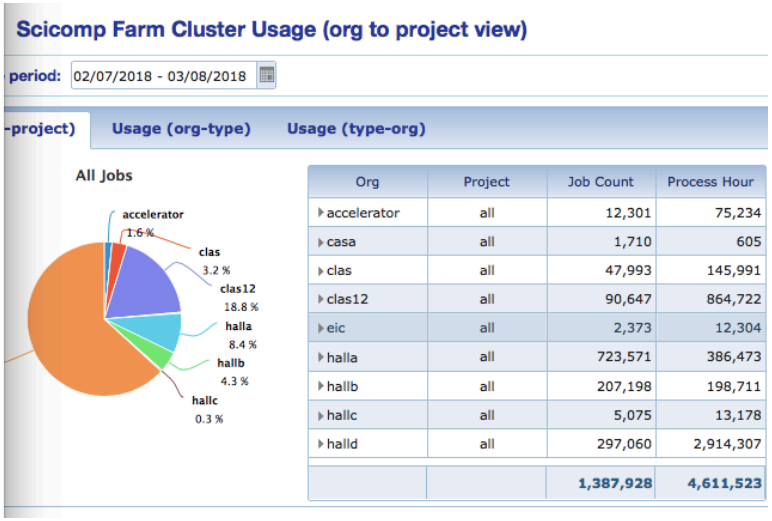
all



The Farm – Hall B



Org	Project	Job Count	Process Hour
accelerator	all	12,452	75,253
casa	all	11,013	1,715
clas	all	1,415,543	587,577
clas12	all	366,404	2,166,643
eic	all	13,740	22,388
halla	all	1,915,402	2,475,142
hallb	all	904,508	1,189,174
hallc	all	57,454	887,477
halld	all	2,105,853	12,178,815
		6,802,369	**19,584,186**



Org	Project	Job Count	Process Hour
accelerator	all	12,301	75,234
casa	all	1,710	605
clas	all	47,993	145,991
clas12	all	90,647	864,722
eic	all	2,373	12,304
halla	all	723,571	386,473
hallb	all	207,198	198,711
hallc	all	5,075	13,178
halld	all	297,060	2,914,307
		1,387,928	**4,611,523**

1. FY18, with Hall B using 20.1% = 3.0 + 11.1 + 6.1%
2. the last month, with Hall B climbing to 26.3% = 3.2 + 18.8 + 4.3%

Physics adjusted fairshare in early February, from

- A 10, B 20, C 10, D 60% to
- A/C 10, B 30, D 60%

The Farm – File Systems

Lustre, over ZFS

400 TB /cache: disk resident copy of tape data; auto-managed

200 TB /volatile: large scratch; auto-managed

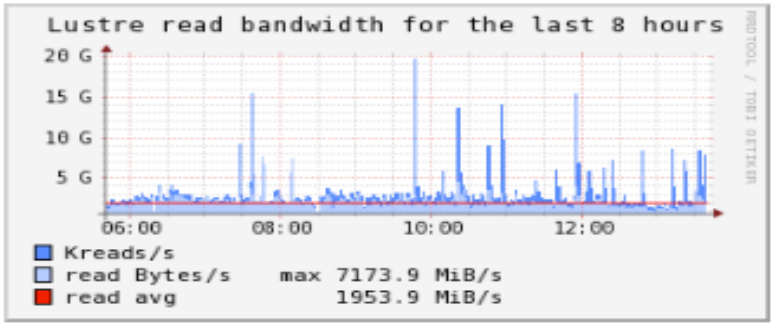
ZFS, via NFS

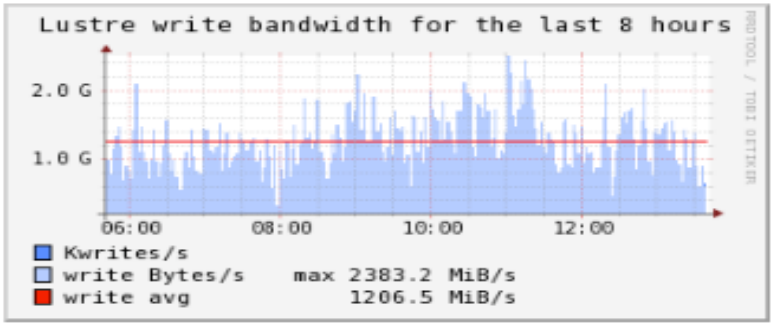
375 TB /work: small files, project managed, NOT BACKED UP

- 40 Gbps QDR Infiniband connectivity

/scratch: local disk, ranges from 0.5 to 1.5 TB

/mss: files on tape





The Farm - Software

CentOS 7; gcc, python, perl, java, …

Physics Software Committee; local site expert for each

ROOT, CERNLIB, GEANT4, CLHEP, EVIO, CCDB, GEMC

http://data.jlab.org

JLab SWIF workflow tool, for running jobs

- Optimize tape access
- Easily create/cancel/modify/retry jobs
- Specify inter-job dependencies
- Script-friendly
- Simplify batch system interface
- Possible interface for offsite resources, develop as needed

https://scicomp.jlab.org/docs/swif

MSS Hardware

IBM TS3500 Library: 25 PB, 24 LTO tape drives, 11 frames

- Bandwidth over 2GB/s

  8 LTO5, 140 MB/s 8 LTO6, 160 MB/s 4 LTO7, 300 MB/s 4 LTO8, 360MB/s
- 12,500 data cartridge slots, LTO4/5/6
- Room for expansion
  - 5 more frames to max of 16, ~1300 slots each
  - Awaiting LTO 8 integration and M8 (9 TB) cartridges
    - can add ~60 PB: 1300 slots/frame * 9 TB ea * 5 frames
  - Regular LTO8 media will hold 12 TB ea (LTO6 is 2.5TB ea)

- Duplicates of raw beam data stored offline, in the tape vault
- Constant migration of old data onto newer media

IBM MSS

IBM® System Storage™ TS3500 Tape Library

IBM.

System Summary

Logout Help

All Frames

Total storage slots	12537
LTO Licensed Capacity	12537
LTO Unlicensed Capacity	0
Total empty storage slots	1947
Offline storage slots	0
Accessors	1
Total I/O slots	16
Empty I/O slots	16
Total LTO data cartridges	10607
LTO Ultrium-2	0
LTO Ultrium-3	0
LTO Ultrium-4	1119
LTO Ultrium-5	5483
LTO Ultrium-6	3984
LTO Ultrium-7	20
LTO Ultrium-8	0
LTO Ultrium Not Labeled	1
Not Supported	0
Cleaning cartridges	2
Drives	24
Node cards	5

MSS Software

/mss: files on tape

Jlab JASMine software: interface between DAQ, MSS, Farm

- jcache – between disk and tape, /cache file pinning
- jmirror – DAQ to MSS, optional /cache file pinning

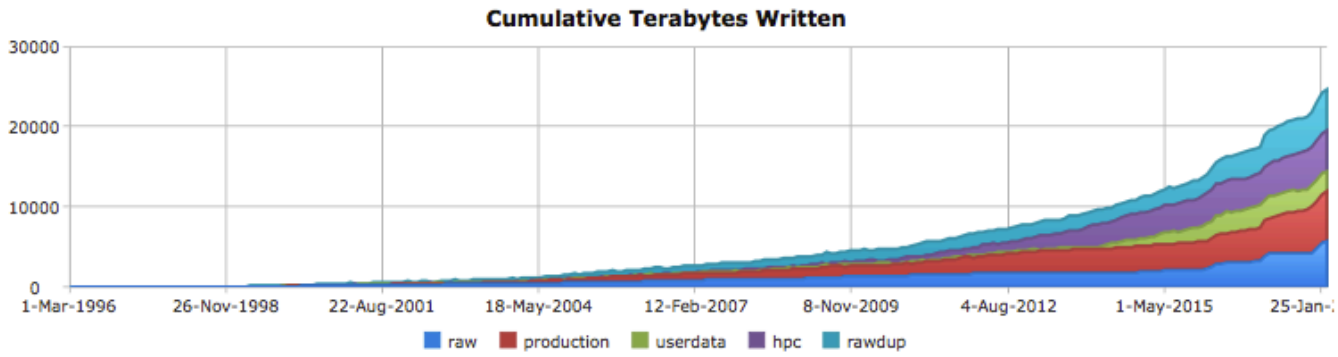
For files outside of DAQ or /cache:

- jput
- jget

MSS Usage, 25 Petabytes

Data Accumulation

Jan 01 1995 - Mar 07 2018



Organization × Category × Volume Set ×

Date	Category	Organiza...	TB	Files
raw, 0 TB, 25,967 Files				
duplicate, 5,024 TB, 2,051,006 Files				
production, 0 TB, 27,953 Files				
hallc, 731 TB, 1,633,144 Files				
hallb, 7,226 TB, 17,957,536 Files				
userdata, 0 TB, 644 Files				
home, 2,337 TB, 10,110,063 Files				
halla, 500 TB, 1,267,598 Files				
accelerator, 3 TB, 103,894 Files				
physics, 16 TB, 147,530 Files				
lattice, 5,066 TB, 17,739,259 Files				
halld, 4,579 TB, 12,614,080 Files				
test, 18 TB, 24,574 Files				
ccc, 1 TB, 6,172 Files				

Human Resources

- helpdesk@jlab.org
  - Trouble tickets, new project setups
  - Received by multiple staff
- Physics Computing Committee
  - ENP/IT regular meeting
  - Hall B Computing Coordinator: Harut Avakian
  - Physics Computing Coordiantor: Graham Heyes
- IT Steering Committee
- UGBOD Computing Contact – Or Hen, MIT
- IT OnCall
  - for after-hours emergencies
  - contact Guard to notify us

Challenges

- Defining computing and storage requirements
  - last formal ones are old - from November 2016 review
- Run regular Data Challenges
  - Verify resources at scale – nodes, DB, web, I/O, …
- Smoothing peaks and valleys, bursting to other resources
- I/O, using the right file system
  - large blocks (>=4MB) to Lustre /volatile, /cache
  - smaller, random to ZFS /work
  - Local /scratch (/scratch/clara)
- No/Old source code; few resources for older environments

Challenges

- Multi-threaded code, exclusive node use
  - reduces memory footprint
- Workflow planning
  - data on disk as needed, not duplicated, use when available
- Data preservation
  - IT: https://scicomp.jlab.org/DataManagementPlan.pdf
  - ENP: https://data.jlab.org/drupal/?q=dataManagement
- Fault tolerance / identify single points of failure risks
  - Documentation, multiple experts, duplicates / spares

Outlook

- HPC 12s cluster to ENP July 1: ~ 4000 cores
  - Run a full data challenge, ahead of fall beam
- FY18 farm node and storage procurement?
- Online/offline farm, for 10% event processing?
- Investigate and use Singularity containers for offsite resources
  - OSG (GlueX)
  - Amazon AWS
  - Microsoft Azure
  - Supercomputing Centers
    NERSC (GlueX)