JCTC Journal of Chemical Theory and Computation

Scaling of Multimillion-Atom Biological Molecular Dynamics Simulation on a Petascale Supercomputer

Roland Schulz,[†,‡,¶] Benjamin Lindner,[†,‡,¶] Loukas Petridis,[†,¶] and Jeremy C. Smith*[,†,‡,¶]

Center for Molecular Biophysics, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, Tennessee 37831, Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, M407 Walters Life Sciences 1414 Cumberland Avenue, Knoxville, Tennessee 37996, and BioEnergy Science Center, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, Tennessee 37831

Received June 5, 2009

Abstract: A strategy is described for a fast all-atom molecular dynamics simulation of multimillion-atom biological systems on massively parallel supercomputers. The strategy is developed using benchmark systems of particular interest to bioenergy research, comprising models of cellulose and lignocellulosic biomass in an aqueous solution. The approach involves using the reaction field (RF) method for the computation of long-range electrostatic interactions, which permits efficient scaling on many thousands of cores. Although the range of applicability of the RF method for biomolecular systems remains to be demonstrated, for the benchmark systems the use of the RF produces molecular dipole moments, Kirkwood G factors, other structural properties, and mean-square fluctuations in excellent agreement with those obtained with the commonly used Particle Mesh Ewald method. With RF, three million- and five million-atom biological systems scale well up to ∼30k cores, producing ∼30 ns/day. Atomistic simulations of very large systems for time scales approaching the microsecond would, therefore, appear now to be within reach.

1. Introduction

Molecular dynamics (MD) simulation is a powerful tool for the computational investigation of biological systems.[1] Since the first MD study of a protein in 1977, which simulated <1000 atoms for <10ps,[2] significant progress has been made in the time and length scales accessible, and it is now common to probe systems of $\sim10^5$ atoms on time scales of ∼100 ns. This increase in scope has allowed many processes of biological interest to be characterized. However, there is clear interest in further extending both the time and the length scales beyond those currently accessible.

Recent algorithmic[3−6] and hardware developments have allowed MD studies to be extended to multimillion-atom systems.[7−9] Current supercomputers, such as the “Jaguar” Cray XT5 at Oak Ridge National Laboratory used for the present study, are beginning to assemble over 1×10^5 cores and in this way reach petaflop nominal speeds. However, the challenge for MD, and other applications, is to achieve efficient scaling up to $\sim1 \times 10^4$ to 1×10^5 cores, i.e., the simulations are limited by the parallel efficiency of the MD algorithms.

The computationally most demanding part of MD simulation of biological systems is the treatment of long-range interactions, which in nonpolarizable force fields is represented by Coulomb and van der Waals (Lennard-Jones) terms.[10] While the van der Waals interaction is almost always truncated at a cutoff distance R_{vdw}, the electrostatic interaction extends to longer ranges. A common method to treat the electrostatics is to directly calculate the Coulomb interaction for any pair of atoms separated by less than another cutoff distance R_{coul} and, outside this distance, to calculate the interactions with the Particle Mesh Ewald[11,12] (PME) method

* Corresponding author. E-mail: smithjc@ornl.gov.

[†] Oak Ridge National Laboratory.

[‡] University of Tennessee.

[¶] BioEnergy Science Center.

10.1021/ct900292r CCC: $40.75 © 2009 American Chemical Society
Published on Web 08/27/2009

(assuming periodic boundary conditions are applied to the system). By using an Ewald summation to split the Coulomb interaction into a short-range part that converges quickly in real space and a long-range part that converges quickly in reciprocal space, the PME method reduces the computational cost of N particles interacting with each other from $O(N^2)$ to $O(N \ln N)$. The reciprocal space sum is performed by using the fast Fourier transformation (FFT).

Most MD calculations have been performed using PME. Although PME suffers from artifacts introduced by the periodicity,[13−17] it is fast on a small number of processors, and FFT-based electrostatics methods are also very successful on special purpose hardware. However, on massively parallel computers, the full electrostatic treatment via the PME method presents a performance barrier, arising from the state-of-the-art implementation of PME, which requires two FFT steps. The FFT algorithm in turn requires one or two global transposes which, on a message passing system, is inherently limited by the bandwidth and the latency of the network. As more nodes are used to simulate a system, each MD time-step can be calculated faster, and thus, the time between communications becomes shorter. If the time for the global transpose is of the same order of magnitude as the computation time, then the required communication becomes a bottleneck for the parallel efficiency. The same reasoning explains why, when running on the same number of cores, the parallel efficiency of a large system (e.g., 1×10^6 atoms) is much better than that of a small system (e.g., 1 000 atoms): the time needed to compute a single time step on a single processor is much longer in case of a large system. Therefore, for a large system, many more cores can be used before the communication bottleneck occurs. As a result, larger systems can often be simulated at about the same speed (in ns/day) as smaller systems.

An alternative method to PME, that avoids the electrostatics bottleneck, is the reaction field (RF).[18−21] In RF it is assumed that any given atom is surrounded by a sphere of radius, r_c, again within which the electrostatic interactions are calculated explicitly. Outside the sphere, the system is treated as a dielectric continuum. The occurrence of any net dipole within the sphere induces a polarization in the dielectric continuum, which in turn interacts with the atoms inside the sphere. Due to the approximate treatment of electrostatics outside the cutoff sphere, the RF method in principle does not yield accurate results for systems that contain full charges. Nevertheless, the RF model allows the replacement of the infinite Coulomb sum by a finite sum modified by the reaction field and, therefore, limits the parallel scaling less than the PME method.

Many simulations have used the RF model for electrostatics in the past.[14,22−30] Testing of the accuracy of RF for charged biomolecular systems has also been performed. A study on the thermodynamics of the folding of a short peptide with a net charge of −1 found that a RF simulation produces results similar to those with the PME calculations and experiments:[27] the free energy surfaces derived from both the RF and the PME calculations exhibit the same single minimum, which corresponds to a β-hairpin, the experimentally determined native state of the peptide. The main difference between the PME and the RF results concerned the structures of the less frequently sampled unfolded configurations. The validity of the RF has also been discussed in an extensive examination of the effects of force fields and electrostatics treatments on the secondary structure propensity and the sampling in the peptide folding simulations.[30] It was found that the electrostatics treatment has little effect on the folding characteristics of the peptides, with the PME exhibiting a slightly lower rmsd relative to the native state, but the RF had slightly better sampling. Earlier studies on a highly charged protein crystal,[23] a RNA hairpin in solution,[24] and a small, highly charged globular protein[25] also found that RF produced similar structures[23−25] and conformational spaces sampled[24,25] compared to PME. However, the studies in refs 23−25 involve relatively short trajectories (between 1.5 to 5 ns) and, therefore, do not provide a rigorous test of the use of the RF in representing longer time scale dynamics. In ref 30, it was shown that large conformational changes in proteins in solution can occur on the microsecond time scale, and these might, in principle, be sensitive to electrostatics.

The present paper outlines a strategy for fast and accurate all-atom simulation of multimillion-atom biomolecular systems that do not contain charged groups. The benchmark systems used in the present study are cellulose in water and models of lignocellulosic “biomass”. Lignocellulosic biomass is a complex material composed of crystalline cellulose microfibrils laminated with hemicellulose, pectin, and lignin polymers.[31] In recent years, there has been a revived interest in biomass structure, as biomass offers a potentially abundant and cheap source of sugar for industrial biofuel production.[32] Due to its complexity, lignocellulose poses significant challenges to MD simulation. Among these are the characteristic length scales (Å−μm) and time scales (ns−μs and beyond) of events pertinent to the recalcitrance of biomass to hydrolysis into sugars.[32] To access these length and time scales, standard MD protocols must be modified to scale up to massively parallel machines. Two technical problems are addressed. First, we compare the accuracy of MD using PME and RF on the benchmark systems, and, second we examine the scaling of MD of large systems on a petascale supercomputer.

The present comparative studies show that the examined properties derived using PME are well reproduced using the computationally less demanding method of RF. Scaling benchmarks on multimillion-atom systems show that the use of the RF drastically improves the parallel efficiency of the algorithm relative to PME, yielding ∼30 ns/day. Consequently, microsecond time scale MD of multimillion-atom biomolecular systems appear now within reach.

2. Methods

2.1. Simulation Setup. The simulations were performed using cellulose[33] and lignin[34] force fields parametrized for the CHARMM energy function[35] using GROMACS 4.0.4[4] as the MD software. The validation of the use of the RF is

Table 1. Sets of Benchmark Simulations Performed[a]

simulation index	electrostatic treatment
1	PME with switch
2	PME with cutoff
3	RF
4	shift

[a] Each set comprises five 20 ns trajectories initiated from the same structure but with different initial velocity distribution.

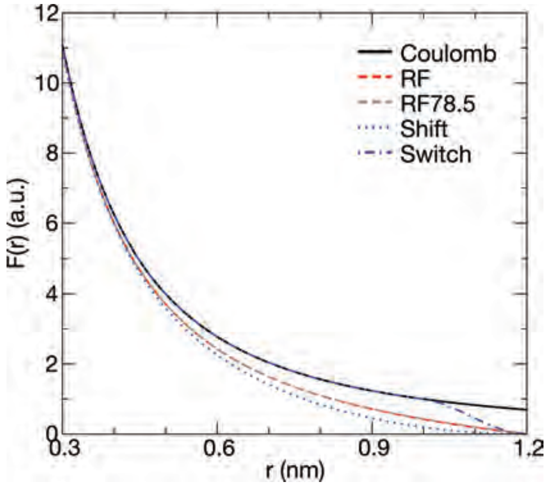

Figure 1. Coulomb force as a function of the distance between opposite charges. Coulomb is the Coulombic force without modification. RF is the reaction field with $\varepsilon = \infty$ outside the cutoff radius. For RF78.5, $\varepsilon = 78.5$ outside the cutoff radius. Shift and switch are computed as described in the user manual of GROMACS.[36] The switch distance after which the Coulomb function is altered is 1nm.

particularly important in the present benchmark systems, since the force fields were not parametrized using the RF method.

The GROMACS simulations were performed with the electrostatic treatments RF, PME-cutoff, PME-switch, and shift and switch (see Table 1). The analytical expression for electrostatic potential, V_{rf} with the RF method is

$$V_{\mathrm{rf}} = \left(1 + \frac{(\varepsilon - 1)r^3}{(2\varepsilon + 1)r_c^3}\right)r^{-1} - 3\frac{\varepsilon}{r_c(2\varepsilon + 1)} \qquad (1)$$

where ε is the dielectric constant outside the radius r_c, and r is the distance separating two charges. In the present work, we employ $\varepsilon = \infty$. This has the advantage that the force is continuous at the cutoff distance, whereas RF with a finite dielectric constant is subject to errors due to force discontinuity. Apart from this significant improvement, the Coulomb forces of the RF with $\varepsilon = \infty$ and 78.5 are almost identical (see Figure 1).

In switch and shift a function S is added to the Coulomb force F_c, giving a total force $F_t = F_c + S$. S is a third-degree polynomial acting over interatomic distances r where $R_1 < r < R_{\mathrm{coul}}$ and is zero otherwise, R_{coul} being the cutoff radius.[36] R_1 is zero for shift and corresponds to the switch-on distance for the switch method (in this study $R_1 = 0.8$ nm and $R_{\mathrm{coul}} = 1.2$ nm, more details on the simulation parameters follow in the next paragraphs). The polynomial is constructed so that $S(r_1) = S'(r_1) = F_t(r_c) = F_t'(r_c) = 0$. The switch, shift, RF, and Coulomb functions are shown in Figure 1. The switch electrostatics was immediately found to produce severe artifacts, including a strong suppression of the fluctuations of the heavy atoms of the cellulose. Consequently the switch simulation was not further considered for detailed analysis. We suspect the switch-induced errors to have been enhanced by the periodicity in the fibril.

In the so-called "atom-based" cutoff, the electrostatic interactions between atoms that are separated by less than the cutoff distance, are explicitly calculated. Some MD codes, such as CHARMM[37] and GROMACS[4] but not NAMD,[3] subdivide molecules into neutral "charge groups". These charge groups are composed of a small number of covalently bonded atoms, such as a methyl group or the HO−C−H moiety of cellulose (see Figure 7). A group-based cutoff can then be defined in which the electrostatic interactions between all atoms belonging to two "charge groups" are calculated explicitly, as long as the geometric centers of the two groups are separated by less than the cutoff distance.

The introduction of the cutoff distance, r_c, is accompanied by a meaningful reduction in the computational cost of the electrostatics only if the list of atoms that are separated by less than the cutoff distance is not calculated at every step. A list of neighbor atoms is created containing all pairs separated by the neighbor list distance, R_{ne}, which is greater than the cutoff distance r_c. The list of neighbors is calculated and updated after n_{list} steps, with n_{list} usually taking values between 10 and 24. The neighbor search distance, R_{ne}, must be large enough to ensure that undesirable events in which atoms jump abruptly in or out of another atom's cutoff sphere do not occur. This would happen, for example, if two atoms are initially separated by $> R_{\mathrm{ne}}$, but after the n_{list} simulation steps move significantly closer, to a distance $< r_c$.

The crystal structure of the I_β allomorph of cellulose[38] was used. This cellulose structure has two chains per triclinic unit cell, which will be referred to as the "origin" and the "center" chains. The shape of the fiber was hexagonal.[39] Figure 2 shows structural details of the model. Details on the models of the lignin molecules are presented elsewhere.[40]

For the simulations in which the effects of varying the electrostatic model were examined, the cellulose contained 80 monomers per chain (36 chains) and was solvated with 70 656 TIP3P[41] water molecules, producing a model totaling 272 556 atoms. A triclinic box was used with a 60° angle between the two short box vectors. The systems were equilibrated for 1ns and then simulated for 20 ns with a time step of 2 fs. For each simulation setup, five simulations with different initial velocities were run. Neighbor searching was performed every 10 steps. All bonds were constrained using LINCS[42] (order: 3, iterations: 2). Temperature coupling was performed with the Nosé−Hoover[43] algorithm (τ = 1ps) during equilibration and the Berendsen[44] algorithm (τ = 0.1ps) during production. Pressure coupling was performed with the Berendsen algorithm (semi-isotropic, τ = 4 ps) during equilibration and the Parrinello−Rahman[45] (isotropic, τ = 4 ps) algorithm during production.

For all lignocellulose simulations with the RF, a dielectric constant of $\varepsilon = \infty$ and a group-based cutoff were employed with the so-called reaction field-zero method, which uses

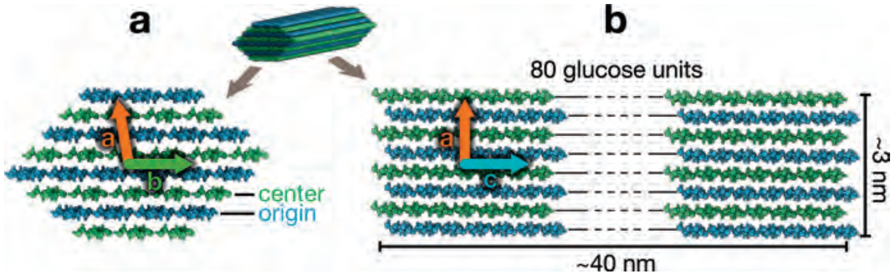

Figure 2. The model of the simulated cellulose fibril showing: (a) the cross-section and (b) a side perspective. The fibril consists of 18 origin chains (blue) and 18 center chains (green). The axes of the unit cell are also indicated.

spline interpolated tables instead of the analytical expression.[36] For the RF, shift, and PME with switch runs a neighbor list-search distance of 1.5 nm, a electrostatic and VDW cutoff of 1.2 nm, and a switch distance of 0.8 nm were used. For PME without switch a neighbor list-search distance of 1.2 nm, a electrostatic cutoff of 1.2 nm, a VDW cutoff of 1.0 nm, and a VDW switch distance of 0.8 nm were used.

In a first analysis step, the simulations were inspected visually. This inspection showed that a strong artifact can arise in the case where only a small buffer region is employed between the cutoff radius and the neighbor list-search distance. To determine the optimal width of the buffer region, a series of simulations was performed varying the width from 0 to 0.3 nm in 0.1nm steps. Simulations using non-PME electrostatics and buffer regionsof <0.3 nm exhibited artificial deterministic linear translation of whole cellulose fibers along their axes with a speed of ~1m/s. Thus, for all further analysis, a buffer region of 0.3 nm was used for non-PME electrostatics.

For the supercomputing performance comparisons, a system was constructed of lignocellulosic biomass containing 52 lignin molecules each with 61 monomers, the same cellulose fibril as described above and 1 037 585 TIP3P water molecules, totaling 3 316 463 (or 3.3 million) atoms. All simulation settings, apart from bond constraints, were the same as the RF settings for cellulose given above. All bonds involving hydrogens were constrained with LINCS[46] (order: 4, iterations: 1). For the sole purpose of extending the scaling tests to a larger system, an additional model system containing 64 000 dipeptide molecules (GLY-PRO) and 1 280 000 water molecules, totaling 5 376 000 atoms, was also constructed. The system was simulated with the same protocol and parameters as the 3.3 million atom lignocellulose system. The detailed system setup is described elsewhere.[47] For the PME simulations in Figure 9, the NAMD multiple time step method was used, in which the long-range electrostatics is computed only every third step and, in addition, a smaller buffer was used.

Topologies were generated in CHARMM[48] and converted using a locally modified version of psfgen[49] (see Supporting Information for details). The correctness of the converted topology and force field was checked by comparison with CHARMM and NAMD (see Supporting Information). All analysis was performed using the tools provided by GROMACS.[50,51] The NAMD trajectories were converted for analysis to GROMACS format and reordered with catdcd.[52] Molecular drawings were made with QuteMol.[53]

2.2. Supercomputer Performance Measurements.

The performance tests were run on JaguarPF, a Cray XT5 massively parallel processing (MPP) computer with over 150 000 Opteron 2.3 GHz cores. JaguarPF has a LINPACK performance of over one petaflop and a SeaStar 2+ interconnect. The internal timings of GROMACS 4.0.4 and NAMD (CVS version) were used. Both CHARM++ and NAMD were built with the Cray-XT4 optimized settings available in the source code. For all tests, all eight cores of each node were used. No special benchmark partition was setup, and as a result, the benchmarks were subject to the regular placement of nodes by the scheduler. The Cray-XT scheduler does not allow consecutive blocks of nodes to be allocated, and the run time varied depending on the node placement. The IO time was included in the benchmarks.

For RF with GROMACS, all parameters were as described in the system setup Section 2.1. For PME with NAMD, a neighbor list-search distance of 1.35 nm, a multiple time step method with a full electrostatic frequency of 3 and a steps-per-cycle of 24, and a grid spacing of 0.13 nm was used (this relatively large spacing was used to ensure good performance by the PME/NAMD simulation). A variation was observed in the speed (in ns/day) of the benchmark runs that used the same number of cores. This variation was caused by the reading input/writing output (IO time) and the task placement.

The IO time was found to be impacted by latency problems caused by Lustre scaling (due possibly to the meta data server). The currently available profiling data do not conclusively identify the relative contribution of Lustre and node placement to the variation of the performance. We have chosen to consider only the best three times (although all 12 times are given in the Supporting Information), as the best three times reflect the optimal performance of the Cray XT5, i.e., when the Lustre system and node placement do not impede the performance of the MD codes. Each MD run was limited to a wall clock time of 10 min. For the thermostat and barostat, the total energy and virial were computed every 10 steps in RF/GROMACS and every 24 steps in PME/NAMD. The calculation of the total energy/virial requires a MPI_Allreduce communication, and therefore, more frequent updates would limit the scaling. For the domain decomposition (DD), the 12 288 cores were arranged in a 3D 96 × 16 × 8 grid. The load balancing works by changing the volume of the DD cells relative to each other. For the minimum DD cell length, 0.77, 0.68, and 0.34 of the average length were used for *X*, *Y*, and *Z* respectively.

3. Results

3.1. Comparison of Simulations with Different Electrostatic Methods. As will be discussed in Section 3.2, fast MD simulation of the 3.3 million atom lignocellulose system can be obtained using the RF method for treating the electrostatic interactions. This section is devoted to examining the accuracy of RF for the biomass test systems. For this, structural and dynamical properties are compared in simulations using different electrostatic methods for both the solvent and the solute. The particular choice of properties for comparison in the case of the solute is based on their structural importance and anticipated sensitivity to possible electrostatic artifacts.

In order to investigate the dependence of dynamical properties on the chosen electrostatics method, the set of MD simulations listed in Table 1 was analyzed. The system of a cellulose fibril in aqueous solution, i.e. without lignin, was chosen for this comparison. Lignin was omitted since significantly longer trajectories would be required for the convergence of dynamical properties due to its amorphous character, thus, complicating the comparison.

Quantities were calculated that were expected to be particularly sensitive to electrostatics. Two functions probing the electrostatic-induced structure and dynamics are the total dipole moment of the fiber and the Kirkwood function between the dipoles of different chains, the latter providing information on the distance-dependent correlation of molecular dipoles. Finally, three specific dihedral angles were selected for comparison due to their structural importance in cellulose.

3.1.1. Solvent. The applicability of the reaction field method for pure water has been discussed in literature (see for example refs 22, 26, and 28). Different implementation schemes exist which can be classified into two groups, namely atom and charge-group based truncation methods.

A suitable function for probing the structural artifacts of water is the Kirkwood factor, $G_k(r)$, which is given by[54]

$$G_k(r) = \sum_{r_{ij}<r} \frac{\vec{\mu}_i \vec{\mu}_j}{|\vec{\mu}|^2} \qquad (2)$$

where $\vec{\mu}_i$ and $\vec{\mu}_j$ are the electric dipole moments of water molecules i and j, respectively. $G_k(r)$ is a measure for the orientational ordering of the dipole moments of the solvent molecules. In earlier work, significant discrepancies were found between PME and RF78.5 (i.e., RF with $\varepsilon_r = 78.5$) in the simulation of bulk water.[26,28] We, therefore, investigated the cause of these discrepancies. To reproduce the earlier results, a simulation setup of bulk water as described in ref 28 was constructed. The resulting Kirkwood factor for water is shown in Figure 3 for four distinct electrostatic treatments with this setup: (a) PME, (b) RF with an infinite dielectric constant ε and group-based cutoff, termed RF in Figure 3, (c) RF with $\varepsilon = 78.5$ and group-based cutoff and (d) RF with $\varepsilon = 78.5$ and an atom-based cutoff. For RF (b) interpolation of tabulated values was used and for RF78.5 (c and d) the analytical expression of RF was used directly.

The RF method with $\varepsilon = \infty$ shows the best overall agreement with PME in terms the of residual difference. For group-based cutoff RF78.5, the Kirkwood function is very different from PME with a deep minimum around the cutoff distance, agreeing with the previous observations.[26,28] We performed several additional simulations (not shown) to find the reason for this difference of $G_k(r)$ for RF78.5. It turns out that the discrepancy arises from the combination of the neighbor-list search with the behavior of the analytical RF expression (eq 1). It is possible to simulate with an atom-based neighbor list, by splitting the water molecule into three charge groups, as proposed originally in ref 22. Using this atom-based cutoff and updating the neighbor list at each step, the disagreement between RF78.5 and PME can be reduced, yielding a general agreement on the shape of the curve. However, in what follows, all lignocellulose simulations were performed using an infinite dielectric constant and a group-based cutoff.

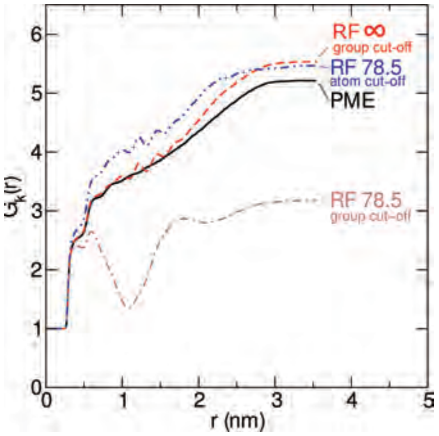

Figure 3. Distance-dependent Kirkwood factor (eq 2) for (a) PME, (b) RF with $\varepsilon = \infty$ and group-based cutoff, (c) RF with $\varepsilon = 78.5$ and group-based cutoff, and (d) $\varepsilon = 78.5$ and atom-based cutoff (where ε is the dielectric constant outside the cutoff radius). RF 78.5 with group-based cutoff produces artifacts that are corrected by using an atom based cutoff and updating the neighbor list at each step.

GROMACS calculates the electrostatic interaction for all atom pairs included in the neighbor list. The distance between atoms for which the electrostatic interaction is calculated can be larger than the cutoff distance in two cases: for a group-based cutoff or a neighbor-list search with frequencies <1/step. In the former case, only the group center needs to be within the cutoff distance for all atoms of the group to be included in the list. In the latter, it is sufficient for the atom to be within the distance at the time of the neighbor-list search, even if it moves outside afterward. The analytical equation (eq 1) of RF78.5 is negative for distances longer than the cutoff distance. Consequently, we conclude that the earlier observed difference between RF and PME[26,28] arises from the negative interaction of atom pairs separated by distances longer than the cutoff distance, caused by the group-based neighbor-list search. Using a spline interpolated table for RF78.5 instead of the analytical expression, as reaction field-zero does by default, allows the interaction for these distances to be set to zero.

3.1.2. Solute/Cellulose. Cellulose was chosen as a benchmark for the solute since it has a high degree of order and,

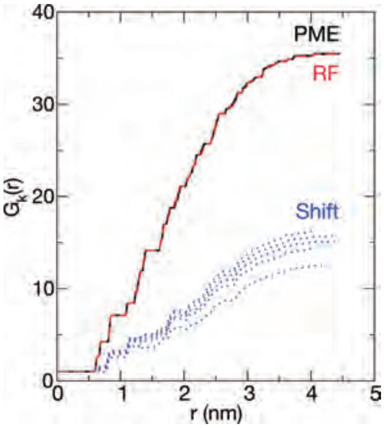

Figure 4. Distance-dependent Kirkwood factor (see eq 2). The two sets performed with PME are shown as indistinguishable black/solid lines. The RF set is red/dashed lines, and the shift set is blue/dotted lines. The profiles of simulations with PME and RF are almost identical, implying very good agreement between the two methods.

therefore, is anticipated to be adequately sampled over the time scales probed.

3.1.3. Dipole Correlation. In the first comparison, shown in Figure 4, the Kirkwood factor, $G_k(r)$, of cellulose is presented. The form of the Kirkwood factor corresponds to that given by Equation 2,[54] where $\vec{\mu}_i$ and $\vec{\mu}_j$ are now the electric dipole moments of glucose chains i and j, respectively, and r_{ij} then becomes the distance between their centers of mass. In this way, $G_k(r)$ is a measure for the orientational ordering of the dipole moments of the cellulose chains in the fibril. It is clear from Figure 4 that the RF method is in very good agreement with the PME method, contrasting with shift in which $G_k(r)$ is much lower. The spread of the shift profiles arises from differences between the individual simulations in the set.

3.1.4. Total Dipole Moment. A further useful test for global changes in dipolar correlation is the total dipole moment for a given macromolecule. Therefore, this should serve as a further benchmark for the accuracy of the electrostatic methods. As seen in Figure 5, the conclusions drawn from this comparison agree with those from the Kirkwood *G* factor: RF and PME show similar features (RF yielding a total dipole moment about 1% lower than PME), whereas the shift method results in a 50% lower magnitude and has slower convergence.

3.1.5. RMSF and RMSD. General dynamical properties examined include the root-mean-squared fluctuations (RMSF) and the modes resulting from principal component analysis (PCA). The RMSF is a measure of the fluctuations of atoms around their equilibrium structure, and PCA provides information on the major collective modes of motion. Both properties are commonly calculated in biomolecular simulations and were, therefore, checked for reproducibility.

Figure 6 shows the difference between the time-averaged RMSF of each atom in the cellulose fibril computed with the RF method minus the RMSF computed with the PME method ($\Delta(\text{RMSF})_{\text{RF}}$). Shown also is the RMSF difference between the shift and PME methods ($\Delta(\text{RMSF})_{\text{shift}}$). The overall good agreement between the RF and the PME is observed once more: the RF enhances fluctuations slightly (with a more pronounced effect for the atomic indices in the range 35 000−40 000), but the shift leads to a much stronger deviation from the PME behavior. In contrast, the rmsd as a function of time, shown in the Supporting Information (Figure S1), shows little difference between the electrostatic treatments. Also in the Supporting Information it is shown that the amplitudes of the eigenvalues obtained from the PCA of the trajectories using the three methods are similar.

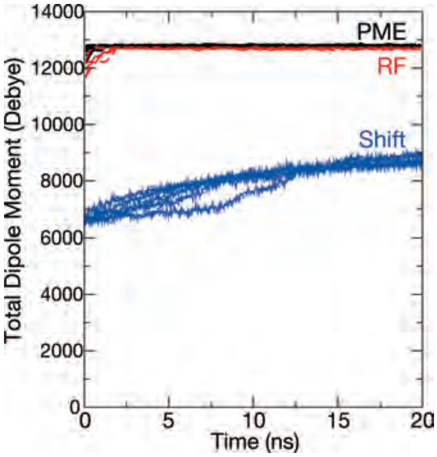

Figure 5. Total dipole moment of the cellulose fibril plotted. The two sets performed with PME are indistinguishable black/solid lines, the RF set in red/dashed lines, and the shift set in blue/dotted lines. The profiles of simulations with PME and RF are almost identical.

3.1.6. Dihedral Angles. The final test focuses on important local structural features of crystalline cellulose. Two sets of dihedrals are examined, as indicated in Figure 7. The particular relation of these dihedrals with respect to cellulose structure is discussed in detail in MD studies of cellulose.[55,56] The ω dihedral (O6−C6−C5−C4) determines the configuration of the primary alcohol group,[55] which affects the hydrogen bonding between adjacent glucose chains within a cellulose fiber and, therefore, is a main determinant for the crystalline phase. When the alcohol lies on the plane of the five-membered glucose ring ($\omega = -60°$ or $\omega = 180°$), single monomers preferentially hydrogen bond to partners within the (010) crystal plane, whereas when the primary alcohol points perpendicular to the five-membered ring plane ($\omega = 60°$) intersheet hydrogen bonds are formed. The Φ and Ψ angles (O5−C1−O1−C4*/C1−O1−C4*−O5*, where * marks atoms on the succeeding monomer) describe the twisting between two consecutive monomers and probe for the frustration in twisting behavior of isolated glucose chains induced by the fiber structure. Unlike the previous properties, these dihedral measures were not necessarily expected to be especially sensitive to differences in electrostatic treatment. They do, however, play an important role in the structure of cellulose. It is, therefore, of interest to determine weather their PMF are not significantly affected by variation of the electrostatic treatment.

The PMFs were calculated according to the equation:

$$W(\theta) = -k_B T \log P(\theta), \qquad \theta = \{\omega, \Phi, \Psi\} \qquad (3)$$

where θ is the dihedral angle in question, and $P(\theta)$ is the associated probability distribution. Since the I_β crystal phase of cellulose has two distinct chains per unit cell, a total of six PMF calculations was performed: for each of the three

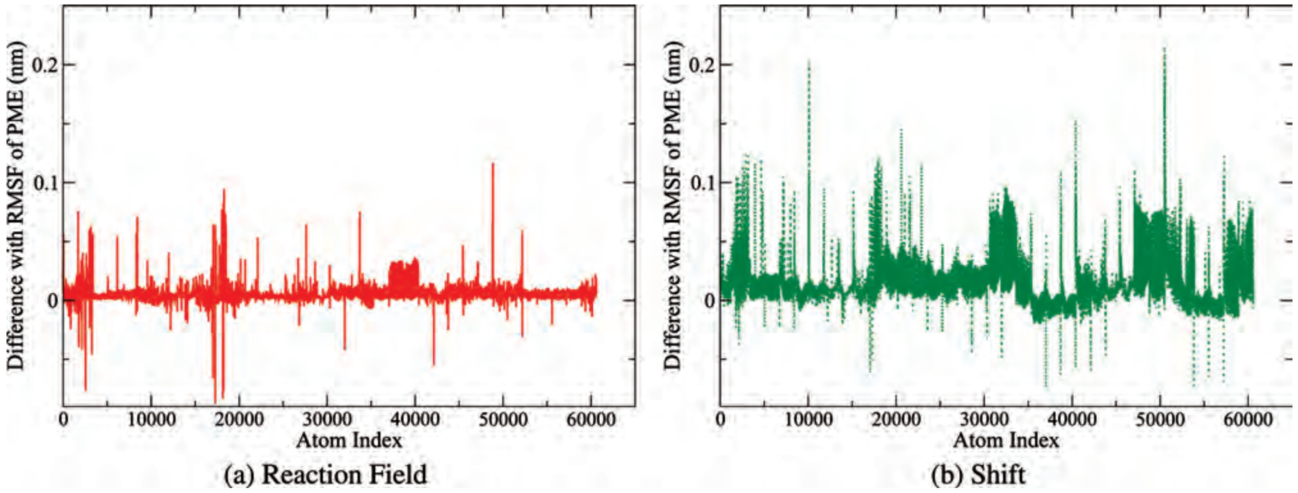

Figure 6. $\Delta(\mathrm{RMSF})_{\mathrm{RF}}$ and $\Delta(\mathrm{RMSF})_{\mathrm{shift}}$ as defined in Section 3.1.2 for all atoms in the cellulose fibril (atomic index on *x*-axis).

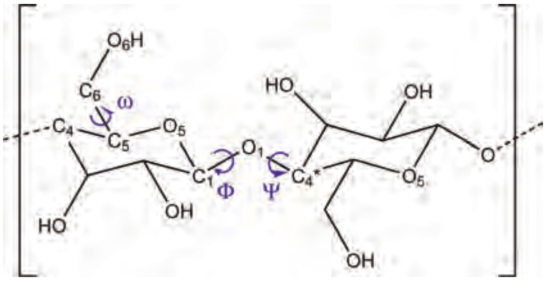

Figure 7. Sketch of cellobiose, the repeating unit of cellulose, indicating the three important dihedrals: the primary alcohol ω dihedral and the Ψ and Φ dihedrals.

dihedrals (ω, Φ, and Ψ), the PMF was calculated for the center and the origin chains. The resulting plots are shown in Figure 8.

The PMFs for the primary alcohol dihedrals follow the same trend as the previous benchmarks, i.e., there is good agreement between the RF and the PME methods but not with the shift method. We note that comparison of the profiles is only meaningful at the relatively low-energy regions that are adequately sampled. In the PMF for the origin chains in 8a, the RF and the PME profiles are almost indistinguishable. However, with the shift method the global minimum moves from 70° to 50°. The difference between the shift and the PME is even more pronounced in the PMF for the center chains (8b), for which the shift introduces a new minimum at −80°, which is only a weak shoulder in the PME calculations.

It is of interest that in the crystal structure of cellulose all primary alcohols have $\omega = -60°$.[38] The transition from $\omega = -60°$ to $\omega = 180°$ observed during the MD simulation is as expected and has been reported in previous MD studies.[55] The origin of the transition is that the force field employed[33] was parametrized for glucose in water and favors the $\omega = 180°$ conformation. Curiously, the shift method appears to "correct" this shortcoming of the force field, and the $\omega = -60°$ conformation is populated in the center chains. However, this effect is probably a cancellation of errors. The present test concerns not the accuracy of the force field with respect to experiment, but rather a comparison between the methods for treating long-range electrostatics.

The PMF for the Φ and Ψ dihedrals shows little variation between the simulations using the three electrostatic treatments and is given in the Supporting Information, Figures S3 and S4.

In summary, the RF electrostatics method is found to accurately reproduce simulation results performed with the widely used PME method. From the set of tests performed, it can be, therefore, inferred that no obvious artifacts are introduced by the proposed RF methodology when applied to the simulation of systems with no net charges.

3.2. Scaling. The parallel efficiency of the RF MD simulation is now evaluated by considering strong scaling. In strong scaling, the system size (i.e., here, the number of atoms in the system) is held constant, while the number of cores used varies. The strong scaling of the 3.3 million atom MD simulation of lignocellulose, using the RF on the ORNL "Jaguar" Cray XT5, is shown in Figure 9. For this system, GROMACS scales well to the 12 288 cores and achieves 27.5 ns/day, running at 16.9TFlops. This performance is made possible by the good scaling of the RF, and a fast particle−particle streaming SIMD (single instruction, multiple data) extensions (SSE) compute kernel running for the lignocellulose system at 4GFlops per opteron core. The RF also improves the parallel efficiency of the MD simulation of even larger systems. Figure 10 shows the strong scaling of a 5.4 million atom peptide solution test system. The same production of 28 ns/day is obtained, this time scaling well to 30k cores.

The load balancing is most critical for the scaling to several thousand cores, as the load per volume of each domain decomposition cell is not equal. The primary cause of load imbalance is the difference in computational speed between the solvent and the solute. The higher speed for the solvent arises from a specially optimized water compute kernel and fewer van der Waals interactions for water. We implemented a new way of performing the dynamical load balancing in GROMACS, detailed in the Supporting Information, improving the average load imbalance from 200 to 75% and leading to an overall 44% improvement of the performance. This improvement resulted in the code obtaining the same production (in ns/day) using half the cores that were used

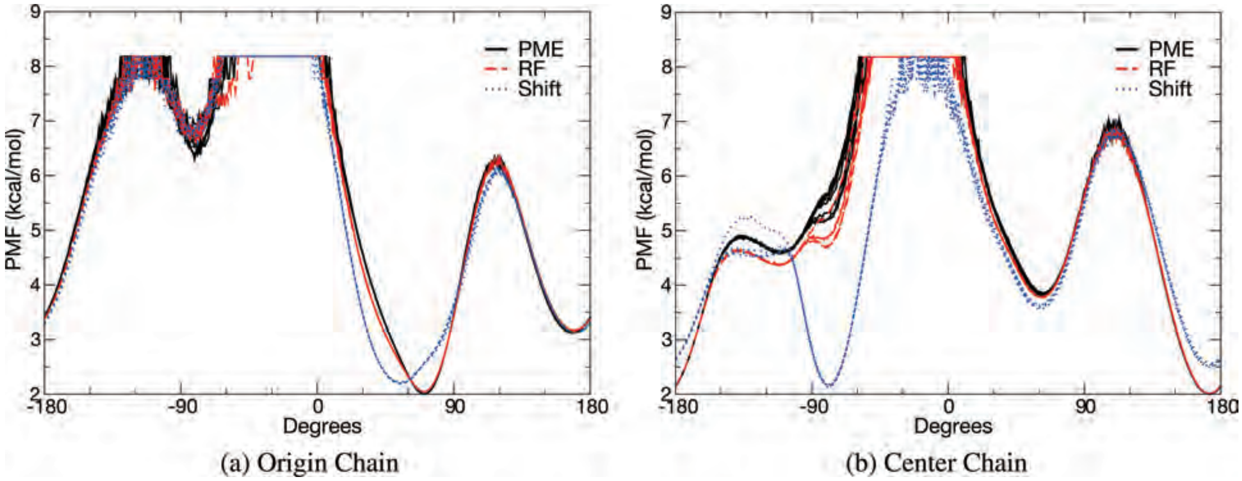

Figure 8. Potentials of mean force for the primary alcohol dihedral ω = O6−C6−C5−C4: (a) results from all 36 origin chains and (b) results from all 36 center chains.

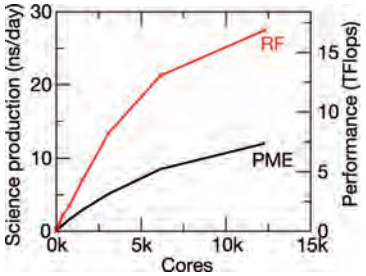

Figure 9. Strong scaling of 3.3 million atom biomass system on Jaguar Cray XT5 with RF. With 12 288 cores the simulation produces 27.5 ns/day and runs at 16.9TFlops. As a comparison, the performance of PME is shown.

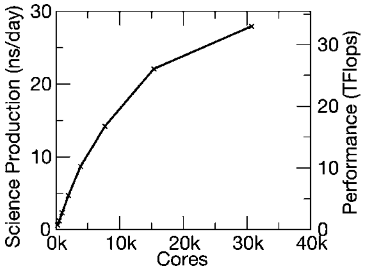

Figure 10. Strong scaling of 5.4 million atom system on Jaguar Cray XT5. With 30 720 cores, 28 ns/day and 33TFlops are achieved.

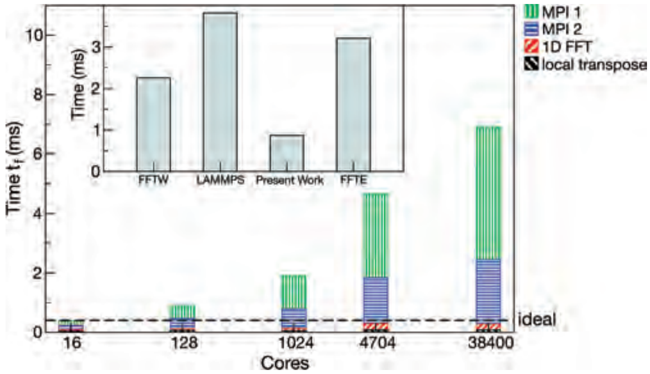

Figure 11. Weak scaling of complex-to-complex FFT on Cray XT5 with FFT implemented as described in Supporting Information, A.3. The 3.3 million atom system requires the 588 × 128 × 128 FFT. The time required to compute one FFT step is represented by t_f.

prior to the improvement. The new implementation of the load balancing is now part of the GIT version of the GROMACS and will also be included in GROMACS 4.1.

To highlight the computational benefit of using the RF, the scaling of a simulation of the 3.3 million lignocellulose system using the PME method is also shown in Figure 9. The PME simulation was run using NAMD, since this MD application is known to have good parallel efficiency.[57] To ensure a "fair" comparison between the two electrostatics methods, some of the parameters of the PME simulation were adjusted to improve its performance (standard 2 fs time step for RF and 6 fs full electrostatics time step and neighbor-list distance update for PME, see Section 2.1 for details). In particular, the reason the RF calculation is faster than the PME at low levels of parallelization is that, on a single processor, the time per step for GROMACS with RF is shorter than for NAMD with PME. However, we stress that the aim of this benchmark is a comparison between the electrostatic treatments and not between the different MD applications. Two different applications were used simply because a direct comparison of simulations using different electrostatics methods with one application is presently not possible: NAMD, which is presently the most scalable code using PME on Cray XT, does not have RF implemented, and GROMACS does not yet have an efficiently scaling PME implemented, with the consequence that PME calculations using GROMACS currently scale up to less than 1 000 cores for large systems (for more details see the Supporting Information).

The significant difference in the parallel efficiency of the PME and the RF electrostatics methods, demonstrated in Figure 9, can be understood by examining the weak scaling of the parallel FFT required for PME, shown in Figure 11. In weak scaling, the ratio of the problem size to the number of cores used in the simulation is held constant. The FFT is a new and improved implementation, the technical details of which are presented in Supporting Information, A.3. The Inset of Figure 11 shows that the new FFT is faster than the FFTs from LAMMPS-FFT,[58] FFTE 4.0,[59] and FFTW 3.2.[60] In ideal weak scaling, the time, t_f, required to perform one

FFT step, indicated by the height of the bars in Figure 11, would remain constant as the number of cores used in the calculation increases from 16 to 38 400. In practice, however, Figure 11 shows that parallel FFT calculations show poor weak scaling, with t_f increasing dramatically on a large number of cores. This increase is a result of the large increase of the required communication time (MPI-1 + MPI-2 in Figure 11). Since, in weak scaling, the number of cores is proportional to the size of the simulated system, Figure 11 demonstrates that the PME method becomes computationally inefficient for large systems.

4. Discussion

This paper presents a strategy for efficient atomistic MD scaling of biological systems on massively parallel supercomputers. The key element of the strategy is to compute the long-range electrostatic interactions with the reaction field (RF) method.

In recent years many simulations have been performed using the PME method. This popularity can be attributed to its high accuracy and fast performance. The most efficient way to perform PME calculations is to balance the work so that one-third to one-fourth of the total CPU-time is spent on the PME part.[4] Hence, removing the PME part improves the overall performance of MD calculations by only ∼25−33%. However, since RF requires a longer neighbor list, as explained in detail previously, the amount of work required for the direct part increases. Since the number of interactions within the neighbor-list distance increases with the volume and, thus, the third power of the radius, the work increases by 2.0 times (or 3.4 times) when increasing the distance from 1.0 to 1.2 nm (or 1.5 nm). This assumes all interactions in the neighbor-list distance are calculated, as is currently implemented in GROMACS, because selecting interactions is expensive on modern CPU architectures. Thus, when comparing PME with a shorter neighbor list to RF with a longer neighbor list (to ensure accurate results), PME is faster on a small number of processors. This picture, however, changes dramatically when MD simulation is run on a large number of processors, where the PME method displays inherent scaling problems.

The scaling of MD codes is restricted by global communications, i.e., instances when all computer nodes exchange information. Although an improvement in the FFT part of PME speed is reported in the Supporting Information, MD simulations using PME still face weak-scaling problems. While for small systems, containing less than 100k atoms, simulations achieving over 100 ns/day are currently possible,[4,6,61,62] for larger systems the global communication for the FFT (MPI_Alltoall) takes longer than the time available for one time step on a large number of cores. This problem worsens as the size of the system increases further, as the MPI_Alltoall global communication does not exhibit efficient weak scaling.

We stress that Figure 9 compares the parallel efficiency of the RF and the PME methods and does not compare different MD applications. As shown in Figures 9 and 10, the use of the RF method greatly improves the strong scaling of million-atom systems, to the point where 28 ns/day are obtained when a 5.4 million atom system is run on 30k cores with a 2 fs time step. Using RF for the electrostatics calculation removes the biggest inherent limitation of the scaling of MD. While all (i.e., irrespective of the method of treating the electrostatics) MD simulations in the NPT ensemble require one global communication (a MPI_Allreduce for a barostat and a global thermostat), this communication is not necessary at each step. The FFT part of PME, however, requires two additional global MPI_Alltoall communications, which take more time than MPI_Allreduce and do not exhibit good weak scaling, see Figure 11. Consequently, the performance, in ns/day, for large systems is inherently limited with PME.

The RF method has been employed in numerous studies,[14,22−30] and there has been some validation of its use on biological systems, including systems with net charges.[14,23−25,27] However, some of these earlier studies involved short (∼5 ns) trajectories, and therefore, it cannot be guaranteed that the RF does not induce a significant bias when dealing with longer time scale dynamics. Further benchmarks are needed to examine the applicability of the RF method for biomolecular systems containing charged groups. Indeed, both the gains in computational efficiency and the possible sources of error arise from the implicit treatment of the Coulomb interaction for atoms separated more than the cutoff distance.

In the present work, the RF method was employed using an infinite dielectric constant outside the cutoff sphere, the benefit of this approach being that the Coulomb force is continuous (and equal to zero) beyond the cutoff distance. In the present study of systems that do not contain net charges, the RF does not appear to compromise the accuracy of MD simulation of the test system under study. This conclusion is drawn after a series of tests in which simulations were performed with different methods for treating the long-range electrostatics interactions. The RF and shift/switch methods are similar in the sense that they do not consider explicitly electrostatic interactions between atoms separated by more than the cutoff distance. Consequently, one might have expected the RF and shift methods to yield similar results. However, our findings suggest a different picture: all benchmarks show very good agreement between RF and PME, while the shift method exhibits several significant artifacts. Also the RF and PME simulations are in very good agreement in tests on bulk water.

In the near future, it is anticipated that the performance of MD using RF might be improved to over 60 ns/day for million-atom systems by using threads and asynchronous communication with neutral territory for improving parallel efficiency. In further benchmarks using the RF together with all-bond constraints and virtual sites, which allow removal of hydrogen atom degrees of freedom enabling integration time steps up to 5 fs,[4] we found that 38 ns/day is obtained when the 5.4-million atom system is run on 15 360 cores (data not shown). Since PME simulations are limited by the time step of the full electrostatics (e.g., 6 fs in the comparison), a longer time step for the short-range interaction does not improve the performance of PME as it would for RF. Thus for large systems, a significant improvement in performance by employing longer time steps is more easily achieved using RF for the electrostatics. However, we stress that carefully designed benchmarks should first be performed before 5 fs time step simulations are routinely applied to biomolecular systems.

This work focuses on supercomputer architectures similar to the Cray XT5. In the future accelerators (GPU, Cell), special purpose and multicore clusters and MPP are likely to become competitive with general purpose supercomputers. For example, recently the special purpose computer Anton was able to simulate a 23k atom protein with a speed of over 10 μs/day[63] using Gaussian split Ewald.[64] Furthermore, accelerators have shown very good performance on MD with small numbers of atoms and processors.[65,66] For the challenging task of using accelerators in highly parallel computers in ref 66, 2 ns/day were obtained for a one million atom system.

Other possibilities for improving the parallel performance of electrostatics are to employ multigrid[64,67−69] or multipole methods.[70] These algorithms are faster for very large systems because the work scales as $O(N)$, compared to $O(N \log N)$ for PME. A larger prefactor, however, can make these methods slower for small systems. A theoretical estimate has predicted the real-space Gaussian split Ewald multigrid method to be faster than the reciprocal FFT method (very similar to PME) for systems larger than ∼30 000 atoms.[64] For the protein ApoA-I (with 92 224 atoms), it has been shown that the multigrid method implemented in PROTO MOL is as fast the PME in NAMD 2.5 (tested up to 66 processors).[68]

Some critical biological phenomena, such as ligand binding and the folding of small proteins, require the simulation of relatively small systems (e.g., $\sim 10^4$ atoms or $\sim 1-10$ nm length scales) for relatively long time scales (e.g., 10^3 s). For this type of application the strategy described here is not applicable. Rather, the present approach permits efficient atomistic MD simulation of larger, multimillion-atom biomolecular systems that do not contain net charges (i.e., on a length scale ∼100 nm) for times of ∼30 ns/day. Using the proposed strategy simulations of these large systems for time scales approaching the microsecond would now seem to be within reach on the Cray XT5. We anticipate that a wealth of structural and dynamical information of biological importance will, thus, be revealed.

Acknowledgment. Graphics were prepared by Thomas Splettstoesser. R.S. thanks the GROMACS developers, especially B. Hess for help with the parallel efficiency and D. van der Spoel for fruitful discussions on RF. We thank Pär Bjelkmar for the CHARMM conversion script. We also extend our thanks to the staff of the ORNL National Center for Computational Sciences for general help and for providing CHARM++ and NAMD binaries. This research is funded by the Genomics: GTL Program, Office of Biological and Environmental Research, U.S. Department of Energy, under the BioEnergy Science Center and FWP ERKP704. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. This research used the resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, supported by the DOE INCITE and Early Access award from the Office of Science.

Supporting Information Available: Correctness of the converted topology and force field that was checked by comparison with CHARMM and NAMD is outlined. The amplitudes of the eigenvalues obtained from the PCA of the trajectories using the three methods are presented. Weak scaling information of the complex-to-complex FFT on Cray XT5 with FFT implemented. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) *Computational Biochemistry and Biophysics*, 1st ed.; Becker, O. M., MacKerell, A. D., Jr., Roux, B., Watanabe, M., Eds.; Marcel-Decker, Inc.: New York,2001.

(2) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.

(3) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(4) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(5) Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1–19.

(6) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. ACM/IEEE Conf. Supercomput.*; New York, 2006.

(7) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; Mcpherson, A.; Schulten, K. *Structure* **2006**, *14*, 437–449.

(8) Sanbonmatsu, K. Y.; Tung, C. S. *J. Struct. Biol.* **2007**, *157*, 470–480.

(9) Zink, M.; Grubmüller, H. *Biophys. J.* **2009**, *96*, 1350–1363.

(10) Mackerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(11) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(12) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(13) Weber, W.; Hunenberger, P. H.; McCammon, J. A. *J. Phys. Chem. B* **2000**, *104*, 3668–3675.

(14) Baumketner, A. *J. Chem. Phys.* **2009**, *130*, 104106+.

(15) Hünenberger, P. H.; Mccammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856–1872.

(16) Villarreal, M. A.; Montich, G. G. *J. Biomol. Struct. Dyn.* **2005**, *23*, 135–142.

(17) Hünenberger, P. *Biophys. Chem.* **1999**, *78*, 69–88.

(18) Gunsteren, W. F.; Berendsen, H. J.; Rullmann, J. A. *Faraday Discuss. Chem. Soc.* **1978**, *66*, 58–70.

(19) Neumann, M. *Mol. Phys.* **1983**, *50*, 841–858.

(20) Neumann, M. *J. Chem. Phys.* **1986**, *85*, 1567–1580.

(21) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(22) Hunenberger, P. H.; van Gunsteren, W. F. *J. Chem. Phys.* **1998**, *108*, 6117–6134.

(23) Walser, R.; Hünenberger, P. H.; van Gunsteren, W. F. *Proteins* **2001**, *43*, 509–519.

(24) Nina, M.; Simonson, T. *J. Phys. Chem. B* **2002**, *106*, 3696–3705.

(25) Gargallo, R.; Hünenberger, P. H.; Avilés, F. X.; Oliva, B. *Protein Sci.* **2003**, *12*, 2161–2172.

(26) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847–10860.

(27) Baumketner, A.; Shea, J. E. *J. Phys. Chem. B* **2005**, *109*, 21322–21328.

(28) van der Spoel, D.; van Maaren, P. J. *J. Chem. Theory Comput.* **2006**, *2*, 1–11.

(29) Gnanakaran, S.; Nussinov, R.; Garcia, A. E. *J. Am. Chem. Soc.* **2006**, *128*, 2158–2159.

(30) Matthes, D.; de Groot, B. L. *Biophys. J.* **2009**, *97*, 599–608.

(31) Cosgrove, D. J. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 850–861.

(32) Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.

(33) Kuttel, M.; Brady, J. W.; Naidoo, K. J. *J. Comput. Chem.* **2002**, *23*, 1236–1243.

(34) Petridis, L.; Smith, J. C. *J. Comput. Chem.* **2009**, *30*, 457–467.

(35) MacKerell, A. D.; et al.*J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(36) van der Spoel, D.; Lindahl, E.; Hess, B.; Kutzner, C.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. *GROMACS USER MANUAL*, Version 4.0; www.GROMACS.org.

(37) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(38) Nishiyama, Y.; Langan, P.; Chanzy, H. *J. Am. Chem. Soc.* **2002**, *124*, 9074–9082.

(39) Ding, S. Y.; Himmel, M. E. *J. Agric. Food. Chem.* **2006**, *54*, 597–606.

(40) Petridis, L.; Xu, J.; Crowley, M. F.; Smith, J. C.; Cheng, X. Atomistic Simulation of Lignocellulosic Biomass and Associated Cellulosomal Protein Complexes. In *Computational Modeling in Lignocellulosic Biofuel Production*; Nimlos, M. R., Crowley, M. F., Eds.; ACS, 2009; p in print.

(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(42) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.

(43) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695+.

(44) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(45) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(46) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(47) McLain, S.; Soper, A.; Daidone, I.; Smith, J.; Watts, A. *Angew. Chem., Int. Ed.* **2008**, *47*, 9059–9062.

(48) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(49) Gullingsrud, J.; Saam, J.; Phillips, J. psfgen; http://www.ks.uiuc.edu/Research/vmd/plugins/psfgen/ 2006 (accessed January 6, 2009).

(50) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.

(51) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(52) Gullingsrud, J. catdcd; http://www.ks.uiuc.edu/Development/MDTools/catdcd/ (accessed January 6, 2009).

(53) Tarini, M.; Cignoni, P.; Montani, C. *IEEE Trans. Visual Comput. Graphics* **2006**, *12*, 1237–1244.

(54) Oster, G.; Kirkwood, J. G. *J. Chem. Phys.* **1943**, *11*, 175–178.

(55) Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2006**, *341*, 138–152.

(56) French, A. D.; Johnson, G. P. *Cellulose* **2004**, *11*, 449–462.

(57) Schulten, K.; Phillips, J. C.; Kal Bhatele., A. Biomolecular modeling in the era of petascale computing. In *Petascale Computing: Algorithms and Applications*; Bader, D., Ed.; Chapman and Hall/CRC Press, Taylor and Francis Group: Boca Raton, FL, 2008; pp 165−181.

(58) Plimpton, S. Parallel FFT Package; http://www.sandia.gov/∼ sjplimp/docs/fft/README.html (accessed January 6, 2009).

(59) Takahashi, D. FFTE: A Fast Fourier Transform Package; http://www.ffte.jp (accessed January 6, 2009).

(60) Frigo, M.; Johnson, S. *Proc. IEEE* **2005**, *93*, 216–231.

(61) Fitch, B. G.; Rayshubskiy, A.; Eleftheriou, M.; Ward, C. T. J.; Giampapa, M.; Pitman, M. C.; Germain, R. S. Blue matter: approaching the limits of concurrency for classical molecular dynamics. *Proc. ACM/IEEE Conf. Supercomput.*; New York, 2006.

(62) Freddolino, P. L.; Liu, F.; Gruebele, M. H.; Schulten, K. *Biophys. J.* **2008**, *94*, 75−77.

(63) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(64) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *J. Chem. Phys.* **2005**, *122*, 054101.

(65) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. *J. Chem. Theory Comput.* **2009**, *5*, 1632–1639.

(66) Phillips, J. C.; Stone, J. E.; Schulten, K. Adapting a message-driven parallel application to GPU-accelerated clusters. *Proc. ACM/IEEE Conf. Supercomput.*; Piscataway, NJ, 2008; pp 1−9.

(67) Sagui, C.; Darden, T. *J. Chem. Phys.* **2001**, *114*, 6578–6591.

(68) Izaguirre, J. A.; Hampton, S. S.; Matthey, T. *J. Parallel Distr. Com.* **2005**, *65*, 949–962.

(69) Skeel, R. D.; Tezcan, I.; Hardy, D. J. *J. Comput. Chem.* **2002**, *23*, 673–684.

(70) Kurzak, J.; Pettitt, B. M. *Mol. Simulat.* **2006**, *32*, 775–790.

CT900292R