AWS and Cerebras Collaboration Aims to Set a New Standard for AI Inference Speed and Performance in the Cloud

March 13, 2026

Deployed in AWS data centers and accessed through Amazon Bedrock, AWS Trainium + Cerebras CS-3 solution will accelerate inference speed

Key Takeaways

1. Fastest inference coming soon: AWS and Cerebras are partnering to deliver the fastest AI inference available through Amazon Bedrock, launching in the next couple of months.
2. Industry-leading speed and performance: With AWS Trainium optimized for prefill and Cerebras CS-3 optimized for decode, this innovative integrated system will provide unmatched performance and speed for AI inference.
3. Pioneering cloud collaboration: AWS is the first cloud provider for Cerebras's disaggregated inference solution, available soon through Amazon Bedrock.

Seattle and Sunnyvale, CA – March 13, 2026 – Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), and Cerebras Systems today announced a collaboration that will deliver the fastest AI inference solutions available for generative AI applications and LLM workloads in the coming months. The solution, to be deployed on Amazon Bedrock in AWS data centers, combines AWS Trainium-powered servers, Cerebras CS-3 systems, and Elastic Fabric Adapter (EFA) networking. Later this year, AWS will also offer leading open-source LLMs and Amazon Nova using Cerebras hardware.

“Inference is where AI delivers real value to customers, but speed remains a critical bottleneck for demanding workloads like real-time coding assistance and interactive applications,” said David Brown, Vice President, Compute & ML Services, AWS. “What we're building with Cerebras solves that: by splitting the inference workload across Trainium and CS-3, and connecting them with Amazon’s Elastic Fabric Adapter, each system does what it's best at. The result will be inference that's an order of magnitude faster and higher performance than what's available today."

“Partnering with AWS to build a disaggregated inference solution will bring the fastest inference to a global customer base,” **said Andrew Feldman, Founder and CEO of Cerebras Systems.** “Every enterprise around the world will be able to benefit from blisteringly fast inference within their existing AWS environment.”

How It Works: Inference Disaggregation

The Trainium + CS-3 solution enables “inference disaggregation,” a technique which separates AI inference into two stages: prompt processing, or “prefill,” and output generation, or “decode.” These two stages have profoundly different computational characteristics. Prefill is natively parallel, computationally intensive, and requires moderate memory bandwidth. Decode, on the other hand, is inherently serial, computationally light, and memory bandwidth intensive. Decode typically represents the majority of inference time in these scenarios because each output token must be generated sequentially.

Because each stage has a different computational challenge, they each benefit from different compute architectures and low-latency, high-bandwidth EFA networking between them. By strategically disaggregating the inference problem — with Trainium optimized for prefill and the Cerebras CS-3 optimized for decode — the two different computational challenges can be optimized in a specialized way.

Built on the AWS Nitro System — the foundation of AWS's secure, high-performance cloud infrastructure — the new solution wil ensure that Cerebras CS-3 systems and Trainium-powered instances operate with the same security, isolation, and operational consistency customers expect from AWS.

AWS Trainium for Prefill and Cerebras CS-3 for Decode

Trainium is Amazon's purpose-built AI chip, designed to deliver scalable performance and cost efficiency for training and inference across a broad range of generative AI workloads. Two of the world's leading AI labs—Anthropic and OpenAI—are committed tc Trainium. Anthropic has named AWS its primary training partner and is using Trainium to train and deploy its models, while OpenAI will consume 2 gigawatts of Trainium capacity through AWS infrastructure to support demand for Stateful Runtime Environment, frontier models, and other advanced workloads. Since its recent release, Trainium3 has seen strong customer adoption, with organizations across industries committing significant capacity.

Cerebras' CS-3 is the world's fastest AI inference system. It delivers thousands of times greater memory bandwidth than the fastest GPU. As reasoning models now represent a majority of inference to compute and generate more tokens per request as they “think” through problems, the need to accelerate this portion of the workflow has grown accordingly. OpenAI, Cognition, Mistral, and others use Cerebras to accelerate their most demanding workloads, especially agentic coding where developer productivity is constrained by inference speed.

In the disaggregated solution, CS-3 will be fully dedicated to decoding acceleration, enabling dramatically higher capacity for fast output tokens. With Trainium handling prefill, the CS-3 handling decode operations, and high-speed EFA networking connecting them, each processor will deliver maximum token capacity for its focused part of the workload.

About Amazon Web Services

Amazon Web Services (AWS) is guided by customer obsession, pace of innovation, commitment to operational excellence, and long-term thinking. By democratizing technology for nearly two decades and making cloud computing and generative AI accessible to organizations of every size and industry, AWS has built one of the fastest-growing enterprise technology businesses in history. Millions of customers trust AWS to accelerate innovation, transform their businesses, and shape the future. With the most comprehensive AI capabilities and global infrastructure footprint, AWS empowers builders to turn big ideas into reality. Learn more at aws.amazon.com and follow @AWSNewsroom.

About Cerebras Systems

Cerebras Systems builds the fastest AI infrastructure in the world. We are a team of pioneering computer architects, computer scientists, AI researchers, and engineers of all types. We have come together to make AI blisteringly fast through innovation and invention because we believe that when AI is fast it will change the world. Our flagship technology, the Wafer Scale Engine 3 (WSE-3) is the world’s largest and fastest AI processor.56 times larger than the largest GPU, the WSE uses a fraction of the power per unit compute while delivering inference and training more than 20 times faster than the competition. Leading corporations, research institutes and governments on four continents chose Cerebras to run their AI workloads. Cerebras solutions are available on premise and in the cloud, for further information, visit cerebras.ai or follow us on LinkedIn, X and/or Threads.

This press release contains forward-looking statements, including statements regarding the expected benefits of our products and the transaction described herein. These statements are subject to risks and uncertainties that could cause actual results to differ materially. Neither we nor any other person assumes responsibility for the accuracy and completeness of forward-looking statements. The forward-looking statements included in this press release relate only to events and information as of the date hereof. Cerebras undertakes no obligation to update or revise any forward-looking statement as a result of new information, future events or otherwise, except as otherwise required by law.