Wafer-scale AI for science and HPC

Cerebras Systems

A Hock

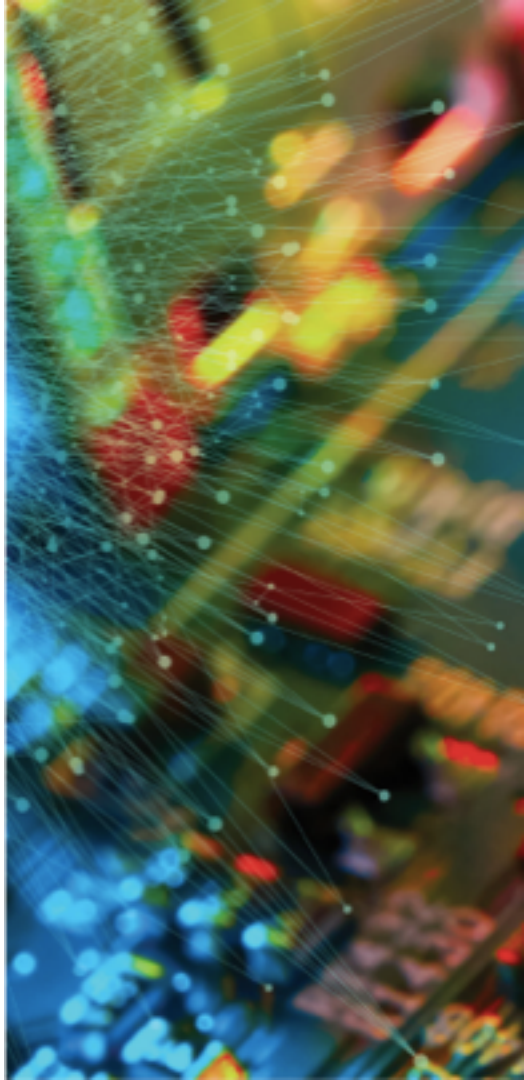
ISC Machine Learning Hardware Workshop 2020
25 June 2020

Introduction

Opportunity

From e.g. fundamental physics to energy, environment, human health

AI has massive potential for science and HPC



Introduction

Opportunity

From e.g. fundamental physics to energy, environment, human health

AI has massive potential for science and HPC

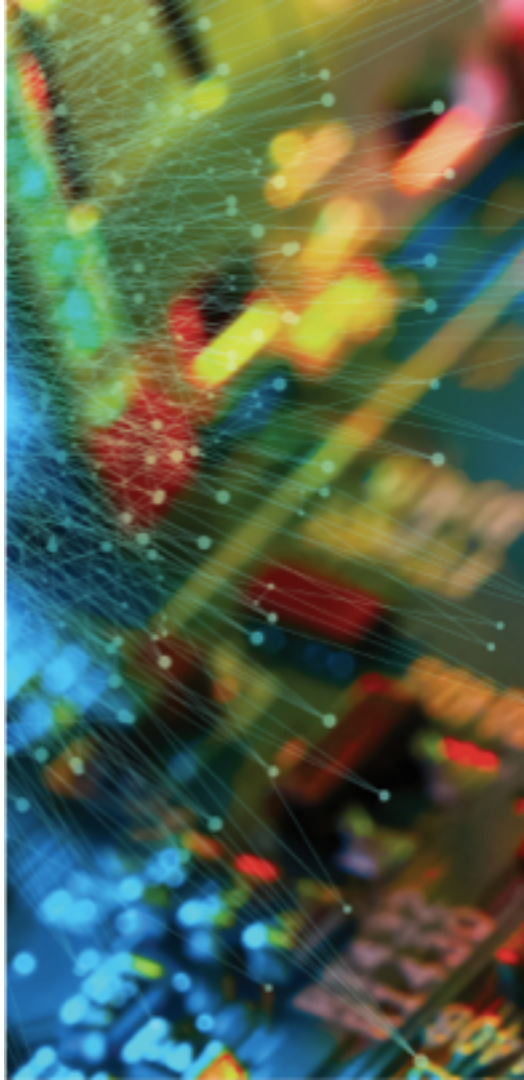
Challenge

NN compute is unique, challenging for legacy processors

Training commonly takes days-weeks, even on clusters of GPU

This is inefficient, expensive, constrains innovation

AI for science is compute-limited today



cerebras

Introduction

Opportunity

From e.g. fundamental physics to energy, environment, human health

AI has massive potential for science and HPC

Challenge

NN compute is unique, challenging for legacy processors

Training commonly takes days-weeks, even on clusters of GPU

This is inefficient, expensive, constrains innovation

AI for science is compute-limited today

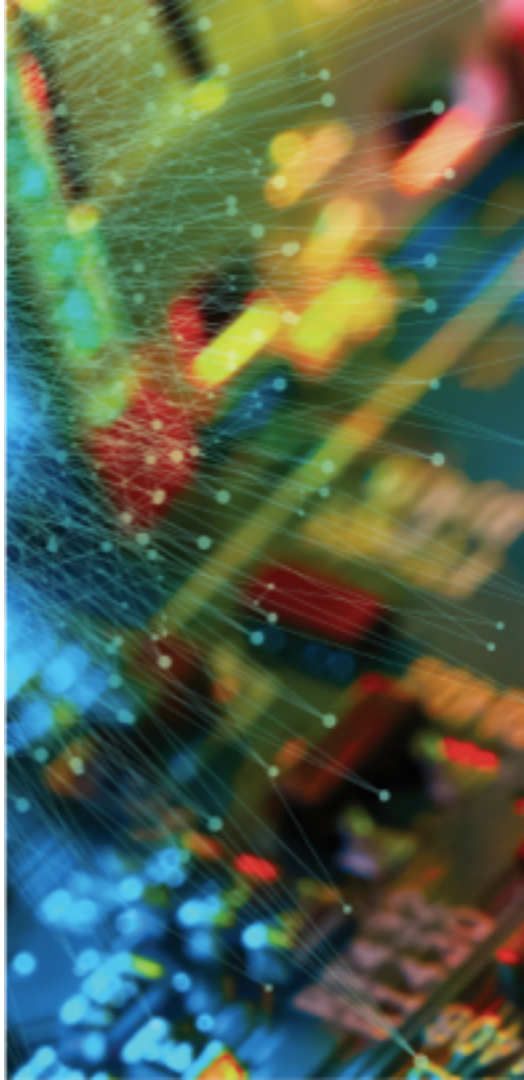
Need

Massive compute: faster wall clock training and inference

Programmability with today's tools, ability to optimize and extend

Systems that can integrate with HPC and scientific instrument facilities

We need a new solution for AI compute



cerebras

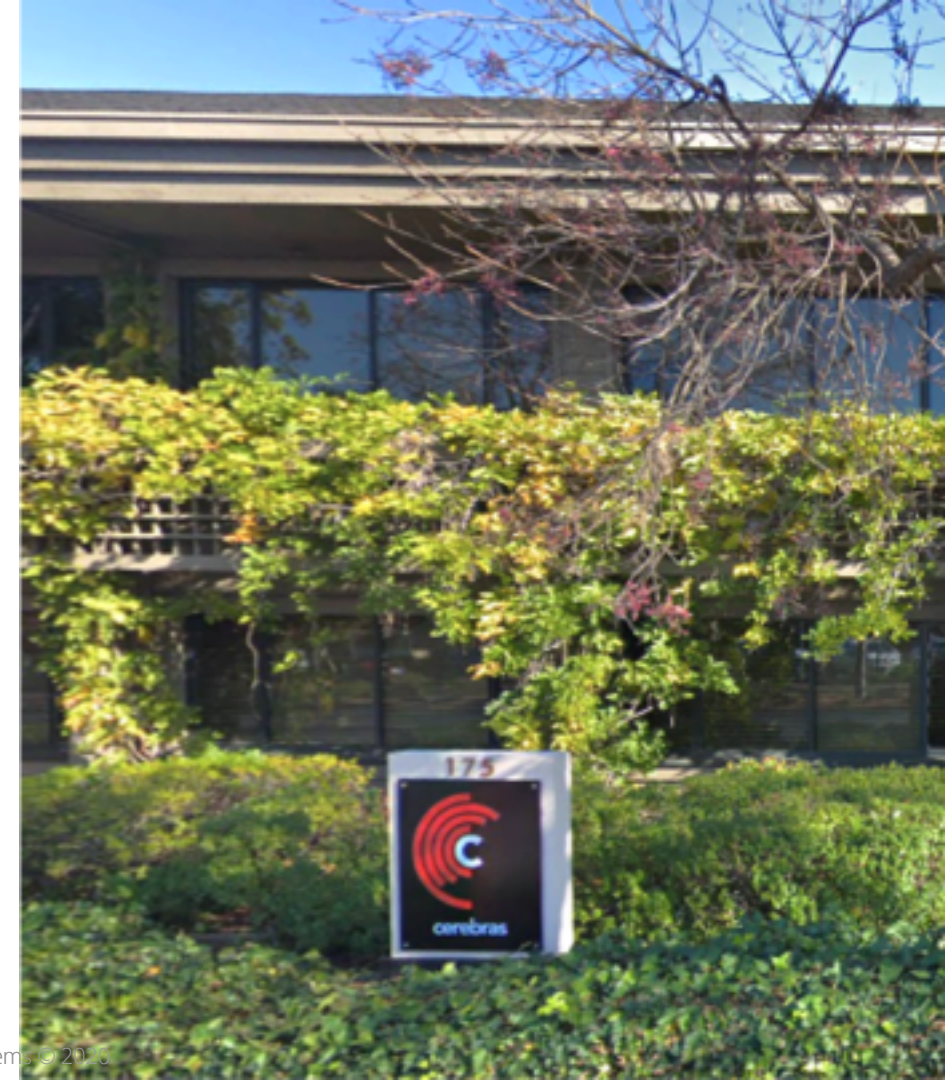
Enter Cerebras Systems

Founded in 2016

Building systems to accelerate and **change the landscape of compute for AI**

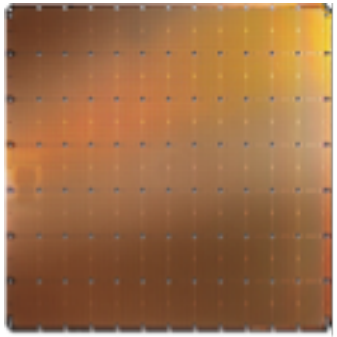
200+ **world-class engineering** team

- HW, SW, ML research



cerebras

Cerebras' solution overview



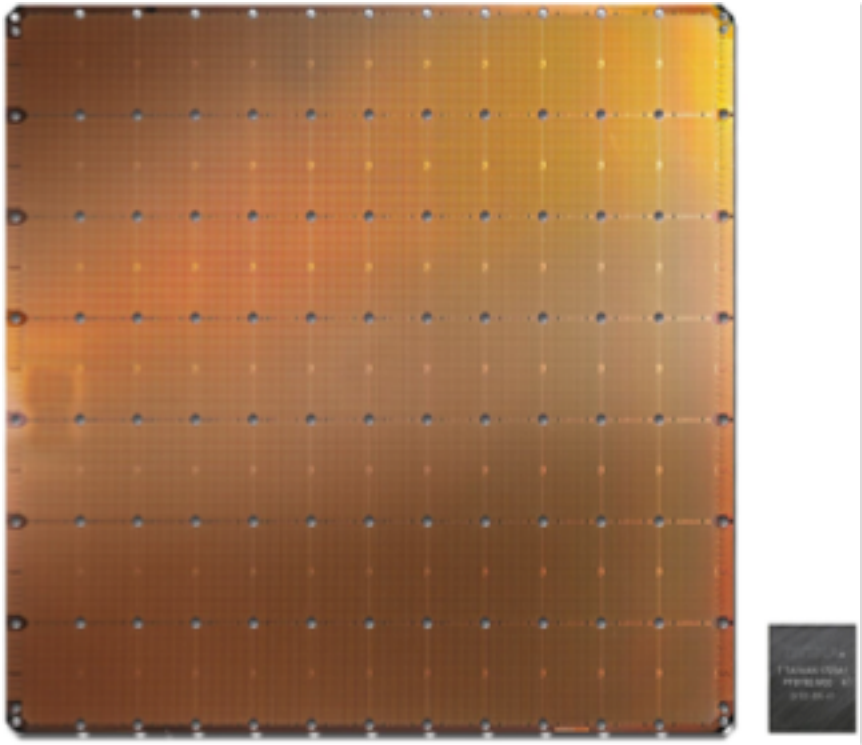
A complete AI compute solution for AI at HPC scale

Unique, massive high performance **processor**
→ Training and inference
→ Orders of magnitude performance gain beyond legacy processors

Software stack to meet users where they are
→ Programmable as a single node with standard frameworks

System
→ Replaces racks of equipment with a single system
→ Straightforward deployment, orchestration

The Cerebras Wafer-Scale Engine



The world's largest chip and most powerful processor for AI.

Designed from the ground-up to deliver orders of magnitude performance gain for deep learning.

- 215 x 215 mm, 1.2 trillion transistor chip
- 400,000 cores
- 18 GB on-chip SRAM
- 100 Pb/s interconnect

cerebras

Optimized architecture for DL compute

Massive compute: cluster-scale resources on a single chip

- Core optimized for neural network primitives
- Flexible, programmable core – support evolving range of neural network architectures
- Dataflow architecture, sparsity harvesting – designed for sparse compute native to NN

Local memory: all on-chip SRAM – efficient local access model weights & activations; datapath has full performance from memory.

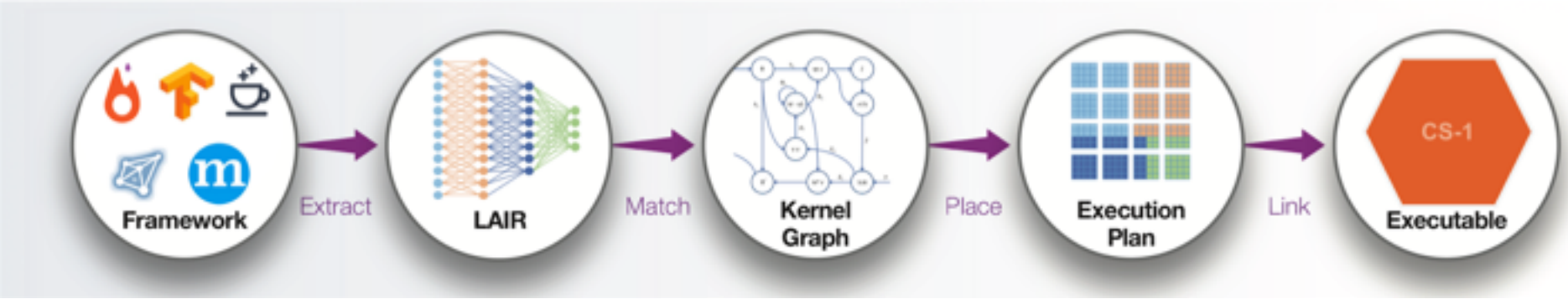
Fast interconnect: configurable on-chip fabric, communicate layer-to-layer with high bandwidth and low latency

Together unlock, e.g. **model parallel** execution on-chip, full utilization training and inference down to **batch 1**.

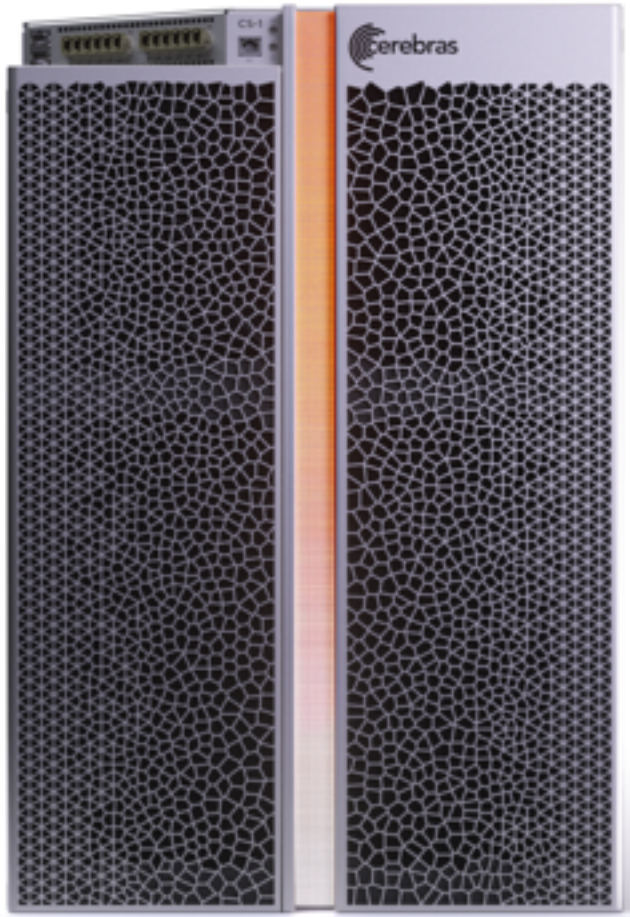
cerebras

The Cerebras Software Platform

Our software stack makes the Wafer-Scale Engine easy to use:



→ Programmable with today's ML frameworks
→ Flexible and customizable with lower level APIs



The Cerebras CS-1

A **full solution** in a single system:
- Powered by the WSE
- Programmable via TF, other frameworks
- Install, deploy easily into a standard rack

Readily **integrate with existing HPC** systems
- 1.2 Tbps I/O via 12x 100 GbE

Orchestrate with standard frameworks

Cluster multiple systems for greater acceleration and scale

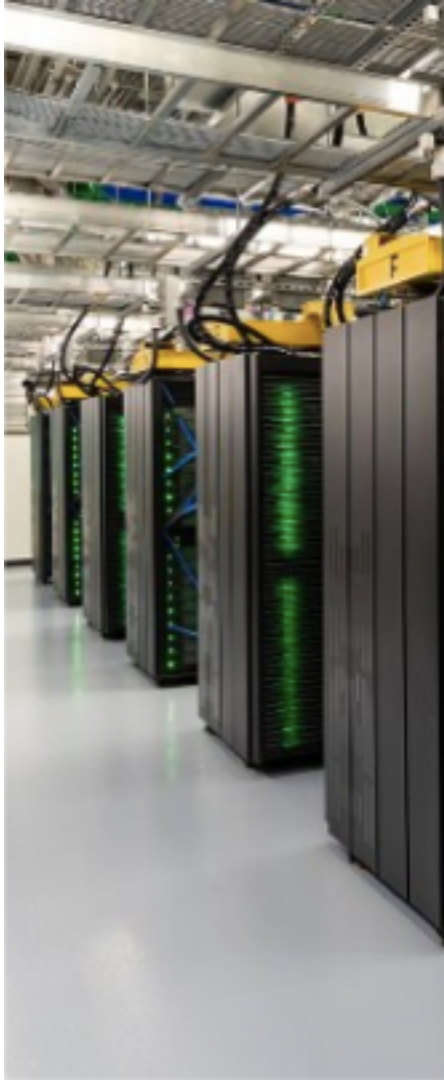
cerebras

Value proposition for HPC+AI for science

Orders of magnitude acceleration in wall-clock **training** time
→ More experiments per unit time; larger datasets, greater accuracy
→ Higher cadence re-training

Orders of magnitude higher throughput, lower latency **inference**
→ Augment physics-based simulation
→ High performance inference for scientific instrument facilities

Flexible compute engine
→ Unlock research into new NN architectures, ML methods



Concluding remarks

Proud to introduce ISC to the **Cerebras CS-1**, the world's most powerful deep learning solution.

Built to **accelerate AI+HPC by orders of magnitude and empower researchers** like you to do more, faster.

Concluding remarks

Proud to introduce ISC to the **Cerebras CS-1**, the world's most powerful deep learning solution.

Built to **accelerate AI+HPC by orders of magnitude and empower researchers** like you to do more, faster.

Multiple systems deployed and **running customer workloads today**, all the way from TF.
- Accelerating AI for science, e.g. with DoE and NSF!

Call to action: bring us your big and different problems, system and partnership interests. Can't wait to work together.

Thank you!

cerebras