Published in final edited form as:
*J Chem Theory Comput*. 2012 July 10; 8(7): 2521–2530. doi:10.1021/ct3003089.

# GPU/CPU Algorithm for Generalized Born/Solvent-Accessible Surface Area Implicit Solvent Calculations

**David E. Tanner**[†,‡], **James C. Phillips**[‡], and **Klaus Schulten**[*,†,¶,‡]

[†]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign

[‡]Beckman Institute, University of Illinois at Urbana-Champaign

[¶]Department of Physics, University of Illinois at Urbana-Champaign

## Abstract

Molecular dynamics methodologies comprise a vital research tool for structural biology. Molecular dynamics has benefited from technological advances in computing, such as multi-core CPUs and graphics processing units (GPUs), but harnessing the full power of hybrid GPU/CPU computers remains difficult. The generalized Born/solvent-accessible surface area implicit solvent model (GB/SA) stands to benefit from hybrid GPU/CPU computers, employing the GPU for the GB calculation and the CPU for the SA calculation. Here, we explore the computational challenges facing GB/SA calculations on hybrid GPU/CPU computers and demonstrate how NAMD, a parallel molecular dynamics program, is able to efficiently utilize GPUs and CPUs simultaneously for fast GB/SA simulations. The hybrid computation principles demonstrated here are generally applicable to parallel applications employing hybrid GPU/CPU calculations.

## Introduction

Laboratory and clinical studies of living cells are increasingly complemented by computational atomic level modeling of cellular processes.[1-8] In fact, modeling has developed today into a computational microscope[9] used to explore nanoscale structures and processes in structural cell biology[10-12], cellular mechanics,[9,13,14] and nanosensor development.[15,16] Small size and short time scales of many sub-cellular processes challenge experimental methodologies, making the molecular dynamics computational microscope an ideal supporting tool for biological research.

Molecular dynamics (MD) has already enabled, in particular, modeling of health-relevant biomolecular systems and processes, e.g., viral infection,[17,18] interactions between tissues and therapeutics,[1,2] blood coagulation,[13,19-25] and amyloid fibril formation.[26-31] Often, however, simulation time scales are too short to be of value; only by continually adopting the latest computing technologies can simulation speeds increase, and extend the reach of MD to the millisecond scale[32] that is necessary to describe many cellular processes.

MD simulations calculate interatomic forces of biopolymer systems and solve the classical equations of motion for each time step,[33] to explore the dynamic behavior of biological systems. The largest computational expense in MD simulations stems from calculating so-called "non-bonded" interactions, forces between atoms which are not covalently bonded, comprising Coulomb and van der Waals forces.

[*] To whom correspondence should be addressed: kschulte@ks.uiuc.edu.

The calculation of non-bonded forces is highly amenable to parallel computing[33,34] because the modeled interaction between two atoms is independent of all other atoms; thus, atom-pair interactions can be calculated independently from one another, then accumulated to determine net atomic forces. Such independent calculations offer a high degree of concurrency, making them ideal for parallel computing.[35] Indeed, NAMD successfully conducted recently a 100,000,000-atom MD simulation on 200,000 CPU cores.[36] Along with multi-core CPUs, MD programs have also harnessed new computing technologies such as graphics processing units (GPUs).[37,38]

## Graphics processing units

Though originally developed only for interactive graphics rendering, GPUs are well suited for accelerating general scientific calculations. A single GPU may today contain 16 multiprocessors, each with 32 cores, yielding 512 cores per GPU. Maximizing the use of such highly parallel processors requires meeting strict criteria such as coalescing access to slow memory and issuing 10,000s of independent calculations. Because many molecular biological calculations can meet the strict criteria, GPUs have been used to accelerate biological computing applications such as molecular modeling,[39] electrostatic calculations,[37,40] molecular orbital display[41] and simulations of protein diffusion in whole cells.[42]

The application of GPUs to molecular dynamics[43] has already demonstrated significant performance benefits[44] for both explicit solvent[38,45] and implicit solvent[46,47] models. Each implementation varies in the use of the GPU, ranging from employing the GPUs for calculating forces only[38] to using them to integrate the equations of motion[48,49] as well. GPU implementations also vary in regard to how restrictive they are of the MD methodologies they admit; for example, some implementations allow a non-bonded interaction cut-off length[37] which is common for large systems, while others do not,[46] thereby being applicable only to small systems.

The great success already achieved in accelerating full molecular dynamics calculations with GPUs inspires additional work in advancing MD calculations which involve both GPU-accelerated calculations and CPU calculations (those not yet adapted to the GPU) as they arise in, for example, steering[50] or grid potentials.[51] Prior efforts at GPU-acceleration of MD simulations had adapted almost all of the computation to the GPU, the CPU’s role remaining secondary.[48,49]

## Hybrid GPU/CPU computing

Each GPU requires a so-called host CPU to send data to and receive data from the GPU as well as instruct the GPU as to which calculations to perform. Therefore, a necessary and difficult component of hybrid GPU/CPU calculations is allowing host CPUs to switch between GPU-hosting responsibilities and performing their own scientific calculations. Because many existing MD methods have not yet been adapted to GPUs, or are not amenable to GPU calculation, it is vital to explore how to efficiently perform hybrid calculations which require full use of both GPUs and CPUs and which, therefore, require coordination between the two.

An ideal application for hybrid computing is the generalized Born (GB) implicit solvent with solvent-accessible surface area (SASA or SA) calculation, commonly abbreviated GB/SA. A GB/SA implicit solvent simulation employs the GB calculation to account for the polar, i.e., hydrophilic, effects of water while the SA calculation models the nonpolar, i.e., hydrophobic, effects.[52] It is known that the hydrophobic free energy of solvation is approximately equal to the product of the molecular surface area and a surface tension parameter.[53,54] The GB/SA calculation consists of three classes of non-bonded interatomic forces: the classical Coulomb and van der Waals forces, the generalized Born[53,55,56] (GB) force, and the solvent-accessible surface area (SA) force, calculated via the linear combination of pairwise overlaps[54] (LCPO) algorithm, as in the Amber MD program.[57] While the Coulomb, van der Waals and GB force calculations are being calculated on the GPU, the LCPO algorithm's SA calculation is best performed on the CPU, thus making the GB/SA calculation an ideal candidate for hybrid GPU/CPU calculation.

The present work explores a variety of performance and functionality issues relevant to GPU accelerated calculations of the GB/SA implicit solvent model in the context of the NAMD parallel MD program.[33] First, we analyze the necessary non-bonded force calculations and describe why each is best suited for GPUs or CPUs. Second, we outline NAMD's algorithmic strategy for GB/SA calculations on hybrid GPU/CPU computers. Finally, we explore the effect of the surface tension parameter as well as demonstrate the performance of the hybrid strategy through ~250 benchmark simulations.

## Methods

The generalized Born/solvent-accessible surface area (GB/SA) implicit solvent model constitutes a fast representation of a solvent's polar and nonpolar effects on biomolecules.[53] Fast simulation speed can be attained by calculating the Coulomb, van der Waals and generalized Born (GB) forces on the GPU while simultaneously calculating the hydrophobic surface area (SA) force on the CPU. First, we present the equations arising in the Coulomb, van der Waals, GB and SA force calculations which motivate the hybrid GPU/CPU algorithm employed. Then, NAMD's implementation of the various calculations will be described.

### Coulomb and van der Waals forces

The non-bonded forces arising in explicit and implicit solvent simulations are the Coulomb and van der Waals (calculated using the Lennard-Jones description) forces. The total system energy contributed by Coulomb and van der Waals interactions, $E_{\mathrm{T}}^{\mathrm{CW}}$, is

$$E_{\mathrm{T}}^{\mathrm{CW}} = (1/2)\sum_{i}\sum_{j\in N(i)}\{\underbrace{4\varepsilon_{ij}\left[\left(\sigma_{ij}/r_{ij}\right)^{12}-\left(\sigma_{ij}/r_{ij}\right)^{6}\right]}_{\text{van der Waals}}+\underbrace{\left(k_{\mathrm{e}}/\varepsilon_{\mathrm{p}}\right)\left(q_i q_j/r_{ij}\right)}_{\text{Coulomb}}\}\ , \tag{1}$$

where $\varepsilon_{ij}$ is the well depth and $2^{1/6}\sigma_{ij}$ equilibrium interaction length parameters of the Lennard-Jones potential, $r_{ij}$ is the distance between atoms $i$ and $j$, $k_{\mathrm{e}}$ = 332 kcal Å/e$^2$ the Coulomb constant, $\varepsilon_{\mathrm{p}}$ = 1 the dielectric constant of the protein interior and $q_i$ the atomic charge; $N(i)$ is the set of all neighbors, $j$, that are within the interaction cut-off, $r_c$, from atom $i$.

The net force, $\vec{F}_i$, on an atom is calculated as

$$\vec{F}_i = -\sum_{j\in N(i)}\left(dE_T/dr_{ij}\right)\widehat{r}_{ij}; \tag{2}$$

by applying Eq. (2) to Eq. (1), the net Coulomb and van der Waals forces on an atom, $\vec{F}_i^{\mathrm{CW}}$, is

$$\vec{F}_i^{\mathrm{CW}} = \sum_{j\in N(i)} \{\underbrace{24\varepsilon_{ij}\left[2\left(\sigma_{ij}^{12}/r_{ij}^{13}\right)-\left(\sigma_{ij}^{6}/r_{ij}^{7}\right)\right]}_{\text{van der Waals}} + \underbrace{\left(k_{\mathrm{e}}/\varepsilon_{\mathrm{p}}\right)\left(q_i q_j/r_{ij}^2\right)}_{\text{Coulomb}}\}\hat{r}_{ij}\,. \tag{3}$$

The summation in Eq. (3) requires an MD program to iterate over all pairs of interacting atoms, *i* and *j*, where $r_{ij} < r_c$; the successful application of GPUs to computing atom-pair interactions, such as Coulomb and van der Waals forces, has previously been demonstrated in NAMD.[37,38]

## Generalized Born implicit solvent forces

The generalized Born implicit solvent model,[53,56] already implemented in NAMD for the CPU,[58] describes water as a bulk solvent acting as a dielectric;[53,56] the dielectric solvent screens,[59,60] i.e., reduces, electrostatic interactions between charged atoms. The total GB energy, i.e., hydrophilic energy of solvation, of the atomic system is given by the sum of pair- and self-energies according to

$$E_{\mathrm{T}}^{\mathrm{GB}} = (1/2)\sum_i \sum_{j\in N(i)} \underbrace{E_{ij}^{\mathrm{GB}}}_{\text{pair}} + (1/2)\sum_i \underbrace{E_{ii}^{\mathrm{GB}}}_{\text{self}}\,, \tag{4}$$

where the pair- and self-GB energies are calculated according to

$$E_{ij}^{GB} = -k_e D_{ij} q_i q_j / f_{ij}^{GB}. \tag{5}$$

Here, $D_{ij}$ is an effective dielectric constant, which depends on the implicit ion concentration,[58,61] between atoms *i* and *j*, and $f_{ij}^{\mathrm{GB}}$ is[53]

$$f_{ij}^{GB} = \sqrt{r_{ij}^2 + \alpha_i\alpha_j \exp\left(-r_{ij}^2/4\alpha_i\alpha_j\right)}. \tag{6}$$

The quantities $\alpha_i$ arising in this expression, the so-called atomic Born radii, describe an atom's exposure to solvent and, thus, characterize the degree to which an atom's electrostatic interaction is screened; $\alpha_i$ is calculated, according to Onufriev, Bashford and Case's (OBC) description,[56] as

$$\alpha_i = \left[(\rho_i - 0.09)^{-1} - (1/\rho_i)\tanh\left(\delta\psi_i - \beta\psi_i^2 + \gamma\psi_i^3\right)\right]^{-1}, \tag{7}$$

where $\rho_i$ is the atomic radius as parameterized by the OBC model; $\delta = 1$, $\beta = 0.8$ and $\gamma = 4.85$ are additional dimensionless parameters of the OBC model[56] which enable calculation of the correct Born radii, i.e., those derived from Poisson-Boltzmann calculations, from the sum, $\psi_i$, of pairwise atomic descreening, $H_{ij}$,

$$\psi_i = (\rho_i - 0.09)\sum_{j\in N(i)} H_{ij}. \tag{8}$$

The pairwise descreening, $H_ij$, between neighboring atoms $i$ and $j$ is given by a distance-dependent piecewise function, defined in four mutually exclusive interaction domains,

$$H_{ij}=\begin{cases} 0 & \left(r_{ij}>r_c+\rho_j\right) \\ f_1\left(r_{ij},\rho_j,r_c\right) & \left(r_{ij}>r_c-\rho_j\right) \\ f_2\left(r_{ij},\rho_j\right) & \left(r_{ij}>4\rho_j\right) \\ f_3\left(r_{ij},\rho_j\right) & \left(r_{ij}>\rho_i-0.09+\rho_j\right) \\ f_4\left(r_{ij},\rho_i,\rho_j\right) & \text{otherwise} \end{cases} \tag{9}$$

where $f_{1–4}$ are also taken from the OBC model.[56]

Because the GB energy defined in Eq. (4) depends on the interatomic distances directly, through $f_{ij}^{\mathrm{GB}}$, and also indirectly, through $\alpha_i$, calculating the GB force on an atom, $\vec{F}_i^{GB}$, requires multiple partial derivatives,[58] namely,

$$\vec{F}_i^{GB}=\sum_{j\in N(i)}\left[\left(\partial E_{ij}^{GB}/\partial f_{ij}^{GB}\right)\left(df_{ij}^{GB}/dr_{ij}\right)+\left(\partial E_T^{GB}/\partial\alpha_i\right)\left(d\alpha_i/dr_{ij}\right)+\left(\partial E_T^{GB}/\partial\alpha_j\right)\left(d\alpha_j/dr_{ij}\right)\right]\widehat{r}_{ij}. \tag{10}$$

Of the various required derivatives, $\partial E_{ij}^{GB}/\partial f_{ij}^{\mathrm{GB}}$, $\partial E_T^{GB}/\partial\alpha_k$ and $d\alpha_k/dr_{ij}$, the most expensive to calculate is $\partial E_T^{GB}/\partial\alpha_k$, as it requires an additional summation over atom-pairs,

$$\partial E_T^{GB}/\partial\alpha_k=(1/2)\sum_i\sum_{j\in N(i)}\left[\partial E_{ik}^{GB}/\partial\alpha_k+\partial E_{kj}^{GB}/\partial\alpha_k\right]+\sum_i\partial E_{ii}^{GB}/\partial\alpha_k. \tag{11}$$

The three summations in Eq. (8), Eq. (10) and Eq. (11), known as the three phases of the GB force calculation, require three successive iterations over all pairs of atoms each time step. The three phases compute, in order, the atomic Born radii, $\alpha_i$, the partial derivatives $\partial E_T^{GB}/\partial\alpha_k$, and, finally, the GB force, $\vec{F}_i^{GB}$.

As in the case of Coulomb and van der Waals forces, iterating over atom-pairs for the three GB phases is amenable to GPU-acceleration. Because calculating the generalized Born force as outlined in Eq. (4)-Eq. (11) is several times more computationally expensive than calculating only the Coulomb and van der Waals forces, and because GPUs are well suited for such arithmetically expensive calculations, the GB calculation stands to benefit strongly from GPU-acceleration as previously demonstrated by other MD programs.[46,47] Even for the case that the computational expense of an implicit solvent simulation is greater than that of an all-atom explicit solvent simulation, because conformational relaxation processes are usually faster in implicit solvent,[58] such a model can still yield an overall benefit to conformational sampling.

## Solvent-accessible surface area forces

The generalized Born model only describes the polar, i.e., hydrophilic, energy of solvation. It is desirable to account also for the nonpolar, i.e., hydrophobic, energy of solvation through a solvent-accessible surface area (SA) calculation, as it is known that the hydrophobic solvation energy is approximately proportional to SA.[53,62] We note here that the hydrophobic energy contributes to every surface element of a protein, even those involving charged side groups; this seemingly counterintuitive issue is discussed below. The linear combination of pairwise overlaps (LCPO) is an approximate method for calculating the SA[54] and, in particular, the spatial derivatives necessary for hydrophobic force calculation.

The LCPO method,[54] which considers only non-hydrogen atoms, is founded on calculating the surface area overlap between two spheres representing atoms; for two spheres, centered on atoms *i* and *j*, with radii $R_i$ and $Rj$, separated by a distance $r_ij$, close enough that their surfaces overlap, i.e., it must hold $r_ij < R_i + Rj$, the surface area of atom *i* overlapped by atom *j*, $A_{ij}$[54]

$$A_{ij}=2\pi R_i\left[\left(R_i-\left(r_{ij}/2\right)-\left(R_i^2-R_j^2\right)\right)/\left(2r_{ij}\right)\right],\quad (12)$$

where the radius, $R_i$, is the atomic van der Waals radius plus the 1.4 Å solvent probe radius.[54] According to the LCPO method, the surface area of atom *i*, SA$_i$, can be evaluated approximately by multiple overlap summations,

$$SA_i=p_{1,i}4\pi R_i^2+P_{2,i}\sum_{j\in N(i)}A_{ij}+P_{3,i}\sum_{j\in N(i)}A_{ij}\sum_{k\in N(i)\cap N(j)}A_{jk}+P_{4,i}\sum_{j\in N(i)}\left[A_{ij}\sum_{j\in N(i)}\sum_{k\in N(i)\cap N(j)}A_{jk}\right],$$

where $k \in N(i) \cap N(j)$ represents all atoms *k* which overlap atoms *i* and *j*, thus requiring the LCPO algorithm to iterate over atom triplets, sets of three atoms whose surfaces overlap (i.e., it must hold $r_{ij} < R_i + R_j$, $r_{ik} < R_i + R_k$, $r_{jk} < R_j + R_k$). The total molecular surface area is the sum of atomic surface areas, $\sum_i \mathrm{SA}_i$. The four atomic parameters $P_{1-4i}$ were fitted by atom type such that the LCPO algorithm most closely approximates the correct surface area as calculated by numerical methods.[54] The hydrophobic, surface area force on atom *l*, $\overrightarrow{F}_l^{\mathrm{SA}}$, can then be calculated employing the LCPO algorithm,

$$\overrightarrow{F}_l^{SA}=-T_S\sum_{i\in N(l)}\left(dSA_i/dr_l\right),\quad (13)$$

where $T_S$, with units kcal/mol/Å$^2$, is the surface tension parameter. The relatively inexpensive derivatives required for the SA force calculation within the LCPO algorithm were previously described.[54]

For the above three force calculations, approximately 13% of the computational cost is due to Coulomb and van der Waals forces, 70% due to GB forces, and the remaining 17% due to SA forces. Having transferred most of the force calculation (Coulomb, van der Waals, GB) to the GPU, the CPU can perform the relatively inexpensive (17% of total) SA calculation in the same time as the GPU's expensive (83% of total) force calculations. The SA algorithm requires iterating over all atom triplets, compared to only pairs of atoms as in the case of GB; the summation is feasible due to the low computational cost involved and because the summation over triplets can be performed well by CPUs.

## Incorporating the GB/SA calculation into NAMD

To achieve a high performance implementation of the GB/SA model, formulated by Eq. (1)-Eq. (13), we thought to optimize use of hybrid GPU/CPU computers as outlined. Aside from fast performance, an ideal implementation should interfere as little as possible with NAMD's already highly scalable internal structure.

To achieve advanced parallel scaling, NAMD decomposes a simulated system (Figure 1A) into a 3D grid of atom-groups (Figure 1B). In the case of Coulomb, van der Waals and GB forces, atom-groups are defined such that atoms in any atom-group interact only with atoms of the same atom-group and with atoms of adjacent atom-groups. Calculating the atom-pair interactions between two neighboring atom-groups (Figure 1D red and green) constitutes an independent force work unit. To illustrate our GB/SA implementation, we apply NAMD to the 13,340-atom glycogen phosphorylase protein, Protein Data Bank (PDB) ID 1GPB with missing atoms placed by VMD's[63] psfgen tool. The atoms of this protein system can be divided into roughly 100 atom-groups, shown in Figure 1B, and 3,000 force work units which can be assigned across hundreds of CPU cores or GPUs.[37,38]

## GB force calculation on GPUs

NAMD has already achieved a sixfold speed-up on its 92,000-atom ApoA1 benchmark system by calculating Coulomb and van der Waals forces on the GPU using the CUDA language.[37,38] Because of the success of the Coulomb and van der Waals calculations on the GPU and the close algorithmic similarities, NAMD's GB calculation on the GPU, outlined below, employs a similar GPU implementation as NAMD's previously reported Coulomb and van der Waals force calculation on the GPU.[37,38]

For parallel calculations, each processor core requires a list of instructions to drive computation; for both GPU and CPU processors, a "thread" is the list of instructions detailing what operations a processor core must perform; multi-core CPUs and GPUs both require multiple threads, i.e., multiple lists of instructions, to drive the many processor cores. In CUDA, a thread block is the group of threads which drives one of the GPU's many 32-core multiprocessors.

In NAMD, a force work unit consists of calculating the atomic interactions between two atom-groups, such as group-I (red) and group-J (green) of Figure 1B and D. The calculation of the force work unit is assigned to a thread block, i.e., is calculated by a single 32-core multiprocessor, as depicted in Figure 1D. Each thread (vertical line) in the thread block calculates the force on a single atom of group-I (red) due to its interactions (blue circle) with atoms of group-J (green). To accelerate memory access, the GPU's 32-core multiprocessors access group-J atoms 32 at a time, in order 1-3 shown in Figure 1D, and intermediately store the data in "shared" memory, a fast memory shared only by the 32 cores of a multiprocessor.

The above pattern of NAMD's GPU calculation of Coulomb, van der Waals and GB forces exploits several features of GPU hardware to achieve fast performance. First, because adjacent threads operate on atoms adjacent in memory, reading atomic coordinates from and writing atomic forces to slow memory, is coalesced, giving significant memory speed-up over non-coalesced memory access. Second, because GPUs operate fastest when threads on a multiprocessor perform the same calculation, NAMD pre-sorts atomic coordinates before transferring them to the GPU to reduce the likelihood that threads on a multiprocessor must evaluate differing pairwise functions, $f_{1-4}$ of Eq. (9), as doing so dramatically reduces performance. Much work has previously been done exploring implementation patterns for optimal GPU performance of molecular modeling applications.[37-39,64,65]

## SA force calculation on CPUs

Because Eq. (13) requires iteration over all atom triplets, the LCPO calculation cannot be carried out using the structure of atom-group pairs which NAMD employs for Coulomb, van der Waals and GB force calculation, as it would be possible for a triplet of three overlapping atoms to belong to three neighboring atom-groups, in which case no atom-group pair would evaluate the triplet. Therefore, an LCPO force work unit instead consists of a 2 × 2 × 2 set of 8 adjacent atom-groups as illustrated in Figure 1C. Each LCPO force work unit calculates the surface area and hydrophobic force on the inner 1/8th fraction of atoms (Figure 1C red) due to surface area overlaps with neighboring atoms (red and blue). LCPO force work units can also be partitioned such that the force work unit in Figure 1C is duplicated with each copy calculating forces for only a fraction of the inner core atoms (red). The ability to partition the SA calculation into many small force work units will be shown to be critical to the efficiency of the hybrid GPU/CPU algorithm.

### Balancing the GPU and CPU calculations

With fast algorithms in place to perform the GB force calculation on the GPU and the SA force calculation on the CPU, the remaining obstacle of combining the two into a hybrid GB/SA calculation requires coordination of the GPU and CPU computation. First, the right balance of GPUs and CPU cores will allow the Coulomb, van der Waals and GB calculation on the GPU and the SA calculation on the CPU to be executed in the same amount of time. Second, partitioning the SA calculation into small force work units will allow the host CPU to switch between SA calculations and GPU interaction with high frequency, thereby minimizing the delay of either calculation.

Because GPUs and CPUs possess different computational power and are responsible for calculating workloads of different size, a computationally efficient simulation requires that one must first determine, through benchmarking the particular system to be simulated, an optimal ratio of CPU cores per GPU that balance the durations of GB and SA calculations. While the thorough benchmarking of our implementation provided below (see Table 2) will aid a user in judging performance a priori, the user may want to verify in case of any particular calculation if a chosen ratio of GPUs to CPU cores yields efficient performance, for example by varying the ratio and judging the resulting overall performance. In order to achieve optimal performance, it must also verified that CPU and GPU calculations overlap; the corresponding analysis will be carried out below.

### Simulations carried out

To analyze the performance of NAMD’s hybrid GB/SA algorithm, four test systems, differing in size, were simulated on 0-4 GPUs and 1-32 CPU cores. Figure 2 depicts the four tested systems and Table 1 lists their SA values; atoms missing in the public PDB structure files were placed by VMD’s[63] psfgen tool.

For benchmarking, three types of simulations were performed, each with increasing computational cost and solvent accuracy. In vacuo simulations evaluate only the Coulomb and van der Waals non-bonded forces as defined through Eq. (1)-Eq. (3), GB simulations additionally evaluate the generalized Born forces, determined through Eq. (4)-Eq. (11), and GB/SA simulations further include the SA calculation following the LCPO algorithm as stated through Eq. (12)-Eq. (13).

The following simulation parameters were employed for all simulations. A value of 16 Å was employed for the Coulomb and van der Waals interaction cut-off as well as for the GB phase 2 calculation, c.f. Eq. (11), while the GB phase 1 and 3 calculations, as defined through Eq. (8) and Eq. (10), were cut off at 14 Å. For GB and GB/SA simulations, an implicit ion concentration of 0.3 M was assumed.[61] The GB/SA simulations employed a surface tension of 0.005 kcal/mol/Å$^2$,[62] unless otherwise specified. A time step of 2 fs was employed, requiring the SHAKE[66] algorithm to restrain covalent bonds to hydrogen atoms, with all forces being evaluated every time step.

Benchmark simulations were run for 620 time steps on a 2.2 GHz 48-core AMD Opteron 6174 computer accelerated with an NVIDIA S2070 GPU system. The first 20 steps of simulation perform conjugate gradient minimization; the next 500 steps consisted of NAMD load balancing to optimize parallel performance; the simulation speed, in units seconds/time step, was averaged over the final 100 steps. Results of the benchmark simulations are presented in Figure 3 and Figure 4.

To verify the computational accuracy of NAMD's new GPU-accelerated GB calculation and multi-core LCPO calculation, energy and surface area calculations were compared to those of prior reference implementations.[57,58] The relative error, $(E_{\text{ref}} - E_{\text{new}})/E_{\text{ref}}$, of NAMD's GB energy calculation, determined through Eq. (4), on the GPU, with NAMD's CPU implementation[58] as reference, falls below $5 \times 10^{-6}$ for all four benchmark proteins. The relative error, $(\text{SA}_{\text{ref}} - \text{SA}_{\text{new}})/\text{SA}_{\text{ref}}$, of NAMD's surface area calculation, determined through Eq. (13), with the Amber[57] implementation as reference, falls below $4 \times 10^{-7}$ for all four benchmark systems; total molecular surface areas of the four systems, as calculated by the Amber and NAMD implementations of LCPO, are listed in Table 1.

Hydrophobic solvation energy is a significant contributor to solvent behavior,[67,68] necessitating it's inclusion along with GB calculations. The effect of the hydrophobic energy contributions, accounted for in the SA calculation, was explored by simulating the 2W5U benchmark system using the GB/SA implicit solvent, employing eight different surface tension parameter values. Additionally, the 2W5U system was simulated in TIP3P[69] explicit solvent to determine which surface tension parameter values most closely reproduce protein behavior in explicit solvent. With the experimentally measured surface tension of hydrocarbons in aqueous solvent being 0.005 kcal/mol/Å$^2$,[62] surface tension parameter values in the range 0.001-0.128 kcal/mol/Å$^2$ were tested. Using otherwise the same simulation parameters previously outlined, the 2W5U system was equilibrated for 1.0 ns, at which time the total surface area, a molecular property closely affected by surface tension, had reached an equilibrium value. The surface areas evaluated during the simulations are plotted in Figure 5; for the TIP3P simulation, only the final protein surface area, as calculated by VMD's[63] high precision solvent-accessible surface area tool, with a solvent probe radius of 1.4 Å, is shown.

## Results

The performed simulations examine the hybrid GPU/CPU algorithm by verifying simultaneous overlap of GB and SA force calculations, analyzing the ideal ratio of GPUs to CPU cores which maximizes computing efficiency, benchmarking performance and evaluating appropriate values for the surface tension parameter.

### Hybrid GB/SA performance analysis

Central to fast hybrid GPU/CPU algorithms is the simultaneous overlap of calculations on both processor types, brought about by the host CPUs coordinating the GB calculation on the GPU with the SA calculations on the CPU. To demonstrate this coordination, Figure 3 illustrates, using the Projections[70] tool, the execution process of a GB/SA simulation of the 1GPB system on multiple GPUs and CPU cores.

Figure 3A illustrates the GB/SA calculation executed on 16 CPU cores without GPUs. The three phases of the GB calculation, described by Eq. (8), Eq. (11) and Eq. (10), are carried out with SA calculations interspersed to utilize the CPU when it would otherwise be idle while waiting for other cores to complete the phase; the lack of idle time (shown in yellow) signifies a highly efficient calculation.

Figure 3B demonstrates how the GB/SA calculation executed solely on 16 CPU cores is accelerated through the addition of 1 GPU. Because of its powerful computing capability, the GPU completes all Coulomb, van der Waals and GB calculations in the same time that the 16 CPU cores need to perform the SA calculation, resulting in a three-fold overall performance increase. This performance increase should be judged on the basis of an approximate 1 GPU/16 CPU cores cost ratio of about 1, i.e., by doubling the hardware cost, one triples performance.

A bottleneck for hybrid GPU/CPU calculations is the CPU's limited ability to switch between performing the SA calculation and GPU-related operations. Figure 3C illustrates the interplay between the GPU execution and the host CPU calculations. For each of the three GB phases, the host CPU initializes the GB calculation on the GPU then, while the GPU calculates the GB phase, the host CPU performs SA calculations. NAMD efficiently overlaps the GB and SA calculations by partitioning the SA calculation into many short, independent force work units, such that at the completion of each SA force work unit, the CPU can engage in GPU-related operations, if needed, then return to the next SA force work unit as highlighted by Figure 3C. Figure 3D illustrates how not partitioning the SA calculation increases GPU idle time during GB phase 1 and CPU idle time during GB phase 2 and 3, thereby decreasing performance.

Because of the performance difference between GPUs and CPUs and the different computational cost of the GB and SA calculations, it was not known what ratio of GPUs to CPUs would be the most efficient. The best ratio would have GPUs and CPUs requiring the same time to perform their respective force calculations. Based on the data in Table 2 and as illustrated in Figure 3, a ratio of 16 CPU cores per GPU allowed highly efficient performance, while deviating from this ratio resulted in either the GB or SA calculation to become rate limiting. For example, Table 2 shows that calculating GB/SA for the 1GPB system on 1 GPU is 1.9 times faster with 16 CPU cores than with 8 CPU cores (adding more cores alleviates the SA calculation bottleneck on the CPU) while 1 GPU with 32 CPU cores is only 1.3 times faster than with 16 CPU cores (GB calculation on GPU begins to be the bottleneck), thereby sharply decreasing computational efficiency. As both GPU and CPU technologies evolve, the ideal ratio of CPU cores per GPU will also evolve, both for our GB/SA algorithm as well as for other hybrid GPU/CPU algorithms.

## Performance benchmark results

The performance of the hybrid GB/SA implementation is demonstrated by the benchmarks, which are reported in full in Table 2 and illustrated for the 1IR2 system in Figure 4. The incorporation of GPUs into the MD simulation computation offers dramatic performance increases; for example, for the GB simulations of the 1IR2 system, simulations on 1 GPU are 25% faster than on 32 CPU cores. Additionally, the three largest systems tested were simulated, with the GB model, 30-35 times faster with 1 GPU than on 1 CPU core alone, while even the smallest system, 1YRI, was simulated 13 times faster with 1 GPU than on 1 CPU core alone. The GB/SA simulation of the 1IR2 system executed 50 times faster on 1 GPU with 16 CPU cores than on 1 CPU core and 3.5 times faster than on 16 CPU cores.

In case of larger simulated systems, one can take advantage of more GPU accelerators; because GPUs operate most efficiently when they are issued a lot of concurrent work, for moderately sized systems one achieves the great simulation speeds on relatively few GPUs (see the GB columns of Table 2). The 582-atom villin headpiece is simulated at the same speed on 1 GPU as on 32 CPU cores, but not any faster on 2 GPUs. The larger 2,412-atom flavodoxin is simulated 27% faster on 1 GPU than on 32 CPU cores, and 60% faster on 2 GPUs. The 13,340-atom glycogen phosphorylase is simulated 3.2 times faster on 4 GPUs than on 32 CPU cores while the 74,926-atom RuBisCO is simulated 4.5 times faster for the same comparison.

To enhance performance for large systems, we employed in the present study an interaction cutoff. Because cutoff lengths affect both accuracy and performance, we compare in Table 3 the computational expense of simulating the four benchmark systems as well as the relative error in calculating the total GB energy using a no-cutoff calculation as a reference, $\left(E^{\mathrm{GB}}_{\mathrm{cutoff}} - E^{\mathrm{GB}}_{\mathrm{no\ \ cutoff}}\right)/E^{\mathrm{GB}}_{\mathrm{no\ \ cutoff}}$. In case of the 16 Å cutoff, compared to a no-cutoff calculation, computational expense is reduced, the relative error is about 0.5%, and the fraction of computational expense associated with the SA calculation is higher (requiring more CPU cores per GPU).

The impressive speed-ups achieved by utilizing GPU technology demonstrate the benefit which GPUs offer to MD computing. While the benchmarks reported here are only for up to 4 GPUs, the hybrid algorithm also operates on supercomputers built from many GPUs and multi-core CPUs as previously described for GPU-acceleration of NAMD on distributed memory computers.[38]

### Testing surface tension parameters

When employing the SA calculation, the surface tension, in units kcal/mol/Å$^2$, is a parameter of the model, controlling hydrophobic energy of solvation; it is not a physical property measured from the simulation. To explore the effect of the parameter on a protein system, the 2W5U system was equilibrated through a GB/SA implicit solvent simulation employing surface tension parameters ranging from 0.001 to 0.128 kcal/mol/Å$^2$; the system was also equilibrated through a standard all-atom simulation employing explicit TIP3P[69] solvent. Figure 5 presents protein surface area arising in the GB/SA equilibration simulations as well as the final surface area of each. The protein surface area is expected to diminish with increasing surface tension as the latter imparts an energy penalty for the protein’s surface exposed to solvent. The simulations utilizing surface tension parameter values of 0.004 and 0.008 kcal/mol/Å$^2$ returned final surface areas closest to the area of the reference TIP3P equilibrated system, suggesting that employing surface tension parameter values between 0.004 and 0.008 kcal/mol/Å$^2$ most closely reproduce protein behavior in explicit solvent. The experimentally measured surface tension of hydrocarbons in aqueous solvent is 0.005 kcal/mol/Å$^2$ [62] which validates our finding.

## Conclusions

The structural biology community has been well served by technological advances of general purpose GPU computing, which have made molecular dynamics more powerful and accurate. Many biological computing programs have already achieved great improvements in performance through GPU acceleration. As it is often neither feasible nor ideal to perform all needed calculations on the GPU, it becomes increasingly important to develop methods for overlapping GPU and CPU calculations. From the present work results an efficient method for performing GB/SA simulations on hybrid GPU/CPU computers, permitting extremely fast and accurate molecular dynamics simulations that can be executed, for example, interactively by researchers.

## Acknowledgments

The authors thank John Stone for helpful discussions regarding implementation and performance optimization for GPUs. This work was supported by the National Institutes of Health (NIH P41-RR005969).

## References

(1). Le L, Lee EH, Schulten K, Truong T. PLoS Currents: Influenza. 2010 2009 Aug 27:RRN1015, (9 pages).

(2). Le L, Lee EH, Hardy DJ, Truong TN, Schulten K. PLoS Comput. Biol. 2010; 6:e1000939. 13 pages. [PubMed: 20885781]

(3). Cheng LS, Amaro RE, Xu D, Li WW, Arzberger PW, McCammon JA. J. Med. Chem. 2008; 51:3878–3894. [PubMed: 18558668]

(4). Acharya R, Carnevale V, Fiorin G, Leviné BG, Polishchuk AL, Balannik V, Samish I, Lamb RA, Pinto LH, DeGrado WF, Klein ML. Proc. Natl. Acad. Sci. USA. 2010; 107:15075–15080. [PubMed: 20689043]

(5). Khurana E, Peraro MD, DeVane R, Vemparala S, DeGrado WF, Klein ML. Proc. Natl. Acad. Sci. USA. 2009; 106:1069–1074. [PubMed: 19144924]

(6). Khurana E, DeVane RH, Peraro MD, Klein ML. Biochim. Biophys. Acta. 2011; 1808:530–537. [PubMed: 20385097]

(7). Newhouse EI, Xu D, Markwick PRL, Amaro RE, Pao HC, Wu KJ, Alam M, McCammon JA, Li WW. J. Am. Chem. Soc. 2009; 131:17430–17442. [PubMed: 19891427]

(8). Fidelak J, Juraszek J, Branduardi D, Bianciotto M, Gervasio FL. J. Phys. Chem. B. 2010; 114:9516–9524. [PubMed: 20593892]

(9). Lee EH, Hsin J, Sotomayor M, Comellas G, Schulten K. Structure. 2009; 17:1295–1306. [PubMed: 19836330]

(10). Trabuco LG, Villa E, Mitra K, Frank J, Schulten K. Structure. 2008; 16:673–683. [PubMed: 18462672]

(11). Davidovich C, Bashan A, Yonath A. Proc. Natl. Acad. Sci. USA. 2008; 105:20665–20670. [PubMed: 19098107]

(12). Poehlsgaard J, Douthwaite S. Nat. Rev. Microbiol. 2005; 3:870–881. [PubMed: 16261170]

(13). Lim B, Lee EH, Sotomayor M, Schulten K. Structure. 2008; 16:449–459. [PubMed: 18294856]

(14). Hsin J, Strümpfer J, Lee EH, Schulten K. Annu. Rev. Biophys. 2011; 40:187–203. [PubMed: 21332356]

(15). Venkatesan B, Polans J, Comer J, Sridhar S, Wendell D, Aksimentiev A, Bashir R. Biomed. Microdev. 2011:1–12.

(16). Carr R, Comer J, Ginsberg MD, Aksimentiev A. J. Phys. Chem. Lett. 2011; 2:1804–1807. [PubMed: 22611479]

(17). Freddolino PL, Arkhipov AS, Larson SB, McPherson A, Schulten K. Structure. 2006; 14:437–449. [PubMed: 16531228]

(18). Arkhipov A, Freddolino PL, Schulten K. Structure. 2006; 14:1767–1777. [PubMed: 17161367]

(19). Tavoosi N, Davis-Harrison RL, Pogorelov TV, Ohkubo YZ, Arcario MJ, Clay MC, Rienstra CM, Tajkhorshid E, Morrissey JH. J. Biol. Chem. 2011; 286:23247–23253. [PubMed: 21561861]

(20). Ohkubo YZ, Tajkhorshid E. Structure. 2008; 16:72–81. [PubMed: 18184585]

(21). Ohkubo YZ, Morrissey JH, Tajkhorshid E. J Thromb. Haem. 2010; 8:1044–1053.

(22). Morrissey JH, Pureza V, Davis-Harrison RL, Sligar SG, Rienstra CM, Kijac AZ, Ohkubo YZ, Tajkhorshid E. J Thromb. Haem. 2009; 7:169–172.

(23). Morrissey JH, Davis-Harrison RL, Tavoosi N, Ke K, Pureza V, Boettcher JM, Clay MC, Rienstra CM, Ohkubo YZ, Pogorelov TV, Tajkhorshid E. Thromb. Res. 2010; 125(Suppl. 1):S23–S25. [PubMed: 20129649]

(24). Morrissey JH, Pureza V, Davis-Harrison RL, Sligar SG, Ohkubo YZ, Tajkhorshid E. Thromb. Res. 2008; 122:S23–S26. [PubMed: 18691494]

(25). Interlandi G, Thomas W. Proteins: Struct., Func., Gen. 2010; 78:2506–2522.

(26). Miller Y, Ma B, Nussinov R. J. Am. Chem. Soc. 2011; 133:2742–2748. [PubMed: 21299220]

(27). Parthasarathy S, Long F, Miller Y, Xiao Y, McElheny D, Thurber K, Ma B, Nussi-nov R, Ishii Y. J. Am. Chem. Soc. 2011; 133:3390–3400. [PubMed: 21341665]

(28). Miller Y, Ma B, Tsai C-J, Nussinov R. Proc. Natl. Acad. Sci. USA. 2010; 107:14128–14133. [PubMed: 20660780]

(29). Fogolari F, Corazza A, Viglino P, Zuccato P, Pieri L, Faccioli P, Bellotti V, Esposito G. Biophys. J. 2007; 92:1673–1681. [PubMed: 17158575]

(30). Buchete N-V, Hummer G. Biophys. J. 2007; 92:3032–3039. [PubMed: 17293399]

(31). Buchete N-V, Tycko R, Hummer G. J. Mol. Biol. 2005; 353:804–821. [PubMed: 16213524]

(32). Shaw, DE.; Dror, RO.; Salmon, JK.; Grossman, J.; Mackenzie, KM.; Bank, JA.; Young, C.; Deneroff, MM.; Batson, B.; Bowers, KJ.; Chow, E.; Eastwood, MP.; Ierardi, DJ.; Klepeis, JL.; Kuskin, JS.; Larson, RH.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, MA.; Piana, S.; Shan, Y.;

Towles, B. Millisecond-Scale Molecular Dynamics Simulations on Anton; SC’09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis; New York, NY, USA. 2009; p. 39:1-39:11.

(33). Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. J. Comp. Chem. 2005; 26:1781–1802. [PubMed: 16222654]

(34). Kalé L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K. J. Comp. Phys. 1999; 151:283–312.

(35). Schulz R, Lindner B, Petridis L, Smith JC. J. Chem. Theor. Comp. 2009; 5:2798–2808.

(36). Mei, C.; Sun, Y.; Zheng, G.; Bohm, EJ.; Kalé, LV.; Phillips, JC.; Harrison, C. Enabling and Scaling Biomolecular Simulations of 100 Million Atoms on Petascale Machines with a Multicore-optimized Message-driven Runtime; Proceedings of the 2011 ACM/IEEE conference on Supercomputing; Seattle, WA. 2011;

(37). Stone JE, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K. J. Comp. Chem. 2007; 28:2618–2640. [PubMed: 17894371]

(38). Phillips, JC.; Stone, JE.; Schulten, K. Adapting a Message-Driven Parallel Application to GPU-Accelerated Clusters; SC’08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing; Piscataway, NJ, USA. 2008;

(39). Stone, JE.; Hardy, DJ.; Isralewitz, B.; Schulten, K. Scientific Computing with Multicore and Accelerators. Dongarra, J.; Bader, DA.; Kurzak, J., editors. Chapman & Hall/CRC Press; 2011. p. 351-371.Chapter 16

(40). Hardy, DJ.; Stone, JE.; Vandivort, KL.; Gohara, D.; Rodrigues, C.; Schulten, K. GPU Computing Gems. Hwu, W., editor. Morgan Kaufmann Publishers; 2011. p. 43-58.Chapter 4

(41). Stone, JE.; Hardy, DJ.; Saam, J.; Vandivort, KL.; Schulten, K. GPU Computing Gems. Hwu, W., editor. Morgan Kaufmann Publishers; 2011. p. 5-18.Chapter 1

(42). Roberts, E.; Stone, JE.; Sepulveda, L.; Hwu, WW.; Luthey-Schulten, Z. Long time-scale simulations of in vivo diffusion using GPU hardware; Proceedings of the IEEE International Parallel & Distributed Processing Symposium; 2009; p. 1-8.

(43). Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. J. Comp. Chem. 2009; 30:1545–1614. [PubMed: 19444816]

(44). Harvey MJ, Giupponi G, Fabritiis GD. J. Chem. Theor. Comp. 2009; 5:1632–1639.

(45). Baker JA, Hirst JD. Mol. Inf. 2011; 30:498–504.

(46). Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS. J. Comp. Chem. 2009; 30:864–872. [PubMed: 19191337]

(47). Eastman P, Pande VS. J. Comp. Chem. 2010; 31:1268–1272. [PubMed: 19847780]

(48). Anderson JA, Lorenz CD, Travesset A. J. Chem. Phys. 2008; 227:5342–5359.

(49). Liu F, Gruebele M. J. Mol. Biol. 2007; 370:574–584. [PubMed: 17532338]

(50). Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. Computational Molecular Dynamics: Challenges, Methods, Ideas. Deuflhard, P.; Hermans, J.; Leimkuhler, B.; Mark, AE.; Reich, S.; Skeel, RD., editors. Vol. 4. Lecture Notes in Computational Science and Engineering; Springer-Verlag; Berlin: 1998. p. 39-65.

(51). Wells DB, Abramkina V, Aksimentiev A. J. Chem. Phys. 2007; 127:125101. [PubMed: 17902937]

(52). Shivakumar D, Deng Y, Roux B. J. Chem. Theor. Comp. 2009; 5:919–930.

(53). Still WC, Tempczyk A, Hawley RC, Hendrickson T. J. Am. Chem. Soc. 1990; 112:6127–6129.

(54). Weiser J, Shenkin PS, Still WC. J. Comp. Chem. 1998; 20:217–230.

(55). Onufriev A, Bashford D, Case DA. J. Phys. Chem. 2000; 104:3712–3720.

(56). Onufriev A, Bashford D, Case DA. Proteins: Struct., Func., Bioinf. 2004; 55:383–394.

(57). Case D, Cheatham T III, Darden T, Gohlke H, Luo R, Merz K Jr, Onufriev A, Simmerling C, Wang B, Woods R. J. Comp. Chem. 2005; 26:1668. [PubMed: 16200636]

(58). Tanner DE, Chan K-Y, Phillips J, Schulten K. J. Chem. Theor. Comp. 2011; 7:3635–3642.

(59). Hassan SA, Mehler EL. Proteins: Struct., Func., Gen. 2002; 47:45–61.

(60). Hassan SA, Mehler EL, Zhang D, Weinstein H. Proteins: Struct., Func., Gen. 2003; 51:109–125.

(61). Srinivasan J, Trevathan MW, Beroza P, Case DA. Theoret. Chim. Acta. 1999; 101:426–434.

(62). Sitkoff D, Sharp KA, Honig B. J. Phys. Chem. 1994; 98:1978–1988.

(63). Humphrey W, Dalke A, Schulten K. J. Mol. Graphics. 1996; 14:33–38.

(64). Rodrigues, CI.; Hardy, DJ.; Stone, JE.; Schulten, K.; Hwu, WW. GPU Acceleration of Cutoff Pair Potentials for Molecular Modeling Applications; CF'08: Proceedings of the 2008 conference on Computing Frontiers; New York, NY, USA. 2008; p. 273-282.

(65). Stone JE, Hardy DJ, Ufimtsev IS, Schulten K. J. Mol. Graph. Model. 2010; 29:116–125. [PubMed: 20675161]

(66). Ryckaert J-P, Ciccotti G, Berendsen HJC. J. Comp. Phys. 1977; 23:327–341.

(67). Zhu J, Shi Y, Liu H. J. Phys. Chem. B. 2002; 106:4844–4853.

(68). Daidone I, Ulmschneider MB, Nola AD, Amadei A, Smith JC. Proc. Natl. Acad. Sci. USA. 2007; 104:15230–15235. [PubMed: 17881585]

(69). Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. J. Chem. Phys. 1983; 79:926–935.

(70). Kalé LV, Zheng G, Lee CW, Kumar S. Fut. Gen. Comp. Sys. 2006; 22:347–358.

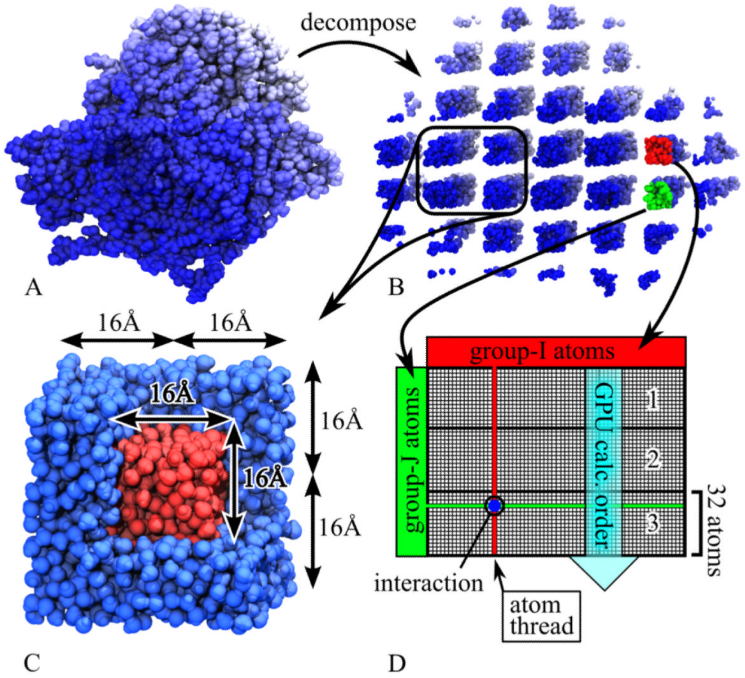

**Figure 1.**

Hybrid GB/SA decomposition. (A) Simulated protein glycogen phosphorylase. (B) Protein decomposed into a 3D grid of atom-groups. (C) LCPO force work unit involving a 2 × 2 × 2 set of 8 adjacent atom-groups. The force work unit calculates forces for the inner 1/8th core of atoms (red) due to overlap with neighbors (red and blue). (D) Thread block design for GB force calculation on GPU multiprocessor. Each thread (vertical line) calculates the force on one group-I atom (red) due to interactions (blue) with group-J atoms (green); group-J atoms are loaded 32 at a time, in order 1-3, for coalesced and, therefore, accelerated memory access.

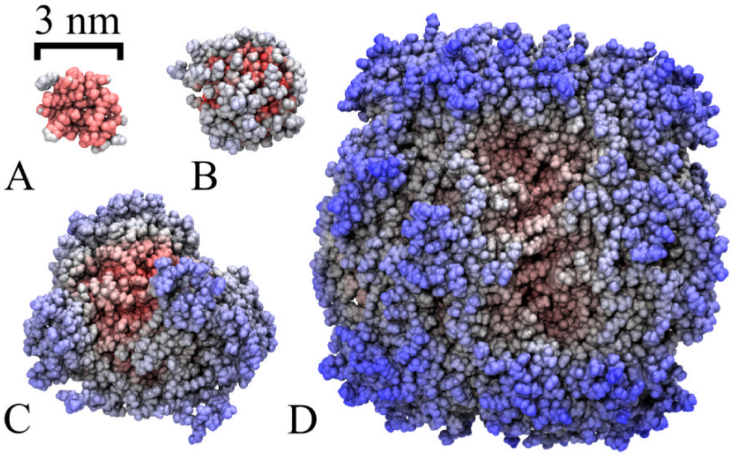

**Figure 2.**
Benchmarked protein systems. (A) Villin headpiece (PDB ID 1YRI) with 582 atoms; (B) flavodoxin (PDB ID 2W5U) with 2,412 atoms; (C) glycogen phosphorylase (PDB ID 1GPB) with 13,340 atoms; (D) RuBisCO (PDB ID 1IR2) with 74,926 atoms.

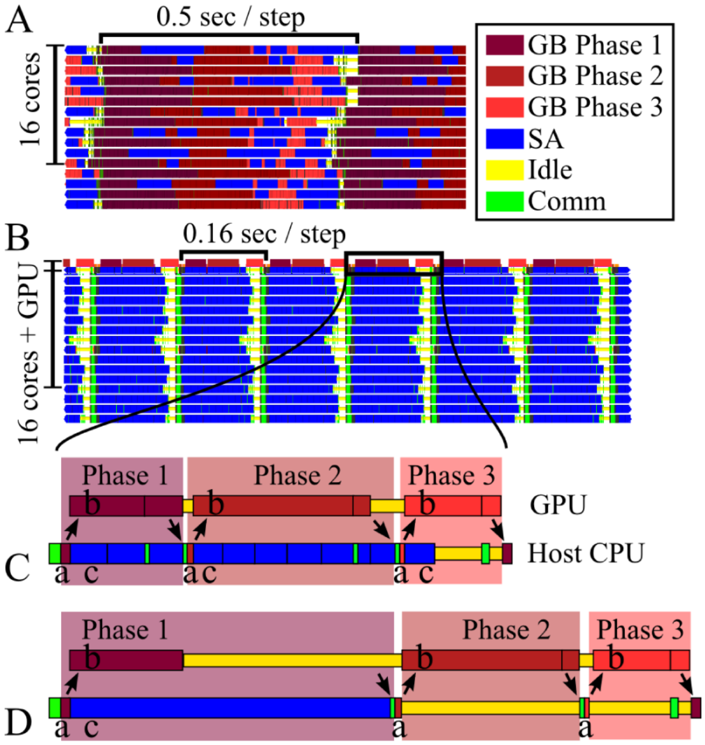

**Figure 3.**

NAMD's GB/SA calculation for the 1GPB system. (A) Performance on 16 CPUs only. (B) Performance on 16 CPUs with 1 GPU added. Colors represent calculations being performed on the processors: three phases of GB calculation (see text); SA calculation; idle time; communication. (C) Detailed view of host CPU switching between communicating with the GPU and performing the SA calculation; arrows represent the transfer of data between host CPU and GPU; for GB phase 1, 2 and 3 calculations: (a) host CPU initializes the GB calculation on the GPU; (b) GPU calculates a GB phase while (c) host CPU performs SA calculations. (D) Detailed view of how switching without SA partitioning slows performance.

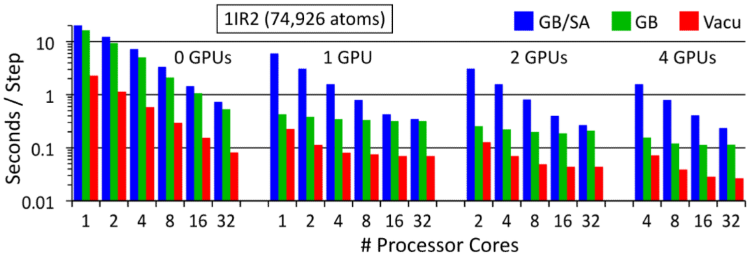

**Figure 4.**

Simulation speed, in seconds/time step, for the 1IR2 benchmark system on 0-4 GPUs and 1-32 CPU cores; speeds for in vacuo (red), GB (green) and GB/SA (blue) calculations are shown. Both the seconds/step and processor cores axes are logarithmic.

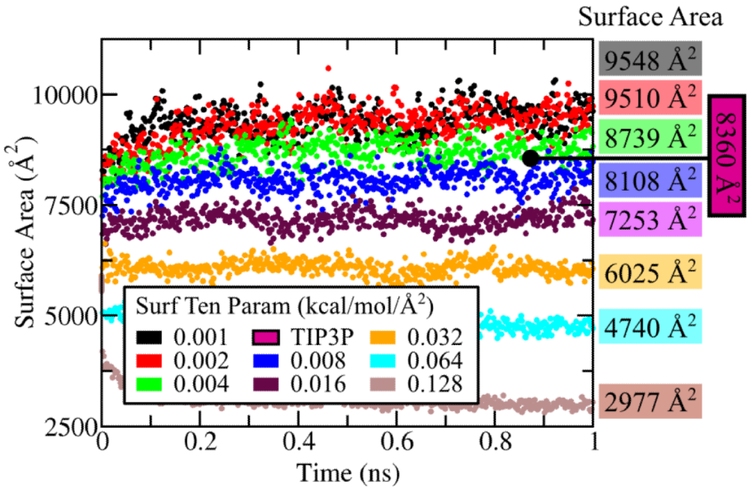

**Figure 5.**

Surface area during eight GB/SA simulations of the 2W5U system employing different surface tension parameter values. The final surface area for each GB/SA simulation is listed at right; shown is also the final surface area resulting from the simulation with explicit (TIP3P) solvent.

## Table 1

Benchmarked protein systems. Listed are associated surface areas, in units Å$^2$, as calculated by NAMD's and Amber's implementations of the LCPO[54] algorithm.

| Name | PDB ID | # Atoms | NAMD SA | Amber SA |
|---|---|---:|---:|---:|
| villin headpiece | 1YRI | 582 | 2,706.72 | 2,706.72 |
| flavodoxin | 2W5U | 2,412 | 9,988.72 | 9,988.72 |
| glycogen phosphorylase | 1GPB | 13,340 | 46,106.66 | 46,106.68 |
| RuBisCO | 1IR2 | 74,926 | 176,335.86 | 176,335.90 |

## Table 2

Benchmark simulation speeds, in units seconds/time step, for the four benchmarked protein systems tested on 0-4 GPUs and 1-32 CPU cores. Speeds are reported for three increasingly accurate and expensive simulation types: in vacuo (Vacu), GB, and GB/SA. The small size of system 1YRI, see Figure 2A, prohibited utilization of 32 CPU cores in some cases.

| | | 1YRI | | | 2W5U | | | 1GPB | | | 1IR2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPU | CPU | Vacu | GB | GB/SA | Vacu | GB | GB/SA | Vacu | GB | GB/SA | Vacu | GB | GB/SA |
| 0 | 1 | 8.4E−3 | 4.9E−2 | 7.3E−2 | 4.9E−2 | 3.5E−1 | 5.2E−1 | 3.4E−1 | 2.4E+0 | 3.4E+0 | 2.3E+0 | 1.6E+1 | 2.2E+1 |
| | 2 | 5.7E−3 | 2.7E−2 | 4.0E−2 | 2.7E−2 | 1.8E−1 | 2.6E−1 | 1.8E−1 | 1.2E+0 | 1.7E+0 | 1.2E+0 | 9.5E+0 | 1.2E+1 |
| | 4 | 3.8E−3 | 1.5E−2 | 2.2E−2 | 1.4E−2 | 9.3E−2 | 1.3E−1 | 9.4E−2 | 6.2E−1 | 8.8E−1 | 5.9E−1 | 5.1E+0 | 7.2E+0 |
| | 8 | 2.6E−3 | 8.7E−3 | 1.3E−2 | 8.0E−3 | 4.7E−2 | 6.8E−2 | 4.8E−2 | 3.2E−1 | 4.4E−1 | 3.0E−1 | 2.1E+0 | 3.4E+0 |
| | 16 | 2.2E−3 | 5.6E−3 | 7.5E−3 | 4.8E−3 | 2.6E−2 | 3.7E−2 | 2.6E−2 | 1.6E−1 | 2.3E−1 | 1.6E−1 | 1.1E+0 | 1.5E+0 |
| | 32 | 1.7E−3 | 3.7E−3 | 4.9E−3 | 3.1E−3 | 1.4E−2 | 2.1E−2 | 1.4E−2 | 8.3E−2 | 1.2E−1 | 8.2E−2 | 5.4E−1 | 7.3E−1 |
| 1 | 1 | 3.6E−3 | 3.8E−3 | 3.1E−2 | 8.1E−3 | 1.1E−2 | 1.7E−1 | 4.3E−2 | 6.7E−2 | 1.0E+0 | 2.3E−1 | 4.3E−1 | 6.0E+0 |
| | 2 | 3.2E−3 | 4.4E−3 | 1.9E−2 | 5.0E−3 | 1.1E−2 | 8.9E−2 | 2.3E−2 | 5.8E−2 | 5.2E−1 | 1.1E−1 | 3.9E−1 | 3.1E+0 |
| | 4 | 2.1E−3 | 3.4E−3 | 1.2E−2 | 3.6E−3 | 9.9E−3 | 4.6E−2 | 1.4E−2 | 5.4E−2 | 2.7E−1 | 8.1E−2 | 3.5E−1 | 1.6E+0 |
| | 8 | 1.8E−3 | 3.3E−3 | 6.9E−3 | 2.8E−3 | 1.0E−2 | 2.7E−2 | 1.3E−2 | 5.1E−2 | 1.4E−1 | 7.5E−2 | 3.3E−1 | 8.0E−1 |
| | 16 | 2.1E−3 | 3.3E−3 | 4.8E−3 | 2.7E−3 | 1.1E−2 | 1.8E−2 | 1.3E−2 | 5.4E−2 | 8.1E−2 | 7.1E−2 | 3.2E−1 | 4.3E−1 |
| | 32 | | | | 2.8E−3 | 8.9E−3 | 1.2E−2 | 1.2E−2 | 5.0E−2 | 6.1E−2 | 7.0E−2 | 3.2E−1 | 3.5E−1 |
| 2 | 2 | 3.0E−3 | 3.9E−3 | 2.0E−2 | 5.0E−3 | 9.2E−3 | 8.9E−2 | 2.4E−2 | 4.4E−2 | 5.2E−1 | 1.3E−1 | 2.6E−1 | 3.1E+0 |
| | 4 | 2.3E−3 | 3.4E−3 | 1.1E−2 | 3.7E−3 | 9.2E−3 | 5.0E−2 | 1.4E−2 | 3.7E−2 | 2.7E−1 | 7.1E−2 | 2.2E−1 | 1.6E+0 |
| | 8 | 1.8E−3 | 3.2E−3 | 7.9E−3 | 2.5E−3 | 8.6E−3 | 2.9E−2 | 1.0E−2 | 3.6E−2 | 1.4E−1 | 4.9E−2 | 2.0E−1 | 8.1E−1 |
| | 16 | 1.7E−3 | 3.2E−3 | 5.8E−3 | 2.7E−3 | 8.7E−3 | 1.9E−2 | 8.9E−3 | 3.8E−2 | 8.1E−2 | 4.4E−2 | 1.9E−1 | 4.0E−1 |
| | 32 | | | | 2.5E−3 | 7.6E−3 | 1.3E−2 | 7.9E−3 | 3.4E−2 | 5.3E−2 | 4.4E−2 | 2.1E−1 | 2.7E−1 |
| 4 | 4 | 2.7E−3 | 3.4E−3 | 1.2E−2 | 3.5E−3 | 1.0E−2 | 5.0E−2 | 1.4E−2 | 3.0E−2 | 2.7E−1 | 7.2E−2 | 1.6E−1 | 1.6E+0 |
| | 8 | 1.8E−3 | 3.1E−3 | 1.2E−2 | 2.5E−3 | 8.1E−3 | 3.0E−2 | 8.5E−3 | 2.6E−2 | 1.4E−1 | 3.9E−2 | 1.2E−1 | 8.0E−1 |
| | 16 | 2.2E−3 | 3.2E−3 | 6.3E−3 | 2.6E−3 | 7.9E−3 | 2.1E−2 | 7.2E−3 | 2.8E−2 | 9.1E−2 | 2.9E−2 | 1.1E−1 | 4.1E−1 |
| | 32 | | | | 2.4E−3 | 6.7E−3 | 1.2E−2 | 6.4E−3 | 2.3E−2 | 4.7E−2 | 2.7E−2 | 1.2E−1 | 2.4E−1 |

## Table 3

Computational expense and accuracy for different non-bonded interaction cutoff lengths. Listed for each of the four benchmarked systems and for each cutoff (no cutoff, 30 Å cutoff, and 16 Å cutoff), are the execution time required for one timestep on one CPU core, the percent of computational expense associated with the Coulomb and van der Waals (Vac), GB and SA calculations and the relative error in the total GB energy calculated with cutoff, namely $\left(E^{\mathrm{GB}}_{\mathrm{cutoff}} - E^{\mathrm{GB}}_{\mathrm{no\ \ cutoff}}\right)/E^{\mathrm{GB}}_{\mathrm{no\ \ cutoff}}$.

| | **no cutoff** | | | | **30 Å cutoff** | | | | | **16 Å cutoff** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PDB** | **Time** | **Vac** | **GB** | **SA** | **Time** | **Vac** | **GB** | **SA** | **Err** | **Time** | **Vac** | **GB** | **SA** | **Err** |
| 1IR2 | 556s | 21% | 76% | 3% | 54.9s | 19% | 69% | 12% | 0.04% | 22s | 10% | 62% | 28% | 0.3% |
| 1GPB | 17.8s | 19% | 74% | 7% | 6.61s | 15% | 72% | 13% | 0.3% | 3.4s | 10% | 60% | 30% | 0.6% |
| 2W5U | 0.61s | 15% | 75% | 10% | 0.57s | 15% | 73% | 10% | 0.04% | 0.52s | 10% | 57% | 33% | 0.7% |
| 1YRI | 0.05s | 15% | 60% | 25% | 0.05s | 14% | 59% | 27% | 0.3% | 0.07s | 8% | 59% | 33% | 0.3% |