



Cerebras and Callosum Partner to Deliver Ultra-Low-Latency Heterogeneous Agentic Inference

August 20, 2026

Strategic collaboration expands Cerebras' customer footprint in Europe

SUNNYVALE, Calif. and LONDON, Aug. 20, 2026 (GLOBE NEWSWIRE) -- [Cerebras](#) (NASDAQ: CBRS) and Callosum are partnering to bring next-generation inference to market, delivering industry-leading AI compute at scale. By working with Callosum, a frontier AI and compute software company headquartered in London, Cerebras continues to expand its European presence, on the heels of last month's announcement that it is delivering [200 MW of European AI compute data center capacity](#).

Agentic inference workloads are becoming increasingly heterogeneous, with multi-agent systems operating over long-horizons in open-ended, dynamic environments. At each stage, distinct agentic reasoning and action requirements demand different computational requirements, driving the need for heterogeneous infrastructure.

The Cerebras and Callosum partnership addresses this challenge by integrating Cerebras' silicon into the Callosum software platform. Callosum's software decomposes and orchestrates heterogeneous agentic workloads, extracting maximum new levels of performance from new models and next-generation silicon. Cerebras' Wafer-Scale Engine delivers the ultra-low-latency inference that makes new categories of agentic AI possible. This collaboration gives customers direct and optimized access to frontier performance, unlocking new territories for Cerebras' silicon to be utilized where it matters most.

"Agentic AI demands infrastructure that can keep up with the speed of reasoning, not just the size of the model," said Andy Hock, Chief Strategy Officer, Cerebras. "By integrating Cerebras into Callosum's platform, we're making ultra-low-latency inference available exactly where it creates the greatest impact, enabling customers to build AI systems that simply weren't practical before. This strategic partnership amplifies Cerebras' reach across Europe, equipping a massive community of innovators to achieve the impossible."

Compute infrastructure remains the key resource for scaling AI. This partnership brings frontier silicon to market through the integration layer built for the next era of AI scaling, reflecting a shared conviction on the next paradigms of scaling compute.

"The next-generation of AI will be defined by how intelligently compute is orchestrated, not simply how much compute is available," said Danyal Akarca, CEO, Callosum. "Cerebras brings unmatched inference performance. Together we've made it easy for customers to harness that capability through the Callosum platform. This partnership gives organizations a faster path to deploying sophisticated agentic AI systems while expanding access to the world's leading AI infrastructure."

Cerebras capacity will be available through [Callosum APIs](#).

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.

About Callosum

Callosum is the Intelligent Systems Company, building the infrastructure for the next-generation of AI. We believe intelligence will not be defined by any single model or chip, but by systems that bring together specialised models, workflows and hardware to solve each problem in the way that best fits its requirements.

Callosum builds the software layer that unifies heterogeneous compute across the full stack, co-evolving models, workflows and silicon to deliver intelligence tailored to each workload. Working directly with leading silicon and infrastructure partners, we make diverse compute usable at scale - unlocking new capabilities while improving cost, speed and efficiency, without reliance on any single provider.

We're building with the teams tackling the world's hardest problems - and with the compute companies developing the silicon that will power what comes next. More at [callosum.com](#).

Contacts

Cerebras

Kriselle Laran
Media Relations
pr@cerebras.ai

Sean Dorsey
Investor Relations
investors@cerebras.ai