Cerebras Systems Fact Sheet

Company Overview

Cerebras Systems, founded in 2016, is a team of more than 350 pioneering computer architects, computer scientists, deep learning researchers, functional business experts, and engineers of all disciplines. Our Silicon Valley headquarters are in Sunnyvale, CA. We also have offices in San Diego, CA; Toronto, Canada; and Tokyo, Japan. As a private company, we have raised roughly $450m from a mix of VC firms, including Benchmark, Foundation Capital, Coatue, Eclipse Capital and investors including Sam Altman (CEO & Founder, OpenAI), Andy Bechtolsheim (Founder, Sun Microsystems), Pradeep Sindhu (Founder, Juniper Networks) and Fred Weber (Former CTO and CVP, AMD).

We have come together to build a new class of computer to accelerate artificial intelligence and deep learning work by orders of magnitude beyond the current state of the art.

Recognition

- 2021 Fast Company's Most Innovative Companies
- 2020 Forbes AI 50
- 2020 IEEE Spectrum's Emerging Technology Awards
- 2020 Global Semiconductor Alliance "Startup to Watch"
- 2019 CBInsights AI 100

Core Competencies

Cerebras Systems builds the world's fastest AI accelerator, the CS-2 system. The CS-2 is based on the largest processor ever built, the Cerebras wafer-scale engine (WSE-2). Core competencies include:

- Accelerated artificial intelligence compute, orders of magnitude faster than contemporary graphics processors
- Reduced training time from days-weeks to minutes-hours; orders of magnitude faster inference in production
- Out-of-the-box support for state-of-the art language and sequence data models like BERT, Transformer and GPT for applications like classification and translation; large graph neural nets for modeling and signal processing
- Easily train massive models on large, real-world domain-specific datasets
- Faster AI research: research idea to model in production in weeks instead of months
- Accelerated sparse linear algebra computation for HPC applications (computational fluid dynamics, molecular dynamics, signal processing) by multiple orders of magnitude beyond legacy computer systems
- Research and development enabled for completely new and differentited AI & HPC capabilities
- Cluster-scale compute in a single machine, easily programmable with standard frameworks as a single node
- Easy to use, simple to deploy, power- and space-efficient AI

CEREBRAS SYSTEMS, INC
1237 E. ARQUES AVE, SUNNYVALE, CA 94085 USA
CEREBRAS.NET

Cerebras Systems Fact Sheet

Announced Customers Include

Argonne National Laboratory
Accelerating deep learning for cancer, COVID-19 research, signal processing for astrophysics and materials research.

AstraZeneca
Enabling rapid, large-scale medical research search, a critical capability for advancing drug discovery. With a CS-1 system, training which previously took 2 weeks to run on a large cluster of GPUs is accomplished in 52 hours.

Cirrascale Cloud Services
Providing the Cerebras Cloud instance with a CS-2 system in a cloud consumption model for enterprise and cloud-native startups.

EPCC
Accelerating AI-powered data science initiatives in the Edinburgh and Southeast Scotland City Region, enabling national-scale genomics research for public health initiatives.

GlaxoSmithKline
Accelerating deep learning for drug discovery, natural language processing, sequence and compound modeling.

Lawrence Livermore National Laboratory
Applying deep learning and high performance computing: integrating a Cerebras system into Lassen, the world's 17th largest supercomputer, for AI-augmented physics simulation and traumatic brain injury research.

National Energy Technology Lab
High performance computing to accelerate stencil codes for computational fluid dynamics 10,000x faster than GPU and 200x faster than an entire state-of-the-art supercomputer.

Pittsburgh Supercomputer Center
Transforming how scientists and engineers develop and test ideas for public health, medicine, healthcare, energy, and the environment.

Tokyo Electron Device
Expanding high performance AI capabilities in Asia via a Cerebras system in the TED AI Lab to dramatically reduce training time for increasingly complex AI and NLP models.

Differentiators

Performance
- The CS-2 is the fastest AI accelerator in existence, orders of magnitude faster than previous state of the art machines
- Shrinks training times from weeks to hours, and inference latency from milliseconds to microseconds

Technology
- The CS-2 system is based on the Wafer Scale Engine (WSE-2), the largest processor ever made
- The WSE is 56 times larger than the nearest competitor, with 123 times more AI optimized compute cores, 1,000 times more on chip memory and 12,733 times more memory bandwidth

Capabilities
- Enables exploration of networks that are impossible with legacy solutions
- Bigger and deeper networks; extraordinarily sparse networks; and very wide shallow networks
- Capable of supporting 1,000x more data in a training set, making using much larger datasets feasible

Ease of use
- Cluster-scale compute performance in a single machine
- Programmed easily as a single node with standard ML frameworks like TensorFlow and PyTorch
- No changes to programming paradigm, models easily imported from or exported to other hardware
- Lower level programming tools and APIs for custom kernel and application development for AI or HPC

CEREBRAS SYSTEMS, INC
1237 E. ARQUES AVE, SUNNYVALE, CA 94085 USA
CEREBRAS.NET