FY2012 Acquisition Plan
for the
SC Lattice QCD Computing Project Extension
(LQCD-ext)

Operated at
Brookhaven National Laboratory
Fermi National Accelerator Laboratory
Thomas Jefferson National Accelerator Facility

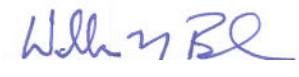
for the
U.S. Department of Energy
Office of Science
Offices of High Energy and Nuclear Physics

Version 1.2

February 23, 2012

PREPARED BY:
Chip Watson, JLab

CONCURRENCE:



William N. Boroski
LQCD-ext Contractor Project Manager

February 23, 2012
Date

FY12 Acquisition Plan
Change Log

Revision No.	Description	Effective Date
Rev 0.1	Initial version.	10/17/2011
Rev. 1.0	Incorporated feedback from first version.	10/31/2011
Rev. 1.1	Updated to reflect new information on funding, hardware availability.	12/09/2012
Rev. 1.2	Added table of dates at the end.	02/23/2012

LQCD-ext FY2012 Acquisition Plan

Version 1.2

Summary

In FY2012 the LQCD-ext project will deploy the following hardware at Jefferson Lab:

1. An Infiniband cluster (without accelerators)
2. An accelerated Infiniband cluster (with accelerator cards such as GPUs)
3. Disk servers

The split between the accelerated and non-accelerated clusters will be determined as late as possible, so as to best match the evolving requirements of the users. For the purposes of this document, the two clusters are assumed to be of roughly equal cost, of scale $0.8M to $1.2M per cluster, with the sum fitting into the roughly $2M combined computing budget. Disk space will be roughly an additional $100K.

Timing for the procurements were planned to take into account operation under a five month continuing resolution (now shorter than originally anticipated), and will take into account anticipated release dates for new hardware. Hardware of high interest in 2012 includes:

1. AMD Interlagos server CPUs
2. Intel Sandy Bridge server CPUs
3. NVIDIA Kepler GPUs
4. Intel Many Integrated Core (MIC) device, a.k.a. Knights Corner
5. Mellanox FDR Infiniband

The AMD CPU has now been released, the Intel CPU is anticipated to be released in Q1 of 2012 (dual socket) and Q2 (quad socket), the NVIDIA GPU in Q3 or Q4 of 2012, and the Intel Knights Corner in late 2012. (All of these dates are from Wikipedia and other web searches and do not use any non-disclosure information.)

Because the CPU refreshes occur earlier, the non-accelerated cluster will be procured first, waiting only long enough to accumulate the necessary funds under a continuing resolution. The accelerated cluster will be bought when at least one of the new accelerators is available, aiming at a delivery on or before Sept 30, 2012, which might preclude the Tesla version of Kepler, and MIC (Knights Corner). If no new accelerators turn out to be viable in FY2012, then an NVIDIA Fermi GPU based solution will be pursued (for which procurement costs should be falling as Kepler nears its release date).

Disk space will be deployed as needed, anticipating increased requirements ahead of each deployment of a cluster, and responding to requirements given in responses to the 2012 call for proposals.

The total performance increase for USQCD coming from the FY2012 procurements will depend upon the split between non-accelerated and accelerated nodes. As explained in the Alternative Analysis document (Aug 11, 2011), an equal split of funds should yield about 14 TFlops non-accelerated and 50 TFlops accelerated (effective

performance, depending upon the detailed mix of applications), for a total of ~ 64 TFlops.

Continuing Resolution Considerations

Under a continuing resolution, we anticipated receiving the equivalent of 1/12 of the FY2011 funds per month. With a budget now passed, we anticipate having all funds available in March 2012, allowing a full award for the non-accelerated cluster. This procurement is now underway and will be awarded in March.

This award will be shaped into a roughly $800K initial procurement (40% of hardware budget) with options to increase that to up to 60% as late as the end of May, by which time we will have better information on the availability of next generation accelerators, and better information on the computing requirements and accelerator software readiness of the collaboration for FY2013.

Deciding on the Split between Accelerated and Non-Accelerated Clusters

The decision on the size of the first cluster will be driven by the software maturity of GPU (or other accelerator) software anticipated for FY2013. Since the accelerated nodes can deliver more science per dollar, they are preferred if sufficient software is or will be ready to exploit them. If insufficient software is ready, the first cluster will be expanded to ~$1.2M in units of whole racks. The maximum expansion of the first cluster would yield of order 25M Jpsi core hours per year, a slightly less than 10% impact on the non-accelerated cluster capacity of USQCD.

The accelerated cluster at $0.8M (the minimum) will add of order 240 GPUs of perhaps 2x the performance of the current GPUs. This yields approximately 3.4M new standard GPU hours per year, on top of an existing 4.9M, for a total of 8.3M. The largest accelerated cluster size would yield an additional 1.7M hours per year, or a 20% increase in USQCD's capacity compared to the minimum sized FY2012 addition. If next generation accelerators are not available for this year, the increases will be cut in half, yielding only a 12% swing depending upon what fraction of the hardware dollars are allocated.

The decision on these minor adjustments of +/- 5% from the average value on the capacities for these architectures will be based upon best data available at the end of May. Input will be sought in April from the Scientific Advisory Committee, who will have at their disposal the allocation requests submitted in March of 2012. Additional input will be sought at the collaboration meeting to be held May 5-6 of 2012. Based upon this data, a decision will be made by the end of May as to how large a fixed priced procurement option (if any) to exercise on the non-accelerated cluster.

Infiniband Cluster

Quad socket Sandy Bridge will likely not be available as early as we would like to start procurement of the non-accelerated cluster. Our guiding principle is that we delay procurements if, for each month of delay, we anticipate a gain in performance of more

than 5%. At this time we do not anticipate such a gain for quad socket Sandy Bridge, and so we will not delay the first cluster.

For Interlagos CPUs, we expect that the price/performance will be better for quad socket servers than for dual socket servers. This was already the case for the Ds cluster at Fermilab in 2010-2011, which has the Magny-Cours processor, the Interlagos predecessor.

Power and Space Estimates

Dual socket systems might be in the $4K range, and quad socket systems in the $5K-$7K price range, so we are looking at a system size of order 180 +/- 60 nodes. Power requirements have consistently come in at 1 watt per $10 spent for non-accelerated nodes, and this trend is likely to continue, so a $1M cluster would require 100 KW +/- 10% spread across about 8 racks. The 7n cluster will be decommissioned to free up sufficient power and cooling for this cluster (~100 KW cooling released, an additional 30 KW cooling available in the same area; ample power available).

Memory Bandwidth

LQCD is always memory bandwidth constrained, and the strength of both Sandy Bridge and Interlagos is their increased memory bandwidth. Both chips will have 4 channel memory controllers supporting up to DDR3-1600 memories. For Intel this is an increase of 33% (from 3 controllers to 4) and for AMD an increase of 50% (from 1066 to 1600). However, the highest end chips (supporting DDR3-1600) may remain too expensive for our purposes. The latest USQCD cluster using Intel was the 10q ARRA cluster with only triple DDR3-1066 buses, even though 1333 was available. If we end up with quad DDR3-1333 buses for 2012 this would be a 67% improvement in 2 years, slower than Moore's Law. Similarly, if we go from 1066 to 1333 on AMD, this will be a 25% improvement in roughly a year (Ds came in 2 phases). Hopefully the 1600 buses won't have too high a price premium so that we can get the larger performance gains per node. Cache sizes will be larger, but memory bandwidth will remain a strong constraint. Benchmarking will be necessary to ascertain the most cost effect CPU and bus speeds for both Intel and AMD.

Memory per Core

If we use a single 4 GB DIMM per memory bus, then an 8 core Intel chip will have 2 GB per core. Interlagos could have up to 16 cores per socket (or 8 core pairs sharing a floating point pipeline), and we could use a single 8 GB DIMM or a pair of 4's per memory bus to reach the same 2 GB / core if we want to count each integer core, or we could stick to the same 4 GB DIMM size and have 2 GB per AVX floating point unit, the same as for Intel. Note that the difference between 64 GB and 128 GB / server would be around $640 ($10/GB), or of scale 10% of the system cost. Since recent acquisitions have leaned towards 2-3 GB/core, we will not go down to 1 GB/core (which has on occasion been a constraint on the 7n cluster). Doubling to 4 GB / core will be too expensive.

Benchmarking

As in previous years, inverter benchmarks will be used to measure performance and price/performance for the non-accelerated cluster: DWF, Asqtad, and Clover. Current

production problem sizes will be used, and the benchmark problems will be scaled to correspond to running a production job at a sustained 1 TFlops, thus something on the order of 512-1024 cores. The appropriate global and hence local problem sizes for the 3 actions will be selected with input from currently running or upcoming USQCD projects prior to the call for proposals so that vendors can be given benchmark applications for testing their hardware.

Infiniband Network

QDR Infiniband, at 40 Gbps, should be adequate for the non-accelerated cluster, and will likely remain less expensive than the newer FDR (56 Gbps). However, for quad socket systems QDR would need to be deployed without oversubscription, and so a 2:1 FDR oversubscribed solution is expected to be better for such systems.

For QDR without oversubscription, 12 leaf switches plus 6 additional switches used as a scalable core could be used to support up to 214 nodes, leaving 2 ports free for an uplink to the file system. Alternatively, a single switch such as the 216 port IS5200 might be a viable choice, although typically large switches are priced too high. Functionally these two solutions are the same, and the selection between the two will be based upon cost (including the cost of cabling). (IS5200 list price for 216 ports is $104K, 18 unmanaged leaf switch @$6K $108K; single switch solution requires longer cables, leaf design requires 2x as many shorter cables; fewer spare parts are needed for the distributed solution).

An FDR solution with 2:1 oversubscription would use 9 leaf switches plus 3 spine switches (there is no appropriate ~200 port core switch). This uses 1/3 fewer switch chips, which might be enough to offset the higher cost, and would simplify cabling. Each node could still have a QDR HCA, with only the inter-switch links running at the ~2x faster speed. FDR switches are available online for $9K.

Since quad socket nodes are of higher cost than previous dual socket nodes, the price/performance advantages of segmenting the IB fabric to reduce the fabric cost are much smaller than they were in 2009-10 for the ARRA cluster, and this approach will not be as compelling as 2 years ago. It will be examined during the procurement for cost, and only be used for a portion of the cluster if the cost advantage is very compelling.

Target Timeline

A best value procurement RFP was issued Jan 24 asking for quotes on a system with an aggregate performance of 8 TFlops, with options for additional racks. Bids will be due by Feb 29 (5 weeks), and an award of ~$800K (rounded to a whole number of racks) will be made by the middle of March, with a fixed price option for up to 3 more racks to be exercised within 30 days of delivery of the first award. The first systems should be installed in mid May, and the option exercised no later than mid June.

Note: the definition of best value for selecting the winning bid will incorporate the options yielding a system of size ~$1M, even if the options are not in the end selected.

Accelerated Cluster

Current Resources

USQCD has the following NVIDIA Fermi based GPU resources:

- 360 GTX480 / GTX580 configured as quad GPU systems (JLab)
- 144 C2050 / M2050 configured as quad GPU systems (JLab)
- 152 C2050 configured as dual GPU systems (Fermilab)

These 3 different configurations are targeted at different applications:

- The 360 GTXx80 GPUs are gaming cards best used for inverters only.
- The 296 x2050 GPUs all have ECC memory, and are suitable for any application for which CUDA software is available.
- The quad x2050 cluster at JLab is better for applications which are either inverter heavy with a mix of non-inverter accelerated sections, or any applications in which the bulk of the run-time has been ported to the GPU (i.e. applications where Amdahl's Law – the impact of code still running unaccelerated on the CPU – isn't a major constraint on performance.
- The dual C2050 cluster at FNAL has twice as many CPUs per GPU, and so can handle applications which have more of the execution time still resident on the CPU. This cluster also has more I/O bandwidth per GPU, and so provides better scaling to larger GPU counts.
- 8 of the JLab nodes (32 GPUs) are dual rail QDR, and so also support higher I/O bandwidth per GPU.

Future Requirements and Benchmarking

Going forward, we anticipate that a larger fraction of applications will be running with more than the inverters on the GPU, which leads to a need to grow the deployment of ECC enabled systems compared to non-ECC (gaming) cards. What is not yet clear is whether enough additional code will have moved to the GPU to allow quad GPUs to deliver better price/performance than dual GPUs for the expanding suite of applications that will exploit GPUs.

To answer this question it will be necessary once again to refine the set of benchmarks that will be used to drive this decision. A survey of then current usage and anticipated future usage will be made in Q2 2012, possibly incorporating input from the 2012 call for proposals process. The Scientific Program Committee will be asked to include a request for information about the anticipated readiness of accelerated software for the July 2012 – June 2013 running period, and, where possible, the observed performance on existing dual and quad GPU configurations. Users will also be invited to provide stand-alone versions of their software in a form suitable for use as benchmarks for the procurement. Scaling considerations will be probed so as to understand how much bandwidth each GPU or node will require.

There is one other accelerator architecture that will be watched over the coming six months, and that is the Intel MIC accelerator. There is not much public information available, so no details will be presented in this document.

System Size, Power & Cooling

Assuming optimistically $2K per accelerator and $4K for the host, networking, etc., a quad GPU system would cost approximately $12K, and a dual GPU system approximately $8K. Assuming a $1M procurement, we end up with something in the range of 250 to 332 GPUs. This leads to a power estimate of around 80 KW, and 6-8 racks.

The combined power draw for the two clusters will bump up against the current maximum installed cooling at Jefferson Lab, but the lab will be expanding the cooling in the room by an additional 30 tons (> 90 KW) in Q3 2012, ahead of a delivery of the second cluster. The expansion plus the decommissioning of 7n will leave more than enough power and cooling for 2012-2013.

Infiniband Network

In order to support scaling to high performance, perhaps including small scale configuration generation, the accelerated nodes are expected to need very high bandwidth, exceeding a single QDR link. There are two viable approaches: (1) incorporate 2 QDR cards per node, and (2) deploy FDR.

FDR cards are PCI gen 3 x8 cards. The next generation of Intel chipsets will have 40 lanes of PCI gen 3, thus enough for 2 GPUs and an FDR card. Quad GPU systems could incorporate dual chipsets or PCI multiplexer chips.

The size of the fabric for this cluster will be 80 – 124 nodes. There are no medium sized FDR switches, so the fabric would need to be built from 36 port leaf switches. Six leaf switches plus 3 switches used as the spine would handle 106 nodes (leaving 2 ports open for a file system uplink). If this turned out to be too small, then 7-8 leaf switches plus 4 spine switches could support up to 126-142 nodes.

Because the use of leaf switches shrinks the length of the longest cable, the entire 7 rack configuration could be done with 5 meter copper cables (the maximum for FDR).

Target Timeline

For a typical cluster at Jefferson Lab, the time from RFP to delivery is typically 12 weeks. To allow some contingency, we will aim at 14 weeks, thus June 15 for the RFP.

This will also be a best value procurement, possibly with the number of accelerators per node still not specified, as was the case at FNAL for 2011. Best value will be defined to include all options needed to spend out the available funds.

An RFI will be issued in early May, alerting vendors of the upcoming procurement, and soliciting feedback to ensure the RFP doesn't exclude any useful options.

A best value procurement RFP will be issued by Jun 15 for all remaining hardware funds after any options on the non-accelerated cluster have been exercised. Bids will be due by July 13 (4 weeks), and an award will be made by Aug 2, with an 8 weeks delivery requirement. Commissioning might extend into November.

For 2012-2013 allocations, the accelerated cluster will be allocated in a separate call for proposals, since not enough information is available today for that purpose.

File Servers

File servers will be procured just-in-time, in 2 phases matched to the deployments of the two clusters. The first phase will be procured roughly at the same time as the Infiniband cluster (and is now underway). The second phase will be procured roughly at the same time as the accelerated cluster, and will take into account user disk allocation requests for the 2012-2013 allocation year.

The added capacity will be in the form of additional nodes to be configured as Lustre OSSes (Object Storage Servers), most likely NAS devices (Network Attached Storage) with 3 TB SATA drives, a recent generation RAID controller, and QDR Infiniband. Assuming approximately $250 / TB, the budget should support a deployment of scale 400 TB for the year, more than doubling the disk resource as the computing resources are also doubled. The servers will be held in a single rack, adjacent to an existing rack with a QDR switch with ample ports available.

Estimated Timeline

Jan 15	RFP for non-accelerated cluster (done)
Mar 23	Award for non-accelerated cluster
May 25	Installation of non-accelerated cluster
June 1	Decision made on split between non-accelerated and accelerated cluster, and decision made on acceptable accelerator choices
June 8	Options exercised for non-accelerated cluster
June 15	RFP for accelerated cluster
July 1	Production running on first phase of non-accelerated cluster
Aug 2	Award for accelerated cluster
Sept 1	Production running on full non-accelerated cluster
Sept 30	Delivery of accelerated cluster
Nov 1	Production running on accelerated cluster