Jefferson Lab LQCD Computing

April 2019 All Hands Meeting

Chip Watson

Scientific Computing Group

Outline

New GPU resource at JLab
File System & Archival Storage
Operations & staff changes
Looking forward to FY2020

Nuclear and Particle Physics LQCD Computing Initiative

Reminder

- Single lab, NP funded, serves all of USQCD, and is complementary to the modified (2 lab, IC based) HEP LQCD project
- $1M per year, about half hardware, half labor
  (equals average NP investment per year at JLab for last 10 years, so no real change in funding for the lab)
- FY2018: upgraded Jlab's KNL resources
- FY2019: upgraded Jlab's GPU resources

  (both of these steps were specified in the initiative proposal)

GPU Resources going into FY2019

Old quad k20 cluster (2012, Kepler)

- Reached 6 years of age in November, servers slowly dying, now down to 41 (more cards than servers can hold); we anticipate we can deliver 32 node years in the coming allocation year.

FY 2019 Possibilities Considered (last October)

- Quad or Octal v100 (Volta Tesla professional)
  roughly 1.5x Pascal performance, 6x 2012 Kepler
- Octal RTX-2080 (newest gamer, released late summer)
- Goal: bring a new GPU resource online before the 12k cluster was retired so as to continue supporting mixed architecture workflows (KNL + GPU)

USQCD 2019-2020 Planned GPU Resources

Exclusive of Jlab's new resource:

1.07M Nvidia k80 hours

Assumed conversions (from the call for proposals)

1 k80 = 0.5 p100 = 2.2 k40 = 3 k20m

Jlab's new resource choices, Fall 2019:

Option 1: v100 = ~1.5 p100 ~ 3k80

Option 2: rtx2080 = ~ 0.3-0.44 v100 ~ 0.9-1.3 k80

(this rating is only for NP multi-grid inverter; call quoted 2080/k80 = 1)

A v100 w/ 32GB memory was considered a higher value than p100 w/ 16 GB memory even though p100 could deliver marginally more flops/dollar.

Factors in GPU Selection

RTX strengths

- Highest memory bandwidth / dollar
- Highest multi-grid inverter performance / dollar

RTX weaknesses

- No ECC memory (should really think “inverters only”)
- No GPU-GPU communications (must use CPU)
- No GPU-Fabric communications (must use CPU)
- Small memory per GPU (8 GB; the 11 GB Ti version was rumored to have overheating issues; v100 has 16 or 32 GB)

Selection Optimization

Within the available budget, Jlab could have added 0.4M v100 based k80 hours, with this cluster thus becoming 27% of USQCD GPU resources.

Because the NP community is typically the larger user of GPU resources, it was reasonable to explore optimizing this 27% purchase for NP users (who in any case are the most prevalent users at Jlab).

Multiple NP users favored the higher aggregate performance of the RTX-2080 despite its weaknesses.

A separate careful evaluation was made as to whether a small number of users could keep this unique resource fully productive.

Answer: yes (based on promises from users).

Final Configuration

Cluster Name: 19g (2019, GPU)

32 nodes, each with

- octal RTX 2080, 8 GB/card
- dual Intel Gold 5118 (12 core),
- 192 GB memory
- 100 Gbps Omnipath
  (single rail, single switch)
- 12 PCIe x16 slots,
  (can upgrade to dual OmniPath)
- 1 TB NVMe SSD (3 GB/s)

(We intentionally compromised CPU performance to reach 2^N nodes and GPUs.)

256 GPUs --- 280 – 380 Gflops/GPU on multi-grid
(> 64 TFlops total; depends on how full the 8 GB is)

Current Status

- April burn-in
  - 3 of 264 cards (including 8 spares) have exhibited high memory faults and have been returned for replacement
  - The remaining cards have 0 reported errors (tempting?)
  - All cards are tested nightly (reduce to weekly after burn-in is done)
- May – June
  - Continue burn-in, moving towards science running + code and workflow optimizations
  - Move some 12k GPU allocations to this cluster and shrink 12k from 41 nodes to 32 nodes, retiring 9 for other lab purposes (these will serve a spares for the coming 8th year for this long lived cluster)
  - We will slowly start to support other projects doing performance testing and even early starts on 2020 allocations.

Jefferson Lab Disk Resources

Today: 2.1 PB Lustre file system

- Shared with Experimental Physics, currently 60% LQCD
- Run 80% full to keep performance, so LQCD pools are ~ 1 PB
- Over 10 GB/s total bandwidth (have yet to see a saturation)
- Auto managed to < 80% full to avoid fragmentation

Soon: Adding an additional 1.2 PB (all for Experimental Physics)

- adds 8 GB/s of bandwidth, so you will see improvements
- New version of Lustre has a much higher performance SSD-based Meta Data Server (you will see the improvements)
- As we did years ago, we will be migrating all 1.7 PB of data from the old system to the new system project by project and server by server. This is expected to take 4-6 months. Please delete files you don't need on disk to accelerate this migration!

Tape Library -> Libraries

Today: 24 PB tape library, shared, with ~ 6 PB LQCD data

- Now doing data migration from LTO-4 and LTO-5 media to LTO-M8 media; includes 0.5 PB of mostly NP LQCD data
- please "delete" data from tape you no longer need, and/or migrate your old data to your host laboratory*; jremove tool can remove one file at a time, submit help ticket to remove large groups of files (we'll do it for you)

New IBM 4500 tape library, with 8 more LTO-8 drives

- Should improve bandwidth for all users by 1.5x in about 2 weeks

* Retention Policy (review) USQCD tape allocations at JLab are NOT for archival data storage, only for medium term (18 months) storage. For archival storage, contact the laboratory or other institution owning your science scope. (JLab will host medium energy NP archive.)

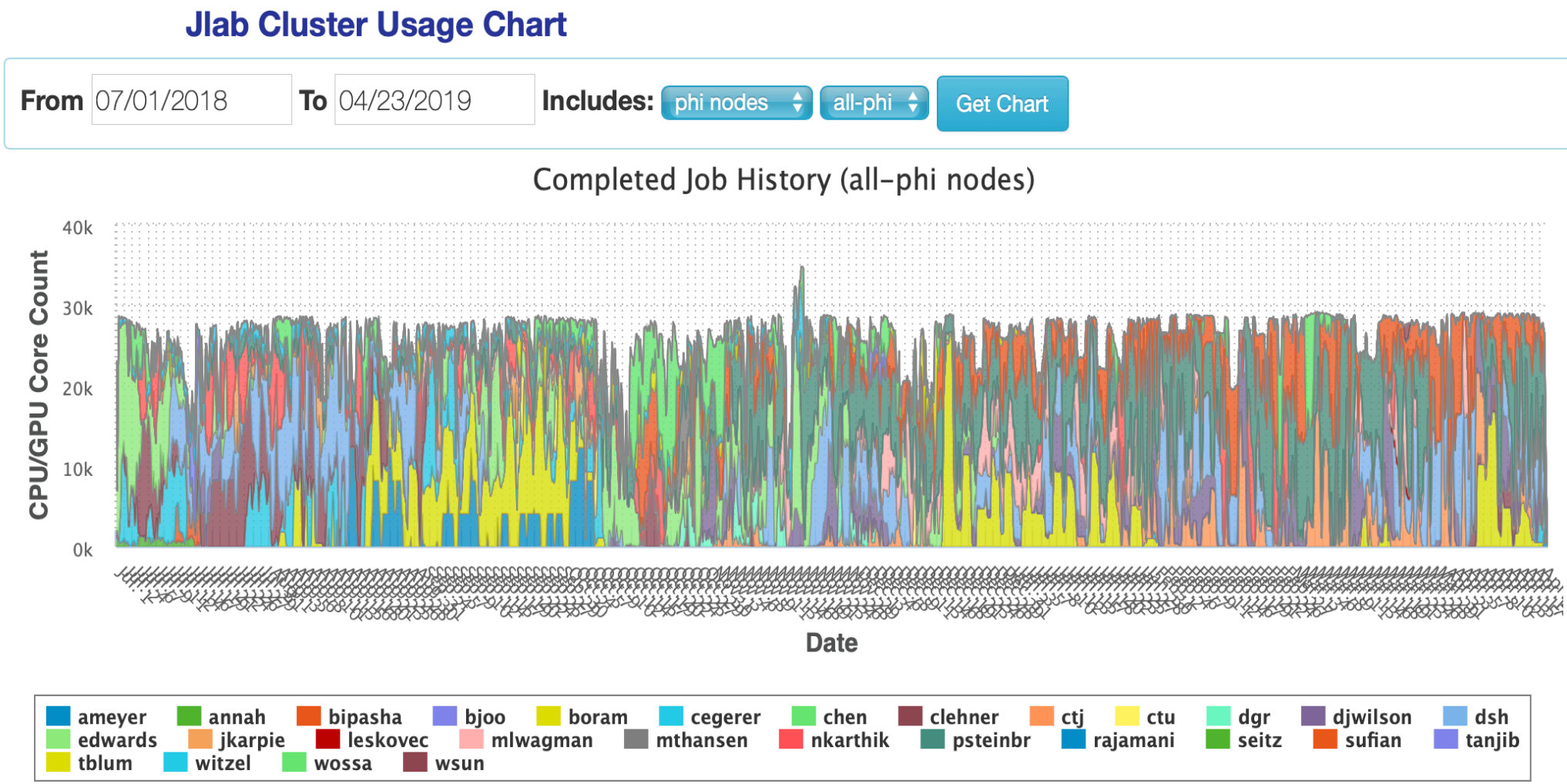
Recommendation

All projects creating tape (or having used tape in the past) should have a long term data management plan.

Operations

See lqcd.jlab.org and click on links to see ***many*** views of operations at Jlab.

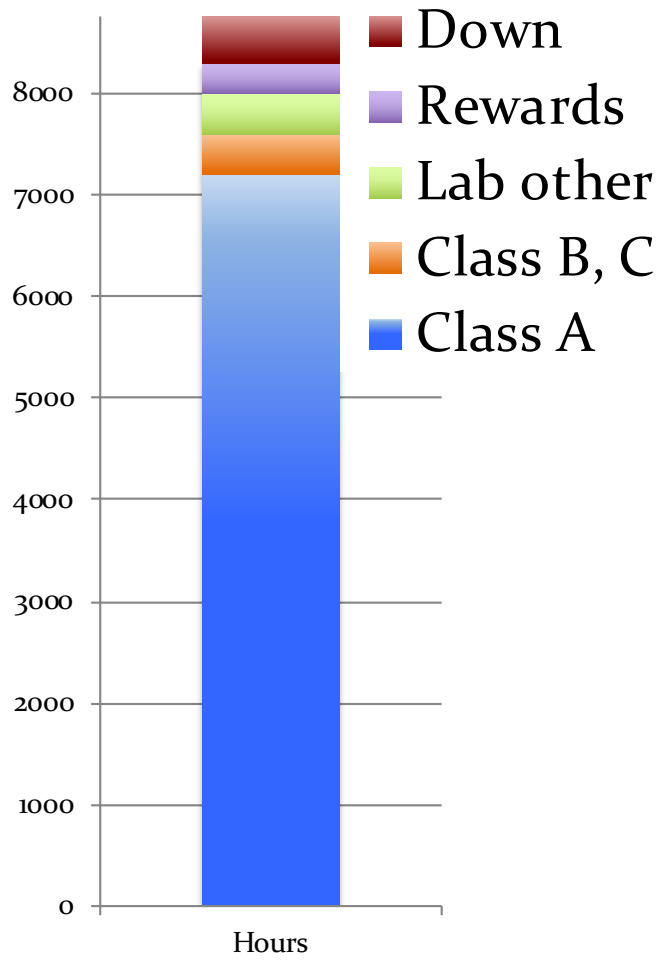
The KNL system remains very popular, and is well used (July 1 – present below)



(dis-) Incentives

Rewards Program

- Usage > 110% of pace in a month is **discounted 50%** (until allocation gone)
- Class A projects can never be more than 2 months behind (penalty);
- New for this year: usage will include draining time for very large jobs (rewards people who keep jobs of same size in the queue)



- Fair share growth is limited to 2x, and reduction is limited to 4x
- Rewards funded by high usage and penalties (aim at 120% pace)

Rewards Statistics this year

Awards given to projects of size > 10M hours:

- 7 of 9 have rewards
- Total rewards exceeds 5% of allocations (>32M KNL hrs to date)
- Utilization is 110% of pace (i.e. 7.92 K hours/year), and this does not (yet) include the time needed to drain for large jobs
- 2 projects have overrun even including their rewards (now at a balance of zero with fair share set to ¼ of initial fair share)
- There still remains ample resources to satisfy all allocations

Staff Changes

IT Division at Jefferson Lab has re-organized…

- Operations has moved over to another department to accommodate my further reduced hours and my growing portfolio (now adding Machine Learning)
- Bryan Hess (here today) is taking over as head of operations for Scientific Computing (be nice, he's still coming up to speed!)
- Sandy Philpott has retired, and last heard is enjoying more time with family

Long Range Thoughts for FY2020

There are multiple ideas under consideration for FY2020. Here's a short list (choose one or more):

- Small v100 cluster, possibly using 8 of the 19g servers to make it affordable (i.e. explore replacing 64 cards)
- SSD burst buffer, e.g. for tying GPU jobs to KNL jobs
- Memory upgrades for 16-32 18p nodes (96 GB -> 384 GB)
- Evaluate AMD Rome (will happen anyway for Experimental Physics this coming month)
- Hold funds until summer 2020 to make larger purchase across the fiscal year boundary (the original plan)

As always, feedback appreciated, especially specific use cases that can drive the design(s).

Questions?

Backup Slides…

Allocations and the Cost of Computing

Cost of operating a node per year is not cost / 6 years.

1. Declining balance depreciation of original cost over 6 years: 45% + 25% + 14% + 8% + 5% + 3% = 100%
2. Labor cost per node ~ FTE-year per 800 nodes plus ¼ FTE per cluster (almost constant per year, and can be approximated as 16% of purchase price per year)

This leads to a cost per node hour that falls like Moore's Law for the first few years, then stalls on labor costs.

With fixed annual investments, the shape of the depreciation curve doesn't much matter, and labor costs approach slightly more than 50% of the total budget.

Allocation Costs, 2019

KNL node hour: $0.14
* 440 nodes * 8.2K hours = $505K

Octal rtx2080 node hour: $1.49 ($0.186 / k80 or 2080 GPU)
* 32 nodes * 8.2K hours = $391K

TB-year of disk: $96
* 1000 TB = $96K

18 month TB of tape: $10
* 1000 TB = $10K

The total value provided by JLab this coming year: **$1,000K**, consistent with our funding stream of $1M / year.