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 mm^2 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

Cerebras Wafer-Scale Engine (WSE-3)

The Largest Chip in the World

900,000 cores optimized for sparse linear algebra

46,225 mm^2 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

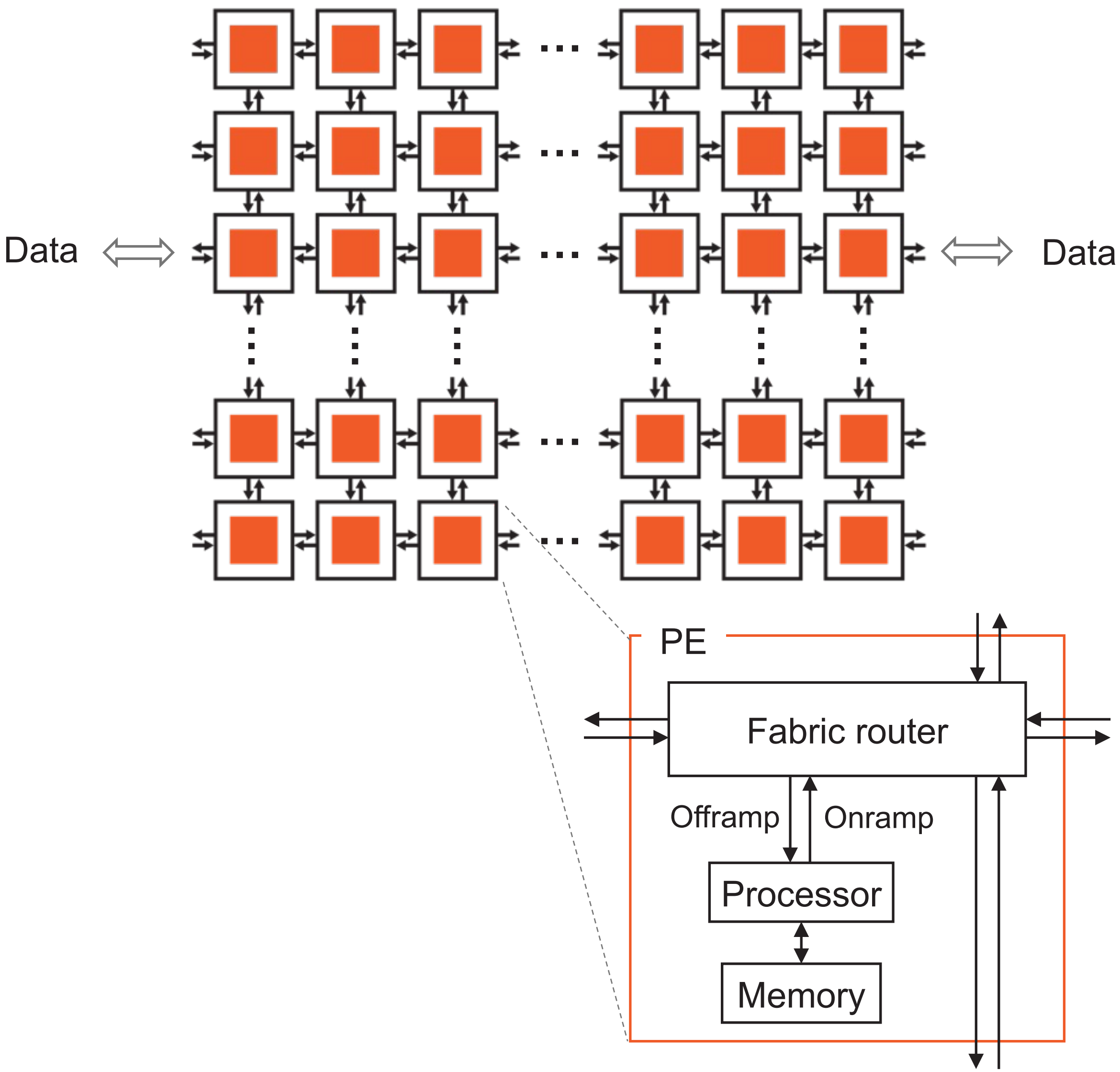
Cluster-scale acceleration on a single chip

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

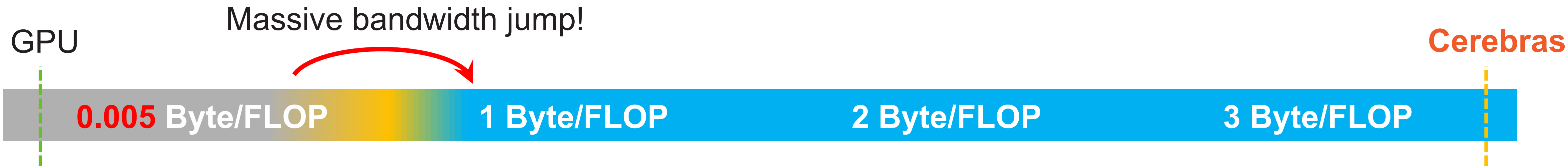
Flexible communication

- 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



Matrix-Matrix

GEMM

$$C \leftarrow \alpha AB + \beta C$$

$$\boxed{C} \mathrel{+}= \boxed{A}\ \boxed{B}$$

Matrix-Vector

GEMV

$$y \leftarrow \alpha Ax + \beta y$$

$$\boxed{y} \mathrel{+}= \boxed{A}\ \boxed{x}$$

Vector-Vector

DOT

$$\alpha \leftarrow (x, y)$$

$$\boxed{\alpha} \mathrel{+}= \boxed{y}\boxed{x}$$

Vector-Scalar

AXPY

$$y \leftarrow \alpha x + y$$

$$\boxed{y} \mathrel{+}= \boxed{\alpha}\boxed{x}$$

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

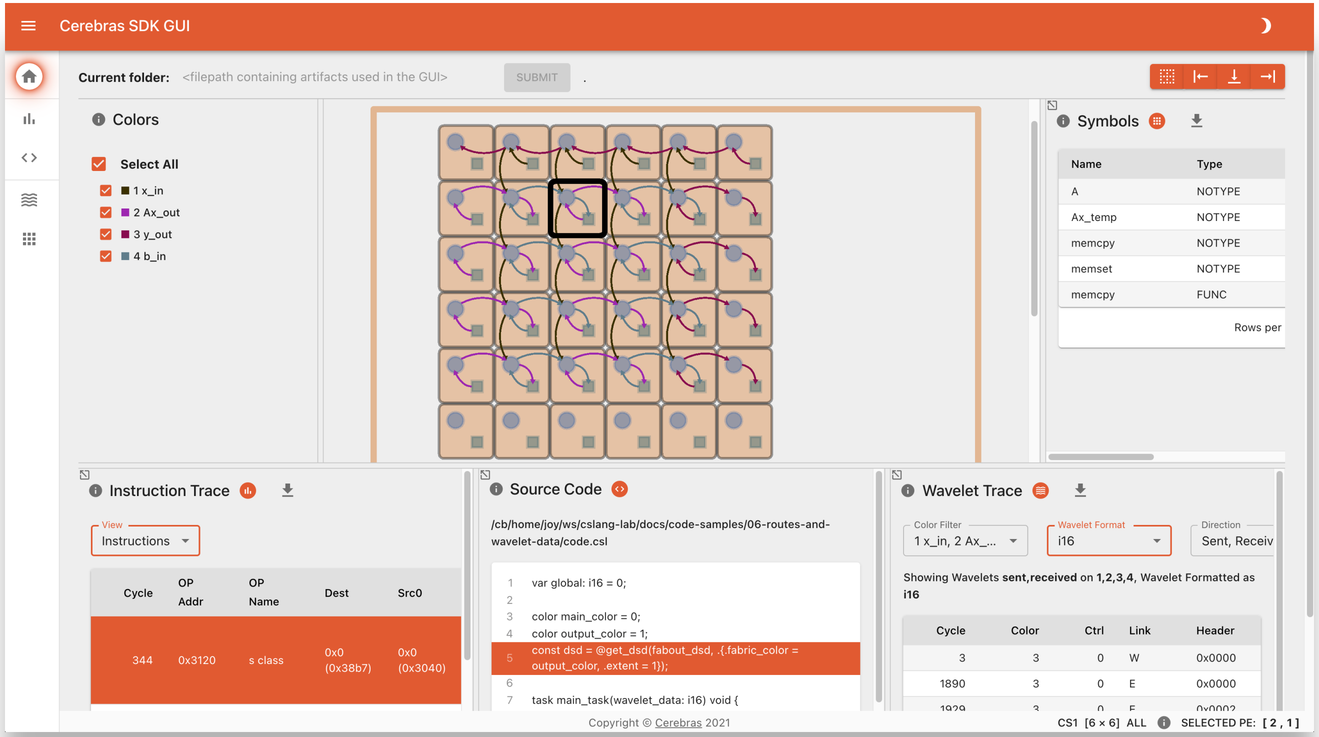
For HPC Users, Cerebras SDK provides **flexible**, **lower-level** programmability and access to HW performance features.

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



Cerebras SDK GUI

Current folder: <filepath containing artifacts used in the GUI> SUBMIT .

Colors

☑ **Select All**

☑ ■ 1 x_in

☑ ■ 2 Ax_out

☑ ■ 3 y_out

☑ ■ 4 b_in

Symbols

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

Rows per

Instruction Trace

View: Instructions

Cycle	OP Addr	OP Name	Dest	Src0
344	0x3120	s class	0x0 (0x38b7)	0x0 (0x3040)

Source Code

/cb/home/joy/ws/cslang-lab/docs/code-samples/06-routes-and-wavelet-data/code.csl

```
var global: i16 = 0;

color main_color = 0;
color output_color = 1;
const dsd = @get_dsd(fabout_dsd, .{.fabric_color =
output_color, .extent = 1});

task main_task(wavelet_data: i16) void {
```

Wavelet Trace

Color Filter: 1 x_in, 2 Ax_...

Wavelet Format: i16

Direction: Sent, Receiv

Showing Wavelets **sent,received** on **1,2,3,4**, Wavelet Formatted as **i16**

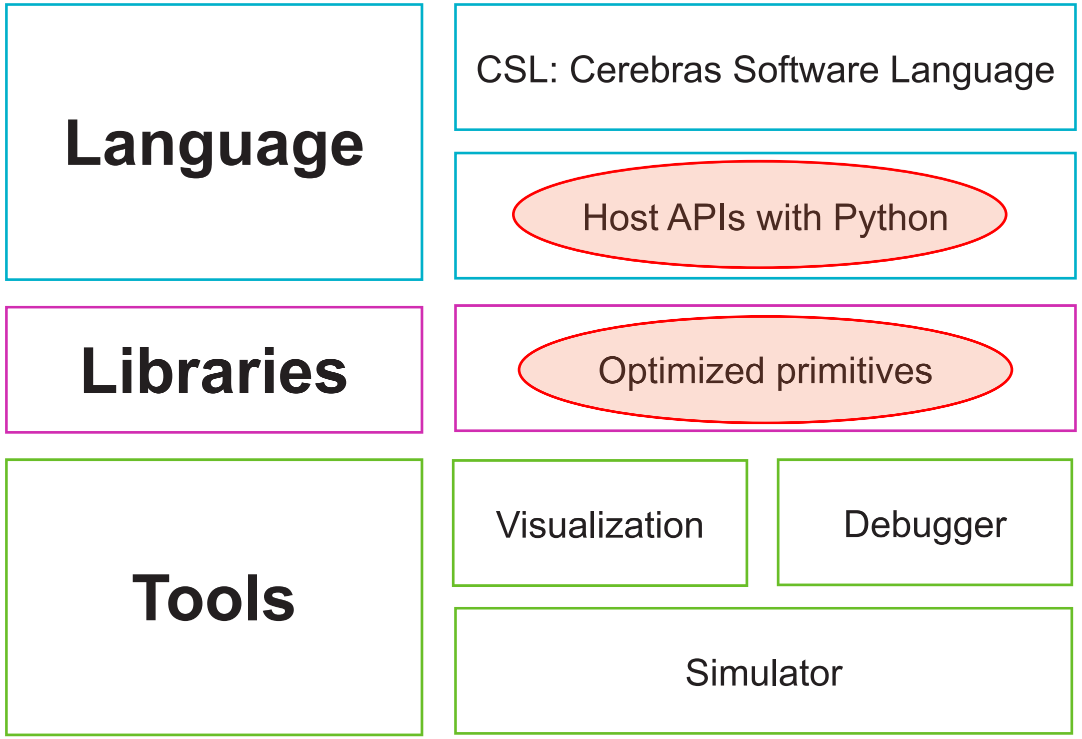
Cycle	Color	Ctrl	Link	Header
3	3	0	W	0x0000
1890	3	0	E	0x0000
1929	3	0	E	0x0002

Copyright © Cerebras 2021

CS1 [6 × 6] ALL SELECTED PE: [2 , 1]

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

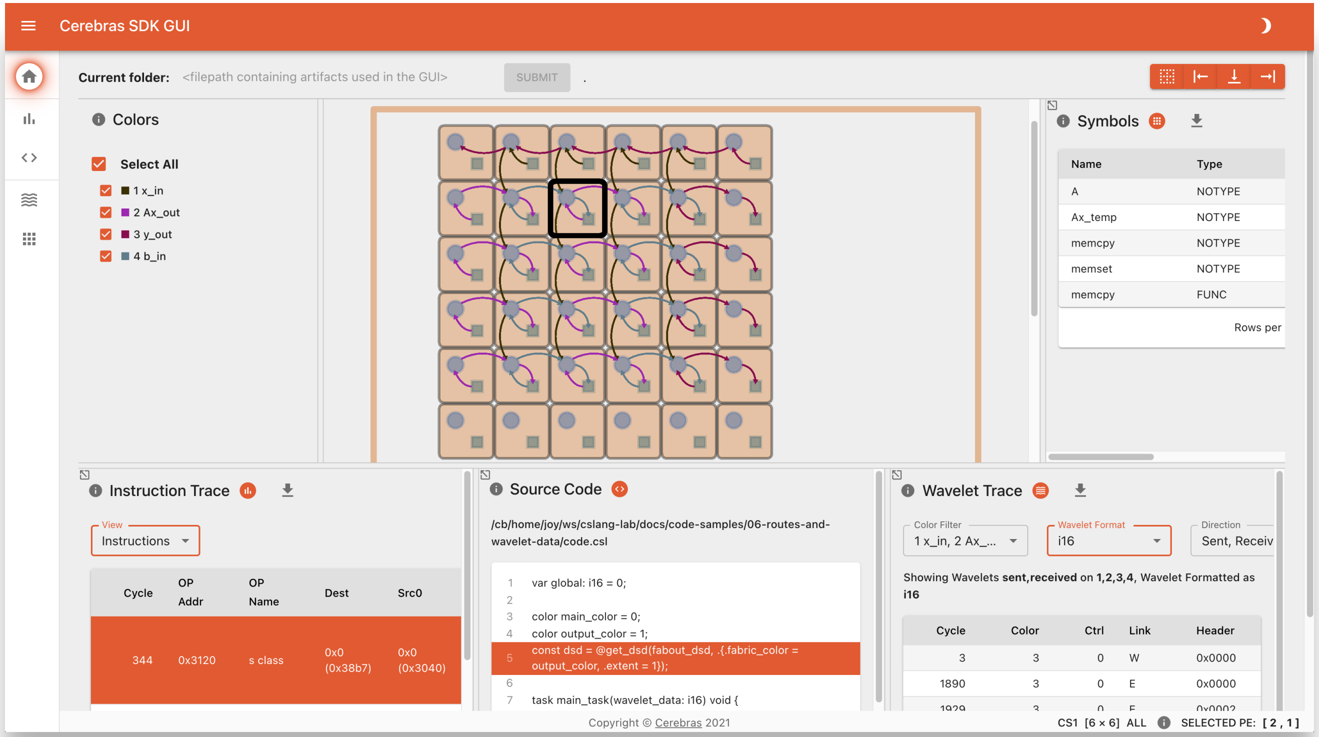
Optimized primitives

Tools

Visualization

Debugger

Simulator



Current folder: <filepath containing artifacts used in the GUI> SUBMIT .

☑ **Select All**

☑ 1 x_in

☑ 2 Ax_out

☑ 3 y_out

☑ 4 b_in

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

Rows per

View: Instructions

Cycle	OP Addr	OP Name	Dest	Src0
344	0x3120	s class	0x0 (0x38b7)	0x0 (0x3040)

/cb/home/joy/ws/cslang-lab/docs/code-samples/06-routes-and-wavelet-data/code.csl

```
var global: i16 = 0;

color main_color = 0;
color output_color = 1;
const dsd = @get_dsd(fabout_dsd, .{.fabric_color =
output_color, .extent = 1});

task main_task(wavelet_data: i16) void {
```

Color Filter: 1 x_in, 2 Ax_...

Wavelet Format: i16

Direction: Sent, Receiv

Showing Wavelets **sent,received** on **1,2,3,4**, Wavelet Formatted as **i16**

Cycle	Color	Ctrl	Link	Header
3	3	0	W	0x0000
1890	3	0	E	0x0000
1929	3	0	E	0x0002

Copyright © Cerebras 2021

CS1 [6 × 6] ALL SELECTED PE: [2 , 1]

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

ering

ence

.sa

12 pages.

ELL

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

Kazutomo Yoshii* Andrew Siegel* Leighton Wilson‡

p learning has highly innova- chitecture, the ures 40 GB of

portance 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 [9]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SPCL

Communication Collectives for the Cerebras Wafer-Scale Engine

Bachelor Thesis

Piotr Luczynski
`pluczynski@ethz.ch`

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

ectives are a (HPC) appli- n of Reduce SE). This ar- performance ational prob- estimate the e our predic- n addition to several new Moreover, we peration on

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

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. Settgast
Lawrence Livermore National Laboratory
Livermore, California, USA

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

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 Yoshii[1], Andrew Siegel[1]

[1]Argonne National Laboratory, Lemont, IL; [2]University of Chicago, Chicago, IL

by ANS]

atency memory, making it an n methods. Recent work has ance gains for the continuous nsport method. In the present od based off of the fractional

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

Anonymous Author(s)

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

Trackable Agent-based Evolution Models at Wafer Scale

Matthew Andres Moreno[1,2,3,*], Connor Yang[4], Emily Dolson[5,6], and Luis Zaman[1,2]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, United States
[2]Center for the Study of Complex Systems, University of Michigan, Ann Arbor, United States
[3]Michigan Institute for Data Science, University of Michigan, Ann Arbor, United States
[4]Undergraduate Research Opportunities Program, University of Michigan, Ann Arbor, United States
[5]Department of Computer Science and Engineering, Michigan State University, East Lansing, United States
[6]Program 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

Near-optimal Reduce on the Cerebras Wafer Scale Engine

Piotr Luczynski
Lukas Gianinazzi (advisor)
Patrick Iff (advisor)
Daniele De Sensi (advisor)
Leighton Wilson (advisor)
Torsten Hoefler (advisor)

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

erouiche
tud.ntnu.no
rondheim
way

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

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

Torsten Hoefler
torsten.hoefler@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

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‡

p learning has highly innova- chitecture, the ures 40 GB of

portance 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 [9]

12 pages.

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

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

ectives are a (HPC) appli- n of Reduce SE). This ar- performance ational prob- estimate the e our predic- n addition to several new Moreover, we operation on

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 ext vector sizes. This means they are

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

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. Settgast
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 Yoshii[1], Andrew Siegel[1]

Chicago, Chicago, IL

by ANS]

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

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

Anonymous Author(s)

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

Trackable Agent-based Evolution Models at Wafer Scale

Matthew Andres Moreno[1,2,3,*], Connor Yang[4], Emily Dolson[5,6], and Luis Zaman[1,2]
[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, United States
[2]Center for the Study of Complex Systems, University of Michigan, Ann Arbor, United States
[3]Michigan Institute for Data Science, University of Michigan, Ann Arbor, United States
[4]Undergraduate Research Opportunities Program, University of Michigan, Ann Arbor, United States
[5]Department of Computer Science and Engineering, Michigan State University, East Lansing, United States
[6]Program 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

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

erouiche
tud.ntnu.no
rondheim
way

Andrei Ivanov
anivanov@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

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

Piotr Luczynski
Lukas Gianinazzi (advisor)
Patrick Iff (advisor)
Daniele De Sensi (advisor)
Leighton Wilson (advisor)
Torsten Hoefler (advisor)

Near-optimal Reduce on the Cerebras Wafer Scale Engine

Pre-order Reduce

tree with pre-order traversal

Results

#PEs P = 512

prediction
t of 3 reduce implementations
he best implementation at runtime

Our code tbd
Recording tbd

1D Allreduce

#PEs P = 512

Reduce followed by broadcast better than classical patterns

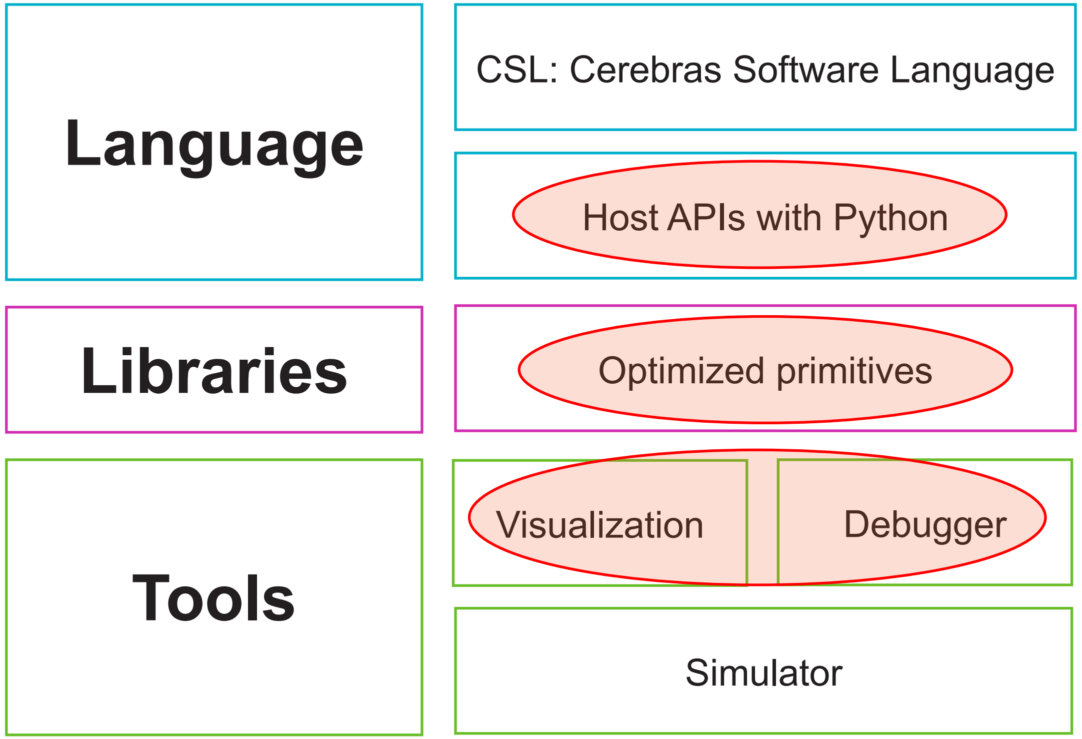
2D

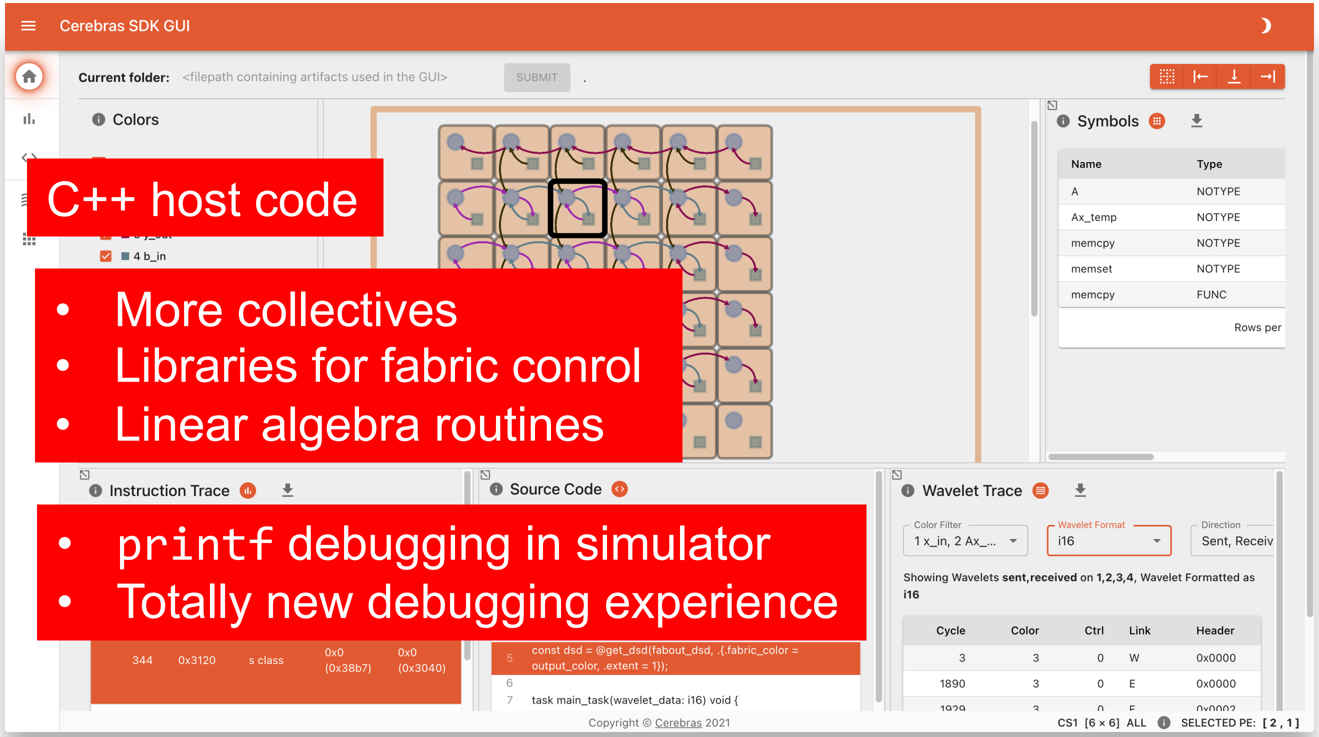
- Need to utilize the whole grid
  1. reduce to the leftmost PE in the row
  2. follow by reduce in the column

#PEs P = 512 x 512

Cerebras SDK Developments

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





Current folder: <filepath containing artifacts used in the GUI>

C++ host code

- More collectives
- Libraries for fabric conrol
- Linear algebra routines

- `printf` debugging in simulator
- Totally new debugging experience

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

```
344    0x3120    s class    0x0         0x0
                            (0x38b7)    (0x3040)
```

```
const dsd = @get_dsd(fabout_dsd, .{.fabric_color =
output_color, .extent = 1});

task main_task(wavelet_data: i16) void {
```

Color Filter: 1 x_in, 2 Ax_...
Wavelet Format: i16
Direction: Sent, Receiv

Showing Wavelets **sent,received** on **1,2,3,4**, Wavelet Formatted as **i16**

Cycle	Color	Ctrl	Link	Header
3	3	0	W	0x0000
1890	3	0	E	0x0000
1929	3	0	E	0x0002

Copyright © Cerebras 2021

CS1 [6 × 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