人工智能行业专题（19）

从Nebius看新兴云与开源模型Token优化

行业研究 · 海外市场专题

互联网 · 互联网II

投资评级：优于大市（维持）

证券分析师：张伦可
0755-81982651
zhanglunke@guosen.com.cn
S0980521120004

证券分析师：陈淑媛
021-60375431
chenshuyuan@guosen.com.cn
S0980524030003

核心结论

- Nebius 是全栈自研 AI 原生云代表厂商，工程基因是核心壁垒。公司脱胎于 Yandex，继承 20 年分布式基础设施能力，形成“裸金属算力→多租户云→Token Factory 托管推理（25年底落地）→智能代理层（规划）”的四层产品架构。**Nebius代表的不仅GPU租赁商，而是一种"全栈垂直整合"的新兴AI云范式——通过硬件自研、编排自控、模型自优化的三层能力叠加，将开源模型的推理成本最高砍掉70%，从而推动AI推理从"闭源模型依赖"向"开源模型替代"的结构性迁移。**

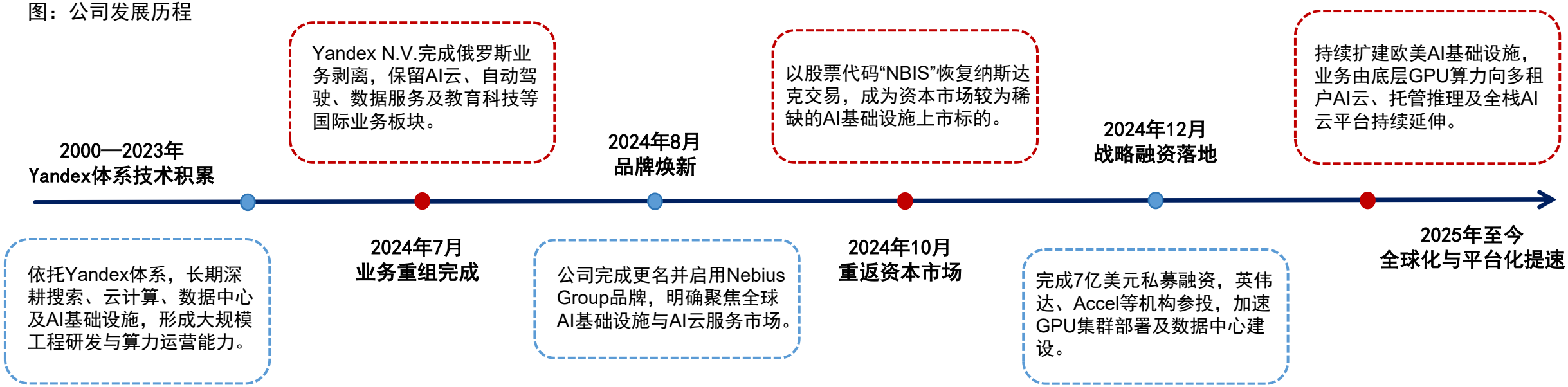
- **底层：自研高密度 ODM 机架与工程基因构建成本护城河。**比较AWS，AWS微调的TCO(总拥有成本)比Nebius贵43%，推理TCO(总拥有成本)贵112%。新兴云具备裸金属去虚拟化、InfiniBand直连、无通用云搭售等成本优势。特殊的是，根据SemiAnalysis和官网，Nebius是唯一一家使用定制ODM机箱的新兴云厂商，预计节省10-15%OEM溢价。根据官方白皮书，自研液冷方案将服务器功耗降低约20%。除此以外，Nebius拥有GPU 弹性打包 + 双栈调度回收侧优化，相比较其他新兴云能够回收约5-15%闲置算力。
- **中层：Token Factory通过收购Eigen AI实现模型层全栈优化，帮助提升单GPU收入，客户部署成本下降。**Eigen 推理优化让单卡 Token 吞吐提升 2-3 倍，从而把单卡Token产出提升至2-3倍，与此同时，Token价格基本和原模型一致。所以，同样的硬件成本Nebius能卖出更多收入。除此以外，对于客户来说，Nebius Token Factory的Token优化能力可有效降低客户AI推理部署的全链路成本。传统部署路径下，客户若选择裸金属自部署，需要用更多的卡完成同等每日Token量需求。
- **Token Factory案例一定程度证实了"开源替代闭源模型"的经济性。**Revolut因成本失控从顶尖闭源模型全面迁移至Token Factory上的Kimi 2.5；Cosine因银行/国防客户的数据不出墙要求，通过Papyrax后训练框架将Llama 3.3 70B调教至OpenAI O3水平。非Nebius案例同样印证趋势：Shopify将Qwen3-32B微调后推理速度2.2倍、成本下降68%；Coinbase形成"开源模型GLM-5.2、Kimi2.7承接常规调用、闭源模型处理复杂任务"的分层架构，AI支出接近减半。

目录

Nebius：欧洲为主的算力租赁厂，工程基因是是底层优势

- Nebius 是一家总部位于荷兰阿姆斯特丹的 AI 原生云基础设施公司，2024 年完成重组后定位为"从硅片到 API"全栈自建的 AI 云。
- **Nebius 拥有接近20年的互联网基础。**它脱胎于纳斯达克上市科技巨头 Yandex（前身为"俄罗斯谷歌+优步），继承了上千名顶尖分布式软件工程师与 20 年的大厂全栈内功；在其他 AI 云在当靠出租 GPU 赚差价时，Nebius 已经凭借超大规模调度、自动化故障自愈以及深度的网络软件拓扑能力，将昂贵的万卡集群打造成了无缝续传、压榨出极限性能的“精装自动化 AI 工厂”。

图：公司发展历程



资料来源：公司公告，国信证券经济研究所整理

Nebius：管理层经验深厚，延续Yandex技术积淀

- Nebius由Yandex联合创始人Arkady Volozh带领，**管理层长期深耕搜索、云计算、分布式系统及数据中心建设，具备大型科技平台从技术研发到商业化运营的完整经验。**
- **根据公司披露，Nebius承接Yandex核心技术团队，公司披露拥有超过1000名具备长期协作经验的AI工程师，覆盖AI基础设施及相关技术研发。**成熟的工程团队与软硬件一体化能力，有望缩短AI基础设施建设周期，并支撑公司由底层算力向多租户云、托管推理及全栈AI云平台持续延伸。

图：Nebius核心管理层情况

姓名	职位	主要职责及背景
Arkady Volozh	创始人、首席执行官兼董事	Yandex联合创始人，长期负责大型科技平台建设；主导Nebius战略规划、全球AI基础设施布局及重大资本决策。
Ophir Nave	首席运营官兼董事	负责集团日常运营、组织管理、资源配置及公司战略落地，与CEO共同参与核心经营决策。
Roman Chernin	首席商务官	负责商业战略、客户拓展及合作伙伴关系建设，推动AI算力与云服务的商业化。
Andrey Korolenko	首席基础设施与产品官	负责数据中心、GPU集群、云基础设施及产品体系建设，推动软硬件一体化AI云平台落地。
Dado Alonso	首席财务官	负责财务管理、融资、资本配置、预算及投资者关系，为数据中心扩张和全球化投入提供支持。
Marc Boroditsky	首席营收官	负责全球销售体系、收入增长及大型客户拓展，推动基础设施资源向订单和收入转化。
Danila Shtan	首席技术官	负责核心技术架构与研发体系，重点覆盖AI云平台、计算基础设施及底层技术能力建设。

资料来源：公司公告，国信证券经济研究所整理

Nebius：从基础设施租赁到Token Factory，软件附加值增加

国信证券 GUOSEN SECURITIES

- **Nebius 的产品战略围绕"四层架构"展开，核心逻辑是：越往上层走，可服务的客户群数量呈十倍增长、差异化以及软件附加值越高。**①裸金属基础设施（按兆瓦计费，服务微软/Meta 级别的超大客户）；②多租户 AI 云（按 GPU 小时计费，面向数百至数千家研究团队）；③Token Factory 托管推理平台（按 Token 计费，支持 60+ 开源模型）；④以及规划中的智慧代理层（按任务计费，让开发者只需提需求、平台自动决策模型调用路径）。

图：Nebius 四层产品架构 — 从物理层到代理层

资料来源：Nebius Youtube访谈，国信证券经济研究所整理

NeoCloud vs 传统云：从资源切分转向算力聚合

国信证券 GUOSEN SECURITIES

- **传统云的很多能力，在AI场景里并不是最核心的，甚至可能拖累 AI 集群性能。传统云（如 AWS、Azure）是为“切分资源”而生；而 AI 算力集群需要的是“聚合资源”。**传统云时代，Amazon 等公司靠虚拟化、安全隔离、自研 NIC、自研 SSD建立优势；但在 AI GPU 集群里，客户往往不是租一颗 GPU、用几个小时就还回去，而是整租、长周期租赁。
  - **有损的虚拟化与安全隔离：**传统云的核心商业模式是把一台物理服务器切分成 100 个虚拟机（VM）卖给不同客户。NeoCloud直接提供裸金属服务器，去除了传统云复杂的虚拟化软件层（Hypervisor）。在 GPU 虚拟化实例里，hypervisor拦截 GPU command、内存请求、网络包带来的性能损耗率，通常在 10-15% 区间，AI 负载下会放大到 20-25%。
  - **RoCE vs InfiniBand：架构分歧。**AI 训练的核心瓶颈不是计算，而是通信。大规模训练采用标准是 NVIDIA Quantum-2 InfiniBand，流控机制基于信用的（credit-based）——发送端有明确的信用额度，不会超发。商业本质英伟达全家桶，极致性能，但存在生态锁死问题。大多数传统云厂商采用RoCE或变种，在标准以太网上"嫁接"RDMA 能力。

图：传统云架构VS AI优先计算架构

维度	传统云计算架构（如 AWS、Azure）	AI 优先计算架构
计算模式	虚拟化 / 多租户 (Multi-tenancy)一台物理机切成几十个虚拟机 (VM)，卖给不同的零散客户。	裸金属 / 单租户独占 (Single-tenancy)客户直接包下整个机柜或整厂，无虚拟化损耗，追求极致原始性能。
安全隔离	高墙隔离在网络、计算、存储层层设防，牺牲部分速度换取绝对的多租户安全。	物理隔离整个集群只跑一个大模型任务，去除软件层审计，网络直通显存。
网络通信	通用网络（如 SRD/EFA）优化高并发、防丢包，适合成千上万个不相关的 App 共享带宽。	AI 专用网络（如 InfiniBand / RoCEv2）超大带宽（400G/800G）、零丢包、超低延迟、支持 GPUDirect RDMA。
存储特征	高随机读写（高 IOPS）自研 SSD 固件和分布式块存储（如 EBS），适合频繁读写的数据库。	高并发顺序吞吐（Throughput）部署分布式并行文件系统（如 Weka），专为几万张卡同时写入 Checkpoint（快照）优化。

资料来源：Nebius architecture YouTube，国信证券经济研究所整理

Nebius租赁成本低，推理效率更高

- **根据Nebius官网，比较AWS，AWS微调的TCO（总拥有成本）比Nebius贵43%，推理TCO（总拥有成本）贵112%。**平均问题解决时间12分钟；3000块GPU能稳定运行近57小时。CoreWeave 在官网同样表示，其基础设施可帮助客户实现高达 20% 的 MFU 提升和96% 的吞吐量，我们总结核心原因包含：
  - 前文所述，裸金属去 Hypervisor（虚拟化软件层）、IB 网络 + Slurm 调度；
  - **无通用云搭售。**在AWS租用带GPU 的虚拟机= GPU 硬件成本 + 数据中心成本+ S3 研发摊销 + VPC 研发摊销 + IAM 摊销 等200+种服务；对比来看，Nebius 的内部成本：GPU 硬件 + 数据中心 + AI 云平台，没有为 200 种服务的开销要分摊。
  - IDC优势：NBIS芬兰75MW 自营（310MW在建），能源成本低。
  - **Token推理侧优化。**

表：以H100举例租赁价格差异：Nebius成本远低于其他平台

Provider	Instance (H100 80GB)	On-Demand ($/GPU-hr)	vs AWS
Nebius	NVIDIA HGX H100 (8-GPU)	$2.95	57.1% cheaper
Nebius (Preemptible)	NVIDIA HGX H100 (8-GPU)	$1.25	81.8% cheaper
Lambda Labs	H100 SXM (8x)	$3.99	42.0% cheaper
Crusoe	H100 80GB HGX	$3.90	43.3% cheaper
CoreWeave	HGX H100 (8-GPU)	$6.16	10.5% cheaper
AWS	p5.48xlarge ($55.04/hr÷ 8)	$6.88	Baseline
Azure	Standard_ND96isr_H100_v5	$12.29	78.6% more

数据来源：Opslyft，国信证券经济研究所整理

Nebius租赁成本低外，采购效率和灵活性优于传统云

- 除了成本优势外，根据Opslyft，Nebius的其他优势包含：
  - **H100 和 H200 的可用性。**热门区域的 AWS p5 容量仍然需要提前数周预订或加入企业折扣计划。Nebius H100、H200 和 Blackwell B200 SKU 可立即按需使用。
  - **采购流程更简便。**创始人只需一张公司信用卡，即可在几分钟内启动一个配备 8 个 GPU 的 Nebius 节点。而 AWS 的同类服务通常需要经过采购、EDP 协商以及长达数月的容量分配谈判。
  - **默认使用 InfiniBand。**Nebius HGX H100 节点出厂时即配备 InfiniBand，用于多节点训练。在 AWS 上，等效的 EFA 网络功能捆绑在特定的实例系列中，需要配置放置组。

图：Nebius官网对公司优势的表述

数据来源：公司官网，国信证券经济研究所整理

传统云受限于组织以及生态绑定

- **对于大模型公司来说，谁能更快把可用算力交付，谁就有价值。**
  - **传统云并非完全无法实现AI原生云的极致性能，其核心瓶颈源于底层架构的基因冲突。**传统云庞大的成本结构与僵化的产品生态也限制了其竞争力。传统云厂商背负着数百种通用云服务的研发与运维成本，这些隐性开支最终都会摊销到GPU租金中，使其在面对垂直AI云厂商时缺乏价格弹性。同时，传统云倾向于通过“全家桶”式的生态绑定来锁定客户。
  - 大云厂商纷纷自研AI芯片、降低对英伟达依赖的趋势。**英伟达主动将NeoCloud作为核心战略抓手**。①AI工厂设计与支持：根据官网，Nebius与英伟达扩展全栈 AI 云，涵盖合作设计资料获取、设计评审与验收流程、早期样品及系统软件支持、系统启动调测支持，以及定期的系统级合作伙伴业务与技术复盘；② 融资与债务背书：为NeoCloud厂商的债务融资提供担保，降低其融资成本；③ 直接股权投资：对CoreWeave、Lambda、Nebius等头部厂商均有战略持股，深度绑定利益。例如，NVIDIA 将向 Nebius 投资 20 亿美元。

图：Nebius拥有业内领先的人工客服工程师首次回复速度

数据来源：公司官网，国信证券经济研究所整理

Nebius的Token Factory模式：软硬件垂直整合

- **Token Factory 是 Nebius 于 2025 年 11 月推出的托管推理平台。**定位不是"模型超市"（像 AWS Bedrock 那样聚合 100+ 个模型），而是"模型工厂"——**把 60+ 个开源模型**（DeepSeek、Llama、Qwen 等）像产品一样精加工，**将每个模型的推理成本砍掉最高 70%**，再以透明的Token定价对外输出。
- **大厂云能做功能对等的产品，但较难Token Factory 的成本结构。**因为 Token Factory 的 70% 降本不是某一层的优化结果，而是 ODM 硬件、Eigen CUDA Kernel、推理调度三层垂直整合的叠加效应，除此以外，大厂硬件团队和推理团队分属不同 P&L 各自优化各自的目标。比如NBIS 拥有"优化吞吐 → 更少 GPU → 赚更多"的激励，AWS 销售 KPI 是卖 GPU-hour，缺少动力帮客户进行优化。

图：NeoCloud和大厂云能力层级

能力层级	Nebius	CoreWeave	Crusoe	Together AI	AWS
自有 AI 数据中心	√	√	√		√
裸金属 GPU 集群	√	√	√		√
自研 ODM 硬件设计	√				部分
多租户 AI 云平台	√	√	√		√
托管推理平台	√		√	√	√
Agent 平台层	√				√
硬件→Token 跨层优化	√				

数据来源：各公司官网，国信证券经济研究所整理

目录

Nebius优化硬件层：自研高密度 ODM 机架 + 液冷

- 根据SemiAnalysis，追踪每家GPU云服务商的总体拥有成本，Nebius是唯一一家使用定制ODM机箱的neocloud服务商。与此同时，还开发自己的专有固件。避免了会推高故障率和排障开销的继承性设计问题。这些设计共同帮助确保集群保持稳定，工作负载运行无意外中断维护快速且可预测。这意味着客户总体拥有成本降低，我们预计节省10－15% 。
- **液冷方面，两套冷却系统并存——空气自由冷却（Nebius 自研）**+ 闭环液冷（NVIDIA 设计）。Nebius 的服务器为高封装密度和可预测的功耗行为设计，能够让服务器在比现成硬件通常容忍的更宽温度范围内可靠运行。根据公司白皮书，配合创新的冷却布局，将服务器功耗相比标准 OEM 服务器降低约 20%。

图：Nebius 自研服务器主板

数据来源：公司官网，国信证券经济研究所整理

Nebius优化硬件层：自研高密度 ODM 机架 + 液冷

- 硬件优化能力主要来自：
- **硬件工程设计能力积累：**工程师大多从 Yandex 转移过来，15-20年拥有建hyperscaler级基础设施基础，包含能设计数据中心、主板、服务器、机架、连接器、液冷的硬件团队；
- **与英伟达深度合作：**Yandex曾是英伟达在欧洲最大的客户之一；
- **算力基本自主控制：**IREN（极重，纯自建）> Nebius（偏重，全自建各占75%） > CoreWeave（由轻转重，土地电力租赁为主转到局部收购)> Oracle（由重转轻，核心自建+边缘租赁）。

图：NeoCloud 背景以及数据中心建设方式

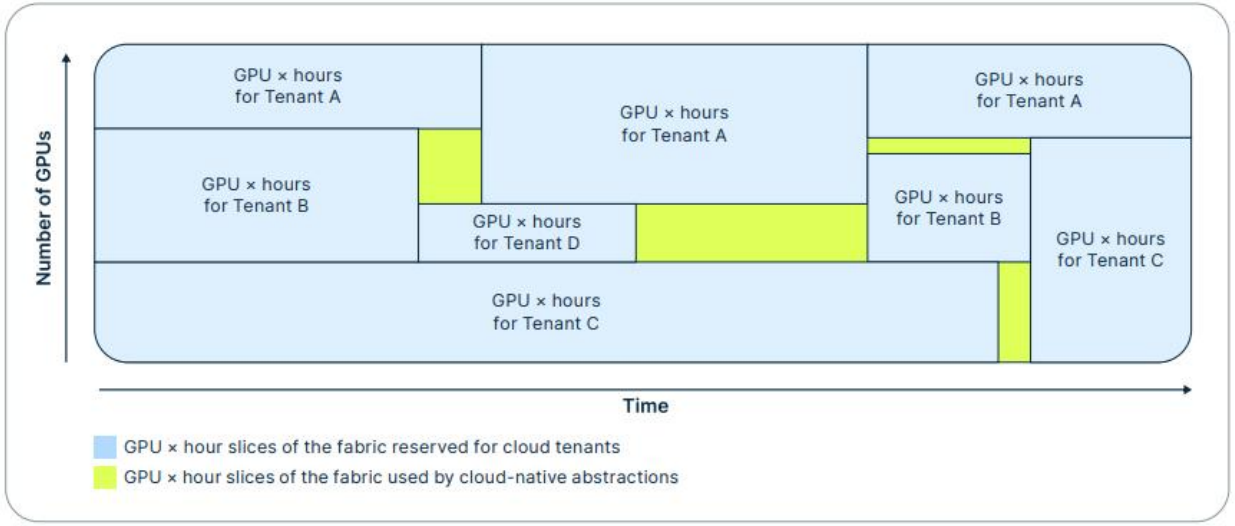
厂商	Oracle	CoreWeave	Nebius	IREN	DigitalOcean
厂商背景	1977年创立数据库软件巨头2010年收购Sun获硬件能力2016年推OCI转型AI云	2017年创立原名Atlantic Crypto以太坊矿工2019年转型AI云计算	1997年Yandex“俄版谷歌”2011年上市2024年拆分成立Nebius带1000+工程师	2018年创立澳大利亚悉尼原名Iris Energy比特币矿工2024年转型AI	2011年创立纽约，专注开发者轻量云主机
数据中心产权	自有+长期租赁	重度租赁→局部自有	混合，自有约75%	100%自有产权	100%全租赁
运营模式	高度垂直整合	轻资产快速扩张	全栈自建导向	完全垂直整合	纯托管运营
全球布局	全球均衡	北美	欧美+中东	北美	全球分散

数据来源：各公司官网，彭博社，国信证券经济研究所整理

Nebius优化编排层：GPU 直通弹性打包× 双栈编排

- **Nebius在GPU直通之上加了一层弹性打包，减少闲置资源浪费，公司表示这是与裸金属服务商的关键差异。**根据官方白皮书，纯裸金属则整租固定块、留下闲置碎片，还需为每个租户单独预留10-20%节点作为节点替换。Nebius在GPU直通之上加了一层 VM 弹性打包：K8s 调度器将用不完的 GPU 立刻分给别人，同时跨所有租户维护一个共享缓冲池，只需 4-5% 容量用于自动修复，空闲功耗降低。最终实现"裸金属的性能 × 云原生的利用率。
- K8s + SLURM 双栈编排让 Nebius用同一套 GPU 硬件同时服务互联网推理用户和 HPC 训练用户的 AI 云厂商，通过Soperator将这两个系统集成起来。两类用户在各自熟悉的工具上零学习成本接入，训练作业填推理低谷，集群利用率最大化。

图：Nebius 如何通过虚拟化捕获裸金属方案中闲置的小型未用容量



数据来源：公司官网，国信证券经济研究所整理

通过收购Eigen推动模型侧优化

- **Eigen AI 收购：Nebius Token Factory 推理优化层的核心拼图。** 根据公司公告，2026年5月，Nebius以约 6.43 亿美元收购硅谷 AI 推理优化公司 Eigen AI，三位出自 MIT HAN Lab 的联合创始人全部加入。
- **技术上，Eigen AI帮助Token Factory以低成本运营开源模型。** Eigen AI创始人拥有两项核心技术——AWQ 4-bit 量化（MLSys 2024 最佳论文，4-bit 模型生产部署的全球标准）和SpAtten 稀疏注意力（2020年以来HPCA最高引论文）。收购前两家已联合优化 DeepSeek、Llama、Qwen 等模型，在 Artificial Analysis 基准上跑出 911 tokens/秒的前列成绩。整合后，Token Factory 将覆盖 GPT-OSS、Gemma、Llama、DeepSeek、GLM、Kimi 等几乎所有主流开源模型。

图：Token Factory：开源模型从实验到生产的一站式平台

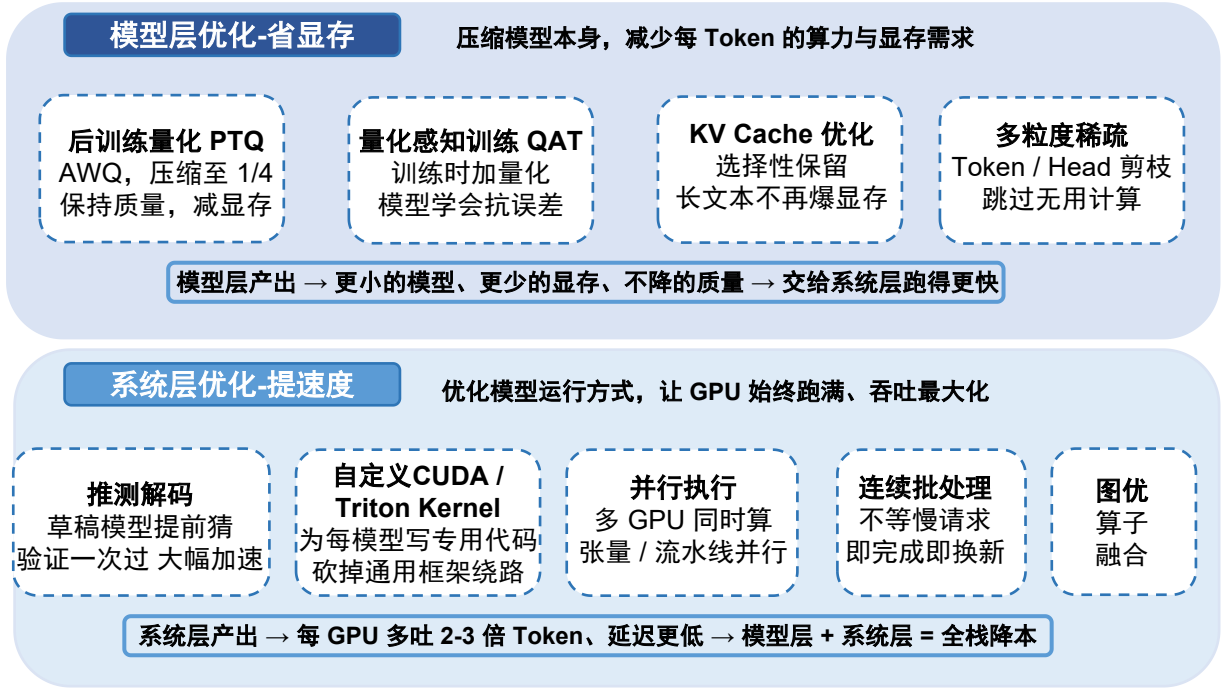
数据来源：公司官网，彭博社，国信证券经济研究所整理

通过收购Eigen推动模型侧优化

国信证券 GUOSEN SECURITIES

- **Nebius 和 Eigen AI 正在共同开发领先的开源模型优化版本。**优化让模型吐 Token 更快、GPU 利用率更高、大规模部署大模型的推理成本更低，这对现代 MoE（混合专家）和推理模型尤为关键，**路由策略、调度效率和显存管理往往直接决定了模型能否在生产环境中稳定、经济地运行。**

图：Eigen AI优化推理的方式



数据来源：公司官网，国信证券经济研究所整理

图：Eigen AI优化对吞吐提升定量计算

Eigen 优化层	做法	对TPS提升（实践值-估计）	对TPS提升（理论值）
AWQ 4-bit 量化	模型压到 1/4，显存释放后可放大 batch size，但受限于计算瓶颈而非纯显存瓶颈	~1.5-2x	~2-4x
SpAtten 稀疏注意力	在已优化 attention 计算的基础上，再剪枝 10－15% 冗余注意力头/FFN 路径	~1.1x	~1.1-1.15x
推测解码	vLLM 基线已有基础草稿验证，Eigen 做更激进的 draft 模型调参，但收益边际递减	~1.3-1.5x	~2-3x
自研 CUDA Kernel	绕过 vLLM 通用调度层 overhead，做算子融合和显存访问优化	~1.15-1.3x	~2-3x

数据来源：公司官网，国信证券经济研究所整理

其他收购：外延并购补齐搜索与推理能力，加速构建全栈AI云平台

国信证券 GUOSEN SECURITIES

- **2026年2月，Nebius拟全资收购Tavily，**将实时网络搜索、事实验证能力接入AI云平台及Token Factory，补齐企业级智能体在“推理—搜索—执行”环节的关键能力。
- **26年5月，公司进一步引入Clarifai核心团队，收购相关专利组合并获得推理及计算编排技术永久许可，强化系统级推理优化、资源调度及生产级部署能力。**继Nebius近期宣布收购Eigen AI之后，此次交易将进一步巩固Nebius Token Factory作为全栈推理平台的地位。
- 两项交易分别从智能体应用层与底层推理基础设施完善技术栈，推动Nebius由算力基础设施提供商向全栈AI云平台升级。

表：Nebius收购案例

时间	收购对象	性质	收购目的	说明
2026年2月	Tavily	完整股权收购	将实时搜索能力注入Nebius AI云平台，完善面向企业级AI代理的全栈基础设施； Tavily的代理搜索模块与Nebius Token Factory（高性能推理）深度耦合，补齐实时数据获取关键环节，形成“推理+检索”闭环。	领先的智能体搜索服务提供商Tavily，该公司服务于财富500强企业和顶尖人工智能公司。
2026年5月	Clarifai涵盖人工智能推理、计算编排及相关技术的专利组合， 吸纳研究人才	核心团队加入 专利组合收购	Clarifai（系统层面优化）+ Eigen AI（模型层面优化），推动构造端到端生产级推理基础设施； Clarifai 的部分工程师和研究人员也将加入 Nebius 的基础设施团队，为 Nebius 带来十多年推理优化和机器学习方面的专业知识。	Clarifai 团队在推理优化和机器学习方面拥有深厚经验。

资料来源：公司公告，国信证券经济研究所整理

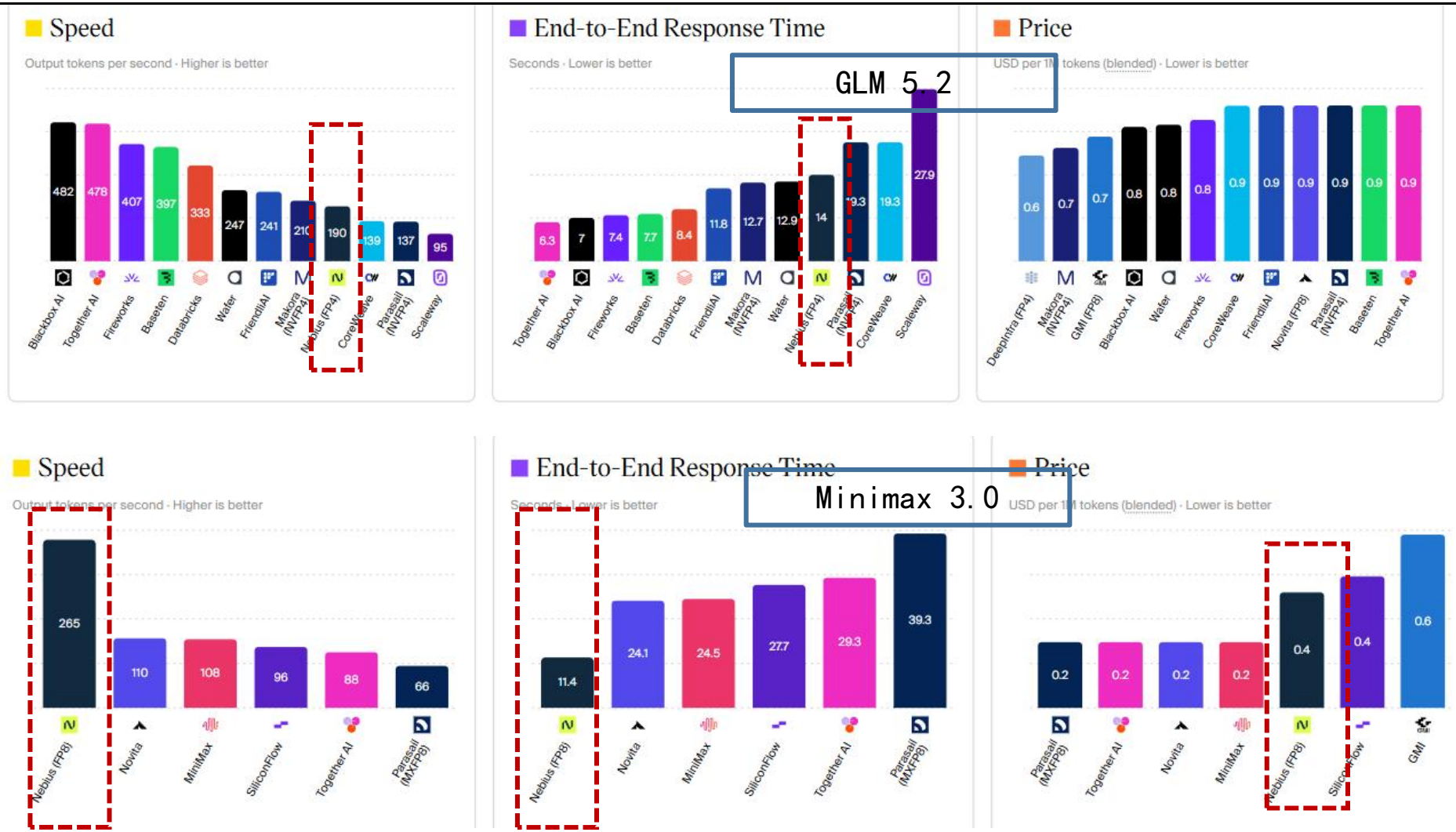
目录

Nebius模型优化让Token吞吐更快

国信证券 GUOSEN SECURITIES

- 实际案例来看，以Minimax 3举例，根据Artificial Analysis，Nebius在模型优化后输出Token速度可以达到原模型的2.6x；根据官网，价格来看，百万Token价格与原模型一致。

图：不同模型提供商对模型速度、反应时间、价格优化成都



资料来源：Artificial Analysis、国信证券经济研究所整理

Token优化帮助Nebius单GPU收入增加，帮助客户部署成本下降

国信证券 GUOSEN SECURITIES

- **技术放大了GPU产能，并且Token价格跟同行平齐。**Nebius靠Eigen优化栈把单卡Token产出提升至2-3倍，在Token价格不变情况下，同样的硬件成本能卖出更多收入。**以Prosus部署Llama-3.3-70B模型测算，Token Factory下的单GPU收入增长60%。**
- **Nebius Token Factory的Token优化能力可有效降低客户AI推理部署的全链路成本。**传统部署路径下，客户若选择裸金属自部署，需要用更多的卡完成同等每日Token量需求，并需承担GPU集群运维、推理内核调优的人力投入，以及10%-20%的冗余缓冲算力闲置损耗。

图：以Prosus 部署Llama-3.3-70B模型案例测算（模型价格 $0.13/$0.40）

	裸部署/裸金属 baseline	Nebius ToKen Factory
H100 TPS token/s/GPU	~2.1K	~9.45K（Eigen、4.5x）
固定200B token/day 需 GPU 数	~1,102 H100	~244 H100
若按裸金属 $2.5/GPU-h客户支出	$66,138/day	—
按 Token Factory $0.13/$0.40客户支出	—	取 $0.15/M（70% 缓存命中 + 80/20 输入输出） ≈ $23,400/day
对应Nebius单 GPU 日收入	$60/GPU/day（裸金属）	$96/GPU/day

数据来源：公司官网，彭博社，国信证券经济研究所整理

图：用户不同方案下产出的Token

方案	具体做法	单 GPU 吞吐	月产 Token	隐性成本
A：自己租 GPU，常规部署	租 1×H100，装 HF+FA2（FP16），默认批处理，不做额外优化	~2.1k token/s	~3.5B	运维精力分摊、20% 冗余 GPU、模型更新维护
B：自己租 + 手动优化	同 A，加 AWQ 4-bit 量化 + 手动批处理调度	~2.8k token/s	~4.2B	运维调优、模型切换测试
Token Factory	一键调用 API，Eigen 全栈优化	~4.2k token/s	~6.5B	

数据来源：公司官网，彭博社，国信证券经济研究所整理

以GLM5.2部署在B200为例，测算单卡TPS以及推理毛利率

国信证券 GUOSEN SECURITIES

表：以GLM5.2部署在B200为例，Nebius基础设施能够将推理毛利率同比+24pct；token优化能让推理毛利率同比+27pct

模型 芯片	GLM5.2原模型 B200	Nebius基础设施优化 B200	Nebius+Eigen全开 B200	备注
TPS = Batch*（输入+输出 Token）/ 总处理时间	4431.71	4431.71	10704.68	
总处理时间 = Decode时间 + Prefill时间	22.18	22.18	13.77	
Prefill时间 = max（Prefill计算时间 ，Prefill内存时间）	1.17	1.17	1.32	
Prefill计算时间 = （2 x batch x 单个请求输入token数量 × 激活参数量）÷计算能力/ 1000	1.17	1.17	1.32	自研CUDA Kernel(1.2x)：绕过vLLM通用调度层overhead，算子融合 SpAttn稀疏注意力(1.1x) ：激活参数减少10%
Prefill内存时间 = （2 × 模型参数量）÷ 显存带宽	0.19	0.19	0.12	AWQ 4-bit量化(1.5x)：权重读取量理论÷4，实际用1.75x
Decode时间 =单个请求输出token数量x（Decode计算时间 +Decode通信时间）	21.02	21.02	12.45	推测解码(Speculative Decoding, 1.30x)： 用轻量Draft模型一次预测N个token，主模型批量验证→一次性确认多个token。
Decode计算时间 = （2 × 激活参数量）/ 内存带宽	0.010	0.010	0.008	自研CUDA Kernel(1.2x)：算子融合+显存访问优化，减少内存读取overhead。
Decode通信时间 = EP通信时间 = Moe层数 x 2 x EP数据量/互联带宽 其中，EP数据量=batch x Topk x HiddenSize x 2	0.00026	0.00026	0.00033	自研CUDA Kernel(1.2x)：算子融合+显存访问优化，减少内存读取overhead。
batch	32	32	48	有效Batch:(×1.5)： AWQ 4-bit量化释放显存
激活参数量（B）	40	40	36	SpAttn稀疏注意力(1.1x)：通过结构化剪枝去除10-15%冗余注意力头/FFN中间层路径，
模型参数量（B）	744	744	744	
收入（美元/h）	8.62	9.91	23.93	
Token均价（美元/百万Token）	0.90	0.90	0.90	
单卡TPS（token/s/GPU）	4432	4432	10705	
利用率	60%	69%	69%	弹性打包 + 双栈调度回收侧优化，回收约15%闲置算力
芯片租赁价格（美元/小时）	6.00	4.55	4.55	使用定制ODM机箱，节省10-15%OEM溢价，自研液冷方案将服务器功耗降低约20%
推理毛利率	30%	54%	81%	

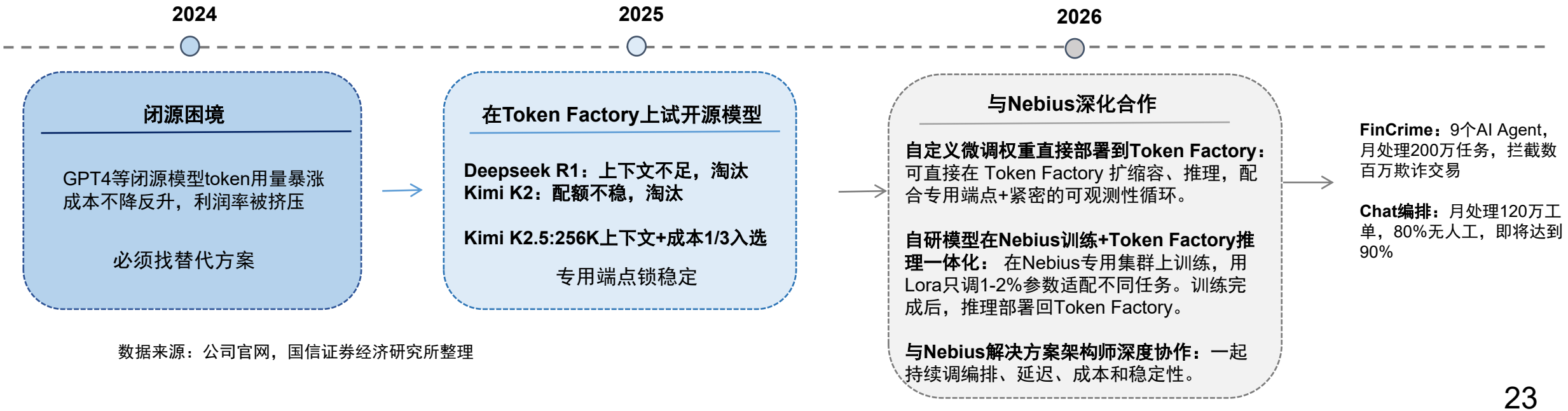
数据来源：公司官网，国信证券经济研究所整理

案例1：成本问题推动Revolut从顶尖闭源模型迁移至Kimi2.5

国信证券 GUOSEN SECURITIES

- **随着时间的推移，任务变得更加具体，token 消耗量暴涨，成本不降反升。业务规模扩大，挑战已从单纯的模型质量转向编排、评估、可观测性、延迟、可靠性与控制力。**
  - Revolut是总部位于伦敦的全球金融科技公司，提供基于应用程序的银行与金融服务，涵盖支付、储蓄、投资、外汇及对公账户等业务。三年内其用户规模从3000万增长至7000万，年度营收达31亿英镑、净利润7.9亿英镑。

图：Revolut用Token Factory托管开源模型带来效果案例



数据来源：公司官网，国信证券经济研究所整理

案例2：运用开源模型和Harness，Nebius让合规Agent成本大幅压缩

• Nebius 用自家 Agents Blueprint 框架构建了监管合规审计 Agent；闭源 GPT-5.5 跑规模化太贵，Nebius用Token Factory托管开源模型（DeepSeek-V4Pro→Nemotron-Ultra），把同一次合规审计从$657/30 分钟压到 $23/12.6 分钟，成本压缩至1/29。模型之间只需在 Token Factory 上一行配置切换、prompt 和检索索引原封不动贯穿四代。

图：Nebius用Token Factory托管开源模型带来效果案例

代次	配置	时间	成本	工单数	解决/暴露问题
1. 原型	GPT-5.5 + Pinecone	29.9 min	$469.99	5	任务能跑 / 时效性差
2. 落地	GPT-5.5 + Pinecone + Tavily	30.7 min	$657.19	12	准确性↑ / 成本暴涨
3. 优化	DeepSeek-V4-Pro on Token Factory + Pinecone + Tavily	53.9 min	$34.63	14	成本↓19x / 运行时变长
4. 生产	Nemotron-Ultra on Token Factory + Pinecone + Tavily + LangSmith + Snowglobe	12.6 min	$23.59	10	全维度最优

数据来源：公司官网，国信证券经济研究所整理

案例3：Cosine因数据合规转用开源模型

国信证券 GUOSEN SECURITIES

- **Token Factory 通过后训练把开源模型拉到接近闭源模型的能力。**Cosine 是做企业级 AI 编程 Agent 的公司，客户是银行/国防/核工业，代码不能出墙，闭源 API 的黑盒 + 数据出境在法律和操作上都不被允许。若想用客户数据微调模型行为，但闭源 API 不开放权重、不支持深度定制。
- Nebius 用 Token Factory 的后训练能力（Papyrax 框架 + SFT + RL）把 Llama 3.3 70B 和 GPT-OSS-120B 调教到 OpenAI o3 水平的 SWE 性能，让 Cosine 能在完全私有环境里交付企业级编程 Agent。

图：Nebius的合作伙伴

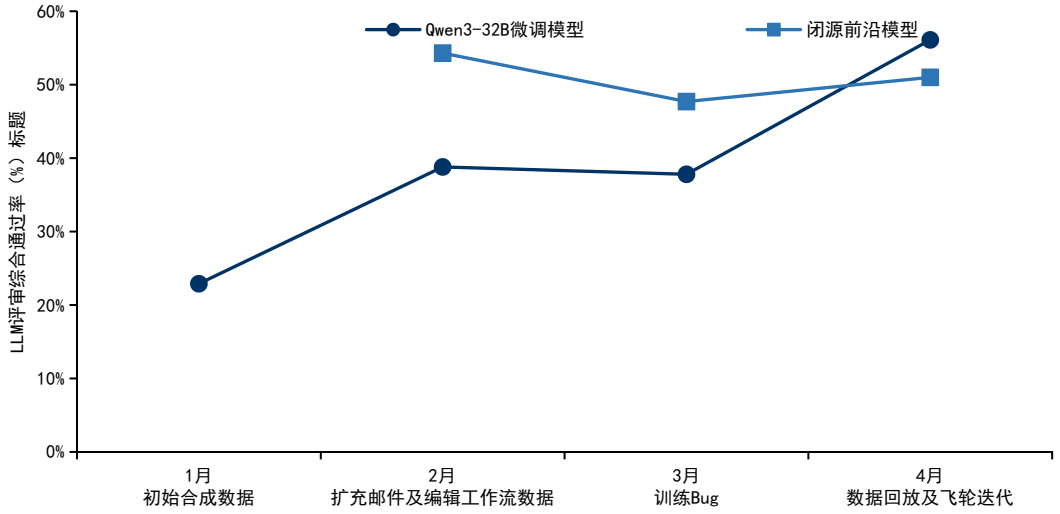
数据来源：公司官网，国信证券经济研究所整理

其他非Nebius案例1：Shopify闭源前沿模型转向Qwen3-32B

国信证券 GUOSEN SECURITIES

- **Shopify此前依赖闭源前沿模型，为AI电商助手Sidekick生成Shopify Flow自动化流程。**随着商户调用规模扩大，通用闭源模型在成本、响应速度及业务适配度方面逐渐受到限制，公司开始尝试以自有数据微调开源模型，提升核心AI功能的成本效率及技术可控性。
- **26年4月，Shopify基于数千条商户实际创建的自动化流程构建训练数据，将开源模型Qwen3-32B微调为具备工具调用能力的专用Agent。**同时，公司将复杂的Flow JSON格式转换为模型更熟悉的Python表达，并建立基于真实商户反馈的周度训练机制，使模型能够持续学习新增业务场景、修正生产环境中的能力缺口。上线后，微调模型相较原闭源前沿模型推理速度提升至2.2倍、成本下降68%，实际效果亦实现超越，目前已承载Shopify Flow Agent的大部分生产流量。

图：Qwen3-32B与闭源前沿模型模型评测表现



资料来源：公司公告，国信证券经济研究所整理

表：模型表现对比

维度	Qwen3-32B微调模型 vs. 闭源前沿模型
Latency	2.2x faster
Cost	68% cheaper
LLM judge score	Higher

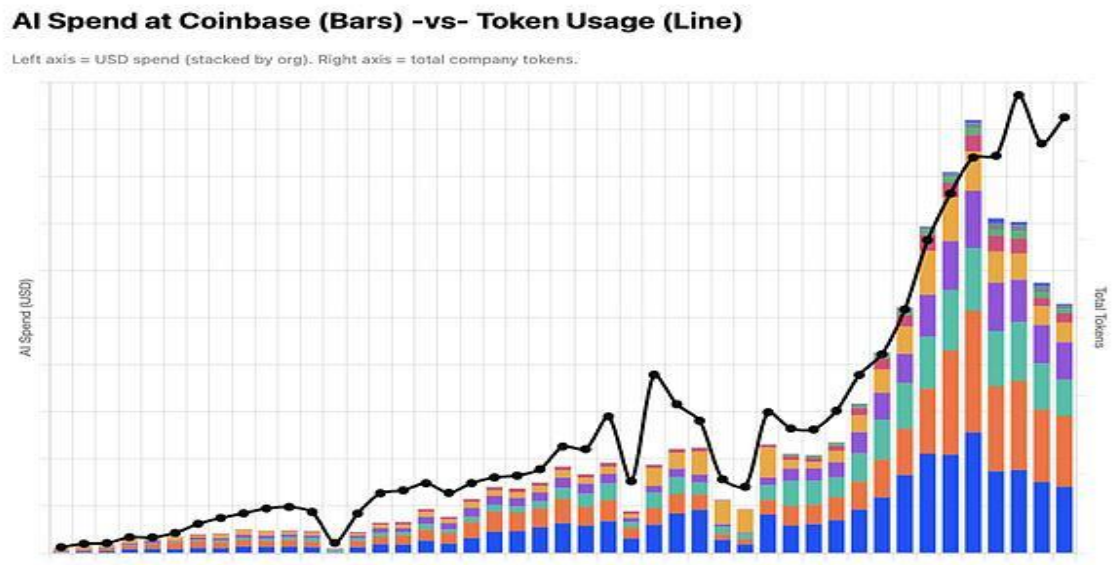
资料来源：公司公告，国信证券经济研究所整理

其他非Nebius案例2：Coinbase将GLM-5.2、Kimi2.7成为默认模型

国信证券 GUOSEN SECURITIES

- **随着内部Token使用量持续增长，Coinbase并未限制工程师使用AI，而是通过统一LLM Gateway优化模型调用结构，将GLM-5.2、Kimi 2.7等开放权重模型逐步设为高频日常任务的默认选项，**并结合自动模型路由、缓存优化、上下文精简及成本可视化等手段控制整体支出。成本控制效果显著，在AI使用量继续提升的情况下，公司AI支出已接近减半。
- Coinbase并非全面替换OpenAI、Anthropic等闭源前沿模型，而是形成“低成本开放模型承接大规模常规调用、前沿闭源模型处理复杂任务”的分层架构。实践表明，该策略在Token使用量持续增长的同时，使AI支出下降近50%。

图： Coinbase 的 AI 支出和Token使用量对比图



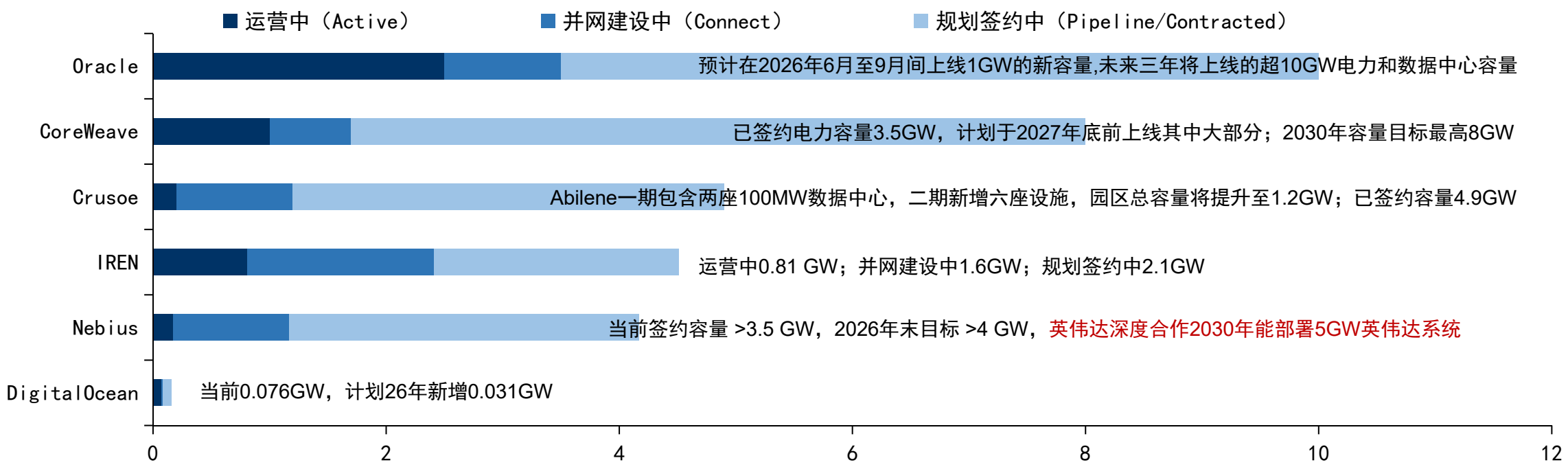
资料来源：X，国信证券经济研究所整理

目录

Nebius：快速扩张数据中心建设

- **签约电力容量：**根据公司公告，Nebius 的签约电力容量从 2025 年 8 月的 >1 GW 起步，每 3 个月跳一档（2.5 → 3 → 3.5 GW），8 个月内翻了三倍多。
- **Nebius 目前签约 3.5 GW（年底指引 4 GW）。**根据业绩会，自有 5 座大基地规划约 3 GW + 租赁补足剩余。但实际已并网运营的有芬兰 75 MW + 新泽西 300 MW + 几个小托管站点，两座1.2 GW美国超级基地要到 2027–2028 年才能逐步投产。

图：数据中心功率增长计划（GW）

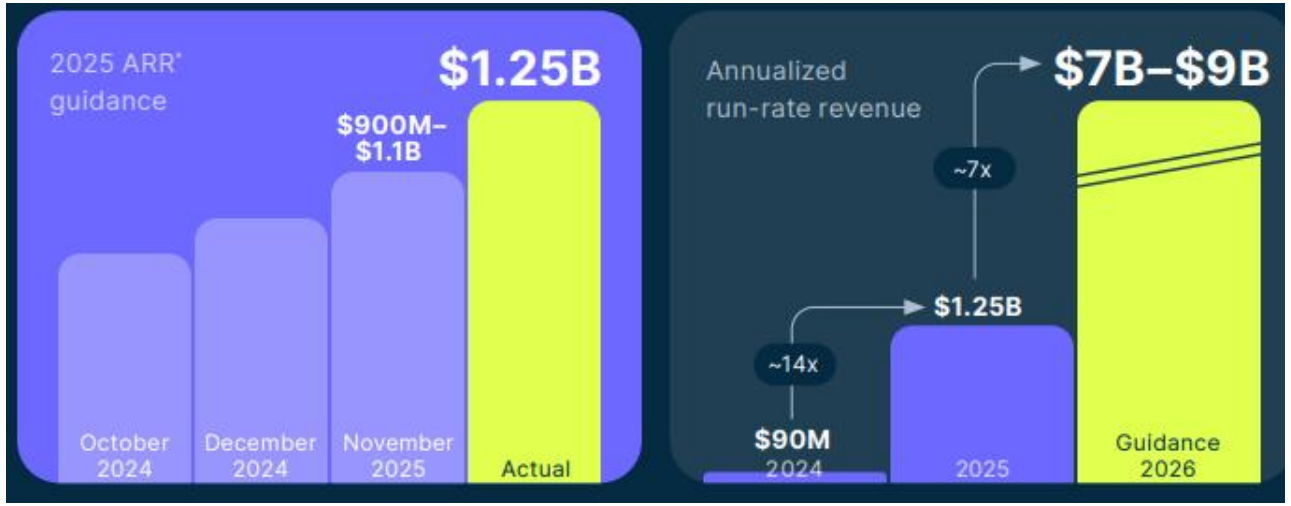


资料来源：各公司官网与业绩会、彭博社、Uptime Institute、华尔街新闻，国信证券经济研究所整理

Nebius：年度ARR预计增长7倍

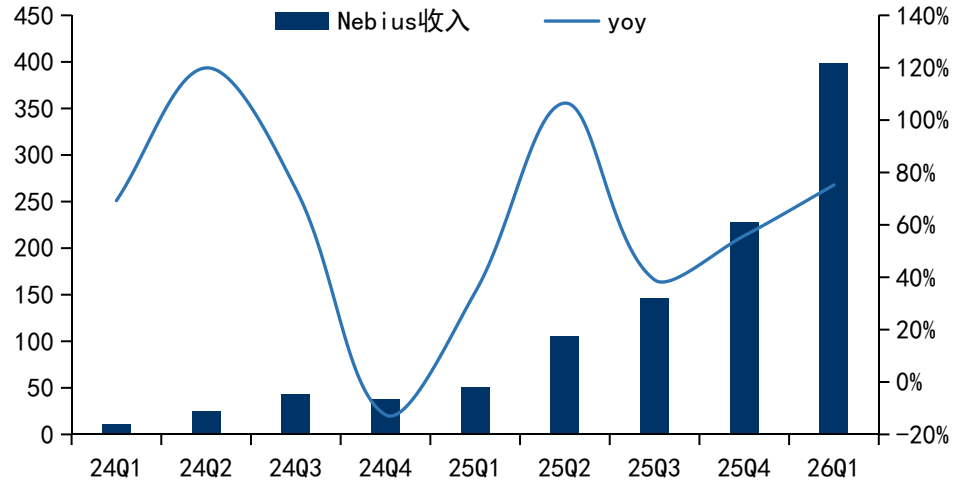
- 26Q1，营收同比飙升684%至3.99亿美元。季末现金储备达93亿美元，为其激进扩张提供了充足弹药。ARR方面，2025年底Nebius的ARR已突破12.5亿美元，公司给出的26年全年营收指引为30-34亿美元，年底ARR目标定在70-90亿美元。
- 软件（含 Token Factory）的定位是 “算力赋能工具”。核心作用是提升GPU硬件的利用率、客户粘性与单卡产出价值，收入打包在 AI 云的整体算力服务中。

图：公司26年业务指引



数据来源：公司官网，国信证券经济研究所整理

图：Nebius收入及增速



数据来源：公司官网，国信证券经济研究所整理

Nebius：背靠两家大客户，非大客户依旧增长迅猛

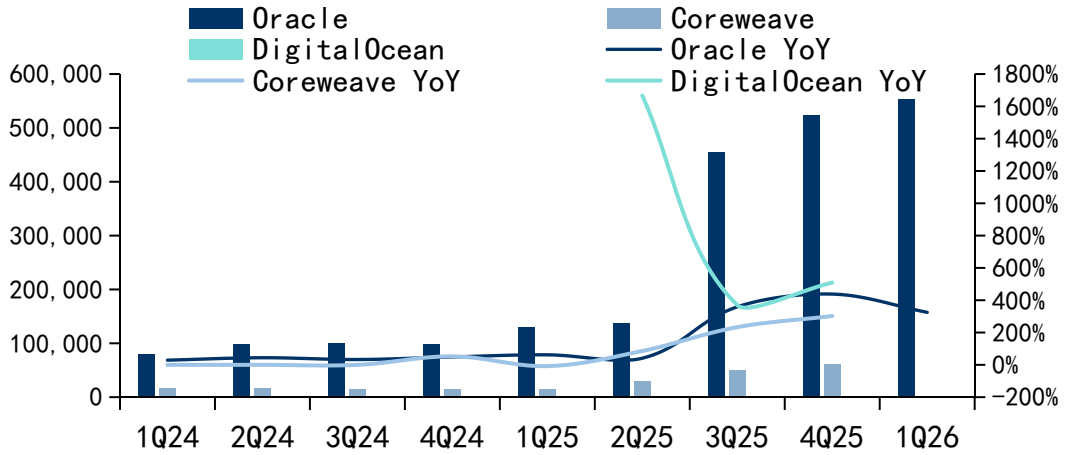
- **绑定微软、Meta大客户。**公司拥有微软五年174亿美元、Meta五年最高270亿美元两份超长期合同，以及英伟达20亿美元战略投资。预计26年两者占收入约50%。
- **与CoreWeave相比，Nebius非超大规模客户增长迅猛。**业绩会表示，26Q1不含Meta等大单的 pipeline环比增长 3.4 到 3.5 倍（整体收入环比增长75%）。CoreWeave10亿美元级承诺客户10家，商业模式建立在少数超大客户的长协之上。

图：公司与微软、Meta签约情况

合作方	合同总规模	订单性质	核心用途	交付节奏
微软	约 174 亿美元（含 70 亿预付款）	100% 确定专属集群订单	补充 Azure AI 算力缺口，支撑 OpenAI 训练、Copilot 全系列推理	分批次滚动交付，已落地 2 批（2025.11、2026.2），后续按季度爬坡
Meta	最高 270 亿美元（5 年期）	120 亿确定专属集群 + 150 亿产能兜底（未售产能由 Meta 承接）	下一代大模型训练、Agent 产品算力，部署 Vera Rubin 新架构 GPU	2027 年初开始批量交付，2026 年为建设期

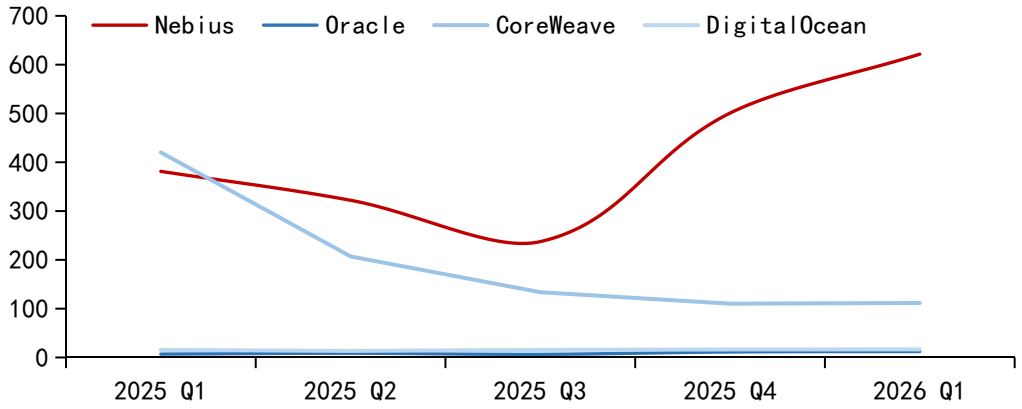
数据来源：公司官网，国信证券经济研究所整理

图：新兴云RPO及同比增速（百万美元/%）



数据来源：各公司官网，国信证券经济研究所整理

图：新兴云收入增速

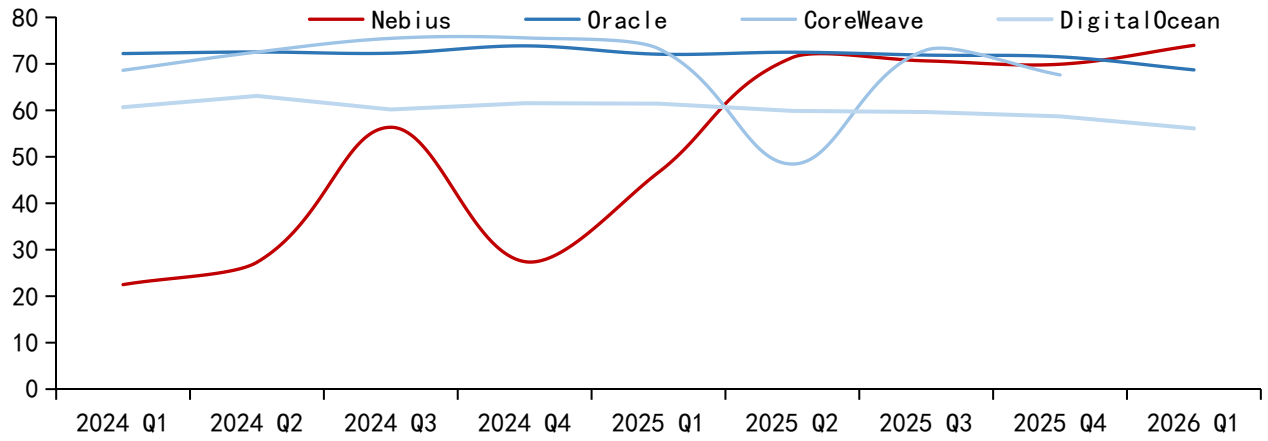


数据来源：各公司官网，国信证券经济研究所整理

Nebius：提价+自研提升+高附加值软件驱动毛利率快速提升

- Nebius毛利率由25Q1 52%提升至26Q1 74%，同比提升12pct，毛利率提升主要系：
- **卖方市场下的直接提价叠加GPU利用率拉满。**26Q1业绩会，公司表示所有投产算力 100% 售罄，每上线 1 台 GPU 有 4 家以上客户争抢。季度内再次上调算力价格，所有型号 GPU 均以更高价格全部售罄。
- **自有算力占比提升，结构性降低成本。**Nebius 自研液冷方案使得电力成本（数据中心最大运营成本之一）比竞品低近20%。
- **高附加值软件产品拉高毛利，是区别于纯算力租赁商的核心利润增量来源。**纯算力租赁属于低毛利的 IaaS 生意，而 Token Factory 等软件化产品属于高毛利 PaaS 服务。Token Factory 按 Token 按量计费，叠加 Eigen 推理优化后，单卡 Token 产出提升2-3 倍，同样的硬件成本能卖出更多收入，变相提升单卡毛利；软件层边际成本几乎为零。

图：新兴云毛利率对比（%）

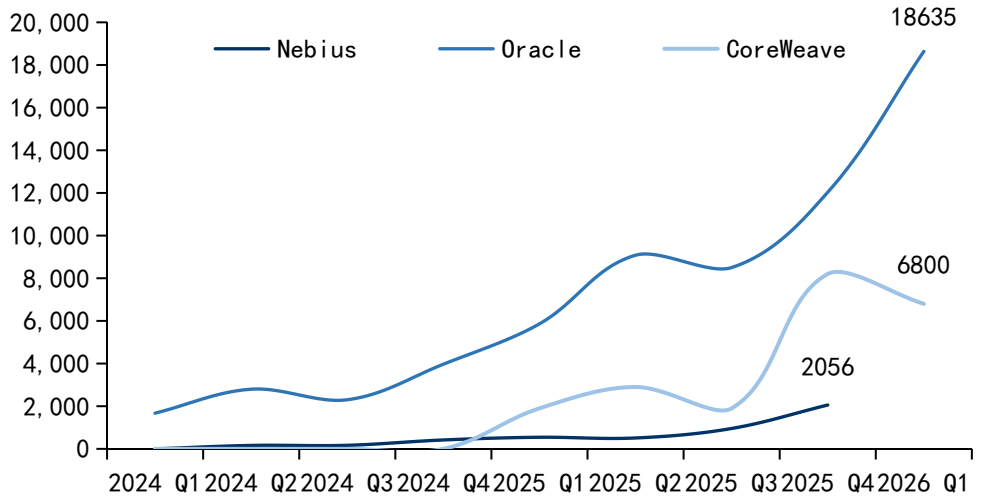


数据来源：各公司官网，国信证券经济研究所整理

Nebius：短期利润承压，高资本开支驱动规模效应

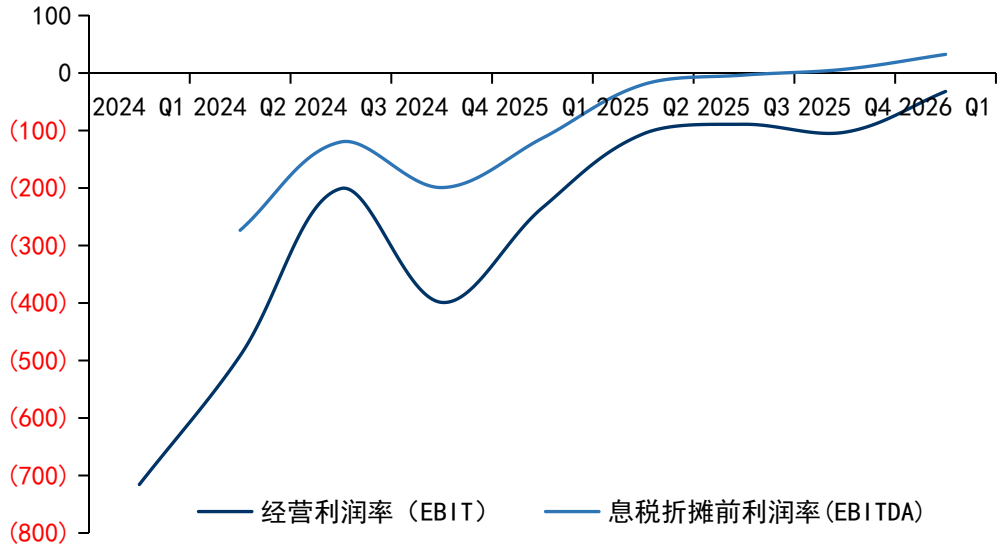
- **新兴云存在AI算力扩张期的高资本开支、折旧压力与利息压力，拖累利润。**新兴云经营利润率显著低于传统云巨头，如Nebius折旧与运营费用极高。后期与微软/Meta的长期合同落地，收入快速放量、规模效应逐步体现，亏损收窄，云业务OPM利润率从-761%（25Q1）减亏至-32%（26Q1）。
  - Nebius 2026年资本开支指引上调至200-250亿美元（2024年仅8亿、2025年41亿），其中60%靠自筹（93亿现金+客户预付款驱动的23亿季度经营现金流）、40%靠外部融资（可转债+以微软/Meta长期合同为担保的债务）。

图：新兴云毛Capex情况（百万美元）



数据来源：各公司官网，国信证券经济研究所整理

图：Nebius 利润率



数据来源：各公司官网，国信证券经济研究所整理

风险提示

第一，宏观经济波动。若宏观经济波动，公司业务、产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧，影响云业务利润率。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所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或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032