Algorithmic Techniques for Atomistic Modeling

Hasan Metin Aktulga
Dept of Computer Science
Purdue University
haktulga@cs.purdue.edu

May 14, 2009

ayg: chage green to a darker color

Outline

1. **Introduction: MD Methods & Reactive Force Fields**

2. **Our Implementation of Reactive Force Fields (ReaxFF)**
   - General structure of a ReaxFF implementation
   - Implementation details of major components
   - Additional features

3. **Applications using the Serial Version**
   - Hexane simulations: Validation and performance analysis
   - Water-Silica Surface Interaction
   - Si/Ge Nanobar Strain Rates

4. **Ongoing & Future Work**

5. **Parallelization of ReaxFF**
   - Challenges
   - Our solution approaches

6. **Current Userbase**

Introduction: Atomistic Modeling Methods

- ***Ab-initio* methods:** Atomistics with electronic degrees of freedom
  - Hartree-Fock(HF) methods $\rightarrow$ misses E_{corr}
  - Post-HF methods
    * E_{corr} incorporated at great computational expense
    * semi-empirical methods to the rescue
  - DFT-based methods
    * completely different approach but similar derivations to HF theory
    * bigger systems, longer simulation times made possible
    * CPMD is the most popular example

- **Classical MD methods:**
  - many approximations: no electronic d.o.f.
  - electronic effects are mimiced through parametrizations
  - static bonds, no reactions!
  - systems at nanoscale, simulation times upto hundreds of ns

- **Continuum Mechanics:** Macroscale systems
  - additional approximations, no atomistic detail!
  - modeled using PDE's.

$\uparrow$
accuracy
complexity
detail
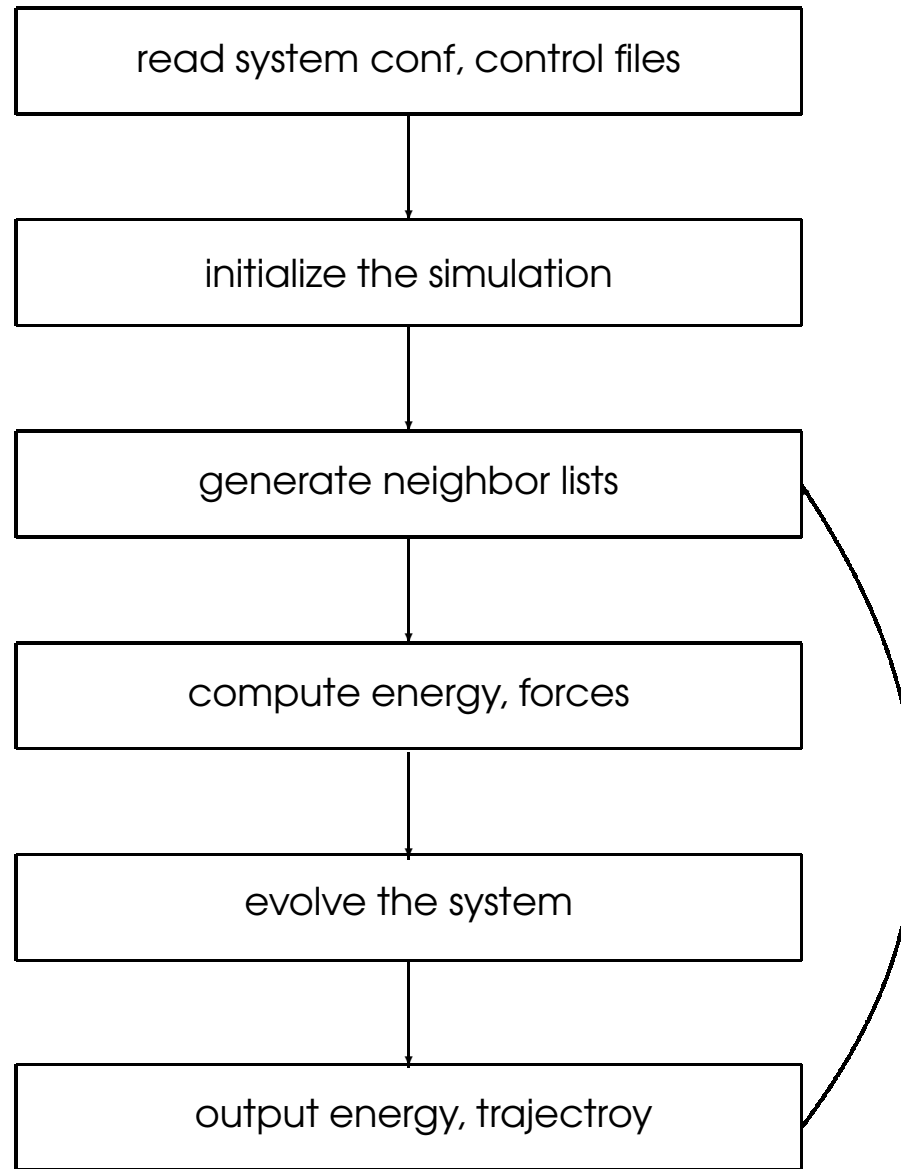comp.
demand

approx's
sys size
sim.
frame

$\downarrow$

Reactive Force Fields (ReaxFF): Bridging the Gap

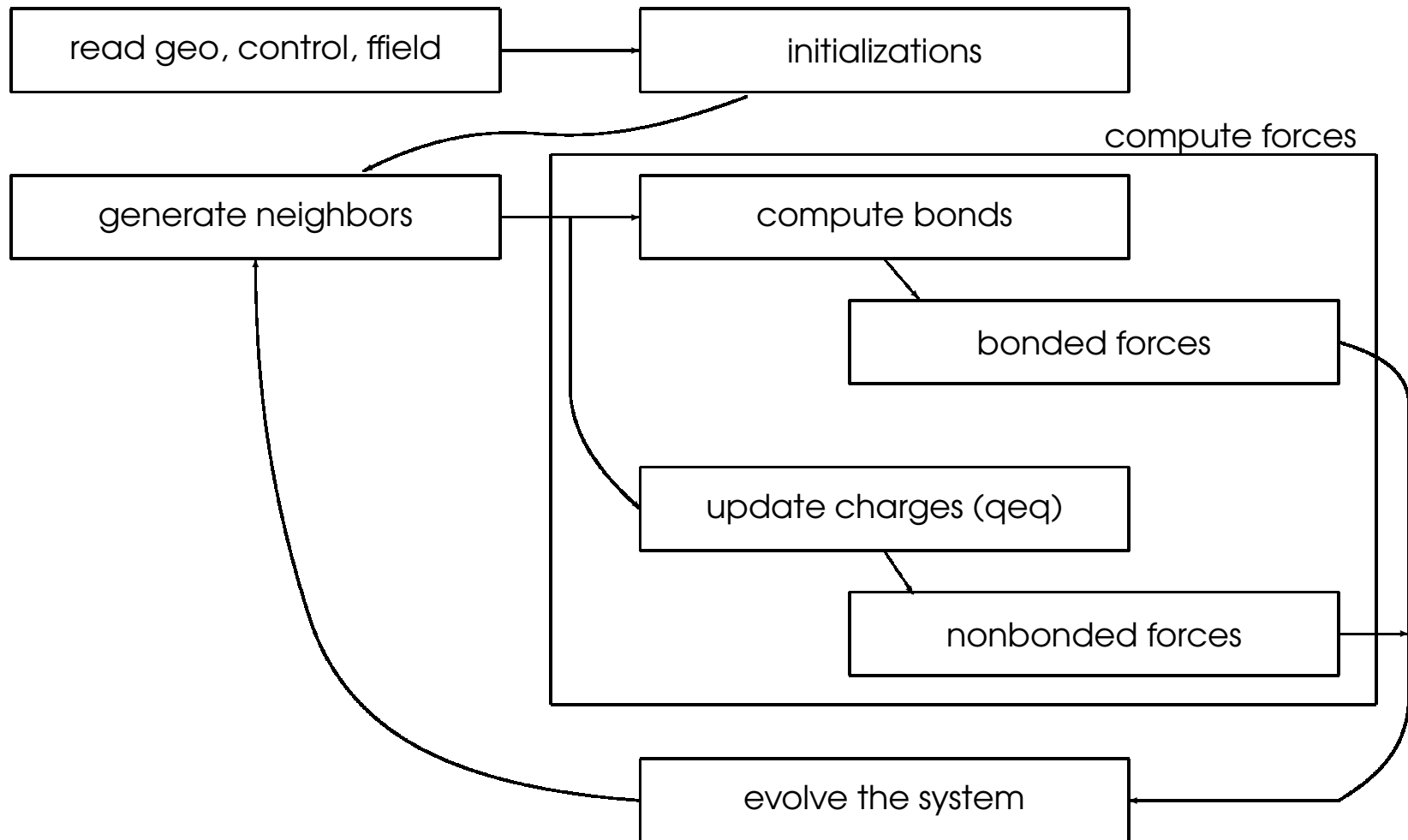|	ReaxFF	Classical MD
advantages	reactive	non-reactive
	dynamic bonds	static bonds
	polarization with QEq	fixed charges in general (except for polarizable FF)
challenges	dynamic interaction lists	static lists
	complex & costly enegies,forces	much simpler energy & force formulas
	frequent update of charges (expensive!)	static charges
	shorter timesteps (≈ 0.25 fs)	longer timesteps ($1-10$ fs)

Interesting applications: large system with reactions and charge-transfer

Simulation of fuel cells, silica crack propogation, corrosion of silica in water, etc.

General Flowchart of Conventional MD Programs

```
┌─────────────────────────────────┐
│   read system conf, control files │
└─────────────────────────────────┘
              │
              ▼
┌─────────────────────────────────┐
│     initialize the simulation     │
└─────────────────────────────────┘
              │
              ▼
┌─────────────────────────────────┐
│      generate neighbor lists      │◄──┐
└─────────────────────────────────┘   │
              │                         │
              ▼                         │
┌─────────────────────────────────┐   │
│      compute energy, forces       │   │
└─────────────────────────────────┘   │
              │                         │
              ▼                         │
┌─────────────────────────────────┐   │
│         evolve the system         │   │
└─────────────────────────────────┘   │
              │                         │
              ▼                         │
┌─────────────────────────────────┐   │
│     output energy, trajectroy     │───┘
└─────────────────────────────────┘
```

ReaxFF Flowchart

read geo, control, ffield	→	initializations

compute forces

generate neighbors	→	compute bonds

bonded forces

update charges (qeq)

nonbonded forces

evolve the system

ayg: What about valence corrections?

Implementation: Neighbor Generation

- 3 different neighbor lists:
  - near_nbrs for bonded forces $\rightarrow$ bond_cut $\approx$ 4-5 Å, full matrix stored
  - hbond_list for hydrogen bonds $\rightarrow$ hbond_cut $\approx$ 6-7.5 Å, only for H
  - far_nbrs for non-bonded forces $\rightarrow$ nonb_cut $\approx$ 10 Å, upper-half only

- Bin atoms into 3D grid cells
  - grid cell dims $\approx \frac{1}{2}$nonb_cut

- Verlet lists with delayed re-neighboring not implemented $\rightarrow$ little benefit

- Compressed adjacency list representation
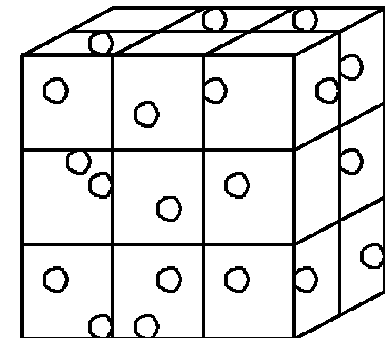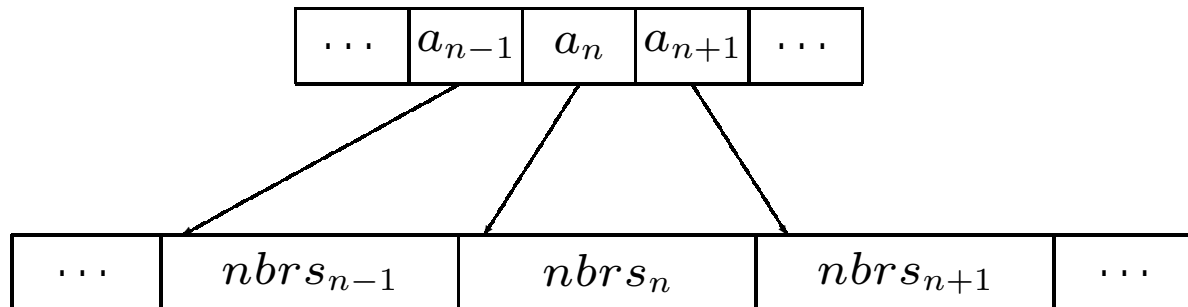
Implementation: Computing Forces and Potentials

- Bonded Interactions: Similar to classical MD but accounts for dynamic bonds
  - precursor: bond orders
  - lone-pair energy, over/undercoordination energies
  - bond energy
  - valence energy (with penalty & 3-body conjugation corrections)
  - dihedral energy (with 4-body conjugation correction)

- Hydrogen Bonds
  - precursor: bond orders
  - H covalently bonded to X and interacting with Z
  - can be considered a bonded interaction

- Non-bonded Interactions
  - precursor: charge equilibration (QEq)
  - electrostatic (Coulomb) energy, van der Waals energy

$$
\begin{aligned}
E_{system} \;=\; & E_{bond} + E_{lp} + E_{over} + E_{under} + E_{val} + E_{pen} + E_{3conj} \\
& + E_{tors} + E_{4conj} + E_{H-bond} + E_{vdW} + E_{Coulomb}
\end{aligned}
$$

Bonded Interaction: Bond Orders

- Prior to bonded forces, compute bond orders based on the new near_nbrs
  - bond_list: subset of near_nbrs, stored in the same way
  - uncorrected bond orders and derivatives
  - store both bo(i, j) and bo(j, i) → efficient construction of angles, dihedrals
  - compute bo(i, j) only if $i < j$, otherwise bo(i, j)=bo(j, i)

$$\mathrm{BO}'_{ij} = \mathrm{BO}^{\sigma'}_{ij} + \mathrm{BO}^{\pi'}_{ij} + \mathrm{BO}^{\pi\pi'}_{ij}$$

$$\mathrm{BO}^{\alpha'}_{ij}(r_{ij}) = exp\left[a_\alpha \left(\frac{r_{ij}}{r_{0\alpha}}\right)^{b_\alpha}\right]$$

$$\mathrm{BO}_{ij} = \mathrm{BO}'_{ij} \cdot f_1(\Delta'_i, \Delta'_j) \cdot f_4(\Delta'_i, \mathrm{BO}'_{ij}) \cdot f_5(\Delta'_j, \mathrm{BO}'_{ij}) \text{ where } \Delta'_i \text{ is the valency of atom } i.$$
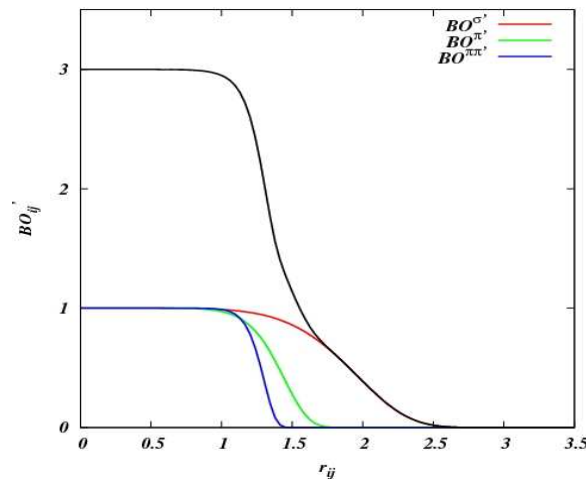
Bonded Interaction: Bond Energy

$$
\begin{aligned}
\mathsf{E_{bond}} \;=\; & -D_e^{\sigma} \cdot \mathsf{BO}_{ij}^{\sigma} \cdot exp\left\{ p_{be1}\left(1 - \left(\mathsf{BO}_{ij}^{\sigma}\right)^{p_{be2}}\right)\right\} \\
& -D_e^{\pi} \cdot \mathsf{BO}_{ij}^{\pi} - D_e^{\pi\pi} \cdot \mathsf{BO}_{ij}^{\pi\pi}
\end{aligned}
$$

- the stronger the bond, the lower the associated energy

- sweep over the bond_list

- compute bond_energy between i, j only if $i < j$

Bonded Interaction: Lone-Pair & Over/UnderCoordination

- Lone-pair energy
  - $\Delta_i^{lp} = n_{opt}^{lp} - n_i^{lp}$

  - energy associated with unpaired electrons of an atom $\rightarrow$ zero for a fully coordinated atom

 - single-body interaction $\rightarrow$ just sweep over atom_list

- Over/undercoordination energy
  - ideal # of bonds = # of valence electrons

  - $\Delta_i = \sum_{j \in nbrs(i)} \text{bo}(i,j) - Val_i$

  - actual # of bonds > ideal # of bonds ($\Delta_i > 0$) $\rightarrow$ over-coordination

  - actual # of bonds < ideal # of bonds ($\Delta_i < 0$) $\rightarrow$ under-coordination

  - actual # of bonds = ideal # of bonds ($\Delta_i = 0$) $\rightarrow$ no over/undercoordination energy

  - functionals of Δ_i and Δ_j's $\rightarrow$ just sweep over atom_list

Bonded Interaction: Valence Angle Energy

$$E_{val} = f_7(BO_{ij}, p_{val3}, p_{val4}) \cdot f_7(BO_{jk}, p_{val3}, p_{val4}) \cdot f_8(\Delta_j, p_{val5}, p_{val6}, p_{val7}) \cdot$$

$$\left(p_{val1} - p_{val1} \cdot exp \left\{ -p_{val2} \cdot \left(\Theta_0 - \Theta_{ijk} \right)^2 \right\} \right)$$

- Θ_0 is the ideal angle, Θ_{ijk} is the actual angle
  - the closer the Θ_{ijk} to Θ_0, the lower the energy

- $f_7(BO_{ij}, \ldots)$ and $f_7(BO_{jk}, \ldots)$ ensure $E_{val} \to 0$ as $BO_{ij} \to 0$ or $BO_{jk} \to 0$

- $\{\forall x, y \in bo\_list_i | x < y\}$ that meet certain criteria, compute the energy of $< x, i, y$

- $\{\forall x, y \in bo\_list_i | x > y\}$ copy $< y, i, x$ into $< x, i, y$ (for dihedrals!)

- E_{pen} and E_{3conj} account for corrections in special cases
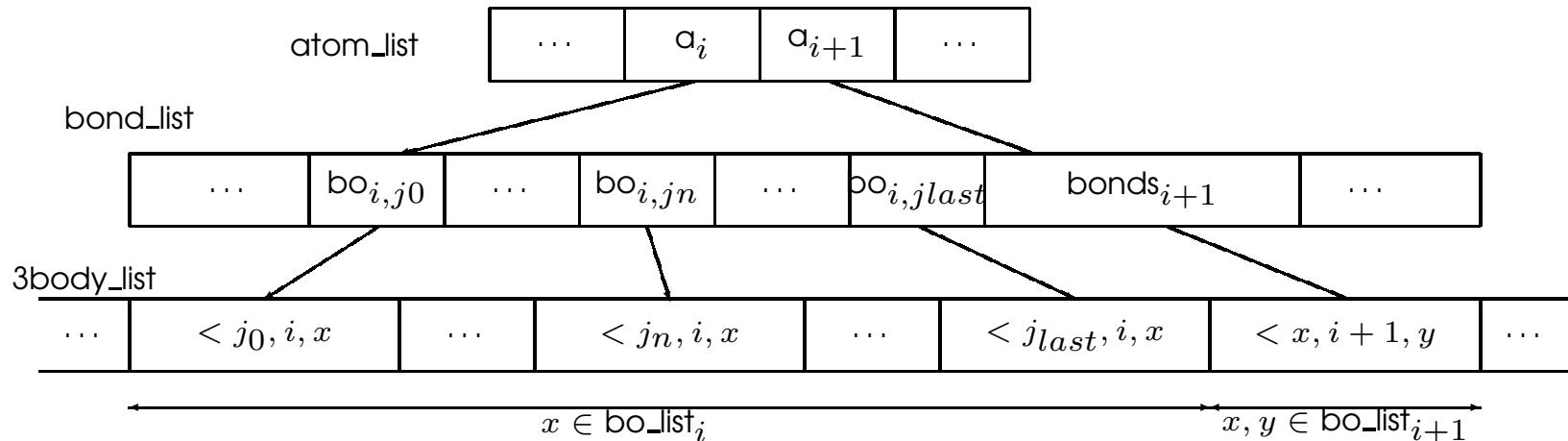
Bonded Interaction: Dihedral Energy

$$E_{tors} \quad = \quad \frac{1}{2} \cdot f_{10}(BO_{ij}, BO_{jk}, BO_{kl}, p_{tor2}, 1) \cdot sin\Theta_{ijk} \cdot sin\Theta_{jkl} \cdot V_{123}(\omega_{ijkl})$$

- Dihedral angle ω_{ijkl} is the angle between planes defined by positions of i, j, k and j, k, l

- $f_{10}(BO_{ij}, BO_{jk}, BO_{kl}, \ldots)$ ensure E_{tors} vanishes smoothly as any of these bonds dissociate

- $sin\Theta_{ijk}$ and $sin\Theta_{jkl}$ ensure that $E_{tors} \to 0$ as $\Theta_{ijk} \to 0$ or $\Theta_{jkl} \to 0$

- $\{\forall i, j, k, l \in$ atoms$|j < k, < i, j, k \in$ 3body_list$_{jk}, < j, k, l \in$ 3body_list$_{kj}\}$ compute the energy associated with ω_{ijkl}

- weak but very important in determining the 3D structures

- no higher order interactions $\to$ no storage necessary

Hydrogen Bonds

$$E_{XHZ} = p_{hb1} \cdot f_7(\text{BO}_{XH}, p_{hb2}, 1) \cdot sin^4\left(\frac{\Theta_{XHZ}}{2}\right) \cdot exp\left\{-p_{hb3} \cdot \left(\frac{r_{hb}^0}{r_{HZ}} + \frac{r_{HZ}}{r_{hb}^0} - 2\right)\right\}$$

- Constraints of a hydrogen bond:
  - Middle atom must be H
  - X, Z must be one of N, O, P, F
  - $X - H$ covalently bonded, $Z \in$ hbond_list$_H$

- $f_7(\text{BO}_{XH}, \ldots)$ ensure $E_{\text{val}} \rightarrow 0$ as the covalent bond breaks

- $sin^4(\frac{\Theta_{XHZ}}{2})$ maximized when $\Theta_{XHZ} = \pi$ ensures alignment on a line

- crucial for accurately describing water, DNA structure, secondary structures in proteins, etc.

Nonbonded Interaction: Charge Equilibriation (QEq)

Minimizing electrostatic energy by redistributing (partial) charges.

$$\text{Minimize } E(Q_1 \dots Q_N) = \sum_A (E_{A0} + \chi_A^0 Q_A + \frac{1}{2} J_{AA}^0 Q_A^2) + \sum_{A<B} (J_{AB} Q_A Q_B)$$

$$\text{subject to } Q_{net} = \sum_{i=1}^{N} Q_i$$

- Solve the optimization problem using the method of Lagrange multipliers
    - gives a sparse linear system of equations for finding charges

- GMRES with restarts, $GMRES(50)$
    - heavy diagonal $\rightarrow$ diagonal preconditioner
    - little configurational change between steps $\rightarrow$ initial guess $q_t = \text{linear\_extrapolation}(q_{t-1}, q_{t-2})$

- Implemented GMRES both with MGS and $Householder$ orthogonalizations
    - virtually no difference

- Implemented CG for comparison
    - GMRES takes fewer number of matvecs and is faster than CG

- Choose tolerance for the norm of the relative residual carefully:
    - too high $\rightarrow$ wrong results!
    - too low $\rightarrow$ QEq dominates the total computation time!
    - more on this later . . .

Nonbonded Interactions: Coulomb & van der Waals Energy

$$E_{Coulomb} = C \cdot Tap(r_{ij}) \cdot \frac{q_i \cdot q_j}{\left[r_{ij}^3 + \gamma_{ij}^{-3} \right]^{\frac{1}{3}}}$$

$$E_{vdWaals} = Tap(r_{ij}) \cdot D_{ij} \cdot$$

$$\left[exp \left\{ \alpha_{ij} \cdot \left(1 - \frac{f_{13}(r_{ij})}{r_{vdW}} \right) \right\} - 2 \cdot exp \left\{ \frac{1}{2} \cdot \alpha_{ij} \cdot \left(1 - \frac{f_{13}(r_{ij})}{r_{vdW}} \right) \right\} \right]$$

- Shielding prevents energies from increasing drastically at close distances

- Long range interactions with cutoffs
  - Taper term ensures smooth vanishing of energies after the cutoff

- no $1 - 2$, $1 - 3$ or $1 - 4$ exclusions $\rightarrow$ smooth bond forming/breaking

- Takes up a large portion of the total computation time
  - tabulate long range energy & forces
  - linear interpolation
  - large table $\rightarrow$ good approximations, reasonable memory usage
  - more on our gains later . . .

Summing Alltogether: Net Force

- Let E_i be the sum of energies from all interactions involving atom i

- Let r_i denote the position of atom i

- $F_i = \frac{\partial E_i}{\partial r_i}$

- Problem:
  - bonded energy expressions include $BO_{ij}(BO'_{ij}, \Delta'_i, \Delta'_j)$ terms

  - $\frac{\partial BO_{ij}}{\partial r_k}$ arise in every bonded interaction

  - $\frac{\partial BO_{ij}}{\partial r_k} = c_1 \cdot \frac{\partial BO'_{ij}}{\partial r_k} + c_2 \cdot \frac{\partial \Delta'_i}{\partial r_k} + c_3 \cdot \frac{\partial \Delta'_j}{\partial r_k}$

  - $\frac{\partial BO_{ij}}{\partial r_k} \neq 0, \forall k \in \text{bonds}_i \cup \text{bonds}_j$, huge memory overhead!

  - even if we choose to store them, very time consuming to compute each single bonded interaction!

Summing Alltogether: Net Force - Solution

- idea: distribution law of multiplication over summation

- let $C_0, \ldots, C_n$ be the coefficients of $\frac{\partial \mathsf{BO}_{ij}}{\partial r_k}$ arising in different interactions

- re-write $\sum_t C_t \times \frac{\partial \mathsf{BO}_{ij}}{\partial r_k}$ as $\frac{\partial \mathsf{BO}_{ij}}{\partial r_k} \times \sum_t C_t$

- while computing interactions, accumulate C_t's in C_{ijk}

- delay computation of $\frac{\partial \mathsf{BO}_{ij}}{\partial r_k} \times \sum_t C_t$ and forces due to them until C_{ijk}'s are determined

- no additional storage, important savings in CPU time

Implementation: Additional Features

- Modular implementation
  - a different force field can be adopted by plugging-in new interaction routines

- NVE, NVT and NPT ensembles **ayg: explain these terms first and what it takes to implement them**

- Compressible custom trajectory format

- Tools for performing common analysis
  - detection of reactions (on-the-fly)

  - property calculations such as drift coefficient, dipole moment (on-the-fly)

  - distributions of bond lengths, strengths, valence angles, charges, etc. (over the trajectory file)

Applications

- Hexane Simulations: Validation and Performance Analysis
  - Preparation of systems

  - Effects of QEq tolerance on accuracy and performance

  - Effects of tabulating long range interactions on accuracy and performance

  - Hexane structure verification

  - Scalabilty of ReaxFF compared to *ab-initio* and classical MD

- Corrosion of silica surface in water (in collaboration with Dr Pandit's group)

- Measuring the strain tensor of Si/Ge nanobar (in collaboration with Dr Strachan's group)

Hexane Simulations: Preparation od the Systems

- Hexane: C_6H_{14}, hydrocarbon, constituent of gasoline

- Initial configuration setup:
  - Very large box compared to the ideal volume → reduces overlaps!
  - Randomly spread copies of a model hexane molecule
  - Rotations of the model molecule around x,y and z axis to increase randomness
  - Various system sizes for scalability analysis (343,512,1000,1728, 3375 molecules)

- Energy minimization and NPT simulations using Gromacs
  - brings the systems to the ideal volume quickly
  - output to be used for ReaxFF studies and scalability analysis

- Energy minimization and NVT equilibration using ReaxFF
  - added H to Gromacs output conf using the Avogadro program
  - energy minimization for 2.5 ps
  - NVT equilibration at 200 K for 2.5 ps
 - QEq tolerance set to $1e - 8$ to be safe

Hexane Simulations: Effect of QEq Tolerance on Accuracy

- Chosen: hexane system with 343 molecules = 6860 atoms

- Restart from the system equilibrated at 200 K

- How to determine the "right" tolerance:
  - observe how the same system evolves over time at different QEq tolerances
  - pick the highest one with reasonable accuracy
  - $\text{tol}_1 = 1e - 3, \text{tol}_2 = 1e - 4, \text{tol}_3 = 1e - 8 \rightarrow$ control run

effect of QEq tol on total energy

effect of QEq tol on system temperature
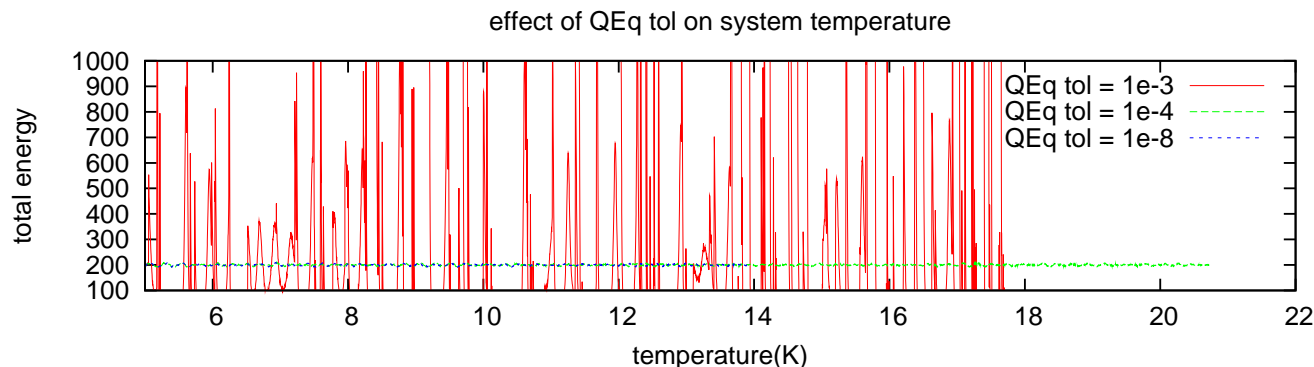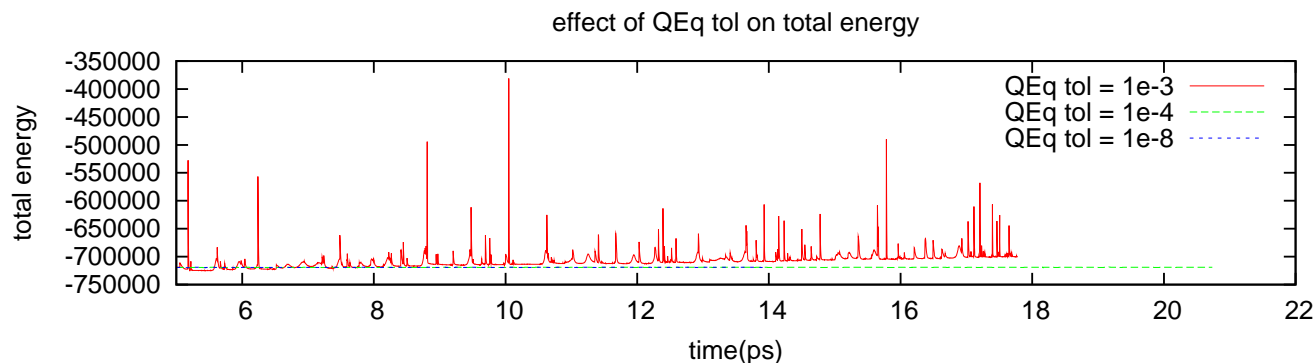
Hexane Simulations: Effect of Tabulation on Accuracy

$tol_2 = 1e - 4$ looks good enough, now turn on tabulation of long range interactions, too!

effect of tabulation on total energy

effect of tabulation on system temperature
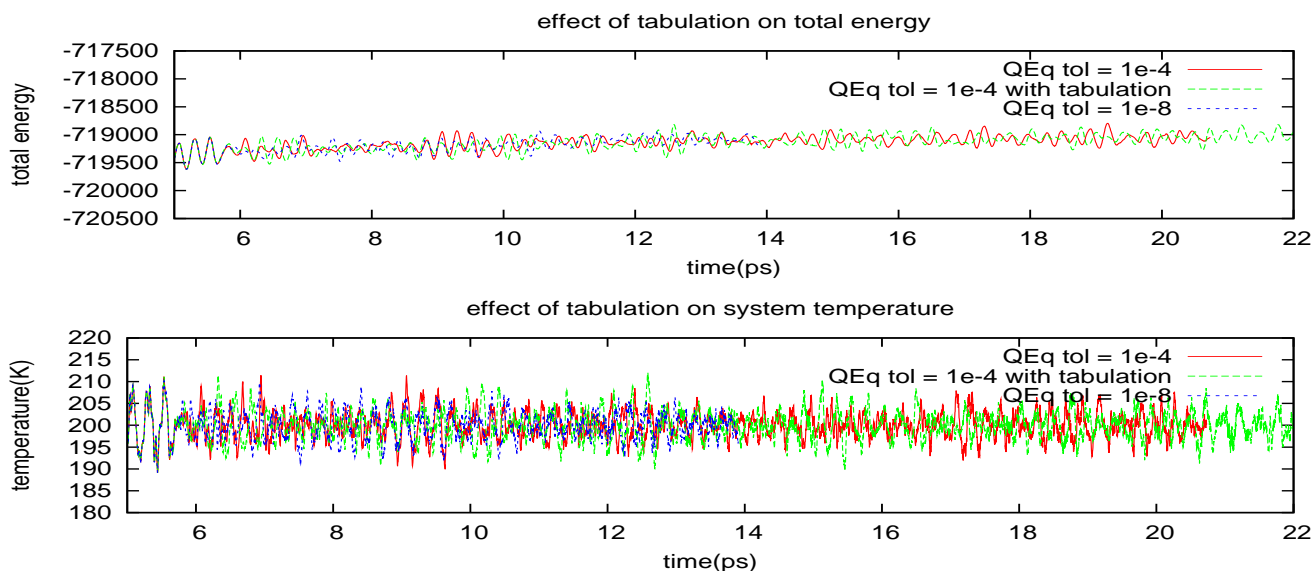
More in depth comparison shows they are almost identical:

property	$tol_3 = 1e - 8$	$tol_2 = 1e - 4$	$tol_2 = 1e - 4$ **with opt.**
C-H bond	1.09 ± 0.01	1.09 ± 0.01	1.09 ± 0.01
C-C bond	1.57 ± 0.01	1.57 ± 0.01	1.57 ± 0.01
<C-C-C	108.0 ± 2.9	107.9 ± 2.9	108.0 ± 2.9
<C-C-H	111.0 ± 0.0	111.0 ± 0.0	111.0 ± 0.0
<H-C-H	106.6 ± 0.0	106.6 ± 0.0	106.6 ± 0.0
q_{C-tip}	-0.171	-0.171	-0.171
q_{C-mid}	-0.080	-0.080	-0.080
q_{H-tip}	0.040	0.040	0.040
q_{H-mid}	0.040	0.040	0.040

Hexane Simulations: Profiling Analysis & Scalability

- Different qeq tolerances, with/without tabulation
- Used the $head.cs$ cluster except for the first case

ayg: What are the numbers here.. are they times in seconds?

	total	neighbors	bonded	nonb	QEq	matvecs	QEq%
tol=$1e-4$ w/opt§	0.93	0.22	0.14	0.22	0.34	9.9	37%
tol=$1e-4$ w/opt	3.26	0.47	0.32	0.65	1.78	9.9	55%
tol=$1e-4$	4.41	0.45	0.31	1.92	1.67	9.9	38%
tol=$1e-6$ w/opt	3.35	0.45	0.31	0.59	1.97	13.6	59%
tol=$1e-6$	4.65	0.45	0.31	1.92	1.95	13.5	42%
tol=$1e-8$	7.56	0.44	0.30	1.91	4.89	46.8	65%

ayg: Dont say Dell Studio.. instead, say what the processor is, etc. §Architecture can make a huge difference. Same system with same parameters on a Dell Studio XPS with $2.67GHz$ quad-core $i7$ processor and $1066MHz$ memory.

- Lessons learnt:
  - QEq tolerance is crucial for accuracy
  - arbitrarily large QEq tolerance might cause QEq domination **ayg: What does the above bullet mean?**
  - QEq is just a precursor to electrostatics yet takes up at least one third of total time!
 - QEq must be improved to make ReaxFF scalable.

Hexane Simulations: Validation

Compare the structure of our hexane molecules to those of experimental results in the literature and *ab-initio* simulation:

property	ReaxFF	experimental†	*ab-initio*‡
C-H bond	1.09 ± 0.01	1.118 ± 0.006	1.100
C-C bond	1.57 ± 0.01	1.533 ± 0.003	1.533
<C-C-C	108.0 ± 2.9	111.9 ± 0.4	114.2
<C-C-H	111.0 ± 0.0	109.5 ± 0.5	109.5
<H-C-H	106.6 ± 0.0	NA	106.5
$q_{C\text{-tip}}$	-0.171	NA	-0.205
$q_{C\text{-mid}}$	-0.080	NA	0.033
$q_{H\text{-tip}}$	0.040	NA	0.047
$q_{H\text{-mid}}$	0.040	NA	$-0.10 \sim 0.10$

†R. A. Bonham, L. S. Bartell, and D. A. Kohl. "The Molecular Structures of n-Pentane, n-Hexane and n-Heptane" J. Am. Chem. Soc., 1959, 81 (18), $4765 - 4769$

‡geometry optimization of an isolated hexane using CPMD v3.13.2 with PBE Troullier-Martins pseudopotentials

Hexane Simulations: Scalability Analysis
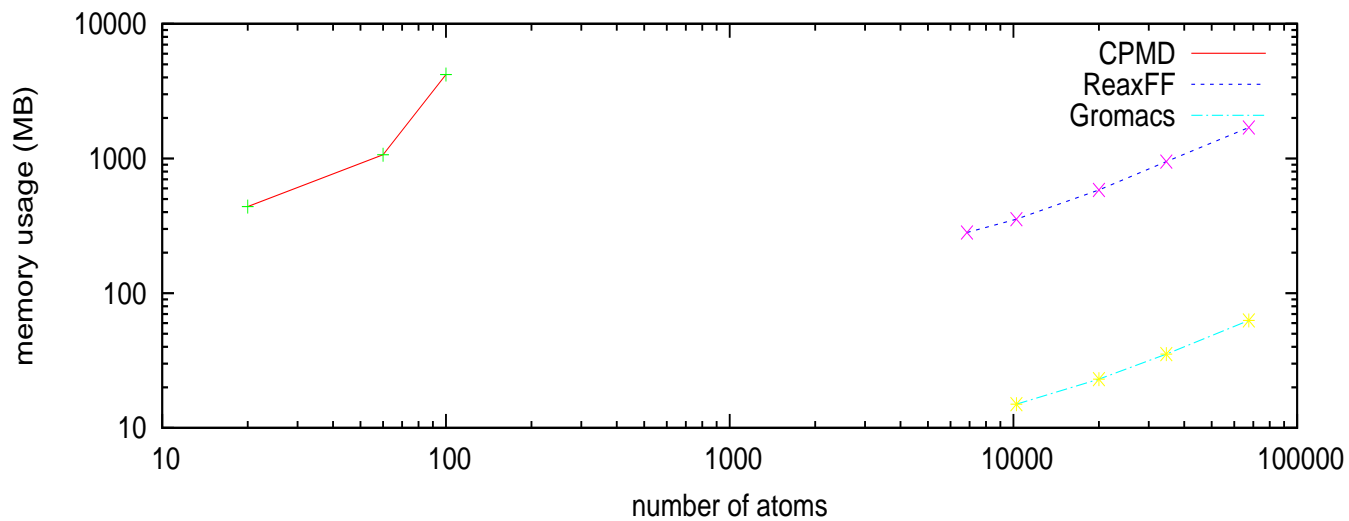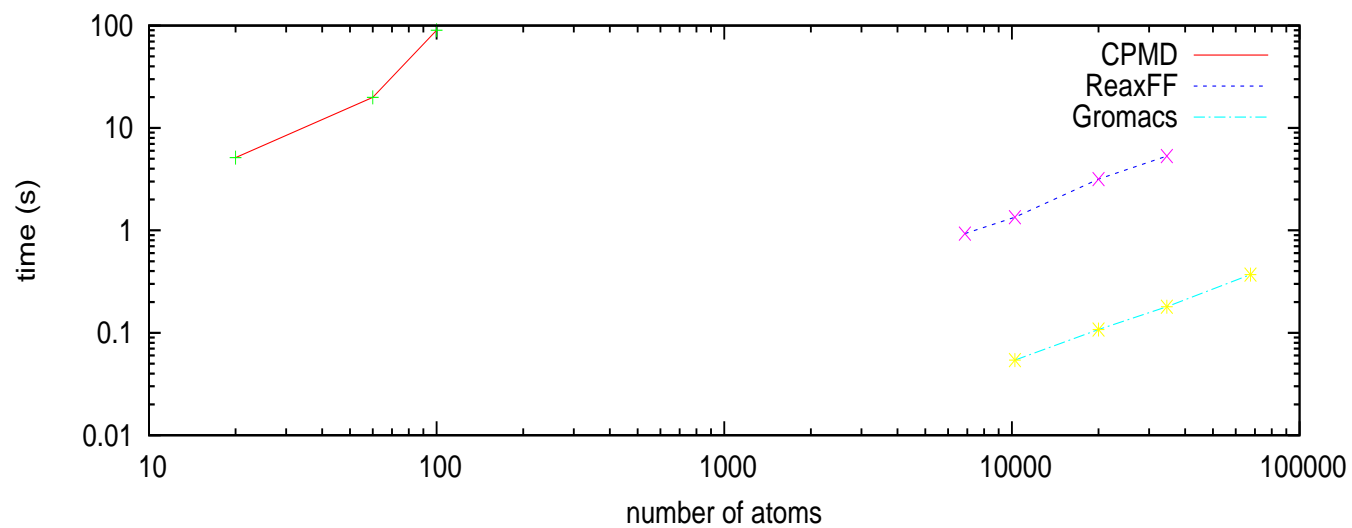
Silica Surface Corrosion in Water

Si/Ge Nanobar

Ongoing & Future Work

- ParallelReax
  - complete & verify the implementation
  - a large scale application (maybe with the PRISM device)

- Integration into LAMMPS (a well-known, widely used MD package from SNL)
  - QEq integrated independently
    * opens the door for polarizable-ff inside LAMMPS
    * compatibility issues to be sorted out!

- Better solvers for QEq
  - block Jacobi type pre-conditioner
  - inner-outer schemes (use a lower cutoff inner solve to precondition outer solve)
  - use a fast multipole-type preconditioner.

- Tight relations among items on the agenda
  - better QEq solvers necessary for scalable ParallelReax
  - better QEq solvers necessary for QEq in LAMMPS
  - completion of ParallelReax necessary for Reax in LAMMPS
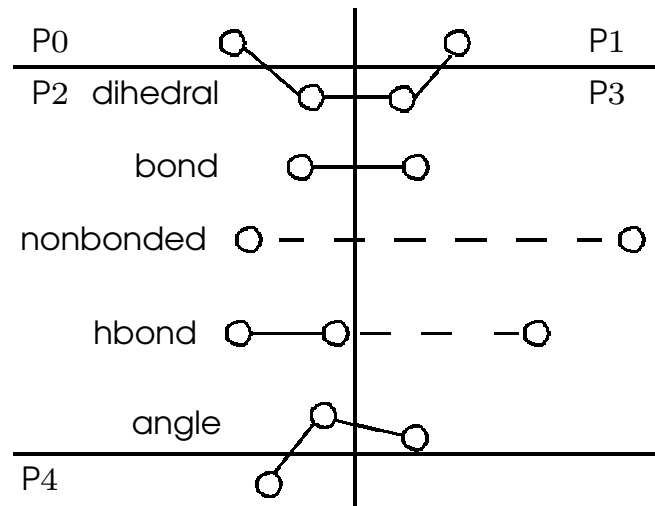
Parallelization of ReaxFF

- A draft version of ParallelReax
  - domain decomposition technique
    * **repeat:** get my share of atoms $\rightarrow$ communicate boundaries $\rightarrow$ compute forces $\rightarrow$ move my atoms
  - not fully verified
  - inefficient handling of processor boundaries

Two big challenges:

- Parallelization of QEq
  - even CG needs at least 4 communications per iteration!
  - parallel GMRES (with *Householder orth.*) is even worse
  - QEq will dominate even more
  - definitely need: better solvers for QEq

- Processor boundaries
  - avoid double computation at boundaries!
  - avoid thick boundaries, retain accuracy for any system!

Parallelization: Our Solution Approaches

- Parallelization of QEq: better solvers for QEq
  - block Jacobi type pre-conditioner
  - inner-outer schemes

- Avoid double computation
  - coordination through the mid-point rule
    * bond(i, j): owner$(\frac{1}{2}(r_i + r_j))$
    * $<(i, j, k)$: owner(r_j)
    * dihedral(i, j, k, l): owner$(\frac{1}{2}(r_j + r_k))$
    * hbond(X, H, Z): owner(r_H)
    * nonbonded(i, j): owner$(\frac{1}{2}(r_i + r_j))$
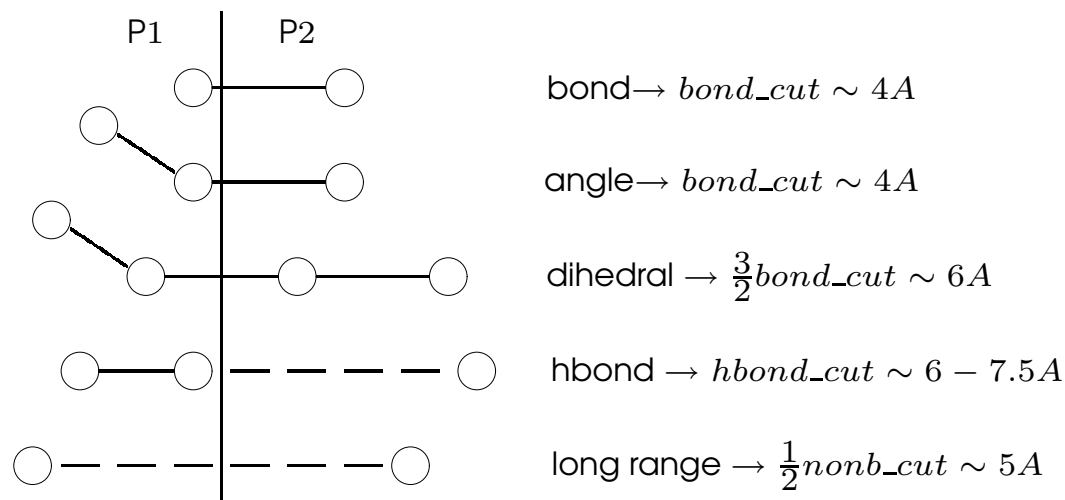
Parallelization: Our Solution Approaches

- Why avoid thick boundaries?
  - a 20 Å cubic box ≈ 1000 atoms
  - assume $nonb\_cut$ long boundaries $\rightarrow$ a few thousand atoms/processor $\rightarrow$ upto a couple of seconds per iteration!
  - definitely need: thin boundaries

- How to avoid thick boundaries?
  - Mid-point rule to the rescue!
  - boundary thickness: $max(\frac{3}{2}bond\_cut, hbond\_cut, \frac{1}{2}nonb\_cut)$
  - gets better if no hydrogen bonds present: $max(\frac{3}{2}bond\_cut, \frac{1}{2}nonb\_cut)$

worst case scenarios

bond$\rightarrow bond\_cut \sim 4A$

angle$\rightarrow bond\_cut \sim 4A$

dihedral $\rightarrow \frac{3}{2}bond\_cut \sim 6A$

hbond $\rightarrow hbond\_cut \sim 6 - 7.5A$

long range $\rightarrow \frac{1}{2}nonb\_cut \sim 5A$

Conclusions

Current Userbase

- Dr Pandit's group at USF
  - silica-water systems

- Dr Strachan's group at Purdue
  - Si/Ge nanobar

- Dr Buehler's group at MIT
  - Silica cracking with strain

- Dr van Duin at PennState

- Dr Goddard's group at Caltech

- Dr Aluru's group at UIUC