Cerebras Cloud @ Cirrascale

The world's fastest AI accelerator is no longer limited to large national research labs and major corporations.

Democratizing High-Performance AI Compute

Now, through Cerebras Cloud @ Cirrascale, anyone can gain access to this powerful AI Accelerator. This complete solution for AI compute, powered by the world's largest chip, was co-designed with the Cerebras software platform so it's simple to program, and packaged in an innovative system so you can tackle your toughest AI workload. Today, the compute resources and expertise needed to efficiently work with large AI models such as BERT and GPT for natural language processing (NLP) are only available to a select few. Using Cerebras Cloud @ Cirrascale, your organization can train models like these in hours or days rather than weeks or months.

A Powerful Cloud Solution for AI Compute

Cerebras Cloud is powered by the groundbreaking Cerebras CS-2 system, which is designed to enable fast, flexible training and low-latency datacenter inference. Now, thanks to the partnership between Cerebras and Cirrascale Cloud Services, you can experience it in the cloud.

Featuring the 2nd generation Wafer-Scale Engine (WSE-2), the Cerebras CS-2 system has greater compute density, faster memory, and higher bandwidth interconnect than any other Artificial Intelligence solution.

Easily programmable with leading machine learning frameworks, the CS-2 system helps industry and research organizations unlock cluster-scale AI performance with the simplicity of a single device. Achieve faster time to solution with greater power and space efficiency .

Software That Integrates Seemlessly

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.

The Cerebras Cloud @ Cirrascale delivers peak performance with no large-scale cluster deployment complexity for you.

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.

Flat-Rate Billing Model

Using the Cerebras Cloud @ Cirrascale ensures no hidden fees with our flat-rate billing model. You pay one price without the worry of fluctuating bills like those at other providers.

Instance	Monthly Rate	Weekly Rate
Cerebras CS-2	$180,000	$60,000

The Cerebras Cloud @Cirrascale is available now by visiting:

https://www.cirrascale.com/cerebras

Cirrascale Cloud Services 5775 Kearny Villa Road, San Diego, CA 92123 USA **Phone** 888-942-3800 **Web** www.cirrascale.com