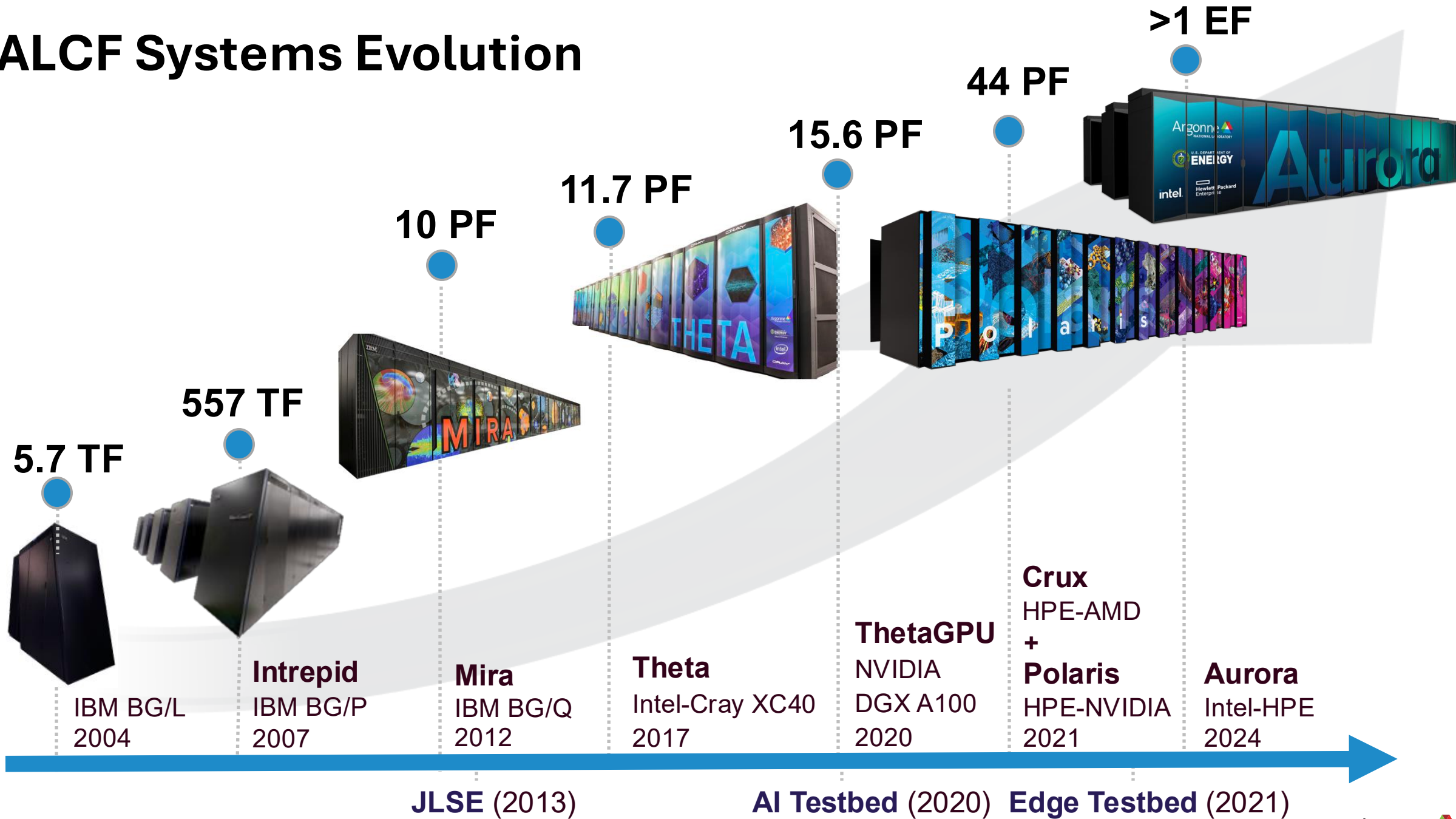


# ALCF AI Accelerators

Murali Emani  
Argonne National Laboratory  
[memani@anl.gov](mailto:memani@anl.gov)

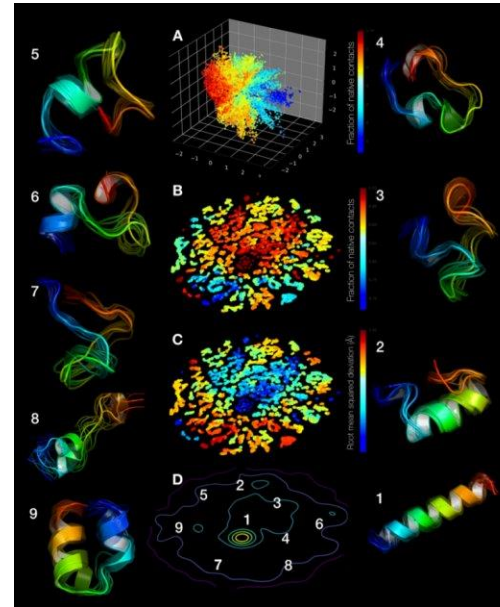
Contributors: Venkat Vishwanath, Varuni Sastry, Bill Arnold, Sid Raskar, Krishna Teja Chitty-Venkata

# ALCF Systems Evolution

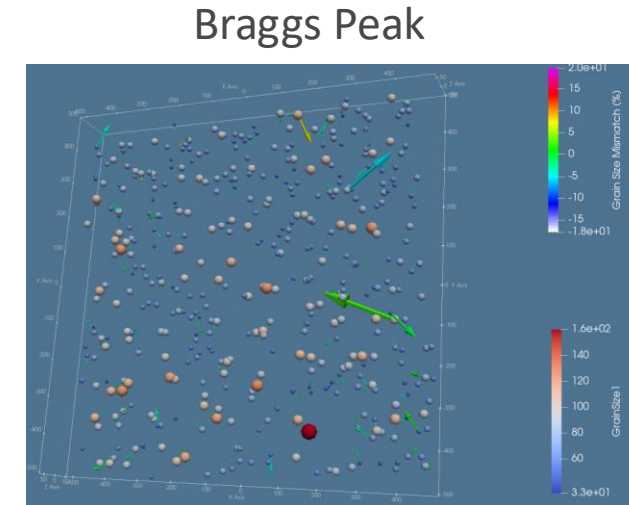


# Surge of Scientific Machine Learning

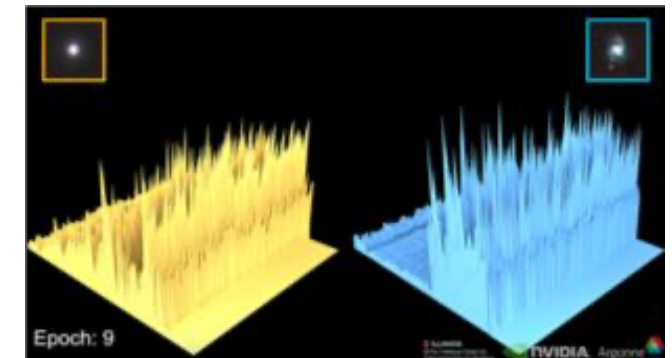
- Simulations/ surrogate models  
Replace, in part, or guide simulations with AI-driven surrogate models
- Data-driven models  
Use data to build models without simulations
- Co-design of experiments  
AI-driven experiments



Protein-folding

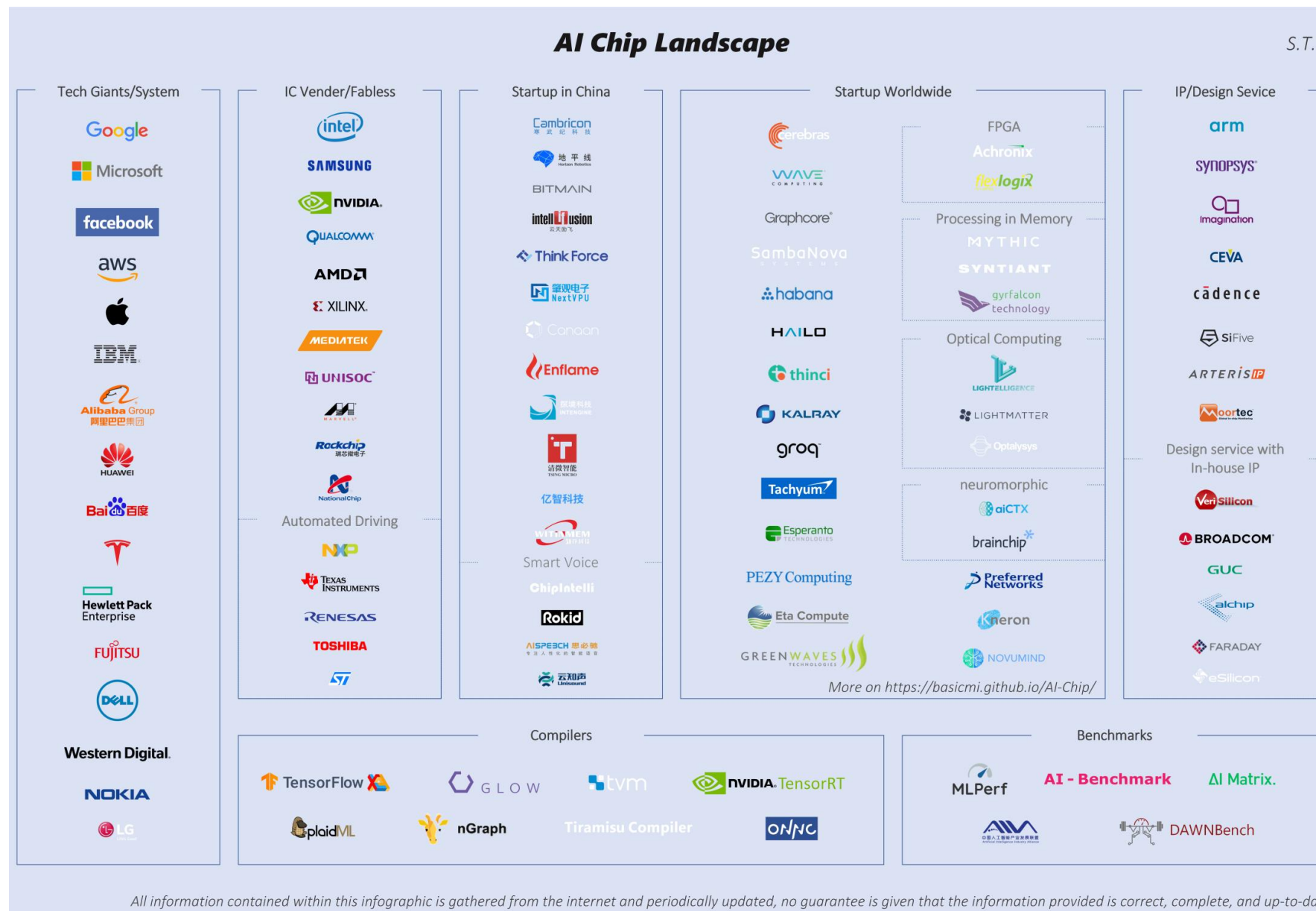


Braggs Peak



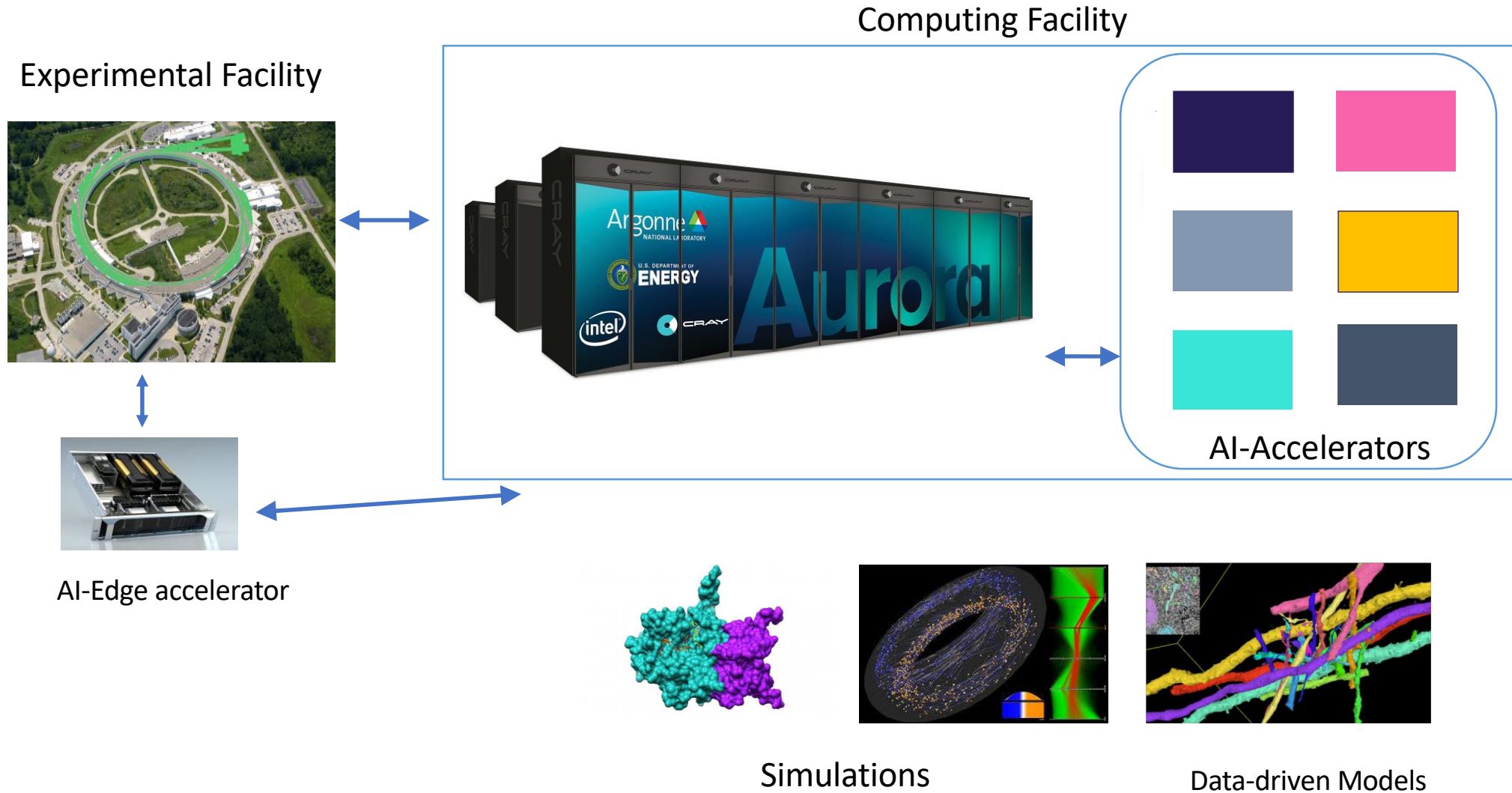
Galaxy Classification

**Design infrastructure to facilitate and accelerate AI for Science (AI4S) applications**





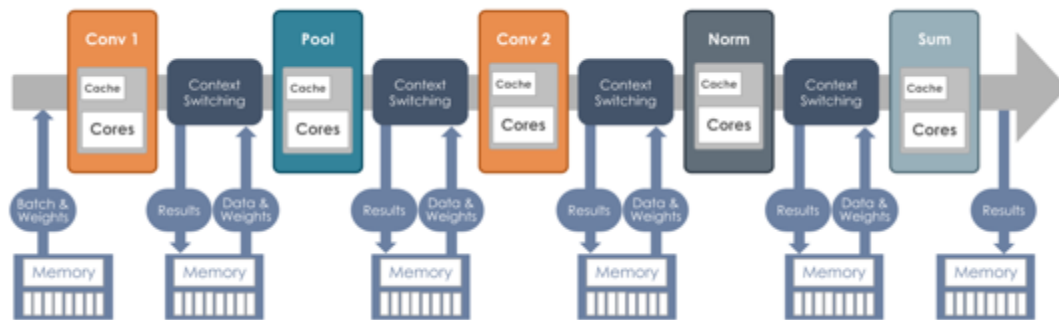
# Integrating AI Systems in Facilities



# Dataflow Architectures



Simple  
Convolution  
Graph



The GPU way: kernel-by-kernel  
Bottlenecked by memory bandwidth  
and host overhead



The Dataflow way: Spatial  
Eliminates memory traffic and overhead

Image Courtesy: SambaNova



# Overview of ALCF AI Testbed



U.S. DEPARTMENT OF  
**ENERGY**

Argonne National Laboratory is a  
U.S. Department of Energy laboratory  
managed by UChicago Argonne, LLC.



**Argonne Leadership  
Computing Facility**



# ALCF AI Testbed

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

- 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

# ALCF AI Testbed

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

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



8 nodes each with 8  
Reconfigurable  
DataFlow Units (RDU)



2 CS-2 Wafer scale  
engines (WSE)

**Upgrading to CS-3**

Cerebras CS-2

SambaNova SN-30



4 nodes each  
with 16 Intelligent  
Processing Units  
(IPUs)



9 nodes each with  
8 GroqChip  
Tensor streaming  
processors (TSP)

Groq

Graphcore Bow Pod64



**Coming Soon !  
Sambanova SN40L  
Inference/Finetuning**

**NSF <https://nairrpilot.org>**



# Argonne Leadership Computing Facility

[ALCF Resources](#) [Science](#) [Community and Partnerships](#) [About](#) [Support Center](#)<https://docs.alcf.anl.gov/ai-testbed>

## ALCF AI Testbed

### 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](#) >

### Table of contents

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

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.

# Getting Started on ALCF AI Testbed

## Available for Allocations

- Cerebras CS-2,
- SambaNova Datascale SN30,
- GroqRack
- Graphcore Bow Pod64

## AI Testbed User Guide

## Director's Discretionary (DD) awards

- Scaling code
- Preparing for future computing competition
- Scientific computing in support of strategic partnerships.

### Allocation Request Form

<https://www.alcf.anl.gov/science/directors-discretionary-allocation-program>

## NAIRR Pilot

aims to connect U.S. researchers and educators to computational, data, and training resources needed to advance AI research and research that employs AI.

<https://nairrpilot.org/>

# Hands-on session

Friday — August 1, 2025

## Track 3 — Machine Learning

8:30 a.m.	Welcome and Introduction	Filippo Simini, ANL
8:40 a.m.	Transition time: splitting into groups (people new to deep learning vs. more experienced)	
	Parallel Session, Part 1 (talk/hands-on):	
	Main room: Introduction to Deep Learning	Bethany Lusch, ANL
	Breakout room: Profiling Deep Learning	Khalid Hossain, ANL
9:40 a.m.	Introduction to Large Language Models (LLMs)	Huihuo Zheng, ANL
10:40 a.m.	Break	
11:10 a.m.	Distributed Deep Learning (talk/hands-on)	Nathan Nichols, ANL Kaushik Velusamy, ANL
12:30 p.m.	Lunch	
1:30 p.m.	Research talk: TBD	Sandeep Madireddy, ANL
2:00 p.m.	AI Testbed (talk/hands-on)	Sid Raskar, PNNL





	<b>Cerebras CS2</b>	<b>SambaNova Cardinal SN30</b>	<b>Groq GroqRack</b>	<b>GraphCore GC200 IPU</b>	<b>NVIDIA A100</b>
<b>Compute Units</b>	850,000 Cores	640 PCUs	5120 vector ALUs	1472 IPUUs	6912 Cuda Cores
<b>On-Chip Memory</b>	40 GB L1, + MemoryX	>300MB L1 1TB	230MB L1	900MB L1	192KB L1 40MB L2 40-80GB
<b>Process</b>	7nm	7nm	7 nm	7nm	7nm
<b>System Size</b>	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)	Several systems
<b>Estimated Performance of a card (TFlops)</b>	>5780 (FP16)	>660 (BF16)	>250 (FP16) >1000 (INT8)	>250 (FP16)	312 (FP16), 156 (FP32)
<b>Software Stack Support</b>	Pytorch	SambaFlow, Pytorch	GroqAPI, ONNX	Tensorflow, Pytorch, PopArt	Tensorflow, Pytorch, etc
<b>Interconnect</b>	Ethernet-based	Ethernet-based	RealScale™	IPU Link	NVLink

# Cerebras Wafer Scale Engine (WSE 2)

**850,000** cores optimized for sparse linear algebra

**46,225 mm<sup>2</sup>** silicon

**2.6 trillion** transistors

**40 gigabytes** of on-chip memory

**20 PByte/s** memory bandwidth

**220 Pbit/s** fabric bandwidth

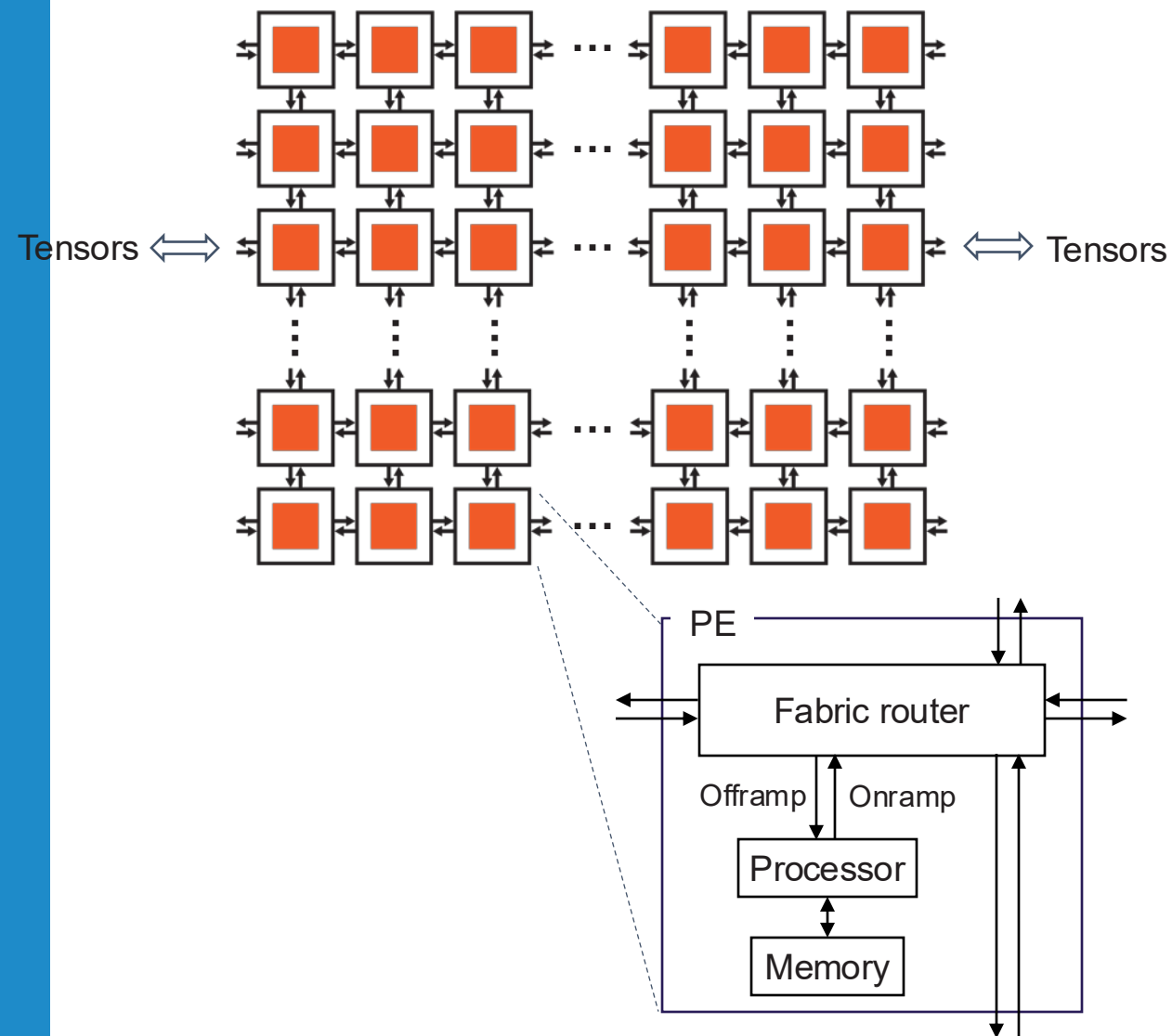
**7nm** process technology



850K cores, 40G on-chip SRAM



# WSE-2 Architecture Basics



The WSE appears as a logical 2D array of individually programmable Processing Elements

## Flexible compute

- 850,000 general purpose CPUs
- 16- and 32-bit native FP and integer data types
- **Dataflow programming**: Tasks are activated or triggered by the arrival of data packets

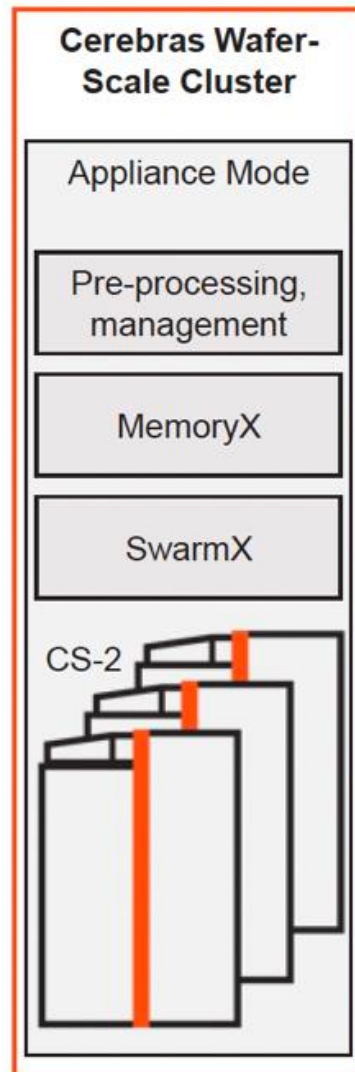
## Flexible communication

- Programmable router
- Static or dynamic routes (**colors**)
- Data packets (**wavelets**) passed between PEs
- 1 cycle for PE-to-PE communication

## Fast memory

- 40GB on-chip SRAM
- Data and instructions
- 1 cycle read/write

# Cerebras Wafer-Scale Cluster



Input preprocessing servers stream training data

MemoryX - Stores and streams model's weights

SwarmX – weight broadcasts and gradient across multiple wafers

Compilation (maps graph to kernels) Execution (training)

Image Courtesy: Cerebras

# Cerebras CS2 Cluster

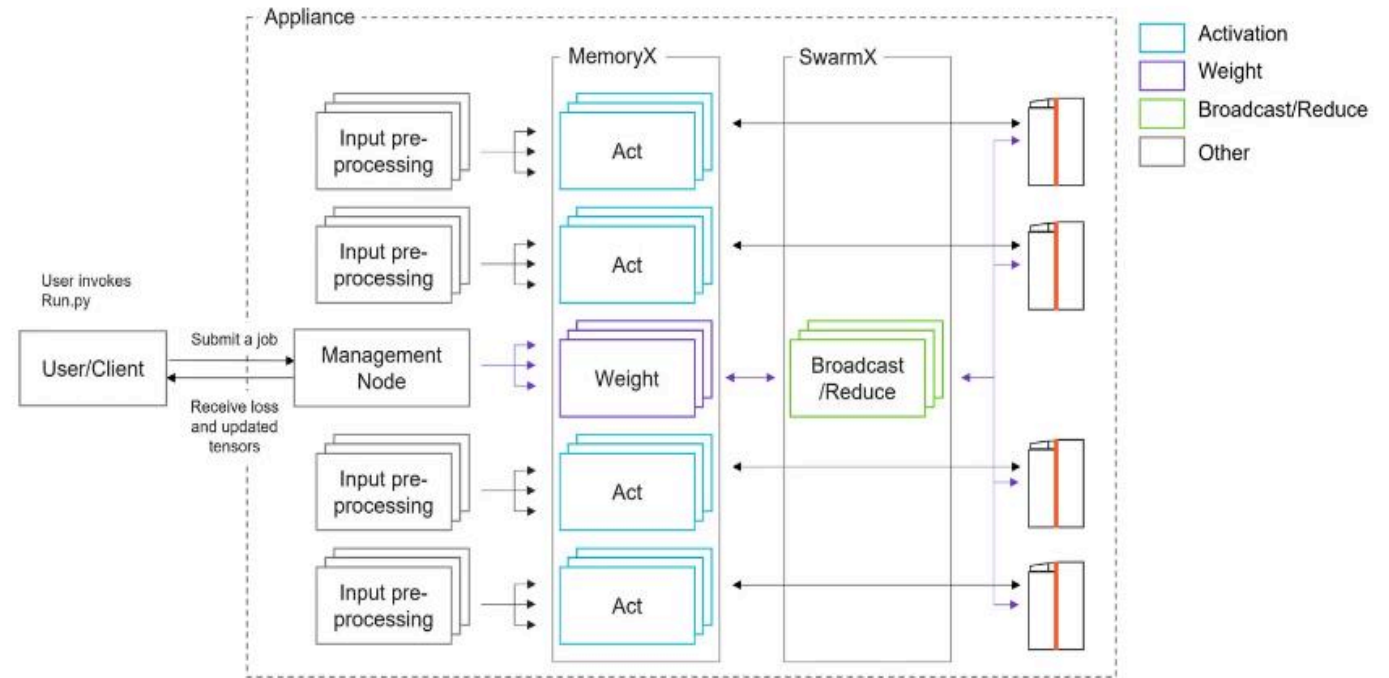
Input preprocessing servers stream training data (16 worker nodes)

MemoryX - Stores and streams model's weights

SwarmX – weight broadcasts and gradient across multiple CS2s

Compilation (maps graph to kernels)  
Execution (training)

Weight Streaming (training) Vs  
Pipeline (Inference)



# Cerebras Software stack

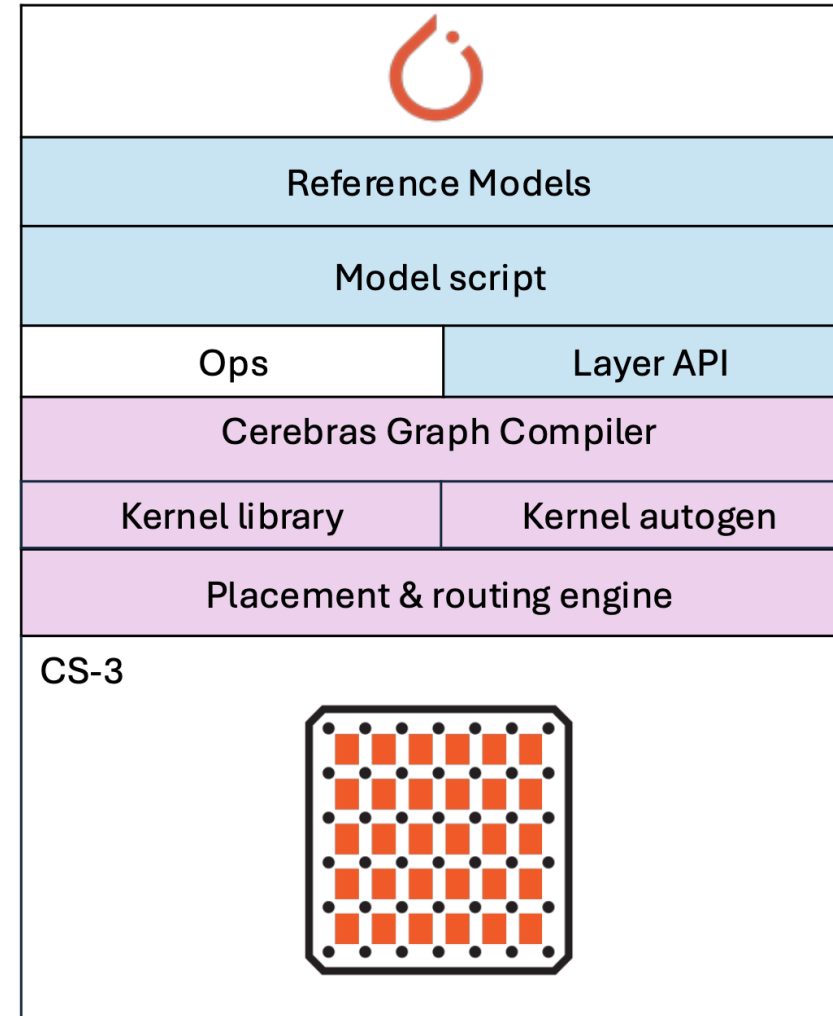
Maps PyTorch code to high performance kernels

Polyhedral code generation or hand written kernels

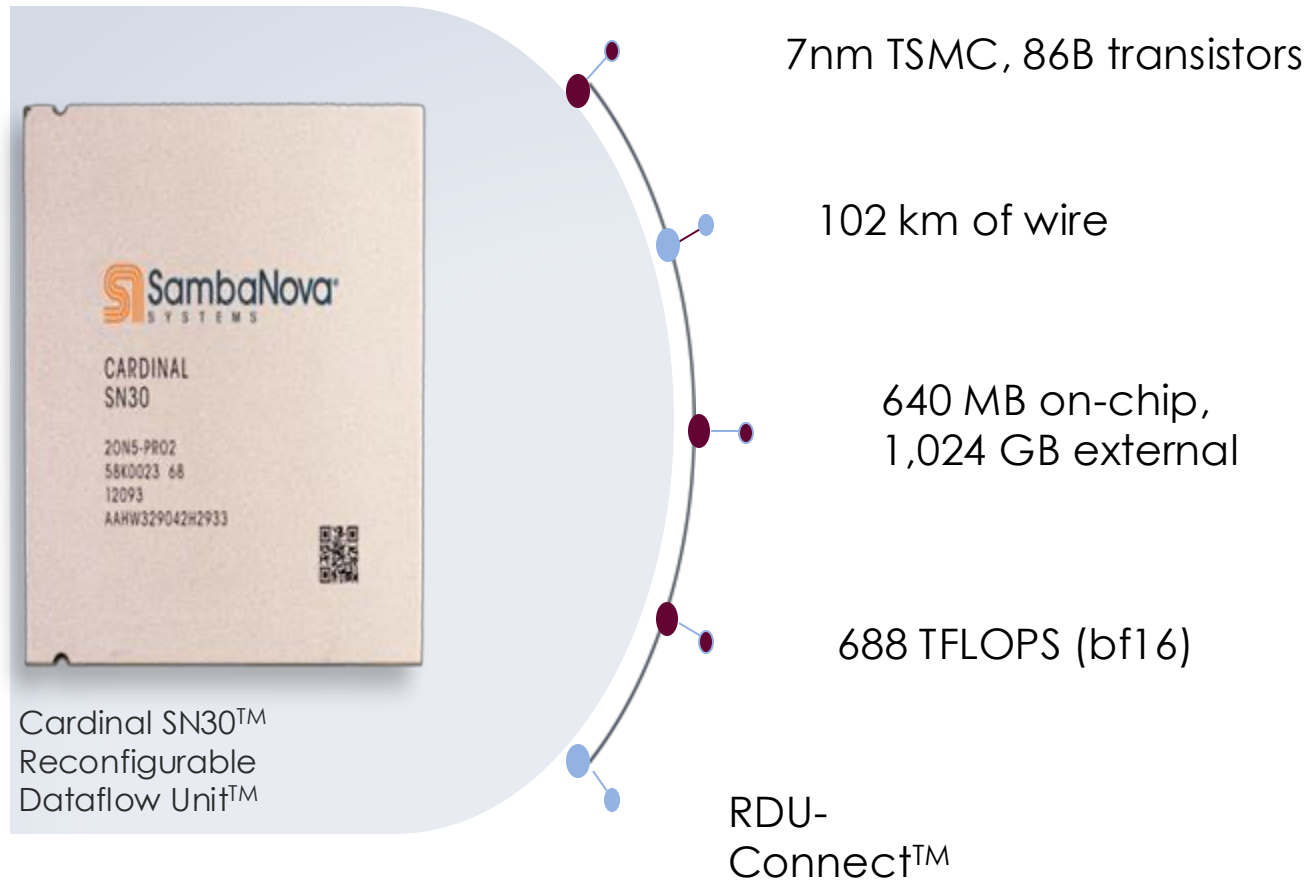
MLIR based compiler

User does not worry about distributed compute or parallelism

Achieves Linear scaling



# SambaNova Cardinal SN30 RDU



## as-a-SERVICE

Pre-trained  
Foundation Models

## SYSTEMS

DataScale®

## SOFTWARE

SambaFlow™

## SILICON

RDU

Image Courtesy: SambaNova



# SN40L “Cerulean” Architecture-based RDU

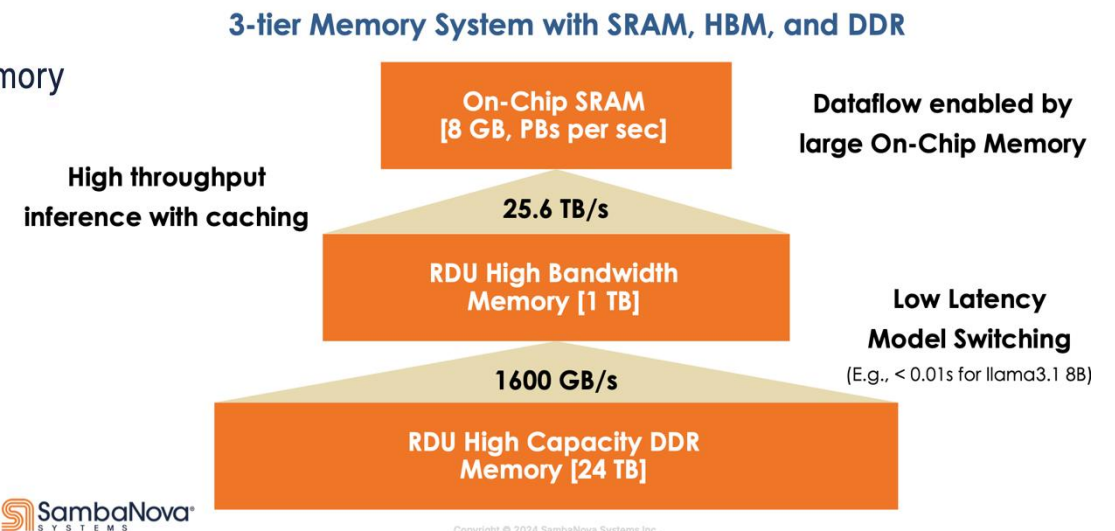
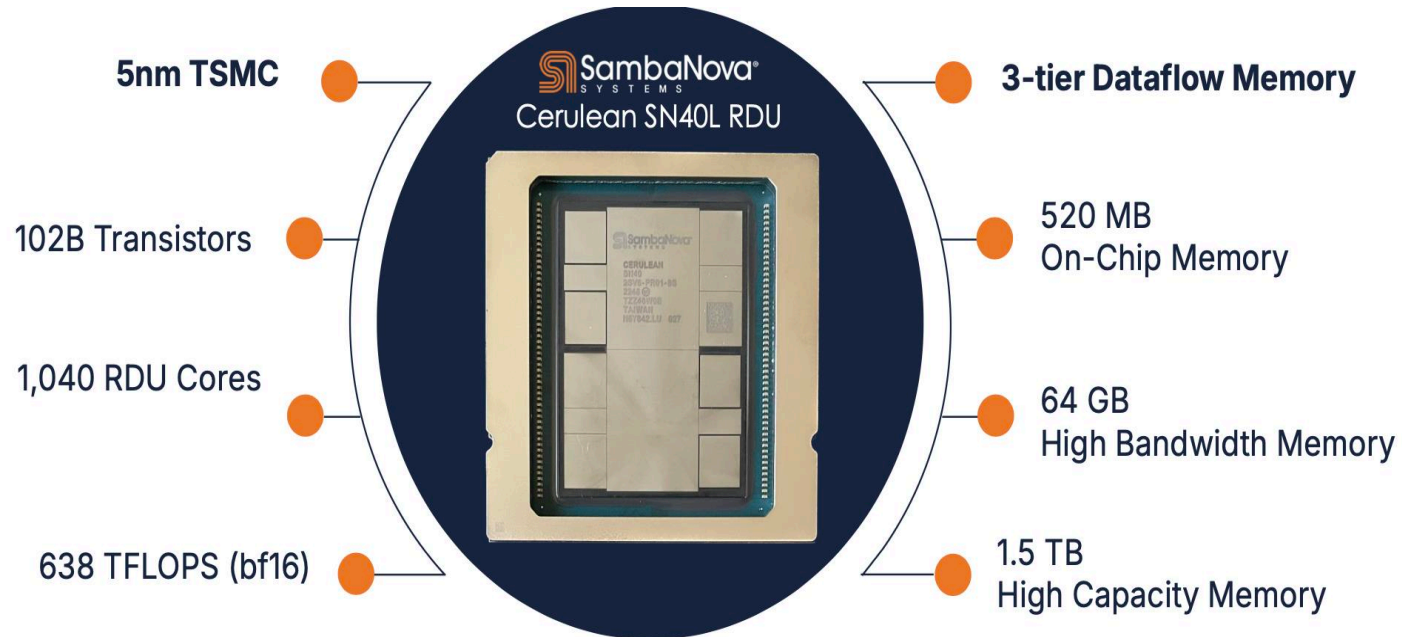
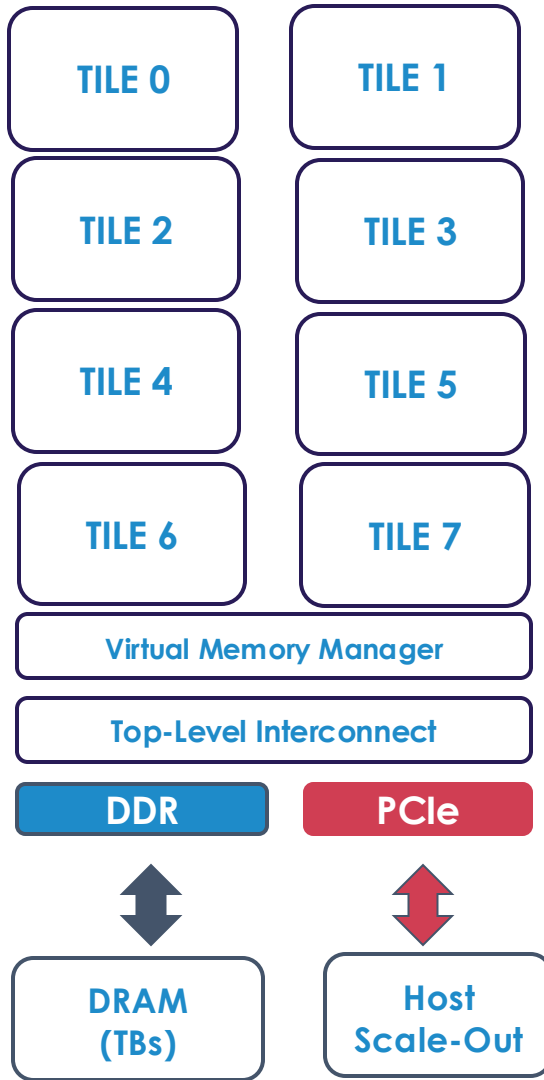


Image Courtesy: SambaNova

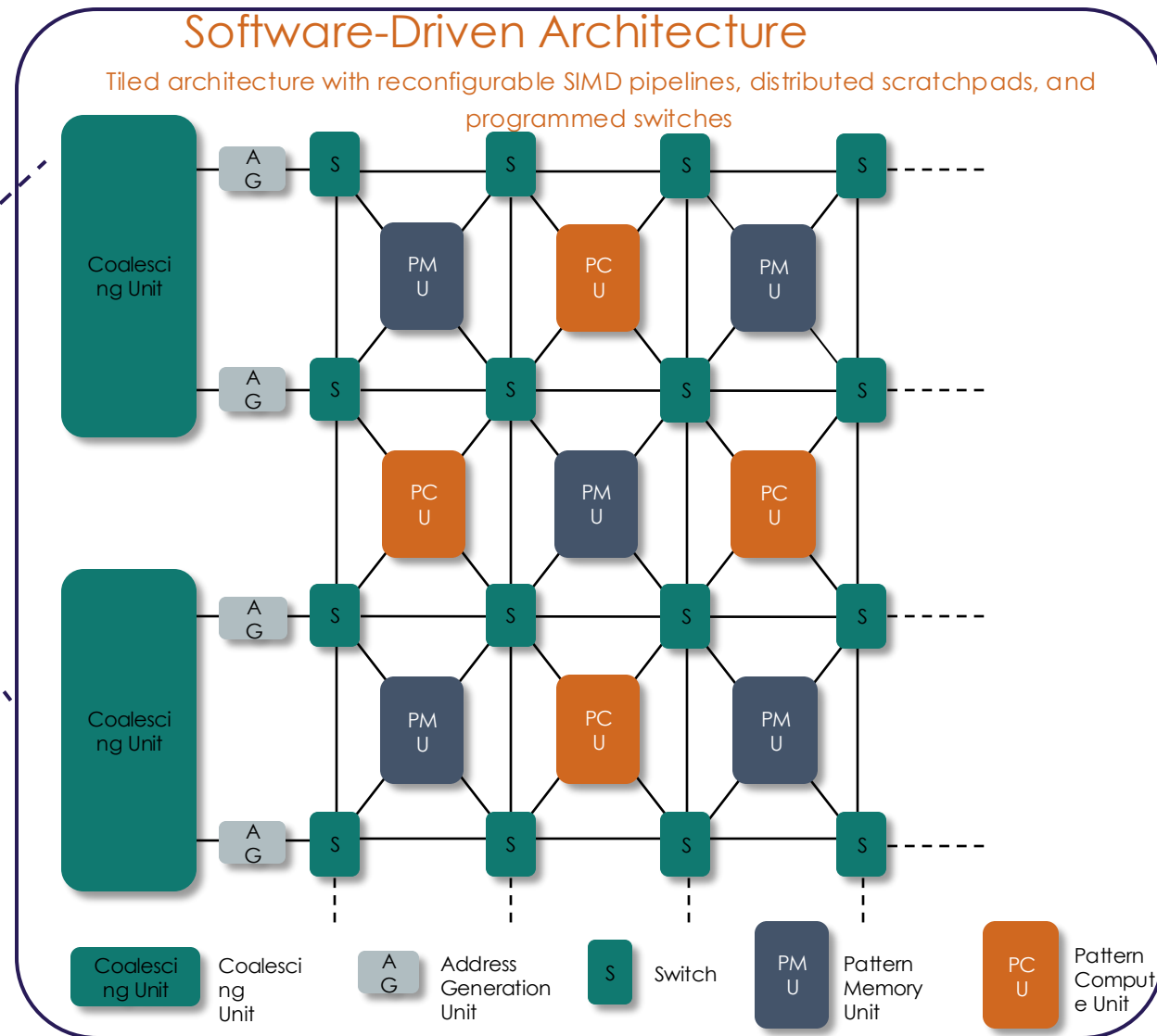
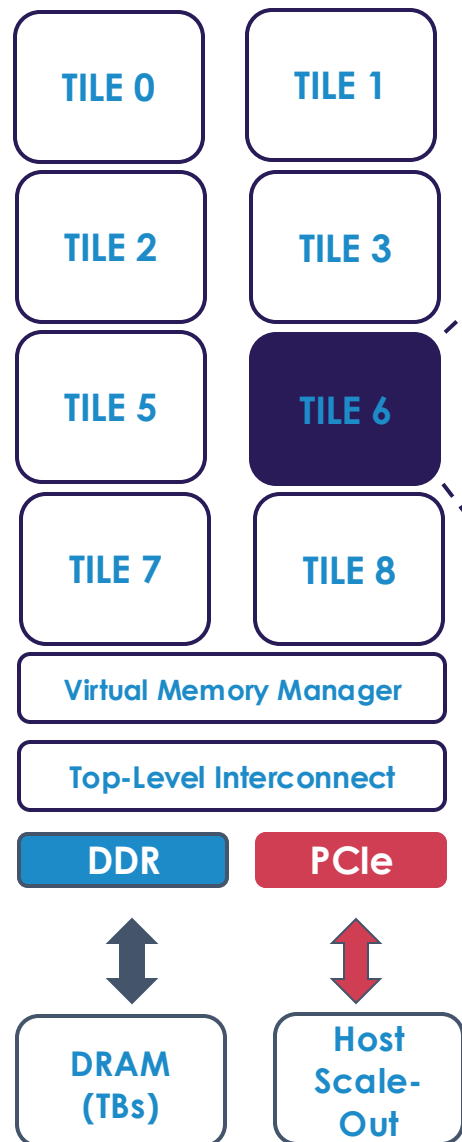
# SN30 RDU Chip and Architecture overview



- RDU broken up into 8-tiles
  - 160 PMU and PCUs per tile
  - Additional sub-components like coalescing units (CU) for connectivity to other tiles and off-chip components, switches to set up communication between PMU, PCUs, and CU
- Tile resource management: Combined or independent mode
  - Combined: Combine adjacent to form a larger logical tile for one application
  - Independent: Each tile controlled independently, allows running different applications on separate tiles concurrently.
- Direct access to TBs of DDR4 off-chip memory
- Memory-mapped access to host memory
- Scale-out communication support

Image Courtesy: SambaNova

# Cardinal SN30: Tile



- Interleaving of compute and memory units
- Routing data through the compute elements
- Dataflow Efficiency + Compute Capability + Large Memory Capacity

Image Courtesy: SambaNova

# SambaFlow Software Architecture

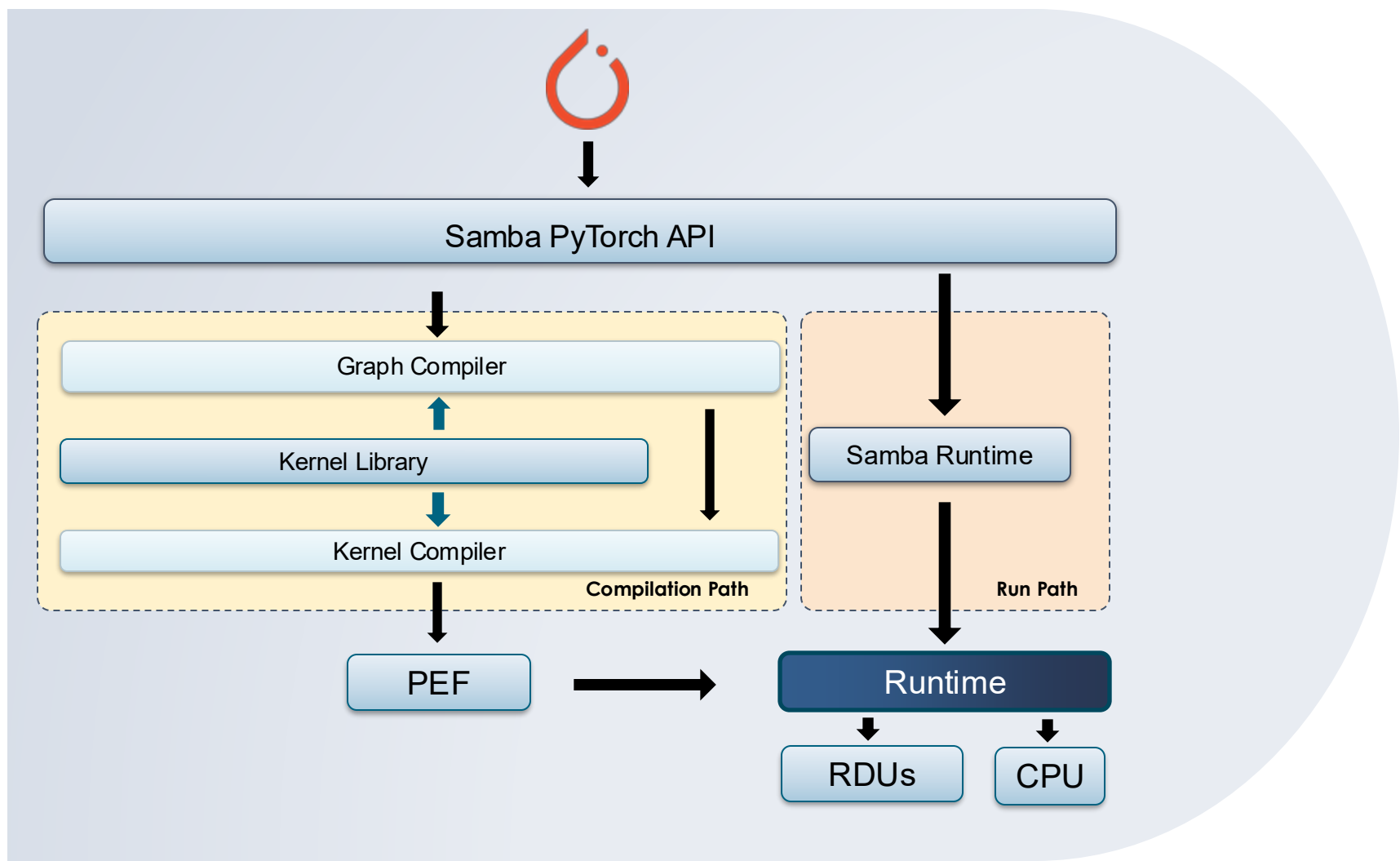
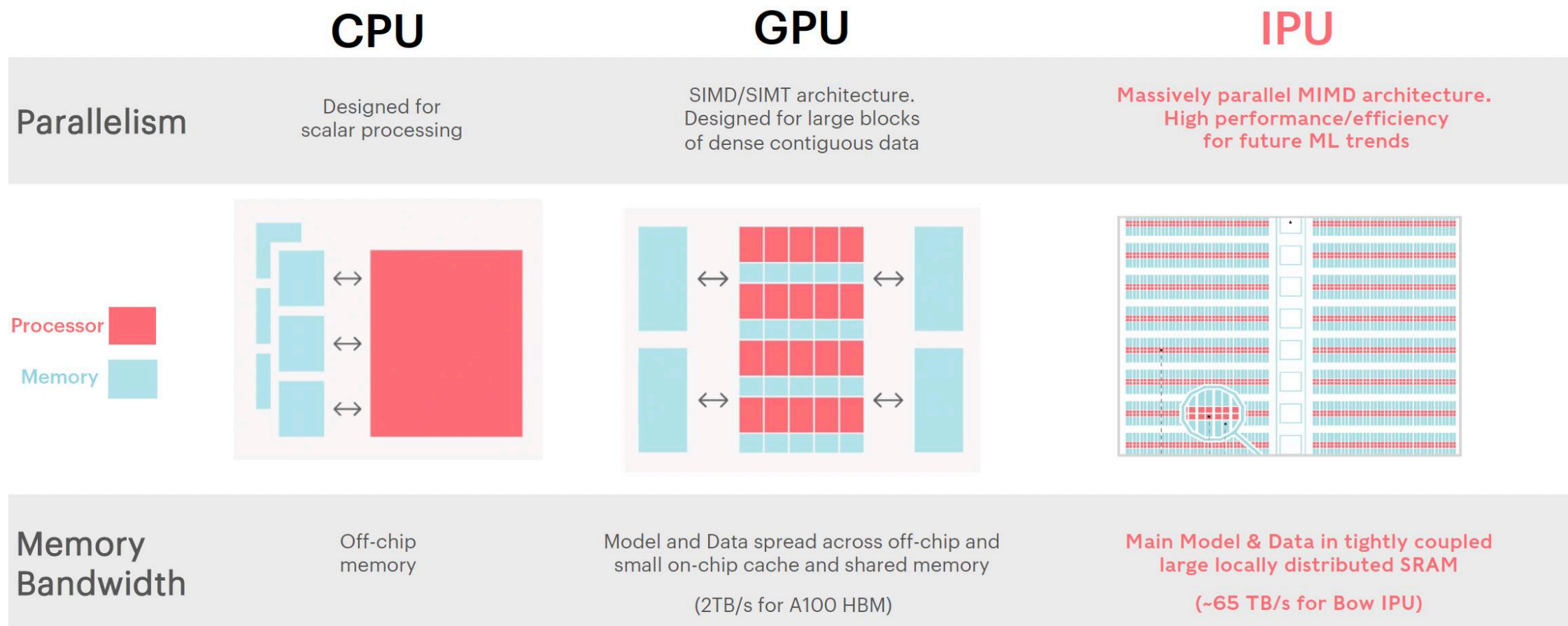


Image Courtesy: SambaNova

# Graphcore Intelligence Processing Unit (IPU)



Slide Courtesy: Graphcore



# Bulk Synchronous Parallel (BSP)

- The IPU uses the bulk-synchronous parallel (BSP) model of execution where the execution of a task is split into steps.
- Each step consists of the following phases:
  - local tile compute,
  - global cross-tile synchronization,
  - data exchange



## IPU-Tiles™

1472 independent IPU-Tiles™ each with an IPU-Core™ and In-Processor-Memory™

## IPU-Core™

1472 independent IPU-Core™

8832 independent program threads executing in parallel

## In-Processor-Memory™

900MB In-Processor-Memory™ per IPU

65TB/s memory bandwidth per IPU

# BOW IPU

## IPU-Exchange™

11 TB/s all to all IPU-Exchange™  
Non-blocking, any communication pattern

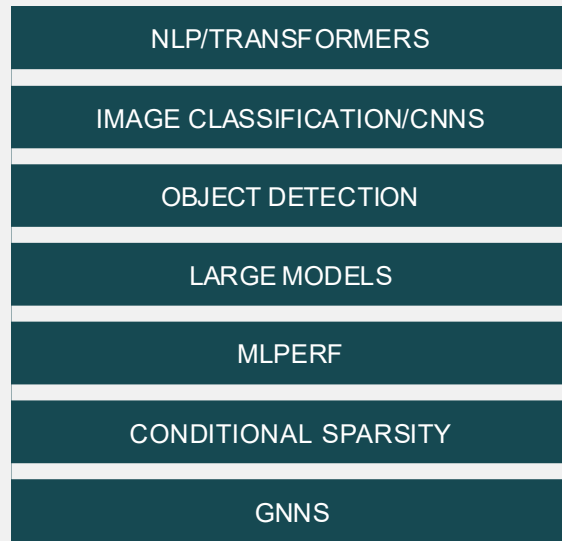
## PCIe

PCI Gen4 x16  
64 GB/s bidirectional bandwidth to host

## IPU-Links™

10 x IPU-Links,  
320GB/s chip to chip bandwidth

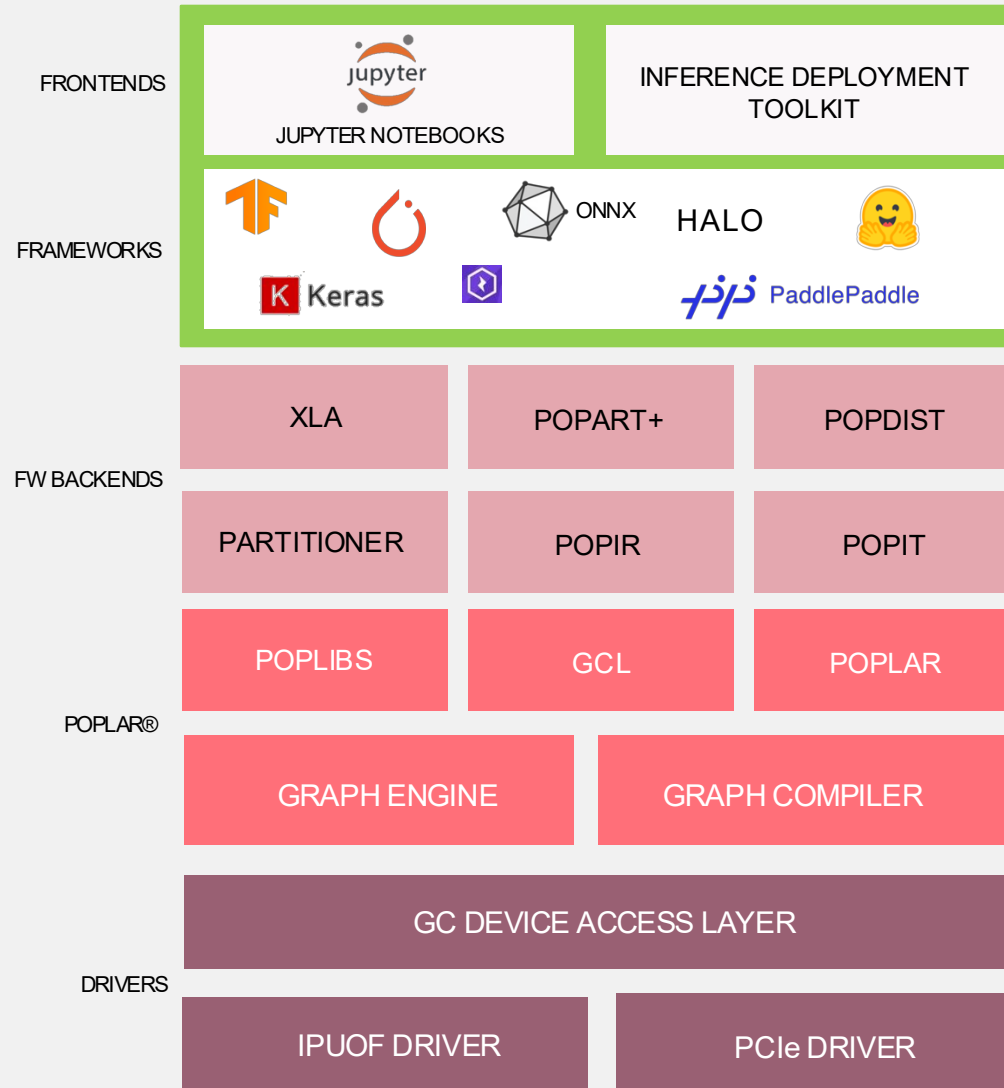
# GRAPHCORE SOFTWARE



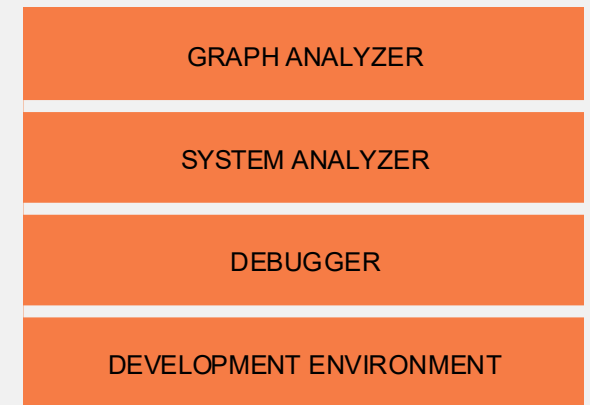
ML APPLICATIONS



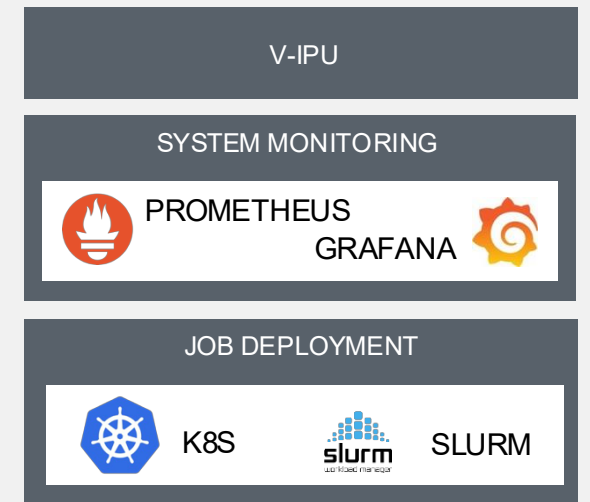
DEVELOPER ECOSYSTEM



POPLAR® SDK



POPVISION TOOLS



SYSTEM SOFTWARE



# Groq LPU Overview

## SRAM Memory

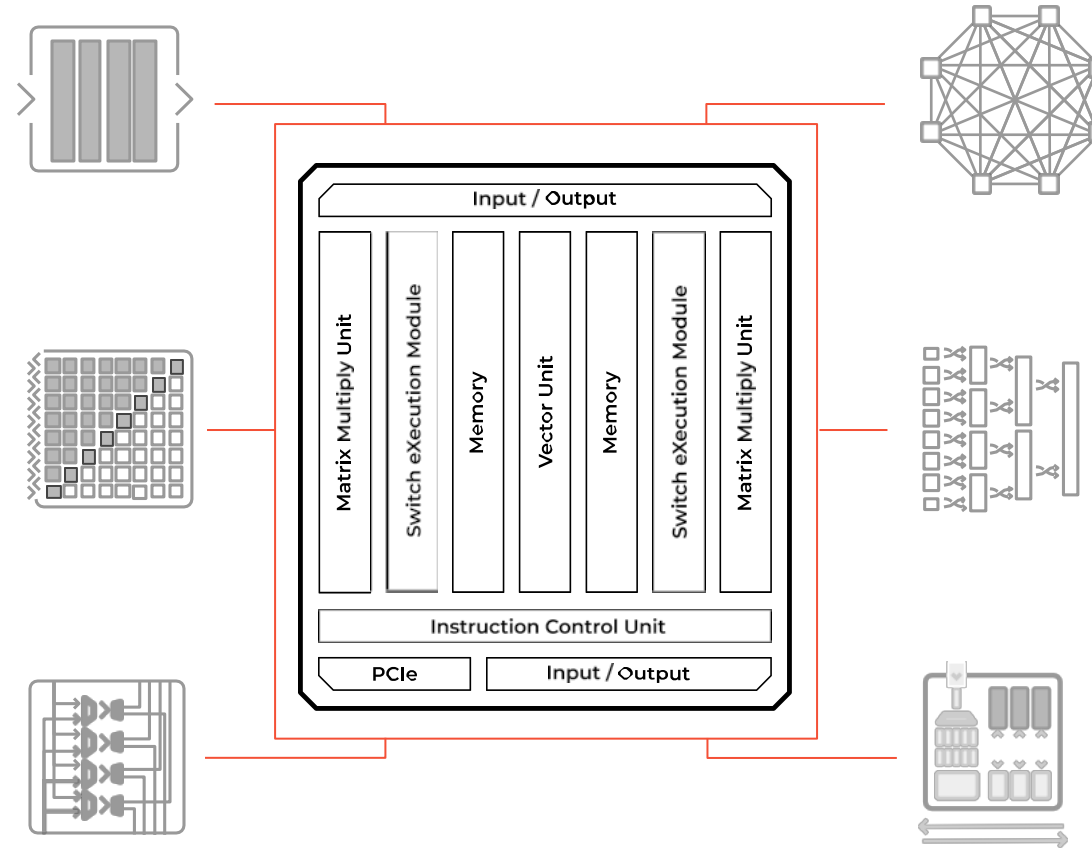
Massive concurrency  
80 TB/s of BW  
230MB capacity  
Stride insensitive

## Groq TruePoint™ Matrix

4x Engines  
750 TOP/s int8  
188 TFLOP/s fp16  
320x320 fused dot product

## Programmable Vector Units

5,120 Vector ALUs for high performance



## Networking

480 GB/s bandwidth  
Extensible network scalability  
Multiple topologies

## Data Switch

Shift, Transpose, Permuter for improved data movement and data reshapes

## Instruction Control

Multiple instruction queues for instruction parallelism

# Groq LPU Building Blocks

Build different types of specialized SIMD units



**MXM**

Matrix  
operations



**VXM**

Vector  
Operations



**SXM**

Data Reshapes



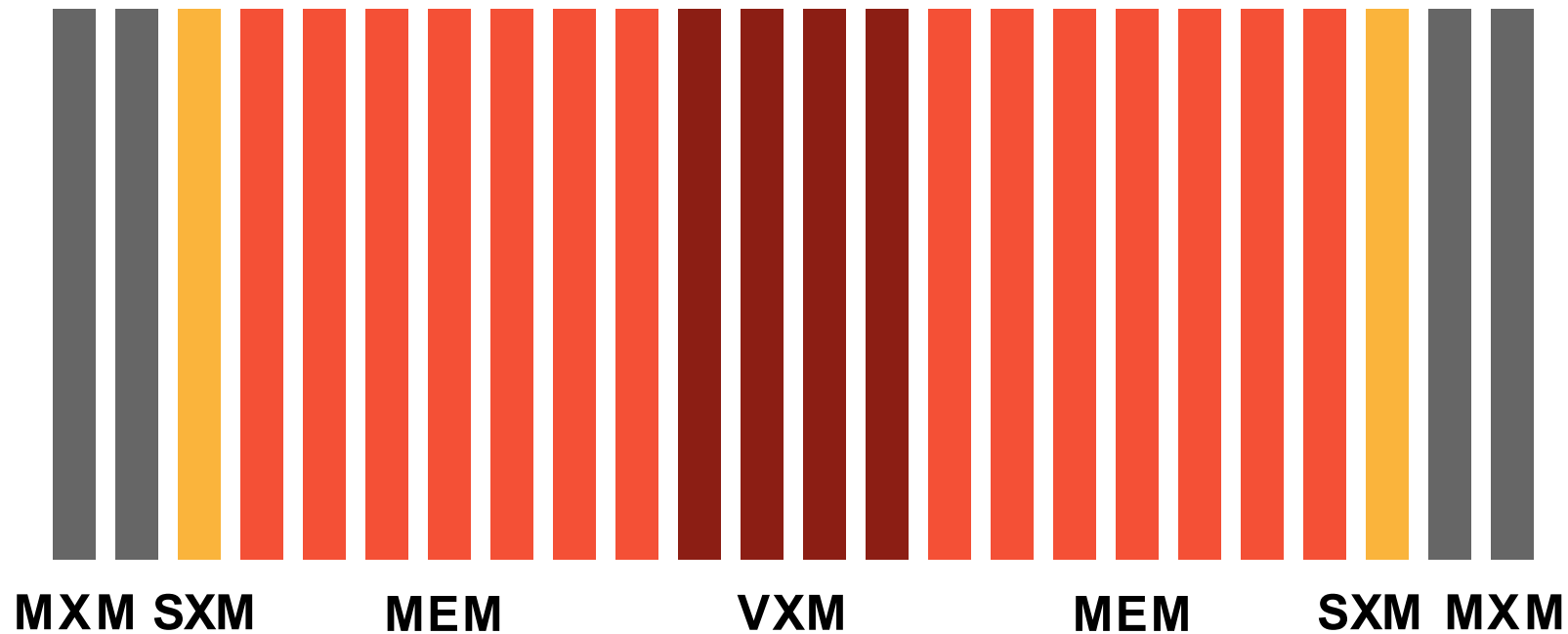
**MEM**

On-chip SRAM



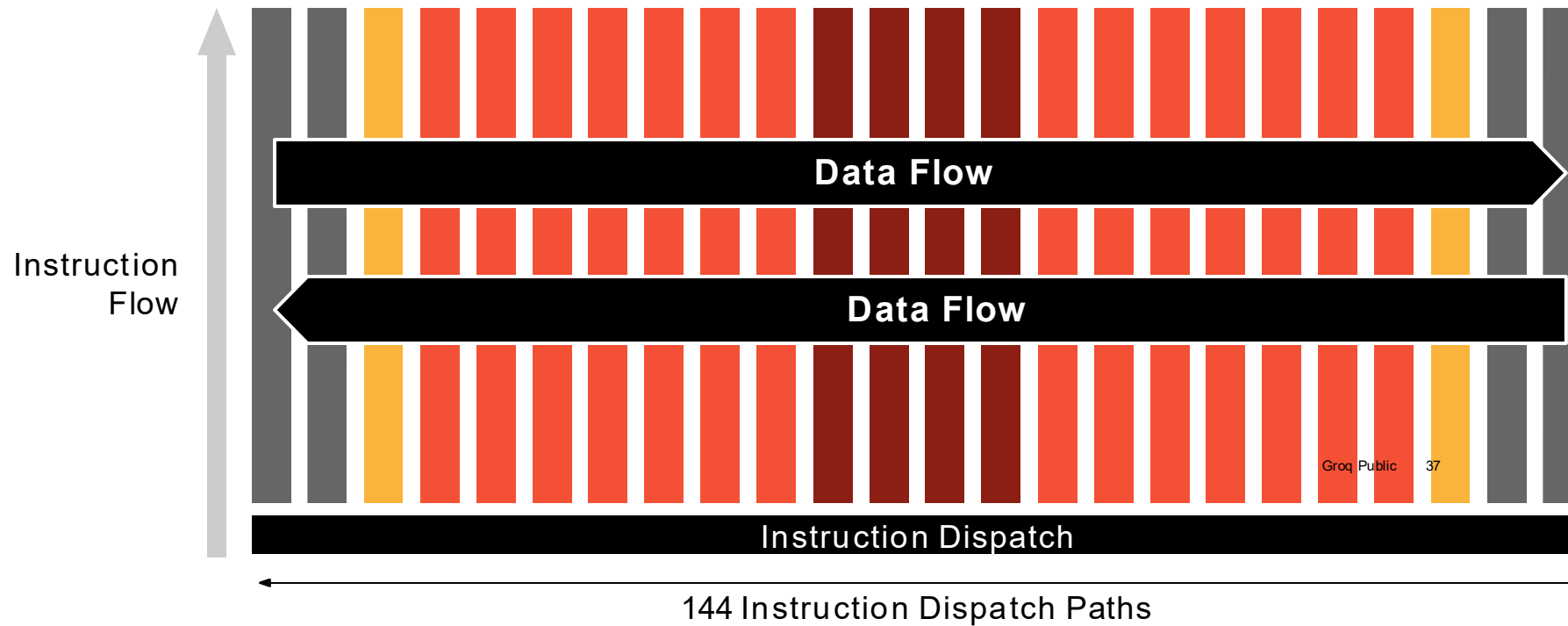
# Groq LPU Building Blocks

Lay out SIMD units across chip area



# Groq LPU Building Blocks

High-bandwidth “Stream Registers” for passing data between units



## Software-controlled memory

No dynamic hardware caching

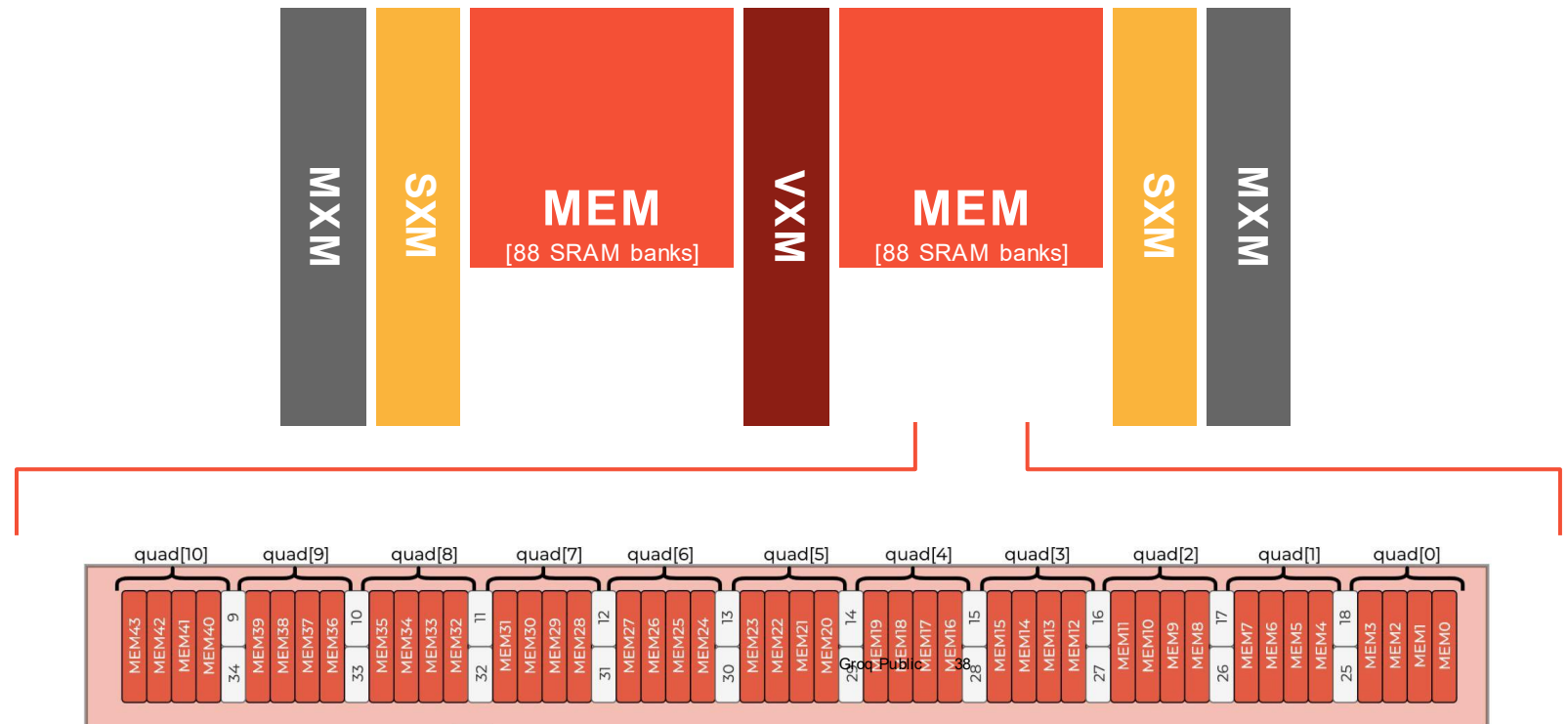
- Compiler aware of all data locations at any given point in time

Flat memory hierarchy  
(no L1, L2, L3, etc)

- Memory exposed to software as a set of physical banks that are directly addressed

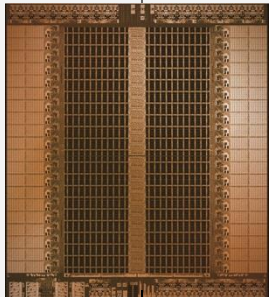
Large on-chip memory capacity (220 MiB) at very high-bandwidth (80 TBps)

- Achieves high compute efficiency even at low operational intensity

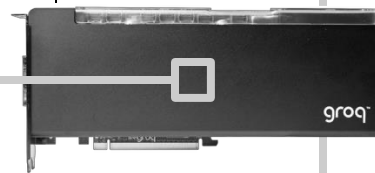


## GroqChip™

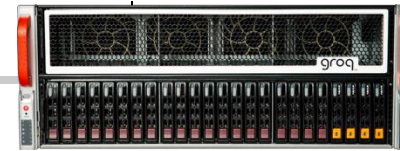
The purpose-built  
Language Processing  
Unit™ Inference Engine



## GroqCard™



## GroqNode™



## Dell Servers



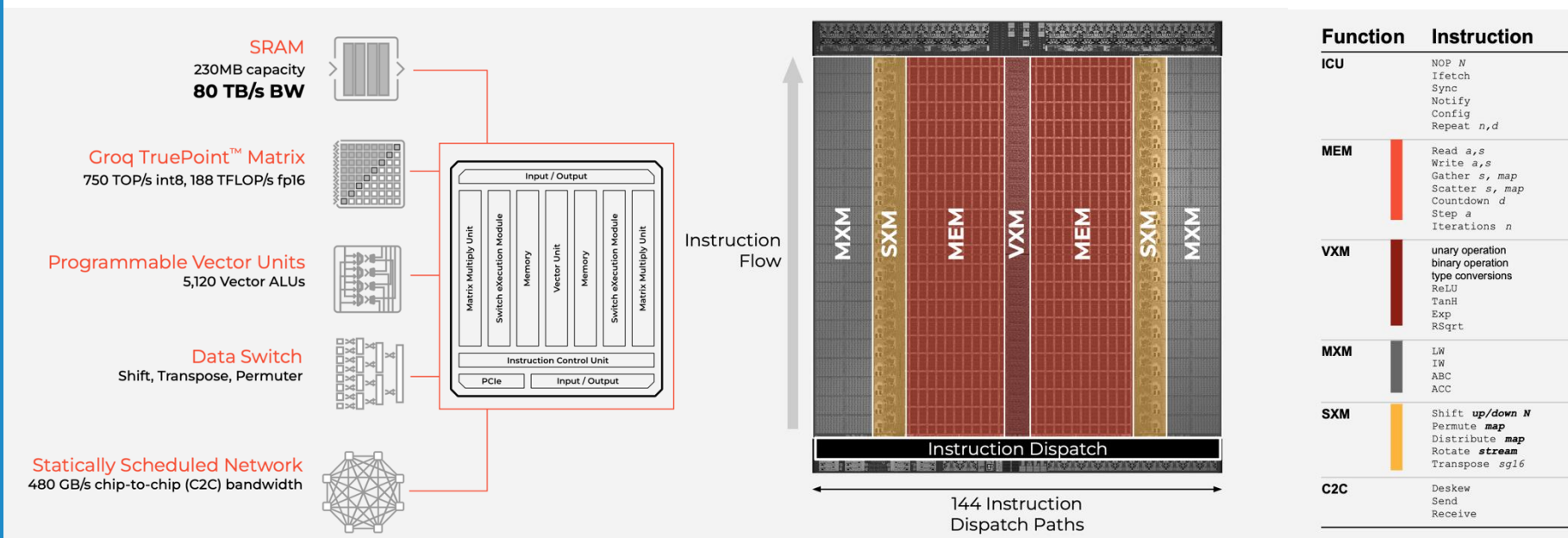
## GroqRack™



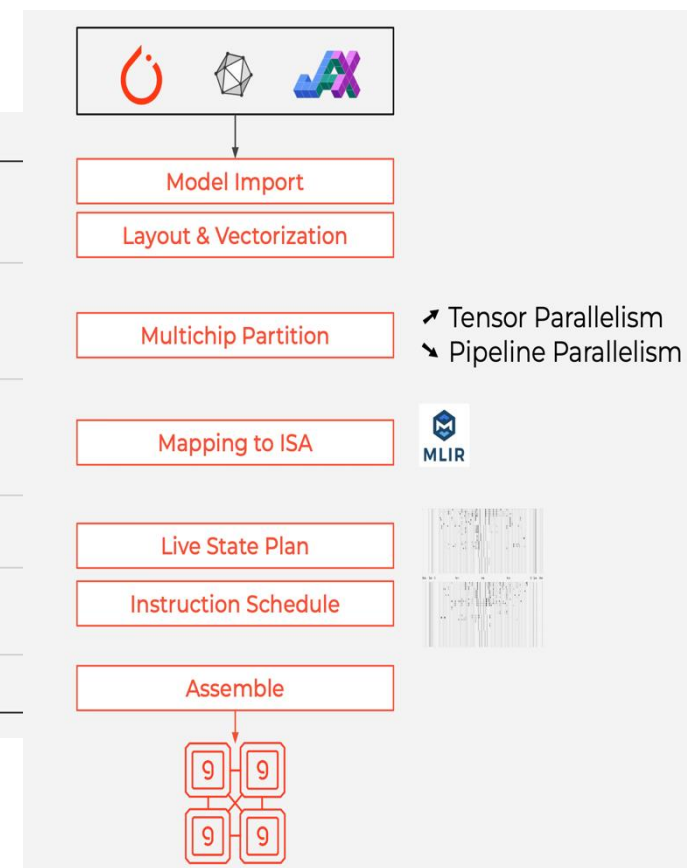
### ≡ EXCEPTIONAL.

at sequential processing. The LPU™ Inference Engine is designed to scale and is more power-efficient, with greater performance, than a GPU for AI applications like LLMs.

# Groq Hardware Architecture







Function	Instruction
ICU	NOP <i>N</i>
	Ifetch
	Sync
	Notify
	Config
	Repeat <i>n, d</i>
MEM	Read <i>a, s</i>
	Write <i>a, s</i>
	Gather <i>s, map</i>
	Scatter <i>s, map</i>
	Countdown <i>d</i>
	Step <i>a</i>
	Iterations <i>n</i>
VXM	unary operation
	binary operation
	type conversions
	ReID
	TanH
MXM	Exp
	RSqrt
SXM	LW
	IW
	ABC
	ACC
C2C	Shift <i>up/down N</i>
	Permute <i>map</i>
	Distribute <i>map</i>
	Rotate <i>stream</i>
	Transpose <i>sg16</i>
C2C	Deskew
	Send
	Receive

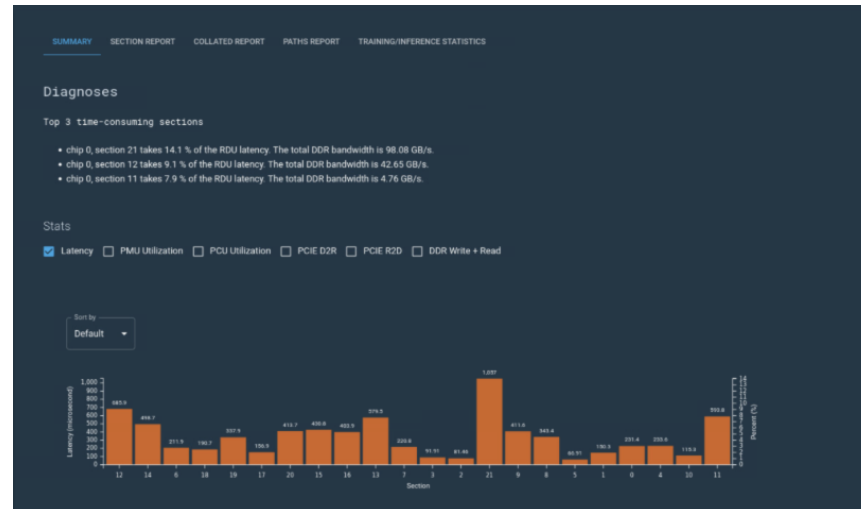
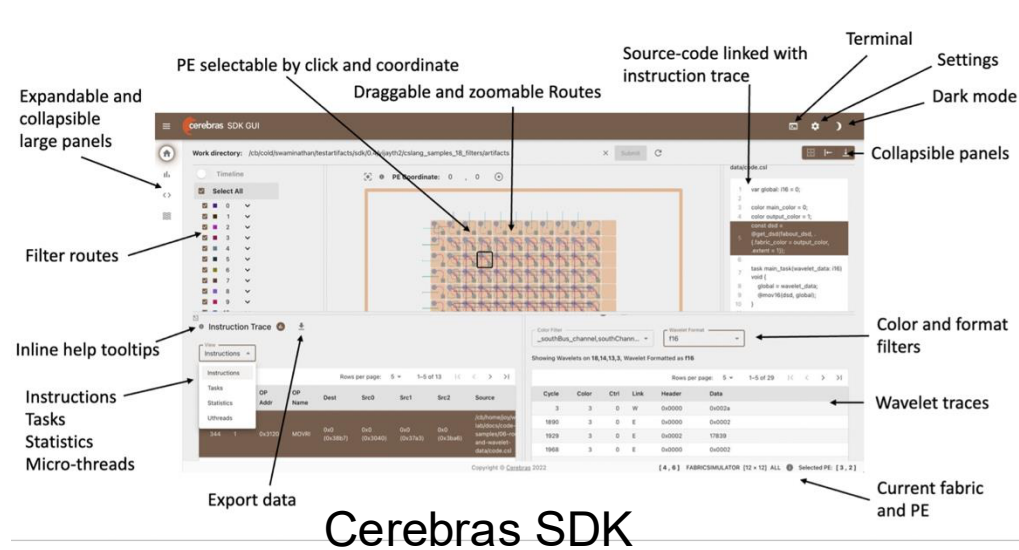




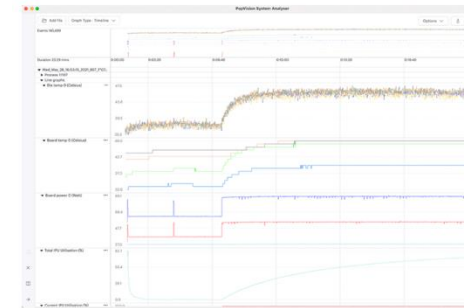
# HPC Software ecosystem on AI Accelerators

	• Cerebras Software Language
	• AI4S (C/C++)
	• Poplar C/C++ API • BSP
	• Groq Runtime API • C/C++

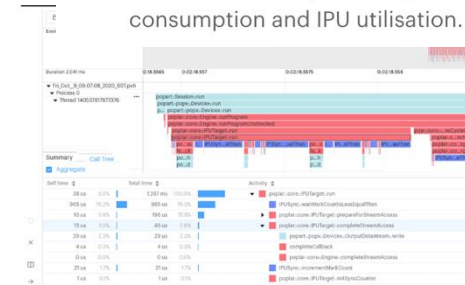
# Tools on AI Accelerators



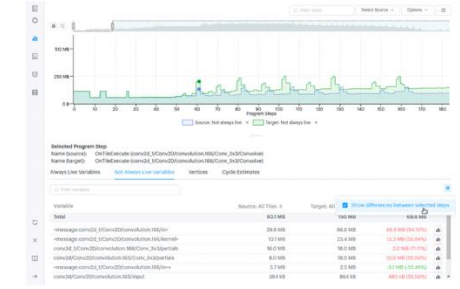
## SambaTune on SambaNova



Plot graph data of any numerical data points from the host or IPU processor systems, such as board temperature, power consumption and IPU utilisation.



Understand the execution of IPU-targeted software on your host system processors. Identify any bottlenecks between CPUs and IPU across a visual interactive timeline.



Open two reports at once to compare their memory, execution, liveness and operations. Visualise where efficiencies can be made with different model parameters.

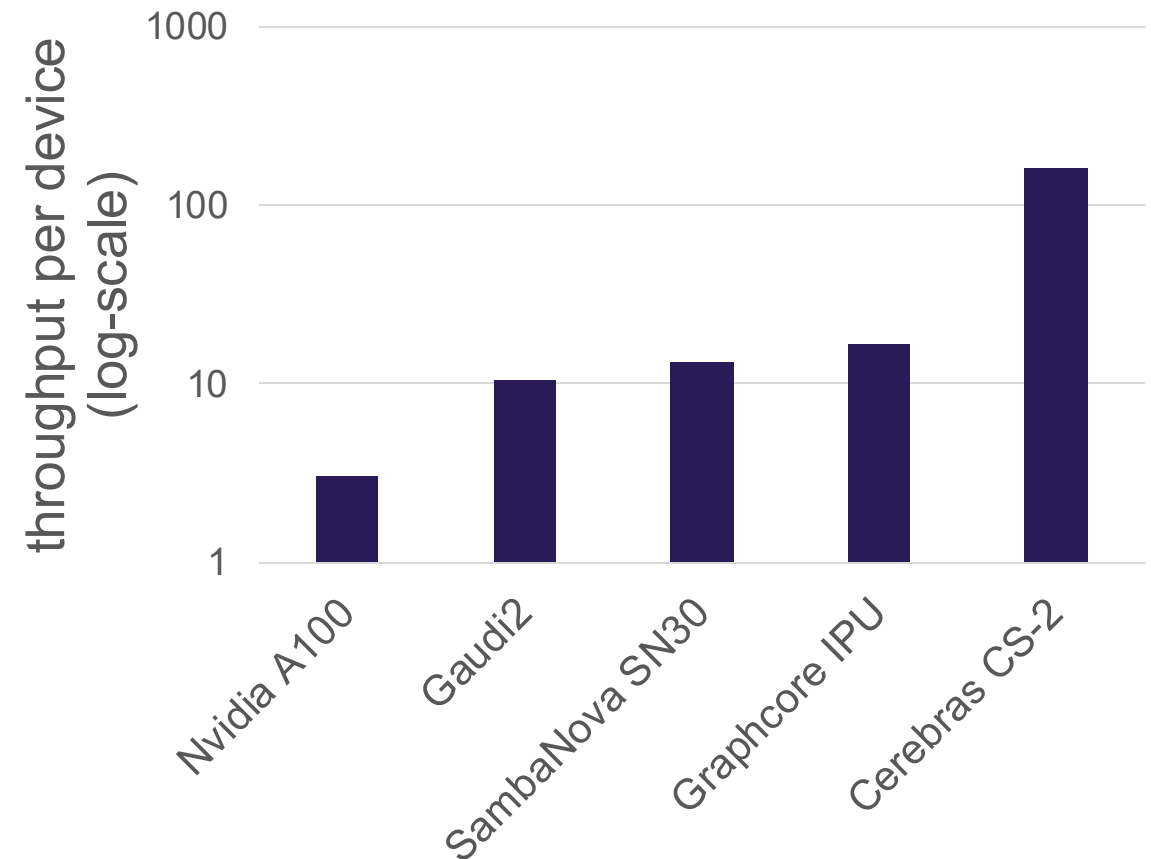


Capture memory information from your ML models when executed on IPUs. Inspect variable placement, size and liveness throughout the execution.

## PopVision on GraphCore

# Training Performance

System	Number of Devices	Throughput
Nvidia A100	64	193
Gaudi 2	16	170
SambaNova SN30	16	212
Graphcore Bow-Pod64	16	266
Cerebras CS-2	2	320



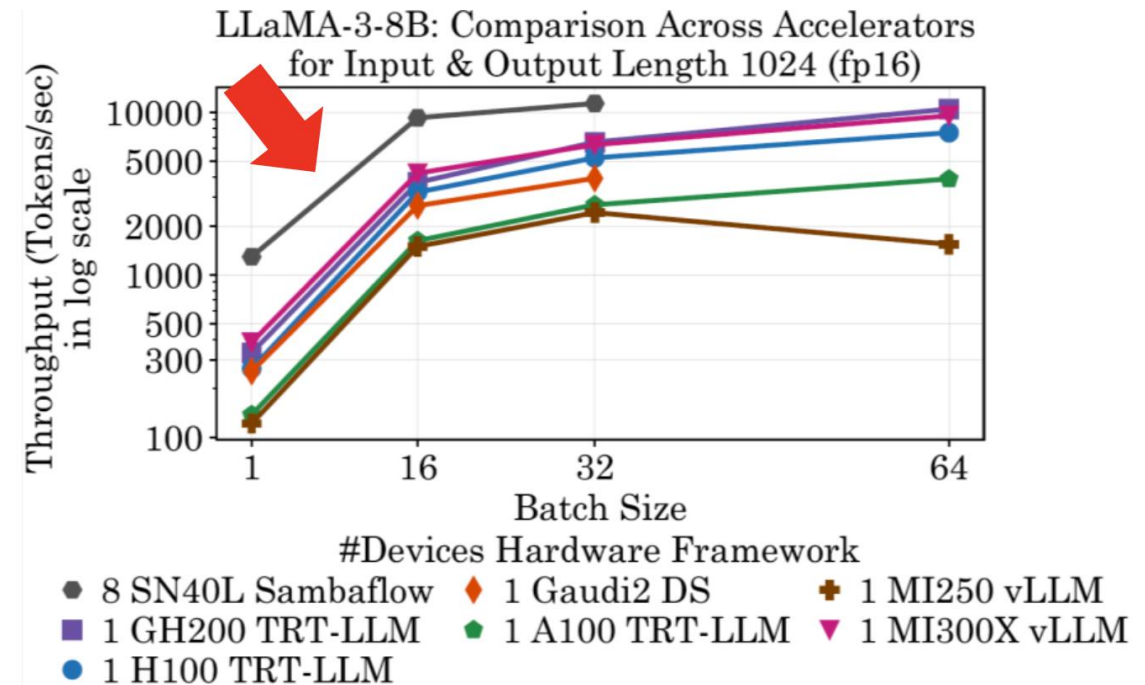
Used GPT-2 XL 1.5B parameter model, OWT dataset

- same sequence length 1k, custom software stack, half-precision
- Runs on A100s used Megatron-Deepspeed, out-of-box runs with no additional optimizations
- 16 SN30 RDUs, 2 CS-2s, and 16 IPU match the performance on 64 A100s

Emani et al. "Toward a Holistic Performance Evaluation of Large Language Models Across Diverse AI Accelerators", Heterogeneity in Computing Workshop (HCW) at IPDPS24.

# Inference Performance

- SambaNova SN40L achieves has the best performance among all the accelerators we benchmarked
- Nvidia GH200 > H100 > A100 (in terms of throughput)
- MI300X and GH200 are comparable
- Habana Gaudi's performance is between A100 and H100
- The performance of AMD MI250 saturates for large batch sizes

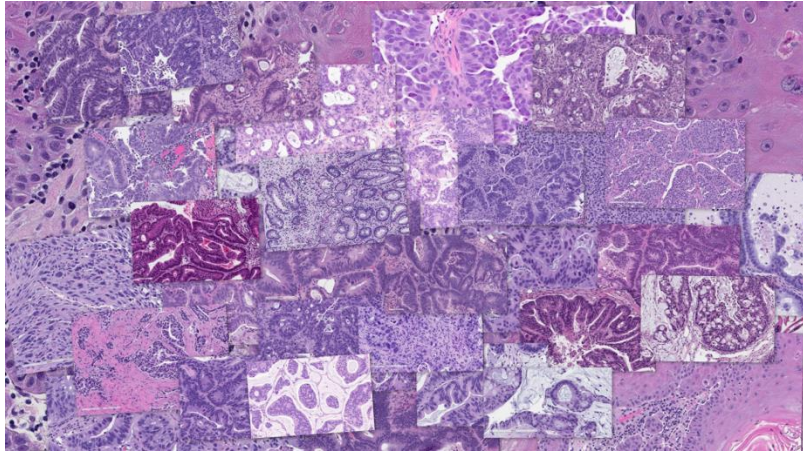


Krishna et. al. 'LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators' PMBS24

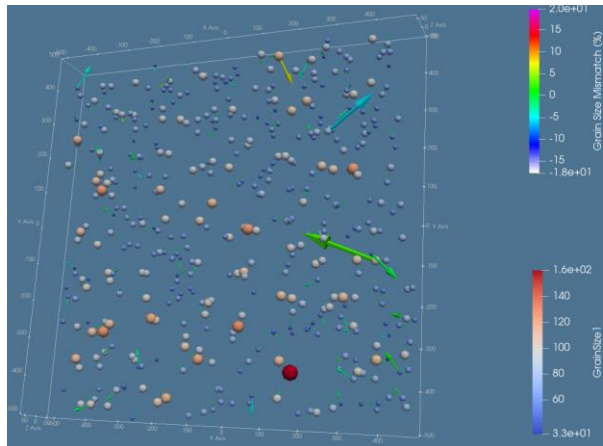
# **Advancing Science through AI Accelerators**



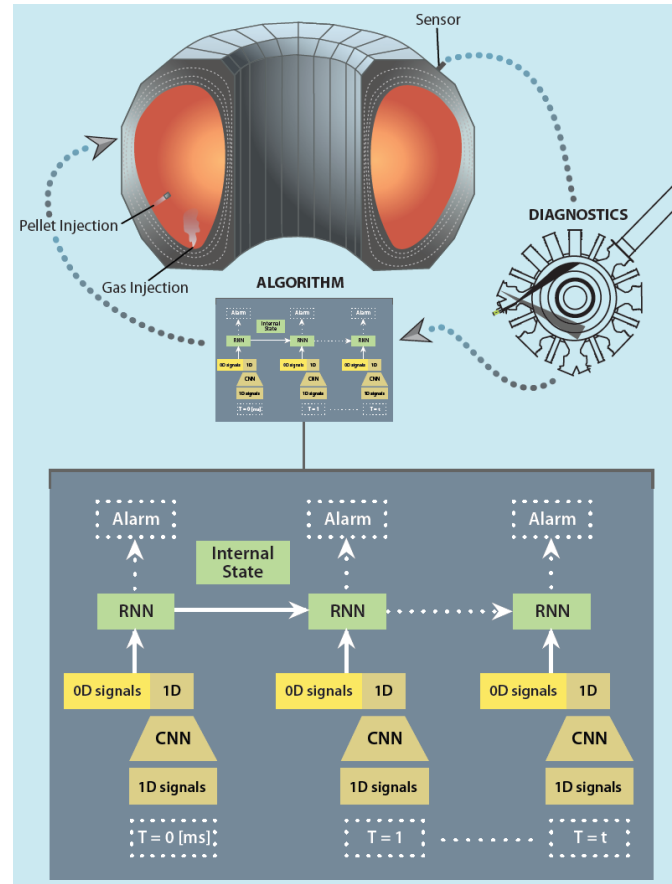
# AI for Science and HPC applications on AI Testbed



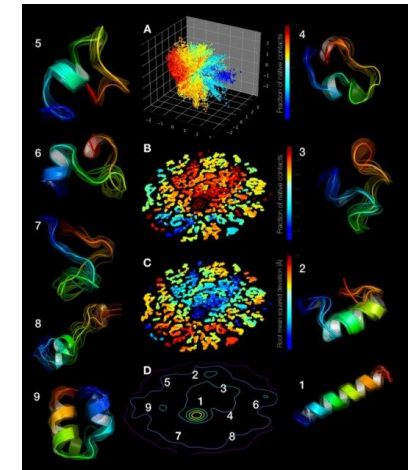
Cancer drug response prediction  
(Credit: Candle)



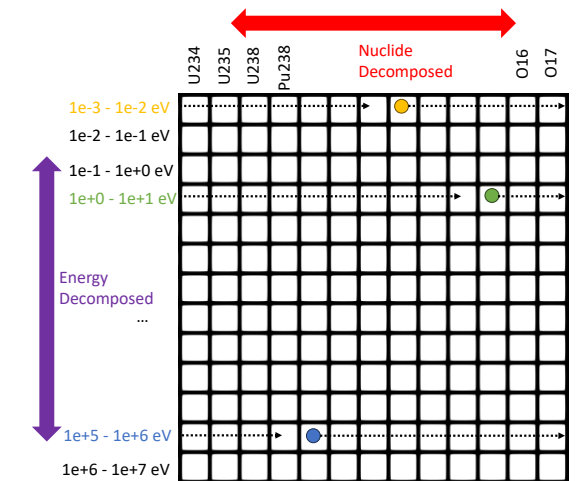
Imaging Sciences-Braggs Peak  
(Credit: Z. Liu)



Tokamak Fusion Reactor operations  
(Credit: K. Felker)



Protein-folding (Image: NCI)



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

and more..

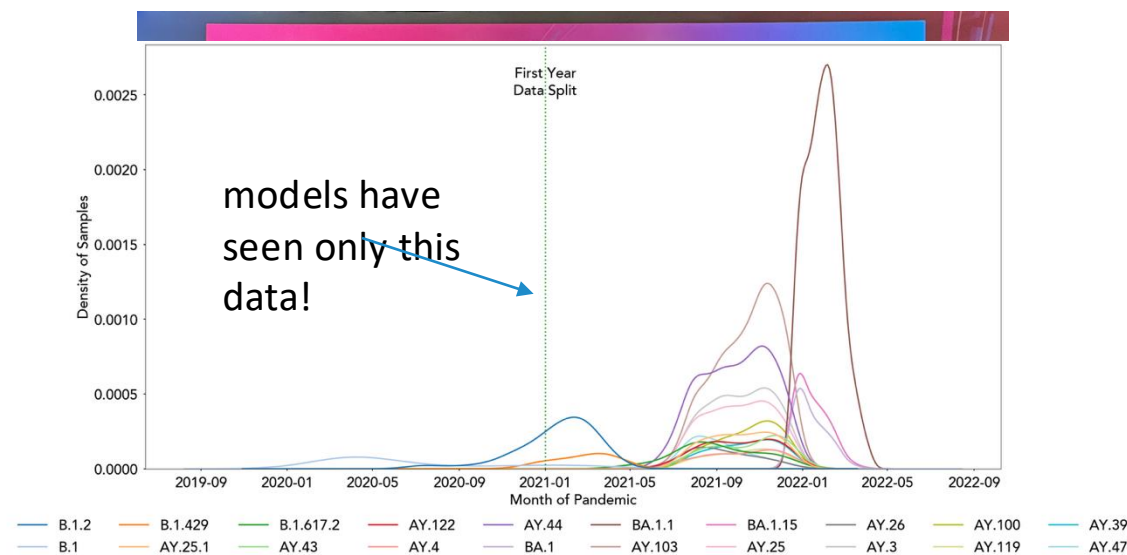
# Genome-scale Language Models (GenSLMs)

## 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,***

DOI: <https://doi.org/10.1101/2022.10.10.511571>

# 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*

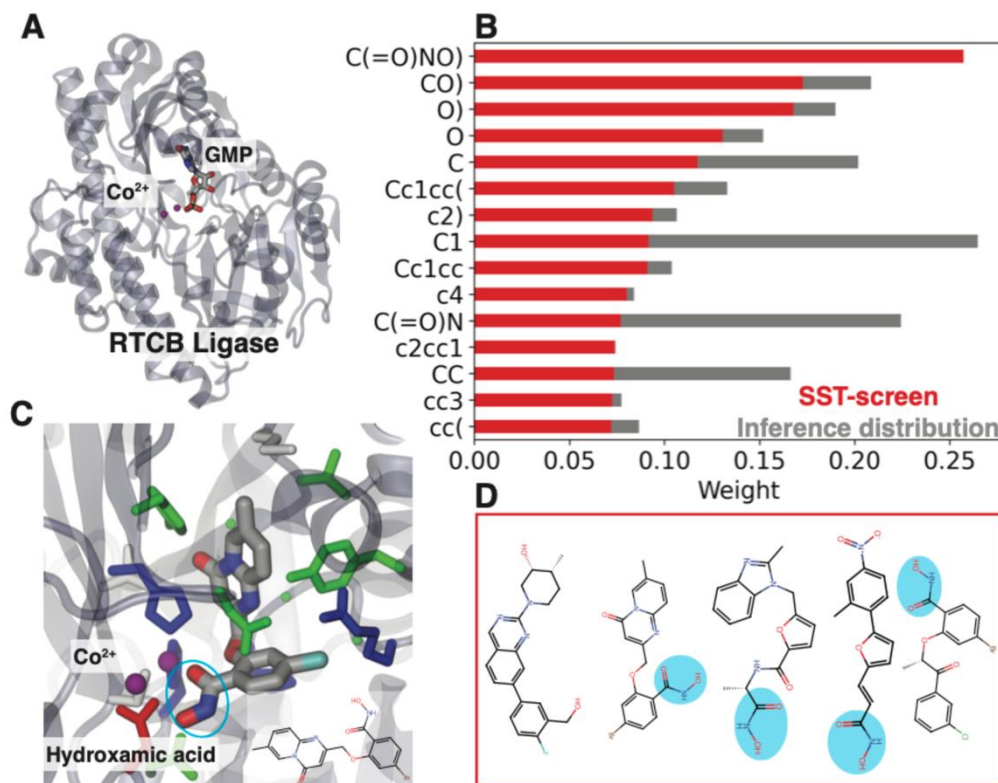
System	Number of Devices	Throughput (tokens/sec)	Improvement	Energy Efficiency
NVIDIA A100	8	1150	1.0	1.0
SambaNova SN30	8	9795	8.5	5.6
Cerebras CS-2	1	29061	25	6.5

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

"Toward a Holistic Performance Evaluation of Large Language Models Across Diverse AI Accelerators", M.Emani et al.,  
HCW workshop, IPDPS 2024

# Accelerating Drug Design and Discovery with Machine Learning

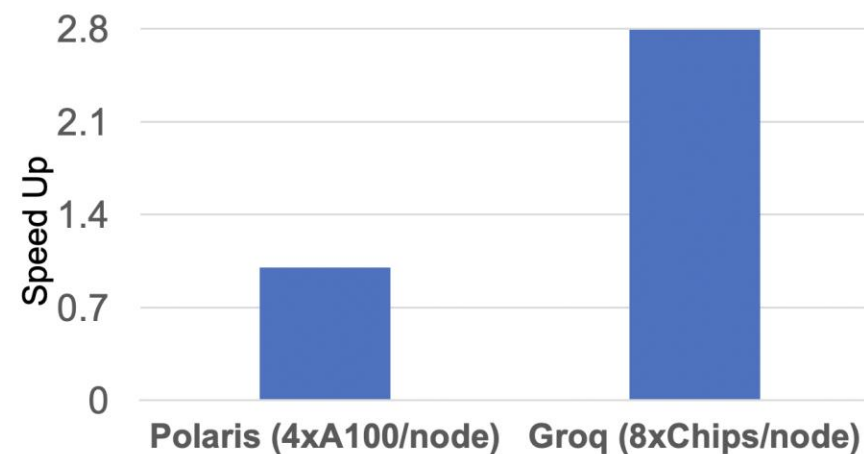
Application code: Simple SMILES Transformer



High performance binding affinity prediction with a Transformer-based surrogate model

Archit Vasan\*, Ozan Gokdemir\*<sup>†</sup>, Alexander Brace\*<sup>†</sup>, Arvind Ramanathan\*<sup>†</sup>, Thomas Brettin\*, Rick Stevens\*<sup>†</sup>, Venkatram Vishwanath\*

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



Courtesy: Archit Vasan

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

Bert based encoder model to identify compounds with high binding affinity directly on the SMILES string input.



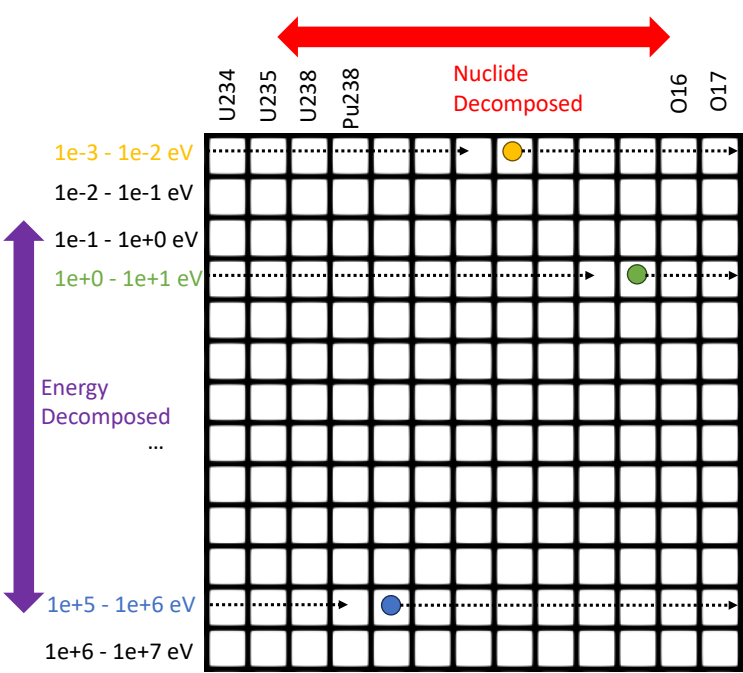
# Monte Carlo with Single Cycle Latency: leveraging the cerebras cs-2 for acceleration of a latency-bound HPC simulation workload

**Challenge:** We examine the feasibility of performing continuous energy **Monte Carlo (MC) particle transport on the Cerebras WSE-2 AI accelerator** by porting XSBench to the Cerebras “CSL” programming model. The MC algorithm has traditionally been bandwidth/latency-bound, making the WSE-2’s 40 GB of 1-cycle SRAM an attractive architecture. The critical challenge is to decompose data and tasks across the WSE-2’s ~750,000 distributed memory processing elements (PEs), each having only 48 KB of memory.

## Outcome:

- Developed several novel algorithms for decomposing data structures across the WSE-2’s 2D network grid, for flowing particles (tasks) through the WSE-2, and for performing dynamic load balancing.
- Developed a method for exploiting the WSE-2’s hardware random number generation capabilities to accelerate kernel by 65%.
- WSE-2 was found to run **130x faster than a highly optimized CUDA version of the kernel run on an NVIDIA A100 GPU.**

Computational Physics Communications  
(<https://doi.org/10.1016/j.cpc.2023.109072>)



MC cross section data decomposition across a 2D grid of WSE-2 processing elements. This diagram shows the third phase of our algorithm where particles are exchanged in a round-robin manner to visit all nuclides in the row.

	Transistor Count [Trillion]	Peak Power [kW]	Monte Carlo XS Lookup FOM [Lookups/s]
A100 GPU	0.0542	0.4	6.43E+07
Cerebras CS-2	2.6	22.8	8.36E+09
Cerebras/A100	48	57	130



# Observations, Challenges and Insights

- Significant speedup achieved for a wide-gamut of scientific ML applications
  - Easier to deal with larger resolution data and to scale to multi-chip systems
- Room for improvement exists
  - Porting efforts and compilation times
  - Coverage of DL frameworks, support for performance analysis tools, debuggers
- Profiling studies to really understand the unique HW and SW capabilities
- Limited capability to support low-level HPC kernels
  - Work in progress to improve coverage

\* Performance/\$ is interesting too!

# Ongoing Efforts

- Evaluate emerging AI models (MoE variants, reasoning models etc)
- Evaluate new AI accelerators offerings and incorporate promising solutions as part of the testbed
- Integrate AI testbed systems with the PBSPro scheduler to facilitate effective job scheduling across the accelerators
- Understand how to integrate AI accelerators with ALCF's existing and upcoming supercomputers to accelerate science insights

# Useful Links

## ALCF AI Testbed

- Overview: <https://www.alcf.anl.gov/alcf-ai-testbed>
- Guide: <https://docs.alcf.anl.gov/ai-testbed/>
- Training:  
<https://www.alcf.anl.gov/ai-testbed-training-workshops>
- Allocation Request: [Allocation Request Form](#)
- Support: [support@alcf.anl.gov](mailto:support@alcf.anl.gov)



# Getting started with ALCF AI Testbed

# Getting Started on ALCF AI Testbed

## Available for Allocations

- Cerebras CS-2,
- SambaNova Datascale SN30,
- GroqRack
- Graphcore Bow Pod64

## AI Testbed User Guide

## Director's Discretionary (DD) awards

- Scaling code
- Preparing for future computing competition
- Scientific computing in support of strategic partnerships.

### Allocation Request Form

<https://www.alcf.anl.gov/science/directors-discretionary-allocation-program>

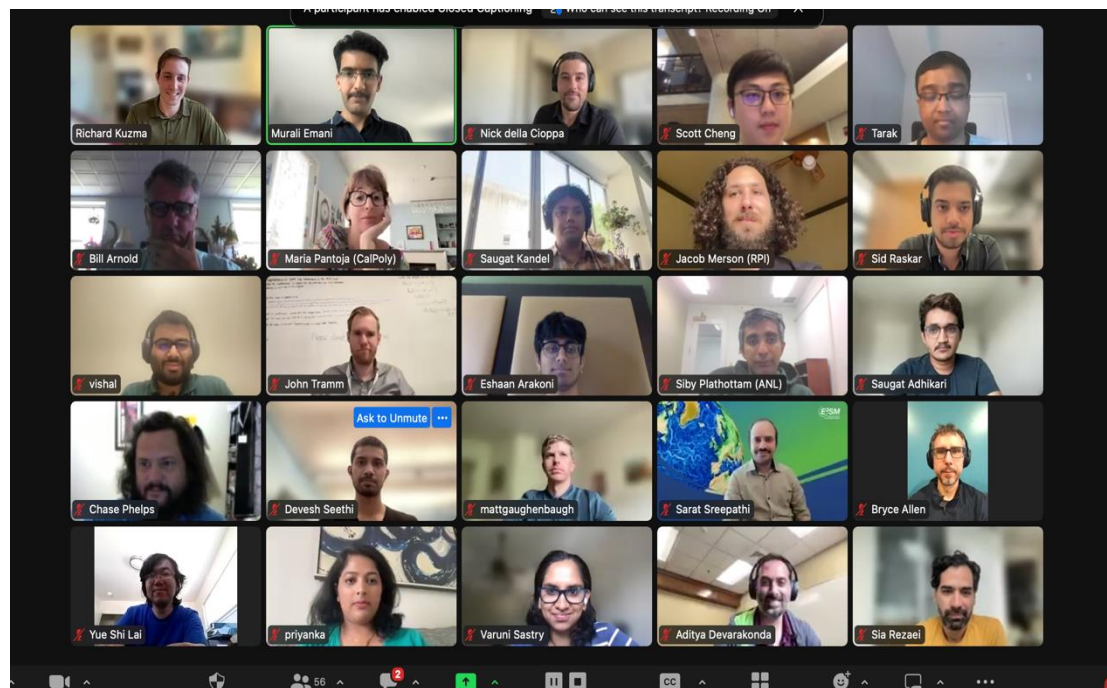
## NAIRR Pilot

aims to connect U.S. researchers and educators to computational, data, and training resources needed to advance AI research and research that employs AI.

<https://nairrpilot.org/>



# AI Testbed Community Engagement



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

- ATPESC Training
- Introduction to AI-driven Science on Supercomputers



[Full Program](#) [My Schedule](#) [Contributors](#) [Organizations](#) [Search](#)

## 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, GPGUs, FPGA, etc.).

NEXT PRESENTATION > ⌚ STARTS IN 118:23:07

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



Murali Emani  
Argonne National Laboratory (ANL)



Leighton Wilson  
Cerebras Systems

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

Next tutorial at **ISC25, June 13, 2025**

# 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 Raffanetti, Valerie Taylor, Murali Emani, Venkatram Vishwanath, "LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators," 2024 IEEE/ACM International Workshop on Performance Modeling, Benchmarking and Simulation of High-Performance Computer Systems (PMBS), Atlanta, GA, USA, 2024.
- **Toward a Holistic Performance Evaluation of Large Language Models Across Diverse AI Accelerators**  
Murali Emani, Sam Foreman, Varuni Sastry, Zhen Xie, William Arnold, Rajeev Thakur, Venkatram Vishwanath, Michael E Papka, Sanjif Shanmugavelu, Darshan Gandhi, Hengyu Zhao, Dun Ma, Kiran Ranganath, Rick Weisner, Jiunn-yeu Chen, Yuting Yang, Natalia Vassilieva, Bin C Zhang, Sylvia Howland, Alexander Tsyplikhin. 2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)
- **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,**  
DOI: <https://doi.org/10.1101/2022.10.10.511571>
- **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*

# Recent Publications

- **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  
DOI: <https://doi.org/10.3389/fphy.2022.958120>
- **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, Hyunseung 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.
- Murali Emani, Michael Papka, William Arnold, Varuni Sastry, Sid Raskar, Krishna Teja-Chitty Venkata, Rajeev Thakur, Ray Powell, John Tramm, 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.