Cerebras CS-2

Purpose-built deep learning system delivering performance at unprecedented speeds and scale

AI Insights in Minutes, not Months

The Cerebras CS-2 system is the industry's fastest AI accelerator. It reduces training times from months to minutes, and inference latencies from milliseconds to microseconds. And the CS-2 requires only a fraction of the space and power of graphics processing unit-based AI compute.

The CS-2 features 850,000 AI optimized compute cores, 40GB of on-chip SRAM, 20 PB/s memory bandwidth and 220Pb/s interconnect, all enabled by purpose-built packaging, cooling, and power delivery. It is fed by 1.2 terabits/s of I/O across 12 100Gb Ethernet links. Every design choice has been made to accelerate deep learning, reducing training times and inference latencies by orders of magnitude.

Powered by the 2nd Generation Wafer-Scale Processor

The CS-2 is powered by the largest processor ever built — the industry's only 2.6 trillion transistor silicon device. The Cerebras Wafer-Scale Engine 2 (WSE-2) delivers more AI optimized compute cores, more fast memory, and more fabric bandwidth than any other deep learning processor in existence.

At 46,225 mm^2, WSE-2 is 56 times larger than the largest graphics processing unit. The WSE-2 contains 123x more compute cores and 1,000x more high performance on chip memory.

Seamless Software Integration

The Cerebras software platform integrates with popular machine learning frameworks like TensorFlow and PyTorch, so researchers can use familiar tools and rapidly bring their models to the CS-2.

The platform is fully programmable and provides both an extensive library of primitives for standard deep learning computations, as well as a familiar C-like interface for developing custom kernels and applications.

Unlock New Paths of ML Research

The performance and scale of the CS-2 unlocks entirely new classes of models, learning algorithms, and researcher opportunities. These include exceptionally sparse networks and very wide, shallow networks. The CS-2 provides faster time to solution, with cluster-scale resources on a single chip and with full utilization at any batch size, including batch size 1.

Datacenter Deployment

Standard install, interface, and management redundancy and carrier-grade reliability built-in

Specifications

Sparse Linear Algebra Compute Cores
850,000

On-chip Memory
40GB SRAM

Memory Bandwidth
20 PB/sec

Core-to-Core Bandwidth
220 Pb/sec

Maximum Power Requirement
23 kW

System IO
12x 100 Gb Ethernet

Cooling
Air- or water-cooled

Dimensions
15 Rack Units (26.25")

Dimensions

15 RU x 445mm x 1005mm
- Fits in standard EIA 19" 1,200 mm (47") deep rack
- Depth-adjustable support rails and brackets

300 kg (660 lbs)
- ~20kg/RU (44lb/RU)

Power

6+6 redundant 4kW power supplies

Inlets: 12x IEC 60320 C20

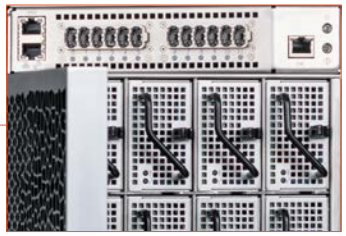
Inputs: 200-240 VAC, 16A, 50/60 Hz
- Independent single-phase inputs
- Protection: each inlet individually protected with external 16A (20A UL) circuit breaker

Network

Integrated optical multimode transceivers

12x 100GbE Data Ports (OM4 MPO/MTP-12)
- 100GBase-SR4 link
- Accepts MPO/MTP-12 fiber strand push-on cables

Use Type-B cross-over OM4 MPO/MTP-12 50/125µm multi-mode fiber patch cable to plug into industry standard 100GBase-SR4 optical module



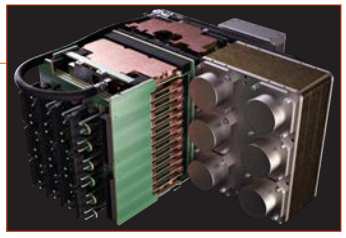
Management

1x 1GbE Management Port (RJ45)

1x Console Port (RJ45)

1x Power Management Port (RJ45)

Cooling



Internal closed-loop, direct-to-chip liquid cooling

Can be deployed with external liquid coolant loop or liquid-to-air cooling

Internal coolant loop: 1+1 redundant hot-swappable pumps

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