RAPIDS Data Understanding Highlights

D. Morozov, P. Balaprakash, J. Ahrens, U. Ayachit, W. Bethel, E. Brugger, A. Buluc, B. Geveci, P. Grosset, H. Guo, C. Harrison, W. Liao, S. Madireddy, J.H. Park, J. Patchett, T. Peterka, S. Philip, D. Pugmire, R. Sadre, O. Ruebel, H.-W. Shen, C. Steed, G. Weber, S. Yoo

Data Reduction for Large Scale Ensemble Cosmological Simulations

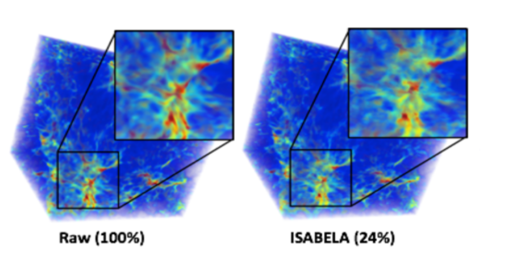
Scientific Achievement
- Enable scientists to reduce the storage space requirement when running large ensemble simulations, while still make it possible to perform full scale simulation parameter exploration for post-hoc analysis
- Enable scientists to compress particle data from large scale N-Body cosmological simulations at a controllable space-quality tradeoff while preserving essential domain

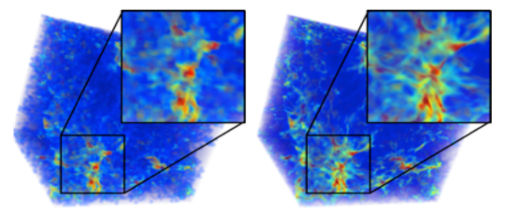
Significance and Impact
- Using GMM-based statistical signatures, it is possible to save only a small portion of data from large scale ensemble run. Post-hoc analysis is done by reconstructing simulation output of novel parameters from the statistics signatures. Experiments shown that the space saving can be more than 99%
- GMMs are shown to be effective to represent particle clusters in cosmological simulations. The reconstructed data from GMMs show very high accuracy in domain specific metrics such as Halo Mass Functions and Power Spectrum, when data are compressed to 1/200 of their original size.

Research Details

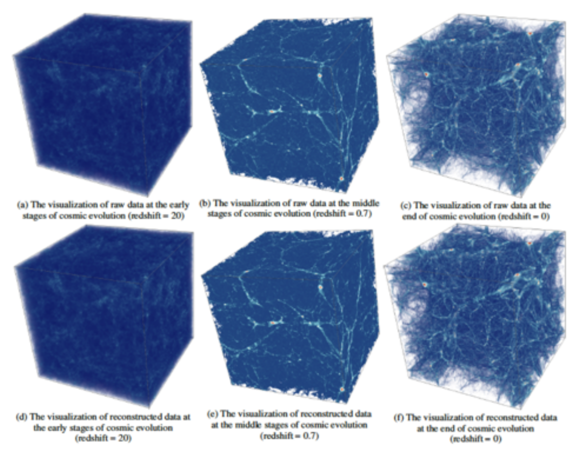
Ensemble Data Exploration
- Store a small number of simulation results at full resolutions into a code book as prior knowledge
- Down sample the remaining data into GMMs as the statistical signatures
- Data at an arbitrary parameter configuration can be reconstructed from the prior knowledge and the statistical signatures
- The priori knowledge only takes 0.44% of the original data for a cosmology simulation using Nyx

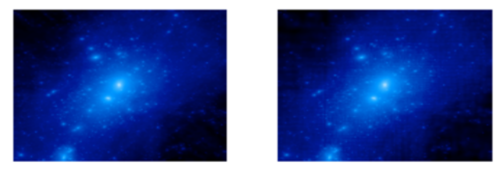
GMM-based Particle Compression
- A k-D partitioning is employed based on the GMM quality requirement and the desired final space consumption
- Spatial GMMs are used to transform the particle data into Gaussian Mixtures. Domain specific metrics are used to verify the quality, and an iterative refinement algorithm is used to adjust the paritions of particles and number of Gaussians.





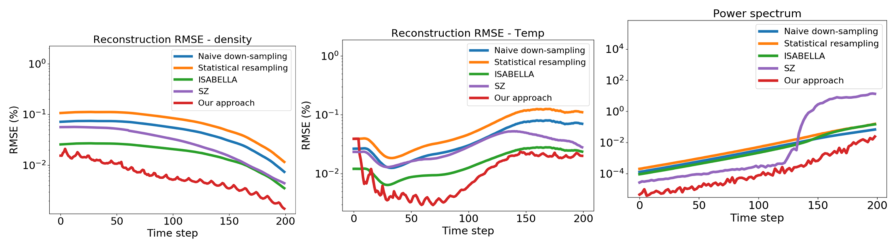
Images produced by our super resolution representations





Above: comparisons between the raw data (top row) and GMM reconstructed data (bottom row).
Below: comparisons of rendering from raw data (left) and our method (right). The space saving is 200 times.

Quantitative Evaluation: Comparing our method with four different compression methods



Quantitative Image Analysis Using Deep Learning

Scientific Achievement

Adaptation of a state-of-the-art deep learning-based image segmentation method enables feature detection in noisy data from an atomic force microscope.

Significance and Impact

The new image analysis pipeline, which includes the DL-based method, will automate a previously manual analysis methodology, enabling more rapid understanding of experimental data.

Research Details

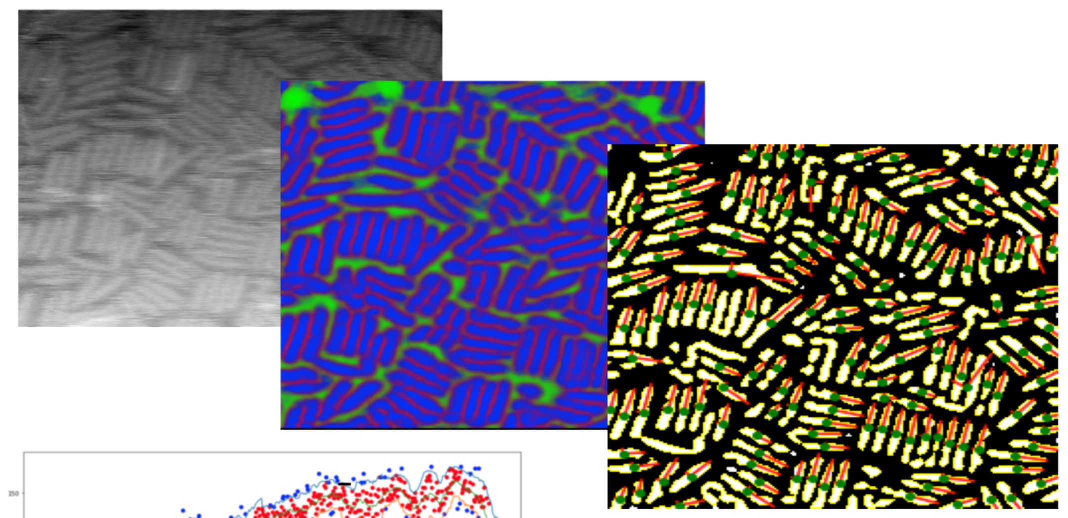
We adapt a deep learning network, U-net, for use in finding features in noisy data from an atomic force microscope, part of the IDREAM EFRC.

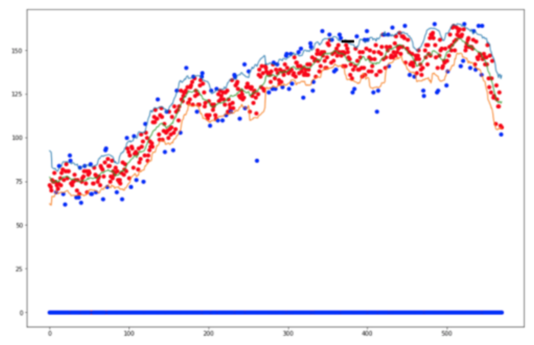
The new process is capable of finding, tracking, and analyzing hundreds-thousands of features in image sequences. Previous results use manual analysis of a handful of features.

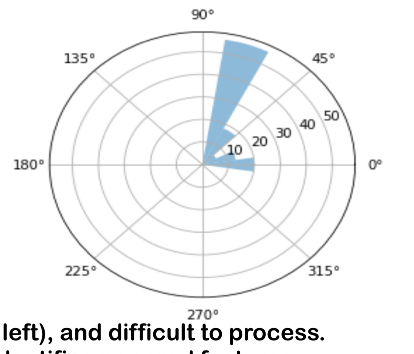
Quote from stakeholder J. De Yoreo: "This is exactly what we need to crack open a bunch of [challenging scientific] problems."

RAPIDS DU Personnel: O. Rübel, T. Perciano, R. Sadre, W. Bethel.

IDREAM EFRC Personnel: J. De Yoreo, S. Zhang

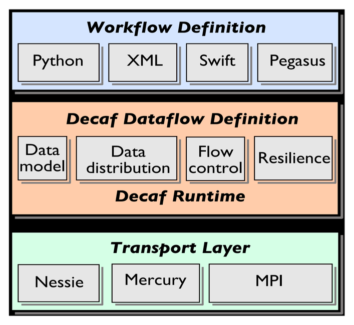




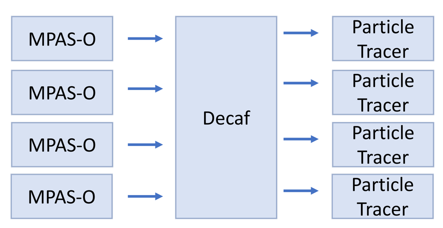


Figures: Raw data from the AFM is quite noisy (upper left), and difficult to process. A deep-learning based image segmentation method identifies nanorod features (upper middle). After computing nanorod position and orientation (upper right), we produce charts showing nanorod sizes over time (lower left) and a radial histogram showing orientation of rods (lower right).

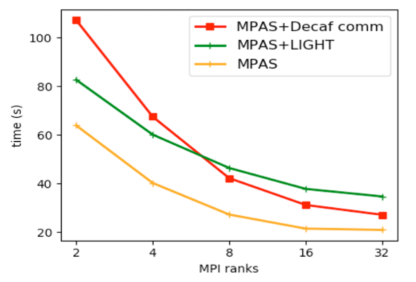
CANGA: Coupling Approaches for Next Generation Architectures



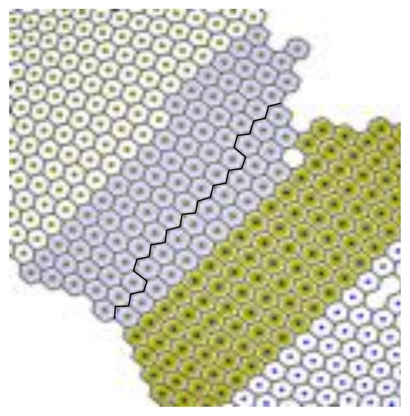
Decaf software organization



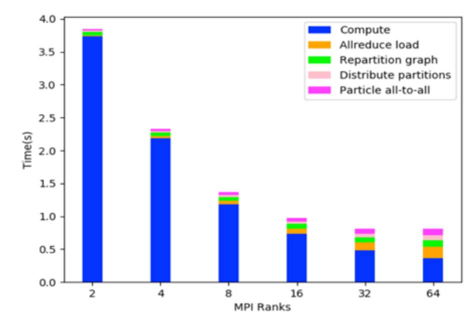
Coupling MPAS-O with particle tracer using Decaf



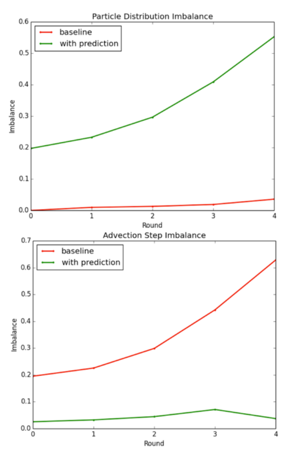
Preliminary timing results for MPAS-Ocean built-in particle tracer (LIGHT), and data movement using Decaf



Constrained graph partitioning for dynamic load balancing particle tracing with unstructured data



Preliminary timing results for constrained graph partitioning with BFS expansion



Imbalance in particle distribution (top) and advection workload (bottom) for synthetic data in prediction and baseline cases

Scientific Achievement

CANGA: New high-performance coupling approaches and capabilities for coupled Earth System Models on next generation computing architectures

Significance and Impact

Coupling external high-performance and load-balanced Lagrangian particle tracing codes with climate models offloads extreme-scale data analysis and visualization. The decoupled workflow based on Decaf simplifies compilation, improves performance, and enables scalable analysis

Research Details
- Decoupled Lagrangian particle tracing for MPAS-Ocean model using Decaf dataflow system
- Load balanced Lagrangian particle tracing based on dynamic, constrained graph decomposition that enables scalable flow analysis and visualization
- Load balancing for unstructured data using graph distance based embedding and constrained k-d tree
- Load balancing for Lagrangian particle tracing using workload prediction based on linear and higher order predictions
- Goal: Develop general methods to achieve better load balancing on unstructured meshes
- Goal: incorporate ensemble and stochastic flow analyses that build upon Lagrangian particle tracing
- Goal: enable coupled data analytics across multiple climate models

[1] Guo et al., IEEE TVCG, 2019 (In Preprint)
[2] Zhang et al., IEEE TVCG, 21(1):954-963, 2018
[3] Zhang et al., IEEE PacificVis, 2018

Work was performed at Argonne under SciDAC CANGA Partnership. Images courtesy of Mukund Raj, ANL.

Mukund Raj, Hanqi Guo, and Tom Peterka (ANL)

Parallel Event Generation and Analysis with DIY

Scientific Achievement

Fermilab researchers developed two HPC parallel codes using the DIY programming model.
- Pythia8 generates Monte Carlo (MC) events [1]
- Feldman-Cousins correction compares experiments with simulated events [2]

Significance and Impact

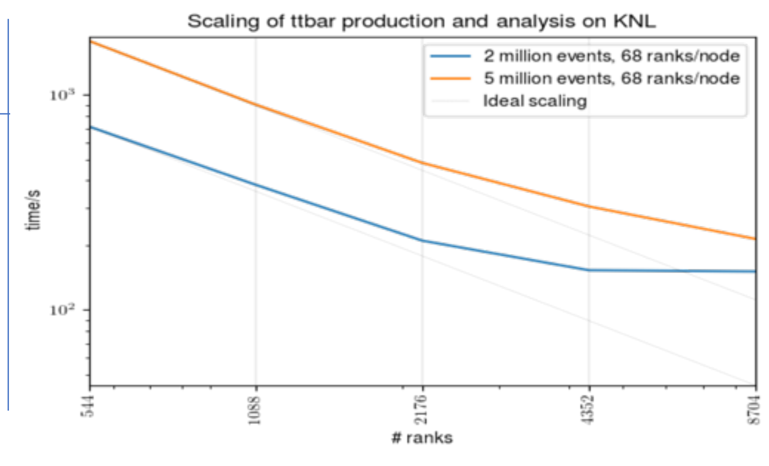
HEP workflows require generating and analyzing vast numbers of MC events. DIY efficiently utilizes HPC resources and HEP community tools.

Research Details

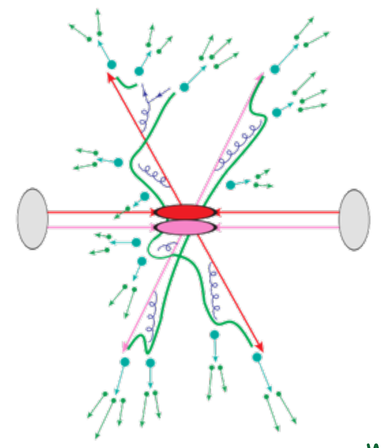
Block parallelism with DIY encapsulates communication in a block-processing application.

Allows for extremely short turn-around of large parameter space explorations (e.g. generator tuning)

Paves the way for new and advanced optimization algorithms, e.g. LHC search analyses.

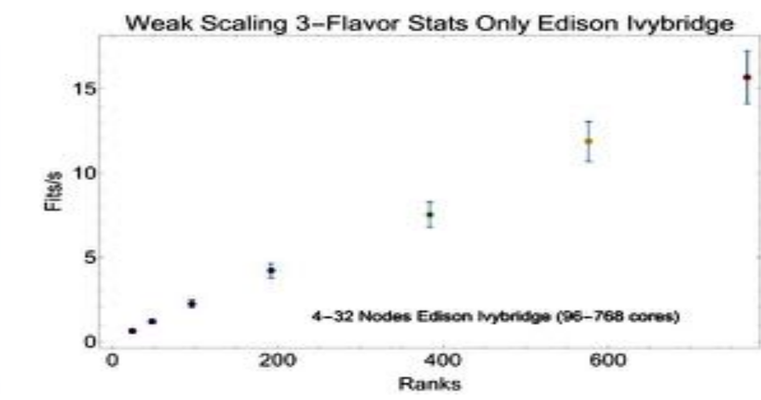


Strong scaling of the HEP's Pythia8 event simulation with ASCR's DIY up to 8704 KNL cores on Cori. The deviation from ideal scaling is due to diminishing work per core at high core counts.



Event generator model for proton-proton collision:

Robust predictions of collider events are needed to search for new physics effects. Much of the dynamics is described by tunable parameters. The calculation of event generator predictions is expensive, and must be done for each choice of parameters. A full detector simulation of these calculations is even more expensive, requiring parallel HPC codes.



Near-perfect weak scaling of Feldman-Cousins DIY code on up 768 Ivybridge cores on Edison.

[1] Hoche et al., *arXiv 2019.*
[2] Sousa et al., *CHEP 2018.*

Work was performed at Argonne and Fermilab under SciDAC HEP on HPC Partnership. Images courtesy of Holger Schulz, U. Cincinnatti and FermiLab

Using Visual Analytics to Understand Neural Network Classifications of Imagery for Genetic Engineering

Scientific Achievement

CrossVis visual analytics capabilities help explain and enhance an Artificial Neural Network (ANN) approach for classifying scanning electron microscope (SEM) imagery.

Significance and Impact

This collaboration demonstrates the advantages of combining interactive visual analytics methods with statistical data analytics to explain and improve artificial intelligence (AI) processes.

Research Details

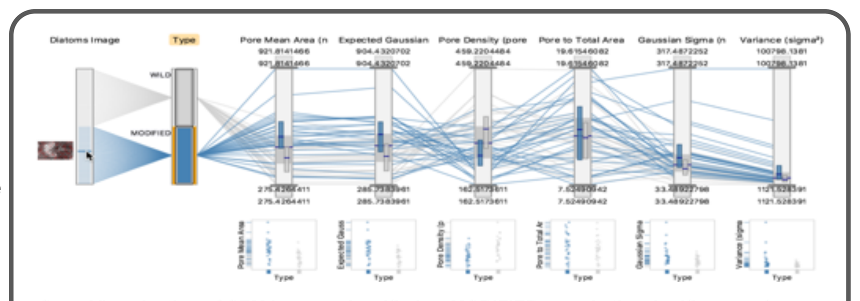
ORNL CNMS scientists used an ANN process to classify whether SEM images corresponded to genetically modified diatoms or not.

Diatoms are unicell alga with significant implications for photonic, filtration, and drug delivery.

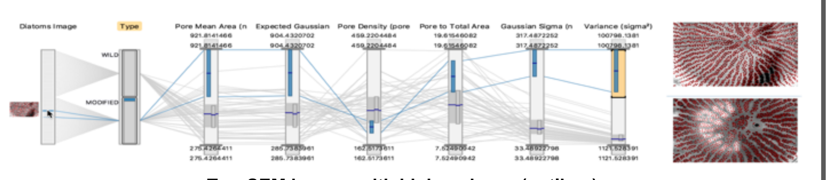
CrossVis interactive visual analysis capabilities reduced the mystery of the "black box" ANN process.

Scientists gained a deeper understanding of the process, improvement ideas, and new trust in AI.

A Nature Partner Journal (npj Comp. Materials) article was recently published (6/13) on this collaborative endeavor.



CrossVis selection of SEM images classified as MODIFIED reveals the significance of pore features.



Two SEM images with high variance (outliers) are visually analyzed for multivariate correlations.

Using Visual Analytics to Explain AI Processes: CrossVis is a visual analytics tool that integrates statistical analytics and an extended version of parallel coordinates to explore large and heterogeneous multivariate data. CrossVis is available at https://github.com/ORNL/CrossVis. (Image Credit: Chad Steed)

Ovchinnikova et al., "Deep Data Analytics for Genetic Engineering of Diatoms Linking Genotype to Phenotype via Machine Learning," npj Comp. Materials, 5:4, 2019. doi:10.1038/s41524-019-0202-3.

Work was performed at Oak Ridge National Laboratory

PI: Chad A. Steed (ORNL)

Visualization of Antarctica LAND Ice

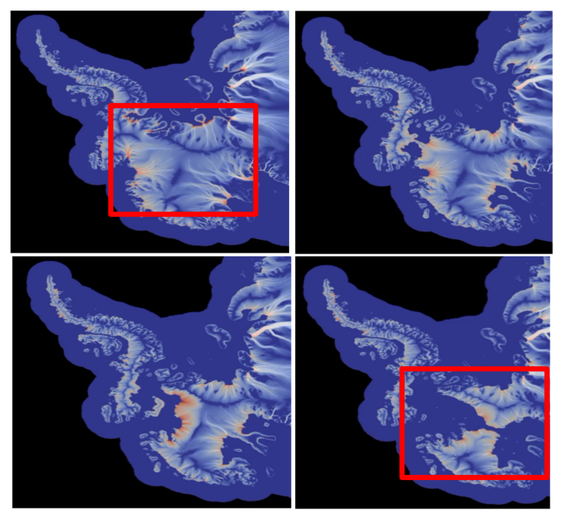
Scientific Achievement

Used time series data, ParaView, and streamlines to show how grounded ice flows and thins on the Antarctic continent in response to ice sheet loss in support of the ProSPect SciDAC.

Significance and Impact

Visualization of key ideas in the science of land ice is key for science understanding within the climate research communities in addition to supporting communication climate science to the general public.

While graphs, such as barcharts, … are useful to show the quantitative implications, they do not show the impact as clearly as visualizations.



Evolution of West Antarctic Ice with an extreme loss of any floating ice over a 200 year period.

Research Details
- Collaborating with the PROSPECT SciDAC to visualize Ice sheet evolution in Antarctica
- Leveraged and improved ParaView, one of the Office of Science tools, for this work
- Improved our ability to support the ProSPect SciDAC with visualizations of polar regions and the ability to represent dynamics of land ice for exploration by land ice scientists.

PI: Jim Ahrens

Performance Analysis for Large Scale Simulations

Scientific Achievement

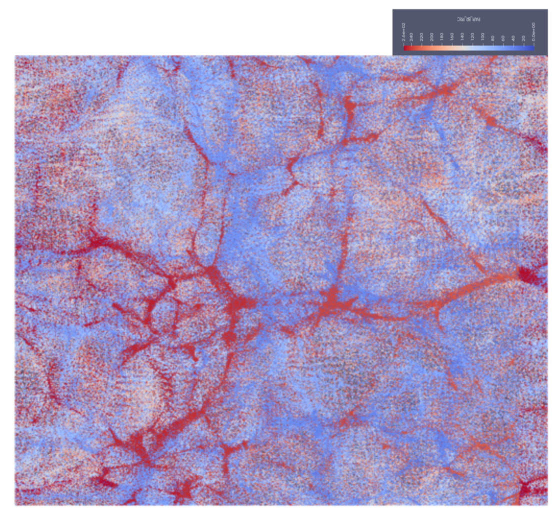
Profiling simulation features instead of function call is a novel way of looking of at simulation performance.

Significance and Impact

Running simulations on supercomputers is expensive (energy usage, time, money, …) but yet essential for better understanding of the world around us.

Profiling tools only allow us to see performance issue related to code

This model can help HPC personnel, software engineers, and scientists better understand simulation performance and figure out how to improve efficiency or simulations by linking performance analysis to simulation features



HACC run using 256 MPI ranks on 16 nodes with papi counter: Conditional branch instructions correctly predicted

Research Details
- Use in situ analysis to look at performance counters instead of simulation features
- Allow users to choose which performance counters are more relevant at different point in time
- Using python for analysis gives users much more freedom for analysis
- Looking into Mochi for easier access to data

Argonne National Laboratory · Brookhaven National Laboratory · Berkeley Lab · The HDF Group · University of Delaware · Kitware · Lawrence Livermore National Laboratory · Los Alamos National Laboratory EST. 1943 · Northwestern University · Oak Ridge National Laboratory · The Ohio State University · University of Oregon · Rutgers · The University of Utah · U.S. Department of Energy · Office of Science