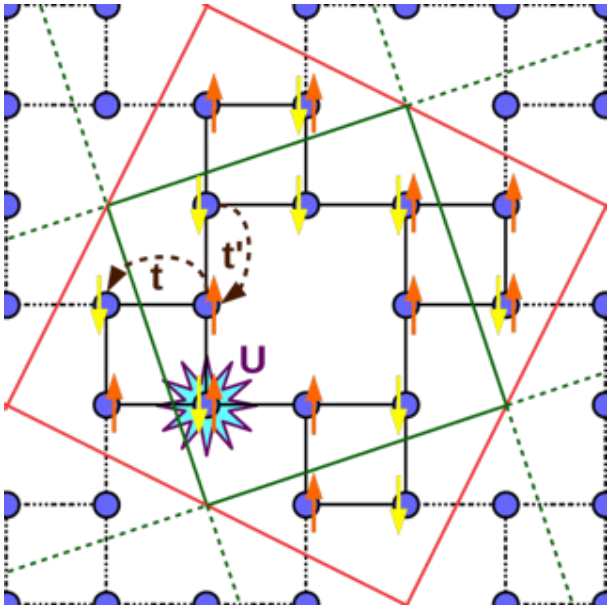
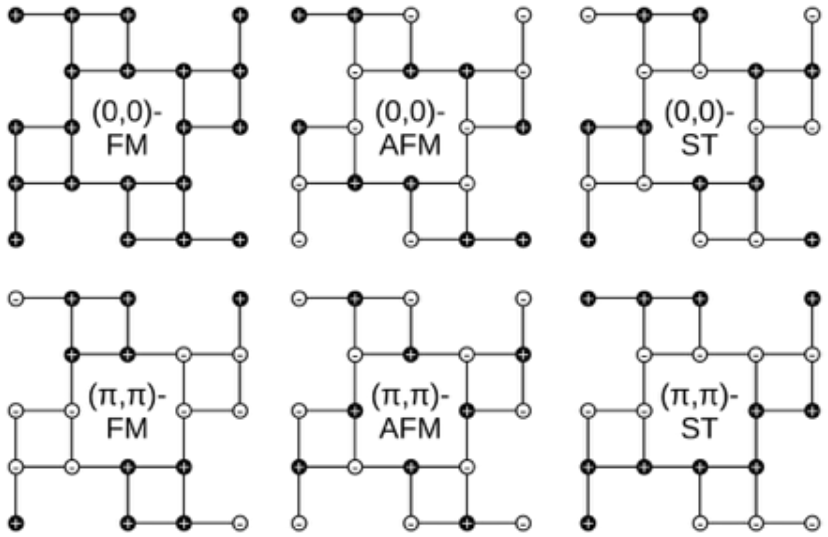
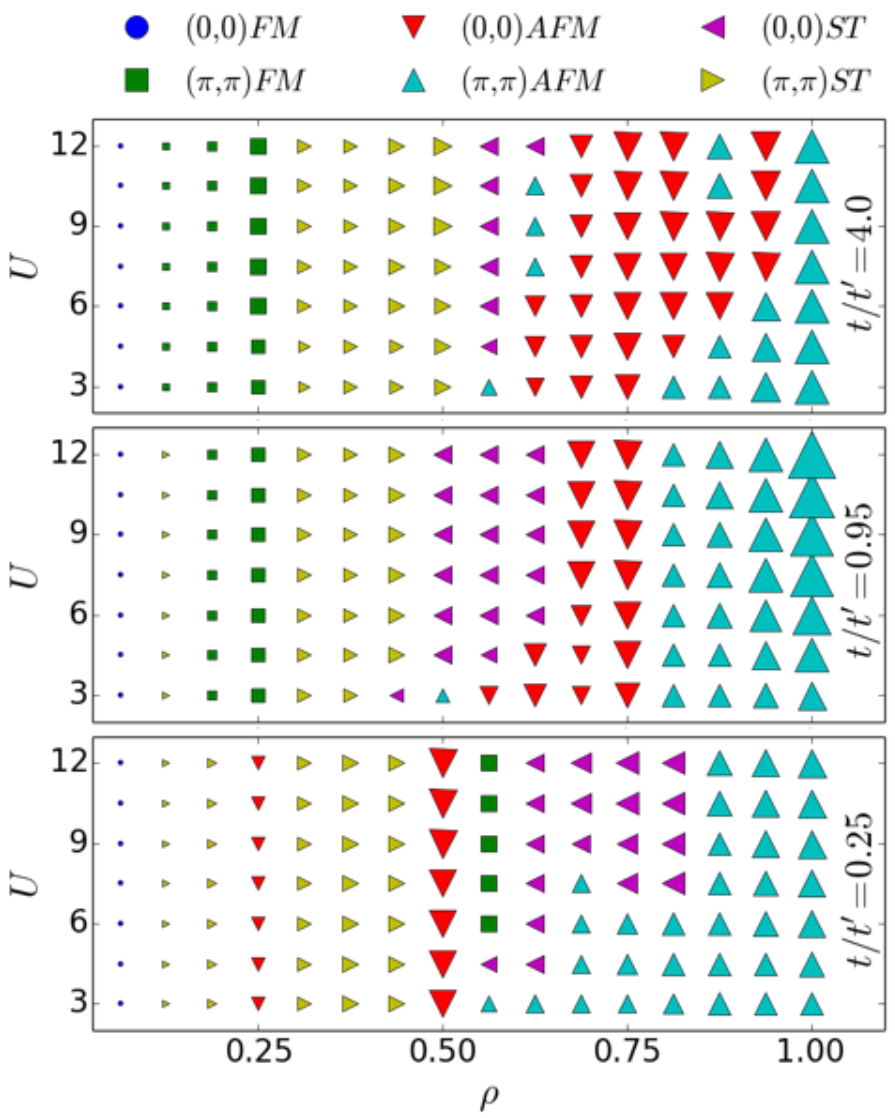
SDSC Summer Institute 2016
Lightning Rounds

Begin with those departing early

Strongly Correlated Materials







SI2016 Takeaway

- Try implementing OpenACC in parts of my next Fortran code
- Use GIT more heavily to make collaboration easier
- Be more conscientious of spatial and temporal locality in cache
- Try using a container to package my research codes so that others can easily run them
- Avoid the cloud for intense distributed computing

Caitlin Ross *(Bob stand in)*

Rensselaer Polytechnic Institute

My research group works on optimistic parallel discrete-event simulations of computer systems. We do simulations of HPC network topologies (torus, dragonfly, fat tree), storage, and workloads.

I'm currently working directly with our simulation engine to add instrumentation to analyze the performance of our simulations and help us tune the parameters to get the best performance so we can perform larger-scale simulations.

I also plan to eventually add more support to help with debugging the sims and provide a way to visually monitor the simulations (both model data and sim engine performance) in situ.

Caitlin Ross
Rensselaer Polytechnic Institute

I learned a lot that will be helpful to me:

- Lots of different tips for improving performance of code
- The Spark session gave me some ideas for how I might use it. We run a lot of different simulations and are starting to collect a lot more data and Spark seems like it may be a good way for us to handle this data and to be able to compare sim results from different parameter settings more easily
- May be able to make use of machine learning in setting all of the parameters for future models we will develop

Mehdi Safari *(not in attendance)*

Fairfield University, Fairfield, CT

I am an assistant professor at Fairfield University. My research is in the area of “large eddy simulation of turbulent reacting flows.” I have developed a novel stochastic model to take into account evolution of “entropy” and “entropy generation” in to my calculations. I develop my own code (Fortran and C++) using OpenMP and MPI and use XSEDE (TACC) for my research. Currently this model I am working on is computationally expensive (~ 30k SU for each simulations). My current research goal is to develop numerical techniques (Monte Carlo based) to reduce computational time and make it affordable for engineering applications.

Mehdi Safari

Fairfield University, Fairfield, CT

I am grateful to SDSC for the opportunity to participate in this summer school.

I learned very good techniques/ideas to implement in my HPC research, specially those sessions on machine learning and visualization. I also was exposed to new ideas (like Science Gateways and virtualization) which I am going to use in my research activities.

This summer institute also provided an opportunity to know SDSC's staff which would be of great importance to seek help through XSEDE collaborative support service. I will consider SDSC resources for future XSEDE proposals.

Carmen Wright
Jackson State University, Campus Champion
Research area: Bioinformatics

Sequence alignment algorithms use parameters to produce an alignment score that represents the similarity of the two DNA or protein sequences. Matches are rewarded while substitutions, insertions, and deletions may come at some specified cost. I seek to optimize these parameters by using SVM to learn good parameter values.

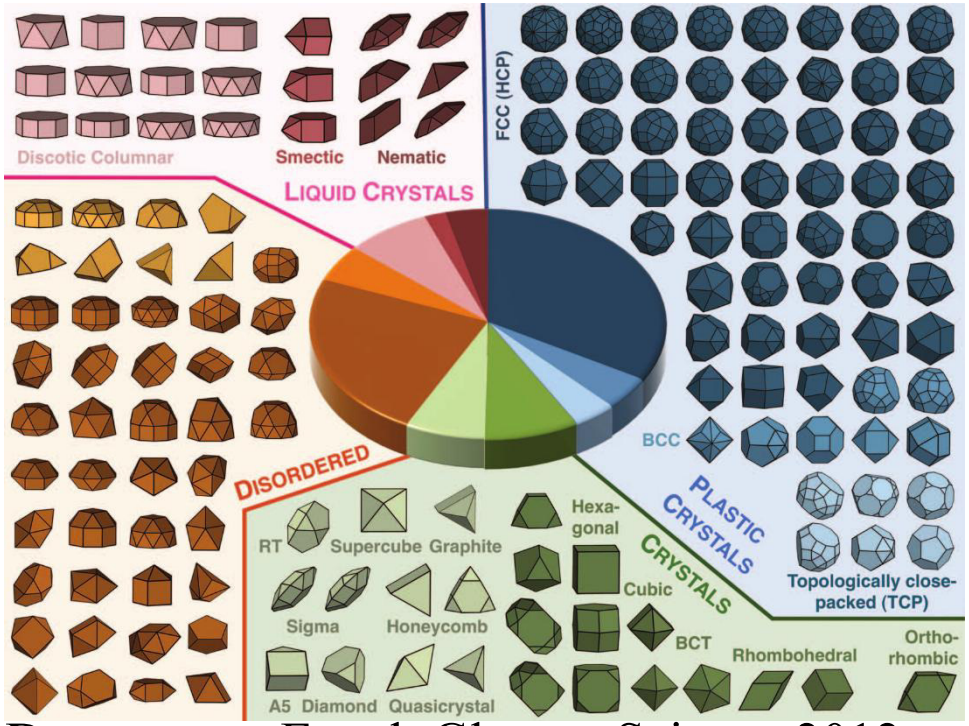
I have been using Python to write my own code – the BioPython package has an alignment module. Currently I am using a synthetic dataset of sequences for testing while I develop my code and algorithms.

Carmen Wright

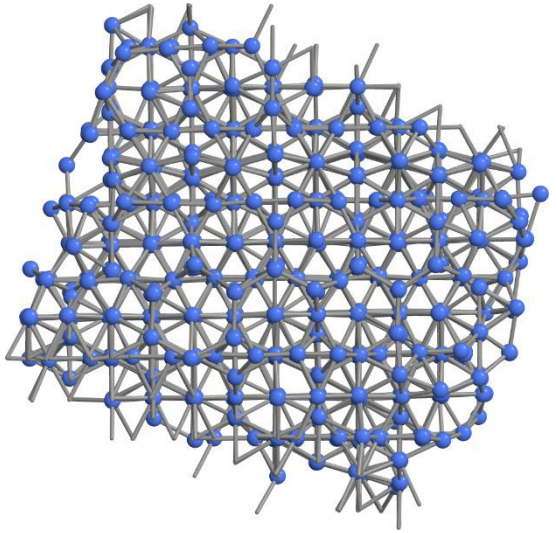
Jackson State University

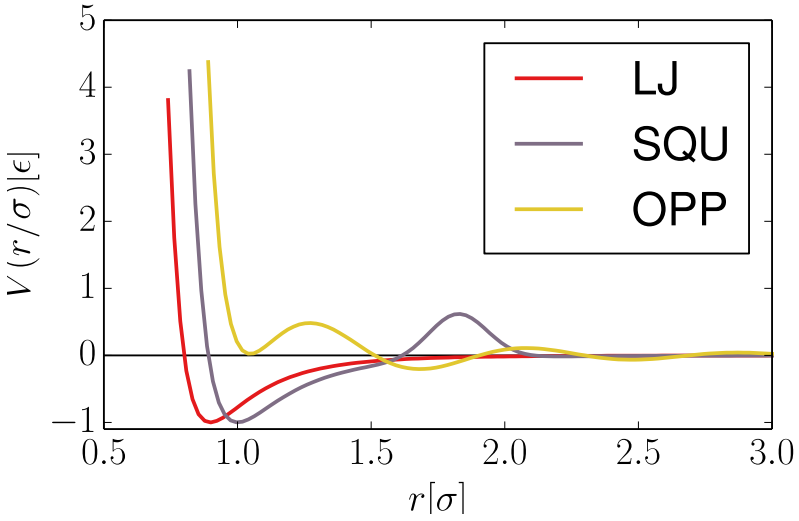
- Code performance optimization
- Github
- Jupyter notebook
- Do's and don'ts: Data management, workflow, reproducibility
- Look at results in R?

Carl Simon Adorf
University of Michigan



Damasceno, Engel, Glotzer, Science 2012



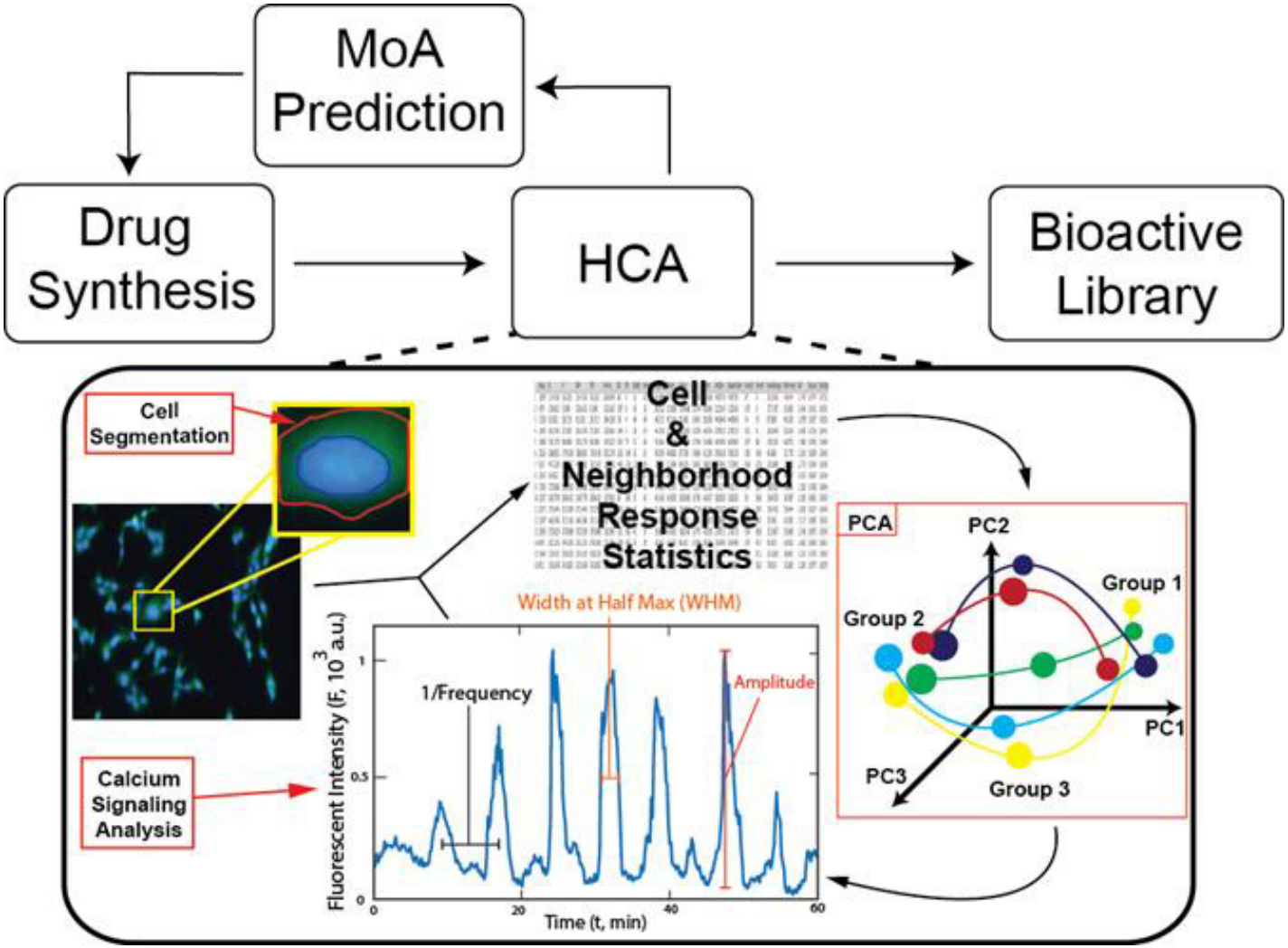


csadorf@umich.edu
http://glotzerlab.engin.umich.edu
http://signac.readthedocs.org

Carl Simon Adorf
University of Michigan

1. A better understanding of **storage hierarchies** and efficient use of in-memory file systems
2. Learned about **dask** and **ipythonparallel** for efficient parallelization of simple tasks in python
3. Learned about **numba** as another alternative for python optimization
4. A better understanding of visualization and will try to use **VisIt** for some of my visualization needs
5. Experiment with **R** and **singularity**

Pavel A Brodskiy
University of Notre Dame

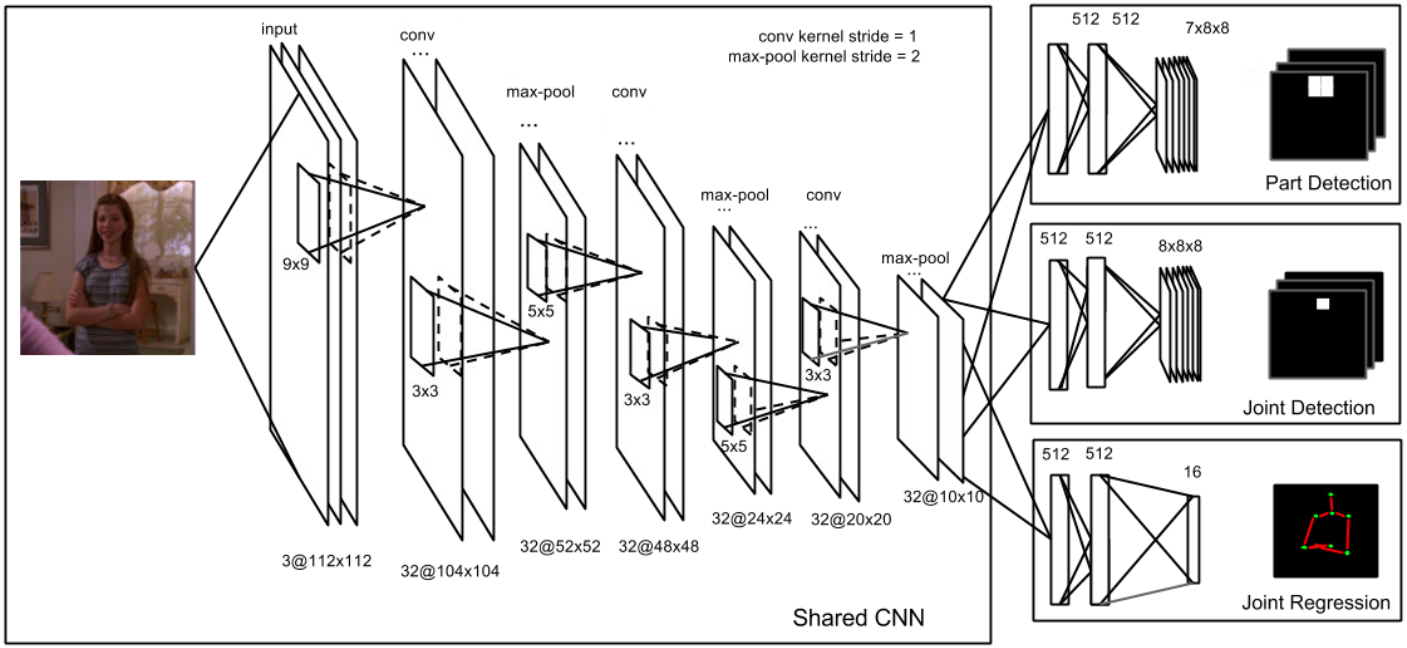


Pavel A Brodskiy

University of Notre Dame

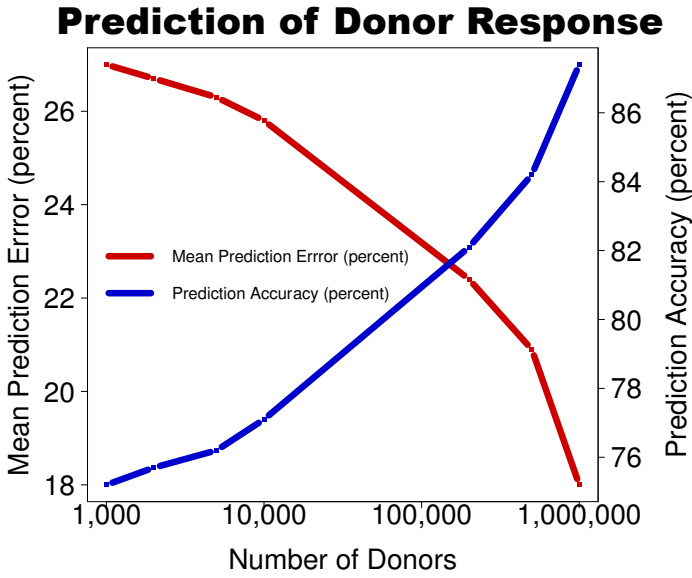
Running R and Visit on the cluster will greatly accelerate data exploration

I would also like to use Convolutional Neural Network algorithm to better segment cells than my current method



http://visal.cs.cityu.edu.hk/wp/wp-content/uploads/deep-pose-conv-r1-for-demo.png

Rico (Federico) Bumbaca
University of California - Irvine



Research: how consumers make choices

Context: big data (100s of million of consumers)

Model: Bayesian hierarchical multinomial logit

Methodology: Markov Chain Monte Carlo (MCMC) simulation

Problem: MCMC is a serial algorithm

Solution: Parallel MCMC algorithm for Bayesian hierarchical models

- R, Rmpi, and C++ on a single multicore computer (1 million consumers)

Rico (Federico) Bumbaca
University of California - Irvine

SDSC resources to scale to 100s of millions of consumers

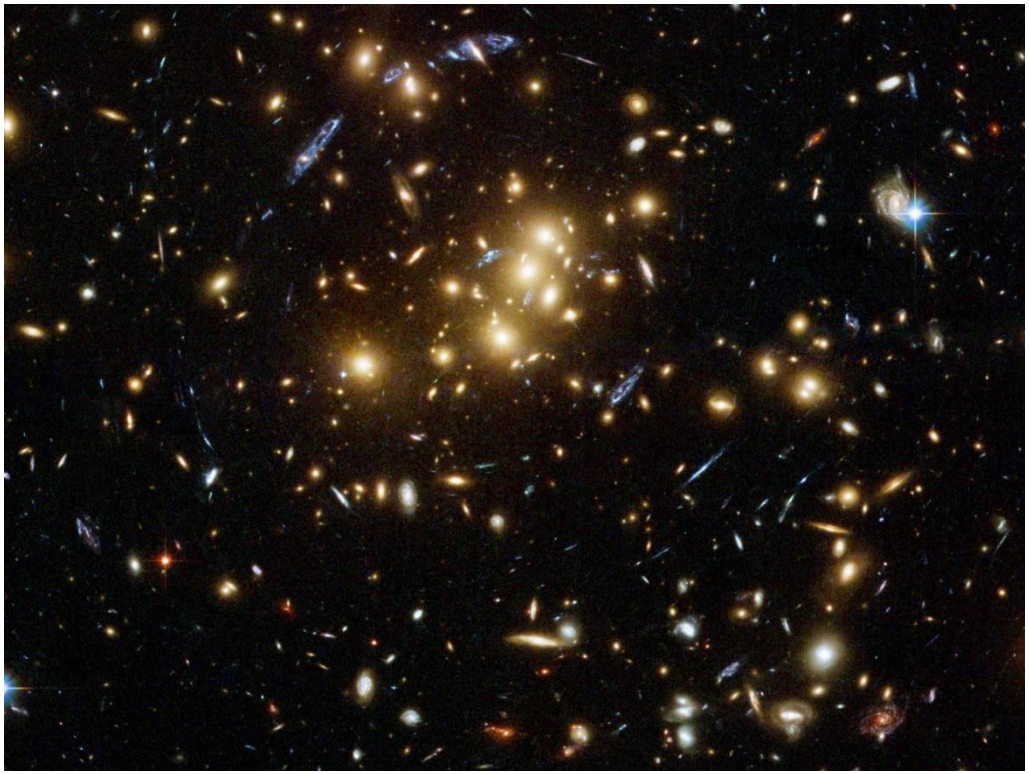
- Trial Allocation
  - Scale R implementation to multiple nodes
  - Port to C++ using hybrid MPI and OpenMP
  - Port to Spark
  - Port to GPUs

- Startup Allocation
  - Extended Collaborative Support Service (ECSS) program

Ryan Foltz
UC Riverside Department of Physics & Astronomy

Galaxy evolution in clusters:

- Data: 12-passband photometry for ~5000 galaxies in each of 10 clusters + spectroscopy
- Pipeline product: galaxy evolutionary models and related extracted parameters (origin, fate, etc.)
- Community tools (Fortran, IDL) and Python (7k+ lines of code in most recent repo)



The Question: What kills spiral galaxies?

How do galaxies form and evolve? How is stellar mass assembled and how does it come to resemble what we see today?

Ryan Foltz

UC Riverside Department of Physics & Astronomy

My pipeline + modeling code can take a whole day to run for a single galaxy cluster!

- Workflow (provenance, reproducibility) should replace my pipeline
- Processing can very much be parallelized
- IOPs can be sped up as well

New experiences:

R, Spark, running code on compute nodes, most of the machine learning.

Bill Fox

HGST

- Solutions Engineer
- Work with customers to improve performance on Genomic Sequencing - (Ilumina, Edico, Rady's Children Hospital)
- Genomic data is lots of small file sizes and low queue depth. Does not work well with SSD drives. Data writes are ranging 200MB – 1GB writes - normal writes 4GB.
- Currently use community code - R Language and Java
- Work with small data files that range hundreds of thousands to millions data sets.
- 2 Goals
  - Interaction of HDFS with Super Computers
  - Is multi threaded python a benefit

Bill Fox

HGST

- Work with SDSC team to further test genomic data with SSD to improve performance.
- Workflow and job management and how it ties in with Big Data and High Performance Compute.
- Implement python scripts to work with Hadoop and Spark and run map reduce jobs.
- Visualization tools (i.e. Visit) to share my findings with customers.
- Look at HDFS as being solution for genomic data and SSD, I will be researching how HDFS writes data to SSD.

Tao Gao

University of Delaware

My Project: Optimize and Extend MapReduce-MPI framework

MapReduce Framework

Various Implementations

Implemented by Java, provide Java or Python interfaces

MapReduce-MPI

Implemented by C++, provide C/C++ interfaces

Shortcommings of MR-MPI:
1.Don't use memory efficiently
2. Scalability Problem

Our work:

MapReduce-MPI++

Characteristics of MapReduce-MPI++
1. Provide C/C++ interfaces of MapReduce
2. Use Memory Efficiently
3. Better scalability

Email: taogao@udel.edu

Tao Gao
University of Delaware

I mainly use Comet to test and evaluate our MR-MPI++ framework with different benchmarks. Currently, we have three benchmarks: BFS, wordcount, octree clusting. We may implement more benchmarks or applications in the Future.

I leaned in this week about how to tune the performance and the knowledge about Spark help me to advance the project. Thanks!

Ed Hall

University of Virginia

Advanced Research Computing Services

I work as a computational research specialist, helping faculty and graduate students optimize and scale up their computational research to run on our local HPC cluster as well as on XSEDE resources.

I’m a Co-PI with an Economics faculty on an XSEDE XRAS allocation for running parallel Matlab code on Comet and Gordon at SDSC to simulate airline markets. The code uses simulated annealing to optimize how airlines achieve a stable market share.

I’m also working a faculty in the School of Medicine to use machine learning on audio signals of breathing recorded during sleep to classify whether someone suffers from sleep apnea.

Ed Hall

University of Virginia

Advanced Research Computing Services

I plan to continue running parallel Matlab code on Gordon and Comet and to use what I learned about visualization at the summer institute to better understand how the optimization objective function converges.

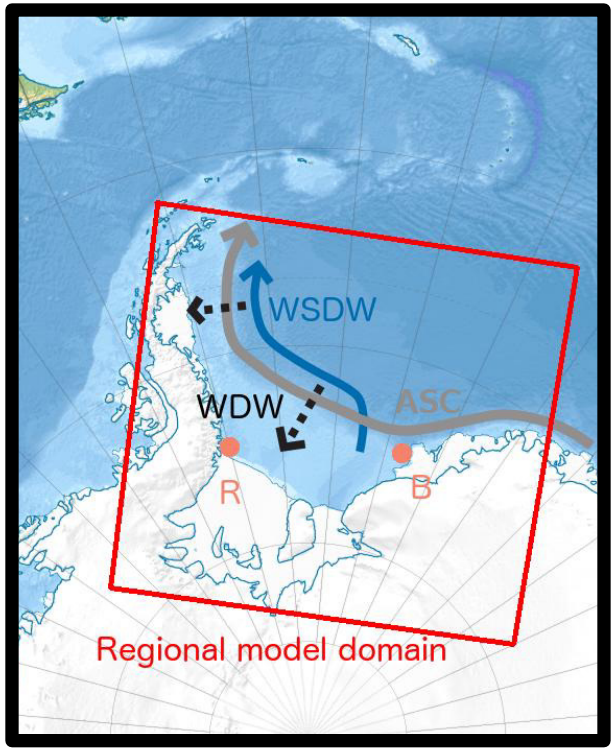
I plan to use what I learned about machine learning to write the Matlab code for the sleep apnea classification project.

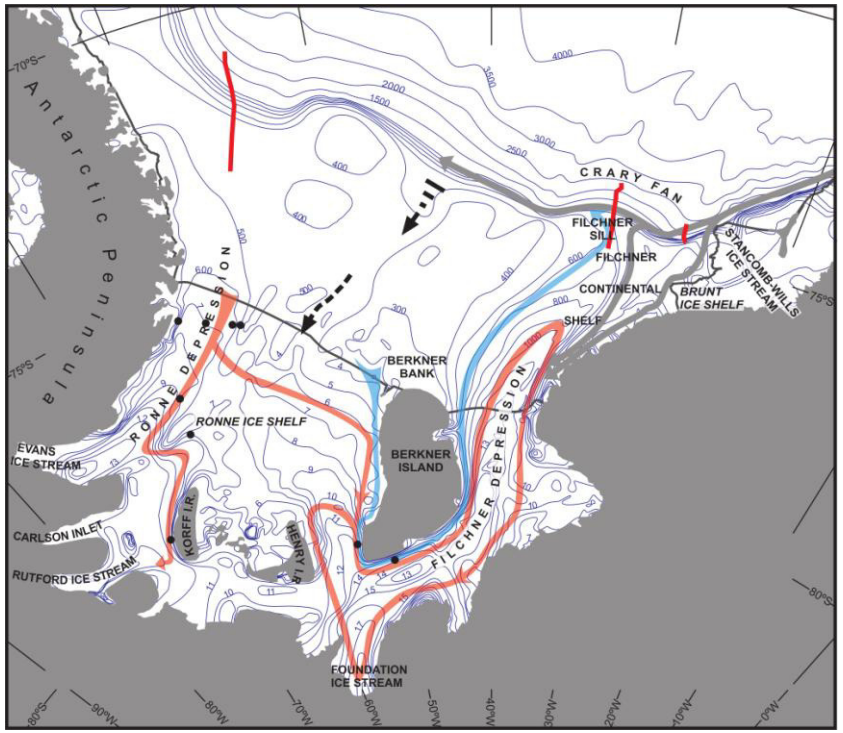
I also plan to learn how to use Python for high performance computing and to start using git for code version control and documentation.

Lastly, learning how to use the Linux top command for threaded applications will be very helpful for code development.

Julia Hazel

UCLA (Atmospheric and Oceanic Sciences)





- Interested in studied in near-Antarctic ocean circulation
- Developing a regional model of the Weddell Sea, the largest established source of Antarctic Bottom Water (AABW)
- Model incorporates the full Filchner-Ronne cavity and simulates the thermodynamics and dynamics of sea ice
- Goal is to refine model to 1/24 degree resolution to capture high-frequency processes (tidal flows, eddys, downslope gravity currents) in the transport/transformation of AABW
- Running the model on Gordon!

Julia Hazel
UCLA (Atmospheric and Oceanic Sciences)

- Making effective use of Gordon = essential for fine-tuning my model to high-resolution
  - Amdahl's Law
  - Will soon need to provide XSEDE with scaling to show model run time improvement as processors are increased (to get more hours)
- Better understanding of Gordon and its internal 'fabric' and 16-compute node structure
  - General understanding of HPC
- Made a github account
  - Will use for future software management Virtualization session and
- learning Viz software
  - 3D imaging
- New but eager to learn Python -> Python for HPC useful

Thakshila Herath

Georgia State University, Atlanta, GA

Hodgkin-Huxley style neuronal model

- $C\frac{dV}{dt} = -\left[g_{Na}m_{Na,\infty}(V)^3 h_{Na}[V - E_{Na}] + g_{K2}m_{K2}{}^2[V - E_K] + g_h m_h{}^2[V - E_h] + g_{leak}[V - E_{leak}] + 0.006\right]$,
- $\frac{dm_{K2}}{dt} = \frac{[m_{K2,\infty}(V,\theta_{K2}) - m_{K2}]}{2}$,
- $\frac{dh_{Na}}{dt} = \left[\frac{1}{1+\exp(500[V+0.0325])} - h_{Na}\right]/0.0405$,
- $\frac{dm_h}{dt} = \left[\frac{}{1+2\ ex}\right.$

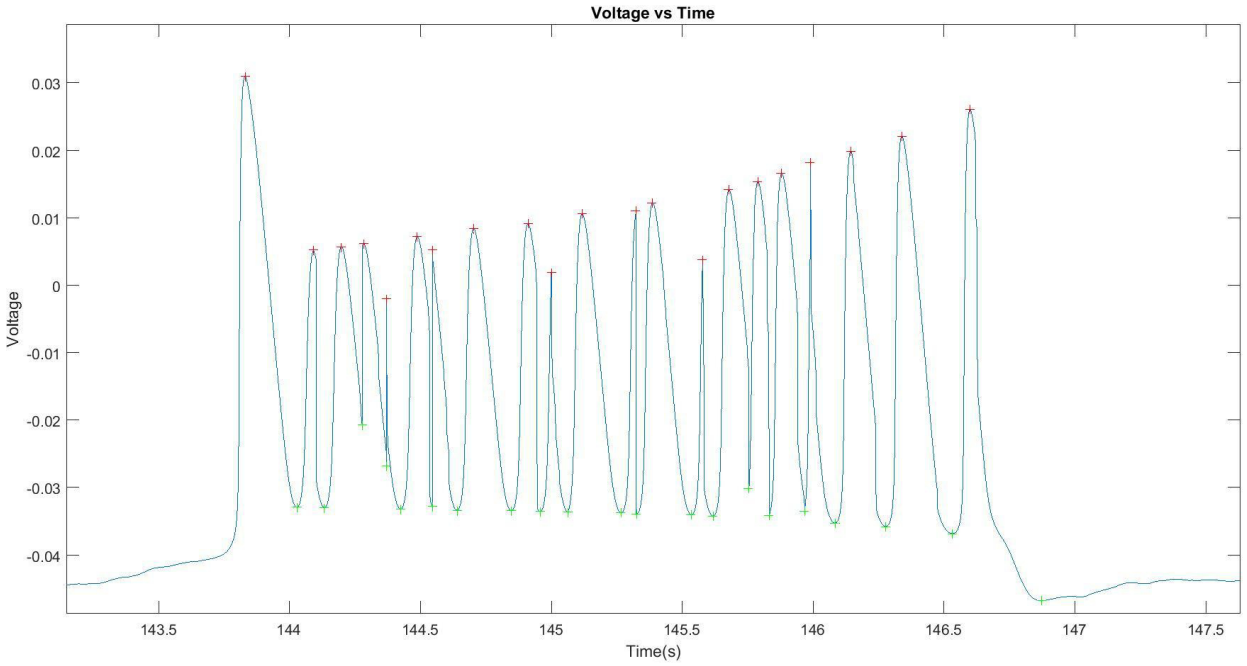
where

$m_{K2,\infty}(V,\theta_{K2}) =$

$m_{Na,\infty}(V) = \frac{}{1+}$

The inactivatic

The activation



1. Barnett WH, Cymbalyuk Controlling Endogenous Responses of a Neuron M
2. http://hyperphysics.phy-a

Thakshila Herath

Georgia State University, Atlanta, GA

- Code performance using optimization techniques
- Effective use of the machine
- Experience in GitHub
- Parallel computing using MPI and Open MP
- Python for HPC
- Good start to learn C and Python

Erin Hodgess
University of Houston - Downtown

- I am a statistician; areas of interest in time series, statistical computing, and geostatistics
- Have been using supercomputing for several years
- Am the “Campus Champion at UHD.
- I do develop a lot of my own code, use R, Fortran, MPI.
- Have been working on Ordinary kriging on Stampede on spatial data
- Would like to expand this to spatio-temporal data sets

Erin Hodgess

University of Houston - Downtown

- So many new topics!
- I got on my Bridges account last night (first time); took one of my Fortran programs. Ran it normally, it took 0.13 seconds. Used the Open ACC directives, it ran in 0.035 seconds! Not too shabby!
- We started a new MS program in Data Analytics in Fall 2016 and I want to build a capstone course including supercomputing, with some machine learning. What I learned here will be a great help.
- The performance optimization session will also be a great help since I do a considerable amount of my own coding (as well as others).

Jackie Huband
University of Virginia

I am a Computational Research Support Specialist within the Advanced Research Computing Services (ARCS) group at University of Virginia.

The role of ARCS is to **advance the culture of computing** among our faculty, staff, and students. So, I

- Provide training sessions to teach our researchers how to write/optimize/parallelize codes/programs;
- Assist with moving our researchers off of their laptops onto our cluster;
- Oversee & encourage use of the Visualization Lab.

My current goal is to fill a void that we currently have within Big Data Analysis.

Jackie Huband

University of Virginia

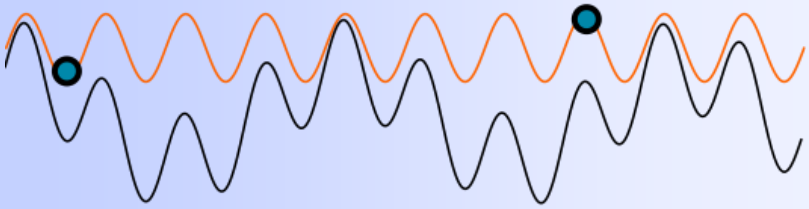
First, this has been an amazing experience. I started working with HPC only four years ago, and have not had a broad exposure to how it all should flow together.

- I am going to learn more about the storage options on the cluster at U.Va., so that I can advise users on how to optimize their I/O.
- I am going to learn more about workflow tools, so that grad students can leave behind a better record of their efforts.
- I am going to use version control when I present at workshops, so that I can set an example for others.
- I am going to work harder to get Spark installed on our cluster, so that all researchers can analyze their big data.

Katharine Hyatt - UC Santa Barbara

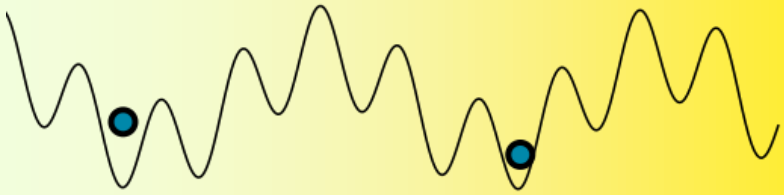
Thermalization

Most quantum systems thermalize
No memory of local initial conditions
Entanglement spreads quickly
Extended eigenstates
Interesting physics is fragile



…and its Discontents

Equilibrium stat mech fails
Long memory of local initial conditions
Entanglement spreads slowly
Localized eigenstates
Protect interesting exotic physics



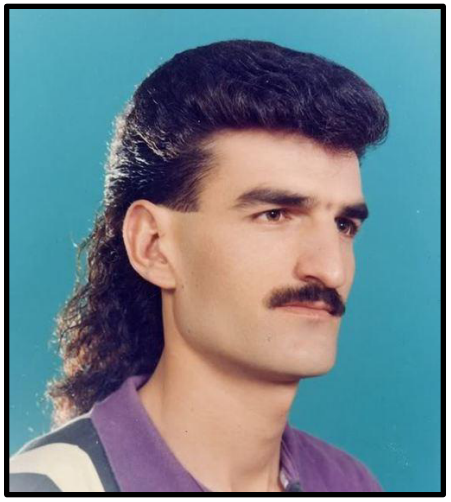
Mid-spectrum eigensolver for **lots** ($n = 2000$) of **big** ($N = 2^{24}$) matrices

Use PETSc and SLEPc to solve, make measurements

Using Julia to manage workflow, do analysis
Coming soon- native Julia parallel eigensolving.

Katharine Hyatt - UC Santa Barbara

- Harder problems / bigger systems with SDSC
- Accessibility for collaborators:
  - Jupyter in the front
  - Julia/PETSc code in the back
  - Test & share code and results
- Use storage hierarchy better
- PETSc has GPU “support”



Ask very nice & smart people for help and advice! :)

Mark Jack

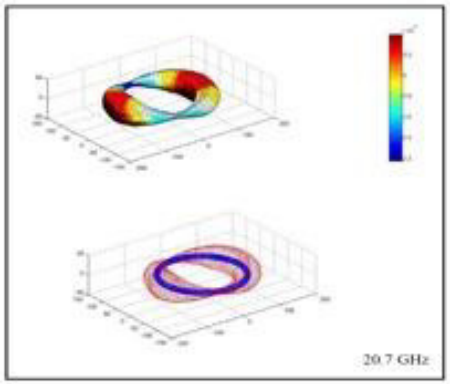
Florida A&M University, Physics

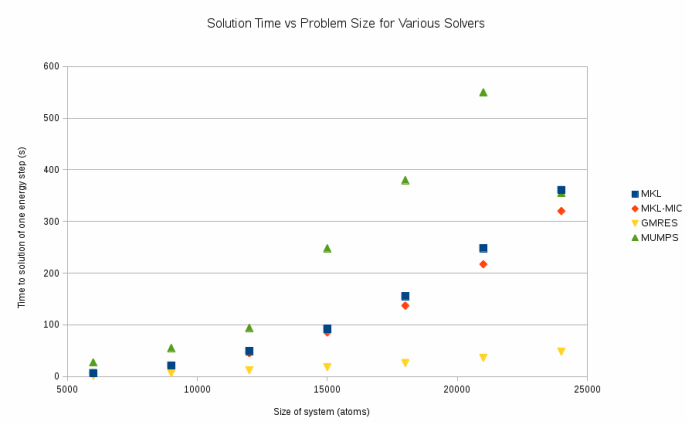
Research: Quantum transport in carbon nanostructures with electron-phonon coupling for low energy phonon modes.

Project: QRing – A scalable parallel software tool for quantum transport simulations in carbon nanoring devices based on NEGF formalism and a parallel C++ / MPI / PETSc algorithm.

Resources: TACC Stampede (XRAC), ORNL Titan (DD), SSERCA (FSU, U Miami).

New directions: OPVs, Materialsproject.org.





Mark Jack

Florida A&M University, Physics

Goal:

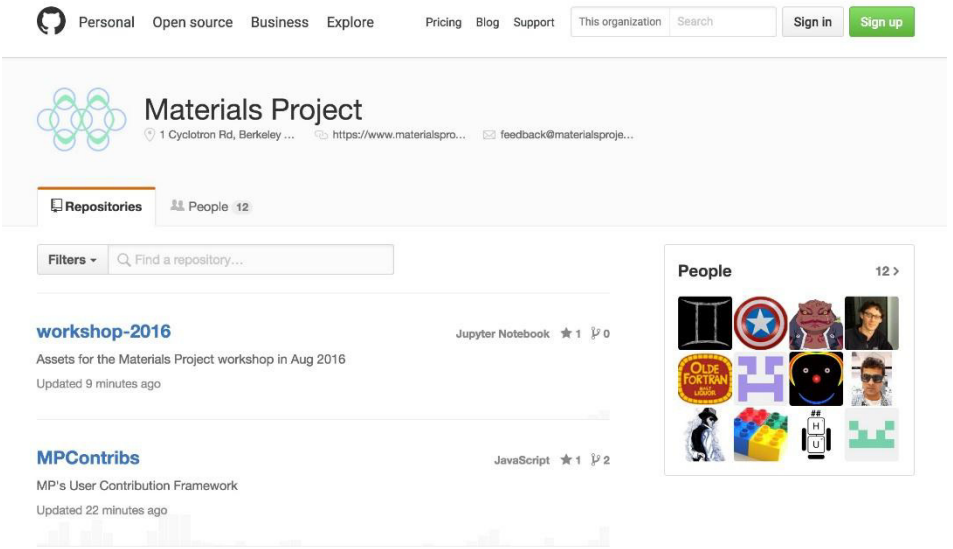
Computational model for Organic Solar Cell materials developed by S. Foo (FAMU-FSU EE) via *Python Materials Genie.*[*http://matgenie.materialsvirtuallab.org*](http://matgenie.materialsvirtuallab.org)

SDSC SI2016 Lesson:

For Big Data and ML, use *Spark* with *iPython Notebook* with *scikit_learn* ML library rather then Spark MLLIB.

Strategy:

Integration of QT and DFT modeling for electronic transport and excitations with '*pymatgen'.*



Resources:

- Trial and startup allocations on Comet (fall 2016). XRAC on Comet (spring 2016). Amazon cloud (fall 2016).
- Application for NSF proposal (Oct '16).

Blake Joyce, Science Informatician Team
CyVerse, BIO5 Institute, University of Arizona

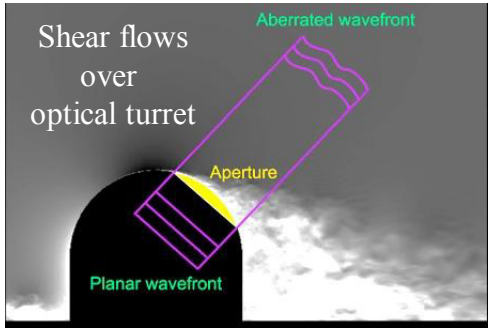
- CyVerse is an NSF-funded, academic group
- I build (free) computational infrastructure for scientists
- My scientific dream:
  - Result reproducibility
    - Provenance of data
    - Traceable analytical metadata
    - Versioning of software
  - Validation of methods
    - Automated parameter testing
    - Analysis of parameter effects and variability
  - Democratization of tools
    - Peer-reviewed workflows
    - Broad scale benchmarking on real experimental data
  - Fungible computational resources

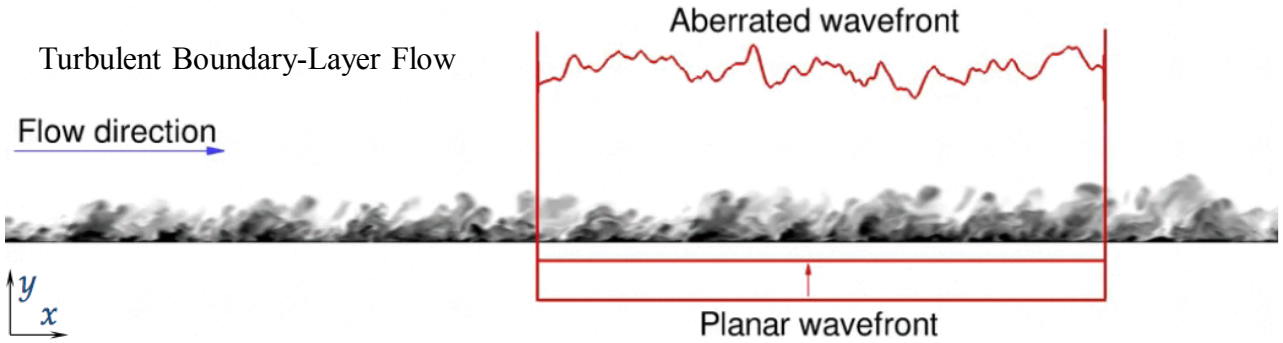
Science Dream	CI to Leverage
Result reproducibility	**Result reproducibility**
Provenance of data	Spark for tracking flow of analyses?
Traceable analytical metadata	Cyberinfrastructure, Spark?
Versioning of software	GitHub versioning (automatic ping?)
Validation of methods	**Validation of methods**
Automated parameter testing	Parallelization of analysis (MPI/Spark)
Analysis of parameter effects	Machine learning
Democratization of tools	**Democratization of tools**
Peer-reviewed workflows	Peer-reviewed workflows
Real data benchmarking	Real data benchmarking
Prediction of comp resources	Machine learning
Fungible comp resources	**Fungible comp resources**
Free movement between platforms	Singularity, Docker, etc
GPU<->CPU?	OpenACC? TensorFlow?

Mohammed Kamel

Graduate Research Assistant (www.nd.edu/~mkamel)

UNIVERSITY OF NOTRE DAME
College of Engineering





- **Aero-optical problem:** Optical distortions due to fluctuating index-of-refraction and density fields of compressible turbulent flows adjacent to an optical aperture
- The current research is aimed at improving the predictive capability for understanding of turbulent flows and the associated aero-optics at realistic flight conditions (high Reynolds numbers, and a wide range of Mach numbers)
- **Wall-modeled Large-Eddy Simulation** (LES) is the employed approach for flow simulations.
- Flow simulations are carried out by **CharLES** code developed at Cascade Technologies Inc. (Khalighi et al. [AIAA-paper 2011-2886])
  - Unstructured finite volume method for spatial discretization
  - Third-order Runge-Kutta for time marching
  - Parallel implementation using MPI
- **Wall-modeled LES shows accurate aero-optical predictions for subsonic & supersonic turbulent boundary layers and subsonic flows over cylindrical turrets (Kamel et al. [AIAA-paper 2016-1462])**

Mohammed Kamel

Graduate Research Assistant (www.nd.edu/~mkamel)

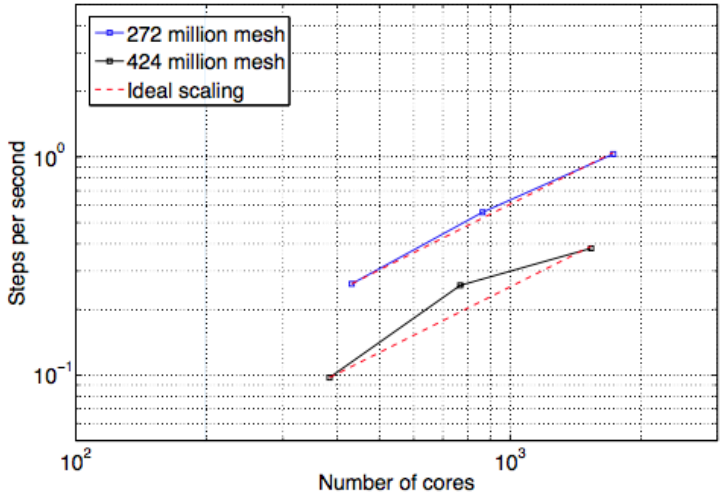
UNIVERSITY OF NOTRE DAME
College of Engineering

How to use what is learned at the summer institute in my current research?

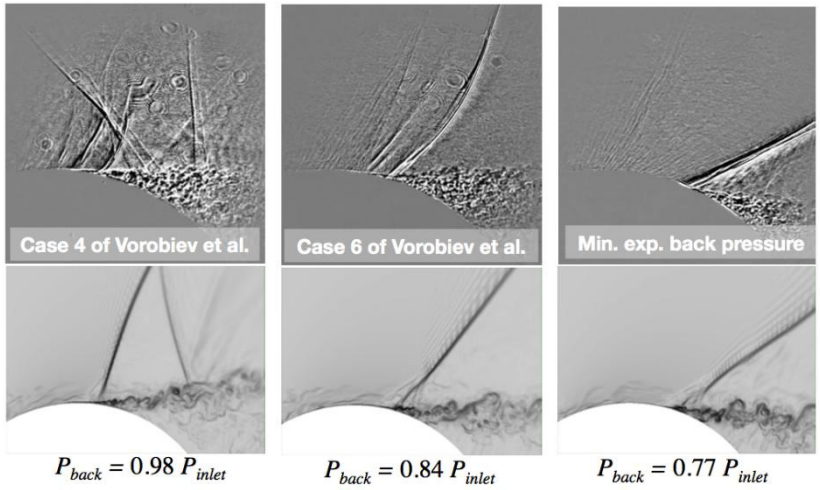
- Managing jobs and data on comet's system
- Data exploration and visualization using VisIt
- Software version management
- Code profiling and optimization
- etc …

How it is expected to use SDSC resources?

- Application to XSEDE research allocation for the next quarter (Oct. 2016) on Comet (PI: Dr. Meng Wang)
- My role in the research proposal is to extend the shear-flow simulations to transonic flow regimes.



Parallel performance of LES code CharLES on Comet at SDSC



Qualitative agreement between the experimental schlieren (Vorobiev et al.) and the numerical Schlieren of the preliminary simulations at different back-pressures

Mohammed Kamel
Graduate Research Assistant (www.nd.edu/~mkamel)
UNIVERSITY OF NOTRE DAME
College of Engineering

Acknowledgments

- Advisor: Dr. Meng Wang
- This research is sponsored by the High Energy Laser Joint Technology Office (HEL-JTO) through AFSOR Grant FA950-13-1-0001 as part of the “Airborne Aero-Optics Laboratory – Transonic” program
- The current available computational resources are provided by:
  - University of Notre Dame Center for Research Computing (CRC)
  - Startup allocation on Comet

Haiqing Li
Beckman Research Institute, City of Hope

- Next generation genomic sequence data analysis for cancer related basic and translational studies
  - Genome-wide genetic (such as DNAseq, Exomeseq, RNAseq, miRNAseq) and epigenetic data (such as ChIPseq, Methly-seq)
- Research informatics infrastructure – “Cyberinfrastructure”
  - High throughput instruments ←→ High performance computing resources ←→ Scientific software virtualization
- My current research projects include tumor stroma analysis using RNASeq, T-cell repertoires analysis using TCR-seq, Triple Negative Breast Cancer recurrence risk factors analysis using Machine Learning, Cancer microenvironment study using EM (Electron Microscope) data and Convolution Neural Network (CNN), crowdsourced OCR correction for clinical narrative, integration i2b2 and tranSMART for translational research.

Haiqing Li

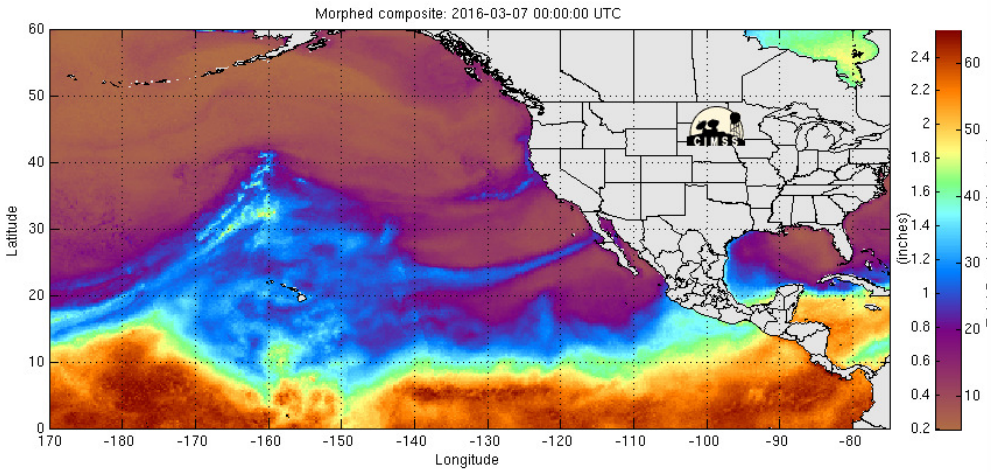
Beckman Research Institute, City of Hope

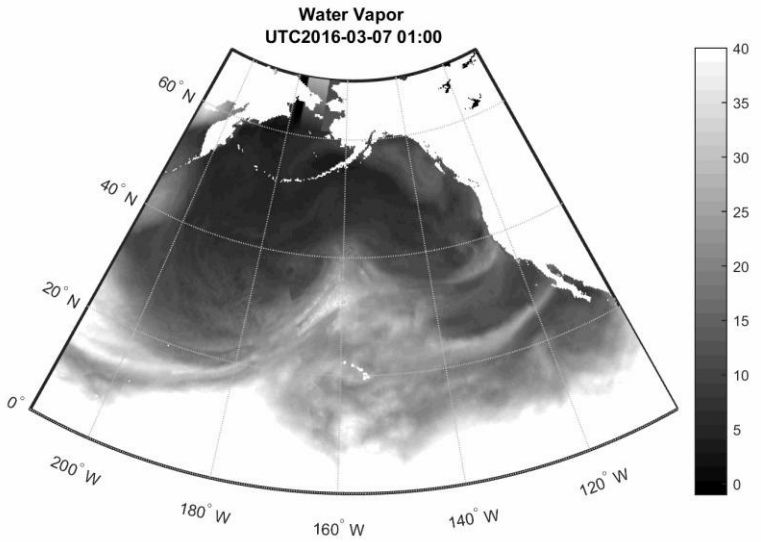
- NGS analysis
  - From primary analysis to downstream analysis
- Image analysis
  - Segmentation and feature extraction using CNN
- Scientific Workflow

Hao Liu,
University of California, Irvine

Result: Interpolation of the Satellite IWV data

New interpolation method overcomes the missing data problem, and is much smoother.





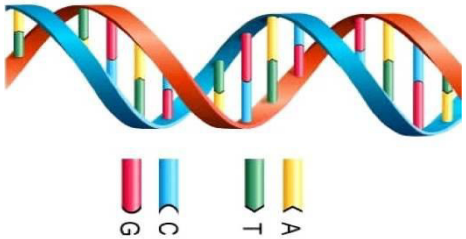
- **I learned a lot to kick-start my work on COMET**
  - **1. GPU programming**
  - **2. Parallel programming using Python (currently using MATLAB but has license issue)**
  - **3. Workflow optimization on limited storage allocation**
- **4. GitHub**

Aalto University

Altti Ilari Maarala

ilari.maarala@aalto.fi

Aalto University, Helsinki, Finland
Department of Computer Science



- Research focus on computational genomics – processing NGS data in parallel on distributed systems. Previous research on IoT, covering intelligent traffic and digital health systems.
- Parallelizing of pipelines and algorithms for bioinformatics – enabling analytics of Pan-Genomes in parallel.
- Pan-Genomics pipeline comprises assembling of consensus genome from multiple individual genomes which is used as a new reference genome for further analysis and variant calling => enables large-scale comparative studies on genomes and phenotypes efficiently, also population-wise.
- Hadoop/Spark used for coarse parallelization => fine grained multithreaded parallelization e.g with MPI, OpenMP needed in some parts of pipelines.

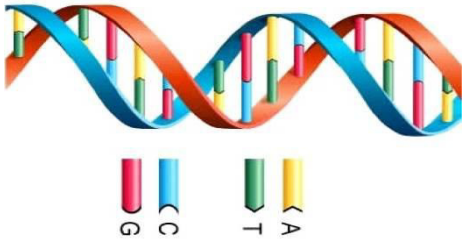
Computational Pan-Genomics: Status, Promises and Challenges, Marschall et al.
http://biorxiv.org/content/biorxiv/early/2016/03/29/043430.full.pdf

Aalto University

Altti Ilari Maarala

ilari.maarala@aalto.fi

Aalto University, Helsinki, Finland
Department of Computer Science



- **Hadoop-BAM**, Java library for manipulating common file formats used in bioinformatics with Hadoop and Spark. Used e.g. in GATK4, Adam, Halvade.
- **SeqPig**, Apache Pig library for analysis of large sequencing datasets stored e.g. to HDFS as BAM files with SQL style queries. New release coming soon with current Hadoop-BAM version.
- https://github.com/HadoopGenomics/
- **SeqSpork**, enables running of SeqPig sequence analysis scripts as Spark jobs https://github.com/ridvandongelci/SeqPig/tree/spork-1.2.0

- Interested in topics of SDSC Institute: HPC management, MPI, GPU programming, OpenACC, Python for HPC, CIPHER Science gateway.

- Currently using HPC resources of Finnish IT Center for Science (www.csc.fi).

- CSC has developed CHIPSTER http://chipster.csc.fi/ containing ~350 tools for NGS data and probably interested on developing Science gateway for bioinformatics.

Interested in collaboration?

Sumukh Sagar Manjunath
San Diego State University

- Research Topic
  - Ocean Science Education Portal
  - Web Interface to Host General Curvilinear Coastal Ocean Model (GCCOM)
  - Meter-scale, fully 3D curvilinear, non-hydrostatic, large eddy simulation (LES) model
  - Implementation of Drop-a-Drifter Water Trajectories over the coast.
- Complexity of GCCOM
  - 16-core Xeon nodes each w/ 64GB RAM.

		6 hours simulation / 10 minutes Assimilation			
nodes:processes	Ensemble Size	Wall-clock time (hours)	Output Size	Total RMSE	Total Spread
2:15	30	6.40	3.93 GB x 2	0.69536	0.56466
4:15	60	7.40	7.38 GB x 2	0.68237	0.56336
6:15	90	8.10	10.84 GB x 2	0.6779	0.56255

- Challenges
  - Data storage and retrieval
  - Analyzing Data
  - Data Visualization
  - Compute and I/O Parallelization
- Website: http://sci.sdsu.edu/csrc-cod/

Sumukh Sagar Manjunath San Diego State University

- Introduction of a powerful Viz Tool like Visit
- Some insights on using spark for our data analysis
- Ease of Using Gateway Portals

Dominique Meroux

UC Davis Institute of Transportation Studies (ITS)

Research focuses on potential (macro-level) risks and benefits of 3 revolutions in transportation: automation, electrification, and shared use mobility

Big data problems I'm interested in

- Realtime API data, e.g. Zipcar and other carshare, bikeshare
  - Answer questions like: what are the most commonly used carshare vehicles (e.g. hybrid / efficient, or SUVs), how are carshare and bikeshare used?
- Optimizations Using Streaming Vehicle Data
  - Optimal routing, especially for future autonomous shared ride services
  - Optimal location of electric charging and hydrogen stations

Dominique Meroux

UC Davis Institution of Transportation Studies (ITS)

Opportunities and benefits for taking advantage of big data sources to inform transportation research are present and growing.

I hope to apply what I've learned to help Davis ITS utilize these opportunities more

•Davis ITS Research informs state policy and industry actions; suboptimal decisions could have a large impacts on air quality

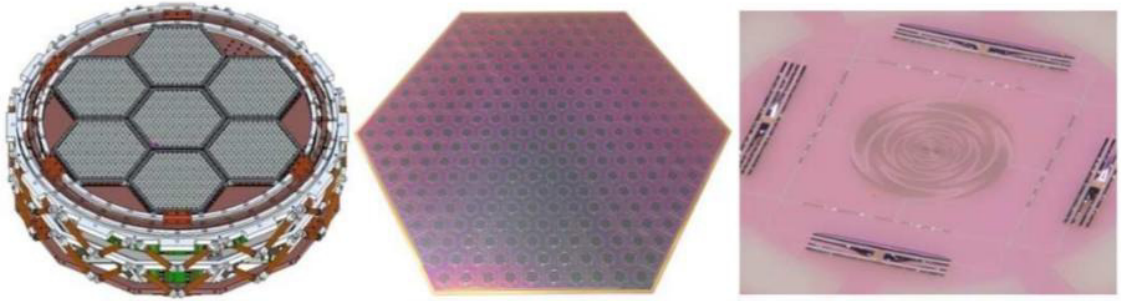
•E.g. one project at Davis ITS streams vehicle data to examine how plug-in vehicles are driven

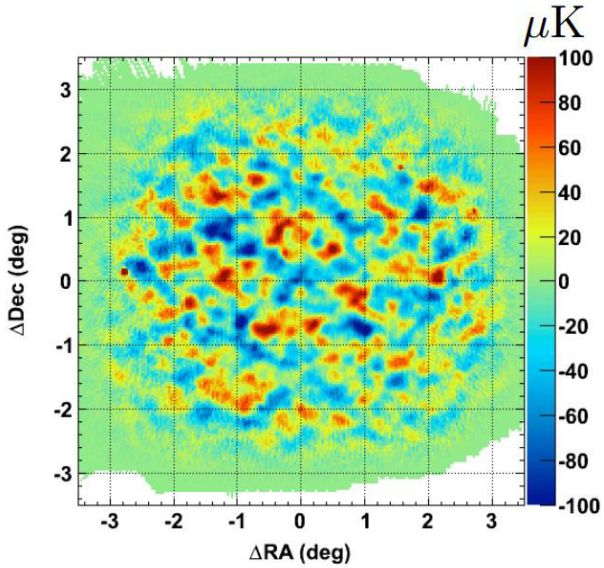
- A survey approach could have been taken, but would have involved human error, potentially low response, etc.

Marty Navaroli
UCSD CASS Cosmology Group

- **POLARBEAR-2/Simons Array**
  - **Atacama Desert, Chile (5200 m)**
- **Searching for signature of B-modes in early universe in Cosmic Microwave Background (CMB)**
- **Use NERSC for timestream data analysis (very large matrices!)**
- **Cryogenics and calibration**

7,588 total bolometers.



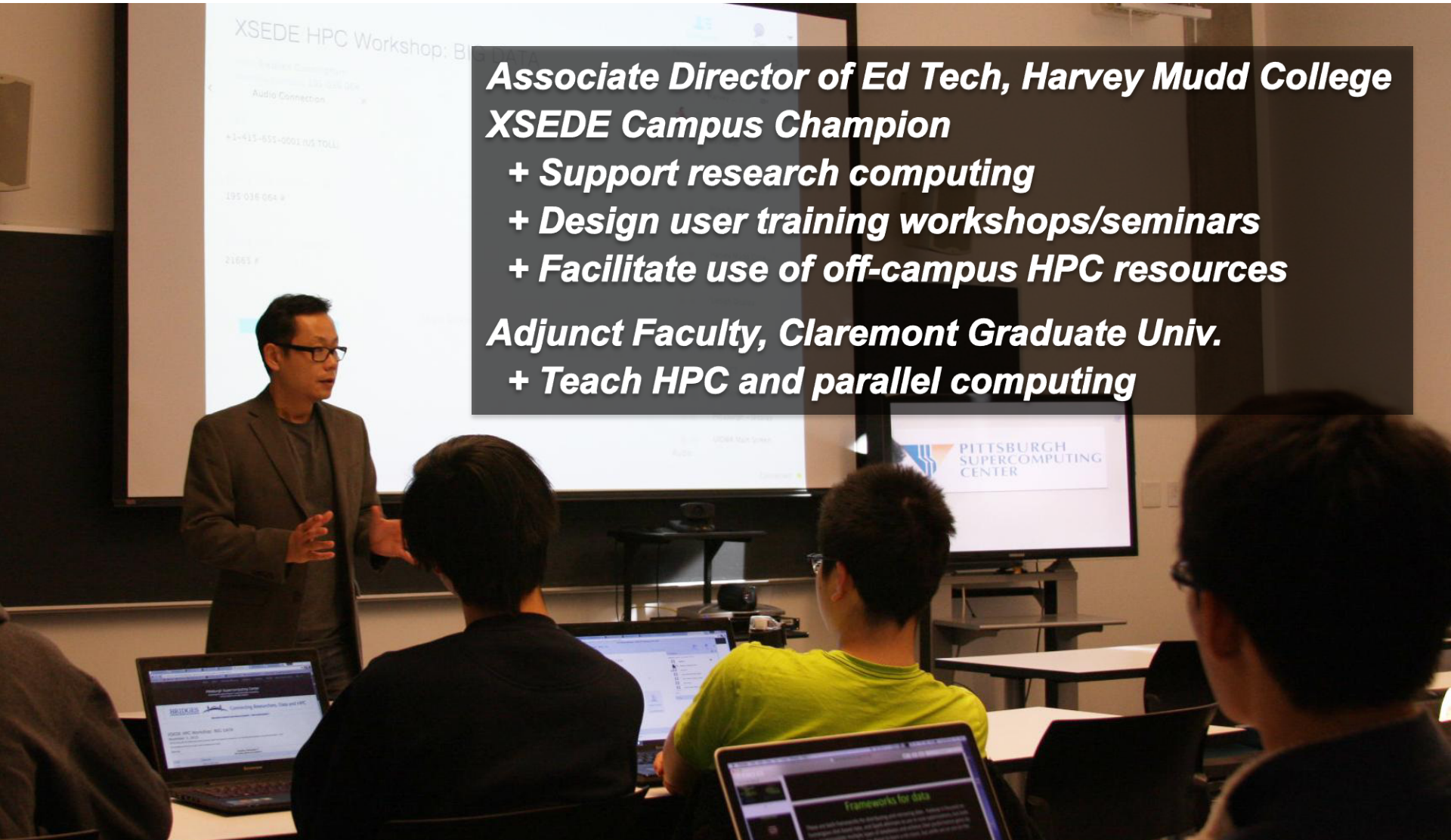


Marty Navaroli

UCSD CASS Cosmology Group

- **Use SDSC for timestream -> temperature map data analysis (along with NERSC)**
- **Talk with group about potential for science gateways (along with potential for undergrad projects)**
- **Virtualization (temperature maps!)**
- **I-python notebooks for quick analysis jobs**

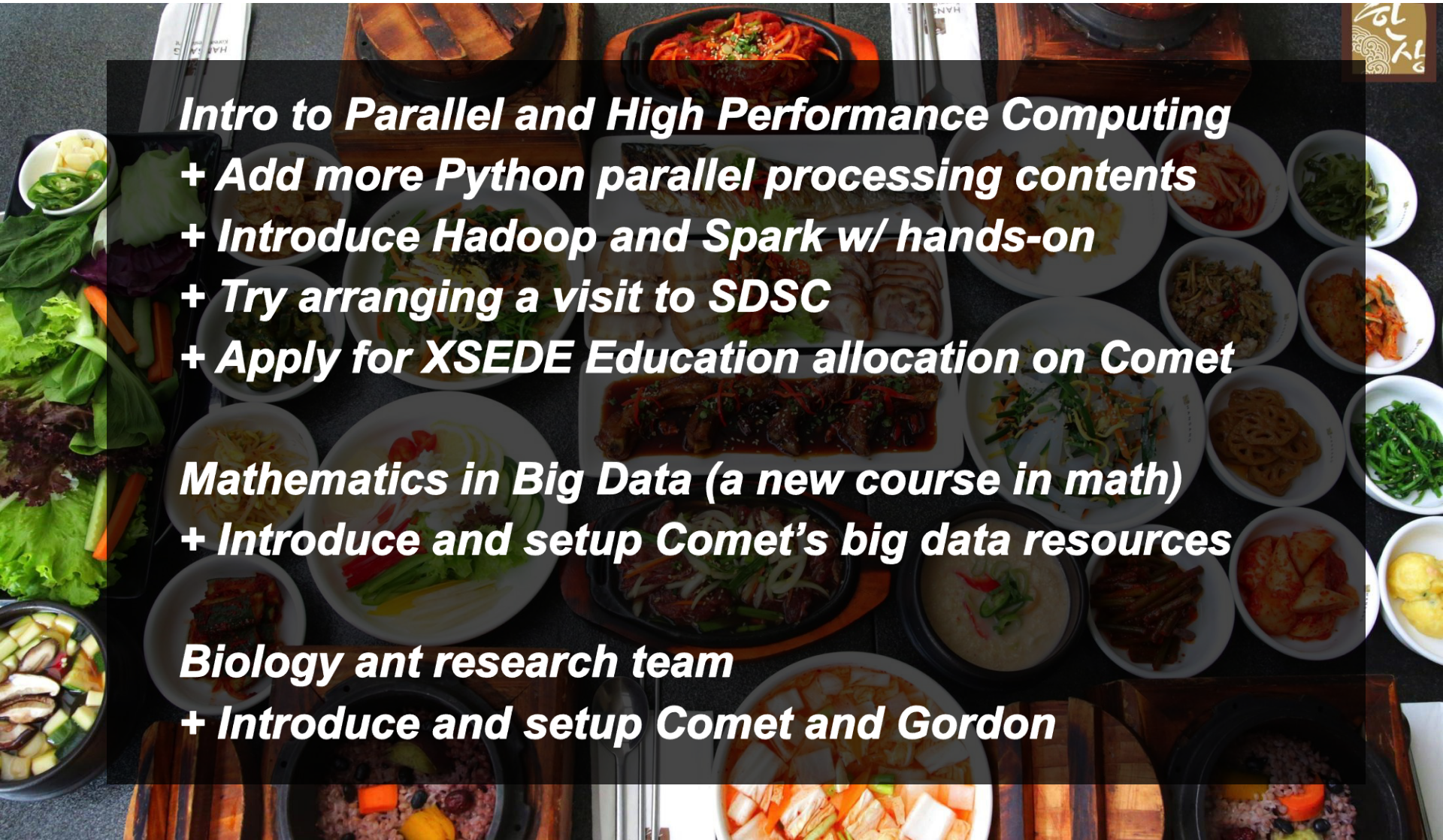
Jeho Park (jepark@hmc.edu)
Harvey Mudd College



Associate Director of Ed Tech, Harvey Mudd College
XSEDE Campus Champion
+ Support research computing
+ Design user training workshops/seminars
+ Facilitate use of off-campus HPC resources

Adjunct Faculty, Claremont Graduate Univ.
+ Teach HPC and parallel computing

Jeho Park (jepark@hmc.edu)
Harvey Mudd College



Intro to Parallel and High Performance Computing
+ Add more Python parallel processing contents
+ Introduce Hadoop and Spark w/ hands-on
+ Try arranging a visit to SDSC
+ Apply for XSEDE Education allocation on Comet

Mathematics in Big Data (a new course in math)
+ Introduce and setup Comet's big data resources

Biology ant research team
+ Introduce and setup Comet and Gordon

Mark Piercy
Stanford University

I work at the Stanford Research Computing Center, helping researchers get up and running with HPC on our main 800 node cluster; Sherlock.

Roles

- Help current users best utilize our resources.
- Work with a diverse user group, some people new to HPC and some who are very experienced.
- Work with users to scale their workflows up from desktop to the cluster.

Needs

- The most common tools/areas; R, Matlab, python, STATA, bioinformatics, molecular dynamics, machine learning, parallelizing jobs, estimating job resources requirements.
- Increasing demand for graphics based tools to interact with the cluster, we use x2go, NoVNC
- Increasing number of users from economics, social sciences and neurosciences that are new to Linux/HPC

Mark Piercy
Stanford University

At the Summer Institute I received a great introduction to many of the technologies and fields that our users work with.

We have many users that need support with Jupyter, Spark, R, Machine Learning, visualization, containers and many others tools.

My group helps users with every subject that was covered in the workshop.

I also was able to compare and contrast how my group does HPC versus SDSC, and that was extremely beneficial.

I now feel much more capable of handling more advanced issues users encounter

My ability to make strategic decisions on what tools our group should or should not use has increased.

More prepared to investigate emerging HPC tools and technologies

Carson Miller Rigoli
Cognitive Science, UCSD

How and why do we move in time as we do?

Statistical models:

(what can we learn from music?)

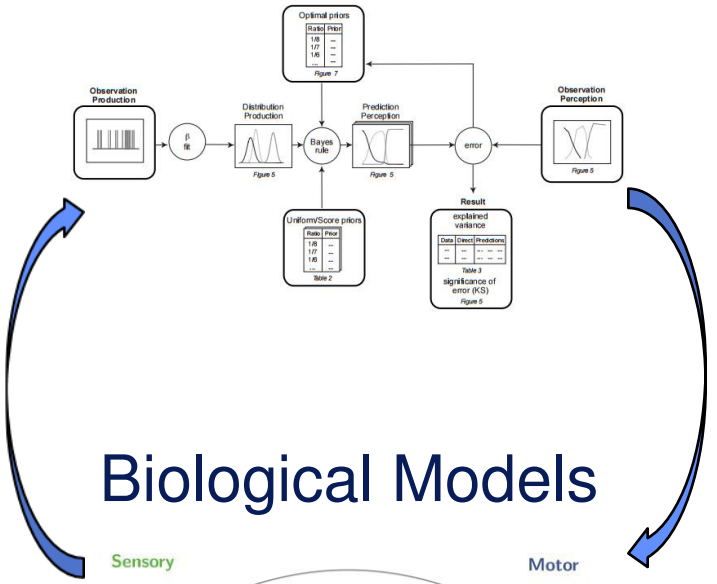
Neural models:

(what is a group of neurons actually capable of?)

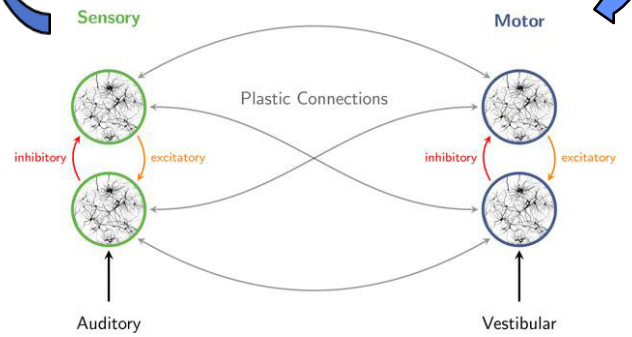
My work:

Bridge the explanatory gap using corpora of real world stimuli to train biological plausible models.

Statistical Models



Biological Models



Carson Miller Rigoli
Cognitive Science, UCSD

HPC v HTC

Different tasks benefit from different systems

Sebastian Reyes Chin Wo
UC Davis

Analysis of genomic data to develop tools and resources for plant breeding

Use a mix of already available bioinformatic tools and development of in-house code

Some tools already support multi-threading, OpenMP or MPI

Mix between high memory and computing demanding jobs

Important growth in the amount of data that is reaching the computational capacity

Sebastian Reyes Chin Wo
UC Davis

Access to more computing resources through SDSC (XSEDE)

- Significant number of tools already available as modules on comet

Optimization of in-house code to increase the throughput

- SPARK
- Parallelizing of Python

Utilization of Singularity for installation of complex software

Matt Settles

UC Davis

Manage UC Davis Bioinformatics Core in the Genome Center [Data Analysis, HPC, and Training]

In the past have mostly worked on BAS, moving work to cluster [slurm] environment.

Conduct data analysis on all types of genomic data

Research is in

- Noise reduction, processing raw data
- Genomics as data science, from sample to interpretation

Matt Settles
UC Davis

Topics found interesting

Spark, Python HPC and Machine Learning

Computational Workshop style, since I give workshops as well

Information on how I can turn those, where our HPC is not the best choice, on to SDSC for resources. Need to discuss further

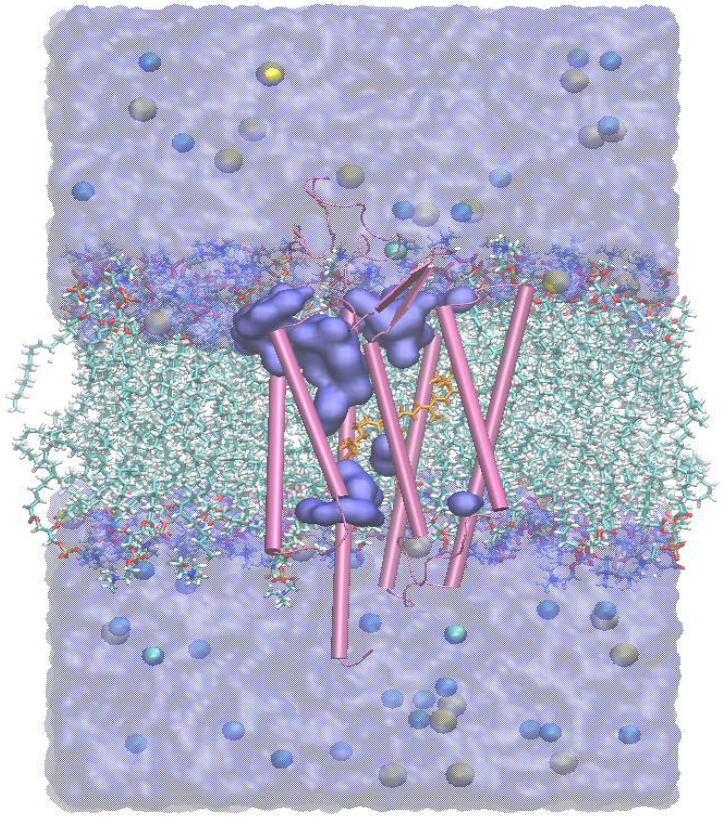
THE UNIVERSITY of NEW ORLEANS

Monika VanGordon

University of New Orleans

Sandia National Laboratories

Molecular dynamics simulations of transmembrane proteins



Snapshot of equilibrated, closed-state channelrhodopsin chimera, C1C2, in DOPC lipid bilayer, NaCl, TIP3 water.

NAMD Scalable Molecular Dynamics

LONI

Visual Molecular Dynamics (VMD)

DCDAnalyzer

python™

MD ANALYSIS

NumPy

matplotlib

THE UNIVERSITY of NEW ORLEANS

Monika VanGordon

University of New Orleans

Sandia National Laboratories

Molecular dynamics simulations of transmembrane proteins

GPU and parallel computing
*effective use of the machines

NAMD Scalable Molecular Dynamics

LONI

VisIt + SeedMe

Visual Molecular Dynamics (VMD)

Python for HPC
GitHub

DCDAnalyzer

python™

MD ANALYSIS

NumPy

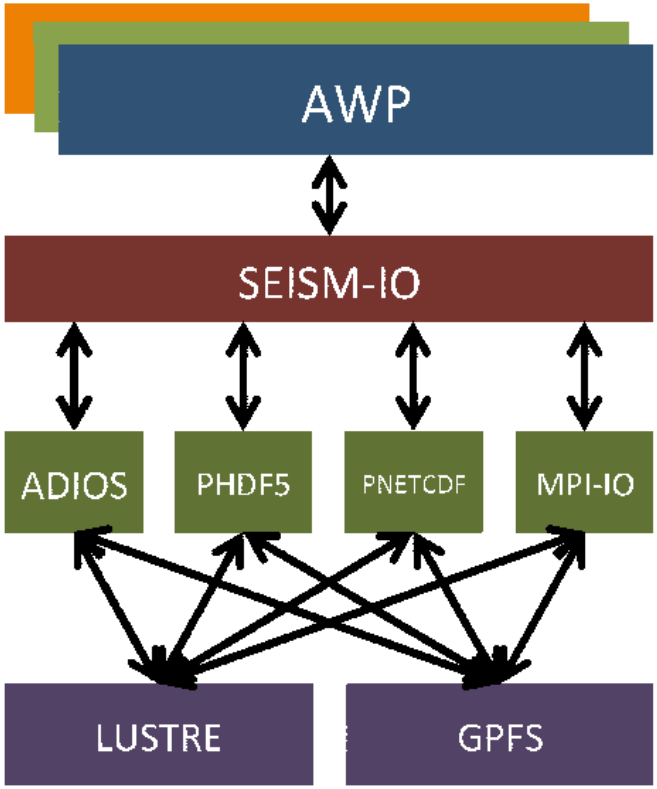
matplotlib

Hui Zhou
SDSC; Second Monitoring and Application Center, CEA

Seism-IO Library debug and development

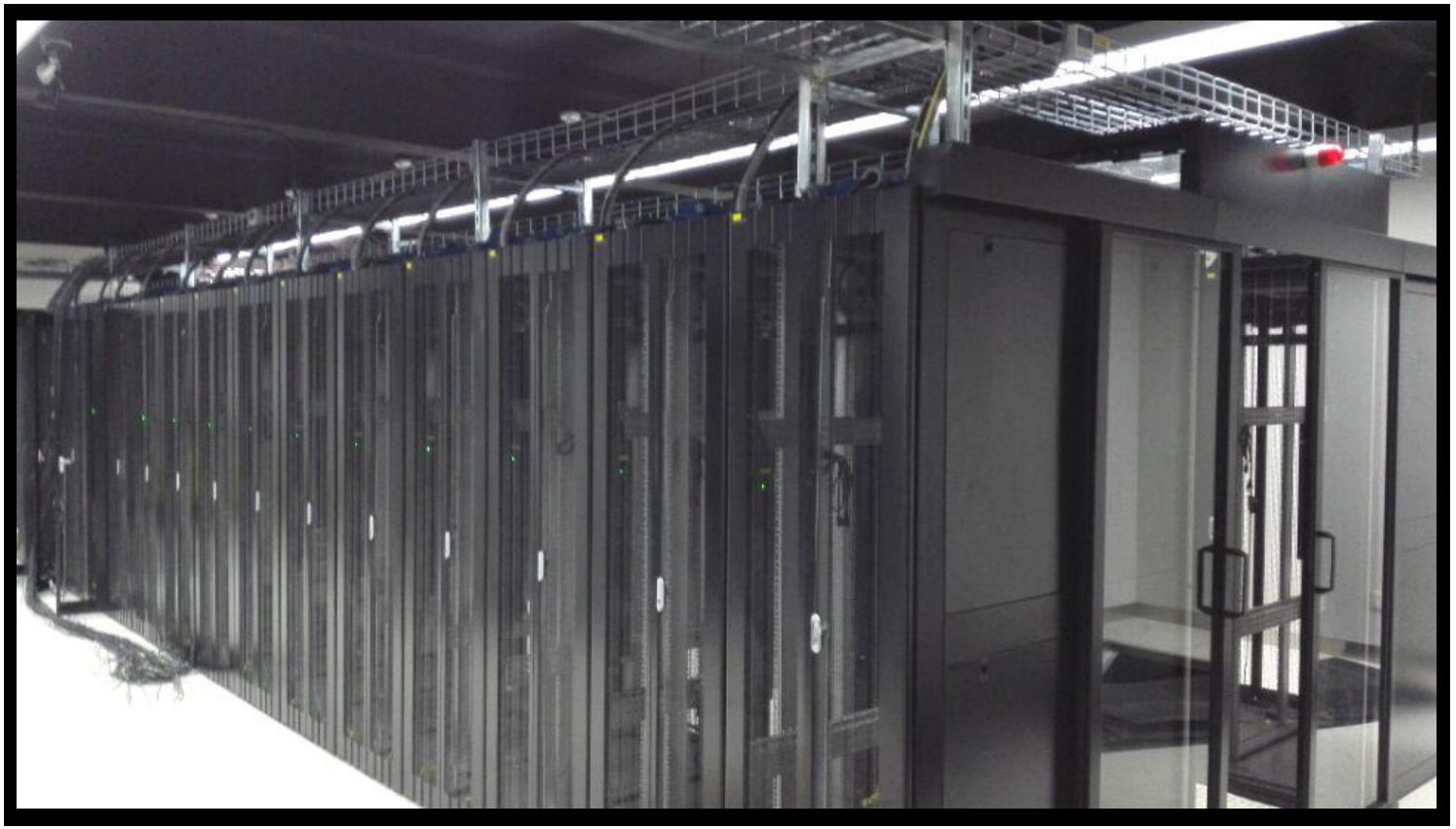
1.Fixed stablility issue of MPI-IO module. Passed 24,000 Cores Test.

2. Intergrating PHDF5, PNETCDF, ADIOS modules.

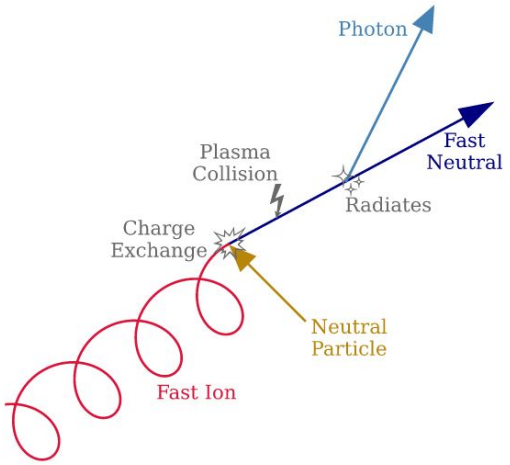


Hui Zhou
SDSC; Second Monitoring and Application Center, CEA

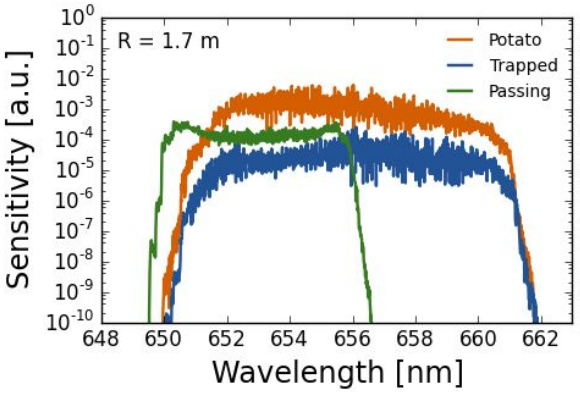
Interesting course related **big data**
We are building China's earthquake disaster recovery data center.

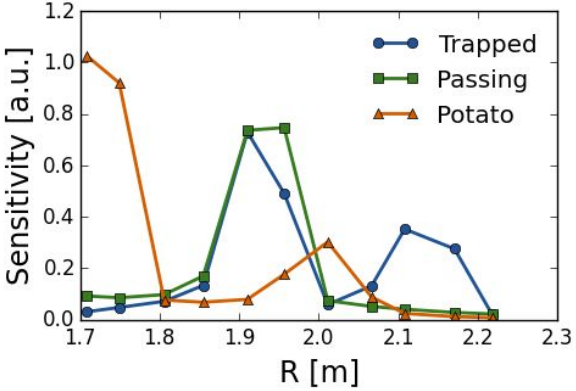


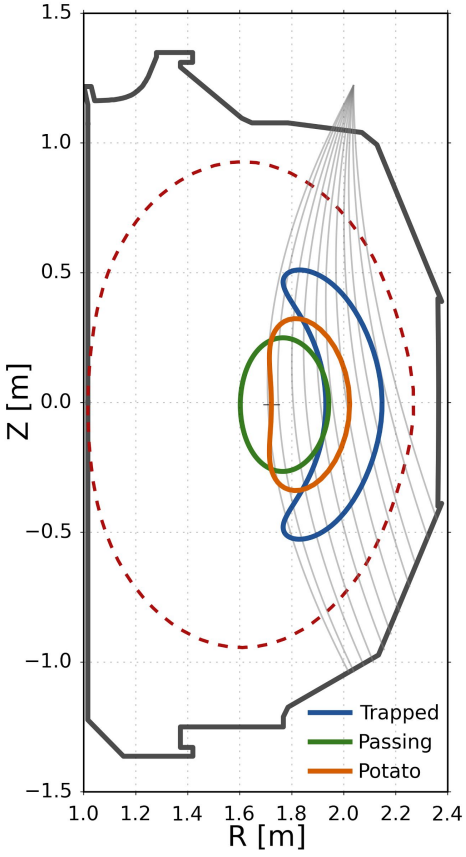
I Simulate Fast-ion Diagnostics in Fusion Devices



Fast-ions charge exchange with neutral particle and emit Doppler shifted photons







There are two reasons I need HPC

Problem: Orbits

I need to calculate a lot of orbits to properly explore the orbit phase space

Solution: Many processes across compute nodes

Problem: OpenMP code is not fast enough for Bayesian methods

I need to run my code hundreds of thousands of times to build up a posterior distribution

Solution: Optimization? GPU?

Sumukh Sagar Manjunath
San Diego State University

- Research Topic
  - Ocean Science Education Portal
  - Web Interface to Host General Curvilinear Coastal Ocean Model (GCCOM)
  - Meter-scale, fully 3D curvilinear, non-hydrostatic, large eddy simulation (LES) model
  - Implementation of Drop-a-Drifter Water Trajectories over the coast.
- Complexity of GCCOM
  - 16-core Xeon nodes each w/ 64GB RAM.

		6 hours simulation / 10 minutes Assimilation			
nodes:processes	Ensemble Size	Wall-clock time (hours)	Output Size	Total RMSE	Total Spread
2:15	30	6.40	3.93 GB x 2	0.69536	0.56466
4:15	60	7.40	7.38 GB x 2	0.68237	0.56336
6:15	90	8.10	10.84 GB x 2	0.6779	0.56255

- Challenges
  - Data storage and retrieval
  - Analyzing Data
  - Data Visualization
  - Compute and I/O Parallelization
- Website: http://sci.sdsu.edu/csrc-cod/

Sumukh Sagar Manjunath
San Diego State University

- Introduction of a powerful Viz Tool like Visit
- Some insights on using spark for our data analysis
- Ease of Using Gateway Portals

Aviv Solodoch

UCLA, AOS department

Physical Oceanography

Eddy shedding in the North Atlantic.
Ocean currents constantly “bleed out” eddies\vortices.

Using a community (hydrodynamical) model, with mpi/openmp parallelization.

Main take-home's

Gamers are right. GPU's are super-cool. I would like to look into utilizing them in physical oceanography models.

Main take-home's

Gamers are right. GPU's are super-cool. I would like to look into utilizing them in physical oceanography models.

Sun Tzu, ~500 BC: “***If you know the enemy, and know yourself, you need not fear the results of a hundred battles***.”

Main take-home's

Gamers are right. GPU's are super-cool. I would like to look into utilizing them in physical oceanography models.

Sun Tzu, ~500 BC: "***If you know the enemy, and know yourself, you need not fear the results of a hundred battles***."

HPC and the long tail of science, ~2016:

"If you k*now your HPC system, and optimize your code, you shall have great FLOPI...*"