THE ACTIVANT *Opinion*

Speed was a niche, now it's strategic

Cerebras paves the way for the ASIC wave

June 2026

AUTHORS:

Jono Vickery

Vice President, Research

About Activant

Activant is a research-led global investment firm that partners with high-growth companies. Since 2015, we have invested in category-defining businesses during their most critical phases of growth, partnering with founders who have won their initial battles and are ready for the next challenge.

Our approach pairs deep, proprietary research with patient, flexible capital and hands-on operational partnership. We work alongside founders and leadership teams to refine strategy, strengthen operations, and accelerate sustainable growth.

Activant Research is dedicated to uncovering the most exciting emerging technologies, sectors, and companies we believe will shape the future. Our research-driven perspective informs everything we do, helping us invest at meaningful inflection points and support founders in building enduring, category-leading businesses.

You can find out more about Activant and our research at https://activantcapital.com/.

Three minutes versus six seconds. That's the difference between a 10,000-token request to Kimi K2.6 on the official endpoint (163.7 seconds) and the same request on Cerebras (5.6 seconds).[1] For coding workflows, that's the difference between a developer staying in flow and pulling out their phone to doomscroll while the agent works. Mechanically, that difference drives a step-change in developer productivity.

Cerebras's May listing was the largest semiconductor IPO on record, oversubscribed roughly 20x, priced at $185, and up 68% on day one to a peak market capitalization of ~$56B before retracing to roughly $45B.[2] While revenue is still modest, it boasts $24.6B of remaining performance obligations.[3] OpenAI signed a multi-year supply agreement that includes a roughly $1B working-capital loan flowing from OpenAI back to Cerebras to fund the build-out.[4] AWS added Cerebras to Bedrock in March.[5]

In January, we wrote that software optimizations on Nvidia hardware would own the inference market, with alternatives like Cerebras remaining niche. Reading everything above, one might say we were wrong. The reality is **that Cerebras is still fundamentally limited, but AI demand is taking more shapes than we appreciated.**

The Architecture Tax Holds

Our initial thesis, Cerebras as niche tool for coding workflows was based on two core limitations: the capex required to host a model on wafer-scale infrastructure, and the software ecosystem advantage that CUDA accrues at every new model architecture.

Six months later, both claims still hold. GPT-OSS 120B has been running on Cerebras for ~12 months and is still the most expensive provider.[6] The Codex deal sends OpenAI's distilled Spark variant to Cerebras, not its frontier Codex 5.3, which scores sixteen points higher on SWE-Bench Pro.[7] Kimi K2.6 on Cerebras sits in research preview for select enterprises with no public pricing and no reference to cost or economics in the announcement.

As we noted in the last piece, Cerebras' SRAM architecture trades memory *capacity* for memory *bandwidth*. It's the Ferrari vs Bus analogy. Unfortunately, the arrow of progress in AI is pointing towards demand for buses. Model sizes are growing with the scaling laws and KV caches increase as agents run for longer and accumulate more context. Roughly half of all coding agent requests exceed Cerebras' max context.[8] Further, datacenter operators are optimizing for system cost, where increasing the number of users on shared hardware drives down the cost per token.

Consider the case of TPU 8i, with an ability to network 1,152 chips and pool 331.8 TB of HBM into a single low-latency domain. A 1 trillion parameter model now occupies 0.3% of system memory, leaving nearly all the memory capacity for KV caches, supporting the high concurrency[9] that collapses cost per token.

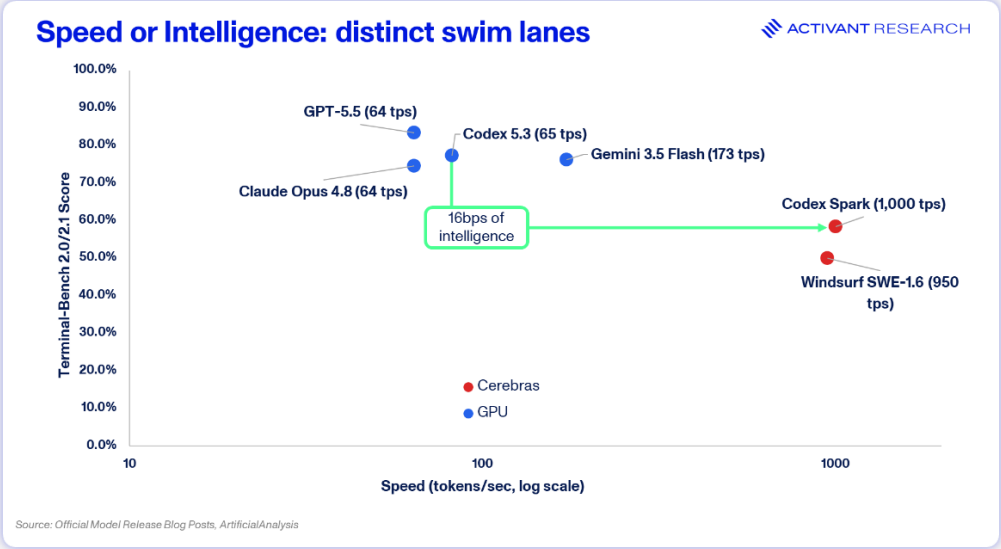
Cerebras simply can't play in the world of high concurrency, frontier intelligence. The market shifted from the capabilities of one wafer to that of the cluster. Nvidia and Google built datacenter-scale systems, with a deep focus on networking. Cerebras, and its obsession with what it could fit on one wafer, got left behind.

So, what's driving its $25bn backlog?

Compute is heterogeneous

SemiAnalysis revealed that 80% of their Claude Code spend was on "fast mode" at 6x the price and critically, their devs didn't want to switch from Opus 4.6 fast to Opus 4.7, a better model that didn't offer fast mode.[10] OpenAI and AWS have decided that their customers are willing to pay a premium for speed if it captures a defined slice of demand.

When we mapped out the **compound AI systems hypothesis** last year, we noted that mature AI systems will emerge to trade off not just cost and performance but speed too. OpenAI now runs two-tier inference openly: Cerebras for the latency-sensitive layer through Codex Spark, GPU for the reasoning-sensitive layer through Codex 5.3. Developers are coalescing - Spark's speed for simple tasks and intelligent models for what matters.



Cerebras and Groq bet on SRAM; Nvidia and Google bet on HBM + networking scale up.[11] Silicon fragmentation is creating distinct swim lanes in the inference market and a new wave of architectures will accelerate this trend. Positron AI and Majestic Labs are both redrawing the memory hierarchy for inference, each placing different bets on where the bandwidth-versus-capacity bottleneck sits. The result will be unique AI workloads that migrate to each. Nvidia itself will soon offer heterogenous compute with its LPX rack. GA in H2 2026, LPX is a purpose-built shape for the speed lane bolted onto the existing GPU stack.

Why the speed lane becomes strategic

For the software companies running on top, heterogeneous compute becomes a product surface, and it operates on three axes at once.

It is a pricing lever. Anthropic's Opus 4.6 Fast charges six times the price for what was 2.5x speed-up and is now closer to 1.75x.[12] The premium accrues directly to the application layer when the workload demands it. It is a demand-generation tool. Codex Spark is a play on capturing developers who value staying in flow. Today those use cases are limited to simple tasks, but Cerebras demonstrated a trillion parameter model on its architecture. If OpenAI can get GPT-5.5-Codex, or even a distilled version, to run at ~1,000 tokens per second; developer share will shift. It is a supply moat. OpenAI has paid for Cerebras' wafer output through 2028, keeping that speed advantage off the market until Cerebras can digest more supply.

Positron, Majestic, Etched, d-Matrix, MatX, a new wave of hardware architectures is likely to hit production in the next two years. The next purpose-built model that arrives on a new memory hierarchy is a wedge for whoever has placed the bet early. The same playbook OpenAI ran with Cerebras, paid capacity, paid construction, model co-design, is one a second-tier lab or an application-layer challenger can run against the next architecture that ships. Allocation to a new shape of silicon is the cheapest it will ever be the quarter before it goes to production.

The companies that win the next two years are the ones that pick the right silicon for each shape of demand and build the model to fit it.

[1] Cerebras, Kimi K2 Enterprise Announcement, 2026. Benchmarked by Artificial Analysis via private endpoint, May 6, 2026.

[2] CapitalIQ

[3] Cerebras, S-1 SEC Filing, 2026.

[4] SemiAnalysis, Cerebras — Faster Tokens Please, 2026. OpenAI–Cerebras 750 MW multi-year compute deal; working-capital facility extended by OpenAI to fund Cerebras build-out.

[5] Cerebras, AWS Bedrock Partnership Announcement, 2026.

[6] Artificial Analysis, GPT-OSS 120B Provider Pricing, 2026.

[7] OpenAI, Introducing GPT-5.3-Codex-Spark, 2026; Scale, SWE-Bench Pro Public Leaderboard, 2026. GPT-5.3-Codex ~72% vs. GPT-5.3-Codex-Spark ~56%.

[8] SemiAnalysis, InferenceX AgentX, 2026. n ≈ 432,000 agentic requests across Claude Code, Codex, Cursor and OpenCode; P50 ISL 96.3k tokens; ~50% of requests >128k.

[9] The number of user requests being processed on the hardware configuration at the same time

[10] SemiAnalysis, Cerebras — Faster Tokens Please, 2026.

[11] Scale-up networking connects a large number of chips into a single high-bandwidth, low-latency domain, each chip can address the full pool of memory across the system. Scale-out distributes workloads across independent nodes connected by lower-bandwidth inter-node fabric. Nvidia's NVLink and Google's ICI are scale-up fabrics; standard Ethernet/InfiniBand clusters are scale-out. For inference, scale-up enables massive KV cache pooling and high concurrency within a single job; scale-out is better suited for running many independent, smaller jobs in parallel.

[12] OpenRouter, Opus 4.6 Performance Telemetry, 2026. Pricing per Anthropic public API.

Disclaimer: The information contained herein is provided for informational purposes only and should not be construed as investment advice. The opinions, views, forecasts, performance, estimates, etc. expressed herein are subject to change without notice. Certain statements contained herein reflect the subjective views and opinions of Activant. Past performance is not indicative of future results. No representation is made that any investment will or is likely to achieve its objectives. All investments involve risk and may result in loss. This newsletter does not constitute an offer to sell or a solicitation of an offer to buy any security. Activant does not provide tax or legal advice and you are encouraged to seek the advice of a tax or legal professional regarding your individual circumstances.

This content may not under any circumstances be relied upon when making a decision to invest in any fund or investment, including those managed by Activant. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by Activant. While taken from sources believed to be reliable, Activant has not independently verified such information and makes no representations about the current or enduring accuracy of the information or its appropriateness for a given situation.

Activant does not solicit or make its services available to the public. The content provided herein may include information regarding past and/or present portfolio companies or investments managed by Activant, its affiliates and/or personnel. References to specific companies are for illustrative purposes only and do not necessarily reflect Activant investments. It should not be assumed that investments made in the future will have similar characteristics. Please see "full list of investments" at activantcapital.com/companies/ for a full list of investments. Any portfolio companies discussed herein should not be assumed to have been profitable. Certain information herein constitutes "forward-looking statements." All forward-looking statements represent only the intent and belief of Activant as of the date such statements were made. None of Activant or any of its affiliates (i) assumes any responsibility for the accuracy and completeness of any forward-looking statements or (ii) undertakes any obligation to disseminate any updates or revisions to any forward-looking statement contained herein to reflect any change in their expectation with regard thereto or any change in events, conditions or circumstances on which any such statement is based. Due to various risks and uncertainties, actual events or results may differ materially from those reflected or contemplated in such forward-looking statements.