Jefferson Lab LQCD Computing

May 2018 All Hands Meeting

Chip Watson

Scientific Computing Group

Outline

New NP Initiative at JLab

Computing Resources

File Systems, Tape & Archival Storage

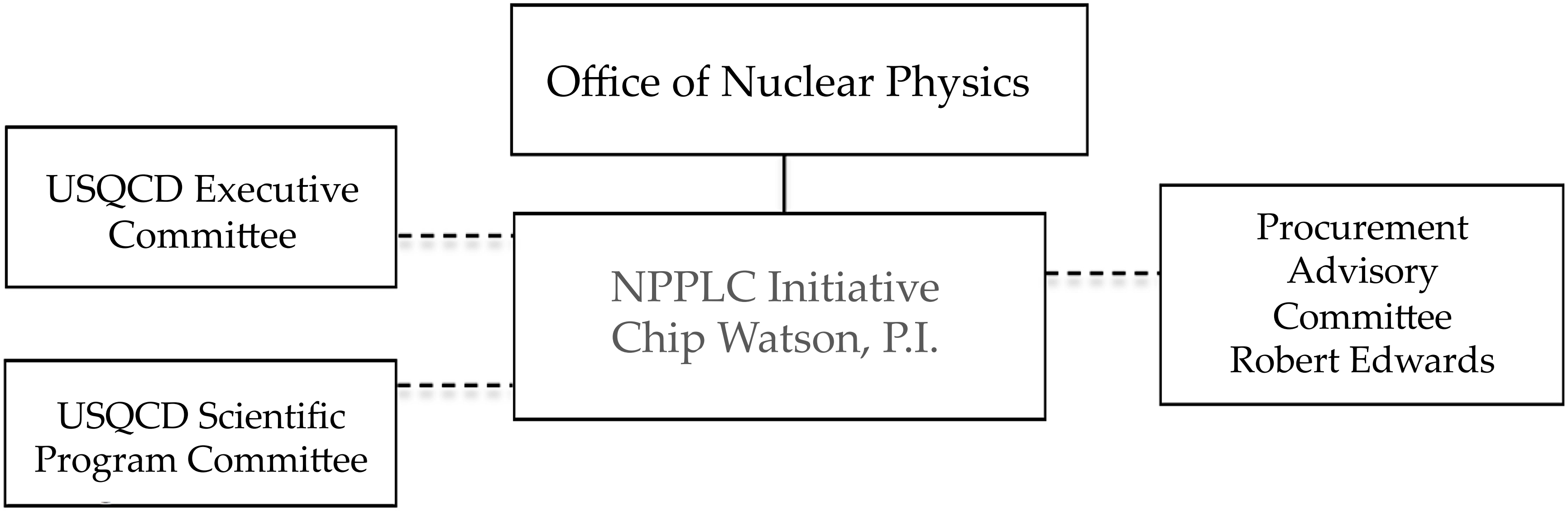
Operations

Nuclear and Particle Physics LQCD Computing Initiative

Summary

- Single lab, NP funded, serves all of USQCD, and is complementary to the modified (2 lab, IC based) HEP LQCD project
- The last time we had 2 concurrent, complementary projects was the NP ARRA LQCD project 2009-2011 (w/ LQCD-ext)
- $1M per year, about half hardware, half labor (equals average NP investment per year at JLab for last 10 years)
- Replaces, for JLab, the previous 3 lab, 2 office project (and takes over operations of existing resources at JLab)
- Focus is unchanged from the previous 3 computing projects: deploy and operate dedicated LQCD optimized resources

Management Organization Chart


Office of Nuclear Physics
USQCD Executive Committee
NPPLC Initiative
Chip Watson, P.I.
Procurement Advisory Committee
Robert Edwards
USQCD Scientific Program Committee

- Uses the same standing advisory bodies (Executive Committee, Scientific Advisory Committee) with a structure identical to the LQCD-ext II project & predecessors and the ARRA project
- Reports to the NP Office at DOE
- Like the ARRA project before it, it should be nearly invisible to the users (i.e. Jefferson Lab remains one of the sites where you do computing)

Quick win for the users: 180 new KNL nodes ! !

The KNL cluster will now have multiple partitions

16p partition: 256+4 nodes (2016)

- 32 nodes per switch, 16 uplinks to core (nominally 2:1 oversubscribed, but not really for regular grid problems)
- Xeon Phi 7230 chips, 64 cores at 1.3 GHz
- 192 GB/node, 1 TB disk

18p partitions: 4 single switch mini-clusters

- 44 or 46 nodes per switch, 2 uplinks to core
- Xeon Phi **7250** chips, **68** cores at **1.4 GHz** (faster!)
- 96 GB/node, 150 GB SSD (leaner)

Allocation charges will be identical on all partitions; choose performance, or memory + network, or use any/all nodes.

Is this the largest dedicated LQCD resource in the world?

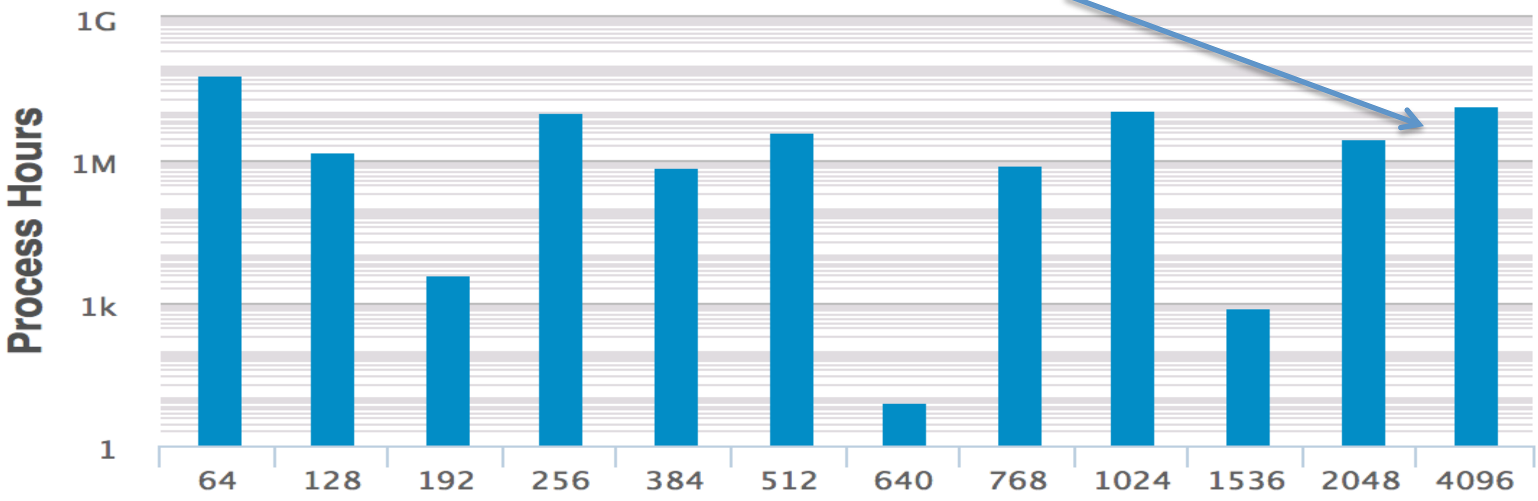
Expanding KNL Resources

Why more KNL?

- **Only option with 2x performance gain per dollar over last 18 months**
- Chosen by 100% of advisory committee as major allocation component
- Easily integrated (experienced ops team, survived initial pain)

Why smaller memory and 4 partitions?

- Memory is 2x as expensive as it was 18 months ago, nodes were much less expensive; as a percent of cost, 192 GB/node was prohibitive
- Network was expensive compared to these very inexpensive nodes
- Job mix has very little usage above 32 nodes (can run in 1st partition)



18p Timeline

- Delivery by early May
- Configuration, testing, burn-in mid to late May
- Released for production for allocations in June (if you have an old allocation, or a new July 1 allocation, you can use these nodes in June)
- Will have the most up-to-date KNL firmware, drivers and software
- 16p will be updated to this newer software in June (rolling updates to minimize downtime)

GPU Resources

Old quad k20 cluster (2012, Kepler)

- Reaches 6 years of age in November, so only 6 months of allocations are in the pool for now; we may keep it running as DNR (do not resuscitate)

FY 2019 Possibility (*selection tbd, input appreciated*)

- Quad v100 (Volta Tesla professional)
roughly 1.5x Pascal performance, 6x Kepler
- Quad GTX-1180 ??? (new gamers, online rumors)
(cheap?, but only suitable for pure inverter workloads)
- Some combination of these?
- Goal: bring a new GPU resource online before the 12k cluster retires (fewer nodes, higher total performance) to support mixed architecture workflows

Disk Resources

Today: 2 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

  Uses ZFS style, with "reserved" quota, and a maximum "quota" that we enforce via our software, but only if Lustre is "full" (over 80%)

  Note that "quota" is over subscribed 50% to allow projects to burst over their normal allocation. "reserved" is under subscribed. We adjust these as needed to keep work flows going so as to get high utilization.

- Plan to add another 0.4 PB in this calendar year for LQCD

We are considering an upgrade to a new version of Lustre this year to improve stability (high loads cause problems today), increase performance, especially for small files (to live in MDS).

New /home, /work Server

Dual Head Node (1 experimental physics, 1 LQCD), each:

- 256 GB memory
- dual SAS3 controllers
- redundant power
- dual 12-core CPU
- FDR Infiniband (56 Gbps)
- 150 GB internal SSD

Disk Arrays, each 44 slots, dual attach SAS3, redundant power

- 2 400 GB SSD (write accelerators, 1 per head node)
- 2 1200 GB SSD (read cache, 1 per head node)
- 6 raid-z2 5+2 10TB disk stripes per array
- 3 stripes for LQCD, 105 TB when @ 80% full (target)

LQCD allocation for /work is 70 TB, currently 40 TB used. This is the best place for large code builds. This server has better small file performance than our current Lustre.

There is also ample space for users' home directories.

(Please clean up old areas to keep space free for active projects.)

Tape

20 PB tape library, shared, with ~ 5 PB LQCD data

- LQCD growing at around 70 TB / month
- Soon to start data migration from LTO-4 to LTO-8 media; includes 0.5 PB of mostly NP LQCD data; please "delete" data from tape you no longer need, or migrate 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)

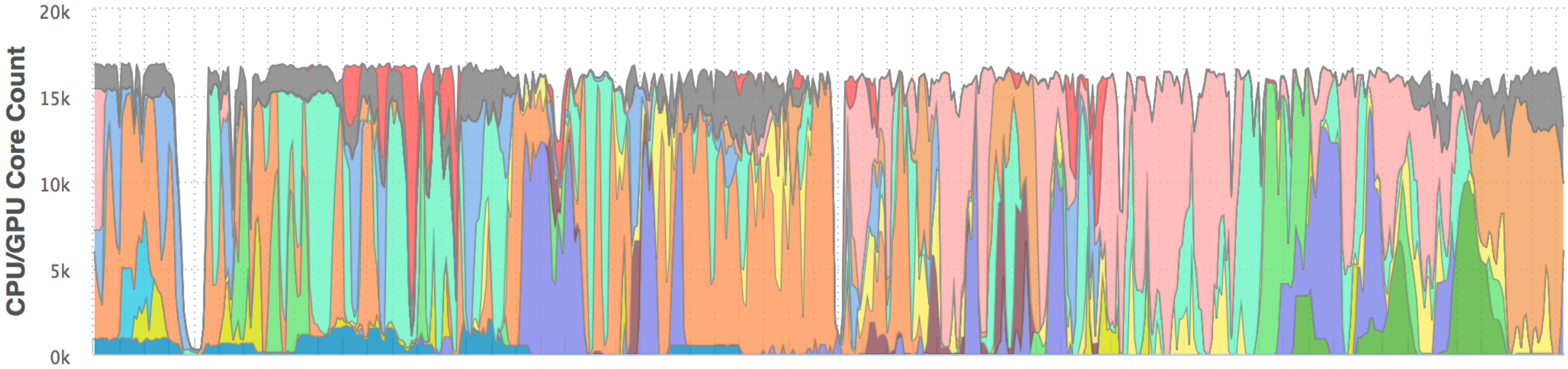
* New Policy USQCD tape allocations at JLab are NOT for archival data storage, only for medium term (18 months) storage. For archival storage, contact the laboratory or other institution owning your science scope. (JLab will host medium energy NP archive.)

Recommendation

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

Operations

- Intentionally lean shared staffing going forward
  - Total ~ 6.5 FTE, with 45% LQCD, 55% experimental physics
  - Leaves more funds for hardware
- Batch system moving to Slurm
  - Better scalability, large feature set, open source
  - Will start with the new 18p nodes, then convert 16p, then the 12k cluster (18p will not be available under PBS)
- KNL system is very popular, and is well used:



JLab LQCD Priorities and Metrics

Plain and simple, the highest priority for this initiative is getting your science done!

Key Metrics

- Delivered computing (utilization, i.e., used by you)
  - Reported to NP regularly, by project
  - Visible on our web pages so you can see it as well
- Benchmarks for systems procured, up-time delivered
  - Multi-grid inverter (dominates NP workloads)
  - Contractions (dominated by batched zgemm)
  - Systems can be evaluated by one or the other or averaged by usage
- Budget and other milestones
  (note that this is our 3rd, not 1st, priority)

Our success is tied to your success !

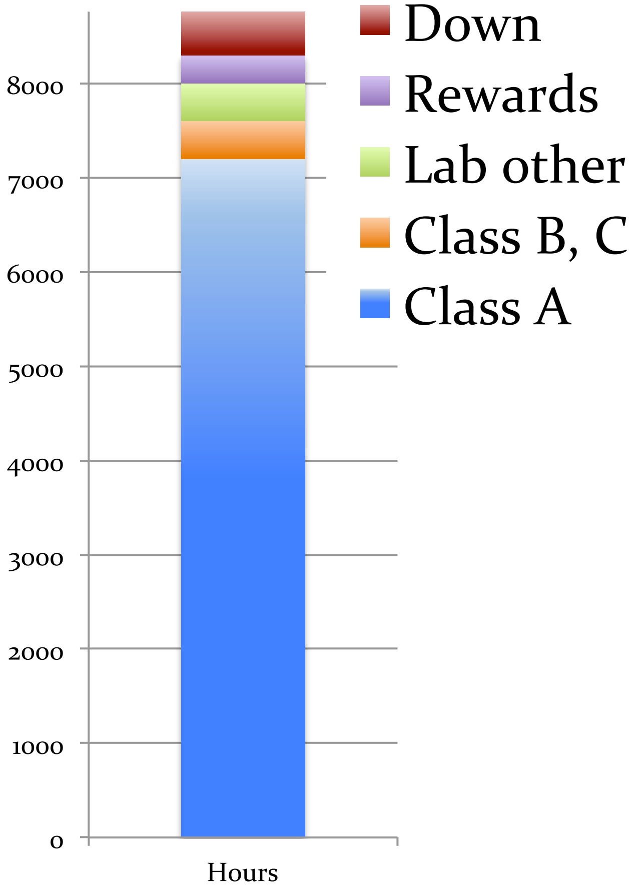
(dis-) Incentives *(proposed, likely)*

Old jeopardy system is poor

- Penalties come too late, are too small
- Fast running projects end up with no fair share in the last quarter (near annual LQCD conference)

New approach: add rewards

- Usage > 110% of pace in a month is **discounted 50%**
- Idle time if usage is below 90% in a month is charged to projects below pace
- Class A projects can never be more than 1 month behind (penalty); B: 2 months



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

Examples: Rewards & Penalties

1. One project runs at 170% of pace for 3 months, gets charged 140% of pace, netting an extra 30% for 3/12 of year => an extra 7.5% bonus for the year.
2. Another project runs at 170% of pace for 9 months, consuming 140%*9/12 > 100%. Its fair share drops 4x and it continues running at 25% of pace for the last 4 months (i.e. is not shut out). Total reward = (30*9 + 25*3)/12 = 29%.
3. A large project's software isn't ready, and they don't run for the first 3 months. First month: other groups use the time, no penalty. Second month: project gets truncated to 11/12. Third month: project gets truncated to 10/12. They would have been better off running for the 2nd and 3rd months with software only 25% efficient.

Other Rewards

Would you be interested in “earning” bonus time in exchange for improving the documentation at JLab? You are the most qualified to write things that help other users!

- Find and suggest a correction to any documentation errors
- Suggest (and write) a new page for the documentation
- Write up an FAQ, or a tip

Rewards will be non-trivial, but “budget” constrained. Productive writers will be given “commit” privileges.

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)

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

These costs were used to assign costs to hypothetical allocations, where the procurement advisory committee members were asked to each "spend" $100K on allocations.

Allocation Costs

KNL node hour: $0.20   (first year, new price, incl. labor)
   * 440 nodes * 7.6K hours = $669K

Quad k20m GPU node hour: $0.50
   * 45 nodes * 7.6K hours = $171K

TB-year of disk: $90
   * 800 TB = $72K

18 month TB of tape: $20
   * 800 TB = $16K

The total value provided by JLab this year: **$1,038K**,
consistent with our funding stream of $1M / year.
(We don't run an IC, but we can generate equivalent costs.)

Science per Dollar FAQs

Jpsi based Allocations

1 SkyLake hour costs the same as 1 KNL hour (192 Jpsi/node)

$$$ comparison

1 SkyLake hour costs >3x 1 KNL hour (purchase & operate)

1. If you ignore $$$, most prefer SkyLake (easier to use), but if you pay the $$$, most prefer KNL.

   What factor (in your allocation) are you willing to give up to get "easier to use"? Should we be doing allocations in $$$ or Jpsi ?

2. "Value" of a node is <= replacement cost of its performance, not whatever you paid for it in the past.

   16p nodes are now "de-valued" to the 18p cost, falling only a little faster than normal depreciation would yield.