LQCD ARRA Computing Project

Final Report

Chip Watson
Jefferson Lab Scientific Computing Group

May 9, 2013

Outline

- Project Scope and Context
- Performance Goals and Evolution to include GPUs
- Technical Performance
- Management
- Budget and Budget Performance

The ARRA LQCD Project

The LQCD ARRA Computing Project directly supports the mission of the DOE's Nuclear Physics Program "to foster fundamental research in nuclear physics that will provide new insights and advance our knowledge on the nature of matter and energy...".

The Project also supports the Scientific Strategic Goal within the DOE Strategic Plan to "Provide world-class scientific research capacity needed to: advance the frontiers of knowledge in physical sciences...[and] provide world-class research facilities for the Nation's science enterprise."

The project scope, management structure, and milestones are defined in the Project Execution Plan, a 17 page document submitted after award, and amended very early in the project to incorporate the evolution in the plans to exploit GPUs.

The technical goals included deploying "resources capable of an aggregate of at least 60 Teraflops of performance sustained in key LQCD kernels" and delivering an integrated performance of 180 TFlops-years, initially planned as 3¼ years of operations after deployment.

ARRA Project Context

The LQCD ARRA project was complementary to the LQCD-ext project

- In 2009 USQCD collaboration requested $24M for "LQCD 2" (5 years)
  - LQCD-ext was funded at $18M
  - LQCD ARRA was funded at $5M, enabling funding of nearly the target amount
- Multi-project co-ordination:
  - Jefferson Lab was to have received the next LQCD cluster in FY 2010
  - A collective decision was made to put LQCD ARRA resource at Jefferson Lab, and to re-locate the FY2010 LQCD-ext machine to Fermilab, shifting it later in the year to create the possibility of a combined FY2010-11 larger machine (which was done)

The LQCD SciDAC project provides the necessary software for both of these computing projects. USQCD proposals for INCITE and NSF allocations address capability computing

The ARRA & LQCD-ext projects primarily target ***high end capacity*** (many jobs < 1 Tflops sustained performance), with some ability to run some of several TFlops.

Project Goals & Phasing

Original performance goal: to nearly double USQCD's resources, at that time 17 Tflops.

As an ARRA project, another goal was to move as quickly as possible to get funding into the economy.

- The project pwas structured to include 2 procurement phases
  - Phase 1, $1.78M in hardware to be awarded by the end of FY2009
  - Phase 2, $1.70M in hardware, to follow by ~3 months

- It evolved to include GPUs
  - By the time the project started, it was clear GPUs would be ready for exploitation by LQCD, enabling a significant performance increase
  - Each phase was adjusted to include a GPU component, so that GPU deployment could match the community's uptake of the technology
  - In late 2011, most of the 2009 GPUs were upgraded to get a highly cost effective performance boost. In 2012, a small Intel Xeon Phi cluster (12 nodes of 4 cards each) was purchased to allow software development and early use of this new, competitive architecture

As a result of the GPUs, the delivered effective performance reached 95 Tflops!

Technical Goals

Performance Goals:

1. To significantly increase the computing resources available to the USQCD collaboration for "analysis"…

   Original target was 16 Tflops sustained aggregate performance averaged over the 3 dominant inverter actions:

   - Domain Wall Fermions (DWF)
   - Staggered (asqtad = a-squared tadpole)
   - Clover, particularly anisotropic clover

   When the project was changed to incorporate GPUs, this goal was raised to 60 TFlops. The final system sustained an effective aggregate performance of 95 Tflops. (Application mix dependent, see below).

2. To deliver an integrated effective performance of 180 TFlops-years.

   Due to the GPUs, this was achieved one year early, and the hardware continues to contribute significantly to USQCD science.

GPU Accelerators

Strategy: buy as much computing capacity for the dollar as possible.

As the ARRA project was starting, USQCD collaborators were finishing up a GPU accelerated implementation of a key kernel (inverter) and were achieving high performance; therefore GPUs were incorporated into the project to increase the total performance.

Phase 1: 25% of compute funds to GPU accelerated nodes

- ✓ Enough software was becoming ready to exploit this capacity, and software development environment (CUDA) was maturing rapidly
- ✓ GPUs allowed this project to double the USQCD total computing capacity

Phase 2: 45% of compute funds to GPU nodes

- ✓ Multiple groups were in production, and were eager to absorb a large increase in capacity; allowed this project to again double the USQCD total capacity
- ✓ Availability of ECC memory on the Tesla GPUs held a promise of expanding beyond inverters to satisfy more of the collaborations computing requirements, so a portion included this capability; many users now exploit this capability.

Quantifying Aggregate Performance

(Reminder) LQCD computing proceeds in 2 phases:

1. Configuration generation (on supercomputers)
   - ❖ Must be produced sequentially, at highest performance
   - ❖ End product: 1000+ configuration files
2. Analysis (propagator generation + observables)
   - ❖ 1000 + jobs able to run in parallel
   - ❖ Target performance: 1% of configuration generation (then at 10's of Tflops)

Analysis is the relevant task for this project. For benchmarking for the LQCD ARRA resources, we selected production lattice sizes for each of the 3 main inverters:

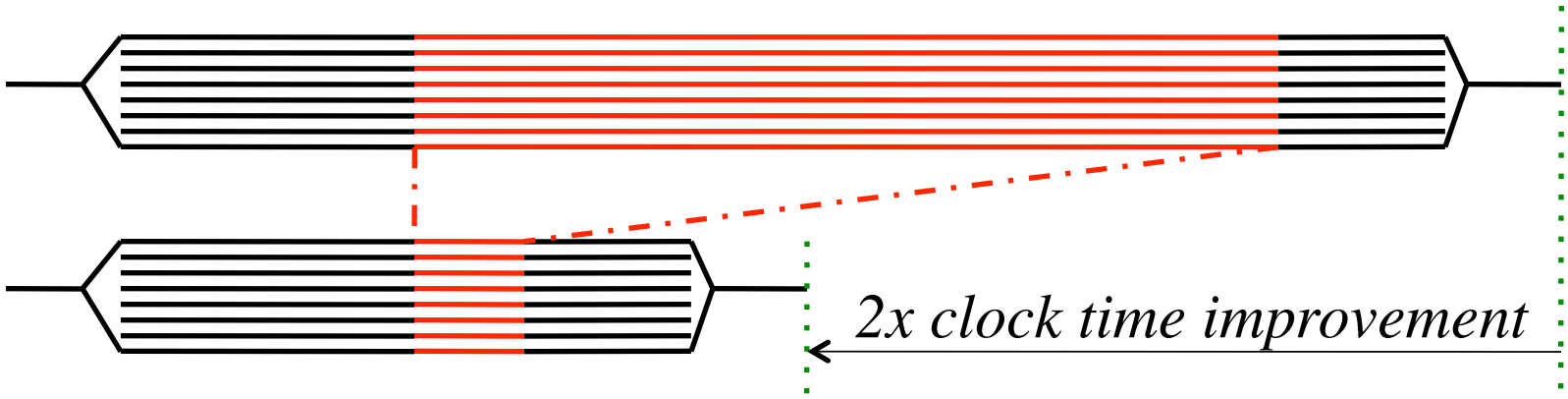
- ❖ Anisotropic Clover: 24^3 x 128
- ❖ Asqtad: 56^3 x 96
- ❖ DWF: 32^3 x 64 x 16

Amdahl's Law (Problem)

A major challenge in exploiting GPUs is Amdahl's Law:

If 60% of the code is GPU accelerated by 6x,

the net gain is only 2x.



Also disappointing in this scenario: the GPU is idle 80% of the time!

Fortunately many LQCD codes spend > 95% of their clock time in a single kernel, a matrix inversion, and so for these applications Amdahl's Law was not (yet) a show-stopper, and gains of 18x using 4 cards were achieved.

Conclusion: GPU node performance is application dependent, and so rating the GPU nodes requires taking usage into account.

Creative Partitioning

At the end of 2009, Robert Edwards made 2 significant changes to his workflow:

1. seeing that inversions were now cheap, he increased the number being done, to improve spectrum statistics
2. to further skirt Amdahl's Law, he broke his application into 2 jobs, one 99% inversions, the other 0% inversions

Using 4 GPUs on one node, the inversions ran 24x faster (pre-Fermi generation cards), for a net job acceleration of 18x for a cost of 1.5x (using gaming cards), thus a price perfomance improvement of 12x!

So instead of a 50% annual increase in allocation (typical), he saw a 600% increase (12x50%) in his workflow, all from the GPUs. The ensuing spectrum calculations could not otherwise have been done.

The non-accelerated jobs still needed a resource more expensive than the new GPU resource. I.e. we were leveraging the existing and new conventional x86 resources to achieve this large gain.

(By 2012, this began to saturate; see later talk.)

GPU Job Effective Performance

We define an "effective" performance to be the x86 inverter performance multiplied by the *job* clock time reduction. I.e. it is the performance of an standard set of nodes yielding the same clock time.

The following table shows the number of core-hours in a job needed to match one GPU-hour in a job. Last project used 32 single GPU nodes and was I/O bound.

The allocation-weighted effective performance of the GPU cluster reached a high of **85 TFlops** in 2012 (inverter heavy running).

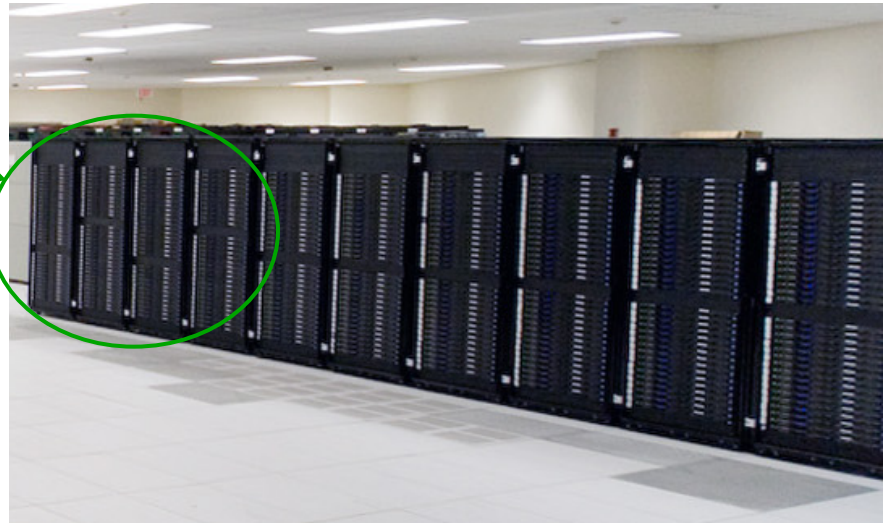
Project	2010-2011 Hours	#GPUs, nodes	Jpsi core hours / GPU hour (job time)	Effective Performance Gflops/node	GPU used
Spectrum	1,359,000	4, 1	180	800	(average)
thermo	503,000	4, 1	90	400	(average)
disco	459,000	4, 1	92	410	C2050
Tcolor	404,000	4, 1	40	175	GTX285
emc	311,000	4, 1	80	350	(average)
gwu	136,000	32, 32	47	50	GTX285

CPU Cluster & IB Fabric Design

The most cost effective conventional nodes were dual Intel systems,
2.4 GHz Nehalem / 2.53 GHz Westmere (phase 1 / 2), about 20 Gflops/node

QDR Infiniband switches have 36 ports, so can hold 32 nodes and still have ports free to connect to the file systems (powers of 2 are best for LQCD). Deploying multiple sets of 32 nodes reduces the cost of the Infiniband fabric while maintaining the highest efficiency for jobs up to 640 Gflops.

17 racks purchased for phases 1 & 2:
13 as single racks non-oversubscribed,
4 interconnected 2:1 oversubscribed
(to support jobs up to ~2 Tflops)



Most jobs on these clusters are 1 node (8 cores), 8 nodes (64 cores), or 32 nodes (1 rack, 256 cores)

Large jobs (256-1024 cores) are moved to a higher priority run queue to prevent starvation by small jobs & backfilling.

All logical partitions have 2 uplinks to a core switch for file services.

GPU Cluster Design

Summary of Key Decisions:

1. NVIDIA CUDA chosen as the most productive software environment.
2. NVIDIA Fermi Tesla cards were the only GPUs supporting ECC memory protection, again keeping us single supplier.
3. For the key inverter kernels, GeForce gaming cards were 3x more cost effective than Tesla cards, with both lower cost and higher performance (ECC on GDDR memory consumes bandwidth, plus GeForce cards are clocked higher).

   The occasional memory errors can be caught on large matrix inversions by a quick test of the residual when the kernel has completed, so running on imperfect hardware is acceptable (early example of fault tolerant computing).
4. The early (and current) workload is mostly 1-4 GPUs with light enough use of the CPU to allow putting 4 GPUs into a single host, yielding a very high performance, modest cost platform. Even today, inverter-only use dominates.

Fall 2011 GPU Upgrade

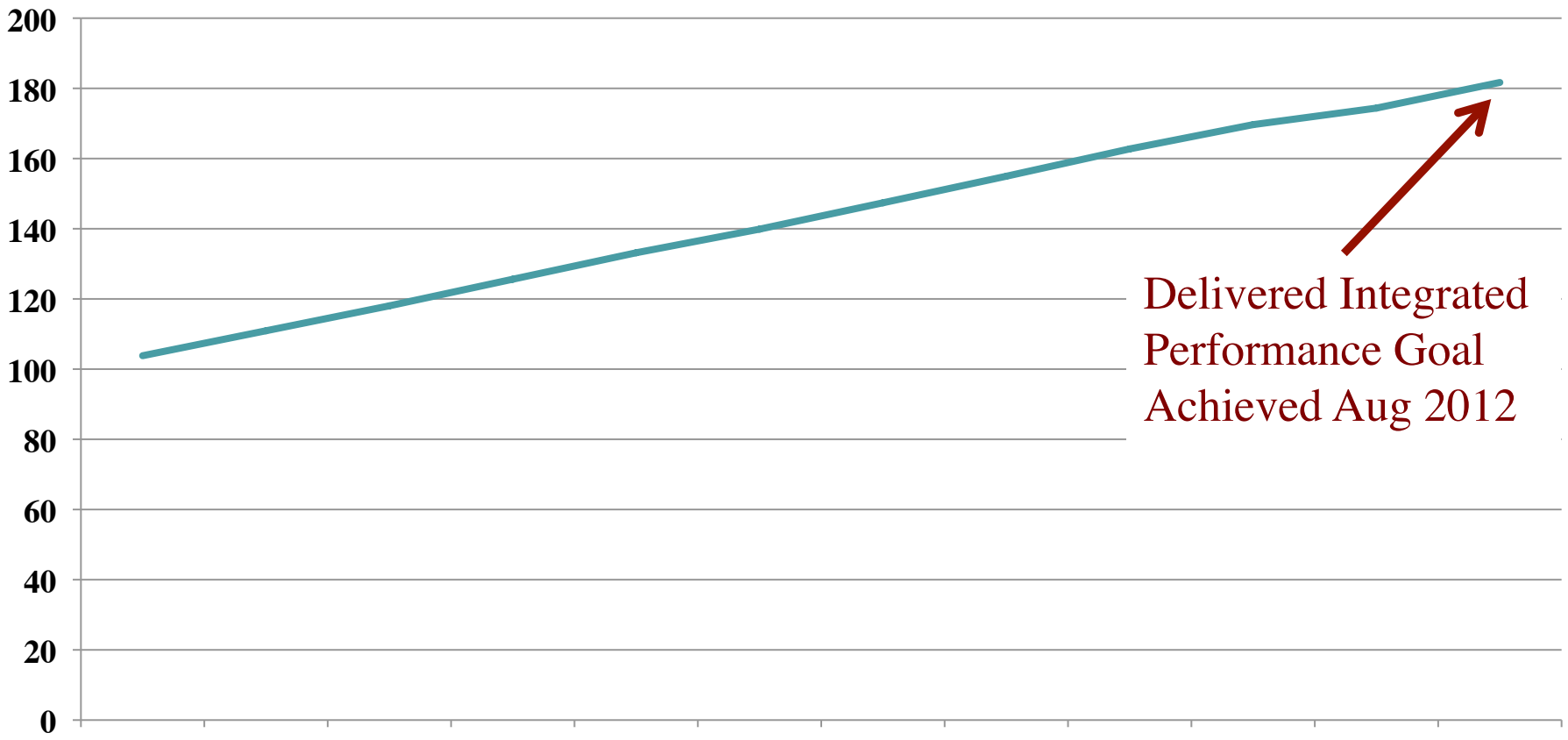
Because of the high performance of the GTX-580 cards, it was highly cost effective to replace the GTX-285s with GTX-580s. This was a level 2 change only, but still discussed with DOE and approved.

The GTX-580s, like the GTX-480s, exhibited higher failure rates than the GTX-285s. The project did its own bin selection (no vendor would guarantee fault free calculations) and set an upper threshold for memory error rates (10 per 2 hours of memory testing). This is checked weekly, and cards that fail are pulled and excessed.

Even taking into account bin selecting the GTX-580s (discarding 20%), upgrading 3 nodes cost the same as buying 1 new node, and increased the performance by the equivalent of 2.5 new nodes!

This type of upgrade was only cost effective because the gaming cards are such a small percentage of the system cost (~33% after bin selecting cards).

“Effective” Tflops-years FY2012



Notes

1. GPU nodes are rated based upon relative performance of equivalent infiniband cluster jobs for the production projects (as reported by users), weighted by the projects allocations, to give “effective” Tflops.
2. Clover performance (Gflops, single-half, per GPU in 4 GPU job):
   GTX-285 = 130, C2050 w/ECC on = 176, GTX-480 = 273, GTX-580 = 300
3. Aggregate performance (“effective”) May 2012: 85 Tflops GPU, 9 Tflops conventional, 94 Tflops total

File System

Open source Lustre was chosen to support a large flat namespace, and to enable scaling out in capacity and performance.

Final ARRA Configuration: 416 TB, ~4 GB/s, $228K

Phase 1: 224 TB across 14 servers (excludes RAID-6 8+2 overhead)

- dual Nehalem 2.26 GHz, 12 GB memory
- 24 * 1 TB disks, 24 disk RAID controller, DDR Infiniband
- bandwidth measured at 1.4 GB/s using 6 nodes (single DDR uplink)

Phase 2: 192 TB across 4 servers

- similar to above, but with 3 RAID-6 (8+2) strips per server instead of 2
- 2 TB disks, QDR Infiniband, higher performance RAID controller
- somewhat lower bandwidth / TB, but still more than necessary

An upgraded Meta-Data Server is now dual head with auto-failover.

Summary of Resources Deployed

Conventional Systems:

544 nodes, ~20 Gflops/node, 10 TFlops

GPU Accelerated Systems:

123 quad GPU nodes

- 32 quad C2050
- 40 quad GTX-285
- 51 quad GTX-480 / GTX-580

(34 single GPU nodes, overlaps conventional count)

Xeon Phi Accelerated Systems:

12 quad 5110P nodes

File System

18 servers, ~400 TB, ~4 GB/s aggregate bandwidth

Technical Summary

The ARRA LQCD Computing project has deployed

- 10 Tflops conventional infiniband systems
- 508 GPUs equivalent to over 100 Tflops sustained capacity for anisotropic clover inverter-heavy jobs, and 65-85 Tflops depending upon the mix of jobs
- 48 Xeon Phi accelerators, with preliminary performance of around 10 TFlops (tbd, not in production)
- 416 TBytes disk, backed by multi-petabyte tape library

Total deployed capacity: 85-105 Tflops (effective), a gain of more than 5x over the original plan of 16 Tflops.

Integrated performance reached 181 TFlops-years in Aug 2012, and exceeded 200 TFlops in Nov 2012.

The total effective Tflops depends upon the efficiency with which the applications use GPUs, and could in principle rise as a larger fraction of the existing code is ported to the GPU (reduced Amdahl's Law problem), or fall as new codes with lower GPU intensity begin to exploit GPUs.

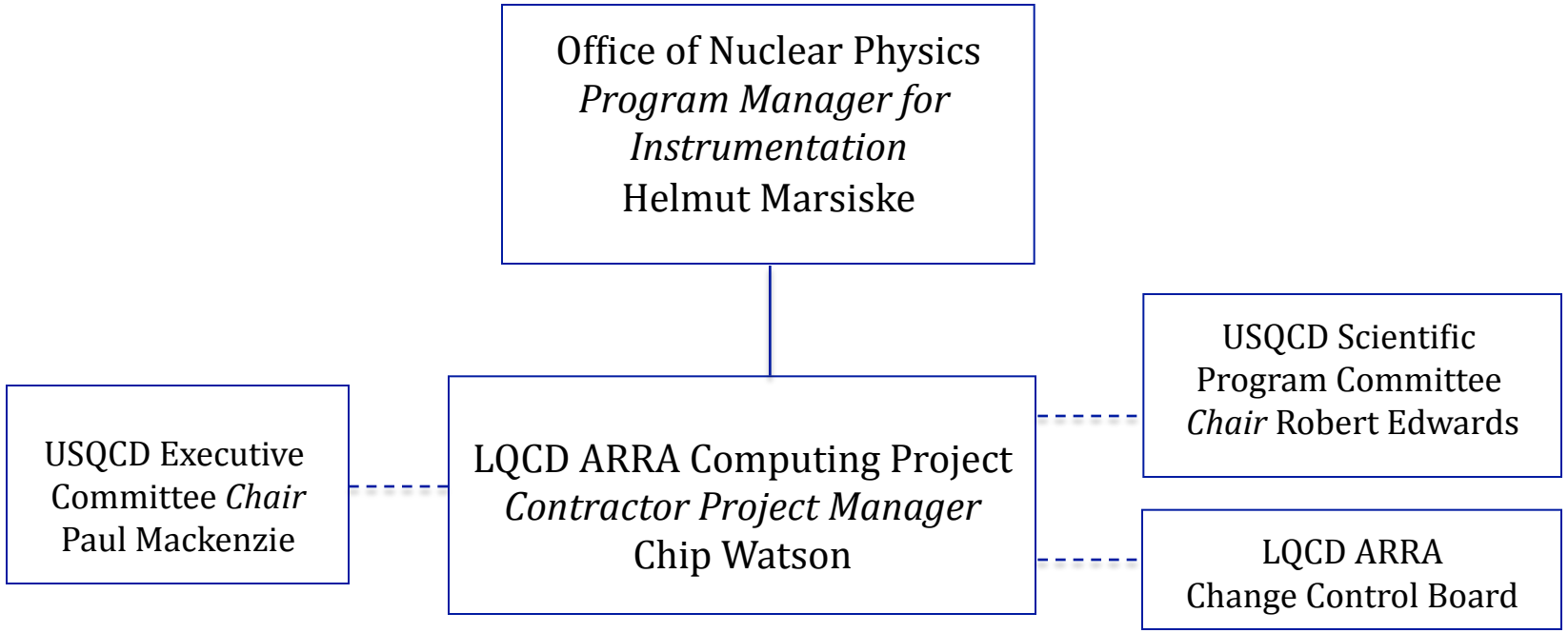
Management

The ARRA LQCD project was in many ways modeled after the existing LQCD Computing project, and re-used the following components or management approaches:

- Relationship to the USQCD Executive Committee
- Relationship to the Scientific Advisory Committee (computing allocations)
- Approach to hardware selection and alternatives analysis, to achieve the greatest performance for dollars invested
- Approach to benchmarking
- Cost model for operations (FTE planning)
- Change control process (but simplified since there is only one site)

Because of the lower total project cost, a single site, and fewer deployment cycles, management was intentionally lighter weight.

Management Organization



Management Organization Chart for the LQCD Computing Project.

Vertical lines indicate reporting relationships. Horizontal lines indicate advisory relationships

Initial Project Budget

WBS	Name	Total Cost K$
1.	Project Planning and Management	97
2.	Deployment	
2.01	Site preparations	140
2.02	Phase 1 deployment	1,830
2.03	Phase 2 deployment	1,777
3.	Operations	
3.01	Year 1	238
3.02	Year 2	283
3.03	Year 3	294
3.04	Year 4	306
	Total Project Cost	4,965

Change Control

Level	Cost	Schedule	Technical Scope
DOE Program Manager (Level 0)		> 1-month delay of a Level 1 milestone date	Change of any WBS element that could adversely affect project performance specifications
LQCD ARRA CCB (Level 1)	A cumulative increase of more than $200K in WBS Level 2	> 1-month delay of a Level 1 milestone date	Any deviation from technical deliverables that does not affect expected project performance specifications.
LQCD ARRA Contractor Project Manager (Level 2)	Any increase of > $50K in the WBS Level 2	> 1-month delay of a Level 2 milestone date	Technical design changes that do not impact technical deliverables.

The GPU upgrade in the Fall of 2011 was a Level 2 change ($77K), as was the purchase of the Xeon Phi cluster in the Summer of 2012. Both procurements traded operating months for higher performance.

Merging with LQCD-ext resulted in a total cost shift of more than $200K for the LQCD ARRA project, and was of a scope sufficient across both projects to be engage oversight all the way up through DOE. Resulting changes among hardware and operations across the combined projects, however, was routine.

FY2012 Project Adjustments

In order to expeditiously finish up all ARRA projects, DOE asked the two LQCD computing projects, LQCD-ext and ARRA LQCD, to transfer ARRA hardware operations to LQCD-ext as soon as ARRA LQCD met all of its targets.

Reasonableness:

ARRA LQCD had finished deployment and was essentially just operations, with a strong overlap in functions with LQCD-ext (operations the same, just for different hardware). All performance milestones would first be met.

Implications:

- Transfer out-year operations labor costs to LQCD-ext (FY13 and FY14)
- Spend the remaining funds on one last small strategic procurement (last 3% of ARRA LQCD budget)

Software Trends Guiding Last Procurement

Early running on the GPUs was mainly matrix inversions, with the largest fraction of that being split half-single precision anisotropic clover with high performance/$ gain.

The low cost of matrix inversions and the large GPU resource has moved the bottleneck elsewhere, and software developments were moving more of the data parallel work onto GPUs (ECC capable).

NVIDIA had figured out how to hamstring the gaming cards compared to the professional line, and the window of opportunity for 10x gains from GPUs was closed – but it was well exploited by USQCD!

Jefferson Lab was already an early access user of Intel's Knight's Corner (Xeon Phi) system, which was showing gains in software development speed. I.e., it might help crack the bottleneck in software development for accelerated system.

Exploiting Intel Accelerators

Extrapolating labor costs to the end of FY2012, the planned date for the transition of the LQCD ARRA hardware to the LQCD-ext project, there was estimated to be approximately $150K of remaining funds (3% of total project funds).

A decision was made and approved by DOE to use these funds to augment the computational resources and to deploy a modest Intel Xeon Phi cluster to support both software R&D and early production running.

Jefferson Lab was one of the first sites to receive the Xeon Phi 5110P accelerators (early FY13). Through a close collaboration with Intel, a single precision d-slash operator and basic Wilson inverter has been developed, and yields the same performance as a K20 (and at a lower cost). Looks competitive, and definitely strategic for preparing to exploit NSF's Stampede machine (first large Xeon Phi machine).

Project Budget Performance

A large fraction of the budget was for computing hardware and disk systems, with a build-to-cost procurement strategy aiming at highest performance meeting specifications at fixed cost.

Labor costs were fairly well understood from the LQCD Computing project (2004-2008). Uncertainly was mostly due to use of new hardware technology (GPUs).

Contingency of ~5% was therefore only carried on the labor (management and operations).

Work Organization and Budget

For budgeting purposes, a simple WBS was adopted (from FY2010 Q2):

WBS	Item	Baseline Total Cost (AY$)	Costed & Committed (AY$)	Estimate To Complete (AY$)	Estimated Total Cost (AY$)	Baseline Contingency (AY$)	Remaining Contingency (AY$)
1.01	FY09 Mgmt	26	26	0	26	0	0
1.02	FY10 Mgmt	25	15	10	25	1	1
1.03	FY11 Mgmt	15	-	13	13	1	1
1.04	FY12 Mgmt	15	-	14	14	1	1
1.05	FY13 Mgmt	16	-	14	14	1	1
2.01	Site Prep	250	195	50	245	16	6
2.02	Phase 1	1,970	1,967	0	1967	0	0
2.03	Phase 2	1,816	864	952	1816	0	0
3.01	FY10 Operations	121	53	68	121	6	12
3.02	FY11 Operations	228	-	228	228	11	12
3.03	FY12 Operations	218	-	218	218	11	12
3.04	FY13 Operations	207	-	207	207	10	12
	Totals	4907	2185	2727	4912	58	58

Contingencies on the Phase 1 and 2 deployments are zero as they are build to cost systems.

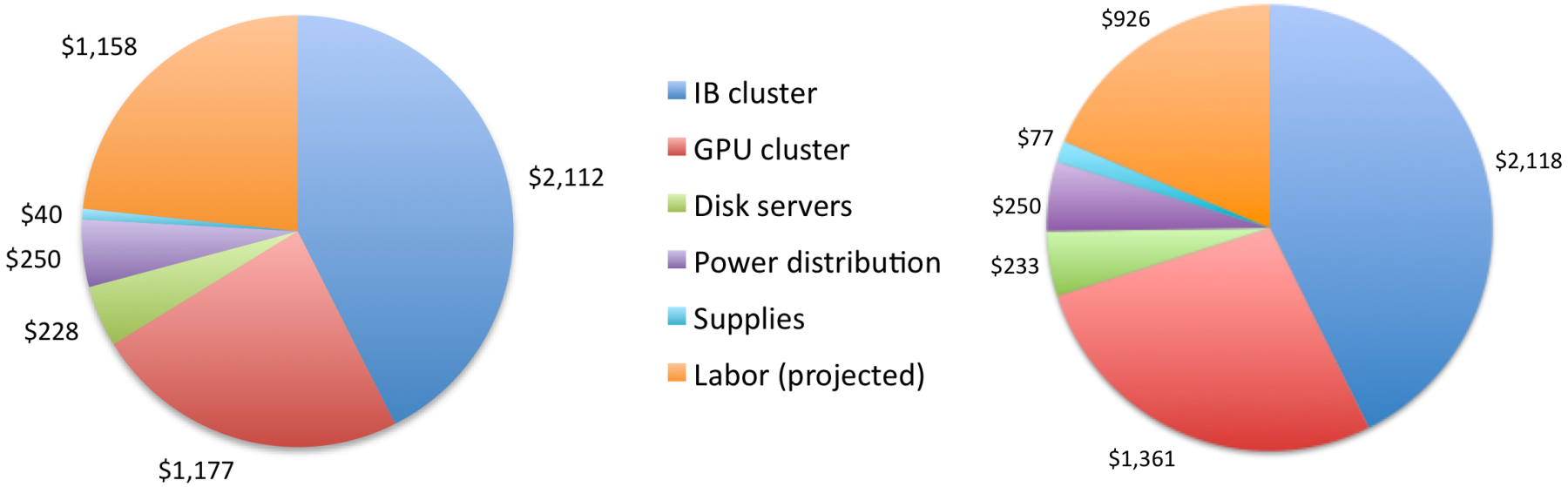
Budget Plan Adjustments

Original plan (left)

- 72% of the funds for hardware (Infiniband & GPU clusters, disk servers)
- 5% for power conditioning and distribution
- 23% for labor

Final (right)

- added GPU upgrades from GTX-285 to GTX-580
- Added additional final hardware procurement
- Reduced labor costs (1 year shorter operations)

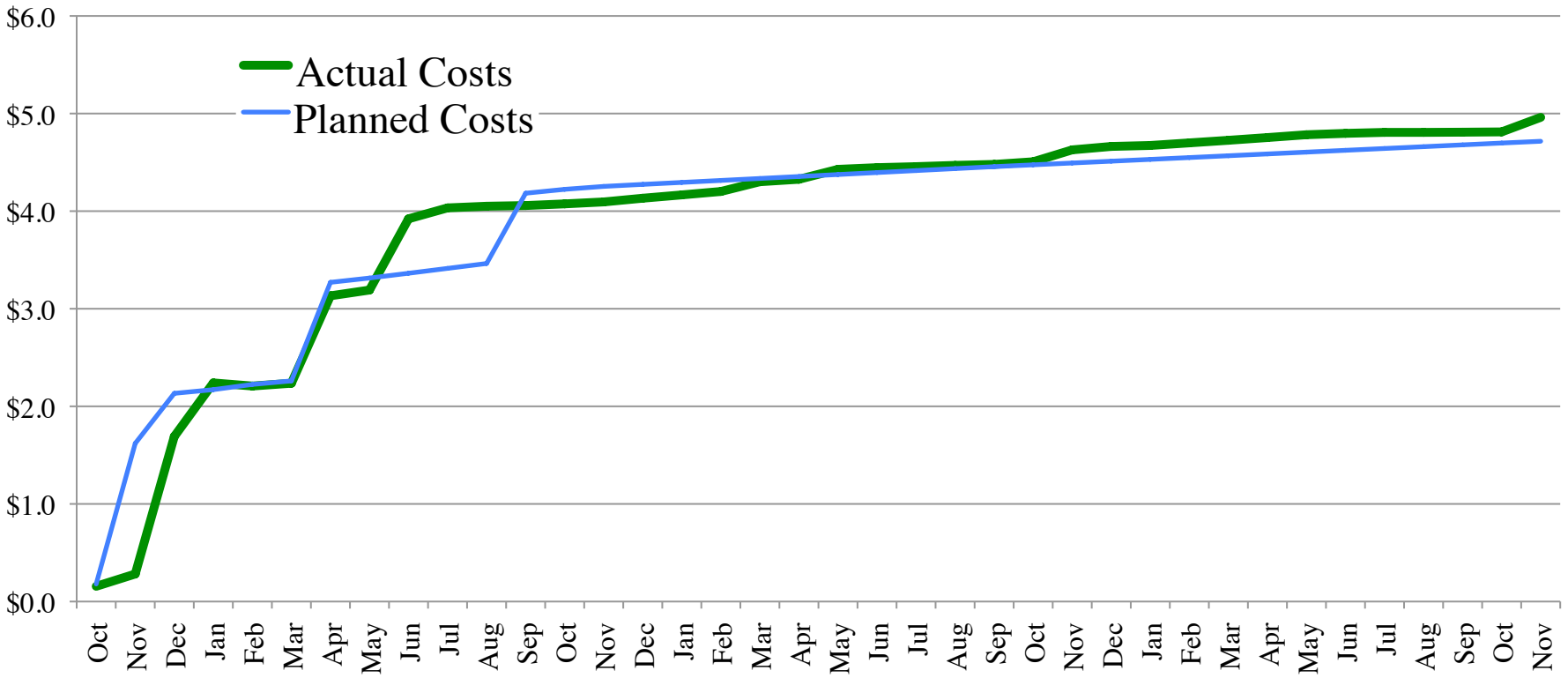


Numbers show at right are final costs

Costs through Nov 2012 ($K)

FY09 Budget	FY09 Actual[1]	FY10 Budget[2]	FY10 Actual	FY11 Budget	FY11 Actual	FY12 Budget	FY12 Actual	FY13 Budget[4]	FY13 Actual
$2,246	$77	$1,962	$3,980	$254	$421	$317	$330	$156	$152

[1] Does not include $1,890 committed for computers and power distribution
[2] After Dec 2009 milestone & level 2 budget changes
[3] Including GPU upgrade
[4] All remaining funds



Summary

The LQCD ARRA project completed on budget and ahead of schedule, exceeding all performance targets.

The project achieved more than a 6-fold increase in total delivered Tflops capacity: 10+85+(10) > 100 Tflops vs 16 Tflops, by moving aggressively to exploit GPUs, while still expanding capacity for non-accelerated codes.

The hardware allowed calculations to be done years early, making a significant impact on the science program of USQCD.

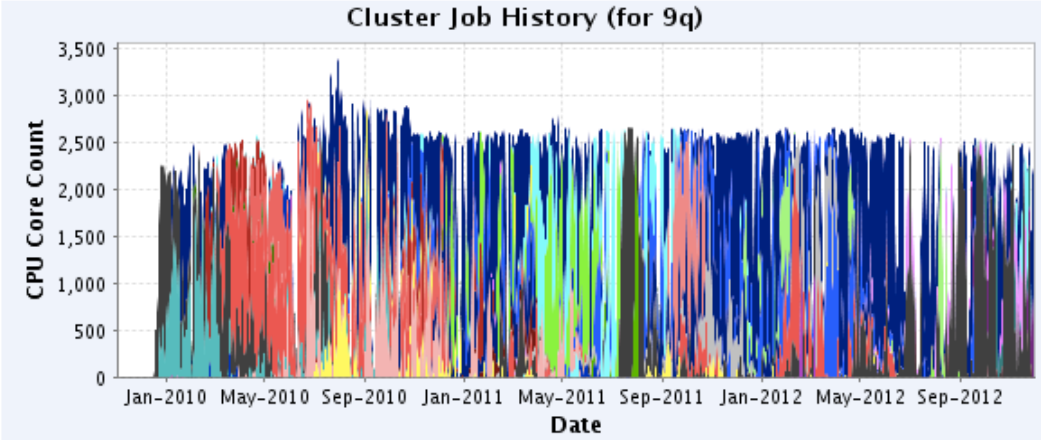
The size of the resource in turn accelerated software developments on GPUs, and has well positioned the community to exploit multiple GPU-accelerated capability machines now available.

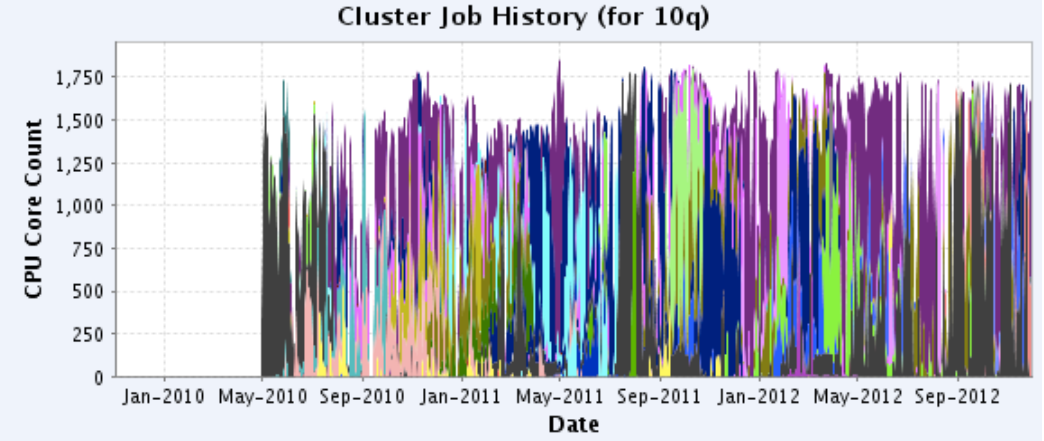
(end of presentation)

Infiniband Running Status and Usage

Infiniband Clusters:

- Clusters ran at > 95% up except for power or cooling disruptions, or system upgrades.
- Utilization was generally high, with dips due to holidays or conferences, or all projects catching their breath at the same time. Some fluctuations are draining for large jobs to start.
- One rack of 10q was dual use for GPUs, so 10q has larger fluctions in the job history as nodes flip between GPU and IB.
- There are some misassignments of jobs from 10q to 9q in Sept 10.





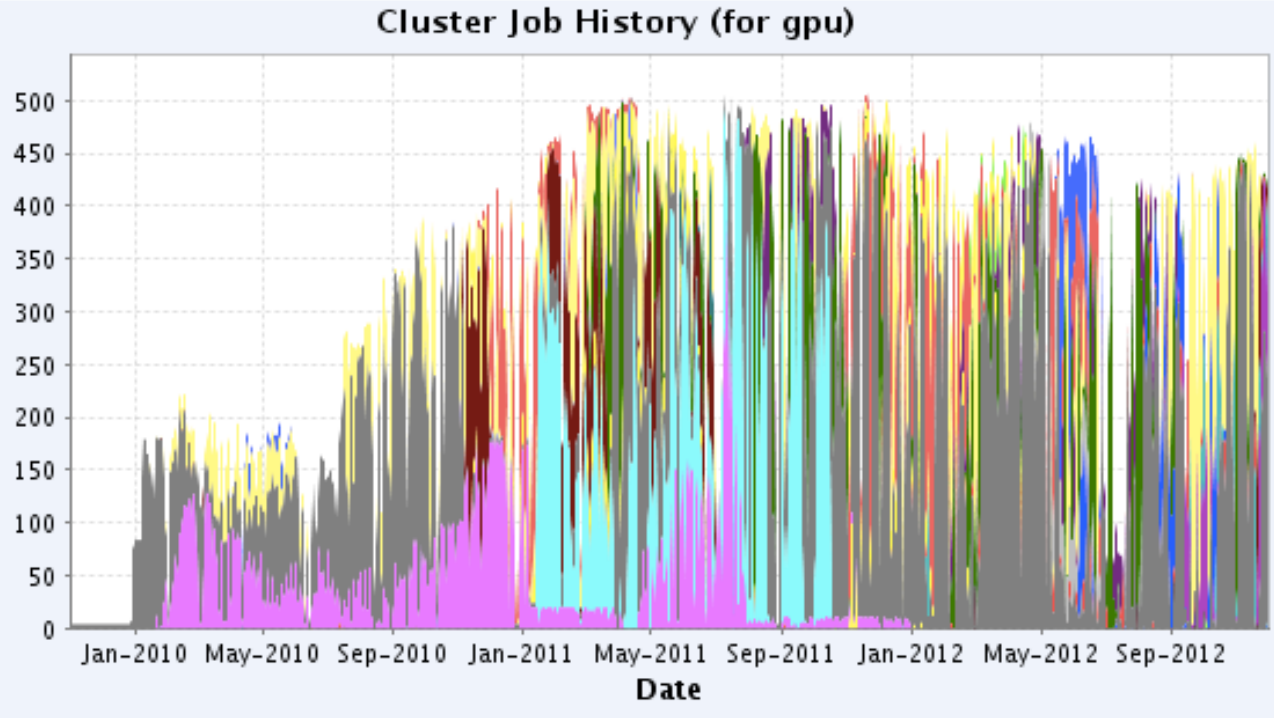
Additional operational charts and graphs can be viewed at http://lqcd.jlab.org/

Running Status: GPU Nodes

GPU Cluster

(Graph is in GPU count, un-normalized for performance.)

- Phase 1 went into production in Jan 2010
- Phase 2 took longer to ramp up, as problems with the GTX-480 GPUs were addressed



- The dips in November 2011 and Jan-Feb 2012 are the upgrades from GTX-285 to GTX-580 cards (and dealing with early failures). Performance was much higher although card count was lower for a couple of months.
- In late 2012, there are some dips in utilization as large GPU projects were not yet ready to run.

Running Status: Disk System

Lustre System

The system consists of

- dual head Meta Data Server
- 23 Object Storage Servers

All nodes are on Infiniband, as are all the compute nodes.

We scale out to increase capacity and bandwidth

- most recent ARRA systems use 2TB disks, with QDR IB

The system performs quite well, with sustained average read throughput of 6 GB/s, aggregate.

Additional expansions planned for this summer and fall. ??? ARRA or LQCD-ext???