July 23, 2026

AMD and Cerebras Announce Industry-Leading Ultra-Low-Latency and High Throughput AI Inference Solution

News Highlights

- *AMD and Cerebras are collaborating to advance a workload-optimized approach to ultra-low-latency AI inference infrastructure.*
- *AMD Helios™ and the Cerebras Wafer-Scale Engine will operate as a single disaggregated inference workflow, combining ultra-high-throughput from AMD Instinct™ GPUs, with ultra-fast token generation of Cerebras Wafer-Scale Engine.*
- *Cerebras plans to deploy AMD Helios in its data centers, with the joint solution expected to be available first through Cerebras Cloud in the second half of 2026.*

SAN FRANCISCO and SUNNYVALE, Calif., July 23, 2026 (GLOBE NEWSWIRE) -- AMD (NASDAQ: AMD) and Cerebras Systems (NASDAQ: CBRS) announced a technical partnership to deliver a new disaggregated AI inference solution that combines AMD Helios™ rackscale solutions with the Cerebras Wafer-Scale Engine. Unveiled at Advancing AI 2026, the solution is designed to deliver the ultra-low latency required for the most advanced AI applications while dramatically increasing the throughput and efficiency.

The joint AMD and Cerebras solution will deploy AMD Helios alongside Cerebras Wafer-Scale Engine technology integrated in a single inference workflow for maximum performance and efficiency. AMD Helios will provide a high-performance, scalable throughput engine. Cerebras Wafer-Scale Engine technology will provide ultra-fast, ultra-low latency decode and token generation. Together, the two compute engines are expected to deliver up to 5x higher tokens per second per watt (T/s/W) [i].

AI inference workloads increasingly have different requirements across latency, throughput, token capacity, cost and scale. High-volume workloads prioritize maximizing token generation, while coding, real-time copilots, live agents and agentic workflows demand faster response times. These differences are driving demand for heterogeneous infrastructure that matches compute technologies to specific workload requirements.

The AMD and Cerebras solution addresses this challenge through disaggregated inference, optimizing the two primary stages of the workflow independently. AMD Helios provides ultra-high throughput, processing prompts and large context windows. The Cerebras Wafer-Scale Engine accelerates the memory-bandwidth-intensive token generation, with ultra-low latency. By connecting these best-in-class engines through one integrated workflow, the companies are creating a differentiated platform for ultra-low-latency inference without sacrificing throughput or scale.

“AI inference is becoming one of the largest infrastructure opportunities in AI, and its growing diversity requires a more flexible approach,” said Dr. Lisa Su, chair and CEO, AMD. “AMD

Helios delivers leadership performance and scale for the broadest range of inference workloads. Together with Cerebras, we are extending that leadership into the most latency-sensitive applications and creating a powerful new platform for real-time agentic AI."

"The demand for ultra-fast inference is growing at an unprecedented pace. Cerebras delivers the world's fastest, ultra-low-latency inference," said Andrew Feldman, CEO and co-founder, Cerebras. "Partnering with AMD gives us an incredible opportunity to bring that performance to even more customers."

Fast token generation is becoming increasingly important as AI moves into software development, autonomous agents, robotics, scientific discovery and other applications where response time directly shapes the user experience and the usefulness of the system. The joint solution brings together complementary architectures purpose-built for these demands.

AMD Helios provides the high-throughput prompt engine, rack-scale efficiency and deployment scale required to process large numbers of complex requests. Cerebras Wafer-Scale Engine technology provides the ultra-low-latency and decode performance needed to return tokens in real time. The result is a solution designed specifically for the ultra-low-latency segment of the inference market, with AMD Helios as the foundation for high-throughput and balanced inference workloads across the data center.

Cerebras plans to deploy AMD Helios systems in its data centers, with the joint solution expected to become available initially through Cerebras Cloud in the second half of 2026.

Supporting Resources

- Follow AMD at Advancing AI 2026 (Press Kit)
- Learn more about AMD Helios™ rackscale solution
- Learn more about AMD Instinct™ Accelerators
- Connect with AMD on LinkedIn
- Follow AMD on X
- Learn more about the Cerebras Wafer-Scale Engine
- Connect with Cerebras on LinkedIn | Follow Cerebras on X

About AMD
AMD (NASDAQ: AMD) drives innovation in high-performance and AI computing to solve the world's most important challenges. Today, AMD technology powers billions of experiences across cloud and AI infrastructure, embedded systems, AI PCs and gaming. With a broad portfolio of AI-optimized CPUs, GPUs, networking and software, AMD delivers full-stack AI solutions that provide the performance and scalability needed for a new era of intelligent computing. Learn more at www.amd.com.

About Cerebras Systems

Cerebras Systems (NASDAQ: CBRS) builds the world's fastest AI infrastructure. The Cerebras team of pioneering computer architects, computer scientists, AI researchers, and engineers of all types came together to make AI blisteringly fast through innovation and invention. We believe that when AI is fast, it will change the world. Leading global corporations, research institutes, and governments choose Cerebras to run their AI workloads. Cerebras solutions are available on premises and in the cloud.
Visit cerebras.ai for more.

AMD CAUTIONARY STATEMENT

This press release contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, scalability, deployment, timing and expected benefits of AMD's collaboration and joint solution with Cerebras, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this press release are based on current beliefs, assumptions and expectations, speak only as of the date of this press release and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and are generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Material factors that could cause actual results to differ materially from current expectations include, without limitation, the following: impact of government actions and regulations such as export regulations, import tariffs, trade protection measures, and licensing requirements; competitive markets in which AMD's products are sold; the cyclical nature of the semiconductor industry; market conditions of the industries in which AMD products are sold; AMD's ability to introduce products on a timely basis with expected features and performance levels; loss of a significant customer; economic and market uncertainty; quarterly and seasonal sales patterns; AMD's ability to adequately protect its technology or other intellectual property; unfavorable currency exchange rate fluctuations; ability of third party manufacturers to manufacture AMD's products on a timely basis in sufficient quantities and using competitive technologies; availability of essential equipment, materials, components (such as memory supply), substrates or manufacturing processes; ability to achieve expected manufacturing yields for AMD's products; AMD's ability to generate revenue from its semi-custom SoC products; potential security vulnerabilities; potential security incidents including IT outages, data loss, data breaches and cyberattacks; uncertainties involving the ordering and shipment of AMD's products; AMD's reliance on third-party intellectual property to design and introduce new products; AMD's reliance on third-party companies for design, manufacture and supply of motherboards, software, memory and other computer platform components; AMD's reliance on Microsoft and other software vendors' support to design and develop software to run on AMD's products; AMD's reliance on third-party distributors and add-in-board partners; impact of modification or interruption of AMD's internal business processes and information systems; compatibility of AMD's products with some or all industry-standard software and hardware; costs related to defective products; failure to maintain an efficient supply chain as customer demand changes; AMD's ability to rely on third party supply-chain logistics functions; AMD's ability to effectively control sales of its products on the gray market; impact of climate change on AMD's business; AMD's ability to realize its deferred tax assets; potential tax liabilities; current and future claims and litigation; impact of environmental laws, conflict minerals related provisions and other laws or regulations; evolving expectations from governments, investors, customers and other stakeholders regarding corporate responsibility matters; issues related to the responsible use of AI; restrictions imposed by agreements governing AMD's notes, the guarantees of Xilinx's notes and the revolving credit agreement; AMD's ability to satisfy financial obligations under guarantees, leases and other commercial commitments; impact of acquisitions, joint ventures and/or investments on AMD's business and AMD's ability to integrate acquired businesses; impact of any impairment of the combined company's assets; political, legal and economic risks and natural disasters; future impairments of technology license purchases;

AMD's ability to attract and retain key employees; and AMD's stock price volatility. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q.

CEREBRAS DISCLOSURE INFORMATION

Cerebras uses its investor relations page (investors.cerebras.ai), its X account (@cerebras), and its LinkedIn page (linkedin.com/company/cerebras-systems/) to disclose material non-public information and for complying with its disclosure obligations under Regulation FD. Accordingly, investors should monitor these channels, in addition to following Cerebras' press releases, Securities and Exchange Commission (SEC) filings, public conference calls and public webcasts.

Forward-Looking Statements

This press release contains "forward-looking statements" within the meaning of applicable securities laws. All statements other than statements of historical fact could be deemed to be forward-looking, including, but not limited to, statements regarding the features, capacity, scalability, performance, timing, data center deployment and implementation, costs and expected benefits and opportunities associated with Cerebras' collaboration and joint solution with AMD, and any assumptions relating to the foregoing. The words "may," "will," "shall," "should," "expects," "plans," "anticipates," "could," "intends," "target," "projects," "contemplates," "believes," "estimates," "predicts," "potential," "objective," or "continue," or the negative of these words or other similar terms or expressions that concern our expectations, strategy, plans, or intentions are intended to identify forward-looking statements, although not all forward-looking statements contain these identifying words. These forward-looking statements are subject to a number of risks and uncertainties, many of which involve factors or circumstances that are beyond Cerebras' control. These risks and uncertainties include, but are not limited to: Cerebras' ability to sustain and manage its growth, access borrowings and other sources of capital on acceptable terms, and deploy available capital to support growth; its history of net losses and ability to achieve and maintain profitability; its limited operating history at its current scale and ability to accurately forecast revenue and appropriately budget and manage expenses; its dependence on a limited number of significant customers, including OpenAI, Group 42 Holding Ltd, Mohamed bin Zayed University of Artificial Intelligence, and AWS, and the potential impact of any reduction in demand from, material adverse development in its relationships with, or failure to meet its obligations to, such customers, including under its Master Relationship Agreement with OpenAI; the timing, execution and expected benefits of its strategic customer, partner and financing arrangements; its historical reliance on sales of hardware systems and the early-stage, rapidly evolving market for its cloud-based offerings and AI infrastructure; its ability to secure sufficient data center capacity and capital to support its cloud-based offerings; its ability to launch new offerings and add new product capabilities; and its ability to compete effectively in the rapidly evolving and competitive market for AI computing solutions.

Cerebras' actual results could differ materially from those stated or implied in forward-looking statements due to a number of factors. Accordingly, undue reliance should not be placed on such statements. These forward-looking statements are made as of the date they were first issued and are based on information available to Cerebras together with Cerebras' expectations, estimates, forecasts, projections, beliefs, and assumptions as of such date. These forward-looking statements should not be relied upon as representing Cerebras'

views as of any date subsequent to the date of this press release. Past performance is not necessarily indicative of future results. Cerebras undertakes no intention or obligation to update or revise any forward-looking statements, whether as a result of new information, future events, or otherwise, except as required by law.

Further information on potential risks that could affect actual results is included in Cerebras' most recent filings with the Securities and Exchange Commission (the "SEC"), including in Cerebras' most recent Quarterly Report on Form 10-Q, copies of which may be obtained by visiting Cerebras' Investor Relations website at investors.cerebras.ai or the SEC's website at www.sec.gov.

Contacts:
Aaron Grabein
AMD Communications
737-256-9518
aaron.grabein@amd.com

Liz Stine
AMD Investor Relations
720-652-3965
liz.stine@amd.com

Kriselle Laran
Cerebras
pr@cerebras.ai

Sean Dorsey
Cerebras Investor Relations
investors@cerebras.ai

---

[i] Based on modelling by AMD Performance Labs and Cerebras in July 2026 to determine tokens per second per kilowatt (TPS/kW) at a comparable interactivity point with Kimi 2.6 1T Model comparing an AMD Helios rackscale solution with Cerebras WSE to a Cerebras WSE-only configuration. System manufacturers may vary configurations, yielding different results. MI400-021

AMD

Source: Advanced Micro Devices, Inc.