# FY2013 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 1.1

April 24, 2013

PREPARED BY:
Don Holmgren, FNAL

CONCURRENCE:

Apr 24, 2013

______________________ __________
William N. Boroski Date
LQCD-ext Contractor Project Manager

## FY13 Acquisition Plan
## Change Log

| Revision No. | Description | Effective Date |
|---|---|---|
| Rev 1.0 | Full version with BG/Q and conventional cluster details. | 10/17/2012 |
| Rev 1.1 | For release at the May 2013 project progress review. | 4/24/2013 |
| | | |
| | | |
| | | |

# LQCD-ext FY2013 Acquisition Plan

*Version 1.0*

## Summary

In FY2013 the LQCD-ext project will deploy the following hardware at Fermilab (FNAL):

1. An Infiniband cluster
2. Disk servers

and the following hardware at Brookhaven (BNL):

1. One half-rack of IBM BlueGene/Q (BG/Q) hardware
2. Disk servers

The split between BG/Q and cluster hardware was finalized in August 2012, following the process outlined in the project's acquisition strategy document. According to a quotation from IBM obtained by BNL, the half-rack of BG/Q will cost approximately $1.317M, including BNL G&A. The Infiniband cluster at Fermilab will be purchased with the approximately $850K remaining in the FY13 project budget for new hardware. Disk space will be approximately an additional $100K at each of the two sites.

The timing for both sets of procurements will potentially be affected by the FY13 continuing budget resolution. As of the start of FY13, the project assumes that funds for operations will be released gradually throughout the year until the CR is resolved, but that all budgeted equipment funds will be available by late in the first quarter of FY13, as communicated by the DOE HEP and NP program managers. Accordingly the project will procure the cluster and BG/Q hardware as soon as possible at each site. Note that this is a change in the plan presented at the 2012 project progress review, which stated that the cluster hardware purchase would be delayed until late in FY13 so that FY13 and FY14 procurements could be combined across the FY13-FY14 boundary; such combined procurements are more cost efficient for the project, as less manpower is used, better leverage is available on a larger buy to obtain good pricing from vendors, and Fermilab G&A costs to the project are lower because they are capped on the first $0.5M of any purchase order. However, given the budgetary uncertainties in this fiscal year, purchasing all hardware as early as possible optimizes the science that can be delivered using the new hardware against the risk that funds could be rescinded or significantly delayed.

## BG/Q and Disk Server Acquisitions (BNL)

### *Justification and configuration*

In the second half of calendar 2012, the next generation of the IBM Blue Gene series, Blue Gene/Q (BG/Q), became available for purchase. This machine, like its BG/L and BG/P predecessors, should perform very well on LQCD applications and is competitive with commodity cluster and accelerated cluster hardware. Significant Blue Gene expertise resides at BNL along with existing infrastructure capable of supporting an additional full rack.

Each BG/Q Compute node has an 18 core chip using a 64 bit Power PC A2 processor cores running at 1.6 GHz. with 16GB of memory. There are 32 compute nodes in each node card and 16 node cards in a half rack. Each node card is water cooled.

*Power and Space Estimates*

The half-rack of BG/Q will be housed in the QCDOC Supercomputer room in Building 515 of the Brookhaven National Lab Information and Technology Division. The maximum power needed is 50kw for the half rack plus 15kw for the disk storage system. Brookhaven has agreed to provide 38 sq ft. of computer floor space, power and cooling for the half-rack BG/Q system, its cooling distribution unit, and 1 Pbyte of disk storage.

The space will be available early 2013 with delivery and installation of the BG/Q expected in the late March early April 2013 time frame

*Schedule*

A purchase requisition will be entered into the BNL requisition system and will include the quotations from IBM dated June 7, 2012 and a completed justification for noncompetitive procurement (JNCP) form designating IBM as the sole source vendor for this system. It is expected that the purchase order will be entered prior to December 1, 2012 with an expected issue date to IBM late February 2013.

It is expected that the BGQ will be delivered to Brookhaven late March 2013 with an early April 2013 installation date.

Acceptance testing will consist of running DWF evolution of $32^3 \times 64 \times 24$ MDWF+ID strong coupling ensemble with $m\,\pi = 140$ MeV for 1.5 days, with 100% reproducibility testing without problems.

It is expected that the BGQ allocations will be available to the USQCD at large on July 1, 2013

*Storage*

A purchase requisition will be entered for additional storage to support the BGQ project. The storage requirements will be an Infiniband-connected storage array of size 0.5 to 1.0 PByte, depending upon pricing, The requisition will be entered into the purchasing system at BNL in June 2013 (subject to funds availability due to the continuing resolution), with an expected delivery date of August 2013.

## Infiniband Cluster and Disk Server Acquisitions (FNAL)

The LQCD-ext project decided to split the FY13 hardware budget between a half-rack of BG/Q at BNL, and either a conventional Infiniband cluster or a GPU-accelerated cluster at FNAL. See the "FY13 Alternatives Analysis for the Lattice QCD Computing Project Extension" document of August 20, 2012 for details.

In September 2012, the project determined that the hardware choice which would best optimize the portfolio of hardware for science production and have the lowest risk would be a conventional Infiniband cluster based on either Intel Sandy Bridge or AMD Abu Dhabi processors. The most recent version of the NVIDIA Tesla GPU, based on the "Kepler" architecture, was not yet readily available and was to be used in the Jefferson Lab (JLab) "12K" cluster, to be delivered in October or November. The anticipated acquisition start for the FY13 cluster in November would therefore occur well before the project had gained experience with the new "K20" Tesla cards or before LQCD software would have been tuned and in production. Further, based on the 2012-2013 USQCD allocations, demand for conventional cluster cycles was as great as or greater than for GPU cycles. Finally, the budget available ($850K) was insufficient to split further into an accelerated and a conventional cluster purchase.

The likely hardware candidates for this cluster are servers based on either dual socket or quad socket Intel Sandy Bridge processors, or quad socket AMD Abu Dhabi processors. Sandy Bridge processors have better memory bandwidth per socket and therefore higher performance on LQCD codes, as confirmed by benchmarking by the project, but are priced considerably higher per processor as determined during the JLab 12s Sandy Bridge-based cluster acquisition and the prior Ds AMD-based cluster at Fermilab.

*Power and Space Estimates*

The FY13 cluster will be housed in computer room A of the Grid Computing Center (GCC-A) at FNAL. The maximum power per rack that can be cooled in GCC-A is 10 KW. For quad-socket systems, which typically draw about 600 W per system, this limits each rack to no more than 16 servers. At roughly 60 GFlops sustained per quad-socket server, an 8.5 TFlops cluster will require at least ninth rack, plus a tenth rack for networking equipment. FNAL has agreed to provide up to sixteen racks of space, power, and cooling in GCC-A for this purchase. The precise system count will depend on brand of processor and number of sockets per host.

Because this cluster will be in a different computer room than the other FNAL LQCD clusters and their Lustre filesystem (both housed in GCC-C), either multiple 10 gigE links or multiple QDR links will be used to provide access to Lustre. FNAL has used a single 10 gigE link previously to provide access to Lustre for an older cluster "Kaon" in another building (Lattice Computing Center, or LCC), and has successfully tested routing Lustre traffic over multiple 10 gigE links.

*Memory Bandwidth*

LQCD is always memory bandwidth constrained, and the strength of both Intel Sandy Bridge and AMD Abu Dhabi 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% over the prior generations (from 3 controllers in Nehalem/Westmere 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. Benchmarking as part of vendor proposals will be necessary to ascertain the most cost effect CPU and bus speeds for both Intel and AMD.

*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 and Ethernet Networks*

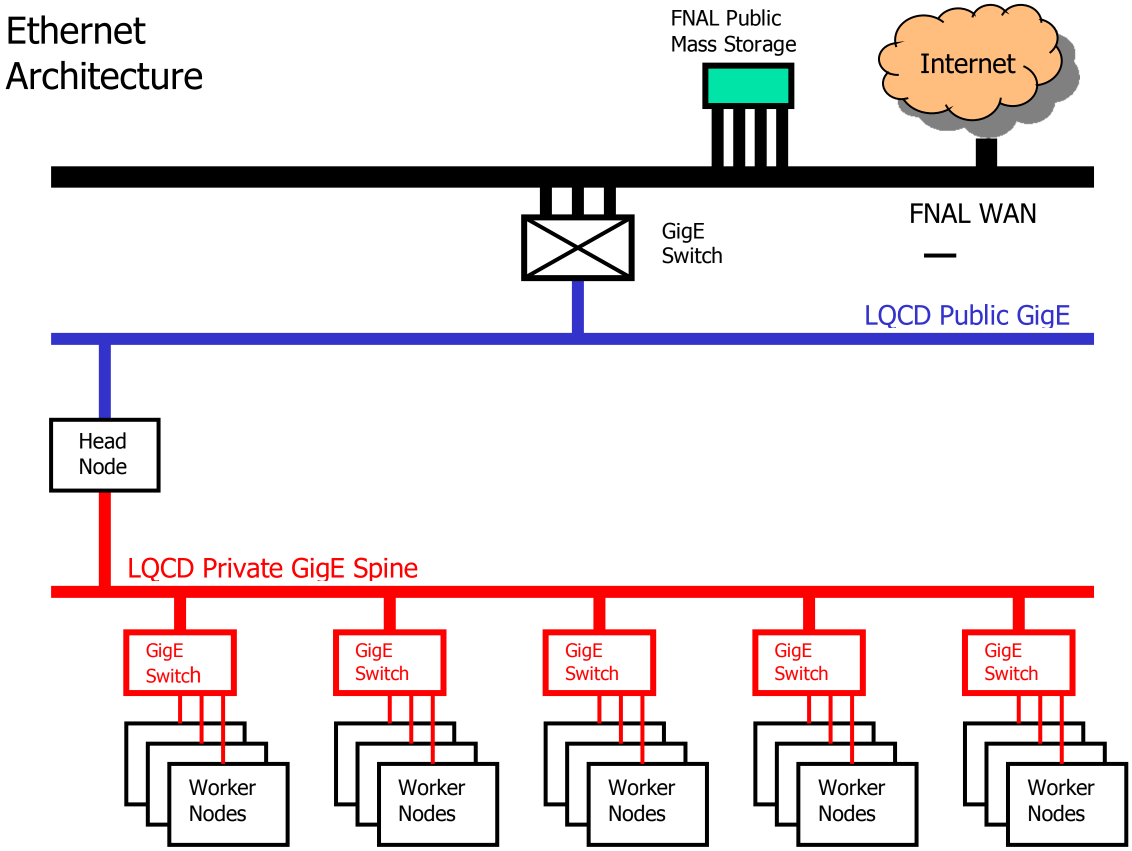
QDR Infiniband, at 40 Gbps, will be adequate for this cluster, and will likely remain less expensive than the newer FDR (56 Gbps). Based on experience with the FNAL Ds cluster (quad-socket AMD), 2:1 oversubscription is well tolerated for LQCD applications. Given that Intel dual-socket systems will have similar throughput to the four-socket AMD systems, 2:1 oversubscription will also be well-tolerated. For quad socket Intel systems QDR may need to be deployed without oversubscription; we will work with Intel to gain access to a remote benchmarking cluster to determine sensitivity to oversubscription.

The cluster will access the FNAL Lustre parallel file system over Infiniband. A gigabit Ethernet network with leaf-and-spine design (*i.e.* top-of-rack leaf switches uplinked to a single spine switch) will be used to provide access to NFS-exported home and local software repository directories. User access to the cluster will be through a head node connected to the Fermilab WAN and thereby to the Internet. All worker nodes on the cluster will use private Ethernet and Infiniband networks, with the dual-homed head node providing access via a batch scheduler (Torque). An dedicated gigabit Ethernet network will be used for remote management using IPMI.

*Ethernet Network Architecture Diagram and Description*

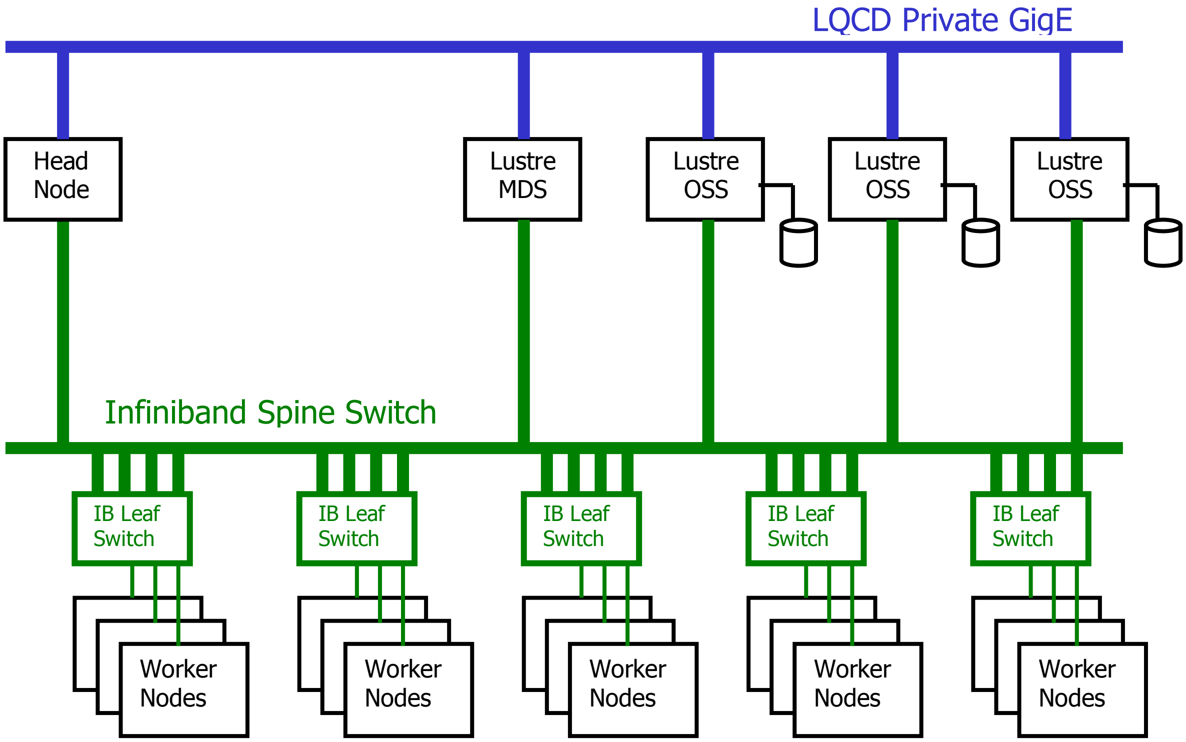
The diagram below shows the Ethernet network architecture of the Ds cluster installed at Fermilab in FY2010 and FY2011. A similar architecture will be used for the FY2014 cluster. Public and private gigE networks will be used, as shown in the diagram. The public gigE network will connect via a Cisco switch to FNAL's wide area network via a set of four channeled gigabit Ethernet connections. The FY2014 facility will access the mass storage facility at Fermilab via the laboratory's WAN. Within the mass storage facility are multiple *tape mover nodes*, each attached to either an LTO-4 tape drive or a Cisco T10K-C tape drive.

Users of the existing FNAL and JLab clusters login to the cluster head (login) node; the scheduler (*Torque* plus *Maui*) runs on either this node or another dedicated node. Approximately 10 Tbytes of local disk are attached to the login node.



The worker nodes will be connected via gigabit Ethernet leaf switches with gigabit Ethernet uplinks to a private spine gigabit Ethernet switch. The head node will communicate via this private network with the worker nodes. This network is used for login access to the worker nodes by the scheduler (using *rsh*). Each worker mounts via NFS the /home and /usr/local directories from the head node. Binaries are generally launched from the /home directory. Each worker node has considerable (120 Gbytes or greater) local scratch space available. High performance I/O transfers to and from the worker nodes and the head node utilize the Infiniband network (see drawing below).

## Infiniband Architecture



*Infiniband Architecture Diagram and Description*

The diagram above shows the Infiniband architecture used on the Fermilab Ds cluster. A similar architecture will be used on the FY2014 cluster.

On the FY2014 cluster, a leaf and spine approach will be used. Each set of worker nodes will be connected to a 36-port leaf switch. Multiple links connect each leaf switch to a central spine switch stack consisting of multiple 36-port edge switches. The Infiniband fabric will be used for internode communications for LQCD applications via MPI (*mvapich* and *OpenMPI* versions will be available). The Infiniband fabric will also be used for high performance file I/O via TCP, using IPoIB.

*Software Deployment and Other Integration Tasks*

To bring the FY2014 cluster into production, the following integration tasks will be necessary (order may vary from that shown):

1. Prepare system installation images for worker nodes (Scientific Linux). These images will include the Infiniband software stack (OpenIB, or commercial) as well as the SciDAC LQCD shared libraries.

2. Install system images on all worker nodes.
3. Unit test worker nodes. These tests will include memory tests, multiple reboot and power cycle tests, disk tests, and LQCD single node application testing and performance verification.
4. Unit test worker racks. This will require configuring the Infiniband fabric within each rack. During these tests, each rack will be operated as an independent cluster. The tests will include LQCD multinode application testing and performance verification.
5. Integrate worker racks. This requires the interconnection of the individual racks to the Infiniband and gigabit Ethernet spine fabrics, and the configuration of the Infiniband subnet manager and monitoring facilities.
6. Configure IPMI facilities on all worker nodes; this includes initializing BMC network parameters (IP addresses, subnet masks, ARP and gratuitous ARP configuration).
7. Test IPMI facilities on all worker nodes.
8. On head node, deploy commercial compilers (Intel, Portland Group as requested by user community).
9. On head node, build and deploy SciDAC libraries.
10. On head and worker nodes, deploy SciDAC common runtime environment.
11. On head and worker nodes, deploy and configure batch system (*Torque* plus *Maui*).
12. On head and worker nodes, create authorized user accounts.
13. Test batch system.
14. Test LQCD applications.

*Schedule*

As with prior cluster procurements at FNAL, we will use a Request for Information (RFI) to solicit vendor input to our design, and then a Request for Proposal (RFP) to solicit bids. The award will be made using a best-value process, with price-performance, power efficiency, space efficiency, lifecycle cost, and vendor experience used to determine the proposal offering the best value.

We will follow this schedule:

- Benchmark processor alternatives (Intel Sandy Bridge, 2P and 4P, AMD Abu Dhabi 4P) – Aug-Dec 2012
- Issue a Request for Information – mid-November 2012
- Issue the Request for Proposal – by mid-January 2013
- Evaluate proposals and award purchase order – by Mar 8, 2013
- Hardware received and integrated in GCC-A – by Apr 30, 2013
- Acceptance testing and friendly user period – begin May 2013
- Release to production – by mid-June 2013

## File Servers

File servers to expand the FNAL Lustre filesystem will be procured at roughly at the same time as the Infiniband cluster. The added capacity will be in the form of additional nodes that will be configured as Lustre OSSes (Object Storage Servers), most likely disk servers with 3 or 4 TB SATA drives, a recent generation RAID controller, and dual Infiniband (QDR for access from the FNAL Ds and Dsg clusters, and DDR for access from the JPsi cluster). Assuming approximately $125 / TB, the budget will support a deployment of scale 256 TB for the year, increasing FNAL Lustre disk resource by 40% as the computing resources are increased by approximately 20%. The servers will be held in a single rack in GCC-C, adjacent to existing Lustre disk arrays.