DE NOVO ULTRASCALE ATOMISTIC SIMULATIONS ON HIGH-END PARALLEL SUPERCOMPUTERS

Aiichiro Nakano[1]
Rajiv K. Kalia[1]
Ken-ichi Nomura[1]
Ashish Sharma[1, 2]
Priya Vashishta[1]
Fuyuki Shimojo[1,3]
Adri C. T. van Duin[4]
William A. Goddard, III[4]
Rupak Biswas[5]
Deepak Srivastava[5]
Lin H. Yang[6]

Abstract

We present a de novo hierarchical simulation framework for first-principles based predictive simulations of materials and their validation on high-end parallel supercomputers and geographically distributed clusters. In this framework, high-end chemically reactive and non-reactive molecular dynamics (MD) simulations explore a wide solution space to discover microscopic mechanisms that govern macroscopic material properties, into which highly accurate quantum mechanical (QM) simulations are embedded to validate the discovered mechanisms and quantify the uncertainty of the solution. The framework includes an embedded divide-and-conquer (EDC) algorithmic framework for the design of linear-scaling simulation algorithms with minimal bandwidth complexity and tight error control. The EDC framework also enables adaptive hierarchical simulation with automated model transitioning assisted by graph-based event tracking. A tunable hierarchical cellular decomposition parallelization framework then maps the $O(N)$ EDC algorithms onto petaflops computers, while achieving performance tunability through a hierarchy of parameterized cell data/computation structures, as well as its implementation using hybrid grid remote procedure call + message passing + threads programming. High-end computing platforms such as IBM BlueGene/L, SGI Altix 3000 and the NSF TeraGrid provide an excellent test grounds for the framework. On these platforms, we have achieved unprecedented scales of quantum-mechanically accurate and well validated, chemically reactive atomistic simulations—1.06 billion-atom fast reactive force-field MD and 11.8 million-atom (1.04 trillion grid points) quantum-mechanical MD in the framework of the EDC density functional theory on adaptive multigrids—in addition to 134 billion-atom non-reactive space–time multiresolution MD, with the parallel efficiency as high as 0.998 on 65,536 dual-processor BlueGene/L nodes. We have also achieved an automated execution of hierarchical QM/MD simulation on a grid consisting of 6 supercomputer centers in the US and Japan (in total of 150,000 processor hours), in which the number of processors change dynamically on demand and resources are allocated and migrated dynamically in response to faults. Furthermore, performance portability has been demonstrated on a wide range of platforms such as BlueGene/L, Altix 3000, and AMD Opteron-based Linux clusters.

Key words: hierarchical simulation, molecular dynamics, reactive force field, quantum mechanics, density functional theory, parallel computing, grid computing

1 Introduction

Petaflops computers (Dongarra and Walker 2001) to be built in the near future and grids (Allen et al. 2001; Kikuchi et al. 2002; Foster and Kesselman 2003) on geographically distributed parallel supercomputers will offer tremendous opportunities for high-end computational sciences. Their computing power will enable unprecedented scales of first-principles based predictive simulations to quantitatively study system-level behavior of complex dynamic systems (Emmott and Rison 2006). An example is the understanding of microscopic mechanisms that govern macroscopic materials behavior, thereby enabling rational design of material compositions and microstructures to produce desired material properties.

The multitude of length and time scales and the associated wide solution space have thus far precluded such

[1]COLLABORATORY FOR ADVANCED COMPUTING AND SIMULATIONS, UNIVERSITY OF SOUTHERN CALIFORNIA, LOS ANGELES, CA 90089-0242, USA (ANAKANO@USC.EDU)

[2]DEPARTMENT OF BIOMEDICAL INFORMATICS, OHIO STATE UNIVERSITY, COLUMBUS, OH 43210, USA

[3]DEPARTMENT OF PHYSICS, KUMAMOTO UNIVERSITY, KUMAMOTO 860-8555, JAPAN

[4]MATERIALS AND PROCESS SIMULATION CENTER, CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CA 91125, USA

[5]NASA ADVANCED SUPERCOMPUTING (NAS) DIVISION, NASA AMES RESEARCH CENTER, MOFFETT FIELD, CA 94035, USA

[6]PHYSICS/H DIVISION, LAWRENCE LIVERMORE NATIONAL LABORATORY, LIVERMORE, CA 94551, USA

The International Journal of High Performance Computing Applications, Volume 22, No. 1, Spring 2008, pp. 113–128
DOI: 10.1177/1094342007085015

Figures 3, 7 appear in color online: http://hpc.sagepub.com

first-principles approaches. A promising approach is hierarchical simulation (Broughton et al. 1999; Nakano et al. 2001; Ogata et al. 2001), in which atomistic molecular dynamics (MD) simulations (Kale et al. 1999; Abraham et al. 2002; Kadau et al. 2002; Nakano et al. 2002) of varying accuracy and computational costs (from classical non-reactive MD to chemically reactive MD based on semi-classical approaches) explore a wide solution space to discover new mechanisms, in which highly accurate quantum mechanical (QM) simulations (Car and Parrinello 1985; Kendall et al. 2000; Truhlar and McKoy 2000; Gygi et al. 2005; Ikegami et al. 2005) are embedded to validate the discovered mechanisms and quantify the uncertainty of the solution.

A simple estimate indicates that a 100 billion-atom MD simulation for one nanosecond (or 500,000 time steps), which embeds 200 million-atom reactive MD and 1 million-atom QM simulations, will require 45 days of computing on a petaflops platform. To enable such ultrascale simulations in the near future, however, nontrivial developments in algorithmic and computing techniques, as well as thorough scalability tests, are required today.

We are developing a ***de novo hierarchical simulation framework*** to enable first-principles based hierarchical simulations of materials and their validation on petaflops computers and Grids. The framework includes:

- An ***embedded divide-and-conquer (EDC) algorithmic framework*** for: 1) the design of linear-scaling algorithms for approximate solutions of hard simulation problems with minimal bandwidth complexity and codified tight error control; and 2) adaptive embedding of QM simulations in MD simulation so as to guarantee the quality of the overall solution, where ***graph-based event tracking*** (i.e. shortest-path circuit analysis of the topology of chemical bond networks) automates the embedding upon the violation of error tolerance.
- A ***tunable hierarchical cellular decomposition (HCD) parallelization framework*** for: 1) mapping the linear-scaling EDC algorithms onto petaflops computers, while achieving performance tunability through a hierarchy of parameterized cell data/computation structures; and 2) enabling tightly coupled computations of considerable scale and duration on distributed clusters, based on ***hybrid grid remote procedure call + message passing + threads programming*** to combine flexibility, fault tolerance, and scalability.

High-end parallel supercomputers such as IBM BlueGene/L and SGI Altix 3000, as well as grid test-beds such as the NSF TeraGrid, are excellent test grounds for such scalable de novo hierarchical simulation technologies. This paper describes scalability tests of our hierarchical simulation framework on these platforms as well as its portability to other platforms such as AMD Opteron-based Linux clusters. In the next section, we describe the EDC algorithmic framework. Section 3 discusses the tunable HCD parallelization framework. Results of benchmark tests are given in Section 4, and Section 5 contains conclusions.

2 Embedded Divide-and-Conquer Algorithms for De Novo Hierarchical Simulations

Prerequisite to successful hierarchical simulations and validation at the petaflops scale are simulation algorithms that are scalable beyond 10^5 processors. We use a unified embedded divide-and-conquer (EDC) algorithmic framework based on data locality principles to design linear-scaling algorithms for broad scientific applications with tight error control (Shimojo et al. 2005; Nakano et al. 2006). In EDC algorithms, spatially localized sub-problems are solved in a global embedding field, which is efficiently computed with tree-based algorithms (Figure 1). Examples of the embedding field are: 1) the electrostatic field in molecular dynamics (MD) simulation (Nakano et al. 2002); 2) the self-consistent Kohn-Sham potential in the density functional theory (DFT) in quantum mechanical (QM) simulation (Shimojo et al. 2005); and 3) a coarser but less compute-intensive simulation method in hierarchical simulation (Nakano et al. 2001).

2.1 Linear-Scaling Molecular-Dynamics and Quantum-Mechanical Simulation Algorithms

In the past several years, we have used the EDC framework to develop a suite of linear-scaling MD algorithms, in which interatomic forces are computed with varying accuracy and complexity: 1) classical MD involving the formally $O(N^2)$ N-body problem; 2) reactive force-field (ReaxFF) MD involving the $O(N^3)$ variable N-charge problem; 3) quantum mechanical (QM) calculation based on the DFT to provide approximate solutions to the exponentially complex quantum N-body problem; and 4) adaptive hierarchical QM/MD simulations that embed highly accurate QM simulations in MD simulation only when and where high fidelity is required. The Appendix describes the three EDC simulation algorithms that are used in our adaptive hierarchical simulations:

- Algorithm 1—MRMD: space–time multiresolution molecular dynamics.
- Algorithm 2—F-ReaxFF: fast reactive force-field molecular dynamics.
- Algorithm 3—EDC-DFT: embedded divide-and-conquer density functional theory on multigrids for quantum-mechanical molecular dynamics.

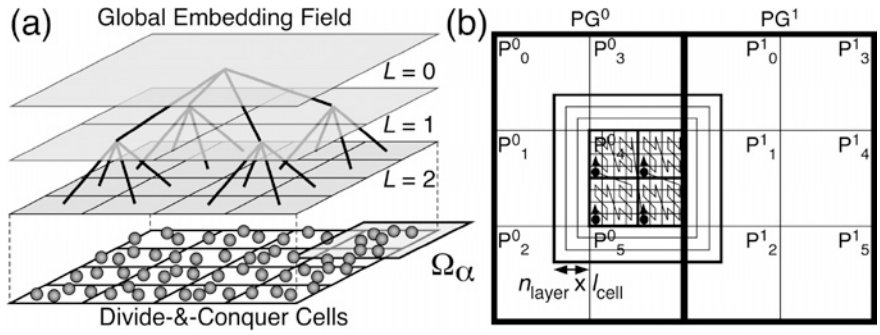

Fig. 1 (a) Schematic of embedded divide-and-conquer (EDC) algorithms. The physical space is subdivided into spatially localized cells, with local atoms constituting sub-problems (bottom), which are embedded in a global field (shaded) solved with a tree-based algorithm. (b) In tunable hierarchical cellular decomposition (HCD), the physical volume is subdivided into process groups, PG^{γ}, each of which is spatially decomposed into processes, P^{γ}_{π}. Each process consists of a number of computational cells (e.g. linked-list cells in MD or domains in EDC-DFT) of size l_{cell}, which are traversed concurrently by threads (denoted by dots with arrows) to compute blocks of cells. P^{γ}_{π} is augmented with n_{layer} layers of cached cells from neighbor processes.

2.2 Controlled Errors of Embedded Divide-and-Conquer Simulation Algorithms

A major advantage of the EDC simulation algorithms for automated hierarchical simulations and validation is the ease of codifying (i.e. turning into a coded representation, in terms of programs, which is mechanically executable by other program components) error management. The EDC algorithms have a well-defined set of localization parameters, with which the computational cost and the accuracy are controlled. Figures 2(a) and 2(b) show the rapid convergence of the EDC-DFT energy as a function of its localization parameters (the size of a domain and the length of buffer layers to augment each domain for avoiding artificial boundary effects). The EDC-DFT MD algorithm has also overcome the energy drift problem, which plagues most $O(N)$ DFT-based MD algorithms, especially with large basis sets ($> 10^4$ unknowns per electron, necessary for the transferability of accuracy; see Figure 2(c); Shimojo et al. 2005).

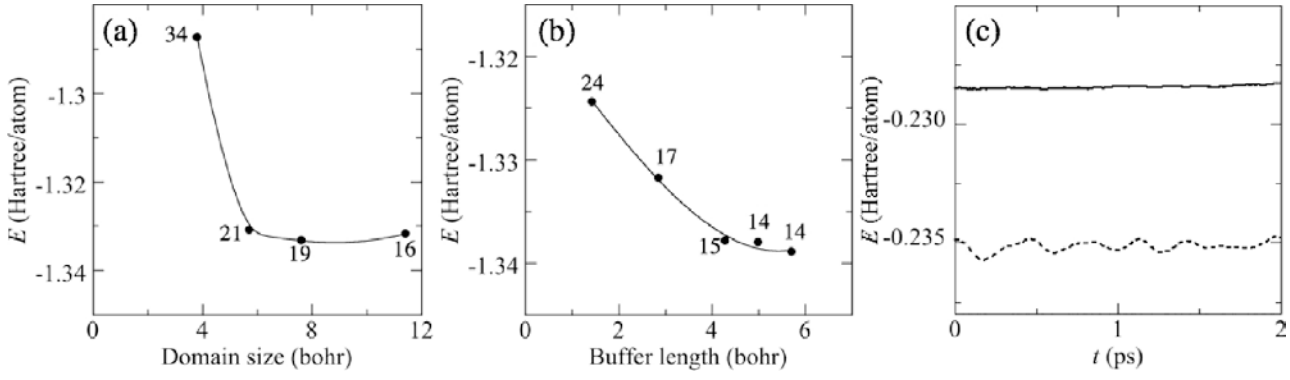

Fig. 2 Controlled convergence of the potential energy of amorphous CdSe by localization parameters: (a) domain size (with the buffer size fixed as 2.854 a.u.); (b) buffer length (with the domain size fixed as 11.416 a.u.). Numerals are the number of self-consistent iterations required for the convergence of the electron density within 10^{-4} of the bulk density. (c) Energy conservation in EDC-DFT based MD simulation of liquid Rb at 1400 K. The domain and buffer sizes are 16.424 and 8.212 a.u., respectively.

2.3 Adaptive Hierarchical Simulation Framework

Adaptive hierarchical simulation combines a hierarchy of MD algorithms (e.g. MRMD, F-ReaxFF, and EDC-DFT described above) to enable atomistic simulations that are otherwise too large to solve, while retaining QM accuracy (Broughton et al. 1999; Nakano et al. 2001; Ogata et al. 2001, 2004). The EDC framework achieves this by using less compute-intensive coarse simulation as an embedding field. We have developed an adaptive EDC hierarchical simulation framework, which embeds accurate but compute-intensive simulations in coarse simulation only when and where high fidelity is required. The hierarchical simulation framework consists of: 1) hierarchical division of the physical system into subsystems of decreasing sizes and increasing quality-of-solution (QoSn) requirements, $S_0 \supset S_1 \supset \ldots \supset S_n$; and 2) a suite of simulation services M_α ($\alpha = 0, 1, \ldots, n$) of ascending order of accuracy (e.g. MRMD < F-ReaxFF < EDC-DFT). In the additive hybridization scheme, an accurate estimate of the energy of the entire system is obtained from the recurrence relation (Dapprich et al. 1999; Ogata et al. 2001),

$$\begin{aligned} E_\alpha(S_i) &= E_{\alpha-1}(S_i) + E_\alpha(S_{i+1}) - E_{\alpha-1}(S_{i+1}) \\ &= E_{\alpha-1}(S_i) + \delta E_{\alpha\ \alpha-1}(S_{i+1}). \end{aligned} \quad (1)$$

This modular additive hybridization scheme not only allows the reuse of existing simulation codes but also minimizes the interdependence and communication between simulation modules. To further expose the data locality of hybrid QM/MD simulation, the EDC framework embeds EDC-DFT simulations of a number of domains within MD simulation of the total system:

$$\begin{aligned} E(\text{total}) &= E_{\text{MD}}(\text{total}) \\ &\quad + \sum_{\text{domain}} [E_{\text{QM}}(\text{domain}) - E_{\text{MD}}(\text{domain})] \\ &= E_{\text{MD}}(\text{total}) + \sum_{\text{domain}} \delta E_{\text{QM MD}}(\text{domain}). \end{aligned} \quad (2)$$

Traditionally, termination atoms are added to the QM and MD domains to minimize artificial boundary effects. However, the solution is sensitive to the choice of the termination atoms, and thus the domains need to be determined manually before the simulation (Ogata et al. 2001). The buffer-layer approach of the EDC-DFT algorithm considerably reduces this sensitivity, and accordingly we are among the first to automate the adaptive domain redefinition during simulations (Takemiya et al. 2006).

2.4 Graph-Based Event Tracking for Adaptive Hierarchical Simulations and Validation

An MD simulation method usually has a validation database that compares MD values with corresponding higher-accuracy QM values for a set of physical properties of selected atomic configurations (i.e. the training set). For example, the ReaxFF potential of RDX (1,3,5-trinitro-1,3,5-triazine) has an extensive training database of DFT calculations not only for 1,600 equilibrated molecular fragments but also for 40 key reaction pathways (i.e. chains of intermediate structures that interpolate the structures of reactant and product molecules; Strachan et al. 2003). Our QM/MD simulation ensures the overall accuracy by using MD simulation only within its range of validity. Our current adaptive hierarchical simulation manages the error based on a simple heuristic, i.e. the deviation of bond lengths from their equilibrium values, for which the MD interatomic potential has been trained (Takemiya et al. 2006). For tighter error management, we extend this approach by encoding the deviation of local topological structures of atoms from those in the training set, based on an abstraction of the structures as a graph and its shortest-path circuit analysis. Though the modeling error is readily quantified by $|\delta E_{\alpha/\alpha-1}(S_{i+1})|$ in equation (1) and can be reduced by incrementally enlarging the size of the high-accuracy subsystem once it is defined, the challenge here is to speculate the error in advance so as to minimize subsequent readjustments (if the subsystem is defined too small) or costly speculative high-accuracy simulations (if it is too large).

We abstract the structure of a material as a graph $G = (V, E)$, in which atoms constitute the set of vertices V, and the edge set E consists of chemical bonds. Bonds are defined between a pair of atoms for which Pauling's bond order is larger than a threshold value or the pair distance is less than a cutoff radius. Each vertex is labeled with its three-dimensional position and auxiliary annotations such as atomic species, and each bond with attributes such as bond length and chemical bond order. Given a vertex x and two of its neighbors w and y of G, a shortest-path circuit generated by triplet (w, x, y) is any closed path that contains edges (w, x) and (x, y) and has a shortest path (w, y) in graph $G-x$ (Franzblau 1991). The shortest-path circuit analysis has been used successfully to characterize topological order of amorphous materials and to identify and track topological defects such as dislocations (Figure 3; Rino et al. 1993; Nakano, Kalia, and Vashishta 1999; Szlufarska, Nakano, and Vashishta 2005; Branicio et al. 2006). Efficient algorithms with near linear scaling are essential for the graph analysis to be embedded as part of simulation in real time. We have developed a scalable parallel shortest-path circuit analysis algorithm with small memory usage, based on dual-tree expansion and spatial

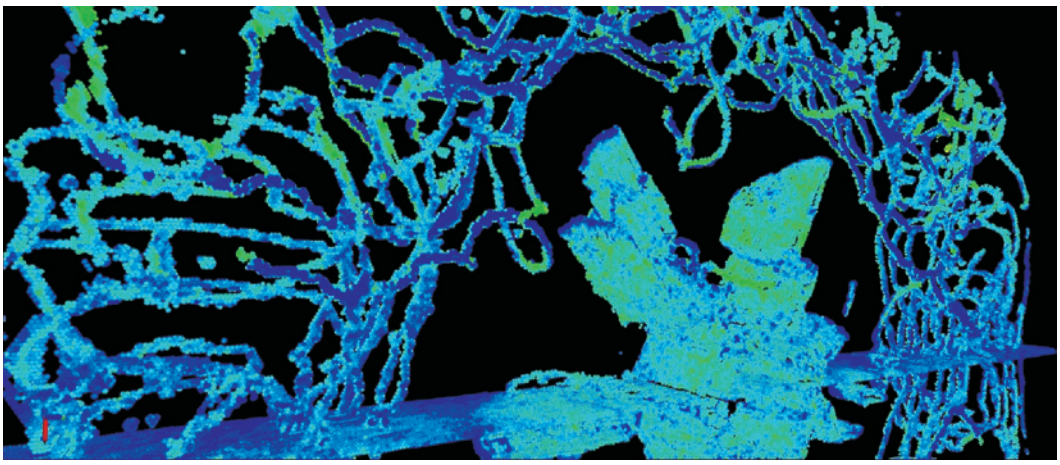

Fig. 3 Network of topological defects (or dislocations) during hypervelocity impact of aluminum nitride ceramic (Branicio et al. 2006). Only atoms that participate in non-6-membered circuits are visualized, where the color represents the pressure value. (A perfect crystal has only 6-membered circuits.)

hash-function tagging (SHAFT; Zhang et al. 2006). SHAFT utilizes the vertex-position label to design a compact, collision-free hash table, thereby avoiding the degradation of cache utilization for large graphs.

The graph abstraction is also used to track discrete events to automate the discovery of mechanisms, i.e. cause-effect relations on a sequence of well-delineated microscopic events that govern system-level macroscopic behavior. An example is a damage mechanism recently discovered by our MD simulation, in which the intersection of an elastic shock-wave front (detected as a discrete jump in pressure) and a high-pressure structural transformation front (the boundary between a sub-graph of vertex degree 4 and that of vertex degree 6) nucleates topological defects and eventually causes the fracture of ceramic under impact (Branicio et al. 2006). Large classical MD simulations involving multibillion atoms are performed to search for events within a wide solution space, and only when distinct events are detected by the graph analysis, QM simulations are invoked only where the local topological anomaly has been detected.

3 Tunable Hierarchical Cellular Decomposition Parallelization Framework

Data locality principles are key to developing a scalable parallel computing framework as well. We have developed a tunable hierarchical cellular decomposition (HCD) framework for mapping the $O(N)$ EDC algorithms onto massively parallel computers with deep memory hierarchies. The HCD maximally exposes data locality and exploits parallelism at multiple decomposition levels, while providing performance tunability (Whaley, Petitet, and Dongarra 2001) through a hierarchy of parameterized cell data/computation structures (Figure 1b). At the finest level, the EDC algorithms consist of computational cells—linked-list cells (which are identical to the octree leaf cells in the fast multipole method; Greengard and Rokhlin 1987) in MRMD and F-ReaxFF (Nakano et al. 2002), or domains in EDC-DFT (Shimojo et al. 2005). In the HCD framework, each compute node (often comprising multiple processors with shared memory) of a parallel computer is identified as a subsystem (P^{γ}_{π} in Figure 1b) in spatial decomposition, which contains a large number of computational cells. Our EDC algorithms are implemented as hybrid message passing interface (MPI; Gropp, Lusk, and Skjellum 1999) + shared memory (OpenMP; Chandra et al. 2000) programs, in which inter-node communication for caching and migrating atoms between P^{γ}_{π}'s is handled with messages, whereas loops over cells within each P^{γ}_{π} (or MPI process) are parallelized with threads (denoted as dots with arrows in Figure 1b). To avoid performance-degrading critical sections, the threads are ordered by blocking cells, so that the atomic n-tuples being processed by the threads share no common atom. On top of computational cells, cell-blocks, and spatial-decomposition subsystems, the HCD framework introduces a coarser level of decomposition by defining process groups ($PG^{\gamma} = \cup_{\pi} P^{\gamma}_{\pi}$ in Figure 1b) as MPI communicators (within a tightly coupled parallel computer) or grid remote procedure calls (Tanaka et al. 2003; on a grid of clusters).

Our programs are designed to minimize global operations across PG^{γ}'s and to overlap computations with inter-group communications (Kikuchi et al. 2002). For

example, the potential energy is computed locally within each group and the global sum is computed only when it needs to be reported to the user. Also our spatial decomposition scheme splits the computations on each processor into those involving only interior linked-list cells and those involving boundary cells. The interior computation is then fully overlapped with the communication of the boundary data. The effect of these latency-hiding techniques on performance is most noticeable on grid environments, since the communication overhead is already very small on each parallel supercomputer as shown in Section 4.

The cellular data structure offers an effective abstraction mechanism for performance optimization. We optimize both data layouts (atoms are sorted according to their cell indices and the linked lists) and computation layouts (force computations are re-ordered by traversing the cells according to a spacefilling curve, a mapping from the 3-D space to a 1-D list). Cells are traversed along either a Morton curve (Figure 1b) or a Hilbert curve (Moon et al. 2001). In a multi-threading case, the Morton curve ensures maximal separation between the threads and thus eliminates critical sections. Furthermore, the cell size is made tunable to optimize the performance. There is also a trade-off between spatial-decomposition/message-passing and threads parallelisms in the hybrid MPI+OpenMP programs (Henty 2000; Shan et al. 2002, 2003). While spatial decomposition involves extra computation on cached cells from neighbor subsystems, its disjoint memory subspaces are free from shared-memory protocol overhead. The computational cells are also used in our multilayer cellular decomposition scheme for inter-node caching of atomic n-tuple (n = 2–6) information (Figure 1b), where n changes dynamically in the MTS or MPCG algorithm (see Appendix). The Morton curve also facilitates a data compression algorithm based on data locality to reduce the I/O. The algorithm uses octree indexing and sorts atoms accordingly on the resulting Morton curve (Omeltchenko et al. 2000). By storing differences between successive atomic coordinates, the I/O requirement for a given error tolerance level reduces from $O(N\log N)$ to $O(N)$. An adaptive, variable-length encoding scheme is used to make the scheme tolerant to outliers and optimized dynamically. An order-of-magnitude improvement in the I/O performance was achieved for MD data with user-controlled error bound. The HCD framework includes a topology-preserving computational spatial decomposition scheme to minimize latency through structured message passing and load-imbalance/communication costs through a wavelet-based load-balancing scheme (Nakano and Campbell 1997; Nakano 1999).

High-end hierarchical simulations often run on thousands of processors for months. Grids of geographically distributed parallel supercomputers could satisfy the need of such 'sustained' supercomputing. In collaboration with scientists at the National Institute for Advanced Industrial Science and Technology (AIST) and the Nagoya Institute of Technology in Japan, we have recently proposed a sustainable grid supercomputing paradigm, in which supercomputers that constitute the grid change dynamically according to a reservation schedule as well as to faults (Takemiya et al. 2006). We use a hybrid grid remote procedure call (GridRPC) + MPI programming to combine flexibility and scalability. GridRPC enables asynchronous, coarse-grained parallel tasking and hides the dynamic, insecure and unstable aspects of the grid from programmers (we have used the Ninf-G GridRPC system, http://ninf.apgrid.org), while MPI supports efficient parallel execution on clusters.

4 Performance Tests

Scalability tests of the three parallel simulation algorithms, MRMD, F-ReaxFF and EDC-DFT, have been performed on a wide range of platforms, including the 10,240-processor SGI Altix 3000 at the NASA Ames Research Center, the 131,072-processor IBM BlueGene/L at the Lawrence Livermore National Laboratory (LLNL), and the 2048-processor AMD Opteron-based Linux cluster at the University of Southern California (USC). We have also tested our sustainable grid supercomputing framework for hierarchical QM/MD simulations on a grid of 6 supercomputer centers in the US and Japan. The codes have been ported without any modifications to all the platforms, except that only the pure MPI implementations have been run on BlueGene/L since it does not support OpenMP.

4.1 Experimental Platforms

The SGI Altix 3000 system, named Columbia, at NASA Ames uses the NUMAflex global shared-memory architecture, which packages processors, memory, I/O, interconnect, graphics, and storage into modular components called bricks (detailed information is found at http://www.nas.nasa.gov/Resources/Systems/columbia.html). The computational building block of Altix is the C-Brick, which consists of four Intel Itanium2 processors (in two nodes), local memory, and a two-controller application-specific integrated circuit called the Scalable Hub (SHUB). Each SHUB interfaces to the two CPUs within one node, along with memory, I/O devices, and other SHUBs. The Altix cache-coherency protocol implemented in the SHUB integrates the snooping operations of the Itanium2 and the directory-based scheme used across the NUMAflex interconnection fabric. A load/store cache miss causes the data to be communicated via the SHUB at the cache-line granularity and automatically replicated in the local cache.

The 64-bit Itanium2 processor operates at 1.5 GHz and is capable of issuing two multiply-add operations per cycle for a peak performance of 6 Gflops. The memory hierarchy consists of 128 floating-point registers and three on-chip data caches (32 KB L1, 256 KB L2, and 6 MB L3). The Itanium2 cannot store floating-point data in L1, making register loads and spills a potential source of bottlenecks; however, a relatively large register set helps mitigate this issue. The superscalar processor implements the Explicitly Parallel Instruction set Computing (EPIC) technology, where instructions are organized into 128-bit VLIW bundles. The Altix platform uses the NUMAlink3 interconnect, a high-performance custom network in a fat-tree topology, in which the bisection bandwidth scales linearly with the number of processors. Columbia runs 64-bit Linux version 2.4.21. Our experiments use a 6.4 TB parallel XFS file system with a 35-fiber optical channel connection to the CPUs.

Columbia is configured as a cluster of 20 Altix boxes, each with 512 processors and 1 TB of global shared-access memory. Of these 20 boxes, 12 are model 3700 and the remaining eight are BX2—a double-density version of the 3700. Four of the BX2 boxes are linked with NUMAlink4 technology to allow the global shared-memory constructs to significantly reduce inter-processor communication latency. This 2,048-processor subsystem within Columbia provides a 13 Tflops peak capability platform, and was the basis of the computations reported here.

The BlueGene/L (http://www.llnl.gov/ASC/platforms/bluegenel) has been developed by IBM and LLNL, and it uses a large number of low power processors coupled with powerful interconnects and communication schemes. The BlueGene/L comprises of 65,536 compute nodes (CN), each with two IBM PowerPC 440 processors (at 700 MHz clock speeds) and 512 MB of shared memory. The theoretical peak performance is 5.6 Gflops per CN, or 367 Tflops for the full machine. Each processor has a 32 KB instruction and data cache, a 2 MB L2 cache and a 4 MB L3 cache, which is shared with the other processor on the CN. Each CN has two floating-point units that can perform fused multiply-add operations. In its default mode (co-processor mode), one of the processors in the CN manages the computation, while the other processor manages the communication. In an alternative mode of operation (virtual mode), both processors can be used for computation. It uses a highly optimized lightweight Linux distribution and does not allow access to individual nodes.

The nodes are interconnected through multiple complementary high-speed low-latency networks, including a 3D torus network and a tree network. The CNs are connected as a 64 × 32 × 32 3-D torus, which is used for common inter-processor communications. The tree network is optimized for collective operations such as broadcast and gather. The point-to-point bandwidth of the 3-D torus network is 175 MB/s per link and 350 MB/s for the tree network.

The Opteron cluster used for the scalability test consists of 512 nodes, each with two dual-core AMD Opteron processors (at 2 GHz clock speeds) and 4 GB of memory (in total of 2,048 cores), which is part of a 5,472-processor Linux cluster at USC (http://www.usc.edu/hpcc/systems/l-overview.php). Each core has a 64 KB instruction cache, a 64 KB data cache, and a 1 MB L2 cache. A front side bus operating at 2 GHz provides a maximum I/O bandwidth of 24 GB/s. The floating-point part of the processor contains three units: a Floating Store unit that stores results to the Load/Store Queue Unit and Floating Add and Multiply units that can work in superscalar mode, resulting in two floating-point results per clock cycle. The 512 nodes are interconnected with Myrinet, which provides over 200 MB/s in a ping-pong experiment. Besides the high bandwidth, an advantage of Myrinet is that it entirely operates in the user space, thus avoiding operating systems interference and associated delays. This reduces the latency for small messages to 15 s.

4.2 Scalability Test Results

We have first tested the trade-off between MPI and OpenMP parallelisms on various shared-memory architectures. For example, we have compared different combinations of the number of OpenMP threads per MPI process, n_{td}, and that of MPI processes, n_p, while keeping $P = n_{td} \times n_p$ constant, on $P = 8$ processors in an 8-way 1.5 GHz Power4 node. The optimal combination of (n_{td}, n_p) with the minimum execution time is (1, 8) for the MRMD algorithm for an 8,232,000-atom silica material and is (4, 2) for the F-ReaxFF algorithm for a 290,304-atom RDX crystal. Since BlueGene/L does not support OpenMP, we will use $n_{td} = 1$ in the following performance comparisons.

Figure 4(a) shows the execution time of the MRMD algorithm for silica material as a function of the number of processors P on Columbia. We scale the problem size linearly with the number of processors, so that the number of atoms $N = 1{,}029{,}000P$ ($P = 1, \ldots, 1920$). In the MRMD algorithm, the interatomic potential energy is split into the long-range and short-range contributions, and the long-range contribution is computed every 10 MD time steps. The execution time increases only slightly as a function of P, and this signifies an excellent parallel efficiency. We define the speed of an MD program as a product of the total number of atoms and time steps executed per second. The constant-granularity speedup is the ratio between the speed of P processors and that of one processor. The parallel efficiency is the speedup divided by P. On 1920 processors, the isogranular

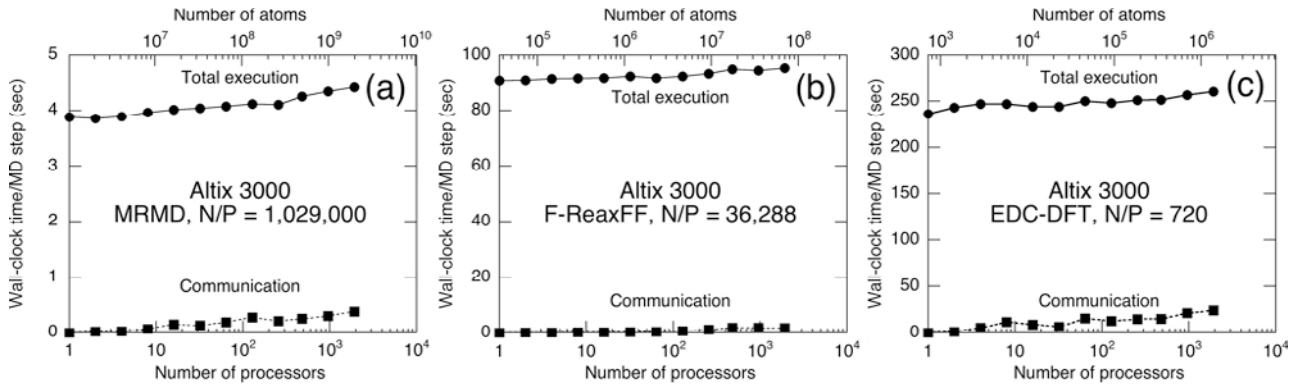

Fig. 4 Total execution (circles) and communication (squares) times per MD time step as a function of the number of processors *P* (= 1, …, 1920) of Columbia, for three MD simulation algorithms: (a) MRMD for 1,029,000 *P* atom silica systems; (b) F-ReaxFF MD for 36,288*P* atom RDX systems; and (c) EDC-DFT MD for 720*P* atom alumina systems.

parallel efficiency of the MRMD algorithm is 0.87. A better measure of the inter-box scaling efficiency based on NUMAlink4 is the speedup from 480 processors in 1 box to 1920 processors in 4 boxes, divided by the number of boxes. On 1920 processors, the inter-box scaling efficiency is 0.977. Also the algorithm involves very small communication time, see Figure 4(a).

Figure 4(b) shows the execution time of the F-ReaxFF MD algorithm for RDX material as a function of P, where the number of atoms is $N = 36{,}288P$. The computation time includes 3 conjugate gradient (CG) iterations to solve the electronegativity equalization problem for determining atomic charges at each MD time step. On 1,920 processors, the isogranular parallel efficiency of the F-ReaxFF algorithm is 0.953 and the inter-box scaling efficiency is 0.995.

Figure 4(c) shows the performance of the EDC-DFT based MD algorithm for 720P atom alumina systems. In the EDC-DFT calculations, each domain of size 6.66 × 5.76 × 6.06 $Å^3$ contains 40 electronic wave functions, where each wave function is represented on $28^3 = 21{,}952$ grid points. The execution time includes 3 self-consistent (SC) iterations to determine the electronic wave functions and the Kohn-Sham potential, with 3 CG iterations per SC cycle to refine each wave function iteratively. The largest calculation on 1,920 processors involves 1,382,400 atoms, for which the isogranular parallel efficiency is 0.907 and the inter-box scaling efficiency is 0.966.

The largest benchmark tests on Columbia include 18,925,056,000-atom MRMD, 557,383,680-atom F-ReaxFF, and 1,382,400-atom (121,385,779,200 electronic degrees-of-freedom) EDC-DFT calculations. We have observed perfect linear scaling for all the three algorithms, with prefactors spanning five orders-of-magnitude. The only exception is the F-ReaxFF algorithm below 100 million atoms, where the execution time scales even sublinearly. This is due to the decreasing communication overhead, which scales as $O((N/P)^{-1/3})$.

The EDC simulation algorithms are portable and have been run on various high-end computers including IBM BlueGene/L and dual-core AMD Opteron. Since the EDC framework exposes maximal locality, the algorithms scale well consistently on all platforms. Figure 5 compares the isogranular parallel efficiency of the F-ReaxFF MD algorithm for RDX material on Altix 3000, BlueGene/L, and Opteron. In Figure 5(a) for Altix 3000, the granularity is the same as that in Figure 4(b). Figure 5(b) shows the parallel efficiency as a function of P on BlueGene/L, where the number of atoms is $N = 36{,}288P$. On 65,536 BlueGene/L nodes (the computation uses only one processor per node in the co-processor mode), the isogranular parallel efficiency of the F-ReaxFF algorithm is over 0.998. Figure 5(c) shows the parallel efficiency of F-ReaxFF on Opteron, where the number of atoms is $N = 107{,}520P$. The measurements on Opteron have been curried out on one core per CPU for $P = 1$ and two cores per CPU for the other cases. The inter-core communication overhead (between $P = 1$ and 2) is negligible. The intranode bandwidth ($P = 4$) and network speed ($P \geq 8$) affect the total execution time by 4~12%. The sharp drop of efficiency in Figure 5(c) above 1000 cores may be attributed to interference with other jobs on the Linux cluster, which was in general use during the non-dedicated scalability test. The parallel efficiency is high for all three platforms, where the higher efficiency is achieved for a platform with the higher communication-bandwidth/processor-speed ratio (in descending order for BlueGene/L > Altix 3000 > Opteron/Myrinet).

Major design parameters for reactive and nonreactive MD simulations of materials include the number of

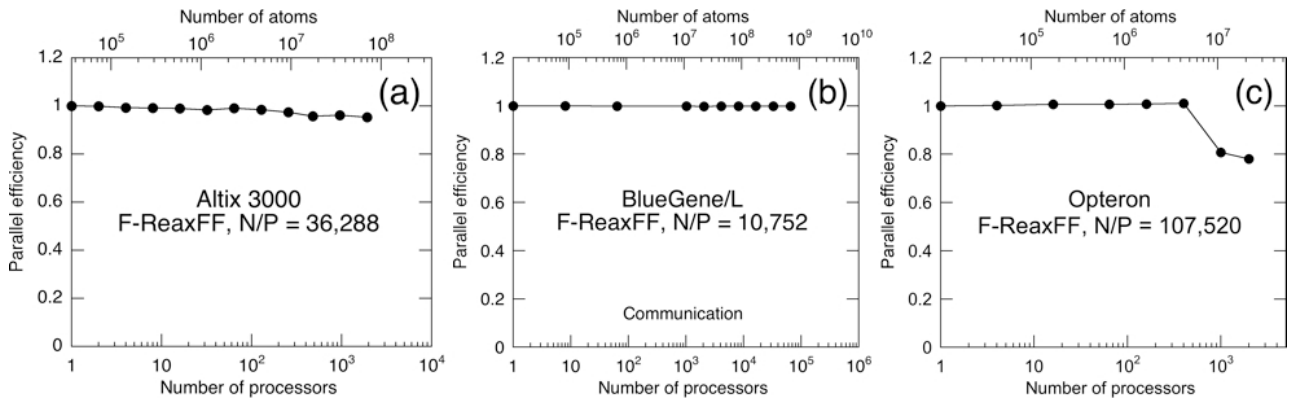

Fig. 5 Isogranular parallel efficiency of F-ReaxFF as a function of the number of processors *P* for RDX material: (a) the number of atoms per processor, *N*/*P* = 36,288 on Altix 3000 (*P* = 1, …,1920); (b) *N*/*P* = 10,752 on BlueGene/L (*P* = 1, …, 65536); and (c) *N*/*P* = 107,520 on an AMD Opteron cluster (*P* = 1, …, 2000).

atoms in the simulated material and the method to compute interatomic forces (classically in MRMD, semi-empirically in F-ReaxFF MD, or quantum-mechanically in EDC-DFT MD). Figure 6 shows a design-space diagram for MD simulations on BlueGene/L, Altix 3000, and Opteron. The largest benchmark tests in this study include 133,982,846,976-atom MRMD, 1,056,964,608-atom F-ReaxFF, and 11,796,480-atom (1,035,825,315,840 electronic degrees-of-freedom) EDC-DFT calculations on 65,536 dual-processor BlueGene/L nodes.

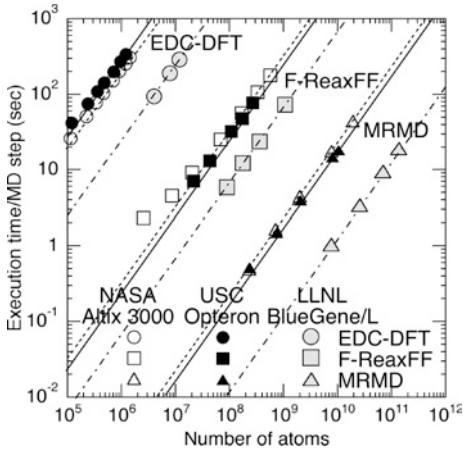

Fig. 6 Benchmark tests of reactive and nonreactive MD simulations on 1920 Itanium2 processors of the Altix 3000 at NASA (open symbols), 2000 Opteron processors at USC (solid symbols), and 65,536 dual-processor BlueGene/L nodes at LLNL (shaded symbols). The execution time per MD step is shown as a function of the number of atoms for: quantum-mechanical MD based on the embedded divide-and-conquer density functional theory (EDC-DFT, circles); fast reactive force-field MD (F-ReaxFF, squares); and nonreactive space–time multiresolution MD (MRMD, triangles). Lines show *O*(*N*) scaling.

Different characteristics of the MRMD, F-ReaxFF and EDC-DFT algorithms are reflected in their floating-point performances. The interatomic potential in MRMD is precomputed and tabulated as a function of the interatomic distance. The MRMD computation is thus predominantly table look-ups for atomic pairs and triplets. The F-ReaxFF algorithm, on the contrary, performs a large number of floating-point operations, but it involves more complex list management for atomic n-tuples (n = 2–6). In contrast to these particle-based algorithms, the EDC-DFT algorithm deals with wave functions on regular mesh points. In all the three $O(N)$ algorithms, however, the data layout and computations are highly irregular compared with their higher-complexity counterparts. The floating-point performances of the MRMD (N/P = 1,029,000), F-ReaxFF (N/P = 36,288) and EDC-DFT (N/P = 720) algorithms on 1,920 Itanium2 processors are 1.31, 1.07, and 1.49 Tflops, respectively, whereas the theoretical peak performance is 11.5 Tflops.

4.3 Grid Test Results

Using our sustainable grid supercomputing framework, we have achieved an automated execution of hierarchical QM/MD simulation on a grid consisting of 6 supercomputer centers in the US (USC and two NSF TeraGrid nodes at the Pittsburgh Supercomputing Center and the National Center for Supercomputing Applications) and Japan (AIST, University of Tokyo, and Tokyo Institute

of Technology; Takemiya et al. 2006). The simulation was sustained autonomously on ~700 processors for 2 weeks, involving in total of 150,000 CPU-hours, where the number of processors changed dynamically on demand and resources were allocated and migrated dynamically according to both reservations and unexpected faults.

5 Conclusions

We have motivated and supported the need for petaflops computing for advanced materials research, and have demonstrated that judicious use of divide-and-conquer algorithms and hierarchical parallelization frameworks should make these applications highly scalable on petaflops platforms. We have illustrated the value of this work with real-world experiments involving quantum-mechanical and molecular-dynamics simulations on high-end parallel supercomputers such as SGI Altix 3000, IBM BlueGene/L and AMD Opteron, as well as on a grid of globally distributed parallel supercomputers.

We are currently applying the de novo hierarchical simulation framework to study deformation and damage mechanisms of nanophase ceramics and nanoenergetic materials in harsh environments, thereby assisting the design of superhard, tough and damage-tolerant nanomaterials as well as nanoenergetic materials with high specific impact and reduced sensitivity.

One application is hypervelocity impact damage of advanced ceramics [aluminum nitride (AlN), silicon carbide, and alumina; Branicio et al. 2006], for which we have recently performed 500 million-atom MD simulations. The simulation has revealed atomistic mechanisms of fracture accompanying structural phase transformation in AlN under hypervelocity impact at 15 km/s. We are extending these classical MD simulations to those involving surface chemical reactions under high temperatures and flow velocities relevant to micrometeorite impact damages to the thermal and radiation protection layers of aerospace vehicles, understanding of which is essential for safer space flights.

Another application is the combustion of nanoenergetic materials. We have performed 1.3 million-atom F-ReaxFF MD simulations to study shock-initiated detonation of RDX (1,3,5-trinitro-1,3,5-triazine, $C_3N_6O_6H_6$) matrix embedded with aluminum nanoparticles (n-Al) on 1024 dual-core Opteron processors at the Collaboratory for Advanced Computing and Simulations of USC (Figure 7). In the simulation, a $320 \times 210 \times 204$ Å^3 RDX/n-Al composite is impacted by a plate at a velocity of 5 km/s. The simulation has revealed atomistic processes of shock compression and subsequent explosive reaction. Strong attractive forces between oxygen and aluminum atoms break N–O and N–N bonds in the RDX and, subsequently, the dissociated oxygen atoms and NO molecules oxidize Al, which has also been observed in our DFT-based MD simulation (Umezawa et al. 2006).

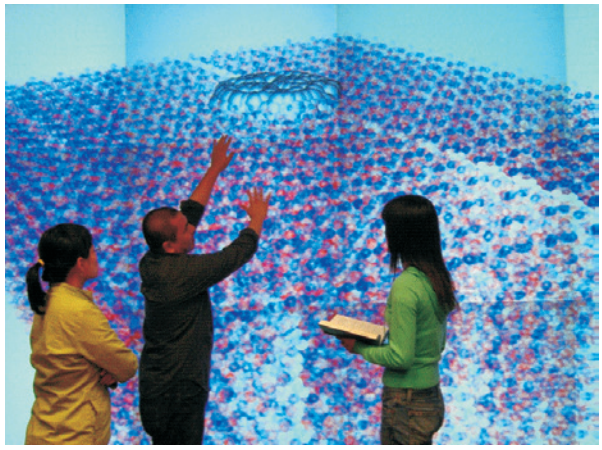

Fig. 7 F-ReaxFF MD simulation of n-Al/RDX simulation shown on a tiled display at the Collaboratory for Advanced Computing and Simulations of USC. Visualization software, with embedded graph analysis algorithms (Zhang et al. 2006), has been developed by Sharma et al. (2003).

De novo hierarchical simulations also have broad applications in nanoelectronics (Ogata et al. 2004). The hybrid QM/MD simulation on the US-Japan Grid described in Section 4 has studied the SIMOX (separation by implantation by oxygen) technique for fabricating high speed and low power-consumption semiconductor devices. The simulation of the implantation of oxygen atoms toward a Si substrate has revealed a strong dependence of the oxygen penetration depth on the incident beam position, which should be taken into consideration in extending the SIMOX technique to lower incident energies.

These applications on high-end computing platforms today are paving the way for predictive, first-principles simulation-based sciences in coming years (Emmott and Rison 2006).

Appendix: Embedded Divide-and-Conquer Simulation Algorithms

This Appendix describes computational characteristics of the three embedded divide-and-conquer simulation algorithms that are used in our adaptive hierarchical simulations: 1) MRMD: space–time multiresolution molecular dynamics; 2) F-ReaxFF: fast reactive force-field molecular dynamics; and 3) EDC-DFT: embedded divide-and-conquer density functional theory on adaptive multigrids for quantum-mechanical molecular dynamics.

Algorithm 1—MRMD: Space–Time Multiresolution Molecular Dynamics

MRMD is used as a template for developing broad particle and continuum simulation algorithms. The MD approach follows the time evolution of the positions, $\mathbf{r}^N = \{\mathbf{r}_i | i = 1, \ldots, N\}$, of N atoms by solving coupled ordinary differential equations (Nakano et al. 2002). Atomic force law is mathematically encoded in the interatomic potential energy $E_{\mathrm{MD}}(\mathbf{r}^N)$, which is often an analytic function $E_{\mathrm{MD}}(\{\mathbf{r}_{ij}\}, \{\mathbf{r}_{ijk}\})$ of atomic pair, $\mathbf{r}_{ij}$, and triplet, $\mathbf{r}_{ijk}$, positions. For the long-range electrostatic interaction, we use the fast multipole method (FMM) to reduce the $O(N^2)$ computational complexity of the N-body problem to $O(N)$ (Greengard and Rokhlin 1987). In the FMM, the physical system is recursively divided into subsystems to form an octree data structure, and the electrostatic field is computed recursively on the octree with $O(N)$ operations, while maintaining spatial locality at each recursion level. Our scalable parallel implementation of the FMM has a unique feature to compute atomistic stress tensor components based on a complex charge method (Ogata et al. 2003). MRMD also utilizes temporal locality through multiple time stepping, which uses different force-update schedules for different force components (Martyna et al. 1994; Schlick et al. 1999; Nakano et al. 2002). Specifically, forces from neighbor atoms are computed at every MD step, whereas forces from farther atoms are updated less frequently. For parallelization, we use spatial decomposition. The total volume is divided into P subsystems of equal volume, and each subsystem is assigned to a node in an array of P compute nodes. To calculate the force on an atom in a subsystem, the coordinates of the atoms in the boundaries of neighbor subsystems are 'cached' from the corresponding nodes. After updating the atom positions due to time stepping, some atoms may have moved out of its subsystem. These atoms are 'migrated' to the proper neighbor nodes. With spatial decomposition, the computation scales as N/P, while communication scales as $(N/P)^{2/3}$. The FMM incurs an $O(\log P)$ overhead, which is negligible for coarse-grained ($N/P >> P$) applications.

Algorithm 2—F-ReaxFF: Fast Reactive Force-Field Molecular Dynamics

In the past 5 years, we have developed a first principles-based reactive force-field (ReaxFF) approach to significantly reduce the computational cost of simulating chemical reactions (van Duin et al. 2001; Strachan et al. 2003). However, its parallelization has seen only limited success, with the previously largest ReaxFF MD involving $N < 10^4$ atoms. We have developed F-ReaxFF to enable ReaxFF MD involving 10^9 atoms (Nakano et al. 2006; Vashishta, Kalia, and Nakano 2006). The variable N-charge problem in ReaxFF amounts to solving a dense linear system of equations to determine atomic charges $\{q_i | i = 1, \ldots, N\}$ at every MD step (Rappe and Goddard 1991; Campbell et al. 1999). F-ReaxFF reduces its $O(N^3)$ complexity to $O(N)$ by combining the FMM based on spatial locality and iterative minimization to utilize the temporal locality of the solution. To accelerate the convergence, we use a multilevel preconditioned conjugate-gradient (MPCG) method that splits the Coulomb-interaction matrix into short- and long-range parts and uses the sparse short-range matrix as a preconditioner (Nakano 1997). The extensive use of the sparse preconditioner enhances the data locality and thereby improves the parallel efficiency. The chemical bond order B_{ij} is an attribute of atomic pair (i, j) and changes dynamically adapting to the local environment. In ReaxFF, the potential energy $E_{\mathrm{ReaxFF}}(\{\mathbf{r}_{ij}\},\{\mathbf{r}_{ijk}\},\{\mathbf{r}_{ijkl}\},\{q_i\},\{B_{ij}\})$ between atomic pairs $\mathbf{r}_{ij}$, triplets $\mathbf{r}_{ijk}$, and quadruplets $\mathbf{r}_{ijkl}$ depends on the bond orders of all constituent atomic pairs. Force calculations in ReaxFF thus involve up to atomic 6-tuples due to chain-rule differentiations through B_{ij}. To efficiently handle the multiple interaction ranges, the parallel F-ReaxFF algorithm employs a multilayer cellular decomposition scheme for caching atomic n-tuples (n = 2–6) (Nakano et al. 2006).

Algorithm 3—EDC-DFT: Embedded Divide-and-Conquer Density Functional Theory on Adaptive Multigrids for Quantum-Mechanical Molecular Dynamics

EDC-DFT describes chemical reactions with a higher quantum-mechanical accuracy than ReaxFF. The DFT problem is formulated as a minimization of the energy functional $E_{\mathrm{QM}}(\mathbf{r}^N, \psi^{N\mathrm{el}})$ with respect to electronic wave functions (or Kohn-Sham orbitals) $\psi^{N\mathrm{el}}(\mathbf{r}) = \{\psi_n(\mathbf{r}) | n = 1, \ldots, N_{\mathrm{el}}\}$, subject to orthonormality constraints (N_{el} is the number of wave functions on the order of N; Hohenberg and Kohn 1964), The data locality principle called quantum nearsightedness (Kohn 1996) in DFT is best implemented with a divide-and-conquer algorithm (Yang 1991; Yang and Lee 1995), which naturally leads to $O(N)$ DFT calculations (Goedecker 1999). However, it is only in the past several years that $O(N)$ DFT algorithms, especially with large basis sets ($> 10^4$ unknowns per electron, necessary for the transferability of accuracy), have attained controlled error bounds, robust convergence properties, and energy conservation during MD simulations, to make large DFT-based MD simulations practical (Fattebert and Gygi 2004; Shimojo et al. 2005). We have designed an embedded divide-and-conquer density functional theory (EDC-DFT) algorithm, in which a hierarchical grid technique combines multigrid preconditioning and adaptive fine mesh generation (Shimojo et al. 2005). The EDC-DFT algorithm represents the physical system as a union

of overlapping spatial domains, $\Omega = \cup_\alpha \Omega_\alpha$, and physical properties are computed as linear combinations of domain properties. For example, the electronic density is expressed as $\rho(\mathbf{r}) = \Sigma_\alpha p^\alpha(\mathbf{r}) \Sigma_n f_n^\alpha |\psi_n^\alpha(\mathbf{r})|^2$, where $p^\alpha(\mathbf{r})$ is a support function that vanishes outside the α-th domain Ω_α, and f_n^α and $\psi_n^\alpha(\mathbf{r})$ are the occupation number and the wave function of the n-th Kohn-Sham orbital in Ω_α. The domains are embedded in a global Kohn-Sham potential, which is a functional of $\rho(\mathbf{r})$ and is determined self-consistently with $\{f_n^\alpha, \psi_n^\alpha(\mathbf{r})\}$. We use the multigrid method to compute the global potential in $O(N)$ time. The DFT calculation in each domain is performed using a real-space approach (Chelikowsky et al. 2000), in which electronic wave functions are represented on grid points. The real-space grid is augmented with coarser multigrids to accelerate the iterative solution. Furthermore, a finer grid is adaptively generated near every atom, in order to accurately operate ionic pseudopotentials for calculating electron-ion interactions. The EDC-DFT algorithm on the hierarchical real-space grids is implemented on parallel computers based on spatial decomposition. Each compute node contains one or more domains of the EDC algorithm. Then only the global density but not individual wave functions needs to be communicated. The resulting large computation/communication ratio makes this approach highly scalable.

Acknowledgments

The work at the University of Southern California (USC) was partially supported by AFOSR-DURINT, ARO-MURI, DOE, DTRA, NSF, and Chevron-CiSoft. The work of LHY was supported under the auspices of the U.S. Department of Energy by the University of California Lawrence Livermore National Laboratory (LLNL) under contract No. W-7405-ENG-48. Benchmark tests were performed using the Columbia supercomputer at the NASA Ames Research Center, the BlueGene/L at LLNL, the NSF TeraGrid, the 5472-processor (15.8 Tflops) Linux cluster at USC, the 11 Tflops Itanium/Opteron cluster at the National Institute for Advanced Industrial Science and Technology (AIST), and Linux clusters at the University of Tokyo and Tokyo Institute of Technology. Programs have been developed using the 2048-processor (4 Tflops) Opteron/Xeon/Apple G5 cluster at the Collaboratory for Advanced Computing and Simulations of USC. The authors thank Davin Chan, Johnny Chang, Bob Ciotti, Edward Hook, Art Lazanoff, Bron Nelson, Charles Niggley, and William Thigpen for technical discussions on Columbia, Shuji Ogata, Satoshi Sekiguchi, Hiroshi Takemiya, and Yoshio Tanaka for their collaboration on the adaptive QM/MD simulations on the US-Japan Grid, and Bhupesh Bansal, Paulo S. Branicio, and Cheng Zhang for their contribution on graph-based data mining.

Author Biographies

Aiichiro Nakano is a professor of computer science with joint appointments in physics & astronomy, chemical engineering & materials science, and the Collaboratory for Advanced Computing and Simulations at the University of Southern California. He received a Ph.D. in physics from the University of Tokyo, Japan, in 1989. He has authored 225 refereed articles, including 140 journal papers, in the areas of scalable scientific algorithms, grid computing on geographically distributed parallel computers, and scientific visualization. He is a recipient of: the National Science Foundation Career Award (1997); Louisiana State University (LSU) Alumni Association Faculty Excellence Award (1999); LSU College of Basic Sciences Award of Excellence in Graduate Teaching (2000); the Best Paper Award at the IEEE/ACM Supercomputing 2001 Conference; Best Paper at the IEEE Virtual Reality Conference (2002); and Okawa Foundation Faculty Research Award (2003). He is a member of IEEE, ACM, APS, and MRS.

Rajiv K. Kalia is a professor in the Department of Physics & Astronomy with joint appointments in chemical engineering & materials science, computer science, and the Collaboratory for Advanced Computing and Simulations. He graduated with a Ph.D. in physics from Northwestern University in 1976. His expertise is in the area of multiscale simulations involving atomistic, mesoscale and continuum approaches on a GRID of distributed parallel supercomputers and immersive and interactive virtual environment. He has authored over 250 papers, which include ultrascale simulations of: sintering, crack growth, stress corrosion, nanoindentation, friction, and hypervelocity impact in ceramics, nanophase composites, and nanoscale devices; oxidation and structural transformations in metallic and semiconductor nanoparticles; and structure/dynamics of self-assembled monolayers. He is a recipient of: FOM Fellowship, the Netherlands (2000); LSU Distinguished Faculty Award (1999); Sustained Excellence Performance Award in Ultra Dense, Ultra Fast Computing Components, DARPA (1997); Japan Society for the Promotion of Science Fellowship Award (1991); and Brazilian Science Research Council Award (1986).

Ken-ichi Nomura is a Ph.D. candidate in the Department of Physics & Astronomy at the University of Southern California (USC). He received an M.S. in physics from Niigata University, Japan, in 2002 and an M.S. in computer science from USC in 2006. He has developed a scalable parallel algorithm for chemically reactive molecular-dynamics simulation and has performed multimillion-atom reactive MD simulations of nanostructured energetic materials. He has authored 14 papers.

Ashish Sharma is a postdoctoral researcher in the Department of Biomedical Informatics of the Ohio State University Medical Center. He received a Ph.D. in computer science from the University of Southern California, Los Angeles, CA in 2005. He received an M.S. in computer science from Louisiana State University, Baton Rouge, LA in 2002. His research interests are in scientific visualization, data mining, high performance computing, and biomedical imaging and grid technologies. He has published 14 papers including 9 journal papers. He is a recipient of the USC Outstanding Student Research Award (2002) and the Best Paper at the IEEE Virtual Reality Conference (2002).

Priya Vashishta received a Ph.D. in physics from the Indian Institute of Technology, Kanpur, India, in 1967. He is the founding Director of the Collaboratory for Advanced Computing and Simulations, with multidisciplinary appointments in the School of Engineering and College of Letters, Arts and Sciences, at the University of Southern California. Before joining USC, he was: Cray Research Professor of Computational Sciences at Louisiana State University, where he was the founding Director of the Concurrent Computing Laboratory for Materials Simulations; and Senior Scientist at the Argonne National Laboratory, where he was the Director of the Solid State Science Division from 1979 to 1982. His research interests include high performance computing and visualization to carry out large multiscale simulations of materials and processes, nanoscale systems, and info-bio-nano interface on massively parallel and distributed computers. His awards and honors include: University of Chicago Award for Distinguished Performance at ANL (1976); Japan Society for the Promotion of Science Senior Fellowship (1985 and 1989); Brazilian Science Research Council Fellowship (1985); United Nations Development Program Fellowship (1990); Sustained Excellence Performer Award in Ultra Dense, Ultra Fast Computing Components, DARPA (1997); Fellow of the American Physical Society (1999); and IEEE/ACM Supercomputing Best Paper Award (2001).

Fuyuki Shimojo is an associate professor in the Department of Physics at Kumamoto University, Japan. He received a Ph.D. in physics from Niigata University, Japan, in 1993. His main research interests are condensed matter physics, computational materials science, and parallel algorithms for large-scale molecular simulations. He has authored 132 refereed articles, including 90 journal papers, in these areas. He is a member of APS and JPS.

Adrianus (Adri) C. T. van Duin is a Senior Research Fellow at the California Institute of Technology, working as Director of Force Field and Simulation Technology at the Material and Process Simulation Center. He received a Ph.D. in chemistry from the Delft University of Technology, the Netherlands, in 1996. He has authored over 50 refereed journal papers. His main area of research involves the development and application of the ReaxFF reactive force field, which has been used to simulate chemical reactions in a wide range of materials (hydrocarbons, proteins, high-energy materials, metals, metal oxides/hydrides/carbides and semiconductors). In addition, he has substantial experience in the field of organic geochemistry, including oil formation and organic/water/mineral interactions. He obtained Marie Curie and Royal Society Fellowships for his organic geochemical and force field development work.

William A. Goddard, III, is Charles and Mary Ferkel Professor of Chemistry, Materials Science, and Applied Physics, and Director of Materials and Process Simulation Center (MSC) at the California Institute of Technology (Caltech). He received his Ph.D. in engineering science (minor in physics) from Caltech in 1964 and has been on the faculty at Caltech since then. He is a member of National Academy of Science (1984), a member of International Academy of Quantum Molecular Science (1988), a fellow of American Physical Society (1988), and a fellow of American Association for the Advancement of Science (1990). He received the ACS Award for Computers in Chemistry (1988), the Richard M. Badger Teaching Prize in Chemistry, Caltech (1995), the Feynman Prize for Nanotechnology Theory (1999), the NASA Space Sciences Award (2000), and the Richard Chase Tolman Prize from the ACS (2000). He was named by ISI as a most highly cited chemist for 1981 to 1999 and was winner of the 2002 Prize in Computational Nanotechnology Design from the Institute for Molecular Manufacturing. He was cofounder of Molecular Simulations Inc. (now named Accelrys) (1984), Schrödinger Inc. (1990), Systine Inc. (2001), Eidogen Inc (now Eidogen-Sertanty) (2000), Allozyne Inc. (2004), and Qateomix Inc. (2005). His current research interests include: 1) new methodology for quantum chemistry, force fields, molecular dynamics, mesoscale dynamics, statistical mechanics; 2) applications of atomistic simulations to chemical, biological, and materials systems, including catalysis, polymers, semiconductors, ceramics, and metal alloys; 3) protein structure prediction, drug design, incorporation of non-natural amino acids; 4) industrial problems in catalysis, polymers, fuel cells, energetic materials, and drug design; 5) computational nanotechnology with applications to nanoelectronics. His has over 687 research publications.

Rupak Biswas is the acting chief of the NASA Advanced Supercomputing (NAS) Division at Ames Research Center. In this capacity, he oversees the full range of

high-performance computing operations, services, and R&D for NASA's primary supercomputing center. He received his Ph.D. in computer science from Rensselaer Polytechnic Institute in 1991, and has been with NASA ever since. He is an internationally recognized expert in parallel programming paradigms; benchmarking and performance characterization; partitioning and load balancing techniques; and scheduling algorithms. He has published more than 130 technical papers in journals and peer-reviewed conferences, and received many NASA awards and two Best Paper prizes (SC99, SC2000). He served as a member of the inter-agency High End Computing Revitalization Task Force, mission partner representative on the DARPA High Productivity Computing Systems Project, one of five expert panelists to assess all high-end computing R&D activities in Japan, and is currently a Director on the OpenMP Architecture Review Board.

Deepak Srivastava is a senior scientist and group lead of computational nanotechnology investigations at the NASA Ames Research Center. He received a Ph.D. in physics from the University of Florida. His research interests focus on simulation of properties of nanoscale materials. His accomplishments include co-winner of Feynman Prize in Nanotechnology (1997), NASA Ames CC Award (1998), Veridian Medals for Technical Paper (1999), NASA Group Excellence Award (2000), The Eric Reissener Medal (2002), and CSC Award for Technical Excellence (2003). Dr. Srivastava serves as associate editor of two peer reviewed journals: Journal of Nanoscience and Nanotechnology and Computer Modeling in Engineering and Sciences, and Editorial Board Member of Composites Science and Technology. He co-founded a nanotechnology start-up company, which was bought by another start-up company, Nanostellar, which focuses on rational design and synthesis of nanomaterials for catalysis and energy sector. He has chaired or co-chaired many conferences on nanotechnology and nanomaterials, and is a founding board member of CANEUS (Canada-Europe-US) Organization for promotion of nanotechnology for aerospace and defense.

Lin H. Yang is a physicist at Lawrence Livermore National Laboratory. He received a Ph.D. in physics from the University of California, Davis, in 1989. He has authored more than 80 refereed journal papers in the areas of large-scale molecular dynamics and quantum molecular dynamics simulations. His current research focus is to develop petascale quantum molecular dynamics simulation algorithm for metals. He is a member of American Physical Society and currently serves as a Member at Large at the APS California Section.

References

Abraham, F. F., Walkup, R., Gao, H. J., Duchaineau, M., De la Rubia, T. D., and Seager, M. (2002). Simulating materials failure by using up to one billion atoms and the world's fastest computer: Brittle fracture, *Proceedings of the National Academy of Sciences of the United States of America*, **99**(9): 5777–5782.

Allen, G., Dramlitsch, T., Foster, I., Goodale, T., Karonis, N., Ripeanu, M., Seidel, E., and Toonen, B. (2001). Supporting efficient execution in heterogeneous distributed computing environments with Cactus and Globus, In *Proceedings of Supercomputing 2001*, Denver: ACM.

Branicio, P. S., Kalia, R. K., Nakano, A., and Vashishta, P. (2006). Shock-induced structural phase transition, plasticity, and brittle cracks in aluminum nitride ceramic, *Physical Review Letters*, **96**(6): 065502.

Broughton, J. Q., Abraham, F. F., Bernstein, N., and Kaxiras, E. (1999). Concurrent coupling of length scales: Methodology and application, *Physical Review B*, **60**(4): 2391–2403.

Campbell, T. J., Kalia, R. K., Nakano, A., Vashishta, P., Ogata, S., and Rodgers, S. (1999). Dynamics of oxidation of aluminum nanoclusters using variable charge molecular-dynamics simulations on parallel computers, *Physical Review Letters*, **82**(24): 4866–4869.

Car, R. and Parrinello, M. (1985). Unified approach for molecular dynamics and density functional theory, *Physical Review Letters*, **55**: 2471–2474.

Chandra, R., Menon, R., Daqum, L., Kohr, D., Maydan, D., and McDonald, J. (2000). *Parallel Programming in OpenMP*. San Francisco: Morgan Kaufmann.

Chelikowsky, J. R., Saad, Y., Ögüt, S., Vasiliev, I., and Stathopoulos, A. (2000). Electronic structure methods for predicting the properties of materials: grids in space, *Physica Status Solidi (b)*, **217**: 173–195.

Dapprich, S., Komáromi, I., Byun, K. S., Morokuma, K., and Frisch, M. J. (1999). A new ONIOM implementation in Gaussian 98. I. The calculation of energies, gradients, vibrational frequencies, and electric field derivatives, *J. Mol. Struct. (Theochem)*, **461–462**: 1–21.

Dongarra, J. J. and Walker, D. W. (2001). The quest for petascale computing, *Computing in Science and Engineering*, **3**(3): 32–39.

Emmott, S. and Rison, S. (2006). *Towards 2020 Science*, Cambridge, UK: Microsoft Research.

Fattebert, J.-L. and Gygi, F. (2004). Linear scaling first-principles molecular dynamics with controlled accuracy, *Computer Physics Communications*, **162**(1): 24–36.

Foster, I. and Kesselman, C. (2003). *The Grid 2: Blueprint for a New Computing Infrastructure*: Morgan Kaufmann.

Franzblau, D. S. (1991). Computation of ring statistics for network models of solids, *Physical Review B*, **44** (10): 4925–4930.

Goedecker, S. (1999). Linear scaling electronic structure methods, *Reviews of Modern Physics*, **71**: 1085–1123.

Greengard, L. and Rokhlin, V. (1987). A fast algorithm for particle simulations, *Journal of Computational Physics*, **73**: 325–348.

Gropp, W., Lusk, E., and Skjellum, A. (1999). *Using MPI*, 2nd edition. Cambridge, MA: MIT Press.

Gygi, F., Draeger, E., de Supinski, B. R., Yates, R. K., Franchetti, F., Kral, S., Lorenz, J., Ueberhuber, C. W., Gunneis, J. A., and Sexton, J. C. (2005). Large-scale first-principles molecular dynamics simulations on the BlueGene/L platform using the Qbox code, In *Proceedings of Supercomputing 2005*, Seattle: ACM.

Henty, D. S. (2000). Performance of hybrid message-passing and shared-memory parallelism for discrete element modeling, In *Proceedings of Supercomputing 2000*, Los Alamitos: IEEE.

Hohenberg, P. and Kohn, W. (1964). Inhomogeneous electron gas, *Physical Review*, **136**: B864–B871.

Ikegami, T., Ishida, T., Fedorov, D. G., Kitaura, K., Inadomi, Y., Umeda, H., Yokokawa, M., and Sekiguchi, S. (2005). Full electron calculation beyond 20,000 atoms: ground electronic state of photosynthetic proteins, In *Proceedings of Supercomputing 2005*, Seattle: ACM.

Kadau, K., Germann, T. C., Lomdahl, P. S., and Holian, B. L. (2002). Microscopic view of structural phase transitions induced by shock waves, *Science*, **296**(5573): 1681–1684.

Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999). NAMD2: greater scalability for parallel molecular dynamics, *Journal of Computational Physics*, **151** (1): 283–312.

Kendall, R. A., Apra, E., Bernholdt, D. E., Bylaska, E. J., Dupuis, M., Fann, G. I., Harrison, R. J., Ju, J. L., Nichols, J. A., Nieplocha, J., Straatsma, T. P., Windus, T. L., and Wong, A. T. (2000). High performance computational chemistry: An overview of NWChem a distributed parallel application, *Computer Physics Communications*, **128**(1–2): 260–283.

Kikuchi, H., Kalia, R. K., Nakano, A., Vashishta, P., Iyetomi, H., Ogata, S., Kouno, T., Shimojo, F., Tsuruta, K., and Saini, S. (2002). Collaborative simulation Grid: multiscale quantum-mechanical/classical atomistic simulations on distributed PC clusters in the US and Japan, In *Proceedings of Supercomputing 2002*, Los Alamitos: IEEE.

Kohn, W. (1996). Density functional and density matrix method scaling linearly with the number of atoms, *Physical Review Letters*, **76**: 3168–3171.

Martyna, G. J., Tuckerman, M. E., Tobias, D. J., and Klein, M. L. (1996). Explicit reversible integrators for extended systems dynamics, *Mol. Phys.*, **87**: 1117–1157.

Moon, B., Jagadish, H. V., Faloutsos, C., and Saltz, J. H. (2001). Analysis of the clustering properties of the Hilbert space-filling curve, *IEEE Transactions on Knowledge and Data Engineering*, **13**(1): 124–141.

Nakano, A. (1997). Parallel multilevel preconditioned conjugate-gradient approach to variable-charge molecular dynamics, *Computer Physics Communications*, **104**: 59–69.

Nakano, A. (1999). Multiresolution load balancing in curved space: the wavelet representation, *Concurrency: Practice and Experience*, **11**: 343–353.

Nakano, A. and Campbell, T. J. (1997). An adaptive curvilinear-coordinate approach to dynamic load balancing of parallel multiresolution molecular dynamics, *Parallel Computing*, **23**: 1461–1478.

Nakano, A., Kalia, R. K., and Vashishta, P. (1999). Scalable molecular-dynamics, visualization, and data-management algorithms for materials simulations, *Computing in Science & Engineering*, **1**(5): 39–47.

Nakano, A., Bachlechner, M. E., Kalia, R. K., Lidorikis, E., Vashishta, P., Voyiadjis, G. Z., Campbell, T. J., Ogata, S., and Shimojo, F. (2001). Multiscale simulation of nanosystems, *Computing in Science & Engineering*, **3**(4): 56–66.

Nakano, A., Kalia, R. K., Vashishta, P., Campbell, T. J., Ogata, S., Shimojo, F., and Saini, S. (2002). Scalable atomistic simulation algorithms for materials research, *Scientific Programming*, **10**: 263–270.

Nakano, A., Kalia, R. K., Nomura, K., Sharma, A., Vashishta, P., Shimojo, F., van Duin III, A. C. T., Goddard, W. A., Biswas, R., and Srivastava, D. (2006). A divide-and-conquer/cellular-decomposition framework for million-to-billion atom simulations of chemical reactions, *Computational Materials Science*, in press.

Ogata, S., Lidorikis, E., Shimojo, F., Nakano, A., Vashishta, P., and Kalia, R. K. (2001). Hybrid finite-element/molecular-dynamics/electronic-density-functional approach to materials simulations on parallel computers, *Computer Physics Communications*, **138**(2): 143–154.

Ogata, S., Campbell, T. J., Kalia, R. K., Nakano, A., Vashishta, P., and Vemparala, S. (2003). Scalable and portable implementation of the fast multipole method on parallel computers, *Computer Physics Communications*, **153**(3): 445–461.

Ogata, S., Shimojo, F., Kalia, R. K., Nakano, A., and Vashishta, P. (2004). Environmental effects of H_2O on fracture initiation in silicon: a hybrid electronic-density-functional/molecular-dynamics study, *Journal of Applied Physics*, **95**(10): 5316–5323.

Omeltchenko, A., Campbell, T. J., Kalia, R. K., Liu, X. L., Nakano, A., and Vashishta, P. (2000). Scalable I/O of large-scale molecular dynamics simulations: A data-compression algorithm, *Computer Physics Communications*, **131**(1–2): 78–85.

Rappe, A. K. and Goddard, W. A. (1991). Charge equilibration for molecular-dynamics simulations, *Journal of Physical Chemistry*, **95**(8): 3358–3363.

Rino, J. P., Ebbsjo, I., Kalia, R. K., Nakano, A., and Vashishta, P. (1993). Structure of rings in vitreous SiO_2, *Physical Review B*, **47**(6): 3053–3062.

Schlick, T., Skeel, R. D., Brunger, A. T., Kale, L. V., Board, J. A., Hermans, J., and Schulten, K. (1999). Algorithmic challenges in computational molecular biophysics, *Journal of Computational Physics*, **151**(1): 9–48.

Shan, H. Z., Singh, J. P., Oliker, L., and Biswas, R. (2002). A comparison of three programming models for adaptive applications on the Origin2000, *Journal of Parallel and Distributed Computing*, **62**(2): 241–266.

Shan, H. Z., Singh, J. P., Oliker, L., and Biswas, R. (2003). Message passing and shared address space parallelism on an SMP cluster, *Parallel Computing*, **29**(2): 167–186.

Sharma, A., Nakano, A., Kalia, R. K., Vashishta, P., Kodiyalam, S., Miller, P., Zhao, W., Liu, X. L., Campbell, T. J., and Haas, A. (2003). Immersive and interactive exploration of billion-atom systems, *Presence-Teleoperators and Virtual Environments*, **12**(1): 85–95.

Shimojo, F., Kalia, R. K., Nakano, A., and Vashishta, P. (2005). Embedded divide-and-conquer algorithm on hierarchical real-space grids: parallel molecular dynamics simulation based on linear-scaling density functional theory, *Computer Physics Communications*, **167**(3): 151–164.

Strachan, A., van Duin, A. C. T., Chakraborty, D., Dasgupta, S., and Goddard, W. A. (2003). Shock waves in high-energy materials: the initial chemical events in nitramine RDX, *Physical Review Letters*, **91**(9): 098301.

Szlufarska, I., Nakano, A., and Vashishta, P. (2005). A crossover in the mechanical response of nanocrystalline ceramics, *Science*, **309**(5736): 911–914.

Takemiya, H., Tanaka, Y., Sekiguchi, S., Ogata, S., Kalia, R. K., Nakano, A., and Vashishta, P. (2006). Sustainable adaptive Grid supercomputing: multiscale simulation of semiconductor processing across the Pacific, In *Proceedings of Supercomputing 2006*, Tampa: ACM.

Tanaka, Y., Nakada, H., Sekiguchi, S., Suzumura, T., and Matsuoka, S. (2003). Ninf-G: a reference implementation of RPC-based programming middleware for Grid computing, *Journal of Grid Computing*, **1**: 41–51.

Truhlar, D. G. and McKoy, V. (2000). Computational chemistry, *Computing in Science & Engineering*, **2**(6): 19–21.

Umezawa, N., Kalia, R. K., Nakano, A., Vashishta, P., and Shimojo, F. (2007). RDX (1,3,5-trinitro-1,3,5-triazine) decomposition and chemisorption on Al(111) surface: first-principles molecular dynamics study, *Journal of Chemical Physics*, in press.

van Duin, A. C. T., Dasgupta, S., Lorant, F., and Goddard, W. A. (2001). ReaxFF: a reactive force field for hydrocarbons, *Journal of Physical Chemistry A*, **105**(41): 9396–9409.

Vashishta, P., Kalia, R. K., and Nakano, A. (2006). Multimillion atom simulations of dynamics of oxidation of an aluminum nanoparticle and nanoindentation on ceramics, *Journal of Physical Chemistry B*, **110**(8): 3727–3733.

Whaley, R. C., Petitet, A., and Dongarra, J. J. (2001). Automated empirical optimizations of software and the ATLAS project, *Parallel Computing*, **27**(1–2): 3–35.

Yang, W. (1991). Direct calculation of electron density in density-functional theory, *Physical Review Letters*, **66**: 1438–1441.

Yang, W. and Lee, T.-S. (1995). A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules, *Journal of Chemical Physics*, **103**: 5674–5678.

Zhang, C., Bansal, B., Branicio, P. S., Kalia, R. K., Nakano, A., Sharma, A., and Vashishta, P. (2006). Collision-free spatial hash functions for structural analysis of billion-vertex chemical bond networks, *Computer Physics Communications*, **175**: 339–347.