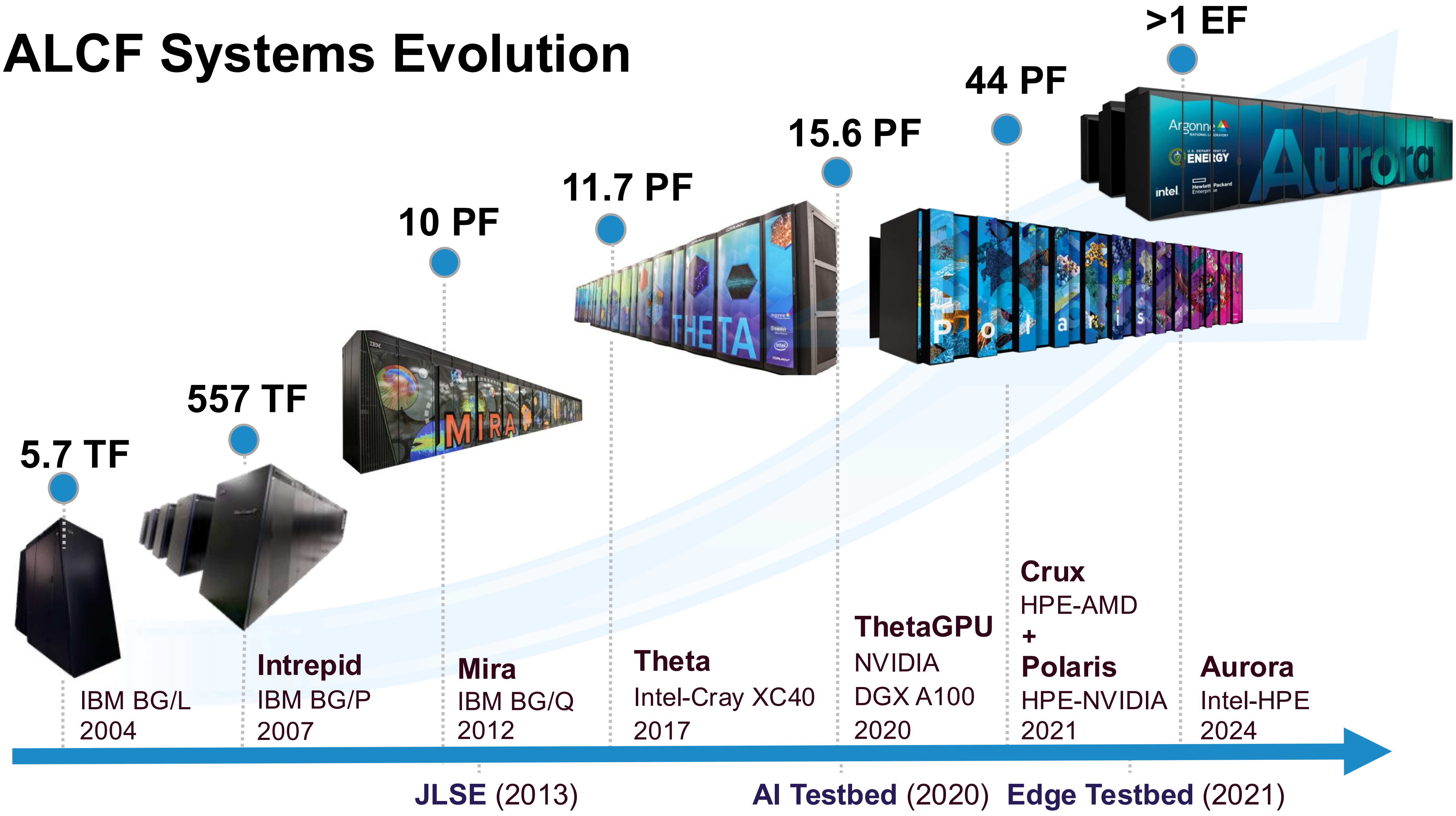
ALCF AI Testbeds

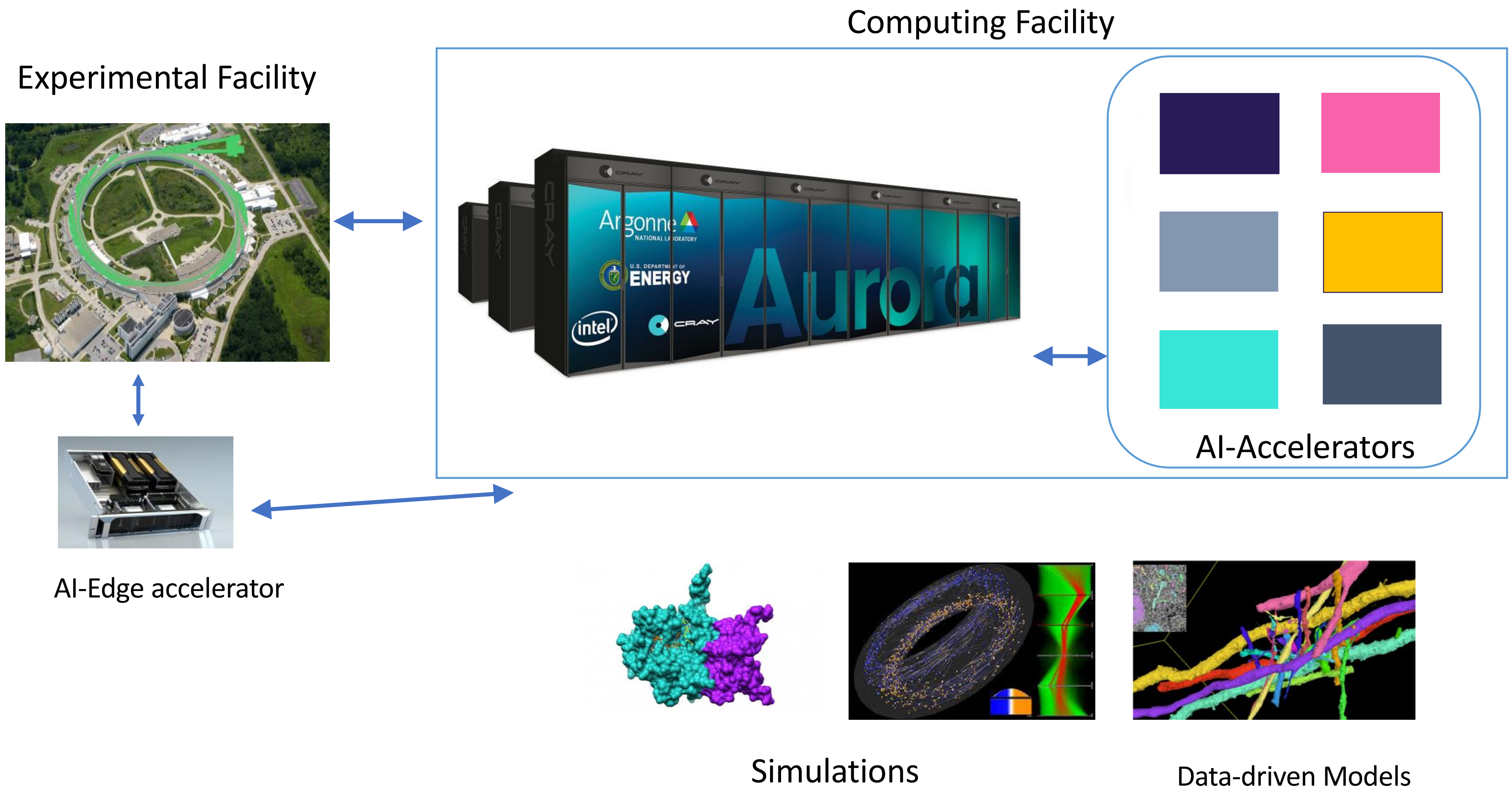
Varuni Sastry
Argonne Leadership Computing Facility
vsastry@anl.gov

ALCF HandsOn Workshop
Oct 8, 2025

ALCF Systems Evolution



Integrating AI Systems in Facilities



ALCF AI Testbeds

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

Cerebras (CS-3)

SambaNova SN30/SN40L

Groq

Tenstorrent

Graphcore

Habana

- 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

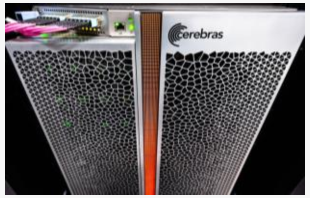
ALCF AI Testbed

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



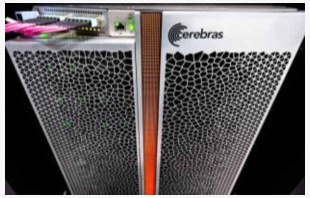
Cerebras CS-3 – 4 WSE

Inference

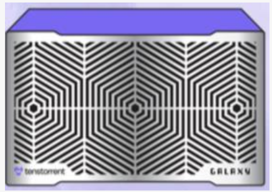
- SN40L – Metis
- Groq
- Cerebras
- Tenstorrent

2 nodes of 16 SL40L RDUs

9 Groq nodes,
8 GroqChip/node (TSPs)



Cerebras CS-3 – 4 WSE

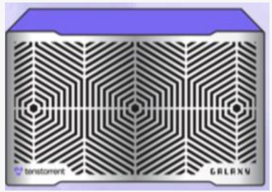


32 Wormhole GU

HPC

- Cerebras
- Tenstorrent

Cerebras CSL



32 Wormhole GU

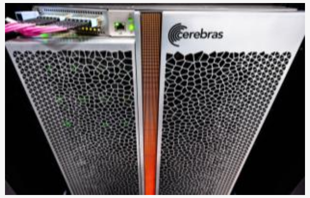
ALCF AI Testbed

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



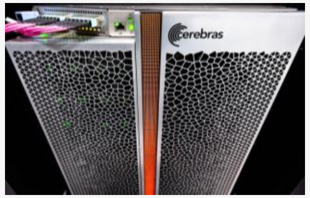
Cerebras CS-3 – 4 WSE

Inference

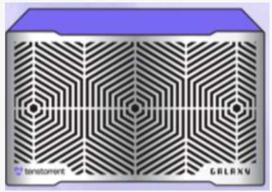
- SN40L – Metis
- Groq
- Cerebras
- Tenstorrent

2 nodes of 16 SL40L RDUs

9 Groq nodes,
8 GroqChip/node (TSPs)



Cerebras CS-3 – 4 WSE

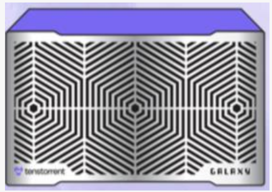


32 Wormhole GU

HPC

- Cerebras
- Tenstorrent

Cerebras CSL



32 Wormhole GU

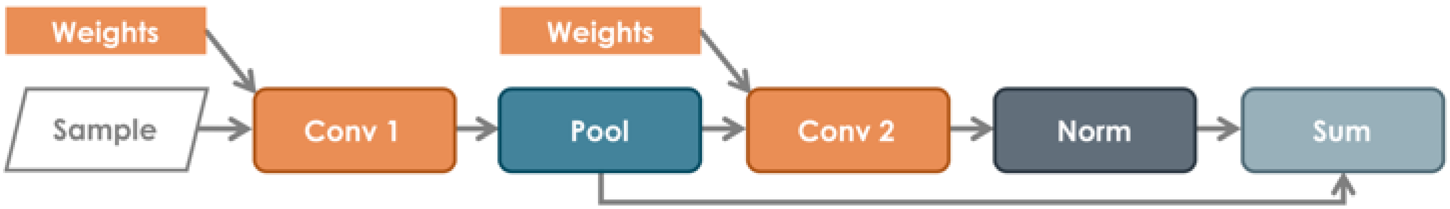
Coming Soon !!

	Cerebras CS3	SambaNova Cardinal SN30 / SN40L	Groq GroqRack	GraphCore GC200 IPU	Habana Gaudi1	NVIDIA A100
Compute Units	900,000 Cores	640/1040 PCUs	5120 vector ALUs	1472 IPUs	8 TPC + GEMM engine	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	24 MB L1 32GB	192KB L1 40MB L2 40-80GB
Process	7nm	7nm	7 nm	7nm	16nm	7nm
System Size	4 Nodes 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)	>660/638 (BF16)	>250 (FP16) >1000 (INT8)	>250 (FP16)	>150 (FP16)	312 (FP16), 156 (FP32)
Software Stack Support	Pytorch	SambaFlow, Pytorch	GroqAPI, ONNX	Tensorflow, Pytorch, PopArt	Synapse AI, TensorFlow and PyTorch	Tensorflow, Pytorch, etc
Interconnect	Ethernet-based	Ethernet-based	RealScale ™	IPU Link	Ethernet-based	NVLink

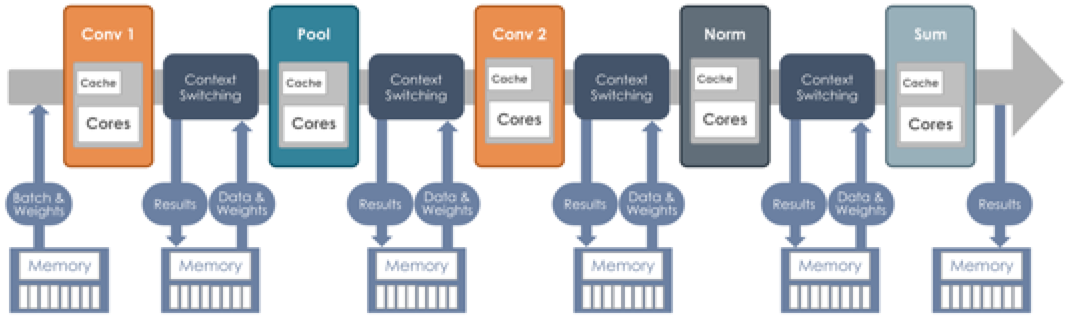
HPC Software ecosystem on AI Accelerators

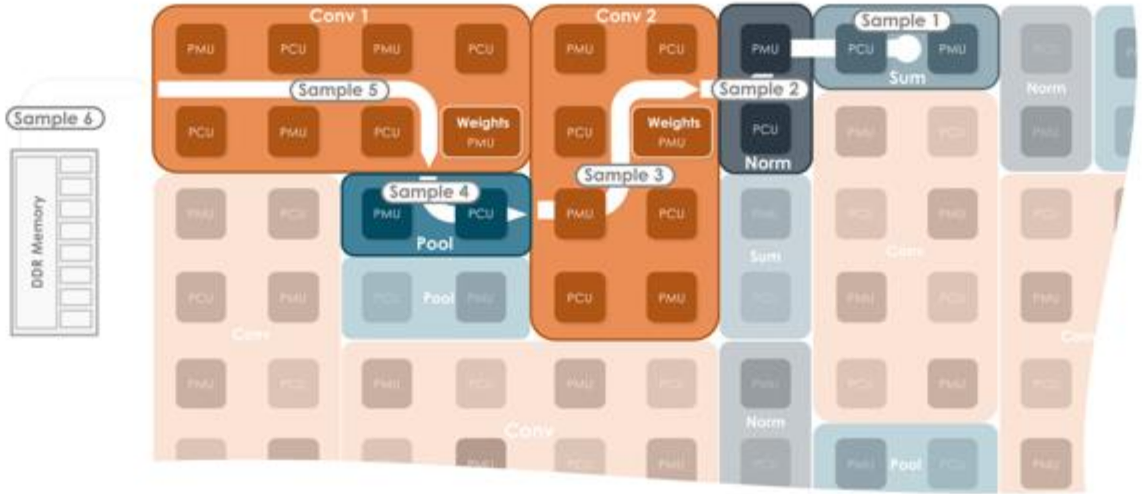
cerebras	• Cerebras Software Language
SambaNova SYSTEMS	• Redwood (C/C++)
GRAPHCORE	• Poplar C/C++ API • BSP
groq	• Groq Runtime API • C/C++
habana An Intel Company	• Habana TPC • C/C++
NEXTSILICON	• OpenMP Pragmas • C/C++

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

Argonne NATIONAL LABORATORY

Dataflow hardware architecture

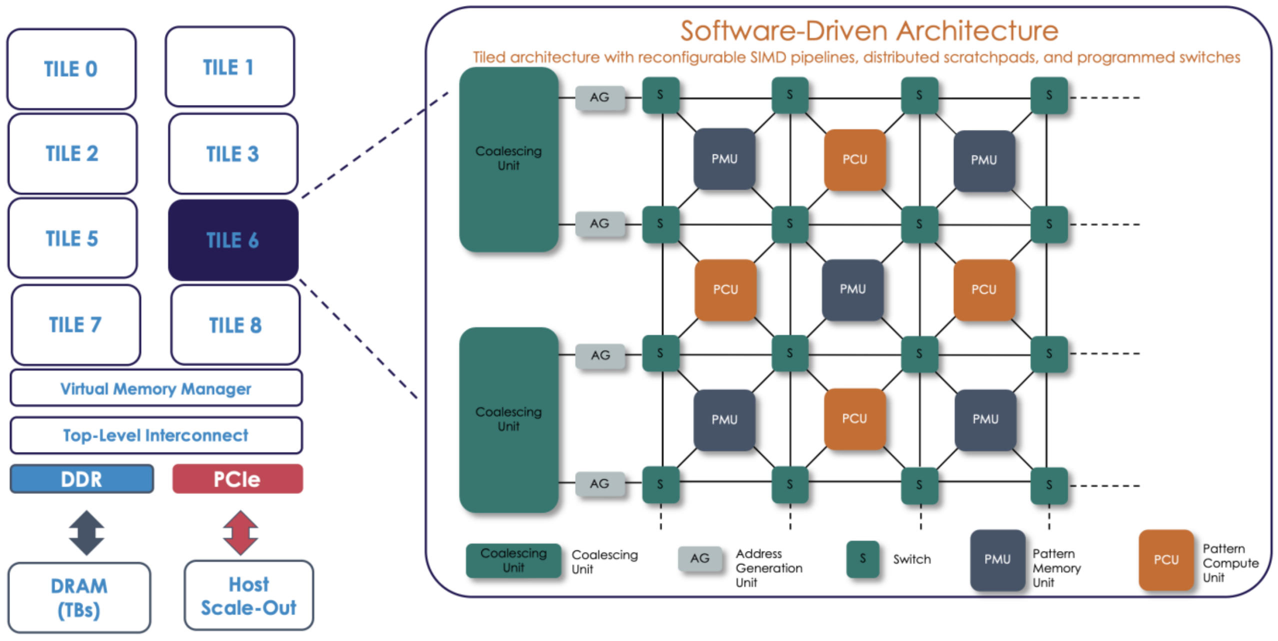


Image Courtesy: SambaNova

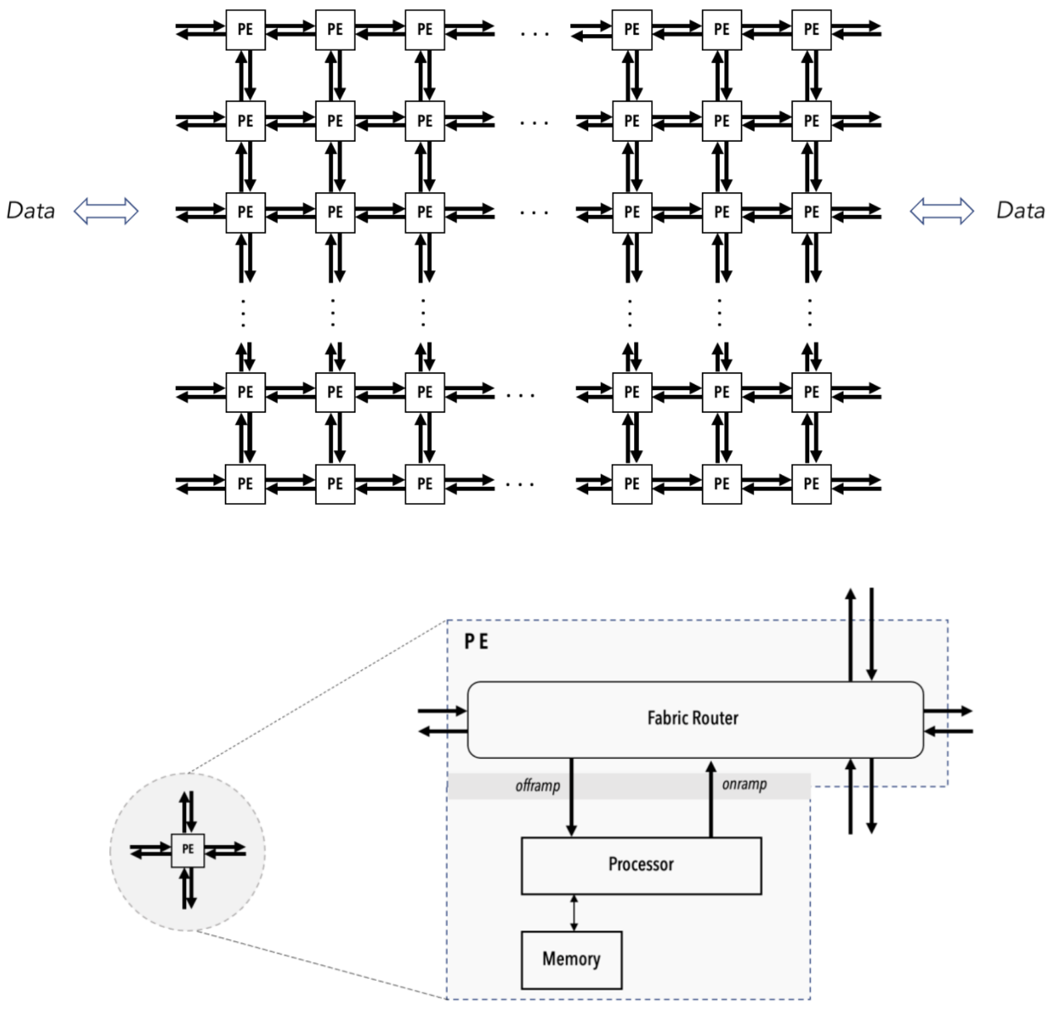
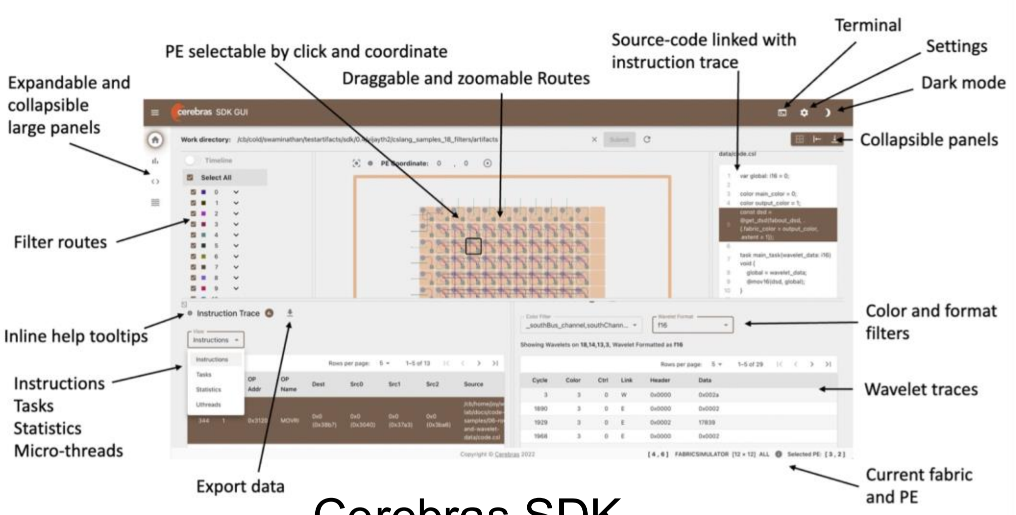


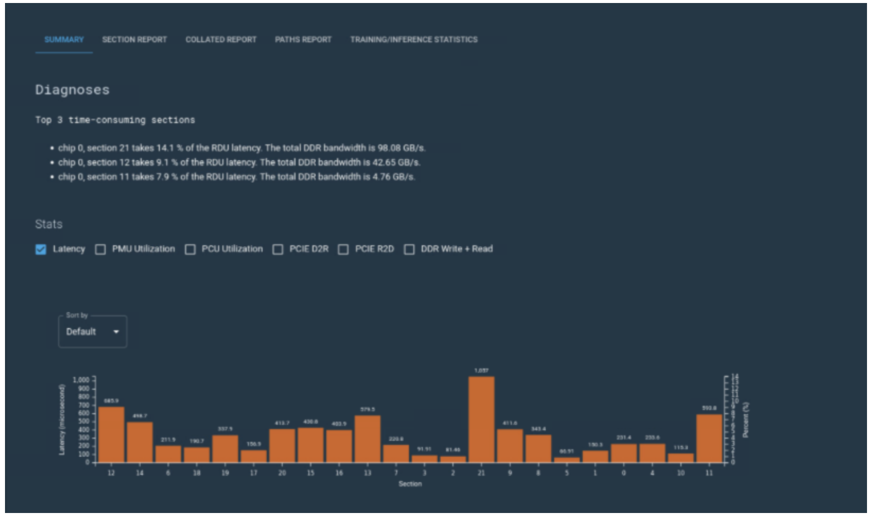
Image coutesy: Cerebras

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

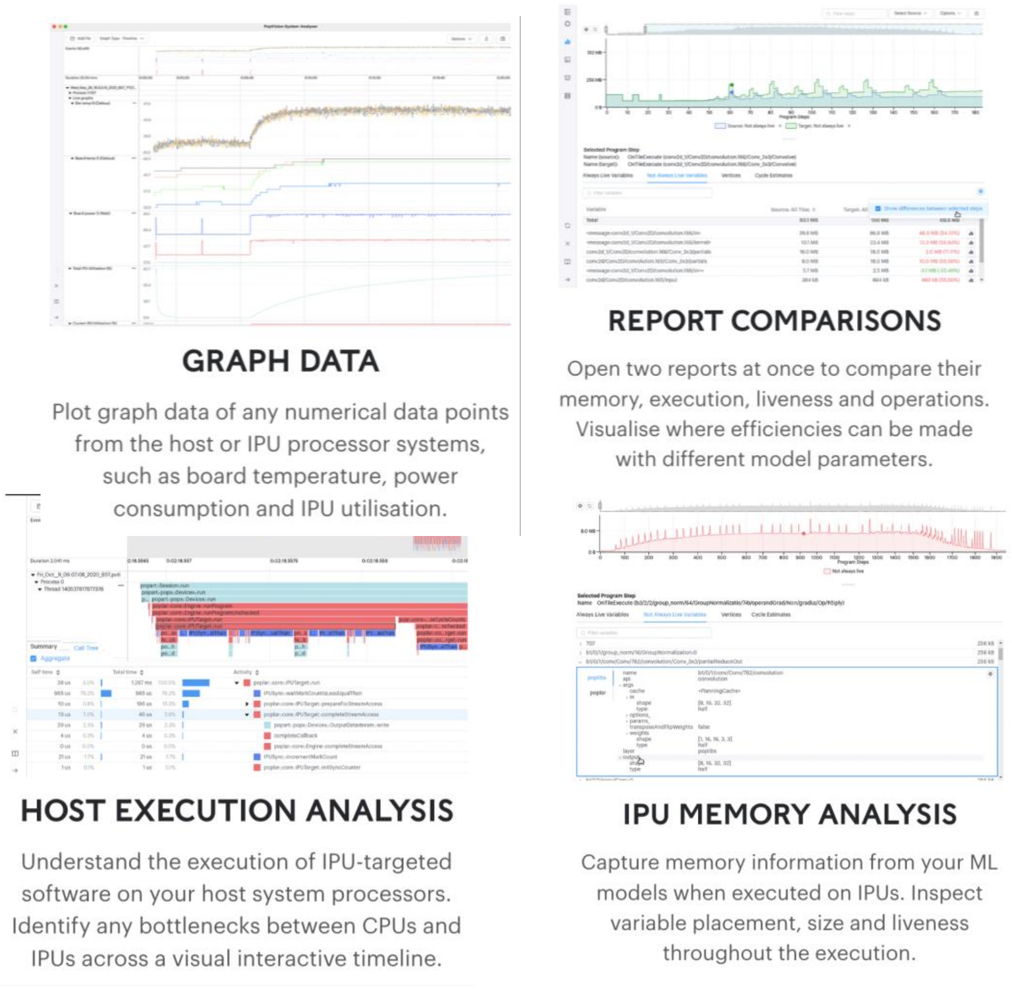
Tools on AI Accelerators



Cerebras SDK



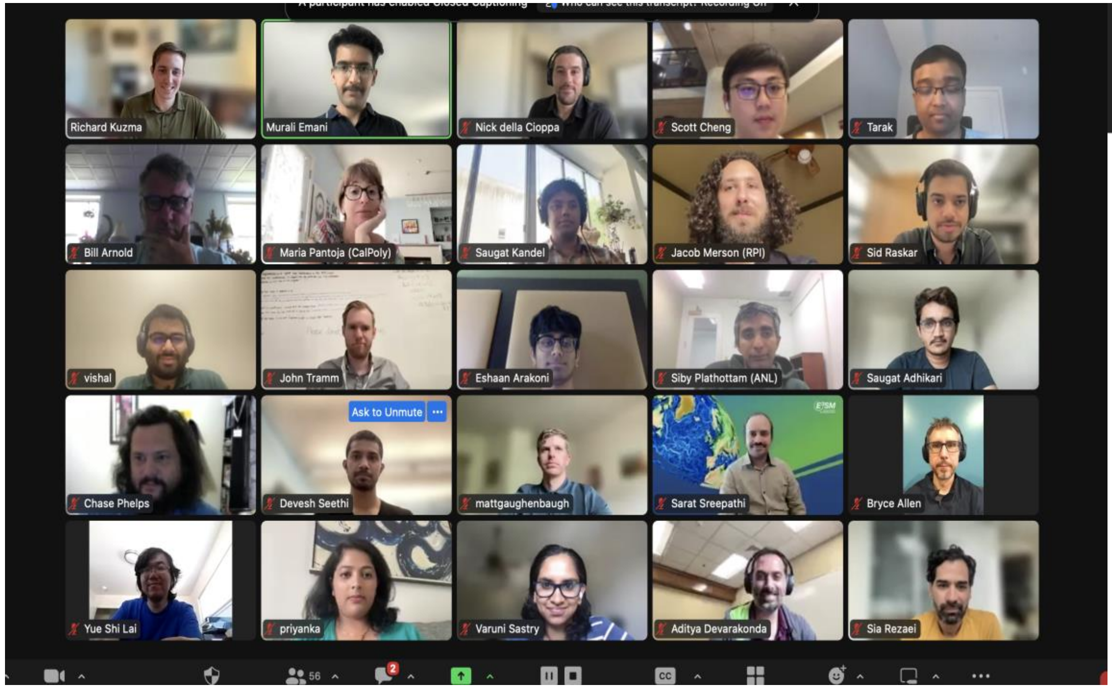
SambaTune on SambaNova

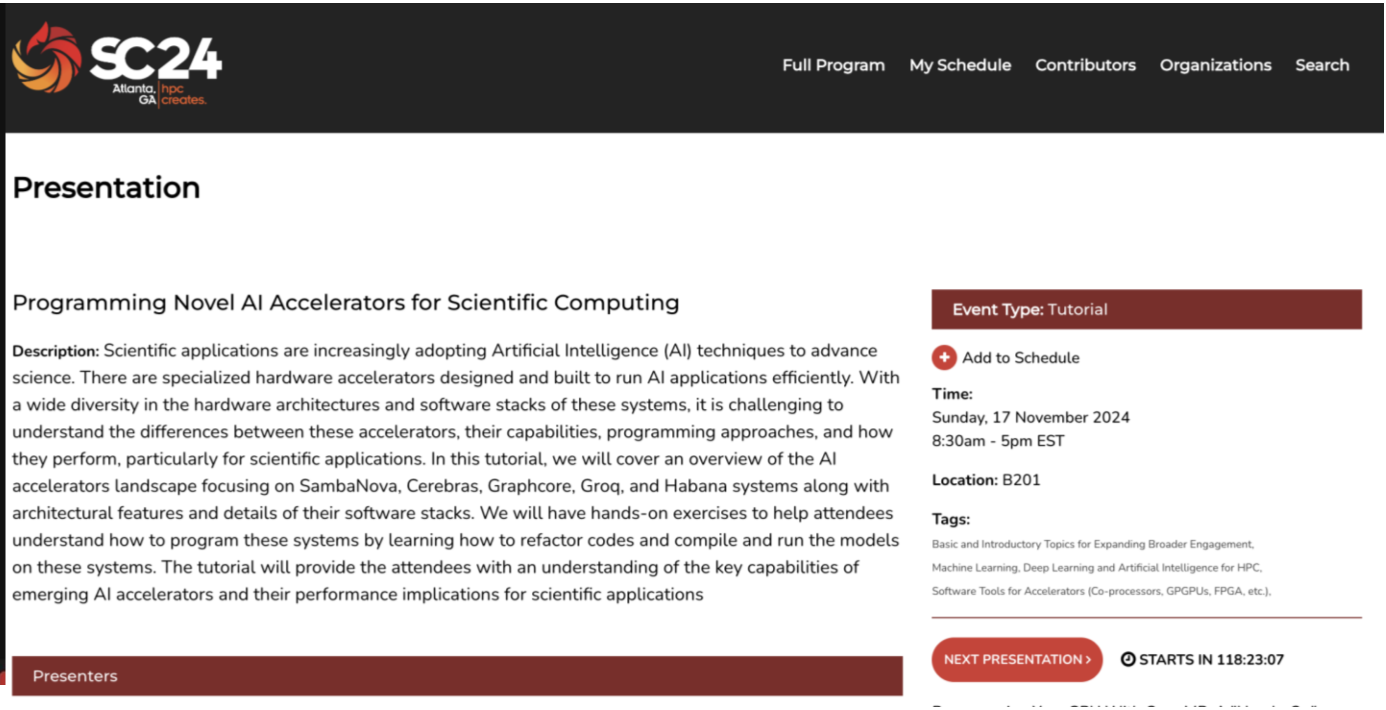


PopVision on GraphCore

Argonne NATIONAL LABORATORY

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

Tutorial at SC24/ISC25 on Programming Novel AI accelerators for Scientific Computing ***in collaboration with Cerebras, Intel Habana, Graphcore, Groq and SambaNova***

Upcoming Tutorial at SC25 St Louis, Missouri

Argonne NATIONAL LABORATORY

Getting Started on ALCF AI Testbed

Available for Allocations

- Cerebras CS-3,
- SambaNova Datascale SN30,
- GroqRack
- Graphcore Bow Pod64
- Sambanova Inference – Metis SN40L

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/

NEWS EVENTS PEOPLE CAREERS Search

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.

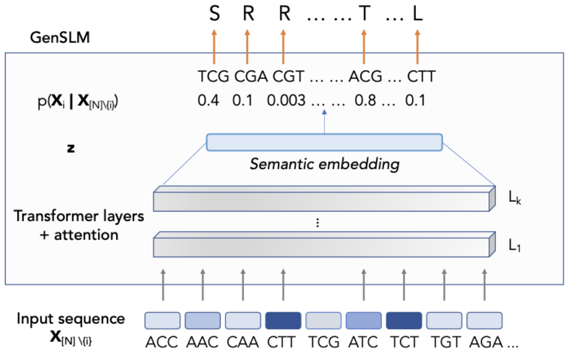
The goal of the testbed is to enable explorations into next-generation machine learning applications and workloads, enabling the ALCF and its user community to help define the role of AI accelerators in scientific computing and how to best integrate such technologies with supercomputing resources.

Table of contents

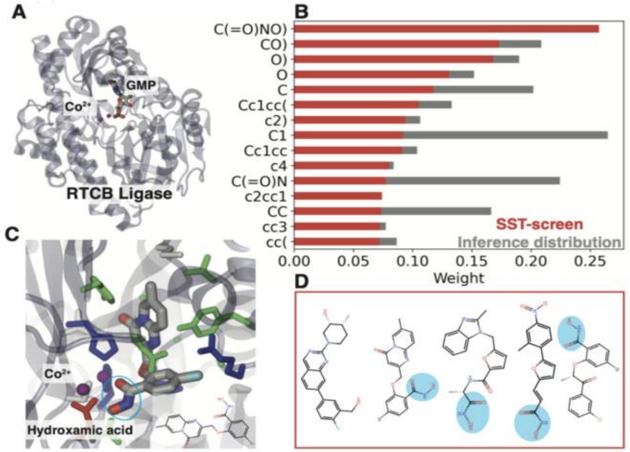
- How to Get Access
- Getting Started
- How to Contribute to Documentation

AI Based Models

Text Based Models

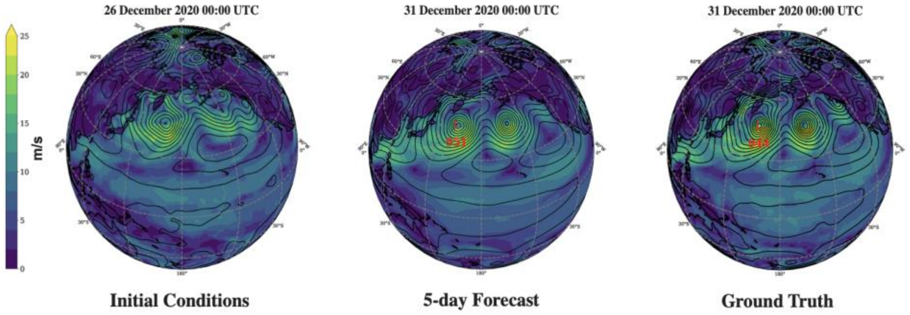


VOC detection

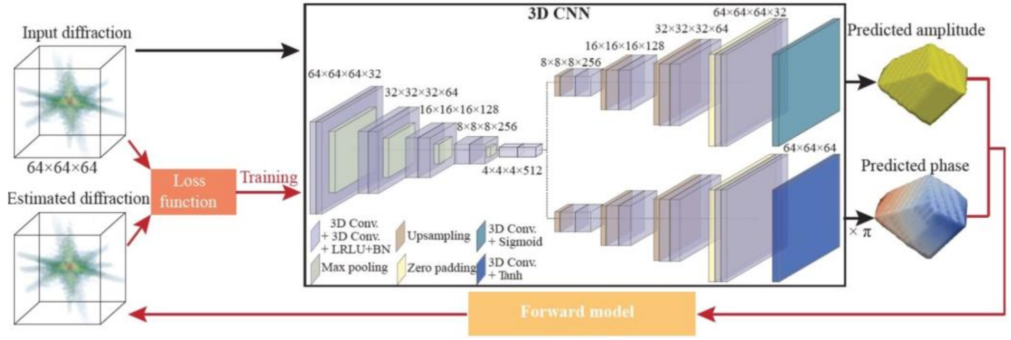


Drug and Molecular discovery

Vision Models



Stormer – Weather Forecasting



Diffraction Imaging

Cosmology and more ..

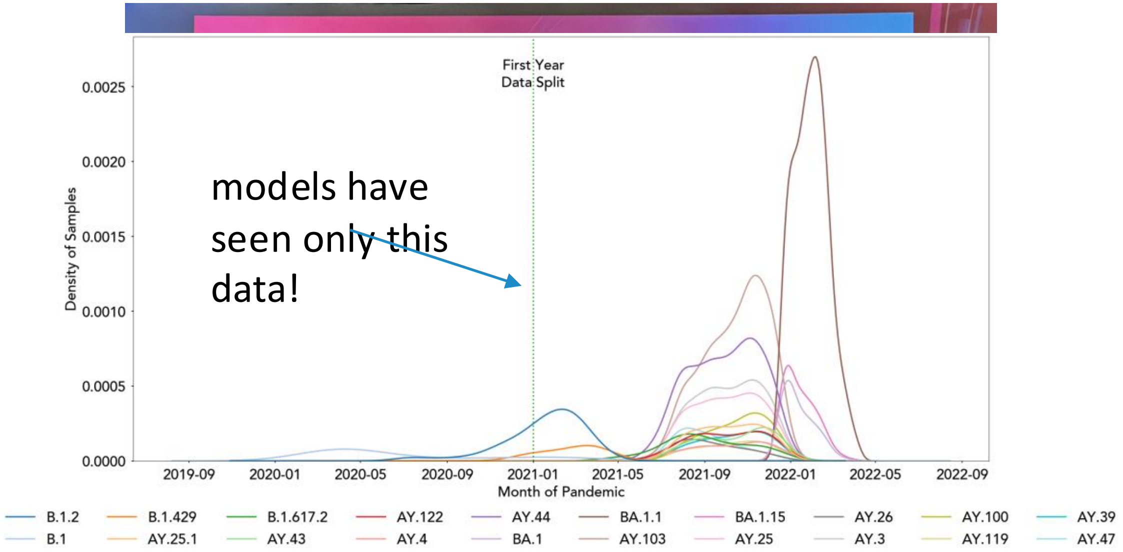
Genome-scale Language Models (GenSLMs)

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

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

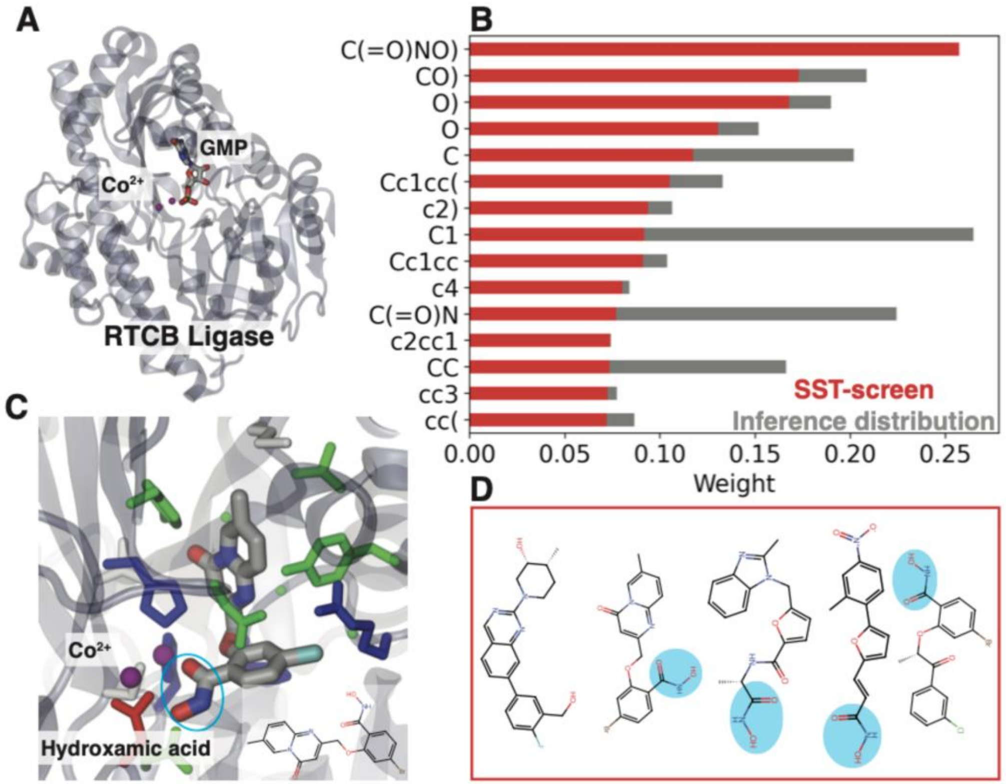
System	Number of Devices	Throughput (tokens/sec)	Improvement	Energy Efficiency
NVIDIA A100	8	1150	1.0	1.0
SambaNova SN30	8	9795	8.5	5.6
Cerebras CS-2	1	29061	25	6.5

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

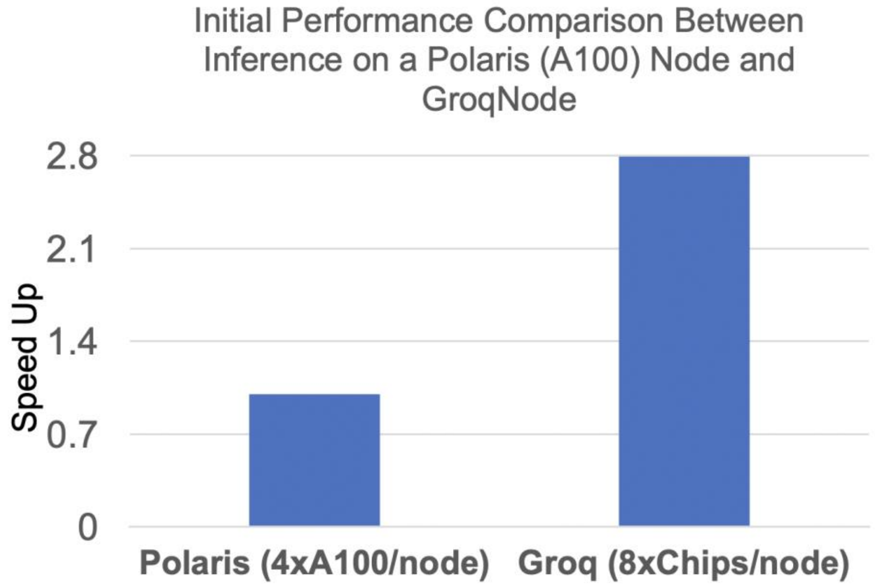
Accelerating Drug Design and Discovery with Machine Learning

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*



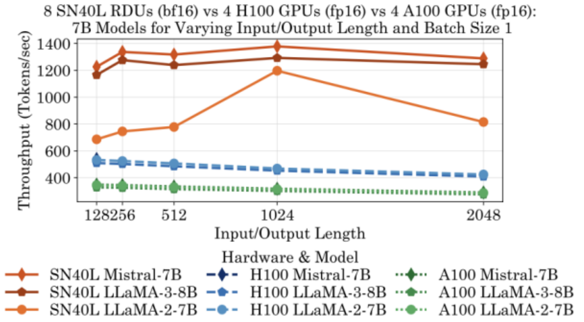
Courtesy: Archit Vasan

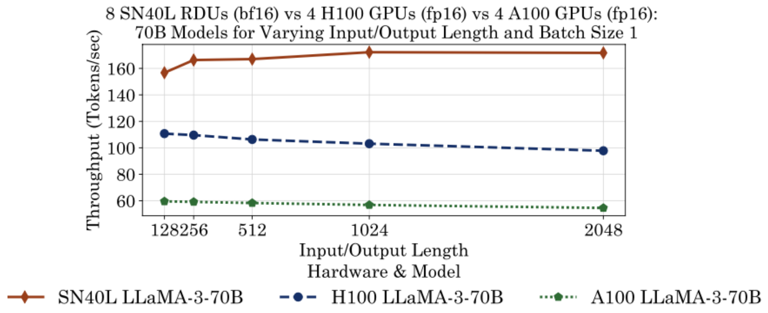
*Simplified Molecular Input Line Entry System (SMILES) - Representation for Molecules

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

Inference Benchmarking

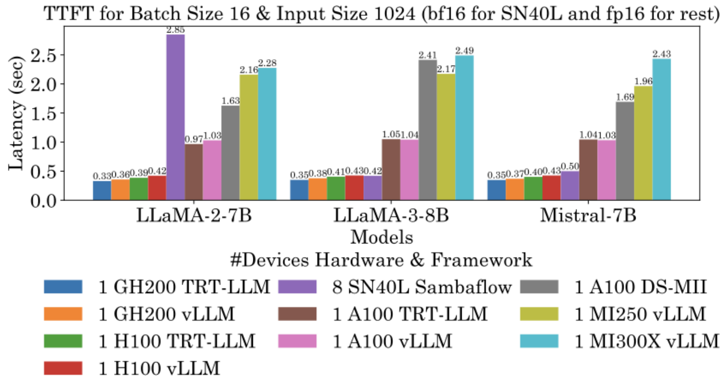
LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators



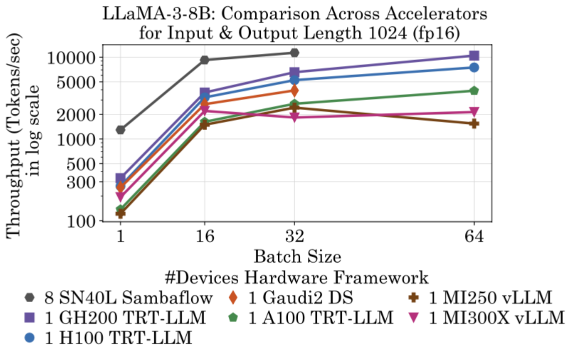


Throughput Comparison of 7B and 70B Llama Models on 8 SN40L RDUs with 4 H100s and 4 A100s GPU

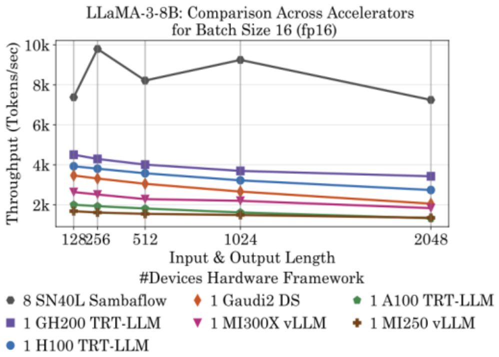
https://arxiv.org/abs/2411.00136



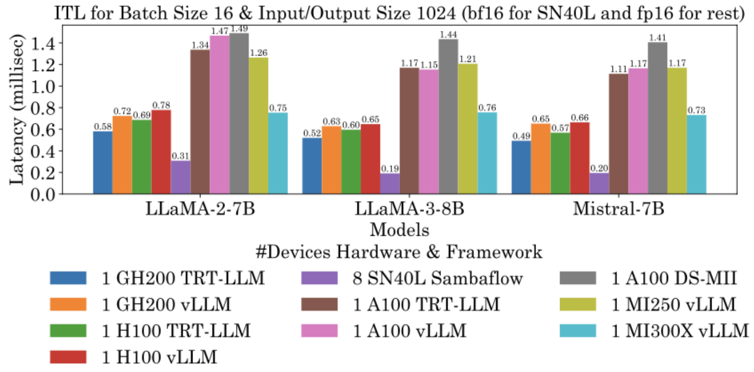
Time to first token (TTFT)



Throughput Vs Batch Size



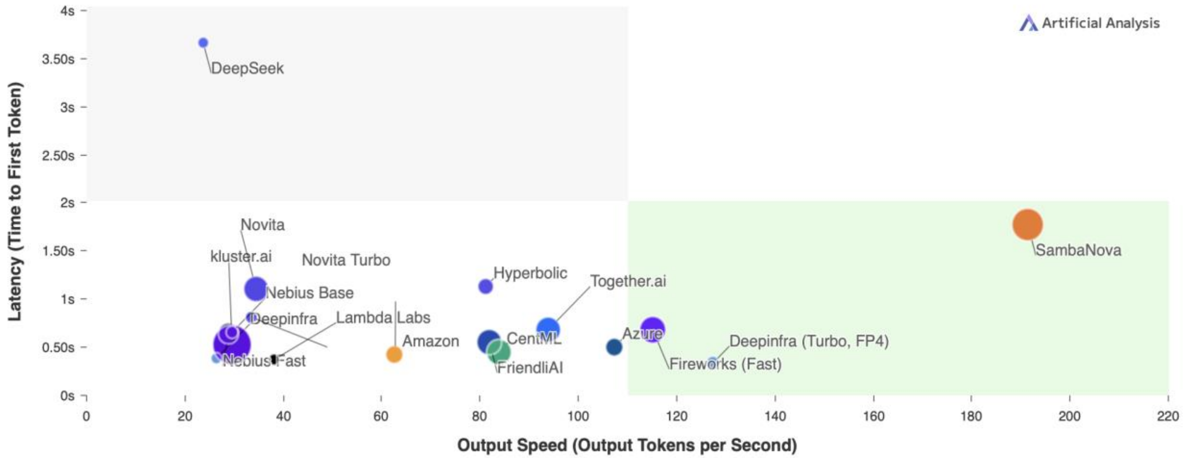
Throughput Vs I/O length



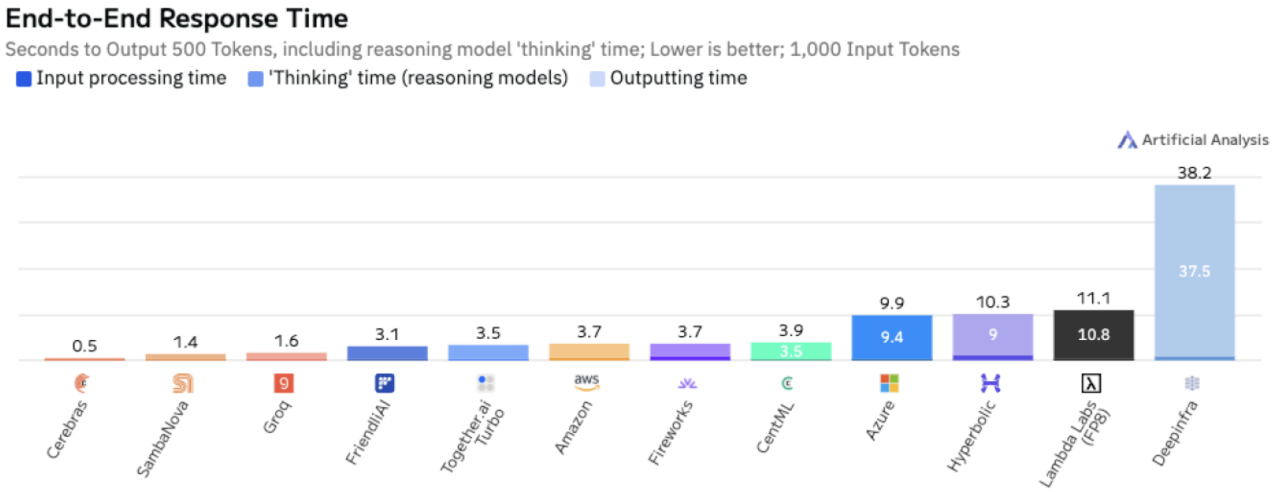
Inter Token Latency (ITL)

Argonne NATIONAL LABORATORY

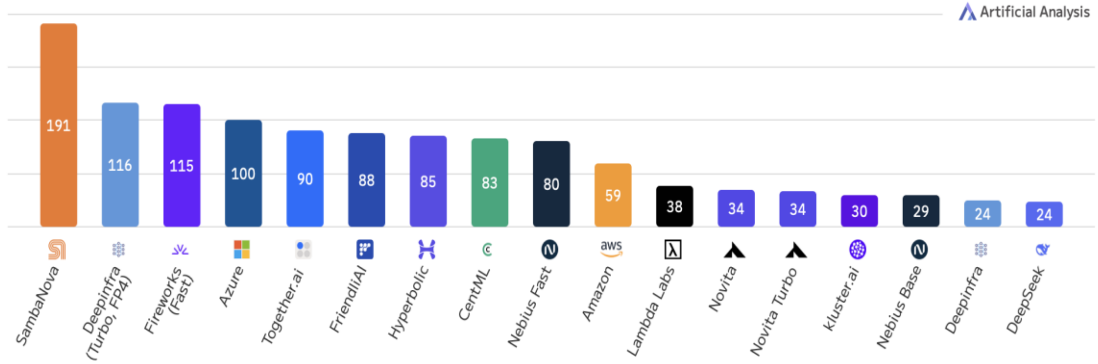
Inference Performance



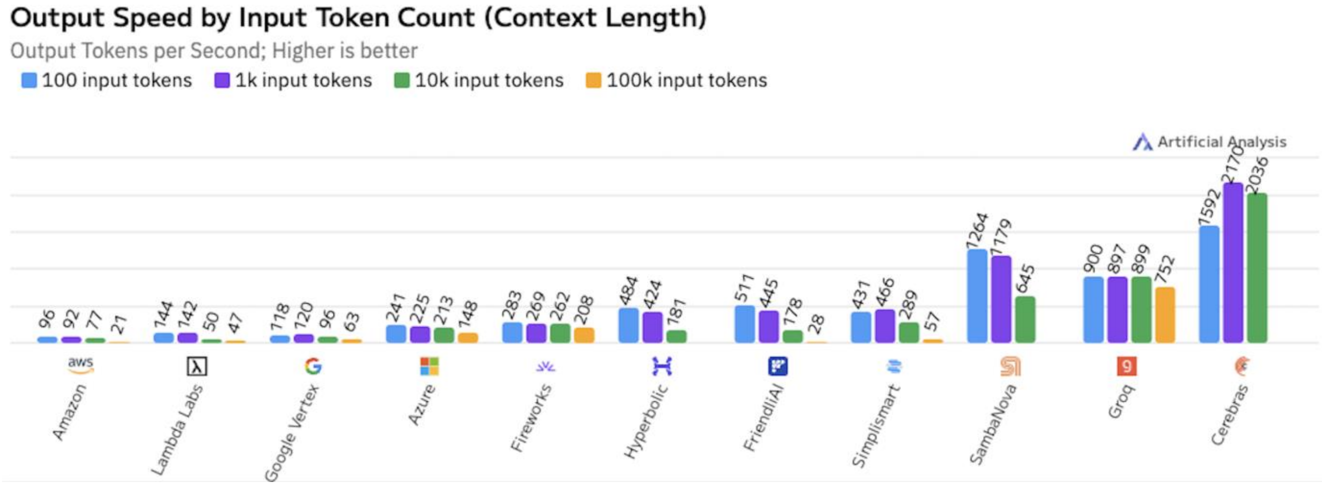
Deepseek R1 latency (TTFT)



Response time for llama 3 70B



Deepseek R1 Output speed



Output speed for llama 8b

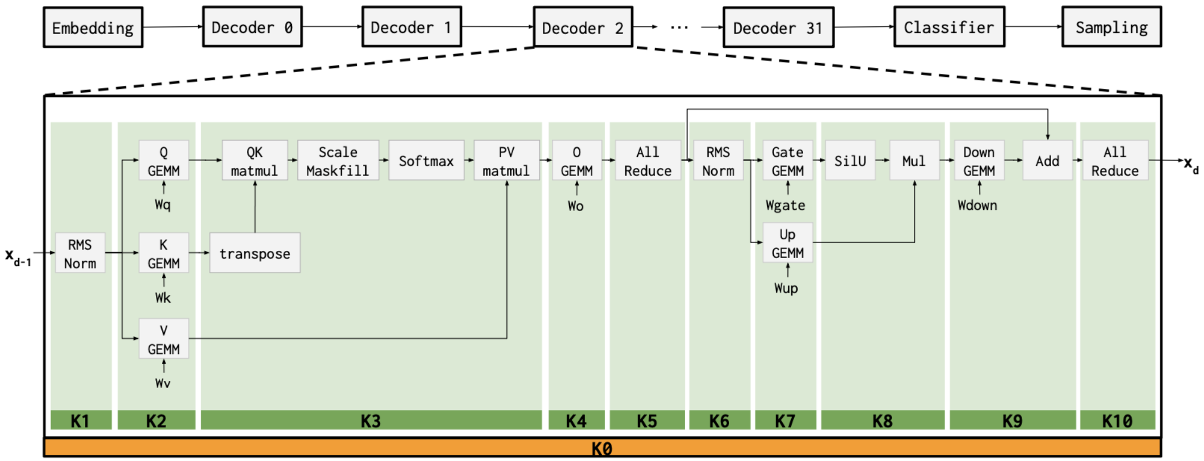
Argonne NATIONAL LABORATORY

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)
```

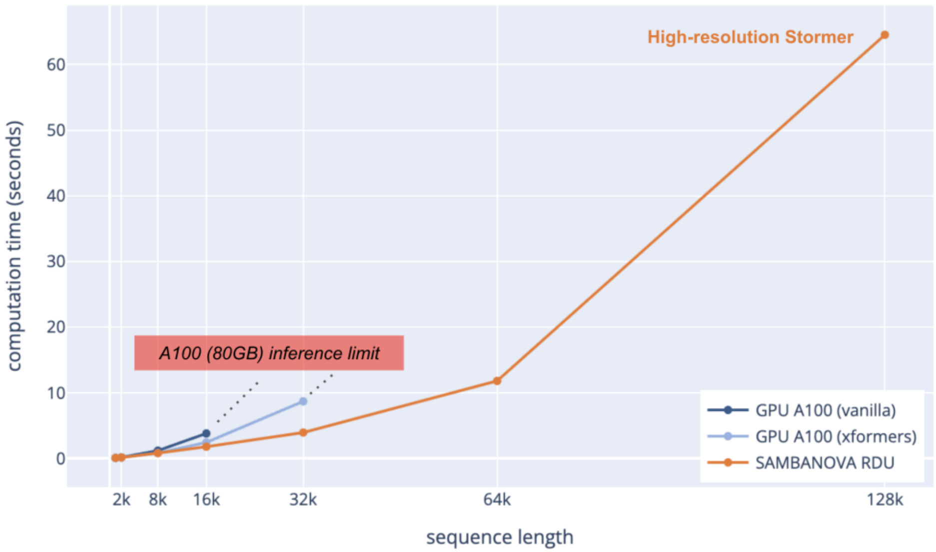
```
// SN40L-8
// Kernel Call Schedule

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

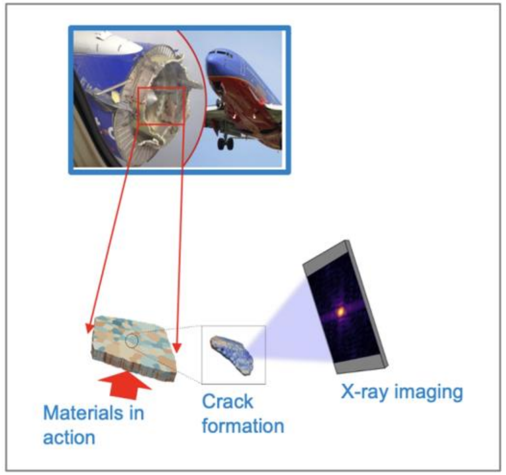
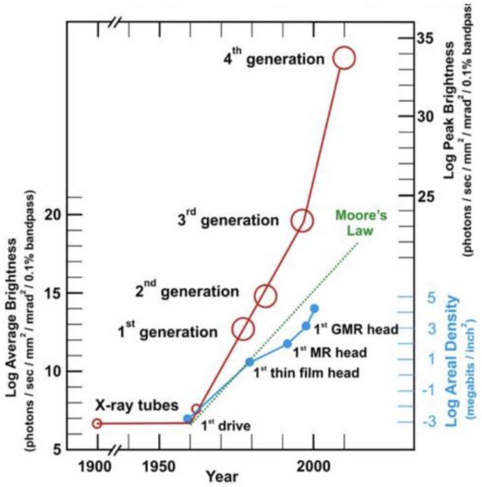
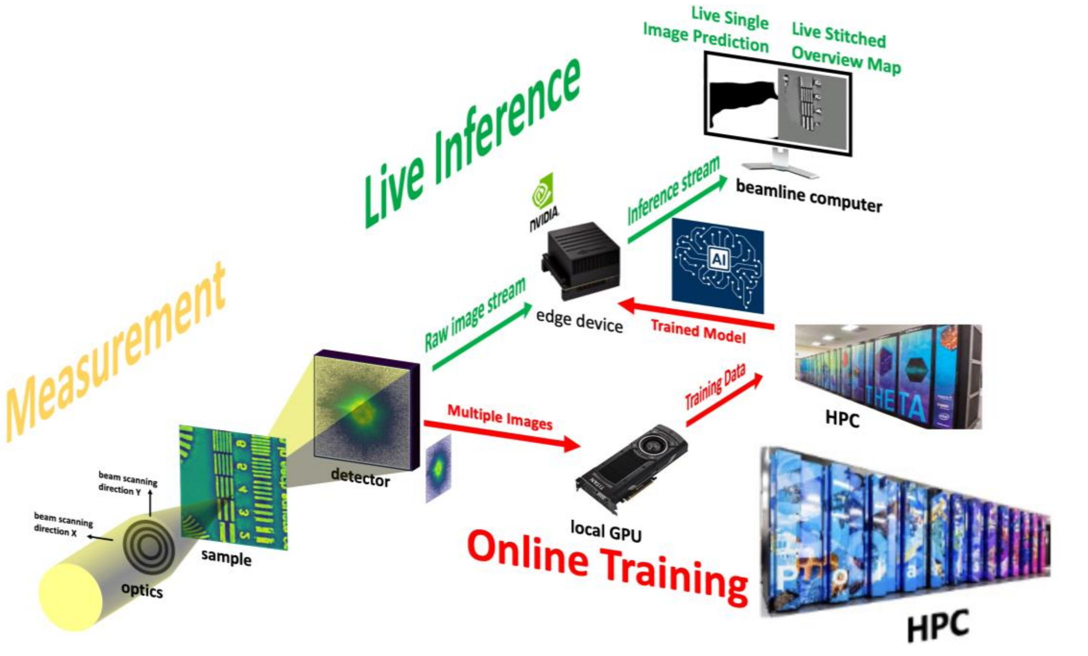


Image adapted from: Jon Almer, Stephan Hruszkewycz et al., ANL



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

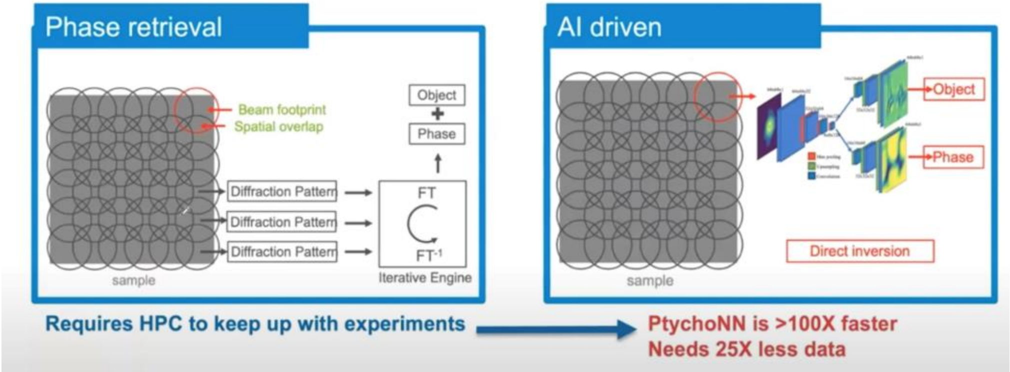
[1] APS Needs and Progress on Machine Learning and Artificial Intelligence Applications

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

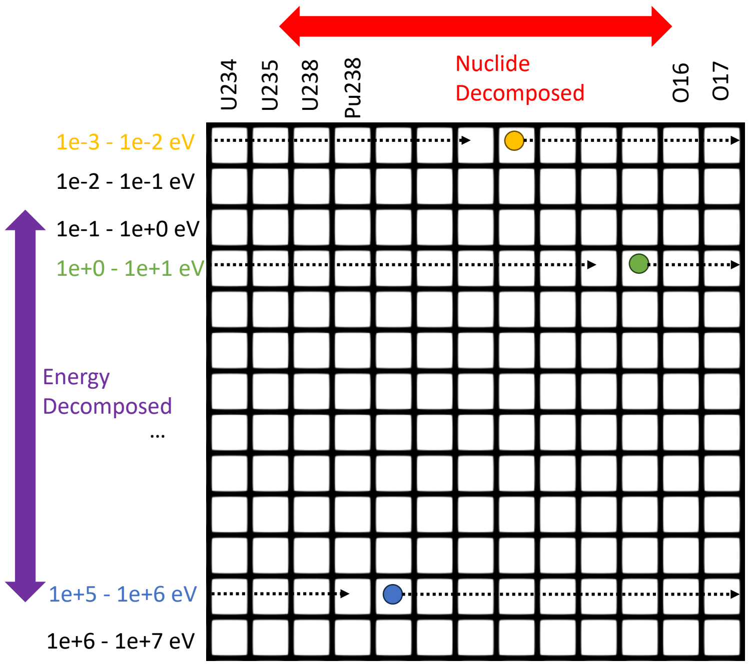
Spec	CS-3 / B200	CS-3 / DGX B200	CS-3 / NVL72
FP16 PFLOPs	28.4	3.5	0.3
Memory (GB)	6,250.0	781.3	88.9
NVLink | Fabric Bandwidth (TB/s)	14,861	1,858	206
Power (Watts)	23.0	1.6	0.2
PFLOPs / W	1.2	2.2	1.8



Argonne NATIONAL LABORATORY

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.



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.

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.**

	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**

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

Data courtesy: John Tramm (Argonne)

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, 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, Intel Habana and SambaNova. There are ongoing engagements with other vendors.

Hands On

https://github.com/argonne-lcf/ALCF_Hands_on_HPC_Workshop/tree/master/aiTestbeds