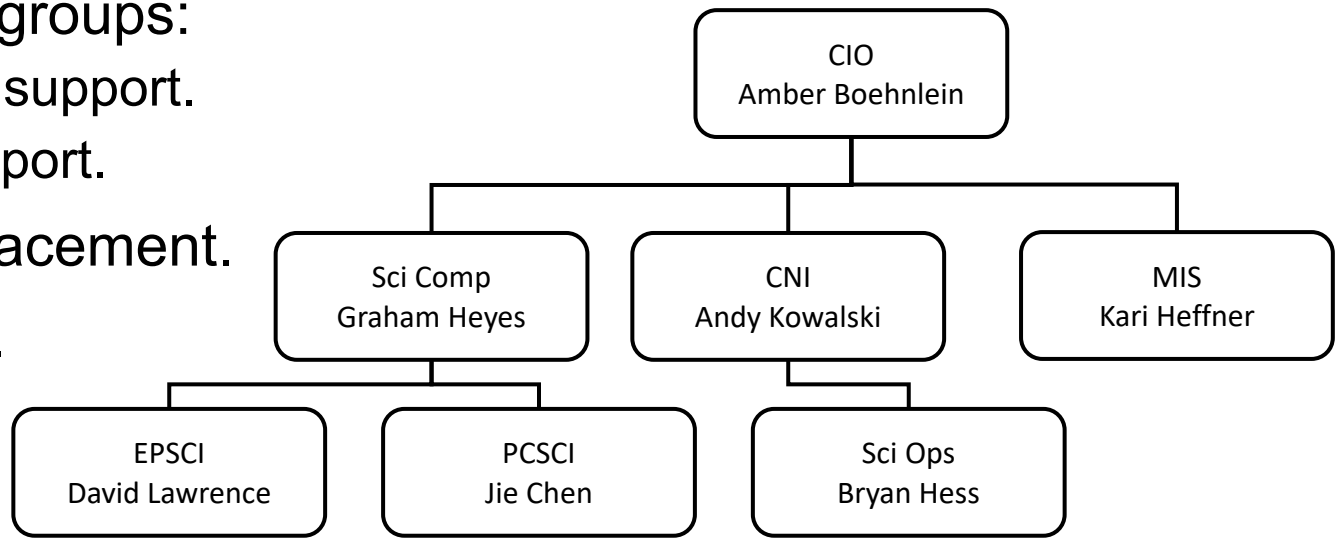
Jefferson Lab LQCD Computing April 2020 All Hands Meeting

Graham Heyes – Scientific Computing Department Head, IT Division, JLab

Under new management and growing

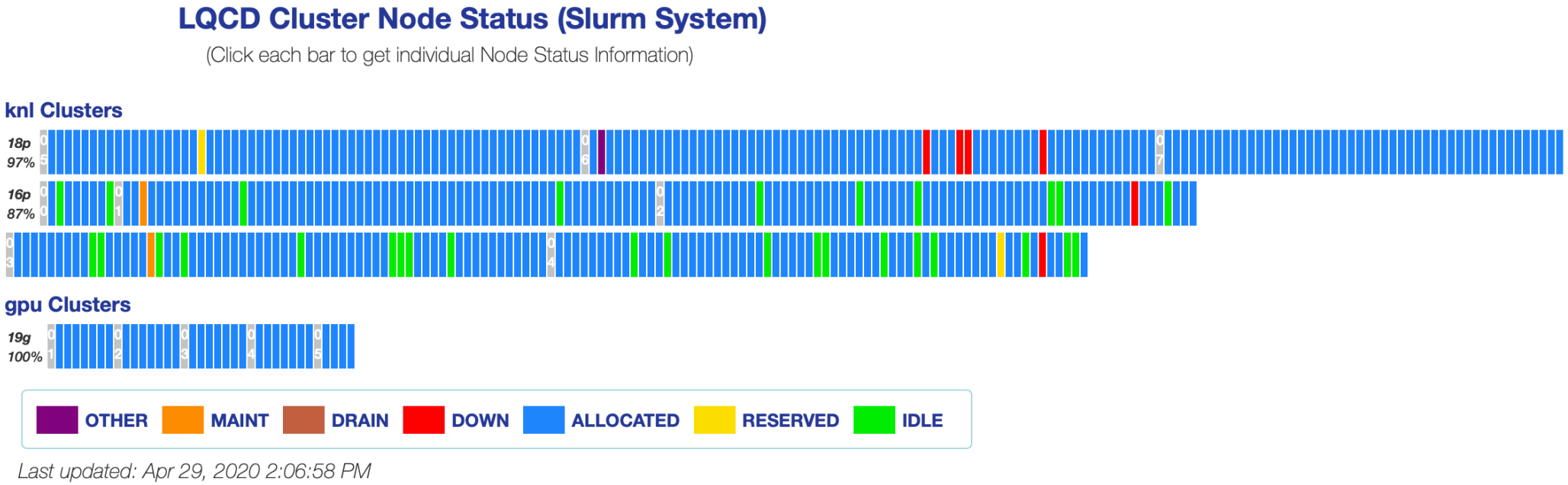
- After a long and illustrious career Chip Watson retired in February 2020.
- I started March 1st and have been working from home since March 15th.
- To better align with networking and the other system admin groups SciOps has remained in CNI.
- Scientific Computing has currently two groups:
  - PCSCI, HPC, LQCD and batch cluster support.
  - EPSCI, software development and support.
- Balint Joo is leaving and will hire a replacement.
- Other hires planned in FY20 and FY21.

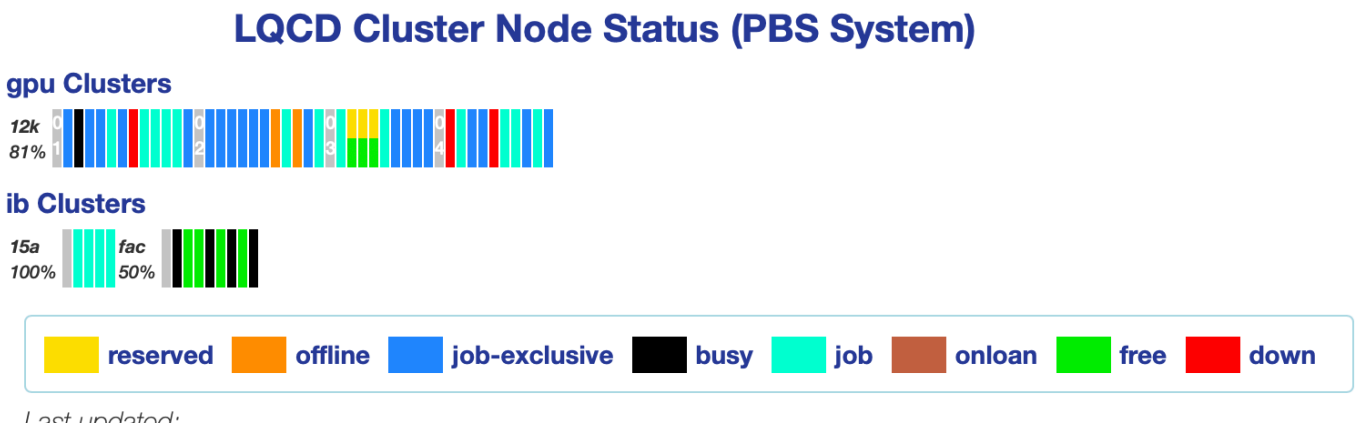


Current resources - clusters

- Jlab currently has two flavors of cluster, with and without GPU.
- 440 node Xeon Phi / KNL cluster (“16p/18p”)
  - Single socket 64 core KNL (with AVX-512 8 double / 16 single precision)
  - 192 (98) GB main memory / node 16p (18p)
  - 32 GB high bandwidth on package memory (6x higher bandwidth)
  - 100 Gbps bi-directional Omnipath network fabric (total 25 GB/s/node)
    - 32 nodes / switch, 16 up-links to core / switch
  - total: 3.168 M node-hours = 202.75 M KNL-core-hours
- 32-node GeForce GPU cluster (“19g”)
  - Eight-GPU RTX-2080 nodes
  - 8 GByte memory per GPU, 192 GByte memory per node.
  - Each on 100g OmniPath Fabric
  - Total: 230.4 K node-hours = 1.84 M RTX2080-GPU-hours
- There are also the remnants of the 12 K cluster that will be retired this summer to make space for 20g.

Current resources - clusters





Jefferson Lab

Current status - clusters

- LQCD Cluster Operations and Status
  - The 12k cluster (2012 Kepler GPUs) is down to only 30 nodes and in its final months. It will be retired in July to make space for new machines in the hot aisle containment system.
    - There are no plans for further allocations for the K20 cluster.
    - The Jlab FY21 facility improvement plan adds a second containment system removing this limit.
  - 19g update on issues reported last year:
    - Has no ecc memory, we have automated gpu memory tests on every RTX2080 card.
    - Since the 07/01/2019, there have been only two bad GPUs out of 256 due to memory issues.
    - Cluster has had excellent uptime – see later slide.
    - Usage exceeded allocation for the year .
  - KNL clusters (16p, 18p):
    - Both clusters are running ahead of annual pace and operating well.

- COVID-19 posture
  - JLab is evacuated except for essential staff.
  - We practice system administration from home.
  - Staff go to the lab to fix hardware problems or install new systems.

- 20g New Cluster Procurement for LQCD:
  - Currently evaluating architectures – delivery of test hardware delayed due to pandemic.
  - Performance based evaluation: Dslash, hadron contractions (batched zgemm).

Current status - disk

- Lustre - Storage for /volatile and /cache
  - For operational and budgetary reasons Lustre is no longer shared with exp. physics.
  - New hardware on order or ready to go :
    - A dedicated all-flash metadata failover server pair has been purchased and should arrive in May.
    - 1 PB of raw disk with better performance has been purchased and is racked up awaiting the arrival of the metadata servers. Lustre servers are in failover pairs for increased reliability.
    - This new hardware is the seed of a new, upgraded Lustre system for LQCD, which will run the latest long-term release of Lustre
  - The oldest Lustre disk shelves are being retired since they have reached their end of life.
  - **The end state will be 1.3PB (usable) for LQCD**
  - New EDR Infiniband switches will be installed for the Lustre core
    - Increases bandwidth between storage and compute.
    - Installation slated for this summer. We do not anticipate a significant outage as groups will be migrated one at a time
- NFS file server - /work and /home on ZFS. No changes this year.
  - /work is not backed up. Quotas are assigned per project.
  - /home is backed up weekly.

Current status - tape

- Tape Storage
  - All data has been migrated to LTO M8 media.
  - New writes are to M8.
  - We anticipate switching writes to LTO8 native media in the next media purchase.
    - This will increase write bandwidth.
  - **LQCD accumulated storage is 1.9PB on tape**
    - 1.3PB on lattice-p "permanent"
    - 0.6PB on lattice-t "temporary"
    - Tape storage for lattice-t USQCD (non-JLab) allocations are retained at Jefferson lab for 18 months after the allocation year ends, then the tapes are re-used.
  - JLab operates two tape libraries, an IBM TS3500 and an IBM TS4500.
    - Soon to be consolidated into a single TS4500.
    - Will add more LTO8 drives in FY20 for 25% bandwidth increase to tape.

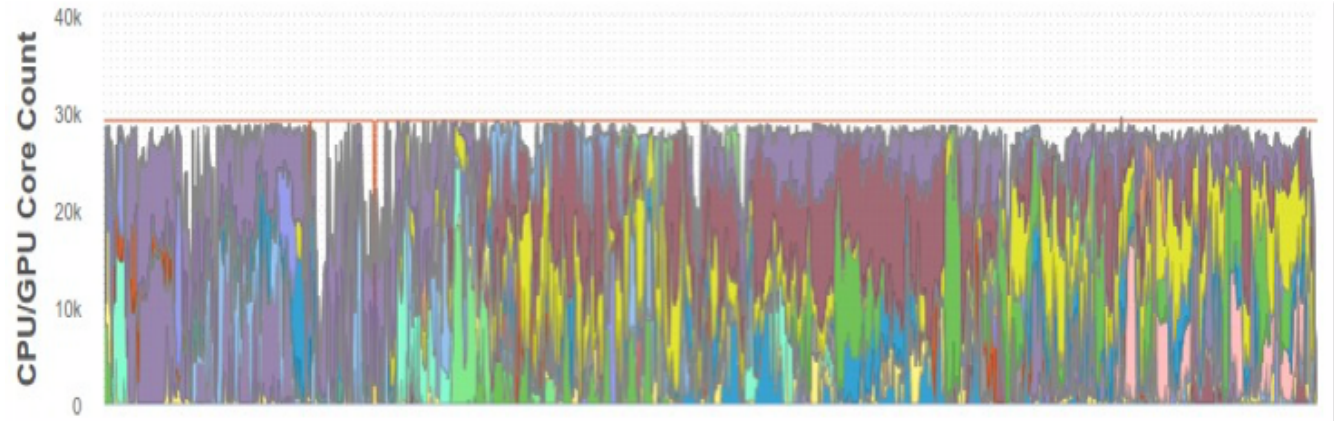
Current status – Batch System

- Batch system
  - After the 12K cluster retirement in July PBS will be removed from service for LQCD.
  - The HPC Slurm server, running 16p, 18p, 19g, and the upcoming 20g cluster, will be migrated to new, more reliable, hardware.
  - This summer, after migration, Slurm will be upgraded to a more recent version.
  - The ENP clusters started using Slurm last year.
    - More complicated mix of workflows and job types.
    - ENP side has features that the LQCD side does not:
      - Scavenger queue – very low priority preemptable jobs.
      - Priority queue – jobs that jump to the front of the line.
      - Ability to request an interactive node from a cluster.
      - Some of the features above are of limited interest since LQCD servers are highly loaded.
      - Workflow manager – users can specify managed workflows to resubmit failed jobs.
  - LQCD job scheduling policy remains Fair Share with Backfill.

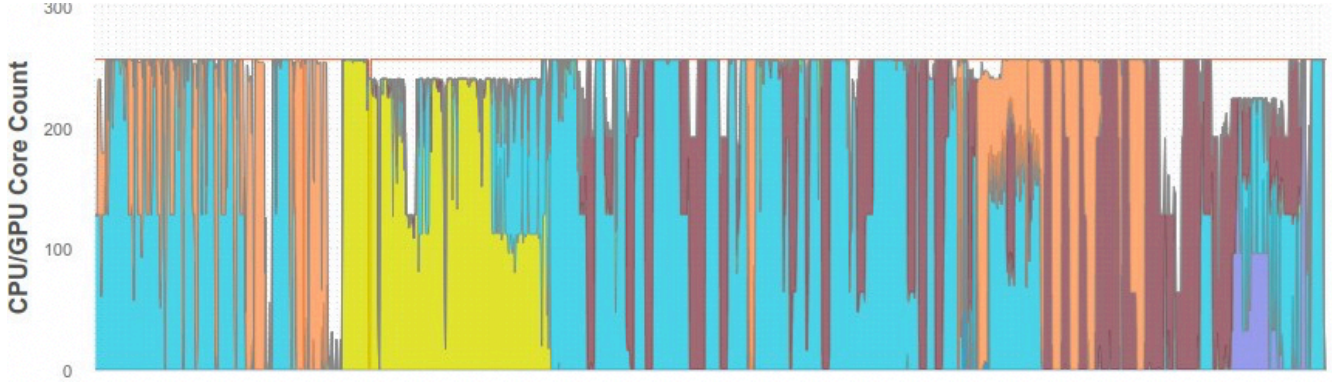
Cluster operation statistics

- Our status is on the web for anyone to see at lqcd.jlab.org
- Plots to the right show load on clusters from July 1st, 2019 to present.
- Orange line shows maximum.
- The recent dip in 19g was due to temporarily removing several nodes from operation to develop code for running COVID-19 simulations.
  - very little unused time.

Jlab LQCD User Jobs (07/01/2019-04/22/2020)



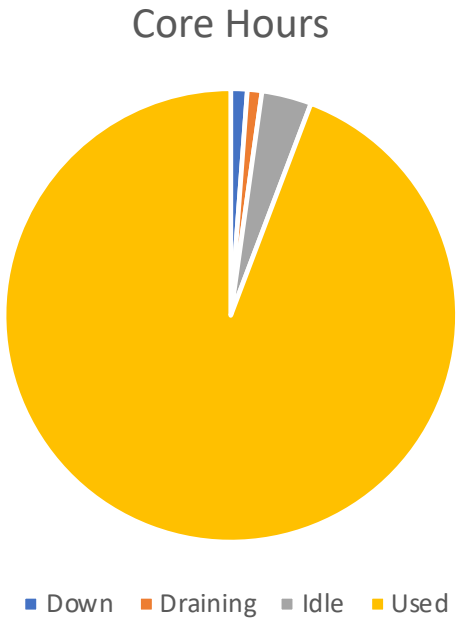
16P and 18P Clusters

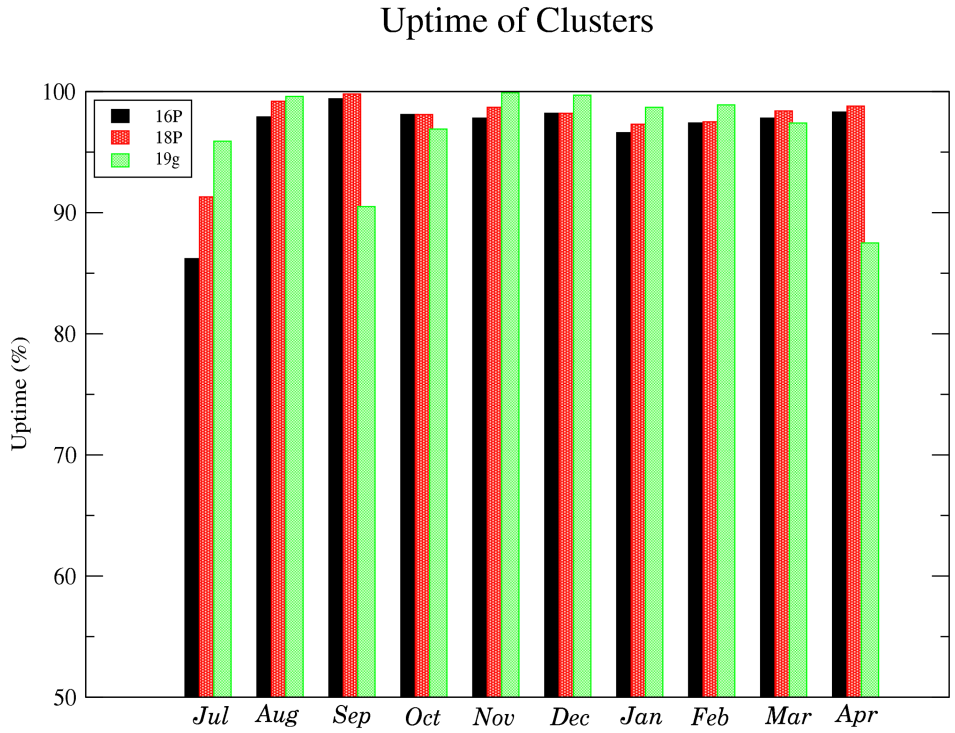


19g Cluster

Cluster operation statistics

- The pie chart shows total used core hours, idle hours, down hours and draining hours for 16p and 18p. The utilization is about 94% and draining hours is around 1% of total core hours.
- Awards are given to projects of size > 10M hours:
  - 5 projects out 12 have rewards.
  - Total of rewards is about 9% of allocations
- Utilization is 103% of pace, currently monthly pace is 111%.
  - One project has overrun even including their rewards.
- We will meet the target to satisfy all this year's allocations.





Jefferson Lab

User statistics

- The table, from the website shows the allocation, used and pace for the various projects.

USQCD Project Allocation Usage (19-20)

(K hours for each cluster are converted to '19-20' weight unit hours based upon measured relative performance)

Project Name	Allocation	Project Used	Annual Pace	Monthly Pace	Adjustments*	Remaining	Overused
thermop	24,000	25,118	126%	213%	1,960	842	0
K2pipiPBC	22,000	11,521	63%	37%	-1,300	9,179	0
NME	21,000	17,121	98%	59%	60	3,939	0
decay0n2b	19,000	12,219	77%	76%	-400	6,381	0
cedmp	17,000	18,060	128%	64%	3,360	2,300	0
Delta	17,000	6,143	43%	125%	-5,660	5,197	0
Spectrump	15,000	36,344	292%	128%	10,860	0	10,484
NPLQCDp	15,000	11,559	93%	196%	-3,280	161	0
cedmpffp	12,000	7,258	73%	41%	-30	4,712	0
lightdist	10,400	2,830	32%	0%	-1,500	6,070	0
LSD4p6	10,000	10,176	122%	118%	1,260	1,084	0
chiQCDp	10,000	5,064	61%	314%	-3,630	1,306	0
flowRG	9,000	9,334	125%	105%	1,390	1,056	0
	201,400	**172,747**	**104%**	**112%**	**3,090**	**42,227**	**10,484**

Project Name	Allocation	Proj. GPU Used**	Annual Pace	Monthly Pace	Adjustments*	Remaining	Overused
Spectrumg	1,100	1,151	126%	127%	155	104	0
Deltag	440	233	64%	0%	-50	157	0
NPLQCDg	300	398	160%	217%	-29	0	127
chiQCDg	197	15	9%	33%	-82	100	0
pqpdfg	105	47	54%	63%	-11	47	0
	2,142	**1,845**	**104%**	**102%**	**-17**	**408**	**127**

Summary

- Despite the pandemic the clusters at Jlab are running and being maintained. We feel that we have a sustainable operations workflow in place for the coming months.
  - We will be able to handle the install and bring up 20g when it is eventually procured.
- The website lqcd.jlab.org allows system status to be monitored.
- Statistics show the systems are utilized at the 94% level, uptime 99%

- After the reorg. and staffing changes our aim is to continue to support your work at the highest possible level.

Please feel free to provide feedback and suggestions

One more thing

Many thanks to Chip Watson!

- Trained as an Expt. Nuclear Physicist - worked at BNL before JLab
- Retired in February after a more than 30-year career at Jlab
- Joined as data acquisition lead in physics division.
- Headed development of CEBAF's EPICS accelerator control system.
- Started the LQCD group in 1999 at the same time that Robert Edwards & David Richards arrived.
- Chip was the principal lead in proposal for original SciDAC project (~2000)
  - Helped shape original USQCD software stack
- Keenly aware of computing trends & harnessing for computations
  - Early proponent/adopter of new systems, including GPUs (ARRA 2009) + KNLs
- Became head of scientific computing in IT division at Jlab.
- Rare talent at combining technical skills + science advocacy in US communities.

- Wish him well on his next adventures (not really retiring of course…)