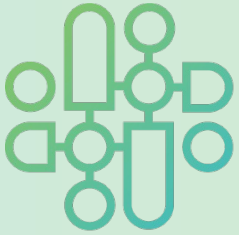


SC23
Denver, CO | i am hpc.

# Programming Novel AI Accelerators for Scientific Computing

Murali Emani

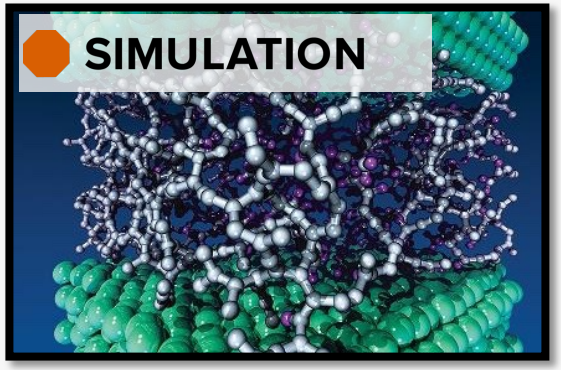
Argonne Leadership Computing Facility

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.

# 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

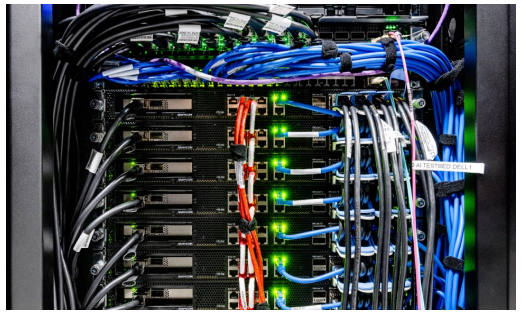
# Recent ALCF AI Testbed Updates

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

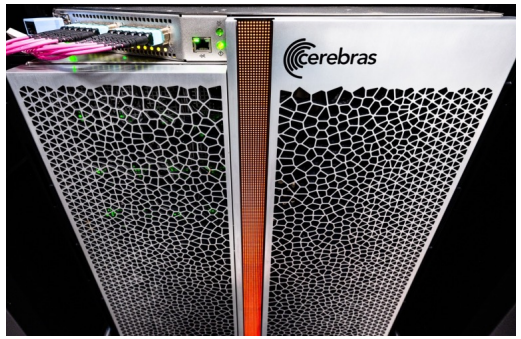
SambaNova upgraded to latest 2nd generation SN30 accelerators and scaled to 8 nodes with 64 AI accelerators

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 recently upgraded to a GroqRack with nine nodes, each consisting of eight GroqChip v1.5 Tensor streaming processors accelerators

GroqRack

# ALCF AI Testbed


The ALCF AI Testbed houses some of the most advanced AI accelerators for scientific research.

The goal of the testbed is to enable explorations into next-generation machine learning applications and workloads, enabling the ALCF and its user community to help define the role of AI accelerators in scientific computing and how to best integrate such technologies with supercomputing resources.

# Tutorial Agenda

https://github.com/argonne-lcf/AIaccelerators-SC23-tutorial

| Time (MST) | Topic |
|---|---|
| 08.30 - 8.35 | Introduction to AI Testbed at ALCF (ANL) |
| 08.35 - 8.50 | Claire Zhang (Cerebras Systems) |
| 08.50 - 9.05 | Petro Junior Milan (SambaNova Systems) |
| 09.05 - 9.20 | Alex Tsyplikhin (Graphcore) |
| 09.20 - 9.35 | Sanjif Shanmugavelu (Groq) |
| 09.35 - 9.50 | Leon Tran (Intel Habana) |
| **10.00 - 10.30** | **Break** |
| 10.30 - 12.00 | Hands session on the AI Testbed (ANL) |

# How to use ALCF AI Testbed

https://github.com/argonne-lcf/AIaccelerators-SC23-tutorial

## Request Account on AI Testbeds At ALCF

- Request an ALCF Computer User Account if you do not currently have one
- If you have an ALCF Account that is currently inactive, submit an account reactivation request*.
- If you have an active ALCF account, click Join Project to submit a membership request. Specify the following in your request: Project Name: `aitestbed_tutorial`

Contact accounts@alcf.anl.gov M-F 9am to 5pm CT. Reach out to us on slack channel `#help-accounts` on ALCF-AIAccelerator-tutorials Slack.

> SC23 Tutorial allocation will stay active till end of November 2023.

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 SN30, Graphcore Bow Pod64, and GroqRack are available for allocations

## Allocation Request Form

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

## AI Testbed User Guide

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

# AI Testbed Talks at SC23

**Sambanova**

- RDARuntime: An OS for AI Accelerators
  13th International Workshop on Runtime and Operating Systems for Supercomputers (ROSS)
  Sunday, 12 November 202310:30am - 10:54am MST
  Location704-706

**Graphcore**

- Reducing Memory Requirements for the IPU Using Butterfly Factorizations
  PMBS23: The 14th International Workshop on Performance Modeling, Benchmarking, and Simulation of High-Performance Computer Systems
  Monday, 13 November 202311am - 11:30am MST
  Location503-504
- Characterizing the Performance of Triangle Counting on Graphcore's IPU Architecture
  Tenth Workshop on Accelerator Programming and Directives (WACCPD 2023)
  Monday, 13 November 202311:20am - 11:40am MST
  Location507

**Cerebras**

- Scaling the “Memory Wall” for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems
  ACM Gordon Bell Finalists Presentations 2
  Wednesday, 15 November 20234:30pm - 5pm MST
  Location501-502

**Groq**

- Exploring Converged HPC and AI on the Groq AI Inference Accelerator
  Exhibitor Forum: Super Intelligence II
  Thursday, 16 November 202310:30am - 11am MST
  Location503-504
- From Stencils To Tensors: Running 3D Finite Difference Seismic Imaging on the Groq AI Inference Accelerator
  Exhibitor Forum: Super Intelligence II
  Thursday, 16 November 202311:30am - 12pm MST
  Location503-504

# 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