

DALALBYTES RESEARCH

Cerebras Systems

The Wafer-Scale Bet on the Inference Era

The largest AI chips on earth, the fastest measured inference on large open-weight models, a \$25.4 billion contracted backlog, and the most concentrated risk profile in public AI infrastructure

NASDAQ: CBRS

October 5, 2026

Reference: \$166.43, Nasdaq close of October 2, 2026.

Founding story, wafer-scale technology, the OpenAI agreement, financials, competition, management, valuation, risks, and 2029 scenarios.

01 BLUNT VERDICT FIRST

68 / 100 : Selective Positive

Cerebras builds the largest chips on earth and sells the fastest AI inference measured on large open-weight models: roughly 3 to 8 times the throughput of Groq's LPU and an order of magnitude above GPUs on long-form generation, independently benchmarked. The commercial proof arrived with the \$20 billion-plus OpenAI Master Relationship Agreement (750 megawatts of inference capacity, 2026-2028, expandable to 2 gigawatts by 2030), a \$25.4 billion remaining-performance-obligations backlog covering more than 60% of the current market capitalization, and a Q2 2026 cloud business growing 287% year over year. The May 2026 IPO raised \$5.55 billion at \$185, the largest US technology IPO of the year at the time, and the balance sheet carries no debt.

The other side is unusually concentrated risk. Two UAE-linked entities (MBZUAI at 62% and G42 at 24%) were 86% of 2025 revenue, and MBZUAI was 78% of receivables at year-end. The customer is also, via OpenAI's \$1 billion loan and 33 million-plus warrants, the lender and a potential large shareholder, with Sam Altman's early personal investment formally disclosed as a conflict. The main lockup expires November 10, 2026, insiders have sold over \$110 million, multiple securities-law investigations are open, and TSMC is the only foundry on earth that can build these chips. At roughly 50 times forward sales, the price assumes flawless execution. A 68/100 Selective Positive says: real technology, real contracted backlog, but a binary risk profile that demands position sizing, not conviction sizing.

\$166.43

Reference price
Nasdaq close, Oct 2
2026

~\$39.5B

Market cap
237.56M shares

\$209.9M

Q2 2026 core revenue
(+103% YoY)

\$25.4B

Remaining performance
obligations

\$95

2029 bear

\$290

2029 base

\$570

2029 bull

~50x

Forward sales
zero tolerance for misses

02 SNAPSHOT: WHAT CEREBRAS IS

Item	Detail
Founded	2015 by Andrew Feldman, Gary Lauterbach, Michael James, Sean Lie, Jean-Philippe Fricker (all ex-SeaMicro)
Listed	Nasdaq, May 14 2026, CBRS; IPO \$185, 30M Class A shares, \$5.55B raise; closed day one at \$311 (+68%)
Business	Wafer-scale AI accelerator systems plus Cerebras Inference cloud
Technology	WSE-3: 5nm TSMC, 4 trillion transistors, 900,000 AI cores, 44GB SRAM, ~21 PB/s bandwidth
Scale	2025 revenue \$510M (+76%); Q2 2026 core revenue \$209.9M (+103% YoY); RPO \$25.4B
Backlog	\$25.4B remaining performance obligations (Jun 30 2026), anchored by OpenAI
Balance sheet	\$5.55B IPO proceeds; no debt; current ratio 5.82; 708 employees
Key customers	OpenAI, MBZUAI, G42, AWS (Bedrock term sheet), Amazon

03 FOUNDING STORY: THE TEAM THAT REFUSED TO CUT THE WAFER

Cerebras was founded in 2015 by five engineers who had already built and sold a company together. Andrew Feldman and Gary Lauterbach started SeaMicro in 2007, sold it to AMD in 2012 for \$334 million, and Feldman then ran AMD's Data Center Server Solutions group. Reunited, they agreed on one contrarian premise: GPUs, designed for graphics and repurposed for AI, were not the optimal semiconductor for the workloads coming.

The idea they chose had been considered impractical for decades: wafer-scale integration, building the processor out of the entire silicon wafer instead of cutting it into chips. The engineering was brutal: new cooling for enormous concentrated power draw, software to route around microscopic manufacturing defects, a custom machine to drill 40 screws into a wafer without cracking it, and packaging problems nobody had faced at that scale. The company burned about \$8 million a month and spent \$200 million before a working product emerged in July 2019.

The WSE generations shipped on cadence, 2019, 2021, 2024, CS-4 in 2026. The \$200 million packaging problem would have killed a lesser team.

DalalBytes Research

The WSE-1 arrived in August 2019 (16nm, 400,000 cores, 1.2 trillion transistors), the WSE-2 in April 2021 (2.6 trillion transistors, 850,000 cores), the WSE-3 in March 2024 (5nm, 4 trillion transistors, 900,000 cores), and the CS-4 system in August 2026. First customers were national labs and pharma (GlaxoSmithKline,

AstraZeneca, Lawrence Livermore, Argonne), the classic beachhead for exotic compute. The IPO road bent twice: a September 2024 filing withdrawn in late 2025 over CFIUS scrutiny of the G42 relationship, then a refiled April 2026 offering that priced at \$185, opened near \$350, and closed at \$311. Bloomberg reported Arm and SoftBank both tried to acquire the company weeks before the IPO; Cerebras chose the public market.

04 THE TECHNOLOGY: WHY WAFER-SCALE MATTERS FOR INFERENCE

Every large language model is bottlenecked by the same physics: generating each token requires moving model weights from memory to compute, and GPUs spend most of their time waiting on that transfer. The WSE-3 keeps the entire model in 44GB of on-chip SRAM with about 21 petabytes per second of memory bandwidth, so weights never shuttle to external HBM. The result is a throughput machine.

Model	Cerebras WSE-3	Groq LPU	Multiple
GPT-OSS-120B	~1,700-3,000 tok/s*	475-493 tok/s	~3.5-6x
Llama 3.3 70B	~2,100-2,500 tok/s	~294-750 tok/s	~3-8x
Llama 3.1 8B	~1,800 tok/s	~750 tok/s	~2.4x

** High end is Cerebras's claimed figure; low end is Artificial Analysis's independent measurement. Groq low end is also the independent measurement. Feldman at RAISE Summit 2026: the fastest inference in the industry, "not by a little bit, but 10, 15, 20, 30 times faster than GPUs" (measured against GPUs).*

Two honest qualifications

First, Groq wins time-to-first-token (0.6-0.9 seconds), the metric interactive chat and voice agents live or die by; Cerebras wins sustained throughput, the metric for long-form generation, batch work, and agentic coding loops. Different races. Second, 44GB of SRAM is finite: it holds today's open-weight models at full 16-bit precision, but frontier models keep growing, which is why the disaggregation work matters.

In September 2026 Cerebras reported a 5x throughput gain from disaggregated inference: partner accelerators (AWS Trainium, AMD Helios Instinct GPUs in the published diagram) handle prompt processing while the WSE pool handles decode, lifting capacity 5x on the same wafer footprint. The AMD partnership targets production in Q4 2026 with an Amazon Bedrock offering in Q1 2027. Strategically, this turns potential competitors' silicon into Cerebras's prefill tier and widens the addressable model size beyond on-chip SRAM.

05 THE BUSINESS: SYSTEMS PLUS CLOUD, MIX SHIFTING FAST

Cerebras sells hardware systems (CS systems for on-premise and sovereign deployments) and Cerebras Inference cloud (API and dedicated capacity). In 2025 the split was \$358.4 million hardware (+69%) and \$151.6 million cloud (+99%), about 70/30. By Q2 2026 it had flipped toward cloud: core cloud and other

services revenue was \$127.7 million (+287% YoY), core hardware \$82.1 million (+17%), roughly 61/39. The mix shift is the whole story: hardware is lumpy and concentrated, cloud is recurring and scaling, and economics improve as utilization rises.

Go-to-market is widening beyond direct sales. The September 2026 General Compute agreement puts Cerebras capacity on a neocloud platform (available Q1 2027, agentic coding first). The Gimlet Labs partnership targets about 100 megawatts for Gimlet Cloud at up to 3,000 tokens per second. A multi-year Finland data center capacity agreement expands the sovereign footprint. The pattern: Cerebras provides the wafer-scale engine, partners provide capital, sites, and distribution.

06 THE OPENAI AGREEMENT: THE CONTRACT THAT REPRICED EVERYTHING

In 2026 Cerebras signed a binding Master Relationship Agreement with OpenAI worth \$20 billion-plus: 750 megawatts of low-latency inference capacity deployed 2026-2028, expandable to 2 gigawatts by 2030. Layer 1 is about \$10 billion-plus of annual compute supply over three years; a second layer added longer-term supply up to \$30 billion with a 1.25 GW expansion option. OpenAI also provided a \$1 billion loan and warrants for 33 million-plus shares, and in January 2026 Cerebras raised \$2 billion in Series H preferred ahead of the IPO.

Read it both ways

Bullish: the largest contracted endorsement of non-NVIDIA inference silicon ever signed; it underwrites the 2027 tripling management guides and converts Cerebras from hardware vendor to OpenAI's inference capacity partner. Bearish: the customer is simultaneously the lender and a potential major shareholder; Sam Altman was an early personal investor in Cerebras while serving as OpenAI's CEO, formally disclosed as a conflict in the S-1; and unconfirmed reports have questioned whether OpenAI's newest models run on Cerebras hardware. The backlog is real (\$25.4B RPO), but its largest plank is one relationship with unusual structural entanglements.

07 FINANCIALS: HYPERGROWTH ON A CASH-BURNING CHASSIS

Year	Revenue	YoY	Gross margin	Op. income	Net income
2023	\$78.7M	-	n.d.	n.d.	n.d.
2024	\$290.3M	+269%	42.3%	n.d.	~(\$500M)
2025	\$510.0M	+76%	39.0%	(\$145.9M)	\$237.8M*
Q1 2026	\$193.4M	-	n.d.	n.d.	n.d.
Q2 2026 core	\$209.9M	+103%	n.d.	neg. 265% margin	(\$0.04/share)

** 2025 GAAP net income of \$237.8M includes a \$363.3M one-time non-cash gain on extinguishment of a financial contract; non-GAAP net loss was \$75.7M. Q2 2026 GAAP revenue was \$180.1M after a \$28M customer-warrant amortization deducted from revenue; "core" revenue excludes that contra-revenue item. R&D consumed 48% of 2025 sales; capex jumped from \$23.4M (2024) to \$382.7M (2025).*

The arc is genuine hypergrowth: revenue compounded from \$78.7 million to \$510 million in two years, and Q2 2026 core revenue beat consensus by 8% with a loss 81% narrower than expected. But the chassis burns cash: the Q2 operating margin was negative 265%, and the OpenAI deployment will demand far more capital than the \$382.7 million spent in 2025. Guidance: full-year 2026 core revenue \$880-\$890 million (raised at Q2), and an ambition to more than triple revenue in 2027, implying roughly \$2.7 billion-plus. The \$25.4 billion RPO is the bridge between today's run rate and that ambition.

08 CUSTOMER CONCENTRATION: THE 86% PROBLEM

In 2025, MBZUAI was 62% of revenue and G42 was 24%: 86% from two UAE-linked entities, with the prospectus describing MBZUAI as "a related party of G42." MBZUAI was also 77.9% of accounts receivable at December 31, 2025. In 2024, G42 alone was 85% of revenue. The direction is improving (G42 fell from 85% to 24%; OpenAI and Amazon are new), but the absolute level remains extreme: lose or delay one relationship and the revenue base breaks.

This is the issue that killed the first IPO attempt, and it is only partially resolved. The CFIUS overhang is gone; the commercial concentration is not. Mitigants to watch: OpenAI revenue recognition as the 750MW deploys (Q2's 287% cloud growth is the first evidence), the AWS Bedrock term sheet converting, and any second hyperscaler or sovereign signing in 2027. Until then, concentration is the thesis risk, full stop.

09 BALANCE SHEET AND CAPITAL ALLOCATION

One of the best-capitalized pre-profitability companies in the market

Year-end 2025: \$930.4M cash and restricted cash plus \$406.5M marketable securities, no debt, current ratio 5.82, debt-to-equity 0.02. January 2026 added \$2B (Series H plus the OpenAI working capital loan); May added \$5.55B from the IPO.

Capital allocation is simple: build. No dividend, no buyback, none expected this decade. Every dollar goes to wafer starts at TSMC, data center capacity, and the OpenAI deployment. The question is not whether the fortress exists (it does, in cash) but how fast the build consumes it. The IPO proceeds are working capital for a construction program, not a cushion.

10 COMPETITION: NVIDIA'S ECOSYSTEM, GROQ'S LATENCY, HYPERSCALER SILICON

Hardware competitors per the filings: NVIDIA, AMD, Intel, Broadcom. Cloud competitors: AWS, Microsoft Azure, Google Cloud Platform, Oracle, CoreWeave. Three axes matter.

First, NVIDIA. The moat is not the chip, it is CUDA, the developer ecosystem, and supply at planetary scale. Cerebras does not beat NVIDIA on ecosystem; it beats NVIDIA's GPUs on inference throughput per the benchmarks, and on power efficiency for sustained generation. Nobody should underwrite Cerebras on beating NVIDIA broadly; the bet is a wedge in inference where latency and throughput economics favor wafer-scale.

Second, Groq and the inference specialists (SambaNova, Together, Fireworks, DeepInfra). Cerebras leads on measured throughput by 3-8x; Groq leads on time-to-first-token and has broader provider mindshare. Price-per-token compression favors the throughput leader on cost but punishes everyone on margin.

Third, custom silicon: Google TPUs, AWS Trainium, Meta MTIA, and OpenAI's own chip ambitions. The structural risk is that the largest buyers vertically integrate; the structural defense is that wafer-scale throughput is genuinely hard to replicate, and the AMD and AWS partnerships co-opt this axis by making partner silicon Cerebras's prefill tier.

11 STRENGTHS

- **The fastest measured inference** on large open-weight models: 3-8x Groq, an order of magnitude above GPUs on throughput, at full 16-bit precision.
- **A \$25.4B contracted backlog**, over 60% of market cap, anchored by the \$20B-plus OpenAI agreement.
- **Cloud revenue growing 287% YoY**, with the mix shifting from lumpy hardware to recurring inference.
- **A fortress cash position:** \$5.55B IPO raise plus \$2B Series H, no debt.
- **A founder-CEO with a prior successful exit** in the same domain and a decade committed to one architecture.
- **Strategic validation:** Arm and SoftBank reportedly tried to buy the company pre-IPO; Mizuho rates it Buy; ARK has been accumulating.
- **The disaggregation roadmap** (5x throughput gain, AMD and AWS partnerships) widens the moat beyond the wafer itself.

12 WEAKNESSES AND RISKS

- **Customer concentration:** 86% of 2025 revenue from two UAE-linked entities; MBZUAI was 78% of receivables.
- **The OpenAI entanglement:** largest customer, \$1B lender, 33M-plus warrant holder, plus the Altman conflict disclosure; unconfirmed reports of OpenAI shifting toward NVIDIA.

- **Lockup expiry November 10, 2026**, with ~19.4M shares already unlocked and heavy insider selling (\$80.8M COO, \$25.1M CTO, \$6.6M CFO, all 10b5-1 but large).
- **TSMC single-source dependency**: the only foundry that can manufacture these chips; any disruption is existential.
- **Cash burn**: negative 265% operating margin, capex exploding, profitability years away; the IPO cash funds construction, not safety.
- **Valuation**: roughly 50 times forward sales on \$885M of guided 2026 revenue; zero tolerance for execution misses.
- **Multiple securities-law investigations** opened since the IPO.
- **Ecosystem**: no CUDA equivalent; developer adoption outside the inference API is thin; hyperscaler custom silicon could obsolete the wedge.

13 MANAGEMENT AND GOVERNANCE

Andrew Feldman, co-founder and CEO: co-founded SeaMicro (sold to AMD for \$334M in 2012), ran AMD's Data Center Server Solutions group, then spent a decade in stealth building wafer-scale computing. The pillar-2 read is strong on vision and persistence and unproven on operating a public company at this scale; this is his first rodeo as a public-company CEO of a \$40 billion business.

The governance flags belong in the report, not the footnotes. The OpenAI relationship stacks customer, lender, and shareholder in one counterparty, with the CEO of that counterparty an early personal investor in Cerebras. Insider selling since the IPO exceeds \$110 million even under 10b5-1 plans. Multiple law-firm investigations into potential securities violations were announced within months of listing. None of this is disqualifying at this stage, but it is why the rating is Selective Positive rather than Positive: governance here is a monitoring item, not a settled question.

14 VALUATION VS PEERS

At \$166.43 (October 2, 2026): market cap ~\$39.5B on 2026 guided core revenue of ~\$885M, about 45x sales; about 50x on forward estimates per Street commentary. There is no P/E; the company loses money on a GAAP and non-GAAP basis.

Company	Sales multiple	Scale	Margins
Cerebras	~45-50x fwd	~\$0.9B rev guide	neg. 265% op. margin
NVIDIA	~25x	\$100B+ revenue	50%+ margins
Nebius	high teens	\$582M Q2 rev (+454%)	loss-making

Nothing is truly comparable: NVIDIA is the scaled incumbent, Nebius rents NVIDIA GPUs rather than manufacturing differentiation, CoreWeave is the closest cloud comp, Groq is private. The honest framing: Cerebras is priced like a company whose \$25.4B backlog converts at high margin, which it has not yet demonstrated. Street: 13 analysts, Moderate Buy (1 Strong Buy, 9 Buy, 2 Hold, 1 Sell), average target \$299.90 (high \$330, low \$275), about 80% upside from the reference price. Mizuho holds Buy; ARK has been buying.

The valuation debate in one line

The backlog (\$25.4B) says the market cap (\$39.5B) is 1.6x contracted revenue, which looks reasonable; the P&L (negative 265% operating margin, 86% customer concentration) says the backlog is one relationship away from a restatement, which looks terrifying. Both are true.

15 2029 SCENARIOS (USD, CALENDAR)

Built off 2026 guided core revenue of ~\$885M and ~295M fully diluted shares by 2029 (up from 237.56M, allowing for OpenAI warrants, employee grants, and a possible raise). Fiscal year is calendar. All three scenarios assume the OpenAI 750MW deploys; they differ on diversification and margin.

	Bear: \$95	Base: \$290	Bull: \$570
2029 revenue	\$4.0B	\$8.5B	\$14.0B
Revenue path	OpenAI partial; no 2nd customer	2027 ~3x, then ~85%, ~70%	OpenAI to 2GW; 2 more hyperscalers
Gross margin	~40%	~50%	~55%
Op. margin	~0%	low teens	~20%
Exit P/S	7x	10x	12x

Bear: diversification stalls, hardware stays lumpy, growth-stock de-rating on concentration. **Base:** 2027 triples per guidance, AWS Bedrock and one more hyperscaler or sovereign convert, cloud mix drives margins positive; 10x sales for 70%-plus growth with a real moat, discounted for concentration. **Bull:** inference demand compounds at the high end, disaggregation opens the prefill market, category-defining franchise at 12x sales. The current analyst average target of \$299.90 sits just above base on a nearer horizon: the Street is already underwriting the 2027 tripling.

16 WHY IT MATTERS

Cerebras is the purest public bet that AI's next bottleneck is inference, not training, and that the fix is architectural rather than incremental. If wafer-scale throughput compounds into the standard for serving large models, the \$25.4 billion backlog is the beginning of a franchise and today's 50x sales will look like 2020 NVIDIA pricing in retrospect. If OpenAI stumbles, diversifies to NVIDIA, or the concentration unwinds disorderly, the same backlog reprices violently and the lockup expiry is the catalyst. At 68/100 Selective Positive, the verdict is: own the technology thesis, respect the binary, size accordingly, and watch three things, OpenAI deployment milestones, the second large customer, and gross margin on the cloud mix, before upgrading conviction.

17 SOURCES

Company S-1/prospectus and IPO pricing release (May 2026): 2023-2025 financials, customer concentration, RPO \$24.6B, OpenAI terms, Altman conflict disclosure; via TechCrunch, Motley Fool, BBAE, Tech Times, TECHi. Q2 2026 earnings (Aug 12, 2026): core revenue \$209.87M (+103%), cloud \$127.7M (+287%), hardware \$82.1M (+17%), FY2026 guide \$880-890M, >3x 2027 ambition, RPO \$25.4B; via Zacks and company materials. Q1 2026 revenue \$193.4M via Financial Modeling Prep. Wikipedia: Cerebras Systems, Andrew Feldman. TechCrunch Disrupt 2026 (Sep 30, 2026). Artificial Analysis benchmarks via Spheron, Unite.ai, Cerebras blog. Market data: Finnhub and MarketBeat (October 2026). News: General Compute (Sep 29, 2026), Finland data center, Gimlet Labs, Mizuho/ARK, insider sales, securities investigations. Full citations with URLs and dates in the research memos filed with this report.