LA-SiGMA

Louisiana Alliance for Simulation-Guided Materials Applications

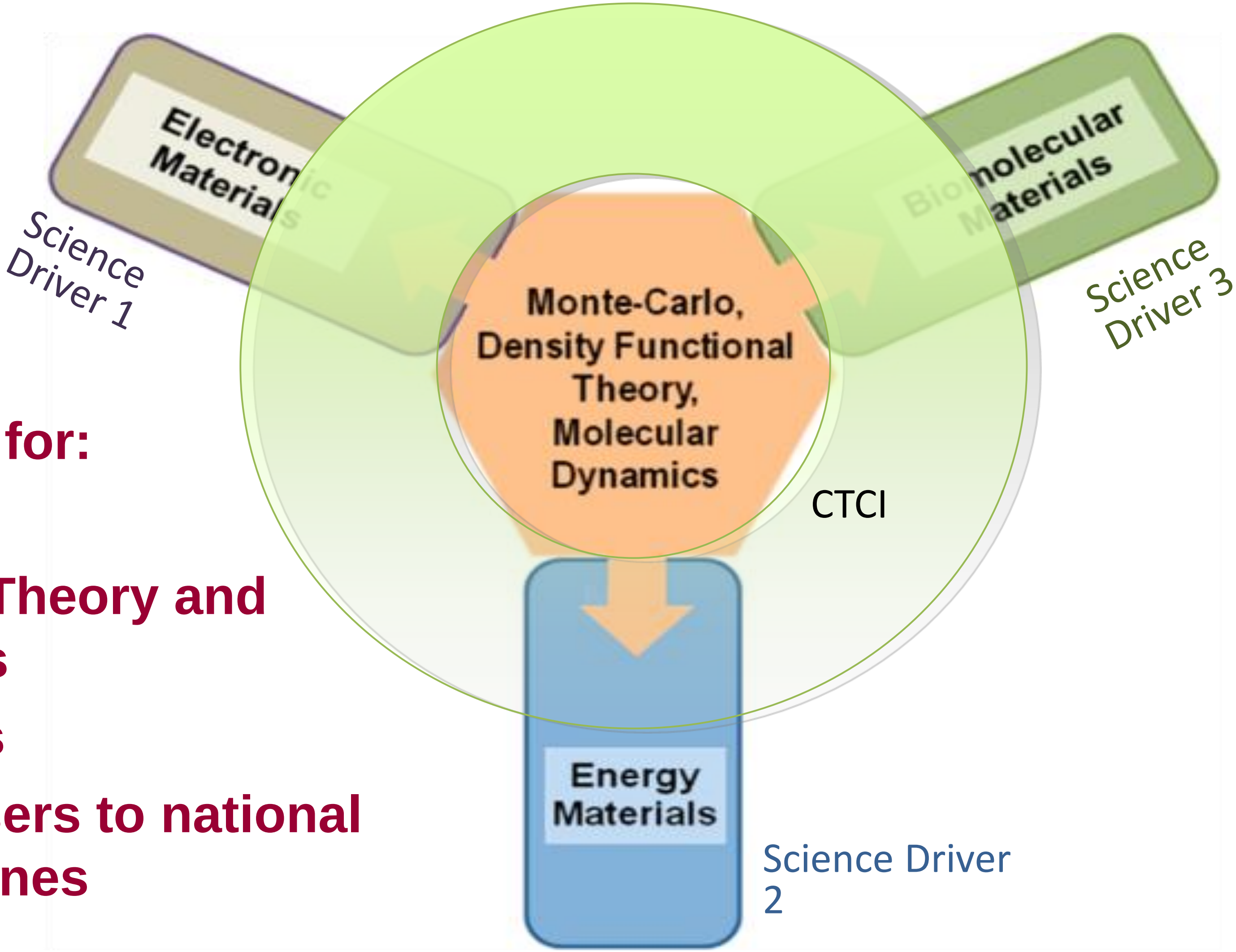
Computational Teams Cybertools / Cyberinfrastructure (CTCI)

“The glue”

LA-SiGMA Symposium, Baton Rouge: July 23, 2012

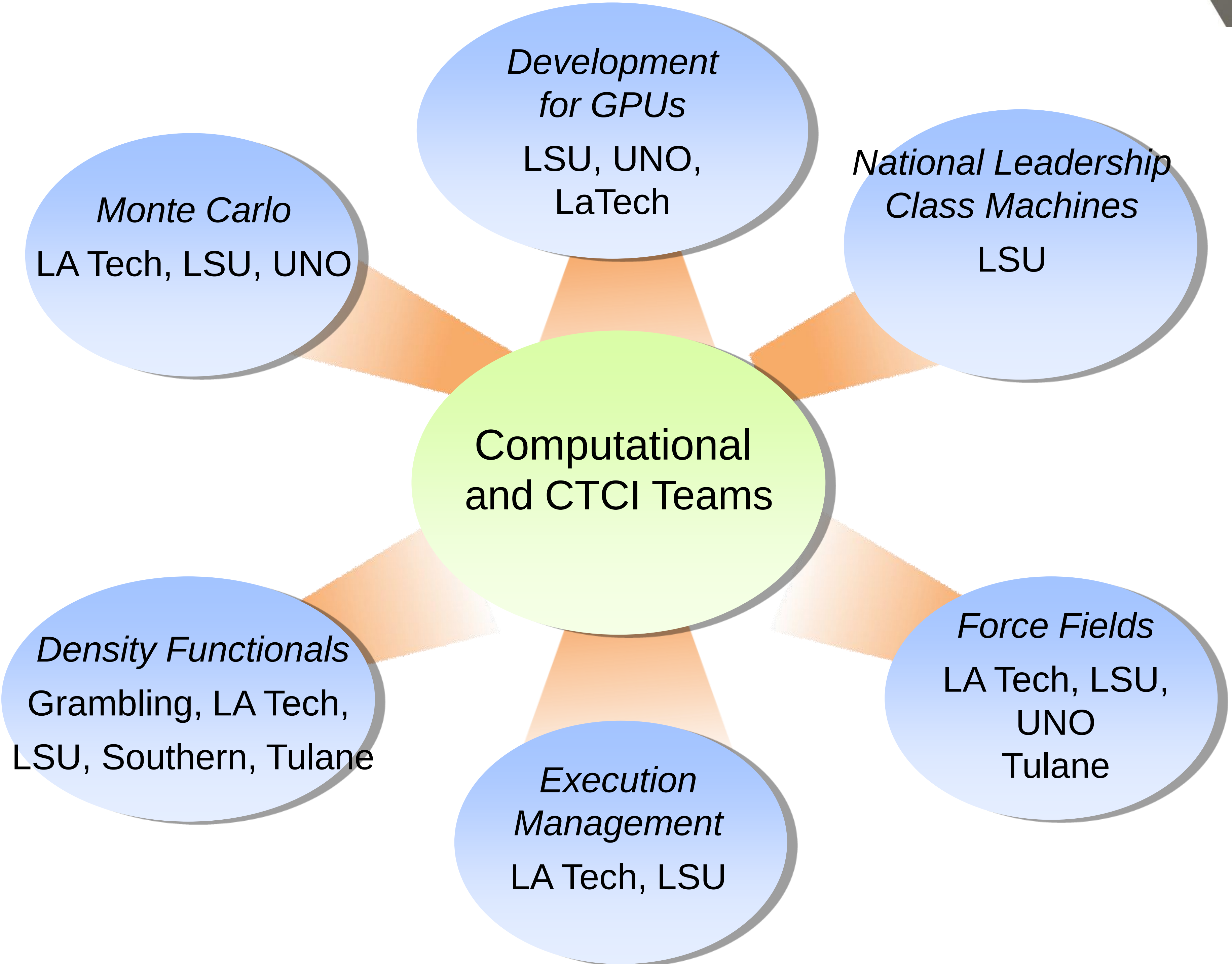
The Goal

Computational and CTCI Teams: Build transformational common toolkits



- **Build common toolkits for:**
  - **Monte Carlo**
  - **Density Functional Theory and Force Field Methods**
  - **Molecular Dynamics**
- **Graduate LA-SiGMA users to national leadership class machines**

Computational/CTCI Research Themes



Computational and CTCI Teams

Development for GPUs

LSU, UNO, LaTech

National Leadership Class Machines

LSU

Force Fields

LA Tech, LSU, UNO Tulane

Execution Management

LA Tech, LSU

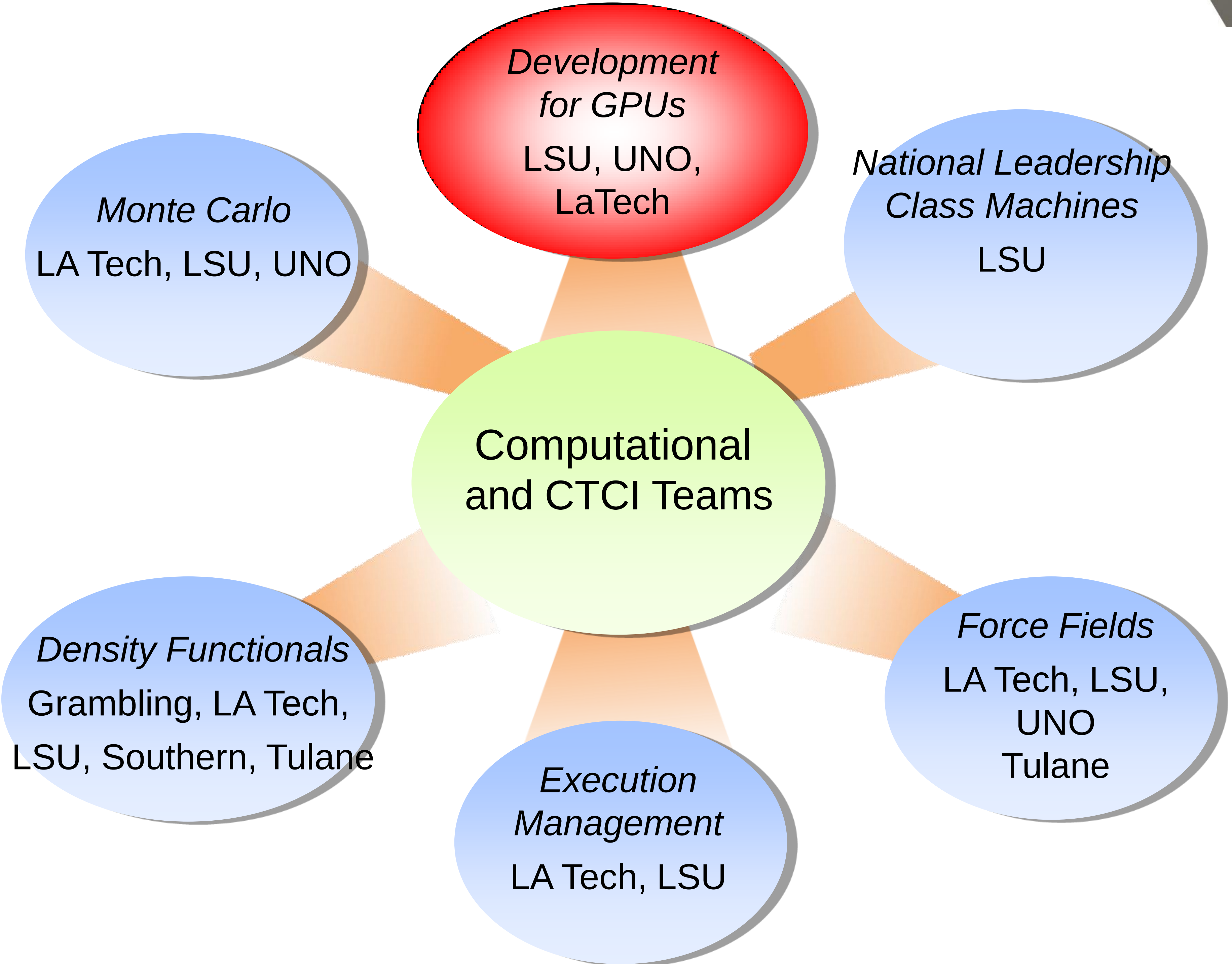
Density Functionals

Grambling, LA Tech, LSU, Southern, Tulane

Monte Carlo

LA Tech, LSU, UNO

Computational/CTCI Research Themes



Computational and CTCI Teams

Development for GPUs

LSU, UNO, LaTech

National Leadership Class Machines

LSU

Force Fields

LA Tech, LSU, UNO
Tulane

Execution Management

LA Tech, LSU

Density Functionals

Grambling, LA Tech, LSU, Southern, Tulane

Monte Carlo

LA Tech, LSU, UNO

GPU: Program Optimization

- Programming GPUs (for example, the NVIDIA GPUs using CUDA or OpenCL) is still tedious:
  - Performance of GPU highly sensitive to the formulation of the kernel; needs significant experimentation
  - Programmers may like this low level of control (suitable for library development; compilers and tools are not highly helpful here)
- **Strategy at LSU (Yun, Ramanujam):**
  - Understand the impact of and interactions among program optimizations for HF-QMC, PT and VMC
  - Develop and use effective transformation and optimization strategies
  - Code partitioning between CPU and GPU
- **Strategy at LaTech (Leungsuksun):**
  - Parallel Programming Tool Development based on Single Assignment C (SaC) toolset that enables parallel application developers expressing their problems in a high-level language

GPU Programming Team

- GPU Programming Team of roughly 25 faculty, students, and postdocs from LSU, LA Tech & Lousiana School for Math, Sciences, and the Arts (HS).
- Housed in the *Collaboratorium* at LSU and containing 12 GPU-enabled desktop computers.
- LSU “Condo” – LA-SiGMA to purchase GPU-enabled nodes on CCT/LSU’s Tezpur upgrade.
- Another GPU cluster (*Shelob*) from NSF CRI funding (significant LA-SiGMA role)

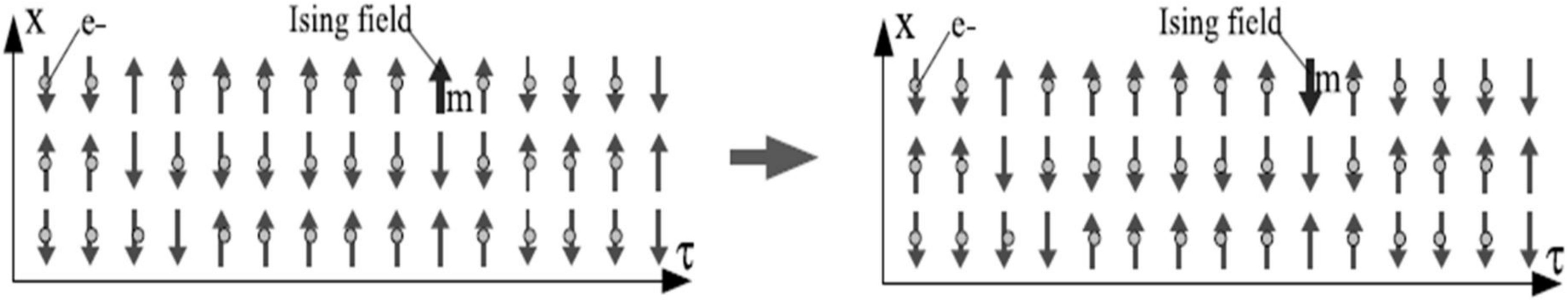
Nvidia GTX 460, http://www.nvidia.com/object/product-geforce-gtx-460-us.html

Parallel Tempering (See poster)

- The main goal here is to develop an efficient Parallel Tempering Monte Carlo code on GPU, with which we can study systems with complex energy landscapes.
- Developed a full-featured Ising model simulation program for CUDA GPUs.
- Studying phase transition of spin glass in a finite magnetization field.
- Results:
  - ➢The FPGA design (custom hardware) of Montovani et al. delivers the best time of 16 picoseconds per spin flip proposal (PS/spin).
  - ➢We achieve 39 picoseconds per spin flip proposal (PS/spin) on a single commodity GPU card, which is 3X better than other 2D GPU implementations.
  - ➢Our GPU version is about 600 times faster than our prototype CPU implementation.

Hirsch-Fye Quantum Monte Carlo-1 (See poster)

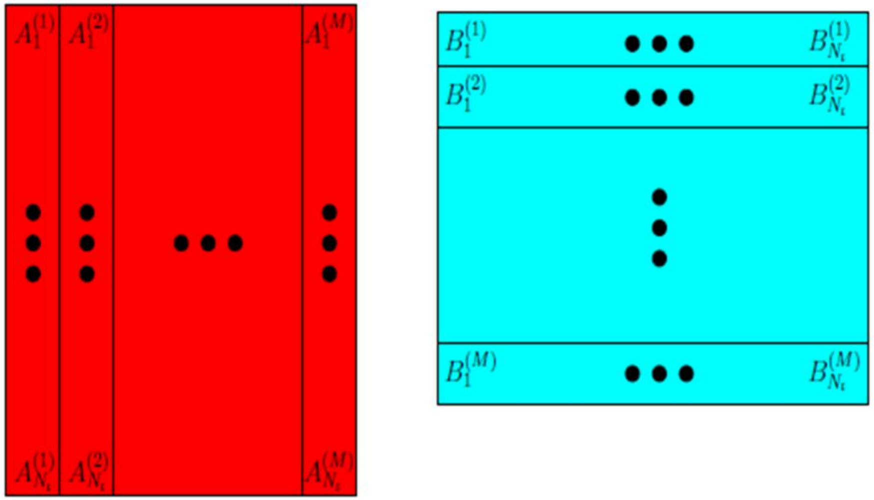
- This project simulates the interaction of conducting electrons in a metal.
- Using the Hirsch-Fye method mapped to a problem of electrons scattering of an Ising field in space and (imaginary) time.
- The configurations of the Ising field are sampled using Monte Carlo techniques.



Hirsch-Fye Quantum Monte Carlo-2 (See poster)

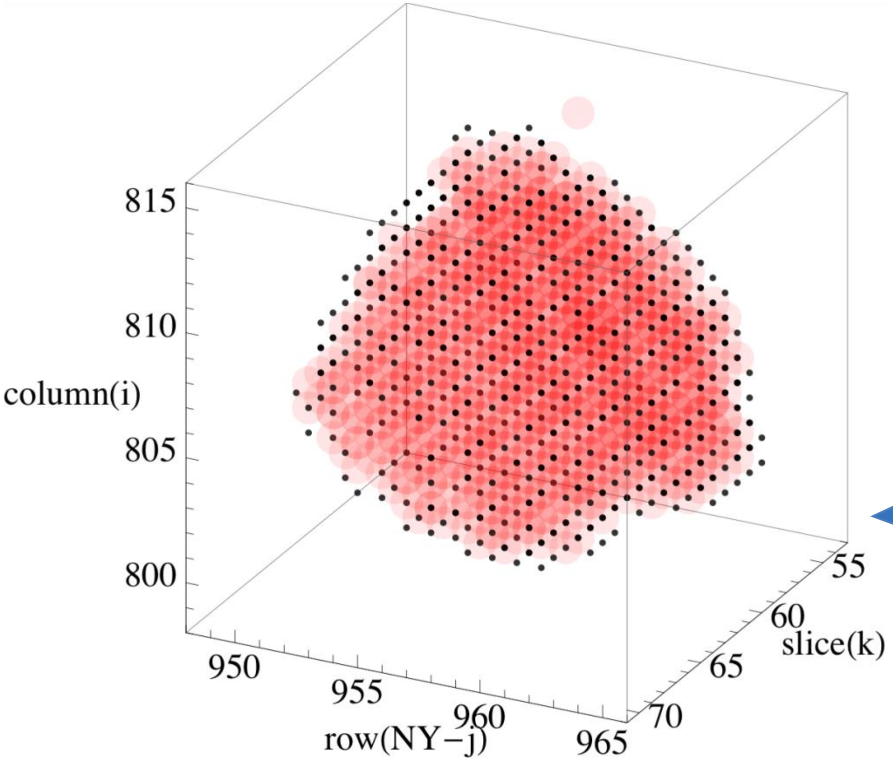
- A key optimization for GPUs is replacing a single outer product with a panel-panel matrix multiplication by employing the technique of delayed updating. This increases the computation to memory access ratio.

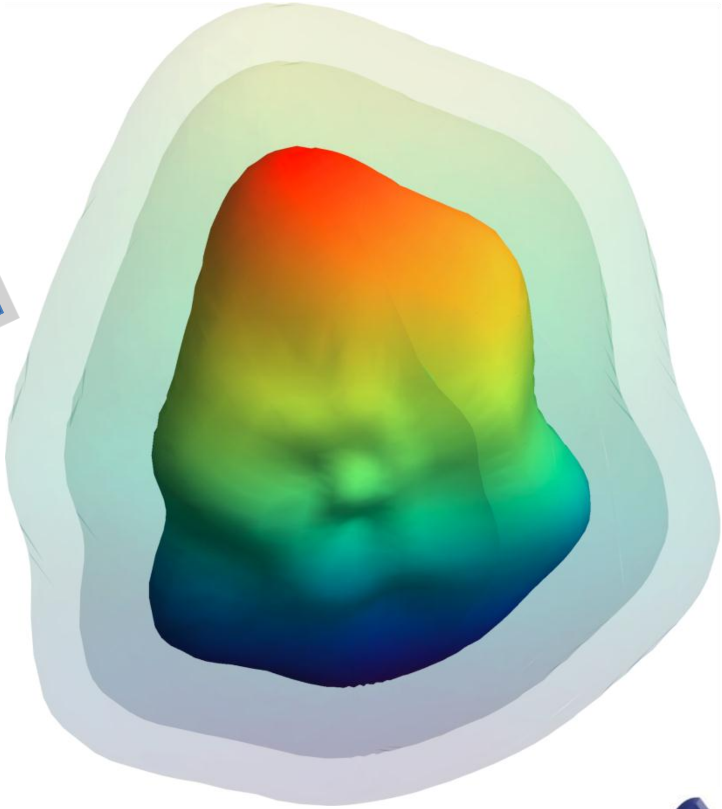
$$G_{ij}^{(M)} = G_{ij}^{(1)} + \sum_{N=1}^{M} A_i^{(N)} B_j^{(N)} \rightarrow$$



Seeding the Experimental Inverse Problem (see Poster)

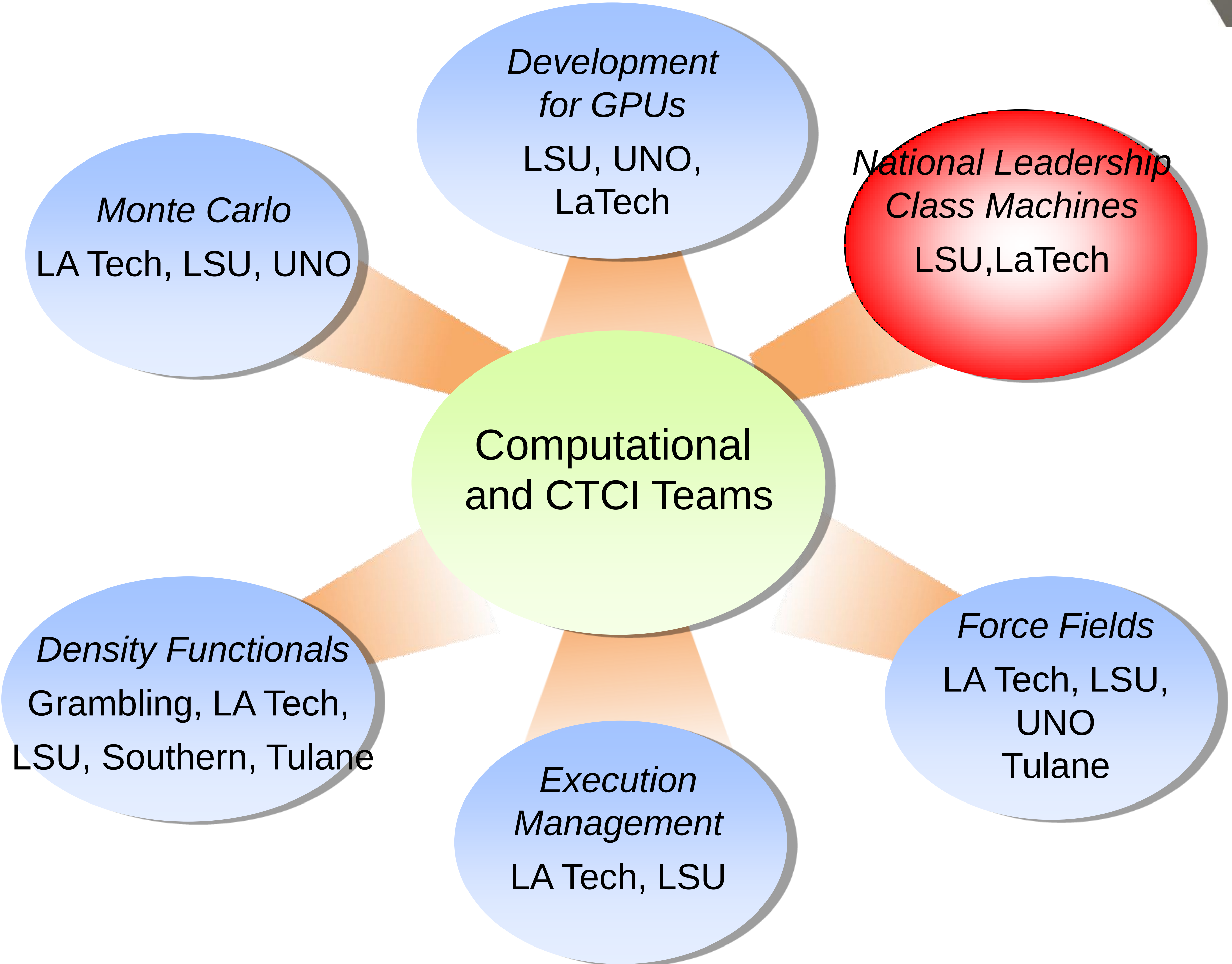
From X-ray and neutron scattering data to images of flame retardants in polymer blends





Using results of actual observations to infer the values of the parameters characterizing the system under investigation

Computational/CTCI Research Themes



Computational and CTCI Teams

Development for GPUs

LSU, UNO, LaTech

National Leadership Class Machines

LSU,LaTech

Monte Carlo

LA Tech, LSU, UNO

Density Functionals

Grambling, LA Tech, LSU, Southern, Tulane

Execution Management

LA Tech, LSU

Force Fields

LA Tech, LSU, UNO Tulane

Graduating to National Leadership Class (NLC) Machines

- Explore parallelism and scalability
- Get experience with code development on smaller clusters such as LONI/TeraGrid-XSEDE
- Demonstrate how your codes will scale to the NLC machines
- Apply for compute time on NLC machines

Current XSEDE Allocations 2012

• Bishop @LaTech:	7.8MSU
• Mobley @UNO:	1.2MSU
• Morena @LSU:	7.5MSU
• Sun @Tulane :	200,000 SU

The Little Fe Project:
Recruiting the next users.

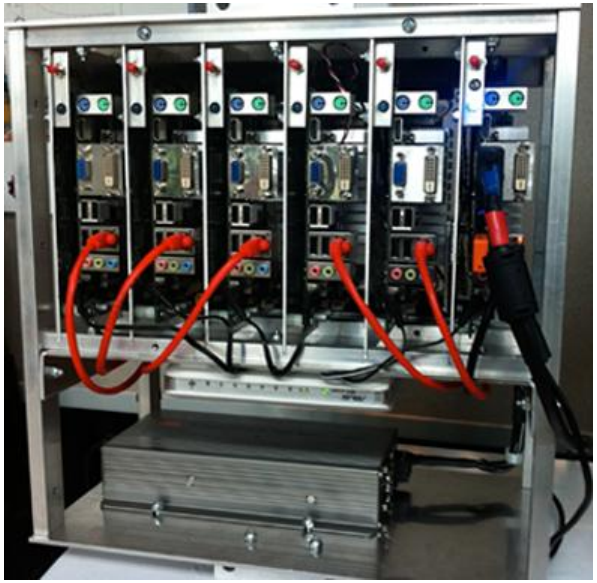
SuperComputing 2011: Little Fe Build Out Session.

- Groups Selected by Application/proposal process
- 12 Little Fe's Awarded: 3 in Louisiana

Louisiana Tech

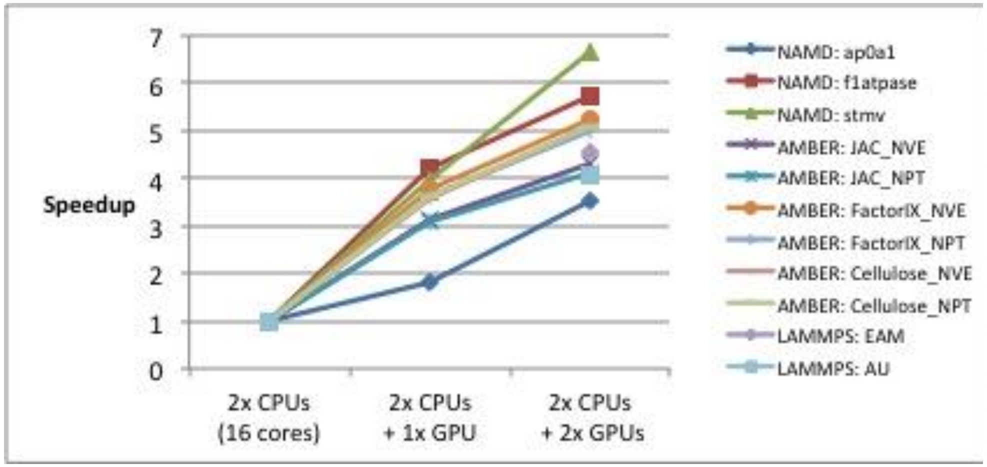
Louisiana School Science Math and the Arts (High School)

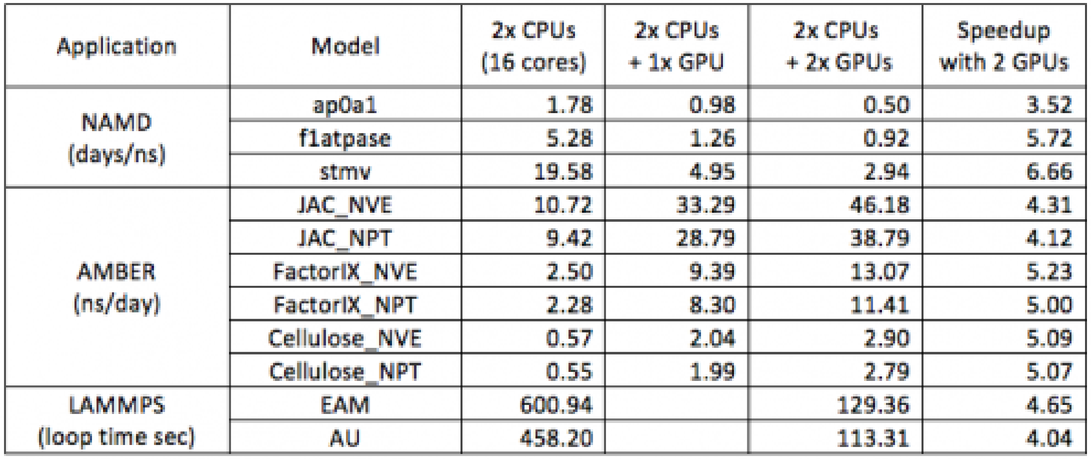
LSU



Computers LA-SiGMA can access

- A 8-node Dell PowerEdge R70 GPU cluster:
  - 2 Sandy Bridge 2.5GHz CPUs, 64GB mem
  - 2 NVIDIA Tesla M2090 6GB GPU
- The new SuperMike:
  - Minimum of 146 CPU Tflops in 382 CPU nodes:
    - 2x Sandy Bridge 8-core 2.6GHz, 32GB mem
  - 66 GPU Tflops from 50 GPU M2090 nodes
    - 2x Sandy Bridge 8-core 2.6GHz, 64GB mem
    - 2x NVIDIA M2090 GPUs



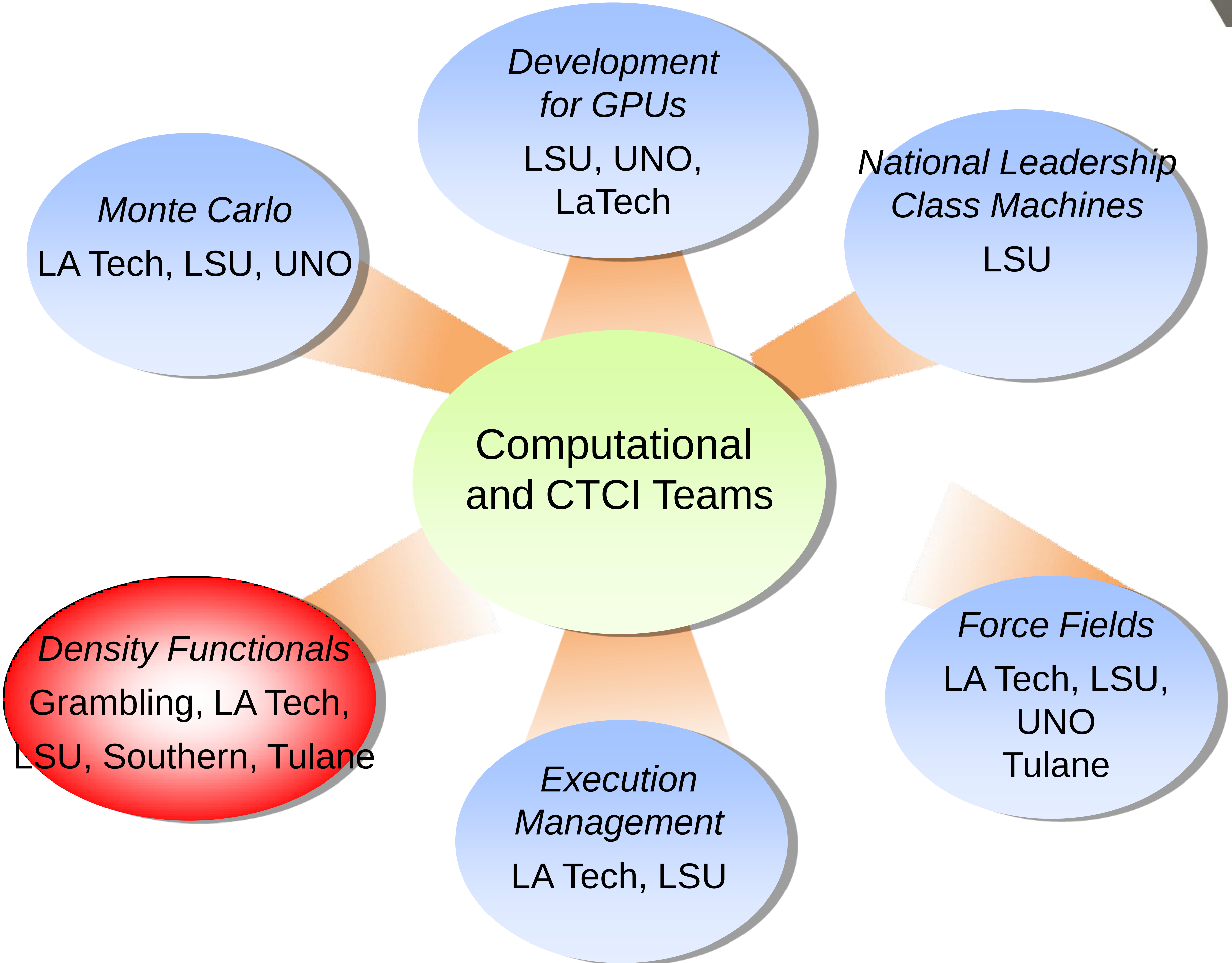


Application	Model	2x CPUs (16 cores)	2x CPUs + 1x GPU	2x CPUs + 2x GPUs	Speedup with 2 GPUs
NAMD (days/ns)	ap0a1	1.78	0.98	0.50	3.52
	f1atpase	5.28	1.26	0.92	5.72
	stmv	19.58	4.95	2.94	6.66
AMBER (ns/day)	JAC_NVE	10.72	33.29	46.18	4.31
	JAC_NPT	9.42	28.79	38.79	4.12
	FactorIX_NVE	2.50	9.39	13.07	5.23
	FactorIX_NPT	2.28	8.30	11.41	5.00
	Cellulose_NVE	0.57	2.04	2.90	5.09
	Cellulose_NPT	0.55	1.99	2.79	5.07
LAMMPS (loop time sec)	EAM	600.94		129.36	4.65
	AU	458.20		113.31	4.04

- NSF CRI (H. Liu, PI) purchase an NVIDIA Kepler (K20) system
  - At least 24 GPU nodes with:
    - 2 Intel Sandy Bridge-EP processors, 64GB memory,
    - At least 2 Kepler K20 GPUs.

R

Computational/CTCI Research Themes



Computational and CTCI Teams

Development for GPUs

LSU, UNO, LaTech

Monte Carlo

LA Tech, LSU, UNO

National Leadership Class Machines

LSU

Density Functionals

Grambling, LA Tech, LSU, Southern, Tulane

Execution Management

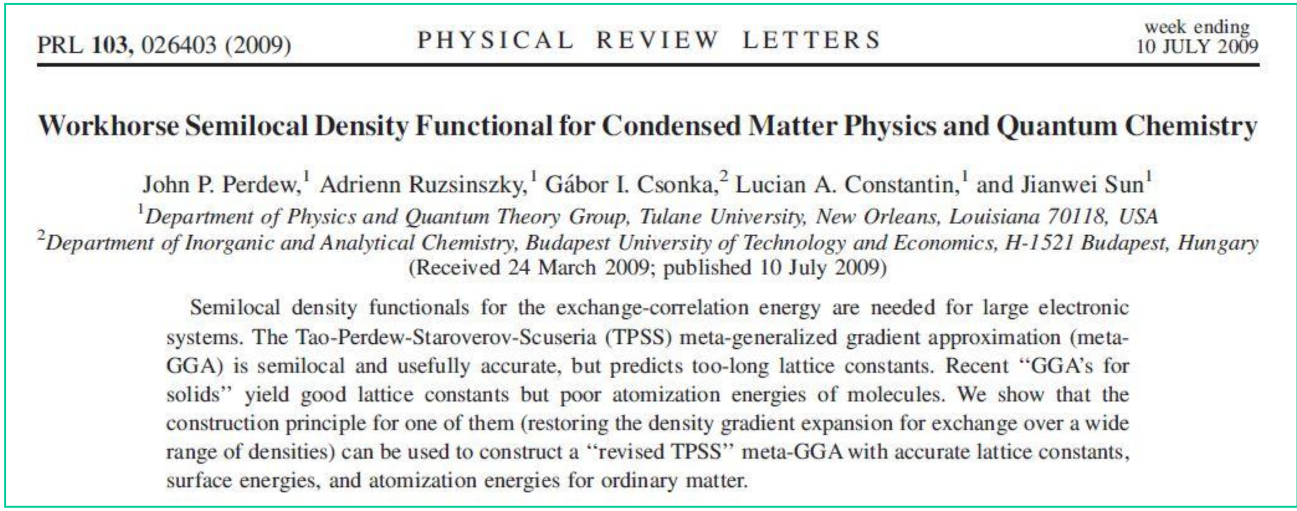
LA Tech, LSU

Force Fields

LA Tech, LSU, UNO Tulane

Density Functionals and Force Fields (rsv)

Perdew group has developed a “work-horse semilocal functional” [*Phys. Rev. Lett*. 103, 026403 (2009)] for large electronic systems that yields accurate lattice constants, surface energies, and atomization energies.



PRL **103,** 026403 (2009) PHYSICAL REVIEW LETTERS week ending 10 JULY 2009

Workhorse Semilocal Density Functional for Condensed Matter Physics and Quantum Chemistry

John P. Perdew,[1] Adrienn Ruzsinszky,[1] Gábor I. Csonka,[2] Lucian A. Constantin,[1] and Jianwei Sun[1]

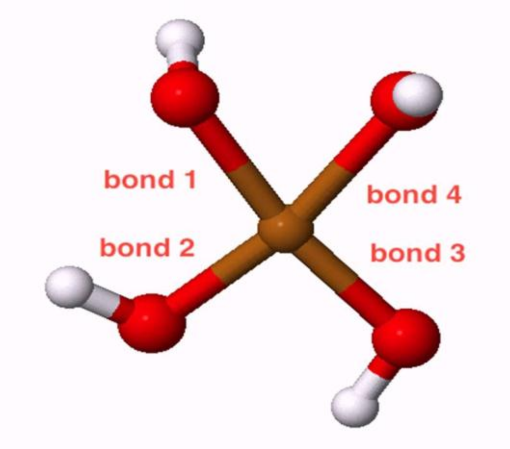
[1]*Department of Physics and Quantum Theory Group, Tulane University, New Orleans, Louisiana 70118, USA*

[2]*Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, H-1521 Budapest, Hungary*

(Received 24 March 2009; published 10 July 2009)

Semilocal density functionals for the exchange-correlation energy are needed for large electronic systems. The Tao-Perdew-Staroverov-Scuseria (TPSS) meta-generalized gradient approximation (meta-GGA) is semilocal and usefully accurate, but predicts too-long lattice constants. Recent “GGA’s for solids” yield good lattice constants but poor atomization energies of molecules. We show that the construction principle for one of them (restoring the density gradient expansion for exchange over a wide range of densities) can be used to construct a “revised TPSS” meta-GGA with accurate lattice constants, surface energies, and atomization energies for ordinary matter.

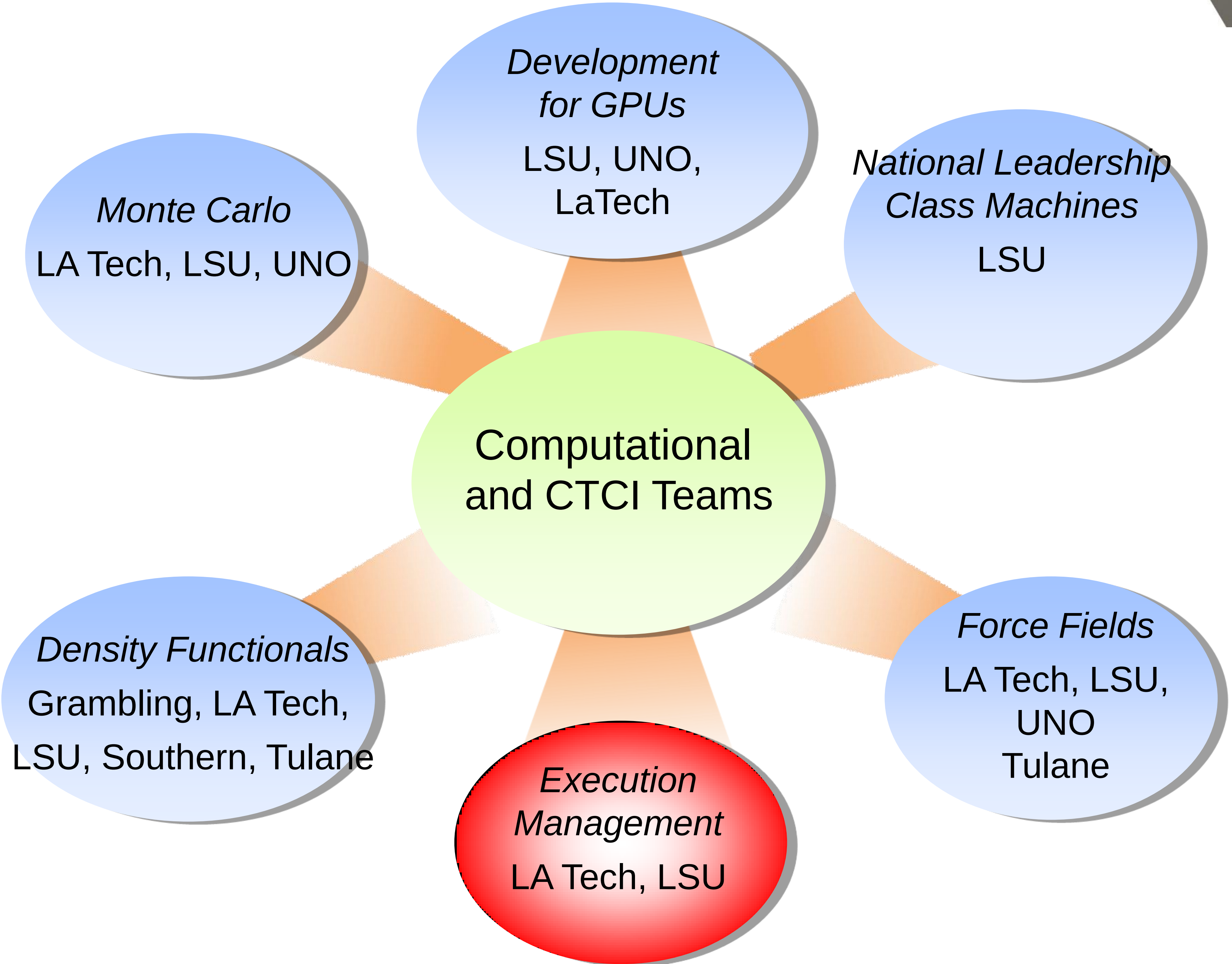
- This functional has been incorporated into VASP, a massively parallel DFT code.
- We are constructing force fields combining *ab initio* calculations of small clusters with different DFT functionals and bulk simulations/calculations.



$Cu_4O_4H_4$ cluster

LSU (Hall, Dellinger), La Tech (Wick, Ramachandran)

Computational/CTCI Research Themes



Computational and CTCI Teams

Development for GPUs

LSU, UNO, LaTech

National Leadership Class Machines

LSU

Monte Carlo

LA Tech, LSU, UNO

Density Functionals

Grambling, LA Tech, LSU, Southern, Tulane

Execution Management

LA Tech, LSU

Force Fields

LA Tech, LSU, UNO Tulane

Ensemble Based MD Simulation Challenges

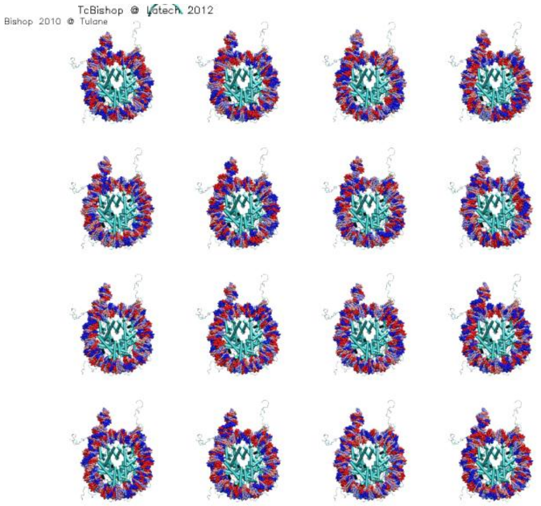
Nucleosome Simulations on XSEDE

336 Systems * 20ns = 6,720 tasks

64-128cpu/task * 8hrs ~ 3.5 MSU

5Gb/task * 6720 ~ 34 TB Data

NAMD with BigJobs



16 nucleosomes * 21 positions = 336 systems
160,000 atoms per system

DNA: Simulations on LONI

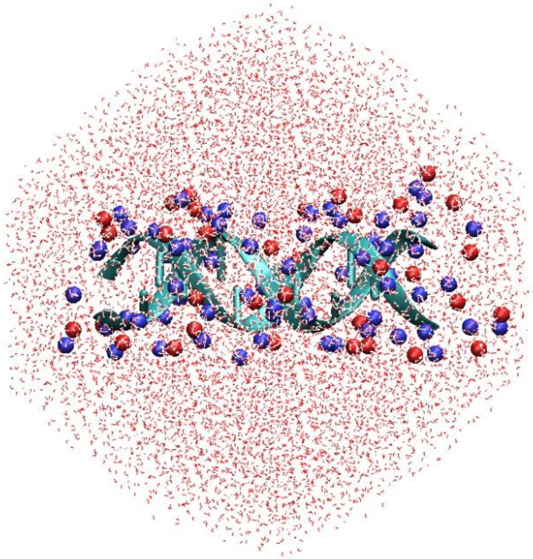
4 Systems * 1000ns = 4,000 tasks

128 cpu/task * 2.5hrs ~ 1.3 MSU

1Gb/task * 4000 ~ 4TB Data

Amber with ManyJobs

ABC: International Collaboration



4 sequences (18bp) = 4 systems
47,000 atoms per system

Scale Across
(Bishop & Jha)

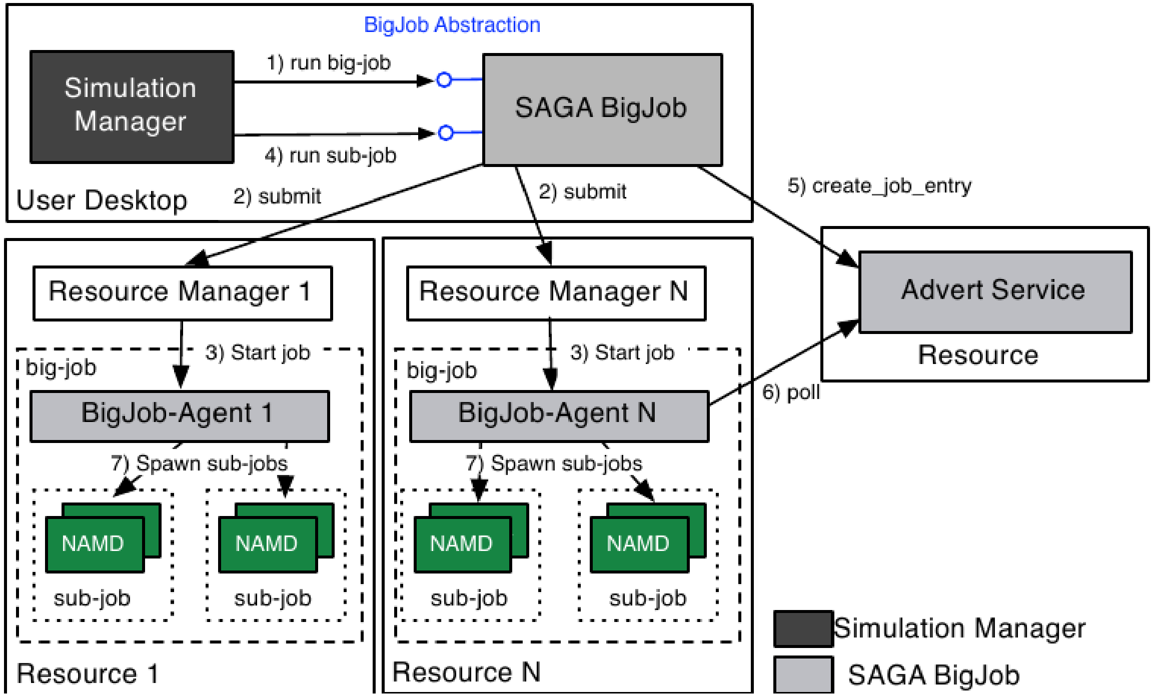
- efficiently distribute computations across computers:
  XSEDE
  LONI
  Local Clusters
  Whatever

Pilot Job Concept

Two Implementations:

BigJobs: SAGA

ManyJobs: Python



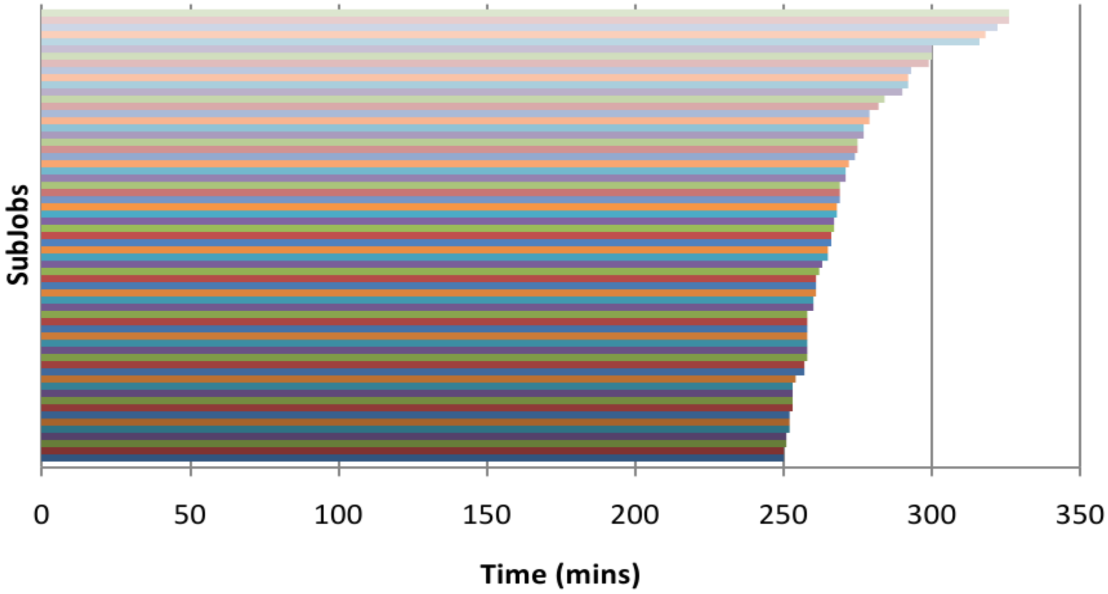
BigJobs

- SAGA: Slide in API for Grid Applications

- LONI, XSEDE and
national grade infrastructure

- Recently restructured and deployed on XSEDE
Advert service on XSEDE VM Data Quarry.

Updated documentation and examples
https://github.com/sagaproject/BigJob/wiki

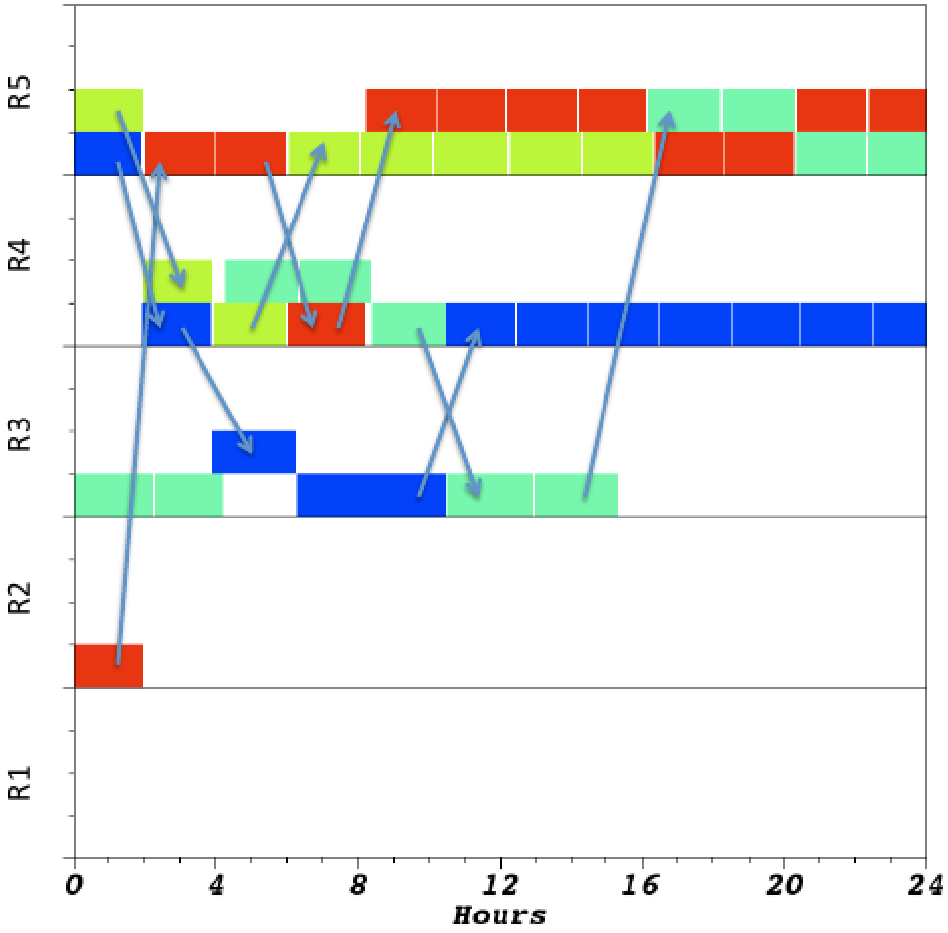


Nucleosome Ensemble
63 Simulations * 192 Core
12,096 CPU on Kraken
Min 4hr Run time
1 ns of 160,000 atom system

ManyJobs

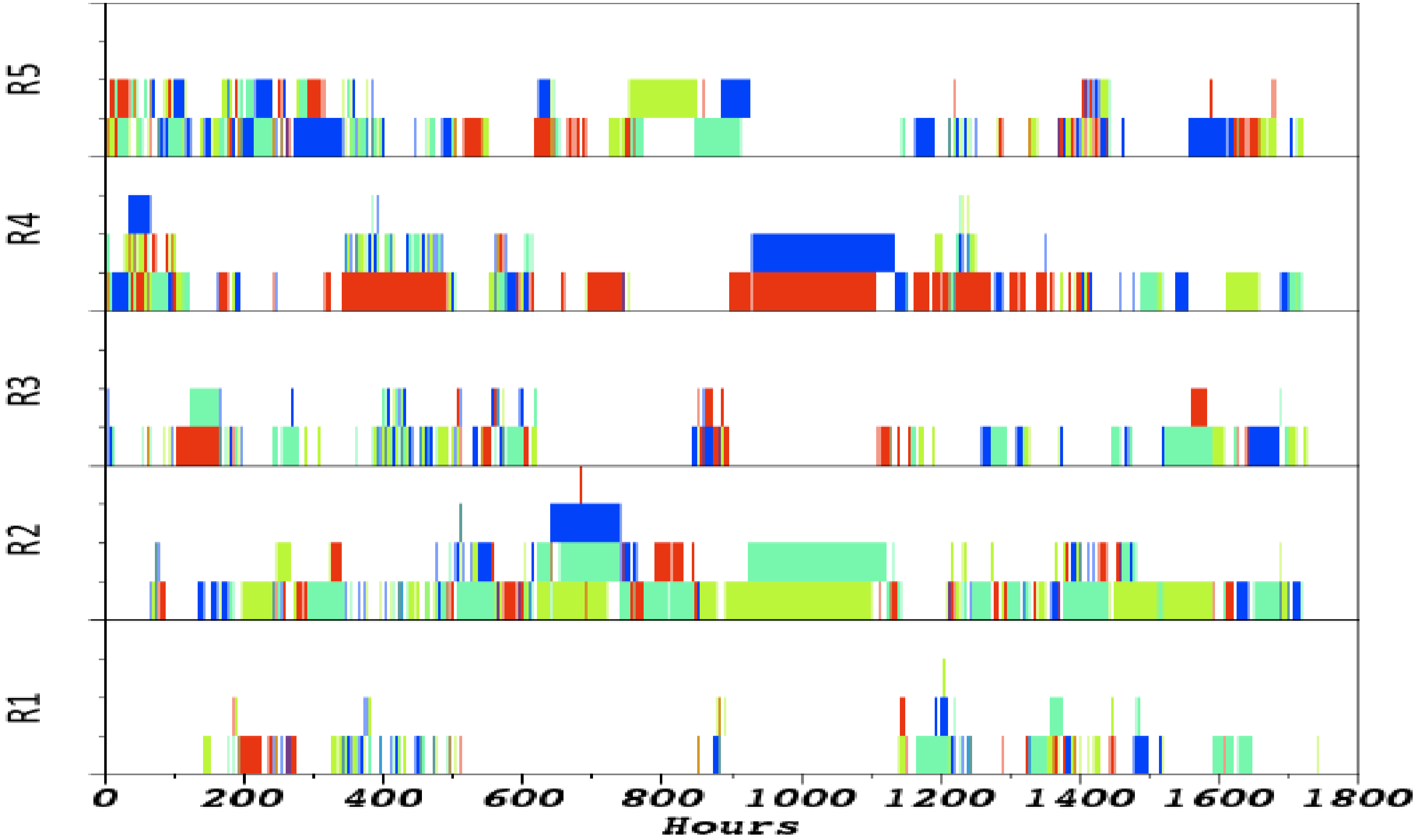
- Python Based with ssh (gsi-ssh)
  - “no prerequisites”
  - Easy deployment

- Clusters, LONI, XSEDE



DNA Ensembles
4 Simulations * 128 Core
5 LONI Machines
Min 2hr run time for 1ns
3,600 ns of 40,000 atom systems

ManyJobs



The Ascona B-DNA Consortium Simulation Effort
Coordinating > 20 Int.'l research groups and > 100,000 Simulations
Bishop: 3 month run period ~ 3500 simulations ~ 900,000SU on LONI

Progress: XSEDE 12

Running Many Molecular Dynamics Simulations on Many Supercomputers

Rajib Mukherjee, Abhinav Thota, Hideki Fujioka, Thomas C. Bishop and Shantenu Jha

The Anatomy of a Successful ECSS Project: Lessons of Supporting High Throughput High Performance Ensembles on XSEDE

Melissa Romanus, Pradeep Mantha, Yaakoub El Khamra, Andre Merzky,Shantenu Jha, Matt McKenzie and Thomas C Bishop

XSEDE Campus Bridge Early Adopter Program:

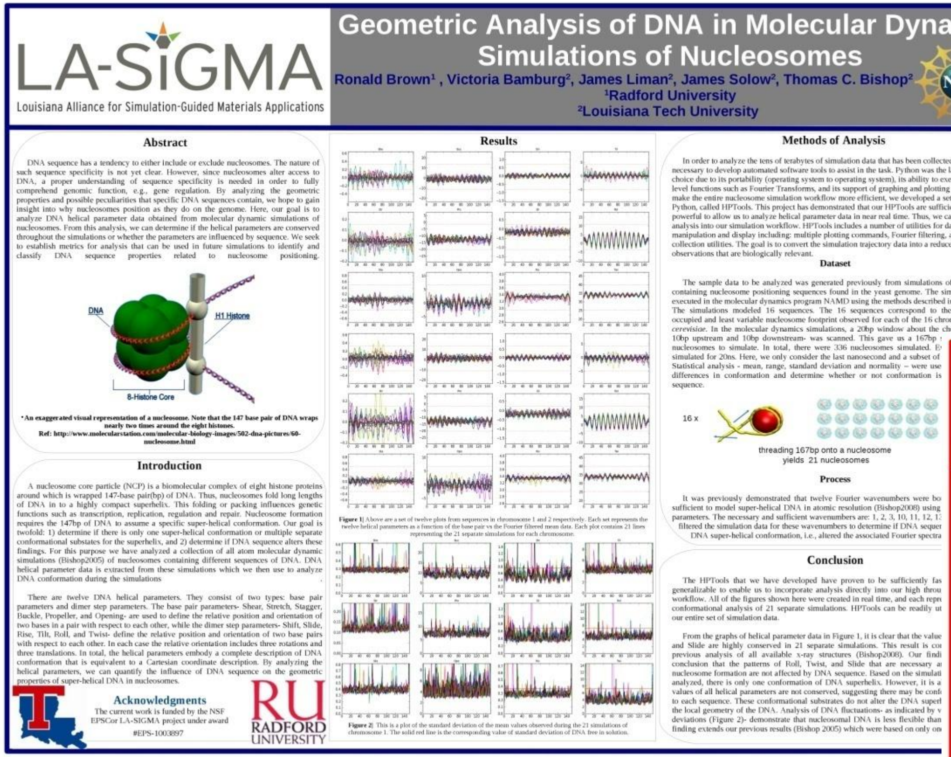
Global Federated File System (GFFS) Pilot Project:

Goal to incorporate GFFS technology into High Performance High Throughput Simulation Workflow.

C.Stewart, R.Knepper, T.Miller, A. Grimshaw, T.C. Bishop, S. Jha.

Posters

Rocky Brown, REU from Radford



LA-SiGMA
Louisiana Alliance for Simulation-Guided Materials Applications

Geometric Analysis of DNA in Molecular Dynamic Simulations of Nucleosomes

Ronald Brown[1] , Victoria Bamburg[2], James Liman[2], James Solow[2], Thomas C. Bishop[2]
[1]Radford University
[2]Louisiana Tech University

Abstract

DNA sequence has a tendency to either include or exclude nucleosomes. The nature of such sequence specificity is not yet clear. However, since nucleosomes alter access to DNA, a proper understanding of sequence specificity is needed in order to fully comprehend genomic function, e.g., gene regulation. By analyzing the geometric properties and possible peculiarities that specific DNA sequences contain, we hope to gain insight into why nucleosomes position as they do on the genome. Here, our goal is to analyze DNA helical parameter data obtained from molecular dynamic simulations of nucleosomes. From this analysis, we can determine if the helical parameters are conserved throughout the simulations or whether the parameters are influenced by sequence. We seek to establish metrics for analysis that can be used in future simulations to identify and classify DNA sequence properties related to nucleosome positioning.

*An exaggerated visual representation of a nucleosome. Note that the 147 base pair of DNA wraps nearly two times around the eight histones.
Ref: http://www.molecularstation.com/molecular-biology-images/502-dna-pictures/60-nucleosome.html

Introduction

A nucleosome core particle (NCP) is a biomolecular complex of eight histone proteins around which is wrapped 147-base pair(bp) of DNA. Thus, nucleosomes fold long lengths of DNA in to a highly compact superhelix. This folding or packing influences genetic functions such as transcription, replication, regulation and repair. Nucleosome formation requires the 147bp of DNA to assume a specific super-helical conformation. Our goal is twofold: 1) determine if there is only one super-helical conformation or multiple separate conformational substates for the superhelix, and 2) determine if DNA sequence alters these findings. For this purpose we have analyzed a collection of all atom molecular dynamic simulations (Bishop2005) of nucleosomes containing different sequences of DNA. DNA helical parameter data is extracted from these simulations which we then use to analyze DNA conformation during the simulations .

There are twelve DNA helical parameters. They consist of two types: base pair parameters and dimer step parameters. The base pair parameters- Shear, Stretch, Stagger, Buckle, Propeller, and Opening- are used to define the relative position and orientation of two bases in a pair with respect to each other, while the dimer step parameters- Shift, Slide, Rise, Tilt, Roll, and Twist- define the relative position and orientation of two base pairs with respect to each other. In each case the relative orientation includes three rotations and three translations. In total, the helical parameters embody a complete description of DNA conformation that is equivalent to a Cartesian coordinate description. By analyzing the helical parameters, we can quantify the influence of DNA sequence on the geometric properties of super-helical DNA in nucleosomes.

Acknowledgments

The current work is funded by the NSF EPSCor LA-SIGMA project under award #EPS-1003897

RU
RADFORD
UNIVERSITY

Results

Figure 1| Above are a set of twelve plots from sequences in chromosome 1 and 2 respectively. Each set represents the twelve helical parameters as a function of the base pair vs the Fourier filtered mean data. Each plot contain 21 lines representing the 21 separate simulations for each chromosome.

Figure 2| This is a plot of the standard deviation of the mean values observed during the 21 simulations of chromosome 1. The solid red line is the corresponding value of standard deviation of DNA free in solution.

Methods of Analysis

In order to analyze the tens of terabytes of simulation data that has been collected, it was necessary to develop automated software tools to assist in the task. Python was the language of choice due to its portability (operating system to operating system), its ability to execute high level functions such as Fourier Transforms, and its support of graphing and plotting utilities. To make the entire nucleosome simulation workflow more efficient, we developed a set of tools in Python, called HPTools. This project has demonstrated that our HPTools are sufficiently fast and powerful to allow us to analyze helical parameter data in near real time. Thus, we can incorporate analysis into our simulation workflow. HPTools includes a number of utilities for data manipulation and display including: multiple plotting commands, Fourier filtering, and file collection utilities. The goal is to convert the simulation trajectory data into a reduced set of observations that are biologically relevant.

Dataset

The sample data to be analyzed was generated previously from simulations of nucleosomes containing nucleosome positioning sequences found in the yeast genome. The simulations were executed in the molecular dynamics program NAMD using the methods described in Bishop2005. The simulations modeled 16 sequences. The 16 sequences correspond to the most highly occupied and least variable nucleosome footprint observed for each of the 16 chromosomes in *S. cerevisiae*. In the molecular dynamics simulations, a 20bp window about the chosen position- 10bp upstream and 10bp downstream- was scanned. This gave us a 167bp nucleosomes to simulate. In total, there were 336 nucleosomes simulated. E simulated for 20ns. Here, we only consider the last nanosecond and a subset of Statistical analysis - mean, range, standard deviation and normality – were use differences in conformation and determine whether or not conformation is sequence.

16 x

threading 167bp onto a nucleosome
yields 21 nucleosomes

Process

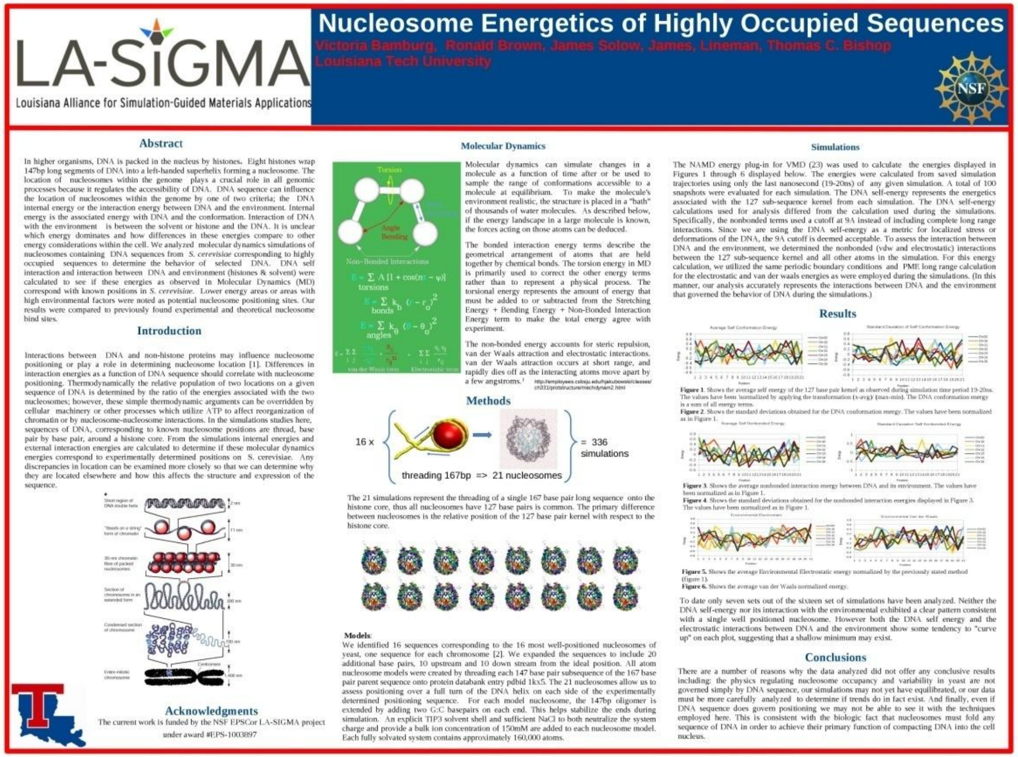
It was previously demonstrated that twelve Fourier wavenumbers were bo sufficient to model super-helical DNA in atomic resolution (Bishop2008) using parameters. The necessary and sufficient wavenumbers are: 1, 2, 3, 10, 11, 12, 1 filtered the simulation data for these wavenumbers to determine if DNA sequer DNA super-helical conformation, i.e., altered the associated Fourier spectra

Conclusion

The HPTools that we have developed have proven to be sufficiently fas generalizable to enable us to incorporate analysis directly into our high throu workflow. All of the figures shown here were created in real time, and each repr conformational analysis of 21 separate simulations. HPTools can be readily ut our entire set of simulation data.

From the graphs of helical parameter data in Figure 1, it is clear that the value and Slide are highly conserved in 21 separate simulations. This result is co previous analysis of all available x-ray structures (Bishop2008). Our findi conclusion that the patterns of Roll, Twist, and Slide that are necessary a nucleosome formation are not affected by DNA sequence. Based on the simulati analyzed, there is only one conformation of DNA superhelix. However, it is a values of all helical parameters are not conserved, suggesting there may be conf to each sequence. These conformational substrates do not alter the DNA superh the local geometry of the DNA. Analysis of DNA fluctuations- as indicated by v deviations (Figure 2)- demonstrate that nucleosomal DNA is less flexible than finding extends our previous results (Bishop 2005) which were based on only on

Victoria Bamburg, REU from LSMA



LA-SiGMA
Louisiana Alliance for Simulation-Guided Materials Applications

Nucleosome Energetics of Highly Occupied Sequences

Victoria Bamburg, Ronald Brown, James Solow, James, Lineman, Thomas C. Bishop
Louisiana Tech University

Abstract

In higher organisms, DNA is packed in the nucleus by histones. Eight histones wrap 147bp long segments of DNA into a left-handed superhelix forming a nucleosome. The location of nucleosomes within the genome plays a crucial role in all genomic processes because it regulates the accessibility of DNA. DNA sequence can influence the location of nucleosomes within the genome by one of two criteria; the DNA internal energy or the interaction energy between DNA and the environment. Internal energy is the associated energy with DNA and the conformation. Interaction of DNA with the environment is between the solvent or histone and the DNA. It is unclear which energy dominates and how differences in these energies compare to other energy considerations within the cell. We analyzed molecular dynamics simulations of nucleosomes containing DNA sequences from *S. cerevisiae* corresponding to highly occupied sequences to determine the behavior of selected DNA. DNA self interaction and interaction between DNA and environment (histones & solvent) were calculated to see if these energies as observed in Molecular Dynamics (MD) correspond with known positions in *S. cerevisiae*. Lower energy areas or areas with high environmental factors were noted as potential nucleosome positioning sites. Our results were compared to previously found experimental and theoretical nucleosome bind sites.

Introduction

Interactions between DNA and non-histone proteins may influence nucleosome positioning or play a role in determining nucleosome location [1]. Differences in interaction energies as a function of DNA sequence should correlate with nucleosome positioning. Thermodynamically the relative population of two locations on a given sequence of DNA is determined by the ratio of the energies associated with the two nucleosomes; however, these simple thermodynamic arguments can be overridden by cellular machinery or other processes which utilize ATP to affect reorganization of chromatin or by nucleosome-nucleosome interactions. In the simulations studies here, sequences of DNA, corresponding to known nucleosome positions are thread, base pair by base pair, around a histone core. From the simulations internal energies and external interaction energies are calculated to determine if these molecular dynamics energies correspond to experimentally determined positions on S. cerevisiae. Any discrepancies in location can be examined more closely so that we can determine why they are located elsewhere and how this affects the structure and expression of the sequence.

Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

Acknowledgments

The current work is funded by the NSF EPSCor LA-SIGMA project under award #EPS-1003897

Molecular Dynamics

Molecular dynamics can simulate changes in a molecule as a function of time after or be used to sample the range of conformations accessible to a molecule at equilibrium. To make the molecule's environment realistic, the structure is placed in a "bath" of thousands of water molecules. As described below, if the energy landscape in a large molecule is known, the forces acting on those atoms can be deduced.

The bonded interaction energy terms describe the geometrical arrangement of atoms that are held together by chemical bonds. The torsion energy in MD is primarily used to correct the other energy terms rather than to represent a physical process. The torsional energy represents the amount of energy that must be added to or subtracted from the Stretching Energy + Bending Energy + Non-Bonded Interaction Energy term to make the total energy agree with experiment.

The non-bonded energy accounts for steric repulsion, van der Waals attraction and electrostatic interactions. van der Waals attraction occurs at short range, and rapidly dies off as the interacting atoms move apart by a few angstroms.[1] http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/mechdynam2.html

Methods

16 x = 336 simulations
threading 167bp => 21 nucleosomes

The 21 simulations represent the threading of a single 167 base pair long sequence onto the histone core, thus all nucleosomes have 127 base pairs is common. The primary difference between nucleosomes is the relative position of the 127 base pair kernel with respect to the histone core.

Models:
We identified 16 sequences corresponding to the 16 most well-positioned nucleosomes of yeast, one sequence for each chromosome [2]. We expanded the sequences to include 20 additional base pairs, 10 upstream and 10 down stream from the ideal position. All atom nucleosome models were created by threading each 147 base pair subsequence of the 167 base pair parent sequence onto protein databank entry pdbid 1kx5. 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. For each model nucleosome, the 147bp oligomer is extended by adding two G:C basepairs on each end. This helps stabilize the ends during simulation. 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.

Simulations

The NAMD energy plug-in for VMD (23) was used to calculate the energies displayed in Figures 1 through 6 displayed below. The energies were calculated from saved simulation trajectories using only the last nanosecond (19-20ns) of any given simulation. A total of 100 snapshots were evaluated for each simulation. The DNA self-energy represents the energetics associated with the 127 sub-sequence kernel from each simulation. The DNA self-energy calculations used for analysis differed from the calculation used during the simulations. Specifically, the nonbonded terms used a cutoff at 9A instead of including complete long range interactions. Since we are using the DNA self-energy as a metric for localized stress or deformations of the DNA, the 9A cutoff is deemed acceptable. To assess the interaction between DNA and the environment, we determined the nonbonded (vdw and electrostatic) interactions between the 127 sub-sequence kernel and all other atoms in the simulation. For this energy calculation, we utilized the same periodic boundary conditions and PME long range calculation for the electrostatic and van der waals energies as were employed during the simulations. (In this manner, our analysis accurately represents the interactions between DNA and the environment that governed the behavior of DNA during the simulations.)

Results

Average Self Conformation Energy
Standard Deviation of Self Conformation Energy

Figure 1. Shows the average self energy of the 127 base pair kernel as observed during simulation time period 19-20ns. The values have been 'normalized by applying the transformation (x-avg)/ (max-min). The DNA conformation energy is a sum of all energy terms.
Figure 2. Shows the standard deviations obtained for the DNA conformation energy. The values have been normalized as in Figure 1.

Average Self Nonbonded Energy
Standard Deviation Self Nonbonded Energy

Figure 3. Shows the average nonbonded interaction energy between DNA and its environment. The values have been normalized as in Figure 1.
Figure 4. Shows the standard deviations obtained for the nonbonded interaction energies displayed in Figure 3. The values have been normalized as in Figure 1.

Figure 5. Shows the average Environmental Electrostatic energy nomalized by the previously stated method (figure 1).
Figure 6. Shows the average van der Waals normalized energy.

To date only seven sets out of the sixteen set of simulations have been analyzed. Neither the DNA self-energy nor its interaction with the environmental exhibited a clear pattern consistent with a single well positioned nucleosome. However both the DNA self energy and the electrostatic interactions between DNA and the environment show some tendency to "curve up" on each plot, suggesting that a shallow minimum may exist.

Conclusions

There are a number of reasons why the data analyzed did not offer any conclusive results including: the physics regulating nucleosome occupancy and variability in yeast are not governed simply by DNA sequence, our simulations may not yet have equilibrated, or our data must be more carefully analyzed to determine if trends do in fact exist. And finally, even if DNA sequence does govern positioning we may not be able to see it with the techniques employed here. This is consistent with the biologic fact that nucleosomes must fold any sequence of DNA in order to achieve their primary function of compacting DNA into the cell nucleus.

Computational Tools for Multi-scale Simulations (Dua, LaTech)

Goal: To develop techniques, algorithms, and strategies for extracting information and knowledge from data generated by Science Drivers and create Computational Tools related efforts.

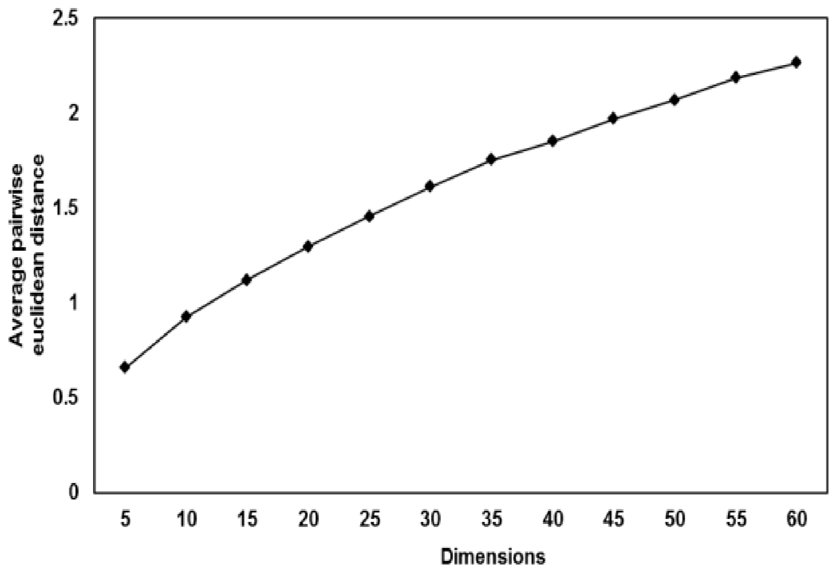


Fig. 1 Avg. pairwise Euclidean distance v/s dimensions

Efforts over the past year:

- Data adaptive rule based approach to supervised learning.
- A grid based agglomerative approach to unsupervised learning.

Future Efforts:

- Distributed data mining frameworks.
- Proposed system architecture.
- Integration of variants of the above approaches to proposed architectur

Data adaptive rule based supervised learning

Goal: To develop a data-adaptive partitioning schema of feature space for rule-based classification.

Objectives :

To develop a data adaptive partitioning scheme

To develop a method for rule extraction

To exploit the extracted rules for supervised learning / classification

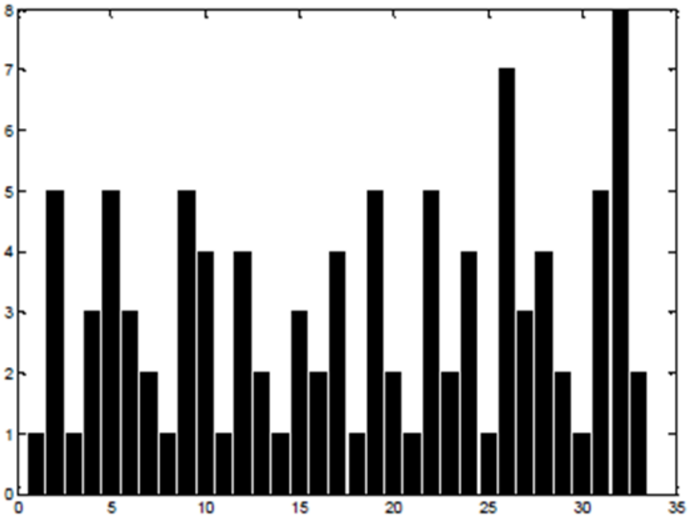


Fig. 2. Histogram plotted for one feature of a dataset.

Significance & Applications:

•Data adaptive partitioning ensures reduction in the number of rules

•Ensure the choice of rules that are both high in sensitivity and specificity

•Modular development of algorithm for easy distribution

Initial Results

Overall accuracy of proposed data adaptive partitioning classification results have compared with rule-based classifiers and non-rule based classifiers.

Classifiers	**Overall Accuracy (%)**
Rule Based Classifiers	
Conjuctive Rule	66
Decision table	77
DTNB	**82**
JRIP	66
NNGE	75
One R	62
PART	77
Ridor	**82**
ZeroR	33

Classifiers	**Overall Accuracy (%)**
Non-rule based classifiers	
Naïve Bayes	84
Logistic	73
Multi Layer Perceptron	59
RBF Network	**86**
Simple Logistic	77
SMO	**86**
Random Forest	80

Classifiers	**Overall Accuracy (%)**
Proposed data adaptive partitioning	
Slope based partitioning	**82.2**
Non-slope based partitioning	**86**

Grid-based agglomerative clustering algorithm

Goal: The goal of this research is to develop a data mining algorithm for clustering multidimensional datasets.

Objectives:

- To develop an algorithm for multi-level data adaptive grid generation.
- To develop a data preprocessing algorithm for sparseness reduction.
- To develop a grid based agglomerative hierarchical clustering algorithm.

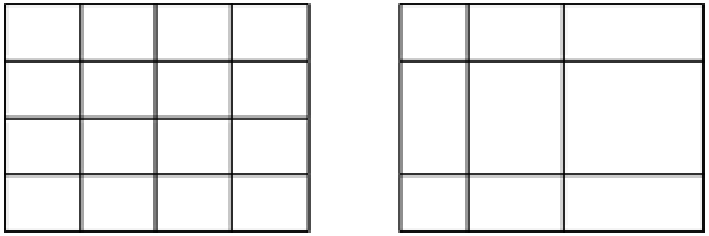
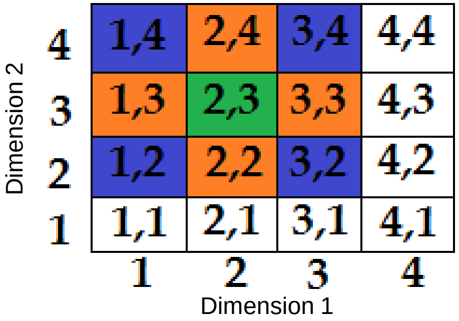


Fig. 7. (a) A uniform grid for 2D data, (b) A non-uniform grid for 2D data



Dimension 2				
4	🟦 1,4	🟧 2,4	🟦 3,4	4,4
3	🟧 1,3	🟩 2,3	🟧 3,3	4,3
2	🟦 1,2	🟧 2,2	🟦 3,2	4,2
1	1,1	2,1	3,1	4,1
	1	2	3	4
	Dimension 1			

Fig. 8. A two dimensional grid with grid cell numbering

Significance & Applications:

- Clustering algorithms augmented with a data preprocessing through sparseness reduction are more accurate and produce better clustering results.
- Our developed algorithm is generic, and is easily adaptable for other scientific pplications.

Initial Results - scalability analysis

Grid generation

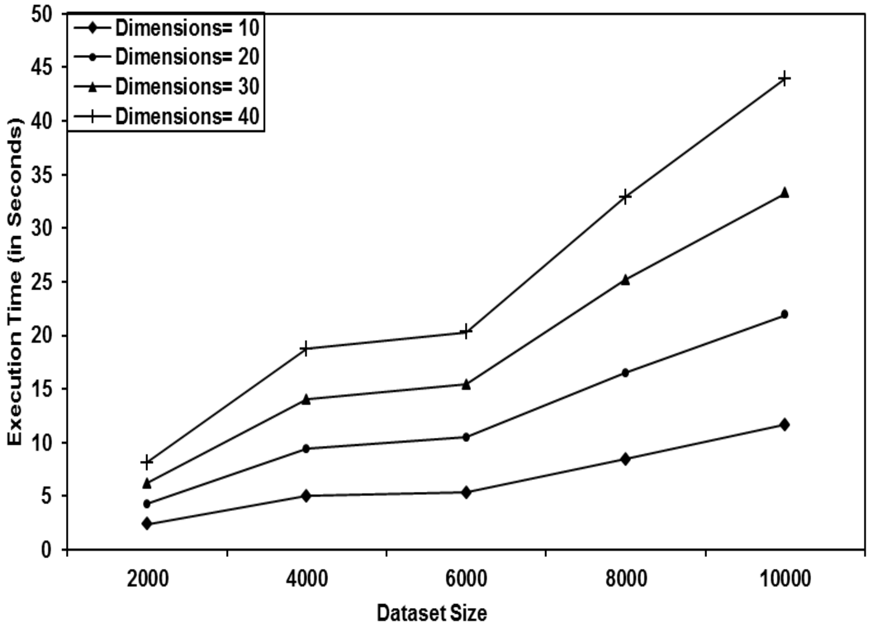


Fig.11. Execution time v/s dataset size

Clustering algorithm

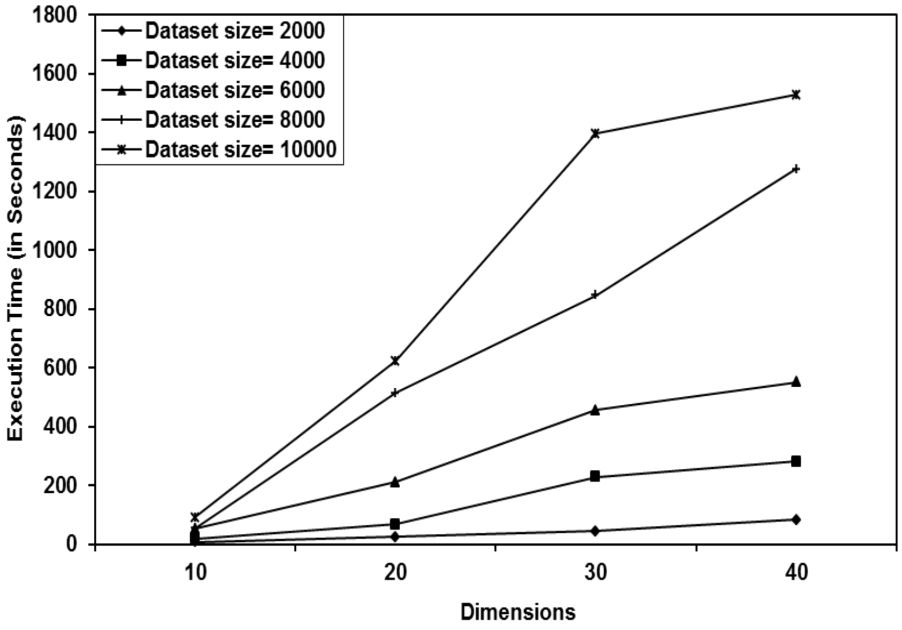


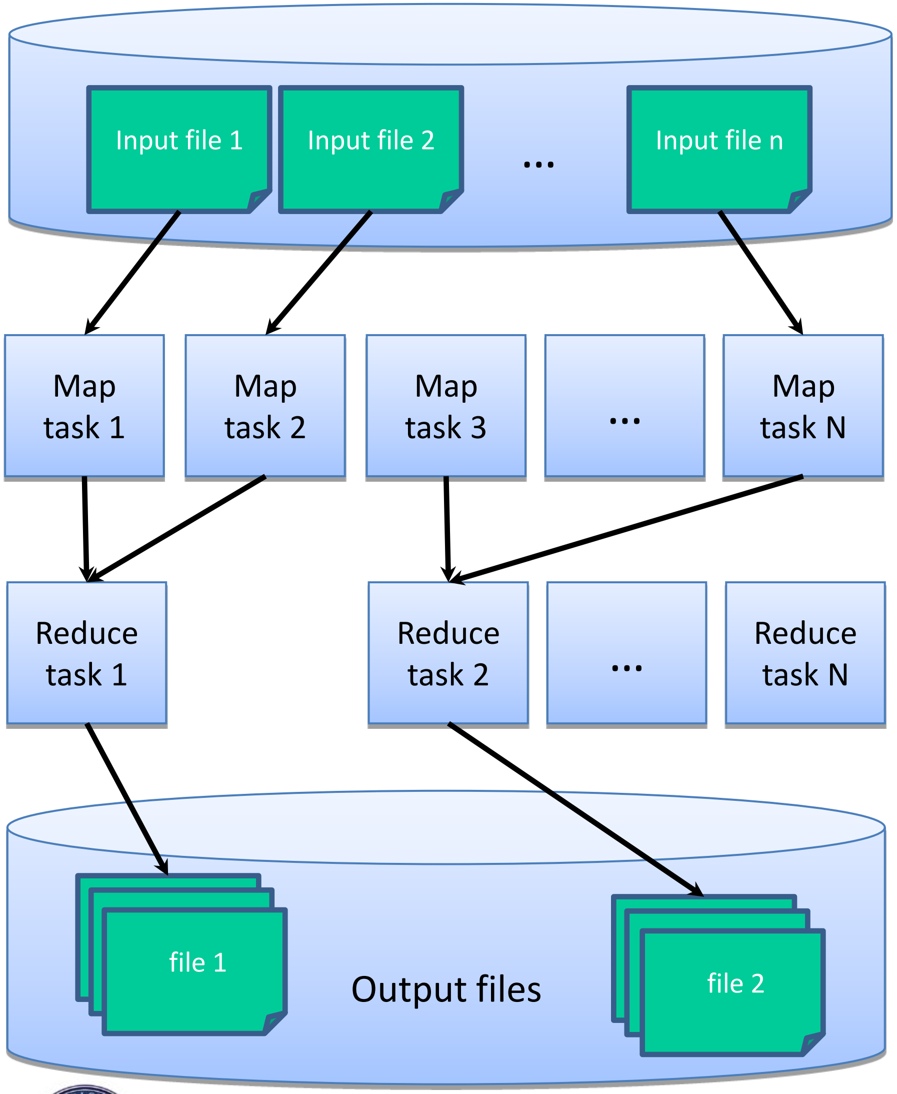
Fig.12: Execution time v/s dimensions

Proposed extension

Based on the MapReduce programming paradigm.

Apache Hadoop

- Hadoop - Distributed File System
- Hadoop - MapReduce.
  - Map function.
  - Reduce function.



Data Mining using MapReduce

Requirements:

- **Scalability:** We mean that the system can easily be altered to accommodate changes in the number of users, resources and computing entities.
- **Reliability:** Difficult to achieve as it is closely related to the complexity of the interactions between simultaneously running components.
- **Availability:** The system can restore operations, permitting it to resume providing services even when some components have failed.
- **Evolution:** Keeping up with changes to the system with newer computational features and newer requirements.

Funding and Outreach

- INCITE proposal (in collaboration with Pacific Northwest National Laboratories) for compute cycles on Jaguar and Titan
- XSEDE Allocation and GFFS incorporation into BigJobs
- ManyJobs with LONI-CS: Hideki Fujioka
- NSF CRI proposal for GPU cluster (***Shelob***) funded (includes several LA-SiGMA faculty members)
- NSF proposal for ScaleMS Bishop and Jha
- Indo-US Center (IUSSTF)
- SCiDAC and other DOE proposals
- Outreach:
  - Summer REU and RET programs
  - Beowulf Boot Camp for High School Students and Teachers
  - Little Fe
  - FEScUE with Colorado State University (Bishop)
  - GPU and Execution Management regular video meetings
  - Conference tutorials on GPUs:
    - International Symp. on Code Gen. & Opt., April 2012
    - Intl. Conf. Parallel Arch. & Comp. Tech., Oct. 2011

Summary

- CTCI: Leveraging the exponential increase in computer power
  - Recruiting new and graduating existing users to national leadership class machines
  - Preparing users for next-generation computers
  - Developing common computational toolkits
  - Expanding collaborations within LA-SiGMA and developing partnerships with national labs

"The glue" that binds the SDs

Thank You