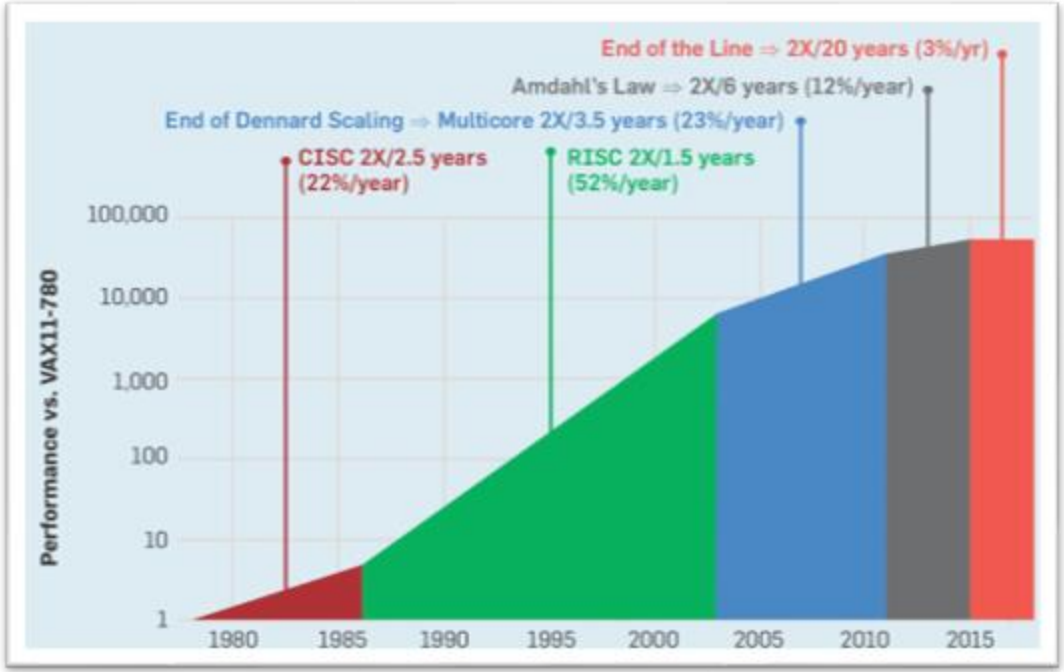
12TH NOVEMBER, 2024

PROGRAMMING NOVEL AI ACCELERATORS

SIDDHISANKET RASKAR
Assistant Computer Scientist
Argonne Leadership Computing Facility
sraskar@anl.gov

Motivation

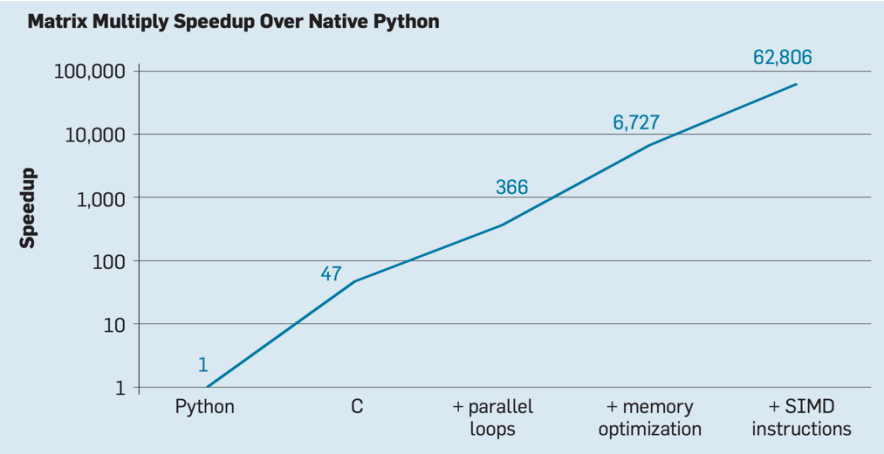


Growth of computer performance

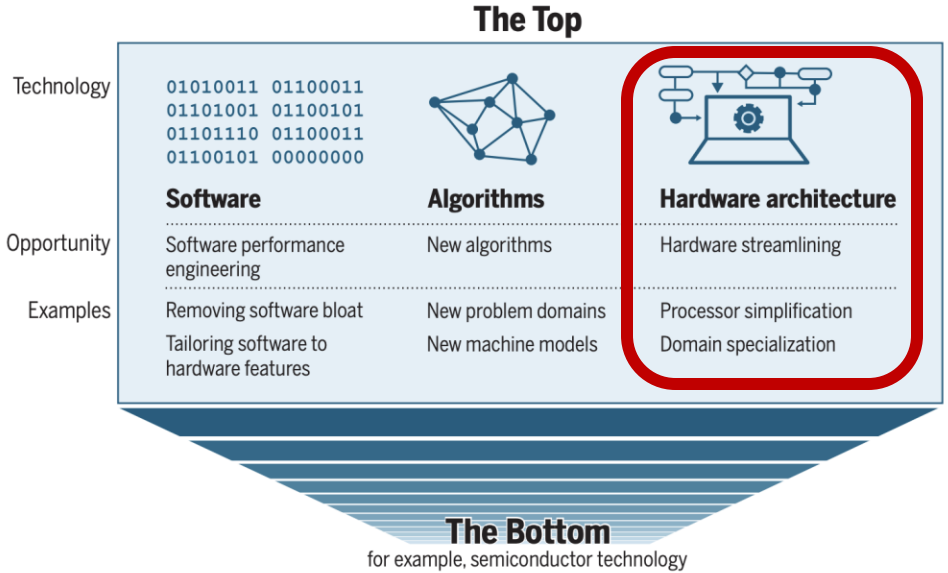
An era without Dennad's scaling along with reduced Moore's law and Amdahl's law is in full effect.

John L. Hennessy and David A. Patterson. 2019. A new golden age for computer architecture. Commun. ACM 62, 2 (February 2019), 48–60.

Motivation



Better Software and algorithms



Domain Specific Architectures and Languages

Charles E. Leiserson et al. ,There's plenty of room at the Top: What will drive computer performance after Moore's law?. Science368, eaam9744(2020).

ALCF AI Testbed

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

Cerebras
CS-2

SambaNova
DataScale SN30

Graphcore
Bow Pod64

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

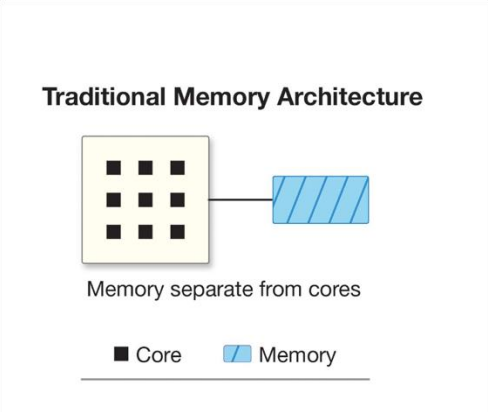
GroqRack

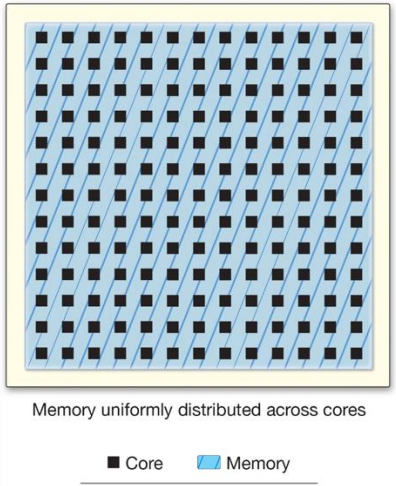
- **Cerebras:** 2 CS-2 nodes, each with 850,000 Cores, compute-intensive models
- **SambaNova:** DataScale SN30 8 nodes (8 SN30 RDUs per node) - 1TB mem per device, total 64 RDUs
- **Graphcore:** BowPod64 4 nodes (16 IPUs per node) - MIMD, irregular workloads, total 64 IPUs
- **Groq:** 9 GroqNodes, 8 GroqCards per node - inference at batch 1, total 72 GroqCards

ALCF AI Testbed Specifications

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

Von Neumann vs Spatial Architectures





- Limitations of Traditional Architectures
- Heavy data movement leads to Increased Energy Cost in GPUs

- Rise of domain-specific dataflow inspired architectures

Von Neumann vs Dataflow

```
(a+b) * (c+d)
```

Von Neumann

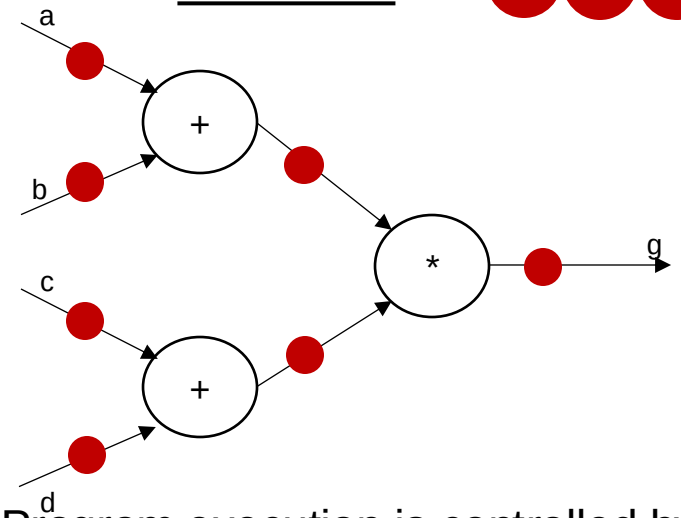
1 2 3

```
add a,b,e
sub c,d,f
mul e,f,g
```

Program execution is controlled using Program Counter

Dataflow

1 2 3

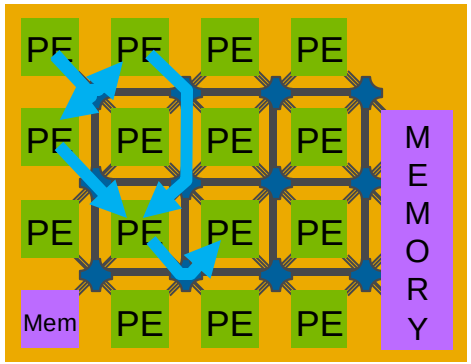


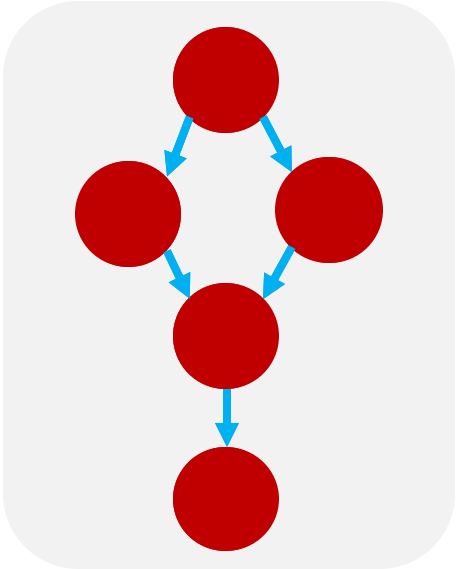
Program execution is controlled by flow of data token through graph

SPATIAL ARCHITECTURES

Workflow

- Program is represented as a graph
- This program graph is mapped on the architecture

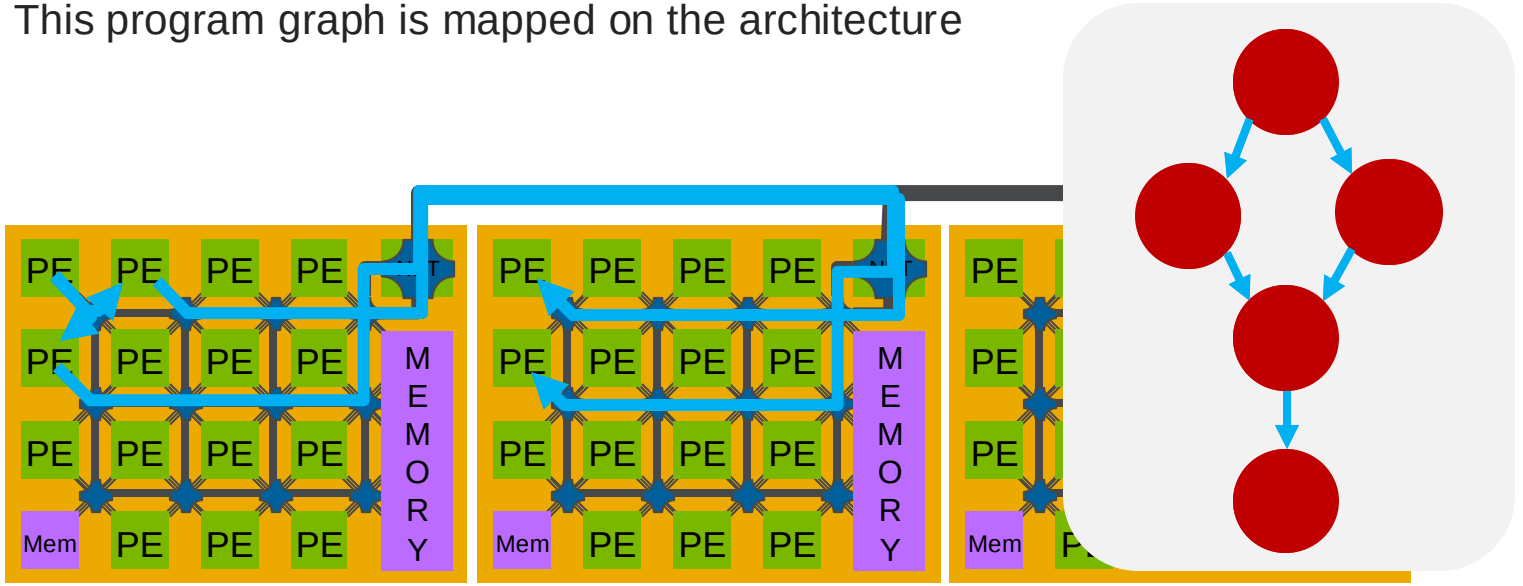




SPATIAL RECONFIGURABLE ARCHITECTURES

Workflow

- Program is represented as a graph
- This program graph is mapped on the architecture



DNN Performance on AI Accelerators

A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads

Murali Emani* memani@anl.gov
Zhen Xie* zhen.xie@anl.gov
Siddhisanket Raskar* sraskar@anl.gov
Varuni Sastry* vsastry@anl.gov
William Arnold* arnoldw@anl.gov
Bruce Wilson* wilsonb@anl.gov

Rajeev Thakur* thakur@anl.gov
Venkatram Vishwanath* venkat@anl.gov
Zhengchun Liu* zhengchun.liu@anl.gov
Michael E. Papka*‖ papka@anl.gov
Cindy Orozco Bohorquez† cindy@cerebras.net

Rick Weisner‡ rick.weisner@sambanova.ai
Karen Li‡ xiaoyan.li@sambanova.ai
Yongning Sheng‡ yongning.sheng@sambanova.ai
Yun Du‡ yun.du@sambanova.ai

Jian Zhang‡ jian.zhang@sambanova.ai
Alexander Tsyplikhin§ alext@graphcore.ai
Gurdaman Khaira§ damank@graphcore.ai
Jeremy Fowers¶ jfowers@groq.com
Ramakrishnan Sivakumar¶ rsivakumar@groq.com

Victoria Godsoe¶ vgodsoe@groq.com
Adrian Macias¶ am@groq.com
Chetan Tekur¶ ctekur@groq.com
Matthew Boyd¶ matt@groq.com

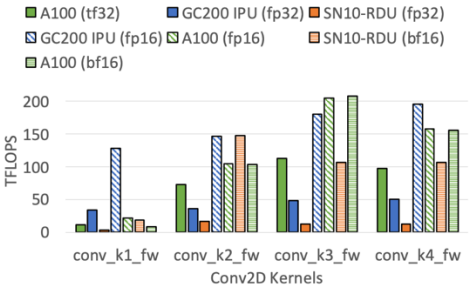
*Argonne National Laboratory, Lemont, IL 60439, USA, †Cerebras Systems, Sunnyvale, CA 95085, USA, ‡SambaNova Systems Inc., Palo Alto, CA 94303, USA, §Graphcore Inc., Palo Alto, CA 94301, USA, ¶Groq Inc., Mountain View, CA 94041, USA, ‖University of Illinois, Chicago, IL 60637, USA

***Abstract*—Scientific applications are increasingly adopting Artificial Intelligence (AI) techniques to advance science. High-performance computing centers are evaluating emerging novel hardware accelerators to efficiently run AI-driven science applications. With a wide diversity in the hardware architectures and software stacks of these systems, it is challenging to understand how these accelerators perform. The state-of-the-art in the evaluation of deep learning workloads primarily focuses on CPUs and GPUs. In this paper, we present an overview of dataflow-based novel AI accelerators from SambaNova, Cerebras, Graphcore, and Groq. We present a first-of-a-kind evaluation of these accelerators with diverse workloads, such as Deep Learning (DL) primitives, benchmark models, and scientific machine learning applications. We also evaluate the performance of collective communication, which is key for distributed DL implementation, along with a study of scaling efficiency. We then discuss key insights, challenges, and opportunities in integrating these novel AI accelerators in supercomputing systems.**

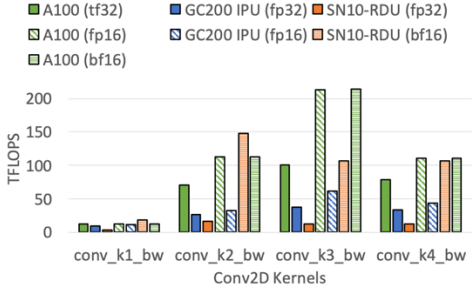
***Index Terms*—Scientific Machine Learning, Deep Learning, Accelerators, Performance Evaluation, Benchmarking**

the above. There will be a surge in scientific applications that require infrastructure to enable in-place data analysis at experimental facilities and AI capabilities integrated with large-scale models. The US Department of Energy (DOE) AI for Science Report [1], put forth by stakeholders from DOE labs, academia, and industry, cohesively highlights the need for tighter integration of the AI infrastructure ecosystem with experimental and leadership computing facilities. There is great emphasis on efficiently implementing Deep Learning (DL) models and exploiting novel architectures, especially reduced-precision AI accelerators. The DOE Advanced Scientific Computing Research (ASCR) report on extreme heterogeneity [2] lists challenges in integrating a broad spectrum of diverse hardware resources for science.

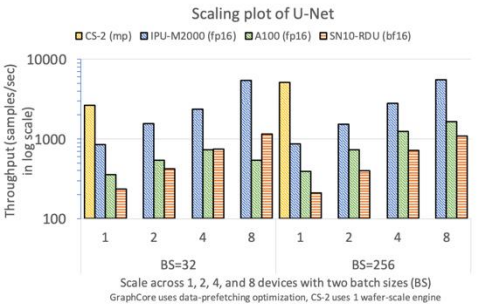
Recent advances in hardware, including heterogeneous systems and AI accelerators, will help researchers to advance the state of the art in scientific applications on powerful exascale supercomputers such as Aurora [3], El Capitan [4],

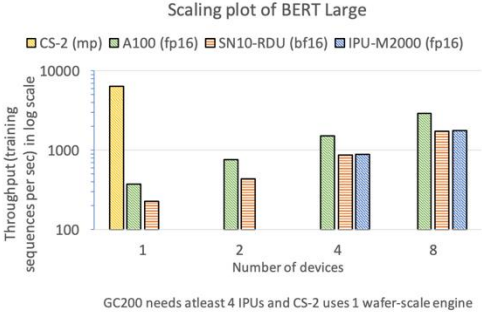


(a) Training mode, forward pass



(b) Training mode, backward pass





https://ieeexplore.ieee.org/abstract/document/10024028

Argonne NATIONAL LABORATORY

LLM Performance on AI Accelerators

Toward a Holistic Performance Evaluation of Large Language Models Across Diverse AI Accelerators

Murali Emani* memani@anl.gov
Sam Foreman* foremans@anl.gov
Varuni Sastry* vsastry@anl.gov
Zhen Xie† zxie3@binghamton.edu
Siddhisanket Raskar* sraskar@anl.gov
William Arnold* arnoldw@anl.gov

Rajeev Thakur* thakur@anl.gov
Venkatram Vishwanath* venkat@anl.gov
Michael E. Papka*‡ papka@anl.gov
Sanjif Shanmugavelu§ sshanmugavelu@groq.com

Darshan Gandhi¶ darshan.gandhi@sambanovasystems.com
Hengyu Zhao¶ hengyu.zhao@sambanovasystems.com
Dun Ma¶ eric.ma@sambanovasystems.com

Kiran Ranganath¶ kiran.ranganath@sambanovasystems.com
Rick Weisner¶ rick.weisner@sambanovasystems.com
Jiunn-yeu Chen‖ jchen@habana.ai
Yuting Yang‖ yyang@habana.ai

Natalia Vassilieva†† natalia@cerebras.net
Bin C. Zhang†† claire.zhang@cerebras.net
Sylvia Howland†† Sylvia.Howland@cerebras.net
Alexander Tsyplikhin** alext@graphcore.ai

*Argonne National Laboratory, Lemont, IL 60439, USA
†State University of New York, Binghamton, NY, 13092, USA,
‡University of Illinois, Chicago, IL 60637, USA, §Groq Inc., Mountain View, CA 94041, USA
¶SambaNova Systems Inc., Palo Alto, CA 94303, USA, ‖ Intel Habana, Santa Clara CA 95054, USA
**Graphcore Inc., Palo Alto, CA 94301, USA, †† Cerebras Systems, Sunnyvale, CA 95085, USA

***Abstract*—Artificial intelligence (AI) methods have become critical in scientific applications to help accelerate scientific discovery. Large language models (LLMs) are being considered a promising approach to address some challenging problems because of their superior generalization capabilities across domains. The effectiveness of the models and the accuracy of the applications are contingent upon their efficient execution on the underlying hardware infrastructure. Specialized AI accelerator hardware systems have recently become available for accelerating AI applications. However, the comparative performance of these AI accelerators on large language models has not been previously studied. In this paper, we systematically study LLMs on multiple AI accelerators and GPUs and evaluate their performance characteristics for these models. We evaluate these systems with (i) a micro-benchmark using a core transformer block, (ii) a GPT-2 model, and (iii) an LLM-driven science use case, GenSLM. We present our findings and analyses of the models' performance to better understand the intrinsic capabilities of AI accelerators. Furthermore, our analysis takes into account key factors such as sequence lengths, scaling behavior, and sensitivity to gradient accumulation steps.**

prediction [2], neutrino particle detection [3], drug design for precision medicine [4], genome-scale foundation model [5] and weather forecasting models [6]. Some of the most commonly used AI techniques include convolutional neural networks, recurrent neural networks, graph neural networks, and large language models (LLMs). These techniques, with their unique architectural characteristics, have become invaluable to assist scientists in their research. Within the AI landscape, the domain of Natural Language Processing (NLP) has experienced a massive surge in growth, fostering usage of LLMs in various tasks such as question-answering, text summarization, and language translation. These models are becoming increasingly critical in scientific machine-learning applications.

LLMs, such as Generative Pre-trained Transformers (GPT) GPT-3 [7], LLaMA [8], Llama 2 [9], and Bloom [10] are diverse in their neural architectures along with the quality of results for these tasks. This growth has been driven in part by

Throughput evaluation of transformer micro-benchmark

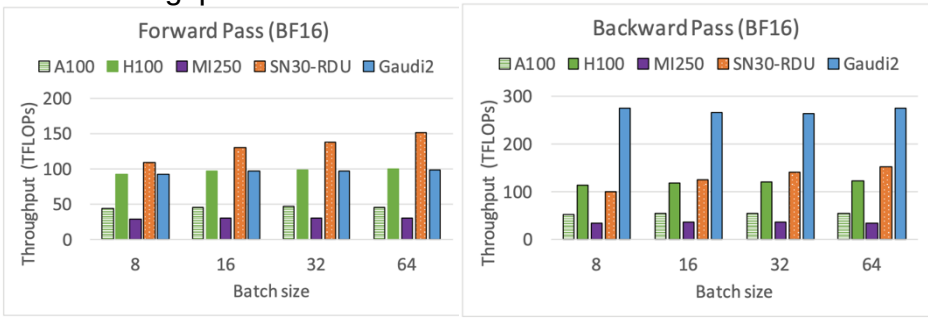


TABLE III: Scaling behavior study with the GPT-2 XL model

System	min #devices	max #devices	scale #devices	scaling efficiency	Speedup
Gaudi2	1	64	64	104%	66.4x
Bow Pod64	4	64	16	100.1%	16x
CS-2	1	2	2	99.87%	1.99x
SN30	1	64	64	97.5%	62.4x
MI250	1	4	4	80%	3.2x
A100	4	64	16	75.8%	12.1x
H100	1	4	4	43%	1.73x

https://ieeexplore.ieee.org/abstract/document/10596441

LLM Inference Bench

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

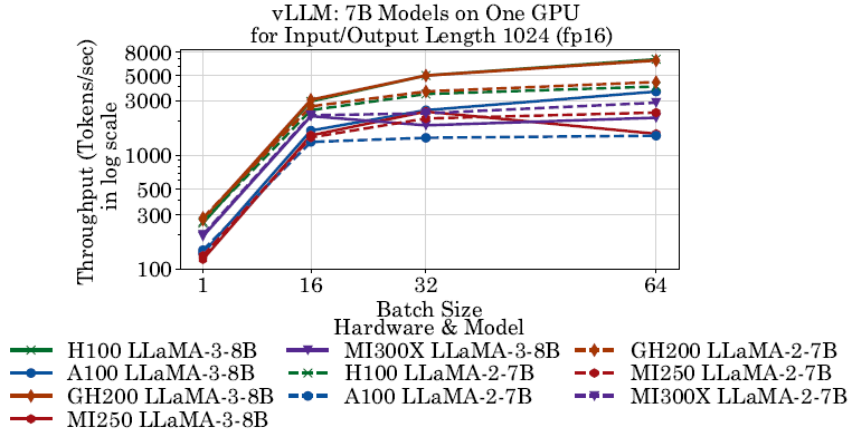
Krishna Teja Chitty-Venkata*† schittyvenkata@anl.gov
Siddhisanket Raskar*† sraskar@anl.gov
Bharat Kale* kale@anl.gov
Farah Ferdaus* fferdaus@anl.gov
Aditya Tanikanti* atanikanti@anl.gov
Ken Raffenetti* raffenet@anl.gov
Valerie Taylor* vtaylor@anl.gov
Murali Emani* memani@anl.gov
Venkatram Vishwanath* venkat@anl.gov
*Argonne National Laboratory, Lemont, IL 60439, USA

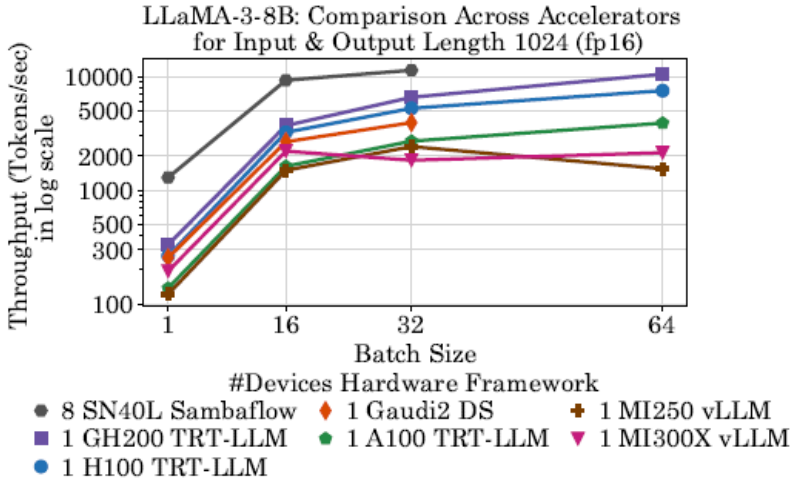
***Abstract*—Large Language Models (LLMs) have propelled groundbreaking advancements across several domains and are commonly used for text generation applications. However, the computational demands of these complex models pose significant challenges, requiring efficient hardware acceleration. Benchmarking the performance of LLMs across diverse hardware platforms is crucial to understanding their scalability and throughput characteristics. We introduce LLM-Inference-Bench, a comprehensive benchmarking suite to evaluate the hardware inference performance of LLMs. We thoroughly analyze diverse hardware platforms, including GPUs from Nvidia and AMD and specialized AI accelerators, Intel Habana and SambaNova. Our evaluation includes several LLM inference frameworks and models from LLaMA, Mistral, and Qwen families with 7B and 70B parameters. Our benchmarking results reveal the strengths and limitations of various models, hardware platforms, and inference frameworks. We provide an interactive dashboard to help identify configurations for optimal performance for a given hardware platform.**

***Index Terms*—Large Language Models, AI Accelerators, Inference Performance Evaluation, Benchmarking**

responses or make predictions. Today, efficient inference is essential for generation capabilities across various applications, such as chatbots, language translation, and information retrieval systems. As LLMs continue to grow in size and complexity, optimizing inference becomes increasingly crucial to balance performance with computational resources, energy consumption, and response times.

In recent years, the development of hardware accelerators for Deep Learning (DL) applications, such as GPUs and TPUs, has been driven to meet the computational demands of large models. These accelerators are designed to enhance performance and energy efficiency, which is particularly crucial for LLMs that consist of billions of parameters. These hardware solutions significantly improve performance, including faster training times, reduced inference latency, and enhanced scalability. This is essential for developing and deploying sophisticated models capable of handling state-of-the-art (SOTA) tasks in NLP, content generation, and decision support systems. The





https://arxiv.org/abs/2411.00136

Argonne NATIONAL LABORATORY

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.

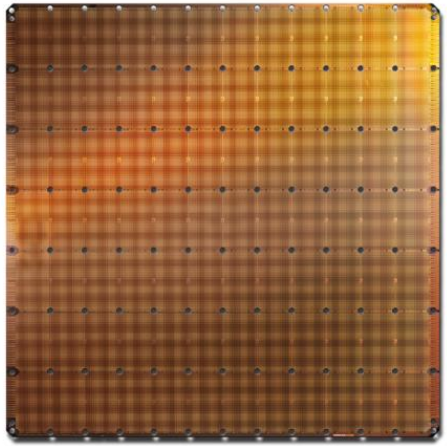
Allocation Request Form

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/

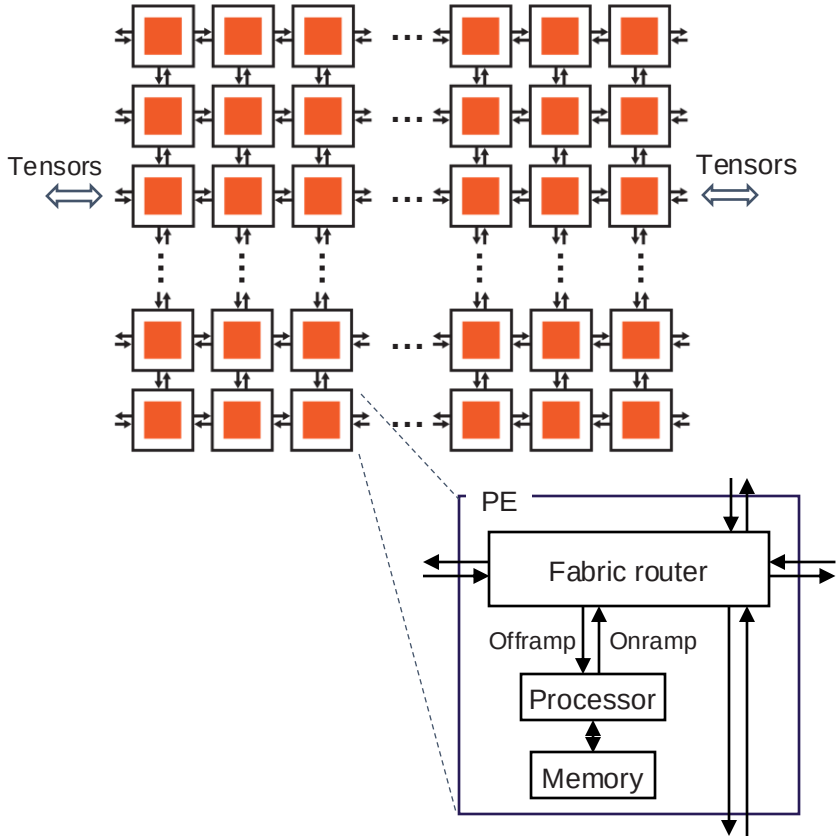


Cerebras WSE
1.2 Trillion transistors
46,225 mm² silicon

Largest GPU
21.1 Billion transistors
815 mm² silicon

- **850,000** cores optimized for sparse linear algebra
- **46,225 mm²** silicon
- **2.6 trillion** transistors, **7nm** process technology
- **40 gigabytes** of on-chip memory
- **20 PByte/s** memory bandwidth **220 Pbit/s** fabric bandwidth

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

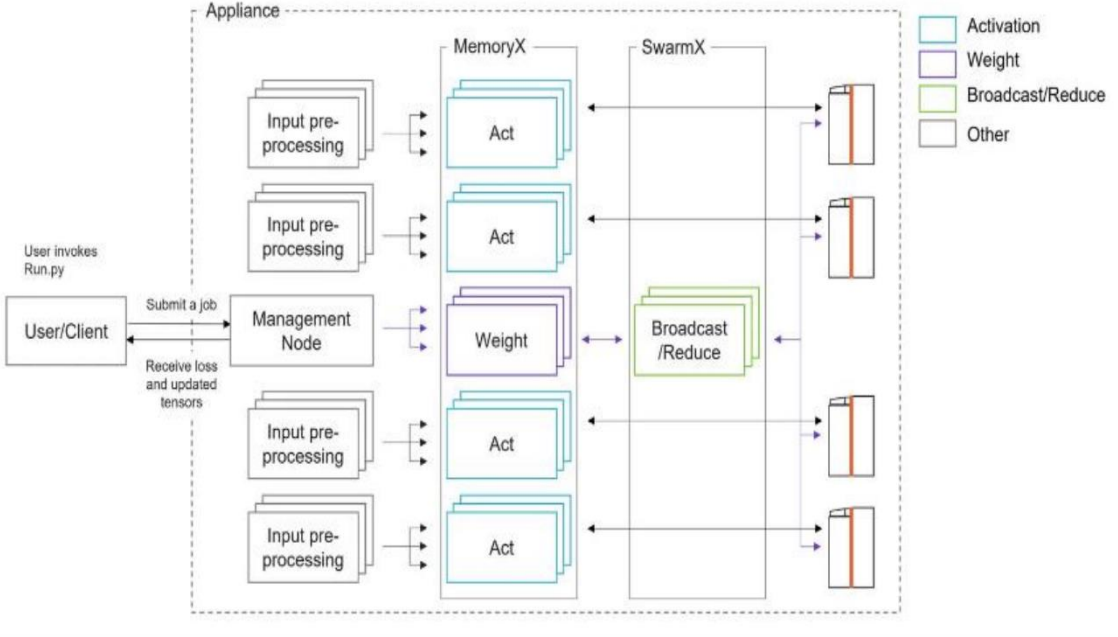
Fast memory
- 40GB on-chip SRAM
- Data and instructions
- 1 cycle read/write

Cerebras 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



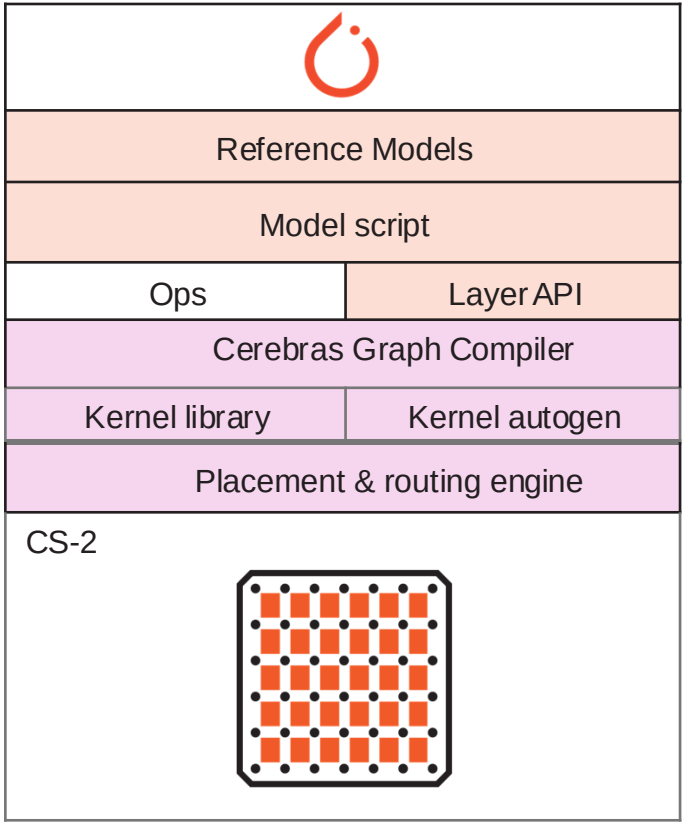
Topology of a Cerebras Wafer-Scale cluster

Training Giant Neural Networks Using Weight Streaming on Cerebras Wafer-Scale Clusters

Lowering from Model to Wafer

Integration with PyTorch

- Models defined in framework + Cerebras API
- Optimally maps from PyTorch to high performance kernels
  - Uses polyhedral code-generation or hand-written kernels
- Compiler using industry standard MLIR framework
  - Cerebras is an active contributor to the MLIR open- source community
- User does not worry about distributed compute or parallelism



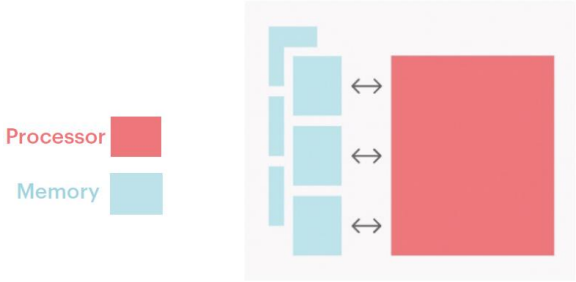
Cerebras Hands-On

https://github.com/argonne-lcf/ai-science-training-series/tree/main/07_AITestbeds

GRAPHCORE

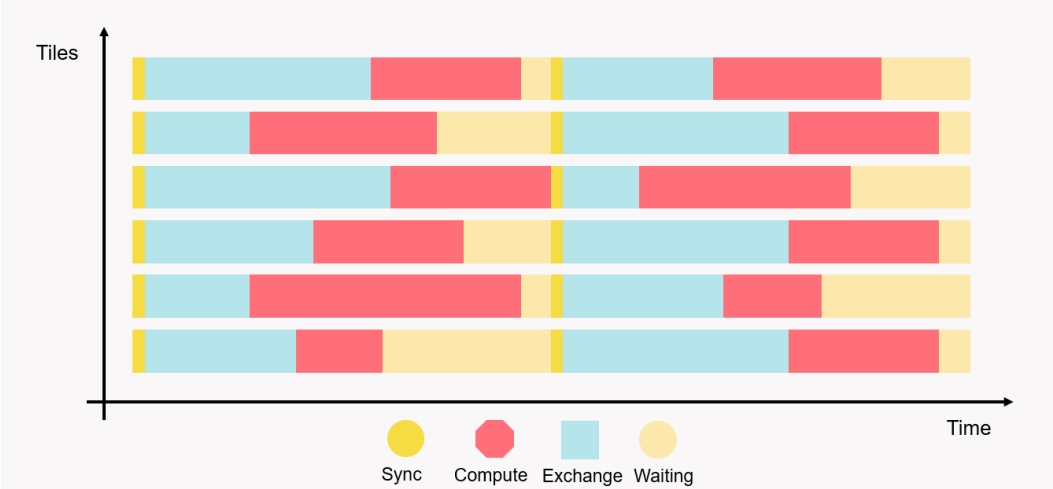
- **1472 independent IPU-Tiles**™ each with an IPU-Core™ and In-Processor-Memory™
  - 8832 independent program threads executing in parallel
- **900MB** In-Processor-Memory™ per IPU
- 47.5TB/s memory bandwidth per IPU
  - 8 TB/s all to all IPU-Exchange™
- 10 x IPU-Links,
- 64 GB/s bidirectional bandwidth to host
- 320GB/s chip to chip bandwidth

Graphcore Intelligence Processing Unit (IPU)

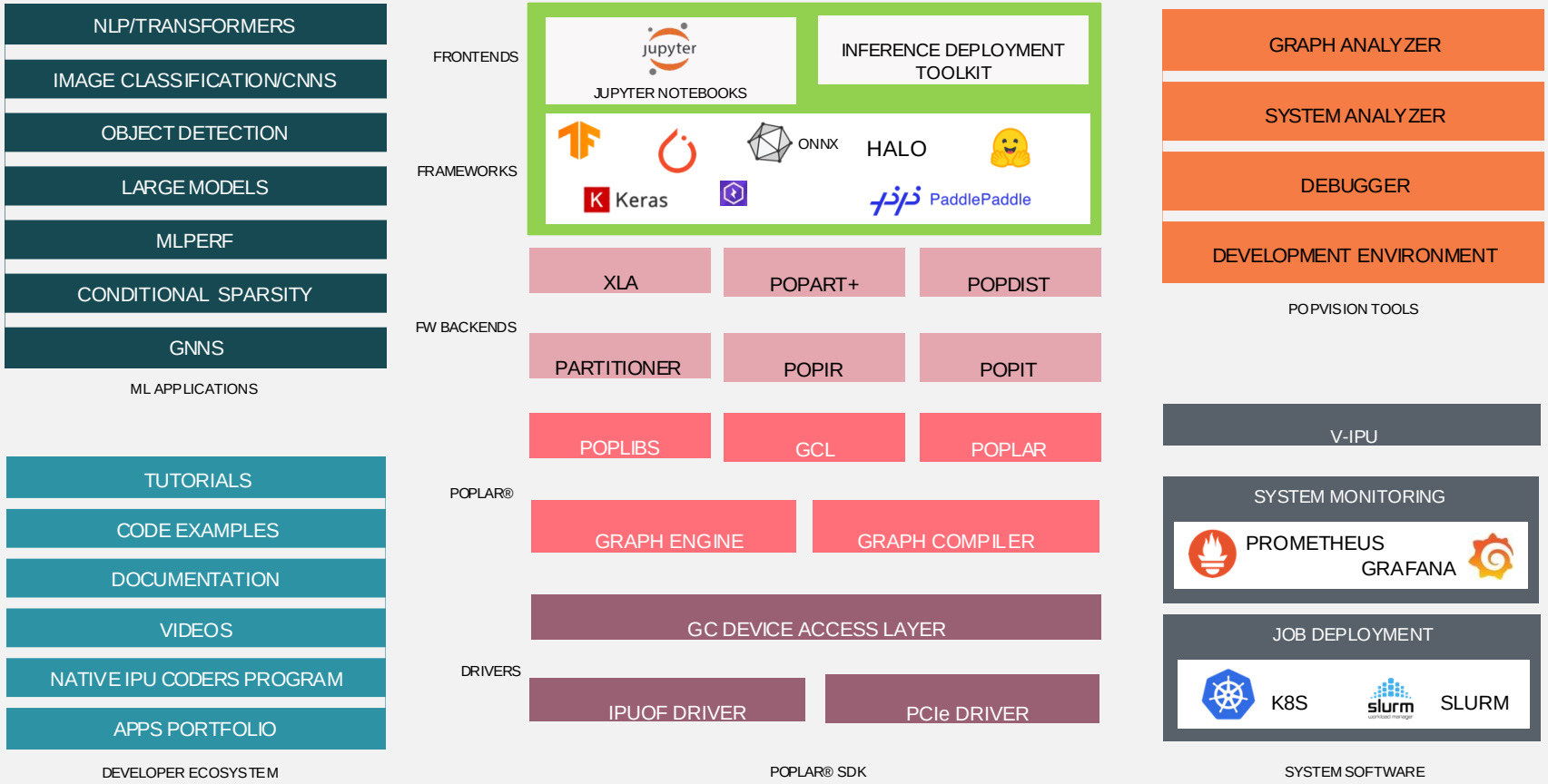
	CPU	GPU	IPU
Parallelism	Designed for scalar processing	SIMD/SIMT architecture. Designed for large blocks of dense contiguous data	Massively parallel MIMD architecture. High performance/efficiency for future ML trends
			
Memory Bandwidth	Off-chip memory	Model and Data spread across off-chip and small 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)

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



GRAPHCORE SOFTWARE



Graphcore Hands-On

https://github.com/argonne-lcf/ai-science-training-series/tree/main/07_AITestbeds

SambaNova Cardinal SN30 RDU

Cardinal SN30™
Reconfigurable Dataflow Unit™

7nm TSMC, 86B transistors

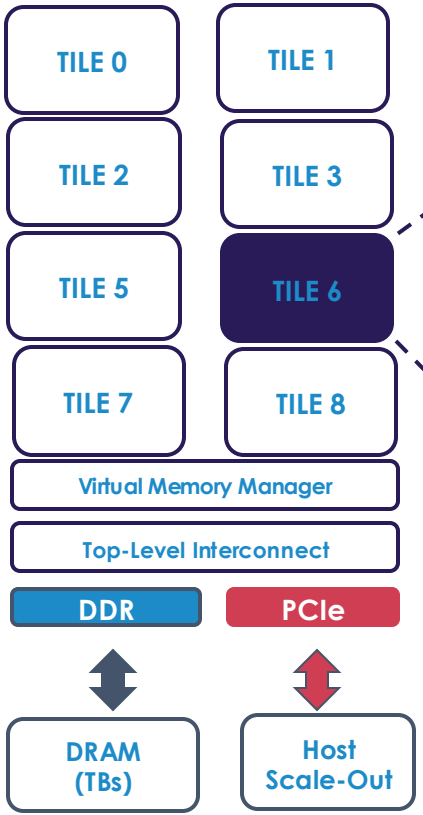
102 km of wire

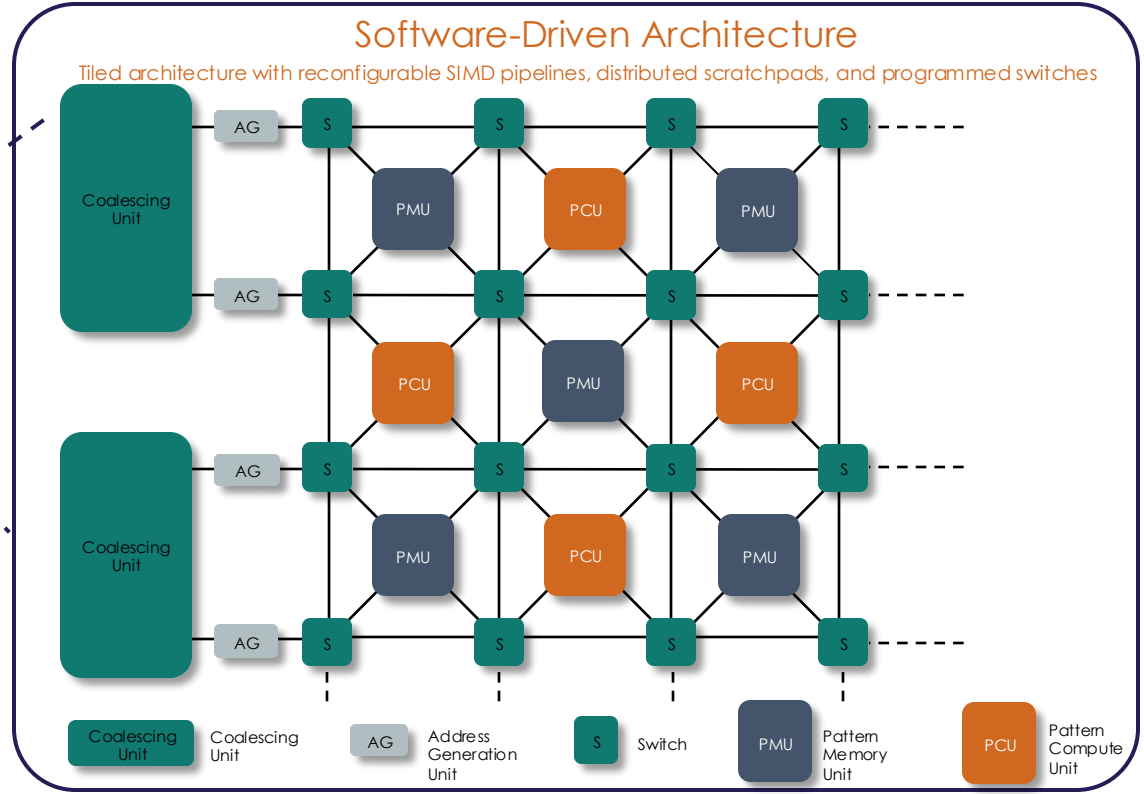
640 MB on-chip,
1,024 GB external

688 TFLOPS (bf16)

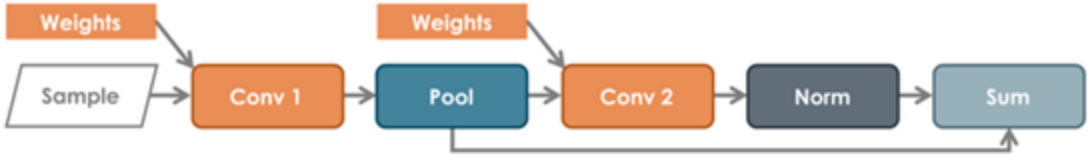
RDU-Connect™

Cardinal SN30: Tile

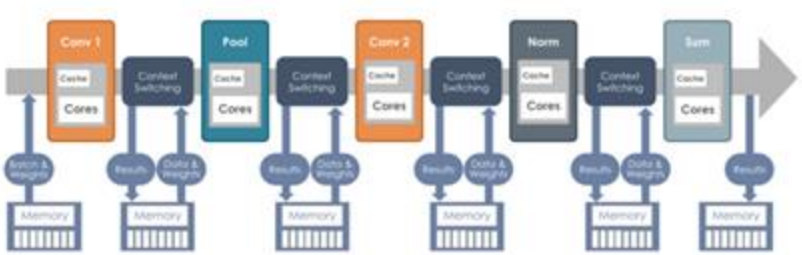




Dataflow Architectures

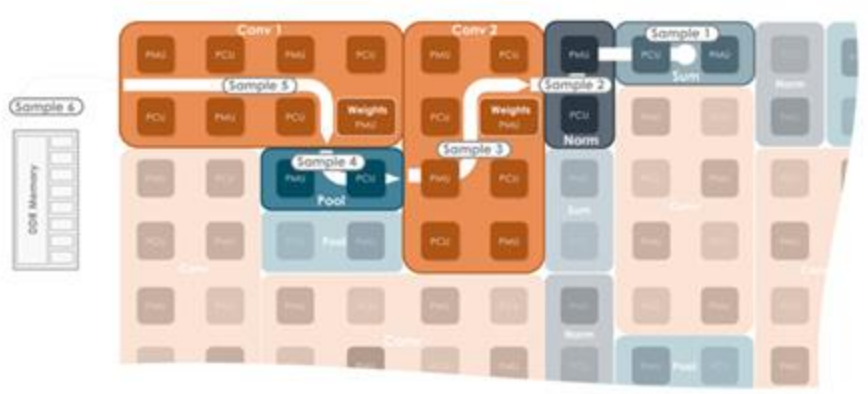


Simple Convolution Graph



The old way: kernel-by-kernel

Bottlenecked by memory bandwidth and host overhead

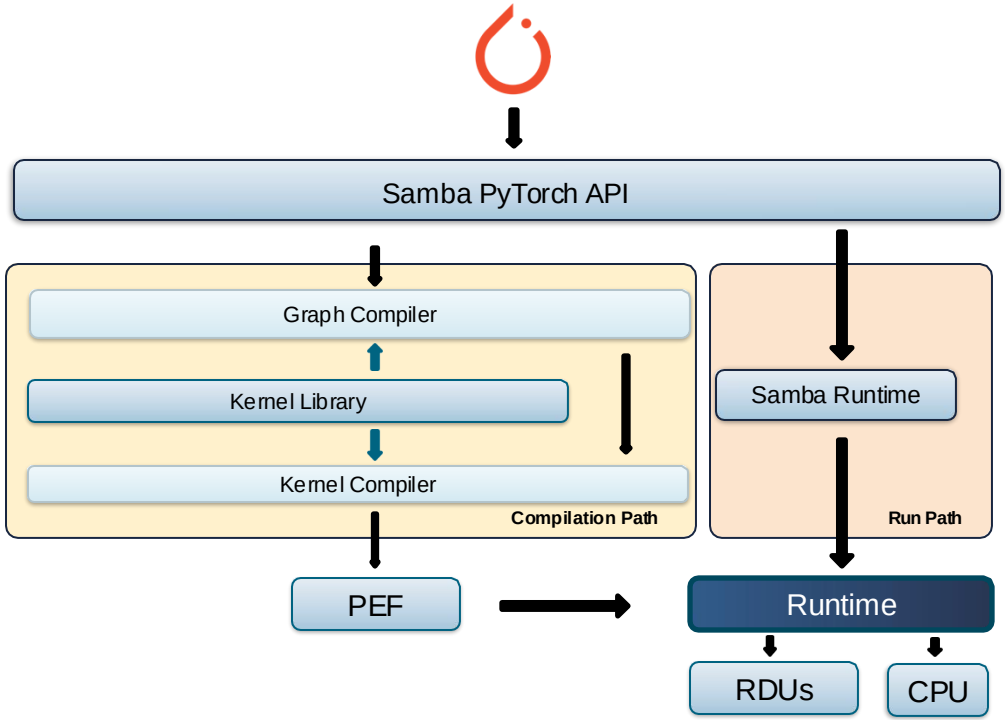


The Dataflow way: Spatial

Eliminates memory traffic and overhead

Samba Compilation Flow

- **Samba**
  - SambaNova PyTorch compilation & run APIs
- **Graph compiler**
  - High-level ML graph transformation & optimizations
- **Kernel compiler**
  - Low-level RDU operator kernel transformation & optimizations
- **Kernel library**
  - RDU operator implementations



Sambanova Hands-On

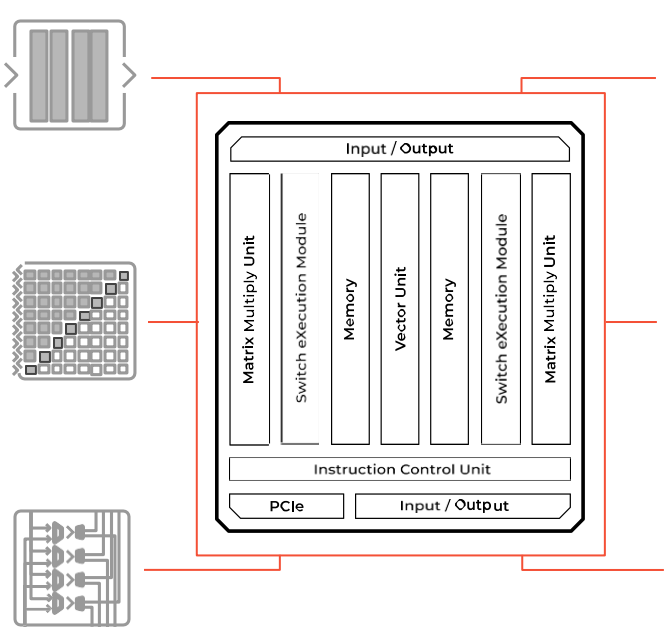
https://github.com/argonne-lcf/ai-science-training-series/tree/main/07_AITestbeds

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-Vector /
Matrix-Matrix Multiply

VXM
Vector-Vector
Operations

SXM
Data Reshapes

MEM
On-chip SRAM

Architecture Empowering Software

Software-controlled memory

No dynamic hardware caching

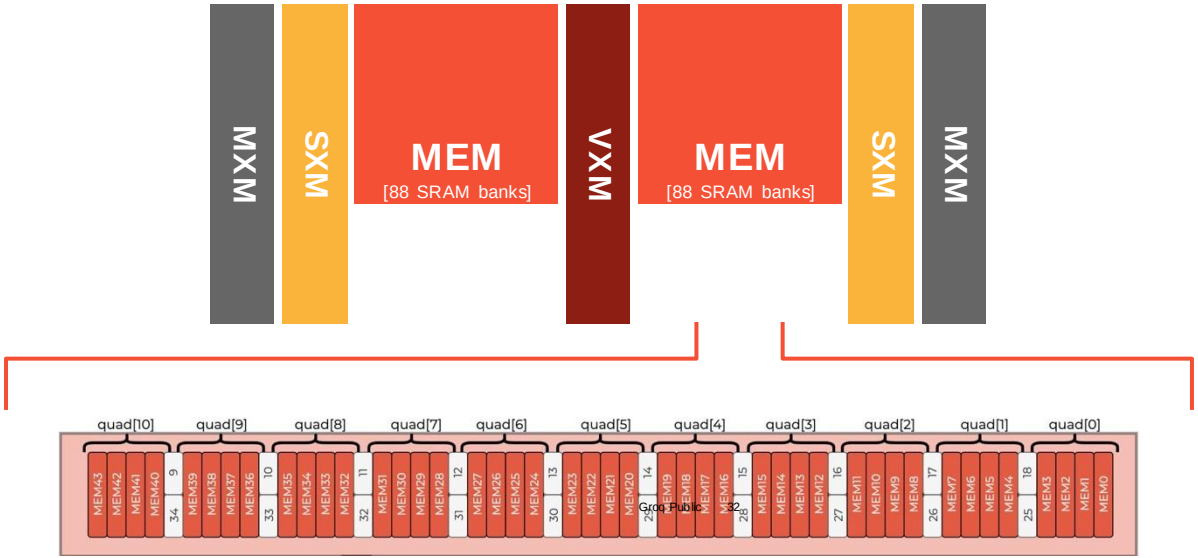
- Compiler aware of all data locations at any given point in time

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

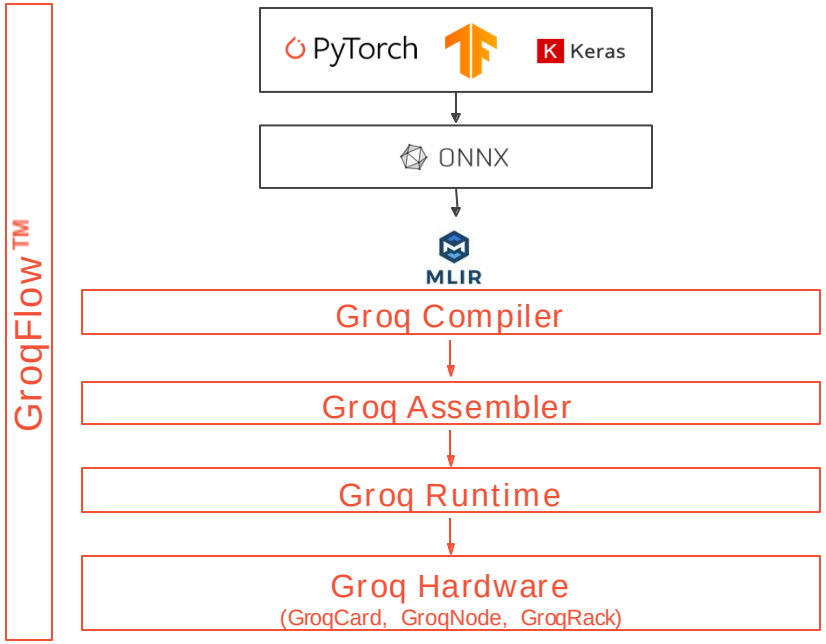
- Memory exposed to software as a set of physical banks that are directly addressed

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

- Achieves high compute efficiency even at low operational intensity



GroqWare™ Suite



DIVERSE SUITE OF DEVELOPMENT TOOLS

Out-of-Box

Groq Compiler provides out-of-box support for standard Deep Learning models

+

Productivity Tools

GroqView Profiler provides visualization of the chip's compute and memory usage at compile time

GroqFlow Tool Chain enables a single line of Pytorch or TensorFlow code to import and transform models through a fully automated tool chain to run on Groq hardware

Groq Hands-On

https://github.com/argonne-lcf/ai-science-training-series/tree/main/07_AITestbeds

Thank You

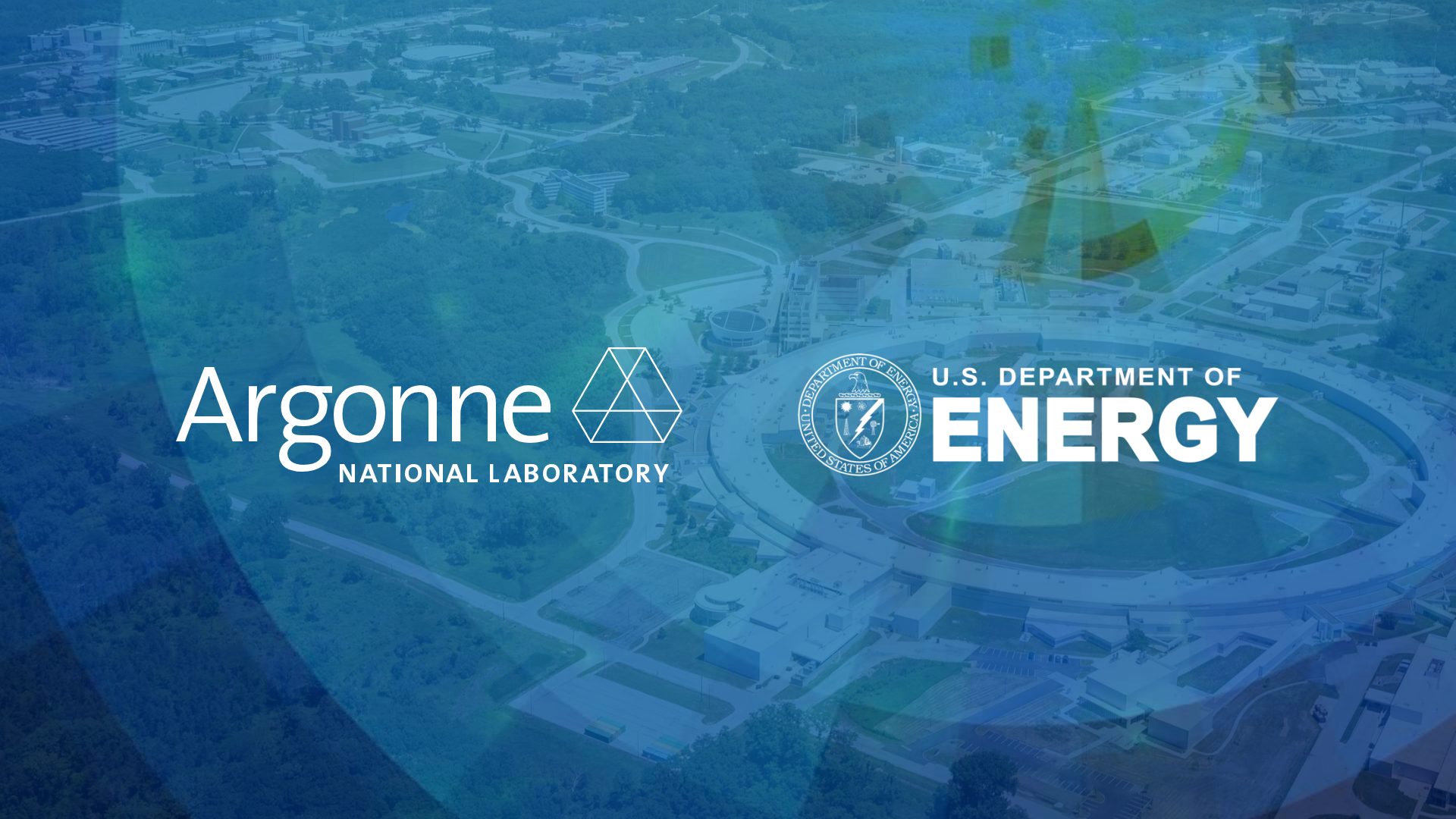
- This research was funded in part and used resources of the Argonne Leadership Computing Facility (ALCF), a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

- Murali Emani, Varuni Sastry, William Arnold, Venkatram Vishwanath

- Our current AI testbed system vendors – Cerebras, Graphcore, Groq, Intel Habana and SambaNova.

- Many slides are courtesy of AI Testbed vendors.

Please reach out for further details
Sid Raskar, sraskar@anl.gov


Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT OF
ENERGY