Cerebras Unveils CS-4: Up to 30 Times Faster than GPU-based Solutions

August 18, 2026

Up to 10x more throughput per watt and 2x the speed of the industry-leading CS-3

SUNNYVALE, Calif., Aug. 18, 2026 (GLOBE NEWSWIRE) -- Cerebras Systems (NASDAQ: CBRS) today introduced the Cerebras CS-4, the fastest AI accelerator in the industry. The CS-4 is a rack-scale solution built from three new Wafer Scale Engines and revolutionary rack and system designs. The CS-4 is the first member of the next-generation Cerebras Nexus rack-scale platform architecture. It is up to twice as fast as the CS-3, bringing the CS-4s advantage in tokens-per-second-per-user over GPUs to up to 30x more.

The CS-4 solution also delivers up to 10x more throughput per watt than the CS-3, vastly improving data center economics. Because fast tokens are more valuable than slow tokens, the CS-4 delivers both higher-value tokens and more total tokens within a given power budget—enabling datacenters to be vastly more profitable.

“In AI, speed is productivity,” said Andrew Feldman, CEO and co-founder of Cerebras. “Historically, fast inference meant using smaller and less capable models. Cerebras CS-4 delivers industry-leading speeds on the largest frontier models, fundamentally changing the paradigm. Every aspect of the design has been optimized to deliver the highest speeds with massive throughput. With the CS-4, AI is so fast that it fundamentally reshapes product experiences.”

The rack scale CS-4 is built from three of the newly released Wafer Scale Engine 3 Turbo (WSE-3T). The CS-4 delivers 750 PFLOPs of AI compute, 7.2 terabits per second of I/O, and 129.6 petabytes per second of memory bandwidth. Total compute fabric bandwidth jumps to 160.5 petabytes per second, and wafer to wafer latency drops as low as two microseconds, enabling the creation of very large clusters and the support of models with over 50 trillion parameters.

“CS-4 makes dramatic improvements in system deployability, reliability, and networking, which enables scaling performance to larger models for large scale token factories,” said Dylan Patel, founder and CEO, SemiAnalysis. “CS-4 will scale ultrafast tokens for larger models and significant user volumes. The Cerebras Backpack is focused on time to market. As ultrafast inference continues to grow, Cerebras’ newest hardware will help ensure frontier speeds.”

A New Leader in AI Inference Speed and Capacity

The CS-4 sets a new high watermark for inference speed. In a head-to-head comparison on GPT-OSS-120B, when given identical prompts, the CS-4 delivers[1] more than 4,400 tokens second per user (TPS/user), up to 30 times faster than GPU solutions.

“Being 30 times faster doesn’t just make a response feel fast. It gives an agentic system room for more than an order of magnitude as much reasoning, verification, or tool use in the same wall-clock time,” said Sean Lie, CTO and co-founder of Cerebras. “That’s the difference CS-4 makes for real production workloads.”

A New Processor

CS-4 is powered by the newly announced WSE-3 Turbo (WSE-3T). Like the WSE-3, the WSE-3T is the largest AI processor ever built, containing four trillion transistors and 900,000 AI-optimized cores across 46,225 square millimeters of silicon, with 44GB of SRAM integrated directly on the wafer.

The WSE-3T doubles AI compute to 250 PFLOPS per wafer and doubles memory bandwidth to 43.2 petabytes per second. And since memory bandwidth is the determining factor in driving speed and throughput, this translates to a step-change improvement in both. The on-chip fabric bandwidth and off-chip I/O both double to 53.5 petabytes per second and 2.4 terabits per second respectively. I/O latency shrinks from five microseconds to as low as two microseconds.

New System and Rack Design

CS-4 is the first iteration of the new Cerebras Nexus Platform Architecture. It is built around a modular concept with three foundational elements: Compute, Power and I/O. Cerebras brings significant innovation to each element. Modularity enables each element to scale independently so innovations get to market faster. The modular architecture also supports extremely rapid deployment and upgrades.

Pluggable Backpack Design: Cerebras has fundamentally re-imagined the compute subsystem into a rear mounted “backpack”

that attaches vertically to the power array. Each Wafer-Scale Backpack is a self-contained assembly that folds power conversion, direct liquid cooling, high-speed I/O, and control electronics into a compact, three-dimensional package built directly around the wafer. By decoupling compute from the power supplies, the Wafer-Scale Backpack simplifies manufacturing and reduces deployment time from days to hours. Compared with the prior-generation system, the Wafer-Scale Backpack has 50% fewer components and uses 60% more automated manufacturing.

High-Density Power Delivery: The Nexus Platform Architecture drives significant power delivery improvements. For example, by moving power conversion 100x closer to the processors – from roughly 50 millimeters away from the processor as on conventional GPU boards to approximately 0.5 millimeters – CS-4 nearly eliminates board-level power loss. This delivers twice as much power to the WSE-3T, enabling higher operating frequencies and faster token generation.

New Modular I/O Subsystem: CS-4 introduces a new programmable I/O subsystem that supports two connectivity modes while doubling I/O bandwidth and reducing latency**.** More bandwidth and lower latency benefits both aggregated and disaggregated solutions.

The fully programmable Wafer I/O Module supports standards-based RoCE v2 RDMA over Ethernet for seamless integration into existing infrastructure and with an ecosystem of heterogeneous systems. Aggregate off-wafer bandwidth doubles to 2.4 terabits per second per wafer and 7.2 terabits per CS-4 rack solution.

The Wafer I/O Module also supports a new communication mode, called Direct Wafer Links, which enables wafers to be linked within and across racks without a switch. Direct Wafer Links enables wafer-to-wafer latency as low as two microseconds. This low latency communication will allow the creation of massive CS-4 clusters and the ability to support models with more than 50 trillion parameters.

Programmable low latency I/O is particularly beneficial for heterogeneous disaggregated inference, in which a purpose-built prefill engine processes an incoming prompt and then hands it to Cerebras for ultra-low latency decode. The combination of programmability, standards-based interfaces, and very low latency will allow the rapid creation of disaggregated solutions from different Cerebras ecosystem partners, like AMD Helios and AWS Trainium.

CS-3 vs. CS-4: Key Specifications

Metric	**CS-3 (one wafer)**	**CS-4 (3 wafers)**
AI compute	125 PFLOPS	**750 PFLOPS**
Memory bandwidth	21.6 PByte/s	**129.6 PByte/s**
On-chip fabric bandwidth	26.7 PByte/s	**160.5 PByte/s**
System I/O bandwidth	1.2 Tbit/s	**7.2 Tbit/s**
I/O latency	5 microseconds	**2 microseconds**

Availability

First CS-4 shipments begin this quarter. Full system specifications are available in the CS-4 datasheet at https://www.cerebras.ai/cs4-datasheet.

[1] Actual throughput varies by model architecture, context length, precision, and serving configuration.

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.

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 nonpublic 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 about the anticipated benefits of CS-4 and WSE-3 Turbo, the rapid creation of disaggregated solutions, and the timing of CS-4 shipments; 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 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

Kriselle Laran
Media Relations
pr@cerebras.ai

Sean Dorsey
Investor Relations
investors@cerebras.ai