Massively-parallel molecular dynamics simulations on New York Blue

David F. Green

Stony Brook University
State University of New York

Department of Applied Mathematics & Statistics
Graduate Program in Biochemistry & Structural Biology

NY Blue Tutorial
Stony Brook University
08/04/08

Many thanks to …

- **Current Research Group**
  - **Noel Carrascal***
  - Jessica Chaffkin
  - Yukiji Fujimoto
  - Ji Han
  - Tao Jiang
  - **Vadim Patsalo***
- **Former Group Members**
  - Jonathan Cheng
  - Ryan TerBush
  - Yong Yu
- **Collaborators**
  - Dan Raleigh (Stony Brook)
  - Steve Skiena (Stony Brook)
  - Steve Smith (Stony Brook)
  - Yaoxing Huang (Aaron Diamond AIDS Research Center)
- **Support:**
  - Stony Brook Dept. of Applied Math & Statistics
  - Microsoft Research
  - New York Center for Computational Science

The Green Lab
http://www.ams.sunysb.edu/~dfgreen/

Computational Biology in AMS
http://compbio.ams.sunysb.edu/

Biomolecular Simulation

- Simulation of biological macromolecules is a key area of interest:
  - Understand the dynamic mechanisms of macromolecular function (protein folding, catalysis, molecular machines)
  - Predict the energetics of various biological processes (ligand association, protein stability)
  - Design novel molecules with particular properties (drug design, protein engineering)

Types of Calculation

- Energy calculations
  - Single point (state) vacuum energies
  - Solvated (explicit or implicit) state free energies
  - Ensemble-averaged free energies
- Conformational search
  - Constant temperature ensembles
    - Molecular dynamics or Metropolis Monte Carlo
  - Brute force enumeration
  - Global optimization/"intelligent" search
    - Simulated Annealing, Dead-End Elimination, Genetic Algorithms

Biomolecular Energetics

- Molecular energetics are properly described by quantum mechanics
  - Much too costly for macromolecules
- Molecular mechanics force fields are classical approximations to the QM energy
  - Vibrations between bonded atoms described by springs; rotation about bonds described as sinusoidal functions; interactions between non-bonded atoms described by Coulomb's Law and Van der Waals interactions.

Molecular dynamics – Integrating Newton's Laws of Motion

- Energetic models give $E = E(\mathbf{x})$, where $\mathbf{x}$ is the Cartesian coordinates of all atoms in the system.
- Force is given by the gradient of the energy, $\mathbf{F}(\mathbf{x}) = -\nabla E(\mathbf{x})$.
- Newton's Laws then relate the dynamics of the system to the forces on each atom:
  - $\mathbf{F}_i(\mathbf{x}) = m_i \mathbf{a}_i(\mathbf{x}(t))$
  - $\mathbf{a}_i(\mathbf{x}(t)) = d\mathbf{v}_i(t)/dt$
  - $\mathbf{v}_i(t) = d\mathbf{x}_i/dt$
- This system of ODE's can be solved by any of many standard integration schemes.
- By the ergotic principle, a converged MD simulation gives a constant temperature, equilibrium ensemble.

Major issues in molecular dynamics: Solvent

- Biology occurs (usually) in a salty, aqueous environment
  - Accurate simulations require the solvent to be treated appropriately
  - Most accurate approach involves explicitly representing both water and mobile ions in the simulation; periodic boundary conditions are generally used to minimize artifacts from a finite-sized simulation size.
    - System sizes become 25,000-100,000 atoms or more, in the unit cell.
  - Alternative approaches replace explicitly represented solvent with implicit continuum models
    - The Poisson-Boltzmann equation describes electrostatic interactions with a polarizable continuum; the Generalized Born model gives an approximation to the PB solution.
    - Cavitation and solute-solvent VDW interactions are often approximated as proportional to Surface-Area.

Major issues in molecular dynamics: Time scales

- Biomolecular events occur over a wide range of time scales
  - Bond vibrations occur on the femtosecond time scale.
  - Rotations of chemical groups happens over picoseconds.
  - Mobile loops sample conformations over nanoseconds.
  - Global conformational transitions may take tens or hundreds of nanoseconds (or more).
  - Protein folding generally takes upwards of milliseconds.
- Accurate descriptions of energetics requires long simulations (100+ ns); accurate simulation of dynamics requires time steps of 1 or 2 fs.
  - At least 10^7 steps are needed (often 10^8 or more).

Major issues in molecular dynamics: Long range interactions

- Non-bonded interactions exist between all pairs of atoms
  - Computational expense of the complete energy (or forces) would increase proportionally to N^2.
  - Van der Waals interactions fall off with $1/R^6$, and thus can be safely truncated at moderate distances.
  - Coulombic interactions fall off with $1/R$, and forces with $1/R^2$; since volume in a shell at R increases with R^2, the total electrostatic energy is not unconditionally convergent. Truncation could lead to artifacts.
  - Particle-mesh Ewald techniques both address the conditional convergence and allow long range electrostatic interactions to be computed with the FFT (charges are mapped on a regular lattice).
  - Fast multipole methods can scale better than the FFT, but are not efficiently implemented in most simulation packages.

Major issues in molecular dynamics: Inter-processor communication

- Forces are dependent on the positions of all other atoms
  - Each time step requires a force evaluation, which would require knowledge of all other atomic positions.
  - With Ewald summation, energy evaluation involves two distinct phases
    - Interactions with near neighbors (bonded interactions, and short range electrostatic and VDW interactions): This term requires knowledge of only near neighbor atomic positions. With neighbor lists, required inter-node communication can be minimized; but only to a point. Theoretical scaling is $O(N)$.
    - Ewald sum for long range electrostatics: All atomic positions are required in updating the Ewald mesh. A 3-D FFT on the grid must them be performed. Theoretical scaling is $O(N_g \log N_g)$.
- In large systems, the Ewald sum may be a significant fraction of the computational cost, and FFT scaling may influence performance.

Choices in Molecular Dynamics

- Choice of force field
  - CHARMM, AMBER, OPLS, Gromos, and more.
    - All modern force-fields are quite reasonable, and perform well
    - AMBER has good support for small molecules (GAFF).
    - CHARMM has good support for lipids and carbohydrates.
- Choice of simulation package
  - Highly integrated, multi-functional simulation packages
    - CHARMM, AMBER
  - Performance-optimized packages with limited functionality
    - NAMD, Gromacs

Choices in Molecular Dynamics – Our Current Workflow

- All-atom CHARMM is used as the force field of choice in all simulations:
  - Param22/27 for proteins/nucleic acids; CSFF for sugars.
- CHARMM (on Seawulf and local servers) is used for system setup and for post-simulation analysis.
- NAMD (on NY Blue) is used for production dynamic simulations.

System setup – Essential steps

Structures obtainable from the Protein Databank

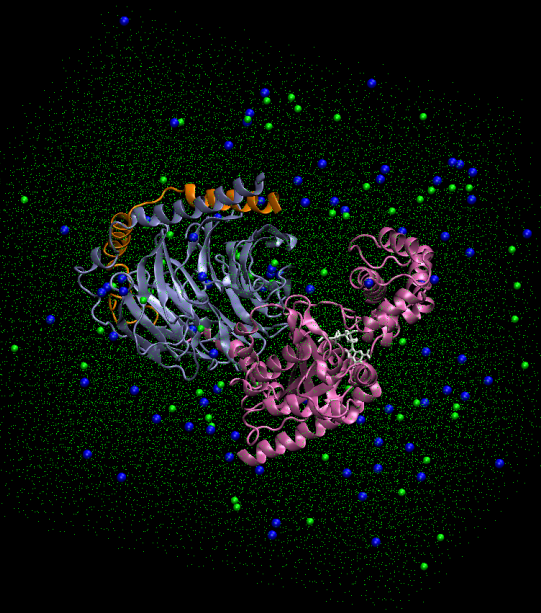
- These generally do not include hydrogen atoms; they can be missing some atoms; ambiguities exist on the orientation of amides (Asn, Gln) and histidine; and protonation states are underdetermined.

1. Assign protonation and amide/His flip states (REDUCE).
2. Place hydrogen atoms, and build missing atoms (CHARMM).
3. Surround system with waters from a pre-equilibrated simulation, giving a minimum 10 Å buffer on each side (CHARMM).
4. Add salt (Na+ and Cl-) in random positions to a total concentration of 0.145 M (1 NaCl per 376 waters); adjust ion concentrations to give the system a neutral net charge (CHARMM).
5. Output XPLOR format PSF and PDB files for NAMD (CHARMM).

Example 1 – Heterotrimeric G-Protein

- A key signaling protein complex

	Residues	Atoms
Gα	349	5578
Gβ	339	5121
Gγ	54	849
GDP	1	40
Na+	80	80
Cl-	63	63
Water	27551	82653
Total	28437	94384



Example 1 – Heterotrimeric G-Protein

- Simulations run on multiple states of the system
  - Full complex (trimer)
  - Unbound Gα (with GDP or GTP.Mg)
  - Unbound Gβγ

	Atoms	Box size
Gαβγ.GDP	94384	112x98x81
Gα.GDP	52123	100x77x65
Gα.GTP.Mg	51563	100x76x64
Gβγ	47101	93x78x61

2 fs time step

12 (14) Å cutoff

SHAKE on H atom bond lengths

Output every 2 ps

Example 1 – NAMD Input Minimization and heating

```
# Adjustable Parameters
coordinates          ../1gia_ions_box.pdb
structure            ../1gia_ions_box.psf

set temperature     100
set outputname      1gia_1R
firsttimestep       0

# Simulation Parameters
#############################################################

# Input
paraTypeCharmm              on
parameters          /gpfs/home2/ncarrasc/usr/par_all27_prot_na_sugar.prm
temperature         $temperature
```

Example 1 - NAMD Input
Minimization and heating

```
# Force field Parameters
exclude                 scaled1-4
1-4scaling              1.0
cutoff                  12.
switching               on
switchdist              10.
pairlistdist            14

# Integrator Parameters
timestep                2.0
rigidBonds              all
nonbondedFreq           1
fullElectFrequency      1
stepspercycle           20

# Temperature Control
reassignFreq            250
reassignTemp            100
reassignIncr            5.
reassignHold            300
```

```
# Periodic Boundary Conditions
cellBasisVector1     101.3   0.0   0.0
cellBasisVector2       0.0  77.4   0.0
cellBasisVector3       0.0   0.0  65.4
cellOrigin             0.0   0.1   0.1

wrapAll              on

# PME   (for full system periodic
 electrostatics)

PME                        yes
PMEGridSizeX               102
PMEGridSizeY               78
PMEGridSizeZ               72
```


Example 1 - NAMD Input
Minimization and heating

```
# Constant Pressure Control

useGroupPressure        yes
useFlexibleCell         no
useConstantArea         no

langevinPiston          on
langevinPistonTarget    1.01325
langevinPistonPeriod    200.
langevinPistonDecay     100.
langevinPistonTemp      $temperature
```

```
# Output

outputName              $outputname

restartfreq             500
dcdfreq                 1000
xstFreq                 1000
outputEnergies          1000
outputPressure          1000
```

```
# Minimization & Temperature equilibration
########################################################

minimize                240
reinitvels              $temperature

run 100000
```

The Green Lab **Computational Biology in AMS**
http://www.ams.sunysb.edu/~dfgreen/ http://compbio.ams.sunysb.edu/

Example 1 - NAMD Input
Changes for production run

```
# Continuing a job from the restart files
set temperature    300

set inputname      ../1R/1gia_1R
binCoordinates     $inputname.restart.coor
binVelocities      $inputname.restart.vel
extendedSystem             $inputname.xsc

firsttimestep      100000
```

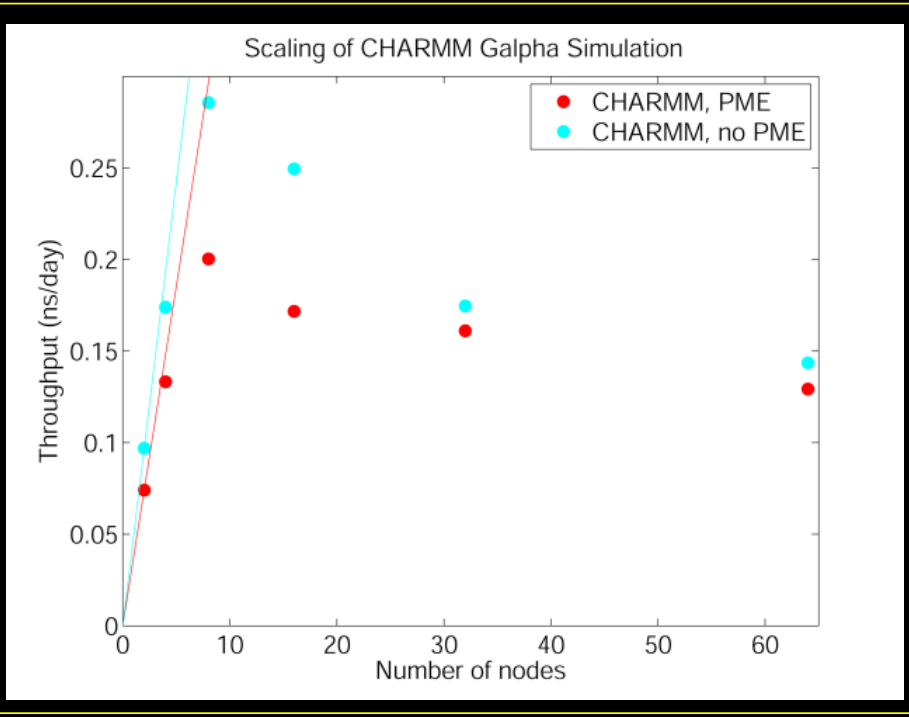
```
# Constant Temperature Control
langevin            on    ;# do langevin dynamics
langevinDamping     5     ;# damping coefficient (gamma) of 5/ps
langevinTemp        $temperature
langevinHydrogen    no    ;# don't couple langevin bath to hydrogens
```

```
# Execute dynamics
run 2000000
```

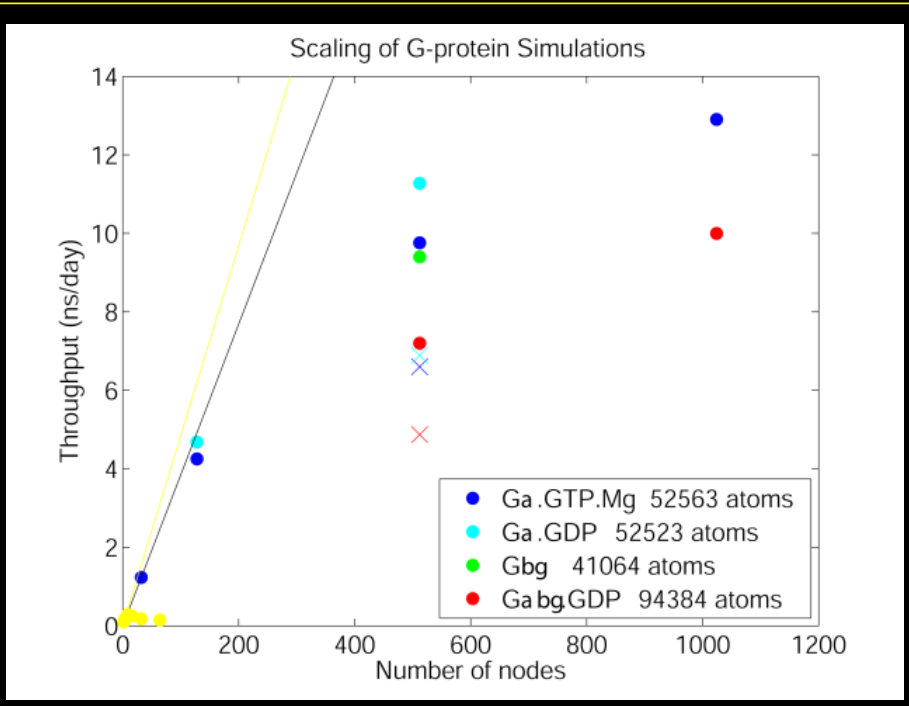

Example 1 - LoadLeveler Input

```
# @ job_type = bluegene
# @ class = normal
# @ executable = /gpfs/home2/ncarrasc/bin/mpirun32
# @ bg_partition = B512TB03
# @ arguments =  -exe /gpfs/home2/ncarrasc/bin/namd2 \
-cwd /gpfs/home2/ncarrasc/G-Prot/full_seq2/1GIA/1R \
-args "/gpfs/home2/ncarrasc/G-Prot/full_seq2/1GIA/1R/1R.in"
# @ initialdir = /gpfs/home2/ncarrasc/G-Prot/full_seq2/1GIA/1R
# @ input = /dev/null
# @ output = $(jobid).out
# @ error = $(jobid).err
# @ wall_clock_limit = 1:00:00
# @ notification = complete
# @ queue
```

Example 1- CHARMM on Seawulf

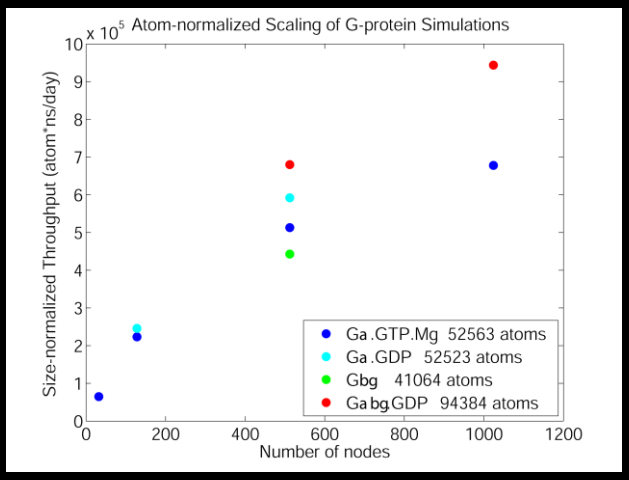


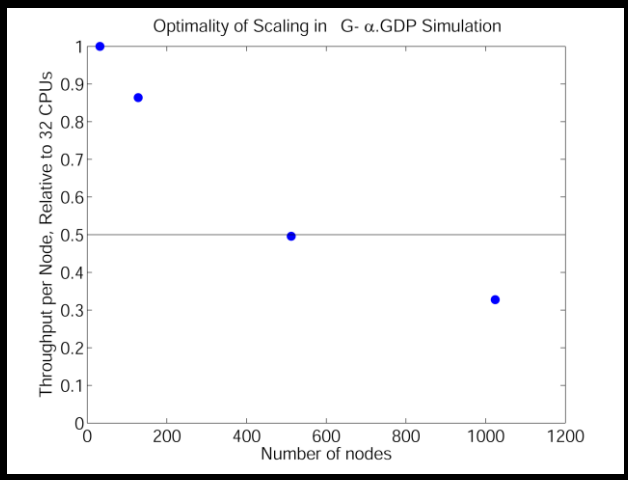
Example 1- NAMD on NYBlue



Caveat:
No PME,
except X

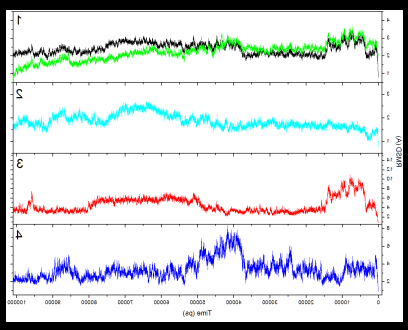
Example 1- Scaling with CPU

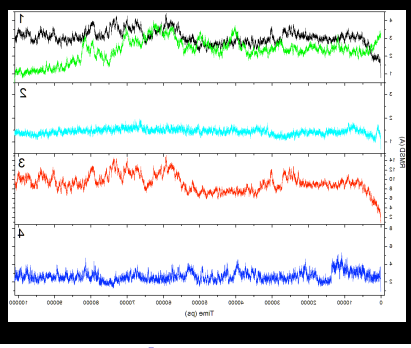


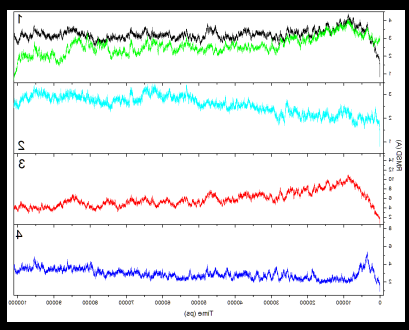


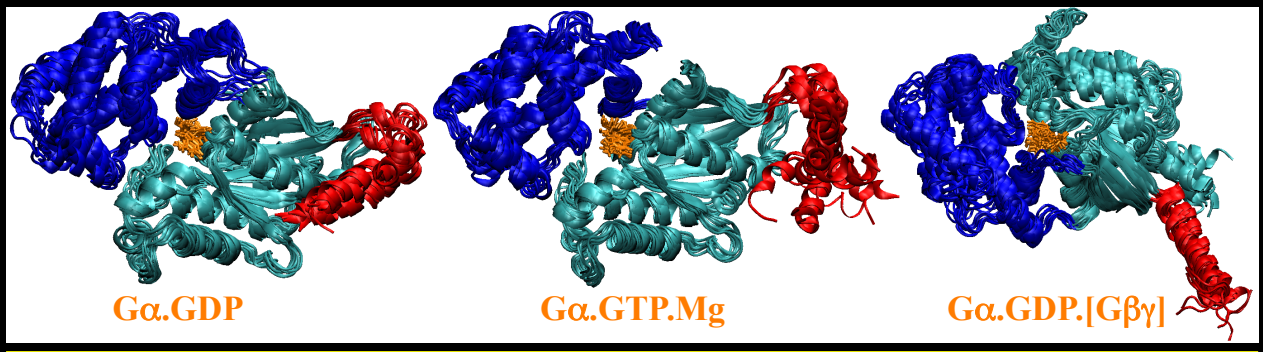
Caveat:
No PME

Example 1- Results





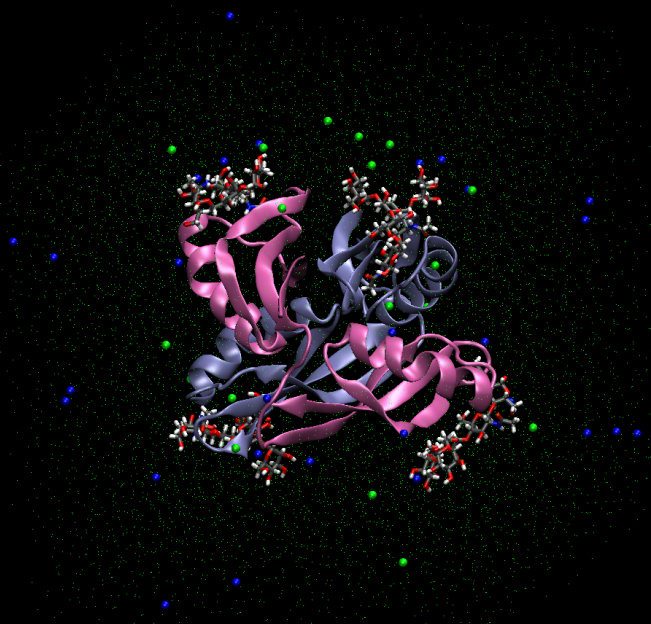




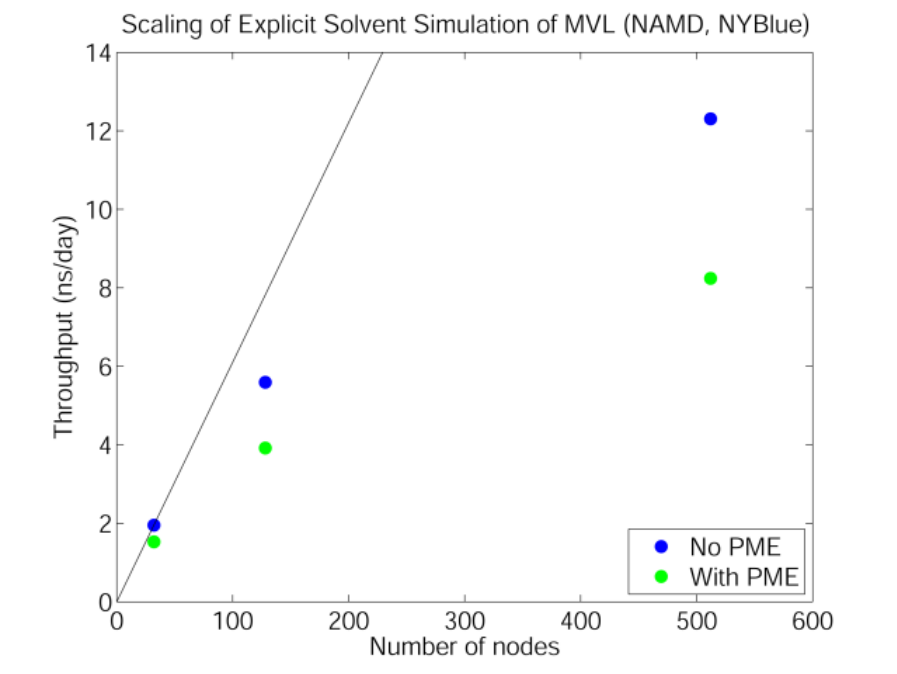
Example 2 - MVL

- An anti-viral carbohydrate binding protein

	Residues	Atoms
Protein	2x113	2x1705
Sugar	4x4	4x120
Na+	29	29
Cl-	19	19
Water	9399	28197
Total	9689	32135



Example 2 - MVL Scaling



2 fs time step

12 (14) Å cutoff

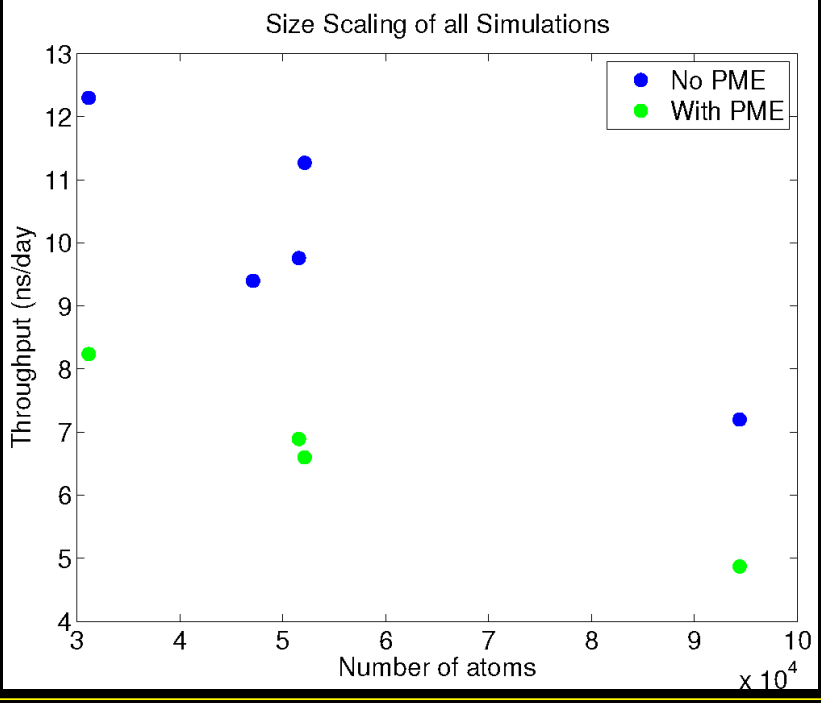
71x66x64 Å box

SHAKE on H atom bond lengths

Output every ps

72x66x66 FFT grid (when used)

Scaling with System Size



All at 512 nodes

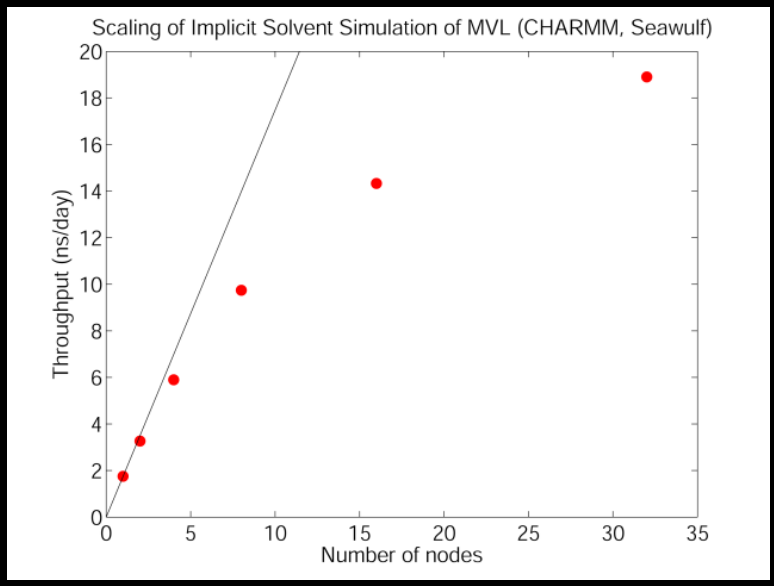
Implicit solvent models.

- In fully explicit solvent simulations, water molecules can consist of ~90% of the total system.
- The implicit-solvent Generalized-Born model thus allows the system size to be reduced by a factor of 10; although each step requires more computation.
  - The GB model involves an all-all calculation for computing effective Born radii; however this radii update need not be done every step.

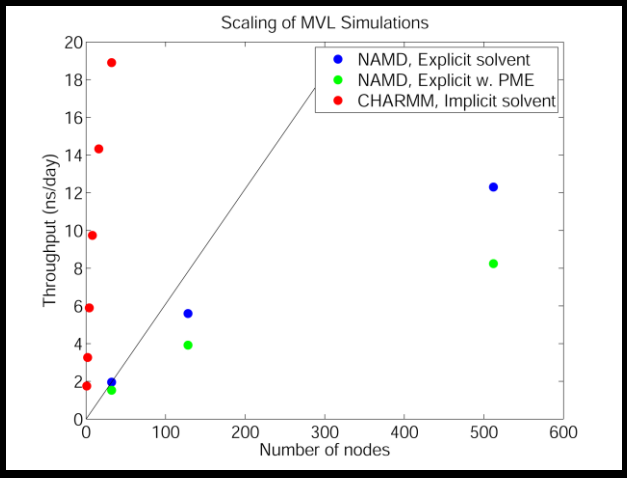
Implicit solvent models in CHARMM.

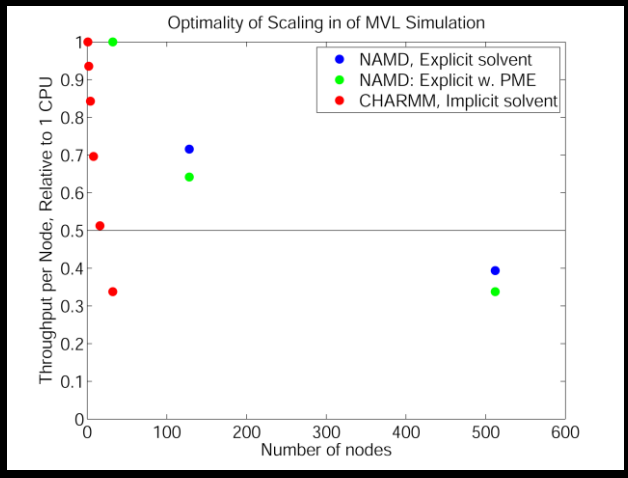
- GBSW module in CHARMM on Seawulf

	Residues	Atoms
Protein	2x113	2x1705
Sugar	4x4	4x120
Na+	0	0
Cl-	0	0
Water	0	0
Total	242	3890



Implicit vs Explicit Solvent Simulations





- Similar performance obtained for 32 times # of CPUs.
  - Seawulf CPUs are 3.4GHz Xeon; NYBlue are 700MHz PPC.

Key points + Future directions

- Key points for current use
  - Running NAMD on NY Blue is straightforward and gives good performance.
  - Care should be taken with system setup, and additional tools are needed.
  - Think carefully about simulation length and size of output.
    - 50,000 atoms, output every 2 ps => 300 MB per ns
- Increasing capability for MD simulations on NY Blue
  - Installation of WORDOM, a suite of MD analysis tools.
  - Compilation & installation of CHARMM.
  - Software for Poisson-Boltzmann calculations (MultiGrid PBE, and the ICE package).
  - Software for protein design (DEE/A*).

Thank you!

The Green Lab **Computational Biology in AMS**
http://www.ams.sunysb.edu/~dfgreen/ http://compbio.ams.sunysb.edu/