

Acquisition Plan – FY2007

Introduction

The 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), as well as 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.

	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	3.0
Planned delivered Performance (JLab + FNAL + QCDOC), Tflop/s-yr (FY06 shows achieved value in parentheses)	6.2 (6.27)	9	12	15

Table I – Performance of New System Deployments, and Integrated Performance (DWF+asqtad averages used). Funding for hardware in FY2009 is ½ prior years.

In FY2007, the project will procure and deploy a cluster at JLab. In FY2008 and FY2009, the new deployments will occur at FNAL. All project hardware procurements utilize firm, fixed-price contracts with vendors specializing in commercial off-the-shelf hardware. The steady state operations 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 “Cloverton”, Pentium “Conroe”) 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.

In the past year, Intel and AMD have switched all new processors of relevance to lattice QCD to dual 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 FNAL “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 their next generation of chipsets and processors. Improvements may increase available memory bandwidth relative to Opteron designs.

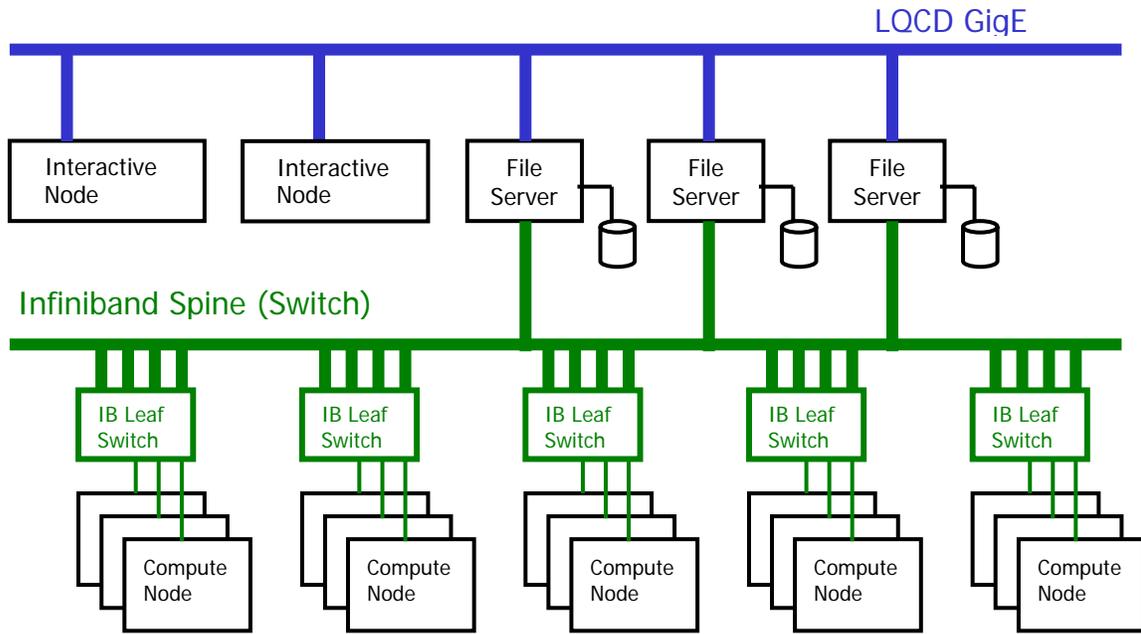
Based on benchmarking and prototyping performed in late 2006, the leading candidates for compute nodes in FY 2007 are those built upon single processor, dual core Pentium, dual processor dual core Xeon using FBDIMMs, and dual processor dual core Opteron motherboards supporting PCI-E.

We have found that hardware management features such as IPMI minimize the operating costs of commodity clusters. We will choose systems in FY 2007 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, and the JLab “6n” and FNAL “Kaon” clusters purchased in FY2006, Infiniband is the preferred choice for the FY 2007 cluster. For dual socket alternatives, the FY2007 JLab cluster will use double data rate (DDR) 4X Infiniband parts, matching the FNAL “Kaon” cluster. Single socket alternatives, such as the Intel “Conroe”, may be more cost effective with single data rate 4X fabric and so SDR will be considered for those alternatives.

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 Infiniband 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 8 ports used to connect to the network spine.



Infiniband Architecture Diagram and Description

The diagram above shows the Infiniband architecture at Jefferson Lab. 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.

Service Networks

Although Infiniband supports TCP/IP communications, standard Ethernet is still preferred for service needs such as booting the nodes over the network (for system installation, or in the case of diskless operating modes, for booting and access to a root file system), IPMI access (IPMI-over-LAN), and serial-over-LAN. 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 with commodity Linux based servers 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 between the file servers and the compute nodes. Utility copy routines have been implemented to abstract the multiple servers (*e.g.*, copy commands refer to */data/project/file*, rather than *serverN:/data/diskn/file*) and to throttle, via queuing, concurrent access to servers from multiple clients..

FNAL uses *dCache* as an alternative to provide 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.

Rather than using *dCache* with its associated manpower costs, JLab will deploy a smaller number of larger higher performance file servers with high performance file I/O via Infiniband. Investigations will be made of systems with potentially more robust NFS support so that applications can read and write directly to the file servers. In this way the cost (latency) of staging input and output files to lower performance local disks on the compute nodes can be avoided.

At JLab the file servers will be connected to the laboratory’s mass storage and wide area networking via 4 trunked gigE connections to the lab’s 10 gigabit ethernet backbone. From that point, data flows to and from multiple *tape data mover nodes*, each attached at 1 gigabit, and controlling a single STK 9940B tape drive.

The compute nodes are connected via fast Ethernet 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 also used for login access to the compute nodes by the scheduler (using *rsh*). Binaries are generally launched from the

/home directory, which is NFS mounted. Each worker node has considerable scratch space available on local disk (60+ Gbytes on newest systems).

Software Deployment and Other Integration Tasks

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

1. Prepare system installation images for the compute nodes (Linux Fedora Core is the current preferred choice). These images will include the Infiniband software stack (OpenIB, or commercial), batch system software, as well as the SciDAC LQCD shared libraries.
2. Install system images on all compute nodes.
3. Unit test compute 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. Attach NFS servers to the Infiniband fabric and perform configuration & testing.
5. Configure IPMI facilities on all compute nodes; this includes initializing BMC network parameters (IP addresses, subnet masks, ARP and gratuitous ARP configuration).
6. Test IPMI facilities on all compute nodes.
7. On the interactive (head) nodes, deploy commercial compilers (Intel, Portland Group, Pathscale as requested by user community).
8. On head node, build and deploy SciDAC libraries.
9. On head and compute nodes, deploy SciDAC common runtime environment.
10. On head and compute nodes, deploy and configure batch system (*Torque* plus *Maui*). Test batch system.
11. Multi-node compute tests. The tests will include LQCD multinode application testing and performance verification. (All nodes, various set sizes).
12. Test production LQCD applications.

Acquisitions in FY2008-FY2009

In FY2008 and FY2009, the specific configurations of the systems acquired in each year will be determined by evaluating the available commodity computing equipment and selecting the most cost effective combinations of computers and networks for lattice QCD codes. Commercial supercomputers will also be evaluated for cost effectiveness.

Acquisitions at JLab and FNAL follow procurement procedures documented at each site. RFP's will be released for each purchase to a number of vendors to ensure competitive pricing. Acceptance tests at each site will include unit tests of the computers and Infiniband components, system tests of the clusters to verify performance on LQCD codes, and reliability tests under heavy compute and I/O loads.