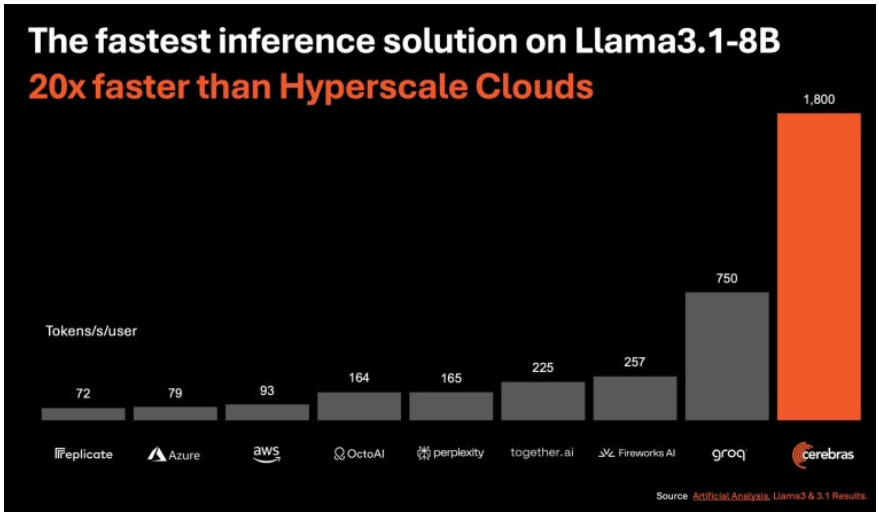
Cerebras Launches World's Fastest AI Inference Technology, 20x Performance Than NVIDIA

Cerebras Systems has just officially announced Cerebras Inference, which is considered the world's fastest AI inference solution.

This Cerebras Inference delivers performance of up to 1,800 tokens per second for Llama 3.1 8B (8 billion parameters) models and 450 tokens per second for Llama 3.1 70B, which is up to nearly 20 times faster than NVIDIA GPU-based AI inference solutions available in today's hyperscale clouds worldwide, including Microsoft Azure.

In addition to its incredible performance, the new inference solution is also very cheap to use, at a fraction of what is offered by popular GPU cloud platforms. For example, customers can get a million tokens for just 10 cents, providing a 100x price performance advantage for AI workloads.

Cerebras' 16-bit precision and 20x faster inference speed will enable developers to build next-generation high-performance AI applications without compromising on speed or cost. This breakthrough price/performance is made possible by the Cerebras CS-3 system and its Wafer Scale Engine 3 (WSE-3) AI processor. The CS-3 delivers 7,000x more memory bandwidth than the Nvidia H100, solving the technical challenge of memory bandwidth for generative AI.



Cerebras Inference is currently available at the following three levels:

1. The Free Tier offers free API access and generous usage limits to anyone who signs up.
2. The Developer Tier is designed for flexible, serverless deployments, providing users with API endpoints at a fraction of the cost of existing alternatives on the market, with the Llama 3.1 8B and 70B models priced

at just 10 cents and 60 cents per million tokens respectively.

3. The Enterprise Tier offers fine-tuned models, custom service level agreements, and dedicated support. Ideal for continuous workloads, businesses can access Cerebras Inference via a Cerebras-managed private cloud or on-premises.

With record performance, competitive pricing, and open API access, Cerebras Inference sets a new standard for open LLM development and deployment. As the only solution capable of delivering both high-speed training and inference, Cerebras opens up entirely new possibilities for AI.

With AI trends evolving rapidly and NVIDIA currently holding a dominant position in the market, the emergence of companies like Cerebras and Groq signals a potential shift in the dynamics of the entire industry. As the demand for faster and more cost-effective AI inference solutions increases, solutions like Cerebras Inference are well-positioned to take a chance on NVIDIA's dominance, especially in the inference space.

You finished reading the article "**Cerebras Launches World's Fastest AI Inference Technology, 20x Performance Than NVIDIA**" edited by the **TipsMake** team. We hope this article has provided you with many useful tech tips and tricks. You can search for similar articles on tips and guides. Thank you for reading and for following us regularly.

© 2019 TipsMake.com