Cerebras Inference Service

STEVE MCDOWELL, CHIEF ANALYST
9/9/24

CONTEXT

Cerebras Systems recently introduced *Cerebras Inference*, a high-performance AI inference service that delivers exceptional speed and affordability. The new service achieves 1,800 tokens per second for Meta's Llama 3.1 8B model and 450 tokens per second for the 70B model, which Cerebras says makes it 20 times faster than NVIDIA GPU-based alternatives.

Let's take a look at what Cerebras announced.

CEREBRAS INFERENCE

The Cerebras Inference Service is a cloud-based AI inference solution from Cerebras Systems, a direct challenger to NVIDIA-based solutions in the inference space. It is powered by the company's WSE-3 (Wafer Scale Engine 3), which Cerabras touts as the world's largest AI chip, which provides significant advantages over traditional GPU-based solutions.

Key Features of Cerebras Inference Service include:

1. **Performance**:
    - It can deliver 1,800 tokens per second for the Llama 3.1 8B model and 450 tokens per second for the Llama 3.1 70B model, up to 20x faster than GPU-based inference services.
2. **Cost Efficiency**:
    - Starting at just 10 cents per million tokens for Llama 3.1 8B and 60 cents per million tokens for Llama 3.1 70B, the service offers 100x better price-performance than traditional GPU services.
    - Its low-cost, pay-as-you-go model makes it attractive to developers and enterprises seeking scalable AI inference without prohibitive costs.

3. **High Performance Powered by WSE-3**:

   - The WSE-3 chip is built on a 5-nanometer process with over 900,000 compute cores and 44 GB of SRAM. The chip offers 52x more cores than Nvidia's H100 GPU and boasts 7,000x greater memory bandwidth, solving a key bottleneck in generative AI tasks.

4. **Tiered Access for Flexible Usage**:

   - **Free Tier**: Allows users to experiment with the platform using generous usage limits and free API access.
   - **Developer Tier**: Offers serverless deployment and API access at a fraction of the cost of competing services, ideal for developers needing scalable AI solutions.
   - **Enterprise Tier**: Targeted at businesses with sustained workloads, it includes custom models, fine-tuning, and dedicated support, available via a private cloud or on-premises deployment.

5. **Agentic AI Workloads**:

   - The service is well suited for agentic AI applications, where AI agents must frequently prompt and interact with models to perform tasks, enabling real-time, multi-step workflows.

6. **API Compatibility and Strategic Partnerships**:

   - The Cerebras Inference API is fully compatible with OpenAI's Chat Completions API, making migration from other platforms seamless, requiring minimal code changes.
   - Cerebras has formed strategic partnerships with AI development and tooling companies such as LangChain, Docker, Weights & Biases, and LlamaIndex, ensuring a rich ecosystem to support developers.

7. **Early Access Customers**:

   - Cerebras disclosed that it's already garnered interest from early adopters like GlaxoSmithKline, Perplexity AI, and DeepLearning AI, showcasing its applicability across various industries, from pharmaceuticals to AI-driven search engines.

BACKGROUND: CEREBRAS WSE-3

The Cerebras Wafer Scale Engine 3 (WSE-3) is the third-generation AI processor developed by Cerebras Systems, recognized as the world's largest semiconductor chip. It's designed specifically for AI and deep learning workloads, offering what Cerebras claims is unparalleled compute performance, memory bandwidth, and scalability.

Here's a detailed breakdown of the WSE-3:

Key Specifications:

- **Size**: The WSE-3 spans 46,225 square millimeters, covering almost the entire surface of a silicon wafer. It’s the largest single chip ever built, much larger than traditional GPUs or CPUs.
- **Cores**: It houses 850,000 AI-optimized cores, enabling massive parallelism, critical for accelerating AI workloads, particularly for LLMs like GPT or LLaMA.
- **Memory**: The WSE-3 integrates 44 GB of on-chip SRAM (a 10% increase from the previous generation), placed close to the compute cores, ensuring minimal latency for data access.
- **Memory Bandwidth**: It boasts a memory bandwidth of 21 PB/sec, allowing for high-speed data transfer between cores and memory.
- **Precision and Throughput**: The WSE-3 includes an eight-wide FP16 SIMD math unit, doubling the computational power from the previous WSE-2’s four-wide SIMD engine. This results in 1.8x more computation capability for matrix operations.

Performance Improvements:

- **2x More Performance**: Compared to its predecessor (WSE-2), the WSE-3 delivers 2x the computational performance, despite only a slight increase in memory and bandwidth. This boost is mainly due to the improved compute architecture, enhanced clock speeds, and the shift to a more advanced 5nm process node.
- **Balanced Architecture**: Unlike GPUs that rely heavily on tensor or model parallelism, the WSE-3 focuses on data parallelism alone, simplifying model scaling and minimizing the complexities in AI model training.

Scalability and Cluster Integration:

- **Cluster Size**: The WSE-3 allows the CS-3 systems (powered by WSE-3) to scale up to 2,048 systems, enabling the creation of extremely large AI supercomputers.
- **SwarmX and MemoryX**: Cerebras introduced the SwarmX interconnect for linking multiple WSE-3 processors and MemoryX technology to decouple memory from compute. MemoryX can scale independently, supporting up to 1,200 TB of memory across a cluster, enabling the processing of models with trillions of parameters.

ANALYSIS

Cerebras' new offering addresses a key challenge in the AI industry: balancing high inference speed with precision. Unlike alternative methods that sacrifice accuracy for speed, Cerebras maintains precision by staying in the 16-bit domain throughout the inference process. This makes it attractive for real-time, high-volume AI workloads, particularly in sectors that cannot afford accuracy loss.

The inference service challenges the dominant position of GPU-based AI solutions by providing a more cost-effective and performance-oriented alternative. With strategic partnerships across the AI development ecosystem—including Docker, LangChain, Weights & Biases, and LiveKit—Cerebras taps into a broad range of use cases, from real-time AI agents to multimodal applications combining voice, video, and text.

Delivering the services with competitively priced *Free*, *Developer*, and *Enterprise* tiers opens the service to a wide spectrum of users. This pricing strategy, combined with ease of integration (compatible with OpenAI's API), significantly lowers the barrier to entry for developers and enterprises seeking to leverage high-performance AI at scale.

Cerebras Inference has the potential to significantly alter the competitive landscape for AI compute, particularly in the inference segment, as enterprises and developers seek to optimize for both performance and cost in deploying next-generation AI applications.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.