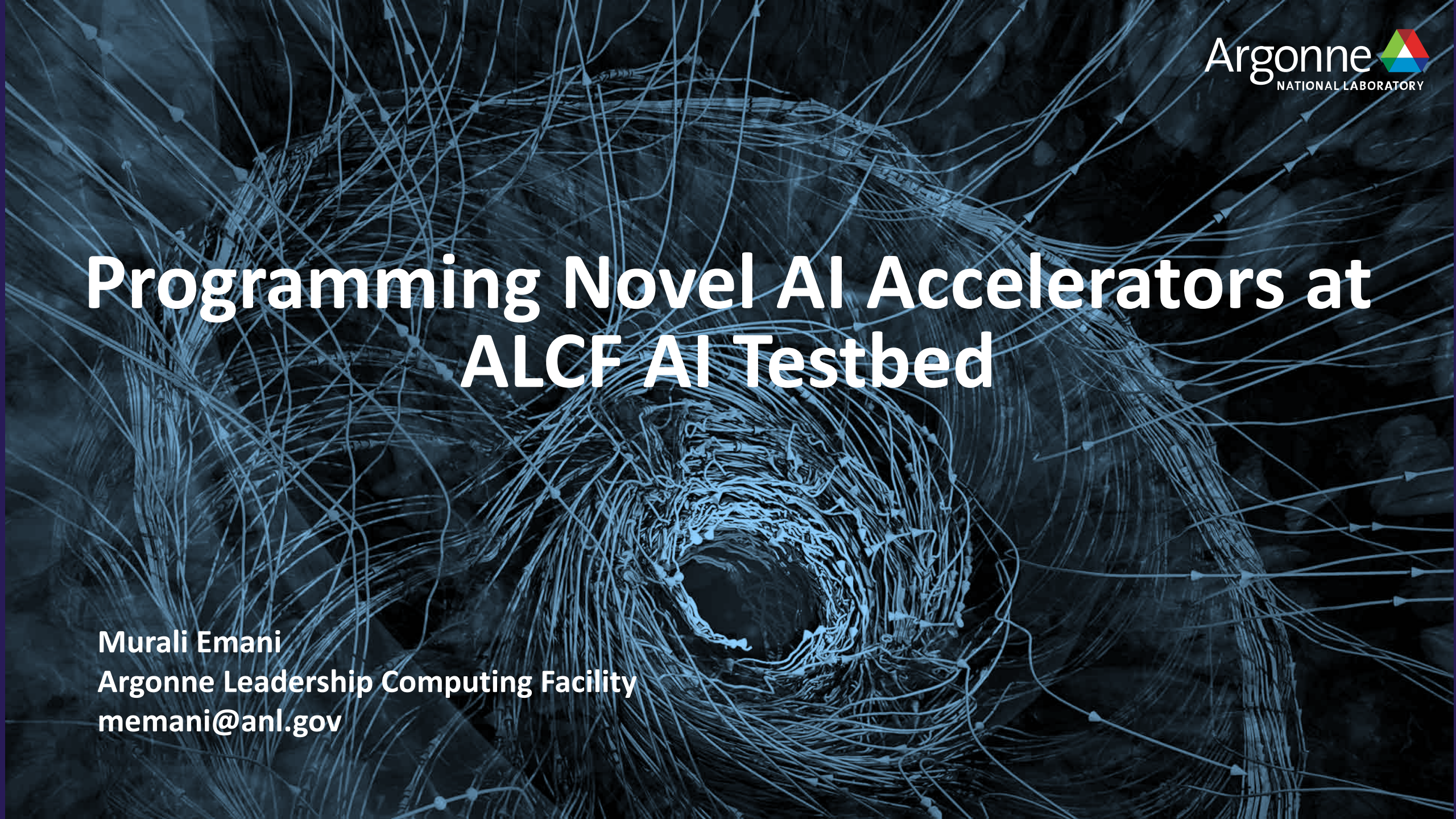
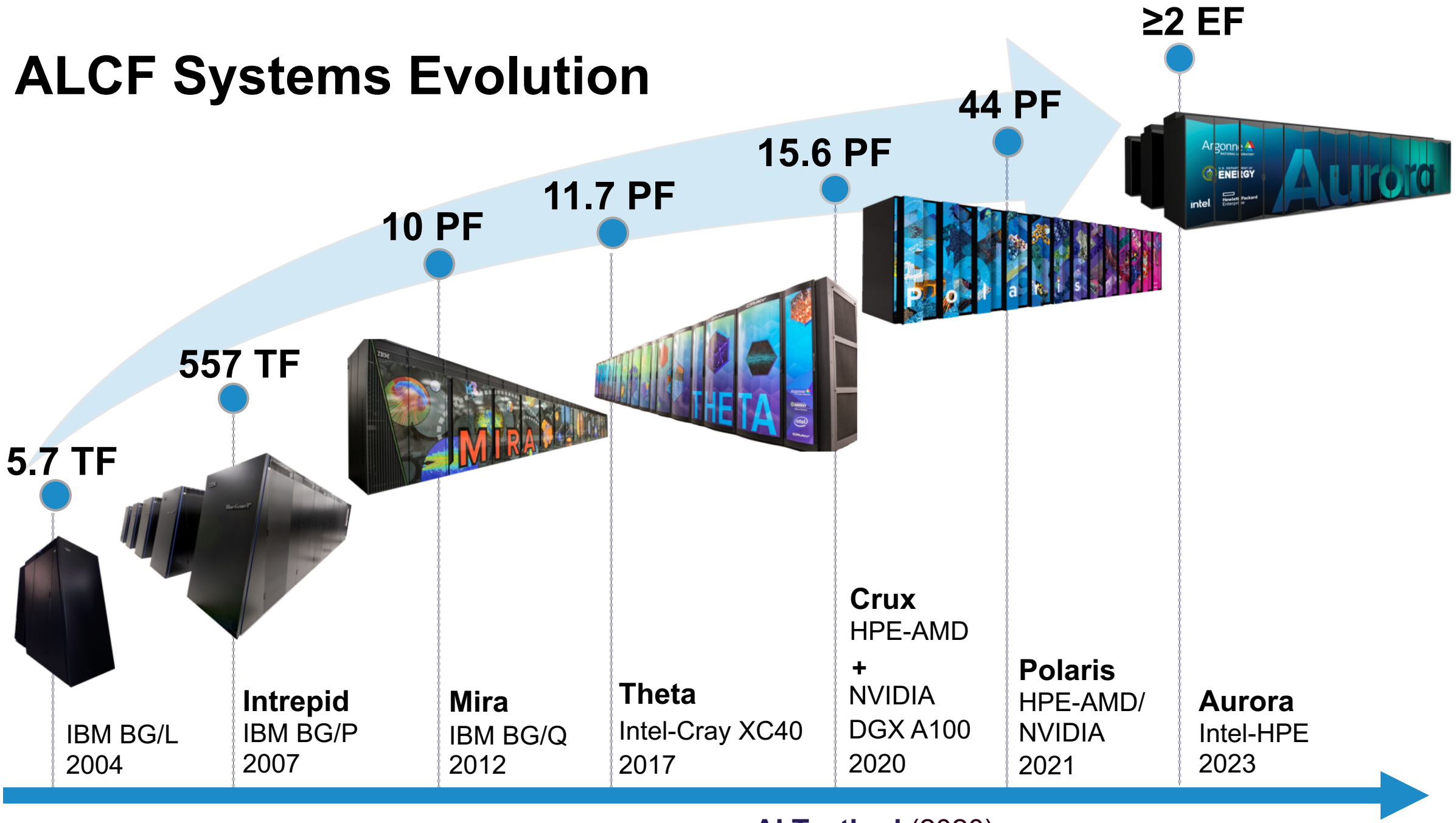


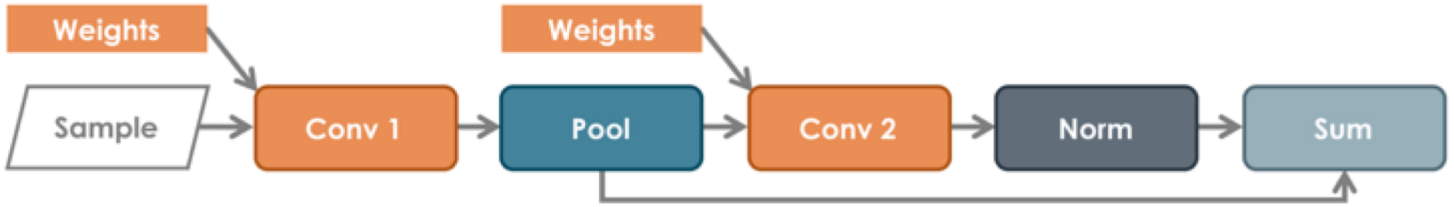
Programming Novel AI Accelerators at ALCF AI Testbed

Murali Emani
Argonne Leadership Computing Facility
memani@anl.gov

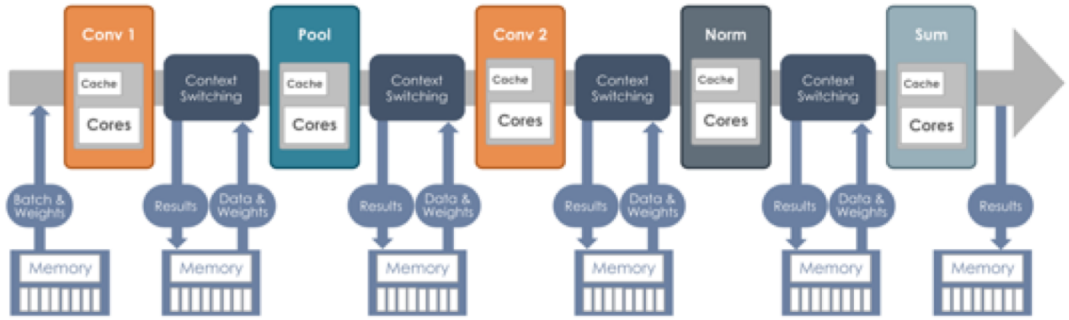
ALCF Systems Evolution



Dataflow Architectures

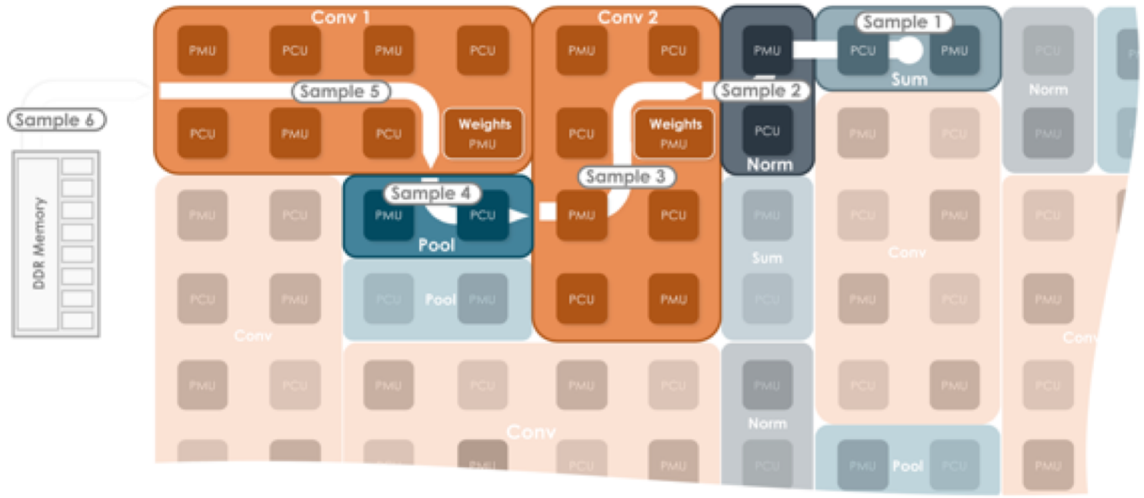


Simple Convolution Graph



The old way: kernel-by-kernel

Bottlenecked by memory bandwidth and host overhead



The Dataflow way: Spatial

Eliminates memory traffic and overhead

Image Courtesy: SambaNova

AI Accelerators

- An AI accelerator is a high-performance parallel computation machine that is specifically designed for the efficient processing of AI workloads like neural networks.
- Types of AI accelerators:
  - Graphic processing units
  - Massive multicore scalar processors
  - Dataflow architectures etc.
- Benefits
  - Improved model performance in throughput and latency
  - potential to deal with large, complex models
  - handle high-resolution datasets
  - power efficiency

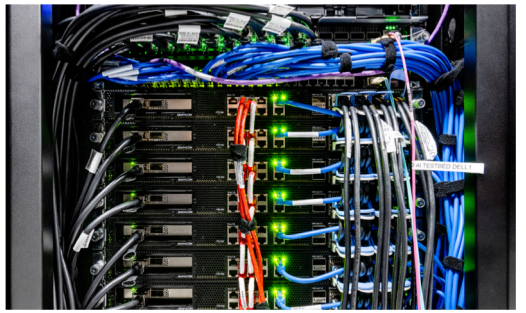
ALCF AI Testbed

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

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

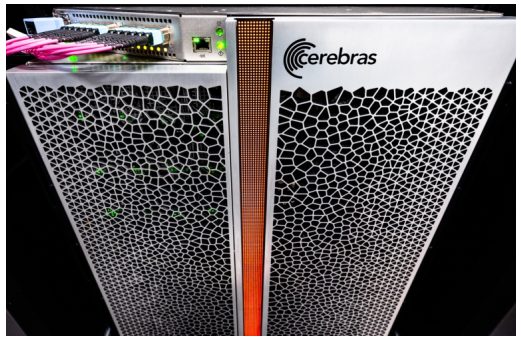
8 nodes each with 8 Reconfigurable DataFlow Units (RDU)

SambaNova SN-30



4 nodes each with 16 Intelligent Processing Units (IPUs)

Graphcore Bow Pod64



2 CS-2 Wafer scale engines (WSE)

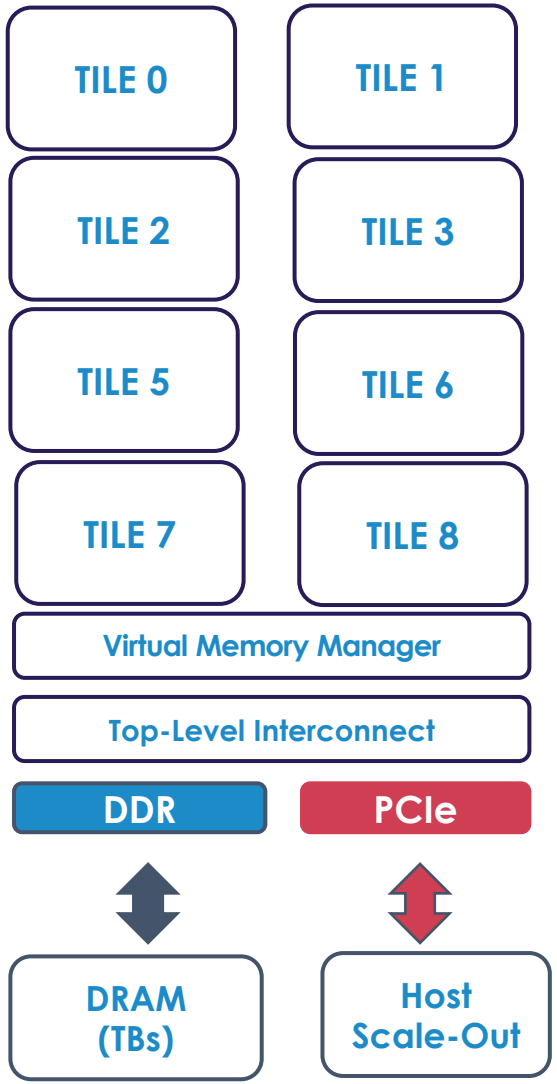
Cerebras CS-2

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

Groq

https://nairrpilot.org

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

Cerebras Wafer-Scale Engine (WSE-2)

850,000 cores optimized for sparse linear algebra

46,225 mm² silicon

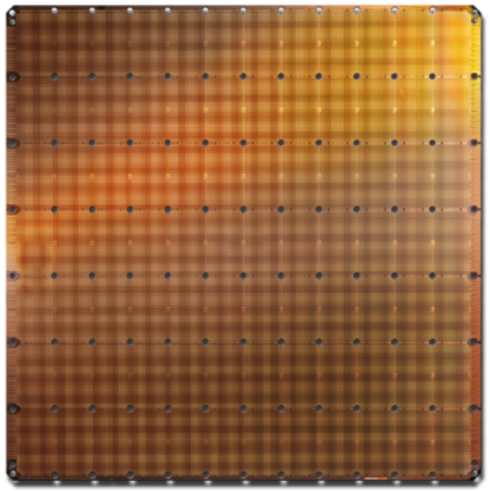
2.6 trillion transistors

40 gigabytes of on-chip memory

20 PByte/s memory bandwidth

220 Pbit/s fabric bandwidth

7nm process technology

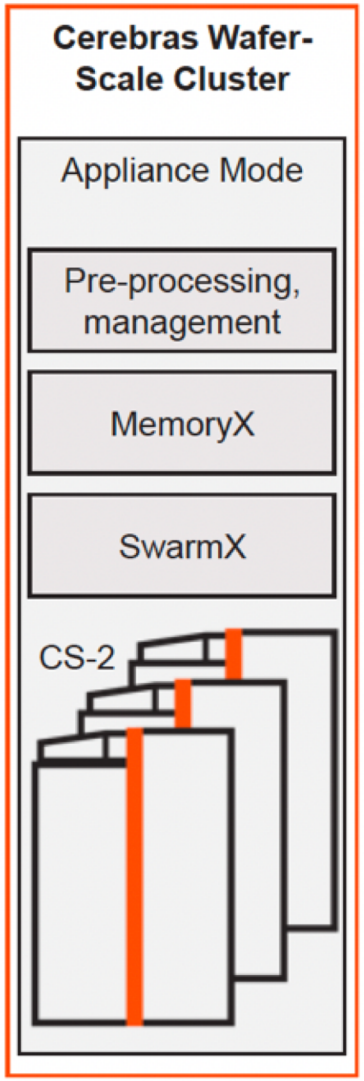


Cerebras WSE
1.2 Trillion transistors
46,225 mm² silicon

Largest GPU
21.1 Billion transistors
815 mm² silicon

https://www.cerebras.net/blog/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

Wafer-Scale Cluster



Input preprocessing servers stream training data

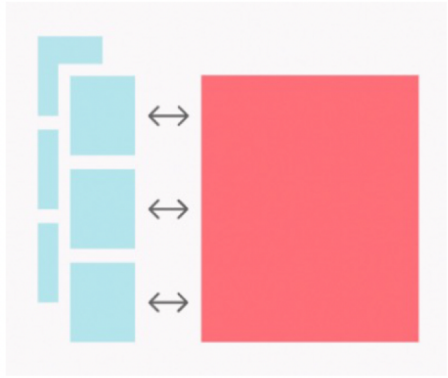
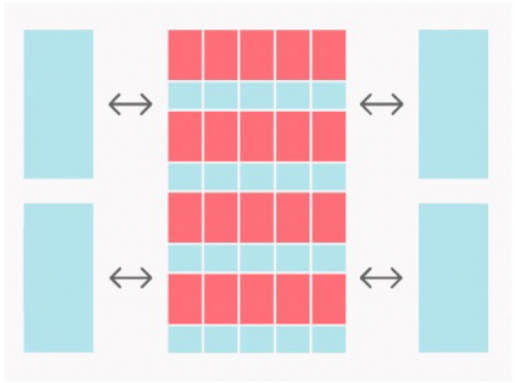
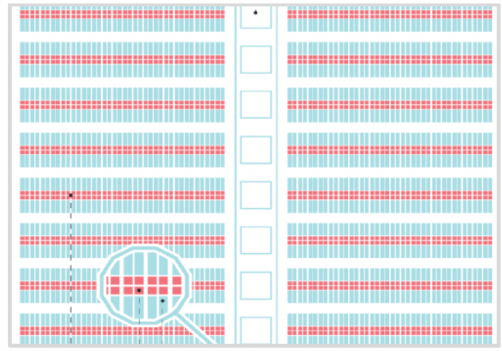
MemoryX - Stores and streams model's weights

SwarmX – weight broadcasts and gradient across multiple CS2s

Compilation (maps graph to kernels) Execution (training)

Image Courtesy: Cerebras

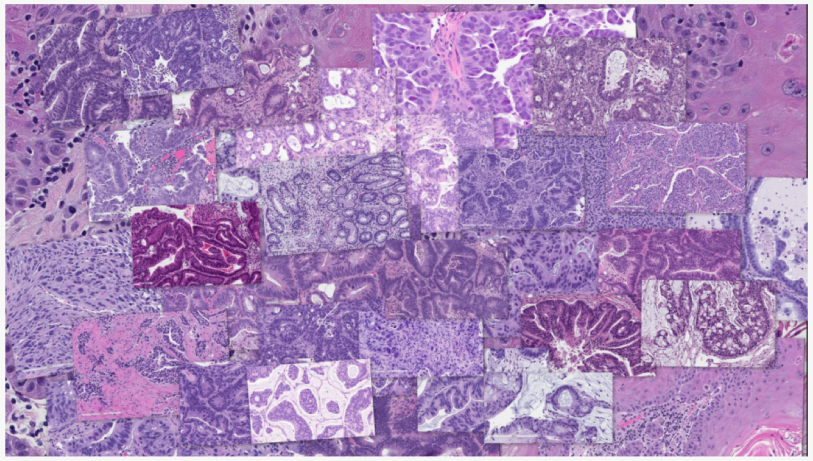
Graphcore Intelligence Processing Unit (IPU)

	CPU	GPU	IPU
Parallelism	Designed for scalar processing	SIMD/SIMT architecture. Designed for large blocks of dense contiguous data	Massively parallel MIMD architecture. High performance/efficiency for future ML trends
			
Memory Bandwidth	Off-chip memory	Model and Data spread across off-chip and small on-chip cache and shared memory (2TB/s for A100 HBM)	Main Model & Data in tightly coupled large locally distributed SRAM (~65 TB/s for Bow IPU)

Slide Courtesy: Graphcore

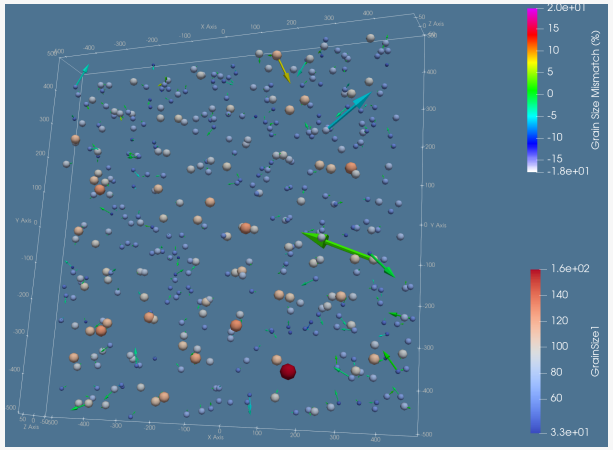
	Cerebras CS2	SambaNova Cardinal SN30 RDU	Groq GroqRack	GraphCore GC200 IPU	Habana Gaudi1	NVIDIA A100
Compute Units	850,000 Cores	640 PCUs	5120 vector ALUs	1472 IPUs	8 TPC + GEMM engine	6912 Cuda Cores
On-Chip Memory	40 GB L1, 1TB+ MemoryX	>300MB L1 1TB	230MB L1	900MB L1	24 MB L1 32GB	192KB L1 40MB L2 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 ™	IPU Link	Ethernet-based	NVLink

AI FOR SCIENCE AND HPC APPLICATIONS ON AI TESTBED



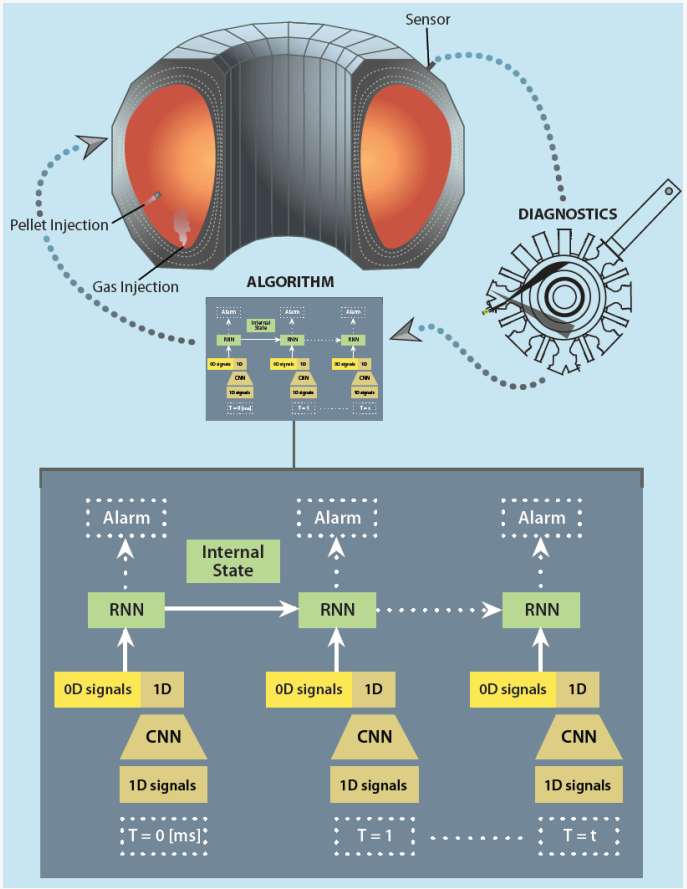
Cancer drug response prediction

(Credit: Candle)



Imaging Sciences-Braggs Peak

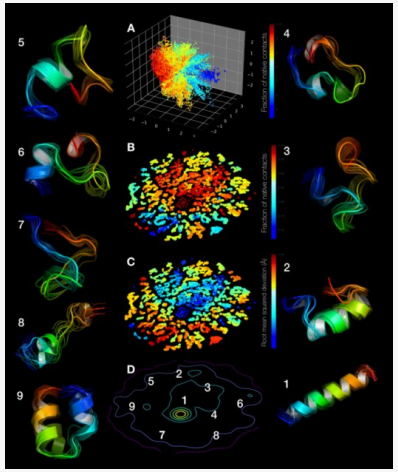
(Credit: Z. Liu)



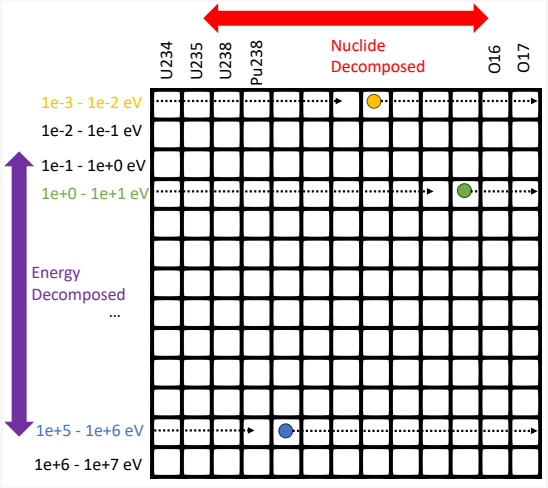
Tokomak Fusion Reactor operations

(Credit: K. Felker)

and more..



Protein-folding (Image: NCI)



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

GenSLM 13B Training Performance

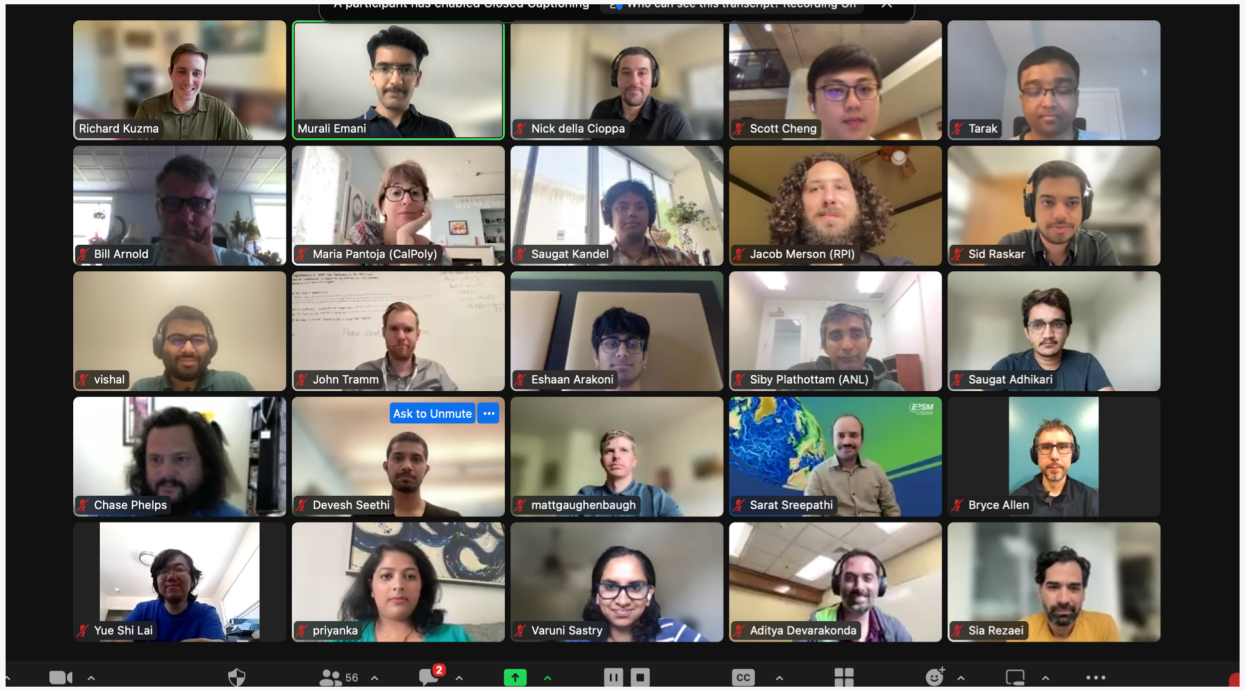
GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics

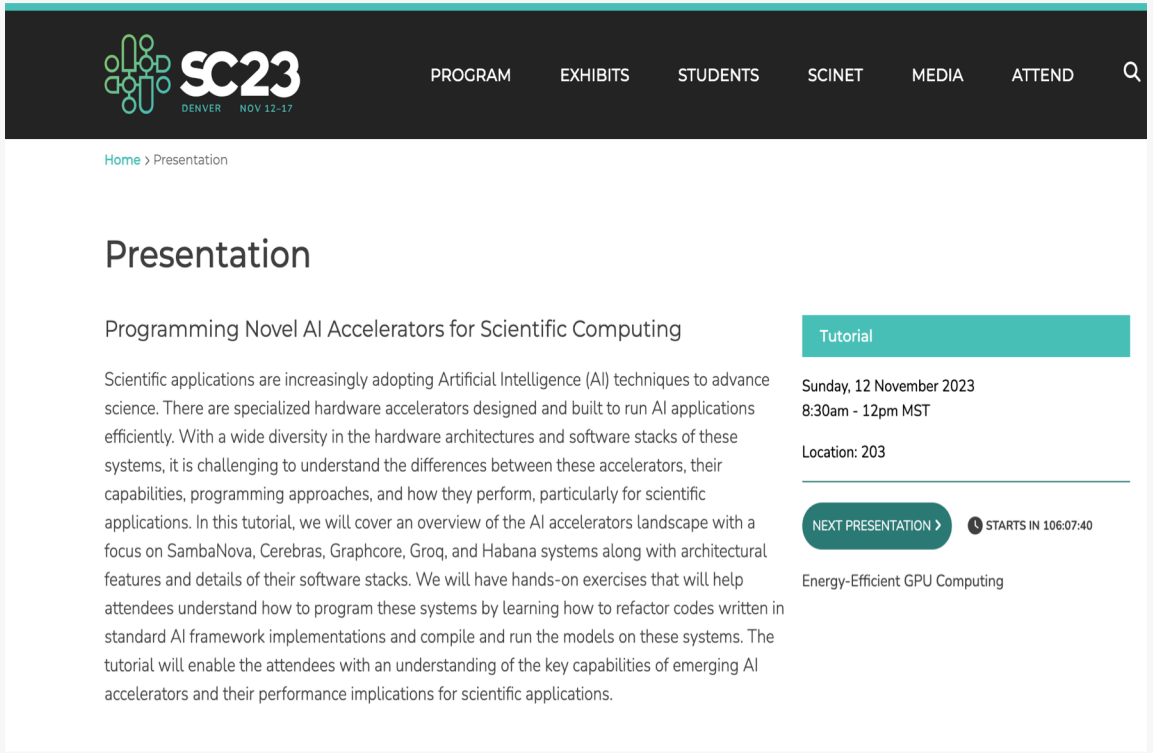
Winner of the ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research, 2022

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

AI Testbed Community Engagement





AI Training workshops

Tutorial at SC23 on Programming Novel AI accelerators for Scientific Computing

AI Testbed Training series (starting April ’24)
https://www.alcf.anl.gov/ai-testbed-training-workshops

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

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, William Arnold, Varuni Sastry, Sid Raskar, Rajeev Thakur, Anthony Avarca, Arvind Ramanathan, Alex Brace, Zhengchun Liu, Hyunseung (Harry) Yoo, Corey Adams, Kyle Felker, Craig Stacey, Ray Powell, Bill Lucnik, Skip Reddy, Tom Brettin, Michael Papka, 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.