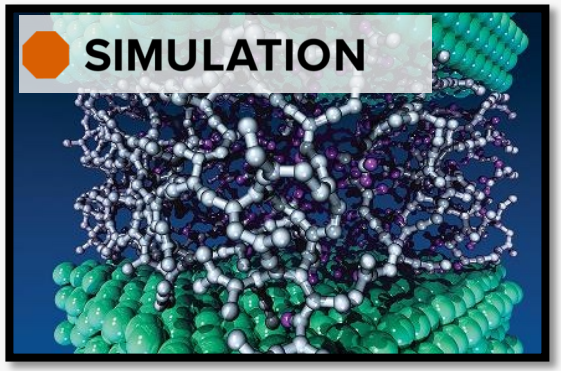
Accelerating Scientific Machine Learning with SambaNova DataScale SN30 at ALCF

Murali Emani,
Argonne Leadership Computing Facility
memani@anl.gov

Argonne Leadership Computing Facility

The Argonne Leadership Computing Facility provides world-class computing resources to the scientific community.

- Users pursue scientific challenges
- In-house experts to help maximize results
- Resources fully dedicated to open science



Architecture supports three types of computing

§ Large-scale Simulation (PDEs, traditional HPC)

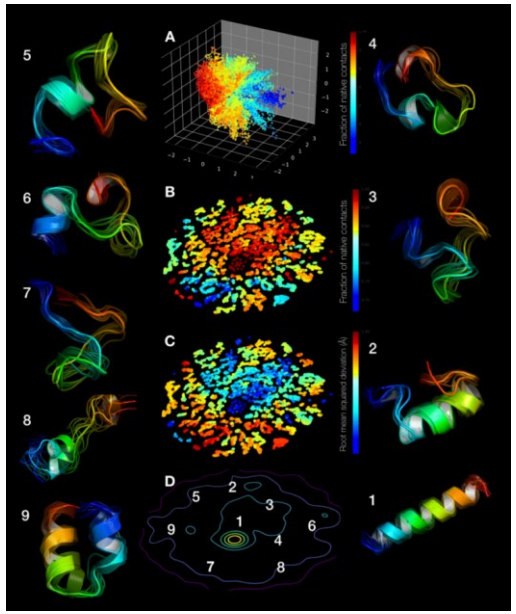
§ Data Intensive Applications (scalable science pipelines)

§ Deep Learning and Emerging Science AI (training and inferencing)

ALCF offers different pipelines based on your computational readiness. Apply to the allocation program that fits your needs.

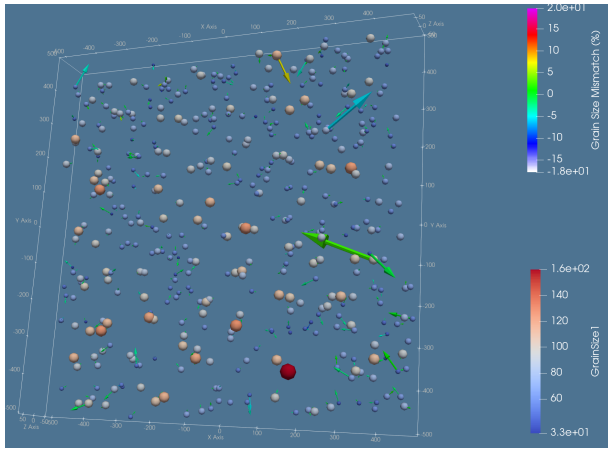
Surge of Scientific Machine Learning

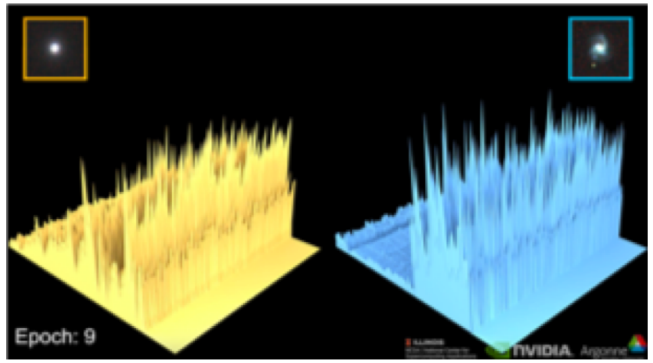
- Simulations/ surrogate models
  - Replace, in part, or guide simulations with AI-driven surrogate models
- Data-driven models
  - Use data to build models without simulations
- Co-design of experiments
  - AI-driven experiments



Protein-folding

Braggs Peak





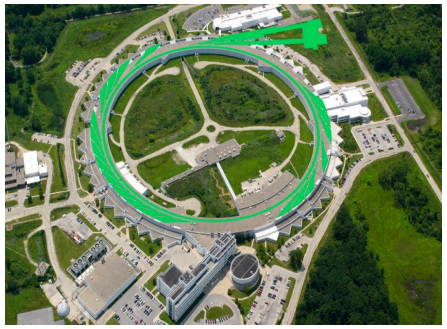
Galaxy Classification

Design infrastructure to facilitate and accelerate AI for Science (AI4S) applications

Integrating AI Systems in Facilities

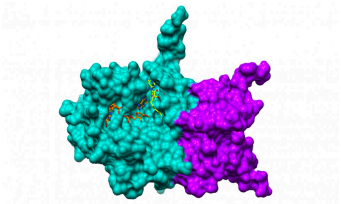
Computing Facility

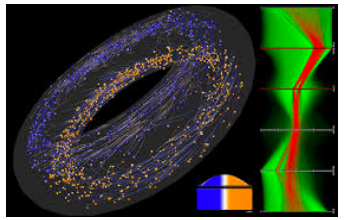
Experimental Facility



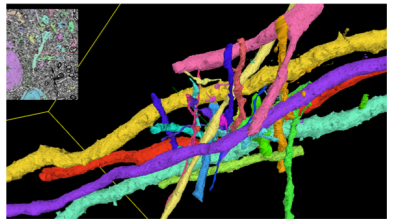
AI-Accelerators

AI-Edge accelerator





Simulations



Data-driven Models

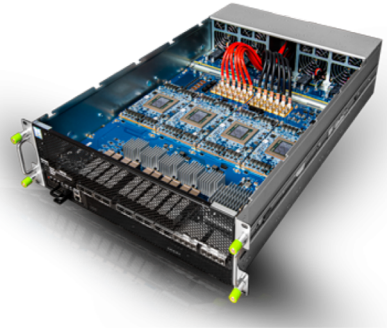
ALCF AI Testbed

https://www.alcf.anl.gov/alcf-ai-testbed

Cerebras CS-2

SambaNova DataScale SN30

Graphcore Bow Pod64



Habana Gaudi1

GroqRack

- Infrastructure of next-generation machines with AI hardware accelerators
- Provide a platform to evaluate usability and performance of AI4S applications
- Understand how to integrate AI systems with supercomputers to accelerate science

Director’s Discretionary (DD) awards support various project objectives from scaling code to preparing for future computing competition to production scientific computing in support of strategic partnerships.

Getting Started on ALCF AI Testbed:

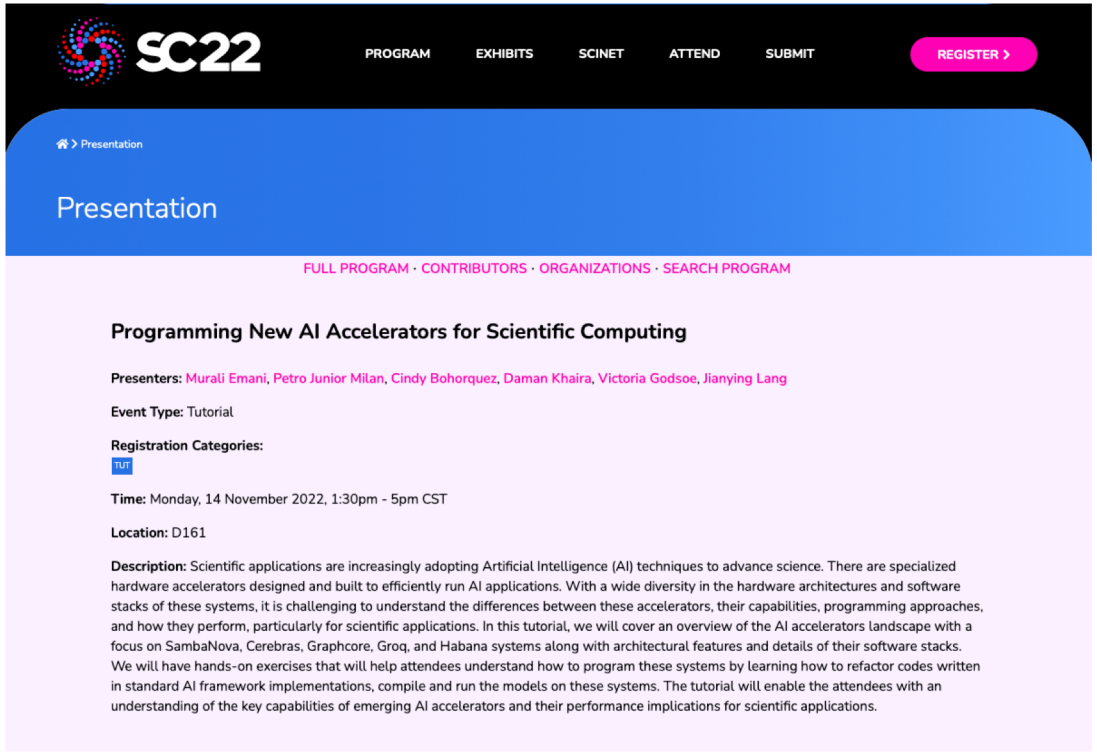
Apply for a Director’s Discretionary (DD) Allocation Award

SambaNova Datascale SN30 is available for allocations

Allocation Request Form

AI Testbed User Guide

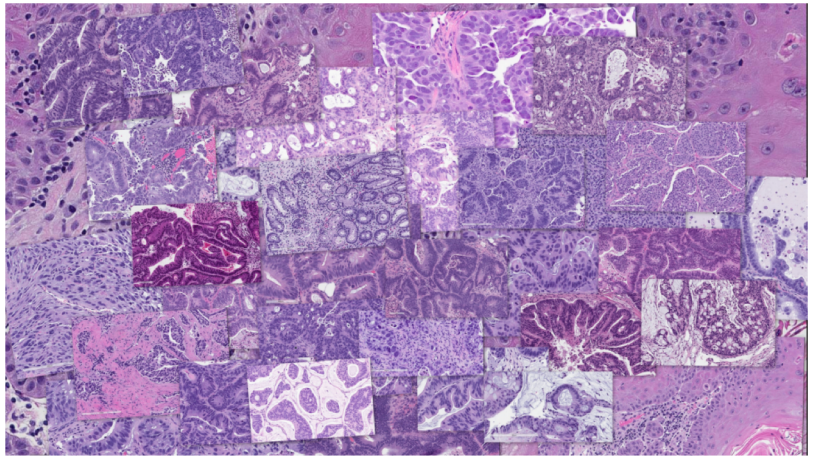
Community Engagement



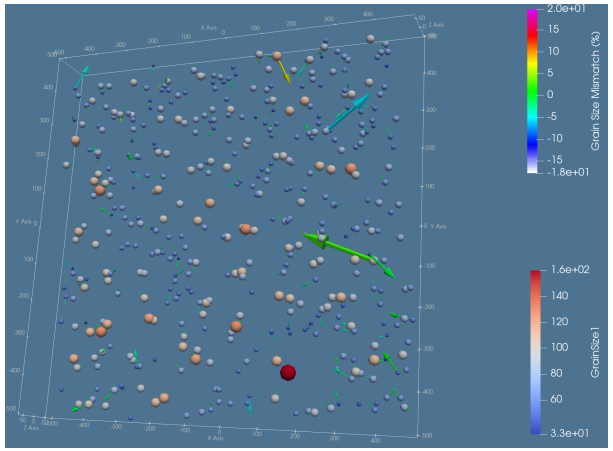
- Regular AI training workshops with SambaNova
- ATPESC H/W Architecture Day
- ALCF AI for Science training series for students

SC'22 Tutorial on Programming AI accelerators for Scientific Computing

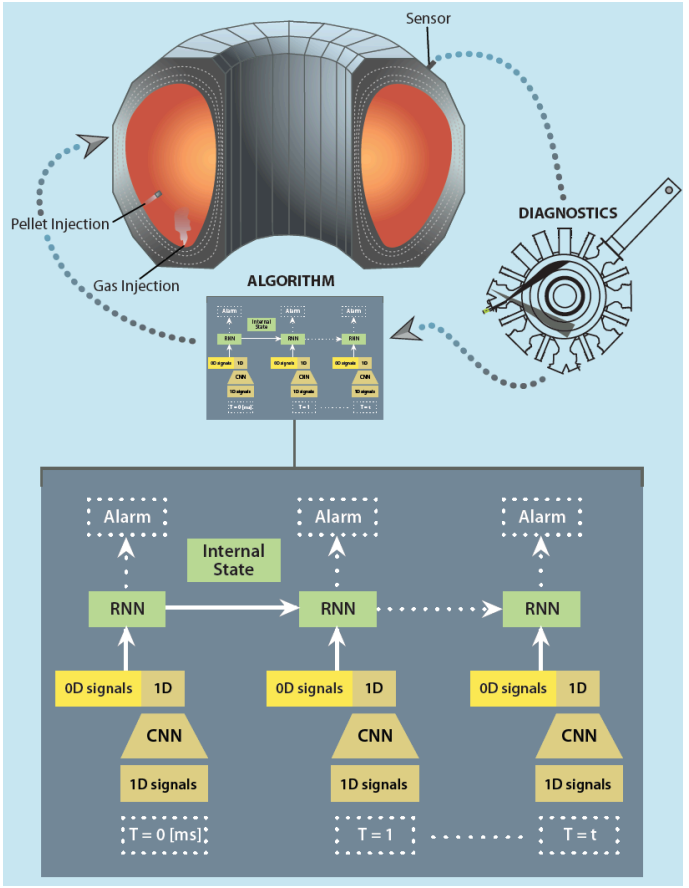
AI FOR SCIENCE APPLICATIONS



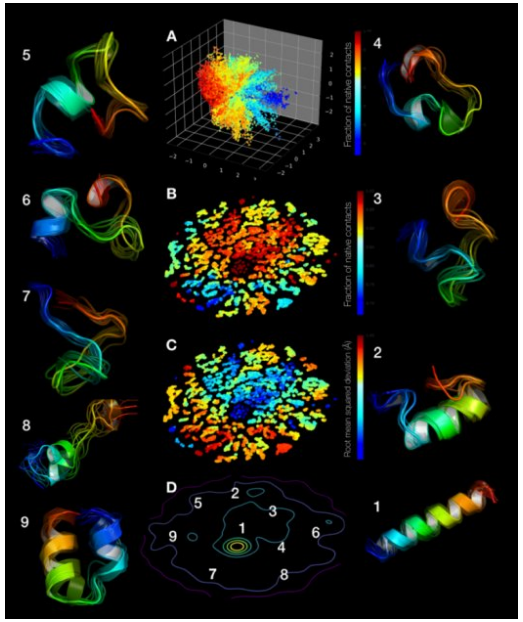
Cancer drug response prediction



Imaging Sciences-Braggs Peak



Tokomak Fusion Reactor operations

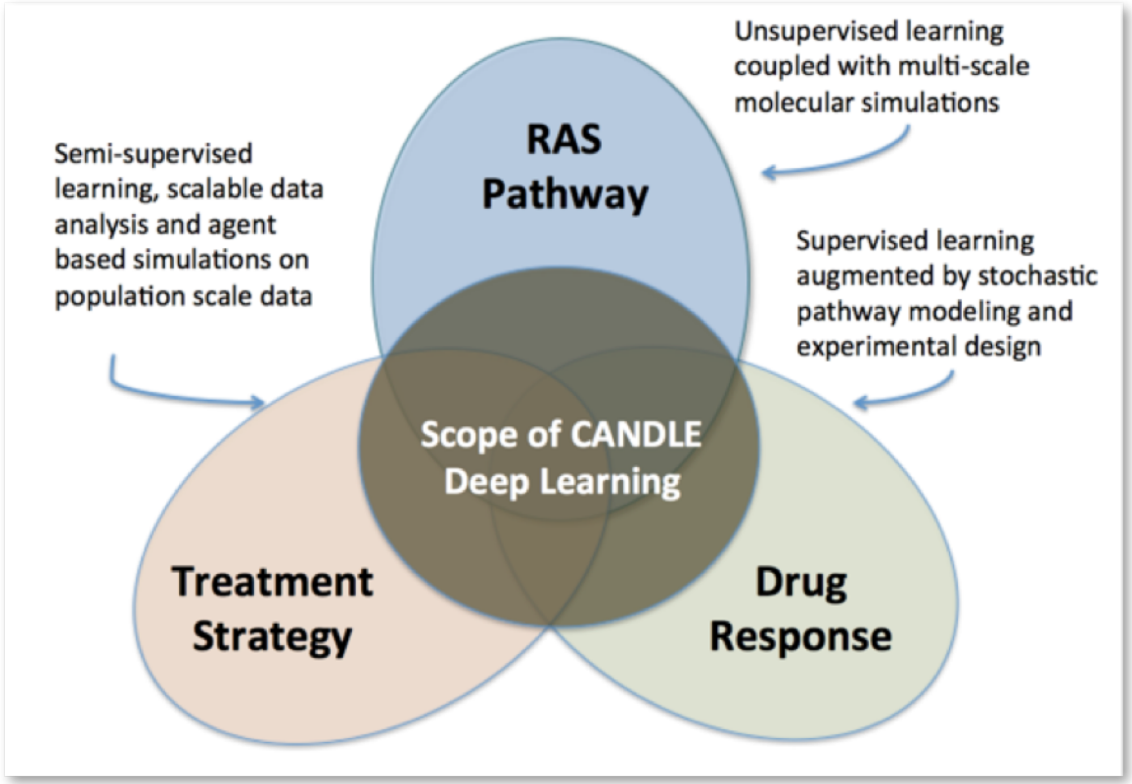


Protein-folding(Image: NCI)

and more..

Drug Discovery - Uno

- CANDLE: Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer
- Implement deep learning architectures that are relevant to problems in cancer.
- Focus on “Uno” application which aims to predict the drug response based on molecular features of tumor cells and drug descriptors.



Drug Discovery - Uno

- Throughput comparison in samples/second:
  SN30: 33392
  A100: 7567

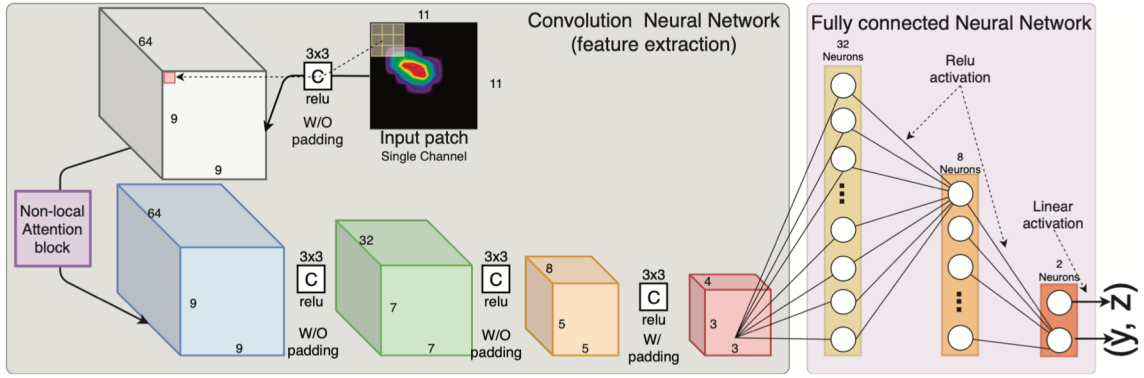
- 4.4x better performance over A100

Fast X-Ray Bragg Peak Analysis

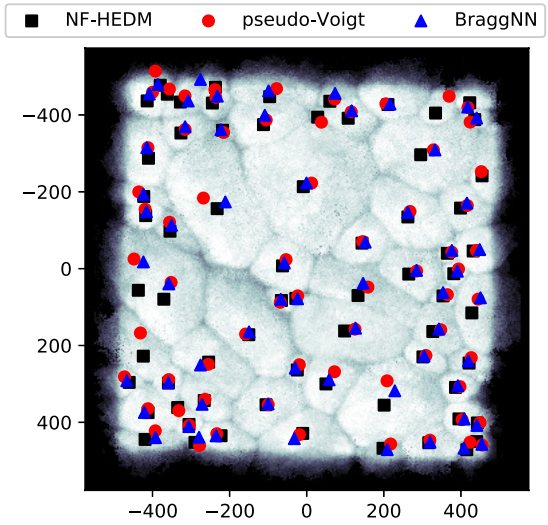
Goal: Enable rapid analysis and real-time feedback during an in-situ experiment with complex detector technologies

Proposed Approach: Deep learning-based method, BraggNN, for massive extraction of precise Bragg peak locations from far-field high energy diffraction microscopy data. BragNN has achieved 200X improvement over conventional pseudo-Voight profiling

Challenges: Model training capability is limited by the hardware



Application of the BraggNN deep neural network to an input patch yields a peak center position (y, z). All convolutions are 2D of size 3 × 3, with rectifier as activation function. Each fully connected layer, except for the output layer, also has a rectifier activation function.



A comparison of BraggNN, pseudo-Voigt FF-HEDM and NF-HEDM. (a) Grain positions from NF-HEDM (black squares), pseudo-Voigt FF-HEDM (red circles) and BraggNN FF-HEDM (blue triangles) overlaid on NF-HEDM confidence map

Courtesy: Z. Liu et al. BraggNN: Fast X-ray Bragg Peak Analysis Using Deep Learning. International Union of Crystallography (IUCrJ), Vol. 9, No. 1, 2022

Fast X-Ray Bragg Peak Analysis

For a batch size of 2048, we measure

(i) end-to-end (e2e) time that includes fixed time for compilation and data-preprocessing along with the model training time

(ii) Model throughput in samples per second

	SN	A100	
e2e time (sec)	133	199	1.55x
throughput (samples/sec)	518	73.7	7x

Ongoing Efforts

Generative AI For Science including large language models

GenSLM:

- LLM-based foundation model trained with gene sequences, focusing on Sars-CoV2 and E-Coli.
- Currently runs with 25M to 25B parameter models on A100 supercomputers.
- We expect massive growth in model sizes besides sequences lengths of ~32K.

CosmicTagger:

- Image segmentation task for liquid argon time projection chamber (LArTPC) detectors in Neutrino Physics experiments to classify each input pixel into one of three classes – Cosmic, Muon, or Background.
- UNet 3D-based model with high resolution image datasets both dense and sparse.

Ongoing Efforts

- Integrate SN30 system with the PBSPro scheduler to facilitate effective job scheduling.

- Evaluate traditional HPC on AI Accelerators

- Understand how to integrate the AI Accelerator with ALCF's existing and upcoming supercomputers to accelerate science insights

Recent Publications

- **GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics**
  Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, Arvind Ramanathan
  ** ***Winner of the ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research, 2022,***
  DOI: https://doi.org/10.1101/2022.10.10.511571

- **A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads**
  Murali Emani, Zhen Xie, Sid Raskar, Varuni Sastry, William Arnold, Bruce Wilson, Rajeev Thakur, Venkatram Vishwanath, Michael E Papka, Cindy Orozco Bohorquez, Rick Weisner, Karen Li, Yongning Sheng, Yun Du, Jian Zhang, Alexander Tsyplikhin, Gurdaman Khaira, Jeremy Fowers, Ramakrishnan Sivakumar, Victoria Godsoe, Adrian Macias, Chetan Tekur, Matthew Boyd, *13th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS) at SC 2022*

- **Enabling real-time adaptation of machine learning models at x-ray Free Electron Laser facilities with high-speed training optimized computational hardware**
  Petro Junior Milan, Hongqian Rong, Craig Michaud, Naoufal Layad, Zhengchun Liu, Ryan Coffee, Frontiers in Physics
  DOI: https://doi.org/10.3389/fphy.2022.958120

Recent Publications

- **Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action***
  Anda Trifan, Defne Gorgun, Zongyi Li, Alexander Brace, Maxim Zvyagin, Heng Ma, Austin Clyde, David Clark, Michael Salim, David Har dy,Tom Burnley, Lei Huang, John McCalpin, Murali Emani, Hyenseung Yoo, Junqi Yin, Aristeidis Tsaris, Vishal Subbiah, Tanveer Raza,J essica Liu, Noah Trebesch, Geoffrey Wells, Venkatesh Mysore, Thomas Gibbs, James Phillips, S.Chakra Chennubhotla, Ian Foster, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, John E. Stone, Emad Tajkhorshid, Sarah A. Harris, Arvind Ramanathan, International Journal of High-Performance Computing (IJHPC'22) DOI: https://doi.org/10.1101/2021.10.09.463779

- **Stream-AI-MD: Streaming AI-driven Adaptive Molecular Simulations for Heterogeneous Computing Platforms**
  Alexander Brace, Michael Salim, Vishal Subbiah, Heng Ma, Murali Emani, Anda Trifa, Austin R. Clyde, Corey Adams, Thomas Uram, Hyunseung Yoo, Andrew Hock, Jessica Liu, Venkatram Vishwanath, and Arvind Ramanathan. 2021 Proceedings of the Platform for Advanced Scientific Computing Conference (PASC'21). DOI: https://doi.org/10.1145/3468267.3470578

- **Bridging Data Center AI Systems with Edge Computing for Actionable Information Retrieval**
  Zhengchun Liu, Ahsan Ali, Peter Kenesei, Antonino Miceli, Hemant Sharma, Nicholas Schwarz, Dennis Trujillo, Hyunseung Yoo, Ryan Coffee, Naoufal Layad, Jana Thayer, Ryan Herbst, Chunhong Yoon, and Ian Foster, 3rd Annual workshop on Extreme-scale Event-in-the-loop computing (XLOOP), 2021

- **Accelerating Scientific Applications With SambaNova Reconfigurable Dataflow Architecture**
  Murali Emani, Venkatram Vishwanath, Corey Adams, Michael E. Papka, Rick Stevens, Laura Florescu, Sumti Jairath, William Liu, Tejas Nama, Arvind Sujeeth, IEEE Computing in Science & Engineering 2021 DOI: 10.1109/MCSE.2021.3057203.

*** Fiinalist in the ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research, 2021**

Thank You

- This research was funded in part and used resources of the Argonne Leadership Computing Facility (ALCF), a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

- Venkatram Vishwanath, Michael Papka, William Arnold, Bruce Wilson, Varuni Sastry, Sid Raskar, Zhen Xie, Rajeev Thakur, Anthony Avarca, Arvind Ramanathan, Alex Brace, Zhengchun Liu, Hyunseung (Harry) Yoo, Corey Adams, Ryan Aydelott, Kyle Felker, Craig Stacey, Tom Brettin, Rick Stevens, and many others have contributed to this material.

Please reach out for further details
Venkat Vishwanath (venkat@anl.gov)
Murali Emani (memani@anl.gov)