# Cerebras — Faster Tokens Please

# Cerebras — 请提供更快的 Token

OpenAI and AWS Partnerships, Tokenomics Explainer, Architecture Deep Dive, Datacenter Ramp, Technical Roadmap

OpenAI 与 AWS 合作伙伴关系、代币经济学详解、架构深度解析、数据中心扩建、技术路线图

MYRON XIE, JORDAN NANOS, MAX KAN, AND 10 OTHERS

MYRON XIE, JORDAN NANOS, MAX KAN 以及其他 10 位作者

MAY 14, 2026 2026 年 5 月 14 日 · PAID · 已付费



It's been nearly 5 years since Dylan wrote a dedicated article about Cerebras in June of 2021 for the newsletter. He shipped 4 articles in 2 days! They could be read inHow times have changed.

自 2021 年 6 月 Dylan 在时事通讯中为 Cerebras 撰写专栏文章以来，已经过去了近 5 年。当时他在两天内发布了 4 篇文章！回首往事，时代确实变了。

One of the other things that has changed is Cerebras's fortunes. With the arrival of fast tokens on the mainstage and a 750MW compute deal with OpenAI notched, Cerebras is feeling ready for the scrutiny of public markets. Up until just 6 months ago, we felt that the Wafer Scale Engine, despite its bold innovations, had some technical weaknesses that were too hard to cover up. Thus, the continued popularity of HBM-based accelerators such as GPU and TPU. The strengths of Cerebras (namely: speed), have been overlooked for years in favor of total throughput. But now, with frontier labs releasing fast, priority, standard and batch tiers of the same model weights, the world has revealed their preference for fast tokens with their wallets. This brings Cerebras's strengths to the fore and is the key reason why OpenAI is willing to fork over tens of

billions of dollars for Cerebras compute.

另一个发生变化的是 Cerebras 的命运。随着“快速 Token”（fast tokens）登上主流舞台，以及与 OpenAI 达成的一项 750MW 计算协议，Cerebras 已经准备好迎接公开市场的审视。直到 6 个月前，我们还认为晶圆级引擎（Wafer Scale Engine）尽管拥有大胆的创新，但在技术上仍存在一些难以掩盖的弱点。因此，基于 HBM 的加速器（如 GPU 和 TPU）持续受到青睐。多年来，Cerebras 的优势（即：速度）为了追求总吞吐量而被忽视。但现在，随着前沿实验室针对同一模型权重发布了快速、优先级、标准和批处理等不同层级的服务，世界已经用钱包证明了他们对“快速 Token”的偏好。这使得 Cerebras 的优势脱颖而出，也是 OpenAI 愿意为 Cerebras 的算力支付数百亿美元的关键原因。

Demand is so strong it’s making everyone look good.

需求如此强劲，以至于让每个人的表现看起来都很出色。

Today, on the verge of Cerebras’s IPO, and because we love the wafer, we are shipping an article that is as long as 4 normal articles. Inside, we will dive deep on:

今天，在 Cerebras 即将 IPO 之际，出于对晶圆技术的热爱，我们将发布一篇篇幅相当于 4 篇普通文章的长文。在文中，我们将深入探讨：

1. Fast inference 快速推理
2. WSE-3, Cerebras’ unique wafer-scale chip

   WSE-3，Cerebras 独特的晶圆级芯片
3. CS-3, Cerebras’ system, with its unique architecture

   CS-3，Cerebras 拥有独特架构的系统
4. Provide a BOM cost analysis

   提供 BOM（物料清单）成本分析
5. Explain when and how the wafer wins for fast inference

   解释晶圆级引擎在何时以及如何凭借快速推理胜出

6. Describe some of the wafer's limitations, showing tradeoffs

描述该晶圆的一些局限性，展示权衡取舍

For paid subscribers we also show the economics of the huge OAI Inference deal that has changed the company's fortunes and share our insights on how far along Cerebras is in becoming a neocloud (i.e. securing the 750MW they need by 2028 for OpenAI). Furthermore, we will talk about Cerebras' future plans of hybrid bonding an wafer scale optical transceiver onto their WSE compute engine, which they claim they are pursuing strictly for the love the game as it is not needed for LLM inference, but is needed for HPC boomer workloads. The HPC customers whom NVIDIA has effectively abandoned after reducing FP64 native hardware on their GPUs to basically nothing.

针对付费订阅者，我们还将展示那笔改变了公司命运的巨额 OAI 推理订单背后的经济账，并分享关于 Cerebras 在成为“新云”（neocloud）道路上进展如何的见解（即在 2028 年前为 OpenAI 锁定所需的 750MW 电力）。此外，我们还将讨论 Cerebras 的未来计划，即在其 WSE 计算引擎上混合键合晶圆级光收发器；他们声称追求这一技术纯粹是出于“对游戏的热爱”，因为 LLM 推理并不需要它，但高性能计算（HPC）的老派工作负载却必不可少。在 NVIDIA 将其 GPU 上的原生 FP64 硬件削减到几乎为零后，这些 HPC 客户实际上已被 NVIDIA 抛弃。

## The Need for Speed 对速度的渴求

Fast inference has arrived.

快速推理时代已经到来。

While SemiAnalysis has historically been an SRAM machine hater, all this changed when Nvidia licensiquihired Groq in December 2025. Clearly Jensen saw at least $20B of value, and he was proven right just a couple months later when we hit the Claude Code Inflection Point. Now, the wafer is here to stay.

虽然 SemiAnalysis 历来对 SRAM 架构并不看好，但当 Nvidia 在 2025 年 12 月“授权式收购”了 Groq 后，一切都改变了。显然，黄仁勋从中看到了至少 200 亿美元的价值，而仅仅几个月后，当我们迎来 Claude Code 拐点时，他的判断得到了证实。现在，晶圆级计算（Wafer-scale）将长久地存在下去。

Many (including Andrej Karpathy) previously believed that raw intelligence/capabilities mattered far more than speed, but our revealed preferences ended up proving that there are times when the opposite is true. Past a certain threshold of intelligence, developers prefer faster tokens to smarter tokens. And in a world where AI is involved in almost every aspect of your workflow, the speed at which tokens are generated can be the bottleneck to "flow state", i.e. how much productive work is completed.

许多人（包括 Andrej Karpathy）此前认为，原始智能/能力远比速度重要，但我们表现出的偏好最终证明，在某些情况下事实恰恰相反。一旦超过某个智能阈值，开发者更倾向于选择"更快的 Token"而非"更聪明的 Token"。在一个 AI 几乎参与工作流方方面面的世界里，Token 生成的速度可能成为进入"心流状态"的瓶颈，即影响实际完成的生产性工作量。

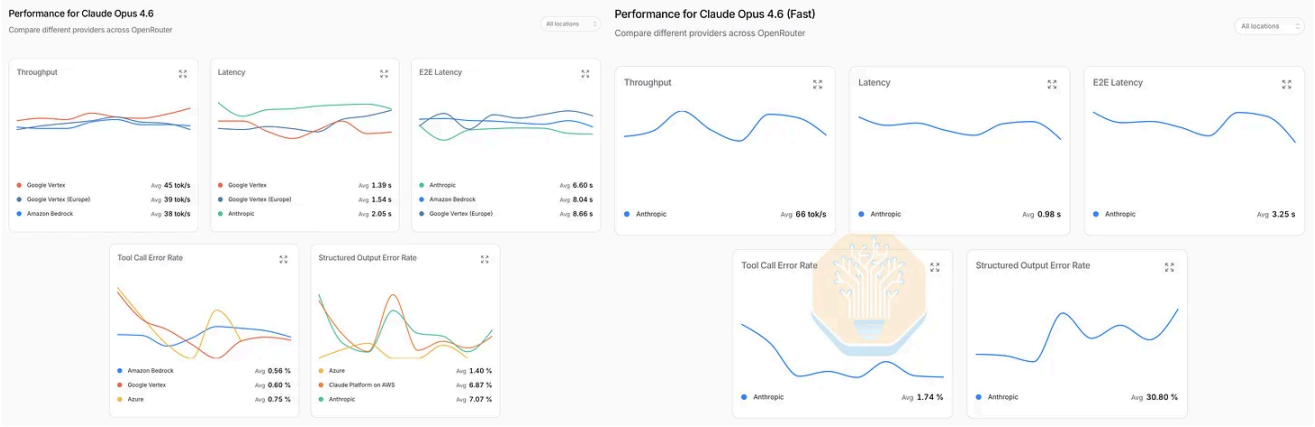
Opus 4.6 fast mode famously charges 6x the price for 2.5x the interactivity (though its now under 2x faster, see chart below). In April, 80% of our AI spend (which peaked at $10M annualized) was on Opus 4.6 fast. When Opus 4.7 came out, many of our engineers refused to switch over because it didn't include fast mode. Notably, this is the first time we've ever decided to forgo frontier intelligence in exchange for faster tokens (and at a significant price premium too!).

众所周知，Opus 4.6 的快速模式（fast mode）以 6 倍的价格提供了 2.5 倍的交互性（尽管现在速度提升已不足 2 倍，见下图）。今年 4 月，我们 80% 的 AI 支出（年化峰值达 1000 万美元）都花在了 Opus 4.6 快速模式上。当 Opus 4.7 发布时，我们的许多工程师拒绝切换，因为它不包含快速模式。值得注意的是，这是我们第一次为了更快的 Token 而决定放弃最前沿的智能（而且还是在支付了巨额溢价的情况下！）。

As an aside, Opus 4.6 fast has become an increasingly worse deal as of late. Standard Opus 4.6 interactivity in Claude Code is consistently around 40 tps (tokens per second). Opus 4.6 fast used to deliver > 100 tps, fulfilling the 2.5 faster guarantee. But it recently degraded to ~70 tps (only 1.75x faster). We recently worked with our friends at

OpenRouter to gather this data on the two operating modes of Claude Opus.

顺便提一句，Opus 4.6 fast 最近的性价比变得越来越低。Claude Code 中标准的 Opus 4.6 交互速度始终保持在 40 tps（每秒 token 数）左右。Opus 4.6 fast 曾经能提供超过 100 tps 的速度，兑现了快 2.5 倍的承诺。但最近它已降至约 70 tps（仅快了 1.75 倍）。我们最近与 OpenRouter 的朋友合作，收集了关于 Claude Opus 这两种运行模式的数据。




Source: OpenRouter 来源：OpenRouter

We believe Opus 4.6 Fast is Anthropic's highest margin SKU and a big reason for their explosion in ARR this year. However, we'll see if this remains true given the slower speeds, delayed 4.7 support, and upcoming Mythos release. For in-depth details on OpenAI/Anthropic revenue broken down by model, see our Tokenomics Model.

我们认为 Opus 4.6 Fast 是 Anthropic 利润率最高的 SKU，也是其今年年度经常性收入（ARR）爆发式增长的重要原因。然而，考虑到速度变慢、4.7 版本支持推迟以及即将发布的 Mythos，这种情况是否能持续还有待观察。有关 OpenAI/Anthropic 按模型细分的收入详情，请参阅我们的 Tokenomics 模型。

# The Throughput-Interactivity Frontier

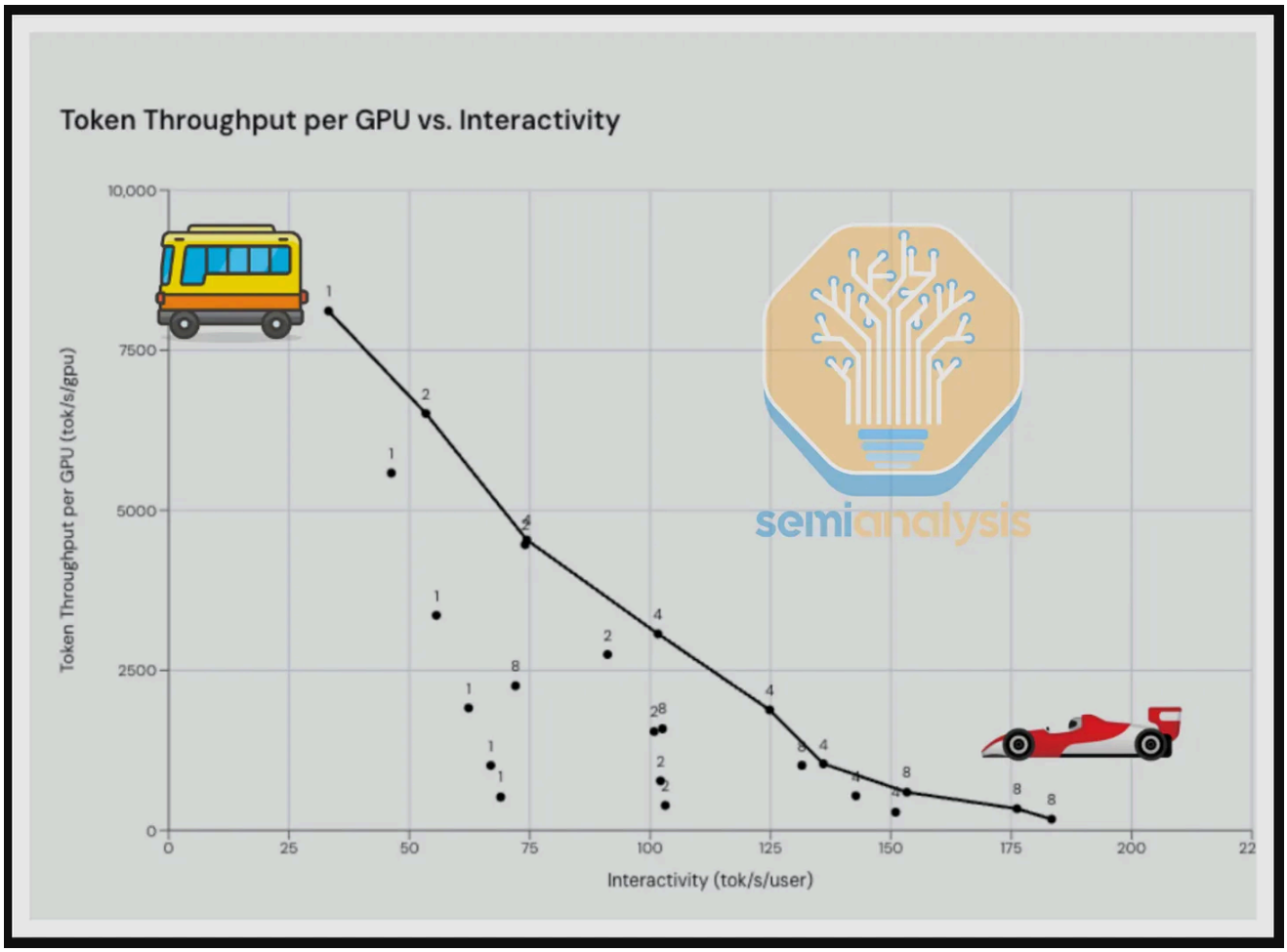
# 吞吐量-交互性前沿

To fully explain the architectural decisions Cerebras has made with their wafer scale chip, we first need to revisit inference fundamentals.

为了充分解释 Cerebras 对其晶圆级芯片所做的架构决策，我们首先需要重新审视推理的基础原理。

As Jensen repeatedly emphasized during this year's GTC, throughput (tokens/sec/gpu) vs interactivity (tokens/sec/user) is the fundamental trade-off for inference. In our original InferenceX writeup, we described it as a bus vs a Ferrari: you can choose to serve lots of users slowly, a single user quickly, or anything in between.

正如 Jensen 在今年的 GTC 上反复强调的那样，吞吐量（token/秒/GPU）与交互性（token/秒/用户）是推理过程中的基本权衡。在我们最初的 InferenceX 文章中，我们将其描述为公交车与法拉利的对比：你可以选择缓慢地服务大量用户，快速地服务单个用户，或者介于两者之间的任何状态。

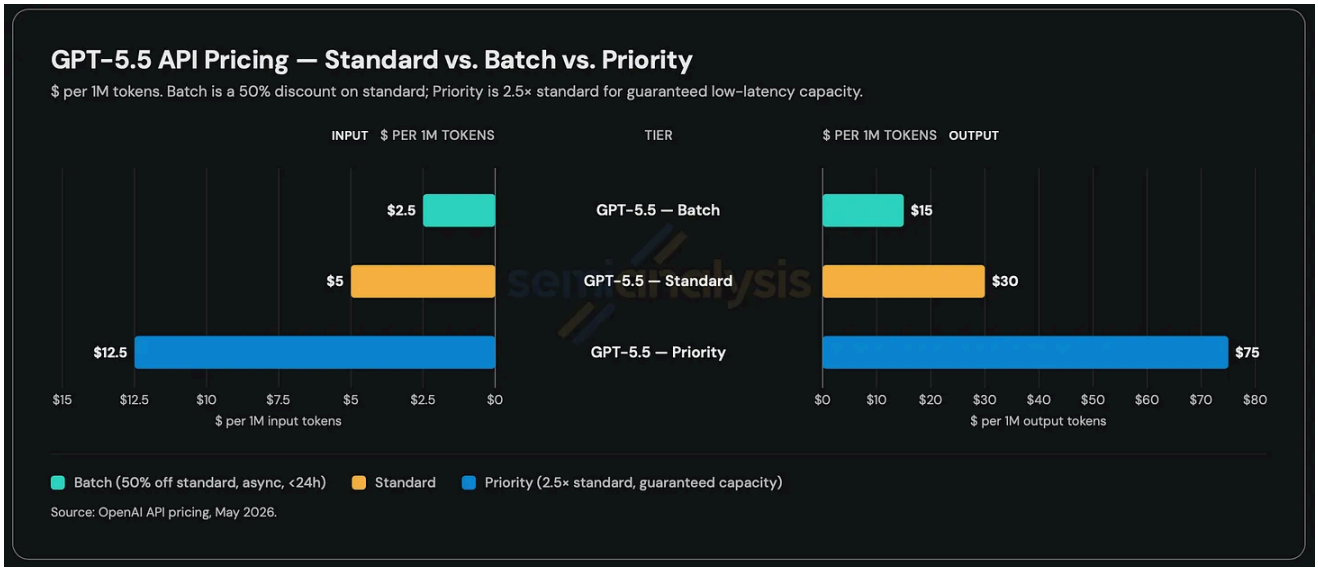



Source: SemiAnalysis InferenceX

SemiAnalysis InferenceX

Of course, users are also willing to pay more money for higher interactivity, so it's currently unclear exactly which spot along the pareto frontier maximizes overall revenue and profitability of inference for a given model provider. In reality, providers are currently deploying multiple options in an attempt to capture the entire market. Fast mode, priority mode, batch pricing, and specific model architectures are all experiments from OpenAI and Anthropic to find the optimal combination for their user base.

当然，用户也愿意为更高的交互性支付更多费用，因此目前尚不清楚在帕累托前沿的哪个点能让特定模型提供商的推理总收入和利润实现最大化。实际上，提供商目前正在部署多种方案，试图占领整个市场。OpenAI 和 Anthropic 推出的快速模式、优先级模式、批量定价以及特定的模型架构，都是为了为其用户群体寻找最佳组合而进行的尝试。



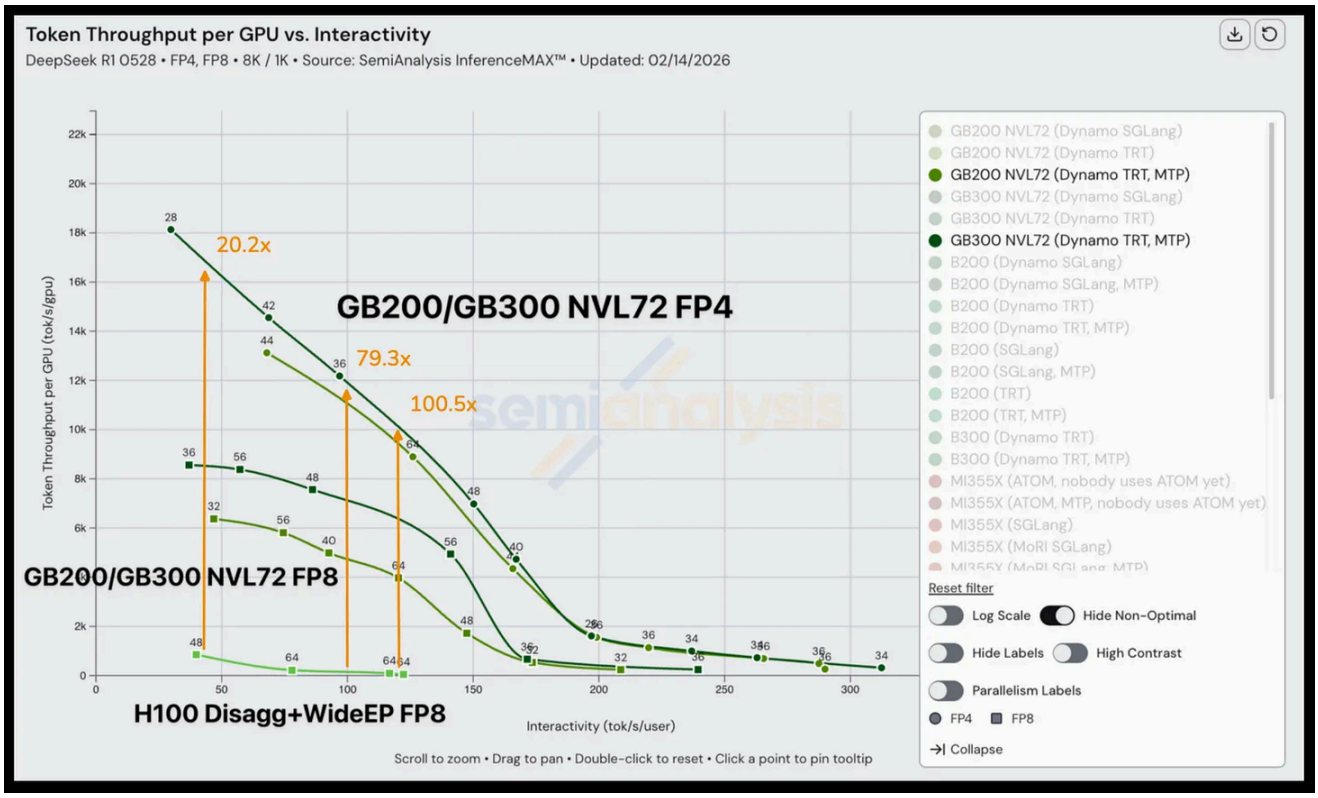

Source: SemiAnalysis Tokenomics Model

SemiAnalysis 代币经济学模型

Manipulating batch size (or "concurrency", the number of users you serve simultaneously) is the primary way to move along the curve given the same hardware. This is the beauty of InferenceX. Whereas most other public inference benchmark only considers a single workload at a single interactivity level, InferenceX builds the entire pareto frontier across 3 different input/output sequence length combos for all the top open-source models. This allows you to make charts like the following, which shows that GB300 NVL72 achieves 20x more throughput than H100s at low

interactivity (40 tps) and 100x more throughput at high interactivity (120 tps).

在硬件相同的情况下，调整 Batch Size（或“并发量”，即同时服务的用户数量）是沿着曲线移动的主要方式。这正是 InferenceX 的魅力所在。大多数其他公开推理基准测试仅考虑单一交互水平下的单一工作负载，而 InferenceX 为所有顶尖开源模型构建了跨越 3 种不同输入/输出序列长度组合的完整帕累托前沿（Pareto Frontier）。这使您能够绘制出如下表所示的图表，该图表显示 GB300 NVL72 在低交互性（40 tps）下的吞吐量是 H100 的 20 倍，而在高交互性（120 tps）下则达到了 100 倍。




Source: SemiAnalysis InferenceX Dashboard

SemiAnalysis InferenceX 仪表板

Alternatively, you can move along the frontier by changing the underlying hardware. This is the promise of SRAM machines like Cerebras and Groq. Their extremely high memory bandwidth allows them to increase throughput at high interactivity, and in the extreme case, achieve interactivity levels that are simply impossible for HBM-based accelerators. Cerebras offers speeds in the thousands of tokens per second, which is literally off the chart compared to the accelerators we benchmark in

InferenceMax

或者，你也可以通过更换底层硬件来沿着前沿移动。这就是像 Cerebras 和 Groq 这种 SRAM 架构机器的潜力所在。它们极高的内存带宽使其能够在保持高交互性的同时提高吞吐量，在极端情况下，甚至能达到基于 HBM 的加速器根本无法实现的交互水平。Cerebras 提供的速度可达每秒数千个 token，与我们在 InferenceMax 中基准测试的加速器相比，这简直是降维打击。

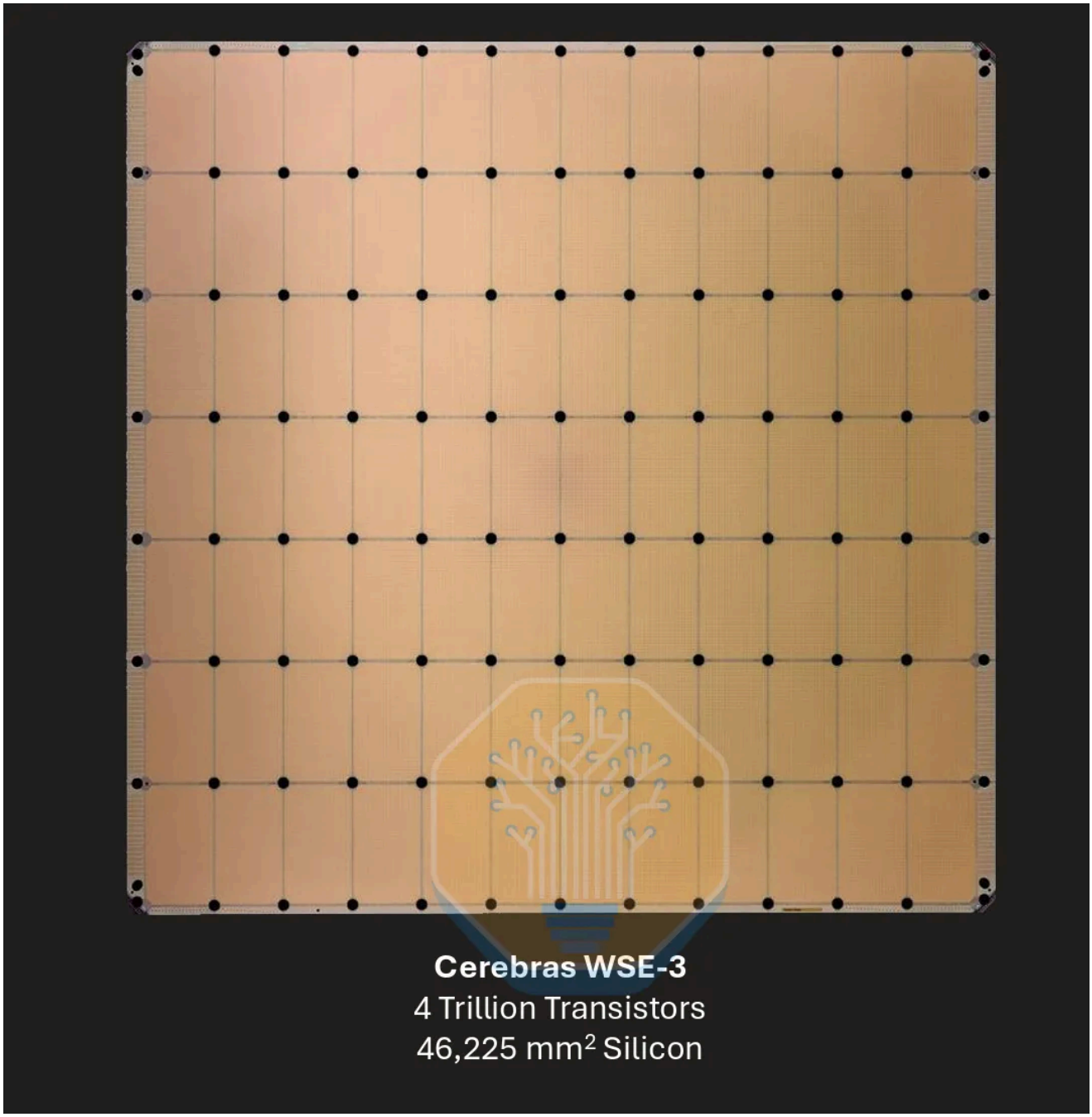
In a world where people are willing to pay more for faster tokens, SRAM machines look quite attractive as they let you both (a) serve more users concurrently at premium speed (pushing the frontier "up") and (b) serve some users at even faster, more expensive speeds (extending the frontier to the right).

在一个人们愿意为更快的 token 支付更高溢价的世界里，SRAM 机器看起来非常有吸引力，因为它们让你既能 (a) 以溢价速度同时服务更多用户（将前沿"向上"推），又能 (b) 以更快、更昂贵的速度服务部分用户（将前沿向右延伸）。

## The Wafer-Scale Engine 晶圆级引擎

Cerebras's fundamental bet has been to go beyond the reticle limit for a single piece of silicon. Instead of splitting a wafer into multiple chips, the goal is to make the entire wafer a chip. This clever scaling was to address a whole host of problems incurred by the slowdown of Moore's law and the hard constraint of silicon being no larger than $858mm^2$; the size of a single reticle pattern in mask-based lithography. This single wafer-sized chip is called their Wafer Scale Engine (WSE).

Cerebras 的根本赌注在于突破单块硅片的掩模版尺寸限制（reticle limit）。其目标并非将晶圆分割成多个芯片，而是将整块晶圆制成一个芯片。这种巧妙的扩展方式旨在解决因摩尔定律放缓以及硅片尺寸不得大于 $858mm^2$（光刻技术中单个掩模版图案的尺寸）这一硬性约束所带来的一系列问题。这种单块晶圆大小的芯片被称为其晶圆级引擎（Wafer Scale Engine，简称 WSE）。

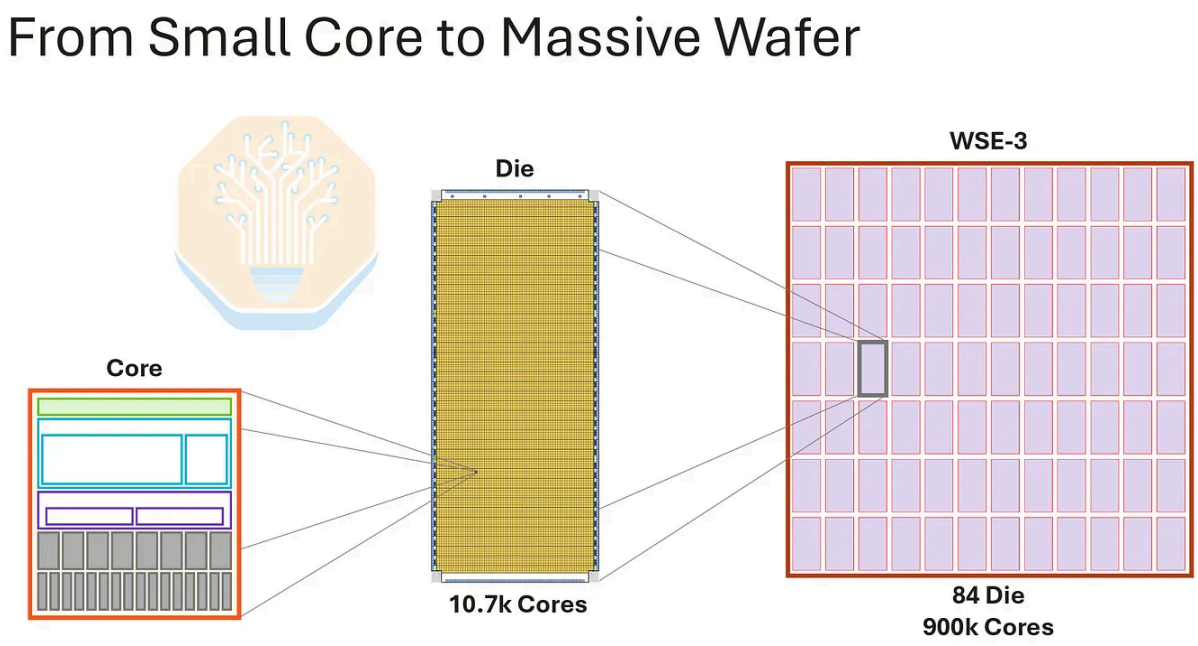



Source: Cerebras Cerebras

The WSE is a 12 x 7 grid of 84 identical steppings/die on a whole wafer that forms one piece of silicon. Each wafer or chip has a large pool of very fast SRAM. 50% of silicon area is dedicated to SRAM cells with the remaining 50% consisting of compute cores. The key innovation is having both the silicon and memory on one piece of silicon instead of interconnecting multiple different chips together. This saves power, latency,

and cost of moving data off-silicon or off-package.

WSE 是由 84 个完全相同的步进/裸片组成的 12 x 7 网格，分布在整个晶圆上，构成了一整块硅片。每个晶圆或芯片都拥有一个巨大的极速 SRAM 池。50% 的硅片面积专门用于 SRAM 单元，其余 50% 由计算核心组成。其核心创新在于将硅片和内存整合在同一块硅片上，而不是将多个不同的芯片互连在一起。这节省了功耗、降低了延迟，并减少了将数据移出硅片或封装的成本。




Source: Cerebras Cerebras

“Traditional” GPUs and XPUs need advanced packaging and networking to achieve greater levels of aggregate compute and memory, which incurs costs in terms of power, speed and more networking equipment. While not a like-for-like comparison, Cerebras compares its on-wafer dataflow speeds to Nvidia’s off-package scale-up bandwidth based on the assumption that data can stay on the WSE whereas GPU data needs to move off-package.

“传统” GPU 和 XPU 需要先进的封装和网络技术来实现更高水平的聚合计算与内存，这在功耗、速度和更多网络设备方面带来了额外成本。虽然这并非完全对等的比较，但 Cerebras 将其晶圆级数据流速度与 Nvidia 的封装外扩展带宽进行了对比，其前提是数据可以保留在 WSE 上，而 GPU 数据则需要移出封装。

| Chip Specifications | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | GB300 | Vera Rubin | TPU v8i | Trainium3 | LPU1 | LPU3 | WSE-2 | WSE-3 |
| Main Memory Type | HBM3E 12-Hi | HBM4 12-Hi | HBM3E 12-Hi | HBM3E 12-Hi | SRAM | SRAM | SRAM | SRAM |
| Main Memory Capacity (GB) | 288 | 288 | 288 | 144 | 0.23 | 0.50 | 40.0 | 44.0 |
| Main Memory Bandwidth (TB/s) | 8.0 | 20.5 | 8.6 | 3.6 | 80 | 150 | 20,000 | 21,000 |
| FP8 FLOPS (TFLOPS)[1] | 5,000 | 17,500 | 10,100 | 2,517 | 750 | 1,200 | 7,500 | 15,625 |
| FP16 FLOPS (TFLOPS) | 2,500 | 4,000 | 5,050 | 671 | N/A | N/A | 7,500 | 15,625 |
| Logic Silicon Area (mm²) | 1,581 | 2,079 | 1,399 | 1,444 | 725 | 813 | 46,225 | 46,225 |
| Scale Out Bandwidth (GB/s uni-di) | 100 | 200 | N/A | 50 | N/A | 50 | 150 | 150 |
| Scale Up Bandwidth (GB/s uni-di) | 900 | 1,800 | 1,200 | 1,200 | 480 | 1,125 | N/A | N/A |
| Aggregate Scale Up and Scale Out Bandwidth (GB/s Uni-di) | 1,000 | 2,000 | 1,200 | 1,250 | 480 | 1,175 | 150 | 150 |
| *(1) INT8 for LPU1* | | | | | | | | |

| Chip Specifications Relative to GB300 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | GB300 | Vera Rubin | TPU v8i | Trainium3 | LPU1 | LPU3 | WSE-2 | WSE-3 |
| Main Memory Type | HBM3E 12-Hi | HBM4 12-Hi | HBM3E 12-Hi | HBM3E 12-Hi | SRAM | SRAM | SRAM | SRAM |
| Main Memory Capacity (GB) | 1.0x | 1.0x | 1.0x | 0.5x | 0.0x | 0.0x | 0.1x | 0.2x |
| Main Memory Bandwidth (TB/s) | 1.0x | 2.6x | 1.1x | 0.4x | 10.0x | 18.8x | 2504.1x | 2629.3x |
| FP8 FLOPS (TFLOPS)[1] | 1.0x | 3.5x | 2.0x | 0.5x | 0.2x | 0.2x | 1.5x | 3.1x |
| FP16 FLOPS (TFLOPS) | 1.0x | 1.6x | 2.0x | 0.3x | N/A | N/A | 3.0x | 6.3x |
| Logic Silicon Area (mm²) | 1.0x | 1.3x | 0.9x | 0.9x | 0.5x | 0.5x | 29.2x | 29.2x |
| Aggregate Scale Up and Scale Out Bandwidth (GB/s Uni-di) | 1.0x | 2.0x | 1.2x | 1.3x | 0.5x | 1.2x | 0.2x | 0.2x |
| *(1) INT8 for LPU1* | | | | | | | | |

Source: Nvidia, Groq, Amazon, Google, Cerebras, SemiAnalysis

Nvidia, Groq, Amazon, Google, Cerebras, SemiAnalysis

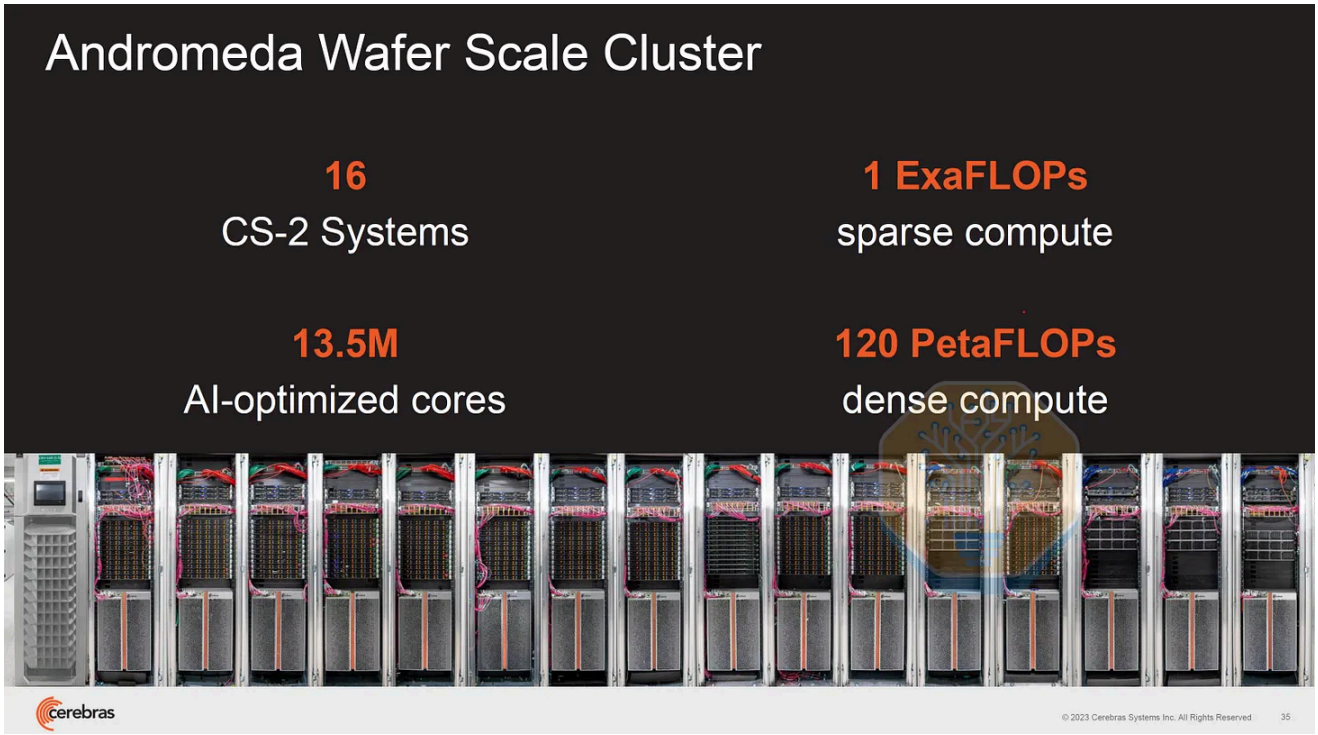
Cerebras is on its third-generation product, WSE-3, which is fabricated on TSMC's N5 node. One WSE-3 has 44GB of SRAM across a wafer or "single chip." This is a lot of SRAM. A typical large processor has on-chip SRAM in the 100s of megabytes. Even the Groq SRAM machine is only 500MB for each LPU3. SRAM is very fast, so it can deliver 21PB/s of bandwidth, thousands of times more than what HBM offers. Again, this is significantly more than the very high bandwidth Groq LPU due to the WSE having several more banks of SRAM and with the bandwidth of individual banks aggregated together.

Cerebras 目前已推出其第三代产品 WSE-3，该产品采用台积电的 N5 工艺制造。单个 WSE-3 在整块晶圆（即"单颗芯片"）上拥有 44GB 的 SRAM。这是一个巨大的 SRAM 容量。典型的超大处理器其片上 SRAM 通常仅为数百 MB。即使是 Groq 的 SRAM 机器，每块 LPU3 也仅有 500MB。由于 SRAM 速度极快，它可以提供 21PB/s 的带宽，比 HBM 提供的带宽快数千倍。同样，由于 WSE 拥有更多的 SRAM 存储体（banks），且各存储体的带宽可以聚合在一起，其带宽也显著高于带宽极高的 Groq LPU。

While Cerebras markets a lot of FLOPs for the WSE-3: 125 PFLOPs of FP16 compute, this is a sparse number, not a dense number. This is taking a page out of the Jensen Math playbook but taking it further. Unlike Nvidia, Cerebras doesn't actually state dense FLOPs in public WSE marketing materials. However, Cerebras assumes 8:1 unstructured sparsity in its sparse number, so dense FLOPS is actually 1/8$^{th}$ or 15.6 PFLOPS of FP16 compute throughput. We call this "Feldman's Formula." For the CS-

2/WSE-2 a 10:1 ratio was assumed – as we see below, the sparse and dense spec is an order of magnitude different. While WSE-3 still wins on absolute compute throughput relative to other chips, compute per silicon area is not that impressive, especially today. This is likely down to each core being much smaller than a GPU's functional array size, which is necessary for the purposes of yield harvesting, which we describe below.

虽然 Cerebras 为 WSE-3 宣传了极高的浮点运算能力（FLOPs）：125 PFLOPs 的 FP16 计算能力，但这是一个稀疏（sparse）数值，而非稠密（dense）数值。这显然是借鉴了黄仁勋（Jensen）的营销策略，甚至更进一步。与 Nvidia 不同，Cerebras 在公开的 WSE 营销材料中并未实际说明稠密 FLOPs。然而，Cerebras 在其稀疏数值中假设了 8:1 的非结构化稀疏度，因此稠密 FLOPS 实际上是 $^{th}$ 的 1/8，即 15.6 PFLOPS 的 FP16 计算吞吐量。我们称之为"费尔德曼公式（Feldman's Formula）"。对于 CS-2/WSE-2，当时假设的是 10:1 的比例——如下所示，稀疏和稠密规格相差一个数量级。虽然 WSE-3 在绝对计算吞吐量上相对于其他芯片仍具优势，但单位硅片面积的计算能力并不那么出众，尤其是在当下。这可能是因为每个核心都比 GPU 的功能阵列尺寸小得多，而这对于实现良率回收（yield harvesting）是必要的，我们将在下文详述。




Source: Cerebras at HotChips 2023

Cerebras 亮相 HotChips 2023

The last part is off-wafer networking, which stands as the weakest part of the WSE. In total there is only 150GB/s of bandwidth, a fraction of GPU/XPU competitors who place huge importance on network to scale capability. We will talk more about the implications of low I/O as well as the structural difficulty in adding more I/O.

最后一部分是片外网络，这也是 WSE 最薄弱的部分。其总带宽仅为 150GB/s，与 GPU/XPU 等竞争对手相比只是九牛一毛，而后者为了扩展性能对网络极其重视。我们将进一步探讨低 I/O 的影响，以及增加更多 I/O 在结构上存在的困难。

In summary, the WSE is a very big chip with a lot of SRAM, a decent amount of compute but not that much relative to silicon area, and almost zero network. We will now talk about the implications of this.

总而言之，WSE 是一款非常巨大的芯片，拥有海量的 SRAM 和相当不错的算力（但相对于其硅片面积而言算力并不算高），且几乎没有网络延迟。接下来我们将探讨这种架构带来的影响。

## SRAM Machines SRAM 机器

Where the WSE is clearly very strong is SRAM capacity. Like Groq's LPU, the WSE is in the class of accelerator we call "SRAM machines," where more silicon area is dedicated to super-fast SRAM, which is used as the primary memory where model weights and KV Cache are stored. In contrast, mainstream GPUs and ASICs such as TPU and Trainium use HBM to store model weights and KV Cache. They still have SRAM, just less of it. In general, trading HBM for SRAM means much higher bandwidth, lower latency and faster token output, but at the cost of capacity and therefore total throughput per {chip, watt, $}. SRAM is also just a lot more expensive per bit. Here is a chart from our recent article on NVIDIA + Groq's use of SRAM

comparing the technologies:

WSE 表现出明显强势的地方在于 SRAM 容量。与 Groq 的 LPU 类似，WSE 属于我们称之为“SRAM 机器”的加速器类别，即更多的硅片面积被分配给超高速 SRAM，并将其作为存储模型权重和 KV Cache 的主要内存。相比之下，主流 GPU 以及 TPU、Trainium 等 ASIC 则使用 HBM 来存储模型权重和 KV Cache。它们虽然也有 SRAM，但容量要小得多。通常情况下，用 HBM 换取 SRAM 意味着更高的带宽、更低的延迟和更快的 token 输出速度，但代价是容量的降低，从而影响了单位 {芯片, 瓦特, 美元} 的总吞吐量。此外，SRAM 的单位比特成本也要高得多。以下是我们最近关于 NVIDIA 和 Groq 使用 SRAM 的文章中的一张图表，对比了这些技术：

| HBM vs. DDR5 vs. GDDR7 vs. LPU SRAM | | | |
|---|---|---|---|
| Memory Type | Capacity (per GPU/XPU/LPU) | Bandwidth (per GPU/XPU/CPU) | Latency |
| HBM4 12-Hi | ~288 GB per GPU/XPU | ~22TB/s per GPU/XPU | ~100–150 ns |
| DDR5 | 128–1024 GB per CPU (~2–16 DIMMs) | ~307–614 GB/s per CPU | ~60–100 ns |
| GDDR7 | ~16–48 GB per GPU (~8–12 chips) | ~1.5–1.8 TB/s per GPU | ~50–80 ns |
| LPU SRAM | ~500 MB per LPU | ~150 TB/s per LPU | ~5–20 ns |

Source: SemiAnalysis SemiAnalysis

Even though the WSE-3's 44GB of SRAM is a huge amount of SRAM relative to any other chip, it is not much more capacity than the 36GB provided by a single stack of HBM3E 12-Hi. With the norm trending towards 8 stacks per accelerator, this is 288GB for a single GPU or TPU package (e.g. the current generation Blackwell Ultra), which is 6.5x more than the SRAM capacity of a WSE.

尽管 WSE-3 的 44GB SRAM 相对于其他任何芯片来说都是巨大的，但与单堆栈 HBM3E 12-Hi 提供的 36GB 容量相比，并没有多出多少。随着每个加速器配备 8 个堆栈成为主流趋势，单个 GPU 或 TPU 封装（例如当前一代的 Blackwell Ultra）的容量将达到 288GB，这是 WSE 的 SRAM 容量的 6.5 倍。

Some readers may have noticed that DRAM has been in demand , and a lot of it is because AI system designers are trying to pack in as much capacity as they can. More memory in a system allows model providers to:

一些读者可能已经注意到 DRAM 一直供不应求，这在很大程度上是因为 AI 系统设计者正试图尽可能多地塞进容量。系统中更多的内存允许模型提供商：

1. fit a larger model (more parameters)

1. 容纳更大的模型（更多参数）

2. serve more concurrent requests, i.e. more users (more KV Cache)

2. 服务更多并发请求，即更多用户（更多 KV 缓存）

3. support larger context windows, i.e. larger sequence lengths per request (more KV Cache)

3. 支持更大的上下文窗口，即每个请求更长的序列长度（更多的 KV 缓存）

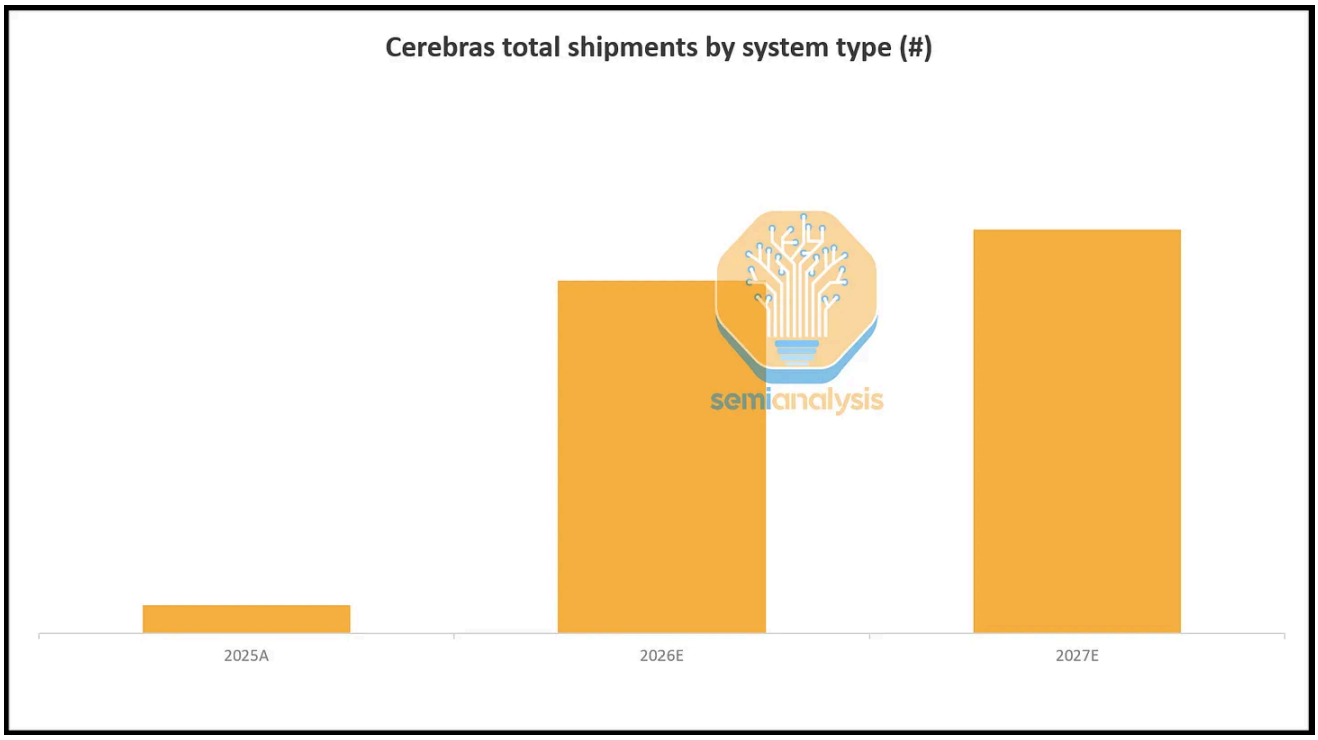
Inference providers make a business out of using all the above, which is why memory capacity per GPU is increasing. Not only that, but usable memory is not limited to a single package, since a workload can be sharded over multiple chips and aggregate memory can be pooled together within a scale up fabric. That's why networking is such a key competitive battleground for all the AI hardware companies. That is, all of them except for Cerebras who have accepted the trade-off of little network and are working around it. So, with on-wafer memory capacity limited, the escape hatch of networking more wafers together is also much narrower for Cerebras. The lack of network bandwidth, while not fatal, is certainly a handicap in the WSE-3 design preventing Cerebras from launching their business to the stratosphere.

推理服务商的业务核心就是利用上述所有特性，这也是为什么每个 GPU 的显存容量在不断增加。不仅如此，可用内存并不局限于单个封装，因为工作负载可以分片到多个芯片上，并且聚合内存可以通过扩展架构（scale up fabric）进行池化。这就是为什么网络技术成为所有 AI 硬件公司关键竞争战场的原因。也就是说，除了 Cerebras 以外的所有公司都是如此，而 Cerebras 接受了网络带宽较小的权衡，并正在设法绕过这一限制。因此，在晶圆级内存容量有限的情况下，通过网络连接更多晶圆的“逃生舱口”对 Cerebras 来说也窄得多。网络带宽的缺乏虽然不致命，但确实是 WSE-3 设计中的一个缺陷，阻碍了 Cerebras 将其业务推向更高峰。

With that said, Cerebras is now on the path to being a healthy and rapidly growing business, with its OAI deal being a game-changer: until 2028 Cerebras will need to ship an order of magnitude more servers than they have since inception. The demand surge is already visible in TSMC's wafer loadings, which step up materially each quarter through the year to meet OpenAI's deployment requirements. We expect

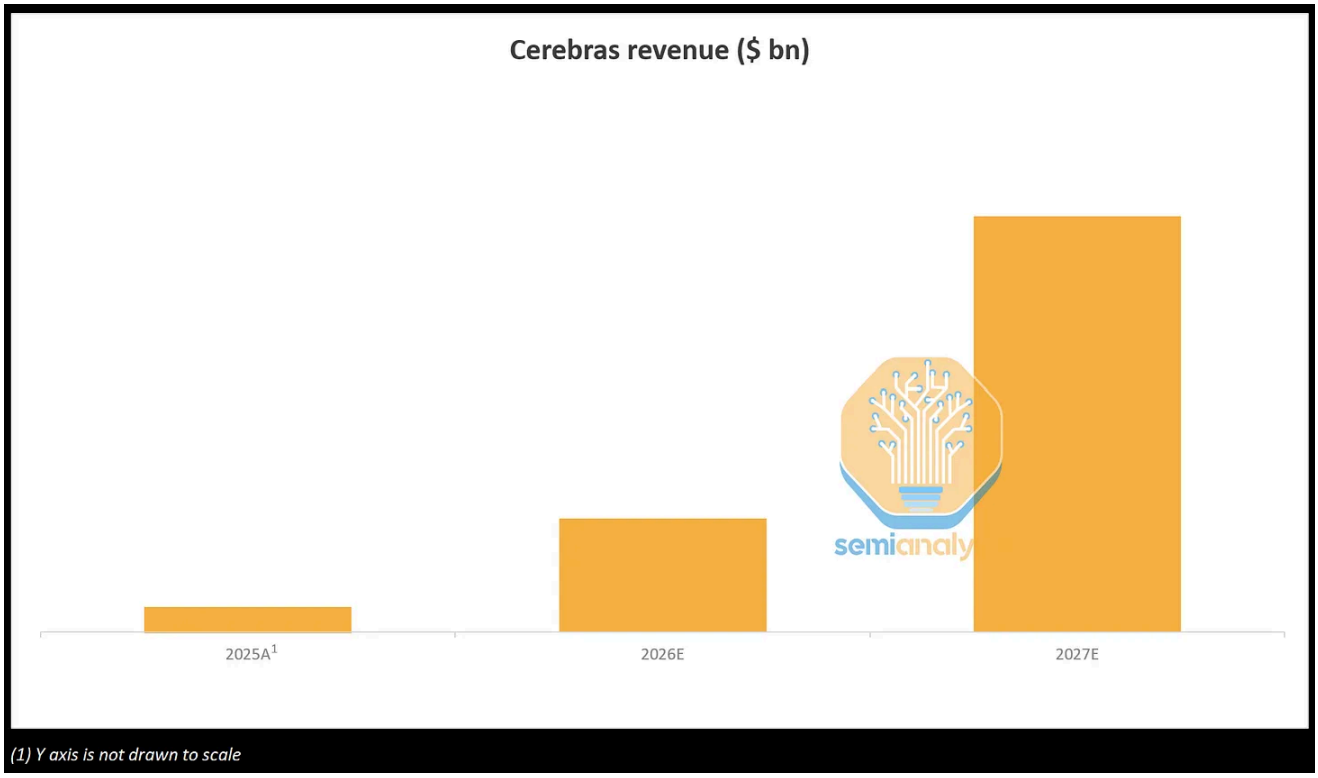
Cerebras revenue to inflect sharply in the coming years, with OpenAI as the primary growth driver.

话虽如此，Cerebras 目前正走在成为一家健康且快速增长的企业道路上，其与 OpenAI 的交易是一个转折点：到 2028 年，Cerebras 需要交付的服务器数量将比其成立以来交付的总量还要高出一个数量级。这种需求激增在台积电（TSMC）的晶圆投片量中已经显现，为了满足 OpenAI 的部署需求，投片量在年内每个季度都在实质性增加。我们预计 Cerebras 的收入将在未来几年大幅增长，而 OpenAI 将是其主要的增长驱动力。




Source: SemiAnalysis Accelerator Model

SemiAnalysis 加速器模型




Source: SemiAnalysis Accelerator Model

SemiAnalysis 加速器模型

# Cerebras's Technology Cerebras 的技术

To reach this point, Cerebras has had to solve many technical problems from silicon to system to software. To their credit, there is a lot of proprietary hardware technology here, especially when compared to the innovations (or lack of) that a lot of other accelerator startups bring to the table. The wafer is a bold bet and not easy for incumbents and competitors to replicate.

为了达到这一目标，Cerebras 必须解决从芯片、系统到软件的诸多技术难题。值得称赞的是，这里包含了大量专有的硬件技术，特别是与许多其他加速器初创公司所带来的创新（或缺乏创新）相比。晶圆级芯片是一个大胆的赌注，现有的巨头和竞争对手都难以轻易复制。

Some of Cerebras's proprietary technologies include:

Cerebras 的一些专有技术包括：

1. Cross-die wiring and routing. Cerebras uses the scribe lines as wiring for the on-wafer data fabric that connects all the dies together. In a typical wafer, these are keep

out zones where the wafer is diced to singulate individual dies.

1. 跨晶圆布线与路由。Cerebras 利用划片线作为片上数据结构的布线，将所有裸片连接在一起。在典型的晶圆中，这些区域是禁区，用于切割晶圆以分离单个裸片。

2. Redundancy and failure routing. For the purpose of having an acceptable level of yield, the ability to route through defective cores is critical. Defects are inevitable especially for near reticle-sized units. Typically, dense processors that are near reticle sized have sort yields of well below 50%. For the sake of redundancy, there are a total of 970,000 cores on the WSE, of which 900,000 are enabled. Each core is deliberately made much smaller for the sake of better yield harvesting. However, this is not simple and there is a significant additional cost required. One of the interesting things done is that **each batch** of wafers will have a custom mask set for the upper metal layers. This is for the purposes of having different wiring for each batch to route around all the defective tiles. The cost of additional masks is a material increase in cost on top of the nominal TSMC wafer cost. Why is this for every batch of wafers? This comes down to intra-batch process variation being lower than across different batches. Read here to learn more about semiconductor manufacturing process variation. The net result of this is that wafer-level yield ends up being high. Nearly 100% of the TSMC wafer output is good enough to be assembled into a production server.

2. 冗余与故障路由。为了获得可接受的良率水平，绕过缺陷核心进行路由的能力至关重要。缺陷是不可避免的，尤其是对于接近光刻掩模尺寸（reticle-sized）的单元。通常，接近光刻掩模尺寸的高密度处理器，其分选良率远低于 50%。出于冗余考虑，WSE 上共有 970,000 个核心，其中 900,000 个被启用。为了获得更好的良率，每个核心都被刻意设计得更小。然而，这并非易事，且需要巨大的额外成本。其中一个有趣的做法是，每批晶圆都会有一套针对上层金属层的定制掩模。这是为了让每批晶圆拥有不同的布线，从而绕过所有缺陷单元。额外掩模的成本在台积电（TSMC）标称晶圆成本之上增加了实质性的支出。为什么每批晶圆都要这样做？这归结于批次内工艺偏差低于跨批次偏差。点击此处了解更多关于半导体制造工艺偏差的信息。其最终结果是晶圆级良率变得很高。几乎 100% 的台积电产出晶圆都足以组装成生产级服务器。

3. Power delivery and cooling. One of the major challenges that Cerebras has solved is getting over 20KW of power into one wafer, and it will be even more next gen. This much power necessitated the need for a custom power delivery solution from Vicor. This power will of course be turned into heat that needs to be removed, which requires

specialized cooling. The power delivery and cooling sub-assembly in each CS server is called the “engine block.” This is another key component which, like the WSE silicon itself, is uniquely architected for Cerebras.

3. 供电与散热。Cerebras 解决的主要挑战之一是如何将超过 20KW 的功率输入到单个晶圆中，而下一代产品的功率还会更高。如此巨大的功率需求促使公司采用了来自 Vicor 的定制供电解决方案。这些电能自然会转化为热量并需要被移除，因此需要专门的散热系统。每个 CS 服务器中的供电与散热子组件被称为“引擎模块（engine block）”。这是另一个关键组件，与 WSE 芯片本身一样，它也是为 Cerebras 独特架构而设计的。

Despite these commendable technical achievements, the WSE architecture runs into a few technical limits that constrain their technical roadmap and ability to serve tokens.

尽管取得了这些令人赞叹的技术成就，WSE 架构仍面临一些技术限制，这些限制制约了其技术路线图以及提供 Token 服务的能力。

## Thermal Design and Cooling

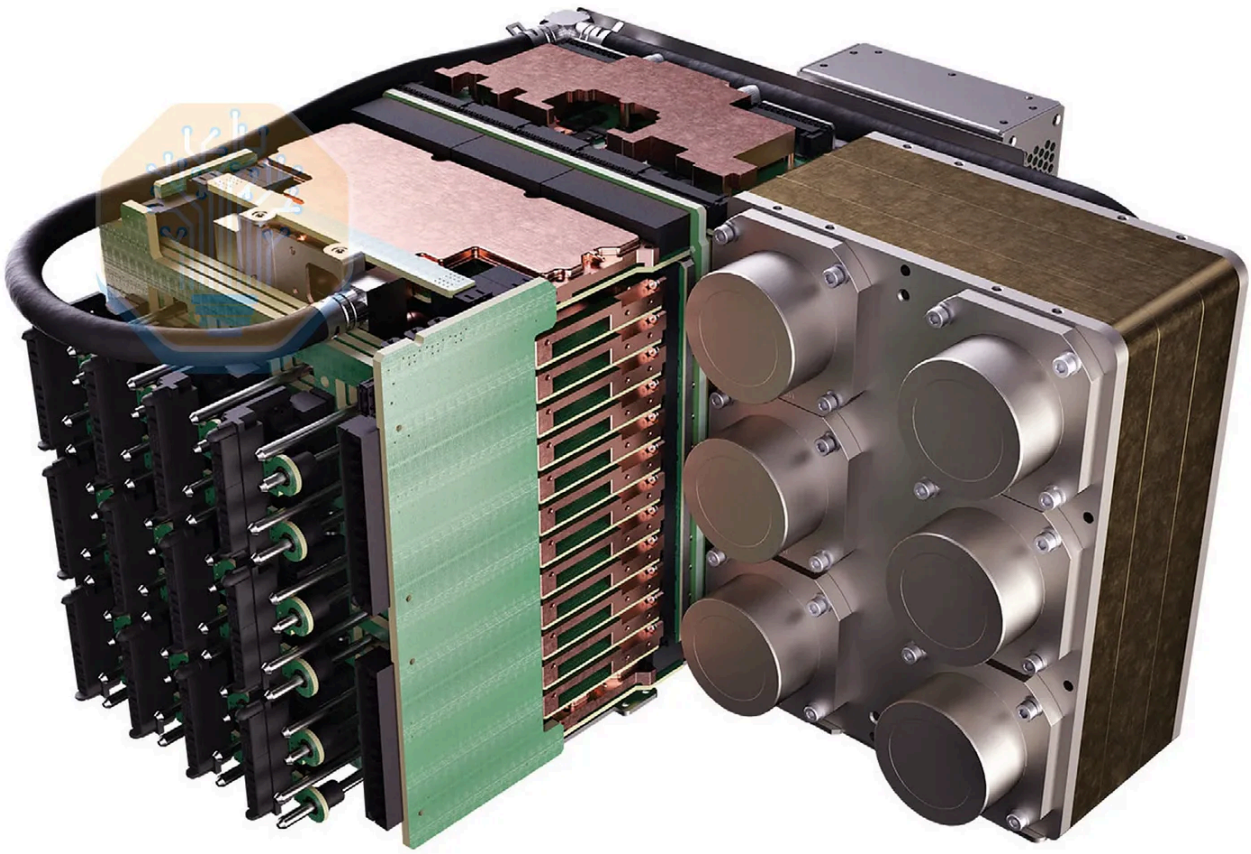
## 热设计与冷却

Cooling 25 kW in a single 46,225 $mm^2$ wafer is the central thermal problem in CS-3 design, which translates into roughly 50 $W/cm^2$ averaged across the die, before accounting for hotspots. Air cooling was rejected because a 3DVC vapor chamber heat spreader (like we see in HGX H100 servers), scaled to span the 21.5 cm die, exceeds its wick’s capillary limit and dries out before working fluid can return to the evaporator. The CS-3 uses a custom liquid-cooled stack that presents architecture, flow rates, and rack-level plumbing different from Nvidia’s more recognizable direct-to-chip single-phase deployments.

在单个 46,225 $mm^2$ 的晶圆上实现 25 kW 的散热是 CS-3 设计中的核心热管理难题。在不考虑热点的情况下，这相当于整个芯片表面平均每平方厘米约 50 W 的功耗。风冷方案被否决了，因为如果将 3DVC 真空腔均热板（如我们在 HGX H100 服务器中看到的那样）扩展到覆盖 21.5 厘米的芯片，会超过其吸液芯的毛细极限，导致工作流体在返回蒸发器之前就已干涸。CS-3 采用了一种定制的液冷堆栈，其架构、流速和机架级管路设计均不同于 Nvidia 常见的直通芯片（direct-to-chip）单相冷却部署。

The thermal solution is 100% custom and co-designed with the wafer. The silicon and the PCB underneath it expands at different rates as they heat up, and across a 21.5x21.5cm wafer that mismatch is large enough to crack a conventional package. The cold plate, the connector that bridges wafer to PCB, and the assembly tooling all had to be built from scratch. Cerebras calls its system the “engine block”, a four-layer sandwich including the cold plate, wafer, compliant connector, PCB, with the cooling manifold mated to the back of the cold plate. We will go over the system architecture in more detail in the next section.

散热方案是 100% 定制的，并与晶圆进行了协同设计。硅片及其下方的 PCB 在加热时膨胀率不同，在 21.5x21.5 厘米的晶圆上，这种失配足以使传统封装破裂。冷板、连接晶圆与 PCB 的连接器以及组装工具都必须从零开始构建。Cerebras 将其系统称为“引擎缸体”，这是一个四层夹心结构，包括冷板、晶圆、柔性连接器和 PCB，冷却歧管与冷板背面相连。我们将在下一节中更详细地介绍系统架构。



Source: Cerebras Cerebras

Heat rejection runs through the cold plate. Coolant flows through micro-fin channels machined into the back of a copper plate. The wafer-facing side of the plate is

polished and held against the silicon under preload, letting the two-slide relative to each other as they expand at different rates while maintaining contact to spread heat.

散热过程通过冷板完成。冷却液流经加工在铜板背面的微鳍片通道。冷板面向晶圆的一侧经过抛光，并在预载力作用下紧贴硅片，使两者在以不同速率膨胀时能够相对滑动，同时保持接触以传导热量。

We find another architectural challenge at the rack-to-CDU interface. The OCP/Nvidia reference design for GB200 NVL72 sizes facility-side flow at ~1.5 LPM/kW. That constant is the one the majority of today's CDU fleet is sized against. The WSE-3 runs at around ~100 LPM at 25kW, roughly 4 LPM/kW, or ~3x the NVL72 reference. That delta forces operators to use larger pumps, larger pipes, oversized CDUs, and quick-disconnects rated for higher flow. We believe that CS-4 should bring rack-level flow back toward 1.5–1.7 LPM/kW, which, if delivered, would converge Cerebras onto standardized infrastructure.

我们在机架与冷量分配单元（CDU）的接口处发现了另一个架构挑战。GB200 NVL72 的 OCP/Nvidia 参考设计将设施侧流量设定在约 1.5 LPM/kW。目前大多数 CDU 机群都是基于这一常数进行设计的。而 WSE-3 在 25kW 功率下的运行流量约为 100 LPM，即约 4 LPM/kW，约为 NVL72 参考值的 3 倍。这一差异迫使运营商必须使用更大的泵、更粗的管道、超规格的 CDU 以及额定流量更高的快速断开接头。我们认为 CS-4 应当将机架级流量降回 1.5–1.7 LPM/kW 左右，如果能实现这一点，Cerebras 将能够向标准化基础设施靠拢。

One of Cerebras's main cooling partners is LiquidStack, which Trane Technologies acquired in March 2026. LiquidStack and Cerebras initially started working on two-phase solutions, and they have jointly developed L2L single-phase CDUs sized to the CS-3's flow and pressure envelope.

Cerebras 的主要冷却合作伙伴之一是 LiquidStack，后者于 2026 年 3 月被 Trane Technologies 收购。LiquidStack 和 Cerebras 最初开始合作研发两相冷却方案，随后共同开发了根据 CS-3 的流量和压力范围定制的 L2L 单相 CDU。

Inlet temperature is a final axis where Cerebras diverges from other chips. Cerebras's Oklahoma facility runs a 6,000-ton chiller plant producing 5°C (42°F) chilled water, which is then warmed across a heat exchanger to ~21°C (~70°F) before reaching the engine block. NVL72, by contrast, is specified up to 45°C (113°F) inlet temperature,

which lets operators run free cooling for larger portions of the year. The CS-3's wafer-level heat flux requires the colder envelope, and the cost is a chiller-heavy facility.

入口温度是 Cerebras 与其他芯片迥异的最后一个维度。Cerebras 位于俄克拉荷马州的设施运行着一座 6,000 吨级的冷水机组，产生 5°C (42°F) 的冷冻水，这些水在到达引擎模块前通过热交换器升温至约 21°C (~70°F)。相比之下，NVL72 规定的入口温度最高可达 45°C (113°F)，这使得运营商在一年中的大部分时间里可以运行自然冷却。CS-3 的晶圆级热通量需要更冷的温包，其代价是设施对冷水机组的高度依赖。


Chiller Plant at Oklahoma City Datacenter. Source: Matthew Berman

俄克拉荷马城数据中心的冷水机组。来源：Matthew Berman

## The CS-3 Architecture and BOM

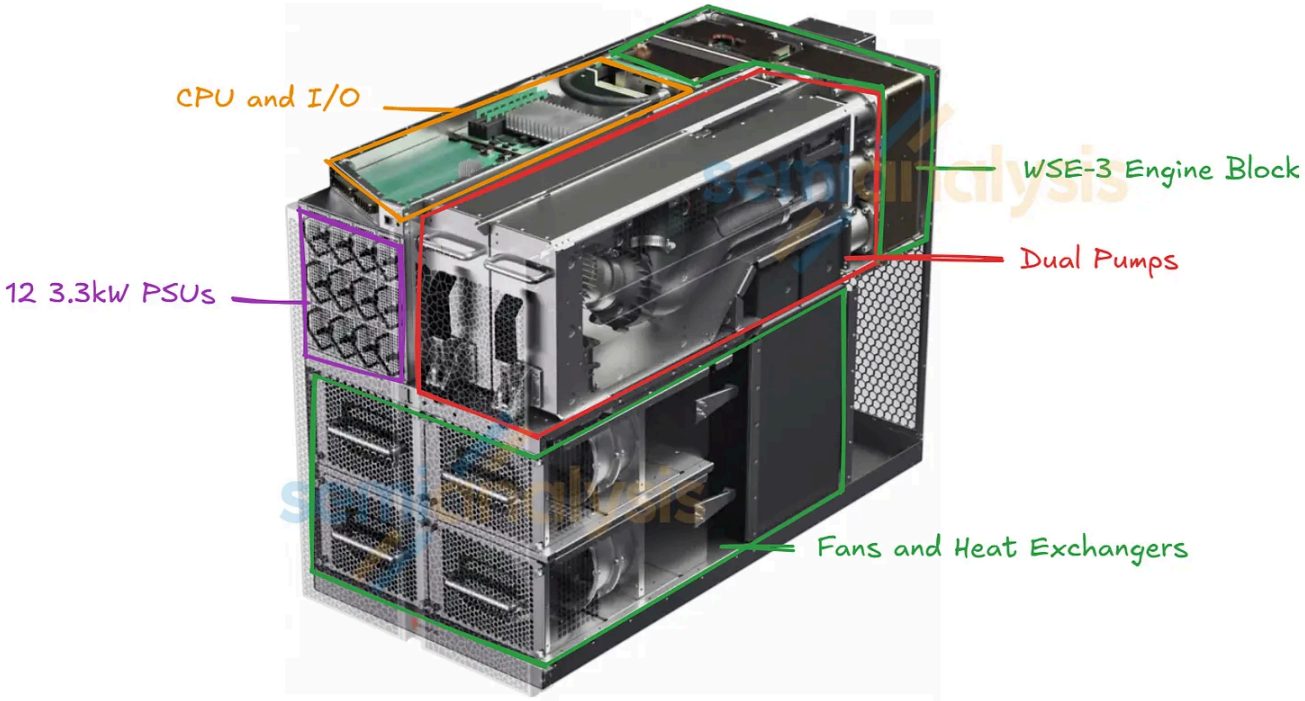
## CS-3 架构与物料清单 (BOM)

Let’s take a step back from liquid cooling and zoom out to the Cerebras CS-3 system.

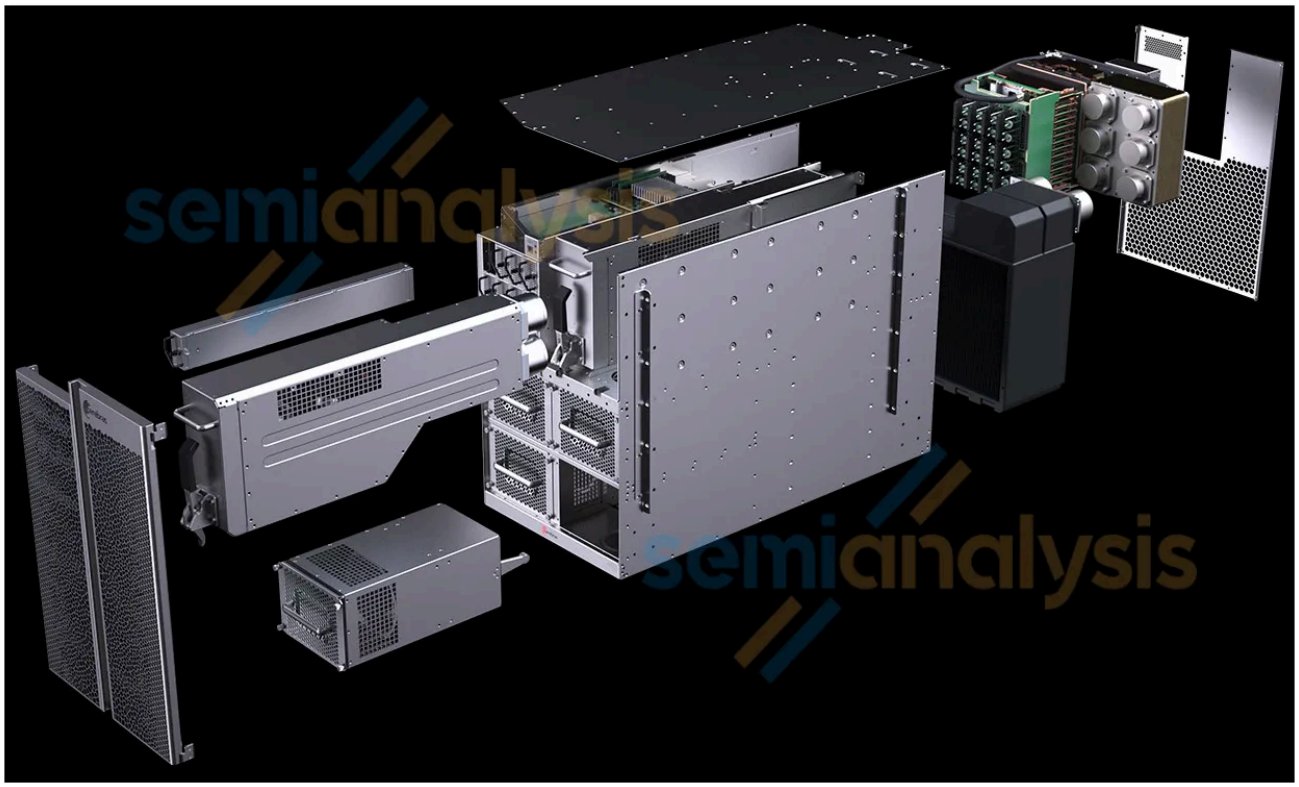
让我们暂时撇开液冷技术，从宏观视角来审视 Cerebras CS-3 系统。

Each CS-3 includes the following: **one WSE-3 engine block**, peripheral compute and I/O modules, two mechanical pumps, 12 3.3kW power supply units, and a liquid-to-air or liquid-to-liquid cooling system.

每台 CS-3 包含以下组件：一个 WSE-3 引擎模块、外围计算和 I/O 模块、两个机械泵、12 个 3.3kW 电源单元，以及一套液对气或液对液冷却系统。



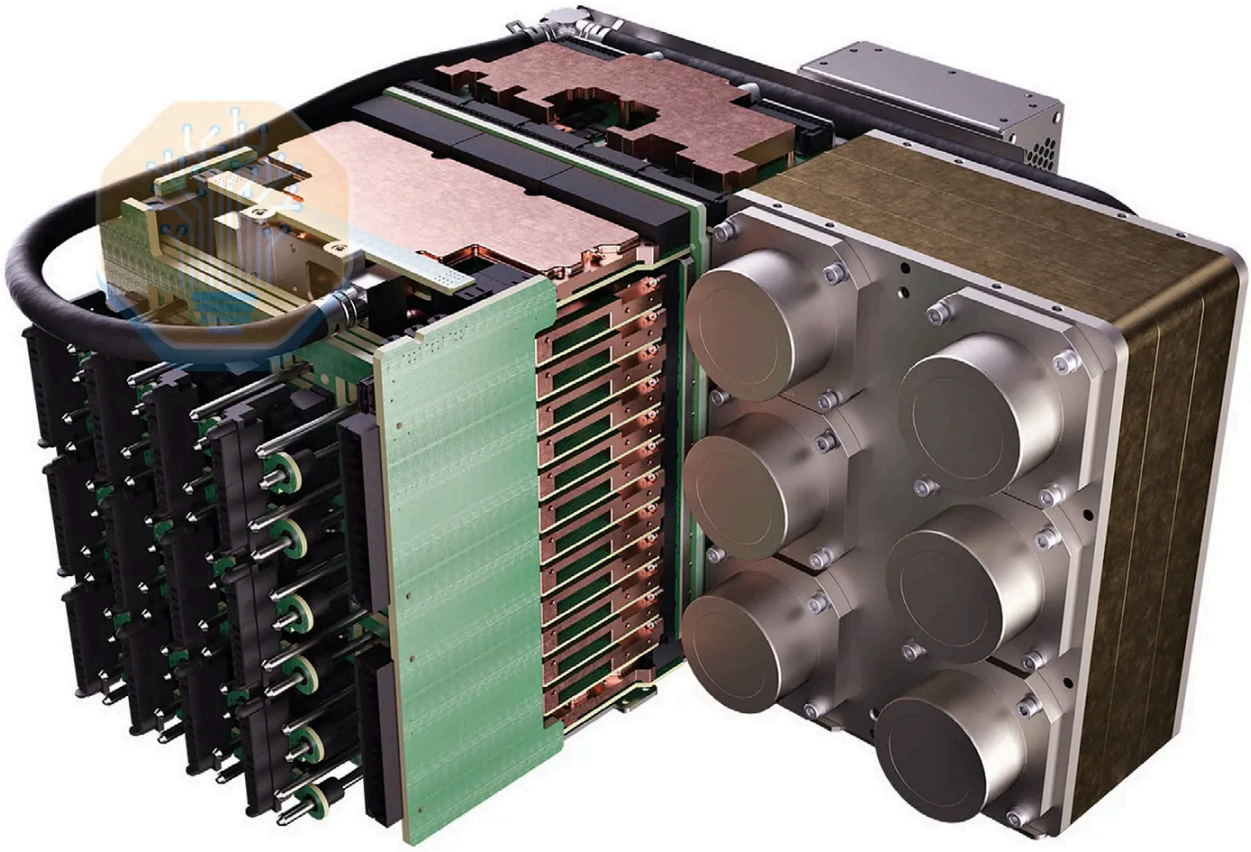

Source: Cerebras Cerebras



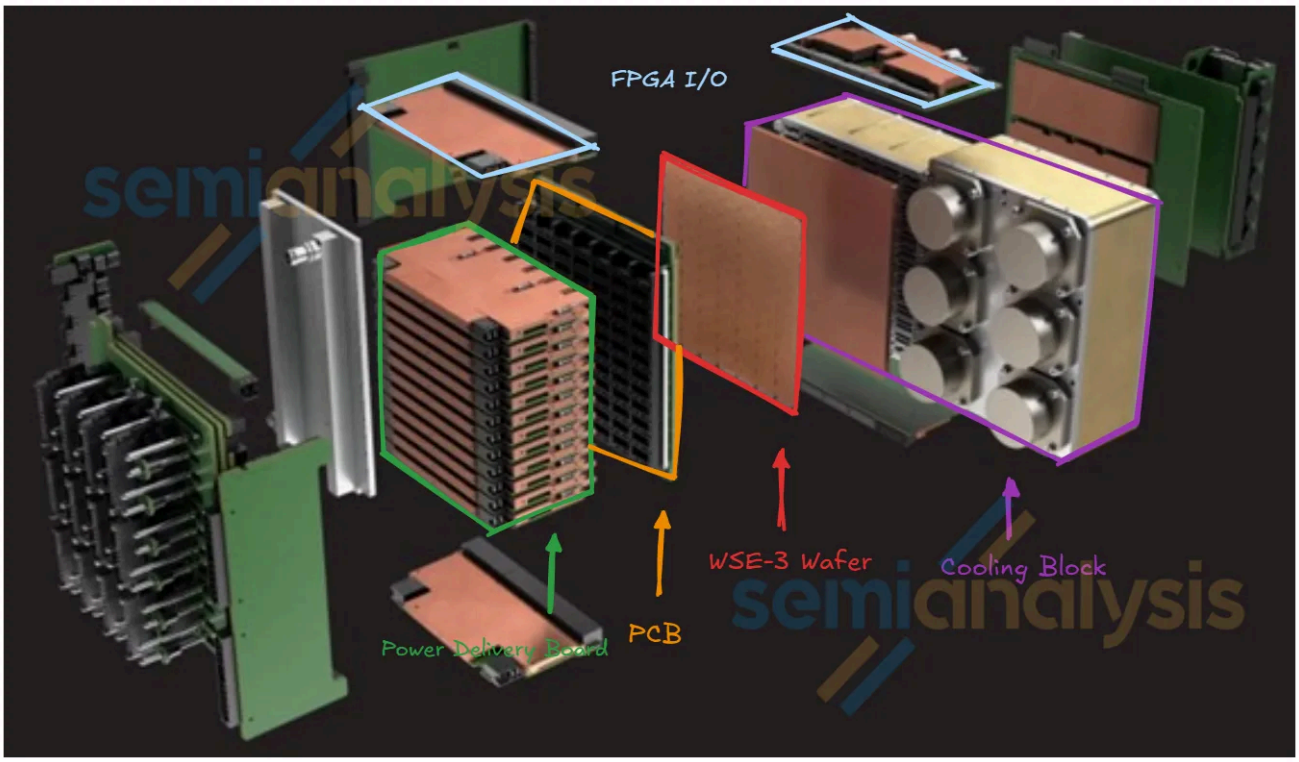

Source: Cerebras Cerebras

Zooming into the WSE-3 engine block, the WSE-3 engine takes in 25kW of power alone. Power delivery and cooling of the WSE-3 wafer is extremely customized and innovated. The power is fed into the WSE-3 engine block via the blind mated power connectors from the 12 3.3kW power supply units. The PSU delivers power at 50V to 12 PDB boards that stack on top of each other horizontally. Each PDB board matches to a row of 7 Vicor power bricks, which matches to a row of 7 blocks on the WSE-3 wafer. With 12 PDB, that is 84 power bricks and 84 blocks on the WSE-3 wafer. Then, 12V power will be delivered to Vicor's power delivery module which is on the PCB with the WSE-3 wafer on the other side, and the Vicor brick will convert the power to 1V before sending it to the wafer. The WSE-3 is socketed onto the customized PCB via an elastomer socket.

深入观察 WSE-3 引擎模块，单台 WSE-3 引擎的功耗就达到了 25kW。WSE-3 晶圆的供电和散热经过了极高度的定制与创新。电能通过盲插电源连接器，由 12 个 3.3kW 电源单元（PSU）输送到 WSE-3 引擎模块中。PSU 以 50V 的电压向 12 块水平堆叠的 PDB 板供电。每块 PDB 板对应一排 7 个 Vicor 电源模块，而这又对应 WSE-3 晶圆上的 7 个区块。通过 12 块 PDB，WSE-3 晶圆上共配置了 84 个电源模块和 84 个区块。随后，12V 电流被输送到位于 PCB 上的 Vicor 供电模块（WSE-3 晶圆位于 PCB 的另一侧），Vicor 模块在将电流送入晶圆前将其转换为 1V。WSE-3 通过弹性体插槽（elastomer socket）安装在定制的 PCB 上。



Source: Cerebras Cerebras




Source: Cerebras Cerebras

At the top of the WSE-3 engine block sits the I/O FPGA module connected to the WSE-3 PCB via board-to-board connectors. These FPGAs essentially serve as NICs that take in the Cerebras proprietary I/O off the wafer and converts it to Ethernet for scale out as well as PCIe. Customized cold plates are attached to the WSE-3 engine, the Vicor power delivery module, the CPUs, and the I/O FPGAs. The cooling loops connect to the manifold on the right side of the WSE-3 engine block. The manifolds have 6 couplings, in which 4 goes to the pump and 2 goes to the liquid-to-air or liquid-to-liquid heat removal system.

在 WSE-3 引擎组件的顶部安置着 I/O FPGA 模块，该模块通过板对板连接器与 WSE-3 PCB 相连。这些 FPGA 本质上充当网卡（NIC）的角色，负责接收来自晶圆的 Cerebras 专用 I/O 信号，并将其转换为用于横向扩展的以太网信号以及 PCIe 信号。定制的冷板安装在 WSE-3 引擎、Vicor 电源模块、CPU 以及 I/O FPGA 上。冷却回路连接到 WSE-3 引擎组件右侧的分流器。分流器配有 6 个接头，其中 4 个连接至泵，2 个连接至液对气或液对液散热系统。

In addition, each CS server has a separate 'KVSS' node. This is a dual socket AMD CPU node with 6TB of DDR5 RDIMM which is used for KVCache offload. We estimated the BoM cost of the CS-3 system and the KVSS CPU node to be $350k USD per rack before the memory price hike that started in Q4 last year. Accounting for the latest memory price hike, we have raised the estimate of the BoM of the CS-3 system and the KVSS CPU node to $450k USD per rack.

此外，每台 CS 服务器都配备了一个独立的"KVSS"节点。这是一个双路 AMD CPU 节点，配备 6TB DDR5 RDIMM，用于 KVCache 卸载。在去年第四季度开始的内存涨价之前，我们估计 CS-3 系统和 KVSS CPU 节点的物料清单（BoM）成本为每机架 35 万美元。考虑到最新的内存价格上涨，我们将 CS-3 系统和 KVSS CPU 节点的 BoM 估算成本提高到了每机架 45 万美元。

This is very high especially relative to silicon content. While nominally the accelerator silicon, usually the most expensive part of the server, is one TSMC N5 wafer that costs around $20k, there are a lot of additional costs. The requirement for masking for each wafer substantially adds to the costs. The other major BOM item is the power delivery modules from Vicor. This is a custom VRM that needs to deliver 25kW to a wafer and uses VPD. The bespoke nature of this also means a high cost, and we believe VICR's content in each WSE is not too far from TSMC's content. The same goes for the customized cooling solution. Assembly and packaging are also performed in-house by

Cerebras rather than at a contract manufacturer. There are also some peripheral components like 12x 100GbE Xilinx FPGAs that effectively act as NICs converting the Cerebras's own I/O into Ethernet for external comms.

这一比例非常高，尤其是相对于硅含量的成本而言。虽然名义上的加速器芯片（通常是服务器中最昂贵的部分）是单片台积电 N5 晶圆，成本约为 2 万美元，但还存在大量额外成本。每片晶圆都需要进行光刻掩模，这显著增加了成本。另一个主要的物料清单（BOM）项目是来自 Vicor 的供电模块。这是一种定制的电压调节模块（VRM），需要向晶圆输送 25kW 的功率并采用垂直供电（VPD）技术。这种定制化性质也意味着高昂的成本，我们认为 Vicor 在每台 WSE 中的价值占比与台积电不相上下。定制化的冷却解决方案也是如此。组装和封装也由 Cerebras 内部完成，而非交给代工厂。此外还有一些外围组件，例如 12 颗 Xilinx 100GbE FPGA，它们实际上充当网卡（NIC），将 Cerebras 自有的 I/O 转换为以太网以进行外部通信。

**Final BoM to Cerebras**

Analysis represents the bill of materials totaling to price paid by Cerebras

| Item | Item Category | Quantity | Unit Cost | Extended Cost |
|---|---|---|---|---|
| WSE-3 Engine Block | | | | |
| WSE-3 Module | | | | |
| FPGA I/O Module | | | | |
| Power | | | | |
| Cooling Distribution | | | | |
| Mechanical | | | | |
| Compute Tray - In-House Assembly and Testing | Assembly and Testing | | | |
| Other Peripheral Modules | | | | |
| I/O and Management Module | | | | |
| CPU Head Node Module (Attached) | | | | |
| CPU Head Node Module | | | | |
| Mechanical | | | | |
| KVSS Server (Attached) | | | | |
| KVSS Server Module | | | | |
| Mechanical | | | | |
| Chassis Level | | | | |
| Power Delivery | | | | |
| Cooling Distribution | | | | |
| Mechanical | | | | |
| Rack Level | | | | |
| Mechanical | | | | |
| Rack Level - In-House Assembly and Testing | Assembly and Testing | | | |
| Total BoM and Power Budget of [Cerebras CS-3 (Pre-Memory Price Hike)] | | | | $341,962 |

Source: SemiAnalysis Estimates

SemiAnalysis 估算

**Final BoM to Cerebras**

Analysis represents the bill of materials totaling to price paid by Cerebras

| Item | Item Category | Quantity | Unit Cost | Extended Cost |
|---|---|---|---|---|
| WSE-3 Engine Block | | | | |
| WSE-3 Module | | | | |
| FPGA I/O Module | | | | |
| Power | | | | |
| Cooling Distribution | | | | |
| Mechanical | | | | |
| Compute Tray - In-House Assembly and Testing | Assembly and Testing | | | |
| Other Peripheral Modules | | | | |
| I/O and Management Module | | | | |
| CPU Head Node Module (Attached) | | | | |
| CPU Head Node Module | | | | |
| Mechanical | | | | |
| KVSS Server (Attached) | | | | |
| KVSS Server Module | | | | |
| Mechanical | | | | |
| Chassis Level | | | | |
| Power Delivery | | | | |
| Cooling Distribution | | | | |
| Mechanical | | | | |
| Rack Level | | | | |
| Mechanical | | | | |
| Rack Level - In-House Assembly and Testing | Assembly and Testing | | | |
| Total BoM and Power Budget of [Cerebras CS-3] | | | | $464,484 |

Source: SemiAnalysis Estimates

SemiAnalysis 估算

# Where the Wafer Wins 晶圆级芯片的优势所在

To understand the extremely high memory bandwidth of Cerebras in context, one must put on the hat of a performance engineer working on LLM inference. To performance engineers, a chip is a tool. Whether you are using 10,000 LPUs, 72 GPUs, or 1 wafer to get the job done, what matters is the "arithmetic intensity" of the chip – how many FLOPs the chip can perform for every byte it transfers to/from memory (FLOPs/byte). Below is a table of chip specs to put the WSE-3 in context. Note that these are theoretical maximum numbers.

为了深入理解 Cerebras 极高内存带宽的背景，必须站在负责 LLM 推理的性能工程师的角度来思考。对于性能工程师而言，芯片就是一种工具。无论你是使用 10,000 个 LPU、72 个 GPU 还是 1 片晶圆来完成任务，真正重要的是芯片的"算术强度"——即芯片每从内存传输一个字节所能执行的浮点运算次数（FLOPs/byte）。下表列出了芯片规格，以便将 WSE-3 置于具体语境中进行对比。请注意，这些均为理论最大值。

| Cerebras vs Others (chips) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | FP16 or BF16 perf | FP8 or INT8 perf | FP4 perf | HBM capacity | HBM bandwidth | HBM perf ratio | SRAM Capacity | SRAM bandwidth | SRAM perf ratio |
| H100 | 0.989 PFLOPS | 1.979 PFLOPS | - | 80 GB | 3.35 TB/s | 591 | 50 MB | 12.8 TB/s | 155 |
| H200 | 0.989 PFLOPS | 1.979 PFLOPS | - | 141 GB | 4.80 TB/s | 412 | 50 MB | 12.8 TB/s | 155 |
| B200 | 2.25 PFLOPS | 4.5 PFLOPS | 9 PFLOPS | 192 GB | 8 TB/s | 1125 | 126 MB | 20 TB/s | 450 |
| B300 | 2.25 PFLOPS | 4.5 PFLOPS | 13.5 PFLOPS | 288 GB | 8 TB/s | 1688 | 126 MB | 20 TB/s | 675 |
| Cerebras WSE-3 | 15.625 PFLOPS | 15.625 PFLOPS | - | - | - | - | 44 GB | 21000 TB/s | 0.74 |
| Groq LP30 | 0.6 PFLOPS | 1.2 PFLOPS | - | - | - | - | 500 MB | 150 TB/s | 8 |
| R200 | 4 PFLOPS | 17.5 PFLOPS | 35 PFLOPS | 288 GB | 13 TB/s | 2692 | ? | ? | ? |
| * perf ratio also known as ridgepoint Arithmetic Intensity, i.e. FLOPs/bw | | | | | | | | | |

Source: public datasheets from NVIDIA, Groq, and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

On a relative basis, the performance of AI applications depends on the performance of individual kernels (i.e. software that runs on the device, not the host CPU) on these chips. The canonical example of a kernel used in AI is GEMMs (general matrix multiplication). GEMMs can have different shapes, dictated by the shapes of the matrices being multiplied. Certain shapes running on specific hardware can be memory bound (i.e. performance is limited by the available bandwidth), or compute

bound (i.e. performance is limited by the available FLOPs).

从相对角度来看，AI 应用的性能取决于这些芯片上单个算子（即运行在设备端而非主机 CPU 上的软件）的性能。AI 中最典型的算子示例是 GEMM（通用矩阵乘法）。GEMM 具有不同的形状，这由相乘矩阵的形状决定。在特定硬件上运行的某些形状可能是内存受限的（即性能受限于可用带宽），或者是计算受限的（即性能受限于可用算力 FLOPs）。

It is striking to see the FLOPs of a WSE-3 compared like-for-like with NVIDIA GPUs. In terms of dense FP16 or INT8 FLOPS (the actual FLOPs that developers using a Cerebras WSE use), an entire WSE-3 is only capable of 15.625 PFLOPS. Compared to NVIDIA GPUs running native FP4, B300 comes in at 13.5 PFLOPS (or 15 for GB300) and the Rubin GPU has 35PFLOPS. Of course, the astute reader will note that FP4 FLOPs and FP16 FLOPs are not always comparable, but with most production inference today shifting to FP4, it's the best real-world comparison. Astute readers should also note the impact of Cerebras product marketing. Cerebras marketing materials, as well as their S1, claim much higher PFLOPs per wafer than our table shows. Thanks to the "Feldman Formula", they use a factor of 8x (claiming 8:1 unstructured sparsity) to get there. An even bigger sparsity factor than the hallmark 2:1 rule of Jensen Math!

将 WSE-3 的 FLOPs 与 NVIDIA GPU 进行同类对比，结果令人震惊。就稠密 FP16 或 INT8 FLOPS（即使用 Cerebras WSE 的开发者实际使用的 FLOPs）而言，整个 WSE-3 仅能提供 15.625 PFLOPS。相比之下，运行原生 FP4 的 NVIDIA GPU 中，B300 可达 13.5 PFLOPS（GB300 为 15 PFLOPS），而 Rubin GPU 则拥有 35 PFLOPS。当然，敏锐的读者会注意到 FP4 FLOPs 和 FP16 FLOPs 并不总是具有可比性，但随着当今大多数生产级推理转向 FP4，这是最符合现实世界的对比方式。敏锐的读者还应注意到 Cerebras 产品营销的影响。Cerebras 的营销材料及其 S1 文件声称的每片晶圆 PFLOPs 远高于我们表中所列的数值。得益于"Feldman 公式"，他们通过使用 8 倍系数（声称 8:1 的非结构化稀疏度）来达到这一数值。这甚至比 Jensen Math 标志性的 2:1 规则所使用的稀疏因子还要大！

To compare Cerebras to alternatives, it is not useful to compare directly, chip-to-chip (or wafer-to-chip). We illustrate a more useful comparison below, with round numbers,

to demonstrate where the wafer fits in.

要将 Cerebras 与替代方案进行比较，直接进行芯片对芯片（或晶圆对芯片）的对比是徒劳的。我们在下方通过整数示例展示了一种更有意义的比较方式，以说明晶圆在其中的地位。

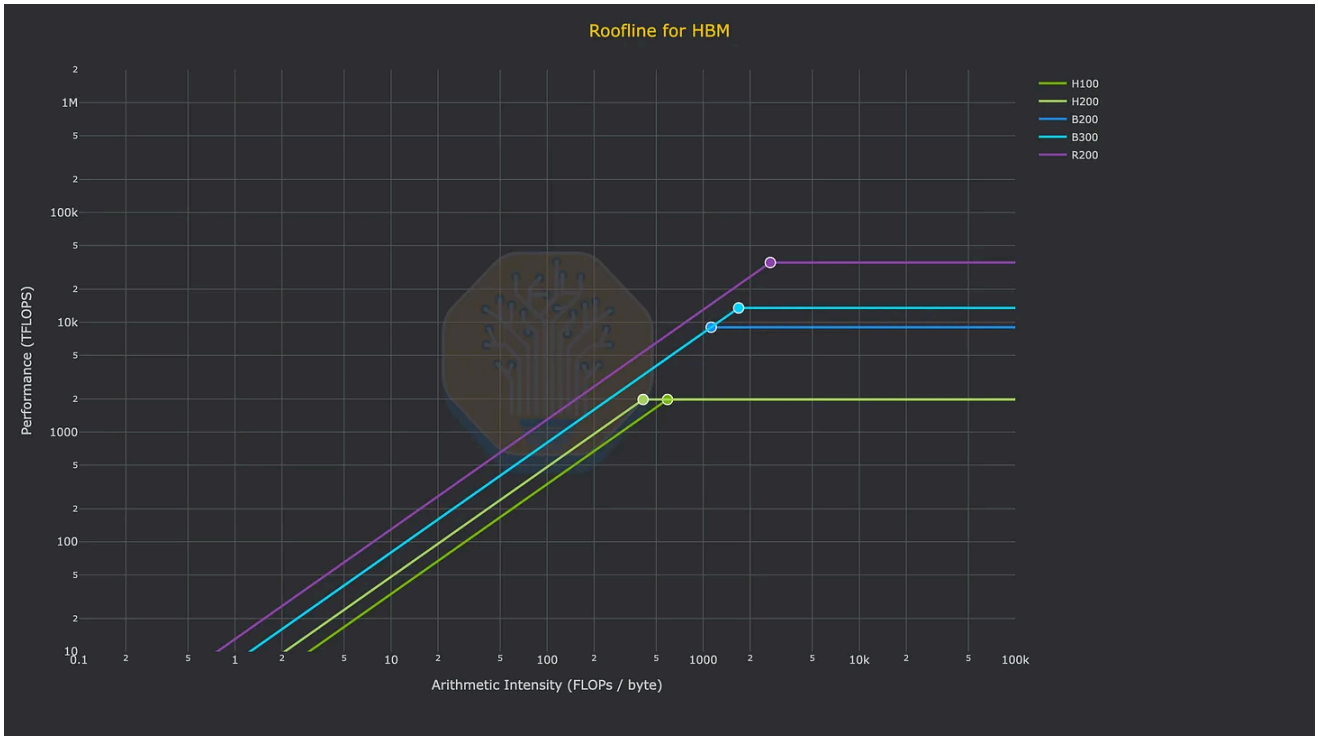
| Cerebras vs Others (chips + systems) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | FP16 or BF16 perf | FP8 or INT8 perf | FP4 perf | HBM capacity | HBM bandwidth | HBM perf ratio | SRAM Capacity | SRAM bandwidth | SRAM perf ratio | rough price |
| **H100** | 0.989 PFLOPS | 1.979 PFLOPS | - | 80 GB | 3.35 TB/s | 591 | 50 MB | 12.8 TB/s | 155 | $ 35,000 |
| **H200** | 0.989 PFLOPS | 1.979 PFLOPS | - | 141 GB | 4.80 TB/s | 412 | 50 MB | 12.8 TB/s | 155 | $ 40,000 |
| **B200** | 2.25 PFLOPS | 4.5 PFLOPS | 9 PFLOPS | 192 GB | 8 TB/s | 1125 | 126 MB | 20 TB/s | 450 | $ 50,000 |
| **B300** | 2.25 PFLOPS | 4.5 PFLOPS | 13.5 PFLOPS | 288 GB | 8 TB/s | 1688 | 126 MB | 20 TB/s | 675 | $ 55,000 |
| **Cerebras WSE-3** | 15.625 PFLOPS | 15.625 PFLOPS | - | - | - | - | 44 GB | 21000 TB/s | 0.74 | $ 1,000,000 |
| **Groq LP30** | 0.6 PFLOPS | 1.2 PFLOPS | - | - | - | - | 500 MB | 150 TB/s | 8 | $ 20,000 |
| **8x H100 (DGX system)** | 8 PFLOPS | 16 PFLOPS | - | 1128 GB | 27 TB/s | 591 | 400 MB | 102 TB/s | 155 | $ 280,000 |
| **8x B300 (DGX system)** | 18 PFLOPS | 36 PFLOPS | 108 PFLOPS | 2304 GB | 64 TB/s | 1688 | 1008 MB | 160 TB/s | 675 | $ 400,000 |
| **72x GB300 NVL72 (rack)** | 162 PFLOPS | 324 PFLOPS | 1080 PFLOPS | 20736 GB | 576 TB/s | 1875 | 9072 MB | 1440 TB/s | 750 | $ 3,960,000 |
| * perf ratio also known as ridgepoint Arithmetic Intensity, i.e. FLOPs/bw | | | | | | | | | | |

Source: public datasheets from NVIDIA, Groq, and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

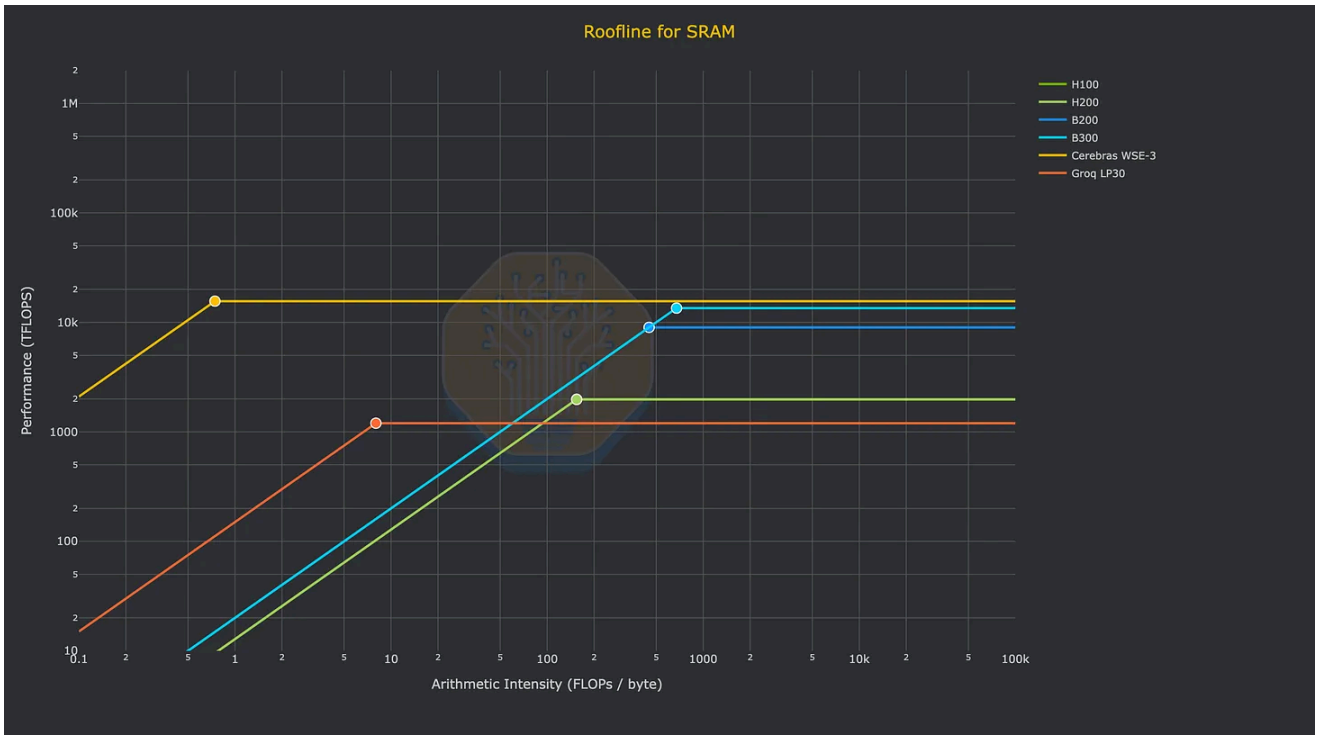
It is most instructive to compare a single wafer's worth of cost and performance to around $1M worth of hardware on both HBM and SRAM. Namely: 2x NVIDIA HGX systems (16 GPUs), 4x NVL72 sleds (16 GPUs), or around 50x Groq LP30s. So, we will progressively add more rooflines to the plot in the following charts.

将单片晶圆的成本和性能与价值约 100 万美元的 HBM 和 SRAM 硬件进行对比是最具启发性的。具体而言，这些硬件包括：2 套 NVIDIA HGX 系统（16 颗 GPU）、4 个 NVL72 机架单元（16 颗 GPU），或大约 50 张 Groq LP30 加速卡。因此，我们将在接下来的图表中逐步向图中添加更多性能上限线（rooflines）。




Source: public datasheets from NVIDIA, Groq and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

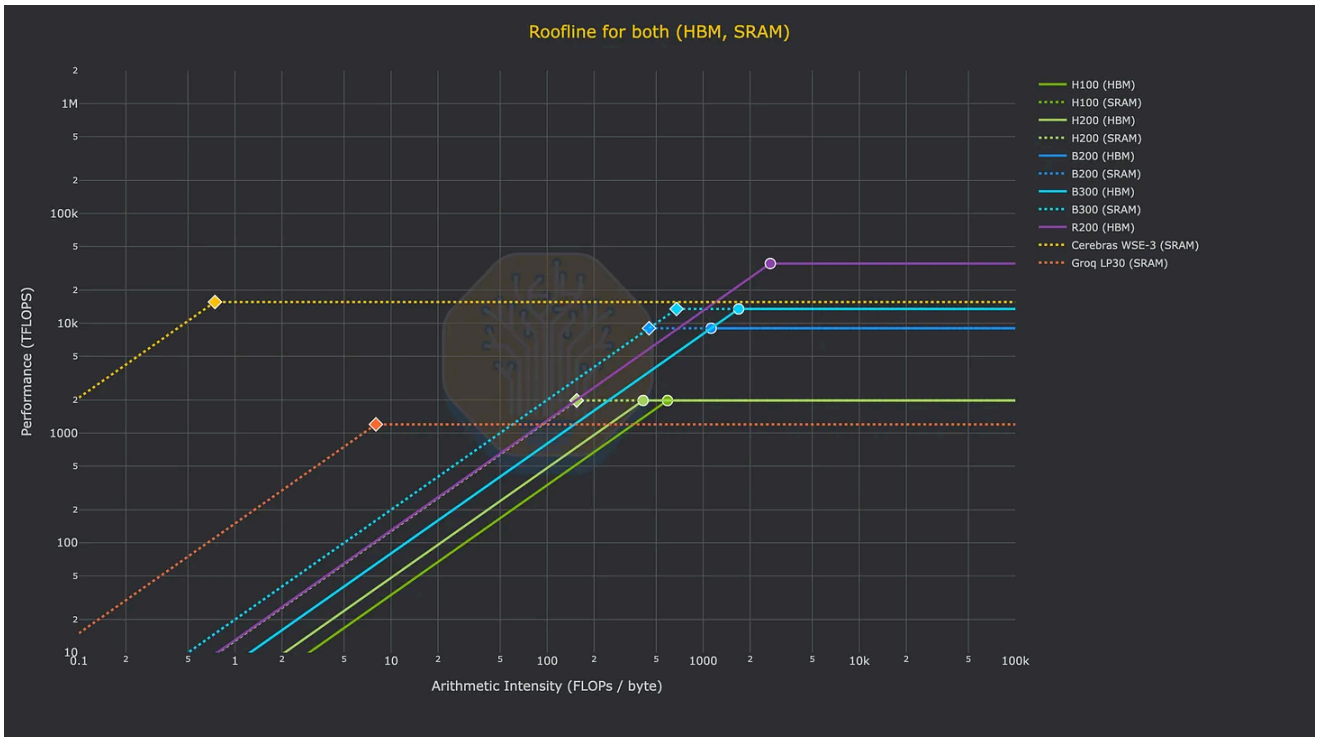



Source: public datasheets from NVIDIA, Groq and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

Here we see a single Nvidia Rubin GPU FLOP mogging an entire WSE-3:

在这里，我们看到单块 Nvidia Rubin GPU 的算力（FLOP）正以压倒性优势胜过整块 WSE-3：

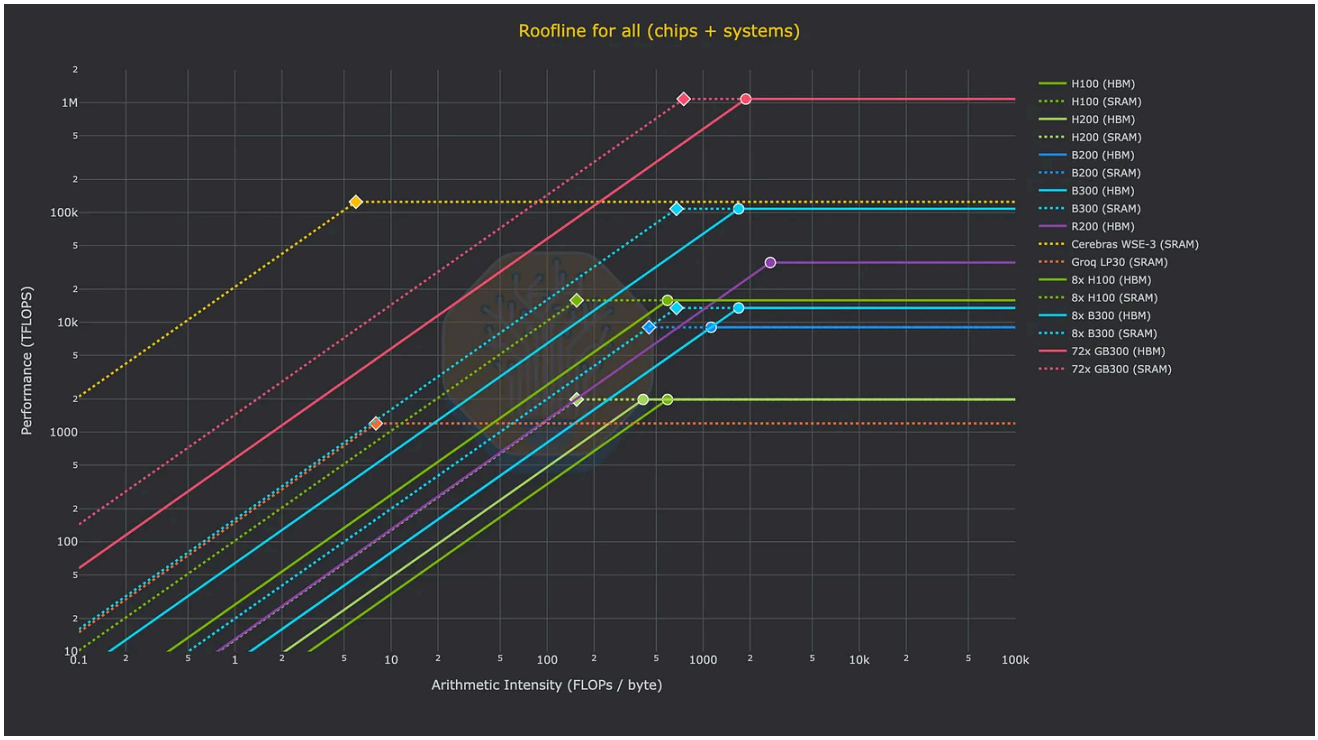



Source: public datasheets from NVIDIA, Groq and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

Finally, this chart demonstrates how this analysis can be extended to the system level (albeit in a naive way), comparing the roofline of a single Wafer's SRAM to DGX systems and a GB300 NVL72 rack. One has to assume zero network overhead and add many racks of GB300 NVL72 just to be able to realize the same FLOPs as Cerebras on kernels with equivalent arithmetic intensity.

最后，这张图表展示了如何将该分析扩展到系统层面（尽管是以一种简化的方式），将单片晶圆（Wafer）的 SRAM 屋顶线（roofline）与 DGX 系统以及 GB300 NVL72 机架进行对比。人们必须假设网络开销为零，并增加许多个 GB300 NVL72 机架，才能在具有等效算术强度的算子上实现与 Cerebras 相同的 FLOPs。

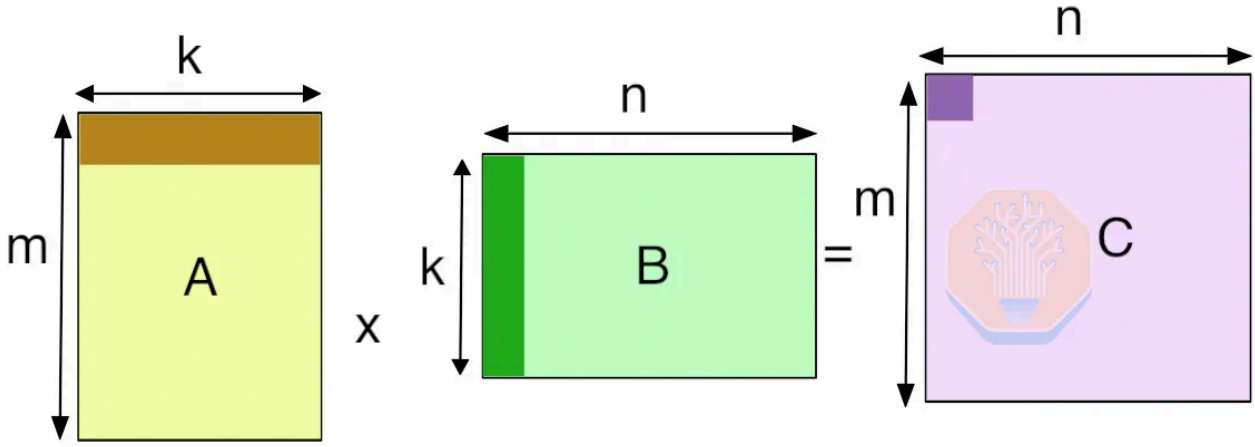



Source: public datasheets from NVIDIA, Groq and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

To finish with a complete understanding of which AI workloads are a good fit for Cerebras, we can just look at common GEMM shapes. GEMMs generally use “mnk” notation, meaning that the input matrices have size “m” and “n” respectively, with a contracting dimension of “k”.

为了全面了解哪些 AI 工作负载适合 Cerebras，我们只需观察常见的 GEMM（通用矩阵乘法）形状。GEMM 通常使用“mnk”表示法，这意味着输入矩阵的大小分别为“m”和“n”，且具有一个收缩维度“k”。



Source: Pete Warden Pete Warden

We can calculate the Arithmetic Intensity of a given GEMM using the following formula:

我们可以使用以下公式计算给定 GEMM 的算术强度：

$$\text{For } C_{M\times N} = A_{M\times K} \cdot B_{K\times N} \text{ in single precision, with bytes per element } b:$$

$$\text{FLOPs} = 2 \cdot M \cdot N \cdot K$$

$$\text{Bytes} = (M \cdot K + K \cdot N + M \cdot N) \cdot b$$

$$\text{assuming all reads/writes go through DRAM}$$

$$\text{AI} = \frac{2 \cdot M \cdot N \cdot K}{(M \cdot K + K \cdot N + M \cdot N) \cdot b} \quad \text{FLOPs/byte}$$

$$\text{For square } M = N = K = n: \qquad \text{AI} = \frac{2n^3}{3n^2 b} = \frac{2}{3} \cdot \frac{n}{b}$$

$$\text{FP8 } (b = 1): \ \text{AI} \approx 0.67n$$
$$\text{BF16 } (b = 2): \ \text{AI} \approx 0.33n$$
$$\text{FP4 } (b = 0.5): \ \text{AI} \approx 1.33n$$

For reference, here are some example GEMM shapes used in LLM inference:

作为参考，以下是 LLM 推理中常用的一些 GEMM 形状示例：

| example GEMM shapes for LLM inference | | | | | | |
|---|---|---|---|---|---|---|
| | Shape (M, N, K) | | | FP8 perf ratio* | HBM bound | FLOPs bound (FP8) |
| **decode batch=1** | 1 | 8192 | 8192 | 2 | all | none |
| **decode batch=32** | 32 | 8192 | 8192 | 64 | H100, H200, B200, B300, R200 | Cerebras, Groq |
| **MoE expert** | 128 | 4096 | 4096 | 241 | H100, H200, B200, B300, R200 | Cerebras, Groq |
| **square 512** | 512 | 512 | 512 | 341 | H100, H200, B200, B300, R200 | Cerebras, Groq |
| **square 1k** | 1024 | 1024 | 1024 | 683 | R200 | Cerebras, Groq, H100, H200, B200, B300 |
| **square 2k** | 2048 | 2048 | 2048 | 1365 | none | all |
| **square 4k** | 4096 | 4096 | 4096 | 2731 | none | all |
| **prefill** | 4096 | 8192 | 8192 | 4096 | none | all |

* perf ratio also known as Arithmetic Intensity, i.e. FLOPs/bw

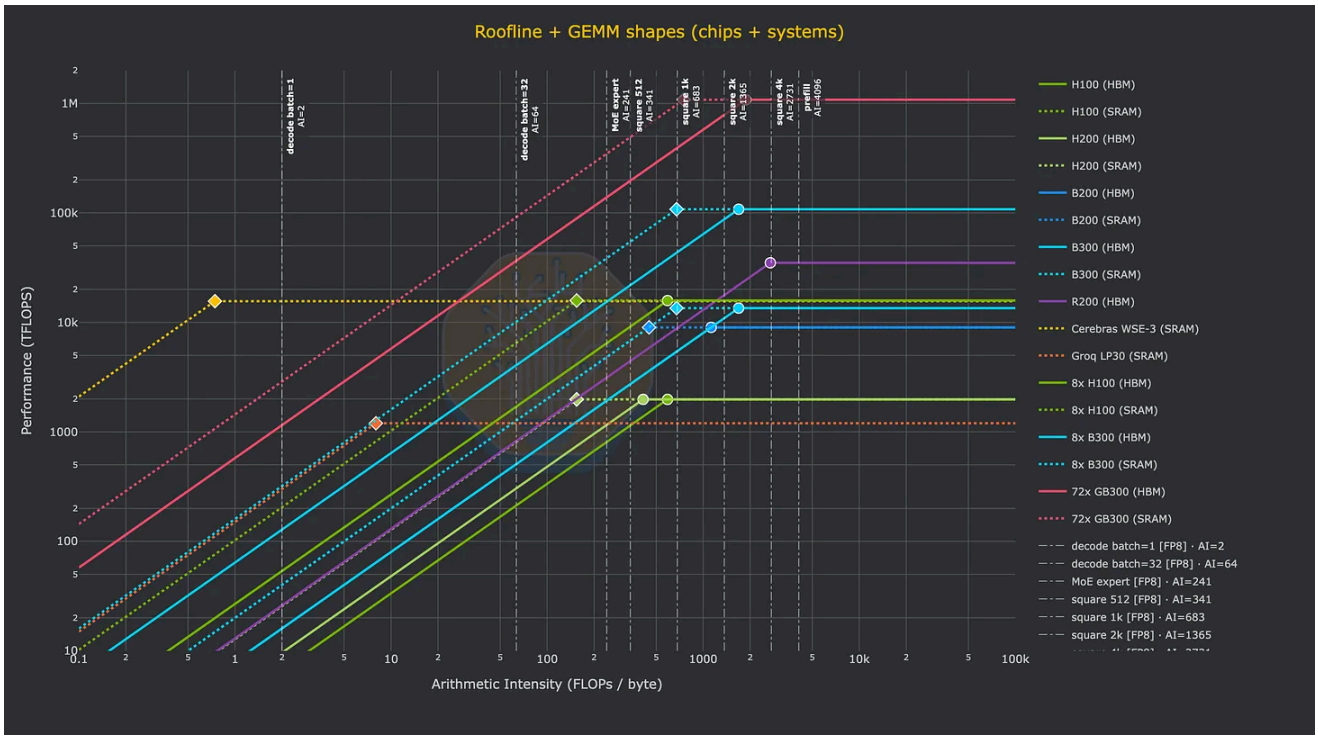
Source: public datasheets from NVIDIA, Groq and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

And finally, here is how those kernels would theoretically perform on different chips. Just trace from bottom to top on one of the vertical lines that represent the arithmetic intensity of a given kernel to see the (theoretical) performance that a given chip will be

able to realize on that GEMM shape (measured in TFLOPs).

最后，以下是这些算子在不同芯片上的理论性能表现。只需在代表给定算子算术强度的垂直线上从下往上追踪，即可查看给定芯片在该 GEMM 形状上能够实现的（理论）性能（以 TFLOPs 为单位）。




Source: public datasheets from NVIDIA, Groq and Cerebras

来源：NVIDIA、Groq 和 Cerebras 的公开数据表

At a high level, it is clear that Cerebras has very unique performance characteristics, with an optimal arithmetic intensity of 0.74 on the WSE-3's SRAM and FP16 or INT8 FLOPs. With HBM-based GPUs going the other direction over time, i.e. an arithmetic intensity increasing to over 1000, there is a clear difference between the GEMM shape (or more generally, which kernels) will make the most effective use of Cerebras hardware.

从宏观层面来看，显而易见的是 Cerebras 具有非常独特的性能特性，其 WSE-3 的 SRAM 以及 FP16 或 INT8 算力在算术强度为 0.74 时达到最优。随着基于 HBM 的 GPU 随时间推移向另一个方向发展（即算术强度增加到 1000 以上），GEMM 形状（或更广泛地说，哪些算子）能最有效地利用 Cerebras 硬件，两者之间存在明显的差异。

For the reader to get a sense of what the realized FLOPs looks like for a given decode kernel, just imagine a decode kernel with (m=batch=1) and arithmetic intensity of (AI=2). This is the leftmost vertical bar on the previous chart. As you trace your finger from bottom to top on that line you will cross many chips before you reach Cerebras: all NVIDIA GPUs and Groq LPUs will only be able to realize dozens or hundreds of TFLOPs in an absolute max, theoretical case. Meanwhile, the Cerebras wafer can (again, theoretically) realize its full 15.625 PFLOPs. This is the key point of the wafer. Absolutely massive amounts of memory bandwidth from the 44GB of SRAM on the wafer mean that decode kernels can realize equally massive amounts of performance.

为了让读者了解特定解码算子（decode kernel）的实际算力表现，只需想象一个（m=batch=1）且算术强度为（AI=2）的解码算子。这就是前一张图表中最左侧的垂直柱状线。当你沿着那条线从下往上移动手指时，在到达 Cerebras 之前你会经过许多芯片：所有 NVIDIA GPU 和 Groq LPU 在绝对理想的理论情况下，也只能实现几十或几百 TFLOPs 的算力。与此同时，Cerebras 晶圆（同样在理论上）可以实现其全部 15.625 PFLOPs 的算力。这就是该晶圆的关键所在。晶圆上 44GB SRAM 提供的极其巨大的内存带宽，意味着解码算子可以实现同样巨大的性能。

Going back to our job as a performance engineer, this means that decode kernels with low arithmetic intensity have a much higher theoretical limit in terms of the amount of FLOPs that can be realized. The SRAM bandwidth can keep up with the compute, while the HBM of a GPU running the same kernel leaves Blackwell SM100 FP4 Tensor Cores starving. And as a result, the types of models and workloads that will be designed to run on the Cerebras WSE-3 in the future, such as GPT-5.3-Codex-Spark (with an architecture that also goes by the name of gptoss-120b), will be developed with the performance characteristics of the wafer in mind.

回到我们作为性能工程师的工作，这意味着算术强度较低的解码内核在可实现的 FLOPs 数量方面具有更高的理论极限。SRAM 带宽可以跟上计算速度，而运行相同内核的 GPU HBM 则会让 Blackwell SM100 FP4 Tensor Cores 处于饥饿状态。因此，未来为 Cerebras WSE-3 设计运行的模型和工作负载类型，例如 GPT-5.3-Codex-Spark（其架构也被称为 gptoss-120b），将在开发时充分考虑晶圆的性能特性。

A perfect example of hardware-software co-design.

硬件与软件协同设计的完美范例。

# The Wafer Taketh and the Wafer Giveth

# 晶圆之失与晶圆之得

The WSE has several clear weaknesses that we have mentioned. It has a lot of SRAM, but given SRAM is inherently not dense on a per-watt or per-dollar basis, HBM-based GPUs and XPUs offer far more memory capacity per watt or dollar. This HBM is currently used to serve larger models with longer context length, as well as more batching of users to drive throughput. Networking more wafers together to overcome the lack of memory per wafer is also constrained by the lack of off-wafer bandwidth. Absent a heroic technical achievement (hybrid bonded optical transceiver wafer anyone?), both these issues are an intentional part of the Cerebras architecture and make it hard for Cerebras to economically serve large models or even medium size models with long context lengths, that are representative of today's agentic workloads.

WSE 存在几个我们提到过的明显弱点。它拥有大量的 SRAM，但鉴于 SRAM 在单位功耗或单位成本下的密度天生较低，基于 HBM 的 GPU 和 XPU 在单位功耗或成本下能提供远高于它的显存容量。目前，HBM 被用于支持具有更长上下文长度的大型模型，以及通过更多的用户批处理（batching）来提升吞吐量。为了克服单片晶圆内存不足的问题而将更多晶圆联网，也受限于晶圆外带宽的匮乏。除非出现史诗级的技术突破（比如混合键合光收发器晶圆？），否则这两个问题都是 Cerebras 架构中刻意权衡的结果，这使得 Cerebras 很难经济高效地支持大型模型，甚至难以支持代表当今智能体（agentic）工作负载的具有长上下文的中型模型。

To illustrate this point, we have made an interactive calculator available at tokenomics.info/cerebras. This is a taste of the kind of research that our Tokenomics subscribers get.

为了说明这一点，我们在 tokenomics.info/cerebras 提供了一个交互式计算器。通过它，您可以初步了解我们的 Tokenomics 订阅用户所能获得的各类研究成果。


Source: Cerebras IPO | Tokenomics.info

Cerebras IPO | Tokenomics.info

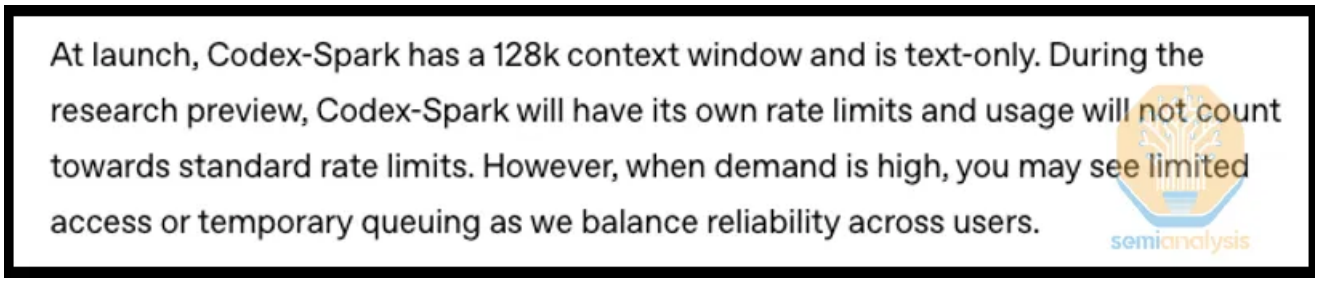
As shown above, when adjusting the average request size, number of concurrent requests supported, model size, and quantization for weights and KV Cache, the total number of WSEs required to run inference varies significantly. This, of course, leads to different performance characteristics on inference or decode, and $/Mtok cost conclusions.

如上所示，在调整平均请求大小、支持的并发请求数、模型大小以及权重和 KV 缓存的量化方式时，运行推理所需的 WSE 总数会发生显著变化。当然，这也会导致推理或解码阶段呈现出不同的性能特征，并得出不同的 $/Mtok 成本结论。

A notable assumption in this calculator is our 96.3k average request size. While Cerebras chooses to build their inference product for their customers around an assumption of 64k avg request size, we believe this is an artifact of running models

with limited context windows of 128k. In other words, confirmation bias in action.

该计算器中一个值得注意的假设是我们设定的 96.3k 平均请求大小。虽然 Cerebras 选择围绕 64k 平均请求大小的假设为其客户构建推理产品，但我们认为这只是在运行上下文窗口限制为 128k 的模型时产生的一种人为现象。换句话说，这是确认偏误在起作用。




Source: OpenAI's GPT 5.3 Codex Spark announcement

OpenAI 的 GPT 5.3 Codex Spark 发布公告

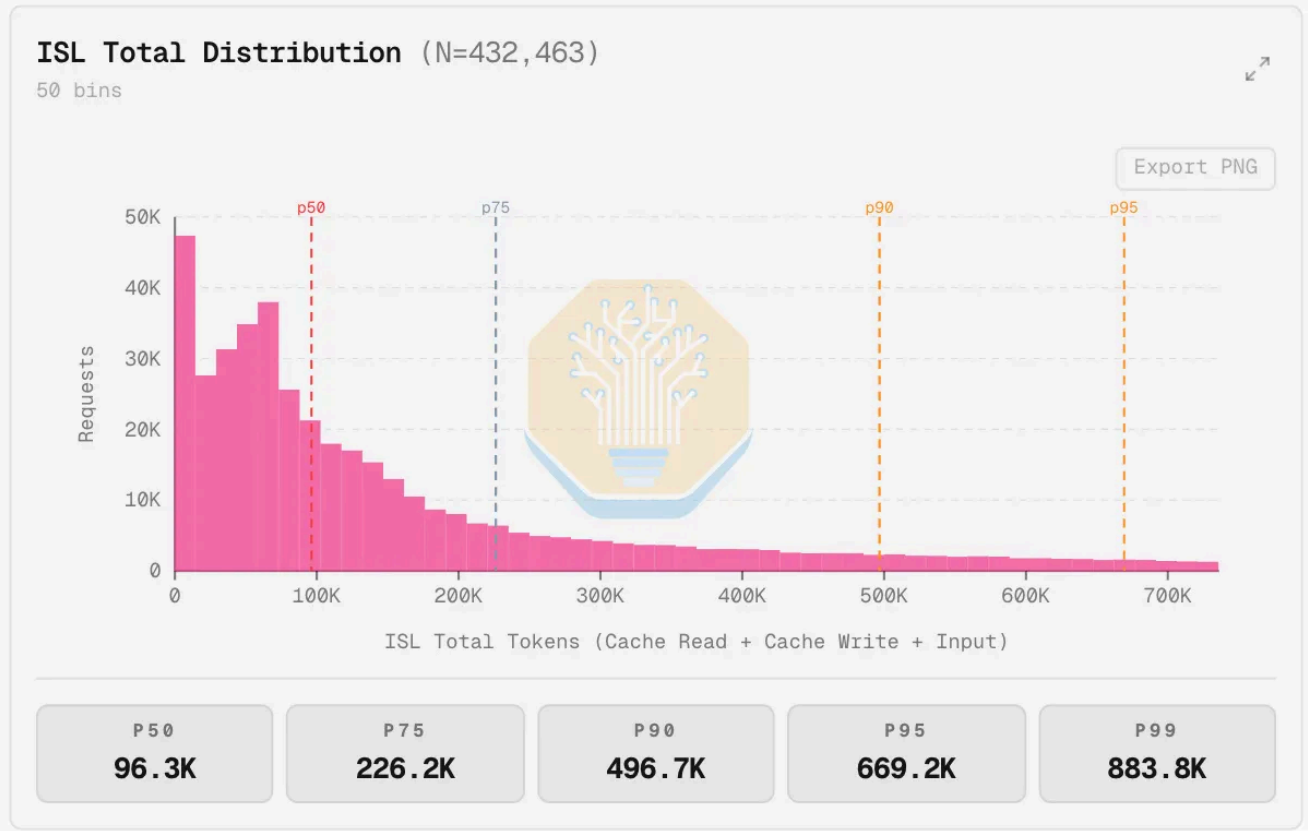
To get an understanding of exactly what real-world traffic patterns are, we built a proxy that collects fully anonymous traces from popular agentic coding harnesses such as Claude Code, Codex, Cursor, and OpenCode. This is part of an ongoing effort to collect production agentic traces for offline replay on InferenceX.

为了准确了解现实世界的流量模式，我们构建了一个代理程序，用于从 Claude Code、Codex、Cursor 和 OpenCode 等流行的智能体编程框架中收集完全匿名的追踪数据。这是我们持续努力的一部分，旨在收集生产环境中的智能体追踪数据，以便在 InferenceX 上进行离线回放。

A relatively large sample size of ~432k requests (about 80B tokens) leads us to believe that a typical P50 ISL is ~96.3k tokens, not 64k or fewer. We also deduce that the P90 or P95 requests can be exponentially more valuable than the initial requests and still critical to support. In total, almost 50% of our requests are over 128k, which is the maximum context window that Cerebras currently supports on public endpoints. Many sessions we see have an initial context length of over 100k tokens due to tool use context, system prompts, and things like skills and various other forms of primer

context.

约 43.2 万个请求（约 800 亿个 token）的较大样本量让我们相信，典型的 P50 输入序列长度（ISL）约为 9.63 万个 token，而非 64k 或更少。我们还推断出，P90 或 P95 请求的价值可能比初始请求呈指数级增长，且对于支持这些请求仍然至关重要。总计有近 50% 的请求超过了 128k，这是 Cerebras 目前在公共端点上支持的最大上下文窗口。由于工具使用上下文、系统提示词以及技能和各种其他形式的引导上下文，我们看到的许多会话初始上下文长度就超过了 10 万个 token。




Source: SemiAnalysis InferenceX AgentX dashboard (public launch soon!)

SemiAnalysis InferenceX AgentX 仪表板（即将公开发布！）

Moreover, the industry is trending towards larger context windows ad infinitum -- 128k context will certainly not be acceptable for long, especially with the prevalence of agentic workloads. The obvious conclusion of this analysis is that to run the latest open-source models with full context windows for real-world traffic patterns, Cerebras

needs to deploy a lot of wafers.

此外，行业正趋向于无限大的上下文窗口——128k 的上下文肯定无法长期满足需求，尤其是在代理型（agentic）工作负载盛行的情况下。这一分析的显而易见结论是：为了在现实世界的流量模式下运行具有完整上下文窗口的最新开源模型，Cerebras 需要部署大量的晶圆。

Just for the DeepSeek v4 example above, with 24 CS-3 a CS-3 customer could get 5 GB300 racks. Each rack has 20TB of HBM which is easily able to absorb the model weights leaving over 19TB for KVCache. That is a lot of KVCache to serve more users and to support long sequence length, and there are 5 of these racks also. While we've shown the speed gap in favour of Cerebras, this is how the throughput gap is well in favour of HBM-based GPUs.

仅就上述 DeepSeek v4 的例子而言，凭借 24 台 CS-3，CS-3 客户可以获得 5 个 GB300 机架。每个机架拥有 20TB 的 HBM，足以轻松容纳模型权重，并留下超过 19TB 用于 KVCache。如此庞大的 KVCache 能够服务更多用户并支持长序列长度，而且这样的机架共有 5 个。虽然我们展示了 Cerebras 在速度上的领先优势，但这正是基于 HBM 的 GPU 在吞吐量差距上占据绝对优势的原因。

## SRAM Scaling is Dead SRAM 扩展已死

Arguably, Cerebras is the company most exposed to the death of SRAM scaling, with Cerebras's key draw being SRAM and 50% of wafer area dedicated to SRAM. It's already showing up on their roadmap. WSE-1 on TSMC 16nm shipped with 18 GB of SRAM; WSE-2 on 7nm jumped to 40 GB, a decent 2.2x gen-on-gen. WSE-3 on 5nm advanced to just 44 GB. That's a 10% increase across a full node transition, while logic transistor count grew ~50%.

可以说，Cerebras 是受 SRAM 缩放停滞影响最大的公司，因为 Cerebras 的核心吸引力在于 SRAM，且其晶圆面积的 50% 都用于 SRAM。这一点已经在他们的路线图中显现出来。采用台积电 16nm 工艺的 WSE-1 配备了 18 GB SRAM；采用 7nm 工艺的 WSE-2 跃升至 40 GB，实现了 2.2 倍的代际增长。而采用 5nm 工艺的 WSE-3 仅增加到 44 GB。这意味着在完整的节点跨越中，SRAM 仅增长了 10%，而逻辑晶体管数量却增长了约 50%。

| SRAM Scaling by Node | | | |
|---|---|---|---|
| Node | HD SRAM Cell ($um^2$) | Density ($Mb/mm^2$) | Shrink vs N5 |
| N7 | 0.027 | 25 | -28.1% |
| N5 | 0.021 | 32 | Baseline |
| N3B | 0.020 | 33 | 5.2% |
| N3E | 0.021 | 32 | 0.0% |
| N2 | 0.021 | 32 | 0.0% |
| A16(1) | 0.021 | 32 | 0.0% |

*(1) Estimated HD SRAM Cell*

Source: SemiAnalysis, TSMC

来源：SemiAnalysis，TSMC

As we look to the future, this only gets worse. We can see that beyond 5nm (what the WSE-3 is currently fabbed on), SRAM scaling basically stops dead. The most common flavour of 3nm, N3E, has zero shrink relative to N5, and this continues to be the case for N2 and beyond. Now, the only way for Cerebras to increase SRAM capacity is by increasing wafer area dedicated to SRAM, sacrificing compute area. It's a strict tradeoff when the chip is wafer scale. This is why the next generation CS-4 system will use the same N5 based WSE-3, but with higher power to sustain higher clock speeds and compute but stuck at the same SRAM capacity.

展望未来，情况只会变得更加糟糕。我们可以看到，在 5nm（目前 WSE-3 所采用的制程）之后，SRAM 的缩放基本停滞了。最主流的 3nm 版本 N3E 相较于 N5 在尺寸上完全没有缩小，而 N2 及更先进的制程也将延续这一趋势。现在，Cerebras 增加 SRAM 容量的唯一方法就是增加专门用于 SRAM 的晶圆面积，但这必须以牺牲计算面积为代价。当芯片达到晶圆级规模时，这便成了一个严格的权衡问题。这就是为什么下一代 CS-4 系统将继续使用基于 N5 制程的 WSE-3，通过提高功耗来维持更高的时钟频率和计算能力，但 SRAM 容量却只能停留在原有水平。

By comparison, this isn't as critical for Groq as they are able to scale in the Z direction: using hybrid bonding to add additional SRAM tiles to vastly expand SRAM per package, which is on the roadmap for the Nvidia Groq LP40.

相比之下，这对 Groq 来说并不那么关键，因为他们能够向 Z 轴方向扩展：利用混合键合技术添加额外的 SRAM 磁贴，从而大幅扩展每个封装的 SRAM 容量，这已列入 Nvidia Groq LP40 的路线图中。

The logical path would be for Cerebras to do the same: wafer-on-wafer bond another wafer to expand SRAM and or compute per system. This is something that Cerebras is seriously exploring, having shown their concept of a DRAM wafer hybrid bonded onto the WSE to add more fast memory capacity. However, the timeline and technical feasibility of this is a concern for us given the litany of thermo-mechanical and bond-wave challenges. Yes, wafer-on-wafer bonding is an established process, but not where the whole wafer is stitched together as a whole chip. Cerebras has overcome these sorts of challenges in the past and will need to continue to innovate.

逻辑上的路径应该是 Cerebras 采取同样的做法：通过晶圆对晶圆（wafer-on-wafer）键合另一片晶圆，以扩展每个系统的 SRAM 或计算能力。这是 Cerebras 正在认真探索的方向，他们已经展示了将 DRAM 晶圆混合键合到 WSE 上以增加更多快速内存容量的概念。然而，鉴于一系列热力机械和键合波（bond-wave）挑战，其时间表和技术可行性令我们担忧。诚然，晶圆对晶圆键合是一项成熟的工艺，但并非针对整个晶圆被缝合为一个完整芯片的场景。Cerebras 过去曾克服过此类挑战，未来仍需持续创新。

## The Island Problem - bandwidth is geometry

## 孤岛问题——带宽即几何学

Despite the SRAM scaling issue, WSE still delivers an overwhelming amount of more compute and SRAM per single piece of silicon compared to other chips. Now comes the biggest tradeoff: the network. As mentioned earlier, each WSE has just 1.2 Tb/s (150GB/s) of off-package bandwidth. This is low compared to the average accelerator, and especially low relative to the amount of compute that the WSE has. No, this is not because the Cerebras architects have missed the importance of I/O for AI compute and overlooked adding more SerDes, this is just an inevitable tradeoff that comes with a wafer-scale chip.

尽管存在 SRAM 缩放问题，但与其他芯片相比，WSE 在单片硅片上提供的计算能力和 SRAM 容量仍然具有压倒性优势。现在到了最大的权衡点：网络。如前所述，每片 WSE 的封装外带宽仅为 1.2 Tb/s (150GB/s)。与平均水平的加速器相比，这个数值很低，尤其是相对于 WSE 所拥有的计算量而言更是如此。不，这并不是因为 Cerebras 的架构师忽略了 I/O 对 AI 计算的重要性，或者忘记了增加更多的 SerDes，这只是晶圆级芯片所带来的必然权衡。

By comparison, each Groq LP30 that NVIDIA will produce includes 96 lanes of 112G SerDes. That's a 9.6 Tb/s pipe in and out of a much smaller chip. It is clearly well prepared for the PDD + AFD inference setup that Jensen debuted at GTC this year.

相比之下，NVIDIA 即将生产的每颗 Groq LP30 都包含 96 通道的 112G SerDes。这意味着在一个小得多的芯片上拥有 9.6 Tb/s 的进出带宽。显然，它已经为黄仁勋在今年 GTC 上首次展示的 PDD + AFD 推理架构做好了充分准备。

| Cerebras vs Groq | | | | | |
|---|---|---|---|---|---|
| | Max FLOPS (PFLOPS) | SRAM Capacity (GB) | SRAM Bandwidth (TB/s) | Scale-Out Bandwidth (Tb/s) | Rough Price ($)[1] |
| Cerebras CS-3 / WSE-3 | 15.6 | 44.0 | 21,000 | 1.2 | 1,000,000 |
| Groq LP30 | 1.2 | 0.5 | 150 | 9.6 | 20,000 |
| Comparison* | 13.0x | 88.0x | 140.0x | 0.1x | 50.0x |

*Cerebras : Groq ratio
(1) Rough estimates for calculation purposes

Source: SemiAnalysis Estimates

SemiAnalysis 估算

So why the bandwidth tradeoff? At the current 150 GB/s (1.2 Tb/s) of off-wafer bandwidth, that's just 0.17 GB/s per mm of edge, so Nvidia's off-chip I/O is 130x denser!

那么为什么要进行带宽权衡呢？在目前 150 GB/s (1.2 Tb/s) 的晶圆外带宽下，每毫米边缘仅有 0.17 GB/s，因此 Nvidia 的片外 I/O 密度要高出 130 倍！

| Chip Type | Location | PHY Interface | Interface type | Edge dimensions (mm) | Uni-directional BW (TB/s) | BW DENSITY (GB/s/mm) |
|---|---|---|---|---|---|---|
| Nvidia Blackwell | East + West | NVLink5 | Serialized | | 0.9 | |
| | | NVLink-C2C | | | 0.5 | |
| | | PCIe Gen 6 | | | 0.1 | |
| | | **Total off-chip B/W** | | **65.0** | **1.5** | **22.7** |
| **Cerebras WSE-3** | **Full Perimeter** | **Total off-chip B/W** | **Serialized** | **860** | **0.15** | **0.17** |

Source: SemiAnalysis, Cerebras, Nvidia

来源：SemiAnalysis, Cerebras, Nvidia

Cerebras's lack of shoreline density comes down to the wafer scale architecture and reticle stepping problem. The WSE is patterned one reticle field at a time, tiling the same reticle pattern across the wafer in an 84-die array (12 columns × 7 rows on WSE-3). For the cross-scribe-line interconnect to work, every reticle exposure has to be identical, with the same logic, the same memory, the same routing, in the same

positions. That's what allows the on-wafer 2D mesh fabric to extend uniformly across die boundaries: every die's east edge connects to its neighbor's west edge with matching pin assignments.

Cerebras 缺乏岸线密度（shoreline density）的原因在于其晶圆级架构和光刻掩模步进问题。WSE 是每次对一个掩模场进行曝光显影，在晶圆上以 84 个芯片阵列（WSE-3 上为 12 列 × 7 行）平铺相同的掩模图案。为了使跨划线互连正常工作，每一次掩模曝光必须完全相同，即在相同位置具有相同的逻辑、相同的内存和相同的路由。正是这一点使得晶圆上的 2D 网格织物能够均匀地延伸跨越芯片边界：每个芯片的东侧边缘都以匹配的引脚分配与其邻居的西侧边缘相连。

This uniformity requirement is non-negotiable, and it has a punishing implication for IO. You cannot dedicate one reticle to PHYs while the other 83 reticles do compute. Every reticle has to be the same reticle. So, if you want more SerDes lanes on the wafer edge, you have to spend reticle area on SerDes in *every* reticle, not just the perimeter ones. Most of those PHYs will be in the middle of the wafer where they cannot reach the outside world, doing nothing. You pay a full silicon cost for IO that's stranded inside the wafer.

这种统一性要求是不可逾越的，而且它对 IO 产生了惩罚性的影响。你不能让一个光刻区域专门负责 PHY，而让其他 83 个区域负责计算。每个光刻区域都必须是完全相同的。因此，如果你想在晶圆边缘获得更多的 SerDes 通道，你就必须在每个光刻区域都分配 SerDes 面积，而不仅仅是边缘的那些。这些 PHY 中的大多数将位于晶圆中部，无法连接到外部世界，处于闲置状态。你为那些被困在晶圆内部的 IO 支付了高昂的硅片成本。

An alternative, putting PHYs only in perimeter reticles, would require a non-uniform stepping pattern, which is unfeasible from a process point of view. It would require swapping out reticles on a partially patterned wafer which would introduce untenable process risk and complexity, especially given all these reticles need to be stitched together which breaks the cross-scribe-line interconnect that makes wafer-scale work in the first place (what we called the "scale-up network" earlier).

另一种方案是仅在边缘光罩中放置 PHY，但这需要非均匀的步进图案，从工艺角度来看是不可行的。这需要在已完成部分图案化的晶圆上更换光罩，这将引入无法承受的工艺风险和复杂性，特别是考虑到所有这些光罩都需要缝合在一起，而这样做会破坏使晶圆级芯片得以实现的跨划线互连（即我们之前提到的"扩展网络"）。

Even if Cerebras accepted stranded silicon and burned area on PHYs everywhere, they would hit a third constraint: on-wafer dataflow blocking. During inference, the on-chip 2D mesh fabric carries the activations, weights, and gradients between cores (again, why we called it the scale-up network). Every PHY block placed inside a reticle is a hole in the mesh, a region where compute and routing cannot exist. PHYs are large (high-speed SerDes are typically 1–3 $mm^2$ each at 5nm, including the analog circuitry that doesn't scale with logic), and their analog circuitry is hostile to neighboring digital logic due to power and EMI concerns, demanding guard regions. Putting PHYs in the middle of the wafer means the 2D mesh fabric has to be routed around that area, increasing latency between reticles and reducing total bandwidth. Too much of this excess routing would defeat the purpose of going wafer-scale, since the whole point is fast and low-power dataflow across tiles.

即使 Cerebras 接受了晶圆上存在闲置硅片以及到处都是 PHY 造成的面积浪费，他们也会遇到第三个限制：晶圆级数据流阻塞。在推理过程中，片上 2D 拓扑网络（mesh fabric）在核心之间传输激活值、权重和梯度（这再次解释了为什么我们称其为纵向扩展网络）。放置在光刻掩模区内的每个 PHY 模块都是网格中的一个“空洞”，即计算和路由无法存在的区域。PHY 的体积很大（在 5nm 工艺下，高速 SerDes 通常每个占地 1–3 $mm^2$，其中包括不随逻辑电路缩小的模拟电路），而且由于功耗和电磁干扰（EMI）问题，其模拟电路对相邻的数字逻辑极不友好，需要设置保护区域。将 PHY 置于晶圆中间意味着 2D 网格网络必须绕过该区域进行路由，从而增加了掩模区之间的延迟并降低了总带宽。过多的这种冗余路由将违背采用晶圆级设计的初衷，因为其核心意义就在于实现瓦片（tile）之间快速且低功耗的数据流传输。

In summary, the uniform tiling that makes wafer-scale possible (one reticle pattern, one mesh fabric) is what makes adding IO bandwidth hard. Cerebras must be looking for ways around this limitation.

总而言之，使晶圆级规模化成为可能的统一平铺设计（单一光掩模图案、单一网格结构），正是导致增加 IO 带宽变得困难的原因。Cerebras 必须正在寻找突破这一限制的方法。

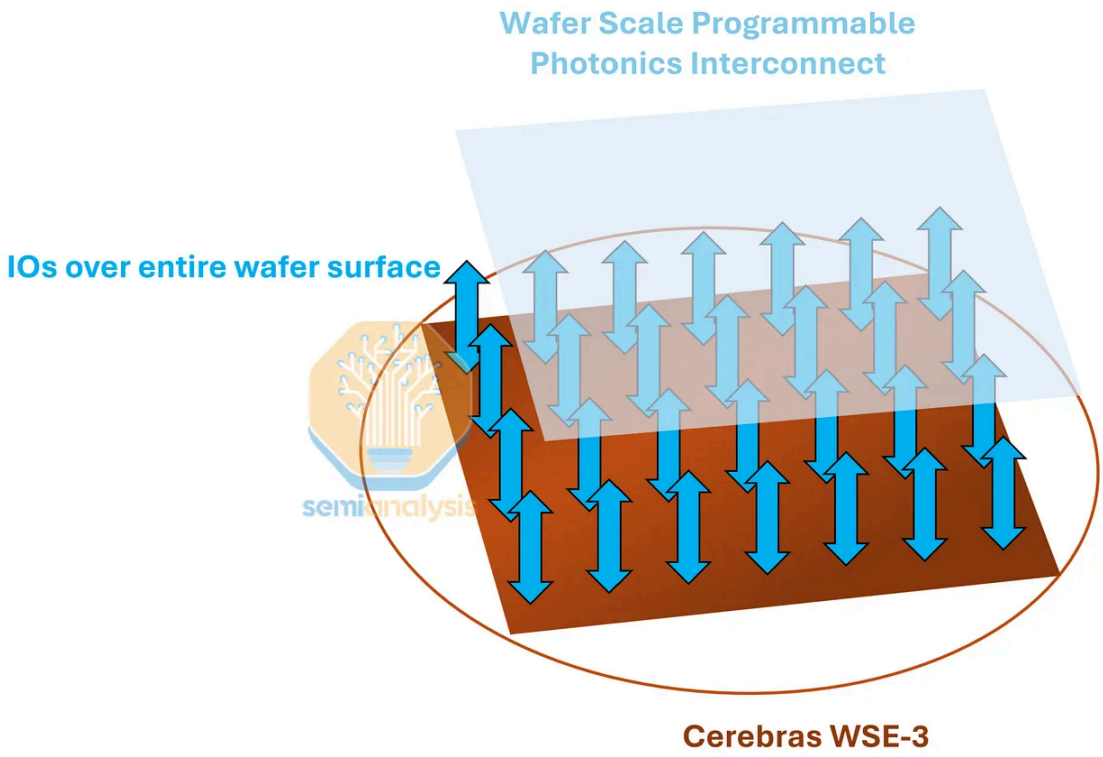
A lot of the issues we just described come from the realities of moving data in the electrical realm, which are circumvented with optical I/O. The solution that Cerebras is working on (again proof that Cerebras recognizes the problem) is a photonic interconnect wafer hybrid bonded onto the WSE. As with the additional DRAM wafer to solve the memory constraint, the bandwidth constraint is also being addressed with

another wafer.

我们刚才描述的许多问题都源于在电学领域移动数据的现实情况，而光子 I/O 则可以规避这些问题。Cerebras 正在研究的解决方案（再次证明了 Cerebras 意识到了这一问题）是将光子互连晶圆混合键合到 WSE 上。正如通过增加 DRAM 晶圆来解决内存限制一样，带宽限制也正在通过另一个晶圆来解决。

Cerebras claims that for LLM inference they don't need any more bandwidth and is only aggressively pursuing hybrid bonding wafer scale photonic I/O to help their HPC boomers. The HPC customers whom NVIDIA has effectively abandoned after reducing FP64 native hardware on their GPUs to basically nothing. This is great that Cerebras is aggressively reinvesting completely back into moonshot R&D instead of doing buybacks. Buybacks is not an good idea for companies that are lots of r&d things to reinvest into, for example, AMD did ~$221million dollars of buybacks last quarter yet internally multiple AMD internal teams continue to lack development interconnected GPU clusters.

Cerebras 声称，对于 LLM 推理，他们不再需要更多的带宽，目前正积极追求混合键合晶圆级光子 I/O，以助力其 HPC（高性能计算）领域的老牌客户。在 NVIDIA 将其 GPU 上的原生 FP64 硬件削减到几乎为零后，这些 HPC 客户实际上已被 NVIDIA 抛弃。Cerebras 将资金积极地重新投入到开创性的研发中，而不是进行股票回购，这一点非常出色。对于有大量研发项目需要再投资的公司来说，回购并非良策。例如，AMD 上个季度进行了约 2.21 亿美元的回购，但其内部多个团队仍持续缺乏用于开发的互连 GPU 集群。




Cerebras's photonic wafer concept. Source: SemiAnalysis, Cerebras

Cerebras 的光子晶圆概念。来源：SemiAnalysis，Cerebras

This allows data to move in/out of the wafer up through the z-axis, rather than having it go through the edges. The photonics partner developing this photonic wafer is Ranovus. This reintroduces the issues of WoW hybrid bonding for wafer scale silicon. Optical components are thermally sensitive (cannot be too hot or too cool) and it will be sandwiched directly against a wafer that runs hot. Lastly, there is the practical difficulty of fibers needing to be perfectly coupled off to the wafer. This is still being figured out at the optical engine level for conventional CPO, let alone for something wafer scale.

这使得数据能够通过 z 轴进出晶圆，而不是通过边缘传输。开发这种光子晶圆的光子学合作伙伴是 Ranovus。这再次引入了晶圆级硅的 WoW 混合键合问题。光学元件对温度非常敏感（不能太热或太冷），而它将被直接夹在运行温度很高的晶圆之间。最后，还存在光纤需要与晶圆完美耦合的实际困难。对于传统的共封装光学（CPO）来说，这在光学引擎层面仍处于摸索阶段，更不用说晶圆级应用了。

With all this in mind, let's look at how the architecture shapes inference workloads

考虑到以上所有因素，让我们来看看该架构是如何塑造推理工作负载的

## Pipeline Parellelism is Forced

### 流水线并行是必然选择

One of the key concerns that we have already highlighted with using Cerebras in any inference deployment is just how big models have gotten. Both in terms of total parameter count (e.g. DeepSeek V4 is 1.6T total parameters), and in terms of KV Cache (256k context is the norm, with DeepSeek V4 debuting 1M context).

在任何推理部署中使用 Cerebras 时，我们已经强调过的一个核心担忧就是模型已经变得多么庞大。这既体现在总参数量上（例如 DeepSeek V4 的总参数量为 1.6T），也体现在 KV 缓存上（256k 上下文已成为常态，而 DeepSeek V4 更是推出了 1M 上下文）。

The combination of limited single wafer SRAM capacity of 44GB in the WSE-3 and low IO bandwidth results in challenges effectively serving models of these sizes.

WSE-3 单晶圆 44GB 的有限 SRAM 容量与低 IO 带宽相结合，导致在有效提供此类规模的模型服务时面临挑战。

Each CS-3 has just 12x100GbE of IO bandwidth -- roughly 150 GB/s for the entire wafer. This is one sixth of the scale-up bandwidth for Blackwell with NVLink5 at 900 GB/s per GPU, and an order of magnitude below the bandwidth of HBM.

每台 CS-3 仅具备 12x100GbE 的 IO 带宽——整个晶圆约为 150 GB/s。这仅为配备 NVLink5 的 Blackwell 单个 GPU 900 GB/s 扩展带宽的六分之一，且比 HBM 的带宽低了一个数量级。

This bandwidth constraint is what makes it difficult for Cerebras to serve larger parameter models. Any large tensors to be used must be resident on the wafer; streaming on/off the wafer is impossible with such a small amount of I/O. Similarly, any sharding strategy that requires high-bandwidth collectives at each layer is categorically ruled out.

这种带宽限制正是 Cerebras 难以服务更大参数模型的原因。任何需要使用的大型张量都必须驻留在晶圆上；由于 I/O 量如此之小，在晶圆内外进行流式传输是不可能的。同样，任何需要在每一层进行高带宽集合通信的分片策略也都被断然排除在外。

The only real option is pipeline parallelism, which slices the model layer-wise across wafers and only transfers activations between stages, relying on the fact that activations are small relative to weights. This reduces network requirements and keeps the capacity-demanding components (the weights, and to some extent the KV cache) stationary instead of moving on or off the wafer. For instance, Cerebras shards Llama3 70B across 4x WSE-3, transferring only the activations between each wafer and staying well within the available 1.2Tbps I/O.

唯一的实际选择是流水线并行（pipeline parallelism），它将模型按层切分并分布在不同的晶圆上，仅在各阶段之间传输激活值（activations），这利用了激活值相对于权重较小的特性。这种方法降低了对网络的要求，并使对容量要求较高的组件（权重，以及在一定程度上的 KV 缓存）保持静止，而不是在晶圆内外移动。例如，Cerebras 将 Llama3 70B 分片部署在 4 个 WSE-3 上，仅在每个晶圆之间传输激活值，从而将带宽需求控制在 1.2Tbps 的可用 I/O 范围内。

As you increase the number of wafers used to host the model, there are several factors to wrestle with to increase scale. First, the **pipeline bubble**: to keep N pipeline stages busy, you need at least N in-flight microbatches. A 4-stage config needs ~4 microbatches in flight; a 16-stage config needs ~16. Second, **each in-flight microbatch carries its own KV cache**, and on Cerebras that KV cache must live in the same 44GB of on-wafer SRAM that's already mostly consumed by weights. Even if there is enough capacity in the SRAM with the heavily compressed KVs of recent models such as DeepSeek V4, the time to transfer the KV cache on or off the wafer is still quite large. Additionally, scaling the model size scales the number of wafers needed to hold the weights and therefore increases the number of times the latency of wafer->wafer activation transfer adds to the decode time.

随着用于托管模型的晶圆数量增加，在扩大规模时需要应对几个因素。首先是流水线气泡：为了保持 N 个流水线阶段处于忙碌状态，你至少需要 N 个在途（in-flight）微批次。4 阶段配置需要约 4 个在途微批次；16 阶段配置则需要约 16 个。其次，每个在途微批次都携带自己的 KV 缓存，而在 Cerebras 上，这些 KV 缓存必须存放在晶圆上仅有的 44GB SRAM 中，而这些空间大部分已被权重占用。即使在 DeepSeek V4 等近期模型采用重度压缩 KV 的情况下 SRAM 容量足够，将 KV 缓存移入或移出晶圆的传输时间仍然相当长。此外，扩大模型规模会增加容纳权重所需的晶圆数量，从而增加晶圆间激活传输延迟累加到解码时间中的次数。

In summary, the way the wafer is being used in production today basically goes against the entire ethos of the wafer. The whole point of the wafer is to run really fast at small batch sizes!

总而言之，目前在生产环境中使用晶圆的方式基本上违背了晶圆设计的初衷。晶圆存在的全部意义在于以小批量实现极高的运行速度！

# Running the Numbers 数据计算

Let's take a look at some napkin math with a few open-source model architectures to better understand how different models map to Cerebras's SRAM footprint. Below are some rough ballpark numbers showing the footprint of several models.

让我们通过一些开源模型架构来进行简单的估算，以便更好地理解不同模型如何映射到 Cerebras 的 SRAM 占用空间。以下是显示几种模型占用空间的粗略估算数据。

**Model Footprints**

| Model | Total Params (B) | Attn Params (B) | FFN Params (B) | Size BF16 (GB) | Size FP8 (GB) | Size FP4 (GB) | Checkpoint Size (GB) | KV Storage | KV@128K (GB) | KV@256K (GB) | KV@1M (GB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Llama 3.1 8B | 8.0 | 1.3 | 5.6 | 16.1 | 8.0 | 4.0 | 16.1 | BF16 | 16.8 | - | - |
| Llama 3.1 70B | 70.6 | 12.1 | 56.4 | 141.1 | 70.6 | 35.3 | 141.1 | BF16 | 41.9 | - | - |
| Llama 3.1 405B | 405.9 | 71.9 | 329.8 | 811.7 | 405.9 | 202.9 | 811.7 | BF16 | 66.1 | - | - |
| Llama 4 Scout | 107.8 | 3.0 | 102.7 | 215.5 | 107.8 | 53.9 | 215.5 | BF16 | 25.2 | 50.3 | 196.6 |
| Llama 4 Maverick | 400.7 | 3.0 | 395.6 | 801.4 | 400.7 | 200.4 | 801.4 | BF16 | 25.2 | 50.3 | 196.6 |
| Llama 4 Behemoth | 2105.8 | 45.6 | 2053.6 | 4211.6 | 2105.8 | 1052.9 | 4211.6 | BF16 | 41.9 | 83.9 | 327.7 |
| gpt-oss-120b | 116.8 | 1.0 | 114.7 | 233.6 | 116.8 | 58.4 | 65.2 | BF16 | 9.4 | - | - |
| DeepSeek V3 (671B / 37B) | 682.5 | 11.6 | 669.1 | 1365.1 | 682.5 | 341.3 | 682.5 | BF16 | 9.0 | - | - |
| DeepSeek V4-Pro (1.6T / 49B) | 1595.2 | 16.3 | 1577.0 | 3190.4 | 1595.2 | 797.6 | 856.1 | FP8 | 0.6 | 1.3 | 5.0 |
| DeepSeek V4-Flash (284B / 13B) | 290.8 | 5.1 | 284.6 | 581.6 | 290.8 | 145.4 | 157.4 | FP8 | 0.4 | 0.9 | 3.5 |

Source: Llama, DeepSeek, OpenAI, SemiAnalysis

Llama, DeepSeek, OpenAI, SemiAnalysis

And now some rough numbers considering the WSE-3 specs. We make some assumptions here, including that the transfers will use the full 12x100Gbps.

现在，让我们根据 WSE-3 的规格来看一些粗略的数据。我们在这里做了一些假设，包括传输将使用全部 12x100Gbps 的带宽。

| Deployment on WSE-3 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Pretrained Size (GB) | Min Wafers | Free SRAM per Wafer | KV Storage | 128K KV Transfer (ms) | 256K KV Transfer (ms) | 1M KV Transfer (ms) |
| Llama 3.1 8B | 16.1 | 1 | 27.9 | BF16 | 111.8 | - | - |
| Llama 3.1 70B | 141.1 | 4 | 8.7 | BF16 | 55.9 | - | - |
| Llama 3.1 405B | 811.7 | 21 | 1.3 | BF16 | 21.0 | - | - |
| Llama 4 Scout | 215.5 | 6 | 0.9 | BF16 | 3.4 | 6.7 | 26.2 |
| Llama 4 Maverick | 801.4 | 24 | 1.8 | BF16 | 7.0 | 14.0 | 54.6 |
| Llama 4 Behemoth | 4211.6 | 96+ | 0.1 | BF16 | 2.9 | 5.8 | 22.8 |
| gpt-oss-120b | 65.2 | 2 | 11.4 | BF16 | 31.5 | - | - |
| DeepSeek V3 (671B / 37B) | 682.5 | 21 | 1.3 | BF16 | 2.9 | - | - |
| DeepSeek V4-Pro (1.6T / 49B) | 856.1 | 21 | 1.2 | FP8 | 0.2 | 0.4 | 1.6 |
| DeepSeek V4-Flash (284B / 13B) | 157.4 | 4 | 4.7 | FP8 | 0.7 | 1.5 | 5.8 |

Source: Llama, DeepSeek, OpenAI, SemiAnalysis

Llama, DeepSeek, OpenAI, SemiAnalysis

Here we define the minimum number of wafers to store the model weights by sharding strictly along layer boundaries, but we don't include the space to store KV caches. In practice, more wafers may be used to give more space for KV caches. Activation transfer times are not included because activations are so small that their transfer will be bound by the propagation time across the I/O path.

在此，我们通过严格沿层边界进行分片，定义了存储模型权重所需的最少晶圆数量，但其中不包括存储 KV 缓存的空间。在实践中，可能会使用更多晶圆以提供更大的 KV 缓存空间。激活传输时间未被计入，因为激活值非常小，其传输将受限于跨 I/O 路径的传播时间。

It is clear from the table that recent KV cache compression techniques such as those published by DeepSeek might significantly alleviate issues Cerebras has with long-context serving. However, the problem of slow I/O does not completely disappear. Firstly, KV transfer times on- and off-chip are still quite large at several milliseconds, both impacting TTFT and making it more difficult to achieve high utilization due to issues of batching, pipelining, and latency-hiding related to KV cache storage and transfer. Secondly, the fixed I/O latency of activation transfer must be paid in proportion to the number of wafers used to host a model instance. This is a fixed cost

in the TPOT that scales linearly with the number of wafers used to host the model.

从表中可以清楚地看到，最近的 KV 缓存压缩技术（如 DeepSeek 发布的技术）可能会显著缓解 Cerebras 在长上下文服务方面面临的问题。然而，I/O 缓慢的问题并未完全消失。首先，片上和片下的 KV 传输时间仍然相当长，达到数毫秒，这既影响了首字延迟（TTFT），也由于与 KV 缓存存储和传输相关的批处理、流水线及延迟隐藏等问题，使得实现高利用率变得更加困难。其次，激活传输的固定 I/O 延迟必须与用于托管模型实例的晶圆数量成比例支付。这是每输出 Token 时间（TPOT）中的一项固定成本，并随托管模型所使用的晶圆数量线性增加。

The key takeaway is that Cerebras, while fast, pays a large latency cost to move data on and off the wafer, and therefore their cost-to-performance ratio (or perf per Joule) will depend on how much of that latency they can hide or minimize. A clue about the difficulty of this in practice may be reflected in Model offerings on Cerebras Inference Cloud. The largest production model is GPT-OSS, which is only 120B total parameters. There are larger preview models, but even those top out at 355B (GLM 4.7). For reference, Sonnet and Opus are 1T and 5T parameters respectively, per Elon. Notably, the formerly popular Llama 70B and 405B models were also deprecated, potentially due to the economics of serving them.

核心结论是，Cerebras 虽然速度极快，但在晶圆内外传输数据时会产生巨大的延迟成本，因此其性价比（或每焦耳性能）将取决于他们能在多大程度上隐藏或最小化这种延迟。在实践中，这一难题的线索可能体现在 Cerebras 推理云（Cerebras Inference Cloud）提供的模型中。目前最大的生产级模型是 GPT-OSS，总参数量仅为 120B。虽然有一些更大的预览版模型，但即便如此，上限也仅为 355B（GLM 4.7）。作为参考，据 Elon 称，Sonnet 和 Opus 的参数量分别为 1T 和 5T。值得注意的是，此前广受欢迎的 Llama 70B 和 405B 模型也已被弃用，这可能是由于提供这些服务的经济效益问题。

Models served on the Cerebras Inference Cloud[1]

| # | Status | Model | Model ID | Vendor | Arch | Total Params | Active Params (MoE) | Weights @ FP16 | CS-3 Needed | Throughput (tok/s) |
|---|---|---|---|---|---|---|---|---|---|---|
| 01 | Production | Llama 3.1 8B | llama3.1-8b | Meta | Dense | 8 B | 8 B | 16 GB | 1 | ~2,200 |
| 02 | Production | GPT-OSS 120B | gpt-oss-120b | OpenAI | MoE | 120 B | 5.1 B | 240 GB | 8 | ~3,000 |
| 03 | Preview | Qwen 3 235B Instruct | qwen-3-235b-a22b-instruct-2507 | Alibaba | MoE | 235 B | 22 B | 470 GB | 15 | ~1,400 |
| 04 | Preview | GLM 4.7 | zai-glm-4.7 | Z.ai | MoE | 355 B | 32 B | 710 GB | 22 | ~1,000 |
| 05 | Partner | GPT-5.3 Codex-Spark | codex-spark | OpenAI | Proprietary | n/d | n/d | n/d | n/d | ~1,000 |
| 06 | Deprecated | Llama 3.3 70B | llama3.3-70b | Meta | Dense | 70 B | 70 B | 140 GB | 5 | ~2,314 |
| 07 | Deprecated | Llama 3.1 405B | llama3.1-405b | Meta | Dense | 405 B | 405 B | 810 GB | 25 | ~969 |
| 08 | Deprecated | Llama 4 Scout | llama-4-scout-17b-16e | Meta | MoE | 109 B | 17 B | 218 GB | 7 | ~2,000 |
| 09 | Deprecated | Llama 4 Maverick | llama-4-maverick-17b-128e | Meta | MoE | 400 B | 17 B | 800 GB | 25 | n/d |
| 10 | Deprecated | Qwen 3 Coder 480B | qwen-3-coder-480b | Alibaba | MoE | 480 B | 35 B | 960 GB | 30 | ~2,000 |
| 11 | Deprecated | DeepSeek R1 Distill Llama 70B | deepseek-r1-distill-llama-70b | DeepSeek | Dense | 70 B | 70 B | 140 GB | 5 | ~1,600 |

*(1) Open-weight and partner foundation models hosted on Cerebras CS-3 inference infrastructure. Throughput shown is the Cerebras-quoted single-stream peak. Weight sizes are computed at native FP16 (2 bytes per parameter); MoE models show total / active for memory and compute respectively.*

Source: Cerebras, Llama, OpenAI, DeepSeek, Llama, Qwen, SemiAnalysis

Cerebras, Llama, OpenAI, DeepSeek, Llama, Qwen, SemiAnalysis

It's also worth emphasizing that two of the most popular frontier open-source models of 2025, DeepSeek V3 and Kimi K2, have never been offered on the public Cerebras Cloud. This is despite the large KV cache size reduction in DeepSeek V3 due to the use of Multi-head Latent Attention (MLA), which would leave it with better serving economics than Llama 3 405B.

还值得强调的是，2025 年最受欢迎的两款前沿开源模型 DeepSeek V3 和 Kimi K2 从未在 Cerebras 公有云上提供。尽管 DeepSeek V3 由于采用了多头潜在注意力（MLA）机制，大幅减少了 KV 缓存占用，使其在推理服务经济性上优于 Llama 3 405B，但情况依然如此。

With that said, our analysis above shows that the even newer DeepSeek V4 Pro can have a similar deployment shape to Llama 405B (which they have already served on Cerebras cloud), with significantly smaller KV cache sizes. For that reason, with modern KV cache compression techniques and enough concurrency, Cerebras might indeed look attractive even for large 1T+ models.

话虽如此，我们上述的分析表明，更新的 DeepSeek V4 Pro 可能具有与 Llama 405B（他们已经在 Cerebras 云上提供服务）类似的部署形态，且 KV 缓存大小显著更小。因此，凭借现代 KV 缓存压缩技术和足够的并发量，Cerebras 即使对于 1T+ 的超大模型也确实可能具有吸引力。

# The Cerebras OpenAI Deal

# Cerebras 与 OpenAI 的交易

OpenAI plays a huge role in Cerebras's future. It is simultaneously the company's secured lender, its largest warrant holder, and the source of essentially all of its $24.6B backlog. OpenAI's financial stake in Cerebras means Cerebras's fortunes are tied to a single counterparty through three interlocking mechanisms that all move in the same direction. If the relationship succeeds, the loan is repaid through capacity delivery rather than cash (with the 6% accrued interest waived on capacity-repaid portions), the warrant vests and aligns incentives, and revenue scales into the billions. On a fully diluted basis, OpenAI could hold as much as 12% of Cerebras shares (not including any new issuances and offerings).

OpenAI 在 Cerebras 的未来中扮演着至关重要的角色。它同时是该公司的担保债权人、最大的认股权证持有者，以及其 246 亿美元积压订单的几乎全部来源。OpenAI 在 Cerebras 的财务权益意味着，Cerebras 的命运通过三个同向变动的连锁机制与单一交易对手紧密相连。如果双方关系取得成功，贷款将通过交付算力而非现金来偿还（以算力偿还的部分将免除 6% 的应计利息），认股权证将行权并使双方利益趋于一致，营收规模也将达到数十亿美元。在完全稀释的基础上，OpenAI 可能持有高达 12% 的 Cerebras 股份（不包括任何新发行和发售的股份）。

Here are the details:

以下是详细信息：

> · In December 2025, Cerebras and OpenAI signed a Master Relationship Agreement (MRA) under which OpenAI committed to purchase 750MW of AI inference compute capacity, deployed in tranches over 2026-2028, with each tranche carrying a 3-4 year term extendable to five years. OpenAI also holds an option (not an obligation) to purchase an additional 1.25GW, bringing the total potential to 2GW. The S-1 discloses $24.6B in remaining performance obligations as of December 31, 2025. More importantly, pass-through costs (data center rent, power, leasehold improvements, security) are reimbursed by OpenAI and recognized as revenue on a

gross basis.

· 2025 年 12 月，Cerebras 与 OpenAI 签署了一份主关系协议（MRA）。根据该协议，OpenAI 承诺购买 750MW 的 AI 推理计算容量，并在 2026 年至 2028 年期间分批部署，每批次的期限为 3 至 4 年，并可延长至 5 年。OpenAI 还拥有一项期权（而非义务），可额外购买 1.25GW，使总潜在容量达到 2GW。S-1 文件披露，截至 2025 年 12 月 31 日，剩余履约义务为 246 亿美元。更重要的是，代收代付成本（数据中心租金、电力、租赁改良支出、安保）由 OpenAI 报销，并按总额法确认为收入。

· OpenAI also provided a $1B Working Capital Loan to Cerebras via a secured promissory note that bears 6% annual interest. Interest is waived if Cerebras repays through delivery of compute capacity or hardware under the MRA. Repayment is scheduled in equal amortized installments over three years, starting after delivery of the final tranche of the initial 250MW. If the MRA is terminated for any reason other than OpenAI's own material uncured breach, Cerebras may be required to immediately repay the full outstanding balance plus accrued interest. OpenAI also retains the right to direct the custodian bank to stop following Cerebras's instructions on deploying the funds and instead control the disposition directly.

· OpenAI 还通过一份年利率为 6% 的担保本票，向 Cerebras 提供了 10 亿美元的营运资金贷款。如果 Cerebras 根据主转售协议（MRA）通过交付算力或硬件进行偿还，则可免除利息。还款计划在交付首批 250MW 的最后一期后开始，分三年等额摊还。如果 MRA 因 OpenAI 自身重大未补救违约之外的任何原因终止，Cerebras 可能会被要求立即偿还全部未偿余额及应计利息。OpenAI 还保留指示托管银行停止执行 Cerebras 资金部署指令，并转而直接控制资金处置的权利。

· Alongside the MRA, Cerebras issued OpenAI a warrant for 33,445,026 shares of Class N (non-voting) common stock at an exercise price of $0.00001 per share, effectively free. The warrant vests in three structurally distinct tranches: 4,459,337 shares vested immediately upon receipt of the $1bn Working Capital Loan in January 2026; 5,574,171 shares vest upon the earlier of Cerebras reaching a $40bn market capitalization or OAI hitting specified fee payment milestones under the MRA; and the remaining 23,411,518 shares vest in sub-tranches tied to capacity delivery, split between *Committed Capacity* (tied to firm delivery dates already in the MRA) and *Additional Capacity* (which vests only if OAI exercises its option to expand the deal to the full 2GW). Per S-1 filings, Cerebras assessed that the working

capital loan tranche, the market capitalization / payment threshold tranche, and the Committed Capacity sub-tranche are *probable* of vesting, while the Additional Capacity sub-tranche is *not probable* (i.e. the 2GW expansion is not yet baseline). OAI also holds demand registration rights, meaning it can force Cerebras to register these shares for public resale at any time. The warrant expires December 24, 2035, or five business days after no binding commitments or payments remain under the MRA.

· 在签署主收入协议（MRA）的同时，Cerebras 向 OpenAI 发行了一份认股权证，涉及 33,445,026 股 N 类（无投票权）普通股，行权价为每股 0.00001 美元，实际上等同于免费。该权证分三个结构截然不同的部分行权：4,459,337 股在 2026 年 1 月收到 10 亿美元营运资金贷款后立即行权；5,574,171 股在 Cerebras 市值达到 400 亿美元或 OpenAI 达到 MRA 规定的特定费用支付里程碑时（以较早者为准）行权；剩余的 23,411,518 股则分为与容量交付挂钩的子部分行权，其中包括承诺容量（与 MRA 中已确定的交付日期挂钩）和额外容量（仅在 OpenAI 行使期权将交易扩大至完整的 2GW 时行权）。根据 S-1 注册文件，Cerebras 评估认为营运资金贷款部分、市值/支付门槛部分以及承诺容量子部分均有可能行权，而额外容量子部分则被视为不太可能（即 2GW 的扩张尚未成为基准计划）。OpenAI 还持有要求注册权，这意味着它可以随时强制 Cerebras 为这些股票办理公开转售注册。该权证将于 2035 年 12 月 24 日到期，或者在 MRA 项下不再存在任何约束性承诺或付款后的五个工作日到期。

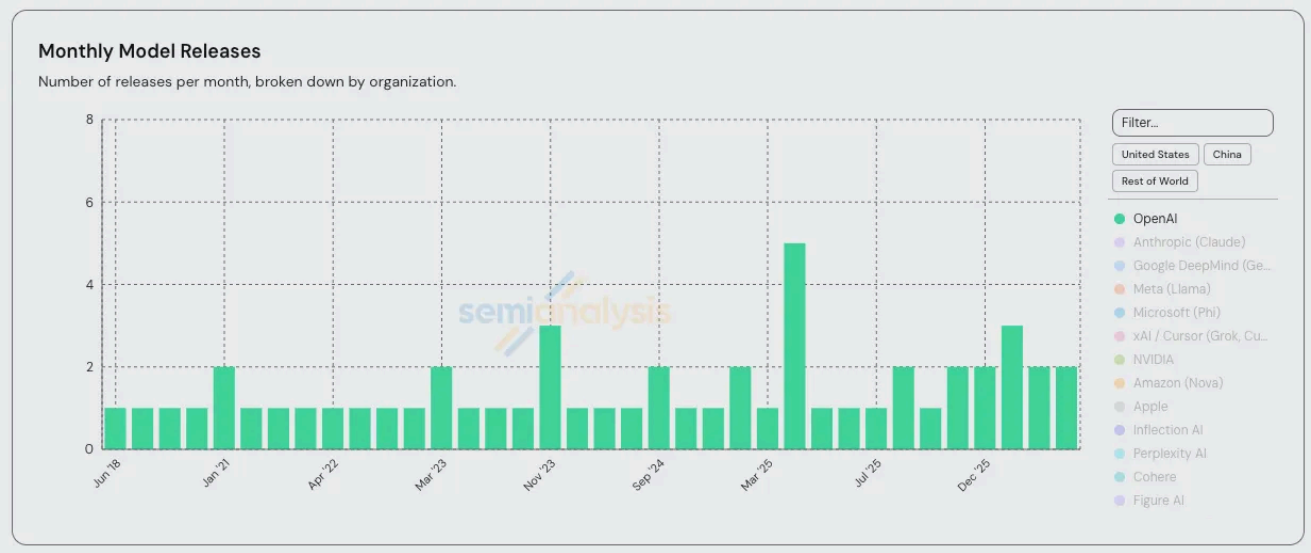
· Under ASC 505-50, equity given to a customer is treated as recognized as contra-revenue over the life of the commercial agreement, not at vesting and not at market value. The number is locked to the grant date fair value, regardless of where the stock trades later. Per S-1 filings, Cerebras values the warrants at $82.02 per share as of December 31, 2025, which serves as a useful proxy for grant date fair value for the OpenAI deal. Applying the $82.02 per share to the full ~33.4M shares, we get a theoretical maximum contra-revenue of ~$2.74bn or roughly 10% of the revenue expected from OpenAI. We assume the reported $24.6bn backlog is NET of the contra-revenue from the warrants. In reality, however, only the *probable* tranches flow through revenue on a sliding-scale basis; the Working Capital Loan tranche (~$366mn, vested January 2026), the market capitalization / payment threshold tranche (~$457mn), and the Committed Capacity sub-tranche (size undisclosed). The Additional Capacity sub-tranche only hits contra-revenue with a cumulative catch-

up adjustment *if and when* OAI exercises the 2GW expansion option.

根据 ASC 505-50 准则，给予客户的股权被视为在商业协议有效期内确认为收入抵减（contra-revenue），而非在归属时或按市场价值确认。该数值锁定在授予日的公允价值，无论股票后续交易价格如何。根据 S-1 文件，Cerebras 将截至 2025 年 12 月 31 日的认股权证估值为每股 82.02 美元，这可作为 OpenAI 交易授予日公允价值的有效参考。将每股 82.02 美元应用于全部约 3340 万股，我们得到理论上的最大收入抵减额约为 27.4 亿美元，约占 OpenAI 预期收入的 10%。我们假设报告的 246 亿美元积压订单是扣除认股权证产生的收入抵减后的净额。然而在现实中，只有“很可能发生”的份额会按比例计入收入抵减；包括营运资金贷款份额（约 3.66 亿美元，2026 年 1 月归属）、市值/付款门槛份额（约 4.57 亿美元）以及承诺容量子份额（规模未披露）。只有当 OAI 行使 2GW 扩容选项时，额外容量子份额才会通过累计追溯调整计入收入抵减。

While Cerebras had been largely left out of the neocloud boom, OpenAI's February release of GPT-5.3-Codex-Spark (a model using the gptoss-120B architecture that was distilled from the real 5.3 Codex) is turning things around. Spark runs on Cerebras at up to 2,000 tok/sec/user and led to the announcement of a long-term deal between the two companies, driving their IPO prospects (and the value of sama's stake) ever higher.

虽然 Cerebras 在很大程度上被排除在了新云（neocloud）热潮之外，但 OpenAI 在 2 月份发布的 GPT-5.3-Codex-Spark（该模型采用 gptoss-120B 架构，是从真正的 5.3 Codex 蒸馏而来）正在扭转局面。Spark 在 Cerebras 上的运行速度高达每用户 2,000 tok/sec，并促成了两家公司长期合作协议的宣布，这使得他们的 IPO 前景（以及 sama 所持股份的价值）持续攀升。



Source: SemiAnalysis Tokenomics Dashboard

SemiAnalysis 代币经济学仪表板

Cerebras's chips are only economically capable of serving relatively small models today, or at least based on what's available to the public. GPT-5.3-Codex-Spark, for example, is NOT at all the same thing as the full GPT-5.3-Codex; it's gpt-oss-120b fine-tuned on GPT-5.3-codex traces. In other words, it's a distilled model that's over 10x smaller.

Cerebras 的芯片目前在经济上仅能支持运行相对较小的模型，至少根据目前公开的信息来看是这样。例如，GPT-5.3-Codex-Spark 与完整的 GPT-5.3-Codex 完全不是一回事；它是基于 GPT-5.3-codex 的轨迹进行微调的 gpt-oss-120b。换句话说，它是一个缩小了 10 倍以上的蒸馏模型。

While GPT-5.3-Codex-Spark is really fast, its tokens likely aren't worth $10B today. For OpenAI to run any model above 1T total params with a 1M context window for modern agentic workload patterns, they will need to accept significant tradeoffs on cost (and recoup it by selling those tokens at a significant premium), and we expect the realized performance to be below 1000 tok/sec interactivity. On the other hand, algorithmic improvements will certainly make small models smarter. We're probably less than a year away from GPT 5.5-level intelligence in a 120B form factor.

虽然 GPT-5.3-Codex-Spark 的速度确实很快，但其生成的 Token 在今天可能并不值 100 亿美元。对于 OpenAI 而言，要在现代智能体（agentic）工作负载模式下运行任何参数总量超过 1 万亿且具备 100 万上下文窗口的模型，他们必须在成本上做出巨大妥协（并通过以极高溢价销售这些 Token 来回收成本），而且我们预计其实际性能将低于 1000 tok/sec 的交互速度。另一方面，算法的改进肯定会让小模型变得更聪明。我们距离在 1200 亿参数规模上实现 GPT 5.5 级别的智能，可能只有不到一年的时间。

As mentioned earlier, many of our engineers were willing to forgo the frontier level intelligence of Opus 4.7 in exchange for faster tokens from Opus 4.6 fast. With GPT-5.5, OpenAI finally has an Opus 4.5 level model. Will people be willing to pay for really fast GPT-5.5-quality tokens a year from now even after the true bleeding edge frontier has moved far beyond it? For the first time ever, we think the answer may be yes. While the first 750MW is locked, there is much more upside for Cerebras if OAI

chooses to take the full 2GW or even more. This is all dependent on the quality of the model they can fit on Cerebras hardware.

正如前文所述，我们的许多工程师宁愿放弃 Opus 4.7 的前沿级智能，以换取来自 Opus 4.6 fast 更快的 Token 生成速度。随着 GPT-5.5 的推出，OpenAI 终于拥有了一款达到 Opus 4.5 水平的模型。一年后，即便真正的技术前沿已经远超于此，人们是否还愿意为极速的 GPT-5.5 级 Token 付费？我们认为答案可能是有史以来第一次肯定的。虽然首批 750MW 的容量已经锁定，但如果 OAI 选择使用全部 2GW 甚至更多，Cerebras 将拥有更大的增长空间。这一切都取决于他们能在 Cerebras 硬件上运行的模型质量。

Behind the paywall, we will go through just how the OAI deal's profitability Cerebras and the major execution risk - how far along is Cerebras in securing the DC capacity.

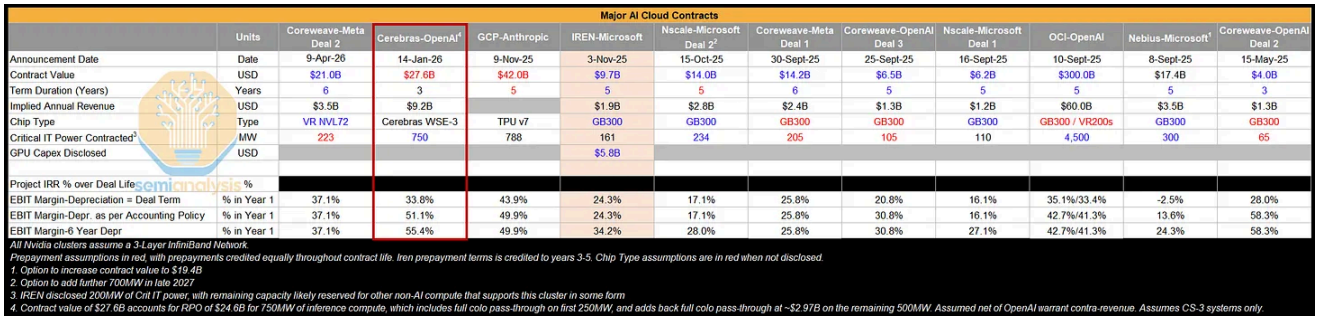
在付费墙之后，我们将详细分析 OpenAI 交易对 Cerebras 的盈利能力影响，以及重大的执行风险——即 Cerebras 在确保数据中心容量方面进展到了何种程度。

## The Cerebras and OpenAI Deal Economics

## Cerebras 与 OpenAI 交易的经济效益

The Cerebras-OpenAI deal earns Cerebras a strong above-average project IRR, driven by the high implied rental rate per CS-3 system. Compared to other major recent cloud deals, which average ~15-25% IRR, the deal looks extremely favorable for Cerebras. One of the big drivers of profitability is Cerebras is hosting their own hardware, unlike a GPU deal where margin is being paid to Nvidia. We will further elaborate on the mechanics of this calculation below.

得益于每台 CS-3 系统极高的隐含租赁率，Cerebras 与 OpenAI 的交易为 Cerebras 赢得了远高于平均水平的项目内部收益率（IRR）。与近期其他主要云交易（平均 IRR 约为 15-25%）相比，这笔交易对 Cerebras 极其有利。盈利能力的一大驱动因素在于 Cerebras 托管的是自有硬件，而不像 GPU 交易那样需要向 Nvidia 支付利润。我们将在下文进一步阐述这一计算机制。



Major AI Cloud Contracts

| | Units | Coreweave-Meta Deal 2 | Cerebras-OpenAI[4] | GCP-Anthropic | IREN-Microsoft | Nscale-Microsoft Deal 2[2] | Coreweave-Meta Deal 1 | Coreweave-OpenAI Deal 3 | Nscale-Microsoft Deal 1 | OCI-OpenAI | Nebius-Microsoft[1] | Coreweave-OpenAI Deal 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Announcement Date | Date | 9-Apr-26 | 14-Jan-26 | 9-Nov-25 | 3-Nov-25 | 15-Oct-25 | 30-Sept-25 | 25-Sept-25 | 16-Sept-25 | 10-Sept-25 | 8-Sept-25 | 15-May-25 |
| Contract Value | USD | $21.0B | $27.6B | $42.0B | $9.7B | $14.0B | $14.2B | $6.5B | $6.2B | $300.0B | $17.4B | $4.0B |
| Term Duration (Years) | Years | 6 | 3 | 5 | 5 | 5 | 6 | 5 | 5 | 5 | 5 | 3 |
| Implied Annual Revenue | USD | $3.5B | $9.2B | | $1.9B | $2.8B | $2.4B | $1.3B | $1.2B | $60.0B | $3.5B | $1.3B |
| Chip Type | Type | VR NVL72 | Cerebras WSE-3 | TPU v7 | GB300 | GB300 | GB300 | GB300 | GB300 | GB300 / VR200s | GB300 | GB300 |
| Critical IT Power Contracted[3] | MW | 223 | 750 | 788 | 161 | 234 | 205 | 105 | 110 | 4,500 | 300 | 65 |
| GPU Capex Disclosed | USD | | | | $5.8B | | | | | | | |
| | | | | | | | | | | | | |
| Project IRR % over Deal Life | % | | | | | | | | | | | |
| EBIT Margin-Depreciation = Deal Term | % in Year 1 | 37.1% | 33.8% | 43.9% | 24.3% | 17.1% | 25.8% | 20.8% | 16.1% | 35.1%/33.4% | -2.5% | 28.0% |
| EBIT Margin-Depr. as per Accounting Policy | % in Year 1 | 37.1% | 51.1% | 49.9% | 24.3% | 17.1% | 25.8% | 30.8% | 16.1% | 42.7%/41.3% | 13.6% | 58.3% |
| EBIT Margin-6 Year Depr | % in Year 1 | 37.1% | 55.4% | 49.9% | 34.2% | 28.0% | 25.8% | 30.8% | 27.1% | 42.7%/41.3% | 24.3% | 58.3% |

*All Nvidia clusters assume a 3-Layer InfiniBand Network.*
*Prepayment assumptions in red, with prepayments credited equally throughout contract life. Iren prepayment terms is credited to years 3-5. Chip Type assumptions are in red when not disclosed.*
*1. Option to increase contract value to $19.4B*
*2. Option to add further 700MW in late 2027*
*3. IREN disclosed 200MW of Crit IT power, with remaining capacity likely reserved for other non-AI compute that supports this cluster in some form*
*4. Contract value of $27.6B accounts for RPO of $24.6B for 750MW of inference compute, which includes full colo pass-through on first 250MW, and adds back full colo pass-through at ~$2.97B on the remaining 500MW. Assumed net of OpenAI warrant contra-revenue. Assumes CS-3 systems only.*

Source: SemiAnalysis AI TCO Model

来源：SemiAnalysis AI TCO 模型

Based on the disclosed information about the Cerebras-OpenAI deal, we calculate an implied rental price of $41.96/hr/CS-3 system. Additional assumptions include a ~4% customer prepay, which stems from the $1B potentially interest-free Working Capital Loan offered to Cerebras. These datapoints allow us to triangulate a high project IRR.

根据披露的 Cerebras 与 OpenAI 交易信息，我们计算出每台 CS-3 系统每小时的隐含租赁价格为 41.96 美元。其他假设包括约 4% 的客户预付款，这源于向 Cerebras 提供的 10 亿美元潜在无息营运资金贷款。这些数据点使我们能够推算出该项目具有较高的内部收益率（IRR）。

**Scenario Settings**

| System Configuration | Cerebras CS-3 System | $478,178 |
|---|---|---|
| Number of Accelerators in Project | [redacted] | |
| Price per Server (as of 2025) | $478,178 | USD |
| Logical GPUs per Server | 1 | |
| Number of Servers | [redacted] | |
| Total Server Capex | | USD |
| System First Production Date | 31-Mar-24 | |
| | | |
| Cost of Capital for NPV | 13% | |
| Customer Prepay % | 4% | |
| Customer Prepay Period Basis | 3 | Years |
| Locked in term | 3 | Years |
| Customer Locked-in Rental | $41.96 | USD/hr/Chip |
| Customer Prepay Amount | $40,000 | USD |
| Physical Chip Expected Lifetime | 5 | Years |
| Shut down EBITDA Margin | -10% | % |
| | | |
| | Proj IRR | [redacted] |

Source: SemiAnalysis AI TCO Model

来源：SemiAnalysis AI TCO 模型

The high rate of return on the Cerebras-OpenAI deal makes more sense when the rental prices are read in the context of the TCO of the CS-3 system.

考虑到 CS-3 系统的总拥有成本（TCO），Cerebras 与 OpenAI 交易的高回报率在租用价格的背景下就显得更加合理了。

The table below outlines the estimated server cost of $453K (~$10K of networking costs have been broken out separately from the previous BOM tables), and an all-in cluster cost of $478K which accounts for miscellaneous server service and installation costs.

下表列出了约 45.3 万美元的预估服务器成本（约 1 万美元的网络成本已从之前的物料清单表中分离出来），以及计入各项服务器服务和安装费用后的 47.8 万美元集群总成本。

**AI Cloud Capital Cost of Ownership**

| AI Cloud Capital Cost of Ownership | | |
|---|---|---|
| | Unit | Cerebras CS-3 System (Internal) |
| Cluster Size | **Chips** | **64** |
| | | |
| **Cluster Capital Costs** | | |
| **Server Cost** | **USD** | **$453,593** |
| Server Service | USD | $5,000 |
| Networking Cost | USD | $10,385 |
| Storage Cost | USD | $0 |
| Software Licenses and Other Costs | USD | $4,200 |
| Other Installation | USD | $5,000 |
| **Service, Networking, Storage, Software, Others** | **USD** | **$24,585** |
| **Total Upfront Cluster Capex, per Server** | **USD** | **$478,178** |
| **Total Upfront Cluster Capex, per Accelerator** | USD | $478,178 |
| | | |
| Equity Cost of Capital | % | 11.0% |
| Debt Cost of Capital | % | 4.5% |
| Equipment Downpayment (Equity Portion) | % | 75.0% |
| | | |
| Weighted Average Cost of Capital | % | 9.4% |
| Useful Life in Years | Years | 5 |
| **Total Cluster Capital Costs per Month per Server** | **USD/mth** | **$10,013** |
| *per Hour per Accelerator* | *USD/hr/Accelerator* | *$13.72* |

Source: SemiAnalysis AI TCO Model

来源：SemiAnalysis AI TCO 模型

Operating costs for the CS-3 system come in at a cluster operating cost per hour of $9.63/hr/CS-3. This largely stems from the 30kW all-in power consumption per CS-3 system, much of which is consumed by the WSE-3.

CS-3 系统的运行成本为每小时每台 CS-3 9.63 美元的集群运行成本。这主要源于每台 CS-3 系统 30kW 的总功耗，其中大部分功耗由 WSE-3 产生。

**AI Cloud Operating Cost of Ownership**

| AI Cloud Operating Cost of Ownership | | |
|---|---|---|
| | Unit | Cerebras CS-3 System (Internal) |
| | | |
| **Cluster Operating Costs** | | |
| Electricity Cost | USD/kWh | $0.0870 |
| Utilization Rate | % | 80% |
| Power Usage Effectiveness (PUE) | Ratio | 1.35 |
| Electricity Cost per kW of Critical IT per mth | USD/kW/mth | $68.6 |
| Colocation Cost | USD/kW/mth | $165.0 |
| **Total Self-Build Monthly Costs** | **M USD/MW** | **-** |
| Total Cost per kW Critical IT Power per Month | USD/kW/mth | $233.6 |
| All-in Power Consumption | kW | 30.00 |
| Total Server Costs per Month | USD/mth | $7,008 |
| Remote Hands + Support Engineer | USD/mth | $16 |
| Internet Connection | USD/mth | $5 |
| **Total Cluster Operating Cost per Month, per Accelerator** | **USD/mth** | **$7,029** |
| **Total Cluster Operating Cost per Month, per Accelerator** | **USD/mth** | **$7,029** |
| *per Hour per Accelerator* | *USD/hr/Accelerator* | *$9.63* |
| | | |

Source: SemiAnalysis AI TCO Model

来源：SemiAnalysis AI TCO 模型

Taken together, the TCO per hour of the CS-3 system adds up to $23.35/hr/CS-3. When compared against the OpenAI-implied rental figure of $41.96/hr/CS-3, it becomes clear why Cerebras is enjoying extremely healthy returns from their OpenAI deal.

综上所述，CS-3 系统的每小时总拥有成本（TCO）累计为 23.35 美元/小时/CS-3。与 OpenAI 暗示的 41.96 美元/小时/CS-3 的租赁价格相比，显而易见，Cerebras 从与 OpenAI 的交易中获得了极其丰厚的回报。

As Cerebras uses the CS-3 both for its internal fleet, as well as for sales to external customers, we tabulate and compare both in the table below. For sales to external customers, we assume an ASP to external customers of $1.3M per CS-3, and assume the OpenAI-implied rental figure of $41.96/hr/CS-3. This leads to a TCO per PFLOP of $1.49/hr/PFLOP for Cerebras's internal fleet, and a TCO per PFLOP of

$2.69/hr/PFLOP for Cerebras's external compute rental customers.

由于 Cerebras 既将 CS-3 用于其内部集群，也将其销售给外部客户，我们在下表中对两者进行了列举和对比。对于向外部客户的销售，我们假设每台 CS-3 的平均售价（ASP）为 130 万美元，并采用 OpenAI 暗示的每小时 41.96 美元/CS-3 的租赁价格。这使得 Cerebras 内部集群的每 PFLOP 总拥有成本（TCO）为 1.49 美元/小时/PFLOP，而 Cerebras 外部算力租赁客户的每 PFLOP TCO 为 2.69 美元/小时/PFLOP。

**AI Cloud Total Cost of Ownership**

| AI Cloud Total Cost of Ownership | | | |
|---|---|---|---|
| | Unit | Cerebras CS-3 System (Internal) | Cerebras CS-3 System (External Customer) |
| | | | |
| Capital Cost per Unit, per Hour | USD/hr/Accelerator | $13.72 | |
| Operating Cost per Unit, per Hour | USD/hr/Accelerator | $9.63 | |
| **Total Cost per Unit per Hour** | **USD/hr/Accelerator** | **$23.35** | **$41.96** |
| Capital Cost as % of Total Ownership Cost | % | 58.8% | |
| | | | |
| Upfront Capital Cost per Server | $ | $478,178 | $1,300,000 |
| Upfront Capital Cost per Accelerator | $ | $478,178 | $1,300,000 |
| | | | |
| Logical GPUs per Server | Accelerators | 1 | 1 |
| | | | |
| Marketed TFLOPS (FP8) | TFLOPS | 15,625 | 15,625 |
| Memory Bandwidth per Accelerator | TB/s | 21,000.0 | 21,000.0 |
| Memory Capacity per Accelerator[1] | GB | 44.0 | 44.0 |
| Marketed TFLOPS (FP8) / Memory Bandwidth | TFLOPS/TB/s | 0.7 | 0.7 |
| | | | |
| Upfront Cluster Cost per PFLOP | $ per PFLOP | $30,603 | $83,200 |
| Upfront Cluster Cost per Memory Bandwidth | $ per TB/s | $23 | $62 |
| | | | |
| TCO per PFLOP | $/hr per PFLOP | $1.49 | $2.69 |
| TCO per Memory Bandwidth | $/hr per TB/s | $0.00 | $0.00 |

*1. Cerebras CS-3 uses SRAM, compared to GB300 NVL72 which uses HBM.*

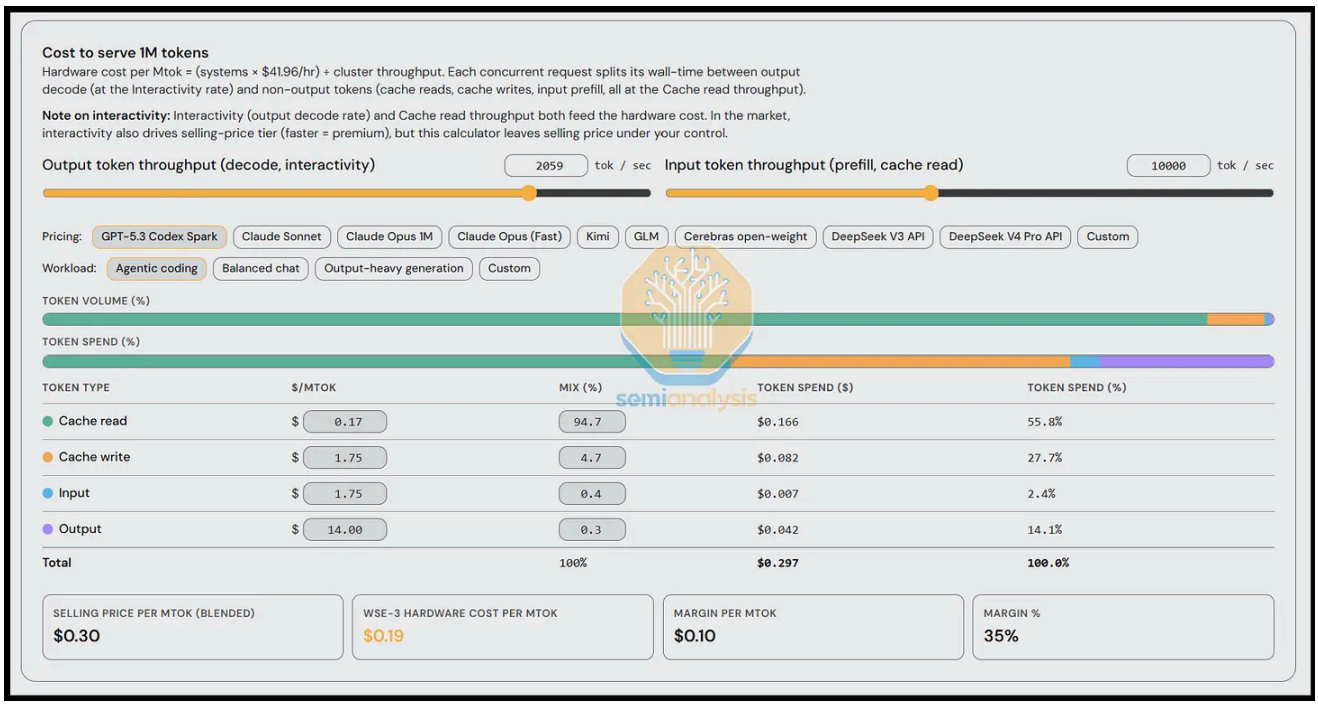
Source: SemiAnalysis AI TCO Model

来源：SemiAnalysis AI TCO 模型

On the flip side, what does this mean for OpenAI's economics? At this assumed hardware rental cost we can see from our dashboard cost per token. This takes into account real values we have observed for these kinds of workloads (sequence lengths, cache read/write ratios). On 5.3 Codex-Spark, for an agentic coding workload, we model cost to serve cost per million tokens as $0.19, compared to $0.30 of token revenue, resulting in a 35% inference gross margin, which is far below frontier model

profitability. There are of course 2 levers to profitability. First is increasing revenue by being able to charge a premium for higher quality tokens served on a higher quality model. The other lever is by decreasing cost by greater software and hardware co-design such as designing workloads to suit the arithmetic intensity of Cerebras's hardware as we mentioned earlier.

另一方面，这对 OpenAI 的经济效益意味着什么？按照这一假设的硬件租赁成本，我们可以从仪表板上看到每 token 的成本。这考虑了我们针对此类工作负载（序列长度、缓存读/写比率）观察到的真实数值。在 5.3 Codex-Spark 上，对于智能体编码工作负载，我们模拟的每百万 token 服务成本为 0.19 美元，而 token 收入为 0.30 美元，这导致推理毛利率仅为 35%，远低于前沿模型的盈利水平。当然，提高盈利能力有两个杠杆：一是通过在更高质量的模型上提供更高质量的 token 来收取溢价，从而增加收入；另一个杠杆是通过加强软件和硬件的协同设计来降低成本，例如像我们之前提到的那样，设计适合 Cerebras 硬件算术强度的工作负载。




Source: SemiAnalysis Tokenomics Dashboard

SemiAnalysis 代币经济学仪表板

# The Trainium / CS-3 Disaggregated PD Setup

# Trainium / CS-3 解耦式 PD 设置

While OAI is the most important development for Cerebras, the partnership that Amazon announced with Cerebras in March 2026 will provide Cerebras with an additional vector of growth. AWS will deploy WSEs in its own datacenters to power Amazon's Bedrock inference service. As part of the partnership, the WSE will be used for decode, with Trainium used for the prefill nodes. Notably, the biggest Trainium customer by far is Anthropic, via Project Rainier. Though Anthropic has not formally been announced as a direct PD Disagg Trainium + Cerebras customer, we expect that to come in due time as demand for fast mode grows. Regardless, Cerebras will likely be used to serve Claude tokens anyway as part of the Bedrock service.

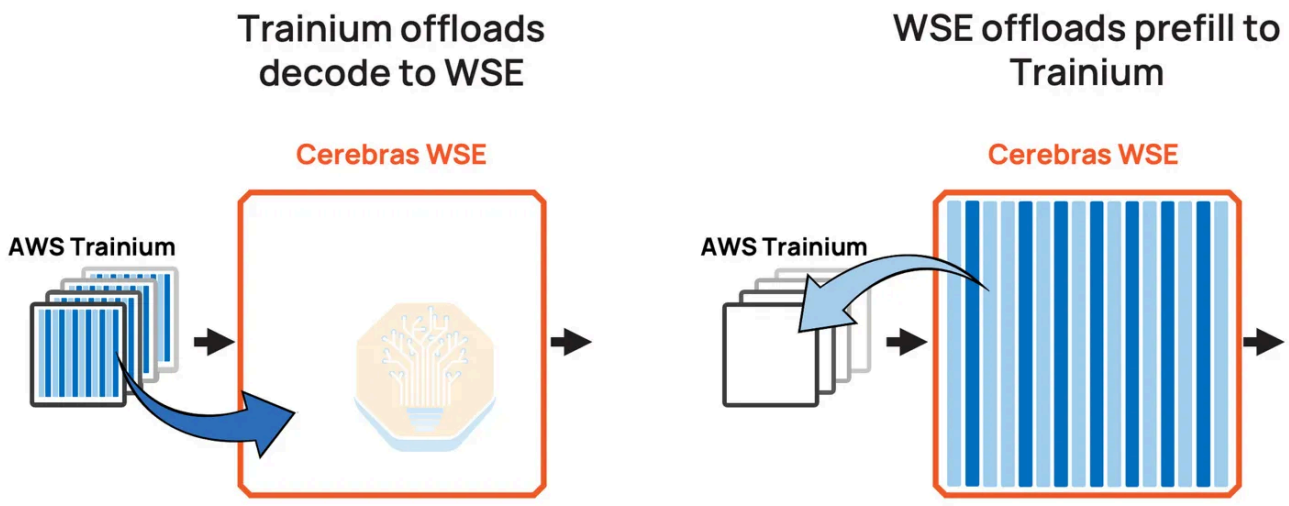
虽然 OAI 是 Cerebras 最重要的发展契机，但亚马逊在 2026 年 3 月宣布与 Cerebras 达成的合作伙伴关系将为 Cerebras 提供额外的增长维度。AWS 将在其自有的数据中心部署 WSE，以为亚马逊的 Bedrock 推理服务提供动力。作为合作伙伴关系的一部分，WSE 将用于解码（decode），而 Trainium 则用于预填充（prefill）节点。值得注意的是，目前 Trainium 最大的客户是通过 Project Rainier 合作的 Anthropic。尽管 Anthropic 尚未正式宣布成为 PD 分离式（PD Disagg）Trainium + Cerebras 的直接客户，但随着对"快速模式"需求的增长，我们预计这只是时间问题。无论如何，作为 Bedrock 服务的一部分，Cerebras 很可能都会被用于提供 Claude 的 Token 服务。

We've already written about disaggregated inference in this article, as well as in both of our InferenceX articles (here, and here), but the quick recap is that decode is memory bandwidth and latency constrained while prefill is compute constrained. Thus, Cerebras will be used for decode while Trainium is used for prefill. This is **not** AFD like NVIDIA's Groq LPU announcement at GTC. It is standard PD disagg. All decode will run on Cerebras.

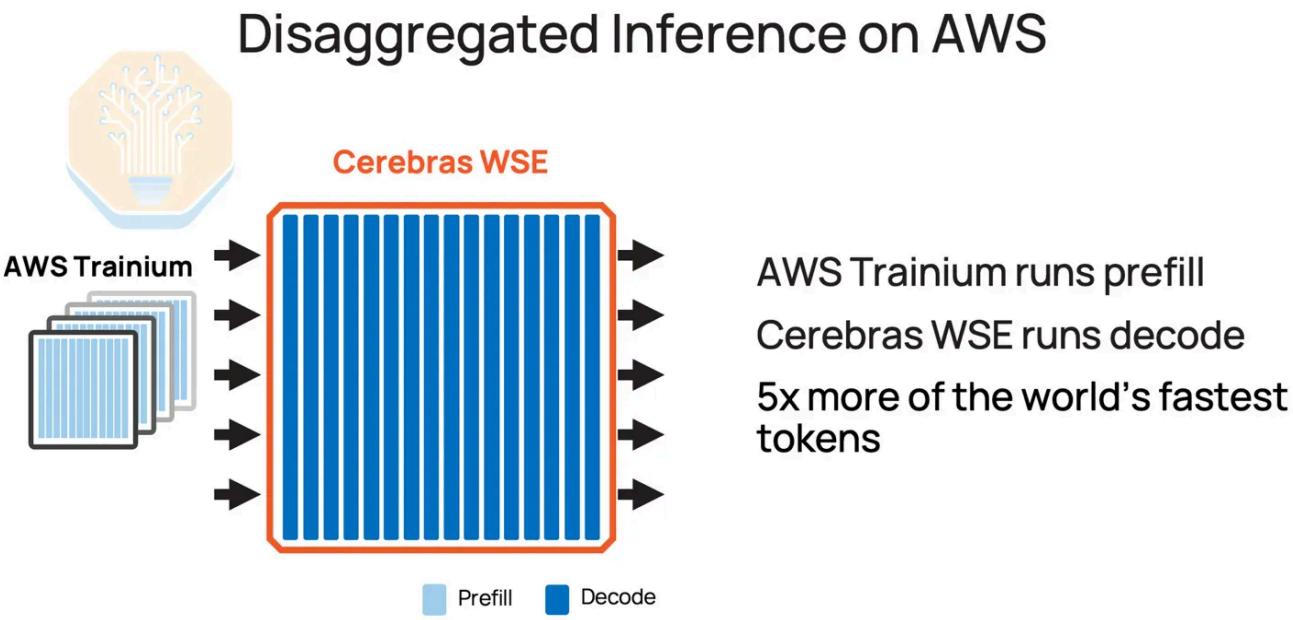
我们已经在本文以及两篇 InferenceX 文章（[此处](here)和[此处](here)）中探讨过分离式推理，简单回顾一下：解码受限于内存带宽和延迟，而预填充则受限于算力。因此，Cerebras 将用于解码，而 Trainium 用于预填充。这与 NVIDIA 在 GTC 上发布的 Groq LPU 式的 AFD 不同，这是标准的 PD 分离架构。所有的解码工作都将在 Cerebras 上运行。

The joint announcement claims a 5x throughput improvement using this setup.

联合公告声称，使用这种配置可实现 5 倍的吞吐量提升。




Source: https://www.cerebras.ai/blog/cerebras-is-coming-to-aws




Source: https://www.cerebras.ai/blog/cerebras-is-coming-to-aws

来源：https://www.cerebras.ai/blog/cerebras-is-coming-to-aws

Similar to the OpenAI deal, Cerebras issued a warrant giving AWS the right to purchase up to a maximum 2.7M shares at an exercise price of $100 per share, contingent on AWS purchasing sufficient volumes. Unlike the OpenAI arrangement which is structured as cloud-based revenue, we believe the AWS deal will be recognized predominantly as hardware sales revenue, as Cerebras is selling CS-3 into

AWS-owned data centers.

与 OpenAI 的交易类似，Cerebras 发放了一份认股权证，赋予 AWS 以每股 100 美元的行权价格购买最多 270 万股股票的权利，前提是 AWS 购买足够的量。与结构为云服务收入的 OpenAI 协议不同，我们认为 AWS 的交易将主要被确认为硬件销售收入，因为 Cerebras 是将 CS-3 销售到 AWS 自有的数据中心。

On the technical merits, the disaggregated architecture needs to be proven at scale. The proposed setup has Trainium chips handle prefill, generate the KV cache, then transfer it to the WSE for decode, which needs to overcome around 5 microseconds of latency on every switch hop.

从技术层面来看，这种解耦架构（disaggregated architecture）的大规模应用仍需验证。拟议的方案由 Trainium 芯片处理预填充（prefill）并生成 KV 缓存，随后将其传输至 WSE 进行解码（decode），这需要克服每一级交换机跳数约 5 微秒的延迟。

As we have been harping on in this article, the potential bottleneck when using the WSE for PD disagg is the I/O, and no amount of disaggregation changes the fact that all KV Cache must pass through the same switches and same network interfaces to get from Trainium to the wafer.

正如我们在本文中一直强调的，在将 WSE 用于 PD 解耦（PD disagg）时，潜在的瓶颈在于 I/O。无论如何解耦，都无法改变这样一个事实：所有的 KV Cache 必须通过相同的交换机和网络接口，才能从 Trainium 传输到晶圆（wafer）上。

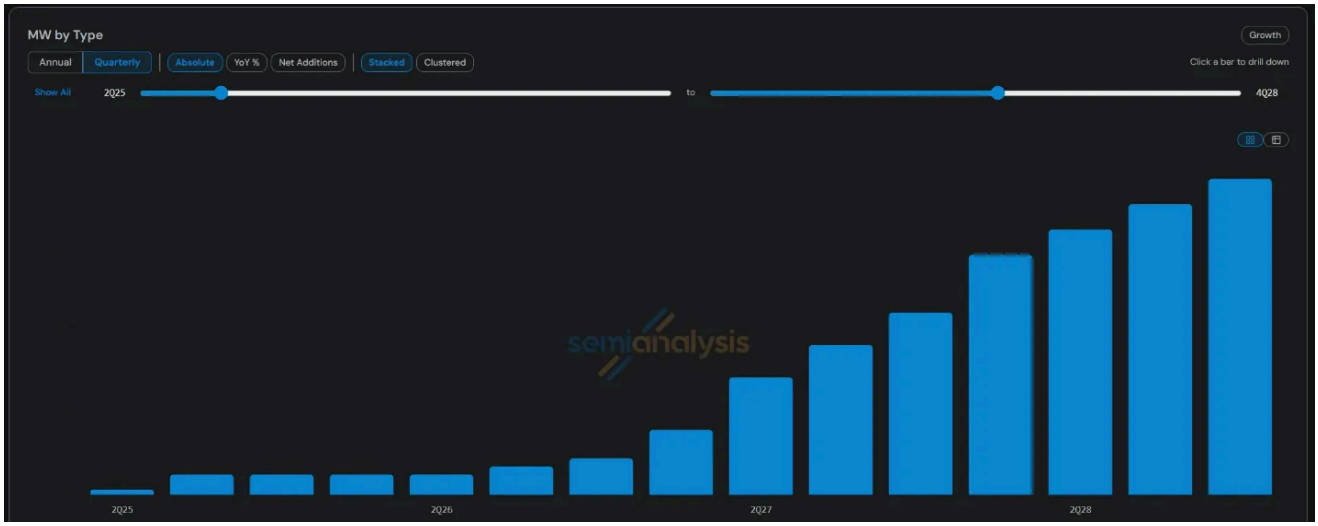
## Can they ship? 他们能顺利出货吗?

As if designing a chip, system, programming model, software runtime, and serverless inference endpoint sales business wasn't hard enough, Cerebras has also decided to become a Neocloud as Cerebras needs to host all this compute for OpenAI, at a minimum the first 250MW of it. It needs to secure a lot of power and soon.

仿佛设计芯片、系统、编程模型、软件运行时以及无服务器推理终端销售业务还不够艰巨，Cerebras 还决定转型为一家 Neocloud（新型云服务商），因为 Cerebras 需要为 OpenAI 托管所有这些计算资源，起步规模至少为 250MW。它需要尽快锁定大量的电力供应。

Unfortunately, we believe that to deliver on their commitments to OpenAI, Cerebras is currently below the mark on near-term datacenter capacity. The OpenAI deal alone requires 250MW each year in 2026–2028, but our Datacenter Model subscribers learned a month ago that Cerebras likely only has ~180MW lined up by YE2027 (note: the OpenAI deal excludes G42, and as a result the 40MW at UAE Stargate, leaving roughly 140MW for OpenAI). That's a meaningful shortfall against contracted demand.

遗憾的是，我们认为 Cerebras 目前在近期数据中心容量方面尚未达到履行其对 OpenAI 承诺的标准。仅 OpenAI 的交易就要求在 2026 年至 2028 年间每年提供 250MW 的容量，但我们的“数据中心模型”订阅用户在一个月前获悉，Cerebras 到 2027 年底可能仅筹备了约 180MW 的容量（注：OpenAI 的交易不包括 G42，因此也排除了阿联酋 Stargate 的 40MW，仅为 OpenAI 留下了约 140MW）。相对于合同需求而言，这是一个显著的缺口。




Source: SemiAnalysis Datacenter Industry Model

SemiAnalysis 数据中心行业模型

In 2026, Cerebras is likely fighting an uphill battle. We've heard 2026 capacity should sit in Cerebras's own (leased) datacenters, and we've identified only ~43MW by YE2026 in leased capacity so far (50MW+ on the high end based on the S-1). Cerebras's AWS capacity is likely already being deployed and serves OpenAI as well, but judging from

the S-1, it's likely no larger than ~10MW.

到 2026 年，Cerebras 可能会面临一场艰苦的战斗。据我们所知，2026 年的产能应该部署在 Cerebras 自己的（租赁）数据中心，而到 2026 年底，我们目前确定的租赁容量仅为约 43MW（根据 S-1 文件，高端估计为 50MW+）。Cerebras 的 AWS 容量可能已经开始部署并同样为 OpenAI 提供服务，但从 S-1 文件判断，其规模可能不超过 10MW。

2027 improves when their Bell AI Labs campus (128MW) comes online, and potentially further if their 100MW Guyana Sovereign AI campus (MoU) begins construction soon — though Guyana has been delayed multiple times. We've heard of a 65MW facility under construction with Cerebras involved but have yet to locate or determine Cerebras's capacity.

到 2027 年，随着其 Bell AI Labs 园区（128MW）上线，情况会有所改善；如果其 100MW 的圭亚那主权 AI 园区（谅解备忘录阶段）能尽快开工，情况可能会进一步好转——尽管圭亚那项目已经多次推迟。我们听说有一个 65MW 的设施正在建设中，且 Cerebras 参与其中，但目前尚未定位该设施或确定 Cerebras 所占的容量。

Bell AI Campus – April 30, 2026. Source: SemiAnalysis Datacenter Industry Model

Bell AI 园区 – 2026 年 4 月 30 日。来源：SemiAnalysis 数据中心行业模型

OpenAI can host Cerebras in existing OpenAI facilities for 2027 capacity, but we've flagged out to Datacenter Model subscribers that even 1H2027 capacity is looking sold out now, with 2H2027 quickly being sold as well. For now, we've exhausted all known, disclosed locations.

OpenAI 可以在现有的 OpenAI 设施中托管 Cerebras 以满足 2027 年的产能需求，但我们已向“数据中心模型”订阅用户发出提醒，目前甚至 2027 年上半年的产能也已售罄，下半年的产能也正在被迅速抢购。目前，我们已耗尽了所有已知且已披露的选址。

What can Cerebras do? Right now, it's a seller's market. Anyone with a credit line measured in billions (see: Neoclouds, AI Labs, Hyperscalers) is taking whatever

capacity (see: powered land, shells, bare metal GPUaaS) they can get. However, we believe that serving inference capacity means Cerebras can take on smaller, disparate capacities instead of larger hyperscale sites. We have heard this is exactly what's happening, with Cerebras asking for multiple sites and being open to modular and pre-fab builds, with less concern on pricing. After all, DC costs are passed through to OAI, and we understand that OAI is allowing Cerebras to pay up to $200/kW/month, which is a good amount above the going rate of $130-140/kW/month. We don't doubt they may find capacity, but it may be expensive (and painful), especially when thinking of the customized liquid cooling infrastructure required. We covered the thermal architecture in 3e; the same cooling constraints (custom CDUs, ~4 LPM/kW facility flow, chiller-heavy inlet temperatures) are why standard liquid-cooled DC capacity is not drop-in for Cerebras.

Cerebras 能做什么？目前这是一个卖方市场。任何拥有数十亿美元信用额度的机构（如：新型云服务商 Neoclouds、AI 实验室、超大规模云厂商）都在抢夺任何能拿到的资源（如：带电土地、机房外壳、裸金属 GPUaaS）。然而，我们认为，提供推理算力意味着 Cerebras 可以利用较小的、分散的资源，而不必非要追求超大规模站点。我们听说这正是目前正在发生的情况：Cerebras 正在寻求多个站点，并对模块化和预制化建筑持开放态度，且对价格不太敏感。毕竟，数据中心成本会转嫁给 OpenAI，据我们了解，OpenAI 允许 Cerebras 支付高达 200 美元/kW/月的租金，这远高于目前 130-140 美元/kW/月的市场价。我们不怀疑他们能找到产能，但这可能会非常昂贵（且过程痛苦），特别是考虑到所需的定制化液冷基础设施。我们在 3e 部分介绍过其散热架构；正是由于这些冷却限制（定制 CDU、约 4 LPM/kW 的设施流量、对冷水机组依赖度高的入口温度），使得标准的液冷数据中心无法直接适配 Cerebras。

---

Recommend 推荐

75 Likes 75 次赞 · 6 Restacks 6 次转推

Previous 上一篇

## Discussion about this post

关于此帖的讨论

Comments Restacks

Write a comment...

Tanj Bennett 1h

Author

Hi Ignacio, thankyou for your very nice blog. I have learned a lot from you over the years.

You definitely have found an arithmetic error. I well check with the author of that chart to see what is going on. We will update it.

LIKE REPLY SHARE

Gregg McKnight Gregg McKnight 2h

I'm surprised you didn't address wafer size scaling and its implications to future system performance? With wafer size at its physical limit and with limited networking, where do future Cerebras iterations go from here? Clearly GPUs, TPUs etc will increase transistor density thru increased die size and packaging. But what does Cerebras do to maintain competitiveness? Or is this why they are moving so quickly to IPO? Its difficult to see a roadmap for sustained leadership when the starting point is end-game wafer size.

LIKE REPLY SHARE

2 more comments...