# Developing Custom Applications at Wafer-Scale

**Leighton Wilson**

leighton.wilson@cerebras.net

**SC24**

# 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

**21 PByte/s** memory bandwidth

**214 Pbit/s** fabric bandwidth

**5nm** process technology

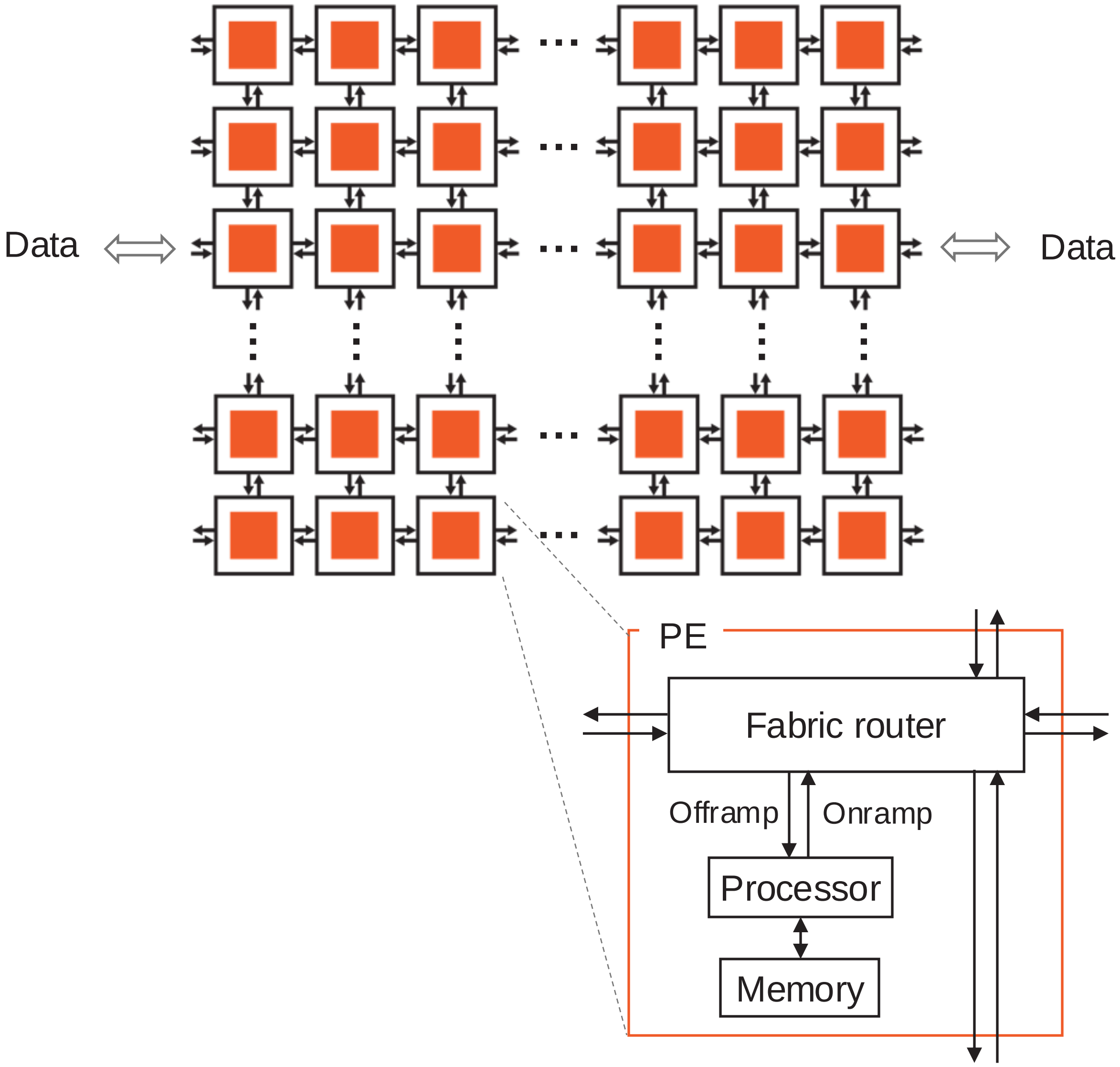
**Cluster-scale acceleration on a single chip**

# 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

- ~900,000 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

# 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

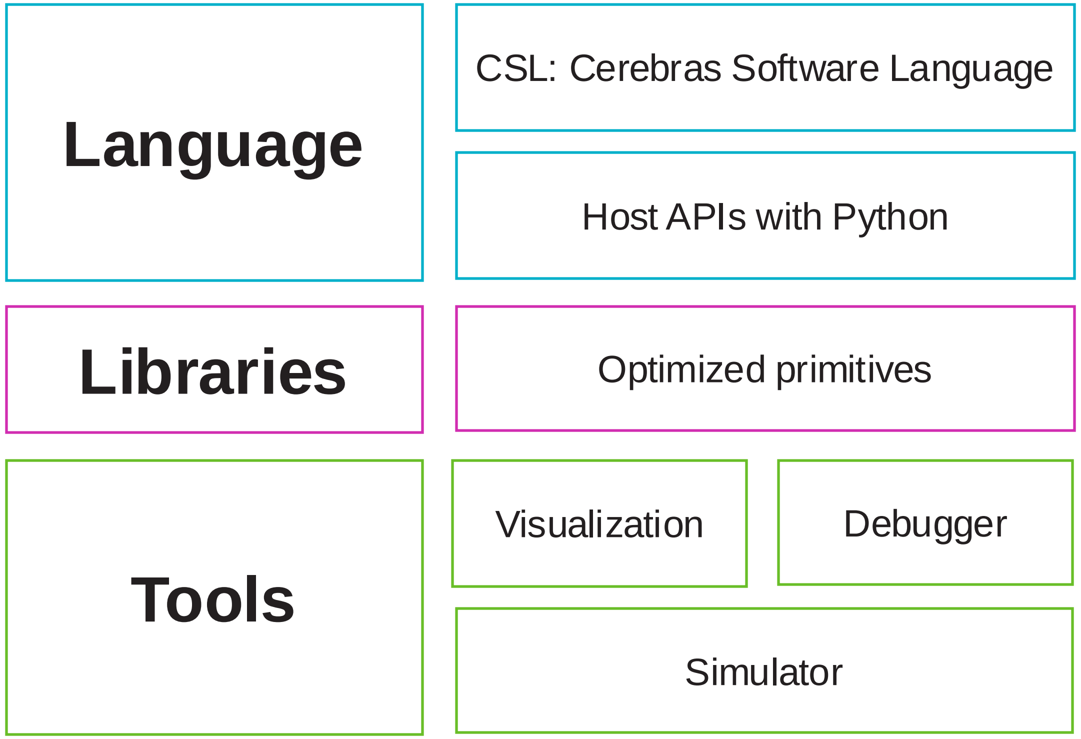
Weights & Biases

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

Cerebras SDK & CSL

# Cerebras SDK

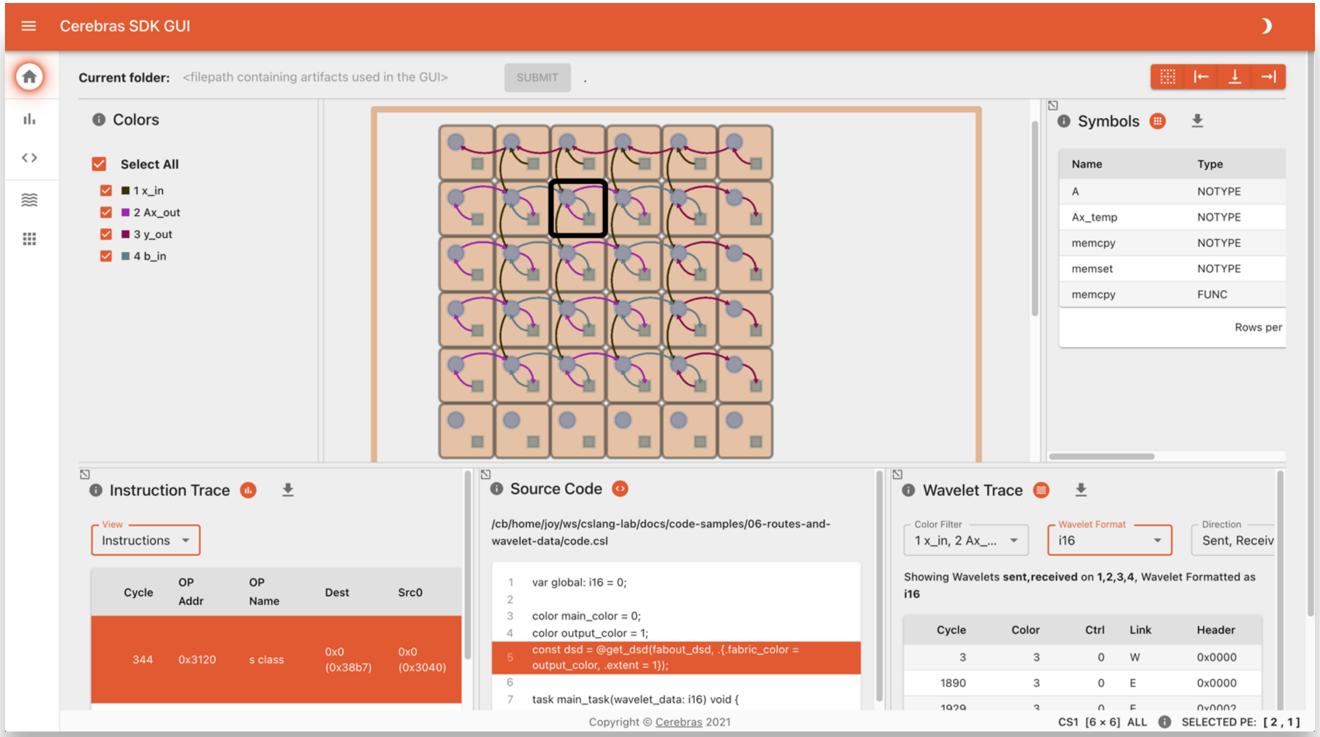
A general-purpose parallel-computing platform and API allowing software developers to write custom programs (“kernels”) for Cerebras systems.



**Language**
- CSL: Cerebras Software Language
- Host APIs with Python

**Libraries**
- Optimized primitives

**Tools**
- Visualization
- Debugger
- Simulator



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

☑ 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 ]

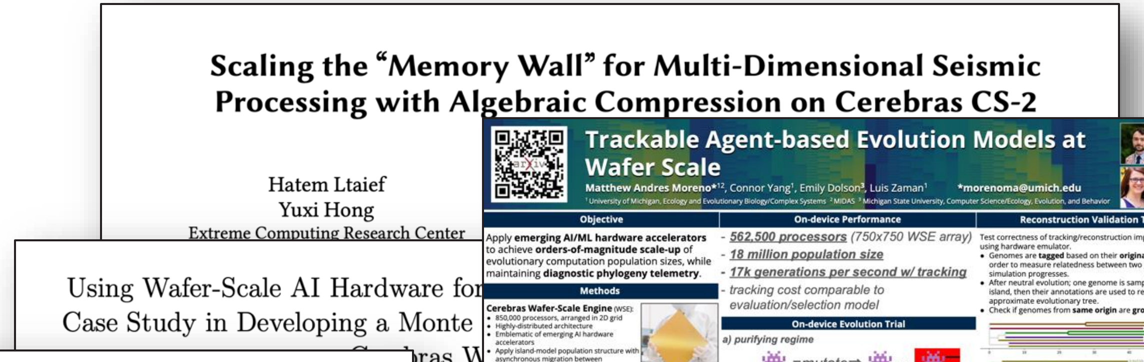
# SDK Example Programs Available

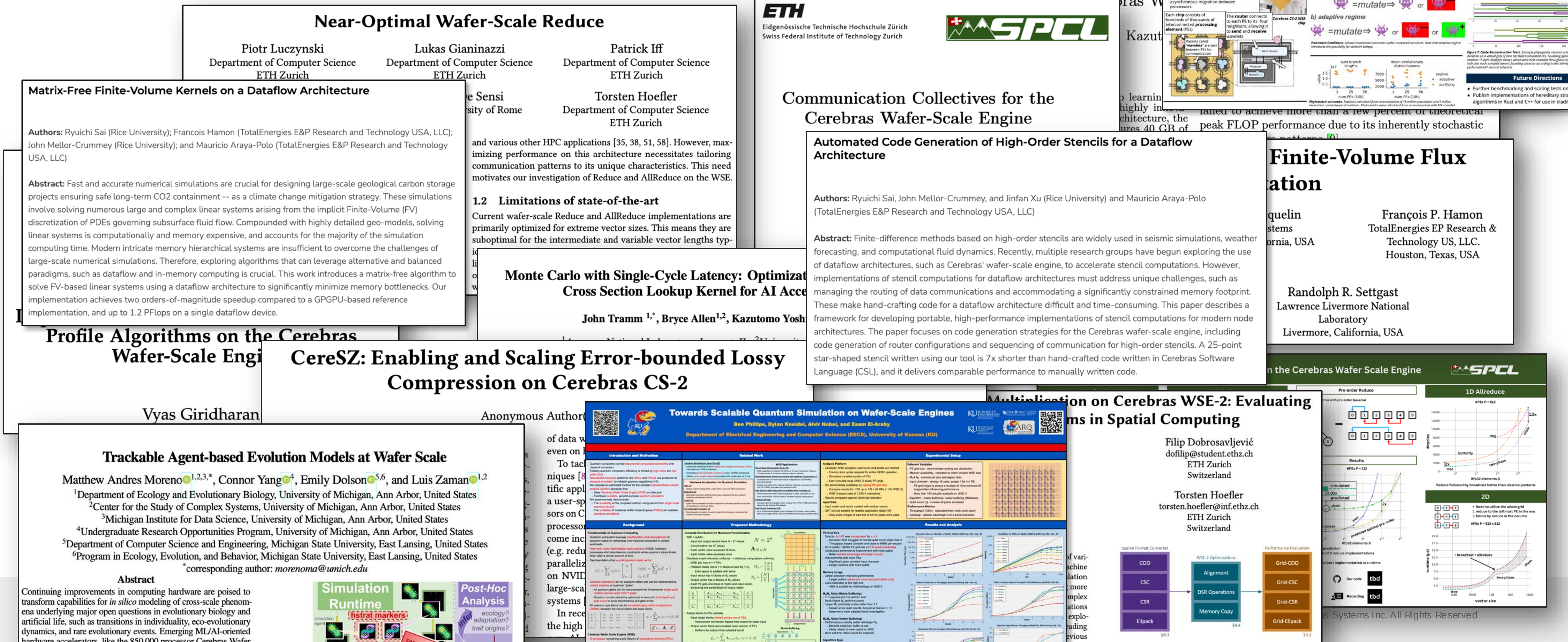
**Repository:** github.com/Cerebras/csl-examples

- 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

Hatem Ltaief
Yuxi Hong
Extreme Computing Research Center

Trackable Agent-based Evolution Models at Wafer Scale

Matthew Andres Moreno*[1,2], Connor Yang[1], Emily Dolson[3], Luis Zaman[1] *morenoma@umich.edu

[1] University of Michigan, Ecology and Evolutionary Biology/Complex Systems [2] MIDAS [3] Michigan State University, Computer Science/Ecology, Evolution, and Behavior

Objective

Apply **emerging AI/ML hardware accelerators** to achieve **orders-of-magnitude scale-up** of evolutionary computation population sizes, while maintaining **diagnostic phylogeny telemetry**.

Methods

**Cerebras Wafer-Scale Engine** (WSE):
- 850,000 processors, arranged in 2D grid
- Highly-distributed architecture
- Emblematic of emerging AI hardware accelerators
- Apply island-model population structure with asynchronous migration between processors.

On-device Performance

- *562,500 processors* (750x750 WSE array)
- *18 million population size*
- *17k generations per second w/ tracking*
- *tracking cost comparable to evaluation/selection model*

On-device Evolution Trial

a) purifying regime

b) adaptive regime

Reconstruction Validation

Test correctness of tracking/reconstruction using hardware emulator.
- Genomes are **tagged** based on their original order to measure relatedness between two simulation progresses.
- After neutral evolution; one genome is sampled per island, then their annotations are used to approximate evolutionary tree.
- Check if genomes from **same origin** are grouped

Future Directions

- Further benchmarking and scaling tests on
- Publish implementations of hereditary stratigraphy algorithms in Rust and C++ for use in tradition

Using Wafer-Scale AI Hardware for Case Study in Developing a Monte Carlo Cerebras W

Kazut

## 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

e Sensi
sity of Rome

Torsten Hoefler
Department of Computer Science
ETH Zurich

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 suboptimal for the intermediate and variable vector lengths typ-

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SPCL

Communication Collectives for the Cerebras Wafer-Scale Engine

learnin highly in chitecture, the ures 40 GB of

failed to achieve more than a few percent of theoretical peak FLOP performance due to its inherently stochastic

### Matrix-Free Finite-Volume Kernels on a Dataflow Architecture

**Authors:** Ryuichi Sai (Rice University); Francois Hamon (TotalEnergies E&P Research and Technology USA, LLC); John Mellor-Crummey (Rice University); and Mauricio Araya-Polo (TotalEnergies E&P Research and Technology USA, LLC)

**Abstract:** Fast and accurate numerical simulations are crucial for designing large-scale geological carbon storage projects ensuring safe long-term CO2 containment -- as a climate change mitigation strategy. These simulations involve solving numerous large and complex linear systems arising from the implicit Finite-Volume (FV) discretization of PDEs governing subsurface fluid flow. Compounded with highly detailed geo-models, solving linear systems is computationally and memory expensive, and accounts for the majority of the simulation computing time. Modern intricate memory hierarchical systems are insufficient to overcome the challenges of large-scale numerical simulations. Therefore, exploring algorithms that can leverage alternative and balanced paradigms, such as dataflow and in-memory computing is crucial. This work introduces a matrix-free algorithm to solve FV-based linear systems using a dataflow architecture to significantly minimize memory bottlenecks. Our implementation achieves two orders-of-magnitude speedup compared to a GPGPU-based reference implementation, and up to 1.2 PFlops on a single dataflow device.

### Automated Code Generation of High-Order Stencils for a Dataflow Architecture

**Authors:** Ryuichi Sai, John Mellor-Crummey, and Jinfan Xu (Rice University) and Mauricio Araya-Polo (TotalEnergies E&P Research and Technology USA, LLC)

**Abstract:** Finite-difference methods based on high-order stencils are widely used in seismic simulations, weather forecasting, and computational fluid dynamics. Recently, multiple research groups have begun exploring the use of dataflow architectures, such as Cerebras' wafer-scale engine, to accelerate stencil computations. However, implementations of stencil computations for dataflow architectures must address unique challenges, such as managing the routing of data communications and accommodating a significantly constrained memory footprint. These make hand-crafting code for a dataflow architecture difficult and time-consuming. This paper describes a framework for developing portable, high-performance implementations of stencil computations for modern node architectures. The paper focuses on code generation strategies for the Cerebras wafer-scale engine, including code generation of router configurations and sequencing of communication for high-order stencils. A 25-point star-shaped stencil written using our tool is 7x shorter than hand-crafted code written in Cerebras Software Language (CSL), and it delivers comparable performance to manually written code.

Finite-Volume Flux
ation

quelin
stems
rnia, USA

François P. Hamon
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: Optimizat Cross Section Lookup Kernel for AI Acce

John Tramm [1,*], Bryce Allen[1,2], Kazutomo Yosh

Profile Algorithms on the Cerebras Wafer-Scale Engi

Vyas Giridharan

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

Anonymous Author(

of data w even on l To tack niques [8 tific appl a user-sp sors on C processo come inc (e.g. redu paralleliz on NVID large-sca systems In rece the high

## 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

Simulation Runtime

Post-Hoc Analysis

ecology? adaptation? trait origins?

Towards Scalable Quantum Simulation on Wafer-Scale Engines

Ben Phillips, Dylan Kneidel, Alvir Nobel, and Esam El-Araby
Department of Electrical Engineering and Computer Science (EECS), University of Kansas (KU)

Introduction and Motivation

Related Work

Experimental Setup

Background

Proposed Methodology

Results and Analysis

Multiplication on Cerebras WSE-2: Evaluating ms in Spatial Computing

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

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

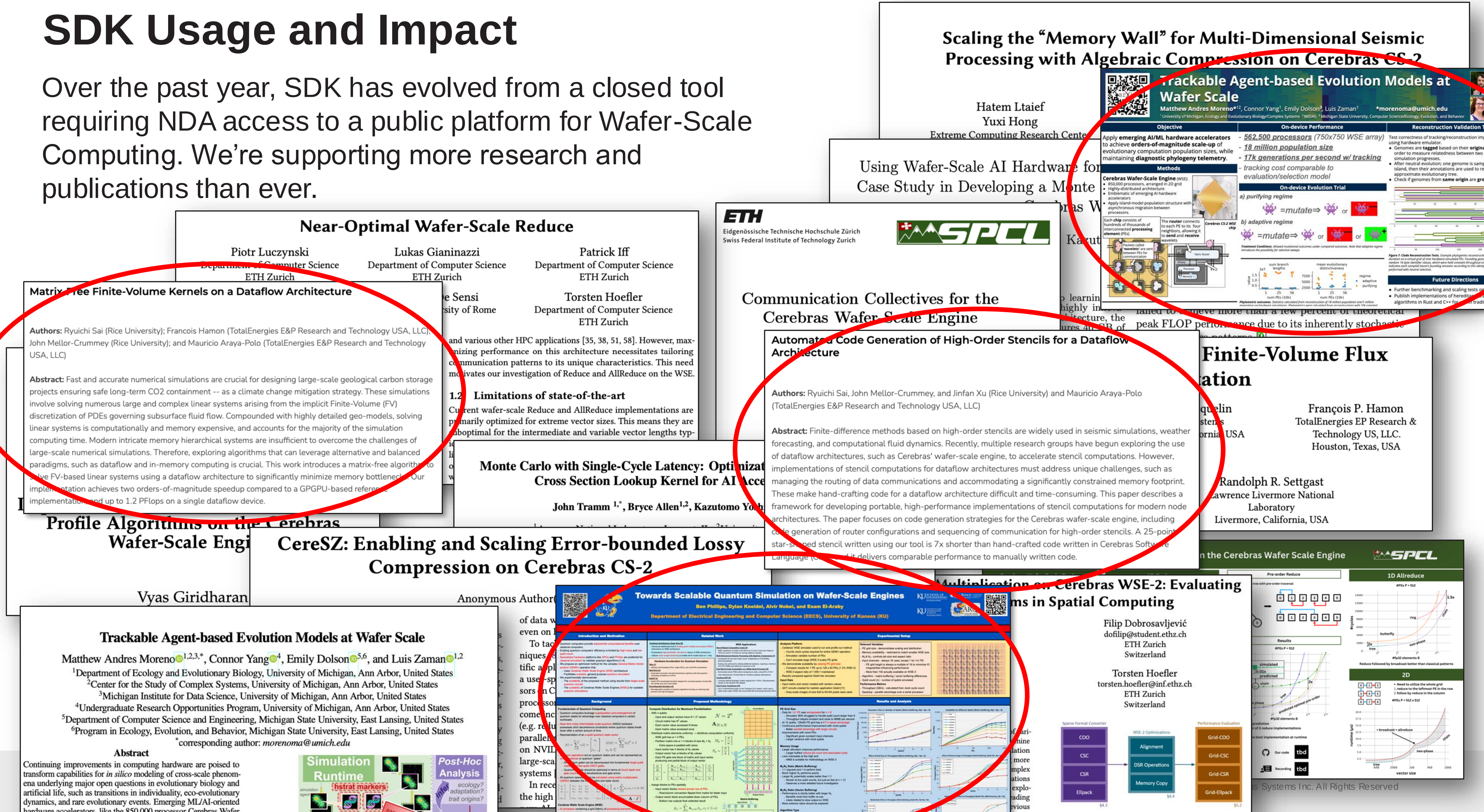
of vari chine lation more mplex lations explo ading vious

n the Cerebras Wafer Scale Engine

SPCL

# 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

Hatem Ltaief
Yuxi Hong
Extreme Computing Research Cente

Trackable Agent-based Evolution Models at Wafer Scale

Matthew Andres Moreno*[1][2], Connor Yang[1], Emily Dolson[3], Luis Zaman[1] *morenoma@umich.edu
[1] University of Michigan, Ecology and Evolutionary Biology/Complex Systems [2] MIDAS [3] Michigan State University, Computer Science/Ecology, Evolution, and Behavior

Objective

Apply **emerging AI/ML hardware accelerators** to achieve **orders-of-magnitude scale-up** of evolutionary computation population sizes, while maintaining **diagnostic phylogeny telemetry**.

Methods

**Cerebras Wafer-Scale Engine** (WSE):
- 850,000 processors, arranged in 2D grid
- Highly-distributed architecture
- Emblematic of emerging AI hardware accelerators
- Apply island-model population structure with asynchronous migration between processors.

On-device Performance

- 562,500 processors (750x750 WSE array)
- 18 million population size
- 17k generations per second w/ tracking
- tracking cost comparable to evaluation/selection model

On-device Evolution Trial

a) purifying regime

b) adaptive regime

Reconstruction Validation

Future Directions

Using Wafer-Scale AI Hardware for
Case Study in Developing a Monte

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SPCL

Communication Collectives for the Cerebras Wafer-Scale Engine

## 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

Torsten Hoefler
Department of Computer Science
ETH Zurich

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 suboptimal for the intermediate and variable vector lengths typ-

Matrix-Free Finite-Volume Kernels on a Dataflow Architecture

**Authors:** Ryuichi Sai (Rice University); Francois Hamon (TotalEnergies E&P Research and Technology USA, LLC); John Mellor-Crummey (Rice University); and Mauricio Araya-Polo (TotalEnergies E&P Research and Technology USA, LLC)

**Abstract:** Fast and accurate numerical simulations are crucial for designing large-scale geological carbon storage projects ensuring safe long-term CO2 containment -- as a climate change mitigation strategy. These simulations involve solving numerous large and complex linear systems arising from the implicit Finite-Volume (FV) discretization of PDEs governing subsurface fluid flow. Compounded with highly detailed geo-models, solving linear systems is computationally and memory expensive, and accounts for the majority of the simulation computing time. Modern intricate memory hierarchical systems are insufficient to overcome the challenges of large-scale numerical simulations. Therefore, exploring algorithms that can leverage alternative and balanced paradigms, such as dataflow and in-memory computing is crucial. This work introduces a matrix-free algorithm to solve FV-based linear systems using a dataflow architecture to significantly minimize memory bottlenecks. Our implementation achieves two orders-of-magnitude speedup compared to a GPGPU-based reference implementation, and up to 1.2 PFlops on a single dataflow device.

Profile Algorithms on the Cerebras Wafer-Scale Engi

Vyas Giridharan

Monte Carlo with Single-Cycle Latency: Optimizat
Cross Section Lookup Kernel for AI Acce

John Tramm[1,*], Bryce Allen[1,2], Kazutomo Yo

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

Anonymous Author(

Automated Code Generation of High-Order Stencils for a Dataflow Architecture

**Authors:** Ryuichi Sai, John Mellor-Crummey, and Jinfan Xu (Rice University) and Mauricio Araya-Polo (TotalEnergies E&P Research and Technology USA, LLC)

**Abstract:** Finite-difference methods based on high-order stencils are widely used in seismic simulations, weather forecasting, and computational fluid dynamics. Recently, multiple research groups have begun exploring the use of dataflow architectures, such as Cerebras' wafer-scale engine, to accelerate stencil computations. However, implementations of stencil computations for dataflow architectures must address unique challenges, such as managing the routing of data communications and accommodating a significantly constrained memory footprint. These make hand-crafting code for a dataflow architecture difficult and time-consuming. This paper describes a framework for developing portable, high-performance implementations of stencil computations for modern node architectures. The paper focuses on code generation strategies for the Cerebras wafer-scale engine, including code generation of router configurations and sequencing of communication for high-order stencils. A 25-point star-shaped stencil written using our tool is 7x shorter than hand-crafted code written in Cerebras Software Language (C... and it delivers comparable performance to manually written code.

Finite-Volume Flux
ation

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

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

Towards Scalable Quantum Simulation on Wafer-Scale Engines

Ben Phillips, Dylan Kneidel, Alvir Nobel, and Esam El-Araby
Department of Electrical Engineering and Computer Science (EECS), University of Kansas (KU)

Introduction and Motivation

Related Work

Experimental Setup

Background

Proposed Methodology

Results and Analysis

Multiplication on Cerebras WSE-2: Evaluating
ms in Spatial Computing

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

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

n the Cerebras Wafer Scale Engine

SPCL

Pre-order Reduce

1D Allreduce

2D

## 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

# 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, LRZ, …

Questions? leighton.wilson@cerebras.net

discourse.cerebras.net

cerebras.net/developers/sdk-request