Accelerating Quantum Monte Carlo Simulations of Real Materials on GPU Clusters

More accurate than mean-field methods and more scalable than quantum chemical methods, continuum quantum Monte Carlo (QMC) is an invaluable tool for predicting the properties of matter from fundamental principles. Because QMC algorithms offer multiple forms of parallelism, they're ideal candidates for acceleration in the many-core paradigm.

In 1929, Paul Dirac, wittingly or unwittingly, issued a grand challenge that continues to engage the scientific community even today:

> The underlying physical laws necessary for a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.[1]

This challenge reflects the understanding that an exact solution of the Dirac equation (or its nonrelativistic counterpart, the Schrödinger equation), would let us predict all of matter's properties and behavior, at least at terrestrial energy scales. For a system containing N electrons, the Schrödinger equation, which governs the quantum mechanical wave function, Ψ, is a partial differential equation in $3N$ dimensions. Because of this high dimensionality, exact solutions can be found only in extremely simple cases, such as in an isolated hydrogen or helium atom. As such, Dirac's challenge has yet to admit defeat. Nevertheless, more than eight decades of development in theoretical physics and chemistry have produced means of finding evermore accurate—albeit approximate—solutions to these equations.

This methodological development, coupled with an exponential increase in computing power, has enabled progress from qualitatively correct models based on empirically determined parameters to truly first-principles calculations with significant predictive capability. Such ab initio methods take as input only the atomic numbers and the atoms' positions in the molecule or material. These methods complement experiment and analytic theory, have helped drive discovery, and advance understanding across a broad range of disciplines.

As we describe here, continuum quantum Monte Carlo (QMC) has proven to be an invaluable tool for predicting the properties of matter from fundamental principles. By solving the many-body Schrödinger equation through a stochastic projection, QMC achieves greater accuracy than mean-field methods and much better scalability than quantum chemical methods, enabling scientific discovery across a broad spectrum of disciplines. The multiple forms of parallelism afforded by QMC algorithms make them ideal candidates for acceleration in the many-core paradigm.

1521-9615/12/$31.00 © 2012 IEEE
Copublished by the IEEE CS and the AIP

Kenneth P. Esler, Jeongnim Kim, and David M. Ceperley
University of Illinois at Urbana-Champaign
Luke Shulenburger
Carnegie Institution of Washington

Here, we present the results of our effort to port the QMCPack simulation code to the Nvidia CUDA GPU platform. We restructure the CPU algorithms to express fine-grained additional parallelism, minimize GPU–CPU communication, and efficiently utilize the GPU memory hierarchy. Using mixed precision on GT200 GPUs and MPI for intercommunication and load balancing, we observe typical full-application speedups of approximately 10 to 15 times relative to quad-core Xeon CPUs alone, while reproducing the double-precision CPU results within statistical error.

Ab Initio Methods

Ab initio methods can be categorized in roughly three general areas. Mean-field methods, including Hartree-Fock (HF) and density-functional theory (DFT), make an approximate mapping of the $3N$-dimensional equation onto a coupled set of N 3D equations. Although these methods often yield great insight and reasonable accuracy, there are many outstanding cases in which a mean-field description is quantitatively, or even qualitatively, incorrect.

A second class of methods, utilized primarily in quantum chemistry, expands the wave function in a linear basis that's truncated intelligently to capture the important physics. Although these can be quite accurate, the computation time scales as N^4 or higher, which limits the most accurate calculations to small systems.

The final class, continuum quantum Monte Carlo (QMC) methods, addresses the Schrödinger equation's high dimensionality by casting it as an integral, which is evaluated through stochastic sampling. By treating electron correlation in a natural and robust way, QMC has produced results significantly more reliable than those provided by mean-field methods.[2] The computational cost of QMC grows as N^3, the same as DFT, although QMC's prefactor is typically two to three orders of magnitude greater. Nonetheless, simulations including tens of atoms and hundreds of electrons are now routine, and landmark simulations beyond a thousand electrons have been performed. The need for greater accuracy has led to QMC's rapidly increasing adoption in areas of chemistry, physics, and materials science.[3]

During the '80s and '90s, computational methods enjoyed a "free lunch" from advances in silicon process technology. Serial codes could expect a doubling of clock speeds about every 18 months. Taking advantage of larger installations required parallelizing the code with a message-passing layer such as the message passing interface (MPI). Although this process was often nontrivial, purchasing this "lunch pass" gave you access to any number of publicly allocated clusters, which have grown over the years to exceed one petaflops of performance. In the past decade, complementary metal-oxide semiconductor (CMOS) technology has hit the power wall and clock speeds have plateaued; consequently, the growth in computational power has come almost exclusively through parallelism. In particular, multicore processors have enabled a rapid increase in the core-count of clusters without a concomitant increase in cost.

In recent years, a new computational meal ticket has hit the market, in the form of computing with GPUs. These processors push beyond multicore to the extreme of many-core by combining numerous floating-point units and relatively simple execution units with a wide memory bus to allow dramatically higher peak throughput than conventional CPUs. Single-GPU performance exceeding 20 CPU cores isn't uncommon, and some applications have enjoyed much higher degrees of acceleration. To make use of these capabilities, algorithms and data structures must be reorganized to expose parallelism and make good use of a given GPU platform's explicit memory hierarchy. Work must be divided between the CPU and GPU, and the transfer of data between the two explicitly managed. As was the case with moving from workstations to clusters, this investment is nontrivial.

The many levels of parallelism afforded by QMC algorithms make QMC intrinsically scalable and ideally suited to take advantage of this many-core architecture. Early work that partially accelerated QMC algorithms on GPUs showed promise.[4] Other electronic structure methods, including HF and DFT, have also benefited from GPU acceleration.[5] As such, we elected to make the investment, and have ported much of our QMC code to Nvidia's Compute Unified Device Architecture (CUDA) GPU platform. We're not yet certain whether the long-term payoff for moving to this new computing paradigm will be as large as it was with distributed-memory clusters, but our hopes remain high.

Quantum Monte Carlo Methods

Many methods fall under the broad heading of QMC. Here, we consider the two continuum methods most widely used: variational Monte Carlo (VMC) and diffusion Monte Carlo (DMC). These methods are stochastic means to solve the time-independent Schrödinger equation—an eigenvalue equation of the form $\hat{\mathcal{H}}\Psi = E\Psi$, where

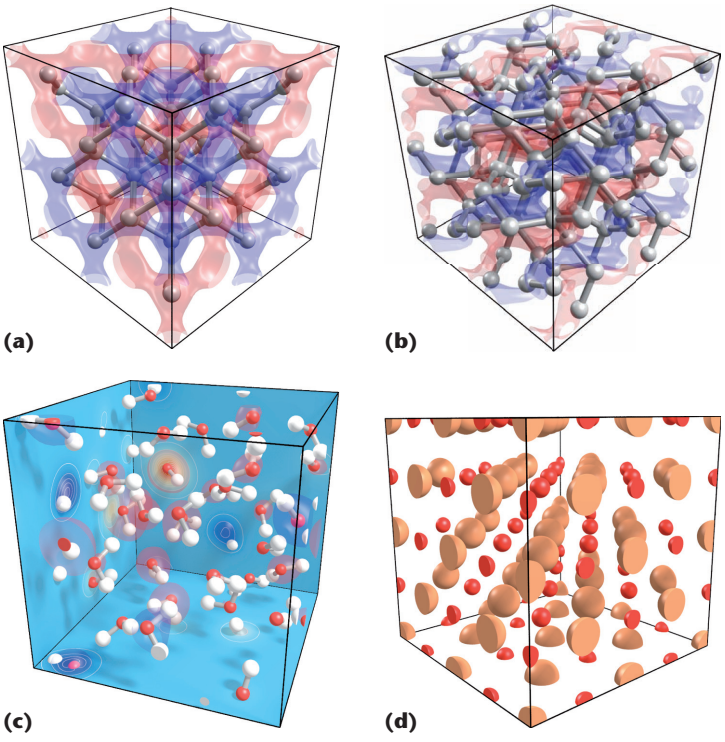


Figure 1. Visualizations of atomic structures and orbitals from systems under study with the quantum Monte Carlo package (QMCPack). Red and blue transparent surfaces are positive and negative isosurfaces of example orbitals: (a) the diamond phase of carbon, (b) the BC8 crystalline phase of carbon, (c) a configuration of 32 molecules of liquid water, and (d) the mineral wüstite (FeO) in its rock salt structure.

$\hat{\mathcal{H}}$ is the *Hamiltonian*, or total energy operator. The Hamiltonian depends on the type of system studied and describes the system interactions.

Here, we consider a system of electrons and ionized atomic cores, so that

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m}\nabla^2 + \sum_{i<j\in\text{elecs}} \frac{e^2}{r_{ij}} + \sum_{I\in\text{ions}} \hat{V}_I^{\text{NL}},$$

where the first term is the kinetic energy, the second is the Coulomb repulsion between electrons, and $\hat{V}_I^{\text{NL}}$ is a nonlocal *pseudopotential* operator.[2] For a system containing N electrons, $\mathbf{R}$ is a $3N$-dimensional vector representing the electrons' positions.

In VMC, an approximation for the lowest energy eigenstate is found by

- parametrizing a trial wave function, $\Psi_T(\mathbf{R})$;
- generating electron positions, $\mathbf{R}_n$, sampled stochastically from the probability distribution $|\Psi_T(\mathbf{R})|^2$;
- finding the parameters that minimize the expectation value of the energy or variance over the set of samples. This yields the best wave function consistent with the parametrization.

DMC begins with an optimized trial function, Ψ_T, and projects out the ground state by repeated application of an imaginary time Green's function or *propagator*. This propagator is applied stochastically by a combination of a *drift-diffusion process* followed by a *branching process*. The drift-diffusion is akin to a force-bias Monte Carlo for classical simulations.

Example Applications

In the initial development of our GPU implementation, we focused on simulating real materials and molecules comprised of electrons and ions.[6] This let us study an extremely wide class of systems. Figure 1 shows some example applications. Figures 1a and 1b show two structural phases of carbon. Diamond is metastable at low pressure, and is believed to transition to a BC8 crystal structure at approximately 10 million atmospheres at room temperature. We're presently using our GPU code to determine the precise transition pressure, which might be of great practical importance to laser-induced fusion experiments.

Figure 1c shows a configuration of 32 molecules of liquid water. Several hundred such configurations were recently simulated using the QMC package (QMCPack) on a Cray-XT5 system, consuming more than 30 million CPU hours. We believe that our GPU version will let us further this research at greatly reduced computational cost. Figure 1d shows the mineral wüstite (FeO) in its rock salt structure. We're using our GPU version of QMCPack to understand pressure-induced changes in its electronic structure. Understanding such transitions in metal oxides is closely linked to a range of phenomena, from superconductivity to metal-to-insulator transitions.

Variational Monte Carlo

VMC computes the expectation value of a quantum-mechanical operator, $\hat{\mathcal{O}}$, defined as

$$\left\langle \hat{\mathcal{O}} \right\rangle = \int d^{3N}\mathbf{R} \left|\Psi_T(\mathbf{R})\right|^2 \mathcal{O}_L(\mathbf{R}),$$

$$\text{where} \quad \mathcal{O}_L(\mathbf{R}) = \frac{\hat{\mathcal{O}}\Psi_T(\mathbf{R})}{\Psi_T(\mathbf{R})}$$

and $\mathcal{O}_L(\mathbf{R})$ is the *local value* of the operator. When the operator is the Hamiltonian, $\hat{\mathcal{H}}$, the average of the *local energy*, $E_L(\mathbf{R})$, gives the total energy's expected value. Other observables commonly computed are the potential energy, pair correlation

functions, the electronic charge density, and, more recently, the forces on the atoms.[7,8] To compute this very high-dimensional integral, VMC uses the standard metropolis Monte Carlo method to generate samples from the probability distribution $P(\mathbf{R}) \propto |\Psi_T(\mathbf{R})|^2$.

To sample this distribution, we generate an ensemble of points in $3N$-dimensional space, $\{\mathbf{R}_i\}$, which we call *walkers*. Each walker is propagated by attempting a random move to R_i' which is accepted or rejected based on the ratio $|\Psi_T(\boldsymbol{R}_i')|^2/|\Psi_T(\mathbf{R}_i)|^2$. In practice, highest efficiency is usually achieved when only a single electron is moved at a time. After the walkers have reached their equilibrium distribution, we can compute the expected value of observable properties as an average of the local values $\mathcal{O}_L(R_i^n)$ taken over all walkers, i, and all Monte Carlo steps, n.

We write the trial wave function in an analytic form with adjustable parameters, p_j, so that $\Psi_T(\mathbf{R}) = \Psi_T(\mathbf{R}; \{p_j\})$. It satisfies a *variational principle*—that is, for any set of parameters p_j, the average energy of $\Psi_T(\mathbf{R}; \{p_j\})$ must be greater than or equal to the true ground state energy, E_0. Thus, the parameters can be optimized to minimize the energy $E(\{p_j\})$. Alternatively, we can minimize the variance of the *local energy*, $E_L(\mathbf{R}_i; \{p_j\})$, over the ensemble of samples.

Diffusion Monte Carlo

DMC modifies the stochastic procedure to repeatedly apply a Green's function operator that projects out the ground state from the trial wave function. Because smaller steps must be taken, DMC is somewhat more expensive than VMC, but provides much more accurate results.

In practice, the stochastic projection is effected through a *branching* process. After a single-particle move is attempted for each electron in turn, $E_L(\mathbf{R}_i)$ is computed for each walker. Each walker carries a weight, w_i^n, which is updated as

$$w_i^{n+1} = w_i^n \exp\left\{\tau\left[E_T - \frac{E_L(\mathbf{R}_i^{n+1}) + E_L(\mathbf{R}_i^n)}{2}\right]\right\},$$

where the superscripts label Monte Carlo generations and τ is an adjustable time step that must be small enough to avoid error. After the weights are assigned, walkers are stochastically replicated or destroyed so that the final number of copies of each is proportional to its weight. The *trial energy*, E_T, is adjusted through a slow feedback process to keep the population size centered around a target.

For bosons (such as ^{4}He atoms), the DMC method is exact (within statistical errors) if the time step, τ, is sufficiently small. However, for fermions, including electrons, the *fixed-node approximation* is introduced to alleviate the *fermion sign problem*. It yields exact results if the nodes of Ψ_T—that is, the $(3N - 1)$-dimensional surface defined by $\Psi_T(\mathbf{R}) = 0$—are exact. With approximate nodes, the energy error is quadratic in the nodal error. In practice, the error is usually quite small if Ψ_T is reasonable. For example, using a trial wave function derived from density functional theory, the accuracy of DMC usually greatly surpasses that of the DFT calculation.

The Trial Wave Function

The most commonly used trial wave function for solid-state systems takes the Slater-Jastrow type,

$$\Psi_T = \det\left[\phi_n^{\uparrow}(\mathbf{r}_j^{\uparrow})\right]\det\left[\phi_n^{\downarrow}(\mathbf{r}_j^{\downarrow})\right]e^{J_1(\mathbf{R};\mathbf{I})}e^{J_2(\mathbf{R})}$$

where ϕ_n are single-particle orbitals, and J_1 and J_2 are one-body and two-body Jastrow correlation factors. Here, $\mathbf{I} = \{\mathbf{i}_1 \ldots \mathbf{i}_M\}$ is the $3M$-dimensional vector giving the coordinates of the nuclei or ions. In the absence of magnetic fields, electrons are assigned a definite spin and the configuration is given as

$$\mathbf{R} = \{\mathbf{r}_1^{\uparrow} \ldots \mathbf{r}_{N^{\uparrow}}^{\uparrow}, \mathbf{r}_1^{\downarrow} \ldots \mathbf{r}_{N^{\downarrow}}^{\downarrow}\}.$$

For simplicity, we'll omit the spin label for the remainder of our discussion.

In general, it's possible to optimize both the single-particle and Jastrow correlation functions. In practice, only a small number of parameters (approximately 20) are typically needed to describe the Jastrow factors, while many might be required to describe the orbitals. At present, then, we parametrize and optimize only the Jastrow functions.

Standard CPU Implementation

We first describe our reference QMC implementation in QMCPack (http://qmcpack.cmscc.org) and discuss the performance issues on common architectures based on multicore CPUs. QMCPack, an open-source QMC simulation code written in C++, implements advanced QMC algorithms for large-scale parallel computers. Designed with the modularity afforded by object-oriented architecture, it makes extensive use of template metaprogramming to achieve high computational efficiency through inlined specializations. QMCPack utilizes a hybrid OpenMP/MPI approach to parallelization to take advantage of the growing number of cores per symmetric

```
for generation = 1 … N_MC do
  for walker = 1 … N_W do
    let R = {r_1 … r_N}
    for particle i = 1 … N do
      set r'_i = r_i + δ
      let R' = {r_1 … r'_i … r_N}
      ratio ρ = Ψ_T(R')|^2 / Ψ_T(R)|^2
      if r → r' is accepted then
        update inverse matrix, distance tables, etc.
      end if
    end for {particle}
    Compute local energy, E_L = ĤΨ_T(R)/Ψ_T(R)
      Kinetic energy = −(1/2)∇^2Ψ_T(R)/Ψ_T(R)
      Electron–electron energy (Coulomb)
      Pseudopotential energy
  end for {walker}
  Reweight and branch walkers
  Update E_T
  if generation > N_eq then
    Collect properties
  end if
end for {generation}
```

Figure 2. Algorithm 1: The pseudocode for diffusion Monte Carlo, with loop ordering optimized for CPUs. For good cache reuse, we attempt to move each particle within a single walker before proceeding to the next walker.

multiprocessing (SMP) node. Finally, it uses standard file formats for input and output—including XML and HDF5—to facilitate data exchange.

The main design goal of the computation kernels and physical abstractions of QMCPack is to minimize the time to complete a *generation*—that is, a step in DMC's stochastic projection process, as summarized in Algorithm 1. The propagation of the walkers within the walker loop can be carried out independently by distributing N_w walkers over parallel processing units (MPI nodes and OpenMP threads in QMCPack). Once a generation has evolved, the properties of all walkers in an ensemble are collected and E_L is averaged to determine E_T and N_w of the next generation. This requires a global communication among the parallel units and redistribution of the walkers to keep the load balanced. To initialize the DMC simulation, uncorrelated walker positions are generated with VMC. The ensemble of walkers must then be propagated for many generations (typically 500–1,000) to reach the steady-state distribution. The properties collected during this equilibration period are discarded for the final results. Figure 2 shows the DMC algorithm as implemented for CPUs.

GPU Implementation

Several differences between CPU and GPU architecture mandate changes in how algorithms are best implemented in each case. Present multicore CPUs have relatively low memory bandwidth relative to GPUs, but this deficit is partly compensated by the inclusion of a hierarchy of large caches. The working set of data required for a single walker is usually small enough to reside entirely within the L2 or L3 cache of a typical workstation CPU. In this case, it's most efficient to make many moves on a single walker before moving onto the next to reuse data in the cache.

In contrast, on GPUs, numerous in-flight threads are required to fully utilize all the functional units on the processor and to hide latency. For this reason, it's greatly advantageous to propagate all walkers in parallel—that is, to interchange the walker and particle loops in Algorithm 1 (see Figure 2). Because the working dataset for a single walker is far too large to fit in the shared memory in present GPUs, the advantage of working on a single walker at a time is also lost. In practice, this loop reordering requires restructuring all the computational kernels.

In typical simulations, only 64–256 walkers are assigned to each GPU. If only walker-level parallelism were exploited, an insufficient number of threads would be available to hide latency. Furthermore, the GPUs we employ load and store data to DRAM in 64-byte chunks. Hence, the highest bandwidth utilization is achieved when memory accesses have unit stride. This memory operation *coalescing* would be difficult to achieve if we assigned adjacent threads to distinct walkers. Thus, in each kernel, we exploit additional parallelism—typically loops over electrons, atomic cores, or orbitals. For kernels to operate efficiently, the data structures employed require reorganization to allow coalesced memory access, and to allow efficient use of shared memory. In some cases, functions are broken into two or more kernels to permit use of less shared memory, different block sizes, and synchronization between blocks.

To the major computational classes in QMCPack, we added routines that act on an ensemble of walkers rather than a single walker. These C++ member functions transfer any necessary data to the GPU, call C functions that launch CUDA kernels, and copy back results required by the host CPU. By incorporating the GPU functionality into an existing code, we've been able to perform cross checks for correctness and accuracy at each stage, essentially providing a facility for

built-in unit tests. The method's stochastic nature, combined with QMCPack's generality and complexity, has made this method of debugging and validation invaluable. Identifying one of dozens of kernels responsible for a statistically erroneous result in a stochastic code might otherwise have proven intractable.

Data Structures

For each walker, temporary data associated with each component of the wave function must be stored to avoid recomputation, such as the inverse matrices associated with determinant evaluation. To avoid spending excessive time allocating, deallocating, and transferring data in small chunks, we aggregate all data associated with a walker into a single anonymous buffer. In a pre-allocation stage, each object that requires per-walker storage makes a storage request for each data member. Each pre-allocation request returns an offset into the buffer. Requests are padded to 64-byte boundaries to assure coalesced access to global memory. After the pre-allocation requests, a single buffer is allocated for each walker. This scheme has particular benefits during DMC simulation, in which walkers are dynamically created, destroyed, and migrated between nodes.

Multi-GPU Parallelization

In both the CPU and GPU versions of QMCPack, MPI is used for parallelization between nodes to do the following:

1. Migrate walker datasets between nodes for load balancing in DMC.
2. Accumulate statistical averages.
3. Update the trial energy, E_T, as DMC progresses.
4. Broadcast large input datasets.

The dynamic duplication and annihilation of walkers during DMC simulation can result in a load imbalance between nodes. To correct this, we redistribute walkers between nodes with point-to-point transfers after each branching step. To effect this transfer, the GPU-resident walker data must first be copied to host memory. By storing the entire dataset in an anonymous buffer, this copy can be accomplished in a single transfer, greatly mitigating overhead. In practice, we've observed very little overhead from the transfers due to load balancing. All other communication (that is, steps 2–4) represents either one-time or extremely small transfers and doesn't impose practical limitations on MPI scaling.

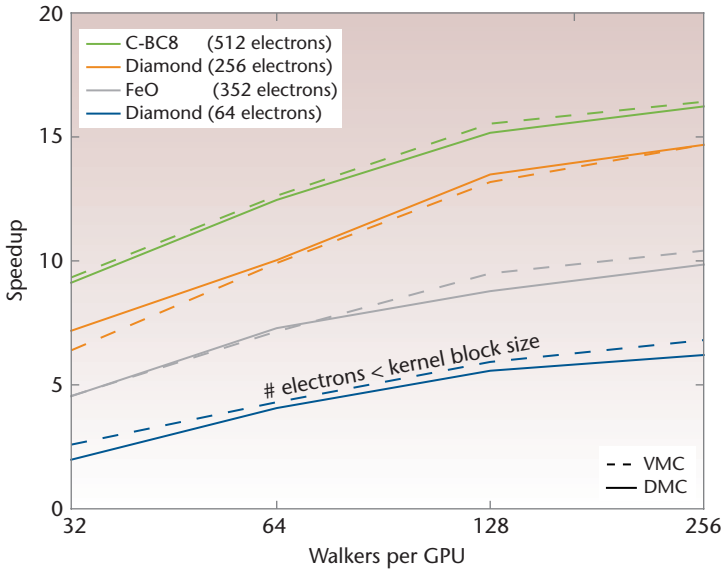


Figure 3. GPU speedup versus quad-core Xeon E5410 with four threads. The speedup is measured by the ratio of the average time to execute a block of Monte Carlo moves.

Overall Performance and Accuracy

Because the Nvidia GT200 GPU has a significant gap between single- and double-precision peak FLOP rates, we elected to utilize single precision where possible. After computing properties for each walker in single precision, results are transferred to CPU memory and ensemble averages are accumulated in double precision. We found that only one GPU kernel—the inversion of determinant matrices—required double precision to retain high accuracy. At present, however, we haven't yet modified the CPU version to take advantage of this property. Thus, in our performance evaluation, we compare double-precision performance on the CPU to a mixed-precision implementation on the GPU.

Figure 3 shows the GPU-to-CPU speedup measured by the ratio of the average time to execute a block of Monte Carlo moves. The vertical axis gives the ratio of execution rates when comparing GT200 GPUs (in Tesla S1070s) to quad-core Xeon E5410 CPUs in dual-socket compute blades. The speedup is truly one GPU to one CPU, with four threads running on each CPU, in contrast to the commonly used comparison to single-core performance.

The relative speedup increases with both the number of electrons and the number of walkers running on each GPU. This reflects the fact that many in-flight threads are required to fully hide memory-access latency. As we discuss below, some further restructuring of our kernels might lower the number of walkers needed to saturate

GPU performance. Nonetheless, for large systems in our present implementation, one GT200 can exceed the performance of 15 quad-core Xeon E5410 processors. Further, our CPU code has been highly optimized, with some critical routines hand-optimized with SSE intrinsics, so that this isn't a "straw man" comparison.

If the mostly single-precision GPU implementation doesn't yield sufficiently accurate results, it's useless. We conducted numerous tests comparing the results of the CPU and GPU implementations. So far, we haven't found any statistically significant discrepancies. Despite the use of single precision, the GPU code appears to be as accurate as the CPU version, at least to the level of statistical error that we've achieved thus far.

Main Computational Kernels Implementation

We implemented approximately 100 CUDA kernels (including a number of specializations), which perform essentially all the computation needed to simulate real materials in periodic boundary conditions and molecules in open boundary conditions. Most can be divided into kernels that evaluate wave function ratios and derivatives, and those that evaluate potential energy.

To determine the acceptance probability for single-particle moves, we need to compute the ratio of the trial wave function values at the new and old positions—that is $\Psi_T(\mathbf{R}')/\Psi_T(\mathbf{R})$, where $\mathbf{R}$ and $\mathbf{R}'$ differ by the position of one electron. To compute the kinetic energy, we also require the quantities $\nabla_i\Psi_T(\mathbf{R})/\Psi_T(\mathbf{R})$ and $\nabla_i^2\Psi_T(\mathbf{R})/\Psi_T(\mathbf{R})$. Because the trial wave function is the product of several components, we expressed the derivatives in terms of logarithms, making the complicated application of the product rule at runtime unnecessary.

Single-Particle Orbital Evaluation

To evaluate the determinant part of the wave function, we must first evaluate the single-particle orbitals, a set of functions defined on the 3D domain of our simulation cell. Other ab initio methods for solids typically expand the orbitals in plane waves, $e^{iG\cdot r}$, allowing the use of fast Fourier transforms (FFTs) to efficiently evaluate the orbitals at all points on a regular grid simultaneously. In real space, however, the number of basis function evaluations required to compute the value of an orbital at a single location grows with system size, and many thousands are needed even for a modest number of atoms.

In QMC, it's more efficient to use a localized basis with compact support. 3D tricubic B-splines provide a basis in which only 64 elements are nonzero at any given point in space,[9] allowing the rapid evaluation of each orbital in constant time. Furthermore, the basis can be improved systematically simply by decreasing the spacing parameter. On the downside, the 3D B-splines require a comparatively large amount of memory.

B-splines are also quite amenable to parallel evaluation on GPUs. When we move an electron in QMC, we must compute the values of all the orbitals at the new position. By arranging the B-spline coefficients so that the orbital index runs fastest in memory, we can ensure coalesced reads, minimizing wasted memory bandwidth. The gradient and Laplacian of the orbitals, needed for the kinetic energy, can be evaluated simply by taking derivatives of the basis functions. We released a separate CPU and GPU library for the construction and evaluation of cubic B-splines in 1D, 2D, and 3D (http://einspline.sf.net).

Determinant Ratios

Consider the Monte Carlo move of the kth electron from $\mathbf{r}_k$ to $\mathbf{r}'_k$. Defining $A_{in} = \phi_n(\mathbf{r}_i)$ and $A'_{in} = \phi_n(r'_i)$, we need to compute the ratio, $\det(A')/\det(A)$. Because we've moved only electron k, A and A' differ only by row k. By expanding in cofactors, the determinant ratio can be written in terms of the inverse of A as

$$\frac{\det[A']}{\det[A]} = \sum_n \phi_n(\mathbf{r}'_k)\left[A^{-1}\right]_{nk}.$$

The logarithmic gradient and Laplacian with respect to $\mathbf{r}_k$ of $\det(A)$ can be written similarly.

Inverses and Updates

Using the aforementioned ratio formulae requires that we store the inverse of A for each walker. To compute the inverse initially, we use a relatively simple in-place Gauss-Jordan inversion with partial pivoting. Because this process involves numerous subtractions, truncation error would accumulate for large systems in single precision, resulting in a bias in the simulation. For this reason, we perform this inversion in double precision.

Because the moves we employ change only one row of the A matrices at a time, we employ a rank -1 update of A^{-1} using the Sherman-Morrison formula. This allows the inverse to be updated in $\mathcal{O}(N^2)$ time rather than $\mathcal{O}(N^3)$ time for a full inversion.

In particular, if we replace row k of matrix $\boldsymbol{A}$ with a vector $\mathbf{v}$, then $\boldsymbol{A}^{-1}$ is updated as

$$\left[\boldsymbol{A}+e_k\otimes\delta^T\right]^{-1}=\boldsymbol{A}^{-1}-\frac{\left(\boldsymbol{A}^{-1}e_k\right)\otimes\left(\delta^T\boldsymbol{A}^{-1}\right)}{1+\delta^T\boldsymbol{A}^{-1}e_k},$$

where δ is the change in row k of $\boldsymbol{A}$, and e_k is the vector with a one as the kth element and zeros for all others. To implement this update, we divide the operations into two kernels. In the first, the product $u\equiv\delta_T\boldsymbol{A}^{-1}$ is computed. While computing this product, we store the k^{th} column of $\boldsymbol{A}^{-1}$—that is, $v\equiv\boldsymbol{A}^{-1}e_k$—contiguously in global memory. This avoids using uncoalesced reads in the second kernel, which adds on the outer product of u and v, scaled by the constant $-1/(1+u_k)$. Because repeated updates in single precision can eventually result in an inaccurate inverse, we occasionally reinvert from scratch using our double-precision Gauss-Jordan routine.

Jastrow Evaluation

Jastrow factors let the wave function have explicit correlation between pairs of electrons and between electrons and ions:

$$J_2(\mathbf{R})=\sum_{i<j}u_{ij}\left(|\mathbf{r}_i-\mathbf{r}_j|_{\min}\right).$$

The function $u_{ij}(\mathbf{r})$ depends on whether i and j have the same or opposite spin label, and is constructed to vanish for distances beyond a cutoff radius, r_c. Because we perform the simulation in periodic boundary conditions, we utilize the minimum image convention when evaluating all distances. Kernel specializations are used when r_c is less than the simulation cell's inscribed radius. This permits a much faster determination of the minimum-image distance.

The one-body, electron-ion term is similarly given by

$$J_1(\mathbf{R})=\sum_{i,j}\chi_j\left(|\mathbf{r}_i-\mathbf{I}_j|\right),$$

where $\mathcal{X}_j$ depends on the atomic number of ion j. Both the u and $\mathcal{X}$ functions are represented with cubic B-splines, whose coefficients have been constrained to give the right behavior at the origin and at r_c. The gradients and Laplacians of J_1 and J_2 are computed analytically in separate kernels.

Only a few B-spline coefficients are required to represent the function, so these can be read into shared memory at the beginning of kernel execution. Because the distances are stochastically generated, however, the index of the coefficient read by each thread is random. If more than one thread attempts to read the same coefficient simultaneously, serialization caused by bank conflicts might result if more than one broadcast is required. Nvidia's next-generation architecture, Fermi, will mitigate this by supporting multiple simultaneous broadcasts from shared memory.

Coulomb Interaction

To compute the energy arising from Coulomb interaction between electrons, we must sum over all pairs of electrons. In periodic boundary conditions, the sum must include not only the particles in the simulation cell, but all of their periodic images. A naive summation over images doesn't converge, but the rapidly converging method given by Peter Paul Ewald divides the sum into a real-space part and Fourier-space part. In practice, we use a more efficient breakup that converges even more quickly,[10] resulting in two functions, $v_s(\mathbf{r})$ and $v_l(\mathbf{r})$, the latter of which is Fourier transformed to reciprocal space, yielding $v_G(\mathbf{r})$. Then $v_s(\mathbf{r})$ is tabulated on a fine grid and used to create a 1D "texture object" that allows linear interpolation with the GPU image-sampling hardware. The long-range part is summed in Fourier space, taking advantage of the fast hardware transcendental instructions. In both cases, we use the fast on-die shared memory to minimize loads from GPU DRAM.

Nonlocal Pseudopotential Evaluation

In many first-principles methods—including DFT, quantum chemistry, and QMC—the core electrons are eliminated and their effects on the valence electrons are replaced by a nonlocal potential operator. This operator can be applied by approximating an angular integral by a quadrature on a spherical shell surrounding the atom, as

$$\frac{\left[\hat{V}_{NL}\Psi_T(\mathbf{R})\right]}{\Psi_T(\mathbf{R})}=V_{loc}(r)+\sum_{\ell=0}^{\ell_{\max}}\frac{2\ell+1}{4\pi}\left[V_1(r)-V_{loc}(r)\right]\times\sum_{r_i'}P\ell\left[\cos(\theta_i')\right]\frac{\Psi_T(r_1,\ldots r_i',\ldots,r_N)}{\Psi_T(r_1,\ldots r_i,\ldots,r_N)}.$$

The vast majority of the computation work involved in evaluating this expression is in the last term's wave-function ratios. Beyond a cutoff radius, r_c, $[V_\ell(r)-V_{loc}(r)]=0$ by construction. The first kernel for this routine simply determines which electrons are within r_c of each ion, and constructs a list of job objects, specifying the ratios

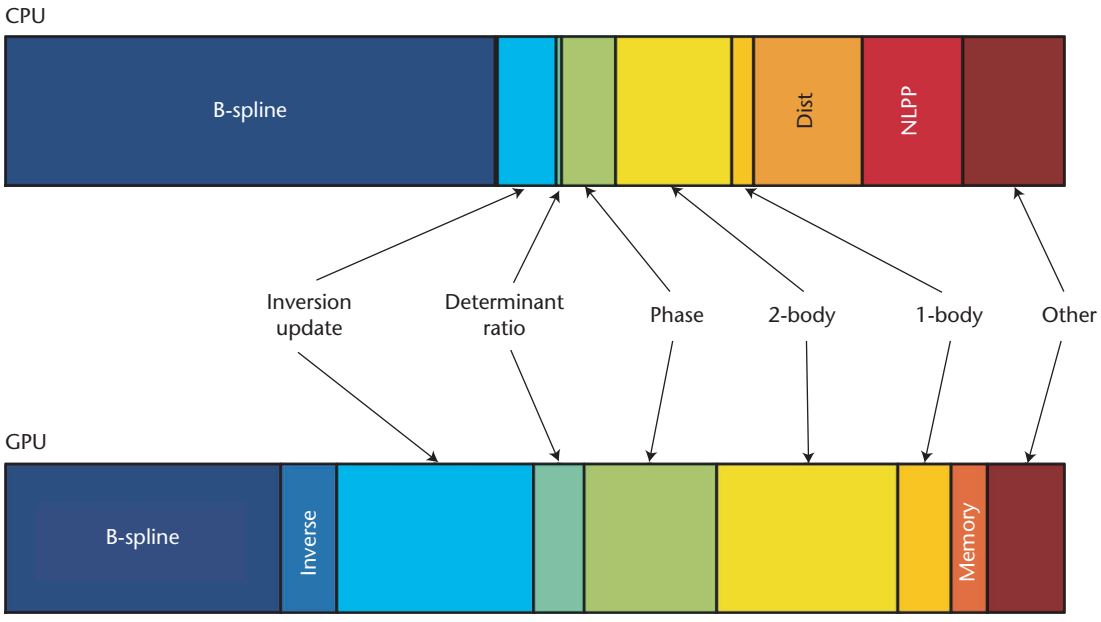


Figure 4. Execution time spent in the main categories of kernels when executed on the same physical system. Because the GPU speedup is more than 10 times, the absolute times are considerably smaller on the GPU for all kernels.

that must be computed in the second phase. These jobs are concatenated into a contiguous list on the CPU, because this can't be done easily in parallel. Finally, the ratios are computed on the GPU and passed back to the CPU, which then performs the summation.

Performance Analysis

It's instructive to consider the execution time of each class of kernels relative to the total execution time, and compare these percentages between the CPU and GPU implementations. Figure 4 shows bar charts representing the proportion of execution time spent in the main categories of kernels when executed on the same physical system. Note that these are percentages, rather than absolute times. Because the GPU speedup is more than 10 times, the absolute times are considerably smaller on the GPU for all kernels. Significant differences between the breakdowns can be readily identified; most can be understood by comparing the difference in memory hierarchy between the GT200 GPUs and Xeon CPUs.

The Harpertown CPUs in this comparison have relatively low memory bandwidth—approximately 10 Gbytes/s per socket in a dual-socket configuration—but feature a large 12-Mbyte L2 cache that often greatly mitigates this disadvantage. In contrast, the Tesla 10 series offers global memory bandwidth exceeding 100 Gbytes/s, but the user-controlled cache (that is, the shared memory) is insufficient to store some larger datasets used in our code. We can identify three classes of kernels, based on the size of their datasets:

- those with working sets too large for both CPU and GPU cache;
- those with working sets that fit in the CPU cache, but not the GPU cache; and
- those with small datasets that fit in both the CPU and GPU caches.

The most expensive kernels are those that perform the B-spline evaluation of the orbitals. Because this involves reading a series of N coefficients from random locations in a 3D table typically more than 1 Gbyte in size, cache is of little benefit. With only two FLOPs per load, this kernel is strongly limited by bandwidth, and our implementation achieves 75 percent of the peak bandwidth available on the Tesla 10. As such, B-spline evaluations constitute a smaller percentage of total runtime on the GPU.

In the CPU version, matrix inverses are computed fully only once at the beginning of a run and are then updated with each move, so that the inverse contributes a negligible amount to runtime. On GPUs, they constitute little time, though it's not completely negligible. On CPUs, the inverse matrices fit comfortably in L2 cache on Xeons and updates are relatively fast. In contrast, on GPUs, these matrices are much too large

for shared memory, so the kernel is again bandwidth limited and takes a larger percentage of time. A similar observation can be made about the determinant ratios.

The Jastrow evaluations require only particle positions and B-spline coefficients, both of which can reside in L1 on CPUs or shared memory on GPUs, so that the kernels are compute limited. In the CPU implementation, distances are computed and tabulated outside the Jastrow evaluation, while they're done inline in the GPU version. Thus, if we combine the CPU times for J_1, J_2, and distance, we find that the total percentage of runtime used for Jastrow and distance evaluation is quite similar on the CPU and GPU.

Considering these kernels, we can make several general observations about relative runtimes. When operating on datasets that can't be contained in cache on either CPU or GPU, the GPU's superior bandwidth leads to the largest observed speedup. On datasets that fit in CPU cache, but not in GPU shared memory, we observed the least speedup. On very small datasets that fit in both CPU cache and GPU shared memory, the speedup is between these two extremes.

Dealing with Limited Memory

At present, our GPU implementation lets us perform our typical production simulations with much less hardware. In the broader picture, however, we also want to use GPUs to extend the range of what's computationally feasible. Until recently, QMC's application to real materials has been mostly restricted to the bulk properties of perfect crystals. Addressing defects and disordered materials requires the ability to simulate larger systems and achieve higher statistical accuracy. At present, several factors limit the size of the systems to which our GPU implementation can be applied.

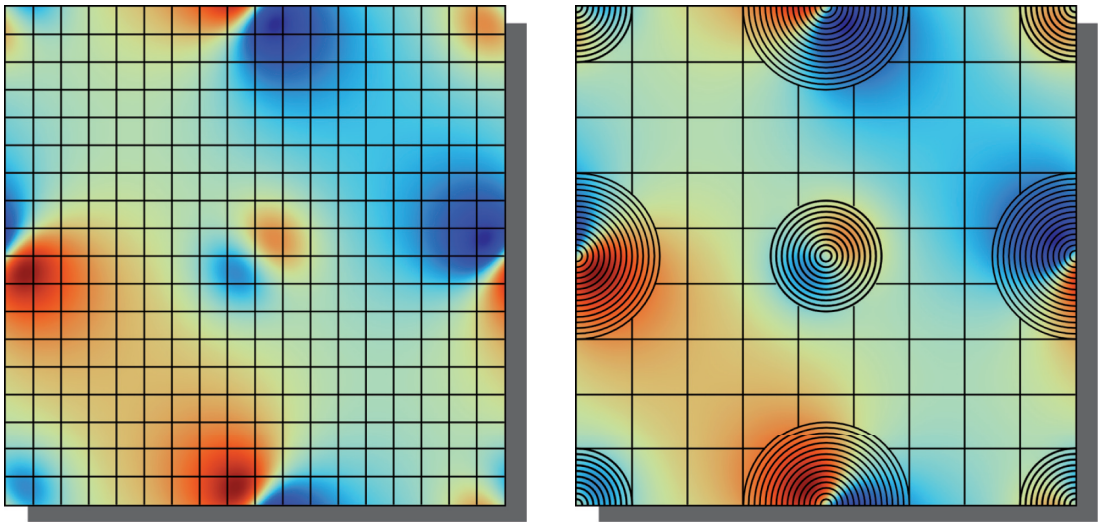
While computationally efficient, the 3D B-spline basis for the orbitals requires a large amount of memory, which grows quadratically with system size. Currently, the largest GPU memory buffer available is 6 Gbytes on the Tesla M2090 card, which limits the system size that we can address. For larger systems, several methods can be used to reduce the required storage. In a perfect crystal, symmetry can be used to let the memory usage grow only linearly with the number of atoms, allowing simulations with more than 500 electrons. In disordered systems, such as liquid water, this type of reduction isn't possible and system size is more limited.

To go further, we implement a mixed-basis approach. The mesh spacing required to accurately represent the orbitals is determined by their smallest feature size. The shortest wavelength features are concentrated around the atomic cores, while in the area between the atoms, the functions are smooth. For this reason, we divide space into spherical regions, or *muffin tins*, surrounding the atoms, and an interstitial region between them. Inside the muffin tins, the orbitals are atomic-like, and thus can be represented accurately and compactly by spherical harmonics as

$$\phi_{MT}^{n,j}(\mathbf{r}) = \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} u_{\ell m}^{nj}(|\mathbf{r}-\mathbf{I}_j|) Y_m^{\ell}\left(\frac{\mathbf{r}-\mathbf{I}_j}{|\mathbf{r}-\mathbf{I}_j|}\right).$$

Because the high-frequency components in the muffin tins have been removed, we can represent the orbitals in the interstitial region by 3D B-splines on a much coarser grid (see Figure 5). The radial functions $u_{\ell m}^{nj}(r)$ are represented as 1D Hermite splines. We didn't use B-splines as elsewhere because, in single precision, truncation error leads to a poor estimate for the second derivative. Utilizing this dual basis in place of 3D B-splines alone typically allows the same accuracy to be achieved with five to 10 times less memory, with about the same performance.

Quantum Monte Carlo simulations have long been, and continue to be, a major consumer of supercomputing power, from vector machines such as the Cray XMP[11] to superclusters such as the Cray XT6. The accuracy they afford in predicting the properties of a diverse set of real materials justifies this consumption. Nonetheless, we've seen that through the use of GPUs, the same scientific results can be achieved at considerably less expense. Furthermore, as the processors develop, the software is refined, and larger GPU clusters are built, this technology will let us progress from perfectly periodic systems to systems with defects, disordered systems, and so on. The speedups afforded by the GPUs can also help us proceed from simulations with the atoms frozen in a single configuration to those that couple the dynamics of the electrons with those of the atomic cores.[12] With further progress, perhaps in a few years, QMC simulations will be as ubiquitous as DFT calculations are today. CiSE



(a) (b)

Figure 5. Schematic of (a) uniform B-spline and (b) mixed-basis representation for orbitals. With a uniform B-spline representation, a dense Cartesian mesh is required. In a mixed-basis representation, a much coarser mesh is sufficient between the atoms.

Acknowledgments

This work was supported by the US Department of Energy (DOE) under contract no. DOE-DE-FG05-08OR23336 and by the US National Science Foundation under contract no. 0904572, and grants EAR-0530282 and DMS-1025392 to R.E. Cohen. Our research used resources of the US National Center for Computational Sciences and the Center for Nanophase Materials Sciences, which are sponsored by divisions within the DOE's Advanced Scientific Computing Research and Basic Energy Sciences under contract no. DE-AC05-00OR22725. Our research was also supported by advanced computing resources provided by the US National Science Foundation, under TG-MCA93S030 and TG-MCA07S016, and used the Abe and Lincoln clusters at the US National Center for Supercomputing Applications and Kraken at the US National Institute for Computational Sciences. Thanks to R.E. Cohen at the Carnegie Institution of Washington for helpful discussions. This work was supported in part by the Carnegie Institution of Washington.

References

1. *Proc of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 123, no. 792, 1929.
2. W.M.C. Foulkes et al., "Quantum Monte Carlo Simulations of Solids," *Rev. Modern Physics*, vol. 73, 2001, pp. 33–83.
3. K.P. Esler et al., "Quantum Monte Carlo Algorithms for Electronic Structure at the Petascale; The Endstation Project," *J. Physics: Conf. Series*, vol. 125, 2008, p. 012057; doi:10.1088/1742-6596/125/1/012057.
4. A.G. Anderson, W.A. Goddard, and P. Schroder, "Quantum Monte Carlo on Graphical Processing Units," *Computer Physics Comm.*, vol. 177, no. 3, 2007, pp. 298–306.
5. I.S. Ufimtsev and T.J. Martinez, "Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics," *J. Chemical Theory Computation*, vol. 5, no. 10, 2009, pp. 2619–2628.
6. K.P. Esler et al., "Fundamental High-Pressure Calibration from All-Electron Quantum Monte Carlo Calculations," *Physical Rev. Letters*, vol. 104, no. 18, 2010, p. 185702; doi:10.1103/PhysRevLett.104.185702.
7. S. Chiesa, D.M. Ceperley, and S. Zhang, "Accurate, Efficient, and Simple Forces Computed with Quantum Monte Carlo Methods," *Physical Rev. Letters*, vol. 94, no. 3, 2005, p. 036404; doi:10.1103/PhysRevLett.94.036404.
8. A. Badinski and R.J. Needs, "Total Forces in the Diffusion Monte Carlo Method with Nonlocal Pseudopotentials," *Physical Rev. Letters*, vol. 78, no. 3, 2008, p. 035134, doi:10.1103/PhysRevB.78.035134.
9. D. Alfé and M.J. Gillan, "Efficient Localized Basis Set for Quantum Monte Carlo Calculations on Condensed Matter," *Physics Rev. B.*, vol. 70, no. 16, 2004, p. 1661101; doi:10.1103/PhysRevB.70.161101.
10. V. Natoli and D.M. Ceperley, "An Optimized Method for Treating Long-Range Potentials," *J. Computational Physics*, vol. 117, no. 1, 1995, pp. 171–178.
11. D.M. Ceperley and B.J. Alder, "Ground State of Solid Hydrogen at High Pressures," *Physics Rev. B*, vol. 36, no. 4, 1987, pp. 2092–2106.

12. M. Dewing and D.M. Ceperley, "Methods for Coupled Electronic-Ionic Monte Carlo," *Recent Advances in Quantum Monte Carlo Methods*, World Scientific, 2002, pp. 218–253.

Kenneth P. Esler *is a senior physicist at Stone Ridge Technology, which he joined after completing postdoctoral work at the University of Illinois at Urbana-Champaign, where the work presented here was done. His research interests include the quantum simulation of materials and high-performance computing, most recently on heterogeneous platforms. Esler has a PhD in physics from the University of Illinois at Urbana-Champaign. Contact him at kesler@stoneridgetechnology.com.*

Jeongnim Kim *is an R&D staff scientist at Oak Ridge National Laboratory. Her interests include electronic structure methods and algorithms, high-performance computing, and software engineering. She has led the development of QMCPack as a research scientist at NCSA at the University of Illinois. Kim has a PhD in physics from the Ohio State University and is a member of the American Physical Society. Contact her at jnkim@illinois.edu.*

David M. Ceperley *is a professor in the Physics Department at the University of Illinois at Urbana-Champaign and a staff scientist at the National Center for Supercomputing Applications. His research interests include using high-performance computers to solve problems for quantum many-body systems, such as systems at low temperature or at high pressure. Ceperley has a PhD in physics from Cornell University and is a member of the National Academy of Sciences. Contact him at ceperley@illinois.edu.*

Luke Shulenburger *is a postdoctoral appointee at Sandia National Laboratories. His research interests include ab initio simulation of materials under extreme conditions. Shulenburger has a PhD in physics from the University of Illinois at Urbana-Champaign. Contact him at lshulen@sandia.gov.*