Codemeta: A Rosetta Stone for Software Metadata

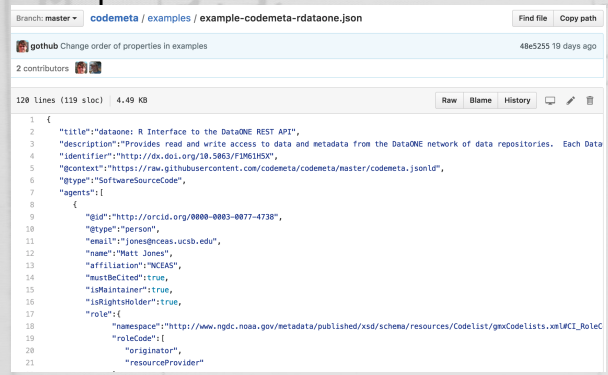
EArly-concept Grants for Exploratory Research (EAGER)

Carl Boettiger, UC Berkeley

Abstract

Research relies heavily on scientific software, and a large and growing fraction of researchers are engaged in developing software as part of their own research (Hannay et al 2009). Despite this, *infrastructure to support the preservation, discovery, reuse, and attribution of software* lags substantially behind that of other research products such as journal articles and research data. This lag is driven not so much by a lack of technology as it is by a lack of unity: existing mechanisms to archive, document, index, share, discover, and cite software contributions are heterogeneous among both disciplines and archives and rarely meet best practices (Howison 2015). Fortunately, a rapidly growing movement to improve preservation, discovery, reuse and attribution of academic software is now underway: a recent NIH report, conferences and working groups of Force11, WSSSPE & Software Sustainability Institute, and the rising adoption of repositories like GitHub, Zenodo, figshare & DataONE by academic software developers. Codemeta is a distributed open source project to improve how these resources can talk to each other.

Example JSON-LD



A Metadata CrossWalk

18 contributors met in Portland, OR on April 15-17, 2016 to agree on a draft crosswalk between major software metadata schema.

Schemas in Crosswalk

DataCite, OntoSoft, Zenodo, GitHub, Figshare, Software Ontology, Software Discovery Index, Dublin Core, R Package Description, Debian Package Metadata, Python Distutils (PyPI), Trove Software Map, Perl Module Description (CPAN::Meta), JavaScript package descption (npm), Java (Maven), Octave, Ruby Gem, ASCL, Schema.org

A JSON-LD Context for Software Metadata

doi 10.5063/schema/codemeta-1.0

- A linked-data representation of codemeta concepts
- Basic types from Schema.org, also Dublin Core and XSD
- 26 new terms introduced in `codemeta` namespace

Project Deliverables

Project Website: codemeta.github.io

JSON-LD Context file: github.com/codemeta/codemeta/codemeta.jsonld

Crosswalk Table: github.com/codemeta/codemeta/crosswalk.csv

`codemeta.json` examples: github.com/codemeta/codemeta/examples

Coming soon

Utilities for automatically generating `codemeta.json` for common software package types (e.g. R packages)

Zenodo parsing of `codemeta.json` into Zenodo metadata and on to DataCite records.

Partner organizations extend metadata representation

Organizations

NSF

DataCite
FIND, ACCESS, AND REUSE DATA

figshare
credit for all your research

DataONE
Data Observation Network for Earth

zenodo

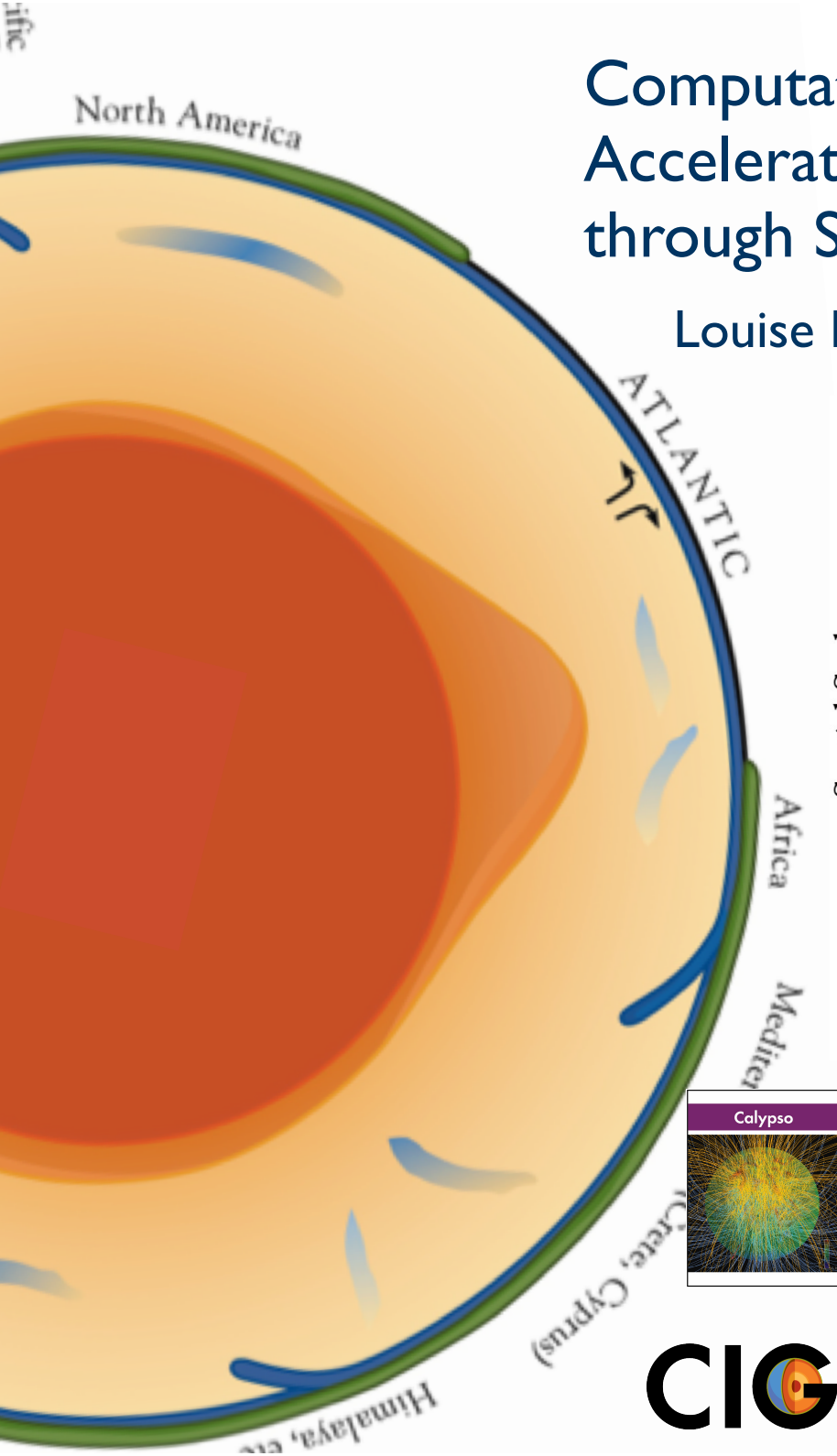
Workshop Participants

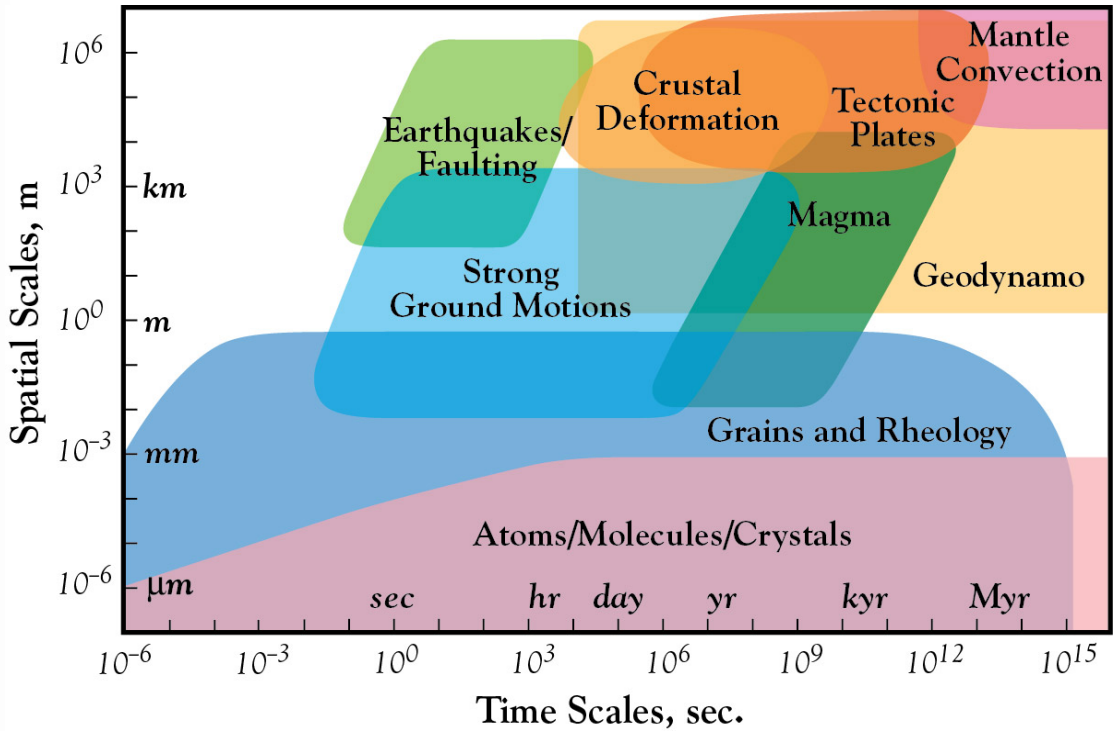
- Carl Boettiger, UC Berkeley
- Matt Jones, NCEAS
- Arfon Smith, GitHub
- Yolanda Gil, USC ISI
- Martin Fenner, DataCite
- Krzysztof Nowak, Zenodo
- Mark Hahnel, figshare
- Abby Mayes, Mozilla Science Lab
- Luke Coy, RIT & MSL
- Kyle Niemeyer, Oregon State
- Alice Allen, ASCL
- Mercè Crosas, Harvard, IQSS
- Ashley Sands, UCLA
- Neil Chue Hong SSI
- Peter Slaughter, NCEAS
- Patricia Cruse, DataCite
- Dan Katz, NCSA
- Carole Goble, University of Manchester

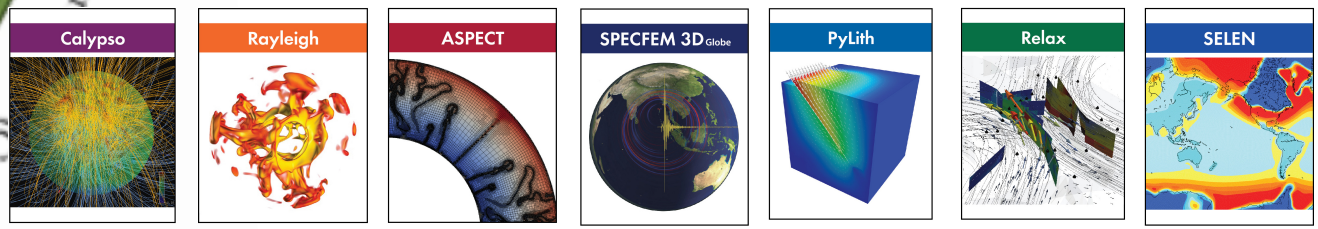
See more contributors at github.com/codemeta/codemeta/graphs/contributors and github.com/codemeta/codemeta/network/members

Computational Infrastructure for Geodynamics: Accelerating Discovery in the Solid Earth through Sustainable Scientific Software

Louise H. Kellogg, University of California, Davis







CIG COMPUTATIONAL INFRASTRUCTURE for GEODYNAMICS

NSF

Scalable Big Data Clustering by Random Projection Hashing

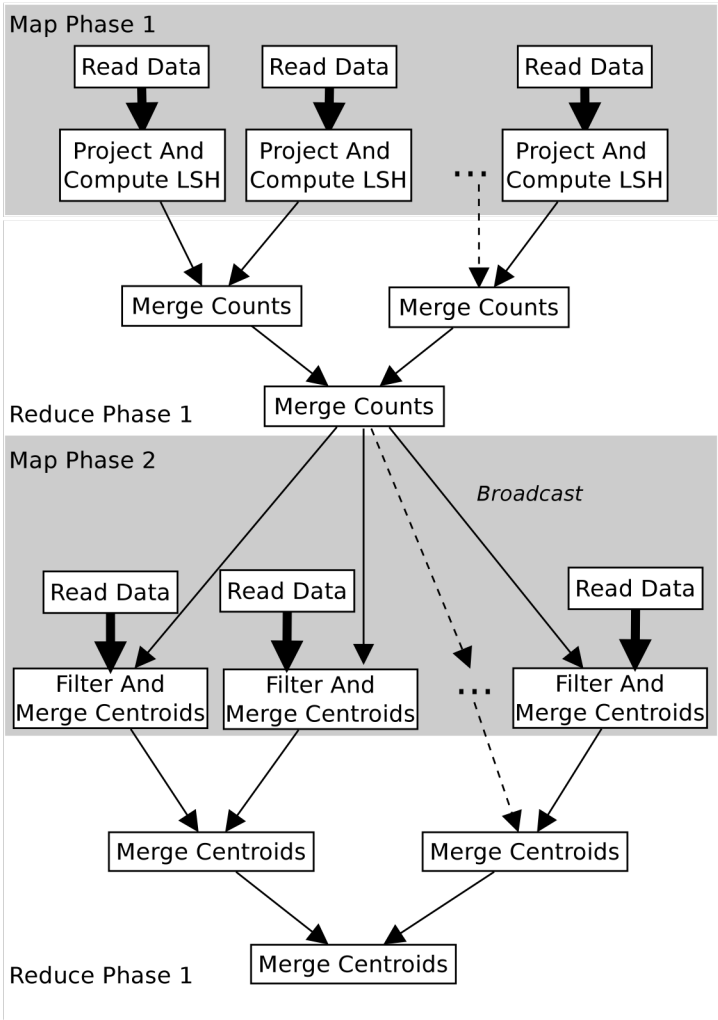
UNIVERSITY OF Cincinnati

Algorithm:

1. Project high-D $\rightarrow$ low-D
2. LSH low-D vectors
3. Count-min sketch
4. Top k counters are centroids

HAR Data	**ARI**	**Time**
Kmeans	0.461	24.746
RPHash	0.363	0.4838
RPHash-Dis	0.48	8.116

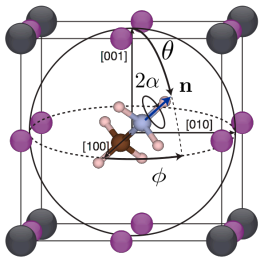
Dim.	**KMeans**	**RPHash**	**RPHash-Dist**
100	3.9656	0.4890	7.54
500	32.92	0.5594	7.87
1000	142.2341	0.7206	8.51
1500	237.0007	0.8233	9.48
2000	366.0743	0.8796	10.73
2500	431.5876	0.9997	11.43
3000	542.0223	1.1122	13.39
3500	631.8423	1.2421	12.97
4000	741.915	1.3157	13.78
4500	811.3911	1.3986	14.40
5000	909.223	1.5095	15.14
5500	975.2703	1.5977	15.98
6000	1076.882	1.6805	17.24
6500	1187.6062	1.7933	18.02



Automated Statistical Mechanics for the First-Principles Prediction of Finite Temperature Properties of Crystals

Jonathon Bechtel, Julija Vinckeviciute, Sanjeev Kolli, S. N. Harsha Gunda, Anton Van der Ven
Materials Department, University of California Santa Barbara

CASM
A Clusters Approach to Statistical Mechanics



input crystal

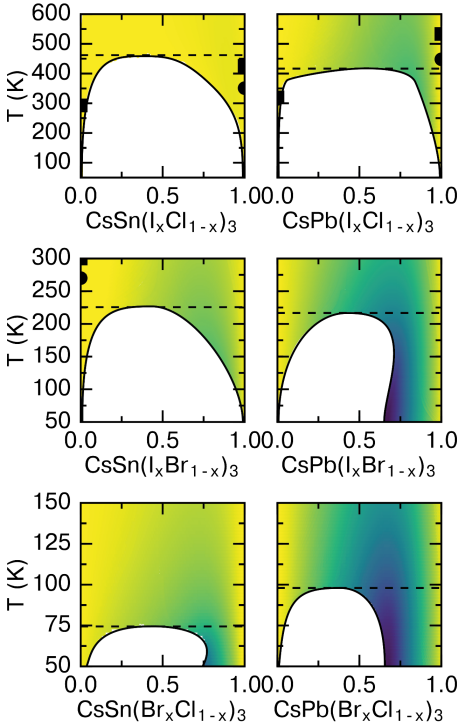
Electronic Structure Methods

$$\left[-\nabla_i^2 + V_{ion}(r_i) + \int \frac{\rho(r')}{r-r'}dr' + V_{xc}(\rho(r_i))\right]\phi_i(r_i) = \varepsilon_i\phi(r_i)$$

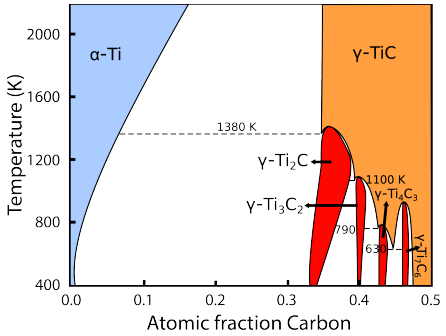
Effective Hamiltonians

(kinetic) Monte Carlo

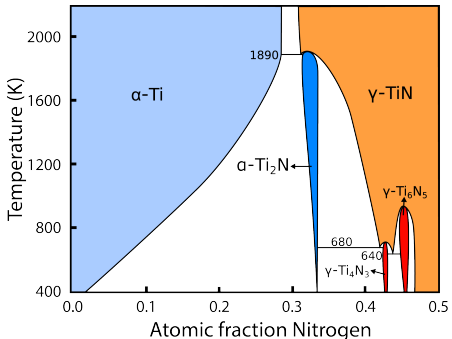
Thermodynamics **Kinetics**



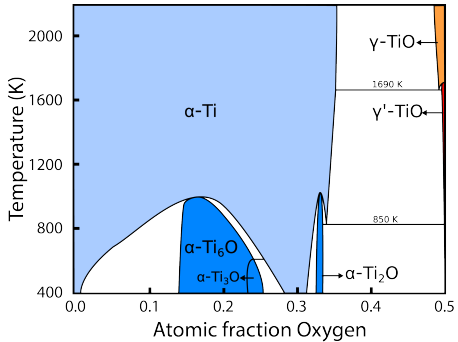
Ti-C



Ti-N



Ti-O

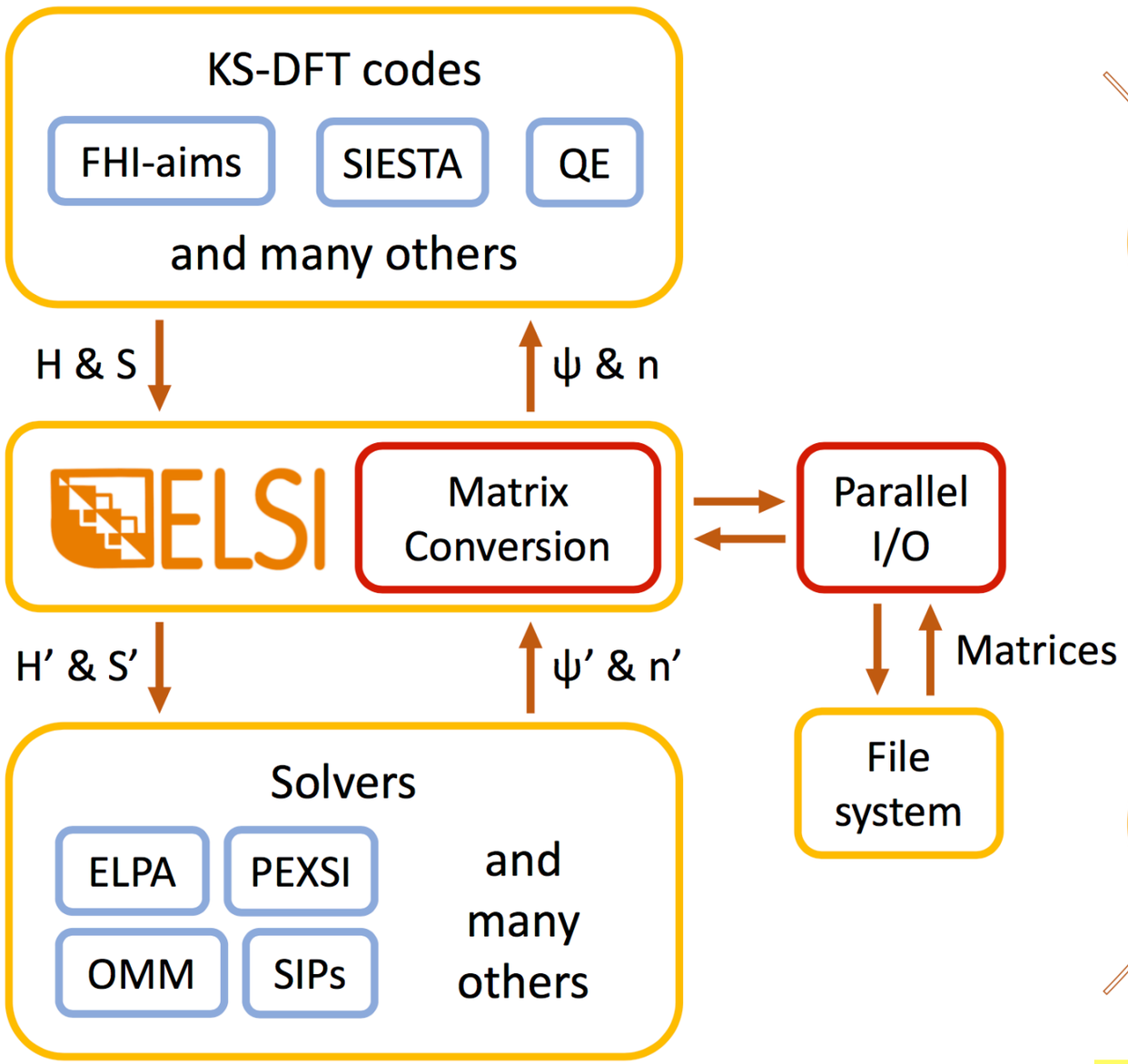


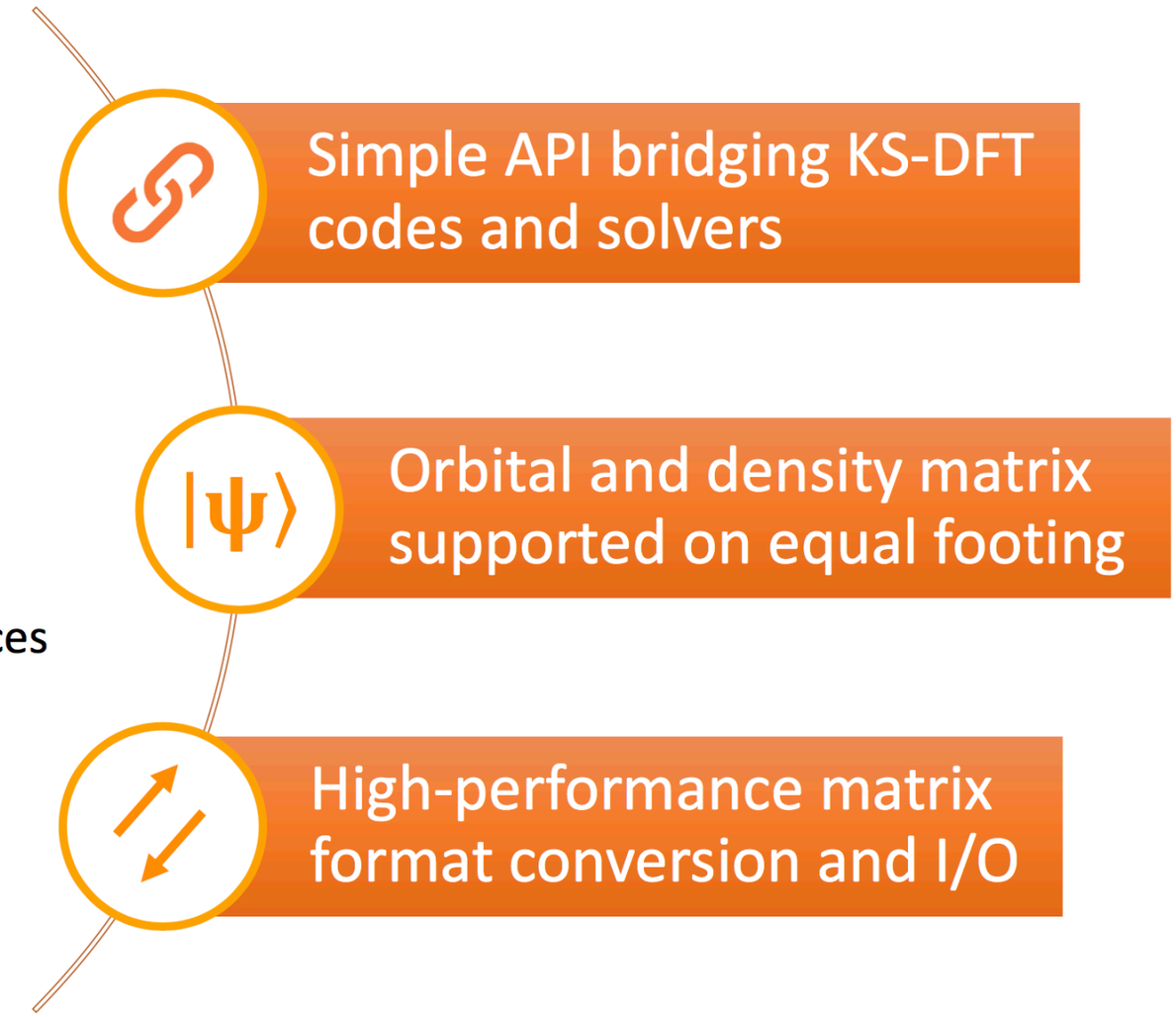
SI2-SSE: GraphPack: Unified Graph Processing with Parallel Boost Graph Library, GraphBLAS, and High-Level Generic Algorithm Interfaces

Andrew Lumsdaine

ELSI - Infrastructure for Scalable Electronic Structure Theory

Victor Wen-Zhe Yu, Fabiano Corsetti, Alberto García, Stefano de Gironcoli, William Paul Huhn, Mathias Jacquelin, Weile Jia, Murat Keçeli, Raul Laasner, Lin Lin, Jianfeng Lu, Yingzhou Li, Álvaro Vázquez-Mayagoita, Chao Yang, Haizhao Yang, and Volker Blum





Yu et al., Comp. Phys. Commun. ***222****, 267 (2018)*

http://elsi-interchange.org

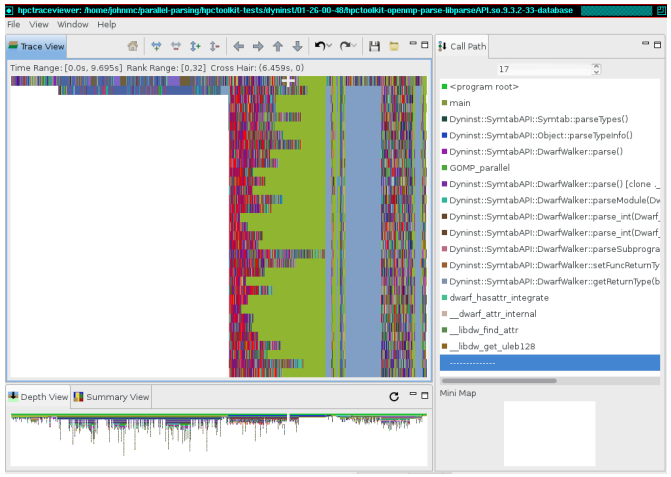
Part of upcoming CECAM
"ESL (Electronic Structure Library) bundle"

Binary Analysis for Performance, Security, and Correctness Tools

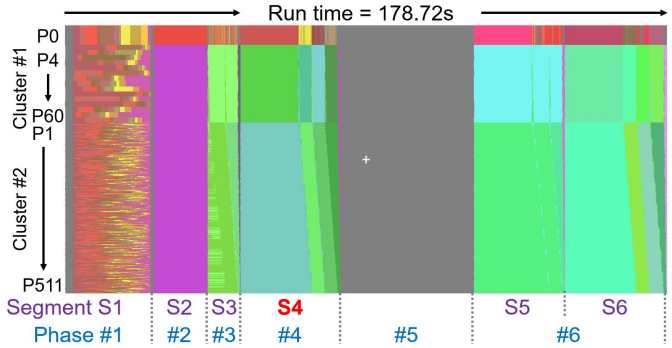
John Mellor-Crummey (Rice)

Barton Miller (Wisconsin)

Parallel Parsing of Application Binaries



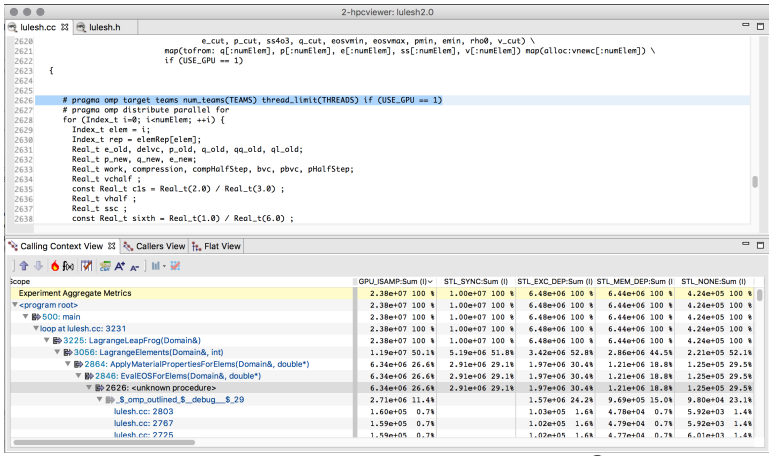
Automated Analysis of Parallel Program Traces



A New Dynamic Data Race Detector for OpenMP

- Instrument data accesses in binaries
- Track task orderings, mutual exclusion, and data environments during execution
- Assess races involving each access by reasoning about locks and orderings

Performance Analysis of GPU Accelerated Applications



RICE

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

ASSISTments™

A Free Public Service of Worcester Polytechnic Institute

SI2 - Adding Research Accounts to the ASSISTments Platform: Helping Researchers Do Randomized Controlled Studies with Thousands of Students

Worcester Polytechnic Institute

Background on ASSISTments

ASSISTments is a free, university-based platform. Each day, teachers assign problems to thousands of students (currently 50,000 students) through problem sets aligned to the Common Core State Standards. Randomized controlled experiments are often embedded in these problem sets to evaluate the efficacy of learning interventions. Heffernan has been funded by the NSF to conduct certain types of research using ASSISTments (i.e., spacing studies). Seventeen peer-reviewed publications have resulted from controlled experiments conducted within ASSISTments.

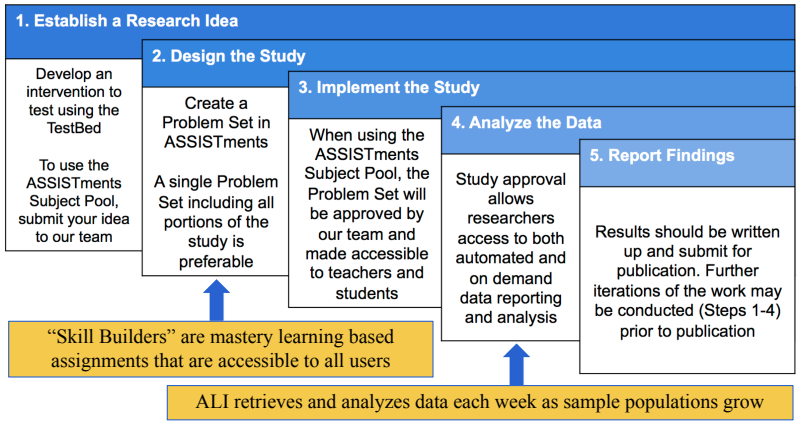
Purpose of the SI2 Grant

This grant proposed the evolution of ASSISTments into a shared scientific instrument that researchers could use to experiment with learning at scale. Software infrastructure modifications were necessary to aid in this transformation. The result is the ASSISTments TestBed (www.ASSISTmentsTestbed.org). The community of potential users include educational psychologists and mathematics education researchers. Developing relationships with schools has traditionally been costly for researchers. The ASSISTments TestBed reduces these costs by bridging relationships with teachers and researchers to conduct noninvasive classroom experiments that improve education.

WPIs Contribution

- Created the infrastructure to allow researchers to design and implement their own RCTs.
- Created a way for teachers to access materials with embedded RCTs but not be distracted by them.
- Ran multiple trainings (AERA, as well as a webinar) to recruit pilot researchers.
- Created a workflow for idea submission. WPI mentors researchers toward study designs compliant with our IRB via 'normal instructional strategies.'
- Created the Assessment of Learning Infrastructure (ALI) to provide easy access to study data. This tool eases the data processing and analysis required of researchers as raw data files can be overwhelming.

ASSISTments TestBed Research Progression



Universal Reporting

Dear Researcher,

Welcome to the data record for problem set PSAMR8Z. You have received this record as part of our weekly reporting. Automated data analysis is featured below, offering a preliminary overview of your sample and a selection of analyses for your consideration. The latter portion of this report contains the raw data files from which you can conduct your own thorough analyses. When publishing your work, please reference this report as a stable location for readers to access your data for review and replication.

By downloading content from this page, you agree to our **Terms of Use**.

Automated Data Analysis

Completion Rates
Students that have started PSAMR8Z: 1556
Students that have completed PSAMR8Z: 1043

Bias Assessment
Before analyzing learning outcomes, we suggest first assessing potential bias introduced by your experimental conditions (i.e., examine differential dropout). The table below reports the number of students that have completed PSAMR8Z, split out by experimental condition. Condition distribution was not significantly different, X^2 (*2, N = 762*) = 4.31, p > 0.05.

Conditions	Started	Finished	% Completed
Text	248	144	58
Video - Human	260	148	57
Video - Pen	254	126	50

Assessment of Learning Infrastructure
A Tool For Researchers

Raw Data Files

Raw data files contain the logged information for each student that has participated in your study. We provide this data in a variety of formats, as shown below, to assist in your analytic efforts.

- *Student Covariate Dataset*
- *Action Level Dataset*
- *Problem Level Dataset*
- *Student Level Dataset*
- *Student Level + Problem Level Dataset*

For a glossary and dataset tutorials, please visit our **Glossary Page**.

Success To Date

- Participation from a dozen researchers, representing:
  - Boston College, Freiburg University, Harvard University, Indiana University, Northwestern University, Southern Methodist University, Texas A&M, University of Colorado - Colorado Springs, University of California- Berkeley, University of Maine, University of Wisconsin, and Vanderbilt.
- Participation from an educational company hoping to evaluate their product.
- Publications:
  - The first manuscript resulting from this pool of research is currently in press.
  - Three researchers have published at international conferences.
  - Two manuscripts are in press to promote this 'research evolution' in similar learning platforms.

Knowledge Gained

Good Experiences…

- Recruiting was easy. This service is in high demand and we filled our yearly researcher quota in just a single day.
- Researchers love our IRB terms. By separating researchers from primary data collection it was much easier to get external IRB approval.

Difficulties…

- We spend time negotiating with researchers about what constitutes 'normal instructional practice' (researchers want extensive pre/post tests).
- We have pivoted our focus to assignments where completion goals are well known to users, rather than letting researchers design new content that we then had to try to entice teachers to use.
- We have noticed a lack of clear, conceptual assessment items aligned to the Common Core State Standards in our content.

Improving Science by Promoting Replication…

- Pre-registration of studies and hypotheses
- Open data and open materials
- Reductions to the 'File Drawer' problem

WPI

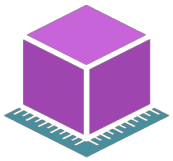
We acknowledge funding from the NSF, 1440753. The ASSISTments System that serves as a base for this work has also been supported by other awards (1316736, 1252297, 1109483, 1031398, 0742503, 1440753), ONR (N00014-13-C-0127), the IES (R305A120125, R305C100024), and GAANN (P200A120238).

For additional information, see: Ostrow, K., Selent, D., Wang, Y., Van Inwegen, E., Heffernan, N., & Williams, J.J. (In Press). The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment. To be Included in the Proceedings of the 6th International Conference on Learning Analytics and Knowledge.

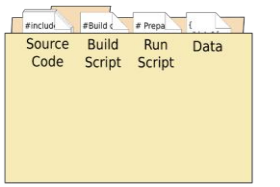
About the Author

Neil T. Heffernan (nth@wpi.edu) is a Professor of Computer Science, the Director of the Learning Science & Technologies Program, and the creator of ASSISTments.

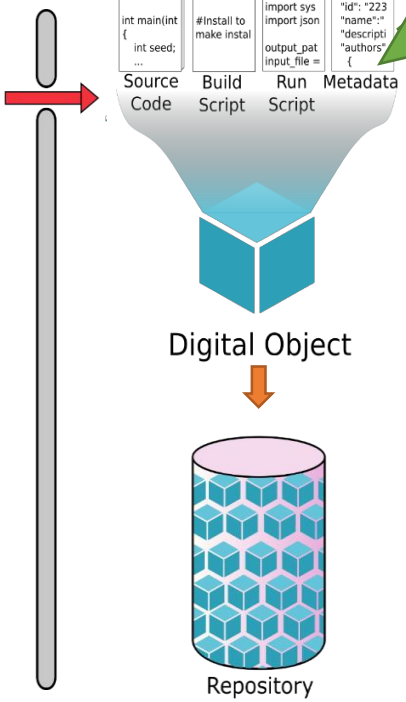
Artifact Execution Curation for Repeatability in Artifact Evaluation



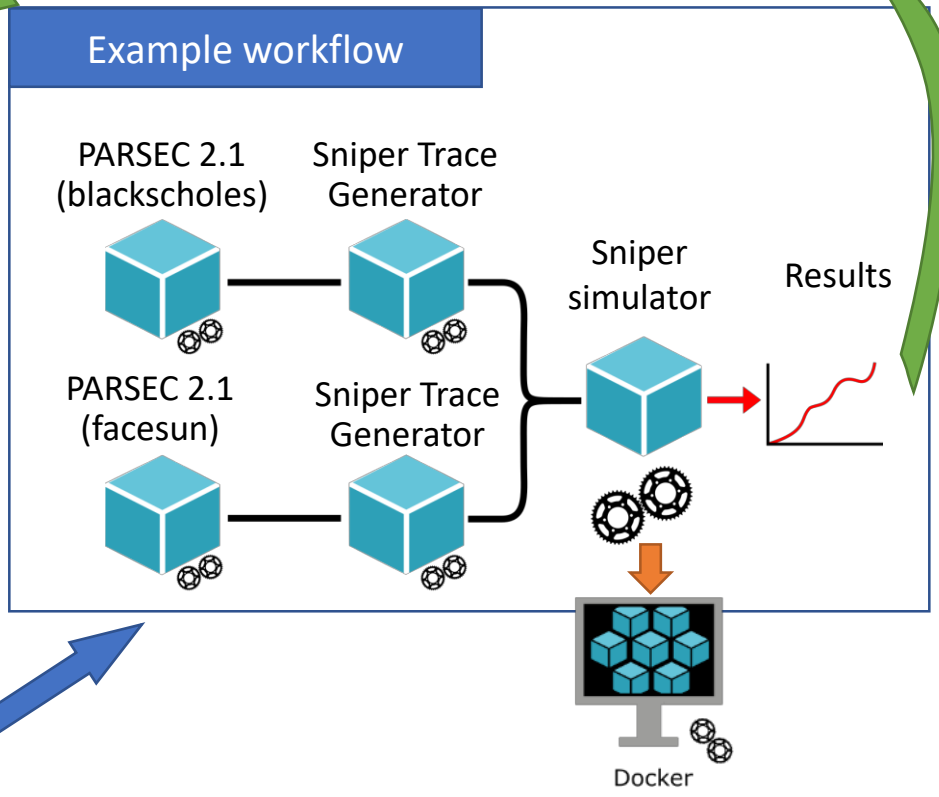
OCCAM



The ephemeral Internet



Track the provenance



Copy artifacts from the ephemeral internet

Use the artifacts in workflows

Execute the artifacts in repeatable environments

CDS&E: Numerical Investigation of Two-Particle Response Functions of Correlated Materials

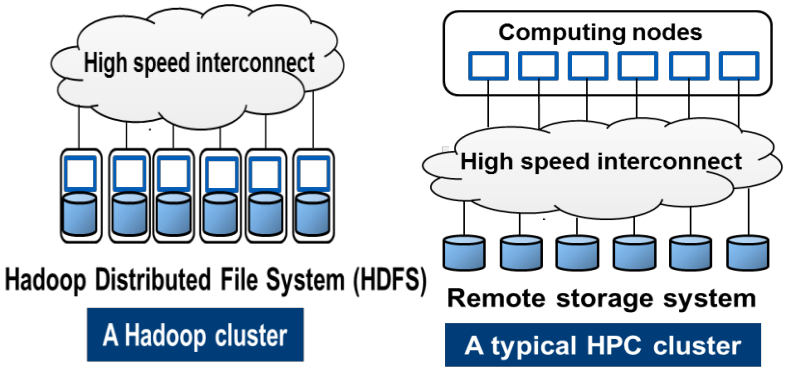
Emanuel Gull

CIF21 DIBBs: PD: Building High-Availability Data Capabilities in Data-Centric Cyberinfrastructure

Haiying Shen, Associate Professor
Computer Science, University of Virginia

U.S. NSF-OAC-1724845

MOTIVATION



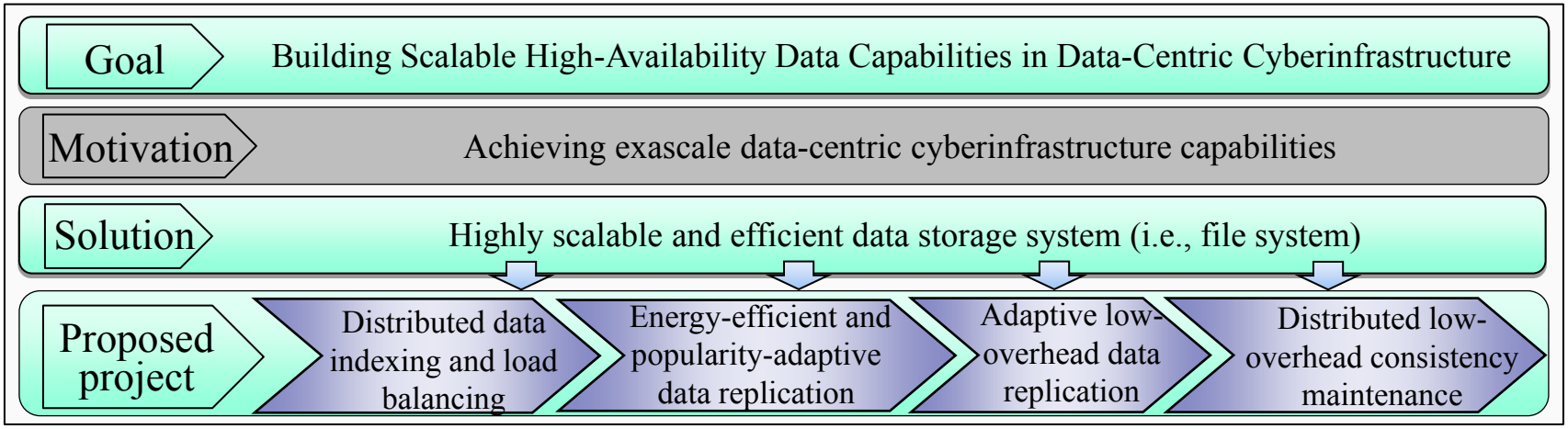
- HPC and HDFS storage architectures are not scalable enough to adapt to exascale systems

INTELLIGENTIAL MERIT

- Distributed data indexing and load balancing
- Energy-efficient and popularity-adaptive data replication
- Adaptive low-overhead data replication
- Distributed low-overhead consistency maintenance

BROADER IMPACTS

- Curriculum development activities
- Student recruiting and mentoring
- Impacts on the research community



CitSci.org: Advancing and Mobilizing Citizen Science Data through Integrated Sustainable Cyber-Infrastructure

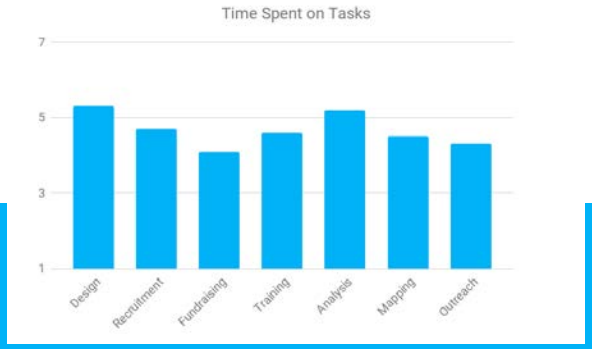
Greg Newman*, Stacy Lynn*, Melinda Laituri*, Russell Scarpino*, Louis Leibenberg†, Sarah Newman*, Justin Steventon†

We power citizen science projects all over the world

- 4,958 people
- 572 projects
- 796,158 data points
- 66,775 locations
- 2,099 protocols

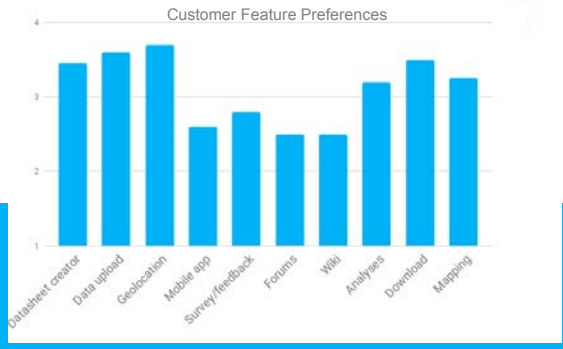
know your audience

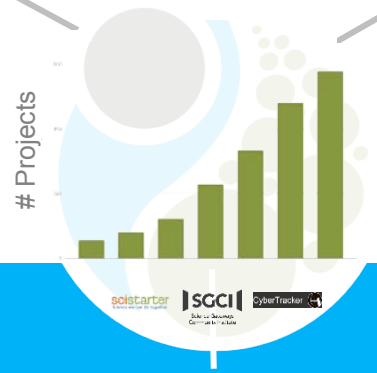
- 120+ customer discovery interviews



build relevant software

- design & analysis tools





partner for sustainability

Goals

- Broaden inclusivity of citizen science (CS)
- Improve CS software
- Mobilize CS data for meta-analyses
- Elevate the value and rigor of CS software and data

Integrations & Interoperability

Future Directions

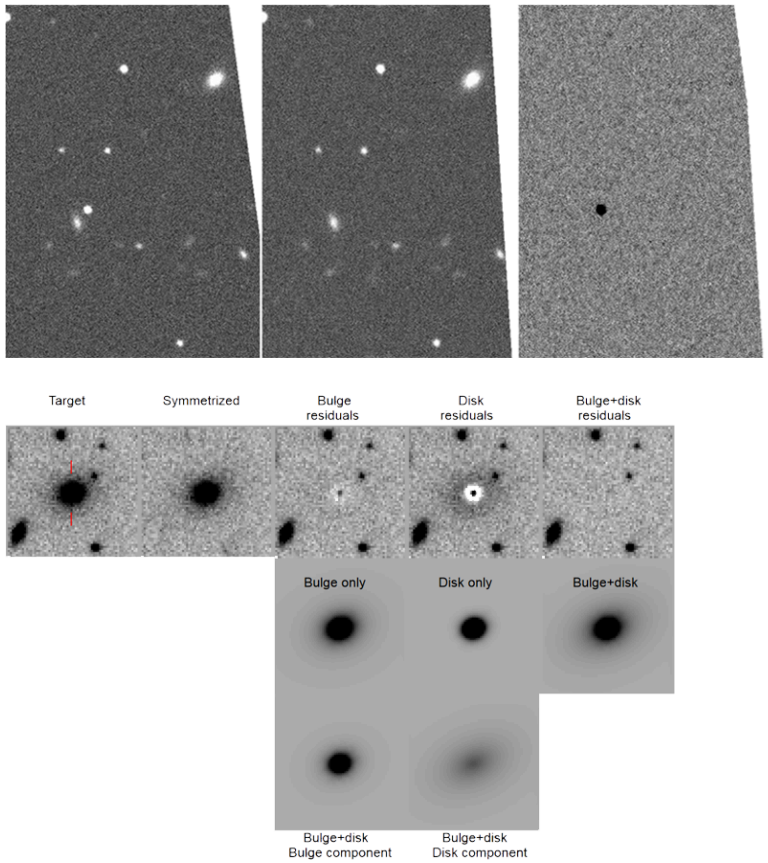
- Integrate real time precipitation data
- Integrate field based with online CS
- Offer design & implementation support services

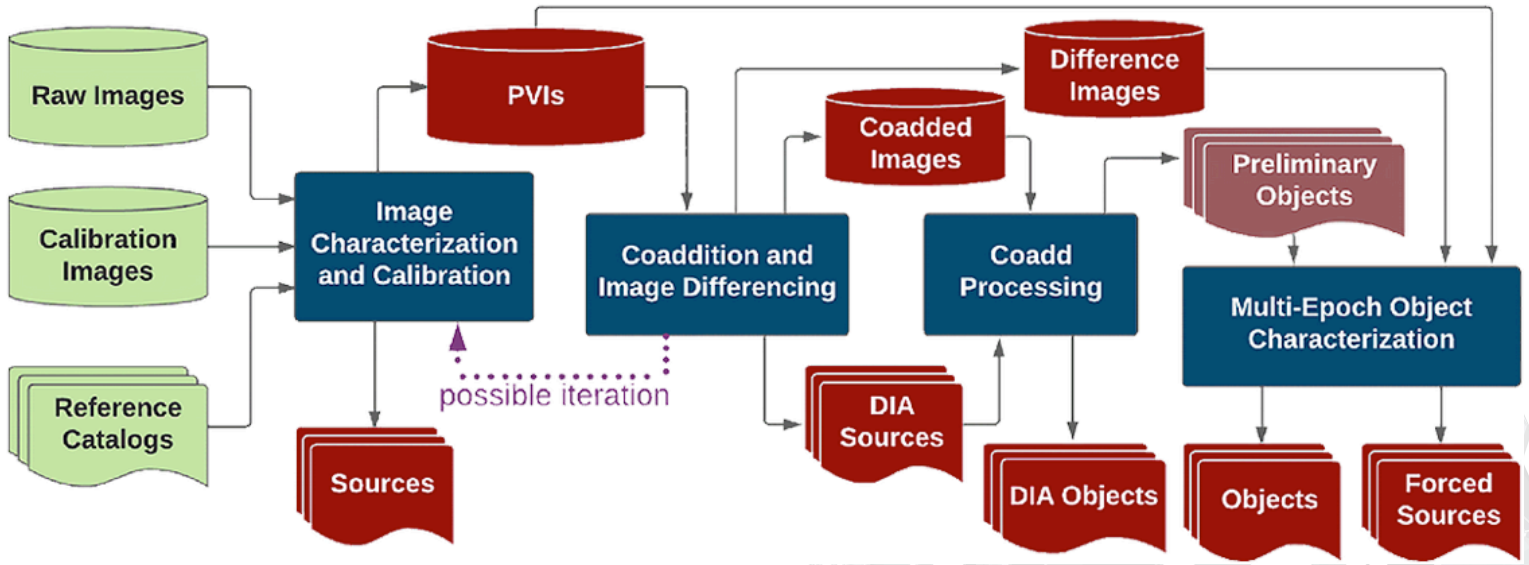
Colorado State University

NREL

SI2-SSE: Reusable Image Analytics Pipelines

Dino Bektesevic (dinob@uw.edu), Magda Balazinska, Alvin Cheung , **Andrew Connolly (PI)**, Mario Juric
University of Washington





Scaling the image processing and analytics pipelines for the Large Synoptic Survey Telescope (LSST) to run efficiently across a broad range of architectures and within the cloud

SI2-SSE: Development of Cassandra, A General, Efficient and Parallel Monte Carlo Multiscale Modeling Software Platform for Materials Research

Jindal K. Shah

INNOVATIVE
COMPUTING LABORATORY

SI2-SSE:

PAPI Unifying Layer for Software-Defined Events (PULSE)

Heike Jagode
Anthony Danalis
UNIVERSITY OF TENNESSEE

GOAL Offer support for **software-defined events (SDE)** to extend PAPI's role as a standardizing layer for monitoring performance counters.

VISION Enable HPC software layers to expose SDEs that performance analysts can use to form a **complete** picture of the entire application performance.

BENEFIT HPC application scientists will be able to better understand the interaction of the different application layers, and the interaction with external libraries and runtimes.

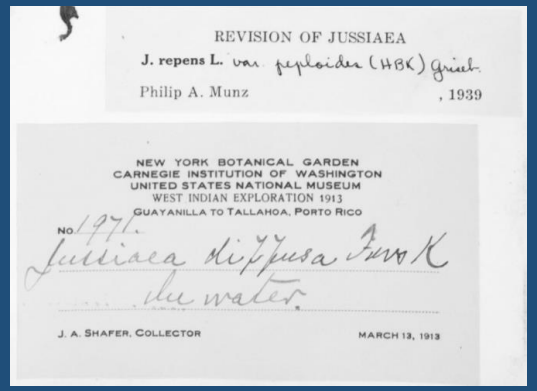
All new PAPI SDE API functions are available in C and FORTRAN, e.g.:

```
void *papi_sde_init(char *lib_name);

void papi_sde_register_counter( …, char *event_name,…, void *counter);

void papi_sde_register_fp_counter( …, char *event_name,…, func_ptr_t fp_counter);

void papi_sde_describe_counter( …, char *event_name, char *event_description);
```

HuMaIN

Human- and Machine-Intelligent Network of Software Elements for Cost-Effective Scientific Data Digitization

Ícaro Alzuru, Andréa Matsunaga, Maurício Tsugawa, Jingchao Luan, and José A.B. Fortes



Problem: Scientific data digitization is a slow process, commonly performed using humans (crowdsourcing).

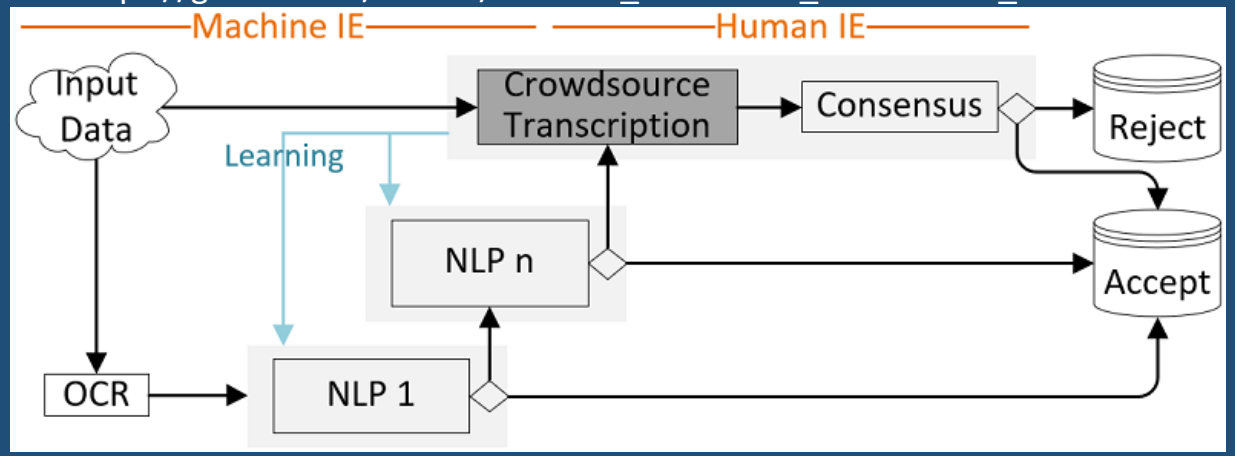
HuMaIN's Objective: Efficient digitization of scientific data.

HuMaIN: Information services framework for the integration of human & machine Information Extraction (IE) tasks.

Use case: Digitization of biocollections' labels.

1) **Self-aware** Information Extraction (SELFIE) workflows
   - IE alternatives are organized in cost-incremental order. They are able to **evaluate** each generated value and **decide** if it is correct or should be extracted by another IE process.
   - Tasks **actively improve** from humans (Active Learning - Humans in The Loop)
2) Smart use of the **crowdsourcing interfaces**.

https://github.com/acislab/HuMaIN_Self-aware_Information_Extraction



We have shown that hybrid approaches can yield **human-like quality** results and **machine-like** execution **speed**.

SEISM A sustainable **Software Environment** for **Integrated Seismic Modeling**

PI: John E. Vidale (USC), Award ID: ACI-1450451
Co-PIs: Yifeng Cui (SDSC), Kim B. Olsen (SDSU), Ricardo Taborda (University of Memphis)

SCEC and SEISM Project Objectives

The Southern California Earthquake Center (SCEC) coordinates fundamental and applied research on earthquake processes using Southern California as its principal natural laboratory. This research program is investigator-driven and supports core research and education in seismology, tectonic geodesy, earthquake geology, and computational science. The SCEC community advances earthquake system science through three basic activities: (a) gathering information from seismic and geodetic sensors, geologic field observations, and laboratory experiments; (b) synthesizing knowledge of earthquake phenomena through physics-based modeling, including system-level hazard modeling; and (c) communicating our understanding of seismic hazards to reduce earthquake risk and promote community resilience.

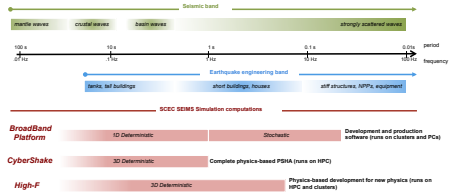
The SCEC SEISM project integrates scientific software elements (SSEs) into a software ecosystem for physics-based seismic hazard analysis (SHA) capable of using current and future extreme-scale computing systems.

- O1: Incorporate into community SSEs the physics necessary to extend deterministic simulations to seismic frequencies of engineering interest (>1Hz).
- O2: Improve the accuracy of SEISM simulations, and work with community of researchers and stakeholders to validate simulation products for practical use.
- O3: Enhance the performance and robustness of the SEISM computational platforms, and prepare them to operate on the next-generation supercomputers.

Ground Motion Simulations

As part of the design process, numerous engineering applications require seismograms as input to soil and structural models to simulate their response. Although recorded earthquake ground-motion datasets are constantly being compiled and updated over time, there is still only a limited number of records available, especially for large magnitude events (**M**>7). Additionally, other source-site effects, such as basin effects or rupture directivity, have not been captured by many of these recorded motions. Because of these limitations, simulated ground motions can provide otherwise unavailable information. However, it is essential to carefully validate those simulations before they can be confidently used for engineering purposes.

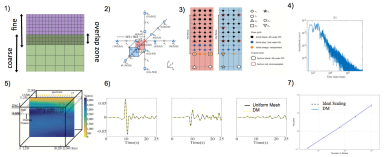
Three main simulation software suites were developed through SEISM to address this need (Broadband, CyberShake and High-F).



High-F

AWP-ODC Discontinuous Mesh

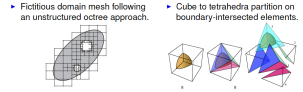
Motivation: uniform-grid methods inefficient for large contrasts in seismic wave speeds, such as basin models.
Challenge: Stability is inherently difficult to obtain in overlap between fine and coarse meshes.
Approach: Factor-of-three contrast in grid spacing along all three dimensions (1), 4th-order staggered grid (2,3).
Status: Stable to 1M+ timesteps for factor-of-three velocity contrast inside overlap zone (4), accurate in realistic velocity models using finite fault sources in overlap zone (5,6), scalable to 1024+ GPUs (7).



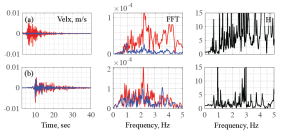
Prototyping "Virtual Topography" Using Hercules

Topography affects ground motions differently depending on the geometry of the topographic features. This effect is frequency dependent.

- We partition the domain using a non-conforming meshing scheme (left) following fictitious domain ideas.
- We developed the so called Virtual Topography (VT) scheme, which accounts for the nonconforming nature of the outer elements that contain the free surface (right).
- We solve the weak form of the elastodynamic equation in space by an octree-based finite element method and a second order central difference step-by-step marching-forward scheme in time.



Results for a sample station below: velocity (Vel), Fourier amplitude spectra (FAS) and transfer functions (H=FAS(VT)/FAS(FLAT)) for a sample station; blue lines for FLAT and red for VT.



Open-source Software

BBP: https://github.com/SCECcode/bbp
Hercules: https://github.com/CMU-Quake/hercules/wiki/Hercules
AWP-ODC: https://github.com/HPGeoC/awp-odc-os
UCVM: https://github.com/SCECcode/UCVM
SeismTools: https://github.com/SCECcode/HighF

CyberShake

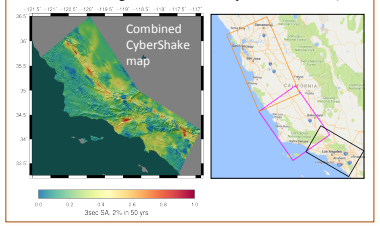
CyberShake is a simulation-based platform for probabilistic seismic hazard analyses (PSHA). CyberShake simulates ground motions by combining finite-fault rupture descriptions of earthquakes and wave propagation in 3D structural models of the Earth. CyberShake ground motions account for complex effects such as the coupling of directivity and basin response that cannot easily be captured with empirical modeling of ground motions.

For the latest CyberShake computation (Study 17.3 for central CA) we utilized Blue Waters and Titan to compute hazard results from over 400,000 earthquakes at 468 locations, using alternative velocity models, producing 285 million two-component seismograms, and computing and storing 43 billion intensity measures (response spectral accelerations and duration measures):

- Averaged 1295 nodes (CPU + GPU) for 31 days, maximum of 5374
- 900,000 node-hours consumed (21.6M core-hours)
- Pegasus, HTCondor, Globus used for workflows
- Workflow tools scheduled 15,581 jobs to both systems
- Transferred 308 TB of intermediate data between the two systems
- Workflow tools managed 777 TB of data, 10.7 TB of output data automatically staged back for archival storage

Porting CyberShake to the new region required non-trivial science investigation and the development of software to conduct them.

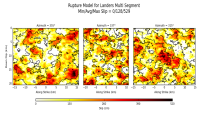
CyberShake Maps – Toward a CA-wide model
The map below shows the 2% in 50 yrs probability of exceedance of 3 sec. pseudo-spectral acceleration (PSA) for the combined region created following Study 17.3 (black box for Socal, magenta for central CA; orange box for future computation of NorCal).



Broadband Platform (BBP)

The BBP simulates broadband (0-20⁺ Hz) waveforms using different suites of scientific modules that model source, path and site effects. The current implementation of the BBP includes source modules relying on kinematic rupture generators and wave propagation is performed through 1-D layered velocity structures. These simplifications make the BBP a resource-efficient tool, ideally suited for a relatively large number of runs, which supports sensitivity studies and uncertainty estimations. The BBP was used recently in the Pacific Earthquake Engineering Research center (PEER) NGA-West2 and NGA-East projects to support ground-motion model development. Recent updates include:

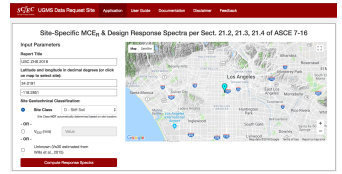
- New Irikura Recipe Method 1, the 7th method implemented
- Updated GP rupture generator code to model multi-segment ruptures
- New Fourier Amplitude Spectra (FAS) calculation module
- New scripts to combine time series from a number of separately-calculated segments, allowing for a multi-segment rupture to be simulated.
- Improvements for running ensembles calculation using cluster computers.



UGMS Committee and Building Code Application

Using simulation-based and traditional empirical-based approaches, the SCEC Committee for Utilization of Ground Motion Simulations (UGMS) is working to develop improved response spectral acceleration maps for the Los Angeles region for possible future inclusion in the NEHRP and ASCE 7 Seismic Provisions and the Los Angeles City Building Code. The UGMS is developing recommendations for the specification of the risk-targeted Maximum Considered Earthquake (MCE_R) design basis for spectral periods larger than 1s.
The committee has developed a criterion based on the combination of values from ground motion prediction equations (GMPEs from the PEER NGA-West2 project) and from CyberShake. The new MCE_R tool is to be released on May 3rd 2018.

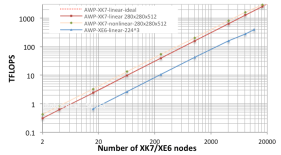
MCE_R Portal



Software Scaling

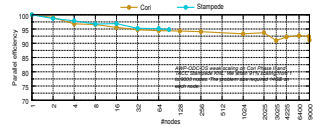
AWP-ODC on Nvidia GPU

- First 4-Hz nonlinear M7.7 simulation on the southern San Andreas Fault conducted using 4,200 Blue Waters GPUs
- 100% of parallel efficiency achieved for both linear/ nonlinear versions of AWP-ODC up to 8,192 GPUs.
- Accelerated time-to-solution from original nonlinear 0.68sec to 0.29sec per iteration (6.5 speedup).
- Blue Waters PAID project provided additional support.



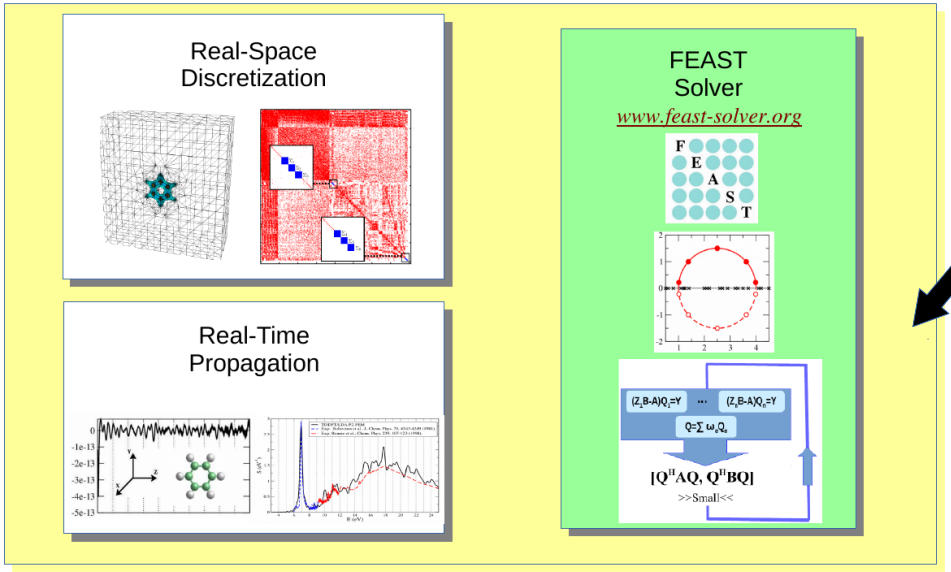
AWP-ODC on Intel Xion-Phi

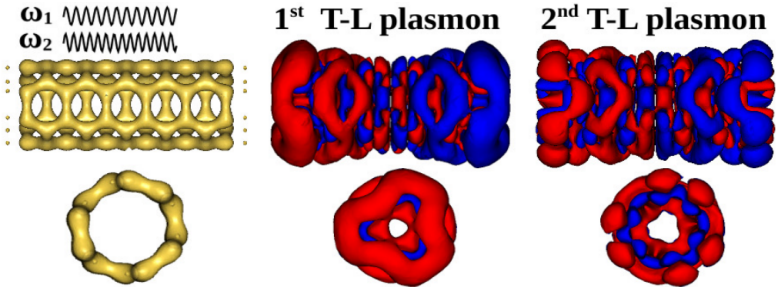
- Stencil generation and vector folding through YASK.
- Hybrid placement of grids in DDR and MCDRAM
- Normalized cross architecture evaluation in Mega Lattice Updates per Second (MLUPS): Xeon Phi KNL 7290 achieves 2x speedup over NVIDIA K20X, 97% of NVIDIA Tesla P100 performance.
- Performance on 9,000 nodes of Cori-II equivalent to performance of over 20,000 K20X GPUs at 100% scaling.



A parallel computing framework for large-scale real-space and real-time TDDFT excited-states calculations

E. Polizzi, UMass Amherst

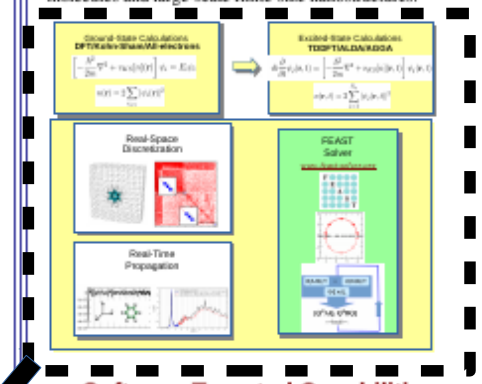




A parallel computing framework for large-scale real-space and real-time TDDFT excited-states calculations

Eric Polizzi (PI), ECE/Math UMass Amherst

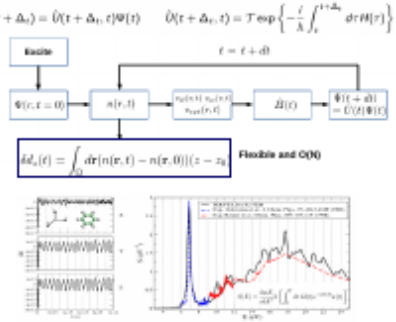
Abstract- This work involves developing a new open source software, NESSIE, for first-principle calculations making use of the FEAST eigensolver kernel. We aim at addressing the numerical challenges in real-space DFT and real-time TDDFT excited states calculations in order to operate the full range of electronic spectroscopy (UV-Vis, X-Ray, mid-IR), and study the nanoscopic many-body effects (plasmons, etc.) in arbitrary complex molecules and large-scale finite-size nanostructures.



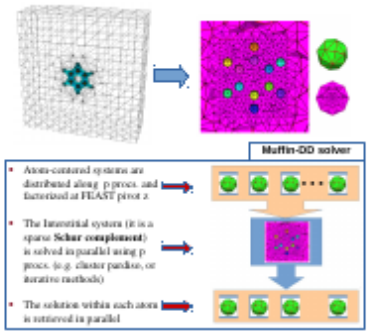
Software Targeted Capabilities

- Beyond pseudopotential approx. →All-electron calculations
- Beyond plane-wave/LCAO discretizations →Finite Element Method and Domain-decomposition- high accuracy, O(N) and parallelism
- Beyond DFT ground state →Capture excited-states
- Beyond frequency domain and linear response→Real-time propagations account for any forms of non-linear response- achieve O(N) and parallel scalability
- Beyond Born-Oppenheimer approx.→Ion dynamics via TDDFT/MD
- Beyond traditional solvers → Optimal use of the FEAST solver for achieving linear parallel scalability

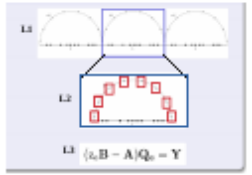
Real Time TDDFT

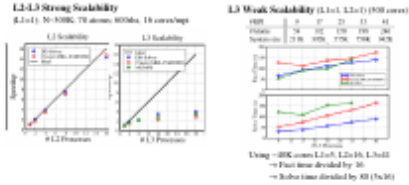


Real-Space DD

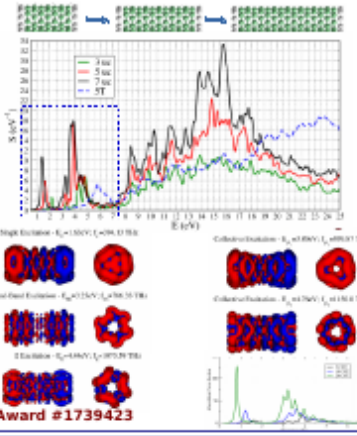


PFEAST: Three Levels of Parallelism





Applications



2018 NSF SI² PI Meeting, Award #1739423

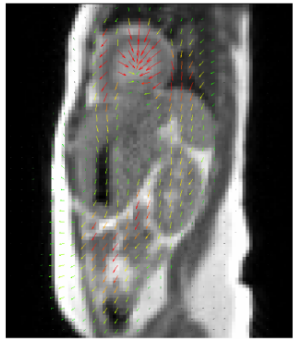
High-Performance Workflow Primitives for Image Registration and Segmentation

James A. Shackleford, Nagarajan Kandasamy, & Gregory C. Sharp

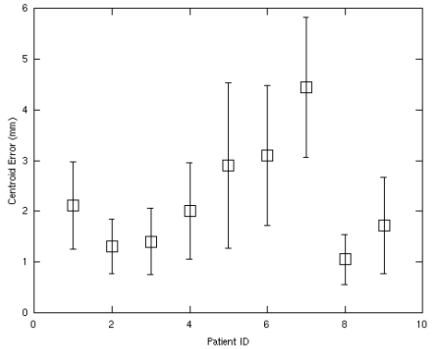
Project Motivation

Image registration and segmentation are vital enabling technologies for addressing many complex, data driven problems. Examples include individualized medical treatment where disease progression is monitored by analyzing MRI, CT, or ultrasound images over time; identifying anatomical structures in medical images; recognizing objects and people in video footage; and extracting imageable biometrics such as fingerprints, faces, and the iris. Images and videos can now be easily acquired at a rate that far surpasses our capacity to perform advanced image analysis. For this reason, advanced registration and segmentation algorithms are not routinely used for many large-scale and time sensitive applications because they require more processing time than is available.

Real-time Image Registration



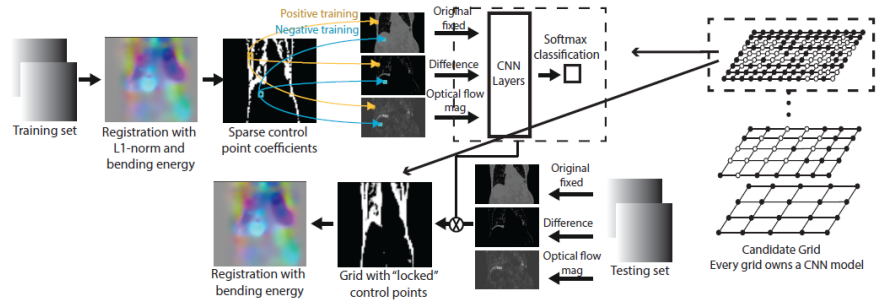
Example real-time deformable image registration result (left)



Difference between target centroid and manual contouring (mm) (right)

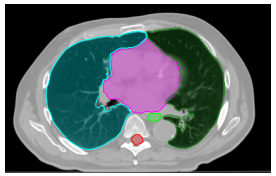
Cine-mode MR is an ideal imaging method for real-time therapeutic control. Our deformable image registration method can be computed in less than 200 ms, to enable real-time tracking of multiple soft-tissue targets. The intended application is automatic optimization of radiotherapy in the thorax and abdomen, to compensate for respiratory motion. *(Submitted, RSNA 2018)*

Multi-grid Image Registration



B-spline deformable image registration is well established for mapping 3D medical imaging volumes. B-spline coefficients can be optimized simultaneously at multiple resolutions using a multi-grid approach. Algorithm performance is boosted by detecting grid sparsity and removing unnecessary parameters from the coefficient grid. *(Accepted, CVPR 2017)*

Work in Progress



Automatic image segmentation methods use distance maps during training and classification. Drexel PhD students Shihao Song and Michael Spanier are developing high-performance methods to accelerate this important algorithm.

Contact and Visit

NSF Award #1642380

James Shackleford
shack@drexel.edu

Naga Kandasamy
kandasamy@drexel.edu

Greg Sharp
gcsharp@mgh.harvard.edu

http://www.libkaze.com

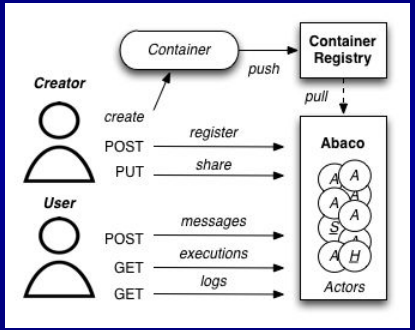
http://www.platimatch.org

CORE INFRASTRUCTURE INITIATIVE

SI2-SSE: Abaco -Flexible, scalable, and usable Functions-as-a-service via the Actor Model

PRIMARY CAPABILITIES

- **Reactors** provide functionality for event-driven programming
- **Asynchronous Executors** enable executing functions on the Abaco cluster from directly within a running application.
- **Data adapters** enable users to create high-quality API access to data from disparate external sources

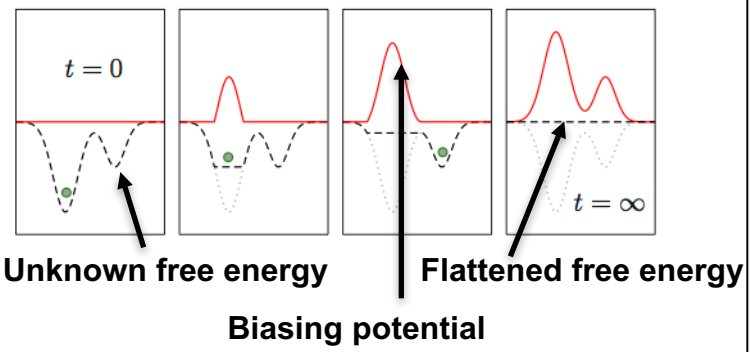


EARLY ADOPTERS

- ETL pipelines in the SD2E platform
- Scheduling containers on an elastic cloud in the IPT on the Web gateway
- Automatically creating Singularity images from new Docker images in the Biocontainers project.
- Scaling out pleasantly parallel Opensees function calls within an engineering Jupyter notebook

SI2-SSE: Enhanced Software Tools for Biomolecular Free Energy Calculations

Celeste Sagui and Christopher Roland
Department of Physics, NC State University



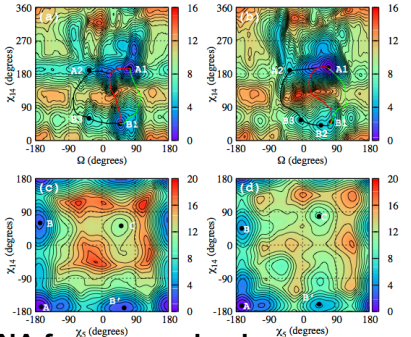
- The free energy is the single-most important quantity describing biomolecular systems at equilibrium
- Biomolecular free energy surfaces are rugged and complex requiring very long time scales to explore
- To address issue, we developed the Adaptively Biased Molecular Dynamics (ABMD) method with belongs to the general category of umbrella sampling with methods with time-dependent potential [Babin et al, JCP 2008]

- ABMD calculates Landau free energies as a function of suitably chosen collective variables
- Implemented with multiple walker and replica exchange enhancements
- Implemented along with Steered Molecular Dynamics (SMD) in AMBER software suite

- Our SEE program aims to develop software tools around this program
- Current extensions:

1. Port ABMD suite from SANDER to PMEMD in AMBER v.16 +
2. GPU compatable
3. Introduce interacting multiple walker algorithms for enhanced sampling
4. Well-tempered ABMD
5. Driven-ABMD with combines SMD and ABMD
6. Introduce swarms-of-trajectories string method (STSM) for calculating minimum free energy path (MFEP)

- Our group has applied methodolgy to variety of biomolecular systems: small peptides, sugar puckering, polyproline systems, guest-host systems, polyglutamine systems, DNA/RNA
- E.g., investigation of hairpin loops and stems for CAG and GAC DNA/RNA systems as implicated in trinucleotide repeat expansion diseases



RNA free energy landscapes as a function of two collective variables – see Pan et al, Biophysical J., 2017 for details

What is OpenAtom

Sohrab Ismail-Beigi
Applied Physics
& Materials
Yale

Sanjay Kale
Computer Science
UIUC

Glenn Martyna
Physical Chemistry
& Materials
IBM

NSF SI2-SSI: Scalable, Extensible, and Open Framework for Ground and Excited State Properties of Complex Systems

NSF ACI 1339804 & 1339715

- Quantum simulation of materials and molecules (DFT)
- **OpenAtom** software package: molecular dynamics now, excited electrons in progress (GW)
- Plane waves and pseudopotentials
- charm++ parallel infrastructure

Mining Seismic Wavefields (EAR-1551462)

PI: Gregory C. Beroza (Stanford); Co-PIs and Contributors: Clara Yoon, Karianne Bergen, Kexin Rong, Hashem Elezabi, Peter Bailis, Philip Levis (Stanford), Yehuda Ben-Zion, Haoran Meng, Philip Maechling , John E. Vidale (USC), Egill Hauksson, Zachary Ross (Caltech), Zhigang Peng, Zefeng Li (Georgia Tech)

Scientific Challenge: Develop new methods and software to detect weak and unusual seismic events that currently go unreported.

Solution Approach: We have developed multiple techniques that use waveform similarity to detect and identify previously undetected signals.

The “large-T” approach: Use waveform similarity of multiple events over long periods of time (T).

The “large-N” approach: Use waveform similarity of single events as recorded on dense seismic arrays featuring a large number (N) of stations

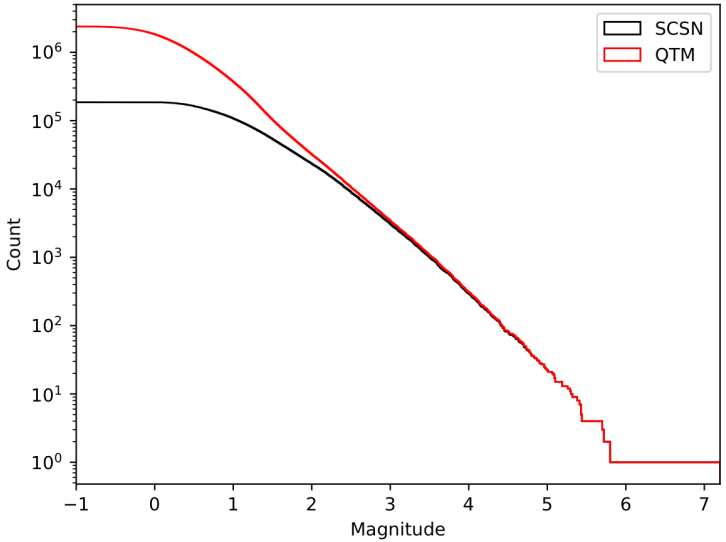


Fig. 1. Using a template matching procedure (Large-T method), we reprocessed the SCSN continuous waveform archive using the seismograms of about 300,000 previously recorded earthquakes as templates which took about 1 million GPU hours. This method identified 2.4 million earthquakes for the period 2008-2017, which is a 13 times increase over the standard SCSN regional catalog.

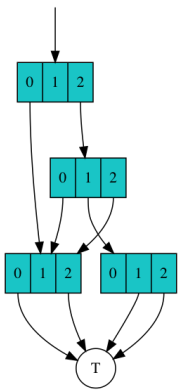
MEDDLY: Multi-terminal and Edge-valued Decision Diagram LibrarY

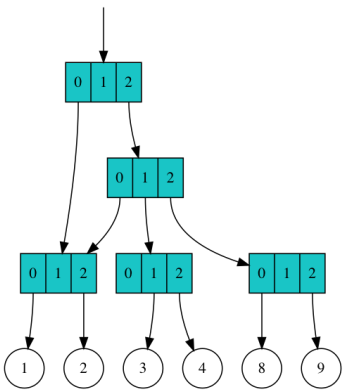
What is MEDDLY?

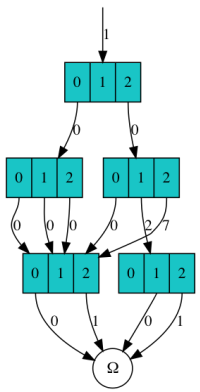
- Library for decision diagrams (BDDs, MDDs, EVMDDs)
- Graph data structures
- Represent functions over discrete variables
- Can be very compact
- State-of-the-art algorithms

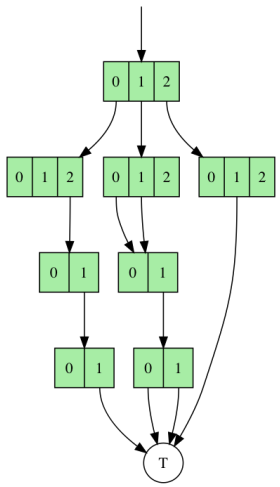
Why use MEDDLY?

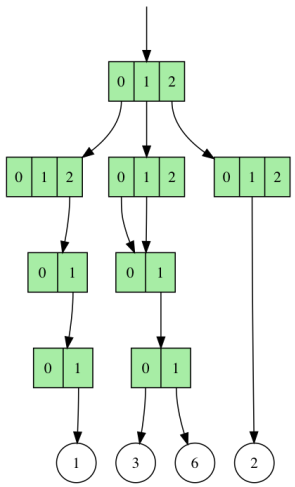
- Digital logic verification (BDDs)
- Model checking (BDDs, MDDs)
- Counterexamples (EVMDDs)
- Integer constraint problems
- Applications needing large sets of vectors of integers

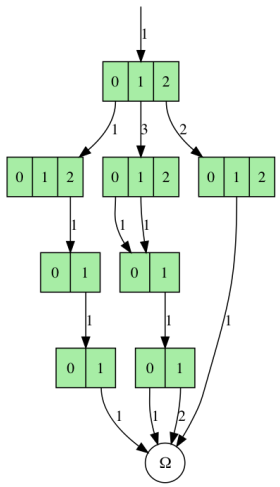












Larch: Integrating Synchrotron X-Ray Analysis Methods

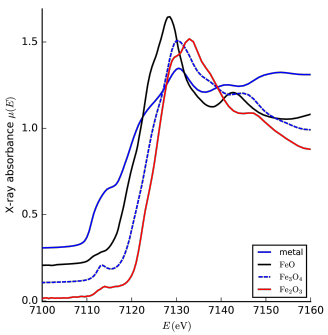
Larch provides an open-source set of libraries and applications for synchrotron X-ray methods

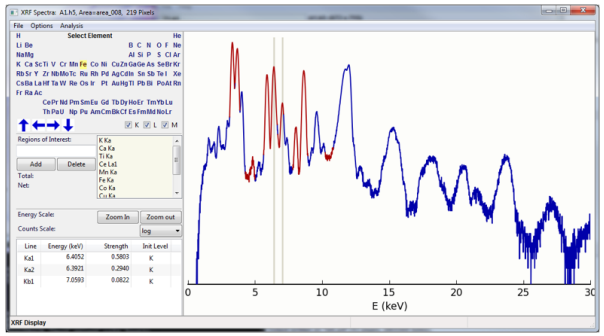
Synchrotron beamlines like X-ray microprobes produce *multi-modal data* including X-ray absorption spectroscopy, fluorescence imaging, X-ray diffraction, diffraction imaging, and tomography.

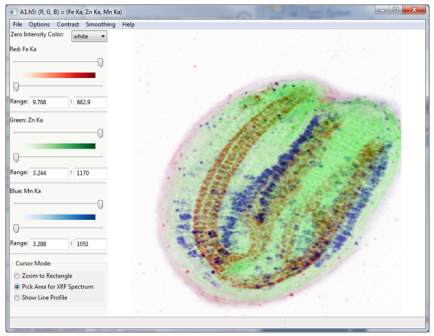
Larch provides a coherent, tested, documented, and extensible library for treating these datasets.

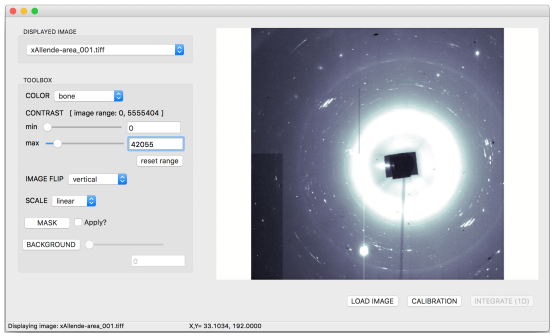
It uses the scientific Python stack, and provides GUI apps for visualization and analysis of X-ray spectroscopies and related methods. This gives a shallow entry for non-expert scientists users, and also full scripting capabilities for more experienced users.

Synchrotron users run thousands of experiments per year, producing complex and heterogeneous datasets in a wide variety of fields.









Distributed MultiThreaded Checkpointing (DMTCP)

NSCI: SI2-SSE: Extensible Model to Support Scalable Checkpoint-Restart ...
Save running computation and restart (possibly on different computers)

Several Use Cases:

1. ***Checkpoint Many Times / Restart Once:*** Save and restart long-running computation (in case of computer crash); **EXAMPLE: MPI for HPC**
2. ***Checkpoint Once / Restart Many Times:***
   Execute long-running initialization once and checkpoint —
   Then restart many times
   EXAMPLE (Emulation of new CPU chips by Intel):
   Start an operating system using hardware emulator, checkpoint, and restart many times to check common applications (e.g., office suite)
3. ***Checkpoint Once / Restart Many Places:***
   EXAMPLE (formal verification): Explore many paths in parallel

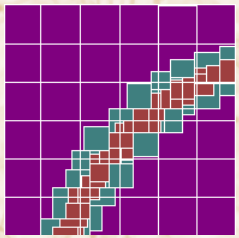
DMTCP transparently supports a wide range of environments:
MPI, InfiniBand, CUDA/GPU, Distributed Software, ...

Empirical Methods for Computational Science

Tim Menzies

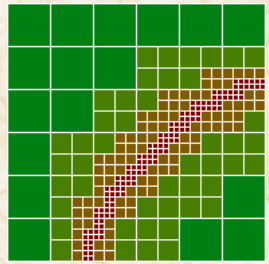
Petascale astrophysics, software infrastructure development, and community engagement

1. ENZO



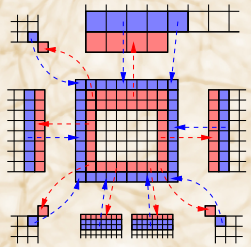
community developed
structured AMR
astrophysics & cosmology
powerful but scaling issues

2. Enzo-P



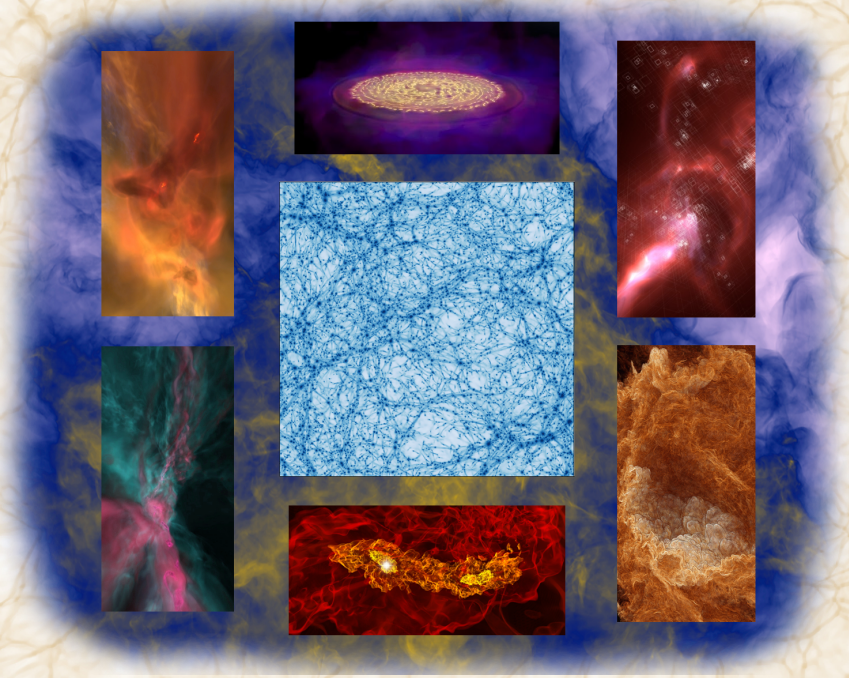
"Petascale Enzo"
ENZO physics
redesigned AMR
Cello framework

3. Cello

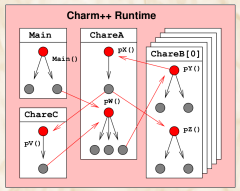


highly scalable
reusable framework
array of octrees
uses Charm++

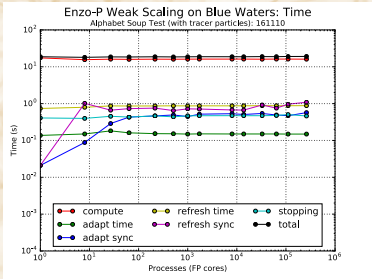
4. Charm++



parallel programming system
data-driven / asynchronous
targeting Exascale apps.

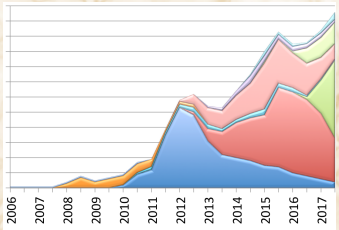


5. Scaling



256K BW cores
hydro+tracer
$\approx$ 98% || effic.
fully distributed
>> ENZO

6. Next phase



Exascale Enzo-E
strong scaling
heterogeneous

Synthesizing Self-Contained Scientific Software

Hassen Saidi and Ashish Gehani

Problem: Reproducibility in scientific computing depends on the ability to run the exact same analysis. Any change in the environment, makes consistency a challenge.

Wholly!: Tool to build reproducible, portable, minimal, and self-contained software packages. Specify environment to minimize external dependencies.