

**System Description
for the
SC Lattice QCD Computing Extension Project
(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 1.2

August 4, 2009

PREPARED BY:
Don Holmgren, FNAL

CONCURRENCE:



Bill Boroski
Contractor Project Manager

August 4, 2009
Date

**LQCD-ext System Description
Change Log**

Revision No.	Description/ Pages Affected	Effective Date
Revision 1.0	Initial Version – Preliminary System Description	April 15, 2009
Revision 1.1	Update to create System Description for CD2/3	July 30, 2009
Revision 1.2	Update to reflect new budget profile	August 4, 2009

1. Scope and Purpose

This document describes the computing systems that will be deployed and/or operated by the Lattice QCD Computing Project Extension (SC LQCD-EXT) and Fermilab (FNAL), Jefferson Lab (TJNAF), and Brookhaven Lab (BNL). These systems include three clusters (7n, Kaon, J/Psi) that were deployed during the prior Lattice QCD Computing Project (SC LQCD), the QCDOC purpose-built supercomputer that was deployed at Brookhaven Lab by the DOE High Energy Physics (HEP) and Nuclear Physics (NP) program offices before the SC LQCD project, and the systems planned for deployment during SC LQCD-EXT.

2. Definitions Specific to SC LQCD-EXT

The following is a partial list of terms specific to SC LQCD-EXT and their definitions:

- *Inverter* – computer code that performs the algorithms necessary to solve, or “invert”, the Dirac operator. Specific inverters have been implemented by lattice QCD physicists for a number of specific actions, including Wilson, staggered fermion, improved staggered fermion (“asqtad”), domain wall fermion (“DWF”), and anisotropic clover.
- *Sustained TFlops* – the performance rating used to characterize lattice QCD computer code on a given computing platform. SC LQCD-EXT follows the practice of the prior project, SC LQCD, in assigning a standardized performance rating, *sustained TFlops*, to specific computer systems defined as the average of the performance of the domain wall fermion (DWF) and improved staggered fermion (asqtad) inverters. Each of these inverters requires a precisely known number of floating point operations during each pass of the iterative Dirac operator solver (“inverter”). The performance measurement for each action is defined as the number of iterations performed per second multiplied by the number of floating point operations per iteration.

3. Systems Existing at the Start of SC LQCD-EXT

The following systems dedicated to LQCD computations will be in existence at the start of SC LQCD-EXT and will be operated for one or more years by the project. All performance figures use the average of the performance of the domain wall fermion (DWF) and improved staggered fermion (asqtad) inverters.

- QCDOC: This is a purpose-built supercomputer deployed at BNL that has 12288 400MHz PowerPC cores with a custom six dimensional mesh interconnect. The system may be partitioned by reconfiguring the mesh and currently consists of twelve 1024-node partitions. Users access QCDOC by logging in with ssh via a BNL firewall machine to one of two head nodes running the AIX operating system. The QCDOC sustains 4.2 TFlops aggregate. This system will be operated through FY2010.
- 7n: This is a commodity cluster deployed at TJNAF that has 396 nodes, each with two quad-core 1.9 GHz AMD Opteron “Barcelona” processors, 8 GBytes of memory, and double data rate (DDR) Infiniband. Users access 7n by logging in with ssh via a TJNAF firewall machine to one of two head nodes running the Linux operating system. 7n

sustains 2.9 TFlops aggregate. This system will be operated through at least the first three quarters of FY2011.

- **Kaon:** This is a commodity cluster deployed at FNAL that has 600 nodes, each with two dual-core 2.0 GHz AMD Opteron processors, 4 GBytes of memory, and DDR Infiniband. Users access Kaon by logging in with Kerberos to a head node running the Linux operating system. Kaon sustains 2.56 TFlops aggregate. This system will be operated through at least FY2010.
- **J/Psi:** This is a commodity cluster deployed at FNAL that has 856 nodes, each with two quad-core 2.1 GHz AMD Opteron processors, 8 GBytes of memory, and DDR Infiniband. Users access J/Psi by logging in with Kerberos to a head node running the Linux operating system. J/Psi sustains 8.40 TFlops aggregate. J/Psi also includes eight nodes each with two Tesla 1070 graphics processors (GPUs). This system will be operated through at least FY2012.

4. Systems to Be Acquired During SC LQCD-EXT

For further details refer to the SC LQCD-EXT Acquisition Strategy. During SC LQCD-EXT at most five independent systems dedicated to LQCD computing will be deployed, one per year in each of FY2010-FY014, and each at a cost varying by budget year between \$1.61M (FY2010) and \$2.46M (FY2013). The goals for aggregate sustained performance on LQCD applications for each of these deployments are given in the Table below.

	FY 2010	FY 2011	FY 2012	FY 2013	FY 2014
Planned computing capacity of new Deployments, Tflop/s	11	12	24	44	57

Each acquisition will deploy a system that will best meet the scientific goals planned by the USQCD community. The typical system will consist of a commodity cluster with a high performance interconnect. However, in each acquisition any other suitable hardware will be considered, such as the IBM BlueGene family, customized commodity hardware, and traditional supercomputer hardware.

The FY2010 and FY2011 systems will be acquired across the FY2010-FY2011 fiscal year boundary using a single subcontract with two separate obligations and corresponding deliveries. The planned architecture for this cluster will use a commodity cluster based on Intel *Nehalem* or *Westmere* processors (or other suitable processor, such as AMD *Shanghai*, *Istanbul*, or *Magny-Cours*) interconnected with Infiniband, unless one of the alternatives discussed above proves to be more effective at meeting the planned scientific goals. The acquisition plan will be reviewed in detail during the spring 2010 project progress review.

Each system acquired by SC LQCD-EXT will be operated for a minimum 4-year period.

5. System Design, Validation, and Operation

Each system acquired and/or operated by SC LQCD-EXT will support the software libraries and physics applications developed by the SciDAC Lattice QCD and SciDAC-2 Lattice QCD projects. For further information, see <http://www.usqcd.org/software.html>. The primary design goal for each system is the correct execution of these physics applications and libraries at the scale and performance required to carry out the planned program of scientific work. “Scale” here refers to both the size of the lattice QCD simulation, determined by the dimensions of the 4-D or 5-D representations of space-time and particle wave functions, as well the number of simulations to be carried out in each year of SC LQCD-EXT. Each new system acquired by SC LQCD-EXT in concert with the existing systems must meet a specific aggregate performance goal, expressed as the total sustained TFlops capacity measured using the domain wall fermion and improved staggered fermion inverters.

As detailed in the SC LQCD-EXT Acquisition Strategy document, acquisitions of new systems by SC LQCD-EXT will utilize “request for information” (RFI) and “request for proposal” (RFP) solicitations to obtain suggestions for system designs and bids for delivered systems. During the RFI and RFP process, vendors are provided with lattice QCD applications in binary and/or source form. Bids responding to RFP solicitations must include measured performance on these applications. Performance measures include the sustained TFlops on the codes and power consumption during these calculations. Acceptance testing of delivered systems includes verification of these performance figures.

Usage of the systems operated by SC LQCD-EXT is controlled by the USQCD Executive Committee and the USQCD Scientific Program Committee. The latter allocates time to projects on an annual basis based on proposals submitted to the committee. These allocations are quantified by a standard candle, typically, core hours or node hours. In 2007-2009, allocations were based on “6n node-hours”, each of which is equivalent to the production achievable in one hour on one node (dual-core computer) of the TJNAF “6n” cluster. All systems operated by the project are rated in terms of their sustained TFlops capacity using the average of DWF and asqtad performances detailed earlier in this document; this rating is used to determine the standard candle equivalence in any given year (currently, “6n node hours”).

The capacity of each system operated by SC LQCD-EXT for sustained performance on lattice QCD applications is established annually by benchmarking lattice QCD applications run as large (hundreds to thousands) of cores. These applications include internal checks (reproducibility, and mathematical checks such as verification of unitarity on representations of physics quantities) sufficient to validate the correct operation of the systems.

All of the systems operated by SC LQCD-EXT utilize the “Torque” (formally known as “PBS”) batch system for the running of jobs. The allocations of the USQCD Scientific Program Committee and any relevant operations policies established by the USQCD Executive Committee are enforced in the configuration of Torque, and in the case of the clusters, the Maui scheduler.