东吴证券 SOOCHOW SECURITIES

传媒行业深度报告

NeoCloud 龙头崛起，AI 算力基础设施价值凸显

2026 年 05 月 31 日

增持（维持）

证券分析师 张良卫
执业证书：S0600516070001
021-60199793
zhanglw@dwzq.com.cn
证券分析师 张家琦
执业证书：S0600521070001
zhangjiaqi@dwzq.com.cn

投资要点

- **AI 算力供需错配催生 NeoCloud，专业化 AI 基础设施价值凸显：**进入 AI 时代后，云计算基础设施的核心矛盾从通用算力供给转向高密度 GPU 集群、电力、数据中心、网络、液冷和工程交付能力的系统性供给。需求侧，大模型训练仍在迭代，多模态模型、Agent 工作流、企业级 AI 应用和推理调用持续放量，推动 AI 算力需求从训练主导逐步演变为训练与推理共同驱动；供给侧，瓶颈已从 GPU 扩散至电力并网、数据中心建设和 AI Factory 交付能力。我们认为，NeoCloud 本质上并非简单 GPU 转租商，而是围绕 AI 训练和推理负载形成的重资产 AI 基础设施运营商，短期受益于供需错配和长约订单，中长期竞争关键在于能否从“卖算力”升级为“算力+软件+平台”的复合服务模式。

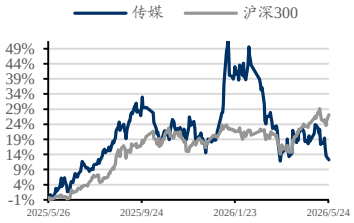
- **商业模式核心在于长约、融资与交付能力循环，平台化决定长期终局：**NeoCloud 商业模式以高强度资本开支换取长期算力合同现金流，收入端依赖大客户长约和 Backlog 提升可见度，成本端则受 GPU 采购、电力、数据中心、网络、液冷、运维和融资成本共同影响。从经营层面看，成熟算力资产在高利用率环境下具备较强 EBITDA 盈利能力，但折旧摊销、利息费用和持续资本开支仍会显著压制净利润和自由现金流。我们认为，NeoCloud 盈利能力可拆分为“单位算力收入-单位算力 TCO-资本成本”，未来胜负不仅取决于 GPU 获取速度，更取决于低成本融资、稳定交付、客户质量和平台化能力。

- **Nebius 平台化空间更优，CoreWeave 短期收入确定性更强：**Nebius 和 CoreWeave 是当前海外 NeoCloud 赛道中最具代表性的两类公司。Nebius 依托 Yandex 工程基因和 AI 原生重构，围绕全球 AI Factory、GW 级电力资源和平台化能力持续扩张，长期有望从底层 GPU 算力供给进一步延伸至推理平台、开发者工具和企业 AI 基础设施服务，成长空间更多来自“算力+平台”升级。CoreWeave 则凭借与 NVIDIA 的深度绑定、大客户长约和资产融资模式实现快速扩张，收入积压较高、合同能见度较强，短期业绩兑现确定性更高。我们认为，两家公司分别代表 NeoCloud 的两条核心路径：Nebius 更强调全球化布局、电力卡位和平台化演进，CoreWeave 更强调 GPU 供给优势、NVIDIA 生态绑定和合约驱动扩张，但二者均需持续关注重资本开支、融资成本、客户集中和产能交付风险。

- **投资建议：**AI 算力供需错配仍是本轮 NeoCloud 行业景气的核心来源，海外建议重点关注 NeoCloud 龙头及算力/存储产业链，国内则建议沿算力基础设施和运营商生态布局。建议关注：1）海外 NeoCloud 龙头：CoreWeave、Nebius 等；2）海外算力/存储产业链：NVIDIA、Broadcom、Marvell、Micron、SK Hynix、TSMC、Arista Networks、Vertiv、Super Micro Computer、Dell Technologies、Lenovo Group 等；3）国内算力基础设施与运营商生态：盛视科技、平治信息、华策影视、协创数据、宏景科技等。

- **风险提示：**AI 算力需求不及预期风险，资本开支过高和融资成本上升风险，产能交付不及预期风险。

行业走势



相关研究

《“十五五”规划定调，看好游戏出海》

2026-03-16

《海外 AI 年度复盘及财报综述：狂欢将尽还是新周期开启？》

2026-01-21

内容目录

图表目录

1. NeoCloud 行业：AI 算力供需错配下的新型基础设施资产运营模式

1.1. 行业定义与定位：AI 时代的专业化算力云

全球云计算行业正在 AI 时代重新分层。过去二十年，云计算竞争主要围绕通用计算、存储、数据库、网络、安全和企业服务展开，AWS、Azure、GCP 等 Hyperscaler 依靠全球数据中心网络、完整云产品矩阵和企业客户基础，形成了全栈通用云的主导格局。进入 AI 时代后，云基础设施的核心矛盾发生变化：客户需求从传统通用算力转向能够支撑大模型训练、推理部署和 AI 应用运行的高密度 GPU 集群。由此，以 AI 算力为核心的新型专业化云厂商——NeoCloud 开始快速崛起。

从产业定位看，NeoCloud 既不是传统 IaaS 厂商的简单延伸，也不是单纯的 GPU 转租商，而是围绕 AI 训练和推理工作负载形成的重要资产 AI 基础设施运营商。其核心商业模式是通过采购、融资、长期供应协议等方式获取 GPU 服务器，并配套数据中心、电力、网络、存储、冷却和运维能力，向大模型公司、云厂商、企业客户和开发者提供可规模化交付的 AI 算力服务。与全栈通用云强调产品矩阵完整度不同，NeoCloud 更关注 GPU 资源获取速度、集群交付能力、单位算力效率、网络与调度优化，以及大规模 AI 工作负载的稳定运行能力。

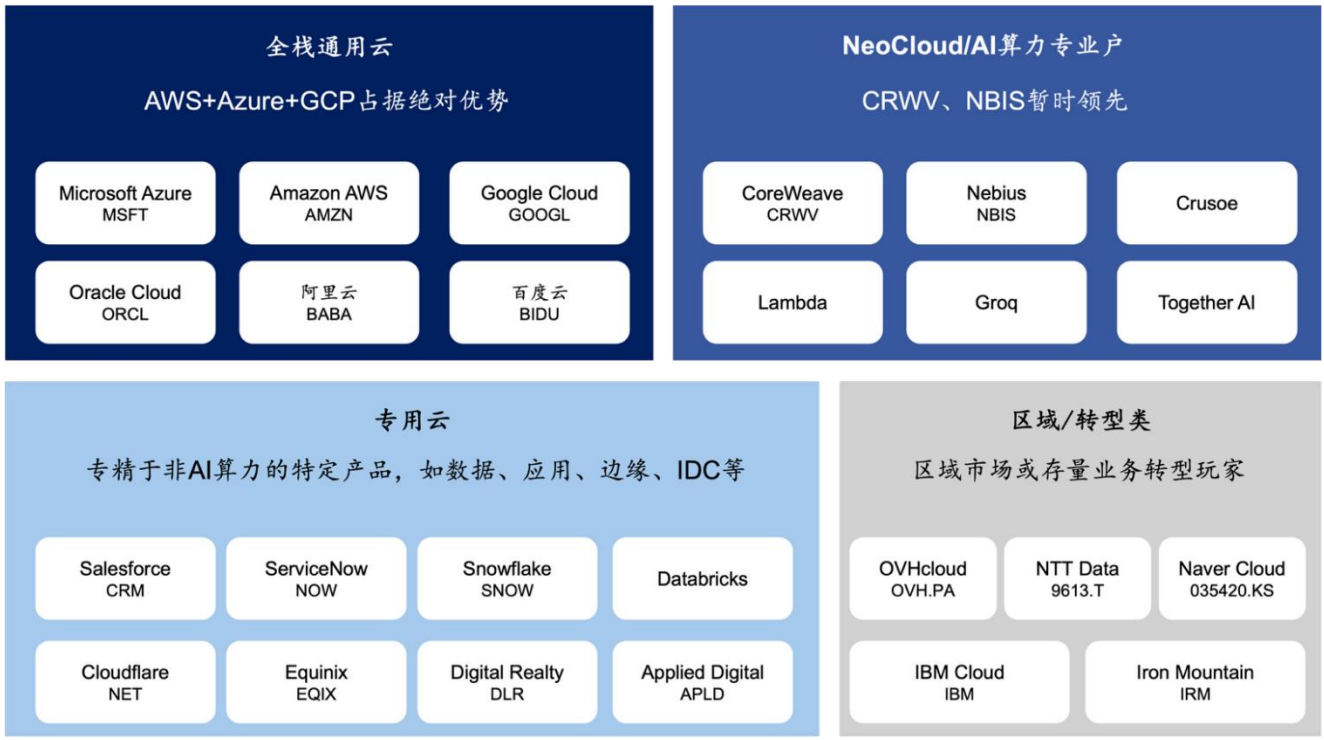
从全球云生态看，当前云厂商大致可以划分为四类。第一类是全栈通用云，包括 AWS、Azure、GCP、阿里云等，覆盖计算、存储、数据库、AI 平台和企业服务等全品类，是 AI 算力最大的建设者、消费者和整合者。第二类是 NeoCloud，包括 CoreWeave、Nebius、Lambda 等，核心聚焦 GPU 算力、AI 训练集群、推理平台和 AI 基础设施服务，是 AI 时代最直接的算力供给层。第三类是垂直云/云软件平台，包括 Snowflake、Databricks、Cloudflare 等，主要围绕数据、应用、边缘网络或特定场景形成专业化能力。第四类是区域云/传统 IT 转型/数据中心基础设施玩家，主要受区域客户、数据主权、合规要求和传统 IT 云化需求驱动。

NeoCloud 与 Hyperscaler 之间也并非简单替代关系，而是阶段性互补与长期竞争并存。传统 Hyperscaler 最初主要服务于通用计算、虚拟化、多租户应用和企业云服务，AI 工作负载早期更多是其云服务体系中的新增品类。但大模型训练和推理对 GPU 密度、低延迟网络、液冷散热、大规模并行调度和集群稳定性提出更高要求，通用云架构叠加 AI 能力，与原生面向 AI 工作负载设计的平台之间开始出现工程效率差异。NeoCloud 正是在这一阶段性窗口中，承接了大模型公司、企业客户乃至 Hyperscaler 自身对于新一代 GPU、大规模集群和可预期 SLA 的迫切需求。

但从长期看，NeoCloud 的出现并不意味着云计算格局将彻底分散。相反，AI 时代可能进一步强化头部全栈云厂商的综合优势。AI 云竞争不仅需要 GPU 和数据中心，还

需要模型、开发工具、数据库、安全能力、企业客户关系、生态合作和全球交付体系。全栈云厂商能够将底层算力与上层平台、应用和企业场景打通，形成更强客户粘性和更完整商业闭环。因此，NeoCloud 未来发展的核心问题，不在于短期能否依靠 GPU 稀缺获取订单，而在于能否从算力资源型公司，进一步升级为具备平台能力、软件能力和客户生态能力的 AI 基础设施平台公司。只有完成这一能力跃迁，NeoCloud 才有机会在 AI 云长期竞争格局中占据独立位置。

图1：云厂商分类



数据来源：各公司官网，东吴证券研究所

1.2. 行业背景：AI 算力需求爆发与物理供给刚性创造 NeoCloud 成长窗口

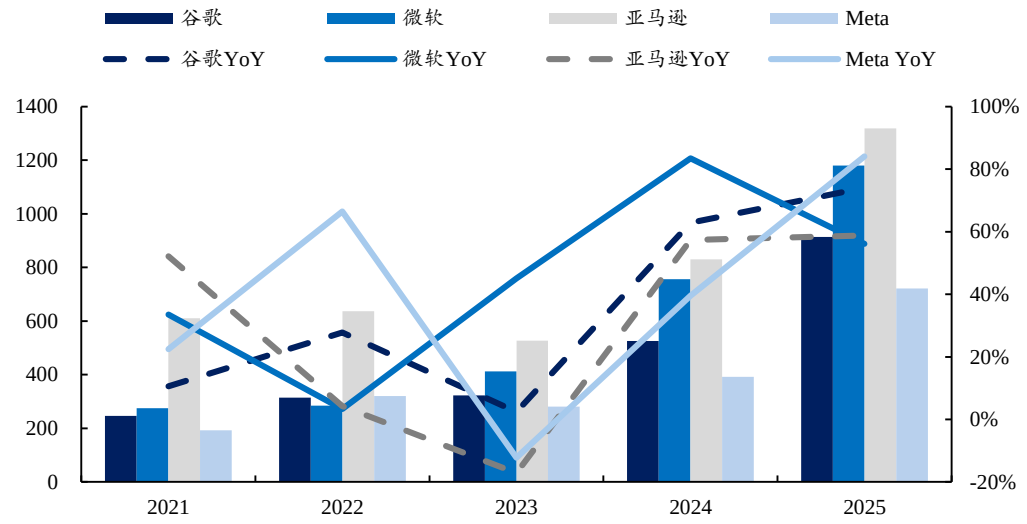
NeoCloud 崛起的核心背景是 AI 算力需求快速扩张与物理产能供给刚性之间的错配。不同于传统云计算周期，本轮 AI 基础设施扩张并不是单纯增加通用服务器、存储和网络资源，而是围绕高端 GPU 集群、高功率密度数据中心、低延迟网络、液冷散热和大规模集群调度能力展开。需求侧可以在数个季度内快速爆发，而 GPU 交付、电力接入、数据中心建设、液冷改造和集群调试均需要较长周期，这为专业化 AI 算力云厂

商创造了阶段性成长窗口。

1.2.1. 需求侧：训练需求延续，推理放量接棒

AI 基础设施建设仍处于高景气阶段，核心体现为头部云厂商资本开支持续上升。2023 年以来，全球主要 CSP 围绕 GPU 集群、数据中心、电力与网络设施持续加大投入，2026 年资本开支指引进一步上行。谷歌在 FY1Q26 财报会上将 2026 年全年资本开支指引由 1750-1850 亿美元上调至 1800-1900 亿美元，并预计 2027 年资本支出将较 2026 年进一步显著增加。微软在 FY3Q26 财报会后预计 2026 年全年资本开支将达到 1900 亿美元，同比增长约 61%。亚马逊在 FY4Q25 财报会上预计 2026 年全年资本开支约 2000 亿美元，同比增长超过 50%，显著高于市场此前预期。Meta 也在 FY1Q26 财报会上将 2026 年全年资本开支指引由 1150-1350 亿美元上调至 1250-1450 亿美元。头部云厂商资本开支的持续上行，表明 AI 算力供给仍未充分满足需求，AI 基础设施仍处于产能快速扩张周期。

图2：四大 CSP 资本开支（亿美元）与增速（%）



数据来源：各公司财报，东吴证券研究所

表1：北美五大 CSP 资本开支指引（2026 年）

公司	2026 年资本开支最新指引	此前指引/市场预期	变化	主要投入方向

亚马逊	约 2000 亿美元	1460-1470 亿美元	显著高于预期	主要用于 AWS、AI、机器人、半导体、卫星等方向；管理层强调 AWS 增长越快，资本开支也会随之提升
谷歌	1800-1900 亿美元	1750-1850 亿美元	上调 50 亿美元	主要由 AI 算力需求、Google Cloud 订单增长、内部及外部 AI 计算需求驱动；公司还预计 2027 年 CapEx 将较 2026 年显著增加
微软	约 1900 亿美元	1470 亿美元	显著高于预期。同比+61%	主要用于 Azure 和 AI 基础设施；微软 FY26 Q3 资本开支为 319 亿美元，约三分之二用于 GPU、CPU 等短寿命资产；公司表示客户需求仍超过供给
Meta	1250-1450 亿美元	1150-1350 亿美元	上调 100 亿美元	主要由于组件价格上涨，以及为未来产能准备的额外数据中心成本
甲骨文	500 亿美元	500 亿美元	不变	主要用于扩建 OCI 云基础设施和 AI 算力产能

数据来源：证券时报，华尔街见闻，新浪财经，科创板日报，东吴证券研究所

英伟达最新业绩进一步验证 AI 基础设施需求的高景气度。FY1Q27，公司实现收入 816 亿美元，同比增长 85%；其中数据中心收入达到 752 亿美元，同比增长 92%，占总收入比重超过 9 成。公司同时给出下一季度约 910 亿美元收入指引，反映下游 AI 基础设施需求仍在延续。管理层将当前行业阶段定义为大规模 AI Factory 建设周期，并强调 Blackwell 平台正在进入规模放量阶段。对于 NeoCloud 而言，英伟达数据中心收入的高增长不仅意味着 GPU 需求强劲，也意味着 AI 服务器、数据中心、电力、网络和冷却等配套基础设施正在同步扩张，行业景气度并非单一芯片周期，而是贯穿 AI 基础设施全链条的系统性建设周期。

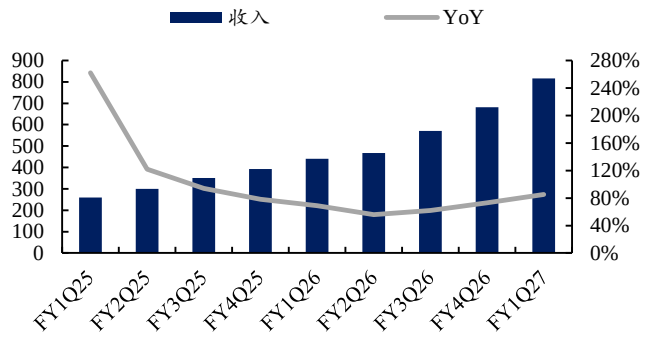
从需求结构来看，AI 算力需求正在由训练主导演逐步演变为训练与推理共同驱动。训练仍然是最直接、最集中、单体规模最大的 GPU 需求来源，大模型参数规模提升、多模态模型迭代以及头部模型厂商持续训练新一代模型，仍将支撑高端 GPU 集群需求。TrendForce 预计，2026 年北美前五大 CSP 通过采购英伟达 GB/VR 系列服务器，将推动其 AI 训练算力同比增长超过 56%。但更重要的变化在于，推理需求正在成为 AI 基础设施新增需求的主要增量。TrendForce 预计，2026 年北美前五大 CSP 的 AI 推理算力将同比增长约 122%，增速明显快于训练算力。与此同时，微软明确表示其 AI 业务快速增长主要来自推理，而非单纯 GPU 训练租赁；AWS Bedrock 在 2026 年第一季度处理的 Token 数量已经超过此前所有年份总和；谷歌也首次将第八代 TPU 拆分为训练芯片 TPU 8t 和

推理芯片 8i。上述变化表明，AI 算力需求正在从模型训练阶段，向应用调用、Agent 工作流、企业级部署和实时推理场景扩散。训练需求决定单次集群规模和短期订单弹性，推理需求则决定算力使用效率、持续性和商业化深度。

这一需求结构变化对 NeoCloud 具有重要意义。训练需求往往以大客户、长约、大规模集群交付为主，有助于 NeoCloud 快速获取订单、提升收入规模，并验证其 GPU 资源获取和 AI Factory 交付能力。推理需求则具有更高频、更分散、更持续的特点，对平台化调度、成本优化、低延迟网络、弹性扩容和多租户服务能力提出更高要求。因此，NeoCloud 短期成长更多受益于训练集群扩张和 GPU 供给稀缺，中长期能否打开更大空间，则取决于其能否承接推理放量带来的持续算力需求，并从项目型训练集群租赁，升级为更稳定的平台型 AI 基础设施服务。

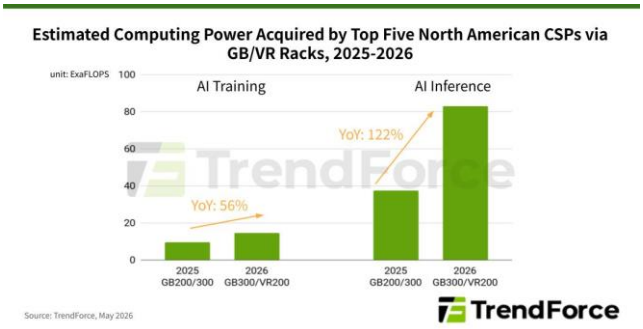
企业 AI 应用落地构成 NeoCloud 中长期成长的另一条主线。大量企业不具备自建大规模 GPU 集群的能力，也未必愿意完全绑定单一 Hyperscaler。随着企业将 AI 嵌入工作流程，模型微调、私有化部署、专属推理服务、行业智能体和企业级 AI 应用将持续释放算力需求。对于这类客户而言，NeoCloud 的价值不只是提供 GPU 资源，更在于提供灵活、可扩容、交付速度快、成本相对可控的 AI 基础设施选择。

图3：英伟达收入（亿美元）与增速（%）



数据来源：公司财报，东吴证券研究所

图4：北美 5 大 CSP 通过 GB/VR 机架获得的算力估计



数据来源：TrendForce，东吴证券研究所

1.2.2. 供给侧：瓶颈从 GPU 扩散到电力、数据中心和 AI Factory 交付能力

AI 算力供给并非买到 GPU 即可形成收入。真正可交付的 AI 产能，需要同时具备 GPU、服务器、数据中心、电力、液冷、网络、存储、调度系统和运维能力。任何单一环节受限，都会影响最终算力交付。因此，AI 算力短缺本质上不是单一 GPU 短缺，而是系统工程短缺。对于 NeoCloud 而言，行业竞争正在由 GPU 获取能力，逐步转向 GPU、电力、数据中心、网络和工程化能力的综合交付能力。

短期看，GPU 仍是 AI 基础设施最核心、最稀缺的资源。由于微软等大型云服务商通常会优先将 GPU 库存分配给内部 AI 团队和大客户，部分 AI 初创公司算力获取难度

上升，只能以更高价格争夺剩余服务器资源。根据现货市场价格显示，租赁 NVIDIA B200 GPU 的每小时成本在 4 月中下旬大幅上涨。部分 AI 公司在续签 Blackwell 算力合同时，面临价格上涨和较长合约期限。在这一背景下，能够更早获得新一代 GPU 的云厂商，更容易在供给紧张、客户需求强劲、价格较高的窗口期获取订单。CoreWeave 与英伟达的深度绑定本质上体现的是 GPU 供给、技术路线图、客户信任和资本市场背书的总和优势。Nebius 也通过英伟达生态支持和大规模 AI Factory 建设获取高端 GPU 和客户合同。进一步看，Blackwell、Vera Rubin 等平台并不只是单一 GPU 升级，而是同步推动服务器形态、网络互联、液冷系统、电力密度和软件栈迭代。对于 NeoCloud 而言，每一轮新平台导入都会带来新的工程适配和交付要求，能够更快获取新平台、完成集群调试并交付稳定产能的厂商，将持续拥有阶段性溢价。

图5：NVIDIA B200 GPU 现货市场租赁价格

数据来源：The Information，东吴证券研究所

中长期看，电力约束已经成为头部 NeoCloud 扩张叙事中的核心指标。根据 IEA 2026 年最新测算，全球数据中心用电量预计将由 2025 年的约 485TWh 提升至 2030 年的约 950TWh，基本实现翻倍，并将在 2030 年占全球电力需求约 3%；其中 AI-focused 数据中心用电量增长更快，预计 2025-2030 年约增长 3 倍。区域上，美国是本轮数据中心电力需求增长的核心区域。IEA 指出，美国将贡献全球数据中心用电增量的最大部分，到 2030 年，美国数据中心用电强度预计将超过 1200kWh/人，约相当于美国家庭年均用电量的 10%。

电力短缺并不只是发电量不足，更体现为并网和输电瓶颈。Lawrence Berkeley National Laboratory 最新数据显示，截至 2025 年底，美国超过 2060GW 的发电和储能容量正在申请并网，大量发电和储能项目已经规划，但受制于并网审批、输电线路、系统影响评估和电网扩容，短期难以转化为可实际使用的电力。因此，未来 AI 云竞争不只

是 GPU 获取能力的竞争，更是可用电力、并网速度、高密度数据中心建设能力和低成本稳定运行能力的竞争。需要注意的是，签约电力容量与可用电力容量并不完全等同。对 NeoCloud 而言，更重要的是已上线电力、已并网容量、电力价格、长期购电协议、备用电源体系、液冷支持能力和高功率密度机柜交付能力。只有能够按期转化为 AI Factory 上线产能的电力资源，才真正构成收入天花板。

1.3. 商业模式聚焦：重资产 AI 基础设施运营，核心是长约、融资与交付能力的循环

NeoCloud 商业模式的本质是以重资产投入换取长期算力合同现金流。其扩张路径通常包括四个环节：首先，与大客户签订多年期算力合同、容量预留协议或 take-or-pay 合约；其次，以客户合同、GPU 资产和未来现金流为基础进行债务、股权、可转债、资产支持融资或客户预付款融资；再次，将资金投入 GPU 采购、数据中心建设、电力锁定和 AI 集群交付；最后，通过长期算力服务收入回收投资并支撑后续扩张。因此，NeoCloud 的商业闭环并不是简单的客户需求驱动收入增长，而是由客户长约、融资能力、资本开支、产能交付和现金流回收共同构成的资本循环。这一模式决定了 NeoCloud 同时具备云服务公司和基础设施资产运营商的双重属性。收入端类似云服务，依赖客户需求、合同期限和算力价格；资产端类似基础设施，依赖资本开支、融资成本、折旧周期和资产利用率。NeoCloud 的核心能力，也不只是获取 GPU，而是能否将客户需求、资金、资产、工程交付和平台能力连接为可持续的商业循环。

1.3.1. 收入模式：长约和 Backlog 决定短期能见度，平台化收入决定长期弹性

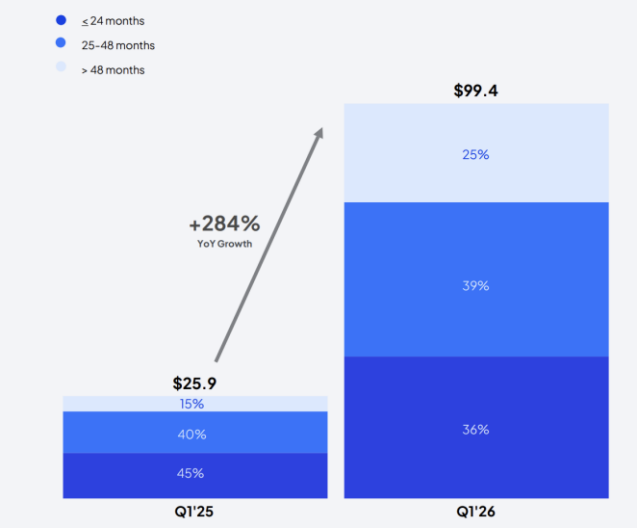
NeoCloud 收入主要来自 GPU 算力租赁和 AI 集群服务。收费形式包括 GPU 小时、节点租赁、专属集群、容量预留、多年期 take-or-pay 合同等。在 AI 算力供给紧张阶段，大客户为了保障训练和推理产能，通常愿意通过多年期合同提前锁定 GPU 资源，并通过预付款、最低采购承诺或 take-or-pay 条款提高供应确定性。CoreWeave 的模式最具代表性，其 Committed Contracts 为 take-or-pay，且历史活跃合同加权平均预付款约为 TCV 的 15%-20%。截至 1Q26，CoreWeave 收入达到 20.78 亿美元，同比增长 112%，Backlog 达到 994 亿美元，较 2025 年底的 668 亿美元进一步提升，大量未来收入已经通过客户合同提前锁定。

但 Backlog 规模并不等同于收入质量。我们认为，判断 NeoCloud 收入质量，至少需要关注四个维度：1）合同是否具备 take-or-pay 或最低采购承诺：若客户必须支付最低采购金额，收入确定性更强；若仅为框架协议或可选容量，实际兑现风险较高。2）客户信用质量：Microsoft、Meta 等高信用客户的合同，不仅收入兑现概率更高，也更容易被资本市场认可为融资基础。3）合同期限与 GPU 折旧周期是否匹配：若 GPU 资产折

旧周期长于客户合同期限，公司将面临续约价格和资产利用率不确定性。4）定价机制是否具备保护：若合同价格相对锁定，公司在未来算力租赁价格下行时更具保护；若客户具备较强重新议价权，收入弹性和毛利率可能承压。

从收入结构看，当前 NeoCloud 仍以 GPU 基础设施收入为主，软件和平台收入占比普遍较低。短期增长主要来自训练集群、大客户容量预留和算力长约。中长期若推理 API、模型优化、托管 AI 平台、开发者工具和企业 AI 工作流收入占比提升，收入结构才有机会由单纯算力租赁转向更高毛利、更高粘性的平台型收入。尤其在 AI 工作负载由训练扩展至推理和 Agent 应用后，客户需求将更加关注单位 token 成本、推理延迟、模型部署效率和多工作负载调度能力。平台化不再只是估值提升选项，而是 NeoCloud 应对 AI 云竞争加剧的必要条件。

图6：CoreWeave Backlog



数据来源：CoreWeave 财报，东吴证券研究所

图7：NeoCloud 收入质量判断

数据来源：东吴证券研究所

1.3.2. 客户结构：Hyperscaler 和 AI Lab 贡献短期需求，企业与开发者客户决定长期天花板

NeoCloud 的核心客户主要分为以下三类：

第一类是 Hyperscaler 和大型科技公司。它们自身 AI 需求增长快，但自建 AI 数据中心、部署 GPU 和调试大规模集群需要时间，因此会通过 NeoCloud 快速获得外部算力，作为短期产能补充。微软、Meta 等客户与 CoreWeave、Nebius 签订大规模长期合同，体现的正是这一逻辑。对这类客户而言，NeoCloud 的价值在于交付速度、可用产能和供应链弹性，而不只是价格。

第二类是 AI Lab 和大模型公司。这类客户对 GPU 需求极大，但通常采取多供应商

策略，以保障训练和推理算力供应，并降低对单一云平台的依赖。对于这类客户而言，NeoCloud 的价值在于能够提供新一代 GPU、大规模集群、较短交付周期和更灵活的云合作关系。尤其是在新一代 GPU 供给紧张阶段，谁能率先交付可用集群，谁就更容易获得模型公司的训练订单。

第三类是企业客户和 AI 应用开发者。随着 AI 应用从模型研发走向商业落地，金融、医药、制造、互联网、软件、自动驾驶、机器人等行业客户会产生模型微调、推理部署、Agent 应用和私有化 AI 服务需求。这部分客户可能不需要超大规模训练集群，但需要稳定、低延迟、可扩展和成本可控的 AI 算力。若 NeoCloud 能够通过推理平台、开发者工具和企业 AI 工作流切入这类客户，其客户结构将从少数大客户长约逐步扩散至更多分散化需求。

现阶段头部 NeoCloud 客户结构仍然偏集中。大客户长约能够提高收入可见度，但也带来客户集中度风险。一旦主要客户削减订单、推迟部署、转向自建产能或迁移至其他平台，公司收入和资本开支回报率都会受到影响。

1.3.3. 融资能力：扩张的前置条件，决定长约能否转化为真实产能

融资能力是 NeoCloud 商业模式中最关键的环节之一。融资能力对 NeoCloud 的意义在于：1）影响产能扩张速度：融资效率越高，GPU 采购、电力锁定和数据中心上线速度越快；2）影响利润留存能力：若债务成本过高，利息费用会显著侵蚀经营利润；3）影响资产负债表安全性：若公司过度依赖短债或高息债，一旦 GPU 租赁价格下行、客户需求延后或资本市场收紧，现金流与债务到期之间的错配风险将迅速放大；4）影响客户信任：大型客户与 NeoCloud 签订多年期 AI 基础设施合同，本质上是在判断供应商未来数年能否持续交付 GPU、电力和数据中心产能，若供应商融资能力不足，即使客户存在需求，也未必愿意签订大额长约。

从融资模式看，NeoCloud 已经从传统科技公司股权融资，转向“客户长约+资产抵押+债务融资”的重资产基础设施融资模式。由于 GPU、数据中心、电力和液冷设施前期投入巨大，单纯依靠股权融资难以支撑快速扩张，头部 NeoCloud 普遍通过多年期限客户合同锁定未来现金流，并进一步将客户预付款、最低采购承诺或 take-or-pay 条款转化为融资基础。债务融资也是 NeoCloud 扩张的核心杠杆。CoreWeave 是最典型案例，公司一方面通过 OEM 融资、设备融资和担保债务融资工具支持 GPU 采购，另一方面发行优先票据和可转债补充长期资金。与此同时，供应商和生态方支持也成为 NeoCloud 融资体系的重要组成部分。因此，NeoCloud 融资能力的核心不只是能否融到钱，而是能否形成客户、资产和资本三者之间的闭环。高信用客户长约决定收入能见度，客户预付款和 take-or-pay 条款降低前期资金压力，GPU 与数据中心资产支持债务融资，供应商背书进一步降低融资难度。但这一模式也放大了杠杆和交付风险，一旦 GPU 到货、数据中心建设、电力接入或客户上线进度不及预期，公司可能同时面临收入延迟、利息压

力上升和资产利用率不足的问题。

表2：NeoCloud 主要融资方式

融资方式	具体形式	案例	核心作用	主要风险
客户预付款	大客户在合同初期支付部分款项，用于支持 GPU 采购和数据中心建设	IREN 与 Microsoft 97 亿美元五年合同包含 20%预付款；IREN 表示将通过现有现金、客户预付款、经营现金流和其他融资支持合同相关 CapEx	降低前期资金压力，把客户需求直接转化为建设资金	若交付延迟，可能触发退款、赔偿或合同终止风险
合同支持债务融资	以 Microsoft、Meta 等高信用客户长约为基础发行债务或获得贷款	Nebius 与 Microsoft 174 亿美元 AI 基础设施合同中，公司明确表示将通过合同现金流及以该合同为担保的债务融资支持 CapEx，融资条件受 Microsoft 信用质量改善	用客户信用提升融资能力，降低债务成本	高度依赖大客户合同兑现，客户集中度和交付风险较高
设备融资/GPU 抵押融资	以 GPU 服务器、网络设备、数据中心设备等作为抵押，获得设备贷款或 OEM 融资	CoreWeave 披露，截至 2025 年 9 月底，公司 OEM Financing Arrangements 相关设备融资名义余额约 19 亿美元	适合 GPU 等可识别资产采购，资金用途明确	GPU 技术迭代快，抵押品价值可能快速折旧
项目贷款/延迟提款贷款	以数据中心项目、客户合同、设备资产和未来现金流为支持，获得大额贷款额度	CoreWeave 此前通过大额延迟提款贷款支持扩张；2025 年 7 月又完成 26 亿美元 secured debt financing facility	与 CapEx 节奏匹配，支持分阶段建设	利率较高、财务杠杆提升，若利用率或交付不及预期，偿债压力上升
高收益债/优先票据	发行 senior notes 或其他债券补充长期资金	CoreWeave 2025 年 5 月发行 20 亿美元 2030 年优先票据，2025 年 7 月发行约 18 亿美元 2031 年优先票据	延长债务期限，补充大规模扩张资金	提高利息负担，对信用评级和资本市场窗口依赖较高
可转债融资	通过可转债获得低于普通债务的票息，同时给予投资者未来转股权	CoreWeave 在 FY2025 后披露完成约 26 亿美元可转债融资；Nebius 2026 年完成约 43.4 亿美元可转债融资；AI 相关可转债发行在 2026 年明显增加	相比纯债降低现金利息压力，适合高成长公司	股价上涨时带来摊薄；股价承压时仍是债务负担
股权融资/IPO/战略投资	私募融资、IPO、增发或战略投资	Lambda 2025 年 2 月完成 4.8 亿美元融资；CoreWeave 已上市；Nebius 曾出售认股权证给 NVIDIA	补充永久资本、降低杠杆，增强市场信用	摊薄股东权益，市场窗口影响较大
供应商/生态方支持	NVIDIA、Google 等通过投资、容量购买、租回或付款担保支持 NeoCloud 融资	NVIDIA 与 Lambda 签署约 15 亿美元 GPU 租回协议；NVIDIA 承诺购买 CoreWeave 未售出容量；Google 在 Fluidstack 相关数据中心租约中提供付款支持	提升客户与融资方信心，降低产能空置风险或融资成本	生态绑定加深，供应商、客户、融资方之间关系复杂
数据中心/电力资产合作	与数据中心运营商、电力资源方、矿企转型公司合作，通过租赁、合资或长期容量协议扩张	IREN 利用其 Childress 园区约 750MW 电力资源承接 Microsoft 合同；Fluidstack、Cipher、TeraWulf 等也通过长期数据中心租约融资扩张	缓解自建数据中心和电力获取压力，加快上线速度	长期租赁负债较重，电力交付与并网风险仍存在

数据来源：新浪财经，证券时报，各公司公告，各公司财报，Davis Polk，路透社，华尔街日报，东吴证券研究所

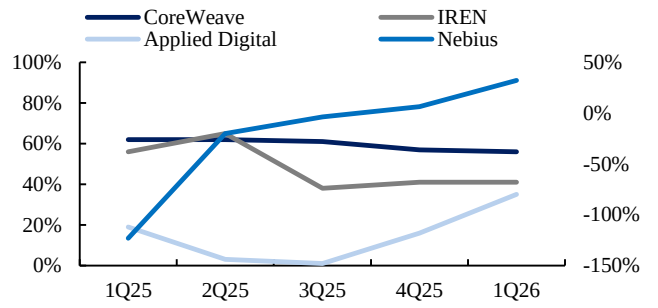
1.3.4. 成本结构与盈利能力：TCO 差异决定经营利润，资本开支与融资成本压制净利润

NeoCloud 的盈利能力具有明显分化特征，本质上取决于算力租赁价格与全生命周期 TCO 之间的差额。收入端看，在 AI 算力供给紧张、GPU 利润率较高、客户愿意通过长约锁定产能的阶段，已上线 GPU 集群能够产生较强经营利润。但成本端看，NeoCloud 是典型的重资产基础设施运营商，其成本结构包括 GPU 服务器采购、网络设备、存储、数据中心租赁或自建、电力、液冷、运维、折旧摊销和融资成本。上述成本中，GPU 和服务器决定初始 CapEx，电力和数据中心决定长期运行成本，折旧和利息费用则直接影响利润表和自由现金流。因此，NeoCloud 即使在 EBITDA 层面表现较强，也可能长期面临净利润和自由现金流压力。

从经营层面看，成熟算力资产在高利用率环境下具备较强盈利能力。1Q26，CoreWeave 收入达到 20.78 亿美元，同比增长 112%，Adjusted EBITDA 达到 11.57 亿美元，对应 Margin 为 56%；Nebius 收入达到 3.99 亿美元，同比增长 684%，Adjusted EBITDA 达到 1.30 亿美元，对应 Margin 为 32%。但 EBITDA 并不等同于最终盈利质量。1Q26，CoreWeave Adjusted Net Loss 仍达 5.89 亿美元，对应 Adjusted Net Loss Margin 为 28%；Nebius Adjusted Net Loss 为 1 亿美元，对应 Margin 约 25%。NeoCloud 在 EBITDA 层面能够较快实现盈利，但折旧摊销、利息费用、减值、股权激励以及持续资本开支仍会显著压制净利润。

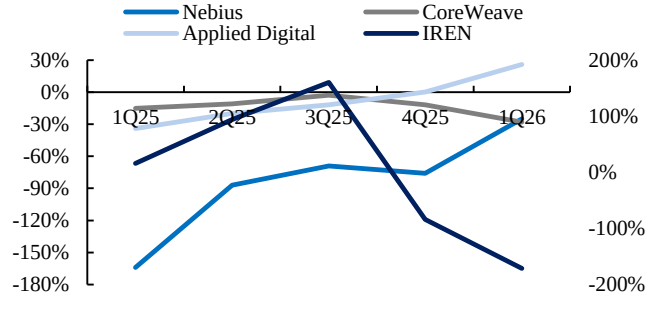
我们将 NeoCloud 盈利能力拆分为：NeoCloud 盈利能力=单位算力收入-单位算力 TCO-资本成本。其中，单位算力收入主要取决于 GPU 租赁价格、长约定价、客户结构和资产利用率；单位算力 TCO 包括 GPU 服务器采购成本、电力成本、数据中心租赁或自建成本、网络设备、液冷系统、运维成本和维护费用；资本成本则包括折旧摊销、债务利息、设备融资成本以及潜在股权摊薄。SemiAnalysis 关于 AI Cloud TCO 的分析也表明，不同 NeoCloud 之间的 TCO 并不相同，CoreWeave 凭借溢价定价成为白金供应商，在大型 LLM 预训练、多模态强化学习研究、推理端点等场景中拥有显著成本优势。其研究还显示，即便 GPU 小时定价相同，超大规模云厂商在 TCO 调整后的成本可能比金牌供应商贵 10%以上，而银牌新兴云厂商可能贵 15%以上。

图8：NeoCloud 厂商 Adjusted EBITDA Magin (%)



数据来源：各公司财报，东吴证券研究所

图9：NeoCloud 厂商 Adjusted Net Income Margin (%)

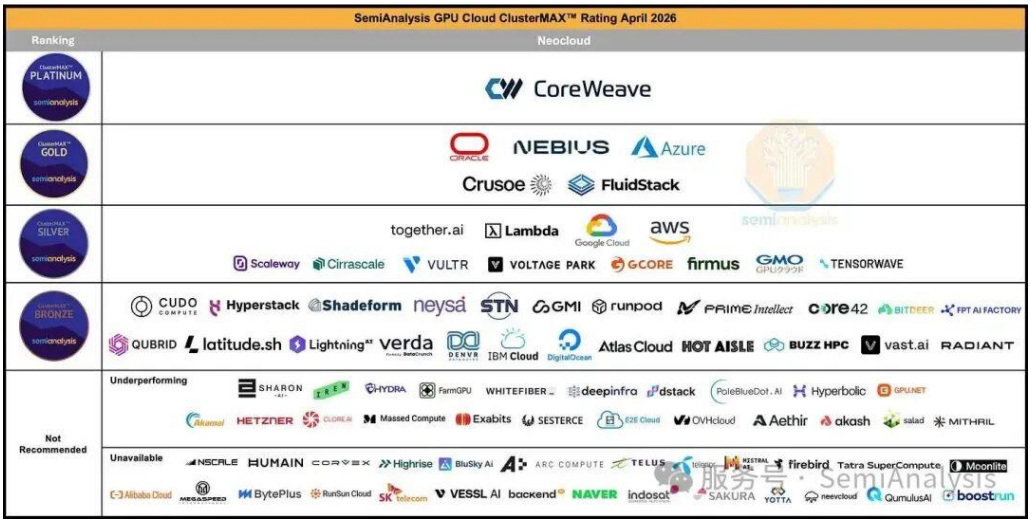


数据来源：各公司财报，东吴证券研究所

注：Nebius 为右轴，IREN/Applied Digital 财务季度调整，1Q25 对应 FY3Q25

注：IREN 为右轴&GAAP Net Income Margin

图10：SemiAnalysis ClusterMAX 2.1 排名，2026 年 4 月



数据来源：SemiAnalysis，东吴证券研究所

1.3.5. 商业模式是否可持续：供需错配创造窗口，平台化影响终局

NeoCloud 商业模式是否可持续，是市场关注重点之一。我们认为，应该分阶段考虑。短期来看，商业模式具有较强支撑。AI 算力需求仍处于高速增长期，GPU、电力和数据中心供给存在刚性约束，头部 NeoCloud 已经通过大客户长约锁定大量未来收入。在供给紧张环境下，NeoCloud 可以维持较高利用率和较强定价能力。

中期来看，可持续性取决于供需缺口能维持多久。如果 AI 训练和推理需求继续高增，而电力、数据中心和高端 GPU 供给仍然紧张，NeoCloud 仍将享受较高景气度。但如果英伟达扩产、Hyperscaler 自建产能释放、替代芯片成熟，当前算力供给短缺可能缓解，GPU 租赁价格和合同条款可能承压。届时，只有具备低成本融资、强交付能力和高质量客户合同覆盖的头部公司能够维持竞争力。

长期来看，商业模式可持续性的核心在于能否平台化。如果 NeoCloud 长期只是借钱买 GPU、签约租出去、用现金流还债，其本质更接近高杠杆基础设施租赁公司，面临资产折旧、融资周期和价格竞争压力。真正可持续的模式，应当是在底层算力基础上形成推理平台、开发者生态、企业 AI 工具和客户工作流嵌入，把收入从单纯 GPU 租赁升级为“算力+软件+平台”的复合收入。

从这一角度看，NeoCloud 行业的核心问题是头部公司能否在供需窗口期完成能力沉淀。若公司只能依赖 GPU 供给紧张赚取阶段性价差，长期将面临供给改善后的估值

重估。若公司能够在重资产扩张过程中沉淀客户、工程、融资和平台能力，则有机会从算力资源型公司升级为 AI 基础设施平台型公司。Nebius 和 CoreWeave 之所以值得重点研究，正式因为二者分别代表了 NeoCloud 赛道的两条典型路径：前者从 AI Factory 基础设施出发，尝试向推理平台和全栈 AI 基础设施服务商演进；后者更偏向合约驱动型 AI Hyperscaler，通过英伟达绑定、大客户长约和资产融资实现快速扩张。基于对行业商业模式分析和未来云计算竞争格局判断，我们更看好 Nebius 长期发展潜力。

2. Nebius：从基础设施出发，迈向全栈 AI 平台

2.1. 公司概览：Yandex 基因与 AI 原生重构

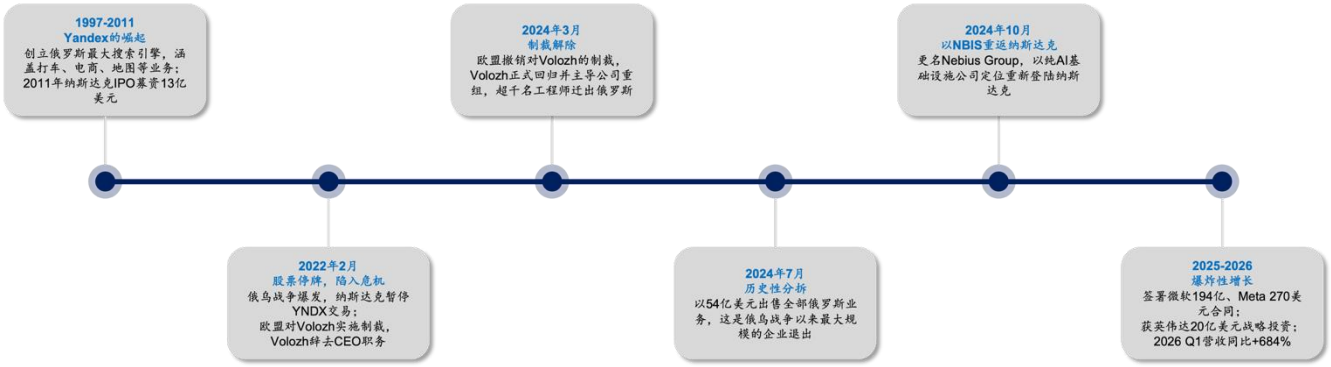
Nebius 是一家总部位于荷兰阿姆斯特丹的 AI 原生云平台公司，专注于为大模型训练、推理及企业 AI 应用提供全栈 AI 基础设施服务。公司的核心产品是面向 AI 和机器学习高强度工作负载深度优化的 GPU 集群平台，覆盖 GPU 算力、高性能网络、存储、数据中心基础设施以及 AI 工作负载管理等完整技术栈。与 AWS、Azure 等追求“大而全”的通用云厂商不同，Nebius 的定位更接近一个 AI Factory 运营商。其核心价值不来自简单的 GPU 租赁，而是将 GPU、服务器、网络、存储、电力、冷却、调度系统和开发者工具整合为可持续运行、可规模化交付的 AI 基础设施能力，并正加速向包含推理服务和 AI 工作负载管理在内的软件平台升级。

公司的战略起点决定了其与众不同的竞争禀赋。Yandex 曾是俄罗斯最大的互联网科技集团之一，长期在搜索、广告、云计算、自动驾驶和数据中心等领域积累了深厚的工程能力。2022 年俄乌冲突爆发后，公司创始人 Arkady Volozh 被列入欧盟制裁名单，公司随即启动了国际业务与俄罗斯核心业务的彻底切割。2024 年，Yandex 俄罗斯业务以约 26 亿美元价格出售给俄罗斯本土买家，国际业务正式以 Nebius 品牌独立运营，由 Volozh 重新出任 CEO，核心工程团队随之转移，并于 2024 年 10 月重返纳斯达克。这一历史背景的关键战略意义在于，Nebius 并非从零起步搭建 GPU 租赁业务的新公司，而是在既有的云计算架构设计、分布式系统工程、数据中心运营与工程研发能力的基础上，针对 AI 工作负载进行全面的原生化重构。Yandex 时期积累的大规模分布式系统、低延迟实时计算和云服务运营能力，均可直接迁移到 GPU 集群调度优化、推理延迟控制和多租户资源管理等 AI 云核心场景。这是普通 GPU 采购型转租商所不具备的底层能力，也是 Nebius 能够在短时间内赢得微软、Meta 等超大规模客户信任的重要基础。

公司已形成多元化、高质量的资本支撑体系。自 2024 年 12 月重返纳斯达克以来，Nebius 通过权益融资、战略认股权证和可转换票据三条渠道，在不足一年半的时间内累计融资逾百亿美元，且每一轮均获超额认购或主动扩容，充分反映机构投资者对公司商业模式的高度认可。权益端，2024 年 12 月，公司完成由 NVIDIA、Accel 及 Orbis 参与

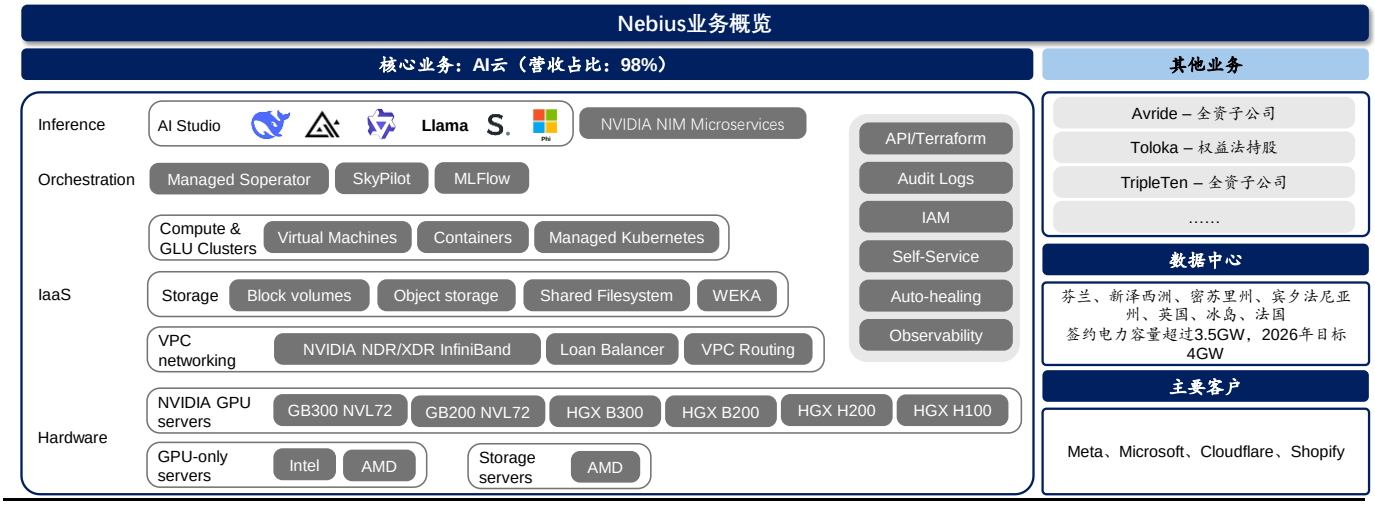
的 7 亿美元私募定向增发，发行价每股 21 美元，NVIDIA 自此轮起以战略投资者身份介入，其背书的核心价值不在于资金本身，而在于为 Nebius 锁定了 GPU 供应链的优先分配权。2025 年 9 月，公司再度完成约 10 亿美元公开增发，与同期可转换票据并行发行。债券端，公司先后完成三轮可转换优先票据发行：1）2025 年 6 月首轮 10 亿美元；2）2025 年 9 月，紧随微软合约落地，第二轮从原定 20 亿美元扩容至 27.5 亿美元，以合约现金流为信用抵押的资产支持结构使融资成本显著低于无担保债务；3）2026 年 3 月，第三轮从 37.5 亿美元进一步扩容至 43.4 亿美元。与此同时，NVIDIA 于 2026 年 3 月以认购预融资认股权证的方式向公司注资约 20 亿美元，以象征性行权价获得可转换为逾 2100 万股 A 类普通股的权证，进一步强化了双方在 GPU 供应链上的战略绑定。

图11：Nebius 发展历程



数据来源：人民日报，纽约时报，财联社，智东西，电子工程专辑，华盛通，路透社，Nebius，东吴证券研究所

图12：Nebius 业务概览



数据来源：Nebius，东吴证券研究所

图13：Nebius 融资梳理（2024 年 12 月-2026 年 3 月）

时间	融资类型	规模	战略意义
权益融资（合计约17亿美元）			
2024年12月	私募权益	7亿美元	复牌后首轮融资，超额认购；NVIDIA战略入股，锁定GPU供应链优先分配权
2025年9月	公开增发	10亿美元	微软合约落地后配套融资，与27.5亿美元可转换票据同期发行
NVIDIA战略认股权证投资（合计约20亿美元）			
2026年3月	认股权证	20亿美元	GPU分配优先权实质性强化
可转换优先票据（合计约87.5亿美元）			
2025年6月	可转债	10亿美元	首轮可转债，用于GPU采购及数据中心建设
2025年9月	可转债	27.5亿美元	紧随微软174亿美元合约落地；资产支持融资结构
2026年3月	可转债	43.4亿美元	与NVIDIA认股权证同步发行，迄今最大单轮

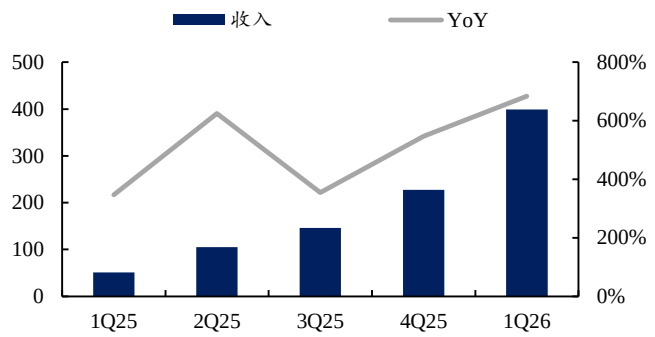
数据来源：Nebius，路透社，东吴证券研究所

2.2. 财务分析：收入兑现与资本开支并行，经营杠杆进入验证期

收入高增已获财务验证，ARR 与客户长约共同夯实能见度。1Q26，Nebius 实现集

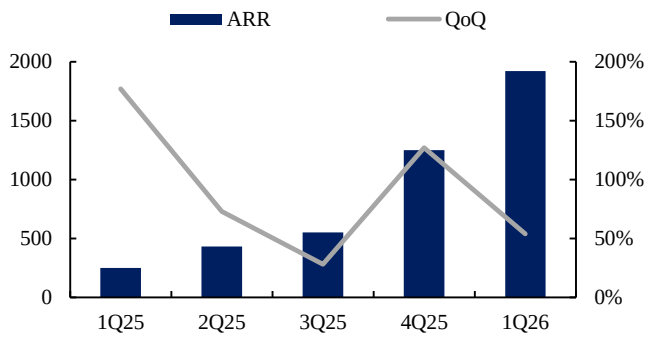
团收入 3.99 亿美元，同比增长 684%，环比增长 75%；核心 AI Cloud 收入 3.90 亿美元，同比增长 841%，环比增长 82%，占总收入约 98%。收入加速并非单纯来自低基数，而是产能扩张、GPU 供需紧张、定价韧性和高利用率共同推动的结果。从客户承诺看，微软和 Meta 构成了公司未来收入可见度的两大核心锚点，两笔合约均为 5 年期，法律约束力明确，且均含分批交付安排，使收入释放具备较高的确定性与节奏可预期性。ARR 快速提升也验证产能正在持续被客户消化。1Q26 末，AI 业务 ARR 达到 19.2 亿美元，较上季度末的 12.5 亿美元环比增长超过 50%。同时，公司维持全年 ARR 指引 70-90 亿美元。ARR 的快速提升表明，新增 AI Cloud 产能正在稳定转化为年化收入，而非依赖一次性交付拉高单季数字。后续判断收入增长质量，不应只看单季增速，更应跟踪 ARR 的绝对水平与环比增速、客户长约覆盖率、各 AI Factory 站点实际投产节奏以及 GPU 利用率。微软合约 2026 年下半年主力批次交付是下一阶段 ARR 能否接近指引上限的最关键单一变量。

图14：Nebius 收入（百万美元）与同比增速（%）



数据来源：Nebius 财报，东吴证券研究所

图15：Nebius ARR（百万美元）与环比增速（%）



数据来源：Nebius 财报，东吴证券研究所

图16：Nebius 主要收入来源

客户/合作方	已披露规模	合同性质	确定性强弱	主要意义
Microsoft	174亿美元，最高194亿美元	5年期AI基础设施协议	高，另有20亿美元潜在扩展	高信用客户长约，验证Hyperscaler交付能力
Meta	120亿美元确定采购，最高270亿美元	5年期AI基础设施供应协议	120亿美元较高，额外150亿美元偏条件性	降低单一客户依赖，并为新增产能提供需求兜底

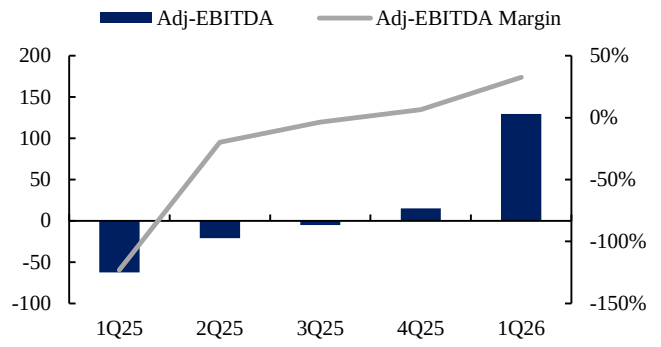
数据来源：Nebius 公告，东吴证券研究所

经调整 EBITDA 由负转正，经营杠杆开始实质性释放。1Q26，公司经调整 EBITDA 由去年同期亏损 5370 万美元转为盈利 1.295 亿美元，对应经调整 EBITDA Margin 由-105.5%提升至 32.5%，AI 云业务经调整 EBITDA Margin 从 4Q25 的 24%进一步扩张至 45%。这一改善并非费用压缩的结果，其底层机制是 GPU 集群、数据中心、电力与网络

等大量前置固定成本随收入放量被快速摊薄，同时新一代 GPU 定价能力更强推动毛利率提升，客户向更长期合约迁移也进一步改善了单位算力的收入产出。

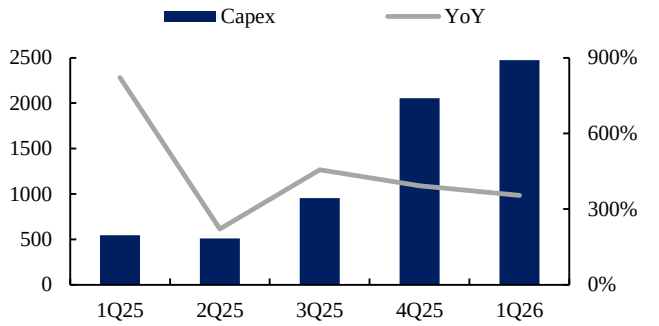
高 Capex 是 Nebius 成长的基础，也是市场估值分歧的核心。公司已将 2026 年 Capex 指引从 160-200 亿美元上调至 200-250 亿美元，约为全年收入指引的 6-8 倍。孤立看绝对金额意义有限，判断 Capex 质量的核心维度在于：新增产能是否属于“订单驱动型扩张”，而非基于市场乐观预期的激进押注。从现有证据看，此轮 Capex 上调的主要驱动是 Meta 在 2026 年 3 月扩容协议中已明确承诺的 2027 年算力需求。客户不仅在合同上承诺采购，更已用真金白银为未来产能买单，部分建设风险由此从公司转移至客户侧，有效降低了前置资本投入的不确定性。若微软、Meta 等高信用客户的长约持续覆盖新增产能，高 Capex 更接近未来收入的前置锁定；反之，若 AI 算力供需关系出现逆转或新增产能未能被充分消化，高资本开支将迅速转化为产能空置、折旧压力与资产回报率下降。因此，后续跟踪的核心指标不是 Capex 绝对金额，而是订单覆盖率、产能上线节奏、GPU 利用率与单位 Capex 产出。

图17：Nebius Adj-EBITDA（百万美元）和 Margin



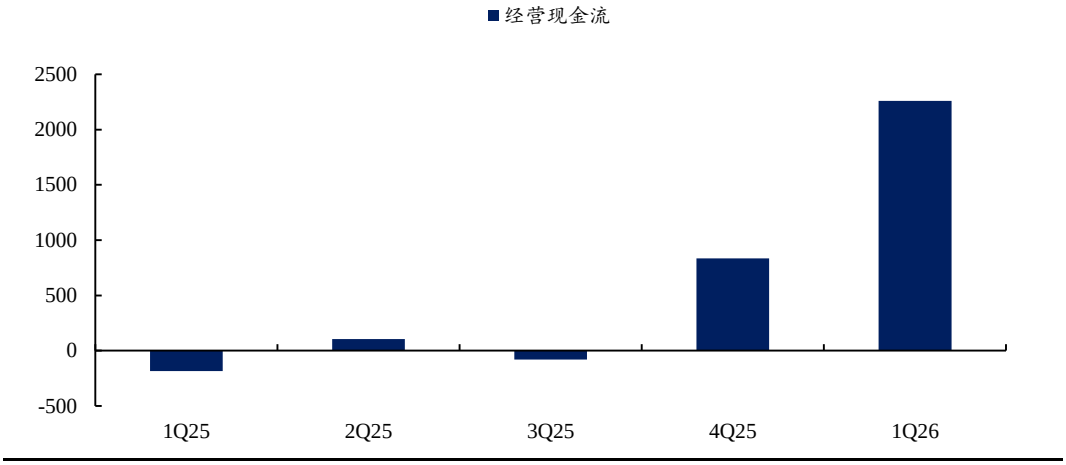
数据来源：公司财报，东吴证券研究所

图18：Nebius Capex（百万美元）和同比增速（%）



数据来源：公司财报，东吴证券研究所

图19：Nebius 经营现金流（百万美元）



数据来源：公司财报，东吴证券研究所

2.3. 核心优势：全球 AI Factory、电力卡位与工程平台能力构筑成长壁垒

2.3.1. 优势一：全球 AI Factory 布局，支撑大客户交付与区域合规需求

全球 AI Factory 布局是 Nebius 区别于其他厂商的重要基础。对于 AI 云厂商而言，GPU 资源只是基础，真正决定公司能否承接大客户需求的是可交付产能。可交付产能不仅包括 GPU 和服务器本身，还涵盖数据中心场地与电力接入、冷却系统与机架设计、运维体系与 SLA 保障能力，以及日益重要的区域合规与数据主权能力。Nebius 正在全球范围内系统性地构建其 AI Factory 网络——从美国东部的新泽西州、宾夕法尼亚州，到中部的密苏里州，再到欧洲的芬兰、法国、英国，以及中东的以色列，形成了覆盖主要 AI 算力需求地区的多站点布局。

美国市场是 Nebius 当前最核心的营收来源。新泽西州瓦恩兰数据中心专门为微软服务，已于 2025 年开始分批上线，是微软 174 亿美元合约的主要交付节点。密苏里州独立城站点于 2026 年 5 月正式奠基，是公司首个 GW 级自建 AI Factory，占地超过百公顷，完工后将成为面向多客户的通用 AI 云主力产能。宾夕法尼亚州新站点是目前已知的最大单体项目，最高可支持 1.2GW 电力容量，主要面向 2027 年的新增产能需求，与此前披露的 Meta 扩容协议形成明确的需求-供给对应关系。

欧洲市场则承担着完全不同的战略功能。欧盟《AI 法案》和 GDPR 构建了全球最严格的数据合规框架，使本地数据主权成为大量欧洲企业和政府机构在采购 AI 基础设施时的优先考量因素。Nebius 总部位于阿姆斯特丹，并在芬兰、法国、英国以及冰岛等地推进布局，具备在欧洲本地提供合规算力服务的独特能力。北欧站点的额外优势在于能源成本。芬兰和冰岛拥有充裕的可再生水电和风电资源，数据中心电力使用效率可接近 1.1，明显低于 Uptime Institute 2024 年披露的全球数据中心平均值 1.56，意味着相同的算力产出所需总能耗更少，单位运营成本更低。

全球布局的深层价值，在于使 Nebius 能够向微软、Meta 等超大规模客户提供跨地区的长期算力保障。这类客户采购的从来不是简单的 GPU 实例，而是能够持续扩容、满足区域合规要求并保障 SLA 的长期算力伙伴。Nebius 通过多区域 AI Factory 布局，向市场证明了其具备工程复制和跨区域交付能力，这是普通 GPU 采购转租商所无法提供的。然而，全球化布局的另一面是执行风险的累积。多区域同步建设大幅提高了项目管理的复杂度，例如宾夕法尼亚州 1.2GW 站点从签约到全面上线，至少需要经历电力并网、变电站建设、GPU 部署和 SLA 验收等多个关键环节。后续各站点的实际投产节奏和客户上线情况值得重点关注。

图20：Nebius 数据中心布局

地点	模式	状态	当前/规划容量	详情
			欧洲	
芬兰·曼采莱	自建	运营中	75 MW	继承自Yandex的原始旗舰数据中心，2024年完成三倍扩容，GPU扩展至约6万张，当前容量完全售罄
芬兰·拉彭兰塔	自建	建设中	310 MW	新建超大规模AI专用园区，满负荷运行后将是欧洲最大AI专用计算设施之一
法国·巴黎	托管	运营中	少量	早期欧洲扩张托管节点，主要服务法国及西欧客户，随着贝蒂讷大型自建节点推进，巴黎节点将逐步迁移整合
法国·贝蒂讷	自建	规划/建设	240 MW	大规模AI专用超大规模数据中心改造项目，将是法国最大AI计算节点
冰岛·凯夫拉维克	托管	运营中	10 MW	100%可再生能源驱动，天然低温提供免费冷却，对注重ESG的欧洲企业客户吸引力显著，绿色数据中心认证优势突出
英国·萨里郡	托管	运营中	~4000 GPU	欧洲首批Blackwell Ultra GPU部署之一，液冷高密度设计，支持英国初创、科研及公共部门，自备发电规避英国电网接入限制
			美国	
密苏里州·堪萨斯城	托管	运营中	5-40 MW	美国首个落地节点，由堪萨斯城星报印刷厂改建，作为“快速上市”验证站点证明了Nebius快速部署能力，二期扩容于2025年Q2完成
新泽西州·瓦恩兰	自建	建设中	300 MW（首期）	按Nebius专有设计建造，分阶段建设，支持大规模专属实例，自备发电架构规避东部电网接入瓶颈，电力资源自主可控
宾夕法尼亚州	自建	规划中	250-250 MW（首期）	满容量1.2 GW是全球最大AI专用数据中心规划之一，Meta已就宾夕法尼亚新容量预签协议（纳入270亿美元、五年框架），2027年首期点亮时已有确定客户承诺
密苏里州·独立城	自建	规划中	大规模	是现有托管节点的规模化升级，具体容量和时间表尚未公开披露
阿拉巴马州·伯明翰	自建	许可申请中	大规模	利用TVA剩余电力，电力成本具竞争力；2026年1月公众听证会引发当地社区异议，聚焦噪音、用水及土地分类问题，Nebius在临时暂停令前提交许可，主张既得权利，但获批存在不确定性
			以色列	
莫迪因	托管	运营中	8 MW	Nebius进入以色列市场首个站点，25% GPU资源贡献以色列创新局国家超算项目，体现主权AI战略布局价值
马斯米耶	托管	建设中	22 MW	与贝特谢梅什打包合同总价约8.8亿美元，两站合计将以色列部署容量从8 MW跃升至逾88 MW（含远期）
贝特谢梅什	托管	建设中	58 MW	以色列规模最大的Nebius站点，远期222 MW将成为中东重要AI算力枢纽

数据来源：Nebius 公告，Data Center Dynamics，路透社，Data Center Magazine，BAXTEL，东吴证券研究所

2.3.2. 优势二：GW 级电力资源卡位，决定中长期产能上限

AI 基础设施的竞争焦点，已从“谁能采购到 GPU”演变为“谁拥有可用的电力和数据中心”。随着 GPU 从 H100 到 H200 再到 B200、B300 的快速迭代，单 GPU 功耗持

续上升，NVIDIA 下一代旗舰平台 GB200 NVL72 机柜的满载功率预计高达 130kW 以上。在这一背景下，电力接入能力已成为数据中心建设最核心的瓶颈。根据 Lawrence Berkeley National Lab 2025 年发布的研究数据，美国电网的并网申请队列等待期中位数已接近 5 年，意味着今天没有签约电力的公司，至少在 5 年内无法新增大规模算力产能。Nebius 通过提前大规模锁定电力资源，本质上是在为未来 3-5 年的产能扩张打下无法被快速复制的竞争基础。

从公司电力储备的演进轨迹看，增速之快令人印象深刻。2025 年 8 月融资时，公司披露签约电力约 1GW；微软合约落地后，2025 年底目标提升至 2.5GW；2025 年 Q4 实际超过 2.5GW，同时有 800MW-1GW 处于连接（即可立即用于部署 GPU）状态；1Q26 末，签约电力已超过 3.5GW，远超原全年目标，促使公司随即将年底目标上调至超过 4GW。宾夕法尼亚州新站点最高 1.2GW 的签约，是这一轮目标上调的主要来源，也是 2027 年新增产能的核心载体。

理解电力资源对收入天花板的意义，可以通过一个简化框架：可用 IT 负载电力 × 单位电力 GPU 部署密度 ×GPU 平均利用率 × 单 GPU 小时价格 × 年运行小时数 = 年化收入上限。以 B200 集群为参考，B200 单 GPU 功耗约 1000W，但实际部署还需考虑服务器、CPU、内存、网络、存储、供电冗余和冷却等配套负载，因此单纯以 GPU 芯片功耗估算容量会高估实际可部署数量。Nebius 芬兰 Mäntsälä 站点披露 75MW 可容纳最多约 60,000 块 GPU，对应约 800 块 GPU/MW，可作为单位电力 GPU 部署密度的参考口径。按 Nebius 官网 B200 NVLink 价格 5.50 美元/GPU 小时估算，若采用 800 块 GPU/MW、80%平均利用率，则 1MW 有效 IT 负载对应年化收入约 3,083 万美元；反推 70-90 亿美元 ARR 约需 227-292MW 有效 IT 负载。Nebius 1Q26 股东信披露，公司 2026 年 ARR 指引为 70-90 亿美元，同时合约容量已超过 3.5GW，并将 2026 年底合约容量目标上调至超过 4GW。由此看，公司提前锁定 GW 级电力资源，确实为 70-90 亿美元 ARR 目标提供了重要资源基础。但需要注意，合约电力并不等于当期可用 IT 负载。真正决定 ARR 兑现的，是电力能否按期转化为连接电力、数据中心是否完成建设、GPU 是否完成部署、客户是否按计划上线以及实际利用率能否维持在较高水平。因此，Nebius 的>4GW 电力目标更准确地说是对中长期产能扩张能力的背书，而非对当期收入的直接等同。

北欧站点的选址策略，还体现了对运营成本的长远规划。芬兰和冰岛站点具备两大结构性成本优势：1）可再生水电和风电资源充裕，电力成本低于欧洲大陆平均水平；2）寒冷的自然气候几乎可以完全替代主动冷却，大幅降低冷却系统的资本开支和运营成本。对于 7×24 小时高负荷运行的 AI 集群而言，运营成本差异在规模化后将转化为数亿美元级别的竞争优势，是竞争对手难以通过短期选址调整来弥补的结构性差异。

图21：Nebius 签约电力容量演进

时间点	签约/目标容量
2025年8月	约1GW
2025年底	目标2.5GW
2025年Q4	超2.5GW，800MW-1GW连接中
2026年Q1	超3.5GW
2026年底	目标超4GW

数据来源：EQS，Data Center Dynamics，Nebius，东吴证券研究所

2.3.3. 优势三：工程基因与平台化升级，从卖算力到卖平台的双轮驱动

工程基因是 Nebius 区别于普通 GPU 转租商最本质的底层优势。AI 云不是简单地采购 GPU 然后出租，而是一项需要长期积累分布式系统架构、数据中心运维工艺、高性能网络工程、大规模集群调度和客户交付管理的综合能力。Nebius 脱胎于 Yandex，继承了这家互联网巨头二十余年的大规模计算系统工程积淀：搜索引擎时期的大规模并发分布式处理能力，可迁移至 GPU 集群调度与多租户任务管理；自动驾驶项目锻炼的低延迟实时控制能力，可迁移至 AI 推理延迟优化和 SLA 保障；Yandex Cloud 的多年运营经验，则直接为 AI Cloud 平台架构和客户运维提供了成熟范本。创始人 Arkady Volozh 亲任 CEO，在资本市场、大客户谈判和内部执行三端保持高度整合能力，对于一家年度 Capex 规模超过当年收入 7 倍的重资本公司而言，这种创始人级别的决策权威是维持战略连贯性的重要组织保障。超过 1000 名 AI 工程师的团队规模，在国际 Neocloud 领域处于领先地位，是公司在高强度扩张阶段保持交付质量和产品迭代速度的核心支撑。

Token Factory 是 Nebius 从 GPU 云向高附加值推理平台延伸的战略抓手。公司于 2025 年 11 月发布 Token Factory，将其定位为面向生产环境的推理平台，支持企业在开源和自定义模型上完成部署、后训练、访问管理和规模化推理，并依托 Nebius 全栈 AI 基础设施和 AI Cloud 3.0 “Aether” 实现低延迟、高吞吐和成本可控。因此，Token Factory 并非简单 API 转售，而是将模型部署、推理优化、平台治理与底层算力资源整合为统一的生产级推理服务。2026 年以来，Nebius 围绕 Token Factory 连续补强软件能力。公司先后宣布收购 Tavily 和 Eigen AI：前者提供 agentic search 能力，补足 Agent 应用对实时网络搜索、外部数据获取和 RAG 的需求；后者以约 6.43 亿美元交易对价收购，主要强化模型层推理和后训练优化能力，提升单位 GPU 的 token 产出效率。与此同时，Nebius 引入 Clarifai 核心工程和研究团队，并获得其推理与计算编排技术许可；公司明确表示，Eigen AI 偏模型层优化，Clarifai 偏系统层优化，二者结合有助于 Token Factory 形成贯穿模型、系统和基础设施的端到端推理能力。从商业模式看，Token Factory 的意义在于把底层 GPU 算力进一步包装成高粘性 API 和推理平台服务。Nebius 披露，早期客户 Prosus 在特定场景中相较前沿闭源模型实现最高 26 倍成本下降。若 Token Factory 能够

持续以低成本推理吸引开发者和企业客户，Nebius 有望通过更低 token 成本绑定客户对自有算力的使用，并将 API 层利润更多内化到自身平台。

3. CoreWeave：以 AI Hyperscaler 为定位的合约驱动型重资本玩家

3.1. 公司概览：从 GPU 矿工到 AI 云基础设施平台

CoreWeave 是美国 AI 云基础设施公司，自我定位为“AI Hyperscaler”，核心是为 AI 工作负载提供原生构建的高性能云平台。与 Nebius 一样，CoreWeave 并不追求覆盖计算、存储、数据库、企业软件等完整云服务品类，而是高度聚焦 GPU 加速计算，围绕大模型训练、推理、高性能计算和企业 AI 部署提供专用基础设施。公司的核心能力由数据中心、电力资源、NVIDIA GPU 集群、高性能网络、Kubernetes 编排栈以及面向 AI 负载的软件工具共同构成，本质上是将 GPU 硬件、数据中心工程、网络架构和软件调度能力打包为可规模化交付的 AI 算力平台。

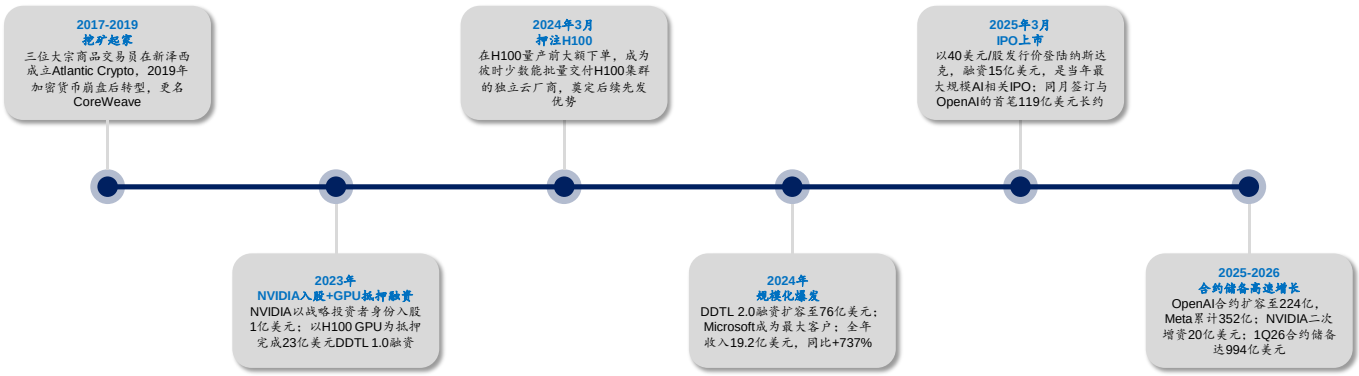
从基础设施规模看，CoreWeave 已经成为独立 AI 云厂商中最具代表性的重资产玩家之一。截至 1Q26，公司活跃运营数据中心数量达到 49 个；可用电力容量超过 1GW，已签约电力容量超过 3.5GW，并计划在 2030 年前建设超过 5GW 的 AI Factory 产能。这意味着 CoreWeave 已经从早期 GPU 云服务商，逐步演化为围绕电力、数据中心、GPU 和客户长约进行系统化扩张的 AI 基础设施平台。

公司的发展路径决定了其与传统云厂商截然不同的组织基因。CoreWeave 由 Michael Intrator、Brian Venturo 和 Brannin McBee 等人在 2017 年创立，早期名为 Atlantic Crypto，主营业务是利用 GPU 进行加密货币挖矿。加密市场下行后，公司并未清算 GPU 资产，而是将高密度 GPU 部署、电力管理、散热运维和集群调度能力迁移至 GPU 云服务，并于 2019 年更名为 CoreWeave。虽然早期挖矿业务并非面向 AI 场景，但其沉淀的底层工程能力，恰好与后续大模型训练和推理所需的 GPU 密集型基础设施高度匹配。

2022 年后，CoreWeave 抓住 NVIDIA GPU 迭代窗口，完成从 GPU 云厂商到 AI 基础设施平台的跃迁。公司在 H100 早期供应阶段大规模押注 NVIDIA 最新 GPU，成为少数能够较早批量交付 H100 集群的非超大规模云厂商之一。此后，公司与 NVIDIA 的关系从供应链合作进一步升级为资本、技术和生态绑定。NVIDIA 在 2023 年参与 CoreWeave 融资；2026 年 1 月进一步以每股 87.20 美元价格投资 20 亿美元，并宣布双方将在 AI Factory 建设、NVIDIA 软硬件验证和大规模 AI 基础设施交付上深化合作。对 CoreWeave 而言，NVIDIA 不仅是关键供应商，也是战略股东和生态背书方；但这种绑定也使公司暴露于 NVIDIA 生态、GPU 供给和产品迭代周期之下。

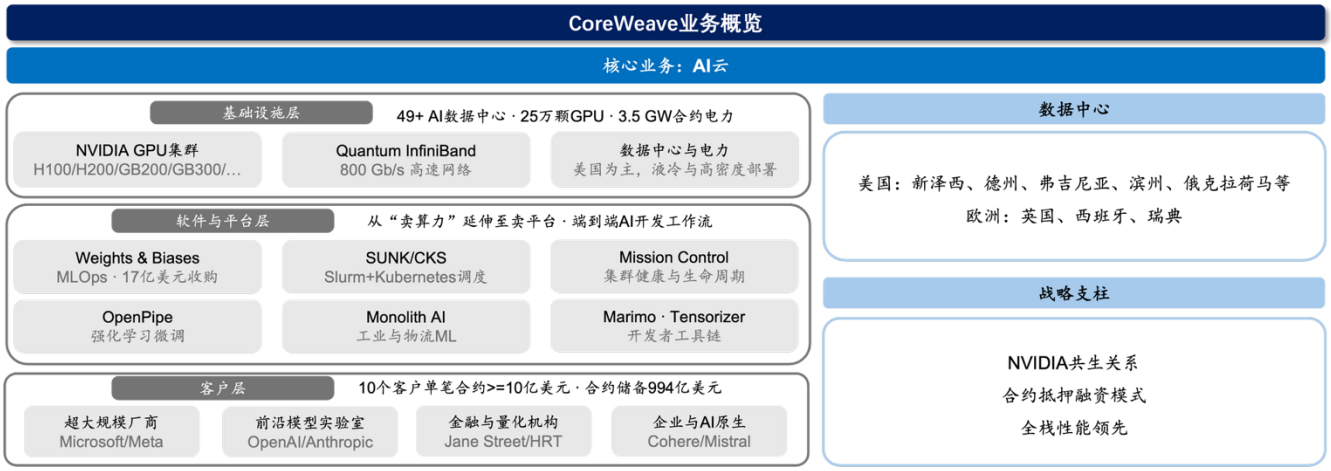
更重要的是，CoreWeave 率先将 GPU 基础设施资产“金融化”，形成区别于多数 NeoCloud 厂商的扩张能力。2023 年，公司曾以 H100 GPU 等资产为抵押完成大额债务融资，开创了以 GPU 硬件和客户合约现金流作为融资基础的模式。此后，公司持续围绕客户长约、GPU 资产和数据中心项目推进延迟提取定期贷款融资，形成“签订客户长约-锁定 GPU 与数据中心资源-以资产和合约现金流融资-建设并交付产能-确认收入和现金流”的闭环。2026 年 3 月，公司完成 85 亿美元 DDTL 4.0 融资，获得 Moody’s A3 和 DBRS A(low)投资级评级，公司称这是首个由 HPC 基础设施和相关客户合约支持、达到投资级评级的融资安排。

图22：CoreWeave 发展历程



数据来源：CoreWeave，Businesswire，36 氪，路透社，证券时报，东吴证券研究所

图23：CoreWeave 业务概览



数据来源：CoreWeave，路透社，OpenPipe，雅虎财经，Marimo，东吴证券研究所

图24：CoreWeave 融资梳理

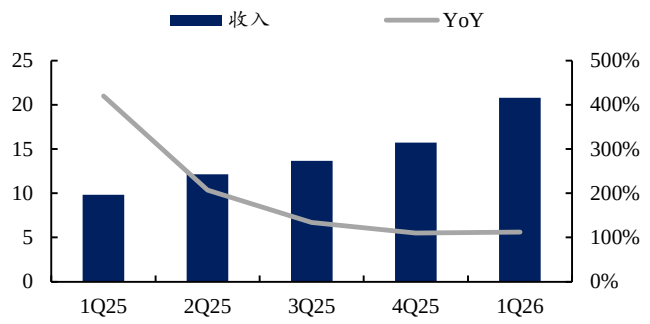
时间	融资类型	规模	战略意义
			早期股权融资及二次出售（合计约28-30亿美元）
2021年11月	B轮	0.5亿美元	Magnetar领投，CoreWeave完成转型为AI云后首次主流机构融资
2023年4-5月	B轮及扩展融资	4.21亿美元	Magnetar继续领投，NVIDIA参与融资，开启产业绑定
2023年12月	二次出售	6.42亿美元	旧股转让，估值大幅抬升
2024年5月	C轮	11亿美元	Coatue领投，IPO前最后一轮私募融资，估值约190亿美元
2024年11月	二次出售	6.5亿美元	Jane Street、Fidelity、Macquarie 等首次入股，预热 IPO
			IPO和IPO后股权融资（合计约45亿美元）
2025年3月	IPO	15亿美元	40美元/股登陆纳斯达克；NVIDIA 锚定2.5亿美元，OpenAI 同步认购3.5亿美元
2026年1月	NVIDIA增资	20亿美元	87.2美元/股私募增资，配套至2030年5GW+ AI Factory联合建设计划
2026年4月	Jane Street增资	10亿美元	109美元/股，配套60亿美元算力使用合约，金融垂直行业的标杆性背书
			高息债与可转换票据（合计约112.5亿美元）
2025年5月	高级票据	20 亿美元	票息9.250%，2030年到期，IPO后首发美元债，定价反映非投资级风险溢价
2025年7月	高级票据	17.5亿美元	票息9.000%，2031年到期，超额认购，规模由15亿美元上调至17.5亿美元
2025年12月	可转换高级票据	22.5亿美元	票息1.75%，2031年到期，转股溢价显著降低名义资金成本
2026年4月	高级票据	17.5亿美元	票息9.750%，2031年到期，配合Meta等客户合约扩容及基础设施建设需求
2026年4月	可转换高级票据	35 亿美元	票息1.75%，2032年到期，初始转股价约119.6美元，为公司迄今最大单笔可转债
			GPU 抵押融资DDTL系列（合计约 209 亿美元）
2023年8月	DDTL 1.0	23 亿美元	Blackstone、Magnetar领投，行业首创以H100 GPU为抵押的债务融资，定价约SOFR+11%
2024年5月	DDTL 2.0	75 亿美元	Blackstone再度领投，14家投资人参与；当时最大规模私募信贷之一
2025年7月	DDTL 3.0	26 亿美元	Morgan Stanley、MUFG牵头，定价SOFR+4%，专项用于交付OpenAI合约
2026年3月	DDTL 4.0	85 亿美元	获Moody's A3/DBRS A(low)投资级评级，定价SOFR+2.25%，行业首次GPU抵押融资获投资级评级

数据来源：CoreWeave，Businesswire，路透社，PR Newswire，Investopedia，东吴证券研究所

3.2. 财务分析：收入加速兑现，利润与现金流仍处重资本扩张期

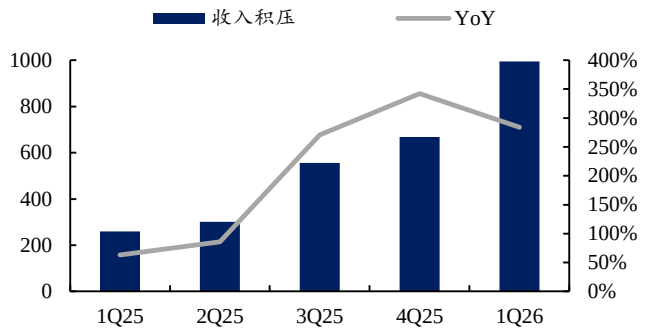
收入端高速增长，合约储备强化未来能见度。CoreWeave 收入规模在过去两年快速放大，2024 年收入 19.15 亿美元，同比增长约 700%；2025 年收入进一步提升至 51.31 亿美元，同比增长 168%；1Q26 实现收入 20.78 亿美元，同比增长 112%，单季收入已接近 2024 年全年水平。公司维持 2026 年全年 120-130 亿美元收入指引，并将 2026 年末年化收入规模指引由 170-190 亿美元上调至 180-190 亿美元，反映已签约订单正随 GPU 集群和数据中心上线逐步转化为收入。更重要的是，截至 1Q26 末，公司 Backlog 达到 994 亿美元，同比增长 284%、环比增长 50%，剩余履约义务达到 988 亿美元。对 CoreWeave 而言，当前收入增长并非单纯来自现货算力需求，而是由多年期大客户合约驱动。

图25：CoreWeave 收入（百万美元）与增速（%）



数据来源：公司财报，东吴证券研究所

图26：CoreWeave Backlog（亿美元）和增速（%）

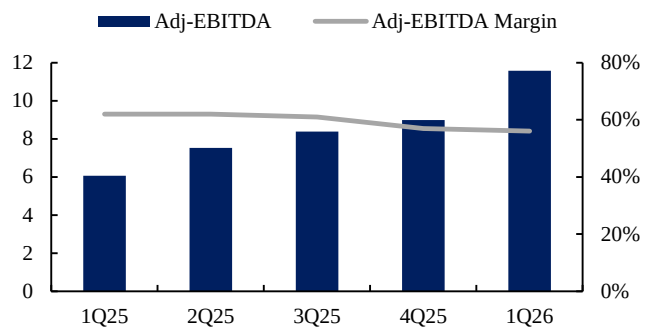


数据来源：公司财报，东吴证券研究所

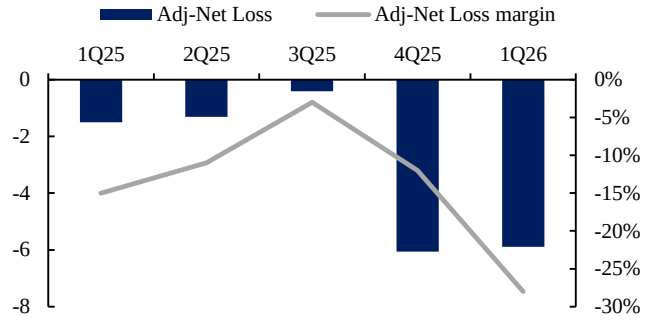
利润表呈现“EBITDA 强、净利润弱”的典型重资产特征。1Q26，公司经调整 EBITDA 达到约 12 亿美元，对应 Margin 为 56%，说明已投入运营的算力资产具备较强经营盈利能力；但同期公司净亏损 7.40 亿美元，净亏损率约 36%。这种分化主要来自折旧摊销和利息费用压力。1Q26，公司利息费用从去年同期的 2.64 亿美元提升至 5.36 亿美元；同时，大量 GPU、服务器和数据中心投入带来高折旧摊销，使净利润明显承压。换言之，公司在 EBITDA 层面已经验证算力资产盈利能力，但高资本开支和高杠杆扩张仍在利润表中形成显著拖累。

图27：CoreWeave adj-EBITDA（亿美元）和 Margin

图28：CoreWeave adj-Net Loss（亿美元）和 Margin



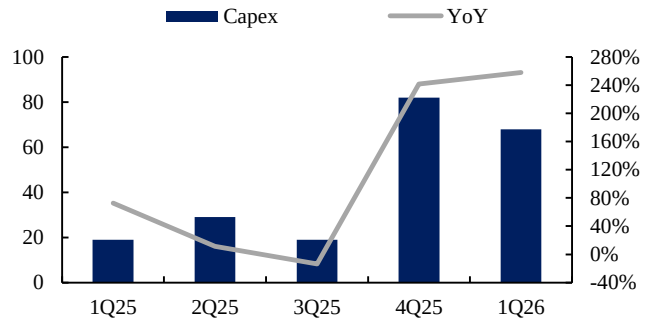
数据来源：公司财报，东吴证券研究所



数据来源：公司财报，东吴证券研究所

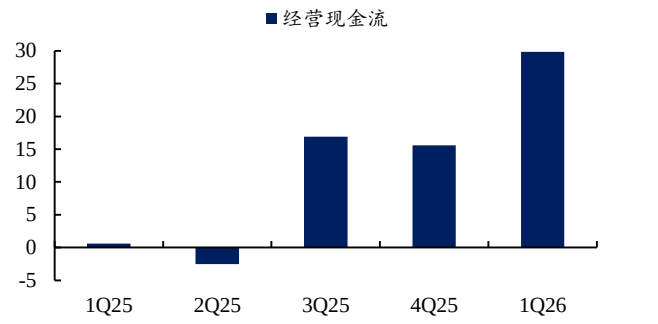
现金流好于净利润，但自由现金流仍显著为负。1Q26，公司经营现金流明显好于同期净亏损，主要受客户预付款、递延收入增加和营运资本变化影响，说明长期合约可以在一定程度上为扩张提供现金支持。但公司仍处于资本开支高峰期，1Q26 资本开支达到约 68 亿美元，全年 2026 年资本开支指引为 310-350 亿美元，主要用于 NVIDIA 芯片采购、数据中心和能源基础设施建设。

图29：CoreWeave Capex（亿美元）和增速（%）



数据来源：公司财报，东吴证券研究所

图30：CoreWeave 经营现金流（亿美元）



数据来源：公司财报，东吴证券研究所

3.3. 核心优势：供应绑定与架构领先，共同构筑 AI 云扩张壁垒

3.3.1. 优势一：与 NVIDIA 深度绑定，从供应链合作升级为 AI 生态共同体

与 NVIDIA 的深度绑定，是 CoreWeave 相较其他 NeoCloud 厂商最核心的差异化优势之一。这种关系并不是简单的“GPU 采购-云服务销售”，而是逐步演化为涵盖资本、硬件供应、技术协同、产品路线图和生态背书的多层次合作。NVIDIA 早在 2023 年即参与 CoreWeave 融资，并在公司上市前后持续持股；2026 年 1 月，NVIDIA 进一步以每股 87.20 美元价格投资 20 亿美元，同时宣布双方将扩大合作，共同推进超过 5GW 的 AI Factory 建设。根据双方公告，CoreWeave 还计划将其云平台扩展至 NVIDIA Rubin 平台、

Vera CPU、BlueField 等下一代基础设施。由此看，CoreWeave 已经不只是 NVIDIA GPU 的重要采购方，而是 NVIDIA 在 Hyperscaler 之外推动 AI Factory 落地的重要合作伙伴。

这一绑定的本质，是 NVIDIA 在 GPU 供给紧张和云厂商加速自研芯片背景下，对核心算力分发渠道的战略选择。AWS、Google、Microsoft 等超大云厂商均在推进自研 AI 芯片，以降低对 NVIDIA GPU 的依赖。相比之下，CoreWeave 并不自研 AI 芯片，而是围绕 NVIDIA GPU 构建 AI 云平台，因此更符合 NVIDIA 扩大 GPU 生态、强化客户触达和加速新平台商业化的需求。对 NVIDIA 而言，CoreWeave 既是 GPU 大客户，也是其面向 AI Lab、企业客户和开发者生态输出完整 AI Factory 能力的重要载体。对 CoreWeave 而言，NVIDIA 则提供了 GPU 供给、技术路线图、性能优化和客户信任背书。

硬件优先供给，是这种绑定带给 CoreWeave 最直接的竞争优势。在 AI 云行业，能否提前获得新一代 GPU，往往决定厂商能否在供给最紧、价格最高、客户需求最强的窗口期抢占订单。CoreWeave 过去多次体现出领先部署 NVIDIA 新平台的能力，公司较早商业化部署 H100、H200、GB200 等产品，并计划将平台扩展至 NVIDIA Rubin 架构。对 AI Lab 和大型企业客户而言，这种领先部署能力意味着更早获得新一代算力、更高训练与推理效率，以及更低排队和迁移成本。因此，CoreWeave 与 NVIDIA 的深度绑定不仅提升了供给确定性，也强化了其在大客户长约谈判中的议价能力。

技术协同和生态背书进一步降低了客户尽调成本。CoreWeave 并不是简单采购 GPU 后出租，而是围绕 NVIDIA GPU 构建适配 AI 训练和推理负载的软件栈、集群调度能力和高性能网络架构，并多次参与 MLPerf 等行业基准测试。对于 AI 云客户而言，选择供应商不只是比较 GPU 数量和单价，更重要的是判断其能否稳定交付大规模集群、持续优化性能，并及时跟进下一代硬件平台。NVIDIA 的资本参与、硬件支持和技术协同，实际上为 CoreWeave 提供了信用增强，使其更容易从单纯 GPU 云服务商升级为核心 AI 基础设施合作伙伴。

但这一优势也伴随显著的单一依赖风险。CoreWeave 的核心竞争力高度围绕 NVIDIA 生态构建，其硬件供给、软件栈、网络架构、性能优化和客户认知均与 NVIDIA 深度绑定。如果未来其他厂商芯片在性能或成本曲线上取得突破，或 NVIDIA 因渠道策略、Hyperscaler 关系、供给紧张等原因调整对 CoreWeave 的支持力度，CoreWeave 的相对优势可能被削弱。

图31：CoreWeave 与 NVIDIA 深度绑定：资本支持、平台协同与生态背书共同强化扩张能力

绑定维度	NVIDIA提供	CoreWeave获得的优势	对CoreWeave商业模式的影响
资本支持	NVIDIA于2026年1月以87.20美元/股向CoreWeave投资20亿美元	增强资产负债表与市场信心	支撑重资本扩张和AI Factory建设
平台协同	双方计划推进Rubin平台、Vera CPU、BlueField等下一代NVIDIA基础设施合作；CoreWeave曾较早接入H100、H200、GB200/GB300 NVL72等平台	更快跟进NVIDIA新一代GPU平台	提升高端算力供给能力和客户黏性
技术验证	双方计划测试和验证CoreWeave的AI-native软件及参考架构，包括SUNK、Mission Control等	提升软件栈、集群管理和平台稳定性的可信度	降低客户对大规模集群交付能力的尽调成本
生态背书	双方共同推进超过5GW AI Factory建设，并强化基础设施、软件和平台层面的合作	增强AI Lab和企业客户对CoreWeave长期交付能力的信任	强化多年期大客户合约获取能力

数据来源：NVIDIA，东吴证券研究所

图32：Hyperscaler 自研芯片趋势下，CoreWeave 成为 NVIDIA 重要算力分发渠道

云厂商类型	代表公司	芯片策略	与NVIDIA关系	对NVIDIA意义
Hyperscaler	AWS/Google/Microsoft	自研AI芯片与NVIDIA GPU并行	核心客户，但同时降低对NVIDIA的依赖	需应对客户自研芯片替代风险
NeoCloud核心伙伴	CoreWeave/Nebius	不自研AI芯片，围绕NVIDIA GPU构建AI云平台	资本、平台与生态深度合作	放大NVIDIA GPU生态，承接Hyperscaler之外的AI算力需求
其他GPU云/AI云	Lambda/Crusoe	以NVIDIA GPU为主要算力底座，但公开披露的绑定深度不同	NVIDIA云合作伙伴与渠道补充	扩大NVIDIA GPU触达范围，分散下游需求来源

数据来源：英伟达，Lambda，Crusoe，东吴证券研究所

3.3.2. 优势二：AI-Native 基础设施架构，从“适配 AI”到“为 AI 而建”

CoreWeave 的基础设施优势，来自其 AI-Native 架构起点。传统超大规模云厂商的基础设施最初主要服务于通用计算、虚拟机、共享存储和多租户工作负载，后续再通过 GPU 实例、专用网络和高性能存储承接 AI 训练与推理需求。相比之下，CoreWeave 从业务转型早期即围绕 GPU 加速计算、AI 训练和高性能计算搭建平台，因此其架构更接近“为 AI 而建”，而不是在通用云架构上追加 GPU 能力。

这种差异体现在资源调度、网络、存储和运维四个层面。CoreWeave 以 Kubernetes 和裸金属 GPU 集群为核心，尽量减少传统虚拟化层对高性能计算的性能损耗；在大规模集群中采用 NVIDIA 高性能网络、NVLink/NVL72 机架级架构和液冷方案，支持低延

迟、高带宽的分布式训练；同时引入 DPU 和网络/存储 offload 能力，将部分网络、存储和安全任务从 CPU/GPU 侧剥离；并通过 Mission Control 等自研运营工具监控节点健康、管理节点生命周期，降低大规模训练中的中断和 GPU 小时浪费。NVIDIA 与 CoreWeave 的合作公告也提到，双方计划测试和验证 CoreWeave 的 AI-native 软件和参考架构，包括 SUNK 和 CoreWeave Mission Control，并推进与 NVIDIA 云伙伴和企业客户参考架构的互操作。

架构优势最终体现为两类结果：性能更高、部署更快。在 MLPerf Inference v5.0 中，CoreWeave 使用 NVIDIA GB200 Grace Blackwell Superchips，在 Llama 3.1 405B 模型上实现 800 tokens/s 的推理性能，较 H200 实现约 2.86 倍单 GPU 性能提升；其 H200 实例在 Llama 2 70B 模型上实现约 33,000 tokens/s，较 H100 提升约 40%。训练侧，CoreWeave、NVIDIA 和 IBM 联合提交 MLPerf Training v5.0 结果，使用 2,496 颗 NVIDIA GB200 GPU，在 Llama 3.1 405B 训练任务中用 27.3 分钟完成训练提交，CoreWeave 称这是当时 MLPerf 中规模最大的 GB200 NVL72 集群提交。公开基准的意义在于，它证明 CoreWeave 并非单纯堆叠 GPU，而是具备将 GPU、网络、存储和调度系统整合为高效集群的工程能力。

但这一壁垒并非技术独占，而更接近“工程债务差”带来的阶段性领先。Kubernetes、NVLink/InfiniBand、DPU offload、液冷、高性能存储和集群监控等关键组件并非 CoreWeave 独有，Hyperscaler 在资金、人才和客户资源上仍具备追赶能力。CoreWeave 的真实优势在于历史包袱较轻：其平台从一开始围绕 GPU 密集型负载设计，不需要在传统通用云架构、多租户虚拟化兼容、老旧客户工作负载和 AI 专用集群之间做过多折中。因此，后续跟踪关键在于每一代 NVIDIA 新 GPU 发布后，CoreWeave 与最快 Hyperscaler、其他 NeoCloud 之间的实际商业化可用时间差，以及 MLPerf 等公开基准中能否持续保持领先。

图33：CoreWeave AI 算力平台

图34：CoreWeave 在 MLPerf 基准测试中领先

数据来源：公司官网，东吴证券研究所

Platform	Benchmark model	CoreWeave Throughput	Performance Improvement
NVIDIA H200 GPUs	Llama 2 70B (fp8)	33K tokens/s	40% faster than H100 GPUs
NVIDIA Grace Blackwell Superchips	Llama 3.1 405B (fp4)	800 tokens/s	2.86X over H200 GPUs

数据来源：公司官网，东吴证券研究所

4. 投资建议

AI 算力供需错配仍是本轮 NeoCloud 行业景气的核心来源。需求端，大模型训练、多模态模型迭代、Agent 工作流和企业级 AI 算力供需错配仍是本轮 NeoCloud 行业景气的核心来源；供给端，GPU、数据中心、电力、液冷、网络和集群交付能力均存在较长建设周期，可交付算力供给仍具备刚性约束。因此，NeoCloud 并非简单的 GPU 租赁模式，而是 AI 时代以高资本开支、长期客户合同和持续交付能力为核心的新型基础设施运营模式。

从盈利能力来看，NeoCloud 厂商的盈利能力可概括为：单位算力收入 – 单位算力 TCO – 资本成本。收入端，在高端 GPU 供给紧张、客户倾向通过多年长约锁定产能的背景下，已上线 GPU 集群能够在高利用率和高定价环境下释放较强收入弹性；成本端，GPU 服务器、数据中心、电力、网络、存储、液冷、运维、折旧摊销和融资成本共同决定全生命周期 TCO；资本端，由于行业处于产能快速扩张阶段，资本开支和融资成本会阶段性压制净利润和自由现金流。因此，短期内头部 NeoCloud 厂商可能呈现“收入高增、Adjusted EBITDA 改善，但净利润和自由现金流仍承压”的财务特征，后续盈利拐点取决于产能爬坡、利用率提升、客户结构优化、融资成本下降以及平台化收入占比提升。

我们预计，2026-2028 年 NeoCloud 行业仍将维持较高增长，但不同厂商之间的盈利分化会进一步扩大。具备三类能力的公司更有望在本轮周期中胜出：一是 GPU 和新一代 AI 服务器资源获取能力，决定其能否在供给紧张阶段优先获得订单；二是电力、数据中心和 AI Factory 交付能力，决定其订单能否转化为真实收入；三是融资能力和客户长约能力，决定其能否支撑重资本扩张并平滑现金流压力。相反，仅依赖短期 GPU 租赁价差、缺乏长期客户合同和资本支持的厂商，在供给缓解或租赁价格回落后，盈利和估值弹性可能明显收缩。

基于上述分析，我们建议重点关注：1）海外 NeoCloud 龙头：CoreWeave、Nebius 等；**2）海外算力/存储产业链：**NVIDIA、Broadcom、Marvell、Micron、SK Hynix、TSMC、Arista Networks、Vertiv、Super Micro Computer、Dell Technologies、Lenovo Group 等；**3）国内算力基础设施与运营商生态：**盛视科技、平治信息、华策影视、协创数据、宏景科技等。

5. 风险提示

1） AI 算力需求不及预期风险：当前 NeoCloud 行业高景气主要建立在大模型训练持续迭代、推理调用快速放量以及企业 AI 应用加速落地的基础上，若大模型能力提升放缓、AI 应用商业化进展不及预期，或客户资本开支出现阶段性收缩，可能导致 GPU 集群需求、算力租赁价格和长约订单增长低于预期。

2） 资本开支过高和融资成本上升风险：NeoCloud 属于典型重资产运营模式，GPU 服务器、数据中心、电力、网络和液冷等环节均需要持续高强度资本开支投入，若公司融资环境持续恶化、债务成本上升，或资本市场风险偏好下降，可能影响厂商持续扩产节奏，并对净利润、自由现金流和资产负债表造成压力。

3） 产能交付不及预期风险：AI 算力供给并非仅取决于 GPU 采购，还依赖数据中心建设、电力接入、液冷改造、网络调试和集群运维能力，若电力并网、数据中心交付、服务器部署或客户验收进度低于预期，可能导致订单无法按期转化为收入，并影响公司经营杠杆释放。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn