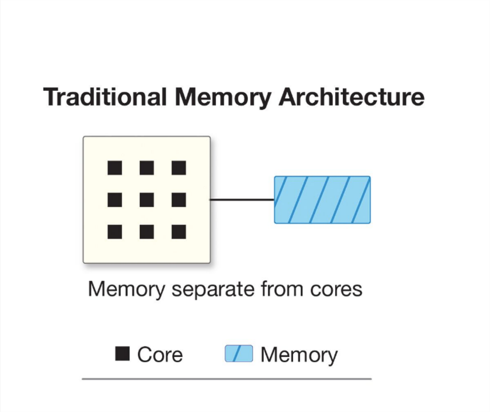
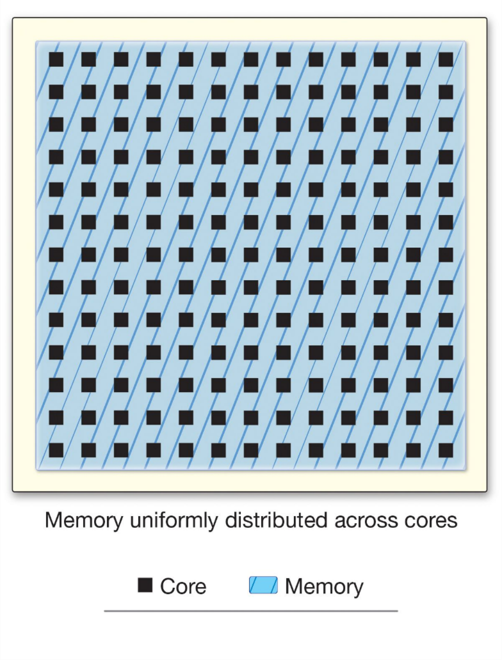
Overview of AI Testbeds at ANL

ALCF Simulation, Data, and Learning Workshop
7th October 2021

By
Sid Raskar

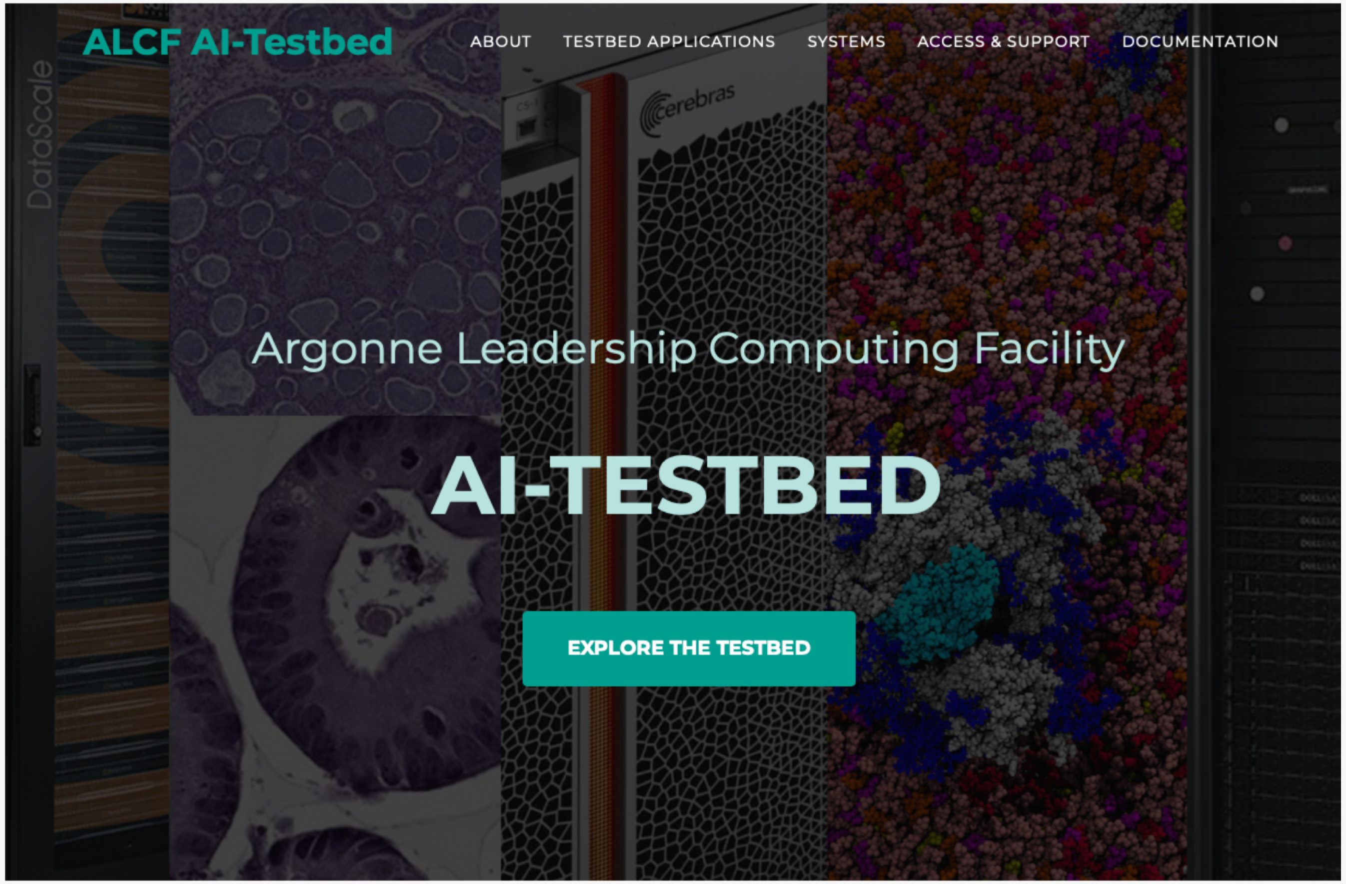
AI Accelerators





- Limitations of Traditional Architectures
- Heavy data movement leads to Increased Energy Cost in GPUs
- Rise of domain-specific dataflow inspired architectures
- Workflow
  - Program is represented as a graph
  - This program graph is mapped on the the architecture

AI Testbeds at Argonne



https://ai.alcf.anl.gov/

- Infrastructure of next-generation machines with hardware accelerators customized for artificial intelligence (AI) applications.
- Provide a platform to evaluate usability and performance of machine learning based HPC 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

AI Testbeds at Argonne

cerebras

SambaNova SYSTEMS

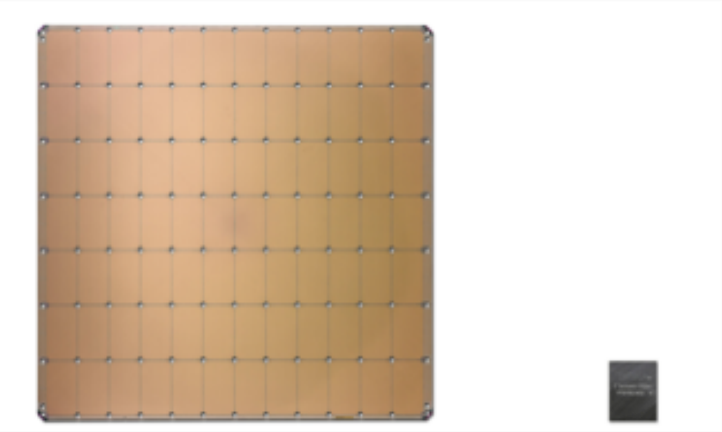
GRAPHCORE

groq

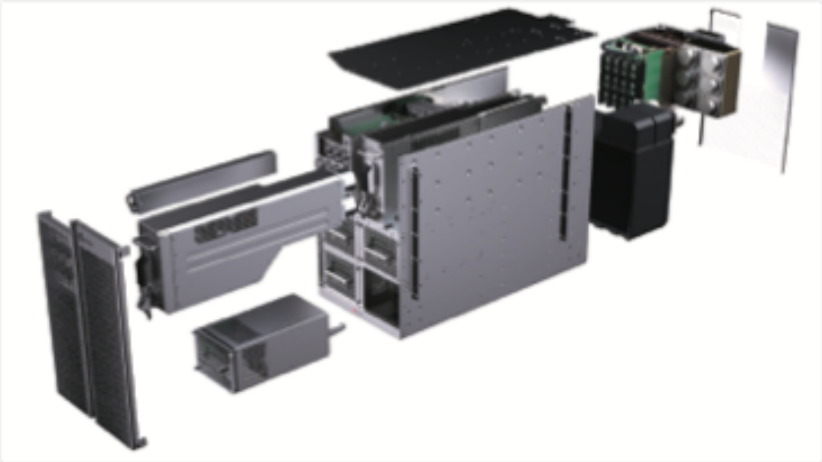
- Architecture
- Applications
- Publications

AI Testbeds at Argonne

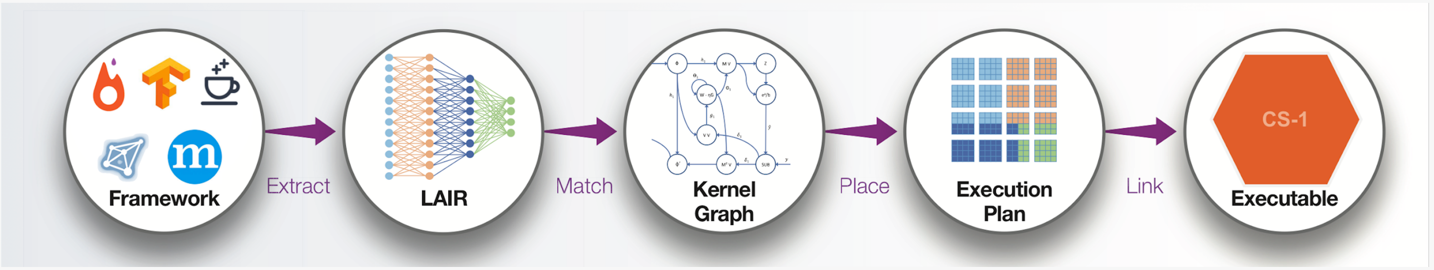
- Architecture
- Applications
- Publications



Wafer Scale Architecture
Cerebras Chip vs GPU



Side view of Cerebras System

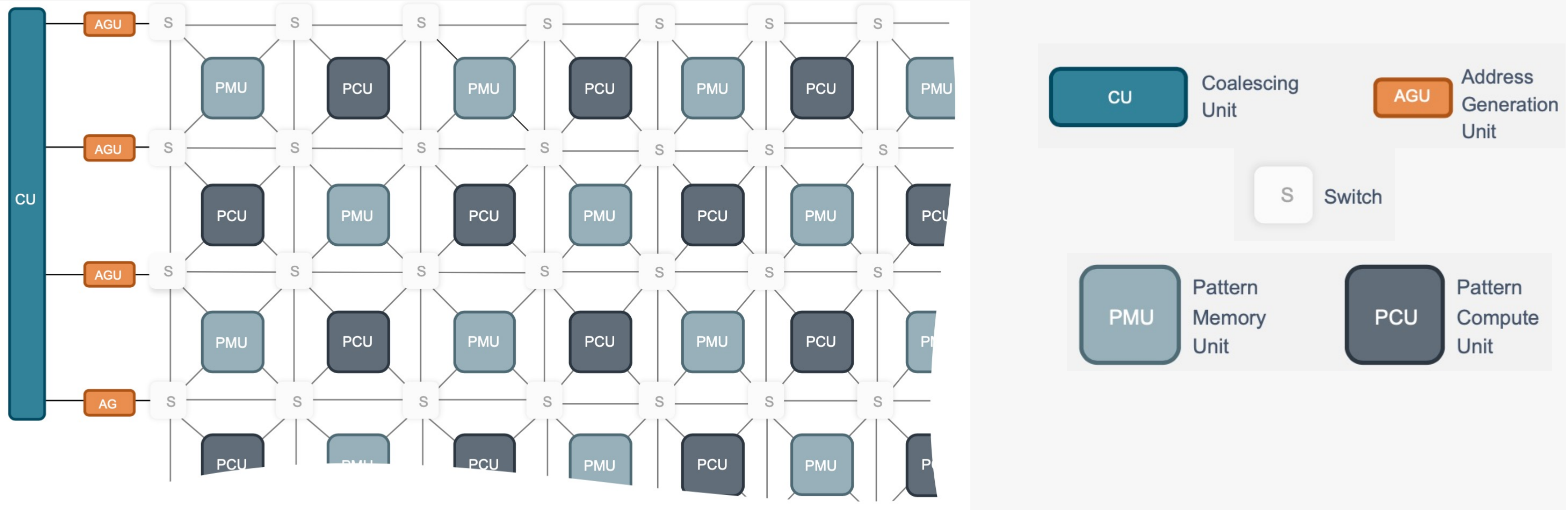


Cerebras LAIR :**L**inear **A**lgebra **I**ntermediate **R**epresentation

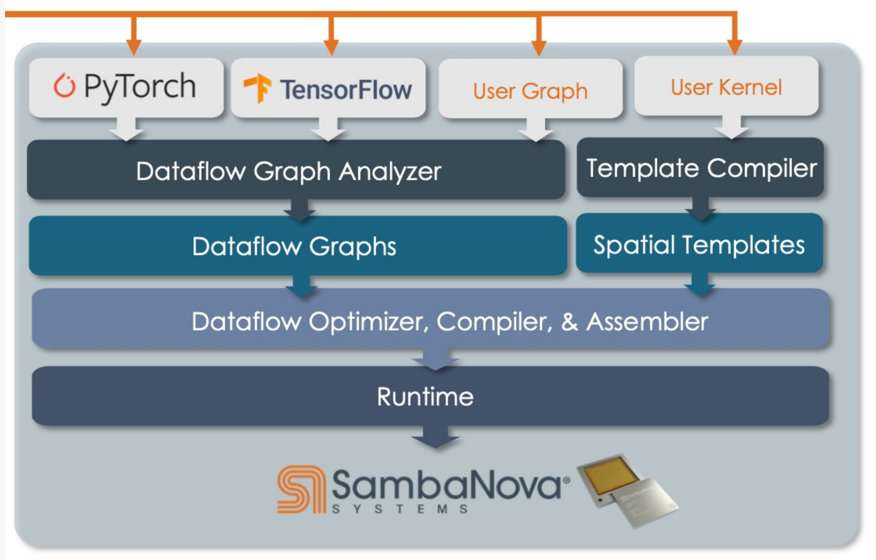
cerebras

- 400,000 Cores
- 18GB of total On-Chip Memory
- 100PBits/s fabric bandwidth
- 9PByte/s Memory bandwidth
- 1.2T transistors, 16nm
- >300 TFLOPS (BF16) of claimed performance
- Supports Tensorflow & PyTorch
- Supports both training and inference

Source : Cerebras Whitepaper



Simplified Reconfigurable Dataflow Unit (RDU) architecture

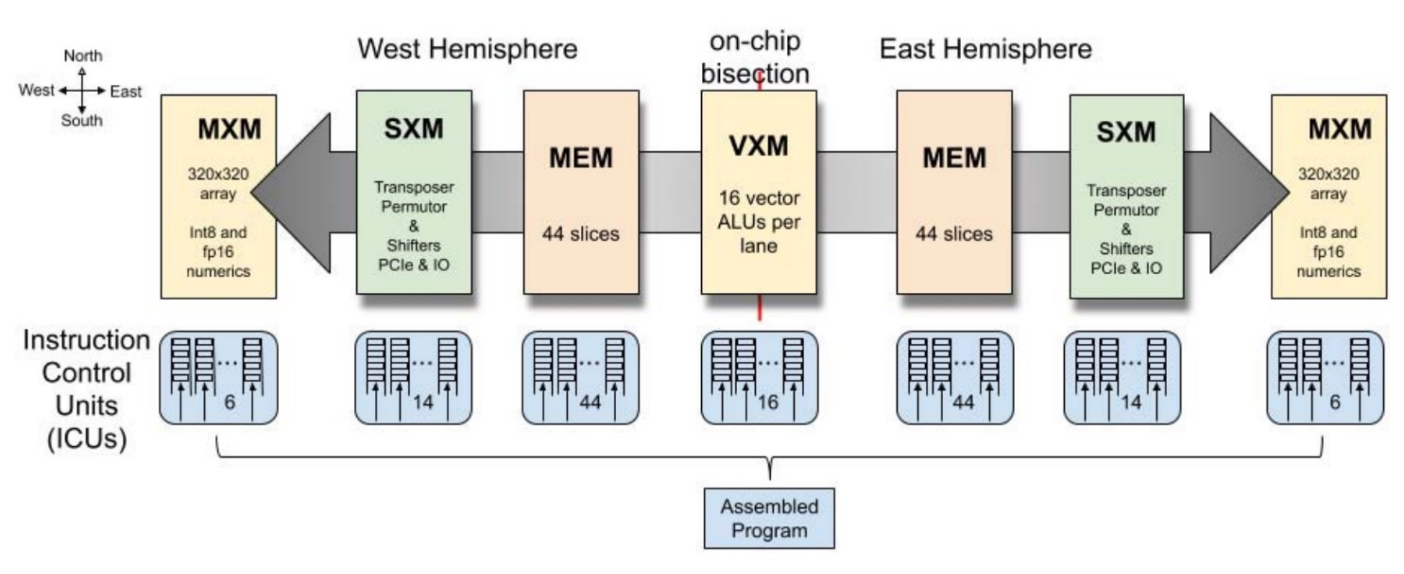


SambaFlow Software Stack

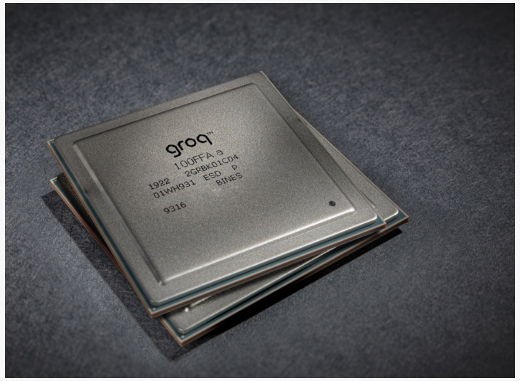
SambaNova SYSTEMS

- 100s MB of on chip memory
- 40B transistors, 7nm TSMC
- 100s TFLOPS of claimed performance
- Support for Sambaflow, PyTorch, Tensorflow
- Support for training and interence

Source : Sambanova Whitepaper



The organization and dataflow within a row in the on-chip network.



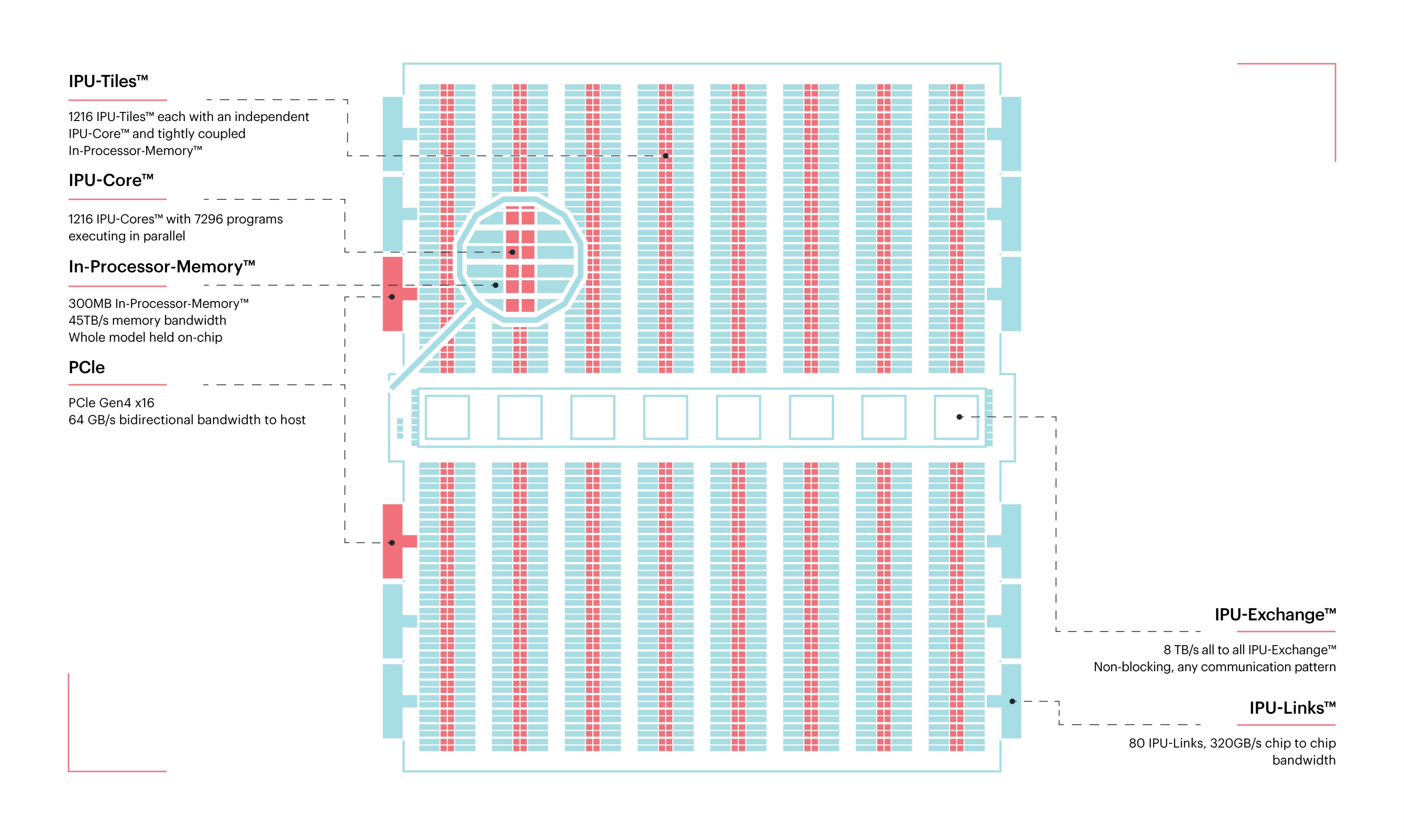
Tensor Streaming Processor

- 220MB of on-chip memory
- 14nm process, 26.8B transistors
- Claimed performance of 250TFlops FP16
  1 PetaOps in int8
- 80TB/s on-die memory bandwidth
- 300W of Power consumption
- Support for Tensorflow, PyTorch, ONNX
- Support for Inference only

Source : Groq Whitepaper

Colossus MK2 GC200 IPU (Intelligent Processing Unit)

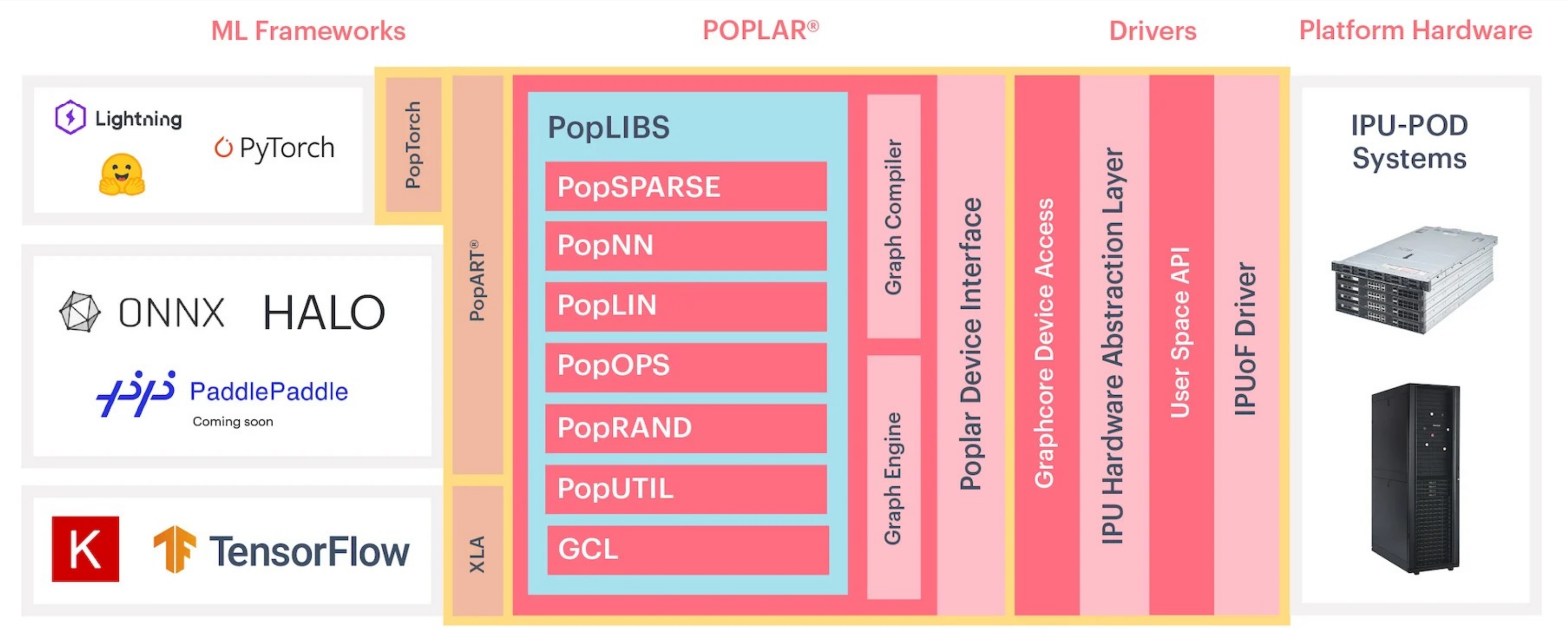
GRAPHCORE



- 900MB of on-chip memory
- 47.5TB/s memory bandwidth
- 7nm process, 59.4Bn transistors
- 250TFLOPS (FP16) of claimed performance
- Support for Tensorflow, PyTorch and PopArt
- Support for Both training and inference

Source : Citadel Whitepaper

GRAPHCORE



Poplar Software Stack

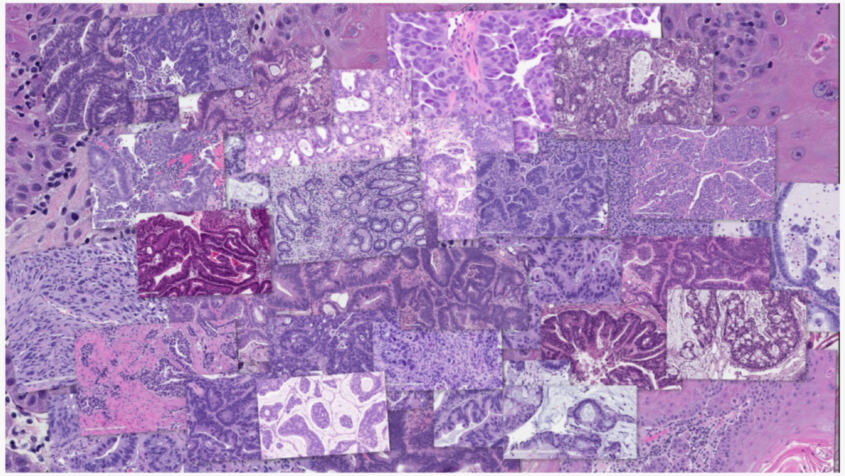
- ➢ 900MB of on-chip memory
- ➢ 47.5TB/s memory bandwidth
- ➢ 7nm process, 59.4Bn transistors
- ➢ 250TFLOPS (FP16) of claimed performance
- ➢ Support for Tensorflow, PyTorch and PopArt
- ➢ Support for Both training and inference

Source : [Poplar Whitepaper]

AI Testbeds at Argonne

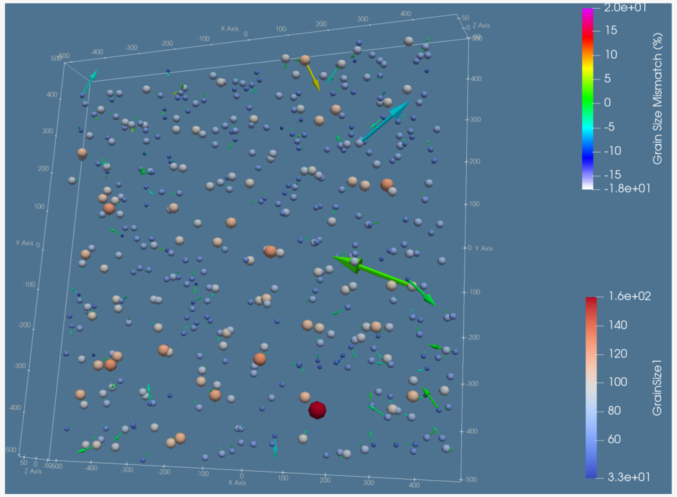
- Architecture
- **Applications**
- Publications

Argonne Science Applications on Cerebras CS-1

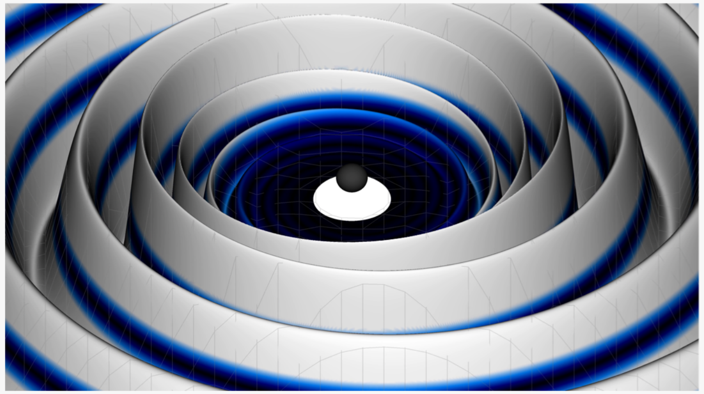


Cancer Drug response prediction

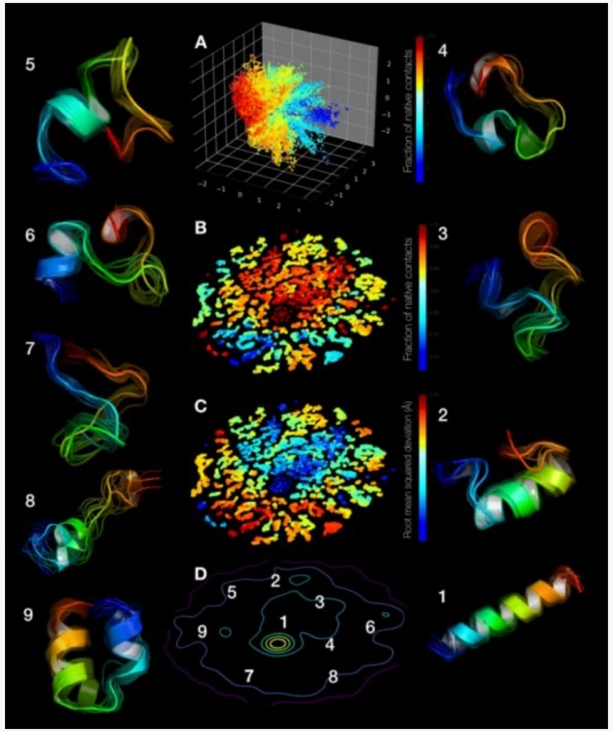
Predicting cancer type and drug response using histopathology images from the National Cancer Institute's Patient-Derived Models Repository.



Fast X-Ray Braggs Peak Analysis



Gravitational waves (Image: NCSA)



Protein-folding (Image: NCI)

Argonne Science Applications on on SambaNova

- Cosmic Tagger : Cosmic Background Removal with Deep Neural Networks in SBND
- SambaWF : Highly Resolved Surrogate Models for Weather Forecasting
- Accelerating Graph Convolution based Deep Learning Framework for Large Scale Highway Traffic Forecasting with Sambanova
- Deep Learning Hamiltonian Monte Carlo
- BraggNN : Fast X-ray Bragg Peak Analysis
- Deep Learning based Scalable and Robust Strong Gravitational Lensing Characterization Pipeline
- Acelerating AI/ML for fusion Sciences
- Deep Learning Atomic Potentials

AI Testbeds at Argonne

- Architecture
- Applications
- **Publications**

Publications

- M. Emani et al., "Accelerating Scientific Applications With SambaNova Reconfigurable Dataflow Architecture," in Computing in Science & Engineering, vol. 23, no. 2, pp. 114-119, 1 March-April 2021, doi: 10.1109/MCSE.2021.3057203.

- Alexander Brace, Michael Salim, Vishal Subbiah, Heng Ma, Murali Emani, Anda Trifa, Austin R. Clyde, Corey Adams, Thomas Uram, Hyunseung Yoo, Andew Hock, Jessica Liu, Venkatram Vishwanath, and Arvind Ramanathan. 2021. Stream-AI-MD: streaming AI-driven adaptive molecular simulations for heterogeneous computing platforms. Proceedings of the Platform for Advanced Scientific Computing Conference. Association for Computing Machinery, New York, NY, USA, Article 6, 1–13. DOI:https://doi.org/10.1145/3468267.3470578