Jefferson Lab LQCD Computing April 2021 All Hands Meeting

Bryan Hess, *bhess@jlab.org* – Scientific Computing Operations Group

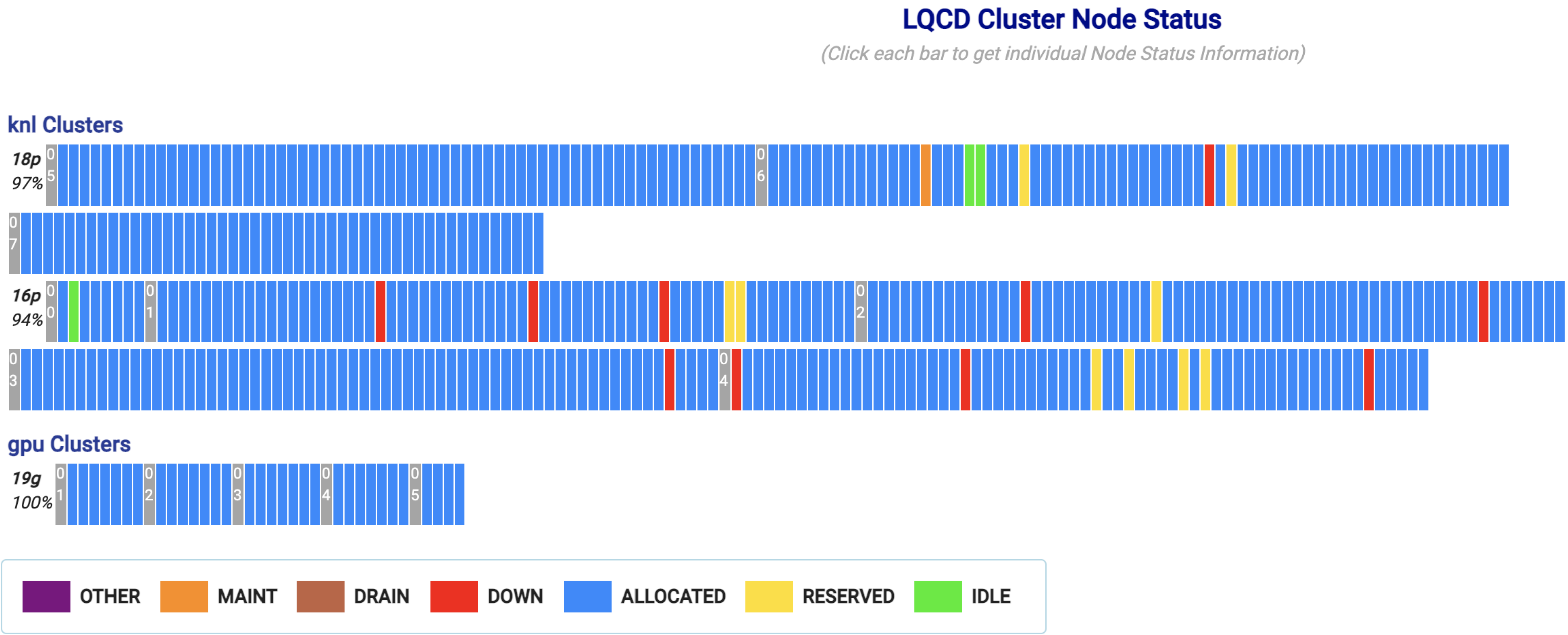
Facility Topics

- Status of Clusters
- Data Center Network Upgrade
- Lustre Upgrade
- Tape Library Upgrade and Status
- Significant Operational Events from the year
- System Administration and Change Control

Current resources – Clusters

- JLab continues to operate two flavors of cluster, KNL and 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)
  - 16GB 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
- The 12k cluster was decommissioned in July 2020.
- 21g procurement has been awarded – More on this from Amitoj Singh shortly

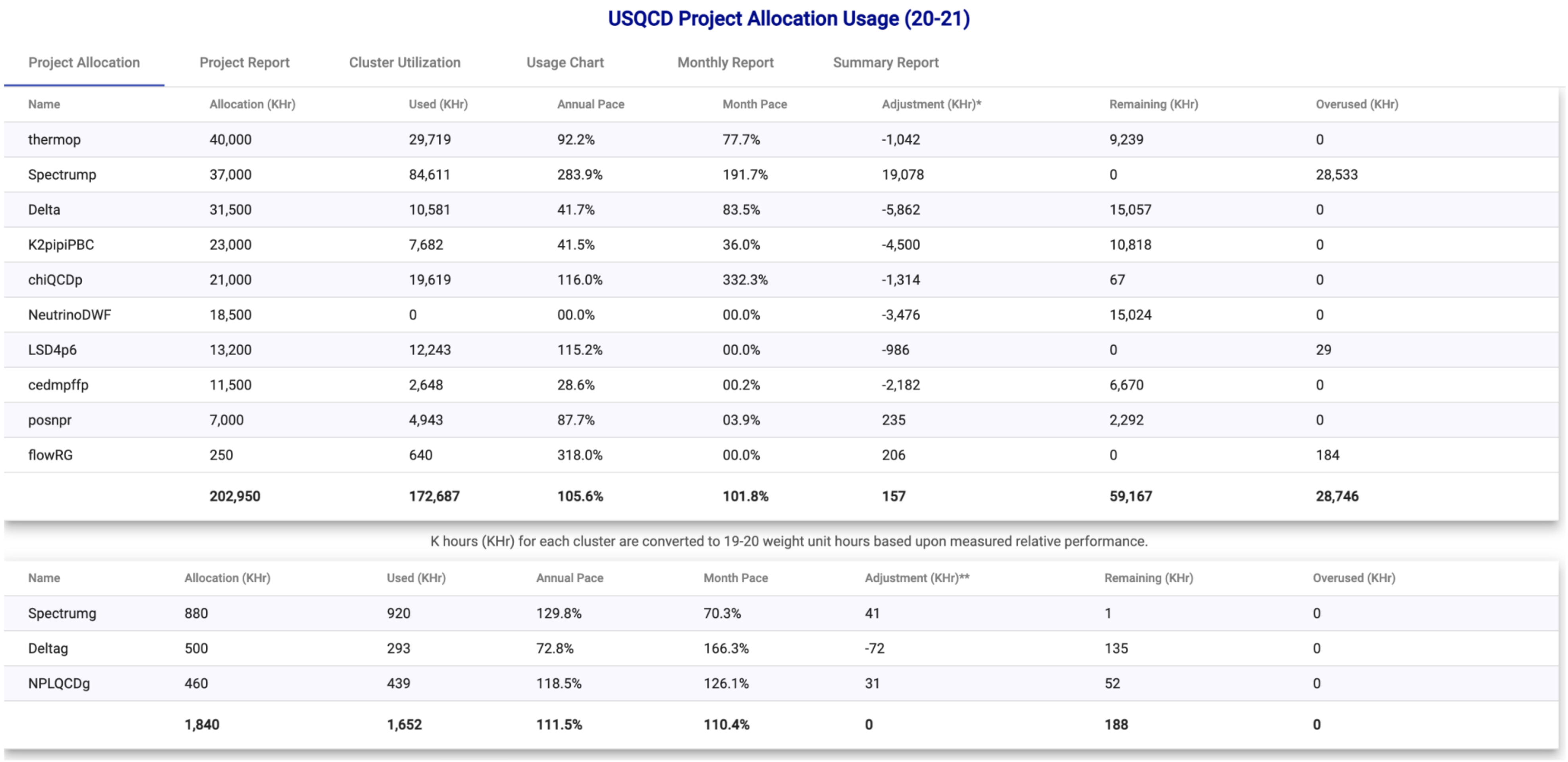
Current resources - clusters



LQCD Cluster Node Status

(Click each bar to get individual Node Status Information)

User Allocation statistics – from lqcd.jlab.org



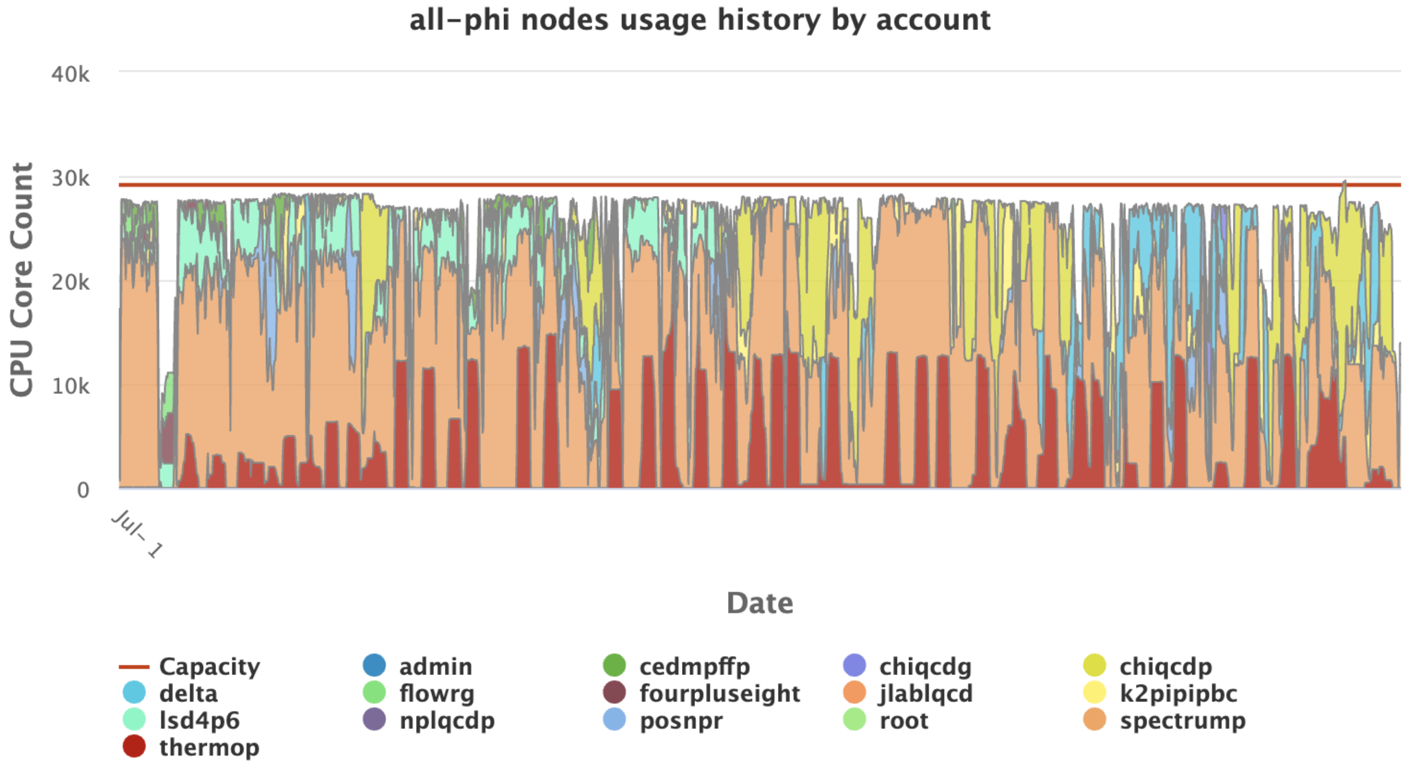
USQCD Project Allocation Usage (20-21)

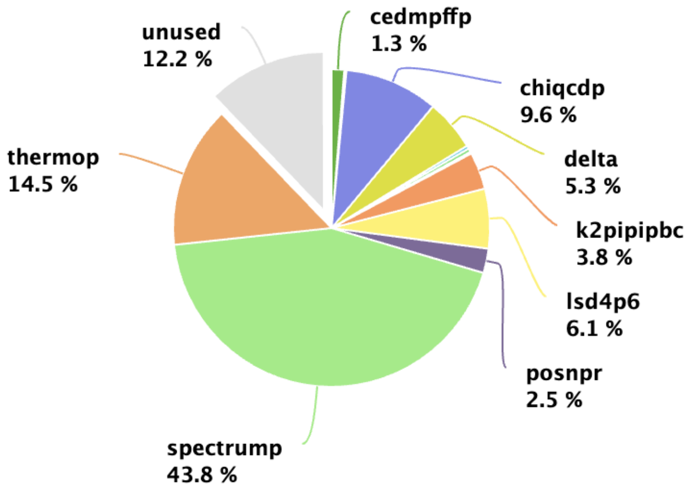
Name	Allocation (KHr)	Used (KHr)	Annual Pace	Month Pace	Adjustment (KHr)*	Remaining (KHr)	Overused (KHr)
thermop	40,000	29,719	92.2%	77.7%	-1,042	9,239	0
Spectrump	37,000	84,611	283.9%	191.7%	19,078	0	28,533
Delta	31,500	10,581	41.7%	83.5%	-5,862	15,057	0
K2pipiPBC	23,000	7,682	41.5%	36.0%	-4,500	10,818	0
chiQCDp	21,000	19,619	116.0%	332.3%	-1,314	67	0
NeutrinoDWF	18,500	0	00.0%	00.0%	-3,476	15,024	0
LSD4p6	13,200	12,243	115.2%	00.0%	-986	0	29
cedmpffp	11,500	2,648	28.6%	00.2%	-2,182	6,670	0
posnpr	7,000	4,943	87.7%	03.9%	235	2,292	0
flowRG	250	640	318.0%	00.0%	206	0	184
	202,950	**172,687**	**105.6%**	**101.8%**	**157**	**59,167**	**28,746**

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

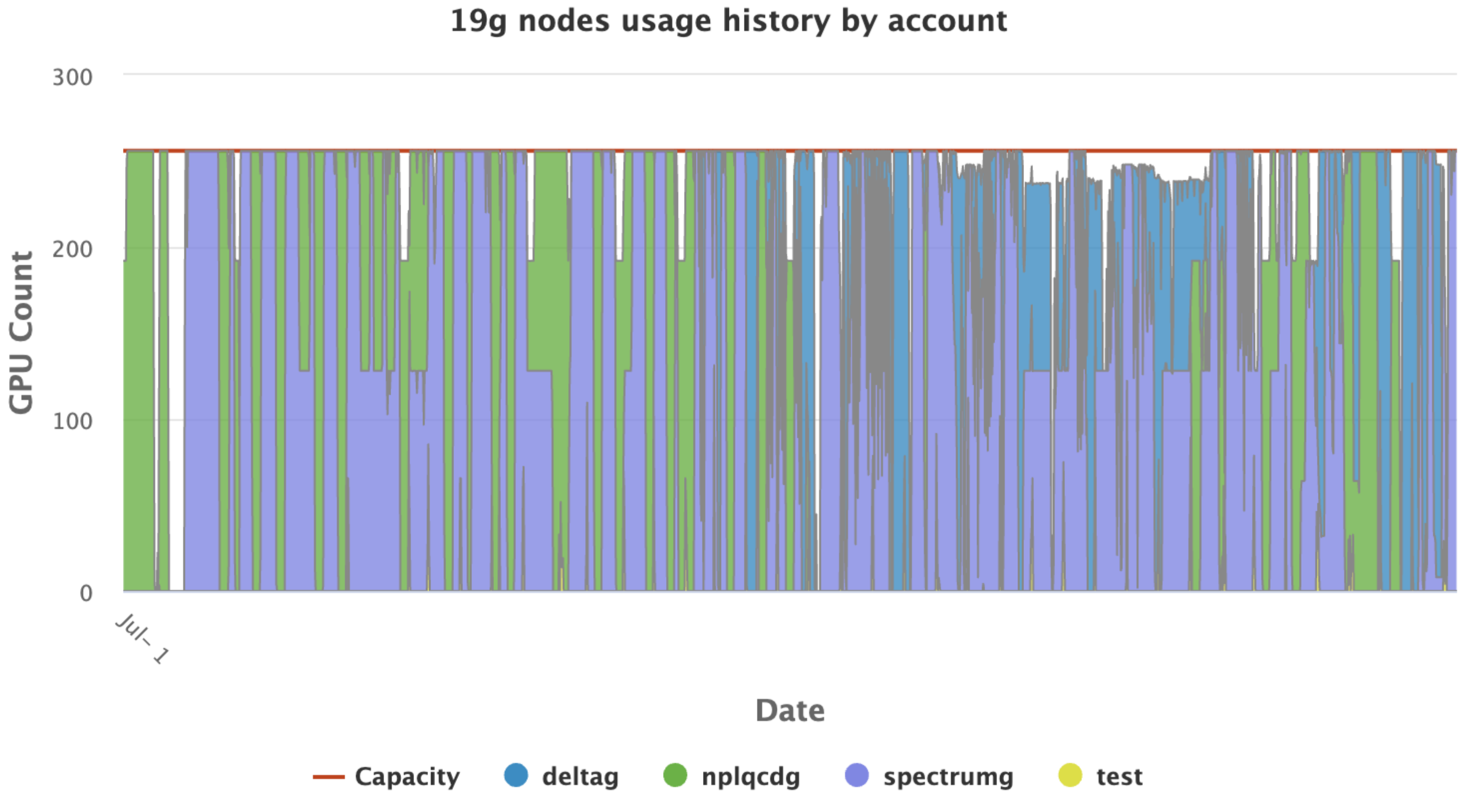
Name	Allocation (KHr)	Used (KHr)	Annual Pace	Month Pace	Adjustment (KHr)**	Remaining (KHr)	Overused (KHr)
Spectrumg	880	920	129.8%	70.3%	41	1	0
Deltag	500	293	72.8%	166.3%	-72	135	0
NPLQCDg	460	439	118.5%	126.1%	31	52	0
	1,840	**1,652**	**111.5%**	**110.4%**	**0**	**188**	**0**

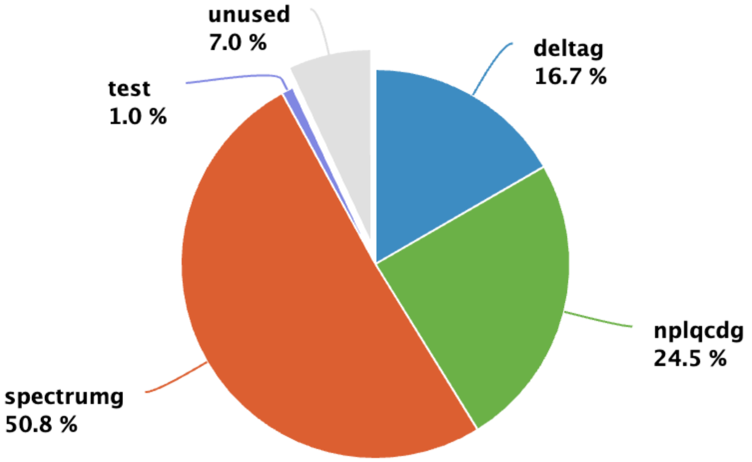
KNL Cluster Usage (16p + 18p) 7/1/2020 - 4/21/2021



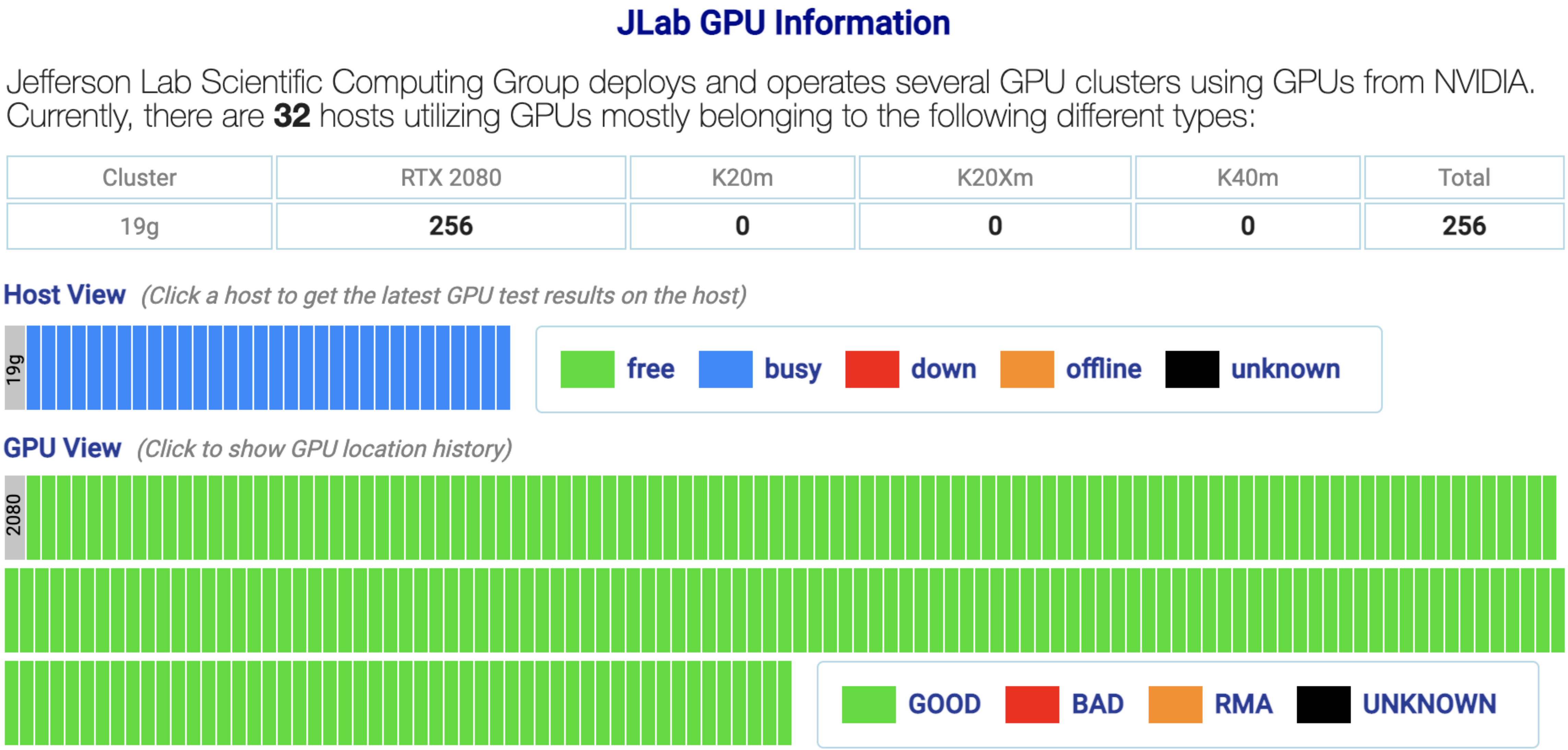


19g Cluster Usage 7/1/2020 - 4/21/2021





GPU Health Monitoring



JLab GPU Information

Jefferson Lab Scientific Computing Group deploys and operates several GPU clusters using GPUs from NVIDIA. Currently, there are **32** hosts utilizing GPUs mostly belonging to the following different types:

Cluster	RTX 2080	K20m	K20Xm	K40m	Total
19g	**256**	**0**	**0**	**0**	**256**

Host View *(Click a host to get the latest GPU test results on the host)*

🟩 free 🟦 busy 🟥 down 🟧 offline ⬛ unknown

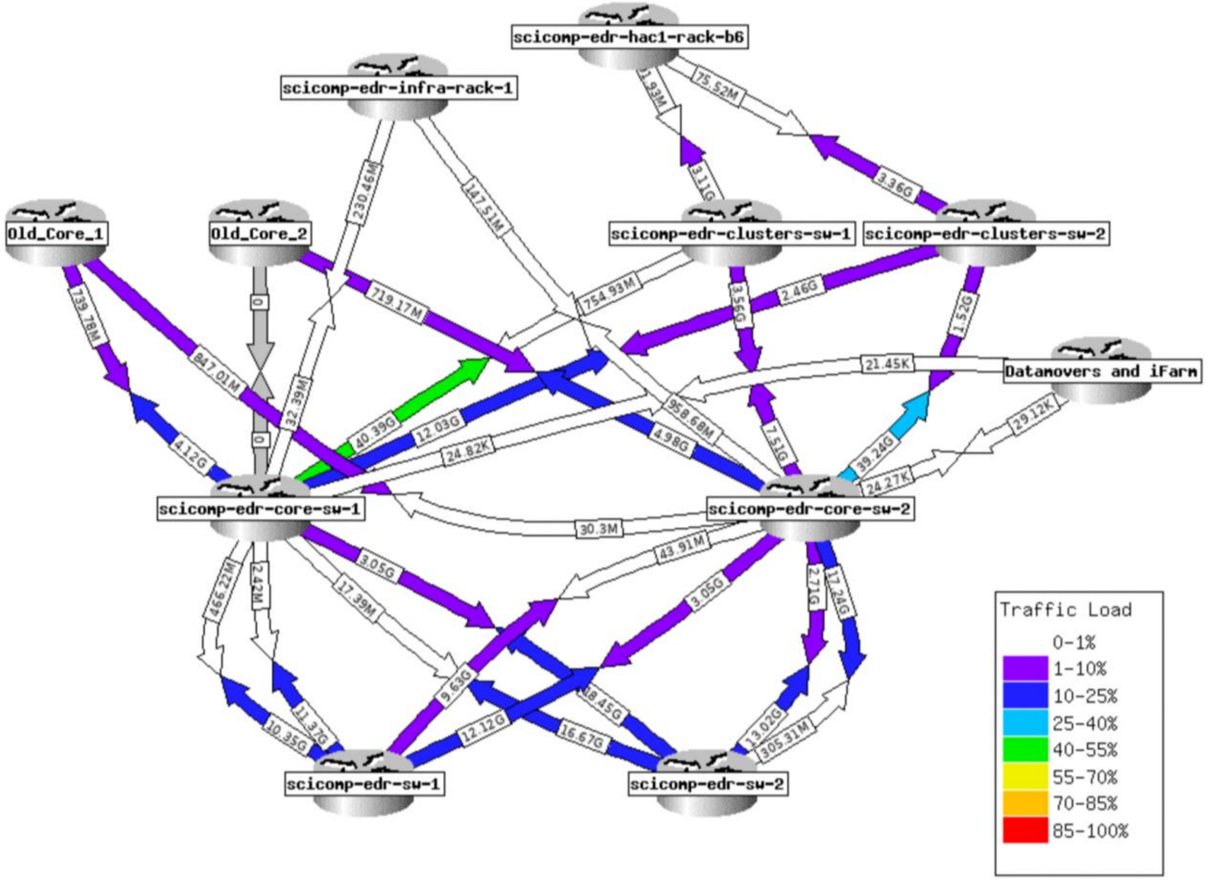
GPU View *(Click to show GPU location history)*

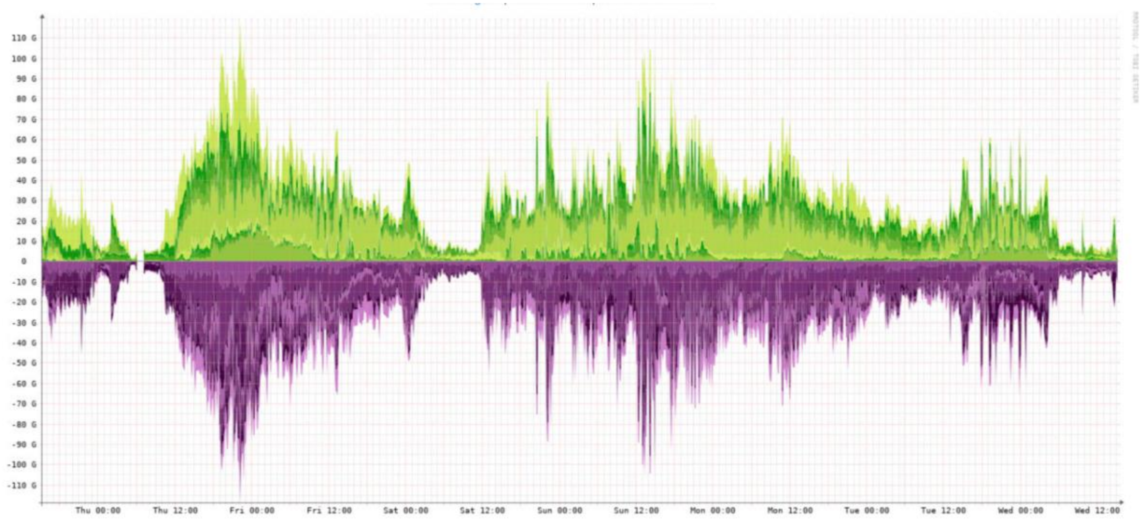
🟩 GOOD 🟥 BAD 🟧 RMA ⬛ UNKNOWN

Network Upgrade – from QDR to managed EDR

- Clusters are on OmniPath
- Connected to the IB core via LNET routers
- The IB Core was recently upgraded from an unmanaged QDR core to a managed EDR core
- Improvements include
  - Lustre on managed switches with switchport speeds better matched to the clusters
  - Clear roles for switches and interconnects
  - Clean tolopology
  - Monitoring and alerting on error conditions including speed mismatch, drops, and down links
- Internet Bandwidth now 20Gbit/sec due to routing improvements

Network Management and Alerting for IB – Faster Response to Faults

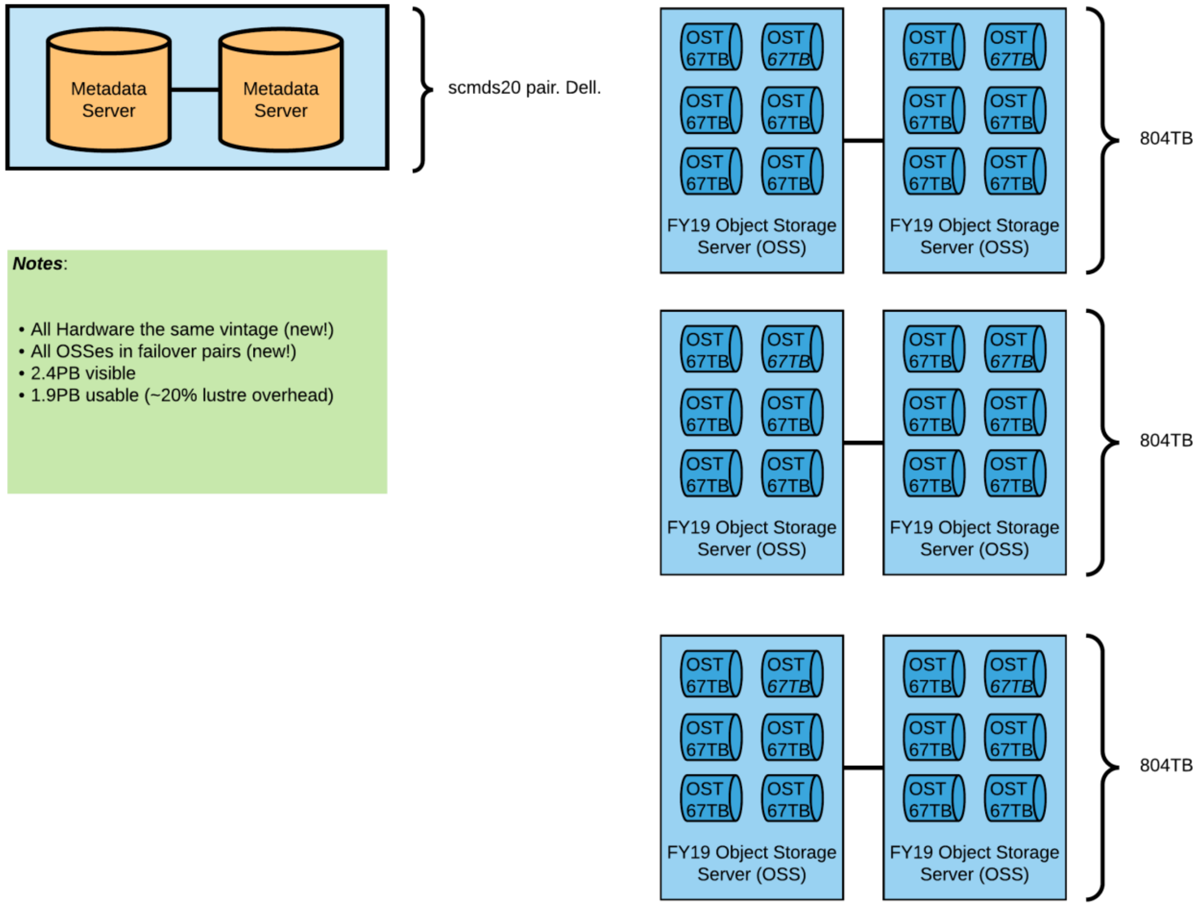




- Network monitoring identifies problems more quickly
- IB core interconnects clusters, tape storage, disk storage, and servers.
- Deprecated QDR core on top left
- User-visible services (including Lustre) dual-connected for redundancy

Lustre Upgrade – Lustre20 upgrade Completed April 2021

LQCD Lustre20 refresh - /cache and /volatile for LQCD



Notes:

- All Hardware the same vintage (new!)
- All OSSes in failover pairs (new!)
- 2.4PB visible
- 1.9PB usable (~20% lustre overhead)

- The Legacy Lustre filesystem has been replaced by a dedicated system for LQCD, lustre20
- Significant Improvements
  - No longer shared with the HTC cluster
  - All hardware the same specification to avoid hot spots
  - All servers in failover-pairs to eliminate single points of failure in hardware components
  - Lustre Upgraded to latest stable release with increased logging and root squash for compute nodes
  - Managed EDR switches to monitor performance and match cluster speed for 20g
- New Capability: Testbed Hardware for testing and staging changes

Disk Storage Summary

Filesystem	Filesystem type	Backup	Quota	Deletion Policy
/work	NFS	None	Per project quota	User managed
/home	NFS	Weekly backups	Per user quota	User managed
/cache	Lustre	Flush to tape within two weeks for files sizes between 3MB and 300GB.	Per project quota	Auto-deletion based on least recently used once on tape
/volatile	Lustre	Not backed up	Per project quota	Auto-deletion based on LRU
/scratch	local to worker node	none	none	Scrubbed after each job

Disk Issue: July 9th Lustre deletion event

- Two filesystems, /cache and /volatile, are Lustre resident, and not backed up by design.
- On Thursday, July 9th at 4:23 pm, a software process began deleting files in the Jefferson Lab LQCD Lustre area, removing the contents of /volatile and /cache
- The Lustre system, dating from 2014, had insufficient logging to determine the source of the I/O. Additionally, root access was enabled from many compute nodes for legacy support reasons.
- Although /cache is backed by tape write, /volatile is not, and the content was lost.
- This is a Community supported Lustre system; There is no support contract with a vendor.

Lustre deletion event lessons learned, mitigations

- We apologize for the loss and interruptions, and have taken measures to limit future events
  - Root access is now limited to select administrative hosts (root squash)
  - The Lustre system has been replaced with a newer one capable logging additional information in the changelog
  - Access to the Lustre system has been trimmed to the essential systems
  - The legacy system build process has been retired.
  - A new change control system has been put in place (more on this)

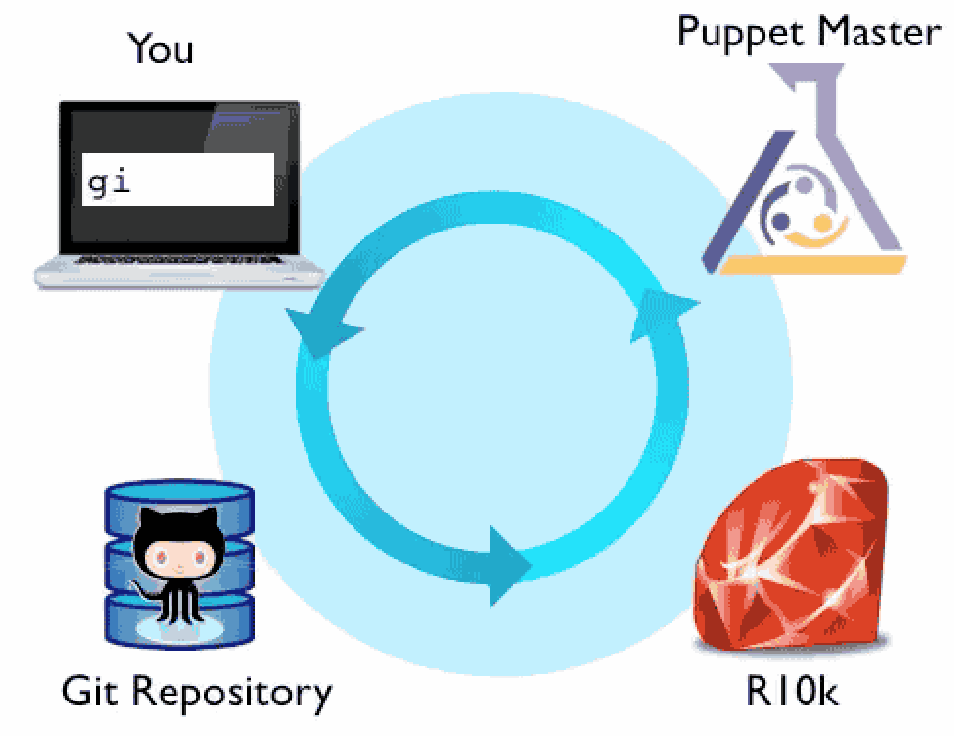
Configuration Management

- A follow-on activity from the Lustre event was an increased emphasis on the configuration management upgrade
- Replaced SaltStack Legacy System
  - Limited on-site expertise
  - No dashboard, no fine-grained monitoring of change deployments
  - Loose version control
- Migrated to Puppet
  - Leveraged on-site expertise and common tools
  - Foreman used for dashboard (Upstream project of RedHat's Satellite)
  - Easily test/stage changes using Puppet environments
  - ***Mandatory*** version control (completely driven from GitLab)

Formalized change control with Puppet, Git, Foreman

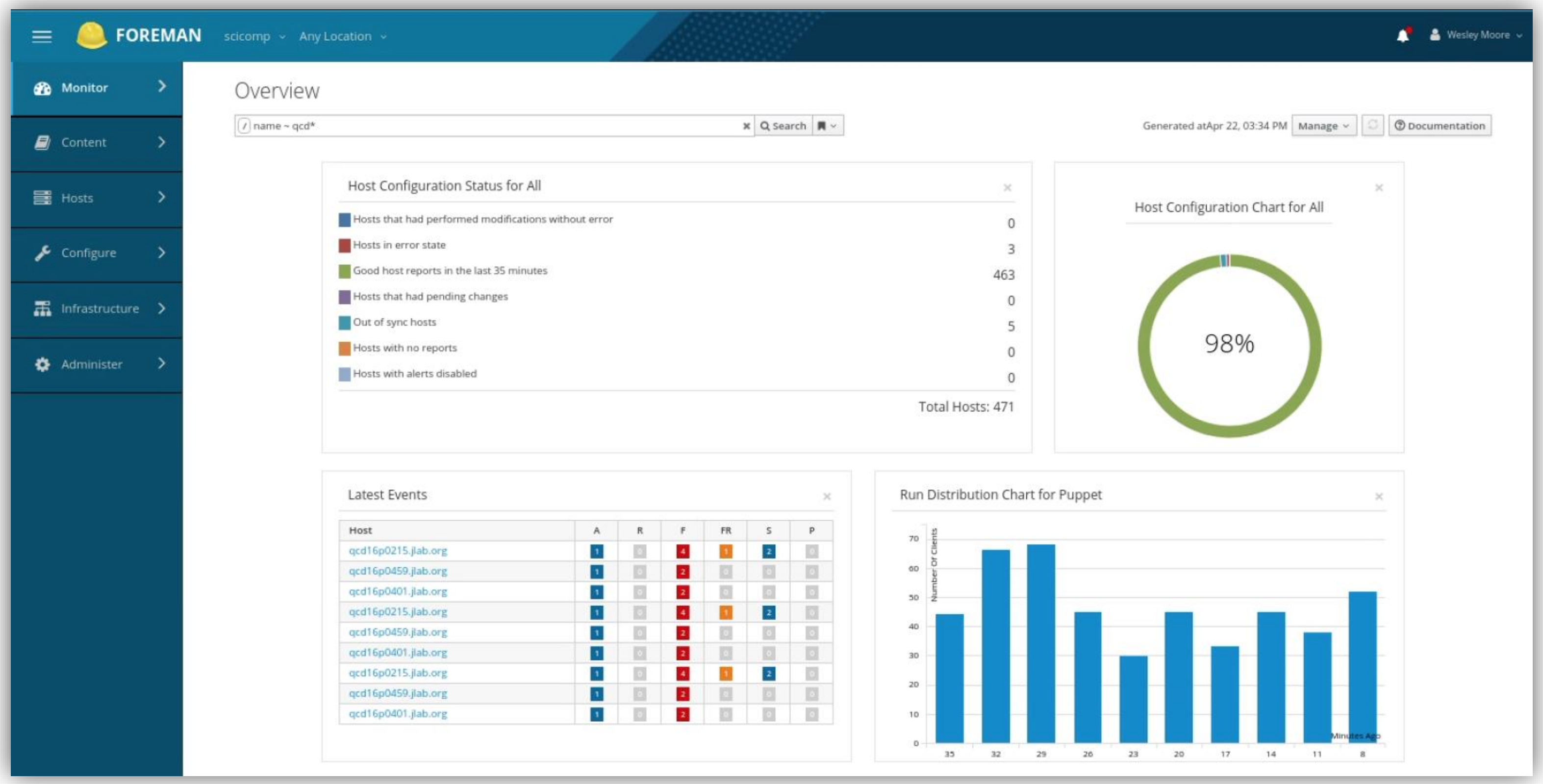
Mandatory version control for uniformity of compute nodes:

- Fully automated, triggered by “git push”
- GitLab server allows for syntax checking before deployment
- Deploys to a cluster of Puppet servers, prioviding load-balancing
- Git and Puppet development environments are useful for testing configurations



Administration Dashboard – quickly identifies irregularities in the clusters

- Same services used for HTC cluster at JLab
- Dashboard for with status of Puppet runs, patching, check-in times
- Provides Yum repository for all rpms available to the nodes
- Easy remote command execution for individual nodes or batches



Overview

name ~ qcd*

Generated at Apr 22, 03:34 PM

Host Configuration Status for All

🟦 Hosts that had performed modifications without error	0
🟥 Hosts in error state	3
🟩 Good host reports in the last 35 minutes	463
🟪 Hosts that had pending changes	0
🟦 Out of sync hosts	5
🟧 Hosts with no reports	0
🟦 Hosts with alerts disabled	0

Total Hosts: 471

Host Configuration Chart for All

Latest Events

Host	A	R	F	FR	S	P
qcd16p0215.jlab.org	1	0	4	1	2	0
qcd16p0459.jlab.org	1	0	2	0	0	0
qcd16p0401.jlab.org	1	0	2	0	0	0
qcd16p0215.jlab.org	1	0	4	1	2	0
qcd16p0459.jlab.org	1	0	2	0	0	0
qcd16p0401.jlab.org	1	0	2	0	0	0
qcd16p0215.jlab.org	1	0	4	1	2	0
qcd16p0459.jlab.org	1	0	2	0	0	0
qcd16p0401.jlab.org	1	0	2	0	0	0

Run Distribution Chart for Puppet

Current status – Tape Library

- Tape Storage - increased bandwidth
  - Data is now written to LTO8 media using a 20 LTO8 drives
  - In November, the two tape libraries were consolidated into a single TS4500 library, and 5 additional tape drives were added to improve bandwidth to tape.
  - **LQCD accumulated tape storage for this project year is 1.7PB**
    - 1.5PB on lattice-p "permanent"
    - 0.2PB 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.
  - **LQCD accumulated tape storage for all years is 9.94PB**
    - 9.06PB lattice-p
    - 0.88PB lattice-t

Tape Verification and Recovery

- Analysis
  - Two problematic drives were torn down for electron microscope evaluation and chemical analysis
  - This work pointed to humidity as the primary contributing factor for the head corrosion
- Monitoring
  - We are validating checksums on all data written to LTO8 media to find tapes with read errors.
- Progress
  - Because this is an ongoing process, we have established a database of the files that are offline and in the queue for recovery at https://lqcdtest.jlab.org/lqcd/badTapeFile
  - Files that are recovered will be removed from this list.
  - Files that are not recoverable will be reported here: https://lqcd.jlab.org/lqcd/lostFile
- Prevention of a repeat
  - We have installed additional environmental monitors in the tape library and are working with facilities management to on new monitoring and controls.

COVID-19 Era Work Adaptations

- On-site data center work proved challenging in the early months of the pandemic because of personnel distancing limitations
  - Installing and removing heavy disk shelves and servers required close work, which was initially prohibited
  - After delays, close work was allowed with N95 masks and training
  - Work coordindation, PPE supplies, and policies have improved over the year
- Over the summer we had minimal staffing in the data center
- By the autumn we transitioned to a schedule where there was always staff on site during business hours
- Though the lab is operating at "maximum telework" we now have at least two system administrators on site every weekday

Summary

- Working remotely has been challenging for installations, but protocols are in place now.
- Cluster operations have continued at pace
- Lessons Learned and new practices stemming from two user-impacting events
- Significant improvments to the facility infrastructure
  - Replced the dated lustre system
  - New managed IB network
  - Increased tape library bandwidth
  - Change control system for system management
  - Improved system monitoring
  - Upgraded slurm and system software
- Preparing for 21g cluster
- Thank you!