FY2014 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

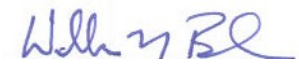
for the
U.S. Department of Energy
Office of Science
Offices of High Energy and Nuclear Physics

Version 1.0

October 1, 2013

PREPARED BY:
Don Holmgren, FNAL

CONCURRENCE:



October 1, 2013

_____ _____

William N. Boroski
LQCD-ext Contractor Project Manager

Date

FY2014 Acquisition Plan
Change Log

Revision No.	Description	Effective Date
Rev 1.0	Initial version.	10/1/2013

LQCD-ext FY2014 Acquisition Plan

Version 1.0

Summary

In FY2014 the LQCD-ext project will deploy the following hardware at Fermilab (FNAL):

1. A conventional Infiniband cluster
2. A GPU-accelerated Infiniband cluster
3. Disk servers

In addition, the project will deploy disk servers at Jefferson Lab (TJNAF).

The decision was finalized in August 2013, following the process outlined in the project's acquisition strategy document. The clusters at Fermilab will be purchased with the $1.95M FY2014 project budget for new hardware. Disk space will be approximately $145K at FNAL, and $50K at TJNAF.

The timing for both sets of procurements will potentially be affected by the FY2014 continuing budget resolution (CR). As of the start of FY2014, the project assumes that, similar to FY2013, 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 FY2014, as communicated by the DOE HEP and NP program managers. Accordingly the project will procure the clusters as soon as practical at FNAL. The disk servers, budgeted on operations funds, will be procured in calendar 2014 as funds become available.

Conventional and GPU-Accelerated Clusters and Disk Server Acquisitions

The LQCD-ext project, as discussed in the "FY2014 Alternatives Analysis for the Lattice QCD Computing Project Extension" document of August 20, 2013, will split the FY2014 hardware budget between a conventional Infiniband cluster and a GPU-accelerated cluster at FNAL. The precise budgetary split between the two types of clusters will be determined after further study, including determining for various hardware the approximate costs and performance on LQCD software, and understanding the probable demand from USQCD users in the next two years for either type of cluster. As presented at the project's annual review in May, 2013, the fraction of the budget planned for the conventional cluster will be 40% to 60% of the total budget of $1.95M, with the remaining funds used for the GPU-accelerated cluster.

The likely hardware candidates for the conventional Infiniband cluster are servers based on either dual socket or quad socket Intel Xeon "Sandy Bridge" or "Ivy Bridge processors", or quad socket AMD Opteron "Abu Dhabi" or "Warsaw" processors. For Intel, some of the "Ivy Bridge" processors are expected to have better memory bandwidth than the "Sandy Bridge" processors used on the TJNAF "12s" cluster. This will be confirmed by benchmarking performed by the project during early FY2014. The AMD "Warsaw" processors may offer better power efficiency and slightly higher memory bandwidth than the "Abu Dhabi" processors used on the FNAL "Bc" cluster.

The likely hardware candidates for the GPU-accelerated cluster are servers based on Intel Xeon "Sandy Bridge" or "Ivy Bridge" processors, with NVIDIA K20X or K40 GPUs. Intel-processor-based hosts have better I/O bus (PCIe) performance than AMD-processor-based hosts. K40 GPUs are expected to support the PCIe Gen3 implementation available on "Ivy Bridge" (the K20-era GPUs did not interoperate correctly with the PCIe Gen3 implementation on "Sandy Bridge" processors and so used Gen2 speeds); PCIe Gen3 is not available on AMD-based hosts. K40 GPUs are expected to have larger on-board memory and to be somewhat faster than K20X parts. Peer-to-peer PCIe communications are expected to have better performance on the "Ivy Bridge" implementation; this performance is relevant to LQCD problems that run across multiple GPUs. The project will work with Intel and NVIDIA to obtain hardware and to benchmark LQCD codes to determine the performance impact of these hardware changes. There will be a tradeoff between cost and performance.

Because of the G&A policy at FNAL, overhead costs to the project will be minimized if the two FY2014 cluster purchases utilize the same purchase order. The project will therefore structure the specifications as given in the Request for Information and Request for Proposal documents so that vendors will provide bids for both clusters in the same proposal.

Power and Space Estimates

The FY2014 clusters will be housed in computer room C of the Grid Computing Center (GCC-A) at FNAL, utilizing space to be made available from the retirement of the "JPsi" cluster purchased by the LQCD project in FY08 and FY09. The maximum power per rack that can be cooled in GCC-C is 14 KW. Following the removal of JPsi, a total of 31 rack positions will be available. The precise system count for each cluster will depend on brand of processor, number of processor sockets per host, and number of GPU accelerators per host. The GPU count per host will very likely be four, but higher counts are feasible and will be examined for price/performance.

Memory Bandwidth

LQCD is always constrained in whole or part by memory bandwidth, and the strength of Intel "Ivy Bridge" processors is their increased memory bandwidth compared with "Sandy Bridge" and AMD "Abu Dhabi". "Ivy Bridge" chips have 4-channel memory controllers supporting up to DDR3-1866 memories. For Intel this is an increase of 16% over the prior generation (4-channel DDR3-1600). AMD "Abu Dhabi" and likely "Warsaw" will support up to DDR3-1866, but have only 2 channels per processor socket; a four-socket AMD-based system will have comparable memory bandwidth to a two-socket Intel-based system, but if cost per socket is half that of Intel then AMD systems will be competitive (as was the case for the AMD-based "Bc" cluster at FNAL in FY2013). 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, for the conventional cluster purchase inverter benchmarks will be used to measure performance and price/performance for each of the DWF, HISQ, and Clover actions. Vendors will be provided single-node binaries that will measure performance for each of these actions using all available cores. Fermilab staff will perform benchmarking on vendor-supplied clusters, with problems using the three actions run at a variety of sizes. These weak-scaling measurements will be used to estimate scaling factors between the single-node measurements performed by vendors and the multi-node problems (128 and greater MPI ranks) that will be typical for production once the cluster is deployed and released to operations.

For the accelerated cluster purchase, vendors will be provided single node benchmark binaries that will measure performance using all available GPUs. The benchmarks will include DWF, HISQ, and Clover action LQCD code, plus synthetic benchmarks such as those in the SHOC suite (originally from the Oak Ridge National Laboratory). The synthetic benchmarks will be used to measure PCIe performance.

Infiniband and Ethernet Networks

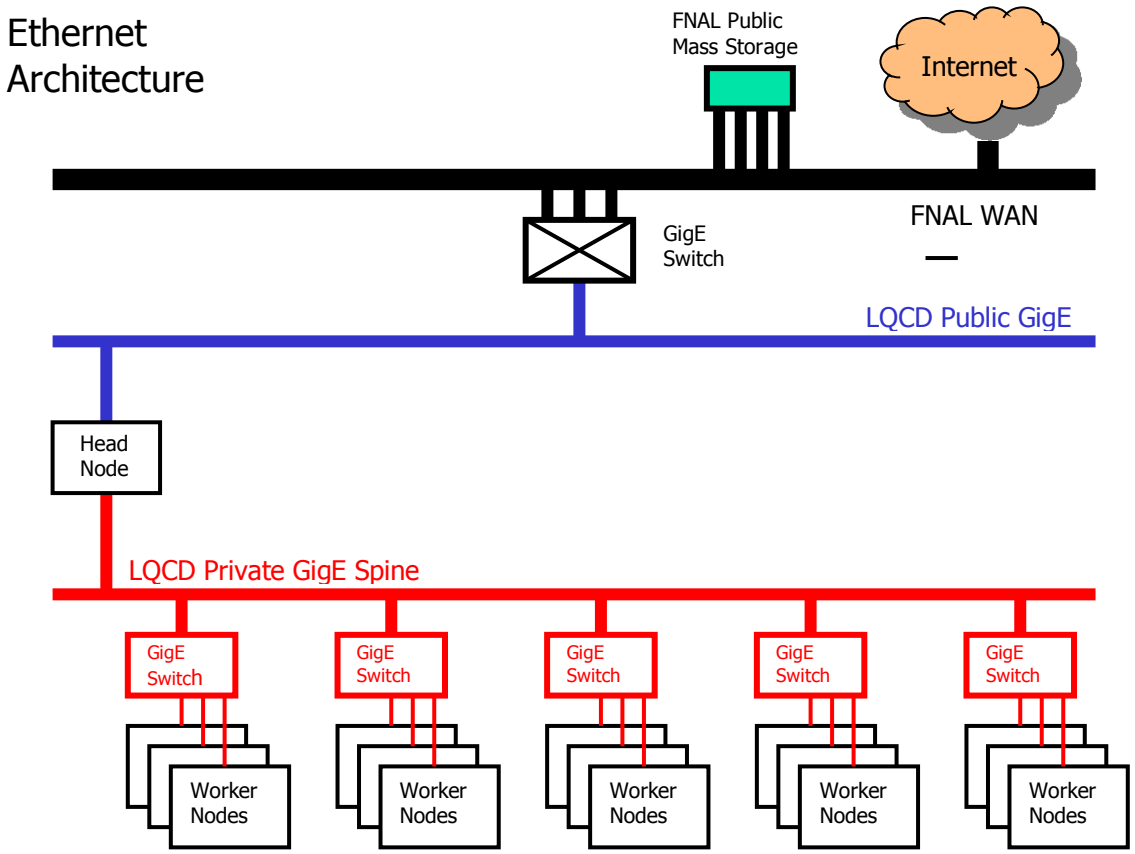
QDR Infiniband, at 40 Gbps, will be adequate for the conventional cluster, and will likely remain less expensive than the newer FDR (56 Gbps). Based on experience with the FNAL Ds cluster (quad-socket AMD), 2:1 oversubscription is well tolerated for LQCD applications where the throughput per host system is as high as 60 GFlop/sec on the various inverters discussed under *Benchmarking* above. The JLab quad-GPU/host K20 cluster "12k" uses non-blocking FDR Infiniband. We will work with USQCD GPU experts to understand whether non-blocking QDR is sufficient, based on data from "12k".

The clusters 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 on each cluster to provide access to NFS-exported home and local software repository directories. User access to the clusters will be through one or more head nodes connected to the Fermilab WAN and thereby to the Internet. All worker nodes on the clusters will use private Ethernet and Infiniband networks, with the dual-homed head node(s) providing access via a batch scheduler (Torque). A dedicated gigabit Ethernet network on each cluster will be used for remote management using IPMI.

Ethernet Network Architecture Diagram and Description

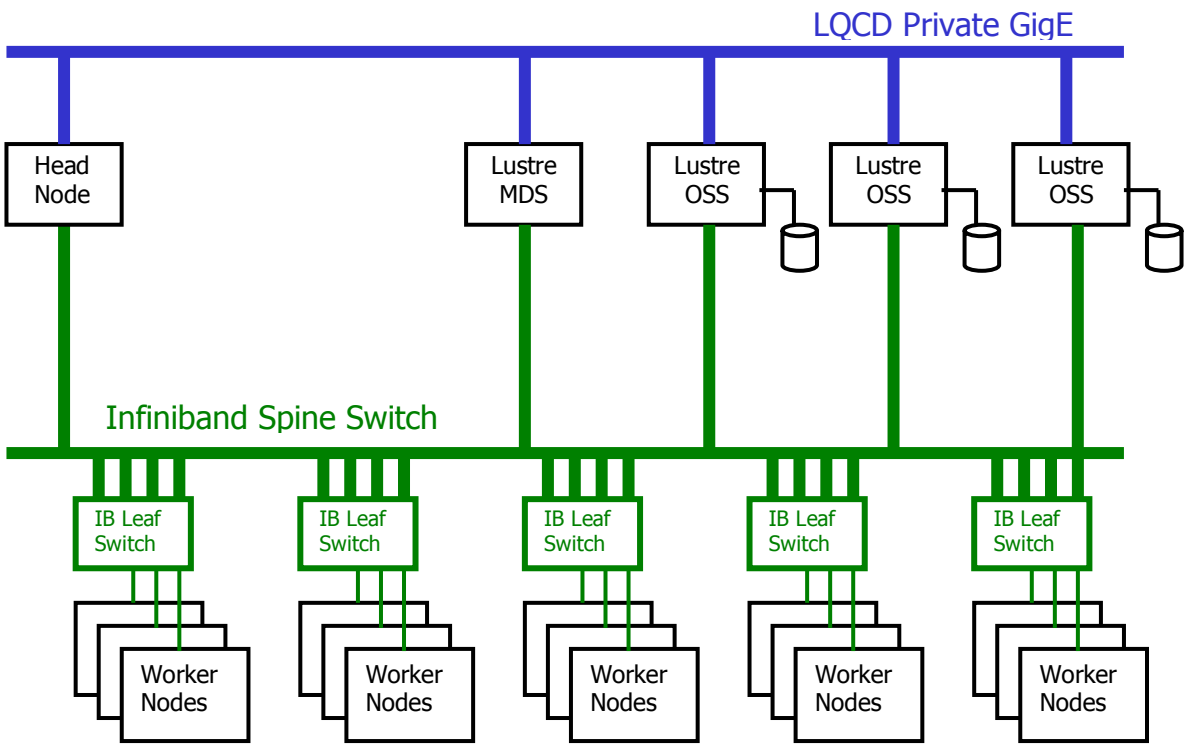
The diagram below shows the Ethernet network architecture of the Bc cluster installed at Fermilab in FY2013. A similar architecture will be used for each of the FY2014 clusters. Public and private gigE networks are used, as shown in the diagram. The public gigE network connects via a Cisco switch to FNAL's wide area network via a set of four channeled gigabit Ethernet connections. Access the mass storage facility at Fermilab occurs 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 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 connect via gigabit Ethernet leaf switches with gigabit Ethernet uplinks to a private spine gigabit Ethernet switch. The head node communicates 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 Bc cluster. A similar architecture will be used on each of the FY2014 clusters.

On the Bc cluster, a leaf and spine approach us used. Each set of worker nodes is 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. An Infiniband fabric is used for internode communications for LQCD applications using MPI (*mvapich* and *OpenMPI* versions are available to users). The Infiniband fabric is also used for high performance Lustre file I/O using native Infiniband protocols.

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).
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. On the GPU-accelerated cluster, single node applications testing and performance verification will be performed on both the host CPUs and on the installed GPUs.
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, build and deploy SciDAC libraries.
9. On head and worker nodes, deploy and configure batch system (*Torque* plus *Maui*).
10. On head and worker nodes, create authorized user accounts.
11. Test batch system.
12. 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 and Ivy Bridge, 2P and 4P, AMD Abu Dhabi 4P and if available, Warsaw 4P) – Aug-Dec 2013
- Issue a Request for Information – early- to mid-November 2013
- Issue the Request for Proposal – mid- to late-December 2013, or later if constrained by a continuing budget resolution
- Evaluate proposals and award purchase order – by Mar 5, 2014
- Hardware received and integrated in GCC-C – by Apr 30, 2014
- Acceptance testing and friendly user period – begin May 2014
- Release to production – by mid- to late-June 2014

File Servers

File servers to expand the FNAL Lustre filesystem will be procured at roughly at the same time as the Infiniband cluster, although procurement could be delayed by a continuing budget resolution. 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 QDR Infiniband. Assuming approximately $160 / TiB including G&A, the budget will support a deployment of scale 900 TiB for the year at FNAL, and 300 TiB at TJNAF.