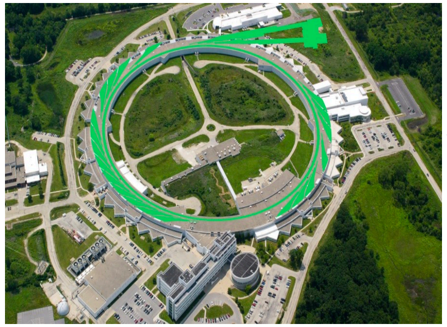
Artificial Intelligence Testbeds at Argonne National Laboratory

Murali Emani
Argonne Leadership Computing Facility
memani@anl.gov

Integrating AI Systems in Facilities

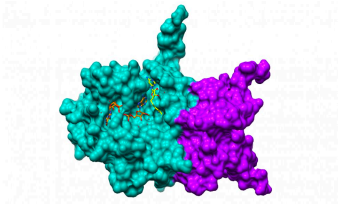
Experimental Facility

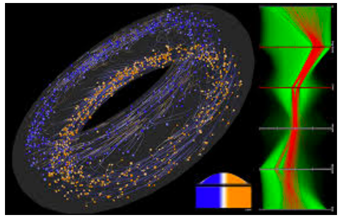


Computing Facility

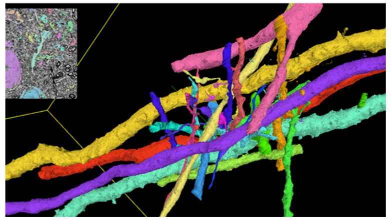
AI-Accelerators

AI-Edge accelerator





Simulations



Data-driven Models

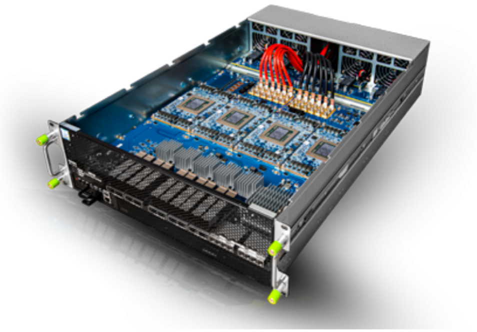
ALCF AI Testbeds

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

Cerebras (CS-2)

SambaNova

Graphcore



Habana

Groq

- Infrastructure of next-generation machines with hardware accelerators customized for artificial intelligence (AI) applications.

- Provide a platform to evaluate usability and performance of various applications running on these accelerators.

- The goal is to better understand how to integrate AI accelerators with ALCF's existing and upcoming supercomputers to accelerate science insights

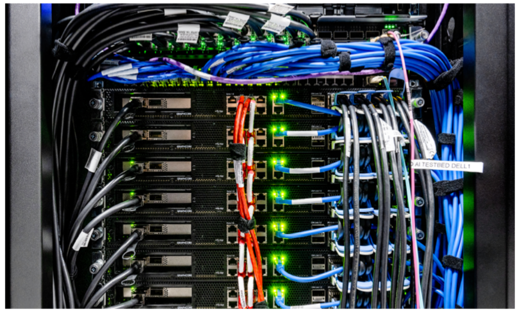
Recent ALCF AI Testbed Updates

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 (RDU)**

SambaNova SN30



Graphcore upgraded to latest Bow generation accelerators and scaled to a **Pod-64 configuration with 64 accelerators (IPU)**

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

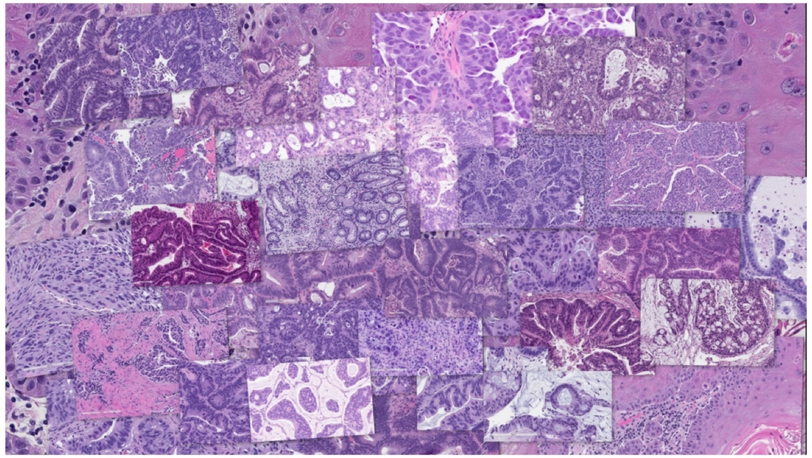
Groq system has been upgraded to a GroqRack with nine nodes, each consisting of eight GroqChip Tensor streaming processors, **72 accelerators**

GroqRack

https://nairrpilot.org

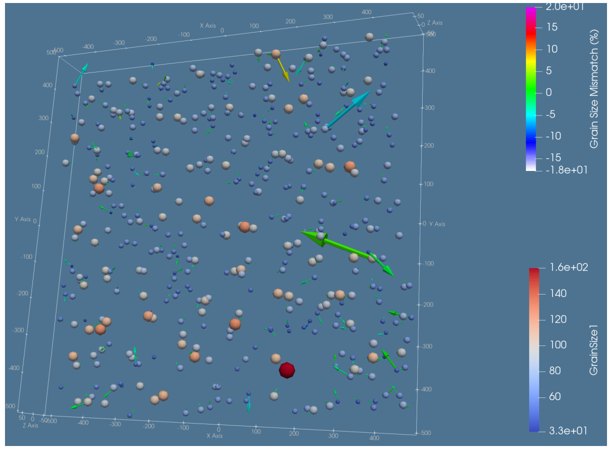
	Cerebras CS2	SambaNova Cardinal SN30	Groq GroqRack	GraphCore GC200 IPU	Habana Gaudi1	NVIDIA A100
Compute Units	850,000 Cores	640 PCUs	5120 vector ALUs	1472 IPUs	8 TPC + GEMM engine	6912 Cuda Cores
On-Chip Memory	40 GB L1, 1TB+ MemoryX	>300MB L1 1TB	230MB L1	900MB L1	24 MB L1 32GB	192KB L1 40MB L2 40-80GB
Process	7nm	7nm	7 nm	7nm	7nm	7nm
System Size	2 Nodes including Memory-X and Swarm-X	8 nodes (8 cards per node)	9 nodes (8 cards per node)	4 nodes (16 cards per node)	2 nodes (8 cards per node)	Several systems
Estimated Performance of a card (TFlops)	>5780 (FP16)	>660 (BF16)	>250 (FP16) >1000 (INT8)	>250 (FP16)	>150 (FP16)	312 (FP16), 156 (FP32)
Software Stack Support	Tensorflow, Pytorch	SambaFlow, Pytorch	GroqAPI, ONNX	Tensorflow, Pytorch, PopArt	Synapse AI, TensorFlow and PyTorch	Tensorflow, Pytorch, etc
Interconnect	Ethernet-based	Ethernet-based	RealScale TM	IPU Link	Ethernet-based	NVLink

AI FOR SCIENCE AND HPC APPLICATIONS ON AI TESTBED



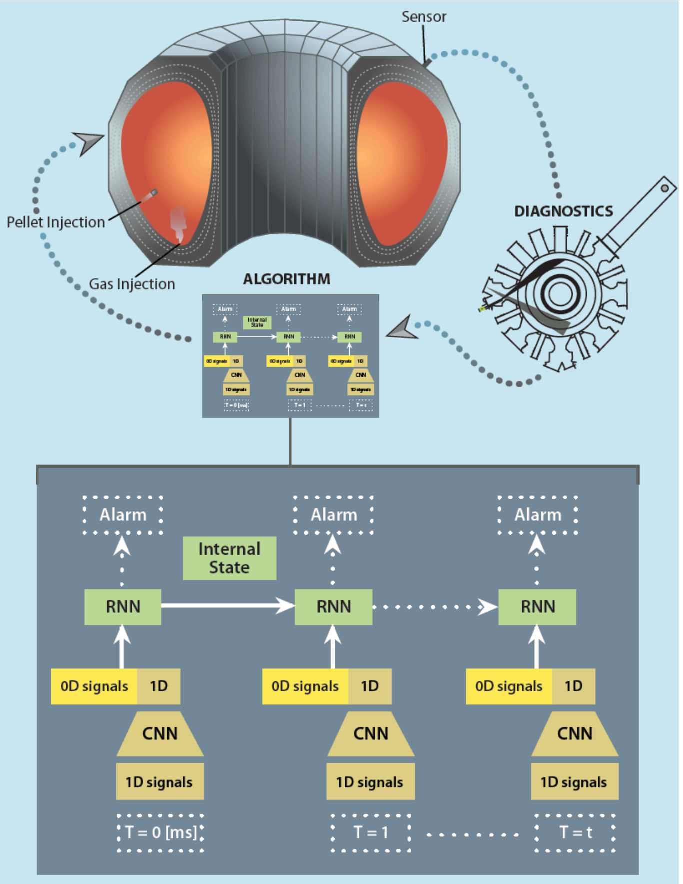
Cancer drug response prediction

(Credit: Candle)



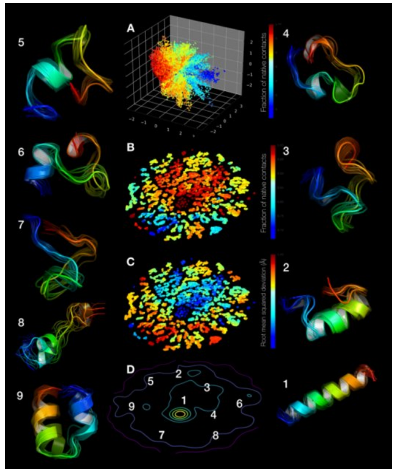
Imaging Sciences-Braggs Peak

(Credit: Z. Liu)

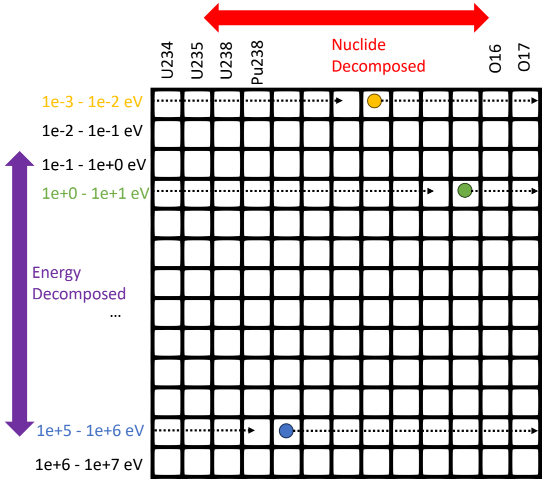


Tokomak Fusion Reactor operations

(Credit: K. Felker)



Protein-folding (Image: NCI)



Monte Carlo Particle Transport for Reactor Simulation (Credit: J. Tramm)

and more..

Argonne NATIONAL LABORATORY

A Traditional HPC Simulation Kernel on an AI Accelerator

Scientific Achievement

The Cerebras WSE-2 is a wafer-scale AI accelerator. Despite not being designed for traditional HPC workloads, we were able to develop new algorithms and performance optimization strategies to allow for a key Monte Carlo particle transport simulation kernel to execute with high efficiency on the device. Significant speed and power advantages compared to GPU were found.

Significance and Impact

- **Developed mini-app** representing key cross section lookup kernel form the Monte Carlo (MC) particle transport algorithm **for the Cerebras CSL SDK**
- Compared results against highly optimized CPU and GPU implementations, and found that the **WSE-2 was >100,000x faster than serial CPU execution**, and **182x faster than A100 GPU execution**
- Results suggest full MC particle transport app on WSE-2 will be possible

Technical Approach

- Leveraged vast quantities (>40GB) of single-cycle latency SRAM on WSE-2
- Developed new hyper domain decomposition techniques to spread simulation data across the >700k processing elements of the WSE-2, and optimized movement of particles through the WSE-2.

PI(s)/Facility Lead(s): John Tramm
Collaborating Institutions: ANL, UChicago
ASCR Program: ANL LDRD Expedition
ASCR PM: N/A
Publication(s) for this work: Tramm, et al., "Efficient algorithms for Monte Carlo particle transport on AI accelerator hardware," *Computer Physics Communications, Volume 298* (2024). doi:10.1016/j.cpc.2023.109072.

Architecture	Monte Carlo Performance FOM Lookups/sec
Cerebras WSE-2	1.17E+10
CPU (single 8180M Xeon Core)	1.15E+05
GPU (single NVIDIA A100 GPU)	6.43E+07

WSE-2 >100,000x speedup over serial CPU

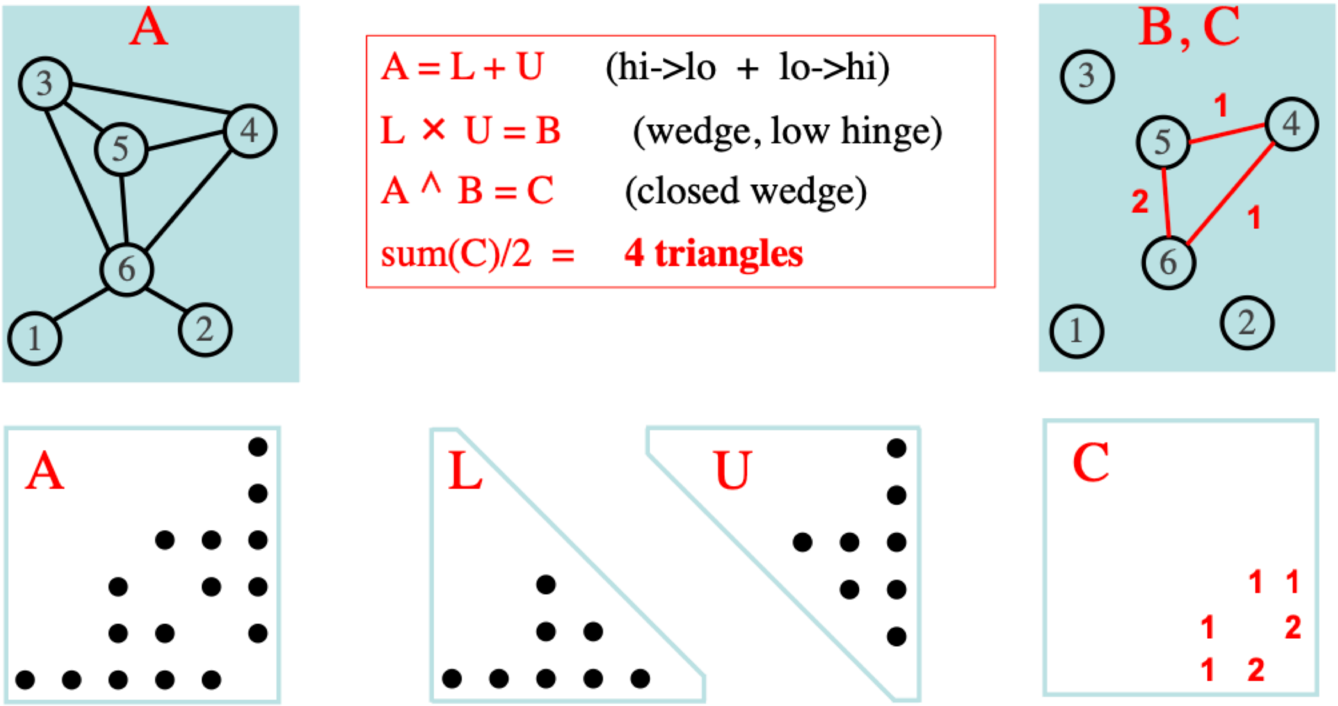
WSE-2 182x speedup over A100 GPU

This table shows the performance figure of merit for our mini-app (based on XSBench). The CPU version is written in C. The GPU version is written in optimized CUDA using architecture-specific optimization strategies. The Cerebras version was written using the CSL Cerebras SDK. The figure of merit represents the number of macroscopic cross sections per second (higher is better) in a typical depleted fuel reactor simulation problem with hundreds of nuclides.

Courtesy: John Tramm, ANL

Linear Algebra-based Triangle Counting on Graphcore's IPU Architecture

Given $G(V, E)$ where V is the vertex set and E is the edge set, count the number of triangles in G. A triangle is a triplet $< u, v, w >$ such that $u, v, w \in V$, and $uv, vw, uw \in E$.



Azad, B., Gilbert. "Parallel triangle counting and enumeration using matrix algebra". *IPDPSW, 2015*

Key Steps:

- LU-decompose Adjacency matrix
- Matrix multiply L , U
- Elementwise multiply
- Reduce

[Characterizing the Performance of Triangle Counting on Graphcore's IPU Architecture]

Reet Barik, Siddhisanket Raskar, Murali Emani, Venkatram Vishwanath

Triangle Counting on IPU: Mapping to architecture

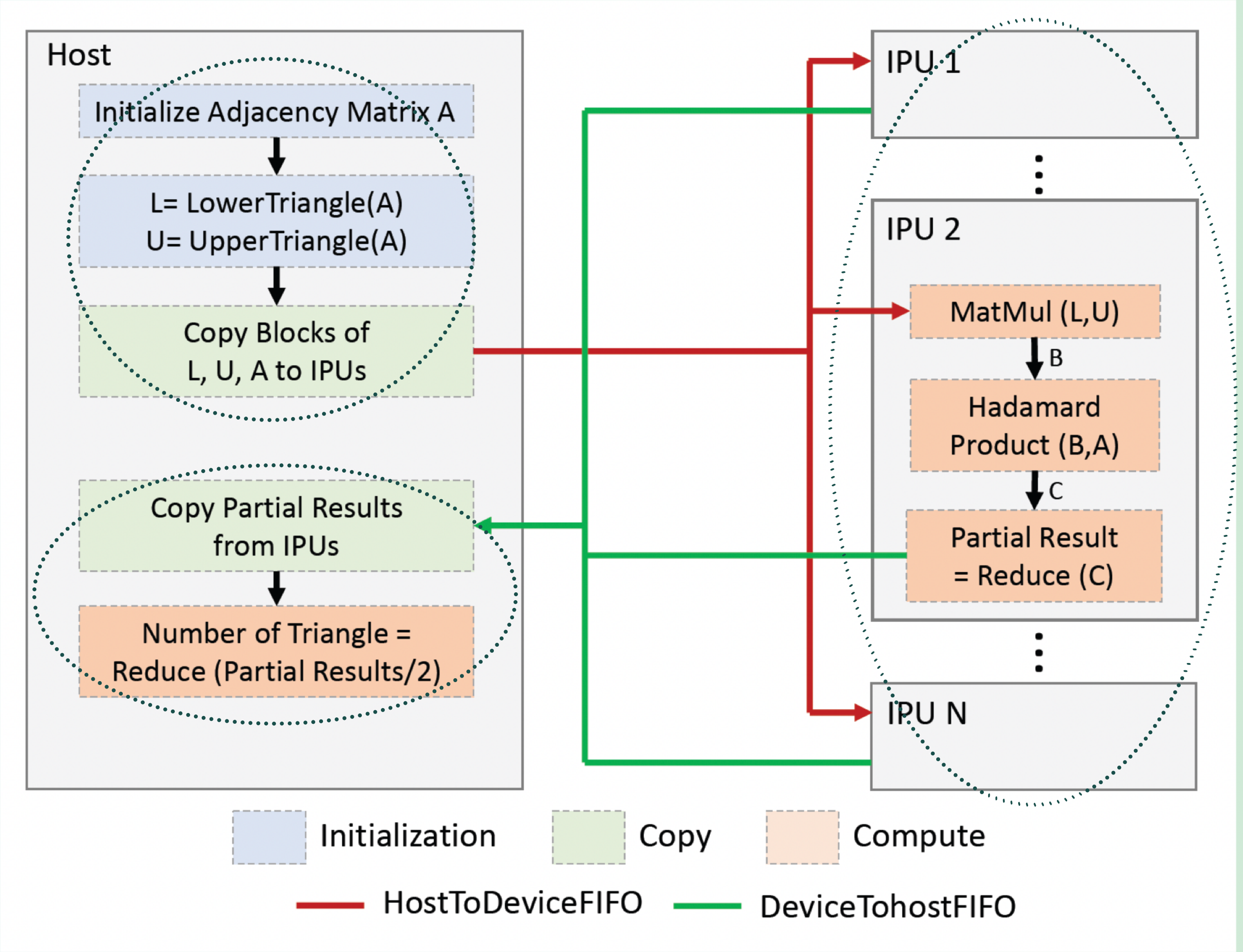
$$\begin{array}{cc} a & b \\ \multicolumn{2}{c}{\begin{bmatrix} \text{IPU 1} & \text{IPU 2} \\ \text{IPU 3} & \text{IPU 4} \end{bmatrix}} \\ c & d \end{array} * \begin{array}{cc} e & f \\ \multicolumn{2}{c}{\begin{bmatrix} \text{IPU 1} & \text{IPU 2} \\ \text{IPU 3} & \text{IPU 4} \end{bmatrix}} \\ g & h \end{array} = \begin{array}{cc} ae + bg & af + bh \\ \multicolumn{2}{c}{\begin{bmatrix} \text{IPU 1} & \text{IPU 2} \\ \text{IPU 3} & \text{IPU 4} \end{bmatrix}} \\ ce + dg & cf + dh \end{array}$$

Block decomposition of input matrix

Triangle Counting on IPU: Mapping to architecture

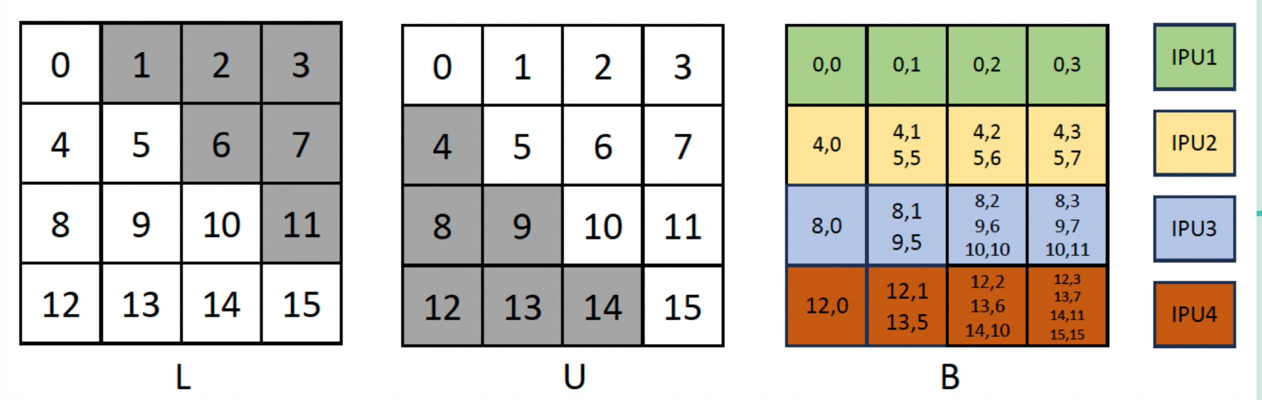
Init and copy From Host

Copy partial results to Host and reduce



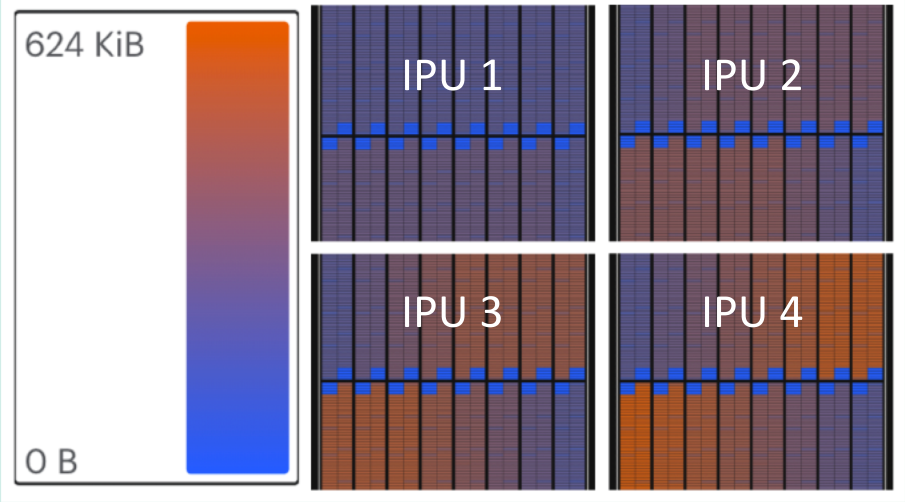
Compute On Device

Triangle Counting on IPU: Optimization

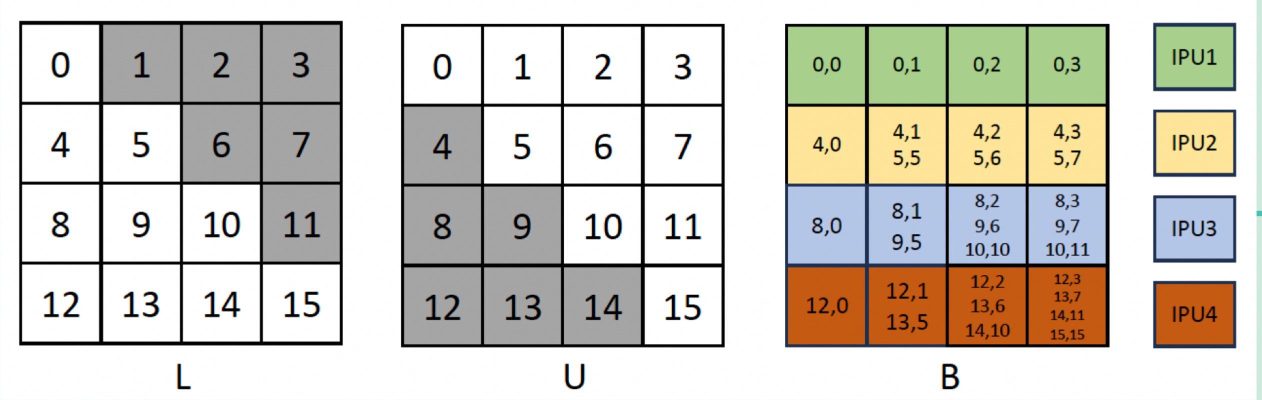


Computation workload pattern

Load imbalance

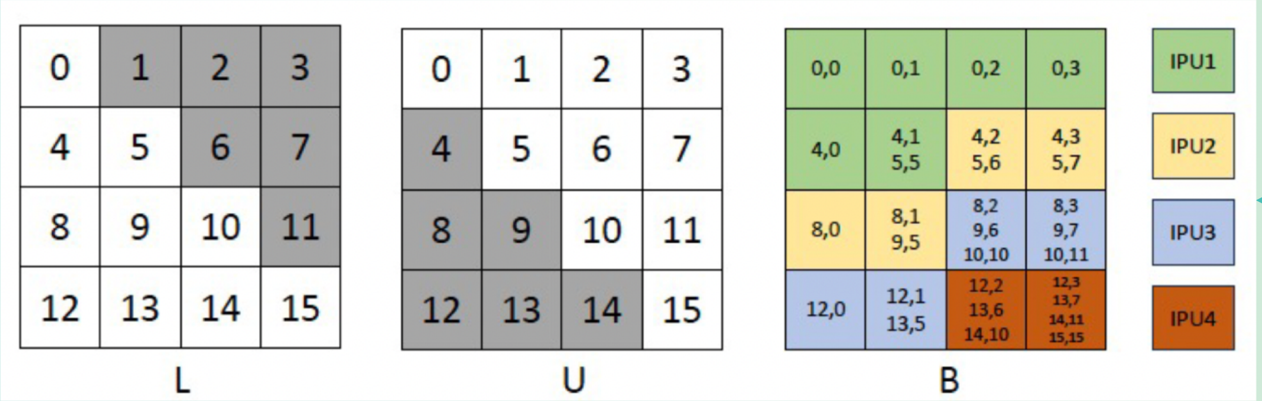


Triangle Counting on IPU: Optimization



Computation workload pattern

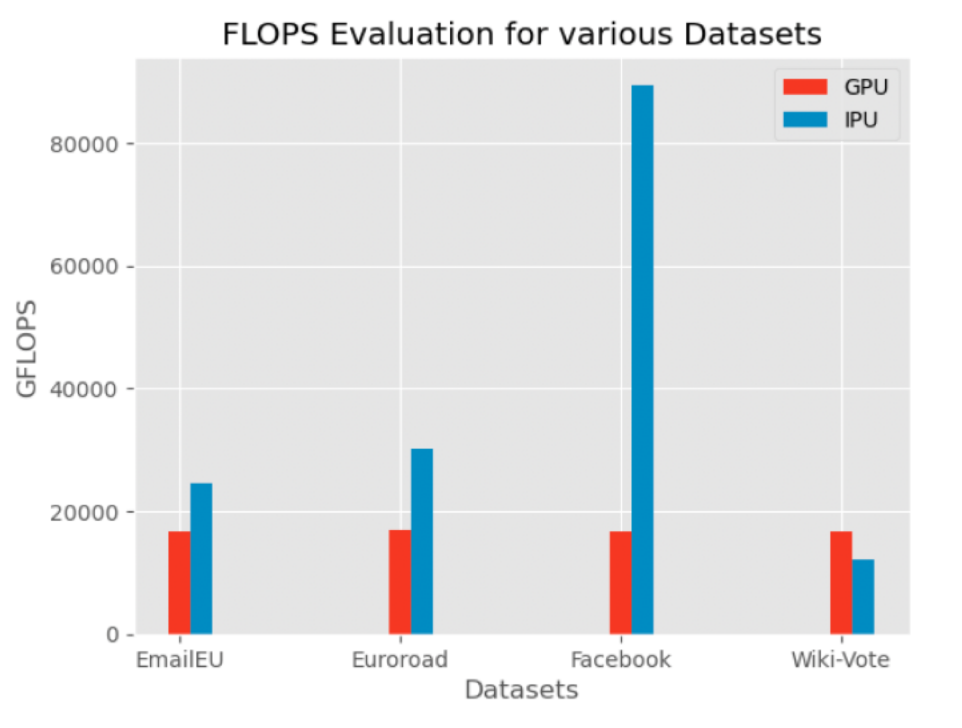
Weighted mapping of workload to IPUs



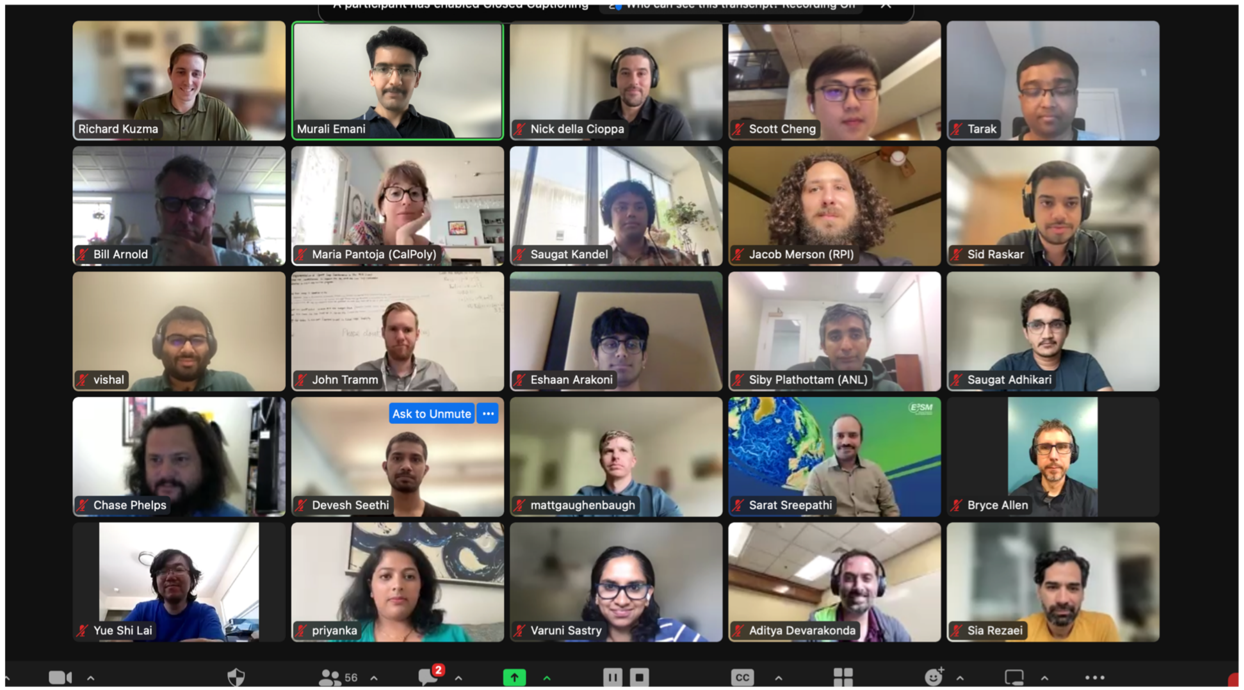
Triangle Counting on IPU: Experiments

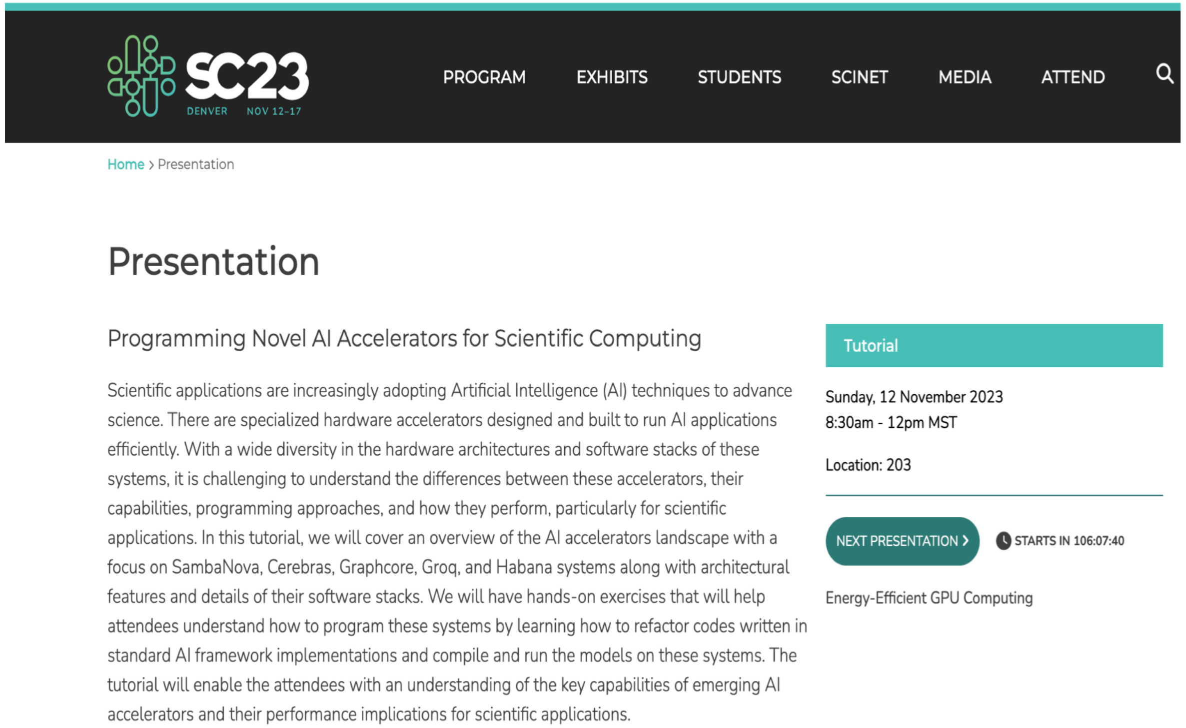
Input	|V|	|E|	Max Deg	Avg. Deg
Kronecker (2^8)	256	2155	163	16.8
Kronecker (2^9)	512	4752	274	18.6
Kronecker (2^10)	1024	10496	471	20.5
Kronecker (2^11)	2048	22709	747	22.2
Kronecker (2^12)	4096	48386	1316	23.6
Kronecker (2^13)	8192	102124	2250	24.9
EmailEU	1,005	25,571	345	31.9
Euroroad	1,174	1,417	10	2.4
Facebook	4,039	88,234	1045	43.7
Wiki-Vote	7,115	103,689	1065	28.3

Table 1: Dataset Characteristics



AI Testbed Community Engagement





- AI training workshops
  Cerebras: https://events.cels.anl.gov/event/420/
  SambaNova: https://events.cels.anl.gov/event/421/
  Graphcore: https://events.cels.anl.gov/event/422/
  Groq: https://events.cels.anl.gov/event/448/

Tutorial at SC23 on Programming Novel AI accelerators for Scientific Computing ***in collaboration with Cerebras, Intel Habana, Graphcore, Groq and SambaNova***

Argonne NATIONAL LABORATORY

Observations, Challenges and Insights

- Significant speedup achieved for a wide-gamut of scientific ML applications
- Early adoption for HPC kernels show promising results
- Recent work on using OpenMP to offload kernels to Graphcore IPUs
- Room for improvement exists
  - Porting efforts and compilation times, custom libraries
  - support for performance analysis tools, debuggers
- Limited capability to support low-level HPC kernels
  - Work in progress to improve coverage

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

- **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 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

- **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