Artificial Intelligence Testbed at Argonne National Laboratory

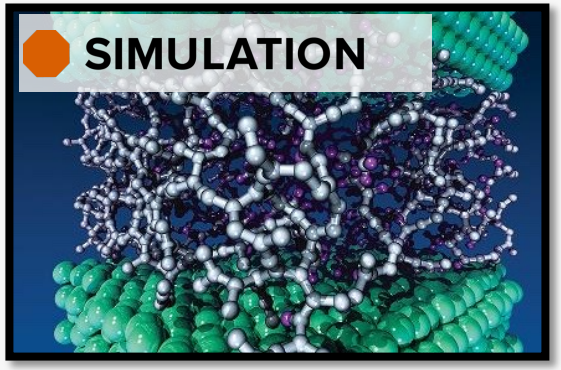
SciFM Summer school
July 17, 2024

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

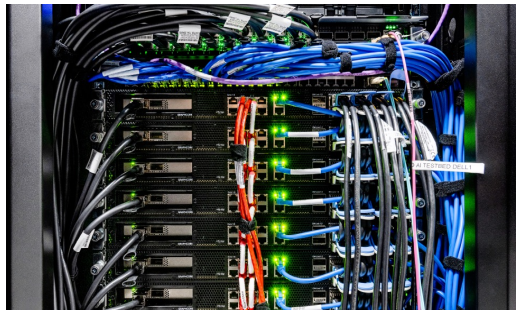
- Infrastructure of next-generation machines with AI hardware accelerators
- Provide a platform to evaluate usability and performance of AI4S applications
- Understand how to integrate AI systems with supercomputers to accelerate science

ALCF AI Testbed

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

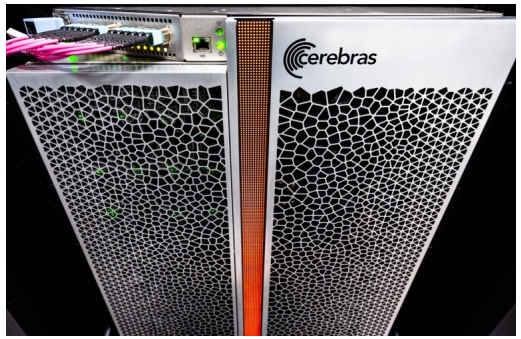
SambaNova SN30

SN30 DataScale
8 nodes with 64 AI accelerators (RDU)



Graphcore BowPod64

Bow generation accelerators **Pod-64 configuration with 64 accelerators (IPU)**



Cerebras CS-2

Two CS-2 WSEs and appliance mode to include Memory-X and Swarm-X technologies to enable larger models

GroqRack

GroqRack **9 nodes with 72 LPU accelerators**

https://nairrpilot.org

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 Pod64, Groq 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://github.com/argonne-lcf/summer-school-2024/

Tutorial at SciFM Summer-school-2024 at University of Michigan

Date	Wednesday, 17 July 2024
Time	10:30 AM - 4:45 PM EST

Agenda

Time (EST)	Topic
10.30 - 10:45	Murali Emani(ANL) [Slides]
10.45 - 11.15	Sylvia Howland (Cerebras Systems) [Slides]
11.15 - 11.45	Vijay Tatkar (SambaNova Systems) [Slides]
11.45 - 12.00	**Break**
12.00 - 12.30	Chad Martin (Graphcore)[Slides]
12.30 - 01.00	Sanjif Shanmugavelu, Hatice Ozen (Groq) [Slides]
01.00 - 02.00	**Lunch**
02.00 - 02:30	Sam Foreman (ANL) (LLMs on Nvidia)
02.30 - 04:00	Hands session on the AI Testbed: Sid Raskar, Varuni Sastry (ANL)
04.00 - 04.15	**Break**
04.15 - 04:45	Open Discussion, Q/A

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

- **Centimani: Enabling Fast AI Accelerator Selection for DNN Training with a Novel Performance Predictor**
  Zhen Xie, Murali Emani, Xiaodong Yu, Dingwen Tao, Xin He, Pengfei Su, Keren Zhou, Venkatram Vishwanath *USENIX ATC 2024*

- **WActiGrad: Structured Pruning for Efficient Finetuning and Inference of Large Language Models on AI Accelerators**
  Krishna Teja Chitty-Venkata, Varuni Katti Sastry, Murali Emani, Venkatram Vishwanath, Sanjif Shanmugavelu,Sylvia Howland *Euro-Par 2024*

- **Toward a Holistic Performance Evaluation of Large Language Models Across Diverse AI Accelerators**
  Murali Emani, Sam Foreman, Varuni Sastry, Zhen Xie, Siddhisanket Raskar, William Arnold, Rajeev Thakur, Venkatram Vishwanath, Michael E. Papka, Sanjif Shanmugavelu, Darshan Gandhi, Dun Ma, Kiran Ranganath, Rick Weisner, Jiunn-yeu Chen, Yuting Yang, Natalia Vassilieva, Bin C. Zhang, Sylvia Howland, Alexandar Tsyplikhin *Heterogeneity in Computing Workshop (HCW'24) at IPDPS24*
  https://arxiv.org/abs/2310.04607

- **Efficient algorithms for Monte Carlo particle transport on AI accelerator hardware**
  John Tramm, Bryce Allen, Kazutomo Yoshii, Andrew Siegel, 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,

Recent Publications

- **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

- **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

Recent Publications

- **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, 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