# 国信通信·算力租赁专题报告

## Neocloud引领算力租赁发展，国内市场发展可期

**行业研究 · 行业专题**

**通信**

**投资评级：优于大市（维持评级）**

证券分析师：袁文翀
021-60375411
yuanwenchong@guosen.com.cn
S0980523110003

证券分析师：张宇凡
021-61761027
zhangyufan1@guosen.com.cn
S0980525080005

# 投资摘要



**全球算力景气度延续，随着英伟达GB系列高密度算力机柜加速出货，全球高端算力景气度进一步提升。**当前AIGC浪潮下，全球服务器出货量持续增长，咨询机构IDC预计2028年全球人工智能服务器市场规模有望达到2,227亿美元，其中生成式人工智能服务器占比将从2025年的29.6%提升至2028年的37.7%。从需求端来看，模型迭代加速背景下训练端需求仍维持高位，推理侧需求随着应用的渗透逐步提升；从供给端来看，以英伟达B/Rubin、AMD MI系列为代表的高性能算力芯片持续迭代，2025年下半年GB300有望加速交付。根据CSP厂商的Capex指引，预计2025年，海外亚马逊、谷歌、微软、Meta四家厂商合计Capex增至3610亿美元，同比增幅超58%；国内字节、腾讯、阿里Capex有望超过3600亿元，AI发展高景气度延续。

**GPU云(算力租赁)或解决目前全球高端AI芯片紧缺问题，GPU云(算力租赁)市场快速发展。**在大模型军备竞赛的背景下，各大厂加速万卡甚至十万卡集群建设。Meta、微软&OpenAI、xAI等多家AI巨头陆续宣布或者完成10万卡集群建设，国内通信运营商、头部互联网、大型AI研发企业等均发力超万卡集群的布局。然而**在全球高端AI芯片供给紧缺背景下，以租赁代替购买的商业模式应运而生，租赁模式因地制宜且性价比更高。**云计算市场历经传统云、混合云阶段后，正在迎来第三次分化浪潮——AI智算云NeoCloud，即GPU云(算力租赁)，预计到2033年全球GPU云(算力租赁)市场规模将增至128亿美元(Verified Market Research预测)。

**AI芯片巨头正在通过GPU云(算力租赁)商业模式布局全球市场，国内GPU云市场发展值得期待。**（1）全球市场来看，英伟达以股权或合作方式辅助GPU云厂（CoreWeave、NBIS、Omniva等）发展，巩固其在高端芯片领域的全球主导地位。三家GPU云厂覆盖区域和发展规模虽有不同，但均受益GPU云市场的高景气度，处于快速增长期，2025Q2CoreWeave和NBIS营收增速分别达到207%/625%。（2）国内方面，国产AI芯片目前主要支持推理业务，部分训练场景英伟达高端AI芯片性能表现更优；国内外算力政策有差异，同时以OPEX租赁算力方式实现训练业务或具备更高性价比，国内算力租赁企业迎来发展契机。**目前国内算力租赁企业的租赁回报较为可观，测算净利率或达15%，与海外GPU云(算力租赁)的商业模式和发展前景具有部分相似之处。**

**投资建议：**AI算力景气度持续，短期看，GPU云(算力租赁)或为解决高端算力供需不匹配的核心解决方案；长期看，GPU云(算力租赁)具备灵活、低成本的解决方案，渗透率有望持续提升。推荐关注国内GPU云相关企业，建议关注【润建股份】及相关产业公司。

**风险提示：**AI发展及投资不及预期，行业竞争加剧，全球地缘政治风险，新技术发展引起产业链变迁。

# 目录

# 第一章 算力高景气度延续，英伟达AI芯片仍领先市场

# AIGC浪潮下，人工智能算力市场规模有望持续扩大

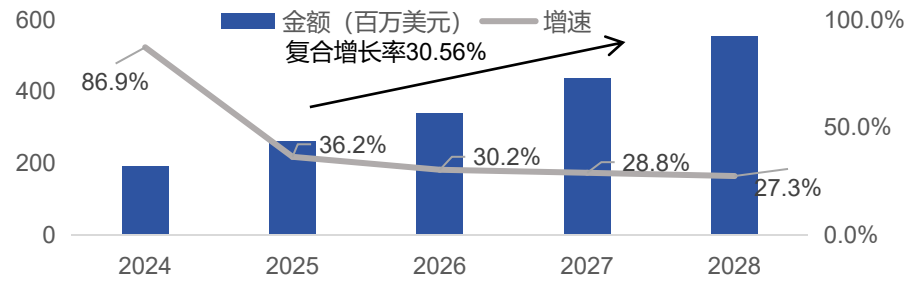
**全球人工智能服务器市场规模持续增长，生成式人工智能服务器占比不断提升。**IDC数据显示，2024年全球人工智能服务器市场规模预计为1,251亿美元，2025年将增至1,587亿美元，2028年有望达到2,227亿美元，其中生成式人工智能服务器占比将从2025年的29.6%提升至2028年的37.7%。

**我国人工智能服务器市场规模持续扩大，2025-2028年CAGR达到31%。**随着我国大模型的迭代和人工智能应用兴起，高性能计算资源的需求显著提升，人工智能服务器作为核心基础设施，市场规模持续扩大。2024年中国人工智能算力市场规模达到190亿美元，2025年将达到259亿美元，同比增长36.2%，2028年将达到552亿美元，2025-2028年CAGR为31%。同时，**推理侧与训练侧市场均呈现扩张态势，推理侧占比显著提高、增速更快。**当前推理场景的需求日益增加，推理服务器的占比有望显著提高。2024年我国推理占比为65%，预计到2028年，推理工作负载占比将达到73%。

图：全球生成式人工智能和非生成式人工智能服务器市场规模预测

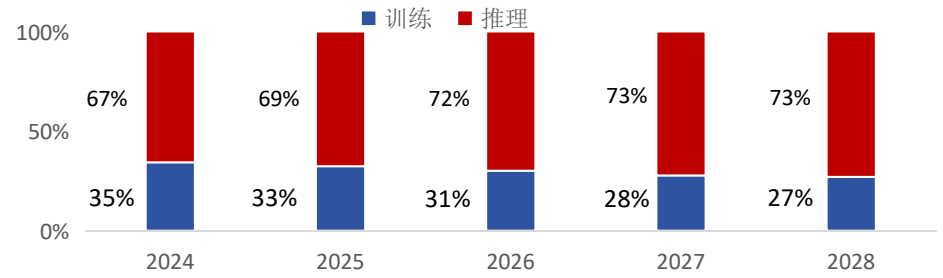

资料来源：IDC，国信证券经济研究所整理

图：中国人工智能服务器市场预测，2024-2028



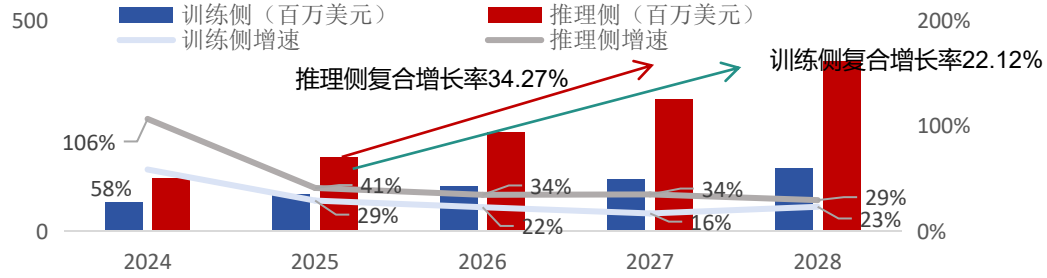

资料来源：IDC，国信证券经济研究所整理

图：中国人工智能服务器工作负载预测，2024-2028




资料来源：IDC，国信证券经济研究所整理

图：中国人工智能服务器市场训练侧和推理侧预测



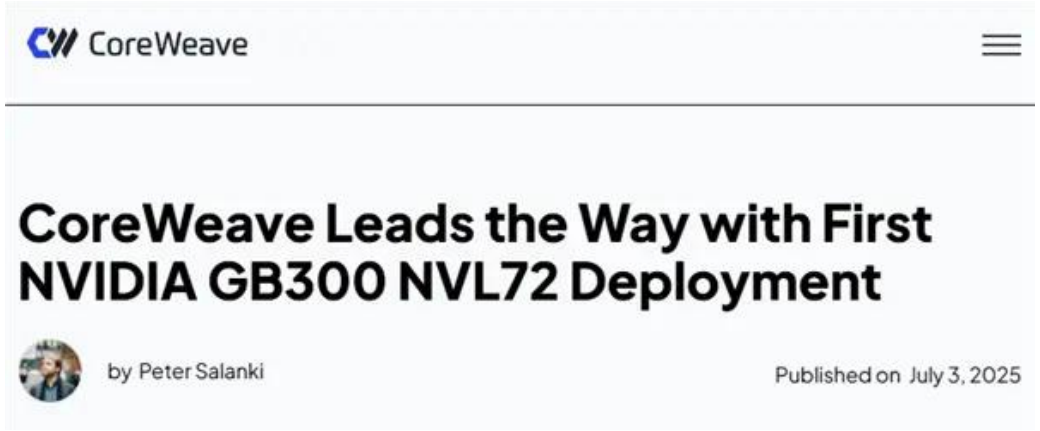

资料来源：IDC，国信证券经济研究所整理

# 算力供给持续优化：GB300预计于2025年Q3交付

**GB300服务器拥有超级算力，助力企业训练和运行AI软件。1）GB300 芯片是定位为高性能计算与 AI 推理的旗舰产品。**2025 年3 月18 日，GTC 大会发布 GB300 芯片，该芯片基于 NVIDIA Blackwell Ultra 架构。NVIDIA Blackwell 架构 GPU 拥有 2080 亿个晶体管，采用专门定制的台积电 4NP 工艺制造，且所有 NVIDIA Blackwell 产品均采用双倍光刻极限尺寸的裸片，通过 10TB/s 的片间互联技术连接成一块统一的 GPU。**2）GB300 NVL72 系统是行业首款达到 ExaFLOPS 级别的企业级算力设备。**系统内建72个英伟达Blackwell Ultra GPU和36个基于Arm架构的英伟达Grace CPU，理论算力可达 1 万亿次浮点运算 / 秒，掀开全球 AI 基础设施升级的新篇章。

**GB300首批已出货至Core Weave， 预计2025Q3开始放量。**7月3日，美国Core Weave公司宣布已收到市场上首个基于英伟达GB300的人工智能服务器系统。该系统采用戴尔 PowerEdge XE9712 服务器，以 Nvidia GB300 NVL72作为基础。CoreWeave计划在今年内持续扩大Blackwell Ultra服务器的部署规模，以满足客户不断增长的AI计算需求。除戴尔外，其他服务器厂商正推动 GB300 服务器出货。广达资深副总暨云达总经理杨麒令表示，GB300 目前按计划推进，正在测试并与客户进行验证，预计2025年 9 月出货。

图：Coreweave获得首批GB300服务器




资料来源：CoreWeave官网，国信证券经济研究所整理

图：NVIDIA GB300 NVL72


资料来源：芯智讯官微，国信证券经济研究所整理

# 算力供给持续优化：Rubin预计于2026-2027年量产

国信证券 GUOSEN SECURITIES

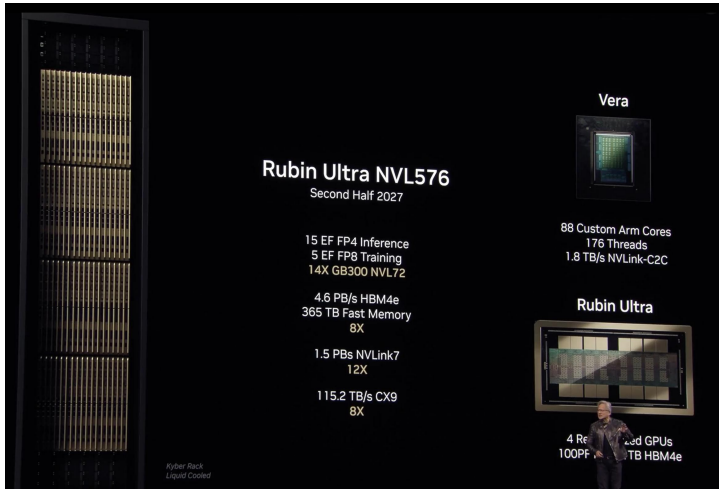
**英伟达最新一代R系列芯片在高强度AI训练和推理任务中更具优势。**英伟达持续优化和迭代GPU,R100将采台积电的N3制程 (vs. B100采用台积电的N4P) 与CoWoS-L封装 (与B100相同)。与此同时，R100采用约4x reticle设计 (vs. B100的3.3x reticle设计)。这一工艺进步增强了R100的能效比，巩固英伟达市场领先地位。

- **2026年，Vera Rubin 144 架构将量产，AI 算力正式迈入 Exascale 时代。**英伟达 CEO 黄仁勋在GTC 2025 大会确认，Vera Rubin 144将于2026年下半年推出。Vera Rubin144包含144个Rubin GPU和多个Vera CPU，采用液冷Oberon机架，功耗600kW，提供3.6 ExaFLOPS的FP4推理性能和1.2 ExaFLOPS的FP8训练性能，其性能约为Blackwell GB300 NVL72的3.3倍。
- **2027年，Rubin Ultra NVL576架构将量产，推动英伟达算力进一步突破。**GTC2025大会，黄仁勋表示，该架构包含576个Rubin Ultra GPU，FP4推理性能达15 ExaFLOPS，FP8训练性能达5 ExaFLOPS，其性能约为Blackwell GB300 NVL72的14倍。

图：Vera Rubin144

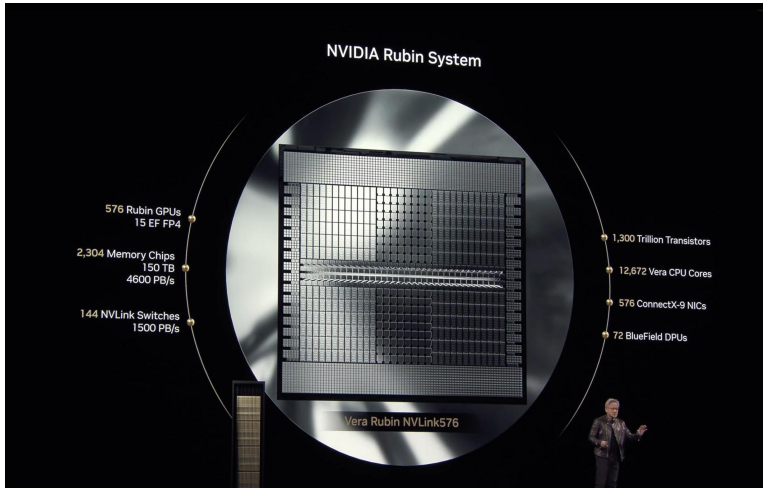

资料来源：NVIDIA官网，国信证券经济研究所整理

图：Rubin Ultra NVL576




资料来源：NVIDIA官网，国信证券经济研究所整理

图：NVIDIA Rubin System




资料来源：NVIDIA官网，国信证券经济研究所整理

# 训练侧算力供给持续优化：AMD新品MI350与B200性能相当

国信证券 GUOSEN SECURITIES

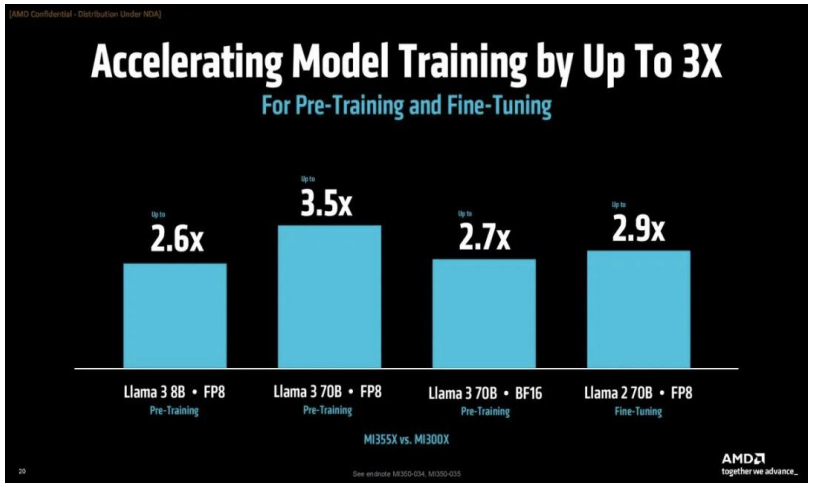
**2025年6月12日，AMD公布新产品AMD INSTINCT MI350系列GPU的研发计划。**MI350系列包括“MI350X”（风冷）和“MI355X”（液冷），将于今年三季度通过云服务公司以云服务器的形式提供给终端用户，2026年则将推出下一代产品“MI400”。

**第五代EPYC（Turin）芯片助力MI350算力提升。**通过搭配AMD第五代EPYC（Turin）芯片，8个GPU通过153.6 GB/s的双向Infinity Fabric链路进行通信，可以组成一个节点。这些节点还将继续组合成风冷或液冷机柜，形成最高128GPU的集群，FP8算力达到1.3EFLOPs。

**训练和微调上，MI350系列拥有更高性能。1)相比MI300X大幅提升**，Llama 2 70B 微调：在 FP8 精度下，MI355X 的性能是 B200 的 1.1 倍、GB200 的 1.13 倍。**2)拥有和B200/GB200相当或更高的性能。**Llama 3 8B/70B 预训练：MI355X 在 FP8 精度下相比前代 MI300X 分别提升 160% 和 250%，与 B200 性能相当。

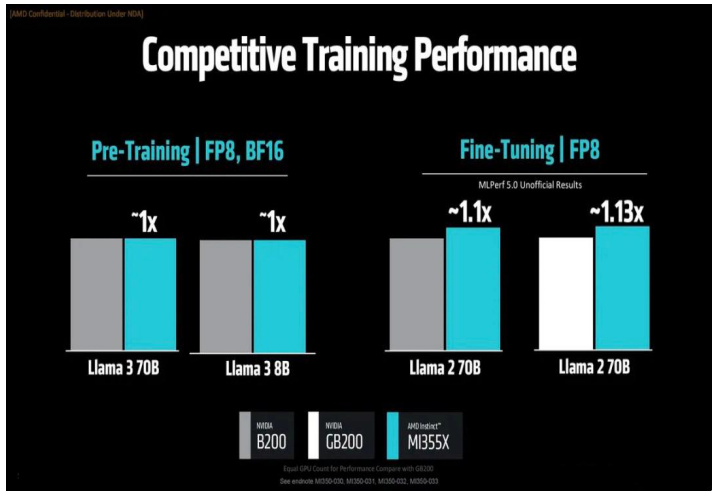
Oracle云基础设施(OCI)宣布将提供泽塔级AI集群，配备多达131,072个MI355X GPU，使客户能够大规模构建、训练和推理AI。

图：MI355X与MI300X对比（训练和微调）



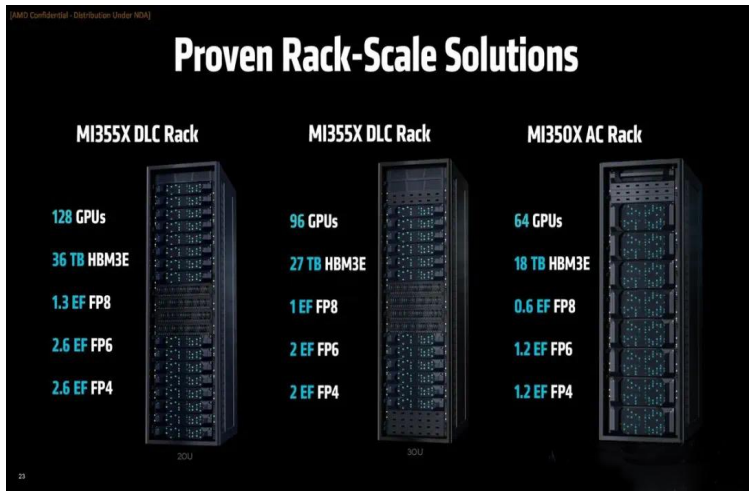

资料来源：AMD，国信证券经济研究所整理

图：MI355X与B200、GB200对比（训练和微调）




资料来源：AMD，国信证券经济研究所整理

图：MI350X和MI355X




资料来源：AMD，国信证券经济研究所整理

# Deepseek推动AI降本、开源，推理算力需求快速提升

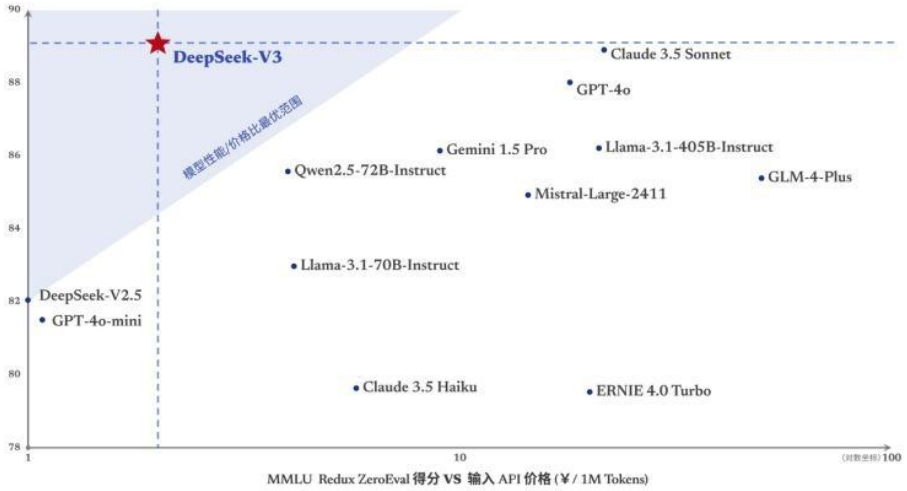
国信证券 GUOSEN SECURITIES

**DeepSeek成本低且训练高效，性价比优势显著。**DeepSeek于2024年12月和2025年1月分别发布V3训练和R1推理模型，Deepseek-V3的训练使用2048块英伟达H800 GPU，花费558万美元，成本不到其他顶尖模型的十分之一。DeepSeek R1训练算力只有Llama3的1/10，推理阶段缓存数据量降低了50倍。而根据DeepSeek性能测试，在数学任务中，DeepSeek-R1 的表现与OpenAI-o1-1217相当，并远超其他模型。

**DeepSeek推动开源趋势，加速推理侧爆发。**IDC预测，2025年，将有55%的企业使用开源人工智能基础型开发应用程序。DeepSeek使企业可以使用水平相当的开源模型，大幅降低了企业的训练和部署成本，使得更多开发者能够负担先进人工智能开发的费用，加速人工智能的普及。

杰文斯悖论指出，当技术进步提高了资源利用效率时，该资源的总消耗量反而可能会增加，Deepseek带来的效率优化不仅未抑制算力需求，反而带来算力需求的增长。同时，DeepSeek通过技术普惠化、场景纵深化和算力泛在化三重路径推动大模型运用，加速推理侧需求的爆发。

图：各大训练模型输出价格对比




资料来源：DeepSeek官网，国信证券经济研究所整理

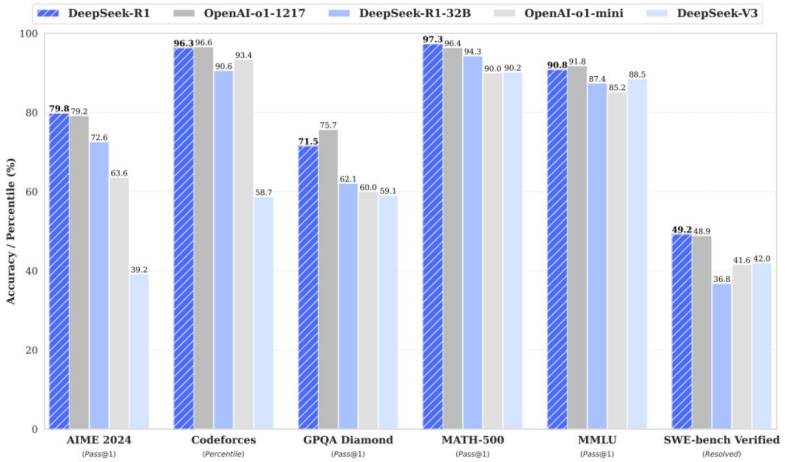
图：DeepSeek-R1与其它大模型的对比

3.1. DeepSeek-R1 Evaluation

| | Benchmark (Metric) | Claude-3.5-Sonnet-1022 | GPT-4o 0513 | DeepSeek V3 | OpenAI o1-mini | OpenAI o1-1217 | DeepSeek R1 |
|---|---|---|---|---|---|---|---|
| | Architecture | - | - | MoE | - | - | MoE |
| | # Activated Params | - | - | 37B | - | - | 37B |
| | # Total Params | - | - | 671B | - | - | 671B |
| English | MMLU (Pass@1) | 88.3 | 87.2 | 88.5 | 85.2 | **91.8** | 90.8 |
| | MMLU-Redux (EM) | 88.9 | 88.0 | 89.1 | 86.7 | - | **92.9** |
| | MMLU-Pro (EM) | 78.0 | 72.6 | 75.9 | 80.3 | - | **84.0** |
| | DROP (3-shot F1) | 88.3 | 83.7 | 91.6 | 83.9 | 90.2 | **92.2** |
| | IF-Eval (Prompt Strict) | **86.5** | 84.3 | 86.1 | 84.8 | - | 83.3 |
| | GPQA Diamond (Pass@1) | 65.0 | 49.9 | 59.1 | 60.0 | **75.7** | 71.5 |
| | SimpleQA (Correct) | 28.4 | 38.2 | 24.9 | 7.0 | **47.0** | 30.1 |
| | FRAMES (Acc.) | 72.5 | 80.5 | 73.3 | 76.9 | - | **82.5** |
| | AlpacaEval2.0 (LC-winrate) | 52.0 | 51.1 | 70.0 | 57.8 | - | **87.6** |
| | ArenaHard (GPT-4-1106) | 85.2 | 80.4 | 85.5 | 92.0 | - | **92.3** |
| Code | LiveCodeBench (Pass@1-COT) | 38.9 | 32.9 | 36.2 | 53.8 | 63.4 | **65.9** |
| | Codeforces (Percentile) | 20.3 | 23.6 | 58.7 | 93.4 | **96.6** | 96.3 |
| | Codeforces (Rating) | 717 | 759 | 1134 | 1820 | **2061** | 2029 |
| | SWE Verified (Resolved) | **50.8** | 38.8 | 42.0 | 41.6 | 48.9 | 49.2 |
| | Aider-Polyglot (Acc.) | 45.3 | 16.0 | 49.6 | 32.9 | **61.7** | 53.3 |
| Math | AIME 2024 (Pass@1) | 16.0 | 9.3 | 39.2 | 63.6 | 79.2 | **79.8** |
| | MATH-500 (Pass@1) | 78.3 | 74.6 | 90.2 | 90.0 | 96.4 | **97.3** |
| | CNMO 2024 (Pass@1) | 13.1 | 10.8 | 43.2 | 67.6 | - | **78.8** |
| Chinese | CLUEWSC (EM) | 85.4 | 87.9 | 90.9 | 89.9 | - | **92.8** |
| | C-Eval (EM) | 76.7 | 76.0 | 86.5 | 68.9 | - | **91.8** |
| | C-SimpleQA (Correct) | 55.4 | 58.7 | **68.0** | 40.3 | - | 63.7 |

资料来源：Daya Guo, Dejian Yang等，《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》，arXiv，2025，卷（2501.12948）13-13

图：DeepSeek-R1测试性能

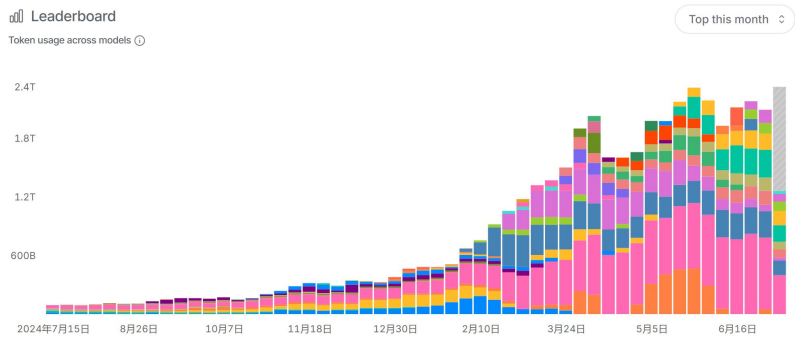



资料来源：Daya Guo, Dejian Yang等，《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》，arXiv，2025，卷（2501.12948）1-1

# 全球大模型tokens消耗量加速增长

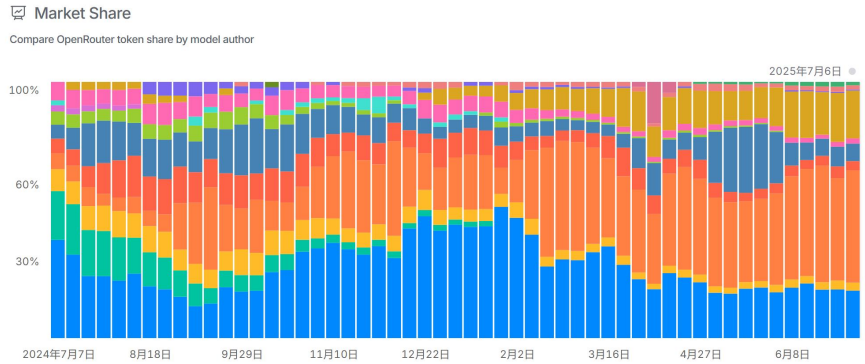
**全球Tokens消耗量加速增长，推理侧算力需求快速提升。**1）海外：根据25年5月Google I/O大会数据，Google的tokens月均调用量从24年4月的9.7万亿增长至25年4月的480万亿，增长50倍。微软 Azure AI 基础设施 2025 年一季度处理超 100 万亿 Token，同比增长 5 倍。2）国内：截至2025年5月底，豆包大模型tokens使用量超过16.4万亿，较去年5月底刚发布时增长超过137倍，月均复合增长率约43.01%，当前推理需求的爆发式增长可能会使得推理算力需求增速远超单位成本下降速度。

图：Tokens调用量全球前十大模型厂商



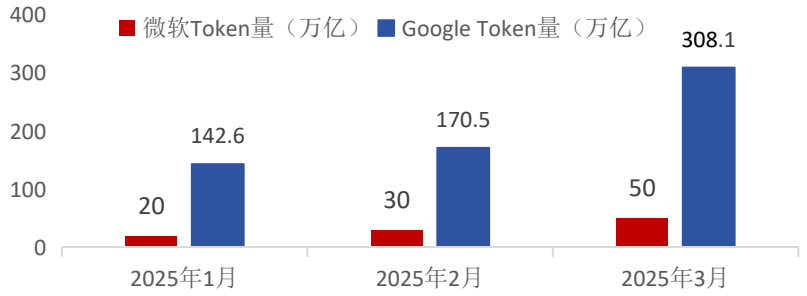

资料来源：LLM Rankings Open Rutor，国信证券经济研究所整理

图：全球大模型厂商调用量Tokens份额



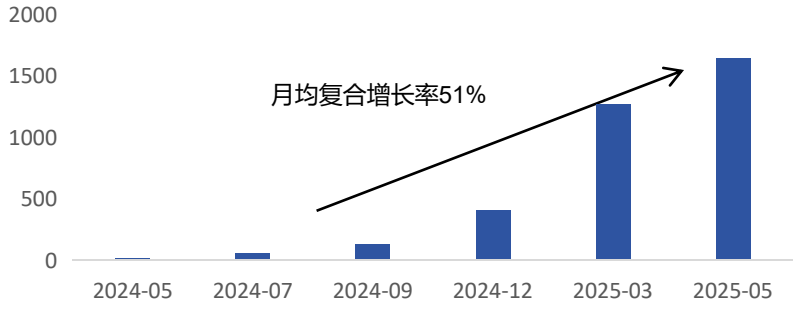

资料来源：LLM Rankings Open Rutor，国信证券经济研究所整理

图：Google和微软月度Tokens量对比




资料来源：Google I/O大会，微软，国信证券经济研究所整理

图：豆包日均tokens消耗量（单位：百亿）




资料来源：火山引擎官微，国信证券经济研究所整理

# 当前芯片供应格局：英伟达占据绝对主导，ASIC发展迅猛

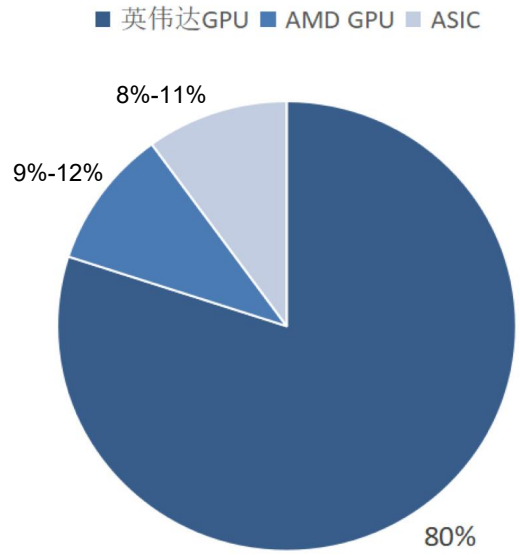
国信证券 GUOSEN SECURITIES

**英伟达占绝对主导，引领AI芯片市场发展。**2025年全球AI芯片市场规模预计突破1200亿美元，年均复合增长率(CAGR)超25%。英伟达凭借CUDA生态的深厚壁垒，长期垄断AI训练市场，英伟达GPU占AI服务器市场80%以上市场份额，成为AI芯片市场的引领者，ASIC仅占8%~11%。

**ASIC 市场发展迅猛，重塑半导体市场格局，推理算力需求增长。**ASIC是专为满足特定应用而设计的芯片，相比GPU，在AI应用中效率更高。ASIC正成为全球云厂商及AI模型企业的核心布局方向。Marvell 作为核心供应商，推动 AI 芯片从 GPU 向 ASIC 的范式转移，重塑数据中心半导体市场格局。2024年，全球ASIC出货量达345万颗，亚马逊Trainium芯片年出货量增速突破200%。这一趋势表明，ASIC的规模化应用正在推动推理算力需求的爆发式增长，并带动产业链上下游协同升级。

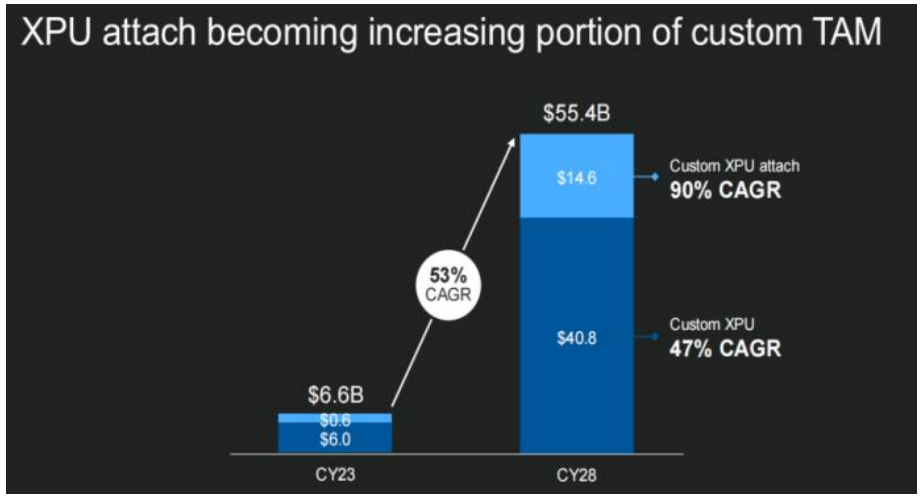
**云厂商自研芯片势头强劲，ASIC架构的时代逐步到来。**海外CSP大厂自研芯片AWS Trainium2、谷歌TPU v5成本较英伟达方案降低40%，自用率超70%。2025年，预计谷歌和亚马逊AWS两家的ASIC芯片出货量合计约为英伟达GPU出货量的40%~60%。到2026年，随着Meta和微软的大规模部署，ASIC出货量有望超越英伟达GPU。

图：当前芯片供应格局




资料来源：中研网，国信证券经济研究所整理

图：Marvell上修AI ASIC市场规模




资料来源：Marvell，国信证券经济研究所整理

# 谷歌TPU最早发力ASIC市场，AWS紧随其后

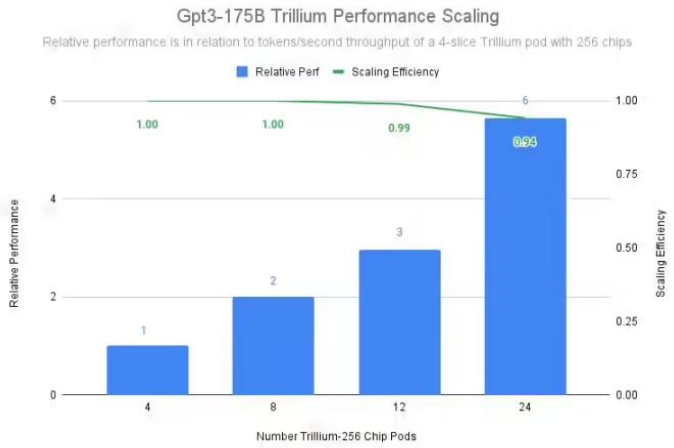
国信证券 GUOSEN SECURITIES

表：Google、AWS的ASIC情况

| | 第一次正式发布ASIC | 芯片命名 | 想达到的初始目标 | ASIC合作方 | 最新一代ASIC表现 | 芯片采购策略 | 对于ASIC的将来愿景和规划 |
|---|---|---|---|---|---|---|---|
| **Google** | 2015 | TPU | 能效提高2-3倍 | 博通、联发科 | TPU v6(最新TPU)性能是v5e的4.7倍， 能效提高67%,单个pod最多可支持256个 TPU。 | 谷歌自2017年开始加大内部定制芯片(TPU) 的采用，目前大部分内部训练和推理工作负载 均由TPU完成，同时NVIDIA的解决方案也可供谷歌云客户使用。 | 谷歌十多年来一直致力于开发ASIC,以推动 规模化和效率的前沿发展。TPU在Google I/0大会上支持了众多创新，并将在未来的 TPU中支持更多创新，包括Gemini 1.5 Flash、Imagen 3和Gemma 2;所有这些模型 都已使用TPU进行训练和运行。 |
| **AWS** | 2018 | Inferentia Trainium Gravtion | 每瓦性能提高50% | 世芯、Marvell | Inferentia 2的性能比Inferentia 1提升高达2.3 倍，且每次推理的成本降低 70%。与Inferentia 2相比，Trainium可节省高达50%的训练成本。 | AWS使用多种芯片来源来支持 客户不同的工作负载。例如， 定制ASIC (Trainium/Inferentia AI 芯片、Graviton CPU)和GP CPU/GPU(Intel、NVIDIA和AMD)。 | GPU的短缺使得构建和使用 Gen AI模型/应用程序的成本 极其高昂。AWS一直在开发 Bedrock等工具，允许开发人 员通过API访问基础模型，并设计ASIC以提供卓越的性 价比，并使其速度更快、成本 更低、能耗更低。 |

资料来源：公司官网，国信证券经济研究所理

图：谷歌TPU v6 Trillium具备近乎线性的扩展能力

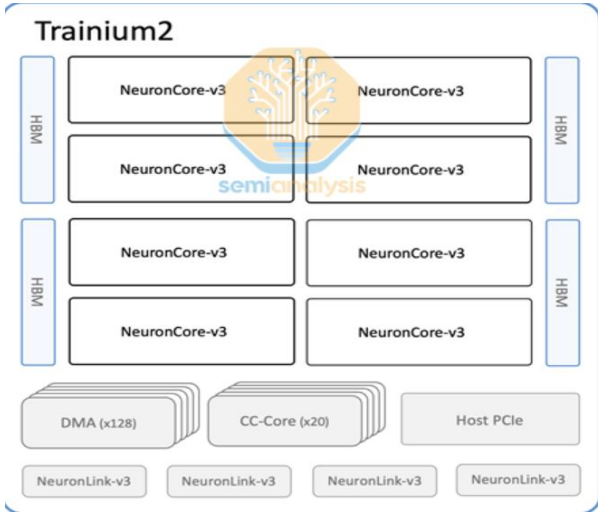



资料来源：IT之家，国信证券经济研究所整理

图：TPU v6 Trillium相较于上一代性能和效率明显提升


资料来源：IT之家，国信证券经济研究所整理

图：Trainium2专为大规模 AI 模型训练和推理设计




资料来源：AWS，semianalysis，国信证券经济研究所整理

# 微软Maia加速扩张，Meta ASIC实现性能跃迁

表：微软、Meta的ASIC情况

| | 第一次正式发布ASIC | 芯片命名 | 想达到的初始目标 | ASIC合作方 | 最新一代ASIC表现 | 芯片采购策略 | 对于ASIC的将来愿景和规划 |
|---|---|---|---|---|---|---|---|
| **Microsoft** | 2023 | Maia | 成本效益 | GUC、Marvell、联发科 | Maia 100的性能和内存带宽均低于NVIDIA H/B系列，但芯片成本也较低。 | 微软的Athena项目采用来自多个来源的AI解决方案，例如内部ASIC,AMD/NVIDIA的GPU,d-Matrix的AI芯片等。Maia 100为运行OpenAI等高级AI工作负载提供了新的选择。 | 微软正在共同设计和优化硬件和软件，以实现1+1大于2。微软已经在设计第二代Maia AI加速器和Cobalt CPU。该公司的使命始终清晰：优化其技术堆栈的每一层，从核心芯片到终端服务。 |
| **Meta** | 2023 | MTIA MSVP | 2倍性能 | 博通、Marvell | MTIA v2的整体性能是MTIAv1的3.5倍。H100的工作量比MTIA v2多5.6倍，但功耗却高出7.8倍，成本高出10-15倍。 | Meta使用内部ASIC进行AI 推理工作负载，但其Llama 4 训练仍使用NVIDIA的H100。 其定制加速器(MTIA)采用 RISC-V内核。 | MTIA是Meta长期路线图的重要组成部分，旨在为其AI工作负载构建最强大、最高效的基础设施。Meta正在设计ASIC,使其能够与现有基础设施以及未来更先进的硬件协同工作。目前，正在进行的几个项目旨在扩展MTIA的应用范围 。 |

资料来源：公司官网，国信证券经济研究所理

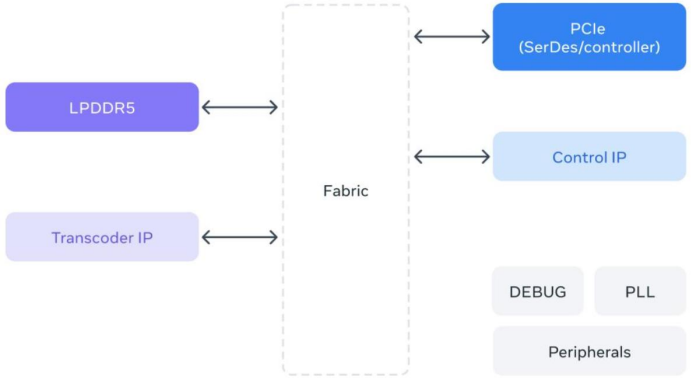
图：Microsoft发布Maia 100


资料来源：Microsoft官网，国信证券经济研究所整理

图：Microsoft Maia 100介绍


资料来源：Microsoft官网，国信证券经济研究所整理

图：Meta MSVP架构示意图




资料来源： Meta，国信证券经济研究所整理

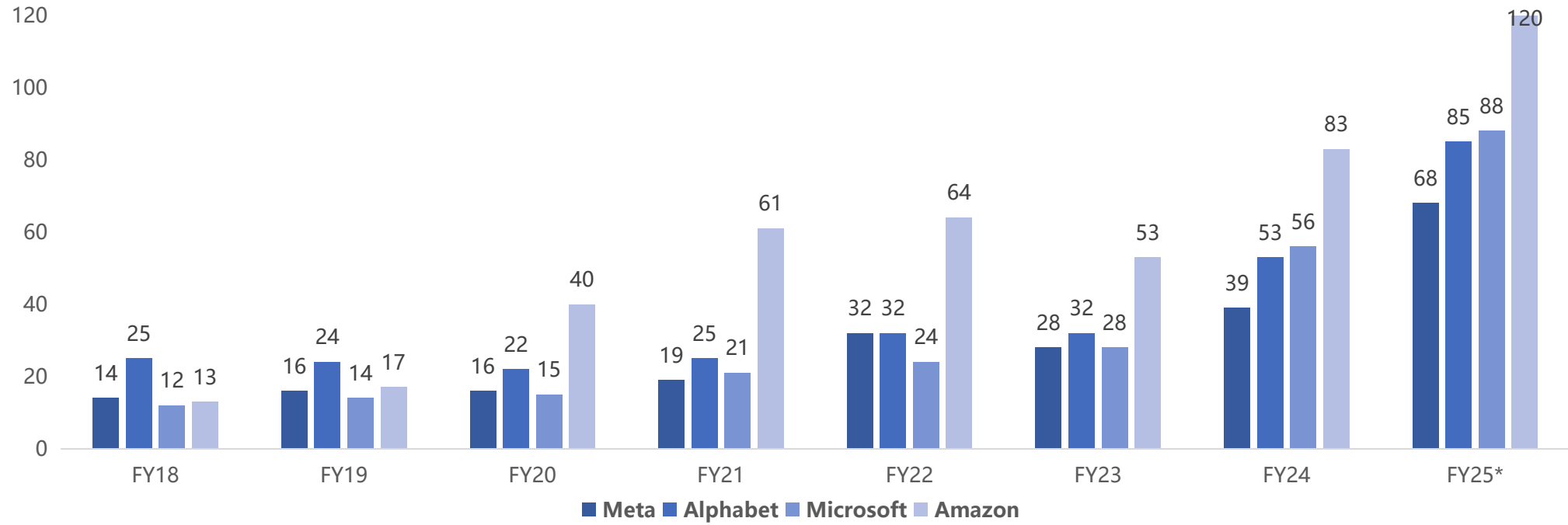
# 海外大厂CAPEX指引乐观，算力维持高景气度

国信证券 GUOSEN SECURITIES

**海外厂商高度聚焦 AI 算力基建，预计25年CAPEX增幅超58%，为全球算力产业注入强劲动力。**当前北美CSP厂商的CAPEX预计呈现快速增长态势，2024 年，亚马逊、谷歌 、微软 、Meta四家厂商合计 Capex 共 2283 亿美元，预计2025年增至3610亿美元，增幅超 58%。

根据最新财报情况，Capex方面，谷歌、微软、Meta、Amazon均进一步加大投入力度，谷歌将2025年全年资本支出预期从750亿美元上调至约850亿美元，并预计2026年将进一步增加；微软预计FY26Q1将资本开支上提至300亿美元以上，预计全年达到880亿美元；Meta将全年资本支出的最低水平从640亿美元上调至660亿美元，资本支出范围在660亿美元至720亿美元之间；Amazon全年资本支出总额预计将达到1185亿美元。

图：海外大厂CAPEX情况（单位：十亿美元）




资料来源：Economy Insights，国信证券经济研究所整理

# 第二章 高端算力资源稀缺，GPU云(算力租赁)市场价值显现

# 海内外CSP万卡集群迅速布局

国信证券 GUOSEN SECURITIES

**在大模型军备竞赛的背景下，国内外巨头加速万卡集群建设。**据IDC研究，预计2022年至2032年全球人工智能产业规模的复合增长率高达42%，2032年将达到1.3万亿美元。国际上Meta、微软&OpenAI、xAI等多家AI巨头陆续宣布或者完成10万卡集群建设，国内通信运营商、头部互联网、大型AI研发企业等均发力超万卡集群的布局。

表：全球范围内部分科技公司智算布局

| | 万卡智算集群布局进展 |
|---|---|
| **谷歌** | 2023年5月，推出AI超级计算机A3，搭载了约26000块H100 GPU，为其在机器学习和深度学习研究中的应用提供强大的算力支持。2024年，深化在万卡智算集群的布局。其基于自研芯片搭建的TPUv5p 8960卡集群高效运转，在人工智能基础研究与应用开发方面不断发力。 |
| **Meta** | 2024年初，Meta建成了两个各含24576块GPU的集群。2024年底，其朝着构建包含35万块H100 GPU的基础设施的目标推进，建成后将大幅提升在元宇宙和AI研究领域的技术竞争。 |
| **微软** | 自构建万卡超级计算机后，微软不断扩充其万卡集群规模与应用范畴。在云计算和ai服务领域，微软利用万卡集群为旗下人工智能产品与服务提供坚实算力支撑，推动Azure云服务中AI功能的持续升级优化。 |
| **亚马逊** | Amazon EC2 Ultra集群采用了2万个H100 TensorCore GPU，为用户在处理大规模数据分析和机器学习任务方面提供强大算力支持。 |
| **特斯拉** | 2023年8月，特斯拉上线集成1万块H100 GPU的集群，将极大提升特斯拉在自动驾驶和车辆智能化方面的研发速度。 |
| **百度** | 百度文心大模型4.0是在万卡AI集群上训练出来的，也是国内首次使用万卡规模集群进行训练的语言大模型。 |
| **腾讯** | 推出的星脉高性能网络能够支持高达10万卡GPU的超大规模计算，网络带宽高达3.2T，为未来的AI和大数据应用提供了广阔的发展空间。 |
| **字节跳动** | 提出的MegaScale生产系统，支撑12288卡Ampere架构训练集群，为字节跳动在内容推荐、图像处理等AI应用方面提供了强大的算力保障。 |
| **阿里巴巴** | 阿里云已服务全国一半的人工智能大模型企业，在通义千问Qwen3训练中，基于1.2万卡H100集群，训练周期仅14天，模型算力利用率(MFU)达68%，较传统架构成本降低53%，彰显出其在万卡智算集群应用上的领先优势。 |
| **华为** | 2023年7月华为昇腾AI集群全面升级，规模从4000卡集群扩展至16000卡，是业界首个万卡AI集群，拥有更快的训练速度和30天以上的稳定训练周期。 |
| **小米** | 2024年12月正在着手搭建自己的GPU万卡集群，将对AI大模型大力投入。 |
| **中国移动** | 2025年，计划商用哈尔滨、呼和浩特、贵阳三个万卡集群，总规模接近6万张GPU卡。 |
| **中国电信** | 计划2024年在上海规划建设一个达到15000卡、总算力超过4500P的万卡算力池。2024年3月，天翼云上海临港万卡算力池已正式启用。 |
| **中国联通** | 计划今年内在上海临港国际云数据中心建成中国联通首个万卡集群，集群建成后将为中国联通在数据中心和云计算市场提供的竞争优势力。 |

资料来源：AI云原生智能算力架构，公司公告，国信证券经济研究所理

# 全球市场英伟达高端AI芯片仍处于供不应求的状态

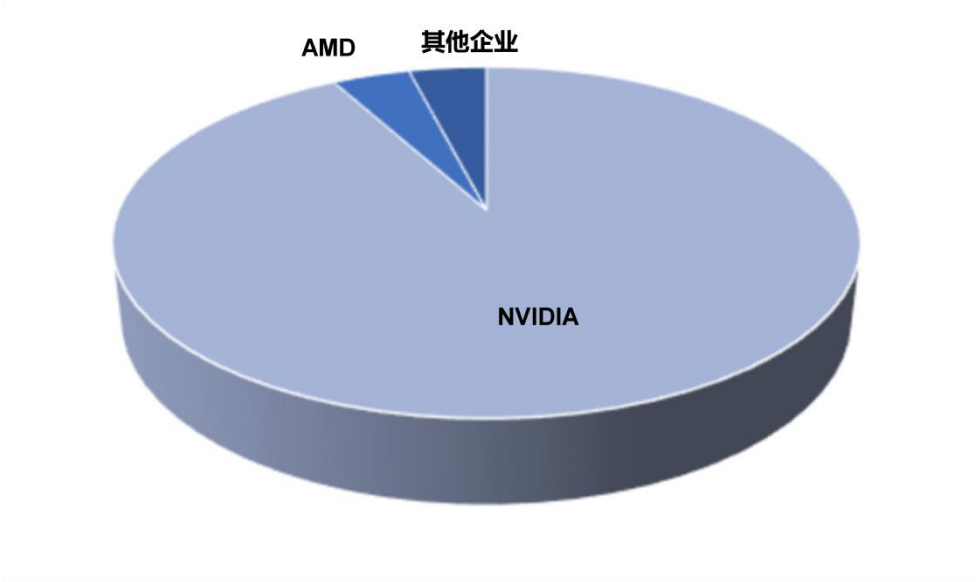
国信证券 GUOSEN SECURITIES

**供给端高度集中且产能受限。**目前AI芯片市场已形成由英伟达和AMD构成的“一超一强”寡头垄断格局，其中英伟达作为行业领导者占据全球超80%的市场份额；在生产端，英伟达的芯片采用4nm工艺与CoWoS封装技术由台积电独家代工。在2024年H100芯片产能为150万块，而台积电的CoWoS产能在2025年每月仅7.5~8万片，尽管计划到2028年提升至每月15万片，但短期内产能紧张的问题仍较为突出。

**需求端爆发式增长且需求旺盛。**在AI芯片的需求端，市场呈现出强劲且持续扩张的态势，微软、谷歌等四大云服务商贡献了英伟达数据中心业务50%以上的收入，而OpenAI等公司对算力的需求更是急剧攀升，以GPT-4V为例，其单次训练需要8000块H100运行60天；从需求规模来看，截至2023年底，全球计算力总规模已突破1300 EFLOPS，其中智能算力占比超63%，预计到2030年全球算力规模将突破16ZFlops，2023-2030CAGR达42%，其中智能算力占比预计超90%。

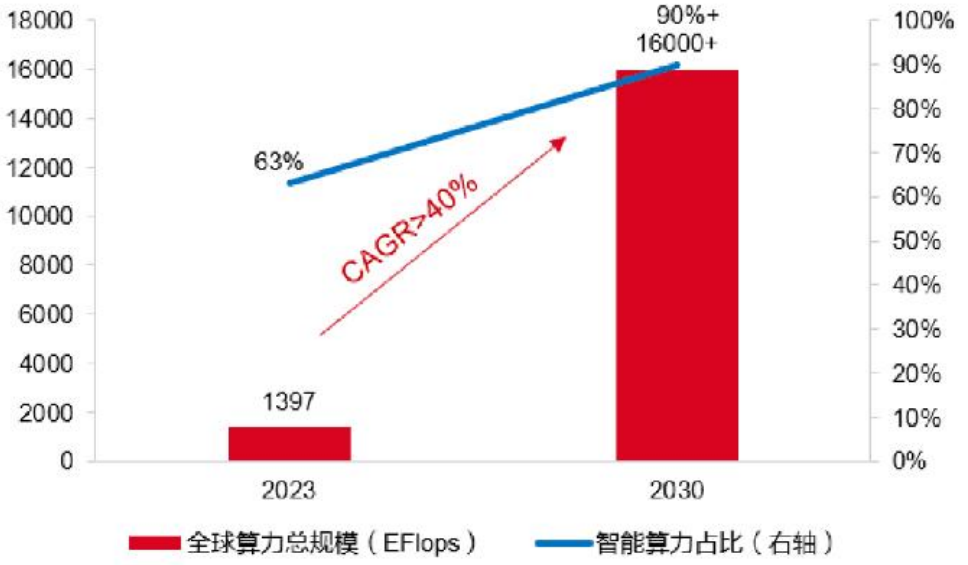
**供需缺口明显且短期难缓解。**GB200原计划2024年9月量产，因液冷等技术挑战推迟至2025年二季度；GB300原计划 2025年三季度量产，后调整为9月批量出货，因此只能复用现有设计缓解供应链压力。

图：2024年全球AI芯片市场份额占比




资料来源：半导体产业纵横，国信证券经济研究所整理

图：全球算力总规模及智能算力占比




资料来源：华经情报网，国信证券经济研究所整理

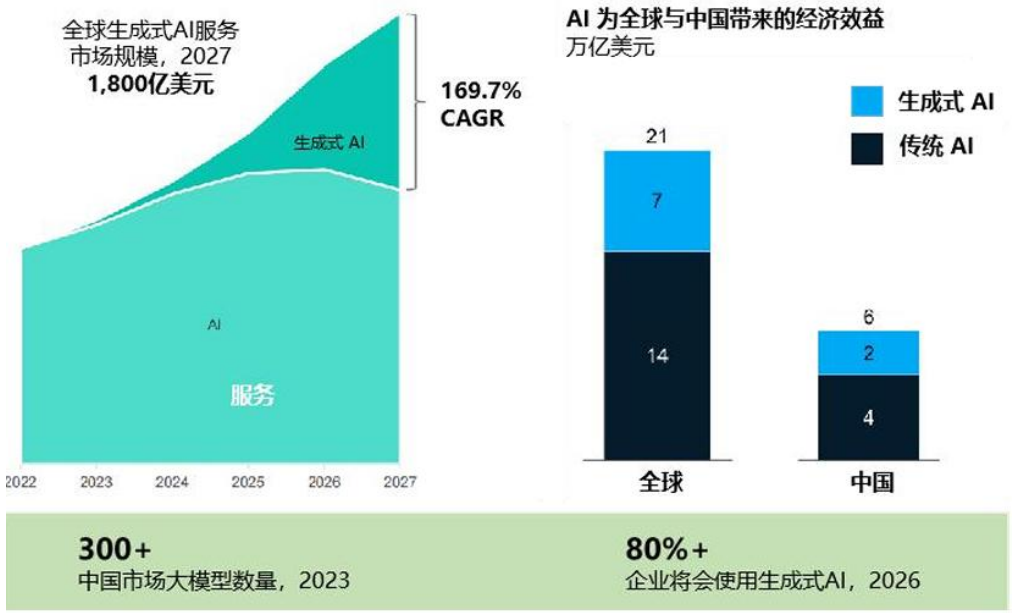
# 云计算市场再迎分化，GPU云(算力租赁)市场规模扩张显著

国信证券 GUOSEN SECURITIES

**AIGC产业趋势下更高规格的算力需求显著提升，生成式AI服务的市场规模快速增长。**当前AIGC产业趋势明晰，生成式AI已成为不可或缺的生产工具，预计2026年80%以上的企业将会使用生成式AI。全球生成式AI服务的市场规模预计在2027年将增长至1800亿美元，2022-2027年的CAGR为169.7%。

**云计算市场历经传统云、混合云阶段后，迎来了第三次分化浪潮——AI 智算云 NeoCloud。**Gartner 预测，到 2026 年，全球云计算市场有望达到万亿美元。而伴随着大模型技术持续突破，云计算市场加速分化，逐步形成多元化格局。在AIGC浪潮下，云服务商开始提供专门为 AI 训练和推理优化的 GPU 。 Verified Market Research数据预测，**预计到 2033 年，全球 GPU 云市场规模将增至 128亿美元。**

图：生成式AI驱动AI市场规模化发展




资料来源：Gartner，麦肯锡，IDC，国信证券经济研究所整理

图：全球GPU云计算市场规模与范围


资料来源：Verified Market Research，国信证券经济研究所整理

# 算力租赁：因地制宜、部署灵活、性价比高的算力解决方案

国信证券 GUOSEN SECURITIES

**算力租赁是指企业或个人通过支付租金的方式，从拥有大量计算资源的服务提供商那里租用所需的计算能力。**对于需要大规模计算能力但又不希望或无法承担高昂前期投资成本的用户来说，算力租赁提供了一种灵活、高效且成本较低的解决方案。服务提供者通常是拥有大量计算资源的公司，如云服务厂商、传统IDC服务厂商以及其他跨界布局的企业。随着AIGC发展，专门针对生成式AI提供高性能算力租赁的厂商被称之为NeoCloud，成为当前节点算力租赁的核心力量。

算力租赁的**产品形态多元**，主要分为服务器租赁、虚拟机租赁、GPU租赁以及存储和网络资源租赁。其中GPU租赁是指针对需要进行大规模并行计算的任务，如人工智能训练和图像处理，用户可以租赁GPU服务器资源。

运营模式可分为**签约制**和**按需使用**两种：签约制能够为服务商提供稳定现金流和长期业务可见性。客户通过预先承诺资源使用量换取较低单价和更高服务保障，服务商则依托合约融资，支撑GPU采购和数据中心建设等高额资本开支。按需使用主要服务于中小型企业或临时需求，灵活性高，但单价较高且可能面临资源竞争。

表：签约制与按需使用的商业模式对比（以CoreWeave为例）

| | 签约制 | 按需使用 |
|---|---|---|
| **定义** | 客户预先承诺2-5年购买固定容量计算资源，签订“照付不议”合同 | 客户根据实际需求临时租用资源，无长期承诺，按使用量付费 |
| **定价方式** | 按合约期限和承诺量阶梯定价，单价较低（较GCP便宜40%） | 按市场价计费，单价较高（溢价20-50%） |
| **合同灵活性** | 低，需提前规划容量；部分合同允许硬件升级 | 高，可随时启停 |
| **客户类型** | 大型AI实验室，云厂商，企业级客户 | 中小型企业、初创公司、测试项目 |
| **现金流周期** | 签约时收15-25%预付款，按月分期收款 | 按使用后结算，无预付款 |
| **资源保障** | 优先分配资源，确保高可用性 | 依赖剩余资源，高峰期可能短缺 |
| **典型案例** | OpenAI与CoreWeave的1119亿美元5年合约 | 临时扩容推理算力、短期模式测试 |
| **风险承担** | 客户承担未使用资源成本；服务商承担折旧和债务压力 | 客户承担价格波动风险；服务商需管理闲置资源 |

资料来源：爱思博格公众号，国信证券经济研究所理

表：算力租赁与自建成本项对比

| 成本项 | 算力租赁（以某头部云服务商为例） | 购买算力（自建） |
|---|---|---|
| **初期投入** | 无（按需付费） | 硬件采购：100×8万=800万元<br>机房建设：约200万元（含电力、冷却、机柜） |
| **单月运维成本** | 约15万元（按100台GPU×150元/台/天×30天计算） | 约8万元（电费+冷却+基础维护） |
| **年综合成本** | 15万×12=180万元 | 800万（折旧摊销，按3年计）+8万×12=96万=896万元 |
| **弹性扩展成本** | 无额外硬件投入，可快速扩容 | 需提前采购冗余设备，闲置成本高 |

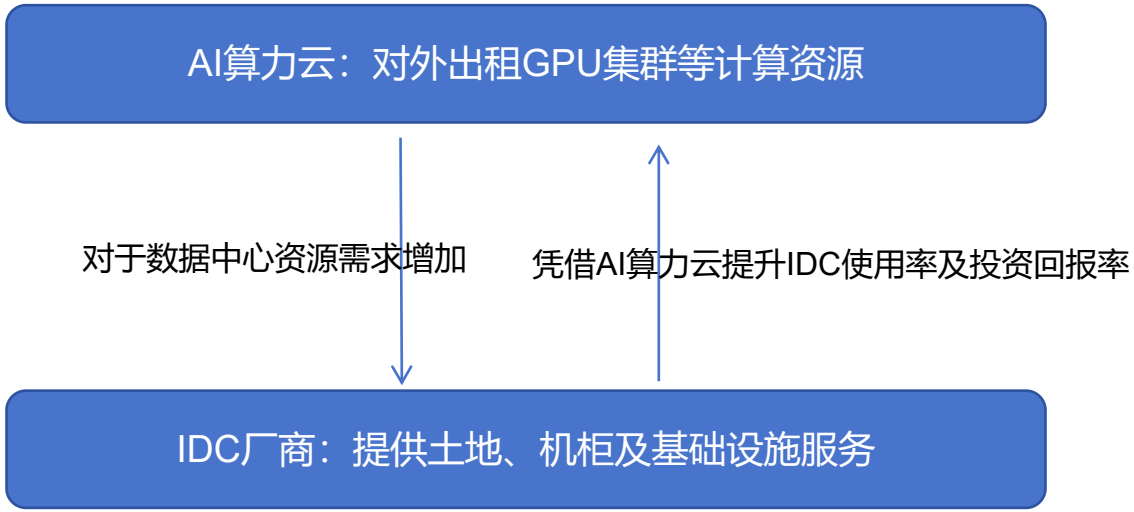
资料来源：JYGPU极智算，国信证券经济研究所整理

# 服务提供者NeoCloud：提供AI高性能GPU云(算力租赁)服务

**Neocloud商业模式与传统云厂商相似，核心区别在于提供以英伟达为代表的最新高性能芯片的能力。**AI算力云专注于提供GPU算力租赁服务，支持AI模型训练、微调和推理等任务。AI算力云对于数据中心资源的需求急剧增加，特别是在液冷基础设施和高速网络互联方面。因此领先的IDC服务商需升级设施，提供高功率密度液冷机柜和卓越的网络解决方案，以满足AI云厂商的需求。同时，AI算力云商寻求与这些IDC服务商建立长期合作，为自己的设备部署提供稳定支持，进一步提高数据中心的使用率和投资回报率。

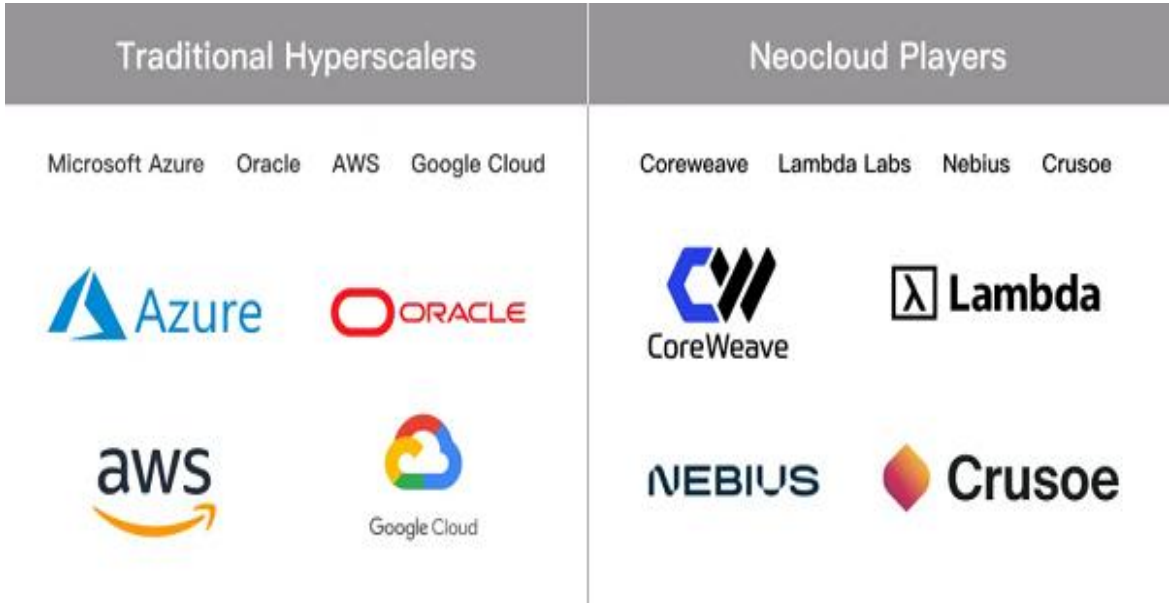
Hyperscalers（如AWS、Azure、Google Cloud等）依托其全球庞大的基础设施，提供全方位的云服务，优势在于大规模的资源整合、全球化的服务覆盖以及强大的市场份额。这些巨头在传统计算和 GPU 云服务等领域均具备强大的技术实力和市场影响力。NeoCloud （如CoreWeave、Lambda Labs、Crusoe等）则专注于提供专为 AI 领域进行工作负载优化的计算资源。NeoCloud 服务商通过深度优化 GPU 资源、专注于 AI 模型训练和推理等特定应用场景，能够为用户提供更灵活、高性价比的计算能力。传统云厂商在按需服务领域更具优势，**Neocloud的竞争力集中在高性能专有集群和长期合约优化上。**

图：AI算力云与IDC在一定程度上形成互补




资料来源：华纳云，国信证券经济研究所整理

图：GPU云市场中传统巨头与新兴玩家分化




资料来源：界面新闻，国信证券经济研究所整理

# 算力租赁相较于IDC自建各有优劣，获卡渠道的稳定性为核心壁垒



IDC核心壁垒在于获取一线城市及周边地区的土地资源与能耗指标，并通过绿色节能技术降低运营成本。政策要求数据中心PUE逐步降低，且在绿色电力比例要求下，一线城市能耗指标审批趋严，数据中心逐步向其他非一线城市部署。

**算力租赁行业的核心壁垒为获卡渠道稳定性**，尤其在英伟达高端芯片如H100等GPU供应受限背景下，资源垄断性成为关键。国产替代（如华为昇腾910系列）虽加速，但短期内无法完全替代英伟达系芯片单次大模型训练需消耗千卡级GPU集群。

表：IDC与算力租赁不同维度对比

| 对比维度 | 算力租赁 | 自建IDC |
| --- | --- | --- |
| **初期成本** | 低（如单卡月租金约18000元，初期投入成本节70%+） | 高（需一次性采购硬件，如单台H100服务器约210万元） |
| **部署时间** | 分钟级响应（如劲速云5000卡GPU集群扩容仅需90秒） | 需6-18个月建设周期（如中国电信甘肃庆阳项目需380天） |
| **管理复杂度** | 由提供商负责（如腾讯云提供7×24小时驻场技术支持） | 需自建团队 |
| **可靠性** | 高SLA保障（如阿里云单实例可用性99.975%，跨可用区99.995%） | 自主控制（如MTN尼日利亚Tier III数据中心提供4.5MW冗余供电） |
| **扩展性** | 即时扩容（如阿里云支持分钟级创建万台实例） | 扩展周期长（需6-12个月新增机柜） |
| **能源效率** | 提供商优化（如贵安美的云数据中心PUE低至1.17） | 需符合政策要求（2025年北京存量IDC PUE需≤1.35） |
| **技术支持** | 专业团队服务（如端脑云提供24×7技术支持） | 依赖内部团队（需50-100人运维团队） |
| **地理位置** | 多区域覆盖（如阿里云全球28个地域） | 本地化部署（如中国移动宁夏中卫数据中心服务区域需求） |

资料来源：劲速运算力，CSDN，数据中心之家，公司公告，国信证券经济研究所理

# 微软与CoreWeave加强合作，CSP自建转租赁或为产业趋势

**算力租赁具备成本效益高、灵活性高、技术更新及时的优势，能够应对算力短缺。**相较于自建，算力租赁具备较强的成本和灵活性优势，同时得益于有专业团队进行维护和技术迭代，GPU集群的利用率有望维持较高水平。

**CSP自建转租赁或为产业趋势：CSP巨头之一微软为CoreWeave第一大客户，算力租赁满足其快速增长的AI算力需求。**微软的Azure云服务在AI领域的需求激增，尤其是在训练和推理大规模AI模型方面。微软考虑到CoreWeave作为一家专注于高性能GPU云服务的公司，能够提供大量的英伟达GPU资源，更高效地支持微软的AI项目。同时微软无需自行大规模投资建设数据中心，不仅优化了成本，还提高了资源利用效率。

表：算力租赁优势

| | 核心优势 |
|---|---|
| **成本效益** | 算力租赁可以降低企业的初始投资成本，避免大规模的资本支出，尤其对于中小企业和初创公司来说更具吸引力 |
| **灵活性** | 租赁模式可根据企业实际需求灵活调整算力资源配置，更好地应对业务的波动 |
| **技术更新** | 租赁服务提供商通常会及时更新硬件设备，企业更容易地获取最新的技术 |
| **应对短缺** | 在高端GPU供应受限的情况下，租赁成为获取算力资源的有效途径 |

资料来源：华纳云，国信证券经济研究所整理

图：微软与CoreWeave加强合作


Microsoft to spend $10bn on CoreWeave by end of decade

Microsoft will be both competitor and customer of the AI cloud provider

November 04, 2024 By: Charlotte Trueman Comment

资料来源：datacenterdynamics，国信证券经济研究所整理

# 第三章 海外算力租赁：新GPU云厂商（Coreweave等）与英伟达深度合作，算力租赁市场快速增长

# CoreWeave起源于加密货币挖矿公司，转型为AI算力服务商

国信证券 GUOSEN SECURITIES

**CoreWeave 前身为加密货币挖矿公司，随着AI浪潮转型为全球领先的 AI 算力基础设施服务商。**CoreWeave崛起与 AI 革命深度绑定，凭借独特的资源整合能力和技术优势，CoreWeave成为全球 GPU 云服务市场的重要参与者。公司主营业务是提供专为人工智能（AI）打造的 CoreWeave Cloud Platform，专注于为 AI 模型训练、推理等高性能计算需求提供优化的算力、网络和存储支持，是定位为 “AI Hyperscaler” 的云计算服务商。**CoreWeave通过股权和债务融资筹集超过120亿美元**，投资方包括Magnetar Capital、黑石、英伟达、摩根大通、高盛、摩根士丹利等。2025年3月底CoreWeave成功IPO。

图：CoreWeave发展历程

**2017**

- **前身为加密货币挖矿公司Atlantic Crypto。**成立于2017 年，由三位交易员创立，专注于以太坊挖矿——CoreWeave的起点

**2018**

- **低价收购大量GPU。**2018 年加密货币寒冬，大量挖矿公司破产，GPU低价抛售，公司低价收购大量二手GPU，构建算力网络，果断转型为 AI 算力服务商

**2019**

- **更名为CoreWeave，转型为提供高性能计算的数据中心GPU云服务商。**大量GPU资源提供了强大的计算能力，满足客户的高性能计算需求

**2020**

- 成为**英伟达**的首选云服务提供商之一

**2022**

- **ChatGPT使各科技公司对英伟达GPU的需求呈爆发式增长。**CoreWeave凭借其GPU资源，为客户提供强大的算力支持，满足AI训练和推理方面的要求

**2023年初**

- **鱼鹰计划Project Osprey：**英伟达承诺在四年内向CoreWeave支付13亿美元，租回自己生产的芯片，持续到2027年8月。英伟达成为仅次于微软的CoreWeave第二大客户

**2023年6月**

- **微软**与CoreWeave签署AI算力协议，用于计算基础设施，未来数年内的投资金额可能数十亿美元

**2024**

- **2023-2024 年，公司营业收入从2.29亿美元飙升至19.15亿美元，增速高达736%。**公司的客户群体也涵盖了微软、Meta、OpenAI等众多科技巨头

**2025**

- **成功IPO：**3 月 28 日（北京时间）在纳斯达克上市，首日开盘破发，市值约 189 亿美元。3 个月内股价飙升近 300%，跻身纳斯达克综合指数上市以来表现最佳的前 30 只股票，巅峰时期市值近 730 亿美元

资料来源：公司官网，国信证券经济研究所整理

# CoreWeave机构股东融合技术巨头与顶级投资机构

国信证券 GUOSEN SECURITIES

**CoreWeave 的股权架构采用典型的“创始人控制型双类股结构”。**其中Class A 普通股：每股1票，可公开交易；Class B普通股：每股10票，不可公开交易，仅创始人及特定股东持有，可随时转换为A类股（无对价）。B 类股由创始人持有，赋予创始人绝对控制权，确保创始团队在套现后仍掌握绝对话语权，而财务投资者（如英伟达）虽持股较多，但投票权被大幅稀释。

**机构股东融合了技术巨头与顶级投资机构，反映出AI算力租赁公司较强的市场吸引力。**NVIDIA作为CoreWeave的核心GPU供应商，通过投资确保CoreWeave优先获取H100、H200、GB300等稀缺芯片，**实现供应链绑定。**Magnetar作为最大外部股东，提供股权融资和债务融资，表现出传统资本市场对 海外AI算力租赁赛道的长期认可。

表：机构股东架构

| 股东名称 | 股份类型 | 持股比例 | 投票权比例 | 特殊权利/角色 |
|---|---|---|---|---|
| **Magnetar Capital** | Class A | 0.3454 | 0.0723 | 最大外部机构股东（B轮领投+债务融资主导） |
| **Fidelity** | Class A | 0.0756 | 0.0153 | 财务投资，投票权受双类股结构稀释 |
| **NVIDIA** | Class A | 0.0597 | 0.0121 | 战略供应商（提供90%以上GPU）+客户，通过B轮融资获得A类股，无特殊权利 |
| **Cisco** | Class A | <1% | <0.2% | 网络设备供应商，战略协同（招股书仅提及其为“客户” |
| **Jane Street** | Class A | <1% | <0.2% | 量化交易客户+财务投资，无特殊权利 |

资料来源：CoreWeave招股说明书，国信证券经济研究所理

表：个人股东

| 股东名称 | 股份类型 | 持股比例 | 投票权比例 | 特殊权利/角色 |
|---|---|---|---|---|
| **Michael Intrator** | Class B | 0.4733 | 0.3831 | CEO |
| **Brian Venturo** | Class B | 0.3161 | 0.2545 | 首席策略官 |
| **Brannin McBee** | Class B | 0.2422 | 0.1938 | 首席开发官 |

资料来源：CoreWeave招股说明书，国信证券经济研究所理

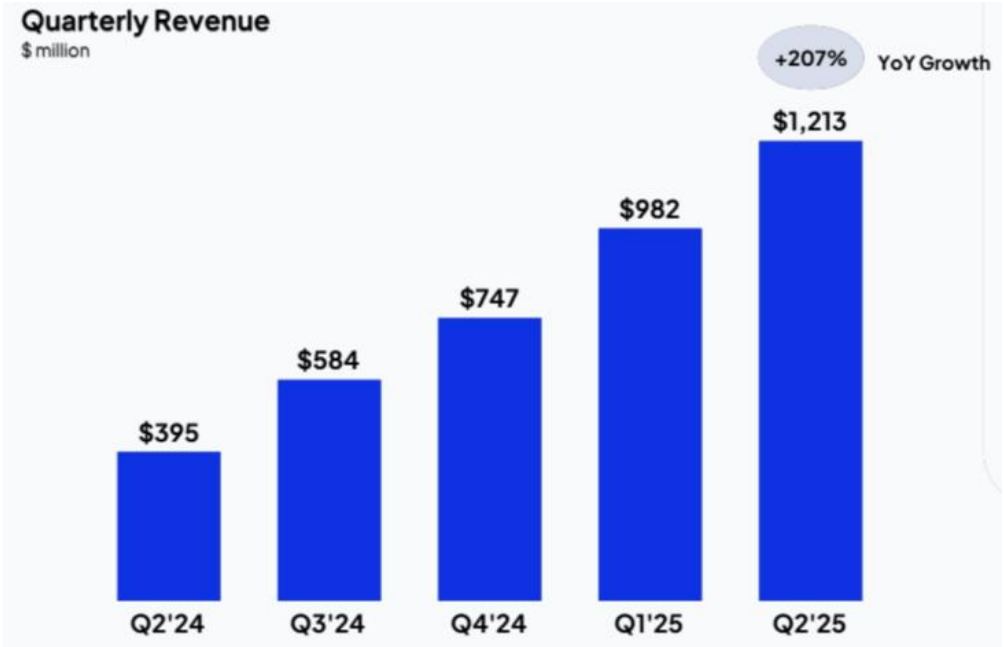
# CoreWeave处于高速扩张期，Revenue Backlog达301亿美元

国信证券 GUOSEN SECURITIES

**CoreWeave营收呈爆发式增长，公司处于高速扩张期。**随着AIGC的快速发展下对高端算力需求持续增加，CoreWeave仍处于高速扩张期，2022年-2024年公司营业收入从0.16亿美元增长至19.15亿美元，2023 /2024年同比增速分别为 1346%/737% 。季度来看，24Q1公司营业收入为1.89亿美元，此后逐季环比增长，25Q1增长至9.82亿美元，同比+420%；25Q2营收环比+23.55%至12.13亿美元。但CoreWeave 归母净利润持续亏损，目前在费用控制方面面临较大挑战。

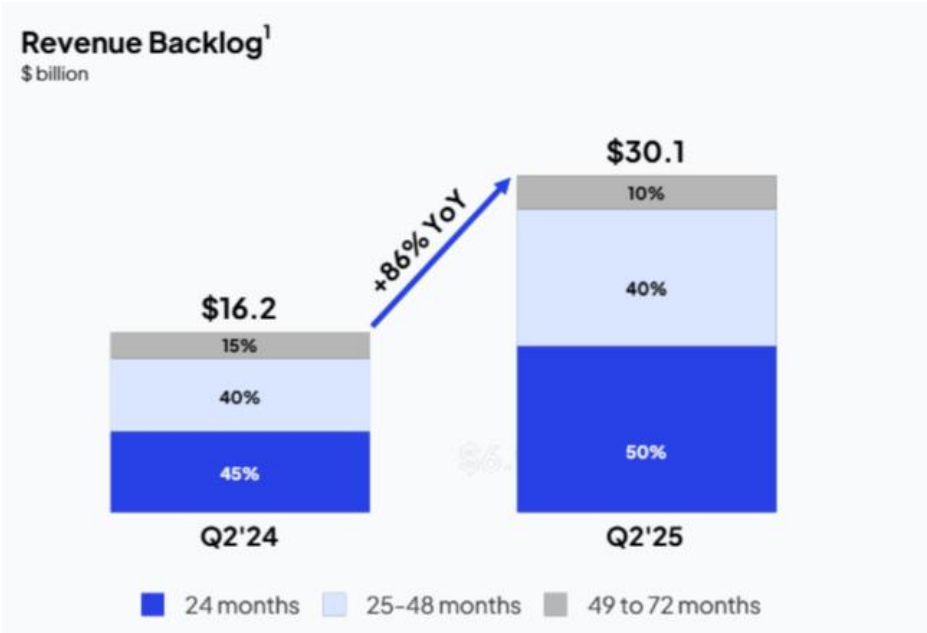
**公司在手订单充裕，截至25Q2已增至301亿美元。**管理层指出，2025Q2的backlog已达301亿美元（同比+86%，环比+16%），同时强调了当前市场需求依旧远超供给。公司新增了来自OpenAI的40亿美元追加订单。同时，2025财年收入预期被上调2.5亿美元，至51.5–53.5亿美元。

图：CoreWeave公司营业收入




资料来源：CoreWeave招股说明书，国信证券经济研究所整理

图：CoreWeave公司订单情况




资料来源：CoreWeave招股说明书，国信证券经济研究所整理

# CoreWeave下游客户集中度高，微软是第一大客户

通过租赁AI算力，CoreWeave的客户众多，包括微软、Meta 、NVIDIA 等全球领先的大型科技公司，以及众多明星AI初创公司。

**客户集中度高，微软是最大客户。**2023年6月，微软与CoreWeave签署AI算力协议，用于计算基础设施，未来数年内的投资金额可能数十亿美元。CoreWeave招股说明书显示，微软是 CoreWeave 的最大客户，2023 年贡献了 35% 的营收，2024 年这一比例上升至 62%。

**英伟达是仅次于微软的CoreWeave第二大客户，2024年为其贡献了15%的营收。**英伟达与CoreWeave签订反向租赁协议，命名为“鱼鹰计划Project Osprey”，英伟达承诺在四年内向CoreWeave支付13亿美元，租回自己生产的芯片，合约将持续到2027年8月。英伟达不仅是CoreWeave主要供应商，同时也是其重要客户，此协议不仅帮助CoreWeave缓解资金压力，而且确保芯片流向可控。

**OpenAI新合同填补空白，CoreWeave逐步减少对英伟达和微软的依赖。**2025年3月，CoreWeave与OpenAI达成119亿美元的五年合同，这将帮助其每年增加约20亿美元收入，CoreWeave将为OpenAI提供计算能力，用于训练和运行其人工智能模型。此协议对OpenAI来讲，有利于摆脱对微软的依赖，寻求计算能力多元化；对CoreWeave来讲，用长期合同锁定客户路径，有利于减少对英伟达和微软的依赖。

图：2022-2024年CoreWeave收入占比超过10%的主要客户列表

| | Year Ended December 31, | | |
|---|---|---|---|
| | 2022 | 2023 | 2024 |
| Customer A | * | 35% | 62% |
| Customer B | * | 21% | * |
| Customer C | * | 17% | 15% |
| Customer D | 16% | * | * |
| Customer E | 13% | * | * |
| Customer F | 12% | * | * |
| Customer G | * | * | * |

* Customer did not represent 10% or more of revenue

资料来源：CoreWeave招股说明书，国信证券经济研究所整理

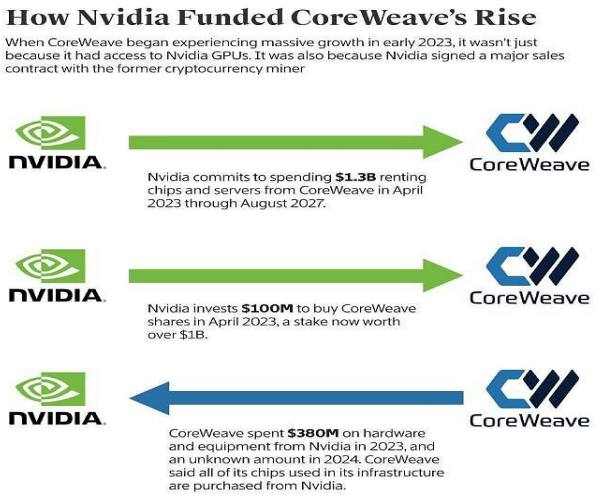
# CoreWeave与英伟达深度绑定及战略协同开发

**CoreWeave拥有英伟达AI 硬件的优先获取权，具备极强的先发优势。**2023年，英伟达向CoreWeave投资1亿美元，并向其优先供应数十万台高端GPU。同年8月，CoreWeave将英伟达GPU作为抵押品，获得了另外23亿美元债务融资。2025年7月3日，CoreWeave宣布其成为全球首家部署 NVIDIA 最新 GB300 NVL72 系统的云服务提供商。这套基于 Blackwell Ultra 芯片的 AI 设备在某些工作负载上比前代产品快 50%，每个系统整合了 72 颗 Blackwell Ultra 芯片、36 个 NVIDIA Grace 处理器以及 18 个 BlueField-3 数据处理单元，代表了当前 AI 硬件的最高水准。

**英伟达与CoreWeave进行战略协同开发,构建技术护城河。**英伟达向CoreWeave提供定制版CUDA、专用优化芯片甚至专属的软件堆栈，共同优化CUDA架构在云计算环境的运行效率，确保客户能够发挥出每台GPU的最大效能。这种深度绑定的合作，使CoreWeave在高性能GPU的供应和技术支持方面获得了优先级，从GPU采购、数据中心优化，到AI模型部署，向客户提供一站式解决方案。

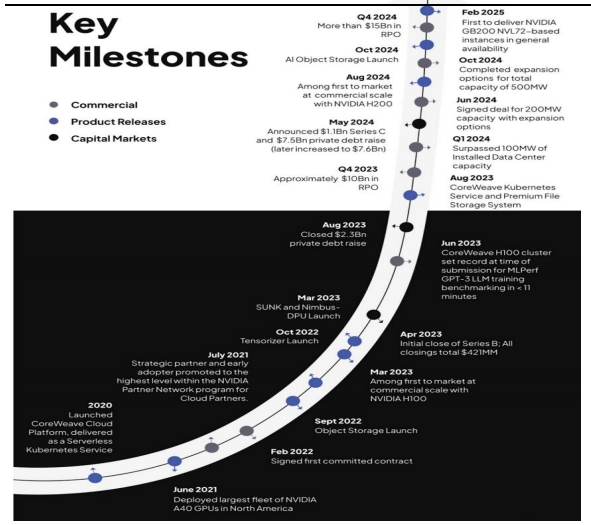
**英伟达与CoreWeave签订反向租赁协议——鱼鹰计划Project Osprey，有利于CoreWeave缓解资金压力，且确保芯片流向可控。**2023年初，英伟达承诺在四年内向CoreWeave支付13亿美元，租回自己生产的芯片，在CoreWeave提交的IPO文件中将其命名为鱼鹰计划”Project Osprey”。这使英伟达成为仅次于微软的CoreWeave第二大客户，2024年为其贡献了15%的营收，英伟达租回自己芯片的合约将持续到2027年8月。

图：CoreWeave与英伟达的合作




资料来源：the Information，国信证券经济研究所整理

图：CoreWeave获得英伟达GPU时间线




资料来源：CoreWeave招股说明书，国信证券经济研究所整理

图：CoreWeave部署 NVIDIA 最新 GB300 NVL72 系统）


资料来源：CoreWeave，国信证券经济研究所整理

# CoreWeave算力资源丰富，GPU利用率较云计算大厂具备优势

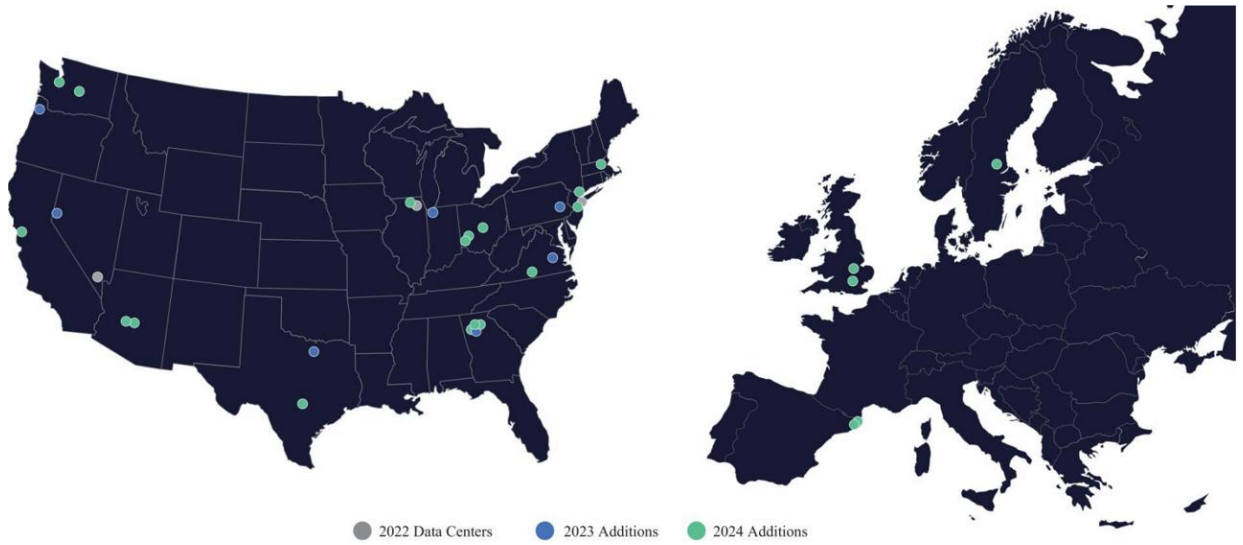
国信证券 GUOSEN SECURITIES

**CoreWeave拥有数十万英伟达GPU，算力资源丰富。**根据招股说明书，截至2024年12月，CoreWeave总共拥有32 个AI 数据中心,25万个英伟达GPU，由超过360 兆瓦的有功功率提供支持，未来计划继续扩大数据中心版图。截至2024年12月31日，总签约电力1.3吉瓦（GW），预计将在未来几年逐步部署。

**CoreWeave通过技术提高GPU利用率，相比云计算大厂GPU 利用率提升20%。**CoreWeave通过自研SUNK 协调系统以及Tensorizer 优化工具，最大限度地提高客户的 GPU 利用率。CoreWeave认为，GPU的计算效率在峰值理论性能的35%到45%之间（这也是大部分云计算大厂的平均水平），若使约 35% 的实际效率与理论上 100% 的效率之间的效率缺口降至最低，将大幅释放AI基础设施的性能潜力。经过技术优化，现已将GPU 利用率提升20%，超过云计算大厂的GPU 利用率。

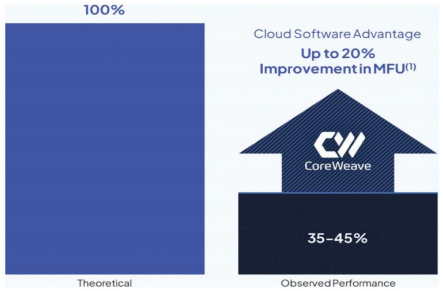
**计划收购 Core Scientific， 扩大基础设施规模，有助于提高在云服务领域的竞争力。**Core Scientific 运营着 1.3 吉瓦的数据中心网络，其中约 40% 的容量仍在进行加密货币挖矿。CoreWeave计划在交易完成后，出售这部分基础设施或者将其改造为 AI 工作负载设施，有利于在AI 云服务市场中建立更强优势。

图：CoreWeave数据中心分布网




资料来源：CoreWeave招股说明书，国信证券经济研究所整理

图：CoreWeave观测性能




资料来源：CoreWeave招股说明书，国信证券经济研究所整理

图：CoreWeave计划收购 Core Scientific


EXCLUSIVE DEALS

CoreWeave in Talks to Buy Core Scientific

A bid last year was rejected by Core Scientific as too low

By Lauren Thomas Follow and Ben Dummett Follow

Updated June 26, 2025 12:58 pm ET

资料来源：CoreWeave，国信证券经济研究所整理

# Nebius——全栈AI原生云平台领导者，融资加速发展

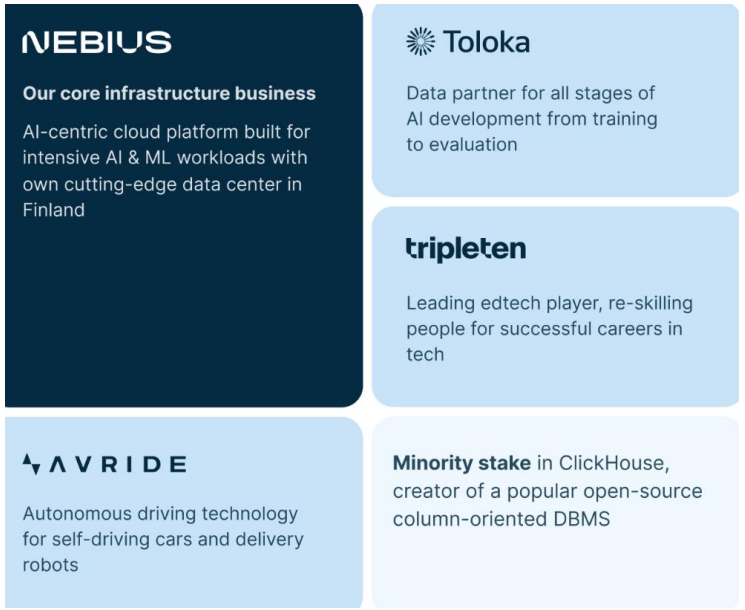
**Nebius是全栈AI原生云平台领导者，提供大规模GPU集群和专为AI与机器学习高强度工作负载优化的基础设施。**Nebius由俄罗斯互联网巨头Yandex拆分而来。2024年7月，Yandex将其俄罗斯业务出售，保留了在俄罗斯境外的业务，并更名为Nebius Group N.V.，总部设在荷兰阿姆斯特丹。Nebius研发中心遍布欧洲、北美和以色列，为全球人工智能行业的爆炸性增长提供服务，包括大规模GPU集群、人工智能原生云平台以及面向开发人员的工具和服务。Nebius集团业务包括nebius.ai GPU 云服务、Toloka数据标注公司、TripleTen IT技能培训、Avride自动驾驶与机器人研发。

**Nebius超额认购7亿美元战略股权融资，加速推出全栈人工智能基础设施。**2024 年12月2 日，该公司宣布完成一轮7亿美元的融资，风险投资公司 Accel、英伟达以及 Orbis Investments 等机构参与。根据计划，Nebius将发行33,333,334股A类股，每股价格为21.00美元。融资所得将用于进一步构建大规模 GPU 集群、扩展云平台以及为开发者提供更多工具和服务，全面支持全球 AI 先驱的创新发展。

图：Nebius公司


资料来源：Nebius官网，国信证券经济研究所整理

图：Nebius集团主要业务




资料来源：Nebius官网，国信证券经济研究所整理

图：Nebius宣布完成一轮7亿美元的融资


资料来源：Nebius官网，国信证券经济研究所整理

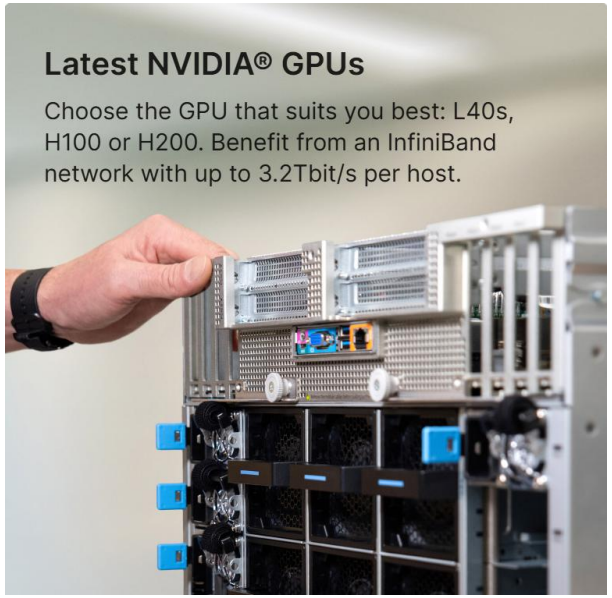
# Nebius与英伟达合作密切，GPU部署全球扩张

国信证券 GUOSEN SECURITIES

**Nebius与英伟达合作密切，具有英伟达GPU优先供应权，是第一个总部位于欧洲的NVIDIA 云合作伙伴。1）Nebius是首批提供NVIDIA Blackwell和Blackwell Ultra的GPU云提供商之一。**2025年三月，Nebius宣布成为NVIDIA Blackwell Ultra平台的早期采用者云提供商，从第二季度开始在美国数据中心全面推出NVIDIA Blackwell GPU，同时被任命为new NVIDIA Dynamo合作伙伴。**2）Nebius是全球为数不多的几家拥有参考平台 NVIDIA 云合作伙伴地位的公司之一，**在设计和部署 NVIDIA 参考架构的全栈硬件和软件基础设施方面具有专业知识。Nebius计划到 2025 年年中，在人工智能基础设施方面投资超过10亿美元，部署数以万计的NVIDIA GPU，成为英伟达在欧洲的 “战略支点”。

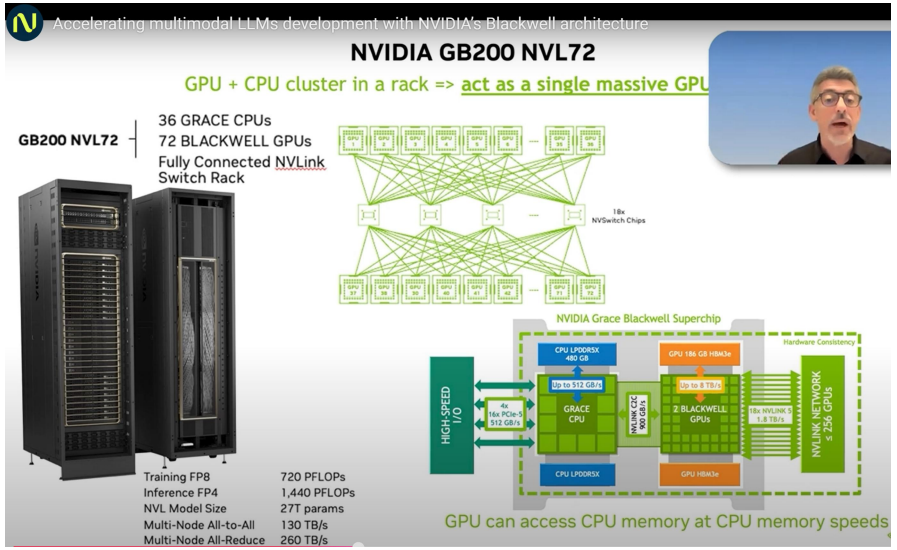
**Nebius算力布局广泛，在美国、英国、冰岛、芬兰等地部署数据中心，快速推动全球扩张。**2024年11月，Nebius宣布在美国堪萨斯城部署首个NVIDIA GPU集群，可容纳约35,000个GPU。2025年3月，Nebius计划在美国新泽西州建设容量高达300MW的新数据中心，同时确认在冰岛凯夫拉维克部署新的托管。2024年10月，该公司将芬兰数据中心的容量增加两倍，达到75MW。2025年6月，Nebius在英国部署GPU集群，预计将于该年第四季度在英国部署数千台最先进的NVIDIA Blackwell Ultra GPU，为2025年美国和欧洲的装机容量做出重大贡献。

图：Nebius可预定最新NVIDIA GPU集群



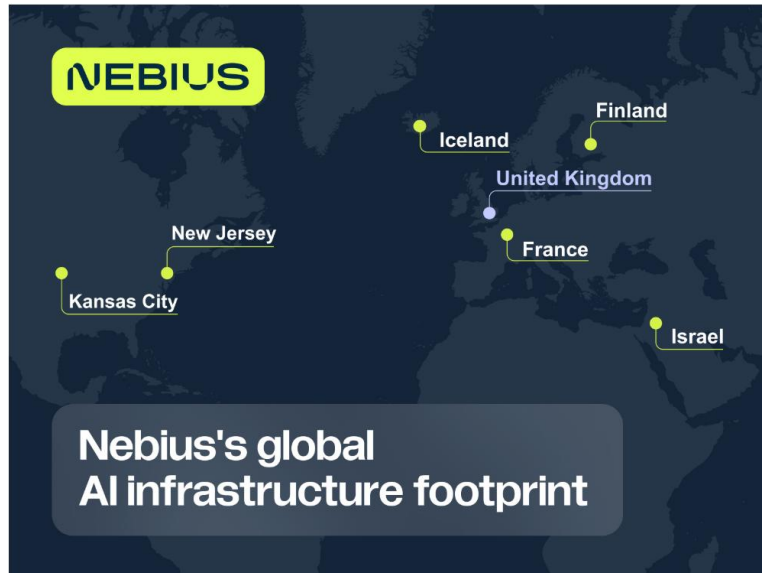

资料来源：Nebius官网，国信证券经济研究所整理

图：Nebius使用NVIDIA的Blackwell架构加速多模态LLM开发




资料来源：Nebius官网，国信证券经济研究所整理

图：Nebius全球数据中心布局



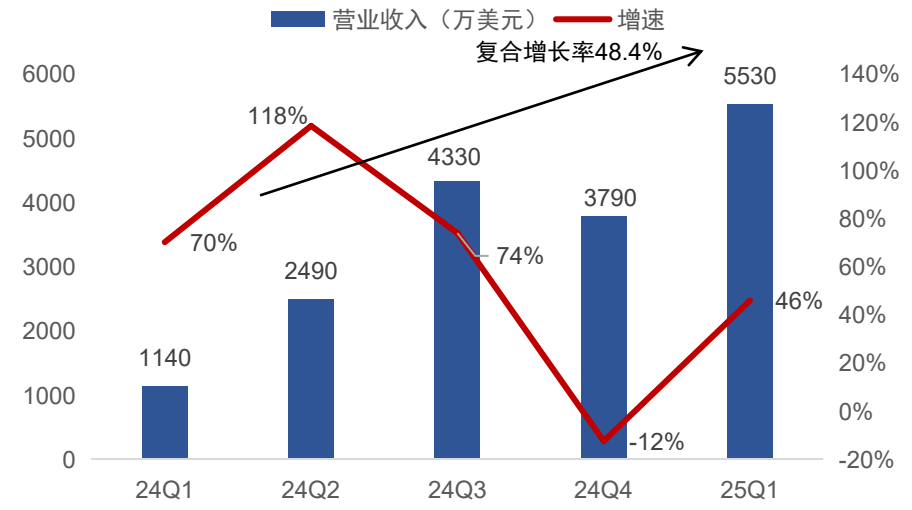

资料来源：Nebius官网，国信证券经济研究所整理

# Nebius处于高速扩张期，2025Q2营收同比+625%

**Nebius营收呈爆发式增长，公司处于高速扩张期。**Nebius公司于2024年7月正式脱离Yandex公司，从2024Q1开始，由于核心AI基础设施业务驱动，公司营业收入高速增长。24年Q1营业收入为1,140万美元，25年Q1为5,530万美元，同比增长385%，2025Q2环比+90%至1.05亿美元。公司核心业务是为密集型AI工作负载打造的 AI 云平台，通过自研的云软件架构和硬件，为 AI 开发者提供计算、存储、托管服务等关键资源，该业务的竞争力和市场需求支撑了收入的快速增长。

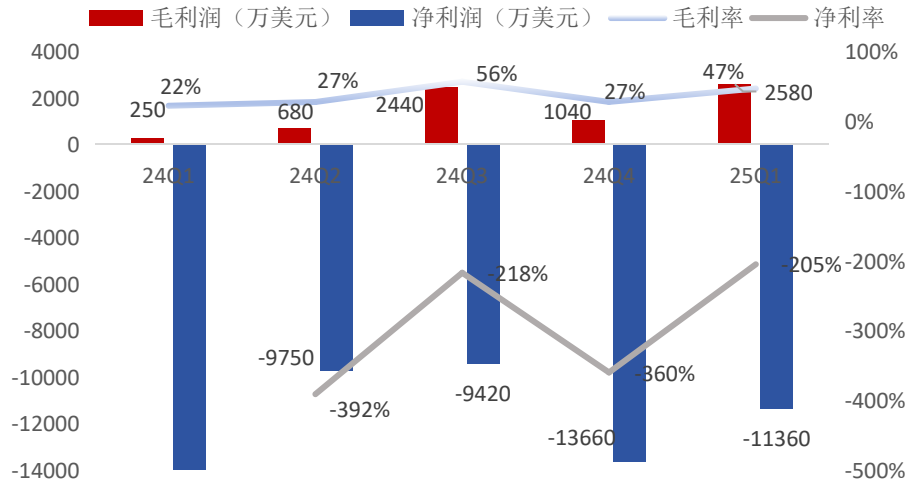
**毛利率阶段性波动增长，25Q2净利润大幅扭亏。公司毛利率阶段性增长，盈利能力有所提升。公司**24Q1毛利率为 22%，此后波动增长，25Q2增至71.36%。**公司持续亏损，净利率波动较大。**25Q2净利率大幅扭亏主要系公司将旗下Toloka的权益转为按权益法核算，非经常收益确收增长所致，不具备参考意义。参考25Q1，公司同比亏损幅度有所收窄，但仍处于亏损状态，主要系1）公司进行高研发与基础设施投入，相关资本性支出及运营成本对利润形成阶段性压制。2）市场扩张与销售团队搭建，拉低净利率水平。3）收入规模仍处于快速上升阶段，尚未形成足够的规模效应以覆盖固定成本，导致净利率承压。3）**未来GPU集群有望形成规模效应，改善利润。**Nebius已在美国、法国部署GPU集群，计划 2025年在美国和芬兰的数据中心部署超22,000个NVIDIA Blackwell GPU，这些硬件扩容将提升算力服务能力。公司预计2025 年12月的年化营收达到7.5亿至10亿美元。

图：Nebius公司营业收入




资料来源：Nebius Earning call，国信证券经济研究所整理

图：Nebius公司净利润与毛利润




资料来源：Nebius Earning call，国信证券经济研究所整理（24Q1净利润-3.17亿美元，净利率-2776%）

# Omniva起源于加密货币挖矿数据中心，转型为AI云服务供应商

国信证券 GUOSEN SECURITIES

**Omniva是AI基础设施与GPU云服务供应商，由大型加密货币挖矿数据中心转型而来。**Omniva于2014年在科威特注册，前身是Moneta United Technologies。母公司是科威特企业集团KMGC，业务覆盖石油、房地产等传统领域，为其提供雄厚的资金支持。Moneta早期建立大型加密货币挖矿数据中心，2023年，随着ChatGPT的迅速崛起，公司更名为Omniva，将业务转向AI云服务，对标谷歌、微软等巨头。

**Omniva引入多位科技巨头前高管，预期对标大型科技厂商。**自2023年起，该公司聘请了前AWS首席财务官Sean Boyle、前微软副总裁Kushagra Vaid、Meta基础设施副总裁T.S. Khurana，此外，亚马逊Project Kuiper的前副总裁Tyson Lamoreaux和台积电Somnuk Ratanaphanyarat也加入Omniva，强化云服务和数据中心运营经验，为高级管理层带来了急需的数据中心专业知识。

图：Omniva数据中心


资料来源：Omniva官网，国信证券经济研究所整理

表：2023年起引入多位科技巨头前高管

| 姓名 | 前职位 | 职责方向 |
|---|---|---|
| **Sean Boyle** | AWS 首席财务官 | 财务战略、资本运作 |
| **Kushagra Vaid** | 微软 Azure 杰出工程师 | 服务器架构、GPU 集群设计 |
| **T.S. Khurana** | Meta 基础设施副总裁 | 超算集群运维、AI 模型部署 |

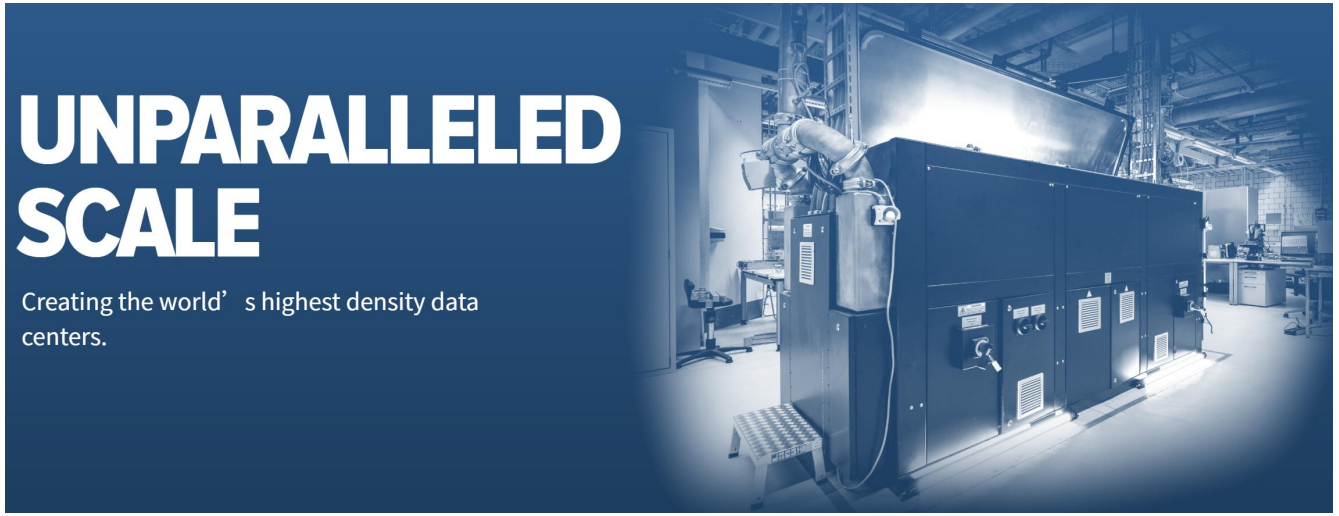
资料来源：半导体产业纵横，国信证券经济研究所整理

# Omniva主动寻求渠道合作、提高技术，支持超大规模算力部署

**Omniva与纬颖再次讨论GB200 NVL72订单，Omniva在高端GPU渠道合作中具有主动性与稳定性。**2025年年初，Omniva与纬颖讨论GB200 NVL72订单，预计在25Q2–25Q3开始量产且首批订单约1,000柜。GB200 NVL72作为NVIDIA面向超大规模AI训练的旗舰级系统，是全球AI算力竞争的核心资源，1,000柜的部署量将形成支撑大规模AI集群（如多模态大模型训练、自动驾驶仿真）的核心算力池，匹配中东地区快速增长的 AI 算力需求。该渠道合作将使Omniva的算力资源既具备 “量”的优势，又具备处理尖端 AI 任务的“质”的领先性。

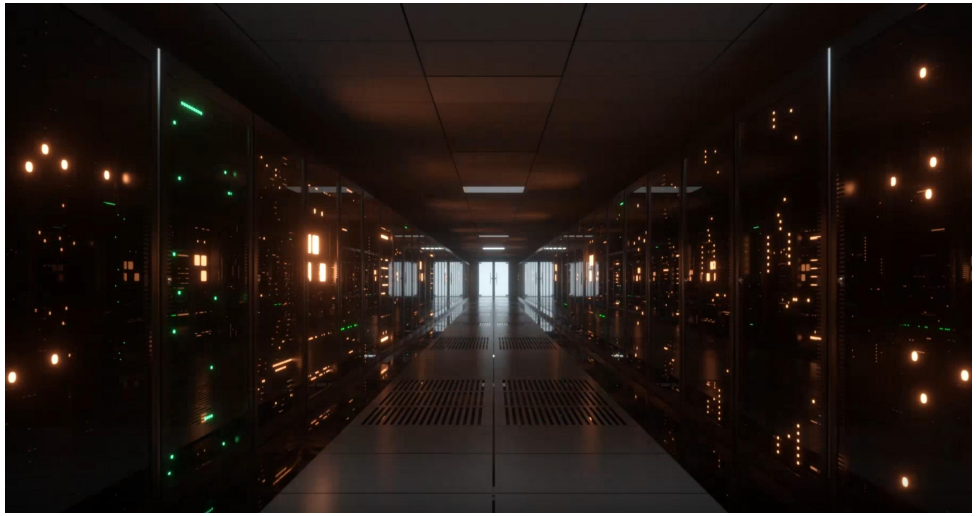
**Omniva数据中心采用“雅典娜Athena”沉浸式冷却系统，依托3M氟化液实现高效散热，支撑超大规模AI算力集群部署。**雅典娜系统基于3M Novec HFE-7100氟化液构建两相蒸发 / 冷凝自然流动散热架构。规划含 1,650 个服务器水槽，总设计功率864MW，占地面积 40,000 平方米。单个服务器水槽支持30张 GPU 部署，单槽 IT 功率密度达500kW，适配高密度 AI 算力需求，支撑 AI 云业务。

图：Omniva公司超大规模数据中心




资料来源：Omniva官网，国信证券经济研究所整理

图：Omniva公司数据中心



资料来源：Omniva官网，国信证券经济研究所整理

# 海外GPU云厂商：英伟达战略为主导，积极开拓全球市场

国信证券 GUOSEN SECURITIES

**通过对比CoreWeave、Nebius、Omniva三家海外GPU云厂商，可以发现部分共通之处，或为实现GPU云的必备要求：**

1、深度绑定英伟达或其ODM厂商，具备一定的优先供应权。Omniva与台企纬颖建立合作关系，间接获得英伟达优先供应权。

2、尽管区域侧重点不同，但是均积极投入开拓全球市场，组网能力较强，预期算力规模持续增长。

表：CoreWeave、Nebius、Omniva对比

| | CoreWeave | Nebius | Omniva |
|---|---|---|---|
| **转向算力租赁时间** | 2019年 | 2024年 | 2023年 |
| **公司区域战略** | 从美国本土向欧洲复制 | 先欧洲、再美/中东 | 聚焦于中东和欧洲 |
| **是否具有英伟达GPU优先供应权** | 有 | 有 | 间接 |
| **英伟达有无持股** | 有，占 Class A 股份总数的 5.97%，并拥有1.21% 的投票权 | 有 | 无 |
| **获得GB200/300方式** | 英伟达 | 英伟达 | 台企 |
| **最新算力规模** | 33个AI数据中心，470MW有功功率，总签约电力2.2GW（截至25Q2） | 预计拥有220MW的连接电力，其中100MW为可用状态（预计截至25Q4） | 与纬颖讨论GB200 NVL72订单，预计在25Q2–25Q3开始量产且首批订单约1,000柜（约72MW） |

资料来源：Coreweave官网，Nebius官网，Omniva官网，国信证券经济研究所整理

# 第四章 国内算力租赁：我国高端算力需求旺盛，国内算力租赁市场逐步打开

# 我国算力租赁市场未来3年规模复合增长率(CAGR)有望超50%

国信证券 GUOSEN SECURITIES

**受制于美国对中国的芯片禁令，中国算力成本上升，算力租赁以轻资产运营满足算力需求，市场规模有望持续扩张。**美国政府对中国的科技封锁不断加剧，从限制高端芯片出口到拟限制中国企业使用美国厂商的云计算服务，严重阻碍了中国获取先进算力芯片及相关技术的渠道，导致高端芯片短缺，算力成本上升。同时，AI大模型的快速发展使得对算力的需求呈指数级增长，训练和推理需要庞大的计算资源支持。自建算力成本高昂，不仅需要购买大量昂贵的硬件设备，还需承担后续的运维和管理成本，算力租赁能够兼顾成本和满足企业算力需求，2023年后算力租赁市场迅速兴起。2024年，中国智能算力租赁市场规模达到377EFlops（FP16精度），同比增长88%。预计未来三年，中国智能算力租赁市场规模复合增长率将保持在53%，至2027年，中国智能算力租赁市场规模将达1346EFlops。

图：特朗普政府撤销《人工智能扩散规则》

Bureau of Industry and Security

Search... Email notificat

Regulations Licensing Learn & Support News & Updates Enforcement About BIS

**Bureau of Industry & Security**
Office of Congressional and Public Affairs

**FOR IMMEDIATE RELEASE** | May 13, 2025 | Media Contact: OCPA@bis.doc.gov

**Department of Commerce Rescinds Biden-Era Artificial Intelligence Diffusion Rule, Strengthens Chip-Related Export Controls**

Download as PDF

**WASHINGTON, D.C.**— Today, the Department of Commerce (DOC) initiated a rescission of the Biden Administration's AI Diffusion Rule, while announcing additional steps to strengthen export controls on semiconductors worldwide.

The AI Diffusion Rule was issued on January 15, 2025, with compliance requirements that were set to come into effect on May 15, 2025. These new requirements would have stifled American innovation and saddled companies with burdensome new regulatory requirements. The AI Diffusion Rule also wou have undermined U.S. diplomatic relations with dozens of countries by downgrading them to second-tier status.

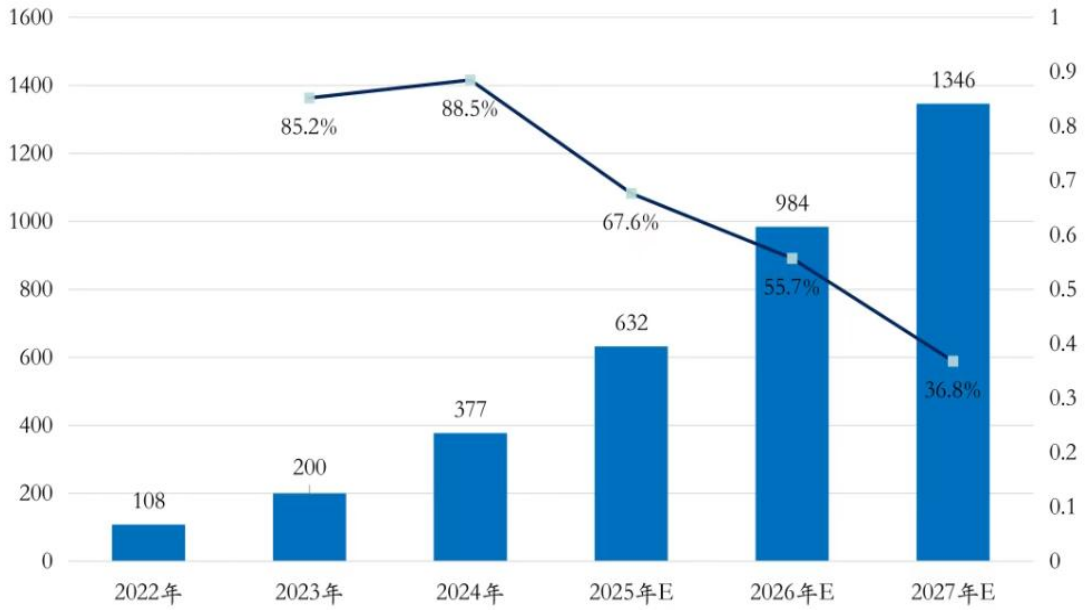
BIS plans to publish a Federal Register notice formalizing the rescission and will issue a replacement rule in the future.

Under Secretary of Commerce for Industry and Security Jeffery Kessler has instructed BIS enforcement officials not to enforce the Biden Administration's AI Diffusion Rule, stating:

"The Trump Administration will pursue a bold, inclusive strategy to American AI technology with trusted foreign countries around the world, while keeping the technology out of the hands of our adversaries. At the same time, we reject the Biden Administration's attempt to impose its own ill-conceived and counterproductive AI policies on the American people."

资料来源：BIS，芯智讯，国信证券经济研究所整理

图：中国智能算力租赁市场规模及预测（EFlops，FP16）



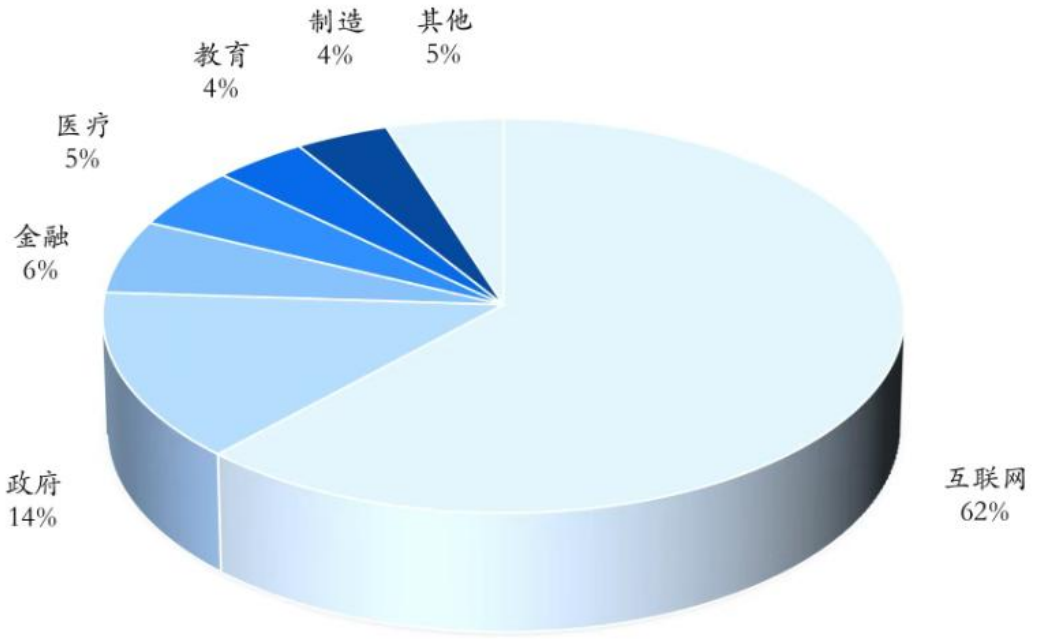

资料来源：科智咨询，国信证券经济研究所整理

# 下游行业集中，CSP互联网云厂占比超六成

**中国智能算力租赁市场下游行业集中，互联网行业以62%的绝对占比成为核心需求方。**互联网行业的需求主要集中于云计算、AI模型训练及实时推理等高算力场景，其对大规模弹性算力的依赖源于海量数据处理、短视频推荐算法优化以及生成式AI应用的爆发式增长。

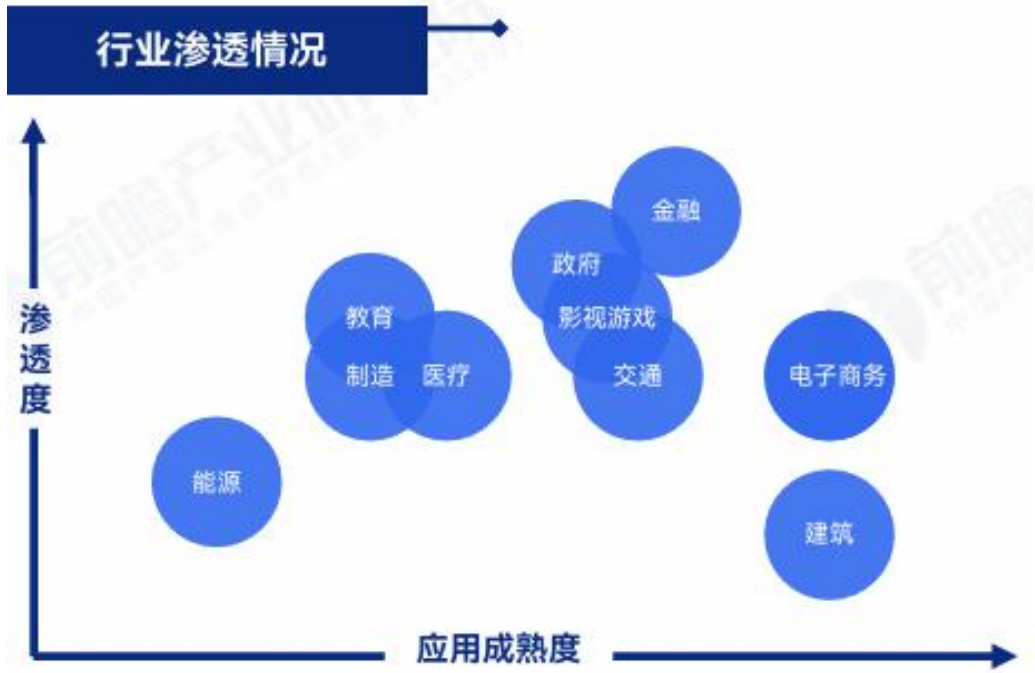
**AI大模型融合千行百业，金融、政务渗透率最高。**除互联网外，金融及政务的智能算力租赁需求较高。政府占比14%，其需求以智慧城市、政务云和公共安全为重心，偏向数据合规性及安全性，并倾向于采用私有化部署与国产化算力解决方案。金融行业占比6%，聚焦高频交易风控、量化建模及反欺诈分析，对低延迟算力和异构资源调度能力要求较高。另外，医疗行业占比5%，制造与教育行业各占4%。

图：中国智能算力租赁市场需求方




资料来源：科智咨询，国信证券经济研究所整理

图：AI大模型行业应用渗透情况




资料来源：前瞻产业研究院，国信证券经济研究所整理

# 国内CSP云厂CAPEX维持高位，OPEX呈增长态势

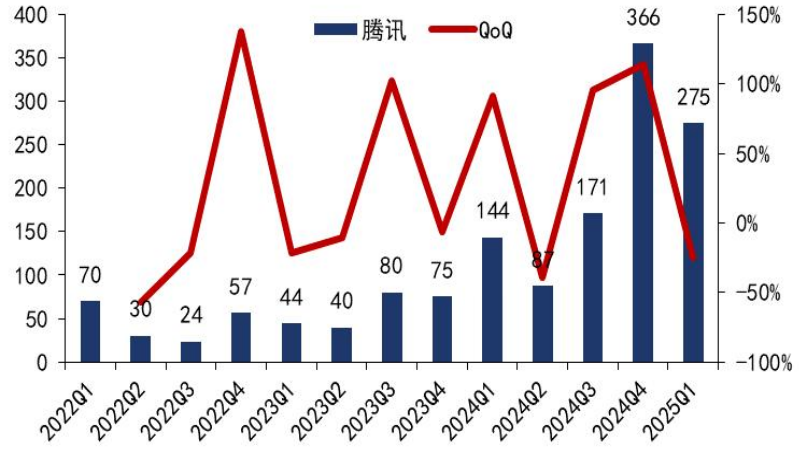
**字节跳动2025年CAPEX预期为1600亿元，字节和阿里均有望突破千亿。**今年年初阿里宣布计划未来三年将投入至少3800亿元人民币用于建设云计算和AI的基础设施，践行其对长期技术创新的承诺，凸显公司对AI驱动增长的聚焦。腾讯也计划在2025年进一步增加资本支出，并预计资本支出将占收入的低两位数百分比。腾讯2025年的Capex预计也将达到千亿级别。2025年5月互联网大厂阿里和腾讯分别发布财报，阿里2025Q1的Capex为246.12亿元，同比+120.68%，环比-22.54%；腾讯2025Q1的Capex同比+91%至为275亿元，但环比-24.93%，2025Q2环比-30%至191亿元。考虑到美国4月15日升级对华AI芯片的出口限制，无限期禁止英伟达H20对华出口，云厂商Capex的下滑极有可能与算力卡供应紧缺相关。

**大厂Opex呈现增长态势，算力租赁业务或成为互联网大厂进行AI投入的重要方向。**以腾讯最新财报为例，2025Q2公司OPEX在 Non-IFRS财报中呈现同比与环比双增长态势。以IFRS为标准情况下，S&M及R&D费用分别环比增长20%/7%，呈现较为明显的增长态势。阿里2025Q1的Opex同比增长明显，产品发展、S&M及R&D费用分别同比+5.7%/15%/4%。

图：阿里资本开支（亿元）


资料来源：wind，国信证券经济研究所整理

图：腾讯资本开支（亿元）




资料来源：wind，国信证券经济研究所整理

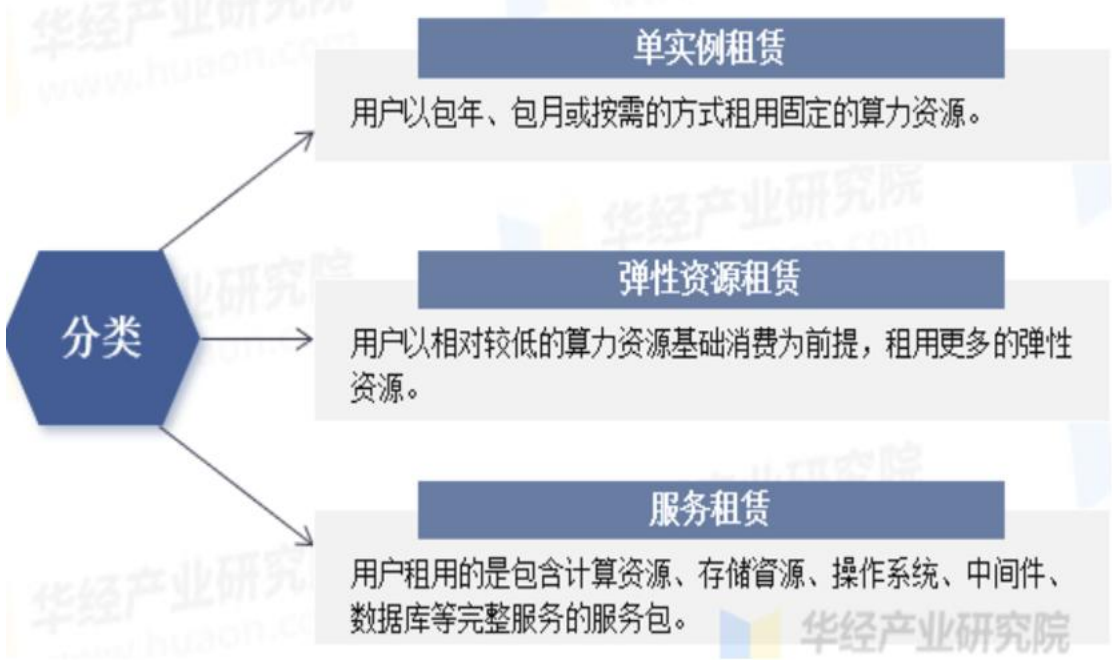
图：腾讯25Q2 OPEX情况


资料来源：公司官网，国信证券经济研究所整理

# 算力租赁灵活性高，市场价格已基本稳定

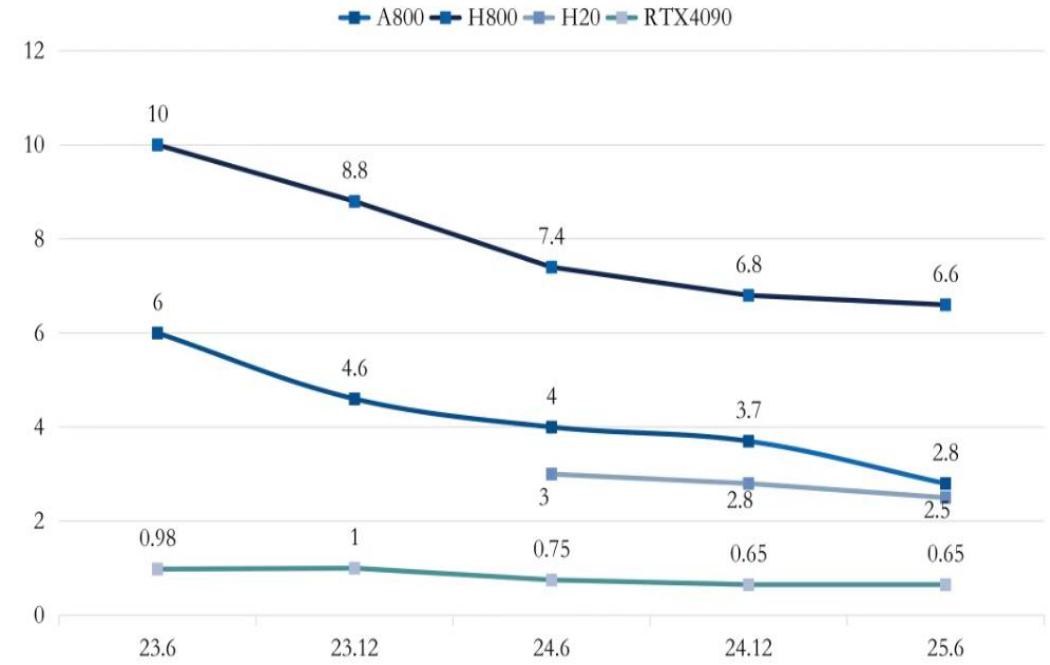
**中国智能算力租赁灵活性高，市场价格保持稳定或进一步降低。**算力租赁灵活性和资源利用率较高，我国算力租赁的模式主要包括单实例租赁、弹性资源租赁与服务租赁。根据科智咨询，当前中国智能算力租赁市场的主流AI服务器整体呈现出逐渐下降的趋势。其中，8卡服务器A800的租赁价格从2023年6月的约6万元/台/月，下降到2025年6月的约2.8万元/台/月，下降幅度53%；H800的租赁价格从2023年6月的约10万元/台/月，下降到2025年6月的约6.6万元/台/月，下降幅度34%；H20为2023年底推出的产品，由于其在推理场景的高适用性和超高的性价比，其租赁价格从2024年6月的约3万元/台/月，下降到2025年6月的约2.5万元/台/月，下降幅度仅为17%。

图：算力租赁行业分类




资料来源：华经产业研究院，国信证券经济研究所整理

图：中国智算服务器租赁价格走势




资料来源：科智咨询，国信证券经济研究所整理

# 当前国产算力芯片主要适配推理场景

国信证券 GUOSEN SECURITIES

**国产芯片持续突破，单卡性能仍存在代际鸿沟。1）华为昇腾系列持续提升。**2023年9月，华为发布昇腾910B芯片，算力为320TFlops(FP16)，是英伟达H20的近2倍。根据Lennart Heim分析，昇腾910C算力为800TFlops(FP16)，性能可达到H100的80%，实现中国AI产业新突破。据《华盛顿日报》，华为将推出昇腾910D，旨在对应英伟达的高端产品，即将进入交付阶段。**2）与海外芯片仍存在代际。**在芯片架构上，昇腾910C的芯片面积约比英伟达H100大60%，在架构效率方面与H100仍存在差距。显存技术上，H200搭载141GB HBM3显存，带宽4.8TB/S；昇腾910B仅64GB HBM2，带宽3.35TB/S，国内AI芯片仍需进一步突破。

**集群生态与海外仍有差距，我国AI芯片大模型集群训练性能大多不及英伟达A100/A800系列的50%。**北京智源人工智能研究院林咏华表示，我国AI芯片公司40余家，但市占率低于10%，各家AI芯片软件各异、生态相对分裂。当前中国AI芯片大模型集群训练性能，大多数不到英伟达A100/A800的50%。

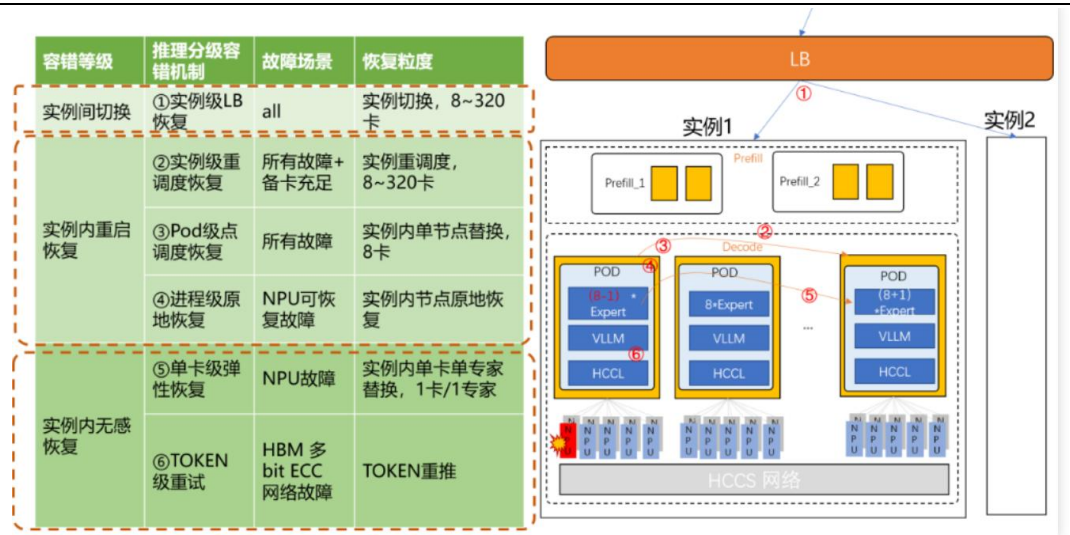
**国产芯片在推理场景的适配性更高，可在AI应用的“最后一公里”上发力，为推理场景提供算力支撑。**1）推理侧更注重能效比，国产AI芯片具有能效比优势。如华为昇腾910B，实算力达320 TFLOPS，功耗仅310W，PUE仅1.25，能效比是H20的3.2倍。2）推理侧对成本较为敏感，国内芯片成本控制较好。如训练千亿参数大模型，昇腾910C集群比英伟达H100节省38%时间，电费直降120万元。3）硬件架构优化：华为昇腾芯片针对超大规模混合专家模型（MoE）部署，从算子、模型和框架入手，开发一整套方案，在100ms时延约束下单卡吞吐达到808 Tokens/s，在推理性能上超越英伟达Hopper架构。

表：国产芯片与英伟达对比

| | 芯片 | 发布时间 | 浮点计算精度 | FP16（TFLOPS） | 功耗（W） |
|---|---|---|---|---|---|
| 华为 | 910A | 2019年 | FP16（TFlops） | 256 | 310 |
| | 910B | 2020年 | | 320 | 310 |
| | 910C | 2025年 | | 800（市场估计） | / |
| 寒武纪 | MLU590 | 2024年 | FP16（TFlops） | 314.6 | 550 |
| | MLU370-S4/S8 | 2022年 | | 72 | **75** |
| | MLU370-X4 | 2022年 | | 96 | **150** |
| | MLU370-X8 | 2022年 | | 96 | **250** |
| 摩尔线程 | MTT S4000 | 2023年 | FP32（TFlops） | 25 | **450** |
| | MTT S3000 | 2022年 | | 15 | **250** |
| | MTT S2000* | 2022年 | | 10.6 | **150** |
| 英伟达 | H100 | 2022年 | FP16（TFlops） | 2000 | **700** |

资料来源：各公司官网，国信证券经济研究所理

图：华为千亿MOE分布式推理分钟级恢复技术




资料来源：华为技术报告，国信证券经济研究所整理

# 英伟达高端AI芯片仍领先全球，主要适用于训练场景

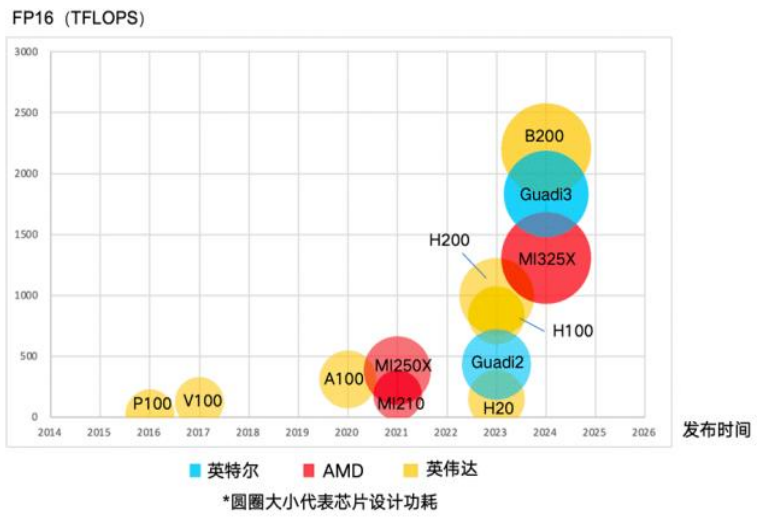
国信证券 GUOSEN SECURITIES

**不论是和北美CSP的ASIC相比，还是国产芯片对比，英伟达芯片的性能仍处于领先地位。**从性能看，英伟达的B200采用4NP制程，晶体管数量达2080亿，FP16算力2200 TFLOPS，远超AMD同期产品及国产芯片（如壁仞BR100 FP16算力1024 TFLOPS）。

**B300作为B200的迭代产品，其性能延续了升级趋势。**B300在制程优化、算力提升（尤其是FP8/FP16等AI核心算力）及互联效率上大概率进一步突破。在技术迭代与互联能力上，英伟达节奏更快，从2020年A100到2024年B200，四年内制程从7nm升级至4NP，NVLink5技术实现1800 GB/s双向互联带宽，远超AMD及国产芯片的互联性能，为大规模集群提供核心支撑。

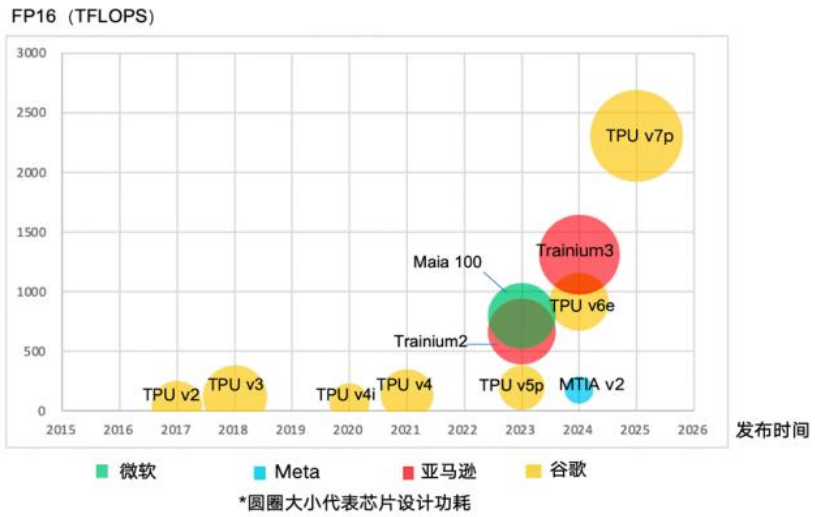
**生态层面，英伟达CUDA生态覆盖全球90%以上AI框架，迁移成本高达3000万元/企业，形成强壁垒。**英伟达或将继续拉开与其他芯片厂商的性能差距，巩固其在全球算力芯片领域的领先地位。

图：英伟达、英特尔、AMD芯片发布时间与设计功耗对比




资料来源：半导体综研，国信证券经济研究所整理

图：部分美国互联网大厂芯片发布时间与设计功耗对比




资料来源：半导体综研，国信证券经济研究所整理

图：部分国产芯片发布时间与设计功耗对比


资料来源：半导体综研，国信证券经济研究所整理

# 国内万卡训练集群需求旺盛，算力GPU云（租赁市场）发展可期

国信证券 GUOSEN SECURITIES

表：上市公司算力租赁相关订单情况（部分）

| 上市公司 | 已公告订单金额（亿元） | 公告订单时间 | 订单合同年限 | 已公告采购合同 | 采购合同订单时间 | 已申请授信额度(亿元) | 申请授信额度公告时间 | 定增 | 定增公告时间 |
|---|---|---|---|---|---|---|---|---|---|
| 宏景科技 | 42 | | | | | 40 | | | |
| | 1.49 | 2025/6/25 | | | | | | | |
| | 5.96 | 2025/5/20 | 5 | | | | | | |
| | 5.63 | 2025/5/14 | 5 | | | | | | |
| | 2.35 | 2025/4/16 | | | | | | | |
| | 3.12 | 2025/4/1 | | | | | | | |
| | 1.61 | 2025/3/19 | | | | 40 | 2025/3/18 | | |
| | 7.21 | 2025/3/12 | 5 | | | | | | |
| | 1.61 | 2025/2/25 | | | | | | | |
| | 4.09 | 2024/11/19 | | | | | | | |
| | 4.86 | 2024/10/30 | 3 | | | | | | |
| | 4.09 | 2024/10/22 | 3 | | | | | | |
| 利通电子 | | | | | | 40 | 2025/4/27 | | |
| 海南华铁 | 36.9 | 2025/3/5 | 5 | | | | | | |
| 蓝耘科技 | 37.07 | 2025/3/5 | | | | | | | |
| 弘信电子 | | | | | | 75 | 2025/2/25 | | |
| 协创数据 | | | | 70 | | | | | |
| | | | | 40 | 2025/5/27 | 205 | 2025/5/27 | | |
| | | | | 30 | 2025/3/7 | 35 | 2025/3/28 | | |
| 中贝通信 | 4.41 | 2025/4/21 | 4 | | | | | 19 | 2025/3/13 |
| 润建股份 | | | | | | 50 | 2025/3/31 | | |
| 智微智能 | | | | 30 | 2025/3/18 | 90 | 2025/3/7 | | |
| 有方科技 | 40 | | | | | 40 | 2025/6/6 | | |

资料来源：wind，公司官微，行业报告研究院官微，国信证券经济研究所理

**普通卡万卡集群：**

**成本：**硬件成本：约260万/台服务器 * 1024台 = 26.6亿；融资成本：约85%资金来自贷款/租赁（年化4-5%），5年总融资成本约3亿+；运维成本：5年约硬件成本的6% = 1.6亿；

**收入：**总收入=租赁收入（签5年长约，总金额约37-40亿）；退税收入：增值税退税约7%，约1.86亿；残值收入：服务器5年后残值保守按5%算 = 1.33亿 (实际可能更高)。

**总成本：**硬件26.6亿 + 融资3亿+ + 运维1.6亿 ≈ 31.2亿+。

**总利润（税前）：**总收入(约43亿) - 总成本(约31.2亿) ≈ 11.8亿。

**年化净利润（税后）：**约2亿/年 (保守1.5-2亿)。

**高级卡万卡集群：**

**成本：**硬件成本：约37-40亿。**租赁收入：**5年长约约60亿。

**年化净利润（税后）：**约3.5-4亿/年。

**结论：年化利润率为15-20%**

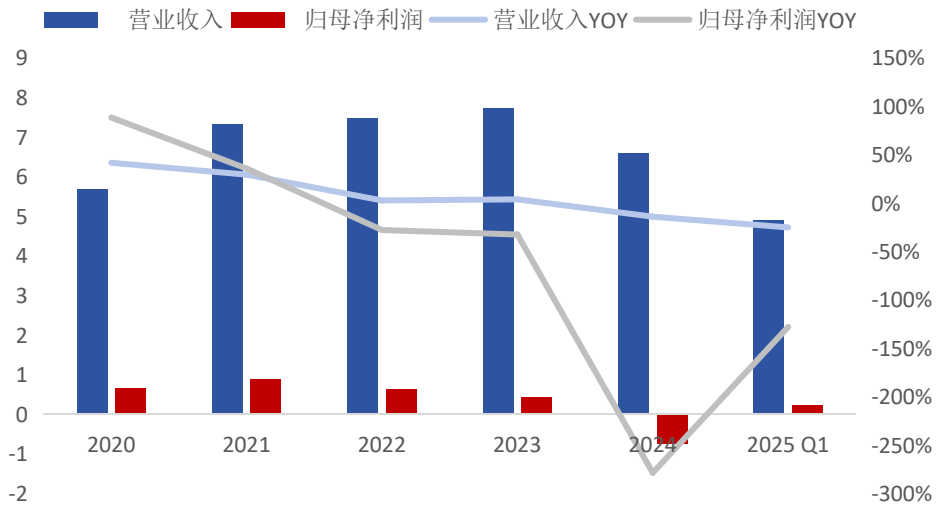
# 宏景科技：实现算力业务板块突破，打开全新发展空间

国信证券 GUOSEN SECURITIES

**宏景科技从智慧城市业务转型为算力服务商，算力服务业务收入高速增长。**宏景科技成立于1997年，专注于智慧城市业务，2023年以“AI+算力”为支点，转型算力服务新赛道。2023年，该公司算力服务业务收入为1.16亿元， 2024年提升至4.66亿元，同比增长302%，算力服务业务收入占总收入比重从15%提升至71%。算力业务量的增长带动营业收入上升。2025年一季度，公司营业收入4.89亿元，同比增长958%，归母净利润2165.12万元，实现扭亏为盈，主要系算力验收业务增加所致。算力服务的扩张为公司打开了新的发展空间。

**宏景科技深化算力业务布局，进行算力业务板块的突破，连续获得大额订单。**2025年以来宣布7笔订单，累计总金额27.5亿元。5月20日，宏景科技发布公告称，近日与Y公司签署了《智算项目服务合同》，按照Y公司要求向Y公司提供服务器、组网配套服务以及对服务器进行必要的改配服务，并提供算力服务。合同总金额为5.97亿元（含税），合同期限为五年。宏景科技表示，本次合同签署是市场对公司在算力业务及客户服务的进一步肯定，是各方展开更大规模战略性合作的开始，对公司算力业务的深化布局、市场的积极开拓和品牌效应建立都有着积极正面的影响。

图：公司营收及归母净利润情况(单位:亿元;%)




资料来源：wind，国信证券经济研究所整理

表：公司已签订的算力业务重大合同(部分)

| 公告订单时间 | 已公告订单金额（亿元） | 提供服务内容 |
|---|---|---|
| 2025/5/20 | 5.97 | 提供服务器、组网配套服务以及对服务器进行必要 的改配服务，并提供算力服务， |
| 2025/5/14 | 5.63 | 提供服务器、组网配套服务以及对服务器进行必要 的改配服务，并提供算力服务， |
| 2025/4/16 | 2.35 | 提供算力组网集成服务 |
| 2025/4/1 | 3.12 | 提供算力服务 |
| 2025/3/19 | 1.61 | 提供算力服务器采购、组网调试、服务器改配调优服务。 |
| 2025/3/12 | 7.21 | 提供服务器、组网配套服务以及对服务器进行必要 的改配服务，并提供算力服务， |
| 2025/2/25 | 1.61 | 求提供算力服务器采购、组网调试、服务器改配调优服务。 |
| 2024/11/19 | 4.09 | 提供硬件资源设备、组网配套服务及对服务器 进行必要的改配服务，并提供算力服务 |
| 2024/10/30 | 4.86 | 提供必要的算力服务 |
| 2024/10/22 | 4.09 | 提供硬件资源设备、组网配套服务及对服务器 进行必要的改配服务的，并提供算力服务 |

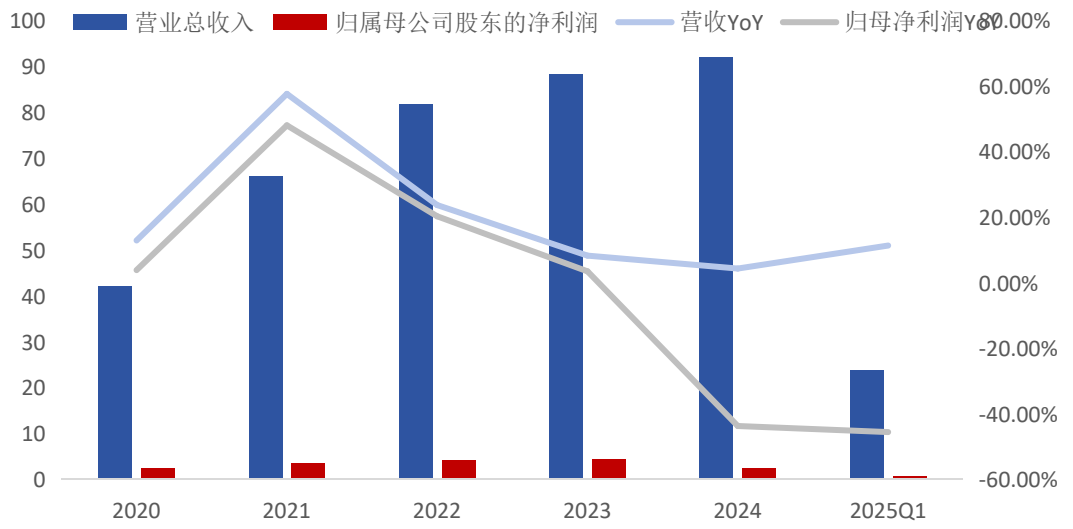
资料来源：公司官网，国信证券经济研究所理

# 润建股份：五象云谷智算中心提供智算云租赁业务

**公司为领先的数字化智能运维（AIOps）服务商、中国软件百强企业、中国服务业500强企业，致力于成为行业领先的人工智能行业模型及算力服务商。**公司在“线上数字化平台+线下技术服务”的经营模式基础上，进一步深化技术应用，发布人工智能发展战略，以算力服务、数据服务为基础，通过公司自主研发的“曲尺”生成式人工智能行业模型开发平台，锻造具有核心竞争力的行业模型、数字化产品、行业解决方案，赋能通信网络、数字网络、能源网络等业务领域。2024年公司实现91.99亿元，同比+4.23%；实现归母净利润2.47亿元，同比-43.77%，主要系公司对存货跌价准备的估计更加谨慎，整体准备计提增加，影响了全年整体净利润表现。2024年公司算力网络业务收入为5.29亿元，同比+71.7%。

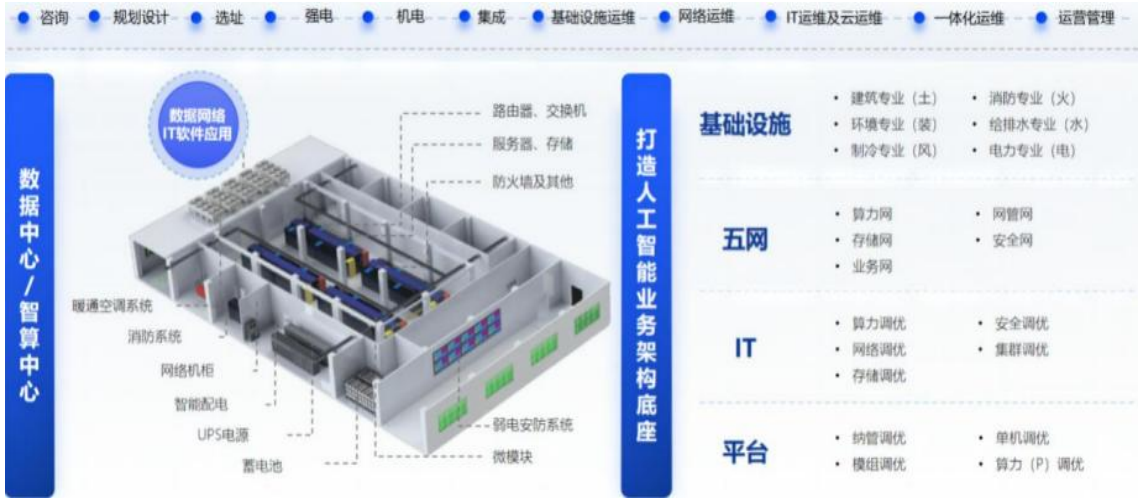
公司公开发行 A 股可转换公司债券募集资金人民币 10.9 亿元用于建设五象云谷智算中心项目，五象云谷智算中心将打造符合国标 A 级、国际 T3 级设计标准、满足国家绿色数据中心标准的标杆人工智能产业园区，是截至目前为止广西最高等级、最大规模的智算中心。在五象云谷智算中心项目基础上，公司加大投入升级算力服务，采购高性能算力服务器，打造润建股份智能算力中心，为客户提供智算云租赁服务，并且根据市场需求持续投入。

图：公司营收及归母净利润情况（单位：亿元；%）




资料来源：wind，国信证券经济研究所整理

图：五象云谷智算云服务



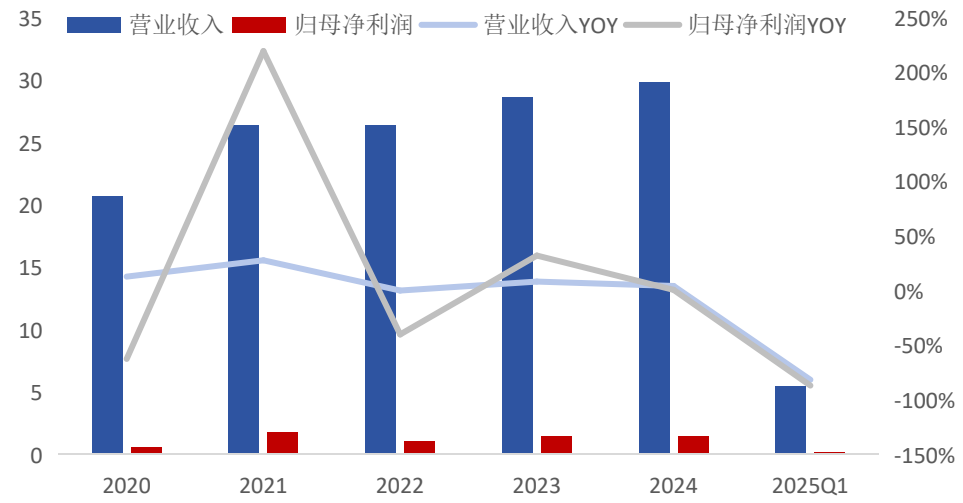

资料来源：公司公告，国信证券经济研究所整理

# 中贝通信：部署智算中心，推动实现智算中心“双万”战略目标

**中贝通信部署智算业务，新建多个智算中心，智算业务营收高速增长。**中贝通信成立于1992年，2023年公司开展算力业务运营，目前主要产品有5G新基建、智算业务、智慧城市及其他服务。公司已在全国多地建设智算中心，如中贝武当智算中心、中贝合肥智算中心、中贝三江源智算中心及中贝太原智算中心等，初步完成智算中心全国布局，其中合肥智算中心规划算力规模最高可达25000P。2020年-2024年，公司营业收入稳步提升，复合增长率10%。2024年整体营收29.84亿元，同比增长4%。25Q1营业收入5.49亿元，其中智算业务收入1.35亿元，同比增长973%。归母净利润整体稳定增长，复合增长率26%，2024年归母净利润为1.45亿元。

**签订算力服务订单，实现智算中心“双万”战略目标，推动算力出海。**2025年4月21日，中贝通信与万界数据科技公司签订《算力服务合同》，中贝通信提供智算中心的算力服务，提供定制款八卡推理服务器集群。合同金额为4.41亿元，服务期限4年。中贝通信董事长李六兵表示，公司正深化智算新算力业务，努力打通上下游产业链，尽快实现智算中心“双万”战略目标（万P算力+万卡集群）。未来国际公司将以新基建、新算力、新能源三大业务为重点，拓宽中东资源网，发展欧洲、东南亚算力业务，提升品牌及公司价值。

图：公司营收及归母净利润情况(单位:亿元;%)




资料来源：wind，国信证券经济研究所整理

图：中贝通信云集群现场

资料来源：中贝通信官网，国信证券经济研究所整理

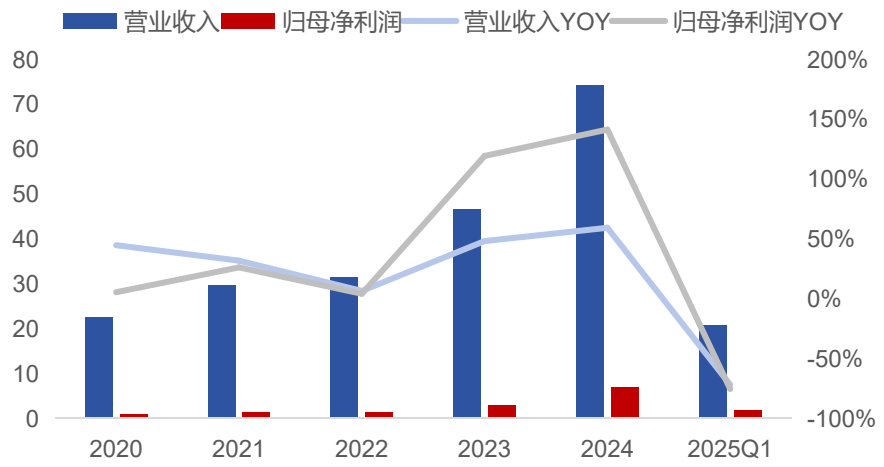
# 协创数据：加速开拓算力市场，大量采购服务器以促高端算力合作

国信证券 GUOSEN SECURITIES

**加速开拓算力市场，进行数据中心布局，近两年营收增长迅速。**协创数据成立于2005年，专注于物联网智能终端和数据存储领域，产品涵盖数据存储设备、AIoT智能终端、云服务等。目前公司正加大算力领域的投入。协创云已在中国、美国、欧洲、亚太四大服务区域的十大节点布局，并各区域设立独立的数据中心。2020年-2024年，公司营业收入稳步提升，复合增长率35%。2024年实现营收74.1亿元，同比增速59%，25Q1营业收入20.77亿元，同比增长18%。其中2023年、2024年增长迅速，同比增速分别为48%/59%，主要系海外客户持续拓展带动数据存储设备等核心产品销量增长所致。归母净利润整体稳定增长，复合增长率26%。2024年归母净利润1.45亿元，其中2023年、2024年增长迅速，同比增速分别为119%/141%，主要系高附加值业务占比提升所致。

**大规模采购服务器，取得NVIDIA CLOUD PARTNER资质，进行以高端算力为基础的云算力租赁。**2024年6月以来，公司先后与优崴超级运算股份有限公司、日本优必达株式会社、中国移动等签署云业务合作协议。目前公司向多家头部企业提供算力应用产品实现业务增值，展开以高端算力为基础的云算力租赁、云安防和大模型的合作。2024年10月，公司控股子公司奥佳软件正式取得NVIDIA CLOUD PARTNER资质，在AI算力租赁和云计算服务方面具备了与英伟达合作的资质和能力。2025年3月7日，公司发布公告称拟采购服务器，合同金额不超过30亿元，主要用于为客户提供算力租赁服务。

图：公司营收及归母净利润情况(单位:亿元;%)




资料来源：wind，国信证券经济研究所整理

图：协创数据中心节点分布


资料来源：公司年报，国信证券经济研究所整理

# 渠道、组网、资金周转能力为算力租赁核心壁垒

国信证券 GUOSEN SECURITIES

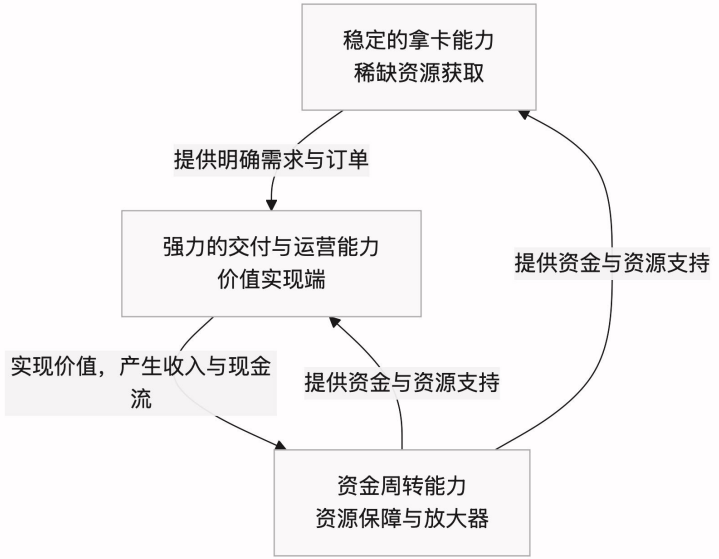
**我国高端算力需求旺盛，渠道+组网+现金流周转能力为核心竞争力：**

**稳定的拿卡能力。**互联网等客户对算力需求较大，单笔订单所需服务器数量较大，对供应商的供应能力有较强要求。如果不能按时交付，可能会有一定的处罚。

**强力的交付、上线和后续运营能力。**AI算力服务器工作条件要求高，到货后需要调试、组网和调优，达到一定要求之后才能正常满足客户使用需求。同时，AI算力服务器上线运营之后，需要专业团队维护和运营。较快的交付、上线时间，较少的损耗和故障情况，可以显著提升业务流畅性、降低运营成本，提升盈利能力。

**资金周转能力。**AI算力服务订单金额较大，上市公司需要各种融资方式来完成订单要求。较强的资金周转能力决定公司能否按时完成客户的订单交付已经上线运营。

图：高端算力获卡能力、交付能力、资金周转能力关系




资料来源：华纳云，行业报告研究院，国信证券经济研究所整理

图：中贝通信定增融资

关于中贝通信集团股份有限公司向特定对象发行股票申请文件的审核问询函之回复报告

**问题 1. 关于募投项目**

**根据申报材料，1）发行人本次向特定对象发行股票募集资金总额不超过 192,223.48 万元，将用于“智算中心建设项目”、“5G 通信网络建设项目”和偿还银行借款。2）“智算中心建设项目”采用“以销定产”模式，拟在庆阳、丹江口等地建设智算集群。3）“5G 通信网络建设项目”计划在多个省市、自治区及直辖市进行 5G 通信网络相关设施的建设。4）公司前次募投项目实际效益不达预期，主要因相关框架合同的实际订单执行率未达预期所致。**

资料来源：公司公告，国信证券经济研究所整理

图：宏景科技募集资金

一、募集资金基本情况

（一）实际募集资金金额、资金到位时间

经中国证券监督管理委员会《关于同意宏景科技股份有限公司首次公开发行股票注册的批复》（证监许可[2022]1325 号）同意注册，公司首次向社会公开发行人民币普通股（A 股）22,844,900.00 股，每股面值人民币 1.00 元，每股发行认购价格为人民币 40.13 元，募集资金总额为人民币 916,765,837.00 元，扣除与发行有关的费用人民币 99,589,223.65 元（不含税），公司实际募集资金净额为人民币 817,176,613.35 元。募集资金总额人民币 916,765,837.00 元，扣除券商剩余应支付的承销保荐费用人民币 80,260,031.38 元后剩余募集资金人民币 836,505,805.62 元已于 2022 年 11 月 8 日全部到账。

上述募集资金净额经华兴会计师事务所（特殊普通合伙）“华兴验字[2022]21000590495 号”《验资报告》验证。公司对募集资金采取专户存储管理。

资料来源：公司公告，国信证券经济研究所整理

# 海内外算力租赁对比：海外发展相对成熟，具备一定参考意义



受制于供给因素，我国算力租赁市场相较于北美仍处于发展早期，尚没有形成大的GPU云服务商，市场格局较为分散。**考虑到中美算力租赁市场的主导因素存在一定差异，预计国内市场短期仍维持较为分散的格局，但是有望逐步向有经验的厂商集中。**海外CoreWeave为代表的算力租赁厂商发展时间也较短，仍在持续扩展市场、积累经验，但是商业模式基本定型，相较于国内更为成熟，具备一定参考意义。

表：海内外算力租赁模式对比

| | 海外 | 国内 |
|---|---|---|
| 供给 | 以英伟达最新高端芯片为主，目前为GB200/300机柜 | 8卡服务器为主 |
| 需求 | 微软、谷歌、Oracle等厂商；OpenAI等大模型厂商；初创公司 | 阿里、字节、腾讯等互联网企业、AI大模型厂商、政府等 |
| GPU云参与者 | 相对集中 | 相对分散 |
| 发展阶段 | 较为成熟 | 较为早期 |
| 受到芯片政策影响程度 | 较小 | 较大 |
| 主导因素 | 英伟达战略规划及下游需求 | 芯片政策及下游需求 |

资料来源：SemiAnalysis，Omdia Research，国信证券经济研究所整理

# 第五章 投资建议

# 投资建议



**全球算力景气度延续，随着英伟达GB系列高密度算力机柜加速出货，全球高端算力景气度进一步提升。**当前AIGC浪潮下，全球服务器出货量持续增长，咨询机构IDC预计2028年全球人工智能服务器市场规模有望达到2,227亿美元，其中生成式人工智能服务器占比将从2025年的29.6%提升至2028年的37.7%。从需求端来看，模型迭代加快背景下训练端需求仍维持高位，推理侧需求随着应用的渗透逐步提升；从供给端来看，以英伟达GB200、GB300、AMD MI350等代表的最高性能算力持续迭代，2025年下半年GB300有望加速交付。根据CSP厂商的Capex指引，预计2025年，海外亚马逊、谷歌 、微软 、Meta四家厂商合计 Capex 增至3610亿美元，同比增幅超 58%；国内字节、腾讯、阿里Capex有望超过3600亿元，下游景气度延续。

**GPU云(算力租赁)或解决目前全球高端AI芯片紧缺问题，GPU云(算力租赁)市场快速发展。**在大模型军备竞赛的背景下，各大厂加速万卡甚至十万卡集群建设。Meta、微软&OpenAI、xAI等多家AI巨头陆续宣布或者完成10万卡集群建设，国内通信运营商、头部互联网、大型AI研发企业等均发力超万卡集群的布局。然而**在全球高端AI芯片供给紧缺背景下，以租赁代替购买的商业模式应运而生，因地制宜且性价比更高。**云计算市场历经传统云、混合云阶段后，或迎来第三次分化浪潮——AI智算云NeoCloud，即GPU云(算力租赁) ，预计到2033年全球GPU云(算力租赁)市场规模将增至128亿美元(Verified Market Research预测)。

**AI芯片巨头正在通过GPU云(算力租赁)商业模式布局全球市场，国内GPU云市场发展值得期待。**英伟达以股权或合作方式辅助GPU云厂（ CoreWeave、NBIS、Omniva 等）发展，巩固其在高端芯片领域的全球主导地位。三家GPU云厂覆盖区域和发展规模虽有不同，但均受益GPU云市场的高景气度处于快速增长期，2025Q2CoreWeave和NBIS营收增速分别达到207%/625%。国内方面，国产AI芯片目前主要支持推理业务，部分训练场景英伟达高端AI芯片性能表现更优；国内外算力政策有差异，同时以OPEX租赁算力方式实现训练业务或具备更高性价比，国内算力租赁企业迎来发展契机。**目前国内算力租赁企业的租赁回报较为可观，测算净利率或达15%，同时与海外GPU云(算力租赁)的商业模式具有部分相同之处。**

**投资建议：**AI算力景气度持续，短期来看，GPU云或为解决高端算力供需不匹配的核心解决方案；长期来看，GPU云具备灵活、低成本的解决方案，渗透率有望持续提升。推荐关注具备提供GPU云能力的相关企业，建议关注【润建股份】及相关产业公司。

**风险提示：**AI发展及投资不及预期，行业竞争加剧，全球地缘政治风险，新技术发展引起产业链变迁。

# 主要公司盈利预测

表：重点公司盈利预测和估值

| 公司代码 | 公司名称 | 收盘价（8月22日） | 总市值（亿元） | 净利润 | | | PE | | | PEG |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 2024A | 2025E | 2026E | 2024A | 2025E | 2026E | 2025E |
| 002929.SZ | 润建股份 | 50.0 | 142.0 | 2.5 | 3.1 | 4.6 | 57.6 | 46.3 | 30.7 | 1.3 |

资料来源：Wind，国信证券经济研究所整理；各公司盈利预测取自Wind一致预期

# 风险提示

国信证券 GUOSEN SECURITIES

- AI发展及投资不及预期
- 行业竞争加剧
- 全球地缘政治风险
- 新技术发展引起产业链变迁

# 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