

Joseph Lee - EPCC

Murali Emani - Argonne National Laboratory

Josef Weidendorfer - Leibniz Supercomputing Centre

Lukas Gianinazzi - ETH Zurich

Leighton Wilson - Cerebras Systems



Democratizing AI Accelerators for HPC Applications: Challenges, Success, and Support

ISC 2024 | MAY 12 – 16 | HAMBURG, GERMANY | #ISC24

Session Goals



Build community of interested & experienced users



Share success stories and experience



Feedback for vendors for support, training, and development

Why is this relevant?

- Lots of new and exciting AI accelerator hardware, e.g.



- High compute capability and large bandwidth
- How can we leverage these hardware for HPC applications?

Success stories

Efficient algorithms for Monte Carlo particle transport on AI accelerator hardware [☆]

John Tramm ^{a,*}, Bryce Allen ^{a,b}, Kazutomo Yoshii ^a, Andrew Siegel ^a, Leighton Wilson ^c

Scaling the “Memory Wall” for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems

Hatem Ltaief
Yuxi Hong

Leighton Wilson
Mathias Jacquelin

Matteo Ravasi
David Keyes

Steering Customized AI Architectures for HPC Scientific Applications

Hatem Ltaief^{1(✉)}, Yuxi Hong¹, Adel Dabah¹, Rabab Alomairy¹, Sameh Abdullah¹, Chris Goreczny³, Pawel Gepner⁴, Matteo Ravasi², Damien Gratadour⁵, and David Keyes¹

PARENDI: Thousand-Way Parallel RTL Simulation

Mahyar Emami

Thomas Bourgeat

James R. Larus

RUNNING JULIA ON GRAPHCORE IPUS

Written By:
Mosè Giordano

High Performance Monte Carlo Simulation of Ising Model on TPU Clusters

Kun Yang*
kuny.work@gmail.com
Google Research

Yi-Fan Chen†
yifanchen@google.com
Google Research

Georgios Roumpos
roumposg@google.com
Google Research

Chris Colby
ccolby@google.com
Google Research

John Anderson
janders@google.com
Google Research

Record Acceleration of the Two-Dimensional Ising Model Using High-Performance Wafer Scale Engine

Dirk Van Essendelft¹, Hayl Almolyki², Wei Shi³, Terry Jordan⁴, Mei-Yu Wang⁵, Wissam A. Saidi⁶

Quick Survey

- Experience with AI accelerators?



Quick Survey

- Experience with HPC applications on AI accelerators?



Cerebras SDK for HPC Research and Applications

Leighton Wilson

leighton.wilson@cerebras.net

ISC 2024



Cerebras Wafer-Scale Engine (WSE-2)

The (2nd) Largest Chip in the World

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

6.8 PetaFLOPS dense fp16

7nm process technology

Cluster-scale acceleration on a single chip



Cerebras Wafer-Scale Engine (WSE-3)

The Largest Chip in the World

900,000 cores optimized for sparse linear algebra

46,225 mm² silicon

4.0 trillion transistors

44 Gigabytes of on-chip memory

24.5 PByte/s memory bandwidth

245 Pbit/s fabric bandwidth

12.5 PetaFLOPS dense fp16

5nm process technology

Cluster-scale acceleration on a single chip

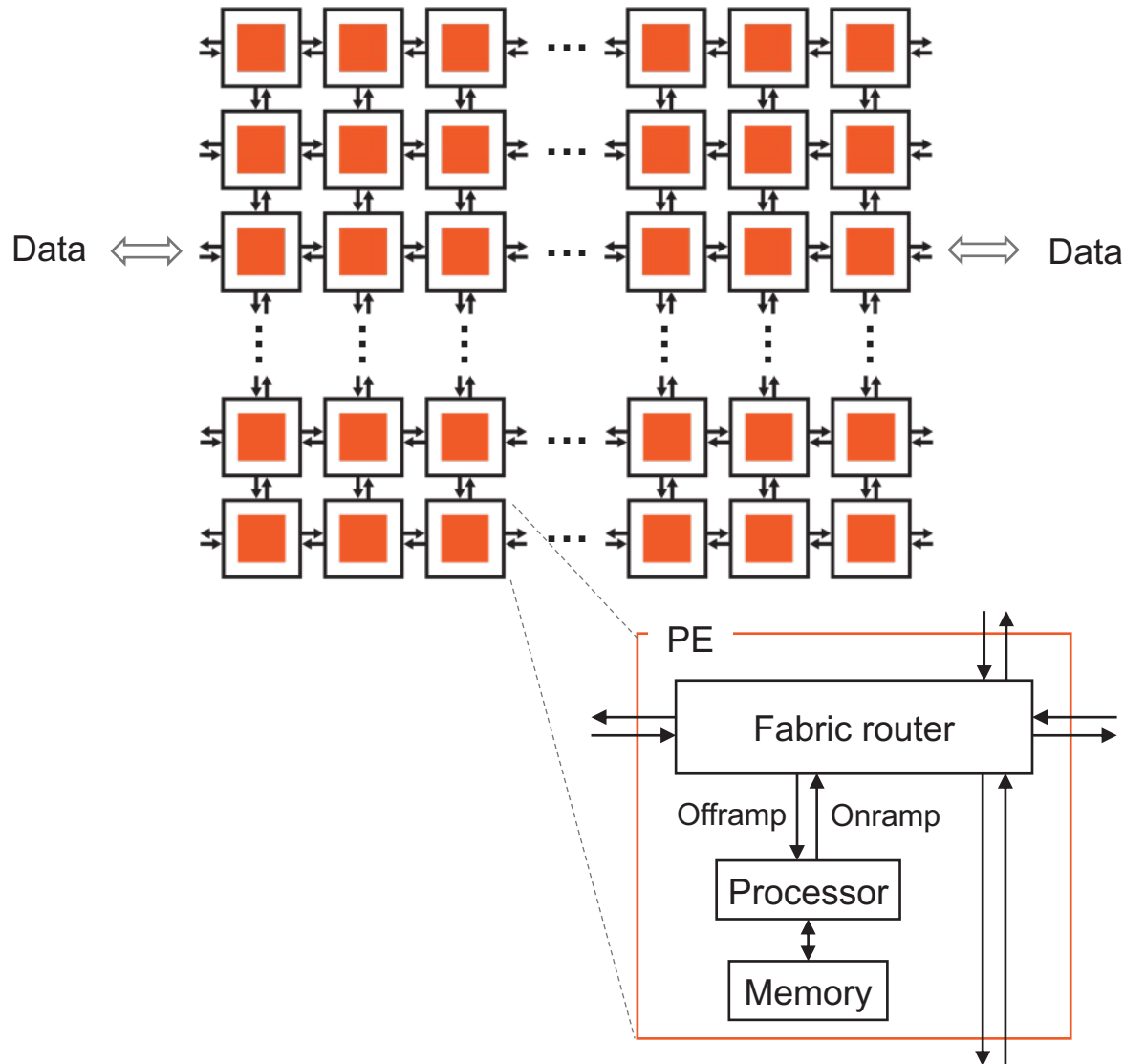
Cerebras CS System

The world's most powerful AI and HPC accelerator

- Powered by WSE
- Install, deploy easily into a standard rack
- Programmable via our SDK or PyTorch



CS Architecture Basics



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

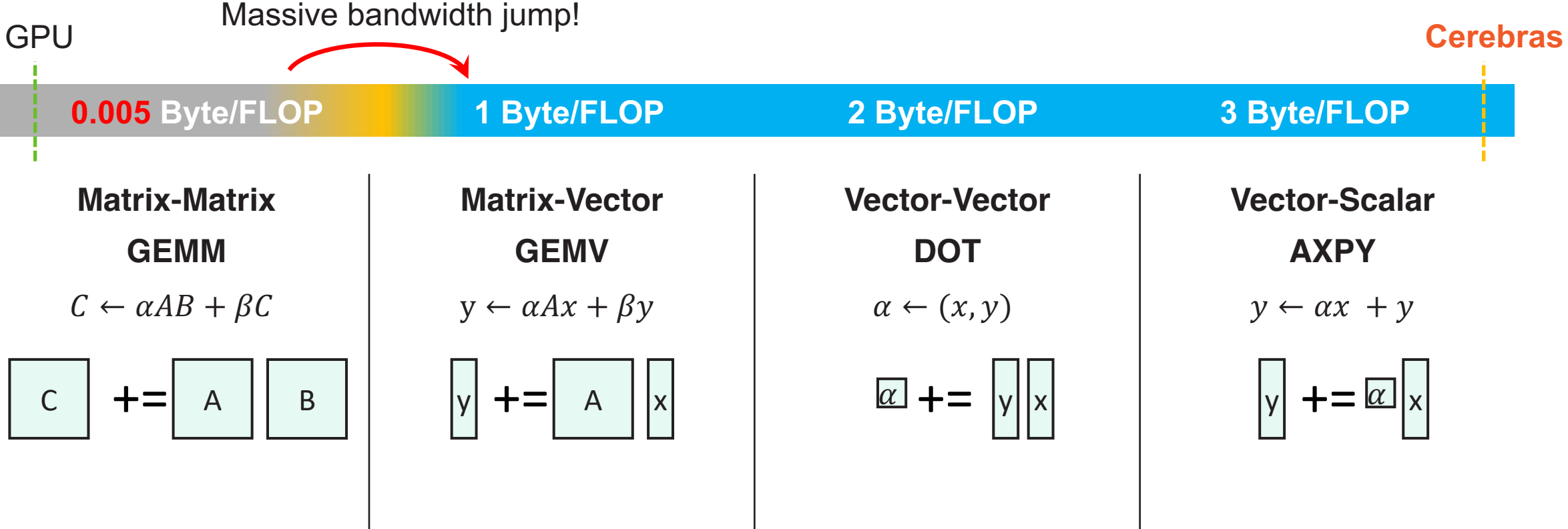
Flexible communication

- Programmable router
- Static or dynamic routes (**colors**)
- Data packets (**wavelets**) passed between PEs
- Single cycle PE-to-PE communication

Fast memory

- 48 kB SRAM per PE for data and instructions
- 1 cycle read/write

Memory performance at all BLAS levels



Cerebras Supports Two Programming Paradigms

For AI Users, Cerebras ML stack provides **familiar, high-level** programmability with popular ML frameworks and compatibility with 3P model repos and ML Ops tools

 PyTorch



Hugging Face



Weights & Biases

For HPC Users, Cerebras SDK provides **flexible, lower-level** programmability and access to HW performance features.

Cerebras SDK & CSL

Cerebras SDK

A general-purpose parallel-computing platform and API allowing software developers to write custom programs (“kernels”) for Cerebras systems.

Language

CSL: Cerebras Software Language

Host APIs with Python

Libraries

Optimized primitives

Tools

Visualization

Debugger

Simulator

The screenshot displays the Cerebras SDK GUI with several panels:

- Colors:** A list of color selection options: Select All, 1 x_in, 2 Ax_out, 3 y_out, and 4 b_in.
- Grid Visualization:** A 6x6 grid of nodes connected by lines, with a black box highlighting a central 2x2 area.
- Symbols:** A table listing symbols and their types:

Name	Type
A	NOTYPE
Ax_temp	NOTYPE
memcpy	NOTYPE
memset	NOTYPE
memcpy	FUNC
- Instruction Trace:** A table showing execution details:

Cycle	OP Addr	OP Name	Dest	Src0
344	0x3120	s class	0x0 (0x38b7)	0x0 (0x3040)
- Source Code:** A code editor showing C code:

```
1 var global: i16 = 0;  
2  
3 color main_color = 0;  
4 color output_color = 1;  
5 const dsd = @get_dsd(fabou_t_dsd, {fabric_color =  
  output_color, extent = 1});  
6  
7 task main_task(wavelet_data: i16) void {
```
- Wavelet Trace:** A table showing wavelet execution:

Cycle	Color	Ctrl	Link	Header
3	3	0	W	0x0000
1890	3	0	E	0x0000
1000	2	0	E	0x0000

Copyright © Cerebras 2021

Cerebras SDK

A general-purpose parallel-computing platform and API allowing software developers to write custom programs (“kernels”) for Cerebras systems.

Language

CSL: Cerebras Software Language

Host APIs with Python

Libraries

Optimized primitives

Tools

Visualization

Debugger

Simulator

The screenshot displays the Cerebras SDK GUI with several panels:

- Colors:** A list of color selection options: Select All, 1 x_in, 2 Ax_out, 3 y_out, and 4 b_in.
- Grid Visualization:** A 6x6 grid of nodes connected by lines, with a central 2x2 area highlighted by a black box.
- Symbols:** A table listing symbols and their types:

Name	Type
A	NOTYPE
Ax_temp	NOTYPE
memcpy	NOTYPE
memset	NOTYPE
memcpy	FUNC
- Instruction Trace:** A table showing execution details:

Cycle	OP Addr	OP Name	Dest	Src0
344	0x3120	s class	0x0 (0x38b7)	0x0 (0x3040)
- Source Code:** A code editor showing C code:

```
1 var global: i16 = 0;  
2  
3 color main_color = 0;  
4 color output_color = 1;  
5 const dsd = @get_dsd(fabou_t_dsd, {fabric_color =  
  output_color, extent = 1});  
6  
7 task main_task(wavelet_data: i16) void {
```
- Wavelet Trace:** A table showing wavelet execution:

Cycle	Color	Ctrl	Link	Header
3	3	0	W	0x0000
1890	3	0	E	0x0000

SDK Example Programs Available

Repository: github.com/Cerebras/csl-examples

- Introductory Tutorials
- GEMV
- GEMM
- Cholesky Decomposition
- 1D and 2D FFT
- 7-Point Stencil SpMV
- Power Method
- Conjugate Gradient
- Preconditioned Conjugate Gradient
- Finite Difference Stencil Computations
- Mandelbrot Set Generator
- Shift-Add Multiplication
- Hypersparse SpMV
- Histogram Computation

SDK Usage and Impact

Over the past year, SDK has evolved from a closed tool requiring NDA access to a public platform for Wafer-Scale Computing. We're supporting more research and publications than ever.

Scaling the "Memory Wall" for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems

Hatem Ltaief
Yuxi Hong
Extreme Computing Research Center

Leighton Wilson
Mathias Jacquelin
Cerebras Systems Inc.

Matteo Ravasi
David Keyes
Extreme Computing Research Center

Using Wafer-Scale AI Hardware for Traditional HPC Simulation Workloads: A Case Study in Developing a Monte Carlo Particle Transport Application for the Cerebras WSE2 AI Accelerator

Kazutomo Yoshii* Andrew Siegel* Leighton Wilson†

importance to both fission and fusion reactor simulation fields, and because the MC algorithm has historically failed to achieve more than a few percent of theoretical peak FLOP performance due to its inherently stochastic memory access patterns [6].

Near-Optimal Wafer-Scale Reduce

Piotr Luczynski
Department of Computer Science
ETH Zurich

Lukas Gianinazzi
Department of Computer Science
ETH Zurich

Patrick Iff
Department of Computer Science
ETH Zurich

Leighton Wilson

Daniele De Sensi
Sapienza University of Rome

Torsten Hoefler
Department of Computer Science
ETH Zurich

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

and various other HPC applications [35, 38, 51, 58]. However, maximizing performance on this architecture necessitates tailoring communication patterns to its unique characteristics. This need motivates our investigation of Reduce and AllReduce on the WSE.

1.2 Limitations of state-of-the-art

Current wafer-scale Reduce and AllReduce implementations are primarily optimized for extreme vector sizes. This means they are

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



Communication Collectives for the Cerebras Wafer-Scale Engine

Bachelor Thesis

Piotr Luczynski
pluczynski@ethz.ch

Massively Distributed Finite-Volume Flux Computation

Ryuichi Sai*
TotalEnergies EP Research & Technology US, LLC.
Houston, Texas, USA
ryuichi@rice.edu

Mathias Jacquelin
Cerebras Systems
Sunnyvale, California, USA

François P. Hamon
TotalEnergies EP Research & Technology US, LLC.
Houston, Texas, USA

Mauricio Araya-Polo
TotalEnergies EP Research & Technology US, LLC.
Houston, Texas, USA

Randolph R. Settgaest
Lawrence Livermore National Laboratory
Livermore, California, USA

Monte Carlo with Single-Cycle Latency: Optimization of a Continuous Energy Cross Section Lookup Kernel for AI Accelerator Hardware

John Tramm^{1,*}, Bryce Allen^{1,2}, Kazutomo Yoshii¹, Andrew Siegel¹

¹Lawrence Livermore National Laboratory, Livermore, CA, USA; ²University of Chicago, Chicago, IL

by ANS/

CereSZ: Enabling and Scaling Error-bounded Lossy Compression on Cerebras CS-2

Anonymous Author(s)

Trackable Agent-based Evolution Models at Wafer Scale

Matthew Andres Moreno^{1,2,3,*}, Connor Yang⁴, Emily Dolson^{5,6}, and Luis Zaman^{1,2}

¹Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, United States

²Center for the Study of Complex Systems, University of Michigan, Ann Arbor, United States

³Michigan Institute for Data Science, University of Michigan, Ann Arbor, United States

⁴Undergraduate Research Opportunities Program, University of Michigan, Ann Arbor, United States

⁵Department of Computer Science and Engineering, Michigan State University, East Lansing, United States

⁶Program in Ecology, Evolution, and Behavior, Michigan State University, East Lansing, United States

*corresponding author: morenoma@umich.edu

Abstract

Continuing improvements in computing hardware are poised to transform capabilities for *in silico* modeling of cross-scale phenomena underlying major open questions in evolutionary biology and artificial life, such as transitions in individuality, eco-evolutionary dynamics, and rare evolutionary events. Emerging ML/AI-oriented hardware accelerators like the 850,000 processor Cerebras Wafer



of data within a short time impose considerable challenges, even on high-performance computers.

To tackle this big data challenge, lossy compression techniques [8, 21, 25, 27, 35] have been commonly used in scientific applications to reduce the data size while maintaining a user-specified error limit. Beyond the traditional compressors on CPU, accelerating data compression on heterogeneous processors, such as FPGA [37] and GPU [13, 38, 42, 43], has become increasingly important for real-time compression tasks (e.g. reducing data stream intensity). For instance, cuSZ [38] parallelizes quantization, prediction, and Huffman encoding on NVIDIA GPU, improving the runtime performance of large-scale cosmic simulation [16] and deep learning training systems [17].

In recent years, there has been a boom in AI chips to meet the high computation demand of AI workloads. Among the

latency memory, making it an methods. Recent work has gains for the continuous sport method. In the present od based off of the fractional

erouiche and.ntnu.no rondheim way

Andrei Ivanov
ivanov@inf.ethz.ch
ETH Zurich
Switzerland

ABSTRACT

Sparse matrix multiplications are a fundamental component of various scientific disciplines, including computational physics, machine learning, and data analysis. They involve efficient manipulation of matrices with a large number of zero elements, enabling more compact and computationally efficient representations of complex data structures. This work optimizes sparse matrix multiplications on a novel architecture, namely the Cerebras WSE-2, through exploration of sparse data formats and optimization strategies, leading to significant performance improvements. In contrast to previous

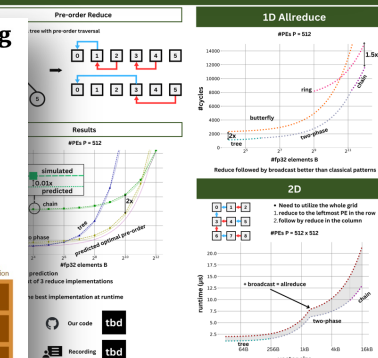
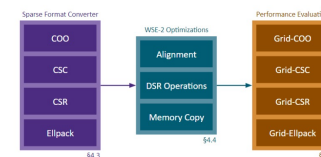
Multiplication on Cerebras WSE-2: Evaluating M Algorithms in Spatial Computing

erouiche and.ntnu.no rondheim way

Andrei Ivanov
ivanov@inf.ethz.ch
ETH Zurich
Switzerland

Filip Dobrosavljević
dofilip@student.ethz.ch
ETH Zurich
Switzerland

Torsten Hoefler
torsten.hoefler@inf.ethz.ch
ETH Zurich
Switzerland



Cerebras Systems Inc. All Rights Reserved

SDK Usage and Impact

Over the past year, SDK has evolved from a closed tool requiring NDA access to a public platform for Wafer-Scale Computing. We're supporting more research and publications than ever.

Near-Optimal Wafer-Scale Reduce

Piotr Luczynski
Department of Computer Science
ETH Zurich

Lukas Gianinazzi
Department of Computer Science
ETH Zurich

Patrick Iff
Department of Computer Science
ETH Zurich

Leighton Wilson

Daniele De Sensi
Sapienza University of Rome

Torsten Hoefler
Department of Computer Science
ETH Zurich

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Implementation and Evaluation of Matrix Profile Algorithms on the Cerebras Wafer-Scale Engine

Vyas Giridharan

CereSZ: Enabling and Scaling Error-bounded Lossy Compression on Cerebras CS-2

Anonymous Author(s)

Trackable Agent-based Evolution Models at Wafer Scale

Matthew Andres Moreno^{1,2,3,*}, Connor Yang⁴, Emily Dolson^{5,6}, and Luis Zaman^{1,2}

¹Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, United States

²Center for the Study of Complex Systems, University of Michigan, Ann Arbor, United States

³Michigan Institute for Data Science, University of Michigan, Ann Arbor, United States

⁴Undergraduate Research Opportunities Program, University of Michigan, Ann Arbor, United States

⁵Department of Computer Science and Engineering, Michigan State University, East Lansing, United States

⁶Program in Ecology, Evolution, and Behavior, Michigan State University, East Lansing, United States

*corresponding author: morenoma@umich.edu

Abstract

Continuing improvements in computing hardware are poised to transform capabilities for *in silico* modeling of cross-scale phenomena underlying major open questions in evolutionary biology and artificial life, such as transitions in individuality, eco-evolutionary dynamics, and rare evolutionary events. Emerging ML/AI-oriented hardware accelerators like the 850,000 processor Cerebras Wafer



Monte Carlo with Single-Cycle Latency: Optimization of a Continuous Energy Cross Section Lookup Kernel for AI Accelerator Hardware

John Tramm^{1,2}, Bryce Allen^{1,2}, Kazutomo Yoshii¹, Andrew Siegel¹

¹University of Chicago, Chicago, IL

²University of Chicago, Chicago, IL

by ANS]

of data within a short time impose considerable challenges, even on high-performance computers.

To tackle this big data challenge, lossy compression techniques [8, 21, 25, 27, 35] have been commonly used in scientific applications to reduce the data size while maintaining a user-specified error limit. Beyond the traditional compressors on CPU, accelerating data compression on heterogeneous processors, such as FPGA [37] and GPU [13, 38, 42, 43], has become increasingly important for real-time compression tasks (e.g. reducing data stream intensity). For instance, cuSZ [38] parallelizes quantization, prediction, and Huffman encoding on NVIDIA GPU, improving the runtime performance of large-scale cosmic simulation [16] and deep learning training systems [17].

In recent years, there has been a boom in AI chips to meet the high computation demand of AI workloads. Among the

latency memory, making it an methods. Recent work has gains for the continuous sport method. In the present od based off of the fractional

ABSTRACT

Sparse matrix multiplications are a fundamental component of various scientific disciplines, including computational physics, machine learning, and data analysis. They involve efficient manipulation of matrices with a large number of zero elements, enabling more compact and computationally efficient representations of complex data structures. This work optimizes sparse matrix multiplications on a novel architecture, namely the Cerebras WSE-2, through exploration of sparse data formats and optimization strategies, leading to significant performance improvements. In contrast to previous

Scaling the "Memory Wall" for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems

Hatem Itjaief
Yuxi Hong
Extreme Computing Research Center

Leighton Wilson
Mathias Jacquelin
Cerebras Systems Inc.

Matteo Ravasi
David Keyes
Extreme Computing Research Center

Using Wafer-Scale AI Hardware for Traditional HPC Simulation Workloads: A Case Study in Developing a Monte Carlo Particle Transport Application for the Cerebras WSE2 AI Accelerator

Kazutomo Yoshii* Andrew Siegel* Leighton Wilson†

Communication Collectives for the Cerebras Wafer-Scale Engine

Bachelor Thesis

Piotr Luczynski
pluczynski@ethz.ch

Massively Distributed Finite-Volume Flux Computation

Ryuichi Sai*
TotalEnergies EP Research & Technology US, LLC.
Houston, Texas, USA
ryuichi@rice.edu

Mathias Jacquelin
Cerebras Systems
Sunnyvale, California, USA

François P. Hamon
TotalEnergies EP Research & Technology US, LLC.
Houston, Texas, USA

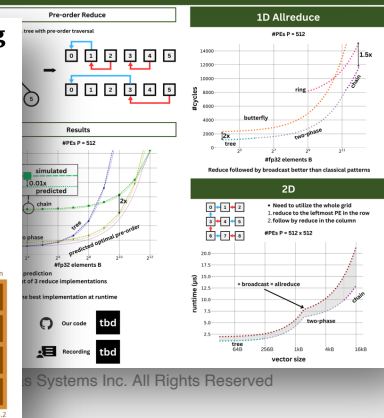
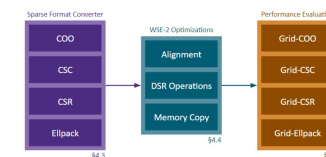
Mauricio Araya-Polo
TotalEnergies EP Research & Technology US, LLC.
Houston, Texas, USA

Randolph R. Settgaest
Lawrence Livermore National Laboratory
Livermore, California, USA

Multiplication on Cerebras WSE-2: Evaluating M Algorithms in Spatial Computing

Filip Dobrosavljević
dofilip@student.ethz.ch
ETH Zurich
Switzerland

Torsten Hoefler
torsten.hoefler@inf.ethz.ch
ETH Zurich
Switzerland



Cerebras Systems Inc. All Rights Reserved

Cerebras SDK Developments

A general-purpose parallel-computing platform and API allowing software developers to write custom programs (“kernels”) for Cerebras systems.

Language

CSL: Cerebras Software Language

Host APIs with Python

Libraries

Optimized primitives

Tools

Visualization

Debugger

Simulator

The screenshot displays the Cerebras SDK GUI with several components:

- Current folder:** <filepath containing artifacts used in the GUI> with a SUBMIT button.
- Colors:** A grid visualization of a neural network with a highlighted cell.
- Symbols:** A table listing symbols and their types.
- Instruction Trace:** A table showing instruction details.
- Source Code:** A code editor showing C++ code.
- Wavelet Trace:** A table showing wavelet data.

Red callouts highlight the following features:

- C++ host code**
- More collectives
 - Libraries for fabric control
 - Linear algebra routines
- printf debugging in simulator
 - Totally new debugging experience

Name	Type
A	NOTYPE
Ax_temp	NOTYPE
memcpy	NOTYPE
memset	NOTYPE
memcpy	FUNC

Cycle	Color	Ctrl	Link	Header
3	3	0	W	0x0000
1890	3	0	E	0x0000
1000	2	0	E	0x0000

```
344 0x3120 s class 0x0 0x0 (0x38b7) (0x3040)
5 const dsd = @get_dsd(fabout_dsd, {fabric_color =
output_color, extent = 1});
6
7 task main_task(wavelet_data: i16) void {
```

Cycle	Color	Ctrl	Link	Header
3	3	0	W	0x0000
1890	3	0	E	0x0000
1000	2	0	E	0x0000

Copyright © Cerebras 2021

CS1 [6 x 6] ALL SELECTED PE: [2, 1]

SDK Access

Get local access to the SDK simulator!

- Email developer@cerebras.net for access

Join the Cerebras Developer Community

- Forums at discourse.cerebras.net

View our public SDK examples GitHub repository

- See github.com/Cerebras/csl-examples

Partner systems at ANL, EPCC, PSC

Questions? leighton.wilson@cerebras.net



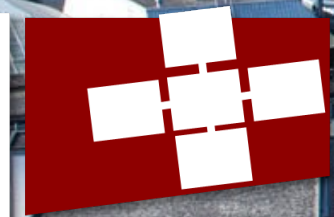
discourse.cerebras.net



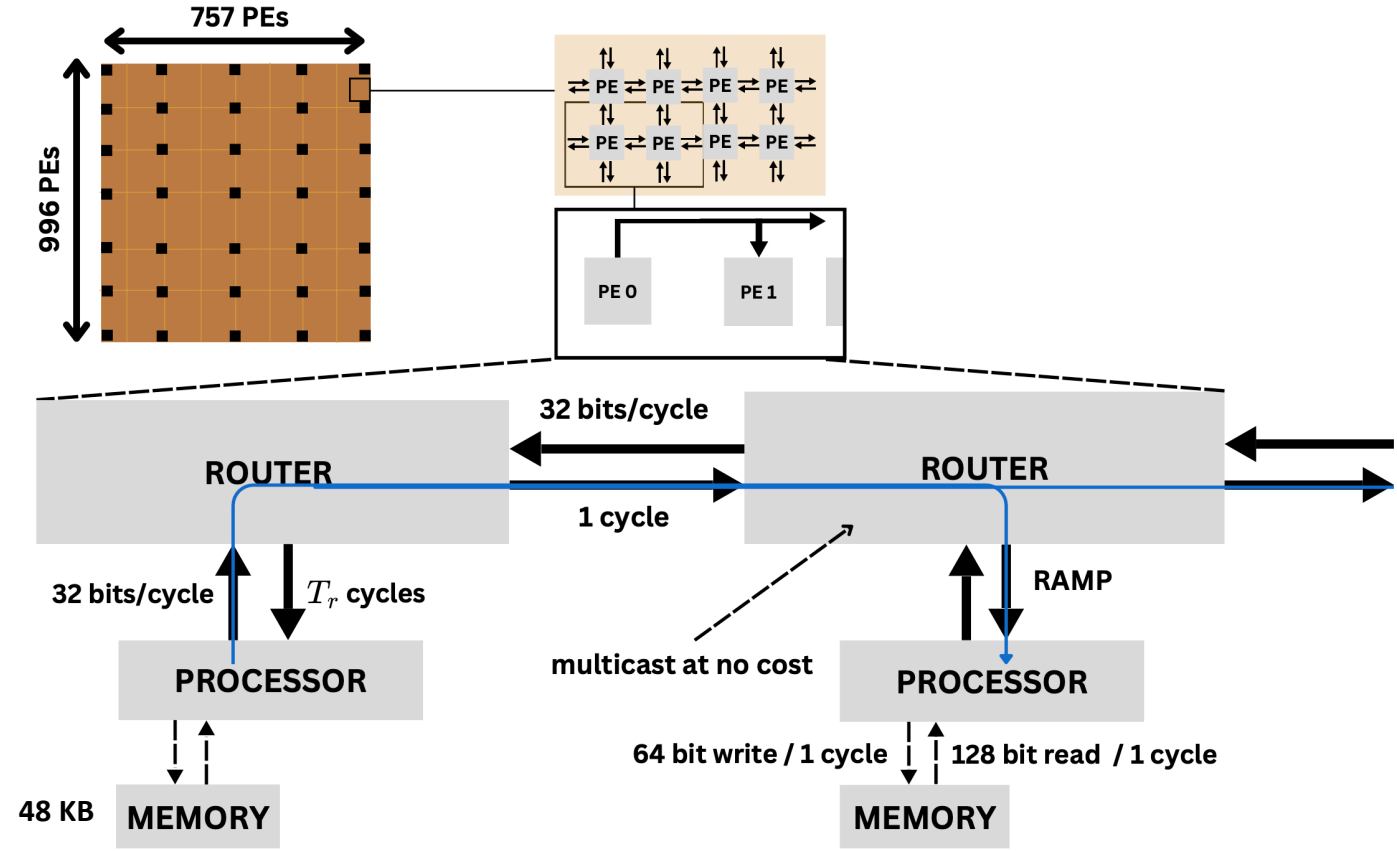
cerebras.net/developers/sdk-request

LUKAS GIANINAZZI, PIOTR LUCZYNSKI, LEIGHTON WILSON, P. IFF, D. DE SENSI, M. BESTA, S. ASHKBOS, Y. BAUMANN, T. BEN-NUN, T. HOEFLER

Performance Models Enable HPC on AI Accelerators

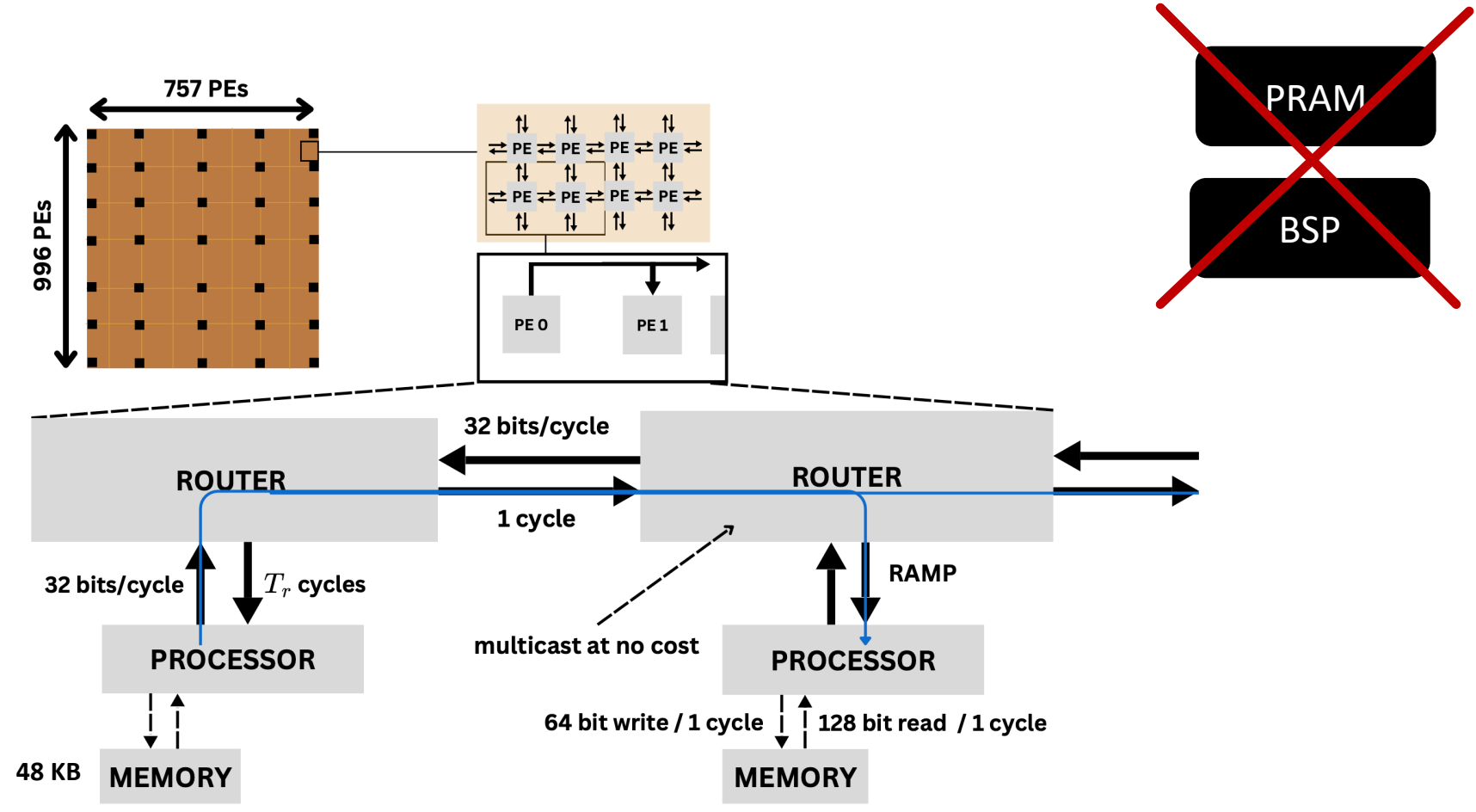


Modeling AI Accelerators



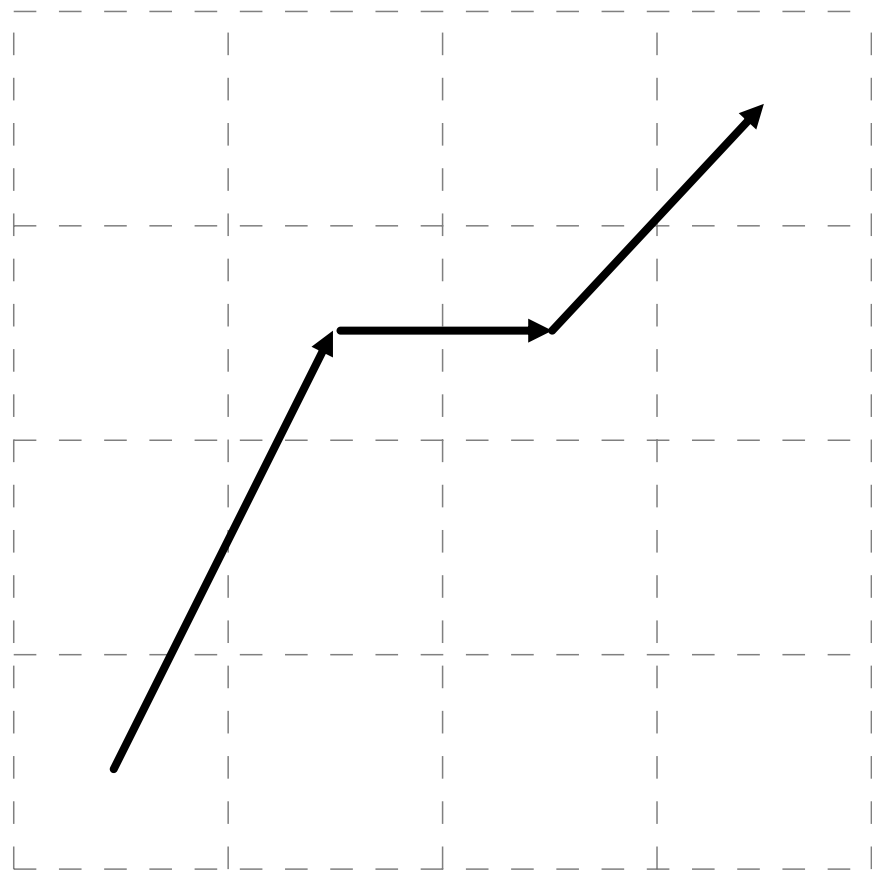
Cerebras CS-2 Wafer-Scale Engine

Modeling AI Accelerators

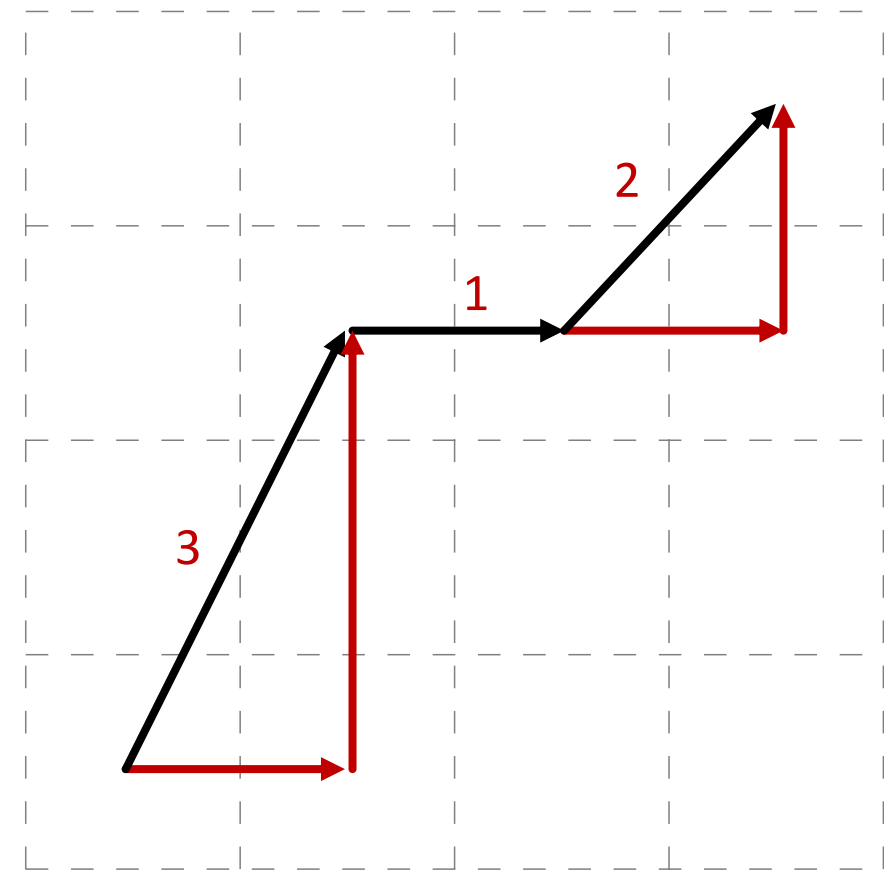


Cerebras CS-2 Wafer-Scale Engine

Modeling AI Accelerators – Spatial Model

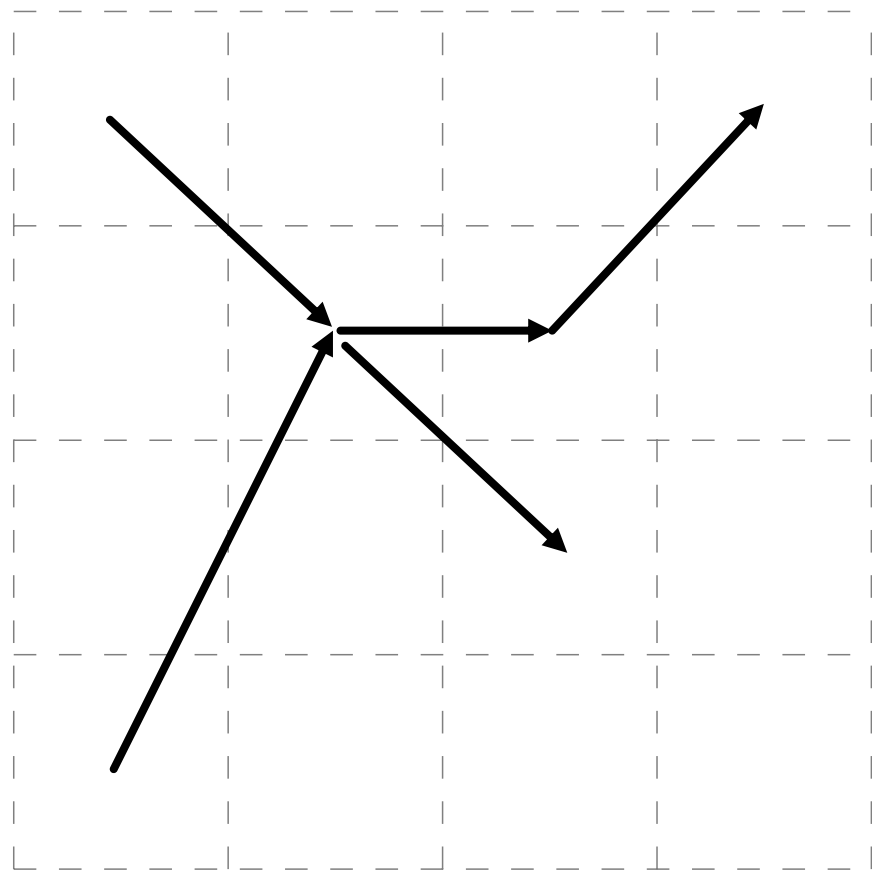


Modeling AI Accelerators – Spatial Model



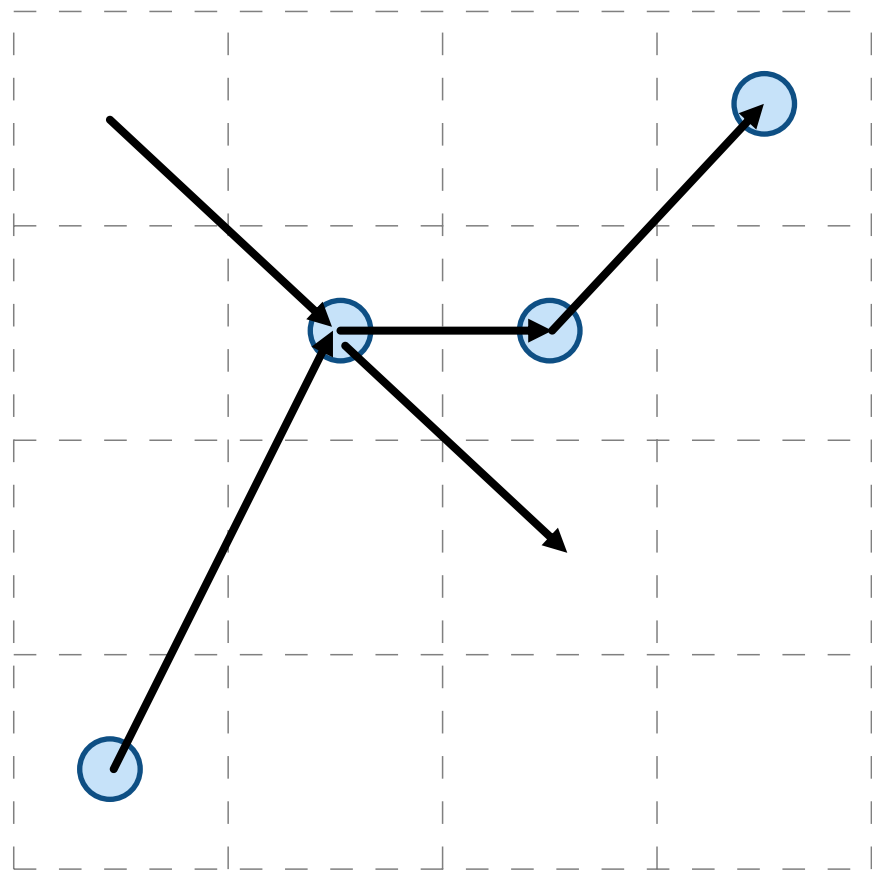
Distance 6

Modeling AI Accelerators – Spatial Model



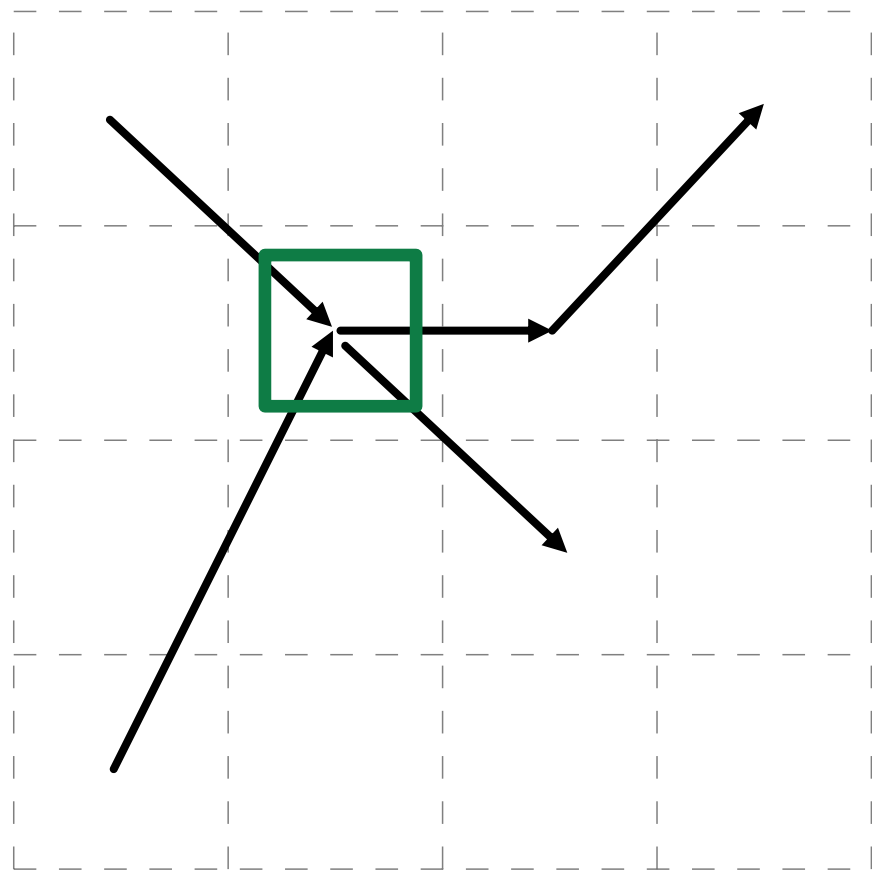
Distance 6	
Maximum 6	Total 10

Modeling AI Accelerators – Spatial Model



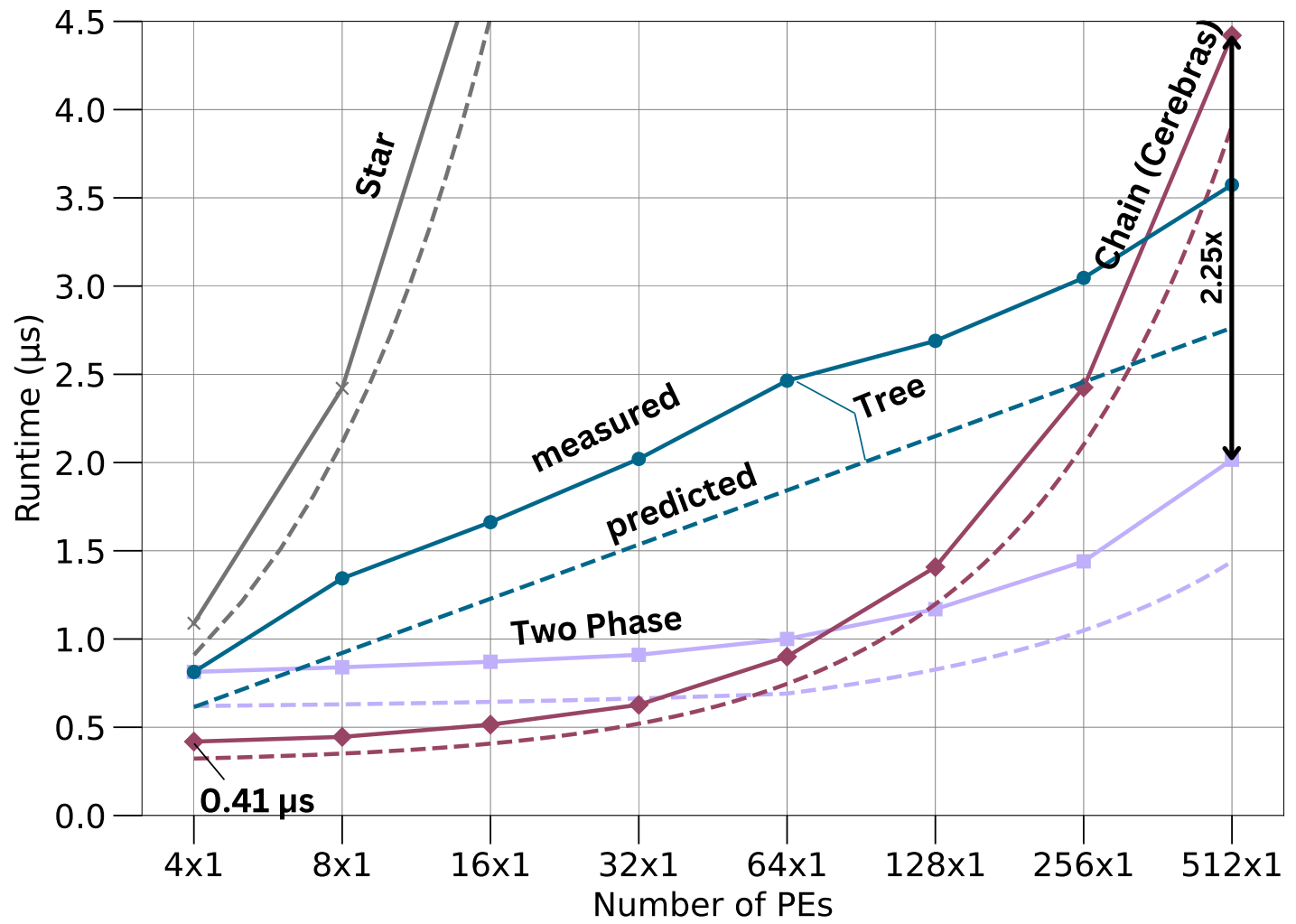
Distance 6	
Maximum 6	Total 10
Depth 4	

Modeling AI Accelerators – Spatial Model



Distance 6	
Maximum 6	Total 10
Depth 4	
Contention 2	

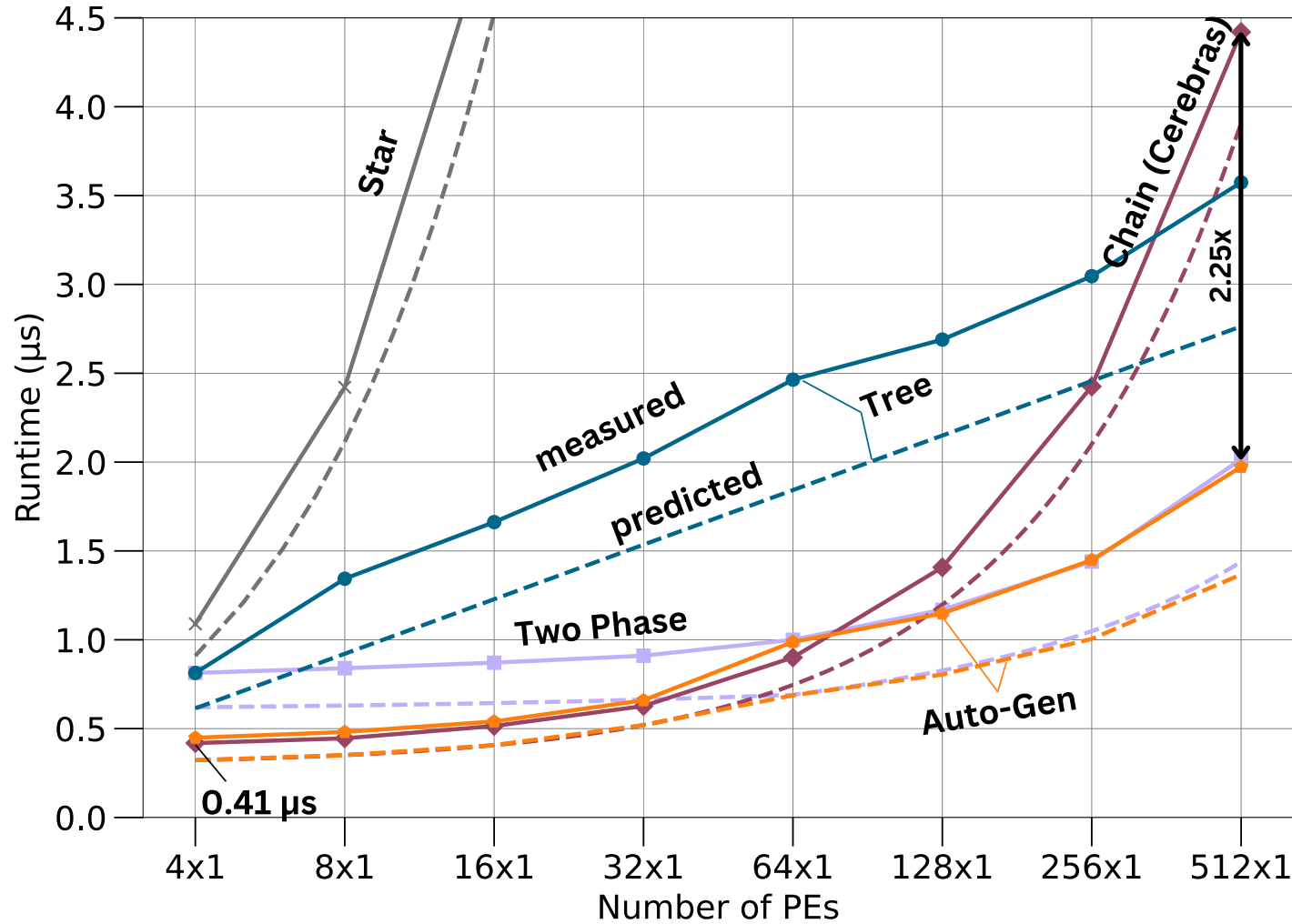
Communication Collectives on the CS-2



Near-Optimal Wafer-Scale Reduce
<https://arxiv.org/abs/2404.15888>
 to appear at HPDC 2024

Reduce 1 KB per PE

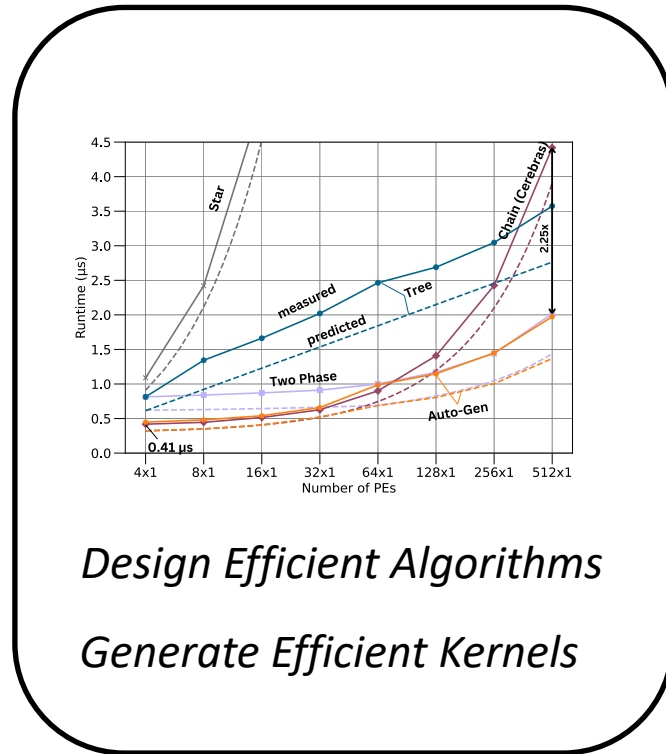
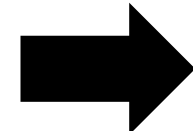
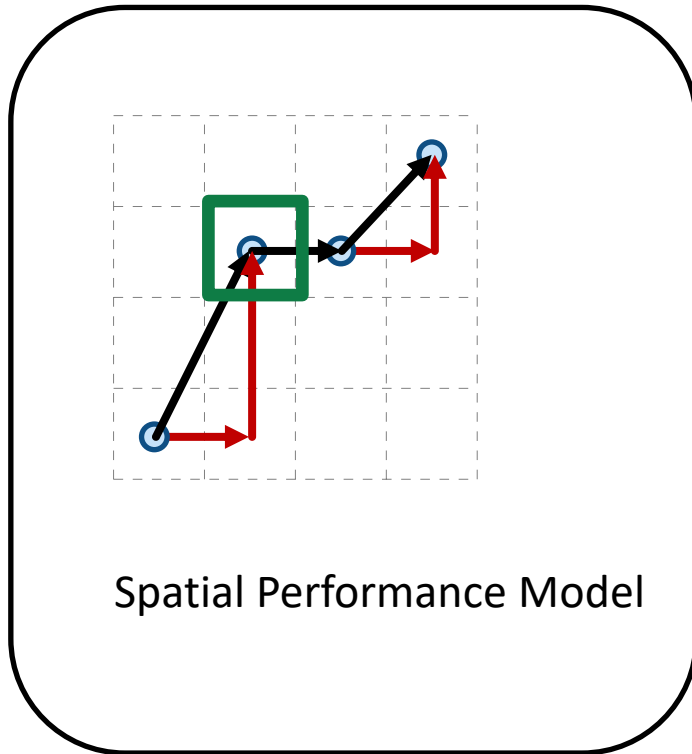
Communication Collectives on the CS-2






Near-Optimal Wafer-Scale Reduce
<https://arxiv.org/abs/2404.15888>
 to appear at HPDC 2024

Reduce 1 KB per PE

Conclusions



More of SPCL's research:

- 
youtube.com/@spcl **180+ Talks**
- 
twitter.com/spcl_eth **1.4K+ Followers**
- 
github.com/spcl **3.8K+ Stars**

... or spcl.ethz.ch



Near-Optimal Wafer-Scale Reduce
 Lucyznski & Gianinazzi et al.
<https://arxiv.org/abs/2404.15888>
 to appear at HPDC 2024

Democratizing AI Accelerators for HPC Applications: Challenges, Success, and Support

ISC 2024 | May 14, 2024 | Josef Weidendorfer

Collaboration between

BDAI Team (Big Data Artificial Intelligence)

Nicolay Hammer, Juan Durillo, Jophin John, Michael Hoffmann

FC Team (Future Computing)

Josef Weidendorfer, Amir Raoofy, Arjun Parab



- Part of German GCS (Gauss Centre for Supercomputing)
 - Compute proposals to be granted
 - German / European
- Academics in Munich area

~ 2.000
Researchers



100s of projects

- Astrophysics, Particle Physics
- Chemistry / Material Science
- Comp. Fluid Dynamics / Eng.
- Environmental / Life Sciences



LRZ user base is diverse

- **codes often developed by users**
- **porting/tuning is not contributing to science, so often not in focus**

SuperMUC-NG

Top500 - Nov 2018: #8
(Nov 2021: #23)

Lenovo Intel

311,040 cores

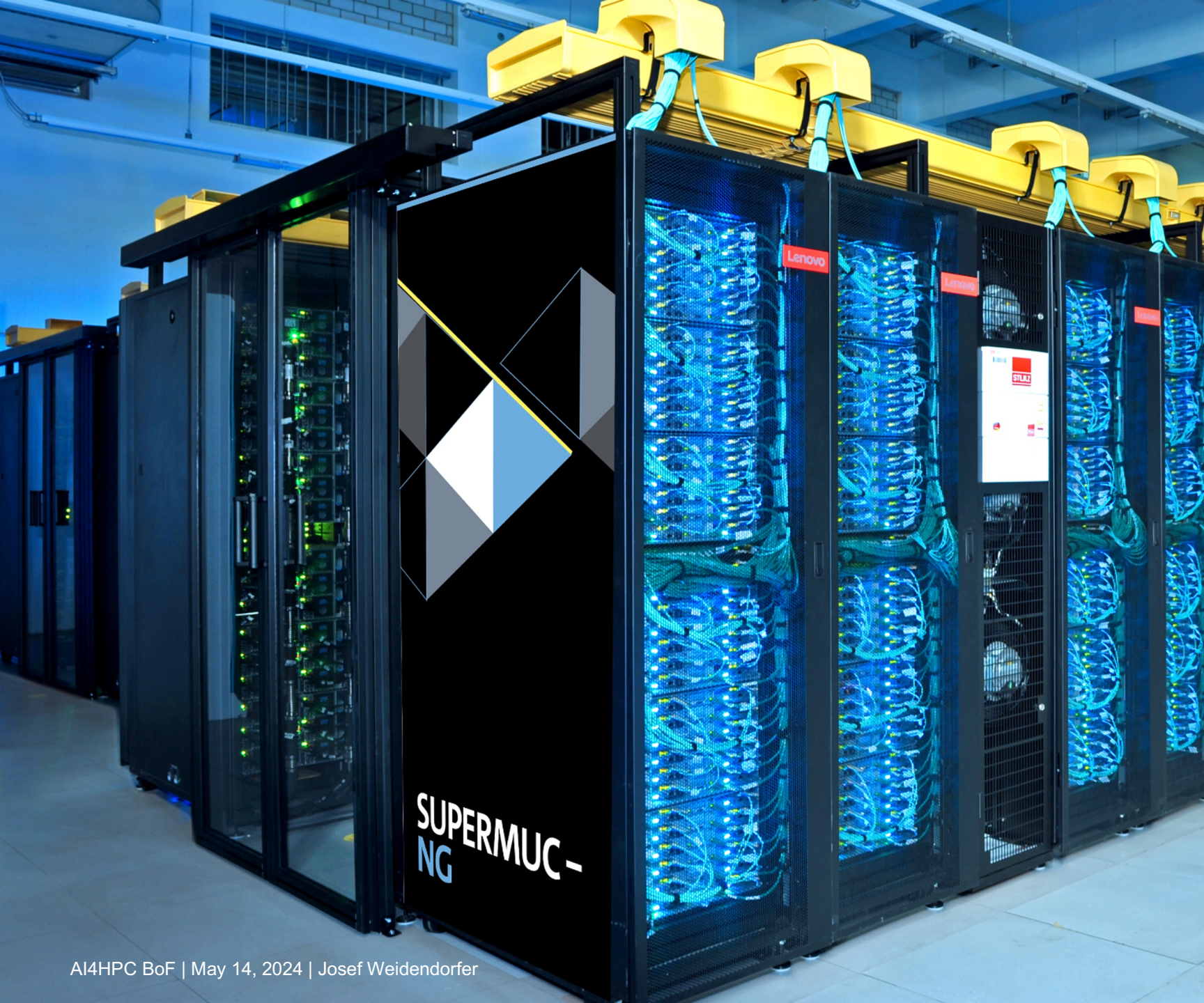
Intel Xeon Skylake

26.9 PetaFlops Peak

19.5 PetaFlops Linpack*

719 TeraByte Main Memory

70 PetaByte Disk



Accelerated node architecture

- 2x Intel® Xeon® Platinum 8480L, 56 cores
- 4x Intel® Data Center GPU Max 1550
- 512 GB DDR5 main memory
- Lenovo's SD650-I v3 platform



Distributed asynchronous object storage (DAOS)

- 1 PB capacity
- > 750 GB/s write bandwidth



High speed interconnect

- Mellanox HDR Infiniband
- fat tree topology
- two uplinks per node
- separated from Phase 1



Integration

- Phase 1 accounts and HOME directories
- Phase 1 WORK and SCRATCH filesystems
- DSS volumes available
- direct warm water cooling



Efforts on Future Technologies: BEAST Experimental Environment



BEAST (Bavarian Energy Architecture Software Testbed)

- Evaluate recent hardware technology options for HPC and AI
 - CPUs : x86 (Intel / AMD), ARM (Marvell, Fujitsu, Nvidia GH)
 - GPUs : Nvidia (A-100/H-100), AMD (MI-100/200), Intel (PVC)
 - Special Purpose Accelerators : NextSilicon DataFlow, Cerebras WSE-2
- Other research infrastructure (FPGAs, Smart Network, Quantum Computing)

BEAST (Bavarian Energy Architecture Software Testbed)

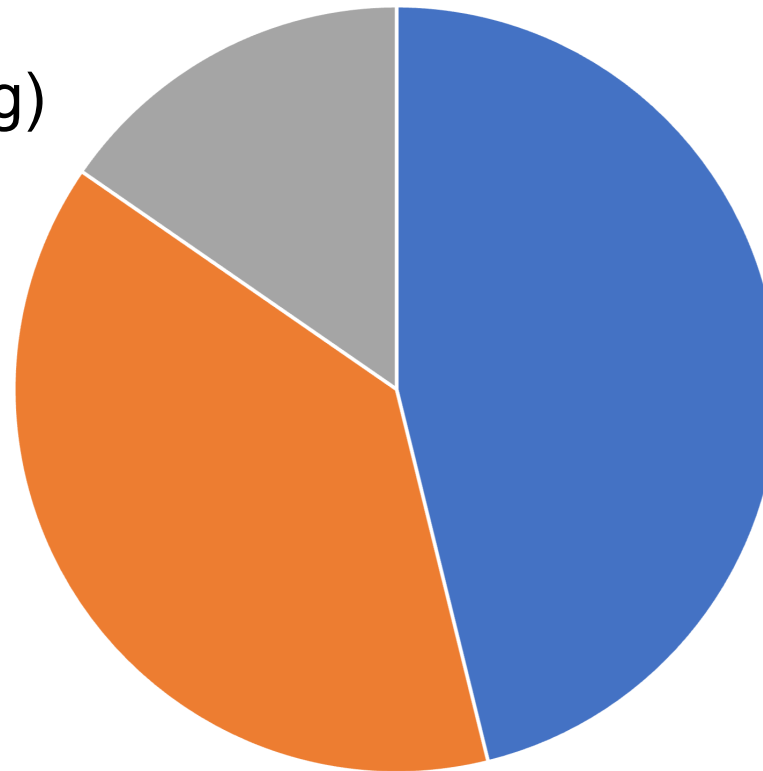
- Benchmarking of hardware located at LRZ allows to understand
 - Performance for different domains (vs. vendor claims)
 - Power/Energy efficiency measurements (using LRZ monitoring solution DCDB)
 - Stability of Software Stack
 - Experience from selected friendly users
 - Effectiveness of vendor-provided support on hardware failures

Evaluation of Cerebras WSE-2 as part of BEAST

In Contact with 13 “Friendly” User Groups

HPC
2 (CFD, Data Mining)

Computer Vision
5 (Medical, ...)



NLP
6 (LLMs, ...)

“Unconventional Usage“ of Accelerators

Can we support HPC via AI Accelerators?

Why does anybody want to use an exotic AI HW for HPC?

- Exceptional On-Chip Memory Capability
 - Size
 - Bandwidth
- Huge number of cores on a single chip promise low-latency synchronization

- Unique architecture of WSE-2
 - 850000 cores
 - 40GBs on-chip SRAM
 - 20 PB/s memory bandwidth
 - Very high NoC bandwidth

Case Study (In Cooperation with TUM)



Master Thesis

- “Implementation and Evaluation of Matrix Profile Algorithms on the Cerebras Wafer-Scale Engine”
- Time Series Mining (Similarity Indexing)

- Matrix profile computations are usually performed on CPUs and GPUs using SCAMP library which is known to be fast
- Can we port SCAMP to CS-2 WSE?

- Experiments were conducted on the CS-2 of the EIDF (EPCC)

Case Study - Preliminary Take Away Lessons



- Cerebras's SDK and Documentation have matured over the study's time allowing to successfully port the SCAMP Kernel to CS-2
 - Functional port successful
 - No optimizations yet done about
 - on-chip data pipelining
 - data transfers to device
 - Challenge to make best use of on-chip per-core memory (both needed by code and data)

Is Cerebras WSE Usable for HPC?



Usecase Study

- First efforts promising
- More research required
- Default software stack may be difficult for HPC usage

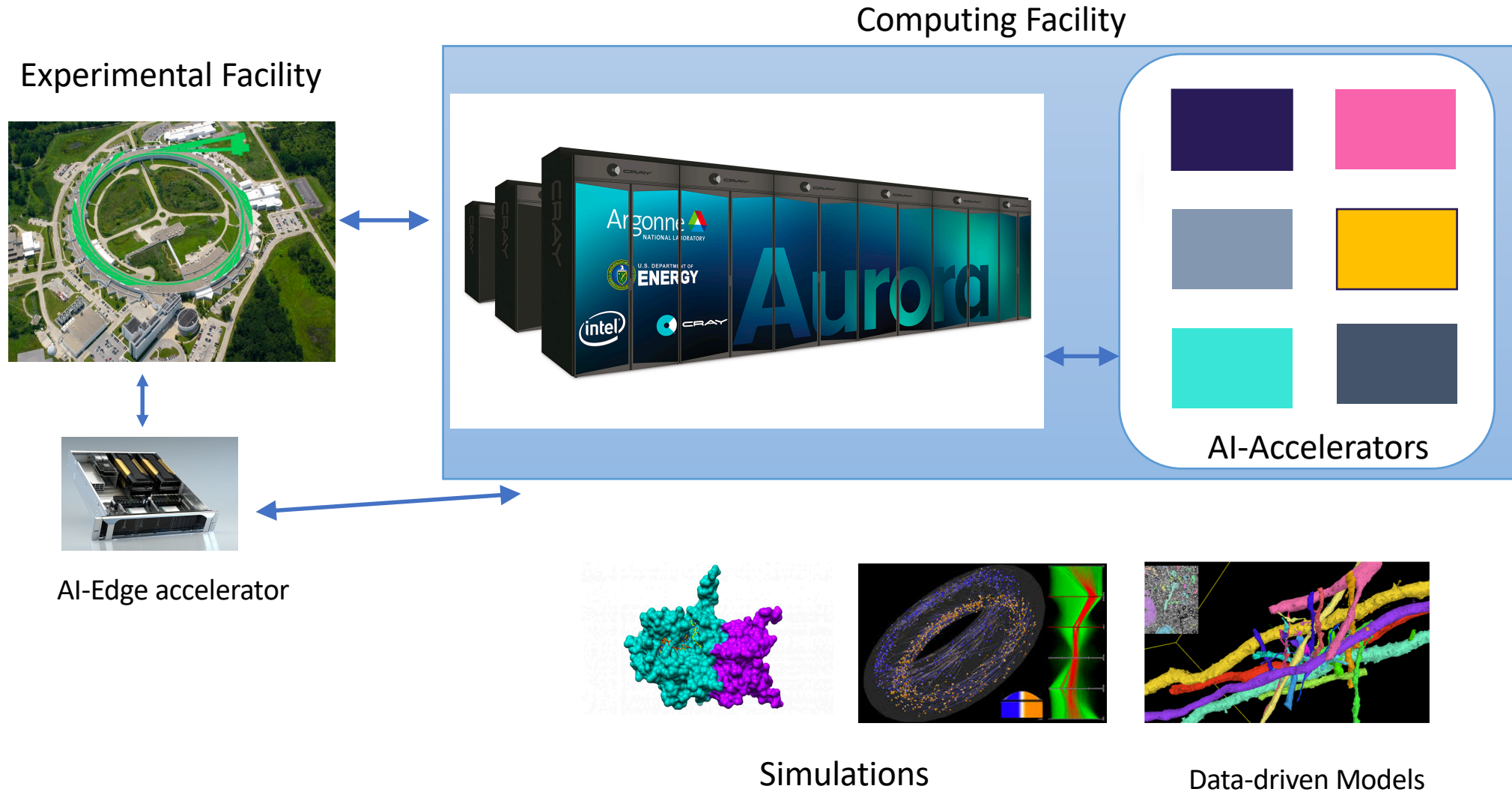
General Perspective

- Ongoing evolution (CS-2 to CS-3)
- Expanded library and customer support enhances usability
- Continuous feature additions to the SDK

Artificial Intelligence Testbeds at Argonne National Laboratory

Murali Emani
Argonne Leadership Computing Facility
memani@anl.gov

Integrating AI Systems in Facilities



ALCF AI Testbeds

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



Cerebras (CS-2)



SambaNova



Graphcore



Habana



Groq

- Infrastructure of next-generation machines with hardware accelerators customized for artificial intelligence (AI) applications.
- Provide a platform to evaluate usability and performance of various applications running on these accelerators.
- The goal is to better understand how to integrate AI accelerators with ALCF's existing and upcoming supercomputers to accelerate science insights

Recent ALCF AI Testbed Updates

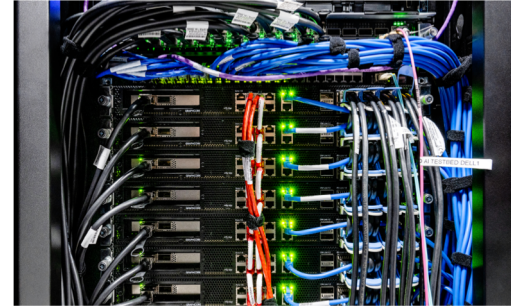
ALCF AI Testbed Systems are in production and available for allocations to the research community

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



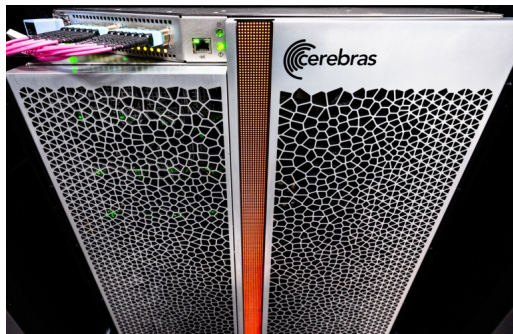
SambaNova upgraded to latest 2nd generation SN30 accelerators and scaled to **8 nodes with 64 AI accelerators (RDU)**

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



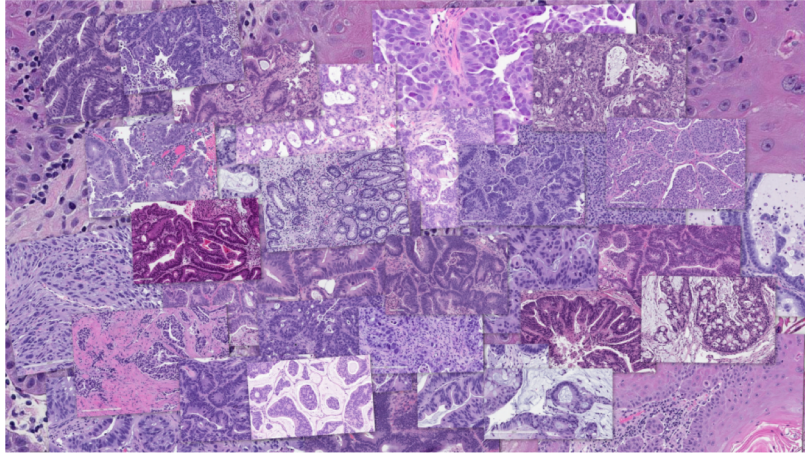
Groq system has been upgraded to a GroqRack with nine nodes, each consisting of eight GroqChip Tensor streaming processors, **72 accelerators**

GroqRack

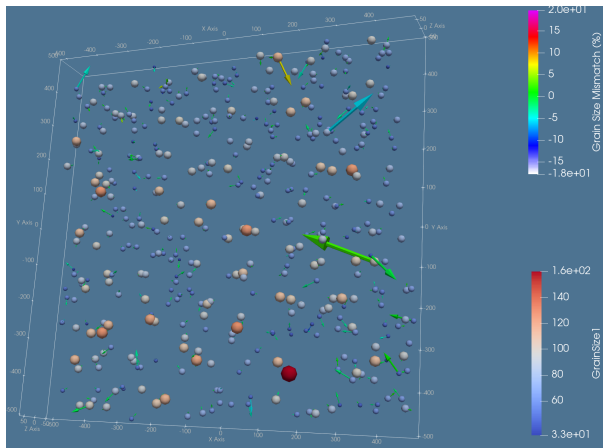
<https://nairrpilot.org>

	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™	IPU Link	Ethernet-based	NVLink

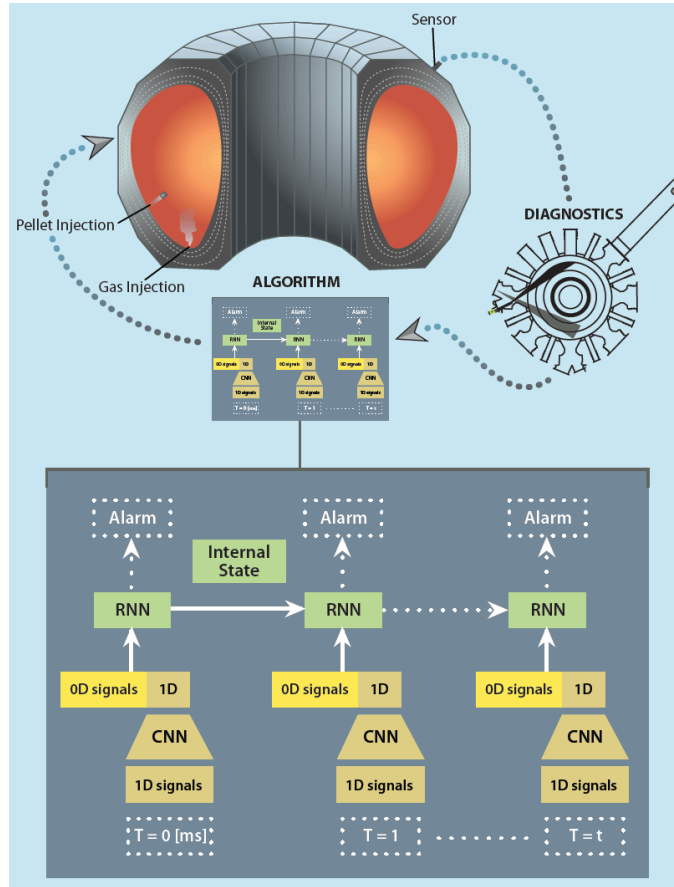
AI FOR SCIENCE AND HPC APPLICATIONS ON AI TESTBED



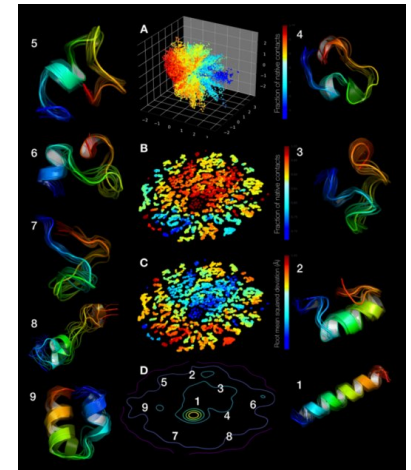
Cancer drug response prediction
(Credit: Candle)



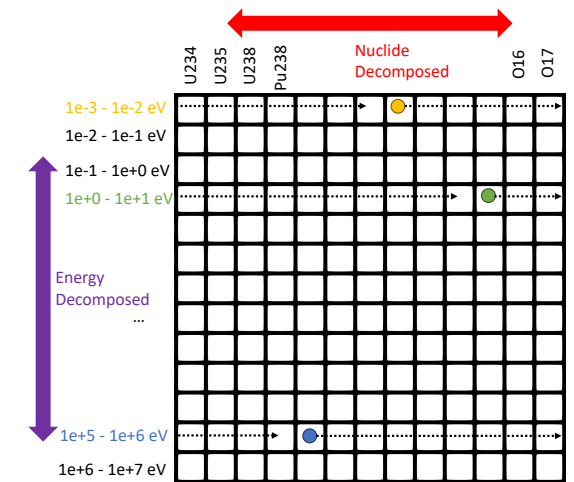
Imaging Sciences-Braggs Peak
(Credit: Z. Liu)



Tokamak Fusion Reactor operations
(Credit: K. Felker)



Protein-folding (Image: NCI)



Monte Carlo Particle Transport for
Reactor Simulation (Credit: J. Tramm)

and more..

A Traditional HPC Simulation Kernel on an AI Accelerator

Scientific Achievement

The Cerebras WSE-2 is a wafer-scale AI accelerator. Despite not being designed for traditional HPC workloads, we were able to develop new algorithms and performance optimization strategies to allow for a key Monte Carlo particle transport simulation kernel to execute with high efficiency on the device. Significant speed and power advantages compared to GPU were found.

Significance and Impact

- **Developed mini-app** representing key cross section lookup kernel from the Monte Carlo (MC) particle transport algorithm **for the Cerebras CSL SDK**
- Compared results against highly optimized CPU and GPU implementations, and found that the **WSE-2 was >100,000x faster than serial CPU execution**, and **182x faster than A100 GPU execution**
- Results suggest full MC particle transport app on WSE-2 will be possible

Technical Approach

- Leveraged vast quantities (>40GB) of single-cycle latency SRAM on WSE-2
- Developed new hyper domain decomposition techniques to spread simulation data across the >700k processing elements of the WSE-2, and optimized movement of particles through the WSE-2.

PI(s)/Facility Lead(s): John Tramm
Collaborating Institutions: ANL, UChicago
ASCR Program: ANL LDRD Expedition
ASCR PM: N/A
Publication(s) for this work: Tramm, et al., "Efficient algorithms for Monte Carlo particle transport on AI accelerator hardware," *Computer Physics Communications*, Volume 298 (2024).
doi:10.1016/j.cpc.2023.109072.

Architecture	Monte Carlo Performance FOM Lookups/sec
Cerebras WSE-2	1.17E+10
CPU (single 8180M Xeon Core)	1.15E+05
GPU (single NVIDIA A100 GPU)	6.43E+07

WSE-2
>100,000x
speedup over
serial CPU

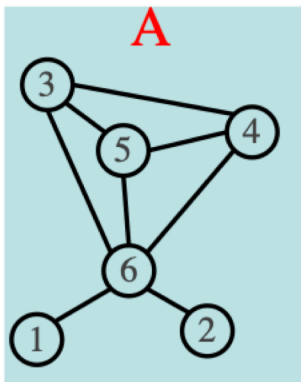
WSE-2 182x
speedup over
A100 GPU

This table shows the performance figure of merit for our mini-app (based on XSbench). The CPU version is written in C. The GPU version is written in optimized CUDA using architecture-specific optimization strategies. The Cerebras version was written using the CSL Cerebras SDK. The figure of merit represents the number of macroscopic cross sections per second (higher is better) in a typical depleted fuel reactor simulation problem with hundreds of nuclides.

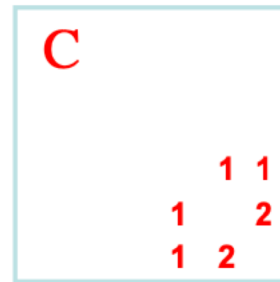
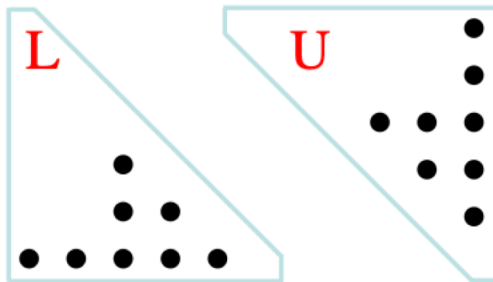
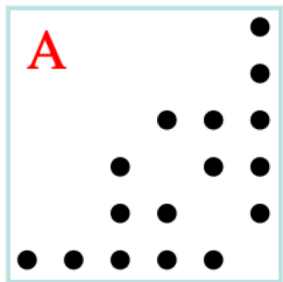
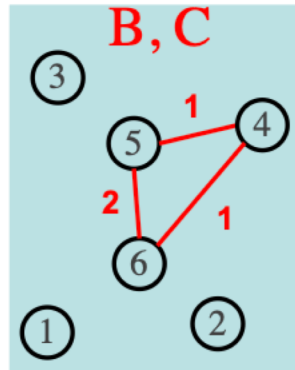
Courtesy: John Tramm, ANL

Linear Algebra-based Triangle Counting on Graphcore's IPU Architecture

Given $G(V, E)$ where V is the vertex set and E is the edge set, count the number of triangles in G . A triangle is a triplet $\langle u, v, w \rangle$ such that $u, v, w \in V$, and $uv, vw, uw \in E$.



$$\begin{aligned}
 A &= L + U && (\text{hi} \rightarrow \text{lo} + \text{lo} \rightarrow \text{hi}) \\
 L \times U &= B && (\text{wedge, low hinge}) \\
 A \wedge B &= C && (\text{closed wedge}) \\
 \text{sum}(C)/2 &= \mathbf{4 \text{ triangles}}
 \end{aligned}$$



Key Steps:

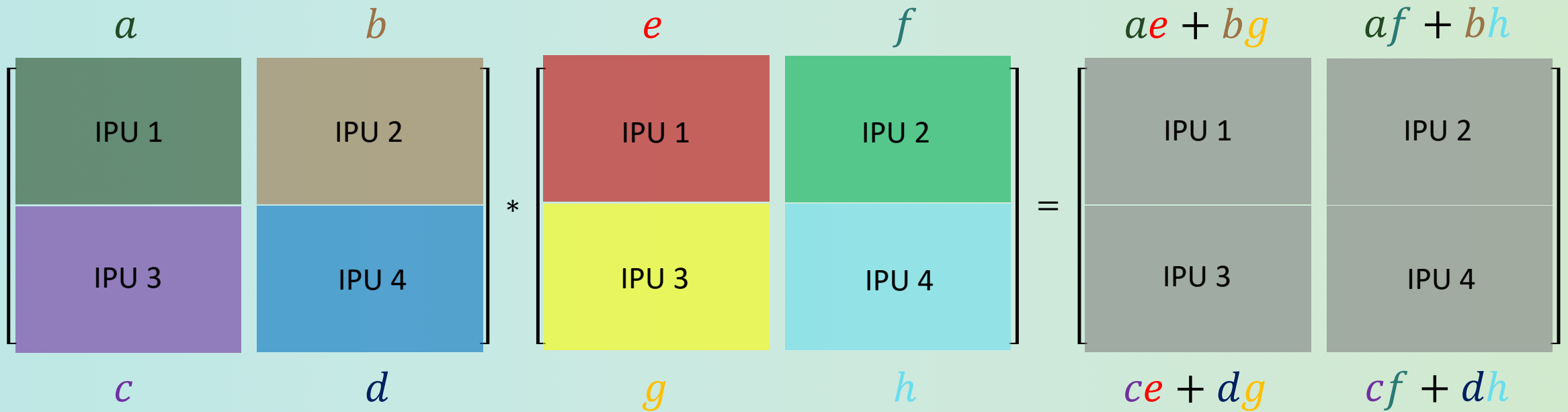
- LU-decompose Adjacency matrix
- Matrix multiply L, U
- Elementwise multiply
- Reduce

[Characterizing the Performance of Triangle Counting on Graphcore's IPU Architecture](#)
 Reet Barik, Siddhisanket Raskar, Murali Emani, Venkatram Vishwanath

Azad, B., Gilbert. "Parallel triangle counting and enumeration using matrix algebra". *IPDPSW, 2015*



Triangle Counting on IPU: Mapping to architecture

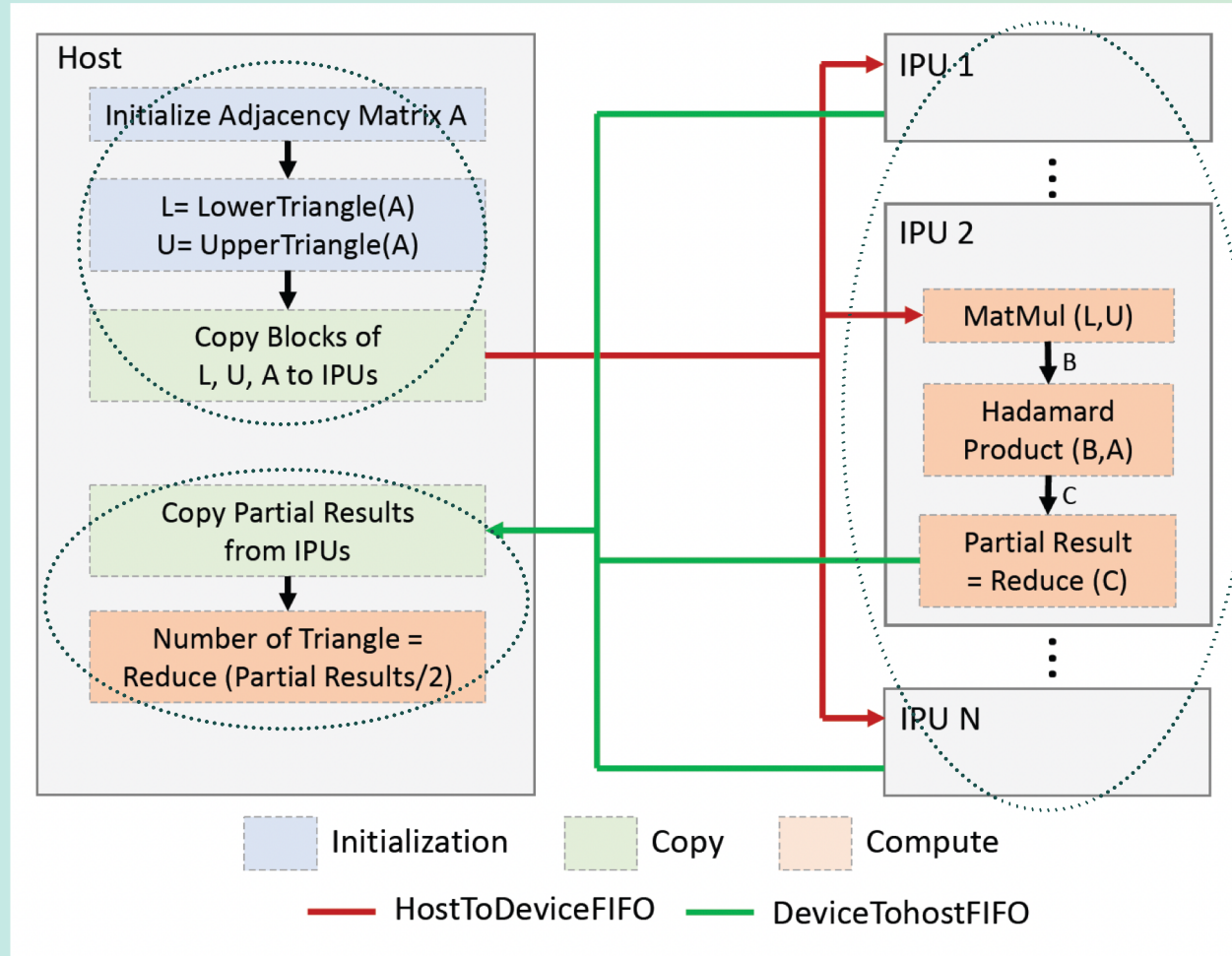


Block decomposition of input matrix

Triangle Counting on IPU: Mapping to architecture

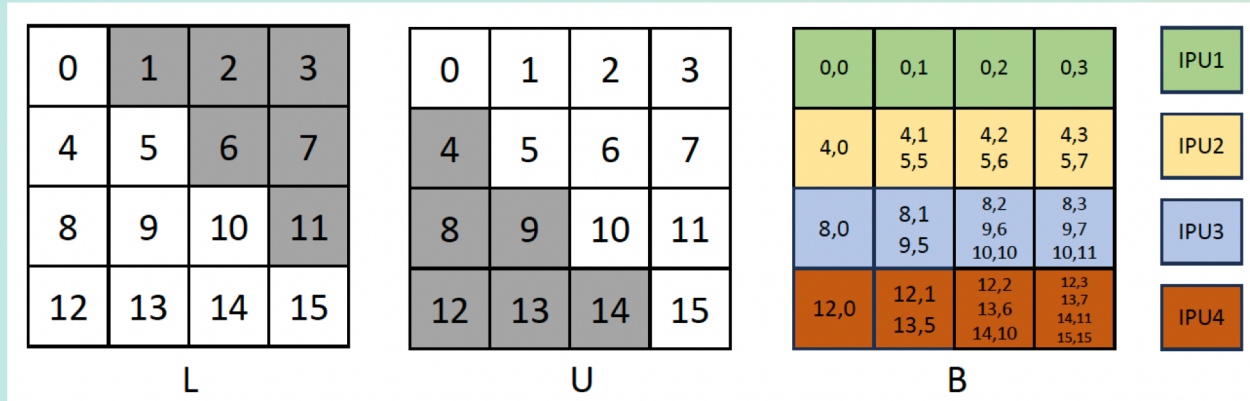
Init and copy
From Host

Copy partial
results to Host
and reduce



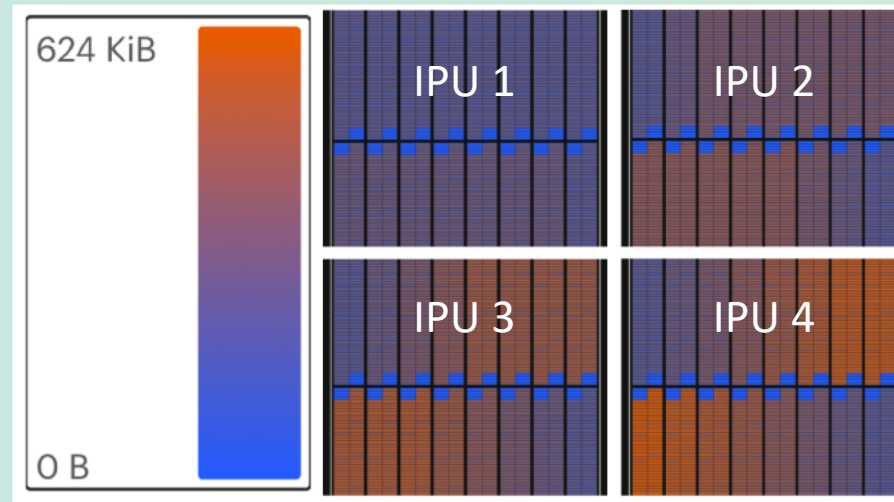
Compute
On Device

Triangle Counting on IPU: Optimization

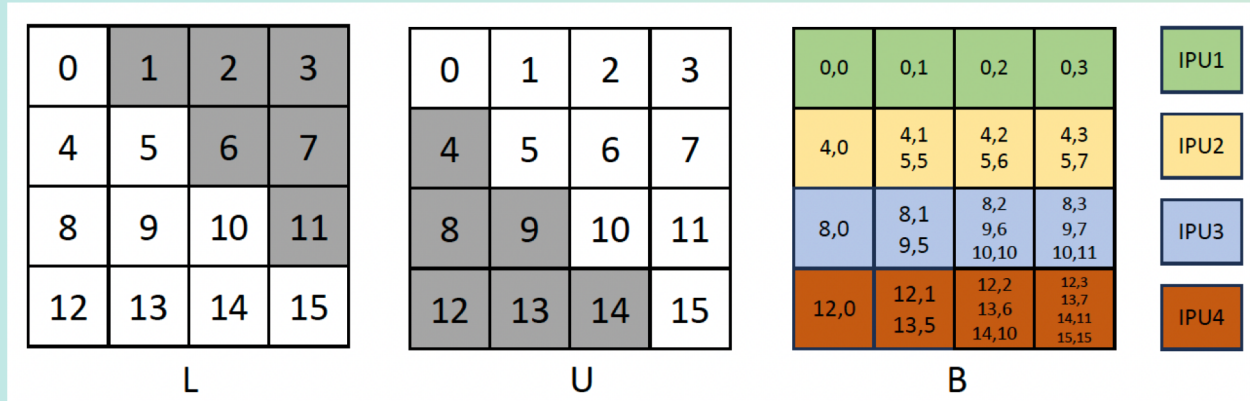


Computation workload pattern

Load imbalance

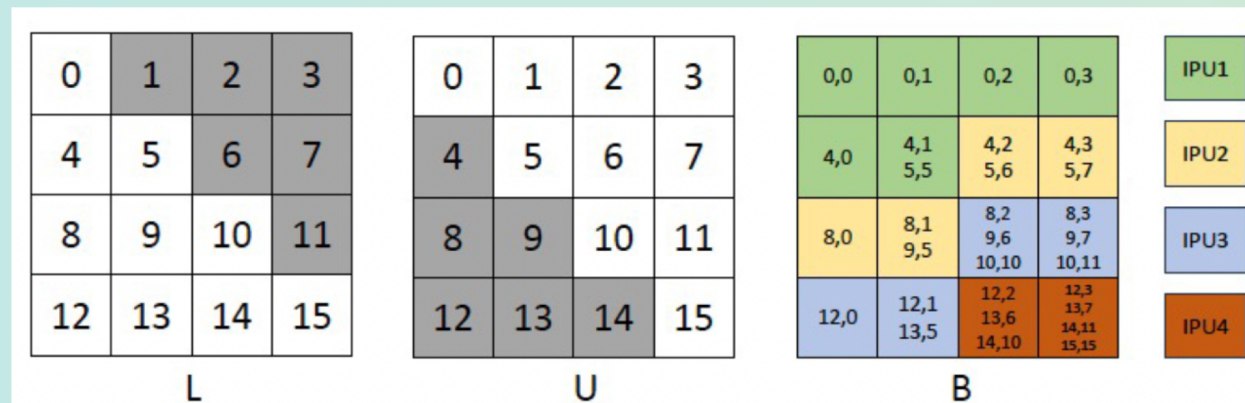


Triangle Counting on IPU: Optimization



Computation workload pattern

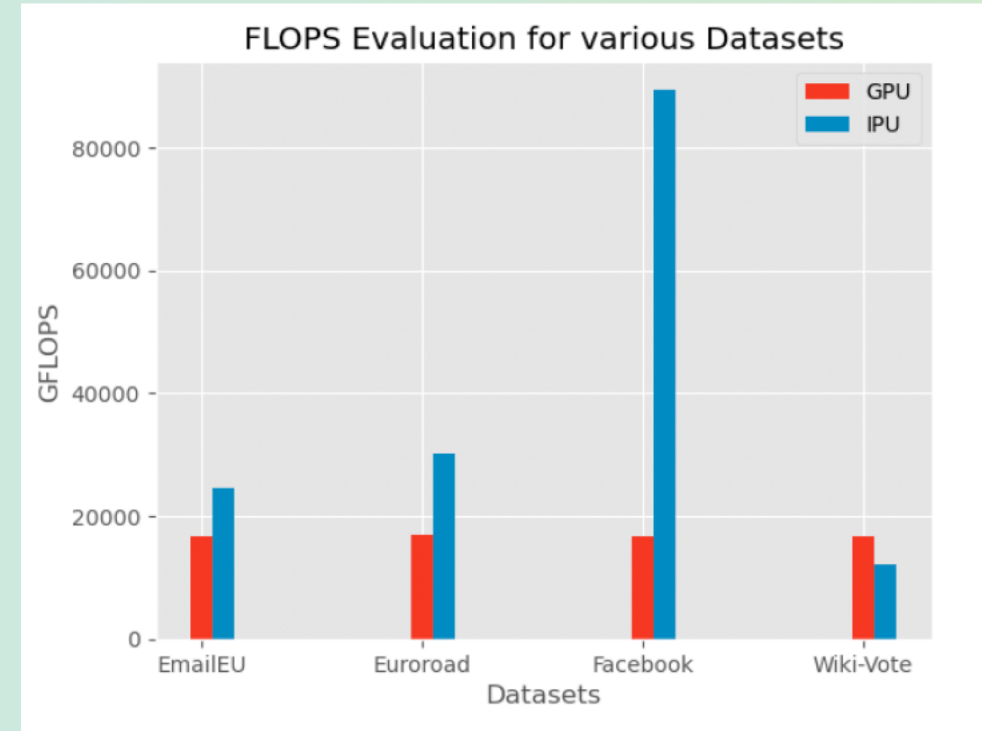
Weighted mapping of workload to IPUs



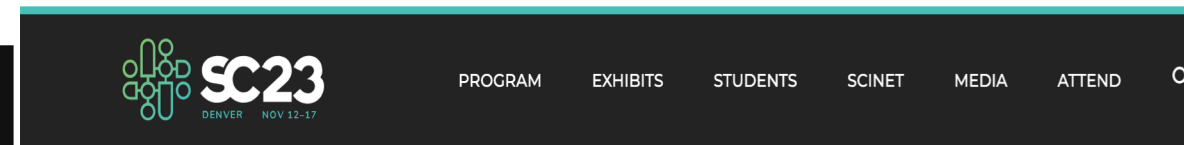
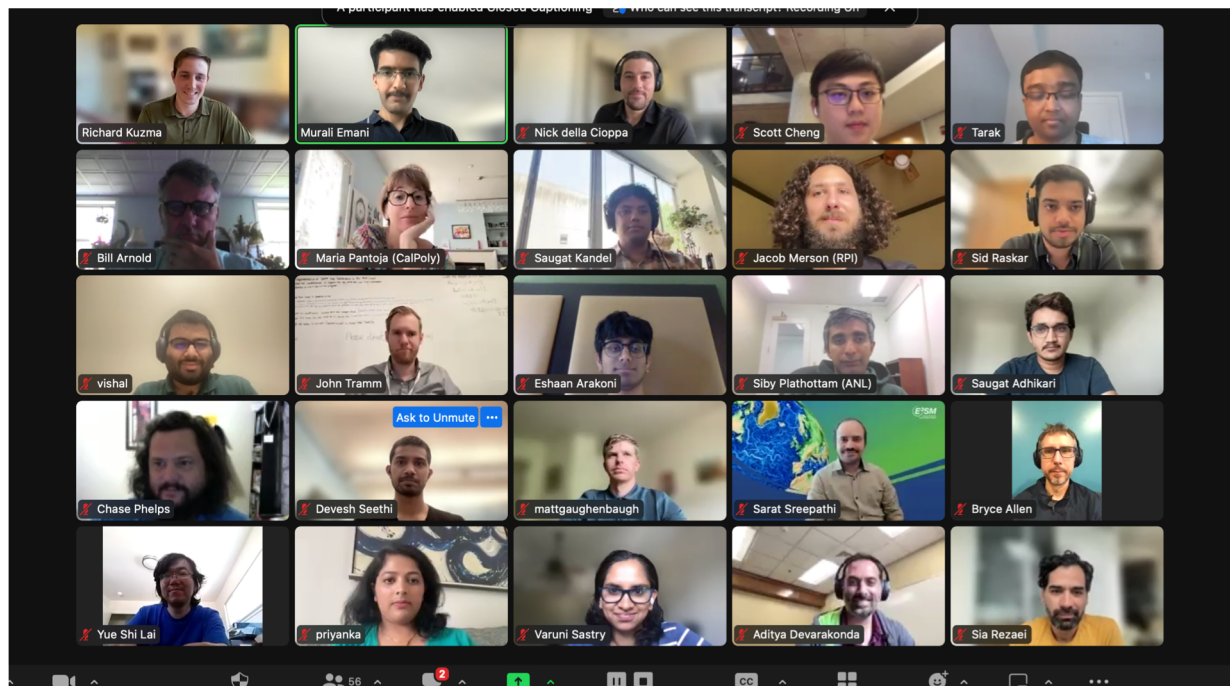
Triangle Counting on IPU: Experiments

Input	V	E	Max Deg	Avg. Deg
Kronecker (2^8)	256	2155	163	16.8
Kronecker (2^9)	512	4752	274	18.6
Kronecker (2^{10})	1024	10496	471	20.5
Kronecker (2^{11})	2048	22709	747	22.2
Kronecker (2^{12})	4096	48386	1316	23.6
Kronecker (2^{13})	8192	102124	2250	24.9
EmailEU	1,005	25,571	345	31.9
Euroroad	1,174	1,417	10	2.4
Facebook	4,039	88,234	1045	43.7
Wiki-Vote	7,115	103,689	1065	28.3

Table 1: Dataset Characteristics



AI Testbed Community Engagement



Home > Presentation

Presentation

Programming Novel AI Accelerators for Scientific Computing

Scientific applications are increasingly adopting Artificial Intelligence (AI) techniques to advance science. There are specialized hardware accelerators designed and built to run AI applications efficiently. With a wide diversity in the hardware architectures and software stacks of these systems, it is challenging to understand the differences between these accelerators, their capabilities, programming approaches, and how they perform, particularly for scientific applications. In this tutorial, we will cover an overview of the AI accelerators landscape with a focus on SambaNova, Cerebras, Graphcore, Groq, and Habana systems along with architectural features and details of their software stacks. We will have hands-on exercises that will help attendees understand how to program these systems by learning how to refactor codes written in standard AI framework implementations and compile and run the models on these systems. The tutorial will enable the attendees with an understanding of the key capabilities of emerging AI accelerators and their performance implications for scientific applications.

Tutorial

Sunday, 12 November 2023
8:30am - 12pm MST

Location: 203

NEXT PRESENTATION > STARTS IN 106:07:40

Energy-Efficient GPU Computing

- 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/>
 - Groq: <https://events.cels.anl.gov/event/448/>

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

Observations, Challenges and Insights

- Significant speedup achieved for a wide-gamut of scientific ML applications
- Early adoption for HPC kernels show promising results
- Recent work on using OpenMP to offload kernels to Graphcore IPU
- Room for improvement exists
 - Porting efforts and compilation times, custom libraries
 - support for performance analysis tools, debuggers
- Limited capability to support low-level HPC kernels
 - Work in progress to improve coverage

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.

* Finalist 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



THE UNIVERSITY *of* EDINBURGH



AI accelerators for HPC

Joseph Lee

EPCC

j.lee@epcc.ed.ac.uk

Who are we?

- Supercomputing centre at the University of Edinburgh, UK
- Providing world-class computing systems, data storage and support services
- Hosts the UK's national supercomputing service: ARCHER2



- ~150 staff
- Research, Education and Training: MSc and PhD programme

Who are we?

- Edinburgh International Data Facility (EIDF)
- High-powered data analytics and storage service that supports research and data-driven innovation
 - CPU & GPU
 - Cerebras CS-2 (x2)
 - Graphcore Bow Pod 64
- We have also been using these machines for HPC research



What are we looking at?

- ExCALIBUR (UK research programme for exascale algorithms and infrastructure)
- Enabling Hardware & Software Programme
 - Accelerators: RISC-V, FPGAs, Cerebras WSE



TensorFlow as a DSL for stencil-based computation on the Cerebras Wafer Scale Engine

Nick Brown¹, Brandon Echols², Justs Zarins¹, and Tobias Grosser³

- xDSL (Python x MLIR) for stencil computation
- PETSc on Cerebras
 - And more we want to do!
- Training: SDK tutorial at HPC Days – Durham UK



Experience

Positive 👍	Challenges 😞
Helpful support	Programming model/language (DSL)
Active community	Training & documentation
Tools (e.g. simulator)	Closed tools & reproducibility
Performance!	Big software changes
	Repeated/similar work (more collaboration?)



What are we looking for?

- Usable and accessible: Language, Library, Compiler
- Documentation and support
- AI Accelerator for HPC tutorials?
- Community!

Q&A and Discussion

Q&A and Discussion

- Useful features
- Challenges
- Interface/language
- Compiler and programming model
- Libraries
- Applications
- Architecture
- Additional support
- Community & knowledge exchange

Wrapping up

- Feedback via swapcard app
- Get in touch if
 - Want to get involved for BoF at SC
 - Comments, ideas, and feedback!
- j.lee@epcc.ed.ac.uk



feedback form



Get in touch!