

规格通缩：当上游短缺开始削顶下游设计

规格通缩——当上游短缺开始削顶下游设计

旗舰 GPU 的内存减配是两笔独立的账：封装执行占大头，内存短缺占后半段。把两段合成为一个百分比记到内存头上，会同时高估短缺的破坏力和低估机架端的时间风险。真正被短缺量化定额的不是 HBM，是 CPU 侧的 LPDDR。

涌现资本产业研究部 · 2026 年 8 月 10 日

30秒结论卡

一句话判断：内存短缺已经硬到能改写旗舰产品的物料清单——这是瓶颈四特征里分量最重的一条证据；但这一代旗舰的容量下修是**两段独立的账**，先进封装占前半段、内存供给占后半段，两段的受益方与推翻条件完全不同。

项目	内容
机会状态	● 观察为主——卡点证真、但价格变化率已过峰，量价须两笔独立记账
市场错在哪	把「封装执行事故」与「内存短缺」合成一个数字，于是既高估了短缺对下游的伤害，也忽略了真正被定额配给的是 CPU 侧 LPDDR 而非 HBM
重点标的	MU \$877.57 ● · SK 海力士 ₩1,436,000 ● · 三星电子 ₩231,500 ● · SNDK \$1,212.21 ● · NVDA \$223.96 ●
最大反证	若 Kyber 机架顺延至 2028 兑现、2027 年实际交付的仍是 Oberon 级机架，则一切以 NVL576 为口径的价值量测算须整体后移一年

现价口径：yfinance 与 EODHD 双源交叉，2026-08-10 02:05 UTC（对应 08-07 美股收盘、韩股 08-07 收盘）。

投资要点

一句话框架：短缺出清有两条路——涨价与降规格。当涨价的一阶导见顶，出清就切换到规格，而规格通缩的账要按驱动因素分开记。

观点	方向
卡点已硬到能反向改写下游旗舰设计，这是瓶颈证真的最强形态	●
容量下修分两段：封装执行占前半段，内存供给占后半段，受益方与推翻条件不同	●
真正被量化定额配给的是 CPU 侧 LPDDR（仅获约 60% 需求量），不是 HBM	●
合约价变化率已过峰（QoQ 90~98% → 13~18%），价格水平仍在高位	●
DRAM 系扩容路径全部同吃一片晶圆，结构上不是泄压阀；只有 NAND 系来自另一晶圆池	●
供给缓解落在 2028—2029，2027 年的供给曲线已基本锁死	●
上游涨价已在下游利润表上显形（元件成本被点名压缩毛利率指引）	●

一、钱流：变化率已经见顶，价格水平还在高位

这一轮内存周期最容易读错的地方，是把「涨价还在继续」与「涨价还在加速」当成同一件事。合约价的环比序列已经给出了明确答案：

季度	传统 DRAM 合约价 QoQ	NAND 合约价 QoQ	读数
2026Q1	+90~98%	—	● 加速期顶点（当季 DRAM 产业营收 +81% QoQ）
2026Q2	+58~63%	—	● 一阶导转负
2026Q3	+13~18%	+10~15%	● 变化率见顶，水平仍在涨

收敛有两个来源：消费端已触及可负担性上限；美系云厂的多年期长约限制了原厂对签约客户的提价空间。价格的水平仍在历史高位，价格的变化率已经过峰——这两条必须分开记账，混算就会同时得出「周期见顶」和「短缺加剧」两个互斥结论。

供给侧的算术则指向相反方向。HBM 占前三大厂 DRAM 晶圆投片比例为 18%（2025 年底）→ 22%（2026）→ 约 30%（2027 年底），但只占 DRAM 位元供给的 8% → 9% → 13%。反解出的单位位元耗片强度约为普通 DRAM 的 2.3 倍以上：**每多做一片 HBM，要从普通 DRAM 池子里抽掉两倍多的产能**。2027 年 HBM 位元出货增长 50~60%，仍不足以跟上需求；DRAM 充足率 2026 年约 -1% ~ -2%，2027 年缺口继续扩大。

供给缓解的时点落在 2028—2029，不在 2027。SK 海力士 8 月 7 日董事会批准约 ~~₩~~54 万亿扩产，其中占大头的龙仁 Y2（~~₩~~35.2 万亿）2027 年 7 月才动工、首座洁净室 2029 年 6 月；清州 M17（~~₩~~19.1 万亿）2027 年 2 月动工、2028 年 12 月洁净室。美光 FY26 资本开支约 \$27B，FY27 季度资本开支超 \$100 亿且增幅逾半来自厂房建设，ID1 首片晶圆 2027 年年中、纽约 2H2028。**2027 年的供给曲线，在今天已经基本锁死。**

二、价值链五层：钱流经过哪几道闸

层	环节	现状	状态
L1 需求源	四大云厂 2026 日历年资本开支合计约 \$725B（微软 \$190B、Alphabet \$195—205B、Meta \$125—145B、亚马逊约 \$220B）	无一家减单，亚马逊明确把上调归因于内存成本	●
L2 一级供应商	NVIDIA 机架平台（Oberon / Kyber）	内存成本挤压设计自由度，路线图重排	●
L3 真卡口	HBM / 传统 DRAM 晶圆产能、先进封装（CoWoS-L 及其面板级后继）	双卡口并存，且互相争夺同一批旗舰机型	●
L4 可投代理	三大原厂 + NAND 系替代层 + 内存接口/控制器 IP	前瞻估值仍低，但拥挤度刚经历一轮清洗	●
L5 已拥挤 / 未定价	已定价：DRAM 涨价本身；未定价：规格通缩的受益方与受损方分布	—	●

微软管理层把 2026 年约 \$250 亿的资本开支增量直接归因于元件涨价。**上游的涨价已经变成下游资产负债表上的一个可识别科目——这是本轮周期与历次内存牛市最大的结构性差别。**

三、两笔账：一次减配，两个不同的原因

旗舰 GPU 的单封装内存容量，从最初公布的 1TB 走到当前主线方案的 192GB。这一段路由两个互不相干的事件叠加而成，把它们合成一个百分比是本轮最容易犯的归因错误。

段	容量变化	驱动因素	时点	受影响的层
第一段	1,024GB → 约 384GB	先进封装执行： 四 die + 16 堆栈的 2×2 大 die 阵列造成基板严重翘曲、计算 die 与基板失去接触导致电性失效，改回双 die + 8 堆栈	2026 年 6 月底	封装 / 基板 / 机架
第二段	384GB → 192GB（-50%）	内存供给： 2027 年 DRAM 短缺挤占 HBM 晶圆配额，叠加 12 层 HBM4E 的验证与良率爬坡不确定	2026 年 Q3 起	HBM 原厂 / 单机内存价值量

官方基准至今未被新规格表替代：四 die、1TB HBM4E、16 堆栈、100 PFLOPS FP4、NVL576 机架 600kW、2027 下半年——这套规格在 2026 年 3 月 16 日仍以实物计算托盘形式公开演示。产业界的建模基准则早在 6 月初就已按每 GPU 384GB 推进，两条线在 6 月分叉。

当前的评估范围是四选一：12 层 HBM4E、8 层 HBM4E、12 层 HBM4、8 层 HBM4，最终规格未定。主线方案落在 8 层 HBM4、192GB，峰值算力不变，芯片级功耗约 1,800W。容量映射有一手规格背书：12

层 HBM4E 为 48GB/堆栈 $\times 8 = 384\text{GB}$ ，8 层 HBM4 为 24GB/堆栈 $\times 8 = 192\text{GB}$ 。相对现役 Rubin 的 288GB，主线方案下修 33%。

分账的投资含义：第一段的钱不流向内存厂，流向的是封装与基板的执行风险；第二段才是内存短缺的真实读数。把两段合并计算，会把封装事故的责任记到内存供给头上，从而在 HBM 需求侧做出过度悲观的推论。

真正被定额配给的是 CPU 侧内存

这一代唯一带配额量化的硬证据不在 GPU 上，在 CPU 上：Vera CPU 的 SOCAMM 模组由 192GB 腰斩至 96GB，机架级 CPU 内存从约 55TB 降到约 28TB，而同机架的 HBM 容量维持 20.7TB 不变。原因写得很直白——2027 年的初步分配下，只能拿到估计 LPDRAM 需求的约 60%，属供给驱动的架构妥协，而非需求端调整。

HBM 的容量在被讨论，LPDDR 的容量已经被砍掉一半。市场的注意力全在前者，定额配给的实锤在后者。

四、规格通缩不是孤例，是本轮短缺的下游结算方式

同一时间窗内并行发生的四件事，指向同一个机制：

1. 旗舰 GPU 的 HBM 配置由单一基准扩为四方案并行评估；
2. 数家云厂同步评估下调下一代自研 ASIC 的 HBM 容量；
3. 云厂与服务器 OEM 在 2026 上半年已下调 RDIMM 配置容量；
4. CPU 侧 SOCAMM 容量直接减半。

当短缺无法通过价格出清时，它就通过**规格**出清。这是瓶颈证真的最强形态：卡点硬到能反向改写下游旗舰产品的设计，而不只是抬高其成本。对内存厂而言，这意味着单机用量的天花板被按住；对下游而言，这意味着同样的资本开支买到的有效内存容量在缩水。

五、卡点评分卡与四象限

指标	HBM 晶圆产能	先进封装	LPDDR / SOCAMM	NAND 系 (QLC SSD)
短缺强度	🔴 极强 (2027 缺口扩大)	🔴 极强 (翘曲问题未解)	🔴 极强 (仅获约 60% 需求量)	🟡 中 (2027 充足率转正)
替代难度	🔴 无同类替代	🔴 面板级方案时点未明	🟡 可降规不可替代	🟢 供应商众多
认证周期	🔴 12 层验证爬坡不确定	🔴 机架级重新设计	🟡 已量产	🟢 已在货
集中度	🔴 三家	🔴 一家为主	🟡 三家	🟡 分散
定价权	🟢 归原厂	🟢 归封装方	🟢 归原厂	🟡 分散
可投性	🟢 三家均上市	🟡 多为代工方	🟢 同三家	🟢 上市标的多

四象限落位：HBM 晶圆产能与 LPDDR = 真卡点且可投（核心）；先进封装 = 真卡点但可投代理有限（proxy 或监控）；NAND 系 = 非最硬卡点但现金流被低估（价值修复）；CXL 内存池化 = 叙事热于兑现（回避追高）。

六、泄压阀检验：五条替代路径，只有一条过三关

买不到 HBM，需求会去哪里？这里有一条决定性的物理约束：**CXL 扩容、LPDDR、SOCAMM 全部吃同一片 DRAM 晶圆**，因此结构上不可能成为短缺的泄压阀——它们与 HBM 抢的是同一份产能。只有 NAND 系来自另一个晶圆池。2027 年 DRAM 与 NAND 的充足率正式分道扬镳：DRAM 缺口继续扩大，NAND 转正。

路径	技术定位	时间关	认证关	产能关	净判定
QLC SSD / KV cache 卸载	把上下文搬到 NAND	2026H2 伙伴供货	15 家具名存储伙伴、8 家具名早期客户	另一晶圆池	● 唯一三关齐过
LPDDR / SOCAMM2	替代 RDIMM, 降功耗与体积	已量产 (192GB、1cnm)	已认证	✗ 同吃 DRAM 晶圆	● 有效但不解渴
HBF (高带宽闪存)	NAND 堆栈做带宽层	样品 2H26、系统级 1H27、商业化 2027 年底—2028	首版规格 2026-08-03 发布	需新建	● 对的层, 错的时间
CXL 内存池化	主机间共享内存池	控制器 3Q26 送样、设计案 2027 出货	未规模验证	同吃 DRAM 晶圆	● 2026 年无承接力
软件侧 (KV 压缩 / 稀疏化)	降单位 token 内存占用	已在用	—	—	● 无证据显示削减了总采购量

CXL 的带宽与 HBM 差约两个数量级，其定位是扩容与共享，不是替代带宽层。HBF 的带宽区间为 0.4—3.0 TB/s、单堆栈上限 512GB、容量为同成本 HBM 的 8—16 倍，是结构上最对的答案——但全面放量要到 2030 年前后，对 2027 年的缺口毫无帮助。

NAND 这条路的代价是它自己也在涨价：2026Q2 NAND 合约价预估环比 +70~75%、企业级 SSD 1Q26 约 +80%，到 3Q26 收敛至 +10~15%。这是一条「贵但买得到」的路，宽松要等 2027 下半年。

七、公司卡

美光 (MU · \$877.57 · PE 19.80 / 前瞻 6.10 / PB 9.84 · ●)

位置：三大 DRAM 原厂之一，HBM 份额目标对齐自身 DRAM 份额。为什么重要：FQ3 FY26 营收 \$41.46B、GAAP 毛利率 84.6%、营业利润率 80.4%、EPS \$24.67，FQ4 指引 \$50.0B ± \$1.0B、毛利率约 86%。为什么被低估：前瞻 PE 6.10，HBM3E 与 HBM4 已售罄至 2027 日历年、需求延伸至 2028。证据：HBM4 累计营收逾 \$10 亿，12 层爬坡速度为 HBM3E 的两倍。估值状态：低前瞻倍数 + 极端 PB 9.84，周期位置高于估值位置。催化：FY26Q4 业绩、HBM4 供价定案。反证：Citi 已把 2027 年 EPS 的适用倍数由 10x 下调至 8x，目标价 \$1,400 → \$1,150。

SK 海力士 (000660.KS · ₩1,436,000 · PE 6.23 / PB 3.95 · ●)

位置：HBM 份额第一，出货占比三家最高。为什么重要：2Q26 营收 ₩79.3 万亿、营业利润 ₩60.5 万亿 (利润率 76%)，DRAM ASP 环比约 +30%、位元环比高个位数增长，3Q26 指引 DRAM 位元再增约 10%。为什么被低估：表观 PE 6.23 含巨额非经营性利得，剔除一次性并按约 23.5% 实际税率还原后的 trailing PE 约 8.8—9.1 倍——仍是周期高盈利下的低倍数。证据：已与约 10 家客户完成长期协议。估值

状态：PB 3.95 为本组最合理。催化：¥54 万亿扩产落地节奏。反证：2Q26 营收与营业利润双双低于共识，当日跌 9.6%——创纪录的绝对水平已经不足以满足预期。

三星电子 (005930.KS · ¥231,500 · PE 10.17 / PB 2.54 · ●)

位置：唯一同时具备 DRAM、NAND、代工与自有 base die 能力的厂商。为什么重要：2Q26 营收 ¥171.5 万亿、营业利润 ¥89.5 万亿，双创历史新高；半导体部门贡献集团营业利润的 99.7%。为什么被低估：PB 2.54 为本组最低，手机部门亏损压制了整体估值。证据：存储业务营收与营业利润双创新高。估值状态：分部估值错配显著。催化：HBM4 份额验证。反证：非存储部门的亏损可能持续拖累合并报表。

闪迪 (SNDK · \$1,212.21 · PE 16.43 / 前瞻 5.71 / PB 11.48 · ●)

位置：NAND 系替代层的最直接代理，同时是 HBF 标准的推动方。为什么重要：FQ4 FY26 营收 \$8.965B、non-GAAP EPS \$39.25，数据中心营收 \$2,977M、环比 +103%、占比由 12% 跃升至 38%。为什么被低估：前瞻 PE 5.71。证据：自估 NAND 市场 CY2026 超 \$3,000 亿、CY2027 接近 \$5,000 亿，数据中心占比由约 30% 升至约 50%。估值状态：PB 11.48 为本组最高，反身性风险最大。催化：HBF 系统级方案 1H27。反证：FQ1 FY27 营收指引 \$10.3—10.8B 低于预期，毛利率指引由 84.6% 降至 83—85%，公司点名「更高的元件成本」——这是上游涨价开始侵蚀下游利润的第一手证据，也是 8 月 6 日韩股单日暴跌的直接触发。

英伟达 (NVDA · \$223.96 · PE 34.30 / 前瞻 22.43 · ●)

位置：内存需求的定义方，也是内存涨价的最大承受方。为什么重要：FQ1 FY27 营收 \$81.6B、GAAP 毛利率 74.9%，FQ2 指引营收 \$91.0B ± 2%、毛利率 74.9% ± 50bp。减配读作毛利率防御——在内存成本占机架物料清单四分之一以上的结构下，规格是唯一还能调节的变量。估值状态：前瞻 22.43 倍，为本组中唯一不便宜的标的，但也是唯一不靠周期价格赚钱的。催化：FQ2 FY27 财报 2026-08-26。反证：Kyber 机架据报因正交背板的中板 PCB 难题推迟逾 12 个月至 2028，双机柜过渡方案因最大客户不接受而取消。

八、受益链与情景

大公司受益链：需求源（四大云厂约 \$725B 资本开支）→ 直接受益（三大 DRAM 原厂：量价两本账，价的贡献远大于量，但量并未停滞）→ 二阶受益（NAND 系承接被挤出的容量需求：SNDK 及企业级 SSD 供应链）→ 隐形受益（内存接口与控制器 IP：规格分叉越多，授权与验证价值越高）→ 伪受益（纯 CXL 概念标的，2026 年无实际承接力）→ 承压方（下游整机与模组：元件成本已在毛利率指引中显形）。

情景	核心假设	结果	方向
 牛市	减配落在 8 层 HBM4E (256GB)、Kyber 按期 2027H2、DRAM 缺口延续至 2028	量价齐升，原厂 2027 盈利再上台阶	三大原厂  
 基准	主线定案 8 层 HBM4 (192GB)、Kyber 顺延至 2028、2027 年交付 Oberon 级机架	单机内存价值量下修，行业位元增长承接，价格变化率继续收敛	原厂  · NVDA 
 熊市	2027 下半年 DRAM 合约价转负、云厂资本开支增速放缓	周期定价从「涨价」切到「见顶」，高 PB 标的先受损	SNDK  · 原厂 

标的	现价	估值 (PE / 前瞻 / PB)	在这条链上的位置	方向
MU	\$877.57	19.80 / 6.10 / 9.84	HBM4 爬坡最快，已售罄至 2027	
SK 海力士	₩1,436,000	6.23 / 3.44 / 3.95	HBM 份额第一，长约覆盖最广	
三星电子	₩231,500	10.17 / 3.59 / 2.54	分部估值错配最大	
SNDK	\$1,212.21	16.43 / 5.71 / 11.48	NAND 泄压阀 + HBF 推动方，但已被元件成本反噬	
NVDA	\$223.96	34.30 / 22.43 / 27.75	规格通缩的执行方，毛利率防御中	
RMBS	\$101.52	46.6 / — / —	内存接口 IP，规格分叉利好授权模式	
MRVL	\$218.72	75.2 / — / —	CXL 控制器与自研 ASIC 双敞口，2026 年无承接力	

标的表仅为方向性评级与观察清单，不构成任何买卖建议。

九、风险与反证

风险	会证明结论错的观测信号	观测方法
减配未获公司确认	最终规格定案为 12 层 HBM4E (384GB)，四方案评估回到原基准	NVIDIA 官方规格表更新；FQ2 FY27 财报电话会 2026-08-26
机架时间表整体后移	2027 年实际交付的仍是 Oberon 级机架、NVL576 低量或顺延	机架级 ODM 出货口径、中板 PCB 供应链动向
价格拐点早于预期	2027 下半年 DRAM 合约价环比转负（当前卖方假设 -3%、NAND -5%）	季度合约价环比序列
短缺被推翻	原厂公开表态供应转向平衡；下游停止下调规格	三大原厂季度电话会的位元出货与产能分配口径
拥挤度重建	高 PB 标的（SNDK 11.48、MU 9.84）在无盈利上修的情况下继续扩张倍数	PB 与前瞻 EPS 修正方向的背离
成本转嫁失效	更多下游公司在指引中点名元件成本压缩毛利率	下游 OEM 与模组厂的毛利率指引

这条结论最脆弱的地方：核心触发事件至今没有公司层面的一手确认。公司发言人称路线图完好，硬件工程负责人在 7 月公开表示内存供应处于稳定状态。可核实的产业口径只到「评估范围扩大到四个方案、最终规格未定」，具体容量数字来自两条相互独立的供应链信源交叉。**规格通缩的机制成立，具体幅度尚未定案。**

十、来源置信度、口径与数据分歧

A / B 级（公司披露与产业一手口径）

结论 / 数字	口径	置信度
现役 Rubin 单封装 288GB HBM4、22 TB/s	公司官方技术文档，单 GPU 封装口径，均带「up to」	A
Rubin Ultra 官方基准 1TB HBM4E / NVL576 / 600kW / 2027H2	官方公开演示，截至 2026-08 未被新规格表替代	A
机架容量 GB200 NVL72 = 13.4TB HBM3E + 17TB LPDDR5X；GB300 = 20TB + 17TB	公司官网规格表	A
SOCAMM 192GB → 96GB、机架 CPU 内存 55TB → 28TB、仅获约 60% LPDRAM 需求	产业调研机构官方稿，供给驱动定性	B
HBM 评估扩至四方案、动因为 2027 DRAM 短缺 + 12 层验证不确定	产业调研机构官方稿， 不含任何 GB 数字	B
合约价 QoQ 序列 90~98% → 58~63% → 13~18%	传统 DRAM，3Q26 一档另有「服务器 DRAM」并行表述	B
HBM 晶圆投片占比 18/22/30%、位元供给占比 8/9/13%	总投片口径，非仅三大原厂	B
主线方案 8 层 / 192GB、峰值算力不变、约 1,800W	两条独立供应链信源交叉，公司未确认	B—

C 级（卖方模型与产业链转述，仅作量级参考）

结论 / 数字	口径	置信度
机架物料清单：整机 \$7.80M / 内存 \$2.00M（占 25.6%）	卖方模型，非公司披露	C
整机 \$9.1M / 内存与存储约 \$3.2M；HBM4 2027 年 \$53/GB	另一家卖方模型，分母与上一行不同，禁止互相代入	C
HBM4 对下游约 mid-\$500/堆栈（36GB 12 层 ≈ \$15/GB），较 HBM3E 高逾 50%	产业链转述；另有 per-Gb 口径给出 HBM4 约 \$2/Gb、HBM3E 约 \$1.5—1.6/Gb（溢价 21~30%）	C

数字三道关自评：①量级一致性——机架级容量与堆栈数逐项复核（144 封装 × 16 堆栈 = 2,304 堆栈；36 CPU × 8 × 192GB ≈ 55.3TB，腰斩后 27.6TB，与 28TB 相符）；单机内存价值量因两套模型分母互斥，**只取相对变化、不取绝对占比**。②单位显式化——容量统一为 GB/单 GPU 封装，带宽区分 TB/s 与 Gb/s，韩元与美元不混算，per-stack 与 per-Gb 单价分列。③双源——现价经 yfinance 与 EODHD 交叉；官方规格取公司文档；产业数据取机构官方稿，未获一手来源者一律标 C 或留白。

真实存在的机构间分歧（三处并列，不做收敛）：

1. **单机架物料清单总额：**\$7.80M 与 \$9.1M 两套模型并存，内存占比分别为 25.6% 与约 29%，分母定义不同（是否含存储、是否含机架结构件），**不可互相代入或相乘**。

2. **HBM4 的 2027 年单价**：\$32–40/GB 与 \$53/GB 两个口径未收敛，差异约四成，直接决定单机内存价值量测算的量级。
3. **HBM4 相对 HBM3E 的溢价**：按每堆栈口径为逾 50%，按每 Gb 口径为 21~30%——两者的差别在于比较基准是否为 HBM3E 涨价前的价格。

主动留白：2027 年 HBM 合约价的具体涨幅倍数无机构原始口径，只可表述为「上涨数倍」；算法优化（KV cache 压缩、稀疏化）对总 HBM 采购量的削减幅度，遍查无可核实证据，此处留白；机架级封装的面板级后继方案量产时点无可核实出处，不予引用。

免责声明

本报告由涌现资本产业研究部独立出品。所载数据均来自公开可核实来源，现价采用实时行情接口于 2026-08-10 取得。报告仅供研究参考，不构成任何证券的买卖要约或投资建议。市场有风险，决策需谨慎。