

人工智能行业专题（18）：上限几何？ 北美科技巨头Capex空间测算及ROI回报拆解

行业研究 · 海外市场专题

互联网 · 互联网 II

投资评级：于大市（维持）

证券分析师：张伦可

0755-81982651

zhanglunke@guosen.com.cn

S0980521120004

证券分析师：刘子谭

liuzitan@guosen.com.cn

S0980525060001

- 从宏观角度，北美科技整体资本开支今年按照万亿美金计算，占比美国GDP约3%。相比“科技资本开支占GDP”在2000年的历史高点约3.27%空间不大。但此轮AI投资已进入“企业资本开支+主权投资+AI Labs专属集群”并行阶段。从英伟达客户收入结构来看，2026Q1北美Hyperscale贡献50%数据中心收入、主权AI占比15%、Neocloud和AI初创厂商15%、行业私有云等其他占比20%。
- ① **主权AI**：预计2026年投入超千亿美元，在手合同超万亿。若类比公用事业投资，资本开支投入GDP的1%，中长期可达1.2万亿以上。
- ② **Hyperscale**：2026年微软、谷歌、亚马逊、Meta、Oracle资本开支指引合计超7500亿美金。资本开支是当年经营性现金流的约1.2倍，因此需要债务融资，今年预期新增债务规模达3000亿。杠杆层面仍有较大空间，微软、谷歌、亚马逊、Meta 26H1净债务/EBITDA仅在0附近，对比北美IG AA评级整体中位净债务/EBITDA可达到1.2倍。若假设经营性现金流年化增速20%，净杠杆在2030年达到1.2倍，每年的资本开支可达经营性现金流的1.35x左右。未来四年资本开支增速CAGR约在25%，四大巨头2030年总资本开支有望超过1.6万亿。
- 从ROI视角，AI对内赋能占比低个位数。如M365 Copilot付费率与微软云收入占比均约4.5%。META AI约拉动广告10%收入增长。
- 算力租赁与云业务ROI可观**。SpaceX的Colossus集群约两年即可收回建设成本，而META同时出租空闲算力和新建算力也为商业化行为，其A/H卡全部租赁情况下，测算年度租金可达110亿美金，可拉动META今年EPS约15%增长。路透社报道Meta 2026年计划部署7GW算力、2027年翻倍，包括全资自建（5GW+）、表外合资（1GW）、第三方租赁（约1GW），硬件混用英伟达、AMD GPU与自研MTIA。
- 云厂商AI收入快速增长，测算26Q1 AI云占比云收入约20%-30%。**短期由于资本前置现金流承压，测算2029-2030年规模效应释放后现金流转正。测算2030年谷歌云业务ROI 38%、亚马逊24%、微软17%，差异主要来源于自研大模型和自研芯片的利润增厚，亚马逊Trainium、谷歌TPU单卡TCO低于英伟达芯片，后续可有效提升云毛利率。
- 大模型厂商是算力核心需求方。根据yipidata数据，26年6月Anthropic ARR约690亿、OpenAI约420亿，全年收入有望达1200亿美元（全年收入增速分别约15倍和3倍）。且两者毛利率已处在60%左右，26Q2 Anthropic已实现经营性现金流转正、OPM约为5%。
- 测算Anthropic、OpenAI 2026年总算力支出超千亿美元（约占比北美CSP今年收入的20%），**其中Anthropic 490亿美金（240亿推理、250亿训练），OpenAI 550亿美金（230亿推理、320亿训练），成为支撑科技厂商资本开支的核心动力。全年两者占用12GW算力，预期年末升至17GW。且每MW算力年度收入今年分别有68%、34%的同比增长，说明其单位算力下模型价值持续提升。长期大模型空间广阔，白领劳动收入为40万亿美金市场（10亿人*年收入4万美元），可替代10万亿（25%工时自动化），若假设10%渗透也对应万亿美金市场。
- 风险提示**：AI商业化落地不及预期、高额发债推升融资成本、芯片交付不及规划、芯片快速迭代引发大额资产减值风险

【 01 】 资本开支水位与可行性区间

【 02 】 科技大厂AI投资的ROI测算

【 03 】 大模型厂商财务模型与ROI

【 04 】 风险提示

- 若以世界银行预期美国2026年名义GDP 32.33万亿美元计算，头部五家（微软、谷歌、亚马逊、Meta、Oracle）资本开支7500亿美金（26Q1公司指引）占比约2.30%，若美国科技整体资本开支按照万亿美金计算，占比约3%。相比2000年美国“科技资本开支占GDP”的历史高点约为3.27%，如果按照“技术设备+软件”口径测得的峰值约为4.46%-4.5%。
- 若只看北美四大CSP/企业资本开支，会低估全球AI资本开支的上限。因为从2026年开始，全球AI投资已进入“企业资本开支+主权投资+AI Labs专属集群”并行阶段。
- Gartner口径显示，2025年全球AI总支出约1.76万亿美元，同比增长67.6%；2026/2027年预计升至2.60万亿/3.49万亿美元，其中2026年AI基础设施投资约1.43万亿美元，占比近55%，人工智能服务约5,855亿美元，数据与模型相关支出增速则更快。

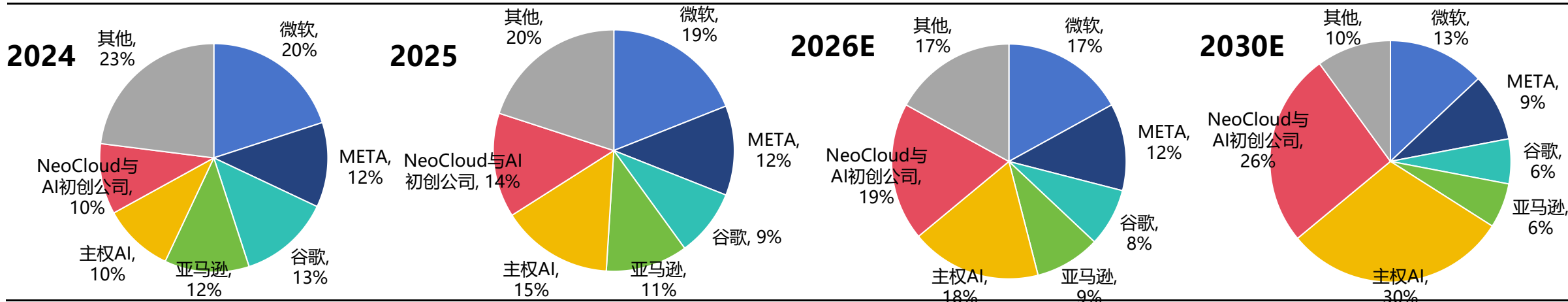
表：2000年投资开支与GDP比值情况			
指标	起点	峰值	含义
科技资本开支/GDP		2000年3.27%	硬件投资强度提升
技术设备与软件投资/GDP	1992年2.9%	2000年4.5%	科技投资强度显著抬升
信息部门投资/GDP	1990年代中期低位	2000年>2.0%	电信/信息行业投资热潮
非住宅投资/GDP	1992年约11%	2000年接近15%	宏观总投资周期上行
Aggregate US capex/GDP	长期均值17.2%	2000Q3 18.5%	宏观Capex周期顶部
企业投资对GDP增长贡献		泡沫前5年年均1.3pct	

资料来源：SEC、彭博、世界银行，国信证券经济研究所整理测算

英伟达客户贡献数据中心收入占比

- 英伟达26FYQ2业绩会表示：2025年全球数据中心基础设施投入预计6000亿美元（两年内翻倍），2030年前有望达到3-4万亿美元。增长驱动来自于推理型AI（agenticAI）需求激增、全球主权AI与企业AI建设、物理AI（机器人/自动驾驶）兴起。
- **Hyperscale（北美四大CSP+Meta）约50%**：26FYQ4（截至2026年1月）英伟达业绩会表示云服务商约贡献了50%的数据中心收入，包括AWS、Azure、GoogleCloud、CoreWeave、Oracle等大型云服务商。其需求主要来自于AI基础设施建设，包括大模型训练、推理等需求。
- **政府与主权AI约15%**：潜力巨大的新市场，2026财年占比数据中心收入的15%。**26财年，主权AI业务营收同比增长超三倍，突破300亿美元。**主要客户来自加拿大、法国、荷兰、新加坡和英国，预计主权人工智能业务的机遇将持续增长。
- **Neocloud与AI初创公司约15%**：新云（如Coreweave）与初创AI公司贡献了数据中心的剩余主要收入。新云当前RPO规模约是年收入的5-10倍，反映了算力建设极高的预期。初创AI公司后续预计成为快速增长点，NVIDIA计划投资OpenAI高达1000亿美元换取OpenAI将部署至少10吉瓦的NVIDIA系统。
- **其他行业公司约20%**：（如行业私有AI工厂）约贡献了数据中心收入的20%-25%，且增速显著。25FYQ4英伟达业绩会表示消费互联网领域营收同比增长3倍，包括推荐系统、视觉语言理解、合成数据生成搜索、多模态内容处理和代理AI等。

图：2026与2030年海外科技厂贡献英伟达数据中心芯片收入占比测算



资料来源：Omdia、公司官网、彭博一致性预期，国信证券经济研究所整理测算

英伟达数据中心收入：主权AI与Neocloud、AI初创公司为主要增量

国信证券
GUOSEN SECURITIES

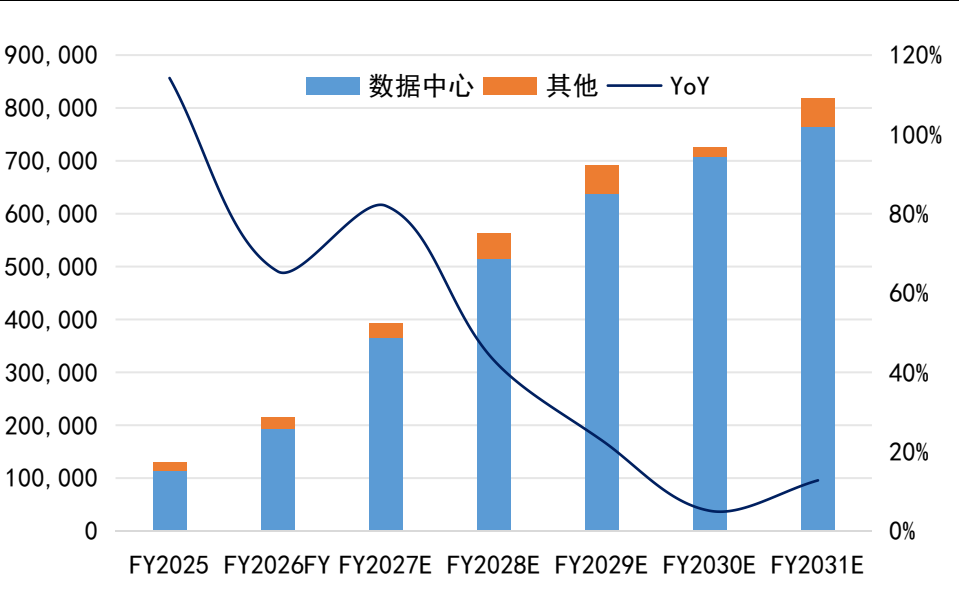
- 英伟达26FYQ2业绩会表示：2025年全球数据中心基础设施投入预计6000亿美元（两年内翻倍），2030年前有望达到3-4万亿美元。增长驱动来自于推理型AI（agenticAI）需求激增、全球主权AI与企业AI建设、物理AI（机器人/自动驾驶）兴起。
- **Hyperscale（北美四大CSP+Meta）约50%**：26FYQ4（截至2026年1月）英伟达业绩会表示云服务商约贡献了50%的数据中心收入，包括AWS、Azure、GoogleCloud、CoreWeave、Oracle等大型云服务商。其需求主要来自于AI基础设施建设，包括大模型训练、推理等需求。
- **政府与主权AI约15%**：潜力巨大的新市场，2026财年占比数据中心收入的15%。**26财年，主权AI业务营收同比增长超三倍，突破300亿美元**。主要客户来自加拿大、法国、荷兰、新加坡和英国，预计主权人工智能业务的机遇将持续增长。
- **Neocloud与AI初创公司约15%**：新云（如Coreweave）与初创AI公司贡献了数据中心的剩余主要收入。新云当前RPO规模约是年收入的5-10倍，反映了算力建设极高的预期。初创AI公司后续预计成为快速增长点，NVIDIA计划投资OpenAI高达1000亿美元换取OpenAI将部署至少10吉瓦的NVIDIA系统。

表：海外科技与相关主体资本开支测算（百万美元）

当年资本开支(百万美元)	2024	2025	2026E	2030E	2026-2030CAGR
微软	75,600	118,000	161,556	430,000	28%
Meta	39,225	72,213	121,763	280,000	23%
谷歌	52,535	91,447	181,801	460,000	26%
亚马逊	77,658	128,320	198,534	468,000	24%
主权AI	25,000	83,000	190,000	950,000	50%
NeoCloud与AI初创公司	25,000	70,000	190,000	700,000	39%

资料来源：Omdia、公司官网、彭博，国信证券经济研究所整理测算

图：英伟达收入变化（百万美元）



资料来源：彭博、公司官网，国信证券经济研究所整理（采用彭博一致性预期）6

主权AI的投资规模：当前千亿美金/年，有望超万亿美金/年



- 根据现有合同我们预计（至2026年）主权AI算力资本开支将突破1000亿美元，中期（未来五年）“合同”的主权AI项目与相关生态投资合计体量有望超过1万亿美元。核心驱动来自于“数据与安全主权”“合规与本地部署”“低时延与产业应用”“地缘与技术管制”“能源成本与可获得性”等五大因素，叠加新型融资模式（PPP、主权基金、芯片租赁+大厂担保、私募信贷）加速落地，使主权AI成为继北美CSP之后全球AI资本开支的“第二增长极”。
- 若类比公用事业投资，年度全欧电力总投资（电源+输配电网）年度电力投资/GDP≈1.3% - 1.5%，中国电力年度投资/GDP长期稳定在1.4% - 1.8%。**假设算力资本开支投入1%（若类比公用事业投资），2026年全球主要经济体GDP约120万亿美元，主权AI投资可达1.2万亿美元/年。**

表：主权国家数据中心建设情况

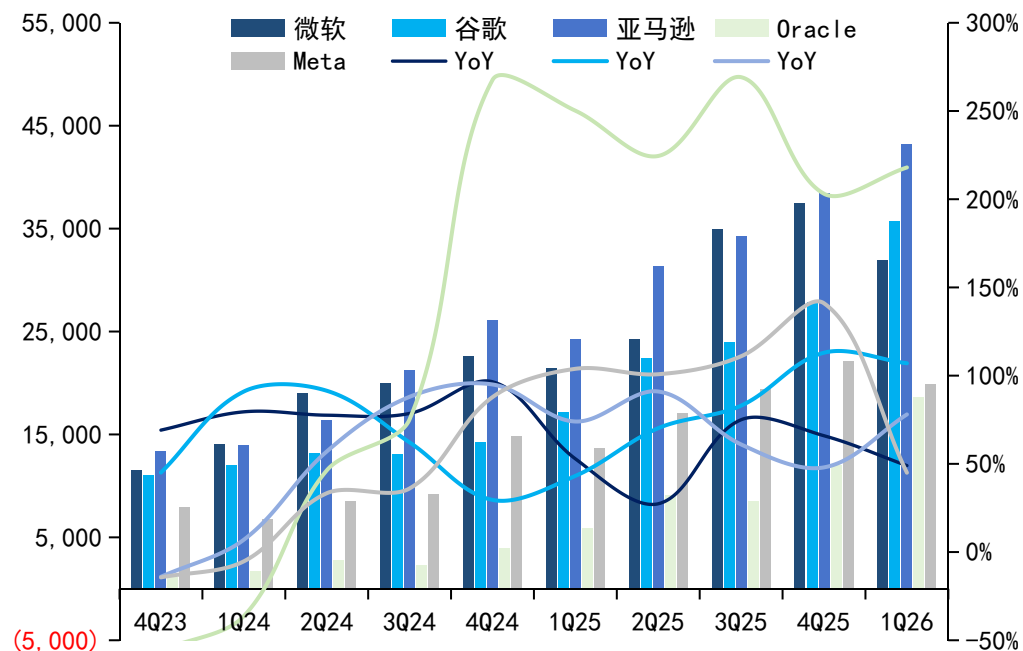
类别	规模/目标	时间维度	依据/备注
美国Stargate	5000亿美元	多年期	国家级AI工程，容量目标GW级。
欧盟Invest AI	2000亿欧元	多年期	“AI超级工厂”+主权云/本地化。
沙特Humain AI	1000亿美元	多年期	主权基金驱动，低成本绿电+园区化布局。
新加坡配额与外资	300MW配额；AWS至2028年S\$120亿；Google累计~50亿美元	2024–2028	政府“绿色增长”政策+区域枢纽定位。
沙特STC×Humain JV	1GW目标（首期250MW）	多年期	从连接/托管到共建AI基础设施的跃迁。
主权AI合计（未来五年）	>1万亿美元	5年	富士康管理层展望，欧洲项目在列。
2026年主权AI投入	>1000亿美元	年度	德勤行业预测：国家数增加、投入破百亿美元大关。
单位造价（GW）	440–600亿美元/GW	现状测算	牛津/英伟达口径，含IT、供电、工程与配套。

资料来源：公司官网、彭博社、华尔街新闻，国信证券经济研究所整理

全球云巨头与META资本开支情况

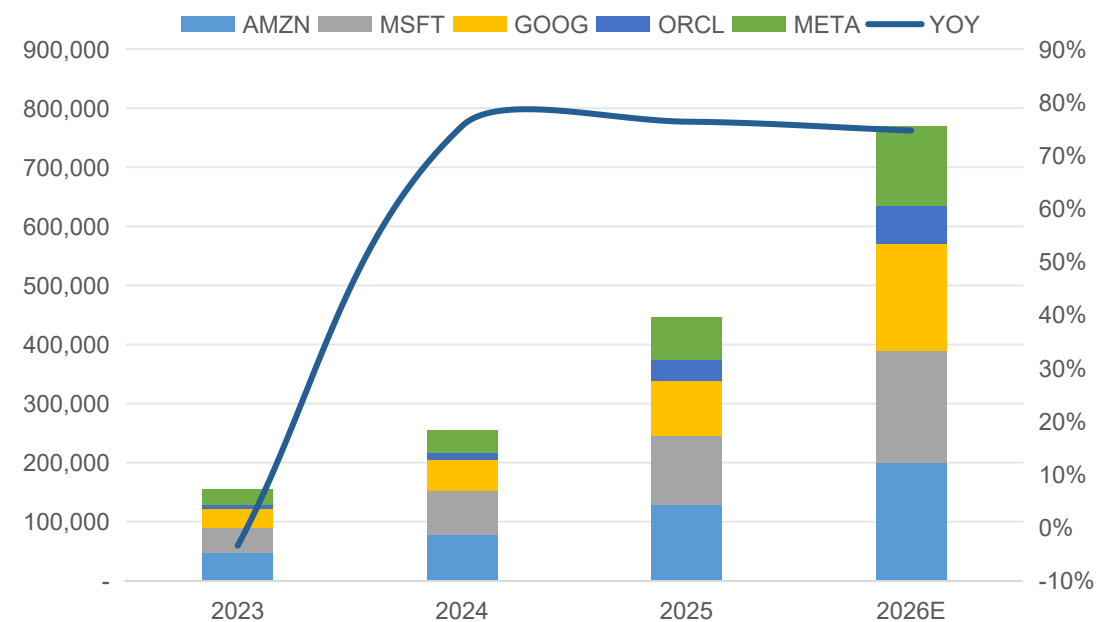
- 2025年海外厂商各家全年Capex约4500亿美金，处于70%以上同比增长：
- 微软：1180亿美元(同比+56%)；亚马逊：1283亿美元(同比+78%)；谷歌：914亿美元(同比+74%)；Meta:722亿美元(同比+94%)；Oracle:355亿美元(同比+300%)；
- 2026年Q1业绩会各家指引资本开支超7500亿，对应四家公司2025年经营性现金流合计的130%：
- 微软：1900亿美元(同比+61%)；亚马逊：2000亿美元(同比+56%)；谷歌：1800亿美元(同比+97%)；Meta:1350亿美元(同比+87%)；Oracle:650亿美元(同比+90%)；

图：各互联网与云厂Capex季度投资与变化情况（百万美元、%）



资料来源：公司财报、国信证券经济研究所整理

图：各互联网与云厂资本开支变化情况（百万美元、%）

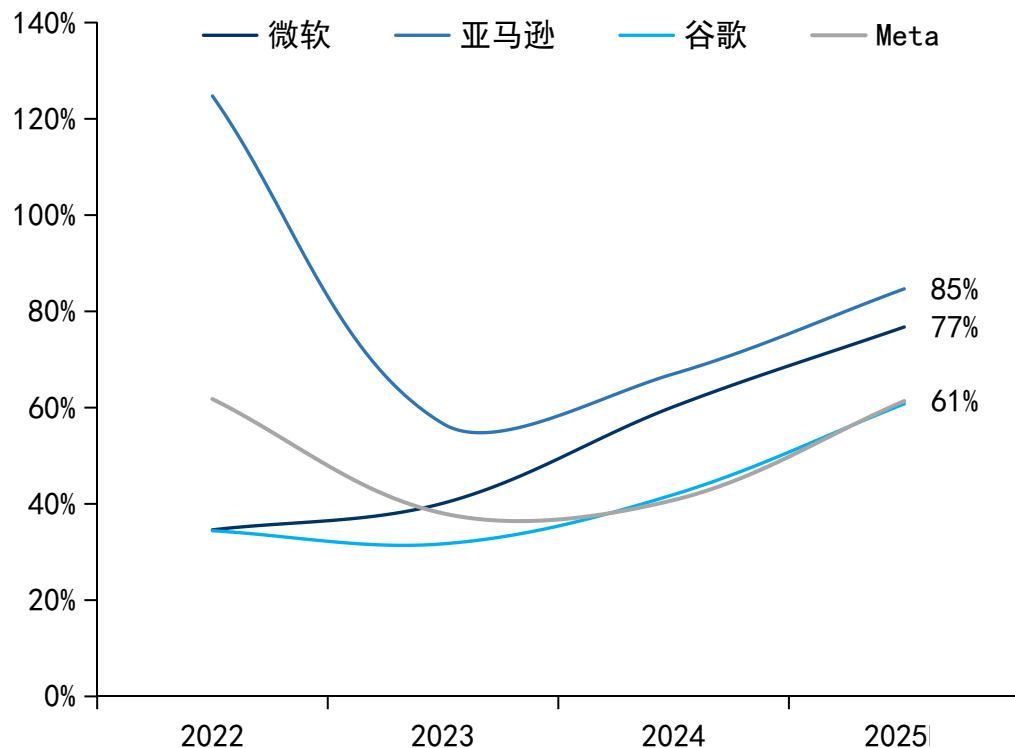


资料来源：公司财报、国信证券经济研究所整理

海内外Capex量级：2026年资本开支约为经营性现金流的1.2倍

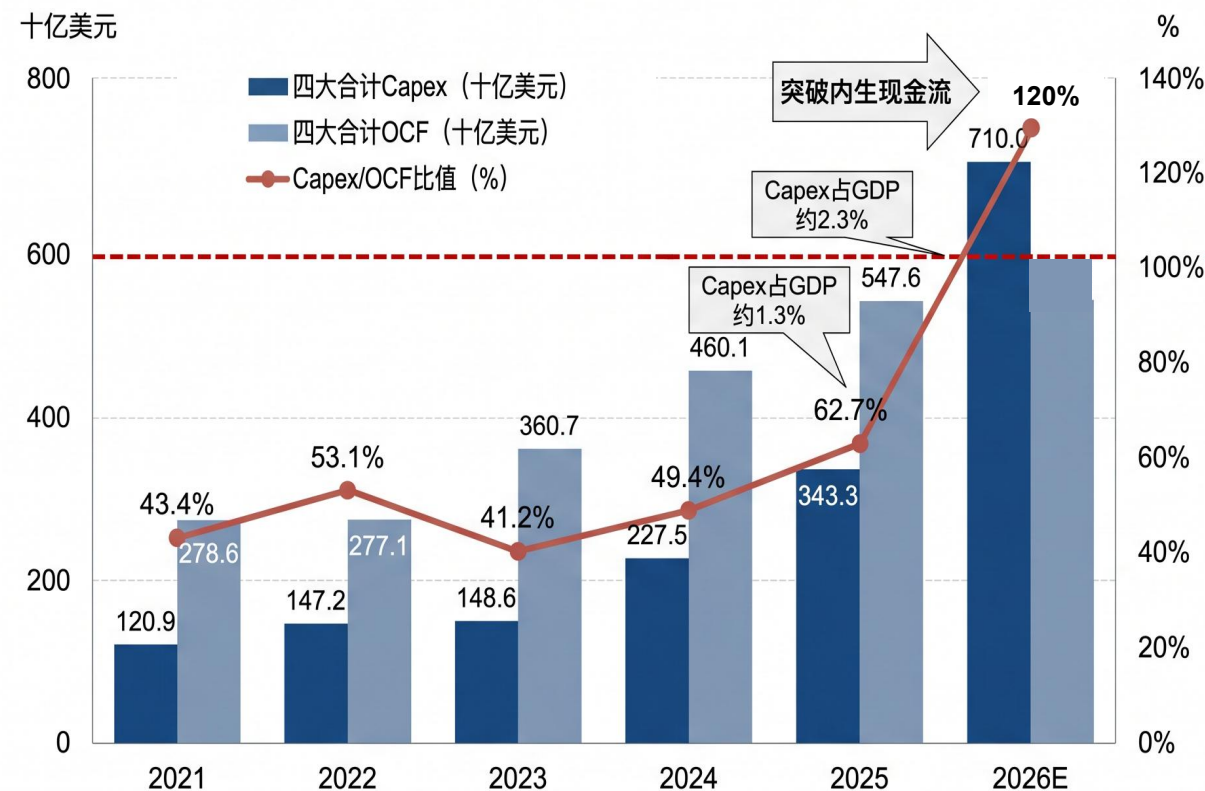
- 过去三年，海外科技厂商的资本开支呈现持续提升趋势，且目前各家资本开支增速指引依然显著高于业务收入/利润增速，我们预计2026年四家资本开支约为经营性现金流的1.2倍。
- 考虑到海外科技厂商（典型如微软、META等）通常还会有约30%净利润进行分红回购，因此未来各家现金流与利润并不足以覆盖资本开支，各家均采用一定融资方案，债券融资成为主力资金来源。

图：海外互联网Capex/经营性现金流变化情况（%）



资料来源：公司财报、彭博一致性预期、国信证券经济研究所整理

图：海外互联网Capex/经营性现金流变化情况（%）

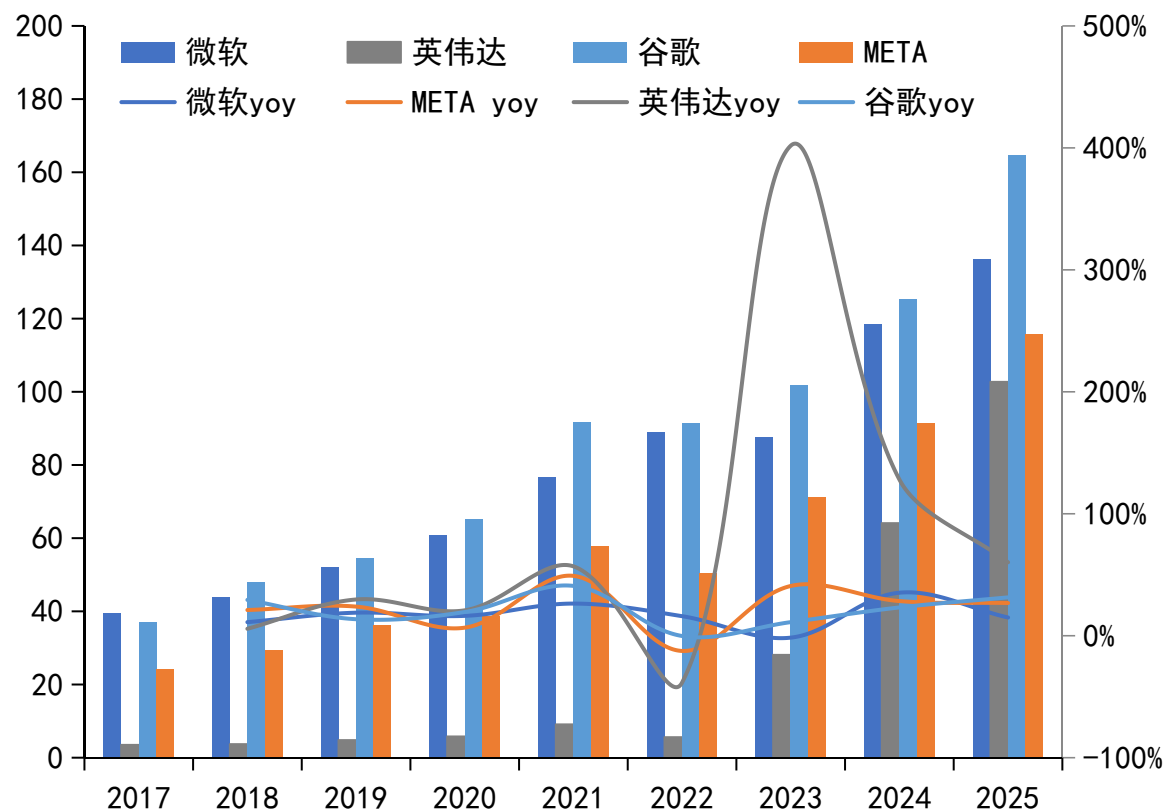


资料来源：公司财报、彭博一致性预期、国信证券经济研究所整理

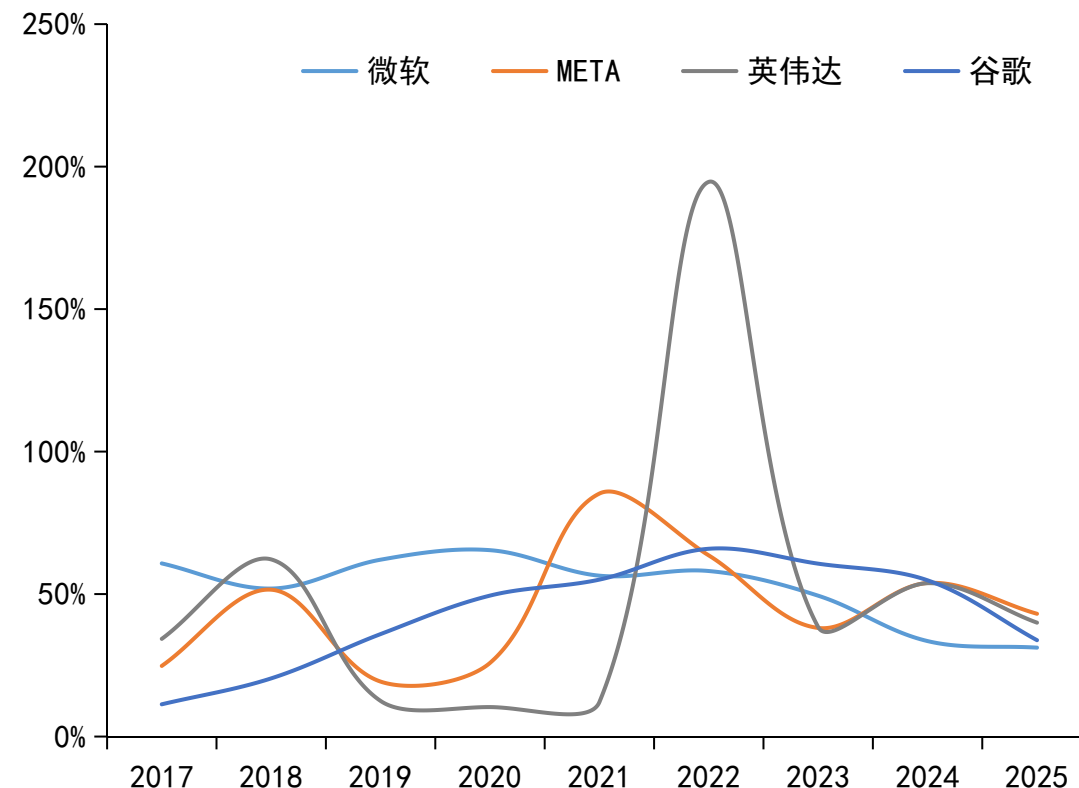
科技巨头回购、资本开支与经营性现金流情况

- 过去8年，海外科技厂商常年将30%-60%的经营性现金流用于股票回购、股利支出，近几年科技厂商由于AI高额资本开支，回购占比提升节奏放缓，稳定在中等区间。
- 英伟达成为新的红利股，2017-2022回购偏低，AI景气兑现之后现金流暴涨，英伟达2023-2025回购占比快速抬升至约50%，2026FY（截止2026-01-25）经营现金流1027亿美元，回购超400亿美金，26年5月新增800亿美元无期限回购授权。

图：科技巨头经营性现金流情况（十亿美元，%）



图：科技巨头分红和回购占比经营性现金流情况



资料来源：彭博，国信证券经济研究所整理

资料来源：彭博，国信证券经济研究所整理

海内外Capex量级：现金流约束下厂商增加融资租赁与金融化操作



- 2020年四大全年发债仅400亿美元，2025年全年发行1210亿美元公司债，**2026年前5个月已发1590亿美元，根据华尔街日报预期全年预计2500 - 3000亿美元。**

表：主要北美科技公司发债情况（亿美元）

时间区间 / 节点	Amazon（AWS）	Microsoft（Azure）	Alphabet	Meta
26H2总债务（长债+短债）	1219	402	795	837
26H2总现金（现金+短期投资）	1067	670	1268	750
2025-2026 年 6月 累计发行总额	820 亿美元以上	约 420 亿美元	约815 亿美元	约550 亿美元
2025年Q4	单笔 150 亿美元多期限美元债（5-30 年期，票面利率 4.6%-4.9%）	两笔合计 145 亿美元，增设超长 30 年期品种	250 亿美元美元债	通过与 BlueOwl 合资 SPV 发行 270 亿美元债务（表外负债，Meta 仅持股 20%，债务不并入母公司报表）
2026年Q1	500 亿美元主力美元债，初始询价 370-420 亿，认购火爆上调发行规模	单笔 70 亿美元美元优先无抵押债	315 亿美元大额债券，内含 10 亿英镑 100 年期超长债（行业稀缺品种）	-
2026年Q2	140 亿加元等值 100 亿美元加元计价债券（加拿大史上最大企业债），同步发行欧元中长期债券	美元债 90 亿 + 欧元等值 20 亿，合计 110 亿美元	-	追加发行 250 亿美元中长期债券，配套数十亿美元私募信贷额度
整体发行特征	大额集中式发行，单笔百亿级融资为主，多币种布局（美元、加元、欧元）	季度小额常态化发债，单笔极少超 200 亿；久期以 10/20 年期为主；30 年期票面 4.2%-4.5%	大额单次发行，创新推出百年超长期限债券	采用母公司直发 + SPV 表外融资双模式，配套公开债券 + 私募信贷多元融资工具

资料来源：公司官网、彭博社、华尔街日报，国信证券经济研究所整理

海内外Capex量级：现金流约束下厂商增加融资租赁与金融化操作

- 海外科技巨头密集发债，认购热度边际回落、融资成本分化：
 - ① **Amazon**：发债成本由2025Q4的4.71%升至2026Q1的5.03%，同期认购倍数由5.33倍连续降至3.41倍和2倍。**亚马逊250亿美元债券发行后，已明确告知承销商2026 全年不再发行任何新增债券**
 - ② **Alphabet**：融资条件相对最优。2025Q4和2026Q1认购倍数分别为5.14倍和4.36倍，需求维持高位；2026Q1发债成本降至1.74%。
 - ③ **Meta**：融资成本维持高位。发债成本由2025Q4的5.83%小幅降至2026Q2的5.75%，认购倍数由4.17倍降至3.84倍。

表：主要北美科技公司发债情况

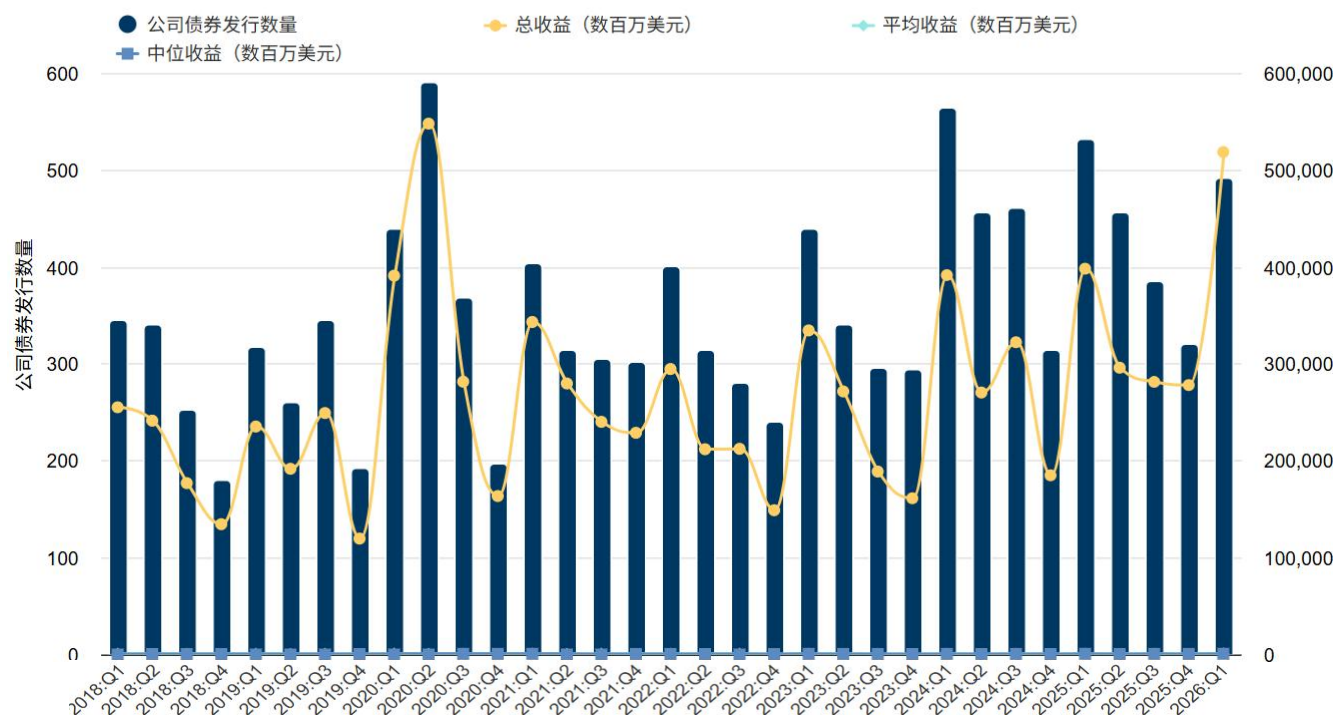
公司	发行类型	核验后规模	加权债务成本	认购倍数
Amazon	2025Q4美元债	150亿美元	YTM 4.709%	5.33倍
	2026Q1美元债	370亿美元	固定债YTM 5.035%	3.41倍
	2026Q1欧元债	145亿欧元	固定债票息 3.704%	约 1.92倍
	2026Q2加元债	140亿加元	YTM 4.344%	约 2.00倍
Alphabet	2025Q4美元债	175亿美元	固定债YTM 4.982%	5.14倍
	2025Q4欧元债	65亿欧元	加权票息 3.438%	未披露
	2026Q1美元债	200亿美元	YTM 4.862%	超过5倍
	2026Q1英镑债	55亿英镑	加权票息 5.313%	100年档 9.5倍
	2026Q1瑞郎债	30.55亿瑞郎	加权票息 1.059%	未披露
Meta	Hyperion项目债	273亿美元	发行收益率 6.58%	无公开订单簿
	2026Q2母公司债	250亿美元	YTM 5.746%	3.84倍

资料来源：公司官网、SEC、路透社，国信证券经济研究所整理

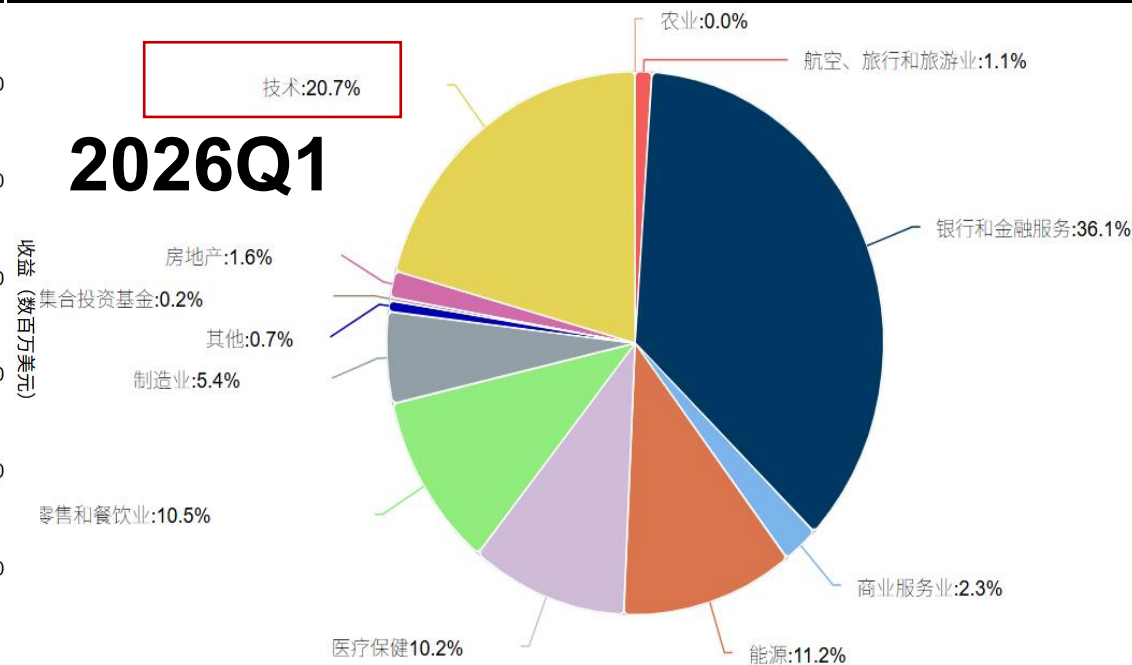
请务必阅读正文之后的免责声明及其项下所有内容

- 根据SEC，2025年美国企业债券发行规模为1.25万亿，2026年一季度达到5180亿美元（同比+30%）。企业债券为构成了美国债券市场重要组成部分，其他组成部分包括美国国债、其他美国政府债券和市政债券。
- 根据SEC行业分类，银行与金融服务占比总规模的1/3以上，其次是科技行业约占比20%（年度约2500亿美金）。科技行业发债规模占比从2025年的19.7%提升到2026Q1的20.7%，体现出AI资本前置背景下各家的融资需求提升。

图：北美企业债券发行数量与规模（百万美元）



图：北美企业债券分行业发行规模占比（%）

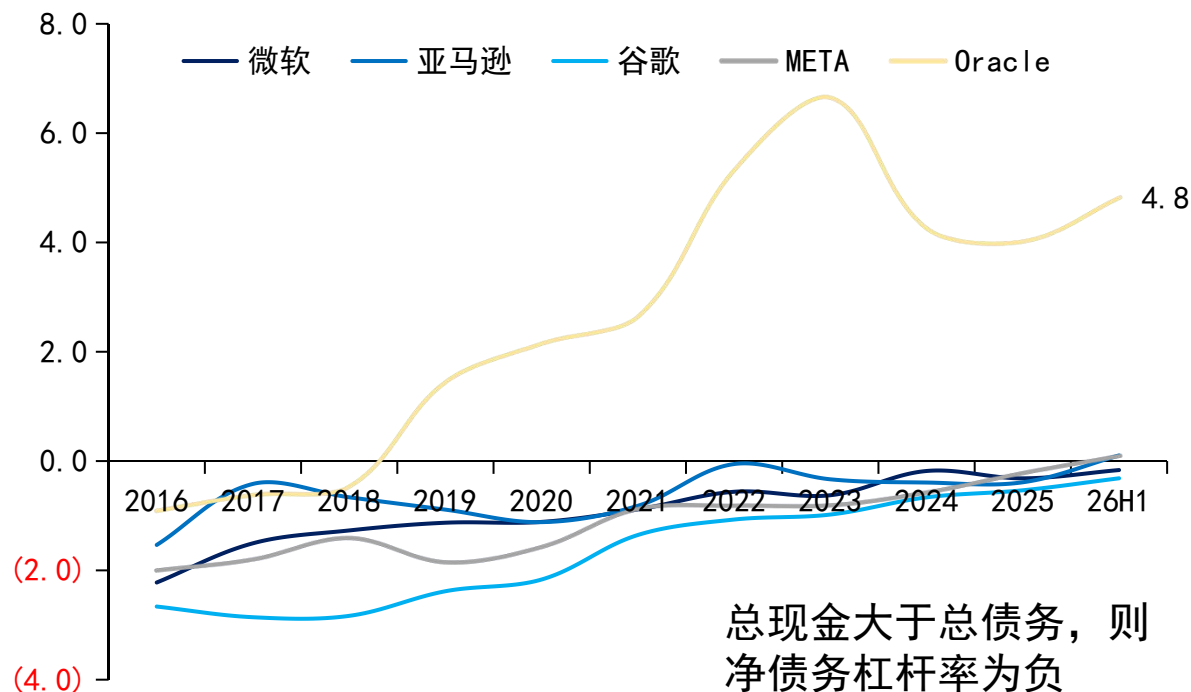


海外科技公司净债务杠杆率变化

- 公司层面“净债务/EBITDA”是反映现金流覆盖与信用承受度的核心指标。
- 当前五大Hyperscaler平均总债务/EBITDA约1.3x，26H1净杠杆约0.3x（不考虑Oracle约0附近），Oracle为个例，净杠杆约4.8x。
- 对标IG基准：北美IG整体中位净债务/EBITDA，A级：1.24x、BBB级：2.38x；投资级中位利息保障倍数(EBITA/当期利息支出)5.8x。

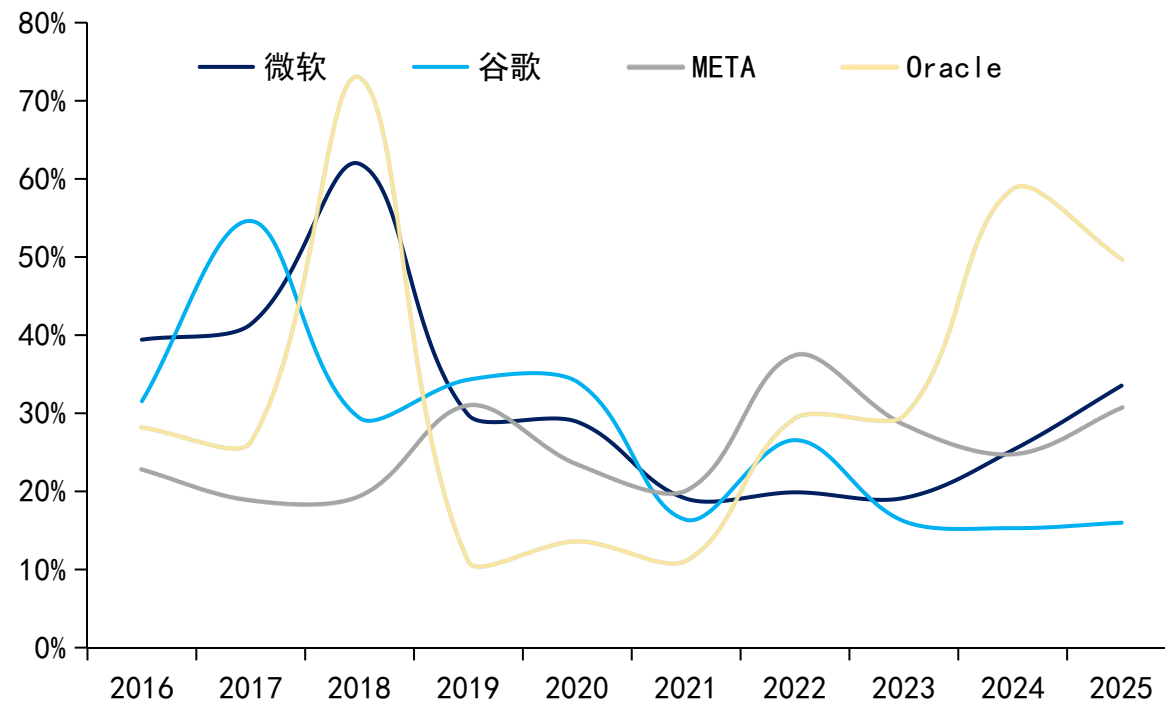
$$\text{净债务杠杆率} = \frac{\text{净债务}}{\text{EBITDA}}$$
$$= \frac{(\text{短期债务} + \text{长期债务} - \text{总现金})}{(\text{营业利润} - \text{折旧摊销})}$$

图：海外互联网净债务杠杆率变化情况



资料来源：公司财报、彭博、国信证券经济研究所整理

图：海外互联网折旧摊销/净利润变化情况



在合理杠杆区间内，未来四年资本开支增速CAGR可达25%

- 以北美IG群体中位数作为安全限制，合理的净债务杠杆率A级约1.24x、BBB约2.38x（若高于该数值，外部融资成本会显著提升）：
 - 对以自有现金流为主的头部Hyperscaler想要处于AA/A评级，**净债务/EBITDA维持在1.2倍以内可兼顾AI高Capex与信用稳定**；
 - 对处于BBB级（如Oracle），净债务/EBITDA最好控制在约2-3倍，超过3倍需以更强的合约现金流与结构性融资对冲风险；
- **可加杠杆空间**：根据彭博2026年微软、谷歌、亚马逊、META约有6000亿美元经营性现金流，假设债务AA评级、即净杠杆不超1.2倍，若假设经营性现金流年化增速20%，**四年累计债务上限1.5万亿美元、即每年可以新增0.33x经营性现金流债务**。2030年经营性现金流到1.24万亿美金，**每年资本开支可达经营性现金流的1.33x左右。未来四年资本开支增速CAGR约在25%，2030年总资本开支有望超1.6万亿。**
- **市场约束**：美元IG指数的单一发行人权重上限（约2.3%）意味着短期内面向公募美元债的容量约在4500亿美元量级，之后需转向非美元、项目融资与私募等渠道。

表：主要北美科技公司资产负债率情况

主体	S&P评级	26H1 净债/EBITDA	资产负债率/负债净资产率	Capex/OCF (2026)	备注
微软	AAA	-0.2x	43%	52.9%	2026年Capex约1900亿美元，AI容量全年+80%、两年翻倍；26Q1 Commercial 待履约订单/RPO 约6270亿美元。
谷歌	AA+	-0.3x	32%	55.5%	Capex/OCF逼近1，26Q1待履约订单/RPO 4620亿美元，Cloud backlog支撑收入可见性。
亚马逊	AA	0.1x	52%	94.5%	AWS 26Q1待履约订单/RPO 3640亿美元，AI年化收入大于150亿美元，保障“以量换现”。
Meta	AA-	0.1x	38%	60.8%	2026年Capex指引1250-1450亿美元；通过“老代际GPU推理商业化+新代际训练自用”的资产配置压缩闲置。
甲骨文	BBB	4.8x	66%	n/a	26Q1待履约订单/RPO 5530亿美元，其中约3000亿美金是与OpenAI的星际之门等合作，落地具备一定不确定性。引入外部CFO、组织与资本结构优化配合重资本周期。

资料来源：公司官网、彭博社、华尔街新闻，国信证券经济研究所整理测算（采用彭博一致性预期）

表：北美科技公司经营性现金流与债务对应

亿美元	经营性现金流	净债务杠杆率1.2倍对应债务	每年新增0.33x债务
2026	6,000	7,200	1,980
2027	7,200	8,640	2,376
2028	8,640	10,368	2,851
2029	10,368	12,442	3,421
2030	12,442	14,930	4,106
总计	44,650		14,734

资料来源：公司官网、国信证券经济研究所整理测算（采用彭博一致性预期，假设经营性现金流年化增速20%，每年资本开支为经营性现金流的1.33x）

海内外科技厂商关于资本开支与算力容量的相关表述

表：海外科技厂商关于资本开支与算力容量的相关表述

	资本开支表述	AI 算力表述	数据中心算力表述	数据中心功耗表述	电力供应表述
微软	2026自然年资本开支约1900亿美元（含约250亿美元组件涨价，占比13%） - 微软每年在AI基建上的固定投资超过800亿美元，这是十年规划。	公司计划FY26（25Q3-26Q2）将AI总容量提升超80%。	26Q1业绩会：本季度新增1GW算力容量，计划2年内整体算力翻倍。	25Q2业绩会：数据中心超400个，过去一年新增超2GW； - 微软CE025年10月发言：微软云全球数据中心总功耗5GW；	
谷歌	2026年26年上修50亿达到1800-1900亿，主要是收购了能源公司Intersect有关。并且指引27年将继续大幅增长				
META	上调2026年资本开支至1250-1450亿美元	24年底计拥有等效60万H100GPU的计算能力；			
亚马逊	预计26年全年将达到2000亿美元，同比+56%。	26Q1：芯片业务ARR200亿美元，独立对外销售，ARR将500亿美元	25Q3业绩会：AWS目标2027年底实现算力翻倍		25Q4业绩会：过去12个月新增3.9GW电力容量，预计到27年底将翻倍，仅25Q4就新增1.2GW。
OpenAI	- 星际之门计划投资超5000亿美元； - OpenAI承诺在未来多年内向微软Azure云服务采购约2500亿美元的服务；	英伟达未来供应10GW数据中心的芯片；AMD未来供应6GW芯片；博通未来供应10GW芯片；		2025：未来四年内星际之门Stargate计划容量达到近10GW；	

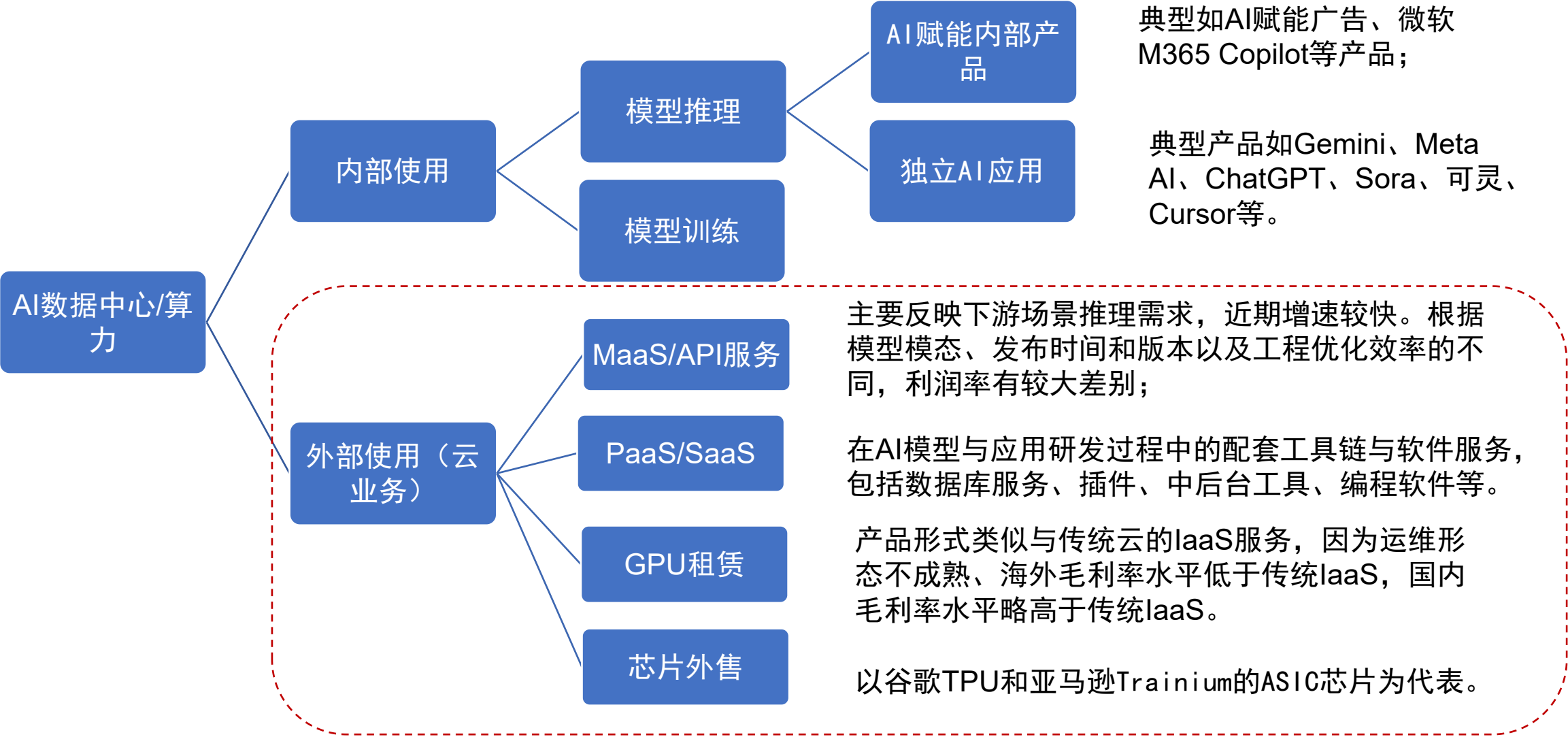
【 01 】 资本开支水位与可行性区间

【 02 】 科技大厂AI投资的ROI测算

【 03 】 大模型厂商财务模型与ROI

【 04 】 风险提示

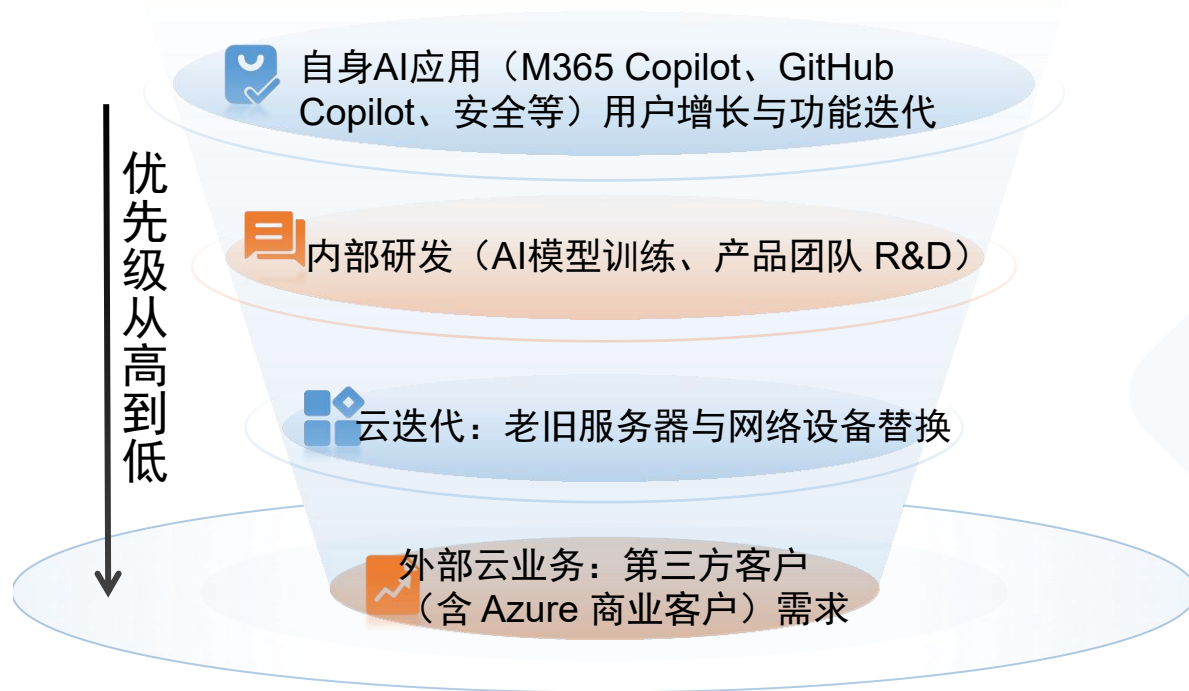
图：AI数据中心算力分配与商业化变现路径



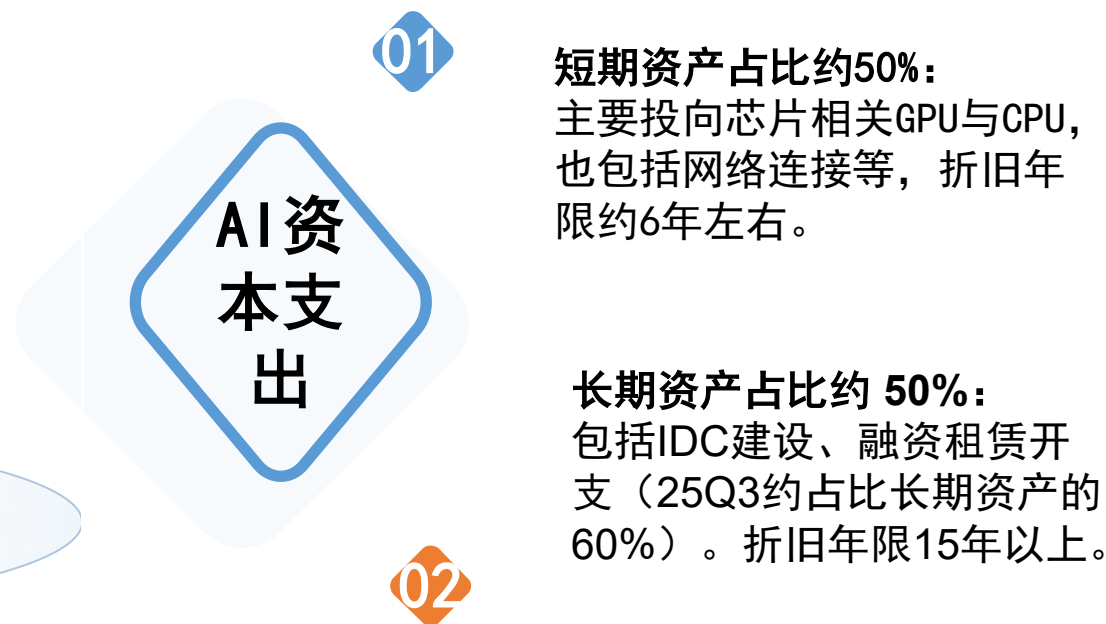
M365 Copilot付费率与云收入占比约4.5%

- 26Q1业绩会，微软表示M365 Copilot 付费席位突破2000万（M365商业付费席位约4.5亿）、**付费率约4.5%**，假设单用户、月付费30美元，对应年化收入约70亿美金，**占比微软云2026自然年收入约4.5%**。
- 26Q1业绩会全面从“席位订阅制”转向 **席位 + 用量消费混合模式**，GitHub Copilot率先切换。企业在基础席位外，为高价值AI任务支付用量费用，Copilot、GitHub、Dynamics 365 均已验证；长期看，将推动营收从“用户数驱动”转向“使用强度驱动”，支撑FY27及后续双位数增长，同时提升客户粘性与长期ROI。

图：微软AI算力分配优先级



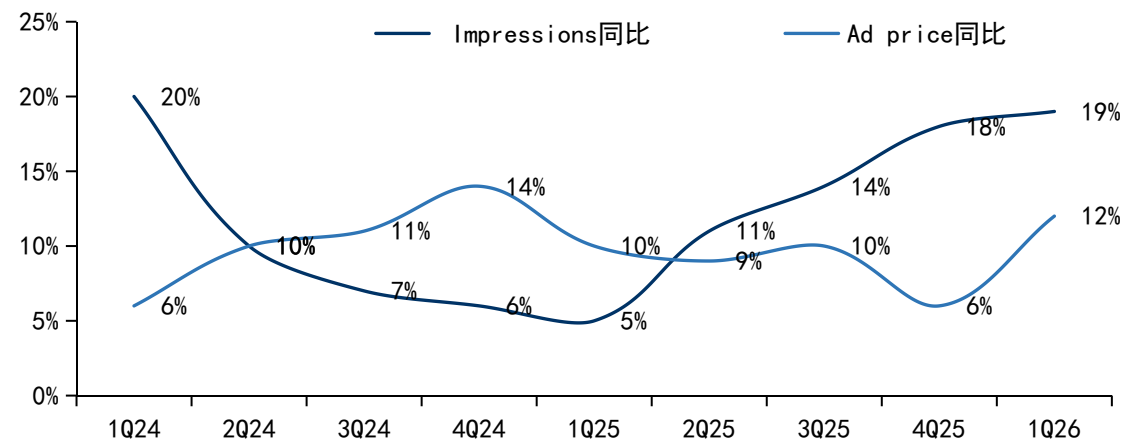
图：微软AI资本开支分配



META Capex分配与ROI：AI约拉动10%广告收入增长

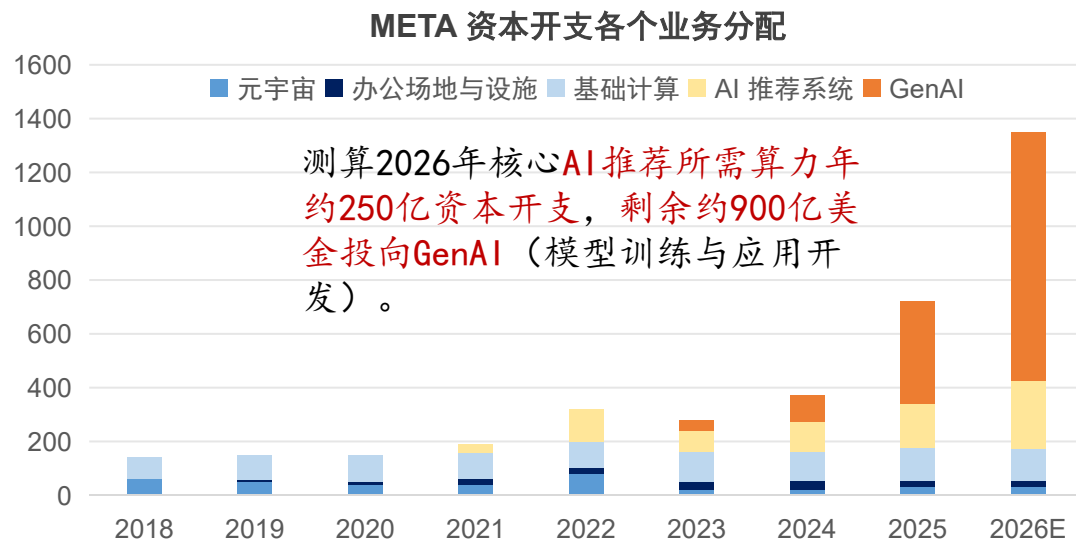
- 过去两年META广告业绩增速显著跑赢广告大盘约10%，测算约10%的广告收入增长主要来自于AI驱动，对应赋能收入今年约200亿美金。AI通过精准推荐带动广告曝光量（Impression）增长，通过Advantage+工具带动广告单价（CPM/AdPrice）增长。
- “26Q1财报会：用AI优化推荐系统，Instagram 用户交互序列长度翻倍，Facebook实时使用时长提升10%，全球视频观看时长增超8%。”

图：广告量价变化



资料来源：公司业绩会、国信证券经济研究所整理

图：METACapex使用分配



资料来源：Meta财报会、国信证券经济研究所整理测算

- 1) 办公场地与设施：非技术资本支出；
- 2) 元宇宙业务：主要与硬件相关，如VR/AR眼镜；
- 3) 核心应用家族（FoA）业务：
 - ① 基础计算（CoreCompute）：主要是维护以托管和运行应用程序，包括像Reels短视频功能上的视频内容；
 - ② AI推荐系统（AI content ranking）：
 - 内容推荐引擎底层硬件更新；
 - 广告技术平台升级，如Advantage+新产品；
 - ③通用人工智能计算（GenAI）：如训练Llama，为MetaAI提供推理算力以及AI Agent等。

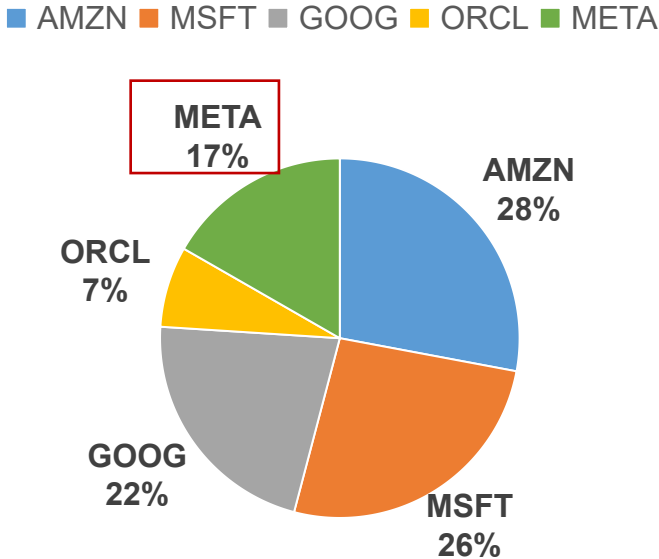
META算力规模：占比全球英伟达卡总等效算力的12%

- 23年底META表示24年底计划购买35万片英伟达H100，加上其他GPU约等效60万H100GPU的计算能力。Omdia报道，2023年Meta和微软各购得15WH100，谷歌、亚马逊、甲骨文和腾讯各购得5WH100，特斯拉1.2WH100。
- 24Q4业绩会，CEO表示2024年底将部署35万张实体H100，整体AI总算力等效近60万张H100。测算25年底等效H100卡为270W张，2026年底预计达等效550W片H100，占比全球英伟达卡总等效算力的11.7%（测算2026年底全球等效H100算力为4700W片）。按照卡类型测算，META的A卡占比全球总量约12%、H卡占比约10%、B卡占比约14%。

表：META采购英伟达的算力规模预测

年份	型号	GPU（万张）	等效H100（万片）	年度资本开支（亿美元）	对应AIDC功耗（MW）	全球出货总量（万）	数量占比
2022	A100	7.6	1	320	69	65	12%
2023	H100/H200	15	15	270	239	125	12%
2024	H100/H200	35	43	372	557	400	9%
2024	B200	4	9		91	30	13%
2025	B200/B300	80	203	722	2182	480	17%
2026	B200/B300	50	127	1350	1507	425	12%
2026	R200/RV200	33	150		1350	383	9%
总计		225	548	3,034	5,994	1,908	

图：头部科技厂四年累计Capex投入份额



资料来源：TNP、Meta财报会、国信证券经济研究所整理

资料来源：各公司业绩会、国信证券经济研究所整理

META 2026年计划部署7GW算力：自建为主，算力租赁约1GW

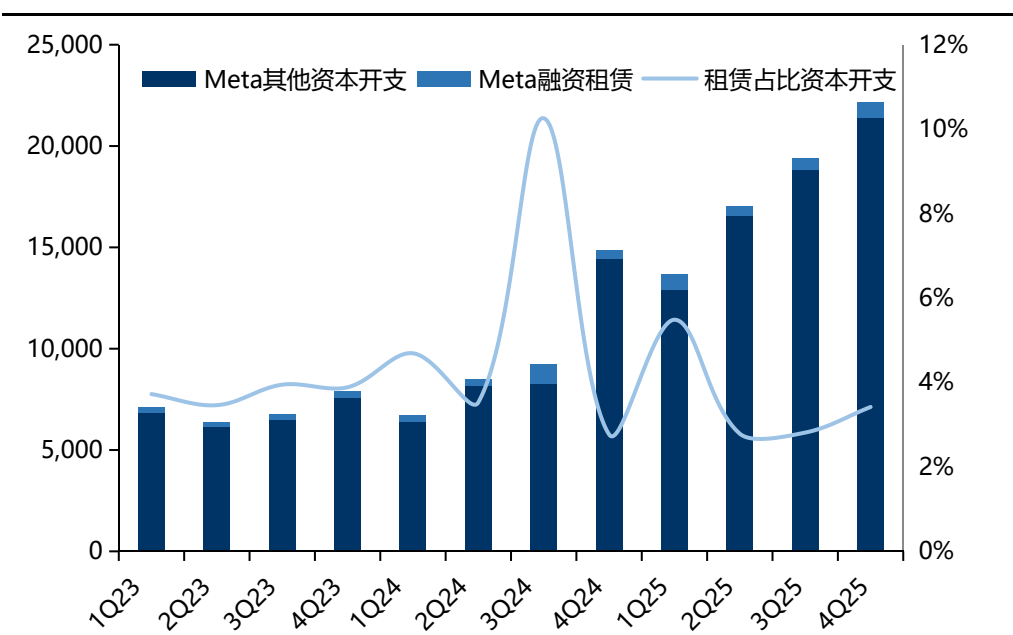
- 路透社近期报道，Meta 2026年计划部署7GW的计算基础设施，并计划在2027年翻一倍。根据测算：
 - ① **META全资自建**：测算截止25年英伟达芯片对应约**3GW**算力，今年采购约2.8GW算力预计今年有**1GW**算力落地（考虑建设周期半年左右）；AMD与Meta 6GW InstinctGPU多代部署协议在26H2首个**1GW**出货；2026年Meta的ASIC芯片MTIA（与博通合作）预计出货小几十万片，主要由MTIA3（Iris）贡献。
 - ② **表外合资共建**：路易斯安那州Hyperion超算园区（长期规划5GW），由BlueOwl Capital等外部资本持股80%，Meta以10年以上长期包租协议获得独家使用权，该项目2026年首批交付约**1GW**，2027年交付1-1.5GW；
 - ③ **算力租赁**：略超**1GW**规模（测算见下表）。根据业绩**META资本开支中融资租赁占比在5%左右**，2025H2开始密签订算力租赁订单。

表：META租赁的算力规模预测

合作方	合同金额	容量口径	2026年状态	关键点
Nebius（首份）	~30亿美元/5年	测算约0.04GW	2026年2月全部上线	两期交付，分别确认12/11个月收入
Nebius（新约）	270亿美元/5年	测算约0.36GW	2027年初起交付，	
CoreWeave	140亿美元	测算约0.19GW	续行	2025年9月签订的长期供给框架。
CoreWeave	210亿美元/至2032年12月	测算约0.23GW	2026H2起交付，含Rubin初始部署	2026年4月新合同
GoogleCloud	~100亿美元/6年	测算约0.11GW	续行	2025年8月签订，2026年持续供给
Crusoe	—	~1.6GW	报道披露，细节待确证	为2026-2027提供大体量的新增功率池

总计 2026年约交付略超1GW算力，2027年约交付2.5GW算力

表：META算力租赁的占比变化



算力租赁模型：Colossus集群预计约两年即可收回成本

- Colossus 2 GPU目标集群建设成本约为**350 - 400亿美元**（Colossus1不到80亿美金），在功率达到1 - 2GW量级，两个集群已签订的算力租赁合约显示出年化**260亿美金**的稳定现金流，且并未出租所有算力，Colossus 形成“快建、低单MW成本、强租金”的可复制收益模型。
- 仅考虑市场散单租赁价格，假设租赁价格H卡 \$2.5/GPU·小时、A卡 \$1.63/GPU·小时，META A/H卡全部租赁情况下，**年度租金可达110亿美金，对META EPS贡献约4.7美元，可拉动META今年EPS 15%增长。**

表：Colossus集群租赁合同

客户	租金	合同期
Anthropic	12.5亿美元/月，年化150亿美金	2026-06至2029-05（36个月）
Google	9.2亿美元/月，年化110亿美金	未披露

资料来源：华尔街新闻、公司官网、国信证券经济研究所整理

表：Colossus集群租赁合同

集群/基准	周期	规模/说明
Colossus I 首簇	122天	~10万H100，约130MW。
Colossus I 二簇	90天	10万张H100/H200与约3万张GB200 NVL72。
Colossus II 首簇	91天	~11万GB200，约210MW。
Colossus II 二簇	64天	~11万GB300，约220MW。

资料来源：华尔街新闻、公司官网、国信证券经济研究所整理

表：META A/H全部租赁下 租赁单价与META EPS弹性测算

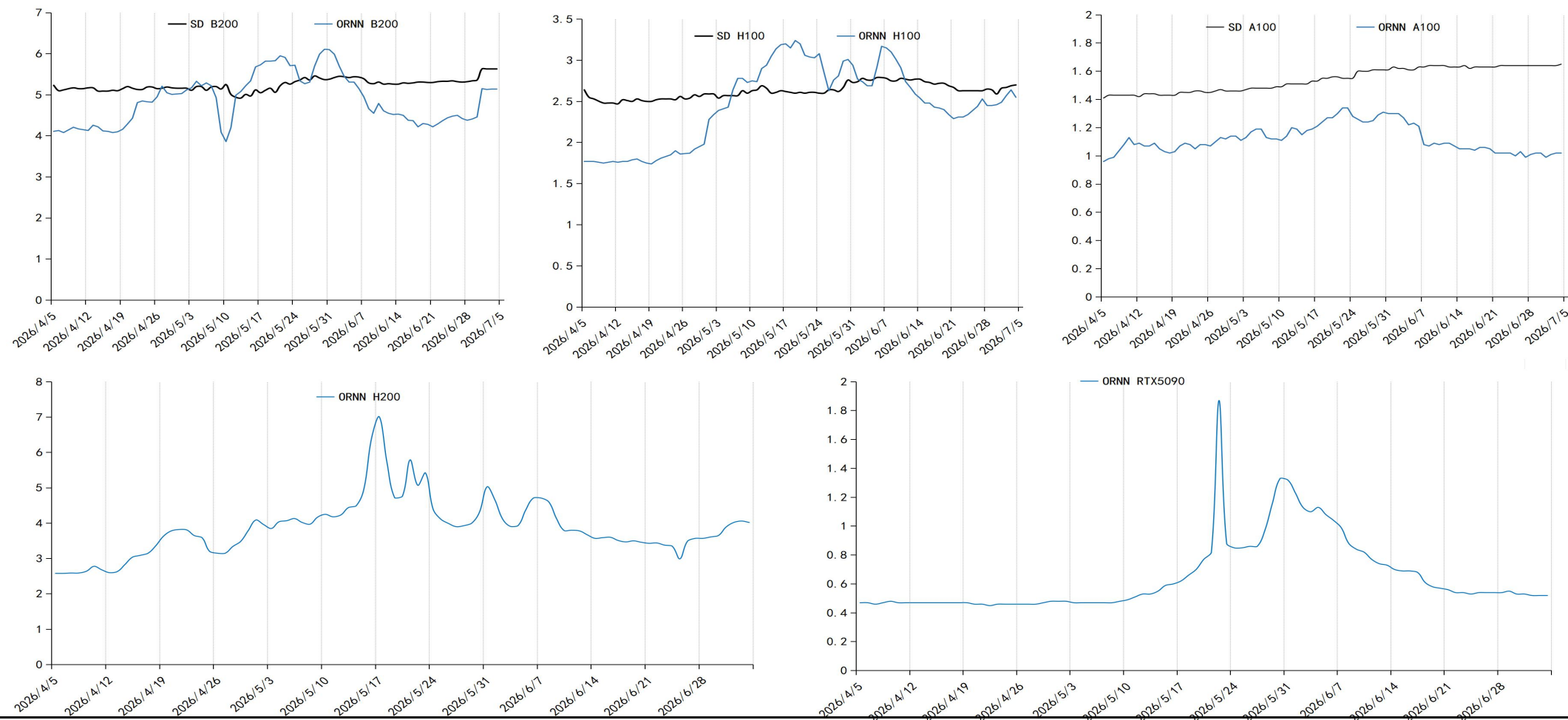
A/H (美元/GPU·小时)	2	2.3	2.5	2.8	3
1.3	3.8	4.3	4.7	5.2	5.5
1.5	3.9	4.4	4.7	5.2	5.6
1.6	3.9	4.4	4.7	5.3	5.6
1.8	3.9	4.4	4.8	5.3	5.7
2	4.0	4.5	4.8	5.4	5.7

资料来源：SiliconData、彭博、Meta财报会、国信证券经济研究所整理

算力租赁价格：边际价格上涨，AWS全线提价20%

- 2026年7月初B200/H100/H200成交价格均出现上涨，其中B200上涨幅度最大，主要是AWS对全线GPU产品从7月1日开始提价20%的影响。

表：算力租赁价格变化（美元/GPU*小时）



资料来源：SiliconData (Neocloud租赁报价)、ORNN (成交价)、国信证券经济研究所整理

云巨头AI年化收入：约占比云收入的20%-30%

图：2026Q1各个公司AI年化收入占比说明

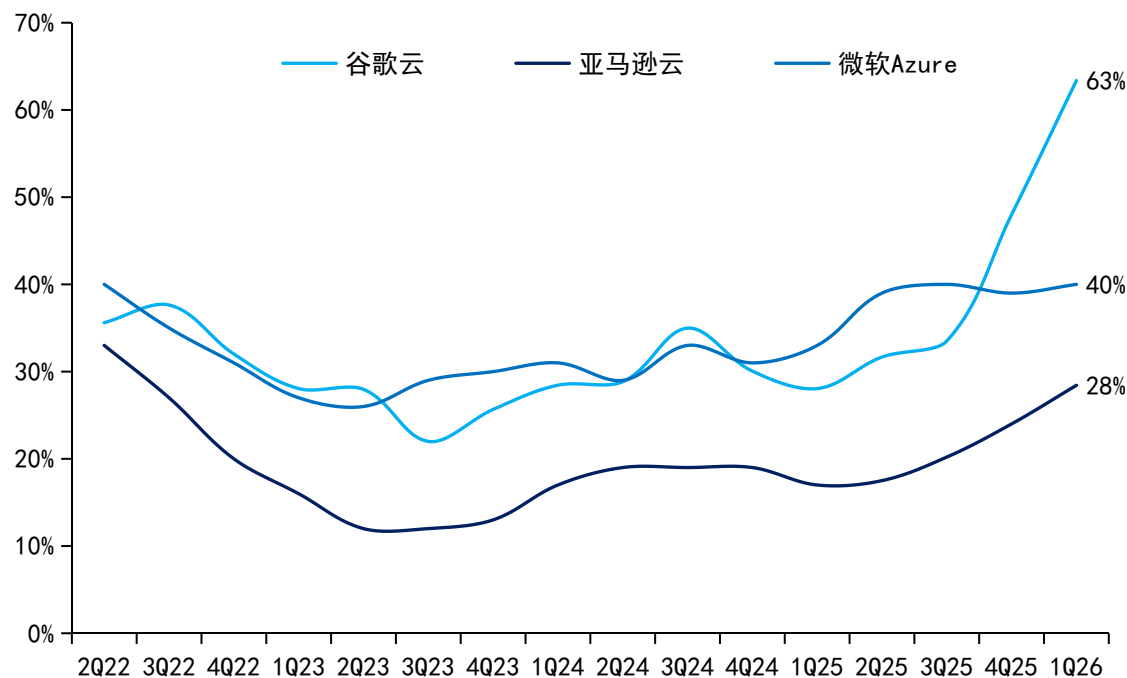
	亚马逊AWS	微软智能云	谷歌云（GoogleCloud）
AI年化营收估算	超150亿美元，同比三位数增长	超370亿美元（全栈AI），同比+123%	约50亿美元，同比近800%，G谷歌云本季度同比63%、AI贡献了超过一半的增长加速
AI营收占比云收入	约占比AWS整体1500亿年收入的10%	约占比智能云25%、占比Azure37%	约占比谷歌云6%
总结	MaaS/API收入（占比云收入高个位数）+PaaS等解决方案（约5%）+GPU租赁收入(占比超10%)≈20%-25% +内部业务赋能（占比低个位数）≈25%-30%		
官方表述原文	AndyJassy：在这波AI浪潮的前三年，AWS的AI收入年化运行率已超过150亿美元，几乎是AWS推出三年后规模的260倍 BrianOlsavsky：我们的AI收入同比实现三位数增长。	AmyHood：我们的商业AI收入年化运行率现已超过370亿美元，同比增长123%，成为微软历史上增长最快的业务	SundarPichai：我们的企业AI解决方案收入同比增长近800%，目前年化运行率超过50亿美元。这是AI首次成为GoogleCloud增长的主要驱动力
AI年化收入口径区别	AI软件与服务（MaaS、开发工具、智能体）等，不包含自研芯片租赁收入。	全栈AI服务：云AI服务+所有B/C端的商业Copilot产品+BingAI等；	仅包括ToB的企业级生成式AI解决方案，如GeminiEnterprise订阅、模型API调用等。
AI年化收入口径业绩会原文	AndyJassy（AI收入包含内容）：包括SageMaker模型构建、Bedrock的前沿模型推理、Strands私有数据智能体开发、AgentCore生产级部署，以及Kiro编码、Transform迁移等开箱即用智能体	AmyHood：商业AI收入包括AzureAI服务、Microsoft365Copilot、GitHubCopilot、Dynamics365Copilot以及所有其他AI驱动的商业产品	SundarPichai：当我们谈论企业AI解决方案收入时，我们特指来自GeminiEnterprise、AIAPI、智能体开发平台以及为企业客户提供的定制AI服务的收入

资料来源：公司财报会、semianalysis，国信证券经济研究所测算整理测算

海外云收入与利润率：研芯片产能爬坡带动云厂利润率优化

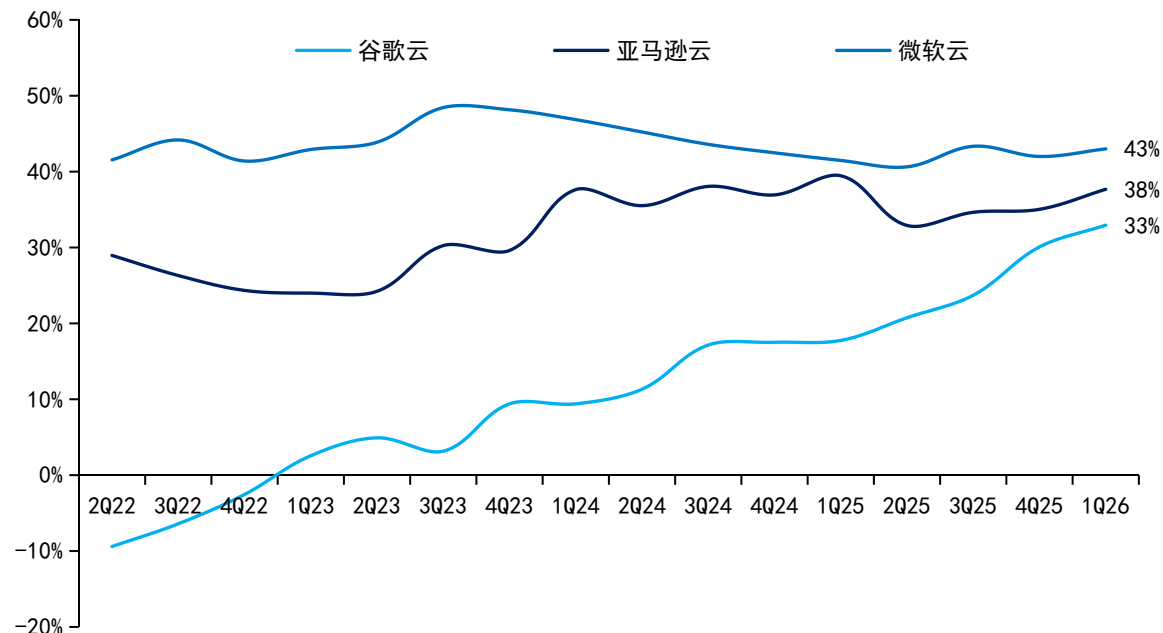
- 研芯片产能爬坡带动云厂利润率优化。26Q1海外云厂收入加速增长，其中谷歌、亚马逊表现亮眼，谷歌云收入增速已达到63%，AWS也连续多个季度加速，但Azure则保持平稳，我们认为核心原因在于TPU和Trainium芯片得到验证后极大缓解了两家公司的供给压力，但微软的自研芯片尚未获得大范围认可。利润端，自研芯片也能够缓解云厂的成本压力（根据公司财报，Trainium合作的世芯毛利率仅20%+，博通在60%，因此谷歌寻求了联发科合作，而英伟达毛利率高达75%）。

图：海外云厂收入增速



资料来源：公司财报、国信证券经济研究所整理

图：海外云OPM



资料来源：公司财报、国信证券经济研究所整理

微软AI云业务投资与回报ROI测算

- 前期高折旧摊销压力下AI云OPM低于当前智能云OPM水平（25Q3为43%），现金流端ROI小于1，后伴随着AI云规模增长与算力成本下降、利润率提升，**2030年现金流接近转正**。资本回报端持续提升，**2030年ROIC可达到17%，即六年可收回投资周期**。

表：微软云AI投资与ROI测算

百万美元/%	2023	2024	2025	2026E	2027E	2028E	2029E	2030E	备注
Capex (含融资租赁)	41,200	75,600	118,000	191,160	248,508	298,210	357,852	429,422	假设依据26FY增速高于25FY，当前微软净利润体量1000亿美金
CapexYOY	46%	83%	56%	62%	30%	20%	20%	20%	
Capex中AI投资占比	40%	72%	90%	90%	90%	90%	90%	90%	
CapexAI增量	16,630	54,080	106,200	172,044	223,657	268,389	322,066	386,480	假设依据：25Q3短期资产占50%，约6年折旧周期。长期资产占比50%、约15年折旧周期。
年中GenAI投入总量	8,360	45,130	125,270	264,392	462,243	708,266	1,003,493	1,357,766	
折旧（综合折旧年限为10年）	873	4,633	12,527	26,439	46,224	70,827	100,349	135,777	
运维成本	0	4,189	10,123	21,364	38,836	60,128	83,836	110,442	假设资本开支投出一年后开始使用每年运维成本为总投入capex的12% 其余内部使用与OpenAI消耗
总成本	873	8,822	22,650	47,803	85,060	130,955	184,185	246,218	
云业务（外部）所用算力占比	40%	55%	55%	60%	65%	70%	75%	75%	
微软智能云收入YOY	96,213	104,859	120,433	160,176	216,237	298,408	396,882	515,947	假设依据：IDC预测28年全球公有云约1.7W亿美元，假设微软份额约18%（3000亿美元）、当前13%，假设Azure占比智能云90%（当前75%）约2700亿美元。
Azure收入	63,789	72,468	88,901	124,461	176,735	256,266	358,772	495,105	
其中AzureYOY	28%	31%	38%	40%	42%	45%	40%	38%	
AI贡献Azure收入	3%	10%	19%	27%	40%	55%	68%	75%	假设依据算力成本下降、规模效应提升。当前云业务OPM为43%
AI云收入	4,441	11,318	16,891	33,605	70,694	140,946	243,965	371,329	
AI云YoY		155%	49%	99%	110%	99%	73%	52%	
整体云营业利润率	46%	44%	42%	42%	43%	45%	47%	48%	假设依据算力成本下降、规模效应提升。当前云业务OPM为43%
AI云营业利润率		10%	27%	28%	30%	33%	38%	43%	
当年AI云营业利润-折旧/当年AI投入			20%	24%	35%	51%	70%	98%	
当年AI云营业利润/累计AI投入			7%	7%	8%	10%	14%	17%	大于100%则当年现金流正贡献 资本回报率：17%则6年收回投资

资料来源：公司财报会、彭博、semianalysis，国信证券经济研究所测算整理测算*口径并未包括对传统云业务、内部其他业务自用以及对OpenAI投资回报

请务必阅读正文之后的免责声明及其项下所有内容

单芯片TCO来看，ASIC芯片更有优势

图：GPU、TPU、Trainium芯片TCO比较

单位：美元	单位	GB200NVL72 (Spectrum)	GB300NVL72 (Spectrum)	Trainium2NL1 62DTorus	Trainium2NL3 2x23DTorus	Trainium3NL1 62DTorus	Trainium3NL3 2x23DTorus	Trainium3NL3 2x2Switched	Trainium3NL7 2x2Switched	TPUv7- 3DTorus- Internal	TPUv7- 3DTorus- External
服务器成本	USD	\$3,179,652	\$3,997,587	\$154,020	\$655,793	\$170,554	\$693,183	\$757,415	\$2,020,274		
服务器维护服务	USD	\$72,000	\$72,000	\$16,000	\$64,000	\$16,000	\$64,000	\$64,000	\$144,000		
网络成本	USD	\$296,733	\$296,733	\$26,300	\$105,202	\$26,300	\$105,202	\$105,202	\$236,700		
存储成本	USD	\$42,021	\$42,021	\$5,800	\$23,200	\$5,800	\$23,200	\$23,200	\$52,200		
软件授权及其他成本	USD	\$20,677	\$23,057	\$2,448	\$9,838	\$2,912	\$11,684	\$12,956	\$30,826		
其他安装成本	USD	\$72,000	\$72,000	\$16,000	\$64,000	\$16,000	\$64,000	\$64,000	\$144,000		
单服务器总前期集群资本开支	USD	\$3,683,083	\$4,503,398	\$220,569	\$922,032	\$237,566	\$961,269	\$1,026,772	\$2,627,998		
单服务器GPU数	#	72	72	16	64	16	64	64	144		
单GPU总前期资本开支	USD	\$51,154	\$62,547	\$13,786	\$14,407	\$14,848	\$15,020	\$16,043	\$18,250		
WACC		9.4%	9.4%	9.4%	9.4%	9.4%	9.4%	9.4%	9.4%		
折旧周期	年	4	4	4	4	4	4	4	4		
单服务器每月总资本成本	USD/mth	\$92,311	\$112,871	\$5,528	\$23,109	\$5,954	\$24,093	\$25,734	\$65,867		
单GPU每小时资本成本	USD/hr/GPU	\$1.76	\$2.15	\$0.47	\$0.49	\$0.51	\$0.52	\$0.55	\$0.63		
电力成本	USD/kWh	\$0.09	\$0.09	\$0.09	\$0.09	\$0.09	\$0.09	\$0.09	\$0.09		
利用率	%	80%	80%	80%	80%	80%	80%	80%	80%		
电源使用效率 (PUE)	Ratio	1.35	1.35	1.35	1.35	1.35	1.35	1.35	1.35		
每月每kWIT负载电力成本	USD/kW/mth	\$68.60	\$68.60	\$68.60	\$68.60	\$68.60	\$68.60	\$68.60	\$68.60		
机房托管成本	USD/kW/mth	\$110.00	\$110.00	\$110.00	\$110.00	\$110.00	\$110.00	\$110.00	\$110.00		
每月每kW总托管成本	USD/kW/mth	\$178.60	\$178.60	\$178.60	\$178.60	\$178.60	\$178.60	\$178.60	\$178.60		
整机功耗	kW	143.05	160.05	12.46	50.16	15.77	63.34	72.43	174.89		
每月服务器托管总成本	USD/mth	\$25,548	\$28,584	\$2,225	\$8,958	\$2,817	\$11,312	\$12,935	\$31,233		
远程运维与支持工程师成本	USD/mth	\$1,179	\$1,179	\$262	\$1,048	\$262	\$1,048	\$1,048	\$2,358		
网络连接成本	USD/mth	\$351	\$351	\$78	\$312	\$78	\$312	\$312	\$702		
单服务器每月总运营成本	USD/mth	\$27,078	\$30,114	\$2,565	\$10,318	\$3,157	\$12,672	\$14,295	\$34,293		
单GPU每月运营成本	USD/mth	\$376	\$418	\$160	\$161	\$197	\$198	\$223	\$238		
单GPU每小时运营成本	USD/hr/GPU	\$0.52	\$0.57	\$0.22	\$0.22	\$0.27	\$0.27	\$0.31	\$0.33		
单GPU每小时成本TCO	USD/hr/GPU	2.28	2.72	0.69	0.71	0.78	0.79	0.86	0.96	1.28	1.60

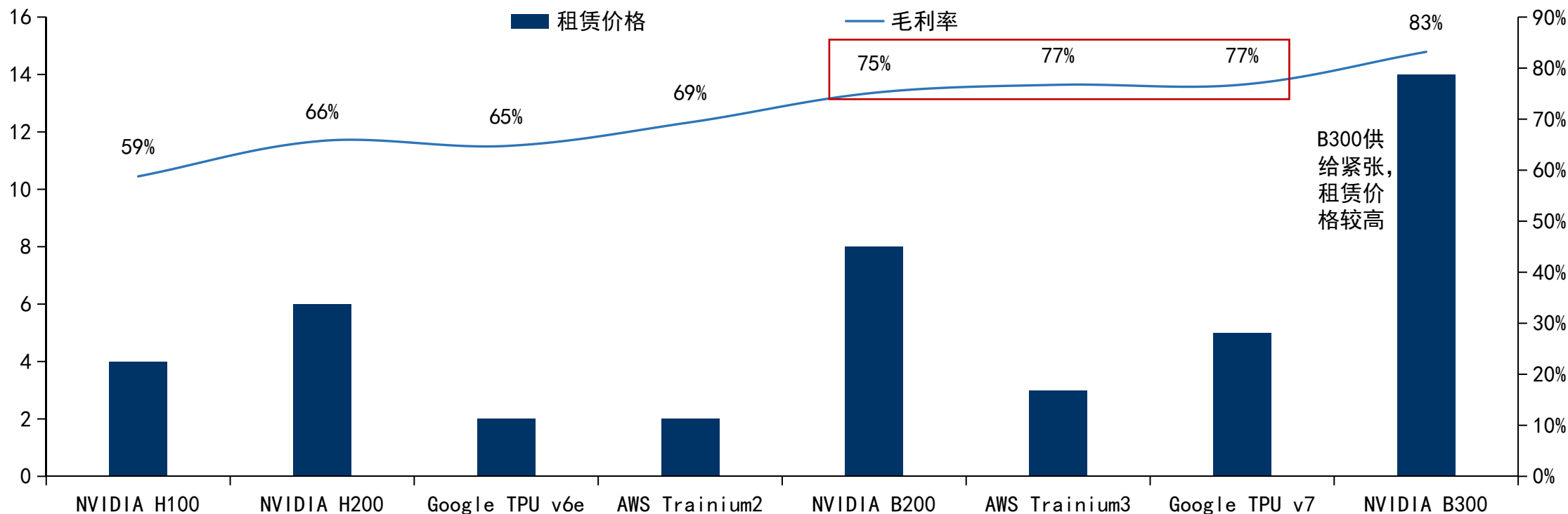
资料来源：SemiAnalysis、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

新产品毛利率水平显著提升，云厂自研芯片能够有效降低成本

- 新产品的毛利率水平显著优于老产品。芯片厂商新产品的单位算力的TCO成本和功耗明显下降，对应租赁的ROI更高。
- 当前，自研ASIC芯片比英伟达GPU芯片毛利率略高、主要解决算力供给的瓶颈。后续，随着自研芯片在性能上逐步追赶英伟达，同时推理需求提升（ASIC更适合推理场景），预计毛利率优势逐步体现。

图：Trainium、GPU、TPU的毛利率对比（注：仅考虑服务器和运营成本，租赁价格未考虑折扣）



资料来源：AWS、GCP、国信证券经济研究所整理测算

谷歌AI云业务投资与回报ROI测算

- 当前现金流端ROI小于1，后伴随着AI云规模增长与算力成本下降、利润率提升，2029年现金流贡献转正。资本回报端持续提升，2030年ROIC可达到38%，即2.5年可收回投资周期。

表：谷歌云AI投资与ROI测算

百万美元/%	2023	2024	2025	2026E	2027E	2028E	2029E	2030E	备注
Capex (含融资租赁)	32,251	52,535	91,447	180,151	234,196	304,454	380,568	456,682	假设依据当前公司资本开支指引
CapexYOY	2%	63%	74%	97%	30%	30%	25%	20%	
Capex中AI投资占比	39%	74%	90%	90%	90%	90%	90%	90%	
CapexAI增量	12,685	38,735	81,958	162,136	210,776	274,009	342,511	411,014	24Q3:技术基础设施投资里60%是服务器(GPU+TPU服务器), 40%是数据中心建设和网络连接设备支出。
年中GenAI投入总量	5,564	34,346	94,693	216,740	403,196	645,589	953,849	1,330,611	
折旧 (综合折旧年限为9年)	714	3,889	10,521	24,082	44,800	71,732	105,983	147,846	
运维成本	0	3,084	7,265	15,838	32,404	52,322	76,595	104,978	假设资本开支投出一年后开始使用, 包含电力、租金等, 数据中心每年运维成本为总投入capex的12% 谷歌数据中心中GPU与TPU算力比例约为6:4。其中GPU70%对外
总成本	714	6,973	17,787	39,920	77,204	124,054	182,578	252,824	
云业务 (外部) 所用算力占比	21%	25%	33%	35%	38%	40%	41%	42%	
谷歌云收入	33,088	43,229	58,705	93,928	143,710	212,691	304,148	425,807	假设依据: IDC预测28年全球公有云约1.7W亿美元, 假设谷歌份额约9%(1500亿美元)、当前6%。
YOY	26%	31%	36%	60%	53%	48%	43%	40%	
AI云收入占比	11%	18%	27%	42%	55%	71%	84%	93%	
AI云收入	3,592	7,798	15,752	39,829	79,225	150,528	255,897	396,640	考虑TPU成本优势, 假设依据算力成本下降、规模效应提升。 Anthropic将合作接入多达100万个谷歌的TPU, 价值数百亿
AI云收入YoY	103%	117%	102%	153%	99%	90%	70%	55%	
TPU收入	2,109	3,832	6,220	16,907	30,433	51,735	85,363	136,582	
YoY	56%	82%	52%	66%	80%	70%	65%	60%	考虑TPU成本优势, 假设依据算力成本下降、规模效应提升。 Anthropic将合作接入多达100万个谷歌的TPU, 价值数百亿
TPU收入占比云收入	6%	9%	11%	18%	21%	24%	28%	32%	
云营业利润率	5%	14%	23%	36%	40%	43%	45%	45%	
AI云营业利润率		10%	25%	38%	43%	45%	48%	48%	
(AI云营业利润+折旧) /当年AI投入			27%	42%	64%	88%	118%	146%	大于100%则当年现金流正贡献
ROIC (AI云营业利润/累计AI投入)			14%	22%	25%	30%	35%	38%	资本回报率: 38% 2.5年收回投资

资料来源：公司财报会、彭博、semianalysis，国信证券经济研究所测算整理测算*口径并未包括对传统云业务、内部其他业务自用测算

亚马逊AI云业务投资与回报ROI测算



- 当前现金流端ROI小于1，后伴随着AI云规模增长与算力成本下降、利润率提升，2030年现金流贡献转正。资本回报端持续提升，2030年ROIC可达到24%，即4年可收回投资周期。

表：亚马逊云AI投资与ROI测算

百万美元/%	2023	2024	2025	2026E	2027E	2028E	2029E	2030E	备注
Capex (含融资租赁)	48,133	77,658	128,320	200,179	260,233	325,291	390,349	468,419	假设依据当前公司资本开支指引
CapexYOY	2%	61%	65%	56%	30%	25%	20%	20%	
Capex中AI投资占比	35%	54%	72%	78%	85%	85%	85%	85%	
CapexAI增量	16,905	41,801	92,246	156,140	221,198	276,498	331,797	398,156	假设资本开支投出一年后开始使用，包含电力、租金等，数据中心每年运维成本为总投入capex的12%
年中GenAI投入总量	9,304	37,224	104,247	228,440	417,109	665,957	970,104	1,335,081	
折旧（综合折旧年限为9年）	1,095	4,342	11,583	25,382	46,345	73,995	107,789	148,342	
运维成本	0	4,732	9,227	18,907	34,598	55,580	79,880	106,761	谷歌数据中心中GPU与TPU算力比例约为6:4。其中GPU70%对外
总成本	1,095	9,075	20,810	44,289	80,943	129,575	187,670	255,103	
云业务（外部）所用算力占比	21%	25%	33%	35%	38%	40%	41%	42%	
亚马逊云收入	90,757	107,556	128,725	167,343	224,239	302,723	402,621	523,407	假设依据：IDC预测28年全球公有云约1.7W亿美元
YOY	26%	19%	20%	30%	34%	35%	33%	30%	
AI云收入占比	4%	7%	14%	19%	24%	29%	35%	47%	
AI云收入	3,228	7,600	17,806	31,198	54,551	88,121	140,333	244,179	考虑Trainium成本优势，假设依据算力成本下降、规模效应提升。
AI云收入YoY		135%	134%	75%	75%	62%	59%	74%	
Trainium收入	511	1,044	3,430	8,088	14,340	23,661	37,858	58,679	
YoY	41%	104%	229%	136%	77%	65%	60%	55%	Anthropic将合作接入多达100万个谷歌的TPU，价值数百亿
TPU收入占比云收入	1%	1%	3%	5%	6%	8%	9%	11%	
云营业利润率	5%	14%	23%	36%	40%	43%	45%	45%	
AI云营业利润率		10%	25%	38%	43%	45%	48%	48%	
(AI云营业利润+折旧) / 当年AI投入			27%	38%	49%	63%	82%	107%	大于100%则当年现金流正贡献
ROIC (AI云营业利润/累计AI投入)			15%	17%	17%	17%	19%	24%	

资料来源：公司财报会、彭博、semianalysis，国信证券经济研究所测算整理测算*口径并未包括对传统云业务、内部其他业务自用测算

[01] 资本开支水位与可行性区间

[02] 科技大厂AI投资的ROI测算

[03] 大模型厂商财务模型与ROI

[04] 风险提示

OpenAI与Anthropic模型厂商对比

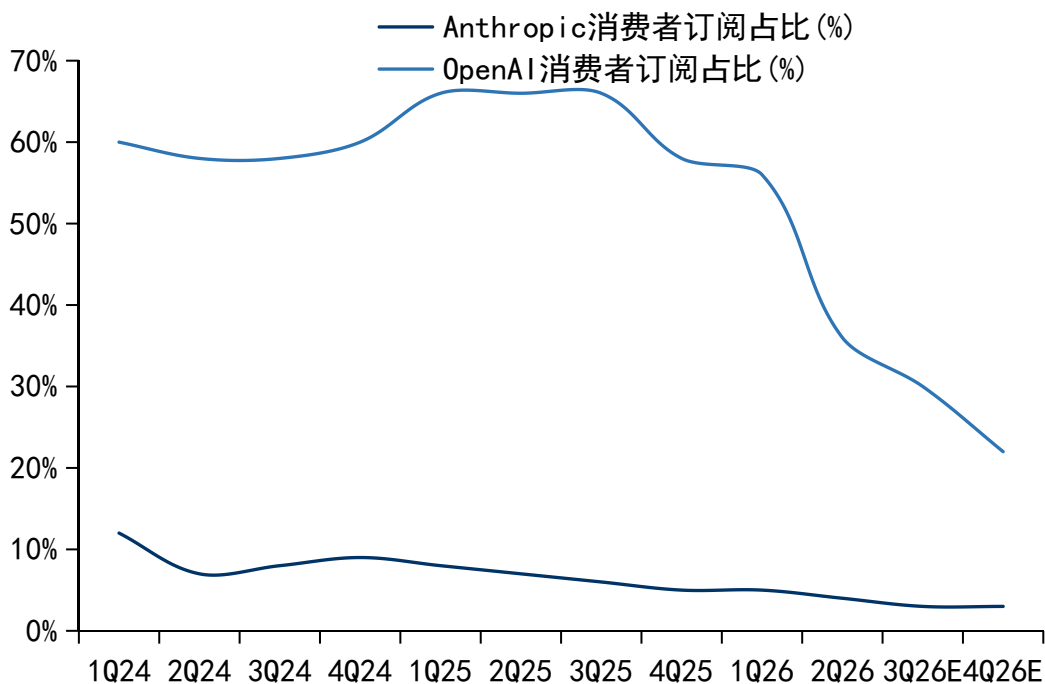
表：OpenAI与Anthropic模型厂商对比

对比维度	OpenAI	Claude（Anthropic）
最新估值	8520 亿美元（2026 年 3 月，融资后）	9000 亿美元（2026 年 5 月，拟融资目标）
融资规模	1220 亿美元（2026 年 3 月，史上最大）	拟融资 500 亿美元（IPO 前最后一轮）
核心投资方	亚马逊、英伟达、软银、微软	谷歌、亚马逊等
年化收入（ARR）	约420 亿美元（2026 年6月）	约690亿美元（2026 年6月）
收入构成	26Q1订阅收入占比65%，其余为API、企业服务、广告	26Q1订阅收入占比15%，其余主要为API业务
增速	2026年收入同比 3 倍增长	2026年收入同比约15 倍增长
用户结构	9.5亿月活，付费率 6%（C 端为主）	6000万月活（不含第三方嵌入和企业调用），付费率9%（企业客户价值高）
旗舰模型	GPT-5.5（多模态）	Claude Opus 4.6（长上下文，低幻觉）
核心优势	多模态能力、C 端用户规模、生态整合	长上下文（1M token）、企业级安全、高毛利
管理层状态	COO/CMO/AGI 负责人全换	稳定
毛利率	26Q1约 60%	70%+（2026 年 5 月）
IPO 计划	2027年 Q1，纳斯达克上市	2026 年 10 月，纳斯达克上市

资料来源：华尔街新闻、彭博、semianalysis，国信证券经济研究所整理

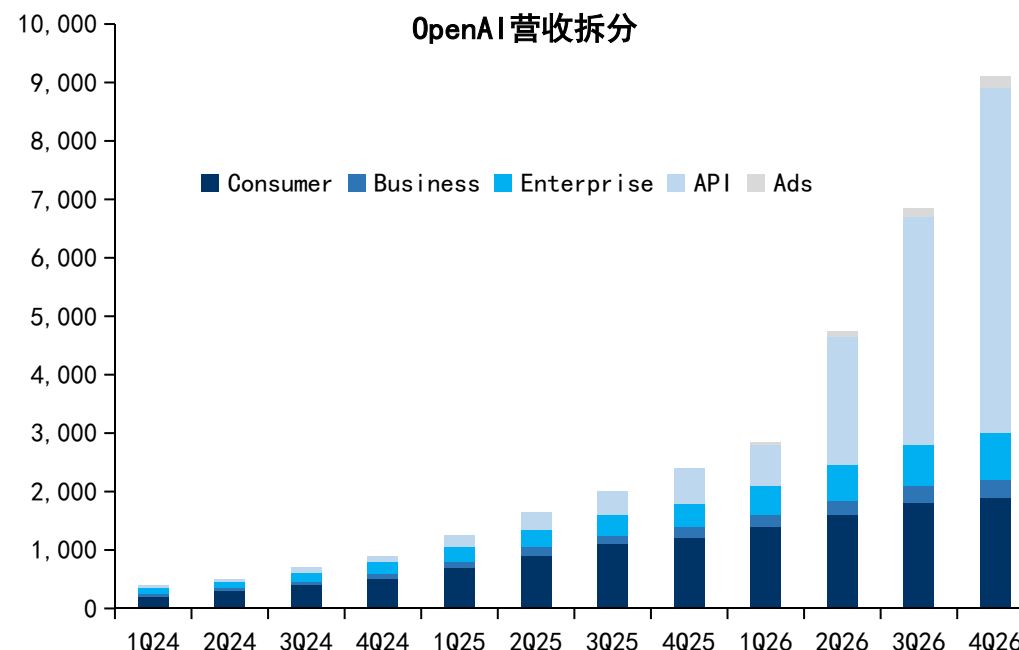
- 订阅业务分为消费者订阅和企业订阅，我们测算Anthropic消费者订阅占比总ARR的15%，而截止26Q1 OpenAI超65%收入源于消费者订阅。
- Anthropic月活跃免费用户约为6000万左右（仅包括C端网页+App MAU），其中约9%付费率。OpenAI从2025年底以来周活跃用户（WAU）一直维持9.5亿左右，付费率约为6%。

图：OpenAI和AnthropicARR变化情况（亿美元）



资料来源：openai、Anthropic、Semianalyse，国信证券经济研究所整理

图：OpenAI和AnthropicARR变化情况（亿美元）

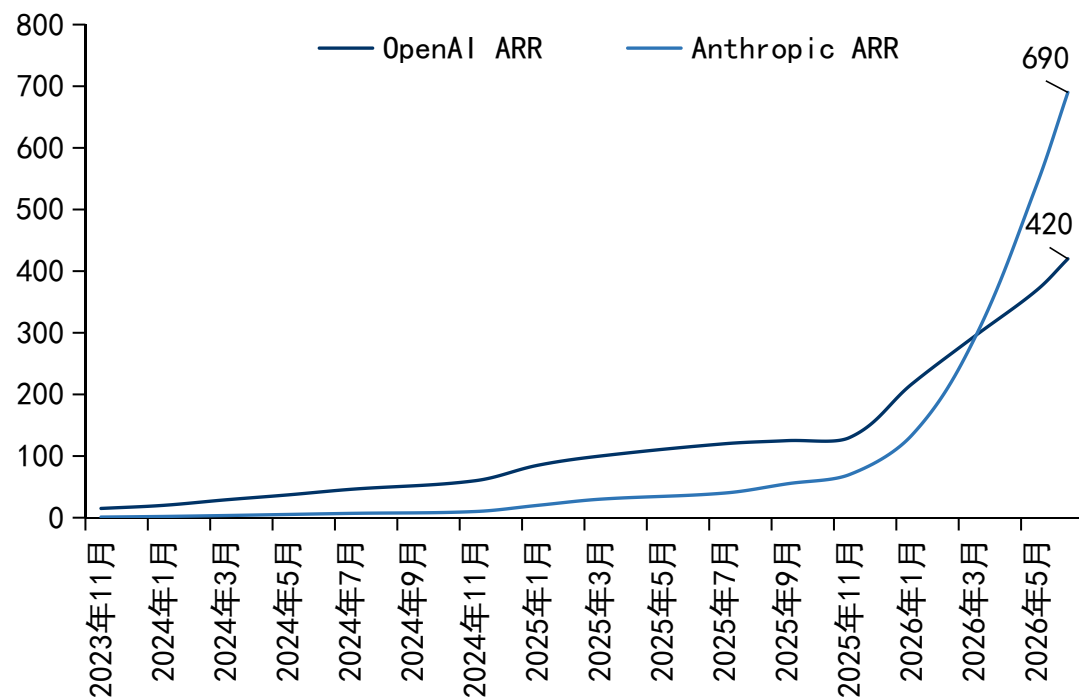


资料来源：openai、Anthropic、Semianalyse，国信证券经济研究所整理

Anthropic或实现单季度盈利，模型厂商迎来商业化闭环

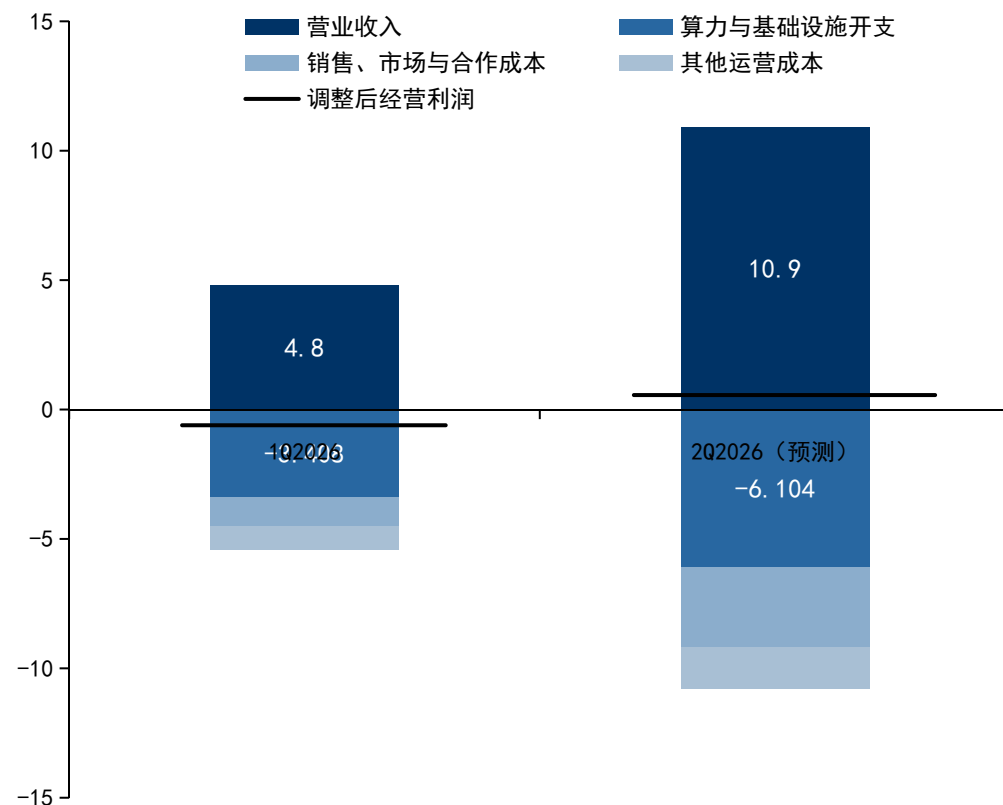
- 26年6月Anthropic ARR预测到690亿（OAI约420亿），26Q2有望实现盈利。根据Theinformation和WSJ的数据，26Q1Anthropic收入48亿美元，OpenAI 57亿美元，但Q2预计Anthropic将达到109亿，并且实现单季度OP盈利（5.6亿）、OPM约5%，而根据公司官网，5月初其ARR已达到470亿美元。

图：OpenAI和AnthropicARR变化情况（亿美元）



资料来源：openai、Anthropic、yipidata，国信证券经济研究所整理

图：WSJ预计Anthropic或将实现单季度经营利润转正（十亿美金）



资料来源：WSJ、国信证券经济研究所整理

Anthropic与OpenAI全年收入有望达1200亿美元



- Anthropic与OpenAI全年收入有望达1200亿美元，全年收入增速分别约15倍和3倍，且两者毛利率已处在60%左右。
- 据TheInformation报道OpenAI预计2030年营收达2000亿美元，2030年起开始产生正向现金流。

表：Anthropic、OpenAI公司财务数据变化

百万美元	Anthropic			OpenAI		
	2024	2025	2026E	2024	2025	2026E
总收入	524	4,129	66,722	3,600	12,905	47,508
YOY	500%	688%	1516%	180%	258%	268%
年末ARR	1,000	7,000	100,000	6,000	13,000	70,000
YoY	1000%	600%	1329%	400%	117%	438%
总运营成本	1,016	2,561	24,144	2,061	8,400	22,640
付费用户推理成本	1,016	2,437	22,870	1,563	4,886	15,692
免费用户推理成本	-	124	1,274	498	3,514	6,948
总毛利	-492	1568	42578	1539	4505	24868
总毛利率	-94%	38%	64%	43%	35%	52%
付费推理毛利	-492	1692	43852	1898	8020	31817
付费推理毛利率	-94%	41%	66%	53%	62%	67%
模型训练成本	1625	4000	24996	2000	7500	32000
其他运营支出	900	2650	15000	4300	12000	18000
运营利润	-3,017	-5,082	2,582	-4,761	-14,995	-25,132
运营利润率	-576%	-123%	4%	-132%	-116%	-53%

数据来源：Exclusive、Semianalyse，公司官网、新浪新闻，国信证券经济研究所整理测算

Anthropic、OpenAI：2026年千亿美金算力支出

- 26年两家头部模型厂商年度算力支出超千亿美金、同比增速5倍以上，相当于北美四家云厂（亚马逊云、微软云、谷歌云、Oracle）2026年云收入（5000亿美金）的20%。其中Anthropic 490亿美金（240亿推理、250亿训练），OpenAI 550亿美金（230亿推理、320亿训练），成为支撑科技厂商资本开支的核心动力。2026年两家头部模型厂商占用超12GW算力，26年末预计超17GW。

表：Anthropic、OpenAI公司算力成本与变现

百万美元	Anthropic			OpenAI			备注
	2024	2025	2026E	2024	2025	2026E	
总收入	524	4,129	66,722	3,600	12,905	47,508	
总算力成本	2,641	6,561	49,140	4,061	15,900	54,640	26年年度算力支出超千亿美金
YOY		148%	649%		292%	244%	
推理算力成本	1,016	2,561	24,144	2,061	8,400	22,640	
模型训练成本	1,625	4,000	24,996	2,000	7,500	32,000	
总算力规模 (MW)	110	779	5963	169	1887	6630	26年占用约12GW算力
推理算力规模	42	304	2930	86	997	2747	
训练算力规模	68	475	3033	83	890	3883	
每MW推理算力年度收入	12	14	23	42	13	17	
YoY		10%	68%		-69%	34%	
每MW推理算力成本	24.0	8.4	8.2	24	8.4	8.2	每MW H100算力租赁约2千万美金/年
推理算力毛利率		38%	64%	43%	35%	52%	
推理算力经营利润率	-297%	-198%	11%	-231%	-179%	-111%	

数据来源：Exclusive、Semianalyse，公司官网、新浪新闻，国信证券经济研究所整理测算

请务必阅读正文之后的免责声明及其项下所有内容

模型公司加大算力采购，与多家云厂商合作

- AI时代，算力已成为企业创造价值的核心生产要素。黄仁勋认为，过去购买算力是为了运行软件，是支持性成本，现在算力就是生产力，算力规模直接决定token产量，进而决定企业收入。
- 26年4月开始，随着Claude用户量快速增长，社区大量反馈Max订阅也会频繁hit limit，5月在与SpaceX达成算力合作后（300MW+、22万+NVIDIA GPU新容量），立刻5小时限额翻倍，对于Pro、Max、Team及按席位计费的Enterprise计划，ClaudeCode的5小时使用限额翻倍，同时取消高峰时段限速。

图：Anthropic去年末以来加速算力建设和采购

合作方	采购/自建	具体细节
谷歌+博通	采购	23年双方开始合作，谷歌一直是Anthropic的主要算力提供方。 25年10月，双方宣布扩展1GW的计算容量，26年上线，Anthropic将获得100万个TPU芯片 26年4月，与谷歌和博通签署了一项新协议，获得5GW的TPU容量，27年开始上线
		- AWS与Anthropic合作开展Project Rainier，25年10月已投入使用，到年底采用超过50万颗Trainium2芯片进行训练和推理任务。其中仅印第安纳州项目就斥资110亿美元，发电量超过2.2GW。 26年4月，与AWS达成协议，未来10年投入1000亿美元，确保5GW的容量，投入涵盖Graviton和Trainium2至Trainium4芯片，包括到26年底上线的接近1GW的Trainium2和Trainium3
亚马逊	采购	
微软+英伟达	采购	25年11月，Anthropic宣布采购300亿美元Azure计算容量，由英伟达Blackwell和Rubin提供支持，购买算力容量上限达到 1GW ，以获得英伟达和微软150亿美元的投资
Fluidstack	自建	2025年11月，Anthropic宣布将投资500亿美元，与Fluidstack合作在德克萨斯州和纽约州建设数据中心
Coreweave	采购	26年4月，达成多年期协议，今年晚些时候开始启用计算服务
SpaceX	采购	26年5月，与SpaceX达成协议，使用其Colossus1数据中心的所有计算能力，一个月内获得 300MW+ ，22万+NVIDIA GPU

资料来源：Anthropic、国信证券经济研究所整理

图：OpenAI数据中心建设计划及算力采购计划

时间	地点	负责方与容量	建设说明
2025-2026	德克萨斯州阿比林市	Oracle负责，目前0.7GW，后续可达1.4GW	两栋建筑完工，每栋约60,000个GB200NVL72服务器
2025-2026	俄亥俄州洛兹敦、德克萨斯州米拉姆县	软银负责，未来18个月内可出售1.5GW	VeraRubin芯片
2026-	阿联酋的StargateUAE、挪威的Stargate等	阿联酋2026年上线200MW，后续计划规模5GW	星际之门阿联酋
2026-2027	德克萨斯州沙克尔福德县、新墨西哥州多纳阿纳县、美国中西部	Oracle负责，超4GW	
总计	原计划未来四年内Stargate计划容量达到近10吉瓦（GW），投资超5000亿美元；最新26年4月已经超过10GW规模 英伟达投资1000亿美元锁定未来10GW数据中心的芯片供应权。 AMD授权最多1.6亿股权（AMD的10%）绑定未来6GW芯片供应权。		
AWS	25年11月，采购380亿美元的英伟达GPU，26年底部署完成 26年2月，新增8年1000亿美元，约2GW AWS Trainium容量，覆盖Trainium3和下一代Trainium4		

数据来源：公司官网、华尔街新闻、彭博新闻、国信证券经济研究所整理

- 根据TrendForce，截至2026H1，Hopper全系列（H100/H200/GH200）累计全球出货总量突破120万颗，测算对应数据中心约60GW负载。
- 根据2026年GTC开发者大会，截至2026年3月，Blackwell全系列（B100/B200/GB200/GB300）已锁定客户长期订单合计150万颗GPU，截止26H1约交付十万量级，对应约10GW，全部交付预计可达100GW。

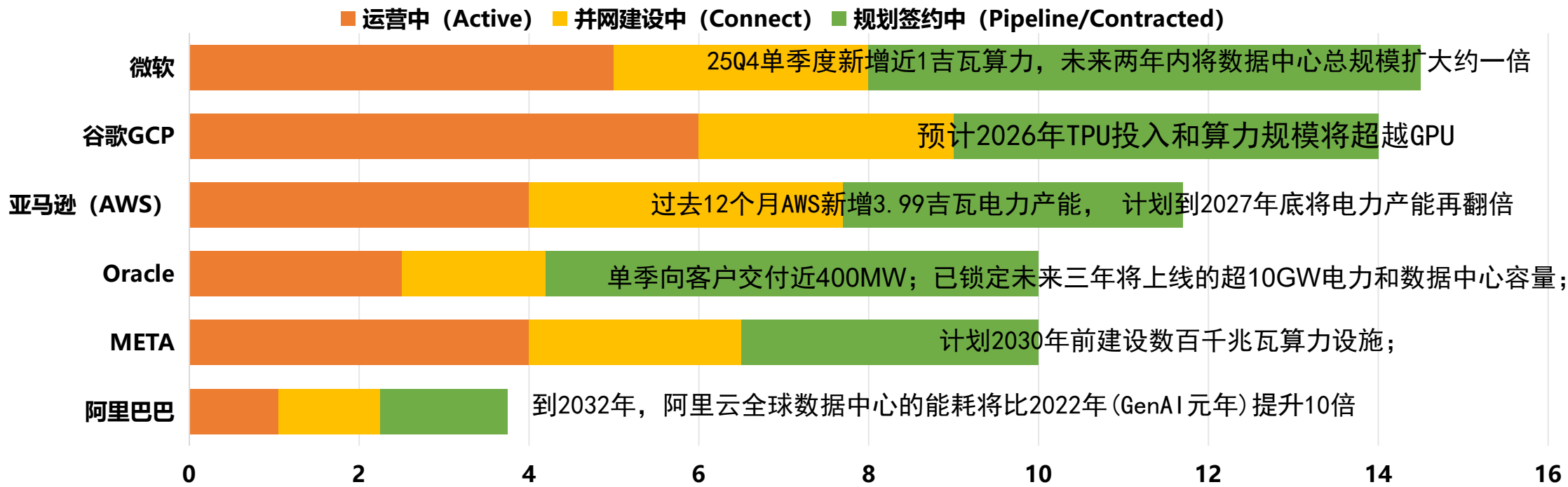
图：全球加速器芯片出货量

加速器/千颗*	2024	2025	2026E	2027E	2028E	2029E	2030E
GPU	5,234	6,110	9,570	11,500	13,300	14,600	16,000
NVDA	4,850	5,700	9,000	10,500	12,000	13,000	14,000
AMD	384	410	570	1,000	1,300	1,600	2,000
ASIC	2,130	2,850	7,700	15,900	35,200	39,500	43,000
TPU	1,280	1,700	5,000	12,000	30,000	33,000	35,000
Trainium	750	750	1,700	2,400	3,200	4,000	5,000
其他ASIC	100	400	1,000	1,500	2,000	2,500	3,000
Total	7,364	8,960	17,270	27,400	48,500	54,100	59,000
YoY		22%	93%	59%	77%	12%	9%
占比							
GPU	71%	68%	55%	42%	27%	27%	27%
NVDA	66%	64%	52%	38%	25%	24%	24%
AMD	5%	5%	3%	4%	3%	3%	3%
ASIC	29%	32%	45%	58%	73%	73%	73%
TPU	17%	19%	29%	44%	62%	61%	59%
Trainium	10%	8%	10%	9%	7%	7%	8%
其他ASIC	1%	4%	6%	5%	4%	5%	5%

资料来源：各公司官网、Semianalyse，国信证券经济研究所整理测算

- 2026年头部三家CSP厂商拥有GW数将来到10GW量级，过去以约每个季度1GW速度增长，预计在2027年底各家规模来到15GW左右。
- 将已完成主体建设、进入电网调试阶段的项目为“并网建设中”，仅将已正式投运、可对外提供算力服务的项目划入“运营中”，将已经合同计划假设的数据中心归为“规划签约中”、基本预计在2027年上线；

图：AI 数据中心功率增长计划（GW）

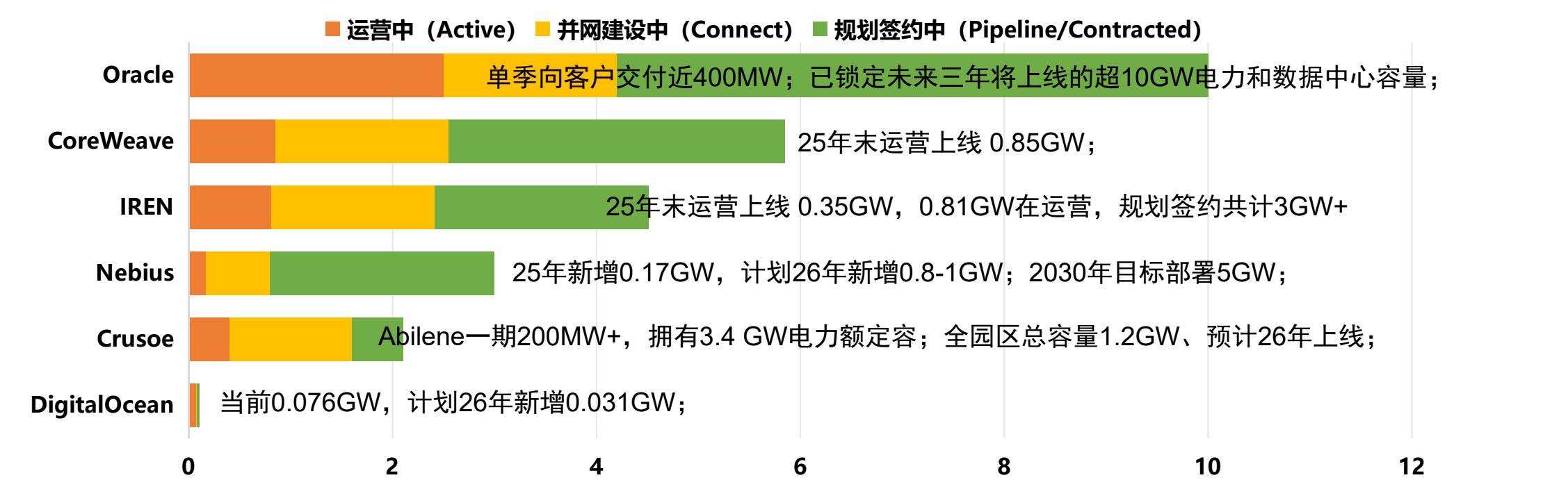


资料来源：各公司官网与业绩会、彭博社、Uptime Institute、华尔街新闻，国信证券经济研究所整理

数据中心建设维度：宏大的中长期规划

- Oracle的基础设施布局最为宏大。26FYQ3单季向客户交付近400MW、最大数据中心达800MW。此外，Oracle还在中东阿布扎比部署了首个OCI超级集群，Stargate总能力提升至7GW口径，已锁定未来三年将上线的超10GW电力和数据中心容量。
- CoreWeave在全球运营43个数据中心，25年底总容量达到850MW，FY26年末计划部署容量达到1.7GW，FY30目标达到8GW。
- Nebius正在快速扩张，截至2025年底，公司在全球拥有16个站点，当年新增0.17GW、超过0.1GW的目标，公司正在建设9个新站点。
- Crusoe的布局聚焦于能源优势，当前电力容量达3.4GW。公司的旗舰项目是德州Abilene的1.2GW数据中心园区，一期容量超过200MW；二期总容量达1.2GW，预计2026年中期通电。

图：数据中心功率增长计划（GW）



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

大模型市场空间：短期10万亿美金市场，收入可达一万亿美金

- Anthropic等大模型短期可以解决/替代知识工作者部分自动化任务，合计约10万亿美金场景。考虑大到企业不会把省下的人力成本全部付给AI厂商，根据a16z《2026企业AI落地实效白皮书》**假设行业付费渗透10%，收入可达一万亿美金。**

表：大模型市场空间

\$120万亿-全球2025年名义GDP规模约为120万亿美元

\$70万亿-全球劳动力市场规模约为70万亿美元

\$40万亿-知识工作者的年度薪酬总额约为40万亿美元
≈白领/知识工作者约10亿人，假设年收入4万美元（占全球70%劳动总收入）

\$10万亿-全球白领平均只有25%的工时可短期自动化，剩余 75% 工时仍需要人类在岗配合、校验、决策。

\$2万亿-编程市场空间2万亿美元
≈3000万人，假设年收入7万美元

【 01 】 资本开支水位与可行性区间

【 02 】 科技大厂AI投资的ROI测算

【 03 】 大模型厂商财务模型与ROI

【 04 】 风险提示

- ① **AI商业化落地不及预期，算力ROI长期承压：**当前巨头持续加码算力资本开支，但AI付费转化、企业客户付费渗透速度低于乐观测算，算力租赁价格存在下行压力，折旧、电力、运维刚性成本高企，或导致云厂商AI业务长期无法实现正向现金流，资本回报周期拉长。
- ② **高额发债推升融资成本，利息支出侵蚀利润：**2026年科技巨头发债规模创历史新高，市场认购热度边际走弱，融资利率上行；Meta、亚马逊等企业债务规模持续扩张，每年利息支出大幅增加，对冲AI业务增量收益。
- ③ **地缘与供应链约束，芯片交付不及规划：**英伟达高端GPU、AMD定制芯片存在出口管制与产能约束，厂商算力采购交付进度存在不确定性；自研芯片量产爬坡进度慢于预期，无法有效对冲外部芯片采购成本与供给缺口。
- ④ **芯片快速迭代引发大额资产减值风险：**AI芯片迭代周期缩短至2-3年，Blackwell、Rubin新架构性能跨越式提升，上代H100/H200硬件快速贬值；头部企业数千亿GPU资产存在大额减值计提压力，压制企业利润水平。
- ⑤ **算力供给过剩，行业陷入价格内卷：**全球云厂商、第三方算力租赁企业同步大规模扩产，中长期算力供给增速高于AI需求增速，机房利用率下滑，算力租赁价格持续下行，削弱算力租赁业务毛利率。

国信证券投资评级

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

分析师承诺

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

重要声明

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

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