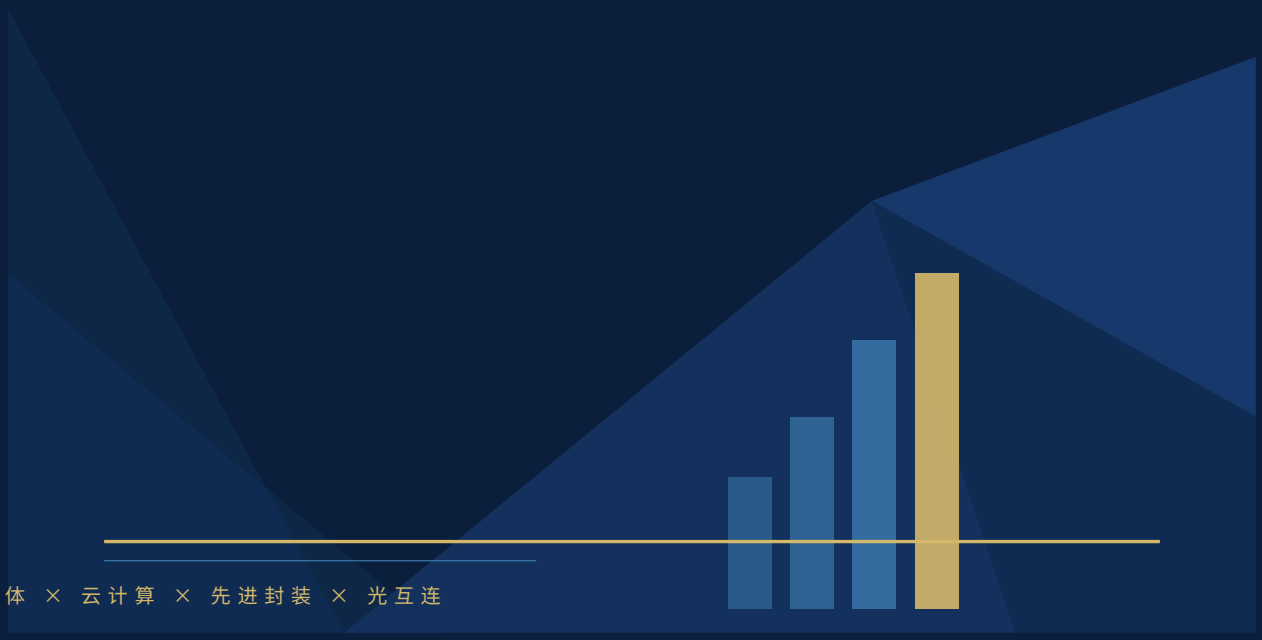


半导体 × 云计算 × 先进封装 × 光互连

定制硅反客为主

ASIC + 开放以太网，正侵蚀英伟达「通用 GPU + 私有互联」护城河

研究问题｜自研 ASIC + 开放以太网，是否正从边缘走向主战场，结构性侵蚀英伟达份额？拐点与受益「卖铲人」在哪里？

01 核心观点

KEY CALLS

一句话总纲：英伟达的护城河不是「GPU 不可替代」，而是**每片 GPU ~72% 的数据中心毛利**。当推理放量把工作负载标准化，超大厂用自研 ASIC 绕开这层毛利——K Research 测算自研硅在标准化推理上**单位有效算力 TCO 低约 41%**；英伟达须把数据中心毛利从 ~72% 砍到 **~38%** 才能在推理打平。份额迁移已经开始，**真正确定性的赢家是"卖铲人"博通与开放以太网阵营**。

01 护城河的本质是毛利、不是硅
英伟达 FY2026 数据中心营收 1,937 亿美元、毛利率 ~72–75%（一手 8-K）。ASIC 与 GPU 共用同样的 HBM、CoWoS、EUV——成本差的大头是英伟达的品牌溢价，而非工艺。

02 推理是 ASIC 的主场，训练仍属 GPU
推理已占 AI 算力约 2/3（2023 仅 1/3）。K Research 测算：推理场景 ASIC TCO 优势 41%、训练仅 28%——拐点先在推理侧发生。

03 份额迁移基准情景：ASIC 价值占比 2024 ~12%→2030 ~38%
TrendForce：2026 定制 ASIC 出货增速 44.6%，为商用 GPU（16.1%）近 3 倍；ASIC 型 AI 服务器占比升至 27.8%。

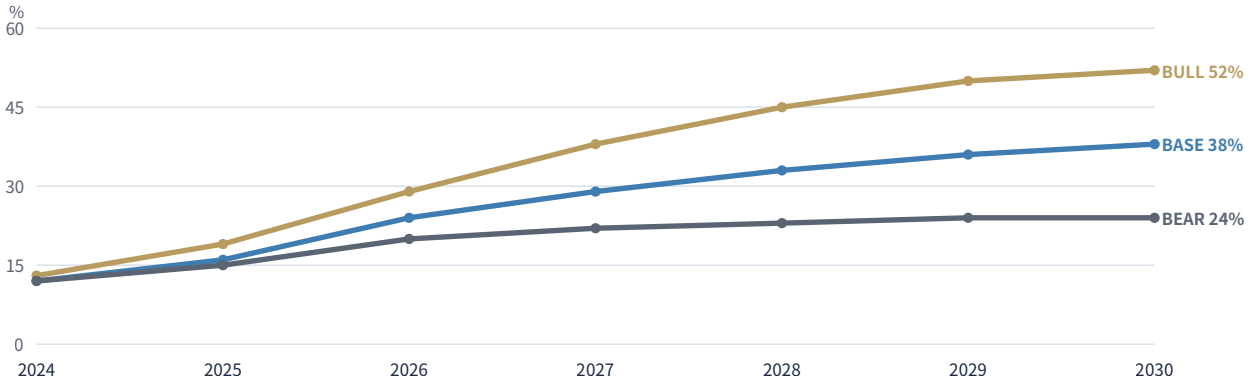
04 开放以太网已经赢下 scale-out
Dell'Oro：以太网 2025 年首超 InfiniBand，2026 Q1 占 AI 后端交换机销售约 2/3。英伟达自己也卖以太网（Spectrum-X 年化 >100 亿美元）——等于承认趋势。

05 最确定的赢家是博通
博通 Q2FY26 AI 半导体 108 亿美元（+143%），FY27 AI 营收指引 >1,000 亿，2027 年定制硅+网络 SAM 600–900 亿美元；与 Marvell 合计垄断 ~95% 定制 ASIC 协同设计。

06 CUDA 护城河在推理侧正被「兼容性」填平
vLLM 经 JAX-XLA 统一后端，PyTorch 模型零改动即可在 TPU 上跑且吞吐 +20%（vLLM 官方）。这是「英伟达无敌」共识最脆弱的一环。

07 先进封装仍是所有人的总瓶颈
NVIDIA 锁定 TSMC CoWoS ~60%、博通 ~15%；HBM4 由 SK 海力士锁定 Rubin 订单 60–70%。GPU 输份额，但 TSMC / SK 海力士「卖铲人通吃」。

主视觉｜ASIC 占 AI 加速器价值份额：三情景渗透曲线，base 2030 触及 ~38%



来源：K Research 自建测算，数据截至 2026年6月，E 为 K Research 预测；校准锚点 TrendForce 2026 ASIC 服务器占比 27.8%

BEAR 25%	BASE 50%	BULL 25%
CUDA+系统锁死，ASIC 困于自用推理	ASIC 吃下推理，GPU 守住训练	开放以太网+TPU 外售加速渗透

02 产业链全景

VALUE CHAIN

数据：AI 加速器产业链可拆为七层。英伟达的统治集中在 L2–L3（系统集成与加速器），靠 CUDA 与 NVLink 把"卖芯片"升级成"卖整机+卖网络"，FY2026 数据中心营收 1,937 亿美元、占总营收 90%（一手 8-K）。**机制：**越往上游，议价权越脱离英伟达——L4 网络已被博通以太网反超，L5 HBM 由 SK 海力士（62%）掌控，L6 CoWoS 近乎 TSMC 独家，L7 EUV 由 ASML 100% 垄断。

产业链全景｜自研 ASIC 改写 L2–L3 的"谁拿毛利"，上游瓶颈 GPU 与 ASIC 共用

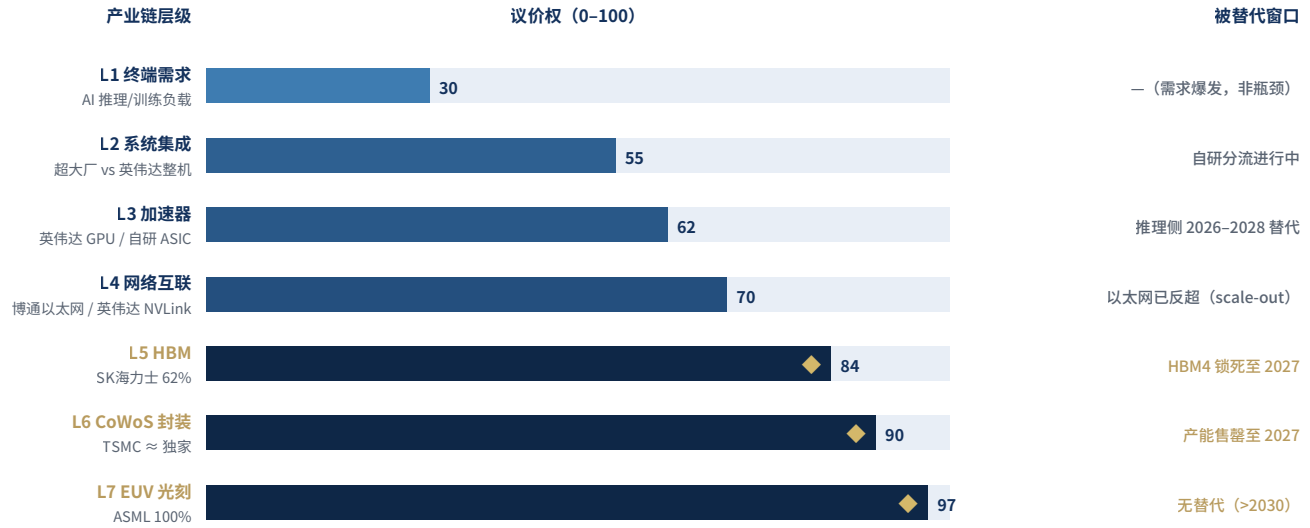
自研 ASIC 改变的是 L2–L3（谁设计、谁拿毛利），但 L4–L7 的瓶颈（HBM／CoWoS／EUV）由 GPU 与 ASIC 共用——卖铲人通吃

来源：公司公告、TrendForce、Morgan Stanley，K Research 整理，数据截至 2026年6月

7 层 choke-point 穿透：议价权与被替代窗口

结论：沿链做满七层穿透，一个反直觉的结论浮现——**份额从英伟达流走，不等于利润池缩小**。被替代的是 L3 的 GPU 溢价，而 L5/L6/L7（HBM、CoWoS、EUV）无论算力跑在 GPU 还是 ASIC 上都必经，议价权反而随总量增厚。**反驳与再结论：**有人认为英伟达可凭 NVLink 把 L2–L4 重新封闭，但 NVLink Fusion（2025/5）反而开放接口、并投资 Marvell 20 亿美元——这是防守而非进攻，等于默认 L4 开放化不可逆。

七层卡位｜议价权随层级加深而上升，◆ 为不可替代的总瓶颈



来源：ASML/SK海力士/TSMC/博通 公告与财报、TrendForce，K Research 整理（议价权为 K Research 评分）

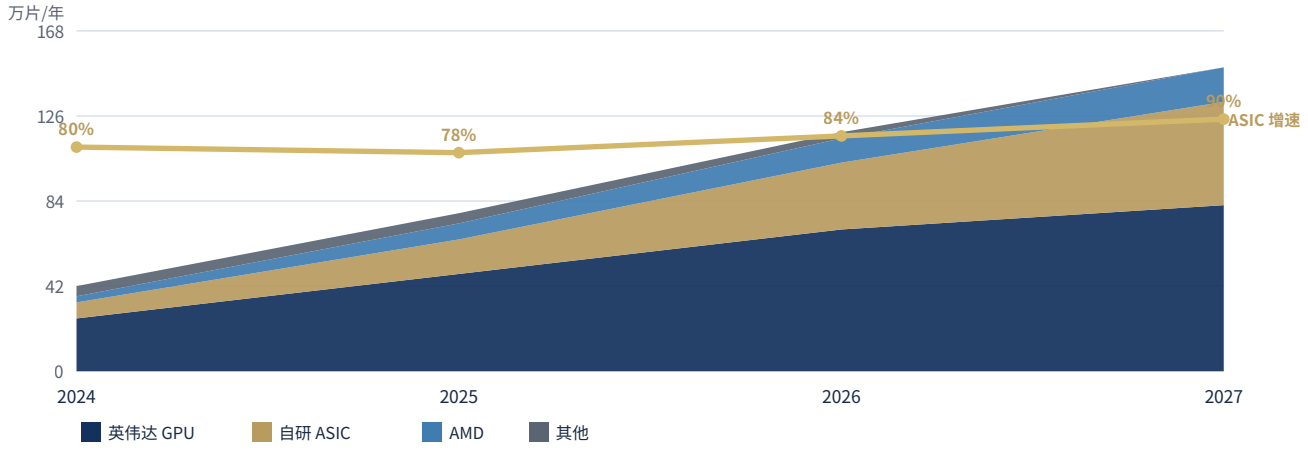
03 供需与价格

SUPPLY & PRICE

数据：先进封装是 AI 算力的真实总闸门。TSMC CoWoS 月产能从 2024 年底 ~3.5 万片爬到 2026 年底 11.5–13 万片（TrendForce），但仍"售罄至 2027"（TSMC 2026Q1 电话会，一手）。2026 年全球 CoWoS 需求约 100 万片，其中英伟达约 60%、博通（代工 TPU/Meta/OpenAI）约 15%、AMD 约 11%、Marvell 约 5.5%（Morgan Stanley）。**机制：**ASIC 对先进封装的需求增速远快于 GPU——MediaTek 为谷歌 TPU 一次性要求 CoWoS 产能增 7 倍（TrendForce）。

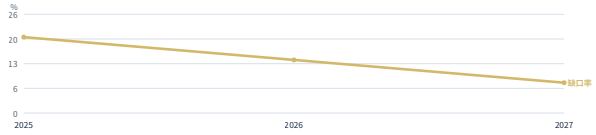
K Research 自建测算 · 先进封装供给堆叠与 ASIC 份额迁移

CoWoS 产能分配｜**ASIC 占比从 18% 升至 34%，蚕食的正是英伟达的封装配额**



来源：Morgan Stanley（2026 分配）、TrendForce（产能/增速），K Research 整理与测算，2026 年后为 E

供需缺口收窄｜**由 ~20% 向 ~10% 收敛**



来源：TrendForce 2026/6/15，E 为预测

瓶颈	掌控方	份额
CoWoS 封装	TSMC	~独家
HBM	SK 海力士	62%
HBM4 (Rubin)	SK 海力士	60–70%
EUV 光刻	ASML	100%

缺口收窄不利于"涨价"逻辑，但封装/HBM 总量随 ASIC+GPU 双扩张而增长——卖铲人受益于**量**，而非单纯的紧缺溢价。

反驳与再结论：若 TSMC 2027 扩产 >60% 兑现、缺口快速消失，先进封装的稀缺溢价会收敛；但 ASIC 渗透意味着**单位算力对封装的消耗强度上升**（更大 fabric、更多 HBM 堆叠，2029 目标单封装 24 颗 HBM），总需求曲线被抬升，对冲了缺口收窄。

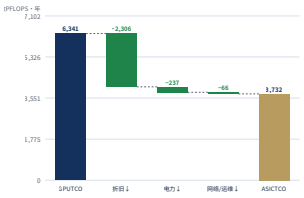
04 财务透视

UNIT ECONOMICS

这是本报告的认知地基。市场争论"ASIC 能否替代 GPU"，却很少把账算到单位有效算力。K Research 自建「单位有效算力年化 TCO 模型」，把 GPU 与 ASIC 拉到同一把尺子：每一有效 PFLOPS·年的总拥有成本（折旧+电力+网络/运维，按 3 年、PUE 1.15、电价 $0.08/kWh）。

K Research 自建测算 · 单位有效算力 TCO（推理场景，base）

TCO 桥｜ASIC 单位有效算力 TCO 比 GPU 低约 41%



来源：K Research 自建测算，数据截至 2026年6月，E 为 K Research 预测

口径	GPU	ASIC
取得成本（ASP/自有）	$28,000	$13,000
峰值算力 FP8 (PFLOPS)	5.0	4.6
有效利用 MFU	42%	40%
功耗 TDP (kW)	1.30	0.65
年化 TCO	$13,315	$6,867
TCO/有效 PFLOPS·年	**$6,341**	**$3,732**

TCO 临界点：英伟达须把数据中心毛利率从 ~72% 砍到 **~38%**（推理）／**~59%**（训练），GPU 才能与自研 ASIC 打平。当前毛利 75%，意味着定价端有巨大让利空间，但会直接击穿其估值核心。

敏感性矩阵｜ASIC 推理 TCO 优势 = f(英伟达毛利率, ASIC 有效 MFU)，中心金框为 base

来源：K Research 自建测算，数据截至 2026年6月，E 为 K Research 预测；负值（灰）表示该组合下 GPU 反而更省，正值（蓝）表示 ASIC 占优，单位 %

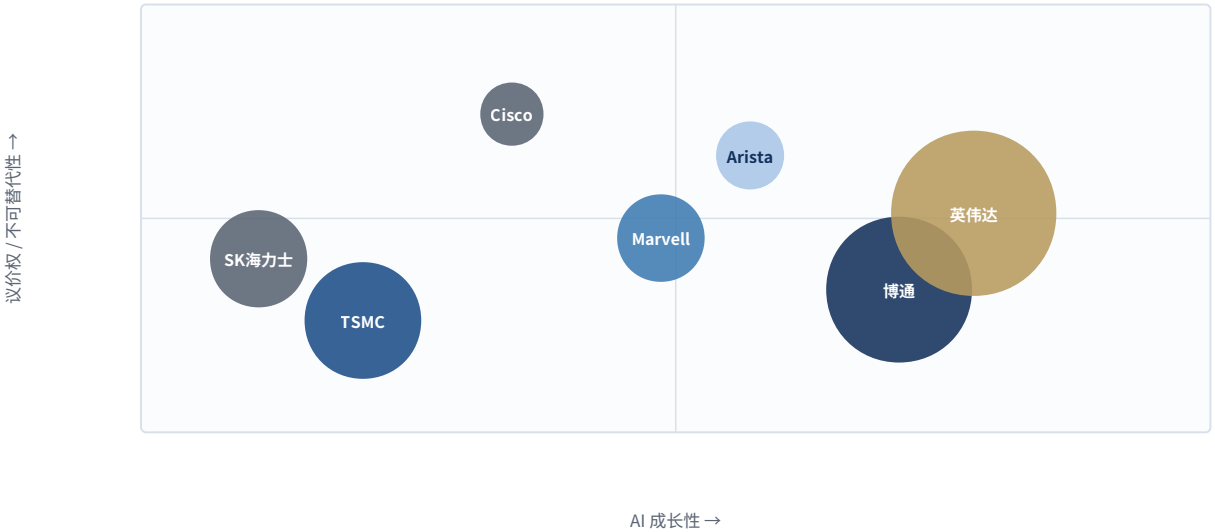
分歧（必须并列）：SemiAnalysis 大规模自用口径给出 TPU v7 每有效 FLOP TCO 较 GB300 低 20–50%，与本模型一致；但 Artificial Analysis 单机租赁口径下英伟达对 TPU v6e 有 ~5× tokens/$ 优势。差异源于规模（>10 万芯片）、自用 vs 租赁、是否计入 JAX 重写工程成本——这是份额迁移叙事的最大不确定性，已纳入 bear 情景。

05 竞争格局

COMPETITION

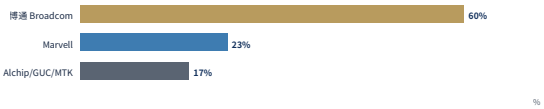
数据：定制 ASIC 的"军备"不在云厂手里，而在设计代工方。博通与 Marvell 合计垄断约 95% 的定制 AI ASIC 协同设计（机构估）：博通绑定谷歌、Meta、OpenAI、字节，Marvell 绑定亚马逊（Trainium）、微软（Maia）。**机制：**云厂自研 = 付给博通/Marvell 一层 ~50–65% 设计毛利，换掉英伟达 ~75% 整机毛利——每美元算力净省。

战略矩阵｜**议价权 × AI 成长性：博通独占右上，英伟达守成长但议价权被侵蚀**



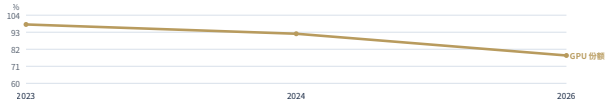
来源：公司财报与电话会，K Research 整理，数据截至 2026年6月；气泡面积≈相关 AI 营收

定制 ASIC 协同设计份额



来源：机构估（IndMoney/Hashrate Index），2026

英伟达数据中心 GPU 份额下滑



来源：TechInsights/Omdia 估，2026E

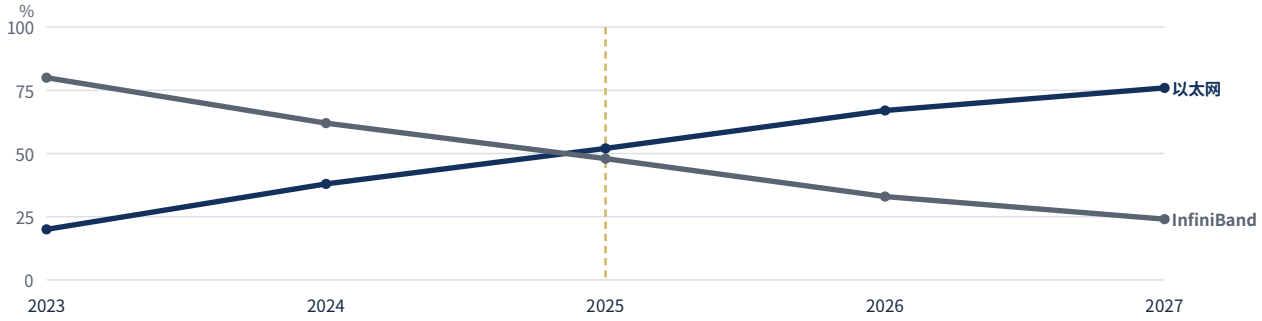
反驳与再结论：英伟达可凭 Blackwell/Rubin 的"通用可编程+迭代速度"守住训练与新架构（黄仁勋："没有 Anthropic，TPU 增长几乎不存在"）；但这恰好承认了 ASIC 已在标准化推理站稳——竞争焦点不是"谁全胜"，而是**推理这块占算力 2/3 的池子里 ASIC 能切多少**。

06 技术路线

TECH ROADMAP

数据：私有互联是英伟达护城河的另一半。但 Ultra Ethernet Consortium 规范 1.0 已于 2025/6 发布（100+ 成员含英伟达）；Dell'Oro 确认**以太网 2025 年首超 InfiniBand**，2026 Q1 占 AI 后端交换机销售约 2/3。**机制：**scale-out（横向扩展）天然适配以太网；英伟达自己的 Spectrum-X 以太网年化已 >100 亿美元、网络部门 FY26 营收 ~310 亿（+142%）——它在用行动承认以太网化。

开放以太网 vs InfiniBand｜2025 黄金交叉，2026 以太网拿下 2/3



来源：Dell'Oro Group（AI 后端交换机销售份额），2026；2027 为 E

网络与光互连里程碑｜拐点节奏

来源：UEC/Broadcom/NVIDIA 公告、SemiAnalysis，K Research 整理

交换芯片	容量	厂商
Tomahawk 6	102.4T	博通
Teralynx T100	102.4T	Marvell
Silicon One P200	51.2T	Cisco
Spectrum-X	—	英伟达

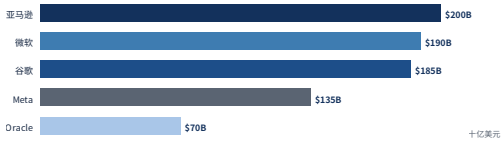
S 曲线定位：开放以太网在 scale-out 已过拐点（导入→放量）；scale-up（对抗 NVLink 的 UALink/SUE）仍在导入早期，拐点滞后至 2027；CPO 处导入期，2028–2030 才上规模——**光互连是下一个被重定价的环节**。

07 资金与持仓

CAPITAL & VALUE

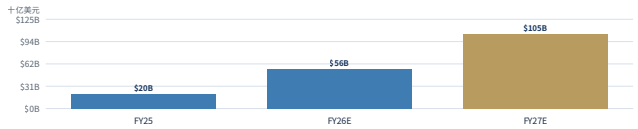
数据：需求侧的确定性来自超大厂资本开支。2026 年微软 ~1,900 亿、亚马逊 ~2,000 亿、谷歌 1,800–1,900 亿、Meta 1,250–1,450 亿、Oracle ~700 亿美元，四大合计 ~7,250 亿（同比 +64–77%，一手指引）。Barclays 测算西方 AI 基础设施年支出 2028 年或破 1 万亿美元，较共识高 3,000 亿+。**机制：**capex 的边际增量正越来越多流向定制硅与开放网络。

超大厂 2026 资本开支



来源：各公司 2026 季报指引，K Research 整理

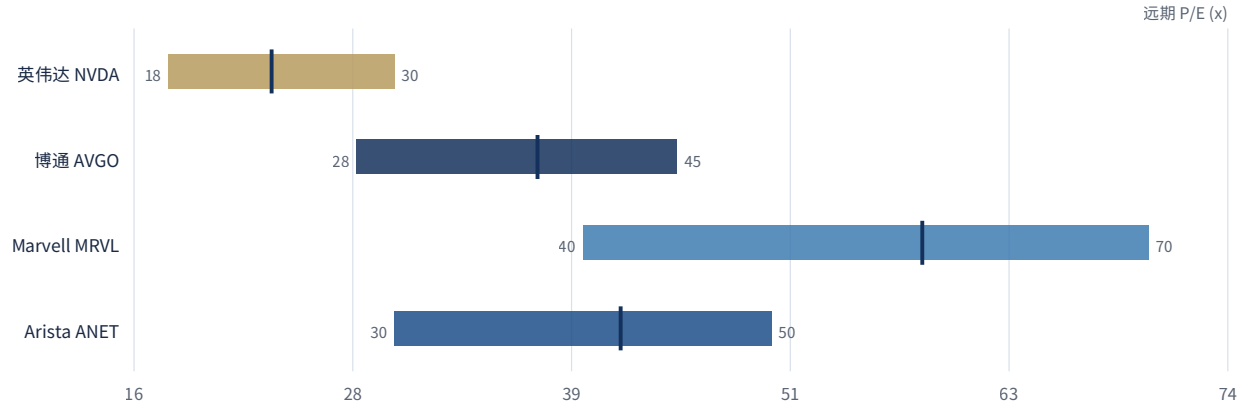
博通 AI 半导体营收轨迹



来源：博通 FY26Q2 电话会指引，>1000亿为管理层口径，E 为预测

K Research 自建测算 · 估值足球场（远期 P/E 区间）

估值足球场｜"卖铲人"已被给到成长溢价，英伟达反成估值洼地



来源：companiesmarketcap/GuruFocus，2026年6月；竖线为当前远期 P/E，区间为 bear–bull 合理区间（K Research）

反共识于估值：市场给博通 37.6×、Marvell 58× 远期 P/E，给英伟达仅 ~23.5×——已经在为"份额迁移"定价。**拥挤度风险：**博通+Marvell 在定制 ASIC 协同设计 ~95%、博通在高速以太网交换 ~80% 份额高度集中，一旦超大厂 capex 预期 reset，二线卖铲标的回撤弹性大于英伟达。

08 反共识深论

NON-CONSENSUS

> **市场共识（白纸黑字）：**"英伟达 = AI 唯一赢家，CUDA 护城河无解，自研 ASIC 只是超大厂的议价筹码、成不了气候。"
> **K Research 攻击其最脆弱假设：**"CUDA 在推理侧同样不可替代"——这一条，正在被"兼容性"而非"性能"击穿。

数据：2025/10 vLLM 官方（谷歌团队撰写）发布 TPU 统一后端，经 JAX-XLA 单一 lowering 路径，**PyTorch 模型代码零改动即可在 TPU 上运行，吞吐反而 +20%**；TPU 第八代原生支持 vLLM/SGLang/PyTorch/JAX。MLPerf 上 AMD MI300X≈H100、MI325X≈H200。**机制：**训练依赖手工 CUDA 核与数值稳定性，迁移成本高；推理是标准算子（attention/GEMM/采样），一旦开源推理栈把 ASIC 后端补齐，CUDA 的锁定力在推理侧迅速衰减——而推理占算力 2/3 且增速是训练 2 倍。

反共识支撑｜**推理是可俘获的池子：占比更大、ASIC TCO 优势更高**

来源：K Research 自建测算，数据截至 2026年6月，E 为 K Research 预测；推理占比来源 Introl/McKinsey/Gartner

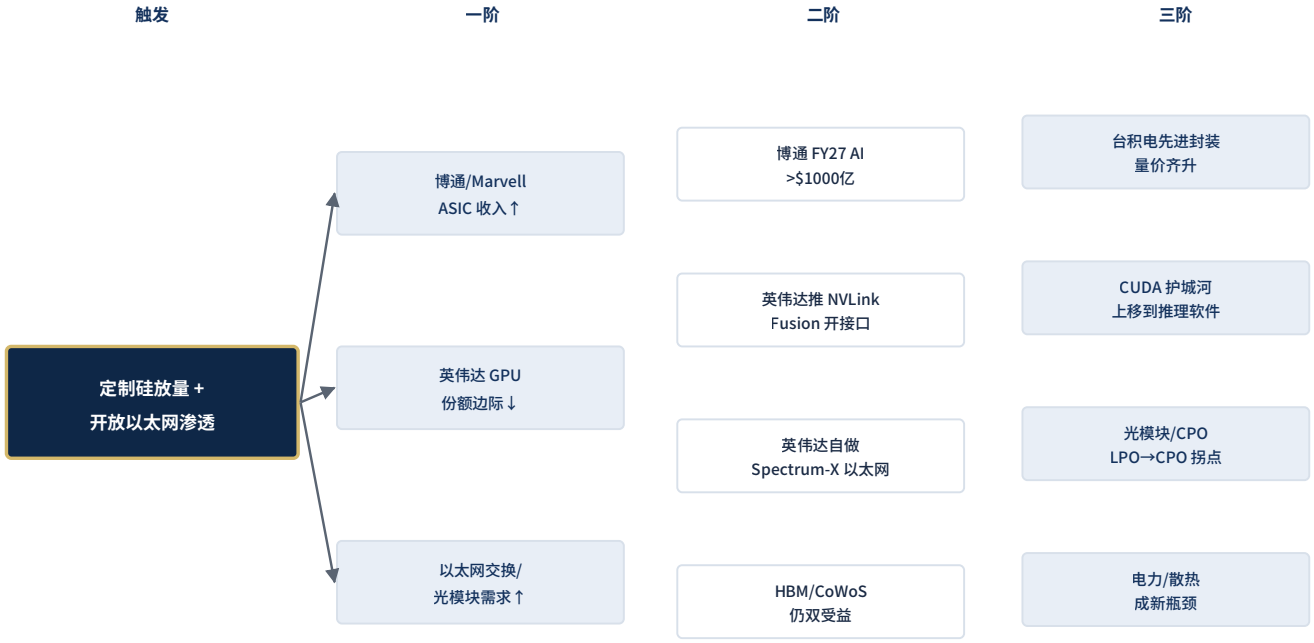
结论：护城河不会崩塌，但会"上移"——英伟达把竞争从"硬件+编译器"上移到推理编排软件（Dynamo/NIM）。**反驳（最强反方，黄仁勋原话）：**"没有 Anthropic，TPU 增长几乎不存在；ASIC 常因做不出比英伟达更好的 velocity 而被取消"；且 TPU 仅限谷歌云、深度依赖博通。**再结论：**这恰恰承认 ASIC 已是一笔需要"被点名解释"的生意；叠加谷歌已宣布向 Fluidstack 等**外部数据中心供货 TPU**（Pichai："deliver TPUs to a select group of customers in their own data centers"）、Meta 2027 起整机外购 TPU——ASIC 正从"自用议价"转为"对外商品"，这才是质变。

09 二阶与三阶效应

2ND/3RD ORDER

一阶是直观的：博通/Marvell 定制硅收入↑、英伟达 GPU 份额边际↓、以太网与光模块需求↑。真正的认知增量在二阶、三阶。

传导链｜**份额从英伟达流走，利润池却向上游与卖铲人重新分配**



来源：K Research 推演，基于公司公告与产业链数据，2026年6月

二阶 · 英伟达的"自我验证"：NVLink Fusion 开放接口、自做 Spectrum-X 以太网、投资 Marvell 20 亿美元——英伟达每一步防守都在确认对手叙事：纯封闭 GPU 已不足以锁定超大厂。CUDA 护城河被迫上移到推理软件层（Dynamo/NIM）。

二阶 · 卖铲人通吃：无论算力跑在 GPU 还是 ASIC，都要过 TSMC CoWoS、SK 海力士 HBM、ASML EUV。份额迁移抬高的是**总封装/HBM 消耗强度**（ASIC fabric 更大、HBM 堆叠更多），台积电与海力士两头受益。

三阶 · 光互连重定价：开放以太网放量→800G/1.6T 光模块需求 2026 翻倍→LPO 过渡→CPO 在 2027–2029 成为新卡位。Coherent、Lumentum、中际旭创、康宁从"配套"升级为"瓶颈"，获结构性份额。

三阶 · 电力成终极约束：算力效率（perf/watt）变成资本效率。ASIC 的能效优势→单位算力电耗下降，但总装机量爆发→电力与散热成为新的 L1 瓶颈，Barclays 把电力/许可证/劳动力列为万亿 capex 的最大证伪点。

10 情景与风险

SCENARIOS

情景	概率	2030 ASIC 价值份额	触发条件	受益主线
BEAR	25%	~24%	CUDA+系统化锁死；ASIC 困于自用推理；开放以太网渗透停滞	英伟达守住溢价
BASE	50%	~38%	ASIC 吃下推理、GPU 守训练；以太网拿下 scale-out	博通/Marvell/光互连
BULL	25%	~52%	TPU 外售加速+开放以太网 scale-up 突破+ASIC 部署占比跳升	卖铲人全面重定价

敏感性矩阵｜2030 ASIC 价值份额 = f(推理占算力比例, ASIC 单位 TCO 优势)，中心金框为 base

推理占 AI 算力比例 \ ASIC 单位 TCO 优势	10%	20%	30%	40%	50%	58%	66%
50%	7	13	22	30	37	40	42
58%	7	14	24	34	41	44	46
66%	8	16	27	37	45	48	50
72%	9	17	29	40	48	51	53
78%	9	18	31	43	51	54	56
84%	10	20	33	45	54	57	60
90%	11	21	35	48	56	60	63

来源：K Research 自建测算，数据截至 2026年6月，E 为 K Research 预测，单位 %

证伪信号清单（出现即说明我们错了）

- **ASIC 部署不及预期：**监测博通 AI 半导体季度营收增速 < +30% YoY、或 FY27 AI 指引下修——若 2026H2 出现，base 降级为 bear。
- **CUDA 壁垒证明不可逾越：**监测开源推理栈（vLLM/SGLang）在 ASIC 上的吞吐若持续低于 GPU 20%+、主流模型回流 GPU 推理——则推理替代假设证伪。
- **开放以太网渗透停滞：**监测 Dell'Oro 季度数据，若以太网在 AI 后端份额回落或 InfiniBand 重夺 50%+——则网络替代叙事证伪。
- **英伟达推理方案胜出：**若 Dynamo/NIM + Rubin 把推理每 token 成本压到 ASIC 之下，TCO 临界点逆转。

11 结论与行动

CONCLUSION

> **结论重申：**英伟达不会被"打败"，但它"通用 GPU + 私有互联 + 75% 毛利"的完整护城河正在被拆成三段——硅可被 ASIC 在推理侧替代、私有互联已被开放以太网反超、唯有上游 HBM/CoWoS/EUV 不可撼动。**份额迁移已不是"会不会"，而是"多快、谁收钱"。最确定的答案是：钱流向卖铲人——博通、开放以太网/光互连阵营、以及无论谁赢都要过路的 TSMC 与 SK 海力士。**

关注时间表

窗口	关键验证点
2026 H2	OpenAI-博通自研芯片量产；博通 FY27 AI 指引；Meta TPU 整机外购落地
2026–27	CoWoS 缺口收窄至 ~10%；HBM4 放量；1.6T 光模块放量
2027	scale-up 开放标准（UALink/SUE）拐点；Rubin vs ASIC 推理对决
2028–29	CPO 大规模部署；Barclays 万亿 capex 见顶窗口

行动指引

沿"份额迁移 → 利润池再分配"配置：**核心**博通（定制硅+以太网双引擎）；**弹性**Marvell、光互连（Coherent/中际旭创/康宁）；**压舱**TSMC、SK　海力士（卖铲人通吃）；**对冲**英伟达估值已计入部分悲观，非单边看空。本段为研究观点，非投资建议。

扫码进入口罩哥知识星球
解锁全部 K Research 独家研报

12 方法论与数据附录

METHODOLOGY

自建模型搭建逻辑

模型一·单位有效算力 TCO：把 GPU 与 ASIC 折算到"每一有效 PFLOPS·年"的总拥有成本（折旧+电力+网络/运维，3 年直线折旧、PUE 1.15、电价 $0.08/kWh、含 ASIC 软件移植税）。核心简化：以单加速器为单位、未完全建模超大 fabric 的网络规模效应（对 ASIC 偏保守）。由此反解"英伟达毛利率临界点"——即 GPU 降价到何种毛利方与 ASIC 打平。**模型二·份额迁移渗透曲线：**以 logistic S 曲线刻画 ASIC 价值份额，由"推理占比 × ASIC 在推理可俘获比例（随 TCO 优势上升、受软件可移植性上限约束）+ 训练侧少量俘获"驱动，校准至 2024 实绩与 TrendForce 2026 锚点。

关键假设表（节选）

变量	取值	来源	级别
GPU 取得成本（Blackwell 级 ASP）	$28,000	Epoch/Silicon Analysts BOM 测算	估计
ASIC 取得成本（TPU v7 级自有）	$13,000	SemiAnalysis/BofA	估计
英伟达数据中心毛利率	72–75%	NVIDIA 8-K（一手）	事实
GPU/ASIC 峰值 FP8 算力	5.0 / 4.6 PF	NVIDIA/Google 官方	事实
推理占 AI 算力比例（2026）	~2/3	Introl/McKinsey/Gartner	估计
ASIC 推理 TCO 优势（base）	41%	K Research 测算（对标 SemiAnalysis 20–50%）	估计
以太网 AI 后端份额（2026）	~67%	Dell'Oro	估计

关键一手信源（≥3）

① **NVIDIA FY2026 / Q1FY27 8-K**（数据中心营收、毛利率，SEC）。② **Broadcom FY26Q2 8-K 与电话会逐字稿**（AI 半导体 108 亿、FY27 >1000 亿、$30 亿要约从 25 亿上调）。③ **vLLM 官方博客 2025/10**（PyTorch 零改动上 TPU、吞吐 +20%）。④ **Alphabet 2026Q1 电话会**（Pichai TPU 对外供货原话）。⑤ **Google Cloud Ironwood / Anthropic / AWS Project Rainier 官方**（部署规模）。⑥ **SemiAnalysis TPU v7 深度拆解**（per-FLOP TCO 20–50%）。

数据口径与已知局限

· 份额、CoWoS 分配、HBM 份额为机构/产业链**估计**级，引用已标注；中国厂商（昇腾/PPU）产量为**传闻**级，未入核心计算。· TCO 模型未完全建模超大 fabric 网络规模效应与二手残值，对 ASIC 偏保守。· 单机租赁口径（Artificial Analysis）与大规模自用口径（SemiAnalysis）存在分歧，已并列并纳入 bear。· 估值区间为相对法，未做完整 DCF。· 截止日 2026年6月19日，此后数据可能变化。**宁写未知，不写大概。**