BY09 Alternatives Analysis
for the
Lattice QCD Computing Project (LQCD)

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 17, 2007

Revision 1.5

PREPARED BY:
Chip Watson, TJNAF
Don Holmgren, FNAL

CONCURRENCE:

William N. Boroski
LQCD Contract Project Manager

8-17-2007
Date

Lattice QCD Computing Project (LQCD)
Change Log: Alternatives Analysis FY09 for the FY06-FY09

Revision No.	Description	Effective Date
1.0	Entire Document	May 1, 2006
1.1	Address e300 PIP	Aug 11, 2006
1.2	Additional explanatory text added	Sept 13, 2006
1.3	Update for FY09 from the FY08 document	June 2007
1.4	Further FY09 updates (Cost avoidance and benefit NPV's)	July 6, 2007
1.5	Revised cost avoidance and estimated benefit NPV's	August 17, 2007

Table of Contents

1 Introduction

This document presents the BY09 analysis of alternatives for obtaining the computational capacity needed for the US Lattice QCD effort within High Energy Physics (HEP) and Nuclear Physics (NP). This analysis is updated at least annually to capture decisions taken during the life of the project, and to examine options for the remainder of the project. The technical managers of the project are also continuously tracking market developments through interactions with computer and chip vendors, through trade journals and online resources, and through computing conferences. This tracking allows unexpected changes to be incorporated into the project execution in a timely fashion.

Alternatives herein are constrained to approximately fit within the current budget guidance of the project, a total of $2.5M / year from the HEP and NP program offices for the first three years of the project (FY06-FY08) and $1.7M for the final year (FY09). This constraint provides adequate funding to meet the basic requirements of the field for enhanced computational capacity, under the assumption of expanding resources at NERSC, ORNL, and ANL already planned by the Office of Science (SC), and under the assumption that a reasonable fraction of those resources are ultimately allocated to Lattice QCD.

All alternatives assume the continued operation of the existing QCDOC at BNL through the end of this project, the operation of clusters at FNAL and JLab procured for LQCD under the first SciDAC LQCD project, and under this project in prior years (FY06-FY07) up to an age of 3.5 years. The allocated project cost of operating these facilities is approximately $0.8M/year (for the three sites combined). Replacing the computational capacity represented by existing resources cannot be done for less than its operating cost.

2 FY08 and FY09 Goals

The proposed objective for the combined FY08 and FY09 procurements is to assemble new computational resources that sustain a total of 6.2 teraflop/s for production lattice QCD calculations. Sustained performance is defined as the average of single precision DWF and single precision improved staggered actions. “Linpack” or peak performance metrics are not considered, as lattice QCD codes uniquely stress computer systems and their performance does not uniformly track either “Linpack” or peak performance metrics across different architectures.

A second goal is to deliver from all resources 12 teraflop/s-years of running in FY08, and 15 teraflop/s-years in FY09.

Beyond FY09, the objective for a second phase of this project is to take advantage of the improvements in technology implied by Moore's law, as well as the specific nature of QCD calculations, to deploy a series of increasingly powerful capabilities for science.

3 Alternatives

The following sections summarize the alternative technologies considered to achieve the stated performance goals of this investment.

3.1 Alternative 1: A Single Optimal Cluster Deployed in Quarter 4, 2008.

Deploy an application-optimized cluster in the fourth calendar quarter of 2008, using funds from FY08 and FY09, to sustain 6.2 teraflop/s.

Based upon current market trends (3.5% average improvement per month in price/performance) this alternative will cost about $0.36 per sustained megaflop/s in FY08 ($2.25M hardware plus $0.25M labor, site prep and overhead). In addition, estimated annual operating costs are approximately 10% of the original hardware purchase cost, or $0.23M/yr, over an expected 3.5 year lifetime.

The incremental lifecycle cost of this alternative is estimated as follows:

- Procure 6.2 TF in FY08 for deployment in FY09-Q1 ($2.5M).
- Operations at 10%/yr ($0.25M/yr)
- Incremental lifecycle cost: $2.5M + ($0.25M/yr*3.5yrs) = $3.38M.

The total lifecycle cost for the investment, with this alternative, is estimated as follows:

- Project costs incurred in FY06-07: $5M
- Operate existing QCDOC machine at BNL and computing clusters at FNAL and TJNAF in FY08/09: $0.89M
- Incremental lifecycle cost for Alternative 1: $3.38M
- Total Lifecycle Cost Estimate: $5M + $0.89M + $3.38M = $9.27M.

Risk adjustment: This procurement will be made using a fixed price contract based upon allocated funds, so there is minimal cost risk. The Moore's Law extrapolation of cost per teraflop/s is only assumed for up to 6 months ahead of the planned procurement date so as to minimize the risk to the project's goals. This is because Moore's Law is not continuous but has discrete steps with occasional 6 month "flat spots". Another way to state this is that there is a planned 23% contingency on the performance goal. In summary there is minimal cost risk, and modest uncertainty in how much the investment may exceed goals. It should be noted that this consideration of risk uniformly applies to all alternatives, and so has no impact upon the alternatives analysis outcome.

Justification for the expectation of increases in cluster performance/dollar: (1) quad core processors will be commoditized; (2) DDR-3 1600 memory will be commoditized; (3) DDR Infiniband will be commoditized; and (4) QDR Infiniband may be commoditized.

3.2 Alternative 1a: An Optimal Cluster Each Year

Deploy application-optimized clusters in the third quarter of each calendar year to sustain 3.4 teraflop/s (FY08) and 2.2 teraflop/s (FY09).

Based upon current market trends (3.5% improvement per month in price/performance) this alternative will cost about $0.41 per sustained megaflop/s in FY08 ($1.4M hardware plus $0.25M labor, site prep and overhead) and $0.27 per megaflop/s in FY09 ($0.6M in hardware and $0.25M in labor, site prep and overhead). In addition, estimated annual operating costs are approximately 10% of the original hardware purchase cost over an expected 3.5 year lifetime.

The incremental lifecycle cost of this alternative is estimated as follows:

- Procure 3.4 TF in FY08 ($1.65M) and 2.2 TF in FY09 ($0.85M).
- Operations at 10%/yr (FY08 machine: $0.17M/yr; FY09 machine: $0.09M/yr)
- FY08 machine lifecycle cost: $1.65M + ($0.17M/yr*3.5yrs) = $2.25M.
- FY09 machine lifecycle cost: $0.85M + ($0.09M/yr*3.5yrs) = $1.17M.
- Incremental lifecycle cost = $2.25M + $1.17M = $3.42M.

The total lifecycle cost of the investment, with this alternative, is estimated as follows:

- Project costs incurred in FY06-07: $5M
- Operate existing QCDOC machine at BNL and computing clusters at FNAL and TJNAF in FY08/09: $0.89M
- Incremental lifecycle cost for Alternative 1a: $3.42M
- Total Lifecycle Cost Estimate: $5M + $0.89M + $3.4M = $9.3M.

Risk Adjustment: This procurement will be a fixed price contract based upon allocated funds, so cost risk is minimal. The Moore's Law extrapolation of cost per teraflop/s is only taken up to 6 months ahead of the planned procurement date to minimize the risk to the project's goals. This is because Moore's Law is not continuous but has discrete steps with occasional 6 month "flat spots". Another way to state this is that there is a planned 23% contingency on the performance goal. In summary, there is very minimal cost risk, and an uncertainty in how much the investment may exceed goals. It should be noted that this consideration of risk uniformly applies to all alternatives, and so has no impact upon the alternatives analysis outcome.

This option does impose a schedule risk not present in Alternative 1. Space to house a cluster deployed in the third quarter of FY08 may not be available to the project, as the current schedule for computer facility construction indicates that the newly-configured space may not be available for beneficial occupancy until early FY09.

Justification for the expectation of increases in cluster performance/dollar: (1) quad core, and later 8-core, processors will be commoditized, (2) DDR-3 1600 memory will be commoditized, (3) DDR Infiniband will be commoditized, and (4) QDR Infiniband may be commoditized.

Additional Comments: Although the cost of this alternative is approximately the same as Alternative 1, it is inferior in that it (1) delivers fewer Teraflop/s; (2) has modest schedule risk; and (3) consumes more high level staff effort, which would be better used elsewhere. In addition, since the FY08 and FY09 funds are used to buy two machines, the phased deployment will somewhat increase the complexity seen by the users, which may somewhat decrease their productivity.

3.3 Alternative 2: Traditional Supercomputers

> *Expand the major DOE Supercomputer Centers, National Energy Research Scientific Computing Center (NERSC, Lawrence Berkeley Lab) and the Center for Computational Sciences (CCS, Oak Ridge National Lab) to meet the needs of the QCD physics calculations (an additional 6.2 teraflop/s roughly in the timescale of quarter 4 of 2008), while continuing to operate the existing systems (QCDOC, FY06 and FY07 project clusters, and remaining SciDAC clusters).*

To estimate the price/performance of general use commercial supercomputers, information from recent upgrades are used. NERSC estimates $6 per sustained Mflop/s for a system installed one year ago (ASCAC meeting, Aug 2006), which would fall to $2.6 per sustained Mflop/s in late FY08 and $1.7 per sustained Mflop/s in FY09.

The incremental lifecycle cost of this alternative is estimated as follows:

- Procure 6.2 TF in late FY08 ($16M), with the cost split across two fiscal years ($8M/yr).
- Operations at 10%/year: (FY08: $0.8M, FY09: $1.6M).
- Incremental 2 year project cost for this option: $18.4M.
- Incremental lifecycle cost: $16M + ($1.6M/yr*3.5yrs) = $21.6M

The total lifecycle cost for the investment, with this alternative is estimated as follows:

- Project costs incurred in FY06-07: $5M
- Operate existing QCDOC machine at BNL and computing clusters at FNAL and TJNAF in FY08/09: $0.89M/yr * 2 yrs = 1.8M.
- Incremental lifecycle cost for Alternative 2: $21.6M
- Total Lifecycle Cost Estimate: $5M + $0.89M + $21.6M = $27.5M.

Analysis: This option falls considerably outside of the investment budget envelope. High-end clusters, such as the NERSC clusters and the XT-3 at ORNL, which are configured to support a wide range of high end computing applications, contain components (cost factors) which are not needed or are not cost effective for lattice QCD, including higher performance message passing fabrics, higher performance and capacity local disks, and much larger memories. In addition, they are usually integrated with much higher performance file services than that required by the LQCD community.

3.4 Alternative 3: BlueGene/P Supercomputer

Purchase (or expand) an IBM BlueGene/P supercomputer of sufficient size to sustain (an aditional) 6.2 teraflop/s in late 2008.

Purchase the most cost effective commercial supercomputer (most likely BlueGene/P), locate it at one of the labs doing the experiments, and dedicate it to LQCD physics calculations. Or, add additional racks at an existing DOE BG/P site (e.g. ANL). In FY08, the BlueGene/P machine will be available. Based upon some knowledge of the planned machine at ANL, this machine will cost $26M for 100 peak teraflop/s. The BG/P is expected to sustain no more than 25% of peak on lattice QCD, based upon conversations with the developers and lattice QCD experience on the BG/L. Thus, estimated cost will be in the $1 per megaflop/s range. A 6.2 Tflop/s machine, or partition of a larger machine, will therefore cost $6.2M. If purchased in late FY08, the cost could be split into two fiscal years ($3.1M per year).

The support contract would be included for the first year, and would be approximately 8% of the purchase price in subsequent years. Other annual operating costs are estimated at approximately 2% of the original purchase price. Power requirements for this machine are modest, about $0.1M/year. Machine lifetime is expected to be about 5 years.

The incremental lifecycle cost of this alternative is estimated as follows:

- Procure 6.2 TF in FY08 ($6.2M)
- Support contract free in FY08, 8% in FY09: $0.25M
- Operations at 2%/yr: FY08 = $0.06M: FY09 = $0.12M: Total = $0.18M
- Incremental 2 year project cost for this option: $6.63M.
- Incremental lifecycle cost = $6.2M + ($0.51M/yr*3.5 yrs) = $8M.

The total lifecycle cost of the investment, with this alternative, is estimated as follows:

- Project costs incurred in FY06-07: $5M
- Operate existing QCDOC machine at BNL and computing clusters at FNAL and TJNAF in FY08/09: $0.89M
- Incremental lifecycle cost for Alternative 3: $8M
- Total Lifecycle Cost Estimate: $5M + $0.89M + $8M = $13.9M.

While more cost effective than traditional supercomputers, this alternative is still too expensive, and if constrained to the project budget, would yield less than half of the incremental capacity that clusters will deliver, at almost twice the average annual cost over its lifetime. Stated another way, it would produce less science per dollar spent.

3.5 Alternative 4: Status Quo (no additional deployments in FY08 and FY09)

Continue to operate the HEP/NP QCDOC deployed at BNL in 2005 and clusters deployed at FNAL and TJNAF through FY2007 for up to 3.5 years age as part of the 3-site facility.

Clusters deployed through FY2007 include SciDAC prototype clusters deployed in 2005, and production clusters deployed at Fermilab and Jefferson Lab in 2006-2007. Note that this option does not meet project goals. It is included only for completeness and would not be capable of providing the necessary computational capacity to achieve the scientific goals of this project. The cost of this choice is $1.6M to operate the existing facilities as a coherent resource for the final two years of the current project (FY08, FY09). The incremental cost of this alternative (new investment) is $0.

The total project cost with this alternative is estimated as follows:

- Project costs incurred in FY06-07: $5M
- Operate existing QCDOC machine at BNL and computing clusters at FNAL and TJNAF in FY08 and FY09: $0.8M/yr * 2 yrs = $1.6M
- Total Project Cost: $5M + $1.6M = $6.6M.

3.6 Other Alternatives

Other alternatives may be relevant for future iterations of this document. These were not considered for detailed analysis at this time, as their current state of maturity was not deemed sufficient. These alternatives include:

- IBM/Sony Cell processor based systems: memory bandwidth considerations do not make the current processors sufficiently promising, but as follow on commercial products emerge additional analysis should be performed. The next generation will have full IEEE floating point, and higher memory bandwidth, but may still not surpass quad core commodity CPU's.
- Other novel accelerators exist, such as ClearSpeed's CSX600 chip on their Advance accelerator board, and on the IBM blade system at Los Alamos. Like the Cell processor, this chip does not appear to have enough memory bandwidth to sustain the high performance necessary for lattice QCD.
- Graphics Processing Units: Potentially more interesting is the emergence of high performance and programmable GPU's (graphics processing units) with highly parallel vector processing capabilities. These may have sufficient memory bandwidth to allow their use in a mixed CPU/GPU cluster system.

4 Net Present Value Considerations

Alternatives 1, 2 and 3 above have the same net present value considerations in that all alternatives have the same computational capacity increment for FY2008 and FY2009, and so yield the same stream of benefits. Thus NPV calculations will have no impact upon the selection of the best alternative, and will give the same rank ordering of the alternatives as comparing costs. For completeness, however, we include NPV calculations for alternatives 1, 2, and 3 based on both cost avoidance and on estimated benefits (see below). Note that the "Baseline: status quo" alternative is rejected in that it does not meet stated goals for capacity. Alternative

1a is very close to alternatives 1, 2 and 3 in benefits, but unnecessarily increases risk and so is not separately evaluated.

For both the cost avoidance and estimated benefits NPV calculations, we assume that the hardware in each alternative is purchased in two phases separated by 30 days (end of FY08 and beginning of FY09, i.e., no lease arrangements), and that the hardware becomes operational at the end of calendar 2008. First, we compare the cost of the chosen alternative (Alternative 1: optimal clusters) to the next most cost effective alternative (Alternative 3: IBM BlueGene/P) and use the difference in each year (acquisition and operations costs in the first year, and operations costs in the subsequent years) as our cost avoidance. In the second analysis, we use the estimated benefits of $10M for the FY08 machines, and $6.8M for the FY09 machine, from the discussion of Return on Investment (see Section 5).

The following table shows the NPV calculated using cost avoidance.

Cost Avoidance NPV						
Discount Rate	4.90%					
	FY08 CA	FY09 CA	FY10 CA	FY11 CA	FY12 CA	NPV
1: Cluster	$1.780	$2.922	$0.646	$0.673	$0.526	$5.882
2: Traditional SC	-$5.390	-$5.880	-$1.021	-$1.064	-$0.832	-$12.900
3: BlueGene/P	$0.000	$0.000	$0.000	$0.000	$0.000	$0.000

Table 2 below shows the NPV calculated using an assumed benefit of $10M over a four year lifetime of the system purchased in FY08, and $6.8M for the system purchased in FY09. See "Return on Investment" below for discussion of this assumed benefit. For the NPV calculation, the $10M and $6.8M benefits are spread evenly across the 3.5 years, with the new systems becoming operational at the end of calendar 2008.

FY08/FY09 Machines Estimated Benefit						
	FY08 Net	FY09 Net	FY10 Net	FY11 Net	FY12 Net	NPV
1: Cluster	-$2.500	$1.900	$3.860	$3.821	$2.835	$8.074
2: Traditional SC	-$9.670	-$6.902	$2.193	$2.083	$1.477	-$10.707
3: BlueGene/P	-$4.280	-$1.022	$3.214	$3.147	$2.309	$2.192

5 Return on Investment

ROI is difficult to quantify, but the following discussion gives an order of magnitude estimate of the benefit of this investment to the HEP and NP programs.

This investment provides two classes of benefits to the High-Energy Physics (HEP) and Nuclear Physics (NP) programs of the DOE's Office of Science (SC). The first class is the direct enhancements to the science itself: these theoretical calculations are important on their own and will lead to new discoveries. The second is that these calculations are in some cases essential to the cost effective exploitation of much more expensive experiments built and operated by the two program offices. In the FY07 Operating Plan, the total HEP and NP programs in SC were funded at $752M and $432M, respectively. Further, both fields of science receive substantial, though smaller, grants from the National Science Foundation. This should be compared to the budget of this project, $2.5M/year in FY06-FY08, and $1.7M in FY09. In HEP, roughly 30% of the Tevatron program at Fermilab has a direct interplay with lattice QCD calculations. Furthermore, the entire PEP-2/BaBar B physics program at SLAC, and the entire (NSF-funded) CLEO-c program at Cornell depends on lattice QCD for a full understanding of the experimental measurements. The whole suite of measurements and calculations are worth much more together than in isolation, so one must conclude that the return on investment for HEP is at least five-fold, possibly even twenty-fold. In NP, the situation is much the same. A significant development at BNL's Relativistic Heavy-Ion Collider (RHIC) is to search for the critical point of the QCD phase transition. Lattice QCD calculations indicate that this search is within RHIC's reach; RHIC would not proceed without this guidance. At Jefferson Lab a key motivation for the upgraded accelerator is the search for hybrid mesons and gluonic excitations, states whose theoretical foundation rests on lattice QCD. One concludes again that the return on investment for NP is at least five-fold, possibly even twenty-fold. With such high rates of return, it is safe to view the calculations as necessary for the DOE to do a sensible deployment of the experiments. But one should then ask whether other computing facilities could do the job. Indeed, all of the experiments in question have computing budgets that rival or surpass this project. However, their communications networks are ill-suited to the data structures of lattice QCD, with a mismatch in efficiency of nearly a factor of 10. In the past, LQCD has, therefore, been carried out at supercomputer centers. Compared to this project's computing facilities, the costs at supercomputer centers are two to eight times greater to deliver the same amount of dedicated lattice QCD computing.

As a conservative estimate of the benefits resulting from the $2.5M investment in FY08, we use only a four-fold ROI and thus a benefit of $10M over the 3.5 year lifetime of the systems. This $10M figure will be used in the BY09 Exhibit 300 submission to the OMB in the Alternatives Analysis section, with the FY08 lifecycle costs cited in the Alternatives sections above as the risk adjusted cost. Similarly, the benefits resulting from the $1.7M investment in FY09 are estimated at $6.8M over the 3.5 year lifetime.