

**Acquisition Strategy
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 1.6

August 17, 2009

PREPARED BY:
Don Holmgren, FNAL

CONCURRENCE:



William N. Boroski
LQCD Contractor Project Manager

August 17, 2009

Date

**Acquisition Strategy
Change Log**

Revision No.	Description/ Pages Affected	Effective Date
Revision 0.9	LQCD II plan draft drawn from LQCD FY08/FY09 Plan	Dec 30, 2008
Revision 1.0	Pre-ARRA version.	March 19, 2009
Revision 1.1	Combined FY2010/FY2011 buy. LQCD II → LQCD-Ext	April 13, 2009
Revision 1.2	Added performance basis, Section 508 and M-07-11 compliance, and cyber security sections.	April 15, 2009
Revision 1.3	Final Version for CD-1	April 15, 2009
Revision 1.4	Reorganize to present strategy and plan separately	June 18, 2009
Revision 1.5	New Version for CD-2 / CD-3	August 3, 2009
Revision 1.6	Expanded storage discussion	August 17, 2009

Table of Contents

Introduction.....	1
Previous LQCD Computing Project	1
Overview of LQCD-Ext Project Deployments	2
Design Considerations and Strategies for LQCD-Ext Deployments	3
Compute Nodes.....	3
High Performance Network	4
Service Networks	5
Network Plan	5
File I/O	5
Procurement Strategy.....	6
Acquisition Plan for FY2010/FY2011	7
Compute Nodes and High Performance Network	8
Ethernet Network Architecture Diagram and Description.....	8
Infiniband Architecture Diagram and Description	10
Software Deployment and Other Integration Tasks.....	11
Computing Room Facility for the FY2010/FY2011 Cluster	11
Schedule	12
Cost Basis.....	13
Performance Basis	15
Section 508 and OMB Memorandum 07-18 Compliance	16
Cyber Security	16

Introduction

The Lattice QCD Computing Extension Project (LQCD-Ext) develops and operates new and existing systems in each year from FY2010 through FY2014. These computing systems are deployed at Fermilab (FNAL), Jefferson Lab (JLab), and Brookhaven (BNL). During FY2010, the project will operate the 4.2 Tflop/s US QCDOC supercomputer at BNL, as well as the clusters developed during the last three years of the prior Lattice QCD Computing Project (FY2006-FY2009) at FNAL and at JLab. Table 1 shows the planned total computing capacity of the new deployments, and the planned delivered (integrated) performance. The integrated performance figures assume at the beginning of FY2010 18.1 Tflop/s of total capacity, equal to the 13.9 Tflop/s aggregate capacity of the existing clusters at JLab and FNAL, plus the 4.2 Tflop/s capacity of the US QCDOC at BNL. Note that in all discussions of performance, unless otherwise noted, the specified figure reflects an average of the sustained performance of domain wall fermion (DWF) and improved staggered (asqtad) algorithms.

	FY 2010	FY 2011	FY 2012	FY 2013	FY 2014
Planned computing capacity of new Deployments, Tflop/s	11	12	24	44	57
Planned delivered Performance (JLab + FNAL + QCDOC), Tflop/s-yr	18	22	34	52	90

Table 1 – Performance of New System Deployments, and Integrated Performance (DWF+asqtad averages used). Integrated performance figures use an 8000-hour year.

All LQCD-Ext Project cluster hardware procurements will utilize firm, fixed-price contracts. Cluster purchases will use contracts with vendors specializing in COTS hardware. The steady state operations of the project computing facilities are performed by the three host laboratories, each of which is a government-owned contractor-operated facility.

In each year of the project, the hardware that best accomplishes the scientific goals for LQCD calculations will be purchased. In FY2010 and FY2011, the project anticipates that a cluster based on COTS computer nodes plus Infiniband high performance networks will be the most effective. In FY2012, the IBM BlueGene/Q supercomputer will be an important hardware candidate. In all years, project personnel will also consider alternative hardware designs, such as custom-integrated COTS-based hardware (*e.g.* SGI Altix ICE systems) and conventional supercomputers (*e.g.* SiCortex systems).

In the rest of this document we first discuss the design considerations and strategies that we will use for all of the procurements of the LQCD-Ext Project. We then discuss the planned acquisition of a cluster in FY2010 and FY2011 by Fermilab. This document will be updated each year to concentrate on the upcoming hardware acquisition.

Previous LQCD Computing Project

From FY2006 through FY2009, the DOE High Energy Physics (HEP) and Nuclear Physics (NP) program offices funded the DOE Office of Science LQCD Computing

Project (SC LQCD). The total project cost of \$9.2M for SC LQCD funded the deployment of four clusters at Jefferson Lab and Fermilab, plus the operations of these clusters, the QCDOC supercomputer at Brookhaven, and several SciDAC LQCD clusters at JLab and FNAL acquired in 2003 through 2005.

The clusters developed during SC LQCD were as follows:

- “6n”, at JLab in 2006, based on single-socket dual-core Pentium processors and single-data-rate Infiniband
- “Kaon”, at FNAL in 2006, based on dual-socket dual-core Opteron processors and double-data-rate Infiniband
- “7n”, at JLab in 2007, based on dual-socket dual-core Opteron processors, upgraded to quad-core processors, and double-data-rate Infiniband
- “J/Psi”, at FNAL in 2008 and 2009, based on dual-socket quad-core Opteron processors and double-data-rate Infiniband

The “J/Psi” cluster was procured using the funds from both FY2008 and FY2009. The FY2008 piece of “J/Psi” was awarded late in the fiscal year under a purchasing contract that allowed, via an option, additional compute nodes and network hardware of the same configuration to be purchased in the first half of FY2009. The FY2008 portion of the cluster was released to physics production at the beginning of January, 2009, and the FY2009 portion from the exercise of the purchase option was released to physics production in mid-April of 2009.

By executing a combined purchase, a single request for information (RFI) and a single request for proposal (RFP) were used, reducing project labor costs, laboratory labor costs, and the overhead (G&A) charged to the project. A similar combined procurement will be used for the FY2010 and FY2011 hardware deployments of the LQCD-Ext project.

Overview of LQCD-Ext Project Deployments

As the LQCD-Ext project begins in FY2010, based on the prior project (SC LQCD) the most effective hardware for the calculations performed on the existing SC LQCD project compute resources in FY2009 will be commodity clusters built using Intel or AMD x86_64 processors and an Infiniband interconnect. We predict this to be the case in both FY2010 and FY2011, and are proposing a combined cluster purchase at Fermilab similar to the SC LQCD FY2008/FY2009 purchase of “J/Psi”.

In FY2012, the new generation of the IBM BlueGene series, BlueGene/Q, is scheduled to be available for purchase. This machine, like the BG/L and BG/P predecessors, should perform very well on LQCD applications and will be competitive with commodity cluster hardware. The LQCD-Ext project will therefore evaluate BG/Q for deployment at BNL in FY2012. Significant BlueGene expertise resides at BNL. If other hardware, such as x86_64 clusters, is determined to be more effective in FY2012, the project will instead deploy that alternative.

In FY2013 and FY2014, the project will procure one or two additional systems, using the most scientifically effective hardware as determined by the anticipated usage.

All procurements will be performed by the host laboratory chosen for the particular hardware deployment. Such purchases will utilize firm, fixed-price contracts. The typical sequence for new deployments will be:

1. In consultation with USQCD community (Executive Committee, Scientific Program Committee) determine usage profile for new deployment (*e.g.*, distribution of job types and sizes, file I/O requirements)
2. Complete preliminary design
3. Issue a Request for Information (RFI) to likely vendors
4. Evaluate the RFI responses and complete a final design
5. Obtain host laboratory purchase approvals via the local requisition process
6. Issue a Request for Proposal (RFP) to likely vendors
7. Evaluate RFP responses and award purchase contract
8. Approve sample node and sample scalable unit (rack)
9. Test and approve vendor-integrated final system
10. Operate final system in “friendly user” mode and tune the configuration
11. Release the final system to users

Design Considerations and Strategies for LQCD-Ext Deployments

Compute Nodes

Lattice QCD codes are floating point intensive, with a high bytes-to-flops ratio (1.45 single precision, 2.90 double precision for SU(3) matrix-vector multiplies). When local lattice sizes exceed the size of cache, high memory bandwidths are required.

As of the beginning of 2009, available commodity processors with the greatest memory bandwidths are Intel x86_64 processors with 1066 and 1333 MHz (effective) front side buses (Xeon “Woodcrest” and “Clovertown”, Pentium “Conroe” and “Kentsfield”) and the AMD Athlon64, AthlonFX, and Opteron processors. The Pentium, Athlon64, and AthlonFX processors can only be used in single processor systems. The Xeon and Opteron processors can be used in dual and quad processor systems. The total cost of quad processor systems of both types, including the cost of the high performance network, exceeds the cost of two dual processor systems with network. At the current cost of Infiniband and competing high performance networks, quad processor systems are not as cost effective as single or dual processor systems.

Since late 2006, Intel and AMD have switched all new processors of relevance to lattice QCD to dual or quad core. The JLab “7n” and Fermilab “J/Psi” clusters purchased and deployed in 2007 and 2008, respectively, use quad core processors; both use motherboards that accommodate two quad-core Opteron “Barcelona” processors. Lattice QCD production on these clusters has shown that quad core processors scale very well on MPI jobs when the cores are treated as independent processors. Dual and quad core processors typically have lower clock speeds than the older analogous single core processors; however, the degree of scaling on MPI jobs is sufficient to make these

processors a more cost effective choice. Roadmaps from both Intel and AMD indicate that all forthcoming designs will be multicore.

In 2007, all commodity dual processor Xeon motherboard designs used a single memory controller to interface the processors to system memory. As a result, the effective memory bandwidth available to either processor was half that available to a single processor system. Opteron processors have integrated memory controllers and local (to the processor) memory buses, with a high-speed link (HyperTransport) allowing one processor to access the local memory of another processor. This NUMA (Non Uniform Memory Access) architecture makes multiprocessor Opteron systems viable for lattice QCD codes. In the 2007 and 2008 LQCD project acquisitions of the “7n” and “J/Psi” clusters, multiprocessor Opteron system were chosen, as this was the most cost effective design.

In 2006, Intel began selling dual processor systems with dual independent memory buses connecting the processors to a memory controller in turn connected to multiple DIMM channels. This memory subsystem is based on FBDIMM (“fully buffered DIMM”) technology. To date, this technology has not proven to deliver as much memory bandwidth as the integrated memory controllers on AMD Opteron systems. However, in the spring of 2009 Intel introduced a new generation of chipset and processor. The processors are fabricated using the same 45-nanometer process used in 2007 and 2008 for the Intel “Penryn” Xeon processor family.

The new Intel generation introduced in 2009 has code-names “Nehalem” for the processor family, and “Tylersburg” for the chipset family. “Nehalem” incorporates a new integrated memory controller design and uses DDR3 memory rather than FB-DIMMs. Improvement of sustainable memory bandwidth over previous Intel dual processor designs are 6 to 7-fold. These improvements greatly impact the performance of LQCD code.

We have found that hardware management features such as IPMI minimize the operating costs of commodity clusters. We will choose systems based upon motherboards that support out-of-band management features, such as system reset and power control.

High Performance Network

Based on LQCD SciDAC prototypes in FY 2004 and FY 2005, the “6n” and “Kaon” clusters purchased in FY2006, the “7n” cluster purchased in FY2007, and the “J/Psi” cluster purchased in FY2008 and FY2009, Infiniband is the preferred choice for the first LQCD-Ext cluster, and likely for any later clusters. The clusters will use quad data rate (QDR) or faster 4X Infiniband parts.

Current double data rate (DDR) and QDR switch configurations from multiple Infiniband vendors include 24, 36, 144, 288, and 324-port switches (QDR switches are 36 or 324 ports). For the large clusters to be built in this project, leaf and spine designs are preferred. Because QDR 4X HCA bandwidths exceed the requirements for lattice QCD

codes, oversubscribed designs will be used. A 3:1 design, for example, would have 24 computers attached to a 32-port switch, with the remaining 8 ports used to connect to the network spine.

Service Networks

Although Infiniband supports TCP/IP communications, we believe that standard Ethernet will still be preferred for service needs. These needs include booting the nodes over the network (for system installation, or in the case of diskless designs, for booting and access to a root file system), IPMI access (IPMI-over-LAN), serial-over-LAN, and NFS access to “home” file systems for access to user binaries. All current motherboard candidates support two embedded gigabit Ethernet ports.

In our experience, serial connections to each computer node are desirable. These connections can be used to monitor console logs, to allow login access when the Ethernet connection fails, and to allow access to BIOS screens during boot. Either serial-over-LAN (standard with IPMI 2.0) or serial multiplexers will be used to provide these serial connections.

Network Plan

For LQCD-Ext Project clusters, we will replicate the network layout currently used on all of the FNAL and JLab lattice QCD clusters. In these designs all remote access to cluster nodes occurs via a “head node”, which connects to both the public network and to the private network that forms the sole connection to the computer nodes. Secure ID logon (Kerberos at FNAL, ssh at JLab) is required on the head node. “R-utility” (rsh, rlogin, rcp) or host authenticated ssh are used to access the compute nodes.

File I/O

Particularly for analysis computing, large aggregate file I/O data rates (multiple streams to/from diverse nodes) are required. Data transfers over the high performance Infiniband network, if reliable, will be preferred to transfers over Ethernet. Conventional TCP/IP over Infiniband relies on IPoIB (“IP over IB”, one of the protocols supported by the Open Fabrics Enterprise Distribution, or OFED, Infiniband software stack)..

NFS has not proven to be reliable on our prior lattice QCD clusters for extensive file reading and writing, though it has been reliable for access to binaries and for smaller writing activities, such as job log files. Instead, command-based transfers using TCP, such as rcp, scp, rsync, bbftp, *etc.*, have been adopted for the transfer of large data files. On the JLab and FNAL clusters, multiple raid file systems available at multiple mount points have been used. Utility copy routines have been implemented to throttle access, and to abstract the mount points (*e.g.*, copy commands refer to */data/project/file*, rather than */data/diskn/file*).

FNAL uses both *dCache* and *Lustre* as alternatives to NFS. *dCache* provides a flat file system with scalable, throttled (reading), and load balanced (writing) I/O; additionally, it supports transparent access to the FNAL tape-based mass storage system. *Lustre* provides a POSIX-compliant filesystem visible from all worker nodes and from the cluster head node. Both *dCache* and *Lustre* have the properties that the storage volume and aggregate performance (instantaneous rate of data movement summed across all active transfers) can be scaled upwards by adding additional storage server nodes (known as *pools* for *dCache*, and *OSTs* for *Lustre*). Each new storage server nodes adds additional independent spindles of disks to the filesystems.

The LQCD-ext project will carefully watch developments in the parallel filesystem area for changes that can impact the deployed systems. LQCD-ext will leverage work in this area performed by the large high energy physics experiments such as Atlas and CMS at the Large Hadron Collider. Relevant issues in this area include concerns over the long term viability of *Lustre* given the pending acquisition of Sun by Oracle, concerns over the long term viability of *dCache* (implemented and maintained by the HEP community for support of experiments), and the emergence and/or maturation of parallel filesystems such as GPFS, pNFS (the parallel version of NFSv4), and Hadoop.

Procurement Strategy

LQCD-Ext will procure at most five separate lattice QCD computing systems, one in each of the five years of the project. The guiding principal of all of these procurements is that the most cost effective hardware will be deployed, where effectiveness is judged by the quantity of science (and of course, quality of science in terms of the reliability of the numerical results) that will produced during the lifetime of the individual lattice QCD system. In addition to commodity hardware clusters, similar to those deployed during the preceding LQCD Computing Project, we will evaluate alternatives such as the IBM BlueGene family of computers, systems based in whole or in part on GP-GPUs (general purpose graphics processing units), traditional supercomputers such as the Cray XT series, purpose built machines such as the QCDOC, and other hardware suitable for lattice QCD calculations that will emerge.

At each of the annual project progress reviews, scheduled in or about the month of May of each fiscal year, LQCD-Ext will present the plans for the deployment that will occur in the next fiscal year. For example, in spring of calendar year 2010, the project will present the plans for the procurement that will occur in FY2011. The only exception to this schedule occurs during the first year of LQCD-Ext; the plans for the FY2010 acquisition will be presented at the CD2/CD3 (Critical Decision 2 / Critical Decision 3) review in August of 2010. The annual presentation of procurement plans will include the selection of hardware designs that will be considered, cost and performance estimates and their justifications, and a detailed schedule.

All procurements will utilize a multistep process:

1. Identify and characterize candidate computer and network hardware
2. Create a preliminary system design

3. Solicit vendor feedback on the preliminary system design through an RFI (Request for Information) solicitation
4. Create a system design based on vendor feedback and any new information that has emerged
5. Solicit vendor cost proposals for the system design through an RFP (Request for Proposal) solicitation
6. Evaluate RFP responses and award purchase order(s) to the winning vendor(s), issuing a final system design as necessary
7. Accept or reject the delivered system(s) based on acceptance testing

Both the preliminary system design and system design may include two or more selections of hardware; for example, in a given year, both commodity clusters and members of the IBM BlueGene family may be included. Throughout the five years of the project, LQCD-Ext personnel will actively monitor the market, identifying and characterizing through benchmarking candidate hardware for upcoming procurements. Project personnel will also interact closely with computer and network manufacturers to understand product features and schedule roadmaps.

The evaluation and selection of hardware for the preliminary system design, and the evaluation of vendor responses to the RFP, will rely on the projected performance of the anticipated lattice QCD applications that will be run on the hardware during its lifetime. The particular mixture of lattice QCD applications used will be determined by LQCD-Ext Project staff in consultation with the USQCD Executive Committee and the USQCD Scientific Program Committee.

All awards will utilize firm, fixed-price contracts. Vendors will be encouraged to include modifications to the system designs in the RFI and RFP in their responses that will maximize the value of the delivered systems. Purchase awards will be based on best value evaluations that will include factors such as price/performance, quality of the vendor, quality of the proposed hardware, power consumption of the proposed hardware, impact on the facility infrastructure of the host laboratory, and usability of the delivered system.

LQCD-ext will procure storage for *dCache*, *Lustre*, and NFS filesystems separately from the computing systems. The amount of storage purchased will be determined in part from the requests that are required for all proposals to the Scientific Program Committee for allocations of time. The incremental storage added at each site annually will be at least as great as the sum of requested storage in the annual allocations proposals. Further, the storage will be deployed using a sufficient number of servers to meet the anticipated I/O bandwidth needs of the coming allocation year.

Acquisition Plan for FY2010/FY2011

As discussed in the section “Previous LQCD Computing Project” above, during the prior LQCD Project a combined FY2008/FY2009 procurement was used for the final hardware deployment. LQCD-Ext will use a similar process and will combine the FY2010 and FY2011 procurements into a single deployment to be housed at Fermilab. The

advantages of a single procurement include minimizing on-project and in-kind manpower as well as reducing the overhead (G&A) costs to the project. In this combined procurement the request for proposal (RFP) will require the vendors' response to include pricing for hardware to be purchased with FY2010 funds, as well as options to buy additional hardware with FY2011 funds and any other FY2010 funds that may be available.

Compute Nodes and High Performance Network

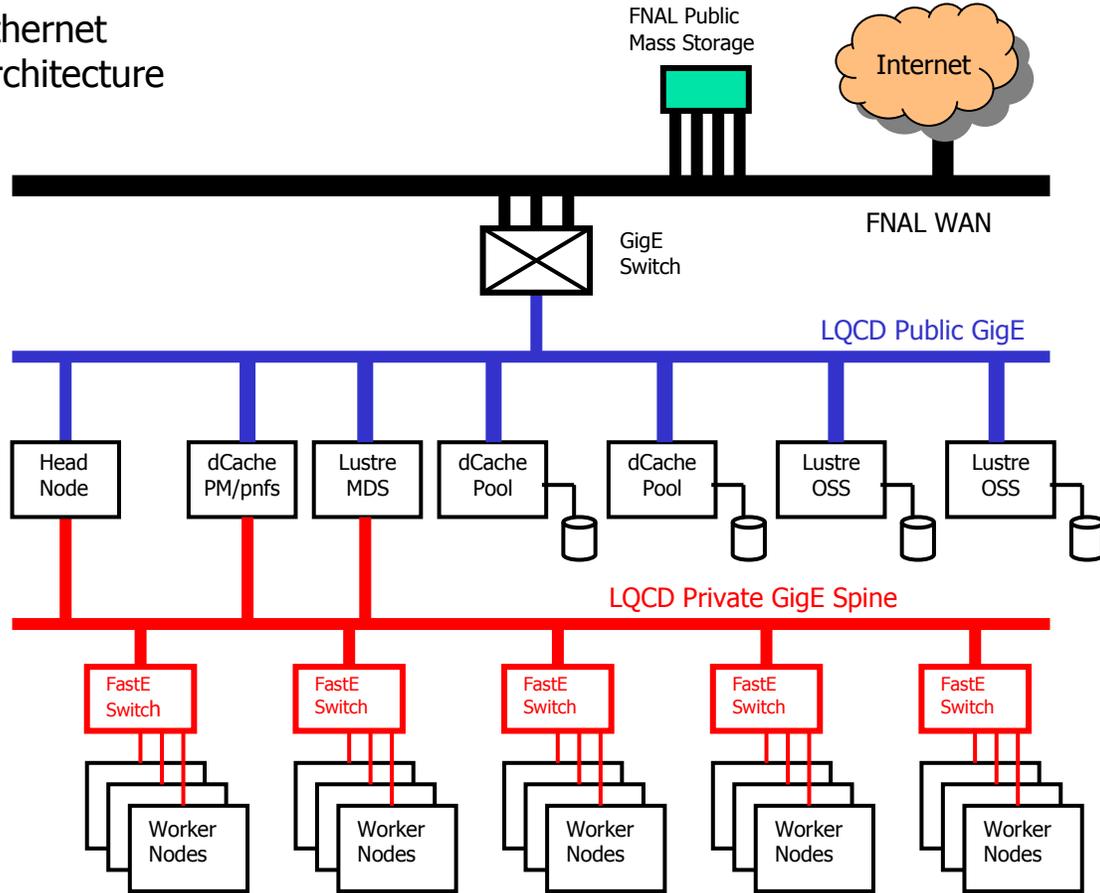
For the FY2010/FY2011 deployment, the follow-on processor to Intel's current "Nehalem" processor, code-named "Westmere", is expected to be available. "Westmere" is projected to have six cores, rather than the four on "Nehalem", and will have faster memory controllers (up to 1600 MHz). Also, a next-generation AMD Opteron processor, with as many as twelve cores per socket, is expected to be available with significantly increased memory bandwidth over current designs through the use of as many as four DDR3 memory channels per processor socket. "Nehalem", "Westmere", "Shanghai" (the current AMD Opteron processor), and "Magny-Cours" (the next-generation Opteron) should provide a rich set of options for the system design. LQCD-Ext personnel will evaluate these options as they become available and will select those that are viable for the preliminary system design to be presented to vendors in the RFI solicitation.

Ethernet Network Architecture Diagram and Description

The diagram below shows the Ethernet network architecture of the J/Psi cluster installed at Fermilab in FY2009. A similar architecture will be used for the FY2010/FY2011 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 FY2010 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 an LTO-4 tape drive.

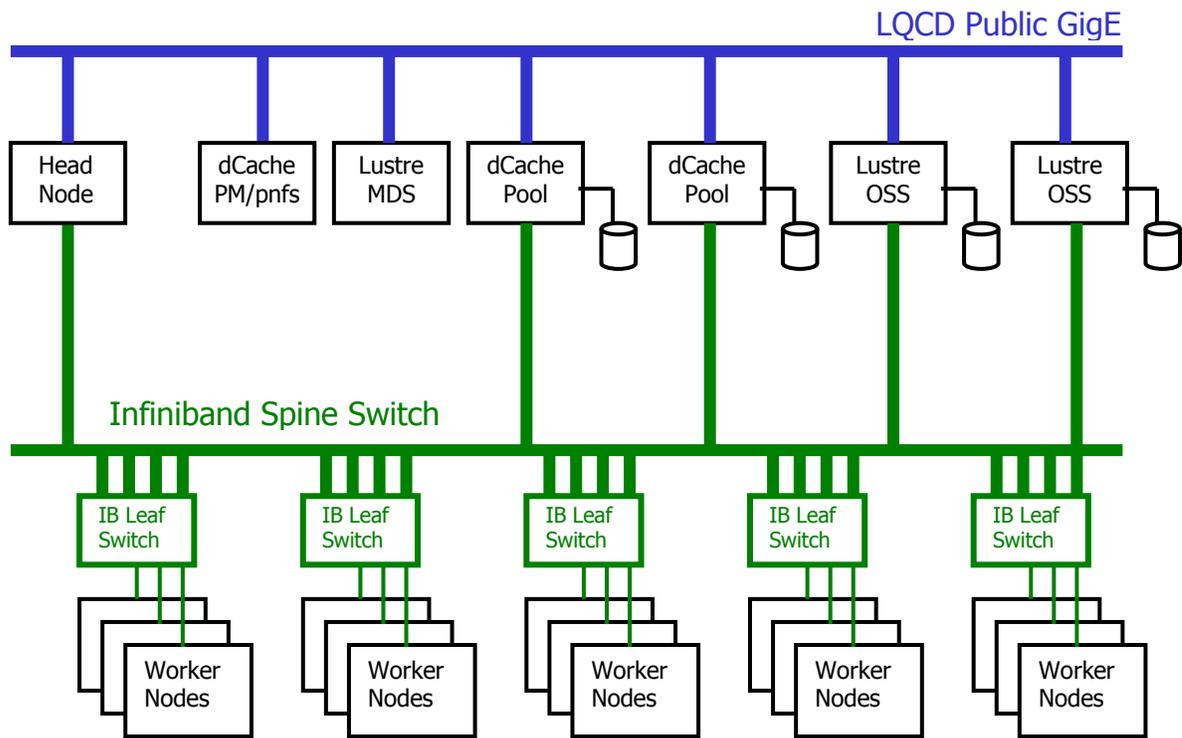
Users of the existing FNAL and JLab clusters login to the cluster head (login) node; the scheduler (*Torque* plus *Maui*) runs on either this node or another dedicated node. Approximately 10 Tbytes of local disk are attached to the login node.

Ethernet Architecture



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 J/Psi cluster. A similar architecture will be used on the FY2010 cluster.

On the FY2010 cluster, similar to the JLab 6N and 7N, and FNAL Kaon and J/Psi clusters, a leaf and spine approach will be used. Each set of worker nodes will be connected to a 24-port or 36-port leaf switch. Multiple links connect each leaf switch to a central spine switch (or switches, depending upon implementation). 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 FY2010 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, Pathscale 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 *Mau*).
12. On head and worker nodes, create authorized user accounts.
13. Test batch system.
14. Test LQCD applications.

Computing Room Facility for the FY2010/FY2011 Cluster

Fermilab will house the FY2010/FY2011 cluster in computer room C of the Grid Computing Center (GCC-C). GCC-C provides a total of 840 KW of power and the corresponding cooling. A UPS is used to condition the power to the computers and to provide a few minutes of operations in the event of a power outage to the building. GCC-C currently houses the SC LQCD J/Psi cluster. J/Psi draws approximately 250 KW of power and occupies 22 rack positions. GCC-C has a total capacity of 64 rack positions.

Schedule

The FY2010 procurement will consist of a number of phases. In the first phase, running from summer 2009 through January 2010, LQCD-EXT Project staff will evaluate the performance of LQCD codes on the various hardware options detailed in this document. These evaluations will determine the configurations of processor, chipset, and networking hardware that will be in the competitive range for vendor bidding. In the second phase, a Request for Information (RFI) will be used to obtain information from prospective vendors about their ability to design and build the cluster. In the third phase, a Request for Proposal (RFP) will be used to solicit vendor bids and to award the subcontract for the FY2010 cluster. This subcontract will be a firm, fixed-price contract competitively awarded based on best value; the subcontract will allow, via an option, the purchase of additional racks of computers and network equipment of the same design and layout in FY2011. In the fourth phase, the vendor will deliver a sequence of hardware, starting with a single machine, then a first integrated rack (approximately 40 machines with an Infiniband and an Ethernet leaf switch, capable of running production LQCD jobs), and then finally the full set of FY2010 equipment. Fermilab will approve the hardware at each step using functional tests before releasing the vendor to deliver the next increment. In the final phase, the FY2010 equipment will be integrated and tested by Fermilab. In FY2011, Fermilab will optionally exercise the purchase option in the contract using FY2011 funds.

The abbreviated schedule for the FY2010/FY2011 cluster deployment is shown below:

- Summer 2009: Test Intel “Nehalem” processors and “Tylersburg” chipsets for price/performance for LQCD codes. Test AMD “Shanghai” processors for price/performance for LQCD codes.
- Fall-Winter 2009: Test Intel “Westmere” processors and corresponding chipsets for price/performance for LQCD codes. Depending upon availability this testing may occur as late as early 2010.
- Feb 15, 2010: Preliminary Cluster Design Document completed
- Mar 15, 2010: Request for Information (RFI) released to vendors.
- Apr 15, 2010: Evaluation of RFI responses complete and documented.
- May 15, 2010: Request for Proposal (RFP) released to vendors.
- June 15, 2010: RFP responses evaluated and award recommendation complete.
- July 1, 2010: Purchase subcontract awarded, committing FY2010 project funds.
- August 1, 2010: Approval of sample unit.
- August 15, 2010: Approval of first rack.
- September 30, 2010: Delivery of remaining equipment.
- Nov 15, 2010: “Friendly User” production begins on hardware integrated from the October 1 delivery.
- Dec 15, 2010: Exercise FY2011 purchase option
- Jan 2, 2011: Release to production of FY2010 cluster.
- Feb 1, 2011: Delivery of FY2011 equipment
- March 1, 2011: Release to production of FY2011 portion of cluster.

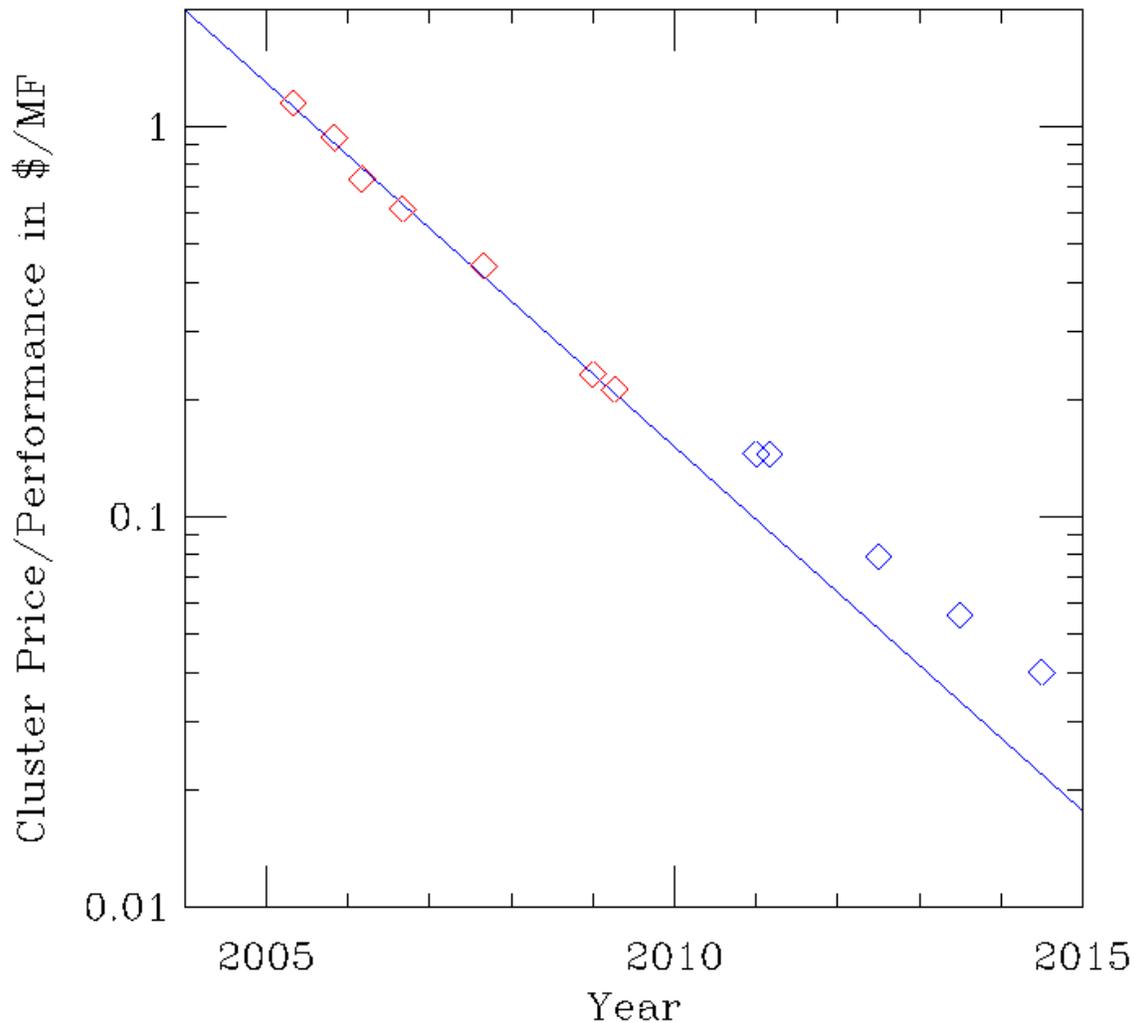
Cost Basis

During the SC LQCD project, four clusters based on Intel and AMD x86-hardware and Infiniband were deployed at Fermilab and Jefferson Lab. Prior to the beginning of that project, one additional similar cluster, Pion, was deployed at Fermilab. Table II below shows the cost per node for each cluster, along with the performance on LQCD applications in MFlops (the standard average of the *DWF* and *asqtad* action inverters). From these data, a price/performance metric of $\$/MF$ is also shown.

Cluster	Price per Node	Performance/Node, MF	Price/Performance
Pion #1	\$1910	1660	\$1.15/MF
Pion #2	\$1554	1660	\$0.94/MF
6n	\$1785	2430	\$0.74/MF
Kaon	\$2617	4260	\$0.61/MF
7n	\$3320	7550	\$0.44/MF
J/Psi #1	\$2274	9810	\$0.23/MF
J/Psi #2	\$2082	9810	\$0.21/MF

Table II – Price/Performance on LQCD Applications of SC LQCD Cluster Deployments at Fermilab and Jefferson Lab. Note that two entries are shown for both Pion and J/Psi, as each of these clusters were purchased in two pieces. In the case of Pion, the second half was purchased 5 months after the first half, and the cost of the computers had dropped. In the case of J/Psi, the second purchase had a lower cost per node for the Infiniband networking because the first purchase included the full Infiniband spine switch.

During the LQCD-Ext project, we plan to deploy computing equipment costing no more than \$2.46M in each of the five years. Further, as discussed above, the FY2010 and FY2011 purchase will be combined into a single contract. The plot below shows the price/performance data from the table above, plus the price/performance that each new hardware purchase would need in order to meet the deployment goals shown in Table I. Each of the purchases is plotted at the planned deployment date. If price/performance trends continue as they have since 2005, we expect to achieve the better price/performance values along the fit line. The separation of the blue diamonds in the plot from the trend line is the project’s performance contingency. In each year, the project will build to cost (spend approximately the full annual budget), and we expect that the resulting computing capacity will be in excess of the project’s “deployed TFlops” goal. This excess is the contingency. The price/performance goals and trend values, as well as the resulting contingency in MFlops and the equivalent contingency in dollars (using the price/performance from the trend line), are shown in Table III for each of the four planned acquisitions.



Price/Performance Trend for Infiniband-based Clusters for LQCD Applications. The red diamonds correspond to LQCD clusters purchased at Fermilab and Jefferson Lab (Pion #1, Pion #2, 6n, Kaon, 7n, J/Psi #1, J/Psi #2). The blue diamonds correspond to the price/performance values necessary to achieve, respectively, systems of capability 23, 24, 44, and 57 TF in 2010/2011, 2012, 2013, and 2014, for costs of \$3.29M, \$1.875M, \$2.46M, and \$2.26M. The blue line is a fit to the red diamonds, with a corresponding Moore's Law halving time of 1.613 years.

Year	Deploy Date	Price/Perf. Goal	Price/Perf. Trend	Goal (TF)	Contingency (TF)	Contingency (K\$)
2010	2011.0	\$0.15/MF	\$0.098/MF	11	4.4	\$437
2011	2011.2	\$0.14/MF	\$0.098/MF	12	4.4	\$428
2012	2012.5	\$0.078/MF	\$0.052/MF	24	11.9	\$616
2013	2013.5	\$0.056/MF	\$0.034/MF	44	26.8	\$900
2014	2014.5	\$0.040/MF	\$0.022/MF	57	42.6	\$932

Table III – Performance Contingency for SC LQCD-EXT Planned Deployments

Performance Basis

The projections of price/performance used in the **Cost Basis** section above assume that computer hardware will continue to improve in effective performance per dollar for LQCD applications. For the FY2010/FY2011 cluster acquisition, the basis for predicting that LQCD application performance will increase includes the following:

- Intel roadmaps indicated that the *Westmere* processor will be available in time for the FY2010/FY2011 deployment. *Westmere* is predicted to have six cores compared to the four cores on *Nehalem*, and it is also predicted to support faster memory (up to 1600 MHz, compared to the 1333 MHz on *Nehalem*). *Westmere* should therefore have higher floating point capability per socket and higher memory bandwidth per socket.
- Mellanox roadmaps indicate that the next generation of Infiniband networking hardware will be released in time for this deployment. Some of the features predicted for the next generation include higher bandwidth (at least 80 Gbit/sec signaling rate compared to the QDR signaling rate of 40 Gbit/sec), better latency, and optimized collectives for parallel programming. At minimum, a new generation of Infiniband will pressure downwards the pricing on the prior generation (QDR).

Cluster	Processor	DWF Performance per Node	Clover Performance per Node	Asqtad Performance per Node
6n	3.0 GHz Single CPU Dual Core Pentium	2900 MFlops	1408 MFlops	1960 MFlops
Kaon	2.0 GHz Dual CPU Dual Core Opteron	4696 MFlops	3180 MFlops	3832 MFlops
7n	1.9 GHz Dual CPU Quad Core Opteron	8800 MFlops	5148 MFlops	6300 MFlops
J/Psi	2.1 GHz Dual CPU Quad Core Opteron	10061 MFlops	7423 MFlops	9563 MFlops
<i>Shanghai</i>	<i>2.4 GHz Dual CPU Quad Core Opteron</i>	<i>12530 MFlops</i>	<i>Not measured</i>	<i>10370 MFlops</i>
<i>Nehalem 1066 MHz FSB</i>	<i>2.26 GHz Dual CPU Quad Core Xeon</i>	<i>23590 MFlops</i>	<i>13240 MFlops</i>	<i>16940 MFlops</i>
<i>Nehalem 1333 MHz FSB</i>	<i>2.93 GHz Dual CPU Quad Core Xeon</i>	<i>29450 MFlops</i>	<i>16210 MFlops</i>	<i>20600 MFlops</i>

Table IV – LQCD Cluster Application Performance on Infiniband-Connected Clusters. Data for clusters 6n, Kaon, 7n, and J/Psi are measured using 128-process parallel runs. Data for the italicized processors are estimated from single node performance, using a 90% scaling factor for *Shanghai* matching the observed scaling on 7n and J/Psi, and using an 85% scaling factor for *Nehalem* matching the observed scaling on an Intel benchmarking cluster in June 2009.

Table IV shows the performance of LQCD applications on current processors interconnected with DDR Infiniband. Of particular note in this table is the significant rise in LQCD application performance from the Opteron processors used for the 7n and J/Psi clusters to the estimated performance for Intel *Nehalem*-based clusters. For the FY2010/FY2011 acquisition, if *Westmere* performance scales with core count, which is plausible since memory bandwidth will also increase, performance per node will range from 30 to 40 GFlops, depending upon the application. At 30 GFlops per node, the 11 TFlop target cluster for FY2010 will require approximately 370 nodes, corresponding to a cost of approximately \$4320 per node including networking and other hardware in the assumed \$1.6M acquisition. This compares very favorably to the per node costs shown in Table II above for SC LQCD project clusters, which ranged from \$1554 to \$3320 per node including networking and other hardware.

Section 508 and OMB Memorandum 07-18 Compliance

The systems that will be acquired during SC LQCD-EXT will be accessed via text-only interfaces without color encoding, over network connections. Any hardware, such as terminals, laptops, or desktop computers, used by scientists accessing these SC LQCD-EXT systems will be provided by those users or their institutions, not by this project. Accessibility issues, and therefore Section 508 compliance, are therefore solely influenced by such accessing hardware and not by the LQCD systems themselves. Any subcontract used by the SC LQCD-EXT will include requirements to comply with Section 508 should user interfaces that are non-text-only or include color-encoded text be included in the purchases.

In addition to their text-only representations, documentation and cluster status will be available to users of the SC LQCD-EXT systems via the World Wide Web. All such web pages will be compliant with Section 508 requirements.

None of the systems that will be deployed or operated by SC LQCD-EXT will utilize Windows Operating Systems. Therefore the requirements of OMB Memorandum 07-18 are not applicable.

Cyber Security

The systems acquired and/or operated by the SC LQCD-EXT project at Fermilab, Jefferson Lab, and Brookhaven will, respectively, belong to those laboratories' existing general computing, scientific computing, and scientific computing enclaves. These enclaves have been certified and accredited (C&A) and have Authority to Operate (ATO). The security plans for these enclaves are prepared and maintained in accordance with NIST Special Publication 800-18, Revision 1: *Guide for Developing Security Plans for Federal Information Systems*.