Scientific Computing for FY2019-23

Scientific Computing Overview

System Architecture and Design Principles for Scientific Computing for the next 3 to 5 years.

Chip Watson

Head of Scientific Computing

Nov 27-28, 2018

Jefferson Lab

U.S. DEPARTMENT OF ENERGY | Office of Science

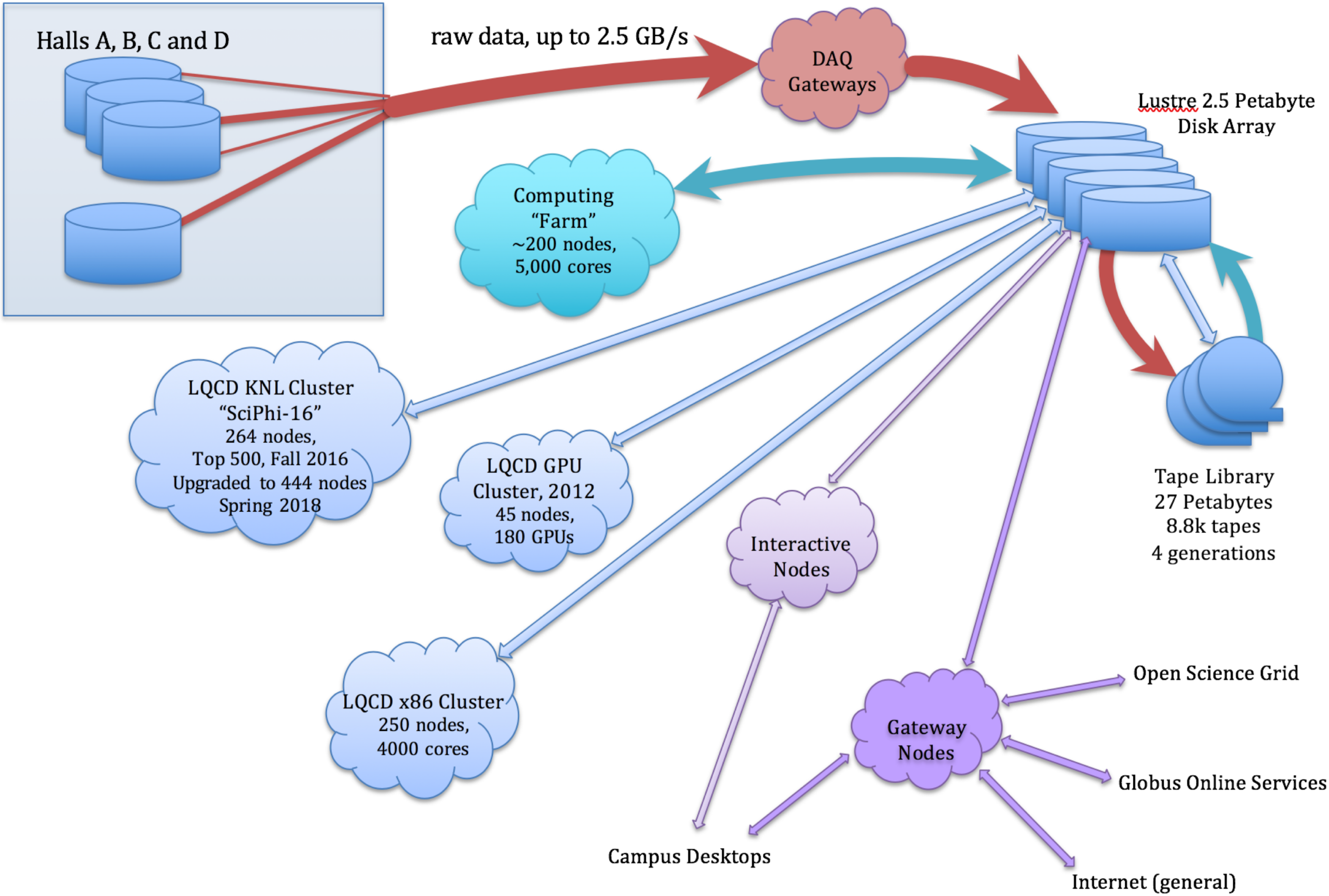
JSA

Outline

- Scientific Computing Scope and Overview
- System Design Principles
- Current Architecture
- Capacity Planning: trend lines + requirements gathering, including known equipment upgrades, anticipated lifetime of equipment
- Near Term Evolution 2019-2020
- Trends:

  Larger Swings in Load,
  Growing Offsite Computing

Scientific Computing


Halls A, B, C and D
raw data, up to 2.5 GB/s
DAQ Gateways
Lustre 2.5 Petabyte Disk Array
Computing "Farm" ~200 nodes, 5,000 cores
LQCD KNL Cluster "SciPhi-16" 264 nodes, Top 500, Fall 2016 Upgraded to 444 nodes Spring 2018
LQCD GPU Cluster, 2012 45 nodes, 180 GPUs
LQCD x86 Cluster 250 nodes, 4000 cores
Interactive Nodes
Gateway Nodes
Tape Library 27 Petabytes 8.8k tapes 4 generations
Open Science Grid
Globus Online Services
Internet (general)
Campus Desktops

Scientific Computing Scope: Programmatic View

- LQCD NPPLCI (NP Initiative, mid range HPC)
  - Currently year 2 of 4 year plan, $1M / year, ~half hardware
  - Hardware details in Sandy Philpott's talk
- Experimental Physics (high throughput computing)
  - Of scale $1.2M / year (loaded), 1/3 hardware (more labor for 10x users, DAQ data flows, offsite data flows)
- Accelerator (smaller scale HPC)
  - Contribute to HPC hardware, gets % of node-hours/year
  - Can use both ENP and LQCD resources, at small scale

"SciPhi XVI" 2016 Top500 #397 Green500 #10

Xeon Phi 7230, 64 cores/node, 264 nodes

192 GB memory, plus

16 GB high b/w memory

100 Gbps Intel OmniPath fabric

1 TB disk (O/S plus scratch)

Spring 2018 added 180 nodes

68 cores/node

96 GB memory

200 GB SSD + 1 TB disk

LQCD is very data parallel, and can exploit advanced architectures such as Xeon Phi and GPUs.

In contrast, x86 clusters are no longer cost effective for the current suite of NP applications.

Consequently, it is not as cost effective for ENP and LQCD to share common hardware.

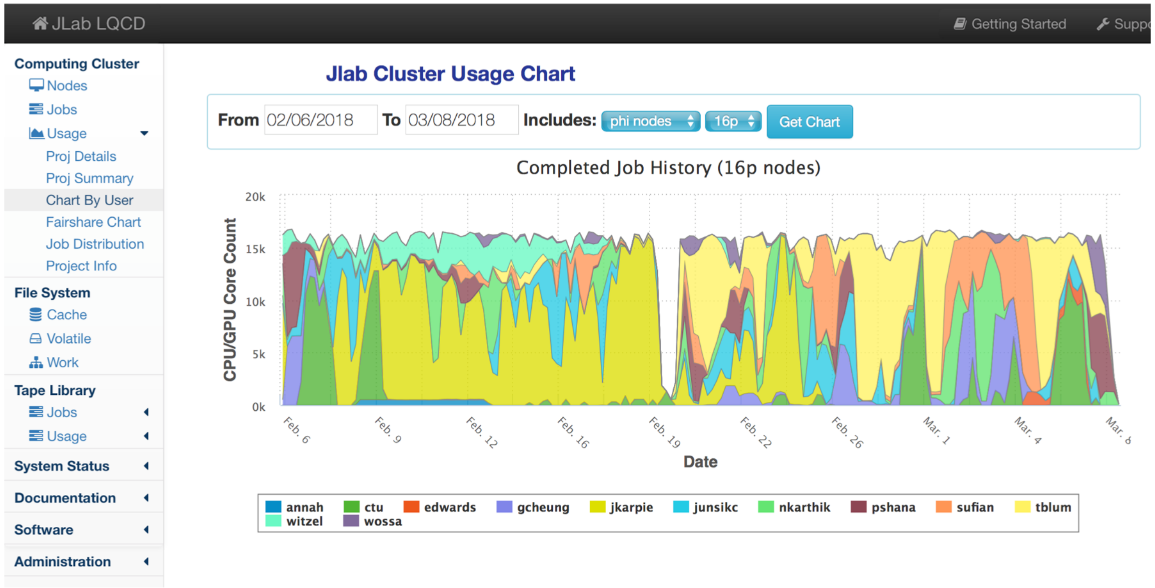
Scientific Computing Scope: Activities View

- Operations
  - Run hardware
  - Support users
  - See Sandy's talk
- System Software Development
  - Tools for users to facilitate using the system
  - Tools for operating the hardware systems
  - Examples below
- LQCD Software Development
  - Part of SciDAC and Exascale Computing Projects
- Planning, Budget and Management, Architecture & Procurement

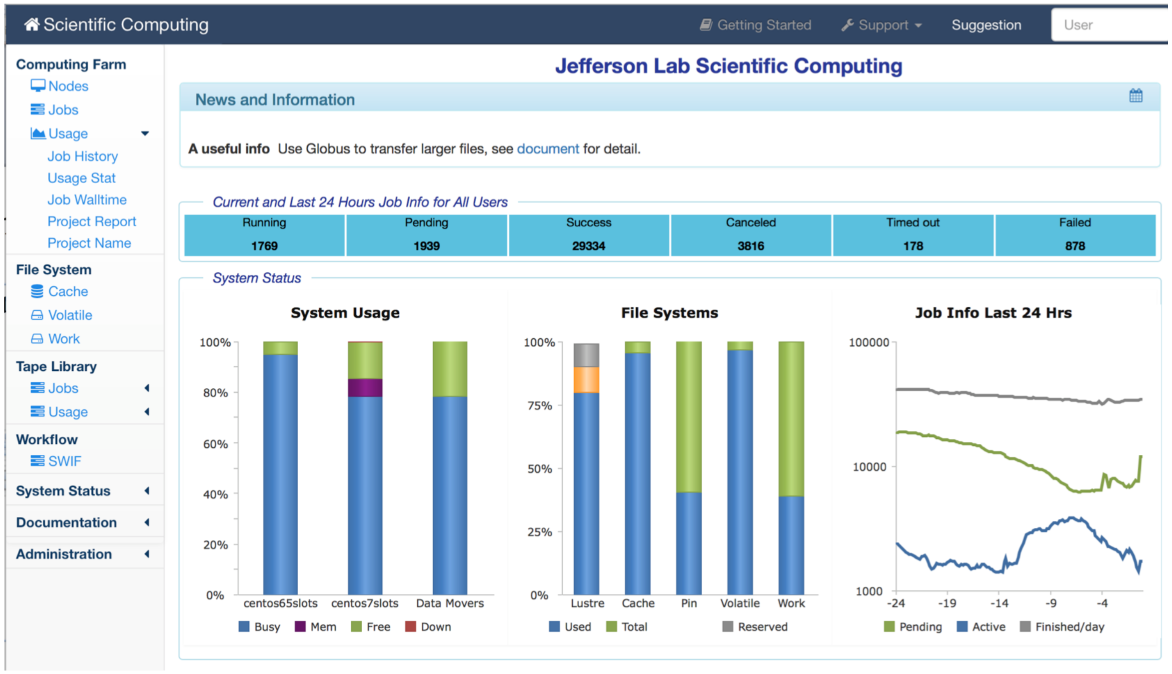
Integration Software

Scientific Computing has two staff working on integration software

- Web apps to view system status (right & next page)
  http://lqcd.jlab.org/
  http://scicomp.jlab.org/
- Custom tape library software
- Custom disk management and file migration software (poor man's HSM)
- (Underway) Remote compute integration / bursting to offsite center / cloud



From 02/06/2018 To 03/08/2018 Includes: phi nodes 16p



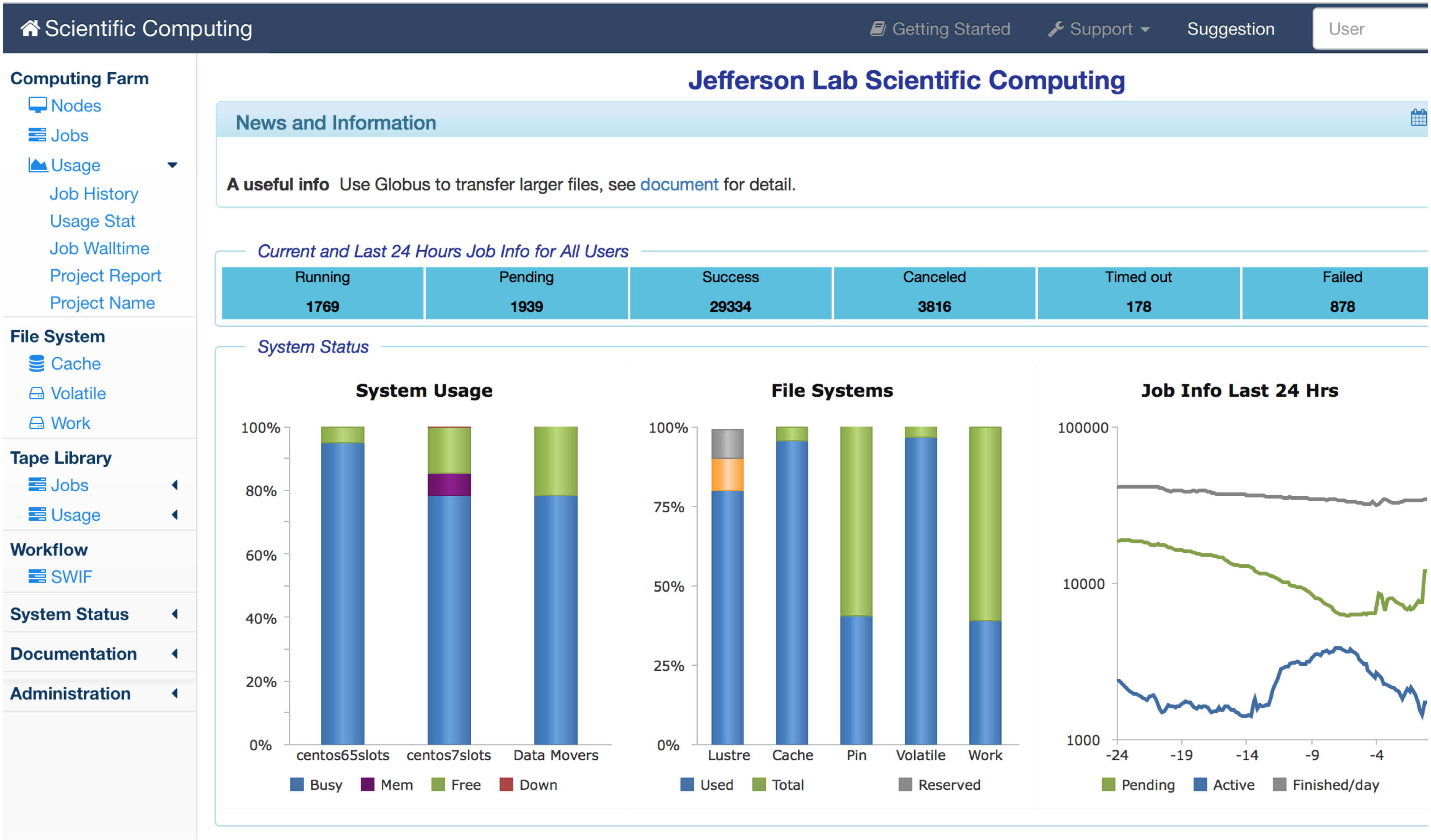
A useful info Use Globus to transfer larger files, see document for detail.

Current and Last 24 Hours Job Info for All Users

Running	Pending	Success	Canceled	Timed out	Failed
1769	1939	29334	3816	178	878

System Status

Scientific Computing Web App



Jefferson Lab Scientific Computing

News and Information

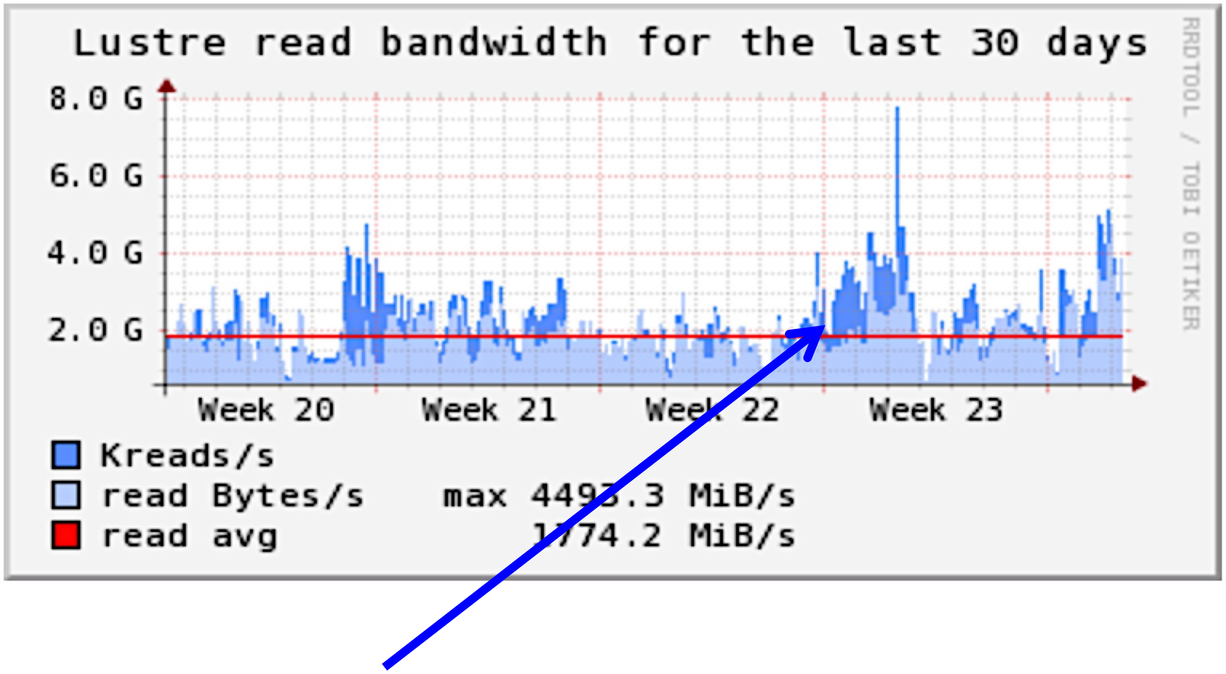
A useful info Use Globus to transfer larger files, see document for detail.

Current and Last 24 Hours Job Info for All Users

Running	Pending	Success	Canceled	Timed out	Failed
1769	1939	29334	3816	178	878

System Status

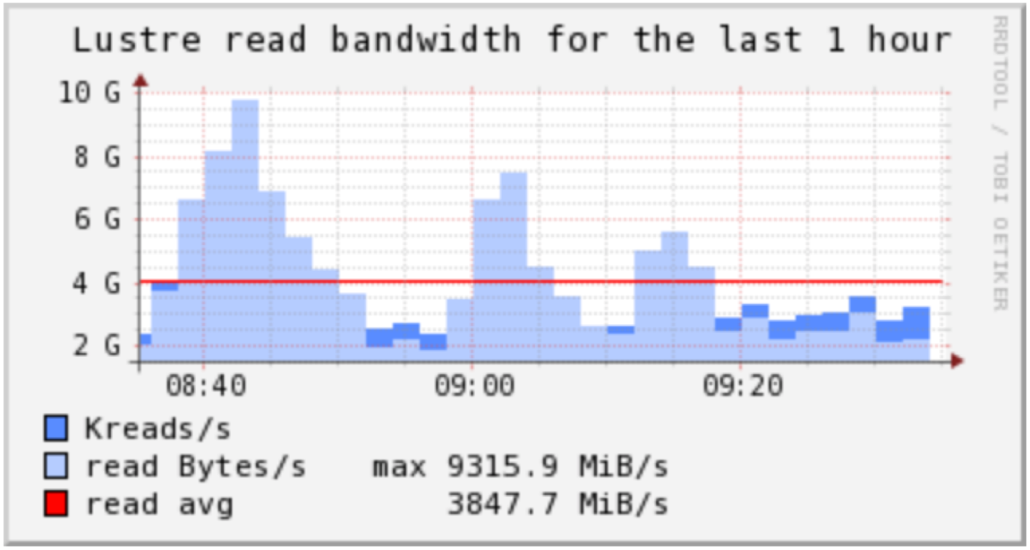
Detecting Common Performance Issues

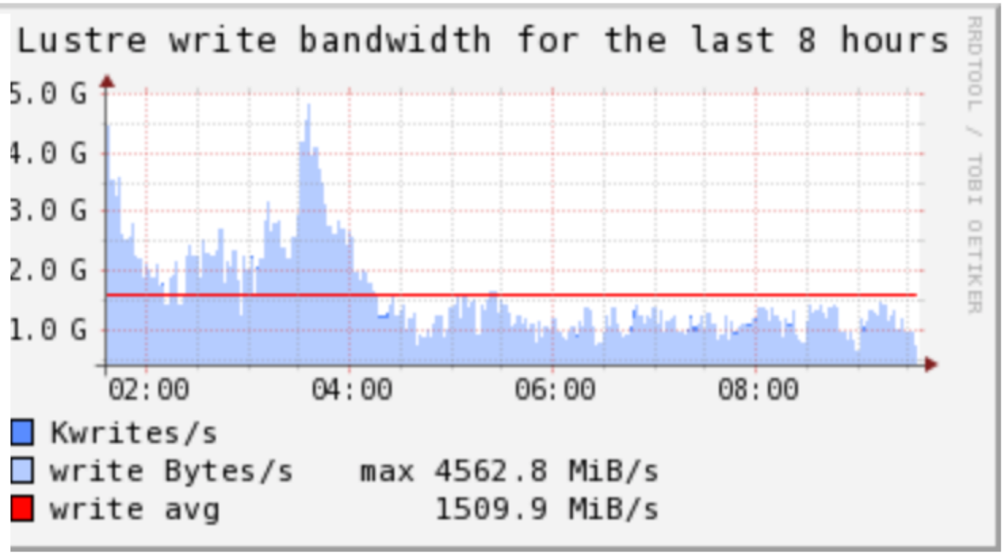


When we see dark blue, we suspect an application is doing I/O poorly: reading in small chunks instead of big chunks.

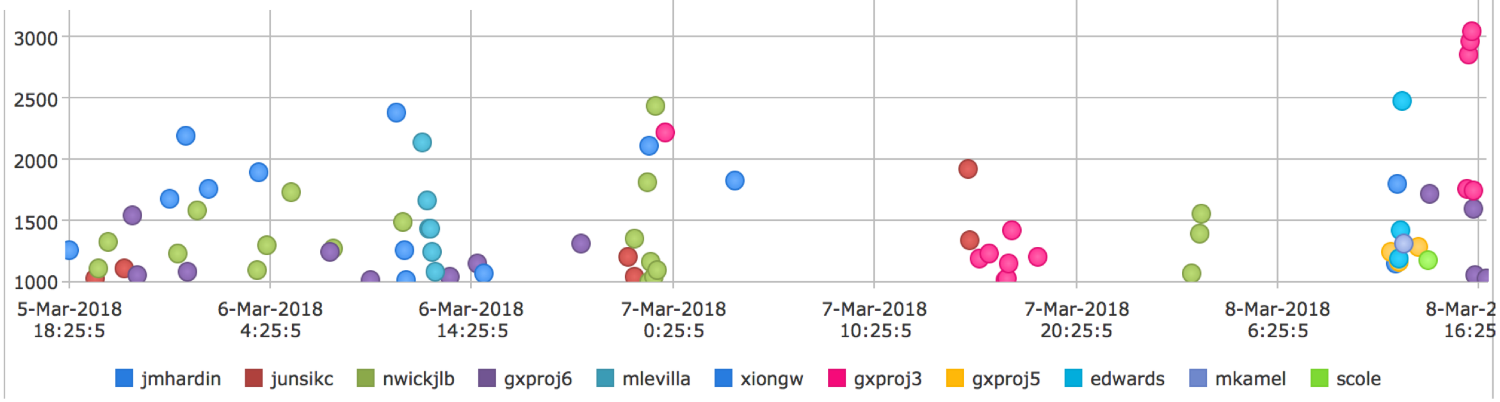
Our servers are configured for streaming large files, not handling thousands of small I/O requests. Good size: 1MB / read or write

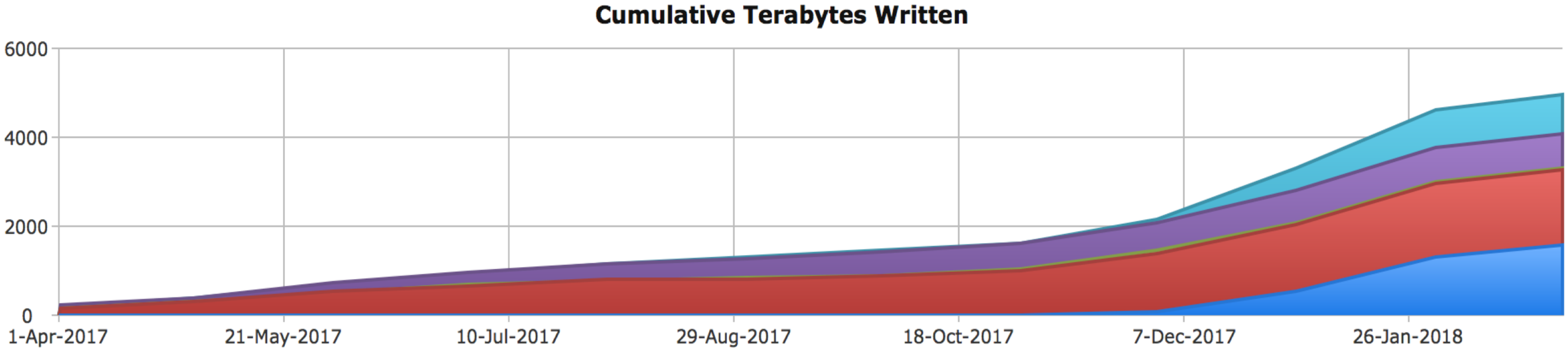
Big brother is watching!





Extensive Monitoring Software





Scientific Computing Design Principles

- **Optimize Science**, not just computing design
- Balanced Design
  - Hardware: balance compute, online storage, offline storage, bandwidth
  - Labor vs Hardware, open source vs. licenses, simple clusters vs. complex grid, cloud, etc. We take into account the full costs of the system and optimize for science / dollar
  - Stay off bleeding edge to control labor costs, unless gain is worth the pain (optimization problem, edge often needed and used for LQCD)
  - Attributes: ease of use (both for users and sys admins) vs. capacity
- Synergy

  Combined procurements and systems for Theory & Experiment, both to get better pricing, to lower labor costs (e.g. shared Lustre system) and to improve utilization (details below)
- Requirements driven, not reactive

  (somewhat aspirational, getting better year by year)

Current Architecture (background)

Past Context:

For the 6 GeV Era (the past 20 years), Experimental Physics Computing pretty much lived in 4-6 racks, plus another few for disk and for tape I/O. It could be managed by <3 people, total.

Theory computing (LQCD) was ~4x the computing footprint.

LQCD influenced the current Experimental Physics architecture in two ways:

1. passed along “end of life” hardware for free
2. enabled the adoption of higher end technology (recycled lower speed Infiniband, Lustre file system)
3. pushed towards more open source choices (especially where licensed software, e.g. batch software, was too expensive for LQCD at their scale)

Current Architecture (1)

- Multiple clusters, nodes interconnected via Infiniband, with uplinks to core file servers and other services via multiple uplinks
  - Fast network allows LQCD jobs to run on ENP hardware if they aren't using the nodes (LQCD has covered almost all of this cost by allowing old network fabric to be reused)
  - Slight exception: LQCD KNL cluster nodes are on Omnipath for higher bandwidth, with routers (Lustre, TCP/IP) to Infiniband
- 2 instances of the batch system (LQCD-HPC, ENP)
  - Batch system parameters are different enough to continue this, although ENP is becoming more like HPC (whole nodes instead of serial single core applications)
- Single shared large file systems (tape, spinning disk)
  - Impact: one side (more often ENP) can use more than their "share" of the bandwidth if they suddenly start many jobs at once

Current Architecture (2)

- Fault tolerance only where needed (science optimization)
  - Only non-compute systems are generator backed w/ dual power to rack
  - Lustre Meta Data Server is dual head active-passive, auto failover (our most robust component, failure impacts everything in the room)
- Redundancy, scale out capacity for most services needing high availability (cheap, does the job)
  - E.g. if one tape drive or its computer fails, that is only a loss of a fraction of capability; everything can still continue using other drives
  - File servers are mostly configured as active-active pairs, operator can flip load away from a failed head to the other head (not automated)
- Predominantly "central computing" but growing in the use of offsite resources to support burst computing
  - In 6 GeV era, cost of labor to support offsite computing exceeded the useful value of the offsite computing

(More details in Sandy's talk)

Forward Planning

Process

- Gather requirements (updates of multi-year estimates by halls)

  Estimates have not been very accurate during the transition from 6 GeV to 12 GeV program, but halls are maturing in their processes

- Usage trend analysis

  See how frequently subsystems are pushed to their limits (cpu, disk, tape I/O, WAN bandwidth)

- Adapt to budget pressures (as science is optimized at an even higher level: detectors vs. accelerator vs. computing)
- Track technologies

  Understand what is available, and what it costs and possibly how those costs are evolving

Based upon all of this data, year by year or even half-year, optimize what can be purchased for the funds available.

Hardware lifetime

1. Retire nodes when they are either too troublesome (labor costs money) or consume too much power per flop (electricity costs the lab money, even if that cost isn't in my budget).

   Over the last 5 years, the reasonable lifetime of a node due to power has grown from 4 yrs. to 6-7 yrs. due to slowing down of Moore's Law.

2. Even if a rack of nodes become troublesome for LQCD multi-node jobs, it is often fine for single node ENP workloads.

3. This year LQCD is better off buying new KNL and GPUs than paying labor to run a 6 year old LQCD x86 cluster, so ENP gets it for the cost of operations, with LQCD still using it when ENP doesn't (win-win). This is more cost effective than all alternatives for ENP provisioning for major campaigns.

4. File servers, however, are only good for 4-5 years due to increasing loads – but we'll re-use the four 2014 file servers as a special cache for the tape library (where failures can be more easily tolerated).

FY2019 Planning (1)

- As Graham described, computing projections fell significantly this year (CLAS-12 track reconstructing plus both halls' simulation needs)
- While the current resources are of adequate size (including the 20% coming from LQCD), that resource and a small amount of the older ENP compute nodes are end of life. So either in FY2019 or early FY2020, $300K will need to go into maintaining the current throughput
- We are currently waiting to see how much growth we get in NERSC cycles as CLAS-12's request is added to GlueX's
- Focus has shifted towards early analysis, thus a desire to keep more data on disk for rapid processing
- Impact: instead of spending only 15% on disk and 85% on cpu, we will spend ~50% on disk subsystems and 50% on compute for this year (and probably return to "normal" next year)

2019 Detailed Plan

- New Lustre MetaData Server
  - Shared cost with LQCD and SciComp, but mostly born by ENP as they will drive the capacity and performance requirements
- Expansion of storage capacity
  - Add ~1.4 PB while retiring 0.3 PB
  - Grows quotas from 0.8 PB to 1.9 PB
- Expansion of SSD fast file system
  - Add 2nd head to flash storage
    (used as raw data buffer on way to tape)
  - Expand current 25 TB capacity to perhaps 50 TB
- Replacement of aging nodes
  - Keep 2012 nodes running until we choose between AMD Rome and Intel's best contemporary Xeon in price/performance
  - Replacement and potentially some growth can be done in early FY2020

2019 Detailed Plan (cont)

- Expansion of bandwidth to the tape library
  - Add 8 more LTO-8 drives to a total 16
    - Each provides 300 MB/s for LTO-M8 media (variant of LTO-7), or 360 MB/s for LTO-8 media
  - Retire 4 LTO-5 drives as their use in writing raw-duplicates and migrating old data to newer tapes is complete
- Double lab WAN bandwidth from single to dual 10g
  - Supports simultaneous use of offline computing for GlueX and CLAS-12, NERSC and OSG
  - Matches growth in disk storage and tape bandwidth
- At scale tests of using Amazon cloud
  - Infrastructure as a service, so managed by our slurm instance
  - Data buffer in the cloud to minimize job's time waiting for data
  - Results will be used for out-year planning

Near Term Evolution (1)

Major initiative now in progress to support larger swings in computing load

- History: 2014-2017 used LQCD as a flywheel (this LQCD resource now given to ENP at end of life)
- LQCD has moved on (mostly) to more advanced architectures, harder to share their newer resources (so far)
- GlueX has been using OSG for most of its simulation work, using OSG tools (tools not really Jlab supported – labor constrained)
- GlueX is now also using NERSC as an offsite resource via a JLab workflow tool derived from the one used for much of local computing
- In 2019 we will prototype bursting to a cloud (not production scale, but enough to learn how to scale it up later)
- To support above changes, Jlab will double WAN to 2*10g this winter to better support NERSC + OSG usage, 100g in 2020

Integrating Offsite Computing

Drivers:

- Experimental physics' peaks and valleys of demand are becoming larger: sharing with LQCD is no longer a large enough flywheel to smooth out load variations, especially as HPC nodes currently no longer a good match
- Provisioning to peaks is expensive (idle time wastes money)

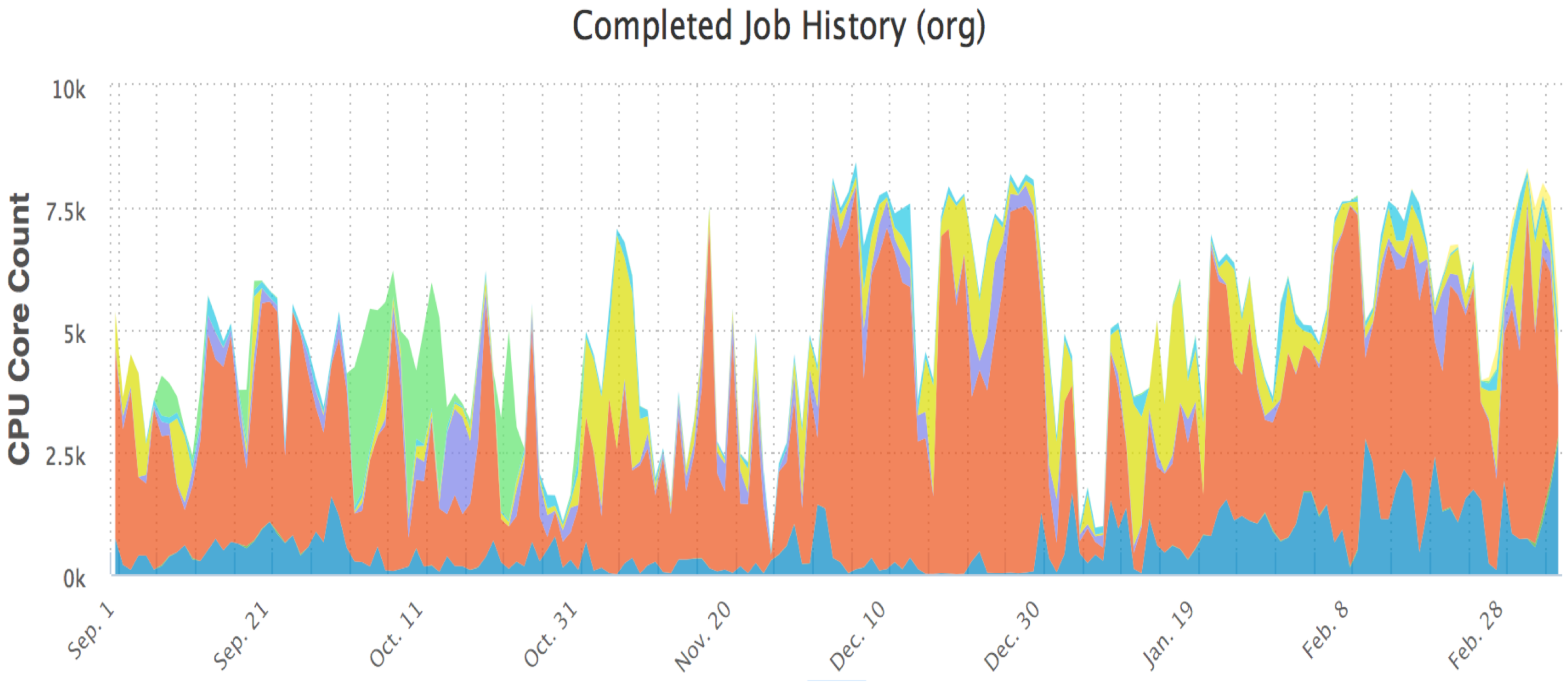
Options:

- Send jobs to OSG, NERSC, Supercomputer Centers, Cloud, …

Considerations:

- Procurement costs for cloud, learning curve for users & for operations, integration costs, etc.
- Wide Area Networking bandwidth constraints (today) will vanish in 2020 when ESNet upgrades us to 100g links

Peaks and Dips



When beam is on, load is high (last Fall, Winter).

Going forward, demand peaks will increase 4x, resources 2x.

Near Term Evolution (2)

1. Support higher peaks on local resources
   - Moving 250 nodes of 2012 LQCD nodes to join 250 mostly newer ‘farm’ nodes, separate from the national LQCD resources. Also moving 42 quad GPU nodes to the ENP slurm system as LQCD purchases a new GPU cluster (2019 Q1)
   - Experimentalists gain easy access to a large system capable of training neural nets, or doing science calculations
   - LQCD can fill x86 and GPU nodes when they are underutilized (reduced waste of cycles), but experimental physics gets more than they paid for during their major campaigns (win-win)

End result: much lower underutilization of JLab resources

Near Term Evolution (3)

- Since NERSC allocations are annual, we may continue a pattern where we request NERSC time each year, and if we get all we need, we prioritize other things (mid-year optimization)
- If NERSC allocations are overly constrained, we expand local resources in the summer, and/or purchase more time in the cloud
- Note: the next NERSC machine (2020) will be mostly GPU accelerated, so we cannot presume NERSC will be able to provide the growth in computing we will need

2. Explore bursting to the Cloud

- *infrastructure-as-a-service* to dynamically increase the size of this local resource (in budget for FY2019)

Current Trend Line Extrapolating Out 3-5 Years

- Continue annual investments in
  - Disk, archival storage capacity and performance
  - Local computing
  - Growing offsite computing
  - Cloud for bursting
- Identify more opportunities for software to improve productivity
- Track new technology
  - AMD processors (Epyc Rome, more to come)
  - ARM processors (making good progress, but not yet ready for us)
  - Advanced memory architectures

(LQCD helps to keep us on the bleeding edge of computing hardware, and to apply many new ideas to ENP needs as they show potential)