Distributed Graph-based Density Matrix Calculation for Quantum Molecular Dynamics using GPUs

S. M. Mniszewski, C. F. A. Negre, M. J. Cawkwell, A. M. N. Niklasson
Los Alamos National Laboratory

smm@lanl.gov

2016 GPU Technology Conference
April 4-7, 2016
Why Molecular Dynamics?

Web of Science Database Topic Search

Article Count (2000-2013)

FD: Fluid Dynamics
CM: Climate Modeling
ML: Machine Learning
PC: Parallel Computing
MD: Molecular Dynamics
SC: Superconductivity
EX: Explosives
GPH: Graphene

Interdisciplinary
Background

• In molecular dynamics simulation, the relative positions of atoms evolve over a series of time steps according to the force acting on each atom.

• Employed in materials science, chemistry, and biology to study structures, defects, and equilibrium and non-equilibrium phenomena.

• Dependence on an interatomic potential to calculate forces and energy.

• Quantum-based models capture the making and breaking of covalent bonds, charge transfer between species of differing electronegativities, and long-range electrostatic interactions - reactions.
Quantum Molecular Dynamics (QMD)

Integrate the equations of motion of classical molecular trajectories

\[ M_I \ddot{R}_I = -\frac{\partial U(R; \rho_{sc})}{\partial R_I} \]

with the forces calculated on the fly from a self-consistent quantum mechanical description of the electronic structure:

\[ H[\rho]\Psi_i = \varepsilon_i \Psi_i \]

\[ \rho = \sum_{occ.} |\Psi_i|^2 \rightarrow \rho_{sc} \]

\[ \text{SCF} \quad \#\text{SCF} \times \mathcal{O}(N^3) \]
Example: Biosynthesis of Histidine

- What is responsible for the biosynthesis of histidine?
- Present in several pathogenic bacteria
- Allosteric mechanism was determined with MD simulation
- Several thousand atoms over timescales of 100s of ns!
Future QMD Simulations

**IAPP dimerization and type-2 diabetes (537 atoms).**

**Beta Amyloid Peptide and Alzheimer’s Disease (410 atoms).**

Computational Cost

- High computational cost and complexity of QMD calculations
- The MD timestep is the most expensive – the density matrix construction
- The second order spectral projection (SP2) algorithm breakthrough
- Use of hybrid parallelism on GPU-accelerated clusters

Quantum Molecular Dynamics

System Size (N)
Wall-Clock Time / MD Time Step

- $O(N^3)$ Diagonalization
- $O(N)$ Regular linear scaling
- $O(N)$ Low pre-factor

Significant pre-factor reduction is required for practical large scale QMD!
The Density Matrix Computation

• Typically, algorithms used in quantum-based models, most notably matrix diagonalization, are not ideally suited to GPUs
  — Due to their complexity
  — Difficulty in extracting thread-level parallelism
  — Difficulty of avoiding branching within warps

• New SP2 approach
  — Computed directly from the Hamiltonian through a recursive expansion of the Fermi Operator with the second order spectral projection (SP2) algorithm
  — Based on a series of generalized matrix-matrix multiplications
  — Only one matrix-matrix multiplication is required per iteration
  — Maps very well to GPUs
The Second Order Spectral Projection Algorithm (SP2) – Reduced Complexity

Recursive Fermi Operator expansion

\[ \rho = \theta \left[ \mu I - H \right] = \lim_{i \to \infty} f_i \left[ f_{i-1} \left[ \ldots f_0 \left[ X_0 \right] \ldots \right] \right] \]

\[ X_0 = \frac{\varepsilon_{\text{max}}}{{\varepsilon_{\text{max}}} - {\varepsilon_{\text{min}}} \left( I - H \right)} \]

\[ f_i \left[ X_i \right] = \begin{cases} X_i^2 & \text{if } 2 \text{Tr} \left[ X_i \right] \geq N_e \\ 2X_i - X_i^2 & \text{if } 2 \text{Tr} \left[ X_i \right] < N_e \end{cases} \]

SP2 Algorithm Using the GPU Approach

Estimate $\varepsilon_{\text{max}}$ and $\varepsilon_{\text{min}}$

\[ X = (\varepsilon_{\text{max}}I - H)/(\varepsilon_{\text{max}} - \varepsilon_{\text{min}}) \]

TraceX = Tr[X] /* Trace kernel on GPU */

Until converged do

\[ X_{\text{tmp}} = X \]

\[ X_{\text{tmp}} = X^2 + X_{\text{tmp}} \] /* CUBLAS xGEMM */

TraceX_{\text{tmp}} = Tr[X_{\text{tmp}}] /* Trace kernel on GPU */

if $|2\text{TraceX} - 2\text{TraceX}_{\text{tmp}} - N_e| > |2\text{TraceX} + 2\text{TraceX}_{\text{tmp}} - N_e|$ 

\[ X = X + X_{\text{tmp}} \] /* CUBLAS xAXPY */

TraceX = TraceX + TraceX_{\text{tmp}} /* CUBLAS xAXPY */

else

\[ X = X - X_{\text{tmp}} \] /* CUBLAS xAXPY */

TraceX = TraceX - TraceX_{\text{tmp}}

end until

$\rho = X$

Cawkwell MJ, Mniszewski SM, Niklasson AMN, Fast Quantum Molecular Dynamics on Multi-GPU Architectures in LATTE, GTC 2013.
Density Matrix Calculation (Nvidia M2090) – Liquid Methane (10 – 1250 molecules)

Cawkwell MJ, Mniszewski SM, Niklasson AMN, Fast Quantum Molecular Dynamics on Multi-GPU Architectures in LATTE, GTC 2013.
Sparse Matrix SP2 – ELLPACK-R Format

- Described by 3 arrays, 2-D values and indices, 1-D non-zero entries per row
- N rows and M max non-zeroes per row, O(Nm^2) computational complexity
- Row-wise storage for parallelism opportunities
- No insertion cost compared to CSR

SP2 Shared Memory – Significant Cost Reduction

Shared Memory SP2 – TRP Cage Protein

303 atom Trp Cage Protein solvated by 2682 water molecules (8349 atoms)

LATTE Simulation – 18.8 ps

NVT ( Thermalization)  NVE

Graph-based Electronic Structure Theory

Divide and Conquer  Sparse Matrix Algebra

Graph Theory
Graph-based SP2

Data dependency Graph $S_\tau$

$$S_\tau \leftarrow [\text{Fermi Operator Expansion}]_{\tau(\text{global})}$$

Recursive Fermi-operator expansion

$$D_\tau = \left\{ \lim_{n \to \infty} f_n(f_{n-1}(\ldots f_0(h[s^{(i)}_\tau])\ldots)) \right\}_{i=1}^N$$

Exact Relation!
Graph-based SP2 – Hybrid approach

On Gpus!

Graph Partitioning SP2

Graph Partitioning
- Structure-based
- Graph-based
- Hypergraph-based
- Community-based

Partitions are sets of core rows/orbitals of $H$

Subgraph Processing
1. Determine core+halo
2. Extract submatrix
3. Run Dense SP2
4. Collect into next $D$

Trivial parallelism based on dense matrix algebra at BLAS3 performance
Subgraph Processing

For all sub-graphs:

1. Determine elements
2. Extract submatrix
3. Run SP2
4. Assemble into $D$

- Trivial parallelism
- Same HOMO-LUMO sequence through SP2
- Dense matrix & communication-free SP2
- Small subgraphs, process single-threaded
- Large subgraphs, process multi-threaded
- Tunable accuracy - thresholds
Graph-based XL Born-Oppenheimer Molecular Dynamics (XL-BOMD)

Liquid water $(\text{H}_2\text{O})_{100}$

DFTB-LATTE

Graph-based XL Born-Oppenheimer Molecular Dynamics (XL-BOMD)

Graph-based SP2 – Tunable accuracy

Graph-based SP2 – Distributed GPU Performance

- 64-4096 METIS partitions
- Single Nvidia Tesla M2090 GPU per node
- Partition core+halo sizes vary, load-balancing required
- 16,384 GPU threads, still perfect strong scaling
- Best - ~25 µs/atom

Density matrix calculations for Polyalanine in water 20,000 atoms

64-4096 METIS partitions
1 CPU SpM Alg (MKL)
1 CPU Graph Part.
1 GPU Graph Part.
16 GPU Graph Part.
32 GPU Graph Part.

Graph-based SP2 – Distributed CPU vs. GPU

- 1024 2048 METIS partitions
- 16-256 CPU/GPU nodes
- MKL vs. CuBLAS
- Similar for 1024 and 2048 partitions
- Near linear scaling
- Speedup of 1.7X on GPUs
- Best – 5 µs/atom

Density matrix calculations for Polyalanine in water 20,000 atoms
Basic Matrix Library (BML) & PROGRESS Library

- **bml C API**
  - bml_multiply()
  - bml_add()
  - ...

- **bml Fortran API**
  - call bml_multiply()
  - call bml_add()
  - ...

Uses BLAS Level 2 and 3 routines.

- dense matrix
- sparse ELLPACK matrix
- sparse CSR matrix

BML available at [http://qmmd.github.io/bml/](http://qmmd.github.io/bml/) under the BSD 3-clause license
Summary

• Distributed graph-based SP2 provides significant speedup using distributed GPU-accelerated architectures

• Available as part of BML & PROGRESS libraries

• Allows for QMD simulations of larger systems and longer timeframes than previously possible