

ALCF AI Testbed

**SC24 BoF: Advanced Architecture "Playgrounds" — Past Lessons,
Current and Future Accesses of Testbeds**

November 21, 2024

Murali Emani

Argonne Leadership Computing Facility (ALCF)

memani@anl.gov

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

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

<https://nairrpilot.org>



SambaNova 2nd generation SN-30 accelerators - 8 nodes with 64 AI accelerators

SambaNova SN-30



Graphcore Bow generation accelerators - Pod-64 configuration with 64 accelerators

Graphcore



Cerebras, 2x CS-2 WSE with Memory-X and Swarm-X technologies

Cerebras CS-2



GroqRack - nine nodes, with eight GroqChip v1.5 Tensor streaming processors accelerators per node

Groq

Coming soon: Tenstorrent and SambaNova SN40L

Argonne Leadership Computing Facility

[ALCF Resources](#) [Science](#) [Community and Partnerships](#) [About](#) [Support Center](#)

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

ALCF User Guides

[Home](#)[Account and Project Management](#)[Data Management](#)[Services](#)[Running Jobs with PBS at the ALCF](#)[Polaris](#)[Theta](#)[ThetaGPU](#)[AI Testbed](#)[Getting Started](#)[Cerebras](#)[Graphcore](#)[Groq](#)[SambaNova](#)[Data Management](#)[Cooley](#)[Aurora/Sunspot](#)[Facility Policies](#)

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.

Table of contents

[How to Get Access](#)[Getting Started](#)[How to Contribute to Documentation](#)

Community Engagement

AI Testbed Training Workshops

Learn how to leverage the ALCF AI Testbed for your science



- AI training workshops

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

Full Program My Schedule Contributors Organizations

Presentation

Programming Novel AI Accelerators for Scientific Computing

Description: Scientific applications are increasingly adopting Artificial Intelligence (AI) techniques to advance science. There are specialized hardware accelerators designed and built to run AI applications efficiently. With a wide diversity in the hardware architectures and software stacks of these systems, it is challenging to understand the differences between these accelerators, their capabilities, programming approaches, and how they perform, particularly for scientific applications. In this tutorial, we will cover an overview of the AI accelerators landscape focusing on SambaNova, Cerebras, Graphcore, Groq, and Habana systems along with architectural features and details of their software stacks. We will have hands-on exercises to help attendees understand how to program these systems by learning how to refactor codes and compile and run the models on these systems. The tutorial will provide the attendees with an understanding of the key capabilities of emerging AI accelerators and their performance implications for scientific applications

Event Type: Tutorial

Add to Schedule

Time:
Sunday, 17 November 2024
8:30am - 5pm EST

Location: B201

Tags:
Basic and Introductory Topics for Expanding Broader Engagement,
Machine Learning, Deep Learning and Artificial Intelligence for HPC,
Software Tools for Accelerators (Co-processors, GPGPUs, FPGA, etc.).

Presenters

Murali Emani
Argonne National Laboratory (ANL)

Leighton Wilson
Cerebras Systems

NEXT PRESENTATION > **STARTS IN 118:23:07**

Programming Your GPU With OpenMP: A "Hands-On Introduction"

Tutorial at SC24 on Programming Novel AI accelerators for Scientific Computing *in collaboration with Cerebras, Intel Habana, Graphcore, Groq and SambaNova*

DNN Performance on AI Accelerators

A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads

Murali Emani* Zhen Xie* Siddhisanket Raskar* Varuni Sastry* William Arnold* Bruce Wilson*
memani@anl.gov zhen.xie@anl.gov sraskar@anl.gov vsastry@anl.gov arnoldw@anl.gov wilsonb@anl.gov

Rajeev Thakur* Venkatram Vishwanath* Zhengchun Liu* Michael E. Papka*^{||} Cindy Orozco Bohorquez[†]
thakur@anl.gov venkat@anl.gov zhengchun.liu@anl.gov papka@anl.gov cindy@cerebras.net

Rick Weisner[‡] Karen Li[‡] Yongning Sheng[‡] Yun Du[‡]
rick.weisner@sambanova.ai xiaoyan.li@sambanova.ai yongning.sheng@sambanova.ai yun.du@sambanova.ai

Jian Zhang[‡] Alexander Tsypikhin[§] Gurdaman Khaira[§] Jeremy Fowers[¶] Ramakrishnan Sivakumar[¶]
jian.zhang@sambanova.ai alext@graphcore.ai damank@graphcore.ai jfowers@groq.com rsivakumar@groq.com

Victoria Godsoe[¶] Adrian Macias[¶] Chetan Tekur[¶] Matthew Boyd[¶]
vgodsoe@groq.com am@groq.com ctekur@groq.com matt@groq.com

*Argonne National Laboratory, Lemont, IL 60439, USA, [†]Cerebras Systems, Sunnyvale, CA 95085, USA,

[‡]SambaNova Systems Inc., Palo Alto, CA 94303, USA, [§]Graphcore Inc., Palo Alto, CA 94301, USA,

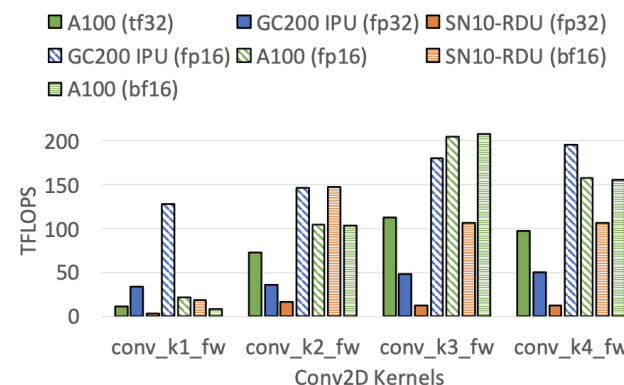
[¶]Groq Inc., Mountain View, CA 94041, USA, ^{||}University of Illinois, Chicago, IL 60637, USA

Abstract—Scientific applications are increasingly adopting Artificial Intelligence (AI) techniques to advance science. High-performance computing centers are evaluating emerging novel hardware accelerators to efficiently run AI-driven science applications. With a wide diversity in the hardware architectures and software stacks of these systems, it is challenging to understand how these accelerators perform. The state-of-the-art in the evaluation of deep learning workloads primarily focuses on CPUs and GPUs. In this paper, we present an overview of dataflow-based novel AI accelerators from SambaNova, Cerebras, Graphcore, and Groq. We present a first-of-a-kind evaluation of these accelerators with diverse workloads, such as Deep Learning (DL) primitives, benchmark models, and scientific machine learning applications. We also evaluate the performance of collective communication, which is key for distributed DL implementation, along with a study of scaling efficiency. We then discuss key insights, challenges, and opportunities in integrating these novel AI accelerators in supercomputing systems.

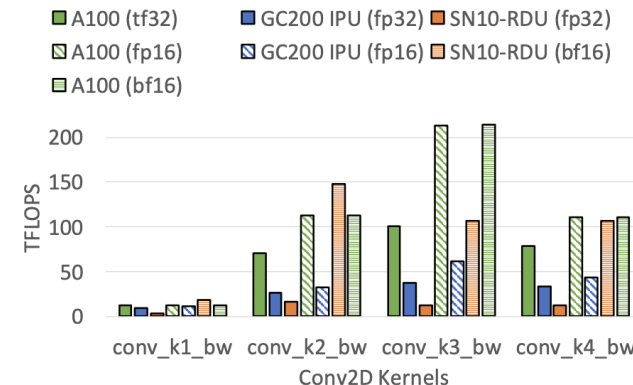
Index Terms—Scientific Machine Learning, Deep Learning, Accelerators, Performance Evaluation, Benchmarking

the above. There will be a surge in scientific applications that require infrastructure to enable in-place data analysis at experimental facilities and AI capabilities integrated with large-scale models. The US Department of Energy (DOE) AI for Science Report [1], put forth by stakeholders from DOE labs, academia, and industry, cohesively highlights the need for tighter integration of the AI infrastructure ecosystem with experimental and leadership computing facilities. There is great emphasis on efficiently implementing Deep Learning (DL) models and exploiting novel architectures, especially reduced-precision AI accelerators. The DOE Advanced Scientific Computing Research (ASCR) report on extreme heterogeneity [2] lists challenges in integrating a broad spectrum of diverse hardware resources for science.

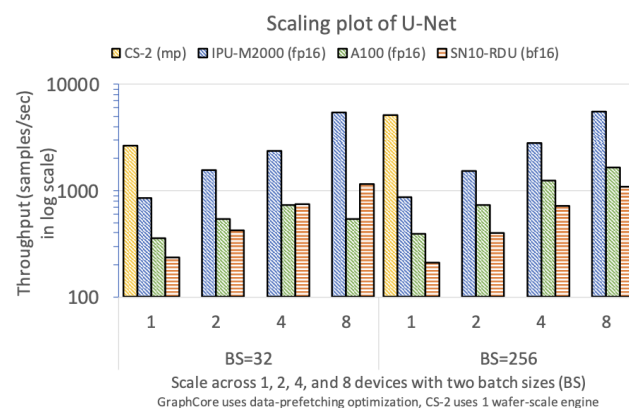
Recent advances in hardware, including heterogeneous systems and AI accelerators, will help researchers to advance the state of the art in scientific applications on powerful exascale supercomputers such as Aurora [3], El Capitan [4],



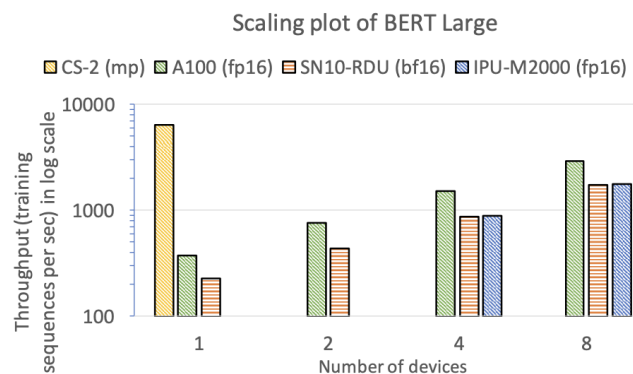
(a) Training mode, forward pass



(b) Training mode, backward pass



Scale across 1, 2, 4, and 8 devices with two batch sizes (BS)
GraphCore uses data-prefetching optimization, CS-2 uses 1 wafer-scale engine



GC200 needs at least 4 IPU and CS-2 uses 1 wafer-scale engine

<https://ieeexplore.ieee.org/abstract/document/10024028>

LLM Performance on AI Accelerators

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*
memani@anl.gov foremans@anl.gov vsastry@anl.gov zxie3@binghamton.edu sraskar@anl.gov arnoldw@anl.gov

Rajeev Thakur* Venkatram Vishwanath* Michael E. Papka*‡ Sanjif Shanmugavelu§
thakur@anl.gov venkat@anl.gov papka@anl.gov sshanmugavelu@groq.com

Darshan Gandhi¶ Hengyu Zhao¶ Dun Ma¶
darshan.gandhi@sambanovsystems.com hengyu.zhao@sambanovsystems.com eric.ma@sambanovsystems.com

Kiran Ranganath¶ Rick Weisner¶ Jiunn-yeu Chen|| Yuting Yang||
kiran.ranganath@sambanovsystems.com rick.weisner@sambanovsystems.com jchen@habana.ai yyang@habana.ai

Natalia Vassilieva†† Bin C. Zhang†† Sylvia Howland†† Alexander Tsyplikhin**
natalia@cerebras.net claire.zhang@cerebras.net Sylvia.Howland@cerebras.net alext@graphcore.ai

*Argonne National Laboratory, Lemont, IL 60439, USA

†State University of New York, Binghamton, NY, 13092, USA,

‡University of Illinois, Chicago, IL 60637, USA, §Groq Inc., Mountain View, CA 94041, USA

¶SambaNova Systems Inc., Palo Alto, CA 94303, USA, || Intel Habana, Santa Clara CA 95054, USA

**Graphcore Inc., Palo Alto, CA 94301, USA, †† Cerebras Systems, Sunnyvale, CA 95085, USA

Abstract—Artificial intelligence (AI) methods have become critical in scientific applications to help accelerate scientific discovery. Large language models (LLMs) are being considered a promising approach to address some challenging problems because of their superior generalization capabilities across domains. The effectiveness of the models and the accuracy of the applications are contingent upon their efficient execution on the underlying hardware infrastructure. Specialized AI accelerator hardware systems have recently become available for accelerating AI applications. However, the comparative performance of these AI accelerators on large language models has not been previously studied. In this paper, we systematically study LLMs on multiple AI accelerators and GPUs and evaluate their performance characteristics for these models. We evaluate these systems with (i) a micro-benchmark using a core transformer block, (ii) a GPT-2 model, and (iii) an LLM-driven science use case, GenSLM. We present our findings and analyses of the models' performance to better understand the intrinsic capabilities of AI accelerators. Furthermore, our analysis takes into account key factors such as sequence lengths, scaling behavior, and sensitivity to gradient accumulation steps.

prediction [2], neutrino particle detection [3], drug design for precision medicine [4], genome-scale foundation model [5] and weather forecasting models [6]. Some of the most commonly used AI techniques include convolutional neural networks, recurrent neural networks, graph neural networks, and large language models (LLMs). These techniques, with their unique architectural characteristics, have become invaluable to assist scientists in their research. Within the AI landscape, the domain of Natural Language Processing (NLP) has experienced a massive surge in growth, fostering usage of LLMs in various tasks such as question-answering, text summarization, and language translation. These models are becoming increasingly critical in scientific machine-learning applications.

LLMs, such as Generative Pre-trained Transformers (GPT) GPT-3 [7], LLaMA [8], Llama 2 [9], and Bloom [10] are diverse in their neural architectures along with the quality of results for these tasks. This growth has been driven in part by

Throughput evaluation of transformer micro-benchmark

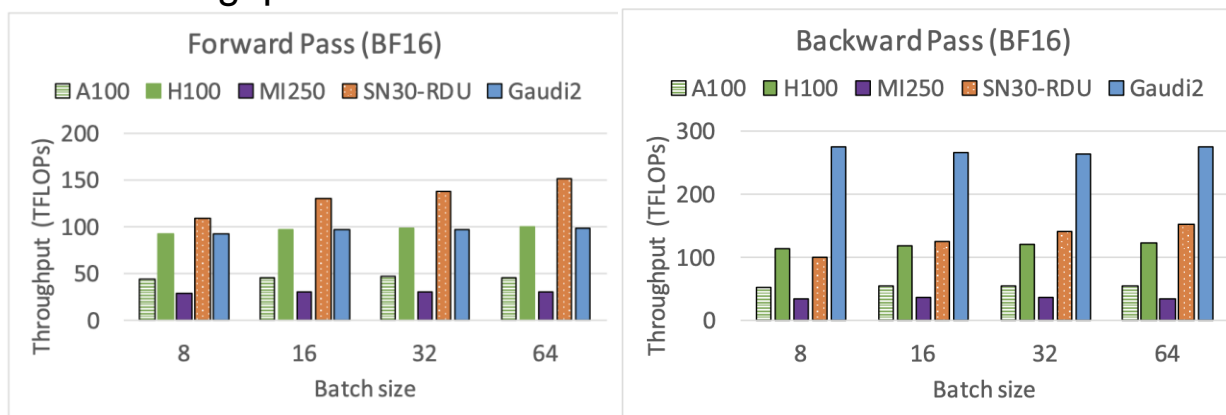


TABLE III: Scaling behavior study with the GPT-2 XL model

System	min #devices	max #devices	scale #devices	scaling efficiency	Speedup
Gaudi2	1	64	64	104%	66.4x
Bow Pod64	4	64	16	100.1%	16x
CS-2	1	2	2	99.87%	1.99x
SN30	1	64	64	97.5%	62.4x
MI250	1	4	4	80%	3.2x
A100	4	64	16	75.8%	12.1x
H100	1	4	4	43%	1.73x

<https://ieeexplore.ieee.org/abstract/document/10596441>

LLM Inference Bench

LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators

Krishna Teja Chitty-Venkata*[†] Siddhisanket Raskar*[†] Bharat Kale* Farah Ferdaus* Aditya Tanikanti*
schittyvenkata@anl.gov sraskar@anl.gov kale@anl.gov fferdaus@anl.gov atanikanti@anl.gov

Ken Raffeneti* Valerie Taylor* Murali Emani* Venkatram Vishwanath*
raffenet@anl.gov vtaylor@anl.gov memani@anl.gov venkat@anl.gov

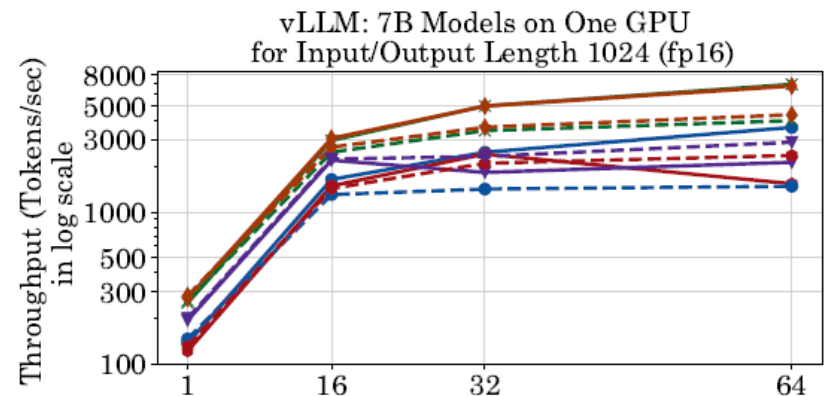
*Argonne National Laboratory, Lemont, IL 60439, USA

Abstract—Large Language Models (LLMs) have propelled groundbreaking advancements across several domains and are commonly used for text generation applications. However, the computational demands of these complex models pose significant challenges, requiring efficient hardware acceleration. Benchmarking the performance of LLMs across diverse hardware platforms is crucial to understanding their scalability and throughput characteristics. We introduce LLM-Inference-Bench, a comprehensive benchmarking suite to evaluate the hardware inference performance of LLMs. We thoroughly analyze diverse hardware platforms, including GPUs from Nvidia and AMD and specialized AI accelerators, Intel Habana and SambaNova. Our evaluation includes several LLM inference frameworks and models from LLaMA, Mistral, and Qwen families with 7B and 70B parameters. Our benchmarking results reveal the strengths and limitations of various models, hardware platforms, and inference frameworks. We provide an interactive dashboard to help identify configurations for optimal performance for a given hardware platform.

Index Terms—Large Language Models, AI Accelerators, Inference Performance Evaluation, Benchmarking

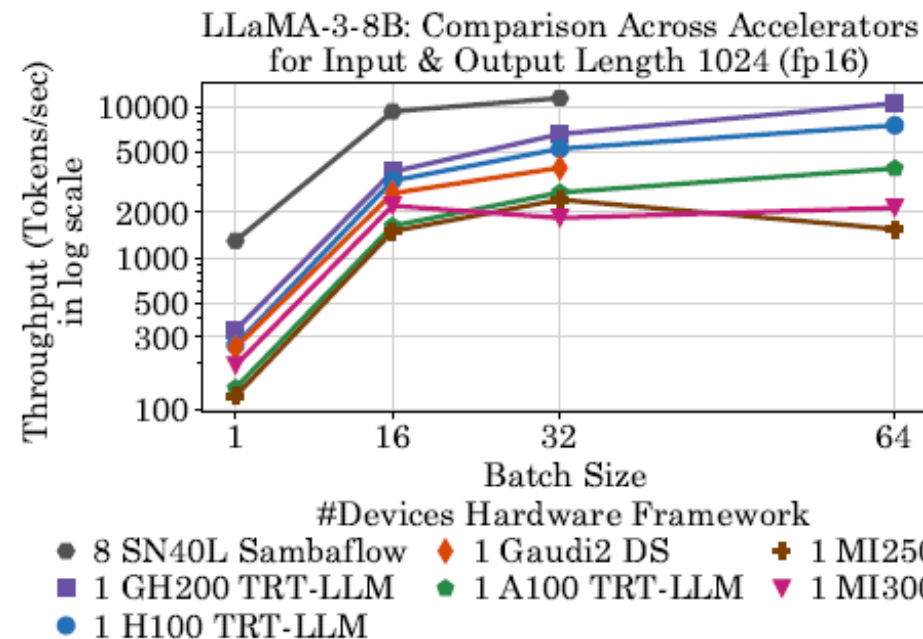
responses or make predictions. Today, efficient inference is essential for generation capabilities across various applications, such as chatbots, language translation, and information retrieval systems. As LLMs continue to grow in size and complexity, optimizing inference becomes increasingly crucial to balance performance with computational resources, energy consumption, and response times.

In recent years, the development of hardware accelerators for Deep Learning (DL) applications, such as GPUs and TPUs, has been driven to meet the computational demands of large models. These accelerators are designed to enhance performance and energy efficiency, which is particularly crucial for LLMs that consist of billions of parameters. These hardware solutions significantly improve performance, including faster training times, reduced inference latency, and enhanced scalability. This is essential for developing and deploying sophisticated models capable of handling state-of-the-art (SOTA) tasks in NLP, content generation, and decision support systems. The



Batch Size
Hardware & Model

- H100 LLaMA-3-8B
- A100 LLaMA-3-8B
- GH200 LLaMA-3-8B
- MI250 LLaMA-3-8B
- MI300X LLaMA-3-8B
- H100 LLaMA-2-7B
- A100 LLaMA-2-7B
- MI250 LLaMA-2-7B
- MI300X LLaMA-2-7B
- GH200 LLaMA-2-7B



#Devices Hardware Framework

- 8 SN40L SambafLOW
- 1 Gaudi2 DS
- 1 MI250 vLLM
- 1 GH200 TRT-LLM
- 1 A100 TRT-LLM
- 1 MI300X vLLM
- 1 H100 TRT-LLM

<https://arxiv.org/abs/2411.00136>



GenSLM 13B Training Performance

GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics

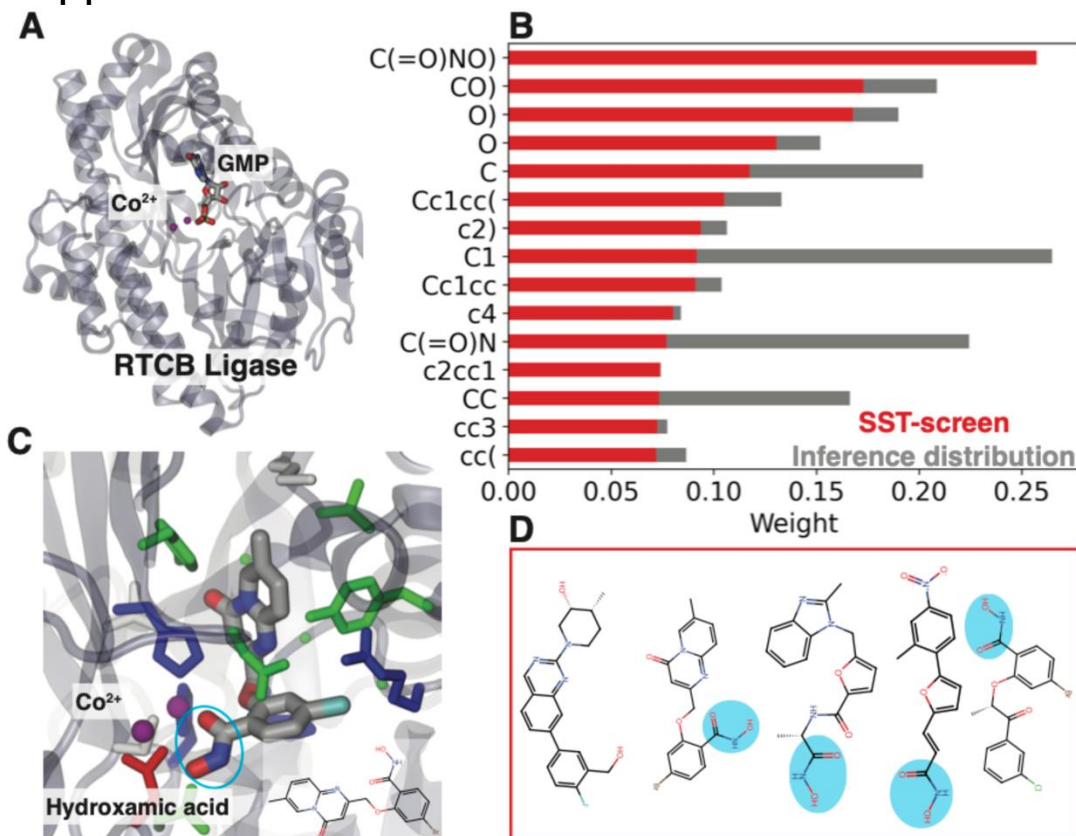
Winner of the ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research, 2022

System	Number of Devices	Throughput (tokens/sec)	Improvement	Energy Efficiency
Nvidia A100	8	1150	1.0	1.0
SambaNova SN30	8	9795	8.5	5.6
Cerebras CS-2	1	29061	25	6.5

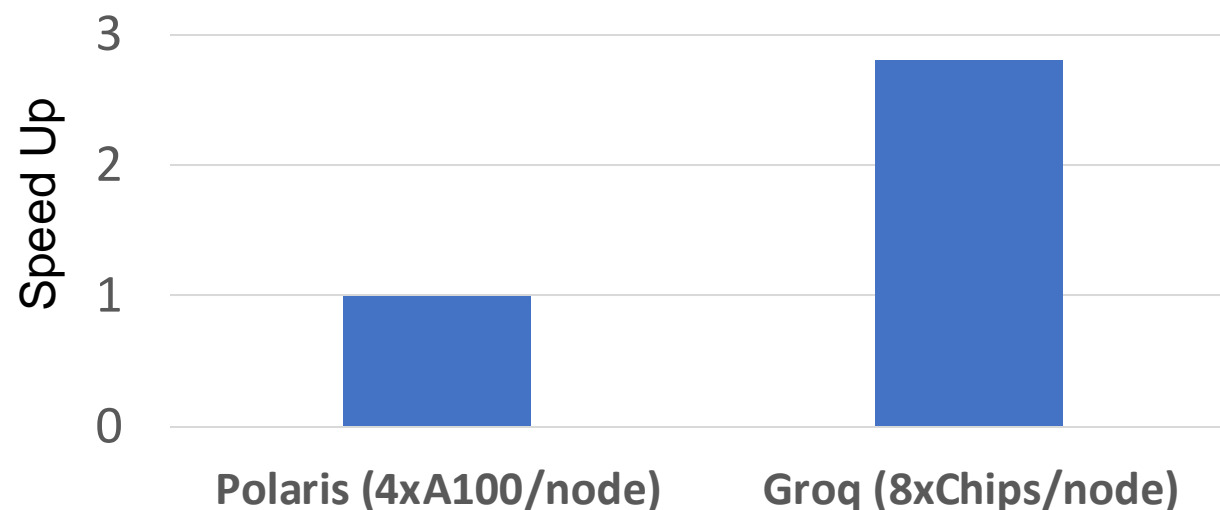
Note: We are utilizing only 40% of the CS wafer-scale engine for this problem

Accelerating Drug Design and Discovery with Machine Learning

Application code: SMILES Transformer



Initial Performance Comparison Between Inference on a Polaris (A100) Node and GroqNode



Note: Ongoing work to optimize inference on A100 using TRT-LLM

Courtesy: Archit Vasan, ANL

A. Vasan, T. Brettin, R. Stevens, A. Ramanathan, and V. Vishwanath.
Scalable Lead Prediction with Transformers using HPC
Resources. <https://doi.org/10.1145/3624062.3624081>

*Simplified Molecular Input Line Entry System (SMILES) - Representation for Molecules

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>
- Allocation Request: [Allocation Request Form](#)
- Support: support@alcf.anl.gov

Recent Publications

- **LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators**
Krishna Teja Chitty-Venkata, Siddhisanket Raskar, Bharat Kale, Farah Ferdaus, Aditya Tanikanti, Ken Raffenetti, Valerie Taylor, Murali Emani, Venkatram Vishwanath *PMBS'24*
- **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>
- **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 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>

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.

* 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, Krishna Teja-Chitty Venkata, 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.
- For any questions, please reach out to (memani@anl.gov).

Backup

Genome-scale Language Models (GenSLMs)

Goal:

- How new and emergent variants of pandemic causing viruses, (specifically SARS-CoV-2) can be identified and classified.
- Identify mutations that are VOC (increased severity and transmissibility)
- Extendable to gene or protein synthesis.

Approach

- Adapt Large Language Models (LLMs) to learn the evolution.
- Pretrain 25M – 25B models on raw nucleotides with large sequence lengths.
- Scale on GPUs, CS2s, SN30.



GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics



Winner of the ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research, 2022,

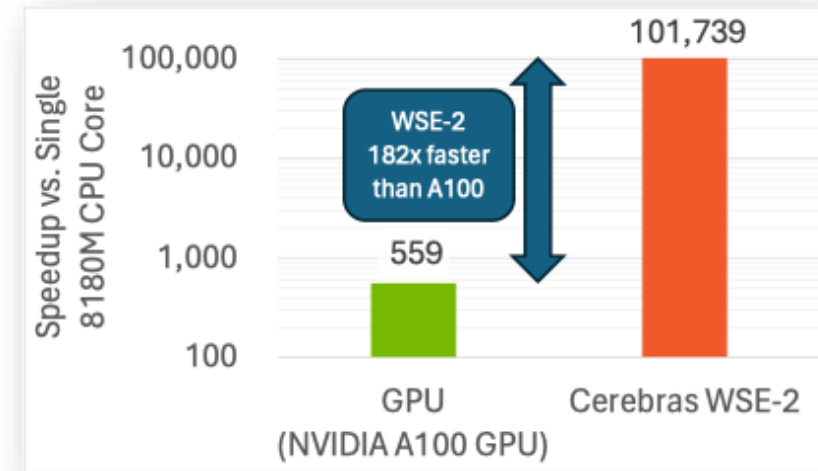
Efficient Algorithms for Monte Carlo Particle Transport on AI Accelerator Hardware

The Science

Artificial Intelligence (AI) accelerators, such as the Cerebras Wafer-Scale Engine 2 (WSE-2), are custom made to efficiently perform AI training tasks at high speed. While not intended for traditional modeling and simulation workloads, aspects of these accelerators make them attractive for some simulation algorithms, nonetheless. A team has developed new algorithms and performance optimization strategies to enable a key Monte Carlo (MC) particle transport simulation kernel to effectively use the device. Speedups of 182x over a single GPU for were observed for the kernel using a production quality workload typical of a nuclear reactor simulation.

The Impact

Significant speed and power advantages compared to highly optimized CPU and GPU implementations were found, suggesting that acceleration of a full MC particle transport code on WSE-2 would be possible. New algorithms for minimizing communication costs and for handling load balancing were developed and evaluated that could aid in the development of optimized algorithms for related simulation methods. AI accelerators, such as the WSE-2, could offer significant advantages to traditional simulation workloads and the development of higher-level programming models to more readily enable software development and exploration could have a tremendous impact for HPC simulations.



Speedup vs. serial CPU execution for macroscopic cross section lookup kernel (adapted from XSBench). (John Tramm ANL).

Contact PI: John Tramm (ANL)
ASCR Allocation PI: John Tramm (ANL)
ASCR Program/Facility: DD/ALCF
ASCR PM: Saswata Hier-Majumder
Date submitted to ASCR: October 7, 2024
Publication(s) for this work:
• J. Tramm, et. al., *Comput. Physics Commun.* **298**, 109072 (2024). doi:10.1016/j.cpc.2023.109072.
• J. Tramm, et. al., *PHYSOR* 2024. (2024). doi:10.13182/PHYSOR24-43696

ANL, U. Chicago, Cerebras