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

Murali Emani
Argonne Leadership Computing Facility
memani@anl.gov

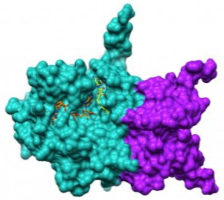
Integrating AI Systems in Facilities

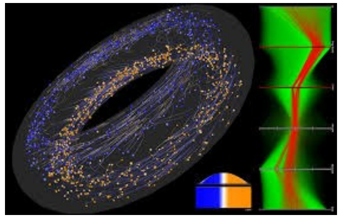
Computing Facility

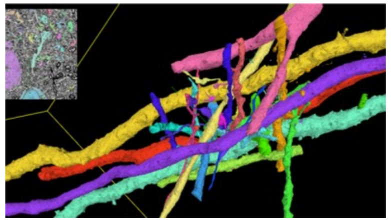
Experimental Facility

AI-Accelerators

AI-Edge accelerator







Simulations

Data-driven Models

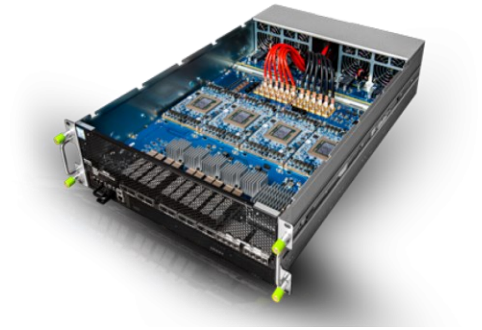
ALCF AI Testbed

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

Cerebras CS-2

SambaNova DataScale SN30

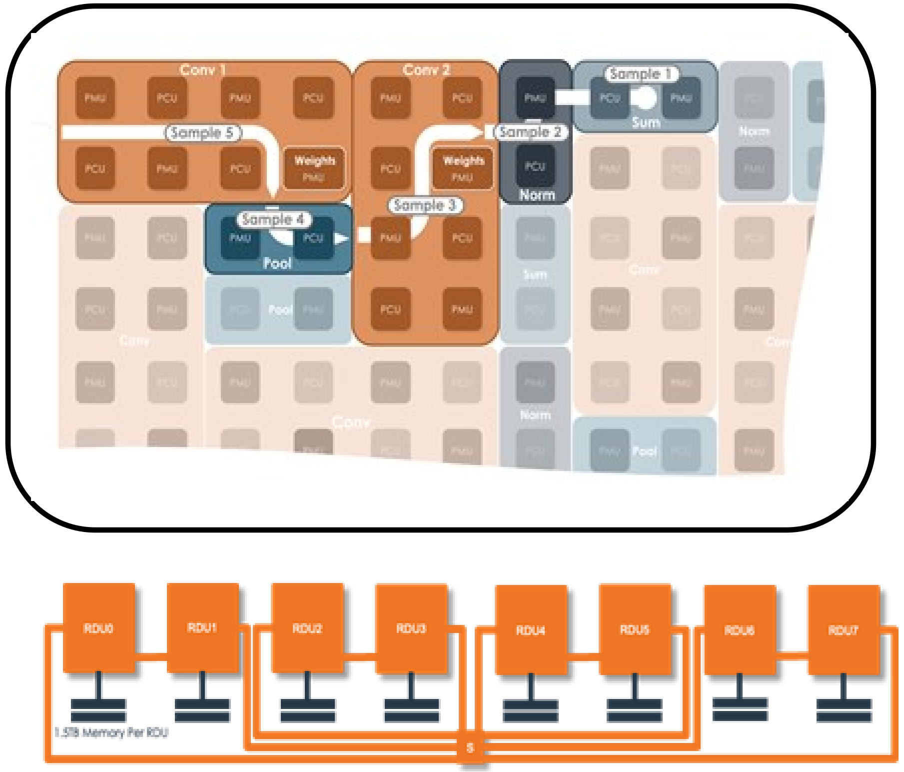
Graphcore Bow Pod64



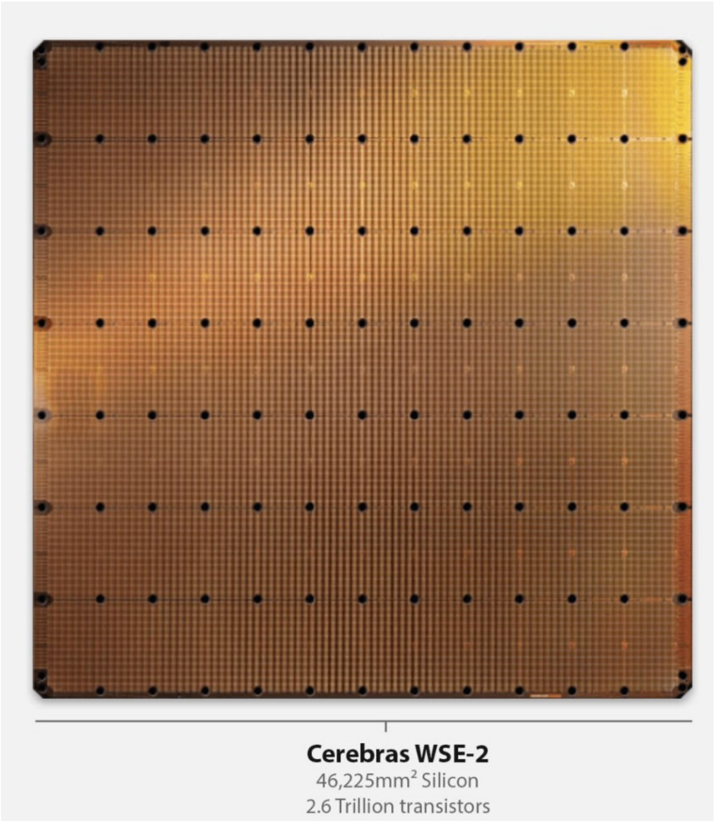
Habana Gaudi1

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



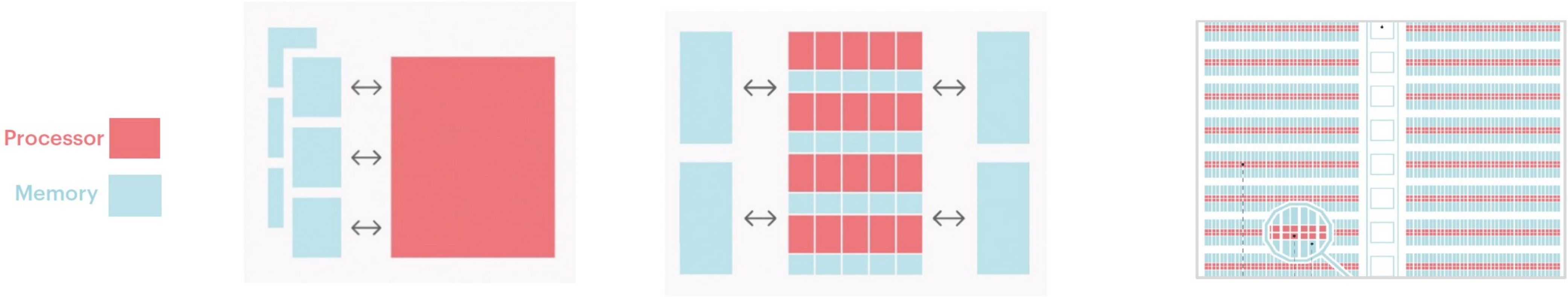
SambaNova Reconfigurable DataFlow Unit (RDU)



Cerebras Wafer Scale Engine

Image source: SambaNova, Cerebras

	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



	CPU	GPU	IPU
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)

Image source: Graphcore

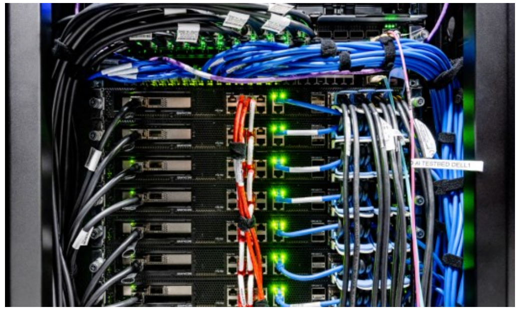
Recent ALCF AI Testbed Updates

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

https://www.alcf.anl.gov/science/directors-discretionary-allocation-program

SambaNova upgraded to latest 2nd generation SN30 accelerators and scaled to **8 nodes with 64 AI accelerators (RDU)**

SambaNova SN30



Graphcore upgraded to latest Bow generation accelerators and scaled to a **Pod-64 configuration with 64 accelerators (IPU)**

Graphcore BowPod64

Cerebras CS-2 upgraded to an appliance mode to include Memory-X and Swarm-X technologies to enable larger models and scaled to **two CS-2 engines**

Cerebras CS-2

Groq system has been upgraded to a GroqRack with nine nodes, each consisting of eight GroqChip Tensor streaming processors, **72 accelerators**

GroqRack

https://nairrpilot.org

TABLE I: Features of evaluated AI accelerators

Feature	**Nvidia A100**	**SambaNova DataScale SN30**	**Cerebras CS-2**	**Graphcore Bow-Pod64**	**Intel Gaudi2**	**AMD MI250**	**Groq TSP**
System Size[1]	$64 (= 16 \times 4)$	$64 (= 8 \times 8)$	$2 (= 2 \times 1)$	$64 (= 4 \times 16)$	$8 (= 1 \times 8)$	$4 (= 1 \times 4)$	$72 (= 9 \times 8)$
Memory (/node)	160 GB	8 TB	1 TB	3.6 GB/128 GB [2]	768 GB	512 GB	16.56 GB
Memory (/device)	40 GB	1 TB	1 TB	900 MB/32 GB	96 GB	128 GB	230 MB
Interconnect	NVLink	Ethernet-based	Ethernet-based	IPU Link	RoCE	AMD CDNA	Groq C2C
Software Stack[3]	TF, PT, ONNX, MxNET, CUDA	SambaFlow™, PT, TF	PT, Cerebras SDK	TF, PT, ONNX, PopArt	Synapse AI, TF and PT	TF, PT, ROCm	PT, TF, ONNX GroqFlow™, GroqWare Suite™
Precision (commonly used)	TF32, FP32, FP16, BF16	FP32, BF16, Int32, Int16, Int8	FP32, FP16, BF16, cbfloat	FP32, FP16	FP32, TF32, BF16, FP16, FP8	FP64, FP32, FP16, BF16, INT8, INT4	FP32, FP16, FP8, INT8, INT4
Compute Units (/device)	6912 Cuda Cores,432 Tensor Cores	1280 PCUs	850,000 Cores	1472 Compute cores	24 TPC + 2 MME	13312 cores, 208 compute units	Single core, with specialized functional slices

TABLE I: Features of evaluated AI accelerators

Feature	Nvidia A100	SambaNova DataScale SN30	Cerebras CS-2	Graphcore Bow-Pod64	Intel Gaudi2	AMD MI250	Groq TSP
System Size[1]	$64\,(=16\times4)$	$64\,(=8\times8)$	$2\,(=2\times1)$	$64\,(=4\times16)$	$8\,(=1\times8)$	$4\,(=1\times4)$	$72\,(=9\times8)$
Memory (/node)	160 GB	8 TB	1 TB	3.6 GB/128 GB [2]	768 GB	512 GB	16.56 GB
Memory (/device)	40 GB	1 TB	1 TB	900 MB/32 GB	96 GB	128 GB	230 MB
Interconnect	NVLink	Ethernet-based	Ethernet-based	IPU Link	RoCE	AMD CDNA	Groq C2C
Software Stack[3]	TF, PT, ONNX, MxNET, CUDA	SambaFlow™, PT, TF	PT, Cerebras SDK	TF, PT, ONNX, PopArt	Synapse AI, TF and PT	TF, PT, ROCm	PT, TF, ONNX GroqFlow™, GroqWare Suite™
Precision (commonly used)	TF32, FP32, FP16, BF16	FP32, BF16, Int32, Int16, Int8	FP32, FP16, BF16, cbfloat	FP32, FP16	FP32, TF32, BF16, FP16, FP8	FP64, FP32, FP16, BF16, INT8, INT4	FP32, FP16, FP8, INT8, INT4
Compute Units (/device)	6912 Cuda Cores,432 Tensor Cores	1280 PCUs	850,000 Cores	1472 Compute cores	24 TPC + 2 MME	13312 cores, 208 compute units	Single core, with specialized functional slices

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 SN30, Graphcore Bow Pod64, and GroqRack at ALCF are available for user allocations

Allocation Request Form
https://www.alcf.anl.gov/science/directors-discretionary-allocation-program

AI Testbed User Guide
https://www.alcf.anl.gov/alcf-ai-testbed

Challenges

- Understand how these systems perform for different workloads given diverse hardware and software characteristics
- What are the unique capabilities of each evaluated system
- Given the advancement of GenAI and foundation models in AI for science applications, focus on LLM performance evaluation

Approach

- Perform a comprehensive evaluation with
  - traditional Deep Learning (DL) models[1]:
    - *DL primitives*: GEMM, Conv2D, ReLU, and RNN
    - *Benchmarks*: U-Net, BERT-Large, ResNet-50
    - *AI4S applications*: BraggNN, Uno
    - Scalability and Collective communications

(1) **Emani et al. "A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads",** Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS) at SC 2022.

.

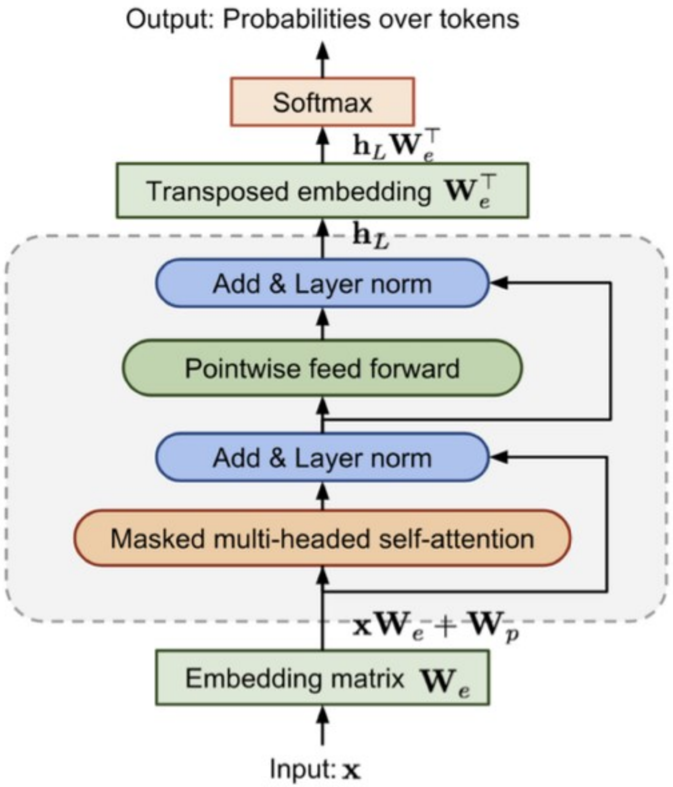
Approach

- Perform a comprehensive evaluation with Large Language Models
  - Transformer block micro-benchmark,
  - GPT-2 XL, scaling study, impact of sequence lengths, gradient accumulation steps
  - Science usecase: GenSLM, foundation model

  - A100, H100, Cerebras CS-2, SambaNova SN30, Intel Habana Gaudi2, Graphcore Bow Pod64, Groq, AMD MI250

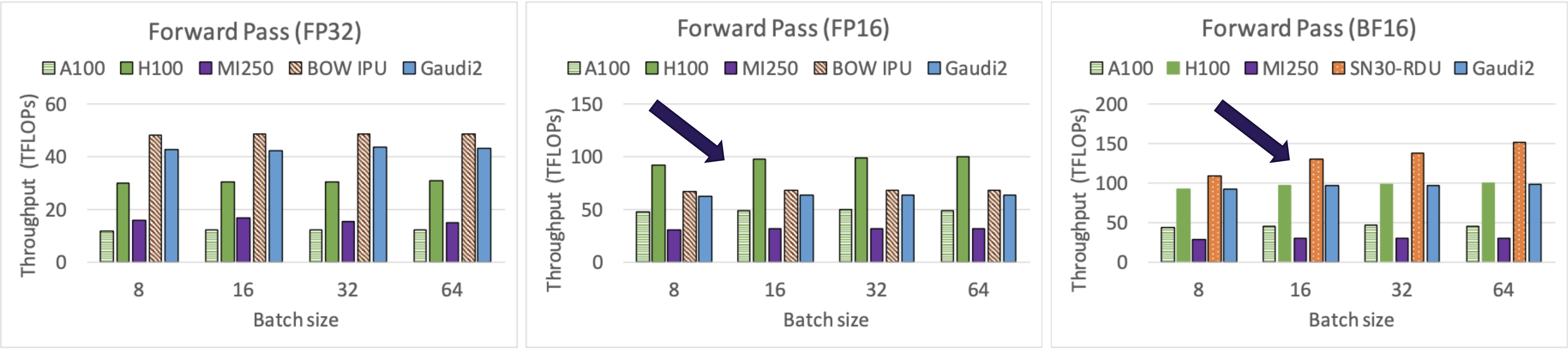
(1) **Emani et al. "A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads",**
Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS) at SC 2022.
(2) **Emani et al. " Toward a Holistic Performance Evaluation of Large Language Models Across Diverse AI Accelerators",**
Heterogeneity in Computing Workshop (HCW) at IPDPS24.

Transformer Block micro-benchmark



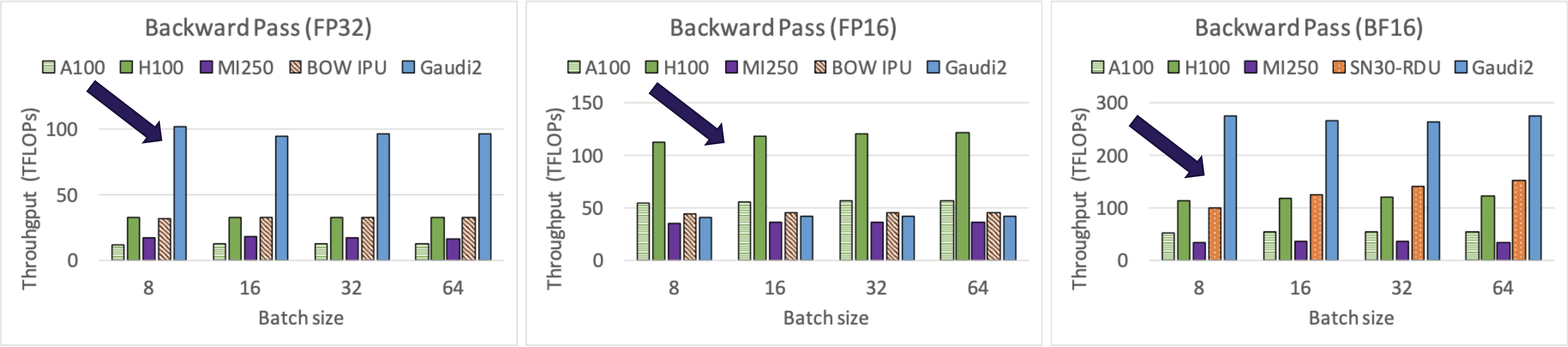
Transformer Block
Repeat x L=12

$$\mathbf{h}_\ell = \text{transformer\_block}(\mathbf{h}_{\ell-1})$$
$$\ell = 1, \ldots, L$$



(a) FP32 (b) FP16 (c) BF16

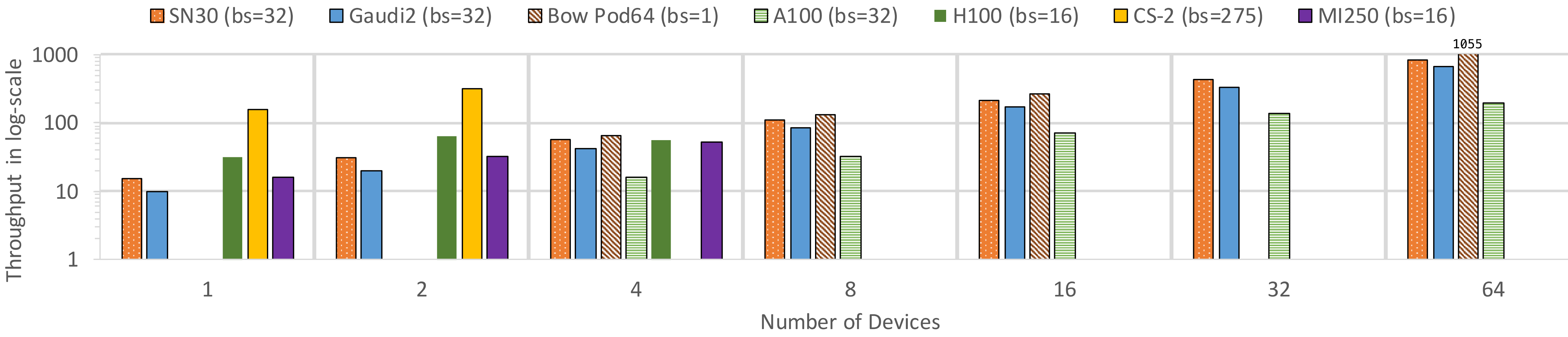
Fig. 1: Throughput evaluation of transformer micro-benchmark in the forward pass with various precision



(a) FP32 (b) FP16 (c) BF16

Fig. 2: Throughput evaluation of transformer micro-benchmark in the backward pass with various precision

GPT Model Performance



Used GPT-2 XL 1.5B parameter model, OWT dataset

- same sequence length, tuned batch sizes, custom software stack
- Runs on A100s used Megatron deepspeed, work with Megatron-core is under progress out-of-box runs with no additional optimizations
- 16 SN30 RDUs, 2 CS-2s, and 16 IPUs match the performance on 64 A100s

Scaling behavior study with the GPT-2 XL model

System	min #devices	max #devices	scale #devices	scaling efficiency
Gaudi2	1	64	64	104%
Bow Pod64	4	64	16	100.1%
CS-2	1	2	2	99.87%
SN30	1	64	64	97.5%
MI250	1	4	4	80%
A100	4	64	16	75.8%

Normalized Throughput per unit local batch size

System	Throughput (tokens/sec)
A100	0.49
H100	3.5
CS-2	2.12
SN30-RDU	1.82
IPU	1.03
Gaudi2	1.33
MI250	1.63

Impact of Sequence length

TABLE III: Impact of Sequence length on model throughput

System (model Size)	Seq Length	Devices	Throughput (tokens/s)
A100 (1.5B)	1024	4	134,144
	2048	4	124,928
CS-2 (1.5B)	1024	1	133,069
	2048	1	114,811
	4096	1	63,488
	8192	1	16,302
CS-2 (13B)	1024	1	20,685
	2048	1	20,173
	4096	1	17,531
	8192	1	15,237
	16384	1	11,796
	32768	1	7537
	51200	1	5120
SN30 (13B)	1024	8	22,135
	2048	8	21,684
	4096	8	17,000
	8192	8	10,581
	16384	8	4936
	32768	8	5021
	65536	8	1880

Genome-scale Language Models (GenSLMs)

Goal:

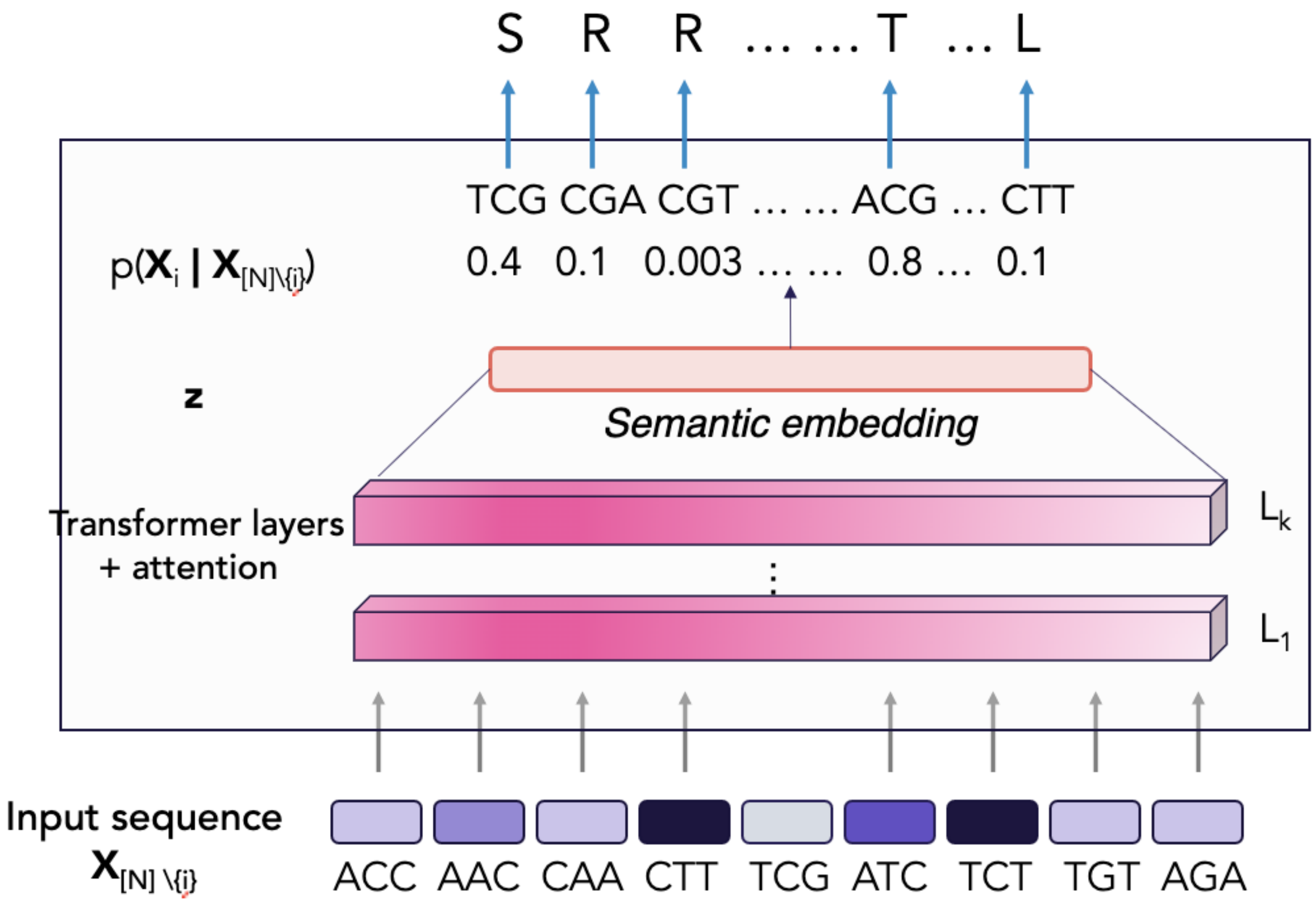
- How new and emergent variants of pandemic causing viruses, (specifically SARS-CoV-2) can be identified and classified.
- Identify mutations that are VOC (increased severity and transmissibility)
- Extendable to gene or protein synthesis.

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:

- High complexity in the attention computation
- Communication overheads

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

Observations, Challenges and Insights

- It is extremely challenging for a fair comparison, devise better methodologies
- Better support to run opensource models (HuggingFace), other architectures (Mixture of Experts)
- To accommodate larger models, focus on optimizations such as sparsity
- Need to focus on memory optimizations, not just compute
- Significant porting efforts and compilation times, getting better over time. Focus on evaluating performance/watt

Ongoing Efforts

- Evaluate new AI accelerators offerings and incorporate promising solutions as part of the testbed

- Work in progress on Inference and Fine-tuning benchmarking on models such as GPT, Llama, Mistral, Mixtral

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

Recent Publications

- **A Comprehensive Performance Study of Large Language Models on Novel AI Accelerators**
  Murali Emani, Sam Foreman, Varuni Sastry, Zhen Xie, Siddhisanket Raskar, William Arnold, Rajeev Thakur, Venkatram Vishwanath, Michael E. Papka
  https://arxiv.org/abs/2310.04607

- **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,***

- **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

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, Murali Emani, Michael Papka, William Arnold, Varuni Sastry, Sid Raskar, Zhen Xie, Rajeev Thakur, Bruce Wilson, 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