Accelerating Scientific Machine Learning with AI Accelerators

Murali Emani,
Argonne Leadership Computing Facility
memani@anl.gov

ARGONNE TRAINING PROGRAM ON EXTREME-SCALE COMPUTING (ATPESC)
July 31, 2023

Surge of Scientific Machine Learning

- Simulations/ surrogate models

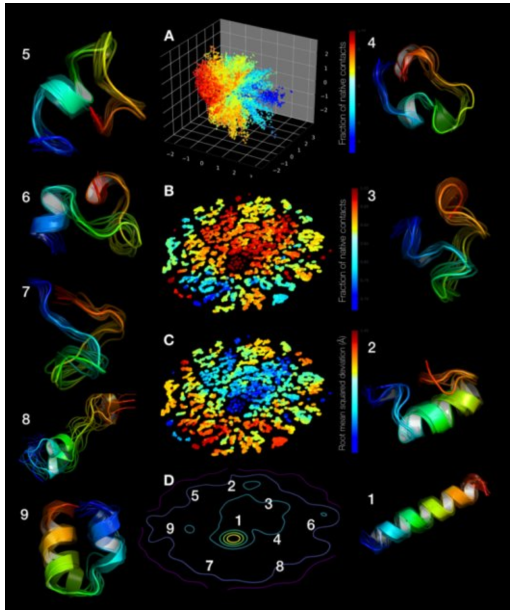
  Replace, in part, or guide simulations with AI-driven surrogate models

- Data-driven models

  Use data to build models without simulations

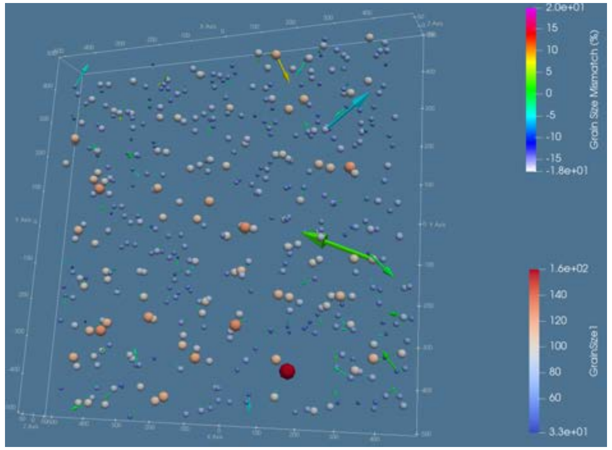
- Co-design of experiments

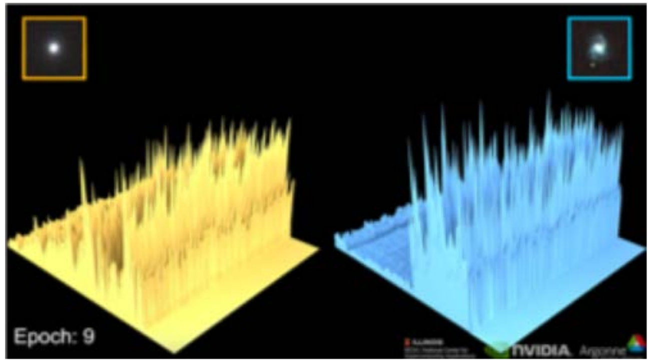
  AI-driven experiments



Protein-folding

Braggs Peak

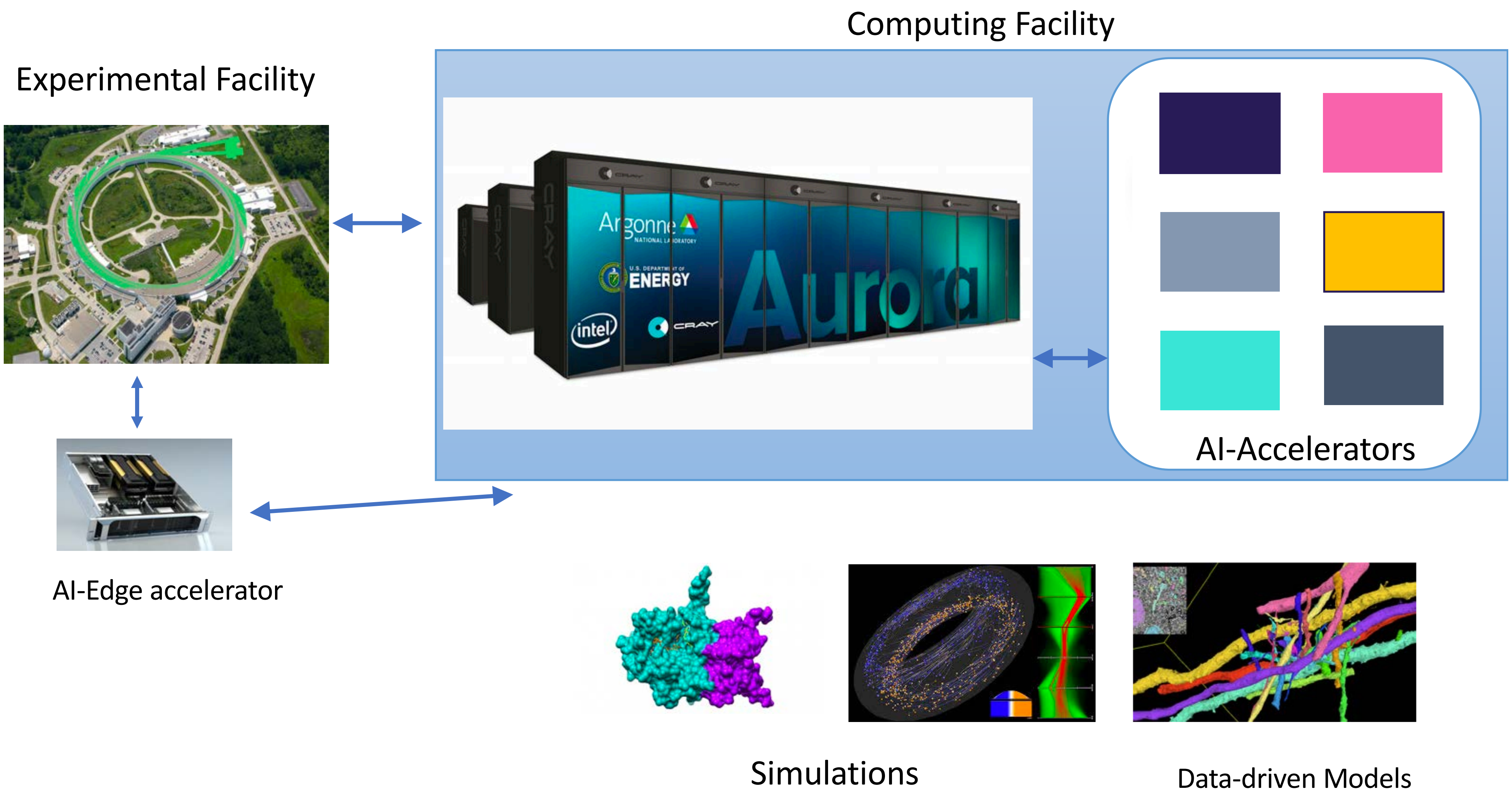




Galaxy Classification

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

Integrating AI Systems in Facilities



Argonne NATIONAL LABORATORY

ALCF AI Testbed

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

Cerebras CS-2

SambaNova DataScale SN30

Graphcore Bow Pod64

Habana Gaudi1

GroqRack

- Infrastructure of next-generation machines with AI hardware accelerators
- Provide a platform to evaluate usability and performance of AI4S applications
- Understand how to integrate AI systems with supercomputers to accelerate science

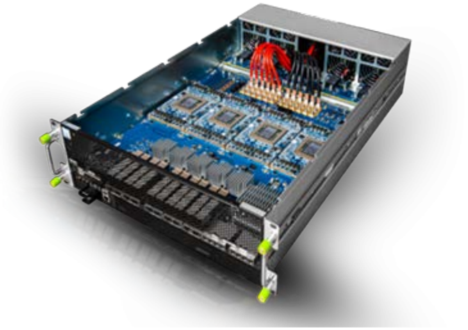
ALCF AI Testbed

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

Cerebras CS-2

SambaNova DataScale SN30

Graphcore Bow Pod64



Habana Gaudi1

GroqRack

- Cerebras: 2 CS-2 nodes, each with 850,000 Cores, compute-intensive models
- SambaNova: DataScale SN30 8 nodes (8 SN30 RDUs per node) - 1TB mem per device, models with large memory footprint
- Graphcore: Bow Pod64 4 nodes (16 IPUs per node) - MIMD, irregular workloads such as graph neural networks
- GroqRack: 8 nodes, 8 GroqNodes per node - inference at batch 1
- Habana Gaudi1: 2 nodes, 8 cards per node - On-chip integration of RDMA over Converged Ethernet (RoCE2), scale-out efficiency

Getting Started on ALCF AI Testbed:

Apply for a Director’s Discretionary (DD) Allocation Award

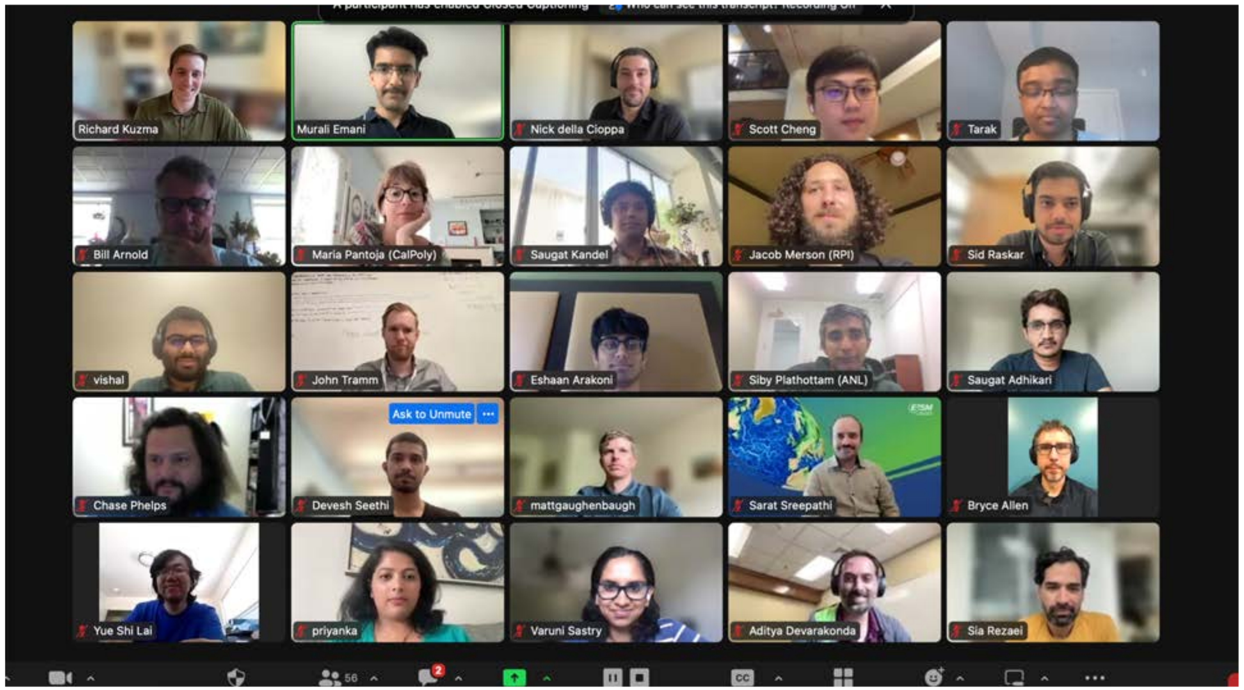
Director’s Discretionary (DD) awards support various project objectives from scaling code to preparing for future computing competition to production scientific computing in support of strategic partnerships.

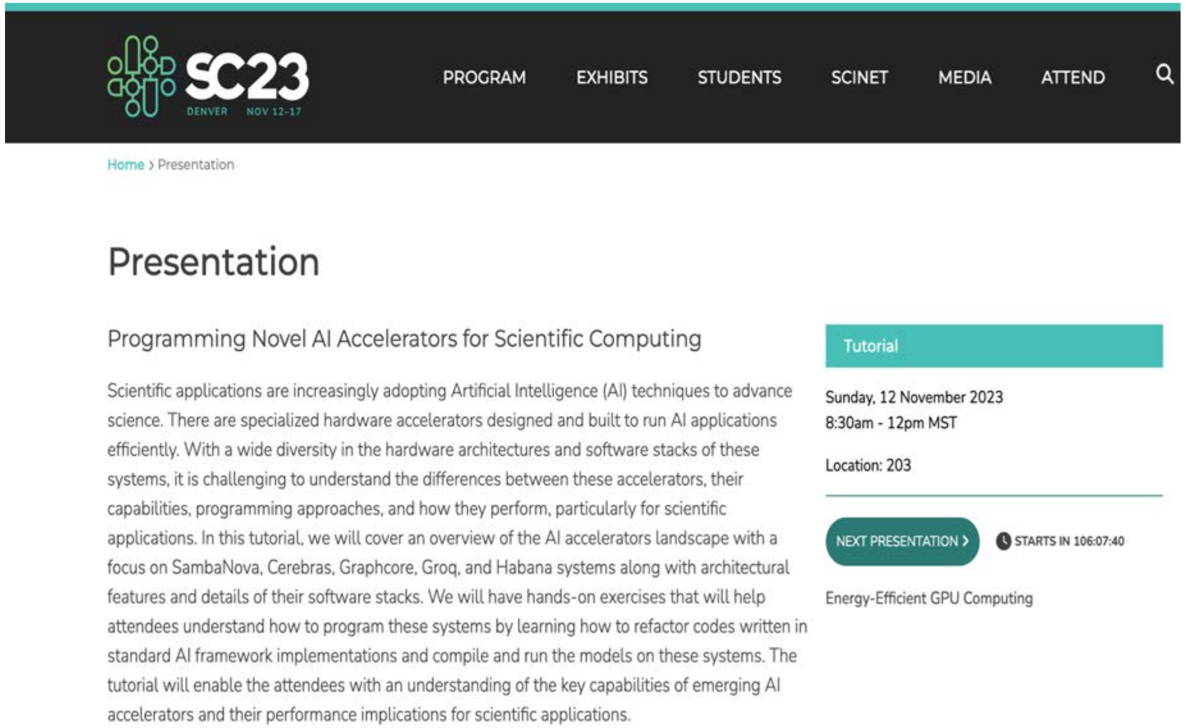
Cerebras CS-2, SambaNova Datascale SN30 and Graphcore Bow Pod64 are available for allocations

[Allocation Request Form]

[AI Testbed User Guide]

AI Testbed Community Engagement

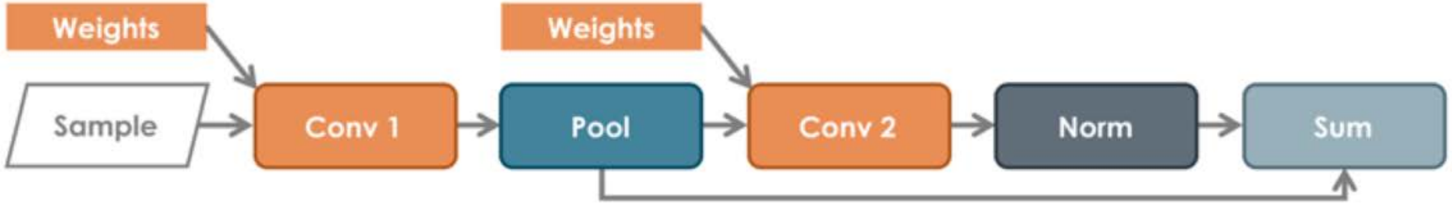




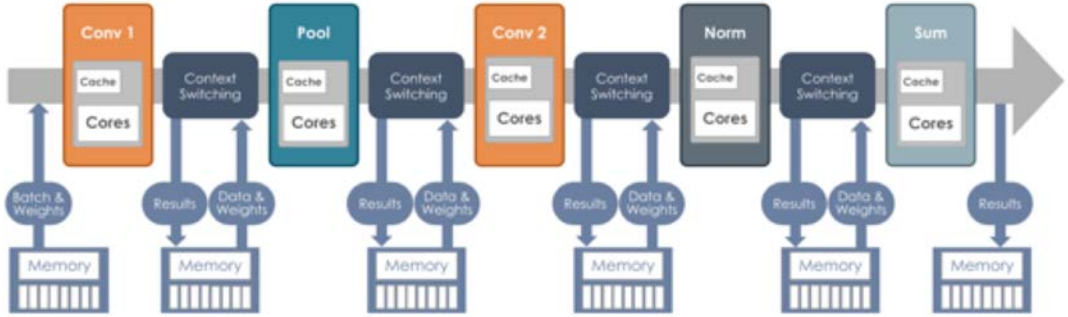
- AI training workshops
  Cerebras: https://events.cels.anl.gov/event/420/
  SambaNova: https://events.cels.anl.gov/event/421/
  Graphcore: https://events.cels.anl.gov/event/422/

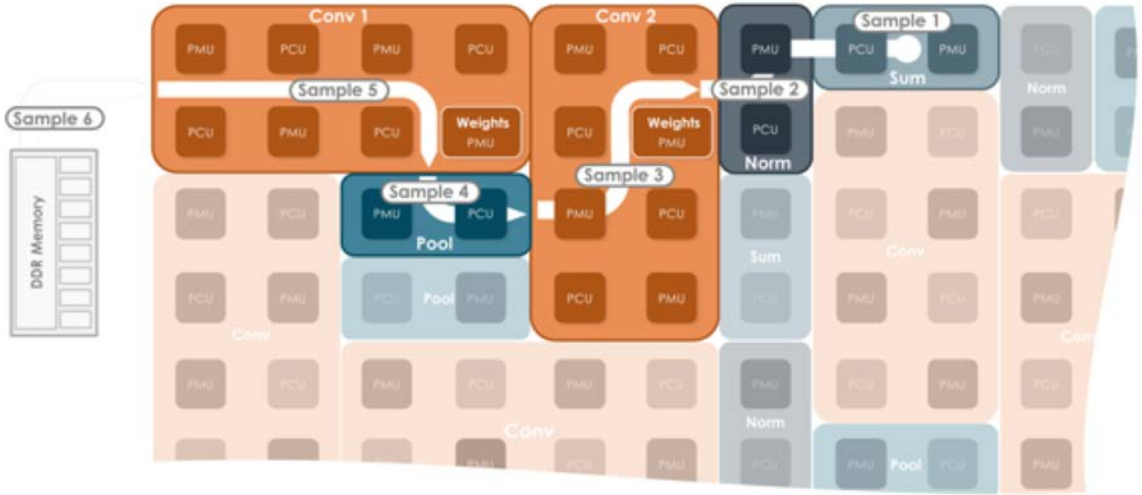
Tutorial at SC23 on Programming Novel AI accelerators for Scientific Computing ***in collaboration with Cerebras, Intel Habana, Graphcore, Groq and SambaNova***

Dataflow Architectures



Simple Convolution Graph



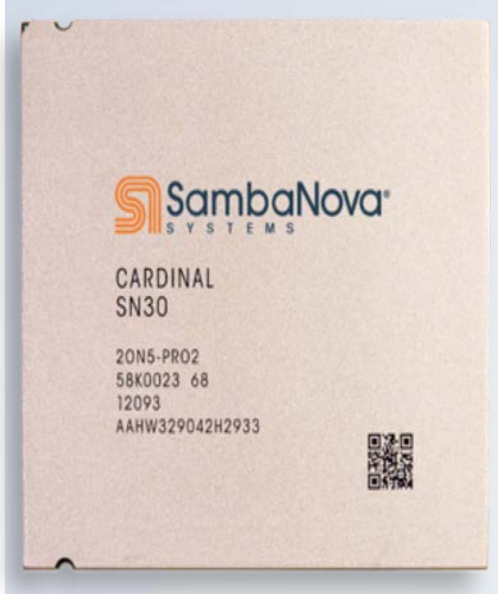


The old way: kernel-by-kernel
Bottlenecked by memory bandwidth and host overhead

The Dataflow way: Spatial
Eliminates memory traffic and overhead

Image Courtesy: SambaNova

SambaNova Cardinal SN30 RDU



Cardinal SN30™
Reconfigurable Dataflow Unit™

- 7nm TSMC, 86B transistors
- 102 km of wire
- 640 MB on-chip, 1,024 GB external
- 688 TFLOPS (bf16)
- RDU-Connect™

as-a-SERVICE
Pre-trained Foundation Models

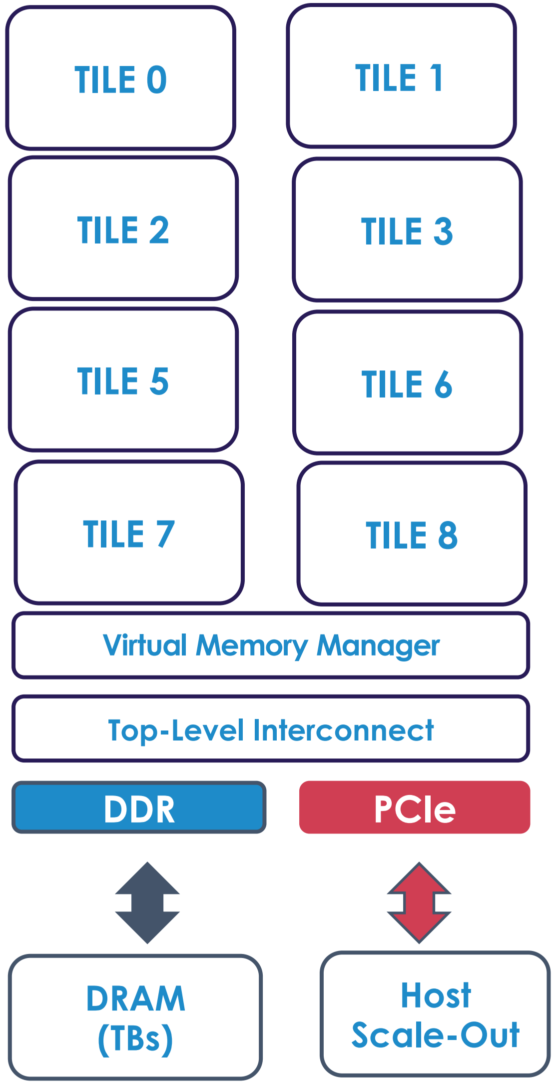
SYSTEMS
DataScale®

SOFTWARE
SambaFlow™

SILICON
RDU

Image Courtesy: SambaNova

Cardinal SN30: 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

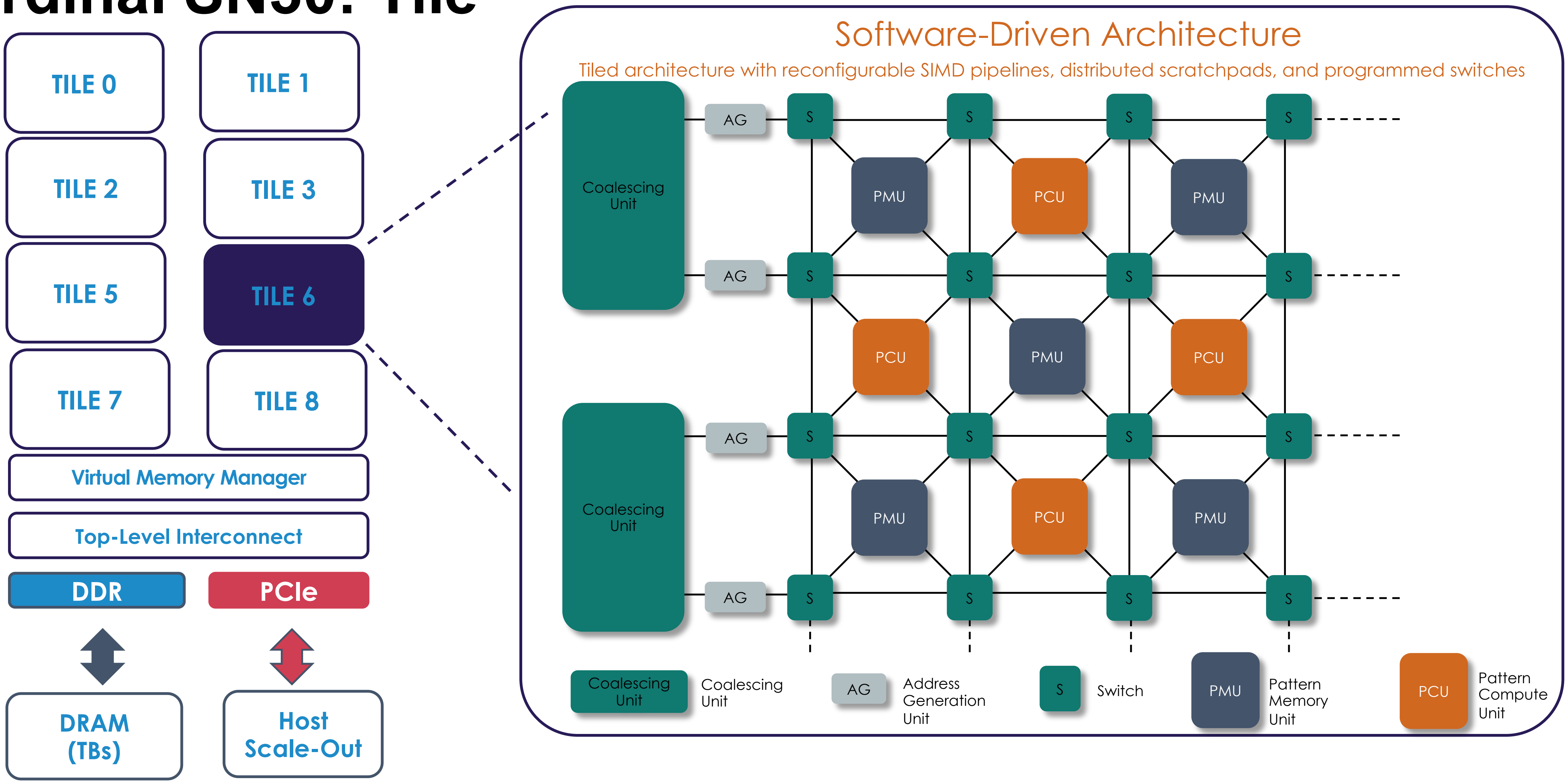


Image Courtesy: SambaNova

Dataflow Architecture for Terabyte Sized Models

DataScale SN30-8R

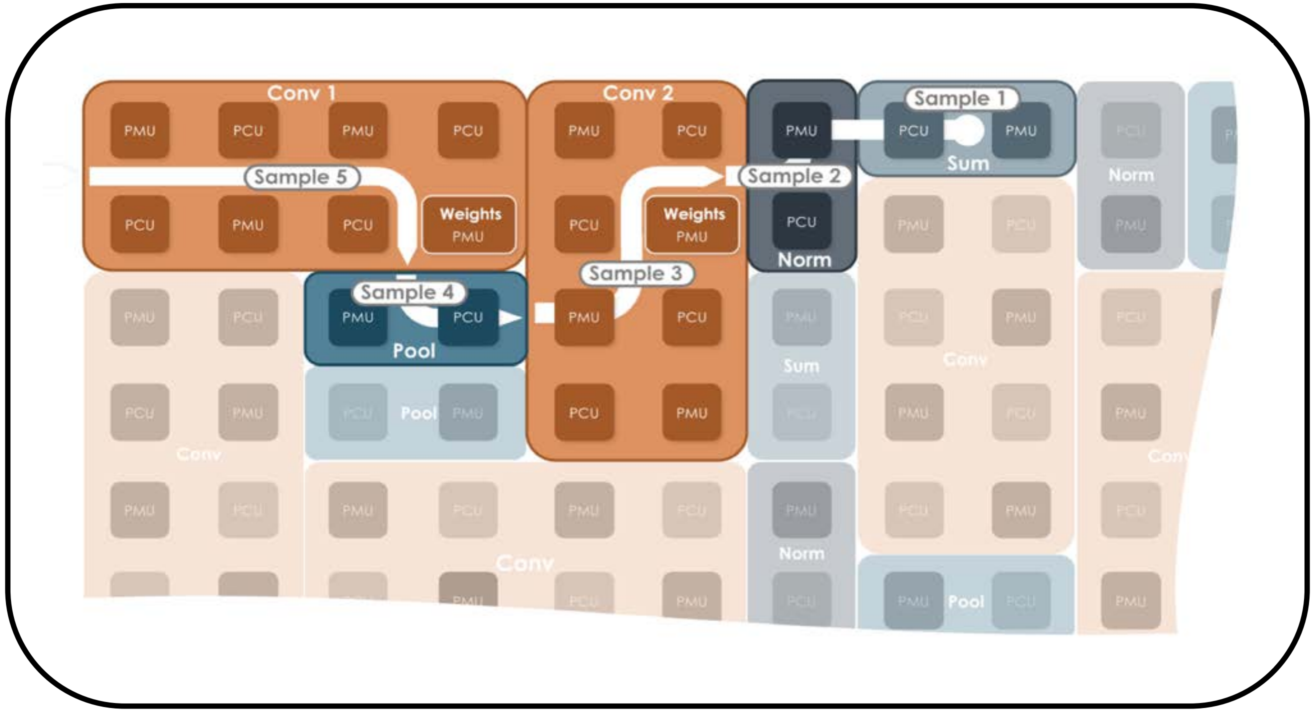
Dataflow Efficiency

+

Compute Capability

+

Large Memory Capacity



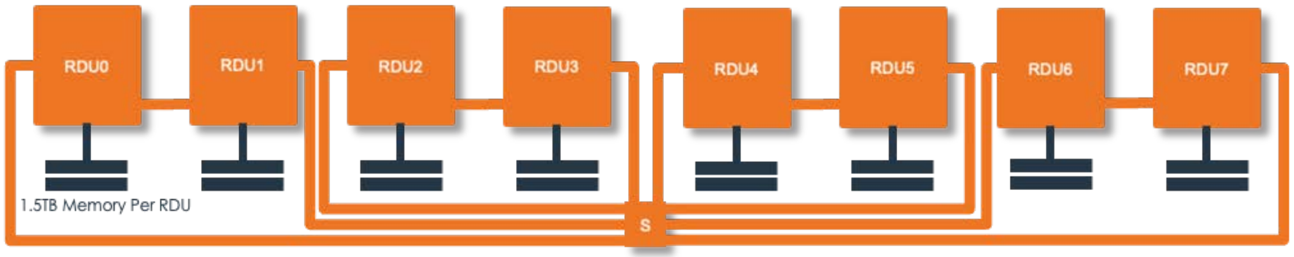


Image Courtesy: SambaNova

SambaNova DataScale SN30-8 System

- 8 x Cardinal SN30 Reconfigurable Dataflow Unit
- 8 TB total memory (using 64 x 128 GB DDR4 DIMMs)
- 6 x 3.8 TB NVMe (22.8 TB total)
- PCIe Gen4 x16
- Host module

Image Courtesy: SambaNova

SambaNova Datascale SN30

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

SambaNova Datascale SN30

- 4 Racks
- 8 nodes of SN30
- 8 RDUs or 4 XRDUs per node
- 8 Tiles per RDU
- Group of 4 tiles

sn30-r1-h1	sn30-r1-h2
sn30-r2-h1	sn30-r2-h2
sn30-r3-h1	sn30-r3-h2
sn30-r4-h1	sn30-r4-h2

SambaFlow Architecture

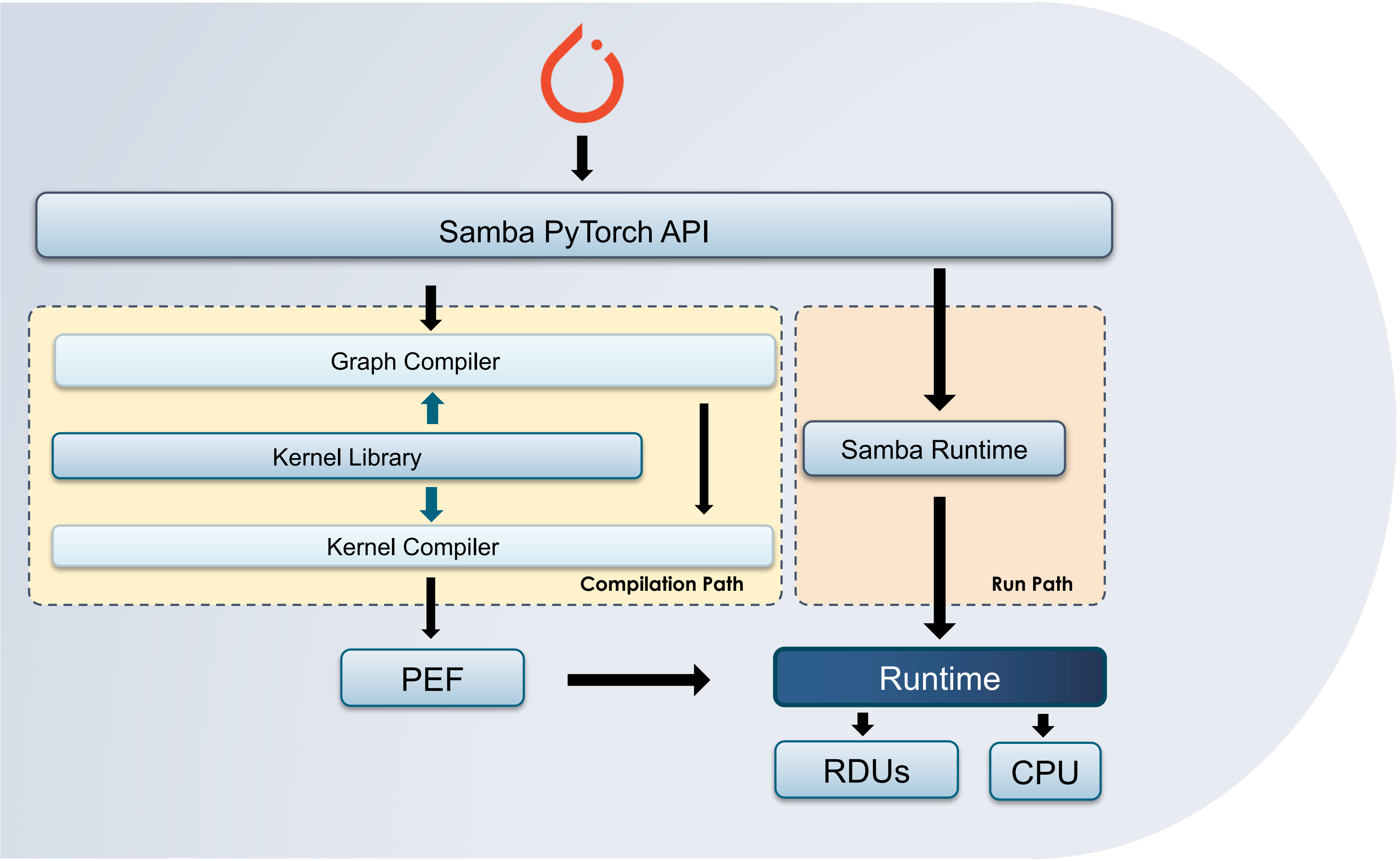


Image Courtesy: SambaNova

Cerebras Wafer-Scale Engine (WSE-2)

850,000 cores optimized for sparse linear algebra

46,225 mm² silicon

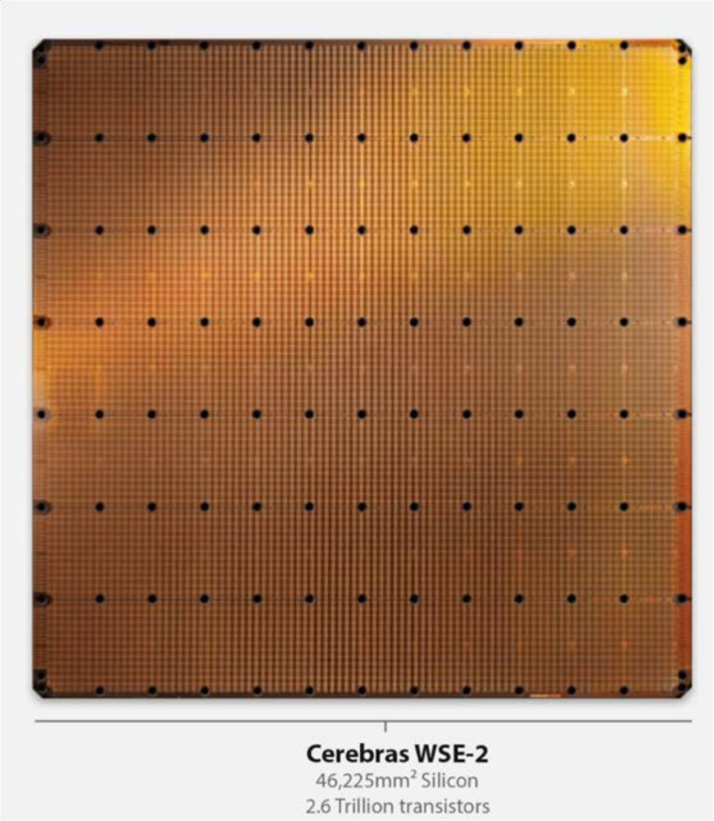
2.6 trillion transistors

40 gigabytes of on-chip memory

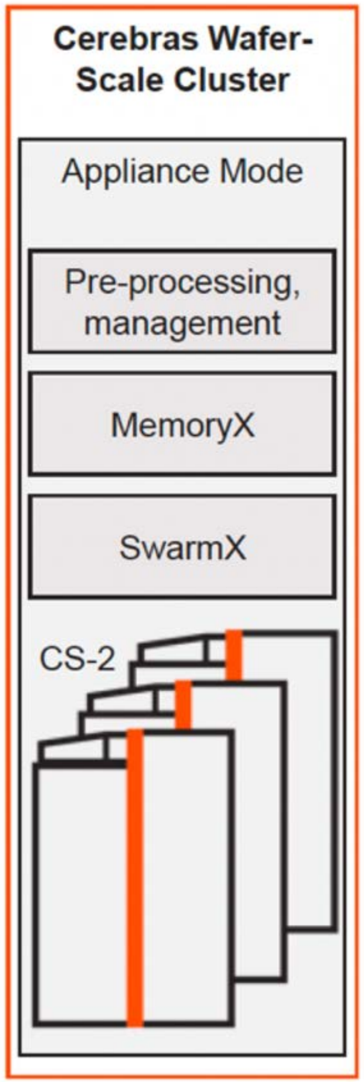
20 PByte/s memory bandwidth

220 Pbit/s fabric bandwidth

7nm process technology



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)

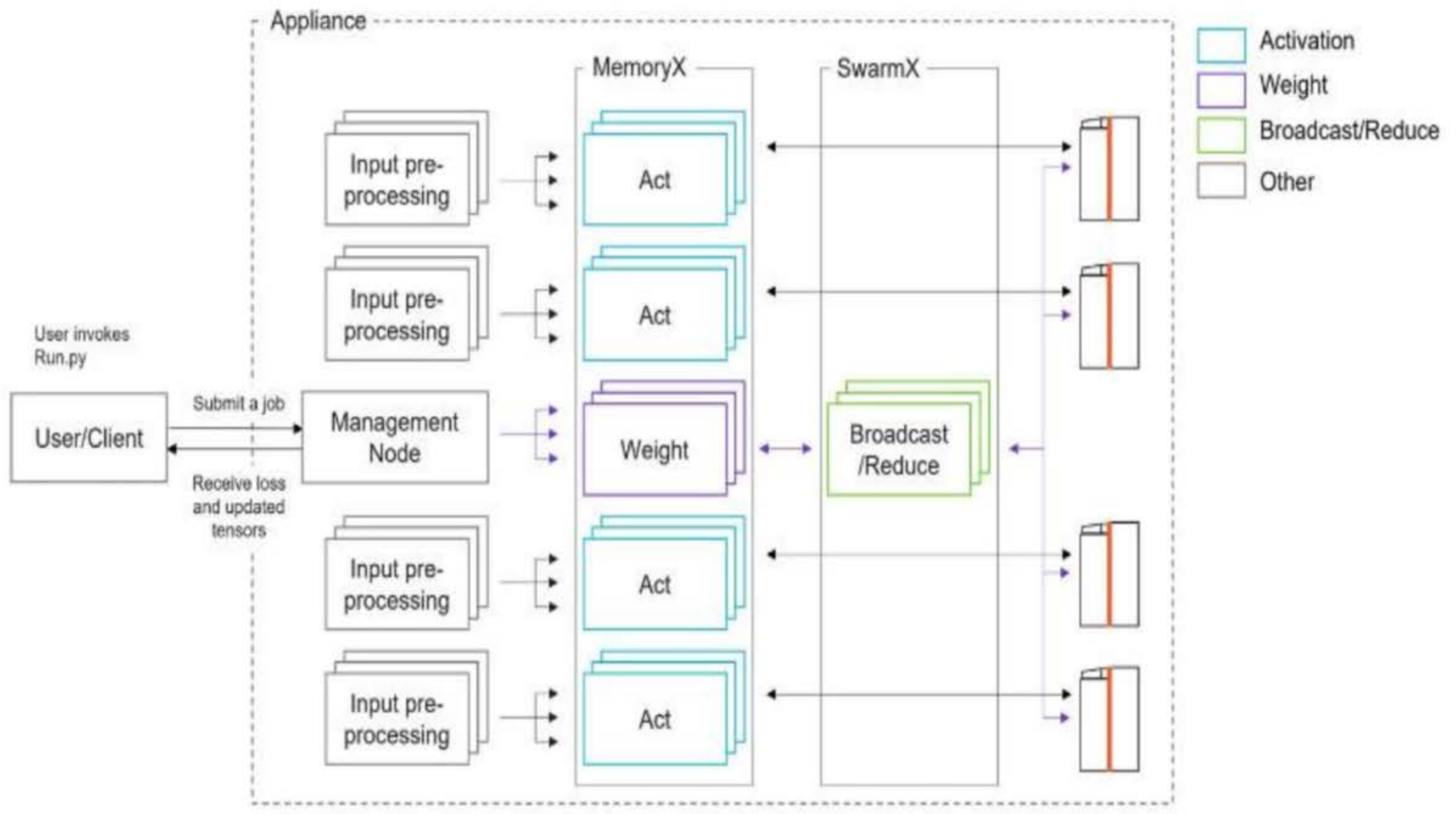
Image Courtesy: Cerebras

Cerebras CS-2 Cluster

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

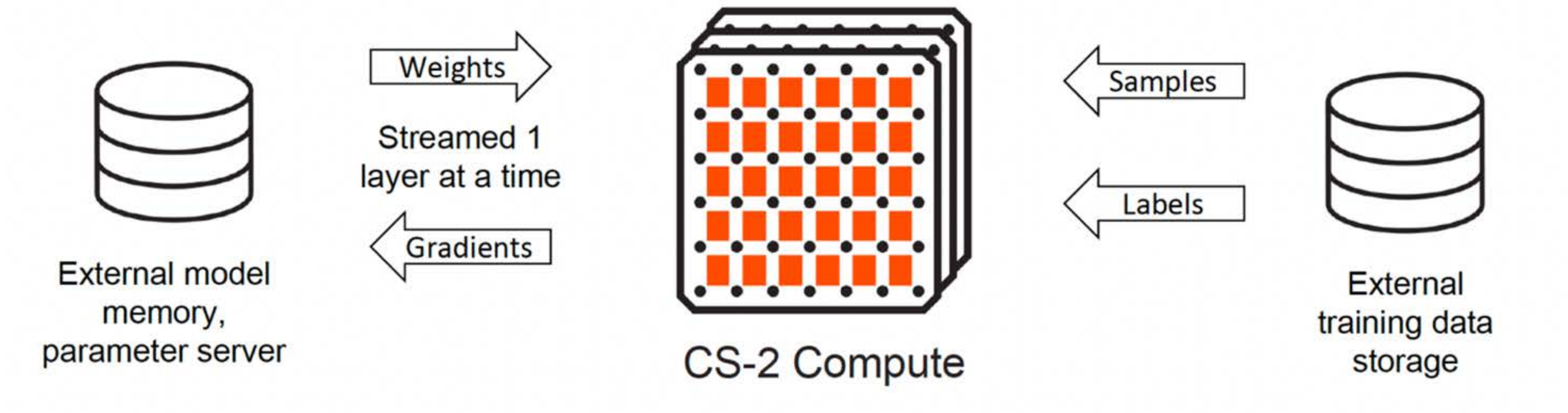
ALCF's CS-2 Cluster

- 2 CS-2 Appliances (each chip 46225 mm^2)
- 1 Management node
- 16 Worker nodes
- 24 MemoryX nodes
- 6 SwarmX nodes
- 3 user login nodes



Topology of a Cerebras Wafer-Scale cluster

Cerebras Weight Streaming Technology

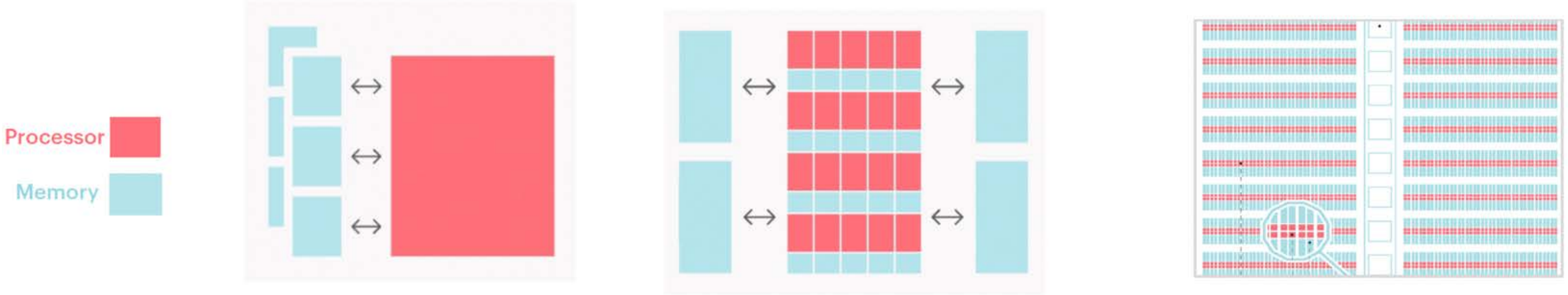


Disaggregate storage and compute
Enable scaling model size

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

BOW IPU

IPU-Tiles™

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

IPU-Core™

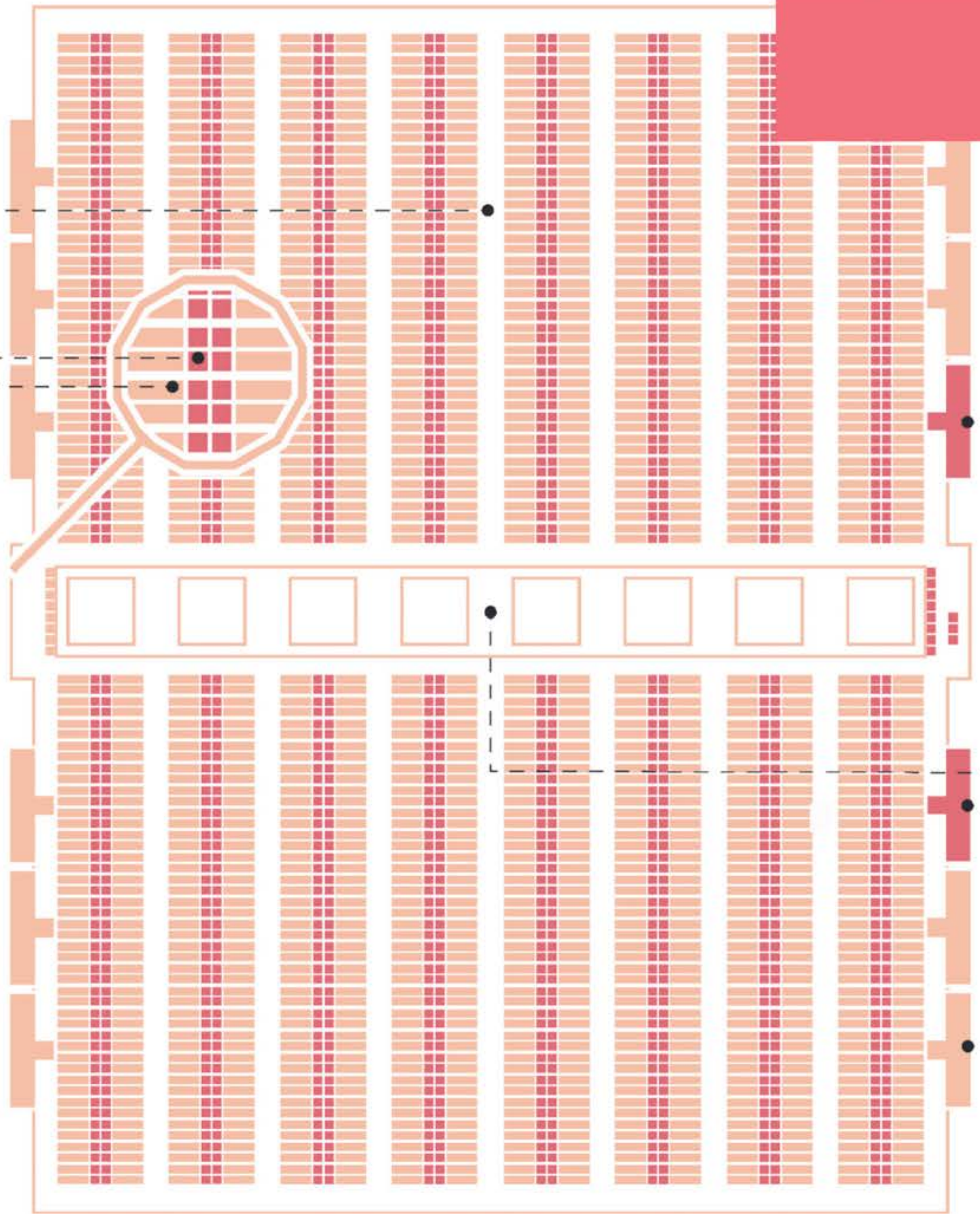
1472 independent IPU-Core™

8832 independent program threads executing in parallel

In-Processor-Memory™

900MB In-Processor-Memory™ per IPU

65TB/s memory bandwidth per IPU



IPU-Exchange™

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

PCIe

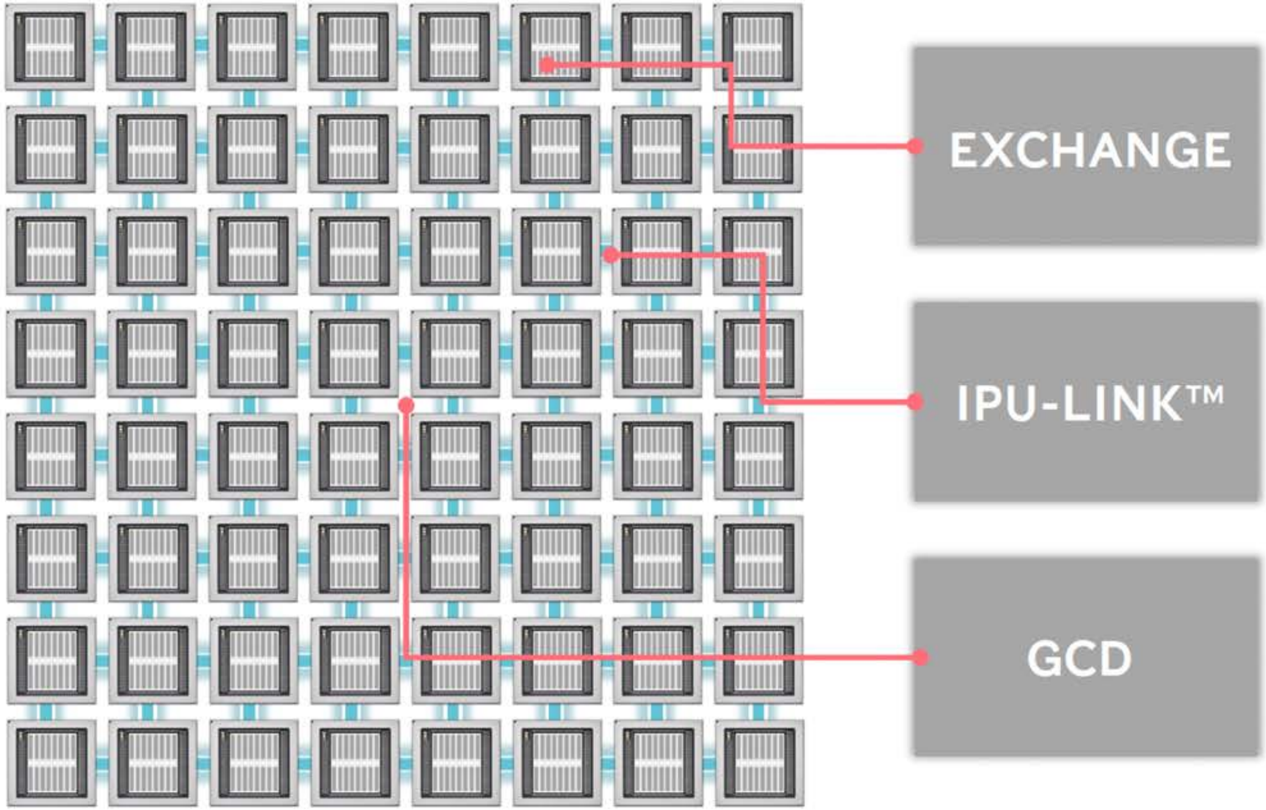
PCI Gen4 x16
64 GB/s bidirectional bandwidth to host

IPU-Links™

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

Slide Courtesy: Graphcore

SCALING ACROSS DEVICES



- THE IPU EXCHANGE MODEL EXTENDS OFF DEVICE
    - ALLOWS TILE MESSAGING ACROSS IPU DEVICES
- SUPPORTS OPTIMIZED CROSS IPU SYNCHRONISATION
- DRIVEN BY COMPILER SUPPORT IN SOFTWARE

- PROVIDES 320 GB/S IPU TO IPU BANDWIDTH
- SUPPORTS COMMUNICATION BETWEEN IPUS
- LAYOUT FULLY SOFTWARE CONFIGURABLE
- SUPPORTS POINT TO POINT TILE MESSAGING

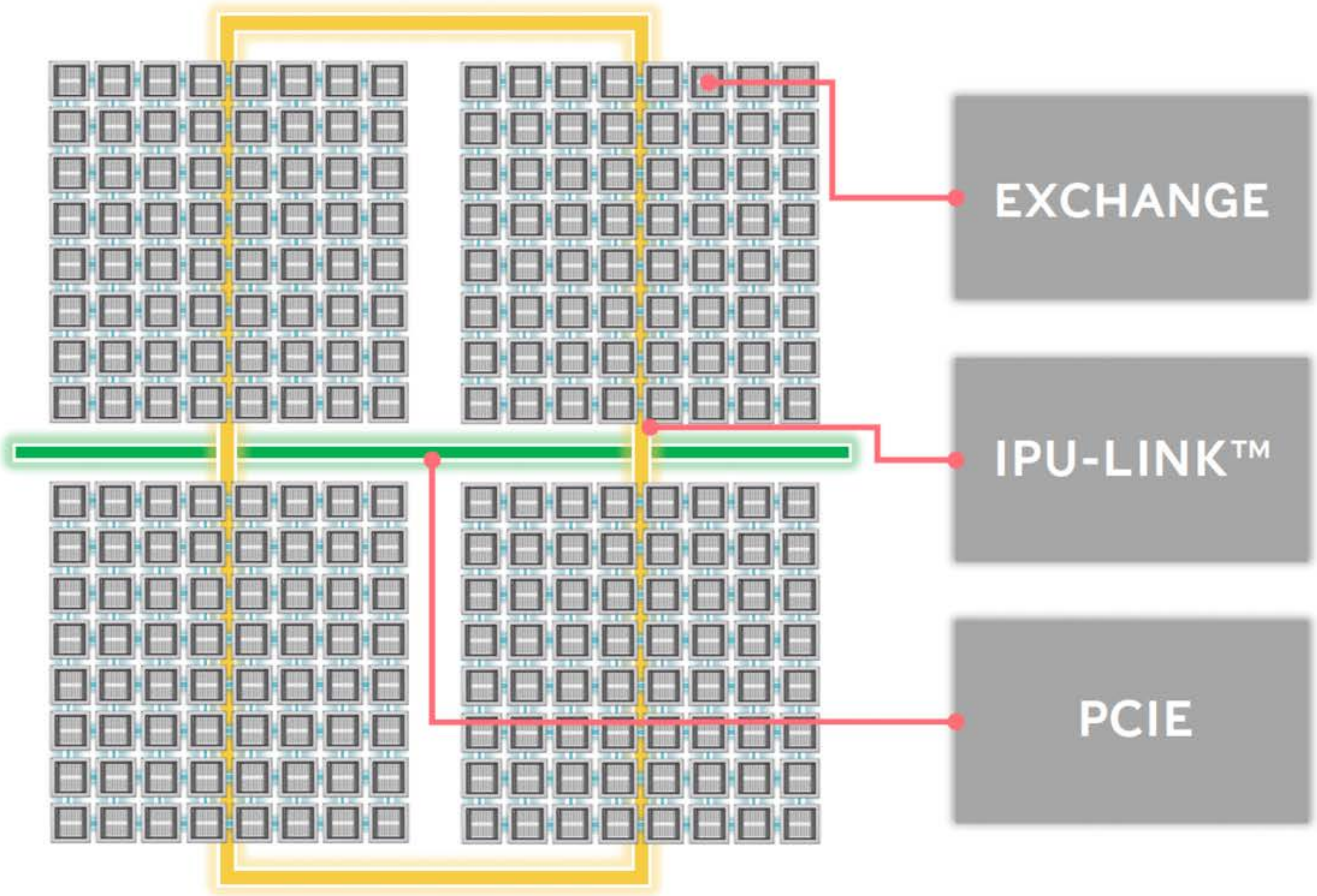
- GRAPH CONTROL DOMAIN FOR APPLICATIONS
- CREATES A SINGLE LARGE IPU SOFTWARE TARGET
- FULLY CONFIGURABLE PARTITIONING OF IPUS
- BOTH DATA PARALLEL AND MODEL PARALLEL

UP TO 64 IPU DEVICES USABLE AS A SINGLE LARGE IPU FROM APPLICATIONS

565248 FULLY INDEPENDENT WORKERS, 57.6GB IN-PROCESSOR MEMORY™, LEVERAGING OVER 3.8 TRILLION TRANSISTORS

Slide Courtesy: Graphcore

SCALING ACROSS SYSTEMS



256 IPU APPLICATION TARGET BUILT FROM INTERCONNECTED 64 IPU DOMAINS

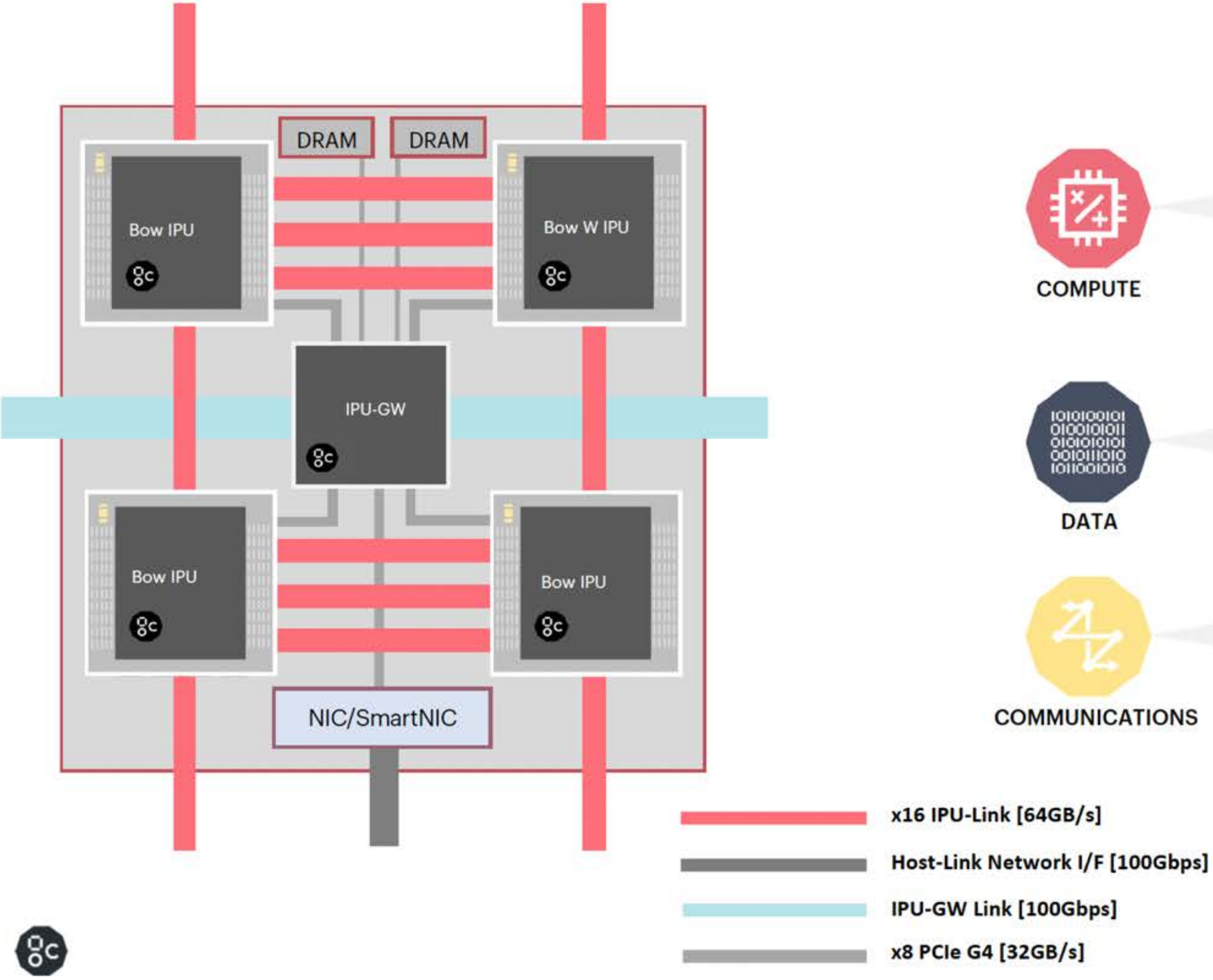
- IPU EXCHANGE SUPPORT ACROSS DOMAINS
  - DRIVEN BY COMPILER SUPPORT IN SOFTWARE
- ENABLES APPLICATION COLLECTIVES SUPPORT
- ALLOWS SCALING UP TO 64000 IPU DEVICES

- IPU LINK™ CAN BE EXTENDED ACROSS DOMAINS
- SUPPORTS OPTIMIZED IPU LINK™ COLLECTIVES
- ALLOWS REPLICATION ACROSS SYSTEMS
- SUPPORTS A STANDARD IPU SOFTWARE MODEL

- IPUS CAN ACCESS MEMORY AND DEVICES OVER PCIE
- ALLOWS INTERFACING WITH HOST BASED SOFTWARE
- APPLICATIONS CAN BUILD ON HOST NETWORKING
- ALLOWS SCALING IN STANDARD SERVER PLATFORMS

Slide Courtesy: Graphcore

BOW-2000: THE BUILDING BLOCK OF LARGE PODS



COMPUTE

4x Bow IPUs

- 1.4 $PFLOP_{16}$ compute
- 5,888 processor cores
- > 35,000 independent parallel threads

DATA

Exchange Memory

- 3.6GB In-Processor Memory @ 260 TB/s
- 128GB Streaming Memory DRAM (up to 256GB) @ 20 GB/s

COMMUNICATIONS

IPU-Fabric managed by IPU-GW

- Host-Link – 100GE to Poplar Server for standard data center networking
- IPU-Link – 2D Torus for intra-POD64 communication
- GW-Link - 2x 100Gbps Gateway-Links for rack-to-rack – flexible topology

Slide Courtesy: Graphcore

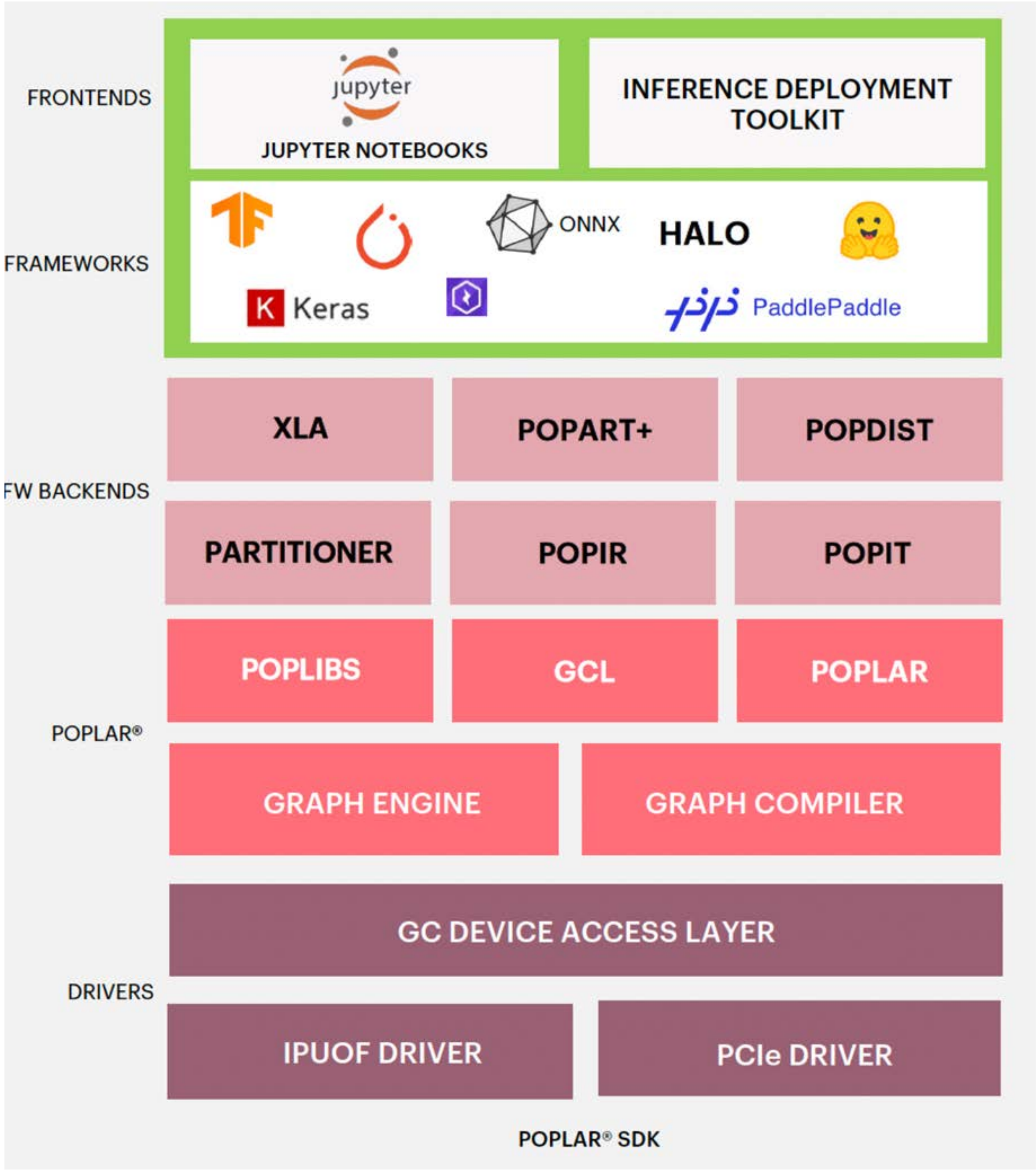
Graphcore POD-64

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

POD64

- 4 Nodes
- 64 IPUs

gc-poplar-01	gc-poplar-02
gc-poplar-03	gc-poplar-04



Slide Courtesy: Graphcore

	Cerebras CS2	SambaNova Cardinal SN30	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 TM	IPU Link	Ethernet-based	NVLink

Challenges

- Understand how these systems perform for different workloads given diverse hardware and software characteristics
- What are the unique capabilities of each evaluated system
- Opportunities and potential for integrating AI accelerators with HPC computing facilities

Approach

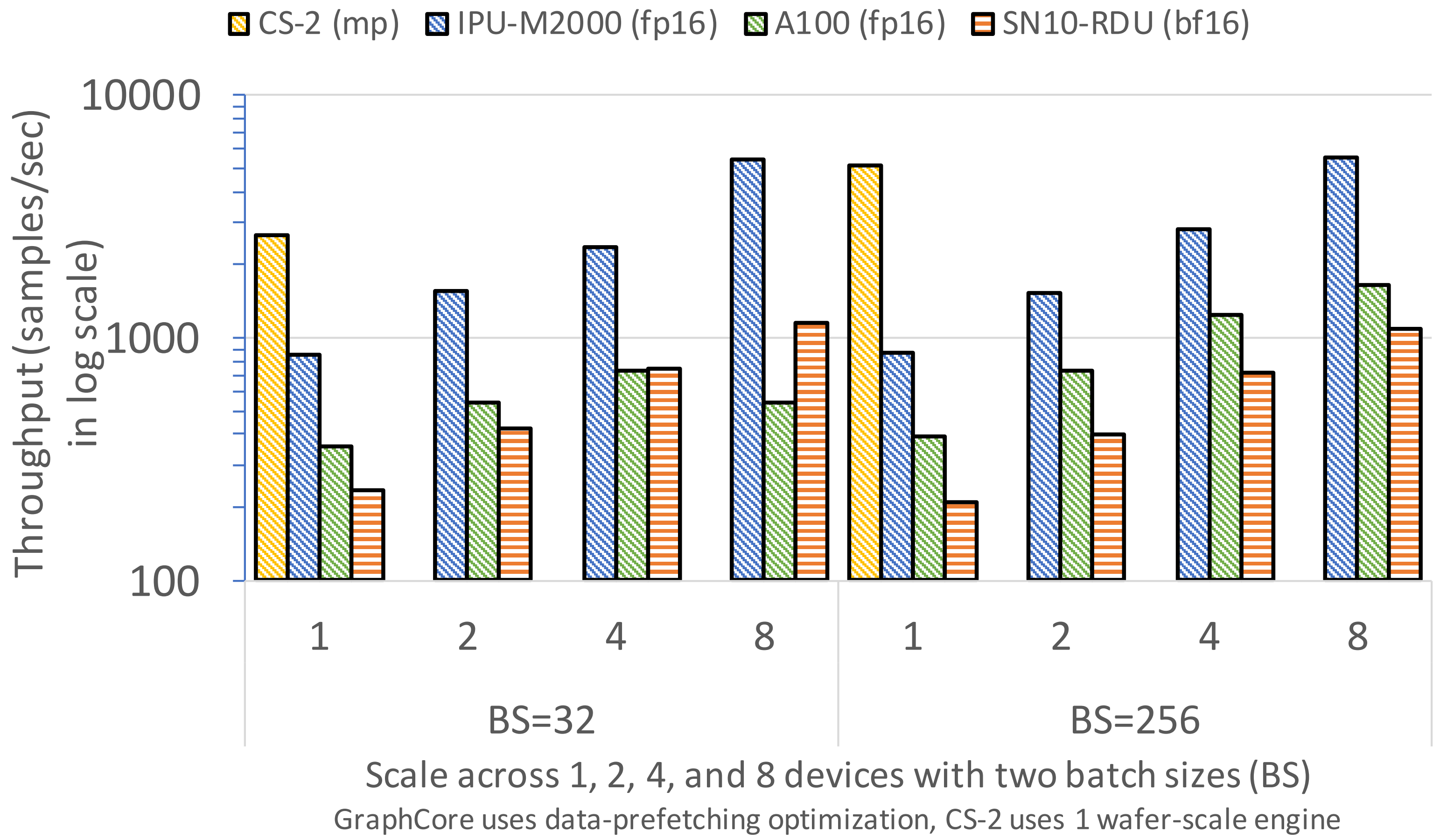
- Perform a comprehensive evaluation with a diverse set of Deep Learning (DL) models:
  - *DL primitives*: GEMM, Conv2D, ReLU, and RNN
  - *Benchmarks*: U-Net, BERT-Large, ResNet-50
  - *AI4S applications*: BraggNN and Uno
  - Scalability and Collective communications

- Evaluated SambaNova, Cerebras, Graphcore, Groq systems and Nvidia A100 as a baseline*

Emani et al. "A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads",
13th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS) at SC 2022.

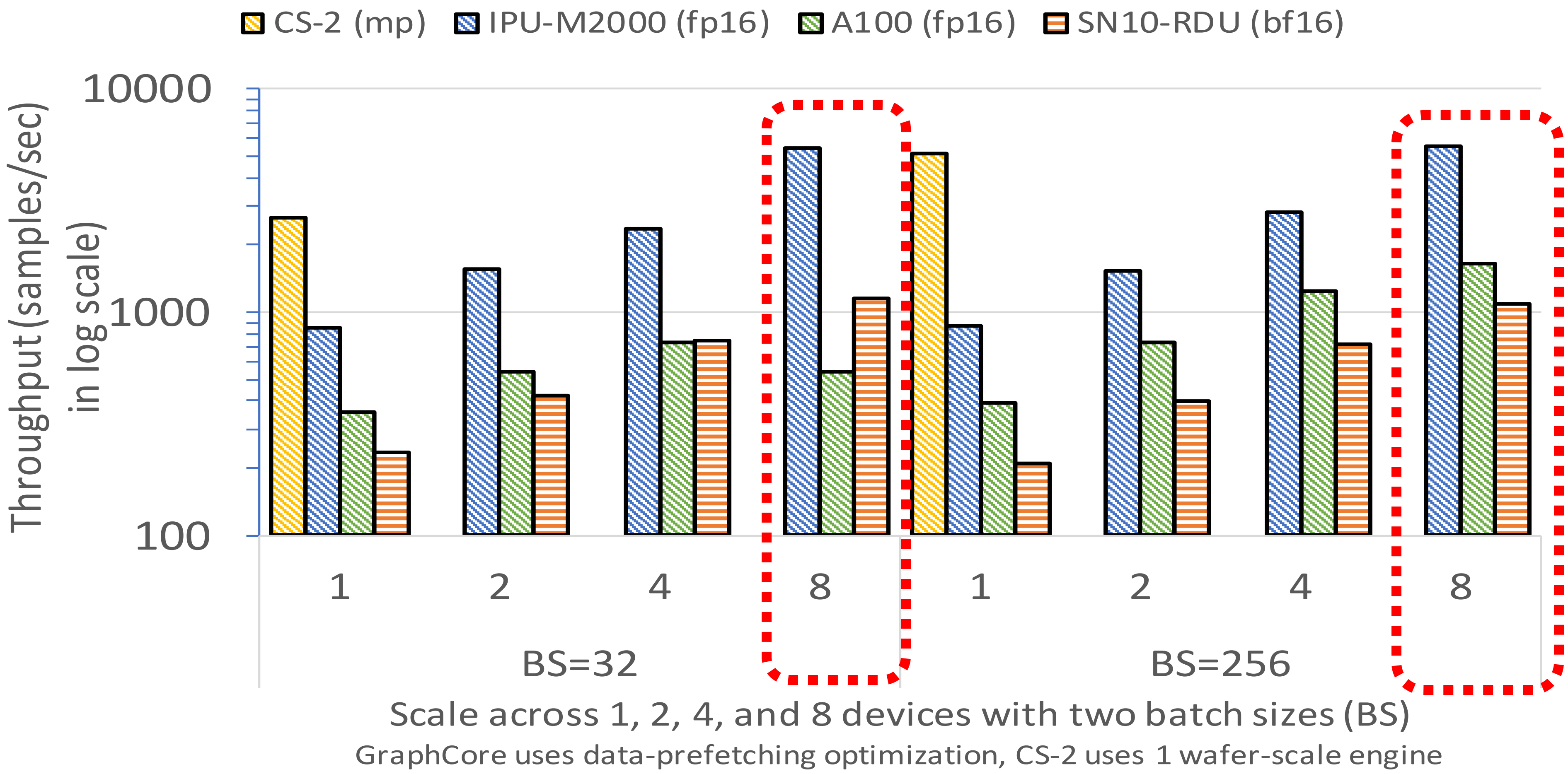
* run out-of-box.

Scaling UNet-2D Training



- 256x256 BMRI dataset
- A100, SN10 – Pytorch
  Graphcore – Tensorflow
  CS2 -- TF estimator
- Accelerators capable of handling much larger image sizes.

Scaling UNet-2D Training

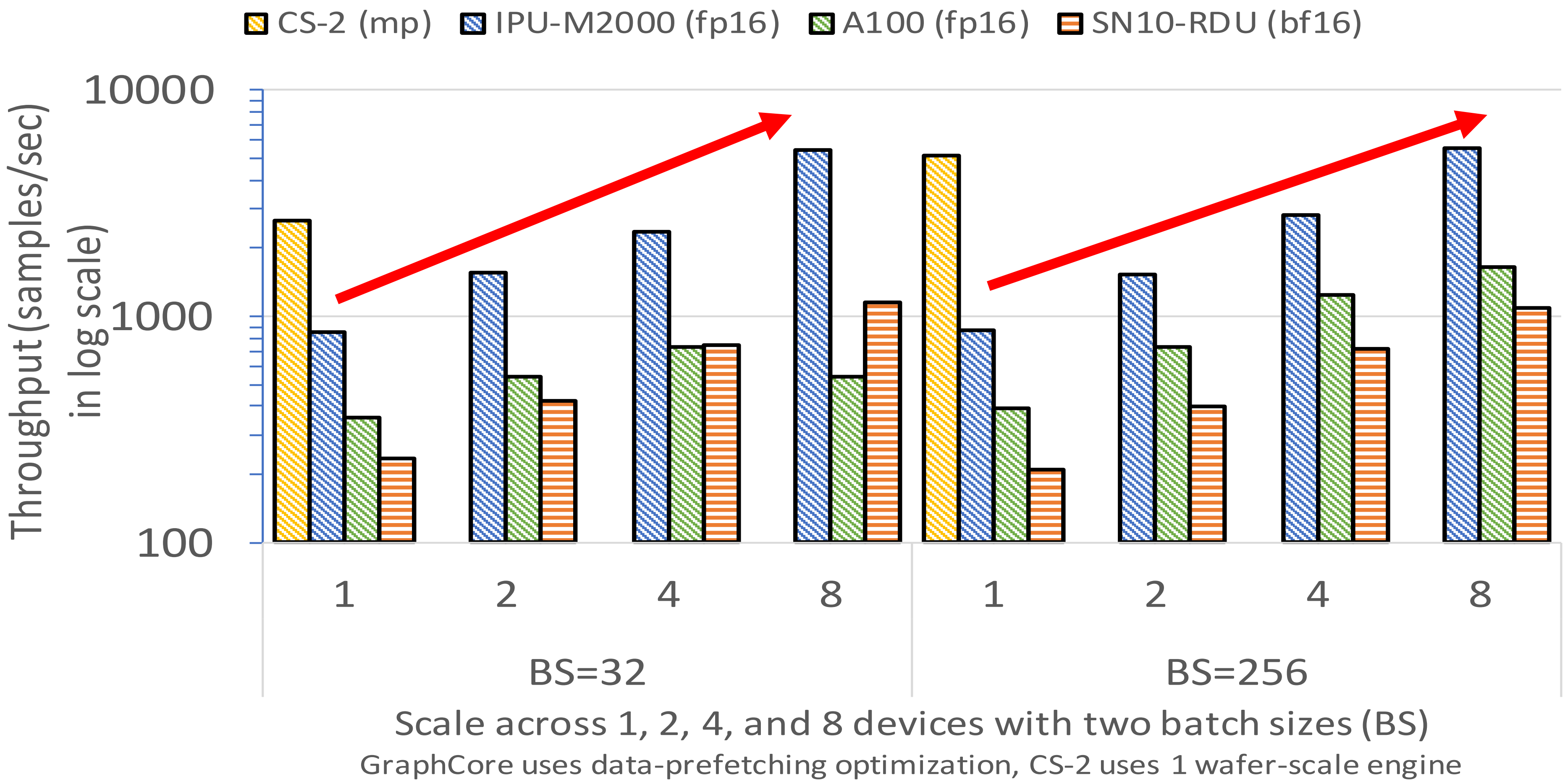


Increased Throughput over 8 A100s

Batch Size	8 SN10 - RDUs	1 CS2	8 GC 200 IPUs
32	2.1x	4.9x	10x

*2x increase in latest sw release

Scaling UNet-2D Training

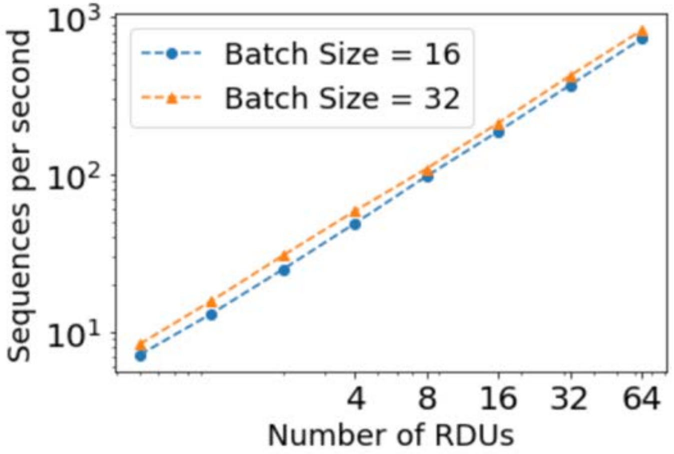


Scaling efficiency

Batch Size	A100	SN10	GC
32	18.8%	42%	79.5%
256	52%	28%	79.6%

GPT Small (1.5B / 2B / xl)

SN30

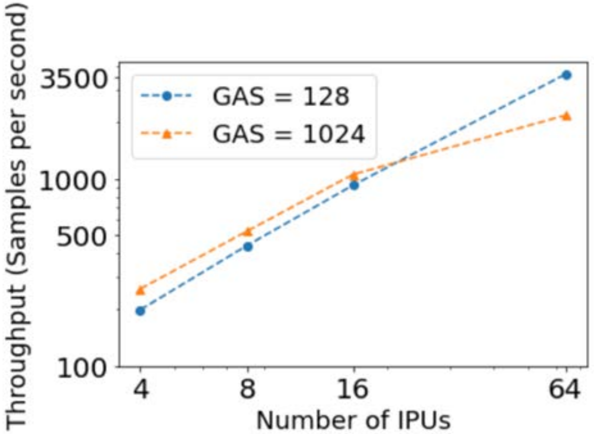


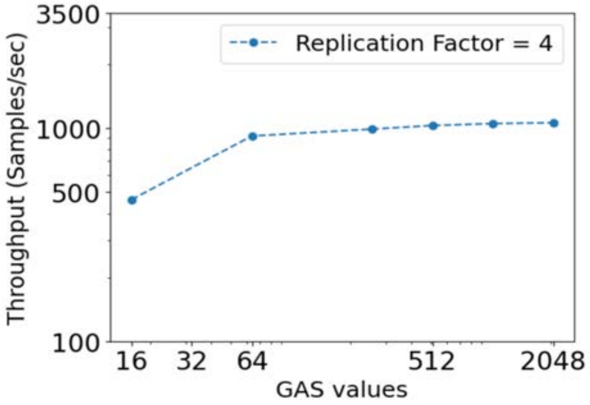
Sequence length = 1024, fits on 4 tiles.

Maximum of 128 instances run on 8 SN30 nodes.

~ 2x speed up observed over the nodes.

Graphcore

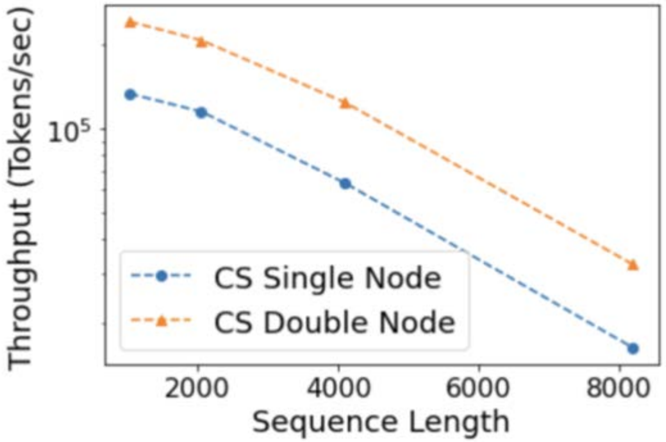




Scaling over replication factor better for smaller GAS values.

poprun for improved scaling efficiency.

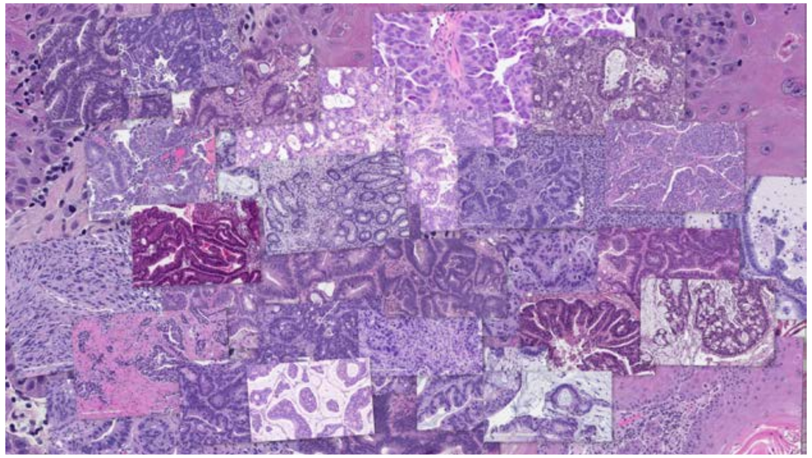
CS2



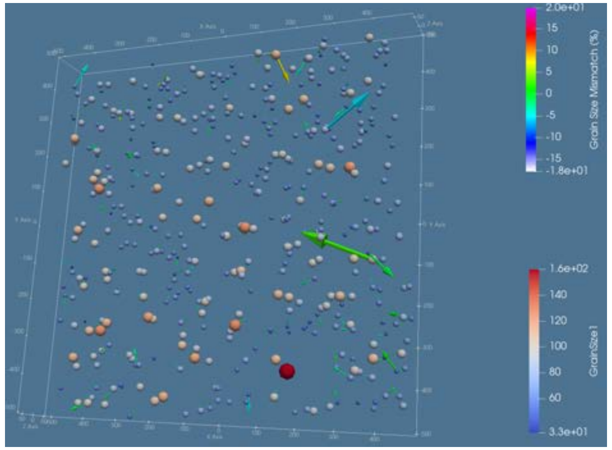
Work in progress for higher sequence length and wider models.

Argonne NATIONAL LABORATORY

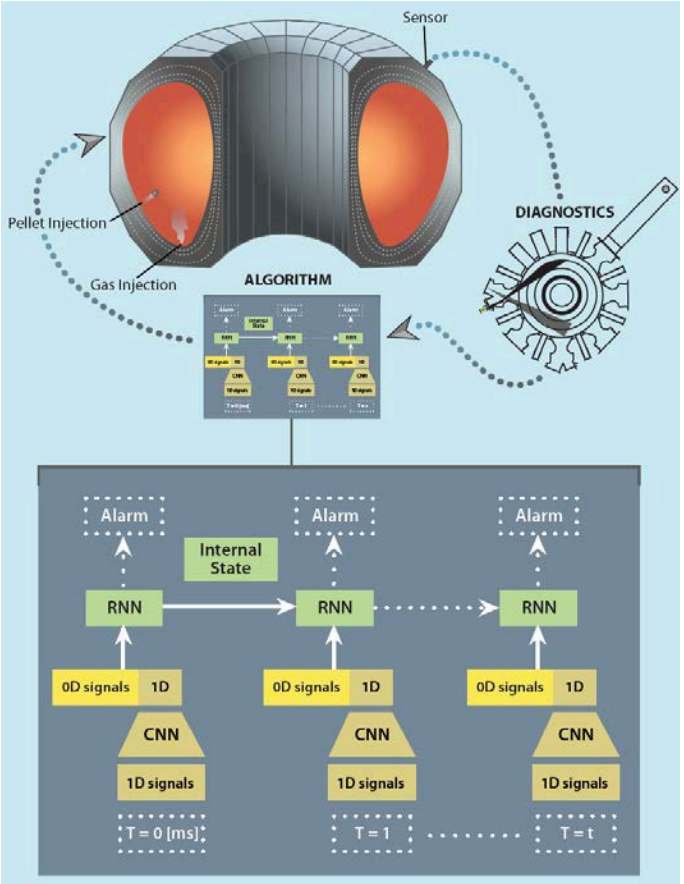
AI FOR SCIENCE APPLICATIONS



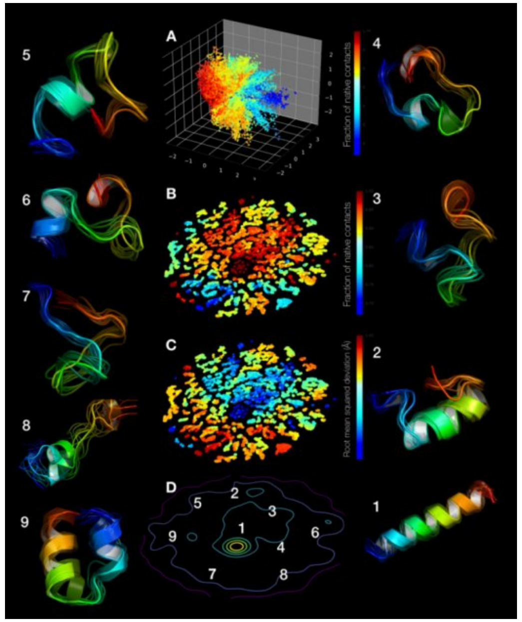
Cancer drug response prediction



Imaging Sciences-Braggs Peak



Tokomak Fusion Reactor operations



Protein-folding(Image: NCI)

and more..

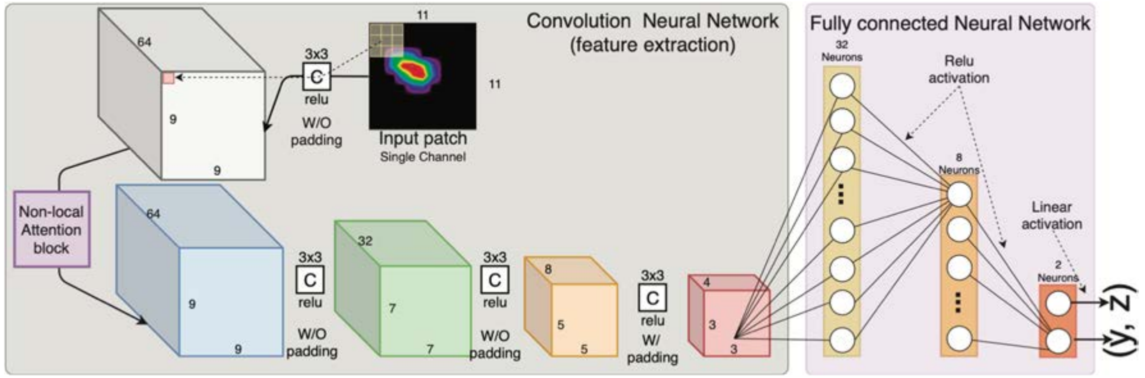
Argonne NATIONAL LABORATORY

Fast X-Ray Bragg Peak Analysis

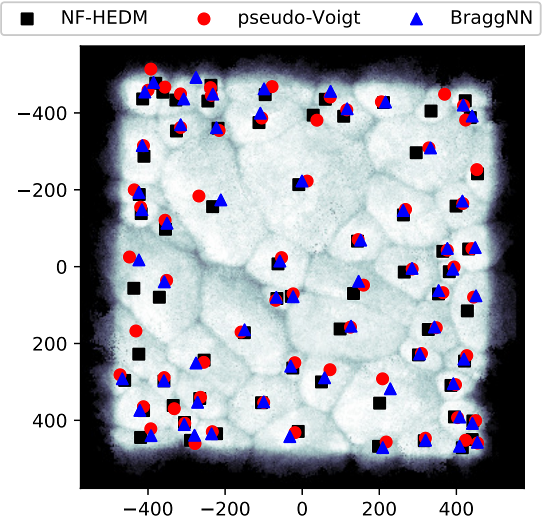
<u>Goal:</u> Enable rapid analysis and real-time feedback during an in-situ experiment with complex detector technologies

<u>Proposed Approach:</u> Deep learning-based method, BraggNN, for massive extraction of precise Bragg peak locations from far-field high energy diffraction microscopy data. BragNN has achieved 200X improvement over conventional pseudo-Voight profiling

<u>Challenges:</u> Model training capability is limited by the hardware



Application of the BraggNN deep neural network to an input patch yields a peak center position (y, z). All convolutions are 2D of size 3 × 3, with rectifier as activation function. Each fully connected layer, except for the output layer, also has a rectifier activation function.



A comparison of BraggNN, pseudo-Voigt FF-HEDM and NF-HEDM. (a) Grain positions from NF-HEDM (black squares), pseudo-Voigt FF-HEDM (red circles) and BraggNN FF-HEDM (blue triangles) overlaid on NF-HEDM confidence map

Courtesy: Z. Liu et al. BraggNN: Fast X-ray Bragg Peak Analysis Using Deep Learning. International Union of Crystallography (IUCrJ), Vol. 9, No. 1, 2022

Fast X-Ray Bragg Peak Analysis

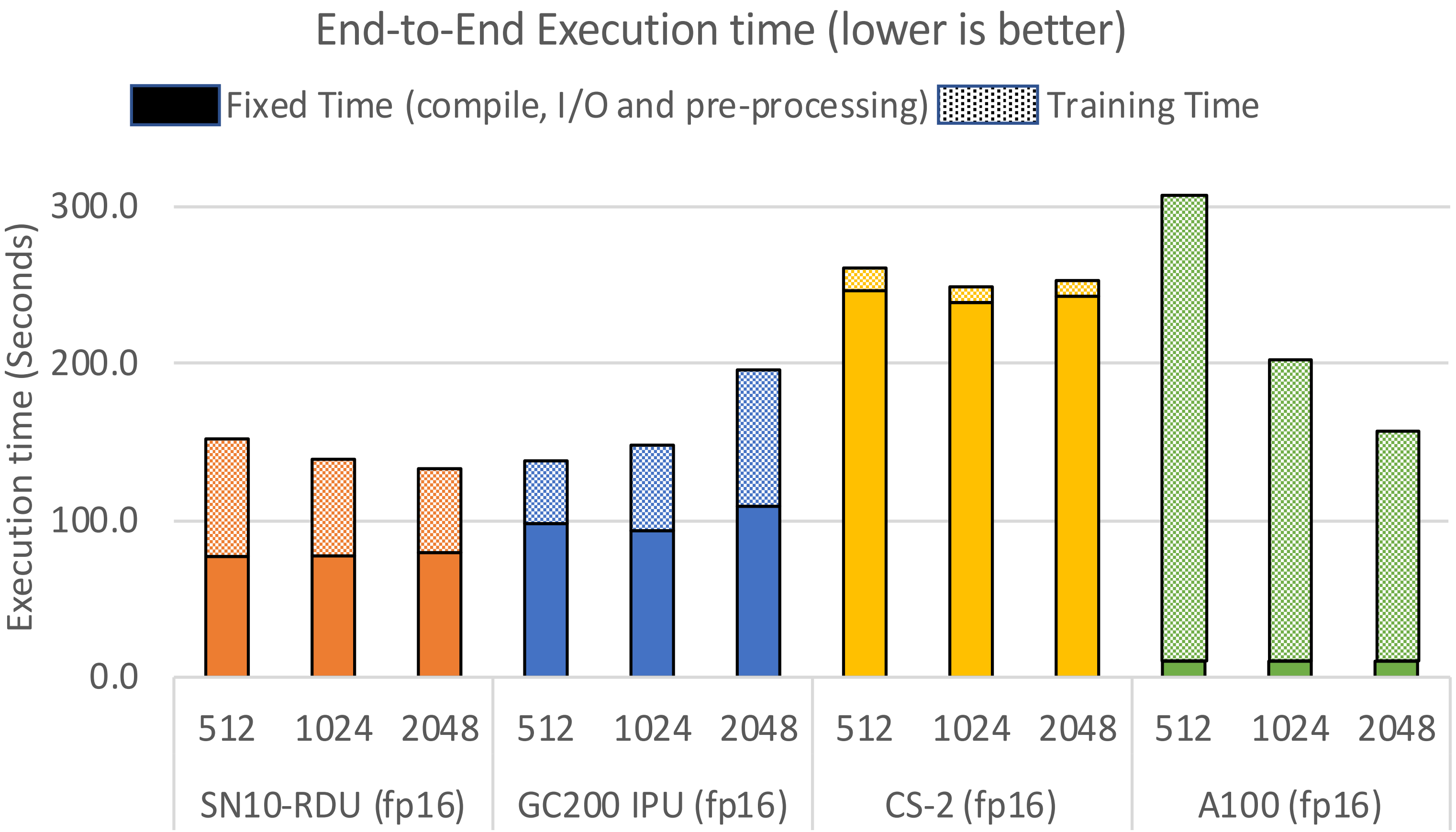


TABLE II: BraggNN Throughput (in order of 1k samples/sec) with various batch sizes (BS)

System	BS=512	BS=1024	BS=2048
CS-2 (FP16)	1365.4	2463	2787.9
GC200 IPU (FP16)	478.0	350.6	219.9
SN10 RDU (BF16)	369.7	449.8	518
A100 (FP16)	53.9	65.5	73.7

- SambaNova and Graphcore achieve lowest time to solution and achieve up to 1.55x and 1.46x speedup in comparison to Nvidia A100 respectively.
- Cerebras achieves up to 37.8x throughput improvement over A100.

Genome-scale Language Models (GenSLMs)

Goal:

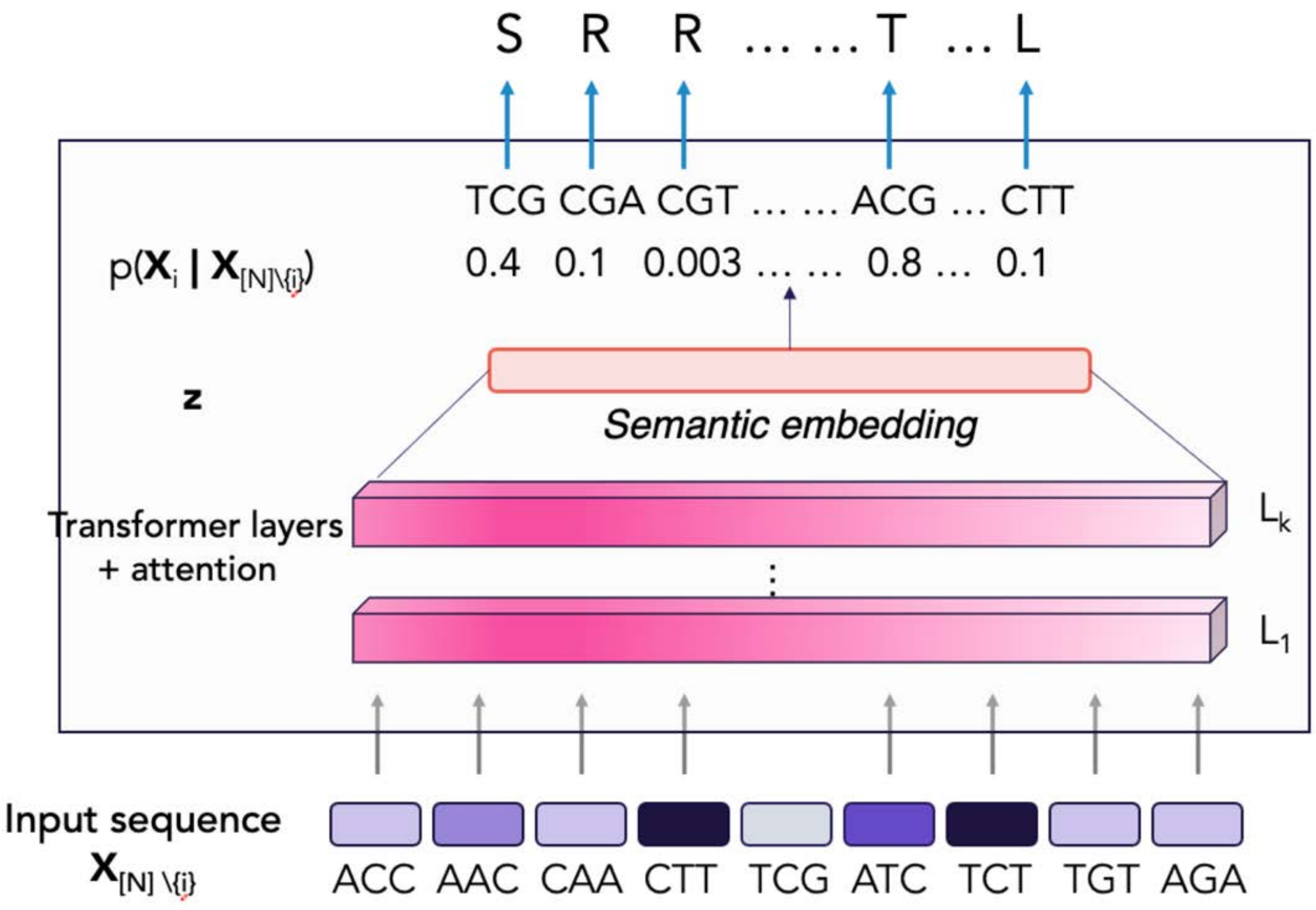
- How new and emergent variants of pandemic causing viruses, (specifically SARS-CoV-2) can be identified and classified.
- Identify mutations that are VOC (increased severity and transmissibility)
- Extendable to gene or protein synthesis.

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

Genome-scale Language Models (GenSLMs)



Model	Seq. length	#Parameters	Dataset
GenSLM-Foundation	2048	25M, 250M, 2.5B, 25B	110M
GenSLM	10240	25M, 250M, 2.5B, 25B	1.5M
GenSLM-Diffusion	10240	2.5B	1.5M

Challenges

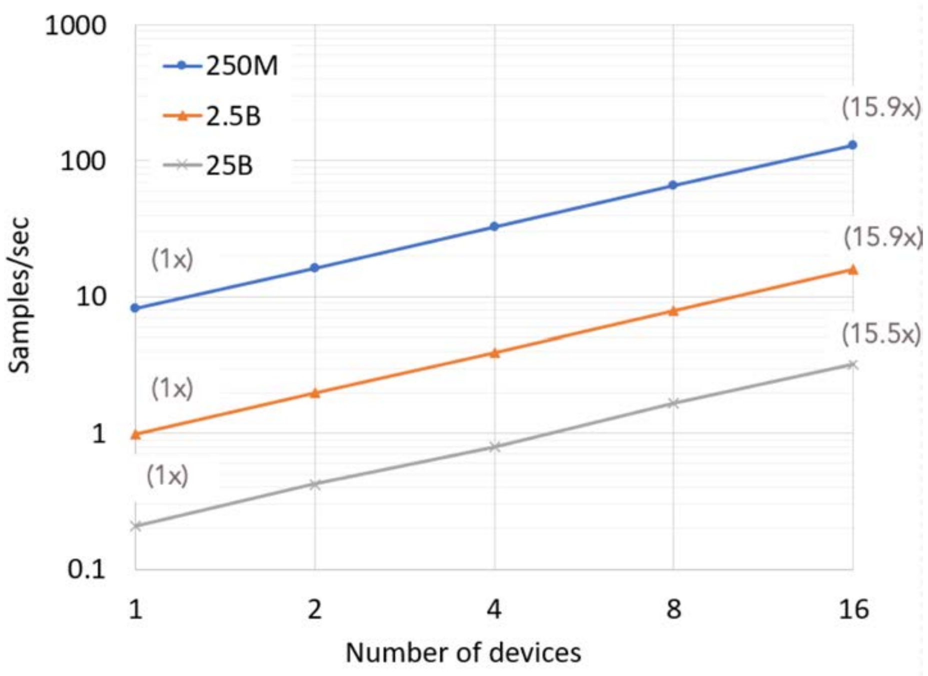
Scaling LLMs with 25B parameters:

- O (L^2) complexity in the attention computation
- Overcome communication overheads
- Sharding and the training time available on GPUs imposing limitations

Solution

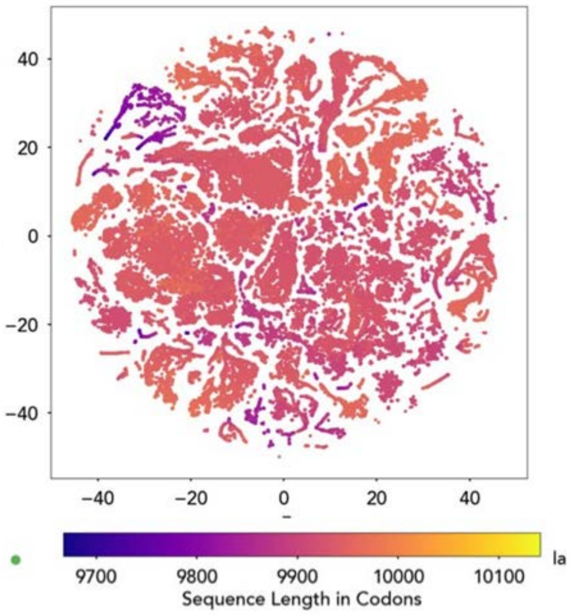
Cerebras CS-2 wafer-scale cluster and Sambanova SN30 enables pre-training and finetuning.

GenSLMs on CS2

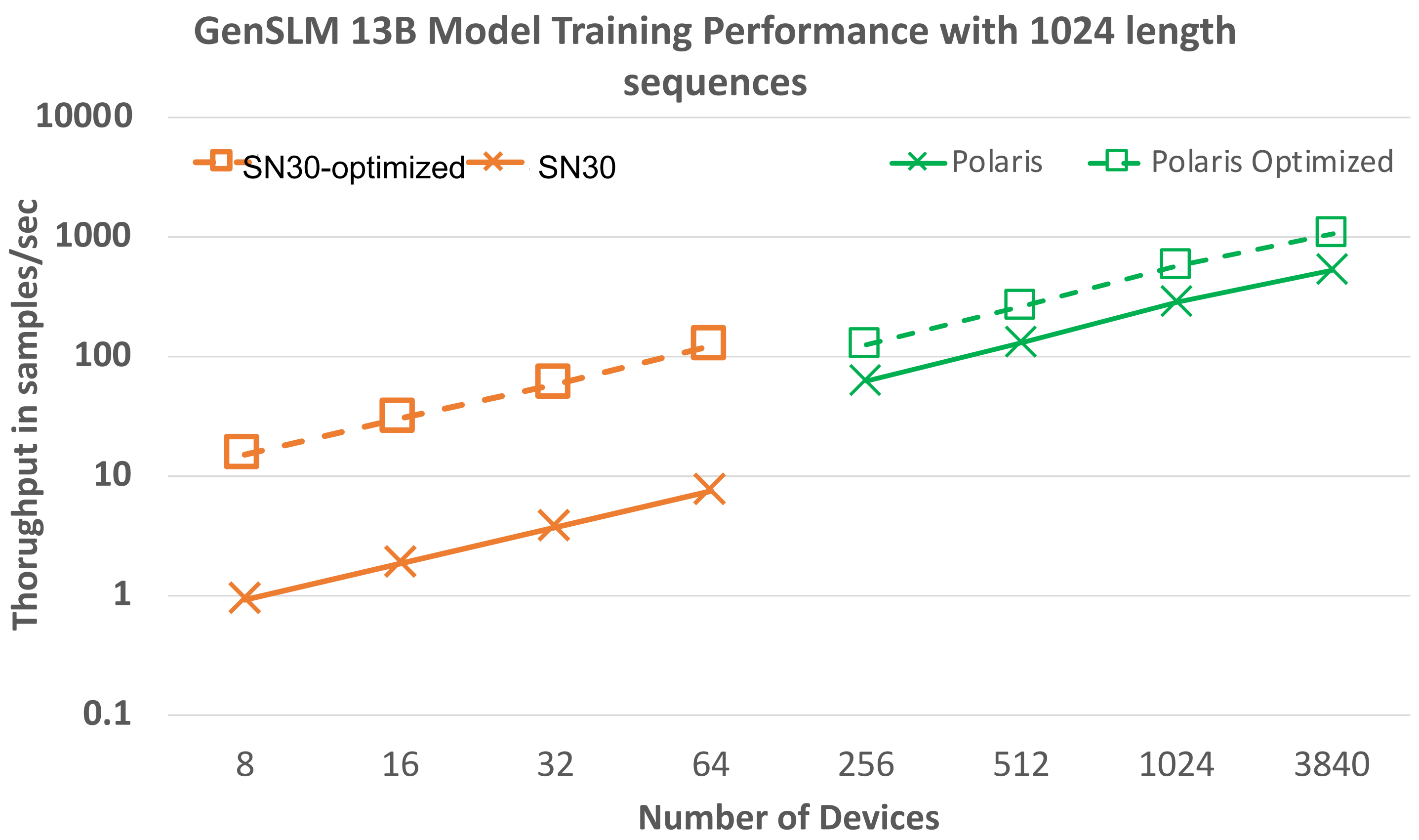


- Sequence Length = 10,240
- Trainable upto GPT3-13b model.
- Training with 4CS2, less than ½ day

	GenSLM 123M		GenSLM 1.3B	
	1 CS-2	4 CS-2	1 CS-2	4CS-2
Training steps	5,000	3,000	4,500	3,000
Training samples	165,000	396,000	49,500	132,000
Time to train (h)	**4.1**	**2.4**	**15.6**	**10.4**
Validation accuracy	0.9615	0.9625	0.9622	0.9947
Validation perplexity	1.031	1.029	1.031	1.025



GenSLMs on SN30



- Sequence Length = 1024
- Model Size 13B
- Achieves linear scaling across nodes.
- SN30 performance similar to 4 A100 on 1.17 release.
- Optimized on 1.18 to get 10x speed-up.
- Pretraining and FineTuning on larger sequence lengths.

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

Observations, Challenges and Insights

- Good progress made in integration of AI accelerators, in production, at a national user facility and significant more work is needed for effective coupling

- Training and Outreach is critical to educate users to effectively use AI systems

- Close collaboration with vendors is necessary to realize the vision of AI for science

Ongoing Efforts

- 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
- Evaluate traditional HPC on AI Accelerators
- Understand how to integrate AI accelerators with ALCF’s existing and upcoming supercomputers to accelerate science insights

Recent Publications

- **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*

- **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

Recent Publications

- **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, Hyenseung 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.

*** Fiinalist 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.

- Venkatram Vishwanath, Michael Papka, William Arnold, Bruce Wilson, Varuni Sastry, Sid Raskar, Zhen Xie, Rajeev Thakur, Anthony Avarca, Arvind Ramanathan, Alex Brace, Zhengchun Liu, Hyunseung (Harry) Yoo, Corey Adams, Ryan Aydelott, Kyle Felker, Craig Stacey, Tom Brettin, 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.

Please reach out for further details
Venkat Vishwanath, Venkat@anl.gov
Murali Emani, memani@anl.gov