CHALLENGES AND OPPORTUNITIES WITH EMERGING AI ACCELERATORS FOR COMPUTATIONAL SCIENCE

Siddhisanket Raskar, Murali Emani, Venkat Vishwanath
sraskar@anl.gov
Argonne National Laboratory

ABSTRACT

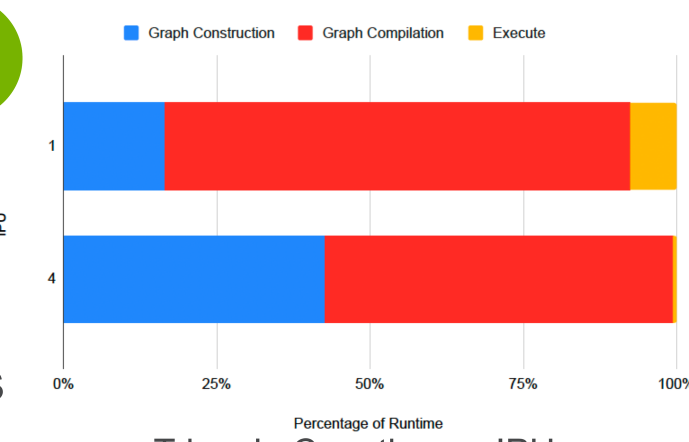
- Provide an overview of the key capabilities and performance implications of emerging AI accelerators
  - Programming approaches and software stacks to take advantage of their capabilities.
  - Performance at scale, particularly for large computational science applications.
- AI accelerators have sparked intense interest in handling traditional HPC workloads and driving algorithmic research due to significant raw compute capability and high bandwidth, often outperforming GPUs.
- Understand foundational questions
  - ideal programming models to port HPC applications to these emerging accelerators
  - the need to build higher-level abstractions to enable portability across them, and compiler support from vendors for active community engagement.

CHALLENGES

Feature	Cerebras CS-2	Sambanova RDU	Graphcore IPU	Groq LPU	Habana HPU	Nextsilicon Maverick	Nvidia GPU
Language	CSL	C/C++	C/C++	C/C++	TPC-C	C/C++, Fortran	C/C++, Fortran
Runtime	N/A	AI4S	Poplar, Poplibs	Groq Runtime	Habana TPC	OpenMP	Cuda & many more
Compiler	LLVM	LLVM, MLIR	LLVM	N/A	LLVM	LLVM	LLVM
Target	HPC Applications	HPC Applications	HPC Applications	Custom ML Kernels	Custom ML Kernels	HPC Applications	HPC Applications

Challenges

- No portability across architectures
- Low level programming models
- Different Optimizations strategies
- Significant Compilation & projection times
- Support for higher precision



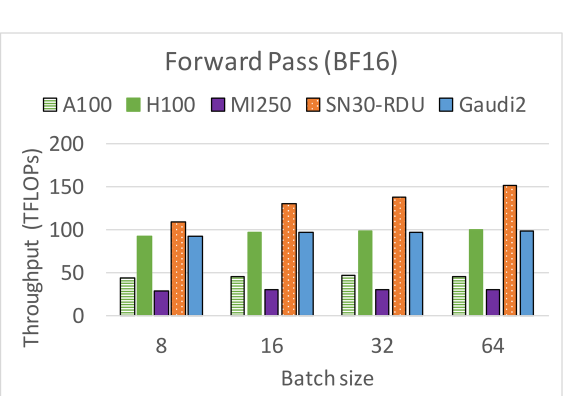
Triangle Counting on IPU Runtime Breakdown [4]

MOTIVATION

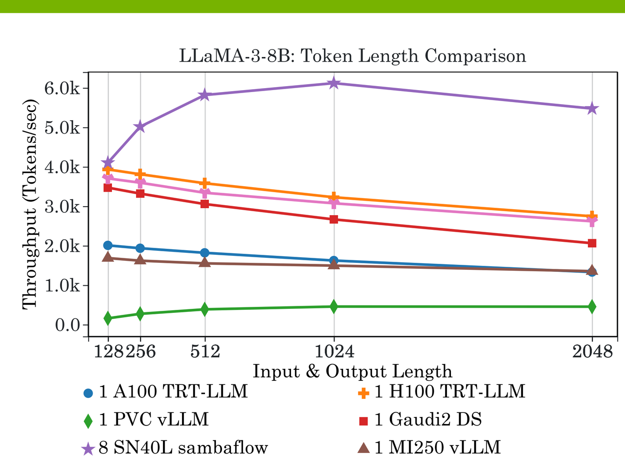
Emergence of Accelerators

cerebras SambaNova GRAPHCORE groq habana NEXTSILICON
... and many more

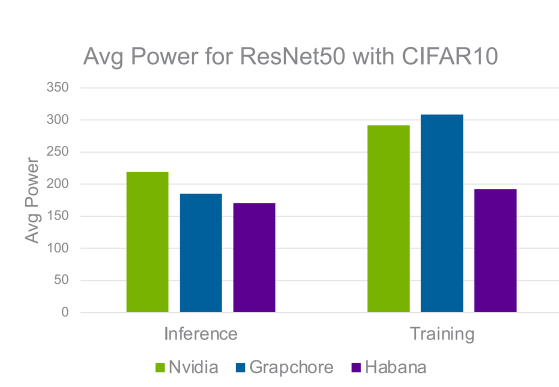
Performance for AI Workloads



Transformer Block [2]



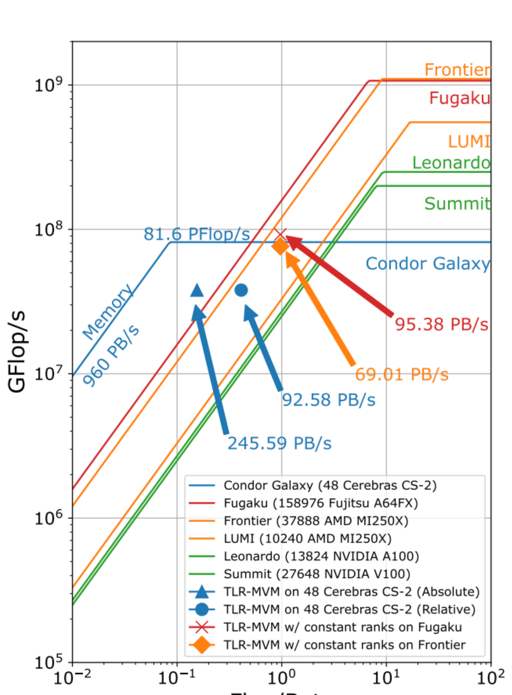
LLM Inference [3]



Power Efficiency

AI Accelerators deliver better performance than GPUs

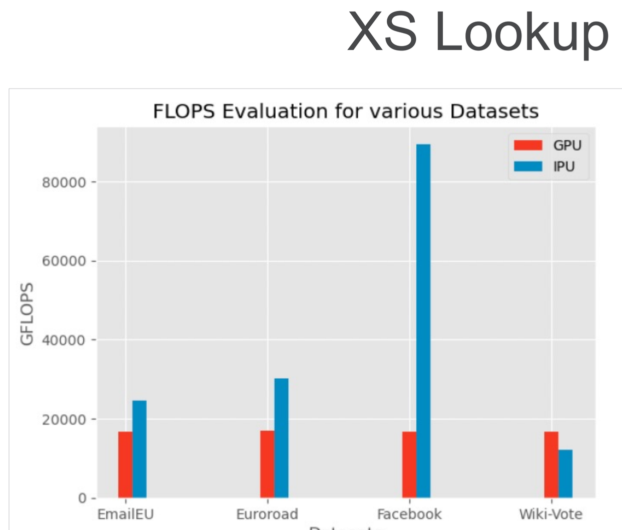
Performance for HPC Workloads



Condor Galaxy vs top 5 Supercomputers [6]

	Transistor Count [Trillion]	Die Area [mm^2]	Peak Power [kW]	Theoretical FP32 Peak [TFLOPS]	Monte Carlo XS Lookup FOM [Lookups/s]
A100 GPU	0.0542	826	0.4	19.5	6.43E+07
Cerebras WSE-2	2.6	46,225	22.8	1,267	8.36E+09
WSE-2/A100	48	56	57	65	130

XS Lookup Kernel [5]



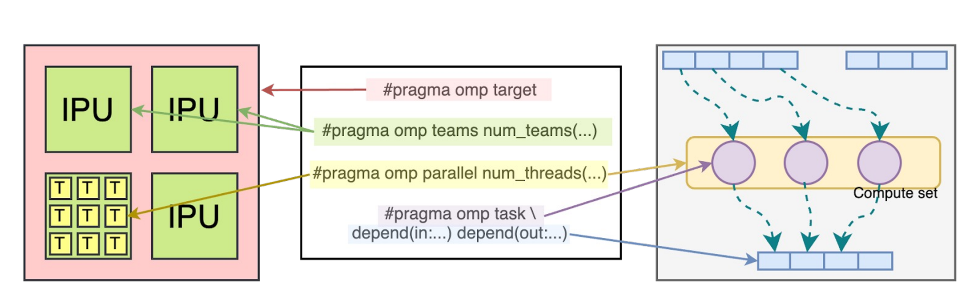
Performance of Triangle Counting on Graphcore [4]

OPPORTUNITIES

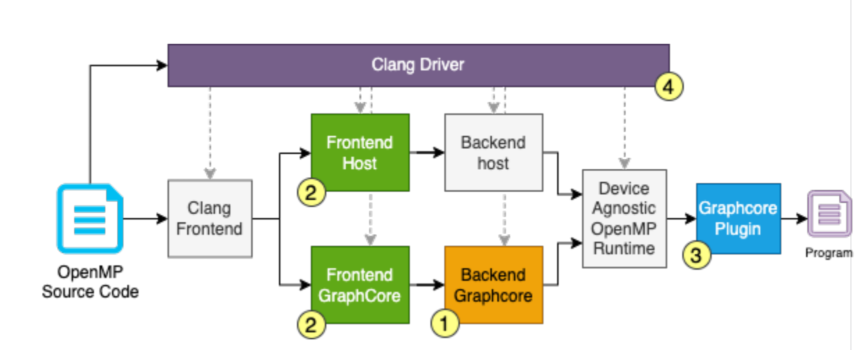
Open Questions

- What is the ideal programming model for these accelerators?
- What is the best way to enable portability across various architectures?
- How optimization techniques differ across architectures and workloads?

OpenMP target Offloading for Graphcore



OpenMP Mapping

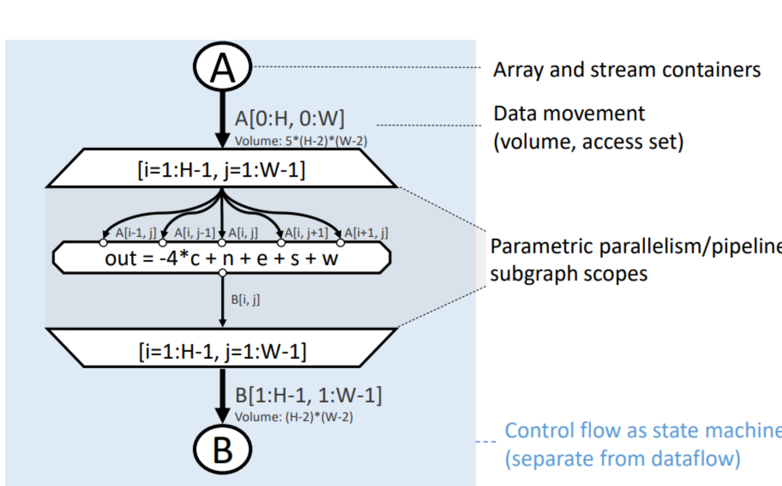


OpenMP compilation Pipeline

Can we support OpenMP on Graphcore's IPU Architecture? [7]

Graphcore backend for DaCE

Data-Centric (DaCe) programming defining a flow-based graph representation for programs, called the Stateful Dataflow Multigraph (SDFG)



DaCE	GraphCore
Control Flow Graph	Poplar::Program
Map-consume (parallelism)	Compute Set
Tasklet	Codelet
Containers	Data variables
DaCE Streams	Poplar Streams
Data Copy	Copy APIs
...	...

Mapping between DaCe [8] & Graphcore Poplar Constructs
(Work in Progress)

FUTURE WORK & CONCLUSIONS

- Continue porting of various computational science applications to AI accelerators and understand optimization strategies.
- Continue efforts to map existing programming models to AI accelerators.
- There is need of higher level common abstraction layer to ease programmability as well as improve portability across architectures.

NOTABLE REFERENCES

[1] M. Emani et al., "A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads," PMBS@SC2022
[2] M. Emani et al., "Toward a Holistic Performance Evaluation of Large Language Models Across Diverse AI Accelerators," HCW@IPDPS2024
[3] Chitty-Venkata, S Raskar et al., "LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators"
[4] Barik, Raskar et al., "Characterizing the Performance of Triangle Counting on Graphcore's IPU Architecture", WACCPD@SC23
[5] Tramm, John et al., "Efficient algorithms for Monte Carlo particle transport on AI accelerator hardware, Computer Physics Communications, Volume 298, May 2024.
[6] Hatem Ltaief et al., "Scaling the "Memory Wall" for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems", ACM Gordon Bell Finalists, SC23.
[7] Monsalve et al., "A Pathway to OpenMP in the GraphCore Architecture", LDRD Expedition Report 2022, 2023
[8] DaCe Resources, https://spcl.inf.ethz.ch/Research/DAPP/#dace

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Acknowledgement: This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357

Argonne NATIONAL LABORATORY
Argonne Leadership Computing Facility