

# 把握 AI 创新，找寻价值扩张方向

**2026 年 01 月 28 日**

➤ **回顾与展望，AI 投资的机遇和挑战：**我们于 25 年初的深度报告《AIDC 电源系列一：“速率+功率”为未来 AI 产业发展的核心矛盾》中首次提出“速率+功率”为未来 AI 产业发展的核心矛盾。在过去的一年内，无论是速率赛道的光+PCB，还是功率赛道中的电源+液冷，都走出了“波澜壮阔”的行情。

那么站在当下，我们怎么看未来一年的算力机遇？我们认为，26 年要重点观察 CSP 及大模型厂商的商业闭环节奏，从而把握整体行业β。同时，积极找寻价值量扩张、资本开支增量倾斜的细分赛道，主线延续“速率+功率”。

**从资本开支到 ROI 测算，解读算力核心变量。**我们认为，算力需求主要看 Tokens 数+Capex。其中，Token 数（包括日活等）主要反映实时的算力需求，而 Capex 则反映云厂商的未来算力预期。部分商业闭环良好的云厂商，如谷歌等，已形成“开支→算力→Token→收入→再开支”正循环。我们主要测算了各大云厂商的 Capex/经营现金流/ROI，从而衡量公司可持续投资，以及 AI 商业闭环能力。

## ➤ 把握 AI 的增量赛道。

海外算力方面，我们延续“速率+功率”的投资思路：

- 1) 速率：光：把握光入柜内的趋势，抓住光模块的业绩线、光芯片的缺货潮、硅光的渗透率提升趋势。关注超节点技术带来的 OCS 等产业趋势。PCB：材料+设备升级是核心焦点。NV 推出全新 PCB 解决方案，M9 等级基材、HVLP4 铜箔与石英纤维布构建的 PCB 正交背板方案成为升级趋势，同步拉动材料和设备升级。
- 2) 功率：单卡和机柜功率密度持续提升，对电力架构提出了新的要求，也使得液冷成为数据中心的标配。

国产算力方面：25 年破局，26 年有望高速增长。需求侧，国产大模型加速追赶，云厂商资本开支展望积极；供给侧、国产先进制程从单点突破走向多点开花。行业供需两强之下，国产算力厂商迎破局元年。

其他方面：半导体，关注 AI 赋能下的存储超级周期，设备受益原厂扩产。消费电子，关注 AI 终端，跟踪华米 OV、OpenAI、Meta 等行业龙头的探索。

➤ **投资建议：**算力产业是科技之基，我们长期看好、深度跟踪。在当前市场对远期增量仍有所担忧之际，我们建议积极寻找价值量扩张、资本开支增量倾斜的细分赛道，主线延续“速率+功率”。同时重点关注国产算力、半导体设备、存储、AI 终端的投资机遇。

➤ **风险提示：**AI 产业发展的不确定性；AI 资本开支不及预期；下游需求不及预期；ROI 测算局限性。

## 重点公司盈利预测、估值与评级

代码	简称	股价 (元)	EPS (元)			PE (X)			评级
			2025E	2026E	2027E	2025E	2026E	2027E	
工业富联	601138	60.69	1.83	3.27	4.32	33	19	14	推荐
胜宏科技	300476	264.39	5.76	9.83	14.37	46	27	18	/
生益科技	600183	74.52	1.41	2.11	2.83	53	35	26	/
中芯国际	688981	128.55	0.67	0.84	1.05	192	153	122	推荐
兆易创新	603986	323.68	2.33	3.24	3.99	139	100	81	推荐
拓荆科技	688072	378.00	3.22	4.70	6.28	117	80	60	推荐
东山精密	002384	74.83	1.57	2.13	2.70	48	35	28	推荐

资料来源：iFind，国联民生证券研究所预测；

（注：股价为 2026 年 1 月 28 日收盘价；未覆盖公司数据采用 iFind 一致预期）

## 推荐

**维持评级**

**分析师 方竞**

执业证书：S0590525120003

邮箱：fangjing@glms.com.cn

**分析师 李少青**

执业证书：S0590525110049

邮箱：lishaoqing@glms.com.cn

**分析师 李萌**

执业证书：S0590525110050

邮箱：l meng@glms.com.cn

**分析师 袁姐**

执业证书：S0590525110053

邮箱：yuanda@glms.com.cn

**分析师 李伯语**

执业证书：S0590525110052

邮箱：liboyu@glms.com.cn

**分析师 王海**

执业证书：S0590524070004

邮箱：wanghai@glms.com.cn

**分析师 王晔**

执业证书：S0590521070004

邮箱：wye@glms.com.cn

**研究助理 蔡濠宇**

执业证书：S0590125110078

邮箱：caihaoyu@glms.com.cn

## 相对走势



## 相关研究

1. 电子行业点评：鸿蒙生态扩容提速，星闪重构无线音频-2025/12/03

## 投资聚焦

### 研究背景

在 AI 成为全球市场成长主线的大背景下，目前市场对 CSP 厂商的资本开支仍有所疑虑，担心 ROI，担心远期增量不明朗等。我们于 **25 年初的深度报告《AIDC 电源系列一：“速率+功率”为未来 AI 产业发展的核心矛盾》** 中首次提出 “速率+功率” 为未来 AI 产业发展的核心矛盾。在过去的一年内，无论是速率赛道的光+PCB，还是功率赛道中的电源+液冷。都走出了“波澜壮阔”的行情。本报告旨在解答，站在当下，我们怎么看未来一年的算力机遇。

### 区别于市场的观点/方法

我们认为，26 年要重点观察 CSP 及大模型厂商的商业闭环节奏，从而把握整体行业  $\beta$ 。同时，**积极找寻价值量扩张、资本开支增量倾斜的细分赛道，主线延续“速率+功率”**。

研究方法方面，我们通过**资本开支和 ROI 测算，解读算力核心变量**。我们认为，算力需求主要看 Tokens 数+Capex。其中，**Token 数（包括日活等）主要反映实时的算力需求，而 Capex 则反映云厂商的未来算力预期**。部分商业闭环良好的云厂商，诸如谷歌等，已形成“**开支→算力→Token→收入→再开支**”的正循环。

### 近期催化

2026 年 CES 展全球 AI 龙头聚焦物理 AI 落地、端云全栈升级、开源生态扩张等；国产大模型和 AI 芯片厂商陆续上市，智谱 AI 和 MiniMax 分别于 1 月 8 日、1 月 9 日在港交所上市，同时壁仞科技、天数智芯也分别于 1 月 2 日、1 月 8 日在港交所上市，百度昆仑芯 1 月 1 日提交港交所上市申请。

### 结论与建议

在当前市场对 AI 远期增量仍有所担忧之际，我们建议**积极寻找价值量扩张、资本开支增量倾斜的细分赛道，主线延续“速率+功率”**。同时重点关注国产算力、半导体设备、存储、AI 终端的投资机遇。

# 目录

<b>1 从资本开支到 ROI 测算，解读算力核心变量 .....</b>	<b>4</b>
1.1 资本开支的贡献者 .....	6
1.2 AI 投资驱动下，云厂商资本开支扩张积极 .....	7
1.3 谷歌引领 CSP，开启下一代 AI 产业浪潮 .....	8
1.4 ROI 及现金流：AI 商业闭环成为 Capex 投资重心 .....	17
1.5 主权 AI 是可观增量 .....	19
<b>2 算力芯片和服务器——大模型的底座 .....</b>	<b>23</b>
2.1 英伟达路线图 .....	23
2.2 ASIC 路线图：自研 AI 算力芯片加速迭代 .....	25
2.3 工业富联——AI 算力领军供应商 .....	28
2.4 速率+功率，算力产业的机遇与挑战 .....	29
<b>3 光：算力时代的核心破局点 .....</b>	<b>31</b>
3.1 光通信的下一站：NPO+CPO .....	31
3.2 OCS 光交换机：全光互联时代的基石 .....	39
<b>4 PCB：材料+设备升级是核心焦点 .....</b>	<b>45</b>
4.1 NV 推出全新 PCB 解决方案，技术升级清晰 .....	45
4.2 英伟达 NVL576 采用 PCB 正交背板实现高速互联 .....	48
4.3 PCB 材料配套 M8/M9 等级升级 .....	51
4.4 M9 材料升级对 PCB 设备及耗材提出更高要求 .....	56
<b>5 功率提升拉动电源+液冷升级 .....</b>	<b>60</b>
5.1 高压直流是未来柜外电源的趋势 .....	60
5.2 液冷从 0-1 实现产业趋势突破 .....	67
<b>6 国产算力：需求侧资本开支展望积极，国产大模型加速追赶 .....</b>	<b>70</b>
6.1 资本开支：25 年受 H2O 扰动，26 年展望积极 .....	70
6.2 大模型：国产大模型弯道超车 .....	72
<b>7 国产算力：供给侧向“芯”而行，国产算力破局元年 .....</b>	<b>76</b>
7.1 晶圆厂：国产算力底座 .....	76
7.2 先进封装助力摩尔定律延续 .....	79
7.3 算力芯片：技术迭代加速，国产替代格局明晰 .....	83
<b>8 AI 驱动存储迎超级周期，设备受益原厂扩产 .....</b>	<b>97</b>
8.1 AI 驱动存储行业迎快速增长 .....	97
8.2 半导体设备受益存储上行周期 .....	101
<b>9 AI 赋能终端，产业范式重构 .....</b>	<b>105</b>
9.1 AI 手机：存量市场内结构性硬件+AI 功能创新 .....	106
9.2 新型 AI 终端：大厂布局新战场，竞逐 AI 时代新入口 .....	109
<b>10 投资建议 .....</b>	<b>118</b>
<b>11 风险提示 .....</b>	<b>120</b>
<b>插图目录 .....</b>	<b>121</b>
<b>表格目录 .....</b>	<b>123</b>

## 1 从资本开支到 ROI 测算，解读算力核心变量

Token 是大模型处理文本的基本计算单位（1 英文 token≈4 字符/0.75 词，非英文 token 字符占比更高），其处理量与计算量呈平方级增长，直接反映算力需求。模型调用时，Token 消耗需依托 GPU/TPU 集群、电力、网络等底层资源，而资本开支正是支撑这些资源建设的核心动力。

**Token 量体现当前算力需求，资本开支反映未来算力预期。**头部厂商通过资本开支投入芯片、数据中心，提升 Token 处理能力；而 Token 增长（尤其是商业化场景 Token）又反推资本开支扩容，形成“**开支→算力→Token→收入→再开支**”的循环。

**主流商业模式（C 端订阅、B 端 API 调用）均以 Token 为定价单位。**本质上是将“智能”与“算力”进行了标准化的货币兑换，传统的软件是零边际成本，但 LLM 每次生成都有显著的 GPU 电力与损耗成本。Token 计费是目前唯一能让收入与物理成本线性对齐的商业单位。

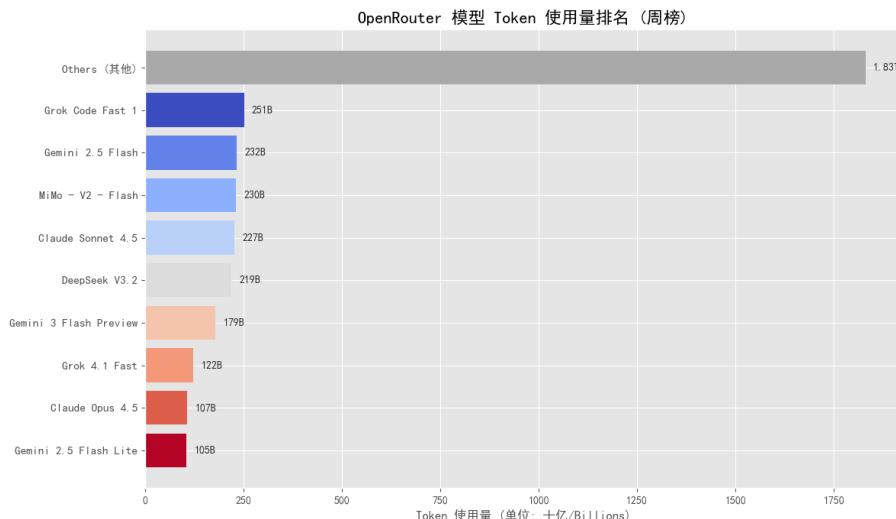
**表 1：头部 AI 大模型定价对比**

服务类型	OpenAI (GPT-4o)	Anthropic (Claude 3)	Mistral	Google Gemini
聊天模型（每百万 Token）	输入：\$5 输出：\$15	Haiku: \$0.25/\$1.25 Sonnet: \$3/\$15 Opus: \$15/\$75	Small: \$2 Large: \$8	1.5 Pro: \$7/\$21
图像生成	DALL-E 3: 每 1024×1024 图片 \$0.04	暂不提供	暂不提供	Gemini Pro Vision: 自定义定价
语音转文字	Whisper: 每分钟 \$0.006	暂不提供	暂不提供	Google 语音转文字: 每分钟 \$0.012
代码辅助	GitHub Copilot: 每月 \$10	暂不提供	暂不提供	Gemini 开发者版: 每用户每月 \$19
视觉能力	Vision: 每千 Token\$0.03	Claude 3 支持视觉功能	暂未广泛提供	Gemini Pro Vision 可用

资料来源：holori, 国联民生证券研究所

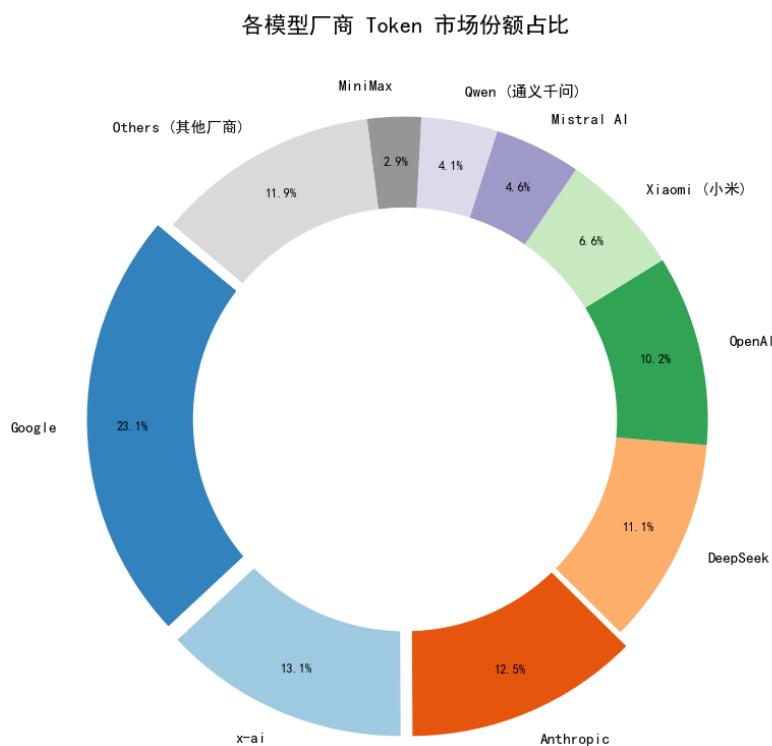
根据 OpenRouter 的数据，在 2025 年 12 月 22 日到 12 月 29 日 OpenRouter 中 Token 使用量排名前三名的是 xAI 的大模型 Grok Code Fast 1、Google 的 Gemini 2.5 Flash 和 小米的 MiMo-V2-Flash。同期，在各大模型厂商中，Google 的 token 数在市场中占比较高，达 23.1%；其次是 X-Ai 和 Anthropic。

图1：OpenRouter 模型 Token 使用量排名 (2025.12.22-12.29)



资料来源：OpenRouter，国联民生证券研究所

图2：各模型厂商 Token 市场份额占比 (2025.12.22-12.29)



资料来源：OpenRouter，国联民生证券研究所

**Token 高增需依赖场景突破，商业化是核心瓶颈。**当前，Token 使用增长主

要来自 AI 搜索、智能聊天工具等场景（如谷歌 AI Overview）；后续会重点推进 AI + 企业服务、智能 Agent 等具备商业化潜力的场景布局，以持续带动业务增长。在资源投入方面，未来相关资本开支的规划，将与高价值应用场景的推进进度相匹配。

**资本开支转向“效率优先”，硬件**（如英伟达 GB200 推理效率较 H100+30 倍）与软件优化降低实际经济投入，企业不再单纯追求 Token 规模，而是通过芯片自研、系统升级提升单位 Token 的算力性价比。

## 1.1 资本开支的贡献者

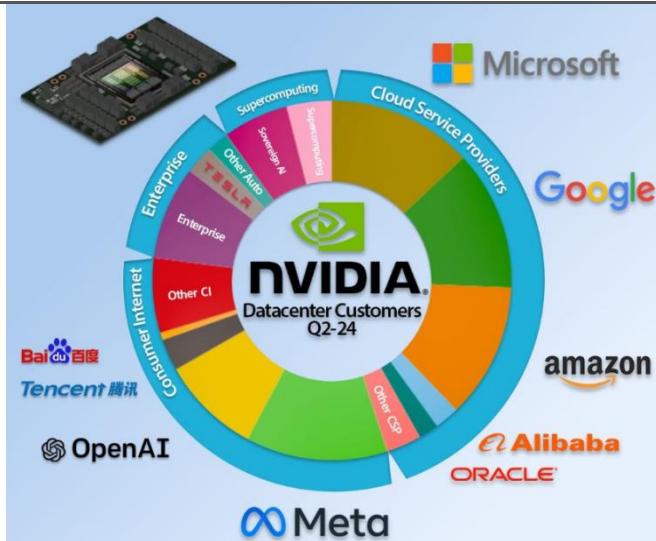
**NV 作为全球算力的绝对龙头，其下游客户涵盖各个行业，我们将英伟达的主力客户，分为 CSP、大模型厂商、主权 AI 和政企客户三大类别。**

**1) CSP 客户：**CSP 厂商主要包括谷歌、Meta、微软、亚马逊、甲骨文等，是英伟达数据中心收入的主要来源，CSP 厂商资本开支快速增长，也成为支撑英伟达业绩持续高增的核心要素。3Q25 全球五大 CSP 厂商资本开支合计达到 3081 亿美元，同比增长 75%，算力军备竞赛下，云厂商资本开支进入加速增长阶段。

**2) 大模型厂商：**除了 CSP 客户以外，大模型厂商同样是英伟达收入贡献的主要来源。全球主流大模型厂商包括 OpenAI 等，此外，谷歌、meta 等厂商在大模型领域也表现突出。

**3) 主权 AI 和政企客户：**地缘政治不确定性进一步加速主权 AI 落地，政策端的技术主权导向与区域化布局，正成为主权 AI 市场扩容的重要支撑。据 Gartner 预测，2027 年全球将有 35% 的国家完成区域专属 AI 平台的锁定，通过本地化专有情境数据构建技术壁垒与竞争优势；据埃森哲预测，至 2028 年，全球 65% 的政府将出台技术主权相关专项法规，进一步规范主权 AI 的场景化应用需求。政策端的密集落地将推动主权 AI 从核心基础设施建设向多行业渗透延伸，为市场规模持续扩容提供坚实保障。

图3：英伟达数据中心收入按照客户拆分



资料来源：Substack，国联民生证券研究所

## 1.2 AI 投资驱动下，云厂商资本开支扩张积极

我们对北美主要云厂商：谷歌、微软、亚马逊、Meta 以及甲骨文的资本开支进行复盘。自 ChatGPT 于 22 年底推出以来，海外云厂商对 AI 基础设施投资力度显著增强，**2023、2024 及 2025 年 Q1-Q3，五家公司资本开支合计分别达 1602 亿、2591 亿 (yoY+62%)、3081 亿 (yoY+75%) 美元，资本开支进入加速扩张周期**，具体来看：

**谷歌：**谷歌的资本开支主要用在 Google Cloud、Gemini 大模型训练，以及支撑广告业务的基础设施等。2024 年谷歌 Capex 达 525 亿美元，yoY+63%；**2025 年 Q1-Q3 谷歌 Capex 进一步增长至 636 亿美元，yoY+66%，其中 3Q25 Capex 为 240 亿美元，yoY+83%，Capex 投入明显加速。**

**微软：**微软的资本开支主要投向两部分：Azure 云和 Copilot 相关的算力需求。2024 年微软 Capex 达 756 亿美元，yoY+84%；**2025 年 Q1-Q3 微软 Capex 增长至 805 亿美元，yoY+52%，其中 3Q25 Capex 达到 349 亿美元，yoY+75%，Capex 投入稳步增长。**

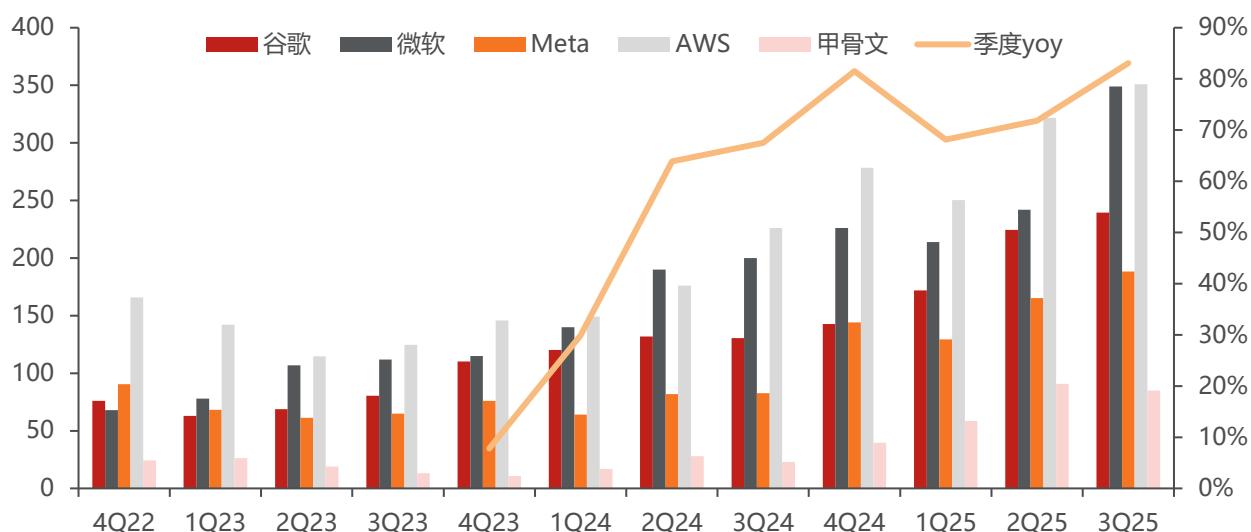
**Meta：**与其他几大 CSP 厂商不同，Meta 没有云业务，其资本开支主要用在广告推荐系统和 AI 模型训练。2024 年 Meta Capex 达 373 亿美元，yoY+37%；**2025 年 Q1-Q3 Meta Capex 增长至 483 亿美元，yoY+112%，其中 3Q25 Capex 达到 188 亿美元，yoY+128%。除体量较小的甲骨文外，Meta 2025 年前三季度 Capex 投入在 5 家 CSP 厂商中增速最快，虽然 Meta 不做云，但其广告和内容分发以及 Llama 系列大模型开发对算力要求较高，Capex 扩张较为激进。**

**亚马逊：**亚马逊 Capex 主要用于大规模的 AWS 云业务扩张。2024 年亚马逊

Capex 达 830 亿美元, yoy+57%; **2025 年 Q1-Q3 亚马逊 Capex 增长至 923 亿美元, yoy+67%, 其中 3Q25 Capex 达到 351 亿美元, yoy+55%。由于亚马逊本身为全球最大的云厂商, 本次 AI 驱动下的资本开支亦为业内最大。**

**甲骨文:** 甲骨文 Capex 主要用于追赶云业务以及支持自身业务。2024 年甲骨文 Capex 达 107 亿美元, yoy+55%; **2025 年 Q1-Q3 甲骨文 Capex 增长至 234 亿美元, yoy+246%, 其中 3Q25 Capex 达到 85 亿美元, yoy+269%。**由于甲骨文云业务体量较小, 仍需大量前置投入扩建算力基础设施, 其 Capex 扩张速度最为激进。

图4：云厂商资本开支情况（亿美元）



资料来源: Bloomberg, 国联民生证券研究所整理

## 1.3 谷歌引领 CSP, 开启下一代 AI 产业浪潮

### 1.3.1 谷歌模型家族

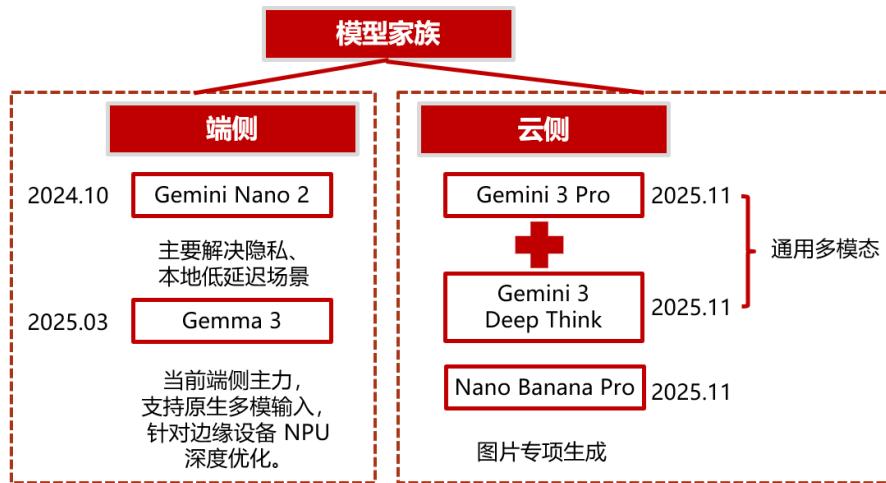
自 2023 年 12 月 Gemini 1.0 发布起, 谷歌大模型体系进入快速迭代: 2024 年 2 月 Gemini 1.5 以百万级上下文窗口树立长序列处理优势; 2024 年底 Gemini 2.0 引入原生 Agent 架构与工具调用, 实现模型—算力—应用的深度闭环。

2025 年, 谷歌同步推进开源与端侧布局, 2025 年 5 月 Gemma 3 在轻量级模型中实现推理能力跨越提升, 2025 年 8 月 Nano Banana 将高质量生成能力下沉至终端侧, 丰富谷歌生态的边缘算力矩阵。

**最新的 Gemini 3.0 Pro 于 2025 年 11 月推出,**其在深度推理、多模态理解及实时交互上全面增强, 被视为谷歌重新占据行业技术高点的关键节点。同时

推出的 Nano Banana Pro 是架构于 Gemini 3 Pro 的新一代图像生成与编辑模型，具有业内领先的文字渲染、精细编辑与现实世界知识，定位为专业级图像生产工具。

图5：谷歌在端侧和云侧大模型的布局



资料来源：谷歌官网，国联民生证券研究所整理

**云侧：Gemini 3 Pro 是谷歌于 2025 年 11 月 18 日推出的新一代大型语言模型。** Gemini 3 Pro 在推理、多模态、Agent 能力明显增强，在数学、代码、长文本等领域显著优于 Gemini 2.5 Pro、Claude Sonnet 4.5 与 GPT-5.1。采用大量自研 TPU 训练。谷歌的 Gemini 3 pro 的主要优势包括：  
**1) 多模态融合：**支持文本、音频、图像、视频等全链路多模态输入，能够理解和处理完整代码仓库、长视频文档等复杂内容，适合高难度推理与跨信息源整合场景。屏幕截图准确率 72.7%（行业领先 2 倍），Video-MMMU 视频推理 87.6%（超 GPT-5.1/Claude 4.5）；  
**2) 逻辑推理能力全面提升：**在 GPQA Diamond (91.9%)超越 GPT-5.1 的 88.1%）、AIME 2025 (95.0%/100%领先)、MMLU (91.8%领先) 强调数学、多步逻辑推理的基准任务上，Gemini 3 Pro 超越 GPT-5.1；  
**3) 长文本能力跨越式升级：**128K 全文长度下仍能保持高准确率，SimpleQA 测试 72%（超 Claude 4.5 的 29.3%）；  
**4) 自研 TPU：**依托谷歌自研 TPU 完成训练，可显著提升大模型训练速度并支持更大模型规模，具有较强的成本优势；

**端侧：Nano banana 是谷歌 2025 年 8 月发布的轻量图像模型 (Gemini 2.5 Flash Image)，主打图像编辑与多模态融合。** 聚焦边缘设备场景的轻量级智能模型，以“小体积、高性能、低门槛”为核心特色，面向消费电子、物联网终端及轻量化业务需求。Nano banana 的主要优势包括：  
**1) 自然语言驱动编辑：**无需复杂工具，仅通过文字指令即可实现“像素级精准修改”，支持模糊指令理解；  
**2) 角色与场景一致性：**可以保持“角色 / 对象一致性”，跨编辑、跨场景保持主体细节统一；多轮修改不“样貌漂移”，还原角色表情、服装纹理；  
**3) 多图像融合：**可同时合成多张图像，同时自动匹配光线方向、视角透视、阴影效果，支持对

已有图像进行局部 / 全局编辑；**4) 低成本：**每张图生成成本约 0.074 美元 (约 0.5 元人民币)。

图6：谷歌 Gemini 3 模型



资料来源：谷歌官网，国联民生证券研究所

图7：谷歌 Nano Banana 模型



资料来源：Nano Banana 官网，国联民生证券研究所

### 1.3.2 谷歌 TPU，当前最强 ASIC

谷歌的 TPU 研发始于 2013 年，自 2015 年起，谷歌就已经开始在内部使用 TPU，**截至目前，谷歌第七代 TPU (Ironwood) 在算力、能效与带宽方面实现全面跃升，是公司当前最强的自研加速器。**能效上，Ironwood 每瓦性能较 TPU v6 (Trillium) 提升约 2 倍；存储规格显著升级，单芯片支持 192GB HBM、7.4TB/s 带宽，分别为上一代的 6 倍与 4.5 倍，可支撑更大模型与更高吞吐；互联方面，双向带宽提升至 1.2Tbps (1.5 倍)，显著增强集群通信效率。整体而言，Ironwood 在能效、存储与互联三项核心指标全面进阶，为谷歌大模型训练/推理及算力外售布局提供更强底座。

表2：谷歌各代际TPU性能对比

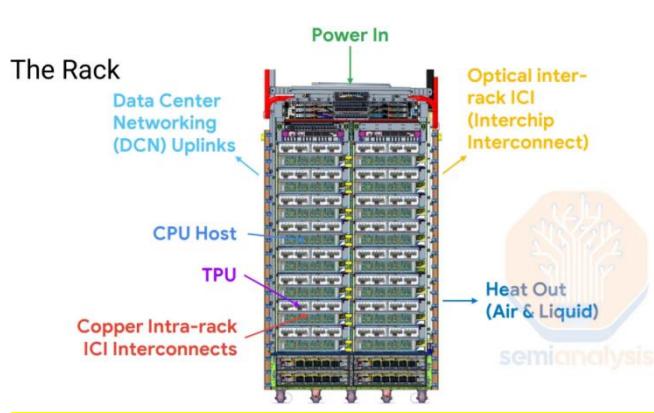
谷歌 TPU 芯片	TPUv1	TPUv2	TPUv3	TPUv4i	TPUv4	TPUv5p	TPUv5e	TPUv6p	Trillium	Ironwood
推出年份	2015	2017	2018	2020	2021	2023	2023	2024	2025	
制程技术 (nm)	28	16	16	7	7	5	5	4	3	
芯片大小 (mm <sup>2</sup> )	330	625	700	400	780	500	350	790	2*445	
芯片内存 (MB)	28	32	32	144	288	48	122	-	-	
时钟速度 (MHz)	700	700	940	1050	1050	2040	1750	2060	1633	
芯片容量 (GB)	8	16	32	8	32	95	16	32	192	
HBM 内存带宽 (GB/s)	300	700	900	300	1228	2765	819	1640	7372	
热设计功耗 (W)	75	280	450	175	300	537	225	383	959	
数据类型	INT8	BF16	BF16	BF16 INT8	BF16 INT8	BF16 INT8	BF16 INT8	BF16 INT8	BF16 INT8 FP8	
INT8 算力 (TOPS)	92	-	-	138	275	918	393	1836	4614	
BF16 算力 (TFLOPs)	-	46	123	69	137.5	459	196.5	918	2307	
FP8 算力 (TFLOPs)	-	-	-	-	-	-	-	-	4614	
单 CPU 主机芯片数量	4	4	4	8	4	8	8	8	8	

资料来源：EET，国联民生证券研究所

**在机柜架构方面，谷歌同样具备优势。**谷歌的 TPU 机柜通常采用 64 卡架构，机柜内部采用铜缆或光纤通信，机柜外部统一采用光通信进行组网。

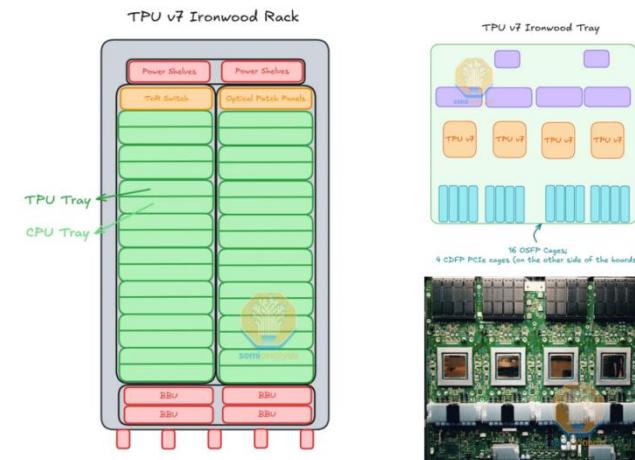
在谷歌 TPU v7 机柜中，一个 TPU 机柜配置 16 个 TPU Tray，以及 16 个或 8 个 CPU Tray，一个 ToR 交换机，两组 Power Shelf 以及 BBU；在每个 TPU Tray 中，PCB 板上安放四颗 TPU 芯片，一个 16 Tray 的 TPU 服务器中配置 64 颗 TPU 加速卡。

图8：谷歌 TPU v7 机柜实物图



资料来源：Semianalysis，国联民生证券研究所

图9：谷歌 TPU v7 机柜及 Tray 内示意图

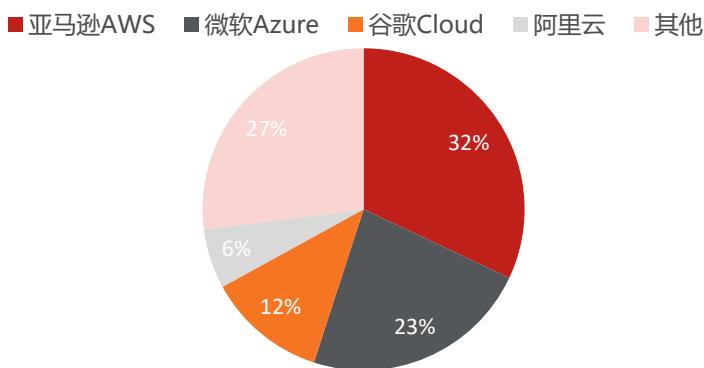


资料来源：Semianalysis，国联民生证券研究所

### 1.3.3 其他 CSP 厂商的 Capex 破局之道

各 CSP 厂商都在探索商业闭环，梳理 Capex 扩张思路，通过其重点投资的方向观察后续 Capex 破局之道。在 CSP 厂商中，目前亚马逊、微软、谷歌市占率靠前，根据 Synergy Research 的数据，2024 年全球云计算市场市占率中，亚马逊占比最高，达到 32%；其次是微软 Azure 和谷歌。

图10：2024年全球云计算市场市占率



资料来源：Synergy Research，国联民生证券研究所整理

#### 1) 微软 Azure：应用生态反向拉动的飞轮式突破

Azure 的正循环核心是“应用生态→算力需求→资本开支→生态强化”，依托 Office、Windows、Teams 等原生应用生态，将 AI 算力投入转化为生态粘性与 ARPU 提升，形成独特的、可持续的正循环逻辑。

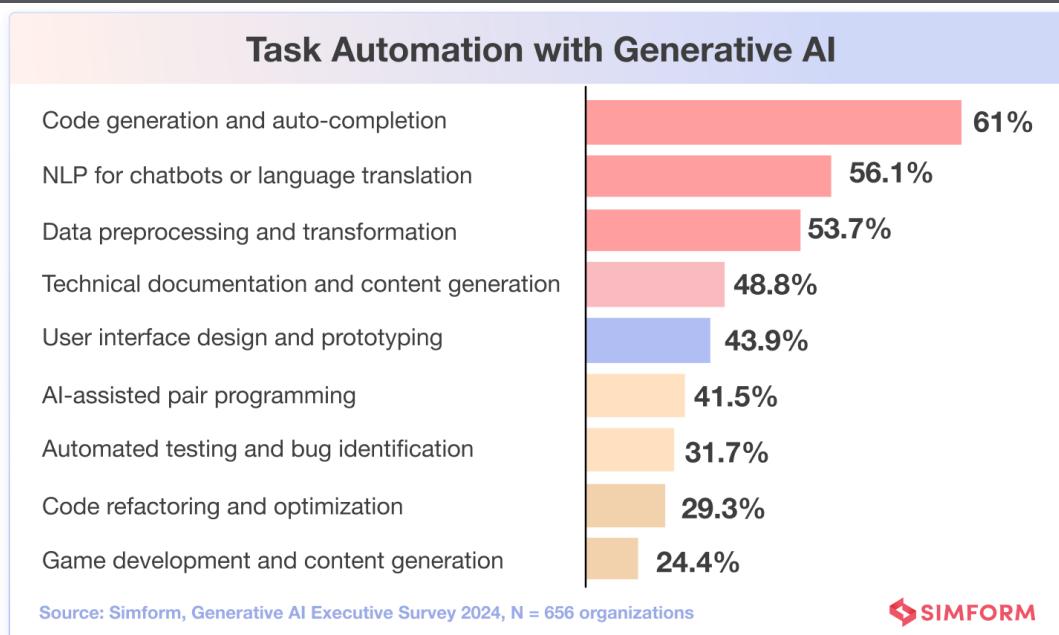
公司将算力重点分配于内部项目，其中 Copilot 为重中之重，此外 Azure 云业务亦为 Capex 的重点投资方向。25 年 10 月，OpenAI 承诺向 Azure 云采购约 2500 亿美元的服务，在未来多年内完成，**3Q25 微软 Azure 业务收入同比增长 39%。**

应用端生态绑定：Copilot 已全面嵌入 Office、Teams、Windows 等核心应用，实现 AI 从“工具”到“生态底层能力”的转变，驱动用户对 Azure 算力的刚性需求——并非为卖云而建算力，而是为强化生态而建算力。

**图11：Azure Copilot 用例**

资料来源：AWS 官网，国联民生证券研究所

需求可预测性突破：Office 等应用用户基数固定、使用场景高频，有望带动 Azure 算力需求稳定增长。

**图12：生成式 AI 在不同任务下的使用占比**

资料来源：SIMFORM，国联民生证券研究所

## 2) 亚马逊：系统级整合驱动成本与自主可控双突破

AWS 正循环的核心并非单一硬件创新，而是全链路自研 + 协同优化的系统级整合能力，通过“需求 - 研发 - 落地 - 反馈 - 迭代”的自循环，将资本开支从“被动投入”转化为“主动调控”，构建可持续的资本开支正循环。

公司为全球最大的云服务厂商，根据 Synergy Research 数据，2024 年亚马逊 AWS 云业务市占率达 32%，AWS 为亚马逊资本开支主要投入方向，预计 4Q25 新增 1GW 以上算力，并计划于 2027 年底实现算力翻倍。公司在 AI 领域的布局相对滞后，3Q25 AWS AI 相关云业务同比增长 18%，远低于 Azure 的 39%，谷歌及微软云分别具备 TPU 及 Open AI 大客户优势，其在 AI 市场的快速渗透给予 AWS 较大的竞争压力，AWS 推出 Trainium3 芯片及绑定 Anthropic 以应对挑战，并与 Open AI 达成合作协议，当前 AWS 在 AI 算力服务方面正加速追赶。

**表 3：AWS 2025 年 AI 芯片业务核心数据一览**

指标维度	核心数据
芯片部署规模	Project Rainier 已部署 50 万颗 Trainium 2
芯片业务收入	AWS 3Q2025 营收：\$33B，年增 20%，QoQ 增 \$2.1B，年化 run rate \$132B；Trainium 2 季度收入环比 +150%，达数十亿美元，产能全部满载
资本开支效率	截至 3Q2025，亚马逊资本支出已达 899 亿美元；2025 年现金资本开支约预计为 1250 亿美元。

资料来源：谷歌，openai，国联民生证券研究所

芯片迭代持续加速，从 Trainium 2 到 3nm 工艺的 Trainium 3（算力、内存容量 / 带宽显著提升），再到预计的 Trainium 4，不断强化核心优势；客户与市场验证中，Anthropic 等核心客户持续涌入，有望推动 2026 年收入提升；资本开支呈正向循环，聚焦 AI、定制芯片与基础设施，资本投入反哺技术研发，形成“投入 - 产出 - 再投入”的良性循环。

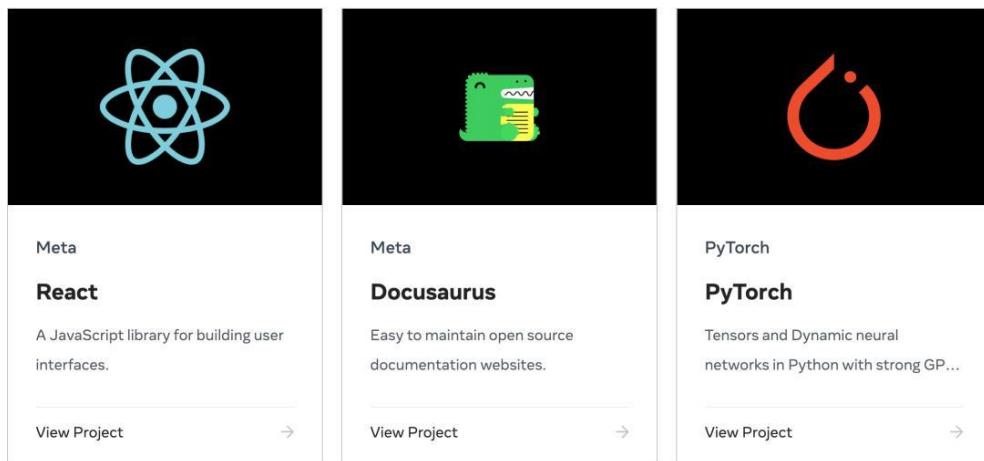
## 3) Meta：开源 + ASIC 重塑行业推理成本体系的颠覆性突破

Meta 的正循环突破是作为外生变量，通过开源模型生态与自研 ASIC 芯片的组合，打破 NVIDIA 主导的推理成本体系，带动全行业生态迁移与成本重置。

与其他几大云厂商不同，Meta 没有对外的云业务，其 Capex 专注于内部项目，如 Llama 系列模型训练、广告内容推荐等。Meta 在 AI 领域的 Capex 已经在对核心广告推荐业务形成明显拉动，在较为激进的 Capex 背景下，Llama 系列模型面临 Open AI、谷歌 Gemini 等顶尖模型的挑战，模型性能将成为关注重点。

双轮驱动布局：一方面保持 Llama 3 及未来 Llama 4 的稳定开源节奏，持续壮大开源模型生态；另一方面公开自研 MTIA v2 ASIC 芯片，采用台积电 5 纳米制程，通过加大芯片 SRAM、带宽和 LPDDR5，实现计算性能的优化方案，核心聚焦效率提升。

图13：META 开源项目图示



资料来源：Meta OpenSource 官网，国联民生证券研究所

**商业价值验证：**推荐业务已率先体现 AI 的边际回报 —— 开源模型驱动的开发者生态迁移，叠加 AI 优化的广告推荐算法，使得算力投入直接转化为商业化成果，形成 “技术投入→生态繁荣→商业变现”的闭环。

表 4：谷歌大模型相关业务发展动态

指标维度	核心数据
开源生态规模	Llama 模型已经被下载了超过 3 亿次
资本开支倾斜	2025Q3 资本支出为 193.7 亿美元。
业务增长动能	2025Q3 旗下所有应用程序的广告展示次数同比增长 14%，平均每条广告价格同比增长 10%，总营收 约 51.2B 美元，同比增长 26%。按固定汇率计算，营收将同比增长 25%。

资料来源：谷歌，openai，openrouter，国联民生证券研究所

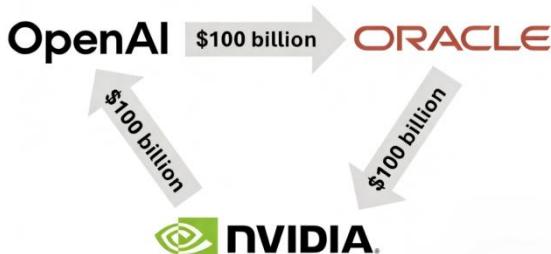
### 1.3.4 OpenAI 的内循环和外循环

**OpenAI 作为全球大模型龙头，与一系列算力基建供应商形成了“供应商投资+订单绑定+股权交叉”的商业和资本内循环，核心逻辑是通过 AGI 的远期成长空间，撬动芯片、云服务、制造等领域巨头的资金与资源，再将这些资源投入自身基础设施建设与模型研发，而合作方则通过投资或绑定订单获得股价上涨、股权收益等回报，资金与资源在生态内循环流转，进而推动整个体系扩张。**

25 年 Q3 以来，OpenAI 持续推进 AI 硬件布局，扩张与 Oracle，软银合作共建的星际之门项目，与 NVIDIA、AMD、Broadcom、AWS 等厂商建立战略合作伙伴关系，与富士康达成了合作意向。

图14：OpenAI、Oracle 与 NVIDIA 的合作循环

THE INFINITE MONEY GLITCH



资料来源：镜报，Reddit，国联民生证券研究所

表 5：OpenAI 在算力领域的合作伙伴

公布时间	合作厂商	领域	合作核心内容
2025/9/22	NVIDIA	芯片	建立战略合作伙伴关系，OpenAI 将用英伟达系统构建至少 10GW（相当于 400 万-500 万块 GPU）的 AI 数据中心；英伟达计划向 OpenAI 投资高达 1000 亿美元，且 OpenAI 的模型和基础设施与英伟达软硬件同步优化。
2025/9/23	Oracle, 软银	云计算	OpenAI、Oracle 与软银为星际之门 (Stargate) 宣布五处 AI 数据中心，确保在 2025 年底前实现 5,000 亿美元投资额与 10 GW 装机容量。
2025/10/6	AMD	芯片	建立战略合作伙伴关系，OpenAI 将部署 6GW 的 AMD GPU，AMD 向 OpenAI 授予可获得至多 1.6 亿股 AMD 普通股的股权。首批 1GW 的 GPU 部署于 2026 年下半年启动。
2025/10/13	Broadcom	加速器	建立战略合作关系，联合开发包 10GW 定制 AI 加速器及博通以太网解决方案的系统；博通计划部署加速器及网络系统机架，预计 2026 年下半年启动，2029 年底前完成。
2025/11/3	AWS	云计算	达成多年期战略合作伙伴关系，总额高达 380 亿美元，OpenAI 将开始 AWS 计算资源，其中包括数十万颗先进的 NVIDIA GPU，并具备扩展至数千万 CPU 的能力，所有算力将在 2026 年底前全面部署，为 ChatGPT 提供推理和模型训练服务。
2025/11/20	富士康	硬件制造	共同打造 AI 数据中心机架；强化并简化美国 AI 供应链，扩大对美国供应商采购，拓展本地测试与组装能力；富士康在美国生产布线、网络、散热和电力系统相关设备。

资料来源：OpenAI 官网，国联民生证券研究所整理

对外，OpenAI 也与企业、政府、消费者等构建了深度合作，跑通了商业模式的外循环。

#### To B: OpenAI 与 B 端客户的合作主要包括两种模式：1) 企业内部使用：

OpenAI 全球企业客户突破 100 万家（截至 25 年 11 月），其中 ChatGPT for Work 使用人数超过 700 万，部分企业通过 API 直接调用大模型，合作伙伴包括金融服务、医疗保健、零售等行业的领军企业。2) 拓展 OpenAI 生态：Shopify、Etsy、Walmart、PayPal 和 Salesforce 融入 ChatGPT，打造“智能电商”新体验；Canva 设计海报、Figma 编辑图像、Zillow 浏览房源等功能均集成在

GPT 中。

**To G: 推出 OpenAI 政府版，助力公用事业提效**，并将与美国国家实验室、国家航空航天局 (NASA) 以及国税局 (IRS) 等政府机构的既有合作纳入该框架；首个政府合作项目将与国防部开展，合同金额为 2 亿美元。

**To C: 目前全球 ChatGPT 每周活跃用户超过 8 亿，付费用户占其每周活跃用户基数的 5%，约为 4000 万。**OpenAI 预计 2030 年每周活跃用户将达到 26 亿人，其中 8.5% (约 2.2 亿人) 为付费订阅用户，使 ChatGPT 成为全球最大的订阅业务之一。

**表 6：OpenAI 各行业的代表性客户**

时间	公司	行业	应用进展
2025/12/1	埃森哲	咨询	埃森哲将为数万名员工配备 ChatGPT Enterprise，并对客户服务、供应链、财务、人力资源的企业职能提供全新 AI 解决方案，帮助包括金融服务、医疗保健、公共部门和零售业企业，将传统流程转化为人工智能驱动流程。
2025/10/14	沃尔玛	零售	打造以人工智能为先导的购物体验，允许顾客通过 ChatGPT 使用即时结账功能在沃尔玛购物。超越了传统的电商搜索栏，推动“智能商务”，人工智能将学习并预测顾客的需求，把购物从被动转变为主动。
2024/6/26	摩根士丹利	金融	摩根士丹利财富管理推出 AI 平台 Debrief，可作为客户会议的记录员、摘要和初稿撰写者，助力财务顾问业务规模化发展。大摩于 2023 年 3 月宣布与 OpenAI 建立合作关系，并于 2023 年 9 月全面推出 AI 助手，可供财务顾问快速访问公司知识资源。
2024/6/10	苹果	消费电子	在 iOS、iPadOS 和 macOS 中集成 ChatGPT，支持 Siri 调用 ChatGPT；在苹果的系统级写作工具集成 GPT，辅助用户生成各种主题的文字或图片。

资料来源：埃森哲官网，摩根士丹利官网，OpenAI 官网，国联民生证券研究所整理

**综合来看，前期市场对算力需求的质疑更多聚焦在 OpenAI、英伟达、Oracle 三者内循环的商业模式，而 OpenAI 在企业、政府、消费者领域的对外合作，则验证了 OpenAI 同样具备外循环的能力，进而减少了市场对大模型 ROI 的质疑。我们认为，伴随 OpenAI 内循环和外循环体系的逐步稳固，算力需求有望进入正反馈阶段，维持快速增长的态势。**

## 1.4 ROI 及现金流：AI 商业闭环成为 Capex 投资重心

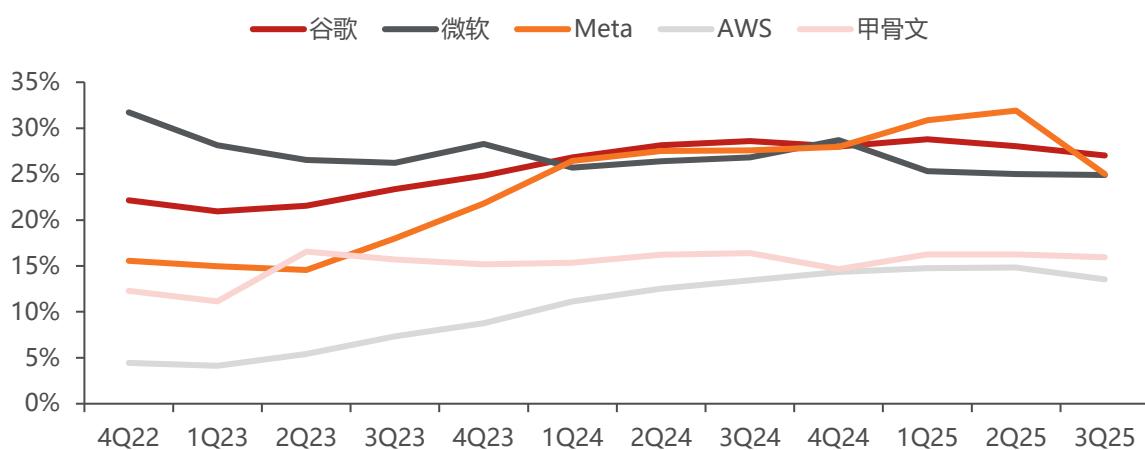
在过去的 AI 发展初期，由于“缺卡”导致的算力短缺，市场更关注 CSP 的资本开支扩展速度，资本开支投入快或意味着潜在的回报就高，对 ROI 及现金流情况较为宽容；现在市场则更加注重 AI 商业模式的闭环，尤其是在当前“AI 泡沫论”的热烈讨论下，交出令市场满意的财务数据从而打破质疑尤为关键，Capex 投入、Token 增长以及业绩兑现形成正循环成为重中之重。

3Q25 谷歌/微软/Meta/亚马逊/甲骨文 ROIC 分别达 27.02%/24.90%/25.02%/13.52%/15.94%，qoq-1.00/-0.12/-6.89/-1.31/-0.30pts，高增的资本开支对各厂商 ROIC 形成一定拖累。具体来看，**Meta 在保持了两年的良好 ROI 后，本季度环比显著下滑，一定程度引发市场对其 Capex 扩张较为激进的担忧；谷歌 ROIC 显著高于其他云厂商，其 TPU 显著降低 AI 投资成本功不可没。**

我们用 Capex/经营现金流衡量公司可持续投资的能力，3Q25 微软/谷歌/Meta/亚马逊/甲骨文 Capex/经营现金流分别为 77%/49%/63%/96%/104%；qoq+20.71/-31.42/-1.93/-0.19/-43.03pts；微软与谷歌该比率长期位于 50%以下（个别季度波动除外），表明其稳健的经营现金流增长可以较好支持资本开支投入；Meta 该比率近几个季度由 40%左右提高至 60%以上，资本开支投入未能同步促进经营现金流增长；而亚马逊及甲骨文则来到 100%左右，亚马逊因为其独特的电商模式拥有大量应付账款，对 Capex 提供一定支持，历史上该比率一直相对较高，甲骨文 Capex 已经超过经营现金流，面临较大的经营压力，需要通过外部资金解决现金流问题。

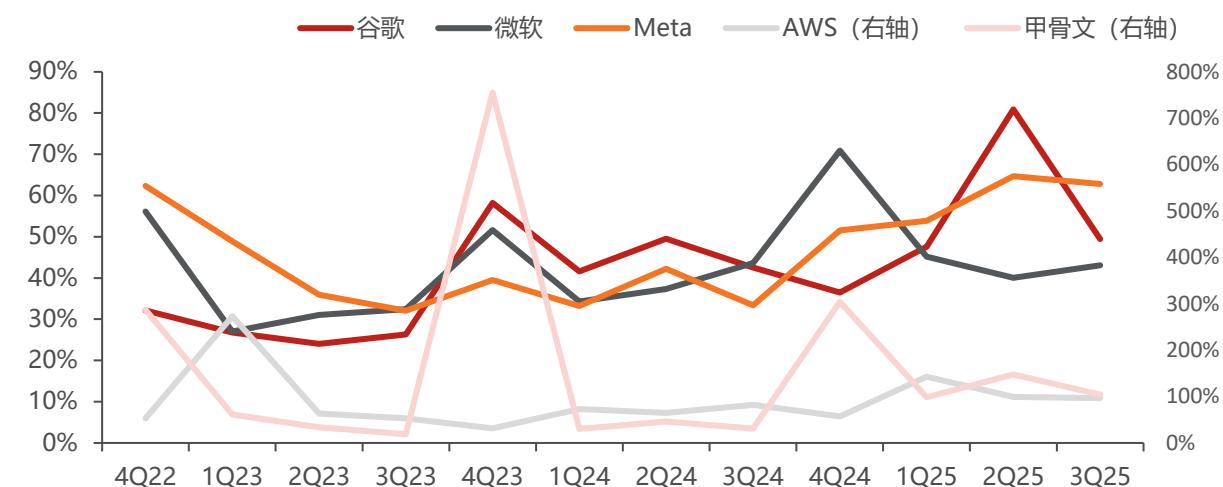
综合 ROI 及现金流情况来看，谷歌现金流充沛，健康的财务指标支持其进一步上修 AI 资本开支，同时 ROI 稳定保持第一，AI 投资回报率高，Capex 投入对经营现金流正向推动明显。**公司 Capex 投入 (TPU 显著降低投资成本)、Token 增长 (Gemini 模型表现优异)、业绩兑现 (AI 推动的 Google Cloud 及广告业务收入增长) 形成正循环。**

图15：4Q22-3Q25 北美云厂商 ROIC



资料来源：Bloomberg，国联民生证券研究所

图16：4Q22-3Q25 北美云厂商资本开支/经营现金流



资料来源：Bloomberg，国联民生证券研究所

## 1.5 主权 AI 是可观增量

主权 AI 的核心是本地部署 AI，依托于本土算力基础设施、数据、人才等，实现技术与应用的全链条自主可控。主权 AI 并非“闭门造车”，其核心是依托本土基础设施、数据、模型和人才，自主开发、部署 AI 并掌控全生命周期的能力，电子制造、核心芯片、电子算力等领域，并提供兼具安全性与实用性的、适配本土需求的智能解决方案，从而保持对 AI 的控制。

表 7：主权 AI 与商业云厂商对比

对比领域	核心电子技术	适配场景	算力部署
主权 AI	聚焦芯片等“卡脖子”技术突破，如本土自主 AI 芯片、自主算力调度系统，摆脱外部依赖。	适配电子领域中关乎国计民生、国家安全的关键场景。	依托本土基建，算力调度优先保障战略电子场景，抗地缘风险能力强。
商业云厂商	依赖开源电子技术或外部芯片，优先保障商业化适配，但核心电子技术自主可控性弱。	聚焦电子领域市场化、高效化的通用商业场景。	全球化电子算力布局，优先保障商业流量，易受地缘政治影响。

资料来源：中国新闻网，人民网，央广网，国联民生证券研究所整理

在主权 AI 重要性日益提升的整体态势下，目前全球已有多个国家进行了主 AI 的布局，以下为典型国家主权 AI 项目：

美国特朗普政府于 2025 年 1 月官宣了国家级 AI 基础设施战略工程“星际之门”(Stargate) 项目，美国联邦政府为此主权 AI 项目主导部门。该项目的核心支持企业为 OpenAI、甲骨文和软银集团，同时，微软、英伟达等企业为该项目提供能源设施、高端 GPU、芯片代工等相关支持。项目总投资规划达 5000 亿美

元，分四年投放，首期 1000 亿美元已投入，项目计划容量接近 7 千兆瓦。在应用领域方面，该项目不仅着眼于军事安全，更广泛涉及产业升级、社会服务与全球经济。

图17：蛋白结构预测式 AI



资料来源：英伟达官网，国联民生证券研究所

**欧盟于 2025 年 4 月提出“人工智能大陆行动计划”，该计划以欧盟委员会为主导，基于 InvestAI 倡议，拟投入 2000 亿欧元资金（欧盟投入 500 亿欧元，提供者、投资者和行业投入 1500 亿），其中专门设立 200 亿欧元用于建设 AI 超级工厂；其中《云与人工智能发展法案》将简化审批流程。该计划旨在大力建设 AI 算力基础设施、革新医疗体系、激发科研创新等，从而提升欧盟整体竞争力。我们认为，欧盟“人工智能大陆行动计划”在目前的建设阶段仍面临资金人才匮乏等挑战。**

**沙特阿拉伯打造国家级主权 AI 基础设施，形成“算力基建 + 场景落地 + 生态赋能”的立体化布局。**沙特公共投资基金（PIF）借助其旗下的 HUMAIN 公司拟依托数十万块英伟达 GPU 打造总算力达 500 兆瓦的超级算力集群，目前项目第一阶段已落地由 18000 块 NVIDIA GB300 Grace Blackwell GPU 驱动的 AI 超级计算机，完成核心算力底座的初步搭建。此外，沙特数据和 AI 管理局依托 5000 块 Blackwell GPU 资源，重点推进智慧城市解决方案研发，并面向科研人员开展 AI 模型开发专项培训，助力能源、制造等支柱产业智能化升级，为沙特数字经济高质量发展注入核心动力。

表 8：不同国家主权 AI 项目介绍

国家	主权 AI 项目	时间	投入	通用大模型依托	应用领域	主导部门
美国	星际之门计划	2025 年 1 月	5000 亿美元	OpenAI	为美国建设支持人工智能发展的基础设施	特朗普政府、微软、OpenAI、甲骨文、软银、英伟达
欧盟	人工智能大陆行动计划	2025 年 4 月	拟投入 2000 亿欧元	-	大力建设算力基础设施，革新医疗体系，激发科研创新，加强人才储备	欧盟委员会、EuroHPC 联盟
沙特阿拉伯	沙特阿拉伯 AI 工厂项目	2025 年 5 月	预计总投资 1000 亿美元	OpenAI	通过数字孪生技术优化制造业、物流和能源行业，助力沙特迈向工业化 4.0	公共投资基金 (PIF)、国家 AI 公司 HUMAIN、沙特数据和人工智能局 (SDAIA)、英伟达、AMD、高通
马来西亚	马来西亚国家级人工智能基础设施战略	2025 年 5 月	59 亿马币	DeepSee	提升公共服务与多领域竞争力，推动马来西亚向数字主权和包容性经济转型	国家人工智能办公室 (NAIO)、通讯与多媒体委员会 (MCMC)、多媒体大学 (MMU)、新兴科技伦理卓越中心 (CEET)
新加坡	Sea-Lion 大语言模型项目	2023 年启动	5200 万美元	通义千问	填补东南亚多语言 AI 鸿沟，解决全球主流 AI 模型以英语为中心，难以适配东南亚语言、日常语码转换等复杂场景的问题	新加坡国家人工智能计划 (AI.SG)、新加坡国家研究基金会 (NRF)、阿里巴巴 (阿里云)

资料来源：人民邮电报，中国科学院，环球网等，国联民生证券研究所整理

### 东南亚国家主权 AI 建设逐渐更加依托于中国大语言模型解决方案。

1) 马来西亚于 2025 年 5 月启动东南亚首个主权全自主全栈 AI 生态项目，由国家人工智能办公室 (NAIO) 统筹规划，包括制定《国家人工智能行动计划 2030》和出台 AI 治理框架。马来西亚通讯与多媒体委员会 (MCMC) 将投资约 20 亿马币 (约 31 亿元人民币) 建设“主权 AI 云” (Sovereign AI Cloud)，并与多媒体大学 (MMU) 及新兴科技伦理卓越中心 (CEET) 合作建立 AI 转型中心。于资金端构建“政府财政 + 私营资本 + 国际投资”多元供给模式，配套税收减免等激励政策；应用端聚焦制造、医疗、公共服务等领域，通过模型本地化部署与智能工厂培育，实现技术主权掌控与产业升级双重目标。

图18：马来西亚宣布启动国家级 AI 基础设施战略



资料来源：马来西亚华文理事会，国联民生证券研究所

图19：新加坡 Sea - Lion 模型概念图



资料来源：阿里云官网，国联民生证券研究所

## 2) 2023年12月新加坡启动开发适配东南亚多语言环境的Sea - Lion模型。

Sea - Lion 模型早期基于美国 Meta 的 Llama2 开发的版本，因英语中心主义的缺陷，在东南亚语言处理和区域常识上问题频发。项目由新加坡国家研究基金会 (NRF) 资助，总投资 5200 万美元。2025 年 11 月 24 日该项目全面改用阿里通义千问 Qwen3 - 32B 作为基座模型，新版本依托 Qwen3 覆盖 119 种语言和方言的优势，结合东南亚 1000 亿个专属语言 token 训练，能精准处理泰语、新加坡式英语等本地化语言场景，填补东南亚多语言 AI 鸿沟。

**全球 AI 资本支出高速增长，国家主权 AI 需求成为其重要驱动之一。**主权 AI 对全球 AI 产业的投资拉动作用将持续深化，其在全球 AI 总支出中的占比有望实现稳步提升。

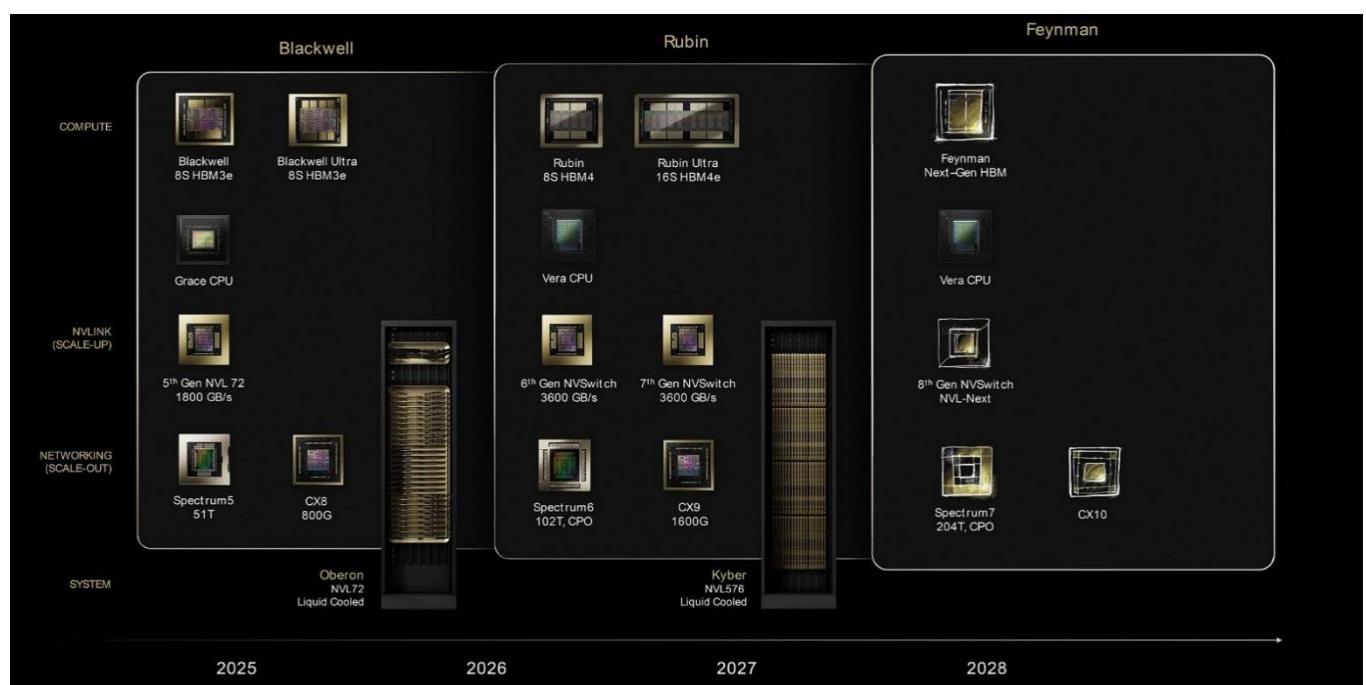
## 2 算力芯片和服务器——大模型的底座

### 2.1 英伟达路线图

#### 2.1.1 加速卡和机柜全面升级

当前英伟达主力芯片从 H 系列逐步升级为 B 系列和 R 系列，B 系列和 R 系列分别对应英伟达 2025 年和 2026 年的标杆产品，算力、存储、带宽方面均获得大幅度提升。GB200 加速卡 2H24 开始放量；2H25 GB300 开始放量；公司预计 2H26 Rubin 系列产品开始出货。

图20：英伟达加速卡升级路线图



资料来源：英伟达官网，国联民生证券研究所

**AI 服务器形态上，英伟达引领了 AI 服务器从 8 卡到 72 卡再到 576 卡的升级路线，带动了算力密度的快速提升。**

**72 卡机柜：**从 GB200 开始，英伟达的主流出货方案转变为 NVL 72 机柜，机柜内部采用高速铜缆互联，单机柜算力密度大幅提升；

**144/576 卡机柜：**Rubin 的第一代产品仍然沿用 Blackwell 的机柜架构，但 GPU 的计数方式从单芯片（两颗 die）转变成了一颗 die 作为一个 GPU，衍生出 NVL 144 机柜，而 Rubin Ultra 开始采用的 NVL 576 则在一个机柜中配置 144 颗芯片，576 颗 die，机柜密度得到了进一步提升。

图21：英伟达机柜升级路线图



资料来源：英伟达官网，国联民生证券研究所

## 2.1.2 Scale-up 互联升级趋势

机柜内部 Scale-up 的升级趋势主要围绕更多 GPU 和更大互联带宽两个目标。

**1) 更高性能的互联方案：**英伟达 8 卡机柜 scale up 采用 PCB 互联，GB200 NVL72 开始采用 DAC，Rubin Ultra NVL576 早期方案采用正交背板+铜连接器方案，而后续的设计方案中也可能考虑将铜互联升级为光通信，从而获取更高性能；

**2) 更大的 GPU 互联数量：**在英伟达的 AI 服务器组网中，Scale-up 与 Scale-out 带宽相差 4.5 倍，因此选择合适的 Scale-up 层 GPU 数量有助于达到最高性价比的组网效果。从 8 卡升级到 576 卡机柜，Scale-up 层的 GPU 数量不断提升，但仍局限于机柜内部，未来 Scale-up 组网可能拓展至多台机柜间，互联 GPU 数量有望进一步提升。

图22：英伟达 Scale-up 升级路线



资料来源：英伟达官网，国联民生证券研究所整理

## 2.2 ASIC 路线图：自研 AI 算力芯片加速迭代

我们认为，云厂商自研加速卡将成为未来AI芯片增量最核心的来源。一方面，云商自研加速卡在成本方面显著优于向英伟达等商业公司外采，3Q24 英伟达毛利率已达到 74%，采用自研加速卡的方式，将帮助云商在有限的资本开支下获得更多的AI算力。另一方面，云商自研 ASIC 更加灵活，云厂商可以根据自身的模型训练和推理需求，进行AI芯片和服务器架构的设计，从而实现更好的训练和推理效果。伴随着云厂商自研 ASIC 产品的逐步成熟，未来云商在AI算力的布局中自研的比例有望逐步提升。

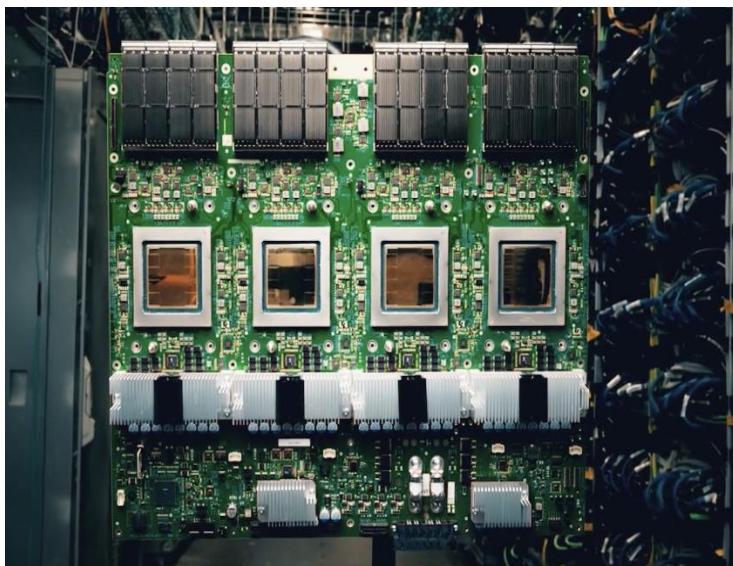
图23：CSP 厂商在 ASIC 领域的路线图

厂商	型号	发布时间	制程 nm	峰值算力 TOPS/TFLOPS			内存信息		互联带宽 GB/s
				INT8/FP8 Dense/Sparse	BF16/FP16 Dense/Sparse	TF32/FP32 Dense/Sparse	类型	容量 GB	
谷歌	TPUv5E	2023	5	394	197	-	HBM2	16	400
	TPUv5P	2023	5	918	459	-	HBM2	95	800
	TPU v6e (Trillium)	2024	4	1836	918	-	HBM3	32	1640
	TPU v7 (Ironwood)	2025	3	4616	-	-	HBM3	192	7372
Meta	MTIA v2	2024	5	354/708	177/354	2.76	-	128	-
微软	Maia 100	2023	5	1600	-	-	HBM3	64	1200
亚马逊	Trainium2	2024	-	1299	667	181	HBM3E	96	-
	Trainium3	2025	3	2517	671	183	HBM3E	144	-

资料来源：各公司官网，Semianalysis，EET，国联民生证券研究所

注：未标注的数据为没有在公开渠道披露的信息

**谷歌在北美云厂商中采取以自研训练型加速卡为核心、外采英伟达加速卡为补充的算力策略，对 GPU 依赖度低。**谷歌自 2013 年开始研发 TPU，相较其他云厂商具备近 10 年的技术与生态积累优势。2023 年 12 月，谷歌推出面向云端大模型训练的 TPU v5p，相较 TPU v4，浮点算力提升约 2 倍、内存带宽提升约 2 倍；在集群层面，TPU v5p Pod 由 8,960 颗芯片组成，单芯片互联带宽达 1200 Gbps。2024–2025 年，谷歌进一步推出 TPU v6e 与 TPU v7 (Ironwood)，训练型 ASIC 路线持续演进。谷歌下一代产品 TPUv8 亦在研发中，预计将分为“Sunfish”及“Zebrafish”两个版，TPUv8ax “Sunfish”主要针对 Gemini 等大型模型训练侧，而 TPUv8x “Zebrafish”则主要针对推理侧场景。

**图24：谷歌 TPU V7 (Ironwood)**

资料来源：EET，国联民生证券研究所

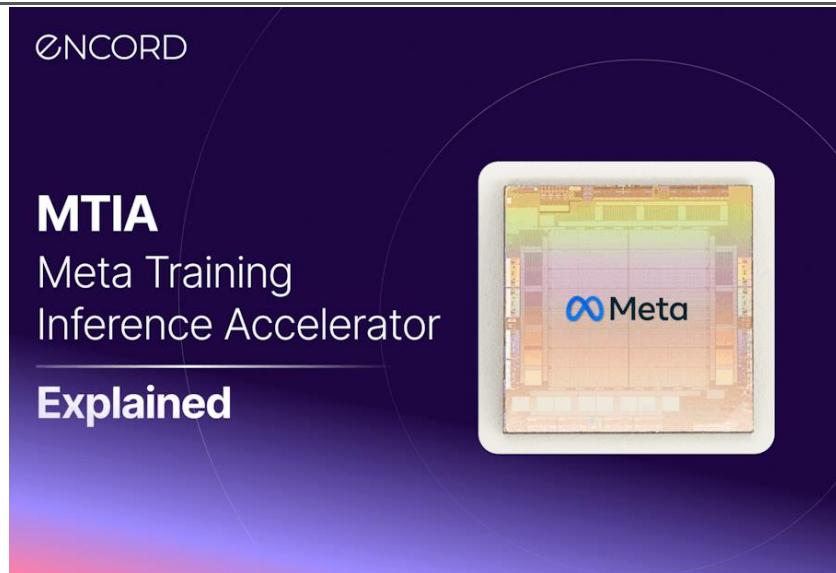
**亚马逊通过自研加速卡构建训练与推理并行推进的算力体系。**亚马逊持续加大在自研 AI 加速卡领域的投入。2024 年 12 月，公司正式推出用于模型训练的 Trainium2 加速卡，补全其 AI 算力布局。Trainium2 的性能显著增强，且架构设计填补了上一代芯片的不足，通过 650 TFLOP/s 的计算能力和 96GB 的 HBM3 内存支持，面向大规模生成式 AI 模型训练与推理。2025 年 12 月，亚马逊发布 Trainium3，采用 3 纳米工艺，提供 2.52 PFLOPs 的 FP8 算力，性能较 Trainium2 提升约 4.4 倍，能效提升 40%，单集群扩展规模达百万颗芯片，旨在显著降低 AI 模型的训练与推理成本。

**图25：亚马逊发布 Trainium3**

资料来源：techpowerup，国联民生证券研究所

Meta 自 2021 年以来将企业发展重心转向元宇宙与人工智能，并持续加大在 AI 基础设施领域的投入。2023 年，Meta 首次推出自研 AI 加速卡 MTIA v1；**2024 年 4 月，Meta 发布 MTIA v2**，新一代产品在算力、内存容量和内存带宽方面表现更优，采用台积电 5nm 工艺，INT8 稀疏算力达 708 TOPS，LPDDR5 内存容量提升至 128GB，主要面向大规模推理场景。目前 Meta 在 LLaMA 等大模型训练中仍主要依赖英伟达加速卡，后续自研芯片的应用重心仍以推理侧为主。

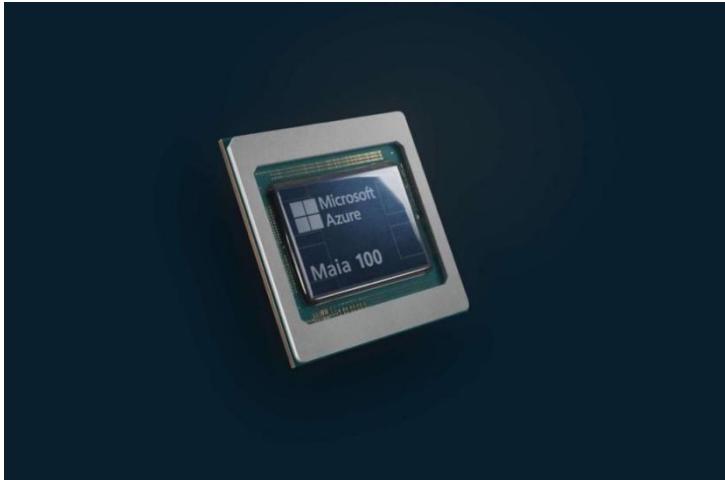
图26：Meta MTIA



资料来源：Encord，国联民生证券研究所

微软 Azure 平台服务的企业客户数量已超过 25 万家，在高端算力需求持续扩张的背景下，公司同步推进自研 AI 加速卡计划。**微软于 2023 年推出自研训练加速卡 Maia 100，专为 Azure 云服务设计。**Maia 100 采用台积电 5nm 工艺，单芯片集成约 1,050 亿个晶体管，FP8 峰值算力达 1,600 TFLOPS，并支持 FP4 运算，峰值算力可达 3,200 TFLOP。

图27：微软 Maia 100



资料来源：TECHradar，国联民生证券研究所

## 2.3 工业富联——AI 算力领军供应商

**AI 算力垂直整合领军者和基础设施架构师。**工业富联成立于 2015 年，是全球领先的高端智能制造和工业互联网解决方案供应商，业务包含云计算、通信及移动网络设备和工业互联网。公司云计算产品涵盖高性能服务器、边缘计算、先进散热和储存设备，凭借液冷技术和精密制造优势，成为英伟达 GB200 的垂直整合供应商，具备 L11（机柜级集成）交付能力。随着谷歌 TPU 和其它云厂商自研芯片迭代，ASIC 服务器市场迎来新一轮增长。2025 年 11 月，公司与 OpenAI 达成合作，共同打造多代 AI 数据中心机架，生产布线、网络、散热和电力系统相关设备。

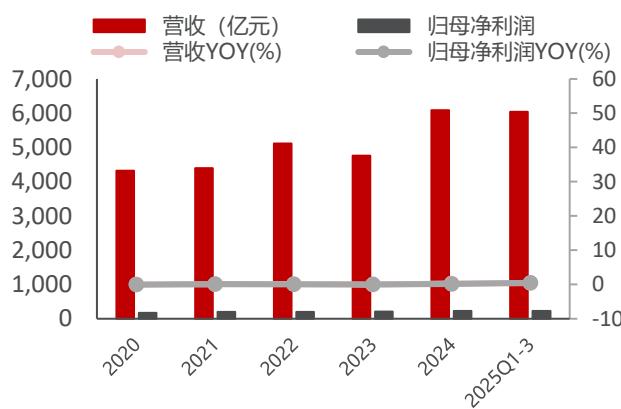
图28：工业富联产品系列布局



资料来源：工业富联官网，国联民生证券研究所

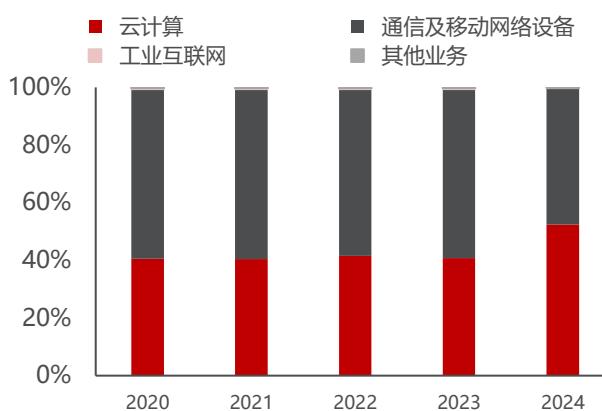
2020-2025 年前三季度，公司财务数据呈现稳步增长—盈利承压—业绩增长的阶段性特征。2020 年-2022 年公司依托通信及移动网络设备稳步增长，2023 年公司主动收缩传统业务，全力押注高附加值 AI 服务器，营收同比下滑 6.9%，净利润却逆势增长 4.8%。2024 年大模型迭代和 AI 技术的规模化应用驱动算力需求高速增长，公司深度参与全球 AI 算力产业链，进入业绩高速增长期，云计算开始成为核心增长极，收入占比首次超过五成。25 年前三季度营收同比激增 38%，突破 6000 亿人民币，净利润增长 48%显著提升，受益于 AI 机柜产品的规模交付，云计算业务营收同比增长 65%。云服务商业务占云计算业务 7 成，表现亮眼，营收同比增速超过 150%，其中 GPU AI 服务器同比增速超过 300%。通信和移动网络设备板块表现稳健：精密机构件受 AI 智能终端新品热销，带动业务增长；交换机业务受 AI 需求持续放量，800G 交换机成长显著。

图29：2020-2025年Q3工业富联营收和利润情况



资料来源：ifind，国联民生证券研究所

图30：2020-2024年工业富联营收结构



资料来源：ifind，国联民生证券研究所

## 2.4 速率+功率，算力产业的机遇与挑战

前文，我们主要讨论了算力芯片的变革。算力需求高增的背景下，英伟达产品加速迭代，而 CSP 自研的 ASIC 则迎来了更快的成长。

展望未来，我们认为，算力的升级要更为侧重“速率+功率”两大赛道。**我们于 25 年初的 AIDC 系列深度中首次提及“速率+功率”，指明二者为未来 AI 产业发展的核心矛盾。**在过去的一年内，无论是速率赛道的光+PCB，还是功率赛道中的电源+液冷。都走出了“波澜壮阔”的行情。足以证明我们观点的前瞻与正确。

**那么站在当下，我们怎么看未来一年的算力机遇？**目前市场对 CSP 厂商的资本开支始终有所疑虑，担心 ROI，担心远期增量不明朗。我们认为，在对 AI 产业保持乐观的同时，要重点观察 CSP 及大模型厂商的商业闭环节奏，从而把握整体行业β。同时，寻找**价值量扩张、资本开支增量倾斜**的细分赛道，**主线延续“速率+功率”。**

**1) 速率，数量的海量增长，急需解决互联瓶颈。**从光模块到 CPO，从 DAC 到 AEC，以及 PCB 的材料和用途创新，都是业内在解决速率问题上的技术演进，将成为速率产业升级的重要组成部分。

**光，把握光入柜内的趋势，**抓住光模块的业绩线、光芯片的缺货潮、硅光的渗透率提升趋势。**关注超节点技术带来的 OCS 等产业趋势。**

**PCB，**规格向更高层数、更高阶 HDI 发展，材料向 M9 升级，并有 PTFE 等潜在新材料在测试中。未来 2-3 年伴随 midplane、正交背板、CoWoP 等新方案的推出，PCB 作为解决速率瓶颈的关键环节，其加工壁垒及价值量有望显著提高。

**2) 功率，功耗提升，带来供电及温控需求。**AI 浪潮下，算力芯片单芯片功耗 (TDP) 快速提高，同时高密度计算要求采用机柜架构，单机柜功率密度提高，对温控和电源系统提出挑战。伴随摩尔定律放缓，制程升级迭代延后，功耗墙成为挡在高算力需求前面的“拦路虎”。**数据中心建设也会随之向液冷+电源倾斜。**

**电源，**单卡和机柜功率密度持续提升，对电力架构提出了新的要求，传统电力架构在效率、可扩展性、稳定性方面已无法满足下一代 AI 基础设施的需求，HVDC 等电力设备升级成为行业必然趋势。

**液冷，**伴随芯片功率提高，传统风冷无法满足散热需求，液冷从 0-1 实现产业趋势突破，目前液冷产业链目前仍以台系+海外厂商为主，但诸多国内本土的液冷厂商正展现出强劲的发展态势。无论技术能力、交付能力、项目经验等方面均可向全球龙头厂家看齐。

### 3 光：算力时代的核心破局点

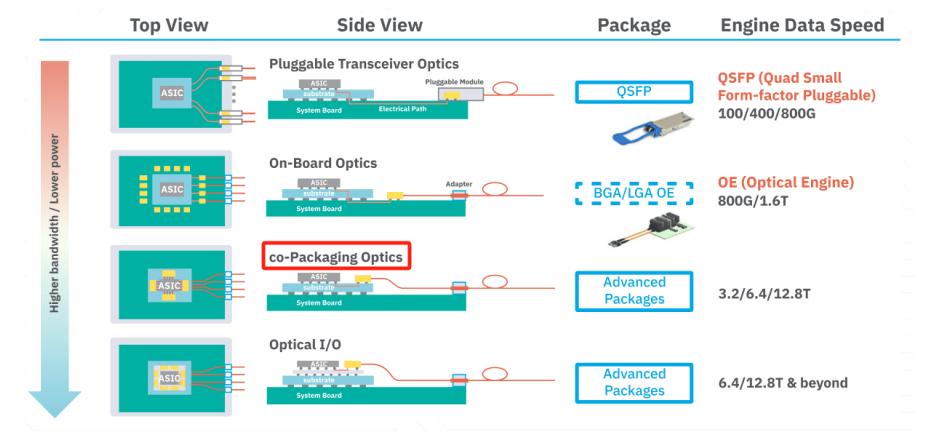
在数据中心 Scale-up 和 Scale-out 带宽升级的过程中，光通信模式的迭代成为贯穿速率演变过程的核心要素。过去光通信主要集中在 Scale-out 环节，而伴随 CPO、NPO 等技术的逐步成熟，Scale-up 也开始出现光通信的身影。OCS 全光交换则是光通信升级的领域核心路径，其颠覆了传统电交换的方式，从而解决了数据中心交换过程中的速率、功耗等核心瓶颈。目前来看，光通信的迭代已经成为算力时代的核心破局点，成为全球主流算力厂商必争的创新高地。

#### 3.1 光通信的下一站：NPO+CPO

##### 3.1.1 CPO 技术演进及封装结构设计探析

CPO 是光引擎和交换芯片共同封装在一起的光电共封装，没有采用可插拔光模块的形式，响应数据中心光模块降耗趋势。技术的核心逻辑在于推动光模块与交换芯片的“持续靠近”。它通过逐步缩短芯片与模块之间的走线距离，最终实现光引擎与电交换芯片的一体化封装，形成一个集成度更高的芯片单元。这种渐进式的集成路径具有明确的替代潜力——在理想场景下，CPO 技术将逐步替代传统的可插拔光模块，通过将硅光子模块与超大规模 CMOS 芯片以更紧密的形态实现封装整合，从而带来系统性的性能跃升。

图31：CPO 指把光引擎和交换芯片共同封装在一起的光电共封装



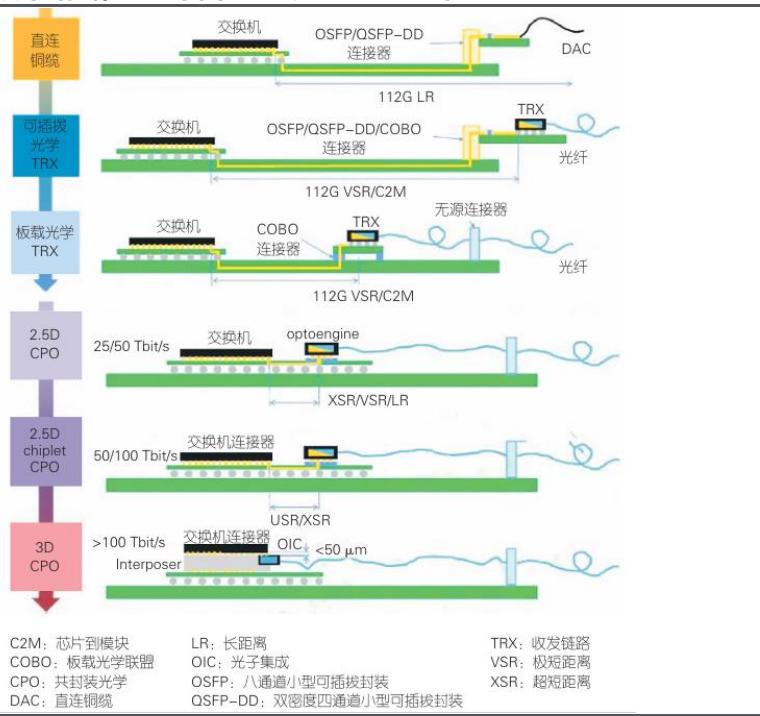
资料来源：ASE 官网，国联民生证券研究所

按照物理结构，CPO 可分为 3 种技术形态。

- 1) **2D 平面 CPO：**基于 2D 封装的 CPO 技术是将光子集成电路 PIC 和集成电路并排放置在基板或 PCB 上，通过引线或基板布线实现互连。
- 2) **2.5DCPO：**2.5D 封装将 EIC 和 PIC 均倒装在中介层 (Interposer) 上。通过中介层上的金属互连 PIC 和 EIC，中介层与下方的封装基板或 PCB 板相连。
- 3) **3DCPO：**3D 封装技术将光电芯片进行垂直互连，在封装过程中会用到硅

穿孔 (TSV)、凸点 (Bumping) 和重布线 (RDL) 等先进封装技术，对传统光模块封装厂商提出新的挑战。

**图32：CPO 技术路线，2D 平面 CPO、2.5DCPO 和 3DCPO**



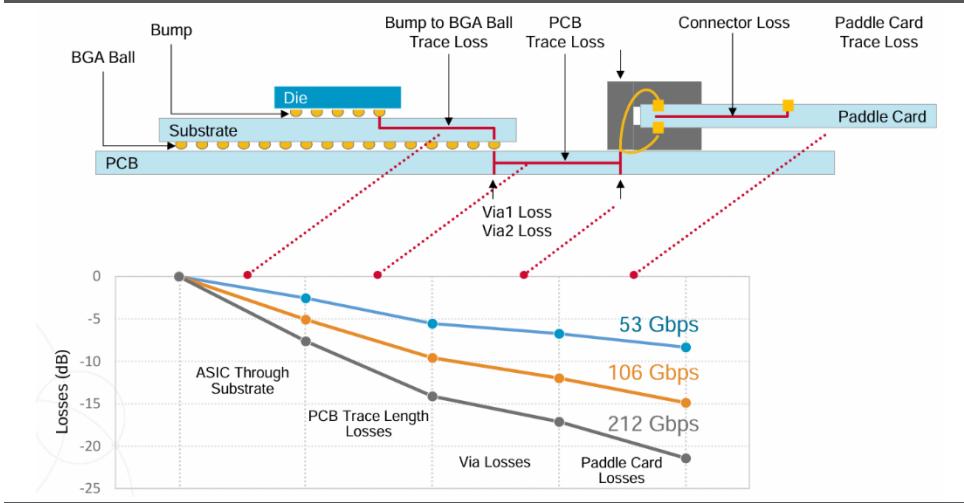
资料来源：张平化等《数据中心光模块技术及演进》，国联民生证券研究所

### 3.1.2 CPO 是破解功耗与速率瓶颈的关键技术

**光通信传输损耗随速率增加而增大，算力驱动带宽激增下功耗问题成下一代高速光互联最大挑战。**光通信传输路径上，基板线路、PCB 线路、通孔、光模块板卡都会产生一定损耗，且传输速率越大损耗越大。当下，算力需求提升带动网络带宽大幅增加，Cisco 数据显示，2010—2022 年全球数据中心的网络交换带宽提升了 80 倍，同时交换芯片功耗增加约 8 倍，光模块功耗增加 26 倍，交换芯片串行器/解串器 (SerDes) 功耗增加 25 倍。此背景下，如何解决功耗问题成为下一代高速光互联应用的最大挑战。

**CPO 通过减少组件和互连器件数量降低成本，提升传输速率，并节省 30% 功耗。**在光模块降耗的发展趋势下，行业围绕驱动器、调制器、激光器及电接口 4 个方面降低功耗，而 CPO 可以在电接口方面缩短交换芯片和光引擎之间的连接长度，直接实现电-光信号转换，显著提升传输速率。CPO 的另一主要优点是能够最大限度地减少所需的组件和互连器件数量，从而大幅降低单位比特成本。此外，与可插拔光模块器件相比，CPO 通过消除电互连功率耗散和变异性，提供了更优越的功率和性能特性。博通官网显示，CPO 能节省 30% 功耗，将每比特光学成本降低 40%，并可支持 1Tbps/mm 带宽密度。

图33：基板线路、PCB 线路、通孔等产生一定损耗，且传输速率越大损耗越大

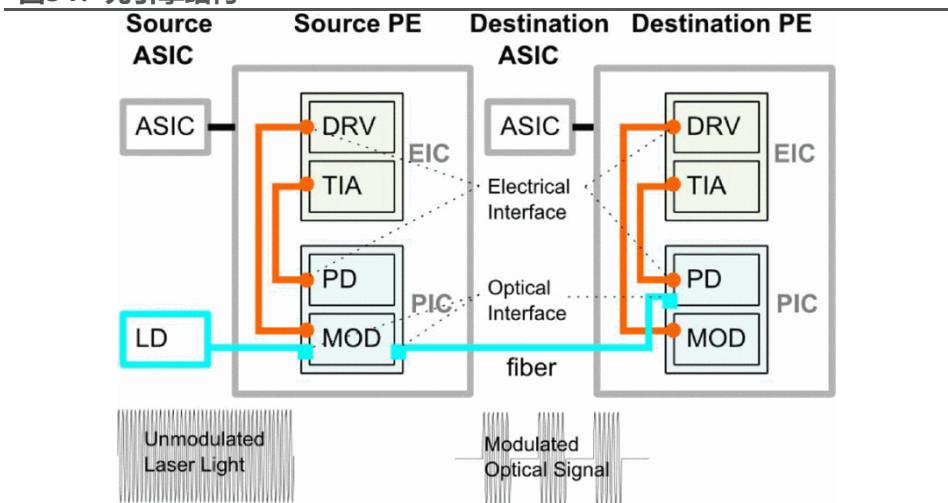


资料来源：博通官网，国联民生证券研究所

### 3.1.3 光引擎和硅光集成：CPO 技术核心

将电信号转换为光信号的高性能光引擎始终是硅光技术的核心。光引擎由光子集成电路 (PIC) 和电子集成电路 (EIC) 组成，这二者通过电气接口相连，光引擎通过其光接口-光纤耦合器接收和传输光信号。CPO 的兴起深刻改变着光引擎的设计理念和技术路径，随着带宽需求的增加，光引擎逐渐靠近 ASIC，以减少铜线的功率损耗和信号衰减，光引擎也从传统热插拔中的分立器件，转换为与电芯片紧密协同的高度集成化光电接口。

图34：光引擎结构



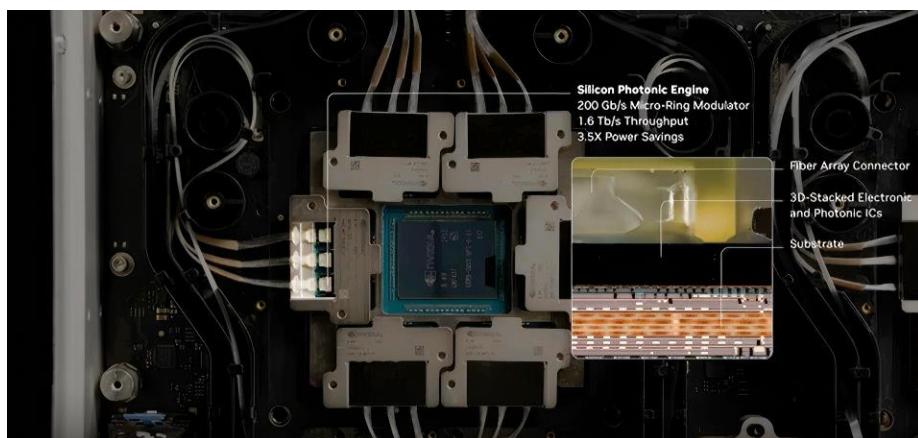
资料来源：H. Hsia et al., "Heterogeneous Integration of a Compact Universal Photonic Engine for Silicon Photonics Applications in HPC," 2021 IEEE 71st Electronic Components and Technology Conference (ECTC)，国联民生证券研究所

PIC 是光引擎的核心光学处理单元，其通过在半导体基片（如硅、磷化铟）上集成激光器、调制器、探测器等光学器件，实现光信号的产生、传输、调制或探测。

EIC 负责驱动 PIC 中的电光调制器、放大探测器输出的电信号，并提供数字信号处理等功能。

**硅基光电子的方案由于具有与成熟的 CMOS 工艺兼容的优势，成为 CPO 光引擎的主流解决方案。**尽管 CPO 应用需要定制结构，但 CPO 芯片的主要制造挑战来自光纤耦合和光源集成。片上光源集成是硅光子学的主要挑战之一，硅基材料本身很难形成高性能激光器，硅光集成技术作为 CPO 的核心使能基础，正在经历制造工艺加速成熟，技术线路融合等一系列升级进程。

图35：硅光引擎



资料来源：英伟达，国联民生证券研究所

### 3.1.4 CPO 驱动 MPO 向更高密度方向演进

**MPO 连接器是支持高速光互联的关键器件，在数据中心布线中扮演着日益重要的角色。** MPO (Multi-fiber Push On) 是一种高密度多芯光纤连接器，是当前主流的并行光传输解决方案，广泛应用于数据中心、5G 网络、云计算等对高速率、低延迟连接有较高要求的场景。与传统单芯 LC 光纤连接器相比，MPO 可在一个接口内容纳 8、12、24 甚至 72 芯光纤，大幅提升单位面积的数据传输能力，是实现 400G/800G 高速通信系统布线不可或缺的部件。其高速传输、高密度布线、低插入损耗和回波损耗的优良性能使其在数据中心布线中占据重要地位。

图36：MPO 连接器组件



资料来源：CablesAndKits，国联民生证券研究所

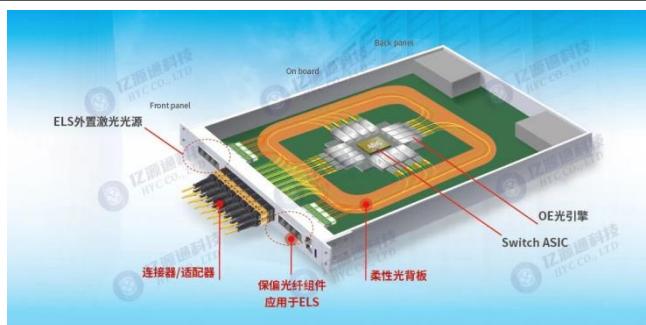
### 光电共封装（CPO）技术的兴起为 MPO 市场更新注入了全新的增长动能。

相较于传统依赖电信号传输的交换机架构，CPO 交换机内部需依靠光纤完成信号传递，外部激光源（ELS）目前是 CPO 光源最常用的解决方案之一，这显著提升了对高密度光纤连接器 MPO 的需求。同时，**CPO 交换机的引入带来了光纤数量激增，驱动 MPO 向更高密度方向演进。**以一台 51.2T 交换机为例，若按照单端口 100G 进行配置，总通道数可达 512，对应 1024 根光纤。若采用 16 芯 MPO 进行连接，则需配备 64 条光缆，等效于 64 个 MPO 接口，显著提升了对 MPO 相关产品的需求。**其中，MMC 创新性地将超小型 MT 型插芯（TMT）和超小型（VSFF）连接器相结合，是 CPO 进一步发展的可选方案。**

#### 3.1.5 Shuffle Box：CPO 的重要增量环节

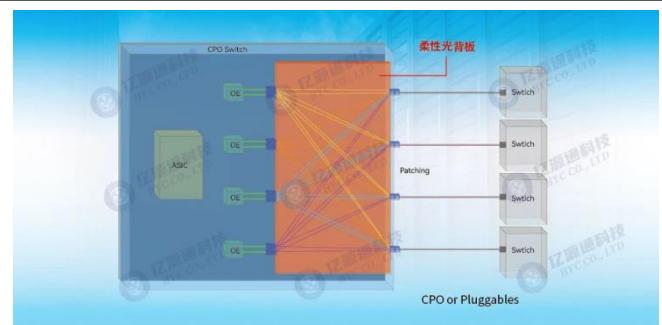
**CPO 布线复杂度不断上升，Shuffle Box 是解决这一问题的一个可行方案。**高速率 CPO 交换机的内部需要精密布局数千根光纤，不仅要解决内部空间狭小的问题，还需解决因板内各光引擎到前面板的距离存在差异、导致光纤长度不一致而引发的制造可靠性问题，这就需要采用光纤柔性光背板 Shuffle 的以光引擎到端面的连接方式解决上述问题。

图37：光纤连接系统图



资料来源：亿源通 HYC 公众号，国联民生证券研究所

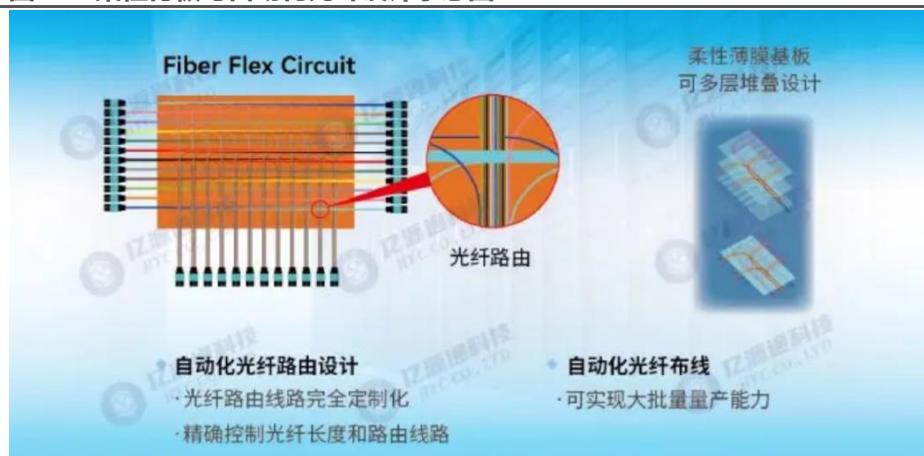
图38：CPO Switch 工作原理展示



资料来源：亿源通 HYC 公众号，国联民生证券研究所

**Shuffle Box 是一种光纤柔性光背板，能够很好地处理 CPO 的布线问题，是实现 CPO 更高密度光路布线的关键解。**Shuffle Box 使用的柔性光背板基于灵活薄膜基板设计，可自定义光纤路由以减少交叉应力，支持复杂信号通道路由，其 1U 空间能实现 600 芯光纤熔接分配，2m 高 40U 机柜总容量达 24000 芯，为常规光纤配线架方案的 20 倍以上。

图39：柔性背板与自动化光纤设计示意图



资料来源：亿源通 HYC 公众号，国联民生证券研究所

**Shuffle Box 在 CPO 交换机内部承担光信号拆分与路由分配的核心任务。**外部光纤通过 MPO 接口接入后，Shuffle Box 将单路信号拆分为多路（如英伟达 Quantum 3400 X800 中拆分为 4 路），分别传输至不同交换芯片。其高效的光信号处理能力不仅优化了数据传输路径，还显著提升了系统整体性能，确保了 CPO 架构的高效运行和稳定性。Shuffle Box 也支持多平面拓扑技术，允许多个独立数据平面并行运行，最终在网卡端汇聚，实现带宽密度提升与信号延迟降低。

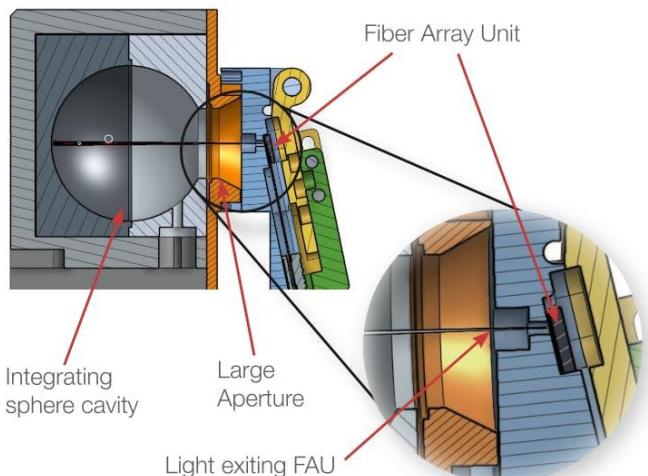
图40：英伟达 Quantum 3400 X800



资料来源：英伟达官网，国联民生证券研究所

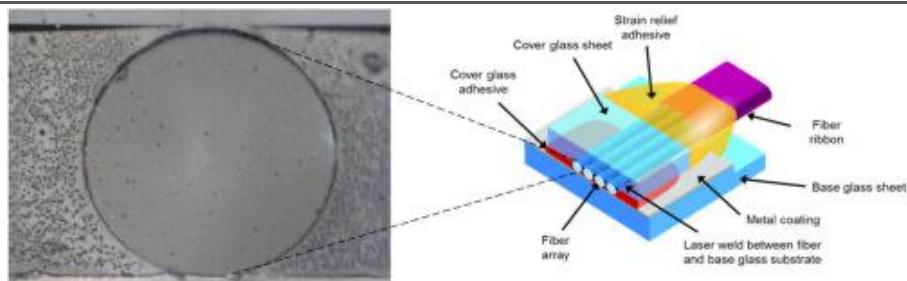
### 3.1.6 FAU 与微透镜阵列——CPO 架构中的核心光学组件

**光纤阵列单元 (FAU)** 是一种由多根光学光纤组成，以一致的间距排列在基板上，通常呈一排的精密对准工具。在 CPO 中，其主要作用是确保多根光纤与光子集成电路 (PIC) 或其他紧凑型光学器件的波导精确对准。在 CPO 系统中，FAU 对于将高通道数光收发器 (光子芯片) 连接到同一硅基板上的集成电路 (IC) 至关重要。它们能够将单模 (SM) 和偏振保持 (PM) 光纤阵列耦合到 PIC 的边缘或表面。

**图41：光纤阵列单元（FAU）的组装与光反射结构剖面图**

资料来源：Santec 官网，国联民生证券研究所

**FAU 在 CPO 系统中的优势在于其能够实现极其精确的光纤定位和低间距误差，从而达成最佳光耦合和高带宽。**这种高精度直接转化为 CPO 系统的卓越性能和可靠性。为了确保 FAU 的稳健性和 CPO 系统的长期寿命，行业正在积极探索和应用先进技术。例如，无胶连接方法，特别是近红外激光焊接，能够将光纤牢固地固定在平面玻璃基板上，形成坚固的玻璃-玻璃键合。这种创新不仅提高了制造精度，还消除了传统环氧树脂可能带来的问题，从而增强了 FAU 的耐用性。

**图42：光纤阵列单元（FAU）的组装与光反射结构剖面图**

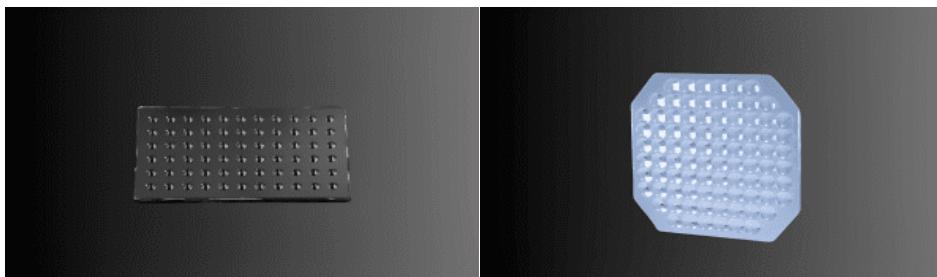
资料来源：ResearchGate，国联民生证券研究所

**微透镜阵列（MLA）对于光纤或波导与光子集成电路（PIC）之间的高效光耦合至关重要。**MLA 是一种微型光学器件，由许多微小的透镜（小透镜）组成，通常尺寸在几微米到几毫米之间，并以一维或二维阵列形式排列在基板上。每个小透镜单独处理光线的方式与大透镜类似，但它们协同工作以实现复杂的光场调控，例如光束变换、整形、准直和匀化。MLA 的特性由焦距、透射波前质量、尺寸、透镜间距和填充因子（可用透镜孔径总面积与阵列总面积之比）等参数来表征。

**目前市场上致力于研发先进 MLA 的企业层出不穷。**Edmund Optics 提供各种尺寸、间距和镀膜的微透镜阵列。AGC 提供以玻璃为基础的 MLA，以其高耐光性和稳定的热特性而闻名。博通得益于其在衍射和折射光学元件方面的科研投入，

也提供用于光纤子组件和收发器的折射式微透镜/阵列。

图43：矩形微透镜阵列（左）与六角形微透镜阵列（右）示意图

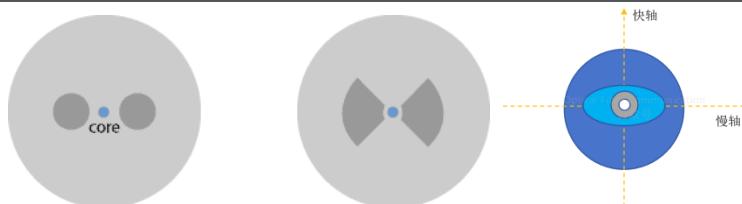


资料来源：Aventier，国联民生证券研究所

### 3.1.7 保偏光纤——CPO 先进光纤与连接器技术

保偏光纤（PMF），是一种单模光纤，设计用于在光线沿其双折射轴之一正确入射时，在传播过程中保持线性偏振。保偏光纤是一种在纤芯两侧引入高膨胀应力区、人为产生强双折射的单模光纤，其快慢轴传播常数差极大，可把线偏振光锁定在选定主轴上传输，即使遭受弯曲、侧压或温度扰动也能保持偏振态稳定，兼具高消光比、低串扰和对准敏感等特点。

图44：保偏 PANDA 光纤（左）、蝴蝶结光纤（中）椭圆形光纤（右）横截面图



资料来源：RP Photonic, Optical Fiber Communication, 国联民生证券研究所

CPO 需要保偏，因为 CPO 内部光学引擎和光子集成电路（PIC）对入射光偏振态具有高度敏感性。CPO 光学引擎的性能对入射光的偏振态高度敏感，需要稳定的激光偏振。而保偏光纤确保激光偏振态保持稳定，从而最大程度地减少失真并保持信号可靠性。CPO 系统需要激光源，可以是集成式的也可以是外部的。虽然集成式激光源提供了密度优势，但外部激光源（ELS）提供了更高的系统可靠性。PMF 对于将激光信号从外部源可靠地传输到 CPO 系统内的光子电路至关重要。博通专为 CPO 设计的 QSFP-DD ARLM-96F8DMZ 激光模块，其激光传输通道就需要偏振保持单模光纤。

**图45：博通 QSFP-DD ARLM-96F8DMZ 激光模块产品图**

ARLM-96F8DMZ: QSFP-DD 800 mW CWDM 激光模块 1×12 SM APC MPO  
连接器

**特征**

- 每通道 CWDM 100 mW 激光输出
- 基于 CPO JDF 标准的数字诊断监控双线串行 (TWS) 接口
- 拉片可轻松插入和提取模块
- 热插拔，易于安装和维修
- 0°C 至 45°C 外壳温度工作范围
- 经过验证的高可靠性 DFB 激光技术
- 1 级眼部安全

资料来源：Broadcom 官网，国联民生证券研究所

## 3.2 OCS 光交换机：全光互联时代的基石

### 3.2.1 OCS 通过全光交换带来性能突破

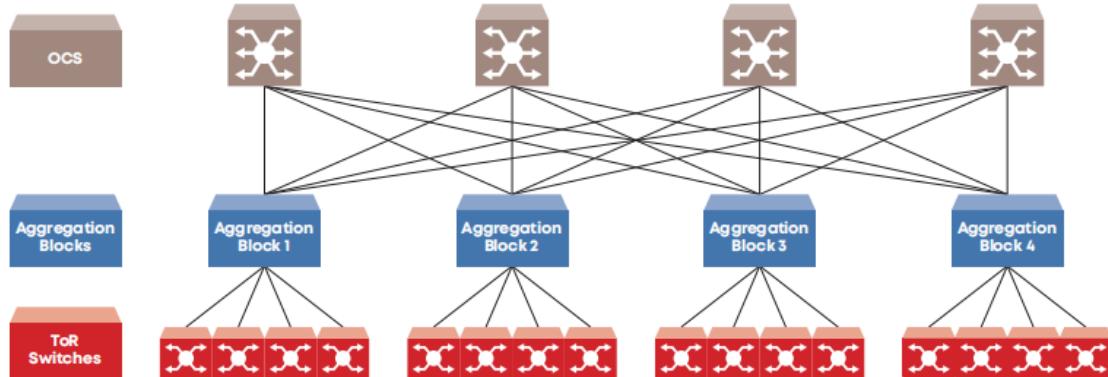
2025 年，谷歌大模型带给世界惊喜，自研芯片超算集群架构核心是 OCS。2025 年 11 月推出了 Gemini 3.0 Pro 是谷歌重新占据行业技术高点的关键节点。谷歌模型基于自研 TPU，依托谷歌自研 TPU 完成训练，可显著提升大模型训练速度并支持更大模型规模，谷歌第七代 TPU (Ironwood) 在算力、能效与带宽方面实现全面跃升，是公司当前最强的自研加速器。谷歌 Jupiter 数据中心网络和 TPU AI 超算集群的核心互联枢纽是 OCS 交换机。

光电路交换机 (Optical Circuit Switches, OCS) 是一种基于纯光信号路径切换的核心技术设备，其工作原理完全规避了传统电交换必需的光电转换环节。OCS 基于光交叉交换原理（在 P 个输入光端口和 M 个输出光端口之间进行切换）的光信号控制交换技术，其核心功能是实现光信号在不同通道间的动态切换，使服务器端口之间直接实现光互连，完成全光路的重构，整个过程不需要光-电-光 (O-E-O) 转换。

现代数据中心往往使用脊叶 (spine-leaf) 架构，OCS 能够在脊柱层 (spine 层) 提供稳定、大带宽直连通道的数据流，为进一步拓展数据传输性能提供了极具吸引力的方案。spine-leaf 架构由两层组成：spine 层（核心骨干交换机）和 leaf 层（接入交换机），leaf 层直接连接数据中心内的服务器、存储设备和其他设备，该层的交换机负责将流量转发到 spine 层。Leaf 层流量的特点是突发性强、连接数量多、但每个连接的数据量小，而 spine 层接收汇聚的数据流，具有大规模、持续性传输的特点。在传统的拓扑结构中，数据中心网络的 spine 层通常采用电交换机，需频繁进行电信号与光信号之间的转换。这一过程不仅消耗大量电力，还会引入显著的数据延迟，此外，若要在数据中心中部署此类架构，必须预先建设大规模的主干层，否则后续扩展将面临整体重新布线的挑战，造成高昂的资本支出。相比之下，OCS 作为光电光的替代方案，能够实现更高效、更低延迟的数据传输，

有助于 spine 层从分组交换向光转换的过渡。

图46：使用 OCS 替代 spine 层的网络架构



资料来源：Polatis，国联民生证券研究所

OCS 凭借其高带宽容量，低能耗，数据速率独立性，低延迟和可扩展性，可以解决电交换机面临的许多限制。

**1) 带宽容量高：** OCS 由于传输速率不受限制，能够充分利用光纤的容量，从而更有效地利用网络资源，满足现代数据中心日益增长的带宽需求。

**2) 能耗低：** 由于消除了光电转换的需求，不需要放大器或中继器等功率密集型组件来进行长距离传输，OCS 的能耗降低，成为更能满足可持续发展目标的技术。

**3) 数据传输速率独立：** OCS 能够以多种速率连接，能够支持更快地扩展规模，其理论交换速度可达传统电交换芯片的 1000 倍以上，美国能源部阿贡国家实验室和普渡大学已在实验中验证该性能指标。

**4) 信号传输快：** OCS 能够通过避免光电转换从而实现接近 0 的延迟，这对于需要实时数据处理和低延迟通信的应用尤其有利。

**5) 可扩展性：** 在架构上，OCS 架构天生具备更强的扩展能力，能够支持更多端口和更高的聚合吞吐，因此 OCS 能够满足现代数据中心动态且不断增长的需求。

### 3.2.2 OCS 光交换方案一览

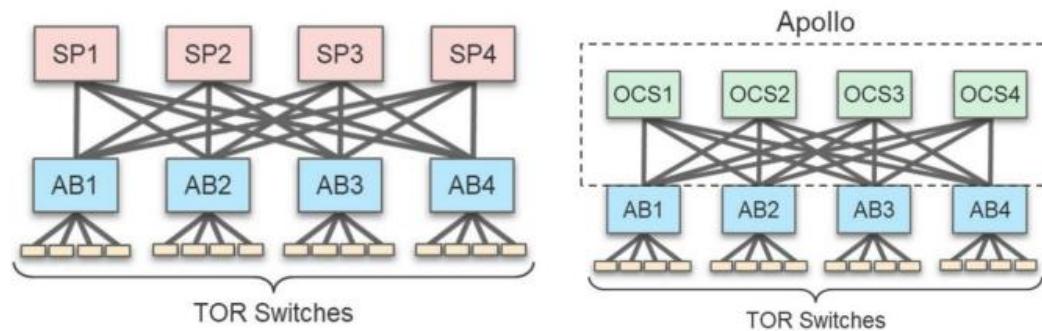
OCS 领域尚未形成统一技术标准，目前有三大技术路线并行发展——3D MEMS，数字液晶 (DLC) 和直接光束偏转 (DBS) 方案，各大厂商正在加速部署。

#### 1) MEMS 方案

MEMS 方案通过在硅晶圆上蚀刻微型反射镜阵列，并利用集成的静电或磁致动器驱动微镜的倾斜，以精确地改变输入光束的传播方向，将其路由至指定的输出端口。Lumentum 在其 R300 产品中使用了 MEMS 方案，其 MEMS 技术已累积

超过 1 万亿小时的现场微镜运行时间，对于提升 OCS 的可靠性和性能有显著作用。

**图47：传统数据中心架构和 Apollo OCS 架构的对比**

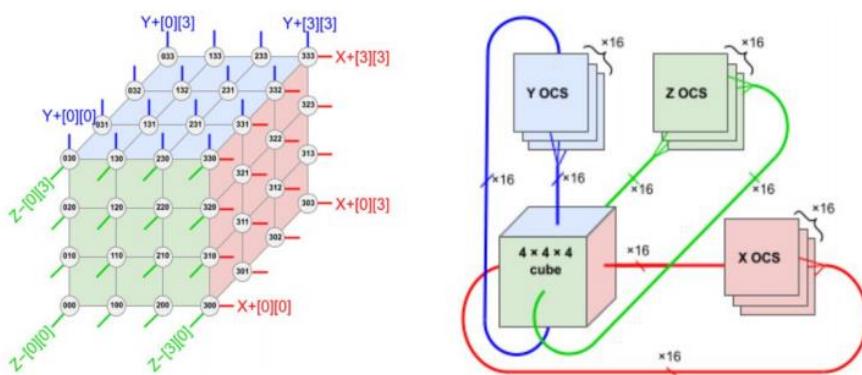


资料来源：“Apollo: Large-Scale Deployment of Optical Circuit Switching for Datacenter Networking,” in Optical Fiber Communication Conference (OFC) 2023, 国联民生证券研究所

**谷歌作为 OCS 领域投入最大、应用最深入的厂商，在 MEMS 方案上已经进行了相当成熟的应用。**谷歌在 OFC2023 中展示的内部项目的“Apollo”OCS 平台基于 MEMS 方案，对于该系统，OCS 降低了 30%的成本和 40%的功耗，能够合理地控制成本。谷歌在脊置换与 AI 集群重构等实际场景中，充分展示了该技术在节能与架构灵活性方面的显著优势，有效推动了市场对该技术前景的关注。

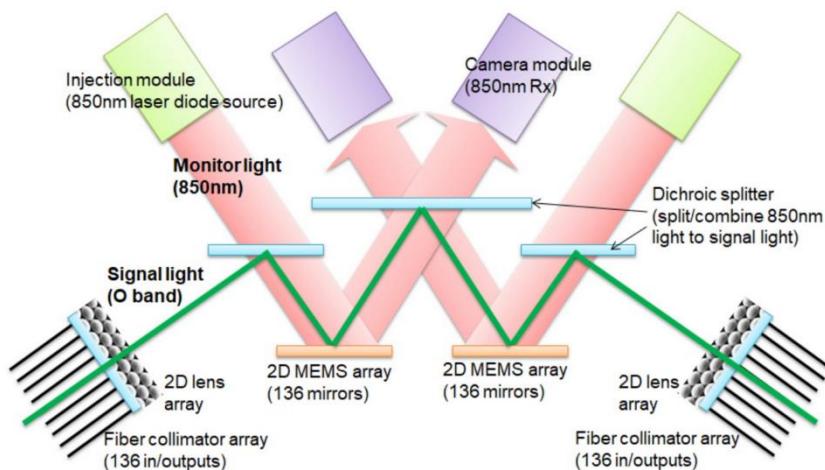
除此以外，谷歌的 TPU v4 Pod 是第一台部署可重构 OCS 的超级计算机，由 4096 个定制设计的芯片组成，与 OCS 链接，专为大规模机器学习工作设计。为了将这些芯片立体结构互联起来，形成更大规模的 Pod，谷歌则依赖于基于 MEMS 的 OCS 和光纤连接，即每个  $4 \times 4 \times 4$  的 TPU 立方体拥有 96 个光链路接口，最外侧六个面上的 TPU 与 48 个独立的 OCS 设备连接，内部的 TPU 之间通过电缆连接，从而构建了一个可重构的高速 3D 环面网络。

**图48：TPU 与 OCS 连接示意图**



资料来源：TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings., 国联民生证券研究所

图49：使用两个 MEM 阵列的 OCS



资料来源：semianalysis,国联民生证券研究所

据 Cignal AI 估计，从 2020 年至 2024 年，谷歌对 OCS 技术的投资累计已超过 5 亿美元，逐步将其深度融合至自身基础设施的多个应用层面。2024 年谷歌已开始量产最新的第六代 TPU v6，使用自主研发的 OCS 取代了脊柱交换机，通过其擅长的软硬件集成，提高了计算集群效率，大大降低了 TPU SuperPOD 的功耗和成本。

## 2) 数字液晶 (DLC) 方案

数字液晶 (digital liquid-crystal, DLC) 是一种非机械的光学交换方案，其工作原理是利用外部电场改变液晶材料的折射率，从而实现对光路方向的精确控制。Coherent 的 OCS 平台便基于此技术，该公司于 18 年前就已经将该项技术用于波长选择开关 (WSS) 中，具有丰富的技术使用经验，同时 DLC 技术只需要极低的驱动电压 (低于 10V) 来切换其液晶单元，能够进一步保障 OCS 运行的可靠性。

图50：Coherent 光开关 (使用 DLC 技术)



资料来源：Coherent 官网，国联民生证券研究所

### 3) 直接光束偏转 (DBS) 方案

直接光束偏转 (DirectLight Beam-Steering) 核心由三个部件构成：光纤准直器 (fiber collimator)、二维压电致动器 (2D Piezo Actuator) 和精确位置传感器 (position sensor)。每个准直器端口的转动位置均经过预先精确标定。系统通过压电效应驱动致动器产生二维伸缩位移，从而实现准直器的精密转动，并借助位置传感器实现闭环反馈控制，最终将两个光纤准直器精准对准至同一直线上。该技术是 Polatis 的独家专利技术，能够帮助 OCS 中的光精确转动与定位，不对光信号做任何其他处理，光性能良好。

**图51：Polatis 单模 576 x 576 矩阵光开关**



资料来源：Luster 官网，国联民生证券研究所

### 3.2.3 OCS 产业协同，供应链不断成熟

据 Cignal AI 估计，至 2028 年，全球 OCS 市场规模有望突破 10 亿美元，展现出强劲的增长潜力与不容忽视的商业价值。在 OCS 设备的供应链体系中，上游光交换机供应商、关键材料与核心元器件供应商以及下游客户共同构成了一个紧密协作且技术驱动型的产业生态。随着全球数据流量持续激增以及数据中心向高速、低功耗方向迭代升级，OCS 技术因其在光网络重构、能耗控制与传输效率方面的显著优势，正迎来广阔的市场前景。

**Coherent 供应 MEMS 和数字液晶两种方案的 OCS 平台。** Coherent 的 MEMS 光交换机的主要优势是可靠性高和寿命长，该光交换机已实现供货。此外，该公司采用数字液晶技术的 OCS 平台已获得 2024 ECOC 展会最佳产品奖，并获得了首个客户订单，预计在 2025 年开始产生收入，这表明 DLC 技术正在赢得市场的认可，成为 MEMS 技术路线的有力竞争者。

**Lumentum 作为光网络和光子解决方案的全球领导者，该公司目前正向多家超大规模客户送样其 R300 光交换机（300x300 端口）。** 其在 FY2Q25 首次通过 OCS 获得了收入，并向两家超大规模云客户发货，同时已获得第三家超大规模云客户的承诺，预计将于 2026 年开始部署。Lumentum 在电信领域深耕多年，积累了深厚的 MEMS 技术和专利组合，为其 OCS 产品的可靠性和性能提供了基础。

**中际旭创通过提供高性能的光模块，以支持 OCS 架构的高速数据传输需求。**为了与 OCS 的平滑互通，光模块需要更高的输出功率、双向性和多波长选项，中际旭创主要为 AI 和数据中心提供 100G、200G、400G、800G 乃至 1.6T 的光模块解决方案，使云运营商能够快速升级计算和网络，满足终端客户的需求。除了生产光模块外，该公司还具备 OCS 交换机整机代工能力。

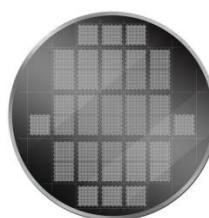
**德科立作为业内领先的光电子器件和光传输子系统供应商，是 OCS 整机供应商及代工商。**德科立建立了光收发模块、光放大器、光传输子系统三大技术平台，产品包括光收发模块、光放大器、光传输子系统等，其 OCS 整机产品主要服务于云计算、大数据中心等领域的核心客户。

**腾景科技已量产的钒酸钇 (YVO4) 单晶适用于隔离器、环形器和偏振器件，广泛应用于光通信领域。**环形器是一种能够实现双向通信的元件，可以使单个光纤股同时进行双向传输。这使得所需的 OCS 端口和光纤数量减半，从而降低了成本和复杂性，对于大规模部署至关重要。

**赛微电子作为核心的 MEMS 晶圆代工厂，其业务涵盖 MEMS 工艺开发和晶圆制造。**赛微电子掌握硅通孔、晶圆键合、深反应离子刻蚀等多种工艺和技术。该公司制造的 MEMS OCS 已通过客户验证，收到了采购订单，现已启动小批量 MEMS OCS 试生产，是 MEMS 技术路线的重要上游支持者。

**炬光科技作为全球领先的光子应用解决方案提供商，其研发的精密光学微透镜阵列为 OCS 小型化奠定技术基础，展现出突破性的创新价值。**该技术同时能够在相同尺寸下大幅提升光通道数量与数据传输带宽，为 OCS 系统性能拓展出更大空间。此外，它还确保了交换机内部光信号的高精度传导，进一步巩固了该公司在高端光通信领域的核心优势。

**图52：赛微电子制造的 8 英寸 MEMS-OCS 晶圆示意图**



资料来源：赛微电子官网，国联民生证券研究所

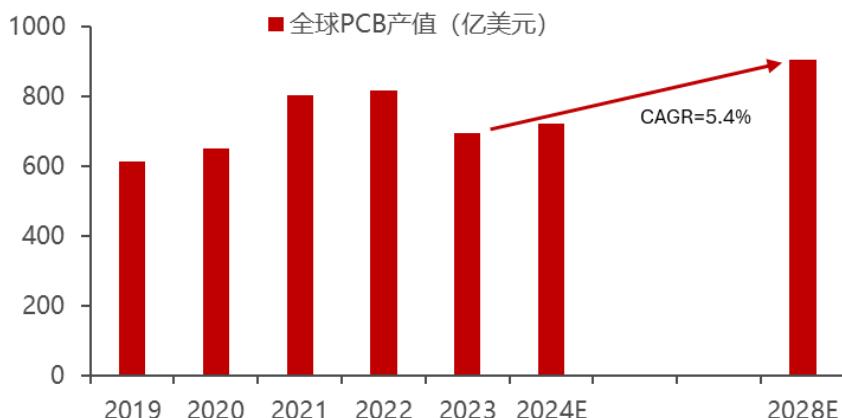
## 4 PCB：材料+设备升级是核心焦点

### 4.1 NV 推出全新 PCB 解决方案，技术升级清晰

#### 4.1.1 AI 拉动 PCB 高速增长，高多层及 HDI 为升级方向

PCB (印刷电路板, Printed Circuit Board) 承担了各电子元器件之间信号互联的功能, 据 Prismark 数据, 2021 年全球 PCB 市场规模达到 804 亿美元, 2023 年受宏观经济影响, 市场规模略有衰退, 市场规模下降至 695 亿美元。Prismark 预测至 2028 年全球 PCB 市场规模有望达到 904 亿美元, 2023-2028 年复合增长率约为 5.4%。2021 年, 中国 PCB 市场规模达 442 亿美元, 创历年来新高。预计 2028 年中国 PCB 产值将达约 461.8 亿美元, 2023-2028 年复合增长率为 4.1%。

图53: 2019-2028 年全球 PCB 市场规模 (亿美元)



资料来源: Prismark, 国联民生证券研究所

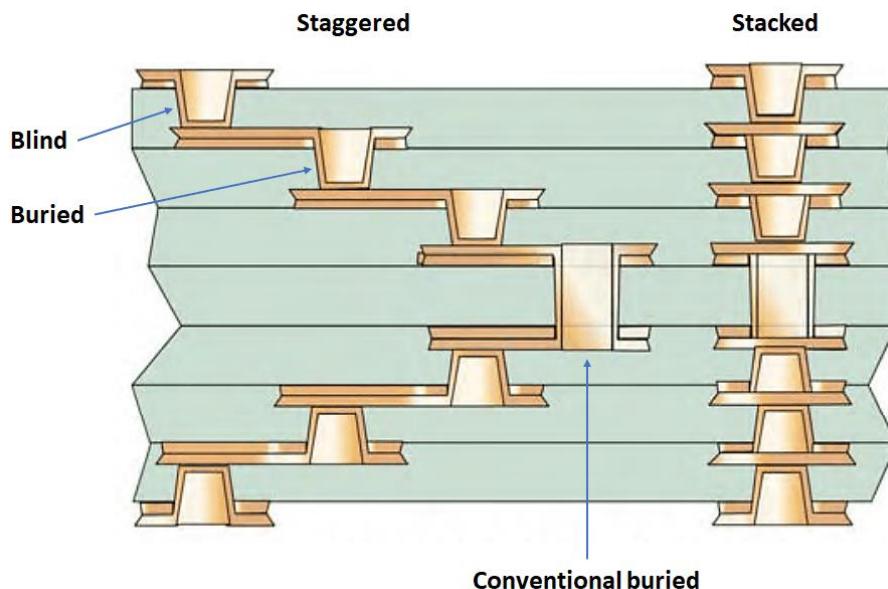
AI 服务器中, 由于 AI 算力芯片计算速度快、数据吞吐量大, 对信号传输有着更高的要求, PCB 的用量、材质及等级亦有提升, 价值量相比普通服务器明显提升, 高多层板和 HDI 板为 AI 领域 PCB 未来发展方向。

**高多层板**采用“信号层 - 电源层 - 接地层”交替排布, 在 AI 相关 PCB 产品中, 通常指 20 甚至 30 层以上通孔板; 相比传统 PCB 对钻孔精度准度及压合工艺等要求更高, **更高的层数也意味着加工难度更大, 价值量更高**。

**高密度互连技术 (HDI)** 采用更小的线宽线距、更精细的盲孔、埋孔和通孔设计, 实现了在有限空间内更高的布线密度和更复杂的电路连接, 与传统 PCB 相比线宽线距更细, 布线密度更高, **更高的阶数意味着加工难度更大, 价值量更高**。

在层间连接方面, 高密度互连技术展现出了卓越的优势。传统的多层线路板层间连接方式往往存在连接点较多、信号传输路径较长等问题, 容易导致信号衰减和传输延迟。而高密度互连技术通过优化孔结构和布线布局, 大大减少了层间连接点的数量, 缩短了信号传输路径。这不仅提高了信号传输的速度和质量, 还降低了信号在传输过程中的干扰和损耗, 确保了电子设备在高速运行时的稳定性和可靠性。

图54：高密度连接板（HDI）示意图

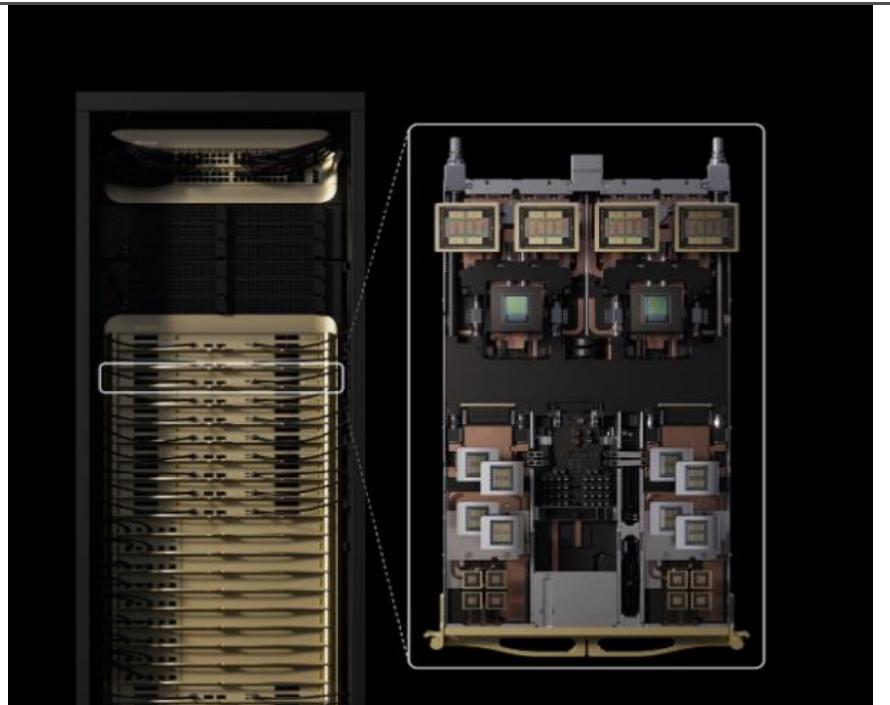


资料来源：the engineering projects, 国联民生证券研究所

#### 4.1.2 CPX 及 Midplane 成为英伟达 Rubin 系列 PCB 全新增量

NVIDIA 于 2025 年 9 月 9 日宣布推出 NVIDIA Rubin CPX，一款针对推理场景而打造的新一代 GPU。Rubin CPX 使得 AI 系统能以突破性的速度与效率，处理百万个词元的软体编码与影片生成。Rubin CPX 与全新 NVIDIA Vera Rubin NVL144 CPX 平台中的 NVIDIA Vera CPU 和 Rubin GPU 协同工作。这款整合式 NVIDIA MGX 系统在单一机架配置下拥有 8 exaflops 的 AI 运算能力，可提供比 NVIDIA GB300 NVL72 系统高出 7.5 倍的 AI 效能，同时配备 100 TB 快速记忆体及每秒 1.7 PB 的记忆体频宽。NVIDIA 也将提供一个专属的 Rubin CPX 运算托盘，以满足客户希望重复利用现有 Vera Rubin 144 系统的需求。

图55：Nvidia Rubin CPX

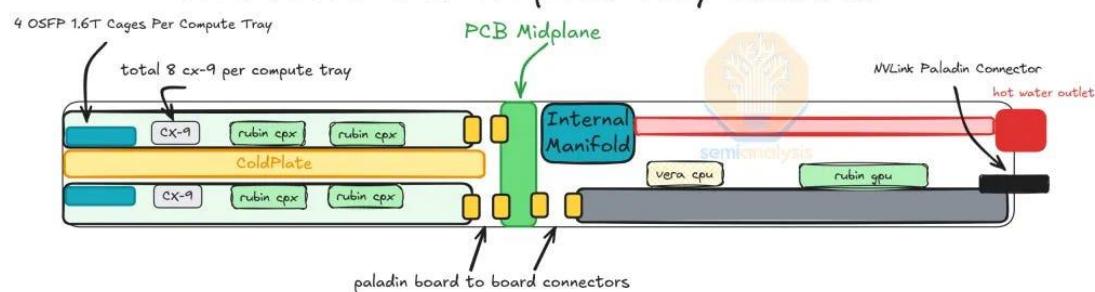


资料来源：Nvidia.Developer, 国联民生证券研究所

在 NVIDIA Vera Rubin NVL144 CPX 平台中，compute tray 上半部分采用类似 GB 系列的 Bianca 板，搭载两颗 Vera CPU 及四颗 Rubin GPU；CPX 芯片以及网卡则布局于 Compute tray 下半部分；**其中 8 颗 CPX 芯片采用夹层设计，左右两侧各放置 4 颗，并搭载于 PCB 之上。**

图56：VR NVL144 CPX Computer Tray 侧视图

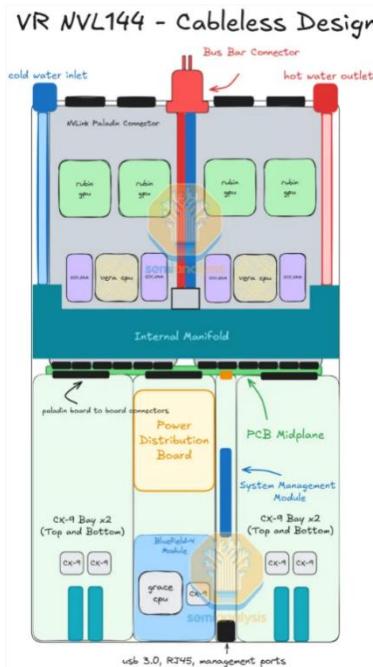
### VR NVL144 CPX Compute Tray Side View



资料来源：SemiAnalysis, Nvidia, 国联民生证券研究所

**Rubin 系列 Compute tray 中另一个重要改动为采取无线缆设计。过去上方 Bianca 板与下方网卡使用线缆实现信号互联，而在 Rubin 系列中将采用 Midplane (中板) 替代线缆实现互联功能，信号通过 Amphenol Paladin 连接器在 Midplane 中完成路由分配。相比线缆方案，PCB 中板具备安装效率高、节省空间、不易损坏等优点，可有效提高机柜良率及生产效率，预计将成为 Rubin 系列标配。**

图57: VR NV144 Computer Tray 中使用 Midplane

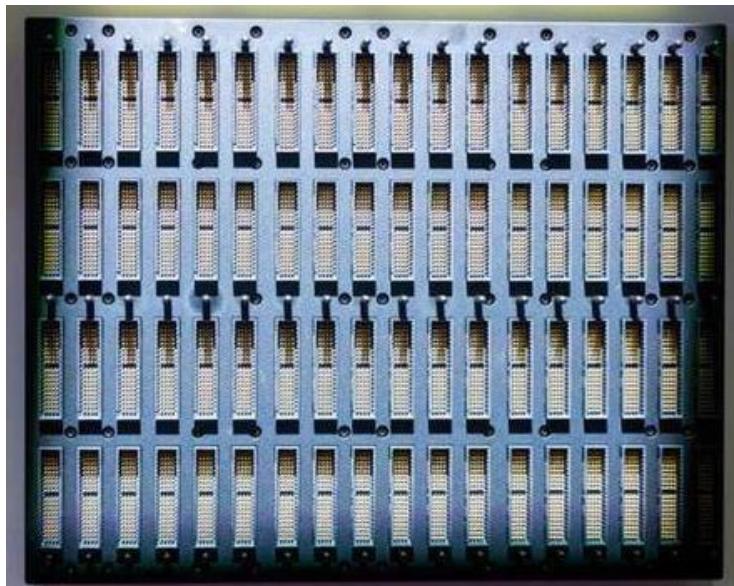


资料来源: SemiAnalysis, Nvidia, 国联民生证券研究所

## 4.2 英伟达 NVL576 采用 PCB 正交背板实现高速互联

**Rubin Ultra NVL576 实现性能显著提升。**英伟达于 2025 年 GTC 大会上宣布, 2027 年下半年将推出 Rubin Ultra NVL576, 采用 PCB 板, 实现了极大的扩展, 浮点运算次数增加了 14 倍, 达到 15 ExaFLOPS, 内存带宽规模提升至 4.6 PB/s, 该系统总共拥有 365 TB 的高速内存, 其中包括 147 TB 的 HBM 和 218 TB 的 LPDDR。英伟达将直接在单个封装中采用 16 堆 HBM, 从 8 堆增加到 16 堆。封装中将有一排 4 个晶圆大小的 GPU, 两侧各有两个 I/O 芯片。计算面积翻倍, 计算能力也翻倍至 100 PFLOPs 的密集 FP4 算力。

图58: Rubin Ultra NVL576 / 背板结构图

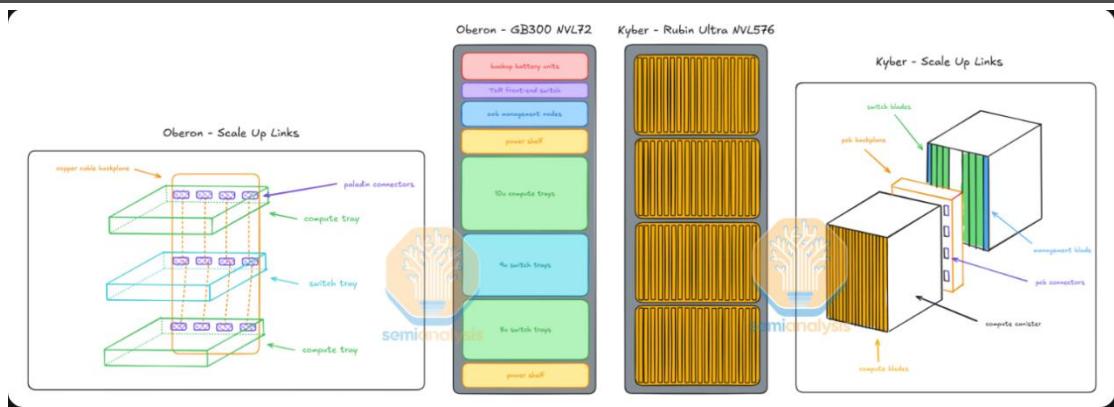


资料来源: Nvidia, 半导体行业观察, 国联民生证券研究所

**出于对系统整体的考量, Rubin Ultra NVL576 采用 PCB 正交背板替代传统铜缆。** NVL576 采用全新的 Kyber 机架架构取代过去的 Oberon 架构, 而传统铜缆互联在信号完整性、布线密度和系统可靠性方面逐渐成为瓶颈, 英伟达选择用 **PCB 正交背板取代铜缆背板作为机箱内 GPU 和 NV Switch 之间的 Scale-Up 连接。**

在结构布局上, NVL576 的计算托盘被旋转 90 度, 以适配刀片式 (blade) 外形, 从而显著提升了整机的机架部署密度。**机架后部的 NV Switch 通过正交背板的反面与前方的计算刀片互联**, 形成一个紧凑而高效的高速互连拓扑。每个机架包含四个罐体 (canister), 每个罐包括两层, 每层 18 个计算盒。每个计算盒配备两个 Rubin Ultra GPU 和两个 Vera CPU, 因此单个罐体内集成了 36 个 R300 GPU (合计 144 个 die) 和 36 个 Vera CPU。四个罐体合计可提供 144 个 GPU (576 个 die), 构成机架级别的 NV Link 大规模互联。这一创新架构为 AI 超算中心提供了更强的可扩展性和更高的带宽密度。

图59：Oberon 机架架构与 Kyber 机架架构对比



资料来源：SemiAnalysis，国联民生证券研究所

#### 传统铜缆互联方案存在多方面制约因素：

- 1) **信号完整性瓶颈** — 随着数据速率提升至 112 G PAM4 及以上，铜缆的趋肤效应和介质损耗迅速加剧，导致信号在传输过程中衰减严重，波形畸变、误码率显著上升。
- 2) **布线密度与空间约束** — 铜缆本身截面积较大，且必须满足特定的弯曲半径要求。在满载 GPU 和 NV Switch 的机柜环境中，若采用铜缆互联，数量庞大的高速线缆不仅占据大量机架空间，还严重影响空气流通与散热效率，增加维护难度。
- 3) **系统可靠性与功耗问题** — 大量连接器与线缆接口是系统故障的高发点。接触不良、振动松脱等问题可能导致系统停机。此外，铜缆长距离传输本身功耗高、信号驱动能耗增大，系统整体能效比下降。
- 4) **组装良率与效率瓶颈** — 从制造与装配角度看，铜缆布线高度依赖人工，工作流程繁琐、耗时，并且线束体积大、布线繁杂，严重制约出货速度和规模化扩产能力。

#### 针对上述痛点，PCB 正交背板方案提供了全方位改进：

**信号完整性与带宽密度显著提升：**通过精密光刻形成的 PCB 传输线可实现严格的阻抗控制与通道特性一致性。结合 M9 级基材、HVLPI4 铜箔、石英纤维布 (Q 布) 等高级材料，其插入损耗与串扰远低于等长度铜缆。同时，PCB 背板可在有限面积内实现极高的布线路密度，多层次设计可支持数万个高速差分对，这是传统铜缆难以企及的。

**系统可靠性与功耗改善：**背板与子卡（计算刀片、交换刀片）通过压接 / 焊接形成刚性结构，抗振动、抗冲击性能远优于传统线缆。连接点数量减少，系统故障率显著下降。同时，优越的信号完整性意味着可用更低的发射功率实现可靠通信，从而减少互联部分的整体功耗。

**产业化与交付效率提升：**PCB 制造采用印刷、蚀刻、层压、测试等高度自动化

流程，高良率、一致性强。机柜组装环节亦简化为“插卡式”操作，极大缩短工时与人工成本。此外，正交背板使系统架构模块化：计算单元、交换单元、电源与冷却可并行制造和测试，最终快速集成，从而显著提升爬坡产能与出货速度。

M9 等级基材、HVLP4 低轮廓铜箔与石英纤维布（“Q 布”）构建的 PCB 正交背板方案通过高性能材料与精密制造技术的结合，突破了铜缆在长距离、高速、大通道场景下的局限，实现了高信号完整性、高带宽密度、结构刚性、模块化制造与自动化组装。

### 4.3 PCB 材料配套 M8/M9 等级升级

覆铜板由基板、铜箔和粘合剂构成，是将增强材料浸以树脂，一面或两面覆以铜箔，经热压而成的一种板状材料，覆铜板是印制电路板（PCB）制造中最核心的材料。

PCB 主要功能是使各种电子零组件形成预定电路的连接，起中继传输的作用，是电子产品的关键电子互连件，有“电子产品之母”之称。单面或双面 PCB 的制造是在覆铜板上有选择地进行孔加工、铜电镀、蚀刻等，得到导电图形电路。在多层印制电路板的制造中，也是以内芯薄型覆铜板为底基，将其制成导电图形电路，并与粘结片交替叠合后一次性层压成型加工，使它们粘合在一起并成为三层以上的图形电路层之间的互联。作为 PCB 制造中的基板材料，覆铜板对于 PCB 整体特性起到十分重要的作用，主要有着导电、绝缘和支撑三方面的功能；PCB 的性能、品质、制造中的加工性、制造水平、制造成本以及长期可靠性等很大程度上取决于基板材料。

图60：覆铜板



资料来源：建滔积层板官网，国联民生证券研究所

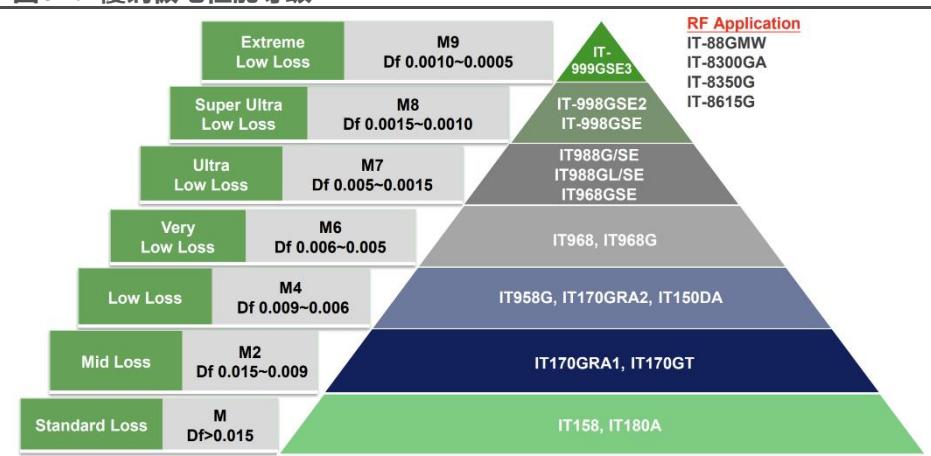
在传统的电子产品应用中，应用频率大多数集中在 1GHz 以下，普通覆铜板的电性能足以满足其要求。但是高频高速环境下，高频信号本身的衰减很严重，另一方面其在介质中的传输会受到覆铜板本身特性的影响和限制，进而造成信号失

真甚至丧失。因此高频高速应用领域对于覆铜板电性能的要求非常高。

业内根据 Df 将覆铜板进行分级，传输速率越高对应需要的 Df 值越低，以 5G 通信为例，其理论传输速度 10-20Gbps，对应覆铜板的介质损耗性能至少需达到中低损耗等级，Df 越低，材料的技术难度越高。

根据应用场景的差异，高频高速覆铜板又可以细分为高速板和高频板两个应用方向，两者都需要更低的 Dk 和 Df，但是侧重点有所差异。其中高速板更侧重 Df，Df 是影响传输损耗和信号完整性的主要因素；高频板更侧重 Dk 的准确性和稳定性，Dk 影响传输时延和特性阻抗。高频高速板主要应用在服务器、存储器、交换机、路由器、基站等对信号传输要求高的场景。

**图61：覆铜板电性能等级**



资料来源：ITEQ，国联民生证券研究所

**聚苯醚 (PPO)** 是世界五大通用工程塑料之一，具有刚性大、耐热性高、难燃、强度较高电性能优良等优点。另外，聚苯醚还具有耐磨、无毒、耐污染等优点。PPO 的介电常数和介电损耗在工程塑料中是最小的品种之一，几乎不受温度、湿度的影响，可用于低、中、高频电场领域。与高分子量 PPO 相比，低分子量 PPO 的熔融粘度小、加工性能更好，因此对低分子量 PPO 的侧链或端基进行化学改性，使其形成可交联基团的热固性 PPO 低聚物。**改性聚苯醚 (MPPO)** 制成的高速覆铜板具有较低的介电损耗因子，同时在耐热性、耐水性、阻燃性及良好的尺寸稳定性方面有一定优势，**目前成为高速服务器覆铜板的主力军，在高速覆铜板广泛使用**。

**碳氢树脂 (PCH)** 是不含任何极性基团的碳链聚合物，仅由 C 和 H 元素组成，具有优异的介电性能。常见的碳氢树脂有苯乙烯-丁二烯-二乙烯基苯共聚物、苯乙烯-丁二烯-二乙烯基苯共聚物、丁二烯均聚物等。由于 C-C 键和 C-H 键的电子极化率小，碳氢树脂在较宽的频率和温度范围内表现出较低的介电常数和超低的介电损耗因数。同时**碳氢树脂具有优异的加工性能，相对于其他高频覆铜板树脂材料，其成型工艺简单、成本低，被认为是下一代高频覆铜板的首选树脂材料**。

**聚四氟乙烯 (PTFE)** 是目前为止发现的介电常数最低的高分子材料之一，具有优良的介电损耗和耐热性。PTFE 介电性能极佳，主要的缺点则是粘接性能和熔

融流动性不好，热膨胀系数较大，导致加工难度较大。当前主要是将其和一些填充材料或者有机聚合物共混实现改性，如无机填充粒子、聚苯酯及聚苯硫醚等。PTFE与碳氢树脂（PCH）凭借优异的低介电性能成为了高频高速基板未来两条核心技术路线。

当前AI服务器以M7等级材料为主，并向M8/M9等级材料升级；更高等级的高速材料价值量更高、加工难度更大，PCB制造壁垒更高、盈利能力较好；后续正交背板/midplane/cpx等全新PCB方案有望搭载更高等级材料，覆铜板及PCB制造商核心受益。

表 10：部分高速覆铜板基体树脂

材料	介电常数Dk (1MHz)	介质损耗因子Df (1GHz)
PTFE	2.1	0.0004
PPO	2.4	0.0007
氰酸酯树脂	2.7-3.0	0.003-0.005
环氧树脂	3.6	0.025
PCH	2.4	0.0002

资料来源：深圳惠科新材料股份有限公司微信公众号，国联民生证券研究所

### 1) 电子布向石英布方向升级

**电子布**，是一种以玻璃纤维为基材，经特殊织造工艺制成的高性能织物，是电子信息产业中关键的基础材料之一。电子布作为生产覆铜板必不可少的材料，为生产印制电路板的专用基本材料，随着5G、AI服务器及高速运算设备对信号传输速度和质量要求越来越高，对介电性能（Dk/Df）提出更高要求，以确保信号传输更快、损耗更小，低介电一代、二代布向石英纤维布升级。材料成分方面，传统玻纤含硼、钙、镁等碱性杂质，低介电一代和二代需降低此类杂质，而石英纤维布99.9%成分为二氧化硅，这是其核心特征。介电性能方面，主要关注介电常数（DK）和介电损耗（DF）。低介电一代在10MHz以下DK值为4.7，DF值为2.9‰；低介电二代DK值为4.0-4.4，DF值为1.7‰-2.3‰；**石英纤维布DK值降至3.74，DF值降至0.2‰**，相比一代和二代显著提升。

图62：玻璃纤维布



资料来源：天津市中天俊达玻璃纤维制品有限公司官网，国联民生证券研究所

在电子布方面，M7 级别 CCL 一般搭配一代布，M8 级别一、二代布混用，M9 级别中则有望加入 Q 布。当前宏和科技、中材科技、菲利华等公司已深耕 low-dk 领域，并向 Q 布进军，已在下游客户认证取得较好进展，由于日东纺、旭化成等外资电子布龙头扩产谨慎，low-dk 布已出现明显的供给紧张，国内相关公司有望在一、二代布实现切入和份额提升，并在 Q 布实现弯道超车。

图63：石英纤维布



资料来源：菲利华官网，国联民生证券研究所

## 2) 铜箔：HVP1-5 升级，材料量价齐升

铜箔是一种由纯铜或铜合金制成的薄片材料，通常厚度在 0.1 毫米以下。它具有良好的导电性、导热性和可塑性，广泛应用于电子、工业和装饰等领域。铜箔按照生产工艺可以分为电解铜箔和压延铜箔，电解铜箔通过电解法沉积成层，压延铜箔通过物理方法反复辊压加工形成。铜箔是制造覆铜板及印制电路板的重要原材料，在 PCB 下游应用领域主要为消费电子、计算机及相关设备、汽车电子、通信设备等行业。

图64：铜箔分类



资料来源：智研咨询，国联民生证券研究所

**标准铜箔根据性能可以分类为常规铜箔和高性能类铜箔两大类。**高性能 PCB 铜箔按照应用领域可以划分为五类，包括高频高速电路用铜箔、IC 封装载板用极薄铜箔、高密度互连电路（HDI）用铜箔、大功率大电流电路用厚铜箔、挠性电路板用铜箔。高频高速电路用铜箔根据粗糙度不同，可以细分为 HTE、RTF 和 HVLP，其中 HVLP 又称高频超低轮廓铜箔，在可满足 AI 服务器等设备对信号传输的高要求。

**HVLP 铜箔具有硬度高、表面平滑、厚度均匀、电流传输稳定高效、信号损耗低等优势，在 AI 服务器及智能汽车、通信设备、消费电子、航空航天等对信号传输要求较高的领域具有广阔应用前景。** HVLP 铜箔根据粗糙度不同可分为 1-5 代，当前主要以 HVLP1 及 HVLP2 居多，部分对电性能要求高的 AI 产品升级为 HVLP3 及 HVLP4 铜箔，而 HVLP5 技术门槛最高，定位下一代产品，尚未批量应用。HVLP 从第一代到第五代，典型粗糙度 (Rz) 逐代降低，信号传输损耗逐代降低，适配更先进的 PCB 技术，同时技术难点和制造成本也在逐代增加。第三代 HVLP 驱动来自 AI 服务器和更高速的网络设备。开始匹配 M7/M8 等级的覆铜板；第四代主要匹配 M8/M9 等级覆铜板，是支撑 800G 光模块、高端 AI 加速卡的关键材料；第五代旨在匹配下一代 M9+ 等级覆铜板，面向未来更高速的计算和通信平台。

**表 11：HVLP 铜箔 1-5 代情况**

产品类别	产品应用	产品特性
HVLP1	用于高速传输的数字设备的基板材料	对基材具有优异的附着力，具有低损耗特性和极低的粗糙度
HVLP2	用于超高速传输的数字设备的基板材料	在 1.0um 以下的极低粗糙度下具有出色的附着力
HVLP3	用于高性能 AI 加速器的基板材料	超低粗糙度 (<0.6um) 和优异的粘合强度适用于具有优异信号传输特性的高性能 AI 加速板
HVLP4	用于高性能 AI 加速器的基板材料	超细结节治疗。信号传输改进
HVLP5	用于高性能 AI 加速器的基板材料	无结节技术。最佳信号传输特性

资料来源：Solus Advanced Materials 官网，国联民生证券研究所

**可剥离超薄铜箔（Peelable Ultra-Thin Copper Foil），简称可剥离铜，是指厚度在 9μm 以下的铜箔，由载体支撑，在使用过程中可剥离。** 可剥离铜具有抗拉强度高、热稳定性好、剥离力稳定可控、表面轮廓低等特点，可用于生产芯片封装基板。可剥离超薄铜箔主要由载体层、剥离层、超薄铜箔层组成，载体层通常使用 18 或 35μm 电解铜箔，其中导电剥离层和超薄铜层的为研发重点。

图65：可剥离超薄铜箔产品结构



资料来源：龙电华鑫控股官网，国联民生证券研究所

**可剥离型超薄载体铜箔，适用于 PCB 制程中 mSAP 半加成法及 Coreless 制程，可大幅降低 PCB 及 IC 载板的厚度和重量，满足终端电子产品轻薄化的需求。** mSAP 工艺特点是在基板表面先铺设一层超薄种子铜，再按电路图形电镀加厚所需铜，再去除种子铜，从而得到精细铜线。由于初始铜极薄，避免了传统蚀刻中的侧蚀问题，导线截面更接近直壁，阻抗一致性好。mSAP 工艺能实现高精度、高密度线路(如 0.018 mm 线宽)且保持量产可行性，已被用于制造类载板 PCB 满足最新智能手机和移动设备对极细线路的需求。

#### 4.4 M9 材料升级对 PCB 设备及耗材提出更高要求

PCB 钻孔机是一种用于在 PCB 上进行精确打孔操作的设备，以便于后续安装电子元件如电阻、电容、IC 引脚以及其他各种连接器。在先进 PCB 制造中，钻孔是实现多层、高密度互连的核心工艺环节，也是设备产业链价质量占比较高的环节。通过机械或激光钻孔形成通孔与微孔，实现电路板内部不同层间以及外部元器件与电路板之间的电气互连。钻孔后进行去毛刺处理，清理孔壁残留的基材碎屑。钻孔的质量影响孔金属化和导通的效果，以及布线的质量和密度，直接影响 PCB 的性能、良率和制造成本，也是高端封装（如 CoWoP）实现量产的基础保障。

而根据不同行业和商家产品设计的需要，3C 产品、医疗设备以及 AI 服务器等公司有在多层板内层导通信号强烈的需求，因此钻孔工艺也有三种常见的技术形式：通孔，导通所有板层的孔；盲孔，导通表层和中间某几层的孔；埋孔，导通内层板的孔。盲埋孔工艺常应用于 HDI 高密度互连板。

钻孔方式和工艺的选择取决于孔径、板材特性及产品设计需求。

表 12：钻孔技术分类

分类	分类依据	主要应用领域
机械钻孔	孔径 $\geq 0.15\text{mm}$ , 需搭配钻针	标准通孔和较厚板材加工, 例如工业或汽车电子中厚板的信号 / 功率通孔
激光钻孔	孔径 $< 0.15\text{mm}$	HDI 设计、微盲孔、灵活电路板等对孔径和密度要求极高的应用场景, 尤其是医疗设备、5G、移动设备、AI 服务器

资料来源：大族数控公司 2024 年年报，国联民生证券研究所

**机械钻孔是 PCB 制造中应用最为广泛的钻孔方式, 主要适用于孔径  $\geq 0.15\text{mm}$  的通孔加工, 也可覆盖部分  $0.05\text{--}0.15\text{ mm}$  的小孔领域。**该方式通过高精度钻头在多层板材上直接钻削, 孔壁光滑、导电性佳, 且加工稳定性高。机械钻孔适用的板材范围广, 可在刚性板、复合板等多种材料上使用, 并能一次叠层加工多块 PCB, 大幅提升生产效率。近年来, 国内企业已突破  $0.01\text{ mm}$  规格钻头生产, 使机械钻孔逐步向极小径延伸, 在稳定性与成本控制上优势明显。

**激光钻孔利用高能量激光束进行非接触加工, 特别适合孔径  $< 0.15\text{ mm}$  的微孔、盲孔及高密度互连 (HDI) 结构。**激光钻孔热影响区小、定位精度高, 可实现  $0.05\text{ mm}$  甚至更小的极细孔径, 但多用于单板材料, 在复合板材上易产生孔形不一致的问题。其加工灵活, 可精准控制钻孔深度, 尤其在加工盲孔和阶梯孔时优势显著。相比机械钻孔, 激光钻孔更适用于高密度布线与尺寸精度要求极高的场景。**在 AI 芯片载板、AI 加速卡核心 HDI 板等应用中, 激光钻孔能够提升互连密度与信号完整性, 满足高速、低延迟的电气性能要求, 为 AI 硬件的小型化和高性能化提供重要工艺保障。**

AI 相关覆铜板材料的升级, 和孔洞厚径比的上升, 导致机械钻孔机效率明显降低, **加工同样面积的 AI PCB 产品所需设备数量显著增加, 对于钻孔设备的加工效率提出更高的技术要求;**而信号完整的提高, 背钻孔数提升的同时加工精度要求更高, 对更高技术附加值的 CCD 六轴独立机械钻孔机的需求量更多。

根据 QYResearch, 2023 年全球 PCB 钻孔机市场销售额达到了 92.5 亿元, 2031 年全球 PCB 钻孔机市场销售额预计将达到 157 亿元, 年复合增长率(CAGR) 为 5.9% (2025-2031)。全球范围内全球激光钻孔机头部企业主要是 Mitsubishi Electric、Via Mechanics、ESI (MKS Instruments)、Han's Laser、EO Technics, 其中 TOP5 企业份额超过了 40%。而就产品类型细分其中典型的型号主要有 Mitsubishi Electric 的 GTW5 系列、Via Mechanics 的 ND 系列、ESI 的 5335 系列。

**大族数控的机械钻机及背钻设备已在 AI PCB 生产中得到广泛应用, 应用于高等级覆铜板加工的新型激光钻孔设备有望实现弯道超车。大族数控新开发的具有 3D 背钻功能的钻测一体化 CCD 六轴独立机械钻孔机, 可实现超短残桩及超高位置精度的背钻孔加工, 已获得行业终端客户的认证及多家高多层板龙头企业的大批量采购; 而针对高多层 HDI 板的加工需求, 需要更多的激光钻孔机来满足多**

阶堆叠盲孔或深盲孔加工，公司研发的高功率及能量实时监测的 CO<sub>2</sub> 激光钻孔机可实现大孔径及跨层盲孔的高品质加工。

**表 13：激光钻孔设备介绍**

产品	产品图片	产品用途	加工效果
CO <sub>2</sub> 激光钻孔设备		采用高功率 CO <sub>2</sub> 激光光源作为加工工具，利用激光烧蚀原理实现微小通、盲孔的加工；主要用于智能手机、平板电脑、光模块、通讯设备、汽车电子等 HDI 及 BT 类 IC 封装基板的加工。	 盲孔 50μm 显微镜效果  盲孔 50μm 切片显微镜效果
UV 激光钻孔设备		采用 UV 冷光源和特有的飞行钻孔模式，实现对挠性线路板及刚挠结合板 PI 材料的微小通孔/盲孔加工；主要用于智能手机、可穿戴设备、PC 及平板电脑、汽车 BMS 电池管理线束等领域。	 盲孔 25μm 显微镜效果  盲孔 25μm 切片显微镜效果
新型激光钻孔设备		采用新型激光钻孔技术，主要针对 mSAP 工艺类载板及 IC 封装基板，热影响效应小，实现对 ABF、BT 及 RCC 等材料微小盲孔/通孔的超快加工，满足新一代 SoC、SiP、CPU 及 GPU 等产品高阶封装领域的需求。	 盲孔 30μm 显微镜效果  盲孔 30μm 切片显微镜效果

资料来源：大族数控公司 2024 年年报，国联民生证券研究所

**在先进 PCB 制造中，钻孔是实现多层、高密度互连的核心工艺环节，也是设备产业链价质量占比较高的环节。**钻孔可通过机械钻孔搭配钻针或激光钻孔实现。在机械钻孔的工序中，钻针是耗材，也是钻孔工具的核心，直接决定了 PCB 的加工质量和生产效率。

当前，PCB 行业正经历着材料升级和层数增加的双重变革。随着人工智能的加速演进与应用深化，新一代信息技术产业对于高算力和高速网络通信的需求呈高增长态势，驱动了下游市场对于大尺寸、高层数、高频高速、高阶 HDI、高散热等高附加值 PCB 产品需求的快速增长，对 PCB 钻针提出了更高的质量要求和数量要求。质量要求方面，英伟达 Rubin 平台有望采用更多的 M9 材料，其核心是 AI 服务器及高速运算设备对信号传输速度和质量要求越来越高，需要材料更优异的介电性能，以确保信号传输更快、损耗更小，**于是 low-dk 玻纤布将向石英布迭代，基材材料硬度较前代大幅提升，导致传统涂层钻针加工磨损增加，寿命大幅缩短，钻针用量增多，市场空间显著增长。**且频繁换针导致生产线停机时间增加，综合加工成本大幅上升。高层数和超厚 PCB 板的需求增加，不仅减少了钻针寿命，对钻针的长径比和加工精度提出了更高要求。

**表 14：鼎泰高科、四方达、沃尔德、中钨高新产品与核心技术介绍**

分类	鼎泰高科	四方达	沃尔德	中钨高新
产品品类	刀具产品(钻针:微型钻针、涂层钻针、长径比钻针、铣刀)、研磨抛光材料	资源开采/工程施工类产品、精密加工类产品、CVD 金刚石业务	超硬刀具、硬质合金刀具及棒材、金刚石功能材料	硬质合金及钨制品(涵盖切削刀具、矿山工具、耐磨零件、数控刀片、PCB 工具等)、钨钼原料及制品、难熔金属、粉末制品等
直径规格范围	钻针产品直径规格覆盖 0.035mm 到 6.75mm 铣刀产品 直径 规 格 覆 盖 0.35mm-3.175mm	0.5mm - 20mm (侧重中大型直径), PCD 厚度 1mm - 10mm	直径范围为 0.1-2mm 的 PCD 微钻系列产品开发;开发 PCD\CVD\单晶金刚石材料的微铣\微钻, 直径 0.5-30mm 的 PCD 铣刀系列, 直径 0.2-12mm PCD 螺旋钻头系列	-
核心技术优势	硬质合金与自研 CVD 涂层、PVD 硬质涂层及 Ta-C 润滑涂层等各类涂层技术	PCD 微钻钻头, 寿命长、加工精度高、光洁度好; 全球能供应超大直径拉丝模的两家生产商之一	有“超硬材料激光微纳米精密加工技术”、“超薄金刚石片、复合片精密研磨及镜面抛光技术”、“系统性的高精密超硬刀具生产技术”、“金刚石功能材料生长技术”	从钨矿资源开发、冶炼加工到硬质合金及钨制品生产的全产业链技术体系; 在硬质合金材料研发、精密成型、涂层技术等方面具备优势; “硬质合金棒材挤压成型及检测关键装备”、“高性能 PCB 微钻用挤压硬质合金棒材关键技术”

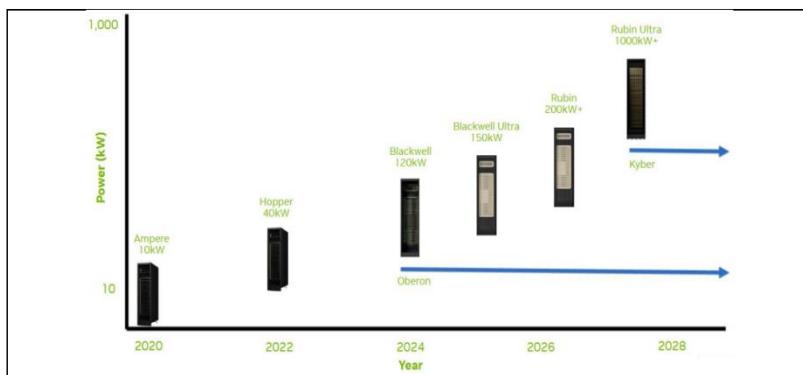
资料来源：各公司公告，国联民生证券研究所整理

内资 PCB 钻针企业迎来快速发展期。鼎泰高科作为全球 PCB 钻针龙头企业，2023 年以 26.5% 的全球市占率稳居第一；鼎泰高科刀具类专用材料企业中营业收入排名第 1 位；中钨高新（金洲）深耕 AI 相关高端 PCB 钻针；而沃尔德和四方达则布局 PCD (Polycrystalline Diamond, 指聚晶金刚石) 技术领域，为高端 PCB 加工提供了新的解决方案。

## 5 功率提升拉动电源+液冷升级

**单卡和机柜功率密度持续提升，对电力架构提出了新的要求。**随着算力的提升，单卡功率也得到了大幅提升，以英伟达为例，芯片功率从 H100 的 700W 提升到 B300 的 1400W。国产卡和 Asic 的功耗也不断提升。同时机柜功率从 Ampere 的 10kW 提升到了 Rubin Ultra 的 1MW，机柜功率提升了 100 倍，AI 工厂同步负载波动对电网和供电系统形成强烈冲击。传统电力架构在效率、可扩展性 (scalability) 和稳定性方面已无法满足下一代 AI 基础设施的需求，电力设备升级成为行业必然趋势。同时，功率密度的逐步提升也使得液冷成为数据中心的标配。

图66：NVIDIA GPU 机柜功率变化进程



资料来源：英伟达《800VDC Architecture for Next-Generation AI Infrastructure》，国联民生证券研究所

表 15：单卡功耗提升进程

	型号	发布时间	功耗 (W)	芯片类型
NVIDIA	H100	2022	700	推理+训练
	B200	2024	1000	推理+训练
	B300	2025	1400	推理+训练
华为昇腾	910b	2023	310	推理+训练
	910c	2025	310-350	推理+训练

资料来源：CSDN，腾讯网，鉴智库公众号，国联民生证券研究所

### 5.1 高压直流是未来柜外电源的趋势

**800VDC 架构成为关键解决方案。**该架构通过更高电压降低电流，同截面铜线传输功率较 415VAC 提升 57%，同时减少转换阶段，降低资本与运营支出，适配 SiC/GaN 等新型功率器件，可支撑 1MW 以上单机架密度，为 AI 基础设施提供高效、经济且具扩展性的电力解决方案。

表 16：机柜功率密度提升带来的挑战

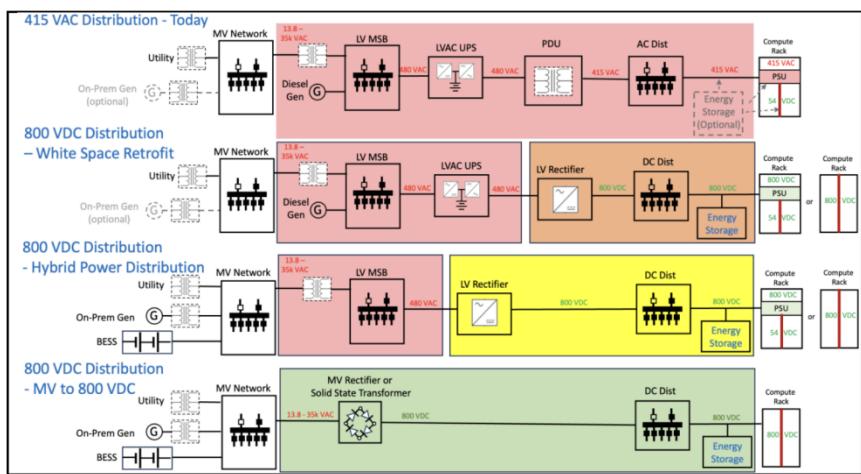
问题	描述
GPU 功率密度激增挑战	GPU 机架功率密度较传统 web 服务器提升近 100 倍。NVLink 技术的发展使单机架 GPU 数量从 32 个扩展到 72 个，功率密度从“代际 20% 增量”跃升至 2 倍、4 倍甚至 8 倍。传统 415/480VAC 架构的电缆、连接器因载流和空间限制完全失效，无法适配当前 GPU 的功率需求。
同步负载波动冲击	AI 工作负载（如 LLM）具有同步特性，导致机架功率在 30%（空闲状态）到 100%（峰值状态）间呈毫秒级波动。若不对电力系统进行升级，未缓解的波动会引发电网电压/频率偏差，甚至导致并网审批延误，严重影响供电稳定性与 AI 工厂的运营。
效率与成本瓶颈	传统 AC 架构存在多重转换环节（如 AC → DC → 低压 DC），不仅损耗大，且电缆体积庞大。在成本与效率层面形成明显瓶颈，难以支撑 AI 基础设施的长期发展。

资料来源：英伟达《800VDC Architecture for Next-Generation AI Infrastructure》，国联民生证券研究所

英伟达白皮书首次将中压整流器和固态变压器（SST）列为未来供电方案参考，推动行业从传统 UPS 向高压直流架构转型，形成三级技术演进路径：

- (1) **过渡方案（白色空间改造）**：在机架外侧增设专用 800VDC 电源架，集成整流器和配电单元，在现有基础设施条件下实现初步转型；
- (2) **可行方案（混合电力）**：从低压 AC 系统直接转换为 800VDC，通过公共 DC 母线整流，平衡转换效率与系统兼容性；
- (3) **未来方案（800VDC 配电）**：采用中压整流器或 SST，直接从中压电力系统转换至 800VDC，大幅简化转换环节，提升供电效率与集成度。

图67：数据中心电源架构



资料来源：英伟达《800VDC Architecture for Next-Generation AI Infrastructure》

**国内技术路线与国际同步**，2025 年 8 月中讯邮电、中数智慧发布的《数据中心 800V 直流供电技术白皮书 1.0》，同样将中压整流器和 SST 明确为未来主流供电方式。数据中心电源系统正从传统 UPS 向 HVDC、巴拿马电源及 SST 实现技术跃迁，核心性能差异如下：

其中，中压整流器（国内对应巴拿马电源）核心是通过移相变压器实现多脉波

**整流**,减少谐波污染,系统效率可达 98% 以上,已在阿里巴巴多地数据中心落地; SST 作为电力电子领域的原理级革命,以 SiC/GaN 等第三代半导体为核心,功率密度达传统变压器的 4 倍以上,体积仅为同容量工频变压器的 1/5-1/3。

**表 17: 数据中心电源演变对比**

电源类型	UPS	HVDC	巴拿马电源	SST
系统效率	92%	93%	97.5%	98.5%
占地面积 (相对比)	100%	97%	70%	<50%
转换环节	AC-DC-AC	AC-DC	AC-DC	AC-DC(直转)
技术成熟度	非常成熟	成熟	应用中期	研究期

资料来源: 张文丽等《数据中心巴拿马电源系统架构应用研究》, 《800V 直流供电技术白皮书 (1.0)》, 国联民生证券研究所

### 5.1.1 需求端: HVDC 星辰大海, 国内云厂商开始招标

从市场商业化进展来看,国内外头部企业均在推动 HVDC 技术落地与生态构建。国内方面,腾讯于 2025 年启动弹性直流一体柜招标;阿里巴巴早在 2018 年、2021 年分别推进高压直流及列头柜设备、巴拿马电源框架项目;中国电信 2025 年开展浙江公司中压直供电采购项目,金额 578 万元,国内云厂商与运营商的布局凸显 HVDC 在本土数据中心的商业化落地加速。海外市场中,Meta 规划 2026 年投运 1GW 级 AI 数据中心“普罗米修斯”,作为其人工智能基础设施大规模扩张的重要部分;谷歌在 2025 年 OCP EMEA 峰会上联合 Meta、微软及 OCP 社区启动“魔鬼山项目”推动±400V DC 标准化,并推广分体式电源机架架构,高效适配高功率密度需求,全球头部企业的技术协同与竞争,正加速 AIDC 供电架构的迭代进程。

**表 18: HVDC 产业进展**

厂商	进展
腾讯	2025年腾讯招标弹性直流一体柜
阿里巴巴	2021年阿里巴巴数据中心2021年巴拿马电源框架项目采购 2018年阿里巴巴数据中心项目高压直流及列头柜设备
中国电信	2025年中国电信浙江公司中压直供电采购项目金额: 578万
Meta	普罗米修斯(Prometheus): 计划于2026年投运的1GW级AI数据中心,被视为人工智能基础设施大规模扩张的重要部分
Google	在2025年OCP EMEA峰会上,谷歌联合Meta、微软及OCP社区启动“魔鬼山项目(Mt.DiabloProject)”,推动±400V DC的标准化进程,这种分体式电源机架将电源组件与计算设备分离,能够更高效地支持未来更高的功率密度。

资料来源:中国储能网,电力招标网,Digital Watch Observatory,OCP官方文档,Storage Review,中恒电气,国联民生证券研究所

### 5.1.2 供给端: 国内进展相对较快, 海外台达、伊顿领先

海外台达和伊顿进展较快,国内中恒电气和科华数据进展较快。台达电子进度相对领先,在 2025 年中国工博会上发布了固态变压器,作为新一代电力电子核心技术,荣获“CIIF 工业自动化奖”;伊顿在 2025 年 10 月推出新一代 800V 直流

电力架构，该设计支持英伟达构建 800V 高压直流的新型 AI 数据中心。维谛技术于 2025 年 5 月 19 日宣布与英伟达（NVIDIA）战略协同，计划 2026 年下半年正式推出其 800VDC 电源产品系列，该系列发布时间将早于 NVIDIA Kyber 和 NVIDIA Rubin Ultra 平台的发布节点；中恒电气参与了 2018 年阿里巴巴数据中心高压直流及列头柜设备项目（金额 3 亿元）、2021 年阿里巴巴数据中心巴拿马电源框架采购项目（金额 8 亿元）以及中国电信浙江公司中亚直供电源采购项目；科华数据在 2025 年加大 HVDC、SST 等关键核心新技术的研发力度及新产品应用，且与腾讯联合开发的直流一体柜已交付 1000 台。

**表 19：HVDC 产业进展**

厂商	进展
台达电子	进度相对领先，在 2025 年中国工博会上发布了固态变压器，作为新一代电力电子核心技术，荣获“CIIF 工业自动化奖”
维谛技术	2025 年 5 月 19 日宣布与英伟达（NVIDIA）战略协同。计划于 2026 年下半年正式推出其 800VDC 电源产品系列，该产品系列的发布时间将早于 NVIDIA Kyber 和 NVIDIA Rubin Ultra 平台的发布节点。
伊顿	伊顿在 2025 年 10 月推出新一代 800V 直流电力架构，该设计支持英伟达构建 800V 高压直流的新型 AI 数据中心 2018 年阿里巴巴数据中心项目高压直流及列头柜设备 金额：3 亿元
中恒电气	2021 年阿里巴巴数据中心 2021 年巴拿马电源框架项目采购 金额：8 亿元 中国电信浙江公司中亚直供电源采购项目
科华数据	2025 年科华加大关键核心新技术（HVDC、SST 等等）的研发力度及新产品应用 和腾讯联合开发的直流一体柜交付 1000 台

资料来源：中恒电气公告，科华数据公众号，科华数据公告，中国电信阳光采购网，台达官网，施耐德官网，伊顿官网，国联民生证券研究所

台达高压直流方案构建了涵盖 HVP 系列高压直流电源、HIP 高集成交流电力模块系统、10kV 中压直供电源、新一代 10kV 供电系统（SST）的产品矩阵。该方案以极致高效（系统效率达 98%，单柜功率最高 960kW，节省占地 60%）、极致可用（模块化设计、冗余控制、无单点失效）、快速交付（标准化预制+现场拼装）、高管理灵活性（全彩触屏+能耗分析、交直流双输出、兼容风光储新能源及多类型电池）为优势。其中，新一代 DP-SST 系列 SST 产品功率覆盖 840-3360kVA，系统效率≥98%，采用模块化冗余设计，支持线上线下进线自由组合，适配大型数据中心与工厂供电场景，为高压直流供电领域提供高效可靠的技术解决方案。

图68：台达高压直流方案



资料来源：台达官网，国联民生证券研究所

图69：台达 SST 方案

### 新一代10kV供电系统-SST

10kV/200Vdc~1000Vdc 840~3360kVA

DPSST系列 840-3360kVA极致高效、安全可靠、灵活组合、预制化设计。单系统采用N+3冗余，实现N+X架构，可靠性更高，简化供电架构、实现更高效率、更快部署、更灵活的供配电方案。内部采用预制化设计，柜体之间采用母排连接，减少了桥架等安装成本，输入功率因数0.99和输入电流谐波 iTHD<3%，满足供电电网要求。整机效率高达98%以上，实现更低的总拥有成本。是针对各种大型数据中心和制造工厂关键电源备份的最佳供配电解决方案。



资料来源：台达官网，国联民生证券研究所

伊顿高压直流方案构建了涵盖 800VDC 核心供电系统、高压直流保护装置、2MW SST 样机、预制舱式直流微网系统的产品矩阵。该方案凭借高效（800VDC 减少转换损耗，SST 效率达 98.3%）、适配（HVDC 适配英伟达 AI 算力设备，SST 适配 10kV 电网标准等优势，覆盖数据中心、新能源基建等多场景需求。其中，800VHVDC 电力系统是伊顿与英伟达合作的标杆产品，支持从电网到芯片的电力适配；2MW SST 样机采用高频变压器技术，适配 10 千伏电网，伊顿近日宣布已完成对 Resilient Power Systems 公司的收购，该公司专注于 SST 技术，这有利于 SST 技术在数据中心、储能等高增长市场的全球规模化应用；配套的高压直流保护装置覆盖 2A-4000A 电流范围；预制舱式直流微网系统整合光储充资源，整站损耗≤5%。该方案为高压直流与固态变压器应用领域提供了高效可靠的技术解决方案。

图70：伊顿 MVSST 数据中心应用场景实例



资料来源：伊顿公众号，国联民生证券研究所

中恒电气高压直流方案以 **Panama 电力模组** 为核心，集成 10kV 配电、变压器、不间断电源等单元，创新融合电路与磁路，优化数据中心供配电链路，提升电能转化效率、降低用电能耗，极简融合智能锂电与新能源，安全可靠性高。方案采用预制化部署、模块化扩容模式，大幅减少现场工程量、缩短配电建设周期，相比原有方案占地面积减少 50%，单套系统支持 2.4M WIT 负载供电，适配新一代智算中心对电源基础设施产品化、快速部署、超高效的高标准要求。**该方案已在阿里巴巴杭州余杭、内蒙古乌兰察布、江苏南通等多地数据中心落地应用，在大型 IDC 供电场景中表现突出。**

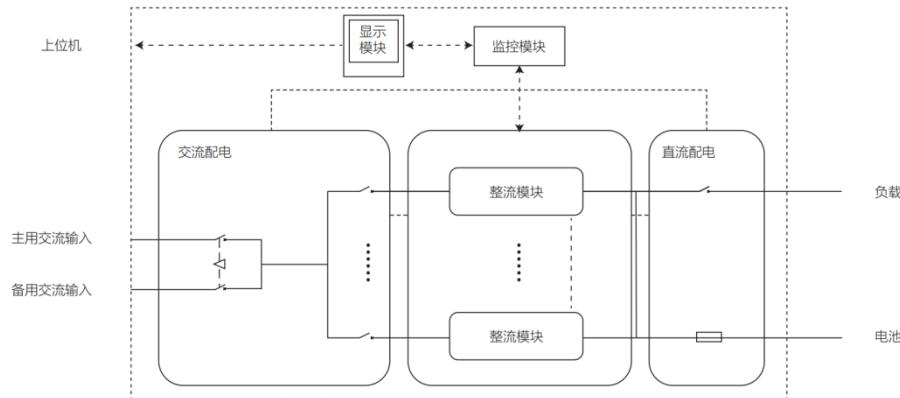
图71：中恒电气高压直流方案成功案例



资料来源：中恒电气官网，国联民生证券研究所

科华数据自主开发 ZL 系列高压直流电源系统。该系统主要由交流配电、直流配电、整流模块、监控模块和显示模块等组成，产品采用模块化设计、全数字化控制技术，具备自动休眠和电池的智能化管理功能，包括 240V、336V 两种电压制式，为客户提供高可靠电源保障。2025 年，科华与腾讯联合开发的弹性直流一体柜 1000 台交付，并对“弹性直流一体柜 2.0 联合开发”举行启动仪式。科华在 2025 年中报中提到“加大关键核心新技术（HVDC、SST 等等）的研发力度及新产品应用”。

图72：科华数据直流方案系统原理图

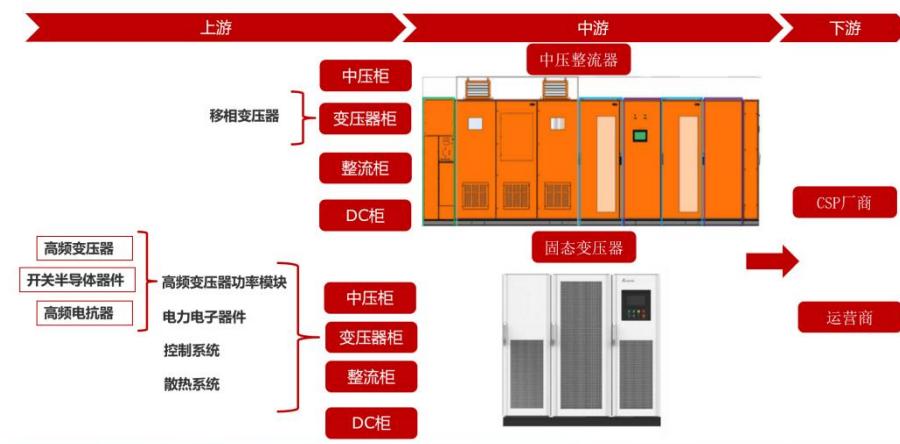


资料来源：科华数据官网，国联民生证券研究所

### 5.1.3 产业链：变压器是核心变化点

中压整流器和固态变压器在总体结构上是相似的，包括中压柜、变压器柜、整流柜和DC柜等构成。整个产业链的中游是系统集成商，下游是CSP厂商或者运营商。中压整流器和固态变压器的核心区别在于变压器柜部分。中压整流器的变压器柜内放置的是移相变压器；固态变压器的变压器柜内放的是高频变压器。

图73：产业链构成



资料来源：ODCC《巴拿马电源技术白皮书》，台达官网，国联民生证券研究所

我们认为，英伟达首次提出柜外直流方案，为解决机柜功率密度与电力损耗问题提供了新路径。这一高效解决方案正获得全球业界的广泛关注，形成明确的产业趋势。供给端，国内产业链发展迅速；海外市场亦在积极跟进，以谷歌为代表的云服务巨头（如其积极推进的“Prometheus”计划所代表的对高效基础设施的探索方向）也在大力布局直流技术。我们看好该赛道未来的发展，建议关注以下方向：

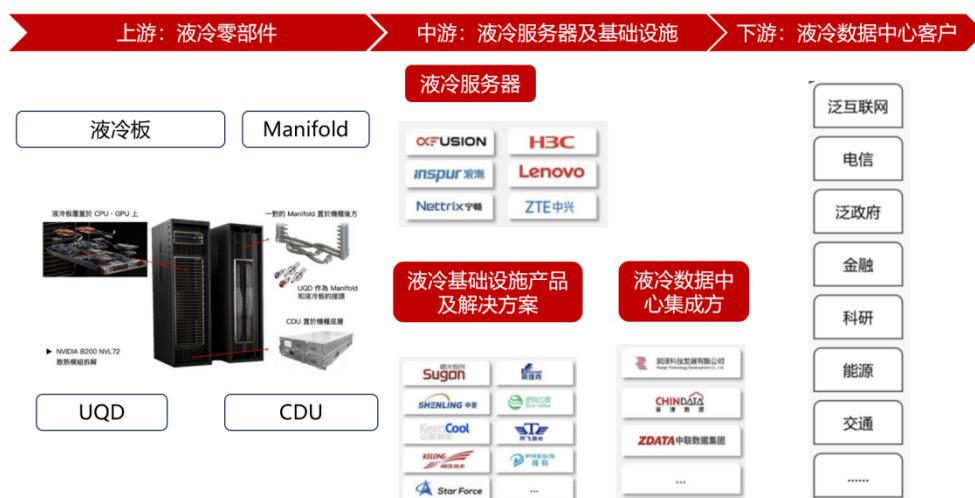
- 1) 终端厂商率先受益，建议关注：中恒电气、科华数据、阳光电源；
- 2) 变压器是核心变化，也最可能供给海外，建议关注：京泉华、伊戈尔、新特电气、金盘科技；
- 3) PCB供应商受益放量，建议关注：威尔高、中富电路；
- 4) 服务器代工厂

商：工业富联、华勤技术、联想集团。

## 5.2 液冷从 0-1 实现产业趋势突破

液冷数据中心产业链由上游液冷零部件、中游液冷服务器及基础设施和下游液冷数据中心用户构成。我们认为机柜液冷是一套完整的系统解决方案，具备系统化解决方案的公司未来会更具竞争力。但同时，零部件厂商也在加速推进认证。液冷零部件主要包括冷板、UQD、manifold、CDU、连接器、电磁阀、TANK 等。中游主要包括浪潮、联想等服务器制造商，以及曙光数创、英维克、申菱股份等液冷解决方案提供商；下游主要客户为互联网厂商、三大运营商、政府、科研院所等机构。

图74：液冷产业链拆分



资料来源：科智咨询《中国液冷数据中心市场深度研究报告》，NVIDIA 官网，CDCC，国联民生证券研究所

**拆分整个上游零部件来看**，冷板是零部件中市场空间最大的，而 UQD (Universal Quick Disconnect) 则是当前市场关注度较高的增量市场部件：

- **冷板是散热模块中的核心部件，因材料成本较高且加工工艺复杂，成本占比较大。** 我们根据 nv 机柜量测算，每个 Compute Tray 6 块冷板，每个 Switch Tray 1 块冷板。每个 NVL72 机柜有 18 个 Compute Tray，9 个 Switch Tray。总计 117 块冷板。
- **UQD 是一种通用快速连接器，主要用于液冷系统，尤其适用于数据中心和超级计算机的热管理应用。** GB300 采用了独立液冷板设计，每个芯片配备单独的一进一出液冷板，根据我们 GTC 拍摄的 GB300 的 compute tray 及 switch tray 所示，共 12 对快接头，相较于 GB200 的 6 对有所增加。
- **Mainifold 垂直放置于整机背部，每个 Tray 均与 Mainifold 连接，实**

现冷却液的分配及汇聚，即过液冷液。

- CDU 将冷却剂从冷源循环传至热源，掌控着冷却液的流量、温度等关键参数，如同液冷系统的“智能管家”，根据系统的实时需求，合理调配冷却液资源，让散热过程始终处于最佳状态。

图75：GB300 NVL72 compute Tray 示意图



图76：GB300 NVL72 Switch Tray 示意图



资料来源：Computex 展会，国联民生证券研究所

资料来源：Computex 展会，国联民生证券研究所

图77：GB300 NVL72 液冷方案：机房内实物拆解



资料来源：Computex 展会，国联民生证券研究所

我们于 2025 年 5 月参加了 Computex 展会，英伟达于会上展示了其“MGX Ecosystem”，并披露了官方认证供应链合作伙伴。

其中，液冷产业链目前仍以台系+海外厂商为主，A+H 供应链仅有工业富联、比亚迪电子和英维克。但诸多国内本土的液冷厂商正展现出强劲的发展态势。无论

技术能力、交付能力、项目经验等方面均可看齐全球龙头厂家，且在海外市场实现0-1的突破。

表 20：液冷产业供应链

产品	Nv 供应链	具备相关产品储备的国内厂商
冷板	工业富联、比亚迪电子、双鸿科技、奇鋐、酷冷至尊、台达、品达、立敏达、Boyd、CoolIT	思泉新材、硕贝德、英维克、申菱环境、奕东电子、科创新源、曙光数创、依米康、铂力特
UQD	工业富联、英维克、比亚迪电子、立敏达、双鸿科技、富世达、Danfoss、Parker、Staubli、Netonx	溯联股份、川环科技、硕贝德、瑞可达
Manifold	工业富联、比亚迪电子、双鸿科技、奇鋐、品达、台达、光宝	英维克、申菱环境、奕东电子
CDU	工业富联、维谛技术、光宝等	英维克、申菱环境、飞龙股份、科华数据、川润股份

资料来源：Computex 展会，公司公告，国联民生证券研究所整理

ASIC 供应链领域，全球云服务厂商如谷歌、Meta、微软、AWS 等大力推进自研 ASIC 布局，对液冷需求明确。Meta 预期在 2027 年发布的 MTIA T-V2，采用更大规模 CoWoS 封装与 170KW 高功率机架设计，可能需要配置液冷。

液冷渗透率大幅提升的当下，市场空间快速扩容，我们长期看好国内本土的液冷厂商的工程师红利。产业浪潮推动下，中国的液冷厂商有望后来居上，获取更高市占率，引领行业发展。

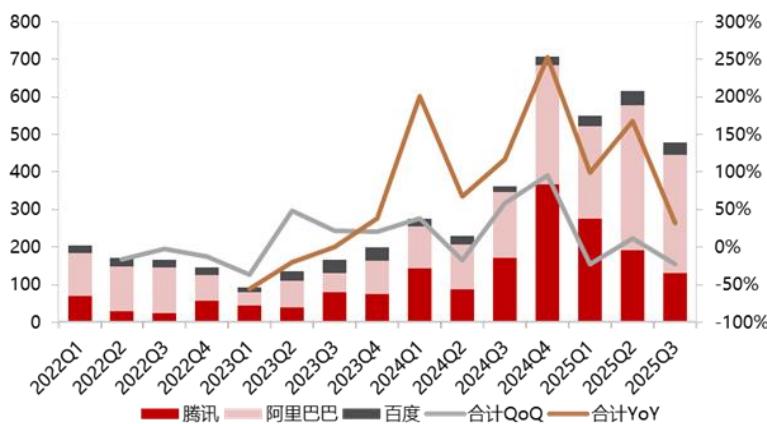
## 6 国产算力：需求侧资本开支展望积极，国产大模型加速追赶

### 6.1 资本开支：25 年受 H20 扰动，26 年展望积极

云商资本开支稳步修复，算力投入持续加码。BAT（百度、阿里、腾讯）合计资本开支从 2022 年 687.4 亿元增长到 2024 年 1574.5 亿元，年均复合增长率（CAGR）约为 51.3%，展现出在 AI 大模型、云计算基础设施和高性能算力部署上的全面提速。2025 年 Q3 BAT 合计资本开支同比提升 32.2% 至 479.0 亿人民币，环比较 25Q2 的 615.0 亿人民币下降了 22.1%。其中，腾讯、阿里巴巴和百度 2025 年 Q3 资本开支分别为 129.8、315.0 和 34.0 亿人民币，同比增速分别为 -24.1%、80.1%、106.8%，环比分别为 -32.1%、-18.6%、-10.5%。

政策扰动致 2025Q3 环比下滑，长期算力投入决心不改。2025 年 Q3 国内互联网厂商 BAT 合计资本开支环比下降，主要原因可能为美国对高端算力芯片出口管制政策的持续影响。但算力作为科技产业迭代升级的核心底层基座这一事实并未发生改变，BAT 或将会持续加码 AI 领域的研发与算力投入，全力角逐下一代算力技术的战略高地。

图78：BAT 资本开支及增速（亿元，%）

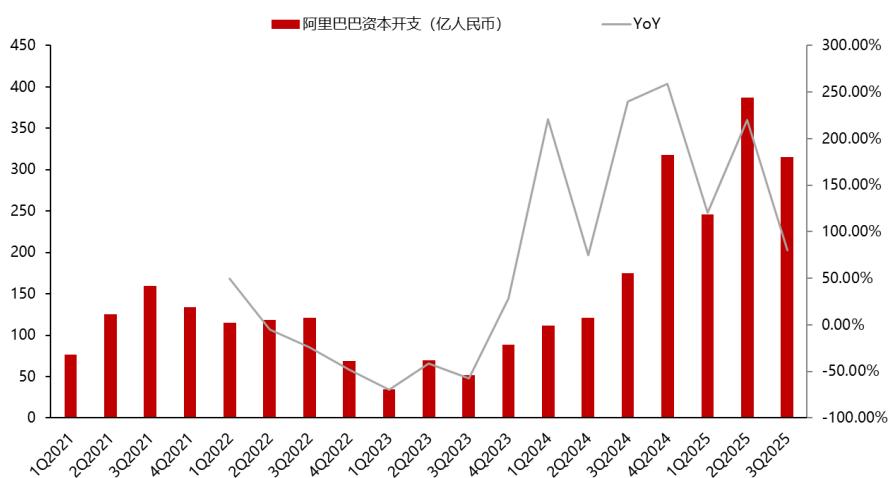


资料来源：各公司公告，国联民生证券研究所

阿里算力投入持续加码，资本开支三年 3800 亿元有望上修。阿里 2025 年 Q3 资本开支为 315.0 亿元，同比增长 80.1%，环比减少 18.6%，重点投向 AI 基础设施和技术先进性建设。阿里巴巴云业务作为 AI 驱动的核心业务，2025 年 Q3 云业务收入达 398.2 亿元，同比增长 34.5%，环比增长 19.2%，AI 相关产品收入连续九个季度实现三位数的同比增长。增长态势的核心推手，是公共云业务的持续放量，其中 AI 相关产品的市场认可度与采用规模正节节攀升，全栈式 AI 业务正

成为驱动阿里巴巴云收入快速增长的核心增长极。阿里原定三年 3800 亿的云和 AI 硬件基础设施资本开支，因服务器上架速度滞后于算力需求增长，公司对 AI 基础设施总体保持积极的投资姿态，存在进一步增投的可能性，展望 2026 年，阿里 Capex 规划有望进一步上修。

图79：2021Q1-2025Q3（自然年）阿里巴巴资本开支



资料来源：阿里巴巴公司公告，国联民生证券研究所

**腾讯资本开支受 H20 影响环比下滑。**腾讯 2025 年 Q3 资本开支为 129.8 亿元，同比下降 24.1%，环比下降 31.9%。对于 2025 年，腾讯预计资本支出将低于之前的指引区间，但将高于 2024 年的实际资本开支水平，维持在高位，调整主要受全球 AI 芯片供应链阶段性紧张的影响，而非战略收缩。

图80：2022Q1-2025Q3 腾讯资本开支



资料来源：腾讯公司公告，国联民生证券研究所

**展望 2026 年，国内主要云厂商的资本开支预计将持续增长，保持高景气加码态势。**国内云厂商预计持续加大对 AI 及算力基础设施的资本投入，补齐供给短板。受英伟达芯片无法供货的影响，国产算力采购的国产化进程预计将加速，国产算力领域不能单纯依赖海外芯片。远期来看，美国对英伟达 AI 算力卡实行出口限制，国内对于自主可控的追求必将掀起国产 AI 算力芯片浪潮，国产算力采购份额具备复合增长潜力。国内主要云厂商预计探索转向国内芯片企业采购产品的可能性，国产算力加速替代的态势逐渐明晰。

## 6.2 大模型：国产大模型弯道超车

### 6.2.1 国产大模型上市：冲刺“全球大模型第一股”

在“全球大模型第一股”的竞争中，智谱和 MiniMax 率先启动上市进程。

**2026 年 1 月 8 日，智谱正式登陆港交所，缔造“全球大模型第一股”。**此次 IPO，智谱发行价每股 116.20 港元，开盘上涨，市值超 550 亿港元。11 家基石投资机构共拟认购约 30 亿港元，同时收获 1164 倍超额认购，为智谱的业务扩张及研发投入提供资金支持。

**作为国内早期布局大模型的企业，智谱收入增长表现突出。**据弗若斯特沙利文数据，2024 年其收入在中国独立通用大模型开发商中排名第一，在所有通用大模型开发商中位列第二，市场份额 6.6%；截至 2025 年 6 月 30 日，累计服务超 8000 家机构客户。

**从营收方式来看，智谱通过其一体化 MaaS 平台提供大模型服务获取收入，根据客户需求提供本地化部署和云端部署两种部署方式。**智谱创业早期，高毛利率的本地化部署模式为其打下了“基本盘”，2022 年以来，智谱本地化部署的毛利率保持在 50% 以上，2025 年上半年，本地化部署业务的毛利率达 59.1%。此外，2024 年，智谱大模型本地化部署服务帮助开拓了海外市场。2025 年上半年，来自新加坡、马来西亚等国的东南亚客户为智谱本地化部署业务贡献了 11.1% 的收入。

**不过，从长期来看，智谱仍然看好并“押宝”云端部署，即大模型 API 调用模式。**据智谱披露，智谱云端 MaaS 和订阅业务呈现指数级增长趋势，中国前十大互联网公司中有 9 家使用智谱 GLM 大模型，付费流量收入超过所有国产模型之和。智谱将持续以 AGI 为奋斗目标，在 AI 编程、多模态、具身智能、智能体等前沿领域深化布局，通过 MaaS 模式实现技术商业化规模化扩张。

**2026 年 1 月 9 日，MiniMax (00100.HK) 在港交所上市。**MiniMax 以 165 港元的定价上限发行，截至 1 月 9 日收盘，报 345 港元/股，涨幅接近 110%，总市值超过 1050 亿港元。

**海外市场是 MiniMax 营收的主要阵地。**2025 年前九个月，公司营收同比增

长超过 170%，其中海外市场收入贡献占比超 70%。

**产品层面**，基于自研全模态模型，MiniMax 已面向全球推出一系列 AI 原生产品，包括海螺 AI、星野、Talkie 等，以及面向企业和开发者的开放平台。截至 2025 年 9 月，MiniMax 已有超过 200 个国家及地区的逾 2.12 亿名个人用户。

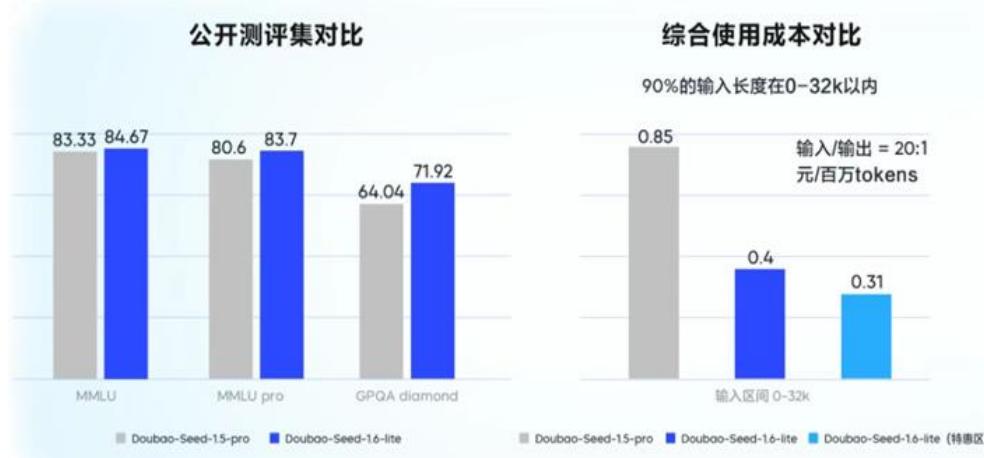
MiniMax 本次上市拟将 90%的募集资金主要用于未来五年的大模型升级与 AI 原生产品开发，提供越来越先进、通用的智能，帮助整个社会提升生产效率、提高发展质量与长期创造力。

### 6.2.2 豆包、DeepSeek 双线突破，国产大模型进入竞速新阶段

2025 年中国大模型产业进入竞速新阶段，以豆包、DeepSeek 为代表的主流国产模型持续领跑，在技术创新、场景落地与市场拓展上均展现出迅猛发展态势。

**豆包模型迭代升级，分档思考适配多元需求。**2025 年 10 月，火山引擎全新发布和升级了四款豆包大模型：豆包大模型 1.6 升级，原生支持 4 种思考长度；豆包大模型 1.6lite、豆包语音合成模型 2.0、豆包声音复刻模型 2.0。豆包大模型 1.6 全新升级，提供 Minimal、Low、Medium、High 四种思考长度，平衡企业在不同场景下对模型效果、时延、成本的不同需求，并进一步提升了思考效率，成为国内首个原生支持“分档调节思考长度”的模型。

图81：豆包和 DeepSeek 与其他大模型的测试对比



资料来源：火山引擎，国联民生证券研究所

**轻量化模型性价比突出，效果成本双优化。**以低思考长度为例，相比模型升级之前的单一思考模式，升级后的豆包 1.6 模型总输出 Tokens 下降 77.5%、思考时间下降 84.6%，模型效果保持不变。此外，火山引擎正式推出豆包大模型 1.6lite (DouBao-Seed-1.6-lite)，相比旗舰模型更轻量、推理速度更快、更具性价比。效果上，该模型超越豆包大模型 1.5pro，在企业级场景测评中提升 14%；价格上，

在 0-32k 输入区间里，综合使用成本较豆包 1.5pro 降低 53.3%。

**调用量与市场份额领先，商业化落地成效显著。**随着 AI 产业落地持续加速，截至 2025 年 12 月，豆包大模型日均 Tokens 调用量累计已突破 50 万亿，居中国第一、全球第三。在企业市场，IDC 报告显示，2025 年上半年中国公有云大模型服务市场，火山引擎以 49.2% 的份额占比位居中国第一。

**DeepSeek 以开源战略为核心，架构优化性能逼近顶尖。**DeepSeek 以开源为战略核心，采用自主改进的 DeepSeek MoE 架构，支持 128k 上下文，在数学、编程与通用逻辑等测试表现上已接近其他国际顶尖模型，成为“国产开源力量”的重要代表。2025 年 12 月，DeepSeek 同时发布两个正式版模型：DeepSeek-V3.2 和 DeepSeek-V3.2-Speciale。

图82：DeepSeek-V3.2 与其他模型在各类数学、代码与通用领域评测集上的得分

Benchmark	GPT-5 High	Gemini-3.0 Pro	Kimi-K2 Thinking	DeepSeek-V3.2 Thinking	DeepSeek-V3.2 Speciale
AIME 2025 美国数学邀请赛	94.6(13k)	95.0(15k)	94.5(24k)	93.1(16k)	96.0(23k)
HMMT Feb 2025 哈佛 MIT 数学竞赛	88.3(16k)	97.5(16k)	89.4(31k)	92.5(19k)	99.2(27k)
HMMT Nov 2025 哈佛 MIT 数学竞赛	89.2(20k)	93.3(15k)	89.2(29k)	90.2(18k)	94.4(25k)
IMOAnswerBench 国际数学奥林匹克竞赛	76.0(31k)	83.3(18k)	78.6(37k)	78.3(27k)	84.5(45k)
LiveCodeBench 世界级编程竞赛	84.5(13k)	90.7(13k)	82.6(29k)	83.3(16k)	88.7(27k)
CodeForces 世界级编程竞赛	2537(29k)	2708(22k)	-	2386(42k)	2701(77k)
GPQA Diamond 理工科博士生测试	85.7(8k)	91.9(8k)	84.5(12k)	82.4(7k)	85.7(16k)
HLE 人类全学科前沿难题测试	26.3(15k)	37.7(15k)	23.9(24k)	25.1(21k)	30.6(35k)

资料来源：DeepSeek 官方公众号，国联民生证券研究所

**双模型定位清晰，推理性能大幅提升。**DeepSeek-V3.2 的目标是平衡推理能力与输出长度，适合日常使用，在公开的推理类 Benchmark 测试中达到 GPT-5 的水平，仅略低于 Gemini-3.0-Pro；相比 Kimi-K2-Thinking，输出长度大幅降低，显著减少计算开销与用户等待时间。DeepSeek-V3.2-Speciale 是长思考增强版，结合了 DeepSeek-Math-V2 的定理证明能力，具备出色的指令跟随、严谨的数学证明与逻辑验证能力，在主流推理基准测试上的性能表现媲美 Gemini-3.0-Pro。

**突破工具调用局限，Agent 能力达开源顶尖。**不同于过往版本在思考模式下无法调用工具的局限，DeepSeek-V3.2 是首个将思考融入工具使用的模型，且同

时支持思考模式与非思考模式的工具调用。通过大规模 Agent 训练数据合成方法，构造了大量难解答、易验证的强化学习任务（1800+环境，85,000+复杂指令），大幅提高了模型的泛化能力。该模型在智能体评测中达到当前开源模型的最高水平，大幅缩小了开源模型与闭源模型的差距，且在真实应用场景中展现出较强的泛化性。

## 7 国产算力：供给侧向“芯”而行，国产算力破局元年

### 7.1 晶圆厂：国产算力底座

在国产大模型密集落地背景下，芯片厂商加速适配国产算力生态。中芯国际 N+1 工艺已逐步成熟，N+2 持续推进，构建国产算力底座；昇腾 910C 量产落地，920 系列研发加快，性能持续逼近国际主流水平；寒武纪、海光等在 AI 训推方向深度布局，硬件端多点突破，生态融合加快。云端 ASIC 正成为算力演进主流，谷歌、亚马逊持续加码自研芯片体系；国内芯原、灿芯等设计企业快速成长，覆盖多元应用场景，并与海内外头部厂商形成紧密合作，成长弹性充足。在软件层面，适配节奏同样加快，助力算力生态向自主可控稳步迈进。

图83：AI 算力产业链



资料来源：《财经》，芯存社，国联民生证券研究所

#### 7.1.1 大陆厂商代工渠道受阻，晶圆制造掣肘国产算力

**出口管制叠加审查收紧，先进制程代工遇阻。**美国政府限制台积电等海外代工厂为中国大陆客户提供先进制程的 AI 芯片代工服务，直接影响大陆芯片厂商相关先进制程芯片的代工供应。同时，美国不断拉高对大陆 AI 相关先进制程芯片的代工限制，台积电也逐步收紧对大陆客户的审查。美国当地时间 2025 年 1 月 15 日，美国商务部工业与安全局（BIS）出台新的对华出口管制法规（EAR），要求前端半导体制造工厂和外包半导体封装与测试（OSAT）厂商对使用 16/14 纳米节点或以下先进制程节点的芯片进行更多尽职调查程序，该新规生效后已影响部分中国芯片厂商的相关先进制程芯片生产与交付。

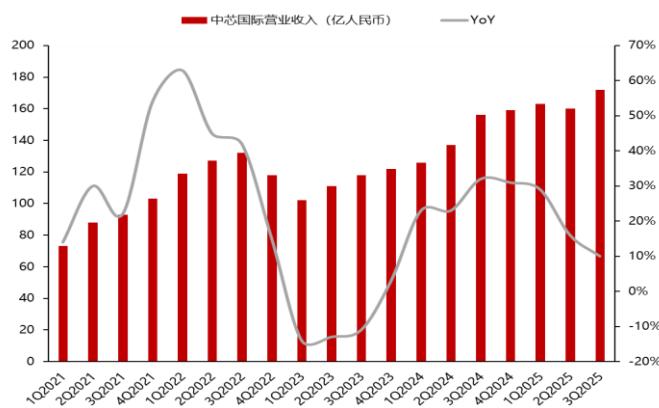
**代工渠道梗阻倒逼自主突破，筑牢产业安全防线。**台积电先进制程的晶圆代工渠道梗阻，会导致大陆相关芯片产能出现缺口，更会倒逼国产晶圆制造产业加速突破技术壁垒。唯有实现先进制程晶圆制造的自主可控，才能筑牢国产算力的供应链安全防线，为 AI 大模型等领域的持续发展提供核心支撑。事实上，先进制程晶圆

制造的自主可控，不仅是解决当前产能缺口的应急之策，更是筑牢国产算力长远发展根基的战略之举。当国产晶圆制造能够稳定供应先进制程的全系列产品时，不仅能为国内 AI 大模型训练、超算中心建设、智能终端升级等领域提供充足的算力支撑，更能带动上下游产业链的协同升级，形成“算力需求牵引制造突破，制造突破赋能算力升级”的良性循环，推动国产算力产业实现从依赖外部到自主引领的质变，在全球算力竞争中构建起独特的产业优势。

### 7.1.2 中芯国际：领航先进制程，铸就国产算力核心底座

**中芯国际营收稳步攀升，净利润大幅回暖。**2024 年中芯国际实现营收 578 亿元，同比+28%，创历史新高，实现归母净利润为 37 亿元，同比-23%。营收增长主要受本土化制造需求带来的产业链的重新组合、客户市场份额的提升、产能规模的扩大、新增产能能够较快地完成验证并投入生产及国家刺激消费政策的拉动；2025 年 Q3 中芯国际实现营收 172 亿元，同比+10%，环比+8%，实现归母净利润为 15 亿元，同比+43.1%，环比+60.7%。营收增长主要受益于国产替代进程中，公司合作客户在产业链中份额提升，家电、消费电子等多品类电子产品供应链的需求持续旺盛增长，进一步助推了订单规模的增长。

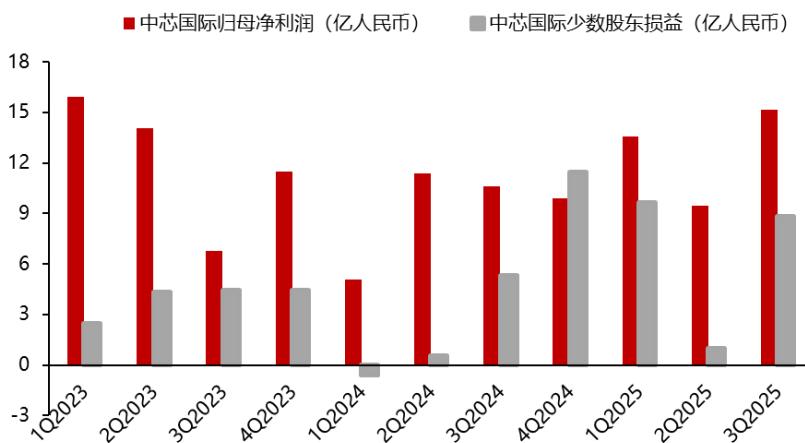
图84：2021Q1-2025Q3 中芯国际单季度营业收入



资料来源：iFinD，中芯国际公司公告，国联民生证券研究所

**少数股东当期损益同比大幅改善，先进制程盈利逐步释放。**25 年 Q3 中芯国际少数股东损益 8.85 亿元，25 年 Q2 1.02 亿元，环比增长 785.7%，较 24 年 Q1 -0.59 亿元大幅改善并维持高位。中芯国际盈利释放，凸显公司先进制程工艺进展明显。展望未来，在地缘政治推动算力自主可控的背景下，先进制程需求有望在 2026 年进一步放量，公司依托先进工艺发展核心逻辑，成长前景依旧向好。

图85：2023Q1-2025Q3 中芯国际单季度归母净利润及少数股东损益

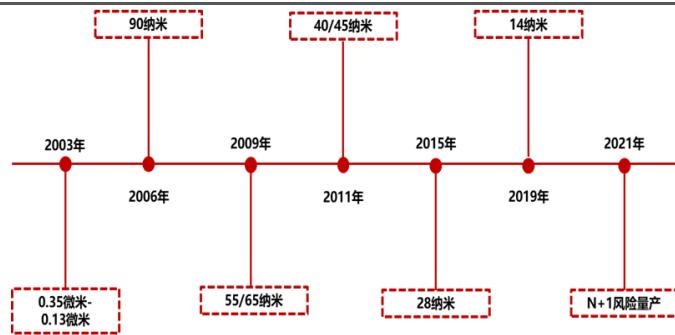


资料来源：iFinD，国联民生证券研究所

**25年Q4指引：**展望Q4，中芯国际指引销售收入将环比持平到增长2%。出货数量预计“淡季不淡”，毛利率指引为18%到20%。中芯国际预计继续保持满载，但受年底客户主动调控出货规模、存储器“超级周期”催生的价格承压态势及供应端波动风险等因素影响，手机、汽车、消费电子等领域的终端厂商对来年生产布局普遍持谨慎态度，这也让代工行业遭遇竞争加剧的挑战。

**在先进制程方面，中芯国际围绕Fin FET路线持续演进，已逐步形成较为完整的代工能力体系。**2019年公司率先实现14nm Fin FET的量产，并主要应用于AI与高性能低功耗计算等领域。公司进一步推出N+1工艺，于2020年完成测试及流片，第一次采用四重图形处理(SAQP)形成Fin架构，并由自对准双重成像技术(SADP)形成Dummy Gate。N+1工艺功耗较14nm下降57%，性能提升20%。最新的N+2制程也在加速突破中，制程有望等同于7nm工艺。尽管中芯国际在技术水平上仍与台积电存在一定差距，但通过对成熟制程的持续优化与重构，逐步推进先进工艺的国产替代。

图86：中芯国际关键技术节点的量产时间

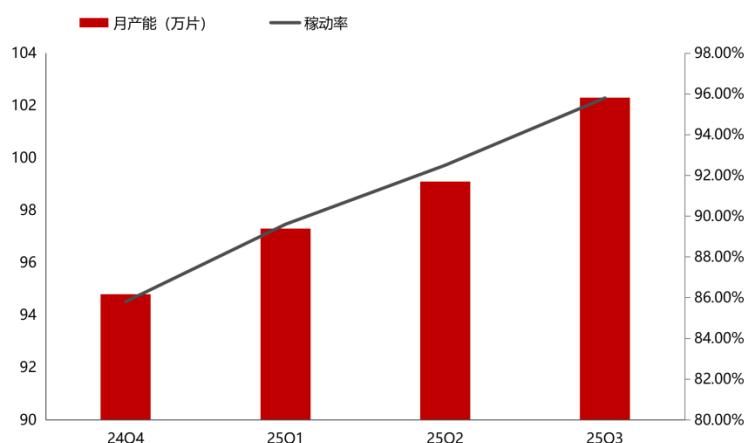


资料来源：iFinD，中芯国际招股书，投资者活动记录表，国联民生证券研究所

**产能规模持续扩大，利用率保持高位。**截至25年Q3，公司折合8英寸月产

能达到 102.3 万片，同比+15.7%/环比 3.2%（计算过程）；稼动率在新产能逐步释放的情况下达到 95.8%，同比+6.0pcts/环比+3.6pcts；25 年 Q3 折合 8 英寸晶圆出货量为 249.9 万片，同比+17.8%/环比+4.6%。

**图87：24Q4-25Q1 中芯国际月度产能（折合 8 英寸，万片/月）及稼动率**



资料来源：iFinD，国联民生证券研究所

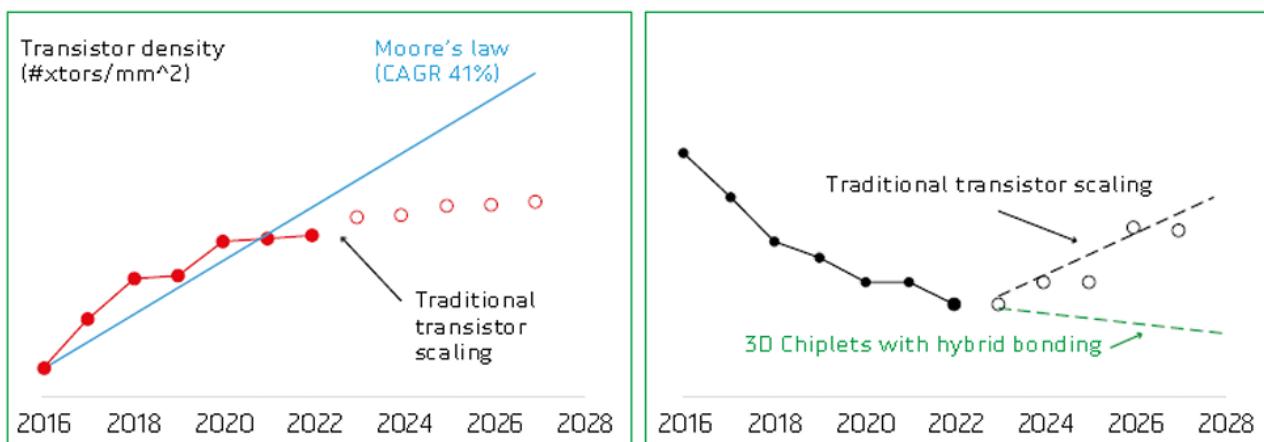
## 7.2 先进封装助力摩尔定律延续

**为什么需要先进封装？**从先进封装的价值和必要性来看，制程层面摩尔定律放缓，晶体管密度提升速度放缓，单晶体管成本触底。因此需要利用先进封装，从系统层面维持摩尔定律。同时芯片面积突破限制，大芯片对封装技术提出更高要求。此外先进封装可以帮助整体芯片方案降本增效，比如 Chiplet 可以将大芯片拆分为小芯片单元，提高部分模块的良率，从而降低成本。

**图88：摩尔定律放缓&单晶体管成本提升**

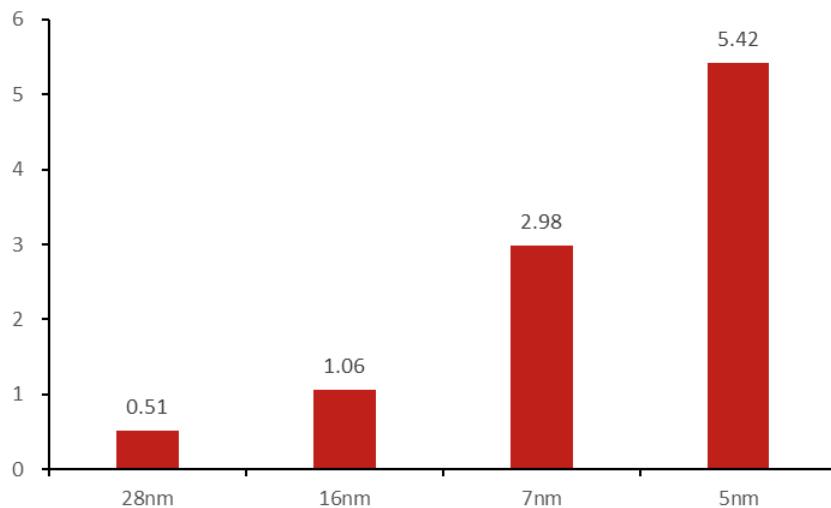
**Moore's Law Scaling Is Slowing**

**Cost Per Transistor Is Increasing**



资料来源：Besi，国联民生证券研究所

图89：不同制程芯片开发成本（亿美元）



资料来源：IBS，国联民生证券研究所

AI 时代，先进封装成为算力芯片的关键技术突破方向。过去的发展历程中，先进封装沿着两大方向发展：

**1、单芯片封装体积减小，密度提升。**从传统的引线框架封装，到倒装芯片封装，到更先进的 WLCSP 封装，单芯片的封装体积逐渐减小，封装密度提升。该技术主要用于消费类产品。

**2、多芯片集成。**从传统的单芯片封装到 chiplet 技术下的 2.5D/3D 封装，芯片封装集成度逐步提升。该技术方向主要应用于算力类产品。

图90：先进封装发展两大路径



资料来源：艾邦半导体，2011 IEEE 13th Electronics Packaging Technology Conference, ansysilicon, anandtech, nexpcb, semiengineering, 国联民生证券研究所

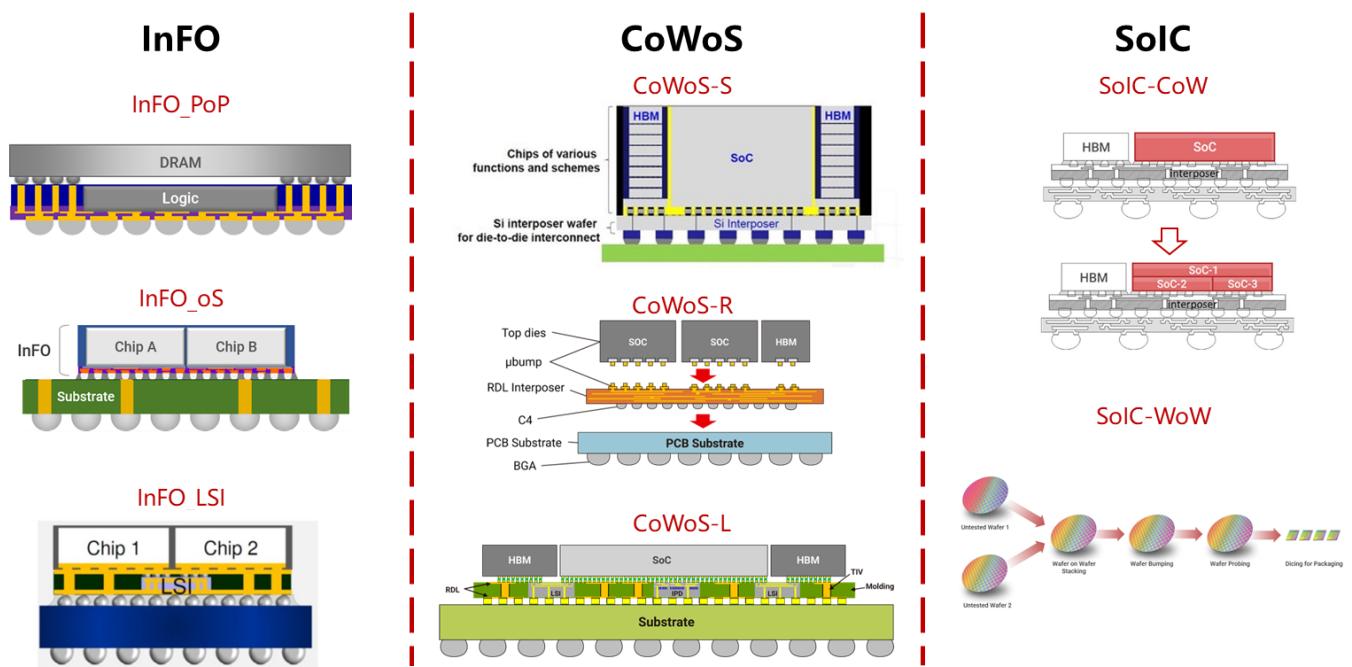
台积电 InFO、CoWoS、SoIC 是具有代表性的高端先进封装技术。

InFO: 无 Interposer, die 直接在载板上互联。成本低、互联速率低。其中 InFO\_LSI 是折中方案，在载板中内嵌了局部硅桥。多用于消费类、通信类，M1 Ultra 采用此种方案。

CoWoS: 有 Interposer, die 在 interposer 上互联, 再封在载板上, Interposer 方案也分硅基、有机两种；CoWoS 造价高，多用于算力芯片（Nvdia H100）。

SiIC: 晶圆键合，前道堆叠，而非后道封装环节堆叠，2021 年发布，目前用于 AMD MI300 系列。

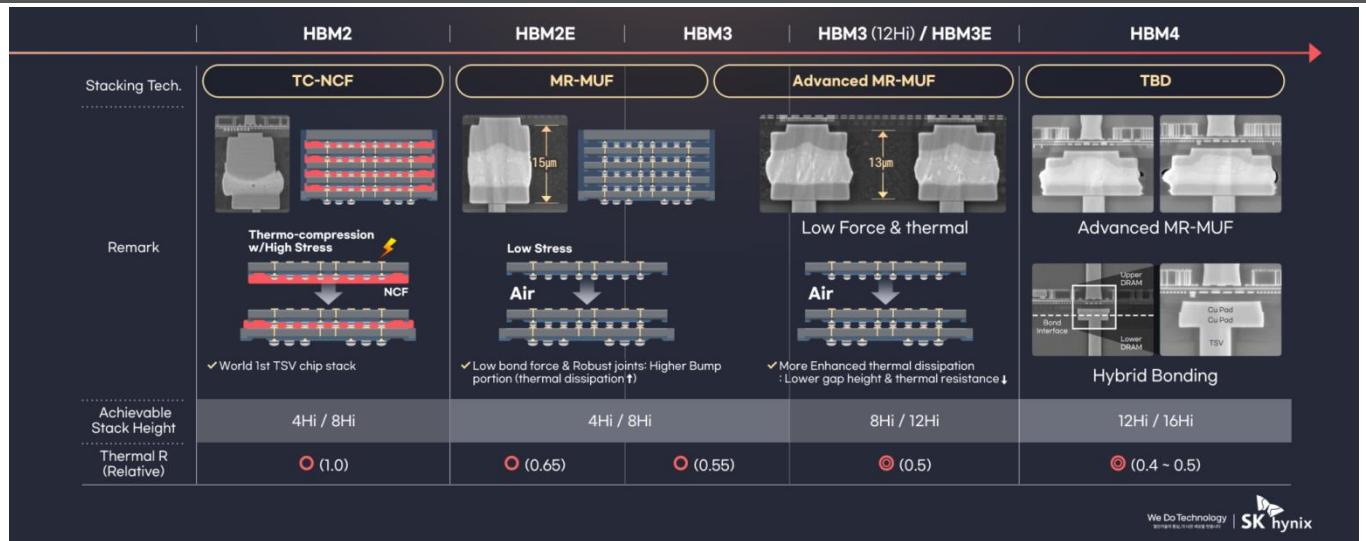
**图91：台积电用于 Chiplet 技术的三种封装工艺**



资料来源：TSMC, eetimes, 国联民生证券研究所

**HBM 成为增速最快的存储芯片，对先进封装提出更高要求。**当前 HBM 的主流方案为 MR-MUF (海力士) 和 TC-NCF (三星、美光) 两种。海力士从 HBM2E 开始从 TC-NCF 切换至 MR-MUF，因 MUF 方案有更低的热阻抗和更好的散热。两种技术路径均有 TCB 需求。TC-NCF 直接使用 TCB 进行堆叠，MR-MUF 则是使用 TC 进行堆叠和 pre-bonding，之后再进行 MUF 填充和 Mass Reflow。三大原厂的技术路径都将带来 TCB 封装的需求增量。

图92：海力士 HBM2-HBM4 的技术路径



资料来源：SK Hynix 2023 Tech Seminar 资料，国联民生证券研究所

**先进封装市场规模不断增长。**根据 Yole 数据，先进封装市场规模 2024 年为 519 亿美元，2028 年有望增长至 786 亿美元，2024-2028 年复合增速约为 11%。未来增长驱动力主要来自于 FC 倒装和 2.5D/3D 封装。FC 倒装是占比最大的先进封装类别，其次是 2.5D/3D 封装和 SiP 封装，FO 和 WLCSP 占比较小。

图93：先进封装市场规模

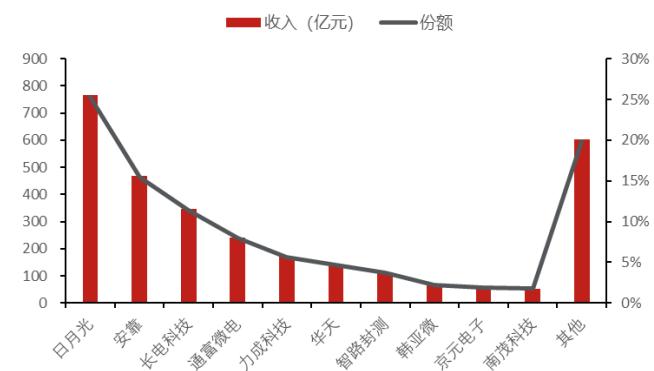


资料来源：Yole，国联民生证券研究所

竞争格局方面，从 OSAT 来看，前十大 OSAT 中，以中国台湾和中国大陆厂商为主，此外 1 家美国、1 家韩国。前十大合计占比 78%。先进封装领域参与者包括 IDM、Foundry、OSAT，前十大内部占比分别为 34%/33%/12%。前十大合计占比 80%。从中国大陆本土 OSAT 来看，中国本土厂商以传统封装为主，10 亿元收入以上本土 OSAT 达到 12 家。先进封装搭配先进制程，受此影响，中国大陆

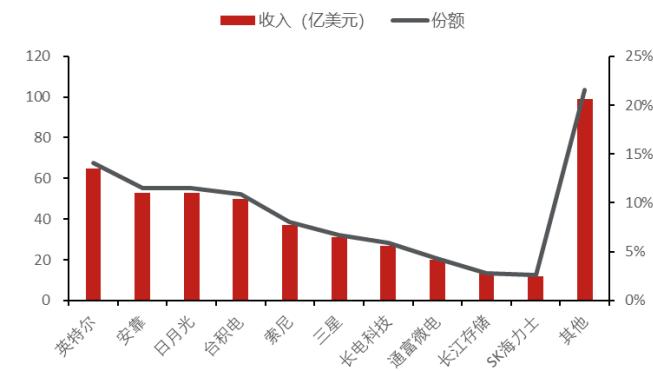
市场先进封装需求和产能扩张均受到限制，近年来增速低于中国台湾地区。

图94：2024年全球前十大OSAT公司



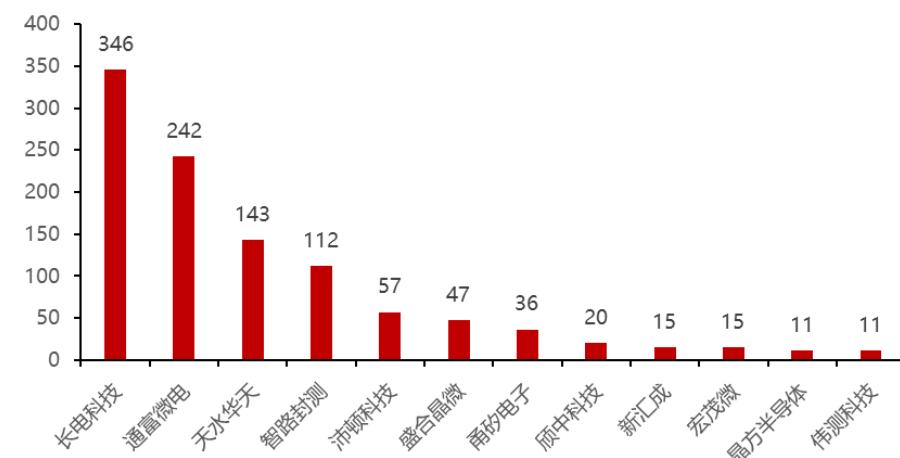
资料来源：TrendForce，国联民生证券研究所

图95：2024年全球前十大先进封装公司



资料来源：Yole，国联民生证券研究所

图96：2024年收入10亿元以上的中国本土OSAT公司



资料来源：芯思想研究院，国联民生证券研究所

## 7.3 算力芯片：技术迭代加速，国产替代格局明晰

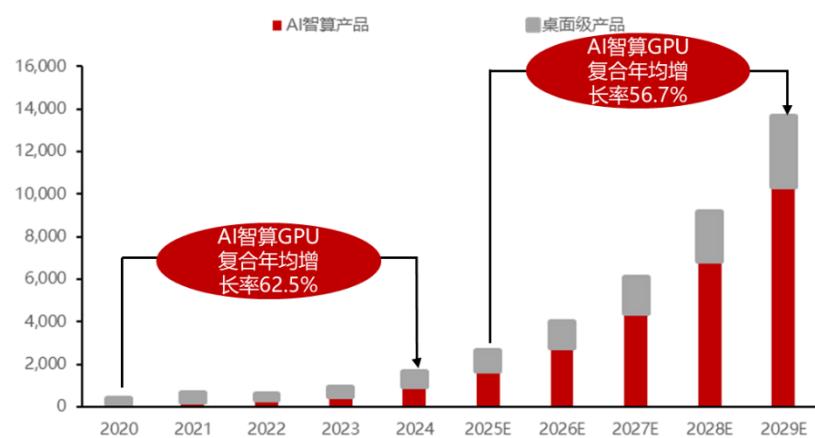
### 7.3.1 算力芯片产品对比

**芯片成算力核心支撑，技术路线多元发展。**当前国产算力芯片厂商主要采用GPGPU芯片与ASIC芯片两种技术路线。GPU依然是国内AI市场的主导芯片。不过，以ASIC和FPGA为代表的其他类型芯片也已实现商业化，并在市场中占据一定比例。

**GPU市场规模快速扩张，AI智算成增长引擎。**过去五年，中国GPU产业呈现快速增长态势，市场规模从2020年的384.77亿元快速增长到2024年的

1,638.17 亿元。GPU 产业下游应用领域可细分为 AI 智算产品和桌面级产品。未来，随着 AI 的应用不断开发，对于 GPU 等算力基础设施的需求预计将会出现快速增长。根据弗若斯特沙利文预测，预计到 2029 年中国 GPU 市场规模将增长到 13,635.78 亿元。中国 AI 智算 GPU 的市场规模从 2020 年的 142.86 亿元迅速增至 2024 年的 996.72 亿元，期间年均复合增长率高达 62.5%。未来，随着 AI 不断发展，对算力的需求预计将呈现指数级增长，根据弗若斯特沙利文预测，到 2029 年，AI 智算 GPU 市场规模将达到 10,333.40 亿元，期间年均复合增长率为 56.7%。此外，桌面级产品的市场规模未来也将保持稳定增长，从 2020 年的 241.91 亿元增至 2024 年的 641.45 亿元，预计 2029 年将进一步增至 3,302.38 亿元。

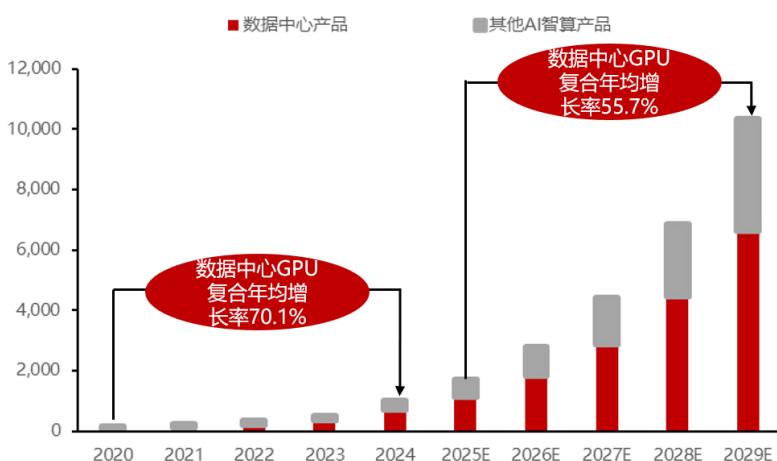
**图97：2020-2029 年中国 GPU 市场规模收入（单位：亿人民币）**



资料来源：摩尔线程招股说明书，弗若斯特沙利文，国联民生证券研究所

**数据中心产品增速领跑，细分市场前景广阔。**在中国 AI 智算 GPU 市场中，数据中心 GPU 产品是过去增速最快的细分市场，其市场规模从 2020 年的 82.00 亿元以 70.1% 的年均复合增长率，快速增长至 2024 年的 687.22 亿元，根据弗若斯特沙利文预测，未来还将以年均复合增长率 55.7% 的高增速增长至 2029 年的 6,639.16 亿元。其他 AI 智算产品（包括边缘计算和云计算产品）预计在未来将以 58.8% 的增速在 2029 年达到 3,694.24 亿元。

图98：2020-2029年中国AI智算GPU市场规模收入（单位：亿人民币）



资料来源：摩尔线程招股说明书，弗若斯特沙利文，国联民生证券研究所

**国产算力芯片生产制造厂商加速追赶，形成梯队竞争态势。**核心代表厂商主要包括：以华为昇腾、寒武纪、芯原、灿芯为代表的国产 ASIC 厂商和以海光信息、沐曦、摩尔线程为代表的国产 GPU 厂商等。

表 21：国产算力芯片主流“玩家”

国产算力卡主流“玩家”				
公司	型号	制程	算力	详细解释
昇腾	910C	7nm	800TFLOPS FP16	昇腾 910C 采用双 Die 封装以及达芬奇架构。由两颗昇腾 910B 垂直拼接形成，采用 2+8 架构（2 个 GPU 带+8 个 HBM）。推理能力强，适合大规模部署。
	950PR	7nm	1PFLOP@FP8/2PFLOPS@FP4	昇腾芯片在微架构方面全面支持 SIMD，NVIDIA GPU 正式均基于 SIMD 架构实现了 GPU 高效并行计算。除此以外，三代昇腾芯片在数值类型方面均支持 MXFP4
寒武纪	思元 590	7nm	314.6TFLOPS@FP16	思元 590 采用 MLUarch05 全新架构，实测训练性能较在售产品有了显著提升，提供了更大的内存容量和更高的内存带宽
海光信息	BW100	7nm	350TFLOPS@FP16	BW100 基于 x86 架构优化，其单卡可用性能实测达到 87% 左右，采用了较为先进的交换芯片设计，在单卡上传方面表现良好
沐曦 MetaX	曦云 C500	7nm	280TFLOPS@FP16	曦云 C500 是一款专为高性能计算和 AI 大模型推理设计的通用计算 GPU，内置 MXC500 芯片，其核心特性体现在强大的多精度混合算力、大容量高带宽显存、先进的多卡互联技术以及全兼容主流 GPU 生态的软件栈
摩尔线程	MTT S4000	7nm	100TFLOPS@FP16	MTT S4000 是基于摩尔线程全功能 GPU 架构专为大模型打造的训推一体通用计算卡，配备 48GB 高性能显存，凭借摩尔线程自研 MTLink 1.0 技术，可实现多卡互联及千卡集群部署，为千亿参数大模型的训练、微调和推理提供强劲算力支撑

平头哥	含光 800	12nm	825TOPS@INT8	含光 800 基于 12nm 工艺与自研架构，性能峰值算力达 825TOPS。主要用于云端视觉处理场景，性能打破了现有 AI 芯片记录，性能及能效比全球第一，在 ResNet-50 测试中，含光 800 推理性能达到 78563IPS，比目前业界最好的 AI 芯片性能高 4 倍
昆仑芯	P800	7nm	256TOPS@INT8	P800 基于 7nm 制程的昆仑芯 2 代 AI 芯片打造，专为深度学习、机器学习算法的云端和边缘端计算而设计，提供强大的 AI 负载运行效率
燧原科技	云燧 T20/T21	12nm	320TOPS@INT8	两款训练卡云燧 T20 和 T21 支持更高性能计算需求，目标市场为大模型推理和视频图像编解码

资料来源：智通财经网，华尔街日报，沐曦官网，芯智讯，巨丰金融研究院，芯存社，21ic 电子网，摩尔线程官网，平头哥官网，EET，国联民生证券研究所整理

### 7.3.2 昇腾：产品加速迭代，超节点引领行业

**昇腾 AI 芯片路线图重磅发布，逐年升级加速追赶。**路线图显示，华为在 2025 年 Q1 已推出了昇腾 910C，后续将在 26 年 Q1 推出昇腾 950PR，Q4 推出昇腾 950DT，27 年 Q4 将推出昇腾 960，28 年 Q4 推出昇腾 970。从时间节奏来看，昇腾后续将保持每年 1-2 次的全新升级：此次更清晰、更长期的芯片战略规划正式拉开了以昇腾、寒武纪为首的国产算力公司与英伟达在高端 AI 市场正面竞争的序幕。从芯片迭代性能指标来看，三代昇腾芯片均搭载自研低成本 HBM 技术，聚焦算力、互联带宽、内存性能等核心指标升级：

- 1) 算力：**昇腾 950 系列单芯片算力达到 1PFLOPS (FP8) /2PFLOPS (FP4)，960 与 970 系列则相较前序型号实现算力翻倍；
- 2) 互联带宽：**950 系列互联带宽较当前主力产品昇腾 910C 提升 2.5 倍，达 2TB/s，960 与 970 系列互联带宽则分别升级到 2.2TB/s、4TB/s；
- 3) HBM 容量与带宽：**昇腾 950PR 为 128GB、1.6TB/s，昇腾 950DT 为 144GB、4TB/s，960 与 970 系列则分别升级到 288GB、9.6TB/s 以及 288G、14.4TB/s。除以上重磅升级以外，我们同样注意到：从 950 系列开始，三代昇腾芯片在微架构方面全面支持 SIMD，NVIDIA GPU 正式均基于 SIMD 架构实现了 GPU 高效并行计算。除此以外，三代昇腾芯片在数值类型方面均支持 MXFP4。我们认为以上产品升级代表着国产 AI 算力芯片已经正式跻身世界一流方阵，与国际产品演进全面接轨，国产算力加速腾飞。

表 22：华为昇腾芯片路线图

芯片	昇腾 910C	昇腾 950PR	昇腾 950DT	昇腾 960	昇腾 970
时间	2025 年 Q1	2026 年 Q1	2026 年 Q4	2027 年 Q4	2028 年 Q4
微架构	SIMD	SIMD/SIMT	SIMD/SIMT	SIMD/SIMT	SIMD/SIMT
数值类型	FP32/HF32/FP16/B F16/INT8	FP32/HF32/FP16/B F16/FP8/MXFP8/Hi F8/MXFP4	FP32/HF32/FP16/B F16/FP8/MXFP8/Hi F8/MXFP4/HiF4	FP32/HF32/FP16/B F16/FP8/MXFP8/Hi F8/MXFP4/HiF4	FP32/HF32/FP16/B F16/FP8/MXFP8/Hi F8/MXFP4/HiF4
互联带宽	784GB/s	2TB/S		2.2TB/s	4TB/s
算力	800TFLOPS FP16	1PFLOPS FP8; 2PFLOPS FP4		2PFLOPS FP8; 4PFLOPS FP4	4PFLOPS FP8; 8PFLOPS FP4
内存	128GB, 3.2TB/s	128GB, 1.6TB/s	144GB, 4TB/s	288GB, 9.6TB/s	288GB, 14.4TB/s

资料来源：华为官网，芯智讯，国联民生证券研究所

**超节点成为 AI 基础设施建设新常态，内部互联能力成为重中之重。**华为在 Coud Matrix 384 超节点基础上，还将推出 Atlas 950 SuperPoD 和 Atlas 960 SuperPoD，分别支持 8192 及 15488 张昇腾卡，将于 26Q4 和 27Q4 上市。我们认为此前市场更多关注芯片算力，但伴随 Scale up 产业趋势崛起，超节点内部的互联能力成为重中之重。此次华为超节点主要有以下核心技术：

**1) 全光互联：**Atlas 950 超节点满配包括由 128 个计算柜、32 个互联柜，共计 160 个机柜组成，柜间采用全光互联，具有高可靠、高带宽、低时延等优势，OCS 产业趋势进一步明确；

**2) 正交背板：**Atlas 950 通过正交架构，实现零线缆电互联，其独创的材料和工艺让光模块液冷可靠性提升 1 倍。

**超节点速率大幅提升，功率迎来新挑战。**对于功率方面，超节点单机柜功耗普遍突破 100kW（如华为 CM384 达 172.8kW，英伟达 GB200 NVL72 约 120-140kW），在算力密度指数级增加的情况下，对超节点机柜的温控和电源系统提出挑战。液冷方面，据 2025 年 7 月 WAIC 展会来看，CM384 架构采用液冷加风冷的模式，液冷覆盖 70%。而 Atlas 950 SuperPoD 即为全液冷数据中心超节点。对于速率方面，Atlas 950 超节点互联带宽达到 16.3PB/s，FP4 算力达 16EFlops，训练性能达 4.91M TPS，推理性能达 19.6M TPS，超节点的互联带宽速率和算力速率迎来大幅提升。

**商用落地成效显著，客户合作持续深化。**昇腾芯片已实现多行业商用落地，基于昇腾 910C 打造的产品 Atlas 900 超节点上市以来累计服务超过 20 家客户，覆盖互联网、电信、制造等重点领域，产品可靠性与场景适配性得到市场验证。云厂

商合作方面，华为云已将昇腾芯片作为核心算力支撑，积极构建基于昇腾的云服务能力；行业内主流云厂商及互联网企业正加速与昇腾芯片的适配对接，针对大模型训练、推荐算法等核心场景的测试与应用筹备工作有序推进。

### 7.3.3 寒武纪：ASIC 架构优势凸显，商用落地持续突破

**技术布局全面，ASIC 架构优势显著。**寒武纪是全球知名的智能芯片企业，聚焦芯片技术突破与商业化落地，具备从云到边、从硬件到系统软件的完整能力。当前全新一代云端 AI 训练芯片思元 590 采用 7nm 制程，采用 ASIC 架构设计，具备低功耗、高效率，单卡算力达 512 TOPS，对标 A100。在存储架构方面，思元 590 的 MTP 与 DDR/HBM 间的带宽大幅优化，支持 L2 Cache 读写缓存，MTP Cluster 的访存带宽较上代提升 4 倍，支持大规模数据处理。在国内维度，寒武纪站在第一梯队。相比海光、摩尔线程，它在算力和显存带宽上占优，仅在部分场景落后于华为昇腾 910B。思元 590 已部署在智能安防、自动驾驶等多元应用场景。值得关注的是，搭载思元 590 的阿里云智能计算集群在 Llama-3 训练任务中已实现训练成本下降 40%，进一步验证其商业应用潜力。

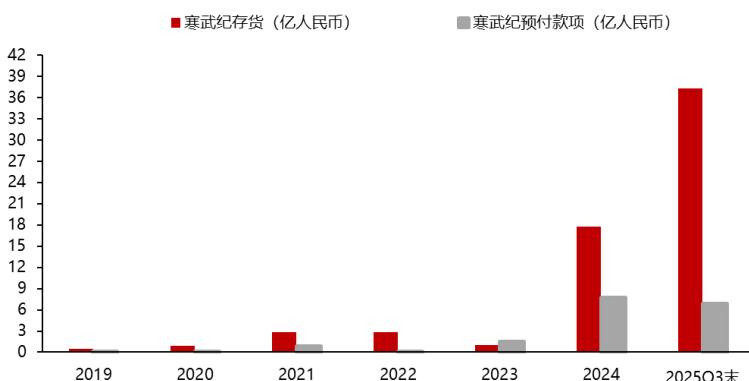
图99：寒武纪产品策略



资料来源：寒武纪公司官网，国联民生证券研究所

**在业绩层面，公司积极把握国产算力需求所带来的市场机遇。**25 年前三季度公司实现营收 46.07 亿元，同比大增 2386.4%；归母净利润为 16.1 亿元，同比大增 321.5%，实现自 2024Q4 以来连续 4 个季度单季度归母净利润为正。25 年 Q3 期末，寒武纪预付款项达 6.9 亿元，较 24 年末减少 0.9 亿元，存货达 37.3 亿元，同比+267.33%，未来良好放量支撑。细分领域方面，公司在大模型、互联网领域深化头部企业合作，人工智能芯片产品优势助力市场份额拓展，持续推动 AI 算力在银行、保险、基金等业务场景中的深度落地。

图100：2019-2025Q3 寒武纪存货与预付款项



资料来源：iFinD，国联民生证券研究所

**寒武纪持续拓展市场，积极助力人工智能应用落地。**2025 年公司的训练软件平台拓展了对 DeepSeek 系列、Qwen 系列、Hunyuan 系列模型的支持。推理软件平台此前已实现大规模专家并行优化等突破，近期进一步降低了通信延时并提升了通信计算并行效率，以 DeepSeek-R1-671B 大模型为代表的推理性能显著提升。

### 7.3.4 海光信息：双产品矩阵覆盖全场景，生态绑定头部企业

**技术壁垒构筑核心竞争力，双产品矩阵覆盖全场景需求。**海光是国内高端处理器研发设计领域的标杆企业，聚焦服务器、工作站等计算存储设备核心芯片，形成海光 CPU（通用处理器）与 DCU（协处理器）双轮驱动的产品格局。公司的 CPU 产品早年获得 AMD 在高端处理器的技术授权及相关技术支持，并基于 X86 架构研发。x86 的生态优势不可替代，通用处理器兼容主流 x86 操作系统、云计算平台及数据库，在金融、电信等对生态兼容性要求高的行业占据主导地位。而“海光三号”作为海光信息最主要产品之一，在参数上处于国产领先地位，性能表现优秀。

**在微体系结构层面，海光持续推进自主创新。**基于完整的 x86 指令集源码实现 C86 架构的国产化研发，持续强化产品竞争优势。同时，海光自研 C86 系列处理器也保持着每代至少 15%-30% 的性能提升。深算 DCU 以 GPGPU 架构为核心，算子覆盖度超 99%，兼容类 CUDA 环境。2025 年 DCU 市占率位列国产 AI 加速芯片头部，深算三号已经投入市场，受到客户认可。海光同时掌握高端 CPU 与 DCU 研发能力，“CPU+DCU”组合可提供一体化算力解决方案，在 AI 大模型训练、算力中心建设场景中，系统适配效率较“第三方 CPU+国产 GPU”方案大幅提升，双产品协同优势明显。

表 23：海光核心产品介绍

产品类型	处理器类型	指令集	主要产品	产品特征	典型应用场景
海光 CPU	通用处理器	兼容 x86 指令集	海光 3000 系列 海光 5000 系列 海光 7000 系列	内置多个核心处理器，继承通用的高性能外设接口，拥有完整的软硬件生态环境和完备的系统安全机制，适用于数据计算和事务处理等通用性应用	云计算，物联网，信息服务等
海光 DCU	协处理器	兼容“类 CUDA”环境	海光 8000 系列	内置大量运算核心，具有较强的并行计算能力和较高的能效比，适用于向量计算和矩阵计算等计算密集型应用	大数据处理、人工智能、商业计算等

资料来源：海光信息招股书，国联民生证券研究所

**业绩稳步增长，研发投入持续加码。**2025 年 Q3，海光实现营业收入 40.26 亿元，同比增长 69.60%；归母净利润为 7.60 亿元，同比增长 13.04%。营收高增受益于市场拓展深化及产业链合作下的客户端快速导入。同时，海光 Q3 单季度研发投入达 12.24 亿元，同比增长 53.83%，公司持续加码研发资源投入，深耕高端处理器核心技术，推动产品迭代升级提速，同时重点布局人工智能计算、科学运算等核心算力赛道，构筑产品护城河夯实长期竞争优势。截至 2025Q3 末，合同负债 28.00 亿元，较 2024 年同期大增 27.85 亿元，创同期历史新高，客户订单需求持续旺盛，后续业绩释放具备较强增长动能。

**生态绑定头部企业，国产化与 AI 双轮驱动。**GPU 推广方面，海光与国内主流大模型全面适配，支持了 AI 场景的多元化落地，公司产品在金融、能源、电信、互联网实现规模出货。在 2026 年，公司的市场份额有望得到进一步提升，逐步成长为国内的 AI 算力的核心供应商。面向互联网客户，海光与字节、腾讯、阿里、百度等大厂在技术联合研发、产品采购、生态共建等方面有深度合作，主要是产品研发和定制化服务为主，未来互联网行业是海光服务的重点行业，将持续加强合作。

### 7.3.5 沐曦：全栈自主可控，国产 GPU 规模化落地

**破局“卡脖子”，打造国产 GPU 领军企业。**沐曦是国内领先的高性能通用 GPU 企业，聚焦训推一体、通用计算及图形渲染三大产品线，构建全栈自主可控技术体系。公司技术团队具备深厚的国际一线 GPU 研发背景，核心产品已在金融、科研、视频处理等行业实现大规模落地应用。

表 24：沐曦核心产品介绍

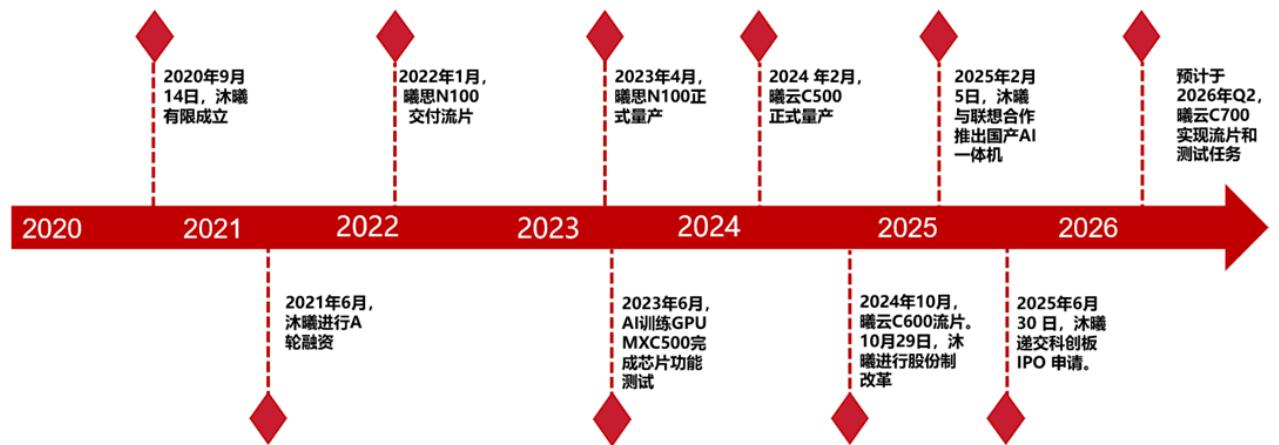
产品类型	型号	产品特征	应用场景
训推一体 GPU	曦云 C500 系列	曦云 C 系列产品拥有多精度混合算力,内置大量运算核心,具有较强的并行计算能力和较高的能效比,适用于向量计算和矩阵计算等计算密集型应用	云端智算(训推一体)、通用计算、AI for Science 等
	曦云 C600 系列		
智算推理 GPU	曦思 N100 系列	曦思 N100 产品系面向传统人工智能场景,内置性能强劲的视频处理器和运算核心	云端及边端推理、视频转码
	曦思 N260 系列	曦思 N 系列后续迭代产品系面向生成式人工智能场景,拥有多精度混合算力、大容量显存和较高的能效比	云端及边端图形处理
	曦思 N300 系列		
图形渲染 GPU	曦彩 G100 系列	G 系列产品系面向图形处理场景,内置性能强大的图形处理器	云端及边端图形处理

资料来源：沐曦集成电路招股说明书，国联民生证券研究所

**针对云端计算场景，沐曦研发出其主力产品曦云 C 系列。**沐曦于 2023 年推出了首款训推一体 GPU 芯片曦云 C500，并在此基础上陆续推出了曦云 C588；该系列基于国产供应链的产品曦云 C600 已回片并点亮。沐曦当前的主力产品曦云 C550 采用公司自研的 XCORE 1.0 架构及指令集，面向云端人工智能训练与推理、通用计算、AI for Science 等计算任务，拥有丰富的标量、矢量和张量计算单元，算力密度更高，支持多种混合精度计算。在研的曦云 C600 采用公司自研的 XCORE 1.5 架构及指令集，增加了 FP8 Tensor 及 Tensor 转置指令，采用 HBM3e 显存技术，支持卡间高速互连。

**公司推出曦思 N 系列实现传统 AI 与生成式 AI 任务的双重落地，产品定位清晰。**2022 年推出的曦思 N100，基于自研 XCORE 0.5 架构及指令集打造，支持 H.264/H.265 等多格式硬件解码及多种混合精度计算，搭配 16GB HBM2e 显存，可高效适配人脸识别、图像识别等传统 AI 任务。当前主力产品曦思 N260 性能再升级，采用自研 XCORE 1.0 架构及指令集，聚焦生成式 AI 下的云端推理场景，64GB HBM2e 大显存与新一代高速 I/O 接口形成硬件优势，支持多种混合精度计算及主流深度学习开发框架，有效适配一机二卡、四卡等灵活配置。产品整体具备强劲的大规模并行计算能力，在通用性、单卡性能及生态兼容性等维度表现优异，依托技术优势可更好满足 AI 任务需求。未来重点推进曦思 N300 的研发工作。曦思 N300 将基于国产供应链构建，采用自研 XCORE 1.5 架构及指令集，延续云端推理的核心定位，搭载更高性能的 HBM3 显存，进一步完善面向生成式 AI 的产品矩阵。

图101：沐曦 GPU 产品系列研发进程



资料来源：沐曦集成电路招股说明书，国联民生证券研究所

**业绩高速增长，规模化落地成效显著。**截至 2025 年前三季度，公司实现营业收入 12.36 亿元，毛利率保持在 55.76% 的较高水平，经营效率持续改善。公司 GPU 产品性能对标国际先进水平，已成功实现千卡集群的商用部署，并加速向万卡规模扩展。随着生成式 AI、大模型推理等新兴场景的快速发展，叠加国家对“算力国产化”的持续支持，沐曦有望在未来三年持续扩大市场份额，成长为国产 GPU 产业链中的关键支点与核心突破力量。

**国产 GPU 竞速新时代，沐曦构建通用算力新格局。**已量产的训推一体 GPU 芯片具备强劲的大规模并行计算能力，在通用性、单卡与集群性能、系统稳定性及生态兼容性等维度达到国内领先水平，综合竞争力突出。依托自主研发的领先 MetaX Link 互连技术，该芯片在大规模集群中可实现优异的线性扩展效果，能精准匹配大模型迭代下激增的集群算力需求，为智算集群实际运营释放强劲性能。沐曦芯片在传统 AI 领域已覆盖智慧交通、教育等行业应用场景，生成式 AI 领域则通过曦思 N260 切入大模型一体机等核心场景，与整机厂商合作潜力显著。截至目前，沐曦 GPU 产品累计销量已突破 25,000 颗，成功在多个国家级 AI 算力平台、运营商智算平台及商业化智算中心实现规模化落地。

### 7.3.6 摩尔线程：MUSA 架构筑基，全功能 GPU 拓展多元场景

**攻坚 GPU 核心技术壁垒，构建自主可控技术和产品体系。**摩尔线程在国内 GPU 领域处于领先地位，基于自主研发的 MUSA 架构，公司率先实现了单芯片架构同时支持 AI 计算加速、图形渲染、物理仿真和科学计算、超高清视频编解码的技术突破，有力推动了我国 GPU 产业的自主可控进程。

**MUSA 架构优势显著，生态兼容性良好。**MUSA 架构具有良好的灵活性与可扩展性，具备与由英伟达主导的国际主流 GPU 生态的兼容性，使得开发者能够以较低成本充分利用目前国际主流生态下的代码资源。基于 MUSA 架构开发的应用程序不仅具有广泛的可移植性，还能够同时在云端及边缘的众多计算平台上运行，其应用领域广泛，涵盖 AI、图形处理、科学计算等多个重要方向。

表 25：摩尔线程 GPU 架构芯片演进情况

芯片架构迭代	升级重点	主要升级情况
苏堤	多引擎统一系统架构	苏堤架构的核心技术形成于公司成立初期，通过首次实现基于多引擎统一系统架构的全功能 GPU，显著提升芯片资源利用率。FP32 精度通用算力达到 5.2TFLOPS，基于 LPDDR4 访存技术实现最高 133GB/s 的带宽，并支持初步 DirectX11 技术规范，完善支持 OpenGL4.6、OpenCL3.0 等技术规范
苏堤→春晓	云计算虚拟化能力和图形应用生态突破	针对云计算场景，春晓架构强化了 GPU 虚拟化能力，支持单芯片 32 路用户并发调用 GPU 资源。在图形渲染兼容性方面，春晓架构完善了对 DirectX11 技术规范的支持，并初步支持 DirectX12 技术规范。在芯片设计方面，基于先进制程和物理设计优化，核心频率由苏堤的 1.4GHz 提升至 1.9GHz，FP32 精度算力从 5.2TFLOPS 提升至 15.5TFLOPS，INT8 精度算力达 62TOPS，访存技术升级至 GDDR6，支持 448GB/s 访存带宽
春晓→曲院	AI 算力与片间互联技术	在 AI 算力方面，曲院架构提升了张量计算核心功能和性能，支持 TF32、BF16、FP16、INT8 等多种计算精度，核心数从春晓的 32 个增至 128 个，单卡 FP16 算力从 15.5TFLOPS 提升至 102TFLOPS，显著增强了 AI 训练与推理效率。在片间互联技术方面，曲院架构研发了 MT-Link 互联技术，单芯片互联带宽达 240GB/S，为千卡级集群(KUAE1)奠定了基础，支持大规模分布式计算
曲院→平湖	AI 算力与计算效率突破	通过全新设计的张量计算引擎，平湖架构实现了片间互联带宽提升至 3 倍，计算和访存效率提升至 4 倍，张量核心数量增至某数量，新增 FP8 精度支持，FP32/TF32/BF16/FP16/INT8 等精度算力较曲院显著提升。新增异步通信引擎，片间互联带宽增至 800GB/S，支持万卡级集群(KUAE2)高速通信满足超大规模模型训练需求。同时，平湖架构在大模型分布式训练中实现了混合精度计算和异步通信的端到端效率提升，效率提升至 80%。
平湖→花港	AI 算力与计算效率突破	最新发布的全功能 GPU 架构“花港”，凭借 FP4 到 FP64 的全精度计算支持，实现密度提升 50%、效能提升 10 倍的技术突破。基于这一架构，摩尔线程将在未来推出两款核心产品：高性能 AI 训推一体“华山”芯片，以及专注高性能图形渲染的“庐山”芯片

资料来源：摩尔线程招股书，摩尔线程公众号，国联民生证券研究所

**新品性能亮眼，应用布局广泛。**摩尔线程最新发布的全功能 GPU 架构“花港”，凭借 FP4 到 FP64 的全精度计算支持，实现密度提升 50%、效能提升 10 倍的技术突破。基于这一架构，摩尔线程将在未来推出两款核心产品：高性能 AI 训推一体“华山”芯片，以及专注高性能图形渲染的“庐山”芯片。联合硅基流动完成的测试数据显示，在 DeepSeekR1671B 全量模型上，MTTS5000 单卡 Prefill 吞吐突破 4000tokens/s，Decode 吞吐突破 1000tokens/s，一举树立国产推理性能新标杆。面向超大规模智算需求，摩尔线程前瞻布局超节点架构，发布 MTTC256 超节点架构规划，以高密硬件设计为基础，打造下一代智算中心的极致性能支撑。与此同时，搭载智能 SoC 芯片“长江”的 AI 算力本 MTTAIBOOK 正式亮相，为“摩尔学院” 20 万开发者与学习者提供端侧算力赋能。

从硬件级光线追踪加速、自研 AI 生成式渲染技术，到具身智能、科学智能（AI4S）、AI for 6G 等前沿领域的深度布局，摩尔线程的技术实践充分展现了全功能 GPU 技术路线的应用广度与面向未来的可扩展潜力。

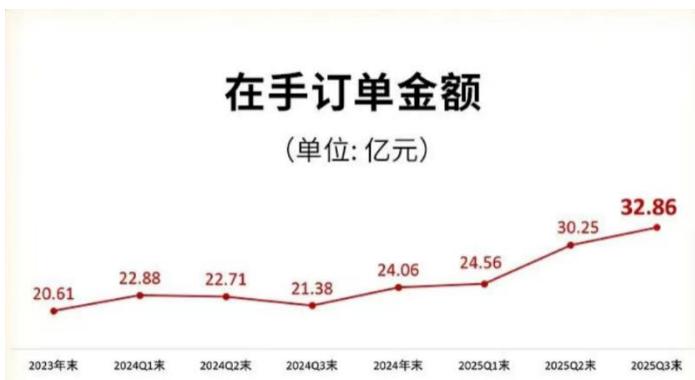
**业绩增长态势良好，亏损规模持续收窄。**摩尔线程营业收入整体呈现增长趋势，主要得益于产品种类及应用领域的丰富、加深客户合作、持续推进产品研发及迭代升级等因素。2025 年前三季度，摩尔线程实现营业收入 7.85 亿元，同比大幅增长 181.99%，主要原因受人工智能蓬勃发展及国产化进程加速的背景影响，公司业绩呈现上升趋势，营业收入快速增加，亏损规模有所收窄。

### 7.3.7 芯原：IP 布局深厚，一站式定制服务承接 AI 算力订单

**IP 资源积淀深厚，芯片定制能力突出。**芯原股份是国内领先的半导体 IP 供应商，拥有包括 GPU IP、NPU IP 在内的多款处理器 IP，具备为客户提供一站式芯片定制服务的能力。公司芯片设计能力覆盖 14nm/10nm/7nm/5nm FinFET 和 28nm/22nm FD-SOI。2023 年，蓝洋智能发布与芯原股份合作打造的基于 Chiplet 架构的高性能 AI 芯片，其中 CC8400 在提供强大算力的同时，还可优化面积和功耗；VIP9400 支持 Transformer 模型，能够为数据中心和汽车应用提供强大的 AI 算力；VC8000D 具有高吞吐量、多格式等特性，可用于视频内容分析。目前，内置芯原 GPU IP 的芯片在全球范围内出货近 20 亿颗，芯原 NPU IP 已被 82 家客户用于其 142 款人工智能芯片中，在全球范围内出货超过 1 亿颗。

**订单转化推动营收增长，盈利能力持续改善。**得益于芯原近两年在手订单持续保持高位，随着订单的不断转化，公司研发资源陆续投入客户项目，芯原 2025 年 Q3 营业收入延续了 Q2 环比大幅增长的趋势，2025 年 Q3 实现营业收入 12.8 亿元，单季度收入创公司历史新高，环比大幅增长 119.3%，同比增长 78.4%；公司 2025 年 Q3 盈利能力大幅提升，单季度亏损同比、环比均实现大幅收窄。2025 年前三季度，为公司贡献营业收入的量产出货芯片共 112 款，公司实现量产业务收入 10.16 亿元，同比增长 76.93%，前三季度收入已超 2024 年全年收入水平。公司量产业务快速增长，主要受益于数据处理、端侧 AI 领域的需求扩张。

图102：芯原在手订单金额



资料来源：芯原公告，国联民生证券研究所

**受益于 AI 浪潮，公司订单饱满。**2025 年 Q3 芯原新签订单 15.93 亿元，同比大幅增长 145.80%，其中 AI 算力相关的订单占比约 65%。公司在手订单已连续八个季度保持高位，截至 2025 年第三季度末在手订单金额为 32.86 亿元，持续创造历史新高。公司 2025 年 Q3 末在手订单中来自系统厂商、大型互联网公司、云服务提供商和车企等客户群体的订单占比为 83.52%。公司 2025 年 Q3 末在手订单中，一站式芯片定制业务在手订单占比近 90%，且预计一年内转化的比例约为 80%，为公司未来营业收入增长提供了有力的保障。

### 7.3.8 灿芯：高速接口 IP 突破，聚焦高性能场景定制

**灿芯面向多领域客户提供一站式芯片定制服务，快速满足客户差异化需求。**基于 28HKC+ 工艺平台的 DDR、SerDes、PCIe、MIPI、USB 等高速接口 IP 已完成验证并实现量产交付，能够为数据中心 AI 加速芯片、车载 SoC 等高性能场景需求提供支持。在 28HKD 工艺平台上，全线 DDR、SerDes、PCIe、MIPI、USB 等高速接口 IP 完成客户小批量验证，新增的 PSRAM 和 EMMC IP 产品线进一步补充了公司在低功耗存储接口领域的布局。公司基于 22nm 工艺平台的 DDR5 IP 完成架构验证，DDR5 IP 核技术是支撑新一代高性能计算芯片的关键模块，主要涵盖控制器、PHY 物理层及完整子系统解决方案，通过高速率、低功耗及创新架构设计，已深度融入 AI 计算、数据中心、移动终端及工业控制等高性能领域。

**多元 ASIC 定制芯片，适配不同应用场景。**基于 RISC-V 内核的边缘计算 ASIC 定制芯片，采用 32 位、12 级流水线深度 RISC-V 处理器，集成 CNN、RNN 等深度学习网络算法；基于 RISC-V 内核的手持终端基带处理 ASIC 定制芯片采用 32 位 RISC-V 处理器，集成灿芯半导体自研 DDR、USB、USIM、CIPHER 等 IP；基于 RISC-V 内核的通信芯片采用 32 位 RISC-V 处理器，在前一代产品基础之上，优化了存储方案，使用 PSRAM 替代 DDR，集成 RISCV/DSP 等控制、计算模块，使得成本降低 50% 以上；高精度导航基带 ASIC 定制芯片采用灿芯半导体 harden CPU core，主频高达 648M，满足 SoC 系统的调度控制使用，并采用自研 Aurora Serdes IP。

**业绩呈现改善态势，在手订单提供支撑。**2025年前三季度，公司实现营业收入4.7亿元，同比下降45.7%，其中，芯片设计业务实现营业收入2.4亿元，同比增长24.2%，芯片量产业务实现营业收入2.3亿元，同比下降65.8%，主要系2024年同期对公司量产业务贡献较大的部分客户因其需求变动减少对公司采购，同时公司新增项目收入尚不足以弥补前述收入变动影响所致。2025年Q3，公司实现营业收入1.9亿元，环比增长30.3%。其中，芯片设计业务实现营业收入1.0亿元，环比增长49.4%，芯片量产业务实现营业收入0.9亿元，环比增长14.2%，呈现改善态势。

**聚焦高潜力领域，拓展市场强化竞争力。**截至2025年9月30日，公司在手订单金额为8.7亿元，其中芯片设计业务在手订单2.9亿元，芯片量产业务在手订单5.8亿元。一方面，灿芯加强市场开拓力度，重点布局汽车电子、端侧AI、AI+IoT等高潜力领域，加速技术研发成果的市场化应用，增强公司核心竞争力；另一方面，灿芯积极拓展销售与服务网络的覆盖度，提升销售团队整体专业素质，优化公司营销模式。

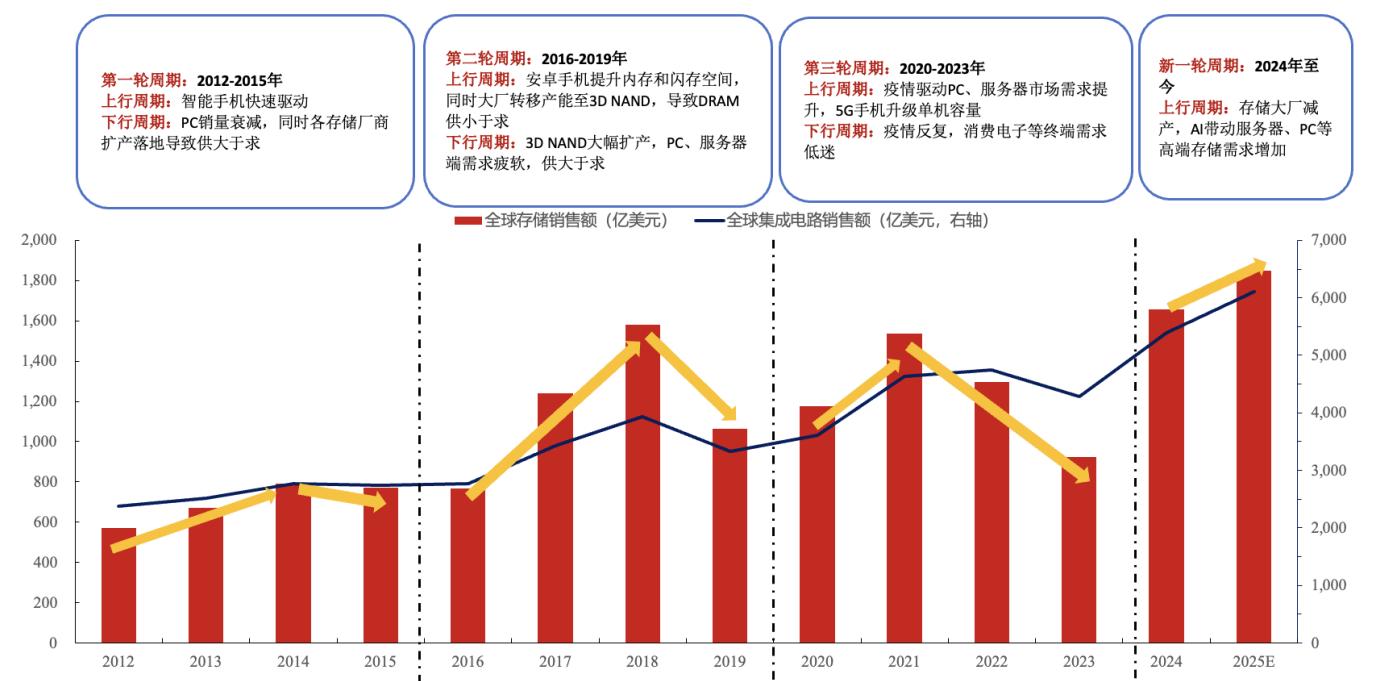
## 8 AI 驱动存储迎超级周期，设备受益原厂扩产

### 8.1 AI 驱动存储行业迎快速增长

#### 8.1.1 存储行业或迎超级周期

存储行业发展呈现明显周期性特征，可划分为多轮周期。第一轮周期（2012-2015年），上行由智能手机增长驱动存储需求，下行因PC销量衰减、存储厂商扩产引发供过于求；第二轮周期（2016-2019年），上行受安卓手机内存/闪存容量提升、厂商转产3D NAND致DRAM供应紧张推动，下行因3D NAND新增产能释放，供大于求；第三轮周期（2020-2023年），上行受益于疫情下PC、服务器需求提升及5G手机单机存储容量升级，下行因疫情反复、消费电子终端需求低迷；**2024年至今进入新一轮周期，上行由存储大厂减产优化供给、人工智能(AI)带动服务器/PC高端存储需求增长驱动。**

图103：存储周期复盘



资料来源：全球半导体贸易统计组织，wind，国联民生证券研究所

**本轮存储涨价的核心原因是供需关系反转和结构性需求快速提升。**一方面头部厂商减产低端产能，另一方面AI算力需求直接推动服务器用存储需求，供需反转下价格出现较大幅度增长。根据TrendForce数据，2025Q4存储价格环比大幅上涨，其中传统DRAM价格环比涨幅达到45%-50%，NAND价格环比涨幅达到25-30%。**展望2026年，TrendForce认为DRAM和NAND产品价格均有望实现较大涨幅，2026Q1传统DRAM价格环比涨幅有望达到15-20%，NAND价格环比涨幅有望达到20-25%。**

表 26：DRAM/NAND 未来一年价格预测

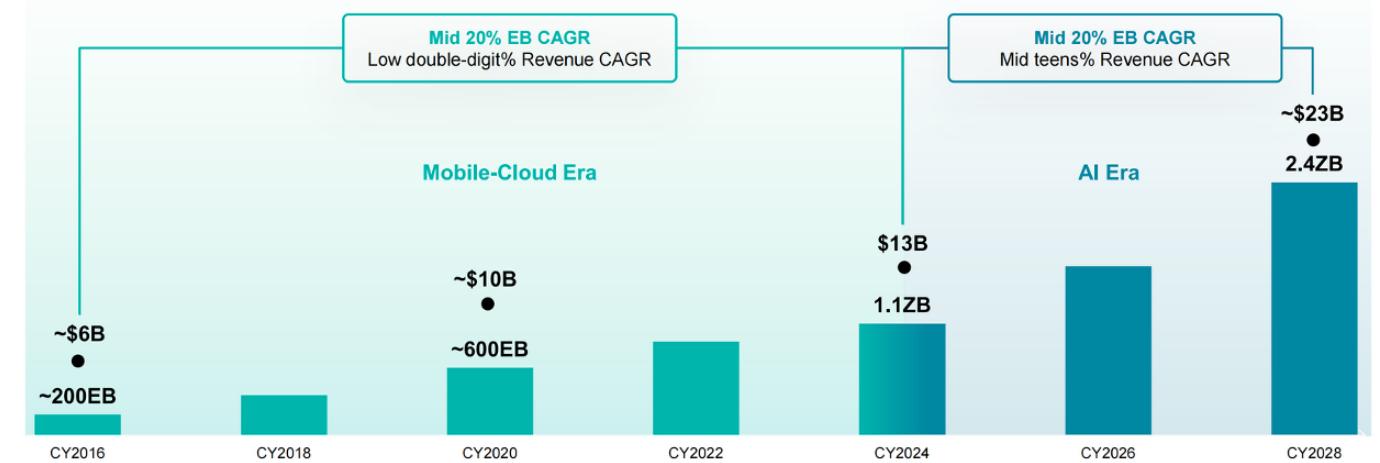
QoQ%	1Q26F	2Q26F	3Q26F	4Q26F
Total DRAM	Conventional DRAM: up 15-20%	Conventional DRAM: up 5-10%	Conventional DRAM: up 3-8%	Conventional DRAM: up 0-5%
Total NAND	up 20-25%	up 13-18%	up 5-10%	up 0-5%

资料来源：TrendForce，国联民生证券研究所

### 8.1.2 AI 驱动存储需求快速增长

随着技术从“移动云时代”迈向“AI 时代”，数据量持续扩张：2016 年数据量约 200EB，2020 年（移动云时代）增至约 600EB；进入 AI 时代后，2024 年数据量达 1.1ZB (1ZB=1000EB)，2028 年预计升至 2.4ZB。整体来看，**数据量实现从 EB 量级到 ZB 量级的跨越，且 AI 时代增长明显加速，直观体现数据量快速增长对存储需求的推动作用。**

图104：数据中心存储需求市场规模



资料来源：希捷科技公告，国联民生证券研究所

随着大模型训练与推理对数据访问需求的增长，大量曾被视为“冷数据”的资源正被重新激活。这些数据因频繁参与模型迭代与实时推理，逐渐转变为“温数据”，甚至因持续调用而成为“热数据”。根据华为发布的《智能世界 2035》，预计 2035 年温数据的占比有望超过 70%，传统的数据三层结构将逐渐演变为“热-温-冷”两层结构，比例趋于 3:7。这一转变不仅显著提升数据利用效率，更意味着企业和社会能够从历史数据中挖掘出前所未有的价值，推动数据资源从“被动存储”走向“主动赋能”。**冷数据主要用 HDD 存储，温数据主要用 HDD 和 SSD 存储，热数据主要用 SSD 和 DRAM 存储。随着冷数据转温，SSD 应用空间有望逐步扩大。**

图105：不同温度数据与数据介质对应关系



资料来源：腾讯云，西部数据，国联民生证券研究所

**AI 服务器在内存、显存、硬盘上均有更高的性能要求。**内存方面，更高传输速率的 DDR5 加速渗透；显存方面，高带宽的 HBM 在高端 AI 加速卡中已占据主导地位，并持续提升渗透率；硬盘方面，AI 服务器中 SSD 硬盘成为首选。根据 TrendForce 数据，普通服务器 Server DRAM 配置约 500-600GB，而 AI 服务器单条模组多为 64-128GB，平均容量达 1.2-1.7TB；Enterprise SSD 方面，AI 服务器更侧重 PCIe 5.0 高速接口以适配高速运算，而非必要扩大容量。**随着未来 AI 模型持续复杂化，Server DRAM、SSD 及 HBM 的需求将进一步同步增长。**

表 27：一般服务器与 AI 服务器平均容量差异

容量	服务器	AI 服务器	未来的 AI 服务器
DRAM	500-600GB	1.2-1.7TB	2.2-2.7TB
SSD	4.1TB	4.1TB	8TB
HBM	-	320-640GB	512-1024GB

资料来源：TrendForce，国联民生证券研究所

在服务器端，随着人工智能和大数据分析等应用快速发展，处理器内核数量日益增多，对内存带宽的需求急剧增长，JEDEC 制定了新型高带宽内存模组多路复用双列直插内存模组 MRDIMM (Multiplexed Rank DIMM) 的相关技术标准。MRCD/MDB 芯片是服务器高带宽内存模组 MRDIMM 的核心逻辑器件，每根 MRDIMM 模组均需要搭配 1 颗 MRCD、10 颗 MDB。MRDIMM 通过 MDB 芯片可以同时访问两个 DRAM 内存阵列（而传统 RDIMM 只能访问一个阵列），从而在标准速率下实现双倍带宽。该产品主要应用于云计算、AI 等对内存带宽要求较高的应用领域。

表 28：内存互连芯片在不同类型的内存模组中的应用及配比关系

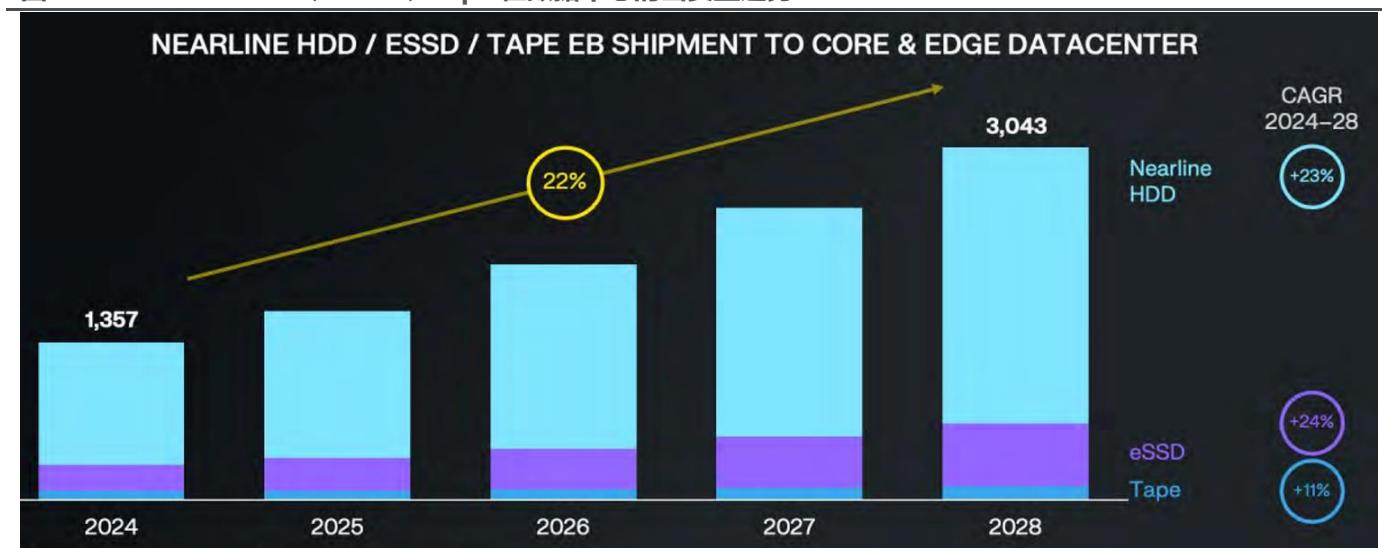
内存模组	RCD	DB	SPD	TS	PMIC	MRCD	MDB	CKD	合计 (颗)
RDIMM (DDR4)	1		1						2
LRDIMM (DDR4)	1	9	1						11
RDIMM (DDR5)	1		1	2					4
LRDIMM (DDR5)	1	10	1	2	1				15
MRDIMM (DDR5)			1	2	1	1	10		15

资料来源：澜起科技公告，国联民生证券研究所

### 8.1.3 SSD 有望加速替代 HDD

根据西部数据公告，2024 年近线硬盘（Nearline HDD）、企业级固态硬盘（eSSD）及磁带（Tape）的合计出货量为 1357 EB。随着数据中心存储需求持续攀升，出货量呈稳步增长态势，预计 2028 年将达到 3043EB，整体增长幅度显著。从 2024-2028 年出货量 CAGR 来看，eSSD 以 24% 的增速领跑，Nearline HDD 为 23%，Tape 为 11%。这一趋势既反映出数据规模扩张对存储能力的迫切需求，也体现出不同存储介质在性能、成本等维度的差异化发展节奏，其中 eSSD 凭借更快的增长速率，凸显出在数据中心存储架构中快速渗透的特点。

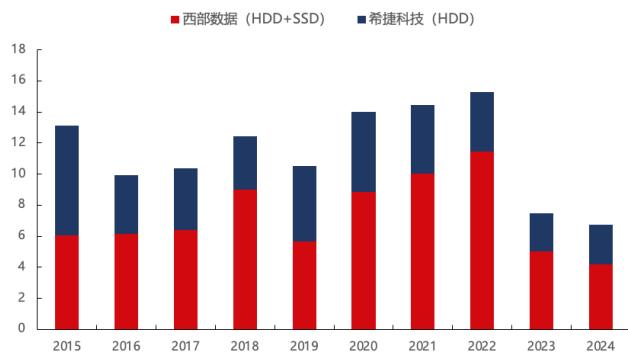
图106：Nearline HDD、eSSD、Tape 在数据中心的出货量趋势



资料来源：搜狐，西部数据，国联民生证券研究所

**HDD 供给受限：HDD 交期延长，扩产周期长且厂商无扩产意愿。** TrendForce 表示，由于全球主要 HDD 制造商近年来规划扩大产能，无法及时满足 AI 刺激的突发性、巨量储存需求。**目前 NL HDD 交期已从原本的数周，延长为 52 周以上，加速扩大 CSP 的存储缺口。** 北美 CSP 早已规划于温数据应用扩大采用 SSD，但因为这波 HDD 缺口严峻，CSP 甚至开始考虑于冷数据采用 SSD，然而，要迈向大规模部署须先解决成本和供应链的双重挑战。

图107：两大龙头 HDD 厂商资本开支变化（亿美元）



资料来源：wind, 国联民生证券研究所

图108：NL HDD 交期延长，加速 SSD 替代

Nearline HDDs vs. QLC SSDs

Product	Lead Time	ASP (US\$/per GB)	Max Capacity	Performance	Energy Efficiency
Nearline HDD	52 weeks	0.015	32 TB	Weak	Inferior
QLC SSD	8 weeks	0.05-0.06	122 TB	Strong	Superior

Source: TrendForce, Sept. 2025

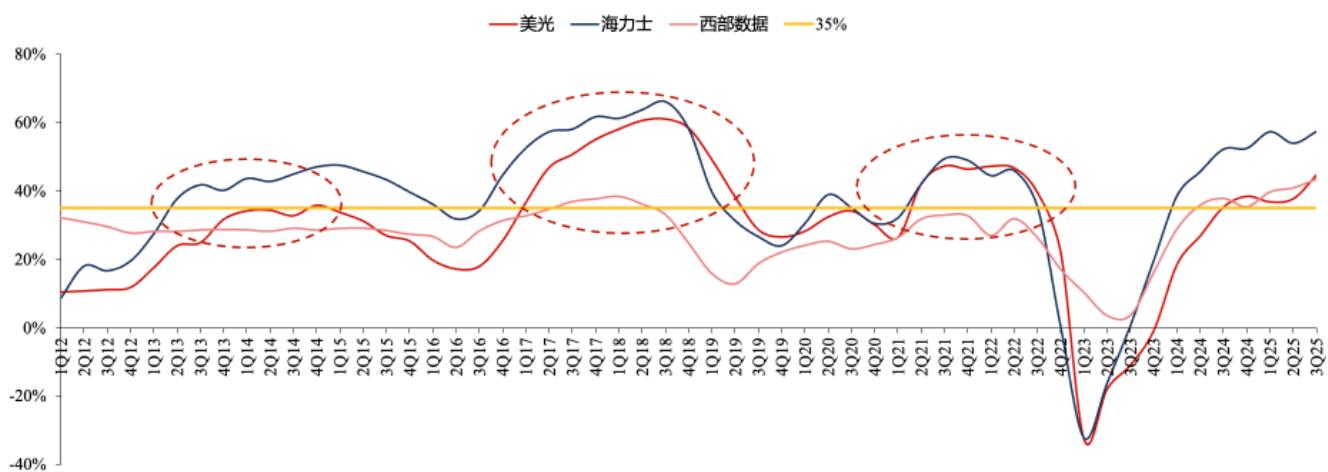
资料来源：Trendforce, 国联民生证券研究所

## 8.2 半导体设备受益存储上行周期

### 8.2.1 上游设备有望受益存储原厂扩产

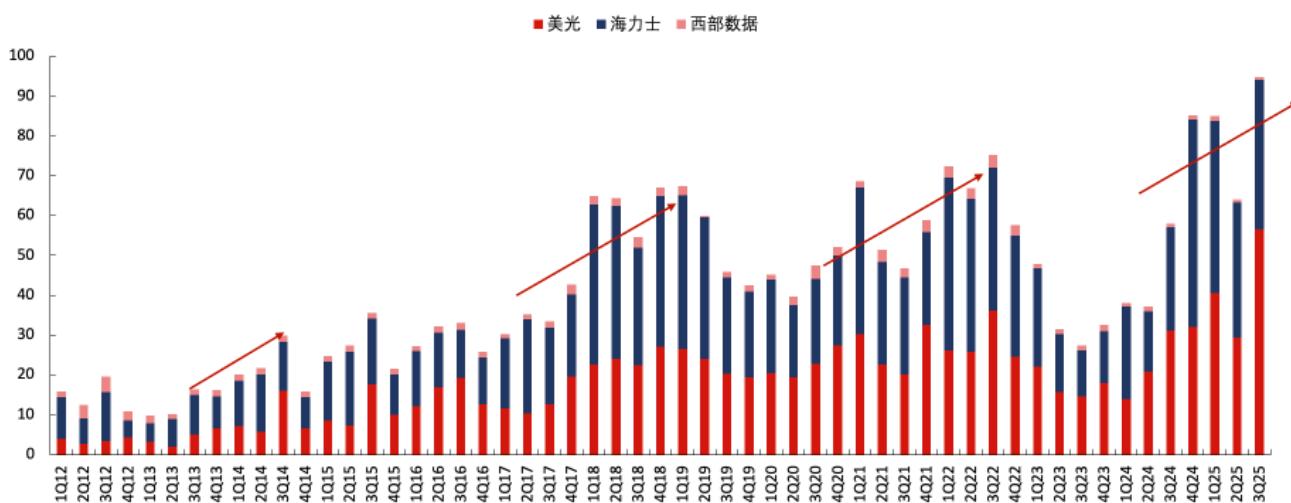
根据美光、海力士、西部数据 2012 年至今的季度财务数据来看，我们发现原厂毛利率达到 35%以上时，对应季度的资本开支增加的概率也会放大。**2024Q3 至今，原厂毛利率逐步提升至 35%以上，资本开支也相应增加。**受益 AI 需求的持续拉动，存储涨价的持续，我们预计原厂毛利率有望持续维持在 35%以上，同时行业面临供需偏紧的状态，原厂或有望提高资本开支以满足持续增长的存储需求。

图109：存储原厂毛利率季度变化



资料来源：wind, 国联民生证券研究所

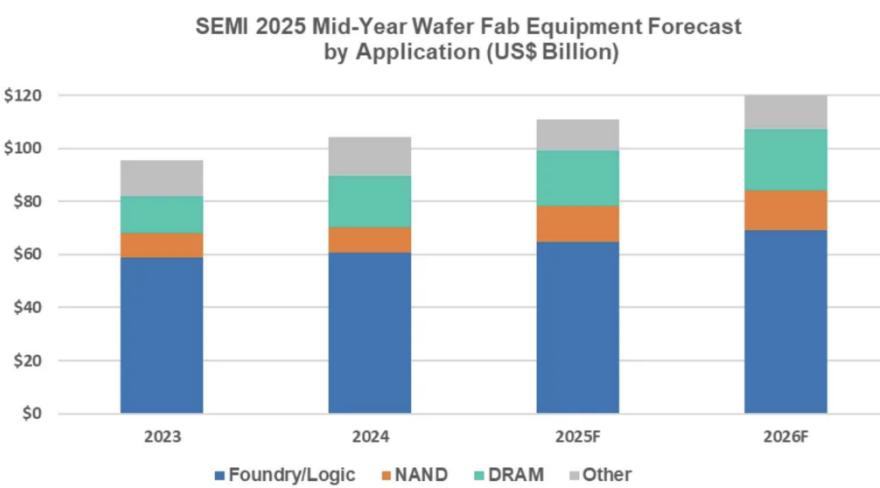
图110：存储原厂资本支出变化（单位：亿美元）



资料来源：wind，国联民生证券研究所

根据 SEMI 数据，预计 2025 年内存相关资本支出将有所增长，并在 2026 年持续增长；NAND 设备销售额将从 2023 年的急剧萎缩中持续复苏。继 2024 年小幅增长 4.1% 之后，预计 2025 年全球 NAND 设备市场规模将增长 42.5%，达到 137 亿美元，2026 年将增长 9.7%，达到 150 亿美元，这主要得益于 3D NAND 堆叠技术的进步和产能扩张。DRAM 设备销售额在 2024 年飙升 40.2%，达到 195 亿美元，预计 2025 年和 2026 年将分别增长 6.4% 和 12.1%，以支持对用于 AI 部署的 HBM 的投资。

图111：全球半导体设备市场规模变化



资料来源：wind，国联民生证券研究所

国内两存有望加速扩产。AI 需求快速提升，数据存储需求也呈现快速增长。存储芯片两大类：一类是 DRAM，其中 HBM 专门用于 AI 的训练和推理，第二类

是 NAND 闪存，负责存放海量数据和模型参数。根据恒运昌公告，长鑫存储作为国领军企业，产能预计从 2024 年底的 20 万片/月增长至 2025 年底的 30 万片/月，同比增长近 50%，2025 年底产能占全球产能比例约为 15.6%。长江存储作为国内 NAND 领军企业，正推进三期扩产计划，三期达产后总产能将达 30 万片/月，目标占据全球 15% 的 NAND 市场份额。

**表 29：长江存储与长鑫存储扩产相关信息**

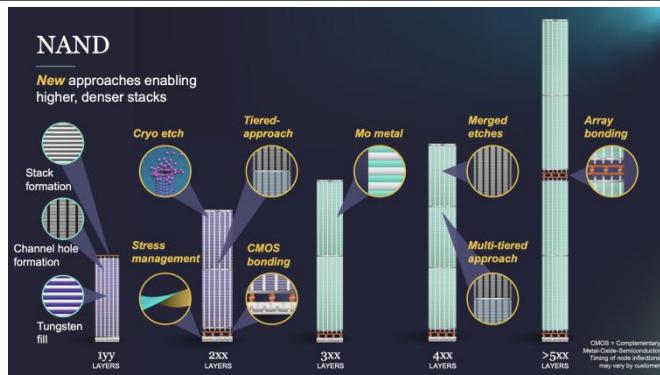
公司	扩产相关信息
长江存储	长江存储三期注册成立，注册资本 207.2 亿元。工商信息显示，长存三期(武汉)集成电路有限责任公司(长存三期)于 9 月 5 日成立，注册资本 207.2 亿元。股东信息显示，长存三期由长江存储持股 50.19%、湖北国资旗下企业湖北长晟三期投资发展有限责任公司持股 49.81%。根据 Trendforce 数据，2025Q1 全球 NAND 市场中，长江存储全球市占率为 8.1%；《电子时报》报道称，长江存储计划到 2026 年底产能在全球占比有望达到 15%。
长鑫存储	据市场分析机构 Counterpoint 预测，长鑫存储 2025 年将在 2024 年大幅增产的基础上产能进一步同比增长近 50%。第三方机构 TrendForce 预测，到 2025 年底，其月产能有望达到 30 万片，届时将占据全球 DRAM 总产能的约 15%。

资料来源：恒运昌公告，国联民生证券研究所

## 8.2.2 刻蚀/沉积设备受益存储架构创新

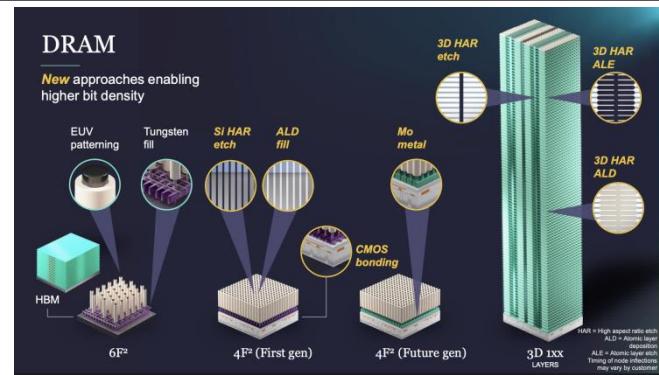
**在存储工艺三维化演进中，刻蚀与沉积设备是 DRAM 和 NAND 密度突破的核心。**DRAM 从 HBM 的 6F<sup>2</sup>迈向 3D 1xx 层，Si HAR 刻蚀、3D HAR 刻蚀/ALE 实现纳米级图形化，ALD 填充保障高深宽比孔薄膜均匀；NAND 从 1yy 层跃至 >5xx 层，深冷刻蚀、分层设计依赖刻蚀设备，ALD/CVD 完成钨填充与介质/金属层原子级构建。二者中，刻蚀是“三维结构雕刻刀”，沉积是“功能层黏合剂”，其性能迭代（刻蚀深宽比破百级、沉积精度达原子级），直接推动存储密度与堆叠层数持续突破，是存储技术三维集成的双引擎。

**图112：NAND 技术发展趋势**



资料来源：LAM 公告，国联民生证券研究所

**图113：DRAM 技术发展趋势**

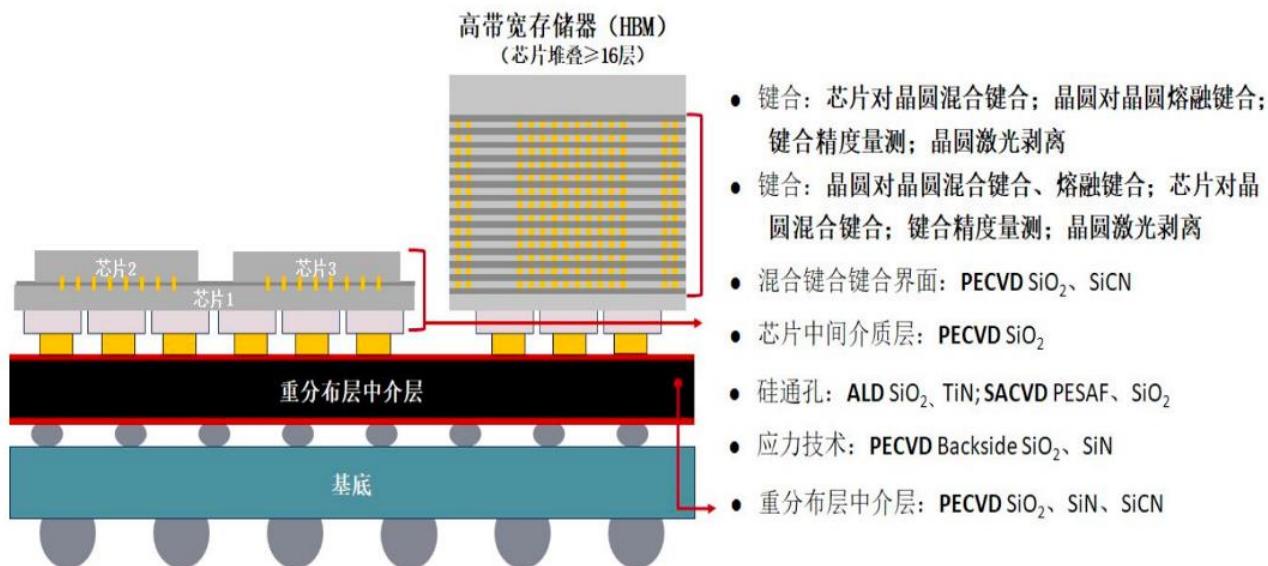


资料来源：LAM 公告，国联民生证券研究所

**键合设备是 3D 集成技术的核心设备。**随着“后摩尔时代”的来临，芯片制程持续缩小并接近物理极限，单纯依赖平面工艺极限已无法实现性能迭代，技术路径

逐步转向新的架构设计及芯片堆叠方式，三维集成技术则是这一技术创新和发展趋势的关键驱动力，而先进键合设备凭借其突破性技术优势成为三维集成技术领域的核心设备。

图114：键合工艺



资料来源：拓荆科技公告，国联民生证券研究所

存储芯片朝着高密度化方向发展，对于刻蚀、薄膜、键合工艺的性能需求越来越高，刻蚀设备、薄膜设备、键合设备的重要性也在逐渐凸显。此外，全球存储景气周期来临，存储原厂扩产意愿加强，**建议关注存储原厂业务占比高的企业**。

## 9 AI 赋能终端，产业范式重构

**AI 大模型跨越式发展，开启终端新一轮产业变革。**一方面，大语言模型 LLM 取得突破性进展，GPT-5、Claude 3、Gemini 3 等大模型实现接近人水平的复杂推理和多轮对话，同时多模态的 AI 融合也逐步成为发展趋势；另一方面，端侧芯片算力的突破为终端产品 AI 化提供了关键支撑，以全新旗舰手机芯片骁龙 8 Gen5 为例，其升级核心架构后，AI 性能相较前代提升 46%、SoC 功耗降低 13%，依托于从算力、内存到感知的全面升级，其支持智能体 AI 助手和多模态 AI 激活，结合高通传感器中枢，能高效处理复杂 AI 任务，并兼容主流生成式 AI 模型。

图115：AI 跨越式发展+硬件成熟，开启终端新一轮产业变革



资料来源：手机中国、中国电子报等，国联民生证券研究所整理

**AI 功能在 AI 手机、AI PC、智能可穿戴等五大智能终端领域均呈现强劲增长势头**，预计未来各类终端硬件在 AI 大模型的改造下，有望迎来智能化重估值机遇。其中以 AI 手机为例，电子发烧友预计 2025 年全球智能手机中 AI 功能渗透率有望达 34%，且 AI 功能正持续下沉至中低端手机市场，后续市场份额将稳步提升。

表 30：2025 年 AI 终端渗透率及驱动因素

品类	2025 年全球 AI 功能渗透率	关键驱动因素	发展情况
AI 手机	约占 34%	端侧模型逐步压缩，次旗舰级的 AI SoC 下沉到中端手机价位	旗舰机先行，千元档机型快速跟进
AI PC	约占 38%	Windows 11 AI+PC 认证，NPU≥40TOPS 的 PC 芯片逐步普及	预计 2028 年 AI PC 有望达到 79% 的占比
智能可穿戴设备	25H1 出货占比 15%	多模态交互、健康 AI 算法及轻量化传感器驱动	2026-2027 年 AI/AR 类眼镜带动智能化发展
智能家居终端	计入存量智能音响升级，约占 65%	低功耗唤醒，云端+本地混合推理	产品以存量升级为主，新增产品增速放缓
车载座舱	> 40%，含 L2+/L3 级车型	座舱大模型逐步上车，车规 AI 芯片普及	中国的 AI 车载渗透率领先全球

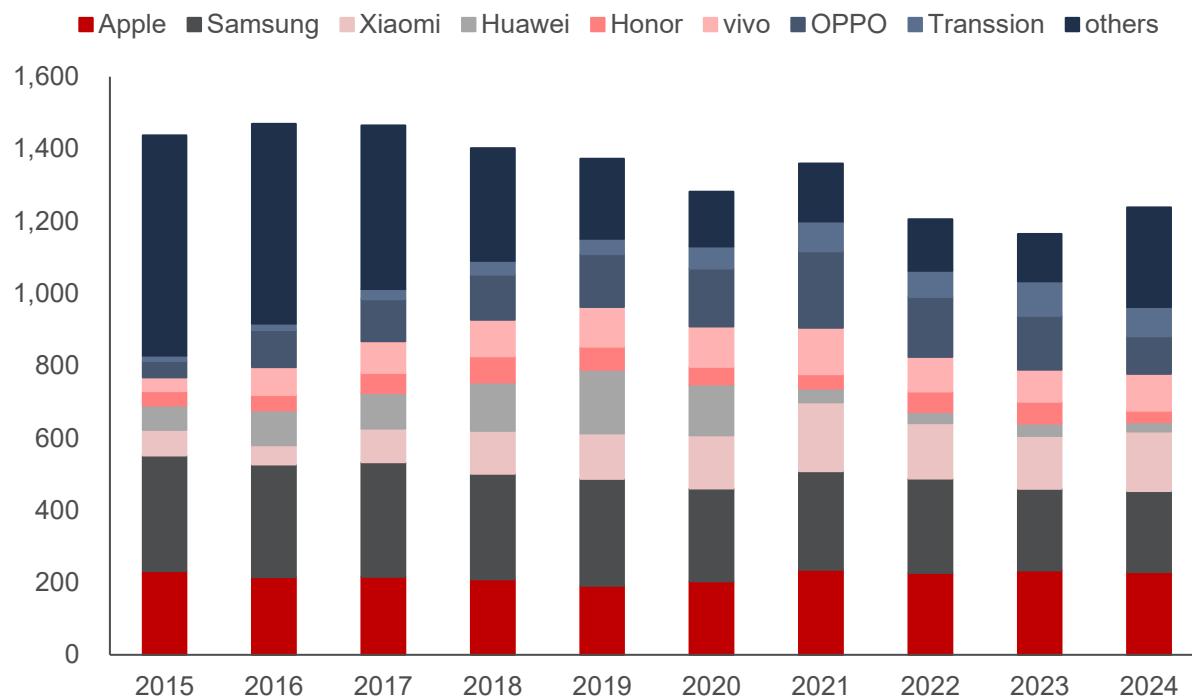
资料来源：电子发烧友，国联民生证券研究所

## 9.1 AI 手机：存量市场内结构性硬件+AI 功能创新

过去几年，智能手机新品硬件创新乏力，市场增速放缓，但步入 2024 年全球，消费电子市场回暖，叠加 AI 功能带来的换机需求，智能手机销量实现反弹。根据 IDC 预测，预计 2025 年全球智能手机出货量将同比增长 1.0%，达到 12.4 亿部，其中主要受 2025 年苹果手机出货量增长 3.9% 的推动。

尽管如此，智能手机的硬件供应链及 AI 功能的创新依旧不止，下文我们将就核心创新点进行展开介绍。

图116：2015-2024 年全球智能手机出货量年度数据（百万部）



资料来源：IDC，国联民生证券研究所整理

**硬件方面，光学仍是目前主流的创新方向之一。**具体来看，玻塑混合镜头在提升进光量、降低镜头高度方面具有优势；潜望式下沉是目前重要创新之一，潜望式镜头通过微棱镜结构实现光学变焦，同时支持远景和微距拍摄；其他光学创新还包括可变光圈、超光谱摄像头、外挂镜头等方向；CIS 方面，超大底（4/3 英寸）+ 大像素是当前的主流升级趋势；TOF 系统方面，可用于生物识别和手势识别。

**散热方向，其重要性正持续提升。**在手机 AI 化的发展趋势下，设备算力与功耗同步攀升，散热能力已成为保障手机稳定运行的核心要素；而轻薄化的设计导向大幅压缩了手机内部的散热空间，高集成度核心零部件的应用又进一步加剧了散热压力，预计 VC 散热板方案后续将逐步下沉至中低端手机领域。

**指纹识别方向，**相较于光学屏下，超声波指纹识别方案更轻薄、更省电，对屏幕要求更低，解锁体验更好，且解锁区域设计更灵活，预计后续将逐步渗透至安卓系中端机型。

**此外，在全球智能手机存量竞争的背景下，折叠屏成为硬件创新的焦点之一，**且中国已成为全球最大的折叠屏手机市场。2024 年 9 月，华为推出开创性三折屏产品 Mate XT；苹果则预计于 2026 年推出折叠 iPhone。价格、重量与厚度，正是制约折叠屏渗透率提升的三大关键要素。

#### 图117：手机核心硬件创新一览

	<b>光学</b>	镜头方面，玻塑混合镜头在提升进光量、降低镜头高度方面具有优势；潜望式下沉是重要创新之一，潜望式镜头通过微棱镜结构实现光学变焦，同时支持远景和微距拍摄；其他光学创新还包括可变光圈、超光谱摄像头、外挂镜头；CIS 方面，超大底（4/3 英寸）+ 大像素是升级趋势；TOF 系统，同3D结构光技术，可用于生物识别和手势识别。	<ul style="list-style-type: none"><li>模组：舜宇光学、丘钛科技、高伟电子、欧菲光</li><li>光学零部件：水晶光电、蓝特光学、瑞声科技、东田微</li><li>CIS：韦尔股份、思特威、格科微</li><li>TOF：力芯微</li></ul>
	<b>散热</b>	手机AI化趋势下，算力与功耗提升，散热成稳定运行关键，同时轻薄设计压缩散热空间，高集成度核心零件进一步加剧散热压力，预计VC散热板方案后续将逐步下沉至中低端手机领域。	<ul style="list-style-type: none"><li>手机散热：苏州天脉、思泉新材、飞荣达、中石科技</li></ul>
	<b>折叠屏</b>	在全球智能手机存量竞争的背景下，中国已成为全球最大折叠屏手机市场。24年9月，华为发布开创性三折屏产品Mate XT；苹果预计26年将推出折叠iPhone。影响折叠屏渗透率提升的三个关键要素分别为价格、重量和厚度。	<ul style="list-style-type: none"><li>折叠屏：精研科技、东睦股份、统联精密</li></ul>
	<b>指纹识别</b>	相较于光学屏下，超声波指纹识别方案更轻薄、更省电，对屏幕要求更低，解锁体验更好，且解锁区域设计更灵活，预计后续将渗透至安卓系中端机型。	<ul style="list-style-type: none"><li>指纹识别：汇顶科技、丘钛科技、欧菲光</li></ul>

资料来源：砍柴网，爱集微，汇顶科技官网等，国联民生证券研究所整理

**在硬件创新之外，AI 技术正成为智能手机行业新的增长驱动。**根据 IDC 数据，2025 年全球搭载生成式 AI 功能的智能手机出货量预计将突破 3.7 亿部，占整体市场份额 30%；随着应用场景扩展及用户认知提升，生成式 AI 功能预计向中端机型普及，2029 年 AI 手机占比有望超过 70%。与此同时，智能手机行业的价值正逐步从终端硬件向 AI 服务入口迁移。

**展望后续，2026-2028 年，AI 手机核心竞争力将从硬件转向 AI 生态整合能**

力，终端的产品价值将形成硬件基础+AI服务增值的双重形态。

图118：手机厂商 Agent 落地模式和进展



资料来源：月狐数据，先见 AI 等，国联民生证券研究所整理

具体来看，当前的 AI 手机 Agent 落地有两种形式，AI 手机+Agent 和 AI 应用+Agent。AI 手机+Agent 方面主要由手机品牌主导，智能手机的头部品牌凭借“自研 OS+端侧大模型+硬件定制”的全链条闭环生态，以及领先的市场份额，牢牢把握 AI 手机的核心定义权。

分析各品牌厂商的 AI 手机战略，苹果计划在系统级层面深度集成 Apple Intelligence，实现设备端实时处理+GPT 集成的端云协同能力；华为在 HarmonyOS 中集成盘古大模型，实现跨设备间的任务调度；小米则通过澎湃 OS 整合自研 NPU 与生态服务，实现人车家全生态 AI 智能；vivo 基于蓝心大模型 +OriginOS，同时自研 7B 参数的端侧模型，具备更强的多模态理解能力，聚焦于

优化影像 AI 能力。

以搭载 MagicOS 10 系统的荣耀 Magic 8 系列手机为例，全新进化的 YOYO 智能体的可自动执行场景已经达到了 3000+，已可以帮助完成用户绝大多数的日常繁琐操作，具备一键比价、一键找餐厅、一键生成工作报告、一键生成旅游攻略等各种功能。

图119：荣耀 MagicOS 10 系统



资料来源：雷科技，国联民生证券研究所

图120：支付宝正式启动智能体生态开放计划



资料来源：机器之心，国联民生证券研究所

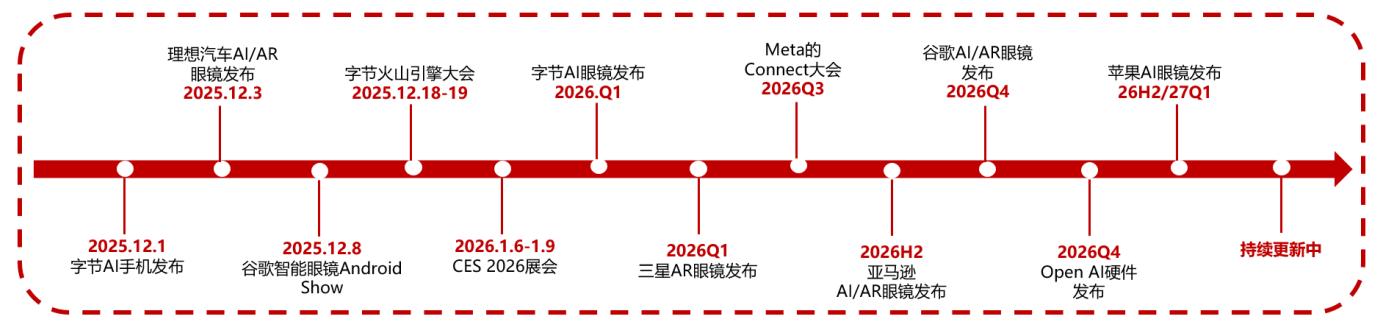
**AI 应用+Agent 方面则是由软件或大模型厂商主导**，借用厂商的应用生态壁垒+AI 能力，辅助应用具备智能感知、决策甚至行动的能力，加速将 AI 转化为生态护城河，抢占 AI 时代的流量入口。例如，微信元宝已实现小程序内智能下单、聊天场景任务提醒，但所有功能均限定在微信生态内；阿里千问则深度适配淘宝购物、支付宝缴费等服务场景，强化“AI+自有生态”协同。

## 9.2 新型 AI 终端：大厂布局新战场，竞逐 AI 时代新入口

**AI 终端新品热潮来袭，大厂竞逐下引领行业新航向。**目前，AI 终端领域正迎来关键发展节点，从发布会节奏看，2025H2-2026 年将是新型 AI 终端密集涌现的阶段，AI 手机、AI 眼镜等硬件产品持续落地，推动 AI 技术从云端向终端侧深度渗透。Open AI、Meta、字节跳动等国内外科技巨头纷纷聚焦 AI 硬件赛道，动作频频。在此格局下，头部厂商对 AI 硬件的核心定义、产品形态及推出节奏，将主导行业方向，成为引领市场预期、重塑竞争格局的关键风向标。

**从大厂的硬件发展方向来看，各家布局领域各不相同。**OpenAI 的初代终端产品外形可能是类似没有屏幕的智能音箱，主打音频交互形态；谷歌、Meta、苹果等厂商纷纷涌入 AI/AR 眼镜赛道；字节跳动则携手中兴通讯，将大模型融入手机系统，推出豆包 AI 手机，后续也将进入 AI/AR 眼镜赛道。

图121：后续 AI 终端发布会时间表



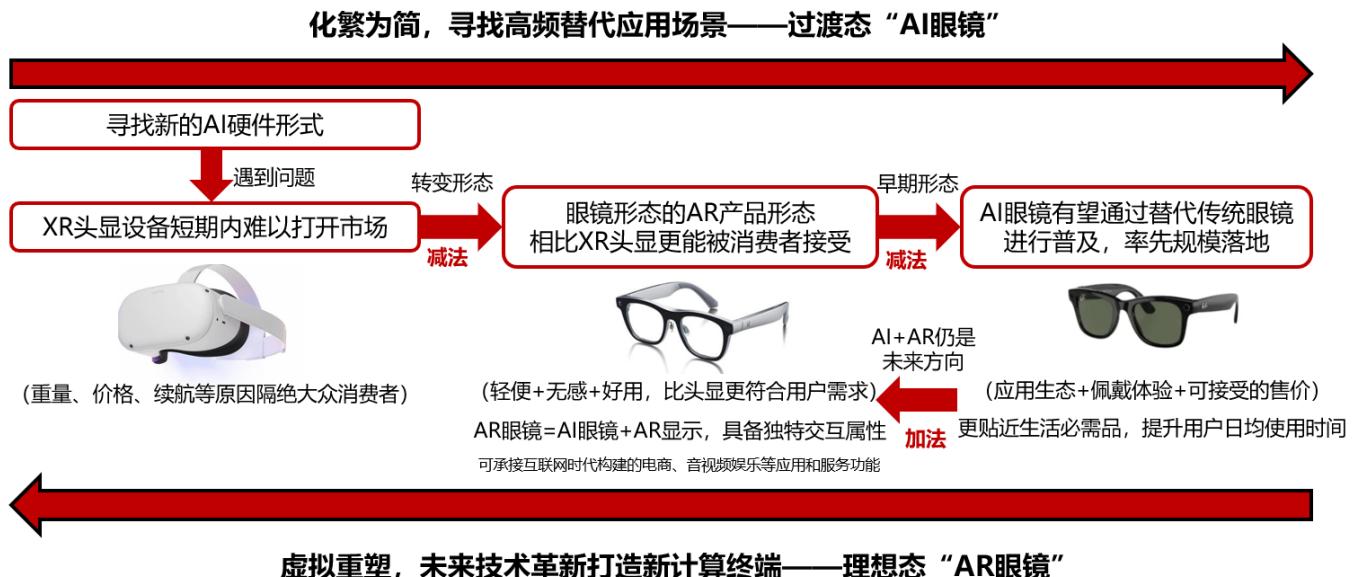
资料来源：VRAR 星球，三次方 AIRX，APPSO 等，国联民生证券研究所整理

**在诸多新智能硬件中，眼镜成为 AI 大模型最受关注的载体之一。**眼镜作为人身上一个重要的传统穿戴设备，具备广阔的智能化改造潜力。眼镜形态靠近人的耳、眼、嘴三个最重要的感官器官，可随时随地、自然、直观地与多模态 AI 进行交流。用户还可通过 AI 智能眼镜听歌，与内置大模型聊天，拍摄图片或短视频等，并将看到的信息与大模型共享。同时，大模型可通过摄像头实现对周围环境的感知，从而实现图像识别等视觉处理操作，进行更全面的信息处理。此外，眼镜的 AR 显示能力为 AI 提供文本和图像输出能力，让用户能接收到更多的图文信息。

随着 2023 年 RayBan Meta 的横空出世，市场也逐步对 AI 眼镜的产品形态达成共识，即眼镜硬件+AI 功能+拍摄+语音的产品形态，并在 AI 功能、价格、实用性上做好了平衡。

**从产品形态来看，AI 眼镜是一种符合当前市场预期和消费者认知的过渡期形态。**当用户对 AI 眼镜的接受度逐步提高，眼镜产品由于其具备贴近用户视觉的特性，在形态上会逐步叠加光学显示模块（光波导+光机），走上 AI+AR 的道路。其中，AI 与 AR 的能力是相辅相成的，AI 可以提升 AR 交互的智能性（如手势识别、眼动跟踪等），AR 则是 AI 合适的显示载体。**我们认为，后续的智能眼镜产品节奏或是 AI 先行，探索 AR。**

图122：智能眼镜演变历史



资料来源：国联民生证券研究所整理

**智能眼镜作为系统级产品，产品体验取决于 AI 大模型+应用生态适配+多元化交互，其中 AI 成为眼镜的核心赋能引擎：**

1) **AI 终端的定价=硬件成本+AI 体验，AI 体验成为终端价值的关键一环：**

与早先的消费电子创新，如 TWS 耳机对比，AI 眼镜具备诸多创新难题，因此产品形态的升级演进需要循序渐进。但优点在于，一旦用户体验成熟，带来便利解决刚需，用户的付费意愿更强。

2) **AI 交互重塑体验：**目前，主流 AI 眼镜均主打语音交互，从而解放双手，未来还将结合手势识别等功能，进一步革新交互体验。未来，手机定位有望转向随身算力终端，交互过程可更多通过 AI 眼镜实现，即：①通过眼镜的摄像头，实现“所见即所得”；②通过麦克风，实现“所言即所行”。

图123：AI 全面赋能智能眼镜硬件，造就全新体验



资料来源：维深 wellsenn XR，国联民生证券研究所整理

**AI 眼镜的硬件需求主要集中在 SoC 处理器上。**通过在传统眼镜上配备摄像头、麦克风等传感器，将其智能化改造，然后将传感器收集到的信息传输给内置的 SoC 处理器进行分析和处理。此外，AI 眼镜由于本身具备空间结构和使用时长的限制，对适用的 SoC 处理器也提出了低功耗、高性能等需求。根据 Wellsenn XR 数据，在 Ray Ban Meta 眼镜的成本拆解中，SoC 处理器是占比最大的单一硬件结构，占比约 34%。

**从眼镜穿戴系统的算力架构来看，其核心逻辑是通过分布式算力网络兼顾功耗、性能、成本与体积的多重需求。**具体而言，该架构由眼镜侧、手机侧、云端三部分算力构成，各端算力彼此协同运作实现 AI 功能。

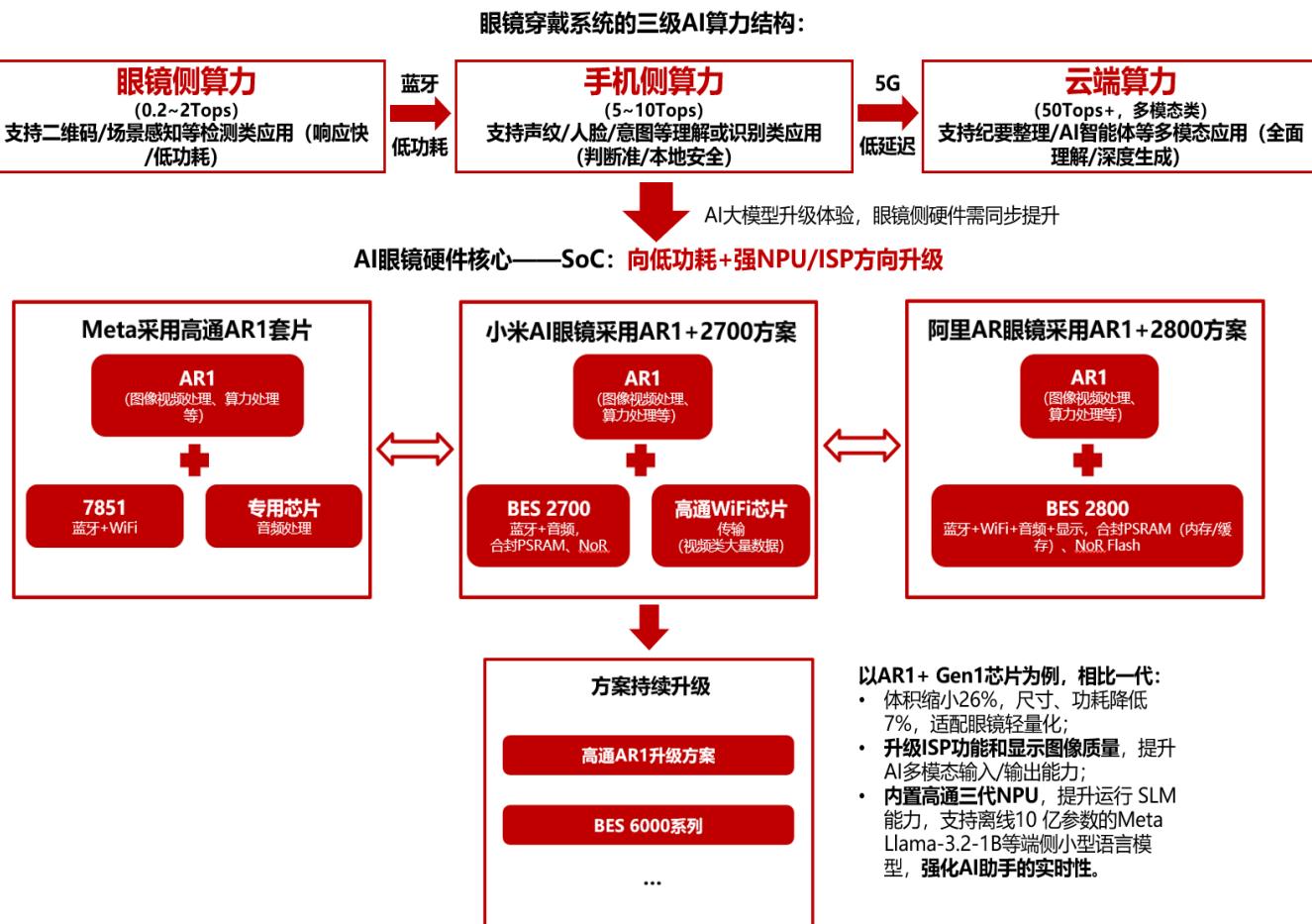
**眼镜侧的算力**在 0.2-2Tops 之间，适合做二维码识别、场景检测等低功耗任务，主要需求是相应速度快+低功耗；**手机侧算力**则会有所提高，至少能达到 5-10 TOPS，从而可支撑复杂的识别和意图理解等应用，具备本地安全可处理隐私；**云端**则能提供 50 TOPS+ 的多模态算力，负责使用眼镜过程中更全面的理解与生成。三层架构系统间分别通过蓝牙、5G 来完成数据传输。

因此 SoC 作为 AI 眼镜的硬件核心，迭代主要是往低功耗、高性能（强 NPU、ISP）的方向发展，从芯片架构来看目前主流有三种差异化方案：

**1) Meta 采用套片形式做功能拆分**，由高通 AR1 主芯片负责算力与显示，搭配单独的 7851 芯片承担蓝牙和 WiFi 连接功能，同时配备专用芯片处理音频；**2) 小米 AI 眼镜则打造了独特的“双星架构”**，将蓝牙和音频功能交由恒玄科技 2700 芯片负责，在低功耗运行模式下实现音频常开、视频休眠，整体功耗表现更优；**3) 阿里夸克 AI 眼镜则选择集成度更高的双星方案**，搭载的恒玄科技 2800 芯片进一步整合 WiFi 与显示处理等功能，实现全系统更低功耗。

**展望后续**，随着高通 AR1 芯片迭代，或者恒玄科技 BES 6000 系列芯片的推出，预计都将从硬件层面为 AI 眼镜续航能力的增强提供支撑，同时带动产品成本下降，推动其价格进入更亲民的区间，逐步迈入大众化普及阶段。

图124：AI需求下全新眼镜硬件架构升级



资料来源：AR洞察，国联民生证券研究所整理

**相比而言，AR眼镜的硬件需求则相对较高，且其光学显示组件无法做到满足性能需求的同时实现轻量化量产。**除了需要功能强大的SoC处理器和各种传感器来支持复杂的图像处理和虚拟信息叠加功能外，AR眼镜还需要额外的显示技术和相应的光学组件来创建虚拟图像，光学组件的集成度和功耗会直接影响AR眼镜的重量和性能。因此，上游的光学模组元件制造工艺复杂且技术难度高，导致其成本高昂，成为制约AR眼镜量产普及的关键因素之一。

**与此同时，功能与产品体验是用户购买AI眼镜的核心驱动力，而AI多模态感知及语义理解技术的突破，叠加基于用户行为数据构建的个性化认知体系，正推动智能眼镜的功能与体验实现高速迭代。**

**个人助手级的交互体验方面，AI眼镜不同于过往的消费电子，依托AI多模态感知、语义理解+个性化认知体系，实现从“被动响应”到“主动服务”的体验跃迁，重构AI眼镜的能力框架，使其从信息显示工具蜕变为主动、智能的个人助手。**

图125：AI需求下全新眼镜硬件架构升级



资料来源：前方智能、新浪新消费等，国联民生证券研究所整理

**功能生态方面**，AI 大模型深度集成在眼镜中，既优化了实时翻译、语音交互等基础功能的实用性，又催生了多模态识别、AI 会议助手、AR 导航、声纹支付等新能力，叠加各类厂商针对不同场景的差异化创新，让 AI 眼镜实现从“可用”到“好用”的跨越。此外，目前智能眼镜功能聚焦，未来可探索特定场景角色型智能体，在会议、翻译等高频场景实现主动服务升级。

以 Meta 在 2025 年 9 月发布的首款消费级 AR 眼镜 Meta Ray-Ban Display 为例，眼镜运行 Meta 基于安卓开发的操作系统，具备拍摄照片和视频、音频和视频通话、消息传递、步行导航、实时翻译和 AI 助理等功能。此外，眼镜还附带 sEMG 腕带，用户可以用不同的手势来与眼镜系统做深度的交互。

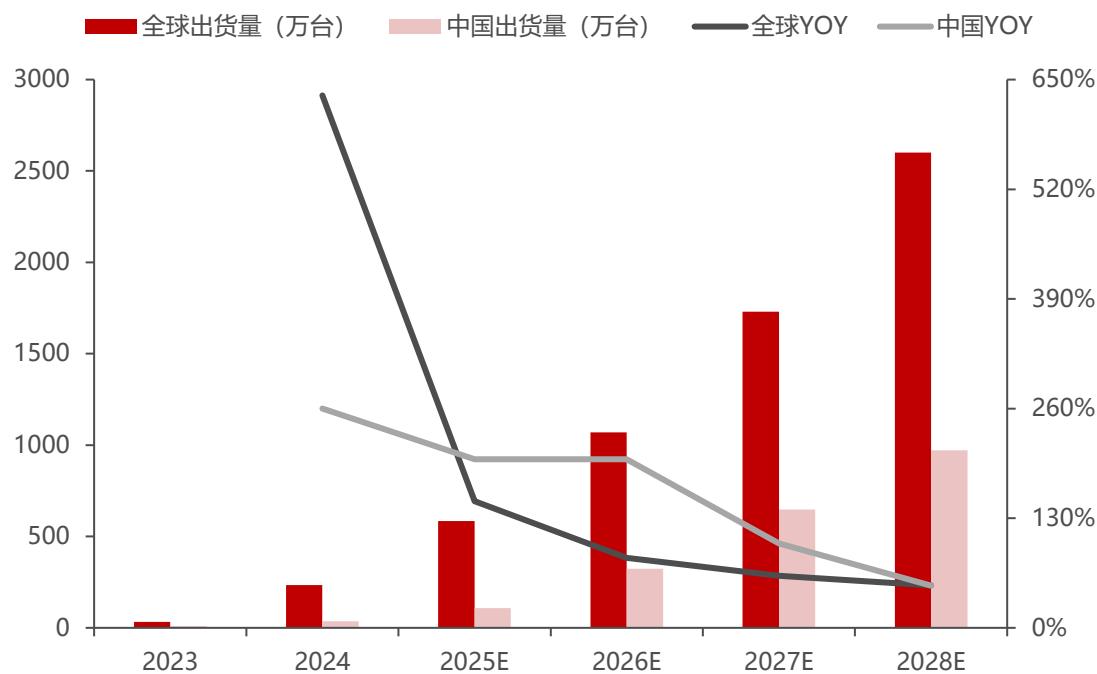
图126：Meta Ray-Ban Display 眼镜及功能生态



资料来源：淘宝 APP，国联民生证券研究所整理

从智能眼镜出货量来看，自 2024 年 Ray-Ban Meta 率先为眼镜植入 AI 能力后，国内外众多品牌纷纷入局 AI 眼镜赛道，直接推动全球智能眼镜出货量迎来井喷式增长，据艾瑞咨询预测，全球 AI 眼镜出货量将于 2026 年突破千万级规模，并在 2028 年攀升至 2600 万副。

图127：2023-2028 年 AI 眼镜出货量规模和预测



资料来源：洛图科技数据，艾瑞咨询，国联民生证券研究所

回到产品本身，智能化升级正催生出多元创新品类并推动其快速落地。从芯片厂商布局看，SOC 作为智能硬件的核心主控，迭代周期短、制程工艺升级快，叠加行业走入上行周期、下游 AI 智能硬件需求攀升，**各家芯片厂商立足于自身优势领域，贴合细分下游市场趋势推进 SOC 芯片迭代升级：**

**瑞芯微：**公司 RK3588 采用高性能 CPU 和 GPU 内核，且带有 6T NPU 处理单元，针对端侧主流的 2B 参数数量级别的模型运行速度能达到每秒生成 10 token 以上，满足小模型在智能座舱、智能大屏、AR/MR、边缘计算等边、端侧部署的需求；同时，公司下一代旗舰芯 RK3688 将会在 CPU、NPU 等领域继续全面升级，适配 AI 时代下游的升级需求。

**晶晨股份：**公司推出 6nm ARM V9 架构芯片 S905X5，具备 4TOPS 端侧算力，拥有 4TOPS 端侧算力，可依托端侧智能能力本地实现同声字幕生成、翻译等功能，显著提升跨语言场景下的用户体验，广泛适配智能电视、机顶盒、AI 音视频终端、智能摄像头等设备。该芯片自 2024 年下半年商用上市后销量持续攀升，公司预计 2025 年全年销量将突破千万颗。

**全志科技：**公司 V881 视觉芯片集成双核高性能 RISC-V CPU 与 1TOPS NPU 算力，实现通用与专用算力协同提升，为前端人形跟踪、场景识别等复杂 AI 算法

提供充沛性能支持；同时 V881 深度融合专业影像处理单元与 AI 感知算法，聚焦 4K 高清及 AI 影像处理，广泛适配 AI 眼镜、运动相机、机器人视觉导航模块等智能视觉设备。此外，公司新推出的 12nm 八核平台芯片 A733，已联动客户开发多款 “+AI” 产品，在智能平板落地多类 AI 功能，未来将持续推动传统智能终端向 AI 终端升级。

**中科蓝讯：**公司 BT897X 系列芯片具备 RISC-V CPU+NPU+DSP 架构，支持浮点运算，基础功耗优化到 4mA 级别，音频指标达行业领先水平；BT897X 系列芯片搭载的 NPU 单元大幅提升 AI 通话算法运算能力并降低通话功耗，后续将配套多家头部客户的耳机项目。

**恒玄科技：**公司的 6nm 智能可穿戴芯片 BES2800 已广泛应用于智能手表、智能眼镜等低功耗智能硬件市场，尤其在智能眼镜领域，已落地阿里夸克 AI 眼镜、理想 AI 眼镜 Livis 等标杆产品。此外，公司的下一代芯片 BES6000 将采用 A+M 核的架构，同步升级 NPU 算力，产品预计在 26H1 进入送样阶段。

**炬芯科技：**公司自研存内计算技术的 ATS362X 芯片，整合高性能 AI 算力、专业音频处理能力与灵活接口设计，可提供更智能、高音质、个性化的 AI 音频解决方案；目前 ATS362X 芯片已在多家头部客户的高端音箱、Party 音箱项目立项导入，在专业声卡、调音台等领域的方案开发也取得良好进展。

**乐鑫科技：**公司的 ESP32-S3 已实现与字节跳动豆包、ChatGPT、百度文心一言等主流大模型的互联互通，新增向量指令可加速神经网络计算与信号处理，赋能图像识别、语音唤醒识别等应用，适配智能家居、AIoT 设备等领域；此外，公司的新产品 ESP32-P4 强化边缘 AI 性能，通过算力与无线连接一体化设计，拓展智能家居、工业控制等场景的端侧智能应用边界。

**泰凌微：**公司的端侧 AI 芯片 TL751X 具备双核 RISC-V CPU+NPU+DSP 架构，兼具 AI 运算能力与 BLE、Zigbee 等主流物联网无线连接协议支持，凭借高性能、低功耗及卓越端侧数据处理优势，已广泛应用于智能家居、智能办公、无线音频等领域。

**星宸科技：**公司第三代异构架构芯片 SSU9366，搭载 4 核 CPU+1.5Tops NPU，双摄满负载功耗仅 1.5W，主要应用于扫地机器人、泳池清洁机器人等小型机器人产品。

表 31：国产 SOC 芯片公司主要终端算力芯片对比

公司	产品型号	主要算力芯片		算力	制程
		芯片架构	应用场景		
瑞芯微	RK3588→RK3688	CPU+GPU+NPU 架构	智能座舱、智能大屏、AR/MR、边缘计算、IPC、NVR、高端平板、ARM PC 等	6 TOPS→ <b>32TOPS</b>	8nm→ <b>5/4nm</b> <b>(2026)</b>
		CPU: 4*A76+4*A55→ <b>8*A730+4*A530</b>			
晶晨股份	S905X5	<b>4 核 ARM V9</b> CPU+GPU+NPU	智能电视、智能机顶盒、AI 音视频终端、智能摄像头等	4 TOPS	6nm
全志科技	V881	双核 RISC-V CPU+NPU	AI 眼镜、运动相机、机器人视觉导航模块等	1TOPS	22nm→ <b>12nm</b> <b>(2025)</b>
中科蓝讯	BT897X	RISC-V CPU+NPU+DSP 扩展 CPU+NPU+DSP (可选)	TWS/OWS 耳机等	-	22nm
恒玄科技	BES2800→BES6000	CPU: 双核 ARM Cortex-M55 → <b>A+M 核</b>	TWS 耳机、智能手表/眼镜等	<b>逐步升级</b>	6nm
炬芯科技	ATS362X	基于自研存内计算的 CPU+DSP+NPU 多核异构	高端音响、Party 音响及专业音频设备等	<b>132 GOPS</b>	22nm
乐鑫科技	ESP32-S3	Xtensa 32 位 LX7 双核处理器+向量指令加速神经网络计算	智能家居、AIoT 设备等	-	40nm
泰凌微	TL751X	双核 RISC-V CPU+NPU+DSP	高端耳机、智能音箱、车载音响系统、游戏外设等 扫地机器人、泳池清洁机器人	-	22nm
星宸科技	SSU9366	<b>4 核 CPU+NPU</b>	人、陪伴机器人和割草机器人等小型机器人产品	<b>1.5TOPS</b>	-

资料来源：高院科研服务、张江芯在线等，国联民生证券研究所

## 10 投资建议

**回顾与展望，AI 投资的机遇和挑战：**我们于 25 年初的深度报告《AIDC 电源系列一：“速率+功率”为未来 AI 产业发展的核心矛盾》中首次提出“速率+功率”为未来 AI 产业发展的核心矛盾。在过去的一年内，无论是速率赛道的光+PCB，还是功率赛道中的电源+液冷，都走出了“波澜壮阔”的行情。

那么站在当下，我们怎么看未来一年的算力机遇？目前市场对 CSP 厂商的资本开支始终有所疑虑，担心 ROI，担心远期增量不明朗。我们认为，26 年要重点观察 CSP 及大模型厂商的商业闭环节奏，从而把握整体行业  $\beta$ 。同时，**积极找寻价值量扩张、资本开支增量倾斜的细分赛道，主线延续“速率+功率”**。

**从资本开支到 ROI 测算，解读算力核心变量。**我们认为，算力需求主要看 Tokens 数+Capex。其中，Token 数（包括日活等）主要反映实时的算力需求，而 Capex 则反映云厂商的未来算力预期。部分商业闭环良好的云厂商，诸如谷歌等，已形成“开支→算力→Token→收入→再开支”的正循环。

我们在第一章中，主要测算了各大云厂商的 Capex/经营现金流/ROI，从而衡量公司可持续投资，及 AI 商业闭环能力。

**把握 AI 的增量赛道：**

**海外算力方面，我们延续“速率+功率”的投资思路：**

**1) 速率：光：**把握光入柜内的趋势，抓住光模块的业绩线、光芯片的缺货潮、硅光的渗透率提升趋势。关注超节点技术带来的 OCS 等产业趋势。**PCB：材料+设备升级是核心焦点。** NV 推出全新 PCB 解决方案，M9 等级基材、HVLP4 铜箔与石英纤维布构建的 PCB 正交背板方案成为升级趋势，同步拉动材料和设备升级。

**2) 功率：**单卡和机柜功率密度持续提升，对电力架构提出了新的要求，也使得液冷成为数据中心的标配。

**国产算力方面：25 年破局，26 年有望高速增长。**需求侧，国产大模型加速追赶，云厂商资本开支展望积极；供给侧、国产先进制程从单点突破走向多点开花。行业供需两强之下，国产算力厂商迎破局元年。

**其他方面：**半导体，关注 AI 赋能下的存储超级周期，设备受益原厂扩产。消费电子，关注 AI 终端，跟踪华米 OV、OpenAI、Meta 等行业龙头的探索。

**投资建议：**算力产业是科技之基，我们长期看好、深度跟踪。在当前市场对远期增量仍有所担忧之际，我们建议**积极寻找价值量扩张、资本开支增量倾斜的细分赛道，主线延续“速率+功率”**。同时**重点关注国产算力、半导体设备、存储、AI 终端的投资机遇。**

**表 32：重点公司盈利预测、估值与评级**

代码	简称	股价 (元)	EPS (元)			PE (X)			评级
			2025E	2026E	2027E	2025E	2026E	2027E	
工业富联	601138	60.69	1.83	3.27	4.32	33	19	14	推荐
胜宏科技	300476	264.39	5.76	9.83	14.37	46	27	18	/
生益科技	600183	74.52	1.41	2.11	2.83	53	35	26	/
中芯国际	688981	128.55	0.67	0.84	1.05	192	153	122	推荐
兆易创新	603986	323.68	2.33	3.24	3.99	139	100	81	推荐
拓荆科技	688072	378.00	3.22	4.70	6.28	117	80	60	推荐
东山精密	002384	74.83	1.57	2.13	2.70	48	35	28	推荐

资料来源：iFind，国联民生证券研究所预测；

(注：股价为 2026 年 1 月 28 日收盘价；未覆盖公司数据采用 iFind 一致预期)

## 11 风险提示

- 1) AI 产业发展的不确定性：**AI 产业技术快速迭代，存在算法瓶颈与算力成本压力，商业化落地进程存在低于预期的可能性。同时，全球监管政策趋严，在数据隐私与伦理等方面的变化，可能制约技术应用的速度与范围，为相关投资带来风险。
- 2) AI 资本开支不及预期：**AI 发展依赖大规模资本投入，若宏观经济走弱、企业盈利承压或 AI 投资回报周期过长，主要云厂商可能削减或放缓在 AI 领域的开支，从而直接影响上游硬件供应链的订单与增长能见度。
- 3) 下游需求不及预期：**全球宏观经济与消费信心仍面临挑战，若手机、PC 等传统消费电子需求复苏乏力，或 AI 硬件、汽车电子等新兴需求渗透不及预期，将导致终端厂商库存去化缓慢、下单谨慎，并通过产业链传导，拖累行业整体盈利修复。
- 4) ROI 测算局限性：**AI 商业模式仍处于早期探索阶段，Tokens 转化收入的效率受大模型降价、技术路线迭代等不可控因素影响，实际 ROI 可能低于模型测算值。

## 插图目录

图 1: OpenRouter 模型 Token 使用量排名 (2025.12.22-12.29) .....	5
图 2: 各模型厂商 Token 市场份额占比 (2025.12.22-12.29) .....	5
图 3: 英伟达数据中心收入按照客户拆分 .....	7
图 4: 云厂商资本开支情况 (亿美元) .....	8
图 5: 谷歌在端侧和云侧大模型的布局 .....	9
图 6: 谷歌 Gemini 3 模型 .....	10
图 7: 谷歌 Nano Banana 模型 .....	10
图 8: 谷歌 TPU v7 机柜实物图 .....	11
图 9: 谷歌 TPU v7 机柜及 Tray 内示意图 .....	11
图 10: 2024 年全球云计算市场市占率 .....	12
图 11: Azure Copilot 用例 .....	13
图 12: 生成式 AI 在不同任务下的使用占比 .....	13
图 13: META 开源项目图示 .....	15
图 14: OpenAI、Oracle 与 NVIDIA 的合作循环 .....	16
图 15: 4Q22-3Q25 北美云厂商 ROIC .....	18
图 16: 4Q22-3Q25 北美云厂商资本开支/经营现金流 .....	19
图 17: 蛋白结构预测式 AI .....	20
图 18: 马来西亚宣布启动国家级 AI 基础设施战略 .....	22
图 19: 新加坡 Sea - Lion 模型概念图 .....	22
图 20: 英伟达加速卡升级路线图 .....	23
图 21: 英伟达机柜升级路线图 .....	24
图 22: 英伟达 Scale-up 升级路线 .....	24
图 23: CSP 厂商在 ASIC 领域的路线图 .....	25
图 24: 谷歌 TPU V7 (Ironwood) .....	26
图 25: 亚马逊发布 Trainium3 .....	26
图 26: Meta MTIA .....	27
图 27: 微软 Maia 100 .....	28
图 28: 工业富联产品系列布局 .....	28
图 29: 2020-2025 年 Q3 工业富联营收和利润情况 .....	29
图 30: 2020-2024 年工业富联营收结构 .....	29
图 31: CPO 指把光引擎和交换芯片共同封装在一起的光电共封装 .....	31
图 32: CPO 技术路线, 2D 平面 CPO、2.5DCPO 和 3DCPO .....	32
图 33: 基板线路、PCB 线路、通孔等产生一定损耗, 且传输速率越大损耗越大 .....	33
图 34: 光引擎结构 .....	33
图 35: 硅光引擎 .....	34
图 36: MPO 连接器组件 .....	34
图 37: 光纤连接系统图 .....	35
图 38: CPO Switch 工作原理展示 .....	35
图 39: 柔性背板与自动化光纤设计示意图 .....	36
图 40: 英伟达 Quantum 3400 X800 .....	36
图 41: 光纤阵列单元 (FAU) 的组装与光反射结构剖面图 .....	37
图 42: 光纤阵列单元 (FAU) 的组装与光反射结构剖面图 .....	37
图 43: 矩形微透镜阵列 (左) 与六角形微透镜阵列 (右) 示意图 .....	38
图 44: 保偏 PANDA 光纤 (左)、蝴蝶结光纤 (中) 椭圆形光纤 (右) 横截面图 .....	38
图 45: 博通 QSFP-DD ARLM-96F8DMZ 激光模块产品图 .....	39
图 46: 使用 OCS 替代 spine 层的网络架构 .....	40
图 47: 传统数据中心架构和 Apollo OCS 架构的对比 .....	41
图 48: TPU 与 OCS 连接示意图 .....	41
图 49: 使用两个 MEM 阵列的 OCS .....	42
图 50: Coherent 光开关 (使用 DLC 技术) .....	42
图 51: Polatis 单模 576 x 576 矩阵光开关 .....	43
图 52: 赛微电子制造的 8 英寸 MEMS-OCS 晶圆示意图 .....	44
图 53: 2019-2028 年全球 PCB 市场规模 (亿美元) .....	45
图 54: 高密度连接板 (HDI) 示意图 .....	46

图 55:	Nvidia Rubin CPX.....	47
图 56:	VR NV144 CPX Computer Tray 侧视图.....	47
图 57:	VR NV144 Computer Tray 中使用 Midplane .....	48
图 58:	Rubin Ultra NVL576 / 背板结构图 .....	49
图 59:	Oberon 机架架构与 Kyber 机架架构对比 .....	50
图 60:	覆铜板 .....	51
图 61:	覆铜板电性能等级 .....	52
图 62:	玻璃纤维布 .....	53
图 63:	石英纤维布 .....	54
图 64:	铜箔分类 .....	54
图 65:	可剥离超薄铜箔产品结构 .....	56
图 66:	NVIDAGPU 机柜功率变化进程 .....	60
图 67:	数据中心电源架构 .....	61
图 68:	台达高压直流方案 .....	64
图 69:	台达 SST 方案 .....	64
图 70:	伊顿 MVSST 数据中心应用场景实例 .....	65
图 71:	中恒电气高压直流方案成功案例 .....	65
图 72:	科华数据直流方案系统原理图 .....	66
图 73:	产业链构成 .....	66
图 74:	液冷产业链拆分 .....	67
图 75:	GB300 NVL72 compute Tray 示意图 .....	68
图 76:	GB300 NVL72 Switch Tray 示意图 .....	68
图 77:	GB300 NVL72 液冷方案：机房内实物拆解 .....	68
图 78:	BAT 资本开支及增速（亿元，%） .....	70
图 79:	2021Q1-2025Q3（自然年）阿里巴巴资本开支 .....	71
图 80:	2022Q1-2025Q3 腾讯资本开支 .....	71
图 81:	豆包和 DeepSeek 与其他大模型的测试对比 .....	73
图 82:	DeepSeek-V3.2 与其他模型在各类数学、代码与通用领域评测集上的得分 .....	74
图 83:	AI 算力产业链 .....	76
图 84:	2021Q1-2025Q3 中芯国际单季度营业收入 .....	77
图 85:	2023Q1-2025Q3 中芯国际单季度归母净利润及少数股东损益 .....	78
图 86:	中芯国际关键技术节点的量产时间 .....	78
图 87:	24Q4-25Q1 中芯国际月度产能（折合 8 英寸，万片/月）及稼动率 .....	79
图 88:	摩尔定律放缓&单晶体管成本提升 .....	79
图 89:	不同制程芯片开发成本（亿美元） .....	80
图 90:	先进封装发展两大路径 .....	80
图 91:	台积电用于 Chiplet 技术的三种封装工艺 .....	81
图 92:	海力士 HBM2-HBM4 的技术路径 .....	82
图 93:	先进封装市场规模 .....	82
图 94:	2024 年全球前十大 OSAT 公司 .....	83
图 95:	2024 年全球前十大先进封装公司 .....	83
图 96:	2024 年收入 10 亿元以上的中国本土 OSAT 公司 .....	83
图 97:	2020-2029 年中国 GPU 市场规模收入（单位：亿人民币） .....	84
图 98:	2020-2029 年中国 AI 智算 GPU 市场规模收入（单位：亿人民币） .....	85
图 99:	寒武纪产品策略 .....	88
图 100:	2019-2025Q3 寒武纪存货与预付款项 .....	89
图 101:	沐曦 GPU 产品系列研发进程 .....	92
图 102:	芯原在手订单金额 .....	95
图 103:	存储周期复盘 .....	97
图 104:	数据中心存储需求市场规模 .....	98
图 105:	不同温度数据与数据介质对应关系 .....	99
图 106:	Nearline HDD、eSSD、Tape 在数据中心的出货量趋势 .....	100
图 107:	两大龙头 HDD 厂商资本开支变化（亿美元） .....	101
图 108:	NL HDD 交期延长，加速 SSD 替代 .....	101
图 109:	存储原厂毛利率季度变化 .....	101
图 110:	存储原厂资本支出变化（单位：亿美元） .....	102
图 111:	全球半导体设备市场规模变化 .....	102

图 112: NAND 技术发展趋势 .....	103
图 113: LAM 技术发展趋势 .....	103
图 114: 键合工艺 .....	104
图 115: AI 跨越式发展+硬件成熟, 开启终端新一轮产业变革 .....	105
图 116: 2015-2024 年全球智能手机出货量年度数据 (百万部) .....	106
图 117: 手机核心硬件创新一览 .....	107
图 118: 手机厂商 Agent 落地模式和进展 .....	108
图 119: 荣耀 MagicOS 10 系统 .....	109
图 120: 支付宝正式启动智能体生态开放计划 .....	109
图 121: 后续 AI 终端发布会时间表 .....	110
图 122: 智能眼镜演变历史 .....	111
图 123: AI 全面赋能智能眼镜硬件, 造就全新体验 .....	111
图 124: AI 需求下全新眼镜硬件架构升级 .....	113
图 125: AI 需求下全新眼镜硬件架构升级 .....	114
图 126: Meta Ray-Ban Display 眼镜及功能生态 .....	114
图 127: 2023-2028 年 AI 眼镜出货量规模和预测 .....	115

## 表格目录

重点公司盈利预测、估值与评级 .....	1
表 1: 头部 AI 大模型定价对比 .....	4
表 2: 谷歌各代际 TPU 性能对比 .....	11
表 3: AWS 2025 年 AI 芯片业务核心数据一览 .....	14
表 4: 谷歌大模型相关业务发展动态 .....	15
表 5: OpenAI 在算力领域的合作伙伴 .....	16
表 6: OpenAI 各行业的代表性客户 .....	17
表 7: 主权 AI 与商业云厂商对比 .....	19
表 8: 不同国家主权 AI 项目介绍 .....	21
表 10: 部分高速覆铜板基体树脂 .....	53
表 11: HVLP 铜箔 1-5 代情况 .....	55
表 12: 钻孔技术分类 .....	57
表 13: 激光钻孔设备介绍 .....	58
表 14: 鼎泰高科、四方达、沃尔德、中钨高新产品与核心技术介绍 .....	59
表 15: 单卡功耗提升进程 .....	60
表 16: 机柜功率密度提升带来的挑战 .....	61
表 17: 数据中心电源演变对比 .....	62
表 18: HVDC 产业进展 .....	62
表 19: HVDC 产业进展 .....	63
表 20: 液冷产业供应链 .....	69
表 21: 国产算力芯片主流“玩家” .....	85
表 22: 华为昇腾芯片路线图 .....	87
表 23: 海光核心产品介绍 .....	90
表 24: 沐曦核心产品介绍 .....	91
表 25: 摩尔线程 GPU 架构芯片演进情况 .....	93
表 26: DRAM/NAND 未来一年价格预测 .....	98
表 27: 一般服务器与 AI 服务器平均容量差异 .....	99
表 28: 内存互连芯片在不同类型的内存模组中的应用及配比关系 .....	100
表 29: 长江存储与长鑫存储扩产相关信息 .....	103
表 30: 2025 年 AI 终端渗透率及驱动因素 .....	106
表 31: 国产 SOC 芯片公司主要终端算力芯片对比 .....	117
表 32: 重点公司盈利预测、估值与评级 .....	119

## 分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接受到任何形式的利益。

## 评级说明

投资建议评级标准	评级	说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；北交所以北证 50 指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级 推荐	相对基准指数涨幅 15%以上
	谨慎推荐	相对基准指数涨幅 5%~15%之间
	中性	相对基准指数涨幅-5%~5%之间
	回避	相对基准指数跌幅 5%以上
行业评级 推荐	推荐	相对基准指数涨幅 5%以上
	中性	相对基准指数涨幅-5%~5%之间
	回避	相对基准指数跌幅 5%以上

## 免责声明

本报告由国联民生证券股份有限公司或其关联机构制作。国联民生证券股份有限公司具有中国证监会许可的证券投资咨询业务资格。本报告的分销依据不同国家、地区的法律、法规和监管要求由国联民生证券于该国家或地区的具有相关合法合规经营资质的子公司/经营机构完成。在遵守适用的法律法规情况下，本报告亦可能由国联证券国际金融有限公司在香港地区发行。国联证券国际金融有限公司具备香港证监会批复的就证券提供意见（4号牌照）的牌照，接受香港证监会监管，负责本报告于中国香港地区的发行与分销。

本报告仅供本公司授权之机构及个人使用，本公司不会因任何人收到本报告而视其为客户。本报告仅为参考之用，并不构成对任何人的操作建议或任何保证，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑获取本报告的机构及个人的具体投资目的、财务状况、特殊状况、目标或需要，客户应当充分考虑自身特定状况，进行独立评估，并应同时考量自身的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见，不应单纯依靠本报告所载的内容而取代自身的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事；本公司自营部门及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。提示客户及公众投资者慎重使用未经授权刊载或者转发的本公司证券研究报告，慎重使用公众媒体刊载的证券研究报告。

本报告的版权仅归本公司所有，未经书面许可，任何机构或个人不得以任何形式、任何目的进行翻版、转载、公开传播、篡改或引用，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。所有在本报告中使用的商标、服务标识及标记，除非另有说明，均为本公司的商标、服务标识及标记。本公司版权所有并保留一切权利。

 无锡 江苏省无锡市金融一街 8 号国联金融大厦 8 楼

 上海 上海市虹口区杨树浦路 188 号星立方大厦 B 座 7 层

 北京 北京市西城区丰盛胡同 20 号丰铭国际大厦 B 座 5F

 深圳 深圳市福田区中心四路 1 号嘉里建设广场 1 座 10 层 01 室

