

ALCF AI Testbed Groq AI Training Workshop

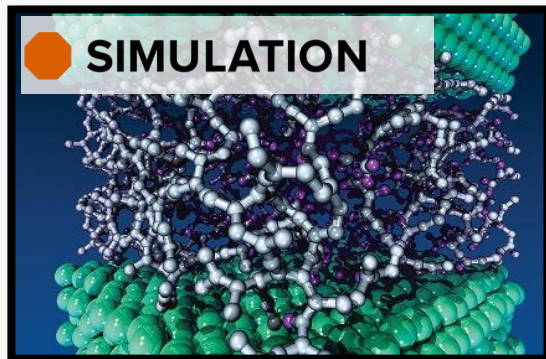
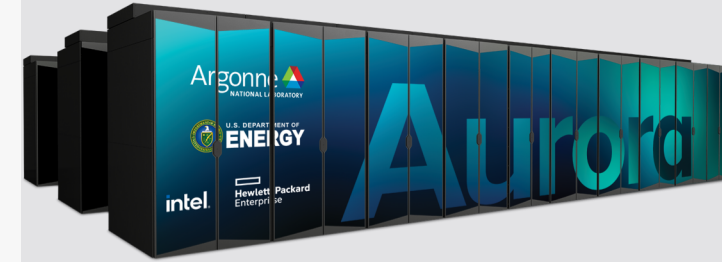
Day 1
December 6, 2023

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



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



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 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

Groq @ ALCF

ANNOUNCEMENTS

Argonne deploys new Groq system to ALCF AI Testbed, providing AI accelerator access to researchers globally

AUTHOR MARIAH LARWOOD AND BETH CERNY
PUBLISHED 10/17/2023
SYSTEMS AI TESTBED



Groq @ ALCF

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



GroqRack (Available for Allocation Requests)

GroqRack Inference

System Size: 72 Accelerators (9 nodes x 8 Accelerators per node)

Compute Units per Accelerator: 5120 vector ALUs

Performance of a single accelerator (TFlops): >188 (FP16) >750 (INT8)

Software Stack Support: GroqWare SDK, ONNX

Interconnect: RealScale TM

Argonne Leadership Computing Facility

ALCF Resources

Science

Community and Partnerships

About

Support Center

tracked from login nodes.

ALCF User Guides

ALCF

Polaris >

Theta >

ThetaGPU >

AI Testbed >

Getting Started

Cerebras >

Graphcore >

Groq >

System Overview

Getting Started

Running a Model/Program

Virtual Environments

Job Queueing and
Submission

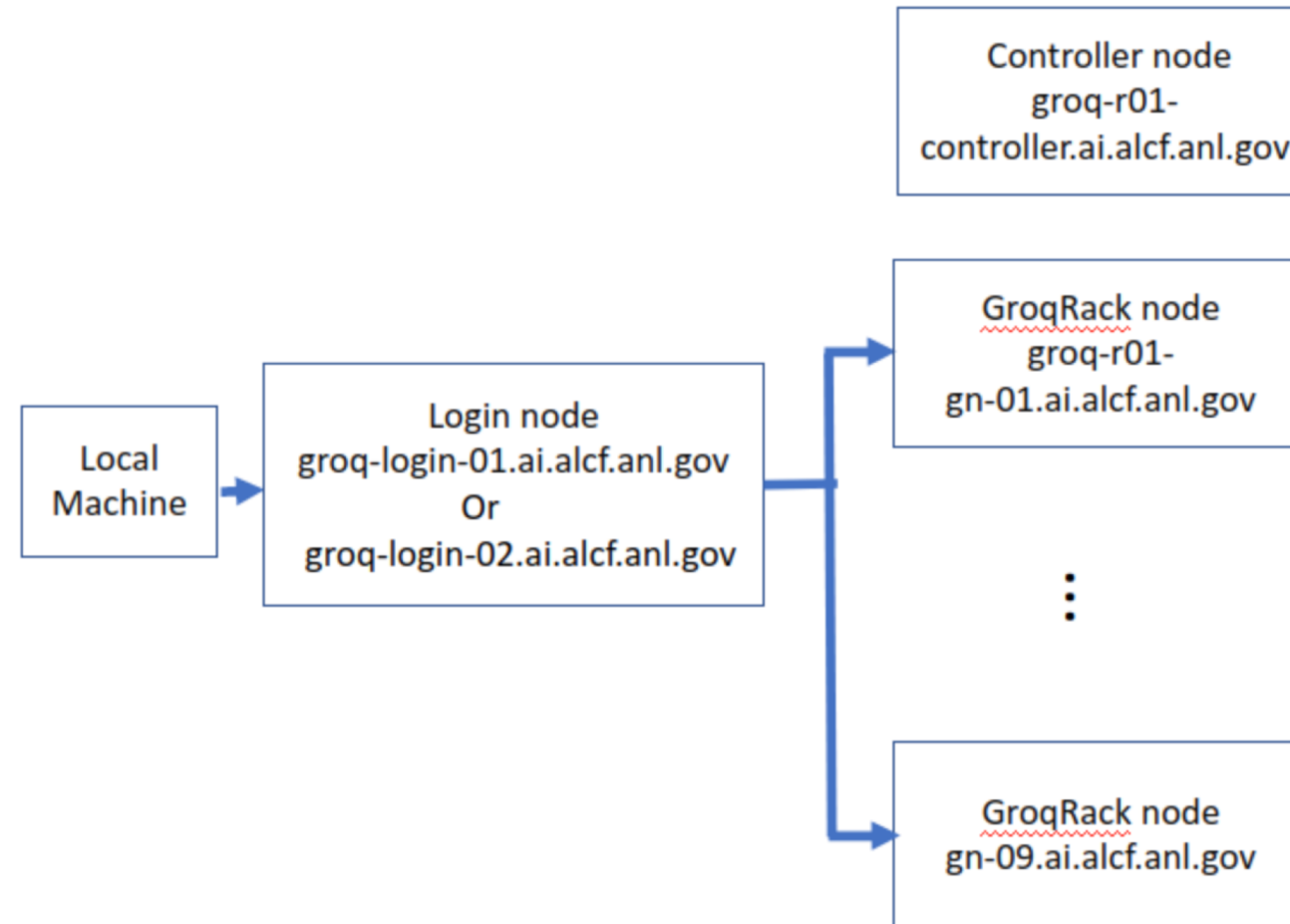
SambaNova >

Data Management

Cooley >



Aurora/Sunspot

Facility Policies >




Agenda - Day 1

<https://events.cels.anl.gov/event/448>

| WEDNESDAY, 6 DECEMBER | | |  |
|-----------------------|-----------|--|---|
| 1:30 PM | → 1:35 PM | Intro to ALCF AI Testbed | 🕒 5m |
| | |  Intro_AITestbed-Gro... | |
| 1:35 PM | → 1:40 PM | Welcome to Groq | 🕒 5m |
| 1:40 PM | → 2:10 PM | Groq Language Processing Unit™ (LPU) Architecture | 🕒 30m |
| 2:10 PM | → 2:25 PM | Intro to MLAGility™ and GroqFlow™ | 🕒 15m |
| 2:25 PM | → 3:25 PM | Porting Models with GroqFlow™ | 🕒 1h |
| 3:25 PM | → 3:40 PM | Break | 🕒 15m |
| 3:40 PM | → 4:25 PM | Benchmarking Models with MLAGility™ | 🕒 45m |
| 4:25 PM | → 5:10 PM | Accessing GroqRack™ at ALCF AI Testbed | 🕒 45m |

Agenda - Day 2

<https://events.cels.anl.gov/event/448>

| THURSDAY, 7 DECEMBER | | |  |
|----------------------|-----------|--|---|
| 1:30 PM | → 1:50 PM | Groq Compiler™ Overview | 🕒 20m |
| 1:50 PM | → 2:10 PM | Groq Runtime™ Overview | 🕒 20m |
| 2:10 PM | → 3:10 PM | Large Language Models (LLMs) and Llama-2 7B Deep Dive | 🕒 1h |
| 3:10 PM | → 3:25 PM | Break | 🕒 15m |
| 3:25 PM | → 4:10 PM | GroqWare Suite™ Developer Tools | 🕒 45m |
| 4:10 PM | → 4:30 PM | Enabling Research with Groq | 🕒 20m |

Getting Started on ALCF AI Testbed:

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

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.

GroqRack at ALCF is available for user 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>

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, Varuni Sastry, William Arnold, Bruce Wilson, 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.
- Our current AI testbed system vendors – Cerebras, Graphcore, Groq, Intel Habana and SambaNova. There are ongoing engagements with other vendors.

Please reach out for further details
Venkat Vishwanath, Venkat@anl.gov
Murali Emani, memani@anl.gov