文件帖 2026-09-21 13:42

来源：刨米机AI精选 | 先知股票数据源

--- 国内大模型调研 —— 参数竞赛加速，推理算力成为商业化核心瓶颈，国产卡涨价压制云厂商盈利，模型厂商和云平台的竞争与合作格局.pdf ---
国内大模型调研 —— 参数竞赛加速，推理算力成为商业化
核心瓶颈，国产卡涨价压制云厂商盈利，模型厂商和云平台
的竞争与合作格局
聚焦：DeepSeek、智谱、Kimi、阿里、腾讯
调研覆盖维度
 大模型研发方向与参数扩张规划
 国产推理算力及成本约束
 国内模型厂商竞争格局
 LLM / Agent / 多模态能力及算力资源配置
访谈对话：Chris（访谈人）、行业专家
1. 大模型整体研发战略与参数扩张目标
专家：整体没有大的战略调整，有两个核心研发重点：
1. 核心大语言模型：当前 Seed 系列参数量偏小，阶段性目标做到 2T 3T 参数。
2. 多模态输入模型：与基础大语言模型属于两条并行研发路线，资源投入互不冲突；
重点推进和手机、汽车等终端硬件厂商合作。
3. 其他模态方向：重点布局 30 帧级别、融合基础知识能力的视频生成模型。
Chris：从 1T 向 2 3T 参数升级背后的驱动原因？
专家：分为商业竞争、商业模式、能力上限、营收价值多重因素。
1. 外部竞争压力：国内云厂商需要对标海外 OpenAI、Anthropic，同时直面 Kimi、
智谱、DeepSeek 等专业模型公司的竞争。
2. 商业模式博弈：
 理想模式：专业模型厂商不做重算力，云厂商接入外部模型做 PaaS 分成，
类似数据库与云厂商的合作模式。
 国内特殊现状：字节、阿里、腾讯、百度均为全链路布局（IDC 算力 云
模型 应用）。如果自有核心模型在中高端场景失去竞争力，AI 云高价值业务将
会流失，只能做低毛利分成业务，商业化损失巨大。

3. 模型能力边界：800B/1T 以内模型仅擅长通识问答、基础对话；2 3T 参数规模
可以带来更强 Agent 能力，能力广度、深度与客户 ARPU 单价强相关。超过该量级
的模型，往往混合视觉等其他模态模块。
4. 营销因素：市场存在参数宣传现象，但只是次要因素。
举例：腾讯早期主打 “小而美” 200 亿参数模型，现已推进 700 800 亿参数模型。
2. 做大参数量模型需要攻克的关键要点
分为技术资源层面、市场商业化层面两大维度。
2.1 技术与资源层面
1. 预训练稳定性与数据质量
 MoE 架构预训练收敛是核心技术难点；Kimi、DeepSeek 均使用 Kimi 的
MoE 参数稳定器。
 2025 年部分企业训练使用广告、自媒体、剧本、电商这类海外巨头舍弃的数
据；2026 年已经进入硬实力比拼阶段，高质量基础知识数据成为关键。
2. 后训练与 Agent 思维链数据
 高质量 Agent 步骤链数据至关重要；智谱 GLM 5.0/5.1 在该类数据积累反

馈较好。
正向飞轮：模型落地越早、用户体量越大，越容易积累优质思维链数据，直接决定编程、复杂任务表现。
3. 算力：训练易解决，推理是卡脖子环节
训练：依靠 “凑卡” 可以完成大模型训练；
推理：推理算力才是商业化落地最大阻碍，大尺寸模型推理算力消耗占比持续走高，例如 Kimi K3 受推理算力约束难以大规模市场化。
4. 模型分层运营：小尺寸高性能模型是盈利终局
API 市场中高利润来自表现优异的小尺寸模型。
案例：Anthropic 针对 Bug 修改、纯代码上下文场景，路由至 20B 30B 小模型，推理毛利最高；
技术特点：小模型做到高性能的技术难度，高于单纯堆叠大参数，是商业化盈利的终局形态。
2.2 市场商业化层面，做大参数模型需要三大要素
1. 合适的市场契机：适合具备云能力、大规模 C 端、重度 ToB 办公项目的企业；没有契机不必强行做大参数。
2. 配套行业生态建设：参考 Kimi 和同花顺、Wind 合作打造金融 Agent；大模型需要垂直场景生态释放价值。
3. 大型 To B 客户基础。
澄清市场传言：网传字节在训练 5T 10T 纯大语言模型属于误解。字节计划在 3 季度末 4 季度初发布多模态输入模型，多模态理解模块本身参数量巨大；该模型短期很难大规模商业化，但手机、汽车终端厂商有强需求。
3. 国产推理算力不足的拆解：供给、性价比、超节点三重困境
Chris：推理算力不足，是国产芯片性能弱于英伟达，还是供给短缺？
专家：多重因素叠加。
1. 供给端：国产芯片供不应求
寒武纪 690 综合表现优秀，但采购渠道壁垒高，多数厂商难以拿到充足供货；昇腾 950 同样紧缺，部分客户补贴资金、超额上报需求换取算力配额。
2. 成本与性价比（HBM 为核心瓶颈）
HBM、卡间互联受限；卡间互联可依靠超节点缓解，但 HBM 供给与价格被外部制约。
HBM 涨价带来连锁反应：国产芯片双向涨价、部分机型 HBM 配置下调；同等电力、机房面积条件下，国产卡算力密度低于英伟达。
商业风险：国产卡持续涨价，叠加行业价格战，有可能出现单 Token 亏损，压力向上传导至云厂商，破坏大模型商业闭环；传统企业 AI 转型预算有限，算力涨价会压制行业增量。
3. 超节点与智算中心基础设施约束
大参数量训练需要定制化超节点，普通机柜无法满足；
当前用户可以容忍 30 秒起步的推理时延，可部署在乌兰察布、贵州等低电力成本地区；
未来商业竞争白热化后推理算力需要贴近用户；国产卡算力密度不足，会抬高智算中心建设需求，带来 IDC 资源的新一轮约束。
4. 国产卡问题带来的影响与采购规划
1. 算力预算制定窗口：国内企业大多 11 12 月制定下一年算力预算，现阶段做明年采购规划尚早。
2. HBM、国产卡涨价会直接压低实际采购量。
3. AI 云厂商：算力成本上涨将在 2026 年末 2027 年初传导到 AI 服务售价，压制

AI + 行业转型的业绩预期。
4. 模型厂商（智谱、Kimi、DeepSeek 等）：推理算力供不应求叠加价格战，短期行业整体利润承压。
5. 国内大模型市场演变与终局格局预判
时间节点重点观察：2026 年底 2027 上半年，行业迎来一轮关键竞争；2027 上半年完成专业模型与云平台规模化接入落地。
5.1 专业模型厂商分化为两大赛道
1. 极致性价比路线：DeepSeek
代表产品 DeepSeek V4.1 Flash，多模态图像输入、推理成本低、价格优势明显；国内主流云平台均完成接入。
后续若推出闭源版本、新授权模式，会和云平台开展分成合作；短期该赛道地位难以撼动。
2. 极致推理能力路线：Kimi、智谱、MiniMax 三方竞争
Kimi：综合思维、知识广度见长，适配金融、设计行业；
智谱：编程、IT 类 Agent 任务口碑突出；
三家短期内不会决出终局，头部厂商会被各大云平台接入。
5.2 AI 公有云格局
市场份额：火山引擎、阿里云合计占据国内 AI 公有云 80% 以上份额；讯飞、华为、运营商为其他参与者。
1. 阿里云：千问系列覆盖 LLM、生图、视频、识图、语音，模型谱系完整。但市场感知上 DeepSeek/Kimi/ 智谱中至少两家综合能力优于千问， 阿里云需要适度妥协引入外部模型。
2. 火山引擎（字节）：Seedance 模型是当前营收主力；重点推进手机系统、车载、端云融合，业务扩张激进。
风险提示：若 2026 年底字节自有大语言模型无法取得质的突破，2027 年在与专业模型厂商博弈中火山引擎需要做出更多妥协，否则市场份额会流向阿里云（该判断不包含 Seedance 类多模态模型）。
5.3 专业模型厂商的新动向：自建算力 + 布局中低端模型
DeepSeek、智谱等企业获取 IDC 资源，选择自建一部分算力，不会完全依赖公有云。
布局中低端、小尺寸高性能模型：该赛道竞争激烈，云厂商分配资源有限，想要争夺营收就必须依靠自有算力。
注：以上分析针对大语言模型；视觉、视频生成、3D 等多模态赛道遵循另一套竞争逻辑，优势厂商会持续放大壁垒。
6. DeepSeek / Kimi / 智谱三者模型能力差异
假设满分 100 为理论通用推理上限。
1. DeepSeek：为极致低成本做架构取舍，架构上限难以突破 3T 参数；主观设定能力上限约 95 分，牺牲部分精度换取成本优势；V5 系列是否调整路线有待观察。
2. 智谱 AI（GLM）：标准 MoE 架构，一次推理激活路由专家 + 2 3 个任务专家，一次核心计算完成推理；速度快、推理成本可控，编程优势短期稳固；后续看 GLM 5.5 是否大规模重构架构。
3. Kimi：多步工程化解构任务，不同步骤激活不同专家，架构可灵活接入视觉子模块；超长任务链、复杂多工具协作场景表现更强，通用能力与知识广度占优。
结论：三家架构底层差异显著，短期内通用推理能力很难完全拉平对齐。
7. 字节大模型对比千问、Kimi、智谱：优劣势拆解
1. 大语言模型短板：属于偏科型，整体能力落后同行接近一代；无论极致推理成本、编程、长复杂任务链均不及智谱、Kimi。

2. 字节核心优势不在基座模型，而在应用工程化：豆包 App、网页端、PC 端的通用 Agent 工程设计优秀；实现国内最大用户规模、较好用户口碑。
7.1 字节大语言模型落后一代的根源
根本是关键时间节点的战略决策失误，累计落后约 15 24 个月。
1. 时间节点：GPT 3 发布后， 国内其他大厂迅速跟进，字节滞后半年才启动相关方向；
2. 人才路线失误：初代大模型由传统机器学习团队开发，没有引进大模型专项人才，项目耗时八九个月宣告失败；后续招聘也走过弯路；
3. 当前状态：依靠加大算力、人力投入追赶，已经弥补一部分差距。
7.2 字节算力分配策略：研发、训练、推理三个阶段优先级不同
1. 研发阶段（前期大量实验）：消耗算力最高，持续 3 6 个月；可使用海外机房算力，大语言模型、多模态之间不存在算力争抢。
2. 训练阶段：大语言模型训练主要放在国内机房，大语言模型训练优先级最高；文生图、文生视频部分训练可在海外完成。
3. 推理阶段：文生视频推理优先级最高，因为该业务毛利率更高。

--- 存储调研DRAMNAND 供给收缩及价格上行，HBM
挤占产能及长协机制升级，原厂客户分层与供货策略.pdf ---
存储调研 ——DRAM/NAND 供给收缩及价格上行，HBM
挤占产能及长协机制升级，原厂客户分层与供货策略
聚焦：三星、海力士、美光、闪迪、Google、Meta、微软、亚马逊等
1 DRAM/NAND 供给收缩及价格上行
2 HBM 挤占产能及长协机制升级
3 原厂客户分层与供货策略
专家问答观点整理
Q1：目前存储行业的整体供给状况如何？尤其是在 DRAM 和 NAND Flash 领域，不同产品线的供应情况是否存在差异？
当前存储行业的供给紧张主要受两大因素影响：首先是东西方供应链的切割，其次是原厂将产能优先分配给 HBM。
具体来看，三星电子已将部分产能迁至越南，SK 海力士正逐步将产能转移回韩国和日本，美光在台北的工厂也有向日本和美国本土迁移的动作。这种策略调整导致消费类产品线全面收缩。美光已停掉了其 To C 品牌 “英睿达”，SK 海力士关闭了三方授权的线上零售店等。
综合来看，所有原厂在消费类产品上都面临晶圆和产能不足的问题，代理商的 \\CSP（客户特定价格）\\ 额度也无法获批，无论是提交 Forecast、DP（需求计划）还是直接申请 CSP，都没有任何结果。
在 DRAM 领域，DDR4 和 DDR5 的供应均受到影响。
在 NAND Flash 领域，TLC 和 QLC 产品同样面临产能紧张。
Q2：中国 CSP 客户对存储产品的需求现状如何，以及这种需求增长背后的驱动因素是什么？
当前市场对 AI 和 HBM 的讨论非常热烈，普遍认为供应链缺口巨大，但这更多停留在

信息传播层面，并未完全反映在终端实际拆箱使用的环节。AI 相关的需求确实是主要驱动力，消耗了大约 60% 的产能，这一点在各类数据上都有体现。
然而，供应紧张的现象并非单纯由需求驱动，而是寡头厂商通过一系列市场操控手段，将原本供应冗余的局面转变为极度稀缺，分为七个演变阶段：
1. 第一阶段：正常的供需交易，例如以 50 美金的价格完成一笔 5K 数量的货品交易。
2. 第二阶段：厂商开始控制额度，向代理商保证持有货物不会亏损，并鼓励其尽快提货，此时市场售价已高于 50 美金。

3. 第三阶段：厂商制造稀缺性，宣称某一笔 5K 的货源仅此一家，迫使多个买家竞标，价高者得，从而将价格推高至 55 或 60 美金。
4. 第四阶段：厂商进一步控制价格和货源，维持局部市场的稀缺状态。
5. 第五阶段：厂商开始对区域代理商施压，要求必须产生销售记录和用户名单。如果半年后某个产品在当区没有销量，该代理商的资格就可能被取消。同时，厂商会进行 “点仓”，检查代理商是否存在囤货行为。如果发现代理商只是为了完成额度将货物结清后囤在自己仓库而不销售，其代理资格在三到六个月后也可能被取消。
6. 第六阶段：厂商强制推行 NCNR 协议，即订单一旦下达，便不可取消或更改，买方必须接受所供应的任何货物。
7. 第七阶段：引入预付款规则。以闪迪为例，除了签署 NCNR 协议，还要求代理商支付 30% 的预付定金，这被视为签订长期供货协议的前奏。
旧长协：属于 “君子协定”，同时锁定价格和数量，周期通常 2 3 年；几乎无预付款，依靠纸面承诺。
新长协：协议年限更长，转变为 take or pay 硬性绑定模式。客户必须支付 20% 30% 的定金，价格由供应商在区间内确定。核心关键条款：即使客户最终不提货，也必须支付 100% 的货款，而不仅仅是损失定金。
Q4：考虑到 SK 海力士、美光和三星电子均计划大幅扩张 HBM 产能，而新的长协协议据称设定的价格下限 (地板价) 较低，这是否意味着存储大厂对 2029 最新的阶段是签订长期供货协议，但与以往锁价锁量不同，现在的长约条款更为复杂。
Q3：美光和闪迪与客户签订的长协协议具体内容是怎样的？新旧长协协议在性质上有何变化？
新旧长协协议在性质上存在显著差异：

年前后的市场价格缺乏信心？
尽管 SK 海力士、美光、三星电子均宣布了 HBM 的扩产计划（分别计划扩产一倍、提升一倍交付能力、提高 60% 80% 制造交付能力），显示出市场对 HBM 的真实需求强劲且供应不足，但市场情绪依然脆弱。目前市场对任何负面消息都反应剧烈，导致客户普遍存在恐慌情绪。
美光签订的 16 份新长协协议中，客户类型和协议期限有所不同：
1. 超大型客户（Google、Meta、微软、亚马逊）：签订五年期长协，属于 HBM 和 DRAM 双料采购方，供应量占美光总销量 70%。
2. 中型客户：数据中心、通过中间商采购的苹果、阿里巴巴、腾讯；同样五年期协议，但合同条款相比超大型客户更为柔和。
3. 小型客户：主要为车规领域客户，包含通用汽车、福特汽车、特斯拉及其方案商。
美光当前战略由过去单一聚焦 HBM 客户的 “单战略” 升级为 \\“双战略”\\，重点同时发展数据中心和车规电子两大领域。
已签订长协产品结构占比：DRAM 类 66% 70%，NAND Flash 类约 33%。
Q5：新的长协协议中，价格区间的设定是怎样的？天花板价和地板价的具体水平如何？
新长协协议的价格区间设定非常宽泛。天花板价大致相当于 2026 年第三季度的市场价格或略高，而地板价则设定得非常低。
Q6：美光签订的 16 份长协协议具体覆盖了哪些客户？
1. 超大型客户：Google、Meta、微软、亚马逊；美光传统 HBM、DRAM 大客户。
其中 Google 订单包含大量 HBM3E，源于美光 HBM3E 良率与供应稳定性优势。
2. 中型客户：苹果、阿里巴巴、腾讯；通过中间商采购，属于中国云服务商渠道。
3. 小型车规客户：通用汽车、福特汽车、为特斯拉提供车载方案的供应商。
Q7：市场普遍预期存储芯片价格在 2026 年第三季度涨幅约为 10% 15%，第四季度涨幅将显著收窄，这与您在市场中观察到的情况是否一致？未来价格走势将如何演变？

市场实际季度涨幅与市场预期存在差异：

2026Q1 涨幅约 15% 19%
2026Q2 涨幅高达 50% 60%
2026Q3 涨幅收窄
预计 2026Q4 涨幅进一步放缓
2027Q1 涨幅预计仅为个位数
核心判断：价格涨幅逐步收窄，但价格绝对值已经处于高位，市场将在高位区间震荡。
Q9：闪迪在 NAND Flash 市场的长协机制是怎样的，与其竞争对手美光相比有何不同？
闪迪与美光在客户层级和合作模式上存在显著差异。举例：同一家分销商，在美光客户体系划分是 Tier3，但是在闪迪体系属于 Tier1 全国最大国代。
根源来自两家企业市场定位差异：闪迪更加看重零售品牌市场，全国大型分销商对其业务至关重要。作为闪迪 Tier1 客户，可以获得更高供应保障、更优合作条款，和美光 Tier3 客户的提货规格、合作待遇差距明显
Q8：您认为存储芯片价格需要很长时间才会回落的原因是什么？
1. 寡头垄断格局：DRAM 由三星、海力士、美光三家主导；NAND Flash 合计五家厂商。NAND 供应商更多，涨价相对平缓；DRAM 受 HBM 挤占约 60% 产能，供给紧张，涨幅更为剧烈。
2. 真实需求难以判断：虽然大量新建厂房、数据中心，但电力不足造成厂房闲置，同时 Meta 等企业出租冗余算力，市场难以分辨真实需求与潜在产能过剩。
3. 资本游戏特征：演变为生产方与使用方的资本博弈。当下游使用方资金承压，上游生产方会通过资金、货物、融资、合股等方式共同维持市场热度，支撑价格长期高位。
。
Q10：存储原厂在 HBM 产能受限的情况下，如何平衡不同产品线的生产与定价，以实现利润最大化？
HBM 占用大量产能后，原厂仅剩约 40% 产能去承载原本 80% 以上市场需求的其余产品线，企业级 DRAM、消费级、车规产品议价空间显著抬升。
原厂逐利逻辑：HBM 长协毛利约 58%；普通 DRAM 市场溢价可达 150%，消费类产品溢价最高至 300%。厂商会利用良率爬坡产生的冗余产能，生产高利润非 HBM 产品优化财报，同时重新谈判 HBM 合约价格。
典型案例：原厂推动客户从 HBM3/HBM3E 转向 HBM4/HBM4E 新产品签约，合约单价由 530 550 美元每颗抬升至 800 1400 美元每颗。若客户坚持采购良率稳定、供货充足的老款产品，原厂利用市场强势地位通过合约条款约束客户，保障自身利润。

--- CPO 龙头调研：CPONPO 放量在即、份额稳固，上游物流保障充足基本锁定.pdf ---
CPO 龙头调研：CPO/NPO 放量在即、份额稳固，上游物流保障充足基本锁定
关键时间标记：2026 年为 CPO 商用元年；英伟达体系 CPO 于 2026 年 8 月小批量出货。
CSP 厂商（谷歌、亚马逊）对英伟达 CPO 持观望态度，倾向自研 OCS 或采购 NPO 规避系统绑定；未来格局大概率为：柜内 OCS 与柜外 CPO 并存。
NPO 定位为 1.6T 之后的过渡方案；预计 2027 年 NPO 量产规模约 1000 万只。
专家问答整理
Q1：请分析当前行业中 1.6T 光模块、NPO 及 CPO 技术各自的推广节奏与市场进展？
1. 1.6T 光模块

2026 年市场由中际旭创占据绝大多数份额，新易盛紧随其后，其余厂商处于小批量、送样阶段。天孚通信光引擎式 1.6T 模块 2026 年 6 7 月开始出货。
2026 年 800G 光模块总规模约 4000 万只；
2027 年预计 800G、1.6T 光模块需求各自达到 6000 7000 万只。
2. NPO 技术
行业尚未形成统一标准，中际旭创、新易盛、华工科技、光迅科技各按自研标准推进。
2026 上半年：样品阶段；
2027 上半年：小批量生产；2027 下半年：大规模量产；
2027 全年 NPO 市场规模约 1000 万只，对当年 1.6T 分流有限；2028 年 NPO 会显著冲击 1.6T 市场。
3. CPO 技术
博通已小批量出货；英伟达体系 CPO 产品 2026 年 8 月开启小批量出货，部分组件来自天孚通信 6 月交付产品。
2026 英伟达 CPO 出货仅几千套；
2027 整体 CPO 交换机市场规模 8 10 万套；

CPO 大规模放量依赖第二代可插拔 DFAU 产品，目标时间点 2028 年，目标规模 100 万套。
应用节奏：2027 上半年 CPO 以柜式交换机场景；下半年第二代可插拔产品落地，拓展柜内连接。
行业判断：2026 是 CPO 商用元年，2027 是过渡与 CSP 客户接受的关键窗口；第二代产品 2027 年 6 7 月推出，供应链开放、成本下降，为 2028 大规模放量铺路。
Q2：在 NPO 市场中，英伟达的供应链采购逻辑是怎样的？天孚通信在其中扮演何种角色，市场份额前景如何？
英伟达 NPO 采购沿用现有光模块供应链，直接向中际旭创、新易盛、菲尼萨、Fabrinet 四家采购；英伟达不制定 NPO 标准，向供应链输出需求；预计 2026 年 10 月做 2027 年度预算时敲定各家 NPO 份额。
天孚通信不是英伟达 NPO 的直接模块供应商，拿不到终端整机订单，但具备完整 NPO 研发制造能力，CPO 技术积累可复用，参与路径两条：
1. 核心器件 / 组件供应商：向旭创、新易盛输出 FAU、ELS 模组、光路耦合组件；FAU 技术行业领先，ELS 模组已在自有 CPO 产品验证成熟。
2. 代工服务商：微软若落地 NPO 方案可能选择外包代工；Mellanox（英伟达旗下）拿到 NPO 订单后，受自身产能约束，可能把订单外包天孚或 Fabrinet。光模块业务中天孚占据 Mellanox 三分之二代工份额；若该合作复制到 NPO，天孚可承接对应代工产能。
小结：无论组件供货还是代工，天孚通信都将充分受益 NPO 量产。
Q3：随着 CPO 在 2027 2028 年逐步放量，天孚通信在其中的市场地位和份额将如何演变？
天孚通信定位为 CPO 核心组件与解决方案提供商。
第一代英伟达 CPO 产品，天孚是无源光纤阵列组件、有源 ELS 光源模组独家供应商。第二代 CPO 落地后供应链放开，引入第二家供应商，但天孚依靠先发技术积累，依旧维持核心供应商地位。

Q4：请详细分析天孚在第一代和第二代 CPO 产品中的市场份额、核心组件供应情况以及整个供应链的演变格局？
1. 第一代 CPO（2026 2027 上半年）
出货规模：数千套级别；天孚份额 70% 80%；
供货部件：ELS 模组、FAU、OE 组件，光路耦合工序由天孚完成；

供应链特征：相对封闭；参与方包含 Mellanox、台湾上诠、波若威、台积电；
三安光电早期参与，后转向第二代产品研发。
2. 第二代可插拔 DFAU CPO（2027 年 6 7 月起切换）
格局变化：原有天孚主导光路耦合模式发生改变；天孚已布局 DFAU 研发，
国内厂商第一梯队，样品完成，正在做制程与良率优化。
供应链更加开放，五大主要参与者：
①天孚通信 + Mellanox 组合；②三安光电（金属头 DFAU）；③Coherent（热
熔工艺 DFAU，技术难度更高）；④康宁（自研玻璃桥技术预计 2028 量产，同
步国内寻找代工厂）；⑤台湾上诠、波若威。
市场判断：供应链多元，单一厂商无法垄断。
Q5：目前主要 CSP 厂商如谷歌和亚马逊对 NVIDIA 推广的 CPO
技术的采购意愿和接受程度如何？
谷歌、亚马逊对英伟达 CPO 态度偏谨慎，没有主动采购，现阶段只是接收英伟达提供
样品开展测试验证。
两大核心原因：
1. 第一代 CPO 产品成本、运维成本居高不下；
2. 采用英伟达 CPO 会造成从芯片、交换机整套系统深度绑定，CSP 厂商希望规避
该风险。
因此谷歌等选择自研 OCS，同时采购 NPO，保障供应链多元化，掌握技术主导权。
Q6：谷歌 OCS 技术的快速推广，是否对 NVIDIA 的 CPO 市场策
略构成了影响？二者未来的竞争格局可能会如何演变？
谷歌 OCS 已经显著影响英伟达 CPO 产品路线。英伟达原本 CPO 主打柜间

（Scale Out）场景，自 2026 年 1 月，受 OCS 柜内竞争压力，英伟达、微软推动
供应商研发重心转向柜内（Scale Up）场景。
2027 上半年 CPO 以柜间应用为主；2027 下半年 CPO 重心切换柜内。
技术对比：CPO 综合成本高于 OCS，路线各有所长。OCS 仅适合柜内场景；CPO 同
时覆盖柜内、柜外，柜外场景优势突出。
最终行业格局预判：柜内以 OCS 为主，柜外采用 CPO，两条路线制衡共存。
Q7：在当前市场环境下，NVIDIA 是否会为了抢占市场份额而同时
推进 NPO 和 CPO 两种技术方案？
是的。英伟达同时推进两套技术：一方面大力推广 CPO；另一方面落地 NPO，以此
快速抢占市场份额。
Q8：业内如何定位 NPO (Near Packaged Optics) 的地位，它是否
因 CPO (Co Packaged Optics) 短期内无法量产而成为一种过渡
性方案？
行业将 NPO 明确定义为过渡方案，填补传统可插拔光模块与下一代 CPO 之间技术
空白。
传统可插拔光模块在 1.6T 遇到物理瓶颈；CPO 面向 100T 级别超高带宽；
NPO 主流规格聚焦 3.2T、6.4T，12.8T 占比较少；3.2T 由 4 颗 800G 或 2 颗
1.6T 光引擎集成，6.4T 采用 4 颗 1.6T 光引擎。
通过规格卡位，NPO 与可插拔模块、CPO 实现场景区隔，完成技术代际之间的补充。
Q9：既然 NPO 被视为过渡方案，为何像旭创等公司仍积极投入并
开发第二代产品，这是否意味着 NPO 具备长期前景？同时，天孚
通信在这类过渡性技术上的投入是否存在回收风险？
1. 对于中际旭创、新易盛、华工科技等模块厂商：CPO 生态由英伟达主导，产业链
相对封闭，国内模块厂商很难深度参与；布局 NPO、XPU 方案是维持行业竞争力的必
经选择。
虽然 CPO 是远期技术方向，但 CPO 良率、成本挑战大，商业化进度具备不确定性。

NPO 对比传统可插拔模块省去 DSP 器件，拥有成本优势，过渡期时间可能长于市场预期，二者是共存而非完全替代。
2. 天孚通信参与 NPO 风险很低，研发投入不会浪费。CPO、NPO 底层光器件技术高度相通，只是产品精度、复杂度差异。同样 FAU、MT 插芯、光引擎组件，分别服务于 CPO/NPO 不同产品。天孚既可以向旭创等销售器件，也有很大概率承接 Mellanox NPO 代工订单。
Q10：DFAU 的技术护城河体现在哪些环节，其产能扩张的主要制约因素是什么？
DFAU 在传统 FAU 基础上新增对接 PIC 光子芯片的专用接口。
核心技术护城河：将原本微透镜模组被动光学对准，集成至 DFAU 接口 FAU 耦合设计；PIC 芯片对接 DFAU 接口即可完成光路耦合。局部部件故障仅报废单个组件，不会造成整个光引擎整体失效，降低报废成本。
产能扩张制约：高度依赖高精度专用封装设备，同时需要大量熟练工艺人员；无法单纯依靠人工完成。
Q11：天孚通信在玻璃加点胶的 DFAU 技术路线上目前进展如何？面对多种技术路线并存的局面，NVIDIA 如何确立接口标准？
天孚通信玻璃加点胶路线 DFAU 已经出样品，处于送样验证阶段；计划 2027 年 6 7 月实现量产，国内处于第一梯队。对比：光库科技、太辰光、仕佳光子尚处于早期样品研发；亨通光电配合客户开发；致尚科技具备代工潜力。
英伟达主导 DFAU 接口标准制定，强制保证 DFAU 接口与英伟达 PIC 芯片精准匹配。外壳可以是金属头、玻璃头等多种技术路线，但是接口参数、内部对接规格统一，多技术路线并行；玻璃桥方案一旦成熟，有概率颠覆现有格局。
Q12：CPO 量产的主要瓶颈是否在于台积电的封装良率？目前进展如何，以及业界是否有其他替代方案？
CPO 量产瓶颈不只台积电 CoWoS 封装，光学路径良率不足、整体成本过高同样是关键阻碍。
行业正在发展封装替代方案：
长电科技给中际旭创 CPO 预研项目提供光引擎，反馈较好；
日月光旗下环旭电子、矽品都在开展相关研发；矽品已经和天孚通信开展测试层面合作。
Q13：NVIDIA 扶持天孚苏州工厂的具体动因、进展以及各方支持情况是怎样的？
建设动因
1. 企业自身：第二代 CPO 切换为 DFAU 产品，原有光路耦合业务会受冲击；天孚必须切入 DFAU 与上游封装环节对冲业务缺口。
2. 客户诉求：英伟达供应链安全诉求，需要打造台积电之外第二供应来源；天孚光通信技术积累满足条件。2024 年底 2025 年初天孚访问英伟达总部后项目正式启动。
3. 地方政策：苏州政府大力支持，用地审批快速落地。同时天孚与晶方科技开展研发合作，引入晶方前研发副总主导项目。
项目进度
苏州新厂房 2026 年 3 4 月动工；预计 2026 年底主体完工；2027 年设备进场；计划 2027 年 6 7 月产出，匹配第二代 DFAU 产品时间节点。
Q14：天孚通信在 1.6T 光引擎模块方面采用何种技术方案？其与 CPO 光引擎有何区别，以及该业务的供应链模式是怎样的？
天孚 1.6T 光引擎全部采用 EML 技术方案。
传统模块用光引擎 vs CPO 内部光引擎：属于完全不同产品。传统模块光引擎做二次封装，稳定性可靠性更强，适配长距离传输、高温高湿恶劣工况。

供应链模式：代工模式。英伟达向代工厂采购光引擎；Marvell 将订单委托天孚生产，专利归属 Marvell，供应链不对外。上游 200G EML 芯片资源由 Marvell 统一协调供给，天孚在此业务没有主导权。

Q15：1.6T 光引擎模块的 EML 技术方案未来是否会被硅光方案替代？以及当前业务的主要制约因素是什么？

硅光集成度性能持续迭代，远期存在替代 EML 可能性，但 2026 2027 年不会发生，观察窗口落在 2028 年。

当前业务最大制约：上游 200G EML 芯片供给，天孚出货量取决于 Marvell 分配芯片资源。

Q16：为实现目标，是否意味着关键物料供应已得到保障？

物料与产能匹配已经基本锁定，EML 芯片采购提前锁定。侧面证据：天孚通信 2026 存货规模相比往年显著提升。

历史背景：天孚过去作为无源器件厂商几乎不储备电子芯片；2025 年 9 月、2026 上半年两次遭遇芯片短缺导致产线闲置，员工临时调配至无源产线，无源业务因此阶段性走强，公司吸取教训主动增加芯片库存。

Q17：当前 EML 芯片的供应问题是否已经解决？DSP 芯片的供应情况以及行业对未来短缺的看法如何？

1. EML 芯片：供应问题基本解决。Lumentum 供给恢复；2025Marvell 紧急协调第一批货源，产能紧张，第二批补货直到 2026 年 5 月到位。

2. DSP 芯片：天孚拓展模块代工业务后新增 DSP 采购需求，DSP 依旧由 Marvell 协调。行业判断 DSP 紧缺短期延续，2027 年仍可能紧张。

缓解逻辑：硅光渗透率提升会用 CW 光源 + 硅光芯片替代 EML+DSP 组合；传统芯片厂商扩产意愿弱，部分产能转向 CW 光源。只有硅光模块市占率达到 50% 60%，DSP 紧缺才能实质性缓解。

Q18：天孚通信的保偏光纤主要采购来源是哪些？未来在 CPO 等新应用上是否有新的供应商计划？

现有两家供应商：康宁、长飞。

CPO 产品保偏光纤采用康宁；

对海思供货的产品使用长飞；早期海思也使用康宁，应客户国产化要求导入长飞；面向海思的采购量很小，一卷光纤可使用数月。

针对 CPO 增量需求，康宁产能有限，但是国内目前暂无其他备选供应商；CPO 标准由英伟达主导，材料选择权在英伟达；NPO 只要参数验证通过就可以导入新供应商，但是暂无明确导入计划。

Q19：天孚通信各生产基地的分工、扩产节奏以及产能规划是怎样的？

基地布局：江西、泰国两大生产基地；苏州研发运营中心；深圳研发副中心。

产能目标：国内、泰国产能各占一半，对冲贸易风险。

产品分工：高人工、高精度无源器件国内生产；之后转运泰国完成光引擎、模块组装；设备驱动、精度要求较低工序放在泰国；有源模块国内泰国双线布局。

扩产策略：无固定产能上限，完全跟随客户订单弹性扩产；预算边执行边动态调整。泰国光引擎设备分多批次采购，依据拿到的客户份额动态增加设备投入。

Q20：公司主要产品的利润率水平如何？未来随着产品结构变化，整体毛利率趋势将怎样演变？

产品线毛利率差异明显：

1. 陶瓷套管：毛利率最高；外销单价约 1 元 / 根，国内约 0.1 元 / 根，年营收规模 1 2 亿元，自用加外销。

2. OSA 产品：2025 年营收超过 12 亿元。

3. 模块代工业务：多数订单毛利率维持 40% 以上；Fabrinet、Marvell 长期合作订单利润较好；Finisar、南通代工订单体量小，毛利率达不到该水平。
4. CPO 业务：2026Q2 已经出货，尚属于研发项目，毛利率水平较高。
未来展望：2027 模块业务营收占比提升，会带来整体毛利率小幅下行，但公司坚持差异化路线规避价格战，整体降幅有限。
Q21：FAU 和 DFAU 产品的毛利率是否存在差异？
FAU 毛利率高于 DFAU。
FAU 高毛利原因：售价高、良率优异；微槽盖板、MT 插芯自主生产，仅外购带状光纤，自制率高。
DFAU 毛利率相对更低：需要外购微透镜模组、MT 插芯、接口部件；公司仅有微槽盖板实现自制；同时摊销高额研发投入。Marvell 作为长期合作研发伙伴，会保障天孚获得合理利润。公司也在推进 DFAU 接口部件自研。

Q22：美国 FCC 近期出台的规则对行业和公司业务的实际影响如何？
实际影响有限。新规核心新增第三方检测环节，只会拉长生产周期，增加检测费用，不会构成实质性壁垒。美国市场 80% 以上光模块依赖中国供应链，完全限制会造成本土供应链断裂。市场普遍预期政策最后仅增加检测流程，不会带来实质阻碍；现有出货产品不受影响。美国本土建厂、进一步限制中国企业的潜在政策尚未落地，影响无法评估。

--- 先进制程 & 封装调研：18AP14AE 良率产能、EMIBT 扩产、客户导入代工订单.pdf ---
先进制程 & 封装调研：18A P/14A E 良率产能、
EMIB T 扩产、客户导入代工订单
聚焦英特尔、台积电、NVIDIA、Google、Amazon、Microsoft、Broadcom、Apple、Qualcomm、AMD、SpaceX 等
1 18A P / 14A E 良率及产能爬坡
1.1 工艺定位与技术对标
目前英特尔在 3 纳米和 2 纳米尺度上的主要先进工艺节点是 18A P 和 14A E，二者为 18A、14A 的后续增强版本，差异显著，可视为两种不同工艺。
18A P：性能等效台积电 3 纳米，凭借背面供电技术具备小幅优势，定位 “3 纳米加” 级别。
14A E：对标台积电 2 纳米平台，部分综合性能接近台积电 A16，定位 “2 纳米加” 级别。
从前道制程来看，英特尔与台积电差距已非常小，基本持平。英特尔在晶体管密度、SRAM 性能具备传统优势；台积电侧重消费类场景，依靠后道封装拉高高性能计算密度。
当前英特尔前道制程良率正快速靠近商业量产盈利水平；全光罩尺寸大芯片良率仍在爬升，预计 2026 年底满足英伟达等新客户要求。14A E 进展与 18A P 深度联动，18A P 成熟后，14A E 可快速切换至量产。
结合后道封装能力，英特尔综合制造实力较强；英特尔比台积电早六七年布局硅桥、大尺寸面板集成，EMIB、Foveros 高端封装技术积累深厚。过去封装优势用于弥补前道短板；当前前道取得突破，前后道协同优势释放，具备直接与台积电竞争的实力。
1.2 14A E 量产时间规划
1. 内部产品：14A E 已实现小规模内部 CPU 产品线量产，样片已经产出，但良率仍需提升。
2. 18A P 对外代工：商业化大规模量产定于 2026 年第四季度；英特尔自用 18A P 至强 6 + 处理器（Clearwater Forest）2026 年 6 月已经量产。

3. 14A E 对外代工：量产时间由原计划 2028 年提前至 2027 年；2027 年下半年启动量产规模爬坡，同期 18A P 进入更大规模商业量产。

加速驱动因素
①英特尔先进后道封装吸引大量外部合作客户；
②美国半导体高端制造回流政策将英特尔作为核心平台，推动美国头部 Tier1 客户规划转产。
谷歌、亚马逊、微软均有 14A E 转产计划，同时仍保留台积电高优先级项目，但规划了特定产品落地；其中博通牵头项目进度最快。
Terafab 计划 2027 年底投产，Q3 Q4 风险试产，技术基于英特尔 14A E 授权，倒逼 14A E 工艺加速成熟。
1.3 14A E 对内 CPU、对外代工时间节奏
对外代工量产：预计 2027 年年中。
对内产品：已在 CPU 小规模量产，样片产出。
客户导入逻辑：18A P 已导入苹果、英伟达、亚马逊、高通；14A E 潜在客户等待英特尔自有产品完成内部验证，验证完成后客户导入进程将提速。
生产布局：14A E 并非仅在 Terafab 生产，英特尔自有产线同步布局。英特尔将 7nm/5nm/4nm 成熟工艺向爱尔兰工厂转移，释放美国本土产能给先进工艺节点。
1.4 英特尔整体产能现状及未来规划
理论总产能上限：每月 200000 片晶圆。
当前实际产能：每月 35000 片，以 18A P 为主；采用柔性产能策略，受良率爬坡、客户数量约束并未开满上限；3nm/4nm/7nm 成熟制程维持原有产能，保障自有 CPU 供给。
2027 年底总产能目标：每月 100000 片。
2028 2029：在美国制造回流政策驱动下，总产能有望突破每月 200000 片理论上限。
35K WSPM 月产能是在约 90% 稼动率条件下实现；稼动率进一步提升还有产能空间，但需要付出其他方面代价。18A P 当前稼动率接近 90%，接近满产；14A E

现阶段属于实验线，稼动率偏低。
1.4.1 35000 片月产能结构
产能结构：约 30% 供给英特尔自用，70% 分配苹果、英伟达代工订单。
苹果 M 系列 die 尺寸接近 400mm²，英伟达硅 die 接近 700mm²，大尺寸芯片消耗晶圆数量高；英特尔自有 CPU 计算裸片面积约 70mm²，单片晶圆可切割上千颗，晶圆消耗少，不挤压自有产品出货。
14A E 当前仅几千片实验产能，无明确固定分配；总产能不是零和博弈，订单增长可以带动总产能扩张。预计 2027Q1 末总产能接近 55000 片；新增产能几乎全部供给代工业务，自有 CPU 产能占比不再提升。英特尔 Foundry 业务独立核算，增量产能归属代工板块，代工业务利润预计 2027 年快速增长。
1.4.2 14A E 产能分配：代工 vs 自有 CPU
14A E 产能规划向代工业务倾斜，重点面向 ASIC、AI 高性能计算芯片，该类芯片 die 尺寸大，晶圆消耗高。
英特尔自有 CPU 对 14A E 依赖度偏低，依靠混合封装策略：
CPU 性能核 P core 使用 14A E；能效核 E core 未来使用 18A P；不同工艺裸片依靠 EMIB 进行混合封装。自有 CPU 裸片面积小，仅消耗少量产能。
2027Q1 末：14A E 代工产能占比仍较低，产能释放取决于博通、谷歌项目进度；谷歌 TPU v10 需要先进后道认证，博通项目存在一定延迟。
2028 年：14A E 产能占比快速攀升至接近 40%，2nm 节点成为主流，产品大量迁移；2029 年维持高位占比。
1.5 良率情况：14A E 中等尺寸芯片
当前 14A E 中等尺寸芯片良率约 55% 60%，预计 2027 年提升至 85%。英特尔策略优先保障中等尺寸 die 生产，大芯片依靠 EMIB T 后段封装堆叠实现，而非单颗

超大 die。
全光罩尺寸芯片主要为英伟达设计思路；英特尔历史上偏好小 die + 封装拼接，规

避大尺寸良率压力。
1.6 EUV 设备配置、保有量、订单交付
一条先进产线标准配置为一线两机（2 台 EUV 光刻机），支撑月产 30K 35K 晶圆。
1. 18A P：采用 Low NA（标准数值孔径）EUV，可复用现有光刻机产能，放弃高 NA 作为主力。
2. 14A E：主力同样 Low NA EUV；仅少数关键光掩模层使用 High NA（高数值孔径）；High NA 设备跨产线共享，不绑定单一产线。扩产主力依靠 Low NA EUV。
Intel7：产能逐步缩减，资源向先进节点转移；产线分布亚利桑那、以色列。
Intel4：产能维持不变，主要服务 Meteor Lake 等 CPU；产线亚利桑那、俄勒冈。
Intel3：需求旺盛，多用于 I/O die、存储类中低端产品；产能小幅微增，主要增量来自爱尔兰工厂升级；俄勒冈 D1X 未来逐步转向 18A/14A E。
稼动率：Intel7/4/3 稼动率维持 85% 90%；良率提升空间有限。2027 成熟制程仅 Intel3 存在增量，Intel7、Intel4 无扩产计划。成熟制程收入变化平缓，可以参考历史数据外推。
2 EMIB T 扩产及大尺寸封装优势
2.1 EMIB T 产能规划

主要供给英特尔自有至强 6 系列
当前后道先进封装有效产能：每月 25K 片晶圆（剔除良率，KGD 可交付口径），
。
扩产目标：2027 年 Q3，EMIB T 月产能达到 100K 片晶圆及以上。
英特尔 EUV 存量：合计约 30 台，其中仅 1 台 High NA，现阶段多用于 18A P 特殊层，稼动率不高，偏向实验用途。
后续订单：2 台 High NA，10 15 台 Low NA；大部分设备 2027 交付，部分 High NA 延迟至 2028 交付。
2027 年后 14A E 部分子工艺会使用 High NA，优先考量成本，仅解决关键层曝光。
1.7 Intel7 / Intel4 / Intel3 成熟制程产能规划

传统 EMIB（不带 T，无源硅桥）：现有月产能 70K 晶圆，用于英特尔历代 CPU，成本已经摊销；部分老产线升级改造转向 EMIB T。
EMIB T 与传统 EMIB 差异：EMIB T 硅桥内部增加 TSV 硅通孔，实现垂直供电，支持更灵活 3D 集成；硅桥工艺从 Intel7 演进至 Intel4，未来引入 Intel3。
2.2 良率、价格与毛利
1. EMIB T 良率约 95%；底层 EMIB 技术已有接近十年历史，量产 7 年以上，已经度过爬坡，2027 年预计维持≥95% 良率。
2. 定价逻辑：不按单颗芯片计价，依据光罩尺寸、订单量确定。
9 倍光罩及以上产品，EMIB T 封装每片晶圆增加成本约 2600 美元；对比台积电 CoWoS>3000 美元，上限仅支持 5.5 倍光罩。
TPU v9 项目观测后道封装附加成本约 2800 美元 / 片晶圆。
叠加 3D 集成技术，封装价格上涨约 50%，良率需要按项目评估。
3. 毛利：EMIB 前期设备折旧已经被自有 CPU 业务摊销完毕，增量成本主要来自玻璃基板以及康宁授权，整体后道封装业务毛利率较高。外部大规模商业化尚未完全展开，暂时没有独立拆分盈利数据。
除 EMIB T 外，英特尔布局 Foveros3D（原 Foveros 3D）3D 堆叠集成，3D 键合进度快于台积电，是吸引谷歌 TPU v10 的关键因素

2.3 客户、订单与大尺寸封装优势
台积电 CoWoS 最大仅支持 5.5 倍光罩，美国头部 Tier1 客户存在更大尺寸封装诉求；英特尔 EMIB T 支持 9.5 10 倍以上光罩尺寸，承接北美 ASIC 增量订单。
重点项目：
1. Google TPU v9：前道台积电，后道封装交由英特尔；芯片接近 10 倍光罩。
2. NVIDIA Rubin 系列。
3. AMD 高端 GPU，签署后道合作备忘录。
4. 微软新项目确定选用英特尔封装方案。
2026 2027 年底，谷歌、英伟达、亚马逊、微软等预期总订单量不低于 500 万颗；具体拆分预计 2027Q1 进一步明朗。
。

2.4 Terafab 项目商业模式
Terafab 由 SpaceX 主导，2027 年底投产，Q3 Q4 风险试产。
英特尔输出前道 + 后道全套工艺 IP，按芯片颗数收取授权费，不采用简单收入分成模式；产线柔性集成改造创新由 SpaceX 主导；该项目收入难以直接量化。
3 客户导入及代工订单
3.1 18A P 主要客户与项目、时间线
1. 苹果：M 系列芯片，内部小批量生产；预计 2027 年分配不少于 20% 的 18A P 产能，2028 维持约 20% 占比，为第一大单一客户。
2. NVIDIA：Rubin GPU（18A P 版本）；2026 年 10 月 Rubin 开始分阶段出货；2027Q1 推出 Rubin Ultra 四芯合封产品，突破台积电 CoWoS 尺寸上限；同时准备 CPU 备份方案规避台积电风险。英伟达 Feynman 芯片现阶段以台积电 A16 为主，14A E 尚处于早期准备。
3. 亚马逊：Trainium3 改版评估转产 18A P，应对台积电产能紧张涨价；Trainium3、Trainium4 属于并行项目非迭代替代。Trainium3 主打推理和微调；改版目标更大面板封装；2026 年 11 月前后会有项目进展。
4. 高通：AI250/AI300 系列计划转产 18A P，面向美国州级数据中心业务。
5. 英特尔自用：至强 6 + 等 CPU，可灵活调配产能。
出货节奏：苹果最先小批量；英伟达初期规模有限，2028 产量明显提升；2027 苹果出货量相对更大。
3.2 Google TPU 系列项目细节
1. TPU v9：计算 die 采用台积电 2nm；评估英特尔 18A 工艺生产 I/O die。英特尔 18A P 背面供电对高速 SerDes 有优势。
2. TPU v10：实验性项目，大量 3D 垂直堆叠键合，存储多层堆叠；die 尺寸约 400mm²，中等尺寸，非全光罩。台积电 3D 键合进度延迟约一年，谷歌将部分业务转向英特尔，依赖 Foveros 3D 能力；技术突破节点预计 2027 年年中。
3.3 14A E 潜在客户

谷歌、微软、亚马逊均规划特定产品，博通项目推进最快；全部需要等待英特尔内部 CPU 完成工艺验证之后启动导入。
3.4 英特尔 CPU 自产 + 外包策略及竞争格局
1. 生产分工：高性能计算 CPU 保留内部生产，消费类 CPU 产品线计划整体转向台积电代工；从 2026 起未来五年英特尔消费 CPU 将大规模使用台积电 3nm/2nm/A16，英特尔也成为台积电前五大客户。
2. 混合异构策略：单颗 CPU 内部混合不同工艺；P core 台积电先进工艺，E core Intel3，I/O die Intel7。消费级 GPU 放弃自研，封装英伟达 RTX 核。
3. 盈利展望：预计 2027 起 CPU 业务回到高毛利区间；虽然份额下滑，但市场缺货，依靠产能抢夺市场。

4. CPU 市场竞争（vs AMD）：AMD 台积电总产能没有扩张，被迫将中低端产能让渡高端，部分老型号停产；英特尔依靠自有产线灵活性复产老至强，抢占中低端市场。英特尔 CPU 最大买家是白牌渠道，囤货行为导致终端实际销量难以精确跟踪，出货与终端消费数据存在偏差。

3.5 Agentic AI 驱动 CPU 需求逻辑

Agentic AI 是 CPU 需求爆发的核心驱动力：GPU 负责算力，CPU 充当 Agent 执行主体，模拟人操作软件工具。

- 算力换算：1 个等效 H100 约可以处理 10 个人的工作量 FTE；一个 FTE 业务负载约等效 8 个 CPU 核心；预计 2030 单 H100 可支撑 30FTE。
- 需求预测：保守预计 2029 2030CPU 总核心数相比当前至少翻一番。即使 GPU 算力 2028 年趋于饱和，执行层 CPU 算力仍然紧缺。
- 架构演进：自至强 6 + 开始，英特尔不再单纯依赖传统超线程，转向超多核，每个物理核心提供独立硬件资源的 “真线程”，适配 AI 沙箱隔离资源需求。
- Astra 等具备 computer use 的长任务大模型进一步抬升 CPU 占用，长任务挂起持续占用 CPU 线程。
- 行业推演：到 2030 年 AI 预计可替代全球约 15% 人类工作量，新增算力当中 CPU 占比很高。

--- 探针卡调研：HBMAI 探针卡市场格局及竞争优势，Socket 国产替代及国内厂商竞争，CPO 硅光探针卡测试方案及耗材增量.pdf ---

探针卡调研：HBM/AI 探针卡市场格局及竞争优势，
Socket 国产替代及国内厂商竞争，CPO 硅光探针卡
测试方案及耗材增量

聚焦 Technoprobe、MJC、Advantest、Teradyne、东京精密、东京电子、长川科技、韬盛科技、Aehr、Keysight、广立微、联讯仪器、强一股份、ficonTEC 等

1 HBM / AI 探针卡市场格局及竞争优势

1.1 完整芯片测试链路及环节

一颗芯片完整测试流程分为五道，部分产品不需要老化、系统级测试：

1. WAT（晶圆切割道参数测试 / Parametric Test）：晶圆厂内部完成；监测制程缺陷、工艺稳定性；测试门级漏电流，在晶圆切割道取样，避免整批报废；全晶圆 100% 覆盖率，采用米字形选点抽样，监控光刻、离子注入等工艺区域均匀性。
2. 中测 Wafer Probing（CP）：晶圆制造完成后，探针卡接触晶圆裸 die 做电学测试；每个客户芯片设计不同，测试方案定制化；AI、GPU 高速信号对设备耗材要求高。
3. 成品测试 Final Test（FT）：晶圆切割封装后；使用分选机替代探针台。
4. 老化测试 Burn in：汽车电子、工业芯片重点环节；高温烘烤模拟生命周期，筛选早期失效器件。
5. 系统级测试 SLT System Level Test：在接近真实整机系统环境验证芯片，不侧重单纯电气参数，验证整机场景表现。

1.2 各环节核心设备、厂商、耗材

1.2.1 WAT 测试

- 测试设备：是德科技（原安捷伦 P9000 系列）；国内：广立微、联讯仪器、概伦电子。

Advantest、Teradyne 不做 WAT：WAT 偏向精密仪表，需要飞安级漏电流检测；

市场规模相比逻辑 ATE 更小，与传统逻辑测试技术路线差异大。

1.2.2 中测 CP 环节

- 测试机：Advantest 爱德万、Teradyne 泰瑞达，合计占据全球逻辑 / 存储测试机 80% 90% 份额；国内长川科技、华峰测控。
- 探针台：东京精密 TSK、东京电子 TEL、韩国 SEMICS；国内矽电半导体。

探针卡 Probe Card（核心耗材）
全球主要厂商：FormFactor（龙头，存储 + SoC 全覆盖，SK 海力士 HBM 大量采用）、欧洲 Technoprobe、日本 MJC。
价格：普通悬臂针卡数万元人民币；AI/HBM 高端探针卡可达数十万美元。
寿命计量单位：touch down（探针接触芯片次数）。
消费类 MPU：200 300 万次。
大电流高敏感芯片：50 100 万次（针尖氧化，需要频繁打磨）。
报废原因：反复打磨探针变短，不是断裂。
AI 大 GPU 场景：单张 12 英寸晶圆 die 数量仅有一两百颗；探针卡寿命 50 100 万次往往超过一代 AI 芯片产品生命周期；实际生产客户会多卡并行保障产能，不等耗材耗尽。
Advantest、Teradyne 壁垒：自研测试板卡、专用 ASIC 硬件 IP；软件调度；整套测试方案、 高精度可重复稳定输出； 高端HBM 测试机单台 300 500 万美元以上。
1.2.3 成品测试 FT 环节
分选机：Cohu（收购 Delta Design、Ismeca）、ASMPT、爱普生；台湾鸿劲科技（转塔式全球龙头）；国内金海通、长川科技（Pick&Place;）。
关键耗材：Load Board 测试 PCB 载板、Test Socket 测试插座。
Socket 全球厂商：Smiths Interconnect、IDI、LEENO 韩国、NHK 日本、台湾 WINWAY。
WAT 设备价格：老型号十几万 20 万美元；P9000 新型号 30 40 万美元。
WAT 探针卡：针数少，漏电流指标 < 1nA，价格低；寿命可达 1000 万次 touch down。

国内 Socket 参与者：
a. 和林微纳：垂直整合，自研弹簧针 + Socket 成品；Socket 属于公司高毛利板块；已经进入英伟达供应链，但英伟达营收占比有限，没有显著拉高整体毛利。
b. 韬盛科技、法特迪：弹簧针主要外部采购。
c. 捷策创等中小型厂商。
1.3 老化测试 Burn in
目的：高温加速应力，筛除早期失效，提升汽车、工业芯片可靠性。
核心设备厂商：Aehr Test Systems（行业龙头），Advantest、新加坡 AEM；国内矽佳半导体等自研。
耗材：老化炉、老化板 Burn in Board、老化插座 Burn in Socket。
Aehr 业务：碳化硅、AI/GPU 大芯片、硅光、HBM 晶圆级老化；设备可压缩老化时长，单台 AI 大芯片老化设备售价 200 300 万美元；正在推进 HBM/HBF 晶圆级老化项目。
1.4 系统级测试 SLT
不需要传统 ATE；搭建接近最终产品的硬件原型平台；芯片通过插座接入模拟整机，运行操作系统与业务，验证系统层面表现。例如手机 AP 芯片 SLT，直接模拟手机整机环境。
2 Socket 国产替代及国内厂商竞争
1. 市场规模对比：探针卡全球约 30 多亿美元；Socket 整体市场规模小于探针卡。
2. 强一股份：营收高度集中于华为；风险点：①非华为客户担心产能被挤压；②华为也会供应链分散；传闻布局 Socket 业务，具体落地形式尚不明确。
3. 和林微纳：Socket 业务毛利高；同时布局 MEMS 探针卡研发；面临韬盛科技、法特迪竞争。
当前行业现状：测试组装设备产能满负荷；交期紧张，新订单没有锁定产能则排到 2027Q1 Q2，至少半年交期，客户普遍提前锁产能。
3 CPO 硅光探针卡测试方案及耗材增量

3.1 光芯片 / PIC/CPO 光引擎测试和传统电测试差异
传统芯片依靠 ATE 做电学信号；光芯片需要完成光电耦合转换，测试光信号输出。
主流两套量产方案：
1. 泰瑞达 + ficonTEC（罗博特科收购）联合方案，向台积电提供演示。
2. Advantest（93K 测试机）+FormFactor 探针台；使用硅光集成探针卡：探针卡上既有传统金属探针处理电信号，又集成光纤阵列对准光口；调试好耦合位置后，整片晶圆一次性完成光电同步测试，提升效率。
光学检测模块集成在爱德万测试机内部。
3.2 CPO 带来耗材增量
CPO 进入大规模量产后，硅光探针卡将成为持续性消耗品；光纤组件会老化衰减，需要定期更换，带来持续耗材需求。

--- 磷化铟调研：衬底产能、价格与长协，多晶料瓶颈及 6 英寸量产进展，良率、温度场与自动化控制等.pdf ---
磷化铟调研：衬底产能、价格与长协，多晶料瓶颈及 6 英寸量产进展，良率、温度场与自动化控制等
聚焦云南鑫耀 / 广东先导 / 住友 / JX 等，2026 09 18
要点
磷化铟衬底产能、价格与长协机制
多晶料 / 高纯原料供给、锁料与扩产节奏
2/3/4/6 英寸产品结构、6 英寸量产与综合良率
云南鑫耀 / 广东先导 / 住友 / JX 竞争格局及长晶自动化
以下是专家观点：
一、产能规划与尺寸分布
1.2 英寸磷化铟衬底产能规划
2024 年：30 40 万片
2025 年：60 80 万片
2026 年规划产能：120 160 万片
截至 2026 年第二季度末，年化产能约130 万片；预计 2026 年年底，年化产能能够达到160 万片。数据基于 Yole 研究以及企业股东电话会承诺：自 2024 年起产能每年翻一番。海外上市公司实际产能通常会比公开数字高出20% 25%。
企业自2025 年 6 7 月开始逐步进入满产，全球主要相关企业大致同一时间段陆续满产。当前产出产品全部售出；受出口许可证严格限制，国内市场需求旺盛，优先保障大客户，小客户基本停止供货。预计相关出口企业 2026 年二、三、四季度财报表现向好。
2027 年产能规划目标320 万片，实际产出存在进一步上行空间。
2.6 英寸产品进展
Lumentum 和 Coherent 已获得英伟达 20 亿美元预付款，用于锁定供应链及材料。
Coherent 因 6 英寸晶圆紧缺，积极推进设备研发合作，直接投资上游。

6 英寸处于全球量产前夕，率先实现稳定量产供货的厂商将成为 Coherent 核心供应商，经济效益显著高于 3 英寸产品。
二、产能扩张瓶颈：磷化铟多晶料紧缺
从当前 160 万片扩张至 2027 年 320 万片，扩产核心约束不是设备，而是合格磷化铟多晶料。设备供给充足，但国内扩产炉子数量激增，多晶料需求爆发。
1. 高纯磷短缺：相对短期，生产门槛较低；部分产品尚未完成认证，提纯难点在于脱除杂质砷。
2. 高纯铟、配比严格的磷化铟多晶料短缺具备长期性。
多晶料短缺的行业解决方案与周期
1. 签订上游长期协议，直接投资上游原料企业扩产；国内已有企业布局多晶料产能建

设。
2. 磷提纯在化工园区完成，磷化铟合成可在非化工园区。
条件顺利：新增多晶料产能最快 6 个月逐步释放
审批、园区复杂场景：需要1 1.5 年
3. 供给缓解节奏取决于高纯磷、高纯铟、合成环节实际产出。
光模块用磷化铟多晶料纯度要求为6N，需要掺杂硫，对其余杂质耐受度相对更高。
原料市场格局
磷化铟多晶料总供应量不超过60 吨。云南鑫耀、广东先导等头部厂商大规模锁料后，市场流通原料极少，剩余流通料质量难以保障。
企业锁料的挑战：部分企业锁料起步晚，上游产能已经被头部客户提前锁定；外资企业普遍锁料时间偏晚。资金充足前提下可以通过商业代价完成产能锁定，障碍并非不可逾越。部分企业具备磷提纯 高纯铟 磷化铟合成全产业链，具备自主扩产条件。
国内多晶料厂商普遍采用水平法：投资低、工艺简单，但晶体均匀性差，易出现富磷 / 富铟。可增加重结晶工序，把头尾料重新生长，改善均匀性；仅增加电费、石英耗材，成本增幅有限。
三、价格走势、长协机制与毛利率

1. 原材料与衬底价格
高纯铟价格上涨约 2 倍以上；高纯磷明显涨价。
磷化铟衬底行业整体价格较 2025 年初上涨 1 1.5 倍。
大客户长协：涨幅约 1 倍
散单小客户：涨幅可达 1.5 倍
行业现状：不再对残次片进行补偿；小客户即使有资金也难以拿货；云南锗业已于 2026 06 01 发布涨价通知。
原材料在总成本中占比较低，价格上涨绝大部分转化为利润，行业毛利率有望进一步提升。
2. 二季报收入和价格涨幅不匹配的原因
1. 已产出产品尚未完成收入确认；
2. 部分外资企业经营策略保守，不会一次性释放全部业绩弹性。
3. 后市价格预判
1. 2026 下半年：长协价格趋于稳定，散单价格继续上行。中小研发项目不计成本拿货支撑散单行情。
2. 2027 年长协：无法锁定固定价格，仅锁定供货优先级；供货时按照当期市场价执行。客户预付获取优先位次，按谈判后的当期市场价结算；顺位靠后的客户只有在前序客户拒绝报价时才轮到谈判。
3. 2027 上半年价格大概率保持平稳，下半年随需求动态调整。
2027 上半年价格维持稳定原因： 多数企业2026 新增产能尚未完成客户认证， 无法有效释放；如广东先导拥有 450 台炉子，但认证未完成，产能无法释放。企业不会过度提价，避免倒逼下游自研，损害长期业务。
2026 年长协：年初涨价 30%，多次调价后累计涨幅 100% 150%。
四、产品尺寸结构与未来扩产方向

1. 当前主力为3 英寸晶圆，良率最高。
4 英寸生长不良降级产出 3 英寸；3 英寸不良降级产出 2 英寸；2 英寸属于生产副产物，不专门投料生长。
2. 未来扩产焦点是 6 英寸晶圆。云南鑫耀等头部企业公开表示 6 英寸进入工业化量产准备。谁率先稳定供货，将取得巨大竞争优势。
3. 4 英寸不会成为主流，良率偏低；部分场景 6 英寸生长难度甚至低于 4 英寸。
五、良率现状

行业公开 60% 80% 良率大多是单工序良率，从投料到出货的全流程综合良率明显更低；如果综合良率达到该水平，会和当前供给紧张的市场现状矛盾。
晶体生长成败高度依赖炉房、温度场 Know how：理论仿真正确的炉体结构不一定长出晶体；更换场地、调整设备参数，原有成熟工艺也可能失效，属于行业核心工艺诀窍。
六、海外厂商格局：住友、JX
1. 住友：2024 年占全球产能约 60%，但产能利用率仅 30%，主要供给内部光器件、射频半绝缘衬底；提升利用率即可实现产能翻倍。衬底电性能弱于国内厂商。2026 供货周期拉长，一部分产能转向射频类半绝缘产品，挤压光模块用衬底供给。很多数据不对外公开。
2. JX：2024 产能约 5 万片，2025 年达到 10 12 万片，实现产能翻倍。
七、晶体炉、自动化技术对行业变革
炉子、温度场细节是晶体生长成败关键。利用晶体相变熔点不变的物理原理，可以完成温度探头校准，消除测量误差，在此基础上搭建自动化控制系统，摆脱高度依赖人工。
行业对标参考：单晶硅领域实现一人看管数百台炉子，效率百倍提升；磷化铟未来也会发生类似变革。普通 PLC 硬件已经可以满足控制，核心是工艺 Know how 与控制系统结合；突破后可提升单晶良率，大幅降低人工。
设备投入：单台炉子约 35 万元，配套设备 15 万，炉房工程 10 万，原材料 15 万；试产 4.5 英寸晶体单套投入约 75 万元。
产出周期：人员方案到位最快 4 个月产出产品；常规 6 个月；完整报批流程需要 1 1.5 年。
行业动态：2026 年行业会议已有 4 家以上自动化厂商参会，包含硅生长温控企业。

八、砷化镓衬底业务现状
1. 行业整体经营承压：受出口管制、低端市场竞争冲击。4 英寸砷化镓价格跌到 70 多元，部分企业生产成本高达 150 元。
2. 盈利关键：良率。半绝缘砷化镓利润率最高，但国内综合合格率低，多数企业达不到成本线。
3. 创新合作模式：衬底厂商进入下游客户工厂内驻场生产，深度绑定，减少运输损耗。
4. 2025 年部分企业毛利率仅 20% 左右的原因：
 当年产品市场价格未上涨；
 出口订单履约受阻，大额订单无法交付；
 异地建厂带来历史包袱拖累当期收益。
注：以上为专家调研观点，仅供参考，不构成投资建议。