Cerebras Systems Fact Sheet

Company Overview

Cerebras Systems, founded in 2016, is a team of 400 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 $720M from a mix of VC firms, including Alpha Wave Ventures, Altimeter Capital, Benchmark Capital, Coatue Management, Eclipse Ventures and Foundation 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.

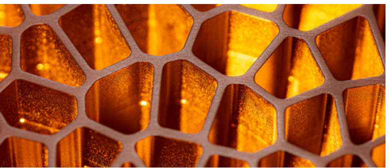
Recognition

- 2022 ACM Gordon Bell Special Prize for HPC-Based COVID-19 Research Award
- 2022 Entrepreneur of the Year, Computer Hardware — The American Business Awards
- 2022 Best Places to Work in the Bay Area – Silicon Valley Business Journal
- 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”

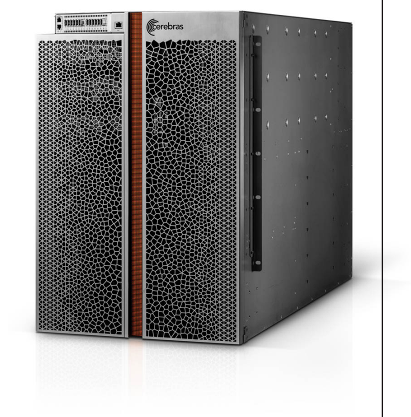
Core Competencies

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). 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; computer vision models for 2D and 3D image segmentation
- Easily train massive models up to 20 billion parameters on large, real-world domain-specific datasets using a single device, something not possible on conventional systems
- 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
- The Cerebras Wafer-Scale Cluster delivers unprecedented near-linear scaling and a remarkably simple programming model
- Easy to use, simple to deploy, power- and space-efficient AI



Cerebras Systems Fact Sheet



Announced Customers Include

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

AstraZeneca
Enabling rapid, large-scale medical research search, a critical capability for advancing drug discovery.

Cirrascale Cloud Services
Providing CS-2 systems in a cloud consumption model for enterprise and cloud-native startups.

EPCC
Accelerating AI-powered data science in Scotland, enabling national-scale genomics public health initiatives.

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

Green AI Cloud
Green AI Cloud is the first cloud computing provider in Europe to offer the industry-leading CS-2 system.

Lawrence Livermore National Laboratory
Integrating with the Lassen supercomputer for AI-augmented physics simulation and brain injury research.

Leibniz Supercomputing Centre
Enabling researchers to accelerate scientific research and innovation in AI for Bavaria.

National Energy Technology Laboratory
Accelerating stencil codes for computational fluid dynamics 200x faster than an entire supercomputer.

National Center for Supercomputing Applications
Providing researchers with a dedicated AI resource for scientific discovery.

Nference
Accelerating self-supervised language model training with longer sequence lengths of data than ever more.

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

Sandia National Laboratories
Investigating stockpile stewardship applications with the National Nuclear Security Administration.

Tokyo Electron Device
Expanding high performance AI capabilities in Asia in the TED AI Lab for increasingly complex AI and NLP models.

TotalEnergies Research & Technology USA
Enabling fast and accurate simulations from batteries to biofuels, to wind flows, drillings, and CO_2 storage.

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

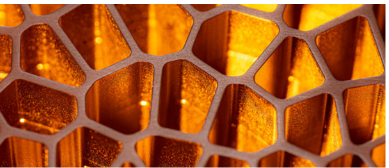
- The CS-2 system is based on the Wafer Scale Engine (WSE), the largest processor ever made
- The WSE is 56 times larger than the nearest competitor, with 123 times more AI optimized compute cores, 1,000 times more on chip memory and 12,733 times more memory bandwidth

Capabilities

- Enables exploration of networks that are impossible with legacy solutions
- Uniquely able to continously pre-train and fine-tune GPT-style language models up to 20 billion parameters on a single device
- 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 using the standard ML frameworks TensorFlow and PyTorch
- No changes to programming paradigm, models easily imported from or exported to other hardware
- No complex distributed programming expertise required
- The Cerebras Software Development Kit allows researchers to extend the platform and develop custom kernels – empowering them to push the limits of AI and HPC innovation
- The Cerebras Wafer-Scale Cluster delivers unprecedented near-linear scaling and a remarkably simple programming model



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