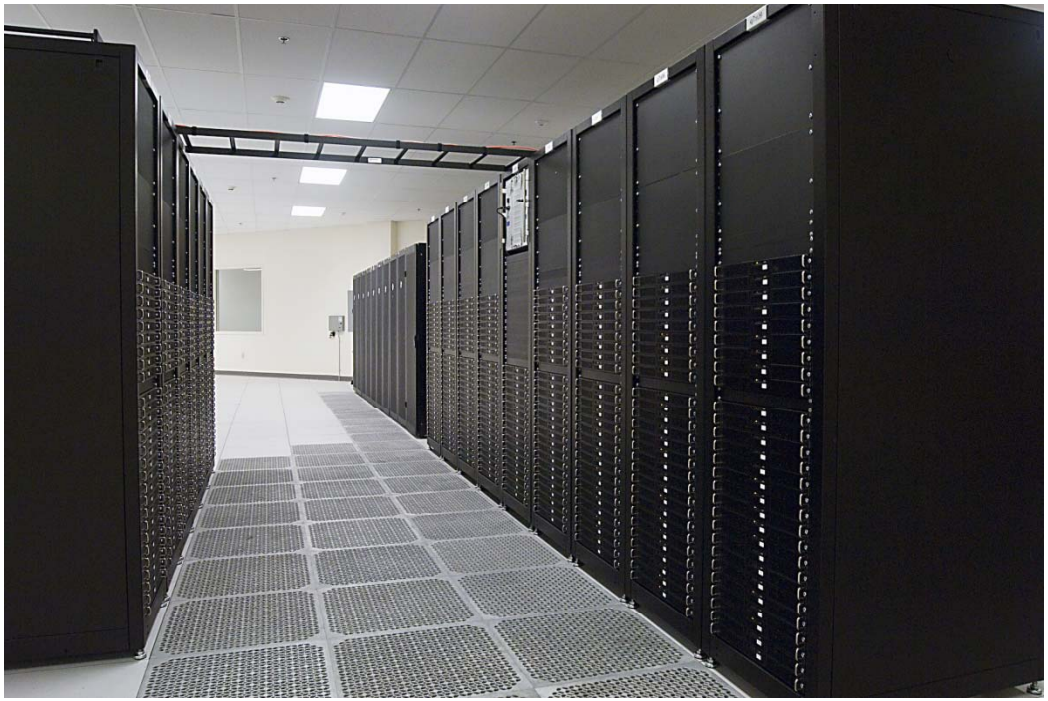
LQCD Facilities at Jefferson Lab

Chip Watson

May 14, 2009

Existing Clusters

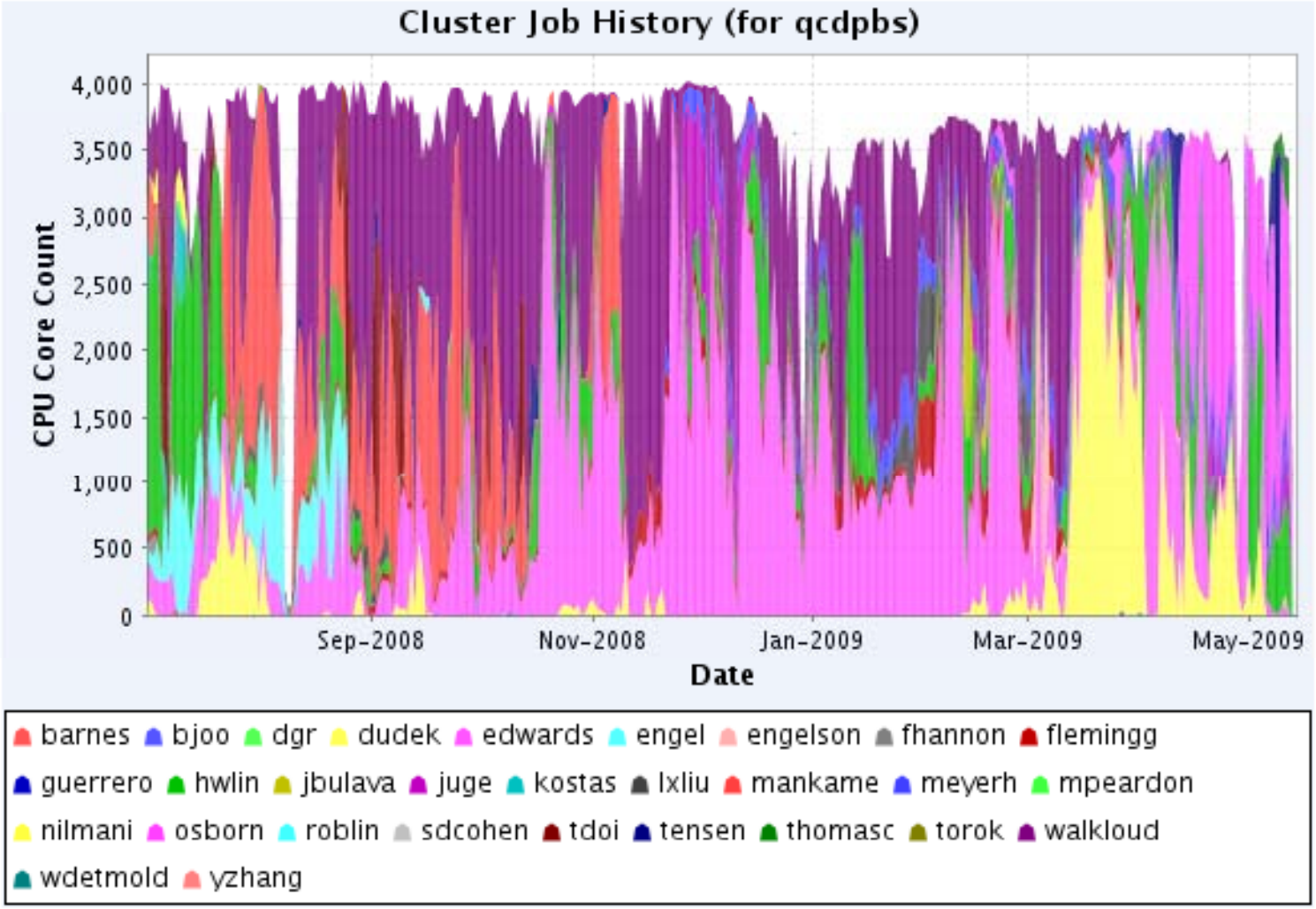
6n 200**6** i**n**finiband
3.0 GHz Pentium-D
1 GB, 0.5 GB/core
256 nodes, 512 cores
Single data rate IB



7n 2007 i**n**finiband
2.0 GHz Opteron
8 GB mem, 1 GB/core
396 nodes, 3168 cores
Double data rate IB

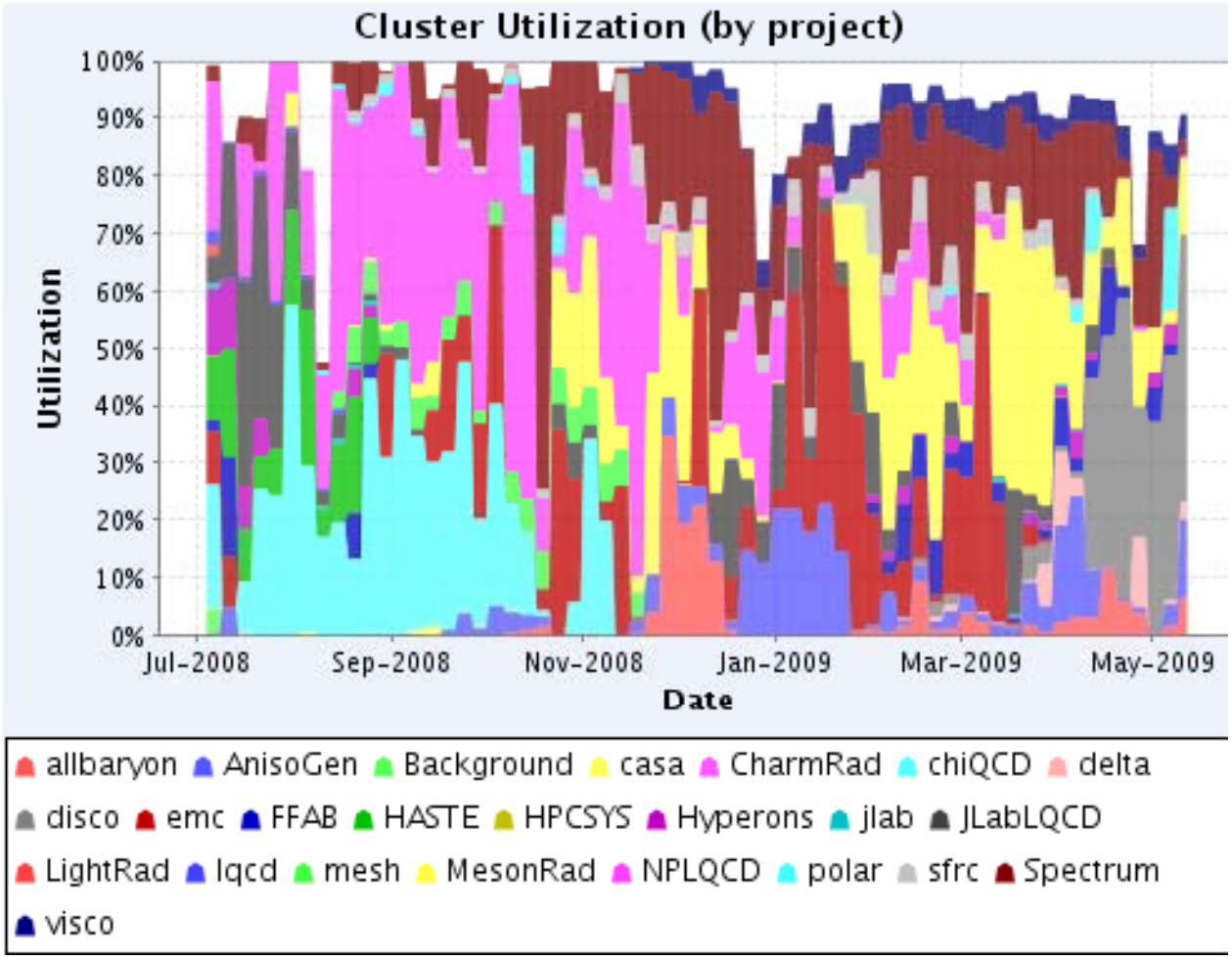
10 month Utilization

From: 2008-7-1 (yyyy-mm-dd) **To:** 2009-05-14 (yyyy-mm-dd) **For:** all

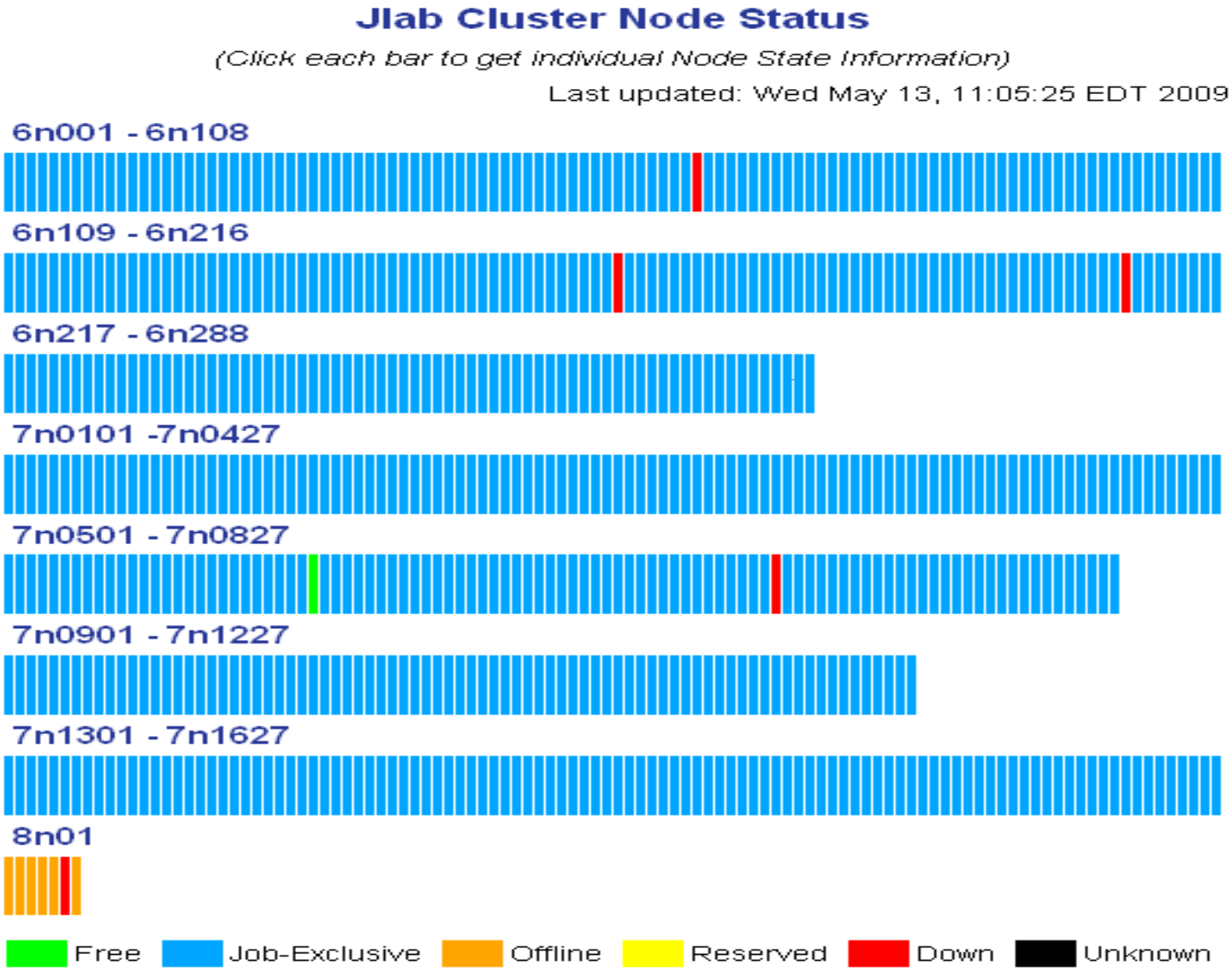


Utilization by Project

From: 2008-7-1 (yyyy-mm-dd) **To:** 2009-05-13 (y



Improved “Nodes Up” ~99%



Operations

Fair share:

- Usage is controlled via Maui “fair share” based on allocations
- Fairshare is adjusted ~monthly, based upon remaining time
- Maui fairshare bug, which divides unused fairshare equally instead of proportional to active account fairshares, was fixed last month.

Disk Space:

- Increased by 67% during the year
- Was tight for much of the year, now releasing additional space to remaining active users
- Space can be user managed, or cache managed (write through cache, with deletion of oldest), at user’s request

LQCD ARRA Proposal

NP requested from JLab a set of proposals for "ready to fund" projects. Included in JLab's mix was a proposal to "forward fund" the entire 5 year $23M LQCD-II national facilities proposal.

With other input and deliberations, NP and HEP decided to keep the LQCD extension in the same shape to which it had evolved (a $17M project extension, not a new project, perhaps 2:1 HEP:NP).

NP then chose to fund a separate LQCD ARRA activity of approximately $5M. (This figure will be reduced by one or more "taxes" of up to 10%).

Good news: NP is now full partner in LQCD, ~1:1 HEP:NP

JLab was selected as the site as it was next in line for a deployment. Note: this has resulted in adjustments to the LQCD-ext project.

This new funding is intended to operate seamlessly as a USQCD resource, using the same allocation process as for the LQCD-ext project extension.

Project Highlights

1. Project budget in round numbers (assuming $4.5M):
   - $3M for a cluster
   - $¼M for disk servers
   - $1¼M for deployment and 4 years of operations

2. LQCD ARRA is a separate project, at Jefferson Lab, with Chip Watson as project manager. Assistance in ARRA specific reporting will be provided by a dedicated ARRA staff at the lab (JLab also received considerable 12 GeV upgrade ARRA funding plus other facilities improvements, total $80M.)

Cluster Expectations: Performance

Intel Nehalem dual socket, quad-core

- 2.66 GHz or 2.8 GHz (lowest cost for fastest memory)
- Each CPU has three memory controllers, DDR3-1333
  - Bandwidth (peak) 25 GB/s per CPU
  - 24 GB planned node memory size (now multiples of 3)
- Cores are hyper-threaded, yielding 10% gain on some codes (appears as 16 cores per box)
- Total performance expected ~15 Tflop/s

Early Benchmarks

JLab early cluster

- 15 nodes 2.66 GHz in-house, with QDR infiniband (one more node coming to allow 16 node running)

For 8x8x8x16: (comparable to cache size)

- 30 Gflop/s single node
- 53 Gflop/s on 2 node, 32 core (hyperthreading on) Chroma run, anisotropic clover, no special tuning
- 160 Gflop/s on 8 node 64 core (hyperthreading off) (not sure how many dims of communication in this)

- Production sized lattices already show 20-23 Gflop/s per node with no special optimizations yet

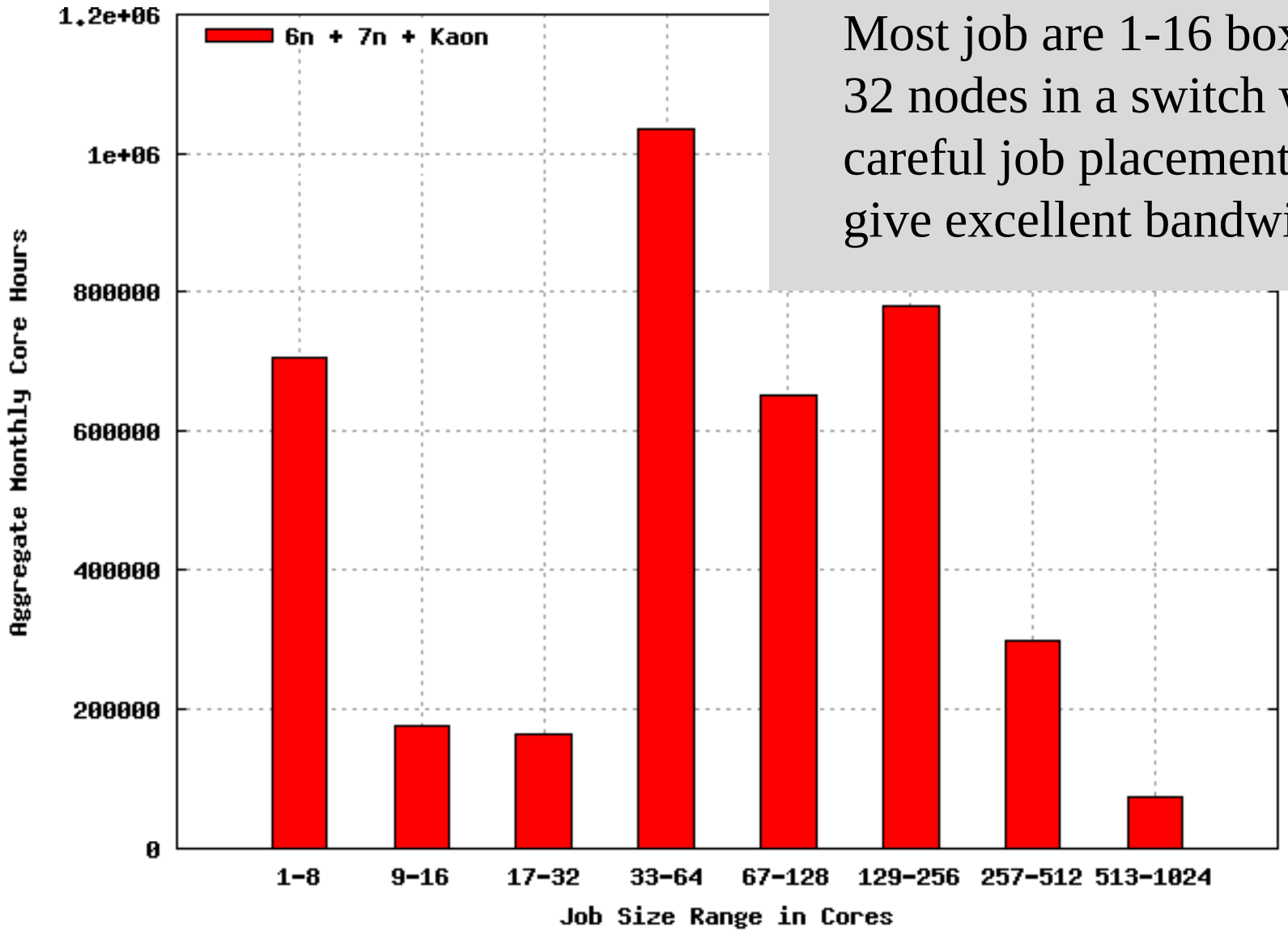
Network Options

Quad Data Rate Infiniband (QDR), 40 Gb/s full duplex

Network Topology Options:

1. 2:1 over subscription, leaf & spine:
   24 nodes per 36 port switch (network is 30% of cost)
2. High over subscription leaf & spine:
   32 nodes per 36 port switch (network is 20% of cost)
3. Mixed:
   Some nodes at 24/switch, 12 uplinks (or big switch)
   Some at 32/switch, 4 uplinks
   Some with no infiniband, dual gigE for file services?
   (network might be 15% of cost?)

Jlab 6n+7n, FNAL Kaon 2008 Job Statistics



Most job are 1-16 boxes, so 32 nodes in a switch with careful job placement would give excellent bandwidth

Discussion Questions

1. Is 24 GB memory per node correct for next few years?
2. Would going down to 12 GB / node be right for some fraction of the nodes – those with low over subscription intended for large jobs (i.e. offset higher network cost with lower memory cost)?
3. If going from 2.66 GHz to 2.8 GHz were to yield 4% gain for 8% cost, would this still be worthwhile if going from 1 node to N nodes were to cost 10%?
4. Does ~20 TB disk per Tflop/s sound about right?

Opinions invited now, and for the next few months!

Disruptive Technology -- GPGPUs

Are GPGPU's reaching the state where one could consider allocating funds this Fall to this disruptive technology?

Probably the answer is "maybe" and "at some scale"…

Integrated node+dual GPU might cost twice as much, and yield 3x performance of two nodes on inverters = 50% gain

Challenges

- Amdahl's law: impact being watered down by fraction of time the GPGPU does nothing
- Software development: currently non-trivial

Using 20% of funds in this way could yield 10% overall gain. Is this too small to bother, or one more good idea?

Disk & Tape

On project:

- ~300 Tbytes of disk
- Servers will be on the new infiniband fabric
- Lustre will be evaluated (likely choice? will learn from FNAL)

JLab contribution:

- Expansion of existing tape library (more slots, more drives)

USQCD / LQCD-ext:

- tape cost funded by LQCD-ext operations

Time Table for ARRA Machine

- June 2009 – issue RFI for cluster, file servers
- August 2009 – issue RFP (after backlog relaxes on Nehalems)
- Sept 2009 – award 50% of cluster, 100% of file servers; option on 2nd 50% for early FY2010
- Nov/Dec 2009 – award second half
- Nov/Dec 2009 – early use on first half
- Jan 2010 – production use on first half
- Mar 2010 – production running on full machine

Dates are high level milestones, and we will work to deploy and release to operations faster than this if no problems are encountered.

Summary

USQCD resources

- 80% - 90% increase in dedicated computing capacity

At JLab

- 5x increase in performance
- 5x increase in disk capacity
- less than 2x increase in staff (i.e. still lean)

QUESTIONS ?