Linear-Scaling Quantum Monte Carlo Calculations

A. J. Williamson, Randolph Q. Hood, and J. C. Grossman

Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, California 94550

(Received 3 May 2001; published 26 November 2001)

A method is presented for using truncated, maximally localized Wannier functions to introduce sparsity into the Slater determinant part of the trial wave function in quantum Monte Carlo calculations. When combined with an efficient numerical evaluation of these localized orbitals, the dominant cost in the calculation, namely, the evaluation of the Slater determinant, scales linearly with system size. This technique is applied to accurate total energy calculation of hydrogenated silicon clusters and carbon fullerenes containing 20–1000 valence electrons.

DOI: 10.1103/PhysRevLett.87.246406 PACS numbers: 71.15.Dx, 02.70.Ss, 71.15.Nc

Over the past two decades, considerable progress has been made using quantum Monte Carlo (QMC) methods to calculate the electronic structure of realistic materials [1]. QMC has two major strengths, first a favorably "soft" scaling, where the cost to calculate the energy per atom scales as N^3, where N is the number of electrons, and, second, very high accuracy. Mean field methods such as Hartree-Fock and density functional theory possess a similar N^3 scaling, but may lack the desired level of accuracy. On the other hand, quantum chemistry approaches such as configuration interaction or coupled cluster may achieve a high accuracy, but typically scale as N^{5-7}.

Despite a number of algorithmic developments [2–7], QMC calculations for realistic systems have typically been limited to a few tens of atoms. This limitation arises from the N^3 scaling and a larger prefactor than present in conventional mean field calculations. The time, T, required to evaluate the energy of a configuration of electrons in the QMC algorithm can be decomposed as

$$T = \alpha N^2 + \beta N^3. \tag{1}$$

The N^2 terms include evaluating the electron-electron and electron-ion interactions and calculating the two-body Jastrow function. The N^3 terms include a term with a small prefactor for updating the value of the determinant after moving all the electrons and a term with a large prefactor for the calculation of the elements of the Slater determinant associated with a given configuration of electronic coordinates. The dominant cost of this Slater determinant evaluation is the focus in this paper.

The cost of evaluating the Slater determinant arises from the requirement to evaluate N orbitals for each of the N electrons. Each orbital evaluation scales as the number of basis functions. If the orbitals are expanded in an extended basis such as plane waves, then the number of basis functions is proportional to N [8] yielding the N^3 scaling. If the orbitals are expanded in a localized basis such as Gaussians, then, in principle, the size of the basis is fixed beyond a given system size and the Slater determinant evaluation would scale as N^2. However, for the system sizes studied until now, the N^3 term has remained.

In this Letter we present, for the first time to our knowledge, QMC scalings that are practically linear in the number of electrons over a wide range of sizes (20–1000 valence electrons). We form QMC wave functions with significant sparsity in the Slater determinant by choosing single particle orbitals derived from the maximally localized Wannier function construction [9]. A combination of a real space truncation of these functions and a representation in a numerical basis that is independent of system size results in an improvement in scaling from N^3 to N. We demonstrate this linear scaling for large hydrogenated silicon clusters and carbon fullerenes. Our results show that it is now feasible for QMC to routinely study systems containing up to 1000 electrons.

The introduction of sparsity into the Slater determinant part of the wave function is similar in concept to the introduction of sparsity into the overlap matrix of the Löwdin states in linear scaling density functional approaches [10,11]. However, the application of linear scaling to the QMC formalism is not fraught with the same technical difficulties that have plagued linear scaling density functional approaches. First, in QMC the Hamiltonian is not constructed from a self-consistent charge density; it is fixed. Therefore, when the orbitals are truncated, this does not affect the conservation of charge and introduce noise into the Hamiltonian. Second, the requirement to invert the overlap matrix in linear scaling density functional calculations imposes a large prefactor on these methods, such that the crossover point with traditional N^3 scaling techniques occurs around 1000 atoms. In our linear scaling QMC approach, the prefactor is the same as for the traditional N^3 implementations, and so the benefit is immediate.

Our QMC calculations were performed using the CASINO code [12], in which we adopt the standard Slater-Jastrow trial wave function [1–5],

$$\Psi = D^{\uparrow} D^{\downarrow} \exp\left[\sum_{i}^{N} \chi(\mathbf{r}_i) - \sum_{i<j}^{N} u(r_{ij})\right], \tag{2}$$

where D represents a Slater determinant, r_i and r_{ij} correspond to electron positions and separations, N is the number of electrons, χ is a one-body function, and u is a

two-body correlation factor from Ref. [13]. The Slater determinant is constructed from a set of single particle states, $D_{ij} = \phi_i(\mathbf{r}_j)$. The states $\{\phi_i\}$ are obtained from calculations of the local density approximation (LDA) to density functional theory using the JEEP code [14], which uses a plane wave basis set with periodic boundary conditions. We then use a conjugate gradient algorithm to search for the unitary transformation of the LDA eigenstates which yields Wannier functions with the maximum localization [9,15]. For finite systems, the spread functional that is minimized becomes equal to $[\langle(\mathbf{r} - \mathbf{R}_n)^2\rangle]^{1/2}$, where $\mathbf{R}_n$ is the centroid of the maximally localized Wannier (MLW) function. For example, in the hydrogenated silicon clusters studied here, the MLW functions represent the Si—Si and Si—H bonds, and the $\mathbf{R}_n$ are located close to the bond centers. As the value of a determinant is unchanged by a unitary transform, the QMC energy evaluated with a set of MLW functions is numerically identical to that obtained from the original LDA orbitals. The orbitals from which the MLW functions are constructed can be occupied or unoccupied states in an insulator, semiconductor, or metal [16], and may correspond to ground or excited states.

Once the MLW functions have been constructed, there are two important steps we implement to take advantage of the localized nature of these functions. Each step reduces the scaling by an order of N.

First, we define a cutoff radius, R_{cut}, beyond which the MLW function is smoothly truncated to zero. This truncation ensures that each electron falls only within the cutoff radius of a subset, M, of the MLW functions. For hydrogenated silicon clusters, this number is typically $M \sim 35$. Therefore, when the system size increases beyond M atoms, the number of orbitals $\phi_i(\mathbf{r})$ required to be evaluated for each electron is fixed at M and the cost of evaluating the Slater determinant then scales as MN^2.

To determine appropriate values for R_{cut}, we require the truncated MLW function to reproduce at least 99.9% of the norm of the original MLW function, and the contribution to the kinetic energy from the Slater determinant should have an error of less than 0.1%. In Fig. 1 we plot the fixed node diffusion Monte Carlo (DMC) energy and the norm of the truncated MLW function as a function of R_{cut}. Since SiH_4 has T_d symmetry, all four MLW functions are symmetry equivalent and the same value of R_{cut} is applied to each MLW function. Figure 1 shows that the DMC algorithm is stable over a wide range of values of R_{cut}; i.e., where the value of the MLW functions is negligible, truncation of these functions does not significantly affect the nodal structure of the Slater determinant. The values of the total energy agree within the 0.0003 hartree statistical error bar for all calculations above R_{cut}. In all the systems we have studied thus far, the DMC energy is well converged with respect to R_{cut} for a value of R_{cut} chosen to reproduce a norm of 0.998. Note, fixed node DMC calculations with local pseudopotentials are variational with respect to the true ground state. After truncating the MLW

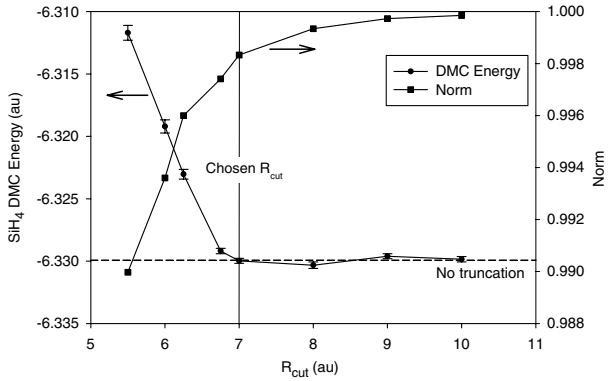

FIG. 1. The dependence of the norm of the truncated MLW function and the fixed node DMC energy on the cutoff radius R_{cut} of the MLW functions. The untruncated DMC energy is marked with the horizontal dashed line.

functions, the fixed node DMC energy is still variational with respect to the true ground state, but could be higher or lower than the original fixed node DMC energy. Figure 1 shows that severe truncation of the MLW functions dramatically increases the fixed node DMC energy. We interpret this increase as resulting from a truncation process which changes the nodal surface so that the electrons are more confined and, hence, have increased kinetic energy [17].

A second step we take to improve the scaling is to remove the dependency of the number of basis functions on the number of electrons. Again taking advantage of the highly localized nature of the MLW functions, we choose to evaluate the MLW functions using three-dimensional cubic splines [18,19]. In so doing, each orbital, $\phi_i(\mathbf{r})$, is evaluated by a single spline interpolation and the cost of evaluating the Slater determinant then scales as MN, i.e., a linear scaling algorithm. The number of spline grid points in each dimension, N_{grid}, was chosen to be consistent with the plane wave cutoff in the original LDA calculation, $N_{\text{grid}} = 2R_{\text{cut}}\sqrt{E_{\text{cut}}}/\pi$, where R_{cut} is the cutoff radius in atomic units and E_{cut} is the plane wave cutoff in rydbergs. For the systems we have studied, R_{cut} is independent of system size and, hence, the total memory required to store the spline grids for all orbitals grows only linearly with system size. This is in contrast to conventional implementations where the total storage grows as N^2, due to the growth of the basis set with system size.

To demonstrate the improved scaling resulting from our spline interpolated, truncated MLW functions, we plot (Fig. 2) the CPU time required to move a single configuration of electrons a single time step for a series of hydrogenated silicon clusters and carbon fullerenes. For the silicon clusters, three different basis sets used to expand the original LDA wave functions are compared: (a) the truncated MLW functions interpolated using cubic splines, (b) a 6-31G* quality Gaussian basis set, and (c) a plane wave basis with 11 Ry energy cutoff. Figure 2 shows that the computational cost of evaluating the LDA wave

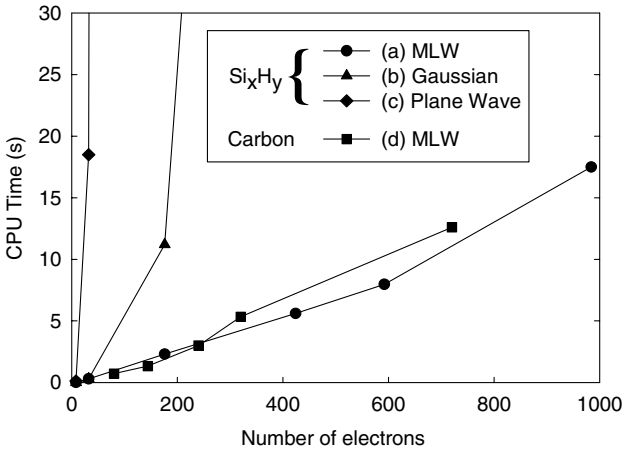

FIG. 2. CPU time on a 667 MHz EV67 alpha processor to move a configuration of electrons within DMC for SiH_4, Si_5H_{12}, $Si_{35}H_{36}$, $Si_{87}H_{76}$, $Si_{123}H_{100}$, $Si_{211}H_{140}$, C_{20}, C_{36}, C_{60}, C_{80}, and C_{180}.

functions in a plane wave basis scales as approximately N^3. The exact scaling is determined by the volume of the supercell chosen for each system. The computational cost of the Gaussian basis also scales as N^3, but with a smaller prefactor as the number of basis functions per atom is much smaller. The calculations using the truncated MLW functions demonstrate that the CPU time required to move a single configuration of electrons scales approximately linearly with the number of electrons. The deviations from linearity in the hydrogenated silicon cluster MLW curves are mainly due to differing ratios of hydrogen to silicon atoms; for the carbon fullerenes, they are due to different strain in the clusters requiring slightly different cutoff radii for the MLW functions. For the systems we compared (SiH_4 and Si_5H_{12}), the DMC energies for all three basis sets agreed within 0.001 hartree per atom statistical error bars.

Once the cost of evaluating the Slater determinant has been reduced to linear scaling, it is interesting to ask how large a system one can study before other parts of the algorithm will begin to dominate and the linear scaling will be lost. For the $Si_{211}H_{140}$ system (984 electrons), approximately 10% of the calculation involves the remaining parts of the algorithm that scale as N^2 and N^3. With relatively minor algorithmic improvements, the cost of these terms could be dramatically reduced, extending the linear regime to several thousand electrons. In particular, we envisage (i) the electron-ion interaction could be rewritten with the sum over ions precomputed so the local part scales linearly. The nonlocal contribution already scales linearly due to the cutoff in the range of the interaction. (ii) The electron-electron interaction could be rewritten to scale linearly by writing it as a sum of short- and long-ranged pieces [1,7], or using Greengard's multipole expansion [20]. (iii) To update the Slater determinant, we adopt the N^3 scaling procedure based on storing the inverse of the transpose of the matrix from Ref. [21]. Our introduction of sparsity into the Slater determinant allows us to significantly reduce the prefactor for this N^3 term. In larger systems where the determinant is increasingly sparse, it should be possible to reformulate the determinant update procedure to utilize this sparsity and obtain a better size scaling.

Note, the discussion thus far involves the scaling of the computational cost of moving a single configuration of electrons. In practice, one calculates either (i) the *total* energy of the system, or (ii) the energy *per atom*, with a given statistical error. The statistical error, δ, is related to the number of uncorrelated moves, M, by $\delta = \sigma/\sqrt{M}$, where σ^2 is the intrinsic variance of the system. Typically, the value of σ^2 increases linearly with system size. Therefore, to calculate the *total* energy with a fixed δ, the number of moves, M, must also increase linearly. When multiplied by our linear increase in the cost of each move, an N^2 size scaling is obtained. For quantities *per atom*, such as the binding energy of a bulk solid, σ^2 still increases linearly with system size, but δ is decreased by a factor of N, and, hence, the number of required moves, M, actually decreases linearly with system size. Therefore the cost of calculating energies *per atom* is now independent of system size.

To illustrate the range of systems that can now feasibly be studied within QMC using truncated MLW functions in the Slater determinant, we have calculated total energies of a series of carbon fullerenes. In Fig. 3 we plot the binding energy per atom of C_{20}, C_{36}, C_{60}, C_{80}, and C_{180} fullerenes. Line (a) shows the binding energies calculated using LDA; line (b) shows the binding energy calculated within fixed node DMC. The LDA calculations were performed at the Γ point of the Brillouin zone, using a cutoff of 40 Ry. The same pseudopotentials were used in the LDA and DMC calculations. Six points were used in the QMC angular integration for the nonlocal pseudopotential.

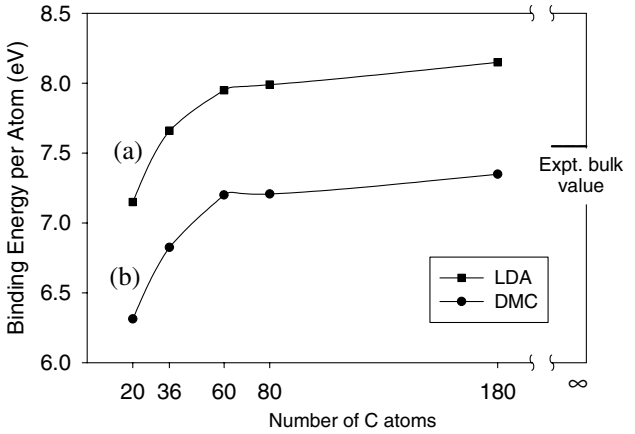

FIG. 3. Binding energy per atom (a) within LDA, and (b) within DMC of carbon fullerenes. DMC statistical error bars are smaller than the symbols. For comparison, we have added 0.18 eV zero point energy to the bulk graphite experimental binding energy [22].

The computational cost of the DMC calculations to determine the binding energy per atom with a statistical error of 25 meV was 1500 CPU hours and, as predicted above, was approximately independent of system size. Figure 3 shows that the LDA binding energies smoothly approach the bulk LDA binding energy for graphite as the size of the cluster increases. Comparison with the DMC results shows that the general trend within DMC is similar to LDA; however, the LDA overestimates the binding energy for each of the clusters.

In conclusion, we have demonstrated a method for using truncated, maximally localized Wannier functions to introduce sparsity into the Slater determinant part of the trial wave function in QMC calculations. When combined with an efficient evaluation of these localized orbitals, the cost of evaluating the Slater determinant is reduced to scaling linearly with system size for hydrogenated silicon clusters and carbon fullerenes containing 20–1000 electrons. Utilizing this method, it is possible to use QMC to study system sizes that could previously be examined only within less accurate methods. This development opens the possibility of accurately studying a range of scientifically and technologically important systems such as quantum dots, nanodevices, biological molecules, and the materials science of solids.

We thank Giulia Galli and Cyrus Umrigar for many helpful conversations. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory, under Contract No. W-7405-Eng-48.

[1] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).
[2] P. J. Reynolds, D. M. Ceperley, B. J. Alder, and W. A. Lester, Jr., J. Chem. Phys. **77**, 5593 (1982).
[3] C. J. Umrigar, K. G. Wilson, and J. W. Wilkins, Phys. Rev. Lett. **60**, 1719 (1988).
[4] S. Fahy, X. W. Wang, and S. G. Louie, Phys. Rev. B **42**, 3503 (1990).
[5] L. Mitas, E. L. Shirley, and D. M. Ceperley, J. Chem. Phys. **95**, 3467 (1991).
[6] J. C. Grossman and L. Mitas, Phys. Rev. Lett. **74**, 1323 (1995).
[7] A. J. Williamson, G. Rajagopal, R. J. Needs, L. M. Fraser, W. M. C. Foulkes, Y. Wang, and M.-Y. Chou, Phys. Rev. B **55**, R4851 (1997).
[8] Note: In periodic systems [7], the number of basis functions varies only with the size of the primitive cell. However, this N^2 scaling is useful only for examining finite size effects, not for comparing different physical systems.
[9] N. Marzari and D. Vanderbilt, Phys. Rev. B **56**, 12 847 (1997).
[10] G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).
[11] E. Hernandez, M. J. Gillan, and C. M. Goringe, Phys. Rev. B **53**, 7147 (1995).
[12] R. J. Needs, G. Rajagopal, M. D. Towler, P. R. C. Kent, and A. J. Williamson, *CASINO Version 1.0 User's Manual* (University of Cambridge, Cambridge, England, 2000).
[13] A. J. Williamson, S. D. Kenny, G. Rajagopal, A. J. James, R. J. Needs, L. M. Fraser, W. M. C. Foulkes, and P. Maccallum, Phys. Rev. B **53**, 9640 (1996).
[14] The code JEEP 1.8.0 (F. Gygi, Lawrence Livermore National Laboratory) was used.
[15] G. Berghold, C. J. Mundy, A. H. Romero, J. Hutter, and M. Parrinello, Phys. Rev. B **61**, 10 040 (2000) [Note: The transpose in Eqs. (24) and (26) should be absent.].
[16] While the unitary transform used to construct MLW functions is well defined even for systems with delocalized states, such as metals, we have yet to determine the extent to which these MLW functions can be truncated.
[17] To accurately reproduce the kinetic energy associated with the truncation of the orbitals, we truncate each orbital outside the given cutoff radius using a polynomial function that ensures a continuous function, first and second derivative.
[18] C. De Boor, *Practical Guide to Splines* (Springer, New York, 1978).
[19] E. Hernandez, M. J. Gillan, and C. M. Goringe, Phys. Rev. B **55**, 13 485 (1997).
[20] L. Greengard and V. Rokhlin, J. Comput. Phys. **73**, 325 (1987).
[21] D. Ceperley, G. V. Chester, and M. H. Kalos, Phys. Rev. B **16**, 3081 (1977).
[22] C. Kittel, *Introduction to Solid State Physics* (Wiley, New York, 1986), 6th ed.