Agent 时代下半场，推理半导体利好谁

Agent 时代下半场，推理半导体利好谁 —— 算力瓶颈从训练 GPU 向推理 CPU、先进封装与材料层迁移

AI 正从训练主导转入推理与智能体时代，算力的瓶颈结构随之重排——CPU 与 GPU 的配比从训练时代的约 1:8 趋向 1:1，封装瓶颈从单一 CoWoS 向 EMIB 硅桥多路线分流，材料层（玻璃基板、磷化铟、人造金刚石、GaN/SiC）成为下一个突破与锁货的战场。沿"推理放量"这条主线，受益纯度最高的是先进封装代工与设备、InP 衬底、外置激光源、玻璃基板量产先行者与金刚石散热；CPU 复兴的最大赢家是 AMD 与 Arm 阵营，Intel 是叙事主角但执行风险大、弹性给中。

涌现资本产业研究部

2026 年 06 月

> **涌现资本产业研究部 · 2026 年 6 月 19 日** · 机构中性

英特尔 CEO 陈立武近期密集发声（2026-02 思科 AI 论坛、2026-06 Computex 主题演讲），核心判断是：**AI 正从训练主导转入推理与智能体时代，算力的瓶颈结构随之重排**——CPU 重回 AI 堆栈的编排中心，先进封装与材料层成为新的物理卡点。我们认同这一方向，并在其基础上结合独立调研，给出涌现资本对"下半场推理半导体"的受益判断与标的落地。

执行摘要：我们的五个核心判断

1. **CPU 推理复兴**：配比从训练时代的约 1:8 趋向推理时代的 1:1，CPU 从摩尔末期的存量替换市场，重新变回跟随 AI 资本开支的增量市场——**AMD 与 Arm 阵营最受益**，server CPU TAM 已被 AMD 官方在半年内上修一倍。
2. **封装瓶颈分流**：先进封装是这一轮少数仍处结构性供不应求的真卡点；瓶颈正从单一 TSMC CoWoS 向 **EMIB 硅桥**多路线演化，EMIB 在推理/ASIC 开辟低成本增量赛道——**承接 CoWoS 外溢的 OSAT（Amkor）与混合键合设备（BESI/ASMPT）弹性最大**。
3. **玻璃基板是终局材料、但放量在 2028 后**：技术优势确凿，真卡点在"高良率 + 碎裂可靠性 + 客户认证"——卡位可提前，兑现需等待，当前最接近商业化的是 Absolics(SKC)/AMD 一线。
4. **CPO 是渗透而非替代**：2026-2028 与可插拔（含 LPO）、铜缆长期共存；价值从"DSP+组装"上移到**硅光引擎、外置激光源、精密光连接**三段不可替代卡点。
5. **材料层高度分化**：**InP（磷化铟）是唯一的真卡点单品**（寡头 + 缺口 >70%）；**人造金刚石迎来散热量产元年**（我们据此补入 universe）；GaN/SiC 虽紧张但夹带 EV 周期、合格供应商多、定价权偏弱。

总纲性结论：下半场的钱流向三处——推理把 CPU 拉回增量市场、封装瓶颈从 CoWoS 向 EMIB 分流、材料层的 InP 与金刚石成为真卡点。**受益纯度最高的不是终端芯片，而是封装代工与设备、衬底与光源、散热材料这些"卖铲子里更上游的一段"。**

一、CPU 推理复兴（约 1:8 → 1:1）

产业逻辑：GPU 只负责单次推理调用；而智能体工作流是"推理→工具调用→取数→评估→再推理"的串行循环，**每两次 GPU 调用之间的编排、tool-calling、RAG 检索、内存管理、安全/网络控制路径全压在 CPU 上**。量化痛点：工具处理可占端到端延迟的 50%~90%；大 batch 下 CPU 动态能耗可达系统动态能耗的 44%（聊天时代 CPU 只承担 5%~10% 算力）。硬量纲：传统 AI 数据中心约 30M CPU 核/GW，智能体时代预计跃升至约 120M 核/GW（约 4 倍）。

> **口径澄清**：这是"CPU 插槽/核数"量级的跃升，**不是"CPU 美元价值超过 GPU"**。GPU 加速机架里 GPU+HBM 仍主导整机美元含量。CPU 复兴的 α 在"出货量 + ASP 双升 + 独立 CPU 机架增量"，而非

> 在 GPU 机架里抢美元份额。

瓶颈已在价格上兑现：server CPU 自 2026 年 3 月起涨 10%~20%，下半年预期再涨 8%~10%；Intel 交期 8~22 周（亚洲达 6~8 个月）、AMD EPYC 部分产品交期 >30 周且 2026 基本售罄；UBS 测算 AI CPU ASP 2025→2030 累计 +56%。

标的	卡位	弹性	要点
AMD `AMD`	server CPU 双寡头·唯一同握 CPU+GPU	**高**	2026-05 苏姿丰把 server CPU TAM 从 $60B/18% 上修一倍至 **2030 >$120B/>35% CAGR**；Q1'26 DC 收入 $5.8B、server CPU 同比 +50%、Q2 指引 +70%；下代 EPYC Verano 定位"纯为 AI 而生"
Arm `ARM`	每瓦性能·云端自研 CPU 收版税	**中**	Neoverse 累计部署 >10 亿核；AWS Graviton5 / 微软 Cobalt / 谷歌 Axion / Ampere 全用 Arm；目标 2030 占 server CPU 出货 40~45%
NVIDIA `NVDA`	亲自下场做"agent 的 CPU"	中（互联）	Vera CPU（88 核·NVLink-C2C 1.8TB/s）定位"the CPU for Agents"，坐实 CPU↔GPU 互联带宽是新卡点
Intel `INTC`	x86 server 主场·叙事主角	**中（执行风险）**	智能体把战场拉回 Intel 仍主导的 x86；DCAI Q1'26 收入 $5.1B(+22%)；18A 的 Xeon 6+/Xeon 7 落地。但兑现取决于 18A 良率/产能与份额防守——**下行保护变厚，不等于份额反转**
Micron / SK 海力士 `MU`	CPU 侧内存（DDR5/MRDIMM/CXL）	高（**强搭车**）	MRDIMM 8800 MT/s（较 DDR5 +39%）。但 2026 内存暴涨主因是 AI 总需求 + HBM 挤占 DDR5 晶圆，归因不能全记在 CPU 复兴账上

二、先进封装瓶颈迁移（EMIB 硅桥 vs CoWoS）

先进封装是这一轮 AI 算力扩张里**真瓶颈纯度最高**的一段（同时满足供需失衡、18–24 月认证替代难、卖方定价权、AI 主驱动四特征）。瓶颈正从单一 CoWoS 向"按场景分流"演化：

- **CoWoS-L（硅中介层 + LSI 桥）守训练 GPU**：最大带宽、超低延迟，NVDA/AMD 旗舰首选；
- **EMIB / EMIB-T（嵌入式硅桥）卡推理/ASIC**：不用大中介层、把硅桥嵌进基板，桥 die 晶圆利用率约 90%（大中介层约 60%），低成本、高良率、大封装（reticle 目标 2026 做到 8×、2028 ⩾12×）。**这正对应"下半场=推理放量"：推理对极致带宽不那么敏感、对每瓦每美元算力敏感，EMIB 的低成本结构与之匹配**（已出现联发科 AI ASIC 同时采用 Intel EMIB + TSMC CoWoS 的双路线策略）。

产能即印证：TSMC CoWoS 月产能 2023 末约 13k → 2025 约 75k → 2026 底目标 120–130k（CAGR >50%），即便如此仍"满订"至 2028；TSMC 主动把 2026 约 24–27 万片/年 CoWoS 外溢给 OSAT（Amkor 约 18–19 万、SPIL 约 6–8 万）。EMIB-T 在桥内加 TSV，打开 HBM4 级供电；Intel 已将先进封装独立为专门业务、由前 SK 海力士 CEO 执掌并直接向 CEO 汇报。

标的	卡位	弹性	要点
Amkor `AMKR`	承接 CoWoS 外溢最纯 OSAT	**高**	JPM 点名为"TSMC 日益依赖的外部 CoWoS-S 伙伴"，2026 承接约 18–19 万片
BESI `BESIY` / ASMPT `0522.HK`	混合键合（hybrid bonding）设备	**高**	3D 堆叠/SoIC 不可替代环节，混合键合机市场 2025→2032 CAGR ~21%
TSMC `TSM`	CoWoS/SoIC 绝对龙头	低（基数稀释）	真卡点但已充分定价、属基础层
Intel `INTC`	EMIB/Foveros 原创方 + Foundry 第二供给	中（兑现待验证）	EMIB-T 是技术差异化，CFO 称接近签"每年数十亿美元"封装大单——属订单叙事，需以实际收入验证
AMAT / Lam `AMAT` / `LRCX`	封装沉积/刻蚀/TSV 设备	中（**泛搭车**）	Lam 先进封装收入 2025 三倍增至 >$30 亿、2026 指引 +50%；但封装只占总盘一小部分，弹性被稀释
通富 / 长电 `002156` / `600584`	国产 2.5D/chiplet 替代	中（政策依赖）	承接被卡的中国 CoWoS 需求，确定性依赖制裁/补贴节奏，与全球真瓶颈逻辑不同，需单独评估

三、玻璃基板（替代有机·放量在 2028 后）

技术优势确凿：平整度 ≤1μm（有机 5–10μm）、CTE 3–8 ppm/°C 匹配硅（解决大面积多 chiplet 翘曲命门）、可做 <2μm RDL + 大方形面板批量化。**但这是典型"卡位已硬、兑现尚远"的远期瓶颈**：2026 仍在中试/送样，2028–2030 才放量。真卡点不在"做出一片玻璃"，而在"**高良率 + 碎裂可靠性 + 客户认证**"——Absolics 良率约 82% vs 成熟有机 >95%，这道关把绝大多数蹭概念标的挡在门外。

- **最接近商业化**：**Absolics（SKC `011790.KS`）/AMD 路线**——佐治亚厂向 AMD 供量产样并进认证、AWS 在测，产能 2026 扩至 8,000 面板/月；
- **玻璃材料双雄**：**Corning `GLW`**（玻璃芯基板约 25% 份额、与 TSMC 共研）、**AGC `5201.T`**（超低膨胀玻璃）；
- **韩系中试**：三星电机 `009150.KS`（世宗中试线、向 2–3 家美国大客户送样）；
- **A 股相对核心**：沃格光电 `603773`（掌握 TGV 全制程）、凯盛科技 `600552`（TGV 中试线规划 2026 量产）；
- **TGV 设备**（真卡点工艺但竞争充分）：海外 LPKF/Philoptics，国产德龙/大族/华工/海目星。

> **受损方**：ABF 载板五寡头 Ibiden `4062.T`(~35%)/Shinko/欣兴 `3037.TW`(~14%)/AT&S(~10%)/南電 `8046.TW`(~5%)。但我们判断短期不会崩塌——2025–2027 ABF 仍结构性紧缺（ASP 一年 +26%、交期峰值 28 周），龙头同时砸钱扩 ABF 并自研玻璃对冲。**短期紧缺、长期（2028 后）高端迁移——这是它们的双面性。**

四、CPO（渗透非替代·价值链重切）

我们的判断：**CPO 不是"光模块消失"，而是"光互联价值链重切"**——价值从"DSP+组装"向"硅光引擎 + 激光光源 + 精密光连接"上移。技术驱动真实（224G/lane 时代可插拔的功耗墙：Broadcom TH6 宣称比可插拔省电 >3.5 倍），但时间表被压缩：**scale-out 交换机侧 2026 先落地，GPU 侧 scale-up 因良率（32 引擎系统级良率示意≈19%）、激光老化、可维护性推迟到 2028-2029**。2026-2028 是可插拔（含 LPO）+ CPO + 铜缆**长期共存期**；Yole 口径 CPO 市场 2024 仅约 $0.46 亿 → 2030 约 $81 亿（CAGR≈137%，高增长但绝对盘子仍远小于可插拔）。

> 🟡 **反身性校准点**：SemiAnalysis 的良率质疑报告曾当日砸盘 AAOI −17%、LITE −8%。判断早期 CPO 标的"真卡点 vs 反身性透支"，这是核心标尺——**别在延迟预期未消化前追高位光器件题材股**。

环节	标的	弹性	要点
硅光引擎代工	TSMC COUPE `TSM`	低（最硬卖铲子）	COUPE（SoIC-X 叠层·5-10x 能效）2H2026 量产，是 NVDA/AVGO 共同底座
CPO 交换芯片	Broadcom `AVGO`	中（确定性高）	第三代 CPO 200G/lane，TH5-Bailly 2025 出货 5 万+台
外置激光源 ELS	Coherent `COHR` / Lumentum `LITE`	**高**	CPO 下激光必须外置成新卡点；NVIDIA 2026-03 对两家**各投 $20 亿**（合 $40 亿）+ 多年采购承诺
光模块龙头（转型中）	中际旭创 `300308` / 新易盛 `300502`	中（组装段泛搭车·硅光自研是真卡点候选）	LPO 中期窗口领先，但 224G+ 后组装环节有被压缩风险
精密光连接	光库 `300620` / 康宁 `GLW`	低-中	保偏 PM-FAU 单价从 ~$15 升至 ~$100，价值量跃升

五、材料卡点（InP 真卡点·金刚石量产元年·GaN/SiC 打折）

四种二三代材料**绝非同一档**，用瓶颈纯度透镜看分化巨大：

🥇 **InP（磷化铟）——唯一的真卡点单品。** 是 800G/1.6T 与 CPO 引擎里 CW 激光/EML 的衬底，CPO 把单系统 InP 用量数量级抬升。供需缺口 >70%；2-inch 价从约 $800（2025 初）涨到 $2,300-2,500（+200%），6-inch 涨到约 $5,000（+250%）；叠加中国 InP 出口管制（中国占铟产量约 70%）。

- **首选 AXT** `AXTI`（InP 衬底约 60-70% 份额寡头，纯度最高、弹性最大但单点 + 反身性风险大）；国产替代 云南锗业 `002428`（2025 出货 +74%、扩至 45 万片/年）。

- 🟡 反身性提示：2026-05/06 中国分批放行 InP 出口，价格出现边际松动——出口管制是涨价主因之一，松绑会削部分溢价。

🥈 **人造金刚石——2026 是散热量产元年（我们据此补入 universe）。** 热导率 >2200 W/m·K（约铜 5 倍），用作 heat spreader 与 GaN-on-Diamond 衬底。**催化剂硬**：NVDA 在 CES 2026 宣布 Vera Rubin 采用"金刚石-铜复合 TIM + 45℃温水直冷"，把金刚石从实验室推向 AI 散热刚需；Element Six×Orbray 2026-06 跑通 3-inch 单晶量产工艺。

- **新纳入 universe**：**黄河旋风** `600172`（国内可量产最大尺寸：6-8 inch 多晶散热片、热导 1000-2200 W/m·K、8-inch 产线 2026-02 投产 30 万片/年；推"金刚石+SiC"复合方案降热阻约 30%）、**力量钻石** `301071`（HPHT 成本较国际低约 40%、2026 拟增 400 台 MPCVD、散热片 50→100 万片/年）。
- ⚠️ 数据提示：券商口径"2030 市场 480-900 亿元""覆盖 15% NVDA 散热需求"为单源测算，量级未二次核验，按谨慎对待。

🥉 **GaN / SiC——AI 是增量但非唯一主驱动，定价权弱、瓶颈纯度打折。** 用于 800VDC 架构（首发 NVIDIA Kyber/Rubin Ultra 机柜·约 2027），但紧张同时由 EV 周期拉动；且合格供应商多、NVDA 800V 批准名单分散采购，削弱单一厂商定价权（与 InP 寡头结构相反）。

- 已覆盖较全：英诺赛科 `2577.HK`（GaN 份额约 29.9% 第一）、Navitas `NVTS`、士兰微、天岳先进 `688234`（SiC 衬底 2025 全球第一）；Wolfspeed `WOLF` 属高位反身性透支、回避。

六、受益地图：按真卡点纯度 × 弹性分层

层级	标的	主线
真卡点·高弹性（核心）	AXT(InP)·Amkor(CoWoS 外溢)·BESI/ASMPT(混合键合)·Coherent/Lumentum(激光源)·**黄河旋风/力量钻石(金刚石·新增)**	封装/材料/光源的稀缺产能
真卡点·高确定性（弹性被基数稀释）	TSMC(COUPE+CoWoS)·AVGO(CPO 交换)·AMD(server CPU TAM 翻倍)·Arm	平台级卖铲子
叙事主角·中弹性（执行风险）	Intel(CPU+EMIB+玻璃，18A 良率/份额防守是变量)	"下行保护变厚 ≠ 份额反转"
强搭车（归因需切割）	Micron/SK 海力士(内存)·AMAT/Lam(封装设备占比小)	弹性高但非纯卡点
受损 / 回避	Ibiden/欣兴/南電/AT&S(玻璃替代·2028 后)·传统可插拔光模块(CPO 渗透)·Wolfspeed(反身性)	—

七、结论

下半场推理半导体的物理卡点地图清晰：**推理把 CPU 从存量替换市场重新拉回跟随 AI 资本开支的增量市场；封装瓶颈从单一 CoWoS 向 EMIB 多路线分流；材料层的 InP 与金刚石成为真卡点。**

沿这张地图，钱会流向三处——① CPU 复兴中 **AMD/Arm 最受益**（server CPU TAM 半年翻倍是最硬的数字印证）；② 封装分流中承接溢出的 **OSAT（Amkor）与混合键合设备（BESI/ASMPT）弹性最大**；③ 材料层的 **InP（AXT）与金刚石（黄河旋风/力量钻石）是真卡点**，前者寡头紧缺、后者迎来 NVDA 背书的量产元年。

纪律提醒：玻璃基板、CPO 是"趋势成立但兑现在 2028 后"的远期瓶颈——卡位可以提前，但别把延迟预期未消化的高位题材股当便宜货；InP 的涨价里有出口管制的反身性成分，松绑是风险点。**Intel 是这套叙事的主角，但叙事主角 ≠ 自动赢家**——它的下行保护因推理放量而变厚，份额与利润率能否反转仍取决于 18A 执行，弹性给中、不给高。
