20TH NOVEMBER, 2024

AI TESTBEDS AT ARGONNE LEADERSHIP COMPUTE FACILITY

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

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

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/

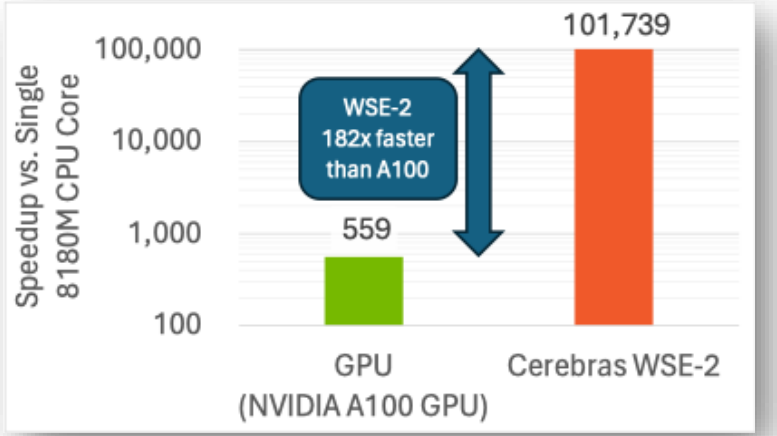
Efficient Algorithms for Monte Carlo Particle Transport on AI Accelerator Hardware

The Science

- Cerebras Wafer-Scale Engine 2 (WSE-2) while not intended for traditional modeling and simulation workloads, aspects of these accelerators make them attractive for some simulation algorithms, nonetheless.
- A new algorithms and performance optimization strategies to enable a key Monte Carlo (MC) particle transport simulation kernel to effectively use the device.
- Speedups of **182x** over a single GPU

The Impact

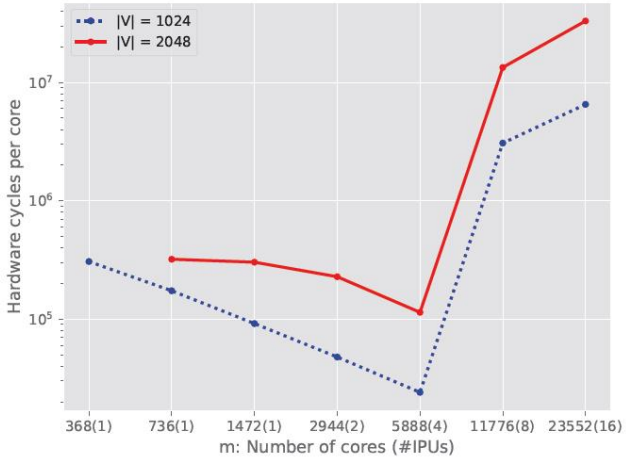
- Acceleration of a full MC particle transport code on WSE-2 would be possible.
- AI accelerators, such as the WSE-2, could offer significant advantages to traditional simulation workloads
- development of higher-level programming models to more readily enable software development and exploration could have a tremendous impact for HPC simulations.



Speedup vs. serial CPU execution for macroscopic cross section lookup kernel (adapted from XSBench, John Tramm ANL).

Characterizing the Performance of Triangle Counting on Graphcore's IPU Architecture

- **Scaling Performance**
  - Strong Scaling: Good performance within a node
  - Weak Scaling: Gooup to 2944 cores.

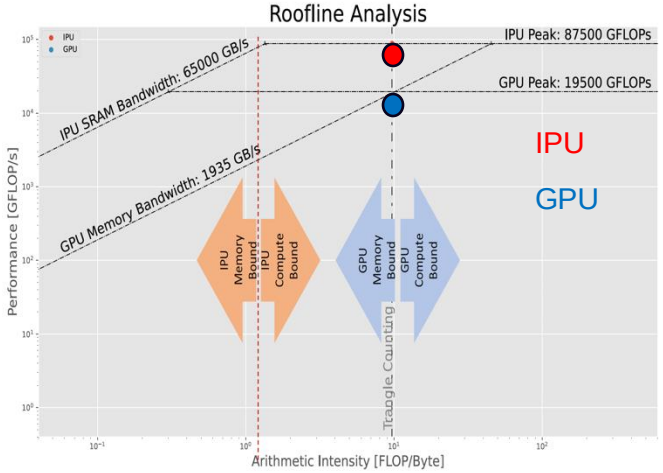


- Average speedup of up to **5.3x** with 4 IPUs over single A100 GPU
- A100 Memory Bound while optimized IPU implementation is Compute Bound
- High Compilation times

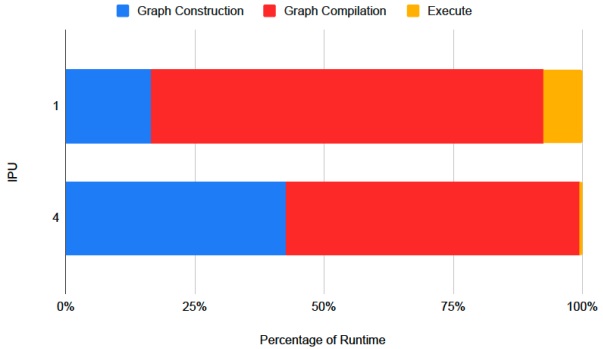
|V|	#Cores	Cycles per core
1,024	368	306,701
2,048	736	320,573
4,096	1,472	359,654
8,192	2,944	376,860

Characterizing the Performance of Triangle Counting on Graphcore's IPU Architecture

- Scaling Performance
  - Strong Scaling: Good performance within a node
  - Weak Scaling: Good up to 2944 cores.

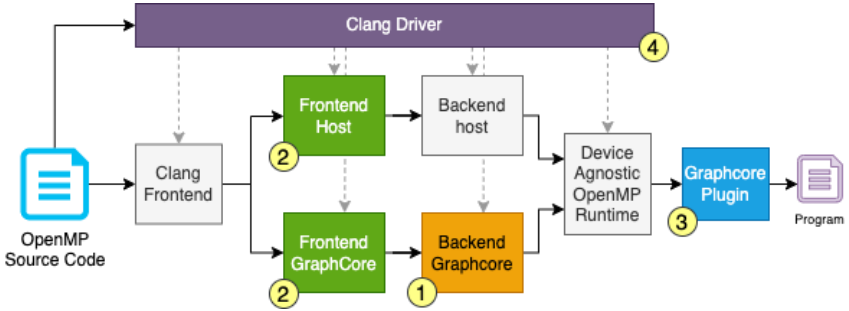


- Average speedup of up to **5.3x** with 4 IPUs over single A100 GPU
- A100 Memory Bound while optimized IPU implementation is Compute Bound
- High Compilation times



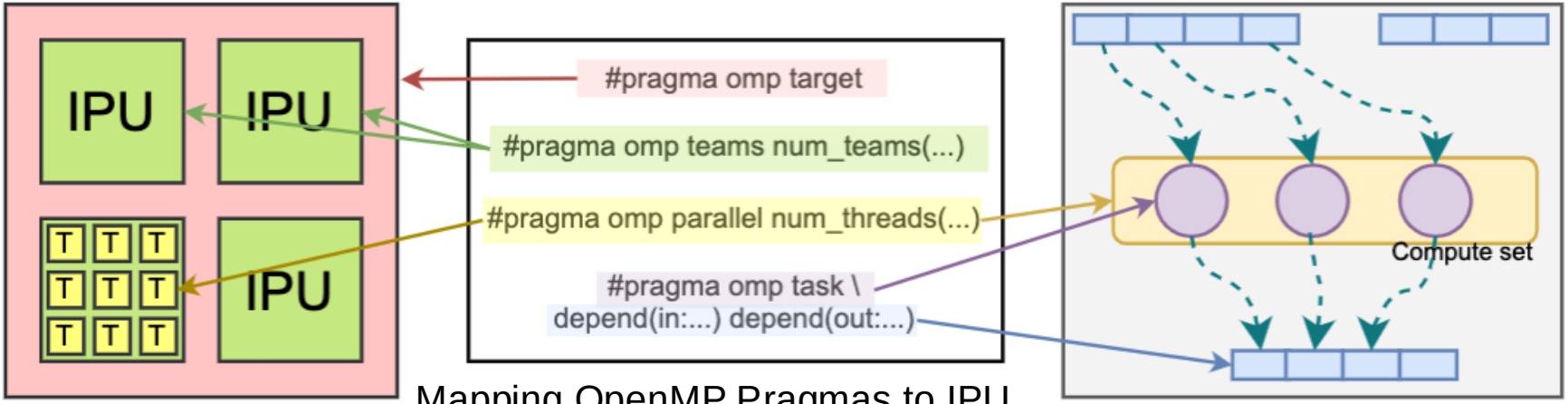
Characterizing the Performance of Triangle Counting on Graphcore's IPU Architecture
Reet Barik, Siddhisanket Raskar et al, Argonne National Laboratory

Exploring Openmp Target Offloading For The Graphcore Architecture



OpenMP compilation pipeline

Explore poplar library to connect to the device using OpenMP Target Offload



Mapping OpenMP Pragmas to IPU

Exploring OpenMP target offloading for the GraphCore architecture
Jose M Monsalve Diaz, Siddhisanket Raskar et al. LLVM Developer's meeting, 2022.

Ongoing work

Porting Applications to AI Testbed Systems

ECP Proxy Applications

- HACCmk
- Xsbench
- NEKbone
- miniQMC

Build Higher Level Abstractions

DaCe is a data-centric parallel programming framework from ETH Zurich that optimizes Python/NumPy code for high-performance execution on CPUs, GPUs, and FPGAs using a transformable intermediate representation called Stateful DataFlow multiGraph (SDFG)

- DaCE Backend for Graphcore

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
...	...

Sameeran Joshi,
University of Utah, Argonne National Laboratory

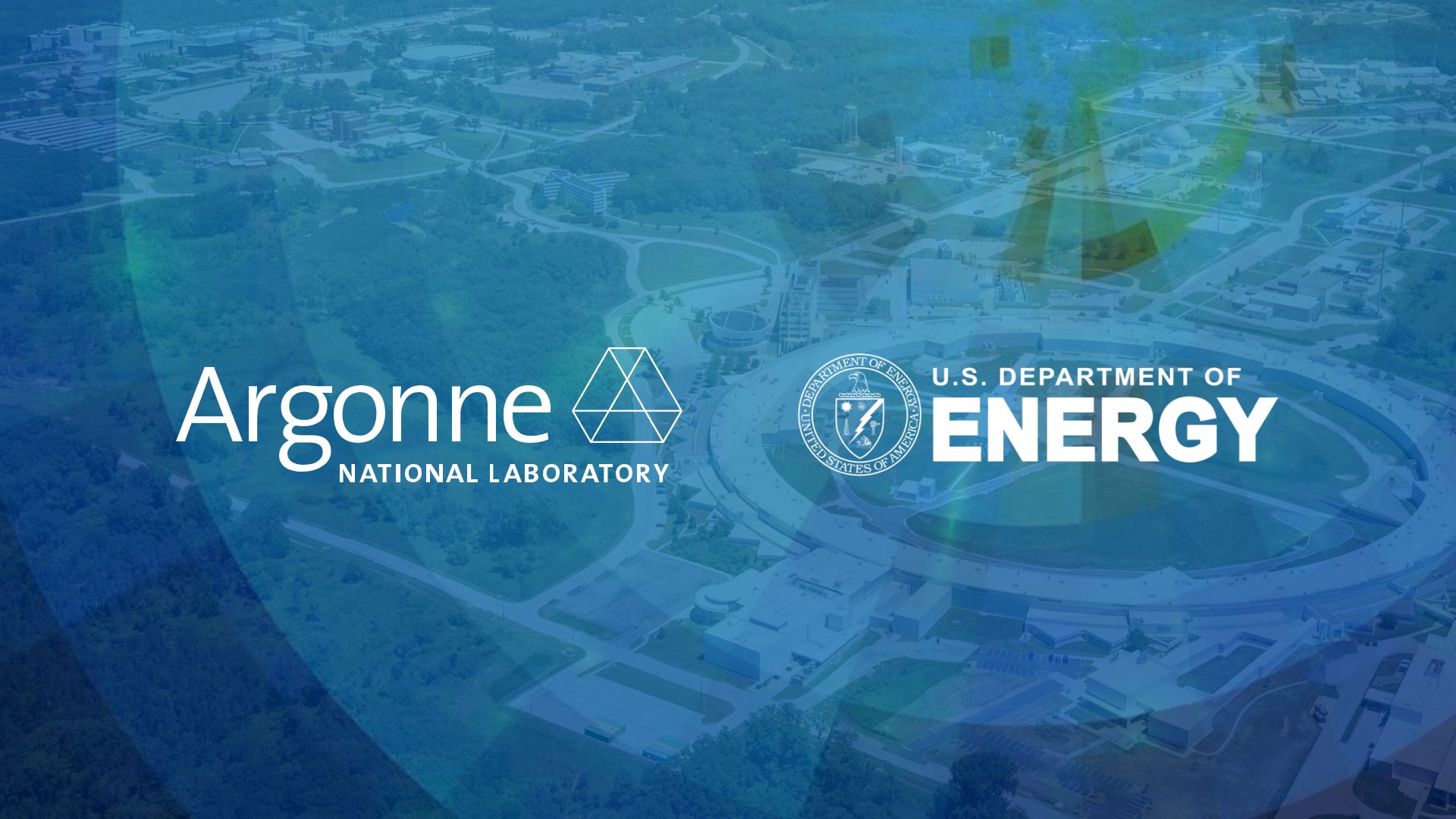
AI Accelerators for traditional HPC

Benefits

- Significant Performance benefits over CPUs and GPUs.
- High memory bandwidth yields to high compute performance
- Programming models allow description of programs in truly parallel and scalable manner

Challenges

- Under development software stack and constantly evolving software stack
- Low level programming gives more flexibility at cost of higher learning curve
- Significant compilation and projection times.
- High Precision Support


Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT OF
ENERGY