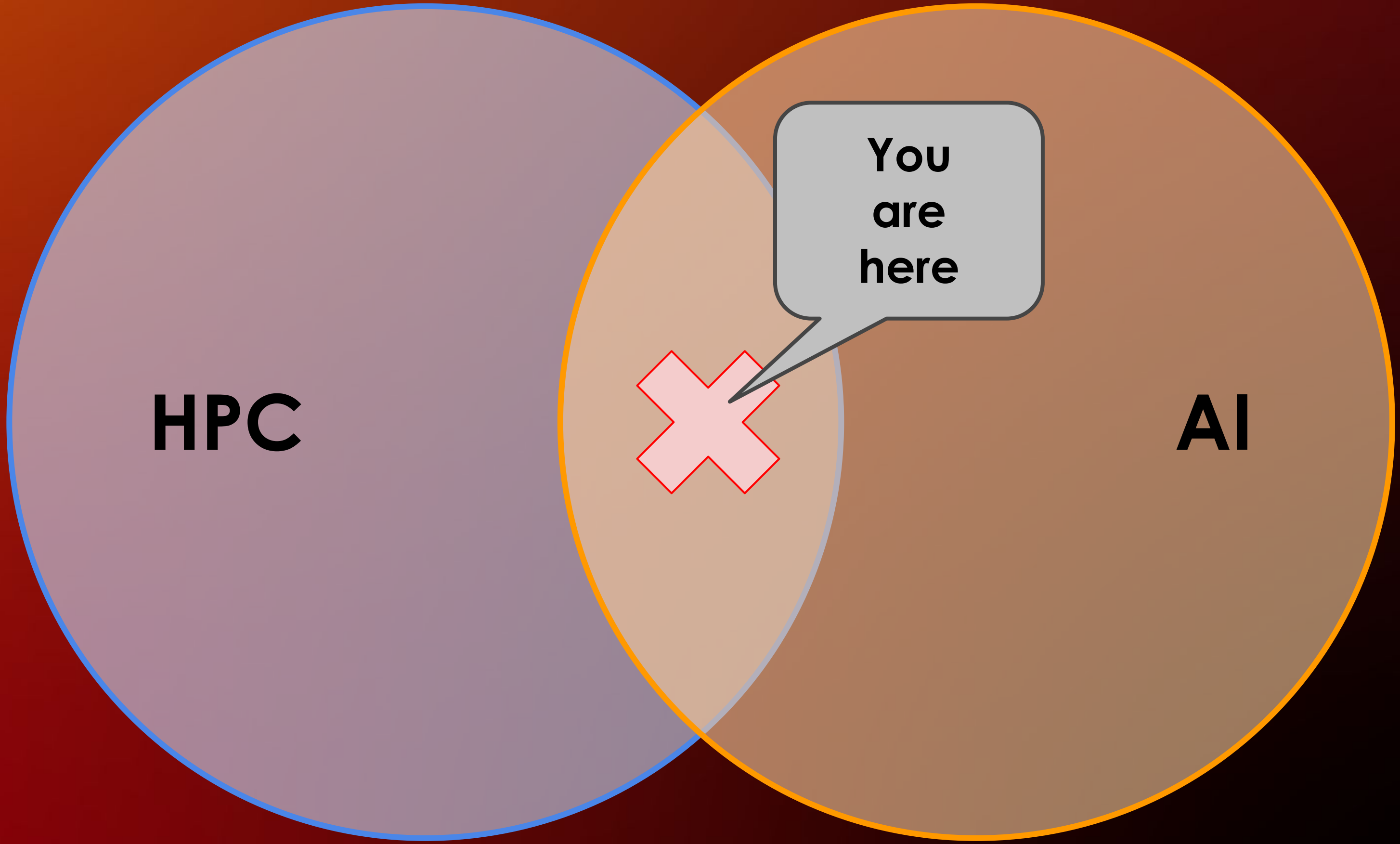
Two Worlds Collide: Trustworthiness and Energy Efficiency for Coupled HPC+AI Simulation Birds of a Feather

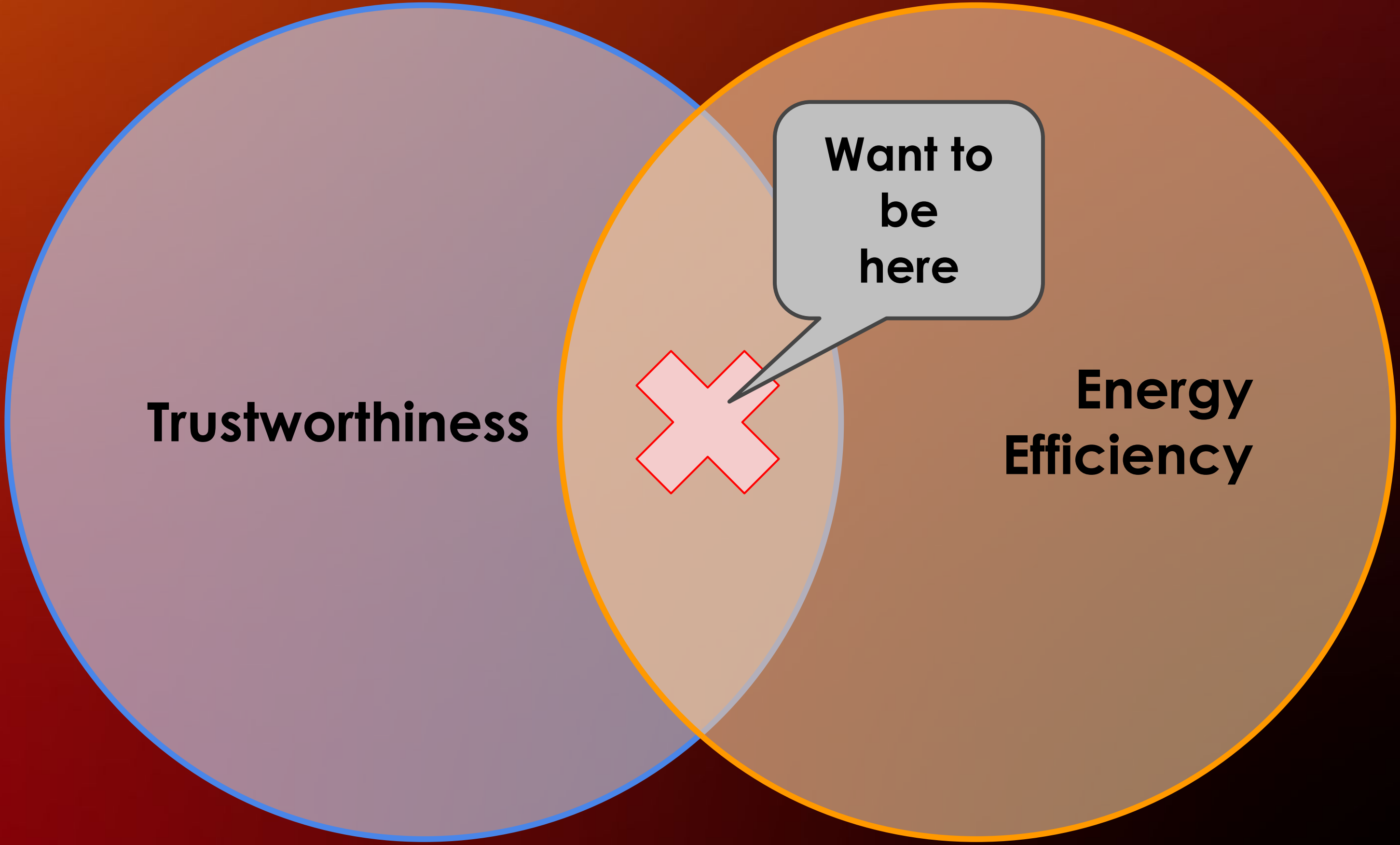
Mark Coletti, *Organizer, Oak Ridge National Laboratory*
Ada Sedova, *Co-Organizer, Oak Ridge National Laboratory*
Venkatram Vishwanath, *Argonne National Laboratory*
Oscar Hernandez, *Oak Ridge National Laboratory*
J. Austin Ellis, *Advanced Micro Devices (AMD)*
Riccardo Balin, *Argonne National Laboratory*
Sanjif Shanmugavelu, *Groq*
Mathieu Taillefumier, *Swiss National Supercomputing Centre*

SC24
Atlanta, GA | hpc creates.

Two Worlds Collide: Coupled HPC+AI Simulation



Two Worlds Collide: Trustworthiness and Energy Coupled HPC+AI Simulation



Please join the conversation

Collaborative white paper on Overleaf:

- https://www.overleaf.com/4196799331zjydbbryznzd#d3f109
  - https://bit.ly/4evnasj

Mailing list to continue conversation:

- https://groups.google.com/g/hpc-ai-two-worlds/about

Agenda

- Introduction
- Brief intros and position statements by panelists
- Open panel discussion
- Invitation for collaboration
  - Mailing list
  - Overleaf-based document

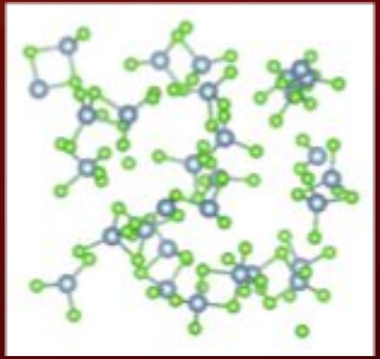
Mark Coletti

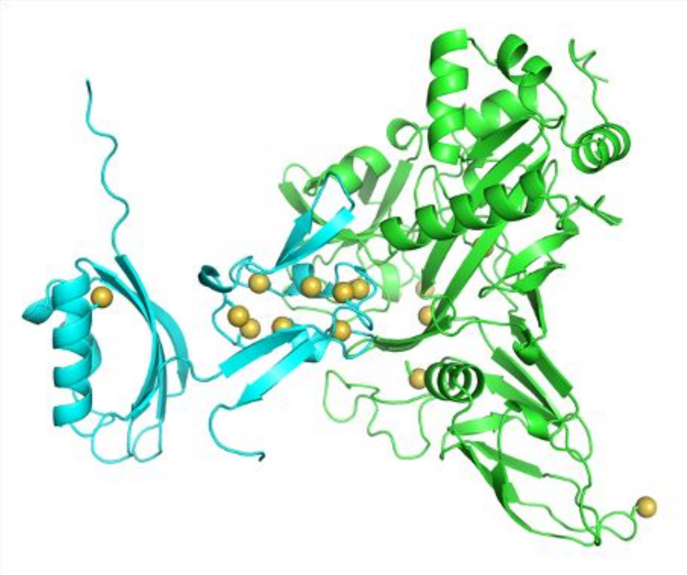
- Computer scientist at ORNL for 9 years
- Used evolutionary algorithms to optimize hyperparameters and neural architectures to maximize accuracy and minimize energy use on DOE's Titan, Summit, Frontier, and Perlmutter systems
- Shifting focus to hardware-based neuromorphic systems that can use considerably less power
- Reduced emphasis on floating point values mitigates floating point related errors
- In several years we should see mass migration to these platforms for HPC
- Will there be new trustworthiness issues? Will new neuromorphic chips actually be energy efficient?

(Video courtesy of Dr. Katie Schuman)

Ada Sedova

- Research scientist in the Molecular Biophysics Group at ORNL
  - Research in HPC, scientific computing, physical chemistry, biophysics, bioinformatics, biochemistry and chemical physics.
- PI of MINNERVVA ORNL project to study reproducibility in deep learning and how it affects simulation
- Also working on performance portability and understanding power consumption in HPC applications
- New projects in materials discovery and biophysics using deep learning

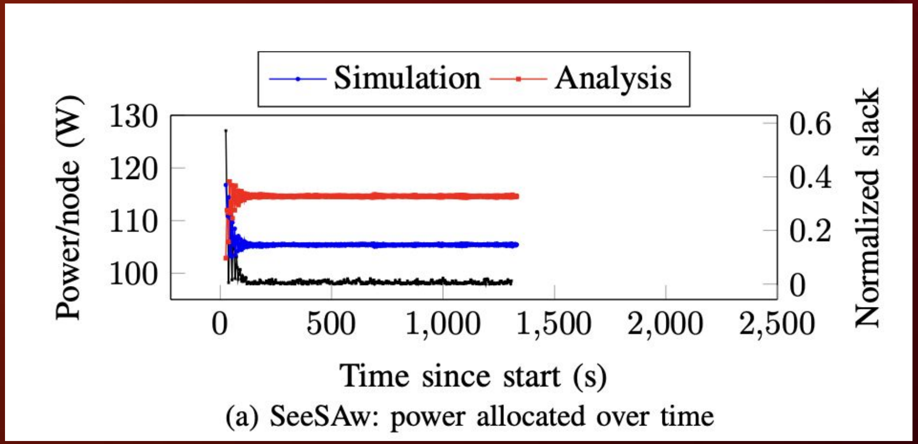




Venkatram Vishwanath

AI ML Lead, Argonne Leadership Computing Facility

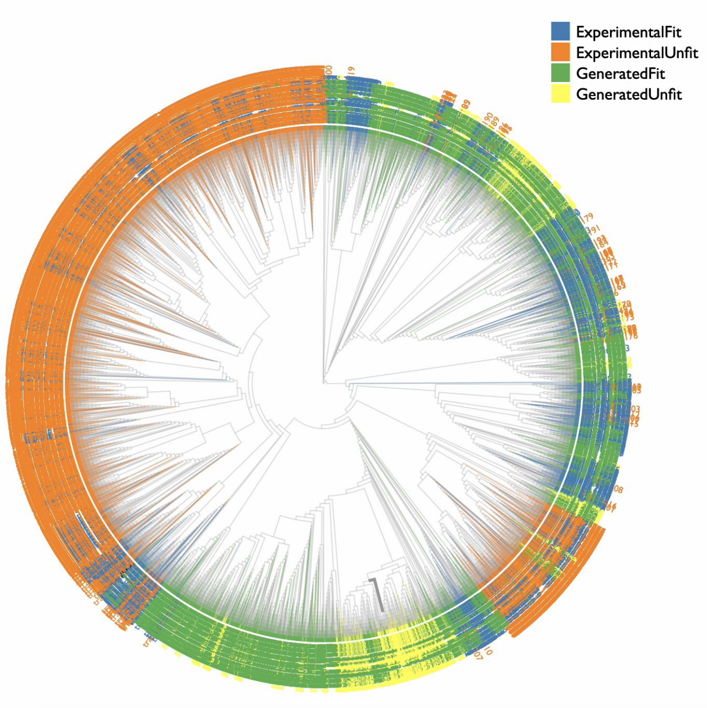
Design of energy-efficient systems for science including telemetry for holistic power-monitoring and knobs for power steering



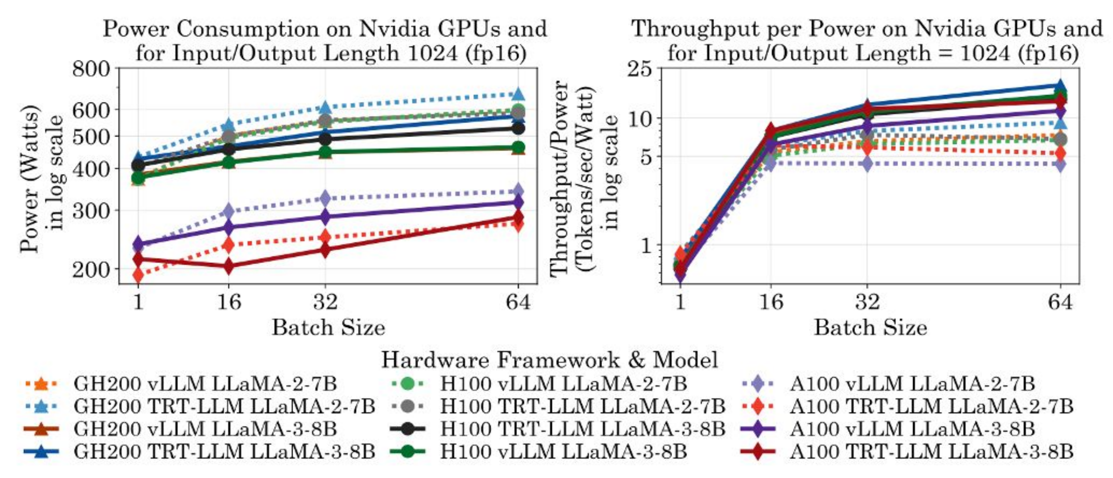
(a) SeeSAw: power allocated over time

Seesaw, Marincic et al., IPDPS’2020

Design of workflows coupling AI, Simulations and Experiments, though spend time fighting challenging correctness issues for applications in production software stacks



MLProt-DPO, Dharuman et al., SC-24



LLAMA-Bench, Teja-Chetty et al.,PMBS@SC’24,

Oscar Hernandez

Senior staff member of the Advanced Computing System Research Section at Oak Ridge National Laboratory

Research and prior work

- Specification and standards for programming models
  - OpenACC, OpenMP, OpenSHMEM
- Communication frameworks - Unified Communication X
- Compiler optimizations and tools
  - Tools ecosystems for performance, debuggers
- Deployment of Leadership class systems - OLCF-X
  - Titan, Summit, Frontier, etc
- Energy Efficient optimizations
- Application-driven exploration of new architectures
  - Testbeds, simulators, etc

J. Austin Ellis

AMD Data Center GPU, Center of Excellence

U.S. DEPARTMENT OF ENERGY
NNSA
Lawrence Livermore National Laboratory
AMD
CRAY
EL CAPITAN

1.74 Exaflops @ 29.6 MW

- Leading AI application engagements for LLMs and training at scale
- AMD AI for Science COE founding member focusing on ML surrogates and low precision strategies

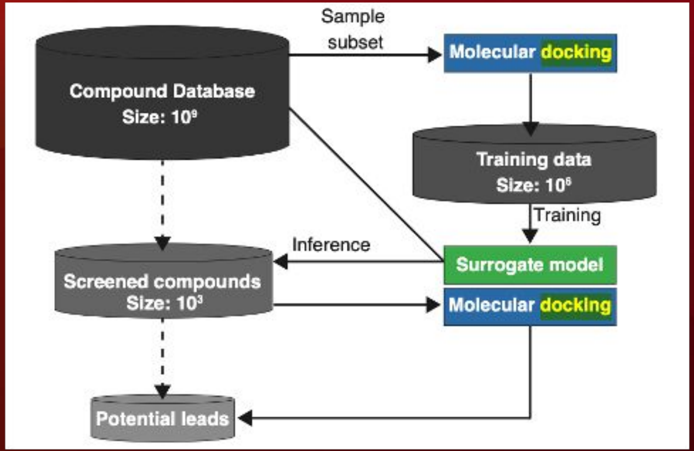
- APU Application Architect for the exascale **El Capitan** system at Lawrence Livermore National Laboratory
- First point of contact for the many LLNL users focused on HPC and AI applications

OAK RIDGE National Laboratory
U.S. DEPARTMENT OF ENERGY
Hewlett Packard Enterprise
AMD
CRAY
FRONTIER

1.35 Exaflops @ 24.6 MW

Accelerating Compound Screening by online Fine-Tuning of LLM Surrogates

Image from Vasan et al. 2024 IEEE IPDPS

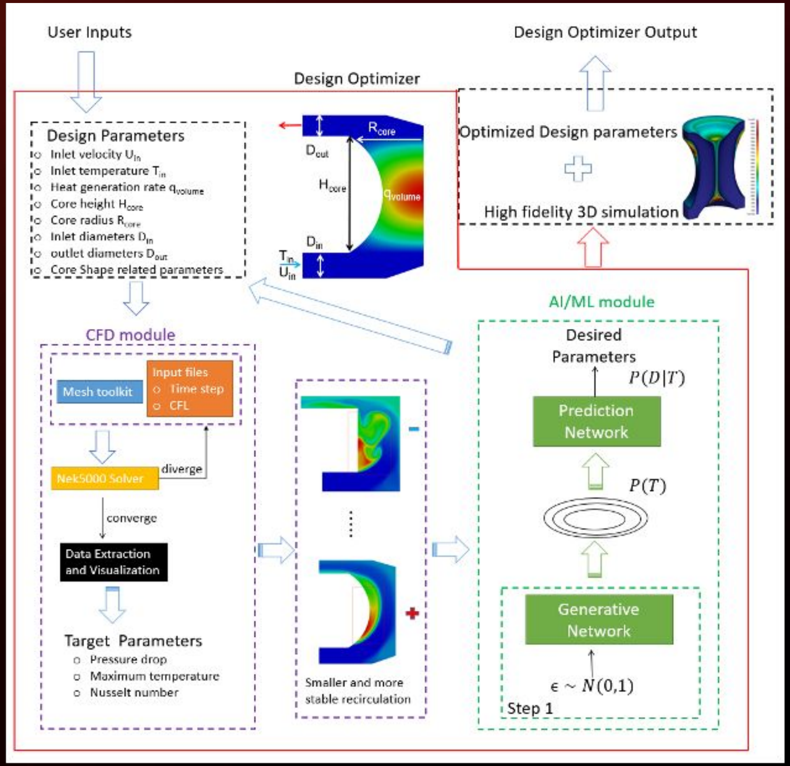


Riccardo Balin

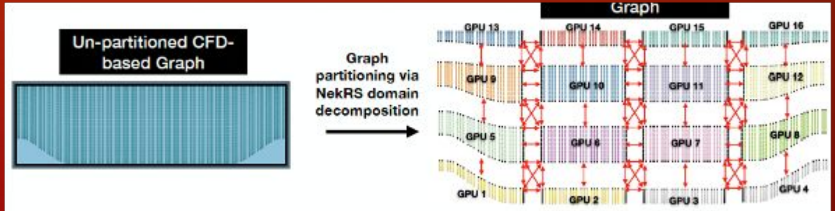
- Assistant Computational Scientist at ALCF
- Help users accelerate science goals through coupled simulation and AI workflows on LCF systems
- Benchmark workflow patterns and tools with focus on scaling and hardware

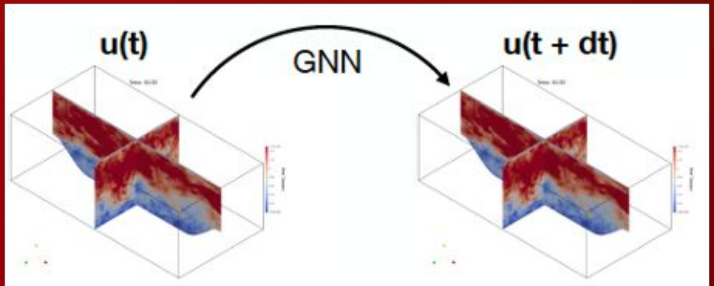
CFD and AI for Molten Salt Reactor Design

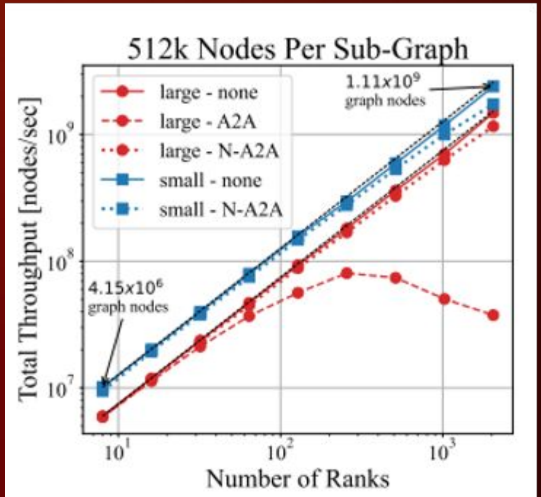
Image from Yiqi Yu, ANL



GNN Surrogates for Extreme Scale CFD



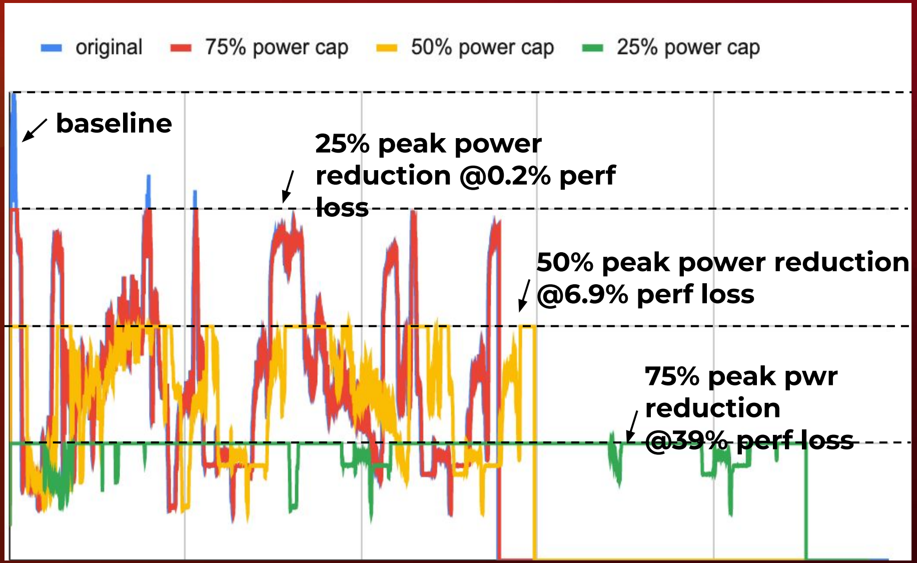




Sanjif Shanmugavelu

- Applications/ Research Engineer at Groq
- Worked on Converged AI & HPC applications with the Groq energy aware compiler
- Formally trained in Statistical Learning Theory, with a focus on ML Robustness
- Recent work includes FPNA impact on ML classification robustness, and device-by-device and run-to-run LLM hallucinations

Groq total Power over time - constraint fixed by the compiler

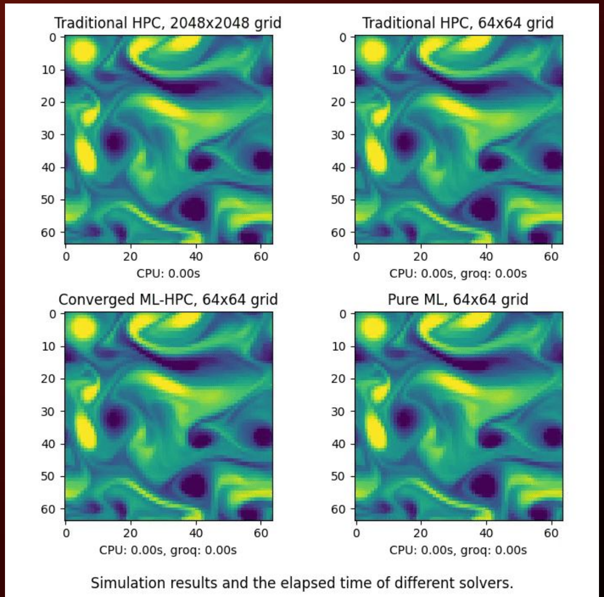


baseline

25% peak power reduction @0.2% perf loss

50% peak power reduction @6.9% perf loss

75% peak pwr reduction @39% perf loss


Traditional HPC, 2048x2048 grid

CPU: 0.00s

Traditional HPC, 64x64 grid

CPU: 0.00s, groq: 0.00s

Converged ML-HPC, 64x64 grid

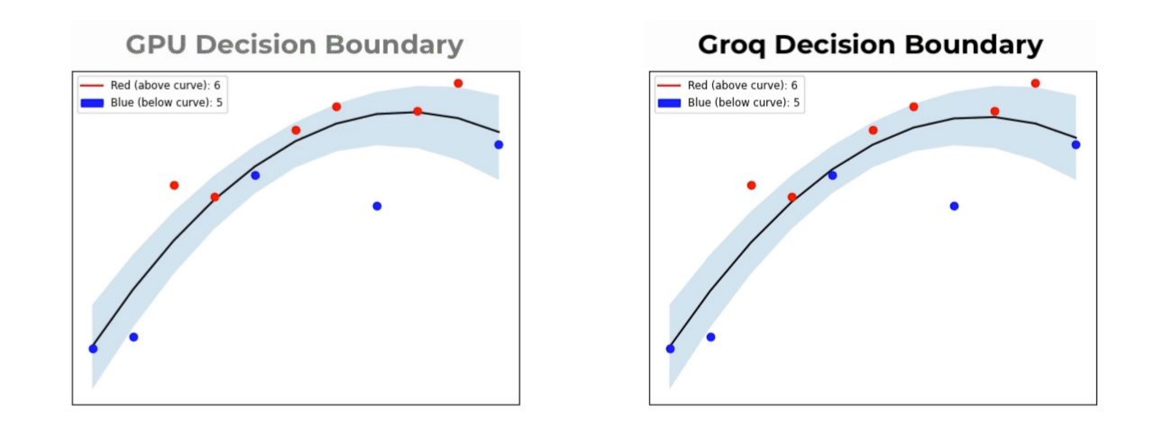
CPU: 0.00s, groq: 0.00s

Pure ML, 64x64 grid

CPU: 0.00s, groq: 0.00s

Simulation results and the elapsed time of different solvers.

Machine learning accelerated computational fluid dynamics (PNAS 2021), running on Groq. Implementation largely credited to Chenyu Zhang

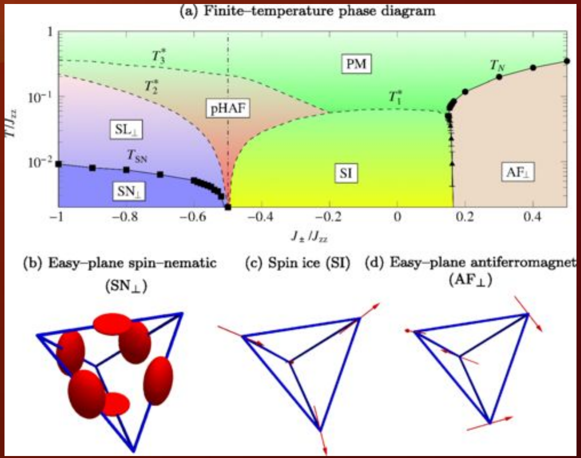

GPU Decision Boundary

Groq Decision Boundary

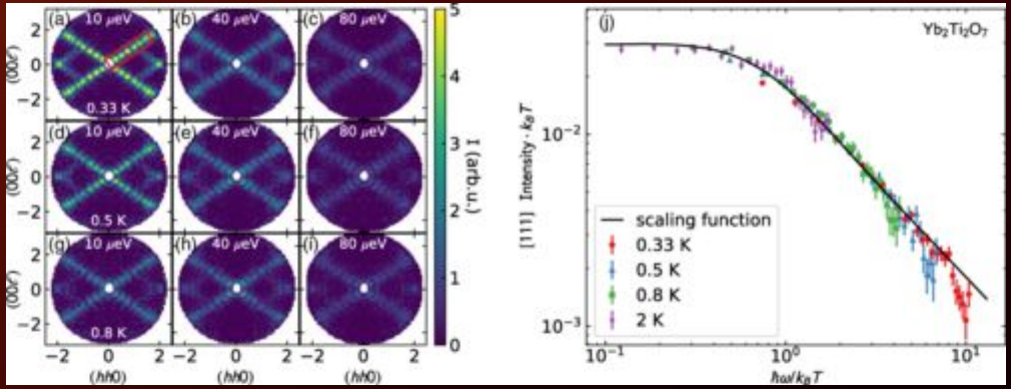
Impact of FPNA on reproducibility for HPC and DL applications (SC24, Correctness workshop)

Mathieu Taillefumier

- Theoretical condensed matter physicist
- Computer scientist / Research software engineer at ETH Zurich / CSCS
- background in magnetism, transport, cold atom physics
- Recent work include impact of FPNA on HPC workloads besides physics



phase diagram of the XXZ model on the pyrochlore lattice extracted from classical monte carlo calculations. (Phys. Rev. X 7, 041057)



Comparison between neutron scattering experiments (ORNL) and theory illustrating the concept of multi-phase competition in $Yb_2Ti_2O_7$ (Phys. Rev. Lett. 129, 217202)

“

BoF Discussion

”

Correctness: What is our new standard of correctness? How do we deal with bugs?

Scientific applications and their requirements for correctness/trustworthiness and how this relates to needs and metrics

- How do ideas about accuracy and correctness differ between traditional HPC simulation and applications that use AI like deep learning?
  - Taxonomy/classification of scientific uses of AI and the associated needs
    - What are the differences in the way that data science treats machine learning vs the way science needs it to be?
  - Disconnects between the two fields in terms of knowledge, software, the problems themselves
  - Hallucinations, black box (lack of analytical estimates of convergence, lack of numerical analysis methods for errors), floating point issues, sensitivity
- Testing approaches
  - How does extensive runtime testing increase power/energy use?
  - What support infrastructure can we build to track and understand bugs and expected accuracy?
- Reproducibility problems
- What support infrastructure can we build?

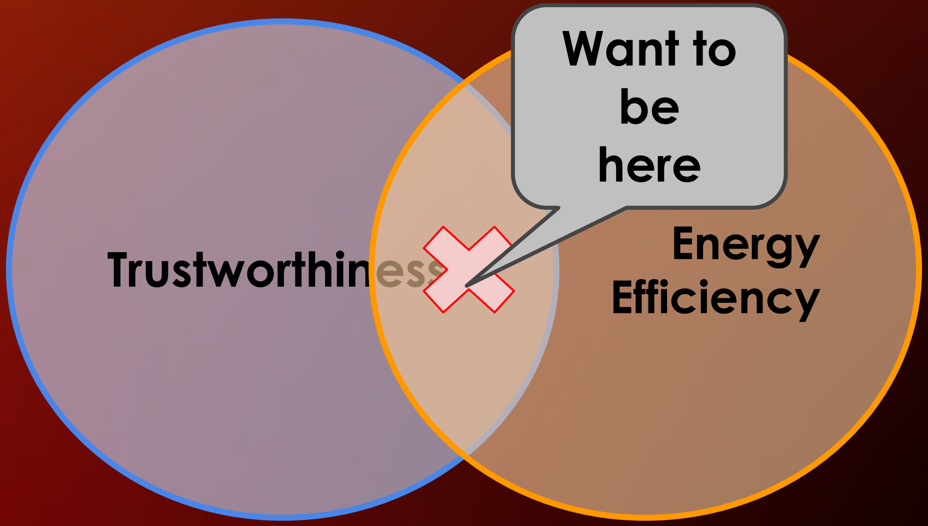
Energy Efficiency

- What metrics are the best? How well do they capture the full training+inference pipeline?
- Granularity of energy use measurement — it is useful to know where in the model using the most energy
- Potential metrics: Inferences per training watt: how often does that same model actually get reused
  - Inferences per training watt as a form of debt – the more a model is used, the more the debt is paid
  - Generalizability, how much UQ is required, etc.
- Specific for AI-augmented simulation: how does the AI really help us to save energy?

Correctness and Energy Efficiency

Ensuring correctness consumes energy. Do we need to reduce our requirements for accuracy and testing to reduce energy consumption?

What solutions can enable us to have the best of both?



Please join the conversation

Collaborative white paper on Overleaf:

- https://www.overleaf.com/4196799331zjydbbryznzd#d3f109
    - ○ https://bit.ly/4evnasj

Mailing list to continue conversation:

- https://groups.google.com/g/hpc-ai-two-worlds/about