国金证券 SINOLINK SECURITIES

市场又低估了 AI 计算机行业研究

2026 年 02 月 07 日

买入（维持评级）

行业点评

证券研究报告

计算机组
分析师：刘高畅（执业 S1130525120005）
liugaochang@gjzq.com.cn

市场又低估了 AI

行业观点：

不要低估科技大厂对 AI 的认知。科技大厂正以“投入不足风险远大于过剩”的认知坚定加码 AI：①业绩方面，2025 年谷歌营收超 4000 亿美元，云业务增速达 48%，Gemini App 月活突破 7.5 亿，证明 AI 进入商业化加速期；② 军备竞赛升级，亚马逊、谷歌、Meta、微软 2026 年总资本开支预计高达 6500 亿美元，谷歌计划约同比翻倍至 1800 亿美元；③融资环境回暖，甲骨文创纪录的 250 亿美元发债获超额认购，标志全球 AI 投资情绪企稳，建设资金充裕。

模型加速迭代，AI 手机驱动入口之战。模型技术与终端应用双轮驱动，AI 手机正通过 Agent 技术重塑流量入口。交互方式从触控转向自然语言指令，行业分化为 API Agent（如苹果）与 GUI Agent（如豆包）两条路线，后者通过视觉模拟人手操作打通应用壁垒，不仅重新定义了移动交互，更引发了巨头对这一超级终端入口的激烈争夺。

AI 应用，2026 年从“选修”到“必修”。2026 年，AI 应用正迎来宏观产业逻辑与微观业绩拐点的双重共振。一方面，行业基本面已于 2025H2 确立拐点，利润弹性显著释放。Wind 软件指数 2025Q3 行业单季营收同比+1.61%，修复通道开启；同期归母净利润 3.77 亿元，同比激增 244.56%。利润增速远超营收增速的“剪刀差”有力验证了降本增效逻辑，板块已步入具备基本面支撑的右侧击球区。另一方面，算力 ROI 正面临市场审视，应用落地成为基础设施后的“必经之路”。

全球巨头抢滩，从模型到应用趋势加快。海外垂直场景价值凸显。英伟达利用 Vera Rubin 芯片深度赋能 AI 药物研发，验证“算力+药企”的 AI for Science 路径；OpenAI 与 Anthropic 相继发布医疗专用模型，精准卡位 B 端研发与 C 端服。国内生态与 Agent 成为竞争高地。阿里千问打通淘宝、支付宝等核心生态实现从对话向执行的跨越；腾讯则全线开启 AI 化，商业侧利用大模型显著提升广告 ROI，体验侧通过 QQ 浏览器“AI 小窗”与游戏 AI 队友实现交互与情感价值供给。

聚焦四类核心应用，超级入口、AI infra、高增长、高壁垒。：①超级入口：大模型量收共振，流量枢纽地位确立。大模型已进化为最具统治力的流量入口，全球领军 OpenAI、Google 及 Anthropic 商业化加速。②AI Infra：软件定义算力，锁定“卖铲子”的确定性收益。Databricks 和 Snowflake AI 验证企业侧在数据治理与算力调度上的强烈付费意愿。③高增长：技术升维推动营销与漫剧率先落地。④高壁垒：数据流与工作流铸盾，垂直场景张力极强。具备深厚 Know-how 与专有数据（如 Palantir）天然免疫通用模型吞噬。AI 医疗蚂蚁“阿福”与 OpenAI Health 验证了 C 端全链路服务刚需。

AI 超级时代，存储、CPU、FAB 加速成长通道。①存储：AI 驱动超级周期，供需结构性失衡。AI 服务器对 DRAM 和 NAND 的需求分别是普通服务器的 8 倍和 3 倍，叠加 DDR5 和 HBM 的高速渗透，推动全球存储市场规模创下历史新高。② CPU： 随着 AI 从 Chatbot 向 Agent 演进，复杂的逻辑编排、操作系统调度压力以及长上下文带来的 KV Cache 卸载需求，使得 CPU 成为继 GPU 后的算力新短板，CPU 有望进入量价齐升阶段。③FAB：全球晶圆代工产业正通过“先进制程扩产+封装技术突破”双轮驱动，为 AI 行业的爆发式增长提供系统性的供给保障。

投资建议：

海外算力/存储：中际旭创、新易盛、兆易创新、大普微、天孚通信、源杰科技、胜宏科技、景旺电子、英维克等；Lumentun、闪迪、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。**国内算力：**寒武纪、东阳光、海光信息、协创数据、华丰科技、星环科技、网宿科技、首都在线、神州数码、百度集团、大位科技、润建股份、中芯国际、华虹半导体、中科曙光、润泽科技、浪潮信息、东山精密、亿田智能、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、欧陆通、杰创智能。**CPU：**海光信息、中科曙光、澜起科技、禾盛新材、中国长城、龙芯中科、兴森科技、深南电路、宏和科技、广合科技。**AI 应用：**1）超级入口：阿里巴巴、腾讯控股、Minimax、智谱、科大讯飞。2）美年健康、德才股份、中控技术、星环科技、卓易信息、昆仑万维等 AI INFRA&高增长&高壁垒。**其他：**空天时代、具身智能等。

风险提示

行业竞争加剧的风险；技术迭代不及预期的风险；特定行业下游资本开支周期性波动的风险。

一、不要低估科技大厂对 AI 的认知

谷歌财报炸裂，AI 需求极度旺盛，AI 投入极其坚决。2026 年 2 月 5 日，Alphabet（谷歌母公司）发布炸裂财报。4Q2025 营收 1138 亿美元（同比+17%），净利润 345 亿美元（同比+30%），2025 年全年营收达 4030 亿美元（同比+15%），搜索广告与云业务成主要增长来源。

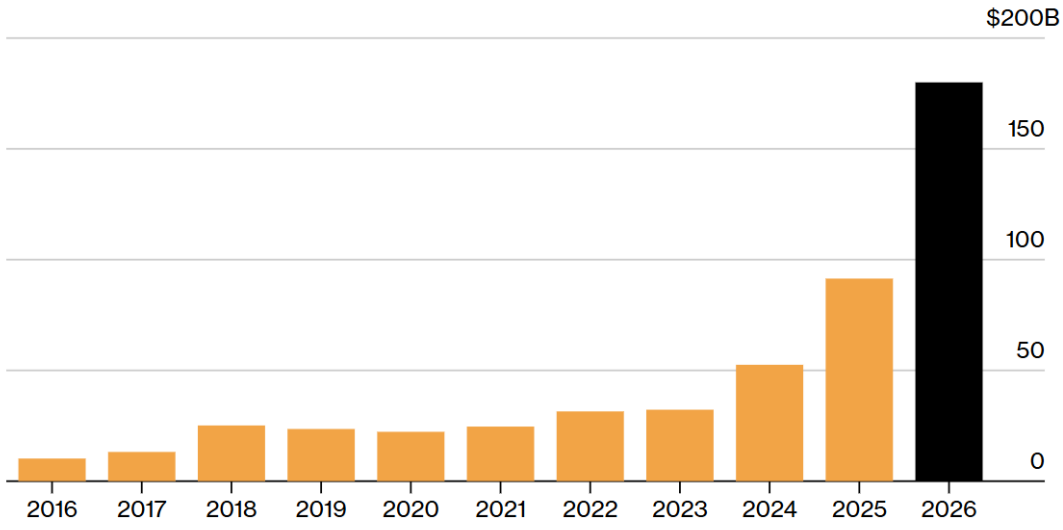
图表1：4Q2025 谷歌发布炸裂财报

来源：Google 官网，国金证券研究所

AI 需求极度旺盛，谷歌云同比增速 48%，Gemini App 月活突破 7.5 亿。

- 谷歌云高歌猛进，营收/利润/在手订单均高增：4Q2025 谷歌云同比增长 48%，营收达到 177 亿美元，高于市场预期 163 亿，盈利能力明显改善，Operating Income 从 21 亿美元提升至 53 亿美元，AI 基础设施进入利润释放阶段。截至 4Q2025，在手订单环比增长 55%至 2400 亿美元，体现出下游的 AI 需求极度旺盛。
- Gemini 月活突破 7.5 亿，C 端和 B 端两开花：C 端用户量激增，Gemini3.0 发布后，4Q2025 Gemini App 用户数量激增，月活用户已突破 7.5 亿，每日 Token 消耗量是上一代模型的 3 倍，且通过模型优化 Gemini 单位服务成本下降 78%，规模效应显著提升。B 端进入 95%主流 SaaS 公司，超过 12 万家企业使用 Gemini，95%的前 20 大和超过 80%的前 100 大 SaaS 公司都在使用 Gemini，包括 Salesforce 和 Shopify，推出仅 4 个月的 Gemini Enterprise 已售出 800 万个席位，Gemini 正成为世界上最成功软件公司的 AI 引擎。
- 搜索业务在 AI 加持下同比增长 17%。Google 搜索和其他广告收入增长 17%至 631 亿美元，实现强劲增长，所有主要垂直行业均持续增长，4Q2025 通过将 Gemini3 集成到 AI 模式和搜索，用户的日查询量翻倍，AI 模式下搜索长度是传统模式的 3 倍，近 1/6 的 AI 模式查询采用非文本模式。

图表2：Alphabet 计划 2026 年 CapEx 提升 1 倍



来源：Bloomberg，国金证券研究所

AI 投入极度坚决，预计 2026 年 CapEx 达 1800 亿美元，同比翻倍。谷歌计划将 2026 年的资本支出将在 1750 亿至 1850 亿美元之间，同比翻倍（2025 年为 914 亿），，投资将在全年逐步增加。公司管理层表示“我们在 AI 方面所做的投资已经转化为所有业务的强劲表现，强劲的业绩增长加强了必须增大 AI 投资力度的信念”。

AI 军备竞赛加剧，硅谷四大科技巨头 2026 年 CapEx 将高达 6500 亿美元。AI 军备竞赛进一步加剧，硅谷四大巨头均不愿掉队，大幅加码 2026 年资本开支，具体看：

- 亚马逊成为四家中投入规模最大的企业，将 2026 年资本支出目标定在 2000 亿美元；
- Alphabet 的资本支出计划高达 1750 亿美元-1850 亿美元，同比接近翻倍；
- Meta 预计全年资本支出将增至 1350 亿美元，同比增幅或达 87%；
- 微软同期公布其第二季度资本支出同比增长 66%，预计其截至 6 月的财年资本支出将逼近 1050 亿美元。

甲骨文创纪录发债，AI 债务“风向标”企稳。据高盛承销部门称，甲骨文(ORCL.US)创纪录的最新债券交易缓解了债务市场的紧张气氛，也为其他希望筹集数千亿美元用于数据中心基础设施建设的科技巨头提供了动力。甲骨文公司在优质债券市场筹集 250 亿美元债券，吸引了众多渴望收益的投资者，认购额超过 1290 亿美元，创下此类发行的最高纪录。在云计算巨头中信用评级最低的甲骨文公司，已成为人工智能投资的风向标，其他超大规模数据中心运营商在财报季结束后也可能很快进入市场，全球 AI 投资情绪有望弥合分歧，坚定投资未来。

我们认为不应低估全球科技巨头对 AI 的认知，AI 投入并非仅是军备竞赛，以谷歌为代表，业绩的强劲增长坚定了大幅加码 AI 投资的信心，当前全球 AI 竞争正处于跑马圈地、需求井喷的阶段，投资不足的风险是远大于投资过剩的风险。

二、模型加速迭代，AI 手机驱动入口之战

模型加速迭代，Anthropic 和 OpenAI 同时上线最新模型。2026 年 2 月 7 日，字节 Seedance2.0 重磅更新，突破“可控性”瓶颈；2026 年 2 月 6 日，Anthropic 和 OpenAI 同时上线了新模型 Claude Opus 4.6 和 GPT-5.3-Codex。

字节 Seedance2.0 重磅更新，突破“可控性”瓶颈，迈向工业级应用。字节跳动旗下多模态模型 Seedance 2.0 于今日（2 月 7 日）完成重磅版本更新，标志着字节系在 AI 视频生成领域已从“追赶”迈向“领跑”。

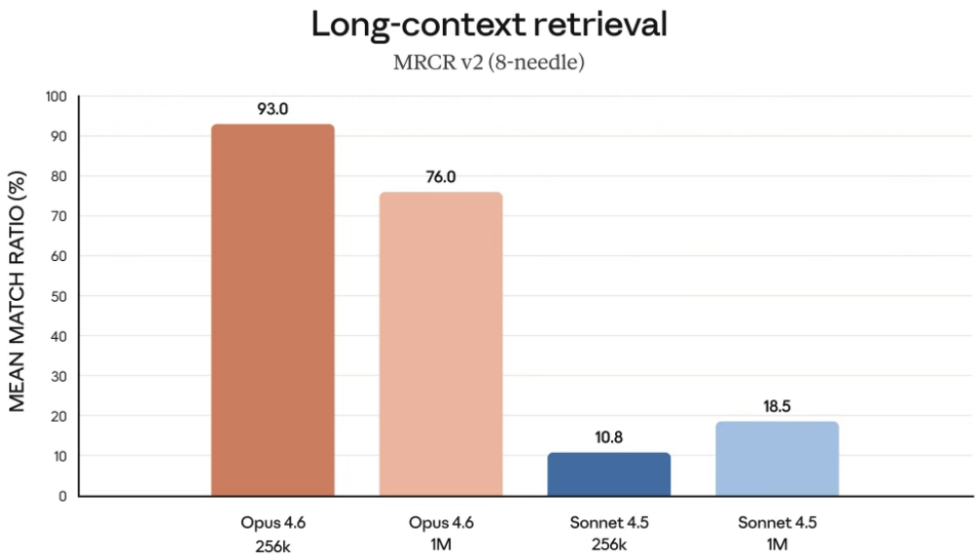
- 角色与场景的一致性质变：有效解决了前代模型在长镜头下的“脸部崩坏”和“风格漂移”问题。在多镜头连续生成中，主体特征保持高度稳定，这是 AI 视频从“短视频素材”走向“叙事性长内容”的关键门槛。
- 复杂的物理动态模拟：在处理大幅度动作（如奔跑、打斗）及光影流转时，物理规律的遵循度显著提升，画面流畅度与真实感逼近实拍效果，大幅减少了“AI 抽搐”现象。
- 语义理解与指令跟随：模型对复杂 Prompt 的解析能力增强，能够精准还原剧本中的情绪

氛围与分镜构图，极大降低了创作者的“抽卡”试错成本。

Anthropic 发布 Claude Opus 4.6：搭载处于 Beta 阶段的 1M 上下文窗口。

- Anthropic 发布了 Claude Opus 4.6，长文本能力提升显著。Claude Opus 4.6 亮点在于 Beta 阶段的 1M 上下文窗口，过去的模型在处理极长文本时，经常出现上下文腐化的问题，即模型性能随着文本长度增加而显著下降，导致它忘记或者混淆较早之前的信息。 Claude Opus 4.6 在著名的大海捞针基准测试 MRCR v2 上，成绩达到了 76%，远超前代 Sonnet 4.5 的 18.5%，这证明它能够真正有效地利用超长下文，在海量文档中精准定位并提取被深埋的关键信息，从而胜任大型代码库分析、多篇论文综述、跨会话长程任务规划等场景。
- 在推理能力和编码能力上，Claude Opus 4.6 在多项权威基准测试中确立了行业领先地位，特别是在需要自主规划和多步执行的智能体编码任务上。比如，在 Terminal-Bench 2.0 中它的测试成绩排名第一。
- 在应用层，Claude 正在深度融入生产力工具链。Claude Code 引入了智能体团体的研究预览功能，允许创建多个协同工作的 AI 智能体来并行处理任务，例如同时对代码库的不同模块进行审查。对于更广泛的办公场景，Claude in Excel 和全新推出的 Claude in PowerPoint 研究预览版，将模型的推理能力和生成能力直接嵌入到电子表格和幻灯片制作中。

图表3：Anthropic 发布 Claude Opus 4.6，长文本能力提升

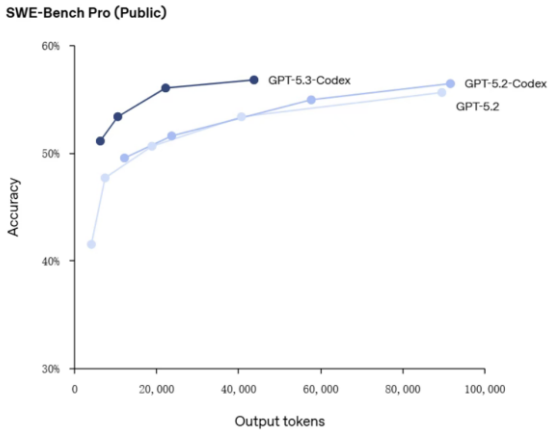


来源：Anthropic，国金证券研究所

OpenAI 推出 GPT-5.3-Codex：可能是目前最强大的智能体编码模型。

- GPT-5.3-Codex 都创下了新的行业纪录，以 77.3%的准确率大幅超越了前代模型在衡量终端编程技能的 Terminal-Bench 2.0 基准测试上的表现，并在更严格的、涵盖多语言的 SWE-Bench Pro 软件工程测评中达到了领先水平。
- GPT-5.3-Codex 的能力边界已经从纯粹的编码拓展到整个知识工作领域。在衡量真实世界职业任务的 GDPval 评估中，其表现能力与 OpenAI 的通用旗舰模型 GPT-5.2 相当。在制作金融分析 PPT、设计零售培训文档以及编写商业计划书中，GPT-5.3-Codex 能够输出专业的可使用内容。

图表4：OpenAI 发布 GPT-5.3-Codex，SWE-Bench Pro 测评中达到了领先水平



来源：OpenAI，国金证券研究所

我们认为模型的迭代还未到瓶颈，2026 年模型将会加速迭代：

- 靠 Scaling law 提升模型能力的路径依旧有效。DeepSeek V3.2 后训练规模扩展到预训练规模 10%，预计未来比例还会提高，而且 DeepSeek 团队在论文表示由于训练算力有限，DeepSeek-V3.2 的世界知识广度还是落后于 Gemini 3 pro 这样的顶尖闭源模型，团队计划未来进一步扩大预训练规模。同时 DeepSeek 大量使用合成数据有效说明不用担心数据会遇到瓶颈。
- DeepSeek-V3.2 提出的 DSA 机制展示出强大算法创新能力，不必担心大模型技术创新已经到达瓶颈。
- 大模型训练的硬件基础升级。英伟达的 Hopper 架构正在转向 Blackwell 架构，Blackwell 相比前代在单卡算力、显存带宽、显存容量、以及集群互联都大幅提升，这对大模型训练的意义一方面是加速和降低成本，另一方面是可以使用更大的 Batch Size（模型更新学习内容前一次性处理的训练样本数量）,这对训练稳定性有帮助，更大的 Batch Size 能够更准确地估计整个数据集的梯度，从而使学习过程更加稳定,而较小的批次则会产生噪声过大且特征过于明显的信号，这可能导致模型的学习路径出现不稳定的跳跃。

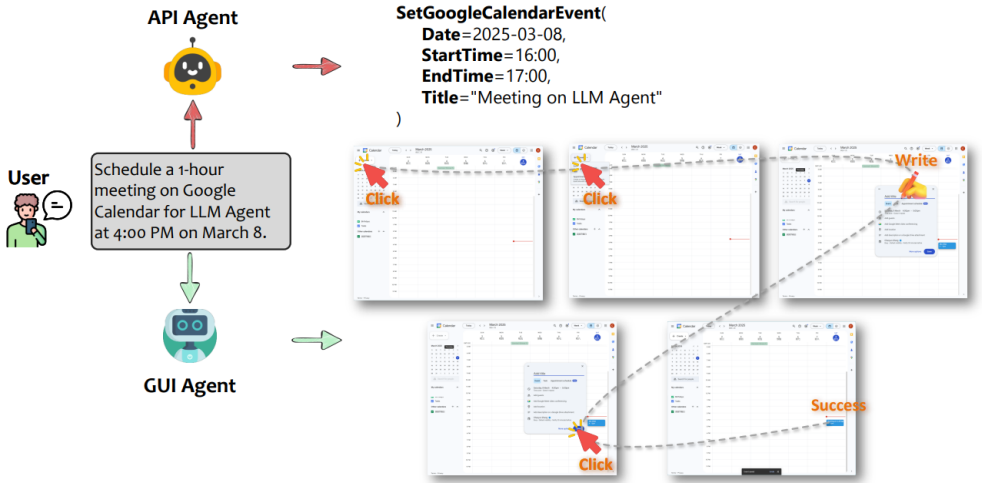
AI 手机作为超级载体，驱动大厂入口之争。手机本身是人们工作与生活信息流的集大成者，全能的 AI 手机助手出现将全面重构所有人的工作与生活。我们认为豆包手机助手是 AI 在 C 端从单一的语音助手进化为真正会行动的助理的重要里程碑，不仅能理解用户意图，还能跨应用自主执行复杂任务，体现了 AI 在真实场景中的巨大进展，也预示着手机交互方式和用户效率将被系统性重塑。

图表5：豆包手机助手执行用户任务

来源：字节跳动，界面新闻，国金证券研究所

基于图形用户界面（GUI）的触控是移动互联网时代的核心交互方式，但也存在操作流程复杂的问题以及 App 间信息孤岛的问题。如今大模型正推动移动端交互革命：用户从复杂的点击跳转中解放，只需下达自然语言指令，AI Agent 即可跨越应用边界完成任务。针对如何打破应用壁垒，行业目前形成了两套演进逻辑：API Agent 路线：通过标准化协议让开发者主动接入。如苹果通过 App Intents 框架让开发者可以将应用接入 Apple Intelligence。GUI Agent 路线：利用多模态模型模拟人类视觉和操作。这种方式无需开发者配合，凭借通用视觉能力直接操作各类 App，如豆包手机助手。

图表6：API Agent 和 GUI Agent 在完成"3 月 8 日下午 4:00 在 Google 日历上安排 1 小时会议"任务时的区别



来源：《API Agents vs. GUI Agents: Divergence and Convergence》，Chaoyun Zhang 等，国金证券研究所

在效率、隐私、可靠性等方面 API Agent 更有优势，但是 GUI Agent 的最核心优势在于灵活通用，不需要应用厂商主动适配，是打通碎片化应用生态的利器。预计模型厂商、应用厂商和手机厂商为了抢夺流量入口，会在两种路线选择上进行激烈的博弈。

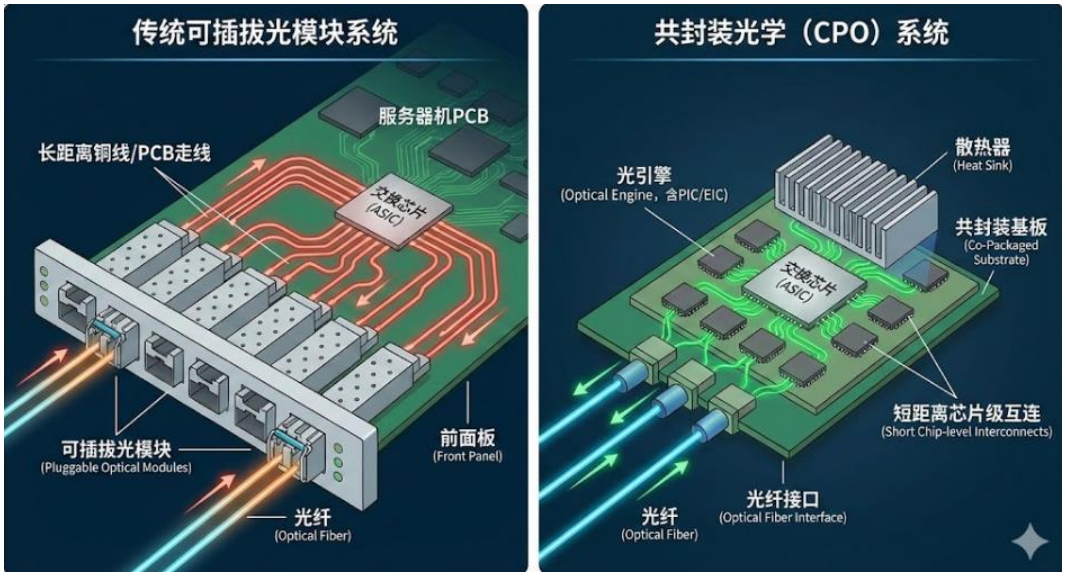
我们认为 AI 手机是其他 AI 终端爆发的前置条件，预计全球顶尖大厂将极致投入到这场新入口之战。

三、CPO 等结构创新加速，空天有望解决能源问题

1.1 CPO 开启“硅光新纪元”，Lumentum 业绩验证高景气

CPO 作为封装新形态，可有效打破“功耗墙”和“密度墙”。1）CPO（Co-Packaged Optics，光电共封装）并非单一的产品，而是一种封装形态的革命。传统的光模块是“可插拔”的（Pluggable），像 U 盘一样插在交换机面板上，数据需要经过长距离的电信号传输才能到达交换芯片（ASIC），这导致了较高的功耗和信号损耗。CPO 技术则是将光引擎（Optical Engine）与交换芯片封装在同一个基板上，尽可能缩短电信号的传输距离。2）在 2026 年的当下，AI 集群正向着万卡、十万卡规模演进，互联速率正从 800G 向 1.6T 乃至 3.2T 跨越，CPO 即可成为打破“功耗墙”和“密度墙”的新选择。

图表7：CPO 和光模块系统对比



来源：国金证券研究所整理

CPO 产业链近期显著变化，结构创新迅速。

1）硅光（SiPh）渗透率加速：硅光技术是 CPO 的基石。相比传统 EML 方案，硅光在集成度、

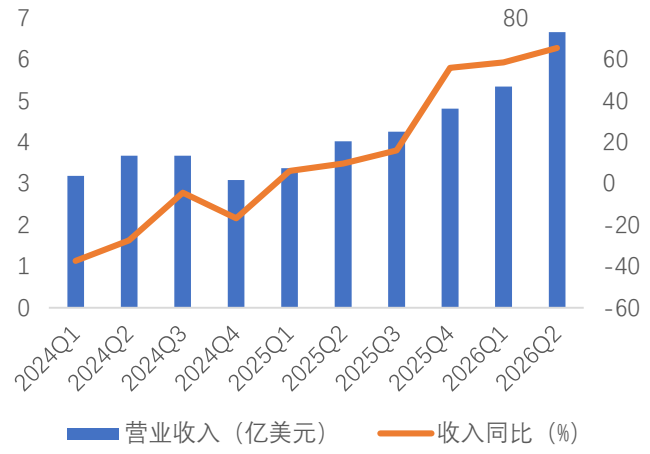
成本可控性上具备天然优势，正成为 CPO 落地的载体。

2）ELS（外置光源）标准趋于统一：CPO 最大的痛点之一是激光器（光源）不耐热，如果封装在芯片旁边容易损坏且难以更换。近期，行业内关于 ELS（External Laser Source）的模块化标准已逐步形成共识，即把“灯泡”（光源）独立出来做成可插拔，坏了可以换，而把“透镜和电路”（光引擎）封在芯片旁。这一方案的逐步成熟，有望大幅降低了 CPO 的维护成本，

3）薄膜铌酸锂（TFLN）的引入：为了进一步降低功耗并支持单波 200G/400G 的超高速率，行业开始在 CPO 中引入薄膜铌酸锂调制器，代表材料端的革新。

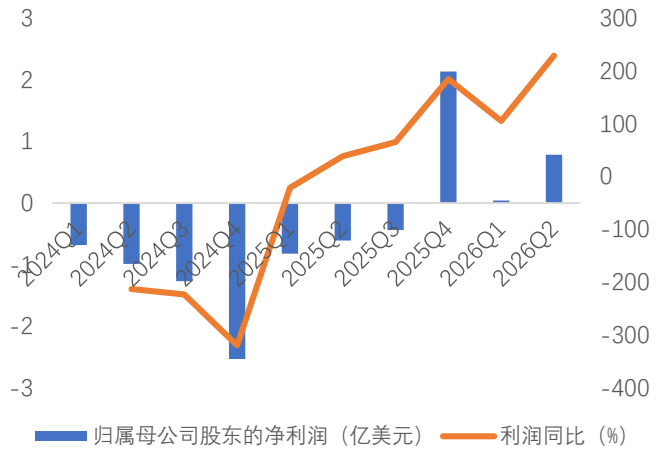
Lumentum 业绩超预期兑现，量价齐升。全球光通信芯片及器件龙头 Lumentum（LITE）于近日公布了 2026 财年第二季度（对应自然年 2025Q4）业绩。报告期内，Lumentum 实现营收 6.66 亿美元，同比增长 65.5%，环比增长 24.7%。

图表8：Lumentum 收入及同比



来源：ifind，国金证券研究所

图表9：Lumentum 利润及同比



来源：ifind，国金证券研究所

我们认为，Lumentum 超预期，预示产业趋势的三个变化：

1）1.6T 迭代速度超预期。Lumentum 指出，云服务商（Cloud Titans）对高速率产品的需求并未放缓，反而加速向 1.6T 迁移。这意味着光通信行业的景气周期被拉长，并未出现此前市场担心的“周期见顶”。

2）CW 激光器（连续波光源）需求井喷。随着 CPO 及硅光模块渗透率的提升，作为硅光方案“心脏”的大功率 CW 激光器需求激增。Lumentum 作为该领域的领军厂商，其业绩超预期印证了硅光/CPO 技术路线的放量。

3）北美云厂商 Capex 意愿依然坚挺。业绩的超预期反映了下游客户（如微软、谷歌、亚马逊等）在 AI 基础设施上的投入依然维持高位，并未受到短期宏观扰动的影响。

1.2 空天有望解决 AI 能源问题，打开每年 100GW 算力空间

算力能耗庞大，地面承载力有限。AI 大模型的参数量与能耗呈非线性增长。根据国际能源署（IEA） 发布的权威报告，2026 年全球数据中心、加密货币和人工智能的电力消耗预计将达到 1000 太瓦时（TWh），相当于日本全年的用电量。在地面建设数据中心，不仅面临严苛的 PUE（电源使用效率）限制，更受到土地与散热资源的物理约束。

商业航天降本，获取通往“太空基建”的入场券。1）太空算力商业化的前提是极低的发射成本。2024-2025 年，以 SpaceX 星舰（Starship）为代表的运载工具取得了里程碑式突破，将低轨发射成本压低至每公斤数千美元甚至更低。2）国内来看，2025 年 11 月国家航天局设立商业航天司，持续推动商业航天高质量发展，确立了其国家级战略地位。这一政策红利在 2025 年集中释放，推动了从火箭制造到卫星互联网的产业链增长。同时，中国商业航天企业如深蓝航天、蓝箭航天在可回收火箭领域的突破，使得大规模部署太空设施成为经济账上算得平的生意。

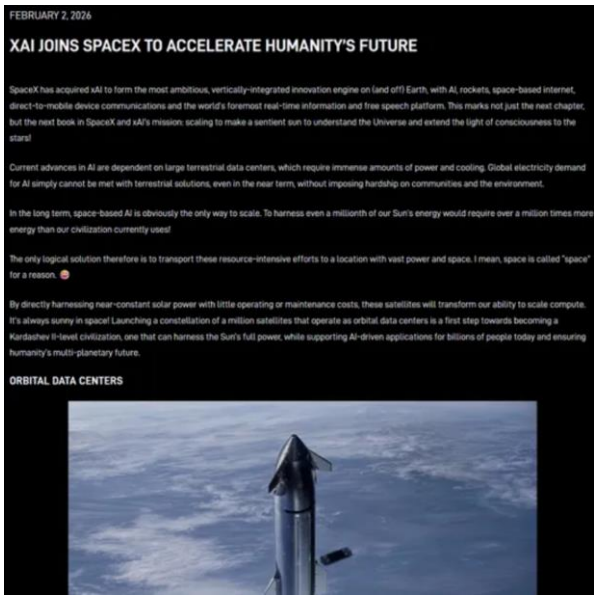
图表10：近十年我国商业航天相关企业注册量及增速

来源：企查查，国金证券研究所

太空有望解决能源问题，成为 AI 算力的“第二增长极”。1）路径一：太空边缘计算。利用太空真空环境（零散热成本）和不间断的太阳能（24/7 清洁能源），直接在轨道上部署数据中心。这不仅解决了地面的“热岛效应”，还能实现数据的全球实时分发。2）路径二：空间太阳能电站。通过在地球静止轨道部署巨大的太阳能阵列，收集能量并通过微波束传输回地球。相比地面光伏，太空光伏不受昼夜和天气影响，效率有望相较地面大幅提升。

马斯克详述太空算力路线图，预计每年新增 100GW 的 AI 计算能力。根据马斯克表述，SpaceX 计划发射 100 万颗卫星组成轨道数据中心星座，直接利用太空近乎恒定的太阳能，且运行和维护成本极低。马斯克估算，每年发射 100 万吨卫星、每吨卫星产生 100 千瓦的计算能力，将每年新增 100 吉瓦的 AI 计算能力。马斯克预计，在两到三年内，在太空中生成 AI 算力将成为成本最低的方式。这一成本效率将使企业能够以前所未有的速度和规模训练 AI 模型和处理数据。

图表11：马斯克旗下 SpaceX 在公司备忘录中确认与 xAI 合并



来源：财联社，国金证券研究所

四、“AI 吞噬软件”过于绝对，错杀带来中期配置良机

模型更新引起恐慌，全球 SaaS 板块普跌。本周，Anthropic 发布了具备“自主计算机操作”能力的 Claude 新版本，展示了 AI 直接接管复杂软件界面的能力。这一进展引发了市场对 SaaS“中间商”价值被吞噬的极度担忧。

图表12：全球代表 SaaS 公司过去一周涨跌幅

代码	简称	区间涨跌幅（%）
		2026/2/2-2026/2/6
CRM.N	Salesforce	-9.86
ADBE.O	Adobe	-8.48
NOW.N	ServiceNow	-13.90
APP.O	Applovin	-14.03
00268	金蝶国际	-14.36
02556	迈富时	-10.61
688111	金山办公	-9.86
002230	科大讯飞	-8.04
688615	合合信息	-12.82

来源：ifind，国金证券研究所

“吞噬论”过于绝对，软件应用公司的核心 knowhow 并未消失。我们认为，“AI 吞噬一切软件”的论调过于线性外推。在思科系统公司于旧金山举办的一个人工智能会议上，黄仁勋表示，人工智能系统的设计初衷是与现有软件工具协同工作，而不是完全取代它们。他直言，担心“人工智能会让软件公司变得不那么重要”的想法是错误的，人工智能将继续依赖现有的软件，而不是从头开始重建基本工具。

黄仁勋强调，人工智能领域的最新突破实际上集中在如何更有效地利用现有工具。而软件工具的设计初衷就是为了支持复杂的操作，因此，它们会是先进人工智能生态系统的重要组成部分。

- 私有数据壁垒：例如 ERP、CRM 厂商掌握着企业最核心的经营数据，这是通用大模型无法触及的“黑盒”。除了经营数据，还包括医疗、工业制造、保险等多个细分领域。
- 工作流深度：复杂的业务逻辑（如复杂的税务计算、供应链调度）需要极高的精度，通用模型的“幻觉”在这些场景下是不可接受的，必须依赖成熟软件的规则约束。

中国特色服务，构筑起 AI 无法逾越的护城河。这一逻辑在中国市场尤为显著。与美国 SaaS 标准化的订阅模式不同，中国计算机公司绝大部分以项目制为主要商业模式。

1）解决方案的复杂性：国内政企客户的需求高度定制化，往往涉及多系统打通、私有化部署及复杂的信创适配。这不仅是写代码的问题，更是对客户业务理解深度的考验。

2）商务与服务壁垒：中国市场的商业成交高度依赖长期的客情关系与驻场服务。大模型无法代替售前团队去搞定复杂的招投标流程，也无法代替实施团队去现场解决突发的业务故障。

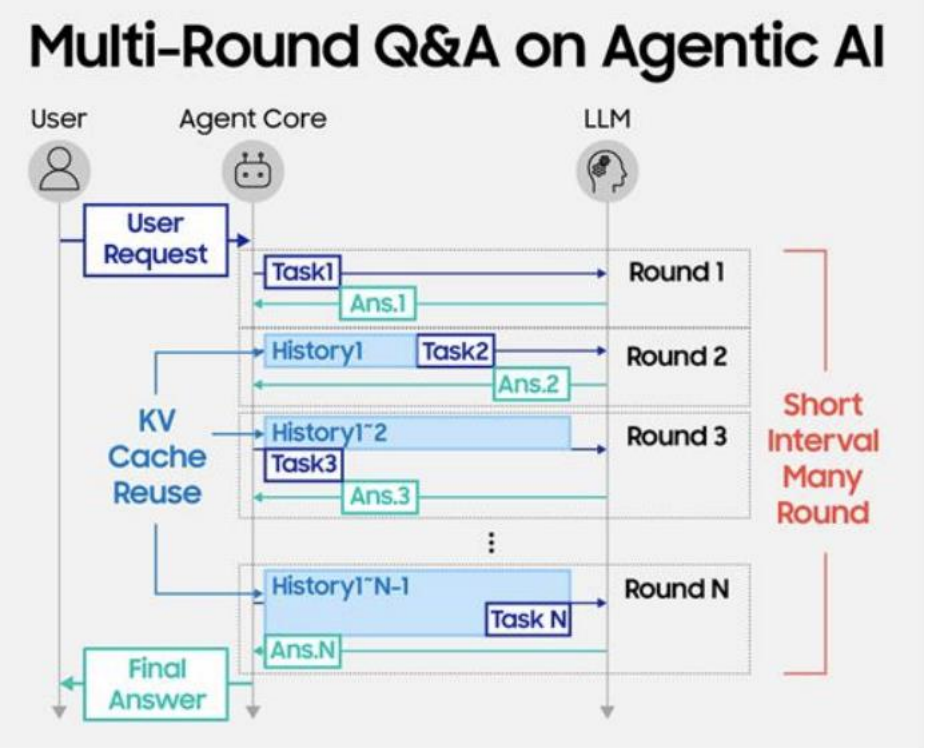
我们认为，对于中国软件公司而言，Anthropic 等基座模型的更新，本质上是生产力工具的升级。它能降低软件公司的开发成本，提升交付效率，而非直接取代其商业地位。当前市场的恐慌性抛售，可提供了低位布局优质行业软件龙头的机会。.

五、存储、CPU、FAB 加速成长通道

5.1 存储：AI 驱动的超级周期，产能扩张与技术迭代将推动长期供需再平衡

Agentic AI 框架依赖于 Agent Core 与大型语言模型之间的迭代式多轮交互。对于每个用户请求，代理核心通常会“解读用户意图—生成执行计划—通过工具检索外部信息—整合中间结果—重复该循环，直至最终输出与用户目标一致”。 由于该循环涉及对同一任务的反复优化，模型常需多次处理相同提示词的不同变体。随着提示词在迭代过程中不断扩展，为每个循环重新计算整个预填充阶段的效率将日益降低。通过利用 KV 缓存，在解码过程中可复用先前计算的键值对，从而避免冗余计算，并在多步推理工作流中保持较低的延迟。在多 Multi-Agent 中，工作负载可能因负载均衡、弹性资源分配或故障恢复而在 GPU 之间迁移。当活动会话迁移至另一 GPU 时，本地存储的键值缓存将丢失，必须在新设备上通过预填充阶段重新生成，这会引入显著延迟。为避免此开销，键值缓存必须超越本地 GPU 内存实现持久化。如图 3 所示，将缓存卸载至外部系统资源（如 CPU 内存或基于 NVMe 的存储）可实现迁移后的快速重载与复用。对于跨节点迁移，缓存必须通过共享资源在多个计算节点间保持可访问性，确保分布式执行过程中的连续性。在这些动态环境中，KV 缓存卸载不仅作为性能优化手段，更是维持稳定延迟、可预测吞吐量和可扩展推理操作的基础架构机制。

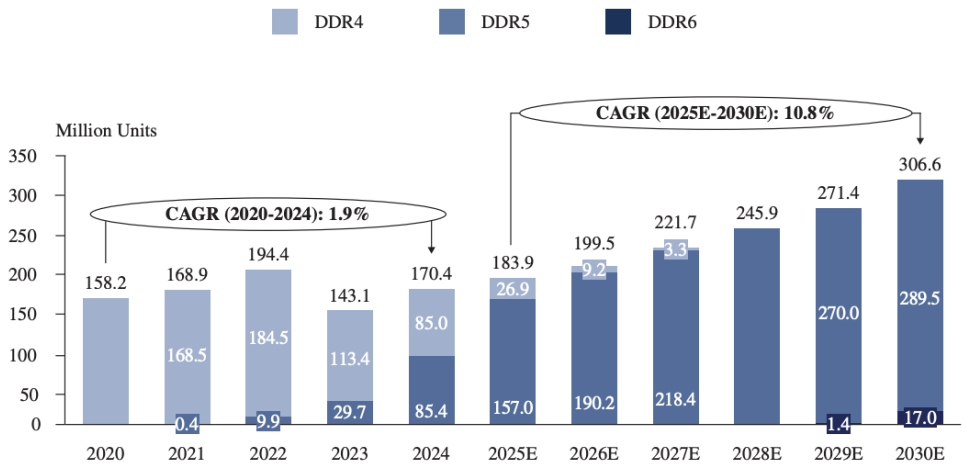
图表13：Multi-Agent 环境下 KV 运行机制



来源：三星，国金证券研究所

当前全球存储芯片市场正处于由 AI 驱动的结构性供需失衡阶段。据 CFM 闪存市场预测，2025 年全球存储市场规模有望达到 1,932 亿美元，创下历史最高纪录。① 需求端：AI 算力需求呈指数级增长，存储缺口具有结构性特征。据美光数据，AI 服务器的 DRAM 容量需求是普通服务器的 8 倍，NAND 容量需求达 3 倍，单台 AI 服务器存储需求高达 2TB。同时，AI 大模型"训练-推理"的正循环使数据存储需求持续放大。根据澜起科技招股书，2025 年全球服务器对 DDR5 渗透率将达 85%，HBM 单价较 DDR5 高约 5 倍，大规模应用进一步打开市场空间。② 供给端：头部厂商产能扩张节奏显著滞后。三星、SK 海力士与美光的主要新增产能要到 2028 年之后才会大规模释放，2026-2027 年内供需紧平衡格局难以改变。经历 2023 年库存危机后，头部厂商转向"精准减产+高端倾斜"策略，DDR4 等传统产能收缩与 AI 存储需求扩张形成结构性失衡。③ 长期展望：随着 3D NAND 层数突破 200 层、HBM4 量产体系逐步建立（2025 年全球 HBM 总产能已增至 54 万片，同比增长 105%）、以及三星 HBF 等新架构研发推进，存储产业正朝着"DRAM 缓存+HBF 加速+NAND 海量存储"的多层架构演进，技术突破将逐步缓解供需矛盾，推动行业进入新的增长轨道。

图表14：2020-2030 年全球服务器内存模组出货量（按 DRAM 代际细分）



来源：Frost & Sullivan，国金证券研究所

5.2 Agent 生态扩张引爆 CPU 性能瓶颈

随着大模型的应用从简单的 Chatbot 向能完成复杂任务的 Agent 演进，计算负载的重心正在发生微妙的偏移。Agent 不仅需要 GPU 进行模型推理，更依赖高性能 CPU 来处理复杂的逻辑编排、工具调用和内存管理。以下是我们认为 Agent 驱动 CPU 需求爆发的三大核心逻辑：

① Multi-Agent 架构引发的 OS 调度压力

传统的 LLM 对话是线性的，而 Agent 的工作流则是复杂的闭环。“推理→执行→评估→反思”的循环机制： Agent 需要在生成 Token 之外，执行大量的逻辑判断和状态管理。，模型需要不断在“思考”和“行动”之间切换。导致操作系统层面的上下文切换和进程调度任务大幅增加。

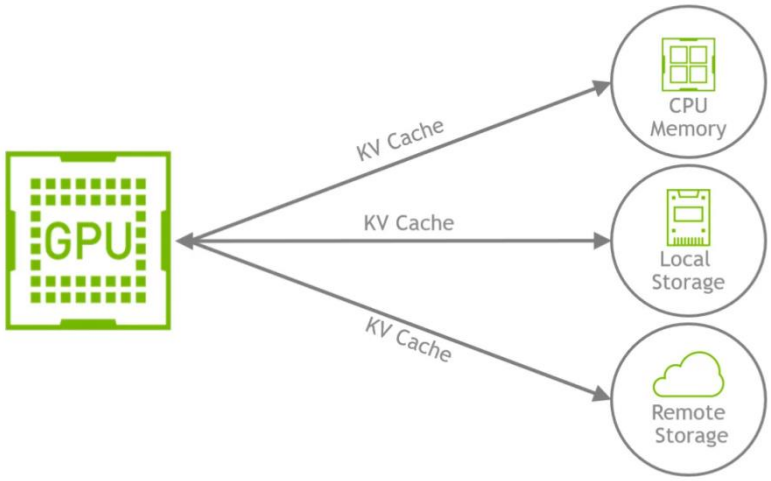
沙盒（Sandbox）需求飙升： Agent 执行代码等操作经常需要在隔离的云端沙盒中运行。这些沙盒环境的启动、运行和销毁依赖 CPU 算力。

② 长上下文场景下的 KV Cache 卸载对 CPU 的挑战

naddod 的技术文章阐述了其原理，键值缓存 (KV Cache) 可以加速 Transformer 推理，但它也会带来一个副作用：消耗大量显存。随着大型语言模型上下文长度的不断增长，这个问题会变得越来越突出。例如当上下文长度达到 8 万个 token 时，仅 KV Cache 本身就可能消耗数十 GB 的显存。更重要的是，GPU 显存不仅要容纳 KV Cache，还要容纳模型权重和中间计算结果。一旦显存耗尽，推理就会崩溃甚至失败。为了解决这一冲突，业界提出了键值缓存卸载（KV Cache Offload）方案。其核心思想是将 GPU 内存中不活跃或暂时未使用的键值数据迁移到其他存储介质例如 CPU 内存或者 SSD。然而 CPU 与 GPU 之间的通信带宽远低于 GPU 内部的 HBM 带宽。而且在进行 KV Cache 传输和管理时，也需要 CPU 进行任务的调度，进一步加剧了 CPU 的负载。

NVIDIA 2025 年 9 月的一篇技术博客《How to Reduce KV Cache Bottlenecks with NVIDIA Dynamo》就专门阐述了在长上下文场景下，利用 NVIDIA Dynamo 等技术将 KV Cache 卸载到 CPU 内存的必要性，并指出这是解决 HBM 瓶颈的关键手段。

图表15：KV Cache 卸载使得 KV Cache 能够从有限的 GPU 内存中传输到更大且性价比更高的存储



来源：Nvidia 官网，国金证券研究所

③ 高并发工具调用带来的 CPU 算力消耗

Agent 的能力不仅在于对话，更在于使用工具，例如检索、写代码、浏览网页。这些非模型推理任务主要由 CPU 承担。前文五大代表性 Agent 工作负载中各项任务的延迟数据证明了这一点。而且在高并发场景下可能有大量 Agent 同时工作，这些任务需要高性能 CPU 进行多线程/多进程处理。
据芯榜 1 月 19 日报道，英特尔将 Intel 3 和 intel 7 产能紧急转向服务器，致使消费电子端交付保证率大幅下滑。

英伟达 Blackwell 架构的 ARM CPU 存在严重瓶颈，因此新一代 Rubin 架构大幅提升 CPU 核心数与超线程；同时英伟达开放英特尔 x86 CPU 用于 NVL72 互联机柜。而 Agent 云端沙盒调用量飙升带动云实例业务增长，进一步加剧了 CPU 供需紧张。

市场研究机构 Jon Peddie Research 2025 年 8 月公布的最新数据显示，全球客户端 CPU 市场已连续两个季度实现增长。2025 年第二季度，客户端 CPU 出货量环比增长 7.9%，同比增长 13%；同期服务器 CPU 出货量同比增长 22%，环比小幅上升 0.6%。

图表16：2025 年 Q2 全球客户端 CPU 市场增长状况以及服务器与客户端 CPU 比例

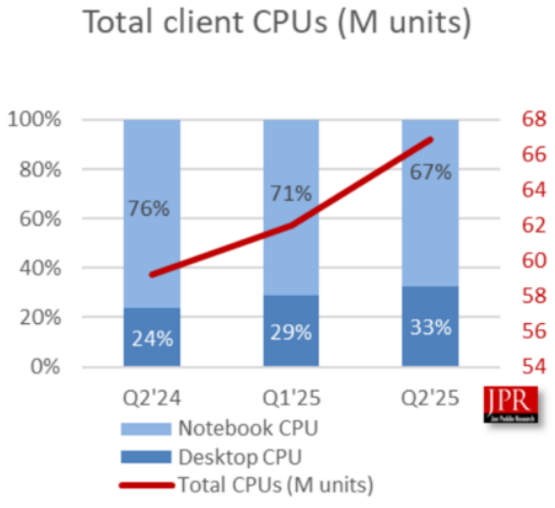


图 1. 今年前两个季度出现了非季节性增长。

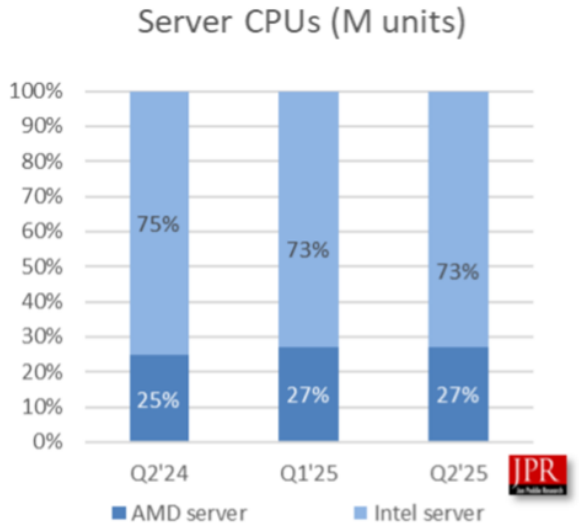


图 2. 服务器在整体市场份额中的比例。

来源：Nvidia 官网，国金证券研究所

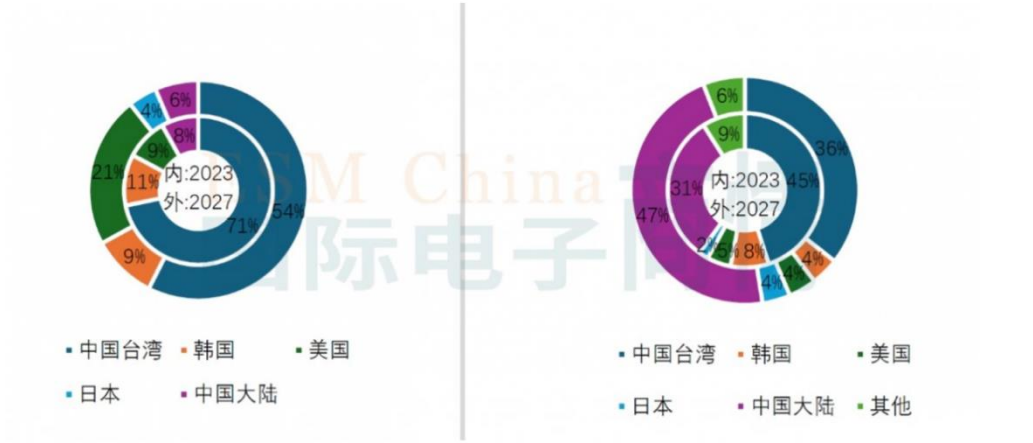
2026 年 1 月 22 日，英特尔 CFO 表示预计第一季度可用供应将降至最低水平，随后在第二季度及以后有所改善，公司正应对整个行业的供应短缺。

我们认为，Agent 时代算力的“木桶效应”已经显现，业界从经历了从GPU 的堆叠，到存储的短缺，目前 CPU 正演变为类似于存储的新短板。补足这一短板将是下一阶段算力基础设施建设的重中之重

5.3 FAB 晶圆代工：全球产能大扩张+先进封装突破将，释放行业增长潜力

晶圆代工作为半导体产业链的核心制造环节，其产能供给直接决定了存储、CPU 及 AI 芯片的交付能力。据 TrendForce 数据，2025 年 Q2 全球前十大晶圆代工厂营收达 417 亿美元，同比增长 14.6%创历史新高。① 先进制程大规模扩产：台积电 2025 年启动 2nm 量产（Fab 20 试产），预计年底月产能达 5 万片）。台积电市占率已达约 70%，2025 年营收占全球晶圆代工的绝对主导。此外，中国大陆晶圆产能 2025 年预计增长 14%至 1,010 万片 wpm，几乎占全球总产能的三分之一。② 先进封装瓶颈加速突破：CoWoS 作为 AI 芯片制造的关键瓶颈，台积电正全力扩产。2025 年 CoWoS 月产能翻倍至 7.5 万片，2022-2026 年 CAGR 超过 50%。台积电董事长魏哲家表示，CoWoS 目前严重供不应求，公司将持续扩产，预计 2025-2026 年实现供需平衡。下一代封装技术（CoPoS、FOPLP）也在加速研发，有望进一步释放产能。③ 长期展望：全球半导体制造版图正在深度重构。先进制程方面，中国台湾占比从 71%降至 54%，美国从 9%升至 21%；成熟制程方面，中国大陆占比从 31%飙升至 47%。我们认为，随着各地区新建晶圆厂陆续投产，叠加先进封装技术的持续突破，制造端的产能瓶颈将逐步解除，为存储、CPU 和 AI 芯片的供给提供系统性保障，驱动整个半导体行业进入新一轮蓬勃发展周期。

图表17：2023 至 2027 年，全球先进制程(左)、成熟制程(右)版图占比



来源：TrendForce，国金证券研究所

风险提示

行业竞争加剧的风险：在信创等政策持续加码支持计算机行业发展的背景下，众多新兴玩家参与到市场竞争之中，若市场竞争进一步加剧，竞争优势偏弱的企业或面临出清，某些中低端品类的毛利率或受 到一定程度影响。
技术研发进度不及预期的风险： 计算机行业技术开发需投入大量资源，如果相关厂商新品研发进程不及预期，表观层面将 呈现出投入产出在较长时期的滞后特征。
特定行业下游资本开支周期性波动的风险： 部分计算机公司系顺周期行业，下游资本开支波动与行业周期性相关性较强，或在个别年 份对于上游软件厂商的营收表现产生扰动。 李可夫 2026-02-08 01:31 李可夫 2026。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

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

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究