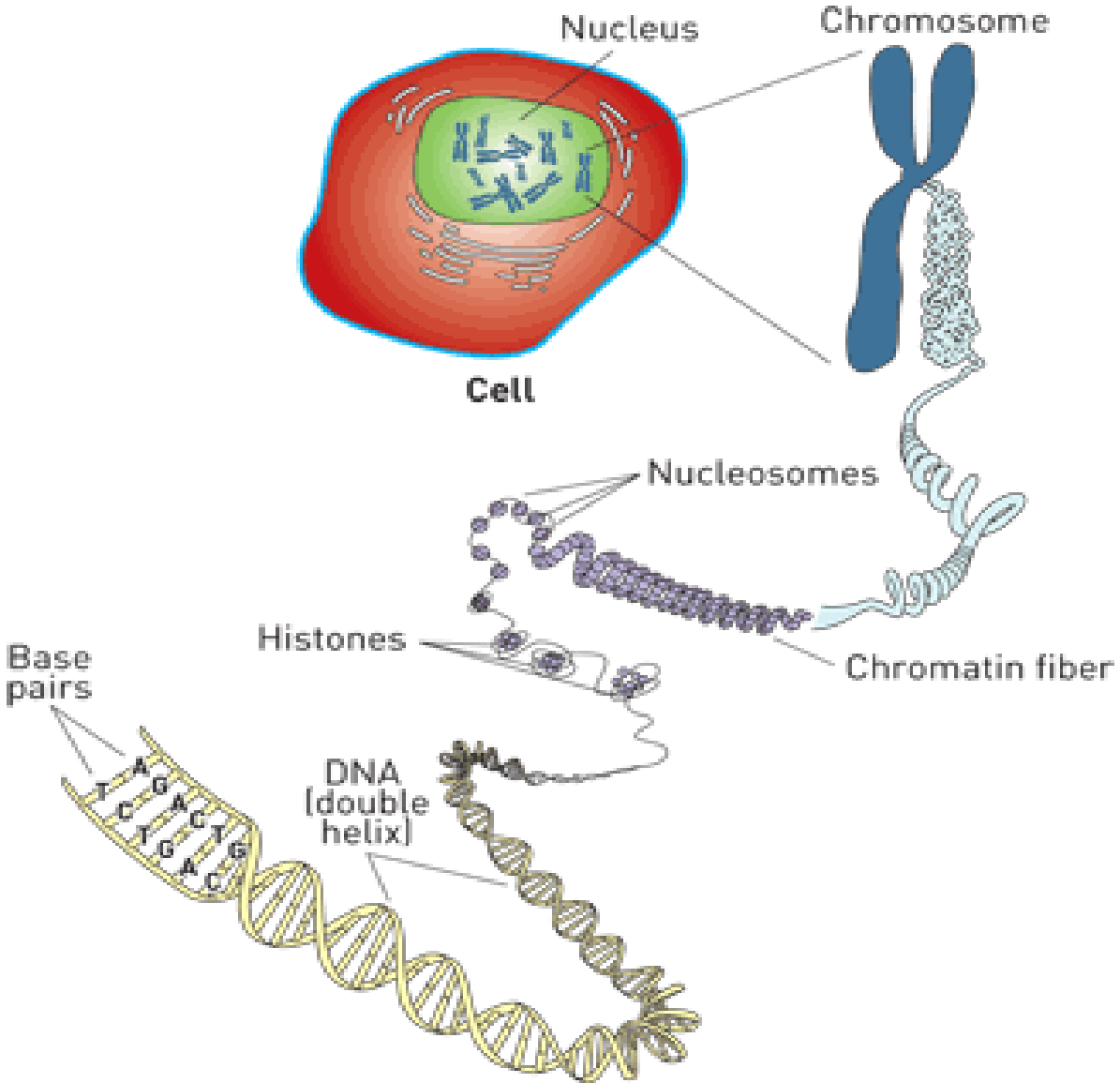
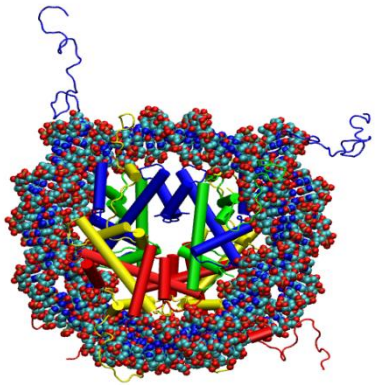
Running Many MD Simulations on Many Super Computers

Rajib Mukherjee[1], Shantenu Jha[2], Abhinav Thota[2], Hideki Fujioka[1], Thomas C. Bishop[1]

[1]Center for Computational Science, Tulane University
[2]Center for Computation and Technology, Louisiana State University

Big Biological Picture





Detailed Biological Picture/Motivation

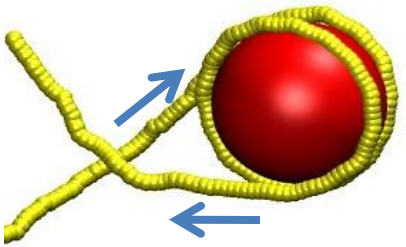
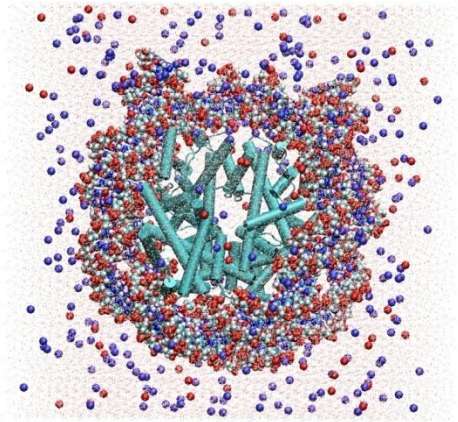
- *In vitro* histones occupy preferred locations on lengths of DNA greater than 147bp (i.e. positioning) and exhibit preferential binding in mixtures containing different 147bp long oligomers (i.e. affinity).

- *In vivo* nucleosome positioning is also observed. But the physical basis for stability and relationships between positioning and affinity remain unclear.

- We have simulated over 300 nucleosomes using all atom molecular dynamics (MD) to investigate the nucleosome structure and dynamics as a function of DNA sequence.

Our Plan: Need Many Simulations

- The 336 nuclesomes modeled represent 16 different segments of DNA, one from each chromosome of *Saccharomyces Cerevisiae*. Each segment contains the highest occupied and least variable nucleosome positioning sequence for the parent chromosome.

16 x   = 336 simulations

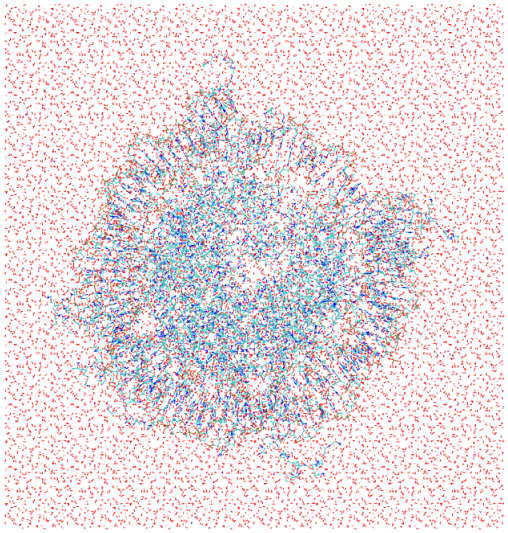
threading 167bp => 21 nucleosomes

•Each segment is 167bp long and includes: the observed 147bp positioning sequence and 10bp on each side. Every 147bp subsequence of each segment is threaded onto the octamer core as observed in xray structure pdbid 1kx5 and solvated, yielding 21 mononucleosomes for each of the 16 chromosomes.

Problem/Need for Many Supercomputers

System Size

13,046 histone protein atoms
152 bp DNA = 9,600 atoms (differs with sequence)
Water = 135,360 atoms, ions = 426 atoms
total: 158,432 atoms



13.7nm x 14.5nm x 10.1nm

Simulation 20ns

52mb Input -> 1ns runs -> 3.7Gb/ns

Total

336 Simulations Threads with 20 Sims/Thread =
6,720 Tasks of 1 ns trajectory: 6.7μs and 25Tb data

Resource Available

- The run is taken on various supercomputing resources, **TeraGrid and LONI**.

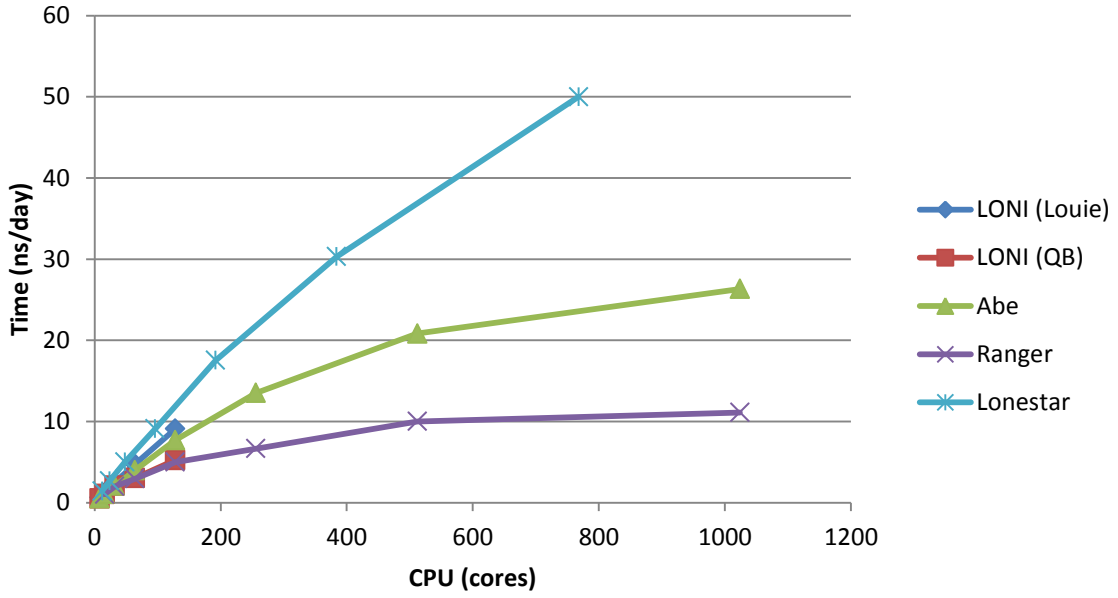
• TeraGrid is digital resource for science and engineering. It has distributed supercomputers all over USA(has 11 partners)

• Louisiana Optical Network Initiative (LONI), is a state-of-the-art, fiber optics network that runs throughout Louisiana 85 teraflops of computational capacity

- The output is stored after simulation and the input is retrieved before beginning of next ns simulation from **PetaShare** Data storage resource – can access data from LONI and TeraGrid machines with command line interface

NAMD 2.7 Performance in Available Resources

We are using NAMD 2.7 for simulation



No. Core	LONI (Louie)	LONI (QB)	Abe	Lonestar	Ranger	Kraken
Max Scratch	100GB	100GB	No Guarantee	250GB	350GB	
Total core	512	5,440	9,600	22,656	62,976	99,072

The Problem

- **Ideally** on Abe: 32 CPU = 2.08 ns/day = 9.6 days for 20 ns

- **Reality**: TOM'S BENCHMARKS ON Abe… simple serial resubmit

  192 cpu (9,600available) benchmark 10ns/day

  throughput is 18ns = 13days => 1.38ns/day

- Extrapolate for our 336 simulations…

- Even in ideal case for 6.7 μs nucleosomal dynamics if in 2 months need to do 112 task/day with 3,584 cores

Scaling Up and Out

Scale Up

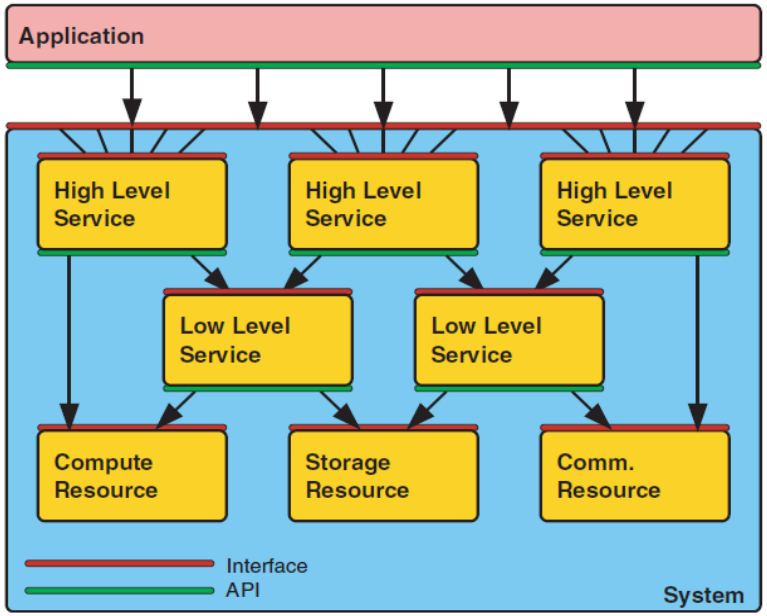
Bigger and Bigger
Heterogeneous Systems
Biology
Materials
Weather

Scale Out

More and More
Parameter Sweeps
Ensembles

1. Shaw, D.E., *et al.* (2009), *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. ACM, Portland, Oregon, pp. 1-11.
2. Schulten, K. NCSA PRAC Award (2010) Publicly accessible Petascale Computing Resource Allocations Award Information.

Getting Across



Simulation

Services

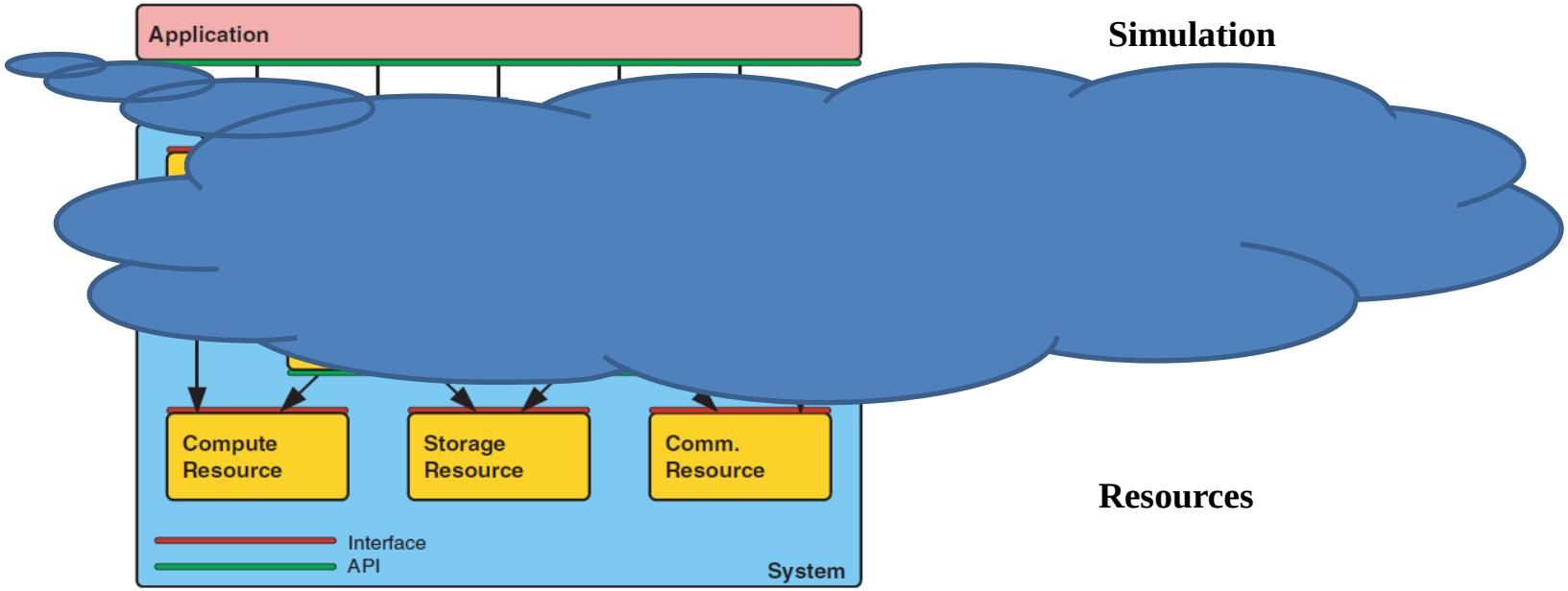
High Level:
- meta-scheduler
- federated fs
- MPI

Low Level:
- batch scheduler
- fs (AFS, GFS)
- TCP

Resources

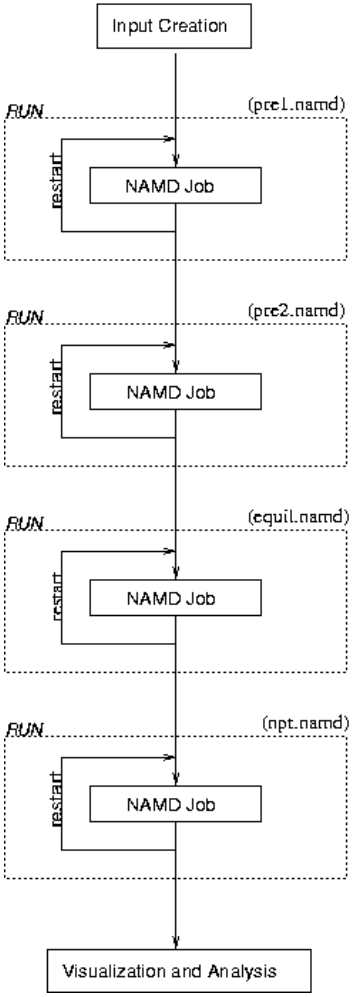
Jha, et al. Concurrency Computat.: Pract. Exper. 2009; 21: 1087-1108

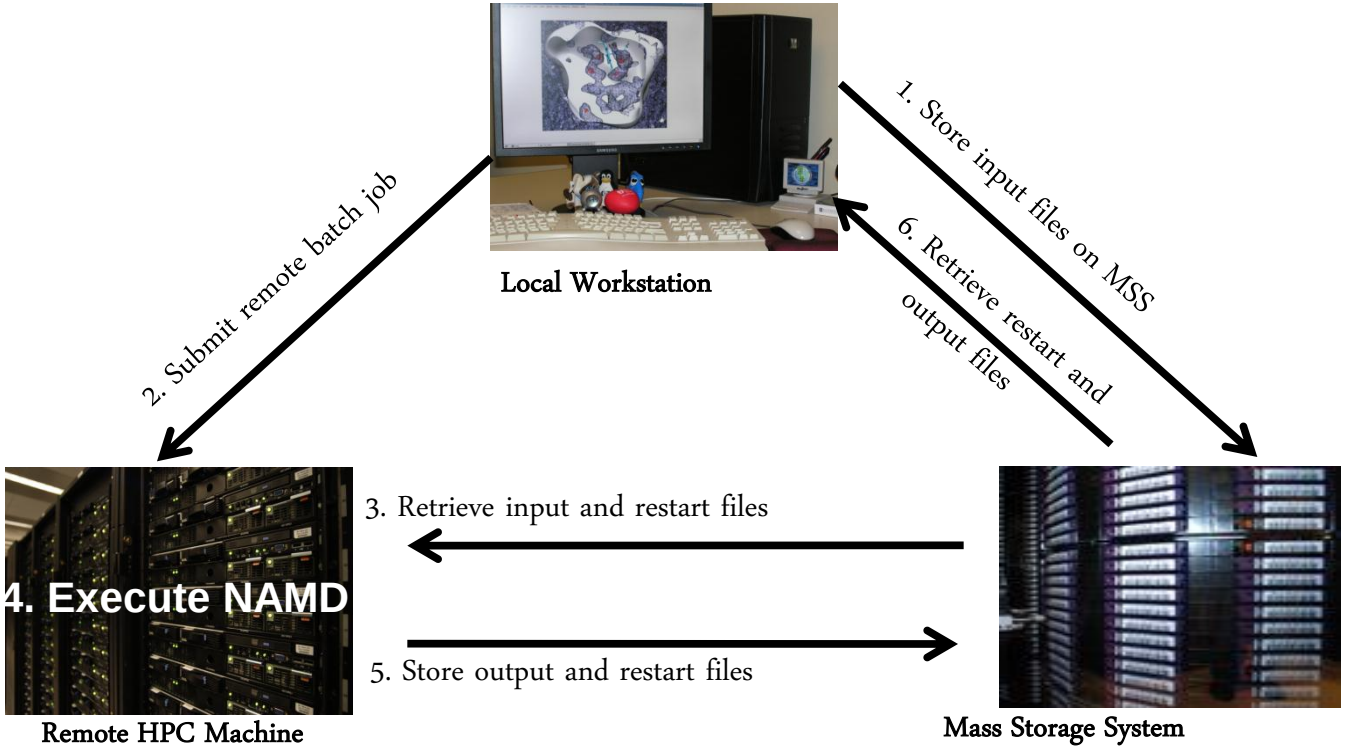
Getting Across (Desktop Computing)



Jha, et al. Concurrency Computat.: Pract. Exper. 2009; 21: 1087-1108

Typical MD 6 Step Work Flow

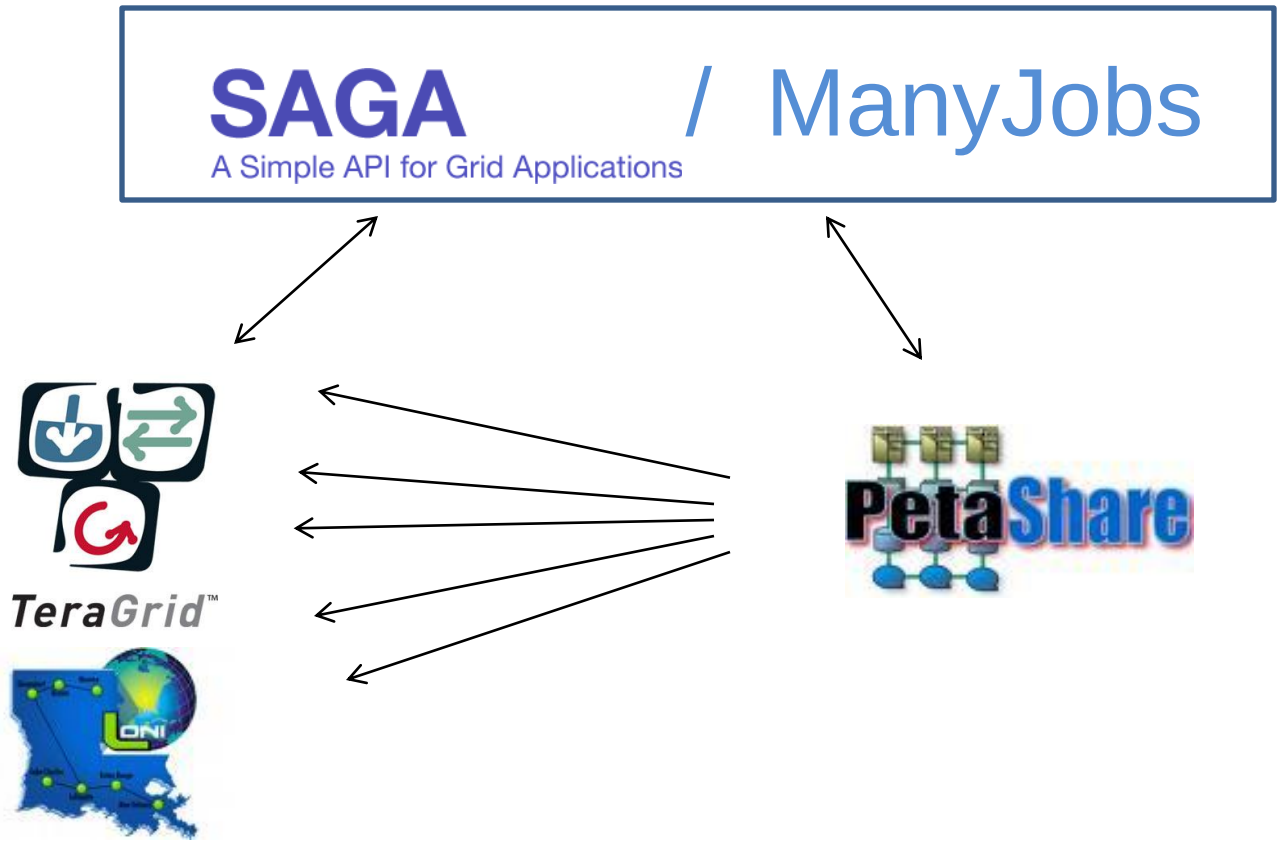




TCBG at UIUC: **NAMD-G**, In Proceedings of the 2006 TeraGrid Conference.

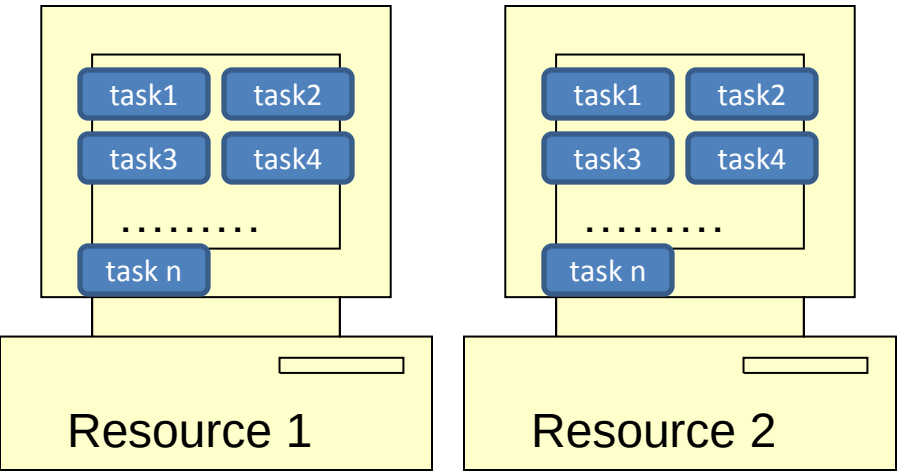
Manual Job Management

Automatic Job Management



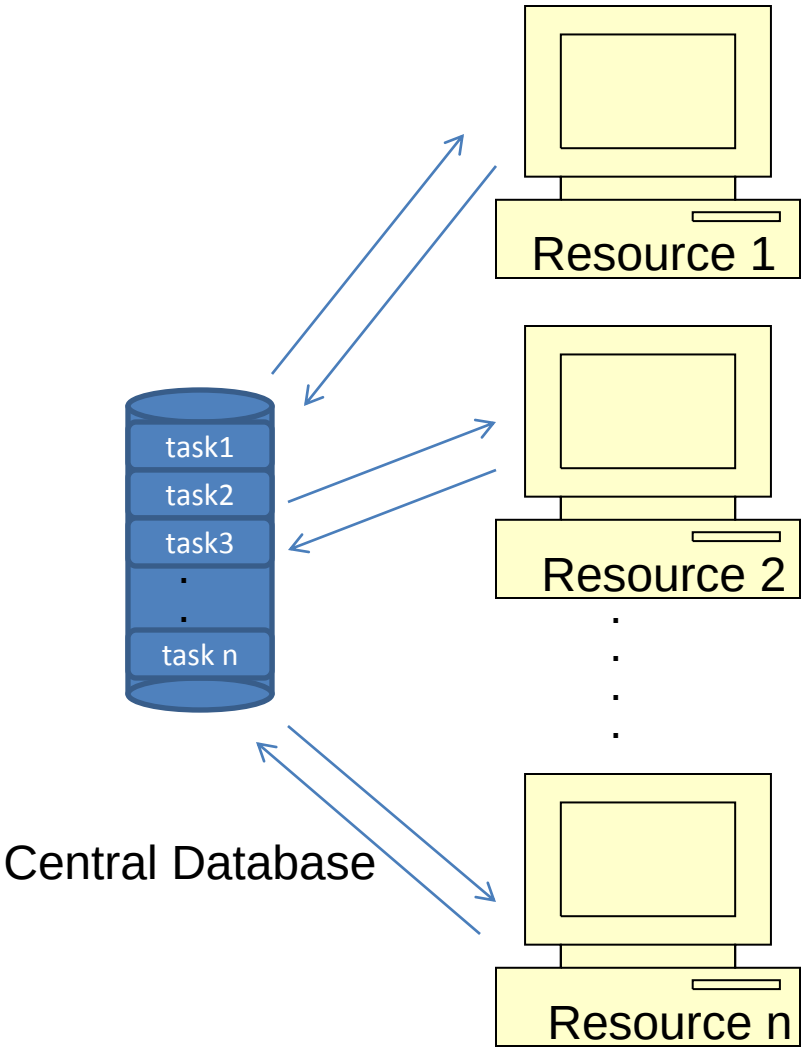
- The run is taken via **BigJob and ManyJobs** on various supercomputing resources.

BigJob vs ManyJobs



……. Resource n

BigJob



ManyJobs

BigJobs Details

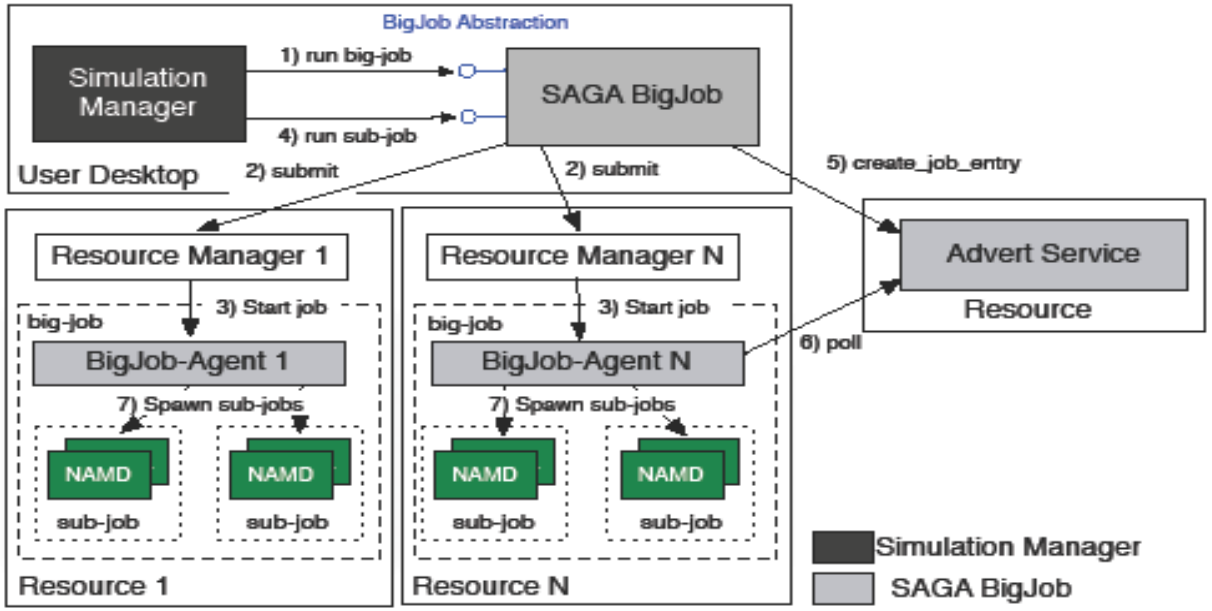
- BigJobs is a Simple API for Grid Applications (SAGA) based pilot job implementation designed to run application across multiple, heterogeneous, distributed grid and cloud resources.

- Individual tasks are dynamically assigned to a larger container of tasks, the pilot job.

- It is successfully implemented for Adaptive Distributed Replica-Exchange Simulations.

- Here we are using it for High Throughput Simulations

SAGA BigJob Manager



- SAGA BigJob Manager access the data using advert services
- Manages and allocates the resources to the sub-jobs.
- It has access to the current state of the sub-jobs and resources
- Uploads the data upon successful completion

ManyJobs Details

•Submits Many Jobs to remote clusters from master.
Requests resources but does not actually assign job.
Job Start: Pull
•Compute node contacts master-host for parameters, environment variables, etc...
Job End: Push
computing node contacts master-host; master-host queues a new job.

•Allows Dependencies:
if job A depends on job B, job A won't run until job B ends.
Communication/Authentication:
'ssh' command w/out password.

Our Usage Results

Direct:

Abe, QueenBee, other LONI machine:
84 systems for 20ns each queued as1,680 individual jobs; each job use 64 cores
Note: Max jobs number of simulations jobs run at one time 63 (QB much less)

BigJobs:

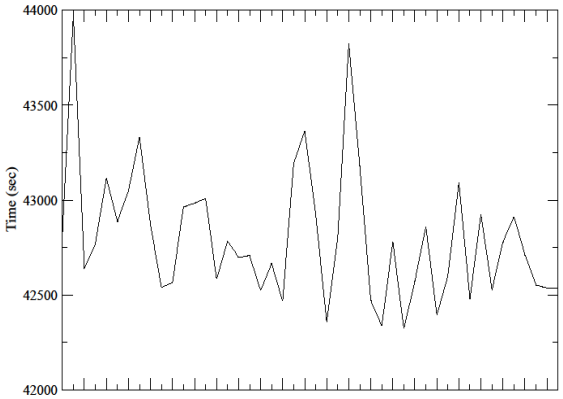
Ranger, QueenBee, other LONI machine:

	BigJob Size (cores)	Runtime Requested (mins)	Typical Wait time (mins)
Ranger	63 X 32 = 2,016	780	323
QueenBee	16 X 32 = 512	720	35
Other LONI (Eric)	4 X 32 = 128	720	1,367

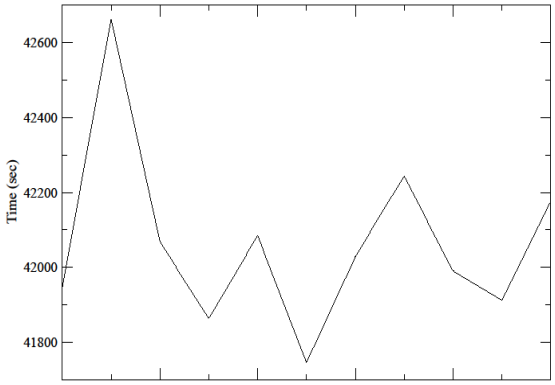
ManyJobs:

Abe, QueenBee, other LONI machine:
336 systems 1ns each for 4 ns; each job 128 cores in Abe, 32 cores in Queenbee and other LONI

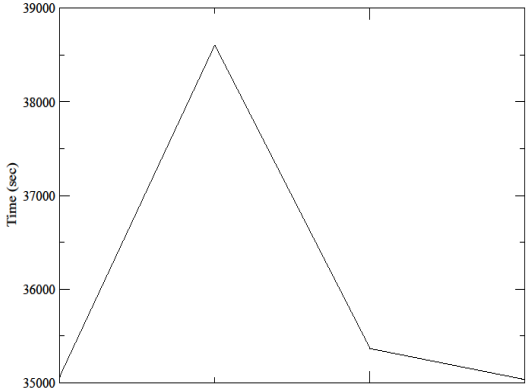
Simulation Time



(a)



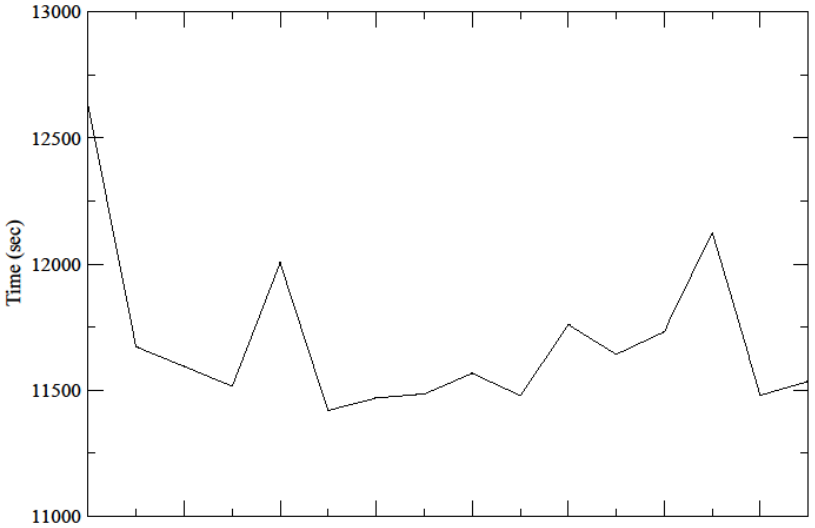
(b)



(c)

- Simulation time of different sub-jobs of BigJob from (a) Ranger (b) Queenbee and (c) Eric, all using 32 cores per job.
- The simulation time varies randomly.

Simulation Time



- Simulation time from Abe with 128 core submitted using ManyJobs.

- Simulation time varies randomly.

Direct Run Problems

- Do not get the benchmark
- Batch schedulers differ by machine (can overcome by BigJobs)
- Queue Limits
- ssh and authentication (can overcome by BigJobs)

Pros and Cons: *Our Experience*

Many Jobs (Hideki's Version)

Pros: lightweight

Pull model allows jobs to run anywhere

Cons:

cannot stage data since don't know where will run

Authentication (ssh) & Scheduling (qsub, bsub)

Big Jobs

Pros: full featured

bundles jobs, shorter time to completion

stages data before run time, no lost run time

Authentication Problems solved

Failure recovery/Fault Tolerance

Cons:

portability (configuration & installation)

impression that it's overkill for today's problem

Project Status: > 2.7μs of nucleosome dynamics

Chromosome	Time (ns) 0 5 10 15 20
I	
I I	
III	
IV	
V	
VI	
VII	
VIII	

Chromosome	Time (ns) 0 5 10 15 20
IX	
X	
XI	
XII	
XIII	
XIV	
XV	
XVI	

Conclusions/TODO

We've started doing production run in SAGA

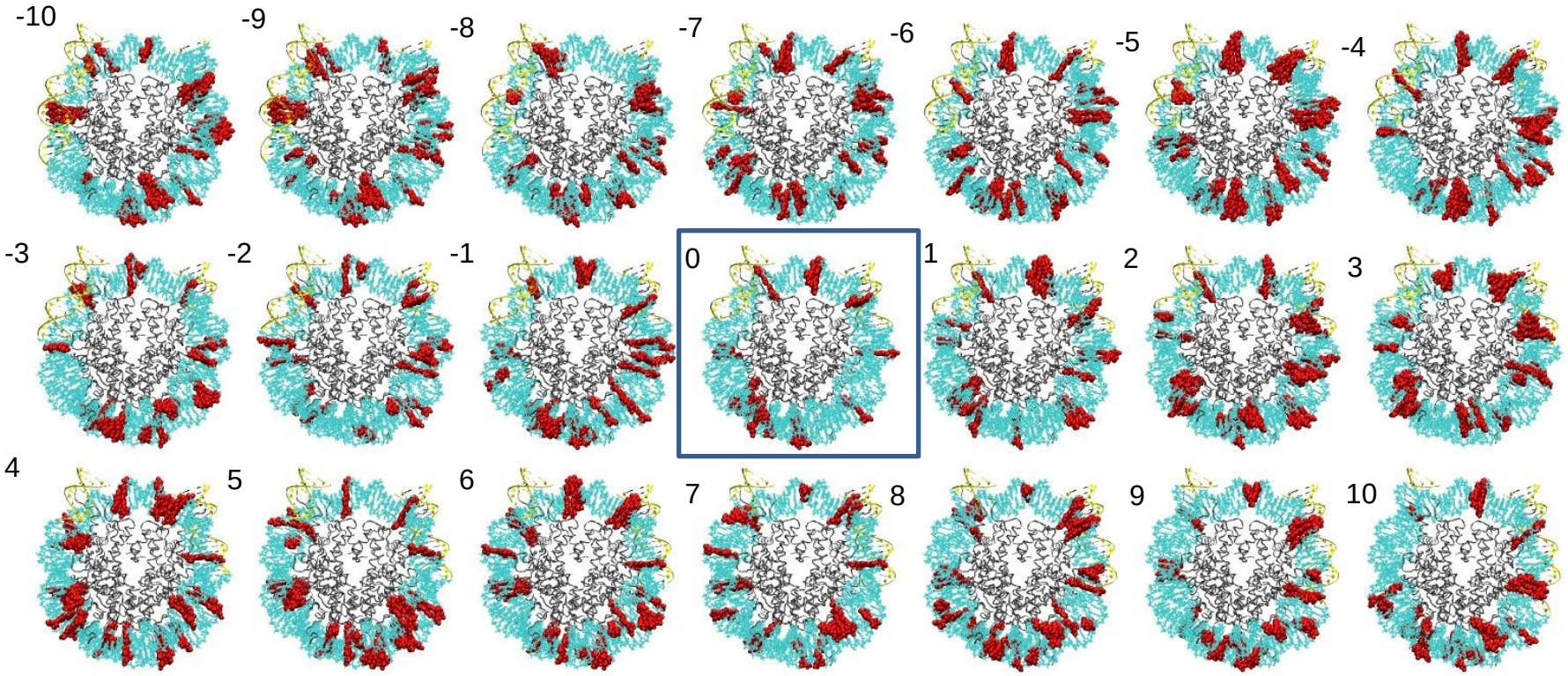
SAGA is needed to Scale up/out/across:

Queue limits, authentication, file access…. Cannot writing all of this in lightweight ManyJobs

Fault Tolerance:
Not yet encountered

Data Management
On-the-fly analysis
Non-Petashare resources

Conclusion



Kinks from **Richmond-Davey Criteria** are highlighted with red

Online Stuff

Hideki's Many Jobs
http://beagle.ccs.tulane.edu/svn/repos/ManyJobs

Saga
http://saga.cct.lsu.edu/

BigJobs
http://saga.cct.lsu.edu/projects/abstractions/bigjob-a-saga-based-pilot-job-implementation

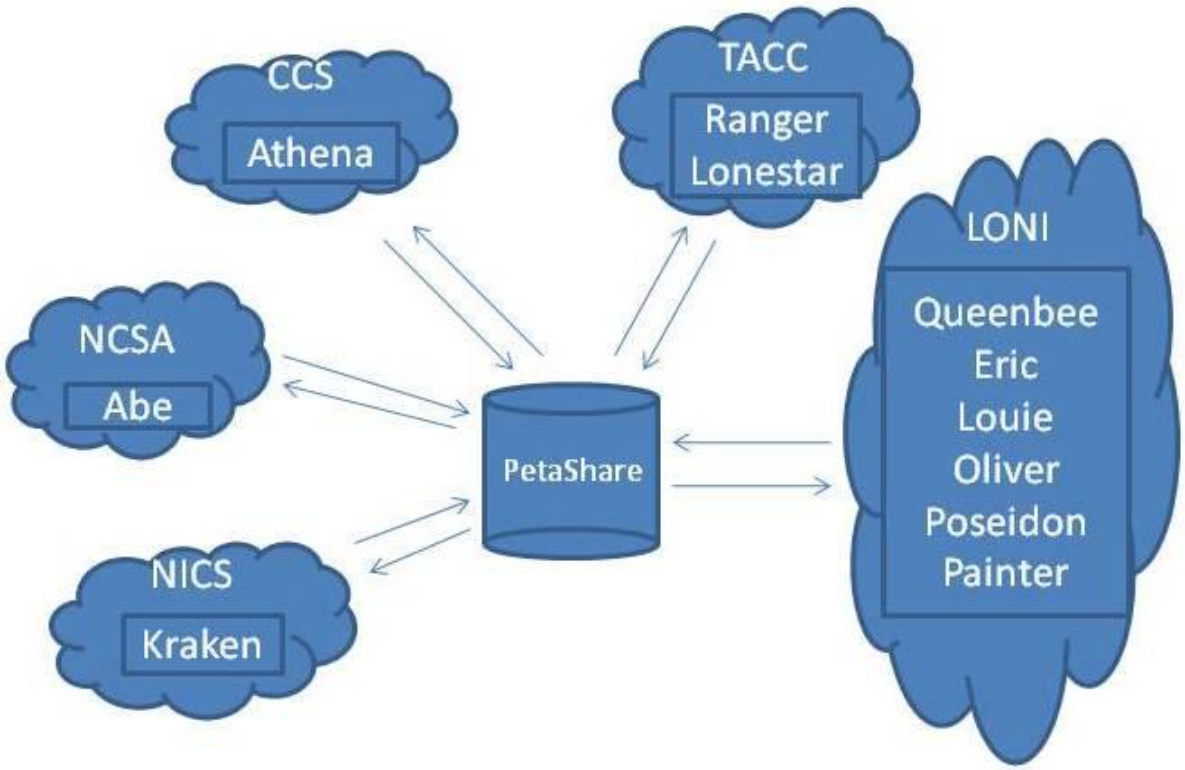
Acknowledgements

CCS: Dr. Hideki Fujioka
CCT: Abhinav Thota & Dr. Shantenu Jha
LONI: Resources
TeraGRID: Allocation

NSF: RII and LaSIGMA
NIH: “Molecular Dynamics Studies of Nucleosome Stability.
“

PetaShare

- PetaShare is a distributed data archival, analysis and visualization cyberinfrastructure for data-intensive collaborative research that enables us to rapidly access petabyte data resources from virtually any remote computing resource via a common command line interface.

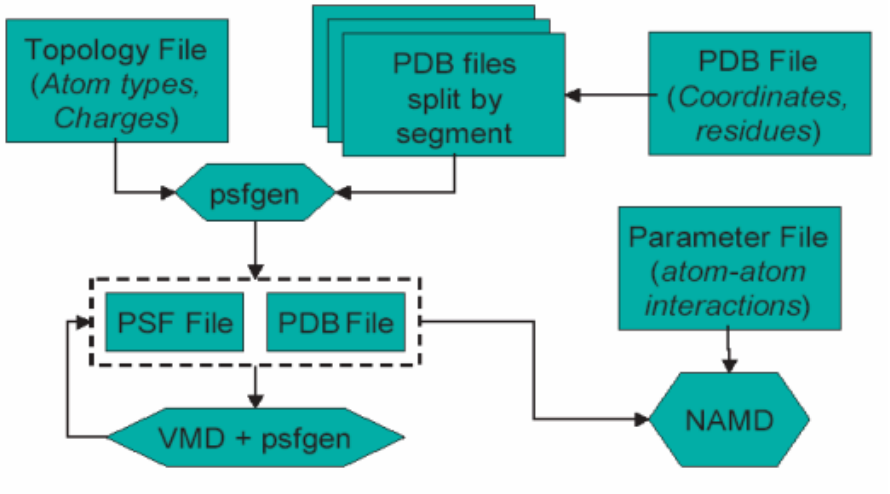


Available MD Codes

- NAMD: http://www.ks.uiuc.edu/Research/namd/
- AMBER: http://ambermd.org/
- CHARMM: http://www.charmm.org/
- GROMACS: http://www.gromacs.org/
- LAMMP: http://lammps.sandia.gov/

(Extra)MD Practical Issues

- Pre-Simulation (1day)

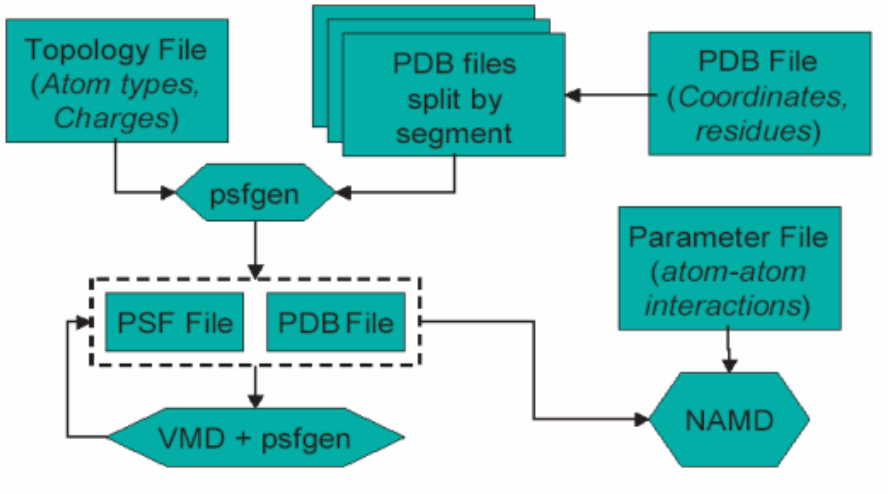


- Simulation (2 mo)
  - Minimization
  - Equilibration
  - Dynamics

- Post Simulation Analysis (1wk)

(Extra)MD Practical Issues

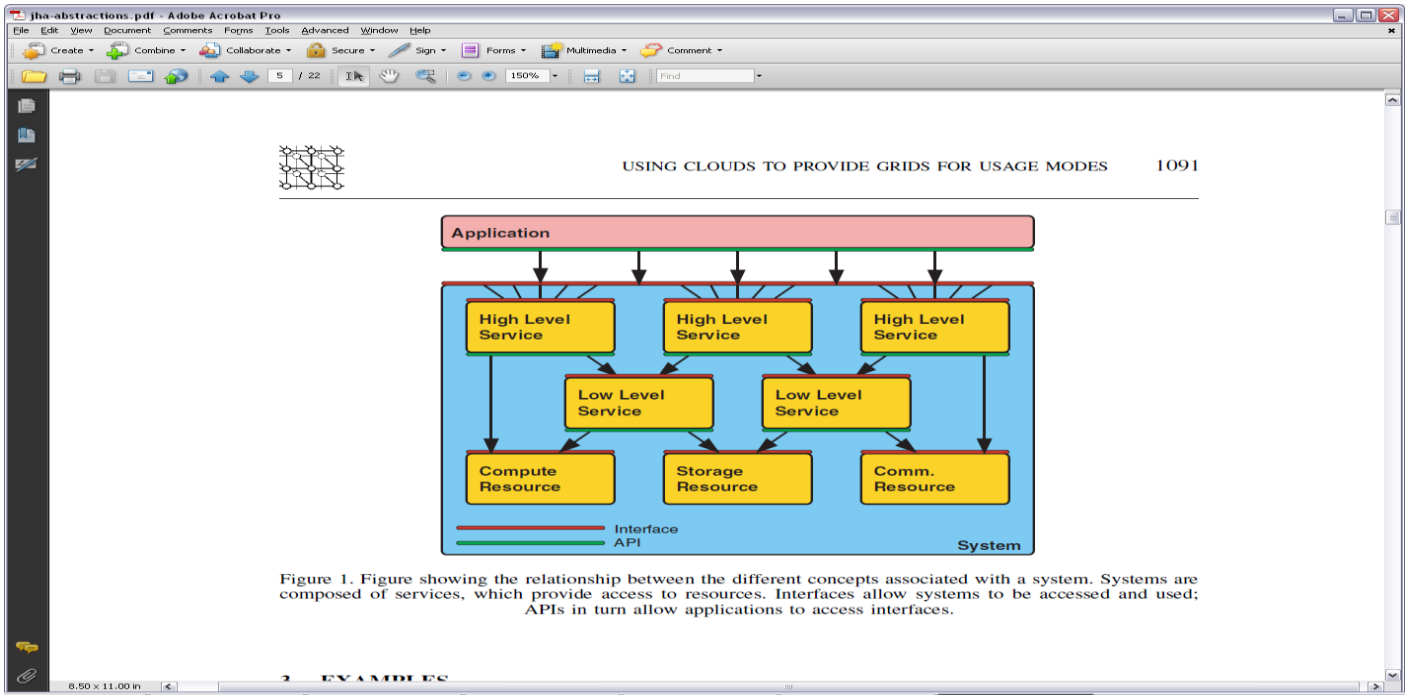
- Pre-Simulation (1day)



- Simulation (2mo)
  - Minimization
  - Equilibration
  - Dynamics

- Post Simulation Analysis (1wk)

SAGA: Simple API for Grid Applications



SAGA is an API that provides the basic functionality required to **build distributed applications**, tools and frameworks so as to be **independent of the details of the underlying infrastructure**. SAGA can be used to provide simple access layers for distributed systems and abstractions for applications and thereby address the fundamental application design objectives **of Interoperability across different infrastructure**, **Distributed Scale-Out**, **Extensibility**, **Adaptivity** whilst preserving simplicity.

TeraGrid XD

•TeraGrid **Extreme Digital Resources for Science and Engineering** (XD)
Phase III of NSF's TeraGrid high-end digital services, providing US researchers and educators with the capability to work with extremely large amounts of digitally represented information.

•Primary Goal of TeraGrid XD is to enable major advances in science and engineering research, in the integration of research and education, and in broadening participation in science and engineering by under-represented groups, by providing researchers and educators with usable **access to extreme-scale digital resources beyond those typically available on a typical campus**, together with the interfaces, consulting support, and training necessary to facilitate their use.

LONI

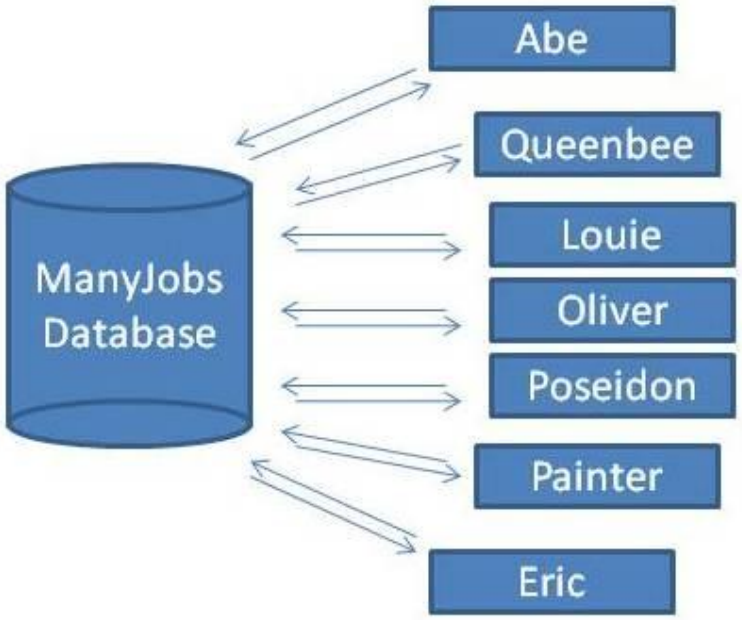
•**Louisiana Optical Network Initiative (LONI), is a state-of-the-art, fiber optics network that runs throughout Louisiana**, and connects Louisiana and Mississippi research universities to one another as well as National LambdaRail and Internet2.

•LONI connects Louisiana's major research universities– allowing greater collaboration on research that produces results faster and with greater accuracy.

•LONI provides Louisiana researchers with one of the most advanced optical networks in the country and the most powerful distributed supercomputer resources available to any academic community with over **85 teraflops of computational capacity**.

ManyJobs

- ManyJobs maintains a database of all compute tasks and the dependencies between tasks.
- ManyJobs submits job requests to all computing resources listed by the user.
- Once a job starts, ManyJobs assigns a specific task to the job and request new job. This process repeats until all tasks in the database are completed.
- The tool is written in Python and utilizes secure shell (ssh) communications.



NAMD 2.7 Performance in Available Resources

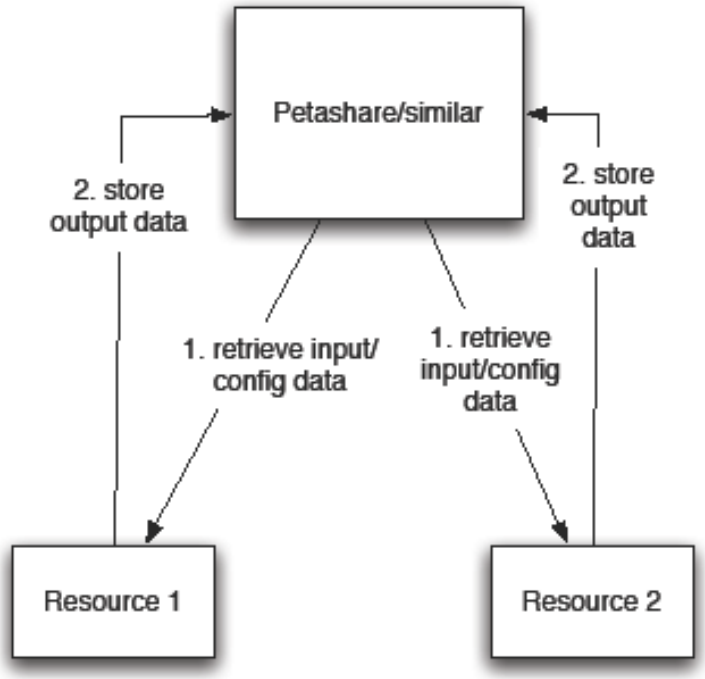
No. Core	atoms/core	LONI (Louie)	LONI (QB)	Abe	Lonestar	Ranger	Kraken
8	19,750	0.625	0.526	0.526			
16	9,875	1.27	1.08	1.06	1.43(12)	1.04	
32	4,937	2.5	2.08	2.08	2.7 (24)	2.0	
64	2,468	4.76	3.03	4.0	5.0 (48)	2.94	
128	1,234	9.09	5.26	7.69	9.09(96)	5.0	
256	617			13.51	17.54(192)	6.66	
512	309			20.83	30.3(384)	10.0	
1,024	154			26.32	50.0(768)	11.11	
2,056							
Max Scratch		100GB	100GB	No Guarantee	250GB	350GB	
Total core		512	5,440	9,600	22,656	62,976	99,072

Time in ns/day

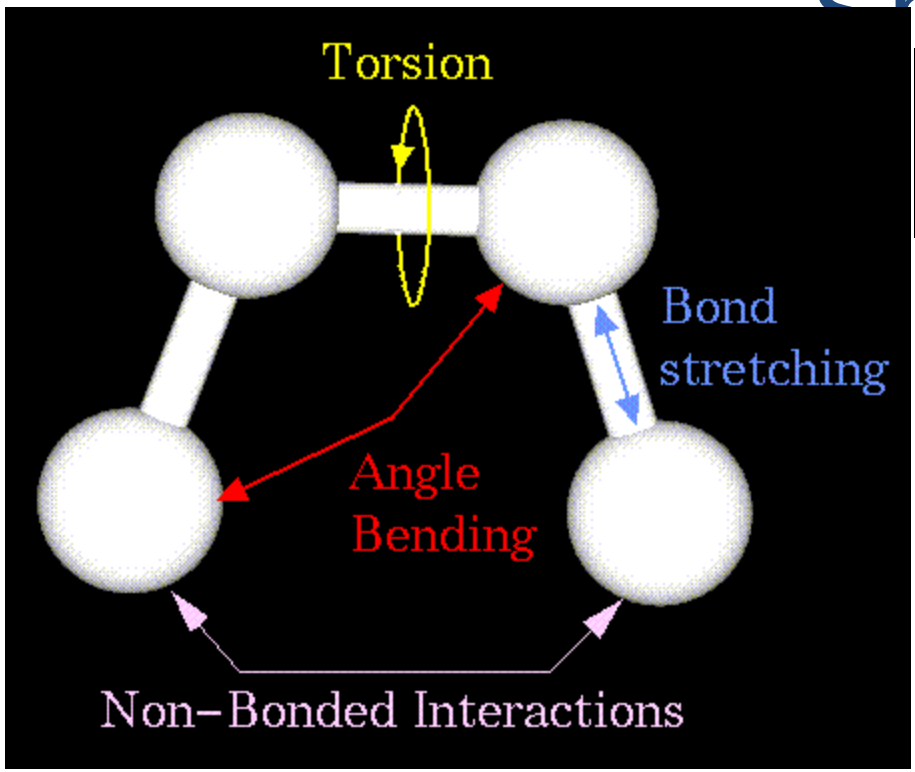
Details Our Plan

- The 21 nucleosomes allow us to assess positioning over a full turn of the DNA helix on each side of the experimentally determined positioning sequence.
- An explicit TIP3 solvent shell and sufficient NaCl to both neutralize the system charge and provide a bulk ion concentration of 150mM are added to each nucleosome model. Each fully solvated system contains **approximately 160,000 atoms.**
- Each nucleosome is subjected to MD simulations with NAMD2.6 and then NAMD2.7. We use the Cornell et al. force field (parm99) with Barcelona corrections for DNA. All systems are minimized and equilibrated to a target temperature of 300K.
- Production simulations used the NPT ensemble with NAMD's Berendsen pressure regulation and Langevin temperature regulation.
- Each trajectory is 20ns and the simulation is divided into 1ns each. This requires simulation of **16 X 21 X20 = 6,720 tasks of 1ns trajectory**.

Work Data Flow Diagram



Molecular Dynamics in a Nut Shell



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

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

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

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.html

From Energy to Dynamics

- Newton's Law

$$-\frac{dE}{dr_i} = F_i$$

- Numerical Integration

1. solve for a_i at t using: $-\frac{dE}{dr_i} = F_i = m_i\, a_i(t)$
2. update v_i at $t + \Delta t/2$ using: $v_i(t + \Delta t/2) = v_i(t - \Delta t/2) + a_i(t)\,\Delta t$
3. update r_i at $t + \Delta t$ using: $r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t/2)\,\Delta t$

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

MD is a Mature Method

- algorithms defined/robust
- efficiently parallelized (10,000 CPUs/ 3+M atoms)
- force fields carefully evaluated
- data formats decided
- tools for visualization well developed

Available MD Codes

- NAMD: http://www.ks.uiuc.edu/Research/namd/
- AMBER: http://ambermd.org/
- CHARMM: http://www.charmm.org/
- GROMACS: http://www.gromacs.org/
- LAMMP: http://lammps.sandia.gov/