A Systematic Approach to Composing & Optimizing Application Workflows: Patterns in Ultrascale Atomistic Simulations

Ewa Deelman, Aram Galstyan, Yolanda Gil, Mary Hall, Kristina Lerman (USC/ISI)
Aiichiro Nakano, Rajiv K. Kalia, Priya Vashishta (USC/CACS)
Joel Saltz (OSU, USC/ISI)

Information Sciences Institute (ISI)
Collaboratory for Advanced Computing & Simulations (CACS)
University of Southern California
Ohio State University

Email: anakano@usc.edu
Systematic Composition & Optimization of Application Workflows

- Systematize the construction & tuning of application workflows from existing components or patterns
- AI search techniques to select optimal application-hardware mappings
Computing Beyond Teraflop

MultiTera-to-Petaflop computers

19B-atom molecular dynamics
1.4M-atom (0.1T-dof) quantum mechanics

51.87 Tflop NASA Columbia

Global Grid of supercomputers

NSF TeraGrid

1,792-proc US/Japan metacomputing
Multiscale FE/MD/QM Simulation

- Embed high-accuracy computations only when & where needed
- Train coarse simulations by fine simulations

Multiscale simulation to seamlessly couple:
- Finite-element (FE) calculation based on continuum elasticity
- Atomistic molecular-dynamics (MD) simulation
- Quantum-mechanical (QM) calculation based on the density functional theory (DFT)
Advanced Atomistic Simulation Methods

\[ m_i \frac{d^2}{dt^2} = -\frac{\partial}{\partial \mathbf{r}_i} E_{\text{MD}}(\{\mathbf{r}_i\}) \]

- **Molecular Dynamics (MD)**
- **Quantum Mechanics (QM)**

**Electron wave function**

\[ \min E_{\text{QM}}(\{\psi_n(\mathbf{r})\}) \]

**Atom**

**Train**

**Quantum Mechanics (QM)**

**ReaxFF**

**Adaptive**

\[ E_{\text{MD}} \]

**First principles-based reactive force-field**

- **Charge equilibration (QEq) \{q_i\}** → Charge transfer
- **Reactive bond order \{B_{ij}\}** → Bond breakage & formation

[Goddard group, Caltech]
To enable billion-atom simulations of chemical reactions, develop:

- **Embedded divide-\&-conquer (EDC) algorithmic framework**: Linear-scaling algorithms for broad scientific \& engineering applications;

- **Hierarchical cellular decomposition (HCD) parallelization framework**: Mapping \(O(N)\) EDC algorithms onto high-end computer architectures;

- **Massive dataset visualization \& analysis**: Interactive visualization of billion-atom datasets \& graph-based knowledge discovery;

Based on “data-locality principles”.
Embedded Divide-\&-Conquer Algorithms

- **N-body problem**: $O(N^2) \rightarrow O(N)$
  - Space-time multiresolution molecular dynamics (MRMD): Fast multipole method \& symplectic multiple time stepping

- **Variable N-charge problem**: $O(N^3) \rightarrow O(N)$
  - Fast reactive force-field (F-ReaxFF) MD: Multilevel preconditioning

- **Quantum N-body problem**: $O(C^N) \rightarrow O(N)$
  - EDC density functional theory (EDC-DFT): Adaptive multigrids
Spatial locality: Fast multipole method (FMM) [Greengard & Rokhlin, ‘87]

1. Clustering: Encapsulate far-field information using multipoles
2. Hierarchical abstraction: Octree data structure
3. $O(N)$ divide-&-conquer algorithm: Locality at each octree level

Temporal locality: Symplectic multiple time stepping (MTS) [Tuckerman, Berne & Martyna, ‘92]

- Different force-update schedules for different force components
  → 1) Reduced computation
     2) Enhanced data locality & parallel efficiency

Atomistic stress tensor by novel complex charge method (CCM)
F-ReaxFF: Variable $N$-Charge Problem

• Charge-equilibration (QEq) in ReaxFF MD: Determine atomic charges $\{q_i \mid i = 1, ..., N\}$ every MD step to minimize $E_{ES}(r^N, q^N)$ with charge-neutrality constraint: $\Sigma_i q_i = 0$
  — Dense linear system: $M q = -\chi$ $O(N^3)$!

• $O(N)$ fast reactive force-field algorithm
  1) Fast multipole method
  2) Temporal locality, $q_i^{(\text{init})}(t+\Delta t) = q_i(t)$

• Multilevel preconditioned conjugate gradient (MPCG) method
  1) Split Coulomb matrix: $M = M_{\text{near}} + M_{\text{far}}$
  2) Sparse near-field preconditioner: $M_{\text{near}}^{-1} M q = -M_{\text{near}}^{-1} \chi$

Results:
• 20% speed up of convergence
• Enhanced data locality: Improved parallel efficiency 0.93 $\rightarrow$ 0.96 for 26.5M-atom $\text{Al}_2\text{O}_3$ on 64 Power nodes
DFT: Quantum $N$-Body Problem

Density functional theory (DFT) [Kohn, Nobel Chemistry Prize, ‘98]
\[
\psi(r_1, r_2, \ldots, r_{N_{el}}) \longrightarrow \left\{ \psi_n(r) \mid n = 1, \ldots, N_{el} \right\} \min E_{QM}\left(\{\psi_n(r)\}\right)
\]
\[O(C^N)\]
\[O(N^3)\]

- Pseudopotential [Troullier & Martins, ‘91]
- Generalized gradient approximation [Perdew et al., ‘96]

Efficient iterative DFT solver: Adaptive multigrids

- High-order finite difference [Chelikowsky, Troullier, Saad, ‘94]
- Multigrid acceleration [Bernholc et al., ‘96; Beck, ‘00]
$O(N)$ Embedded Divide-\&-Conquer DFT

“Quantum nearsightedness” locality principle \cite{Kohn, '96}

- Overlapping domains:

$$\Omega = \sum_{\alpha} \Omega_\alpha = \sum_{\alpha} \Omega_{0\alpha} \cup \Gamma_{1\alpha} \cup \Gamma_{2\alpha}$$

- Linear combination:

$$\rho(\mathbf{r}) = \sum_{\alpha} p^\alpha(\mathbf{r}) \sum_{n} f(\varepsilon_n^\alpha) |\psi_n^\alpha(\mathbf{r})|^2$$ \cite{Yang, '91}

- Robust convergence, controlled error bounds, \& energy conservation during MD simulation

- Spatial decomposition

  > Inter-processor communication of density only $\rightarrow$ scalability
Hierarchical Cellular Decomposition

Mapping $O(N)$ EDC algorithms onto deep memory hierarchies

- Spatial decomposition with atom “caching” & “migration”
- Computational cells (linked-list cells in MD or domains in EDC-DFT) < cell blocks (threads) < processes ($P^\gamma$, spatial decomposition subsystems) < process groups ($P^\gamma$, Grid nodes)
- Data & computation re-ordering & tunable granularity at each decomposition level
- Multilayer cellular decomposition (MCD) for $n$-tuples ($n = 2–6$) in F-ReaxFF
- Hybrid MPI+OpenMP implementation
Locality-Preserving Parallelization

• Topology-preserving “computational-space decomposition” with structured messages in curved space, $\xi(x)$

Irregular data-structures/processor-speed

Map

Parallel computer

• Simulated annealing optimizes $\xi(x)$ for load balancing
• Wavelet representation speeds up the optimization

Regular mesh topology in computational space, $\xi$
Curved partition in physical space, $x$

Load-imbalance + communication costs

Plane wave
Wavelet

CPU time (min)

$10^{-3}$ $10^{-1}$ $10^{1}$ $10^{3}$
Locality in Data Compression

Massive data transfer via wide area network:
75GB/step of data for 1.5 billion-atom MD!
→ Compressed software pipeline

Scalable encoding:
• Store relative positions on spacefilling curve: \( O(N\log N) \rightarrow O(N) \)

Result:
• Data size, 50Bytes/atom → 6 Bytes/atom
Data Locality in Massive Visualization

- Octree-based fast view-frustum culling
- Probabilistic occlusion culling
- Parallel/distributed processing

Interactive visualization of billion atoms

*IEEE Virtual Reality 2002 Best Paper*
Scalable Simulation Algorithm Suite

Design-space diagram on 1,920 Itanium2 (1.5GHz) procs. of NASA Columbia

- 19 billion-atom MRMD of SiO$_2$
- 0.56 billion-atom F-ReaxFF MD of RDX
- 1.4 million-atom (0.12 trillion grid points) EDC-DFT MD of Al$_2$O$_3$

Significant improvement over 1,024-proc. IBM SP3 results in ‘01:
- 6.4 billion-atom MD of SiO$_2$
- 110,000-atom DFT of GaAs

IEEE/ACM SC2001 Best Paper Award
Subcritical crack growth rate: \[ v = v_0 \exp\left(-\Delta(\sigma, T, pH, \ldots)/k_B T\right) \]

Researchers from the University of Southern California, Caltech and Purdue University, led by USC’s Priya Vashishta, are studying how materials fail as a result of mechanical stresses and harsh chemical environments.

**Credit:** Priya Vashishta, Rajiv Kalia and Alichiro Nakano, USC

**Collaborators:** William Goddard, Michael Ortiz (Caltech) Ananth Grama (Purdue)
Multiscale MD/QM (DFT) Algorithm

**Embedded D-&C additive hybridization** [Morokuma et al., ‘96]
- Extrapolation in meta-model space (accuracy vs. size)
- Modular
  → Reuse of existing MD & QM (Density Functional Theory) codes
  → Minimal inter-model dependence/communication

**Divide-&-conquer**

\[ E = E_{\text{system}}^{\text{MD}} + \sum_{\text{cluster}} \left[ E_{\text{QM}}^{\text{cluster}} \left( \{ r_{\text{QM}} \}, \{ r_{\text{HS}} \} \right) - E_{\text{MD}}^{\text{cluster}} \left( \{ r_{\text{QM}} \}, \{ r_{\text{HS}} \} \right) \right] \]
ARL-MURI: Nano-energetic Materials

Million-atom QEq-MD simulation: Al nanoparticles for rocket-fuel ignition

Atomic charges in canonical MD: Stable oxide scale ~ 40Å

Temperature in microcanonical MD: Explosive energy release
Inverse problem: design materials with desired properties

209 million atom MD of hypervelocity impact in AlN for the design of lightweight ceramic armors
Study effect of controlled changes in gene expression on tissue structure

Computational pipeline to construct a 3D model of tissue layers from microscopic images

Associate genotype with phenotype

Big science animal experiments on cancer, heart disease, pathogen host response

- Basic small mouse is 3 cm$^3$
- 1 μm resolution – very roughly $10^{13}$ bytes/mouse
- Molecular data (spatial location) can add additional significant factor; e.g. $10^2$
  - Multispectral imaging, laser captured microdissection, flow cytometry
- Vary genetic composition, environmental manipulation, systematic mechanisms for varying genetic expression; another factor of $10^3$

Total: $10^{18}$ bytes per big science animal experiment

Fig. 1. An example of the microscopic image of a mouse placenta section from the Rb- placenta.
Data Middleware Services

- **Filter-stream based distributed execution middleware** (*DataCutter, STORM*)

- **Grid based data virtualization, data management, query, on demand data product generation** (*STORM, Active ProxyG, Mako*)

- **Distributed metadata management** (*Mobius Global Model Exchange*)
  > Track metadata associated with workflows, input image datasets, checkpointed intermediate results
Artificial Intelligence Search Techniques to Manage Numerous *Interdependent* Choices for Workflow Design

- Many variants/implementations of application components, each with different storage and computational requirements
- Many alternative data collections with different degrees of pre-processing
- Many possible replicas in the distributed environment
- Many possible resources to execute workflow
Workflows & AI Optimization

• **CAT** [Gil, *et al.*, ‘02]: Interactive workflow template editor with intelligent assistance & completion capabilities uses OWL ontologies to model components
  > **Planned**: Representing code variants

• **MCS/Artemis** [Gil, Deelman, *et al.*, ‘04]: Semantic representations of metadata for data catalogs
  > **Planned**: Reasoning about data sizes and computation

• **Pegasus** [Deelman, Gil, *et al.*, ‘03, ‘04, ‘05]: Automated refinement of workflow instance to assign appropriate resource types & add data movements
  > **Planned**: Reason about fine-grained differences in target architecture

• **Reinforcement learning for distributed scheduling** [Lerman, Galstyan, *et al.*, ‘03]
  > **Planned**: Integrated, iterative search framework to support reasoning about interrelated decisions
Component Optimization

• **Model-guided empirical optimization** [Chun, Chame, Hall, ‘05; Lee, Diniz, Hall, Lucas, ‘05]: Use collaboration between compiler models & empirical search to derive hand-tuned performance levels for memory hierarchy.

• **Design space exploration for FPGAs and embedded devices** [So, Hall, Diniz, ‘02; Ziegler, Hall, ‘03, ‘05]: Use models & empirical search for a different class of parallel, distributed-memory architectures.

• **Other work on SIMD compilation**, optimizing for complex memory hierarchies, communication & pipelining.

• **Planned**: Multi-core optimization, generalizing model-guided empirical optimization.
Conclusion

• Ultrascale atomistic simulations can be developed with:
  1) Embedded divide-&-conquer (EDC) algorithmic framework;
  2) Hierarchical cellular decomposition (HCD) parallelization framework; based on data locality principles.

• AI search techniques will be used for optimal workflow design.