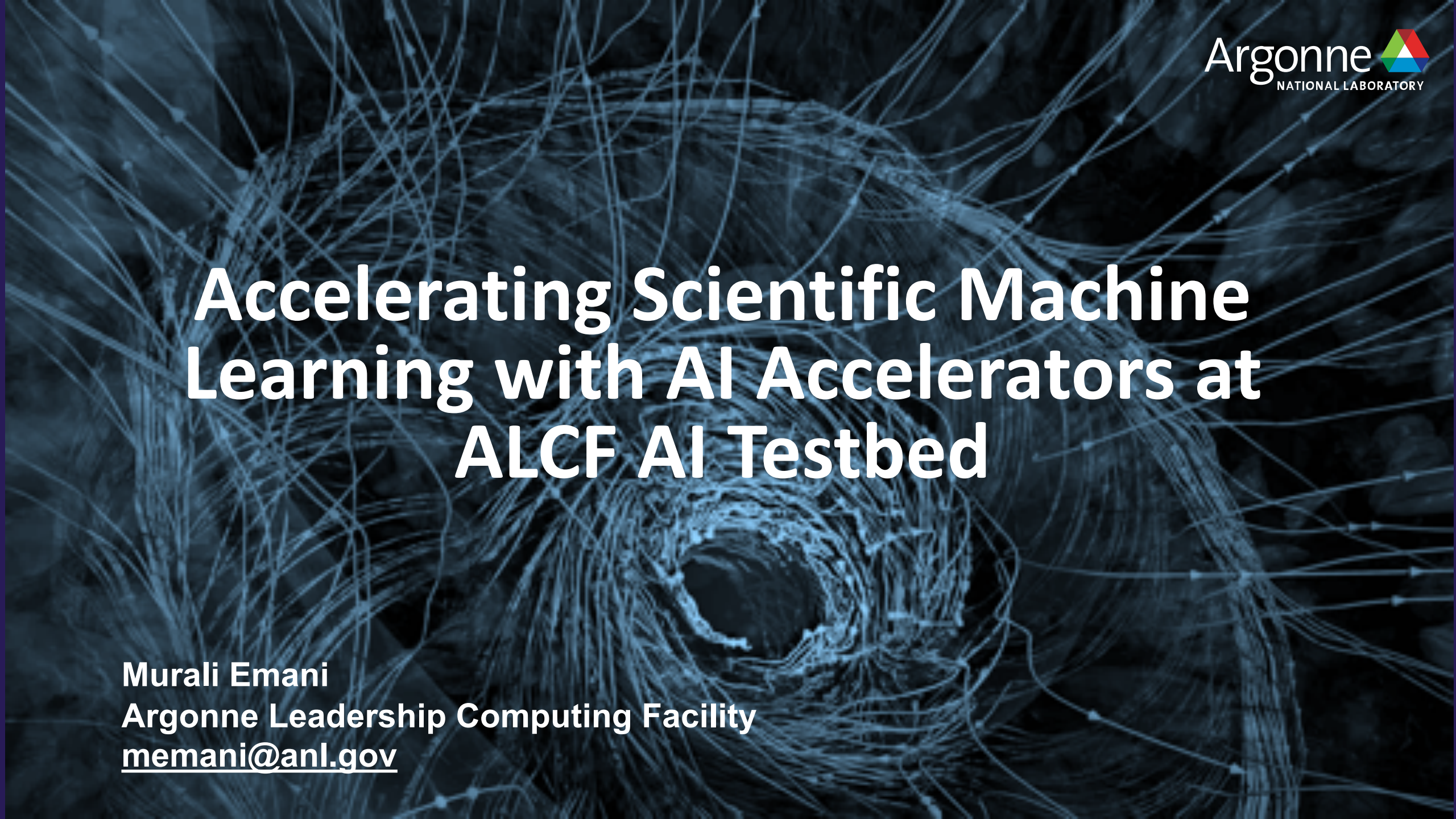


Accelerating Scientific Machine Learning with AI Accelerators at ALCF AI Testbed

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

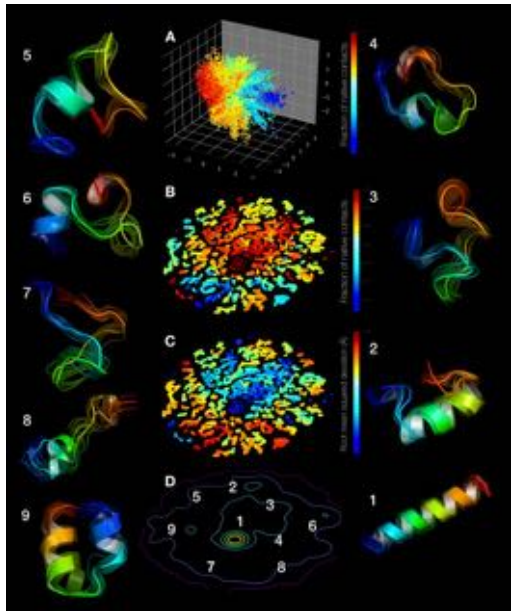
§ Large-scale Simulation (PDEs, traditional HPC)

§ Data Intensive Applications (scalable science pipelines)

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

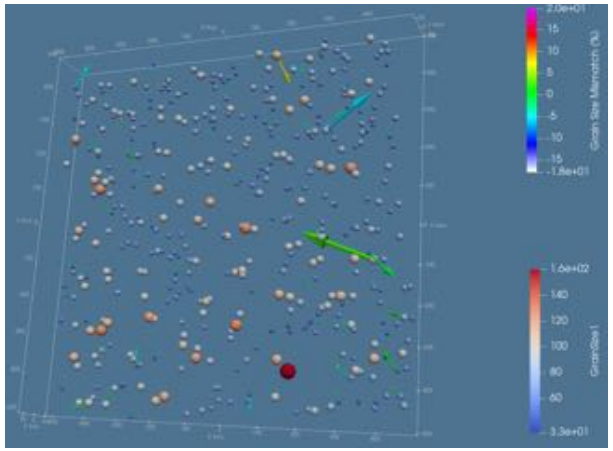
Surge of Scientific Machine Learning

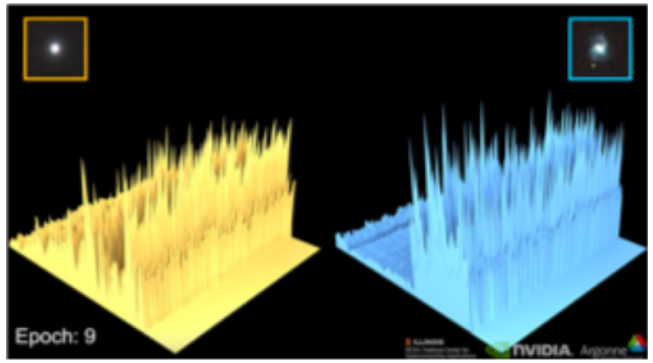
- Simulations/ surrogate models
  - Replace, in part, or guide simulations with AI-driven surrogate models
- Data-driven models
  - Use data to build models without simulations
- Co-design of experiments
  - AI-driven experiments



Protein-folding

Braggs Peak





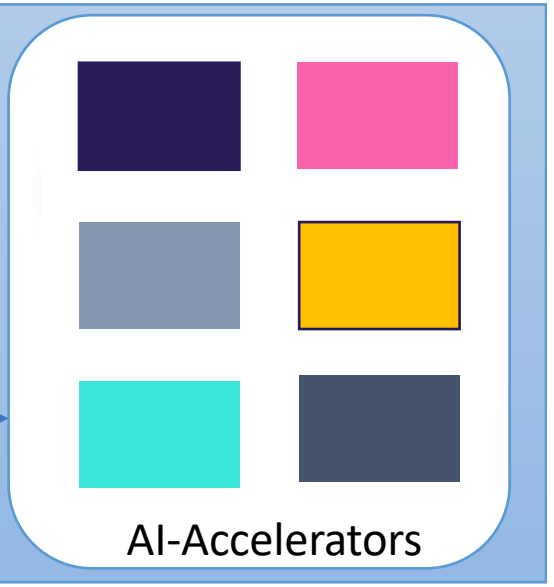
Galaxy Classification

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

Integrating AI Systems in Facilities

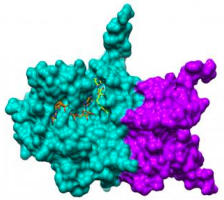
Computing Facility

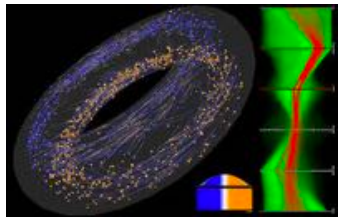
Experimental Facility

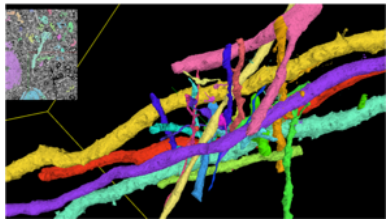


AI-Accelerators

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

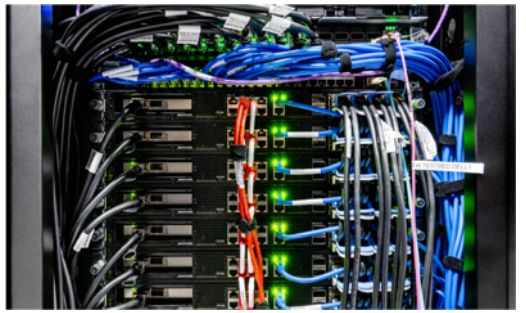
Recent ALCF AI Testbed Updates

ALCF AI Testbed Systems are in production and available for allocations to the research community

https://www.alcf.anl.gov/science/directors-discretionary-allocation-program

SambaNova upgraded to latest 2nd generation SN30 accelerators and scaled to 8 nodes with 64 AI accelerators

SambaNova SN30



Graphcore upgraded to latest Bow generation accelerators and scaled to a Pod-64 configuration with 64 accelerators

Graphcore BowPod64

Cerebras CS-2 upgraded to an appliance mode to include Memory-X and Swarm-X technologies to enable larger models and scaled to two CS-2 engines

Cerebras CS-2

Groq system has been upgraded to a GroqRack with nine nodes, each consisting of eight GroqChip Tensor streaming processors accelerators

GroqRack

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

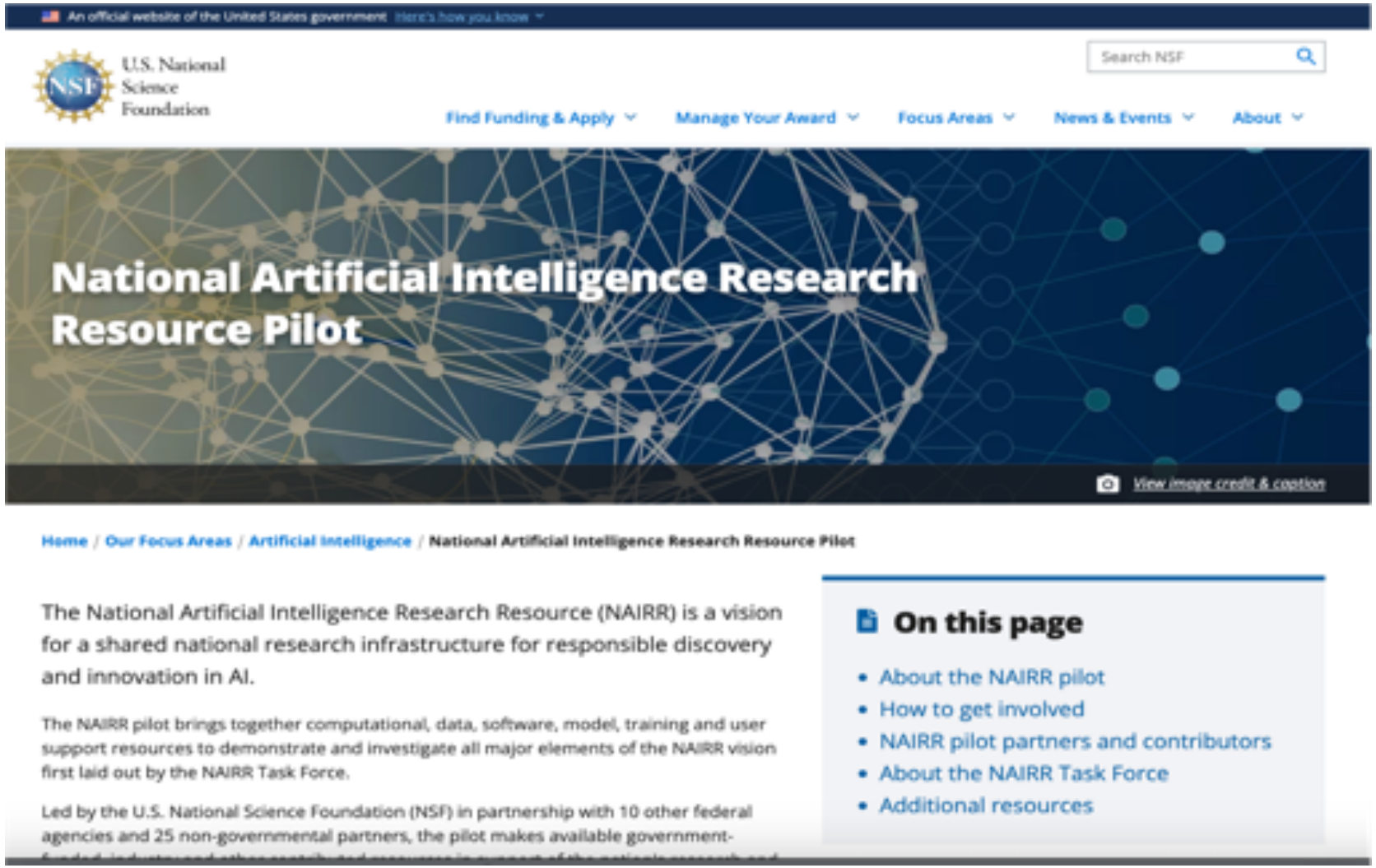
Cerebras CS-2, SambaNova Datascale SN30, Graphcore Bow Pod64, and GroqRack are available for allocations

Allocation Request Form

AI Testbed User Guide

NAIRR

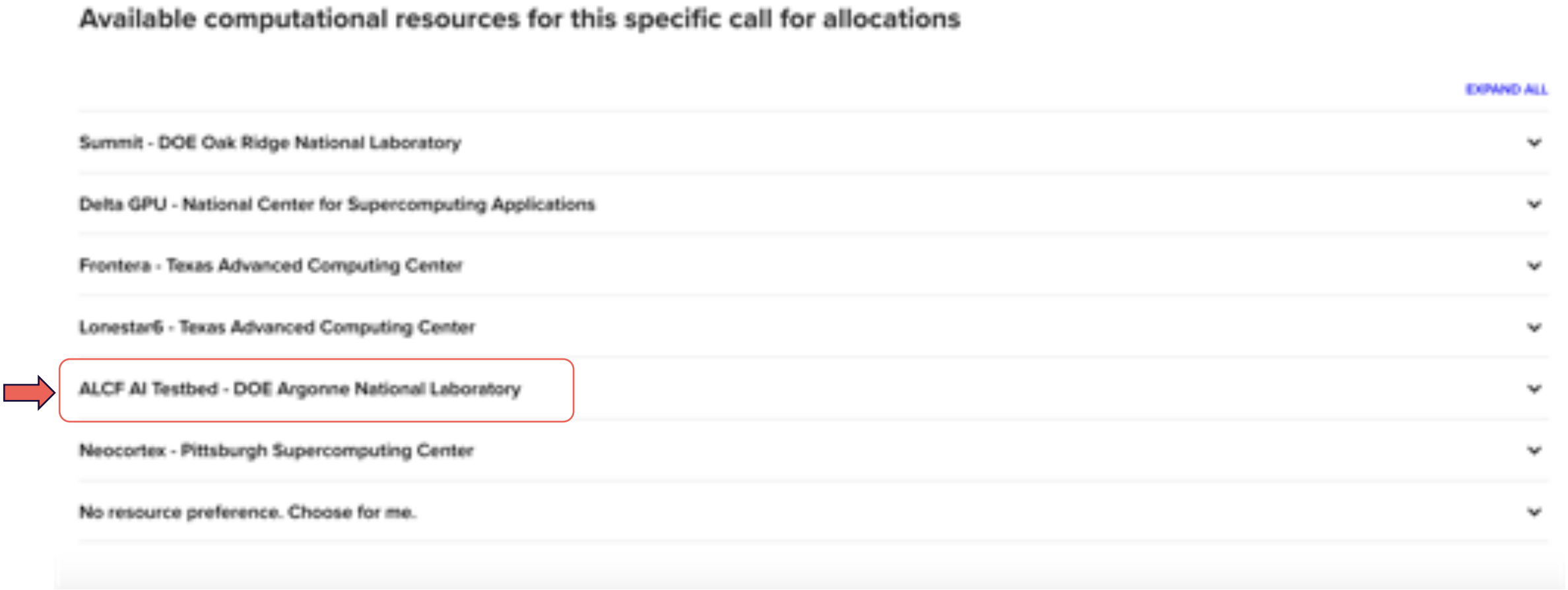
https://new.nsf.gov/focus-areas/artificial-intelligence/nairr



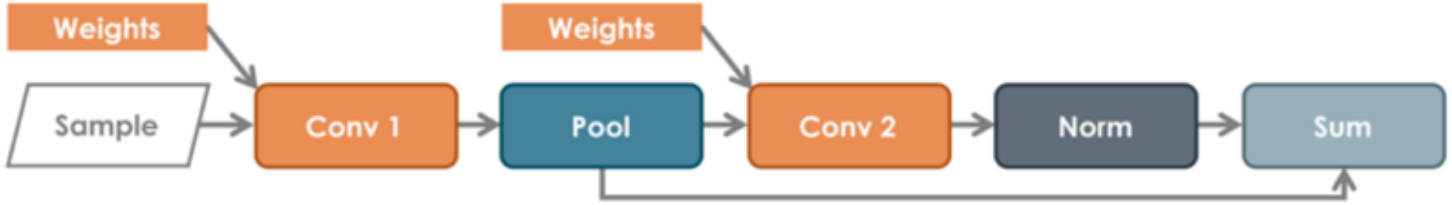
NAIRR

https://nairrpilot.org/allocations

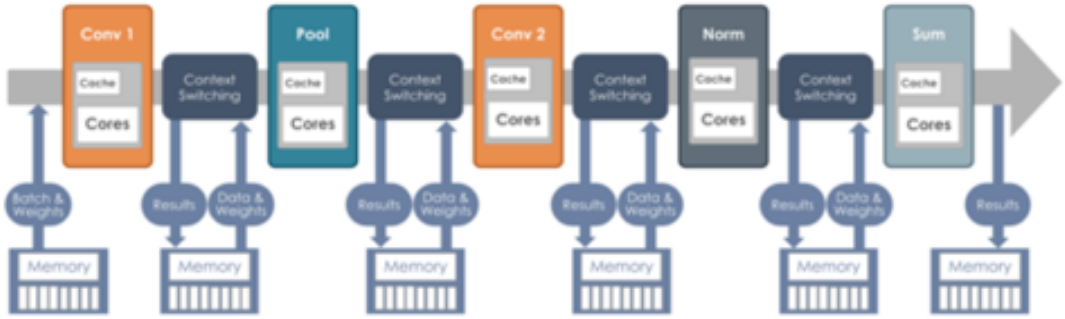
The initial call is open from **January 24 to March 1, 2024**



Dataflow Architectures



Simple Convolution Graph



The old way: kernel-by-kernel

Bottlenecked by memory bandwidth and host overhead

The Dataflow way: Spatial

Eliminates memory traffic and overhead

Image Courtesy: SambaNova

Dataflow Architecture for Terabyte Sized Models

DataScale SN30-8R

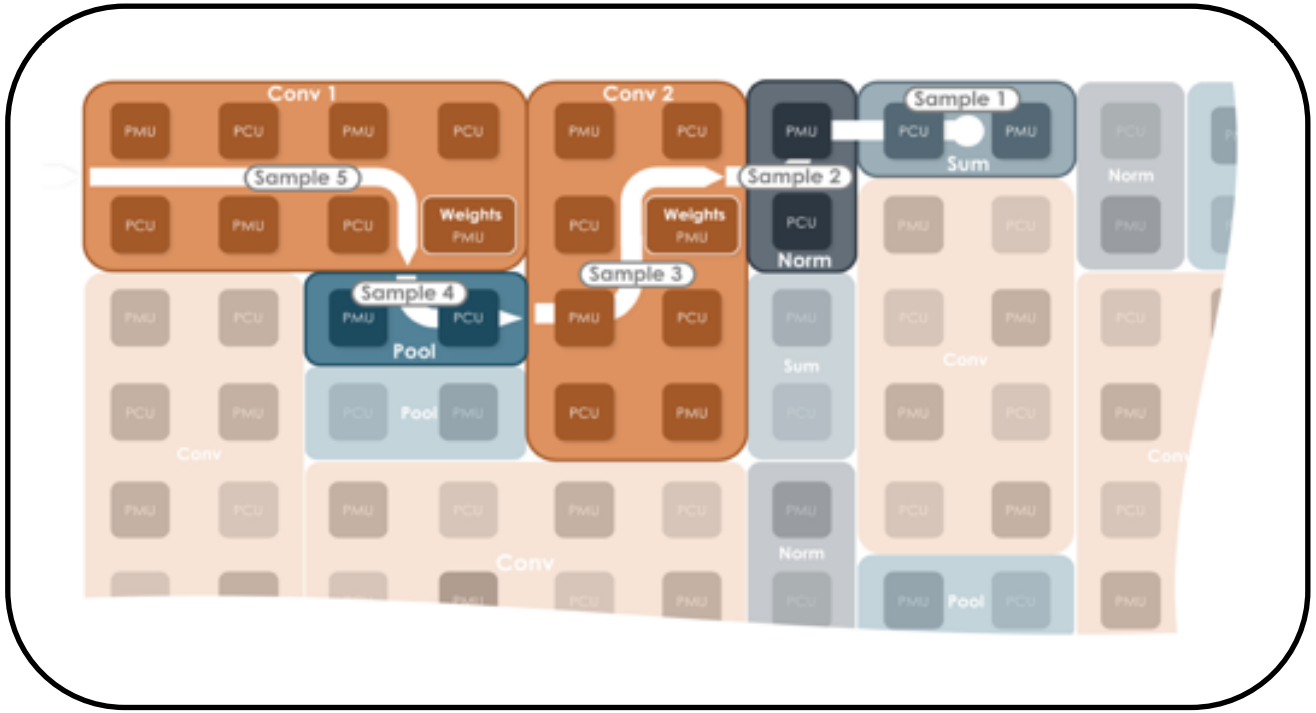
Dataflow Efficiency

+

Compute Capability

+

Large Memory Capacity



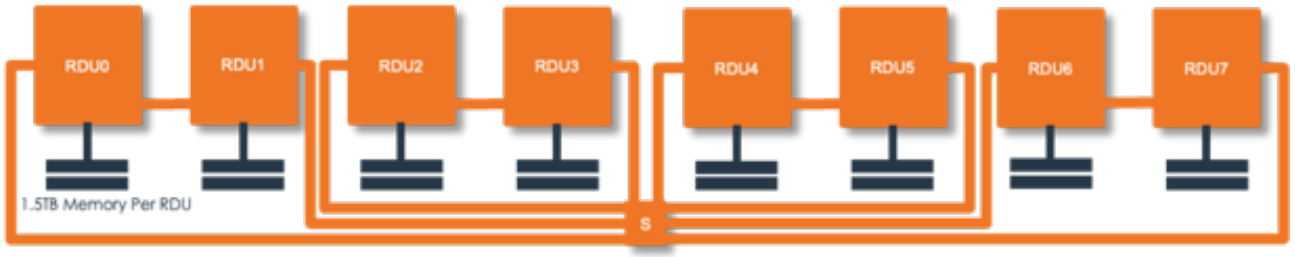


Image Courtesy: SambaNova

Cerebras Wafer-Scale Engine (WSE-2)

850,000 cores optimized for sparse linear algebra

46,225 mm² silicon

2.6 trillion transistors

40 gigabytes of on-chip memory

20 PByte/s memory bandwidth

220 Pbit/s fabric bandwidth

7nm process technology

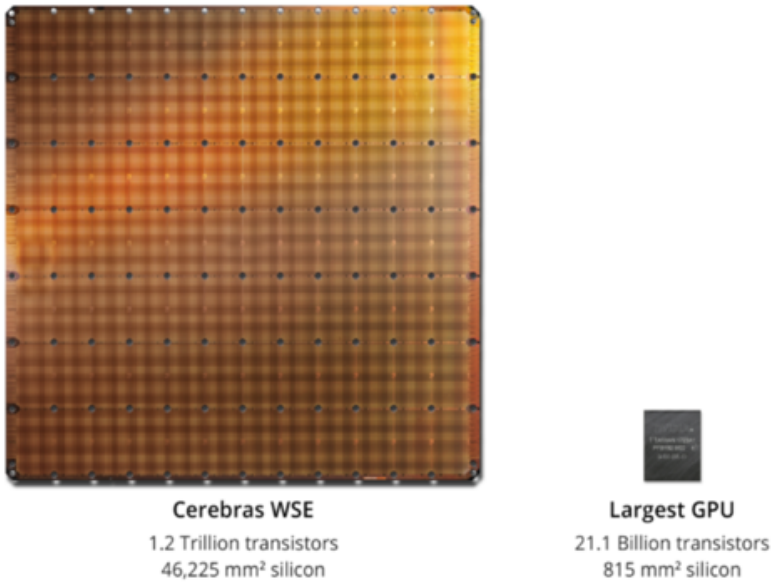
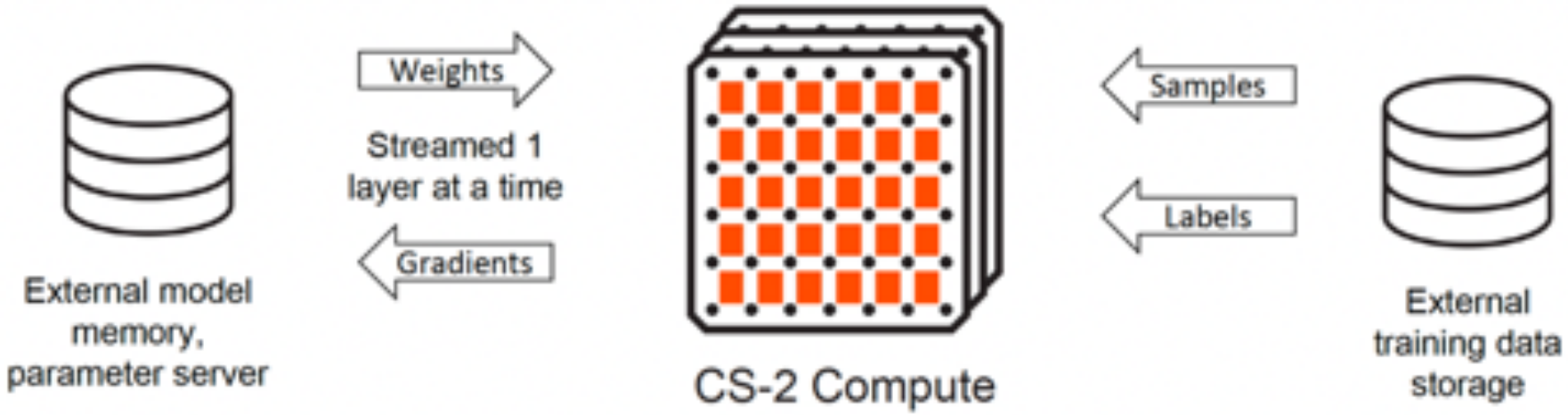


Image Courtesy: Cerebras

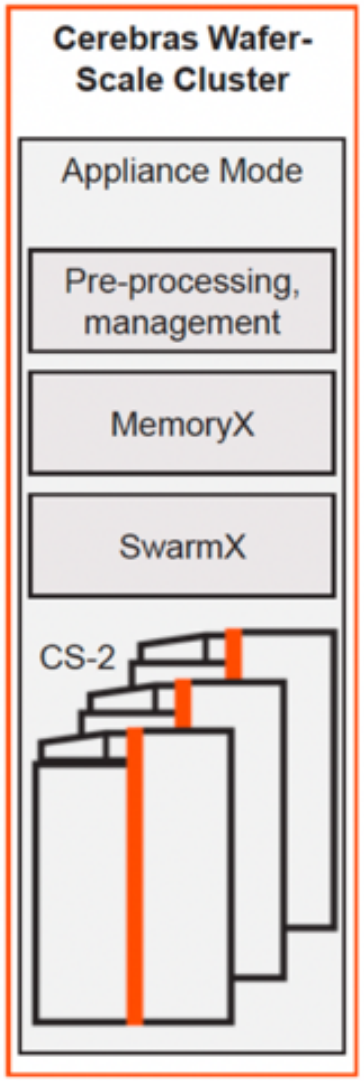
Cerebras Weight Streaming Technology



Disaggregate storage and compute
Enable scaling model size

Image Courtesy: Cerebras

Wafer-Scale Cluster



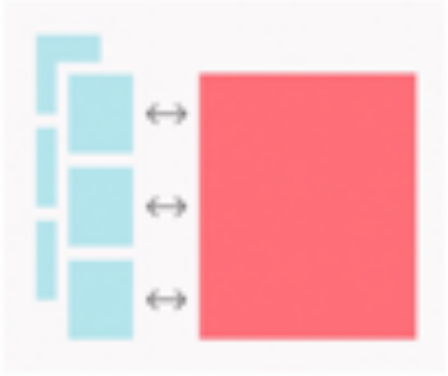
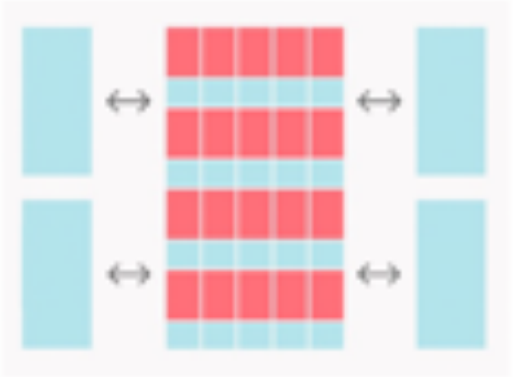
Input preprocessing servers stream training data

MemoryX - Stores and streams model's weights

SwarmX – weight broadcasts and gradient across multiple CS2s

Image Courtesy: Cerebras

Graphcore Intelligence Processing Unit (IPU)

	CPU	GPU	IPU
Parallelism	Designed for scalar processing	SIMD/SIMT architecture. Designed for large blocks of dense contiguous data	Massively parallel MIMD architecture. High performance/efficiency for future ML trends
			
Memory Bandwidth	Off-chip memory	Model and Data spread across off-chip and small on-chip cache and shared memory (2TB/s for A100 HBM)	Main Model & Data in tightly coupled large locally distributed SRAM (~65 TB/s for Bow IPU)

Slide Courtesy: Graphcore

Challenges

- Understand how these systems perform for different workloads given diverse hardware and software characteristics
- What are the unique capabilities of each evaluated system
- Opportunities and potential for integrating AI accelerators with HPC computing facilities

Approach

- Perform a comprehensive evaluation with a diverse set of Deep Learning (DL) models*:
  - *DL primitives*: GEMM, Conv2D, ReLU, and RNN
  - *Benchmarks*: U-Net, BERT-Large, ResNet-50
  - *AI4S applications*: BraggNN, Uno
  - Scalability and Collective communications

- Evaluation of Large Language Models
  —Transformer block micro-benchmark, GPT-2, and GenSLM

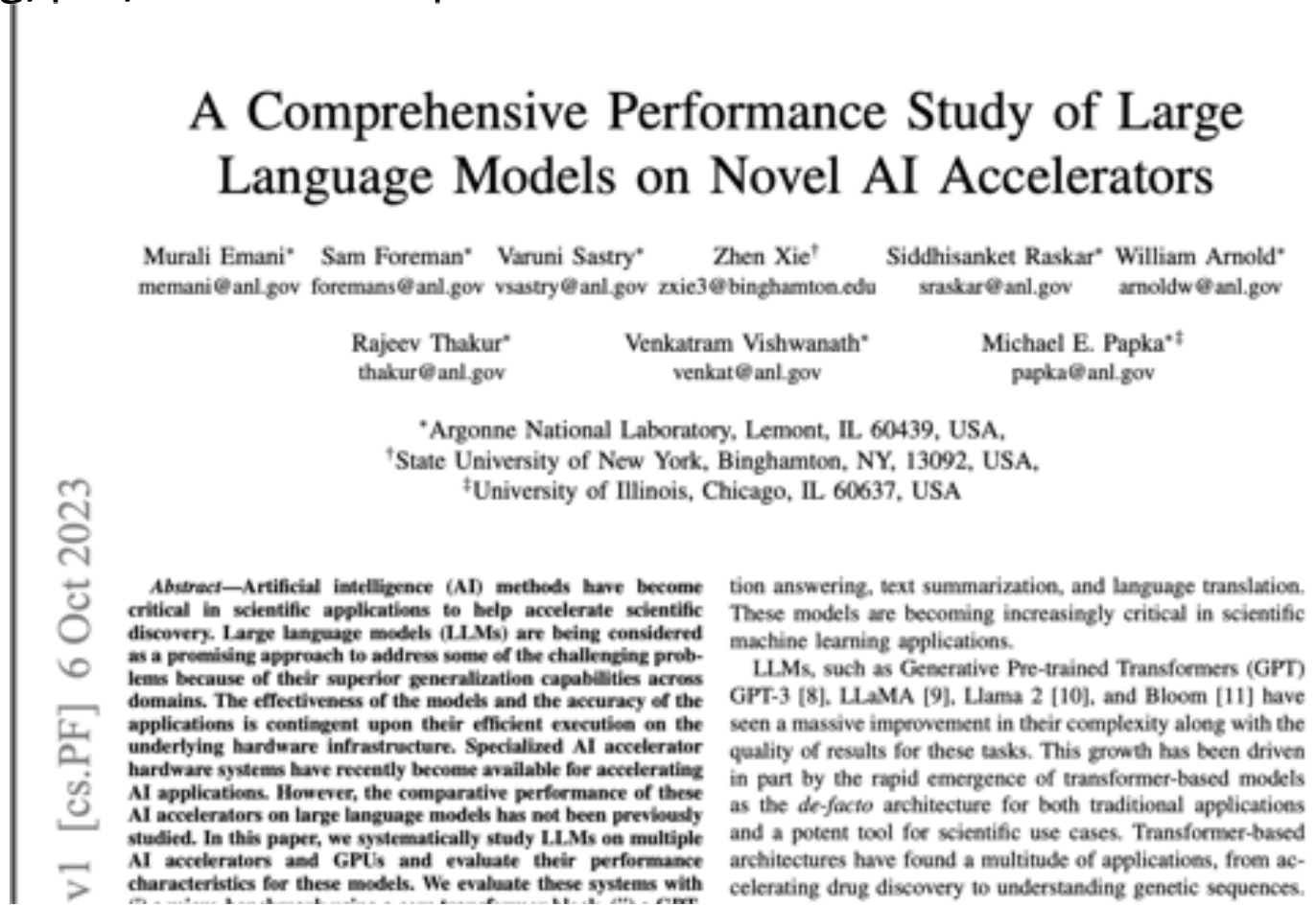
*** Emani et al. "A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads",**
13th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS) at SC 2022.

.

Performance Evaluation

“A Comprehensive Performance Study of Large Language Models on Novel AI Accelerators”

https://arxiv.org/pdf/2310.04607.pdf



A Comprehensive Performance Study of Large Language Models on Novel AI Accelerators

Murali Emani* memani@anl.gov
Sam Foreman* foremans@anl.gov
Varuni Sastry* vsastry@anl.gov
Zhen Xie† zxie3@binghamton.edu
Siddhisanket Raskar* sraskar@anl.gov
William Arnold* arnoldw@anl.gov
Rajeev Thakur* thakur@anl.gov
Venkatram Vishwanath* venkat@anl.gov
Michael E. Papka*‡ papka@anl.gov

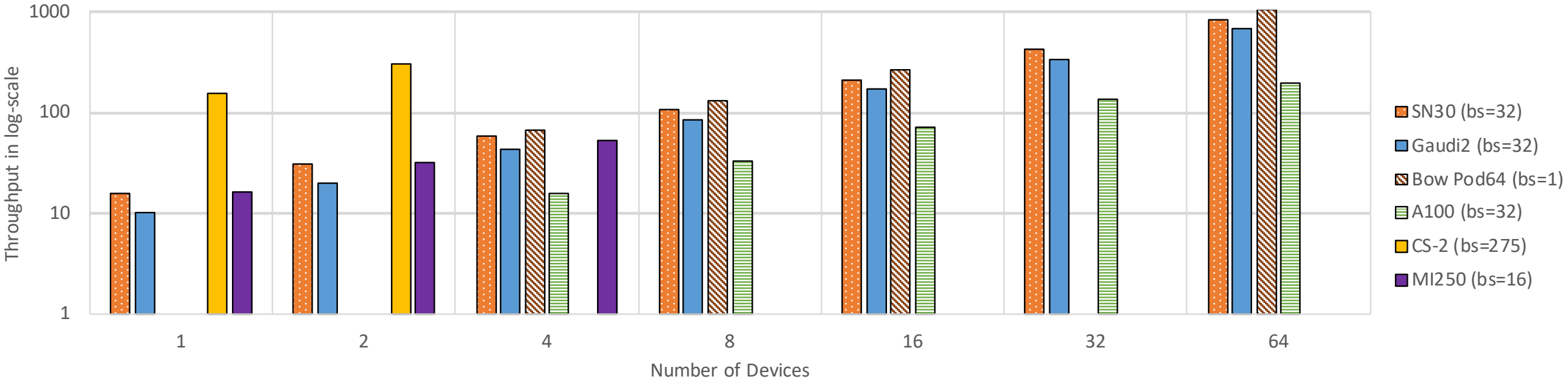
*Argonne National Laboratory, Lemont, IL 60439, USA,
†State University of New York, Binghamton, NY, 13092, USA,
‡University of Illinois, Chicago, IL 60637, USA

***Abstract*—Artificial intelligence (AI) methods have become critical in scientific applications to help accelerate scientific discovery. Large language models (LLMs) are being considered as a promising approach to address some of the challenging problems because of their superior generalization capabilities across domains. The effectiveness of the models and the accuracy of the applications is contingent upon their efficient execution on the underlying hardware infrastructure. Specialized AI accelerator hardware systems have recently become available for accelerating AI applications. However, the comparative performance of these AI accelerators on large language models has not been previously studied. In this paper, we systematically study LLMs on multiple AI accelerators and GPUs and evaluate their performance characteristics for these models. We evaluate these systems with**

tion answering, text summarization, and language translation. These models are becoming increasingly critical in scientific machine learning applications.

LLMs, such as Generative Pre-trained Transformers (GPT) GPT-3 [8], LLaMA [9], Llama 2 [10], and Bloom [11] have seen a massive improvement in their complexity along with the quality of results for these tasks. This growth has been driven in part by the rapid emergence of transformer-based models as the *de-facto* architecture for both traditional applications and a potent tool for scientific use cases. Transformer-based architectures have found a multitude of applications, from accelerating drug discovery to understanding genetic sequences.

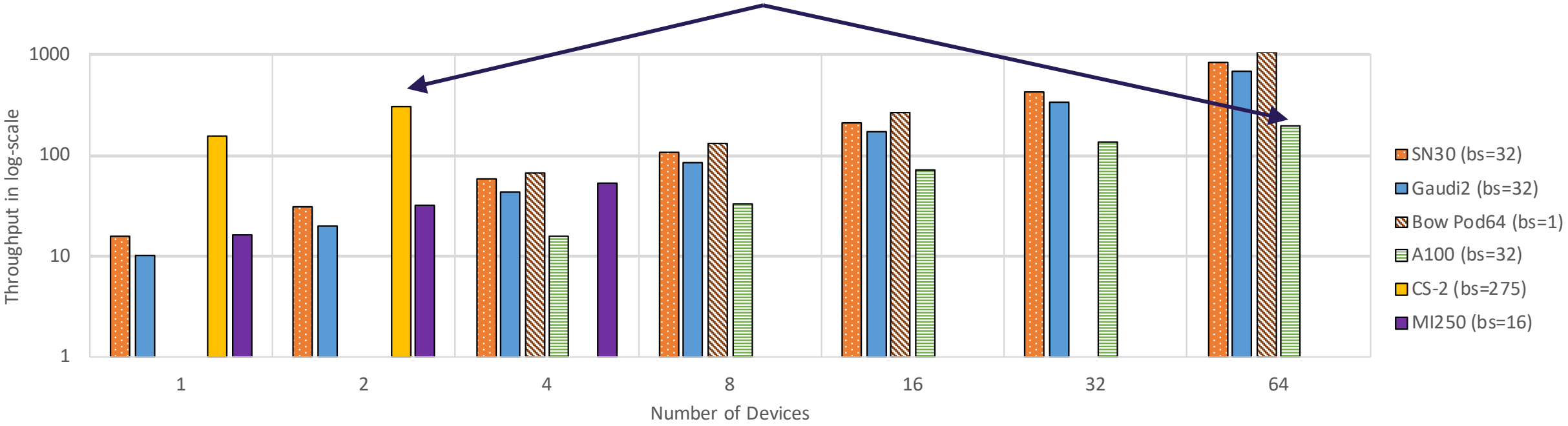
GPT Model Performance



Used GPT-2 XL 1.5B parameter model

- same sequence length, tuned batch sizes
- 16 SN30 RDUs, 2 CS-2s, and 16 IPUs outperformed the runs on 64 A100s
- Scaling efficiencies range from 78% to 104%

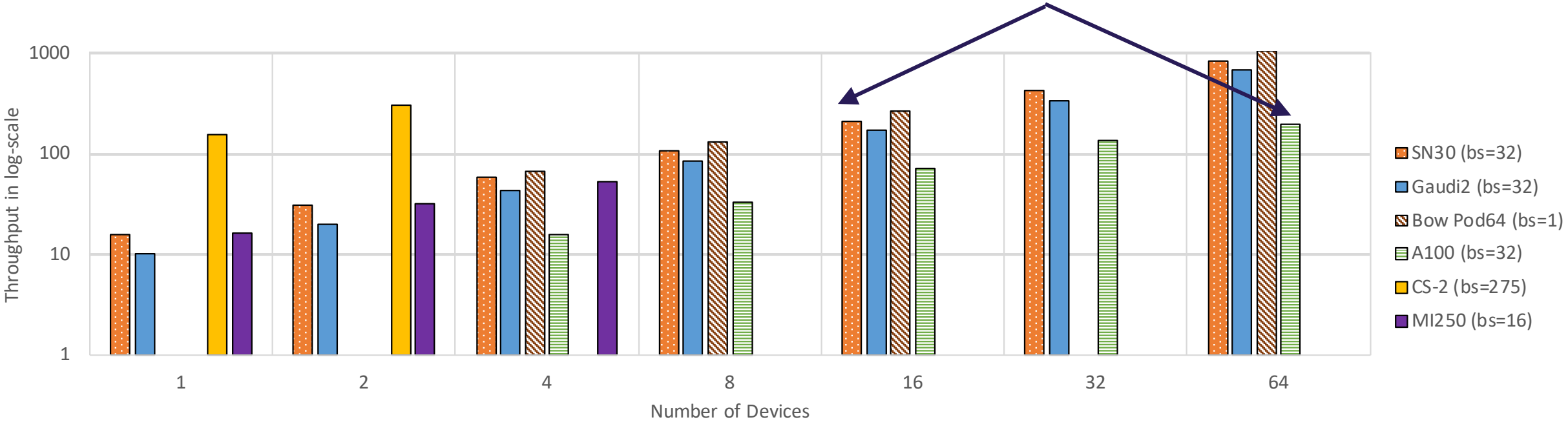
GPT Model Performance



Used GPT-2 XL 1.5B parameter model

- same sequence length, tuned batch sizes
- 16 SN30 RDUs, 2 CS-2s, and 16 IPUs outperformed the runs on 64 A100s
- Scaling efficiencies range from 78% to 104%

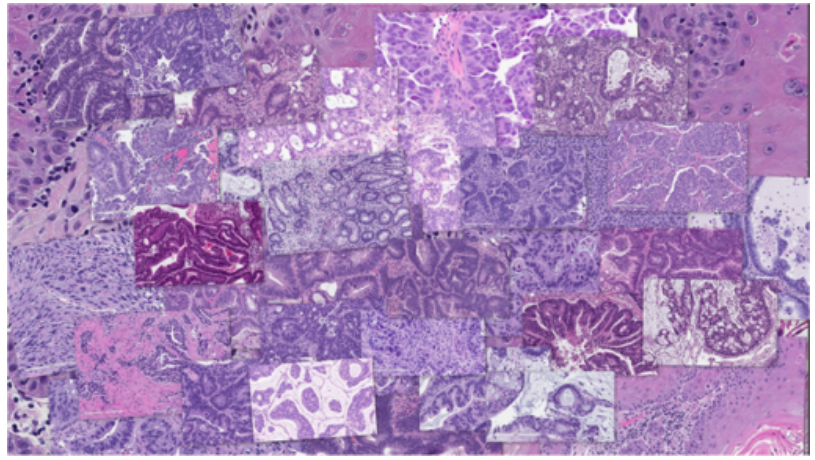
GPT Model Performance



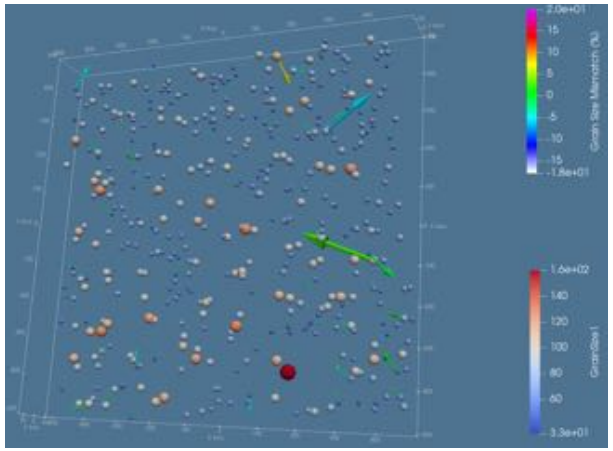
Used GPT-2 XL 1.5B parameter model

- same sequence length, tuned batch sizes
- 16 SN30 RDUs, 2 CS-2s, and 16 IPUs outperformed the runs on 64 A100s
- Scaling efficiencies range from 78% to 104%

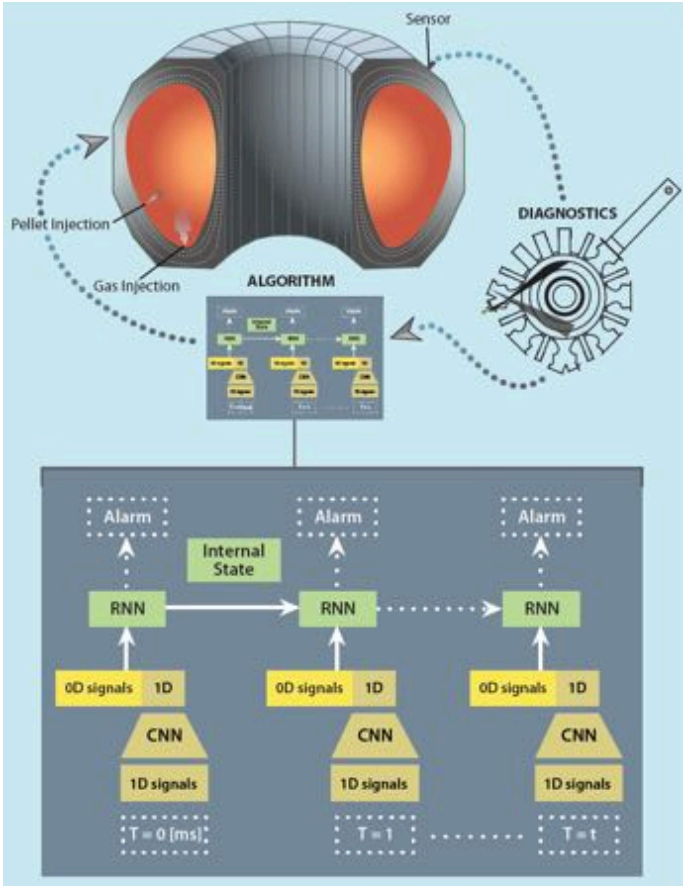
AI FOR SCIENCE APPLICATIONS



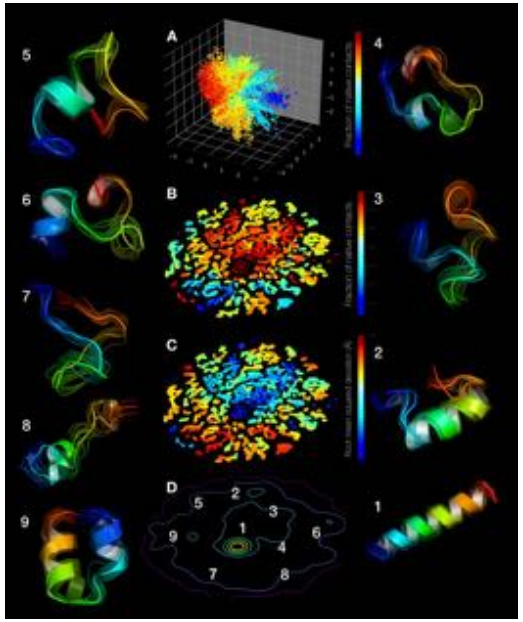
Cancer drug response prediction



Imaging Sciences-Braggs Peak



Tokomak Fusion Reactor operations



Protein-folding(Image: NCI)

and more..

Genome-scale Language Models (GenSLMs)

- Identify new and emergent variants of pandemic causing viruses, (eg. SARS-CoV-2)
- Identify mutations that are VOC (increased severity and transmissibility)
- Extendable to gene or protein synthesis.

- Adapt Large Language Models (LLMs) to learn the evolution.
- Pretrain 25M – 25B models on raw nucleotides with large sequence lengths.
- Scale on GPUs, CS2s, SN30.

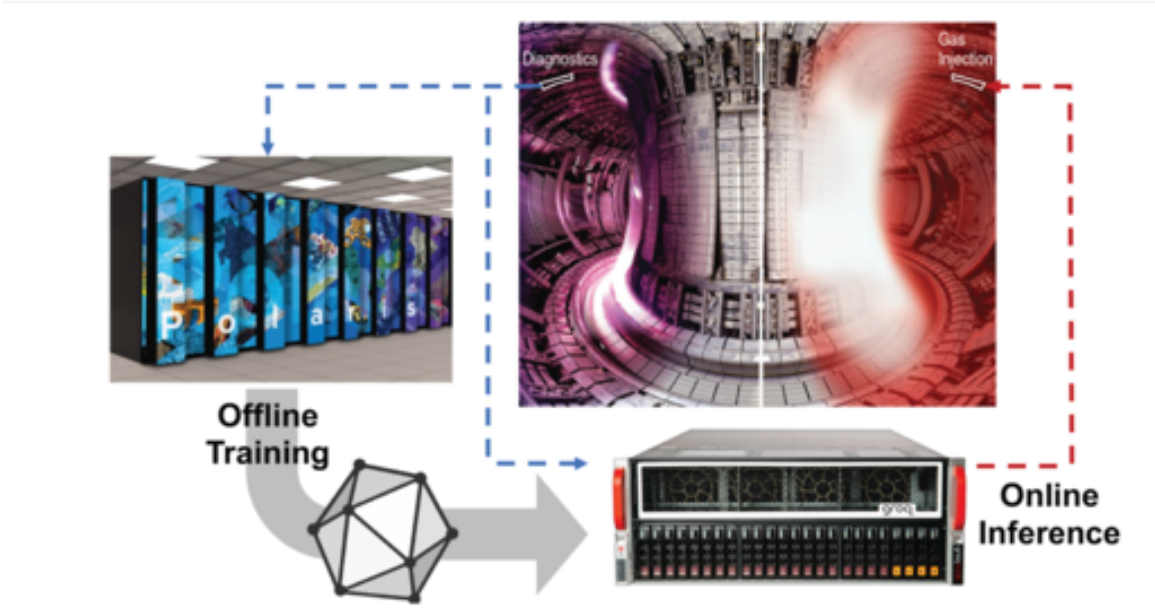
GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics
Winner of the ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research, 2022,
DOI: https://doi.org/10.1101/2022.10.10.511571

GenSLM 13B Training Performance

Accelerator System	Number of Devices	Throughput (tokens/sec)	Improvement (relative to A100-based system)
Nvidia A100	8	1150	1
SambaNova SN30	8	9795	8.5
Cerebras CS-2	1	21811	19

Accelerating Fusion Research

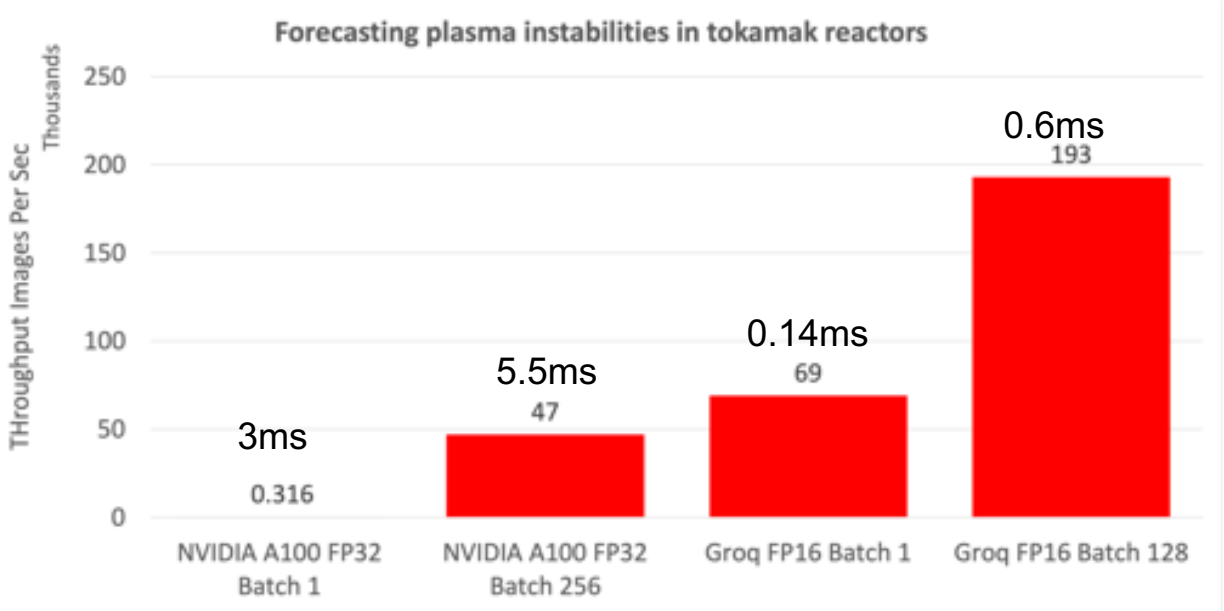
- Deep learning-guided investigations aimed at informing the operation of future fusion energy devices.
- Improve predictive capabilities and mitigate large-scale disruptions in burning plasmas in experimental tokamak systems such as ITER
- Application to the prediction of instabilities in real-world plasma-discharge experiments being conducted as part of a path to viable fusion energy



To improve predictive capabilities for fusion energy science, researchers are developing a workflow that integrates ALCF supercomputers for AI model training with the ALCF AI Testbed's Groq system for inference. Image: Kyle Felker, Argonne National Laboratory.

https://www.alcf.anl.gov/news/researchers-accelerate-fusion-research-argonne-s-groq-ai-platform

Accelerating Fusion Research



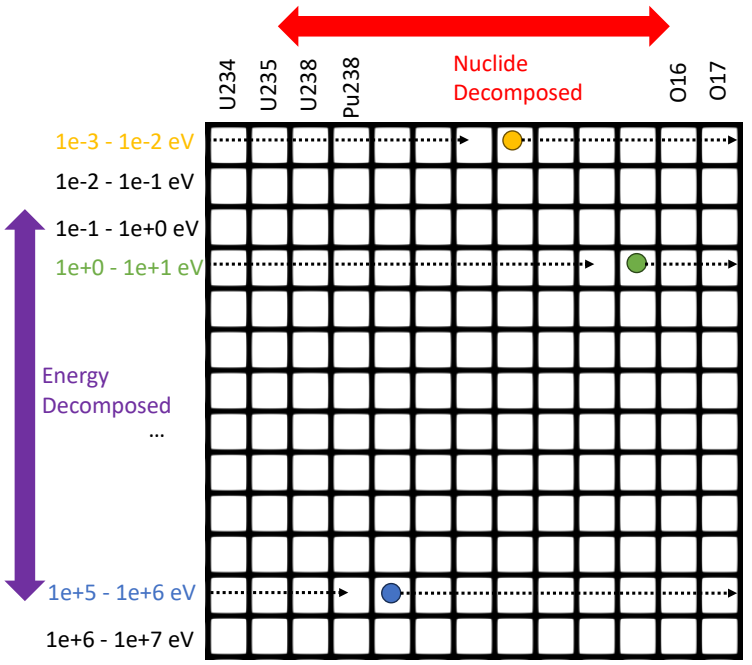
Forecasting Plasma Instability in Tokamak

Promising results using GroqChip for science inference use-cases with respect to latency and throughput in comparison to GPUs

20x improvement in latency
218x improvement in throughput

Monte Carlo with Single Cycle Latency: leveraging the Cerebras CS-2 for acceleration of a latency-bound HPC simulation workload

Challenge: We examine the feasibility of performing continuous energy Monte Carlo (MC) particle transport on the Cerebras WSE-2 AI accelerator by porting XSBench to the Cerebras "CSL" programming model. The MC algorithm has traditionally been bandwidth/latency-bound, making the WSE-2's 40 GB of 1-cycle SRAM an attractive architecture. The critical challenge is to decompose data and tasks across the WSE-2's ~750,000 distributed memory processing elements (PEs), each having only 48 KB of memory.



MC cross section data decomposition across a 2D grid of WSE-2 processing elements. This diagram shows the third phase of our algorithm where particles are exchanged in a round-robin manner to visit all nuclides in the row.

Outcome:

- Developed several novel algorithms for decomposing data structures across the WSE-2's 2D network grid, for flowing particles (tasks) through the WSE-2, and for performing dynamic load balancing.
- Developed a method for exploiting the WSE-2's **hardware random number generation** capabilities to **accelerate kernel by 65%.**
- WSE-2 was found to run **130x faster than a highly optimized CUDA version of the kernel run on an NVIDIA A100 GPU.**

	Transistor Count [Trillion]	Peak Power [kW]	Monte Carlo XS Lookup FOM [Lookups/s]
A100 GPU	0.0542	0.4	6.43E+07
Cerebras	2.6	22.8	8.36E+09
Cerebras/A100	**48**	**57**	**130**

Computational Physics Communications (https://doi.org/10.1016/j.cpc.2023.109072)

Data courtesy: John Tramm (Argonne)

Observations, Challenges and Insights

- Significant speedup achieved for a wide-gamut of scientific ML applications
  - Easier to deal with larger resolution data and to scale to multi-chip systems

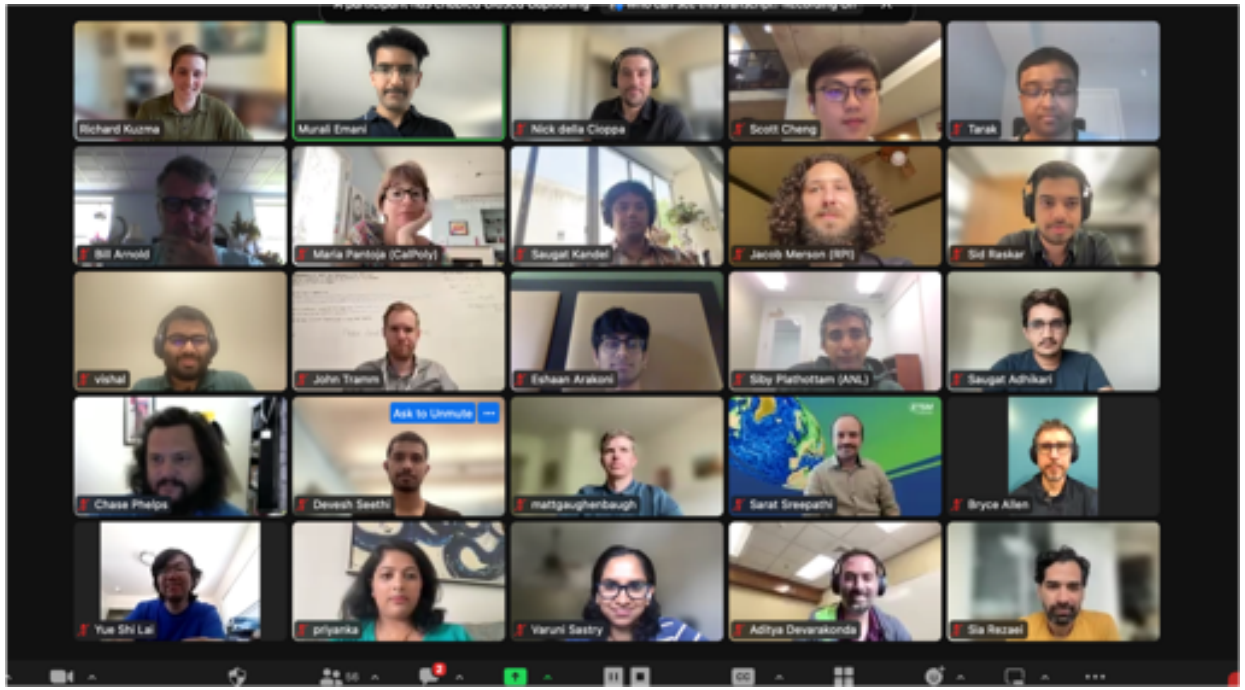
- Room for improvement exists
  - Porting efforts and compilation times
  - Coverage of DL frameworks, support for performance analysis tools, debuggers

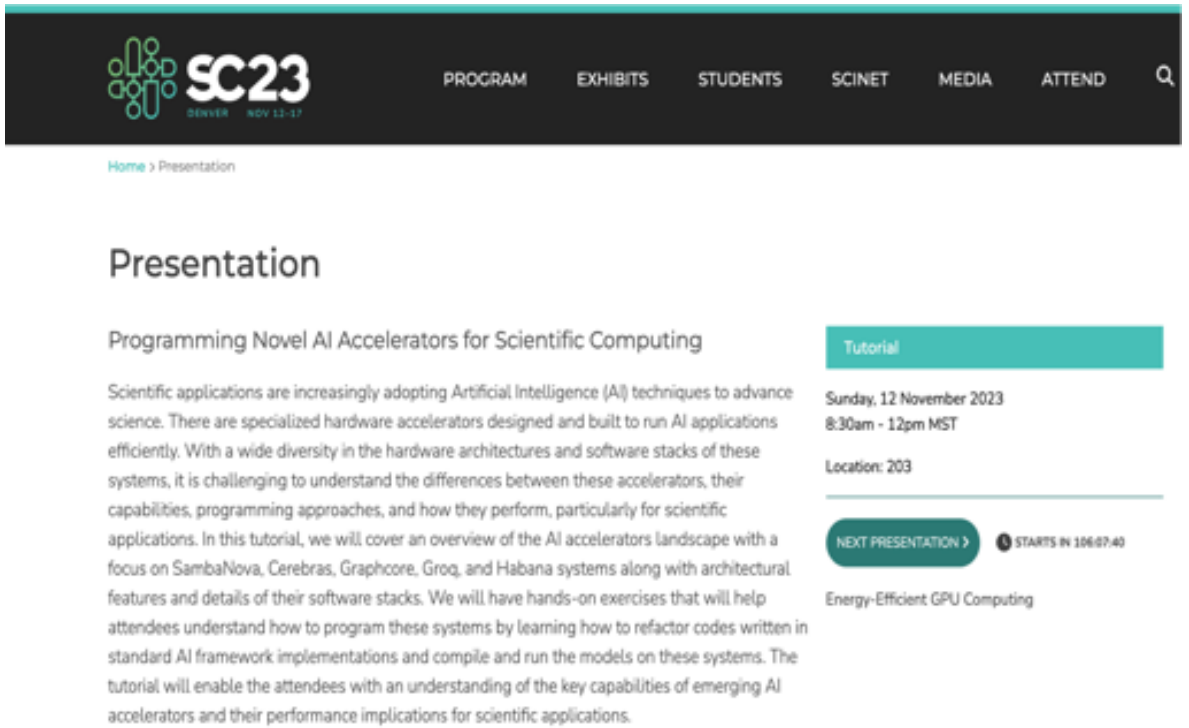
- Limited capability to support low-level HPC kernels
  - Work in progress to improve coverage

Ongoing Efforts

- Evaluate new AI accelerators offerings and incorporate promising solutions as part of the testbed
- Integrate AI testbed systems with the PBSPro scheduler to facilitate effective job scheduling across the accelerators
- Evaluate traditional HPC on AI Accelerators
- Understand how to integrate AI accelerators with ALCF's existing and upcoming supercomputers to accelerate science insights

AI Testbed Community Engagement





- AI training workshops
  Cerebras: https://events.cels.anl.gov/event/420/
  SambaNova: https://events.cels.anl.gov/event/421/
  Graphcore: https://events.cels.anl.gov/event/422/
  Groq: https://events.cels.anl.gov/event/448/

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

Tutorial at SC23 on Programming Novel AI accelerators for Scientific Computing ***in collaboration with Cerebras, Intel Habana, Graphcore, Groq and SambaNova***

ALCF AI4Science Training Series

https://www.alcf.anl.gov/alcf-ai-science-training-series

Intro to AI-driven Science on Supercomputers: A Student Training Series

February 6 - March 26, 2024
3-4:30 p.m. CT

Useful Links

ALCF AI Testbed

- Overview: https://www.alcf.anl.gov/alcf-ai-testbed
- Guide: https://docs.alcf.anl.gov/ai-testbed/getting-started/
- Training:
  - Slides: https://www.alcf.anl.gov/ai-testbed-training-workshops
  - Videos: https://t.ly/X0fOj

- Allocation Request: Allocation Request Form
- Support: support@alcf.anl.gov

Recent Publications

- **A Comprehensive Performance Study of Large Language Models on Novel AI Accelerators**
  Murali Emani, Sam Foreman, Varuni Sastry, Zhen Xie, Siddhisanket Raskar, William Arnold, Rajeev Thakur, Venkatram Vishwanath, Michael E. Papka, **https://arxiv.org/abs/2310.04607**

- **Efficient algorithms for Monte Carlo particle transport on AI accelerator hardware**
  John Tramm [a], Bryce Allen [a b], Kazutomo Yoshii [a], Andrew Siegel [a], Leighton Wilson, Computer Physics Communications

- **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***

- **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

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, Varuni Sastry, Sid Raskar, Zhen Xie, Rajeev Thakur, Bruce Wilson, 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

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