Extreme-scale AI computing with Cerebras

A Hock *
Cerebras Systems

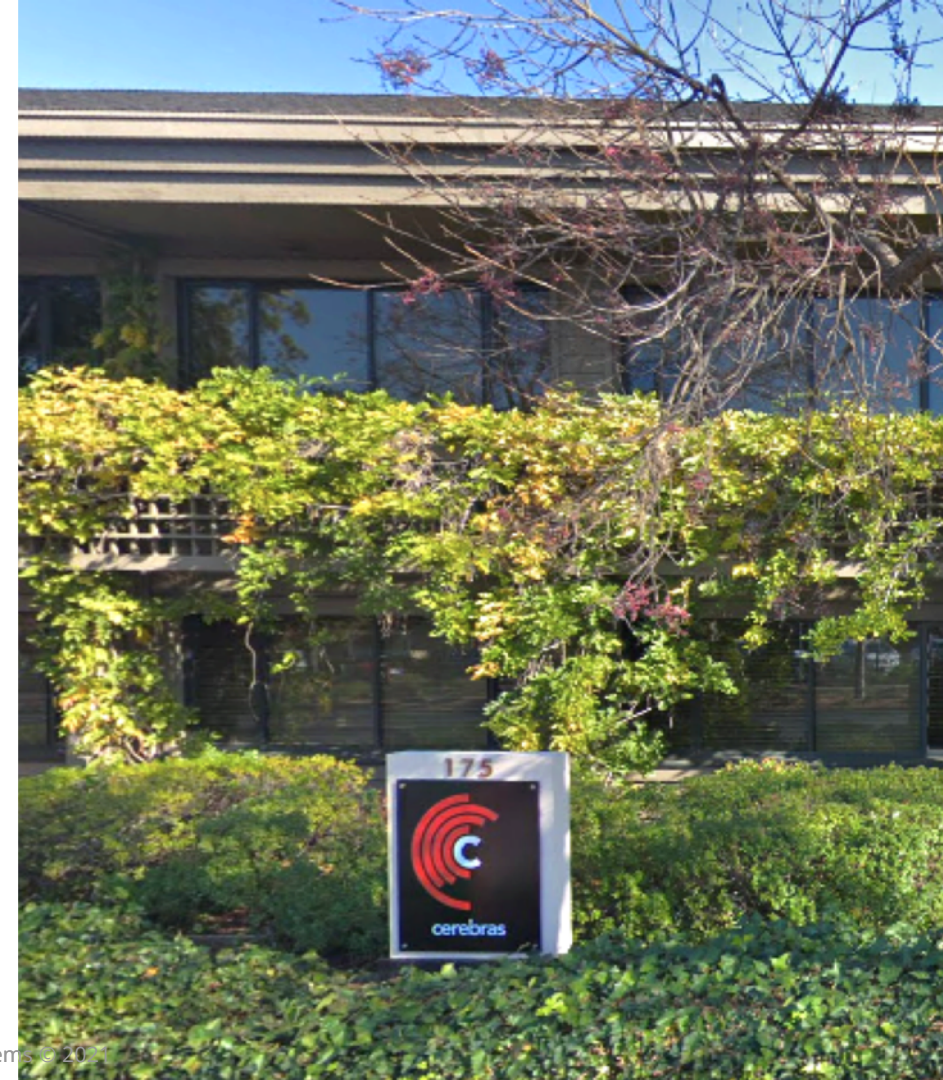
* andy@cerebras.net

Argonne Training Program on Extreme-Scale Computing (ATPESC)
02 August 2021

Cerebras Systems

AI computer systems company
Founded 2016, Silicon Valley HQ
Transform the compute landscape
Radically accelerate AI
Now 300+ world-class engineers
Hardware, software, ML/AI research

BENCHMARK | foundation capital | ECLIPSE
coatue | VY capital | ALTIMETER



cerebras

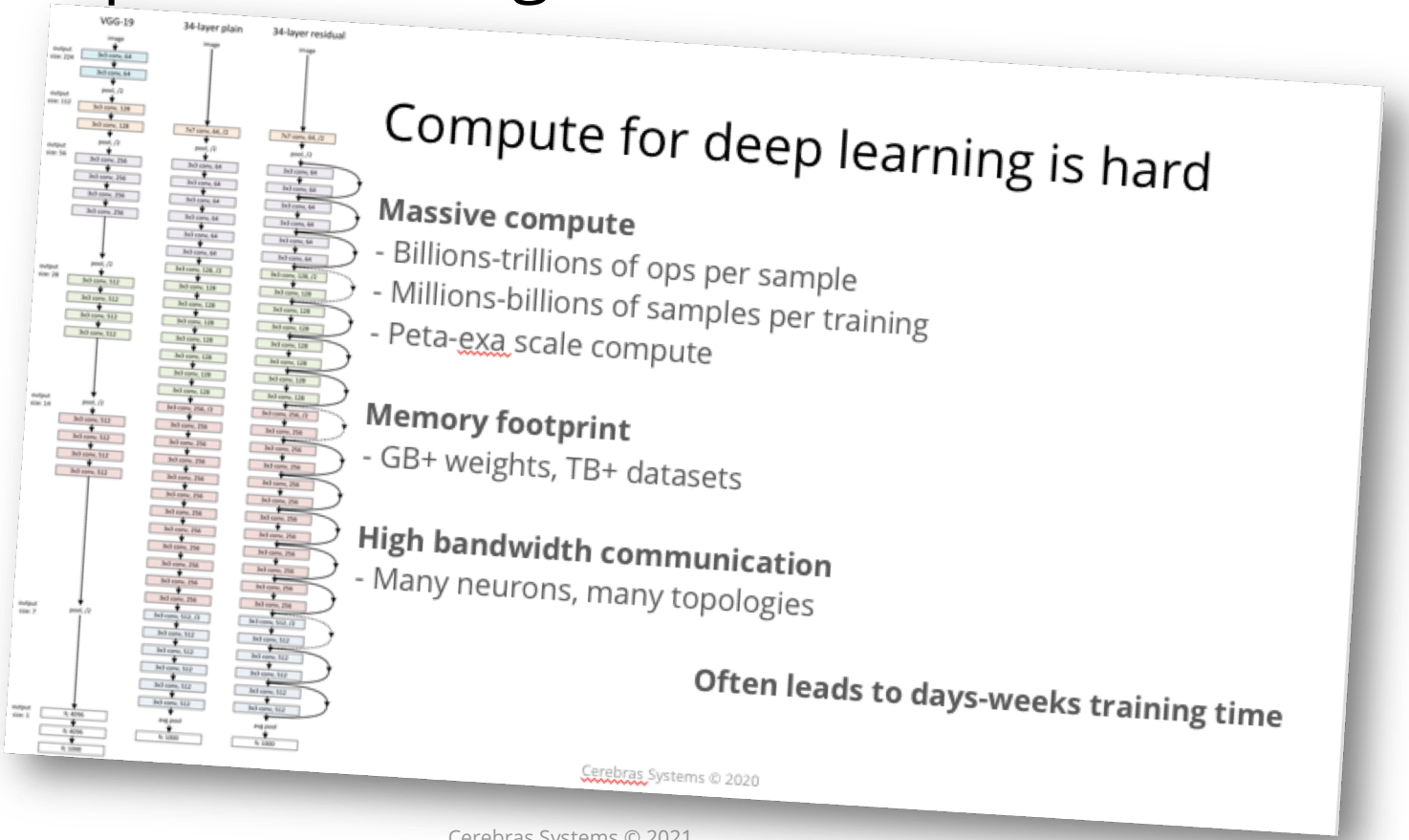
AI is compute-limited today.

Existing processors can't keep up,
were built for other work.

We **need a new compute solution**
to accelerate deep learning.

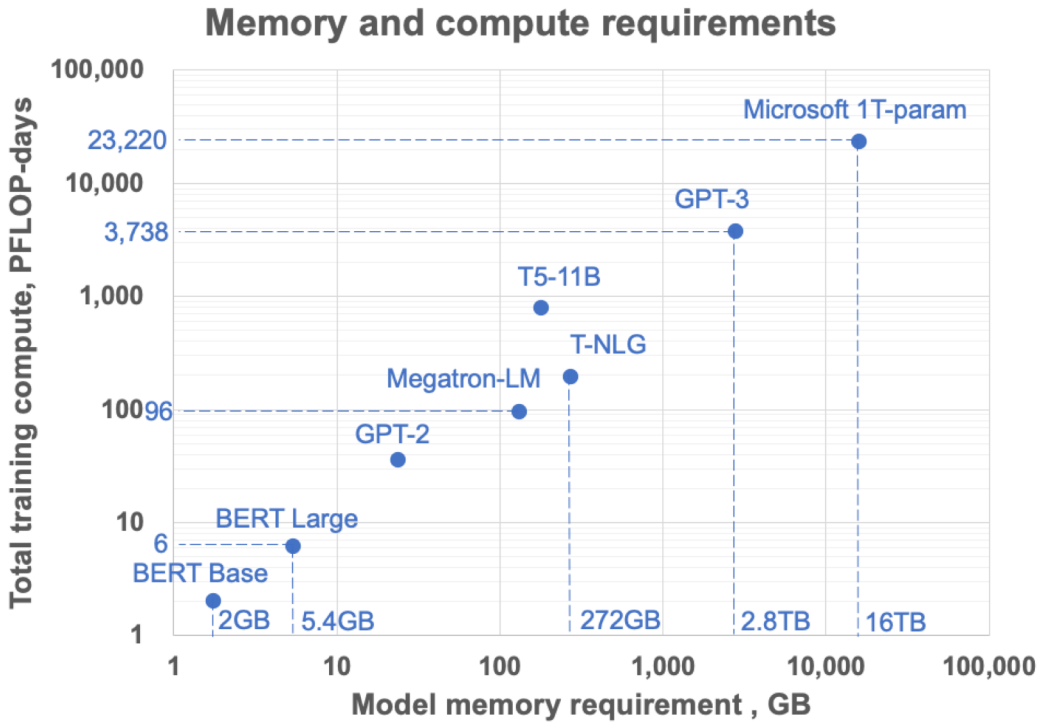
The AI compute challenge…

Last year:



The AI compute challenge…continues to grow:

Memory and compute requirements for modern NLP



1 PFLOP-day is about 1 x DGX-2H
or
1 x DGX-A100 busy for a day

NVIDIA Megatron-LM:
trained on **512 V100** (32 DGX-2H)
for **about 10 days**

OpenAI GPT-3:
trained on **1024 V100** (64 DGX-2H) for **about 116 days**

cerebras

Need a solution that supports larger *and smarter* models

Brute-force scaling is historical path to better models.
Presents a **challenge to traditional architectures and clusters**:

- **Memory** capacity and bandwidth needs to grow;
- More **compute** and communication bandwidth;
- Simpler programming model to enable more iterative R&D.

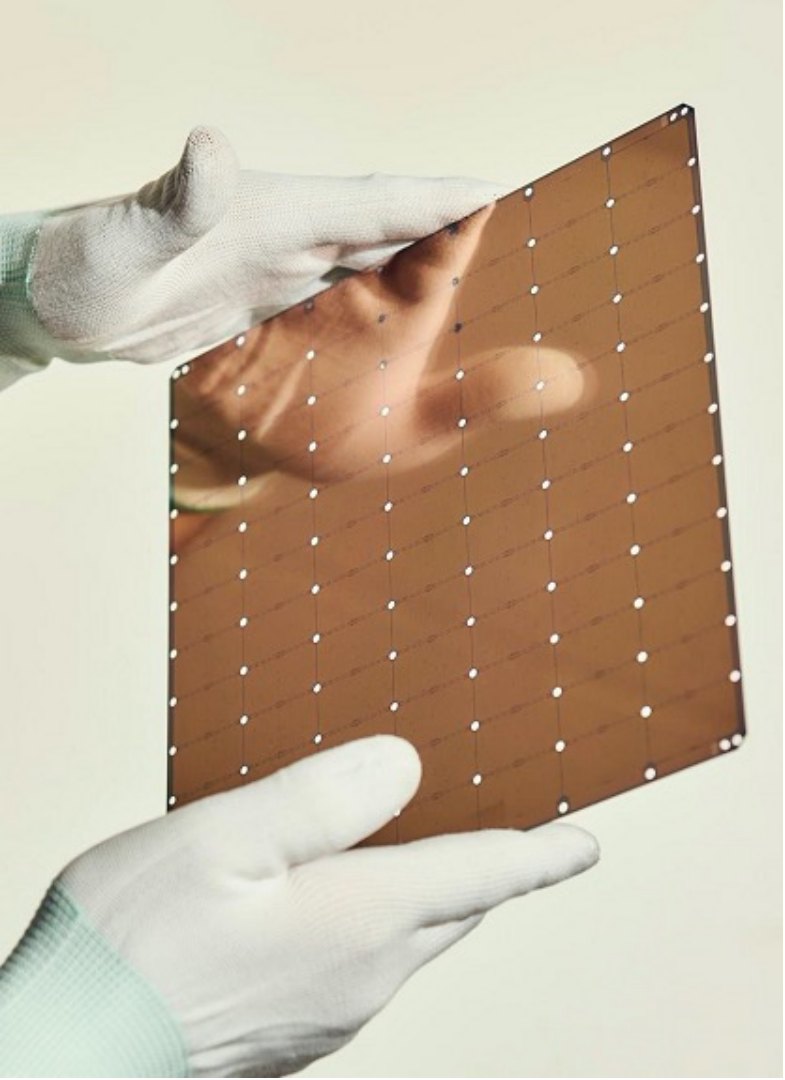
Meanwhile, **algorithmic innovations are opening a path to more efficient models**:

- Sparse models, models with dynamic or conditional compute;
- These are **promising but challenge existing hardware**.

We **need a solution that delivers both**: extreme scale with fewer nodes, flexible compute for smarter, efficient models.

cerebras

Our solution

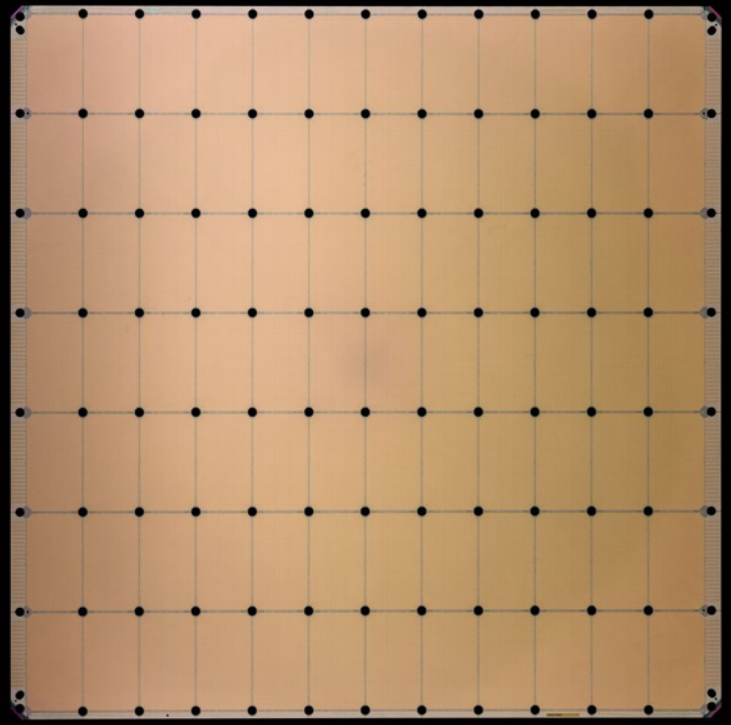


Cerebras Wafer-Scale Engine

	WSE-2
Fabrication process	**7 nm**
Silicon area	**46,225 mm²**
Transistors	**2.6 Trillion**
AI-optimized cores	**850,000**
Memory on-chip	**40 GB**
Memory bandwidth	**20 PB/s**
Fabric bandwidth	**220 Pb/s**

A cluster of flexible, sparse linear algebra compute resources, all on a single chip

Cerebras Wafer Scale Engine



Cerebras WSE-7nm

2.6 Trillion Transistors

46,225 mm² Silicon

Largest GPU

54.2 Billion Transistors

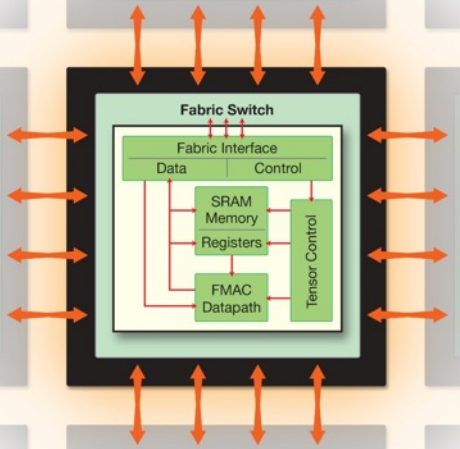
826 mm² Silicon

	Cerebras WSE-2	A100	Cerebras Advantage
Chip size	46,225 mm^2	826 mm^2	56 X
Cores	850,000	6912 + 432	123X
On-chip memory	40 Gigabytes	40 Megabytes	1,000 X
Memory bandwidth	20 Petabytes/sec	1555 Gigabytes/sec	12,733 X
Fabric bandwidth	220 Petabits/sec	600 Gigabytes/sec	45,833 X

The CS WSE architecture is built for deep learning

AI-optimized **compute**

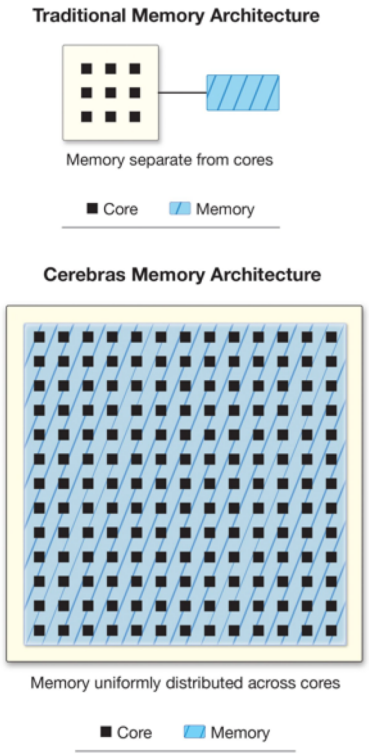
- Fully-programmable core, ML-optimized extensions
- Fine-grained dataflow compute for sparse, dynamic workloads



The CS WSE architecture is built for deep learning

AI-optimized **compute**
- Fully-programmable core, ML-optimized extensions
- Fine-grained dataflow compute for sparse, dynamic workloads

AI-optimized **memory**
- Traditional memory architectures shared memory far from compute
- The right answer is distributed, high performance, on-chip memory

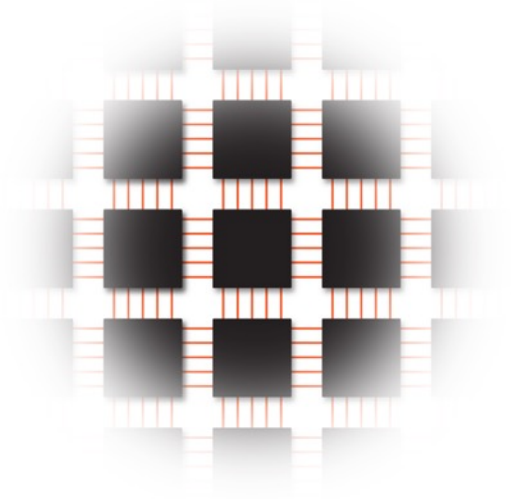


The CS WSE architecture is built for deep learning

AI-optimized **compute**
- Fully-programmable core, ML-optimized extensions
- Fine-grained dataflow compute for sparse, dynamic workloads

AI-optimized **memory**
- Traditional memory architectures shared memory far from compute
- The right answer is distributed, high performance, on-chip memory

AI-optimized **communication**
- High bandwidth, low latency cluster-scale networking on chip
- Fully-configurable to user-specified topology



cerebras

The CS WSE architecture is built for deep learning

AI-optimized **compute**
- Fully-programmable core, ML-optimized extensions
- Fine-grained dataflow compute for sparse, dynamic workloads

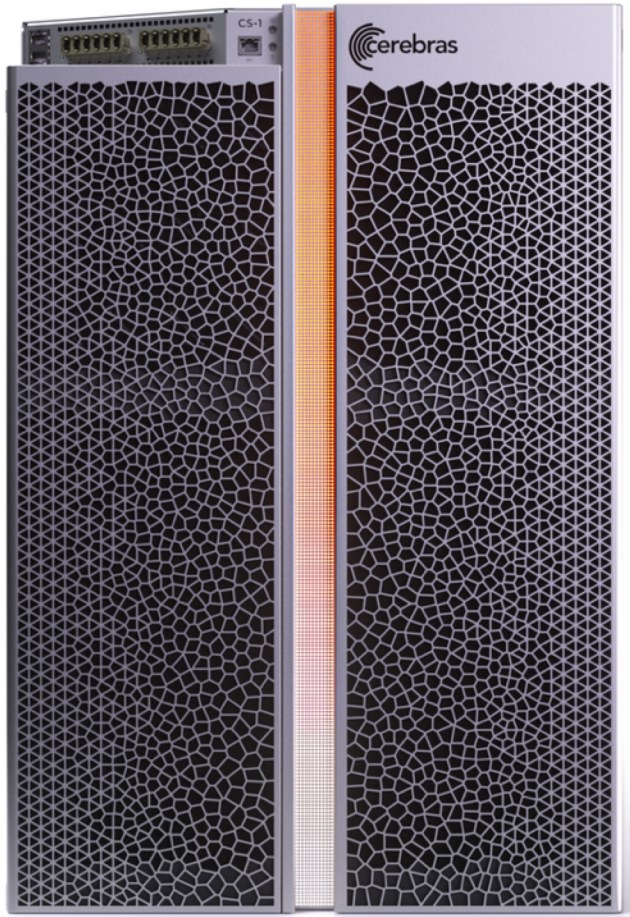
AI-optimized **memory**
- Traditional memory architectures shared memory far from compute
- The right answer is distributed, high performance, on-chip memory

AI-optimized **communication**
- High bandwidth, low latency cluster-scale networking on chip
- Fully-configurable to user-specified topology

Together, orders of magnitude performance and efficiency gain
Native model parallel execution
Full utilization at small batch, accelerated sparse compute

cerebras

An integrated approach to AI compute: system and software



The Cerebras CS-2

The world's most powerful AI computer

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

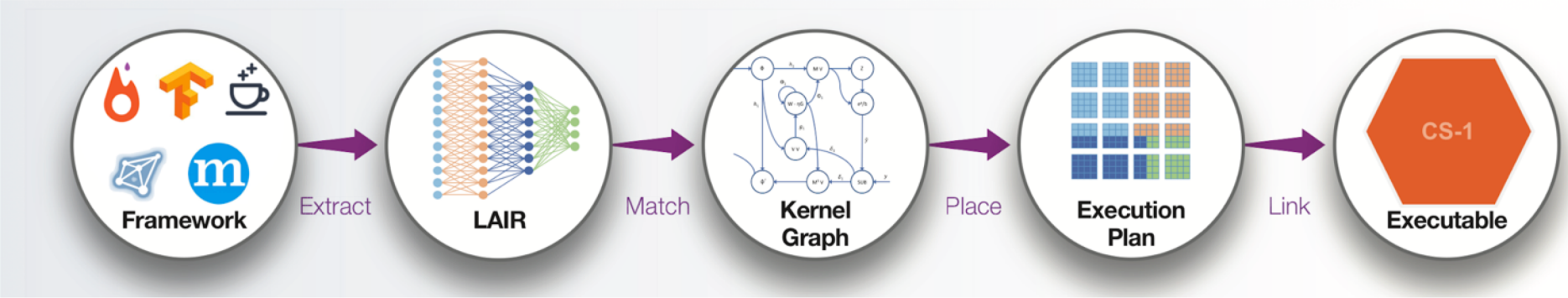
15 RU standard rack-compliant server
1.2 Tbps I/O via 12x100GbE
23 kW power

Replace aisles of legacy general purpose servers with a single system <1 rack; unlock exascale AI + HPC in a cluster

cerebras

The Cerebras Software Platform

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



→ Programmable with today's ML frameworks
→ Library of high performance DL ops
→ Customizable and extensible for other applications with Cerebras SDK

Cluster-scale AI with the programming ease of a single node

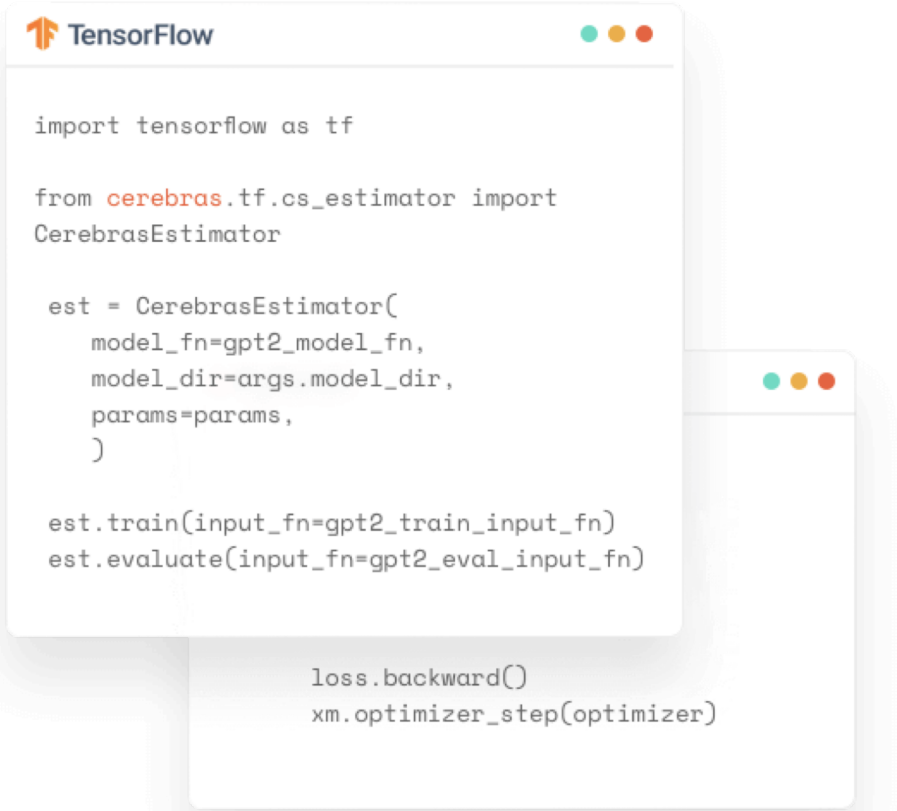
cerebras

Framework integration

Straightforward integration
with TF, PyTorch.

Train large scale models
in minutes-hours, rather than
days-weeks

With the programming simplicity
of a single machine.



Applications and the value of performance

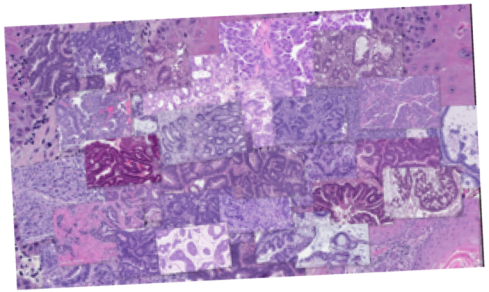
Accelerate innovation, reduce cost of curiosity

Large graph NNs and language models for drug discovery & biology research. We enable faster R&D and development of new, differentiative DNN architectures. “**Because of the sheer size of the memory we have on a single chip, we can build models we can't build as easily on other distributed architectures**." - *Kim Branson, EVP AI, GlaxoSmithKline*

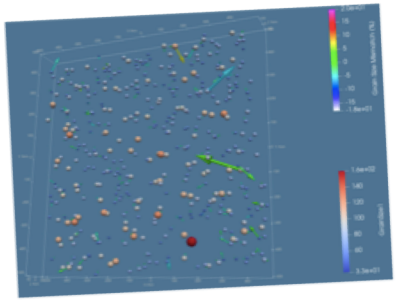
Training BERT-style model for web and social media text classification. Enables higher frequency re-training and better service at lower cost. **Accelerated BERT training time from >2 weeks on customer cluster to <2 days on single CS-1. >100x acceleration beyond GPU**.

Proud to be a part of the ALCF AI platform at Argonne

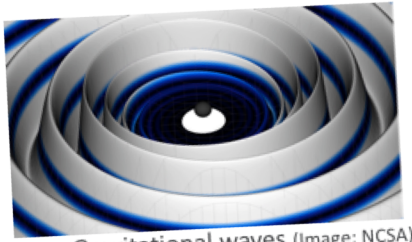
ARGONNE SCIENCE APPLICATIONS ON CEREBRAS CS-1



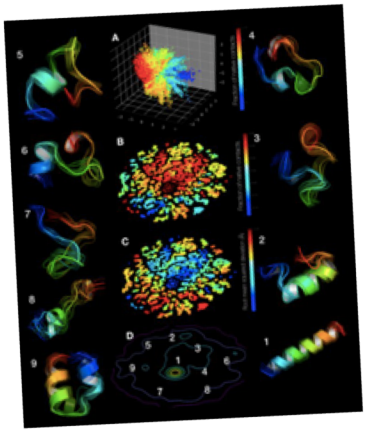
Cancer Drug response prediction



Fast X-Ray Braggs Peak Analysis



Gravitational waves (Image: NCSA)



Protein-folding
(Image: NCI)

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

cerebras

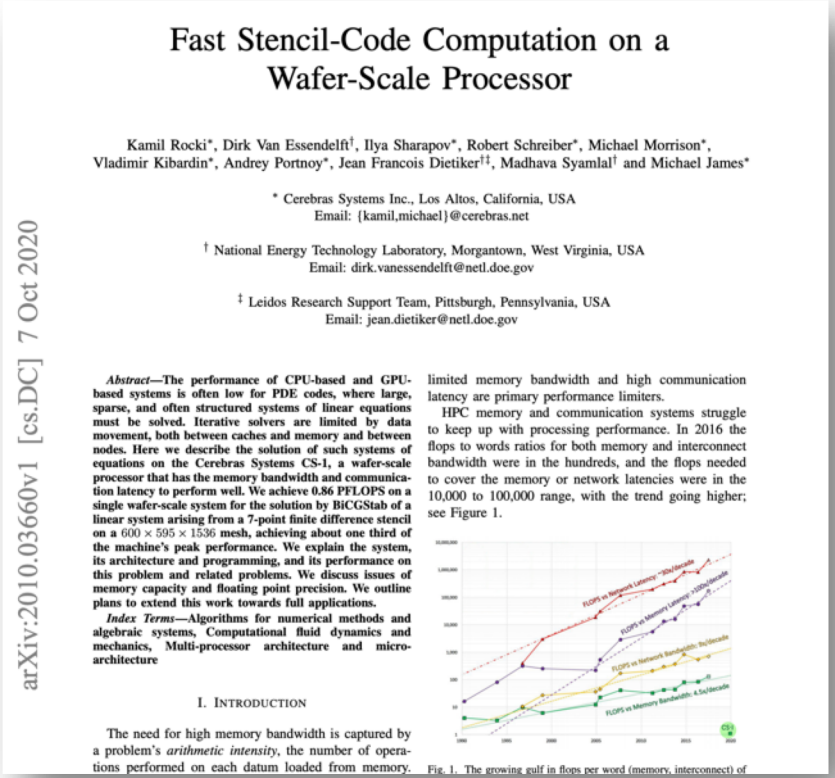
Slide: Emani et al. ANL

Supercomputer-scale HPC modeling and simulation

NETL: CS-1 used to accelerate sparse linear algebra PDE solvers for CFD.

Achieved *faster-than-real-time* simulation complex, nonlinear 3D systems: CS-1 demonstrated 10,000x GPU performance on biconjugate gradient stabilized solver for CFD on a 3D mesh.

Performance in excess of full Joule 2.0 supercomputer, in a single CS machine.



arXiv:2010.03660v1 [cs.DC] 7 Oct 2020

Fast Stencil-Code Computation on a Wafer-Scale Processor

Kamil Rocki*, Dirk Van Essendelft†, Ilya Sharapov*, Robert Schreiber*, Michael Morrison*, Vladimir Kibardin*, Andrey Portnoy*, Jean Francois Dietiker†‡, Madhava Syamlal† and Michael James*

* Cerebras Systems Inc., Los Altos, California, USA
Email: {kamil,michael}@cerebras.net

† National Energy Technology Laboratory, Morgantown, West Virginia, USA
Email: dirk.vanessendelft@netl.doe.gov

‡ Leidos Research Support Team, Pittsburgh, Pennsylvania, USA
Email: jean.dietiker@netl.doe.gov

***Abstract*—The performance of CPU-based and GPU-based systems is often low for PDE codes, where large, sparse, and often structured systems of linear equations must be solved. Iterative solvers are limited by data movement, both between caches and memory and between nodes. Here we describe the solution of such systems of equations on the Cerebras Systems CS-1, a wafer-scale processor that has the memory bandwidth and communication latency to perform well. We achieve 0.86 PFLOPS on a single wafer-scale system for the solution by BiCGStab of a linear system arising from a 7-point finite difference stencil on a $600 \times 595 \times 1536$ mesh, achieving about one third of the machine's peak performance. We explain the system, its architecture and programming, and its performance on this problem and related problems. We discuss issues of memory capacity and floating point precision. We outline plans to extend this work towards full applications.**

***Index Terms*—Algorithms for numerical methods and algebraic systems, Computational fluid dynamics and mechanics, Multi-processor architecture and micro-architecture**

I. Introduction

The need for high memory bandwidth is captured by a problem's *arithmetic intensity*, the number of operations performed on each datum loaded from memory. limited memory bandwidth and high communication latency are primary performance limiters.

HPC memory and communication systems struggle to keep up with processing performance. In 2016 the flops to words ratios for both memory and interconnect bandwidth were in the hundreds, and the flops needed to cover the memory or network latencies were in the 10,000 to 100,000 range, with the trend going higher; see Figure 1.

Fig. 1. The growing gulf in flops per word (memory, interconnect) of

AI-augmented HPC simulation

Researchers at LLNL have integrated CS-1 with the Lassen supercomputer to accelerate their work on **cognitive simulation, physics-based HPC simulations on Lassen with accelerated AI aboard CS-1**. This system provides a unique facility for AI-augmented HPC to improve simulation quality and performance with a first-in-class:

"…heterogeneous [HPC+AI] system architecture. Most supercomputers have the same node over and over…by adding the CS-1, we now have a volume that is specifically tuned and intended for running machine learning models." – Bronis de Supinski, CTO Livermore Computing at LLNL

What’s next?

Path to massive models and inputs

Our team has developed a **new execution mode** in software

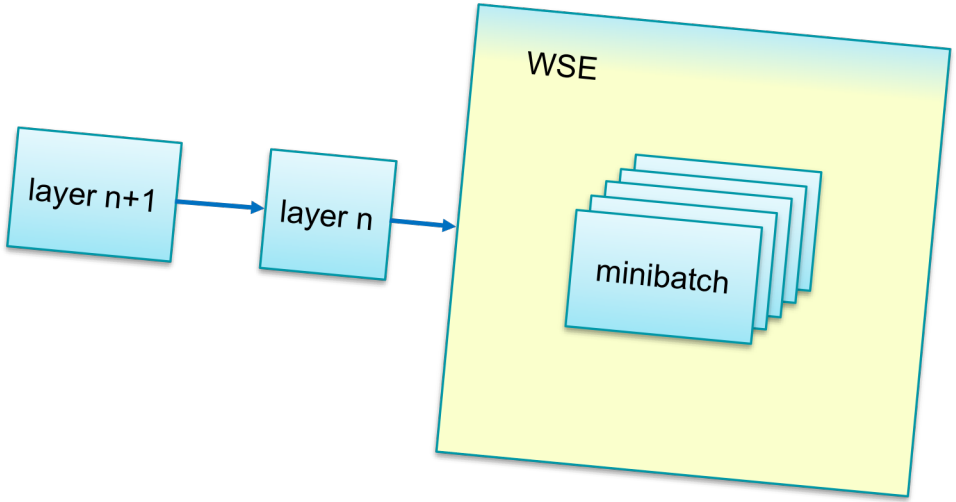
- Fully complementary to existing, *activation streaming* mode
- Load activations, *stream weights*

Enable **billion-trillion parameter-scale model training** on a single machine

Leverage **weight and activation sparsity** for greater acceleration

Throughput scaling in **compact, high performance clusters**

Stay tuned for more



cerebras

Thank you

Pleased to introduce Cerebras and CS-2 to ATPESC.

Please reach out to connect and bring us your biggest AI and HPC challenges.

We look forward to working together to unlock and accelerate exascale computing for science and industry applications.

Thank you! Welcome questions or comments.