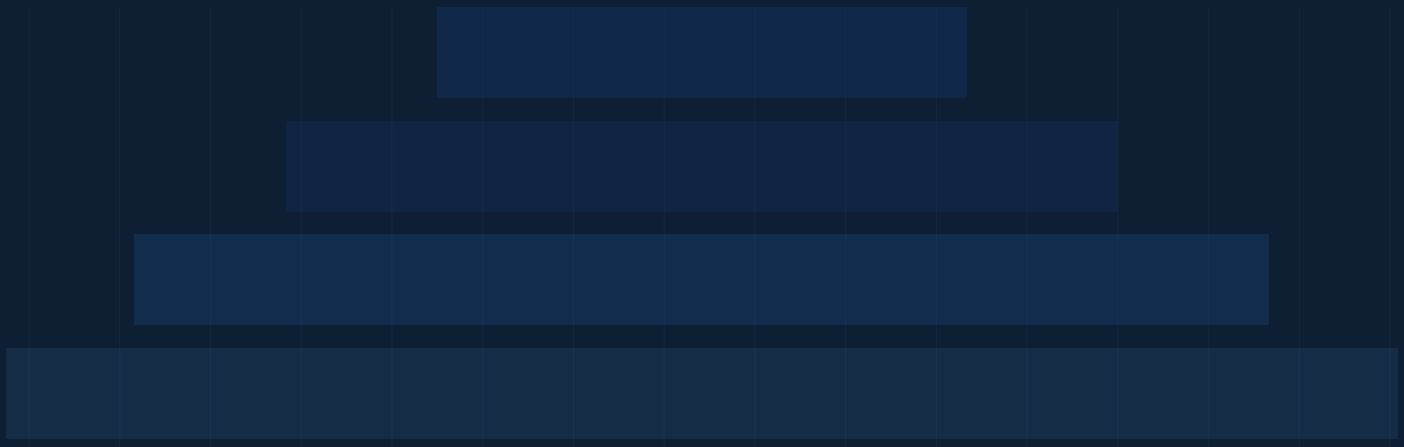
AI推理 · 企业级SSD · NAND · 存储软件

HBM之后，推理开始吞噬NAND

企业级SSD从“低价值冷存储”升级为GPU扩展内存；当单卡HBM装不下百万token的KV Cache，NAND需求第一次获得一个不依赖手机换机的AI锚点——本档自建“每百万token存储账单”与“企业NAND供需缺口”双模型，量化其成色、拐点与证伪线。



K Research ◆ GOLD · 第 007 期

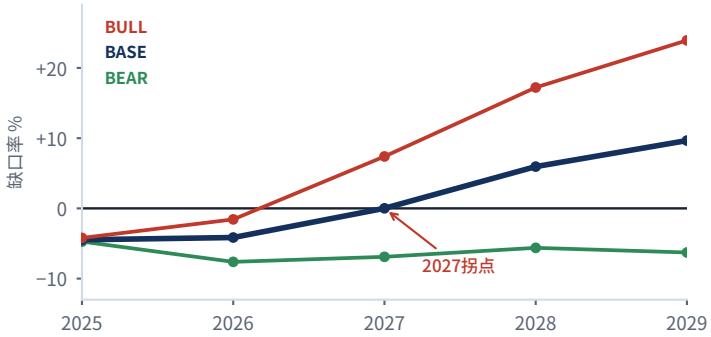
2026年6月19日 · 口罩哥知识星球独家卖方级研报

独家 · 深度 · 可证伪

01 核心观点

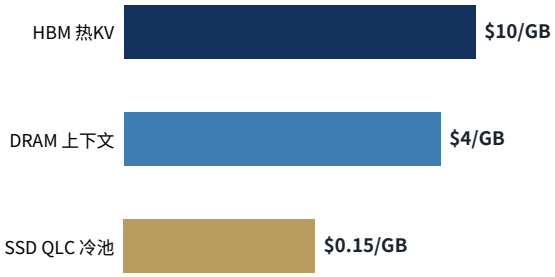
> 训练堆HBM，推理向下分层。当单卡HBM装不下百万token的KV Cache，企业SSD第一次成为GPU内存层的结构性组件，NAND需求获得AI锚——但成色取决于每GPU配置TB数与缓存命中经济。

基准情景：企业NAND缺口2027年由负转正



来源：K Research 自建缺口模型；E 为预测，数据截至 2026年6月

每GB成本：SSD比HBM便宜数十倍



来源：SiliconAnalysts/TrendForce/Solidigm；涨价期估计，对数刻度

三情景速览

BEAR • 25%	BASE • 50%	BULL • 25%
缓存压缩+更大DRAM，SSD offload停留小众；缺口持续为负	RAG与长上下文规模采用，缺口2027转正、2029达+10%	持久KV普及、单GPU SSD翻倍，2029缺口+24%结构性短缺

七条 Key Calls

01 推理时代内存账单从“全HBM”转向“HBM热缓存 + DRAM上下文 + SSD持久层”三层；单百万token为70B级稠密模型生成约293GB KV Cache，是模型权重的两倍，单卡HBM装不下——分层是工程刚需，不是选择题。

02 自建测算：把70%冷KV下沉SSD，每百万token内存账单由约2,930美元降至约590美元，省约80%；HBM越涨、节省越大。这是SSD offload的经济地基。

03 NAND首次出现可穿越手机周期的AI锚：基准情景企业NAND缺口2027年转正（0%→2029年+10%），AI占NAND总需求由2025年8%升至2027年22%、2029年38%。

04 卡位不在裸NAND，而在控制器固件 + GPU直连协议：NVIDIA CMX/Storage-Next + BlueField-4把以太网闪存变成池级KV缓存层，DOCA Memos/Dynamo软件栈才是真choke-point。

05 Kioxia明确押注“AI推理时代”：DC+企业收入占比目标超60%（FY28），三年每年资本开支约4,700亿日元、研发2,300亿日元，CM/GP/LC三线分别对应KV缓存/GPU内存扩展/RAG向量库。

06 分层决定赢家：高带宽TLC(CM9)吃KV热写、XL-FLASH SLC(GP)吃GPU直连超高IOPS、245TB QLC(LC9)吃RAG与历史冷池；低端消费NAND未必同步受益。

07 最脆弱共识：市场把“AI内存机会=HBM+DDR”。变体观点是增量价值向SSD分层迁移，NAND估值变量由手机出货转向“每GPU配置TB数 × 每token存储流量”。

02 产业链全景 · 七层穿透

AI从训练转向长上下文推理与Agent，存储价值链被重排：上游裸NAND晶圆仍是周期品，但真正卡位的环节上移到“控制器固件 + GPU直连协议 + SCM器件”。下图标注瓶颈节点（◆金角标）。

◆ = 七层穿透中议价权最高的瓶颈节点（卡脖子）

七层 Choke-point 穿透：议价权与替代窗口

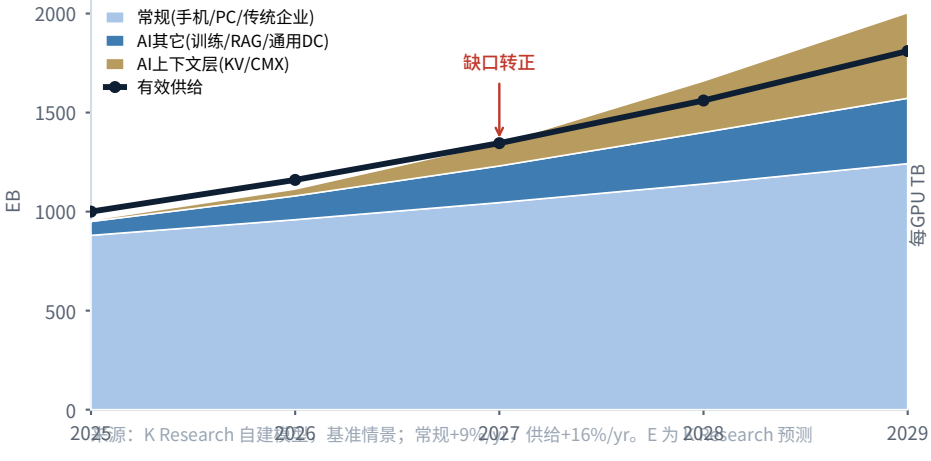
层	渗透对象	代表环节	议价权量化	议价权
L1	终端需求	长上下文/Agent推理token量	需求增速>40%（Kioxia指引DC NAND CAGR 20–46%）	低
L2	系统/整机集成	NVIDIA机架 + CMX/STX参考架构	高：协议与参考设计由NVIDIA定义	高
L3	关键器件	企业SSD：CM(TLC)/GP(XL-F)/LC(QLC)	中高：高容量QLC良率+高带宽产能集中	中高
L4	上游材料/IP	NAND晶圆 + 主控固件 + DOCA软件栈	裸NAND低、固件/软件高（价值上移）	分化
L5	设备/工具	键合(CBA)、高层数沉积/刻蚀	高：先进键合与高深宽比刻蚀交期长	高
L6	设备的设备	刻蚀气体/前驱体/EDA/IP核	高：隐形瓶颈，少数供应商	高
L7	最深不可替代	XL-FLASH/SCM低延迟器件 + 直连协议	最高：低延迟+512B粒度+协议绑定	最高

结论（数据→机制→结论→反驳→再结论）：①事实——CMX/Storage-Next由NVIDIA定义、CM/GP/LC由Kioxia锚定，软件与协议层集中度最高；②机制——裸NAND可被多供应商替代，但低延迟SCM器件(XL-FLASH)、先进键合(CBA)、DOCA/Dynamo软件栈替代窗口以年计；③结论——价值与议价权由晶圆上移至固件/协议；④反驳——若NVIDIA将协议开放或标准化(NVMe扩展)，L2议价权下降；⑤再结论——即便协议开放，L5–L7的SCM器件与键合设备仍是最深护城河，裸NAND环节最先被周期与新增产能稀释。

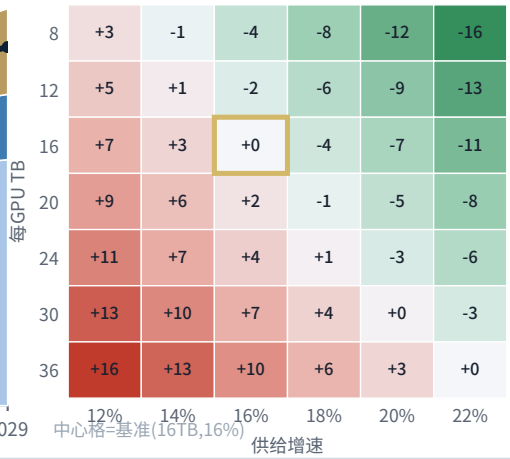
03 供需与价格 · 自建缺口模型

K Research 自建测算 · 企业NAND供需缺口（EB，2025–2029）

AI需求把NAND总需求推过供给线，2027交叉



敏感性：2027缺口率%



来源：K Research 自建测算。基准情景；常规+9%/yr，AI上下文层供给+16%/yr。E 为 K Research 预测　　中心格=基准(16TB,16%)

关键假设（节选）

• 2025 NAND供给基数	≈1,000 EB	TrendForce/IDC	估计
• NAND供给增速	15–17%/yr	TrendForce/IDC	事实
• 每Rubin GPU配置NAND	16 TB	产业(STX/ICMSP)	估计
• 上下文层(2026/27)	34.6 / 115 EB	Rubin出货×16TB	估计
• 常规需求增速	9%/yr	历史均值	估计

三情景输出（缺口率%）

情景	2026	2027	2028	2029
BEAR	-8	-7	-6	-6
BASE	-4	+0	+6	+10
BULL	-2	+7	+17	+24

读法：基准情景缺口2027年由−4%翻正至0、2029年+10%；牛市2029达+24%结构性短缺；熊市持续为负（−6%），对应SSD offload停留小众。三情景与P11情景章一一对齐。

数学链（可复现）

有效供给_t = 供给_2025 × (1+g_supply)^(t−2025)

上下文_t = GPU出货_t × 每GPU_TB ÷ 1e6 (TB→EB)

总需求_t = 常规_t + AI其它_t + 上下文_t

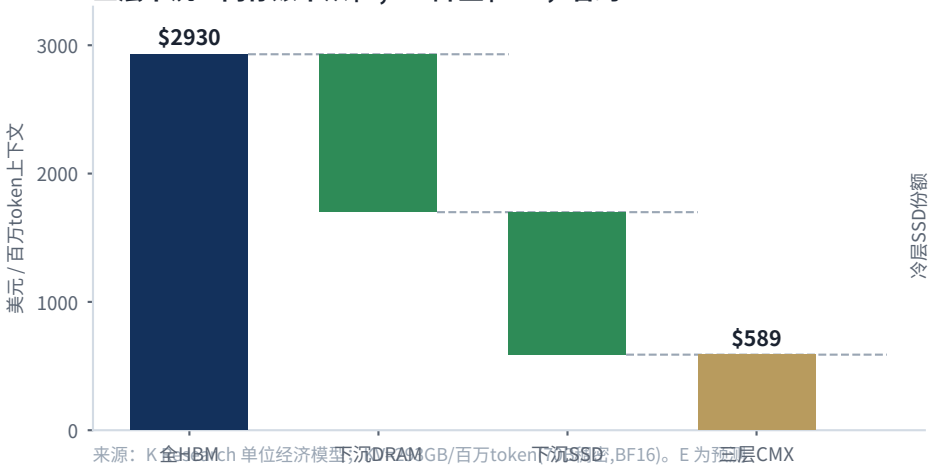
缺口_t = 总需求_t − 有效供给_t；缺口率 = 缺口/总需求

结论：①数据——上下文层NAND从2025年约7 EB跃至2027年115 EB（每Rubin GPU 16TB×出货）；②机制——常规需求温和（+9%）但AI增量陡峭，供给受NAND资本开支克制仅+16%；③结论——基准2027缺口转正，定价权由买方移向卖方；④反驳——若MLA等压缩把KV体积砍至1/4、或供给增速回到20%+，缺口推后；⑤再结论——压缩降低单token体积但放大可服务并发与上下文长度（杰文斯悖论），总bit需求方向不改，仅斜率变化。

04 财务透视 · 每百万Token存储账单

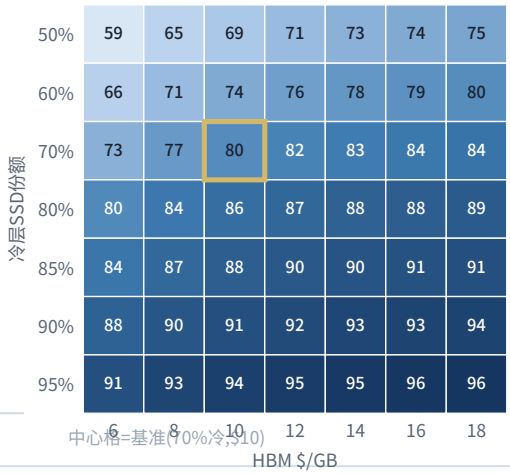
K Research 自建测算 · 每百万Token的KV Cache存储账单（美元）

三层下沉：内存账单从$2,930降至$589，省约80%



来源：K Research 单位经济模型；假设：KV=293GB/百万token，FP8,BF16)。E 为预测

敏感性：节省% vs 全HBM



中心格=基准(70%冷,$10)

分层成本假设（$/GB，2026涨价期）

· HBM3e	$8–10	SiliconAnalysts	事实
· 服务器DRAM(DDR5)	≈$4	TrendForce/Tom's	估计
· 企业TLC SSD	≈$0.40	渠道/估计	估计
· 企业QLC SSD	≈$0.15	Solidigm 122TB	事实

单位经济要点

· 单百万token KV=293GB，>2×模型权重(140GB)，>1×Rubin单卡HBM；这是“装不下”的根因。
· 每百万token上下文约拉动205GB NAND（70%冷层）——即“每token存储流量”锚。
· 厂商已在为KV复用定价：Anthropic/DeepSeek缓存读取=输入价10%（省90%），是offload价值的市场背书。

数学链

KV_GB = KV_MB/token × 1e6 ÷ 1024 = 293 GB

账单_arch = Σ_tier 份额_tier × KV_GB × 成本_tier

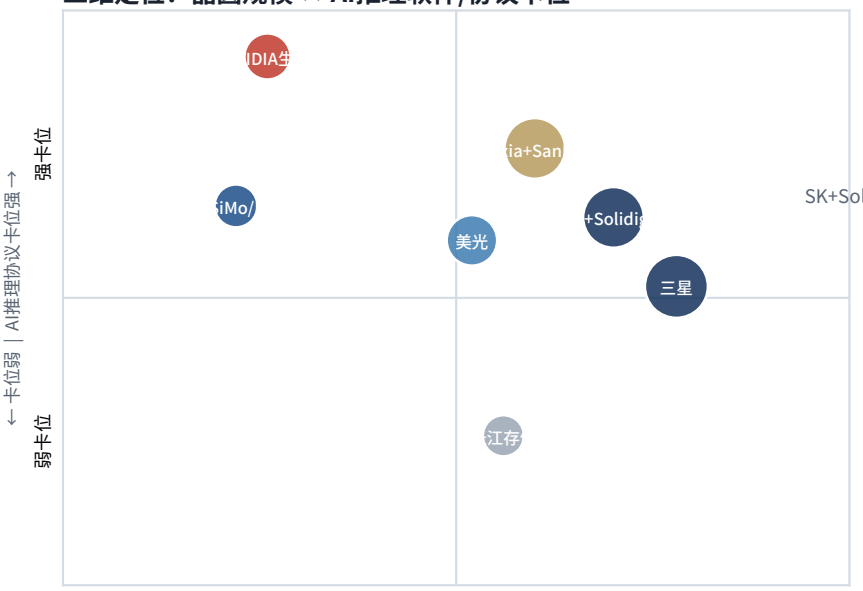
节省 = 账单_全HBM − 账单_三层 = $2,341 (79.9%)

结论：①数据——全HBM承载百万token上下文约$2,930，三层架构降至$589；②机制——SSD每GB比HBM便宜约50–70×，把命中率低的冷KV换成容量换延迟，单位账单塌缩；③结论——offload经济在HBM涨价期更划算（敏感性矩阵右上角省>90%）；④反驳——延迟、耐久(QLC 0.25–3 DWPD)与软件改造是成本，命中率低时SSD收益打折；⑤再结论——CMX/Dynamo把命中率与放置交给软件层，且冷层只需读密集低DWPD，QLC足矣——成本天平仍倒向分层。

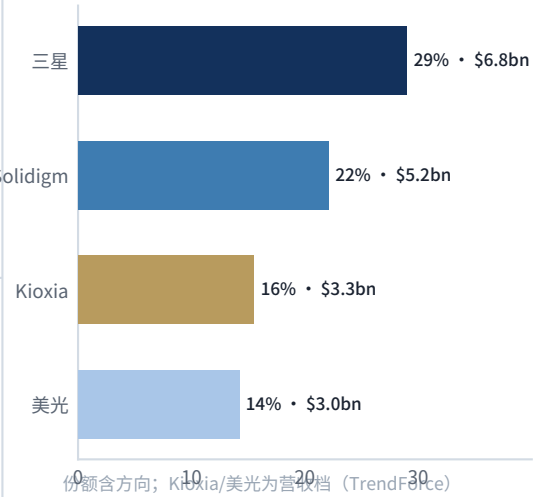
05 竞争格局

NAND六强份额(4Q25)：三星约29%居首，SK海力士+Solidigm 22.1%（营收$5.21bn）逼近，Kioxia($3.31bn)第三、美光($3.03bn)第四。但AI推理存储的胜负手不在份额，而在“晶圆规模×软件协议卡位”二维。

二维定位：晶圆规模 × AI推理软件/协议卡位



NAND份额与营收(4Q25)



份额含方向；Kioxia/美光为营收档（TrendForce）

来源：TrendForce、公司资料、K Research 卡位评分；气泡∝AI存储敞口

卡位拆解（choke-point视角）：Kioxia以CM(TLC高带宽)/GP(XL-FLASH SLC)/LC(245TB QLC)三线 + 10代BiCS，直接对位CMX三层，软件协议卡位最强（金色）；SK海力士借Solidigm在高容量QLC企业SSD份额领先、吃RAG/冷池；三星晶圆规模与HBM协同最强但推理存储协议参与相对弱；美光以G9 PCIe Gen6 DC SSD切入、且在财报点名“向量库与KV cache offload驱动NAND bit需求加速”。

真正的隐形赢家在控制器与固件：超高IOPS(100M)、512B粒度、GPU直连DMA、低写放大与掉电恢复，全靠主控固件与DOCA/Dynamo软件；Silicon Motion、Marvell、群联等主控厂与NVIDIA BlueField-4构成裸NAND之上的高毛利夹层。反驳——若云厂自研主控(类似自研网卡/DPU)，夹层被内部化；再结论——自研需多年验证与生态，短中期协议绑定仍使NVIDIA+头部主控占据议价高地。

06 技术路线

推理存储不是一种SSD，而是按延迟/IOPS/容量分三层落位的器件谱系。下图为分层路线与GPU直连协议时间表（事实层为已发布产品与NVIDIA官方里程碑）。

GPU直连协议时间表（NVIDIA官方）

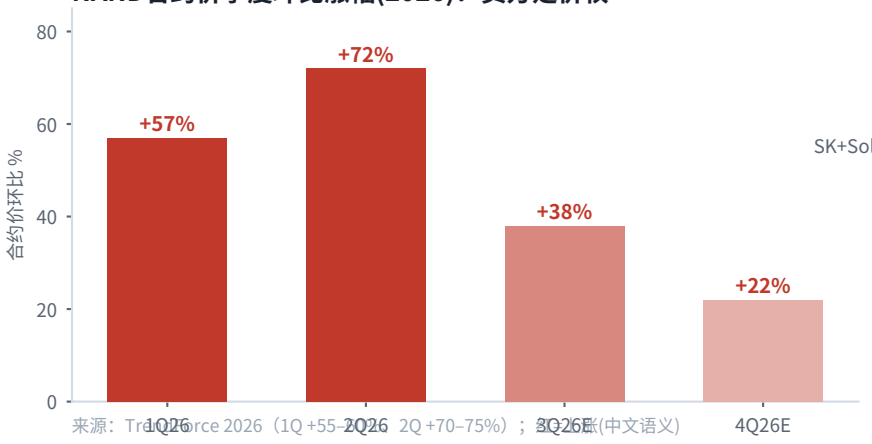
CES'26	GTC'26	'26H2	Rubin
ICMSP三层KV标准化	BlueField-4 STX / CMX	STX量产出货	每GPU≈16TB NAND

S曲线位置：XL-FLASH/SCM处于陡升前夜——2026年10M IOPS、2027目标100M IOPS，512B访问粒度是为GPU细粒度读取定制；这是Optane退场后重生的低延迟非易失层。TLC(CM9)与QLC(LC9)已在量产爬坡区。反驳——SCM曾因Optane败北证明“中间层”叫好不叫座；再结论——本轮不同点在于：需求由GPU直连协议(CMX/Storage-Next)自上而下定义、且KV Cache是天然的“热-温-冷”可分层负载，载体从CPU缓存变为GPU内存扩展，使用场景比Optane时代明确得多。

07 资金与定价 · 估值

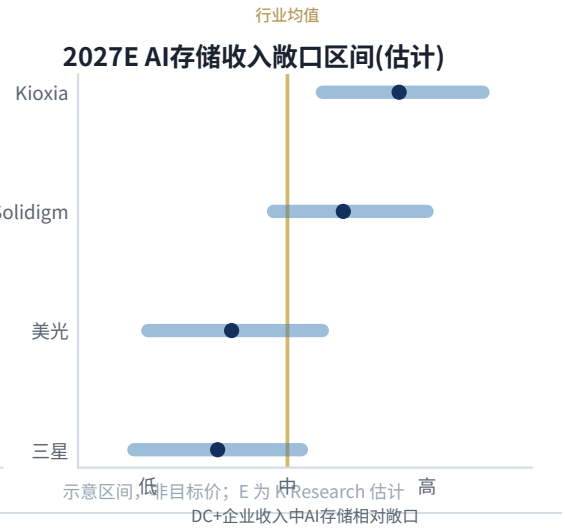
涨价由“产能向服务器/HBM/企业SSD结构性再分配”驱动，非临时扰动。合约价连续跳涨，云厂转向3–5年LTA锁量，定价权回到卖方，并以长约平滑NAND周期性。

NAND合约价季度环比涨幅(2026)：卖方定价权



来源：TrendForce 2026（1Q +55–60%；2Q +70–75%）；3Q26E(中文语义)

2027E AI存储收入敞口区间(估计)



示意区间，非目标价；E 为 KR Research 估计

资金与定价机制：①事实——三星/SK海力士将大客户合约从1年转为3–5年LTA（微软、谷歌、亚马逊、Meta、阿里、字节），经营利润率升至40–50%；Kioxia预计FY26 Q1转为净现金，三年资本开支约1.4万亿日元；②机制——LTA锁量锁价，把过去“量价齐杀”的NAND周期改造为“可见度高、波动低”的合约簿；③结论——具备云厂长约+高容量QLC良率优势者盈利波动下降、估值中枢上移；④反驳——长约价仍可能在下行周期被重谈或塞条款，且新增产能(2027–28)可能重启过剩；⑤再结论——本轮capex克制(NAND增速仅15–17%)+AI需求结构化，使过剩风险后移，但2028–29仍是最大证伪窗口。

估值含义（非投资建议）：当前价对NAND厂普遍隐含“AI推理存储为期权而非主线”。本档变体观点是，若基准缺口路径兑现，企业NAND将从“周期附庸”重定价为“AI产能配套”，估值变量由手机出货切换为“每GPU配置TB数 × 每token存储流量”。读者可用本档双模型自行代入：给定每GPU TB与命中经济，推回各厂AI存储收入弹性与隐含预期差。本节仅提供框架与事实，不构成任何买卖建议。

08 反共识深论

市场共识

AI内存机会 ≈ HBM + DDR。NAND是低价值冷存储，跟随手机/PC换机周期；推理对存储的需求增量有限，KV Cache应尽量留在HBM/DRAM。

K Research 变体观点

推理(非训练)是增量主战场。HBM只承载最热数据，长上下文、Agent状态、向量索引与历史KV向SSD分层；NAND估值锚从“手机出货”转向“每GPU配置TB数 × 每token存储流量”。

市场已为KV复用定价：缓存读取=输入价10%

来源：Anthropic/DeepSeek 缓存读取定价(约90%折扣)；DeepSeek 3FS盘上缓存TTFT 13s→0.5s

逻辑桥：若KV Cache毫无持久化价值，主流API不会把“缓存命中”定价为输入价的10%（省90%）。这是买卖双方对“复用已算KV”支付意愿的市场背书——而复用的载体，正在从昂贵的HBM/DRAM下沉到SSD。DeepSeek把上下文缓存放到分布式盘(3FS)，把首token时延从13秒压到0.5秒、token成本降一个数量级。

五段链：攻击最脆弱假设（“SSD offload经济不成立”）

数据	单百万token KV=293GB>2×权重；SSD每GB便宜50–70×；CMX/ICMSP把三层KV标准化；厂商已为缓存命中打9折。
机制	推理的KV是天然热-温-冷负载，冷层命中率低、读密集、可容忍ms级延迟——正是QLC(0.25–3 DWPD)的甜区。
结论	把冷KV从HBM换到SSD，单位内存账单塌缩约80%，且释放HBM/电力预算给算力（CMX宣称省电5×）。
反驳	压缩(MLA/V4将KV砍至~2%)、更大DRAM、HBM4扩容，可能让SSD层可有可无；Optane中间层曾失败。
再结论	压缩降单token体积却放大可服务上下文与并发(杰文斯悖论)；HBM/DRAM单GB仍贵1–2个数量级。总bit需求方向不改，SSD层由协议自上而下锁定，非自发中间层——这正是与Optane的关键差异。

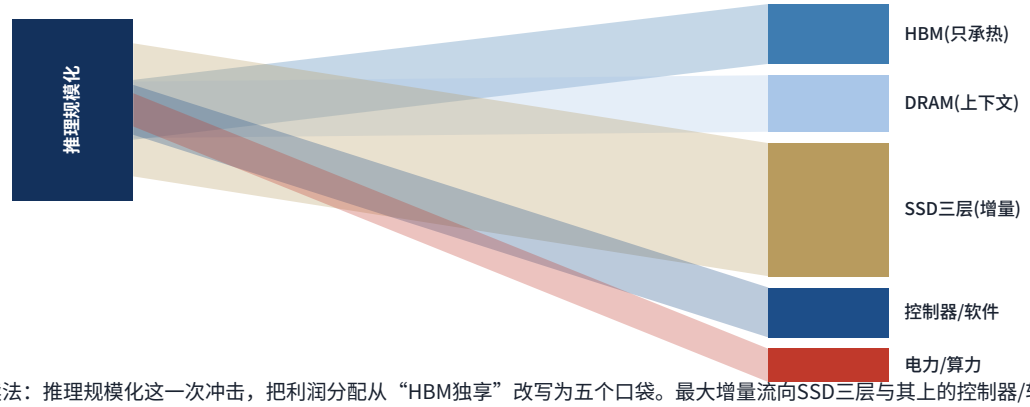
对立视角（evenhanded）：看空者有三点硬理由——其一，模型架构进步(MLA/压缩注意力)可能把KV体积压到边际，使SSD层经济性消失；其二，企业SSD增量可能主要来自传统数据库/HDD替换而非纯AI新增bit，AI锚被高估；其三，SCM/中间层历史失败率高。这些都是真实风险，本档以P11证伪信号逐条挂钩监测。

09 二阶与三阶效应

把“推理吞噬NAND”沿产业链向外推两到三跳：受益与受损者、它们的对手与客户如何被改变、再触发什么连锁。

一阶 · 直接	二阶 · 传导	三阶 · 连锁
● 企业NAND/SSD厂(Kioxia/美光/SK+Solidigm)获AI增量需求	● 价值从裸NAND上移到控制器固件+软件栈，毛利夹层放大	● NAND资本开支超级周期→2028–29潜在过剩与价格反转风险
● GPU直连协议方(NVIDIA CMX/BlueField-4)定义新存储层 →	● 高容量QLC良率+云厂LTA者盈利波动下降、估值中枢上移 →	● 云推理成本结构中“存储占比”上升，改写token经济
● “AI内存=HBM+DDR”的单一叙事被补全为含SSD三层	● DRAM面临KV从DRAM向SSD迁移的部分替代压力	● 低端消费NAND被挤压、价格外溢，手机/PC成本上升
	● 近线HDD受QLC高容量替代与AI冷数据双向拉扯	● SCM/低延迟非易失层(XL-FLASH类)在Optane退场后重生
	● 电力预算从存储重新分配给GPU(CMX宣称省电5×)	● HBM4扩容 vs SSD层的边界将被重新谈判

传导链（桑基式）：一次冲击如何重排五个口袋的利润



读法：推理规模化这一次冲击，把利润分配从“HBM独享”改写为五个口袋。最大增量流向SSD三层与其上的控制器/软件夹层；HBM仍受益但只承热数据；DRAM被部分替代；电力预算回流算力。这解释了为何“AI=只买HBM”的组合会系统性低估存储与软件环节。

10 情景与风险

三情景矩阵（概率合计100%，与双模型输出对齐）

情景	概率	触发条件	2029缺口	AI占NAND	每GPU TB	定价含义
BEAR	25%	压缩(MLA/V4)+更大DRAM/HBM4，SSD offload停留小众；云厂不提SSD配置	−6%	32%	8TB	NAND随手机周期，无AI重定价
BASE	50%	RAG与长上下文规模采用，覆盖部分高并发；CMX逐步铺开	+10%	38%	16–20TB	缺口转正，卖方定价权回归
BULL	25%	2027前持久KV普及、单GPU SSD翻倍、Agent爆发	+24%	44%	28–36TB	结构性短缺，AI产能重定价

证伪信号（出现即下调结论置信度）

① **主流云厂连续3个季度不提升GPU节点SSD容量（每GPU TB停滞≤8TB）**
监测：月度/季度机型拆解、招标配置

② **持久KV Cache的单位token节省 < 10%**
监测：厂商/论文基准、API缓存定价

③ **企业SSD增量主要来自传统数据库/HDD替换，而非AI新增bit**
监测：厂商终端拆分、出货结构

④ **NAND供给增速重回20%+ 且2028产能集中释放**
监测：各厂capex/产能爬坡披露

数据最脆弱处（第二轮红队）：全篇结论对“每GPU配置TB数”与“冷层命中/节省率”两个估计级变量最敏感（见两张敏感性矩阵）。二者若同时走弱（≤8TB且节省<10%），基准缺口将退回负值、AI重定价证伪。故本档把这两项列为最高优先监测，宁可标注未知，不以单点数字伪装确定性。

11 方法论与数据附录

模型搭建逻辑

本档搭两个互锁模型。模型一（单位经济）以“每百万token上下文”为单元，按KV体积×分层成本，量化全HBM vs 三层CMX的内存账单与节省；模型二（供需缺口）自下而上以“GPU出货×每GPU TB”估上下文层NAND需求，叠加常规与AI其它需求，对比受capex克制的有效供给，得季度/年度缺口。关键简化：①以70B级稠密GQA为KV体积基准，未对全模型谱系加权；②供给以行业总量近似，未逐厂逐线建模；③价格用2026涨价期水平，未对未来价格路径内生化。两模型共用同一套bull/base/bear假设并与情景章对齐。

关键一手信源（已读并显式引用）

- **Kioxia Holdings《AI推理时代增长战略》Investor Day** 2026-06-02
  DC+企业收入>60%(FY28)、capex≈4,700亿/研发≈2,300亿日元、CM/GP/LC三线、10代BiCS今夏送样
- **NVIDIA CMX Context Memory Storage Platform 官网** 2026-06
  BlueField-4 + DOCA Memos + Spectrum-X，把以太网闪存做成池级KV缓存层；宣称吞吐/能效各达5×
- **Micron FQ2'26 Prepared Remarks（官方IR）** 2026-03
  NAND营收$5.0bn(+169%)；点名向量库与KV cache offload驱动DC NAND bit需求加速；G9 PCIe Gen6 DC SSD量产
- **DeepSeek API Docs：Context Caching on Disk** 2025–2026
  盘上缓存把token成本降一个数量级、TTFT 13s→0.5s——KV持久化价值的一手背书
- **TrendForce 多篇新闻稿** 2026-01~06
  NAND合约价1Q+55–60%/2Q+70–75%；企业SSD成最大应用；供给+15–17% vs 需求+20–22%；3–5年LTA

数据口径、级别与已知局限

口径：所有金额为美元或日元名义值，数据截至2026年6月。级别——“事实”取自公司披露/官方文档/TrendForce统计；“估计”为K Research基于一手锚点的推演（每GPU TB、分层份额、价格水平、2025供给基数）；传闻线索（“某云厂签长约”“某型号供不应求”）仅用于反查，不入base核心计算。局限与未知（诚实声明）：①真实KV驻留时长、云厂单GPU实际SSD配置、持久缓存命中率为最大未知；②企业SSD中AI工作负载占比、QLC实际DWPD与长约价格公式未公开；③$147bn vs $59bn等NAND市值口径在各机构间存在冲突，本档取方向而非点值。凡达不到颗粒度处，宁写未知，不写大概。

12 结论与行动

结论重申

推理规模化让企业SSD第一次成为GPU内存层的结构性组件。基准情景下企业NAND缺口2027年转正、AI占NAND需求2029年达38%，NAND获得首个可穿越手机周期的AI锚。但这是“条件性重定价”——成色押在每GPU配置TB数与缓存命中经济两个变量上；卡位在控制器固件与GPU直连协议，而非裸NAND。

关注时间表（催化与验证窗口）

- **2026年夏** Kioxia 10代BiCS送样→CM系列；观察KV带宽与良率
- **2026 H2** NVIDIA BlueField-4 STX量产出货；CoreWeave/Oracle等首批落地
- **2026 Q3–Q4** NAND合约价见顶节奏；美光/SK/Kioxia财报中DC SSD与AI bit占比
- **2027** GP系列冲100M IOPS；每GPU TB是否翻倍——基准缺口转正的胜负手
- **2028–2029** 新增产能释放 vs AI需求：最大过剩/证伪窗口

行动框架（非投资建议）：把本档双模型当作可复用脚手架——代入你自己的每GPU TB、冷层命中率与分层成本，即可推回各厂AI存储收入弹性与“市场定价 vs 测算”的预期差，并用P11四条证伪信号做月度校验。读完应获得的不是一个点位，而是一套可证伪、带敏感性的分析框架。