

省下来的钱去了哪一层

省下来的钱去了哪一层——效率跃迁改写的是 AI capex 的构成表，不是它的总额

开源模型的效率跃迁击穿的不是 AI 基建需求，而是「训练算力稀缺」这一个定价因子。需求从训练侧算力迁往推理侧内存、互联与电力；更要紧的是，2026 年 capex 增量里有相当一块是内存涨价而非算力增量——这既削弱了「capex 上行 = 算力需求上行」的推论，又直接坐实了价值正在向内存层沉淀。

效率跃迁改写的是 AI capex 的构成表，不是它的总额

涌现资本产业研究部 · 2026 年 7 月 27 日

30秒结论卡

一句话判断
开源模型的效率跃迁没有削减 AI 基建的总需求，它把需求从「训练集群的浮点算力」搬到了「推理侧的内存容量、互联带宽与电力」——受损的恰好是过去两年涨得最凶的那个定价因子：训练算力稀缺租金。

机会状态
● 观察为主 · 内存与电力两层 ● 算力租赁层 ●

市场错在哪
三件事被同时忽略了。其一，市场把「单位智能变便宜」直接读成「基建少花钱」，跳过了中间一步——省下来的钱不会消失，它流向下一个没被效率触及的物理闸门。其二，2026 年 capex 增量里有相当一块是内存涨价而不是算力增量（微软自述约 250 亿美元、Meta 自述上调主因是元件价格），这意味着「capex 上行」既不能直接证明算力需求上行，又直接指出了价值沉淀在哪一层。其三，市场对 capex 的态度已从奖励转为惩罚——Alphabet 上调指引当日股价下跌约 5%，这才是效率叙事最容易点火的位置。

重点标的
● MU（内存是效率跃迁的结构性净受益层，但须过周期峰值护栏）· ● TSM（无论前沿模型跑在谁的芯片上都绕不开的必经节点）· ●● GEV / CEG（电力闸门与模型效率完全解耦）· ● CRWV / NBIS / IREN（定价锚正是被击穿的那条腿）· ● ORCL（资本开支约为经营现金流 1.7 倍，链条最脆一环）

最大反证
推理 token 总量增速首次跌破单位成本下降速度，且 NVIDIA 数据中心收入出现环比负增长。两者目前均无迹象；任一出现，本报告的杰文斯前提即失效。

观测窗口尚未闭合的声明：微软、Meta、亚马逊的日历第二季度指引预计在未来两周内发布，此前三家的现行数字均为 4 月 29 日口径。「效率跃迁之后超大厂是否变调」这一命题，在三家最大买方身上此刻还没有观测值。本报告第三节的结论建立在 7 月已发生的四条披露之上，并把这三家的下一次指引列为第一检验点。

一、投资要点

框架一句话：效率跃迁作用在模型层，闸门却在内存、互联与电力层——所以它重新分配卖铲子的收益，而不是消灭它。

#	核心观点	方向
1	「低算力追平前沿」这一前提在可核实的事实层面不成立——K3 是 2.8 万亿参数、100 万上下文、需 64 卡以上超节点服务的重型模型，且发布即撞上 GPU 产能挤兑	●
2	中国开源模型「便宜一个数量级」的旧叙事在这一代已经失效：K3 定价 \$3/\$15 每百万 token，与 Claude Sonnet 5 标准档持平，较上代涨约 3.2 倍	●
3	稀疏化削减的是每 token 的浮点算力，不削减权重驻留——全部专家权重仍须常驻显存，HBM 容量需求是升不是降	● 内存
4	2026 年 capex 增量中相当一块是内存涨价而非算力增量 ——这一条既削弱多头的「capex 上行即需求上行」，又坐实空头最不愿承认的「价值正在向内存层沉淀」	● 内存
5	三道物理闸门中只有先进封装出现相对松动（缺口由约 20% 收窄至约 10%）；内存因供给侧主动转产而反向收紧，电力与模型效率完全解耦	● 内存/电力
6	7 月已披露的四条（Alphabet、台积电、SK 海力士、OpenAI）全部指向上调；因「模型效率」下调 capex 的事件截至今日尚未出现——但这是「尚未出现」，不是「已被证否」	● 空头论
7	市场对 capex 的定价态度已由奖励转为惩罚；真正的裂缝在融资结构：债务占增量 capex 的比重三年内由 9% 升至约三分之一	● 久期
8	承压最确定的是靠「训练算力稀缺」定价、且客户具备自建能力的算力租赁层；A 股「K3 概念」篮子无一条可核实敞口	●

二、事实底座：被讲反的那个前提

市场这一周的叙事是「用更少的算力追平前沿」。把可核实的部分单独拎出来，得到的画像与这句话方向相反。

官方口径能确认的只有三项：总参数 2.8 万亿、每 token 激活 896 个专家中的 16 个、上下文 1,048,576 token，外加一条官方推荐——生产部署建议 64 张加速器以上的超节点配置。生态中流传的「解码提速 6.3 倍」「训练效率 +25%、开销 <2%」在官方技术博客中查无实据；就连「效率提升 2.5 倍」这一条，二手源之间也存在「2.5 倍」与「25%」两种相差一个数量级的记载与口径混淆。**本报告因此只作定性表述——厂商自述效率显著提升——不引用任何具体倍数。**

每 token 激活参数：官方未披露。二手源至少给出三组互斥值。这是「低算力追平前沿」论证的分母，任何建立在其上的成本比值都不成立。需要提示的是，「 $16 / 896 \approx 1.8\%$ 」是**专家个数的稀疏度**，不是激活参数比，更不是算力比——注意力与共享参数每 token 全部激活。把这个 1.8% 读成「只用了 1.8% 的算力」是本轮最常见的误读。

训练算力与训练成本，一个绝对数字都没有披露。「低算力追平前沿」这个命题，到今天为止没有一次可核实的绝对成本披露支撑。值得一记的是，该厂商负责人本人曾公开否认上一代模型流传的某个训练成本数字为官方口径——连模型方自己都不认领这类数字。

口径参照也必须交代。上一轮同类宣称是 DeepSeek V3 的 557.6 万美元，该数字只覆盖一次预训练 run 的 GPU 租用，不含实验、数据、人力与硬件 TCO；第三方全口径估计另有服务器资本开支约 16 亿美元、集群运营约 9.4 亿美元等分列口径（不同研究框架不可合并成单一倍数）。**结论只需一句：宣称成本与全成本之间隔着约两个数量级，而 K3 连一个可被折扣的数字都没有给出。**

排名。在第三方智能指数上 K3 位列前四，落后于当前两个闭源前沿模型；「登顶」只在前端代码单项 arena 成立。全域第一不成立。

定价。K3 为 \$3 / 百万输入 token、\$15 / 百万输出 token、\$0.30 / 百万缓存命中输入，与 Claude Sonnet 5 标准档持平，是中国实验室迄今定价最高的模型。逐档对比上代 K2.6：输入价由 \$0.95 涨至 \$3（约 3.2 倍），输出价由 \$4 涨至 \$15（约 3.75 倍），缓存命中价由 \$0.16 涨至 \$0.30（约 1.9 倍）——三档口径不同，不可混用为单一倍数。这一条本身就足以推翻「效率跃迁 = 单价崩塌」的推论链：**能力上去之后，中国实验室第一件事是提价。**

两条最反直觉的事实。其一，K3 上线约 48 小时内即暂停新订阅——需求压垮了自有推理基础设施，更强的开源模型直接转化为更多推理算力需求。其二，在第三方评测中，K3 跑完整套评测的输出 token 量约为全模型中位数的 2.06 倍，同时又比自家上一代少用约 21% 的输出 token。**这两个数放在一起，恰好就是本报告的主线：单位智能成本在降，总 token 消耗在升。**

还有一点须诚实说明：本轮真正有一手论文支撑的效率跃迁证据，其实不在 K3，而在同期的另外两个开源模型——一个在论文中给出了将算力降至约 27%、KV cache 降至约 10% 的机制级硬数，另一个给出了固定 KV 预算下的量化提速。**K3 是这轮叙事的触发器，却是其中证据等级最低的一个。**

三、钱流：效率提升的这半年，capex 往哪个方向走

判断「效率是否击穿 capex」，不能问模型厂商，要问买硬件的人。这里必须先把观测窗口讲清楚——**这是本节最容易被误用的地方。**

3.1 时间线：属于 7 月这一轮的四条披露

主体	披露日	2026 capex 口径	方向	置信度
Alphabet	07-22	全年指引 \$1,950—2,050 亿（前值 \$1,800—1,900 亿）	▲ 单季上调约 \$150 亿，并预告 2027 年继续显著增长	B
台积电	07-16	\$600—640 亿（前值 \$520—560 亿）	▲ 上调约 15%，全年营收增速指引上修至「略高于 40%」	B
OpenAI	07-23	2030 年前累计算力支出 \$7,500 亿（由约 \$6,000 亿上修）	▲	C
SK 海力士	7 月	完成美股 ADR 上市募资约 \$265 亿，明确用于新晶圆厂、新封装厂与 EUV 采购	▲ 供给侧融资	B

3.2 尚未观测的三家（本命题的第一检验点）

微软的日历年 2026 年约 \$1,900 亿、Meta 的 \$1,250—1,450 亿，均为 4 月 29 日给出；亚马逊约 \$2,000 亿为维持值。三家的日历第二季度指引尚未发布。把 4 月的数字当作「效率跃迁之后仍在加码」的证据是不成立的——那时 K3 尚未问世。诚实的表述是：三家最大买方在这个问题上此刻还没有数据点，未来两周内的指引是第一次真正的检验。

自下而上重建合计口径：亚马逊约 2,000 亿 + 微软约 1,900 亿 + Alphabet 1,950—2,050 亿 + Meta 1,250—1,450 亿 = 约 \$7,100—7,400 亿（区间而非点值，各家财年与日历年口径混合，不含融资租赁，不含甲骨文等非四大）。2025 年同口径按分项反推约 \$4,000 亿。方向是大幅上行且年内多次上调，这一点不受观测窗口影响。

3.3 最要紧的一条：capex 涨的是价格，还是算力

这是本报告认为市场普遍漏掉的一层，也是全篇的枢纽。

微软把日历年 capex 中的约 250 亿美元明确归因于内存与元件涨价；Meta 把指引上调的主因表述为元件价格上涨与额外数据中心成本，而非更多算力。也就是说，2026 年 capex 同比高增的分子里，塞进了一大块存储涨价，物理算力增量被系统性高估。

这一条是双刃剑，两边都必须诚实承认：

- 对多头：它削弱「capex 仍在上行，所以算力需求仍在上行」这个推论——涨价撑起来的 capex 不能证明算力需求。
- 对空头：它直接坐实了「省下来的钱去了哪一层」这个问题的答案——去了内存。买方自己在财报电话会上把这笔钱的去向说清楚了。

这两件事指向同一个结论：capex 的构成表被改写了，而改写的方向正是本报告的主线。

3.4 物理闸门没有因效率而下修

台积电在 7 月 16 日的电话会上把资本开支上调 15%、把全年营收增速指引上修至「略高于 40%」。一家真正相信「模型变高效会削减需求」的公司，不会在同一场电话会上做这两件事。

3.5 市场已经从奖励 capex 转为惩罚 capex

必须把「capex 加码」与「市场奖励 capex 加码」分开：Alphabet 上调指引当日股价下跌约 5%，Meta 上调当日跌逾 6%。**基本面叙事没崩，与持仓不痛，是两件事。**这个张力恰恰是效率叙事最容易点火的位置——市场此刻正在寻找一个理由，去给已经不受欢迎的资本开支重新定价。

四、杰文斯：证据方向明确，但有一条必须交代的方法论裂缝

效率提升会不会减少总需求，是一个可以用数据回答的问题。但这里有一处专业读者一击致命的地方，本报告选择主动交代。

第一条曲线（价格）。固定能力档位的单位 token 成本，按跨模型同能力包络线口径年降约 10 倍（该口径原始发布于 2024 年 11 月，早于推理模型大规模商用，仅作保守下限）；按分档基准口径，全样本中位约 50 倍/年，若只取 2024 年之后的子样本，中位升至约 200 倍/年——**降价在加速。**

⚠ **方法论裂缝（本报告主动披露）：**上述价格数据集**明确排除推理模型**（因 token 生成模式不同）。而本报告用来解释「降本被需求吃掉」的传导机制，恰恰全部押在推理与 Agent 负载上。**两条腿站在互斥的样本上。**因此价格年降曲线不能直接外推到以推理时计算为主导的 2026 年负载结构——这是全篇最脆弱的一处推理链，任何据此做的精确弹性测算都不成立。本报告因此**不给弹性数值**，只给方向。

第二条曲线（数量）。可核实的一手口径有两个：某超大厂平台自述月处理量 3.2 quadrillion token、同比约 7 倍（单一厂商口径，非全球）；开发者路由平台官方自述「全年处理量将超过 1 quadrillion」，按每周 25 万亿 token 反推年化约 1.3 quadrillion，半年增约 5 倍（该口径对开源与低价模型有结构性偏向，不可外推为全行业）。此前流传的更大总量数字存在内部算术矛盾，本报告不予采用。

两条曲线不能直接相乘。售价含竞争性补贴与毛利压缩，不等于硬件成本；同时 Agent 负载把「一次请求」的 token 消耗显著抬高（该抬升幅度目前**没有可核实的一手测算**，本报告只作定性表述，不给量化倍数）。判断总算力开销方向，必须回到采购端：**NVIDIA 数据中心收入 FY2026 全年 \$1,937 亿（同比 +68%），FY2027 第一季度单季 \$752 亿（同比 +92%、环比 +21%）。**这是全篇口径最硬的一组数。

上一次的兑现率是零，但恢复很慢。2025 年 1 月 27 日 DeepSeek R1 冲击，NVIDIA 单日下跌 17%、蒸发约 5,890 亿美元；次日反弹 8.9%。但**市值创出新高是在约五个月之后**（2025 年 6 月下旬），期间还叠加 4 月的二次重挫。此后 15 个月，两个可检验的硬指标——NVIDIA 数据中心收入与超大厂 capex——全部朝预测的反方向走。

这一条对读者的实操价值大于对错本身：基本面推翻了崩塌叙事，不等于持仓不痛。上一轮估值修复花了五个月。

边界。R1 冲击发生在预训练规模主导算力去向的阶段，2026 年的算力重心已转向推理时计算。「上次没崩」只证明市场对这一叙事的即时定价在历史上过度，不证明这次的结构相同（置信度 C）。

五、闸门体检：效率跃迁到底松动了哪一道

这是全篇的分层核心。模型层的效率提升，作用到物理供应链上是三道闸门的差异化传导，不是齐涨齐落。

闸门	效率跃迁的传导	2026 下半年状态	方向
先进封装	每 token 浮点算力下降，边际削弱单位需求	唯一相对松动：供需缺口预计由约 20% 收窄至年底约 10%；产能口径应含委外封测伙伴	●
内存	反向加剧：稀疏化削算力不削驻留，长上下文让 KV cache 成为「第二个模型」	供给侧主动收紧：产能向利润更高的通用 DRAM 倾斜；HBM 预计占 2027 年 DRAM 投片量约 30%，对通用 DRAM 形成挤出	●
电力与配 电	完全解耦，效率跃迁无法缓解	最硬一道：美国 2026 年数据中心公告管线约 16 GW、实际在建约 5 GW	●

5.1 稀疏化削的是算力，不是显存

MoE 每 token 只激活少数专家，但全部专家权重仍须常驻显存。K3 用 2.8 万亿总参跑 16/896 的专家稀疏度——每 token 浮点算力大幅下降的同时，**权重驻留需求是史上最大的**。按一篇 2026 年 7 月的预印本情景测算，当代旗舰级模型 FP8 权重驻留在 750 GB 量级（尚未计入 KV cache），单副本约需十张上代旗舰卡、或约三张当代机架级模块（非单卡可承）。效率跃迁对内存是**容量强需求**。

5.2 长上下文把带宽变成新天花板

长上下文解码为内存带宽受限，KV cache 随上下文线性增长。一旦溢出显存落到主机侧，**显存与外围链路的带宽差约两个数量级**（与主机内存的差距约一个数量级），跨层搬运即主导延迟。稀疏注意力与 KV 压缩确实缓解了这一压力，但缓解出来的空间随即被更长上下文与更多推理时计算吃满——这正是把上下文推到 100 万 token 的直接后果。

5.3 供给侧的选择比需求侧的预测更硬

最能说明问题的不是任何价格预测，而是**内存厂自己把产能投向了哪里**：自 2026 年第一季度起，HBM 的晶圆营收与获利能力已被通用服务器内存产品反超。这解释了为什么会出现「把本应转向下一代 HBM 的产线改回通用 DRAM」的报道（该产线调整为**媒体转述、公司未正式确认**，置信度 C）。当**最赚钱的产品不再是 AI 专用件**时，供给侧的理性选择会自动收紧 AI 侧的**可得产能**——这一层与模型效率毫无关系。

（价格口径须分清：某机构对 2026 年第一季度 DRAM 整体合约价的**预估**为环比 +90—95%、PC 用途 +105—110%，而该季 DRAM 产业营收的**实际表现**为环比 +81%。三个数字口径不同，不可混用。）

5.4 电力：口径必须说清楚

美国 2026 年数据中心公告管线约 16 GW、实际在建约 5 GW，隐含三到五成**延期或取消**——须注意业内对「取消」比例存在明确反驳，把延期与取消混用会得出过度悲观的结论。电网侧另有一个常被误引的数字：约 2,300 GW 的并网申请积压是**发电与储能侧队列**（截至 2024 年底口径，且已自 2023 年约 2,600 GW 的峰值回落），**不是数据中心负载接入队列**，只能作为电网拥堵的宏观背景，不能用来直接论证「GPU 订单被挡在门外」。负载接入以年为单位排队（二手口径 4—7 年，未见一手监管文件确认，置信度 C）。

即便把口径收紧到最保守，这一层的结论不变：它是三道闸门里唯一与模型效率完全解耦的一道。

5.5 小结

效率跃迁不是压缩了闸门需求，是把需求从「训练集群的算力」搬到了「推理侧的内存容量与带宽」，并把总量交给了使用强度放大。对 capex 逻辑的正确修正不是「击穿」，而是「**换层**」。

六、真正的裂缝：不在效率，在融资结构

如果本轮 AI 基建有一条会先断的绳子，它不在需求侧。

Alphabet 2026 年第二季度自由现金流转负 59 亿美元，为 2004 年上市以来首次。 同季资本开支 449 亿美元，超过经营现金流 391 亿美元；营收 1,198 亿美元（同比 +24%），云业务 248 亿美元（同比 +82%），订单储备单季增加 500 亿美元至 5,140 亿美元。需求侧毫无问题——问题是现金流缓冲正在被消耗，而公司在同一场电话会上把全年指引再上调约 150 亿美元。

两处必须扣除的水分：该季净利润包含约 990 亿美元的权益投资估值收益，属非经营性，任何「虽然自由现金流转负但盈利仍强」的表述都会踩在这笔一次性收益上；账上 2,425 亿美元现金及有价证券中约 871 亿美元为权益类证券，不能当作可即时投入资本开支的干火药。扣除之后，这仍是一个**久期问题**而非流动性问题——滚动十二个月自由现金流仍为正约 530 亿美元。

融资结构的迁移比单季数字更重要。 债务在增量 capex 中的占比，三年内由 9% 升至约三分之一（截至 2026 年 6 月的滚动十二个月口径 32%，单一数据机构口径，本报告显式标注其为支柱数的单源风险）。已一手确认的表外案例是一笔 270 亿美元的私人信贷合资（结构为信贷方持股八成、企业方两成）。甲骨文是杠杆最极端的样本——本财年资本开支 557 亿美元、经营现金流 320 亿美元，**比值约 1.7 倍**，自由现金流缺口约 237 亿美元，下一财年净资本开支指引约 700 亿美元，且对单一客户的合约集中度极高。

这才是市场应该重新定价的东西。 效率跃迁不改变 AI 的需求曲线，但它改变了**回本时间**的分布假设；而当 capex 由债务而非自由现金流承担时，回本时间拉长直接等于股权价值缩水。

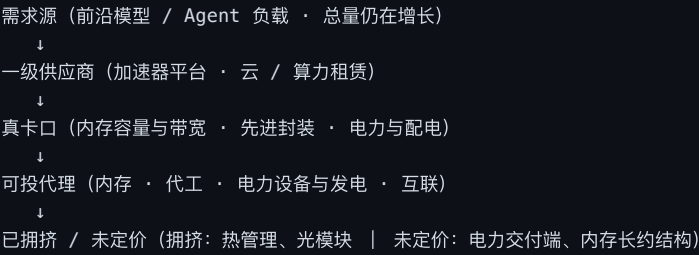
7 月 24 日的盘面已经在做这件事：费城半导体指数当日下跌 4.40%，同目标普 500 上涨 0.05%、恐慌指数 18.58——**这是板块内部的定价因子重估，不是流动性危机**。杀的是高久期、高杠杆、靠稀缺租金定价的那一段，龙头几乎没动（NVDA -0.92%，对比 CRWV -11.37%、NBIS -15.02%）。放到更长的窗口看，费半自 6 月末高点的回撤已逾 20%，进入技术性熊市。

这里必须把观点与事实分开。 价格事实成立，但「因为 K3 所以杀硬件、因为杰文斯所以反弹」这个闭环是本研究部的解读，**不是已被验证的机制**：本轮回撤在 K3 发布之前就已启动，而 7 月下旬那次反弹的市

场归因主要落在韩国出口数据与卖方对内存股的看多重申上。把叙事当机制，是这轮行情里最容易犯的错。

七、受益标的落地

7.1 价值链五层



7.2 标的表（美股为 2026-07-24 收盘 · A 股为 07-27 盘中）

标的	现价	PE	位置	效率跃迁下的机制	评级
MU	\$920.95 (-6.99%)	20.8 (前 瞻 5.9)	HBM + 通用 DRAM	稀疏化推高权重驻留、长上下文推高 KV cache；买方已在电话会上把 capex 增量的一部分明确归因于内存涨价	● ●
TSM	\$403.41 (-2.93%)	—	先进制程 + 先进封装	无论前沿模型跑在谁的芯片上都是必经节点；同场上调 capex 与营收指引	●
NVDA	\$206.84 (-0.92%)	31.6 (前 瞻 23.6)	推理平台	训练算力叙事被削弱，但高显存容量平台承接稀疏大模型分片（约三张机架级模块承一个副本），平台优势反被强化	● ●
GEV	\$1,014.75 (-1.59%)	29.1	电力设备	长交期设备与电网拥堵是效率完全触不到的一层	● ●
CEG	\$274.35 (-0.45%)	23.8	发电	同上；当日仅微跌，已显现防御属性	●
AVGO	\$381.92 (-2.69%)	63.7 (前 瞻 20.4)	定制 ASIC + 以太网	单价压力倒逼云厂自研推理芯片	● ●
VRT	\$290.36 (-4.50%)	73.0 (前 瞻 46.9)	供电与热管理	方向对但估值已高度定价，当日跌幅大于龙头	●
ALAB	\$291.58 (-10.82%)	195.7	内存互联	机制正确（池化与带宽），但缺一手收入敞口拆分， 线索级	●
ORCL	\$114.99 (-4.21%)	—	云 / 算力承包	资本开支约为经营现金流 1.7 倍，单一客户集中度极高，链条最脆一环	●
CRWV	\$71.88 (-11.37%)	无（亏损）	算力租赁	定价锚正是被击穿的那条腿；大客户同时是潜在竞争者	●
NBIS	\$187.77 (-15.02%)	72.5	算力租赁	同上；单客户集中度须按准确口径看——3 月签署的协议为 120 亿美元专属产能硬承诺 + 至多 150 亿美元采购选择权，不可按 270 亿全额计集中度	●
IREN	\$37.07 (-8.65%)	48.1	算力租赁	同源承压，合同结构披露不足	●
中际旭创 300308	¥1,048.88	78.3	光互联	推理侧东西向流量受益，但估值已充分反映	●
澜起科技 688008	¥227.43	108.5	内存接口	机制与内存层同向，估值高位	●

关于内存层的护栏（必读）：前瞻市盈率降到个位数，在周期股上通常不是估值折价，而是**周期高点利润造出来的「假便宜」**。内存层的方向判断成立，但估值锚必须用穿越周期的口径，不能用峰值年度盈利直接乘倍数。

7.3 伪受益（已完成代码↔公司名↔市场三一致核验）

被冠以「K3 概念」的 A 股篮子，逐一核对主营与本次需求落点：

标的	主营	与云端稀疏大模型推理的因果链
星宸科技 301536	边缘 AI SoC	✗ 基本面是真的，但与 K3 无关（见下）
三未信安 688489	商用密码	✗ 无
汇金股份 300368	金融机具 / 数据服务	✗ 无
金达莱 688057	环保水处理	✗ 无

关于星宸科技，本报告要把话说准。 公司 7 月 18 日业绩预告的数字经核是**真实且自洽的**：预告归母净利润 8.20 亿至 9.00 亿元（同比 +583.72% 至 +650.42%），扣非 7.0 亿至 7.8 亿元，营收 26.2 亿至 27.2 亿元（同比 +86.74% 至 +93.86%），对应净利率约 31% 至 33%——量级关与子项关均通过。**问题不在业绩真假，而在归因：**这份高增长来自端侧与边缘 AI SoC 的出货，而 K3 这类云端稀疏大模型的需求落点是内存、先进封装、光互联与电力，两者之间不存在因果链。**说它「基本面好但故事挂错了」，比暗示它数字有问题更站得住，也更经得起反驳。**（另须更正一处常见转述：该股 7 月 20 日为上涨约 16.11%，并非一字涨停。）

统一反证判据：拿不出「与模型厂商或其推理供应商的合同、适配或收入占比」证据的，涨的都是情绪不是订单。评级 ● 回避——回避的是这条归因链，不是这些公司的经营本身。（四家公司代码↔名称↔市场身份已经行情接口双向核验。）

八、风险与反证

#	反证条件（出现即证明本报告结论错）	当前状态
1	推理 token 总量增速首次跌破单位成本下降速度	未出现，但 两者的观测样本互斥 （价格样本不含推理模型），该比较本身精度有限
2	NVIDIA 数据中心收入出现环比负增长	未出现（最新单季环比 +21%）—— 全篇口径最硬的一条
3	微软 / Meta / 亚马逊未来两周的指引出现下调或明确放缓	观测窗口未闭合 ，此为本题第一检验点
4	剔除内存涨价后，capex 的实物算力增量转为下降	需拆分，目前买方仅给了归因未给完整拆分——本报告最想要而尚未得到的一组数
5	先进封装缺口转正（供过于求）且扩产计划下修	未出现；缺口收窄但仍为负
6	内存合约价见顶回落、2027 配额出现退订	未出现
7	债券市场对 AI 债认购倍数持续下滑并出现发行失败	部分出现 ——本报告第六节核心监测项
8	模型厂发布可核实的绝对训练算力与成本，且显著低于同代前沿	未出现；至今零绝对披露

本报告最脆弱的两处，已在正文显式标注：其一，杰文斯价格曲线的观测样本不含推理模型，而传导机制全押在推理负载上（第四节）；其二，三家最大买方的观测窗口尚未闭合（第三节）。若第 1、2 项同时出现，第五节的「换层」结论应立即改写为「总量收缩」。

九、行动清单

- **现在重点观察**：内存层（效率跃迁的结构性净受益方，买方已自述 capex 增量部分流向此处，但须过周期峰值护栏）；电力与配电层（与模型效率完全解耦的最硬闸门）。
- **只监控不参与**：算力租赁层。定价锚被削弱、客户具备自建能力、重资产已按吉瓦级前置锁定——敞口最不对称的一类。
- **绝不追**：A 股「K3 概念」情绪篮子；任何以「训练成本便宜多少倍」为核心论据、却拿不出绝对算力披露的叙事。
- **下次看这五个数**：①微软 / Meta / 亚马逊未来两周的 capex 指引（第一检验点）；②下一季 NVIDIA 数据中心收入环比；③capex 中内存涨价与实物算力增量的拆分；④AI 相关债券的认购倍数与利差；⑤开放权重模型的 API 定价方向（提价 3.2 倍是否成为行业趋势）。
- **哪个信号出现要重写结论**：token 总量增速跌破单位成本降幅，或超大厂首次以效率为由下调 capex。

十、来源置信度表、口径与免责

结论 / 数字	来源	日期	口径	置信度
K3 规格（2.8T 参数 / 16-of-896 / 1M 上下文 / 推荐 ≥64 卡超节点）	厂商官方技术博客	2026-07-16	官方；16/896 为 专家个数稀疏度 ，非算力比	B
每 token 激活参数	官方未披露 ，二手至少三组互斥值	—	技术报告与权重同日（今日）发布，为待验钩子	不予采用
效率提升倍数（2.5x / +25%）	二手源互相打架且口径不同	2026-07	仅作定性	D
「解码 6.3x / 训练效率 +25%、开销 <2%」	二手转述，官方博客查无实据	2026-07	不予采用	D
K3 定价 \$3/\$15/\$0.30；对上代输入 3.2x、输出 3.75x、缓存命中 1.9x	厂商价目逐档比对	2026-07	官方；三档不可混用	B
K3 评测输出 token 为全模型中位数约 2.06x，同时较上代少约 21%	第三方评测机构模型页	2026-07	跨模型中位数比 vs 同族对比，两个基准	B
上一代开源训练成本 557.6 万美元	模型方论文	2025-01	仅一次预训练 run 的 GPU 租用	A
其全口径另有服务器资本开支约 16 亿、集群运营约 9.4 亿美元等分列估计	第三方研究	2025	不同框架不可合并为单一倍数	B
Alphabet Q2 自由现金流 -59 亿 / 指引上调至 \$1,950-2,050 亿 / 净利含约 990 亿权益收益	季报、8-K 与电话会	2026-07-22	GAAP 季度	B
微软日历年约 \$1,900 亿，其中约 \$250 亿系内存涨价	电话会 CFO 表述	2026-04-29	日历年， 非 7 月口径	B
Meta \$1,250-1,450 亿，上调主因元件涨价	电话会	2026-04-29	非 7 月口径	B
四大 2026 合计约 \$7,100-7,400 亿	自下而上分项重建	2026-07-22	财年/日历年混合，不含融资租赁	B
台积电 capex 上调 15% 至 \$600-640 亿	电话会	2026-07-16	年度	B
先进封装 2026 年底目标 12-13 万片/月（另有委外伙伴 5-6 万片/月）	产业链调研口径， 非台积电披露	2026 上半年	月产能	C

结论 / 数字	来源	日期	口径	置信度
供需缺口 20% → 10%	产业研究机构	2026-06-15	年底预估	B
HBM 占 2027 年 DRAM 投片约 30% / 1Q26 起 HBM 晶圆获利被通用产品反超	产业研究机构	2026 上半年	投片量与获利口径	B
内存产线转产调整	韩媒转述，公司未正式确认	2026-06	未证实	C
债务占增量 capex 9% → 32% / 表外 270 亿合资	单一数据机构 / 企业公告	2026-06 / 2025-10	滚动十二个月	B-
甲骨文 capex 557 亿 / 经营现金流 320 亿 (1.7x)	公司披露	FY2026	财年	B
NVIDIA 数据中心收入 1,937 亿 / 单季 752 亿美元	公司财报	FY2026 / FY2027Q1	GAAP	B
单位成本年降（包络线约 10x；分档中位 50x，2024 后子样本约 200x）	第三方研究	2024-11 / 2026	样本排除推理模型	B
token 量：某平台月处理 3.2 quadrillion、同比约 7x；路由平台年化约 1.3 quadrillion、半年 5x	厂商自述 / 官方新闻稿	2026-05 / 2026	单一厂商，非全球	B
R1 冲击：单日 -17%、蒸发 5,890 亿、次日 +8.9%、约五个月后创新高	市场数据	2025-01 至 2025-06	百分比为准，市值增减额系推算	B
电力管线 16 GW vs 在建 5 GW；并网队列约 2,300 GW	二手汇编 / 实验室报告	2026 / 2024 年底	队列为发电+储能侧，非负载接入	C
各标的现价 / PE / 市值	实时行情接口	2026-07-24 收盘 (A 股 07-27 盘中)	美股为 TTM PE	A

数据冲突与口径处理：①先进封装信号并存「缺口收窄」与「产能售罄」两说，本报告按「相对缓解、绝对仍紧」表述；②2025 年四大 capex 实际值存在多个口径（3,880 亿与 4,100 亿两说），本报告采用分项反推的约 4,000 亿并显式标注区间性质；③内存产线调整仅有媒体转述，未采信任何由其派生的利润率数字；④「登顶前沿」仅在单项 arena 成立，全域为前四，本报告按后者陈述且不写精确名次分母；⑤本报告剔除了五处内部算术不自洽或无法溯源的流传数字（全球 token 月总量、开发者平台年化量、需求价格弹性区间、某云厂订单储备值、「当周跌近 10%」的周度口径），不以任何形式引用；⑥Agent 负载的 token 放大倍数、capex 中投向 AI 的具体百分比，因缺一手测算，本报告只作定性表述，不给数值。

本报告主动记录的一处自我修正：内部稽核初判某 A 股标的业绩预告数字「量级异常、疑为单位误植」，回查公司公告后确认该判定有误——真值自洽且来自一手公告，误差发生在「亿元 / billion」的换算环节。**过度保守也是一种失真**，故在正文第 7.3 节按真值陈述，并把反证逻辑从「数字存疑」修正为「基本面真、归因错」。

免责声明：本报告由涌现资本产业研究部独立出品，仅供研究交流，不构成任何证券的买卖要约或投资建议。报告所载数据来自公开渠道，我们力求但不保证其准确性与完整性。市场有风险，决策需自主。

涌现资本产业研究部 · 2026 年 7 月 27 日