方正证券 FOUNDER SECURITIES

正 在 你 身 边

证券研究报告 | 新兴产业 | 2026年09月10日

新兴产业团队·行业深度报告

算力租赁深度报告：海内外业绩与需求共振，技术路径向TaaS演进

分析师 刘明洋 登记编号：S1220524010002

孔德璋 登记编号：S1220525120001

李鲁靖 登记编号：S1220523090002



报告摘要

算力租赁是AI时代连接上游算力硬件与下游AI应用的核心桥梁，也是AI时代的重要基建。算力租赁指服务商将搭载GPU等高性能芯片的服务器算力资源，通过云化方式以按需付费、弹性租用的模式出租给有AI训练、推理需求的企业或机构，按交付形态可分为整机租赁、算力规模租赁与GPU时租三种主流模式。其核心价值在于客户可以OPEX替代CAPEX，通过算力以租代建的方式降低前期投入成本和减值风险。算力租赁厂商的核心竞争力可归结为算力卡采购/调优能力、低成本资金、优质客户三类，三者构成“拿卡→绑定客户→稳定现金流→再融资→拿更多卡”的正向循环，其中拿卡与算力调优能力是目前卖方市场下最关键、最稀缺的能力。

国内算力租赁正处于总量扩容+商业模式迭代阶段，呈现出供需缺口扩大、商业模式向TaaS演进、推理需求主导的态势：

市场供需缺口持续扩大，市场规模有望于2030年接近万亿。供给端：国内算力建设正加速扩容，截至2026年6月底我国智能算力规模达2185EFLOPS、同比+177%，全国算力设施整体上架率达71.4%，国家层面正筹备2026-2030年投入约2万亿元建设全国性AI基础设施，运营商与互联网大厂资本开支集体加码。**需求端：**国内日均Token调用量两年增长超千倍、已接近175万亿/天，互联网大厂与大模型厂资本开支集体加码。**但供需缺口在持续扩大，**26Q1国内AI算力需求同比+417%，显著高于高端GPU供给+128%的增速。BIS审查持续收紧，高端算力交付周期已延至2027Q1-Q2，导致H200、B300等整机价格与租金大幅上行。**26Q1国内算力租赁市场规模680亿元、同比+62%，全年预计突破2600亿元，我们测算2030年国内算租市场规模有望接近万亿规模。**

推理需求正在成为主赛道，开源模型的发布进一步印证。中国大模型周调用量达34.25万亿Token，已连续15周超过美国稳居全球首位，国内已进入算力大规模消耗阶段。IDC预计到2028年，推理工作负载占比将达到73%，推理需求逐渐成为主流。DeepSeek V4、Kimi K3等开源模型接连发布，借1M上下文+Agent等特性将进一步放大推理消耗。并且开源模型大幅降低使用与部署门槛，目前已有超20家央企接入DeepSeek、多地政府同步跟进部署，大量原本无法承担闭源API成本、也无法接受数据外流风险的企业开始具备部署条件，客户群体逐渐向研发机构、初创企业扩散，直接拉动推理算力需求。

算力租赁模式正向TaaS与Token工厂演进。TaaS以Token作为统一标准化计价标的对外销售推理能力，将模型能力变成可量化、可采购的业务消耗品。相较于注重硬件资源出租的IaaS模式，TaaS直接交付Token产出与结果，输出质量更高，且可根据客户需求弹性部署。不同于训练需求大多集中于头部模型公司，推理需求会扩散至调用API做垂直应用的SaaS厂商、传统企业IT部门、科研机构等大量长尾主体，**而TaaS能够以更细颗粒度的计量和更低门槛的接入，把分散的推理需求汇聚成可持续收入，是更适配推理主导型需求结构的商业模式。**Token工厂有望成为TaaS的工程化落地的体现，其核心逻辑是输入电力、算力、数据与算法，输出可计费的Token。凭借Token分成模式及算力优化，相较于传统算租20%左右的毛利率，Token工厂的毛利率可达50%以上，投资回收期可由传统算租的3-5年缩短至2年左右，收入有望达传统算力租赁的5-10倍。目前北京、厦门、无锡等地已启动Token工厂建设，产业将实现快速发展。

更多免费行业报告 并购家 www.ipoipo.net

海外算力租赁已进入需求外溢与业绩兑现期。

Hyperscaler：三大厂商26Q2业绩验证Capex-营收正循环，客户由Frontier Labs向更广泛B端扩散。微软FY26Q4 Intelligent Cloud收入393亿美元、同比+32%，其中Azure与其他云业务同比+43%，Azure全财年收入已突破1000亿美元、同比+41%；backlog 6780亿美元、同比+84%，环比增加的510亿美元全部来自非AI模型开发商客户，云侧订单并未因巨额Capex被消化完，需求端正从Frontier Labs向更广泛企业端扩散。**AWS** 26Q2收入422.3亿美元、同比+37%，创18个季度以来最快增速，backlog4960亿美元、环比增加1320亿美元；公司上调全年Capex至2200亿美元，并直接指出“供给建设速度<需求增长速度”，2026-2027年仍无法满足全部需求。**Google Cloud** 26Q2收入247.7亿美元、同比+82%，营业利润率提升至35.6%，backlog达5140亿美元、环比增加540亿美元；公司将上调全年Capex指引并明确表示Q3将扩大第三方算力租赁作为过渡，侧面证明Neocloud等新算租厂商已进入顶级云厂商的算力供给体系。

Meta近期动向验证Hyperscaler与Neocloud的共存关系。公司于8月5日正式启动算力商业化战略、向外部客户出租自有GPU资源，并同步筹建类AWS Bedrock的托管模型服务与类Neocloud的裸金属出租业务，为巨额Capex补充变现路径。同时公司对高端算力需求并未放缓，截至26H1已签约超5GW云+托管容量、2024年初至今累计近10GW，且在几大Neocloud订单已签满的背景下，其新增供给并未对Neocloud形成挤出，反而有望向Neocloud追加订单以加速扩建。这印证了在供需缺口持续存在的阶段，**Hyperscaler与Neocloud之间并非竞争，而是互存共生的生态关系：**Neocloud为Hyperscaler提供弹性算力补充与过渡产能，Hyperscaler的外溢需求构成Neocloud的核心订单来源，新增的每GW供给都能被市场直接吸收。

Neocloud方面：26Q2订单与业绩同步爆发，逐渐向平台能力过渡。CoreWeave 26Q2营收25.75亿美元、同比+112%，backlog1040亿美元、同比+246%，且Q3初新增超250亿美元客户承诺。7月全线SKU提价约25%，客户名单新增多家应用端企业，验证算力需求正从Frontier Labs流向应用端。Nebius 26Q2营收同比+454%，在手订单400亿美元，经营性现金流22.46亿美元大幅转正，全年客户预付款将超90亿美元、覆盖资本开支的50%-60%，公司并表示“能把2027年全部产能按现价卖光但选择不卖”，暗示对未来供需缺口将持续扩大。Cloudflare 26Q2营收同比+35.9%、超出指引上限，Agentic Internet叙事开始兑现订单。**整体来看，海外算租的订单高增、提价落地与旧卡第二生命周期（CoreWeave的A100集群获续签至2029年、仅低于原始价格5%）共同验证了行业景气的持续性。海外算租也呈现供需缺口扩大、推理需求高增、商业模式从租赁向平台/TaaS过渡、与国内高度共振。**

目前可从三大方向寻找算力租赁板块机会：

供需缺口：国内AI算力需求同比+417% vs 高端GPU供给+128%,BIS审查持续收紧,租金与整机价格大幅上行—**手里有卡、有电、有直接算力交付能力**的算力租赁商有望受益。其存量GPU资产具备保值属性、增量高端卡具备定价权。

TaaS/Token工厂：算力需求中推理需求占比将提升至73%，国内Token调用量两年千倍增长，商业模式从IaaS向MaaS/TaaS升级,Token工厂毛利率(50%+)与回本周期(约2年)具备优势，且能承接AI初创企业、研究单位、地方政府等长尾客户—**具备推理优化、调度运营能力并率先落地TaaS模式/Token工厂**的厂商将受益。

国产算力适配：开源模型(DeepSeek V4/Kimi K3)全面适配国产芯片架构，BIS收紧下国产卡从政策替代品转向成本/可得性优势选择。在供需缺口扩大的背景下**具备国产卡适配能力(通过调优等方式优化国产卡算力输出能力)的厂商将充分享受国产卡的替代&成本优势**。

建议关注：
算力租赁/智算服务：利通电子、并行科技、中贝通信、恒为科技等；
Token工厂/TaaS：协创数据、宏景科技、行云科技、杰创智能、润建股份、首都在线、优刻得-W等；
IDC转型AIDC：东阳光、润泽科技、数据港、世纪互联、奥飞数据、光环新网、科华数据、云赛智联等；
运营商：中国移动、中国电信、中国联通等。

风险提示：出口管制政策变化风险、GPU价格下行与资产减值风险、算力租赁价格下降超预期风险、行业竞争加剧风险、技术迭代风险、测算偏差风险等。

更多免费行业报告 并购家 www.ipoipo.net

方正证券
FOUNDER SECURITIES
正 在 你 身 边

一、国内算租供需缺口持续加大，推理时代到来推动TaaS模式加速落地

1.1 算力租赁连接上游AI硬件与下游应用，核心价值在于需求方可以用OPEX替代CAPEX

算力租赁是AI时代连接上游算力硬件与下游AI应用的核心桥梁，也是AI时代的重要基建。算力租赁指拥有算力基础设施的服务商，将搭载GPU、CPU等高性能芯片的服务器算力资源，通过云化的方式，以按需付费、弹性租用的模式出租给有AI训练、推理或高性能计算需求的企业或机构。算力租赁可被视为把上游资本品加工成下游可即时调用服务的中枢平台。训练时代的核心指标是卡时数与集群稳定性；推理时代则更强调单卡Token产出、延迟、调用弹性与客户运营效率。也正因为推理业务权重上升，算力租赁的竞争逻辑已经从"谁机房多"扩展到"谁能更高效调度、谁能更便宜融资、谁能更稳地绑定客户"。

对于需求方而言，算力租赁的核心价值在于以OPEX(运营支出)替代CAPEX(资本支出)。目前单台8卡A100的GPU服务器采购成本达160万元(8卡B300达1400万元)，配套机房也将产生建设、电力、运维成本，对于中小企业而言，一次性采购多台高端GPU服务器的总成本过高，投资回报周期长达36个月，资金压力巨大。同时，GPU硬件更新周期较短，需求方将承担技术快速迭代所带来的减值风险。**而通过算力租赁，**客户无需一次性投入巨额资本购置GPU服务器，而是通过签订租赁合同将固定资本支出转化为运营支出，实现以租代建的OPEX模式，对需求方的价值显著放大：1.成本节约（避免硬件快速迭代带来的减值风险）；2.部署灵活（按需弹性扩缩容）；3.运维简化（服务商负责硬件运维与网络组网）。**2025年，国内AI企业中，采用算力租赁模式的企业占比达73%，其中中小企业占比89%；科研机构、高校采用算力租赁的比例达68%，较2023年提升42个百分点，需求稳定向上。**

按交付形态与计费方式，目前算力租赁可分为三种主流模式：①整机租赁—租用固定规格的服务器，资源配额相对稳定，可实现物理隔离，适配需要持续稳定算力支撑的训练场景，按台数/时间计费；②算力规模租赁—以算力资源消耗为前提，可额外调用弹性算力资源，算力使用规模随业务需求灵活调整，计费方式通常为PFlops/时间；③GPU时租—按单个GPU卡每小时使用量计费，属于算力共享模式，按每小时计费。

图：三种算力租赁模式(数据截至26年3月)

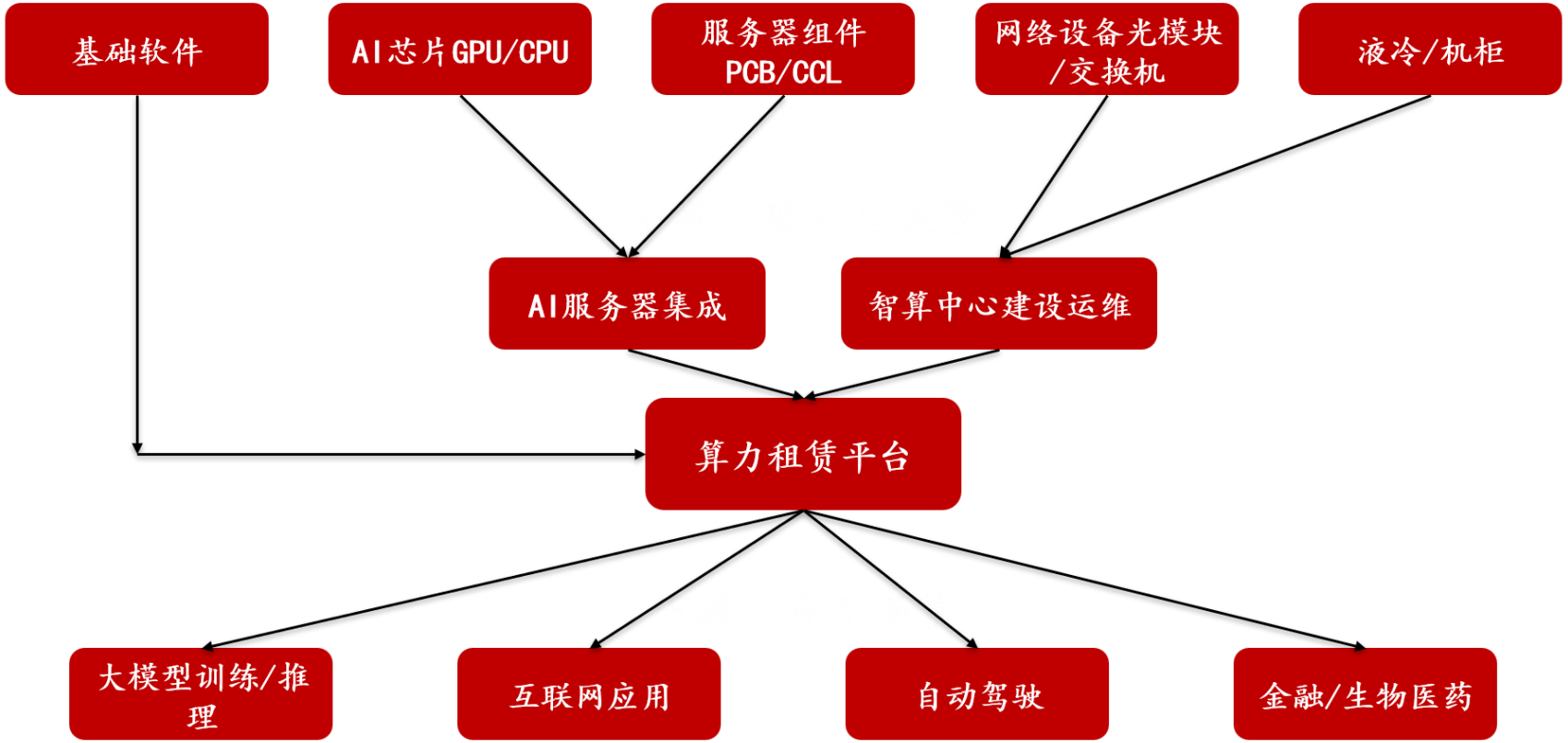
商业模式	定义	定价方式	算力利用率(%)	客户群体	优势	劣势	典型场景
整机租赁（物理隔离）	以搭载特定数量GPU的物理服务器为单位，按月/年长租，资源物理隔离	按月计费、按年计费，长期租赁价格较低	85-90	头部AI企业、国家级科研项目、大型制造企业	性能独占、安全隔离，稳定性强，支持定制化配置	单价较高，短期租赁性价比低，灵活性不足	千亿参数大模型训练、高精度工业仿真、核心数据计算
算力规模租赁（虚拟化）	通过虚拟化技术将算力池化，以计算能力单位（PFLOPS/月）计费，按需弹性分配	按算力规模计费，弹性扩容/缩容，按需结算	90-95	中型AI企业、互联网公司、科研机构	算力弹性灵活，资源利用率高，成本可控，适配需求波动	存在虚拟化损耗（5%-8%），核心数据安全性低于整机租赁	中大型AI模型训练、大数据分析、数字孪生
GPU时租（高弹性）	按单个GPU每小时使用量计费，实现算力"即取即用"，属于算力共享模式	按小时计费，不同GPU型号单价不同	95-97	中小企业、个人开发者、高校学生、创业团队	门槛极低，灵活性极强，按需付费，无闲置成本	高峰时段可能出现资源排队，长期使用总费用偏高	模型微调、算法测试、轻量级推理、科研实验

资料来源：星宇智算，方正证券研究所

更多免费行业报告 并购家 www.ipoipo.net

算力租赁产业链的**上游为基础设施：**包含提供AI芯片、服务器硬件、网络设备、液冷机柜、IDC基础设施的硬件与软件商；**中游为算力集成与服务提供：**包含数据中心集成、建设运维、算力调优、云计算/算力租赁平台运营商(大多数算力租赁公司处于中游)；**下游为应用层：**包含各类将算力用于大模型训练/推理、AI应用、自动驾驶、金融、生物医药等企业与机构。综合来看，算力租赁厂商大多处在产业链中游，主要负责将上游的重资产转化成下游应用端可直接调用的生产力。

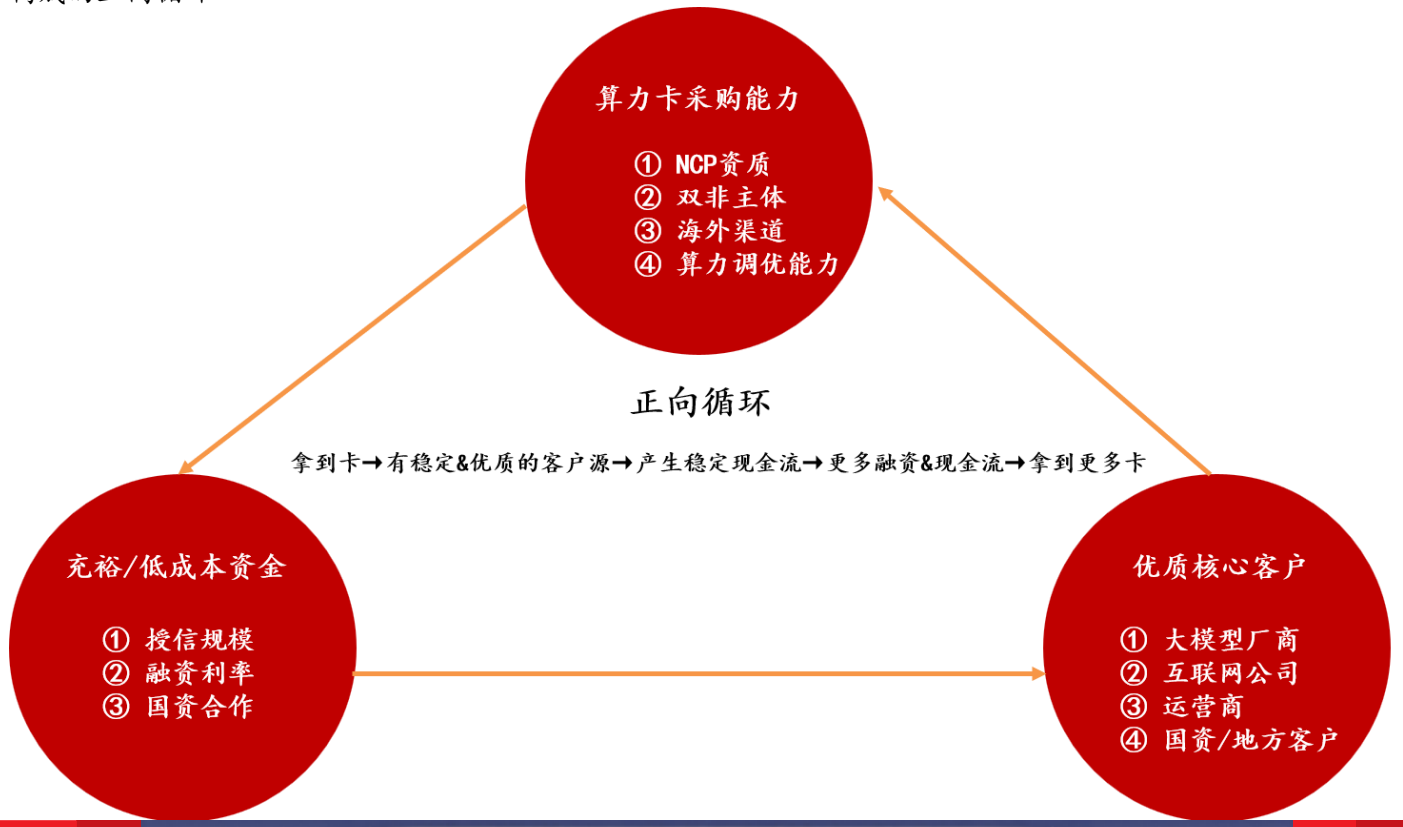
图：算力租赁产业链



1.3 算力租赁厂商的三大核心竞争力可以构成正向循环

算力租赁厂商的核心竞争力可归结三类：算力卡采购能力/算力调优能力+充裕/低成本资金+优质核心客户。这三者构成正向循环—拿到卡→有稳定&优质的客户源→产生稳定现金流→更多融资&现金流→拿到更多卡。正向循环一旦形成，企业不仅规模扩张更快，单位资金效率也会更优：拿卡能力意味着企业具备承接高端需求的资格，在BIS审查逐渐收紧的背景下，拿卡能力/通过算力调优最大化算力输出决定了厂商的供给能力，是在算力租赁成为卖方市场背景下最关键、最稀缺的能力。资金成本决定企业能否以更低杠杆成本和更高速度扩大设备规模，客户锁定则决定订单深度、现金流质量与续投意愿。

图：算力租赁厂商三大核心竞争力构成的正向循环



资料来源：方正证券研究所整理

更多免费行业报告 并购家 www.ipoipo.net

1.4 供需缺口扩大推动景气度上行，国内算租市场规模有望于2030年接近万亿规模

供给端：国内算力租赁行业已进入加速扩容阶段。截至2026年6月底我国智能算力规模达2185EFLOPS，同比增长177%，全国算力设施整体上架率达71.4%。**运营商侧**也持续加大算力投入以扩容：中国联通乌镇智算中心一期算力规模为2000PFLOPS、计划2026年完成万卡集群建设，长三角（吴江）智算中心项目总投资达51亿元；中国联通明确2026年全年资本开支500亿元，其中算力相关投资超175亿元、占比达35%。**国家层面来看，**国内正筹备在未来五年（2026—2030年）投入约2万亿元人民币用于建设全国性人工智能基础设施，以强化本土AI产业能力，并在人工智能领域与美国展开长期竞争。

BIS审查持续收紧，高端算力紧缺，算力采购/租赁价格上行。26年5月31日，**BIS进一步收紧审查政策，**明确总部设于中国的企业，即使通过境外子公司采购先进AI芯片，也必须事先向BIS申请出口许可证。审查政策的不断收紧导致了**海外高端GPU机柜价格持续上涨。**26年7月，B300八卡整机现货价格超过1400万人民币，约为美国市场的三倍。SMM表示，H100运营资源多用于既有客户项目、空余供给有限，散租报价持续走高；H20作为合规供应主力型号，虽有批量入市，但市场流通资源整体仍然偏紧，短期难改。高端算力交付周期已延至2027Q1-Q2，客户采购受配额限制且需排队等待。**从算力采购/租赁价格来看，**H200、H100、B300等高端算力价格较25年年底大幅上行，侧面反映目前算力供需缺口仍旧存在。

图：国内算力租赁景气度供需缺口维度

维度	关键数据/进展	含义
需求缺口	26Q1需求+417%，供给+128%，BIS审查持续收紧	高端算力持续紧张，价格有支撑
推理需求	中国模型周调用量超34万亿Token，连续15周全球领先	推理需求真实落地，带动算租需求上行
建设加速	国家拟在未来5年投入2万亿人民币建设数据中心、阿里云计划将全球数据中心产能提升两倍以上等。	国内算力租赁加速扩容

图：国内部分GPU整机价格与租金变化

机型/整机	2025年底价格/租金	2026年近期价格/租金	变化
H200 八卡整机采购价	约210万—220万元/台	5月已到约340万元/台	大幅上行
H200 八卡整机月租	约6万元出头/月	12-13.5万/月	接近翻倍
H100 八卡整机月租	约4.5万元/月	7.5万—7.8万元/月	上涨约50%+
H100 八卡整机采购价	—	7月初274万元-8月300万	高位
B300 八卡整机采购价	约400万元/台	1400万元/台	翻三倍多

资料来源：SMM，晨涧云，南方都市报，方正证券研究所整理

1.4 供需缺口扩大推动景气度上行，国内算租市场规模有望于2030年接近万亿规模

需求端：Token调用量大幅增长，需求加速释放，供需缺口持续扩大。根据信通院的披露，国内26年3月底的日均Token调用量超140万亿，较2024年初增长超千倍，2026年6月初日均Token调用量已达近175万亿/天。市场规模来看，2026年一季度国内算力租赁市场规模已达680亿元，同比增长62%，全年预计突破2600亿元；同期AI算力需求同比增长417%，显著高于高端GPU供给128%的增速，**供需缺口仍在扩大。**

根据OpenRouter数据，2026年8月3-9日，中国大模型周调用量达34.25万亿Token，环比增长21.76%，为同期美国调用量的3.74倍，已连续15周超过美国、稳居全球首位。当周全球调用量前四名均为中国模型，国内已进入算力消耗阶段。同时开源模型降本使得大量开发者、中小厂加入市场，激发大量推理需求，有望直接拉动算力租赁景气度上行。

互联网大厂资本开支大幅增加，大模型厂短期算力供不应求。阿里在2025年2月宣布未来三年投入3800亿元用于云与AI硬件基础设施建设，超过其过去十年相关投入总和，截至2026年6月，该计划进度已接近一半、约1900亿元；公司8月份表示阿里云正计划将模块化数据中心的全球产能提升两倍以上。2027财年第一季度（截至26年6月30日）单季资本开支达677亿元、同比增长75%。从回报侧看，阿里公开披露其AI相关产品年化收入已超过495亿元人民币，前期投入的商业回收已逐渐启动。从大模型厂来看，KIMI在发布K3模型后公开表示用户请求量已逼近现有集群的承载极限，并暂停C端新用户订阅，并且正在全速推进算力扩容。

整体来看，国内算力租赁市场当前正处于总量扩容的过程中，但供需缺口持续存在且在加大。主要原因为需求(Token调用量千倍增长+大厂运营商资本开支集体加码)的释放速度快于高端算力供给(受芯片获取渠道约束+国内算力建设仍在过程中)的扩张速度，二者之间的剪刀差直接反映在持续上行的整机与租赁报价上。

资料来源：OpenRouter，阿里，KIMI，中[illegible]

更多免费行业报告 并购家 www.ipoipo.net

1.4 供需缺口扩大推动景气度上行，国内算租市场规模有望于2030年接近万亿规模

我们认为，由于供给端受多方因素影响(单卡算力迭代，海外GPU进口政策缩紧，算租市场渗透率受国家大规模基建影响)国内算力租赁市场中期的市场规模更多取决于需求端，也就是下游AI应用对Token的持续消耗。我们构建测算公式：市场规模$_{需求}$=日均Token调用量×日均Token收入系数。

根据中国信通院数据，2026年一季度国内算力租赁市场规模达到680亿元，预计全年将突破2600亿元。从Token调用量来看，2025年底国内日均Token调用量为100万亿，2026年3月突破140万亿，较2024年增长超千倍，假设2026年底日均调用量达200万亿。则我们能够得出，**2026年国内算力租赁日均Token收入系数为2600/200=13亿元/万亿/年(国内每万亿全部日均Token调用量对应的算力租赁市场规模)**

增速：日均Token-根据高盛预测，2030年全球Token消耗量将达到每月约120千万亿,对应日均调用量4000万亿。同时，根据TokenInsight，截至26年3月，国产大模型在全球主流AI调用平台的Token占比达61.2%，**假设国内Token调用量与国产大模型Token调用量占比相符，**则国内2030年Token调用量为全球的60%，则国内2030年日均Token调用量为2400万亿，2026-2030CAGR为86.12%。**收入系数-**随着芯片、模型、推理架构和缓存优化进步，单位Token的算力成本会快速下降。高盛认为虽然现在Token价格正在稳定上涨，算力**成本将每年下降60%—70%**。考虑租赁价格下降显著慢于底层计算成本下降，且算租渗透率有望加速，假设**日均Token收入系数每年下降25%**。

图：国内算力租赁市场规模测算

年份	日均Token调用量（万亿）	算租日均Token收入系数（亿元/万亿/年）	市场规模（亿元）
2025	100	17.33	1733
2026	200	13.00	2600
2027	372	9.75	3627
2028	693	7.31	5068
2029	1289	5.48	7069
2030	2400	4.11	9872

根据测算结果，假设2030年国内日均token调用量为2400万亿次/天，算力租赁日均token收入系数为4.11亿元/万亿/年，则2030年国内算力租赁市场规模将达到9872亿元，接近万亿规模。

资料来源：TokenInsight，方正证券研究所测算

1.5 AI进入规模化推理时代，TaaS将成为新发展方向

TaaS(Token as a Service,Token即服务)，是AI产业进入规模化推理时代诞生的新型商业化模式。服务商将大模型推理能力封装，以Token作为统一标准化计价标的对外销售;下游企业按照实际消耗Token数量按量结算，无需采购服务器、GPU等硬件算力，也不用独立部署、运维大模型。相较于早期算力租赁的IaaS模式更注重于对GPU、机柜等硬件资源的出租，**TaaS直接通过token交付模型推理输出结果，**把模型能力以统一、可量化、可采购的方式变成业务消耗品，推动产业进一步向精细化资源计量化升级。

AI需求的重心从训练逐步转向推理，Token调用量大幅提升，推动TaaS高速发展。目前众多大模型需持续响应大量用户对话、搜索、编码、多模态内容生成及Agent任务需求，产生大量的推理需求，累计Token消耗总量数倍于训练需求，根据中国工程院院士邬贺铨的判断：**全球推理算力将占AI算力总负载的70%至80%，中国推理需求为训练的8倍。IDC预计到2028年，推理工作负载占比将达到73%。**2024年初我国日均Token调用量为1000亿，到2025年底跃升至100万亿，到2026年3月已突破140万亿，两年增长超1000倍，行业收入、发展预期逐渐向由Token消耗驱动。同时，大模型训练需求大多集中在头部模型公司，而推理需求会扩散到调用API做垂直应用的SaaS厂商、传统企业IT部门、科研机构及国家单位，部分中小企业难以具备采购整租集群、部署环境和维护模型等能力，**而TaaS能够把算力转化为稳定输出结果、且可以通过更细颗粒度的计量和更低门槛的接入，把分散需求汇聚成可持续收入，将成为算力租赁新商业模式的发展方向。**

图：IaaS-MaaS-TaaS对比

模式	交付产品	计费单位	核心要素	服务商能力
IaaS	算力资源	规模/时间	服务器规格、价格、交付日期	拿卡、机房、上架、运维
MaaS	模型能力/API	Token用量、API调用	模型效果、接口可用性	模型聚合、接口服务、平台接入
TaaS	Token产出，结果	Token用量	单位成本、吞吐、稳定性、ROI	调度、推理优化、模型适配、运营

Token重量更重质，TaaS的核心竞争力在于Token输出质量。对于TaaS服务商来说，不仅需要提供大量的Token，且要**保证调用Token的交付质量。**而交付质量是由首字延迟(回答多快出来)、生成速度(模型每秒生成多少字)、并发承载能力(同一时刻能服务多少用户)、服务稳定性(每天宕机时间长短)四个指标共同决定。在推理需求为主流的背景下，AI价值由交付Token的智能密度与交付时效共同决定。同时，目前AI Agent是持续长程执行任务，模型调用次数远高于普通聊天，负载呈碎片化、高频化、对延迟严苛，Token质量更为关键。

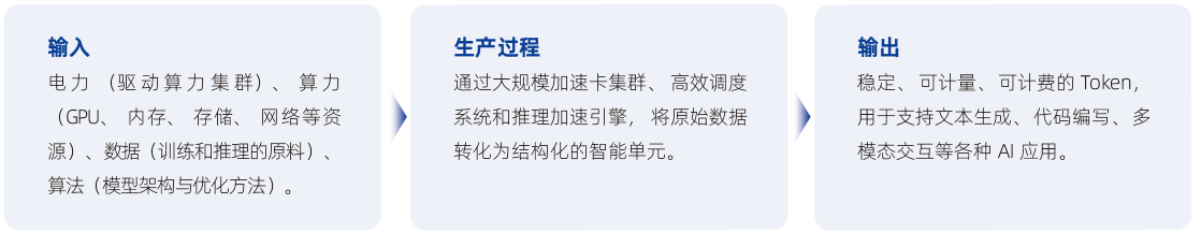
更多免费行业报告 并购家 www.ipoipo.net

1.6 Token工厂有望成为TaaS的实际落地方案

Token工厂有望成为TaaS的实际落地方案。Token工厂将存储数据和提供通用计算的传统数据中心转变为以大规模、高效率、低成本生产Token为核心目标的智能化生产设施，将算力出租场地转变为一条完整的智能生产线。**其核心逻辑是输入电力、算力、数据和算法，输出可量化、可计费的Token。**Token工厂集成了算力基础设施、模型与算法层、调度与分发系统MaaS平台等，构成了高效生产、调度和交付Token的智能化生产系统，通过对算力、模型、推理框架与调度系统进行统一优化，从而持续提升单位算力的有效Token产出和交付能力。因此，Token工厂本质上是产业从按量计费逻辑(MaaS)走向高质量交付体系(TaaS)的关键载体，有望成为TaaS商业模式的工程化体现。目前，北京、厦门、无锡等地已启动Token工厂的建设，产业将实现快速发展。

相较于传统算租模式，Token工厂毛利率更高，回本周期更短。根据黄仁勋的描述，**Token工厂的经济模型为：收入=Tokens/W×可用千兆瓦数。**所以Token工厂**盈利的两大核心驱动因素**即为效率-Tokens/W(反映了将电力转化为智能产出的效率。效率越高，单位成本越低，利润空间越大)。规模-可用千兆瓦数(决定工厂的产能天花板，受限于电力供给、芯片供应和基础设施建设。规模越大，总产出越高，收入上限越高。)**由于Token分成模式可以通过优化拉高单卡Token产出效率，且能分享下游模型API的终端溢价，相较于传统算租模式，利润率显著更高。**传统算力租赁模式的投资回收期通常为3—5年，而根据润建股份等企业的公开测算，在Token分成模式下，其投资回收期可缩短至2年左右。Token服务收入有望达到传统算力租赁的5-10倍。由于固定成本占比极高，Token分成模式的超额收益几乎全部转化为利润，平均毛利率可达50%以上，净利率可达30%以上，约为传统算力租赁的两倍。

图：Token工厂工作流程



图：Token工厂是第三代算力租赁基础设施

	传统IDC-第一代	云厂商-第二代	Token工厂-第三代
核心产品	机柜、带宽、电力	云服务器、数据库	Token、模型服务
计费方式	按机柜/带宽/月	按资源/小时	按Token/次
价值定位	卖基础设施	卖资源服务	卖AI能力
技术门槛	低	中高	高
毛利率	15%-25%	30%-50%	30%-70%
	IDC-卖基础设施 提供机房和电	云计算-卖资源 提供服务器和数据库	Token工厂-卖结果 直接提供大模型回答

资料来源：杰创智能，《2026年TokenFactory词元工厂技术与产业发展白皮书》，方正证券研究所

1.7 国内多款核心开源模型发布，推理需求大幅提升，国产GPU有望迎来机会

2026年4月-8月，DeepSeek、Kimi接连发布多个开源模型，其中DeepSeek V4 Flash代表标准化、大规模、低成本模型；DeepSeek V4 Pro代表高端Agent任务模型；Kimi K3代表超大规模、长程、多模态前沿能力模型。三者共同推动开源模型整体成为企业默认选型之一。

DeepSeek V4 Flash：2026年7月31日开启正式版API公测，采用MoE架构，总参数量达2840亿，每次推理激活约130亿参数，支持100万Token上下文，原生支持Responses API格式，可直接适配Codex工作流。**Flash版本**在Agent能力上大幅跃升，Terminal-Bench 2.1得分达82.7，具备出色的端到端复杂任务执行能力。DeepSeek V4 Flash聚焦极低推理成本与快速响应速度，**API输入价格低至约1元/百万Token，**具备极高的性价比，成为承担编程、长文本及Agent工作负载的主力模型，**在低成本推理场景中确立了稳固的商业定位。**

DeepSeek V4 Pro：正式版于2026年8月12日上线。模型采用MoE架构，总参数量达1.6万亿，每次推理激活约490亿参数，支持100万Token上下文。首创压缩稀疏注意力与重度压缩注意力机制，将单Token推理计算量降至前代的27%，KV Cache压缩至V3.2的10%，大幅优化长序列推理成本及效率。**Pro在编程与智能体细分赛道上接近SOTA级别。**模型在生态兼容与商业定价上具备极强竞争力。随着V4PRO版本的正式上线，DeepSeek上调了API定价，且将采取峰谷定价。**目前DeepSeek-V4-Pro的价格在高峰时段为每百万tokens输入（缓存命中）0.04美元、输入（缓存未命中）1.32美元、输出3.96美元。**

Kimi K3：2026年7月开源，总参数量2.8万亿，为全球首个近三万亿参数级的落地开源模型，支持100万Token超长上下文窗口。创新混合注意力机制，有效解决长序列显存占用问题并保留全局检索能力。引入注意力残差机制，大幅提升深层模型训练稳定性。采用Stable LatentMoE超稀疏架构，896个专家中单Token仅激活16个，整体扩展效率较前代提升2.5倍。Kimi K3在Frontend Code Arena榜单以1679分登顶，成为首个在该权威编程榜单上超越全部海外闭源模型的开源大模型，且可在48小时内完成45nm工艺AI推理芯片全流程设计，**综合能力接近旗舰闭源大模型。其单任务成本仅约0.94美元，推理成本约为Claude Fable 5的三分之一、略低于GPT-5.6 Sol。**

图：国内三大开源模型具体信息

模型	发布时间	总参数	激活参数	上下文长度（Tokens）	原生模态	开源具体状态	核心技术	定位
DeepSeek V4-Pro	2026-08-13	1.6T	49B	1M	文本	API+权重开源	MoE、CSA+HCA/DSA稀疏注意力、FP4+FP8混合精度	Agent编码开源SOTA、世界知识领先多数开源、数理与代码对标头部闭源
DeepSeek V4-Flash	2026-07-31	284B	13B	1M	文本	API+权重开源	MoE、轻量化注意力压缩、FP4+FP8混合精度	推理能力、简单Agent任务表现接近Pro；响应更快、API更经济
Kimi K3	2026-07-17	2.8T	—	1M	原生视觉/多模态	权重完整开源	KDA混合线性注意力、AttnRes、Stable LatentMoE、896专家激活16个	综合性能与多模态理解

图：DeepSeek V4 Flash/PRO峰谷定价详情

模型与时段	缓存命中输入	缓存未命中输入	输出
V4-Flash 调整前	0.0028	0.14	0.28
V4-Flash 非高峰（新）	0.007	0.22	0.66
V4-Flash 高峰（新）	0.014	0.44	1.32
V4-Pro 调整前	0.003625	0.435	0.87
V4-Pro 非高峰（新）	0.022	0.66	1.98
V4-Pro 高峰（新）	0.044	1.32	3.96

资料来源：DeepSeek，Kimi，方正证券研究

更多免费行业报告 并购家 www.ipoipo.net

1.7 国内多款核心开源模型发布，推理需求大幅提升，国产GPU有望迎来机会

开源模型MoE+长上下文+Agent三大特性进一步推动总算力&推理需求增加。1.头部开源模型普遍采用MoE架构，MoE虽将单次推理计算量压缩至传统模型20%，但日均50亿次的调用规模的扩张远超单次成本的降幅，推理需求总量不降反升。**2.**部分开源模型的上下文长度已迭代至1M Tokens，每生成一个新Token都需要读取更多历史信息，KV Cache显存占用将大幅提升，难以抵消上下文数十万甚至百万级别的Tokens带来的推理消耗增量。**3.**Agent已成为大模型应用落地主流形式，Agent需要反复执行规划、调用工具、读取结果、纠错，且多Agent系统还会增加任务拆分、Agent间交接和结果汇总。Anthropic披露其单Agent约消耗普通聊天4倍的Token，多Agent系统约为15倍，直接放大了推理侧的总算力需求。

开源模型大幅降低使用与开发门槛，客户群体更加丰富，TaaS逻辑进一步加强。闭源时代，企业调用AI能力往往受制于高价格的API、较弱的定制化能力以及敏感数据和工作流难以私有部署。开源模型带来了权重开放、许可友好等优势，企业可以在本地/私有云自行部署，大量原本无法承担闭源大模型长期API成本、也无法接受数据外流风险的企业，开始具备部署条件。目前有超20家央企接入DeepSeek，涉及能源、通信、汽车、金融、建筑等多个领域，同时多地政府同步跟进部署。根据前文我们的判断，客户群体的扩散有望进一步加强TaaS落地进度及逻辑，整体看，开源模型的部署➡丰富客户群体➡TaaS逻辑加强。

开源模型适配国产芯片架构，供给侧格局有望迎来改革。GLM-5已适配华为昇腾、寒武纪、海光、摩尔线程等七款国产芯片；DeepSeek-V4在发布时已同步在昇腾等国产算力集群完成部署，可同时兼容英伟达CUDA及国产算力框架。并且，国内已出现依靠国产算力完成训练与推理的万亿参数模型，LongCat-2.0在国产算力上完成全流程训练，峰值规模超过5万张国产算力卡。在推理与部署端，国内算力芯片与开源模型已有一定适配度，在英伟达等海外芯片进口持续收紧的背景下，国产芯片有望从政策替代品转向具备成本和可得性优势的新选择。短期内高端训练&复杂推理任务仍可能采用海外GPU，而成本敏感型场景更容易向国产芯片迁移，且国产芯片份额有望随技术进展不断加大，算租供给将迎来大幅增长。

从右图可以看出，除新推出的匿名模型Ox Alpha外，DeepSeek V4 Flash、MiMo V2.5、Hy3、GLM5.2等国产开源模型近期的调用量位居世界前列。随着下游应用端的发展，开源模型的调用渗透率有望进一步增加，届时能够部署和调用模型的主体数量有望成倍增加，而每个主体又在各自场景中的用量持续加深，Token调用量有望呈指数级增长。

图：OpenRouter上周度大模型调用Token量排名(数据截至8月26日)

1.	Ox Alpha by stealth	23.2T tokens new	11.	MiniMax M3 by minimax	1.57T tokens ↓6%
2.	DeepSeek V4 Flash 0731 by deepseek	11.6T tokens ↑3%	12.	Laguna S 2.1 (free) by poolside	1.53T tokens ↓8%
3.	MiMo-V2.5 by xiaomi	9.92T tokens ↑82%	13.	Claude Opus 5 by anthropic	1.52T tokens ↓43%
4.	Hy3 by tencent	7.16T tokens ↓27%	14.	GPT-5.6 Sol by openai	1.51T tokens ↑84%
5.	DeepSeek V4 Flash 0423 by deepseek	5.58T tokens ↑17%	15.	Kimi K3 by moonshotai	1.42T tokens ↑8%
6.	Nemotron 3 Ultra (free) by nvidia	5.4T tokens ↑100%	16.	Claude Sonnet 5 by anthropic	1.12T tokens ↑3%
7.	GPT-5.6 Luna by openai	4.16T tokens ↓28%	17.	Nemotron 3.5 Lightning (free) by nvidia	1.07T tokens ↑28%
8.	GLM 5.2 by z-ai	3.22T tokens ↓26%	18.	DeepSeek V4 Pro 0813 by deepseek	1.01T tokens ↑7%
9.	Gemini 3.7 Flash by google	2.58T tokens ↑210%	19.	Gemini 3 Flash Preview by google	830B tokens ↑1%
10.	DeepSeek V4 Pro 0423 by deepseek	1.84T tokens ↓29%	20.	GLM 5.3 by z-ai	803B tokens ↑>999%

资料来源：OpenRouter，方正证券研究所

1.8 1GW数据中心GPU价值量占比接近56%，GPU等短期资产为AI算力扩张主要投入方向

根据Epoch AI最新的拆分，建设一个1GW数据中心(搭载NVIDIA GB200 NVL72GPU)的成本大概为378.8亿美元，其中GPU成本大概为211.88亿美元，占比56%，其次是机房设施(30%)、网络设施(13%)、土地与变电站等配套(1%)。

微软修改数据中心使用寿命，数据中心资本开支重心将从长期基建转向GPU等短期资产，后续短期资产有望能真实反映算租利润节奏。近期，微软宣布自 FY27 起将数据中心和办公楼预计使用寿命由**15年延长至25年**，**并使更多未来数据中心租赁由融资租赁转向经营租赁，**通过这一调整，微软将2026日历年的CapEx账面预期从此前的约1900亿美元缩减至约1750亿美元，但剔除使用寿命变化带来的租赁重分类后，总CapEx保持不变，Q4单季CapEx与融资租赁合计410亿美元，27Q1指引超过500亿美元。微软表示，目前CapEx约三分之二投向短期资产（CPU/GPU），而土地和数据中心建设只是较小部分，核心短期资产需求仍旧强劲。**微软此次的调整明确了数据中心内部资产已分为两类，1.楼宇、园区、机房和机电等长寿命基础设施资产，适合以更长折旧周期和经营租赁方式承接，其CAPEX本身属性被弱化；2.CPU/GPU等折旧期限在5年以内的短寿命算力设备，仍是AI算力扩张的主要投入方向，是算力租赁景气、利润率和估值弹性的真正决定因素。因此，除CapEx外，经营租赁承诺、短资产投放、利用率等指标后续也将成为判断行业景气度的额外核心指标。**

根据测算，假设中长期资产按照25年折旧，GPU等短期资产按照5年折旧，GPU折旧年化CapEx为42.38亿美元。根据Epoch AI的测算，1GW数据中心年度运营支出包含能源电费，税费等，则GPU在年化支出(CapEx+Opex)中的占比接近73%，远超其在前期支出的占比，侧面验证了GPU等短期资产是AI算力扩张的主要投入方向。

图：1GW数据中心成本拆分与年化费用测算

资产属性(折旧年份)	1GW数据中心建设成本项	总CapEx(亿美元)	占总CapEx比例	年化折旧(亿美元)	占年化CapEx+Opex比例	年化Opex(亿美元)		占年化CapEx+Opex比例
短期资产(5年)	服务器（含 GPU，采用 NVIDIA GB200 NVL72）	211.88	55.93%	42.38	72.96%	能源电费	5.94	10.23%
中长期资产(25年)	机房设施（建筑外壳、机电系统、液冷配套等）	114.33	30.18%	4.57	7.87%	各类税费	1.43	2.46%
	网络基础设施	49.25	13.00%	1.97	3.39%	设施维护	1.2	2.07%
	土地	1.72	0.45%	0.07	0.12%	人工成本	0.4	0.69%
	公用工程（变电站等配套）	1.64	0.43%	0.07	0.11%	用水成本	0.06	0.10%

资料来源：Epoch AI，Microsoft，方正证

更多免费行业报告 并购家 www.ipoipo.net

1.9 GPU等短期资产是景气核心，折旧到期后将成为高利润率资产

CoreWeave的A100集群仍获新签长约，旧GPU仍具商业寿命。根据公司26Q2业绩会口径：其2020年推出的A100在原始合约到期后，被重新签下一份延续至2029年、价格极具吸引力的合同。管理层同时强调Ampere/Hopper老卡已基本售罄，老一代GPU的售价维持在或高于约一年前的水平。公司整体商业模式并不依赖初始合约到期后的再租赁，每一次转租/续租都是已赚回收益之上的增量。这意味着GPU的真实经济寿命并不等于最前沿训练寿命，A100等旧卡在推理、企业部署、托管推理等场景中仍能持续货币化。同时因为老款GPU均为已完成折旧，续签收入几乎全额转化为纯利，大幅拉高整体利润率。

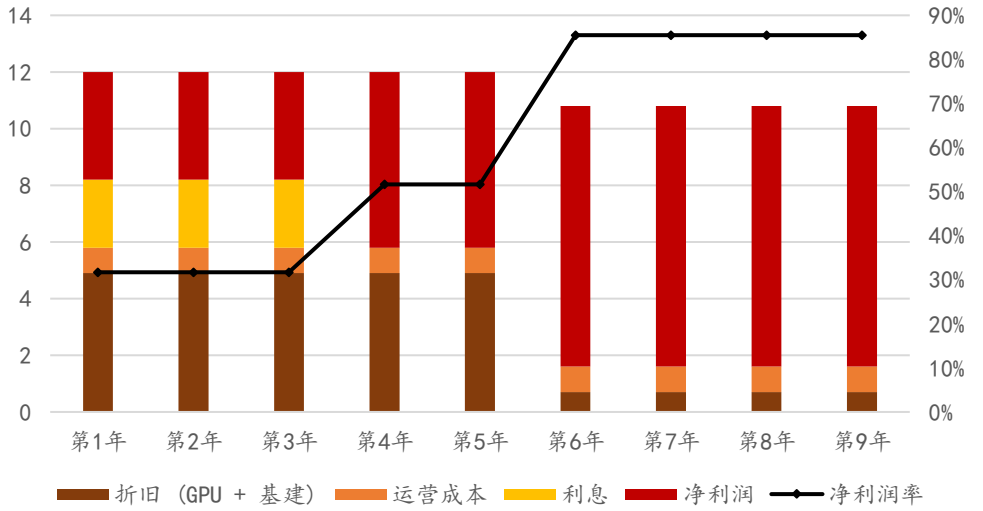
大量推理需求推动旧卡“第二生命周期”。根据英伟达在2023年的表态，当时大部分实时推理已经在上一代GPU上运行。推理场景可以容纳旧代GPU，只要模型能够在既定延迟与成本约束下正常运行即可。**推理(托管)、轻量训练、企业私有化部署等场景正在承接大量旧卡需求**，目前数据中心GPU的生命周期已经从发布 → 训练主力 → 淘汰，逐渐转换为前沿训练 → 次前沿训练/混合负载 → 推理 → 企业场景/托管推理 → 低成本长尾客户需求，可分为：折旧期(前5年，take-or-pay长约)负责还清债务、赚回全部Capex；折旧到期后进入“第二生命周期”(转租/短约/托管推理)，此时收入虽会小幅下降，但因折旧清零、债务已偿、Opex极低，单位利润率反而抬升，有望成为算租的新利润增长。

根据测算，旧GPU在折旧到期后若获得新签合约，利润率将翻倍。

模型关键假设：投资与成本按前文的测算。GPU5年直线折旧（年折旧≈$4.24B/GW），折旧期满后GPU账面成本归零、仅余基建长寿命折旧(基建+网络按 25 年折旧，≈$0.67B/GW)；假设债务比例70%、利率9%，经营现金流在第3年偿清资产债务、利息归零。年收入取行业口径$12B/GW；续约期收入保守取90%（较CoreWeave95%口径略低）。

图：1GW数据中心逐年经济性+GPU5年折旧模型(单位：十亿美元；左表右图)

年份	年收入	折旧(GPU+基建)	运营成本	利息	净利润	净利率	阶段
1	12	4.9	0.9	2.4	3.8	31.70%	折旧期
2	12	4.9	0.9	2.4	3.8	31.70%	折旧期（GPUcapex回本）
3	12	4.9	0.9	2.4	3.8	31.70%	折旧期
4	12	4.9	0.9	0	6.2	51.60%	折旧期（capex回本、债务偿清）
5	12	4.9	0.9	0	6.2	51.60%	折旧期末
6	10.8	0.7	0.9	0	9.2	85.50%	续约期（GPU折旧清零）
7	10.8	0.7	0.9	0	9.2	85.50%	续约期
8	10.8	0.7	0.9	0	9.2	85.50%	续约期
9	10.8	0.7	0.9	0	9.2	85.50%	续约期（对应A100 2020→2029）



资料来源：CoreWeave，英伟达，方正证券研究所测算

方正证券
FOUNDER SECURITIES
正 在 你 身 边

二、从Hyperscaler到Neocloud：北美算租进入需求外溢与业绩兑现期

更多免费行业报告 并购家 www.ipoipo.net

2.1 北美算力租赁参与者可分为四类

北美由于GPU性能更强、云厂商积累较深等原因，算力租赁市场更为成熟，可以被分为Hyperscaler、Neocloud、挖矿转型厂商及边缘/推理云四类厂商。

Hyperscaler(超大规模服务商)：以极大规模运营数据中心和云计算服务的企业，通常包括头部公有云与互联网巨头。Hyperscaler在全球部署大量的大型数据中心，能够按需弹性地扩展算力、存储与网络，并把算力以云服务形式对外出租或自用。同时，**Hyperscaler不仅外租算力，同时也会配套相关SaaS/PaaS能力，例如**AWS通过Bedrock等实现从底层到应用层的收入传导；微软则通过Azure+Copilot生态，将GPU资源内化为更高毛利的软件与平台收入。

Neocloud：专注提供GPU算力租赁服务的新型云服务商。这类纯GPU云服务商能为客户提供顶尖的性能与部署灵活性。相较于Hyperscaler，Neocloud专注GPU基础设施交付，但不追求全栈公有云生态，其更擅长以长约和专属集群绑定大客户，客户群体甚至包含部分Hyperscaler。

挖矿转型厂商：此类公司更类似于高功率数据中心&托管资产供应商，在比特币挖矿时代积累了大量的电力、土地、并网许可和已接入电网的数据中心，在AI时代将BTC挖矿场转型为GPU托管与AI数据中心出租。

边缘/推理云：在全球推理需求高增的背景下，边缘云厂商围绕低延迟、分布式调用与token计费做推理层服务。

北美算力租赁的收费模式主要分为长协/预留合约、裸金属、按需租赁和现货/抢占式四类。长协和预留合约通常采用多月到多年的take-or-pay机制，以更低单价换取容量保障和价格确定性，适合长期训练与稳定推理负载。**裸金属模式**则以整节点、整服务器或专属集群交付，强调独占物理资源、低虚拟化损耗和高性能互联，更适合大规模训练；**按需租赁**以按GPU小时计费为主，灵活性最高，但价格显著高于长约；**现货/抢占式租赁**则利用闲置产能提供折价算力，成本最低但无稳定性保障，主要适用于可中断、容错型任务。整体看，海外市场的定价逻辑是保障越强、期限越长，单位价格越低；灵活性越高、即时性越强，单位价格越高。

图：北美算租厂商分类

分类	代表公司	核心产品	客户	收入模式	核心竞争力
Hyperscaler	微软/AWS/谷歌/Oracle	全栈云+AI服务	企业、Frontier labs、开发者	云订阅+算力+平台	生态、客户、分发能力
Neocloud	CoreWeave/Nebius	GPU基础设施租赁	Frontier labs、云厂商、量化/企业	长协+按需+预付款	快速交付、NVIDIA关系、融资
挖矿转型厂商	IREN/Hut8	AI数据中心托管/长期租约	hyperscaler/neocloud	长期租约+托管费	电力、土地、审批、交付
边缘/主权云	Cloudflare	推理/主权算力	区域政府、企业、开发者	token/API/合规云服务	合规、低延迟、区域资源

图：算租厂商收费模式分类

收费模式	合同形式	计费方式	价格特征	适用场景
预留合约&长协/Reserved&committed	3个月、6个月、1年或多年的take-or-pay	按GPU小时或按固定包量计费	通常较按需低30%-60%，部分长约折扣更高	7×24小时推理、长期大规模训练、稳定产能需求
按Token收费/Per-token	随时使用，按照实际调用的Token收费	按消耗的Token收费	无额外成本，无最低收费	多种推理需求，开发，难以承担GPU等硬件投入
裸金属/Bare Metal	专属服务器/整节点/整机柜租赁	按节点、整机、整集群或包月包年计费	单卡折算价格未必最低，但性能损耗更小	大规模训练、低虚拟化开销、高带宽互联场景
按需租赁/On-Demand	无长期承诺，随开随用	按GPU小时计费	通常较预留价格显著溢价	短期实验、弹性扩容、开发测试
现货/抢占式/Spot	无容量保证，可被中断	按GPU小时动态计费	通常较按需低50%-80%	容错批处理、非关键训练任务

资料来源：AWS，微软，packet.ai，readaitime，方正证券研究所

2.2 三大Hyperscaler业绩验证Capex-营收正循环，客户由Frontier Labs向B端渗透

微软：Capex投资回报逐渐兑现，客户向B端扩散。公司Intelligent Cloud收入393亿美元、同比+32%；**Azure(公有云)与其他云业务同比+43%，**均高于市场预期。**云业务：**公司指引FY27Q1（自然季26Q3）增速预计达到45%。Microsoft Cloud全年收入已超过2140亿美元，其中90%来自前沿模型公司之外的客户。未履行合同达到6780亿美元、同比+84%(扣除OpenAI订单仍+25%)，且约30%将在未来12个月确认，且环比增加的510亿美元全部来自非AI模型开发客户。云侧订单并未因为巨额Capex而被消化完，且订单储备仍在快速累积，同时需求端客户已从Frontier Labs主导向更广泛企业端扩散。同时，365 Copilot付费席位超过3000万，GitHub Copilot用户达5000万，Agent 365两个月内注册近4000万智能体，训练+推理+应用层消费需求并行增长。**Capex：**26年全年Capex约1750亿美元。主要系更多未来数据中心租赁由融资租赁转为经营租赁、不再计入CapEx账面，导致预期由1900亿美元下修，但实际支出不变。

AWS：需求端仍然旺盛，逐渐迈入芯片+平台+算力集成式服务商。AWS26Q2收入422.3亿美元，同比增长37%，创下18个季度以来最快增速；营业利润达到166亿美元，同比增长63%，营业利润率提升至39.4%。公司AI业务年化收入规模已超250亿美元，并保持三位数增长；自研芯片业务年化收入同样超过250亿美元。Q2在手订单4960亿美元，为历史新高。Capex层面，AWS上调全年Capex至2200亿美元，并表示即便上调规模，2026年仍无法满足全部需求，且这一动态在2027年也将持续。AWS直接指出“供给建设速度<需求增长速度”，Capex投入仍将持续加大，且再次印证在推理需求盛行的背景下，云厂商已经开始把芯片、平台与推理服务集成，提供全方位服务。

谷歌：云业务高速增长，传统云厂商仍需要Neocloud算力供给。Google Cloud26Q2收入达到247.7亿美元，同比增长82%，云业务营业利润达到88.1亿美元，同比大幅提升，盈利能力显著增强，AI基础设施供给扩张在规模效应与高附加值订单推动下将放大利润。谷歌云将26年全年Capex指引由原来的1800-1900亿美元上调至1950-2050亿美元。Q2单季Capex达到449.2亿美元，同比接近翻倍。谷歌云backlog达到5140亿美元，环比增加540亿美元，且超过50%预计未来24个月确认，订单储备同样高位增长。Gemini处理量约220亿tokens/分钟，App月活9.5亿(市场预期9.2亿)，推理需求正从实验阶段走向规模化商业阶段。同时，谷歌明确表示Q3将扩大第三方算力租赁作为过渡，会对短期Cloud利润率形成压力，侧面证明了Neocloud等新算租厂商已进入顶级云厂商的算力供给体系，双方并不全部为竞争关系。

图：北美三大Hyperscaler26Q2业绩总结

	微软	亚马逊	谷歌
26Q2云收入(亿美元)	Intelligent Cloud 393	AWS 422.3	Google Cloud 247.7
云收入增速	32%	37%	82%
云利润率	40.70%	39.40%	35.60%
backlog/RPO(亿美元)	6780	4960	5140
当季capex(亿美元)	410	542	449
全年capex指引(亿美元)	1750	2200	1950-2050

三大Hyperscaler上半年云业务增速全面加速，虽然自由现金流由于高投入短暂承压，但在手订单非常充足，资本开支仍处高位，代表着AI算力需求目前并未见顶，27年投入仍将加速，且前期Capex投入逐渐兑现为业绩，商业化模式初步跑通。同时，推理需求大幅加速，客户逐渐从Frontier Labs向企业端渗透，推动大云厂将服务模式拓展至芯片研发、平台能力、推理服务等方向，逐渐向MaaS/TaaS转型。并且向Neocloud等第三方算力租赁的需求外溢仍然存在，Neocloud有望相应受益。

资料来源：AWS，微软，谷歌，方正证券研

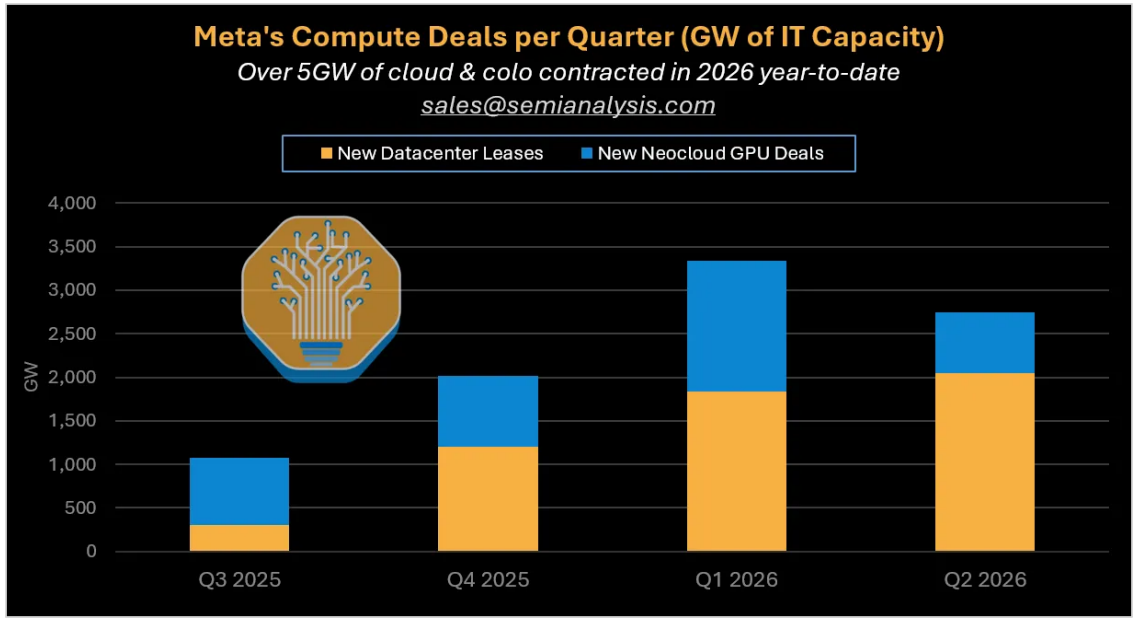
更多免费行业报告 并购家 www.ipoipo.net

2.3 Meta转向算力运营，Neocloud与Hyperscaler互存生态逐渐建立

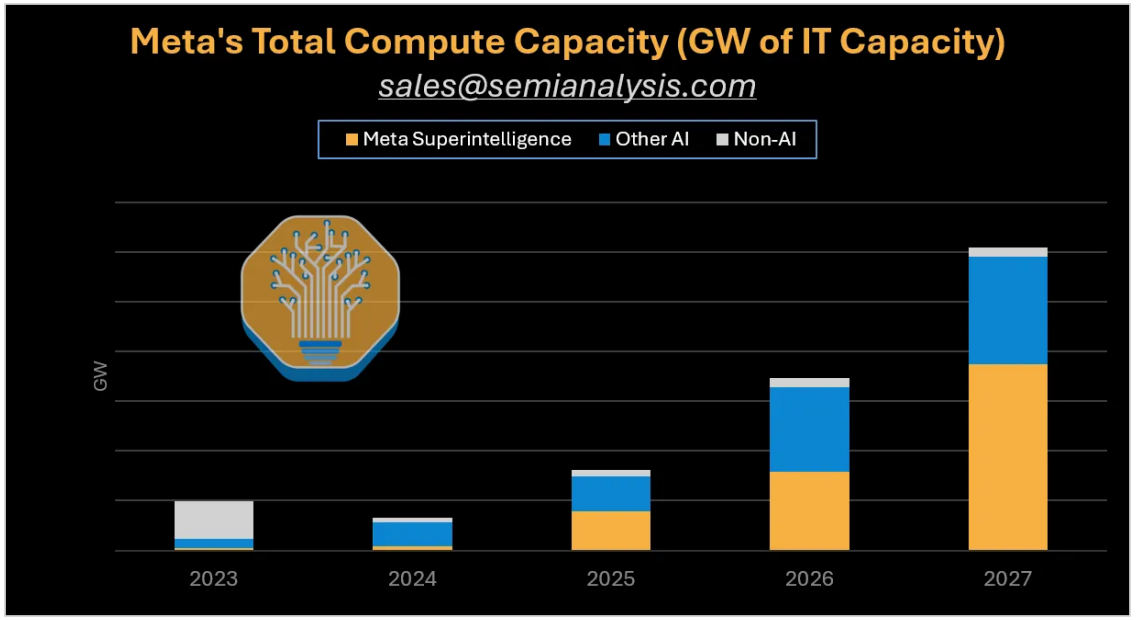
Meta转向算力买方与机会型卖方的双重身份。26年7月1日，据彭博社报道，Meta正筹建云业务Meta Compute，计划提供两种业务模式—类AWS Bedrock的托管模型/API服务，以及类Neocloud的裸金属算力出租。7月中下旬，纽约时报表示Meta正与Anthropic洽谈两年最高约100亿美元、90天可取消的算力合约。7月29日，MetaQ2财报将全年Capex指引下限上调(1300-1450亿美元)，扎克伯格表示Meta算力将有“相当大一部分”用于驱动其自有模型和产品，并且Meta将与BlackRock合资在德州建1GW数据中心。8月5日，Meta正式对外宣布启动算力商业化战略，开始向外部客户出租自有GPU资源，自此，Meta正式成为算力需求方与供应方，可以理解为将上一代闲置算力资产化，为大量Capex补充变现路径，但同时对高端算力需求仍未变化，算力总需求仍在高速增长。

SemiAnalysis表示Meta数据中心与算力采购只会加速、分层业务使得公司逐渐向算力运营商转型。根据SemiAnalysis的报告，截至26H1，Meta已签约超5GW云+托管容量(不包含2.5GW自建项目)、2024年初至今累计近10GW。Meta利用这些算力提供四大差异化的高毛利业务：自身的前沿模型训练(Meta Superintelligence Labs)、广告推荐系统复杂度10倍扩容、类Bedrock的前沿模型私有部署服务以及“SpaceX式”的大规模高溢价按需算力出租(仅需几百MW就能带来每年超过100亿美元的营收)。四类高价值业务使得Meta敢于激进签约算力，即便前沿大模型进度不及预期，仍有大量高毛利替代方案来出售算力，Meta甚至有望向其他Neocloud签约订单来加速算力规模扩建，在几大Neocloud明确订单已经签满的背景下，供需缺口仍然存在，Hyperscaler中短期内并不会与Neocloud就瓜分算租市场产生正面冲突，而是与Neocloud互相合作建立共存生态。中长期来看，几大Hyperscaler逐渐向平台化(推理/软件/调度)转型，海内外TaaS逻辑被进一步强化，核心竞争力逐渐从买不买得到GPU转向可用电力、数据中心工程交付与设备上线速度。

图：Meta各季度算力签约规模



图：Meta每年算力扩张进度



2.4 Neocloud业绩开始大规模释放，商业模式逐渐向平台化服务过渡

CoreWeave：在手订单大幅增长，客户群体开始扩散。26Q2公司实现营收25.75亿美元，同比+112%，净利润-6.26亿美元，同比-115.86%，调整后EBITDA15.1亿美元，同比翻倍，Capex94亿美元，略高于预期；上半年收入46.53亿美元，同比+112%，净利润-13.66亿美元，同比-125.79%；调整后EBITDA26.68亿美元，同比+67.69%。Q2末公司在手订单1040亿美元，同比增长246%，Q3初新增超过250亿美元客户承诺订单，总量接近1300亿美元，**存量订单中超过50%已经进入交付状态**，公司预计年底前这一比例将超过三分之二，且新签合同利润率比近期合同高5%-10%，价格端，7月公司全线SKU提价约25%，并将上游硬件涨价向下传导，CFO明确表示算力交付创造的价值增幅已超过硬件成本涨幅。二季度公司客户名单新增卡特彼勒(机械)、Isomorphic Labs(医药)、Flow Traders(量化交易)、Leidos(国防)等应用端企业，进一步验证AI算力需求已从Frontier Labs流向应用端。全年指引：营收124亿-132亿美元、资本开支350亿-390亿美元、有效电力1.85GW以上。

Nebius：营收高速扩张，现金流十分健康。26Q2公司实现营收5.82亿美元，同比+454%，净利润-1.9亿美元，同比-132.58%，调整后EBITDA2.36亿美元，同比实现扭亏，Capex56.6亿美元；上半年收入9.81亿美元，同比+529%，净利润4.31亿美元，同比-8.52%；调整后EBITDA3.66亿美元，同比扭亏。公司现金流十分健康，26Q2经营性现金流22.46亿美元，大幅转正，预计将随业务规模化显著增加；全年客户预付款将达到90亿美元以上，覆盖对应合同资本开支的50%-60%。公司在手订单达400亿美元，且在Q2签下四笔超过10亿美元的标杆合同，分别来自Reflection、Cohere、某Neolab及一家美国大型量化交易公司。全年指引：营收30-34亿美元、年化收入70-90亿美元、资本开支200亿-250亿美元、调整后EBITDA利润率约40%。

Cloudflare：营收增速再上台阶，Agentic Internet叙事开始兑现订单。26Q2公司实现营收6.96亿美元，同比+35.9%（超出指引上限6.65亿美元），净利润-1.70亿美元（含1.51亿美元裁员重组费用，去年同期亏损0.50亿美元），自由现金流0.56亿美元，同比+69%；Capex约占营收7%（约0.49亿美元），延续轻资本开支模式。上半年收入13.36亿美元，同比+34.74%，净利润-1.93亿美元，同比-116.99%。Q2末公司在手订单27.3亿美元，同比+38%、环比+7%，其中当期RPO同比+35%、占比64%，订单增速连续跑赢收入增速。客户结构持续向上突破：年化收入超10万美元大客户达4698家，同比+27%，环比净增282家（过去一年净增986家创新高），贡献营收占比73%；单季新增付费客户超8万家，新客签单增速创五年新高。**公司明确**互联网正为机器间流量重写，Q2公司网络流量中超过一半已为非人类流量，公司推出智能体支付与变现基础设施，并与OpenAI开展AI公司与内容方关系的研究试点。全年指引上调：营收28.64-28.70亿美元（前值28.05-28.13亿美元，对应同比+32%），non-GAAP运营利润4.43-4.45亿美元。

稀缺的是可交付算力，商业模式从租赁逐渐向平台过渡，Neocloud进入量价齐升的收获期。目前，**海外算力租赁需求的真实性与广度均已得到验证，**CoreWeave、Nebius等Neocloud在手订单的上涨不只因为Frontier Labs的模型军备竞赛，更多因为应用端大量企业入场，意味着AI算力正从模型层的资本开支转化为全行业的生产要素采购，需求结构的多元化显著降低了单一客户砍单的尾部风险。**同时，目前算租供需缺口持续加大，叠加前文提及的旧卡仍存在价值，说明当前稀缺的是可交付的算力，**使得存量GPU资产具备保值能力、增量高端GPU具有独特的定价权(CoreWeave、Nebius的高价卡需联系销售才能定价)，GPU通缩情况暂未出现。**并且商业模式正在发生质变，推理托管、存储网络软件等高毛利业务放量，意味着头部Neocloud正从重资产租赁商转向AI云平台(MaaS/TaaS)，**利润率结构的改善有望对冲逐年的Capex投入。厂家之间的竞争已从GPU采购能力转向电力资源锁定能力—GW级电力储备可以锁定未来三年的收入上限，谁能把签约电力按时转化为有效电力，谁就能承接溢出订单，电力交付节奏将同样成为反映行业供给的核心指标。**综合来看，本轮Neocloud中报确认了海外算力租赁“量价齐升+客户扩散+电力卡位”的三重逻辑，板块核心矛盾正从需求真实性转向交付能力与资金成本的竞赛。**

更多免费行业报告 并购家 www.ipoipo.net

2.5 CoreWeave：深度绑定英伟达的Neocloud龙头

公司将自己定位为The Essential Cloud for AI(AI必备云)，专注为AI计算密集型工作提供GPU/CPU架构云平台。 CoreWeave成立于2017年，最初作为加密货币挖矿公司运营，2019年更名为CoreWeave，利用其庞大的GPU库存开始为企业提供云计算基础设施。2022年，公司投入一亿美元用于购买英伟达H100 GPU，并且英伟达在2023年向公司投资一亿美元。在Chatgpt推动大量训练需求的背景下，凭借独特资源实现业务的高速扩张。2025年3月，公司在纳斯达克完成上市，代码CRWV。

公司主要产品为大规模GPU集群的按需租赁和长期合同服务，提供“硬件 + 调度 + 运维 + 推理”的全套AI算力服务体系。硬件层面，公司提供 GB300、GB200、B300、B200、H100、H200、L40S、A100 等全套GPU产品，多数产品为8-GPU集群，并通过InfiniBand实现高速互联。**软件层面，**公司提供CoreWeave Kubernetes Service（CKS），让客户可以像管理软件一样管理GPU集群，并且推出CoreWeave SUNK(业内首款面向最高负载AI工作负载的一体化训练系统，可为大型长周期训练任务提供可靠性与全链路运维可视化）。SUNK实测可实现最高96%training goodput、97－98%有效训练时间、千卡集群平均故障间隔提升约10倍，使得公司的GPU利用率较行业平均水平高出20%。软硬一体的能力使得公司在大规模训练环境下能维持较高的运行效率。

公司深度绑定英伟达。截至目前，英伟达持有CoreWeave超过7%的股份，并与其签订了63亿美元的算力容量兜底采购协议。26年6月，公司宣布已成功完成NVIDIA Vera Rubin NVL72系统的业界首次机架级部署与全面验证，进一步加深与英伟达的绑定关系。

公司2025年拥有43个数据中心、60万张 GPU、活跃电力850MW、签约电力3.1GW；管理层在26Q2财报会上表示，签约电力已达4.2GW，另有超1.5GW潜在储备。

公司的核心客户包含英伟达(长期合作)、微软、Meta（累计约350亿美元）、OpenAI(累计约224亿美元)、Anthropic(26Q2新签)等Hyperscaler及Frontier Labs，同时也包含众多金融机构、AI初创公司等应用端客户。

图：公司部分GPU租赁价格

GPU 型号	GPU 数量	显存（GB）	vCPU 数量	系统内存（GB）	本地存储（TB）	按需价格（每小时）	竞价实例价格（每小时）	推理单 GPU 价格（每小时）[2]
NVIDIA GB300 NVL72	4^1	279	144	960	61.44	联系销售	N/A	联系销售
NVIDIA GB200 NVL72	4^1	186	144	960	30.72	$42.00	N/A	$10.50
NVIDIA HGX B300	8	270	192	4096	61.44	联系销售	$35.84	联系销售
NVIDIA HGX B200	8	180	128	2048	61.44	$68.80	$34.11	$8.60
NVIDIA RTX PRO 6000 Blackwell Server Edition	8	96	128	1024	7.68	$20.00	$11.09	$2.50
NVIDIA HGX H100	8	80	128	2048	61.44	$49.24	$19.71	$6.16
NVIDIA HGX H200	8	141	128	2048	61.44	$50.44	$20.93	$6.31
NVIDIA L40S	8	48	128	1024	7.68	$18.00	$7.88	$2.25
NVIDIA A100	8	80	128	2048	7.68	$21.60	$9.65	$2.70

图：公司SUNK产品优势



资料来源：CoreWeave，方正证券研究所

2.6 Nebius：全球性全栈AI云

Nebius Group原为Yandex NV，是俄罗斯科技公司Yandex的荷兰母公司，自2011年起在纳斯达克上市，但在2022年2月由于俄乌冲突停牌，24年7月，Yandex将所有俄罗斯资产出售，保留下来的国际业务重组为Nebius Group，专注于AI基建，并于2024年10月以Nebius身份恢复交易，代码NBIS。

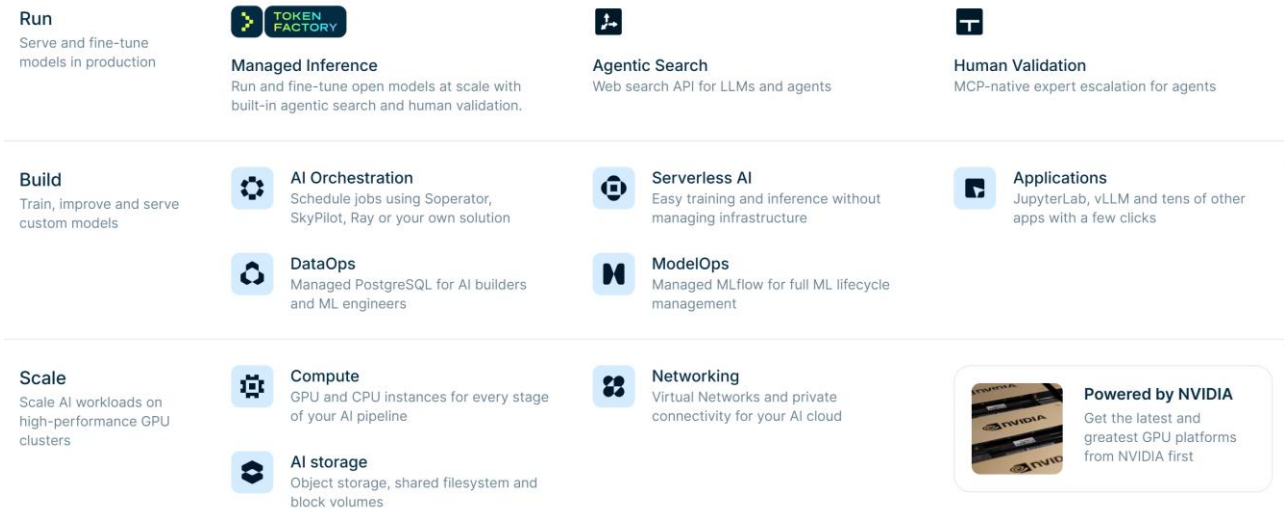
公司定位为全栈AI云平台，提供Token工厂、模型调优等服务。 Nebius的核心产品是Nebius AI Cloud“Aether”，定位全栈AI云平台，基于NVIDIA加速计算平台提供按需GPU计算、高性能存储与推理托管服务。公司配套提供模型调优平台、Token Factory托管推理平台以及Nebius Echo AI agent运维产品。其中，Token Factory在2026Q2接入Eigen AI、Clarifai，并新增Kimi K3、GLM 5.2、Nemotron Ultra等开放模型，推理工作量环比增至3倍以上。

截至目前，公司总共在全球运营&规划了大量数据中心，包括美国的新泽西Vineland、密苏里Independence、宾夕法尼亚；芬兰Mäntsälä、Lappeenranta、法国Lille等，使得公司具备全球跨区域的布局能力，具备更灵活的电力调配和主权云适配能力。同时，公司年底已签约电力目标上调到5GW，2027年起每年新建1GW以上。

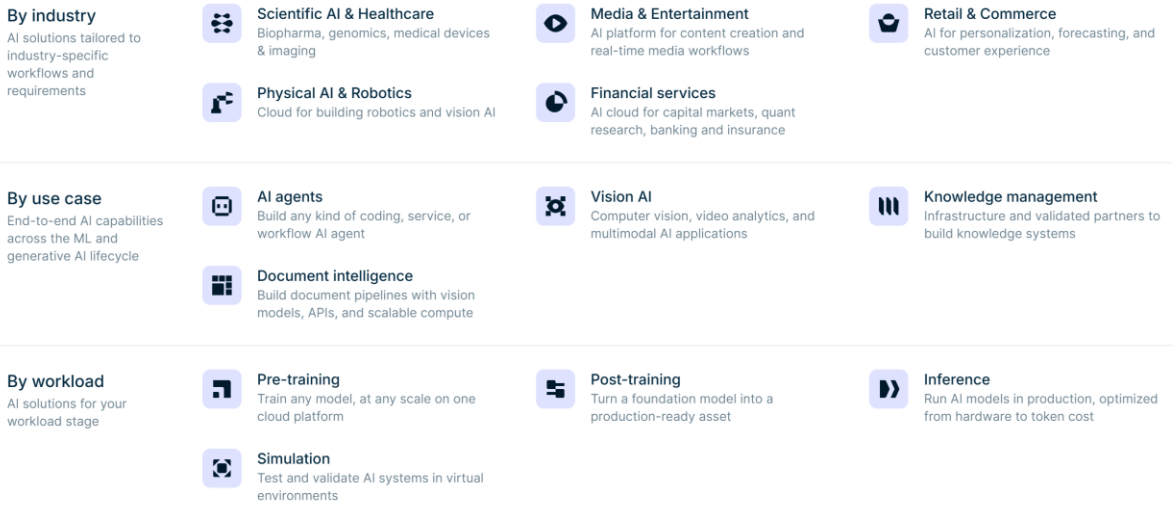
公司锚定微软及Meta两大客户，已签总值超460亿美元的订单。微软：公司于25年9月签下5年期，基础价值174亿美元(若加附加选项达194亿美元)的订单。Meta：公司于26年3月签下5年期，120亿美元专属容量+150亿美元额外容量期权的订单。

公司合同主要分为三类，若有需求能将2027年的全部产能按现价卖光。根据公司描述，**第一类订单**为核心AI云的中期合同，期限1到3年，客户大多为高速扩张的AI公司，定价年化每兆瓦2000-2500万美元，预付款覆盖对应资本开支的50%-60%；**第二类**是短期高溢价合同，通常不超过6个月，客户大多需求很急、要求很高、愿意付溢价，报价区间年化每兆瓦4000-5000万美元甚至更高；**第三类**是与投资级客户的长期合同，大多被作为融资抵押物。公司表示，目前如果想的话，能把2027年的全部产能按现价卖光，但公司选择不卖，暗示公司对未来供需缺口持续扩大、价格持续上行的信心。

图：公司全栈云产品



图：公司拥有多行业，多应用方向的解决方案



资料来源：Nebius，方正证券研究所

更多免费行业报告 并购家 www.ipoipo.net

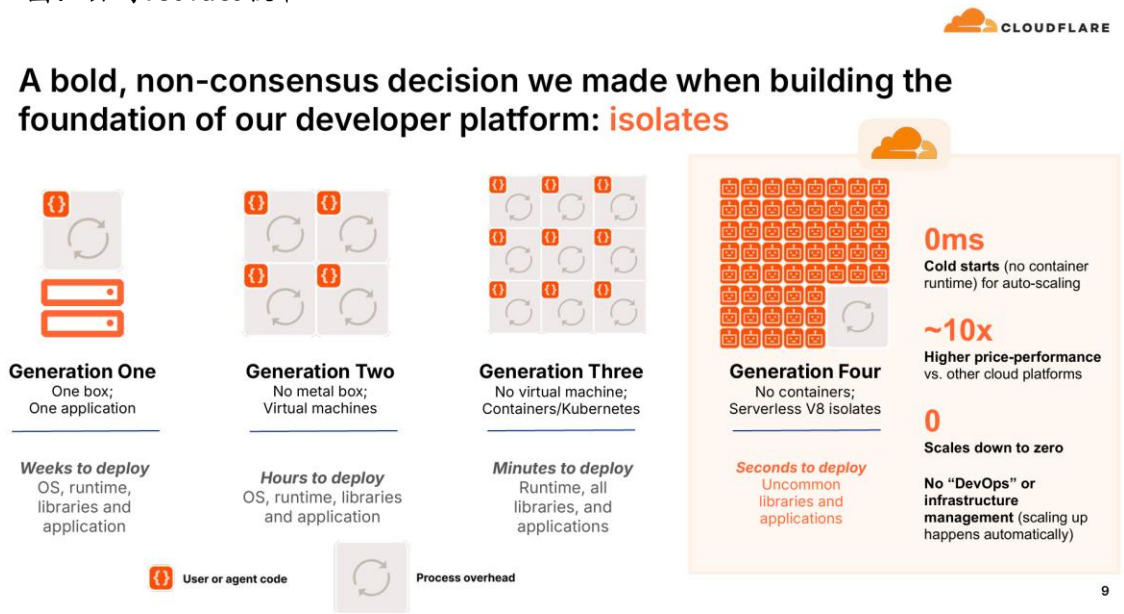
2.7 Cloudflare:专注边缘推理与Agent工作流

产品层面，公司已在全球部署大量边缘计算节点，通过以Workers AI为代表的软件栈将边缘算力封装成可调用的平台能力。软件层面，公司的核心产品为Workers AI，它允许开发者在Cloudflare的全球边缘节点上直接调用托管模型进行推理，而无需自己管理 GPU 集群。**公司也配套AI计算开发平台、AI安全、Agent经济体系三大类产品，**覆盖模型运维管理、Agent开发、全链路安全防护、身份权限管理及机器间支付等环节。**硬件层面，公司在全球拥有335+个PoP，其中190+个节点已部署GPU，已满足边缘推理需求。**

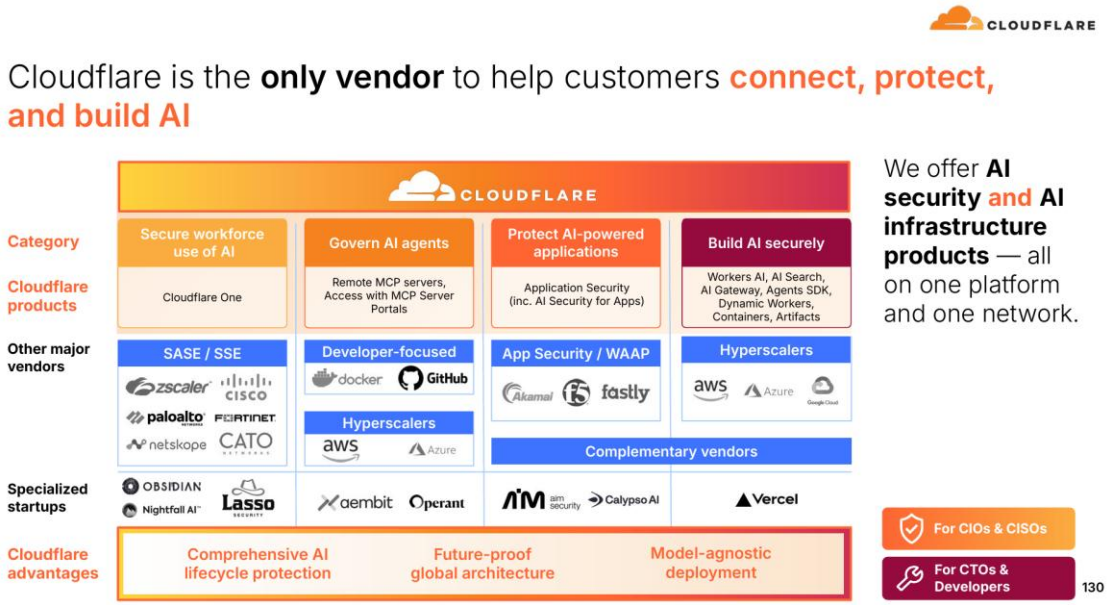
根据大摩的调研报告，在AI推理平台中，Cloudflare排名第三，仅次于AWS及Azure，在边缘计算/边缘推理中市占率第一。并且根据调研结果，52%的调研人员表示，Cloudflare Developer Platform 的10%—30%工作负载来自 AWS、Azure、GCP 的迁移；另有14%认为迁移比例达到30%—50%，代表公司已从Hyperscaler中获取部分推理需求份额，后续增量推理可能流向Cloudflare等边缘平台。

公司自研V8 isolate技术以适配Agent工作流，成本优势显著。公司表示，AI时代代码数量大幅增长， Cloudflare平台上的开发者在26Q1增加了100万个，等同于2025年全年增长量，且大部分开发者为AI Agent，其工作流包括大量CPU执行，通常短暂、突发、持续，难以提前知道需要多少服务器，公司Isolate技术可以快速创建大量轻量级隔离空间，特别适合大量短生命周期任务，在自建模型中，部署1万个Agent生成应用，Cloudflare成本比AWS Lambda低63.6%；运行100万Agent sessions，成本比AWS AgentCore Runtime低74.7%，对比普通云厂商优势显著。

图：公司Isolate技术



图：公司产品分类，优势及定位



方正证券
FOUNDER SECURITIES
正 在 你 身 边

三、部分国内算力租赁公司介绍

更多免费行业报告 并购家 www.ipoipo.net

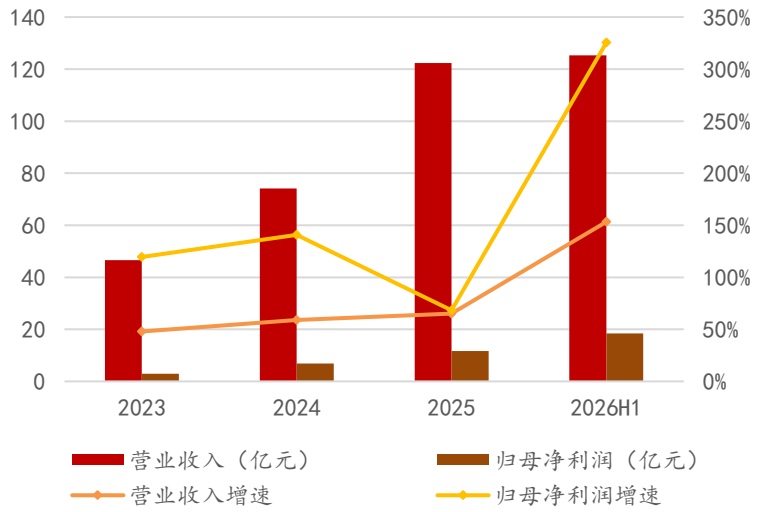
公司定位为“智能终端+算力基建+云端服务”三位一体的全球化AI算力基础设施厂商。公司致力于成为全球领先的智能物联终端制造商、综合性云服务提供商、智能存储设备制造商、服务器再制造行业的先行者和领导者，正沿着算力基础设施+核心部件+平台能力+行业应用的路径，将既有智能制造能力升级为全栈AI能力。

算力服务为公司核心增长引擎，英伟达NCP认证+多地算力中心运营构筑供给能力。子公司已取得英伟达NCP高级别认证，具备异构算力纳管与多用户租赁、算力资源实时调度、主流AI模型适配及高速网络运维调优能力。公司在北京、上海、乌兰察布、廊坊、中卫、广州、深圳、宁波等多地部署并运营算力中心，并在东南亚、美国等海外市场建立分布式算力网络，2025年以来公司公告多笔服务器采购合同、累计金额超300亿元。目前公司已形成万卡级集群的规划建设、网络组网和运营管理能力，客户覆盖头部互联网厂商模型训练推理、自动驾驶、生物医药等领域。

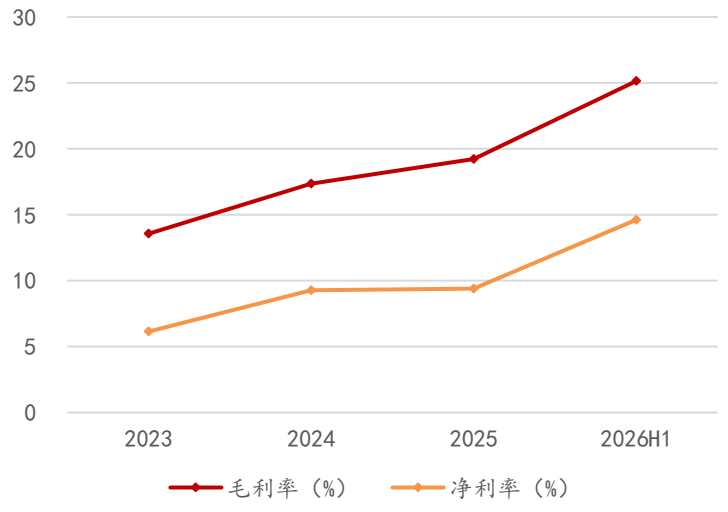
存储、平台、解决方案、服务器再制造协同，公司已构建资源—平台—应用算力业务结构。资源层：公司通过GPU算力集群、数据存储设备、高速网络与光通信、AIoT智能终端等能力，为平台和应用提供算力、存储、连接及数据入口。服务器再制造业务可以较低成本获取硬件资源、延长设备服役周期。**平台层：**公司自研的一体化算力云平台FCloud统一纳管异构算力资源，并且提供训练推理、模型调用、资源调度和AI工具链服务，将分散的硬件资源转化为可配置、可计量、 可持续优化的标准化能力。**应用层：**公司以FCloud为基础，围绕通用AI开发、具身智能、智慧零售、生物医药及模型服务等产业场景，构建OmniBot(具身智能)、Rimo(智慧零售)、BioViMo(生物医药)、**TokenShare(TaaS)**等平台。

公司26H1业绩大幅增长，盈利逐季加速、利润率显著抬升。公司26H1实现营业收入125.23亿元，同比+153.29%；归母净利润18.38亿元，同比+325.51%；经营活动现金流净额18.25亿元，同比转正。26Q2单季实现营收64.38亿元，同比+124.6%，归母净利润10.88亿元，同比+314.0%。26H1毛利率25.15%、净利率14.62%，较2025年全年分别提升5.91pct、5.21pct。公司Q2固定资产220亿，较25年底增长108亿，服务器采购订单逐渐到位。

图：公司2023-2026H1营收、归母净利润及同比增速



图：公司2023-2026H1毛利率及净利率



图：公司TokenShare TaaS平台



资料来源： ，公司公告，TokenShare，公司官网，方正证券研究所

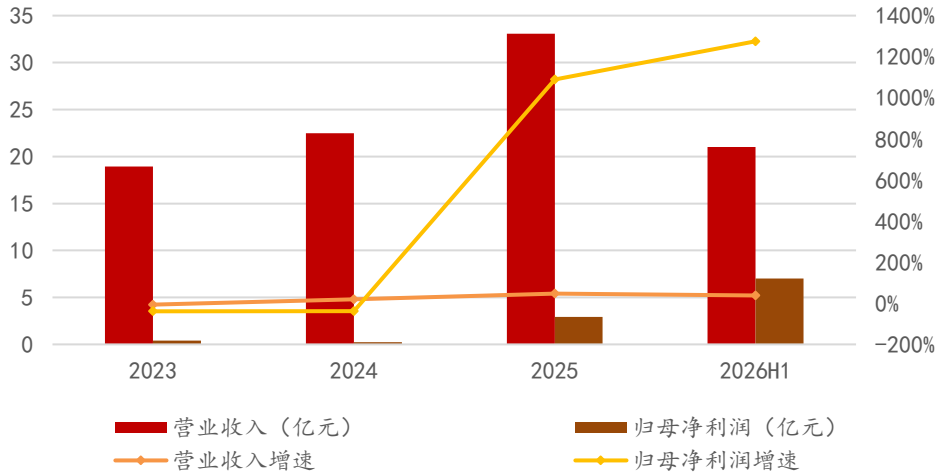
公司起家于液晶电视精密金属结构件，2023年切入AI算力云服务，确立“精密金属结构件+AI算力云服务”双主业格局。公司传统主业为液晶电视等显示领域的精密金属冲压结构件、电子元器件，客户涵盖TCL、海信、三星、小米、索尼及鸿海、京东方等品牌与代工厂。2023年6月，公司以4.5亿元增资取得世纪利通90%股权、投入5亿元运营AI算力项目，正式切入算力租赁赛道，同年11月拟通过新加坡子公司投资5000万美元租赁算力服务器，拓展海外算力租赁业务。

公司算力云规模已达38000P，集群利用率接近100%且长约订单排至2030年以后，将于26H2分批交付。公司以世纪利通为核心平台，围绕高端GPU服务器获取、算力集群建设、云算力租赁、设备经销、运维维保和技术服务构建综合能力。截至2026H1，公司AI算力云服务规模达38000P，其中直租算力16000P、转租算力22000P；已签约在手订单预计于2026年下半年分批交付，在执行算力云服务合同期限均在36个月以上，金额可完全覆盖设备投资及对应回报，算力长期租赁服务已排至2030年以后，现有算力利用率接近100%。公司已与NVIDIA、超聚变等GPU服务器制造商建立战略合作，客户覆盖国内头部互联网企业、AI大模型企业及数字化转型行业客户。

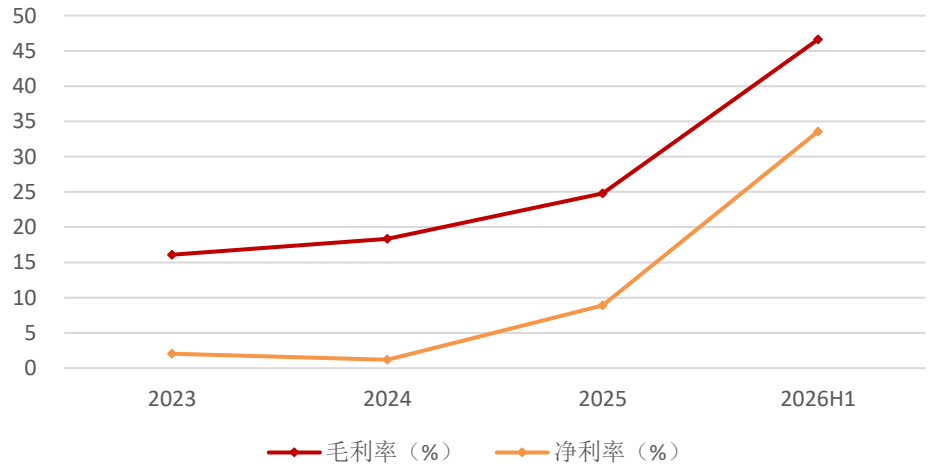
拟定增募资不超50亿元加码中卫智算中心，巩固算力租赁头部梯队地位。26年8月，公司发布定增预案，拟募资不超50亿元，其中40亿元投向智算中心建设项目(总投资59.06亿元，建设期6个月)、10亿元补充流动资金。项目落地“东数西算”枢纽节点，宁夏中卫市具备低电价、高绿电占比与低网络时延优势，公司拟购置高性能GPU服务器及配套网络设备，形成训推一体的算力租赁服务能力。公司表示现有集群保持高出租率，新增产能将与头部互联网客户长期租赁协议和既有长约相衔接。

26H1业绩高度兑现，盈利能力大幅跃升。公司26H1实现营收21.01亿元，同比+38.58%；归母净利润7.02亿元，同比+1275.15%；经营活动现金流净额25.73亿元，同比+733.23%。其中26Q2单季归母净利润4.31亿元，同比+1893.5%，环比增长58.93%，盈利逐季加速。分业务看，AI算力业务实现收入12.74亿元，同比+162.38%，净利润6.34亿元，同比+488.81%：其中算力设备及配件销售业务收入同比+2455.25%，算力租赁收入同比+44.92%（租赁订单多集中于下半年交付）。26H1公司毛利率46.60%、净利率33.56%，同比大幅提升，进入业绩兑现期。

图：公司2023-2026H1营收、归母净利润及同比增速



图：公司2023-2026H1毛利率及净利率



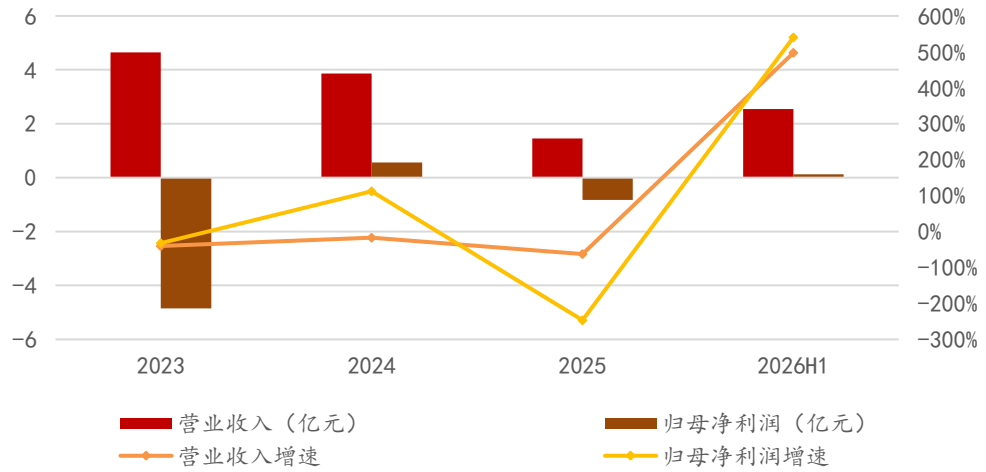
更多免费行业报告 并购家 www.ipoipo.net

公司确立AI全球跨境数字贸易+AI算力基础设施双主业核心战略，形成跨境数字贸易、AI算力基础设施、数智软件运维、液冷算力配套服务四大协同业务板块。其中，跨境数字贸易是公司稳健经营基本盘；AI算力服务为公司核心增长主业，聚焦算力租赁、弹性推理算力服务。公司创新落地高成长TaaS商业化模式，未来盈利质量有望进一步提升。

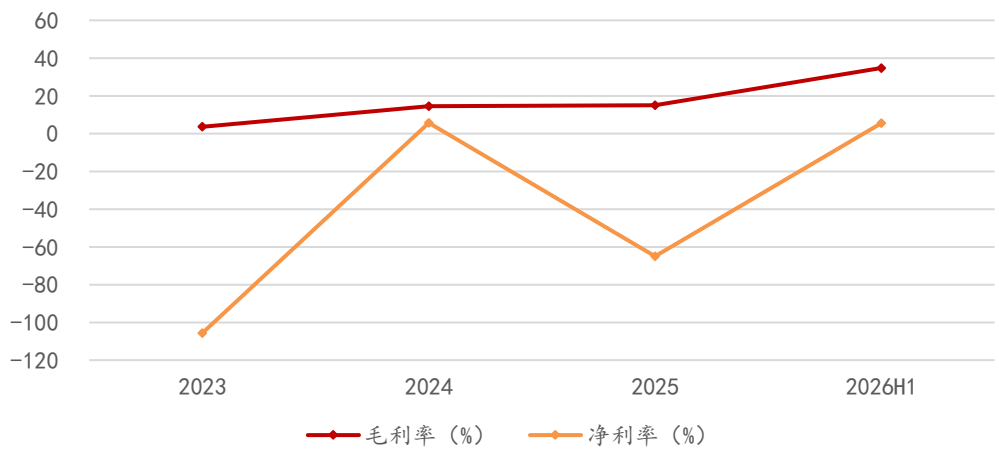
公司卡位异构GPU集群整合运营，提供算力租赁+TaaS弹性计费。公司规模化落地AI算力运营服务板块，并且整合多型号异构GPU算力集群，完成算力资源统一部署、封装、调度与运维，提供算力贸易(服务器整机销售)、算力租赁业务、推理算力资源供给与全栈基础设施解决方案。**公司在手大量长单，**截至26年7月末，公司在手算力长期框架订单规模达150.82亿元，全部为5年长周期协议，客户覆盖头部央企上市公司、一线大模型厂商等优质主体。**公司具备较强的融资能力，**截至26年7月，公司及子公司已向金融及非金融机构申请授信180.24亿元，其中已经批复的授信为102.64亿元，已签署融资及授信合同为52.72亿元，在手订单及产能扩张能力丰富，已形成订单—融资—产能扩张的良性发展循环。**公司自研AlayaJet垂直推理引擎，**可针对性优化大模型推理显存占用与数据传输开销，有效提升单卡Token产出效率、降低整体推理运营成本，提升算力产能和整体盈利水平，**推动公司算力业务从硬件租赁向高效TaaS模式升级。**

公司上半年经营业绩实现跨越式反转，经营质量得到根本性改善。公司26H1实现营业收入2.54亿元，同比+497.11%；归母净利润1201.54万元，同比+540.15%；扣非归母净利润671.74万元，同比+174.64%；经营活动现金流净额1.94亿元，同比转正，公司主业自我造血能力与经营稳健性大幅提升。公司算力服务器业务规模化放量，上半年销售67台服务器确认收入，增加当期毛利润约6500万元，且成为利润增厚的重要来源。**同时，公司已验收供应商交付的算力服务器470台(约50000P)，另有413台正在交付客户压测验收中，这些设备后续将以租赁方式确认收入，结合公司在手150亿长期订单，后续算力租赁业绩放量可期。**

图：公司2023-2026H1营收、归母净利润及同比增速



图：公司2023-2026H1毛利率及净利率



资料来源：，公司公告，方正证券研究所

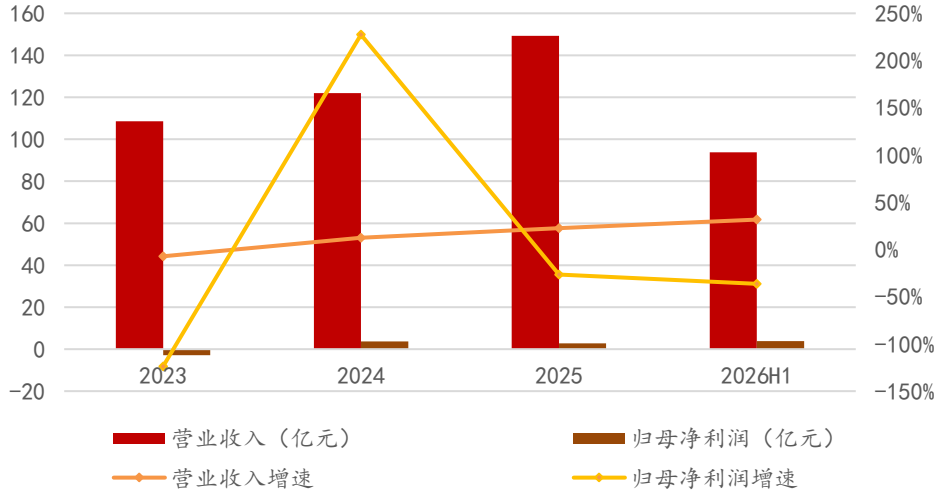
公司传统业务为氟氯化工、电子元器件和高端铝箔制造平台，依托电子元器件及材料、氟氯化工、全链路液冷解决方案等先进制造为基础，以 AIDC+算力运营服务为核心，致力于成为"材料筑基、算力驱动"的一体化科技生态的龙头企业。

公司收购国内头部第三方超大规模算力基建运营商秦淮数据，并设立子公司承接算租服务业务。25年9月，东阳光联合控股股东及多家机构组建收购平台，拟以280亿元整体对价收购秦淮数据中国100%股权，26年1月完成首期交割，26年6月拟发行股份收购剩余70%股权并配套募资不超80亿元投向秦淮数据三座零碳智算基地，完成后秦淮数据将全资并表。秦淮数据投运及在建IT容量近1GW、储备及规划4GW，投产项目上架率超90%，是头部互联网大厂的核心算力供应商，将构成公司算力运营的底仓。26年4月，公司设立控股子公司东阳光云智算，为公司算力服务业务的直接签约与执行主体，采购部署高性能服务器对外出租。东阳光云智算5-7月已签署三笔五年期算力服务合同，总金额达390-460亿元，均提供从采购服务器到上架运维的全周期服务。

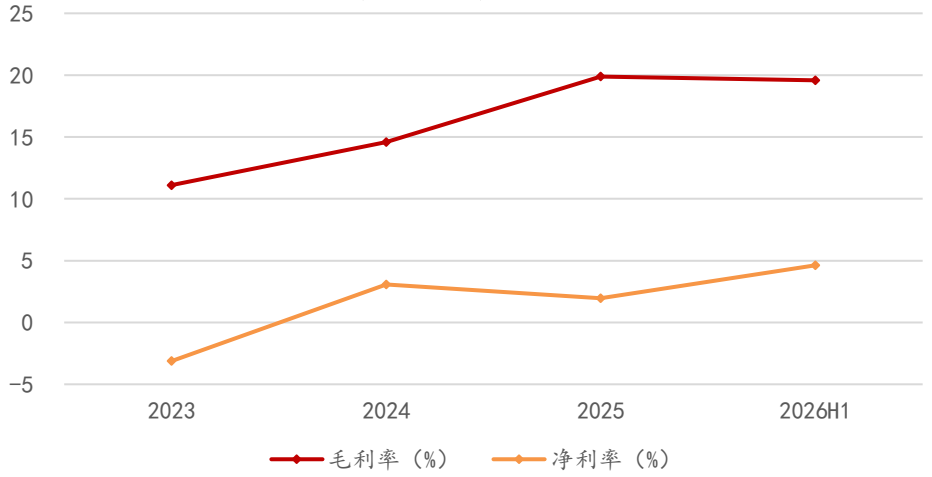
除算力服务外，公司具备液冷、被动元器件等AI算力硬件服务能力，与算租业务实现软硬协同。液冷：公司拥有服务器级—机柜级—系统级全栈式液冷体系，产品覆盖各类冷板、CDU、Manifold等核心组件，可实现专用冷却液、终端散热硬件至整机液冷系统解决方案的自主研发与量产交付。**被动元器件：**公司研发积层化成箔与积层箔电容器、超级电容器等核心产品，全面适配AI服务器与智算中心整套供电链路。**自此，公司已形成AI硬件—算力基建与运维—算力服务全链条贯通，算力服务场景持续为液冷业务及秦淮数据AIDC业务输送增量需求，成熟液冷配套体系和AIDC能力也将赋能算力服务提质扩容，形成双向联动。**

公司26H1算力业务起量，液冷、积层箔、超级电容器协同扩产。公司26H1实现营业收入93.76亿元，同比+31.61%；归母净利润3.89亿元，同比-36.46%，主要系确认了股权激励费用以及股权公允价值变动，剔除该影响后的归母净利润为7.46亿元，同比+40.20%。经营活动现金流净额25.12亿元，同比+843.66%。26H1公司AI业务实现营收2.89亿元，毛利率45.39%。公司积层箔、**超级电容器、液冷等硬件业务持续推进。**

图：公司2023-2026H1营收、归母净利润及同比增速



图：公司2023-2026H1毛利率及净利率



更多免费行业报告 并购家 www.ipoipo.net

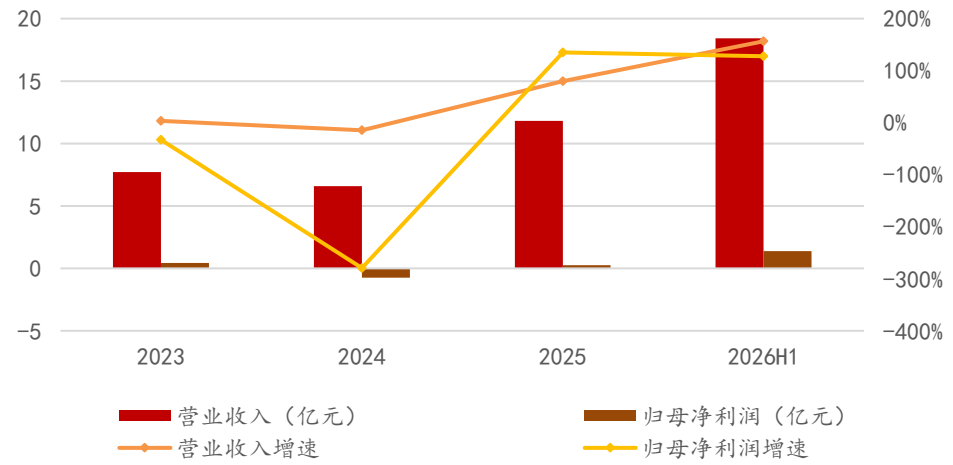
公司持续探索人工智能、大数据、智算等技术的行业应用和成果转化，打造新质生产力，致力于为更广泛的客户实现数智化转型。在算力服务业务方面，公司通过向AI客户提供全链条、多元化、高效益的智能算力服务，全方位赋能客户在大模型训练、推理、科学计算等不同业务场景的各类需求。在智慧城市领域，为遍布各行业的客户提供专业和系统的全流程解决方案。

公司定位于高性能算力服务市场，具备IaaS全栈交付与运营能力。公司算力服务业务可分为算力设备集成服务、算力设备运营服务及算力平台开发应用服务三类。公司明确将自身定位于高性能智算服务市场，客户覆盖大型互联网公司、大模型公司与云厂商等对技术性能、交付时效和订单规模要求更高的客户。公司依托异构计算、高速互联网络、分布式存储等技术积淀，能够提供从基础设施架构设计、硬件集群部署、网络拓扑优化、并行文件系统调优，到虚拟化/容器化平台搭建的完整服务。公司拥有多年数据中心渠道资源与可靠供应链，可及时采购并交付高性能AI芯片及服务器，持续获取交付时效要求高、订单规模大、应收账款回款及时的算力服务业务；截至26H1末，公司已陆续签订并交付多项重大算力业务合同，业务交付规模持续提升。

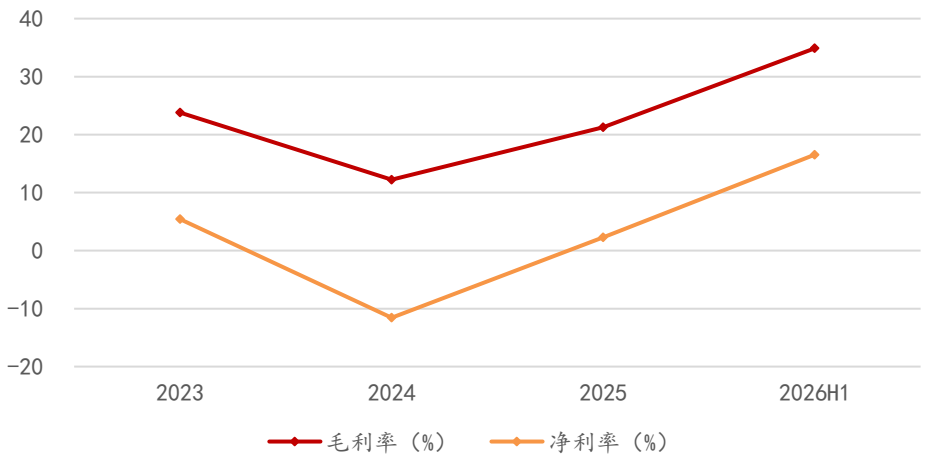
公司算力服务已成主业，Token工厂+算力出海+设备循环打开模式升级空间。公司26H1算力服务实现收入16.89亿元，同比+55.78%，营收占比达91.64%，转型已基本完成。公司打造高效能Token工厂，以标准化、可计量、高性价比的Token产能为核心产品，提升每瓦电力Token产出效率与成本竞争力，并加快全球算力节点布局与跨境通道建设、组建本地化服务团队，推动Token与算力服务规模化出海。26年8月，公司宏景纵横AI创新服务基地开业并启动全国首个算力设备循环产业示范园。

26H1业绩高增且逐季加速，算力资产规模跃升。26H1公司实现营收18.43亿元，同比+156.43%；归母净利润1.37亿元，同比+127.43%；经营活动现金流净额17.69亿元，同比+3105.55%。其中26Q2单季营收14.66亿元，同比+299.97%，环比+289.09%；归母净利润1.07亿元，同比+176.57%，环比+253.09%，交付放量驱动盈利逐季加速。26H1公司毛利率34.88%，其中算力服务毛利率37.98%。截至26H1末，公司固定资产达60.73亿元，较25年底+327.5%、在建工程17.26亿元、长期借款30.38亿元，"借款—采购—交付—运营"的扩表链条清晰，为后续收入放量奠定算力资产基础。

图：公司2023-2026H1营收、归母净利润及同比增速



图：公司2023-2026H1毛利率及净利率



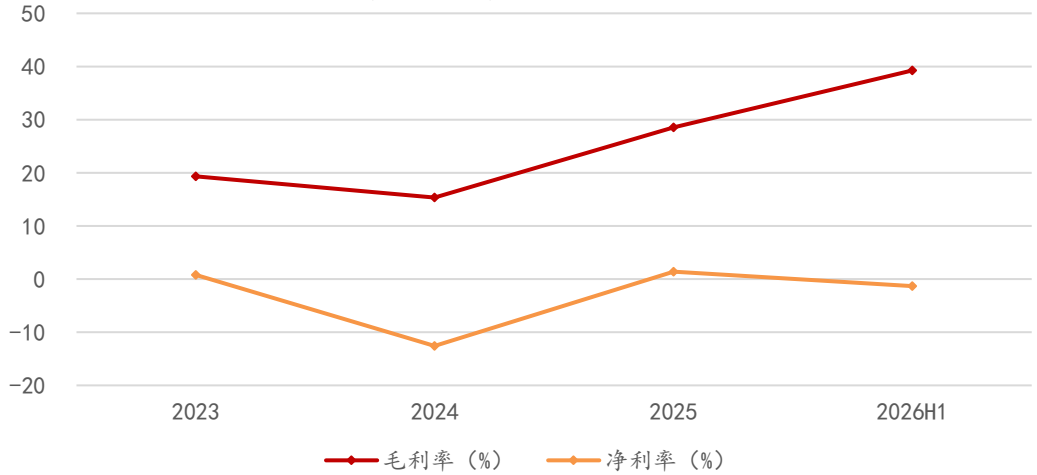
公司专注于人工智能产品的研发与制造，构建了覆盖AI基础设施与“AI+”行业应用的全栈技术体系。旗下常青云品牌聚焦AI算力底座，推出AI大模型一体机、超融合一体机、智算平台等核心产品，助力各行业智能化升级；掌安星品牌深耕智慧安全领域，创新推出AI安全机器人、智能电磁安保设备等产品，广泛应用于公安、政企、工业、教育、医疗等领域。

公司已形成“智算云服务+常青云专有云/混合云”业务模式。智算云：基于自有技术及算力设备，建设和运营智算云集群，面向CSP云厂商、大模型企业、运营商及垂类AI应用客户，根据业务场景、资源用量及服务周期，提供按需、弹性的算力服务，并配套资源调度、系统运维、技术支持等服务。26H1公司共审议不超过140亿元的IT设备及零配件采购计划，目前已在华南、华东、华北等地建成共计约9000P（FP16）算力的智算云集群(设备资产超过25亿元)，均已进入运营期，形成智算云服务收入1.72亿元。在建集群算力超40000P(FP16)，规模有望进一步加大。**常青云：**深耕垂类行业场景客户，全栈云产品在多家用户落地使用，超融合软硬件+AI智能体的“本地化部署+AI行业应用”组合方案成为客户接受度高的新模式，覆盖政务云合作、AI智算平台、国产算力服务器及专有云基础设施等场景。常青云已与昇腾、天数智芯、沐曦、昆仑芯等国产+海外GPU芯片完成适配，进一步加强异构算力的调度、管理、计费能力开发。目前已签订国产算力订单超4.6亿元，取得业务收入约4000万元，超过去年全年业务量。

公司已开始探索Token工厂模式，下半年有望落地。公司拟依托智算云集群、算力调度平台、云平台运营能力及行业客户资源，与模型厂商、算力资源方、云服务商和应用合作伙伴就Token计费、算力服务及联合运营等商业模式开展探索。下半年，公司计划试点打造多个垂类场景的Token工厂，为AI影视制作、商业广告、代码编写、日常办公等场景客户提供高性价比的Token生产能力，涵盖算力供给、模型部署、推理优化和应用接口的一体化服务，逐步构建集算力资源整合、模型部署优化、推理服务运营及Token计量交付于一体的Token工厂业务模式。8月28日，公司“常青云Token工厂”与核心产品“常青云AI一体机”、“常青云KuberCloud算力管理平台”已在数博会展出。

公司Q2单季收入与利润出现拐点，业绩已开始释放。公司26H1实现营业收入3.43亿元，同比+8.72%；归母净利润-0.03亿元，同比转亏，主要系投建智算集群后，利息和折旧费用大幅增加。公司EBIT/EBITDA约0.44/1.41亿元，同比+130%/+340%，实际盈利能力大幅增长。经营活动现金流净额约2.98亿元，同比大幅改善。26Q2实现营业收入1.93亿元，同比+34.83%；归母净利润0.11亿元，同比+115.23%，随着智算业务进入运营期，业绩逐渐开始释放。

图：公司2023-2026H1毛利率及净利率



图：公司常青云Token工厂



更多免费行业报告 并购家 www.ipoipo.net

4 风险提示

1. **出口管制政策变化风险**:海外先进AI芯片出口审查政策变化可能影响国内高端算力供给与租金价格体系。

2. **GPU价格下行与资产减值风险**:芯片迭代加速可能导致存量GPU租金下行及资产减值。

3. **算力租赁价格下降超预期风险**:行业竞争或技术进步可能使租赁单价下降快于预期,影响市场规模与盈利水平。

4. **行业竞争加剧风险**:参与者快速增加,新增供给集中释放可能引发价格竞争。

5. **技术迭代风险**:模型效率提升可能降低单位应用算力需求;TaaS等新模式落地存在不确定性。

6. **测算偏差风险**:市场规模等测算基于多项假设,实际结果可能与测算存在较大偏差。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，保证报告所采用的数据和信息均来自公开合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法存在局限性。特此声明。

免责声明

本研究报告由方正证券制作及在中国（香港和澳门特别行政区、台湾省除外）发布。根据《证券期货投资者适当性管理办法》，本报告内容仅供我公司适当性评级为C3及以上等级的投资者使用，本公司不会因接收人收到本报告而视其为本公司的当然客户。若您并非前述等级的投资者，为保证服务质量、控制风险，请勿订阅本报告中的信息，本资料难以设置访问权限，若给您造成不便，敬请谅解。

在任何情况下，本报告的内容不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求，方正证券不对任何人因使用本报告所载任何内容所引致的任何损失负任何责任，投资者需自行承担风险。

本报告版权仅为方正证券所有，本公司对本报告保留一切法律权利。未经本公司事先书面授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容，不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。如需引用、刊发或转载本报告，需注明出处且不得进行任何有悖原意的引用、删节和修改。

更多免费行业报告 并购家 www.ipoipo.net

类别	评级	说明
公司评级	强烈推荐	分析师预测未来12个月内相对同期基准指数有20%以上的涨幅。
	推荐	分析师预测未来12个月内相对同期基准指数有10%以上的涨幅
	中性	分析师预测未来12个月内相对同期基准指数在-10%和10%之间波动。
	减持	分析师预测未来12个月内相对同期基准指数有10%以上的跌幅。
行业评级	推荐	分析师预测未来12个月内行业表现强于同期基准指数。
	中性	分析师预测未来12个月内行业表现与同期基准指数持平。
	减持	分析师预测未来12个月内行业表现弱于同期基准指数。
基准指数说明		A股市场以沪深300 指数为基准;香港市场以恒生指数为基准，美股市场以标普500指数为基准。



方正证券研究所

专注·专心·专业

北京市朝阳区朝阳门南大街10号兆泰国际中心A座17层
长沙市天心区湘江中路二段36号华远国际中心37层

上海市静安区延平路71号延平大厦2楼
深圳市福田区竹子林紫竹七道光大银行大厦31层

广州市天河区兴盛路12号楼隽峰院2期3层方正证券

更多免费行业报告 并购家 www.ipoipo.net