

**Acquisition Plan – FY2008 and FY2009
for the
SC Lattice QCD Computing Project (LQCD)**

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

August 14, 2007

Revision 1.1

PREPARED BY:
Don Holmgren, FNAL

CONCURRENCE:



William N. Boroski
Contractor Project Manager

8-14-2007
Date

**Acquisition Plan - FY2008 and FY2009
Change Log**

Revision No.	Description/ Pages Affected	Effective Date
Revision 1.0	Final version for May 14, 2007 DOE External Review	May 14, 2007
Revision 1.1	Applied standard project template to revision 1.0	August 14, 2007

Date
8-14-07

William N. Bennett
Contract Project Manager

Table of Contents

Introduction.....	1
Compute Nodes.....	2
High Performance Network	3
Service Networks	4
Network Plan	4
File I/O	4
Ethernet Network Architecture Diagram and Description.....	5
Infiniband Architecture Diagram and Description	6
Software Deployment and Other Integration Tasks.....	7
Computing Room Facility for the FY2008/FY2009 Cluster	7
Schedule.....	8

Introduction

The SC Lattice QCD Computing Project develops and operates new systems in each year from FY2006 until 2009. These computing systems are deployed at Fermilab (FNAL) and at Jefferson Lab (JLab). In addition, the project operates the 4.2 Tflop/s US QCDOC supercomputer at Brookhaven National Lab (BNL), and the prototype clusters developed under the SciDAC program at FNAL and at JLab in 2004-2006. Table 1 shows the planned total computing capacity of the new deployments, and the planned delivered (integrated) performance. For FY2006, the table shows the achieved figures as well as the planned numbers (In FY2006, FNAL deployed the “Kaon” cluster with 2.3 TFlop/s capacity, and JLab deployed the “6N” cluster with 0.32 TFlop/s capacity). The integrated performance figures assume at the beginning of FY2006 1.6 Tflop/s of total capacity of the SciDAC prototype clusters at JLab and FNAL, and 4.2 Tflop/s capacity on 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. On clusters, DWF code sustains approximately 30% greater flop/sec than asqtad code.

Table 1 - Performance of New System Deployments, and Integrated Performance (DWF+asqtad averages used)

	FY 2006	FY 2007	FY 2008	FY 2009
Planned computing capacity of new Deployments, Tflop/s (FY06 shows achieved value in parentheses)	2.0 (2.6)	2.9	4.2	2.0
Planned delivered Performance (JLab + FNAL + QCDOC), Tflop/s-yr (FY06 shows achieved value in parentheses)	6.2 (6.27)	9	12	15

In FY2007, the project will procure and deploy a cluster at JLab. In FY2008 and FY2009, the new deployments will occur at FNAL. The project plans to execute a combined FY08/FY09 acquisition, with a single subcontract and two delivery dates: one in October 2008 with the hardware and services purchased using FY08 funds, and the second smaller delivery in December 2008 with the hardware and service purchased using FY09 funds. By combining the two planned purchases into a single subcontract, a single larger homogenous cluster will be deployed. Combining the purchases also reduces total labor costs associated with the procurement, delivers integrated scientific computing capacity at a greater rate during the first three years of operation, allows the project to consider a new generation of processors and chipsets, and results in a larger homogenous lattice QCD (LQCD) cluster.

All LQCD Computing Project hardware procurements utilize firm, fixed-price contracts with vendors specializing in commercial off-the-shelf (COTS) hardware. Steady state operation of the project clusters are performed by the three host laboratories, each of which is a government-owned contractor-operated facility.

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.

The currently available commodity processors with the greatest memory bandwidths are Intel ia32 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 “6n” and Fermilab “Kaon” clusters purchased and deployed in 2006 use dual core processors; “6n” uses single-socket dual-core Pentium D 830 motherboards, and “Kaon” uses dual-socket dual-core Opteron 270 motherboards. Lattice QCD production on these clusters has shown that dual core processors scale very well on MPI jobs when the cores are treated as independent processors. The dual core versions 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, moving predominantly to four or more cores in 2008 and beyond.

All current commodity dual processor Xeon motherboard designs use a single memory controller to interface the processors to system memory. As a result, the effective memory bandwidth available to either processor is 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 2006 project acquisition of the “Kaon” cluster, a multiprocessor Opteron system was 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 second half of 2007 Intel will introduce two new generations of chipsets and processors. The processors will be fabricated using a new 45-nanometer process.

The first of the two generations has code-names “Penryn” for the processor family, and “Seaburg” for the chipset family. The combination of the new processors and chipsets allows for faster memory buses, up to 1600 MHz from the current 1333 MHz, as well as improved

achievable memory bandwidth due to modifications of the memory controller design and the “snoop filter.” Improvements in sustainable memory bandwidth of 25% over the current designs are predicted. Further, because of improvements in controlling leakage currents, the “Penryn” family will be capable of higher clock speeds, up to 3.2 GHz compared to the current generation’s 2.67 GHz maximum speed. Systems incorporating Penryn/Seaburg will be on the market in August 2007. The LQCD project will have access to sample systems for benchmarking and analysis prior to their commercial release.

The second of the two generations has code-names “Nehalem” for the processor family, and “Tylersburg” for the chipset family. According to industry rumors, substantiated by conversations with Intel, “Nehalem” will incorporate a new memory controller design and will use DDR3 memory rather than FB-DIMMs. Estimates of improvement in sustainable memory bandwidth over current designs are 6 to 7-fold. If realized, these improvements would greatly impact the performance of LQCD code. Intel schedules indicate that the first “Nehalem” processors should be available for the LQCD project to benchmark by early calendar year 2008, if not sooner.

The new AMD quad core processors, code-named “Barcelona”, were available in time for the FY2007 7N acquisition. Preliminary LQCD code benchmarking in late April on engineering samples of the Barcelona family showed very strong performance. New versions of this processor should be available in time for the FY2008 procurement, with larger L3 cache memories expected.

We have found that hardware management features such as IPMI minimize the operating costs of commodity clusters. We will choose systems in FY 2008 based upon motherboards that support out-of-band management features, such as system reset and power control.

High Performance Network

Based on SciDAC prototypes in FY 2004 and FY 2005, the “6n” and “Kaon” clusters purchased in FY2006, and the “7n” cluster purchased in FY2007, Infiniband is the preferred choice for the FY08/FY09 cluster. The cluster will use double data rate (DDR) or quad data rate (QDR) 4X Infiniband parts.

Current switch configurations from multiple Infiniband vendors include 24, 96, 144, and 288-port switches. For the large clusters to be built in this project, leaf and spine designs are required. Because DDR 4X HCA bandwidths exceed the requirements for lattice QCD codes, oversubscribed designs will be used. A 2:1 design, for example, would have 16 computers attached to a 24-port switch, with the remaining 8 ports used to connect to the network spine.

A 2:1 oversubscribed design supporting 1024 compute nodes would employ 64 24-port leaf switches, and either six 96-port, four 144-port, or two 288-port spine switches. A 5:1 oversubscribed design supporting 1024 compute nodes would employ 52 24-port leaf switches, and either two 144-port or one 288-port spine switch.

Starting in 2008, new Infiniband switch configurations are expected based on 32-port rather than 24-port building blocks.

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

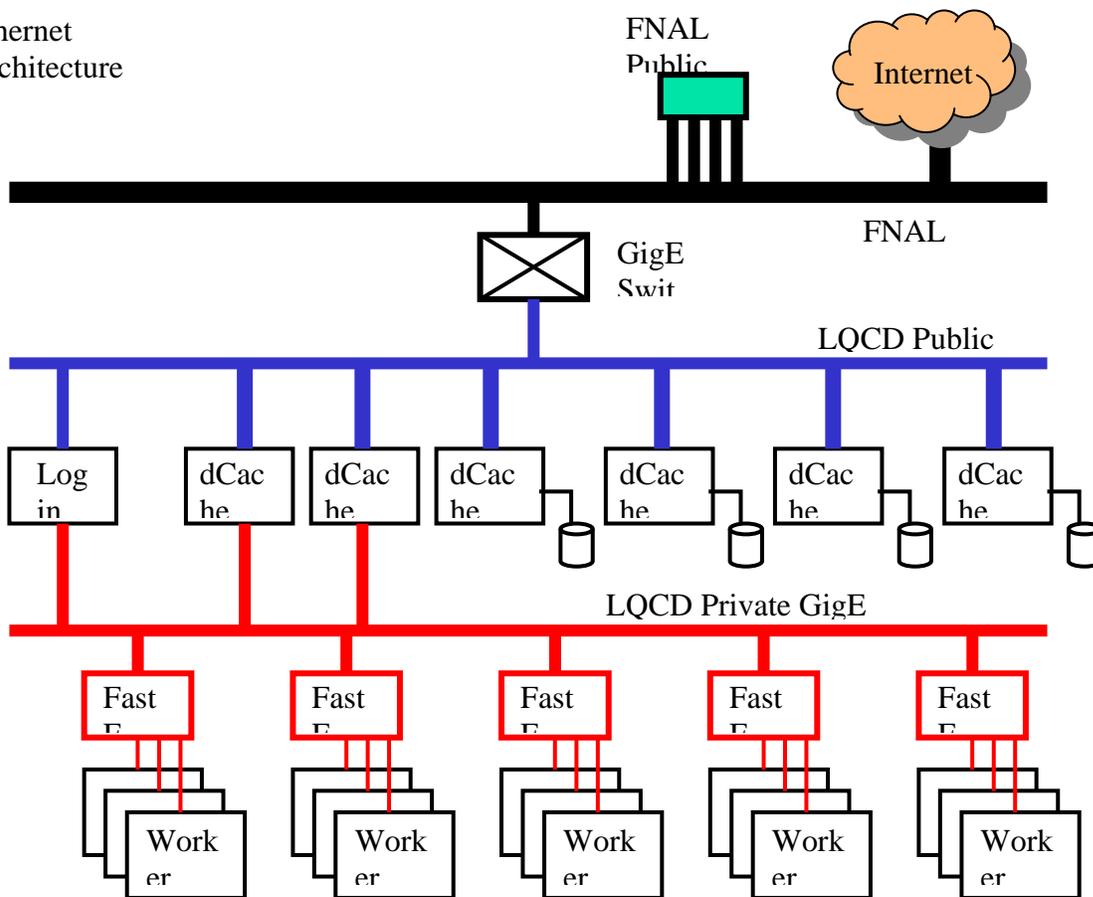
We will replicate the network layout currently used on all of the FNAL and JLab lattice QCD clusters. In these designs all remote access to cluster nodes occurs via a “head node”, which connects to both the public network and to the private network that forms the sole connection to the computer nodes. 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, with SDP (Socket Data Protocol) available as an attractive alternative that incurs less processor overhead.

NFS has not proven to be reliable on our prototypes 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 *dCache* as an alternative. *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.

Ethernet Architecture



Ethernet Network Architecture Diagram and Description

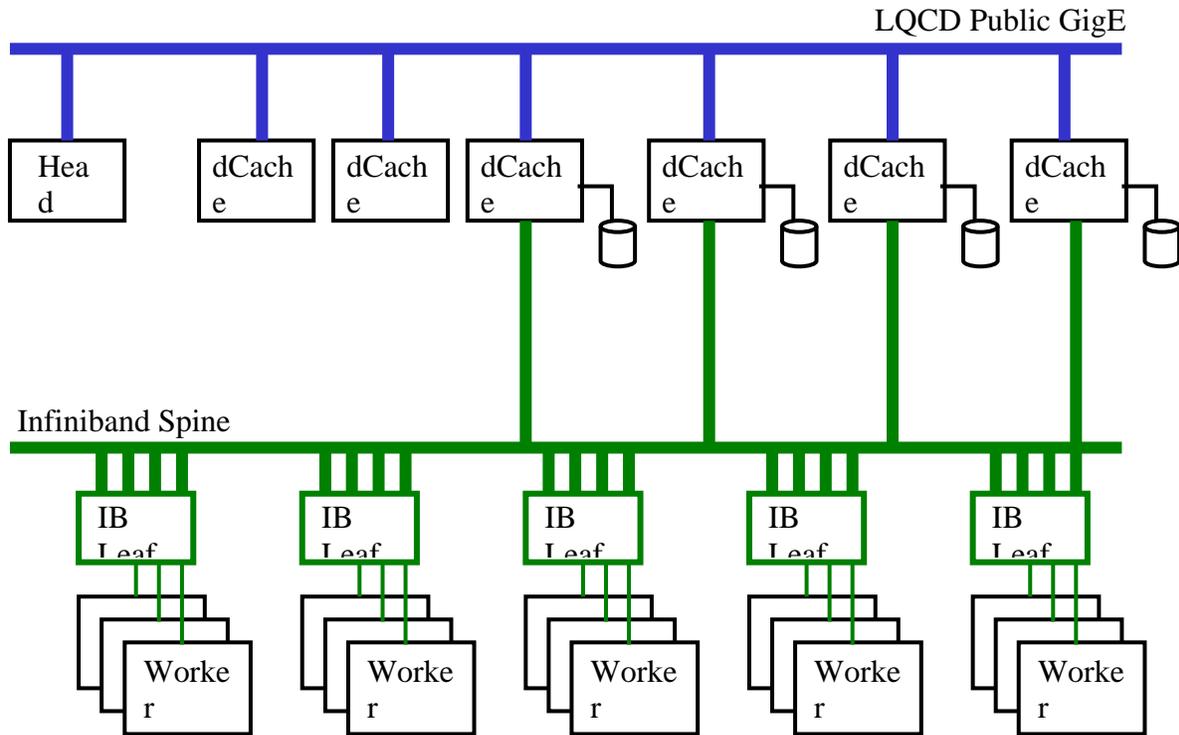
The diagram above shows the Ethernet network architecture of the Kaon cluster installed at Fermilab in FY2006. A similar architecture will be used for the FY08/FY09 cluster. Public and private gigE networks will be used, as shown in the diagram. The public gigE network will connect via a Cisco switch to FNAL's wide area network via a set of four channeled gigabit Ethernet connections. The FY08/FY09 facility will access the Fermilab mass storage facility via the Fermilab WAN. Within the mass storage facility are multiple *tape mover nodes*, each attached to an STK 9940B tape drive (in the future, LTO-3 drives will be available as well). Also within the mass storage facility are multiple *dCache pool nodes*, which provide a disk cache on top of the tape storage.

Users of the FNAL facilities log in to the 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 fast Ethernet 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 NFS-mounts 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 utilize the Infiniband network (see drawing below); the head node is bridged to the Infiniband network via a *router node* (not shown in drawing) that is connected to both the private gigabit Ethernet network and the Infiniband fabric.

Infiniband Architecture



Infiniband Architecture Diagram and Description

The diagram above shows the Infiniband architecture used on the Fermilab Kaon cluster. A similar architecture will be used on the FY08/FY09 cluster.

On the FY08/FY09 cluster, similar to the JLab 6N and FNAL Kaon clusters, a leaf and spine approach will be used. Each set of worker nodes will be connected to a 24-port or 32-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 *mpich-vmi* versions will be available). The Infiniband fabric will also be used for high performance file I/O via TCP, using IPoIB. A high performance I/O path from the head node to each worker node will be available through the use of a *router node* which will route packets from gigabit Ethernet to the Infiniband fabric and vice versa (not

shown in diagram). Infiniband fabric connections to each of the NFS server nodes at JLab will allow for high performance I/O by the worker nodes.

Because the FY08/FY09 cluster will be housed in a different building than Kaon, but will share the same dCache pool nodes, an Infiniband range expander such as Obsidian's "Longbow Campus" will be used to connect via a dedicated fiber the Infiniband fabrics of the new cluster and the existing Kaon cluster. This range expander will yield 8 Gbit/sec of bandwidth between the clusters. Only file I/O traffic will be routed through the range expander, as all LQCD executed jobs will use nodes exclusively on only one of the clusters.

Software Deployment and Other Integration Tasks

To bring the FY08/FY09 cluster into production, the following integration tasks will be necessary (order may vary from that shown):

- 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.
- Install system images on all worker nodes.
- 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.
- 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.
- 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.
- Configure IPMI facilities on all worker nodes; this includes initializing BMC network parameters (IP addresses, subnet masks, ARP and gratuitous ARP configuration).
- Test IPMI facilities on all worker nodes.
- On head node, deploy commercial compilers (Intel, Portland Group, Pathscale as requested by user community).
- On head node, build and deploy SciDAC libraries.
- On head and worker nodes, deploy SciDAC common runtime environment.
- On head and worker nodes, deploy and configure batch system (Torque plus Maui).
- On head and worker nodes, create authorized user accounts.
- Test batch system.
- Test LQCD applications.

Computing Room Facility for the FY2008/FY2009 Cluster

Fermilab will house the FY08/FY09 cluster in the "GCC-C" computer room. GCC is the acronym for "Grid Computing Center". Currently GCC consists of 2 computer rooms, GCC-A, and GCC-B, each of which provides 840 KW of usable power with cooling for computer

systems. GCC-C will similarly provide 840 KW of power. The building for GCC-C exists, and the construction project that will build out GCC-C will begin in early FY2008. Construction of the facility, based on the earlier GCC computer room “A” and “B” constructions, will take four to six months. To take into account the risks of a late construction start due to a continuing budget resolution and to delays associated with any other budget constraints, the LQCD Project is assuming that GCC-C construction will actually start in March 2008 and last for six months. With the inclusion of an additional month of float, beneficial occupancy will be available on October 1, 2008. The LQCD Project schedule will track these external dependencies.

Schedule

The FY08/FY09 procurement will consist of a number of phases. In the first phase, running from June 2007 through February 2008, LQCD 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 FY08/FY09 cluster. This subcontract will be a firm, fixed-price contract competitively awarded based on best value. The subcontract will stipulate two delivery dates, the first corresponding to hardware funded by FY08 project funds, and the second corresponding to hardware funded by FY09 funds. 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 FY08 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 FY08 equipment will be integrated and tested by Fermilab, and the FY09 portion of the subcontract will be executed. The FY09 hardware will be integrated with the FY08 hardware, and after a period of “friendly user” production the full system will be released to production.

The abbreviated schedule for the FY08/FY09 cluster deployment is shown below:

Summer, 2007: Evaluate Intel “Penryn” processor family and “Seaburg” chipset family for price/performance for LQCD codes. Evaluate AMD “Barcelona” processor family for price/performance.

Fall-Winter 2007: Test Intel “Nehalem” processors and “Tylersburg” chipsets for price/performance for LQCD codes. Depending upon availability from Intel, this testing may occur as late as early 2008.

February 15, 2008: Preliminary Cluster Design Document completed

March 15, 2008: Request for Information (RFI) released to vendors.

April 15, 2008: Evaluation of RFI responses complete and documented.

May 15, 2008: Request for Proposal (RFP) released to vendors.

June 15, 2008: RFP responses evaluated and award recommendation complete.

July 1, 2008: Purchase subcontract awarded, committing FY08 project funds.

August 1, 2008: Approval of sample unit.

August 15, 2008: Approval of first rack.

October 1, 2008: Delivery of remaining equipment.

October 15, 2008: Exercise of option in subcontract for additional identical hardware, committing FY09 project funds.

November 15, 2008: "Friendly User" production on hardware integrated from the October 1 delivery.

December 15, 2008: Delivery of FY09-funded hardware.

January 1, 2009: Release to production of FY08 cluster.

January 15, 2009: "Friendly User" production on hardware integrated from the December 15 delivery.

January 31, 2009: Release to production of full combined FY08/FY09 cluster.