Cerebras Cloud @ Cirrascale

The world's fastest AI accelerator is now more accessible than ever! The Cerebras CS-2 system is designed from the ground-up for the most demanding AI training and inference needs. Cerebras Cloud provides access to a blazing fast AI solution, right at your fingertips.

High-Performance AI Compute in the Cloud

Cerebras Cloud @ Cirrascale provides access to latest Cerebras technology hosted at Cirrascale Cloud Services, a specialist in deep learning infrastructure. This joint solution delivers the wall-clock training performance of many 10s to 100s of processors in a single instance, so you can develop new neural network architectures, ML methods and algorithms that were previously impractical – all without worrying about infrastructure management.

Flexible Consumption

Utilize Cerebras Cloud instance for the work you need, when you need it. Pricing is based on weekly, monthly or annual fees; providing flexibility, predictability, and an opportunity to easily scale as you grow. If your data is stored with another cloud vendor, you can easily integrate the Cerebras Cloud with your current cloud-based workflow to create a secure, multi-cloud solution.

Software That Integrates Seamlessly

The Cerebras software platform integrates with popular machine learning frameworks like TensorFlow and PyTorch, so researchers can use familiar tools and effortlessly bring their models to the CS-2 system. No distributed training or parallel computing experience needed. The Cerebras Software Platform makes massive-scale acceleration easy to program. A programmable low-level interface allows researchers to extend the platform and develop custom kernels – empowering them to push the limits of ML innovation.

Acceleration with No Communication Bottlenecks

The WSE-2 packs 850,000 cores onto a single processor, enabling the CS-2 system to deliver cluster-scale speedup without the communication slowdowns that come from parallelizing work across a massive cluster of devices.

Real-Time Inference for Large Models

Keeping compute and memory on chip means extremely low latencies. On the CS-2 system, you can deploy large inference models in a real-time latency budget without quantizing, downsizing, and sacrificing accuracy.

Optimal Network and Hardware Utilization

The Cerebras Graph Compiler (CGC) translates your neural network to a CS-2 executable. Every stage of CGC is designed to maximize WSE-2 utilization. Kernels are intelligently sized so that more cores are allocated to more complex work. CGC then generates a placement and routing map, unique for each neural network, to minimize communication latency between adjacent layers.

CEREBRAS SYSTEMS, INC.
1237 E. ARQUES AVE, SUNNYVALE, CA 94085 USA
cerebras.net/cloud