



Disclaimer

By downloading from or viewing material on this website you agree to the following Terms of Service. Use of Culper Research's ("Culper") research is at your own risk. In no event should Culper or any affiliated party be liable for any direct or indirect trading losses caused by any information on this site. You further agree to do your own research and due diligence, consult your own financial, legal, and tax advisors before making any investment decision with respect to transacting in any securities covered herein. You should assume that Culper (possibly along with or through our members, partners, affiliates, employees, and/or consultants) along with our clients and/or investors has a position in any securities covered herein. You should assume that following publication of any research, we are likely to immediately transact in these securities (by increasing and/or decreasing positions and/or risk exposure), continue to trade in these securities for an undefined time period thereafter, and be long, short, or neutral at any time hereafter regardless of our initial recommendation, conclusions, or opinions. We trade securities in conjunction with risk tolerances and management practices, and such trading may result in the derisking of some or all of the positions in the securities covered herein, at any time following publication of any report, depending on security-specific, market, portfolio, or other relevant conditions. Culper does not undertake to update or supplement this report to reflect changes in its position. Research may contain estimated fair values of securities, utilizing various valuation methods. Estimated fair values are not price targets, and Culper does not commit to hold securities until such time as the estimated fair values are reached. Culper may change its estimates of fair values at any time hereafter without updating its research or otherwise disclosing updated fair values publicly. We may transact in the covered securities for various reasons, none of which may relate to Culper's estimates of fair value. Research is not investment advice nor a recommendation or solicitation to buy securities. To the best of our ability and belief, all information contained herein is accurate and reliable, and has been obtained from public sources we believe to be accurate and reliable, and who are not insiders or connected persons of the securities covered herein or who may otherwise owe any fiduciary duty or duty of confidentiality to the issuer. However, such information is presented "as is," without warranty of any kind – whether express or implied. Culper makes no representation, express or implied, as to the accuracy, timeliness, or completeness of any such information or with regard to the results to be obtained from its use. Certain statements and opinions in this report are based, in part, on information provided by individuals who requested anonymity due to concerns about professional or personal repercussions. These individuals include former executives, employees, contractors, and other persons with direct knowledge of the matters discussed. We believe these sources to be credible based on their prior positions, involvement, and demonstrated familiarity with the matters discussed, as well as corroboration among multiple independent accounts and public records. In all cases, the opinions expressed herein reflect our own assessment of the credibility and reliability of the information received. The inclusion of such information is intended to inform readers of our reasoning and analysis, and does not purport to assert undisclosed facts. All of the opinions in this report are those of the authors, formed from a combination of publicly available information and corroborated source testimony. Research may contain forward-looking statements, estimates, projections, and opinions with respect to among other things, certain accounting, legal, and regulatory issues the issuer faces and the potential impact of those issues on its future business, financial condition and results of operations, as well as more generally, the issuer's anticipated operating performance, access to capital markets, market conditions, assets and liabilities. Such statements, estimates, projections and opinions may prove to be substantially inaccurate and are inherently subject to significant risks and uncertainties beyond Culper's control. All expressions of opinion are subject to change without notice, and Culper does not undertake to update or supplement this report or any of the information contained herein. You agree that the information on this website is copyrighted, and you therefore agree not to distribute this information (whether the downloaded file, copies / images / reproductions, or the link to these files) in any manner other than by providing the following link — <http://www.culperresearch.com> The failure of Culper to exercise or enforce any right or provision of these Terms of Service shall not constitute a waiver of this right or provision. If any provision of these Terms of Service is found by a court of competent jurisdiction to be invalid, the parties nevertheless agree that the court should endeavor to give effect to the parties' intentions as reflected in the provision and rule that the other provisions of these Terms of Service remain in full force and effect, in particular as to this governing law and jurisdiction provision. You agree that regardless of any statute or law to the contrary, any claim or cause of action arising out of or related to use of this website or the material herein must be filed within one (1) year after such claim or cause of action arose or be forever barred.

Cerebras Systems Inc. (CBRS): Hot Garbage

“It was about a year in that one region failed, so they had to replace the whole thing.”
– Customer #1

“Our first system broke down, and then there was an issue with the replacement... Two replacements were made.” – Customer #2

“The whole wafer was down and wouldn’t come back up in a state that was usable.”
– Customer #3

“The wafer will break... So they need to keep the temperature in the wafer constant.”
– Customer #4

“Once I bought it and realized what it was, I couldn’t believe that they could continue to sell it... we just shut it down.”
– Customer #5

We are short Cerebras Systems Inc (“CBRS”, “Cerebras”, “the Company”) because we believe that the Company’s wafer scale engines (“WSEs”) last just 1.5 to 2.5 years in the field – a lifespan that undermines the entire business. Cerebras claims that “*years ago*” it solved the “*75-year-old compute industry problem*” of building, powering, cooling, and commercializing a chip the size of an entire silicon wafer. CEO Andrew Feldman says the Company solved thermal expansion in particular by having “*invented*” a new connector material.¹ We believe that Cerebras never actually solved thermal expansion, and that the material is a 40-year-old product from a small Massachusetts vendor. Instead of fixing the problem at the hardware level, we believe Cerebras instead programmed its wafers to run “dummy operations” such that the wafers *never actually cool down*.² This jerry-rigged workaround has only traded one fatal flaw for another: the wafers never cool, and break down rapidly in the field.

Feldman has repeatedly claimed that Cerebras “*had to invent*” a connector material that “*could absorb some of that difference*” in expansion rates between wafer and board, hence solving thermal expansion. He has dated that breakthrough to 2018, to 2019, and implied it came as early as 2015. Contrary to Feldman’s claims, however, we believe Cerebras’s connectors rely on PariPoser, a material developed in the 1980s and sold by Paricon Technologies, which stated that PariPoser is used in Cerebras’s “*delivered product*.” One former employee we spoke with also confirmed that the material came from an outside vendor: “*you can spin it the way you want when you market it. Everything can be an invention... it’s not a science invention... it exists in other places.*” PariPoser is not new, Cerebras didn’t invent it, and we don’t believe it ever solved thermal expansion.

We believe that Cerebras has instead programmed its wafers to run dummy operations such that wafers *never actually cool down*. Don’t just take our word for it: in June 2026, at a Café Compute event in Berlin, a *current Cerebras solutions architect* stated explicitly that wafers run dummy operations so that they “*stay functional*,” because without them, it “*can break your chip*.” Our decompilation of Cerebras’s own client-side software corroborates: the code includes a field explicitly labeled “*dummy tokens*” that appears to be distinct from standard padding tokens. We found nothing comparable from Nvidia, Google, or Intel.

¹ Thermal expansion poses two problems at wafer-scale. First, the wafer and the board beneath it expand at different rates as the wafer heats and cools, straining the connections between them. Second, each WSE is stitched together from 84 dies, while uneven workloads heat the full wafer unevenly, which can bend it. Both problems worsen with size, and WSEs are more than 50 times larger than GPUs.

² To be clear, dummy operations are a known technique – chips sometimes use them to ramp power gently between jobs, while AI clusters use similar functions to occasionally smooth power draw. We believe Cerebras goes much further, however, running them constantly such that its wafers never cool down and stay evenly heated across their full surface.

Dummy operations force Cerebras's wafers to draw near-peak power constantly. A 2022 paper co-authored by *one of Cerebras's own engineers* found that when idle, the CS-2 continues to draw 70% of peak power draw. This only seems to have worsened for the CS-3. In April 2026, Harvard and Meta researchers found *the CS-3 draws 80% of peak power when idle*, compared to just 20% of peak for AMD and Nvidia GPUs. Multiple customers described their experiences to us: Cerebras systems draw power constantly, wear out, and break down far sooner than expected – in some cases requiring full system replacements within just a year or two. We include their accounts herein.

GPUs throttle down when idle and run at roughly half power during most inference work. Cerebras wafers draw close to full power all the time, so they wear out much faster. Engineers use the Arrhenius model to estimate a given chip's mean time to failure ("MTTF") based on its operating temperature and activation energy. Our own Arrhenius model – which relies on only measured power draw, in the Company's favor – suggests that Cerebras wafers age 3x to 3.5x faster than GPUs.³ If a GPU lasts 5 to 7 years, wafers can be expected to last just 1.4 to 2.3 years. Two changes have conveniently obscured this problem from investors' view.

First, in Cerebras's preliminary IPO prospectuses filed through December 2025, the Company assigned its computer equipment useful lives of just 1.5 to 3 years – in-line with both our Arrhenius modeling and what many of the Company's hardware customers have described to us.⁴ Just four months later, in the Company's April 2026 S-1, Cerebras nearly doubled that range to 3 to 5 years – a change that significantly inflated reported margins. The Company appears to have never explained the reason for this change. Second, Cerebras shifted from selling hardware to selling cloud compute. In a hardware sale, customers see hardware failures and Cerebras books warrantied replacements in its cost of sales. When a cloud system fails, Cerebras swaps it out behind the scenes and can record the replacement as capital spending, depreciated over its newly extended useful life. Former employees described this to us directly: *"Cerebras gets their money, but it's still a problem for Cerebras... the customers won't see reduced reliability, but Cerebras will have to put in more equipment as a buffer", "that was one of the big pivot points is they realized they need to take it off the customers", "because they need to back it up, if something fails along the way, it's their [Cerebras's] money."*

Cerebras's pitch rests on the claim that Feldman and his co-founders solved what the semiconductor industry could not solve for decades. Their track record, however, gives us little confidence. Most of Cerebras's co-founders and patent holders came from Feldman's last company, SeaMicro. Investors treat that as a positive. We view it differently. AMD bought SeaMicro in 2012. Feldman left in June 2014, calling the team *"stronger than ever,"* but less than a year later, AMD shut it down and took \$76 million in charges. Before that, Feldman was VP of Marketing at Riverstone Networks, which the SEC charged with hiding side agreements that gave customers return rights. Feldman himself pleaded guilty to one felony count of violating internal accounting controls. Riverstone went bankrupt and returned just \$1.06 per share to shareholders – 92% lower than its \$14 spin-off price.

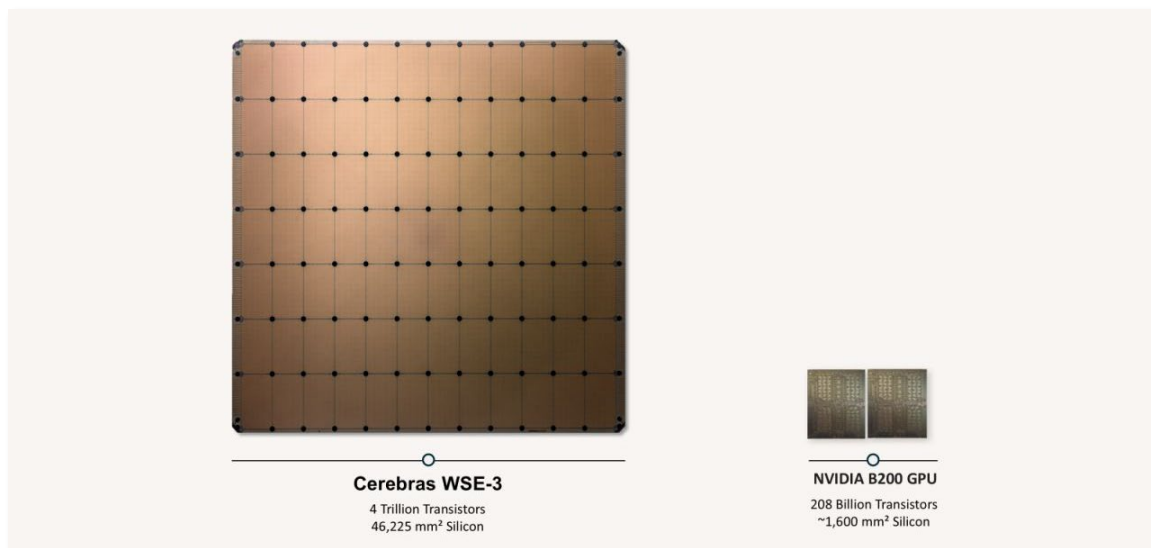
Investors view Cerebras's December 2025 deal with OpenAI as a stamp of credibility. We think it's the opposite. By our estimates, OpenAI is buying compute at a 20% to 77% discount to what hardware customers pay, and at up to an 84% discount once short wafer life is factored in. We estimate the contract's NPV to Cerebras at between roughly break-even and negative \$13.8 billion at a 2.5-year wafer life. Even OpenAI chose Nvidia GPUs, not Cerebras, for its new UltraFast tier. In return, the deal helped Cerebras go public – an IPO that also benefited OpenAI's Sam Altman and Greg Brockman, who had each invested in Cerebras personally. By our estimates, Cerebras's private round investors paid a weighted average price of just \$23 per share, 79% lower than where shares trade today. Insiders have already sold over \$400 million in stock since the Company went public, while another 19.4 million shares unlocked today, and over 50 million more will unlock in early November. We are short and think shares are headed lower.

³ We believe Cerebras wafers run hotter than GPUs, based on conversations with customers and former employees, as well as the Company's own cooling requirements, as more fully described herein. If we assume that Cerebras wafers do run hotter, these figures worsen meaningfully.

⁴ Notably, the December 2025 prospectus was filed well *after* the Company had launched both the CS-3 and the cloud business, suggesting that neither of these dynamics were responsible for the accounting change.

Introduction to Cerebras: the Company that Claims to Have Solved the Unsolvable

Cerebras was [founded](#) in 2015 and sells Wafer Scale Engines (“WSEs”), which the Company calls “*the world’s first and only commercialized wafer-scale processor... 58 times larger than NVIDIA’s B200 chip.*” Each WSE is housed in a system that includes cooling, power, and interconnected technology. The Company introduced its first system, the CS-1, in 2020, the CS-2 in 2021, and its current CS-3 in March 2024. CEO Feldman calls the CS-3 “*the fastest AI processor ever built.*”



Source: [Cerebras CEO Andrew Feldman, LinkedIn](#)

The semiconductor industry has long considered wafer-scale processors effectively impossible to manufacture and commercialize. Feldman himself calls it “75-year-old compute industry problem of wafer-scale integration.” Per the Company’s [Form S-1](#), “*nobody knew how to package such a big chip without cracking it.*”

Thermal expansion, in particular, poses two problems, both of which grow far harder to solve at wafer scale: mismatched expansion rates between the wafer and the board, and uneven heating across the full surface of the wafer.⁵ Cerebras claims, however, to have solved these dilemmas, with Feldman saying that “*mismatches in thermal expansion*” among other problems, are ones “*that Cerebras solved years ago.*”⁶ Central to that claim is a connector material that Feldman said Cerebras had to invent, one that would “*absorb*” mismatches in expansion rates between board and wafer. Contrary to Feldman’s claims, however, we traced that material to a 40-year-old product sold by a small Massachusetts company.

⁵ First, when chips are running, they heat up, and expand. When they are not running, they cool down, and contract. The chip and the surface that it is mounted on, however, are made of different materials, and thus expand and contract at different rates. Each cycle of heating and cooling strains the connections between the chip and underlying surface, and over many cycles, those connections break. As the Company’s Form S-1 put it, these “mismatches in thermal expansion” become greater with the size of the chip. On a conventional GPU, the corners shift ~16 micrometers relative to the board. On a Cerebras wafer, however, they shift ~122 micrometers – nearly 8 times as far. Second, a chip that is heated unevenly will expand unevenly across its surface. If one region of the chip is working while others sit idle, the chip itself is liable to bend and crack. Much like the above, addressing this problem grows in difficulty with the size of the chip. A conventional GPU is a single die, while a Cerebras WSE is a grid of 84 tiles, each of which can run hot or cool based on varying workloads.

⁶ See Feldman’s comments on the Company’s June 2026 conference call.

Cerebras Claims It “Invented” a Connector Material to Solve Thermal Expansion. We Traced It to a Small Massachusetts Company Selling Tech from the 1980s.

Cerebras CEO Andrew Feldman has repeatedly claimed that the Company “*had to invent*” a connector material that would absorb the difference in expansion rates between board and wafer, thereby solving thermal expansion. Our research suggests, however, that the Company’s connectors are actually made from PariPoser – a material that was developed by Bell Labs in the 1980s and is now sold by Paricon Technologies, a small company in Taunton, Massachusetts. We believe the connector material is neither new, nor was it invented by Cerebras, nor did it solve the thermal expansion problem at wafer scale.

In August 2019, Feldman [told TechCrunch](#) that Cerebras “*had to invent a material that could absorb some of that difference*” in expansion rates between board and wafer.⁷ As to how the Company managed to package the wafer-scale chip, Feldman said bluntly, “*Well, the answer is you invent a lot of shit.*” On that much, we agree...

Tellingly, Cerebras never appears to have publicly named this supposed material. When [the EE Times asked](#) about the connector, the outlet reported that, “*Cerebras did not provide further details on this connector, unfortunately.*”

A former Cerebras employee we spoke with, while declining to name the material explicitly, confirmed to us that it came from an outside vendor.⁸

“It was a component that was used previously for other applications, and we basically applied it differently to this application, pushing the component much bigger than it was before. We worked with this vendor to design and qualify, with respect to the pitch on each contact, the thickness, the thermal properties, and at the end of the day, it’s basically a matrix that fixes it in place. These flexible electrical columns are tiny and need to be aligned so that the pads get in contact with the columns on each side of it.”

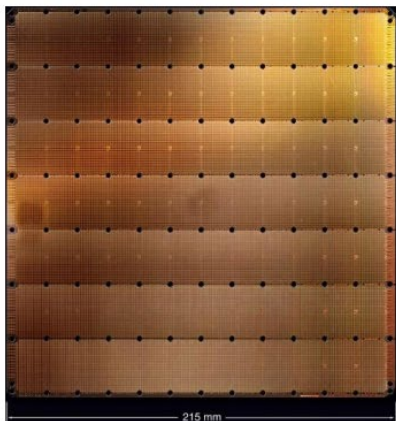
“When you’re saying, ‘invent a new material’, you need to define ‘invent’ and define ‘material’. This is not really on the level of material science. There was some material science thinking going into it, but I would call it just an additional component. You can spin it the way you want when you market it. Everything can be an invention. I’ll put it this way, it’s not a science invention... it exists in other places.”

We believe that contrary to Feldman’s claims of having invented a new connector material, the Company’s connectors are made with PariPoser, a product owned by Paricon Technologies, based in Massachusetts. In May 2020, Paricon [presented a poster](#) at the TestConX conference, describing a connector for a 215mm x 215mm data center wafer – specifications that match Cerebras WSEs – and indeed, Paricon explicitly stated that its wafer connectors are **“used by Cerebras for their mammoth IC... in their delivered product.”**

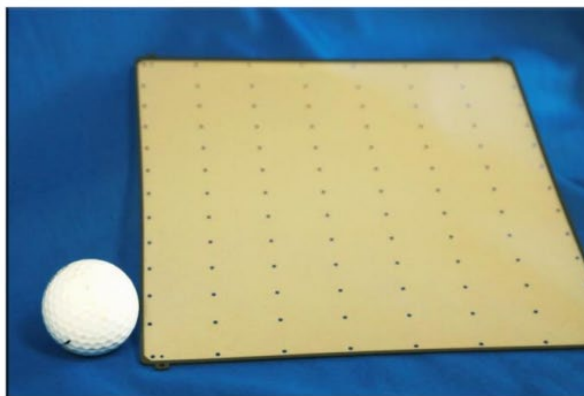
⁷ In a February 2020 [interview](#), Feldman reiterated this “invention” claim, stating, “*We’ve invented materials that allow us to mitigate some of the challenges that you described of thermal expansion.*” We also find it interesting that Feldman cannot seem to keep his timelines straight. In some cases, he dates the supposed breakthrough [to as early as 2015](#), while elsewhere, [he suggests](#) that Cerebras was still trying to solve the problem as late as 2019.

⁸ ... and provided a description that matches PariPoser.

The 219mm x 219mm (8.62" x 8.62") wafer connector shown below is used by Cerebras for their mammoth ICs. (Ref: www.cerebras.net) It's made with a continuous sheet of PariPoser® material mounted on a thin aluminum frame. It is used in their delivered product.



Cerebras 215mm x 215mm wafer



The Paricon wafer connector

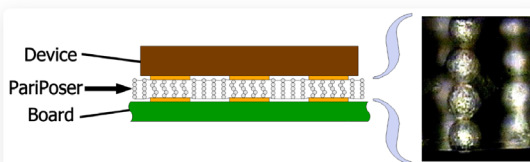
500 Myles Standish Blvd. Taunton, MA 02780 (508) 676-6888 Fax (508) 676-8111 www.paricon-tech.com sales@paricon-tech.com

Source: [Paricon – Wafer Connectors](#)

PariPoser is not a new technology. Paricon's [own website states](#) that the technology “was developed at Bell Labs in the 1980s” and that “when the lab was broken up in the 1990s, Paricon acquired the technology...” PariPoser consists of its trademarked “BallWire” columns of silver-plated nickel particles that form a silicone sheet, the same structure described by Cerebras patents.⁹ These metal balls are then pressed together, and can lean sideways so as to maintain conductivity between the board and wafer under varying levels of thermal expansion.

What Is PariPoser® Technology

PariPoser® material is an anisotropic conductive elastomer. BallWire® columns conduct vertically, while an insulating silicone matrix provides electrical isolation laterally.



Source: [Paricon Technologies](#)

Feldman claimed that Cerebras's supposed invention had to “absorb the expansion mismatch” between wafer and board. Paricon's own product page, however, warns that PariPoser doesn't entirely solve this problem, as differences in thermal expansion between PariPoser, the board, and the device (i.e., the wafer) “might result in horizontal misalignments” and that, “typically, the material CTEs have to match, or otherwise there needs to be an aggressive cooling apparatus engaged.” Given our research suggests that Cerebras runs dummy operations on its wafers, we don't think PariPoser is has fully solved the problem.

⁹ Cerebras holds [several patents](#) related to its connectors, all of which appear to us to describe the connectors without naming the material. In particular, US 10,468,369 – filed in July 2018 and granted in November 2019 – describes a membrane “preferably made using silicon material” filled with “metal ball wires” that “may be made of any suitable conductive material”

The Team That Supposedly Solved the Unsolvable: Responsible for SeaMicro (a Total Bust), and Riverstone Networks (Which Made CEO Feldman a Felon)

We think it's worthwhile to consider the backgrounds of the management team at Cerebras making these claims. The Company was founded by CEO Andrew Feldman alongside Michael James (Chief Architect, Advanced Technologies), Sean Lie (CTO), Gary Lauterbach (CTO, Emeritus), and Jean-Philippe Fricker (Chief System Architect). The Company's key patents name Fricker, as well as Frank Jun and Paul Kennedy. Nearly all of them, along with COO Dhiraj Mallick and [other key employees](#), came from Feldman's last company, SeaMicro. Fricker also worked with Feldman at the company before that, Riverstone Networks.

Feldman himself has made that shared history central to his pitch, practically [bragging that](#) "*all of the founders were with us at SeaMicro*" and that of Cerebras's roughly 200 employees at the time, "*maybe 65 or 70*" had worked with him before.

Investors now treat that history as an indicator of success. For example, one podcast [touts that](#) Feldman "*generated over \$1 billion from selling his startups*" while another piece [refers to](#) SeaMicro as "*a pioneer of energy-efficient, high-bandwidth microservers.*" We view things a bit differently.

Feldman founded SeaMicro in 2007 to build "microservers" that it claimed would disrupt the traditional server market. Intel considered buying the company, yet Intel's data center head Diane Bryant said Intel "*was not impressed, and declined.*" In March 2012, SeaMicro was acquired by AMD in a deal [Forbes called](#) "*a jab at Intel.*" In June 2014, Feldman left AMD, claiming his departure had nothing to do with SeaMicro's prospects, and that the team was "*stronger than ever.*" Less than a year later, AMD shut the business down entirely, took \$76 million in restructuring and impairment charges, and stated that it had no further plans for SeaMicro's "Freedom Fabric" technology. In other words, by the time Feldman was referring to SeaMicro in promotional interviews as "*a very successful previous startup*", AMD had already shut it down. It thus seems to us that SeaMicro was judged a success based primarily on Feldman's having sold it before it imploded.

Prior to SeaMicro, Feldman was VP of Marketing at Riverstone Networks, Inc., which [received SEC charges](#) and went bankrupt in 2006. In particular, the SEC alleged that Riverstone hid side agreements that provided customers with resale and return rights. Feldman and EVP of Sales John Kern each pleaded guilty to one felony count of violating public company internal accounting controls. Riverstone shares had traded at \$14 on their 2001 spin-off from Cabletron, but shareholders received just \$1.06 per share in bankruptcy.

Feldman and co now ask investors to believe they solved a 75-year-old problem that no one else could. We don't.

Rather Than Solve Thermal Expansion, We Believe Cerebras Jerry-Rigged its Wafers to Run Dummy Operations and Never Cool Down

We believe Cerebras has *never* solved thermal cycling at the hardware level, and conceals this by running dummy operations that ultimately come at the cost of significantly shortening the life of its systems in the field.

To be clear, dummy operations on their own are not unusual. At the [SC25 conference](#) in November 2025, Cerebras Chief Architect Michael James [described](#) that the CS-3 will "*do some dummy operations – just multiply random numbers together*" to give the chip has a "*gentle power ramp*" between inference jobs. Smoothing power draw is not an unusual practice among chipmakers. We believe Cerebras has had to go much further, running dummy operations all the time to hold its wafers within a narrow temperature range, both over time and across their surface.

Don't just take our word for it. At a June 2026, "café compute" community meetup in Berlin, a current Cerebras Solutions Architect [stated this explicitly](#), admitting that **without dummy operations, wafers curve and break.**

“And then heat, temperature, yes. An example that I learned from our engineers, and I loved, is that you know, the chip is so big. So if you run some useful computation in the left corner of the chip – left bottom corner – then if you don’t run anything elsewhere, then you know, there it’s going to curve. So it can break your chip. So you need to run some dummy computations on the rest of the chip while you’re running useful computation in the left bottom corner, just to make sure that, you know, the whole chip expands. So there is no curvature and it stays functional.”

Our own analysis of the Company’s client-side software corroborates this dynamic. Interested readers can find the Company’s software packages [here](#) and [here](#). After decompiling, we found that Cerebras has explicitly included “dummy tokens” in the “InferenceUtilizationStats” field.¹⁰

```
message InferenceUtilizationStats {  
  uint64 dummy_tokens_delta = 1;  
  uint64 prompt_tokens_delta = 2;  
  uint64 gen_tokens_delta = 3;  
  uint64 draft_tokens_delta = 4;  
  uint64 total_tokens_delta = 5;  
  int64 time_delta = 6;  
  uint64 utility_tokens_delta = 7;  
  uint64 padding_tokens_delta = 8;  
  uint64 padding_g_tokens_delta = 9;  
  uint64 draft_success_tokens_delta = 10;  
  uint64 padding_bottlenecked_tokens_delta = 11;  
  float useful_token_ratio = 12;  
  uint64 padding_bottlenecked_concurrent_prompts_limit_tokens_delta = 16;  
  uint64 catchup_delta = 13;  
  float time_delta_s = 14;  
  int32 draft_id = 15;  
}
```

Source: Culper Analysis of Cerebras Code

In plain terms, the Cerebras inference runtime appears to report what the wafer’s token slots were spent on during a given window of time. Note that the field appears to be distinct from “padding” tokens – i.e., the standard practice placeholder values added to shorter text sequences to match the length of the longest. To that end, our review of code from Nvidia, Google, and Intel did not uncover any similar mention of “dummy” tokens.

In our view, Cerebras has only traded one problem for another. While dummy operations may keep the wafers from cracking, they force the wafers to draw massive amounts of power around the clock. The wafer itself is 84 dies, and when a job only uses some of those dies, the rest must run dummy operations to keep the wafer evenly heated, as explained by the Company’s own solutions architect. As such, the entire wafer stays hot whether one region is working, or all of them are. Real-world power data bears this out.

In April 2022, Cerebras engineer Mathias Jacquelin and two TotalEnergies researchers [published a paper](#) showing the results of seismic simulations run on a Cerebras CS-2 system. To measure the CS-2’s energy efficiency, the authors extended one test to 754 seconds, and found that during the run, the system averaged 22.8 kW of power, peaked at 22.9 kW of power, and drew baseline power of 16.1 kW. In other words, while running zero workload, the CS-2 continued to draw 70% of its peak power.¹¹

¹⁰ Coincidentally, this is also [cited by industry sources](#) allegedly [citing](#) Cerebras employees.

¹¹ Research [published in January 2026](#), co-authored by another Cerebras engineer, further corroborates.

Similarly, in April 2026, researchers from Harvard and Meta FAIR published a paper comparing eight AI hardware systems – including the Cerebras CS-3, Nvidia A100 and H100, and AMD’s MI300 – on speed, power, and energy efficiency. They found that the CS-3 drew 80% of its rated maximum power while idle – the highest of any platform – and far higher than the 20% of peak for Nvidia and AMD GPUs. Moreover, AI inference runs in two phases: prefill (prompt processing) and decode (response generation). To that end, the researchers found that each platform drew 75% to 100% of its rated power during prefill. During decode, Nvidia GPUs dropped to roughly 50% to 60%. Cerebras wafers, however, drew 100% of rated power in both prefill and decode.

We additionally examine the relative idle power states of each platform as compared to power under high compute-bound workloads. We find idle power is 20% of TDP across NVIDIA GPUs (A100 and H100) and AMD MI300. Gaudi has slightly higher idle power, at 30% of TDP, meanwhile SambaNova idle power sits at 40% of TDP. Cerebras has the highest idle power, at 80% of TDP and note it reaches energy-per-token parity with a 32-GPU H100 cluster at 34% duty cycle. We find that specialized hardware such as Cerebras, Gaudi, and SambaNova have higher idle power than more general-purpose platforms. This has important implications for utilization – as these systems must be highly utilized in order to be efficient.

Source: [*The xPU-athalon: Quantifying the Competition of AI Acceleration*](#)

The Harvard and Meta researchers concluded that high idle power “makes achieving high utilization critical” for Cerebras’s hardware. Keeping the wafers busy, however, does not fix the problem – wafers draw massive power whether idle, prefilling, or decoding. In practice, the wafer never rests.

Cerebras Wafers Last Less than Half as Long as GPUs, Breaking Down After Just ~2 Years in the Field

We estimate that Cerebras wafers last just 1.5 to 2.5 years in the field, less than half the life of competing GPUs and of the Company’s own 3- to 5-year depreciation schedule. Our view is based on conversations with the Company’s hardware customers and former employees, its own historical disclosures, and our Arrhenius modeling. We begin with the people who have run these systems – former employees and customers – who described wafers that draw near-full power even when idle, and that break down far sooner than expected.

“Even in a standby situation, it will use a significant amount of power. The power actually will not change dramatically when it’s idle versus when it’s running.” – Former Employee #1

“If the chip is sitting idle, you’re drawing power and over time that can be expensive.”
– Former Employee #2

“If you’re running inference when there are no tokens to be consumed, the wafer is still active. It’s just the way it’s set up. That’s been the whole thing about it, that the certain amount of power it’s using all the time... Now the CS-3 wafers tend to run at full power regardless. The GPUs may run at half power or something, but the CS-3s run at full power all the time... Our ‘upgrade’ was done because of reliability issues and the processor... The chip would just be down... the whole wafer was down and wouldn’t come back up in a state that was usable... the old ones go back to Cerebras.” – Customer #1

“The machine was way too expensive to operate... it all just cost a lot of money to keep it running for the electricity and the staff time... Even compared to GPU-based systems, which are power hungry, this machine was significant. This was an expensive system compared to everything else. Even if the chip is not running, it soaks up power. You still need the cooling solutions where you still have to keep pumping liquid onto the chip to remove the heat, even when it’s not running. Nvidia GPUs when they idle, they consume power, but significantly less. The system has to work all the time to remove that excess heat... We spent two

and a half years trying to work with it... When we bought it, we were hoping to have it for 5 years... after several attempts of trying to get some use of it, we just shut it down, and we haven't been using it for a while..." – Customer #2

"The wafer will break, physically. So they need to keep the temperature constant... the minimum is 15 kW, so our understanding is they do that to prevent breaking of the wafer. You need dummy operations running on the wafer to keep the 15 kW." – Customer #3

"It needs 16 kW of idle power, so it's always running, and you have this giant wafer, and you need to cool this massive machine, because of this architecture, there's no margin for error... It's a very sensitive design... So we experienced a couple of times, problems in the sense that it's very sensitive to temperature spikes. It had to be really in a certain temperature range... So eventually [within 1.5 years] we had our first system broke down, and then there was an issue with the replacement... Two replacements were made. They had to replace the whole systems." – Customer #4

"The wafer, the system, it's constantly running, but if you're not using it, your baseline is really high... It was about a year in that one region failed, so they had to replace the whole thing. The guarantee covered it, and they replaced the whole wafer." – Customer #5

"Once we had the hardware, we want to keep the device running continuously, because we can't shut it down... Even if you have nothing running, it costs you like 70% of the costs anyway." – Customer #6

Our Arrhenius Modeling Suggests That Cerebras Wafers Age 3 to 3.5 Times Faster Than GPUs, Even When Fully Utilized

Cerebras's dummy operations may keep a wafer from breaking due to thermal expansion, but they create a different problem: the wafer never cools down. In semiconductors, the Arrhenius model predicts a given chip's mean time to failure ("MTTF") from its operating temperature and activation energy. A common rule of thumb holds that every 10°C increase in temperature cuts a chip's life in half.

To that end, we believe Cerebras wafers run meaningfully hotter than GPUs, based on our conversations with customers, former employees, and the Company's own cooling requirements.¹² Our analysis, however, sets that aside *in the Company's favor*. Instead, we estimate the difference in useful lives owing *only* to the fact that a Cerebras wafer runs close to full power around the clock, while a GPU will throttle down far more considerably when idle and runs at roughly half power during most inference work. **Even so, our Arrhenius modeling suggests that Cerebras wafers wear out 3 to 3.5 times faster than GPUs, even at 100% utilization.**¹³

Fleet Utilization	50%	75%	100%
CS-3 aging (% of full-load rate)	81%	91%	100%
GPU aging (% of full-load rate)	23%	28%	33%
Cerebras wears out faster by	3.5x	3.2x	3.0x

¹² E.g., when we asked one former employee whether Cerebras systems run hotter than GPUs, they told us, "Totally" – a view that was shared by multiple additional experts we spoke with. Second, Cerebras required its wafers to stay below 85°C and to vary by no more than 7°C across their surface (Siemens, 2023). Third, per SemiAnalysis, its Oklahoma City data center cools the wafers with water at about 21°C, versus up to 45°C for Nvidia's NVL72, and each CS-3 needs roughly three times the coolant flow per kilowatt. Finally, one of the Company's university customers told us their system required 18°C coolant, while its facility normally runs at about 30°C.

¹³ Assumed activation energy of 0.7 eV. Both chips are assigned a full-load temperature of 50°C, the upper end of what Latif et al. (2025) measured for liquid-cooled Nvidia H100s, with 20°C coolant. Temperature above coolant is scaled by measured power draw (Golden et al., 2026). At idle, the CS-3 draws 80% of peak and GPUs 20%. While working (decode), the CS-3 draws 100% and Nvidia GPUs draw 55%. Each chip's wear is measured against its own rate at full load.

In other words, if GPUs last 5 to 7 years, Cerebras wafers should be expected to last just 1.4 to 2.3 years. That lifespan is consistent both with what multiple customers told us, and with the useful lives that the Company itself assigned to its computer equipment until just before its IPO, when it nearly doubled useful lives without explanation.

Cerebras is Betrayed by its Own Historical Disclosures: the Company Nearly Doubled Useful Lives Just Ahead of IPO Without Explanation

We think Cerebras has already told on itself. In its initial draft registration statements filed June 2024 to December 2025, the Company recorded the useful life of its computer equipment at 1.5 to 3.0 years (2.25-year midpoint).

Asset Category	Useful Life (Years)
Computer equipment	1.5 – 3
Furniture and fixtures	3
Software	3
Machinery and equipment	7
Leasehold improvements	Lesser of estimated useful life or remaining lease term

Estimated useful lives are periodically assessed to determine if changes are appropriate. When assets are retired or otherwise disposed of, the cost of these assets and related accumulated depreciation or amortization are removed from the accounts and any resulting gains or losses are included in loss from operations in the period of disposal.

Source: [*Form DRS filed December 22, 2025*](#)

Just ahead of its IPO, however, the Company increased the useful lives of computer equipment to 3 to 5 years.

Property and equipment are depreciated using the straight-line method over the estimated useful lives of the property and equipment as follows:

Asset Category	Useful Life (Years)
Data center and computer equipment	3–5
Machinery and equipment	7
Leasehold improvements	Lesser of estimated useful life or remaining lease term

Estimated useful lives are periodically assessed to determine if changes are appropriate. Such revisions may result, for example, from changes to plans, demand, or strategy for the Company's inference solutions.

Source: [*Form S-1 filed April 17, 2026*](#)

Notably, Cerebras did not provide any explanation for the change, nor did the Company disclose any reconciliation of the impact of the change on its financial statements.¹⁴ We note that the change is not easily explained by the Company's introduction of the CS-3 system, nor of its shift to a cloud model, which occurred in March 2024 and August 2024, respectively. The Company's Form 10-Q filings for Q1 and Q2 2026 subsequently disclosed "*no significant changes*" in accounting policies since year-end 2025, suggesting that Cerebras continues to depreciate data center and computer equipment on a 3- to 5-year life – double what we think is justified.

¹⁴ This contrasts to e.g., WhiteFiber (WYFI), which also changed its useful life estimates, but provided explicit disclosure of the change and the resulting impact on its financial statements.

By our estimates, at a 2.25-year useful life, the Company would have reported just 24% gross margins and -140% operating margins in 1H 2026.

1H 2026 Financials	As Reported	2.25-year Useful Life
Total revenues	374	374
Cost of revenues	262	284
Gross income	112	90
Gross Margin	29.9%	24.0%
Operating expenses	604	614
Operating income	(492)	(524)
Operating Margin	-131.8%	-140.3%

Cerebras's Shift to Cloud Model Helps Keep Failing Hardware Out of Customers' Sight and Out of Gross Margins

We believe these dynamics also shed light on the Company's August 2024 shift from selling hardware to selling under a cloud model – a change that allows the Company to keep the immediate cost of failing hardware out of customers' sight while blunting its impact on gross margins.

In a hardware sale, Cerebras replaces any failed system covered under warranty – a cost that the Company records in its cost of revenues, while the system failure itself damages the Company's relationship with the customer.

In the cloud model, however, Cerebras retains ownership of the hardware and sells compute as a service. When a system fails, the Company replaces the system ***without the customer – or investors – ever seeing it directly.***

“The [cloud] contracts are to purchase a certain number of tokens over a certain period and price... Cerebras gets their money, but it's still a problem for Cerebras... the customers won't see reduced reliability, but Cerebras will have to put in more equipment as a buffer.” – Former Employee #1

“When its entirely hosted by Cerebras, they can handle all the temperamental issues with the hardware on the back end... They've kind of obfuscated this challenge from the customer... That was one of the big pivot points is they realized they need to take it off the customers.” – Former Employee #2

“[Cerebras can] bake in reserve capacity, and we can hot swap it when the nodes go down... Because they need to back it up, if something fails along the way, it's their [Cerebras's] money.” – Former Employee #3

The cloud model also conveniently allows Cerebras to temporarily conceal the impact on gross margins. Per its SEC filings, purchases of property and equipment are “*primarily for systems to deliver Cerebras Cloud services...*” i.e., systems are booked as capex, rather than directly in cost of revenues, and the resulting purchase is then depreciated over the Company's inflated schedules.

OpenAI Deal Is No Stamp of Credibility: A Massive Discount to Go Public, at an NPV as Low as Negative \$13.8 Billion

Investors view Cerebras's December 2025 deal with OpenAI as a stamp of credibility. We think that is a mistake. By our estimates, OpenAI is buying Cerebras compute at an effective discount of 20% to 77%, and up to an 84%

discount once short wafer life is factored in. At that pricing, we estimate the contract's NPV to Cerebras ranges from negative \$6.3 billion to just \$4.1 billion, even if its wafers last their full term, and from roughly breakeven to negative \$13.8 billion if they last just 2.5 years. In our view, Cerebras signed what will prove to be a value-destructive deal in order to go public, and common shareholders are now left holding the bag.

Cerebras needed a new anchor customer to go public after its previous IPO plans stalled. In January 2024, Bloomberg [reported](#) that Cerebras was weighing an IPO at a discussed offering price above its \$4 billion valuation as of its latest 2021 funding round. In June 2024, Cerebras filed a draft registration statement ("DRS").

At the time, Cerebras depended almost entirely on two UAE-based customers: G42 and the Mohamed bin Zayed University of Artificial Intelligence – related parties of one another – which together would continue to account for a combined 86% of 2025 revenues. G42 in particular had faced [intense scrutiny](#) over their ties to China, including from the New York Times, which reported on G42's "work with large Chinese companies that U.S. officials consider security threats" and that "U.S. officials fear G42 could be a conduit by which advanced American technology is siphoned to Chinese companies or the government..." In October 2024, Reuters [reported](#) that Cerebras was "likely to postpone IPO" due to CFIUS review of G42's investment in the Company. A year later in October 2025, Cerebras raised \$1.1 billion at an \$8.1 billion valuation, and [withdrew](#) its registration statement. Two months later, it signed a deal with OpenAI.

Under the December 2025 Master Relationship Agreement ("MRA"), OpenAI committed to purchase 750 MW of compute, lent Cerebras \$1.0 billion, and received warrants for 33.4 million shares, roughly 12% of the Company on a fully diluted basis. OpenAI CEO Sam Altman and President Greg Brockman had each reportedly invested in Cerebras personally, pre-IPO. In our view, the MRA terms favor OpenAI at nearly every turn: Cerebras is forced to repay OpenAI's \$1 billion loan if the agreement ends for almost any reason, OpenAI can direct how the loan is spent, and is provided the option to purchase more capacity on terms no less favorable than today's, while Cerebras reportedly [agreed](#) that it also would not sell to certain OpenAI competitors for a period of time.

Cerebras CEO Feldman claims that "the market has recognized that fast tokens are more valuable tokens. This supports higher pricing..." yet the Company's single largest deal suggests the exact opposite: by our estimates, OpenAI is receiving a massive effective discount, while the contract terms prevent Cerebras from raising prices.

We estimate that OpenAI's 750 MW of compute requires about 20,000 wafers. Across committed tranche terms of three, four, and four years, that equates to roughly 73,000 wafer-years.¹⁵ At \$20 billion, OpenAI pays about \$273,000 per wafer, per year – or about \$800,000 per wafer in present value at a 10% discount rate.

OpenAI Contract (Culper ests.)	
Compute capacity (three, 250 MW tranches)	750 MW
Wafers (total fleet, all three tranches)	20,000
Term (years, avg of tranches)	3.7
Wafer-years	73,333
Total contract value (\$ billions)	\$20
OpenAI implied price per system, per year of acc	\$272,727
PV discount rate	10%
PV of compute access, per system basis	\$802,417

Cerebras doesn't publish hardware prices, but based on the Company's prior comments, our conversations with multiple Cerebras hardware customers, and multiple third-party [estimates](#), we believe the Company sells systems

¹⁵ Cerebras discloses of the OpenAI deal that "the Company expects to recognize approximately 22% of this revenue [OpenAI contract revenue] in the initial 24 months ending June 30, 2028, 43% between months 25 and 48, and to recognize the remaining balance thereafter."

for \$1 to \$3.5 million each. In 2023, for example, Feldman [told EE Times](#) that G42's purchase of 64 CS-2 systems was a deal "in excess of \$100 million," implying a price of \$1.56 million per system.¹⁶ Hardware customers then pay to run those systems themselves.

Thus, we estimate that – even if each system lasts through its full term – OpenAI receives an effective discount of 20% to 77%.¹⁷ If, as we believe, systems last just 2.5 years, Cerebras must supply roughly 1.5 systems per wafer slot over each term, and OpenAI's effective discount rises substantially.

OpenAI Contract (Culper ests.)	Low	Base	High
Actual life of system (years)	1.5	2.0	2.5
OpenAI contract term (years, avg of tranches)	3.7	3.7	3.7
Systems needed per compute term	2.4	1.8	1.5
OpenAI effective price per system (Culper est.)	\$802,417	\$802,417	\$802,417
Effective price (adj. for useful lives)	\$328,261	\$437,682	\$547,102
Reported hardware pricing	\$3,500,000	\$2,250,000	\$1,000,000
Effective OpenAI discount	-91%	-81%	-45%

We think that the OpenAI deal will ultimately prove value destructive. In 2025, before Cerebras began renting its systems back from G42, its hardware segment generated gross margins of just 43%. At system prices of \$1 to \$2 million, that implies each system costs Cerebras roughly \$570,000 to \$1.1 million. At OpenAI's effective pricing, the contract's revenue is worth about \$14.5 billion in present value terms. Even if Cerebras systems last their full term, the Company will need to supply \$10.4 to \$20.8 billion in hardware, in PV terms, implying an NPV of negative \$6.3 billion to just \$4.1 billion. At our estimated 2.5-year wafer life, Cerebras must replace systems, and NPV falls to as low as negative \$13.8 billion.

Culper Estimates (staggered tranches)	5-year life	2.5-year life
PV of OpenAI Revenues (\$ billions)	\$14.5	\$14.5
Hardware PV (\$571,214 cost per system)	\$10.4	\$20.8
Hardware PV (\$1,142,428 cost per system)	\$20.8	\$41.6
NPV of OpenAI contract (high)	\$4.1	\$0.4
NPV of OpenAI contract (low)	(\$6.3)	(\$13.8)

Further, Cerebras does not appear able to price its way out. Per the MRA, OpenAI has the option to procure additional capacity at terms that are at least equivalent and no less favorable to it than the existing terms.

b. **Additional Capacity.** OpenAI will have the option to procure additional Capacity, and if OpenAI exercises such option, Cerebras shall deliver such additional Capacity at prices, performance, quality and service levels that are at least equivalent to, and no less favorable to OpenAI than, those applicable to the Capacity delivered under Section 2.a, as follows:

Source: [December 2025 Master Relationship Agreement](#)

¹⁶ Alternatively, see estimates from: [Contrary Research](#) (\$2 million per system), [GPU.net](#) (\$2 to \$3 million), [BenchGecko](#) (\$3.5 million), [7wData](#) (\$2 to \$3 million). Our customer conversations corroborate – per one hardware customer, "\$1 million per WSE" not including supplemental equipment or any other costs. Per another customer, "\$1 to \$1.5 million per system." Per another, "an order of magnitude higher" than GPUs at \$100,000 each.

¹⁷ OAI effective price of \$802,417 as compared to low-end of \$1 million and high end of \$3.5 million per system.

Tellingly, on September 30, 2026, SemiAnalysis reported that OpenAI would serve its newest UltraFast tier on Nvidia GPUs rather than on Cerebras systems as expected. Cerebras CEO Feldman has [since claimed](#) that OpenAI opted for Nvidia because Cerebras did not have enough data center capacity. Feldman's own prior words undercut that explanation, however. On the Company's Q1 2026 conference call, CEO Feldman stated explicitly that the chips needn't be hosted in Cerebras's own data centers.

*"I think OpenAI has a choice. **They can deploy it in their data centers in a model where they buy the hardware or they can receive the compute via cloud service.** I think it will depend on OpenAI sort of a portfolio decision of their data centers and their various capacity versus what we can bring in data centers. I think that's likely to be the determining factor, but I think that's really an important question for them."*

The MRA confirms explicitly that the services may be hosted at facilities owned by *either* Cerebras or OpenAI, or their subcontractors.

f. Data Center Identification. The Parties will work together in good faith to identify data centers where the Services will be deployed and may include facilities managed and/or controlled by (i) Cerebras, (ii) OpenAI, or (iii) Subcontractors of either Cerebras or OpenAI.

Source: [December 2025 Master Relationship Agreement](#)

OpenAI could have simply put Cerebras systems in its own buildings. It didn't.¹⁸ Cerebras bulls have since dismissed the news after SemiAnalysis later [reported that](#) "*all of the Cerebras chips at OpenAI are used to serve Ultrafast to Jane Street and a few other customers.*" In our view, that raises an entirely new problem...

Competition is Already Here: Etched is Poaching Cerebras Talent and Customers

The end users of Cerebras capacity at OpenAI include Jane Street, which is now buying inference hardware from a Cerebras competitor, Etched. Cerebras management has consistently pointed to Nvidia and AMD as competitors, and the Company's Form S-1 names Nvidia, AMD, Intel, the hyperscalers, and CoreWeave as competitors. We think they'd prefer not to discuss Etched.

Etched was founded by three ex-Harvard Thiel Fellows, and in October 2026 [reportedly](#) reviewed fundraising offers at a \$40 to \$50 billion valuation. We think Etched poses a significant threat to Cerebras. Etched's transformer-only inference chip targets Cerebras's key weakness in memory capacity. Its rack pools fast SRAM with high-capacity HBM, aiming to combine Cerebras-like speed with GPU-like capacity for multi-trillion-parameter models. Etched also claims that its low-voltage design sustains over 80% of peak performance without throttling. Further, Etched is not wafer-scale, avoiding the thermal issues that we have outlined extensively. In July 2026, Jane Street received their first rack from Etched. According to Etched, Jane Street had [tested](#) the hardware and was "*pleased with the early results.*" One month later, Jane Street led a \$700 million fundraising round for Etched at a \$21 billion valuation.¹⁹

Etched also appears to be attracting significant talent. Several former Jane Street engineers and Cerebras employees have moved to Etched, including Cerebras's "*founding engineer of the inference platform team.*"

¹⁸ One former employee also expressed their view that OpenAI's using Nvidia rather than Cerebras did not stem from a lack of data center capacity, but instead Cerebras's inability to produce the actual systems: "*what's going on now is figuring out how to, as quickly and reliably as possible, how to ramp up production. They just can't make them as fast. Cerebras cannot make thousands of machines per month.*"

¹⁹ Kleiner Perkins, Sequoia, a16z, Peter Thiel, BCV, and Blackstone also participated.


Kamil Khan
 Supercomputing @ Etched | Ex-Cerebras, Ex-Google AI

Experience


Member of Technical Staff
 Etched · Full-time
 Apr 2026 - Present · 7 mos
 San Jose, California, United States · On-site
 Supercomputing @ Etched


Senior Member of Technical Staff
 Cerebras Systems · Full-time
 Jan 2025 - Jul 2025 · 7 mos
 Sunnyvale, California, United States · On-site
 - Founding Engineer of the Inference Platform team, chartered with providing native support for inference workloads on our proprietary wafer-scale compute appliances... more

Source: *LinkedIn*

As one former Cerebras employee warned us, “*If Etched is mature by 2028, they might beat Cerebras for business. Cerebras is currently the leader. That may not be the case if Etched continues to deliver what they have.*”

Conclusion

Cerebras has convinced investors that it accomplished what no one else in semiconductor industry could accomplish for 75 years. We don’t believe it. In our view, Cerebras built a fast chip, but one that cannot survive long enough in the field to make its business economic. Conveniently, the Company’s extended depreciation schedules and cloud shift then concealed this dynamic, while the OpenAI deal allowed it to go public at an over \$50 billion valuation. Its management team has a less than enviable record, in our view: SeaMicro ended in a write-off, and Riverstone Networks ended in bankruptcy, with Cerebras CEO Feldman pleading guilty to felony fraud charges. This same team is now selling their shares in Cerebras. Today, another 19.4 million shares were released from lock-up. In early November, all remaining shares not previously eligible for sale – over 50 million shares, by our estimates – will also be released from lock-up.²⁰ We are short and think shares are headed lower.

6:00 a.m. Eastern Time on October 14, 2026.	An aggregate of up to approximately 19.4 million shares held by Directors and Officers, Non-Executive Employees, and Non-Employee Holders.
6:00 a.m. Eastern Time on October 28, 2026.	An aggregate of up to approximately 19.4 million shares held by Directors and Officers, Non-Executive Employees, and Non-Employee Holders.
The earlier of (i) 6:00 a.m. Eastern Time on the second trading day following our release of earnings for the quarter ending September 30, 2026 or (ii) 180 days after the date of this prospectus (the “Lock-up Period”).	All remaining shares held by our stockholders not previously eligible for sale, subject to applicable limitations under Rule 144, including for “affiliates” and compliance with other applicable law, as described below.

Source: *Form 424B4 filed May 14, 2026 (page 189)*

²⁰ Lock-up expires November 9, as Q3 results are set for November 12, 2026. Per the Q2 2026 Form 10-Q, 237.6 million shares outstanding, less 34.5 million IPO shares, 171.1 million early releases, and 1.2 million RSU sell-to-cover shares, leaving 31 million locked shares remaining, plus 21.5 million vested options that will also become eligible for sale.