

*Department of Energy*

*Review Committee Report*

on the

**LATTICE QUANTUM  
CHROMODYNAMICS  
(LQCD)  
PROJECT**

May 2005

**Intentionally Blank**

## EXECUTIVE SUMMARY

The LQCD project supports the development and operation of a large scale dedicated computing facility capable of sustaining over seventeen teraflop/s for the study of Quantum Chromodynamics (QCD) that will play an important role in expanding our understanding of the fundamental forces of nature and the basic building blocks of matter. The hardware will be housed at Brookhaven National Laboratory (BNL), Fermi National Accelerator Laboratory (FNAL), and Thomas Jefferson National Accelerator Facility (TJNAF), and operated as a single distributed computing facility, which will be available to lattice gauge theorists at national laboratories and universities throughout the United States. The project will start in FY 2006 and be completed in FY 2009. These funds will support the acquisition and operation of ~13 Tflops that, when combined with existing hardware, will yield a system capable of over 17 Tflops. The total cost is estimated to be approximately \$9.2M. The President's FY 2006 budget requests \$2.5M to begin the project.

Over the past six years members of the United States lattice gauge theory community have worked together to plan the computational infrastructure needed for the study of QCD. Virtually all members of the community have been involved in this effort. Research and development performed during this period has provided the groundwork for the construction of production hardware beginning in FY 2006. With support from the Department of Energy's (DOE) High Energy Physics (HEP), Nuclear Physics (NP), Advanced Scientific Computing Research and SciDAC Programs, prototype hardware has been designed, constructed and tested, and the software needed to use it effectively has been developed.

Historically, by taking advantage of simplifying features of lattice QCD calculations, as has been done in designing the prototype hardware, it has been possible to build computers for this field that have significantly better price/performance than commercial machines. Two tracks for the construction of massively parallel computers for QCD have been studied. One involves the design and fabrication of key components, while the other makes use of carefully chosen commodity parts. The latest example of the former is the QCD on a Chip (QCDOC), which was designed by lattice gauge theorists at Columbia University in collaboration with colleagues at IBM. The design incorporates cpu, memory and communication on a single chip. As part of the

research and development effort a 12,288 processor QCDOC was recently constructed at BNL. The operation of the QCDOC for use by U.S. lattice gauge theorists will be an important component of the current project. In the commodity track, prototype clusters optimized for the study of QCD have been developed and tested at FNAL and TJNAF under a grant from DOE's SciDAC program, as well as with the support of these labs' base programs. This work indicates that commodity clusters will be the hardware of choice for the project in FY 2006. The guiding principle will be to build or purchase whatever hardware best advances the science at each stage of the project. Another important aspect of the SciDAC project is the creation of a common software environment that spans the QCDOC system and the clusters. This should allow relatively easy transitions between the systems.

In a February 26, 2005, memorandum, Robin Staffin, Associate Director of Science for High Energy Physics and Dennis Kovar, Associate Director of Science for Nuclear Physics requested that the Office of Advanced Scientific Computing Research organize and lead a DOE review of the LQCD Project plans for FY 2006 – FY 2009. The purpose of the review was to evaluate the LQCD team's plan for procuring and operating high-performance computer hardware and appurtenant services necessary to deliver new science.

Overall the Committee found the LQCD Team's approach to procuring high-performance computing hardware and maintenance and operations services to be reasonable. The team documented a clear scientific need for the proposed computational resources and explained the expected results with each increase in computational facilities. The team has done a very thorough job of evaluating the different hardware options in processors, memory subsystems and networks to deliver the most hardware capability within the available budget. The LQCD project is supportive of the Office of Science's goal to provide its theoretical high energy and nuclear physics scientific user communities with powerful high-end computing resources.

The plan advances the LQCD vision by continuing to operate the QCDOC computer at BNL and acquiring a series of Linux clusters to support the mix of anticipated scientific needs, which range from very compute-intensive generation of lattice configurations to throughput-driven analysis tasks. The strategy presented includes 8 procurements of new computers over the 4-year life of the project, with the first installation at TJNAF in March 2006, and the first installation at FNAL in July-August 2006, after completion of construction of expanded computer space.

The management team and LQCD personnel are well qualified and experienced in the installation of high-end computers for this community. In addition the project is leveraged by almost \$8 million in contributions from the three laboratories' base programs and other efforts such as SciDAC.

The review committee is supportive of the goals of this project and believes that there are significant science opportunities that will both optimize the effectiveness of existing and planned experiments and enable new applications of QCD to problems at the forefront of high energy and nuclear physics. In addition, the plan to use the most capable new computers to generate configurations and the older computers as analysis engines seems very well thought out. The committee was concerned with the number of computer installations being planned, especially given the very constrained operations budgets proposed by the project, and recommends that the team seriously consider reducing the number of installations to one per year, alternating between TJNAF and FNAL. In addition, if the FNAL construction schedule presented at the review, which delays the release of the computer there until September 2006, is accurate, the Committee recommends that the first computer be delivered to TJNAF. If the revised FNAL schedule is accurate, which would enable the computer to be released to operation there in April 2006; the Committee recommends that the team decide where the computer is installed based on where it can deliver the most science for the dollars invested. The Offices of High Energy and Nuclear Physics have made it clear that there are no barriers on how the allocation of project funds be distributed between the three laboratories; the procurement and installation strategy needs to be based on a plan that will optimize the delivery of science.

# CONTENTS

1. Introduction.....	1
2. The significance and merit of the proposed initiative.....	3
3. The status of the technical design.....	5
4. The feasibility and completeness of the proposed budget and schedule .....	8
5. Relevance of prototyping efforts and plans for developing the required software.....	12
6. The effectiveness of the proposed management structure.....	14
Appendices	
Appendix A - Charge Memorandum .....	18
Appendix B - Review Participants.....	21
Appendix C - Review Agenda .....	23
Appendix D - Installation Schedule.....	26
Appendix E - Acronyms .....	29

# 1. Introduction

Over the past six years members of the United States lattice gauge theory community have worked together to plan the computational infrastructure needed for the study of Quantum Chromodynamics (QCD). Virtually all members of the community have been involved in this effort. Research and development performed during this period has provided the groundwork for the construction of production hardware beginning in FY 2006. With support from the DOE's High Energy Physics (HEP), Nuclear Physics (NP), Advanced Scientific Computing Research and SciDAC Programs, prototype hardware has been designed, constructed and tested, and the software needed to use it effectively has been developed.

Historically, by taking advantage of simplifying features of lattice QCD calculations, as has been done in designing the prototype hardware, it has been possible to build computers for this field that have significantly better price/performance than commercial machines. Two tracks for the construction of massively parallel computers for QCD have been studied. One involves the design and fabrication of key components, while the other makes use of carefully chosen commodity parts. The latest example of the former is the QCD on a Chip (QCDOC), which was designed by lattice gauge theorists at Columbia University in collaboration with colleagues at IBM. The design incorporates cpu, memory and communication on a single chip. As part of the research and development effort a 12,288 processor QCDOC computer was recently constructed at BNL. The operation of the QCDOC for use by U.S. lattice gauge theorists will be an important component of the current project. In the commodity track, prototype clusters optimized for the study of QCD have been developed and tested at FNAL and TJNAF under a grant from DOE's SciDAC program, as well as with the support of these laboratories' base programs. This work indicates that commodity clusters will be the hardware of choice for the project in FY 2006. The guiding principle will be to build or purchase whatever hardware best advances the science at each stage of the project.

The LQCD project, which is the subject of this review, supports the development and operation of a large scale dedicated computing facility capable of sustaining approximately twenty teraflop/s for the study of QCD that will play an important role in expanding our understanding of the fundamental forces of nature and the basic building blocks of matter. The

hardware will be housed at Brookhaven National Laboratory (BNL), Fermi National Accelerator Laboratory (FNAL), and Thomas Jefferson National Accelerator Facility (TJNAF), and operated as a single distributed computing facility, which will be available to lattice gauge theorists at national laboratories and universities throughout the United States. The project will start in FY 2006 and be completed in FY 2009. The total cost is estimated to be approximately \$9.2M. The President's FY 2006 budget requests \$2.5M to begin the project.

The primary customers of the LQCD project are the DOE/SC programs in High Energy Physics and Nuclear Physics. Specific scientific projects in Quantum Chromodynamics request access to LQCD resources and receive an allocation of computer time. Each allocation is based on criteria determined by the LQCD Scientific Program Committee in cooperation with the science community served by these systems.

In a February 26, 2005 memorandum, Robin Staffin, Associate Director of Science for High Energy Physics and Dennis Kovar, Associate Director of Science for Nuclear Physics requested that the Office of Advanced Scientific Computing Research organize and lead a DOE review of the LQCD Project plans for FY 2006 – FY 2009. The purpose of the review was to evaluate the LQCD team's plan for procuring and operating high-performance computer hardware and appurtenant services necessary to deliver new science. Specifically the review team was asked to evaluate:

- The significance and merit of the proposed initiative;
- The status of the technical design, including completeness of technical design and scope, feasibility and merit of technical approach;
- The feasibility and completeness of the proposed budget and schedule, including availability of manpower; and
- The effectiveness of the proposed management structure.

In addition the committee was asked to evaluate the appropriateness and effectiveness of relevant R&D and prototyping efforts outside the scope of the initiative, and the status and plans for developing the required software for Lattice QCD computing. The review was held May 24 – 25, 2005 at MIT's Laboratory for Nuclear Sciences.

## 2. The significance and merit of the proposed initiative

The Committee was impressed with the potential significance of the proposed initiative. It builds on the efforts under SciDAC to develop a uniform software infrastructure that can take best advantage of emerging computers to increase the accuracy of a number of important calculations, such as CKM matrix elements, and open new areas of QCD physics ranging from hot nuclear matter to nucleon form factors. The timing of this project is excellent, as it comes at a moment that the Lattice QCD community has demonstrated that the planned computations will lead to important physics results with reliable and sufficient control over all systematic errors. The effectiveness of this initiative will be enhanced by international cooperation in the generation of configurations.

### 2.1. Findings

- 2.1.1. Within the duration of the project, several calculations that are essential to do justice to important experiments in high energy and nuclear physics will be performed. These include, notably, calculations of matrix elements that figure into standard model predictions for flavor and CP violating processes, in order to bring the precision of the predictions up to the level achieved by recent experiments. Several ideas for extension of the standard model predict discrepancies at this level. This confrontation between theory and experiment is a major frontier of nuclear and high energy physics, to which enormous human and capital resources have been devoted over the last decade.
- 2.1.2. Also very notably, aspects of the thermodynamic behavior of QCD, including the equation of state at small but significant values of the chemical potential, will be calculated. These computations will inform the interpretation of ongoing and future experiments at heavy ion colliders.
- 2.1.3. Finally, calculations of nucleon form factors will illuminate some recent surprising, and partially discrepant, measurements in that field.
- 2.1.4. Many other questions, including such classic problems of high energy and nuclear physics as the origin of the  $\Delta I = 1/2$  rule, the possible existence and properties of exotics, the proton-neutron and other electromagnetic mass differences, and many others, will become accessible as the available computational power increases. This will occur on a timescale of several years, given a reasonable extrapolation of current trends.
- 2.1.5. In order to do justice to these opportunities, one must both generate the necessary data, and analyze it. The former need drives requirements for capability, the

latter drives requirements for capacity. At any given time, these requirements must be balanced against one another.

## **2.2. Comments**

- 2.2.1. The committee enthusiastically endorses both the short- and long-term scientific potential of this endeavor. The committee notes that the field has reached the stage where for the computation of many quantities all errors are under quantitative control and that the techniques and error estimates have been validated on well-measured quantities.
- 2.2.2. The committee thinks that the group has maintained a sensible balance between capability and capacity. This issue will be discussed in depth below.
- 2.2.3. The committee believes this project should be encouraged and generously supported.

## **2.3. Recommendations**

- 2.3.1. In addition to exploiting existing opportunities, the group should facilitate exploratory studies in algorithms and comparative quantum field theory by allocating some time on the facility to this type of project. By comparative field theory, the committee means both variants of QCD (e.g., varying the number of colors and flavors, and quark representations, as well as quark masses) and also more radically different field theories (e.g., theories in different space-time dimensions, theories containing scalars, chiral gauge theories).
- 2.3.2. Visualization ought to be a powerful tool for understanding and finding surprises within the vast data set being generated. It also affords an opportunity to present the results to non-experts, including the interested public, in a memorable and attractive way. The team should develop a plan to incorporate specific visualization goals and approaches, as well as ensure sufficient visualization resources to make the approach feasible.
- 2.3.3. It is vital to the long-term health of the subject that young researchers get attracted into it. The team should consider ways in which this facility can be used to help the development of young researchers.

### **3. The status of the technical design, including completeness of technical design and scope, feasibility and merit of technical approach and appropriateness and effectiveness of relevant R&D**

The Committee was impressed with the thorough analysis presented on different hardware configurations and their potential for QCD research. The benchmarking strategy, which focuses on the average of the two major algorithms used, seems very well tailored to predicting realistic performance levels.

However, it appeared to the Committee that the efforts at the three laboratories were not as well integrated as they could have been and that the strategy of procuring a large system at FNAL and a smaller system at TJNAF was not optimal. The LQCD team stated that they needed about 50% of the total resources on capability class computers with performance of at least 2 Tflops, for computation of configurations. These resources needed to be complemented by about 20% of the resource on 1 Tflop-class computers with the rest in capacity for smaller analysis runs.

One aspect of integration that was included was a plan to turn the three laboratories into a Metafacility, where users eventually would see the three labs as a unified entity. However, this relies on software being developed outside the project and this risk needs to be carefully tracked by the project.

#### **3.1. Findings**

- 3.1.1. The LQCD project presented a coherent four year plan for the acquisition and usage of computing resources for the LQCD community. The plan includes approximately equal investment in capability and capacity resources. The plan envisions adding additional capability resources over time and older resources would be utilized as capacity. The projected budgets and anticipated Moore's Law improvements in computational power should allow for the yearly acquisition of new clusters at about the same delivered performance on LQCD applications as the aggregate of existing computing resources.
- 3.1.2. The computational resource plan starts off utilizing the QCDOC machine at Brookhaven and prototype clusters with more than six separate clusters and 1,348 CPUs (counting only those dated after 1/2003) at FNAL/TJNAF.
- 3.1.3. The acquisition plan includes competitive procurements for additional computing resources at FNAL and TJNAF each year. This is a total of eight separate

procurements. The procurements will probably focus on the acquisition of Linux clusters, at least for the first two years.

- 3.1.4. Doing multiple procurements every year also means that the project will support nine or more systems in the final years of the project.
- 3.1.5. The acquisition plan has most of the new computing resources procured and installed at FNAL.
- 3.1.6. In the schedule presented at the review, the first major procurement in FY06 had its delivery delayed until late in FY06 due to facilities upgrade at FNAL to handle anticipated hardware acquisitions in FY08 and FY09.
- 3.1.7. After the review, an error in the FNAL schedule was uncovered which could allow installation of a computer at FNAL in April 2006. In addition, FNAL has more archival storage capacity than TJNAF.
- 3.1.8. A high level integration plan was presented that allowed for multiple sites to field and use the computing resources and share data amongst the international LQCD community.
- 3.1.9. The integration plan included the building of a Metafacility that integrated the computing resources of the three sites into a single logical facility for the LQCD community over timeframe of the project.

### **3.2. *Comments***

- 3.2.1. The computing resource acquisition plan will meet the overall project scientific objectives, within the budget limitations.
- 3.2.2. The computing resource acquisition plan appears to have artificial constraints that inhibit the prompt delivery of hardware to meet the scientific objectives. The timing could be improved.
- 3.2.3. The acquisition of eight additional clusters over the four year period seems excessive.
- 3.2.4. The project's expectation to support 9 to 11 different systems in the latter years is not realistic.
- 3.2.5. The integration plan did not adequately show the architecture and the required hardware and software components for the LQCD simulation environments at the three Laboratories or how they could be integrated.

### **3.3. *Recommendations***

- 3.3.1. The committee recommends that the acquisition plan be modified to allow for a single joint acquisition, possibly every other year, alternating between the TJNAF and FNAL that would allow the delivery of resources to the program promptly in FY06 and beyond. The number of procurements should be reduced from eight to three or four.
- 3.3.2. If the FNAL construction schedule presented at the review, which delays the release of the computer there until September 2006, is accurate, the first computer delivered in FY 2006 should be put at TJNAF. If the revised FNAL schedule is accurate, which would enable the computer to be released to operation there in April 2006, the team should decide on the site for the computer based on where it can deliver the most science for the dollars invested.

- 3.3.3. The cluster integration plan should be written down and an architectural diagram with hardware and software components clearly indicated. The plan should also include the software development and integration work items necessary to bring these resources into production. This plan should be presented to the LQCD scientific advisory board for review and approval.
- 3.3.4. The LQCD project plan should be expanded to identify dependencies on SciDAC and other projects for technology necessary for building the Metafacility. A clear set of Level 1 and/or Level 2 deliverables and milestones (e.g., single integrated login, single batch system, file and data sharing) for the Metafacility should be included in the plan. This will facilitate overall risk assessment and mitigation in the project.

## **4. The feasibility and completeness of the proposed budget and schedule, including availability of manpower.**

The total project cost of approximately \$10M over the four year period is leveraged by approximately \$8M in contributions from other sources, principally laboratory base programs at FNAL, BNL, and TJNAF, as well as some funding from SciDAC. This leverage is a significant strength of the project and enables it to deliver significantly greater resources than a similar scale effort at a “green field” site. However, this leverage also presents a risk to the project, which must be tracked and managed. The team has based its plan on the assumption that only High Energy Physics funds can go to FNAL and only Nuclear Physics funds can go to TJNAF, which was contrary to direction from the DOE program Offices. The Committee was concerned that this assumption may have distorted the planning and resulted in a larger number of procurements than is advisable, as well as a schedule that is suboptimal in delivering science. The Committee urges the team to consider alternate procurement plans that improve the integration of efforts across the sites and improve the schedule for delivery of resources.

### **4.1. Findings**

- 4.1.1. The project plan has two procurements per year, one each at TJNAF and FNAL, with BNL operating the QCDOC system through the length of the project. This results in 9 to 11 separate systems by 2009.
- 4.1.2. The hardware procurement strategy plans to make available an aggregate 17.5 sustained Tflop/s in 2009, including 4.2 Tflop/s from the QCDOC system.
- 4.1.3. QCDOC will be ready for production runs by June 2005, and several SciDAC prototypes are also doing production level computing.
- 4.1.4. The project expects substantial subsidy funding from base and SciDAC and grid efforts. Using 2006 costs, the subsidy is at least, 2.3 FTEs from the base and 1.75 FTE from SciDAC, \$433K in electrical costs, unknown costs for facility modifications and space charges. The electrical costs are expected to increase at least as fast as Moore’s Law due to increased capacity. Thus, the subsidy will be close to \$2.5M in 2009 just for electricity, not including possible increases to the power rate itself.
  - 4.1.4.1. The project is relying on 1.75 FTE from SciDAC to provide the prototyping hardware evaluation that is critical to making good procurement decisions.
  - 4.1.4.2. The project expects an additional 2.3 FTEs contributed from the base programs across the laboratories.
- 4.1.5. The operational budget of \$677-745K contains \$70K for tape and disk costs. The storage needs for the project are 700TB from mid 2005 to mid 2006.

- 4.1.6. The project is heavily reliant on contributions from grid efforts (PPDG, OSG, etc), from CDF (Dcache), and from other sources. Failure of any of these projects will impact the ability of the project to meet its expectations. However, the reliance on Dcache is much more integral to success than reliance on grid software.
- 4.1.7. Based on the schedule presented at the review, the computing facility at FNAL is scheduled to be ready very late in FY 06, so no system can be planned at FNAL before September 2006. FNAL did discover, after the end of the review that there was a five month error in the construction schedule, which would allow a FNAL system to be installed in April 06.
- 4.1.8. The project presented qualitative arguments for the distribution of system across three sites, with little quantitative analysis.

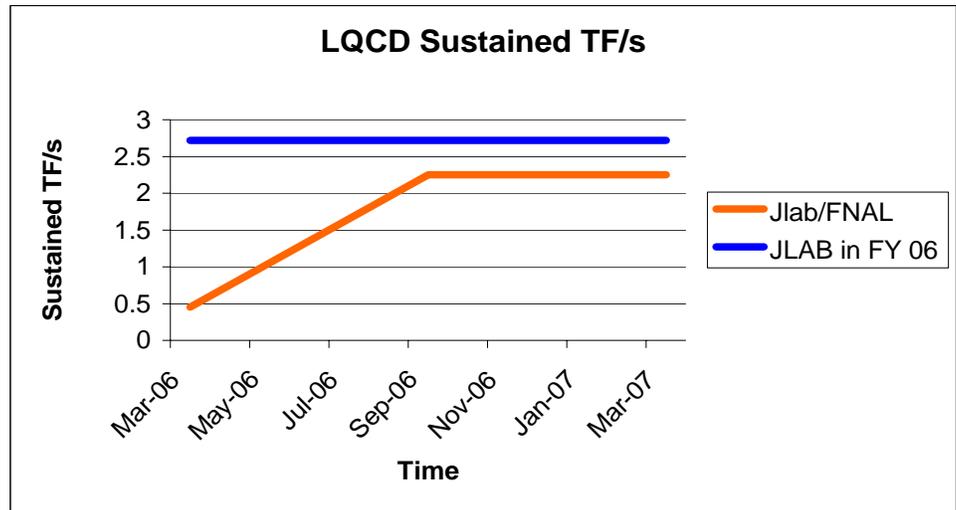
## **4.2. Comments**

- 4.2.1. The LQCD project is expecting substantial operational support from the host laboratories. The Committee estimated the value of these efforts for the LQCD project to be \$8M between 2006-2009.
- 4.2.2. The LQCD project will not be able to support 9 or more systems with the staffing proposed by the project. Even assuming that the proposed architecture minimizes the staffing to support users the staffing required for system integration and optimization for systems of this scale is too low.
- 4.2.3. If the number of systems is cut in half, the support by the 6.4 FTE (project and base) is minimally sufficient.
- 4.2.4. The project properly views sustained performance and sustained price performance as the primary metric for the selection of computer systems. Nevertheless, high rate LQCD production runs also need an appropriate overall computing environment for adequate quark propagator staging (see 3.2.5 and 3.3.2)
- 4.2.5. Because the prototyping is covered in the FY 05 SciDAC activity and will influence the FY 06 procurements, these are on track. Procurements for FY 07 and beyond may have increased risk depending on whether the SciDAC funding for this work continues. Examples of the type of increased risk include increasing the time needed to bring a system into production as well as procuring systems with suboptimal performance on QCD applications.
- 4.2.6. The budget for storage and consumables, at \$70,000 per year, appears low. The "Computational Requirements" presentation documented a storage need of 700 Terabytes (TB) between mid 2005 to mid 2006. Assuming that only 50% of this data has to be kept for at least a year and 80% of the 350TB is stored in an archive and the archive were able to use 500 GB tapes, which cost about \$80/tape in 2005/2006, then just tape media would cost \$38,400. That leaves only \$31,400 for all the disk, all the I/O servers, and all the tape drives. This is not feasible. The problem could be worse, since the project is expects to replicate the configuration data at both sites, which would substantially increase the archive storage needs.
- 4.2.7. The proposed budget for the hardware will in most likelihood be able to achieve the performance levels the project projects.

- 4.2.8. The project has the opportunity to adjust the timing of the procurements, enabling the project to making the most advantageous choice of the latest technology.
- 4.2.9. If the operational support from the host laboratories or the SciDAC-supported prototyping efforts were discontinued, the risk that the LQCD project not achieve its scientific goals will be significantly increased. The support from the host laboratories, for electricity alone is over \$2M in FY 2009. In addition, the ability of the team to conduct procurements and integrate systems in a timely fashion depends on the prototyping activities supported by SciDAC.

**4.3. Recommendations**

- 4.3.1. The project should consider alternative deployment strategies that result in fewer, larger systems over the same time period. This will reduce the required support effort to a feasible level within the project budget and associated subsidies.
  - 4.3.1.1. An example of an alternative deployment strategy is to have single system delivery once a year, alternating between FNAL and TJNAF.
    - 4.3.1.1.1. Because the facility work at FNAL was presented as being completed late in FY06, it appeared more effective to place a single larger system at TJNAF in early FY 06, and then a single larger system at FNAL in early FY 07. This would provide twice as much sustained computing between March 2006 to March 2007 as the schedule proposed by the team. The team should use the amount of science delivered per dollar as the guiding principle for making system siting decisions.



- 4.3.2. The project should provide a cost benefit analysis for one site, two sites and three sites as part of the planning.
- 4.3.3. The cost projections for storage and consumables should be done to the same level as the costs for computational resources in order to ensure the user requirements are met in a balanced manner.
- 4.3.4. The team should ensure wide impact of the valuable SciDAC-funded prototyping work with more timely publication of their results, both on the web site, but also

in more widely shared publications and conferences. This effort should also seek out collaborations with other architectural and performance evaluation efforts.

- 4.3.5. The project team should reevaluate the principles used to determine which costs are included within the project to ensure an accurate presentation of the overall cost of the effort to DOE.

## 5. Relevance of prototyping efforts outside the scope of the initiative and the status and plans for developing the required software for Lattice QCD computing.

The Committee was impressed with the thorough evaluation of computational hardware options carried out by the team. The team has a deep understanding of the impact of different hardware features on the performance that can be delivered. In addition the work on community software for QCD provides a high performance, portable set of software that complements the hardware efforts. The Committee strongly endorses the continuation of these efforts.

### 5.1. Findings

5.1.1. The goal of the prototyping efforts was to define a balanced hardware design tailored to the requirements of LQCD and to find the most cost-effective solution.

The hardware options investigated fall into three categories:

- 5.1.1.1. Processors: Virtually all current processors were benchmarked. The Intel IA32 architecture emerged as the most cost-effective solution at present, closely followed by the AMD Opteron. To avoid memory and I/O bottlenecks, it is currently best to use only one CPU in dual-CPU, server-class motherboards. The other processors investigated are listed below with comments: - PowerPC (excellent FPU performance, but memory bottleneck); Itanium (VLIW compilers not good enough yet, would require too much hand-coding effort); IBM Blue Gene Light BG/L (too expensive, watch future developments (BG/P?)); emerging commercial supercomputers (likely too expensive); QCDOC++ (perhaps in 2009); dual cores (just beginning to emerge, likely a future option); and exotic options such as Graphics Processing Units and Cell-based architectures.
- 5.1.1.2. Memory subsystem: As the memory subsystem is often the bottleneck, it is important to choose a CPU with a large cache and a fast memory (front-side) bus. The memory options investigated were DDR, DDR2, and Rambus. The clear winner is DDR2.
- 5.1.1.3. Network fabric: The communications network determines the sustained performance that can be achieved on a parallel machine. The two important factors are latency and bandwidth, with latency being the dominant factor for the LQCD algorithms on a machine with a large number of nodes. Several options were investigated, including GigE (switched, mesh), Myrinet, and Infiniband. Infiniband currently provides the best performance. In addition, it is essential to use PCI Express to avoid I/O bottlenecks.

- 5.1.2. The LQCD team also presented the status and the plans for developing the required software for lattice QCD. The low-level libraries (QMP, QLA, QDP/QDP++, QIO, optimized kernels) were written with SciDAC support. In addition, there is high-level applications software written in C (MILC) and C++ (Chroma).
- 5.1.3. Most of the required software already exists, is portable to all important architectures and provides good performance for typical lattice QCD applications on the machines considered. On the Intel IA32 architecture, it is essential to use SSE instructions for critical core sections of the applications.

## **5.2. Comments**

- 5.2.1. The prototyping efforts of the LQCD team are very professional. They have identified the most cost-effective solution for LQCD now, but they also recognize that this is a moving target.
- 5.2.2. The coherent software effort by the LQCD team is extremely important for the success of the project. Internationally, the U.S. lattice community is clearly the leader in this area, with a strong collaboration with the UK. The software funded by this project and by SciDAC is portable to all important architectures, transparent and easy to use for QCD physicists, and eliminates duplication of code. The low-level code (written jointly by the QCDOC group and the cluster experts) is essential to get the best performance out of the hardware. Most of this code exists and is already in production, with work on application software ongoing.

## **5.3. Recommendations**

- 5.3.1. The LQCD team should continue to monitor the market and benchmark the available options. The team should build on the existing collaboration of the participating labs in the prototyping effort to develop an integrated prototyping activity for LQCD. In the software area, the committee recommends the use of vendor-provided drivers to increase the communications performance (e.g., QMP over VAPI instead of MPI).
- 5.3.2. SciDAC support has been, and continues to be, absolutely essential for the success of the LQCD project. The leadership of the DOE in this area is recognized in the international lattice QCD community. Although this is somewhat outside the scope of this review, the committee recommends that the DOE consider continuing these efforts.

## **6. The effectiveness of the proposed management structure.**

The proposed management scheme is best described as a federation of efforts at the three laboratories. There are facilities for unified reporting and there is an advisory committee, which is responsible for the entire effort. The day to day management of the effort is primarily done by the individual laboratories. This approach has a number of strengths and ensures close coupling of the individual efforts to their separate laboratories. However, the Committee believes that there are significant benefits for the science in moving beyond the federated model described into a jointly planned model. Activities such as hardware evaluation, prototyping and procurement would be stronger if they were done jointly. In addition, the project relies on a number of external sources such as laboratory contributions, SciDAC software, and International Lattice Data Grid (ILDG) software. Even though these activities are not included in the budget for the project they must be included in the Work Breakdown Structure (WBS) so that their schedule can be tracked and appropriate contingency managed.

### **6.1. Findings**

- 6.1.1. The management is based on a three-laboratory scheme, with project monitoring handled by a central Project Manager. Funding goes directly to the three laboratory partners, based on distributions articulated in a Project Execution Plan (PEP) that will be approved by DOE HEP and NP. Day-to-day operations, including issues like cyber security and ESH compliance, are handled at the each site. The sites plan to make in-kind contributions to the project.
  - 6.1.1.1. Each laboratory site has a site manager.
  - 6.1.1.2. Each Lab does internal review of its own procurement RFPs, etc.
- 6.1.2. The management scheme includes an LQCD Executive Committee, a Scientific Program Committee to evaluate requests for time, and a Change Control Board (CCB).
  - 6.1.2.1. Allocations of computer resources will be managed by scientific program committee
  - 6.1.2.2. Executive Committee oversees entire project.
- 6.1.3. The three labs have established an integrated project management system.
  - 6.1.3.1. Change control procedures have been established with review by scientific program committee and change control board.
  - 6.1.3.2. Project reporting is monthly, coming from the sites to the Project Manager. The Project Manager is responsible for cost, schedule, technical performance reviews, and reporting to the Federal Project Manager.
- 6.1.4. The project has defined rough timetables for machine acquisition and delivered Tflops.

- 6.1.4.1. Project management plans to date are conceptual; with many details of the WBS, change control, risk management, and specific relations of the laboratories to the project remain to be defined.
- 6.1.4.2. Detailed FY 2006 Baseline still to be done.
- 6.1.5. External dependency on FNAL construction.
- 6.1.6. External dependency on SciDAC and other software projects.

## **6.2. Comments**

- 6.2.1. Control of operations costs is important to project goals. The project must understand the risks associated with the in-kind contributions of the sites.
- 6.2.2. Whether or not additional costs are associated with three-site model is difficult to assess due to the in-kind contributions of the sites. The committee concluded that the primary justifications for three-site model are not cost related, but connected to the importance of the interface with experiment. A multiple-site model is also attractive from a long-term view, as a regular cycling of machine upgrades among the three sites would be an efficient way to address community needs.
- 6.2.3. The project management systems establish a good basis for integrated reporting of progress and cost.
- 6.2.4. Use of scientific program committee to make allocations of computing resources is positive.
- 6.2.5. No detailed schedule for deployment of ILDG software and metafacility with contingency for schedule.
- 6.2.6. No explicit SciDAC contingency was displayed even though the team asserted that continued SciDAC support for prototyping was critical.
- 6.2.7. No schedule for detailed reviews of each outyear plan were displayed.
- 6.2.8. Sustained Teraflops, using the measurement proposed by the team, is a valid performance measure for these computers.
- 6.2.9. The relationship between Executive Committee and Scientific Program Committee and the relationship between the Executive Committee and the Contract Project Manager, who has budgetary and performance responsibility, are not clear.
- 6.2.10. The Project Management Team was a new organization that had only been in existence months prior to the review. The Contract Project Manager will need to develop an ownership and oversight of the entire project that is not limited by the perspective and needs of the laboratory at which he is sited.
- 6.2.11. In the WBS and project plan, Metafacility operations were only shown at TJNAF
- 6.2.12. Management plan includes close coupling with scientific community.
- 6.2.13. The committee strongly supports the DOE HEP/NP partnership that is supporting this Project.

### **6.3. Recommendations**

- 6.3.1. Operations agreements with the sites over the lifetime of the project should be executed, which cover all contributions that are not included in the project cost, so that risks associated with escalation of operations costs can be reduced.
- 6.3.2. The project should develop, and update on a yearly basis, a project-wide system deployment plan that optimizes the opportunity to deliver new science without artificial constraints on which programs can fund work at the three partner laboratories.
- 6.3.3. The strongly site-based management scheme reflects in part the history of forming this project. The laboratories should integrate their planning, prototyping and procurement activities. The approach to this should be in the revised PEP.
- 6.3.4. Ensure the WBS is a tool for integrated planning as well as integrated reporting. The reporting should also document the actual physics output measured in terms of the allocations made by the Scientific Program Committee.
  - 6.3.4.1. Incorporate schedules for integrated review of outyear plans into WBS to occur no later than June preceding beginning of FY.
  - 6.3.4.2. Expand procurement processes to include all three sites and possible external experts, including evaluation of joint procurements.
  - 6.3.4.3. Consider integrating technology tracking, hardware and software prototyping across all three sites.
  - 6.3.4.4. Since there are strong dependencies on some external efforts (SciDAC, ILDG, FNAL construction) schedule and contingency for these needs to be in WBS.
  - 6.3.4.5. The laboratories should report the monthly progress of each laboratory in providing the capabilities and capacity agreed to by the Scientific Program Committee.
- 6.3.5. Consider moving metafacility operations to integrated project office.
- 6.3.6. Charters for executive committee and Scientific Program Committee including how members are chosen should be produced and included in PEP.
- 6.3.7. Try to integrate CCB and Scientific Program Committee review of change proposals.

**Intentionally Blank**

# **APPENDIX A**

## **Charge Memorandum**

# memorandum

February 26, 2005

DATE: Office of Science  
REPLY TO: Technical, Cost, Schedule and Management Review of the Proposed Lattice Quantum  
ATTN OF: Chromodynamics (QCD) Computing Initiative  
SUBJECT:

TO: Ed Oliver, Associate Director, SC-30

This memorandum is to request that you organize and conduct a Technical, Cost, Schedule and Management Review of the proposed Lattice Quantum Chromodynamics (QCD) Computing Initiative. This review should appropriately involve the input and participation of the science programs in the Office of High Energy Physics and Office of Nuclear Physics responsible for the effort.

QCD successfully describes the fundamental strong interactions between quarks and gluons. Although the equations that define this theory are exact, none of the analytical methods that are successful elsewhere in theoretical physics are adequate to solve them for all regions of QCD's domain of validity. The lack of precision in current QCD calculations now limits the understanding of many experimental results in high-energy and nuclear physics, including many measurements at the Stanford Linear Accelerator Center B-Factory, the Fermilab Tevatron, the Brookhaven Relativistic Heavy Ion Collider (RHIC) and the Thomas Jefferson National Accelerator Facility, as well as at non-DOE facilities such as the Cornell Electron Storage Ring (CESR) and Japan's KEK B-factory. It has long been known that some aspects of QCD can be simulated on a space-time lattice to high precision, given enough computational power. Recent advances in numerical algorithms coupled with the ever-increasing performance of computing have now made a wide variety of QCD calculations feasible.

Starting in FY 2006, the Office of High Energy Physics and the Office of Nuclear Physics plan to create a large-scale (~20 TFlops, preliminary Total Estimated Cost range ~ \$7-10 million) Lattice QCD computing capability based on the most cost-effective technology available. Our office must be assured that the technical approach and planning for all aspects of Lattice QCD computing are optimized to maximize scientific productivity in the context of other efforts world-wide and constrained budgets.

In particular, it is requested that your review evaluate:

- The significance and merit of the proposed initiative;
- The status of the technical design, including completeness of technical design and scope, feasibility and merit of technical approach;
- The feasibility and completeness of the proposed budget and schedule, including availability of manpower; and
- The effectiveness of the proposed management structure.

In addition, it is requested that you assess the appropriateness and effectiveness of relevant R&D

and prototyping efforts outside the scope of the initiative, and the status and plans for developing the required software for Lattice QCD computing. The report should be submitted to the Office of High Energy Physics and the Office of Nuclear Physics by May 15<sup>th</sup> in order to influence planning for the FY 2007 budget.

Robin Staffin  
 Associate Director  
 Office of High Energy Physics

Dennis Kovar  
 Associate Director  
 Office of Nuclear Physics

SC-93:JSimon-Gillo:cls:2/25/05:3-  
 3613:q:\simongillo\Projects\LQCD\Review05\LQCD\_charge\_022505.doc

SC-93	SC-92	SC-92	SC-90	SC-20	SC-20	SC-20
JSimon-Gillo	DKovar	SCoon	GHenry	GCrawford	JMandula	RStaffin
2/ /05	2/ /05	2/ /05	2/ /05	2/ /05	2/ /05	2/ /05

# **APPENDIX B**

## **Review Participants**

**Department of Energy Review of the  
Lattice Quantum Chromodynamics (LQCD) Project**

**REVIEW COMMITTEE PARTICIPANTS**

**Department of Energy**

Daniel A. Hitchcock, DOE/SC-21, Chairperson

**Consultants**

Maarten Golterman, San Francisco State University  
Wick Haxton, University of Washington  
William T.C. Kramer, LBNL  
Dr. Klaus Schilling, University of Wuppertal, Germany  
Mark Seager, LLNL  
Tilo Wettig, Regensburg University, Germany  
Frank Wilczek, Massachusetts Institute for Technology

**Observers**

Sidney A. Coon, DOE/SC-26.1  
John Kogut, DOE/SC-25.2/ University of Illinois Urbana Champaign  
Jeffrey Mandula, DOE/SC-25.1  
Jehanne Simon-Gillo, DOE/SC-26.2

# **APPENDIX C**

## **Review Agenda**

Agenda for LQCD Computing Project Review  
MIT LNS, Boston, MA  
May 24-25, 2005

May 24, 2005		
<b>Executive Session</b>	8:00 am	Dan Hitchcock
<b>MIT LNS Welcome</b>	8:45 am	June Matthews, MIT LNS Director
<b>LQCD Team Presentations</b>	9:00 AM	
<b>Scientific Overview</b>	20 min	Bob Sugar Scientific case for the project
<b>Project Overview</b>	20 min	Chip Watson High level WBS and project context
<b>Technical Presentations</b>	9:40 AM	
Computational Requirements	20 min	Steve Gottlieb Deliver high level concepts on computing reqs.
New Systems	30 min	Don Holmgren Technology and procurement strategy, performance, near term tech. expectations
<b>BREAK</b>	10:30 AM – 10:45AM	
FY06 Clusters	20 min	Don Holmgren Details of WBS for clusters
Operations	40 min	Chip Watson Staffing, meta-facility operations, user support
<b>LUNCH</b>	12:00 – 1:00	
<b>Related Projects</b>		
QCDOC Development	20 min	Norman Christ Current QCDOC machine, future R&D
SciDAC Prototypes	30 min	Don Holmgren Hardware prototyping
SciDAC Software R&D	10 min	Richard Brower Software context
ILDG	5 min	Chip Watson International Lattice Data Grid
<b>Project Management</b>	30 min	Don Holmgren
<b>Cost and Schedule</b>	30 min	Don Holmgren
<b>Executive Session</b>	3:00 PM	
<b>Questions to LQCD Team</b>	5:30 PM	

May 25, 2005		
<b>Responses to Questions</b>	8:00 AM	
<b>Executive Session</b>	9:30 AM	
<b>Debrief to LQCD Team</b>	1:00 PM	
<b>Adjourn</b>	2:30 PM	

# **APPENDIX D**

## **Installation Schedule**

## LQCD Proposed Cluster Installation Schedule (Based on FNAL Building completed 7/1/2006)

FNAL		Capacity (TF)	Period (Months)	Delivered (TF- yr)
FY06		0.7	12	0.70
		0.15	12	0.15
		1.8	0	0.00
			Total	0.85
FY07		0.7	12	0.70
		0.15	12	0.15
		1.8	12	1.80
		2.2	7	1.28
			Total	3.93
FY08		0.15	3	0.04
		0.7	3	0.18
		1.8	12	1.80
		2.2	12	2.20
		3.6	7	2.10
			Total	6.31
FY09		1.8	12	1.80
		2.2	12	2.20
		3.6	12	3.60
		2.9	7	1.69
			Total	9.29

TJNAF		Capacity (TF)	Period (Months)	Delivered (TF- yr)
	FY06	0.5	12	0.50
		0.3	12	0.30
		0.05	12	0.05
		0.45	7	0.26
			Total	1.11
	FY07	0.5	12	0.50
		0.3	3	0.08
		0.45	12	0.45
		0.6	7	0.35
			Total	1.38
	FY08	0.5	3	0.13
		0.45	12	0.45
		0.6	12	0.60
		0.9	7	0.53
			Total	1.70
	FY09	0.45	12	0.45
		0.6	12	0.60
		0.9	12	0.90
		0.7	7	0.41
			Total	2.36

# **APPENDIX E**

## **Acronyms**

### Acronyms

AMD Opteron	64 Bit CPU Chip from Applied Micro Devices
CKM	Cabibbo-Kobayashi-Maskawa (CKM) matrix elements related to CP violation
Dcache	Distributed Data Cache software, a joint effort of FNAL and DESY
DDR	Double Data Rate memory with speeds from 200 MHz to 333 MHz
DDR2	A new memory standard promoted by Intel. Potentially, it enables to reach higher frequencies and higher bandwidth.
FPU	Floating Point Unit
GigE	Gigabit Ethernet
IA32	Intel 32 Bit instruction architecture
ILDG	International Lattice Data Grid project
Infiniband	InfiniBand is an interconnect or I/O architecture that connects servers with remote storage and networking devices, and other servers. It can also be used inside servers for inter-processor communication. InfiniBand is a channel-based, switched fabric, point-to-point interconnect, which provides scalability and performance for a wide range of platforms and price performance points. InfiniBand provides a scalable performance range of 500 MB/s to 6 GB/s per link, meeting the needs from entry level to high-end enterprise systems
Itanium	Intel 64 Bit CPU Chip
MPI	Message Passing Interface
Myrinet	Myrinet, ANSI/VITA 26-1998, is a high-speed local area networking system designed by Myricom to be used as an interconnect between multiple machines to form computer clusters. Myrinet has much less protocol overhead than standards such as Ethernet, and therefore provides much better throughput and less latency while using the host CPU much less frequently.
PCI Express	an emerging (2004/2005) standard for high-speed graphics, likely to result in a 20% boost over 2003-era AGP 8x performance. The standard, supported by ATI and other vendors, delivers better power management, bi-directional simultaneous I/O and 4GB/s bandwidth
PowerPC	A family of RISC-based computer processors (chips) developed jointly by IBM, Apple Computer, and Motorola Corporation and used in IBM RS/6000 systems and Apple Macintosh computers
PPDG	Particle Physics Data Grid
OSG	Open Science Grid

QCDOC	QCDOC architecture has been designed to provide a highly cost-effective, massively parallel computer capable of focusing significant computing resources on relatively small but extremely demanding problems. This new design is a natural evolution of that used in our earlier QCDSF machines. The individual processing nodes are PowerPC-based and interconnected in a 6-dimension mesh with the topology of a torus. A second Ethernet-based network provides booting and diagnostic capability as well as more general I/O. The entire computer is packaged in a style that provides good temperature control and a small footprint. Central to this design is the IBM Blue Logic technology which makes possible the high-density, low-power combination of an industry standard RISC processor with 64-bit floating point, embedded DRAM, six-dimensional interprocessor communications and the wide array of predesigned functions needed to assemble a complete, functional unit.
QDP/QDP++	SciDAC Data-Parallel Programming Interface for C and C++ computer languages
QIO	SciDAC QIO/C intermediate level input-output package
QLA	SciDAC QLA linear algebra library
QMP	The QMP project is a national effort to provide a high performance message passing interface on various hardware platforms for Lattice QCD computing. This message passing interface aims to provide channel oriented communication end points to communication readers and writers with low latency and high bandwidth. QMP is tailored to the repetitive and predominantly nearest neighbor communication patterns of lattice QCD calculations.
Rambus	Rambus is a high-speed memory technology that uses a narrow 16-bit bus (Rambus channel) to transmit data at speeds up to 800MHz
SSE	SSE instructions are SIMD for single-precision floating-point numbers. SSE instructions operate on four 32-bit floats simultaneously.
VAPI	InfiniBand verbs applications programming interface
VLIW	Very Long Instruction Word, instruction sets with large-sized complex instructions encoded into one instruction.