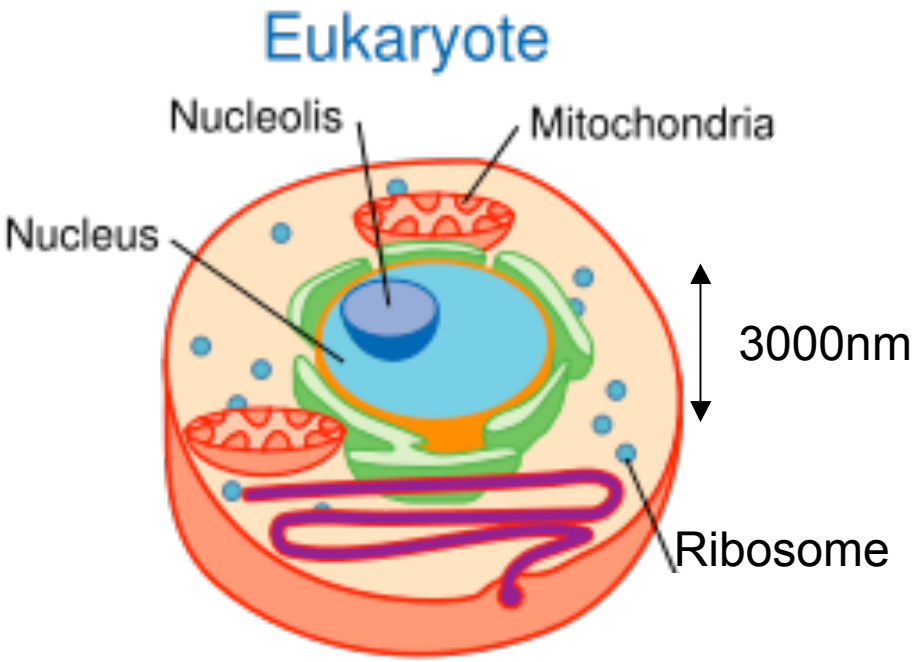
High Throughput High Performance Molecular Dynamics Simulations

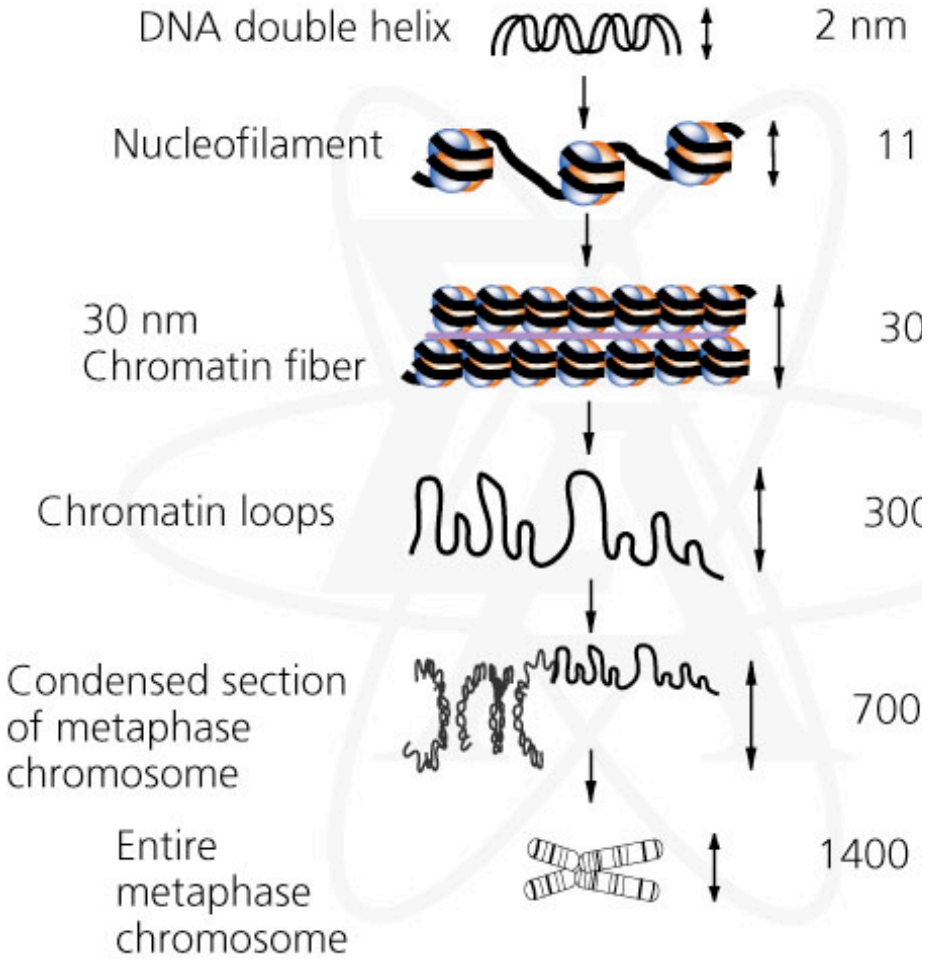
PETA Share
All Hands Meeting
LSU March 3, 2008

Thomas C. Bishop
Center for Computational Science
Tulane University
New Orleans, LA

Cell Biology 001

Cell Organization and Chromatin



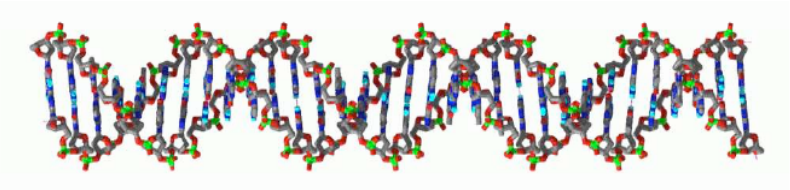


Wikipedia

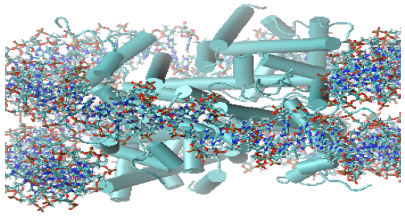
- *"Molecular Biology of the Cell" Alberts B., Bray D., Lewis J*

Folding DNA into Nucleosomes

Free DNA



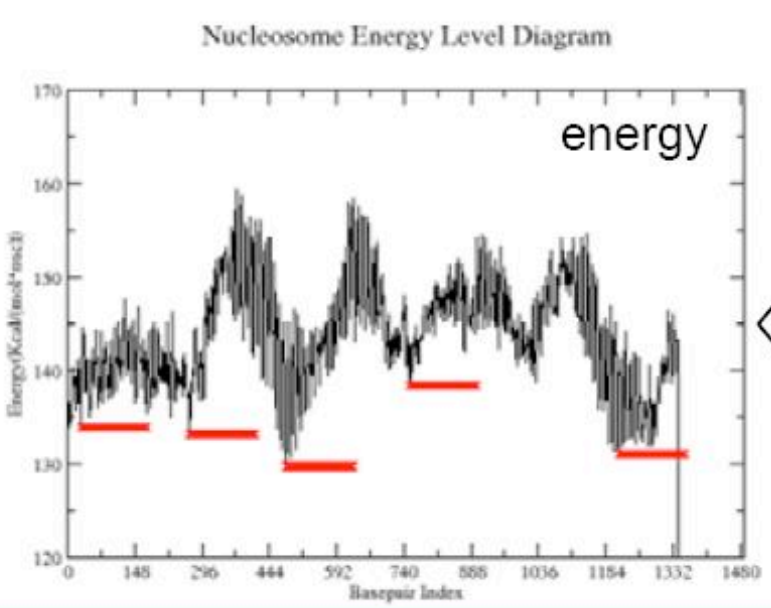
Folded DNA

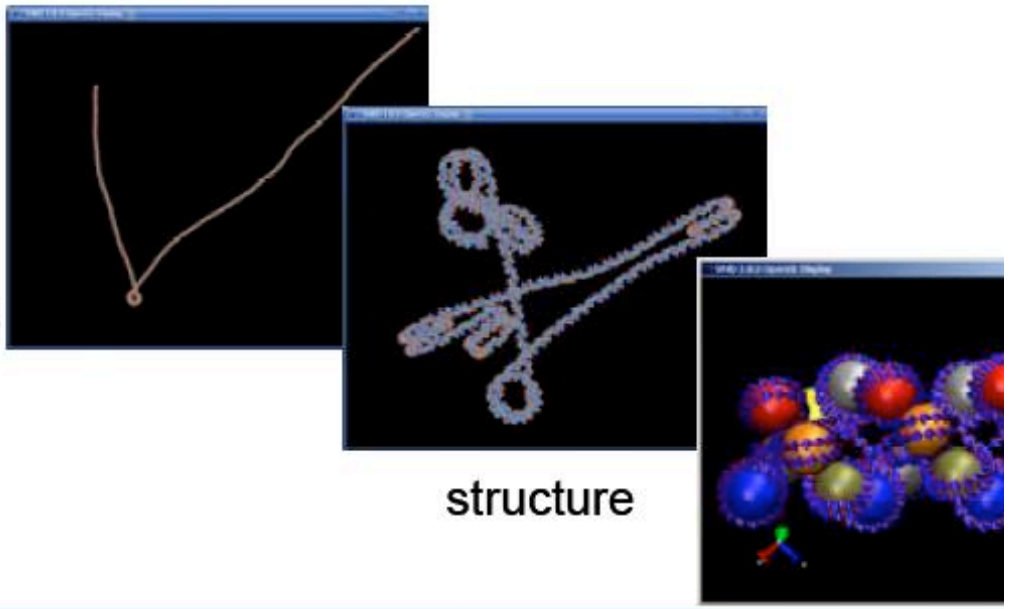


- any sequence will fold
- stability depends on ability of sequence to fold

Chromatin Structure: Prediction & Analysis

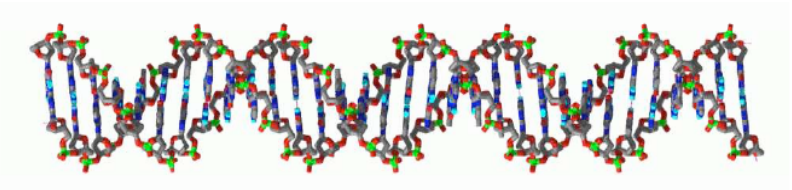
```
TGCTTCTTTT AAAAAAGAAA AAAGGGGGAA ATGCCGCGCC TTCCCGAGAG TGTCCTACAC TTAGGGGAGA AGCAGC
GGGACCCGTC TGCGCACAAA CGGATGAGCC CATCAGACAA AGAAAACTTG GCATAGCTCT GCTTTGCCTA GGGCTA
GGGAAGGGCT CTCACCCTTG ATTCTTTCAA TAATAACTCT TCTGTGATTC GGAGAACTCG ACCTTCCTCT TGGGGC
ACCACACCGC ATCGAATTTG TCCTTTAGAA ATAGAAATAA GAATGCTACC AATAAGACCA ATCCAATAGG TCGATT
GTCACTATTC TTTTAGTACT ATTTTTACTC AAATTCAGAA GTTAGAAATA GACGCTCAAA ATCAATTGAA GAGCAG
AAGGACTATT AAAAAGGGAA AAAAGAGTGT GTTTGTCAAA ATAGGAGACA GTATAGGGGA CCTTACATCT ACAGAC
AGATGCCCCC TTGACTTAAA TTGGGATAGG TGGGTCACAA TCAACGGCTA TAACCCTTTC GTGAAAGACT CGCCAG
AGACCTCCTT GGTGAAAGAC GACATGAAAC AACAGGTACA TGATTATATT TATCT
```



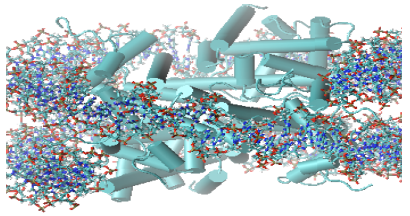


Fine Grain and Coarse Grain

Free DNA



Folded DNA



Atomic Model

molecular dynamics

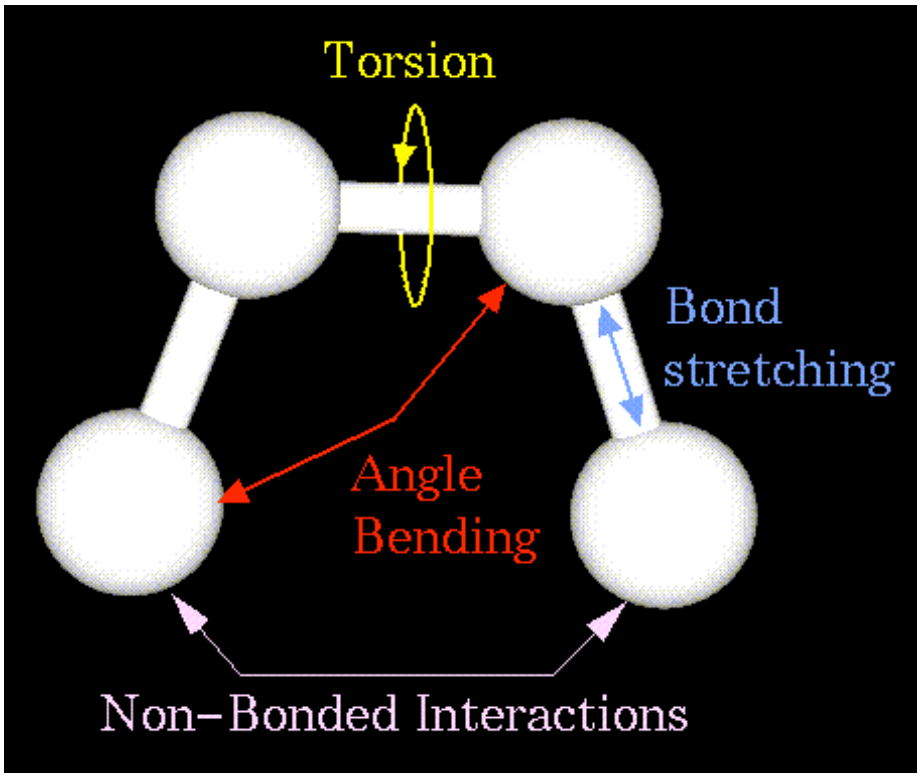
1 to 1000 nucleosomes

Coarse Grain

elastic deformation

entire genomes

Molecular Dynamics in a Nut Shell



$$E = \sum_{\text{bonds}} k_b \, (r - r_o)^2$$

$$E = \sum_{\text{angles}} k_\theta \, (\theta - \theta_o)^2$$

$$E = \sum_{\text{torsions}} A \, [1 + \cos(n\tau - \phi)]$$

Non-Bonded interactions take all the CPU time

$$E = \sum_i \sum_j \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + \sum_i \sum_j \frac{q_i \, q_j}{r_{ij}}$$

van der Waals term Electrostatic term

http://cmm.cit.nih.gov/modeling/guide_documents/molecular_mechanics_document.h

MD is a Mature Method

- algorithms defined/robust
- efficiently parallelized (1000 CPUS/ 3M atoms)
- force fields carefully evaluated
- data formats decided
- tools for visualization well developed

MD Practical Issues

- Pre-Simulation (1day)

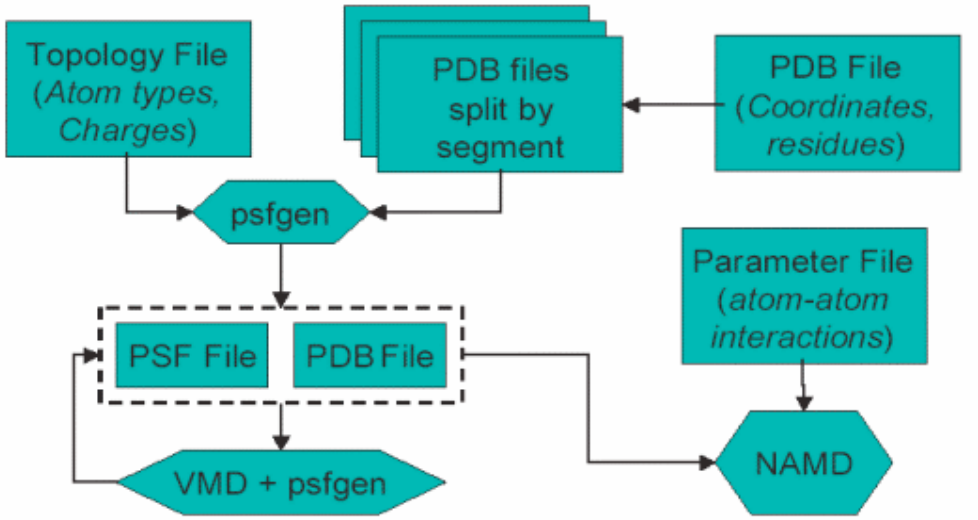


image from NAMD tutorial files
www.ks.uiuc.edu

- Simulation (1n
  1)Minimization
  2)Equilibration
  3)Dynamics

- Post Simulation Analysis (1wk)

Specific Data Notes

- High Throughput

  1300 systems/24,000ea

  ~40Gb

  Initialize Sys. ~1min/sys

  2000 step minimization

  2min/sys

  3 days compute time

- High Performance

  1system/155,000

  ~200Gb

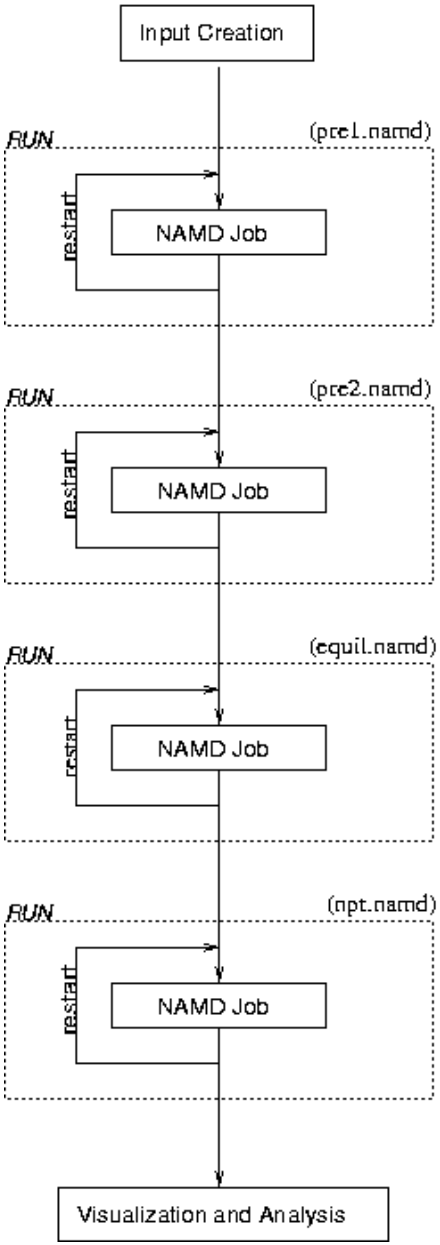
  Initialize Sys. 1day

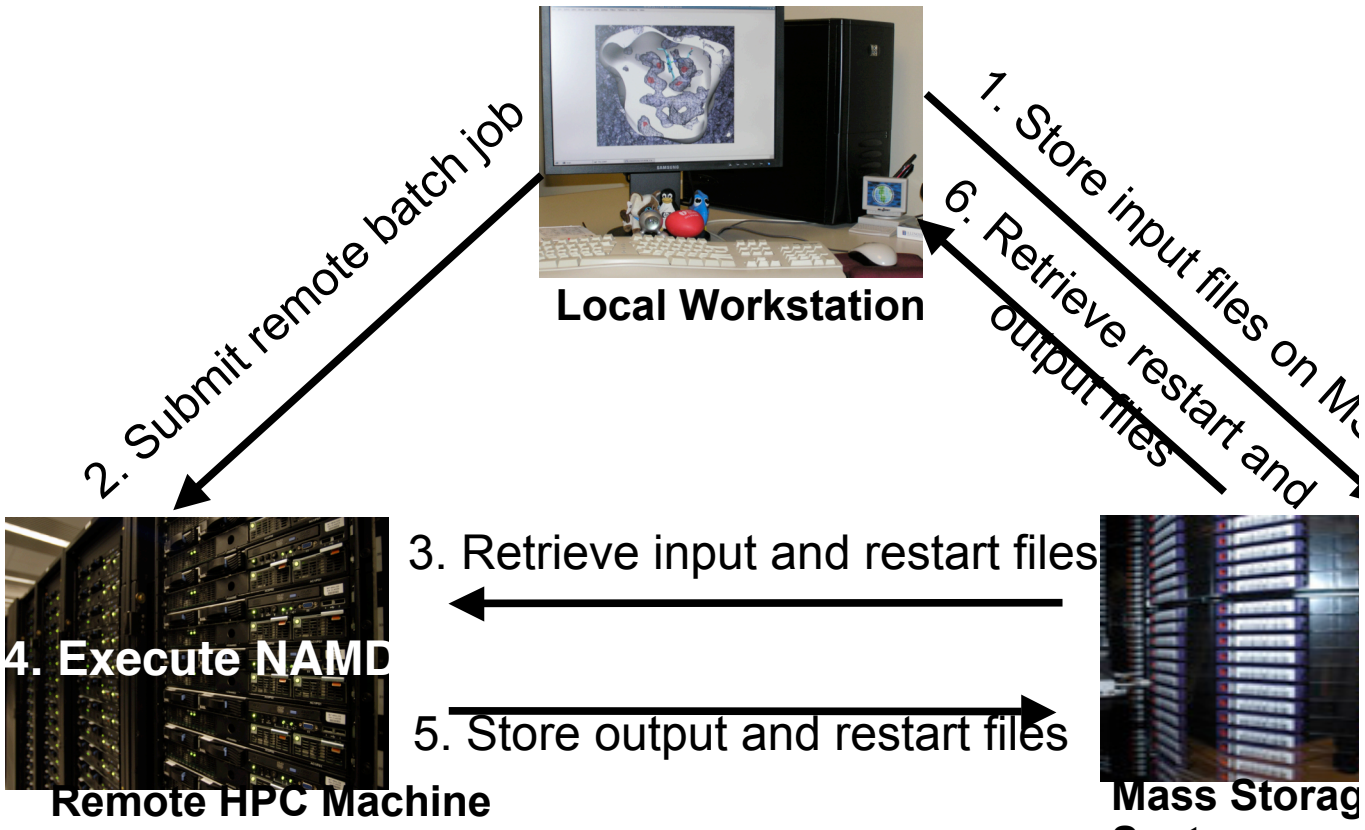
  20ns Simulation

  1day/run

  30 days on 64cpu

The NAMD-G Solution





images from NAMD-G presentation: Theoretical and Computational Biophysics Group at U

-Images of remote HPC machine (Mercury) and mass storage system (UniTree) courtesy of the National Center for Supercomputing Applications (NCSA) and the Board of Trustees of the

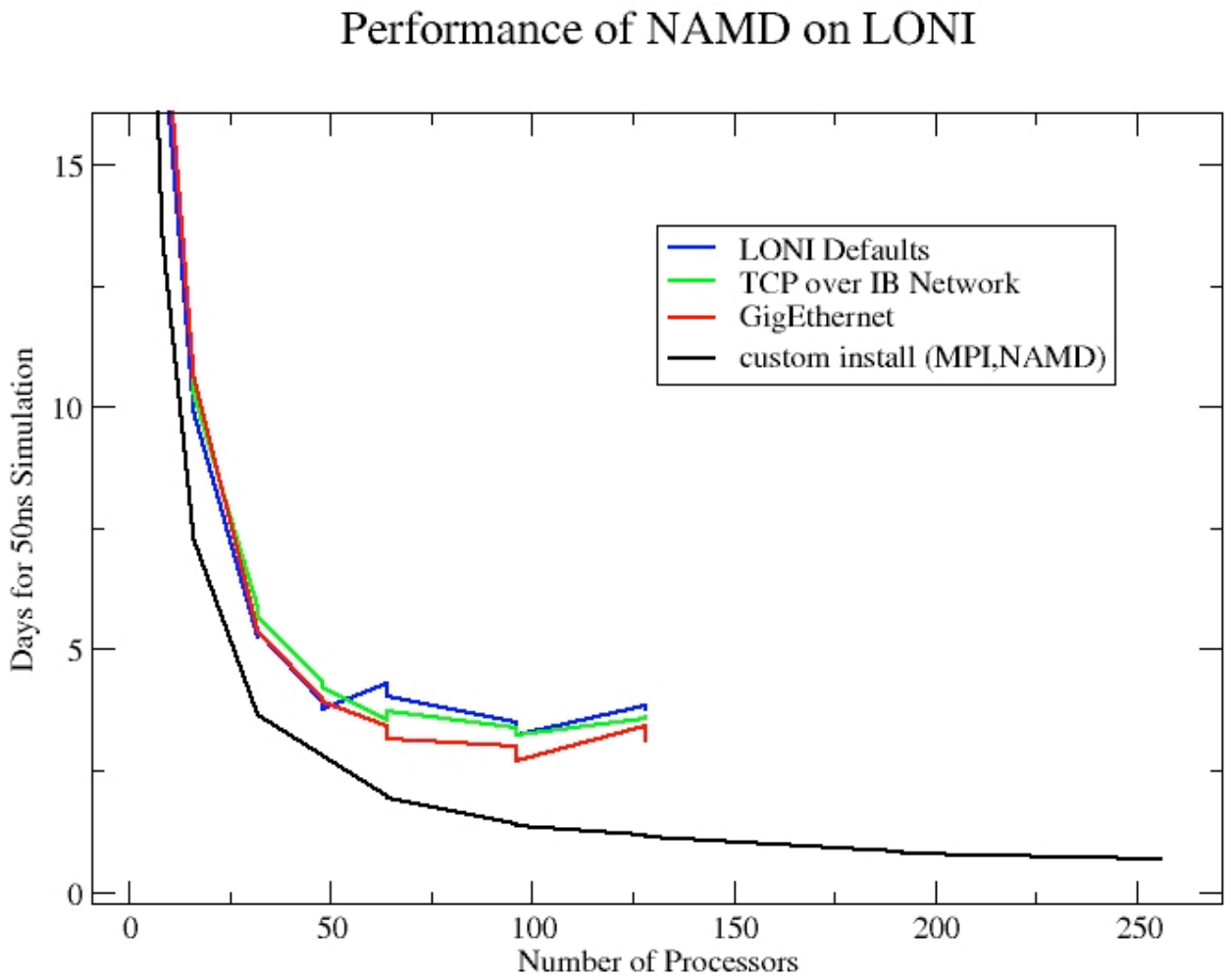
LONI Data Challenges

- Data Location
- Software Availability/Performance
- Authentication & Computing Environments

Data Location

- DNA, CCS , LONI
- Simulation Initiation
  - build on fly vs. build locally then transfer
    - 1 simulation build locally and transfer to compute server
    - 1000 simulations build on fly at compute server (30Gb)
- Queuing
  - data discovery (queue load, inputs, executables)
  - once queued committed to single resource?
- Analysis
  - where did I put that simulation?

Software Availability/Performance



Benchmarks based on one of the sample simulations distributed with NAMD demonstra
that careful optimization for LONI's Dell based infiniband machines provides significant
improvement over default download. This required compiling MPI libraries as well as N

Authentication & Environments

- Each LONI machine is independent
  - common
    - uid/gid/password
  - different
    - /home
    - /work
    - .login
    - .cshrc
    - software load
- DNA, CCS, LONI also differ

Meta-queuing solves my problems

- submit from desktop run anywhere
  - anywhere includes: DNA, CCS, LONI, TERAGRID
  - based on availability and load ... automatically!

- simulation protocols (workflows)
  - build system
  - run simulation
  - analyze

Questions

- Data visualization ?
  - how do I do this remotely?
  - how do I let others see my data?
- Data mining ?
  - comparing data from different simulations
- Re-analyzing ?
  - analysis protocols

THE END