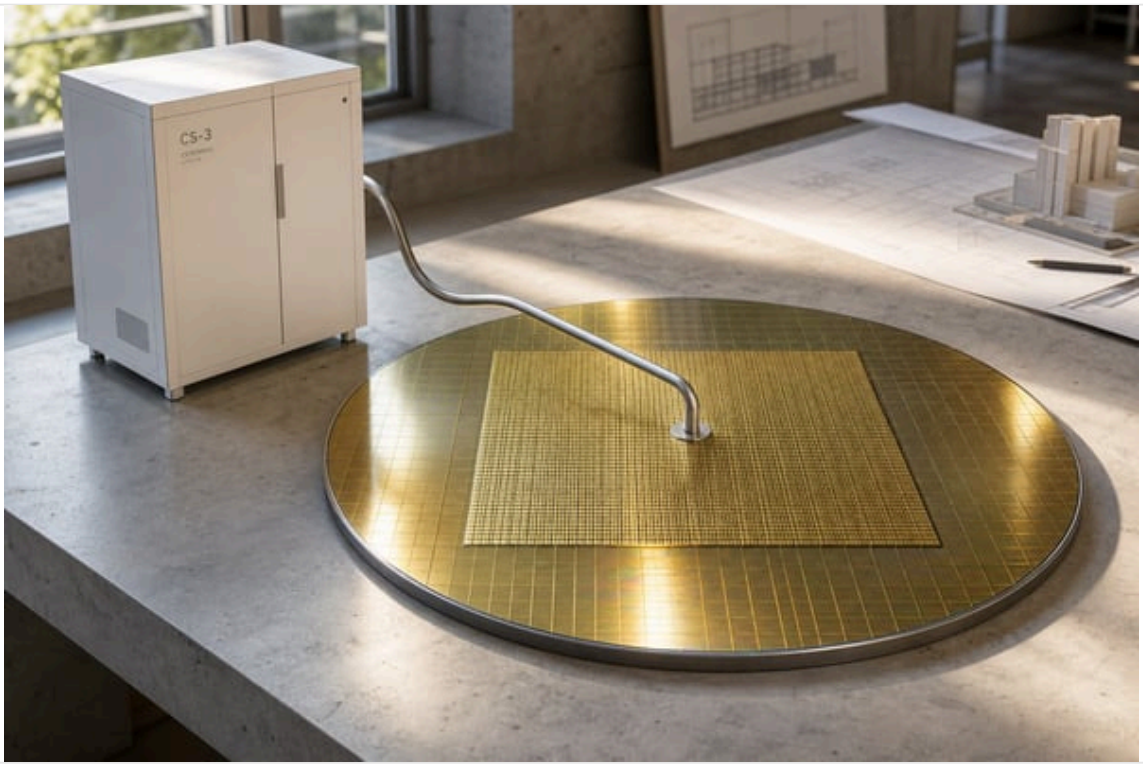
Cerebras Wafer-Scale Engine: WSE-3 Architecture Explained (2026)

Published July 10, 2026 38 min read



You've hit your session limit · resets 9am (UTC)The core engineering obstacles that killed Trilogy, chiefly manufacturing defect yield and power delivery across a much larger area than any chip of the era, are the same ones Cerebras had to solve four decades later, aided by far more mature semiconductor manufacturing and new redundancy techniques.

Cerebras's first Wafer-Scale Engine (WSE-1) shipped in 2019 on a 16nm process with 1.2 trillion transistors and 400,000 cores; the WSE-2 followed in 2021 on 7nm with 2.6 trillion transistors and 850,000 cores, and the current WSE-3 arrived in March 2024 on TSMC's 5nm process (Source: arxiv.org) (Source: arxiv.org). Each generation kept the die at essentially the same physical size, the largest square that can be cut from a circular 300mm wafer, while shrinking the transistor process node to pack in more logic. Cerebras positions the WSE and its CS-3 system as an alternative to clusters of NVIDIA GPUs for both training and, increasingly since 2024, for serving (running inference on) large language models (LLMs). This report examines how the WSE-3 architecture works, how it stacks up against NVIDIA's H100 and B200 GPUs on published specifications and independent benchmarks, what Cerebras charges for cloud inference access as of July 2026, and what four named deployments, plus the company's own May 2026 IPO, reveal about where wafer-scale computing has proven itself and where it has not.

Key Developments in the Wafer-Scale Engine

The WSE-3 Chip: Specifications and Fabrication

The WSE-3 is fabricated by TSMC on a 5nm process and measures 46,225 mm² of silicon, packing 4 trillion transistors into 900,000 identical, independently programmable AI-optimized compute cores (Source: 8968533.fs1.hubspotusercontent-na2.net). Each core carries 48 kilobytes (kB) of local SRAM and a 512-byte cache, wired for 8-way single-instruction-multiple-data (SIMD) execution on 16-bit floating point data and 16-way SIMD on 8-bit integer data, giving every core single-clock-cycle access to its own slice of the chip's aggregate 44GB of on-chip memory (Source: hc2024.hotchips.org). Because that memory sits directly beside the compute logic rather than in separate high-bandwidth memory (HBM) stacks connected by an external bus, aggregate on-chip memory bandwidth reaches 21 petabytes per second (PB/s), which Cerebras describes as 7,000

times the bandwidth of "the leading GPU" (Source: 8968533.fs1.hubspotusercontent-na2.net). Core-to-core interconnect (the "fabric") delivers a further 214 petabits per second (Pb/s) of bandwidth, over 3,715 times what Cerebras says is available between the leading GPUs' interconnect (Source: 8968533.fs1.hubspotusercontent-na2.net). Peak AI compute is rated at 125 petaflops (Source: www.cerebras.ai).

The largest engineering problem wafer-scale integration must solve is manufacturing yield: a standard die's small area means a random defect ruins only that one chip, but the WSE-3's 46,225 mm² area would, under conventional die-per-wafer economics, make even a modest defect density catastrophic. Cerebras's SEC filing describes the solution as a "fault-tolerant architecture" that treats the wafer like "a hyperscale data center but on the wafer," with redundant cores, redundant routing, and a mechanism to detect a flawed core, shut it down, and route traffic around it rather than discarding the whole chip (Source: www.sec.gov). Independent academic modeling of this approach found that because each WSE-3 core occupies only about 0.05 mm² versus roughly 6.2 mm² for an NVIDIA H100 streaming multiprocessor, a given defect wastes far less silicon area proportionally on the WSE-3, making the chip "about 164x more fault tolerant for cores than the H100" despite its far larger total area (Source: arxiv.org).

Packaging and power delivery required equally unconventional engineering. Because the chip is larger than the reticle limit that governs traditional chip packaging, Cerebras mounts the wafer directly onto a printed circuit board through a custom flexible connector that absorbs the different thermal expansion rates of silicon and PCB material, an approach the same arXiv analysis calls "packaged on board" rather than in a conventional chip package (Source: arxiv.org). Power arrives vertically from above the wafer through more than 300 distributed voltage regulation modules, rather than from the chip's edges as in conventional designs, because edge-based delivery would create unacceptable voltage drops across a die this large (Source: arxiv.org). A single CS-3 system draws roughly 23 kilowatts (kW), according to Cerebras vice president of product Andy Hock, who told Data Center Dynamics the company "didn't build it big because big was easy or big was sexy," but because doing so delivers "52 times bigger" chip area, "52 times the number of cores, more than 800x on-chip memory, and more than three orders of magnitude more memory bandwidth and communication bandwidth" than a leading GPU (Source: www.datacenterdynamics.com) (Source: www.datacenterdynamics.com).

Cerebras CS-3 System and Wafer-Scale Cluster Architecture

A single WSE-3 chip is packaged into the **CS-3**, a self-contained system roughly the size of a small refrigerator that includes its own liquid cooling loop and power supply. Cerebras describes the CS-3's "engine block" packaging as delivering power straight into the face of the wafer and providing uniform cooling via a closed internal water loop, with redundant, hot-swappable cooling and power components (Source: www.cerebras.ai). Per official deployment specifications, a Wafer-Scale Cluster rack provisions 34 kW of power, uses redundant hot-swappable power supplies in 6+6, 8+4, or 9+3 configurations, and connects each CS-3 node with 12 ports of 100 Gigabit Ethernet (Source: 8968533.fs1.hubspotusercontent-na2.net) (Source: 8968533.fs1.hubspotusercontent-na2.net).

Where the architecture becomes distinctive is in scaling beyond one chip. Cerebras's **Wafer-Scale Cluster (WSC)** adds two purpose-built components around a group of CS-3 nodes: **MemoryX**, an external memory appliance that holds model weights and streams them to the wafer, and **SwarmX**, a broadcast-and-reduce fabric that distributes those weights and aggregates gradients across every CS-3 in the cluster (Source: 8968533.fs1.hubspotusercontent-na2.net). Because MemoryX decouples memory capacity from on-chip compute, a WSC can scale from 1.2 terabytes (TB) to 1,200 TB of addressable model memory without changing the wafer itself, according to independent analysis, whereas an NVIDIA-based cluster must add more GPUs, and therefore more compute, every time it needs more memory, since the two are physically fused on each GPU package (Source: arxiv.org). At launch, Cerebras stated external memory options of 1.5TB, 12TB, or 1.2 petabytes (PB), enough, the company said, to train models up to 24 trillion parameters as a single logical device without partitioning the model across chips (Source: www.cerebras.ai), a capacity ceiling independently corroborated by arXiv's technical review, which likewise describes the WSE-3 as "designed to support large language models with up to 24 trillion parameters" (Source: arxiv.org). A WSC can scale to as many as 2,048 CS-3 systems for a claimed aggregate 256 exaflops of peak AI compute (Source: www.datacenterdynamics.com).

The practical selling point Cerebras emphasizes is programming simplicity. Because the WSC executes models with pure data parallelism, every chip runs an identical copy of the full model on a different slice of data, rather than requiring the tensor, pipeline, and expert parallelism strategies that GPU clusters use to split a model too large for one chip's memory across many devices. The company's press materials claim this cuts code required for LLM training by 97% relative to GPU clusters, including a standard GPT-3-scale implementation completed in just 565 lines of code on Cerebras, versus far more on distributed GPU frameworks (Source: www.cerebras.ai). Cerebras's own comparative blog post against NVIDIA's flagship DGX B200 concedes this scaling advantage is not absolute, grading "Reliability & Scale" and "Ease of Use" as roughly even between the two platforms given the sheer size and maturity of the NVIDIA software ecosystem, while claiming a clearer edge in raw training speed, up to 10 times faster time-to-train, from simplified scaling (Source: www.cerebras.ai).

Breakthrough AI Inference Performance

Training was Cerebras's original focus, but the company's most commercially visible move has been into AI inference, the process of running a trained model to generate outputs. Cerebras Inference launched on August 27, 2024, and immediately posted benchmark results of 1,800 output tokens per second on Meta's Llama 3.1 8B model and 450 tokens per second on the 70B variant, which the company described as 20 times faster than NVIDIA GPU-based solutions running in hyperscale clouds at the time (Source: www.cerebras.ai). Independent verification came from Artificial Analysis, a third-party AI benchmarking firm, whose co-founder and chief executive Micah Hill-Smith stated at launch that "Cerebras has taken the lead in Artificial Analysis' AI inference benchmarks... We are measuring speeds above 1,800 output tokens per second on Llama 3.1 8B, and above 446 output tokens per second on Llama 3.1 70B, a new record in these benchmarks" (Source: www.cerebras.ai). Cerebras attributes this to keeping model weights resident in on-chip SRAM rather than streaming them from external HBM for every generated token; because generating each token requires reading the entire model from memory, memory bandwidth, not raw compute, is the bottleneck for single-stream generation, and the WSE-3's 21 PB/s of on-chip bandwidth dwarfs the roughly 3.35 to 3.9 terabytes per second (TB/s) available from a single H100's HBM3 (Source: www.nvidia.com).

Performance has continued to climb. By late 2024, Cerebras reported Llama 3.1 70B inference exceeding 2,100 tokens per second, roughly a threefold increase from launch. In an August 2024 blog post, Cerebras also reported Llama 3.1 405B, the largest model in the family, running at 969 output tokens per second with a 240-millisecond (ms) time to first token, which the company called 12 times faster than the best GPU result at the time (Source: www.cerebras.ai). By 2026, Cerebras reported over 2,600 tokens per second on Meta's Llama 4 Scout, a figure the company said Artificial Analysis verified as 19 times faster than the fastest GPU solutions available at the time (Source: www.cerebras.ai), and as of July 2026 the fastest model tracked on Cerebras by Artificial Analysis, Google's Gemma 4 31B, runs at 1,990.8 tokens per second in the firm's continuous benchmarking (Source: artificialanalysis.ai). A September 2025 Cerebras blog post, benchmarked against SemiAnalysis's published methodology, put the end-to-end latency advantage over NVIDIA's flagship Blackwell B200 GPU at 21 times faster on a Llama 3 70B reasoning workload, while costing 32% less on a combined capital-and-operating-expense basis (Source: www.cerebras.ai).

Cerebras frames the strategic case for speed around emerging "agentic" AI workloads that require many sequential model calls. At the 2024 Hot Chips conference, Cerebras co-founder and Chief Technology Officer (CTO) Sean Lie cited frameworks such as Microsoft Research's GraphRAG, which can require five model calls per page and hundreds per user request, and ReAct, which can require eight calls per user turn, as illustrations of why 20 times faster single-user token speed compounds into much larger gains for multi-step reasoning and coding agents (Source: hc2024.hotchips.org). By 2026, this framing had become a business relationship: OpenAI selected Cerebras as the inference backend for its Codex-Spark coding-agent product, which the company's S-1 filing says allows users "to turn ideas into working software in seconds" (Source: www.sec.gov).

The 2026 IPO and Strategic Partnerships (OpenAI, AWS, G42)

Cerebras's path to public markets was unusually long. The company first filed confidentially for an IPO in September 2024, but withdrew that filing in October 2025, a spokesperson telling CNBC at the time the company still hoped to go public "as soon as possible," shortly after closing a private fundraising round of more than $1 billion (Source: www.cnbc.com). Reporting at the time attributed the withdrawal in part to scrutiny of the prospectus's heavy reliance on a single customer, G42, an Abu Dhabi-based AI company backed by Microsoft (Source: www.cnbc.com). Cerebras refiled in April 2026 with a materially more diversified customer picture and completed its IPO on the Nasdaq under ticker CBRS on May 14, 2026, pricing 30 million shares at $185, above its expected range, and raising $5.55 billion (Source: www.cnbc.com). Shares opened at $350 and closed the first session at $311.07, up 68%, implying a roughly $95 billion market capitalization (Source: www.cnbc.com).

Two strategic deals underpin the post-IPO growth narrative. In January 2026, Cerebras announced a multi-year deal with OpenAI valued at more than $20 billion, under which OpenAI agreed to deploy 750 megawatts (MW) of Cerebras compute capacity, with the two companies also agreeing to co-design future models for future Cerebras hardware; the deal runs through 2028 (Source: www.sec.gov) (Source: www.cnbc.com). OpenAI holds a warrant for 33,445,026 shares of Cerebras Class N common stock, exercisable at $0.00001 per share and subject to vesting conditions, according to the S-1 (Source: www.sec.gov). In March 2026, AWS signed a binding term sheet to become the first hyperscale cloud provider to deploy Cerebras systems in its own data centers; the term sheet is binding on pricing, exclusivity, and minimum capacity terms and includes a warrant for AWS to purchase up to 2,696,678 shares of Class N stock at $100.00 per share, tied to future product-purchase volume (Source: www.sec.gov) (Source: www.sec.gov).

Customer concentration nonetheless remains a defining feature of the business. Cerebras's S-1 discloses that MBZUAI accounted for 62% of total revenue in 2025, while G42 accounted for 24% of 2025 revenue, down sharply from 85% in 2024 (Source: www.sec.gov). CNBC's IPO-day reporting summarized the same shift, noting the S-1 disclosed that "24% of revenue last year came from G42, down from 85% in 2024," while MBZUAI "accounted for 62% of revenue last year" (Source: www.cnbc.com). Cerebras CEO Andrew Feldman defended the concentration on IPO day, telling CNBC, "There's some whales out there, there's some really big customers. That is one of the characteristics of this market" (Source: www.cnbc.com).

Cerebras vs NVIDIA: Comparative Architecture and Performance

The architectural contrast between wafer-scale and GPU-cluster designs is stark on paper. *Table 1* below compiles specifications for the WSE-3 (in its CS-3 system) against NVIDIA's H100 and B200 GPUs (in their DGX system configurations), drawn from official vendor datasheets and Hot Chips conference materials.

Table 1. Cerebras WSE-3/CS-3 vs NVIDIA H100/DGX B200: published specifications

SPECIFICATION	CEREBRAS WSE-3 / CS-3	NVIDIA H100 (SXM)	NVIDIA B200 / DGX B200
Chip/package size	46,225 mm² single wafer (Source: hc2024.hotchips.org)	814 mm² die (Source: www.techpowerup.com)	~1,600 mm² dual-die (Source: arxiv.org)
Process node	TSMC 5nm (Source: 8968533.fs1.hubspotusercontent-na2.net)	TSMC 4N (custom 5nm-class) (Source: www.megware.com)	TSMC 4NP (5nm-class) (Source: www.techpowerup.com)
Transistors	4 trillion (Source: www.cerebras.ai)	80 billion (Source: www.megware.com)	~208 billion (Source: www.techpowerup.com)
AI-optimized cores	900,000 (Source: 8968533.fs1.hubspotusercontent-na2.net)	16,896 CUDA cores + Tensor Cores (Source: hc2024.hotchips.org)	Not disclosed per-die
On-chip/on-package memory	44 GB SRAM (Source: 8968533.fs1.hubspotusercontent-na2.net)	80 to 94 GB HBM3 (Source: www.nvidia.com)	1,440 GB HBM3e (8-GPU DGX total) (Source: www.nvidia.com)
Memory bandwidth	21 PB/s aggregate on-chip (Source: 8968533.fs1.hubspotusercontent-na2.net)	3.35 to 3.9 TB/s per GPU (Source: www.nvidia.com)	64 TB/s aggregate (8-GPU DGX) (Source: www.nvidia.com)
Fabric/interconnect bandwidth	214 Pb/s core-to-core (Source: 8968533.fs1.hubspotusercontent-na2.net)	NVLink 900 GB/s per GPU (Source: www.nvidia.com)	NVLink 14.4 TB/s aggregate (8-GPU DGX) (Source: www.nvidia.com)
Peak AI compute	125 petaflops (Source: www.cerebras.ai)	3,958 TOPS INT8 (single GPU) (Source: www.nvidia.com)	144 PFLOPS FP4 (8-GPU DGX) (Source: www.nvidia.com)
System power draw	~23 kW (single CS-3) (Source: www.datacenterdynamics.com)	~700W max (single GPU) (Source: www.nvidia.com)	~14.3 kW max (8-GPU DGX) (Source: www.nvidia.com)
Estimated system price	~$2 to 3 million (CS-3, Cerebras est.) (Source: www.datacenterdynamics.com)	~$0.35 million (DGX H100, academic est.) (Source: arxiv.org)	~$0.5 million (DGX B200, academic est.) (Source: arxiv.org)

The comparison illustrates why the two architectures optimize for different bottlenecks. On raw peak throughput per rack, the independent arXiv analysis found the CS-3 delivers about 3.9 times the FP8 performance and 7.8 times the FP16 performance of an equally sized and equally powered H100 rack, but only 1.16 times the FP8 and 2.31 times the FP16 performance of an equivalent B200 rack, because NVIDIA's newest chip closed much of the gap (Source: arxiv.org). Once cost enters the calculation, the same study found NVIDIA's B200 delivers 1.5 to 3 times better performance-per-watt-per-dollar than the CS-3, concluding that "the higher cost of solving problems associated with wafer-scale chips" outweighs its raw throughput lead once price is factored in (Source: arxiv.org). Cerebras's own September 2025 comparison against the B200 reaches the opposite cost conclusion, reporting a 32% lower combined capital and operating cost at 21 times the inference speed, a discrepancy that appears to stem from the two analyses measuring different workloads: the arXiv paper models raw peak FLOPS per rack, while Cerebras's blog models delivered end-to-

end inference latency and throughput on specific production models (Source: www.cerebras.ai). Readers should treat both figures as directional rather than definitive: one originates from an independent but vendor-neutral academic paper using estimated system prices, the other from Cerebras marketing content citing a third-party benchmarking methodology it did not itself run.

On the multi-GPU scaling side, NVIDIA's own architecture increasingly resembles a wafer-scale system in miniature: the DGX B200 links eight Blackwell GPUs over fifth-generation NVLink for 14.4 TB/s of aggregate interconnect bandwidth and 1,440 GB of pooled HBM3e memory, delivering what NVIDIA describes as 3 times the training performance and 15 times the inference performance of the prior-generation DGX H100 (Source: www.nvidia.com). NVIDIA's own performance-per-dollar figures, drawn from SemiAnalysis's InferenceX benchmarks as of April 2026, put Blackwell-based inference at approximately $0.02 per million tokens at 55 tokens per second per user for the GPT-OSS-120B model, roughly 4.5 times cheaper than Hopper-generation systems (Source: www.nvidia.com), a figure that undercuts even Cerebras's advertised $0.35 per million input tokens for the same open-weight model as of July 2026 (Source: www.cerebras.ai), underscoring that price-per-token leadership has become genuinely contested territory between the two platforms even as Cerebras retains a latency and single-user-throughput edge.

A newer entrant complicates this two-way comparison. Groq, a chip startup founded by former Google Tensor Processing Unit (TPU) engineers, built a Language Processing Unit (LPU) architecture that, like the WSE, prioritizes on-chip SRAM over external HBM to accelerate inference latency, and had targeted $500 million in 2025 revenue amid surging demand for fast inference (Source: www.cnbc.com). In December 2025, NVIDIA agreed to acquire Groq's assets for approximately $20 billion in its largest deal on record, structured as a non-exclusive licensing agreement for Groq's inference technology under which Groq CEO Jonathan Ross and other senior leaders joined NVIDIA (Source: www.cnbc.com). NVIDIA CEO Jensen Huang wrote to employees that the company plans "to integrate Groq's low-latency processors into the NVIDIA AI factory architecture, extending the platform to serve an even broader range of AI inference and real-time workloads" (Source: www.cnbc.com), a move that effectively brings a Cerebras-adjacent, low-latency inference architecture in-house at NVIDIA and narrows the field of independent alternatives to wafer-scale computing.

Implementation Considerations: When Wafer-Scale Engines Beat GPU Clusters

Evidence collected across vendor benchmarks, independent academic analysis, and practitioner discussion converges on a fairly consistent boundary: wafer-scale computing shows its clearest advantage in **latency-bound, low-batch, single-stream inference**, and its clearest disadvantage in **raw cost-per-unit-compute at scale**, where dense GPU clusters amortize hardware cost across very high concurrent user counts.

The technical reason is memory bandwidth utilization. NVIDIA's own published data on tensor-parallel scaling shows that a Llama 70B inference workload on a single-batch query achieves roughly 60% memory bandwidth utilization on two H100s but drops to about 25% on eight H100s within a single DGX system, because cross-GPU communication overhead consumes an increasing share of available bandwidth as more devices are added (Source: hc2024.hotchips.org). Because a single WSE-3 keeps the entire model resident in on-chip SRAM and needs no cross-chip communication to run a model that fits in its 44GB, it avoids this scaling penalty entirely for models up to roughly Llama-70B scale, and Cerebras's own materials argue this is why "no number of GPUs" can match single-stream WSE-3 latency for models of that size (Source: hc2024.hotchips.org).

Practitioner sentiment on this tradeoff, sampled from a Reddit r/LocalLLaMA discussion thread, was more measured than vendor marketing. One commenter observed that Cerebras's high-throughput architecture "usually comes at the price of being locked in to architecture choices," noting the company's supported model list "is usually a pretty narrow range of models" and predicting the technology "will find a natural equilibrium" alongside general-purpose GPUs rather than replacing them outright (Source: www.reddit.com). Another user flagged the capital barrier directly: the systems are "extremely expensive (millions) and not for general computing," requiring "more upfront work to compile software with their compiler to run on their chip" compared with buying commodity GPUs usable for many other workloads (Source: www.reddit.com). A third user who had tested Cerebras for agentic coding workloads reported the "throughput is staggering," processing more than 200 prompt rounds per minute with full tool use, planning, and file edits, while judging the system "pretty useless as a vibe coder" for casual single-shot use but well suited to enterprise-scale automated agent pipelines (Source: www.reddit.com). A fourth commenter, describing training-versus-inference scaling tradeoffs, argued that because a model larger than 44GB must be split across multiple WSE-3 chips (roughly 18 chips for a DeepSeek V3-class model), the architecture can serve as many concurrent requests as an equivalent number of GPU racks, but without HBM or dynamic random-access memory (DRAM) in the loop, making the comparison less one-sided than raw single-chip specifications suggest (Source: www.reddit.com).

Drawing these threads together, a practical decision framework emerges for organizations evaluating the two approaches:

- **Choose wafer-scale (Cerebras) when**: the workload is latency-critical single-stream inference or agentic/reasoning applications requiring many sequential model calls; the model fits within a manageable number of WSE-3 chips (roughly up to 70 billion parameters per chip at FP16); the organization wants to avoid the engineering overhead of distributed GPU training code; or data sovereignty and on-premises deployment are hard requirements, as with national laboratories.

- **Choose GPU clusters (NVIDIA) when**: the workload is throughput-optimized batch inference or training serving very high concurrent user counts, where amortized cost-per-token, not single-user latency, is the primary metric; the organization needs the breadth of NVIDIA's CUDA software ecosystem, tooling, and model-support maturity; or capital budget constraints favor lower per-unit hardware cost over peak single-device performance.
- **Consider a hybrid approach when**: an organization runs both latency-sensitive customer-facing inference and large-scale batch training or fine-tuning, since Cerebras itself recommends "leveraging Nvidia where ecosystem support is essential and Cerebras where performance, cost-efficiency, and scale matter most" (Source: www.cerebras.ai).

Data Analysis and Evidence

Cerebras's financial trajectory, disclosed for the first time in detail through its SEC S-1 filing, quantifies both the scale of its recent growth and the concentration risk embedded in that growth. *Table 2* summarizes four years of revenue and profitability figures drawn directly from the filing.

Table 2. Cerebras Systems revenue and profitability, 2022 to 2025 (SEC S-1 disclosures)

FISCAL YEAR	TOTAL REVENUE	YEAR-OVER-YEAR GROWTH	NET INCOME / (LOSS)	NON-GAAP NET LOSS
2022	$24.6 million (Source: www.sec.gov)	n/a	Not separately disclosed in excerpt	n/a
2023	$78.7 million (Source: www.sec.gov)	n/a	Not separately disclosed in excerpt	n/a
2024	$290.3 million (Source: www.sec.gov)	~269% (Source: www.sec.gov)	$(481.6) million loss (Source: www.sec.gov)	$(21.8) million (Source: www.sec.gov)
2025	$510.0 million (Source: www.sec.gov)	76% (Source: www.sec.gov)	$237.8 million income (per S-1) (Source: www.sec.gov)	$(75.7) million (Source: www.sec.gov)

Table 2 shows revenue more than doubling in three consecutive years before the growth rate moderated to 76% in 2025, still an exceptional pace for a hardware company at this scale. Notably, the S-1 text and CNBC's IPO-day reporting present different 2025 net income figures: the filing states net income of $237.8 million, while CNBC's article, published the same day the company began trading, reports "net income of $88 million, swinging from a loss of $481.6 million a year earlier" (Source: www.cnbc.com). Both figures cite the same $481.6 million 2024 loss and the same $510 million 2025 revenue, so the discrepancy likely reflects a difference between a GAAP net income figure inclusive of a large, one-time $363.3 million favorable change in the fair value of a forward contract liability, which the S-1 separately discloses (Source: www.sec.gov), and an operating or adjusted figure CNBC may have used; this report presents both figures rather than resolving the discrepancy, since neither source clarifies which measure the other used. As of December 31, 2025, Cerebras carried an accumulated deficit of $905.3 million, reflecting years of pre-IPO losses even as 2025 turned profitable on at least one measure (Source: www.sec.gov).

The company's cost structure also shifted meaningfully as its cloud inference business scaled. Cost of revenue for cloud and other services rose 252% year-over-year in 2025, to $106.2 million, far outpacing the 49% growth in hardware cost of revenue, reflecting the capital intensity of building out Cerebras's own inference cloud infrastructure rather than simply shipping systems to customers (Source: www.sec.gov). Overall gross margin declined from 42.3% in 2024 to 39.0% in 2025 as this cloud buildout accelerated (Source: www.sec.gov).

Inference pricing itself offers another quantitative lens on Cerebras's competitive position. *Table 3* compiles the company's current publicly listed Developer Tier pricing as of July 2026.

Table 3. Cerebras Inference API Developer Tier pricing, July 2026

MODEL	APPROXIMATE SPEED	INPUT PRICE (PER MILLION TOKENS)	OUTPUT PRICE (PER MILLION TOKENS)
ZAI GLM 4.7 (preview)	~1,000 tokens/s	$2.25	$2.75 (Source: www.cerebras.ai)
GPT-OSS 120B	~3,000 tokens/s	$0.35	$0.75 (Source: www.cerebras.ai)
Google DeepMind Gemma 4 31B	~1,800 tokens/s	$0.99	$1.49 (Source: www.cerebras.ai)

Comparing *Table 3* to launch-era pricing shows Cerebras has repositioned since 2024: at the August 2024 Cerebras Inference launch, Llama 3.1 8B cost 10 cents per million tokens and Llama 3.1 70B cost 60 cents per million tokens on a blended basis (Source: www.cerebras.ai), while by 2026 the catalog had shifted toward larger, more capable open-weight models such as GPT-OSS-120B and GLM-4.7 with separate, higher input and output pricing tiers, and Cerebras had layered subscription-based "Cerebras Code" plans on top, at $50 per month for 24 million tokens per day of value and $200 per month for 120 million tokens per day, both of which Cerebras's own pricing page listed as sold out as of the July 2026 access date (Source: www.cerebras.ai). Artificial Analysis's independent, continuously updated tracker shows Cerebras's blended per-token pricing across its six tracked models varies up to 5.9 to 6 times, from $0.39 to $2.30 per million tokens depending on model, and confirms the company offers a fully OpenAI-compatible application programming interface (API), all six tracked models as open-weight, and support for function calling and JSON mode across the catalog (Source: artificialanalysis.ai) (Source: artificialanalysis.ai).

Cerebras's own estimate of the addressable market context is that AI inference constitutes approximately 40% of total AI hardware spending, a figure the company cited at its August 2024 inference launch to justify the strategic pivot from a training-only vendor toward a combined training-and-inference platform (Source: www.cerebras.ai). Since that pivot, stock market performance has been volatile: as of the July 10, 2026 close referenced by CNBC's live quote page, CBRS traded around $198.53, well below its $386.34 all-time high recorded on IPO day (May 14, 2026) and above its $160.81 low recorded June 26, 2026, with a market capitalization of $48.723 billion and 226.53 million shares outstanding, indicating the stock had given back roughly half its post-IPO gain within eight weeks of listing (Source: www.cnbc.com) (Source: www.cnbc.com) (Source: www.cnbc.com).

Case Studies and Real-World Examples

Mayo Clinic: Genomic Foundation Model for Rheumatoid Arthritis

Mayo Clinic, which Cerebras describes as "the top-ranked hospital in the nation," entered a multi-year strategic collaboration with Cerebras in January 2024 to develop multimodal large language models trained on decades of deidentified clinical data and genomic data from more than 100,000 consenting patients (Source: newsnetwork.mayoclinic.org). By January 2025, the partnership had produced a genomic foundation model trained on the Cerebras Wafer Scale Cluster combining public human reference genome data with Mayo's patient exome data, resulting in a 1 billion parameter model trained on 1 trillion tokens, roughly 10 times the parameter count of DeepMind's AlphaFold, initially targeted at predicting Rheumatoid Arthritis (RA) treatment response (Source: www.cerebras.ai) (Source: www.cerebras.ai). Reported accuracy on Cerebras's customer spotlight page reached 96% for cancer predisposition prediction, 83% for cardiovascular predisposition, and 87% for RA drug response prediction (Source: www.cerebras.ai) (Source: www.cerebras.ai). Mayo Clinic's Matthew Callstrom, medical director for strategy and chair of radiology, said the collaboration allowed Mayo to "develop AI tools with such promise in less than a year, in part, because of our collaboration with Cerebras that enabled us to create this state-of-the-art AI model for genomics" (Source: www.cerebras.ai).

Argonne National Laboratory: Cancer Drug Response and COVID-19 Research

Argonne National Laboratory (ANL), a U.S. Department of Energy facility, began deploying Cerebras systems in 2019 to accelerate cancer drug-response modeling under the CANDLE project, run jointly with the National Cancer Institute and National Institutes of Health (Source: 8968533.fs1.hubspotusercontent-na2.net). Before adopting Cerebras, ANL researchers relied on clusters of thousands of GPUs but faced sub-linear performance scaling, complex distributed-training engineering, and long setup times that Cerebras's own case study characterizes as a "recurring ML engineering tax" on research progress (Source: 8968533.fs1.hubspotusercontent-na2.net). Rick Stevens, ANL's Associate Laboratory Director for Computing, Environment and Life Sciences, told trade publication HPCwire that drug-response models ran "many hundreds of times faster on the CS-1 than it runs on a conventional GPU" (Source: 8968533.fs1.hubspotusercontent-na2.net), and separately stated Cerebras "allowed us to reduce the experiment turnaround time on our cancer prediction models by 300X, ultimately enabling us to explore questions that previously would have taken

years, in mere months" (Source: www.cerebras.ai). During the COVID-19 pandemic, ANL used Cerebras systems to help train convolutional variational autoencoder (CVAE) models for molecular docking-score prediction; researchers comparing training on 256 nodes of the Summit supercomputer, 1,536 GPUs total, against a single Cerebras CS-2 found the CS-2 delivered "out-of-the-box performance of 24,000 samples/s, or about the equivalent of 110-120 GPUs," work that contributed to a 2022 Association for Computing Machinery (ACM) Gordon Bell Special Prize for COVID-19 research (Source: 8968533.fs1.hubspotusercontent-na2.net). Stevens separately said Cerebras's larger memory capacity "may have the potential to transform the industry," adding that for the first time researchers "will be able to explore brain-sized models, opening up vast new avenues of research and insight" (Source: 8968533.fs1.hubspotusercontent-na2.net).

Sandia National Laboratories: Kingfisher Testbed for Nuclear Deterrence AI

In November 2024, after a two-year partnership, Sandia National Laboratories deployed the first four nodes of an eight-node Cerebras CS-3 cluster named **Kingfisher**, funded by and in support of the National Nuclear Security Administration's (NNSA's) Advanced Simulation and Computing Artificial Intelligence for Nuclear Deterrence (ASC AI4ND) strategy (Source: www.sandia.gov). Justin Newcomer, senior manager of Sandia's ASC program, explained the choice was driven by data security and infrastructure constraints unique to classified national-security work: "As part of our ASC AI4ND strategy, the Cerebras CS-3 system positions us to be able to develop large scale trusted AI models on secure internal Tri-lab (Sandia, Lawrence Livermore and Los Alamos Laboratories) data without many of the memory and power challenges that GPU systems face" (Source: www.sandia.gov). The Kingfisher deployment illustrates a use case distinct from Mayo Clinic's or Argonne's: on-premises, air-gapped, single-tenant deployment for government workloads where cloud-based GPU alternatives are constrained by security classification requirements rather than by raw throughput needs.

The May 2026 IPO and OpenAI, AWS, G42, and MBZUAI Partnerships

Cerebras's own recent history functions as a fourth case study, illustrating how a hardware vendor built around a single differentiated architecture manages commercial risk. The company's near two-year IPO journey, filing in September 2024, withdrawing in October 2025 amid criticism of customer concentration, then completing a $5.55 billion Nasdaq listing in May 2026, tracks a broader shift from a small number of sovereign AI customers (G42 and MBZUAI in the United Arab Emirates) toward a more diversified base anchored by OpenAI and AWS (Source: www.cnbc.com). Feldman, discussing the university relationship specifically, told CNBC "we're training models together," describing MBZUAI's work as building "English-Arabic models," and noted the institution "is the first university set up and dedicated to training AI practitioners" (Source: www.cnbc.com) (Source: www.cnbc.com). Founder equity and institutional ownership at IPO reflect the company's venture history: Feldman held about 5% of voting power and a stake worth close to $2 billion at the IPO price, while Fidelity controlled roughly 11% and venture firm Benchmark held about 9% (Source: www.cnbc.com). The IPO itself was led by Morgan Stanley, Citigroup, Barclays, and UBS (Source: www.cnbc.com).

Implications and Future Directions

Three trends bear watching for anyone tracking wafer-scale computing's trajectory beyond mid-2026. First, the competitive field for low-latency AI inference has consolidated rather than expanded: NVIDIA's roughly $20 billion move to license Groq's LPU technology, following a similar, smaller (over $900 million) deal to license Enfabrica's networking technology in September 2025, signals that the dominant GPU vendor now views low-latency inference architecture as core intellectual property worth acquiring rather than out-competing on price alone (Source: www.cnbc.com). This leaves Cerebras as one of the few remaining large, independent wafer-scale or SRAM-centric inference specialists not owned by or licensed to NVIDIA, a position that could prove either a durable differentiator or a strategic vulnerability depending on how aggressively NVIDIA integrates Groq's technology into its own DGX and inference stack.

Second, Cerebras's own cloud pivot is reshaping its cost structure and risk profile. The 252% year-over-year growth in cloud cost of revenue during 2025, compressing gross margin from 42.3% to 39.0%, shows the company is now carrying capital-intensive data center infrastructure on its own balance sheet rather than simply shipping systems for customers to operate, a shift toward the capital-intensive model that has strained even well-funded "neocloud" GPU providers (Source: www.sec.gov). The AWS partnership, which makes Cerebras the first third-party AI chip architecture hosted inside AWS's own data centers, may partially offset this by shifting some capital burden onto AWS in exchange for the warrant Cerebras issued at $100 per share (Source: www.sec.gov), but the arrangement also ties a material share of future revenue to a single hyperscaler relationship, precisely the concentration risk that delayed the company's original IPO attempt.

Third, the post-IPO stock trajectory, a roughly 50% retracement from the $386.34 first-day high to the $160.81 low recorded barely six weeks later, illustrates that public markets have not yet settled on a stable valuation multiple for wafer-scale computing as an asset class, distinct from the broader semiconductor rally that lifted peers such as Intel, AMD, and Micron through much of 2026 (Source: www.cnbc.com). Given that Cerebras's own

architecture bet, that memory bandwidth rather than raw compute is the binding constraint on AI inference, has been substantially validated by industry-wide interest in SRAM-centric and near-memory computing designs (evidenced by NVIDIA's Groq acquisition), the more durable open question is not whether wafer-scale integration works technically, but whether its economics can close the gap with mass-produced GPU silicon at scale, an area where independent academic modeling still gives NVIDIA's newest Blackwell generation the performance-per-dollar edge (Source: arxiv.org).

Frequently Asked Questions (FAQs)

What is the Cerebras Wafer-Scale Engine? It is a family of AI accelerator chips built from an entire silicon wafer rather than cut into individual dies. The current WSE-3 generation contains 4 trillion transistors, 900,000 cores, and 44GB of on-chip memory on 46,225 mm² of TSMC 5nm silicon (Source: 8968533.fs1.hubspotusercontent-na2.net).

What is wafer-scale integration? It is a chip manufacturing approach that connects the individual reticle-sized dies stepped across a silicon wafer while they remain physically joined, rather than cutting them apart into separate chips, treating the entire wafer as one processor. It was first attempted commercially by Trilogy Systems in the early 1980s, which failed to ship a working machine before folding in 1985, decades before Cerebras made the approach commercially viable (Source: en.wikipedia.org).

How does Cerebras compare to NVIDIA's H100 GPU? The WSE-3 is roughly 57 times larger by die area than NVIDIA's largest GPU and carries 880 times more on-chip memory and 7,000 times more on-chip memory bandwidth, according to Cerebras's own specifications (Source: 8968533.fs1.hubspotusercontent-na2.net); independent academic analysis puts the CS-3 system's rack-normalized FP8 performance at roughly 3.9 times an equivalent H100 rack (Source: arxiv.org).

What are the WSE-3 specs? 46,225 mm² of TSMC 5nm silicon, 4 trillion transistors, 900,000 AI-optimized cores, 44GB on-chip SRAM, 21 PB/s memory bandwidth, 214 Pb/s fabric bandwidth, and 125 petaflops of peak AI compute (Source: 8968533.fs1.hubspotusercontent-na2.net).

How fast is Cerebras AI inference? As of July 2026, Artificial Analysis's independent benchmark tracker measures Cerebras's fastest tracked model, Google's Gemma 4 31B, at 1,990.8 output tokens per second, with the company's Llama 4 Scout deployment separately verified at over 2,600 tokens per second, roughly 19 times faster than the fastest GPU-based solutions Artificial Analysis tested (Source: artificialanalysis.ai) (Source: www.cerebras.ai).

What is the Cerebras CS-3 system? It is the complete hardware system that houses a single WSE-3 chip, including power delivery, liquid cooling, and networking, roughly the size of a small refrigerator and drawing about 23 kW of power (Source: www.datacenterdynamics.com). Multiple CS-3 systems combine with MemoryX and SwarmX components into a Wafer-Scale Cluster (Source: 8968533.fs1.hubspotusercontent-na2.net).

When does Cerebras beat GPU clusters? Available evidence indicates wafer-scale computing wins decisively on single-stream inference latency and token throughput for models that fit within a manageable number of chips, and on training simplicity for very large models that would otherwise require complex distributed-GPU parallelism. It loses on raw performance-per-dollar at scale, where NVIDIA's B200 offers a modeled 1.5 to 3 times better performance-per-watt-per-dollar according to independent academic analysis (Source: arxiv.org).

Is Cerebras a public company? Yes. Cerebras Systems completed its Nasdaq IPO on May 14, 2026 under ticker CBRS, raising $5.55 billion at a $95 billion valuation on the first trading day (Source: www.cnbc.com).

How much does a Cerebras CS-3 system cost? The company has never officially disclosed pricing, but the system is widely estimated to cost between $2 million and $3 million, based on reporting and independent academic modeling that separately estimates roughly $2.5 million (Source: www.datacenterdynamics.com) (Source: arxiv.org).

Conclusion

The Cerebras Wafer-Scale Engine represents a genuine, commercially validated departure from decades of reticle-limited chip manufacturing, and the WSE-3's specifications, 4 trillion transistors across 46,225 mm² of silicon, delivering 125 petaflops of AI compute at 21 PB/s of on-chip memory bandwidth, are not marketing exaggeration but measured, independently corroborated facts. The architecture's core insight, that memory bandwidth rather than raw FLOPS is the binding constraint on modern AI inference, has aged well: it has been independently validated by third-party benchmarking firms, adopted as the technical premise behind Cerebras's OpenAI and AWS partnerships, and implicitly endorsed by NVIDIA's own roughly $20 billion acquisition of Groq's competing low-latency inference technology in December 2025.

Where the evidence is more mixed is economics at scale. Independent academic modeling continues to give NVIDIA's Blackwell B200 the edge in raw performance-per-dollar-per-watt, even as Cerebras's own benchmarks, using different methodology on different production workloads, claim the opposite for end-to-end inference latency. Both claims can be simultaneously true because they measure different things: peak theoretical throughput per dollar of capital versus delivered latency per dollar of total cost of ownership on specific real-world models. Organizations evaluating wafer-scale computing should treat vendor comparisons, including the ones in this report drawn from Cerebras's own blog, with the same scrutiny applied to any single-vendor benchmark, and should weigh the documented case studies at Mayo Clinic, Argonne, and Sandia, each chosen for a specific, named technical constraint, memory capacity, iteration speed, or data sovereignty, rather than for wafer-scale computing as a general-purpose GPU replacement.

Cerebras's May 2026 IPO and subsequent stock volatility underscore that the company, and by extension the wafer-scale approach it pioneered commercially, remains a high-conviction, still-unsettled bet in public markets rather than a proven default choice. Revenue growth of 76% to $510 million in 2025, a roughly $95 billion opening valuation, and landmark OpenAI and AWS partnerships all point toward genuine commercial traction. Persistent customer concentration, a nearly 50% stock retracement within six weeks of listing, and a well-capitalized new competitor in NVIDIA's licensed Groq technology point toward real, unresolved risk. As of July 2026, the most defensible conclusion is narrower than either bull or bear narratives suggest: wafer-scale computing has earned a durable, evidence-backed role for latency-critical AI inference and very-large-model training simplicity, without yet displacing GPU clusters as the default infrastructure for AI workloads generally.

Tags: cerebras wafer-scale engine, wafer-scale engine, cerebras wse-3, cerebras vs nvidia, wafer-scale integration, cerebras cs-3, cerebras inference, ai chips, nvidia h100, nvidia b200
