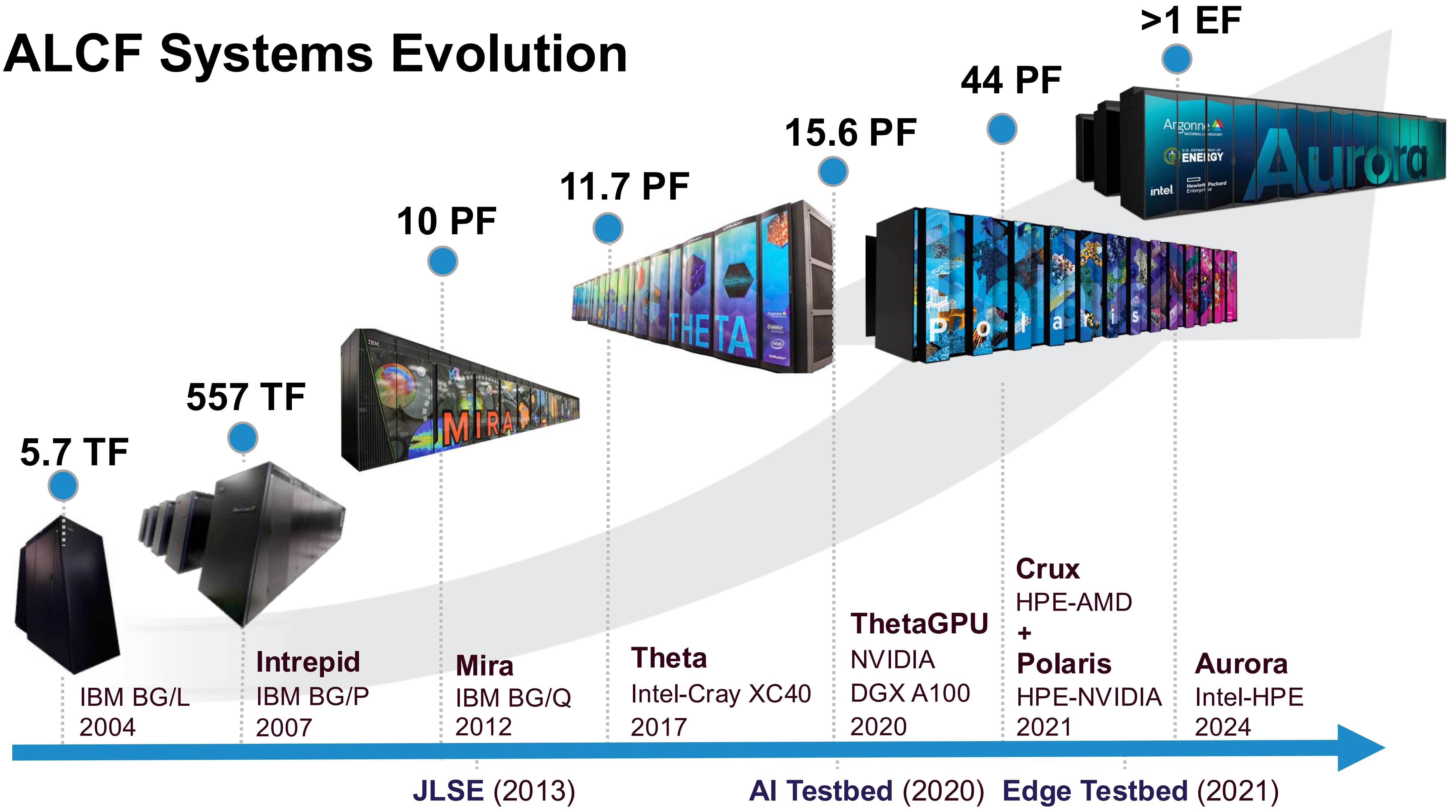
ALCF AI Testbeds

Murali Emani, Varuni Sastry
Argonne Leadership Computing Facility
{memani,vsastry}@anl.gov

ALCF AI Science training series
Nov 11, 2025

ALCF Systems Evolution



Argonne NATIONAL LABORATORY

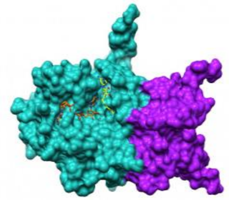
Integrating AI Systems in Facilities

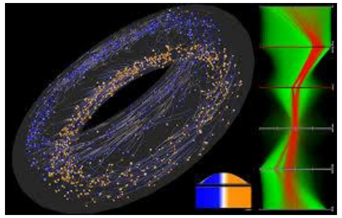
Computing Facility

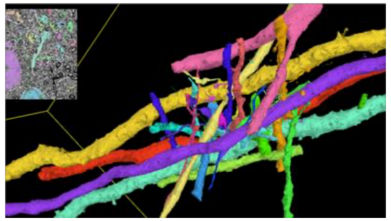
Experimental Facility

AI-Accelerators

AI-Edge accelerator







Simulations

Data-driven Models

ALCF AI Testbeds

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

Cerebras (CS-3)

SambaNova SN30/SN40L

Groq

Graphcore

Tenstorrent

Coming soon!!

- 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

Training

- Cerebras
- Sambanova SN30

SN-30 8 nodes of 8 RDUs

Cerebras CS-3 – 4 WSE

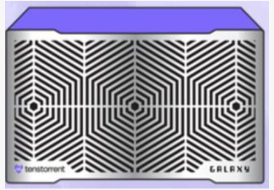
Inference

- SN40L – Metis
- Groq
- Cerebras
- Tenstorrent

2 nodes of 16 SN40L RDUs

9 Groq nodes,
8 GroqChip/node (TSPs)

Cerebras CS-3 – 4 WSE

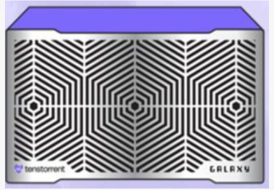


32 Wormhole GU

HPC

- Cerebras
- Tenstorrent

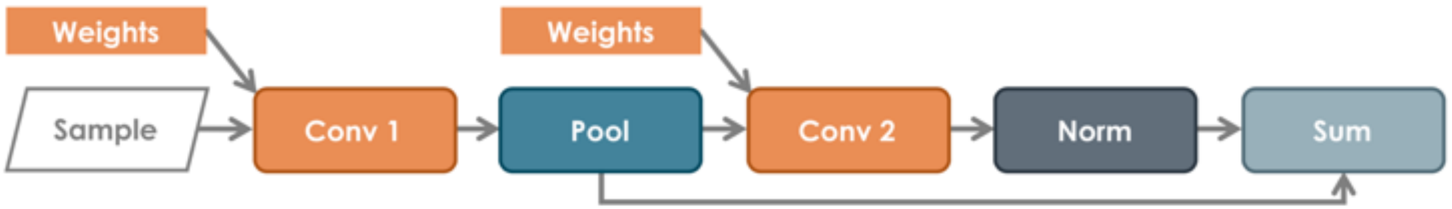
Cerebras CSL



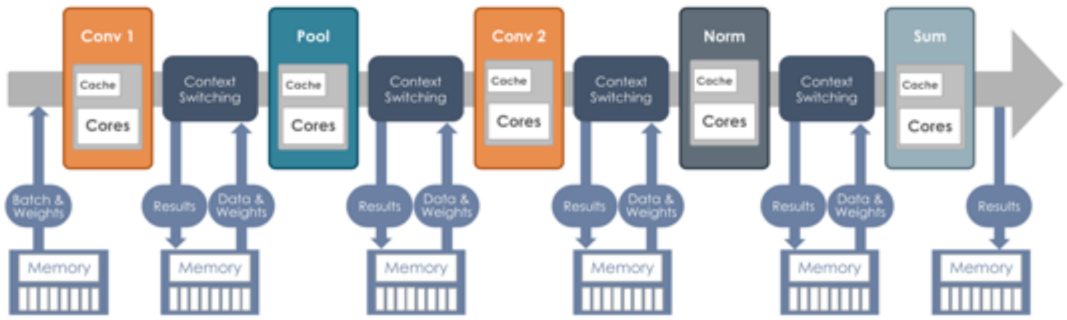
32 Wormhole GU

	Cerebras CS3	SambaNova Cardinal SN30 / SN40L	Groq GroqRack	GraphCore GC200 IPU	NVIDIA A100
Compute Units	900,000 Cores	640/1040 PCUs	5120 vector ALUs	1472 IPUs	6912 Cuda Cores
On-Chip Memory	44 GB SRAM, MemoryX	300/520MB Sram 0/64 GB HBM 1/1.5TB DDR	230MB L1	900MB L1	192KB L1 40MB L2 40-80GB
Process	7nm	7nm	7 nm	7nm	7nm
System Size	4 Nodes Memory-X and Swarm-X	8 nodes (8 RDUs per node)/2 nodes (16 RDUs per node)	9 nodes (8 cards per node)	4 nodes (16 cards per node)	Several systems
Estimated Performance of a card (TFlops)	>5780 (FP16)	>660/638 (BF16)	>250 (FP16) >1000 (INT8)	>250 (FP16)	312 (FP16), 156 (FP32)
Software Stack Support	Pytorch	SambaFlow, Pytorch	GroqAPI, ONNX	Tensorflow, Pytorch, PopArt	Tensorflow, Pytorch, etc
Interconnect	Ethernet-based	Ethernet-based	RealScale TM	IPU Link	NVLink

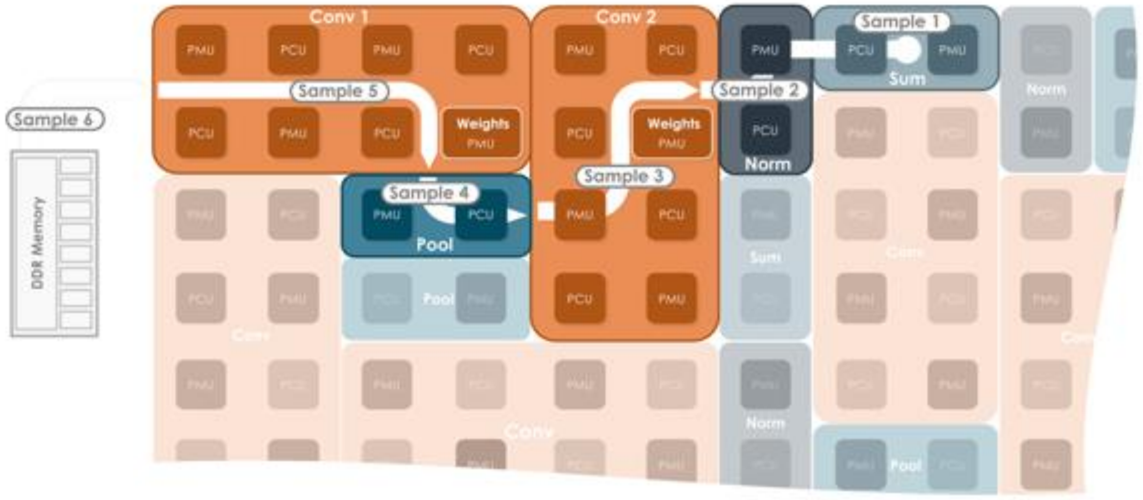
Dataflow Architectures



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

Dataflow hardware architecture

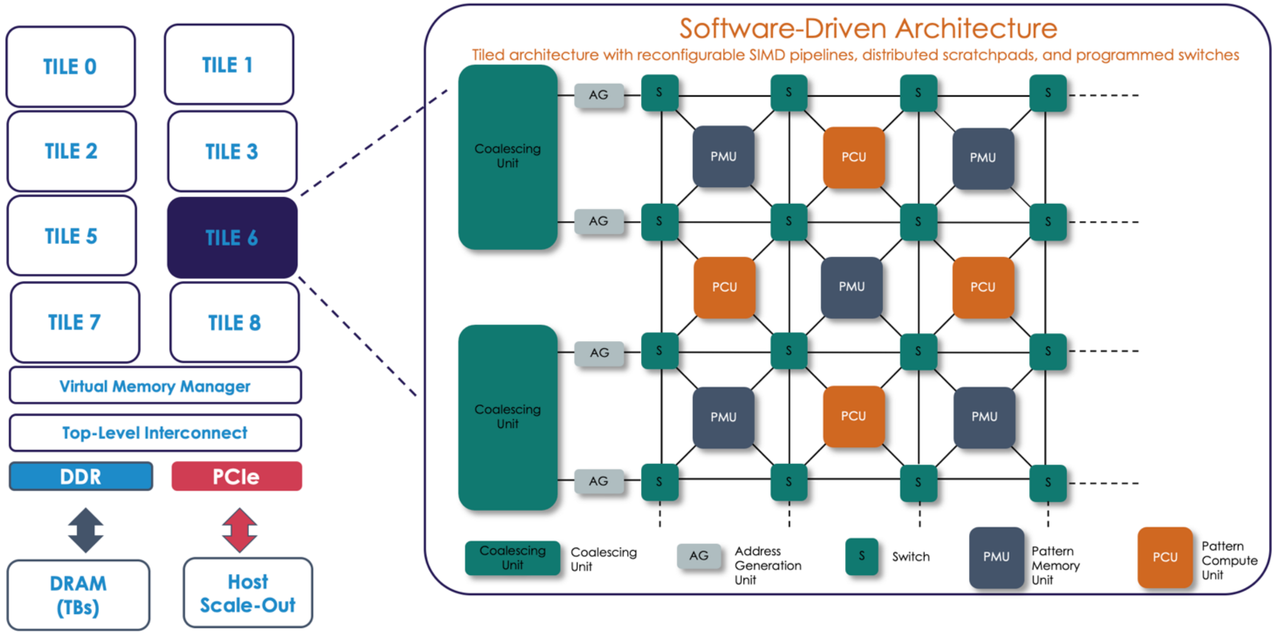


Image Courtesy: SambaNova

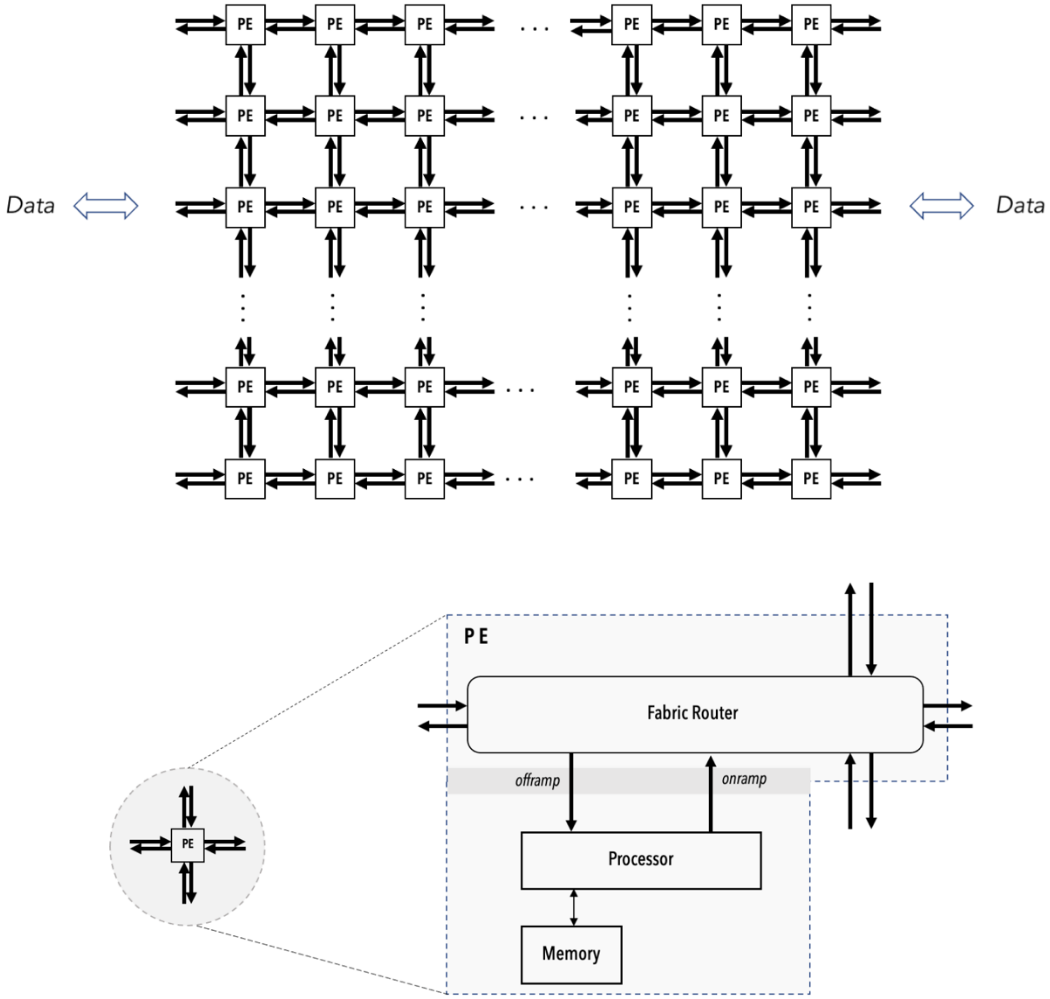


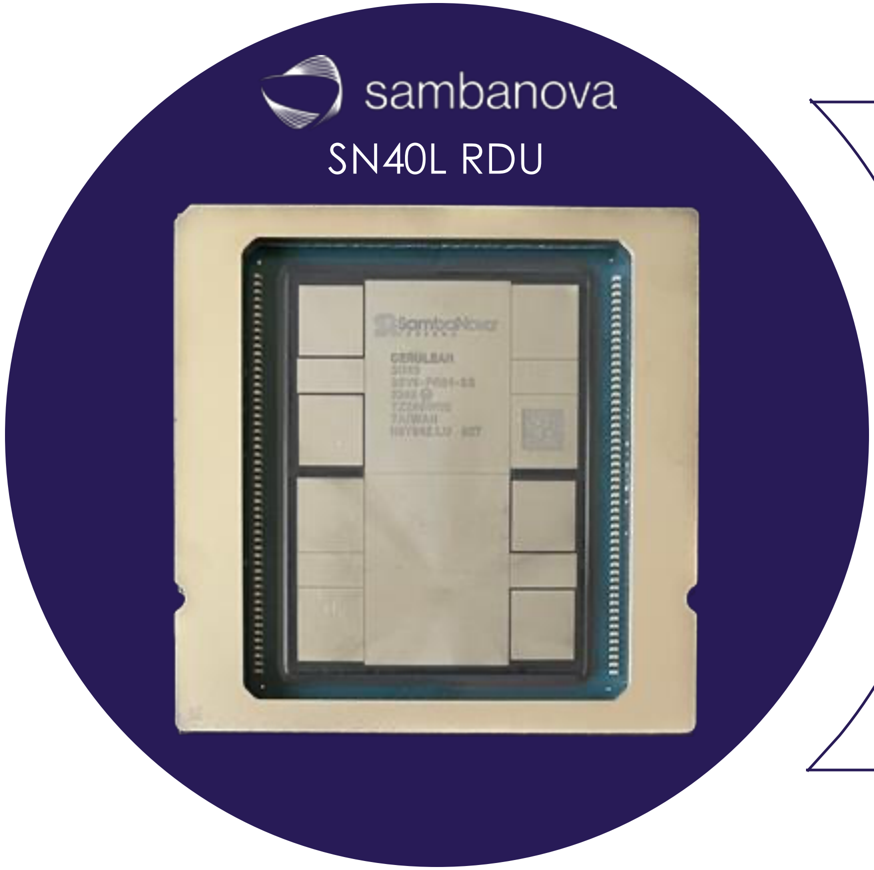
Image coutesy: Cerebras

- Interleaving of compute and memory units
- Routing data through the compute elements

Argonne NATIONAL LABORATORY

SambaNova SN40L

Reconfigurable Dataflow Unit (RDU)
Native multi-tenancy support with fast model switching
Ideal for production inference, multi-tenancy, agentic workflows



- 3-tier Dataflow Memory
- 520 MB On-Chip SRAM Memory → Very fast memory for high speed infere with caching
- 64 GB High Bandwidth Memory → Switch between models in as little as 2 milliseconds
- 1.5 TB High Capacity DDR Memory → Hold large number of models in memory

Image source:, SambaNova

SN40L-16: Node Details

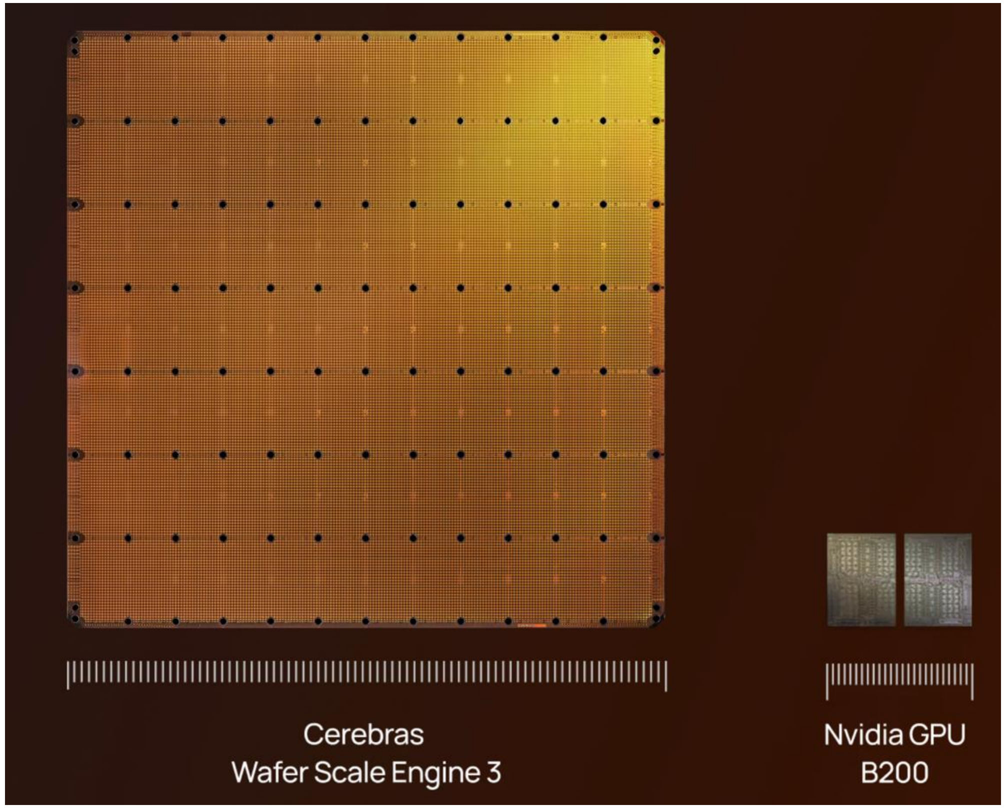
Total of 16 RDUs per node

- SN40L-16
  - 1x SN40L-H (middle)
  - 8x SN40L-2 (aka XRDU) (4x above host, 4x below host, each with 2 RDUs)
  - RDUs connect all-to-all and with Host
  - 12TB of memory w/ 64GB DIMMs
  - 1TB of high-bandwidth memory (HBM)

- 1x SN40-H Host (Server)
  - Standard Linux-based OS server
    - 2x AMD EPYC 64-core CPUs
    - 10TB usable NVMe storage
    - 1TB of DDR Memory
  - Connects to all 16 RDUs

Image source:, SambaNova

Cerebras Wafer Scale Engine (CS-3)



- 900,000 compute cores
- 44G on-chip SRAM
- 21 PB/s Memory bandwidth

- Decoupled compute and memory
- Data-parallel implementation
- External memory can be added independently of compute,

allowing for massive model sizes

Image source:, Cerebras

Cerebras CS-3 cluster

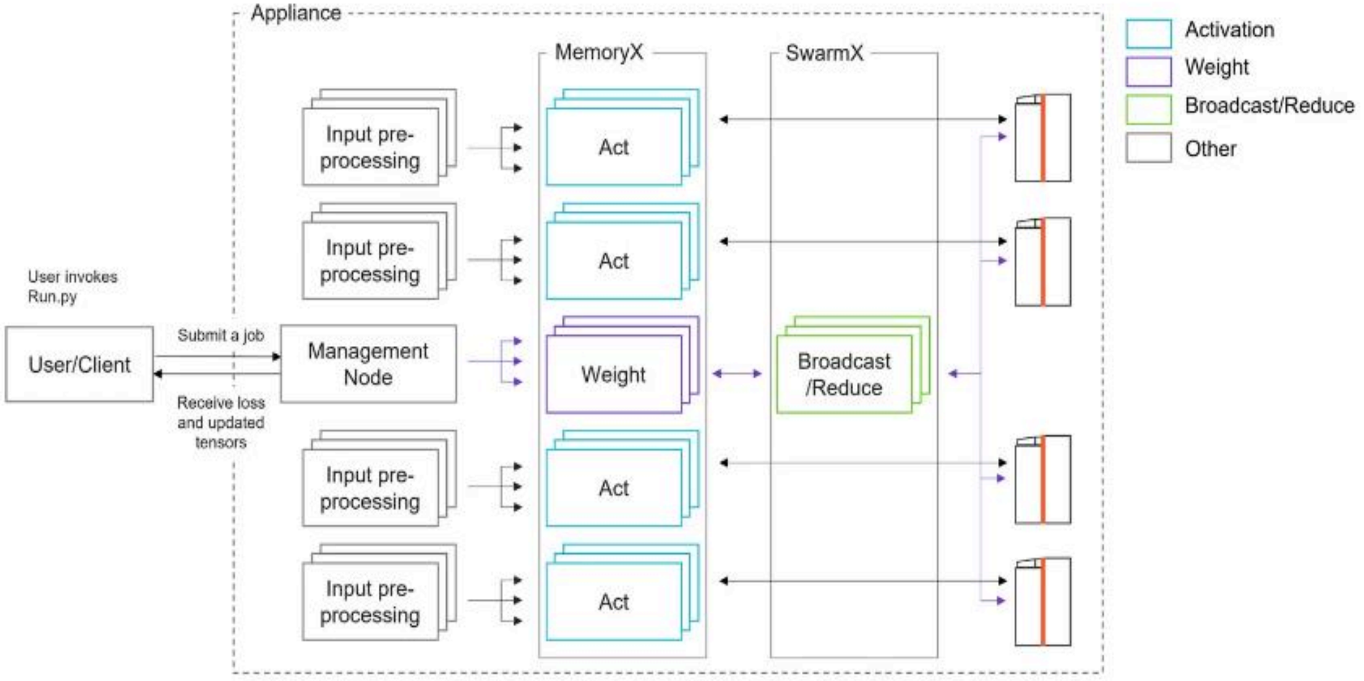
Input preprocessing servers stream training data

MemoryX - Stores and streams model’s weights

SwarmX – weight broadcasts and gradient across multiple CS-3s

Compilation (maps graph to kernels)
Execution (training)

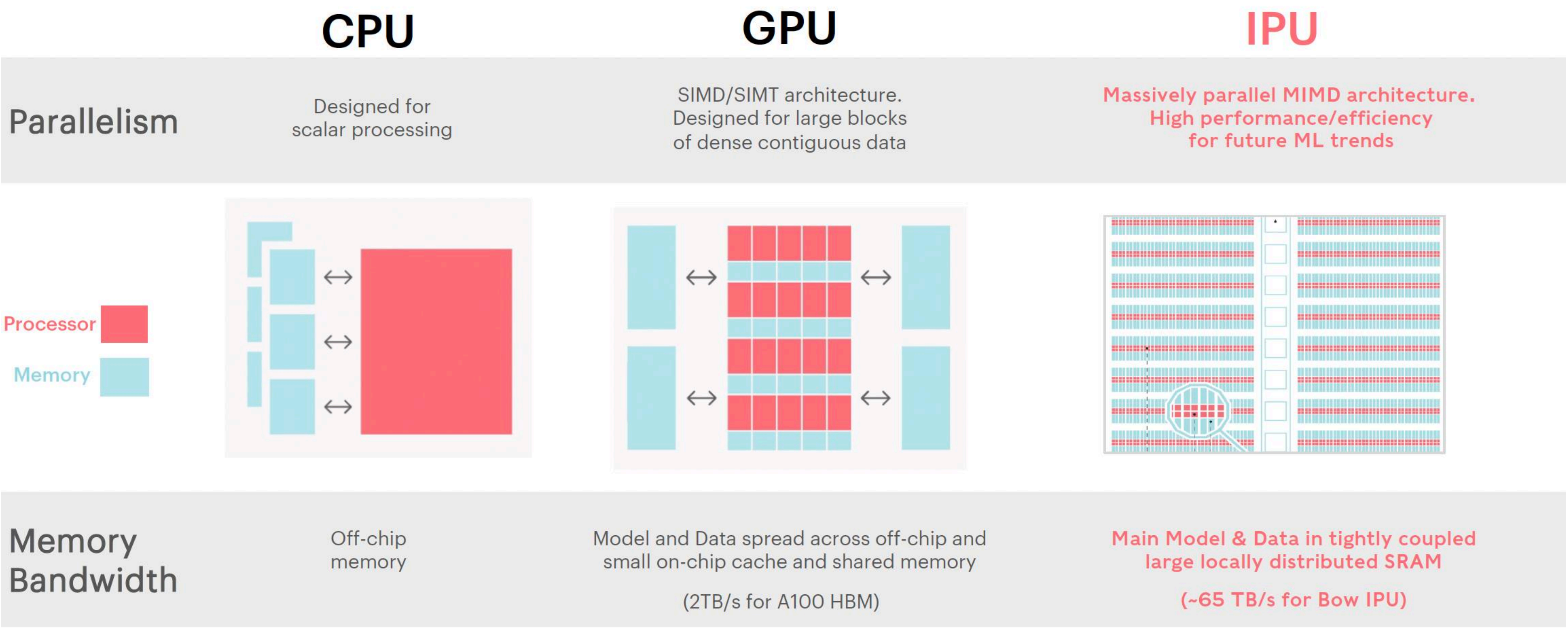
Weight Streaming (training) Vs Pipeline (Inference)



Argonne NATIONAL LABORATORY

Graphcore BowPod64

Intelligence Processing Unit



BowPod64 configuration: 64 IPUs

Image source: Graphcore

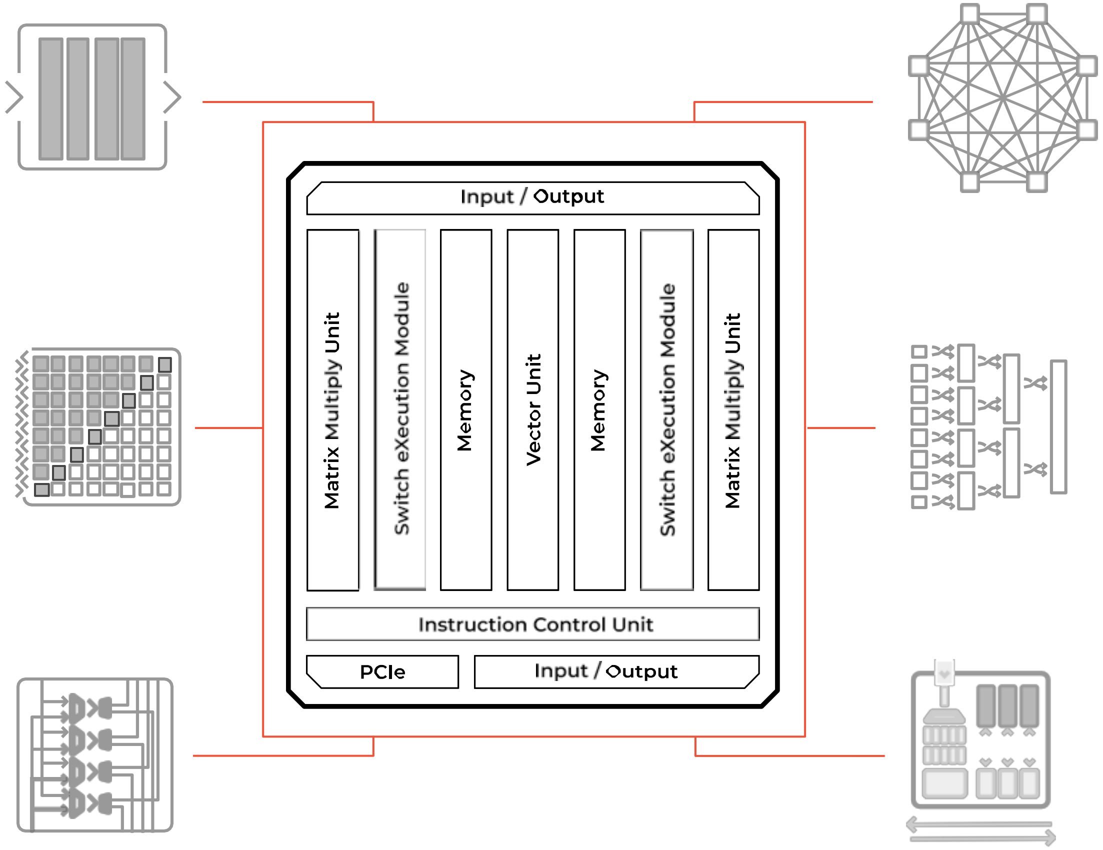
Groq

GroqRack configuration: 72 Groqchips

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



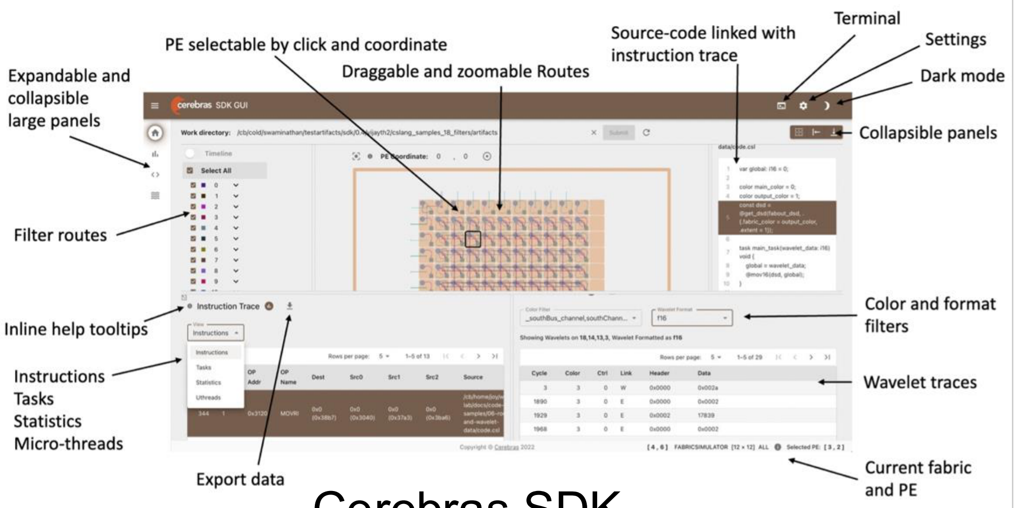
Networking
480 GB/s bandwidth
Extensible network scalability
Multiple topologies

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

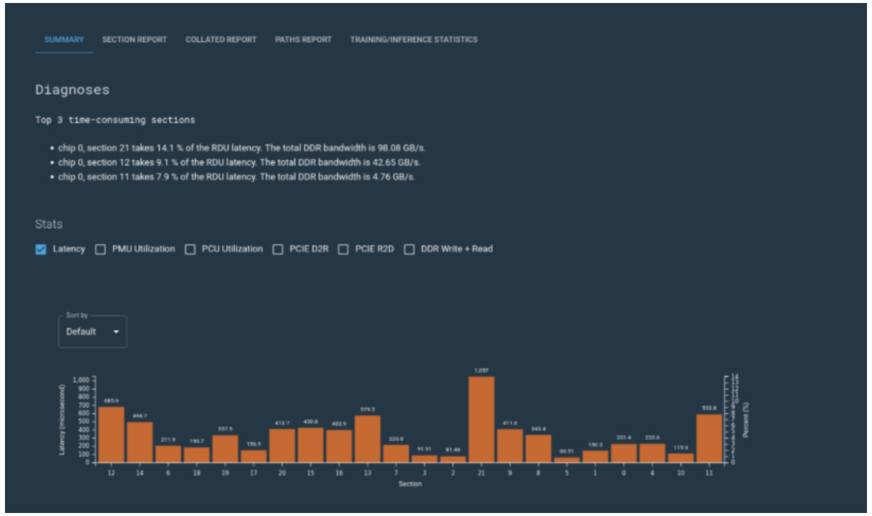
Instruction Control
Multiple instruction queues for instruction parallelism

Courtesy: Groq

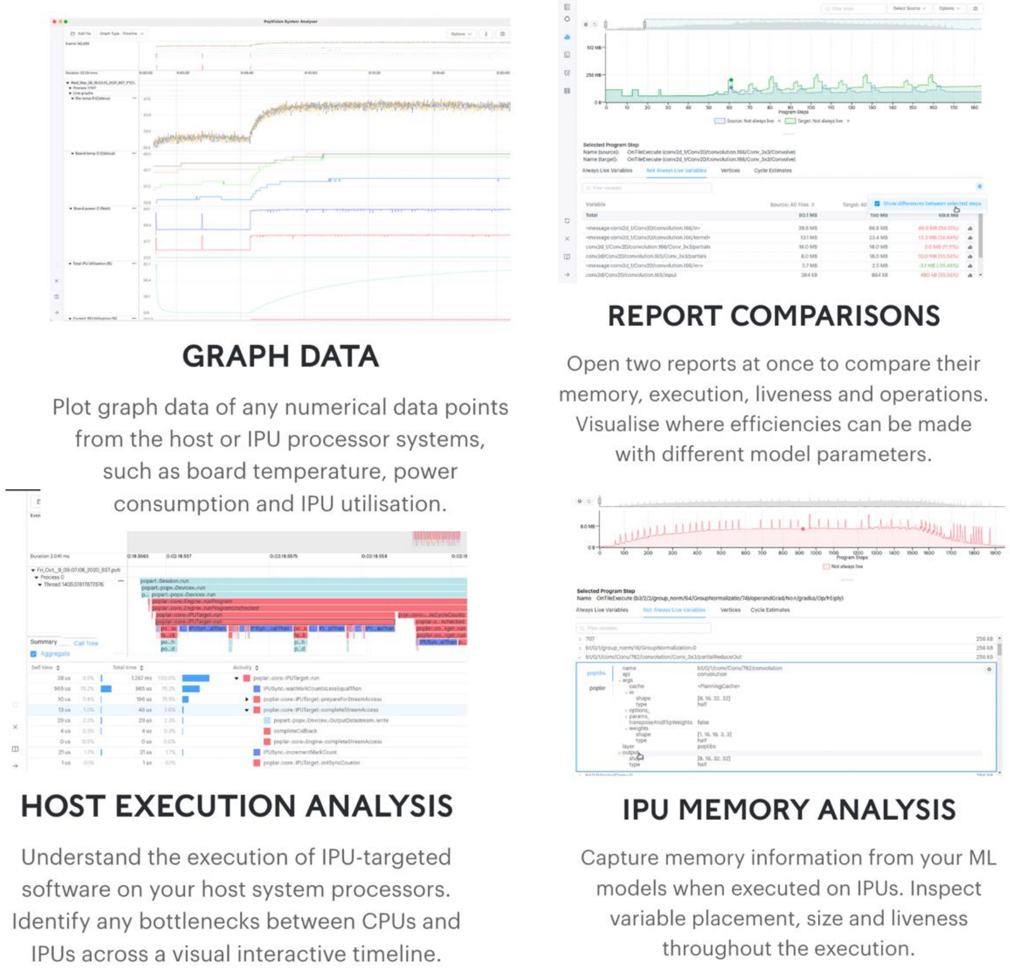
Tools on AI Accelerators



Cerebras SDK

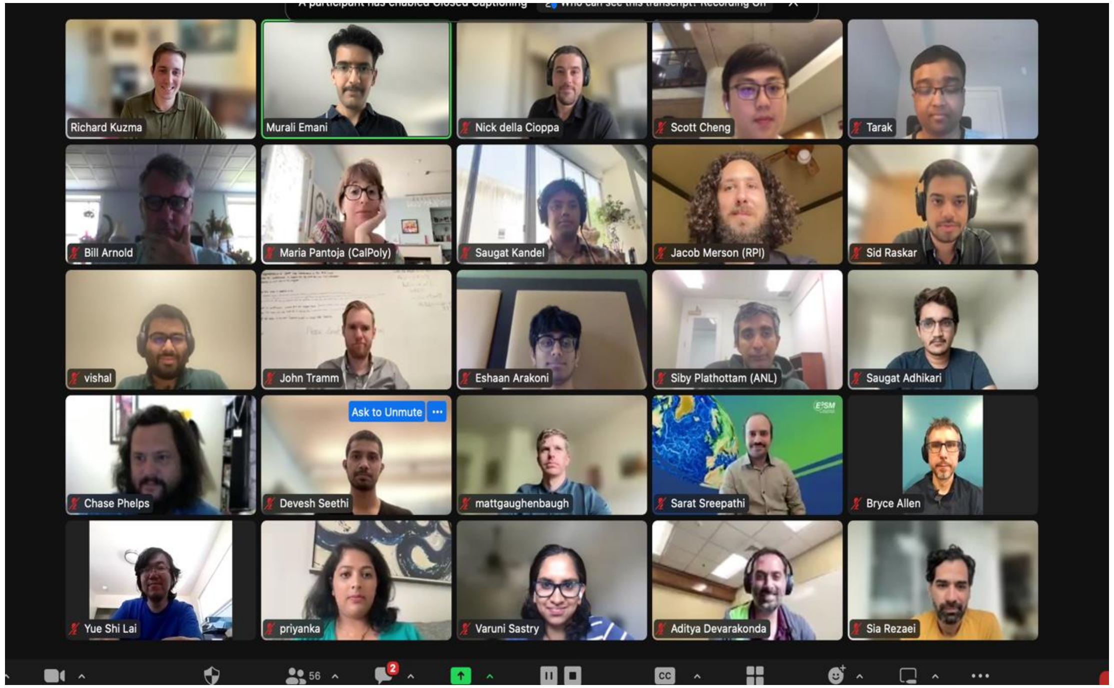


SambaTune on SambaNova



PopVision on GraphCore

AI Testbed Community Engagement



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 Intel Gaudi systems along with architectural features and details of their software stacks. Through hands-on exercises, attendees will gain practical experience in refactoring code and running models on these systems, focusing on use cases of pre-training and fine-tuning open-source large language models (LLMs) and deploying AI inference solutions relevant to scientific contexts. The tutorial will provide 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, 16 November 2025
1:30pm - 5:00pm CST

Location: 121

Registration Categories:
TUT

NEXT PRESENTATION >

- AI training workshops
https://www.alcf.anl.gov/ai-testbed-training-workshops
- ATPESC Training
- Introduction to AI-driven Science on Supercomputers

Upcoming tutorial at SC25 on Programming Novel AI accelerators for Scientific Computing

Nov 16, 2025

Getting Started on ALCF AI Testbed

Available for Allocations

- Cerebras CS-3,
- SambaNova Datascale SN30,
- GroqRack
- Graphcore Bow Pod64
- Sambanova Inference – Metis SN40L (Available for all ALCF users via Inference Service Endpoints)

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/

Argonne NATIONAL LABORATORY

NEWS EVENTS PEOPLE CAREERS MyALCF Search

Argonne Leadership Computing Facility

ALCF Resources | Science and Engineering | Community and Outreach | About | Support

Machines > AI Tes...

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

ALCF User Guides

- Getting Started
- Contribute to User Guides

User Support

- Submit a Ticket
- Get Help & Connect

Machines

- Aurora
- AI Testbed
  - Cerebras
  - Graphcore
  - Groq
  - SambaNova
  - Data Management
- Crux
- Polaris
- Sophia

Running Jobs with PBS

- Example Job Scripts
- Machine Reservations

Data Management

- File System & Storage
- Transfer & Sharing

Services

- Inference Endpoints
- JupyterHub
- Continuous Integration

ALCF AI Testbed

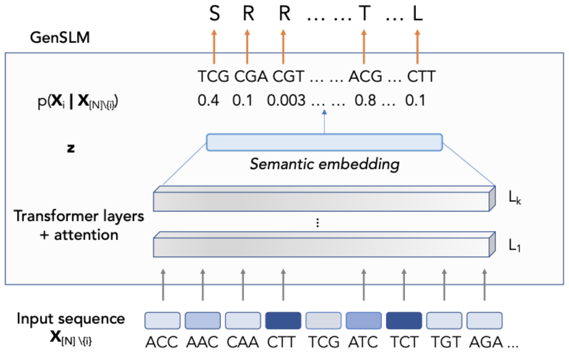
The ALCF AI Testbed houses some of the most advanced AI accelerators for scientific research.
The goal of the testbed is to enable explorations into next-generation machine learning applications and workloads, enabling the ALCF and its user community to help define the role of AI accelerators in scientific computing and how to best integrate such technologies with supercomputing resources.
The AI accelerators complement the ALCF's current and next-generation supercomputers to provide a state-of-the-art computing environment that supports pioneering research at the intersection of AI, big data, and high performance computing (HPC).
The platforms are equipped with architectural features that support AI and data-centric workloads, making them well suited for research tasks involving the growing deluge of scientific data produced by powerful tools, such as supercomputers, light sources, telescopes, particle accelerators, and sensors. In addition, the testbed will allow researchers to explore novel workflows that combine AI methods with simulation and experimental science to accelerate the pace of discovery.

Table of contents

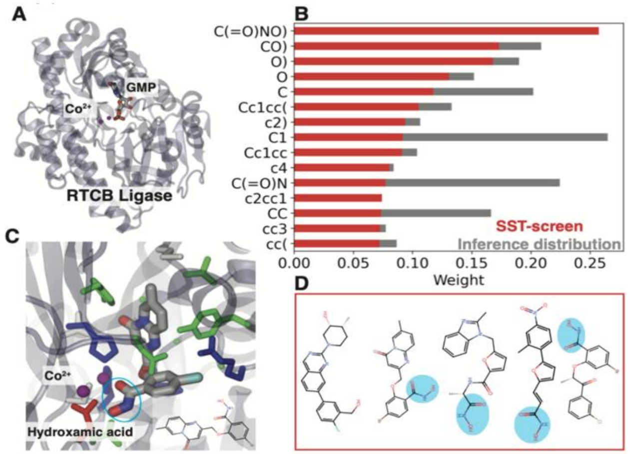
Does this doc need an update? Open an Issue on GitHub

AI Based Models

Text Based Models

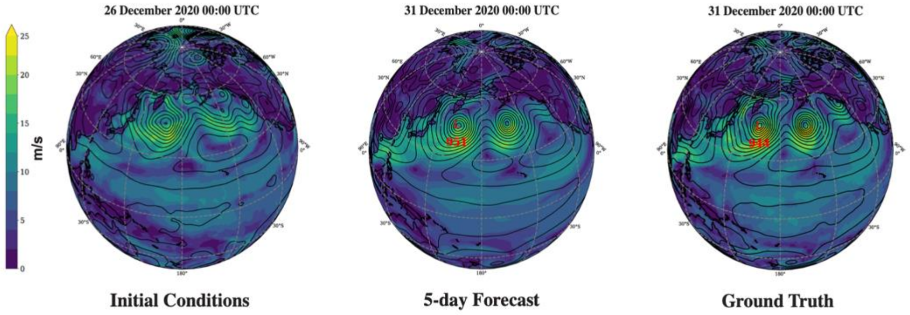


VOC detection

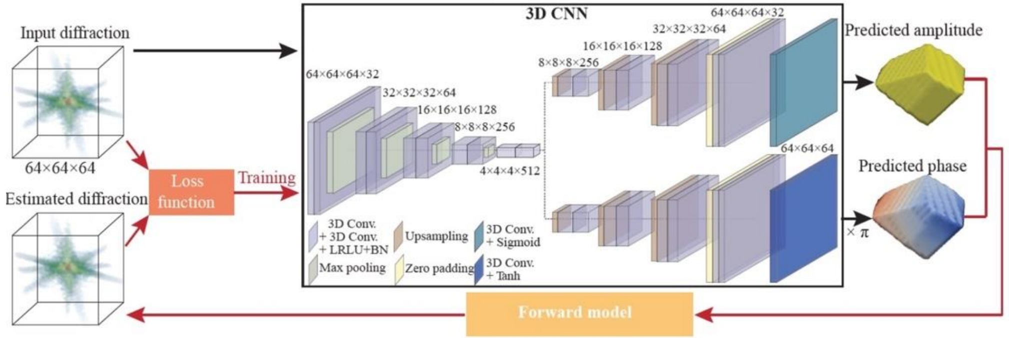


Drug and Molecular discovery

Vision Models



Stormer – Weather Forecasting



Diffraction Imaging

Cosmology and more ..

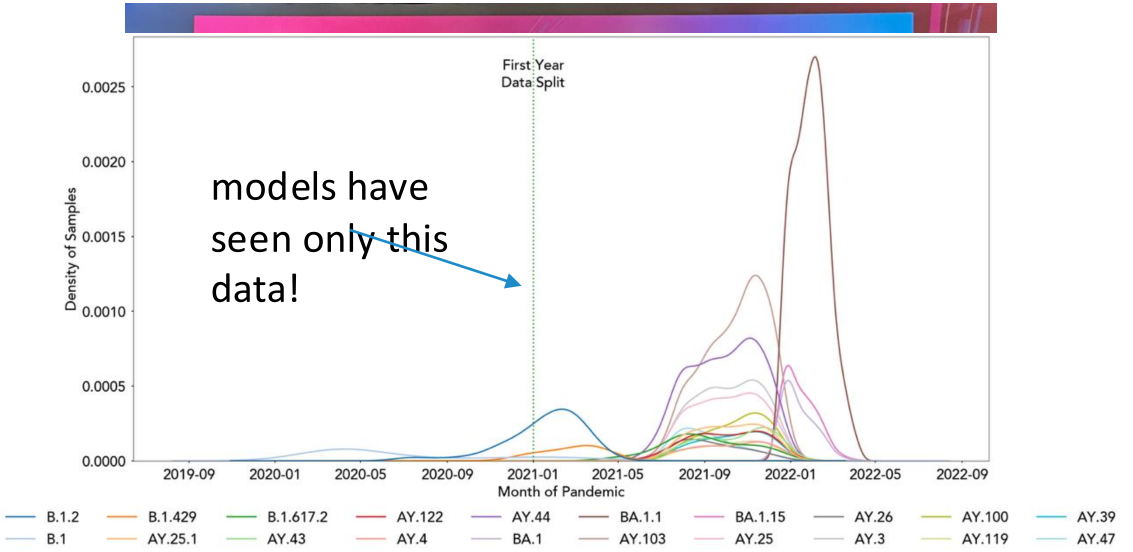
Genome-scale Language Models (GenSLMs)

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

Argonne NATIONAL LABORATORY

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

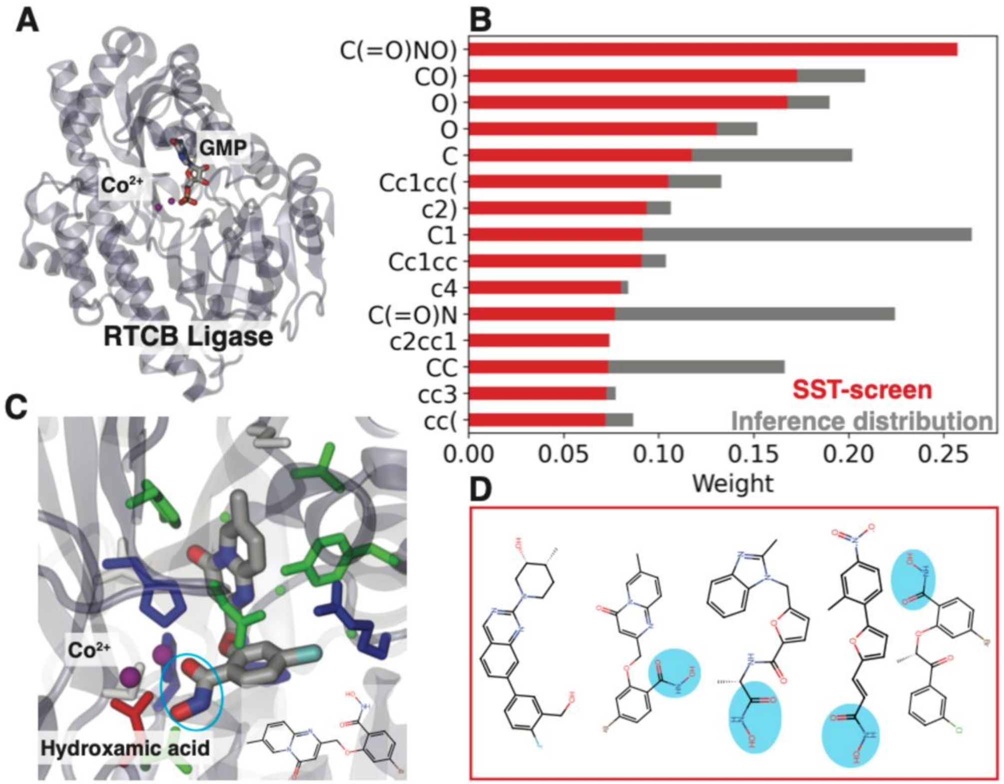
System	Number of Devices	Throughput (tokens/sec)	Improvement	Energy Efficiency
NVIDIA A100	8	1150	1.0	1.0
SambaNova SN30	8	9795	8.5	5.6
Cerebras CS-2	1	29061	25	6.5

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

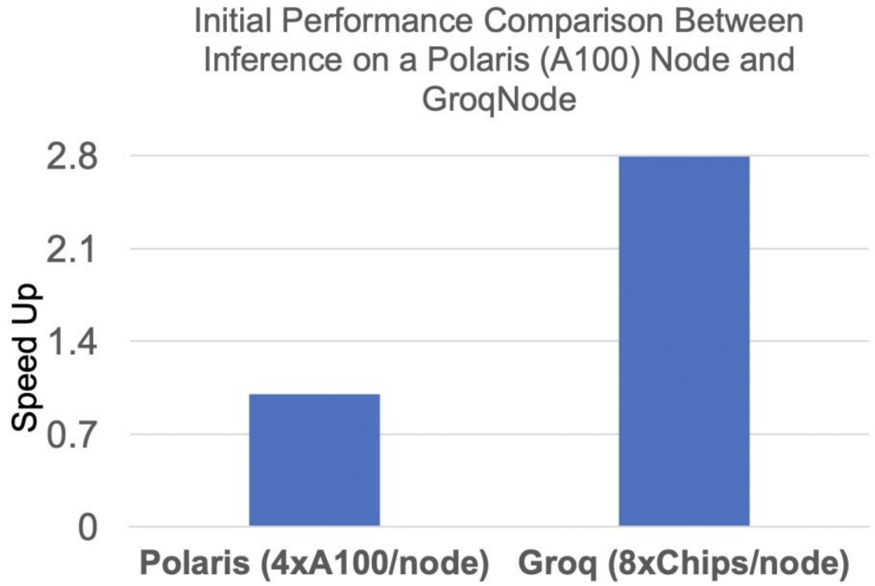
Accelerating Drug Design and Discovery with Machine Learning

Application code: Simple SMILES Transformer



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

Archit Vasan*, Ozan Gokdemir*†, Alexander Brace*†, Arvind Ramanathan*†, Thomas Brettin*, Rick Stevens*†, Venkatram Vishwanath*



Courtesy: Archit Vasan

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

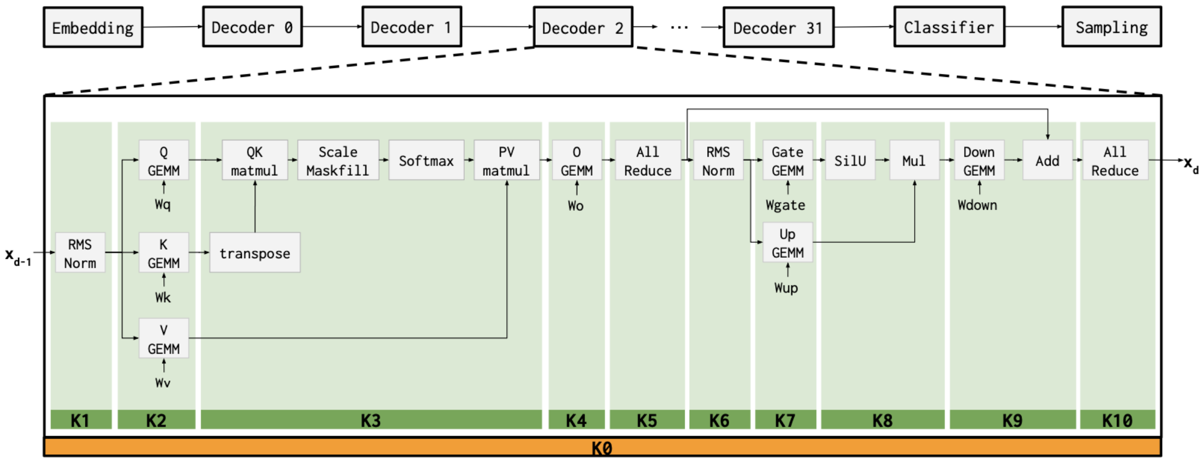
Bert based encoder model to identify compounds with high binding affinity directly on the SMILES string input.

Weather Forecasting

Goal: Achieve faster weather predictions at large scale rollouts 0.25° ERA5 data.

Approach: Sambanova's large memory capacity encourages training on high dimensional data (large context lengths).

Dataflow architecture with kernel looping reduces latency.




```
// DGX H100
// Kernel Call Schedule

x_in = embedding()
for decoder in range(0, 32):
  tmp_k1 = K1(x_in)
  tmp_k2 = K2(tmp_k1)
  tmp_k3 = K3(tmp_k2.q, tmp_k2.k, tmp_k2.v)
  tmp_k4 = K4(tmp_k3)
  tmp_k5 = K5(tmp_k4)
  tmp_k6 = K6(tmp_k5)
  tmp_k7 = K7(tmp_k6)
  tmp_k8 = K8(tmp_k7.gate, tmp_k7.up)
  tmp_k9 = K9(tmp_k5, tmp_k8)
  x_out  = K10(tmp_k9)
  x_in   = x_out
cls_out = classifier(x_in)
out     = sampling(cls_out)
```

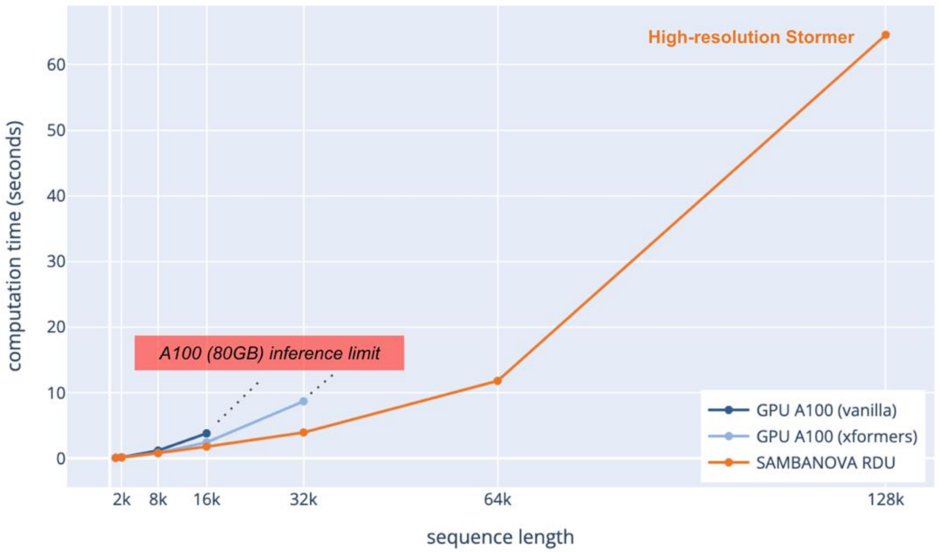
```
// SN40L-8
// Kernel Call Schedule

x_in = embedding()
for decoder in range(0, 32):
  x_out = K0(x_in)
  x_in  = x_out

cls_out = classifier(x_in)
out     = sampling(cls_out)
```

```
// SN40L-8 + Kernel looping
// Kernel Call Schedule

x_in    = embedding()
x_out   = all_decoders_nosync(x_in)
cls_out = classifier(x_in)
out     = sampling(cls_out)
```



Argonne NATIONAL LABORATORY

Diffraction Imaging

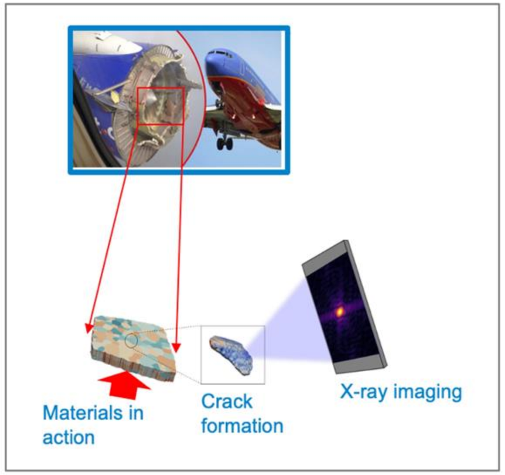
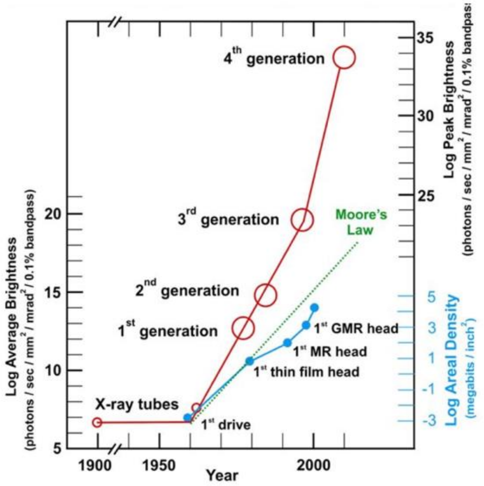
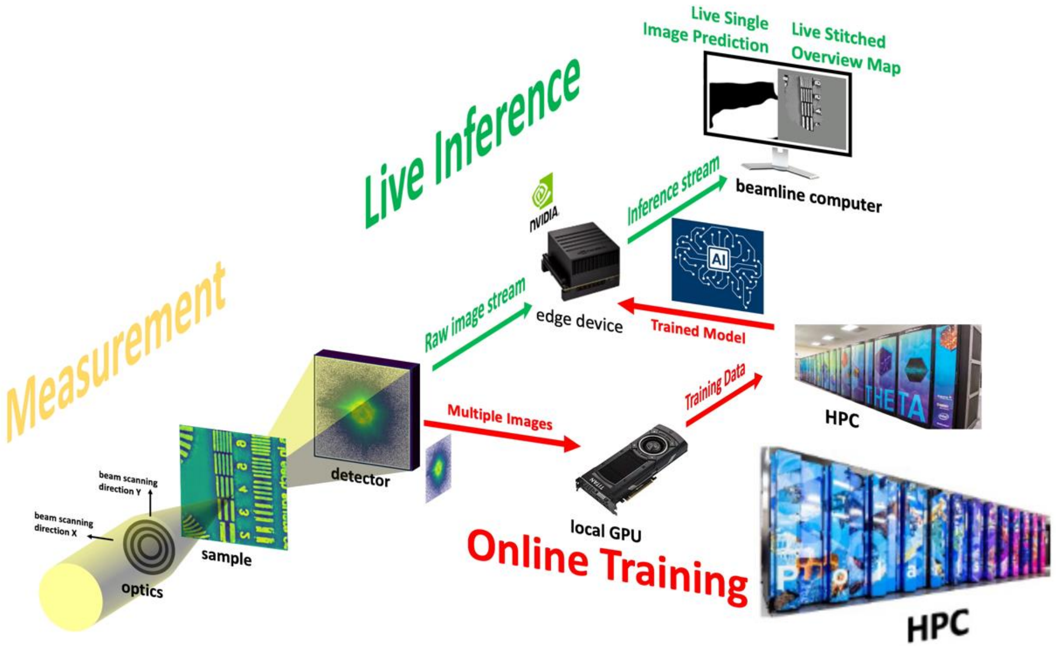


Image adapted from: Jon Almer, Stephan Hruszkewycz et al., ANL



http://archive.synchrotron.org.au/images/AOF2017/Boland---AOF---Future-light-sources-2017-05-29.pdf



A. V. Babu, T. Zhou, S. Kandel, T. Bicer, Z. Liu, W. Judge, D. Ching, Y. Jiang, S. Veseli, S. Henke, R. Chard, Y. Yao, E. Sirazitdinova, G. Gupta, M. V. Holt, I.T. Foster, A. Miceli and M. J. Cherukara, "Deep learning at the edge enables real-time, streaming ptychography", *Nature Communications*, 14, 7059

- Real time feedback and reconstruction time in order of msec.
- APS-U will have 10-100x increase in data rates.
- AI-steered experiments to target 10^12 voxels.

Each technique presents a unique challenge

BCDI

- Today: ~GB (memory for phasing)
  - 256-512 cubed arrays
  - ~ 5 nm
- APS-U: ~TB
  - 2560-5120 cubed 3D FFTs
    - Or equivalent NN network
  - ~ 5 A

Ptycho[1]

- > GB/s data rates
- > PFLOPS of peak computing power to keep up
- Today: ~5 Ptycho beamlines
- APS-U: ~10 Ptycho beamlines

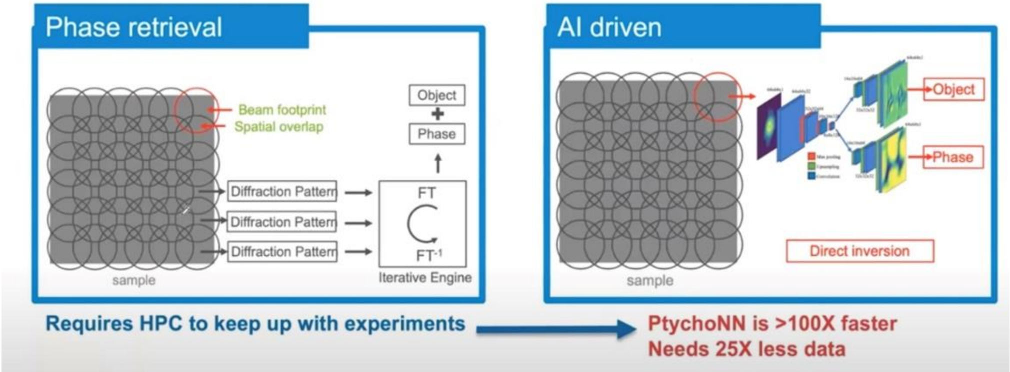
[1] APS Needs and Progress on Machine Learning and Artificial Intelligence Applications

Accelerators for Imaging

- Larger compute fabric and memory footprint enables better throughput and large resolution imaging with almost double the power efficiency.
- Leveraged Sambanova SN30 hardware to bring up the BCDI AI workflow for native resolution upto 256^3 voxels, avoiding the need for downsampling.
- Used Cerebras CS-2 for continual pre-training of PtychoNN model.
- Challenges : FFT and vision support, Compile times, Ease of portability.
- Focused efforts on developing AI methods and frameworks for large resolution APS-U data.

https://cerebras.ai/blog/cerebras-cs-3-vs-nvidia-b200-2024-ai-accelerators-compared

Spec	CS-3 / B200	CS-3 / DGX B200	CS-3 / NVL72
FP16 PFLOPs	28.4	3.5	0.3
Memory (GB)	6,250.0	781.3	88.9
NVLink | Fabric Bandwidth (TB/s)	14,861	1,858	206
Power (Watts)	23.0	1.6	0.2
PFLOPs / W	1.2	2.2	1.8



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
    - energy efficient
    - low latency critical applications
    - Off the shelf models for inference

- Room for improvement exists
    - Porting efforts and compilation times
    - Coverage of DL frameworks, support for performance analysis tools, debuggers

- Limited capability to support low-level HPC kernels
    - Work in progress to improve coverage

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.

- Venkat Vishwanath, Murali Emani, Varuni Sastry, Michael Papka, William Arnold, 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, SambaNova. There are ongoing engagements with other vendors.