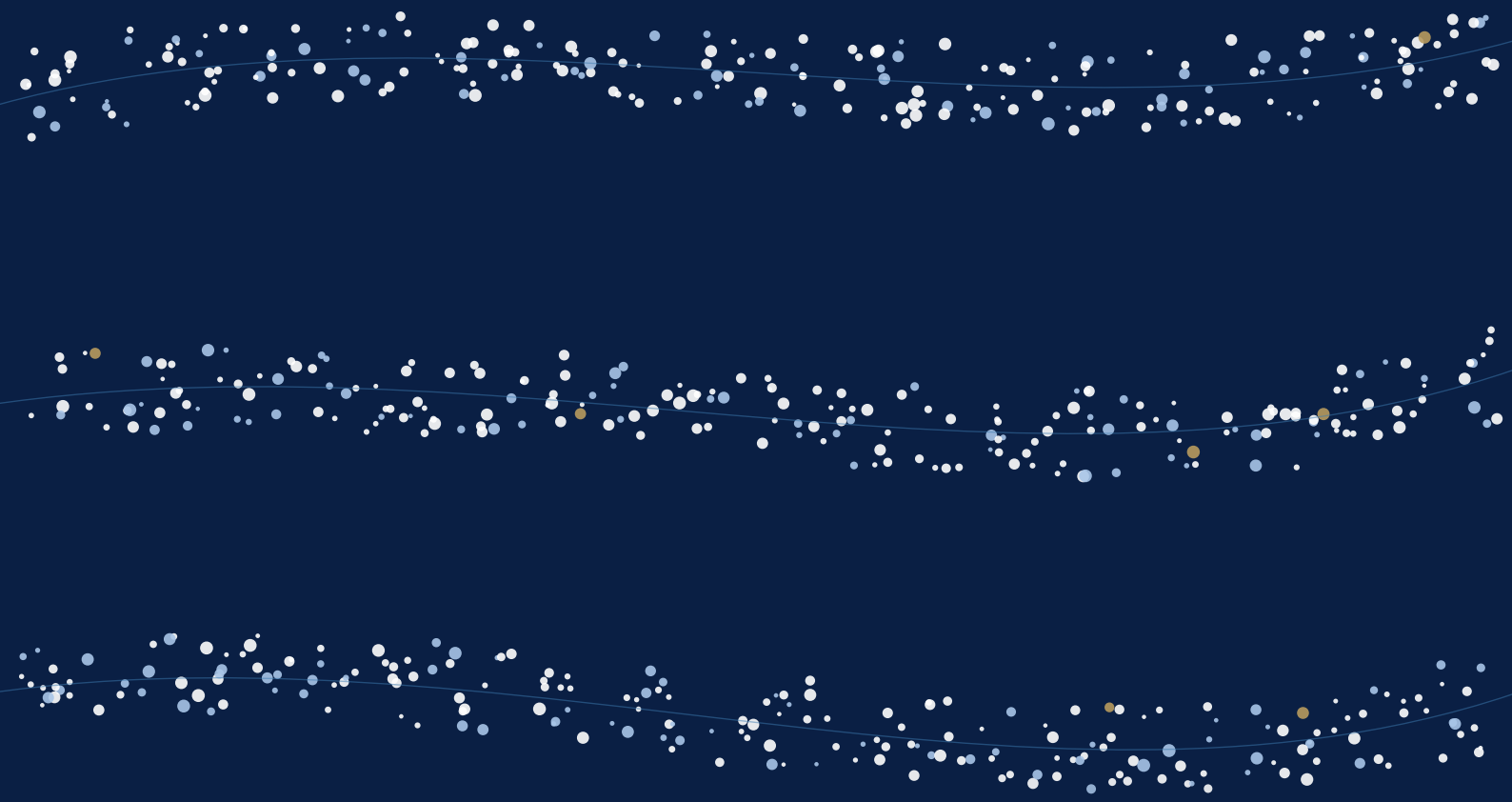
K RESEARCH No.013 · 2026年6月15日 · AI · 算力



推理压倒训练，通吃者却未必是 GPU

价值正以季度级速度从训练迁往推理。真正的赢家藏在 HBM、光网络、电力与定制 ASIC 这些「卖铲」环节——而通用 GPU 的推理定价权，正逼近拐点。

01 核心观点

市场仍用训练 GPU 给整条算力链定价。但推理已在 2026 年占到 AI 算力的过半，且对成本与功耗极度敏感——价值正从通用 GPU，向 HBM、光网络、电力与定制 ASIC 二阶迁移。GPU 守得住推理，却守不住推理的定价权。

01 **价值迁移是现在时，不是将来时。**按电力口径推理已占 AI 算力约 60%（IEA）；Gartner 测 2026 年 55% 的 AI IaaS 支出投向推理。

02 **通用 GPU 的推理份额正逼近拐点。**Barclays 测其从约 80% 降至 2028 年约 50%；ASIC 出货增速 +44.6%，三倍于 GPU 的 +16.1%。

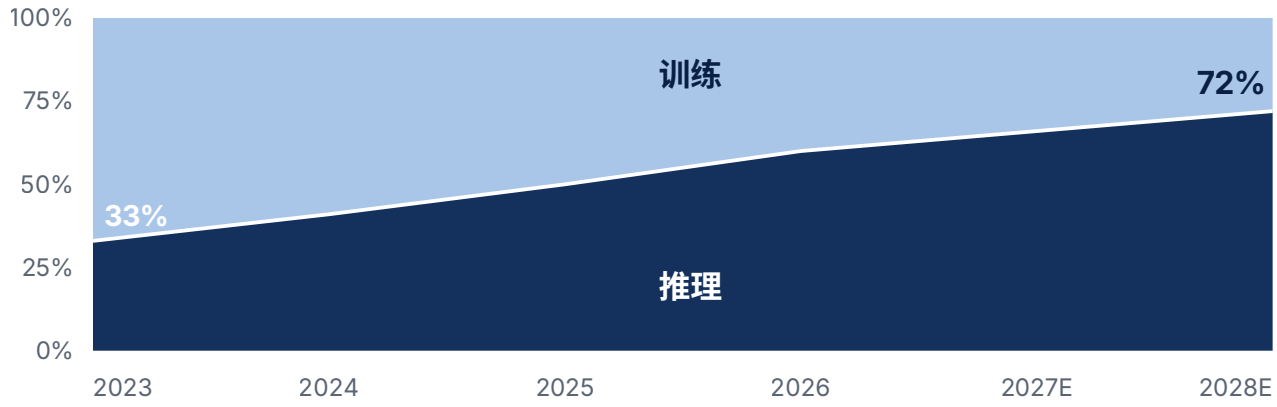
03 **最确定的赢家是「卖铲的二阶环节」。**HBM、光互联、电力、ASIC 协同设计——无论谁赢推理，都按算力规模等比受益。

04 **英伟达已亲口承认推理在分化。**斥 200 亿美元收购 Groq、推出专用推理芯片 Rubin CPX——护城河在推理侧变薄的自证。

05 **Jevons 是底层引擎。**token 成本 3 年降约 1000×，调用量同期暴涨（谷歌月 token 两年 +330×）：越便宜，总支出越大。

推理正在吞下 AI 算力：份额已过半，2028 年逼近七成

推理 vs 训练占 AI 算力比重，2023–2028E（综合算力与电力口径）



来源：Deloitte、Gartner、IEA，K Research 整理与估算，数据截至 2026年6月；E 为 K Research 预测

BEAR · 25%

GPU 推理份额 ≤50%

定制硅与云自研接管推理，推理商品化、API 毛利受挤

BASE · 50%

GPU 推理份额 ~60%

分层共治；HBM/光/电力/ASIC 二阶环节最受益

BULL · 25%

GPU 推理份额 ≥75%

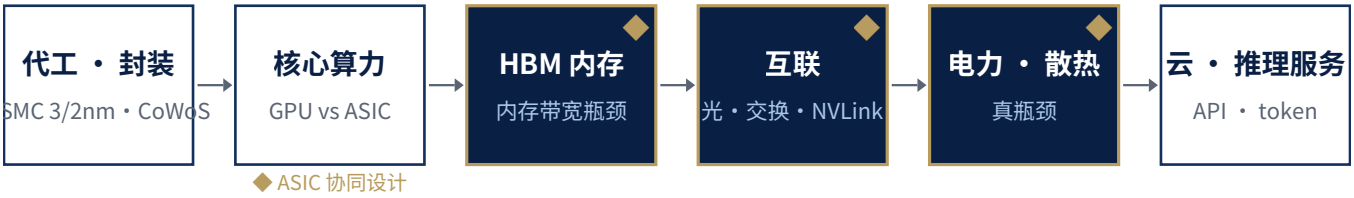
推理复杂度＋Jevons 让全栈 GPU 续霸定价权

02 产业链全景

把推理算力链拉平看，价值不再均匀分布在「造铲人」身上，而是向四个不可绕过的咽喉节点聚集。市场仍按训练 GPU 给整条链定价，却忽略了推理的物理学与训练截然不同：推理是内存带宽瓶颈、是功耗瓶颈、是规模化互联瓶颈，而非纯算力瓶颈。

推理算力价值链：四个 choke point 决定价值归属

上游→核心算力→内存→互联→电力→服务；◆ 为 K Research 标注的咽喉节点



◆ HBM 内存

SK 海力士约 62% 份额，独家先发 HBM4；
推理内存带宽瓶颈，2026 提价约 20%。

◆ 光互联

每 GPU/XPU 约 6 光模块，1.6T 放量；
AI 光市场 $16.5B(25)→$26B(26)。

◆ 电力 & 散热

并网 4–10 年 > 建厂 2–3 年；
推理占 AI 用电约 60%（IEA）。

◆ ASIC 协同设计

博通＋Marvell 约 95% 协同设计份额；
博通 AI 收入 $12.2B(FY24)→$56B(FY26E)。

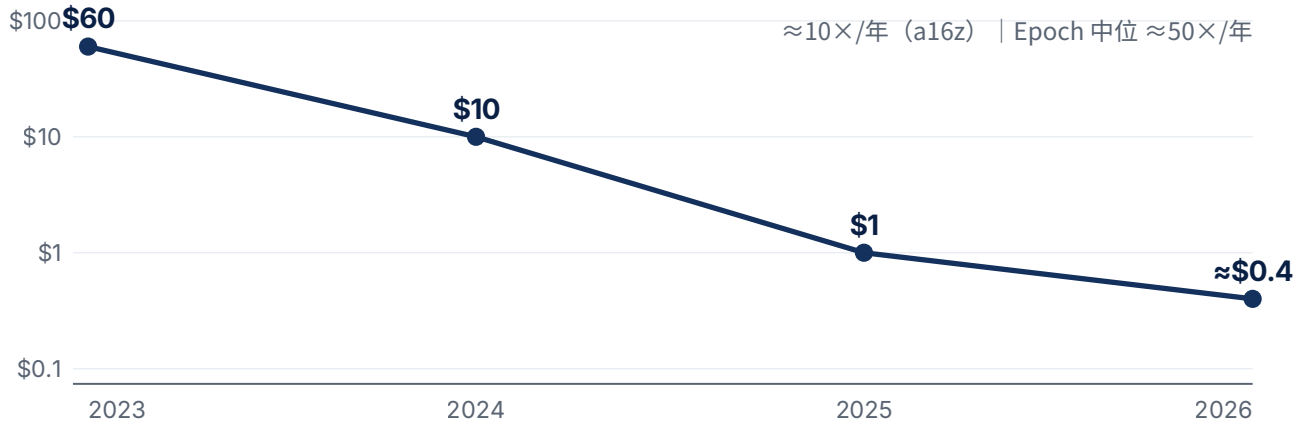
来源：各公司财报与官网、TrendForce、Dell'Oro、IEA，K Research 整理，数据截至 2026年6月

用 choke point 思维逐节点扫描：① **HBM**——每代推理芯片的内存墙，SK 海力士握约 62% 份额并独家先发 HBM4，议价权随长上下文与 KV 缓存上升；② **光互联**——每颗 GPU/XPU 附着约 6 个光模块，1.6T 进入放量；③ **电力与散热**——并网周期 4–10 年远长于建厂 2–3 年，电力已取代芯片成为真正的硬约束；④ **ASIC 协同设计**——博通＋Marvell 合计约 95% 份额，是云厂「去英伟达化」的唯一外部杠杆。这四节点共同特征是 architecture-agnostic：无论英伟达 GPU 还是云厂自研 ASIC 胜出，它们都按算力规模等比受益。

03 供需与价格

单位智能在坍塌：等能力 token 价格每年降约一个数量级

达到 GPT-4 级能力的最低 token 价格（美元/百万 token，对数轴，非零起点）

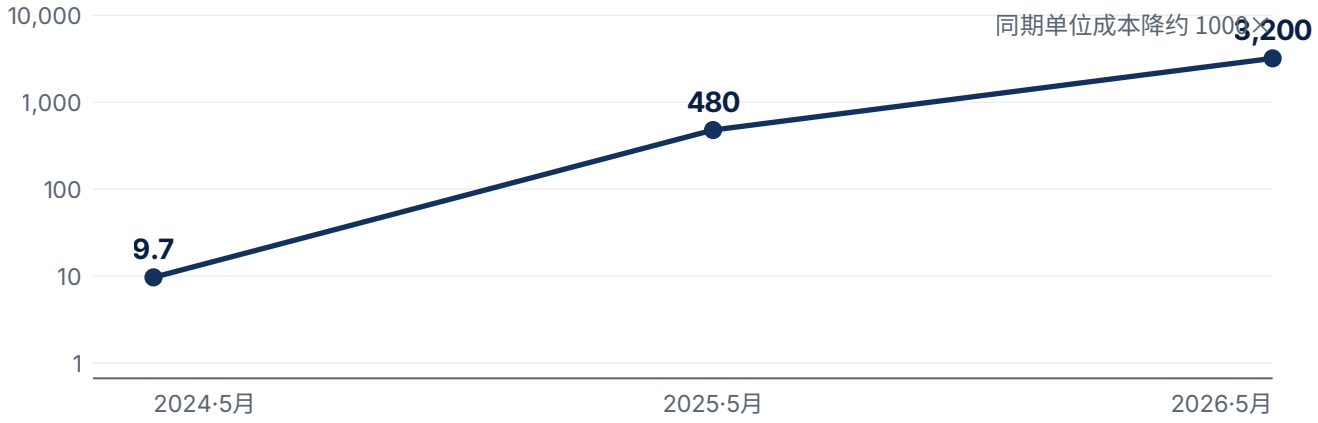


来源：a16z「LLMflation」、Epoch AI、各厂 API 定价页、Introl，K Research 整理，数据截至 2026年6月

成本曲线只是硬币的一面。单位 token 越便宜，需求就越不「珍惜」地消耗它——这正是 Jevons 悖论在算力上的复演。

而调用量在爆炸：谷歌月处理 token 两年增长约 330×

谷歌月处理 token 量（万亿/月，对数轴）；微软 FY25 经 Foundry 处理 500 万亿



来源：Google I/O 2026、微软 FY25 财报、OpenAI DevDay，K Research 整理，数据截至 2026年6月

两条曲线叠加，得到推理时代的底层方程：单位成本以每年约一个数量级坍塌，调用量以更快速度膨胀。结果是推理「总支出」不降反升，且支出结构从「买最贵的训练卡」转向「买最省的推理系统」——后者恰是 ASIC、HBM、光与电力的主场。**把便宜的 token 误读为算力链利空，是当下最大的认知错位。**

04 竞争格局

共识假设 CUDA 护城河会从训练平移到推理。但二者软件依赖结构不同：训练需与 CUDA 深度耦合的算子级优化，推理部署却被 vLLM、TensorRT-LLM、ONNX 等中间层抽象，硬件可替换性显著更高。叠加推理负载「固定、可预测、对单位成本敏感」的特征，定制 ASIC 的 40–65% TCO 优势在此被放大。

证据已从纸面走向账单：Midjourney 将推理迁至谷歌 TPU，月算力成本降约 65%；Uber 采用 AWS Trainium3 推理，成本降约 50%。Barclays 据此预测通用 GPU 在推理的份额将从约 80% 降至 2028 年约 50%。

最有力的反证来自英伟达自己：斥 200 亿美元拿下 Groq 的低延迟推理、推出专为长上下文设计的 Rubin CPX、以 NVLink Fusion 向第三方芯片开放互联——皆是「单一 GPU 通吃不了推理」的自证。我们的判断：英伟达守得住前沿大模型、MoE 与超长上下文这类高价值推理，但份额见顶是基准情形。

定制 ASIC 出货份额五年翻倍，逼近四成

AI 加速器出货份额，按通用 GPU vs 定制 ASIC，2024–2028E

来源：TrendForce、Bloomberg Intelligence、Barclays，K Research 整理与估算，数据截至 2026年6月；E 为预测

反共识

共识把英伟达当作「推理的唯一通道」，但推理恰恰是定制硅经济学最成立的场景。博通 AI 收入从 FY2024 的 122 亿美元跃至 FY2026 指引 560 亿美元，管理层「看到」FY2027 突破 1000 亿——这条曲线本身，就是超大厂把推理负载搬离英伟达的实时读数。

05 财务透视

四大云厂 capex 2026E
$700 B
▲ +69% 同比

博通 AI 收入 FY2026E
$56 B
▲ +177% 同比

谷歌月处理 token
3,200 万亿
▲ +567% 同比

HBM 市场 2026E
$58 B
▲ +53% 同比

推理链受益环节清单：谁按规模等比受益，谁被通缩反噬

环节	受益逻辑	确定性	弹性	代表公司
HBM 内存	推理内存带宽瓶颈，长上下文/KV 缓存推升用量	▪ ▪ ▪	▪ ▪ ▪	SK 海力士、美光、三星
光互联	每 GPU/XPU 附着约 6 光模块，1.6T 放量	▪ ▪ ▪	▪ ▪ ▪	中际旭创、新易盛、Coherent
ASIC 协同设计	云厂去英伟达化的唯一外部杠杆	▪ ▪ ▪	▪ ▪ ▪	博通、Marvell
电力 & 散热	并网周期＞建厂周期，power 是真瓶颈	▪ ▪ ▪	▪ ▪	发电、电网、液冷
通用 GPU	守高价值推理，但份额与定价权见顶	▪ ▪	▪ ▪	英伟达、AMD
推理优化/ neocloud	单位经济杠杆，软件层加速降本	▪ ▪	▪ ▪ ▪	CoreWeave、Fireworks、vLLM 生态
API 模型层	token 通缩挤压毛利，靠量补价	▪	▪ ▪ ▪	OpenAI、Anthropic

注：确定性/弹性为 K Research 三档定性评级（▪ ▪ ▪ 优 / ▪ ▪ 中 / ▪ 弱）。来源：各公司财报、TrendForce、Dell'Oro，数据截至 2026年6月

用二阶效应推演收束：推理需求每增长一分，最确定的现金流不在最显眼的 GPU，而在其上下游的**内存、光、电与协同设计**。它们是「卖铲人的卖铲人」——下游谁赢都要向它们采购；而 API 模型层虽享 token 量爆发，却要直面通缩对毛利的反噬，OpenAI 毛利率 2025 年从约 40% 降至 33% 即是注脚。

06 情景与风险

三情景矩阵：锚定 2028 年通用 GPU 的推理算力份额

维度	BEAR 定制硅通吃	BASE 分层共治	BULL GPU 续霸
核心假设	CUDA 在推理失效，推理商品化	GPU 守高价值推理，ASIC 拿固定批量	推理复杂度＋Jevons 让全栈续霸
GPU 推理份额(2028)	⩽ 50%	~ 60%	⩾ 75%
受益主线	云自研＋博通/Marvell＋电力	HBM/光/电力/ASIC 二阶环节	英伟达全栈＋NVLink 生态
概率	**25%**	**50%**	**25%**
触发信号	超大厂自研占其推理 >50%	ASIC 出货份额升至 35–40%	reasoning 需求超 NVLink 红利

我们错了的信号

- ☐ 英伟达连续两季推理收入份额回升，且 ASIC 出货增速明显放缓 — 2026Q4 前
- ☐ 主要超大厂推迟/缩减自研 ASIC 部署（Maia、Trainium 出货不及指引） — 2027Q1 前
- ☐ HBM/光模块价格因供给释放大幅回落，价值无法在二阶环节留存 — 2026Q4 前
- ☐ token 调用量增速显著放缓（谷歌/微软月 token 同比 <100%），Jevons 引擎熄火 — 2027Q2 前

风险列表

1. **需求证伪风险**——AI ROI 不及预期致 capex 下修——监测：四大云厂 capex 指引、稼动率
2. **英伟达反制超预期**——Rubin CPX＋Groq＋NVLink Fusion 锁定推理——监测：推理份额、NVLink 附着率
3. **政策/出口管制**——HBM 与先进制程对华管制扰动供需——监测：BIS 规则、HBM 对华出货
4. **产能集中反转**——光/HBM 供给释放快于需求致价格战——监测：800G/1.6T 出货与价格、HBM 合约价

07 结论与行动

不要只买「造铲人」，更要买「卖铲人」与「卖电人」。推理时代的超额收益，属于无论谁赢推理都按规模等比受益的二阶环节——HBM、光互联、电力、ASIC 协同设计；而通用 GPU 的合理定位，是「仍在增长但份额见顶」的核心持有，而非「通吃一切」的唯一选择。

接下来盯什么

- **2026Q3 博通 FY26Q3 AI 收入指引 $16B（+200%）兑现度**
  XPU 需求温度计——验证云厂自研推理芯片采购斜率
- **2026H2 英伟达 Rubin / Rubin CPX 出货 + GB300 放量**
  推理性价比被重新定义；观察 FP4 与长上下文专用化能否守住份额
- **2026H2 OpenAI「Titan」（博通）首批流片/部署**
  头部模型厂去英伟达化里程碑——一手验证 ASIC 替代节奏
- **2026Q4 HBM4 放量 + 2026 HBM3E 约 20% 提价兑现**
  内存价值留存验证——二阶环节定价权的硬指标
- **2027 四大云厂 capex 是否破 $700B + 自研芯片占比**
  价值迁移确认——GPU 推理份额读数（Barclays 路径）落地

扫码进入口罩哥知识星球 · 解锁全部 K Research 独家研报

本报告由 K Research 独立制作，所载信息均来源于公开资料，K Research 不对其准确性与完整性作任何保证。报告内容仅代表制作时点的研究观点，不构成任何证券、加密资产或其他金融产品的投资建议，亦不构成任何买卖要约。投资者据此操作，风险自担。市场有风险，投资需谨慎。本报告版权归 K Research 所有，未经书面授权，任何机构和个人不得以任何形式转载、复制或引用。