Welcome to the Tutorial

Programming Novel AI Accelerators for Scientific Computing

ISC25
June 13, 2025

Murali Emani (ANL)
memani@anl.gov

Varuni Sastry (ANL), Sid Raskar (ANL), Leighton Wilson (Cerebras), Petro Jr. Milan (SambaNova), Sanjif Shanmughavelu (Groq)

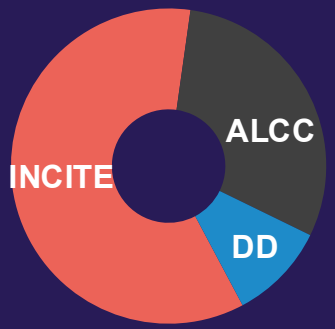
Argonne Leadership Computing Facility

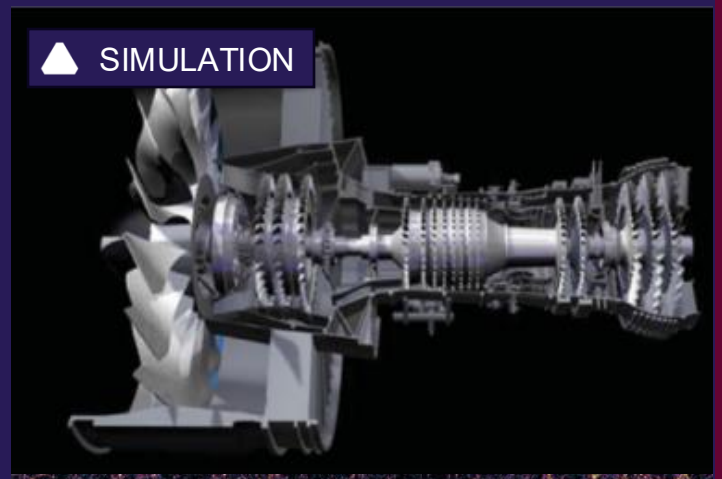
ALCF provides supercomputers to enable scientists to:

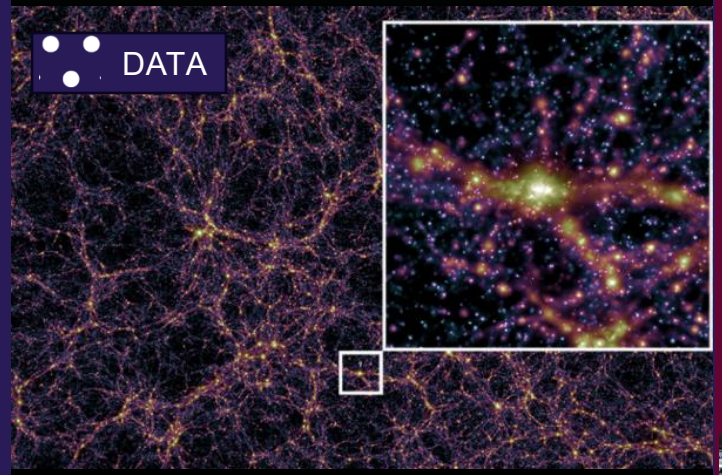
- To address grand challenges for the nation
- Perform research that is too complex and expensive to do in a laboratory setting
- Keep the nation safe and competitive

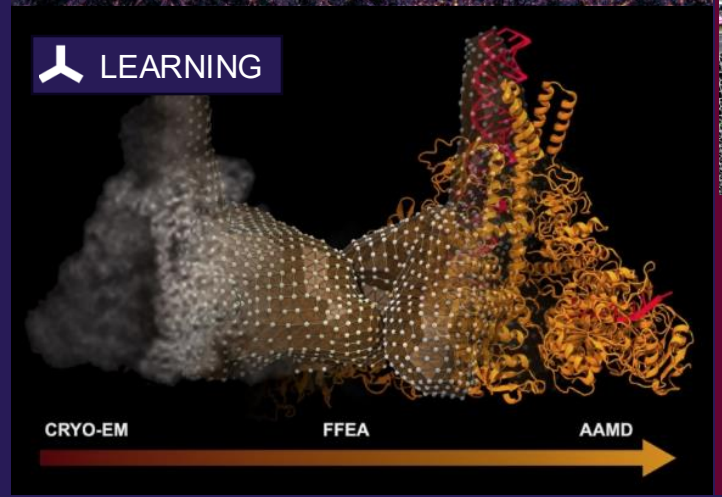
Researchers with a large-scale computing problem can apply to use ALCF resources.

- Multiple allocation award programs available to fit your needs









COMPUTING RESOURCES

#2 AI (HPL-MxP) supercomputer and > 1 exaFLOPS.

System supports 3 types of computing:

- Large-scale Simulations
- Data Intensive Applications
- AI for Science

ALCF AI TESTBED

Next-generation AI platforms to rapidly deploy and accelerate state-of-the-art AI for science

AI Accelerators

- An AI accelerator is a high-performance parallel computation machine that is specifically designed for the efficient processing of AI workloads like neural networks.
- Types of AI accelerators:
  - Graphic processing units
  - Massive multicore scalar processors
  - Dataflow architectures etc.
- Benefits
  - Improved model performance in throughput and latency
  - potential to deal with large, complex models
  - handle high-resolution datasets
  - power efficiency

Overview of ALCF AI Testbed

ALCF AI Testbed

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

Cerebras

SambaNova

Graphcore

Intel Gaudi

Groq

- 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

ALCF AI Testbed

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

https://accounts.alcf.anl.gov/#/allocationRequests

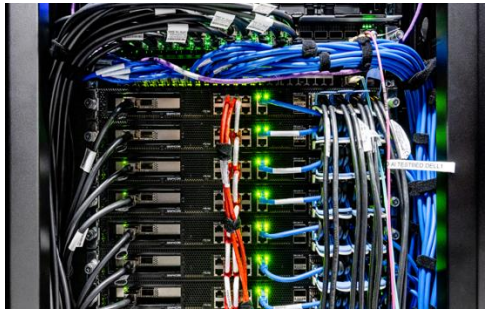
8 nodes each with 8 Reconfigurable DataFlow Units (RDU)

SambaNova SN-30

2 CS-2 Wafer scale engines (WSE)

Upgrading to CS-3

Cerebras CS-2



4 nodes each with 16 Intelligent Processing Units (IPUs)

Graphcore Bow Pod64

9 nodes each with 8 GroqChip Tensor streaming processors (TSP)

Groq

Coming Soon ! Sambanova SN40L Inference/Finetuning

NSF https://nairrpilot.org

Getting Started on ALCF AI Testbed

Available for Allocations

- Cerebras CS-2,
- SambaNova Datascale SN30,
- GroqRack
- Graphcore Bow Pod64

AI Testbed User Guide

Director's Discretionary (DD) awards

- Scaling code
- Preparing for future computing competition
- Scientific computing in support of strategic partnerships.

Allocation Request Form

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

NAIRR Pilot

aims to connect U.S. researchers and educators to computational, data, and training resources needed to advance AI research and research that employs AI.

https://nairrpilot.org/

https://docs.alcf.anl.gov/ai-testbed/getting-started/

ALCF AI Testbed

The ALCF AI Testbed provides an infrastructure for the next-generation of AI-accelerator machines.

The AI Testbed aims to help evaluate the usability and performance of machine learning-based high-performance computing applications running on these accelerators. The goal is to better understand how to integrate with existing and upcoming supercomputers at the facility to accelerate science insights.

We are currently offering allocations on our Groq, Graphcore Bow IPUs, Cerebras CS-2, and SambaNova DataScale systems.

AI Testbed Links

- Request an Allocation on Groq, Graphcore, Cerebras and/or SambaNova
- AI Testbed Guide
- AI Testbed Training
- Email us for more information

Systems

GroqRack (Available for Allocation Requests)

GroqRack Inference

Cerebras CS-2 (Available for Allocation Requests)

Cerebras CS-2 Wafer-Scale Cluster WSE-2

SambaNova Dataflow (Available for Allocation Requests)

SambaNova DataScale SN30

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:
  - https://www.alcf.anl.gov/ai-testbed-training-workshops
  - https://github.com/argonne-lcf/AIaccelerators-ISC25-tutorial/

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

Tutorial Agenda

https://github.com/argonne-lcf/Alaccelerators-ISC25-tutorial/

Agenda

Time	Topic	Duration (minutes)	Speaker
9:00 AM - 9:15 AM	Welcome and Overview of the ALCF AI Testbed	15	Murali (ANL)
9:15 AM - 10:15 AM	Cerebras: Talk + Hands-on [Slides]	60	Leighton (Cerebras)
10:15 AM - 11:00 AM	Sambanova: Talk + Hands-on [Slides]	45	Petro (SambaNova)
11:00 AM - 11:30 AM	Coffee Break	30	
11:30 AM - 11:45 PM	Sambanova: Hands-on [Slides]	15	Petro (SambaNova)
11:45 PM - 12:45 PM	Groq: Talk + Hands-on [Slides]	60	Sanjif (Groq)
12:45 PM - 1:00 PM	Q&A and Conclusion	15	Varuni (ANL)