National Computational Infrastructure for Lattice Gauge Theory

Principal Investigators

N. Christ (Columbia U.), M. Creutz (BNL), P. Mackenzie (Fermilab),
J. Negele (MIT), C. Rebbi (Boston U.), S. Sharpe (U. Washington),
R. Sugar (UCSB) and W. Watson, III (JLab)

October 27, 2005

Abstract

This initiative will develop the software and hardware infrastructure necessary for large scale numerical simulations of QCD, the fundamental theory governing the strong interactions. Following the U.S. Department of Energy's Scientific Discovery through Advance Computing Initiative (SciDAC) Notice 01-11 and Lab 01-11, the U.S. Lattice QCD community in association with Brookhaven National Laboratory, Fermi National Accelerator Laboratory, and Thomas Jefferson National Accelerator Facility will undertake a software development and hardware prototyping effort needed for lattice QCD. In the future, a proposal will be made for a distributed multi-terascale topical computing center using the infrastructure developed within this proposal. The calculations made possible by these new, terascale machines will dramatically advance our quantitative understanding of QCD, providing important theoretical insights and vital support for the Department of Energy's large experimental efforts in high energy and nuclear physics.

Contents

1 Background and Significance: Introduction

The long range objective of our collaboration is to create the hardware and software infrastructure needed for terascale simulations of quantum chromodynamics (QCD), the sector of the Standard Model of elementary particle physics that describes the strong interactions. Such simulations are necessary for understanding some of the most fundamental quantities in high energy and nuclear physics, thus supporting the Department of Energy's large experimental efforts in these fields. In this proposal we request funds for a three year software development and hardware prototyping effort. In the future we plan another proposal to the Department of Energy for a distributed topical computing center to provide the terascale platforms on which these simulations will be performed.

The twentieth century was an era of striking progress towards comprehending the fundamental structure of matter, beginning with the discovery of quantum mechanics and atomic physics, progressing to nuclear physics, and culminating with the Standard Model of elementary particle physics. However, traditional analytical tools have proven inadequate to extract many of the predictions of QCD. Our understanding of nature will remain fundamentally deficient until we know how the rich and complex structure of strongly interacting matter, which comprises most of the known mass of the universe, arises from the interactions among quarks and gluons.

At present, the only method to extract non-perturbative predictions of QCD from first principles and with controlled systematic errors is through large scale numerical simulations of lattice gauge theory. Recent refinements of numerical algorithms coupled with major increases in the capabilities of massively parallel computers have brought these simulations to a new level. It is now possible to calculate a few crucial quantities to an accuracy comparable with their experimental determination. The strong coupling constant and the masses of the c and b quarks are notable examples. Furthermore, the experience we have gained allows confident predictions for the computing resources required for accurate determinations of a broad range of fundamental quantities.

United States physicists invented lattice gauge theory and play a leadership role in the field. To maintain our status and provide needed support for the U.S. high energy and nuclear experimental programs, we must move quickly to construct the substantial computational infrastructure required for terascale simulations. Unfortunately, the U.S. is falling behind Europe and Japan in building such infrastructure. This proposal represents a first step towards a coherent national plan in support of the next level of scientific discovery in lattice gauge theory. The envisioned terascale computing resources will enable calculations essential to precision tests of the Standard Model, to understand the structure of nucleons and other hadrons, and to determine the properties of hadronic matter under extreme conditions.

Massively parallel computers are ideally suited for lattice gauge calculations, having been successfully exploited for many years. Furthermore, it has proven far more cost effective for lattice gauge theorists to build their own computers than to make use of general purpose supercomputers. In doing so they can take optimal advantage of simplifying features of lattice QCD, such as regular grids and uniform, predictable communications. General purpose machines must perform well for a wide variety of problems, including those requiring irregular or adaptive grids, non-uniform communication patterns, and massive input/output capabilities. Thus, commercial supercomputers require considerably more expensive communication systems than are needed for lattice QCD. For this reason, our plans include the development of both hardware and software infrastructure, pursuing both customized clusters of commodity components and the development of a fully custom lattice computer. The concept of a topical computing facility as set out in the Office of Science's computing plan is particularly well suited for lattice gauge theory.

The objective of this proposal is to support the software development necessary to productively exploit the envisioned multi-teraflops computing facilities. Two architectures will be targeted: large

commodity clusters and the QCDOC, the next generation of the highly successful Columbia/ Riken/ BNL special purpose computers. (Development of the QCDOC has been funded separately).

A flexible, user-friendly software environment is critical to the development of efficient new algorithms and computational methods. We will collaborate closely with our colleagues in computer science, computer engineering and applied mathematics on the software and algorithms research and the hardware prototyping needed for terascale QCD systems. The interplay between hardware, software, and algorithms research is central to the proposed work.

The planned software infrastructure will enable members of the national lattice gauge community to achieve high performance on future terascale systems, while focusing their efforts on frontier questions in physics. This will require a variety of components for high application performance, including standardized libraries with highly tuned code for common computationally intensive tasks and optimized communications primitives, implemented for each target architecture. At a higher level, the project will deliver standard programming models, portable applications, and user-friendly interfaces. The software infrastructure will also incorporate components necessary to schedule and monitor jobs, standards for data formats, and other tools to support wide access and management of large computational and data resources.

Scalability of lattice applications to large clusters of commodity computers, including symmetric multiprocessors (SMPs), will be a key development area. Systems containing hundreds of processors are required to provide adequate test-beds for the software and for the viability of multi-teraflops clusters for QCD. To this end we propose expanding the clusters currently under development at Fermilab and by the MIT/Jefferson Lab consortium. These systems will be made available to lattice researchers for their scientific work and to further test and "harden" the software infrastructure. In this activity, the clusters will be available to the full U.S. lattice gauge theory community, as will all of the hardware and software infrastructure created under this proposal.

This work will provide the basis for a future proposal requesting distributed topical computing resources for the study of QCD. To meet the full scientific challenges of QCD, and to allow U.S. lattice gauge theorists to compete effectively with researchers in Europe and Japan, we plan to propose the coordinated development of three open national facilities of 10 Tflops scale in the period 2003-2007. The first terascale facility will be a QCDOC based machine to be located at Brookhaven National Laboratory, at a cost of less than $1 per sustained Mflops. The architectures used for the 10 Tflops facilities at Fermilab and Jefferson Lab will be chosen on the basis of the experience gained with clusters and the QCDOC, on their cost effectiveness, and on their ability to support the scientific program set out here. This plan will optimally position us to exploit future technologies for fundamental physics calculations, and this multi-pronged approach will be crucial to maintaining flexibility in the future.

This project is expected to interact positively with a variety of other scientific endeavors. In computational science, massive parallelism is the only route to substantial improvements in capabilities, but exposure to very large scale parallelism remains limited. Lattice gauge theory is particularly well suited to such machines, and will provide much needed experience. We expect the infrastructure we develop to be useful to other fields, such as condensed matter physics, in which simulations are also carried out on regular grids. We will make a portion of our resources available to members of those fields for testing.

2 Background and Significance: Physics Goals

2.1 Electroweak Matrix Elements

A central focus of experiments at all U.S. high energy physics facilities is precision testing of the Standard Model. The ultimate goal of this work is to find deviations from which we can learn about the properties of matter at even shorter distances. Many of these tests require, in addition to precise experimental measurements, accurate evaluation of the effects of the strong interactions on processes induced by the electroweak interactions. Such an evaluation requires lattice QCD, the only known method which can systematically reduce all sources of error. Such computations are one of the major physics focuses of this proposal. Technically, one needs to compute matrix elements of certain operators between hadronic states, which are known as "electroweak matrix elements", or "weak matrix elements" for short.

Terascale computers will lead to an enormous advance in weak matrix element calculations. We will be able to calculate certain key matrix elements with an accuracy that is comparable to the experimental accuracy in the corresponding measurements. At present, the precision of the tests of the Standard Model which rely on these parameters is limited by the uncertainties in the lattice calculations. The situation is summarized in Table 1 and Figure 1 [1]. These uncertainties will be significantly reduced as a result of the calculations we propose. We estimate that a computer sustaining 0.5 TFlops for a year would lead to a reduction in the uncertainties in all three quantities by about a factor of two, and that a machine sustaining 10 TFlops for a year would halve the uncertainties again. For B_K, this estimate is explained in Appendix A.3.1. The tests of the Standard Model will then be much more stringent, and in some cases limited by experimental errors.

Matrix element	Present lattice error	Quark transition	Experimental quantity	Experimental Error
$f_B^2 B_B$	35%	$t \to d$	ΔM_B	4%
$\xi^2 = f_{B_s}^2 B_{B_s} / f_B^2 B_B$	10%	$t \to d / t \to s$	$\Delta M_B / \Delta M_{B_s}$	not measured
B_K	20%	CP violation in $t \to d$	ϵ, V_{cb}	10%

Table 1: Comparison of present errors in lattice results for three key weak matrix elements and the experimental errors in the quantities determined by the weak matrix elements. The matrix elements are characterized by the quark transition induced by the weak interactions.

There will be corresponding improvements in the calculations of many other matrix elements. For these, the calculations are not as advanced as for the key parameters mentioned above. For such matrix elements, a Terascale facility will make substantial improvements, but further improvements in computer power, or in algorithms, will be required to reach the ultimately desired accuracy.

A fairly complete list of electroweak matrix elements that have been studied using lattice QCD is given in Appendix A.2.1, along with a description of the status of their calculation, and a summary of their importance to the determination of parameters of the Standard Model. We also give a partial list of related matrix elements whose calculation can serve as tests of the lattice methods since they are accurately known experimentally. Particle physicists will discover other interesting matrix elements, and one of the advantages of the lattice methodology is that it often allows the calculation of new matrix elements with relatively little overhead once the gauge configurations are generated. Thus, our proposed Terascale facility will provide a flexible database which can be reused repeatedly as new ideas appear.

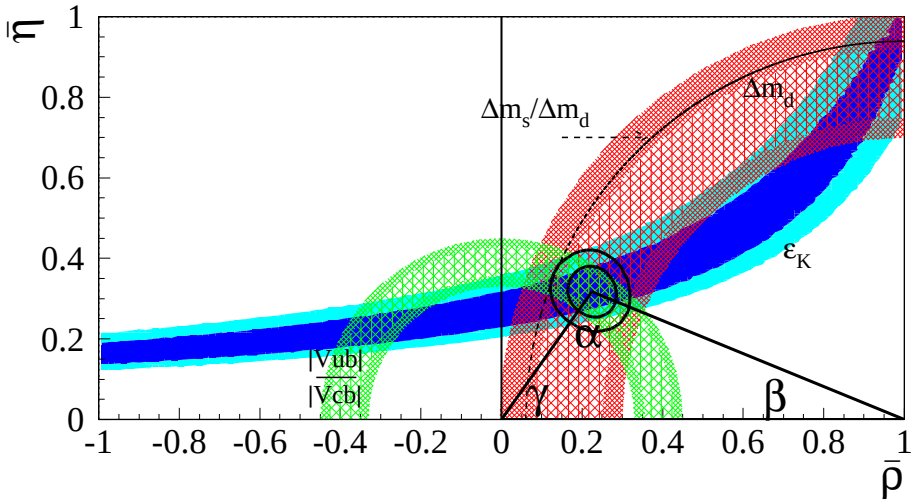

Figure 1: Present constraints on $\bar{\rho}$ and $\bar{\eta}$, the least well known quark mixing parameters in the Standard Model. Uncertainty bands (each at 68% and 95% probability) are shown from the measurements of ε_K (which requires knowledge of the theoretical parameter B_K), Δm_d (requiring $f_B^2 B_B$), and $|V_{ub}| \,/\, |V_{cb}|$. The experimental limit on $\Delta m_s/\Delta m_d$ leads (given the theoretical value of ξ) to the constraint that the values must lie to the right of the dotted curve. The bands are consistent, and lead to the allowed region shown for $\bar{\rho}$ and $\bar{\eta}$. Uncertainties in lattice calculations dominate the width of the bands for ε_K and Δm_d. The calculations proposed here will reduce these widths by about a factor of four, leading to a much more accurate determination of $\bar{\rho}$ and $\bar{\eta}$, or an inconsistency between measurements.

For the key quantities the basic methodology of estimating errors is well known and thoroughly tested. The issue is how to extrapolate from the values of certain parameters used in simulations to their physical values. The extrapolations are in the light quark masses, the volume, the lattice spacing, and, in some calculations, the heavy quark masses. We stress that we are considering unquenched calculations—the internal quark loops will be included for up, down and strange quarks. Based on our experience, we have a good idea of how close to the physical parameters we need to get in order that the extrapolated result has the desired error. We also know, to reasonable approximation, how the required computer time scales, with present algorithms, as one varies the parameters. Putting this together allows us to estimate how the errors will scale with computer power. Examples are given in Appendix A.3.1 We stress that these are conservative estimates in that they assume no improvement over present algorithms.

Finally, we note that our calculations will also provide precise values for other fundamental parameters of the Standard Model—the quark masses and the strong coupling constant, α_S. Precise results for these are needed to differentiate between competing models of flavor physics and electroweak symmetry breaking, and lattice simulations provide the only method for doing such calculations. Indeed, the lattice results for the c and b quark masses are already very accurate (e.g. the error is 2% for m_b). The light quark masses—m_u, m_d and m_s—are more difficult to calculate, requiring extensive simulations with light dynamical quarks. Present errors, estimated to be 25% [2] will be substantially reduced by a Terascale facility.

2.2 The Quark Gluon Plasma

At low temperatures and densities, quarks and gluons are confined in elementary particles, such as neutrons and protons. At very high temperatures and densities one expects a phase transition or crossover from this ordinary strongly interacting matter to a plasma of quarks and gluons. Such a plasma is believed to have been a dominant state of matter in the early development of the universe, and a possible central component of neutron stars today. In a major new DOE experimental initiative, the Relativistic Heavy Ion Collider (RHIC) began operation at the Brookhaven National Laboratory last year. A primary physics goal is the discovery and characterization of the quark-gluon plasma. In order to confirm such an observation, it is important to determine the nature of the transition, the properties of the plasma, including strange quark content, and the equation of state. Lattice gauge theory has proven to be the only source of *a priori* predictions about this form of matter in the vicinity of the phase transition, with members of our collaboration playing a major role in the worldwide effort. We propose to use improved actions, which we have developed, to embark upon a detailed study of these issues.

While lattice gauge theory has already provided a wealth of qualitative information (including an indication of the order and an estimate of the temperature of the phase transition as well as a characterization of the equation of state) simulations have, for the most part, only included the up and down quarks, whereas it is expected that the somewhat heavier strange quark also plays a crucial role. From universality considerations we expect that with two flavors of quarks there is no phase transition at all for physical values of the up and down quark masses – merely a crossover. However, a strange quark could induce a first order transition, or move a second order critical point closer to physical quark masses. These effects may be of considerable importance to the phenomenology of the phase transition.

We propose to take advantage of our recent successes with improved actions and the enormous power of the proposed facilities to carry out a definitive study of the quark-gluon plasma with a realistic quark ensemble and vastly reduced discretization artifacts. The results are expected to give valuable assistance to the RHIC experimental program.

Our scientific objectives are the following:

1. Mapping of the phase diagram in temperature and quark mass for up, down, and strange quarks, including determining the order of the phase transition and temperature of the crossover.
2. Determining the equation of state of the plasma, including strange quark content.
3. Predicting real-time excitations of the plasma.
4. Understanding the role of instantons in the phase transition.
5. Measuring the strength of the axial $U(1)$ anomaly.

The first three objectives have obvious relevance to the analysis of experimental results. The last two are needed for formulating phenomenological models that extrapolate to situations inaccessible to lattice gauge theory. Based on our extensive experience in the study of QCD thermodynamics with simpler actions, this will be a multi-year project, which will require long-term use of the proposed 10 Tflop/s Facility.

As with all lattice simulations, systematic errors primarily arise from three sources: (1) the lattice discretization, (2) finite volume, and (3) unphysically large quark masses. For example, previous thermodynamic studies of the effects of strange staggered quarks [3] and strange Wilson quarks [4] have been done with most of the pion masses far above their real-world values. While

these studies are the best that could be done with the algorithms and computing power of the time, it is hard to do a realistic study of the effects of a (strange) K meson with its correct physical mass of 500 MeV, when most of the (non-strange) pions have comparable masses, well above their physical mass of 140 MeV.

To achieve our scientific objectives in a realistic simulation requires a combination of significantly improved algorithms and significantly enhanced computing power. Recently, we have developed substantial algorithmic improvements in the two major formulations for quarks (fermions) on the lattice: (1) Without requiring an excessively small lattice spacing, our improved staggered fermion actions [5] give vast improvement in the zero-temperature hadron spectrum, a feature essential on the low temperature side of the phase transition, and they give significant improvement in the quark and gluon dispersion relations, essential at high temperature [6]. Also especially important, on the low temperature side, dispersion relations and flavor symmetry in the critically important pion sector are vastly improved. (2) The recently implemented domain wall fermion approach dramatically improves the Wilson fermion scheme. This new method preserves the flavor symmetry of the Wilson formulation, improves its relatively good scaling properties by removing all $\mathcal{O}(a)$ errors both on- and off-shell, realizes the complete chiral symmetry group and supports the physics of the axial anomaly and the Atiyah-Singer theorem [7]. This physical chiral symmetry (both anomalous and non-anomalous) of the domain wall method may be particularly important for an accurate study of the QCD phase transition.

As for objective 2, the strange quark content of the plasma is of vital interest, since excessive strange quark production could be a signal of plasma formation. The strange quark content is determined by measurement of the contribution of strange quarks to the energy density of the plasma, so emerges from a study of the equation of state.

Unfortunately, the determination of the equation of state is very costly, since it involves computing the difference between the energy density at a nonzero and zero temperature. The loss of significance in the subtraction makes the computational effort scale in the lattice spacing as a^{-4} relative to the effort for mapping out the phase diagram. Thus algorithmic improvements that suppress lattice artifacts and permit simulation on a coarser lattice are of critical importance. Recent studies on smaller lattices have shown dramatic changes resulting from algorithmic improvement in the Wilson fermion scheme [8, 9]. We expect similar improvements in the domain wall approach.

Real time excitations of the plasma (objective 3) are very poorly understood, but have a direct relevance, for example, for the measurement of the dilepton emission spectrum. Maximum entropy methods, now being explored by a number of groups, offer the promise of extracting real-time spectral information from finite-temperature lattice Green's functions [10].

2.3 Structure and Interactions of Hadrons

A third goal of our collaboration is to achieve a quantitative, predictive understanding of the structure and interactions of hadrons. The internal structure of the nucleon is a defining problem for hadron physics just as the hydrogen atom is for atomic physics. Indeed, the DOE Strategic Plan specifically highlights the goal of developing a quantitative understanding of how quarks and gluons provide the binding and spin of the nucleon based on QCD. Major experimental efforts at Bates, Jefferson Lab, SLAC, Fermilab, the HERMES experiment at DESY, and the EMC, SMC, and NMC experiments at CERN provide rich and precise measurements of the quark and gluon structure of the nucleon, and proposed experiments such as the RHIC spin program promise to reveal even greater detail. With recent advances in lattice field theory, it is now possible to calculate this nucleon structure directly from QCD, so that multi-Terascale lattice calculations are now an essential tool to obtain the full physics potential of major accelerators and detectors.

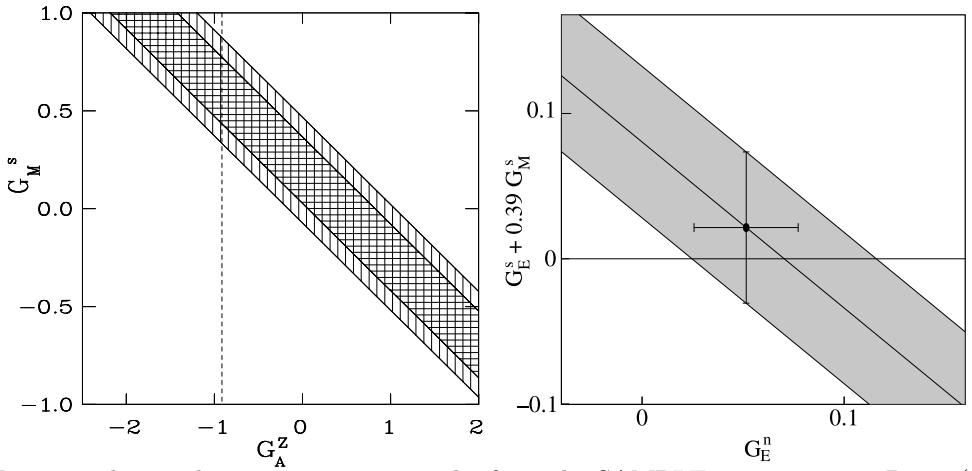

Figure 2: Parity-violating electron scattering results from the SAMPLE experiment at Bates (*left*) and the HAPPEX experiment at JLab (*right*). The error bands show the first-generation experimental constraints on the strange quark magnetic form factor and the Z^0 form factor and on a combination of the strange electric and magnetic form factors and the neutron form factor. Subsequent experiments will soon provide precise measurements of the strange quark contributions to the nucleon magnetic moment and charge radius.

A wealth of experimental observables can be calculated on the lattice. Electromagnetic form factors measured in elastic electron scattering characterize the distribution of charge and magnetization arising from all the quarks in the nucleon. Parity violating electron scattering experiments exploit the fact that the neutral weak current couples to a different linear combination of up, down, and strange quarks than the electromagnetic current to measure the strange quark contributions to electric and magnetic form factors. As shown in Fig. 2, recent experiments at Bates and JLab [11] have now measured these form factors, and future experiments are expected to determine them with high precision [12]. Deep inelastic scattering of electrons, muons, and neutrinos measures structure functions characterizing the light cone quark density, quark spin density, and gluon density as a function of momentum fraction, and the moments of these distributions can be calculated on the lattice. A particularly important example is the lowest moment of the spin density, which measures the fraction of the nucleon spin carried by the spin of quarks. Indeed, the only way to fully resolve the so-called "spin crisis" which arose when experiments showed that only about 20% of the spin of the nucleon originates from quark spins is to calculate in lattice QCD how the total spin is divided between quark and gluon spin and orbital angular momentum.

Prototype calculations with limited computational resources have already established the methodology to calculate the nucleon form factor [13], the axial charge [14], the contributions of strange quarks in the nucleon [15], and moments of the quark density, spin, and transversity distributions in both quenched [16] and full [17] QCD. These calculations also clearly show the need for computational resources by highlighting the major role the long-distance pion cloud plays in hadron structure. As explained in the Appendix A.2.3, quantitative calculations require calculation on a spatial lattice dimension in excess of 4.3 fm and extrapolation from quark masses corresponding to a pion mass below 230 MeV [18], which necessitate multi-Terascale computational resources.

Spectroscopy is the classic tool for discovering the relevant degrees of freedom of a physical system. One of the fascinating features of QCD is that it offers the possibility of a richer range of hadronic states than has yet been observed experimentally. Lattice calculations will study the number and structure of hadronic excited states, as well their transition form factors. The presence or absence of hadrons with exotic quantum numbers, the nature of glueballs, and the overlap between model trial functions and exact hadron states will provide insight into the role of flux

tubes, dibaryons, and the inner workings of QCD. Again, exploratory calculations of the lowest negative parity N^* state [19], a comprehensive calculation of the glueball spectrum [20], calculation of hybrid mesons [21], and a study of the existence of the H particle [22] show that the requisite methodology is ready. Precision multi-Terascale calculations will provide crucial insight into current and future hadron spectroscopy.

Currently, there is no fundamental understanding of the very foundation of nuclear physics, the nucleon-nucleon interaction. Significant insight into the role of gluon exchange, quark exchange, meson exchange, and the origin of short range repulsion will be obtained by lattice calculations of the adiabatic potential between heavy-light systems, [23] that is, mesons or baryons containing a single heavy quark in addition to other light quarks or antiquarks.

In addition to calculating observables to compare with experiment, lattice calculations are invaluable to obtain insight into fundamental aspects of QCD. Current lattice techniques can study the role of instantons [24] and their associated zero modes [25] in chiral symmetry breaking, the role of center vortices [26] and magnetic monopoles [27] in confinement, and calculation of the parameters entering chiral perturbation theory. The lattice also allows theorists to answer interesting theoretical questions inaccessible to experiment, such as how the properties of QCD change with the number of colors, quark flavors, or quark masses. New techniques under development may also enable study of the phases of dense hadronic matter and the transitions between them.

A more detailed description of proposed research in hadron structure and interactions by members of this collaboration may be found on the web[1].

3 Preliminary Studies: Computing Goals

3.1 Computational Resource Requirements

The computational resources required to reach the scientific goals set out in the last section can be estimated from the extensive experience members of our group have had in performing calculations in all of the areas we propose to study. The scale of the hardware we propose to build is driven by these estimates.

The required computational resources vary significantly among the proposed studies as can be seen in Appendix A.3. However, a few features common to all lattice gauge theory calculations determine their approximate scope: the physical dimensions of the lattices must be large to avoid finite size effects; calculations must be performed over a range of small lattice spacings, using improved actions, in order to make accurate extrapolations to the continuum limit; and dynamical up, down and strange sea quarks must be included in the simulations. The first two requirements force one to work on very large lattices, but it is the requirement to include dynamical or sea quarks that plays the most important role in driving up the computational resources. Indeed, in most lattice calculations, the inversion of the Dirac operator takes over 90% of the computer time. The number of floating point operations that go into the inversion grows roughly as the inverse of the quark mass. As a consequence, it has been customary to ignore the effects of strange sea quarks, and either ignore those of up and down sea quarks or to work with unphysically large values of their mass. In this project we propose to work at the physical value of the strange quark mass, and at small enough values of the up and down quark masses to make reliable extrapolations to their physical values. This will require working at π to ρ mass ratios in the range 0.3 to 0.4, with lattice spacings between 0.05 and 0.10 fm, and lattice spatial dimensions of order 3.0 to 5.0 fm. In Appendices A.3.1, A.3.2, and A.3.3 we give detailed estimates of the resources required for several

[1] `ftp://www-ctp.mit.edu/pub/negele/LatProp/`

typical projects. The time requirements range from a few months to over a year with a throughput of 10 Tflop/s.

In comparison to their needs, the resources available to U.S. lattice gauge theorists are meager. The QCDSP computer at Columbia University[2], which was constructed specifically for the study of lattice gauge theory, sustains 120 Gflop/s on production code, and its sister machine at the Riken in Brookhaven sustains 180 Gflop/s. These two machines provide important tools for physicists at Columbia and BNL. However, the remaining 90% of the U.S. lattice community, which includes some of the most highly respected individuals and strongest groups in the world, has access to resources that provide an aggregate throughput of approximately 30 Gflop/s.

These resources are small in comparison to those available to our overseas competitors. The two major Japanese groups, CP-PACS and JLQCD, are currently running computers with a combined throughput of approximately 660 Gflop/s. As of the beginning of 2001, French, German and Italian physicists had APEmille computers with a total throughput of approximately 720 Gflop/s. These represent the latest in the line of special purpose machines built by the APE group in Italy. The APE group has announced plans to have a total installed performance of 10 to 30 Tflop/s for its next generation machine within the next two to five years,[28] and our colleagues in Japan are making their own plans for terascale computers. Thus, unless the U.S. lattice community aggressively constructs the infrastructure needed for multi-terascale computing, it will not be able to carry out the research program outlined above, nor will it continue to lead the field.

3.2 Custom Machines

Lattice QCD requires enormous computing capacity, and the national lattice computing plan is to deploy three machines of scale 10 Teraflop/s sustained in FY2003-2005 as a distributed Lattice QCD Topical Computing Facility.

Because of its regular structure, lattice QCD is well suited to take advantage of machines consisting of thousands of processors. Its communication patterns are highly regular and predictable, allowing overlapping communications and computation. This overlap, and the regular grid structure, makes it an application which does not need the extremely high performance and flexible backplanes of modern supercomputers – nor their high cost. This is one of many reasons for building a computing resource targeted to this application domain.

The lattice QCD community both in the U.S. and internationally has for a number of years pursued an approach of using highly parallel custom machines. The U.S. effort will continue this approach with a two pronged attack: (1) the development of a fully custom parallel machine, and (2) the optimization of commodity clusters. The national software infrastructure presented in this proposal will hide most of the machine differences from the applications and users.

Columbia University is collaborating with IBM in the design of a "computer on a chip" which incorporates processor, memory, Ethernet, and fast communications which directly supports the requirements of lattice QCD.[3] This "QCD on a Chip" (QCDOC) design will allow construction of machines of 20 thousand processors, each sustaining ≈ 0.5 Gflops, double precision. Especially important is the 4-D mesh network which provides 1 Gbyte/sec off-processor bandwidth and communication latencies of hundreds of nanoseconds. Large scale deployment of QCDOC machines is expected in FY03 at a cost of $1 per sustained megaflops[29]. Details of the design can be found in Appendix A.4.

Simultaneously, we are also aggressively exploiting dramatic market driven developments in commodity components to develop cost optimized clusters for Terascale lattice QCD. MIT and

[2] `http://phys.columbia.edu/~cqft/qcdsp.htm`

[3] `http://phys.columbia.edu/~cqft/qcdoc.htm`

Jefferson Lab are collaborating with Compaq to utilize Alpha technology that is expected to lead to commodity "computer on a chip" technology running QCD at over 3 Gflops per chip in the 2003-05 time frame, and Fermilab is exploring analogous Intel technology. Details can be found in Appendix A.5.

We believe this dual exploration of custom and commodity Terascale technology provides the most robust foundation for a QCD topical center that will establish and maintain into the future world leadership in lattice QCD physics.

4 Research Design and Methods: Software Infrastructure

The proposed coherent national effort in QCD physics has two essential software components: (i) achieving very high efficiency on multi-Terascale hardware and (ii) developing a unified QCD programming environment that allows the entire U.S. lattice community to devote their time to physics. Fortunately we have a unique community which has already accomplished a substantial number of the essential tasks in this program, including running QCD applications on multi-thousand node computers at the Terascale. In addition, we have obtained the participation of a number of computer scientists and engineers who will bring new software technology to our enterprise. We propose to complete this program to enable seamless utilization of three 10 Terascale machines and to establish a uniform methodology to amortize the individual software efforts across this national effort. A list of project milestones and a timeline can be found in Appendix A.1.

In the first year, a major focus is the development of software kernels and tools adequate to scale from the 500 Gflops testbed clusters at Fermi and Jefferson Laboratories (to be proposed in FY02), to the 10 Terascale systems (FY03-05). In addition, in the first year we will establish a subset of MPI[4] (Message Passing Interface) required for QCD so that all MILC (MIMD Lattice Computation) code as well as other existing C/MPI based QCD code runs on all platforms, especially the 10 Tflops QCDOC installation. For high performance kernels to take advantage of optimized communications hardware, we will also define a QCD communications layer which will avoid the latencies in MPI. Optimization of MILC performance will provide a valuable test of our general software methodology.

There are several mature and well developed software platforms created over the years by individual collaborations within the U.S. lattice community. They represent a wide range of research aims, identified by the major national laboratories and universities involved in this proposal, based on very similar core routines targeted for high performance on a variety of platforms. Each code base typically represents about 100K lines involving about 10 person-years of work.

The MILC code is an extensive C/MPI based code, available on the Web[5] that runs on all the major platforms except the QCDSP. It supports 25+ major physics applications with about 140+ variants including weak matrix elements, high temperature QCD and spectrum codes. The Fermilab code started from the extensive Canopy system for the ACPMAPS computer[6] and is now evolving into a sophisticated C++ code system with strong support for weak matrix elements. The JLab/MIT/BU package is a largely C based system including use of the UKQCD (United Kingdom QCD Collaboration) code [7] as well as the SZIN[8] (color in Hungarian) methodology supporting hadron spectrum, structure functions, eigenvalues and thermodynamic calculations, and which generates efficient QCD code targeted to many different architectures including the QCDSP. There is special support for structure functions, nucleon excited states, topology and

[4] `http://www-unit.mcs.anl.gov/mpi`
[5] `http://www.physics.utah.edu/~detar`
[6] `http://www-isd.fnal.gov/acpmaps/acpmaps.html`
[7] `http://www.ph.ed.ac.uk/ukqcd/`
[8] `http://www.jlab.org/~edwards/szin`

eigenvector calculations. The Columbia Physics System (CPS) in use at Columbia and BNL is an extensive C++ package with support for full QCD simulations with many different fermion types and includes the measurements needed for hadron spectrum, heavy quark, weak matrix element, Dirac eigenvector and thermodynamic calculations. The low level kernels are particularly optimized for the QCDSP, while the global structure readily supports the architecture of QCDSP and the future QCDOC.

These existing code repositories provide the raw material and experimental data for the software infrastructure project. Indeed the reason we are confident that we can achieve simultaneously exceptional performance and portability is based on the intrinsic advantages that QCD presents to the software designer. The dominant computational time (better than 90% in most instances) is spent in a few key routines, notably the Dirac inverter, and the pattern of communications between processors is both regular and deterministic so that latency hiding strategies are effective. This explains in part past success: On a $O(10^4)$ processor QCDSP machine QCD code sustains 30% peak performance and optimized code on an single Alpha SMP node runs at 45% of peak; the MILC application code based on C and MPI standards has been ported to practically every existing high performance platform and the SZIN QCD code runs efficiently across dissimilar architectures with very different memory hierarchies by remapping data layout. This proposal will continue this effort at the national level on a distributed nation wide multi-Terascale QCD facility.

4.1 Algorithms for lattice QCD

The lattice actions used for QCD, the algorithms employed to sample the system's phase space and the linear algebra technique for solving the Dirac equation have undergone a quarter century of evolution with no end in sight. Obviously no code is truly efficient, independent of its nominal floating point "performance", if it does not implement the best available algorithmic approach. For example, recently there has been a major breakthrough in the formulation of the lattice Dirac field that preserves the full chiral symmetry of the continuum quarks. Like most past developments, we are fortunate that the algorithmic consequences of this revolution has not radically affected the architectural constraints for an efficient QCD machine. However, it serves to emphasize that our software infrastructure must allow rapid implementation and testing of important new ideas. Indeed, more radical methods are certainly needed to extend QCD simulations to problems not presently feasible with existing methods. An example is the exploration of ways to "cure" the notorious "fermion sign problem" encountered in QCD at non-zero chemical potential (or RHIC physics at non-zero baryon number density). The search for better algorithms and improved actions is a worldwide enterprise comprised of theoretical physicists and applied mathematicians.

We now turn to more detailed discussions of the software layers and tasks required to develop an improved QCD infrastructure. A timeline is given in Appendix A.1.

4.2 Platform Software Infrastructure

Although small clusters have shown great promise for extremely cost effective QCD calculations, they have not yet been scaled to tens of Teraflops. Hence, we propose to advance cluster technology to a new regime by developing the software infrastructure required to perform frontier QCD calculations on highly optimized, cost effective, multi-Terascale commodity clusters. This section describes the development of the low level software components, below the QCD-API interface, that are essential to optimization of the QCD kernels, optimized communications, and the execution environment. To be credible, these components must be tested in physics production on actual hardware that is of a scale roughly within an order of magnitude of the three future 10 Terascale

facilities. Hence it is essential to build the large prototype clusters described in the hardware section. A corresponding testing platform for the QCDOC software development will be provided by the QCDOC simulator now partially complete and the initial QCDOC machines being built in 2002 at Columbia. While added QCDOC hardware is not requested in this proposal, a major focus of the software effort proposed here is the close integration of the QCDOC operating and application code into the planned national framework.

4.2.1 Task 1: Optimized Lattice QCD Kernels

The critical section of QCD application code is the application of Dirac operator, which is inverted iteratively to form propagators. There are classes of Dirac actions, operators and inversion methods which form a suite of closely related algorithms and are built out of more primitive kernels. Because the Dirac operator also places the highest demands on the communications network, the precise impact of the primitive kernels depends strongly on the characteristics of the communications hardware as well as the floating point unit. Thus the QCD-API will include a specification for these kernels, allowing the high level code to achieve optimized performance for each multi-Terascale hardware platform. An efficient and hierarchical implementation of the primitive kernels, starting with fundamental linear algebra routines for complex 3 by 3 matrices (in SU(3)), will be available to allow flexibility in developing and testing new approaches. Each platform will provide a QCD library of specific linear algebra routines, some of which are likely to be coded in assembly language.

For the Alpha cluster on a single SMP node (the ES40 with four 667MHz processors), Dirac inverters have been tested that achieve better than 45% of peak when the data set fits in the L2 cache. A major goal in FY01 is to scale up this software infrastructure to a full network implementation. This requires detailed benchmarks and optimization on the SMP clusters as a function of granularity and data layout. For typical problem sizes that do not fit in cache, it has been found from studies with the MILC code that substantial improvements in performance can be achieved by carefully scheduling cache line prefetching. In the SZIN code alternative data layout schemes also appear to improve cache scheduling. Another major goal in FY01 in cluster development will be to develop the combination of processor hardware and software diagnostics necessary for understanding efficient memory management. The development of scalable performance analysis and visualization tools by the Illinois computer science group will play an important role in this work. In addition, they will implement automatic cache optimization for the lattice Dirac operator along the lines of Dongarra's ATLAS system.[9] Lessons learned will be incorporated into the API kernels.

4.2.2 Task 2: Optimized Network Communications

Lattice QCD on multiple processors exhibits a deterministic and highly repetitious communications pattern. (This is the major reason that it can execute well without the need for a conventional expensive supercomputer interconnect.) This permits optimizations in calling syntax and underlying support not possible in a more general system such as MPI, a standard message passing library.

This project will identify an efficient API for nearest neighbor communications, global sums, and any other frequently used operations, and provide optimized implementations of this API on each supported system. (An implementation atop MPI will also be supported for backward compatibility.) This communications layer will in particular support the optimized lattice kernels described above. Optimization efforts will utilize outside expertise of computer scientists as much as possible. For example, computer scientists at the University of Karlsruhe have undertaken extensive software development to obtain significantly higher efficiency for scientific applications on

[9] `http://netlib.uow.edu.au/atlas/`

Myrinet than possible with conventional Myricom software, and have collaborated with computational physicists in obtaining high efficiency for QCD on a 128 node Alpha cluster in Wuppertal. Walter Tichy from Karlsruhe will participate in a cluster workshop organized by Fermilab, JLab, and MIT March 26-28, and we propose to support future travel and collaboration expenses for him and other collaborators as part of this project.

4.2.3 Task 3: Execution Environment

In order to provide a productive environment for research on the scale envisaged in this proposal, the software infrastructure will need to encompass more than just raw application performance. It must provide an execution environment which has the run-time capabilities needed for QCD applications and which provides a batch system to simplify work flow.

• **Run-time Environment**: Beyond standardized lattice kernel libraries and communications protocols, applications will need to read and write files, log error and status messages, and perhaps provide a socket based interface for distributed computing capability. These capabilities must be provided in a standard way to support portability of applications across various systems.

Since both the clusters and QCDOC have ethernet connections to every node, the hardware support for these features is available and very similar. This project will therefore first identify the various run-time capabilities needed by lattice QCD applications and users, then prioritize their standardization and deployment. For clusters, these features will be built atop a conventional multitasking operating system. For QCDOC they will be implemented in the real-time operating system for the machine as it is developed. The end result will be a standardized set of calls for all users, for both the QCDOC and the clusters.

• **Distributed Batch Environment**: It is important that the final system run at high utilization, and that it implement the resource allocation policies set by a program committee. Hence the system must include a flexible batch system making it easy for the user to submit, monitor, and control batch jobs. Both command line and web based interfaces will be deployed to meet this objective, and efforts will be made to leverage work currently in progress in other DOE and NSF funded activities, such as the Portable Batch System (PBS) and the Maui job scheduler.

It will be of additional value to seamlessly integrate multiple computing resources located not just at the three primary national facilities, but also at a number of university facilities, utilizing resources funded through state or other non-DOE funding agencies. This model of distributed resources is being adopted by virtually all large scale computing efforts, such as the Large Hadron Collider computing program, and Fermilab's Collider Run II simulation and analysis effort. As for the basic batch capabilities, these "meta-facility" capabilities may incorporate software from other projects.

• **Optimized Parallel File I/O**: In a typical project, gauge configuration and propagator files are archived and reused. On the proposed 10 Terascale machines, data sets are expected to be 100 GBytes or larger. Thus it is important to pay attention to efficient parallel I/O in cluster and QCDOC design. To implement parallel I/O at this scale on clusters requires a careful integration of I/O nodes into the communications fabric. A coordinated hardware and software effort is required. In FY01 we will be investigating parallel file I/O schemes, file formats, and issues of network integration on prototype cluster machines. We will take advantage of the data handling experience of FNAL experimental groups in this effort. In addition we have enlisted the help of computer scientists Dan Reed, Celso Mendes and Rob Pennington of the University of Illinois, to work on a

range issues for large networks of high performance cluster, such as parallel file systems, performance monitoring and optimization, etc.

For QCDSP, a parallel file system has been implemented, utilizing the high internode bandwidth available from the four-dimensional machine network. A similar strategy will be employed for QCDOC, so little research is required on this issue. However, the format for parallel files must be coordinated between the clusters and QCDOC to allow for easy interchange.

4.3 High Level Software Infrastructure

To operate and optimize the resources of three 10 Teraflops facilities as a single national QCD program calls for a new level of organization and planning. An appropriate reference model might be the particle/nuclear experimental community with its nationwide network of DOE laboratories. Each lab has its own focus but exchange of manpower and data is essential to the whole enterprise and to avoid duplication. For lattice QCD the national co-ordination must be flexible enough to foster the creativity that is a hallmark of the present diverse community. Here we consider the software development at or above the QCD-API interface.

4.3.1 Task 4: Porting and Optimization of Application Code

The first need is to achieve a common methodology so that existing application codes can be run on all facilities with good performance. Due to the unique structure of QCD, providing a library of optimized Dirac inverters (Wilson, clover, staggered, domain wall, overlap, etc.) and a communication layer with the semantics of a subset of MPI will go a long way toward this goal of portability at reasonable performance. This is a practical solution to the short-range goal of "porting" MILC code, for example. Moreover it is also of great practical advantage for debugging purposes to be able to run all application codes on a variety of platforms including short runs on Terascale hardware. However to successfully compete for the expensive resources for production runs on the Terascale facilities, one will have to demonstrate performance at the highest level of all comparable codes. The optimization steps required will be the object of on-going software experimentation and development.

At this higher level, data layout can be critical. Consequently, to accommodate the variety of programming styles in the existing application code suites, it will be important to develop remapping kernels between alternative data layouts and to study the impact of remapping on overall efficiency relative to an alternative approach of implementing a repertoire of data schemes for the fundamental QCD kernels. Depending on the placement in the code hierarchy, a combination of both methods may be required. It may also be necessary to consider the SZIN methodology (easily implemented in an object oriented language such as C++), which automatically generates source code with distinct data layouts appropriate for each target architecture.

4.3.2 Task 5: QCD API and Code Library

The porting and optimization task described above is a useful (and necessary) step in the early stages. But clearly one should move to a software model that avoids duplicating this task on each extension to the application codes. For this we plan to define a QCD-API to act as minimal standard QCD programming model. It will require comparing existing standards and agreeing on naming conventions and syntax for the fundamental components of the QCD kernels library. This includes linear algebra routines, basic communication primitives, Dirac operators, a class of inverters, etc.

The interface will guarantee that application code written to the QCD-API will run **as is** on all the Terascale platforms. This will encourage conversion to this API and will result in a

"toolkit" of compatible application modules. This application "toolkit" will then be archived and made available, just as is the current practice with MILC code. It is the responsibility of all facilities to contribute libraries consistent with QCD-API. We will not impose the QCD-API on the application physicists by fiat. We believe that the advantage and efficiency of the API will in practice ensure its adoption in future lattice QCD applications. One tremendous advantage over time of this more uniform methodology is that future facilities can quickly conform, thus freeing applications programmers from much of the distraction of writing long QCD codes from scratch.

We also propose the development of a common set of software tools for the analysis and fitting of data. This is a time consuming task that will greatly benefit from a more uniform and flexible methodology. For example we need methods to perform multiple-parameter fits to multiple, partially correlated, data sets with difficult-to-invert correlation matrices.

4.3.3 Task 6: Data Management and Documentation

The Terascale computational resources described in this proposal represent a significant investment by the DOE, and one major product of this investment will be a large set of lattice configurations and other valuable data. These files can be re-used to study a number of physics phenomena, and will therefore need to be cataloged and made accessible to the research community.

It is the intent of the proposers to leverage the developments of the DOE Particle Physics Data Grid project (PPDG) [10] to provide location independent, high speed access to replicated data files, with the ability to discover files based upon a description or by a global file name (a distributed data grid). The three national laboratories in the current proposal (BNL, FNAL, and JLab) are all participants in PPDG, and Jefferson Lab and MIT are already working to deploy data grid technologies for lattice QCD.

Crucial to the efficient use of multiple lattice gauge facilities is the ability to run a job at one of several sites according to the resources currently available. We will therefore integrate the distributed batch system described above with the data grid software to provided automated staging of data or executables to and from remote compute or data grid nodes. As mentioned previously, attempts will be made to leverage existing efforts in this area, such as those of our colleagues at NCSA. The clusters at Jefferson Laboratory and at MIT will be the test-bed for a lattice QCD "metacenter".

A further responsibility of the software project is to provide support to users running at the Terascale installations. The creation of documentation to make the operating features and software environment easily accessible to a remote user (both high level software elements as well as cluster and QCDOC specific aspects) will be an important part of this support. This project will maintain the library of software versions, and act as a first point-of-contact for new users of the multi-Terascale facilities.

5 Research Design and Methods: Hardware Infrastructure

As previously stated, our plan is to deploy three machines of scale 10 Teraflop/s sustained in FY2003-2005 as a distributed Lattice QCD Topical Computing Facility. These machines are planned to be sited at BNL, FNAL, and Jefferson Lab so as to best leverage the extensive computing infrastructure at those sites (networking, robotic tertiary storage, and operations expertise). Two strategies are being pursued for these machines: the QCDOC custom machine, and commodity clusters. The software infrastructure presented in this proposal will hide most of the machine

[10] See: http://www.ppdg.net/

differences from the applications and users. This section gives a description of the machine design for each approach, as additional context for the proposed software development. Details can be found in Appendices A.4 and A.5.

5.1 QCDOC

A key element of the five-year plan supported by this proposal is the creation of a very promising 10 TFlops (sustained) computational facility based upon a custom machine, which is being designed as a joint development project by Columbia University and IBM. This project will produce a new 1.0 Gflops "computer on a chip" which, together with an adjacent slot for an industry-standard memory card with up to 0.5 Gbyte of memory, constitutes the complete computing node in a massively parallel (20K-node) computer architecture. These nodes are interconnected as a six-dimensional torus with four dimensions of this grid intended to support the usual four-dimensional space-time mesh used in lattice QCD calculations. The remaining two dimensions are available to permit flexible software partitioning of the machine. In addition to the CPU, this "computer on a chip" contains the twelve nearest-neighbor, high-bandwidth, low latency links needed to realize this mesh communication pattern. A similar computer architecture was developed for and is used in (with considerably less powerful node elements) the 0.6 TFlops QCDSP supercomputer developed at Columbia University that is currently operating at BNL in the RIKEN/BNL Research Center (RBRC). The RBRC machine is optimized for lattice gauge physics applications and consequently provides an excellent architecture for the new QCD on-a-chip (QCDOC) machine.

The IBM-manufactured computer on a chip is currently under development at IBM's T. J. Watson Research Center and Columbia University under a cooperative agreement between IBM and Columbia. The R&D necessary to incorporate this new chip in the massively parallel QCDOC computer architecture is proceeding in parallel at IBM and at Columbia with DOE-HEP funding support. All this development work is expected to result in a first working computer module in 2002 and is consistent with production of a custom machine in the 2002-3 time frame.

This time table assumes that the current plan for the development and production of the Columbia-IBM design is followed and that production quantities of the resulting chips and component parts are purchased in joint procurements with other partners planning to build QCDOC computers using this technology. At the present time, significant QCDOC machines are planned to be built and installed at Columbia University in New York, the RBRC at BNL, the University of Edinburgh as well as the BNL component of the distributed topical computing center we propose. The Edinburgh machine is already funded at the 10 Tflops (peak) level.

A very important aspect of this program is that the code currently operating on the QCDSP achieves a 30% efficiency as it runs today. This represents a high efficiency for any practical computing application. It is anticipated that lattice gauge codes will run on the QCDOC at 50% (or better) efficiency, leading to a remarkably effective lattice gauge physics output capability. A more detailed description of the QCDOC machine together with a construction cost breakdown is provided in Appendix A.4 to this proposal.

The QCDSP has already shown that a system with ten thousand or more compute nodes can deliver useful computing cycles better than 97% of the time. An essential ingredient in achieving this reliability is the hardware monitoring and debugging capabilities of the operating system software for QCDSP. This software identifies specific hardware culprits for a large percentage of machine faults, even during production physics running, allowing faulty hardware to be quickly replaced. Using similar techniques, high reliability from QCDOC should also be achieved. Also with the recent implementation of a parallel disk system on QCDSP, concrete experience with the I/O needs of QCD on multi-thousand node systems will be available to guide such systems on QCDOC.

Given the exciting potential for physics discovery that flows from large-scale lattice QCD computing resources, there is strong motivation to employ highly cost-effective machines for this research. Just as the QCDSP computers represented a significant advance in supercomputer cost / performance (winning the 1998 Gordon Bell prize for $10/Mflops), the new QCDOC architecture promises a further factor of 10 advance in science per dollar. By creating the software infrastructure outlined in this proposal, we will be able to exploit this impressive cost/performance to provide critical support for the research program of the larger U.S. lattice QCD community.

5.2 Optimized Commodity Clusters

A second essential element of the five-year plan envisioned in this proposal is exploiting the rapid advances and cost advantages of commodity technology for multi-Terascale lattice QCD computation. To maintain leadership throughout the 2003-2005 time frame, it is essential to follow the initial investment in the QCDOC machine with additional 10 Teraflops machines in the topical center that follow the rapidly evolving technology curve. This section provides a brief overview of the relevant cluster design parameters and the extremely promising market trends relevant for a 10 Teraflop facility.

5.2.1 Performance Factors

Lattice QCD is a highly regular grid application: the compute and communications requirements for present applications are well known, and new algorithms under consideration are quite similar. Each processor holds a 4 dimensional sub-lattice, and for each step in a typical calculation, it performs a known number of floating point operations on each site while exchanging data for its 3-d hyper-surfaces with 8 nearest neighbors. Computation and communication are overlapped, so key performance factors are (a) sustained single processor floating point performance, (b) memory bandwidth, and (c) messaging bandwidth. At certain points in the calculation, a global sum is performed (sensitive to (d) message latency), and periodically the lattice is written to disk (disk bandwidth (e) being of only minor concern). As the number of processors is increased, the ratio of surface (messaging) to volume (compute) goes up. This trend, however, is offset by a tendency to increase the number of lattice points, in order to optimize the total physics output of the machine rather than the time to do a specific calculation; i.e. harder problems (larger lattices) are tackled on bigger machines.

The key point is that the communications requirements (bandwidth, latency, messaging pattern) do not require the high speed backplanes of current day supercomputers, and are in fact well matched to high performance clusters. (The QCDOC is in fact a custom cluster). One need only to specify the range of expected lattice sizes and the total number of processors to be able to derive the network bandwidth (to send hyper-surfaces in time for when they are needed) and latency (so that global sums constitute only a small fraction of the clock time). Note that the latency tends to constrain the size of the largest system that would achieve good price performance. This limitation currently becomes important at the several million dollar cluster level.

On a more modest sized cluster today consisting of 128 dual processor Compaq alpha processors, (667 MHz, dual issue, 50% of peak sustained, 170 Gflops sustained aggregate) one would typically use a $16^3 \times 32$ lattice. For this system (which achieves $8/Mflop today), link bandwidth of 80 MB/s and latency of 8 microseconds preserves about 85% of the single processor performance. These requirements are easily achievable by a moderate cost switched network such as Myricom's Myrinet. Because the messaging requirements of lattice QCD are somewhat different from other applications, custom software for some cluster interconnects (including Myrinet) can extend the

maximum sized cluster supportable by reducing message latency, a factor which is driving one of the software development tasks in this proposal.

5.2.2 Clusters of SMPs

Building clusters from multi-processor nodes has two advantages: (1) the link cost per processor is reduced (amortized over multiple processors), and (2) the latency requirement for fixed processor count is loosened – there are fewer boxes participating in a global sum (the in-box global sum step is comparatively negligible as it proceeds at full memory speed). Constraining the use of large processor count SMP's is (1) today these have very expensive backplanes (up to 90% of total system cost) and (2) bandwidth out of the box is constrained by the I/O bus. Optimal cluster solutions today exist at the 2 and 4 processors per box point, since these boxes are popular for building web servers and other e-commerce applications. Within two or three years, 8-processor boxes are expected to be equally or more cost effective, allowing even larger clusters for future generations of this architecture.

5.2.3 Industry Trends

A number of trends are relevant to the optimistic outlook for commodity clusters:

- The most significant trend is Moore's law: performance/price doubling every 18 months for commodity boxes. The silicon roadmap for both Intel IA-64 and Compaq Alpha will deliver processors achieving several gigaflops peak within a couple of years. From this trend it is expected that the 10 Tflop sustained cluster can be built with only 4096 processors. By 2004, both Intel and Compaq will be delivering chips which issue 4 floating point operations per clock cycle, pushing performance even higher.

- Commodity processors will soon integrate memory controllers and even cache-coherent backplanes on-chip, reducing system integration costs even as performance rises. Case in point: the Alpha EV7 chip (21364) integrates 2 memory controllers, 4 high speed inter-processor links, and an I/O bus controller, making it a commodity "computer on a chip", and allowing large processor count systems (up to 128) to be constructed with a minimum of glue logic.

- I/O bus speeds will soon reach the Gbytes/s level (PCI-X, Infiniband standards), necessary to support large SMP clusters at the multi-terascale level. In addition, SMP systems will support multiple I/O buses per box.

- Commodity low latency switched networks (Myrinet, giganet, etc.) are available today at high port density, leading to a near-linear cost per port for large clusters. This cost per integrated node for cluster interconnects is falling, even as all performance specifications are improving.

5.2.4 Conceptual Cluster Facilities

Within 2-3 years, a 4096 processor system (10 teraflops sustained) could be constructed either as 1024 quad processors (low-end commodity) or as 32 128-way SMP's (supportable by the next generation alpha chip). The first solution requires clusters no larger than are already running today, and would be only mildly demanding on the network latency – less than 2.5 microseconds assuming a $32^3 \times 64$ lattice. The second solution, if it proves cost effective, would not be as demanding on latency, and may require multiple links per box to reach the required bandwidth of 2 GB/s (e.g. 4 x 500 MB/s). The particular choice of processor, cluster interconnect, and number of processors

per box cannot be known today, as it will depend upon market forces and who brings the next generation chips into commodity systems sooner, but it is clear that multi-terascale commodity systems will be viable in the very near future.

By performing the cluster software development outlined in this proposal, we will be optimally positioned to make the extraordinary developments in commodity components and their cost effectiveness available to the full U. S. lattice QCD community.

6 Research Design and Methods: Phase I Projects

6.1 First Year Software

For the first year, we have chosen tasks that balance the immediate need to sustain the physics program and provide access to the hardware as it becomes available with the general goal of an improved QCD software infrastructure to optimize the utilization of Terascale platforms. This software infrastructure project is a three year project with a reduced budget expected in the third year as the work load shifts to running the Terascale hardware platforms.

Using the partitioning of tasks in Section 4, we detail the FTEs required for the first year's work in Table 2. The FNAL and JLab activities involve FTEs to support optimization and development work on clusters, an activity closely related to the hardware platforms at these locations. Support for such work is not needed for the BNL QCDOC effort. The FTEs for the execution environment will implement the common standard on the various hardware platforms, as well as contributing to its specification. The applications porting and optimization will evolve existing codes to the common QCD-API and insure optimized Dirac inverters on a given hardware platform conforming to the QCD-API. The requested personnel will also contribute to the QCD-API and have responsibility for maintaining the code libraries. Coordinated work on data management will take place across all sites, concluding in common storage formats for lattice configurations and propagators as well as the most commonly produced applications data. Finally support is requested for documentation, which is particularly important for the QCDOC project, since it involves purpose-built hardware. Strong university contributions to all these tasks is critical both in terms of manpower and to continue the interaction with the majority of QCD physicists who make up the user community for the multi-Terascale facilities.

The overview of manpower requested for the software infrastructure project is summarized in the table below. The University participation at BU, Arizona, and Utah supports one postdoctoral staff each at 0.75 FTE time, at Univ. of Illinois Computer Science department one faculty at 0.2 FTE plus 2 students. The Software Manager is 0.3 FTE or 0.4 release of 9 month academic salary (see budget for details).

Brookhaven will provide from its own resources an additional 0.25 FTE of expert software / management consultation from the top members of the RHIC and U.S. ATLAS computing organizations at BNL: Dr. Bruce Gibbard, Director of the RCF and Manager of the U.S. ATLAS Tier-1 Computing Center; Dr. Torre Wenaus, Director of the U.S. ATLAS Software Group; Dr. Richard Baker, Deputy Director for U.S. ATLAS Facilties and Dr. Razvan Popescu, Grid Computing Manager. This group of experts will provide a valuable, interdisciplinary resource to the software effort both at Brookhaven and elsewhere within the collaboration.

We now give more details on the specific tasks to be accomplished by the manpower in Table 2.

- *Task 1: Lattice QCD Kernels*:

  In the first year, we will optimize essential QCD kernels for Intel and Compaq Alpha based cluster architectures at FNAL and JLab/MIT respectively, conforming to the QCD-API. To

Task	FNAL	JLab	BNL	Univ.	Total
Platform Software Infrastructure					
1. Lattice QCD Kernels	0.75	0.50	0.50	1.25	3.00
2. Optimized Network Communications	0.75	0.75		1.00	2.50
3. Execution Environment	0.50	0.75	0.50	0.70	2.45
High Level Software Infrastructure					
4. Applications Porting & Optimization	0.50	0.25	0.50	1.00	2.25
5. QCD-API & Code Library	0.25	0.25	0.25		0.75
6. Data Management and Documentation	0.25	0.25	0.50	0.50	1.50
Software Project Manager				0.30	0.30
Totals	3.00	2.75	2.25	4.75	12.50

Table 2: FTEs required for first year projects

achieve performance comparable to that of a single node on a multi-node machine, we will optimize the data layout to minimize the effect of internode communication and SMP memory sharing on cache. The investigation of SMP performance through local threads and the study of message passing protocols will be performed jointly at JLab, FNAL and participating Universities. Scalable performance analysis and visualization tools will be developed and applied to the QCD kernels by computer scientists at the University of Illinois. They will implement automatic cache optimization for the lattice Dirac operator. Working closely with the group at Columbia, the Brookhaven effort will implement the required changes/enhancements to the QCDOC kernels to meet QCD-API developed in Task 5 below.

- *Task 2: Optimize Network Communication*:

  In spite of MPI's emergence as the *de facto* standard for cluster communication, the fastest internode communication typically requires access to lower-level communication routines that are network-card specific. We will perform a detailed study of the card-specific communication routines for the SCI-standard Dolphin card (Dolphin Interconnect) at FNAL and for the Myrinet card (Myricom Inc.) at JLab and MIT in collaboration with industry. Based on the results, we will implement the communications specified by the QCD-API making use of performance gains from the low-level communication calls. This will require one staff scientist at each of FNAL, JLab and MIT/BU. Funds are also requested to support university computer science faculty at the University of Illinois working on cluster optimization and to support travel by computer scientists in Karlsruhe working on network optimization to enable us to capitalize on these recent developments.

- *Task 3: Execution Environment*:

  The clusters will start with a standard implementation of Linux and evolve towards a lean, robust Linux kernel suitable for Terascale computing. The chosen publicly available batch queuing system (PBS) will be tested on the clusters. The basic QCDOC operating system will be developed by existing personnel, incorporating the hardware debugging features of the QCDSP operating system which have been shown capable of monitoring and diagnosing hardware failures for systems of thousands of compute nodes. Personnel at FNAL, JLab and BNL

will closely co-ordinate and collaborate with affiliated universities, and will work to achieve a uniform operating software platform conforming to a common execution environment as discussed in Section 4.3.2.

- *Task 4: Application Porting and Optimization*:

  To achieve the greatest physics impact from the outset of the project, it is essential to capitalize on existing lattice code efforts. We will ensure that these codes can make use of the developments in QCD kernels and optimized network communications at each facility, and provide routines for data remapping should these be necessary for optimal performance. This will require support of personnel at FNAL, JLab, computer scientists at the University of Illinois, and physicists at associated universities. The software group at BNL will also make the CPS system conform to the QCD-API, and thus usable on any hardware platform supporting the communications conventions.

- *Task 5: QCD-API and Code Library*:

  The experience gained through the first year in developing optimized Lattice QCD kernels and an optimized network communication library will enable us to define a QCD Application Program Interface (QCD-API). The specification of the QCD-API will be defined concurrently with the development of the *Optimized QCD Kernels* and the *Optimized Network Communication*. The definition of the specification, together with some optimal QCD primitives satisfying the specification will require on order of a single man year of effort shared between FNAL, JLab and BNL.

- *Task 6: Data Management and Documentation*

  The data generated by the facilities will be an extremely valuable resource which must be accessible and usable by the widest possible base of collaborating users. To accomplish this, we will in the first year of the project specify a common, machine-independent data format for the gauge configuration data, agree on a common file-naming convention and directory structure for the data archive, and provide a suite of data conversion routines so that existing data can be used in the current project, and new data accessed by legacy codes. This will require about a person-year of effort at FNAL, JLab and BNL. An effort co-ordinated with the associated universities will insure all researchers ease of access to the lattice QCD data archives and QCD code libraries.

6.2 Phase I Hardware

To harness the most cost effective technology in the rapidly evolving marketplace for multi-terascale QCD calculations, it is essential to aggressively build and test two large clusters that exploit the two most promising but dissimilar architectures. An Alpha based cluster will explore the scaling and performance advantages the highest performance processors currently available. A complementary Intel based cluster will explore the advantages that the mass market can bring in reducing the cost per chip. Investing the requisite software effort to optimize the kernels and communications for both of these two dissimilar clusters will ensure that we are optimally positioned to identify and exploit the sweet spot of the commercial marketplace for the machines to follow on the QCDOC in the lattice QCD topical center.

Key research and development tasks for these large clusters are to achieve high single processor efficiency, and maintain that efficiency within an SMP machine, and within a large cluster of SMP's. In order to test the scalability of clusters arranged as 3- and 4-dimensional meshes, it is desirable

to have systems of at least 4^4, or 256 boxes, so that each box has non-degenerate nearest neighbors in all dimensions. For each processor choice, it is necessary to identify the most cost effective interconnect option.

Two prototype platforms will be targeted in the first two years of this proposal: an Intel based cluster at Fermilab exploiting their expertise with Intel clusters and an Alpha based cluster at JLab built on the joint experience at MIT and JLab in their ongoing collaboration in Alpha clusters. Half of each prototype cluster will be procured in FY01, and the remaining half in FY02. As a cost savings measure, the more expensive alpha cluster will be reduced to 2x4x4x4 boxes for these studies. In FY03, prototype hardware based on subsequent generation chips will be evaluated.

Fermilab will test Myricom's Myrinet interconnect on its main cluster and will also investigate the use of Scalable Coherent Interconnect (SCI) based cards from Dolphin. Jefferson Lab will test and optimize the use of Myrinet and custom Myrinet firmware. Alternative link technologies will be evaluated as appropriate.

6.2.1 Jefferson Lab Cluster

For the Jefferson Lab system we propose to procure 32 dual processor boxes, to be augmented by a laboratory contribution to 64 dual processors. These will be connected both by 100 Mbit switched ethernet and by Myricom's Myrinet 2000 switched network. As a point of reference (with final decisions to be made when funds are available), Alpha Processor, Inc. manufactures a 19" rack mount, 1 U, 833 MHz dual processor alpha 21264 system (the CS20), allowing the entire prototype to be mounted in two standard racks. Myricom offers a chassis + card switch that is expandable to 128 ports in a single box, and delivers latency and bandwidth appropriate for systems of the size expected in the first two years (and even much larger). Proposed procurement details:

1. 32 2-way SMP nodes, minimum characteristics: 833-MHz Alpha 21264, 4 MB cache per processor, 512 MB memory per box
2. Myrinet-2000 switch with eight 8 port cards (expandable to sixteen cards)
3. 64 Myrinet-2000 PCI interconnect cards (64 bit, 66 MHz, 120 MB/s) and cables

It is expected that this system (augmented by JLab contributions) will achieve nearly 50% of peak, or roughly 100 gigaflops, at a cost of less than \$7/Mflop. (Better price performance, less than \$6/Mflop, could be achieved using slower processors, but it is important for this phase in the project to stress the ability to achieve a high fraction of peak performance on an extremely high performance processor.)

In the second year of this project (FY02), we expect to be able to procure systems with processors running at 1.25 GHz or better, aiming at less than \$4/Mflop. We propose to procure 64 of these dual processor nodes, with Jefferson Lab either procuring an additional 64 nodes, or upgrading the original nodes to the faster clock speed, yielding a cluster of 128 nodes. This 256 processor system will serve both as a prototype for scalability towards a significant production activity, and will serve a portion of the science needs of the lattice community.

Infrastructure and Matching Contributions: Jefferson Lab will contribute 32 cluster nodes and associated cluster and network cards to bring the first year system to 64 nodes. It will purchase four additional nodes, two to provide a development platform for optimization efforts at MIT, and two to provide a separate small development system at Jefferson Lab, as well as to serve as spares to ensure high availability of a 64 box system. Jefferson Lab will also provide access to its existing alpha clusters (12 duals and 16 singles) for applications software development and testing (as well as production running). In addition, the lab will supply at least one terabyte (TB) of disk space,

and tertiary storage of 10+ terabytes (TB) within its existing tape silo, which has a total capacity of up to 300 TB. The laboratory's OC3 ESNET connection will provide wide area access to this prototype lattice computing facility, and to its repository of lattice configurations and other data. System management, power and cooling, repair, network security and other functions inherent in a shared computing facility will be handled by the Jefferson Lab computing staff.

Partnering with Jefferson Lab, MIT will provide access to its existing cluster of 12 quad-processor systems, a one TB disk cache, and will work with Jefferson Lab to develop and demonstrate multi-site operation.

6.2.2 Fermilab Cluster

For the Fermilab cluster, we propose to procure 200 dual processor boxes, to be attached to our existing 80 computer cluster via expansion of our Myrinet network. The current performance-to-price leaders on our codes are single processor Pentium 4 systems. We anticipate that by the time of the procurement, Intel and other manufacturers will be supplying dual Pentium 4 Xeon systems. Procurement details:

1. 200 2-way SMP nodes, minimum characteristics: 2 GHz Intel Pentium 4 Xeon, 256 MB memory per box.
2. 2 Myrinet-2000 switches, 128 ports per switch
3. 6 Myrinet-2000 switch port cards, to expand existing Myrinet-2000 switch.
4. 400 port (minimum) ethernet switch (100 Mbps ports, at least 2 gigabit uplinks).

Based on our preliminary work on existing codes on Pentium 4 systems, we expect each of these computers to sustain 300 Mflops on problems spanning many nodes, for a total of 90 Gflops, at a cost of approximately $13/Mflop (including networking) for unoptimized code. Through the use of prefetching and vector floating point assembly language instructions, we see a path to improve this to $7/Mflop or better.

Infrastructure and Matching Contributions In the event this proposal is successful, Fermilab will contribute computer-ready space and infrastructure for the cluster. It will contribute an FTE toward procurement, installation, administration, and maintenance of the cluster. It will contribute a budget of $300,000 for hardware and operations for the project. Ten terabytes or more of long term tape storage will be available for users.

7 Subcontract or Consortium Arrangements

All funded participation in this initiative is via direct funding to each institution and there are no subcontracts or funded consortium arrangements.

7.1 Management Plan

Overall responsibility for this effort will be vested in a Committee of Principal Investigators (PIs): N. Christ (Columbia U.), M. Creutz (BNL), P. Mackenzie (Fermilab), J. Negele (MIT), C. Rebbi (Boston U.), S. Sharpe (U. Washington), R. Sugar (UCSB) and W. Watson, III (JLab). They will ensure that the software infrastructure and hardware prototyping is completed in a timely fashion, establish procedures for the equitable use of the infrastructure by the national lattice gauge theory community, arrange for oversight of progress in meeting the scientific goals set out in this proposal, and decide on the distribution of funds among the various projects to be undertaken under this proposal. Decisions will be made by majority vote of the Committee of Principal Investigators,

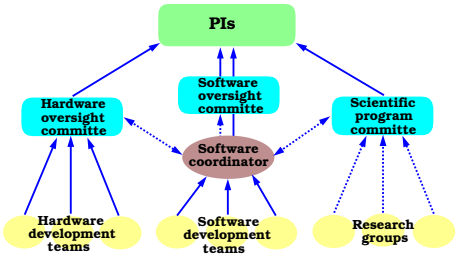

Figure 3: Management structure. The solid lines indicate supervision, the dotted lines indicate overseeing and coordination responsibilities.

with the spokesperson's vote deciding the outcome in case of a tie. Robert Sugar will serve as spokesperson and principal contact for the Department of Energy. The principal investigators at the institutions receiving funds for software development and hardware prototyping will have first level responsibility for this work. The spokesperson will submit to the DOE semiannual reports on the progress of the project.

The PIs have appointed R. Brower to serve as Software Coordinator, and will appoint three committees to assist them in the management of the project:

- **The Software Coordinator** will supervise the work of all software development teams, providing for direction and coherence of effort. The Software Coordinator will develop a detailed set of tasks and milestones, will keep track of who is working on the various components of the project, and will prepare quarterly progress reports, updating the milestones as necessary. The Software Coordinator will report to the PIs. He will keep the Software Oversight Committee informed of the progress in software development, and will consult with the Hardware Oversight Committee to ensure that the design of prototype clusters, the (separately funded) design of the QCDOC supercomputer and the software development are properly matched.

- **The Software Oversight Committee** will oversee the software development, reviewing progress on a quarterly basis. It will report its findings to the PIs, the Software Coordinator and the software developers. This committee will consist of seven senior scientist selected by the PIs, and will include at least two software experts from outside the collaboration.

- **The Hardware Oversight Committee** will oversee the hardware development effort, reviewing progress on a quarterly basis. It will report its findings to the PIs and to the hardware developers. This committee will consist of five senior scientists selected by the PIs, and will include at least two hardware experts from outside the collaboration.

- **The Scientific Program Committee** will monitor the scientific progress of the project, and provide leadership in setting new directions. It will organize an annual meeting of all lattice gauge theorists working on or planning to participate in the project to review progress and plan future directions. If, as is hoped, the efforts of the participants in this project will eventually lead to the development of Terascale machines for lattice gauge theory calculations, then the Scientific Program Committee will be entrusted with the allocation of time for major projects on these machines. The Scientific Program Committee will consist of seven senior members of the collaboration. They will be chosen by the PIs, and will serve two year, renewable terms.

The software coordinator will receive support for partial academic release. The members of the

software and hardware oversight committees who are not collaborators in this project will receive appropriate honoraria for their services.

The overall effort will be supported by the established management structure at the three DOE laboratories (BNL, FNAL, JLab), which will be major participants in the project. These three laboratories will also provide for long term maintenance and support of the developed software after the end of this project.

All of the hardware and software infrastructure developed under this proposal will be made available to the U.S. lattice gauge theory community. In addition, the large lattices and quark propagators generated in major research projects will be stored in a common format, and made available to the entire community for studies in order to maximize the physics obtained from these computationally expensive data sets.

7.2 Budget

The overall budget for the three year project we propose is summarized in Table 3. Detailed budgets for the institutions which are to receive funds, along with the description of the work these budgets will support, can be found in the separate budget sheets. The bulk of the software development work will be done in the first two years. We anticipate that work will taper off during the third year, and thereafter the software will be maintained by the participating laboratories. We have included a 5% cost of living increase in the software budget for year two. The hardware budget would enable the construction of substantial prototype clusters at Fermilab and JLab over the first two years of the project. These clusters are needed to test the software, and to put us in a position to build terascale clusters for the distributed topical computing center. The hardware funds requested for the third year of the project are to enable us to purchase and test limited quantities of new processors and interconnects which we anticipate will become available at that time.

Institution	FY01	FY02	FY03
Software Budgets			
BNL	338	355	248
Boston U.	152	159	97
FNAL	386	405	284
JLab	386	405	284
MIT	148	156	109
U. Arizona	62	65	42
U. Illinois	133	140	98
UC Santa Barbara	46	48	34
U. Utah	62	65	42
Total Software	1713	1798	1238
Hardware Budgets			
FNAL	375	500	100
JLab	375	500	100
Total Hardware	750	1000	200
Total	2463	2898	1438

Table 3: Software and Hardware Budgets in $1,000

References

[1] M. Ciuchini *et al.*, hep-ph/0012308.

[2] V. Lubicz, Nucl. Phys. Proc. Suppl. **94**, 116 (2001) [hep-lat/0012003].

[3] F.R Brown *et al*, Phys. Rev. Lett. **65**, 2491 (1990).

[4] Y. Iwasaki *et al.*, Phys. Rev. **D54**, 7010 (1996).

[5] G.P. Lepage, Nucl. Phys. (Proc. Suppl.) **60A**, 267 (1998); Phys. Rev. D **59**, 074501 (1999); K. Orginos, R.L. Sugar and D. Toussaint, Phys. Rev. D **60**, 054503 (1999); Nucl. Phys. (Proc. Suppl.) **83-84**, 878, 2000.

[6] The MILC collaboration: C. Bernard *et al.*, hep-lat/9912018, to be published in Phys. Rev. D.

[7] P. Vranas, Lattice 2000 Review, hep-lat/0011066.

[8] The MILC Collaboration, C. Bernard, *et al.*, Phys. Rev. **D56**, 5584 (1997); R.G. Edwards and U.M. Heller, Phys. Lett. **B462**, 132 (1999).

[9] Y. Iwasaki *et al.*, Phys. Rev. Lett. **78**, 179 (1997); Nucl. Phys. B (Proc. Suppl.), **63**, 397 (1998).

[10] M. Jarrell and J.E. Gubernatis, Phys. Repts. **269**, 135 (1996); M. Asakawa, T. Hatsuda, and Y. Nakahara, hep-lat/0011040.

[11] B. A. Mueller *et. al.* Phys. Rev. Lett. **78** (1997) 3824.

[12] TJNAF experiments E91-010, E91-017, and E91-004.

[13] S. Capitani *et al.*, Nucl. Phys. Proc. Suppl. **73**, 294 (1999)

[14] S. Gusken *et al.* Phys. Rev. D **59**, 114502 (1999).

[15] S. J. Dong, K. F. Liu, and A. G. Williams, Phys. Rev. **D58** 074504 (1998).

[16] M. Gökeler *et al.*, Phys Rev. D53, 2317 (1996).

[17] D. Dolgov *et al.*, hep-lat/0011010, to appear in Nucl. Phys. B (Proc. Suppl.) (2000).

[18] W. Detmold *et al.* preprint (2001), E. J. Hackett-Jones *et al.*, Phys. Lett. **B 489**, 143, (2000).

[19] D. G. Richards, hep-lat/0011025, to appear in Nucl. Phys. B, (Proc. Suppl.) (2000). S. Sasaki, hep-ph/0004252. S. Sasaki *et al.*, hep-lat/0102010.

[20] C.J. Morningstar and M. Peardon, Phys. Rev. D 60, 34509 (1999).

[21] P. Lacock *et al.*, Phys. Lett. **B401**, 308 (1997); P. Lacock and K. Schilling, Nucl. Phys. B, (Proc.Suppl.), **73**, 261 (1999). C. Bernard *et al.*, Phys. Rev. D **56**, 7038 (1997); Nucl. Phys. B (Proc. Suppl) **53**, 228 (1997); **60A**, 61, (1998); **63**, 206, (1998); **73**, 264, (1999). K.J. Juge, J. Kuti, and C.J. Morningstar, Phys. Rev. Lett. 82, 4400 (1999); P. Chen, X. Liao, T. Manke, hep-lat/0010069, to appear in Nucl. Phys. B, (Proc. Suppl.) (2000).

[22] J. W. Negele, A. Pochinsky, and B. Scarlet, Nucl. Phys. B (Proc. Suppl.) 73, 225 (1999).

[23] D. G. Richards, D. K. Sinclair, and D. Sivers, Phys. Rev. **D42** 3191 (1990). A. Mihály *et al.*, Phys. Rev. **D55** (1997) 3077. R. Koniuk, C. Stewart, Phys. Rev. **D57** (1998) 5581. C. Michael, and P. Pennanen, Phys. Rev. **D60** 054012 (1999).

[24] J.W. Negele, in Unfolding the Matter of Nuclei, Course CXXXVIII, Varenna, Italy, July sd(1998).

[25] T. L. Ivanenko and J. W. Negele, Nucl. Phys. B (Proc. Suppl.) **63** 504 (1998).

[26] L. Del Debbio, M. Faber, J. Greensite, S. Olejnik, Phys. Rev. D55, 2298 (1997).

[27] A.S. Kronfeld, M.L. Laursen, G. Schierholz, U.J. Wiese, Phys. Lett. B198, 516 (1987).

[28] The APE Collaboration, hep-lat/0102011.

[29] D. Chen, *et al.*, Nucl. Phys. B (Proc. Suppl.) **94** 825 (2001) [hep-lat/0011004].

[30] S. Hashimoto, A. X. El-Khadra, A. S. Kronfeld, P. B. Mackenzie, S. M. Ryan and J. N. Simone, Phys. Rev. D **61**, 014502 (2000) [hep-ph/9906376].

[31] L. Lellouch, hep-ph/9912353.

[32] G. Martinelli, Nucl. Phys. Proc. Suppl. **73**, 58 (1999) [hep-lat/9810013].

[33] J. Flynn and C. J. Lin, hep-ph/0012154..

[34] C. Bernard, Nucl. Phys. Proc. Suppl. **94**, 159 (2001) [hep-lat/0011064].

[35] V. Lubicz, hep-ph/0010171.

[36] L. Lellouch, Nucl. Phys. Proc. Suppl. **94**, 142 (2001) [hep-lat/0011088].

[37] L. Lellouch and M. Luscher, hep-lat/0003023.

[38] M. Pospelov and A. Ritz, hep-ph/0010037.

[39] K. Kanaya *et al.* [CP-PACS Collaboration], Nucl. Phys. Proc. Suppl. **73**, 189 (1999) [hep-lat/9809146].

[40] H. L. Lai *et al.*, hep-ph/9706502.

[41] C. Bernard *et al.* [MILC Collaboration], Phys. Rev. D **55**, 6861 (1997) [hep-lat/9612025].

[42] S. Aoki *et al.* [JLQCD Collaboration], Phys. Rev. Lett. **80**, 5271 (1998) [hep-lat/9710073].

[43] T. Blum, *et al*, Phys. Rev. D (to appear) (2001), hep-lat/0007038.

[44] P. Chen, *et al*, Phys. Rev. D (to appear) (2001), hep-lat/0006010.

[45] T. Lippert, S. Gusken and K. Schilling, Nucl. Phys. B (Proc. Suppl.) **83** 182 (2000).

A Appendices

A.1 Project Milestones

The tasks laid out in Section 4 and the first year plans described in Section 6.1 are part of a longterm strategy to achieve full utilization of the proposed terascale machines through the development of a comprehensive software infrastructure. In the first year of this project, the first three tasks outlined in Section 6.1 consume the largest fraction of the effort. In the second and third years, this emphasis moves to the fifth task, particularly the development of widely used data analysis tools. In Table A.1 are the milestones that expected to be achieved. In Figure **??** is another representation of the timeline of the expected milestones.

Task	Date	Description
Lattice QCD Kernels	Dec 2001	Kernels optimized for latency, yielding high efficieny on small clusters of "thin" SMPs.
	Jun 2002	Standardized kernels optimized on QCDOC.
	Dec 2003	Kernels optimized for "fat" SMPs, newer interconnects.
Network Communication	Dec 2001	1st version low level messaging API implemented on clusters, QCD-API communications standard adopted.
	Jun 2002	2nd version low level messaging API running on clusters, QCDOC.
	Dec 2002	high efficiency demonstrated on large clusters.
	Dec 2003	high efficiency on clusters of "fat" SMPs, newer interconnects.
Execution Environment	Jun 2002	Requirements defined for parallel I/O, batch execution, and other aspects of execution environment.
	Jun 2002	Uniform batch system running on clusters, QCDOC prototype.
	Jun 2003	1st version parallel I/O operational on clusters, QCDOC.
	Jun 2004	Final execution environment finished, including fast parallel I/O, distributed batch with load balancing and web portal.
Application Porting and Optimization	Jun 2002	MILC ported to all platforms (not fully optimized).
	Jun 2003	MILC, others, optimized on all platforms.
QCD-API and Code Library	Dec 2001	QCD-APS defined.
	Jun 2002	Requirements for data analysis defined.
	Jun 2003	First versions of most data analysis tools available.
	Jun 2004	Data analysis tools finished.
Data Management	Dec 2001	Data file formats defined.
	Jun 2002	Converters for legacy data available.
	Jun 2002	1st version of data grid operational.
	Jun 2003	Production data grid operational.

A.2 Prospects for Physics Calculations

A.2.1 Prospects for Electroweak Matrix Elements Calculations

In this appendix we indicate of the breadth of weak matrix element calculations that can be studied using a Terascale facility, and describe their status. We first discuss the three key electroweak matrix elements described in the main proposal.

- $f_B\sqrt{B_B}$. (Appears squared in ΔM_B.) Present estimate is 230 ± 40 MeV [34]. Very well understood in quenched approximation—calculated using relativistic and non-relativistic formalisms for b-quark. Unquenching (with unphysically heavy sea quarks) raises results by $\sim 10\%$. Main uncertainties: chiral extrapolation and discretization errors in unquenched theory; effect of dynamical strange quark; perturbative errors in the normalization of operators. The latter errors, which are theoretical, are presently about 5%, might become dominant for calculations on terascale machines. It may be possible to reduce this error by further theoretical work.

- $\xi = f_{B_s}\sqrt{B_{B_s}}/f_B\sqrt{B_B}$ (Appears squared in $\Delta M_{B_s}/\Delta M_B$.) Present estimate is 1.14 ± 6 [35]. Has smaller errors than $f_B\sqrt{B_B}$ since it is a ratio. Well understood in quenched approximation, and unquenching (with unphysically heavy sea quarks) has no significant effect. Main uncertainties same as for $f_B\sqrt{B_B}$, except that perturbative errors are smaller.

- $\hat{B}_K$ (renormalization group invariant form of B_K). Present estimate is 0.86 ± 0.15 [36]. Very well understood in quenched approximation—using staggered (most accurate), (improved) Wilson and domain wall fermions. Unquenching (with unphysically heavy sea quarks) has small effect. Main uncertainties: continuum and chiral extrapolations in unquenched theory; perturbative error in normalization of the operator, The latter error, presently about 5%, will likely become dominant for terascale machines. It may be possible to reduce this error by further theoretical work.

We now list other electroweak matrix elements whose calculation could help probe the Standard Model or distinguish between models of physics beyond the Standard Model, and for which a lattice method is known. Most of these are less accurately known than the key matrix elements listed above. Terascale resources will lead to great improvement in the accuracy of the results for these quantities.

- $f_D\sqrt{B_D}, f_{D_s}\sqrt{B_{D_s}}$. Needed to estimate $D \leftrightarrow \bar{D}$ and $D_s \leftrightarrow \bar{D}_s$ mixing, which, though predicted to be unobservably small in the Standard Model, may be enhanced by new physics. Status of lattice calculation similar to that for $f_B\sqrt{B_B}$ [34].

- $B \to (\pi, \rho)\ell\nu$ form factors. Can use to extract V_{ub} from experimental decay rate. Might become most accurate method for obtaining V_{ub} in a few years. Only quenched results available, and results from relativistic and non-relativistic quarks not in complete agreement. Main uncertainties: quenching errors, normalization of currents [34].

- $B \to (D, D^{(*)})\ell\nu$ form factors. Can use to extract V_{cb} from experimental decay rate. Very accurate lattice results available in quenched approximation using ratio method—essentially removes normalization errors [30]. Main uncertainties: quenching errors.

- $D \to \pi\ell\nu$ form factor. Can use to provide an independent determination of V_{cd}. Only quenched results available.

- $B \to K^*\gamma$ and $B \to K(K^*)\ell^+\ell^-$ form factors. These two processes provide sensitive probes of new physics, and are complementary to each other. Only preliminary quenched results available [31].

- Matrix elements of four-fermion operators arising from physics beyond the standard model and which cause $B \leftrightarrow \bar{B}$, $D \leftrightarrow \bar{D}$ and $K \leftrightarrow \bar{K}$ mixing [32]. These operators can have a different chiral structure from those induced by the weak interactions. Calculations are no more difficult than those for B_B and B_D, or, in the light quark case, for B_K, but have not yet been pursued beyond the quenched approximation. Main uncertainties: extrapolation in heavy quark mass, quenching errors, normalization of operators.

- CP-violating part of $K \to \pi\pi$ decay amplitude. Needed to see if Standard Model is consistent with measured value of ϵ'/ϵ. Dominant contribution comes from two operators—O_6 and O_8. Final state interactions and chiral symmetry both very important. Lattice methods known in principle, but only severe approximations have been studied in practice [36]. Terascale facility should allow first complete quenched study.

- $B \to \pi\pi$ and other non-leptonic decays of B and D mesons. Lattice methods known in principle [37], but not yet implemented. Calculations of relative strength of "penguin" and "non-penguin" contributions would greatly aid studies of CP-violation at B-factories.

- Electric Dipole Moment of the neutron (d_N) induced by operators of the form $\bar{q}\sigma_{\mu\nu}F^{\mu\nu}q$ and $\bar{q}\sigma_{\mu\nu}G^{\mu\nu}q$, where F and G are respectively the electromagnetic and gluon field strengths. These operators are present in supersymmetric extensions of the standard model, and could induce a measurable d_N [38]. Not yet studied on the lattice.

Finally, we list hadronic matrix elements whose calculation serve as important tests of lattice methods because they are known from experiment. The hadron spectrum itself is, of course, such a quantity. Almost none of the quantities listed below have been calculated beyond the quenched approximation. Terascale resources will allow unquenched calculations, and thus give dramatic reductions in the systematic errors in the results.

- f_π. A simple quantity which serves as a first benchmark for the reliability of light quark matrix elements. Accurately calculated in the quenched approximation, and lies about $\sim 10\%$ below experimental value [39]. Main uncertainties: quenching errors and chiral extrapolation.

- f_K/f_π. Ratio can be calculated more reliably than f_π itself, and comparison with experiment tests correct inclusion of "pion cloud" around hadrons. Quenched result lies significantly below the experimental value, 1.22 [39]. Main uncertainties: quenching errors and chiral extrapolation.

- f_ρ. Determines $\rho \to e^-e^+$ decay rate. Tests light quark matrix elements in vector meson sector. Only quenched results available.

- $K \to \pi\ell\mu$, $D \to K\ell\mu$ and related form factors. Since relevant weak mixing angles (V_{us} and V_{cs}) are known independently, these provide tests of the methodology for calculating form-factors, and thus serve as partial benchmarks for the B meson form-factor calculations discussed above. Only quenched results available.

- $\mathcal{A}(K \to \pi\pi)$. Calculation of the $K^\pm$ and K^0 decay amplitudes from first principles will test whether the $\Delta I = 1/2$ rule is explained by QCD. This is a challenging calculation because it

involves two final state particles and chiral symmetry must be maintained. Only recently has a lattice method been developed for including final state interactions correctly [37]. Appropriate chiral symmetry can be maintained using staggered or domain wall/overlap fermions. Only approximate calculations have so far been done.

- Structure functions determined experimentally using deep inelastic scattering with electroweak probes. These are discussed in Section 2.3.
- Radiative and weak transition rates between baryons. Very limited quenched calculations done so far, although straightforward in principle.
- f_{D_s}. Present estimate 255 ± 30 MeV [34]. Status is similar to that for $f_B\sqrt{B_B}$. Important quantity because it can be measured (present result is 270^{+30}_{-34} MeV [35] and tests lattice methods for heavy-light systems.
- $\tau(\Lambda_b)/\tau(B_d)$ and other lifetime ratios of hadrons containing b-quarks. Experimental ratios differ from unity by more than expected using heavy quark effective theory (HQET) analysis. Resolution requires large matrix elements of $1/m_Q^3$ contributions. Only preliminary quenched lattice results available [33].
- $\Delta\Gamma_{B_s}/\Gamma_{B_s}$. Width differences between two $B_s - \bar{B}_s$ eigenstates may be measurable at B-factories and can be evaluated using HQET, up to certain matrix elements. Only quenched results for these matrix elements are available [35].

A.2.2 Prospects for Quark Gluon Plasma Calculations

For completeness we summarize the elements mentioned in the main body of the proposal.

- Order of the phase transition with strange quarks. At present there are contradictory results using the two popular fermion schemes. Agreement is expected, once effects of lattice artifacts are removed. We expect to do much better with the improved fermion schemes.
- Phase diagram in temperature and quark mass including the strange quark. A fair amount is known about the phase diagram with only up and down quarks. Very little outside model predictions is known about the phase diagram with strange quarks. This study will remedy the situation.
- Equation of state. Up to now studies have mostly focused on a plasma without a strange quark. Studies with the new actions have shown significant changes around the crossover, showing the importance of reducing lattice artifacts.
- Strange quark content of the plasma. There is some preliminary work on coarse lattices by the Bielefeld group, but nothing is known about how the results scale with smaller lattice spacing.
- The axial $U(1)$ anomaly. Studies with domain wall fermions have shown the importance of getting the chiral properties right. The next step is to assure that lattice artifacts are also under control.
- Role of instantons. Recent lattice studies at low temperature have shown that with a combination of improvement and reasonably small lattice spacing some elements of the instanton liquid model may be found in lattice calculations. However, it remains to be established whether instanton molecules are important at the crossover.

- Real-time excitations. Nothing much is known from lattice calculations about real-time excitations.

A.2.3 Prospects for Hadron Structure Calculations

A national group of physicists has been collaborating since August 1998 to build a coherent effort to use lattice QCD to understand hadron structure and interactions. The resulting Lattice Hadron Physics Collaboration, which is open to all U.S. physicists interested in collaborating to share national resources for hadron structure research, presently includes 24 physicists beyond the postdoctoral level from 15 institutions. Major research projects they have proposed, and would carry out at a QCD topical computer center in the early years include the following:

- Calculation of the strange quark contribution to the nucleon's electromagnetic form factors. These results will elucidate the strange quark content of the nucleon and complement fundamental parity-violating electron scattering experiments.

- Precision calculation of nucleon form factors with sufficiently light quark masses to include the physics of the pion cloud. This analysis based on QCD is needed, for example, to understand the recent precise measurements of the ratio of the proton's electric and magnetic form factors and to understand measurements of the neutron's electric form factor.

- Calculation of moments of nucleon parton distributions and nonleading twist operators in order to understand the quark-gluon structure of the nucleon. The results will provide fundamental understanding for HERMES and RHIC-spin high-energy experiments, and for experiments at JLab using 6–12 GeV electron beams.

- Calculation of leading light-cone quark-distribution amplitudes of the nucleon. These amplitudes determine the normalization of perturbative contributions to form factors and large momentum-transfer Compton scattering.

- Calculation of the spectrum of lowest lying N^* resonances for several spins in order to determine the lattice QCD spectrum in a mass region where states are predicted to exist in the quark model, but none have been observed experimentally.

- Calculation of N^* and Δ transition form factors. These calculations would be provide essential information about how any deformation of the nucleon is manifested in transition form factors and complement the new experimental N^* program.

- Calculation of the Born-Oppenheimer potential between two baryons, each of which contains one heavy quark and *two* light quarks, would address the central problem of establishing the link between QCD and the nucleon-nucleon interaction. The use of one heavy quark in each baryon allows for a clean definition of the relative coordinate.

- Exploratory studies of chiral phase transitions of staggered fermions at finite temperature and chemical potential using the meron cluster algorithm in order to develop a practical algorithm to study quantum field theory at finite baryon density. Exploration of the quantum link D-theory formulation of QCD with cluster algorithms.

A.3 Computational Resource Estimates

In this appendix we estimate the resources required to carry out a few next generation lattice gauge calculations. We focus on the improved staggered quark action, Asqtad, recently developed by members of our group,[5], on the Wilson action, and on the domain wall quark action. Our estimates for the staggered and domain wall quark actions are based on extrapolations from recent production or tuning runs. For these extrapolations we assume that at fixed physical lattice size and π to ρ mass ratio, the computing resources increase approximately as a^{-7} as the lattice spacing a decreases. Four powers of a^{-1} arise from the change in number of lattice points, and one power each from the inversion of the Dirac operator, the decrease in step size, and the increase in autocorrelation time. We envision working at the physical value of the strange quark mass, and extrapolating in m_π/m_ρ by varying the bare up and down quark mass. If this is done at fixed lattice spacing, then the computing resources are expected to vary as $m_{u,d}^{-2.5}$, one power of $m_{u,d}^{-1}$ being associated with the inversion of the Dirac operator, the second with the step size, and the square root with the correlation length, which is inversely proportional to the pion mass.

A.3.1 Electroweak Matrix Elements

• **The Matrix Element B_K**

As discussed in the text, a terascale facility would allow a dramatic reduction in the errors in various weak matrix elements. Here we elaborate on a particular case, B_K. The present state-of-the-art quenched calculation [42] uses unimproved staggered fermions, and requires the use of lattice spacings down to $a \approx 0.05$ fm ($\beta = 6.5$) for controlled continuum extrapolation. Such a small lattice spacing is required to disentangle the expected order a^2 discretization error. Using of order 100 configurations per mass value/lattice spacing, the statistical error is much smaller than that due to the one-loop perturbative matching factor, which is about 5%. Aside from this error, the dominant systematics arise from quenching and the use of a kaon composed of degenerate quarks. Preliminary small-scale results with dynamical fermions bring the total error estimate up to about 20%. We consider here a simulation to reduce these systematics to the point that the perturbative errors are dominant—i.e. about a four-fold reduction in uncertainties. Bringing the theoretical errors down to the level of the experimental ones would substantially tighten the constraints on the CKM matrix. Further improvement is presumably possible with additional theoretical work.

To proceed we will use the parameters of Ref. [42], but will generate configurations with up, down and strange sea quarks, using improved staggered fermions. In addition, non-degenerate quark masses will give control over all errors. To achieve the desired errors with an improved action requires a minimum lattice spacing $a \approx 0.1$ fm (using $\alpha_s \approx 0.2$). This may be overly optimistic, because to improve the staggered fermion operators themselves may be quite difficult. So, to be conservative, we assume a minimum lattice spacing of $a \approx 0.067$ fm. We assume a minimum m_π/m_ρ of 0.4 is sufficient, based on the expected dependence on quark mass differences from chiral perturbation theory. From preliminary studies we know that on $30^3 \times 60$ lattices with $a = 0.1$ fm and $m_\pi/m_\rho = 0.4$, approximately 7.5×10^{15} flops are needed to generate an independent lattice. This would require 0.21 hours on a 10 Tflop/s computer. At $a = 0.067$ fm each lattice would take 3.4 hours. There is a substantial overhead from running at a sequence of quark masses and lattice spacings, which we take to be a factor of three. Therefore this calculation would require 1,020 hours, or 1.4 months, of running on a 10 Tflop/s computer.

• $K \to \pi\pi$ **Decay**

An important opportunity provided by a multi-teraflops facility would be a domain wall fermion calculation of $K \to \pi\pi$ decay, providing lattice predictions for the $\Delta I = 1/2$ and $\Delta I = 3/2$

amplitudes, A_0 and A_2, and the operator matrix elements necessary to understand the CP violating parameters ϵ and ϵ'. This could be done assuming chiral symmetry, as in the present domain wall studies underway at the RIKEN BNL Research Center and at Tsukuba. However, the effects of dynamical fermions could now be included. Scaling from earlier domain wall fermion quenched [43] and dynamical work [44], we conclude that for a lattice spacing $a = 0.1$ fm, a quark mass giving $m_\pi/m_\rho = 0.46$, and a $24^3 \times 32$ lattice volume, an independent gauge configuration requires ≈ 12 hours on a machine sustaining 10 teraflops. A reasonable ensemble of such configurations would require several months.

These parameters would allow a meaningful comparison with the present quenched studies of A_0, A_2 and ϵ'/ϵ being carried out a $16^3 \times 32$ lattice volume with the same lattice spacing and pion mass.

A.3.2 The Quark Gluon Plasma

The cost of an equation of state study at $N_t = 10$ with up, down and strange quarks, using the improved Asqtad action, can be estimated by comparing with a conventional-fermion $N_t = 6$ study [41] which required approximately 0.5 Gflop/s-years. Scaling with lattice spacing a^{-11}, allowing for the inclusion of the strange quark and the cost of improvement, and allowing for an improvement in statistics and lattice volume gives approximately 3-5 Tflop/s-years.

Similarly, scaling from a 40 Gflop/s-year exploration of the transition region on a $16^3 \times 6$ lattice using domain wall fermions with three degenerate pions of mass ≤ 300 MeV, currently underway at the RIKEN/BNL Research Center, suggests that a significant finite-size scaling study of the order of the transition for $N_t = 6$ using domain wall fermions with $N_f = 2$ and 3 on $16^3\times$, $24^3\times$ and $32^3 \times 6$ volumes will require $\approx 4 - 5$ Tflop/s-years.

A.3.3 Structure and Interactions of Hadrons

• The Hadron Spectrum

The determination of the spectrum of QCD, including the masses of light hadrons, excited nucleons, glueballs, and particles with exotic quantum numbers, has long been a goal of lattice gauge theory. In order to reveal the physical spectrum, it is desirable to study up and down quark masses light enough for a ρ meson to decay into two pions. This requires both a decrease in the quark mass and an increase in the spatial size of the system relative to current simulations. The ρ at rest must decay into pions with nonzero momentum, and the energy of the lowest nonzero momentum state in a periodic lattice decreases as the spatial size increases. Therefore we need a π to ρ mass ratio significantly less than one half.

Extrapolating the ρ and π masses from preliminary studies with the Asqtad action at $a = 0.1$ fm, and using the formula for the energy of the lowest nonzero momentum state

$$aE(K = 2\pi/L) = \sqrt{(am_\pi)^2 + (2\pi a/L)^2}$$

we find an optimal combination of quark mass and lattice size of $am_{u,d} = 0.005$ and $L = 5$ fm. This corresponds to $m_\pi/m_\rho = 0.33$. The spatial extent is chosen to be large compared with the pion Compton wavelength $1/m_\pi$, and of sufficient size to minimize the squeezing of higher orbital angular momentum states.

Extrapolating from our preliminary runs at larger quark masses, we estimate that with these parameters 6.9 hours will be required to generate an independent lattice with a throughput of 10 Tflop/s. Lowering the up and down quark masses to $m_\pi/m_\rho = 0.184$, the physical value, with $a = 0.1$ fm, would require 18 hours to generate an independent lattice. Alternatively, holding

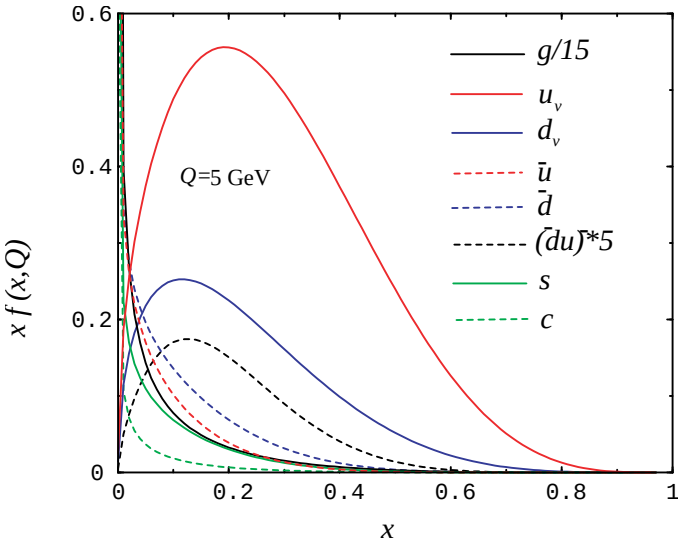

Figure 4: Parton distributions at $Q = 5$ GeV determined from a global fit to experimental data [40]. Lattice calculations will enable us to calculate from first principles moments of the distributions of gluons, g, up and down valence quarks, u_v and d_v, up and down sea quarks, $\bar{u}$ and $\bar{d}$, and strange and charmed quarks, s and c shown in this figure.

$m_\pi/m_\rho = 0.33$, and reducing the lattice spacing to $a = 0.05$, holding the physical size of the lattice fixed, requires 37 days on a 10 Tflop/s machine per independent lattice. Given that one needs at least 200 independent lattices for a spectrum calculation, these numbers make clear that such a project would require use of a 10 Tflop/s machine for well over a year. Of course the lattices generated would be used for a wide variety of applications, as would the smaller lattices generated in the calculation of B_K.

• **Moments of Nucleon Structure Functions**

One of the fundamental opportunities multi-Terascale computer resources would provide in understanding hadron structure is the calculation from first principles of the quark and gluon structure of the nucleon. Several decades of intense experimental study of the nucleon using high energy electromagnetic and hadronic probes is succinctly summarized by the experimental quark and gluon distributions shown in Fig 4 as a function of the momentum fraction x. Moments of these distributions can be calculated directly on the lattice by evaluating local operators in the ground state of the proton.

Physically, one knows that the long-distance pion cloud contributes substantially to the nucleon magnetic moment, axial charge, and moments of structure functions. Accurate calculation of the pion cloud is computationally expensive for several reasons. One must calculate at very light quark mass in order to build up the quark-antiquark excitations producing light pions, thereby encountering the algorithmic slowdown and the requirement for large physical volume described above. In addition, one must use dynamical quarks to include all the physical processes contributing to the pion cloud, precluding the economy of the quenched approximation. Although calculations do not need to be performed at the physical pion mass, it is necessary to use sufficiently light pions that reliable chiral extrapolations can be performed to the physical point. Recently, chiral extrapolations [18] were performed for unquenched calculations of moments of structure functions [17] obtained using the Wilson fermion action showing that lattice measurements of the order of 5% accuracy are needed down to a ratio $\frac{m_\pi}{m_\rho} = 0.3$ or $m_\pi = 230MeV$ in a box of linear dimension 4.3 fm for reliable

extrapolation.

The computational requirements for hybrid Monte Carlo calculations of gluon configurations with Wilson fermions are determined by the cost function obtained by the SESAM collaboration, with whom we have been collaborating in the calculation of moments of structure functions. For present purposes, the number of floating point operations per independent gluon configuration, N in Teraflops-years, may be conveniently written [45],

$$N \simeq .038 \left[\frac{L}{4}\right]^{4.55} \left[\frac{.08}{a}\right]^{7.25} \left[\frac{.3}{\frac{m_\pi}{m_\rho}}\right]^{2.7} .$$

Because the spatial derivatives and non-zero momentum projections required to calculate moments of structure functions require high Monte Carlo statistics, it is necessary to calculate 400 independent configurations. Including equilibration and calculation at higher quark masses, the total computer time is approximately twice that required for 400 configurations at the lowest quark mass. Hence, a calculation with a modest lattice spacing $a = 0.1$ fm and $\frac{m_\pi}{m_\rho} = 0.3$ requires 8 Teraflops-years, *ie*, dedicated use of a computer that sustains 8 Teraflops on QCD for one year. A more accurate calculation at $a = 0.08$ fm would require 40 Teraflops years.

Once again, we emphasize that the lattices generated in this calculation would be employed to answer a variety of other physics questions.

A.4 QCDOC Design and Construction

A.4.1 Introduction

The QCDOC machine now being designed is a natural extension of the successful QCDSP machines in operation at Columbia and the RIKEN BNL Research Center since 1998. The QCDSP architecture employs a simple, cost effective technology. At the time of that design, optimal cost performance was achieved by a small, low-power processing node capable of 32-bit floating point arithmetic, tightly coupled to a limited amount of commodity DRAM. A large number of these nodes could be packaged in a small volume and, if interconnected in a four-dimensional mesh, these processors could all be effectively used in a realistic lattice QCD calculation.

The rapid advances in silicon feature size, single-chip functional integration and communications technology, offer tremendous opportunities to exploit this successful QCDSP architecture to build an even more powerful and economical machine. For the past year and one half the group at Columbia has explored a number of alternatives and has now chosen an especially attractive technology and, with U. S. Department of Energy, RIKEN and Edinburgh/UKQCD funding, begun serious design of such a machine.

The new design is based on a high-performance, highly integrated applications specific integrated circuit (ASIC) that will combine all the functions of the QCDSP processing node on a single chip. Since this trend toward integrating all functions on a single chip is often referred to as "system-on-a-chip" technology, this new new architecture has been named QCDOC for "QuantumChromoDynamics-On-a-Chip".

The IBM corporation has agreed to support the design and manufacture of this ASIC and our design effort has been underway since December 1999. This integrated chip will contain an industry standard RISC processor capable of double-precision floating point arithmetic with a peak speed of 1 Gigaflops, 4 Mbytes of on-chip memory, and 8 Gigabits/sec of interprocessor communication. Additional memory expansion will be provided for each node to allow the architecture to be useful to a wider range of applications beyond the most demanding lattice QCD problems. The resulting processor node will cost less than $300 and consume 1-2 Watts, allowing the design of a computer

with a cost/performance more than ten times improved over the record set by the QCDSP machine in 1998. In this new machine we expand the four-dimensional QCDSP network to six dimensions. The additional two dimensions will be used to allow software configuration of the machine into a variety of partitions with the geometry of 4-dimensional tori.

A.4.2 Computer Design

• Overview

The QCDOC architecture is a natural evolution of that used in the QCDSP machines. Individual processing nodes are now 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 will be packaged in a style that provides good temperature control, a small footprint and easy accessibility. 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, 500 MHz communications and the wide array of predesigned functions needed to assemble a complete, functioning unit.

Node architecture. Each node is made of a single applications specific integrated circuit (ASIC) chip and an industry standard DIMM memory. The ASIC contains a 440 PowerPC processor core with a 64-bit floating point unit. Part of IBM's embedded PowerPC offering, this is a highly functional 32-bit processor with a 32 Kbyte, prefetching instruction cache, and a 32 Kbyte data cache with flexible cache control. The processor includes memory management with a 64-entry translation-lookaside-buffer which supports variable page sizes from 1 Kbyte to 256 Mbyte. This RISC PowerPC core is Book E compliant with dynamic branch prediction and a dual issue pipeline.

In addition to the 440 processor core, this single ASIC chip contains 4 MBytes of on-chip memory, referred to as "embedded DRAM" or EDRAM. This is sufficient to hold the code and data for a standard lattice QCD calculation and provides a large bandwidth to the processor, up to 8 GBytes/sec. (In fact, for most QCD kernels, the entire code will easily fit in the 32 Kbyte instruction cache.) In addition, this ASIC contains the DMA capability needed to move data automatically between EDRAM and external memory, the circuitry to support internode communication and an Ethernet controller for the boot-diagnostic-I/O network described below.

The single DIMM memory card, which will be part of each node, will be 64-bit-wide, 166 MHz DDR SDRAM with an additional 7 bits of ECC. The memory size is determined by the particular memory modules acquired, ranging between 32 to 512 MBytes per node. The budget in this proposal targets a 128 Mbyte DIMM card per node for a total machine memory of 2.6 Tbytes.

Inter-node Communications. Each processor will have the capability to send data to and receive data from each of its twelve nearest neighbors in six dimensions at a rate of 500 Mbits/sec. This will provide a total off-node bandwidth of 1.5 GByte/sec. Each communication link will have a phase locked receiver and single-bit error detection with automatic resend. Each of these twenty-four communication channels will have its own direct memory access capability allowing autonomous reads/writes from either EDRAM or external SDRAM. Instructions controlling each of these DMA transfers will be stored as 32 sequences of block-strided-moves located in 24 separate, on-chip register blocks. (Note, since two of these six dimensions will be used to partition the machine, only two-thirds of this communications bandwidth or 1 Gbytes/sec will be available for a typical QCD calculation.)

Low-latency, global functionality in the form of an automatic "store and forward" capability will be provided to enhance the speed of global sums and broadcasts. While this is a subset of the global operations built into the QCDSP machines, it is well matched to the intrinsic latency of the QCDOC communications network and PowerPC execution speed. This store-and-forward operation

will introduce a latency of 120ns per node in the communications path. A double-precision global sum on a 8K node partition is expected to take $\approx 15\mu$sec.

Booting, diagnostics and I/O. The SCSI booting, diagnostic and I/O network of the QCDSP machines will be replaced by 100 Mbit/s Ethernet. The Ethernet connections of four processors will be joined together with a Fast Ethernet switch whose output will be fed to a higher level switch which includes a Gbit Ethernet link. This Ethernet tree will be used in broadcast mode to provide boot code to the processors, will allow individual processors to be interrogated for diagnostic purposes and permit easy connection to industry standard RAID disks, providing a large aggregate I/O bandwidth. A fully-functional, IBM RISCWatch debugger will be provided, allowing a multi-node, window-per-processor, source-code-based graphical debugging interface.

Mechanical design. As in the QCDSP machines, we will exploit the homogeneity of this style of massively parallel machine to achieve a high degree of mechanical modularity. The individual processors will be mounted two per daughter card, one being impractical given the 5 inch width of the DDR SDRAM cards. We will mount 32 such daughter cards on a mother board and then 8 mother boards in a rather large crate with a single backplane. These crates will be cooled by vertical air flow passing through a water-cooled radiator below each crate. Cable connections will be provided on the backplane for the off-node communications of each motherboard.

• **ASIC components**

An effective overview of the QCDOC ASIC is provided by Figure 5. To a large extent, this ASIC is created from already existent IBM macros that are simply interconnected in the specified way to create the larger unit. Special to our design, in addition to our particular choice of components, is the serial communications unit (SCU), the EDRAM controller and the DMA controller permitting direct transfers between external and embedded DRAM. The various components of this diagram are briefly described below.

440 Core. A standard PowerPC Core which is specified to run at 400MHz worst case. By deliberate temperature control we expect to be able to achieve 500MHz operation. In addition to the usual PLB interface, we will design a special on-chip-memory controller, implemented following a OCM2 strategy, to efficiently use the EDRAM.

Boot/Clock support. This can be assembled out of IBM-provided components. It creates the 500MHz clock and sequences the power-up of the chip as the voltages appear.

DMA controller. Logic that we will provide, allowing automatic transfers over the processor local bus (PLB) between external SDRAM and EDRAM. This should use the full 128-bit width of the PLB and exploit whatever caching capability the DDR SDRAM controller provides.

DMA Ethernet controller. A MADMAL unit specifically configured to load and unload the Ethernet controller. This is an IBM-supplied macro.

EDRAM. This embedded 4 Mbyte DRAM provides code and data storage on-chip. In the most demanding problems, we expect to hold the entire local data and code in this on-chip location.

Ethernet controller. This is a standard EMAC3 unit that is available as an IBM macro. It has a standard interface to the OPB and the DMA Ethernet controller. We do not at present plan to implement a PHY layer on chip. Instead, this controller will drive a standard MII interface to an off-chip PHY controller.

Bootable Ethernet Interface. This second Ethernet connection permits hard-wired Ethernet control of the JTAG interface to the 440. This provides complete control of the processor to the host computer through the Ethernet interface allowing processor reset and bootcode loading directly to the 440 instruction cache. In addition, it supports the IBM RISCWatch debugging tool. This Ethernet/JTAG interface has been developed at IBM and an FPGA hardware version is currently under test at Columbia.

FPU. A 64-bit, IEEE floating point unit that is being designed by IBM as a co-processor to the 440 core. Its physical and software integration with the 440 core will be automatic and seamless.

HSSL. The physical units managing the 500 MHz serial communication in our 6-dimensional network. After an appropriate training period, the input 500 MHz serial data is byte aligned and provided to the serial communications unit at 66 MHz. Similarly, input 8-bit data is serialized and clocked out at eight times the frequency. This is a high-performance, IBM-provided macro that we will use without modification. Each HSSL controls four independent serial inputs and outputs.

Interrupt controller. A standard IBM macro enhanced with our own logic, this unit will process and combine the interrupts generated by the components of this ASIC and additional external interrupts generated elsewhere in the machine.

Location number slave. This unit will be connected to wires strapped in such a way as to uniquely determine the location of the ASIC within the larger machine. This will include location on the motherboard, location of the motherboard within the crate and of the crate within the larger machine. This location information will be used to construct a MAC address that can be used at boot-up to locate the node.

OPB. The standard IBM on-board peripheral bus. It is included here in order to remove inessential loads from the more time-critical PLB and to provide the standard interface required by the Ethernet controller.

OPB-PLB bridge. A standard IBM macro connecting the OPB and PLB. The only version of this bridge that we will need appears as a slave on the PLB and a master on the OPB.

OPB arbiter. A standard IBM macro necessary to manage the control and arbitration signals on the OPB.

PLB arbiter. A standard IBM macro necessary to manage the control and arbitration signals on the PLB.

PLB. This is the main on-chip bus. It is 128-bits wide, will run at 166 MHz and contains three somewhat independent sub-busses. Note there is a limitation of 8 masters total on the PLB and the 440 necessarily uses 3, leaving 5. At present we expect to have only three more: the SCU, the EDRAM DMA and the MADMAL DMA required for the Ethernet controller.

PLL. An IBM-supplied hard macro required by the HSSL units. Its location and wiring pattern must carefully follow IBM specifications so as to meet the requirements of the HSSLs.

PEC. This memory controller provides an independent, buffered interface for the EDRAM to the PLB, the DMA controller and the 440 core. The later connection is provided through a dedicated 500 MHZ, 128-bit PLB bus. Prefetching and buffering are included in the PEC, allowing two independent streams of sequential data to be efficiently read from EDRAM from each of these three ports. In particular, this unit provides an 8Gbyte/sec bandwidth between EDRAM and the 440's data cache. The PEC also provides standard double-bit error checking and single-bit correction for the EDRAM.

SCU. Next to the EDRAM controller, this serial communications unit requires the most design effort. It contain 12 semi-autonomous units that will assemble the incoming bytes from the HSSL units into 64-bit words which will then be fed, under DMA control, to their destination in memory. A further 12 units will parse 64-bit words fetched under DMA control from memory and provide them as a sequence of bytes to the sending components of the HSSL. Sending and receiving units will each have 3-word deep buffers and will synchronize by exchanging ACK packets. A ninth byte will be sent with each transmitted word allowing a start-up protocol to be implemented as well as parity. Bad data will be retransmitted. Finally, the SCU will provide the store-and-forward function that supports low-latency global sums and broadcasts.

A.4.3 QCDOC Schedule

Here we outline the major steps and milestones for the ongoing QCDOC design and construction effort:

- Complete Initial ASIC Design - June, 2001 At this step the design of the ASIC is complete and is released for manufacture. It is expected that manufacture will require 12 weeks after which prototype construction and testing can begin.
- Complete Final ASIC Design - December, 2001 This step permits a second design step for the ASIC allowing possible errors in the original design to be corrected. While we anticipate that this step will not be required, it is important that the schedule be arranged to accommodate such a "re-spin" step if it is needed.
- Complete operating software - December 2001 As described in Section 6.1, software development follows closely the machine design and construction schedule. The initial development of single node and then multi-node high performance code is required for architectural verification and to provide the most demanding test environment. This must be followed by low-level operating system development to support the creation of the Ethernet-based boot-kernel. Finally, complete operating software must be finished during before the 8 motherboard machine is assembled to permit its debugging and thorough testing.
- Complete Prototype - March, 2002 The prototype construction takes place in two steps. First individual daughter boards must be constructed, including the installation of the ASIC, and tested. Next two mother boards are constructed, populated with daughter boards and tested. Both of these steps require the construction of special test fixtures. We anticipate that this prototype phase will have adequately demonstrated the validity of the design and suggested necessary improvements after the completion of initial motherboard testing by November, 2001.
- Complete 8 motherboard machine - March, 2002. This larger-scale construction will follow the successful testing of the prototype daughter and motherboards in November, 2001. This 8 motherboard computer will have a sustained performance for QCD which exceeds 1/4 Teraflops and represents the first step in the large-scale production of this new computer.
- Complete 160 motherboard machine - September, 2002. This large-scale construction will follow the successful completion of the 8-motherboard machine. We anticipate that this construction will involve also machines funded by RIKEN. This 10 Tflops peak speed computer will have a sustained performance for QCD which exceeds 5 Teraflops. Note, this first 10 Tflops machine is not a component of this proposal.
- Complete QCDOC machine for the LGC at BNL - October, 2003 This construction will begin after the first large-scale construction described above is nearly complete. We anticipate that much will be learned and assimilated during the initial construction that will be important for this next step. This next step will be carried out at Brookhaven largely by the Brookhaven technical staff that has been so successful at constructing and maintaining the present RBRC QCDSP machine.

A.4.4 Construction Budget

In Table 4 we present the budget for the construction of the 20 Tflops machine (10 Teraflops sustained performance for QCD) supported by a 20 Tbyte disk storage system and an auxiliary

group of four crates of 2048 processors available for code debugging and the running of small calculations. Approximately 10% spares are included as well to cover initial failures and to provide for maintenance of the machine over an $\approx$ 6-year life span. Since we have not completed the full design of the computer, there is some uncertainty associated with the costs listed in Table 4. However, we believe that the figures presented in this table represent a reasonable estimate of the costs of components and assembly and are very consistent with our experience in building the present QCDSP machines.

Item	Unit cost	Cost ($K)
24600 nodes	$270	$6,642
370 motherboards	$3,300	$1,221
48 backplane	$5,500	$264
20 cabinets	$15,200	$304
2220 cables	$265	$588
4 crates	$5,175	$21
Ethernet switch		$440
20 TByte disk storage		$400
host computer		$120
Total		$10,000

Table 4: Breakdown of the $10M QCDOC machine cost.

A.4.5 Design Collaboration

The design and construction effort is centered at Columbia University and involves the active participation of a number of important collaborators. At present the core design group includes:

- Columbia University: Norman Christ, Zhihua Dong, Robert Mawhinney, Azusa Yamaguchi and six graduate students.
- RIKEN BNL Research Center: Shigemi Ohta(KEK/RBRC) and Tilo Wettig (Yale/RBRC)
- IBM: Alan Gara and Dong Chen.
- Edinburgh: Peter Boyle and Balint Joo.
- Fermilab: Sten Hansen.

As the design effort evolves we hope to be joined by other collaborators to critique and contribute to the design and to participate in the development of software, following a style similar to the earlier QCDSP project.

As described above, the IBM corporation is making a essential contribution to this project. By providing the design support and chip manufacturing resources need by this project, IBM supplies the advanced technology that lies at the heart of this new machine. Columbia signed a research agreement with the Research Division of IBM which provides the framework for Alan Gara's and Dong Chen's collaboration in the project and permits Columbia to contract for additional design support from the Research Division of IBM. At present we are receiving substantial assistance from five or six additional IBM scientists and engineers.

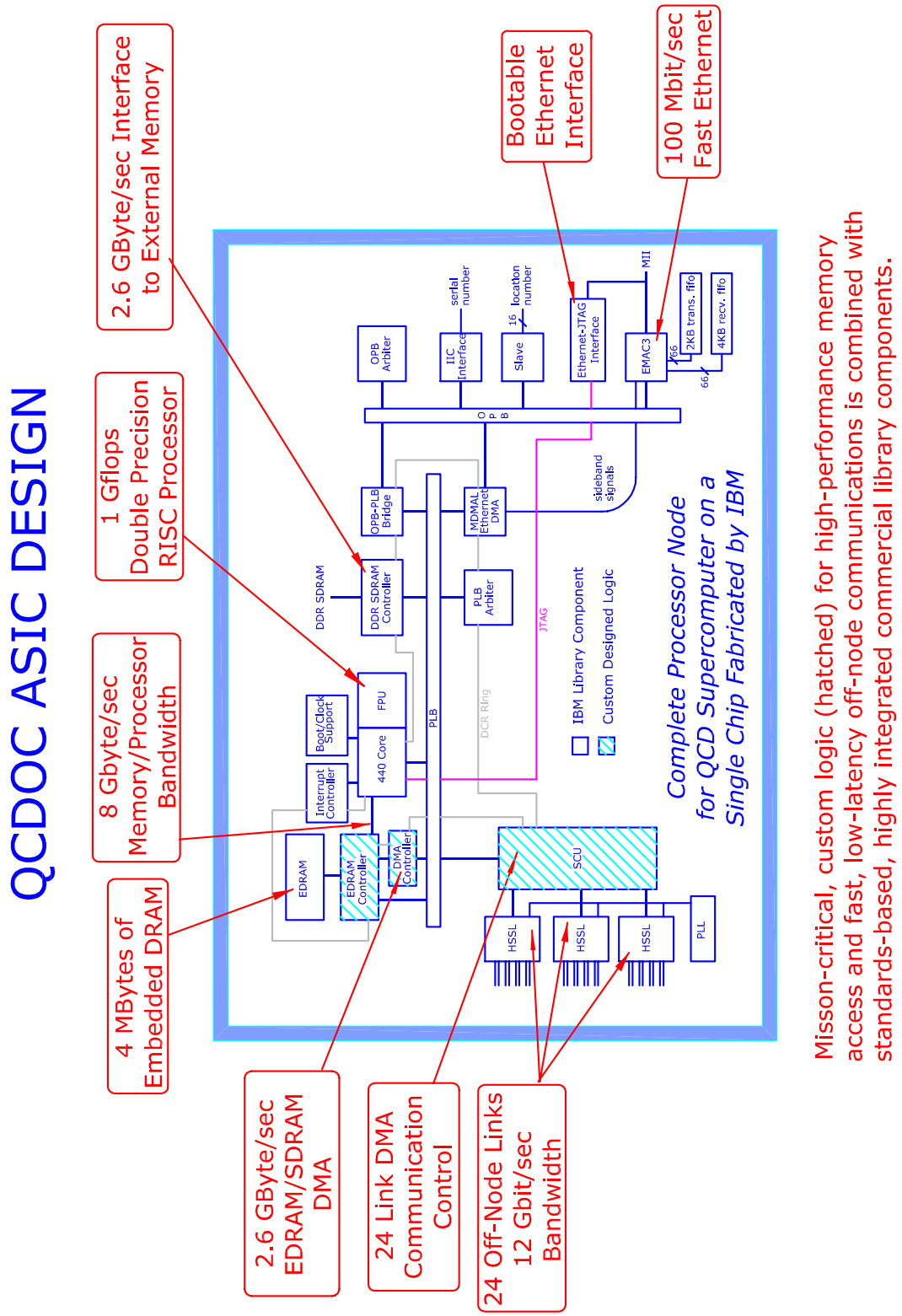

Figure 5: Block diagram representing outline of the QCDOC ASIC design

A.5 Alpha and Intel Clusters

Over the last several years, the intense competition for high performance desktops has lead to enormous gains in processor performance. At the same time, a market for specialized interconnects supporting parallel computing on clusters has made assembling supercomputer performance from commodity components relatively straight forward. In this section we sketch out the important design criteria and market trends, and describe a small number of conceptual clusters which demonstrate the potential for applying state-of-the-art commodity components to lattice QCD.

The following discussion focuses on a tentative 10 Tflop sustained cluster based upon expected evolutions of the current technology choices. Specific choices for all components, and even the overall architecture, would of course be evaluated closer to procurement to take advantage of all developments.

The design of large clusters for lattice QCD must balance a number of factors to optimize price/performance, including processor speed, cache size, memory bandwidth, and cluster interconnect bandwidth and latency. Currently the market offers a number of attractive alternatives for building clusters in the several hundred gigaflops to one teraflop range, with market evolutions supporting 10 Tflop clusters within a couple of years.

Both Intel (Pentium) and Compaq (Alpha) offer promising microprocessors for our application. The Alpha chip is currently the high performance market leader for floating point intensive applications. The Intel chips are lower performance (primarily due to their smaller caches), but their lower cost keeps them competitive. For both of these chips, there are interesting developments under way which will improve their price performance for lattice computing: (1) (for the Pentium) larger cache sizes, (2) (Alpha and Pentium) issuing 4 floating point operations per clock cycle (today's chips do at most 2), and (3) integration of glue logic onto the processor chip, yielding a lower system cost per processor.

A.5.1 Relative Performance Requirements

As a rough guide, local memory bandwidth requirements per processor for key lattice kernels are of order a byte per clock cycle (1 Gbyte/sec on current systems), a number which will fall as cache usage is optimized. For SMP systems, one must select systems containing appropriate glue logic to deliver this bandwidth to multiple processors. This is currently an area where Alpha chipsets excel.

To achieve 50% processor efficiency, lattice kernel implementers must optimize the ordering of assembly level instructions, including the use of memory prefetch instructions (included on all modern processors). Compilers alone cannot achieve this level of performance today in a way which can be scaled to real problem sizes.

Communications requirements are a function of the number of processors per box, the speed of the processors, and the size of the local lattice. For boxes sustaining 2.5 Gflops on a local lattice of 8^4, the performance drops as the cluster interconnect bandwidth is decreased much below 200 Mbytes/second (100 MBytes in each direction, achievable with Myrinet).

The network latency requirement is a function of the number of boxes, as well as the aggregate performance and lattice size, as that dictates the messaging rate. For clusters of 4096 boxes (2.5 Gflops, local lattice 8^4), a network latency below 5 μsec will deliver 90% of the aggregate single box performance for lattice QCD. (Note: the efficiency is not a strong function of the latency, dropping to only 85% if the latency is doubled.) This latency requirement is already exceeded on high performance (expensive) interconnects today, and will be achieved by inexpensive interconnects within a year.

A.5.2 Market Trends

As a point of reference, clusters today integrating dual processor alphas and Myrinet deliver a price/performance for lattice QCD of under \$7/Mflop. The processor chip price/performance is improving like Moore's law, and the system integration costs per processor (both the box and the cluster interconnect) are falling steadily, leading to a better than Moore's law improvement for large clusters. While it is difficult to accurately predict system pricing years into the future, it is safe to say that commodity clusters will deliver \$1/Mflop within a few years.

A.5.3 Previous Cluster Development Accomplishments

The proposal for the large Alpha cluster is based on the results of a sustained, ongoing research and development collaboration between MIT and JLab to exploit cost/performance optimized commodity clusters for lattice QCD. This effort has been supported by substantial investment of institutional startup funds and manpower as well as \$100K of SDAC funds in FY00. At the beginning of the project in 1998, the DEC Alpha was selected as the optimal processor for scaling to the largest clusters because of its high floating point performance and high memory bandwidth. Cost-performance was evaluated for XP1000, DS10, DS20, ES40, and UP2000 nodes, and three development clusters were purchased with institutional start-up funds: a 16 Gflop cluster of 16 single processor XP1000's, a 32 Gflop cluster of 12 dual processor UP2000's, and a 64 Gflop cluster of 12 quadruple processor ES40's.

These small development clusters have served as the testbeds for the optimization of QCD kernels at the single processor and single SMP level that provides an essential foundation for extension to the proposed large Alpha cluster. At the single processor level, the key conjugate gradient inverter for QCD already runs at 58% of peak for a small 4^4 lattice that fits in cache using hand optimized assemble routines from the UKQCD collaboration. Limitations of the memory system bandwidth decrease this performance to 43% of peak for a 16^4 lattice. Optimization at the SMP level is presently carried out using C code rather than hand-optimized assembler code, so the single processor performance is presently somewhat lower. However, as shown in Fig 6 the threaded C code scales nearly linearly from one to four processors on an ES40, with the only slight degradation for single precision occurring in the transition from one to two processors and with purely linear increase being obtained thereafter. Based on this experience, achieving 50% of peak in single precision on a 4-processor SMP appears to be a realistic goal for completely optimized code. The balance between computation and communication on the present small development clusters is such that they cannot provide a meaningful testbed for scaling to terascale clusters. Thus, it is essential to build the large proposed cluster in order to take this next step.

Another essential feature of the present development clusters is that they are presently being used for physics production by the Lattice Hadron Physics Collaboration, and thus being fully stressed and tested by demanding, real life users. Significant development effort was devoted to establishing robust operation of four-processor SMP's with the Linux operating system and the PBS batch system, and in creating a user environment adequate for a national collaboration of users. This groundwork and experience will be essential to establishing the proposed Alpha cluster as a national user facility, and thereby fully stressing and testing it for real physics QCD applications.

The proposed Intel cluster arises from a long-term effort by Fermilab, the MILC collaboration, and the Cornell collaboration to develop powerful, highly cost effective computing systems for lattice gauge theory calculations from commodity processors and interconnects. MILC code has been benchmarked and optimized on a wide variety of processors and clusters[11]. At Fermilab, an eighty

[11] `http://www.physics.indiana.edu/~sg/milc/benchmark.html`

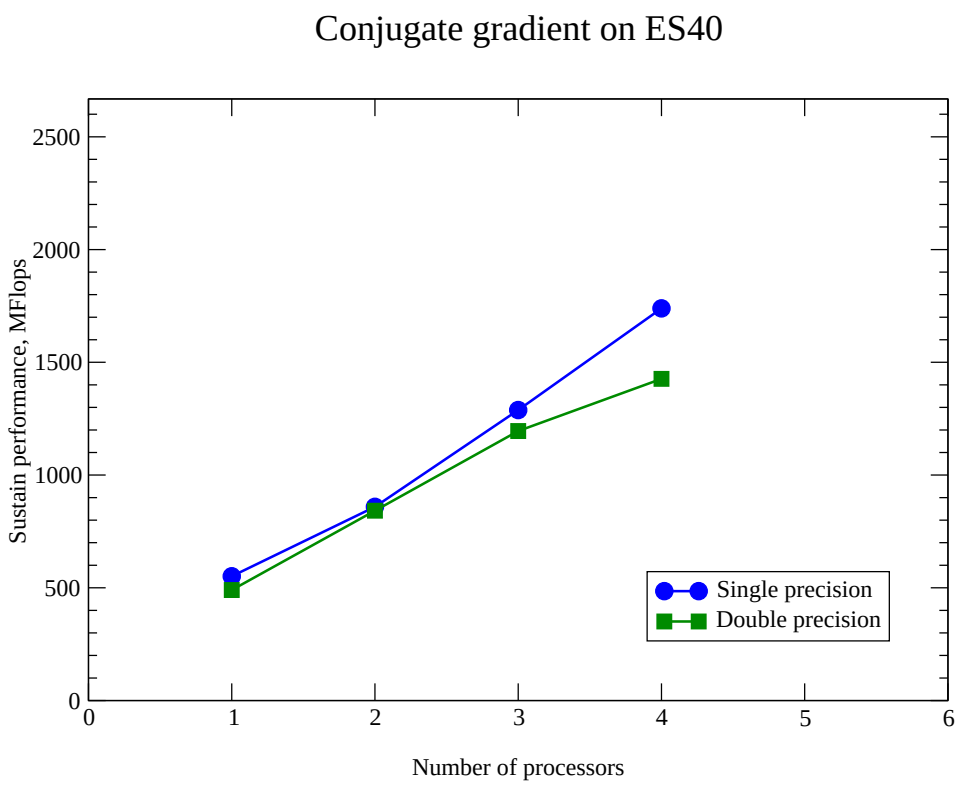

Figure 6: Scaling of conjugate gradient inverter showing nearly linear speedup in single precision when scaled from one to four processors on an ES40 with a peak performance of 1333Mflops per processor.

node prototype cluster has been deployed for this project. It was supported by Fermilab funds, augmented by $100K of SciDAC from the DOE in FY00. The cluster is constructed from Pentium 3's and a Myrinet interconnect. It is currently running and being used for physics calculations. It is also being used to study scaling of the performance with numbers of nodes. Performance on the system is being used to optimize the design of the proposed system.

Fermilab and its coworkers in the MILC collaboration have been studying and optimizing codes on a variety of processors and clusters. Performances of partially optimized full lattice QCD MILC production codes of 450 and 120 megaflops (in cache, out of cache) have been obtained on the Pentium 3's in the existing cluster, 830 Mf and 300 Mf on the Pentium 4, and 1000 Mf on the forthcoming Intel Itanium processor. This assures us of a Moore's law upgrade path for some time into the future.

The eighty node cluster is in production for physics codes of the Fermilab, MILC, and Cornell groups, and thus serves as a testbed not only for performance optimization of large clusters, but also for the user environment of a much larger cluster serving as a national facility for lattice QCD. A production batch system has been installed on the eighty node cluster and is currently running scientific code. The PBS batch system and Fermilab's enstore mass storage system have been installed and are in use, helping to harden the components of the user environment of future larger clusters.

A.5.4 Commodity "Computer on a Chip"

The large clusters of "thin" (2-way or 4-way) SMP's described above will soon be superseded by more highly integrated technology. Within 2 years, Compaq will introduce systems based upon a new generation of the alpha processor, the EV7 [12] (21364). Compaq is following the path of integrating board level functionality onto a single chip (a commodity "computer on a chip"). A block diagram of this chip is shown in Figure 7. This processor chip will have a number of key performance enhancements over the current EV68 including: (1) greater than 1 GHz processor speed, (2) 1.75 MB integrated level 2 cache, (3) dual integrated Rambus memory controllers (6 GBytes/sec), (4) integrated 3 GBit/sec I/O port (for disks or other devices), and (5) 4 integrated networking links (NEWS or north, south, east, west; 3.2 Gbytes/sec each direction with 15 ns latency). These processor-to-processor links are useful for building processor arrays. A 2D mesh of 128 processors can be directly built with no additional logic per processor other than memory. The follow-on product EV8 will continue to use the same memory and I/O structure as the EV7, but will double the floating point issue per clock and increase the memory bandwidth to 16 Gbytes/sec to realize more cost-effective systems of a fixed size.

The high integration of features on this chip will allow a large number of processors to be integrated onto a single board, driving the total system cost per processor much closer to the cost of a single processor chip. Using the NEWS links, Compaq plans to offer large SMP based systems with processors configured in a two-dimensional array. Cache coherency is implemented over these network links with transparent support for multi-hops and routing. These large SMP systems can be clustered together to make powerful terascale systems.

Based upon details provided by Compaq in a private correspondence, the extremely high bandwidth NEWS connections, coupled with the extremely low processor-to-processor latency, will extend scalability into the 100 teraflops range (limited, as today, only by funding resources, not by architectural constraints). For QCD a four dimensional grid can be overlayed on the underlying NEWS grid. A calculation of the bandwidth and latency requirements for a $32^3 \times 64$ problem on a 6144-cpu EV7 machine running at 1600 MHz indicates there is more than sufficient capacity in the system to sustain 50% efficiency in applications of the Dirac operator with a resultant 10 Teraflop sustained performance. An EV8 machine would achieve that performance with only 3072 processors. Given expected pricing it is likely in the 3 year time frame that a $1 per sustained megaflop can be achieved.

The trend within micro-processors to integrate cache, networking and the control of disk and other I/O onto a single piece of silicon is part of a larger strategy to reduce system cost in a competitive marketplace. This trend will yield direct benefits not only to the commodity marketplace, but also to high performance computing. The EV7 processor and the follow-on more powerful EV8 will form the basis of future large processor count SMPs from Compaq, and will also be the basis for proposals to the ASCI program. All other major micro-processor manufacturers are expected to develop architecturally similar offerings.

In a similar way, Intel continues to push the performance of their processor line, integrating cache and moving to quad-issue floating point ("McKinley" IA-64 chip, 2002). Intel has also indicated support for the Infiniband standard, including chipsets and switches, in the 2002 timeframe.

[12] See http://www.digital.com/alphaoem/microprocessorforum.htm for a presentation given at the 1998 Microprocessor Forum. Additional details were provided by Compaq in a private correspondence.

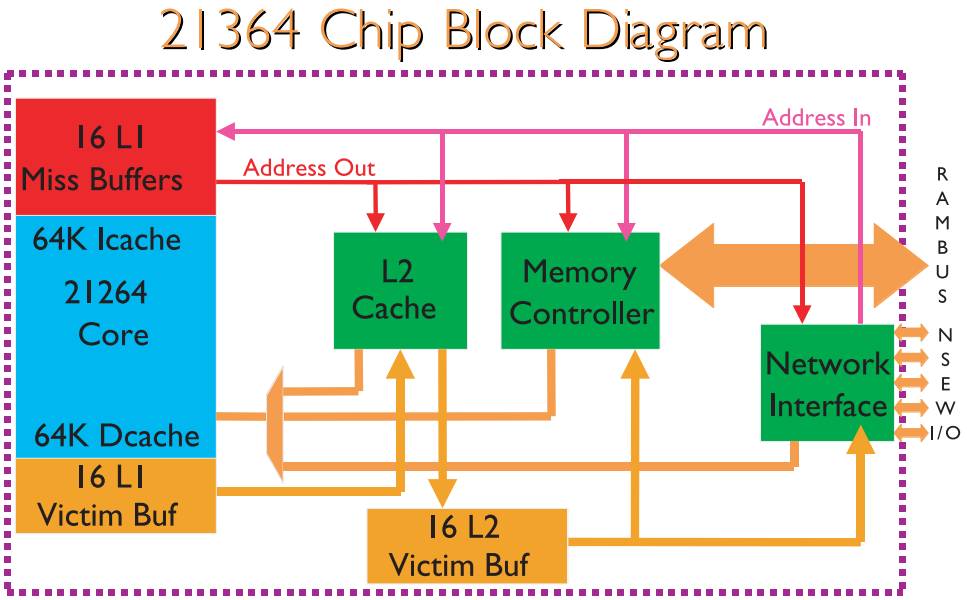

Figure 7: Block diagram of TV7 computer on a chip, integrating the Alpha 21264 core, 1.75 MB of L2 cache, memory controller for additional RAMbus memory, four direct processor-to-processor NEWS interconnects at 2×3.2 GB/sec, and a 3GB/s I/O interface suitable for intra- SMP communication. The EV8 architecture is similar, but with 8 functional units providing 4 floating point operations per clock cycle.

A.5.5 Industry Relationships

MIT and Jefferson Lab have a close working relationship with Compaq that these institutions will continue to utilize in the development of QCD clusters. The 64 Gflops cluster of twelve 4-processor SMP's acquired at MIT under a cooperative research agreement has already played an essential role in performance optimization at the processor and SMP level. Nondisclosure briefings on EV7 and EV8 technology provide the foundation for our planning for the next generation, computer-on-a chip cluster technology, and an agreement is in place for simulation of our QCD application software on EV7 and EV8 architectures.

We have also had non-disclosure meetings and helpful discussions with Alpha Processor, Inc. (API), the manufacturer of the CS20 1U box (an alpha clone), which currently yields the highest floating point performance per rack for commodity systems.

A.6 Senior Personnel

In this appendix we list the senior personnel who will participate in the development of this project towards its long term goal – the establishment and production use of the distributed terascale computing facility. They comprise nearly all of the senior lattice gauge theorists in the United States, as well as senior computer scientists and engineers who have agreed to participate in the project.

Claude Bernard	Washington University
Tanmoy Bhattacharya	Los Alamos National Laboratory
Richard Brower	Boston University
Thomas Blum	Brookhaven National Laboraoty
Matthias Burkardt	New Mexico State University
Shailesh Chandrasekharan	Duke University
Dong Chen	T.J. Watson Laboratories, IBM
Jie Chen	Thomas Jefferson National Accelerator Facility
Norman Christ	Columbia University
Michael Creutz	Brookhaven National Laboratory
Thomas DeGrand	University of Colorado
Carleton DeTar	University of Utah
Shao-Jing Dong	University of Kentucky
Zhihua Dong	Columbia University
Terrence Draper	University of Kentucky
Patrick Dreher	Massachusetts Institute of Technology
Anthony Duncan	University of Pittsburgh
Robert Edwards	Thomas Jefferson National Accelerator Facility
Estia Eichten	Fermi National Accelerator Laboratory
Aida El-Khadra	University of Illinois, Urbana
Rudolf Fiebig	Florida International University
Alan Gara	T.J. Watson Laboratories, IBM
Steven Gottlieb	Indiana University
Rajan Gupta	Los Alamos National Laboratory
Anna Hasenfratz	University of Colorado
Urs Heller	Florida State University
James Hetrick	University of Pacific
Donald Holmgren	Fermi National Accelerator Laboratory
Nathan Isgur	Thomas Jefferson National Accelerator Facility
Xiangdong Ji	University of Maryland
Gregory Kilcup	Ohio State University

A.7 Computer Science Participation

All of the work performed at the University of Illinois will be by computer scientists. A letter of intent from the Illinois principal investigator, Professor Daniel Reed, is attached, as is the Illinois work statement and budget.

Joseph Kiskis	University of California, Davis
John Kogut	University of Illinois, Urbana
Julius Kuti	University of Californa, San Diego
Andreas Kronfeld	Fermi National Accelerator Laboratory
Frank Lee	George Washington University
Peter Lepage	Cornell University
Keh-Fei Liu	University of Kentucky
Paul Mackenzie	Fermi National Accelerator Laboratory
Robert Mawhinney	Columbia University
Celso Mendes	University of Illinois, Urbana
Colin Morningstar	Carnegie Mellon University
John Negele	Massachusetts Institute of Technology
Shigemi Ohta	KEK and Riken BNL Research Center
Robert Pennington	National Center for Supercomputer Applications
Donald Petravick	Fermi National Accelerator Laboratory
Andrew Pochinsky	Massachusetts Institute of Technology
Claudio Rebbi	Boston University
Ronald Rechenmacher	Fermi National Accelerator Laboratory
Daniel Reed	University of Illinois, Urbana
David Richards	Thomas Jefferson National Accelerator Facility
Stephen Sharpe	University of Washington
Junko Shigemitsu	Ohio State University
James Simone	Fermi National Accelerator Laboratory
Donald Sinclair	Argonne National Laboratory
Amarjit Soni	Brookhaven National Laboratory
Robert Sugar	University of California, Santa Barbara
Eric Swanson	University of Pittsburgh
Harry Thacker	University of Virginia
Doug Toussaint	University of Arizona
Steven Wallace	University of Maryland
William Watson, III	Thomas Jefferson National Accelerator Facility
Tilo Wettig	Yale University
Uwe-Jens Wiese	Massachusetts Institute of Technology
Walter Wilcox	Baylor University