Lattice QCD Site Report

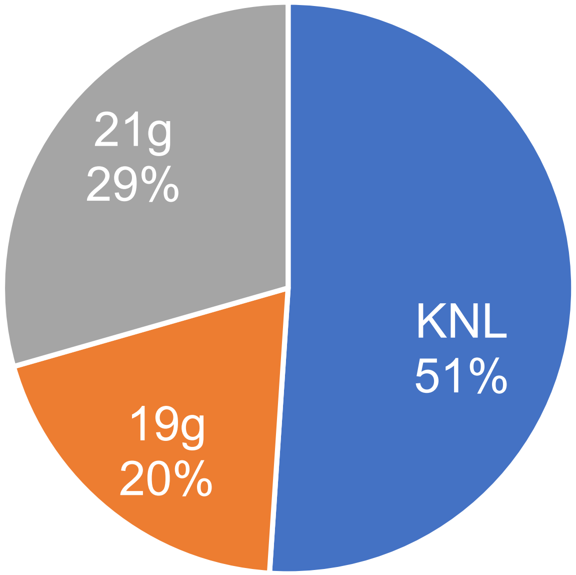
Jefferson Lab

Amitoj Singh

Thursday, April 18, 2024

Current resources as pledged for 2023-24 USQCD allocations

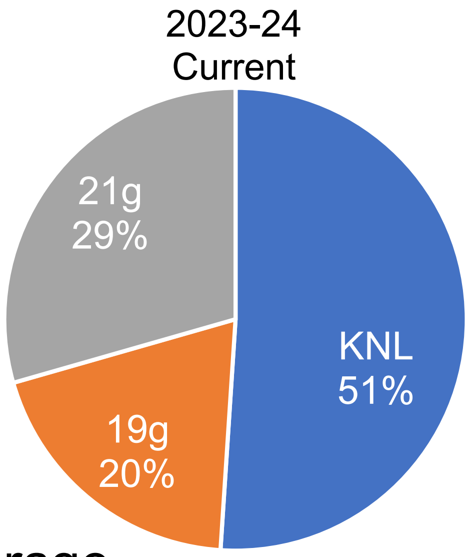
- Compute
  - 400 node Xeon Phi / KNL cluster (“16p/18p”)
    - Single socket 64 core KNL (with AVX-512 8 double / 16 single precision) 192 (98) GB main memory / node 16p (18p)
    - 32GB high bandwidth on package memory (6x higher bandwidth)
    - 100 Gbps bi-directional Omni Path network fabric (total 25GB/s/node) 32 nodes / switch, 16 up-links to core / switch
    - 93.3M Sky-core-hours
  - 32-node GeForce GPU cluster (“19g”)
    - Eight-GPU RTX-2080 nodes
    - 8GB memory per GPU, 192GB memory per node. Each on 100g Omni Path
    - 35.7M Sky-core-hours
  - 8-node AMD GPU Cluster (“21g”)
    - Eight-CPU AMD MI100 nodes with Inter-GPU Infinity interconnect. 32GB memory per GPU, 1TB memory per node.
    - Each on 100g InfiniBand Fabric
    - 53.8M Sky-core-hours

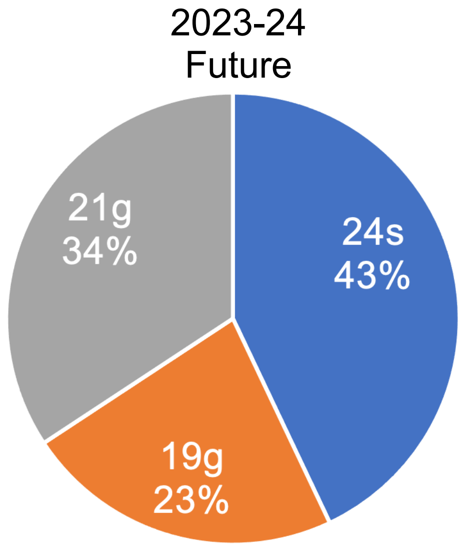


- Storage
  - 1.8PB total of shared disk space and 1.0PB of tape storage

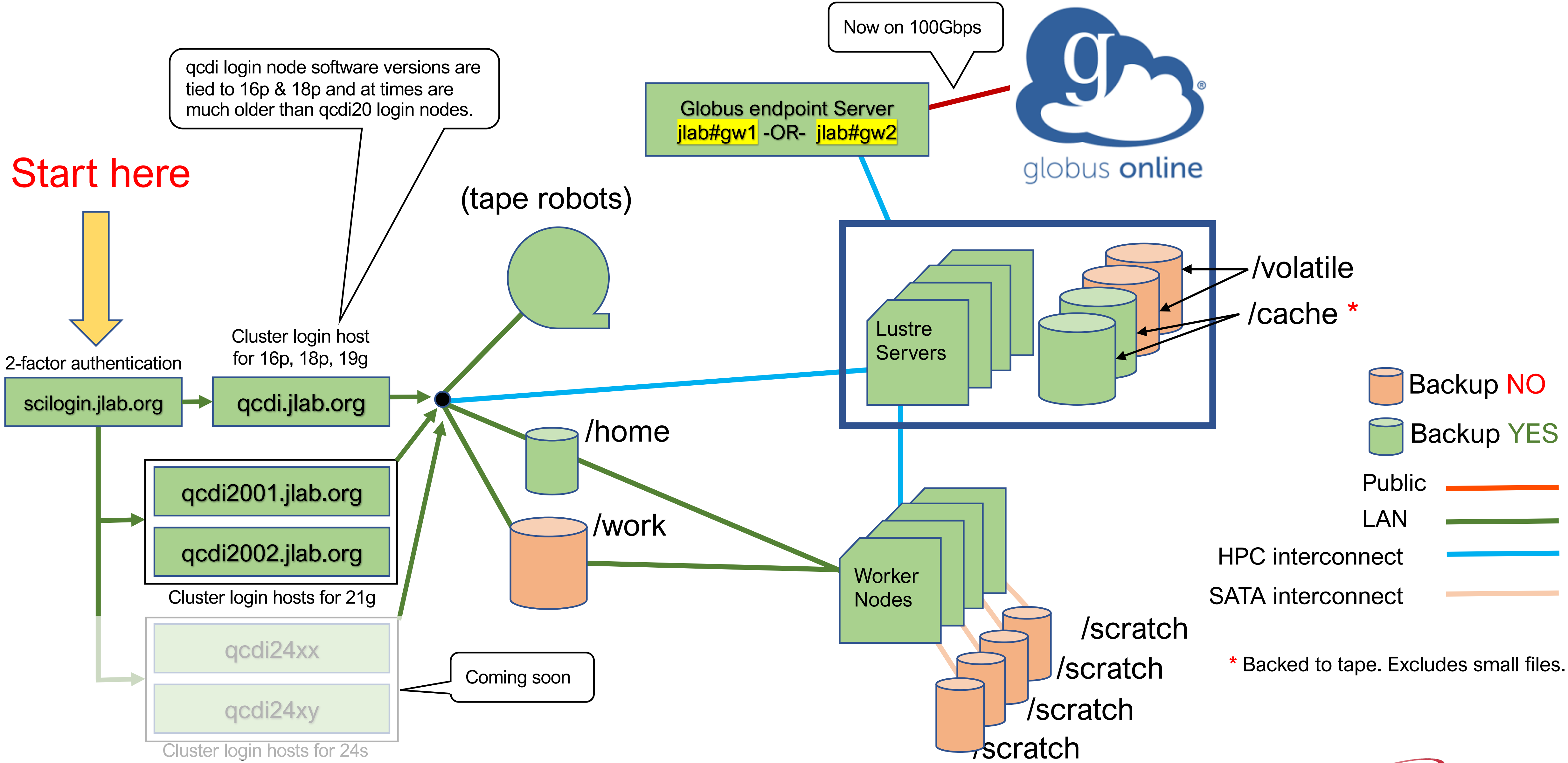
Future resources as pledged for 2024-25 USQCD allocations

- Compute
  - 32-node GeForce GPU cluster (“19g”)
    - Eight-GPU RTX-2080 nodes
    - 8GB memory per GPU, 192GB memory per node. Each on 100g Omni Path
    - 35.7M Sky-core-hours
  - 8-node AMD GPU Cluster (“21g”)
    - Eight-CPU AMD MI100 nodes with Inter-GPU Infinity interconnect. 32GB memory per GPU, 1TB memory per node.
    - Each on 100g InfiniBand Fabric
    - 53.8M Sky-core-hours
  - 100-node CPU Cluster (“24s”)
    - Thirty-two-core, dual-socket, 2.8 GHz Intel Xeon 8462Y+ (Sapphire Rapids) nodes
    - 64 cores per node
    - 1 TB memory/node
    - Each on NDR200 Infiniband Fabric
    - Total: 67.43 M Sky-core-hours
- Storage
  - 1.8PB total of shared disk space and 1.0PB of tape storage



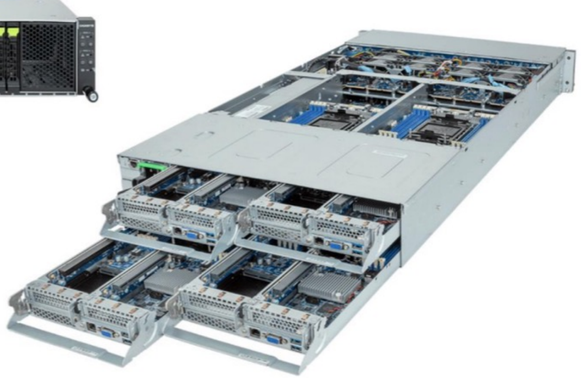


JLab Cluster Layout Summary



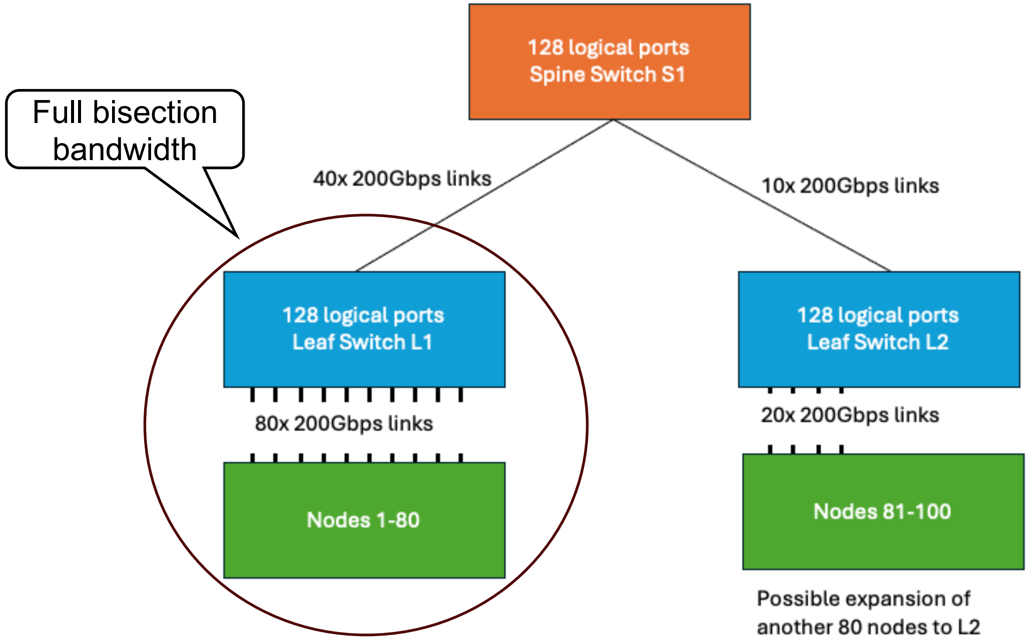
24s

- 100 nodes @ 4 servers/2U chassis
- 2 racks @ 50KWatts/rack (100KW total)
- Machine sound level is a concern: 90dB at idle and 110dB under load, requiring hearing protection
- 32-core dual-socket, 2.8GHz Intel Xeon 8462Y+ “Sapphire Rapids”
- 1TB 4400 MT/s Memory per node
- 64,000 cores total
- 3 InfiniBand NVIDIA Quantum-2 QM9700 32-port NDR switches
- 78 TFlops Average(DWF+Clover)

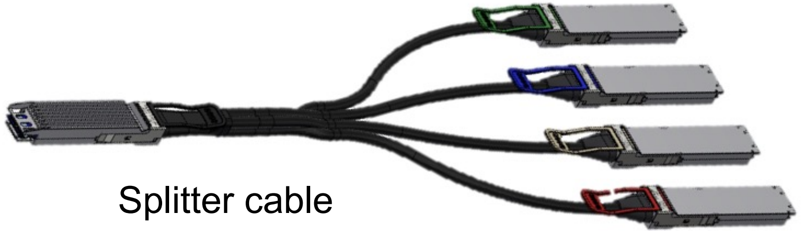


24s Infiniband NDR200 network

- What is NDR200? A version of NDR (capable of up to 800Gbps) running at 200Gbps using splitter cables.
- 24s has NVIDIA Quantum2 QM9700 NDR switches with 2:1 oversubscription in a leaf and spine configuration.
- The QM9700 switch carries an aggregate bidirectional throughput of 51.2Tb/s, with more than 66.5 billion packets per second (BPPS) capacity. Incorporates technologies such as RDMA, adaptive routing, and NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP).
- A single port of the leaf switch is connected to four single-port NDR200 Infiniband ConnectX-7 Host Channel Adapters (HCAs) using 800Gb/s to 4x 200Gb/s (NDR200) passive copper splitter cables. This high-density switching solution allows 80 nodes to share a single leaf switch. Each leaf switch furthermore has 40 200Gb/s uplinks to the spine switch.
- ConnectX-7 results in a substantial boost in the message passing rate, from 215 million messages per second for CX6 to an impressive 330-370 million messages per second.

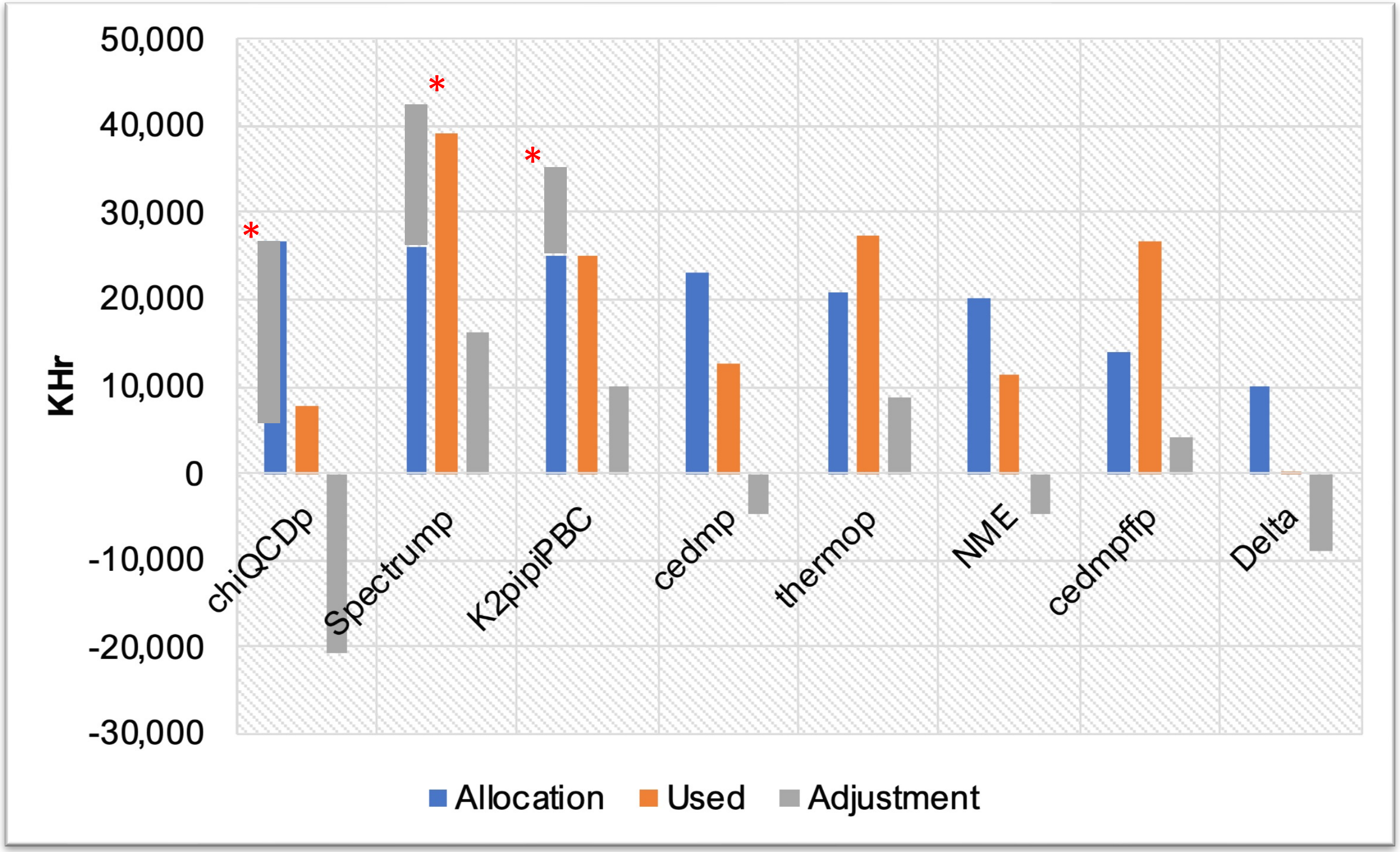


QM9700 Front View



Splitter cable

2023-2024 Allocations Summary – KNL Clusters

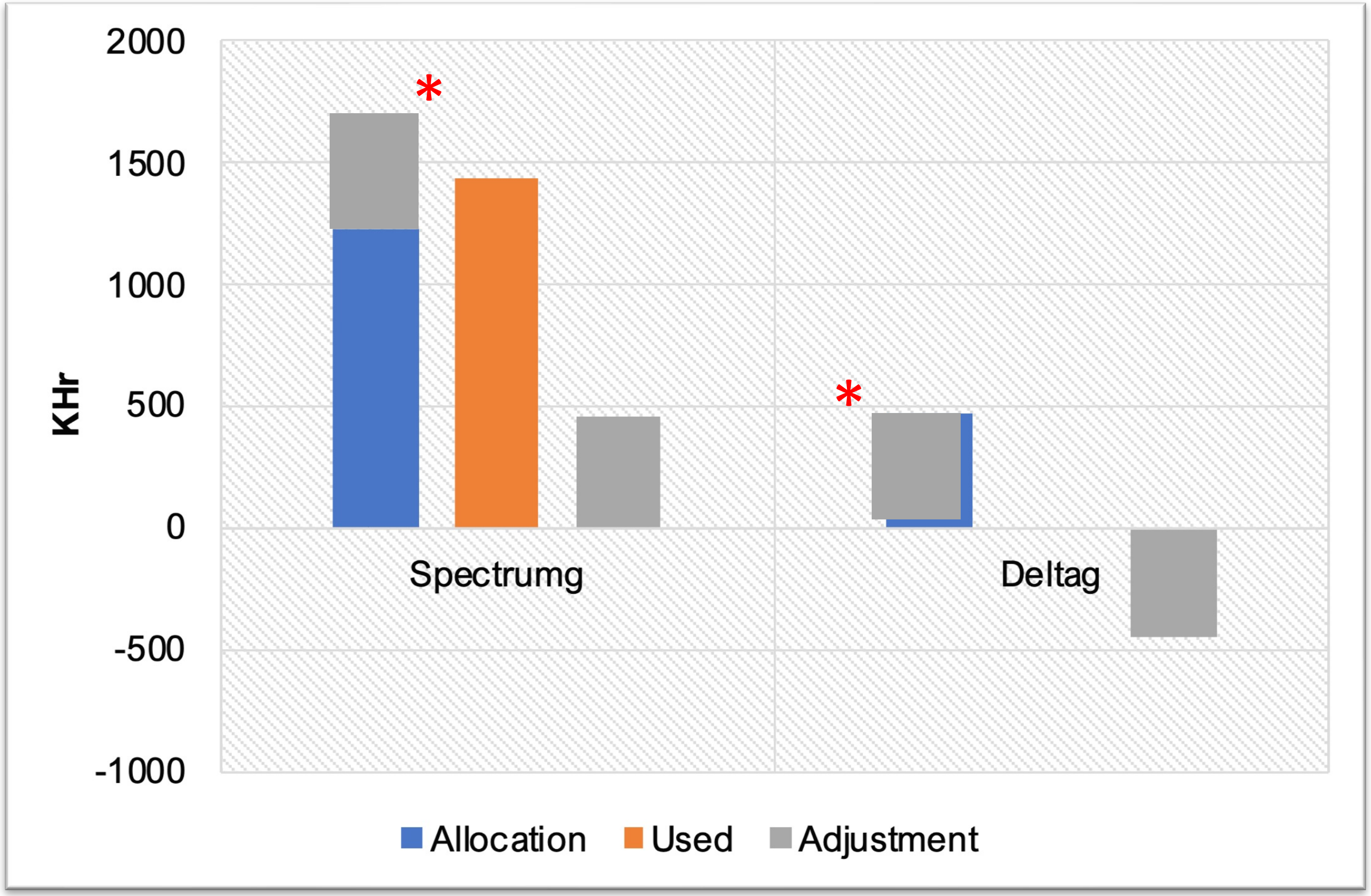


Total allocations used thus far with 80% of allocation year completed = 91%

Adjustment = USQCD jeopardy policy in action
(https://www.usqcd.org/jeopardy.pdf)

* Adjustment bars (gray) added either on top or next to original allocations (blue) to reflect awards or penalties per jeopardy policy rules

2023-2024 Allocations Summary – “19g” GPU cluster

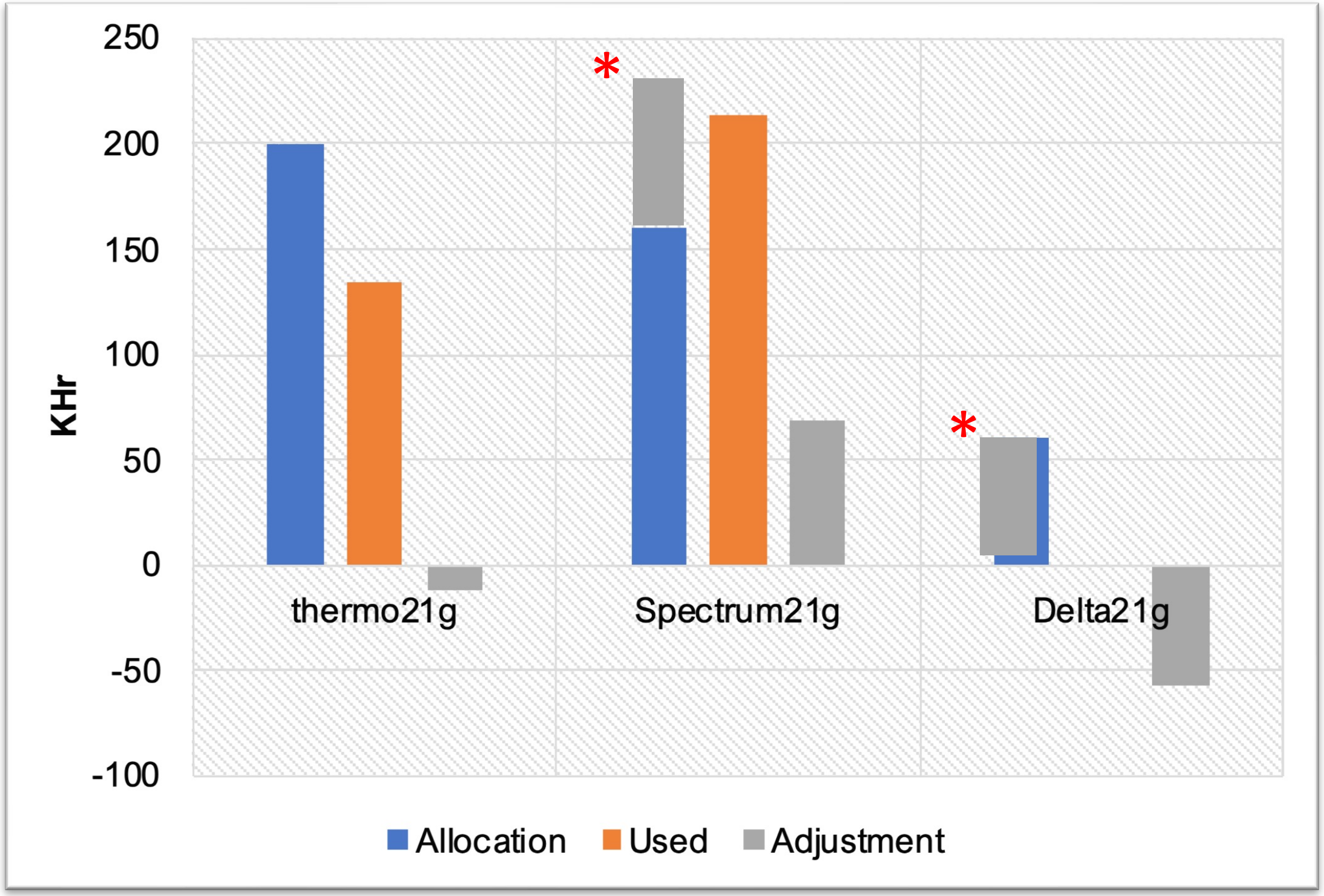


Total allocations used thus far with 80% of allocation year completed = 84%

Adjustment = USQCD jeopardy policy in action
(https://www.usqcd.org/jeopardy.pdf)

* Adjustment bars (gray) added either on top or next to original allocations (blue) to reflect awards or penalties per jeopardy policy rules

2023-2024 Allocations Summary – “21g” GPU cluster



Total allocations used thus far with 80% of allocation year completed = 83%

Adjustment = USQCD jeopardy policy in action
(https://www.usqcd.org/jeopardy.pdf)

* Adjustment bars (gray) added either on top or overlap original allocations (blue) to reflect awards or penalties respectively

Current disk and tape status

Lustre - Storage for /volatile and /cache

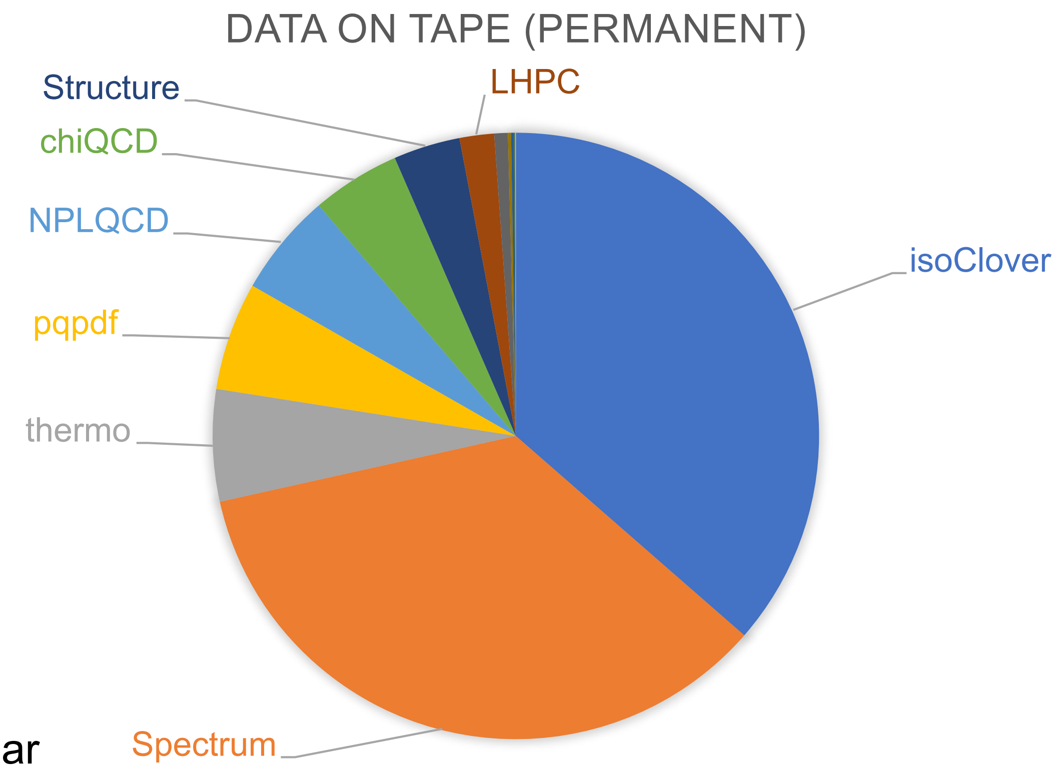
- 2.3PB (actual available 1.9PB) parallel and distributed Lustre file-system
- /cache gets backed to tape automatically when quota exceeded
- /volatile – working on changing file deletion policy

NFS file server - /work and /home on ZFS.

- /work NFS on ZFS and is not backed up (https://lqcd.jlab.org/lqcd/workDisk)
- /home is flash storage and is backed up

Tape Storage

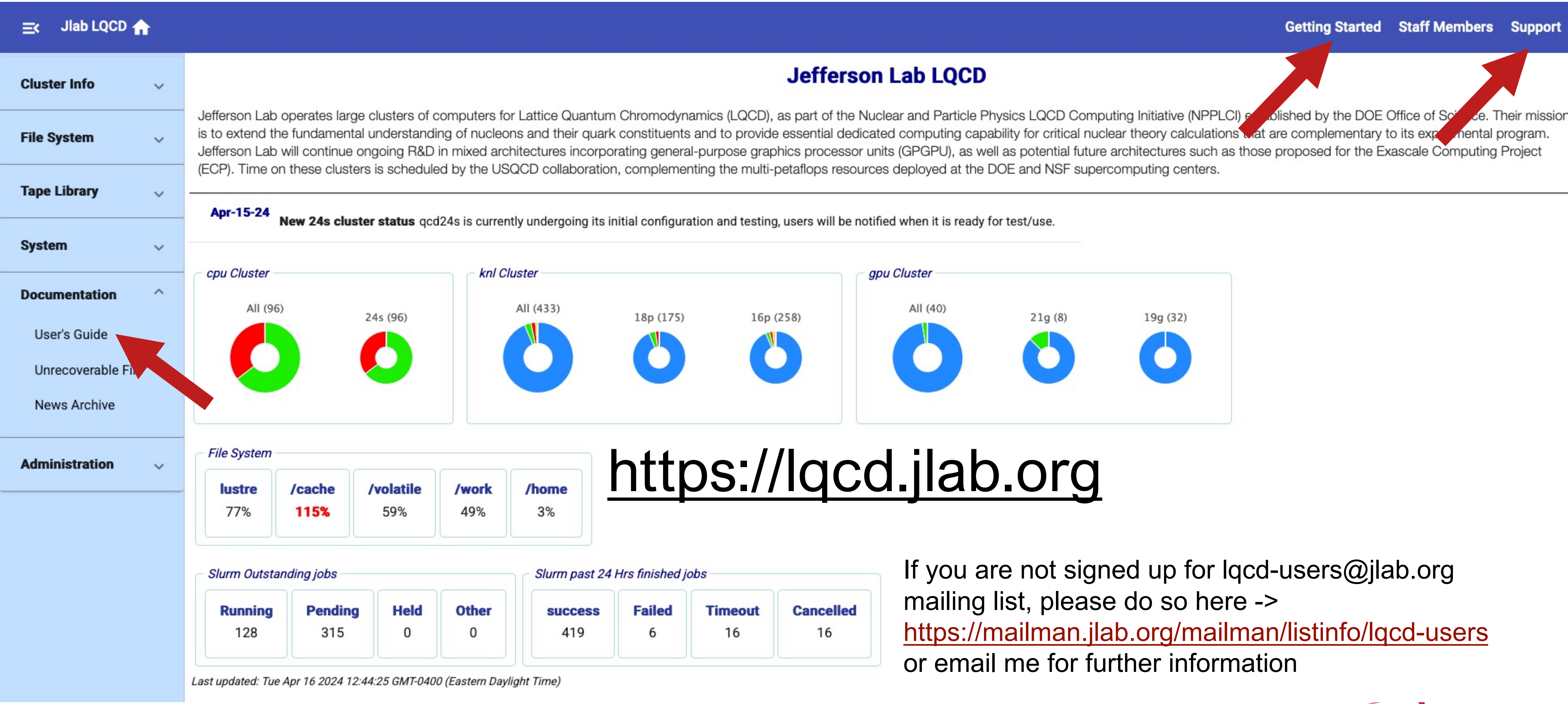
- To date LQCD accumulated storage is 17PB on tape (https://lqcd.jlab.org/lqcd/cacheDisk/project)
  - 13.5PB on lattice-p "permanent"
  - 3.5PB on lattice-t "temporary"
  - Tape storage for lattice-t USQCD (non-JLab) allocations are retained at Jefferson lab for 18 months after the allocation year ends, then the tapes are re-used
- All tape related costs (minus media) are an in-kind contribution by JLab



A few words from the operations team – 2023-24

- SLURM was upgraded to prepare for AlmaLinux 9, which is the target for 24s and future clusters. RHEL 9 will be available on 21g to stay in the AMD ROCm support matrix, but the upgrade path and configuration management are essentially the same between Alma and RHEL.
- Jlab's internet connection was upgraded to 2x100Gbps. This allows for faster transfers using the two Globus Data Transfer Nodes.
- We made changes to our system change management processes using puppet, which we hope was largely user-invisible. This aligns the LQCD cluster management practices with the rest of the scientific computing environment.
- We updated the bridge (LNET routers) between Omni Path and Infiniband to EDR (100Gbps) as part of their lifecycle replacement.
- For 2023 we had 60 support tickets and 92% of the tickets were resolved within 3 days or less.
- Reminder
  - /volatile and /cache are on Lustre. /volatile is not backed up. /cache is written to tape.
  - Infiniband core network is HDR (200Gbps) Infiniband.

User Documentation & how to ask for support



Jefferson Lab LQCD

Jefferson Lab operates large clusters of computers for Lattice Quantum Chromodynamics (LQCD), as part of the Nuclear and Particle Physics LQCD Computing Initiative (NPPLCI) established by the DOE Office of Science. Their mission is to extend the fundamental understanding of nucleons and their quark constituents and to provide essential dedicated computing capability for critical nuclear theory calculations that are complementary to its experimental program. Jefferson Lab will continue ongoing R&D in mixed architectures incorporating general-purpose graphics processor units (GPGPU), as well as potential future architectures such as those proposed for the Exascale Computing Project (ECP). Time on these clusters is scheduled by the USQCD collaboration, complementing the multi-petaflops resources deployed at the DOE and NSF supercomputing centers.

Apr-15-24 **New 24s cluster status** qcd24s is currently undergoing its initial configuration and testing, users will be notified when it is ready for test/use.

File System

lustre	/cache	/volatile	/work	/home
77%	115%	59%	49%	3%

Slurm Outstanding jobs

Running	Pending	Held	Other
128	315	0	0

Slurm past 24 Hrs finished jobs

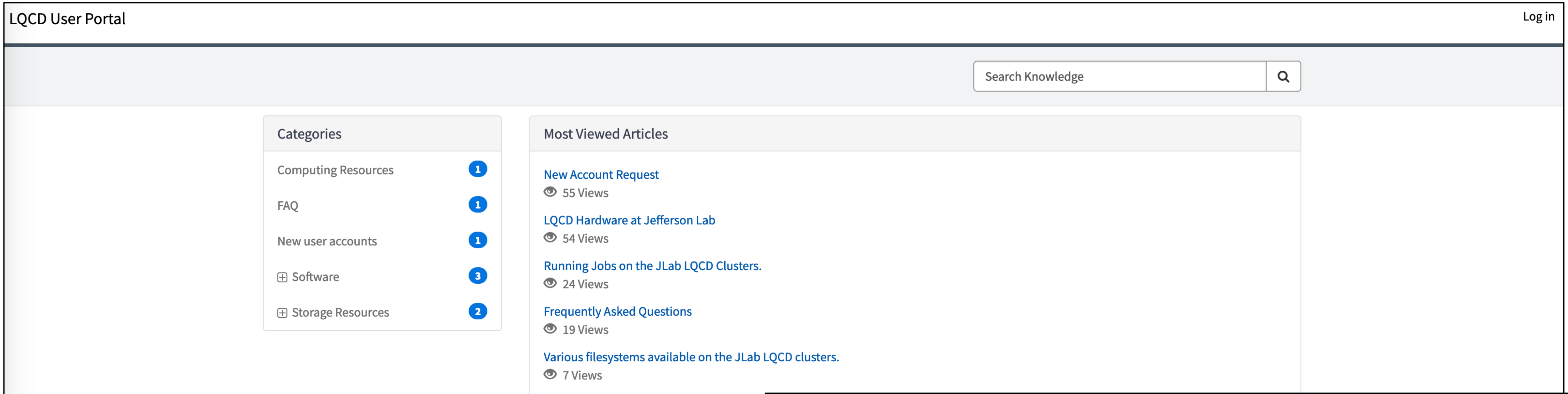
success	Failed	Timeout	Cancelled
419	6	16	16

Last updated: Tue Apr 16 2024 12:44:25 GMT-0400 (Eastern Daylight Time)

https://lqcd.jlab.org

If you are not signed up for lqcd-users@jlab.org mailing list, please do so here -> https://mailman.jlab.org/mailman/listinfo/lqcd-users or email me for further information

User Documentation in the form of Service NOW Knowledge base articles



LQCD User Portal

Log in

Search Knowledge

Categories

Computing Resources	1
FAQ	1
New user accounts	1
⊞ Software	3
⊞ Storage Resources	2

Most Viewed Articles

New Account Request
55 Views

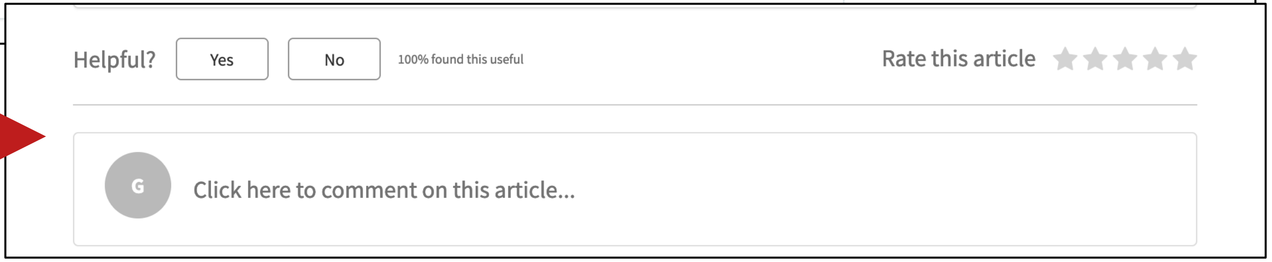
LQCD Hardware at Jefferson Lab
54 Views

Running Jobs on the JLab LQCD Clusters.
24 Views

Frequently Asked Questions
19 Views

Various filesystems available on the JLab LQCD clusters.
7 Views

To help us improve the quality of our documentation You can rate or comment on each article !!



Helpful? Yes No 100% found this useful

Rate this article ☆☆☆☆☆

Click here to comment on this article...

https://jlab.servicenowservices.com/lqcd?id=kb_view2

Questions?

Amitoj Singh

amitoj@jlab.org

JLab LQCD Main Page	https://lqcd.jlab.org/
New User Accounts	https://jlab.servicenowservices.com/lqcd?id=kb_article_view&sysparm_article=KB0014813
Submit a helpdesk ticket form	https://lqcd.jlab.org/lqcd/support
Knowledgebase articles	https://jlab.servicenowservices.com/lqcd?id=kb_view2
Mailing List	https://mailman.jlab.org/mailman/listinfo/lqcd-users
FAQ’s	https://jlab.servicenowservices.com/lqcd?id=kb_article&sysparm_article=KB0014827