

ALCF AI Testbed Cerebras AI Training Workshop

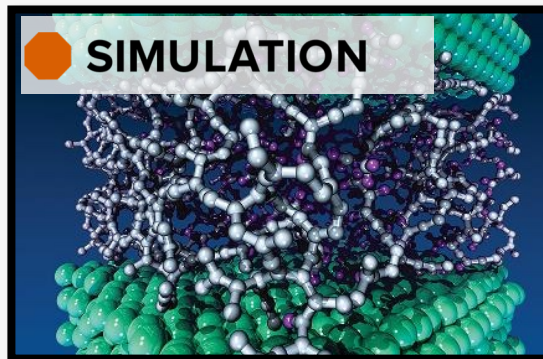
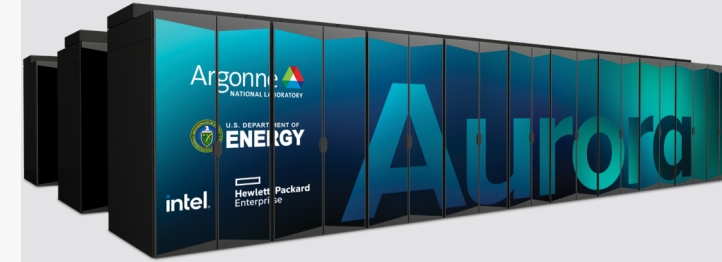
May 6, 2024

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 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

ALCF AI Testbed

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



Cerebras CS-2



SambaNova DataScale
SN30



Graphcore
Bow Pod64




Habana
Gaudi1



GroqRack

- **Cerebras: 2 CS-2 nodes, each wafer-scale engine (WSE) with 850,000 Cores, weight-streaming technology**
- SambaNova: DataScale SN30 8 nodes (8 SN30 RDUs per node) - 1TB mem per device
- Graphcore: Bow Pod64 4 nodes (16 IPUs per node) - MIMD
- GroqRack: 9 nodes, 8 GroqNodes per node -
- Habana Gaudi1: 2 nodes, 8 cards per node - On-chip integration of RDMA over Converged Ethernet (RoCE2)



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 allocation :

* Director's Discretionary (DD) Allocation Award

* <https://nairrpilot.org>

Cerebras CS-2, SambaNova SN30, Graphcore Bow Pod64, and GroqRack at ALCF are 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>

Agenda

<https://events.cels.anl.gov/event/495/>

TUESDAY, 7 MAY			
1:00 PM	→ 1:20 PM	Cerebras CS-2 Introduction	🕒 20m 
1:20 PM	→ 1:35 PM	Hardware and Systems	🕒 15m 
1:35 PM	→ 1:50 PM	Software and Programming	🕒 15m 
1:50 PM	→ 2:00 PM	Break	🕒 10m
2:00 PM	→ 2:30 PM	How-to: Model porting, layer API, data loaders	🕒 30m 
2:30 PM	→ 2:45 PM	Huggingface to CS-2 overview	🕒 15m 
2:45 PM	→ 3:05 PM	How-to: Monitoring and profiling	🕒 20m 
3:05 PM	→ 3:15 PM	Break	🕒 10m
3:15 PM	→ 4:00 PM	Hands-on session for training at ALCF	🕒 45m 
4:00 PM	→ 4:30 PM	Release 2.2.1 highlights	🕒 30m 

Documentation

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

Argonne Leadership Computing Facility

ALCF Resources

Science and Engineering

Community and Outreach

About

Support Center

ALCF User Guides

Home

Account and Project Management >

Data Management >

Services >

Running Jobs with PBS at the ALCF >

Polaris >

AI Testbed >

Getting Started

Cerebras >

System Overview

[Getting Started](#)

Running a Model/Program

Customizing Environments

Job Queuing and Submission

Example Programs

Tunneling and Forwarding

Ports

Miscellaneous

Graphcore >

Groq >

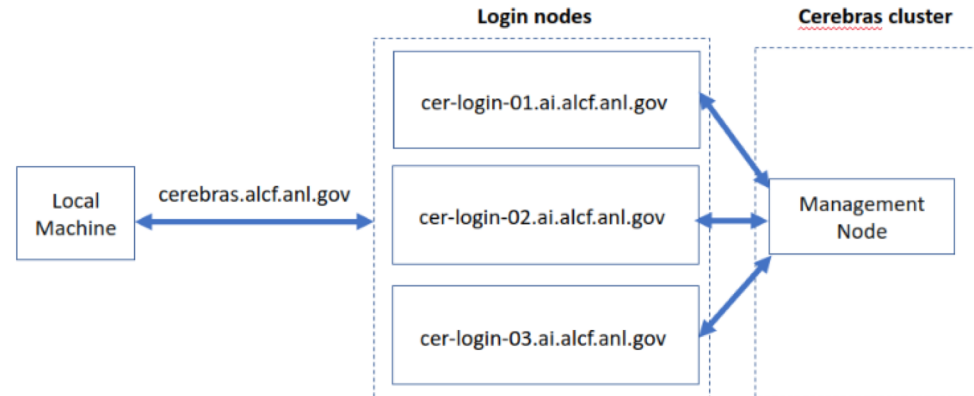
SambaNova >

Data Management >

Aurora/Sunspot >

Getting Started

Connection to a CS-2 node



Connection to one of the **CS-2** cluster login nodes requires an MFA passcode for authentication - either an 8-digit passcode generated by an app on your mobile device (e.g. MobilePASS+) or a CRYPTOCARD-generated passcode prefixed by a 4-digit pin. This is the same passcode used to authenticate into other ALCF systems, such as Polaris.

*In the examples below, **replace ALCFUserID with your ALCF user id.***

To connect to a CS-2 login:

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>
- **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 Hardy, Tom Burnley, Lei Huang, John McCalpin, Murali Emani, Hyenseung Yoo, Junqi Yin, Aristeidis Tsaris, Vishal Subbiah, Tanveer Raza, Jessica 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.

* Finalist 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, Murali Emani, 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.
- 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