

# JLab Status

April 2016 All Hands Meeting

*Chip Watson  
Jefferson Lab*

## *Outline*

Resources

Operations

File System Upgrades

# JLab Resources Summary

2 Infiniband clusters, 6.2k cores, *shrinking to 4.4k July 1*

3 GPU partitions, 344 GPUs, *shrinking to 180 July 1*

- 29 nodes, quad Fermi C2050, *will be retired July 1*
- 12 nodes, quad Kepler GTX690, *will be retired July 1*
- 45 nodes, quad K20 (augmented by trade in of 9q cluster)

Xeon Phi KNC test cluster, 12 nodes, 48 cards,

*these nodes will be converted to 12s as KNL becomes available nationally*

1.7 PB Lustre file system

- shared with Experimental Physics, 60% LQCD
- 26 servers
- 10+ GB/s aggregate bandwidth

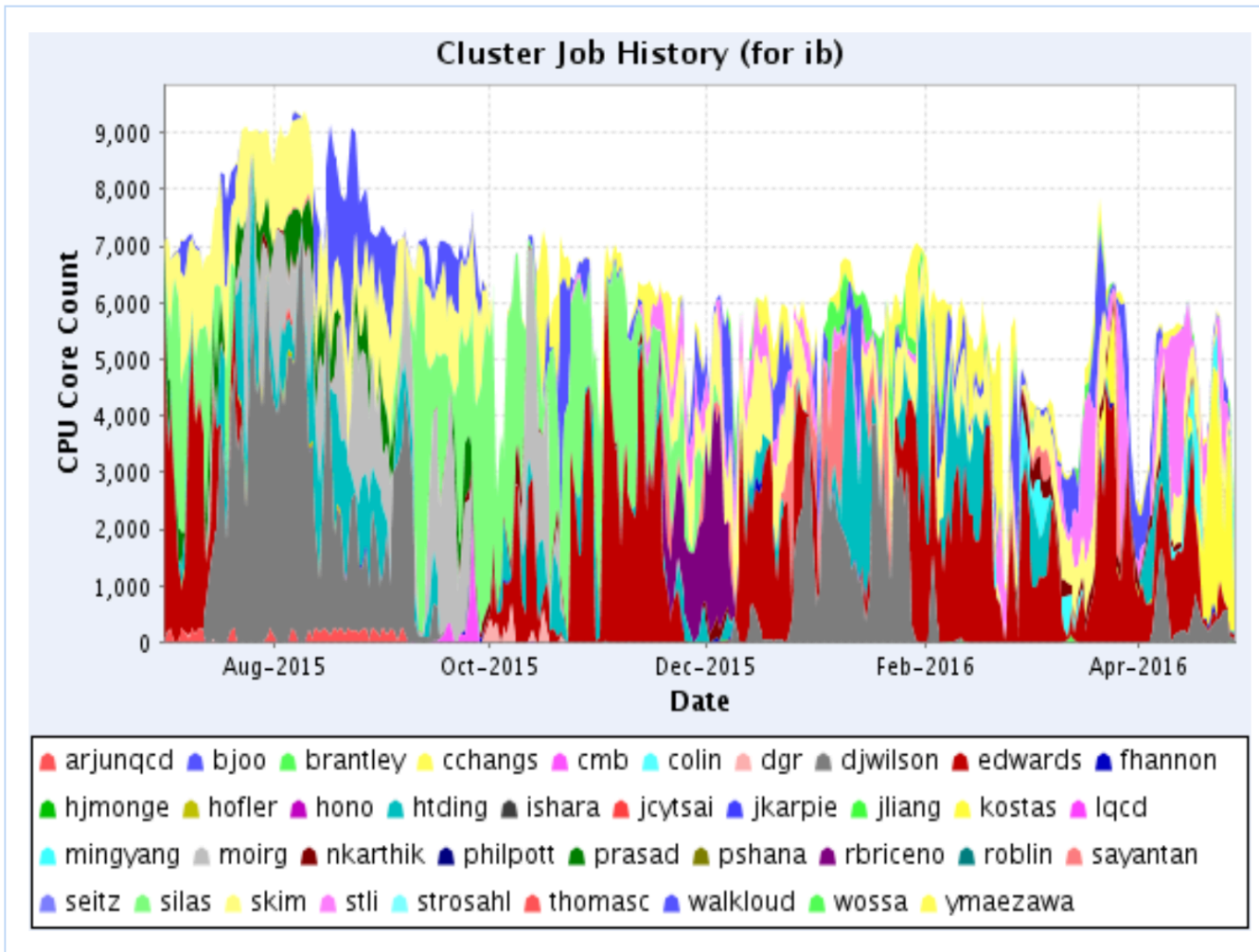
12 PB tape library, shared, ~ 3 PB LQCD

- LQCD growing at around 50 TB / month

# Operations & Utilization

## Cluster Usage Chart

From: Jul 1, 2015 To: Apr 28, 2016 Include Cluster:  ib  all [Get Chart](#)



LQCD running well  
Colors are different  
USQCD projects/users

Note that capacity varies  
day by day, with an  
overall downward trend  
as ARRA clusters are  
retired.

JLab load balances with  
Experimental Physics,  
which can consume nodes  
during our slow months.

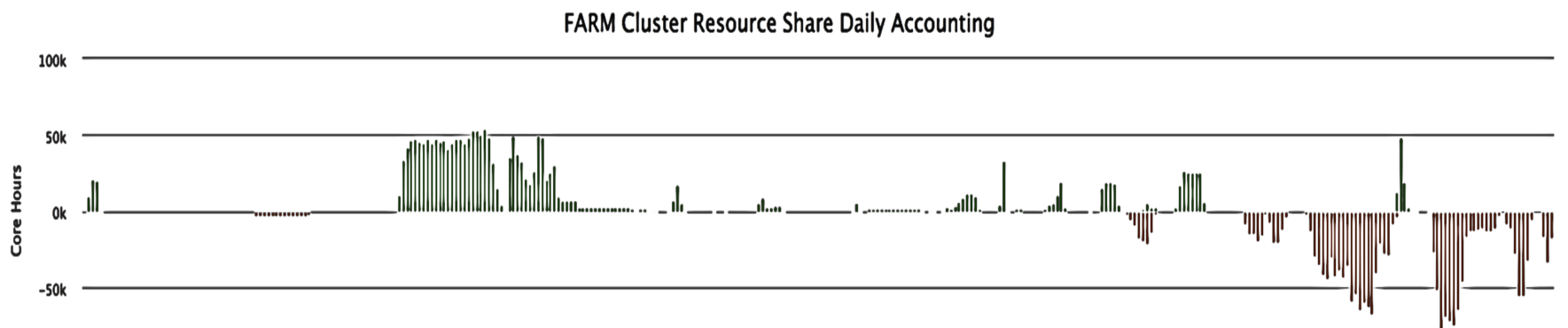
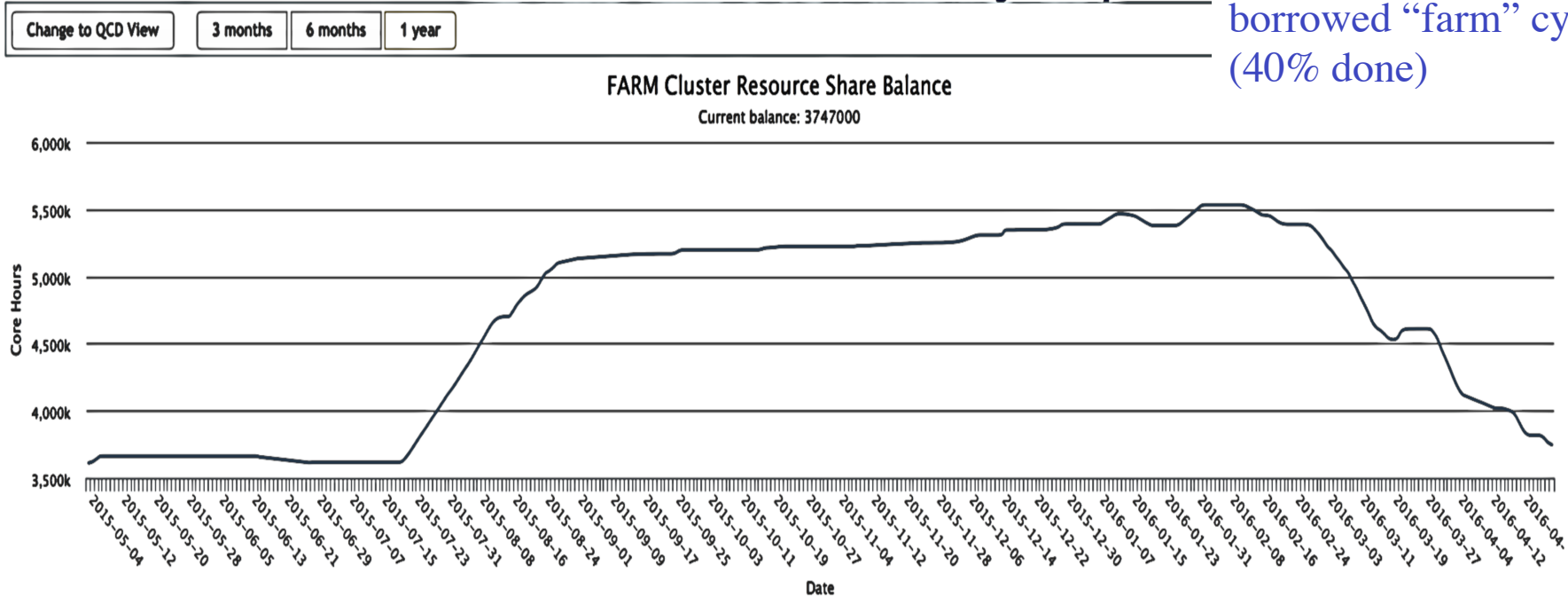
Note: In this chart, all cores are treated identically; differences in performance are ignored.



# Load Balancing

## Jlab Cluster Resource Share Accounting History

LQCD is now returning borrowed "farm" cycles (40% done)



# Allocations Progress (good)

## USQCD Project Allocation Usage (15-16)

*(Core hours for each cluster are converted to 15-16 hours based upon measured relative performance)*

Project Name	Allocation *	Proj. Used Hours	Annual Pace	Monthly Pace	Hour Remaining
Spectrum	34,660,000	40,457,769	141%	90%	0
thermo	32,420,000	33,527,486	125%	102%	0
NPLQCD	28,240,000	13,823,546	59%	64%	14,416,454
FFAB	6,480,000	0	0%	0%	6,480,000
c51	4,070,000	6,956,893	207%	170%	0
PESRadius	3,200,000	3,446	0%	1%	3,196,554
su2	1,000,000	1,313,669	159%	268%	0
chiQCD	1,000,000	810,674	98%	466%	189,326
t-charge	500,000	433,011	105%	3%	66,989
gflow	100,000	0	0%	0%	100,000
FormFactor	100,000	0	0%	0%	100,000
StochLapH	50,000	385	1%	0%	49,615
<b>Total</b>	<b>111,820,000</b>	<b>97,326,879</b>	<b>105%</b>	<b>86%</b>	<b>24,598,939</b>

IB / x86 delivery will slightly exceed 100% total at end of allocation year (after paying back experimental physics; note less than 100% monthly pace now)

*(GPU Hours are converted to allocation hours based upon measured relative performance)*

Project Name	Allocation *	Proj. GPU Hours	Annual Pace	Monthly Pace	Hour Remaining
NPLQCDg	2,589,000	2,083,374	98%	418%	505,626
Spectrumg	1,400,000	1,369,876	119%	173%	30,124
c51g	590,000	1,424,832	293%	186%	0
FormFactorg	10,000	0	0%	0%	10,000
chiQCDg	10,000	0	0%	0%	10,000
<b>Total</b>	<b>4,599,000</b>	<b>4,878,082</b>	<b>129%</b>	<b>312%</b>	<b>555,751</b>

GPU hours already exceeds plan for the year due to late turn-off of gaming cards.

# Lustre File System

1.7 PB across 26 servers shared with Experimental Physics

- Completed upgrade to version 2.5.3 last year
- This included switch to OpenZFS with RAID check on read
- LQCD added 256 TB new and retired 176 TB of 2009 servers

Will upgrade to Intel Lustre on June 21

- Lab funded, response to top request from survey of experimental physics users (we get it for free)
- Contains bug fixes not yet in the open source version
- Gains us support and technical help, for example in setting up dual-head active-active configurations
- Will require a 1.5 day outage

Modest capacity increase this summer

- adding another 256 TB LQCD, turning off 160

# Computer Room Upgrades (update)

To meet DOE goal of PUE of 1.4, power and cooling are being refurbished in 2015-16 (announced last year, slow implementation)

- New 800 KW UPS (done)
- 3 new 200 KW air handlers (+ refurbished 180)
- All file servers, interactive, etc. will move to dual fed power, one side of which will be generator backed (99.99% uptime)

## Transitions yet to come

- Rolling cluster outages to relocate and re-rack to 18-20 KW/rack as opposed to 10-12 KW today (starts in June / July)
- Anticipate 2 days outage per rack (3-4 racks at a time) plus 4 days full system outage over the next 8 months, so <2% for the year. JLab will augment x86 capacity by 3% to compensate



# New Web Pages Coming

- Web Application instead of pages
  - Faster navigation
  - Return to a page instantly (similar to using many tabs)
  - Maintain complex state more easily (user selected table filters and sort orders, time intervals, etc.)
- More complete view of system (e.g. tape system)
- Can be customized (filtered) to one user or one project
- Mobile friendly
  - Resizes and re-stacks vertically for narrow screens



Computing Cluster

- Jobs
- Nodes
- Usage
- By User
- By Project
- Project Info

File System

- Cache
- Volatile
- Work

Tape Library

- Jobs
- Usage

System Status

- Lustre
- Ganglia
- Environment
- Node Share
- GPU Info

Documentation

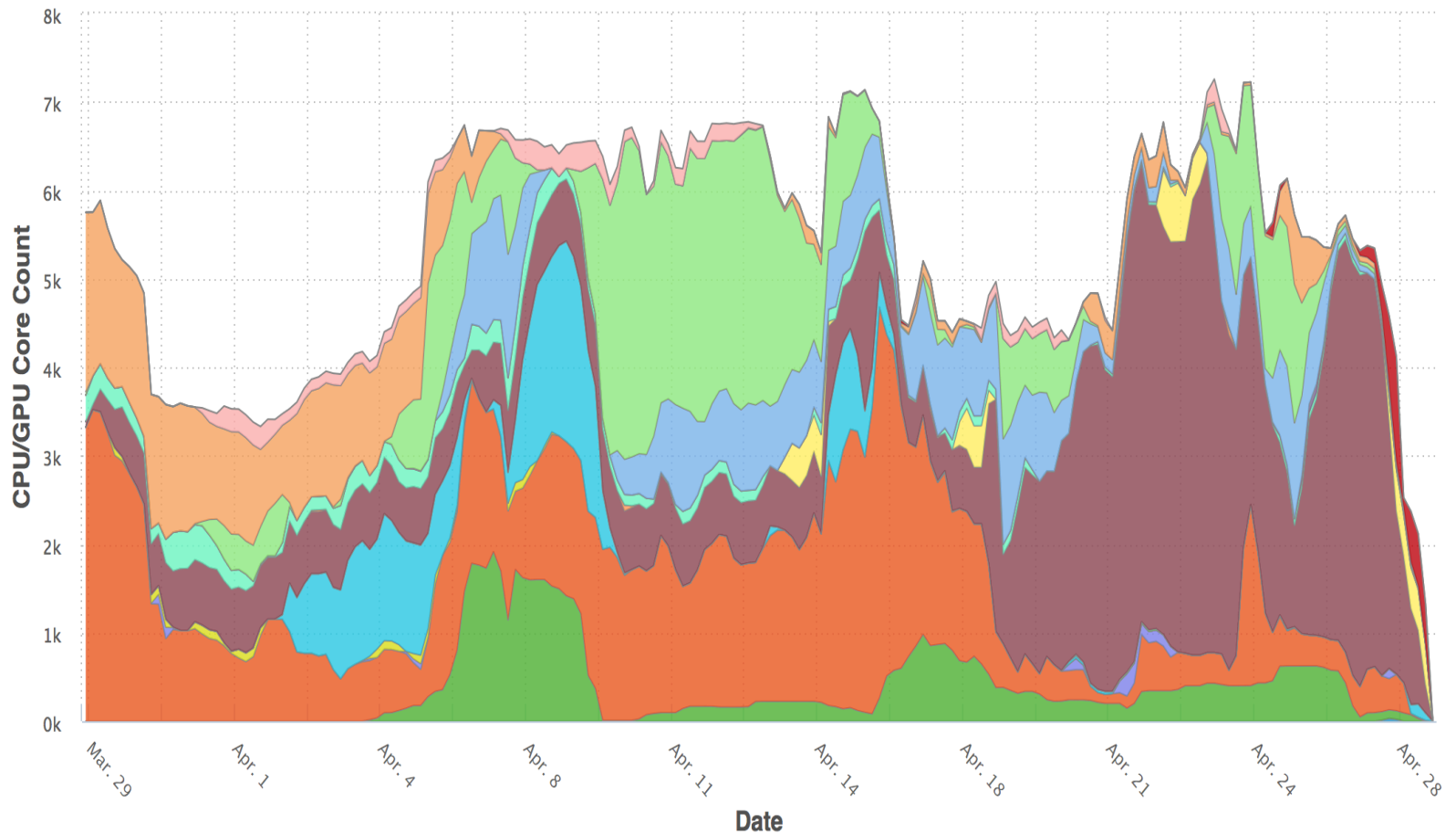
- User Guide
- Resources
- News Archive

Software

### Jlab Cluster Usage Chart

From 03/29/2016 To 04/28/2016 Includes: all nodes Get Chart

Completed Job History (all nodes)



# Documentation

(slowly improving; never perfect)

Drupal “book” based

- Chapter, section, subsection; next/up/back navigation
- Searchable

We will be updating the style this year, and adding more naming of pages so links are more obvious.

Three Books:

1. Scientific Computing Resources
  - Covers both shared and mission specific
2. LQCD User Guide
3. Experimental Physics User Guide

## Data Management FAQ

### What is the best/fastest way to move files from /volatile to /cache?

/volatile and /cache are actually remounts of /lustre/volatile and /lustre/cache. If you use the `mv` command to move files between the two and if you fully qualify the files, i.e., use `mv /lustre/volatile/PROJECT/src/files /lustre/cache/PROJECT/dest/files`, then the move will be instantaneous. Doing `mv /volatile/PROJECT/src/files /cache/PROJECT/dest/files` will require a copy of all of the source files vs just a rename.

### When will a file under /cache be backed up?

By default, files in /cache and 12 days old will be automatically backed up. Files with size larger than 1MB will be migrated to the tape library. A user can use `srmPut` to flush files to tape at anytime.

### Why is it a good practice to tar small files?

Although /cache is designed to store large data files, a user can create a file with any size. However, it is recommend to pack small files into a large file if there are more than 100 files in a group. To put/get many many small files to/from the tape system might take many days, whereas a single tar file could be written or retrieved in a small fraction of that time.

### Can a file/directory be renamed or moved after it is backed up?

No, a user shouldn't rename or move a file after it is put onto tape. There is a one to one mapping between the file path in cache and that in the tape library. Once a file is put into the tape library, the content, location and name should not be altered. If a file is moved/renamed, it will be treated as a new file and will be backed up again. This will not only waste tape, it will also create confusion for future users who may want to use the data. Please send a CCPR if you want to rename a cache directory after the files is backed up to tape system.

## LQCD User's Guide

- ▶ [Getting Started](#)
- ▶ [Batch System](#)
- ▶ [MPI](#)
- ▶ [Using GPUs](#)
- ▶ [Using MIC cards](#)
- ▼ [File Systems](#)
  - [/work from within a batch job](#)
  - [Cache manager policy](#)
  - ▶ [Cache manager utilities](#)
  - [Data Management FAQ](#)
  - [Volatile management](#)
  - [Typical usage pattern](#)
- [FAQ](#)