Cerebras 财报与公司深度研究

Cerebras 自己设计芯片，委托台积电制造晶圆芯片，再与供应链伙伴做成整套计算设备；它可以把设备卖给客户，也可以使用设备提供云算力，另收系统支持和模型工程服务费。2025 年 5.10 亿美元收入扣除交付成本后留下 1.99 亿美元毛利，尚不足以覆盖 3.45 亿美元研发、销售和管理费用；2026 年上半年又投入 5.49 亿美元现金购建设备。理解它的关键，是先看收入能留下多少利润、这些收入需要占用多少资源，再判断技术和订单能否改善回报。[1][2]

研究截至 **2026 年 9 月 12 日**。公司为 Cerebras Systems Inc.，NASDAQ：**CBRS**，2016 年在美国特拉华州注册，总部位于加州 Sunnyvale。财务基准为 2026 年第二季度 10-Q 和最终招股书；行情截至 9 月 11 日美股收盘。全文按“利润表总貌 → 收入与成本结构 → 资产投入与现金 → 商业模式和竞争 → 估值”展开，金额除特别注明外均为美元。[1][2]

---

1. 从主表看：收入留下了多少利润

主营业务已有毛利，但尚未覆盖研发、销售和管理开支。 先读下面这条链：收入减营业成本得到毛利；再减期间费用得到营业利润；加上其他损益、减所得税，才得到公司净利润。把这几层分开，才能解释为什么 2025 年营业亏损与净利润为正同时出现。[1][2]

百万美元，毛利率除外	2024全年	2025全年	2026上半年	2026第二季度
收入	290.3	510.0	373.5	180.1
营业成本	167.5	310.9	261.8	154.6
毛利	122.7	199.1	111.7	25.6
毛利率	42.3%	39.0%	29.9%	14.2%
研发费用	158.2	243.3	395.6	320.2
销售及营销费用	21.0	70.6	101.7	87.0
管理费用	45.0	31.0	106.7	95.7
期间费用合计	224.2	344.9	604.0	502.8
营业利润	-101.4	-145.9	-492.3	-477.2
其他净收益 / 损失	-378.2	390.7	29.5	27.0
税前利润	-479.7	244.9	-462.8	-450.3
所得税费用	1.9	7.1	1.8	0.3
公司净利润	-481.6	237.8	-464.5	-450.5

来源：[1：MD&A第99页及经审计损益表][2：第7页损益表]。单位百万美元，毛利率除外；年度、半年和单季并列展示规模，不能直接比较为环比增长，Q2 已包含在 H1 中。表中净利润为公司整体口径，归属普通股口径另在第7节解释。

2025 年每取得 **100 美元收入**，约有 **61.0 美元营业成本**，留下 **39.0 美元毛利**；还需承担 **47.7 美元研发、13.9 美元销售营销、6.1 美元管理费用**，因此营业层面约亏损 **28.6 美元**。各比例按原始金额计算后独立舍入。研发费用本身已经超过全部毛利，这是现阶段经营结果最直接的解释。[1；研究计算]

净利润转正主要来自营业以外的项目：2025 年其他净收益为 3.91 亿美元，其中包含 3.63 亿美元远期合同负债终止收益，后者不带来当期现金。到 2026Q2，3.77 亿美元股权激励又使当季费用集中增加，其中有 IPO 触发的确认。两类事项都需要在主表基础上解释，不能把净利润直接当作稳定挣钱能力，也不能把单季巨亏机械年化；详细对账保留在第7节。[1][2][3]

---

2. 拆收入与成本：客户买什么，毛利从哪里来

2.1 先看一台计算设备怎样生产、怎样使用

Cerebras 先把自己的芯片设计变成整套计算设备，再通过出售设备、运营设备提供算力，以及配套工程服务收钱。“晶圆级”描述芯片的制造和集成方式；“推理”指用已经训练好的模型处理请求、生成结果。晶圆芯片装进设备后可以反复进行计算，一次推理消耗的是电力、设备使用时间和相关运维资源，不能把一次推理理解为消耗一片原材料晶圆。[1：业务、Manufacturing and Suppliers][2：MD&A第30–32页；机制解释]

制造环节的实际分工很明确：Cerebras 负责设计，**台积电（TSMC）制造 WSE 晶圆芯片，日月光（ASE）承担重布线层等特定工艺**，Cerebras 在 Sunnyvale 管理最终晶圆封装、装配和测试。它还向第三方采购服务器、网络交换机、互连、冷却单元、机箱等部件；代工伙伴完成整机组装与测试，Cerebras 做最终整机老化测试和质量验证。因此，对 Cerebras 而言，制造投入主要体现为向供应商购买芯片、部件及加工服务，同时承担部分自有装配、测试和工程人员成本。[1：第32页供应链风险、第145页Manufacturing and Suppliers]

这里的 **fabless** 是不自建晶圆制造厂，不代表没有制造采购，也不代表提供云服务时不需要持有设备资产。台积电制造厂的投入会通过供货价格影响 Cerebras；Cerebras 留作自营云服务的计算设备，则是另一层资产投入。[1：Manufacturing and Suppliers、固定资产会计政策；机制解释]

2.2 三条收费路径，对应三种资源消耗

实际生意	客户得到什么、谁持有设备	Cerebras 具体向谁付钱	成本怎样进入损益
做出整套设备并卖给客户	客户买下系统，放进自己的数据中心使用；按合同交付或验收	台积电的晶圆芯片、日月光等加工服务商、部件供应商、整机代工厂、物流商，以及装配测试和质保人员	合格生产采购支出先形成存货，交付时结转成本；还承担运输、质保、良率损失及存货减值等
保有或取得设备使用权，提供云推理	客户买可用算力或模型调用；Cerebras 承担约定的设备和服务可用性，设备不因一次调用而转移给客户	设备供应商、设备或数据中心出租方、电力与冷却等设施服务商、运维及客户支持人员	自有设备分期折旧；租入设备和场地发生租赁费用；持续承担电力、设施运营和人员成本
帮客户管理系统、支持软件或定制模型	可以在客户持有的系统上提供服务；客户购买工程交付和持续支持	软件和模型工程师、系统运维及支持人员，以及完成调试、测试所需的计算资源	随工程和支持服务交付发生人员、计算和设施等成本；合同决定收入按期还是验收确认

上表结合公司披露解释生产和交付过程。前一条对应“硬件”收入；后两条主要合并在“云及其他服务”，未分别披露完整收入与成本金额。公司同时披露自有设备资产和租赁费用，不能据此确定每个云项目的设备都自有、或都租入；自有设备折旧与租入设备租金是不同资源取得方式。[1：存货及固定资产会计政策][2：第6–7页、附注8和15、MD&A第31–32页]

云服务也有两种不同账单。**专用容量（Dedicated Capacity）**通常按 take-or-pay 合同收费：客户即使没有充分使用，也须按约支付容量费用；**按需调用（On-Demand）**则主要按 token 用量等收费。对前者，关键是容量能否按约交付并被有效合同覆盖；对后者，调用量会更直接影响收入。同一台机器服务更多付费请求，可以摊薄固定投入，但仍会增加电力、资源占用等支出。[2：MD&A第31页；末句为成本机制解释]

还有设备出租这一种合同安排：公司2026年6月与 AWS 签署硬件租赁协议，但截至6月30日租赁尚未开始、尚未确认相关收入。不能用这个后续安排推断现有全部云服务收入都是出租设备所得。[2：MD&A"Recent Developments—AWS"]

2.3 先确认每条生意留下多少毛利

期间	收入类别	收入	营业成本	毛利	毛利率
2024全年	硬件	212.0	137.3	74.7	35.2%
2024全年	云及其他服务	78.3	30.2	48.1	61.4%
2025全年	硬件	358.4	204.7	153.7	42.9%
2025全年	云及其他服务	151.6	106.2	45.4	29.9%
2026上半年	硬件	164.7	118.1	46.6	28.3%
2026上半年	云及其他服务	208.8	143.7	65.1	31.2%
2026第二季度	硬件	54.1	53.1	1.0	1.8%
2026第二季度	云及其他服务	126.0	101.4	24.6	19.5%

单位百万美元；数据来自[1：损益表、MD&A第101–102页][2：第7页损益表]，毛利及毛利率按同一类别收入减对应成本计算。它们尚未扣除公司层面的研发、销售和管理费用。

把2025年还原为生意：客户购买整套硬件支付 **3.5844 亿美元**，对应制造采购、交付、质保等成本合计 **2.0475 亿美元**，留下 **1.5369 亿美元毛利**。云算力、系统支持和其他服务取得 **1.5155 亿美元收入**，相应设备使用、机房运营和服务交付成本合计 **1.0617 亿美元**，留下 **0.4538 亿美元毛利**。2025年硬件仍贡献约77%的总毛利；这些毛利还需要负担下一代产品开发和获客组织。[1；研究计算]

这些是业务成本合计，不是完整材料清单。 原始财报没有把2.0475亿美元继续拆成"晶圆多少钱、其他部件多少钱、运费多少钱"，也没有把1.0617亿美元完整拆成"机房租金、电费、折旧、运维工资各多少钱"。因此可以确认投入的种类和总额，不能给每项采购编出占比。[1][2]

2.4 研发、销售和管理费用究竟买了哪些东西

主表中的费用	实际发生的活动和资源	能确认的金额与边界
研发	聘用芯片、系统和软件工程师；支付芯片设计版图、掩膜和流片试制费用；做原型、可靠性和系统测试；为模型适配及软件开发使用计算设备、云算力和AI工具	2025年2.4332亿美元；2026H1为3.9565亿美元，其中股权激励2.2785亿美元。现金薪酬、流片、原型和其他开支未完整分拆
销售及营销	销售人员工资、佣金和股权激励；面向客户的产品演示与免费试用算力；展会、公关、广告及差旅	2025年7,064.5万美元；2026H1为1.0167亿美元，其中股权激励7,284.7万美元。广告投流金额、销售薪酬和试用成本未单独完整披露
管理	财务、法务、HR和IT团队；审计、合规、律师与咨询服务、办公软件和保险等	2025年3,096.9万美元；2026H1为1.0669亿美元，其中股权激励6,960.5万美元。各类现金服务费及薪酬未完整分拆

来源：[1：MD&A第96–103页][2：附注13、MD&A第32及36–37页]。股权激励是人员报酬的一种形式，包含IPO触发的确认；并非同期现金工资，也不能据此把扣除股权激励后的余额全当作现金支出。

这也解释了为什么同样是“用机器跑模型”，可能进入不同科目：**为付费客户交付云服务的资源进入营业成本，为潜在客户做免费试用的资源可以进入销售费用，为下一代产品和软件做实验的资源进入研发费用。** 公司在H1费用解释中明确提到客户试用的云支出下降。分类由经营目的决定，不能看见“销售费用”就认为全部是广告获客。[2：MD&A第32和36页；按披露归纳]

2.5 数字还能拆到哪里，哪些仍缺披露

下面统一使用2026年上半年，避免把不同期间的成本拼在一起。表中同时存在“按用途分类”和“按费用性质分类”，**各行不是可以直接相加的成本清单**。

已披露项目，百万美元	金额	可以回答什么、还不能回答什么
硬件营业成本	118.072	整套系统采购制造与交付总成本；未披露晶圆、部件、代工、物流的逐项金额
云及其他服务营业成本	143.709	算力与配套服务交付总成本；未将纯云推理、运维和模型服务分别拆开
公司租赁费用合计	115.008	含长期经营租赁55.641、短期54.586、可变4.781；范围含总部、办公室和数据中心，不能全归到云业务，也不等同现金租金付款
公司折旧合计	42.551	其中28.353计入营业成本、11.568计入研发、1.639计入销售、0.991计入管理；没有进一步披露全部硬件/云业务分摊
计入营业成本的股权激励	16.303	是上述261.781营业成本中的一部分；现金工资及福利未单列
存货减值与报废相关费用	约19.2	产品代际切换等导致此前投入不能充分收回；并非新增现金采购
产品质保准备新增	8.430	当期新增预计质保负担；本期准备使用额5.751，期末准备12.047，均不能直接代替现金维修支出

来源：[2：损益表、附注8、13、15]。租赁总额涵盖多个用途；电力等可能已包含在某些租赁或设施合同收费内，另加“电费”前必须核对合同范围。公司尚未披露足够信息来完整还原“机房租金＋电费＋折旧＋运维工资”的单独金额与相互归属。

OpenAI部分合同还要求公司先支付机房装修、月租、安保、电力等，再按约向客户收取相关款项。它们可能同时进入收入和营业成本。2026Q2公司对账列出相关代垫收入14.503百万美元、成本14.075百万美元；这组数字反映特定合同的重收费方式，不是全公司的机房成本，也不是应在租金、电力之外重复加一遍的成本。[1：收入附注F-19–F-20][3：Revenue及Operating income对账]

下一步应追的是具体采购和产能指标：晶圆与整机的单位采购成本、良率和质保负担；已投产设备原值及寿命；租赁容量、有效电价和实际用电量；每位工程师能维护的容量；付费合同覆盖与调用量。它们把费用和真实业务活动连接起来。现有公开资料尚不足以给出完整的单台系统、每百万token或每MW成本，报告不以未经支持的占比填补。[研究归纳]

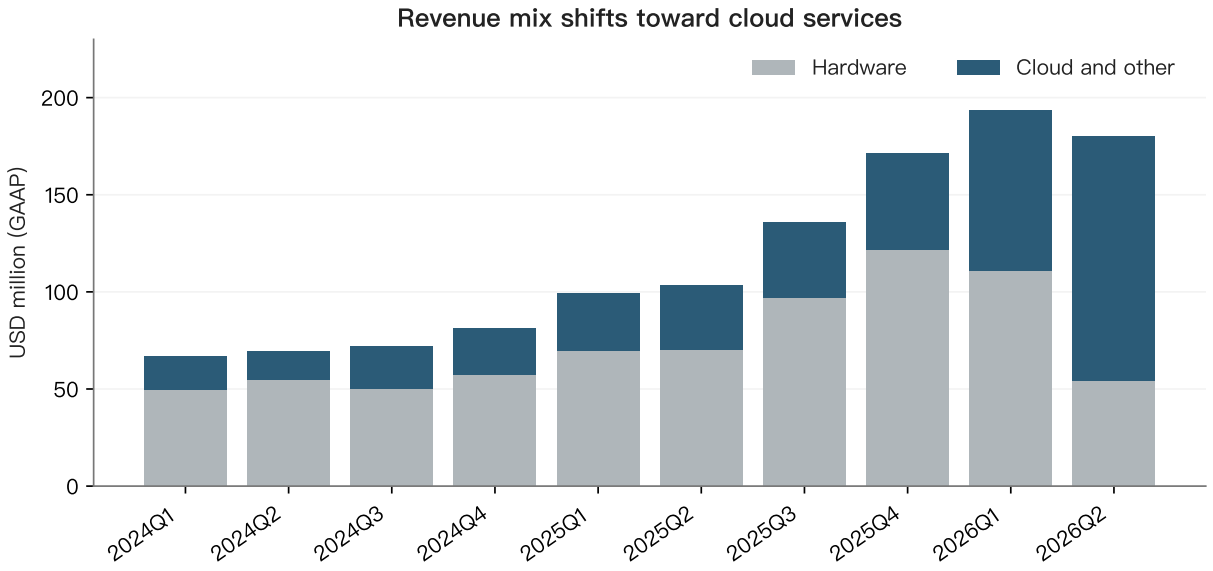


图1：硬件与云及其他服务收入，单位百万美元，均为GAAP历史值。2026Q1由半年值减第二季度得到。来源：[1] [2]。云类别包含其他服务及特定代垫收入，不应整体视为标准化软件订阅收入。

3. 对照资产负债表：这些收入需要投入什么资源

3.1 从收入类别找到对应资产

设备和机房投入正在增加，现金充裕主要反映融资能力。 卖硬件需要先备料、生产，再把存货交付给客户；自己运营云服务，则需要把设备留在资产端，在提供服务的期间承担折旧及场地成本。资产负债表把这两种交付方式的资源占用区分开来。[1：会计政策与MD&A][2：第6页、附注8及MD&A；机制解释]

资源项目，百万美元	2025-12-31	2026-06-30	与经营的连接
现金及短期投资	1,108.2	7,921.5	支持后续采购和运营；主要由融资补充
受限现金	228.7	684.7	用途受约束，不能等同于自由现金
应收账款净额	50.4	123.4	交付和确认收入后尚待收款
存货净额	63.6	32.7	备料、在制与待交付产品；有减值风险
预付及其他流动资产	31.8	191.7	提前支付费用及其他经营性权利
固定资产净额	437.4	986.8	设备、装修及在建工程；分期消耗
经营租赁使用权资产	249.0	528.3	租赁资源的使用权，伴随付款义务
客户权证资产	152.4	1,128.4	客户权益对价，后续摊销抵减收入

来源：[2：第6页资产负债表、附注8和12]。单位百万美元；表中资源项目为选列，不是完整资产负债表。现金与短期投资已合并，客户权证包含流动及非流动部分；其余未列示项目及完整资产、负债与权益恒等式见 Excel 的 Operating assets 表。

固定资产（Property and equipment）净额由 **4.37 亿美元增至 9.87 亿美元**。其中，设备原值、在建工程与累计折旧应进一步分开：

固定资产明细，百万美元	2025-12-31	2026-06-30	经营含义
数据中心及计算机设备原值	277.4	403.6	已归类到设备的投入，不等于可计费容量或满负荷运转
机器设备及租赁装修原值	47.7	70.2	生产、测试及场地配套资源
在建工程	163.5	604.0	已投入但尚待达到预定可用状态，需跟踪转固、验收和计费
固定资产原值合计	488.6	1,077.8	在建工程已包含在内，不能再次相加
减：累计折旧	51.2	91.0	历史投入已经分期计入成本费用的部分
固定资产净额	437.4	986.8	原值减累计折旧后的账面余额

来源：[2：附注8第20–21页]。6 月底在建工程约占固定资产原值的 **56.0%**，因此账上资产增长还需经过投产、交付和计费，才能证明收入回报。达到可用状态后的设备折旧又会进入后续成本；本期低折旧不等于全生命周期成本低。原文各原值细项合计与列示总额有 0.001 百万美元差异，模型保留原数并注明，表中按一位小数显示。[1：固定资产会计政策][2；研究计算与机制推断]

3.2 存货和无形资产分别告诉我们什么

6 月底存货净额 **3,267 万美元**，包括原材料约609万、在产品75万、产成品2,582万美元。存货由年底6,363万美元下降，但上半年同时确认了约 **1,920 万美元存货减值及报废相关费用**，公司解释主要与产品代际切换有关；期末减值准备约2,010万美元。余额下降不能直接解释为周转改善，减值费用也不能当作新增现金采购支出。[2：附注8]

主表没有单列无形资产金额，**这不表示没有技术积累**。公司披露研发支出在发生时费用化，软件在达到技术可行性前的开发成本也计入费用。芯片设计、软件适配、流片、原型和测试等资源，大量体现为前述研发费用，而没有形成同等金额的账面无形资产。2025 年研发费用2.43亿美元说明投入规模，但不能直接当作技术资产估值，也不能再次加进现金资本开支。[1：MD&A第96–97页][2：MD&A第32页]

客户权证资产另需辨认：6 月底余额 **11.28 亿美元**，代表给予客户的权益对价按会计规则递延，并随相关收入摊销、抵减收入；它不是现金、机器或研发成果。经营租赁使用权资产 **5.28 亿美元**则代表租赁资源的使用权，背后有已确认租赁负债及未来付款，不能因场地不是自有就忽略资本与合同约束。[2：第6页、附注12及15]

3.3 利润、资产和现金怎样连起来

三张表的连接是：**先投入设备与营运资金 → 交付系统或算力并确认收入 → 将存货成本、折旧及运营费用计入损益 → 按合同收款或结算。** 这些时点可以不同。6 月底应收账款净额1.23亿美元，说明部分已确认收入尚待收取；客户预付款2.43亿美元和递延收入4.18亿美元则对应尚待履行的义务。递延收入还包含融资安排产生的非现金调整，不能全当作已经收到的预付款。[2：第6页、附注3及10]

上半年现金购买固定资产5.49亿美元，而折旧费用仅4,255万美元，其中2,835万美元进入营业成本。现金支出和当期费用的差异来自资本化及跨期消耗；固定资产净额增加也不等于现金资本开支，因为还有应付设备款、资产转入转出等。经营租赁的费用、付款与使用权资产变动同样不能重复相加。[2：第11–12页现金流量表及附注8]

2026上半年经营现金流为−47.488百万美元，购买物业及设备支出548.873百万美元，合计FCF为**−596.361百万美元**。过去十二个月同口径FCF约−680.213百万美元。云业务需要先安装设备和取得场地，因此“经营现金流接近零”不能代表资本需求很小。[1][2；研究计算]

2024年的正经营现金流也需要拆解：当年客户预付款变动带来约640.3百万美元流入，超过452.0百万美元经营现金流本身。2025年客户预付款变动则为−285.9百万美元。预收资金能帮助融资，但随后履约仍要付出成本，不能把预付款带来的现金流优势当作已经实现的利润。[1：现金流量表]

2026-06-30流动性 / 义务	百万美元	估值和风险处理
现金及现金等价物	6,742.2	与短期投资合计7,921.5
短期投资	1,179.4	计入未受限流动资产
受限现金	684.7	估值桥未作为自由现金扣除
OpenAI贷款本金	918.2	作为债务；部分通过服务抵扣偿还
已确认经营租赁负债	568.8	可比EV口径加入；另列不含租赁EV
截至6月底尚未开始的租赁最低付款	约1,500	未折现、跨合同期，不是即期负债
第三季度追加签署租约最低付款	753.4	与上一项合计约2,300，不再重复加一次

来源：[2：资产负债表、附注10、15、17及MD&A]。未开始租赁未来会上表；不能把未折现长期付款与即时融资缺口等同，但也不能在看净现金时忽略它们。

现金主要来自资本市场和融资。2026上半年IPO扣除承销折让后的融资现金流为62.33亿美元，H轮约10.14亿美元，OpenAI贷款约10.05亿美元；同时存在IPO及要约相关税款代扣现金流出约4.17亿美元。公司另有8.5亿美元循环信贷额度，截至季末未借款，额度不等于账上现金。[2]

在解释资本强度时，应同时看两件事：现有合同覆盖多少已投产容量，以及这些合同带来的毛利能否覆盖研发、设备更新和租赁付款。专用容量有最低采购承诺时，客户使用率下降未必直接减少当期计费；未能按约交付、性能不足和未被合同覆盖的闲置容量仍会影响经济回报。[2：MD&A第31页及风险因素；分析推断]

---

4. 从财务结构归纳商业模式与技术壁垒

前面的报表显示，Cerebras 同时承担硬件设计交付、持续研发和云基础设施运营三类任务。硬件销售通常在转移控制权后取得收入；云容量则需要跨期使用设备、场地和软件。技术差异只有在带来更高售价、更多稳定计费容量或更低全周期交付成本时，才会改善上述财务结果。下面沿这条因果关系讨论晶圆级计算。[1][2；分析推断]

4.1 晶圆级芯片为什么可能更快

大模型生成文本通常分为两步：先处理输入提示，称为 **prefill**；再逐个生成输出 token，称为 **decode**。在低批量、追求单个用户响应速度的 decode 工作负载中，每一步反复读取模型权重，内存和通信常成为瓶颈。把更多运算单元放进系统，只有在数据也能及时送达时才有效。[8：§2]

Cerebras 的 **Wafer-Scale Engine（WSE）**把大量计算核心、片上 SRAM 存储和互连铺在一整片晶圆上，软件再把模型计算映射到这些核心。直觉上，它缩短了计算与存储之间的数据搬运路径。代价是片上存储容量有限、核心间访问距离不均匀，必须处理复杂的模型分片、路由、散热、供电和容错。[8：§§3–6][10]

截至本次研究，公司已发布 CS-4：Nexus 机架平台包含三颗 WSE-3 Turbo，每颗标称 44GB 片上 SRAM。**带宽很高与容量很大是两个不同维度。** 例如，一个 700 亿参数模型若每参数使用 2 字节，仅权重就约 140GB；这是教学计算，尚未包含 KV cache 和运行时空间。大模型需要多晶圆映射，不能据单颗晶圆容量断言 Cerebras 只能运行小模型。厂商跨晶圆设计强调把通信密集的运算尽量保留在晶圆内，跨晶圆传递较少的激活数据。[10][11；教学计算]

4.2 原始论文支持了什么

OSDI 2025 论文 **WaferLLM**由爱丁堡大学与微软研究人员完成，提出 PLMR 硬件模型、MeshGEMM、MeshGEMV 和 KV cache 均衡移动方法。研究说明：晶圆级硬件必须配合针对其存储和互连特点设计的软件；把 GPU 软件的假设直接搬过来，会浪费大量硬件能力。[8：§§3–6]

论文在 WSE-2 与采用 SGLang 的 A100 集群间进行比较。完整端到端模型主要为 Llama 3 8B 与 Llama 2 13B，输入/输出长度和 GPU 数量不同，结果也不同。4K 输入、4K 输出的 Llama 3 8B 测试中，计入 prefill 的每请求吞吐约为 2,459 tokens/s，对照八张 A100 约为 256 tokens/s。较大的 34B、72B 模型有采用部分层实验并按层数放大的结果，不能当作完整模型部署证明。[8：表2及§7]

一个特别有价值的反面结果是：**prefill 和 decode 的能效结论不同。** 在表7的 Llama 3 8B prefill 测试中，八张 A100 相对 WSE-2 的能耗比为 0.34；到表8 decode 测试，该比值为 2.22。因而“低延迟 decode 有优势”有实验证据，“所有阶段都更省电”超出了实验范围。旧代硬件、小模型、单请求速度指标，也不能直接证明 2026 年生产系统相对最新 NVIDIA 系统的单位成本优势。[8：表7–8、§8]

公开代码的当前固定版本已迁移到 WSE-3 / SDK 2.10，论文原实验为 WSE-2 / SDK 1.x。本次检查了论文方法、实验及限制和仓库 README，没有运行硬件复现。对芯片面积和理论带宽的倍数，也没有直接转换成模型速度或利润倍数。[9]

4.3 用户价值如何转化成付费

低延迟适用于交互式编程、实时助手和连续调用模型的 Agent。对这类任务，用户等待时间有价值；但最终应比较同等任务质量下的完成时间、成功率和成本。

一个教学例子：若任务输出 5,000 token，生成速率从 50 提升到 750 tokens/s，纯输出阶段从 100 秒变成约 6.7 秒。如果工具与其他等待另需 60 秒，总时间从 160 秒变为约 66.7 秒，实际加速约 2.4 倍。因而 Cerebras 的商业价值还取决于模型质量、工具等待、并发与价格，而不只是 token/s。[研究计算]

---

5. 市场结构与竞争规模

招股书引用 Bloomberg Intelligence，估计 AI 训练基础设施市场从 2025 年 1,850 亿美元增长至 2029 年 3,800 亿美元，Cerebras 可服务的推理市场从 660 亿增长至 2,920 亿美元。两者合计从 2,510 亿增至 6,720 亿美元。这是**发行人转引的市场预测**；本次没有取得 Bloomberg 底层模型，也没有把它视为已验证收入机会。[1：市场机会章节]

最大的限制来自可服务范围：高吞吐离线推理、实时低延迟服务、模型训练和专用系统，具有不同的购买标准。客户排他条款、模型覆盖、机房供给和资本也会缩小可争取市场。把 Cerebras 收入除以整个 AI 资本开支，无法得到严格意义上的市场份额，因为设备供应商和云运营商处于不同交易层，部分支出会在产业链中重复出现。[1][2；分析推断]

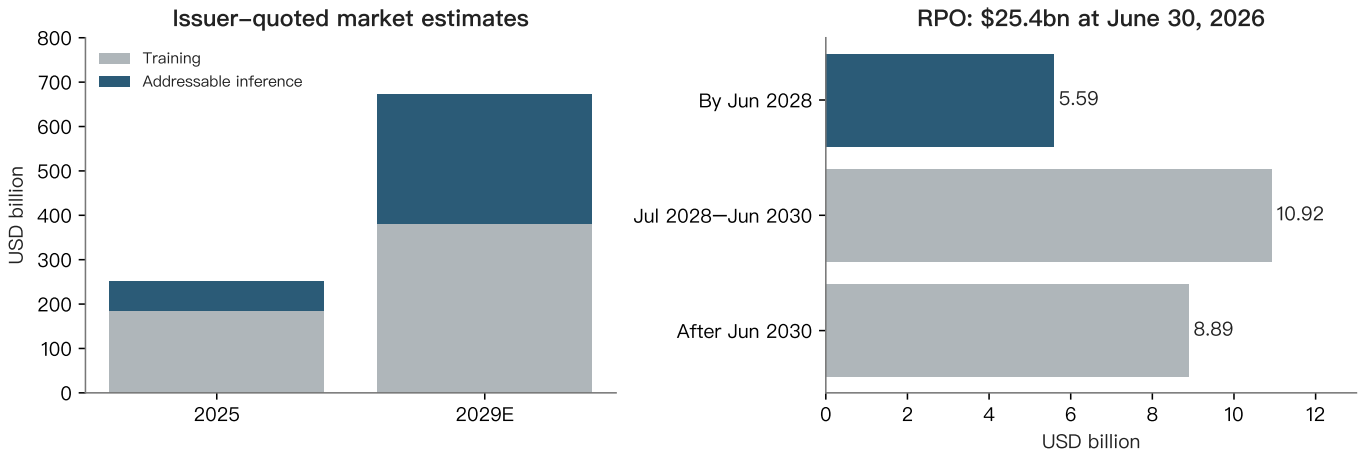


图2：左侧为招股书转引的市场预测，右侧为截至2026-06-30的RPO确认时间估计，单位均为十亿美元，但经济含义不同。RPO未来两年预计确认约22%，再两年约43%，其余约35%。来源：[1][2：附注3]。预测时间不是交付保证。

公司 / 替代方案	实际规模指标	优势与边界
Cerebras	2026Q2收入1.80亿美元；云及其他1.26亿美元	专用晶圆架构与低延迟；客户集中，扩张中的资本效率未定
NVIDIA	FY2027Q2收入962.21亿美元，数据中心约890亿美元	GPU、网络、软件、渠道及资金规模；Groq 3 LPX使其也参与专用低延迟decode竞争
AMD	2026Q2收入115.36亿美元，数据中心约67.18亿美元	GPU/CPU生态和规模；同时与Cerebras合作，竞争与互补并存
CoreWeave	2026Q2收入25.75亿美元；已运行功率约1.5GW	大规模GPU云交付；重资产、债务及利息负担显著
超大云厂商自研芯片、客户自建	本次不构造未披露的可比利润	掌握流量、采购及既有设施，可以选择自建、GPU或专用芯片

NVIDIA季度截至7月26日，AMD截至6月27日，其余截至6月30日；上表为规模比较而非严格同周期业绩排名。功率也不是跨架构等效算力。来源：[2][14][23][24][25]。

竞争正在转向各阶段分别优化。 AMD 与 Cerebras 计划由 AMD Helios 处理输入及高吞吐 prefill，由 WSE 处理低延迟 decode。NVIDIA 8月宣布 Groq 3 LPX 量产，也把专用 SRAM 路径带入竞争。Cerebras 的技术差异仍在，但“GPU阵营始终缺乏针对低延迟场景的回应”已不足以构成投资论据。[12][14；分析推断]

---

6. 价值链、议价权与OpenAI关系

6.1 谁提供稀缺资源，谁承担成本

价值链环节	稀缺资源与议价位置	Cerebras的收益与约束
晶圆制造、系统组装	TSMC晶圆制造；供应链与系统制造能力	自有架构形成差异化，但依赖外部制造。公司称已锁定晶圆供给，不等于所有组件和交付风险消失
数据中心、电力与冷却	场地、并网、容量与长期租约	交付速度成为瓶颈；硬件和场地可能先于收入投入
Cerebras软件与系统	模型映射、编译、供电冷却、稳定运行	是复制壁垒的重要部分；模型变化也可能要求持续开发
OpenAI等大客户	模型、流量、资本、长期采购能力	提供订单和融资，同时获取权证、合同保护和部分排他安排
云与开发者渠道	分发、采购入口、企业信任	扩大可达客户，也需要让渡部分经济利益和客户关系

表中强弱为基于依赖关系的分析判断，不是量化测量。来源：[1][2][3][10][12][13][26]。

6.2 OpenAI的三重角色

关系	已披露安排	应如何理解
客户	750MW承诺容量；服务期3或4年，可延长至总计5年；另有1.25GW选择权	承诺、选项和最终履约收入分别跟踪
融资方	2026年1月提供约10亿美元有担保贷款，名义利率6%	用途受限；可用现金、服务及其他约定抵扣偿还
取得股份的股东	7月行权约1,003万股Class N	相当于8月5日股本约4.22%，但无投票权

来源：[2：附注3、10、12、17]。公司2026年6月公告称OpenAI多年商业协议金额超过200亿美元；截至6月底公司整体RPO为**254亿美元**，其中很大部分来自OpenAI，也包含其他客户安排。[5][2]

OpenAI 在第二季度贡献约 5,680 万美元收入，上半年合计 7,440 万美元，分别约占公司对应收入的 31.5% 和 19.9%；金额已扣除相关客户权证摊销。这使商业合作超出了公告阶段。8月发布的 GPT-5.6 Sol Ultrafast 也提供了后续产品证据，但最高输出速度属于特定产品的公司描述。[2][7]

贷款不能同时被理解为无负担现金和新增销售回款。2026年6月底本金余额约9.18亿美元，上半年已通过服务及其他非现金抵扣偿还8,633万美元，尚未现金偿还。合同在特定违约或终止情形下可以触发提前到期；截至季末，公司披露没有违约。贷款及服务安排还产生重大融资成分：上半年确认约3,860万美元非现金利息费用并增加递延收入，以后随服务履约确认收入。[2：附注10]

OpenAI 的权证最多涉及约3,344.5万股，行权价接近零，归属与容量、市场价值或付款等条件相关。已行权部分与尚有条件的剩余部分必须分开。若全部OpenAI权证最终行使，且忽略其他一切股本变化，其股份约占调整后股本12.8%；这只是稀释敏感性，不代表当前持股。[2；研究计算]

此外，10-Q明确披露，包括OpenAI MRA在内的部分协议限制公司向客户指定的竞争对手提供某些产品和服务。AWS安排也有定价、排他和最低制造容量等保护条款。这支持一个重要判断：战略合作可换取增长，但也可能限制客户多元化与未来议价空间。具体被限制名单、全部价格与性能条款并未完整公开。[2：风险因素]

6.3 客户集中没有消失

2024年G42占收入85%；2025年MBZUAI占62%、G42占24%，合计86%。2026Q2附注中，前三个达到披露门槛的匿名客户分别占34%、32%、10%。公司并未把全部匿名代号与客户名称逐项对应，本报告保留匿名口径；OpenAI贡献另用明确披露金额计算。[1][2：客户集中]

客户构成变化改善了对单一历史客户的依赖，却没有证明客户议价权已经分散。公开资料还不足以验证广泛客户的付费留存、毛利和持续扩容。AWS在2026上半年尚未确认相关商业协议收入，财报称相关租赁开始条件尚未满足；销售渠道已存在，与新合作收入开始确认属于不同事实。[2]

6.4 团队、融资与治理

CEO Andrew Feldman及CTO Sean Lie有SeaMicro和AMD的相关经历，CFO为Robert Komin；2025年底公司披露员工708人。硬件、软件与系统一体化经验对公司有价值，而几百人的组织扩展到大型云基础设施交付，也提高了财务、采购与运维管理要求。[1：管理层、人力资源][2]

资本筹措经历了2025年9月11亿美元G轮、81亿美元投后估值，2026年初约10亿美元H轮、230亿美元投后估值，以及5月IPO。H轮公告列有AMD等参与者。这些交易支持扩张能力，但私人融资估值的优先权、摊薄口径和时间，与上市普通股市值不同。[15][16]

治理需要单独跟踪。B类股每股20票，公众A类股每股1票，N类无投票权。管理层披露截至6月底披露控制与程序在合理保证水平上无效，重大缺陷涉及IT控制、职责分离、人员及复杂交易处理等，整改尚在进行。[2：附注12、Item 4]

招股书还披露，Feldman在此前Riverstone相关案件中，2007年就规避发行人会计控制的一项罪名认罪，2008年另与SEC达成民事和解。该历史事项的性质应按原文限定；当期研究应继续直接核对收入确认、客户权证与内控整改，而不从历史事项推断本期财报存在错报。[1：管理层“法律程序”]

---

7. 财务历史与盈利质量

7.1 四个完整年度

公司可获得的公开经审计年度历史为2022—2025年。本报告未补造第五年数据。2022—2023年来自2024年S-1，2024—2025年来自2026年最终招股书；早期“services and other”与后期“cloud and other services”的业务内容并不完全相同。[1][4]

百万美元，除比率	2022	2023	2024	2025
收入	24.6	78.7	290.3	510.0
毛利润	2.9	26.4	122.7	199.1
毛利率	11.7%	33.5%	42.3%	39.0%
研发费用	155.4	140.1	158.2	243.3
营业利润	−178.8	−133.9	−101.4	−145.9
公司净利润	−177.7	−127.2	−481.6	+237.8
归属普通股的净利润	−177.7	−127.2	−484.8	+87.9
经营现金流	−164.4	−79.0	+452.0	−10.1
现金资本开支	10.5	6.6	23.4	382.7
FCF＝经营现金流−现金资本开支	−174.9	−85.6	+428.5	−392.8

来源：[1：经审计财务报表][4：F-4、F-7]。负号为亏损或现金流出；FCF是本文统一的现金口径，不代表公司的全部非GAAP定义。

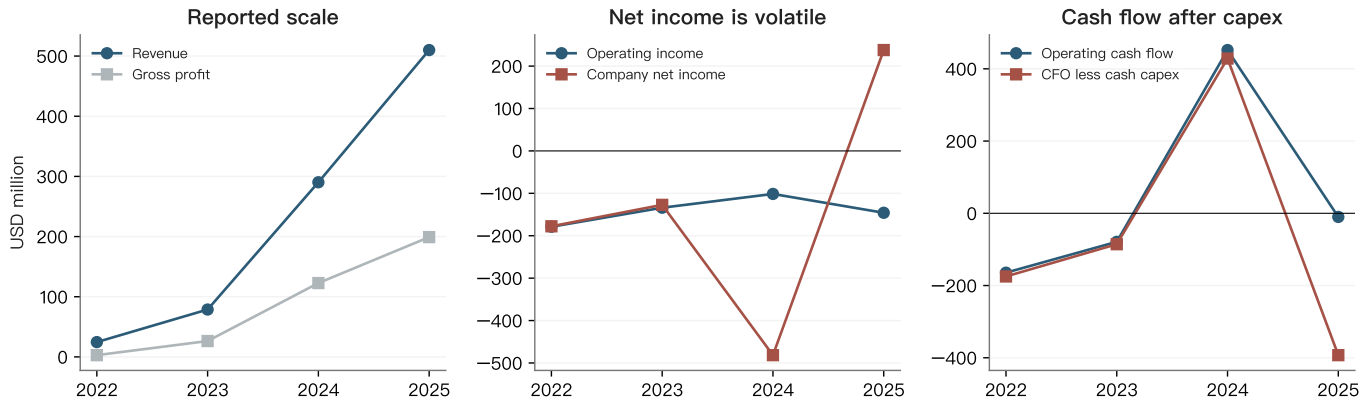


图3：单位百万美元。2024经营现金流受到客户预付款显著影响；2025净利润受到一次性非现金收益影响。图中净利润为公司总净利润，未扣除对可参与分配证券的利润分配。来源：[1][4]。

收入从很小的基数迅速增长，但2025年毛利率从42.3%降至39.0%，营业亏损从1.01亿扩大到1.46亿美元。发行人将毛利率下降主要归因于推理云容量相关数据中心成本增加。硬件的研发费用加上云基础设施投入，构成了两层成本压力；仅用收入增速评价经营杠杆并不充分。[1：MD&A]

7.2 2025年的2.38亿和8,788万，为什么都正确

2025年公司净利润为**237.827百万美元**，其中**149.952百万美元**分配给可参与收益分配的证券，归属普通股的利润为**87.875百万美元**。这是两种利润归属口径，不能把两个数字视为相互矛盾，也不能把差额直接视为现金分红。[1：损益表、每股收益附注]

更重要的是，公司当年确认了**363.336百万美元的远期合同负债终止收益**，属于非现金、非主营收益。2024年相应远期合同公允价值变动又带来大额费用，使前后两年净利润剧烈波动。2025年营业利润仍为负；公司在招股书中的非GAAP净亏损为75.742百万美元，较2024年的21.774百万美元扩大。因此，2025年GAAP净利润不能作为稳定盈利基准。[1]

7.3 Q2收入和利润应分三层阅读

第一层看**法定结果**：收入180.110百万美元，同比增长74.3%，环比下降6.9%；毛利率14.2%；营业亏损477.233百万美元。硬件收入环比波动较大，云收入仍增长，二者同时发生。[1][2]

第二层看**公司Core对账**：Core收入等于GAAP收入，减代垫/转付收入14.503百万美元，再加回客户权证摊销44.262百万美元，得到209.869百万美元。Core并非单纯剔除一次性费用，它还改变了收入分母。[3]

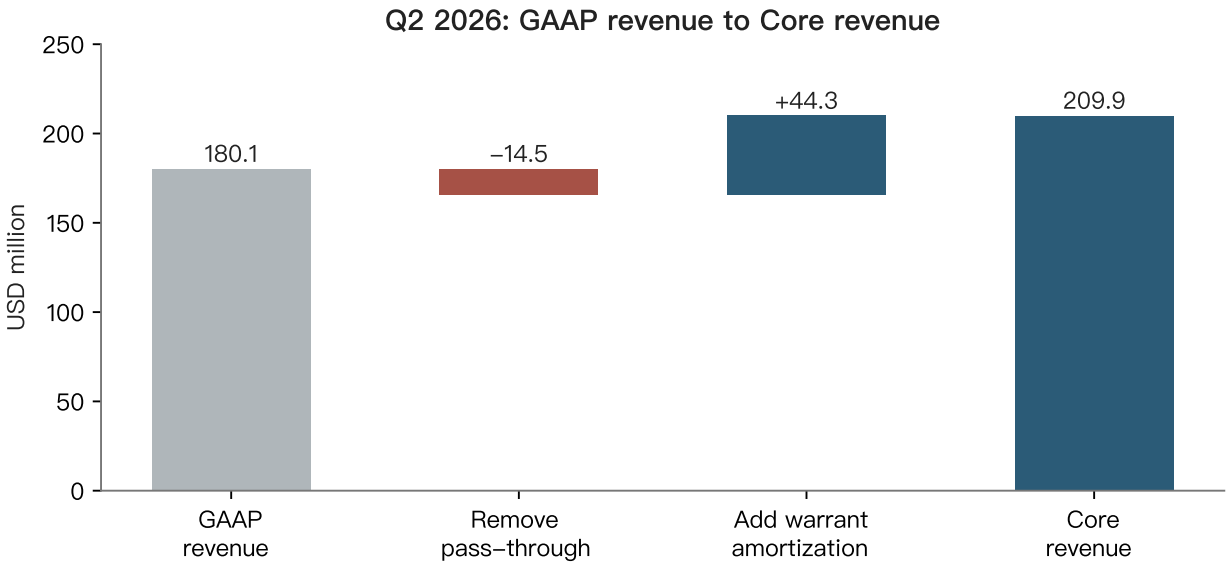


图4：单位百万美元。客户权证摊销在GAAP下抵减收入，Core将其加回。来源：[3：Revenue reconciliation]。

2026Q2，百万美元	金额
GAAP营业亏损	−477.233
减：代垫/转付收入	−14.503
加：对应代垫/转付成本	+14.075
加：客户权证摊销	+44.262
加：股权激励费用	+377.008
加：IPO相关雇主工资税	+22.778

2026Q2，百万美元	金额
Core营业亏损	−33.613

第三层看**普通股经济成本**：377.0百万美元股权激励中有IPO触发的集中确认，不能全部视为每季度正常现金支出；但员工股权激励造成稀释，客户权证也是取得商业关系所付出的经济利益。反过来，股价波动及权证会计使GAAP短期波动变大，也不适合机械年化这一季巨亏。[2][3；分析判断]

Core毛利率40.6%与GAAP14.2%之间的差异，还包括成本端股权激励及IPO工资税调整；不能把整段差额都归因于机房利用率。第二季度OpenAI自身权证摊销约2.5百万美元，公司全部客户权证摊销44.3百万美元，不能全部归到OpenAI。[2][3]

另一个容易混用的指标是**Adjusted EBITDA**。公司第二季度该指标为−53.070百万美元，从GAAP营业亏损加回折旧摊销、股权激励和IPO工资税得到，仍保留客户权证对收入的影响。因此它并不等于“Core营业利润＋折旧”。Core净亏损只有6.908百万美元，也不表示主营业务已接近盈亏平衡，因为仍有正的营业外净收益。[3]

8. IPO、股价与市值历史

2026年5月IPO定价每股185美元，发行3,000万股，承销商全额行使450万股超额配售后，募集总额约63.83亿美元。5月14日首个交易日开盘350美元、收盘311.07美元。到9月11日收盘191.93美元，相对发行价上涨约3.7%，相对首日收盘下跌约38.3%。购买时点决定了非常不同的回报经历。[2][17][19][28；研究计算]

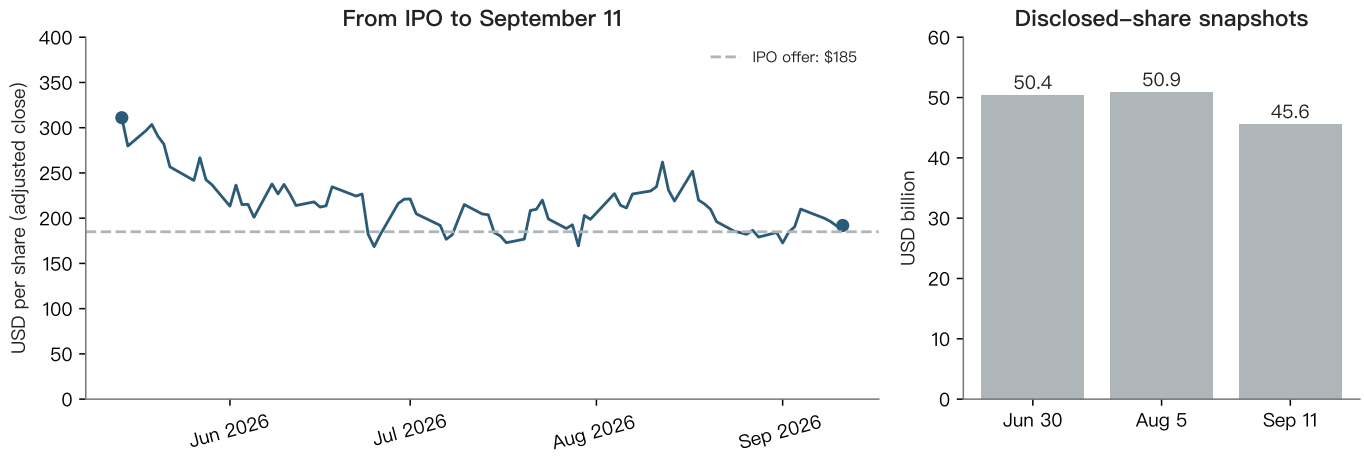


图5：左侧为复权收盘价，美元/股；本次取得期间复权值与收盘价一致。5月15日至9月11日来自[19]，5月14日补充自[28]。右侧市值为[2]披露股本与相应日价格的计算，不连成每日市值曲线。

价格日期	股价	所用全部股本（百万股）	股本日期	计算市值（十亿美元）
2026-06-30	221.00	228.204	2026-06-30	50.43
2026-08-05	214.30	237.564	2026-08-05	50.91
2026-09-11	191.93	237.564	最新可得：2026-08-05	45.60

这是依据后来披露股本重建的历史快照，不代表每个交易日当时市场已掌握全部信息。2026年8月5日之后可能发生的未披露股本变化未纳入；也没有用最新股本回填上市以来每日市值。公司上市不足四个月，无法形成有意义的五年股价历史或长期估值分位数。

9. 估值与同业比较

9.1 先统一分子和分母

企业价值 **EV（Enterprise Value）**在这里以权益市值加债务、再减未受限现金及短期投资计算。为与包含大量租赁的云基础设施同业接近，主比较口径再加入已确认的经营租赁负债；未开始租赁的承诺另列。不同数据供应商对EV仍可能存在处理差异。

Cerebras估值桥	十亿美元
$191.93 × 237.564百万股	45.596
加：贷款本金	+0.918
减：现金和短期投资	−7.922
EV，不含经营租赁	**38.592**
加：已确认经营租赁负债	+0.569
EV，含经营租赁	**39.161**

股本截至8月5日、资产负债表截至6月30日、价格截至9月11日，均为各自最新已取得信息。[2][19][27；研究计算]

TTM收入由2025全年509.991百万美元，加2026H1的373.516，减2025H1的202.834，得到**680.673百万美元**。据此，含租赁EV/收入为57.5倍，不含租赁为56.7倍，P/S为67.0倍。2026全年Core收入指引880—890百万美元，中值对应44.3倍EV/Core收入；它使用未来非GAAP分母，与TTM GAAP倍数分别呈现。[1][2][3；研究计算]

2025盈利受一次性项目影响，2026亏损又受上市股权激励集中确认影响，**P/E不适合作为当前核心估值工具**。EV/收入虽方便，也不能回答资本开支、权证稀释和未来利润率能否支撑估值；这些问题必须通过经营情景补充。

9.2 可比公司：同一天价格，不强行抹平业务差异

2026-09-11收盘口径	Cerebras	NVIDIA	AMD	CoreWeave
股价，美元	191.93	218.29	516.13	88.99
权益市值，十亿美元	45.60	约5,270	842.57	49.08
EV，十亿美元	39.16	约5,250	833.73	95.11
TTM收入，十亿美元	0.681	302.97	41.31	7.59
EV/TTM收入	**57.5x**	**17.3x**	**20.2x**	**12.5x**

来源：[2][19][20][21][22]。Cerebras由原始财报计算；其余为数据供应商汇总且有四舍五入，并以最新季度官方业绩核对规模。[23][24][25]。各公司TTM截止日随最新季报不同，不能视为完全同步的财务期间。

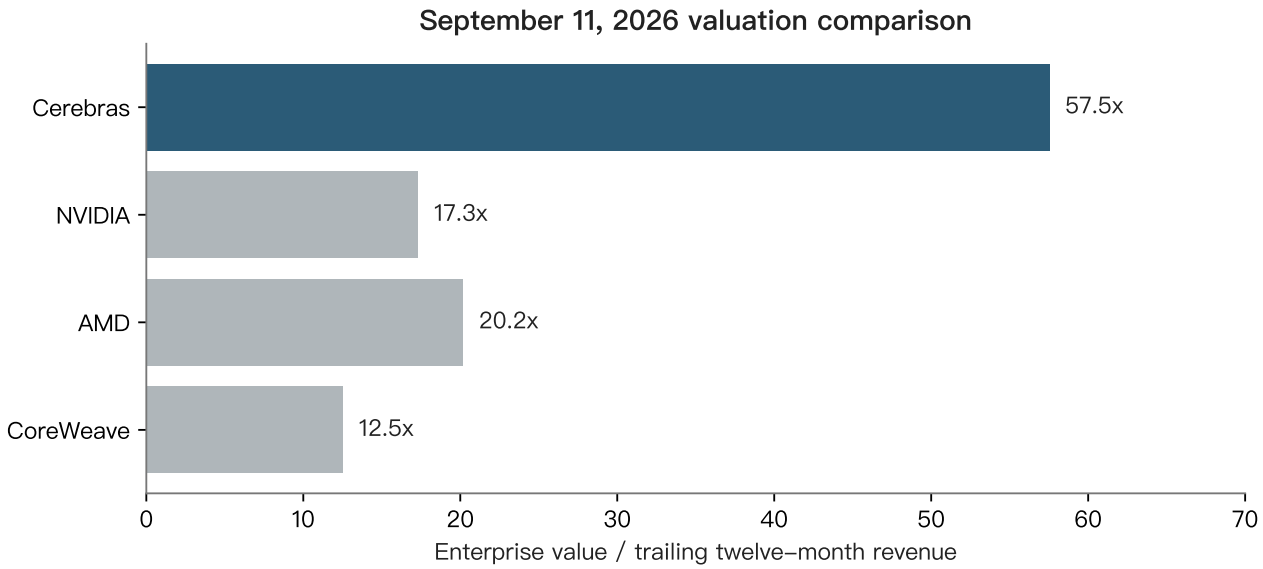


图6：同日市场数据、最新可得TTM。Cerebras EV加入已确认经营租赁，CoreWeave供应商债务口径包含租赁；归类差异和舍入仍存在。来源同上。

NVIDIA最新季度GAAP毛利率约75%，已经能够以极大规模产生经营利润；AMD最新季度GAAP毛利率约54%。Cerebras的Core毛利率约41%仍低于这两者，且Core排除了部分经济成本。因此，当前更高收入倍数需要更快的未来增长来支持，而不是由现有盈利能力直接解释。[2][3][23][24；分析判断]

CoreWeave提供另一种参照：其第二季度收入25.75亿美元，营业亏损0.49亿美元，但净利息费用6.40亿美元。市场评价这类公司的关键包括设备融资、租赁、折旧和利用率。其部分折旧和基础设施成本归入技术基础设施费用，不宜直接将数据网站所列毛利率与Cerebras毛利率比较。[25]

私人公司方面，本次使用Cerebras自身G/H轮作为资本市场定价演化的参照，未把不同优先权、不同收入可见度的私人芯片企业估值硬拼成精确倍数表。缺乏可核验收入和资本结构时，这种精确性会误导。

---

10. 估值要求怎样的经营结果

10.1 2028年条件化情景

以下是**期末收入倍数法的敏感性分析**，不构成盈利预测或目标价。云服务收入被拆为“全年平均实际计费的等效MW × 每MW年净收入”；前者同时包含交付和利用率，后者包含价格、组合与折扣。另加硬件及其他收入。公开合同不足以验证每MW真实价格，因此这些输入全部是分析假设。

2028年假设 / 计算	Bear	Base	Bull
平均实际计费等效容量，MW	300	550	800
每MW年净收入，百万美元	4.0	6.0	7.5
硬件及其他收入，百万美元	500	700	1,000
总收入，十亿美元	**1.7**	**4.0**	**7.0**
假设GAAP毛利率	35%	45%	50%
假设营业利润率	–10%	10%	20%
期末EV/收入假设	8x	12x	16x

2028年假设 / 计算	Bear	Base	Bull
期末净现金，十亿美元	2	3	4
期末全部股本，百万股	270	285	305
期末每股权益价值，美元	57.8	178.9	380.3
按12%、2.30年折现的等价值，美元	44.5	137.9	293.1

公式：收入＝平均计费MW×每MW年净收入＋其他收入；期末权益＝收入×EV/收入倍数＋净现金；每股价值＝权益÷届时股本。假设的未来净现金已概括期间经营及资本开支影响，没有再另加中间自由现金流；也没有假设分红。选取8/12/16倍用于测试增长溢价收缩程度，不是认定未来同业应值这些倍数。

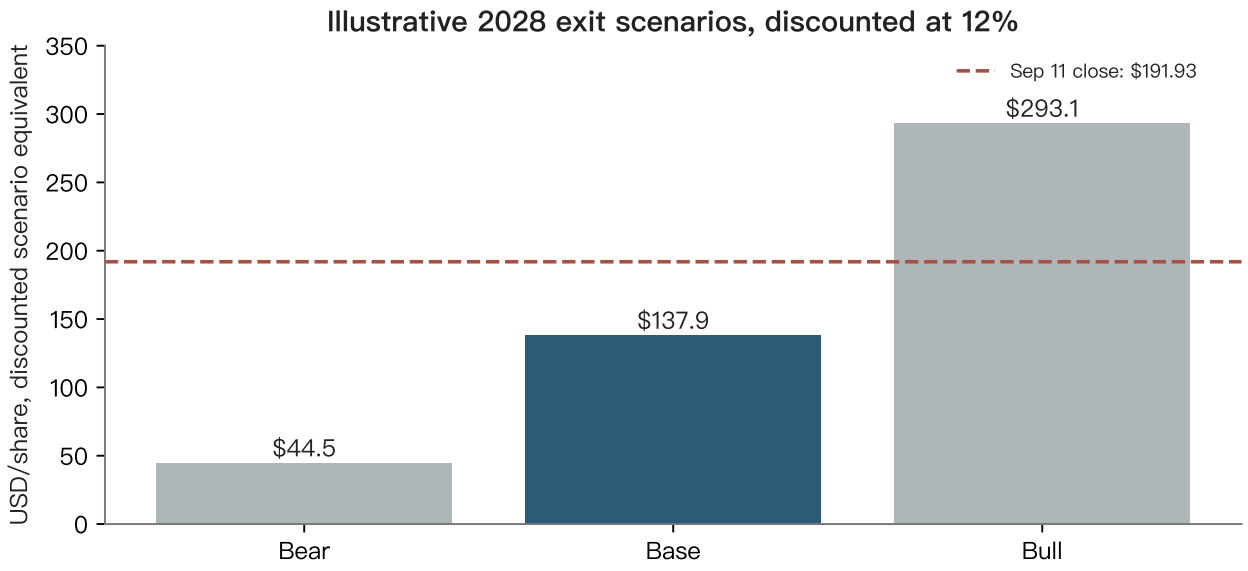


图7：全部为研究假设和计算；虚线为2026-09-11收盘价。参数见上表与Excel。图示是模型敏感性，不包含情景概率。

Bear体现交付或利用率不足、议价压力和资金消耗；Base要求数十亿美元收入与盈利改善；Bull还要求向OpenAI承诺容量之外扩张，并维持较高价格和利润。三者不同的未来股本覆盖部分员工和客户权益稀释，但没有声称已精确预测所有权证及资本募集。

10.2 当前价格的反推门槛

固定Base中的2028年股本2.85亿股、净现金30亿美元、12倍EV/收入与12%折现率，若要支持今天191.93美元的价格，所需2028年收入约为**56.7亿美元**。这个门槛约为2026年Core收入指引中值的**6.4倍**；两者分别是假设GAAP等价收入和公司Core指引，存在口径差异。

这一计算说明，单纯"收入将快速增长"还不够。投资结果同时取决于增长幅度、交付时间、利润率、最终估值倍数与稀释。假如收入扩张但资本开支或股权代价更大，企业可以成功，普通股回报仍可能有限。[研究推断]

管理层在Q2业绩公告中表示计划2027年收入超过三倍增长，但未提供足够公开细节让外部研究者重建这一计划。它应作为需检验的经营目标，而不是确定现金流或订单兑现率。[3]

11. 近期事件、催化剂与风险

日期	可核验事件	下一步观察
2026-05	IPO，随后完成超额配售	募集资金转为运营能力的速度与回报
2026-07-23	AMD与Cerebras宣布分离prefill/decode方案	联合系统实际投产、吞吐和每任务成本
2026-08-12	Q2财报；全年Core收入指引升至8.80—8.90亿美元	GAAP与Core差额、毛利及产能兑现
2026-08-13	GPT-5.6 Sol Ultrafast预览	广泛可用性、持续付费、服务稳定性
2026-08-18/25	CS-4发布及Hot Chips技术披露	批量交付、跨模型适用性、实际能效
2026-08-24	NVIDIA宣布Groq 3 LPX量产	低延迟溢价与竞争架构的可替代程度
2026-09-01	芬兰165MW项目公告	分期建设、并网和验收；首期50MW仍在建设

来源：[2][3][7][10][11][12][14][17][18]。芬兰公告不能自动与此前“已运行及已签约超过600MW”相加，二者是否重叠尚不明确；600MW也不是全部已运行容量。[3][18]

公司预计AMD联合方案于2026Q4投入生产，并计划于2027Q1将相关能力带到Amazon Bedrock。两者是后续催化剂；原有Marketplace渠道与尚未交付的新Bedrock集成应分开跟踪。[3]

市场价格显示，Q1财报后的6月24日收盘下跌19.6%，Q2公告后的8月13日下跌11.9%。这只能说明相应交易日的市场反应，无法仅凭涨跌识别投资者究竟在交易估值、指引、持仓、融资还是宏观因素。[19]

主要风险相互关联：工厂或电力延期使设备闲置，推迟履约；客户集中和排他安排限制需求分散；GPU及其他专用架构压低低延迟溢价；模型变化增加软件适配支出；大额租约和资本开支削弱净现金；员工与客户权益安排稀释股东。对G42等国际客户的业务，还受出口许可和跨境监管影响，具体适用以届时合同及法规为准。[1][2；机制分析]

12. 应持续跟踪的指标与证据缺口

问题	当前基准	判断改善所需证据	反证
RPO能否变成收入	254亿美元，约22%预计未来24个月确认	分期上线、验收、对应收入和回款	持续推迟；新增承诺快于实际履约
扩张能否创造利润	Q2 Core毛利率40.6%；Core营业亏损3,361万美元	收入增长同时提升含折旧利润、缩小GAAP/Core差额	毛利受价格与闲置成本挤压
对客户的让利是否可控	H1客户权证摊销4,632万美元	新合同的现金、股权、折扣总成本	靠增加客户股权激励维持增长
资本是否有效使用	H1现金capex5.49亿美元；FCF−5.96亿美元	实际投产容量、利用率、单位容量现金回收	设备/租赁承诺上升，收入兑现落后
客户能否多元化	历史G42/MBZUAI集中，OpenAI快速增长	非关联客户付费扩容及利润贡献	增长转向另一个单一大客户
速度优势能否保持	公开论文与特定产品证据	同模型、精度、上下文、并发和P95服务目标的独立TCO测试	最新替代方案取得更好的质量/成本组合
治理是否跟上规模	6月底控制缺陷仍未解决	财报披露完成整改、后续审计验证	缺陷反复或出现需重述的重大会计问题

Q3公司指引为Core收入2.14—2.16亿美元、毛利率38%—40%、营业利润率−25%至−23%；全年Core营业利润率仍预计−19%至−17%。按全年与Q3收入中值，隐含Q4 Core收入约2.69亿美元，因此后半年的收入爬坡与投入节奏都值得检验。[3；研究计算]

尚未取得的关键证据包括：OpenAI及AWS全部价格、取消和性能救济条款；实际逐季已上线MW及利用率；分客户贡献利润；全口径每成功任务成本；客户留存；未来所有权益工具的最终稀释；以及8月5日后精确的全部股本。RPO比一般市场宣传更有信息量，但依旧受履约条件和确认时间约束。

研究的条件化判断是：**若Cerebras能把已承诺容量按期变成高利用率服务，扩大独立客户收入，并在计入设备、场地和权益代价后改善现金回报，它就具备成为重要AI基础设施供应商的基础。以当前估值，证明技术有效已经不足以消除投资风险，下一阶段需要证明规模化交付的经济回报。**

---

13. 方法、原始财报与来源

本报告优先使用SEC文件、发行人业绩对账、合作方官方公告和原始论文。监管文件中的历史财务与合同摘要不等同于对未来的保证；技术性能、市场规模和产能计划按公司或作者原有范围呈现。未将另一份AI摘要作为证据。

原始财报已下载并登记哈希：2024年S-1补充2022—2023年历史，2026年最终424B4提供2024—2025年经审计财务，2026Q2 10-Q提供最新法定财务、合同及期后事项，Q2 8-K附件提供非GAAP对账与指引。独立研究论文采用USENIX正式发表版，避免混用预印本不同版本的数字。

所有图表的数据位于工作簿和data.json。主要表格保留百万美元口径；正文亿美元转换及百分比按显示精度四舍五入。价格与股本日期不同已明确标注；情景未赋予概率。报告不提供无法复核的私募cap table、长期估值百分位或同口径全产业市场份额。

下列来源统一访问于2026-09-12；财务数字的报表期、新闻发布日期与股价日期分别标明。

1. **Cerebras final IPO prospectus (424B4)** — Cerebras / SEC，2026-05-14。定位：MD&A pp. 89–110; business pp. 112–148; management pp. 149–152; audited statements F-1 onward。`src-20260514-cerebras-prospectus`；本地PDF。
2. **Cerebras Q2 2026 Form 10-Q** — Cerebras / SEC，2026-08-12。定位：Cover; financial statements pp. 6–12; Notes 3, 9–17; MD&A Item 4 controls; risk factors。`src-20260812-cerebras-10q`；本地PDF。
3. **Cerebras Q2 2026 earnings release and reconciliations (8-K Exhibit 99.1)** — Cerebras / SEC，2026-08-12。定位：Business highlights; Q3/FY guidance; GAAP-to-Core and EBITDA reconciliations。`src-20260812-cerebras-q2-results`；本地PDF。
4. **Cerebras original 2024 Form S-1** — Cerebras / SEC，2024-09-30。定位：2022/2023 audited statements F-4, F-7; historical document, not current offering terms。`src-20240930-cerebras-s1`；本地PDF。
5. **Cerebras Q1 2026 results** — Cerebras / SEC，2026-06-23。定位：OpenAI >$20 billion deal; Q1 results and prior guidance。`src-20260623-cerebras-q1`；本地指针。
6. **OpenAI partners with Cerebras** — OpenAI，2026-01-14。定位：750 MW staged deployment。`src-20260114-openai-cerebras`；本地指针。
7. **GPT-5.6 Sol Ultrafast** — OpenAI，2026-08-13。定位：Product preview; per-token speed, not whole-task speed。`src-20260813-openai-ultrafast`；本地指针。
8. **WaferLLM: Large Language Model Inference at Wafer Scale** — Congjie He et al. / USENIX OSDI，2025-07-07。定位：OSDI pp. 257–273; §§3–8, Table 2, Tables 7–8. Read methods/results/limits; no hardware reproduction.。`src-20250707-waferllm-osdi`；本地PDF。

9. **WaferLLM official code README at fixed commit** — MeshInfra / paper authors，2026-06-12。定位：README prerequisites: current WSE-3 / SDK 2.10 differs from paper WSE-2 / SDK 1.x; README read only。src-20260612-waferllm-code；本地指针。
10. **Ultrafast Frontier Inference: Cerebras at Hot Chips 2026** — Cerebras，2026-08-25。定位：Multi-wafer mapping, interconnect, CS-4; future roadmaps are company claims。src-20260825-cerebras-hotchips；本地指针。
11. **Cerebras CS-4 product specifications** — Cerebras，2026-08-18。定位：Three WSE-3 Turbo per rack; 44 GB SRAM per wafer; specifications are company-reported。src-20260818-cerebras-cs4；本地指针。
12. **AMD and Cerebras disaggregated inference partnership** — AMD / Cerebras，2026-07-23。定位：AMD prefill / Cerebras decode; availability and efficiency are forward-looking。src-20260723-amd-cerebras；本地指针。
13. **Cerebras is coming to AWS** — Cerebras，2026-03-13。定位：Commercial distribution and complementary prefill/decode; distinguish Marketplace from future Bedrock service。src-20260313-cerebras-aws；本地指针。
14. **NVIDIA Groq 3 LPX now in full production** — NVIDIA，2026-08-24。定位：Competing SRAM decode architecture; speed claims model/configuration-specific。src-20260824-nvidia-groq3；本地指针。
15. **Cerebras raises $1.1 billion Series G at $8.1 billion valuation** — Cerebras，2025-09-30。定位：Financing amount, post-money valuation, participants。src-20250930-cerebras-series-g；本地指针。
16. **Cerebras raises $1 billion Series H** — Cerebras，2026-02-03。定位：Body dated February 3; IR header February 4; $23 billion post-money。src-20260203-cerebras-series-h；本地指针。
17. **Cerebras announces pricing of IPO** — Cerebras，2026-05-13。定位：30 million shares at $185; 4.5 million overallotment option。src-20260513-cerebras-ipo-pricing；本地指针。
18. **Cerebras and Compute Nordic Finland announce 165 MW site** — Cerebras / Compute Nordic Finland，2026-09-01。定位：165 MW staged IT capacity; first 50 MW under construction; seven-year service orders。src-20260901-cerebras-finland；本地指针。
19. **CBRS historical prices** — Stock Analysis / S&P Global data，2026-09-11。定位：82 daily OHLCV records 2026-05-15–09-11 extracted; same-day adjusted close; first trading day supplemented with source 28。src-20260912-cbrs-price-history；本地指针。
20. **NVIDIA statistics and market valuation** — Stock Analysis，2026-09-11。定位：September 11 closing price, TTM financial aggregates and vendor EV; rounded。src-20260912-nvda-valuation；本地指针。
21. **AMD statistics and market valuation** — Stock Analysis，2026-09-11。定位：September 11 closing price, TTM financial aggregates and vendor EV; rounded。src-20260912-amd-valuation；本地指针。
22. **CoreWeave statistics and market valuation** — Stock Analysis，2026-09-11。定位：September 11 closing price, TTM financial aggregates and vendor EV; debt includes leases。src-20260912-crwv-valuation；本地指针。
23. **NVIDIA Q2 FY2027 financial results** — NVIDIA，2026-08-26。定位：Quarter ended July 26, 2026; GAAP revenue, margin and operating profit。src-20260826-nvidia-q2fy27；本地指针。
24. **AMD Q2 2026 financial results** — AMD，2026-08-04。定位：Quarter ended June 27, 2026; GAAP revenue, data center revenue, gross margin。src-20260804-amd-q2；本地指针。

25. **CoreWeave Q2 2026 financial results** — CoreWeave，2026-08-11。定位：Quarter ended June 30; revenue, interest, operating loss; infrastructure cost accounting differs。`src-20260811-coreweave-q2`；本地指针。

26. **Cerebras public inference pricing** — Cerebras，undated; accessed 2026-09-12。定位：Rendered developer table: GPT OSS 120B $0.35 input / $0.75 output per million tokens; enterprise quotation separate。`src-20260912-cerebras-pricing`；本地指针。

27. **CBRS historical closing quote cross-check** — MarketMinute / WSIL，2026-09-11。定位：September 11 close $191.93; quote corroboration only。`src-20260912-cbrs-price-crosscheck`；本地指针。

28. **Cerebras IPO: first trading-day prices** — Kiplinger，2026-05-14。定位：First trade $350; intraday high $385; close $311.07; not investment opinion。`src-20260514-cbrs-first-close`；本地指针。