ISSUE 08 TIERED MEMORY & CREDIT TRANSMISSION

08

HBM 太贵之后，AI 服务器把数据放在哪里

KV Cache 冷热分层卸载、四级存力重构与算力信用传导



00 / EXECUTIVE SUMMARY · 核心研究结论与关键指标总览

存力重构核心判断：以四级带宽与容量分层替代单一 HBM 堆叠，以长协锁量隔离局部信用波动

5% 热索引常驻	MRDIMM 1+10	NVMe FDP + QLC	LTA vs 1,100bp
95% 温冷历史 KV Cache 按需外迁主机内存与闪存池	双 Rank 并行复用实现带宽翻倍，CXL 单机扩展率先落地	按生命周期物理分块消除写放大，打开百 TB 闪存空间	头部云厂长协锁定远期产能，局部高息裸租与消费端承压

01
重算税与四级卸载

长上下文多轮 Agent 交互若丢弃历史状态将付出平方级重算税，而全量驻留 HBM 又受制于 775 微米封装限高、3 倍以上等效晶圆消耗与近六成加速卡物料占比。依托稀疏注意力将约 5% 高频检索索引常驻 HBM、95% 温冷主干缓存向主机内存与闪存分级卸载，vLLM 百万 Token 实测将单请求显存占用从 17 GB 压降至 2 GB，首字延迟缩短 7.8 倍。

02
接口裂变与闪存分流

主机内存承接溢出缓存需同步突破容量与带宽瓶颈。新一代 DDR5 MRDIMM 在颗粒不超频前提下通过双 Rank 并行复用将速率推升至 8,800–12,800 MT/s，单条接口套片由 1 颗 RCD 增至 1 颗 MRCD 加 10 颗 MDB；持久化层通过 NVMe FDP 按数据存活周期物理分块写入，将写放大系数压回 1 附近，使百 TB 级 QLC eSSD 顺畅进入推理热路径。

03
三季报成色分化

2026 年三季报验证了价值链的结构性迁移：上游晶圆原厂凭借产能稀缺性掌握定价权，长鑫科技前三季度营收达 2,690.24 亿元、单季毛利率升至 89.53%；不背负晶圆库存的澜起科技前三季度净利率达 59.79%；模组环节则告别单纯囤货价差逻辑，向企业级导入与晶圆级封测制造分化，佰维存储正累计投入 75.9 亿元建设松山湖封测基地。

04
长单护盾与信用传导

本轮存储周期由覆盖 2027–2028 年的照付不议 LTA 长协与大客户预付款托底，阻断了传统现货周期的无序波动。但存储涨价向下游传导已呈现显著分化：32GB DDR5 渠道价升至约 450 美元挤压消费终端拉货。而融资利率达 12%–15%、5 年期 CDS 利差突破 1,100bp 的高杠杆算力租赁商，在租金回落情景下面临偿债覆盖率跌破违约线的压力。

00 / TABLE OF CONTENTS · 报告全局研究框架与视觉资产导航

六章研究架构：从 KV Cache 膨胀与 HBM 物理约束穿透至四级存力分层、三季报核验及算力信用传导

CHAPTER 01 HBM COST & MANUFACTURING WALLS

HBM 成本倒挂与 AI 服务器内存系统的底层重构

- 1.1 三季报利润率倒挂：存储原厂盈利反超算力与代工龙头
- 1.2 重算税约束、冷热访问分化与 775 微米堆叠制造边界

CHAPTER 02 FOUR-TIER KV CACHE OFFLOADING

从 GPU 显存到主机内存与闪存池：历史缓存的四级分层卸载

- 2.1 稀疏注意力解耦：5% 高频检索索引常驻与 95% 温冷缓存外溢
- 2.2 跨层异步预取机制、四级存力边界与百万 Token 机柜成本核算

CHAPTER 03 MRDIMM (1+10) & CXL POOLING

主机侧带宽与容量突围：MRDIMM（1+10）架构与 CXL 内存池

- 3.1 板载焊死与插槽扩展的取舍：RCD 信号中继与主机扩容底座
- 3.2 MRDIMM 双 Rank 多路复用增量与 CXL 两阶段商业成熟度

CHAPTER 04 NVME FDP, QLC ESSD & HBF BOUNDARIES

企业级闪存进入推理热路径：NVMe FDP、大容量 QLC 与 HBF 边界

- 4.1 混合生命周期写入引发的传统 SSD 垃圾回收与尾延迟抖动
- 4.2 NVMe FDP 物理分流、百 TB 级 QLC 能效与 HBF 只读权重定位

CHAPTER 05 VALUE CHAIN SHIFT & Q3 2026 AUDIT

AI 存储价值链重分配：2026 三季报财务核验与三阶段演进路标

- 5.1 接口芯片、企业级主控模组与晶圆原厂的三季报业绩分化
- 5.2 围绕降低静态 HBM 依赖的三阶段商业化兑现节奏

CHAPTER 06 LTA FIREWALLS, $450 DDR5 & NEOCLOUD DSCR

存储周期的护盾与裂缝：LTA 长单锁量、终端挤出与算力信用传导

- 6.1 HBM 三倍晶圆消耗与 2027–2028 照付不议长协的周期重塑
- 6.2 450 美元内存的消费端挤出与高杠杆 Neocloud 偿债覆盖率测试

VISUAL & QUANTITATIVE ASSET INDEX · 14 HIGH-RES MEMORY HIERARCHY DIAGRAMS, Q3 EARNINGS MATRICES & CREDIT STRESS CHARTS

全报告图表与量化矩阵清单

- 图01–图04：产业链利润率倒挂与 BOM 拆解、HBM 775 微米堆叠剖面、KV 冷热分化、四级卸载实证
- 图05–图08：板载与插槽内存对比、MRDIMM 1+10 套片原理、CXL 两阶段架构、NVMe FDP 与 QLC 能效
- 图09–图10：2026 三季报存储全链财务分化矩阵、存力重构三阶段商业化路标
- 图11–图14：LTA 照付不议长协护盾、原厂保证金约束机制、Neocloud 信用利差与 15% 利率 DSCR 测试

01 CHAPTER 01 · HBM BOM INFLATION & 775-MICRON PACKAGING WALL

HBM 成本倒挂与 AI 服务器内存系统的底层重构

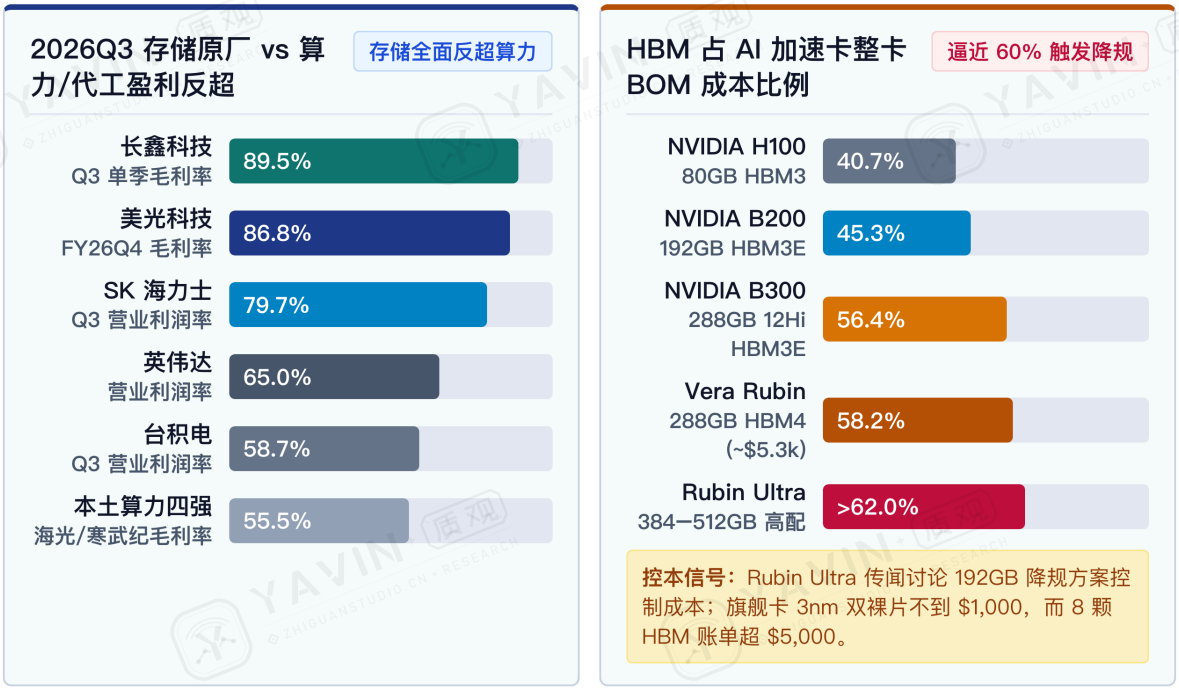
1.1 三季报利润率倒挂：存储原厂盈利反超算力与代工龙头

2026 年第三季度财报披露完毕后，全球半导体产业链出现了一组标志性的盈利结构倒挂：上游存储晶圆原厂的利润率全面超越下游逻辑算力龙头与先进制程代工厂。SK 海力士、美光与三星半导体部门的单季营业利润率悉数跨入 72.4%–82.3% 区间，本土 DRAM 原厂长鑫科技单季营收达 1,187.14 亿元、归母净利润 544.35 亿元，单季毛利率升至 89.53%。由于晶圆厂折旧与洁净室固定制造费用相对刚性，当产能满载且合约价持续上行时，原厂新增营收的边际毛利率超过九成。相比之下，同期英伟达营业利润率约为 63%–66.5%，台积电为 58.7%，海光信息、寒武纪、摩尔线程与沐曦股份四家科创板算力芯片企业的毛利率则分布在 53.7%–56.6% 区间。

这种盈利倒挂直接源于新一代 AI 加速卡的物料成本失衡。拆解 NVIDIA B300 或 Vera Rubin VR200 等旗舰板卡，台积电先进工艺代工的两颗核心计算裸片合计制造成本不足 1,000 美元，而环绕计算核心排布的 8 颗 HBM 显存子系统采购成本高达 4,850–5,300 美元，占整卡物料总成本的 56%–60%。面对持续膨胀的显存开支，供应链已开始评估是否将下一代 Rubin Ultra 部分主力配置的 HBM 容量由原规划的 384–512GB 回调至 192GB；各大模型实验室与开源推理社区则同步上线分层缓存卸载方案，将低频访问的历史状态移出昂贵的 GPU 显存。

下游架构重心的切换，已在 A 股存储产业链的三季报现金流上投射出清晰分化。前期依赖低价囤积消费级晶圆博取涨价价差的模组厂，在三季度遭遇手机与 PC 终端砍单降配冲击，佰维存储单季归母净利环比回落 62%，德明利期末存货升至约 290 亿元、前三季度经营现金流净流出逾 115 亿元。反观不背负晶圆库存、卡位服务器内存接口芯片的澜起科技，三季度归母净利同比增长 161.98% 至 12.40 亿元；跑通企业级 eSSD 与服务器 RDIMM 导入的江波龙，也在三季度单季实现经营活动现金流转正至 8.77 亿元。

EXHIBIT 01 2026 三季报产业链利润率倒挂与旗舰加速卡 HBM 物料占比拆解



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

表 1：历代旗舰 AI 加速卡 HBM 物料成本占比演进与 2026Q3 产业链利润率倒挂全景对标

旗舰 AI 加速卡 / 对标主体	核心存储配置 / 财务指标	计算裸片成本 / 营收规模	HBM 子系统成本 / 利润表现	核心占比 / 利润率指标	产业经济含义与硬件演进指向
NVIDIA H100	80GB HBM3	~$300	~$1,350	**BOM 占比 40.7%**	训练时代 HBM 成本尚处可控区间
NVIDIA B200	192GB HBM3E (8-Hi)	~$680	~$2,900	**BOM 占比 45.3%**	单卡显存翻倍，HBM 成为第一大单一物料
NVIDIA B300	288GB HBM3E (12-Hi)	~$750	~$4,850	**BOM 占比 56.4%**	12 层堆叠良率溢价推升显存成本突破五成
NVIDIA Vera Rubin (VR200)	288GB HBM4 (8-Hi)	~$950	**~$5,300**	**BOM 占比 58.2%**	两颗 3nm GPU 核心不足千美元，8 颗 HBM 超五千美元
Rubin Ultra（高配 vs 传闻控本）	384–512GB vs 传闻讨论 192GB	~$1,800	~$8,500–$10,000	**BOM 占比 >62.0%**	逼近六成触发下游算力厂与云厂开启“HBM 减负”
长鑫科技 / 美光 / 海力士	2026Q3 存储晶圆原厂盈利	长鑫 Q3 营收 1,187 亿元	长鑫 Q3 净利 544 亿元	**89.5% / 86.8% / 79.7%**	存储原厂利润率全面反超英伟达（65%）与台积电（58.7%）

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

当云服务商与模型厂商不再愿意为每一份历史上下文支付高昂的 HBM 溢价，AI 服务器的存力层级必然迎来系统级重设计。理解这场重构的起点，在于从底层算子逻辑厘清长上下文多轮 Agent 交互下的“重算税”

与 KV Cache 膨胀机制。

1.2 重算税约束、冷热访问分化与 775 微米堆叠制造边界

大模型在线服务包含计算密集型的预填充与访存带宽密集型的逐词解码两个阶段。在多轮工具调用、全量代码库扫描与长文档分析场景中，模型每解码生成一个新 Token，GPU 都必须回溯此前全部上下文的特征信息。若不保留中间状态而选择每步从头重算，计算复杂度将随序列长度呈平方级增长，百万字级上下文每吐出一个词都需耗费数十秒等待，形成难以承受的算力“重算税”。工程上规避重复计算的唯一通用路径是以空间换时间，将历史 Token 在每一层神经网络生成的键值中间向量保存为 KV Cache，其驻留总体积由模型层数、键值头数、序列长度与并发批次严格决定：

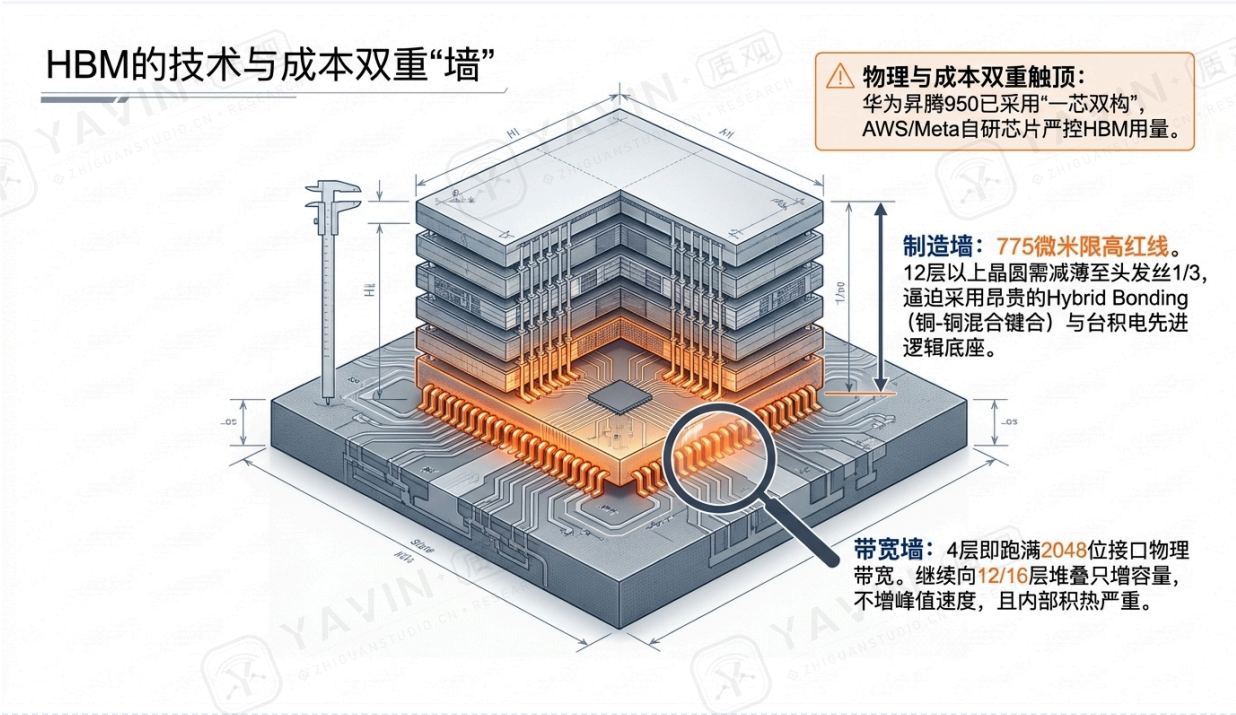
EQUATION · CORE QUANTITATIVE & ECONOMIC MECHANISM

$$V(\mathrm{KV_{Cache}}) = 2 \times L \times n_{\text{kv heads}} \times d_{\text{head}} \times S_{\text{seq}} \times B_{\text{batch}} \times \mathrm{BytesPerElem}$$

KV Cache 驻留体积随网络层数、上下文序列长度及并发批次线性膨胀，长文本高并发场景下动态缓存规模远超静态模型权重本身。

即便算法端引入了稀疏混合专家架构与低精度量化，7×24 小时在线 Agent 持续调用外部工具并保留长周期会话状态，仍使线上驻留数据体积快速膨胀。以 78 层 MoE 架构的智谱 GLM-5.3 为例，单个 100 万 Token 长上下文会话产生的完整 KV Cache 约达 54 GB；若线上同时承载 128 路活跃并发请求，仅动态产生的历史缓存总量便高达 6.4 TB，数倍于模型参数本身的体积。面对这一容量缺口，单纯将单卡 HBM 从 8 层一路堆叠至 16 层乃至更高层数，已同时撞上访问效率与物理制造两道边界。

EXHIBIT 02 **HBM3E 与 HBM4 三维堆叠剖面：775 微米封装限高与制造良率边界**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

第一道边界来自长上下文数据的冷热访问错配。在百万字级多轮交互中，数据访问频率高度不均：当前解码步真正需要纳秒级高频读取的，仅是激活权重、轻量级检索目录与最近生成的一小段热窗口；而占历史缓存总体积 90%–95% 的早期对话与背景资料虽须保留以备跨段落回溯调用，在当前计算步却处于低频待命的温冷状态。用每 GB 采购价高达 15–20 美元的 HBM 静态存放海量温冷数据，甚至为凑够数 TB 显存容量而额外采购十几台 8 卡 GPU 服务器，在数据中心经济账上极不合算。

第二道边界源于 HBM 高层堆叠的物理制造极限与晶圆消耗代价。在硬件工程上，HBM4 单堆栈峰值带宽翻倍依靠的是将接口位宽由 HBM3E 的 1024-bit 拓宽至 2048-bit，每层核心裸片提供 512 个数据引脚，堆叠至 4 层即可跑满全部物理位宽；继续向 12 层、16 层垂直堆叠仅能扩充单颗容量，反而因硅通孔链路拉长、寄生电容增大及内部积热而限制高频信号速率。与此同时，JEDEC 标准将整颗 HBM 封装高度严格限制在 775 微米以内，塞入 12 至 16 层晶圆意味着单片晶圆须背面减薄至 25–30 微米以下，晶圆翘曲与微裂纹风险急剧上升，键合工艺被迫由 MR-MUF 与 TC-NCF 向高门槛的铜对铜无凸点混合键合演进，HBM4 底层控制裸片亦转向台积电 12nm/5nm 等先进逻辑工艺定制代工。综合裸片尺寸扩大、TSV 通孔开销与堆叠良率折损，每生产 1 GB HBM 消耗的等效晶圆面积高达标准 DDR5 的 3 倍以上。

受此约束，AWS Trainium2 与 Meta MTIA 等北美云厂商自研推理芯片均严格控制片上 HBM 配额，华为昇腾 950 亦采用“一芯双构”设计，将计算密集的预填充与带宽密集的解码拆分为两款芯片按需配存。而在占市场主力的通用 AI 服务器上，将温冷缓存向片外分级卸载已成为全行业的共同演进方向。

02 CHAPTER 02 · KV CACHE COLD-HOT SPLIT & FOUR-TIER OFFLOADING

从 GPU 显存到主机内存与闪存池：历史缓存的四级分层卸载

2.1 稀疏注意力解耦：5% 高频检索索引常驻与 95% 温冷缓存外溢

在不触发重算税、也不损伤长文本跨段召回精度的前提下将历史缓存移出 GPU，核心前提是在算法层完成冷热数据的物理拆解。以 DeepSeek 动态稀疏注意力与智谱 GLM-5.3 的 IndexShare 机制为代表，新一代模型为庞大的历史 KV Cache 建立了一套紧凑的检索目录索引。

EXHIBIT 03 长上下文推理的冷热解耦：5% 高频检索索引常驻 HBM 与 95% 温冷缓存外溢

热目录索引层（常驻 HBM）	温冷历史 KV 卸载比例	存力结构重塑收益
~5%	~95%	降本 80%+
GLM-5.3 稀疏注意力 Indexer KV 实测约 5.1%	外溢至 Host DDR5 温层与企业级 NVMe SSD 冷层	终结“拿五千美元 HBM 当普通仓库”错配

MoE 与稀疏注意力下的“5% 热目录常驻 HBM + 95% 温冷缓存卸载”三级存力模型

HBM + Host MRDIMM + NVMe FDP SSD

L1 热层 · GPU HBM3E/4（约 5%）

仅存 Indexer 目录索引

极速访问：>8 TB/s 带宽仅服务于当前激活权重与 Top-k 目录检索索引。

止损意义：避免昂贵高带宽显存被海量沉睡多轮对话 KV 挤占爆仓。

L2 温层 · Host DDR5/MRDIMM

2 TB–4 TB 弹性缓冲

精准唤回：按热索引定位后，通过 PCIe Gen5/6 或 C2C 仅拉取命中的 KV Chunk。

扩容刚需：引爆双路 x86 主存插槽与 RCD/MDB 接口芯片量价齐升。

L3 冷层 · 企业级 FDP SSD

PB 级持久化前缀池

以存代算：承接跨会话长周期历史缓存与 RAG 知识库，跳过重复 Prefill。

成本差：单位容量成本仅为 HBM 的零头，撑起长上下文商业闭环。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

在每一步解码生成新词时，GPU 不再遍历全部百万字主干历史缓存，而是分两步执行：先用计算开销极低的轻量检索器扫描一遍目录索引——在 GLM-5.3 中，每 4 层网络共享 1 个检索器，使目录索引 **Indexer KV** 仅占总缓存体积的约 5.1%——快速筛选出与当前语境最相关的前 2,048 个历史 Token 位置；随后仅抽取这部分约占总量 0.2% 的主干缓存参与注意力运算，其余 99.8% 未被抽中的历史数据在当前步完全无需读取。即便是阿里 Qwen3-Next 等采用的线性混合注意力架构，因纯线性注意力在长距离精确检索上存在精度衰减，生产模型仍须保留 12.5%–25% 的全注意力层及多轮分支状态快照。因此，算法演进省下的是单步访存数据量，却省不掉总历史保存容量，但它成功在物理上划出了每步必读的 5% 热索引与按需抽调的 95% 温冷主干缓存。

2.2 跨层异步预取机制、四级存力边界与百万 Token 机柜成本核算

依托算法层的冷热剥离，推理引擎得以在不拖慢生成吞吐的前提下调用主机内存资源。vLLM 团队在单台 8×H200 服务器、固定 512GiB 主机内存预算下运行 GLM-5.3 百万 Token 推理的 Hybrid HiSparse 实测，清晰展示了生产级动态卸载的工程机制：每步检索必用的 **Indexer KV** 始终 100% 常驻 GPU HBM；主干历史缓存在显存充裕时同样优先驻留片上，仅当百万长文本或高并发挤压显存容量时，系统才按冷热优先级将较冷缓存块逐步驱逐至 CPU 主机内存。一旦后续步骤中某块片外冷数据再次被检索目录命中，推理引擎便利用 GPU 计算第 N-1 层的数十微秒窗口，经主板 PCIe 通道将第 N 层所需数据异步预取回显存，用计算流水掩盖传输延迟，将单请求显存占用从 17 GB 压降至 2 GB、首字延迟缩短 7.8 倍，避免了传统无差别搬运造成的总线拥堵。

结合月之暗面 Mooncake、华为 UCM、阿里 KVCacheStore 与 DeepSeek 3FS 等大规模生产集群部署，新一代 AI 推理服务器的存力体系正收敛为涵盖 GPU HBM、CPU DDR5/MRDIMM、CXL 扩展内存与企业级 NVMe SSD 的四级协同架构，各级介质的访问延迟与单位成本边界如下表所示。

表 2：新一代 AI 推理服务器四层存力金字塔硬件介质、延迟成本量级与核心职责对比（注：并非所有现有服务器均配齐全部四层，但这构成了理解新一代推理基础设施演进的核心架构模型）

存储层级	主要硬件介质	典型访问延迟与单位成本量级	在 AI 推理系统中的核心职责与数据驻留内容
第一层（L1）	**GPU HBM**（HBM3E / HBM4）	30–80 纳秒 **~$15–$20/GB**	**极速核心计算层**：常驻模型权重、每步必读的高频检索目录（如 **Indexer KV**，约占 5%）及当前活跃热窗口
第二层（L2）	**CPU DDR5 RDIMM / MRDIMM**	70–110 纳秒 **~$4–$8/GB**	**大容量主机工作内存层**：在 GPU 显存承压时承接动态卸载的温冷主干历史缓存（约占 95%），支持跨层高速预取
第三层（L3）	**CXL 扩展内存**（PCIe 5.0/6.0）	150–250 纳秒 **~$4.5–$8.5/GB**	**插槽外扩展与池化层**：突破主板本地 DIMM 插槽物理上限，实现单机内存扩容、旧 DDR4 压缩复用或探索跨机池化
第四层（L4）	**企业级 NVMe SSD（含 FDP / HBF）**	10–100 微秒 **~$0.12–$0.25/GB**	**跨轮次持久化与“以存代算”层**：保存暂停会话的历史记忆、全局共享长提示词前缀与只读冷权重，免去重复预填充计算

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

将这套四级分层架构放入数据中心机柜投资账本，其降本逻辑尤为直接。以同时承载 128 路 100 万 Token 智能体活跃会话、实时产生 6.4 TB 历史 KV Cache 的企业级场景为例：若沿用不卸载的纯 HBM 驻留方案，云厂商被迫采购 17 台 8×H200 服务器，硬件购置成本高达 544 万美元；其中最严重的资源错配在于每台服务器都必须先在本地 HBM 内重复存放一份约 700GB 的静态模型权重，17 台机器合计有 11.9 TB、即总显存容量的 66.5% 消耗在冗余权重副本上，GPU 计算核心利用率不足 15%。

改用分级卸载方案将 95% 温冷缓存迁入大容量主机内存与高速企业级 SSD 后，仅需 2 台 8×H200 服务器即可跑满同等并发计算负载，三年期总拥有成本降幅超过 85%。单次长文本推理成本的大幅下降随即诱发杰文斯效应：过去因算力账单过高而难以规模上线的全天候代码库重构、百万字跨轮次精读等 Agent 负载快速放量，促使云厂商将单机并发度推至极限，并为每台高并发推理服务器同步配齐 1.5TB–4TB 高速主机内存与数十至上百 TB 企业级 SSD。

EXHIBIT 04 四级存力分层协同架构与 vLLM 百万 Token 推理卸载实证

AI 推理服务器四层存储协同模型：按冷热梯度重塑内存架构 L1→L4 冷热分级协同

L1 · GPU HBM 显存	L2 · DDR5 / MRDIMM	L3 · CXL 扩展内存	L4 · NVMe FDP SSD
极高带宽 / 纳秒级	TB 级容量 / 异步预取	突破插槽 / 一致性协议	PB 持久化 / 以存代算
常驻模型权重、高频 Indexer KV 与活跃热缓存。	显存承压时逐步承接卸载的温冷历史主干缓存。	突破主板 DIMM 插槽物理上限，支持扩容与池化。	跨轮次会话记忆与全局共享长前缀持久化，免重算。

vLLM GLM-5.3（100 万 Token 上下文）生产实测对比 单台 8×H200 · 512 GiB 锁页主存

传统 KV Cache 全量卸载基线	Hybrid HiSparse 分层协同方案
配额方式： 512 GiB 主存无差别接收全量历史块	**分级配额：** 384 GiB HiSparse 主存 + 128 GiB 常规前缀主存
核心瓶颈： 不区分索引与主干冷热特征，高并发下 PCIe 通道易被大块数据反复搬运阻塞	**GPU 常驻：** 高频 Indexer KV + 热块留守 HBM，显存承压时按需将冷块动态溢出至主存与 SSD

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

当数以 TB 计的温冷历史缓存从 GPU 显存向片外逐级下沉，硬件瓶颈随即传导至服务器主板侧的内存容量、随机读取带宽与物理插槽上限。

03 CHAPTER 03 · DDR5 MRDIMM (1+10 CHIPSET) & CXL MEMORY POOLING

主机侧带宽与容量突围：MRDIMM（1+10）架构与 CXL 内存池

将海量溢出缓存交由服务器 CPU 侧承接时，硬件工程团队会依次遭遇三道递进的物理约束：首先是单机主机内存容量如何从常规的 512GB 跃升至 1.5TB–4TB；其次是单机 8 张 GPU 同时发起跨层预取时，主机内存读取带宽如何匹配；最后是主板本地内存插槽全部插满后，如何继续横向扩容。这三重工程约束，直接决定了 DDR5 RDIMM、MRDIMM 与 CXL 的商业落地次序。

3.1 板载焊死与插槽扩展的取舍：RCD 信号中继与主机扩容底座

在双路服务器主板上，16 至 32 个内存插槽密集挂载着上千颗 DRAM 存储颗粒。若由 CPU 内部内存控制器直接通过主板长走线驱动全部颗粒，过大的电容负载会导致地址、命令与时钟信号严重衰减失真，因此数据中心大容量内存必须采用寄存式模组 DDR5 RDIMM。

EXHIBIT 05 **主机内存扩容路线对比：Grace 板载焊死 LPDDR5X 与双路 x86 服务器 RDIMM 插槽架构**

① NVIDIA Grace：480GB LPDDR5X 焊死主板

容量触顶 · 整板返厂风险

固定物理封顶：单颗 CPU 标配 **480 GB LPDDR5X** 直接焊死在主板 PCB 上，无法插拔扩容。

卸载撞墙困境：遇到 TB 级长上下文 KV Cache 卸载或大型企业并发时即刻触及容量上限。

单点运维诅咒：由于无独立内存模组插槽，**任意 1 颗 DRAM 颗粒损坏即迫使整张主板返厂维修**，停机与备件成本极高。

② 双路 x86：24–32 插槽 RDIMM + RCD 中继

2 TB–4 TB 扩容 + 信号中继红利

TB 级弹性扩容：Xeon 6 / EPYC Turin 提供 **24–32 个可插拔 DDR5 RDIMM/MRDIMM 插槽**，单机平滑跃升至 **2 TB–4 TB** 主存池。

RCD 寄存时钟驱动物理刚需：驱动单机上千颗 DRAM 颗粒时电容负载急剧恶化，必须依赖 **RCD 芯片** 作为核心信号中继站防止时钟指令失真。

接口芯片含金量重估：单根 MRDIMM (1 颗 RCD + 10 颗 MDB) 将高增速“过路费”固化在内存接口芯片层。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

DDR5 RDIMM 在内存条中央增设了一颗寄存时钟驱动器芯片 RCD，并搭配 PMIC 电源管理、SPD 集线器与温度传感器组成标准接口套片。CPU 发出的控制与时钟信号先统一送达 RCD 完成缓冲整形与驱动放大，再向模组两侧的 DRAM 颗粒分发，从而依靠信号中继稳定支撑单根 64GB、128GB 乃至 256GB 的大容量模组。与之形成鲜明架构对照的是英伟达用于 GH200 与 GB200 机柜的 Grace CPU：为换取极致每瓦能效与紧凑走线，Grace 取消了可插拔插槽，直接将 LPDDR5X 颗粒焊死在计算主板上。这一取舍虽借助 NVLink-C2C 提供了极高的片间互连带宽，却在长上下文卸载场景下暴露出两项工程短板：一是单颗 Grace CPU 的内存容量出厂即永久锁死在 480GB 上限，无法随推理负载插拔扩容至 2TB–4TB；二是可维护性受限，主板上焊死的数十颗颗粒一旦在长期满载下出现单点失效，整块高价值计算板卡均须停机返修。正因如此，拥有 24 至 32 个可插拔插槽的双路 x86 服务器平台（如 Intel Xeon 6 Granite Rapids 与 AMD EPYC Turin）在大容量推理节点中展现出显著的配置灵活性。

3.2 MRDIMM 双 Rank 多路复用增量与 CXL 两阶段商业成熟度

插满大容量 RDIMM 解决了容量驻留问题，但当单机并发度拉高、GPU 显存缓存未命中率跨过 10%–20% 临界点时，8 张 GPU 在数十微秒级跨层窗口内并发向主机内存发起异步预取，常规 DDR5-4800/6400 的随机读取带宽很快出现排队拥堵。若将 CPU 内存通道数强行翻倍，会导致封装面积、引脚数与主板层数失控；若强行翻倍 DRAM 颗粒内部电容的充放电频率，又受制于存储工艺物理极限。MRDIMM（多路复用 Rank 内存模组）由此采用“内部常速并行、外部倍速复用”的工程折中：在传统 RDIMM 中，模组正反两面的两组独立阵列 Rank 0 与 Rank 1 分时串行共用 64 位外部总线，同一周期仅一侧传输 64B 数据而另一侧闲置；MRDIMM 则在不提高 DRAM 颗粒原生工作频率的前提下，令 Rank 0 与 Rank 1 在同一时钟周期各自安全输出 64B 数据，交由模组下沿的数据缓冲芯片以双倍接口速率多路复用打包，单周期向 CPU 聚合传输 128B。借此，主机侧传输速率跃升至第一代 8,800 MT/s 与第二代 12,800 MT/s，双路 24 通道服务器的主机内存总带宽突破 1.69–2.46 TB/s。

这一复用机制重塑了单根内存模组的逻辑芯片含量。MRDIMM 将中央驱动器升级为 1 颗多路复用寄存时钟驱动器 MRCD，同时在金手指上方对应 8 组数据字节与 2 组 ECC 校验通道排布 10 颗多路复用数据缓冲器 MDB，构成“1+10”套片架构。单根服务器内存条上的核心接口主芯片由传统 RDIMM 的 1 颗直接增至 11 颗，一台插满 24 根 MRDIMM 的双路推理服务器所需高速接口主芯片随之从 24 颗跃升至 264 颗，为澜起科技、Rambus 与瑞萨电子等掌握高速混合信号技术的厂商打开了确定性的单机增量空间。当然，单条新增 10 颗高速有源芯片也抬升了模组功耗与散热密度，对机箱风道或冷板设计提出更高要求，且必须由支持该协议的新一代服务器 CPU 内存控制器、主板走线与 BIOS 固件协同配合。

EXHIBIT 06 **DDR5 MRDIMM 双 Rank 多路复用原理与 1+10 接口芯片套片架构**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

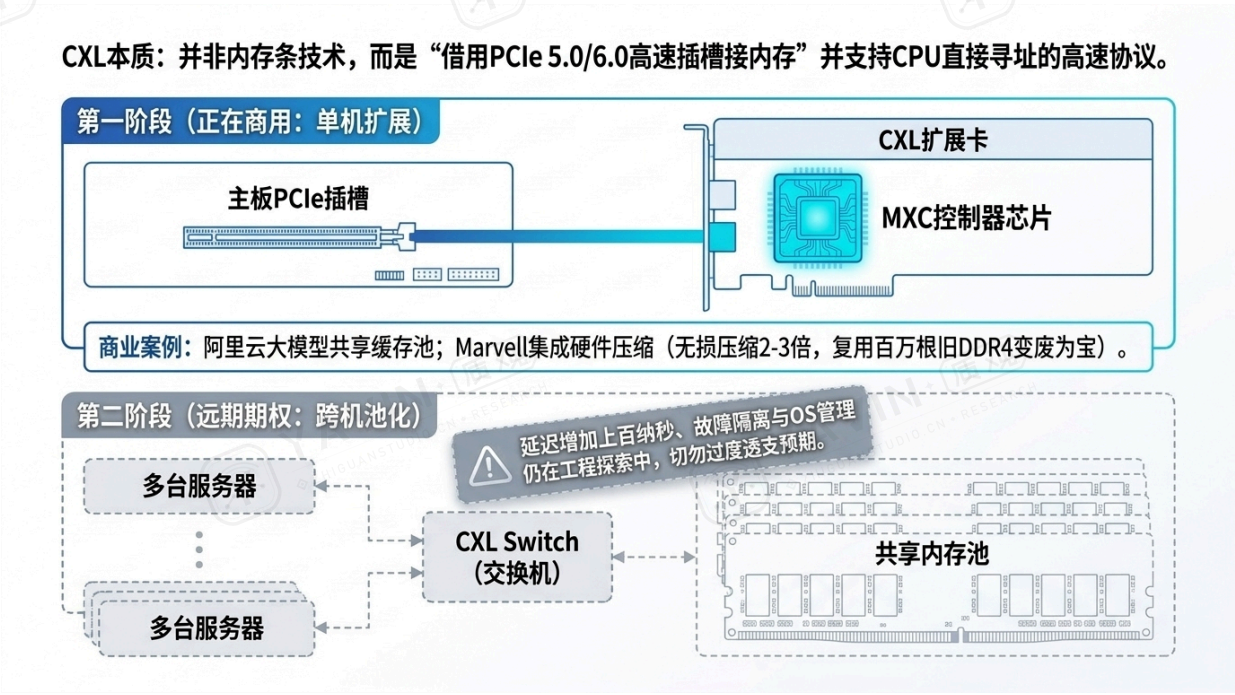
表 3：传统 DDR5 RDIMM 与新一代 MRDIMM 传输速率、双路带宽及“1+10”接口芯片配置演进对比

内存模块代际	主机侧传输速率 (MT/s)	单通道峰值带宽	双路 24 通道理论总带宽	单根内存条的核心接口芯片配置	底层工作机制与工程代价	平台支持与商业化阶段
常规 DDR5 RDIMM	4,800 – 6,400	38.4 – 51.2 GB/s	0.92 – 1.23 TB/s	**1 颗 RCD** + PMIC + SPD + TS	双 Rank 分时串行访问，单周期传 64B	已大规模成熟商用，子代持续迭代
MRDIMM 第一代 (Gen1)	**8,800**	**70.4 GB/s**	**1.69 TB/s**	**1 颗 MRCD + 10 颗 MDB** + 配套	双 Rank 并行 + 缓冲芯片复用传 128B	Intel Xeon 6 / AMD Turin 已支持导入
MRDIMM 第二代 (Gen2)	**12,800**	**102.4 GB/s**	**2.46 TB/s**	**1 颗 Gen2 MRCD + 10 颗 Gen2 MDB**	接口速率翻倍，模块功耗与散热要求提升	澜起科技、Rambus 已完成研发与试产送样
MRDIMM 远期演进 (Gen3)	规划 16,000	128.0 GB/s	3.07 TB/s	1 颗 Gen3 MRCD + 10 颗 Gen3 MDB	需下一代服务器 CPU 与高速信号完整性配合	处于标准与早期预研阶段

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

受制于主板物理空间与高速并行信号走线长度，双路服务器的本地 DIMM 插槽数量通常封顶于 24 至 32 个。当本地插槽全部插满仍需扩容，或希望跨服务器调配闲置内存时，基于高速串行链路的 CXL 协议成为突破本地插槽边界的第三级方案。

EXHIBIT 07 **CXL 演进路径：第一阶段单机内存扩展与第二阶段跨服务器内存池化**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

不同于 DDR5 与 MRDIMM 等并行内存模组，CXL 是构建在 PCIe 5.0/6.0 串行物理接口之上、具备硬件缓存一致性的互连协议，包含设备管理子协议 **CXL.io**、加速器缓存协议 **CXL.cache** 与内存扩展协议 **CXL.mem**。**CXL.mem** 允许主机 CPU 直接使用原生 **Load/Store** 内存读写指令，穿过 PCIe 串行通道访问外部扩展卡上的 DRAM，将其直接映射进系统物理地址空间。在评估产业落地进度时，必须将 CXL 严格区分为商业成熟度截然不同的两个阶段。

第一阶段的单机内存扩展正率先走向商用落地。服务器通过空闲 PCIe 插槽或 E3.S 接口直接挂载内置 MXC 控制器的 CXL 扩展卡，即可为单机额外增加数百 GB 至数 TB 容量与独立串行带宽通道；经过控制器虽会引入约 80–150 纳秒附加延迟，但在大模型数十微秒级的跨层块级异步预取窗口下可被计算流水有效掩盖。例如阿里云已在共享缓存架构中探索 CXL 单机扩展，Marvell 的 Structera 控制器更集成了硬件级 LZ4 实时内联压缩引擎，支持云厂商将退役服务器拆下的旧 DDR4 内存条插上 CXL 扩展卡继续复用以压低池化成本。相比之下，第二阶段通过引入 CXL 交换机连接独立内存机箱、跨多台服务器动态划拨容量的内存池化方案，虽能化解数据中心内存搁浅难题，但跨交换机访问会将延迟推高至 250 纳秒以上，且牵涉复杂的内核 NUMA 动态调度与跨节点故障隔离，目前仍处于小规模概念验证阶段，兑现节奏显著慢于单机扩展。

04 CHAPTER 04 · NVME FDP, ENTERPRISE QLC SSD & HIGH-BANDWIDTH FLASH
企业级闪存进入推理热路径：NVMe FDP、大容量 QLC 与 HBF 边界

顺着缓存外溢链路继续向下延伸，当多步 Agent 进入长达数十秒至数分钟的外部工具调用等待，或成千上万名并发用户共享同一份数万字的系统提示词与行业知识库前缀时，若长期占用断电即失且每 GB 价格仍在数美元以上的 DRAM，机柜经济账依然承压。工程上的自然演进是将跨轮次温冷缓存与公共前缀下沉至第四级企业级 NVMe SSD，在后续会话命中时经高速 PCIe 直接读回显存、免去 GPU 从头重做预填充计算的“以存代算”。华为 UCM 在金融机构长文本实测中借此将 48K 上下文首字延迟从 51 秒压降至 3.8 秒并消除逾半数重复计算，月之暗面 Mooncake 与 DeepSeek 3FS（180 个存储节点实现 6.6 TiB/s 聚合读取吞吐）同样深度依赖这一闪存层。

4.1 混合生命周期写入引发的传统 SSD 垃圾回收与尾延迟抖动

然而，若直接用传统企业级 SSD 承接大模型高并发 KV Cache 卸载，很快就会遭遇严重的寿命磨损与读取卡顿。症结在于 NAND 闪存底层的物理不对称特性：闪存可以按 4KB 或 16KB 的较小“页”写入数据，却无法在原位置直接覆盖改写，重写前必须按数 MB 至数 GB 的“物理闪存大块”执行整块擦除。偏偏在生产级推理环境中，推理引擎下刷至 SSD 时混杂三类寿命悬殊的数据流：保留数天至数周的静态系统前缀与公共知识库、维持数分钟至数十分钟的多轮会话状态，以及显存瞬时承压挤出、数十秒后请求结束立即作废的瞬态溢出缓存。

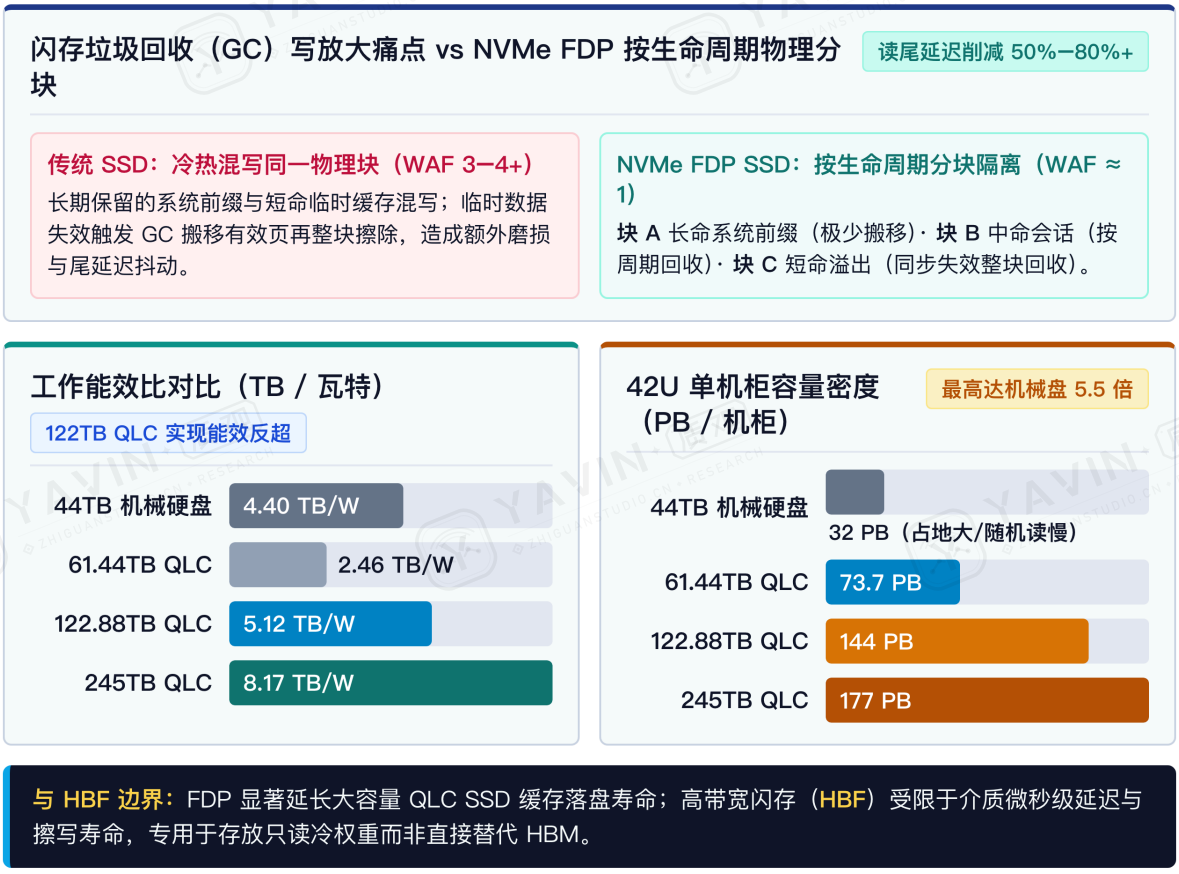
传统企业级 SSD 主控无法识别上层数据的存活周期，只能按写入时间先后将长命前缀、中命会话与短命溢出数据混写在同一个物理闪存大块内。几十秒后，占比很高的短命溢出页批量失效，但同一物理块内仍零星散落着尚在有效期内的长命前缀与中命会话。SSD 主控为回收空间，被迫在后台高频启动垃圾回收：先把旧块内残存的有效页逐一读出并搬迁写入新块，再对旧块执行耗时数毫秒的整块擦除。这种读改写循环带来两重代价：一是严重的写放大，主机仅下发 1GB 缓存写入，闪存内部因搬家实际写入 3–4GB 以上，既侵占内部通道带宽又以数倍速度消耗闪存擦写寿命；二是读取长尾延迟剧烈抖动，一旦 GPU 急需读回某段历史缓存时恰好撞上该通道执行数毫秒级后台擦除，读取请求便被阻塞排队，导致 P99.9 尾延迟骤增数十倍，整机 8 张 GPU 被迫空转等待。

4.2 NVMe FDP 物理分流、百 TB 级 QLC 能效与 HBF 只读权重定位

为从根源消除混写干扰，Meta 与 Google 联合各大原厂及主控厂商在 NVMe TP4146 标准中引入了灵活数据放置技术 FDP。FDP 无需改动 NAND 颗粒物理结构，而是打通了上层推理调度框架与底层 SSD 主控之间的语义隔离墙：**LMCache**、华为 UCM 等缓存框架在写入 KV Cache 时只需增加十余行代码，按预计存活周期附带轻量级放置标识符；支持 FDP 的新一代企业级主控收到标签后，自动将长命前缀、中命会话与短命瞬态溢出流分别引导写入彼此独立的物理回收单元。当几十秒后短命请求结束，对应物理块内的数据同步失效，主控无需搬移任何有效数据即可直接整块擦除回收。生产实测表明，引入 NVMe FDP 可将推理卸载场景下的写放大系数从 3–4 倍以上压平至 1.01–1.16 附近，并将 P99.9 读取长尾延迟削减 50%–80% 以上。

FDP 机制落地直接重塑了第四级闪存的产品格局。一方面，在核心机房供电指标紧缺的约束下，61.44TB 至 245TB 超大容量 QLC 企业级 SSD 凭借 4.9–6.8 TB/W 的高能效密度加速替代传统机械盘与小容量闪存；过去云厂商顾虑 QLC 颗粒仅约千次级的物理擦写寿命经不起 4 倍写放大损耗，如今在 4nm PCIe 5.0/6.0 高速主控与 FDP 减磨机制协同下，寿命顾虑大幅化解，单台 1U/2U 机架即可高密度收纳 PB 级温冷语料与历史缓存池。另一方面，针对近期闪迪与 SK 海力士联合推出的高带宽闪存 HBF——借用类似 HBM 的晶圆减薄与 TSV 垂直堆叠工艺将 16 层 3D NAND 堆叠于逻辑底座之上，实现单颗 512GB 容量与 1.6 TB/s 读取带宽——亦须客观界定物理边界：HBF 底层仍是 NAND 闪存，物理读取延迟达 10–25 微秒，比纳秒级 HBM 慢数百倍，最小访问粒度为 4KB 页且擦写寿命仅 10^4–10^5 次。若强行用 HBF 承载高频动态改写的活跃 KV Cache，不仅会迅速耗尽擦写寿命，微秒级延迟更会拖垮解码吞吐；其真正合理的工程定位是专门存放固定不变的只读模型权重与冷门 MoE 专家参数，释放片上 HBM 空间专注服务高频计算。

EXHIBIT 08 **企业级 SSD 的 NVMe FDP 生命周期物理分流机制与大容量 QLC 机柜能效对比**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

表 4：第四层闪存技术演进（传统 TLC SSD、NVMe FDP SSD、百 TB 级 QLC SSD 与高带宽闪存 HBF）物理边

界与定位对比

第四层闪存技术路线	核心硬件与协议特征	读写带宽与延迟量级	擦写寿命与写放大特征	在 AI 推理服务器中的真实定位	商业化进展
传统企业级 TLC SSD	PCIe 4.0/5.0，常规混写架构	7–14 GB/s ~80–120 μs	约 3,000–10,000 P/E，混写下 WAF 达 3–4+	传统通用存储，高并发 KV 卸载下易受 GC 尾延迟困扰	存量成熟产品
新一代 NVMe FDP 企业级 SSD	**4nm PCIe 5.0/6.0 主控 + FDP 协议**	**14–28 GB/s** 尾延迟削减 50%–80%+	**按生命周期物理分块隔离，WAF 压低至接近 1**	**AI 推理跨轮次 KV 缓存与公共长前缀持久化主力**	**正在头部云厂加速放量**
61.44TB–245TB QLC 企业级 SSD	高密度 3D QLC + FDP 减磨协同	14 GB/s 级读速能效比 **4.9–6.8 TB/W**	~1,000–1,500 P/E，依靠 FDP 解决擦写寿命痛点	**高密度温冷缓存池、RAG 向量库与多模态语料底座**	**北美与本土云厂快速导入**
高带宽闪存（HBF）	16 层 3D NAND **TSV 垂直堆叠 + Base Die**	单颗 **512GB**，读带宽 **1.6 TB/s** **10–25 μs**	10^4–10^5 P/E，延迟比 HBM 慢数百倍，不可高频改写	**专存只读模型权重与冷门 MoE 专家，不可直接替代 HBM**	早期标准与送样验证阶段

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

05 CHAPTER 05 · Q3 2026 STORAGE VALUE CHAIN AUDIT & THREE-STAGE ROADMAP

AI 存储价值链重分配：2026 三季报财务核验与三阶段演进路标

沿着四级存力分层架构核验 2026 年三季报，产业链资本开支与高毛利订单正顺着“服务器内存基础扩容、高速接口与互连芯片升级、企业级 SSD 主控及模组分化、晶圆原厂长约锁量与先进封测外溢”的次序扩散。审视各环节上市公司成色时，核心原则是将已经兑现在三季报利润表中的当期收入来源，与尚处平台导入期的新技术期权严格区分。

EXHIBIT 09 **2026 三季报全球与本土存储产业链业绩验证及现金流分化矩阵**

1. 内存互连、企业级主控与利基存储龙头（2026Q3 率先兑现） 高毛利 + 强现金流验证

代表企业	2026Q3 核心业绩与财务指标	底层驱动与下一步验证路标
澜起科技 RCD / MRDIMM	互连芯片毛利率 **68.4%**；Q3 净利 **+162%（12.40 亿元）**；Q4 在手待交付订单 **25–27.9 亿元**	当前主要由 DDR5 子代 RCD 迭代驱动；后续跟踪 **MRDIMM 套片收入占比**与 CXL 出货
Astera Labs & Rambus 海外连接与接口	ALAB Q3 营收指引 **+138%**；Rambus 产品毛利率高达 **78.3%**	PCIe/CXL 互连芯片与服务器 DDR5 接口芯在北美云厂商 AI 机柜同步放量
平头哥 / 大普微 / 联芸科技 企业级主控 & eSSD	平头哥镇岳 510 出货 **>50 万颗**；大普微进入头部云厂供应；联芸科技 Q3 **+66%**	PCIe 5.0 企业级主控与百 TB 级 QLC eSSD 正在加速导入 AI 推理服务器集群
东芯股份 & 北京君正 利基存储弹性	东芯 Q3 净利环比 **+59%**、毛利率 **79.3%**；北京君正营收 **14.2 亿元**	三大原厂保 HBM 退出 SLC 与 DDR3/4 产线后，利基存储供需缺口兑现

2. 本土晶圆原厂与 A 股存储模组四强分化 重金投建 vs 存货现金流大考

企业阵营	2026Q3 法定财务表现	战略卡位与结构分化信号
长鑫科技 & 长江存储 本土制造原厂	全球份额双升至 **~10%**；长鑫 Q3 单季净利 **544 亿元**、毛利率 **89.5%**、广义锁单 **467 亿元**	全产业链定价咽喉所在地，进入高毛利同价竞争与长单锁量阶段
江波龙 & 香农芯创 企业级放量派	江波龙 H1 企业级 **+209%**、Q3 经营现金流 **+8.77 亿元**；香农芯创年化营收破千亿	企业级 eSSD 导入见效；下一步核心验证 **企业级业务独立毛利率** 提升持续性
佰维存储 & 德明利 封测重资产 vs 备货	佰维斥资 **75.9 亿** 建松山湖晶圆级封测，存货 **240.5 亿**、Q3 净利环比 **−62%**；德明利存货 **~290 亿**、前三季 OCF **−115 亿**	高杠杆备货冲刺企业级或重金投建“准晶圆封测”期，面临短期折旧与存货周转检验

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

5.1 接口芯片、企业级主控模组与晶圆原厂的三季报业绩分化

内存接口与高速互连芯片厂商属于纯逻辑与混合信号设计企业，完全不持有大宗存储晶圆库存，因而免疫消费电子降配引发的存货跌价风险，直接受益于服务器内存通道加密、子代速率迭代与单条接口芯片倍增。澜起科技三季度单季实现营收 20.79 亿元、归母净利 12.40 亿元，互连类芯片毛利率达 68.4%，预计四季度交付的互连在手订单达 25.0–27.9 亿元；配套 DDR5 SPD 集线器芯片的聚辰股份与海外接口龙头 Rambus 单季产品盈利同步创下历史新高，主攻 PCIe 5.0/6.0 Retimer、交换矩阵与 Leo CXL 控制器的 Astera Labs 单季营收指引亦同比翻倍至 5.40–5.60 亿美元。但严谨归因可见，上述企业三季报的高增主力仍来自 DDR5 第三至第五子代 RCD 渗透率抬升及 PCIe Retimer 放量，而非 MRDIMM 或 CXL 已大规模确认收入；未来四个季度真正的验证焦点，在于 MRDIMM“1+10”套片与 CXL 内存控制器在云厂商新平台中的实际出货占比拉升节奏。

第四级企业级闪存与模组赛道则在三季报呈现出深刻的现金流与商业模式分野。随着三季度智能手机与 PC 终端客户因存储涨价启动 10%–20% 砍单并将主流 PC 配置从 1TB 回退至 512GB，单纯依靠低价囤积消费级晶圆赚取价差的模式面临周转压力，期末账面存货分别达 240.53 亿元与约 290 亿元的佰维存储和德明利短期均须消化库存与现金流流出。行业突围正沿两条高壁垒路径展开：一是以自研主控和企业级产品导入穿透云厂供应链，除慧荣、群联与韩国 FADU 外，阿里平头哥自研镇岳 510 主控累计出货已逾 50 万颗，大普微携自研主控及百 TB 级 QLC SSD 导入头部互联网云厂，联芸科技加速推进 PCIe 5.0 企业级主控验证，而率先跑通企业级 eSSD 与服务器 RDIMM 组合放量的江波龙则依托上半年企业级收入增长 208.8%，在三季度单季实现经营活动现金流转正至 8.77 亿元；二是向晶圆级封测制造延伸，佰维存储正累计投入 75.9 亿元建设东莞松山湖“芯成汉奇”晶圆级先进封测基地，主攻多层柔性堆叠与存算芯粒合封工艺，以制造一体化壁垒跨越单纯模组加工的周期局限。

在最上游晶圆原厂与利基存储端，全球市场份额已分别升至约 10% 且对外报价看齐韩系原厂的长鑫科技与长江存储，正与 SK 海力士、三星、美光同步采用附带预付押金与价格走廊的多年期长约锁定云厂需求。与此同时，由于海外三大原厂将洁净室产能优先拨付给高毛利的 HBM 与服务器 DDR5，主动关停退出部分 SLC NAND 与 DDR3/DDR4 成熟产线，承接工控、网络通信与车载刚需的东芯股份（三季度毛利率达 79.31%，单季归母净利环比增长 59% 至 8.30 亿元）与北京君正直接受益于供给收缩格局。从扩产周期推演，2026–2027 年新建晶圆厂与后道 TSV、混合键合产线将持续拉动前道设备与减薄键合材料需求，而全球绿地晶圆厂形成实质性规模晶圆产出并促使价格高位回落的拐点，预计将在 2027 年第四季度至 2028 年上半年出现。下表梳理了核心产业链公司的当期收入结构与下一阶段验证指标。

表 5：AI 服务器存力重构核心产业链公司收入来源、2026Q3 财报证据与下一步核心验证指标总表

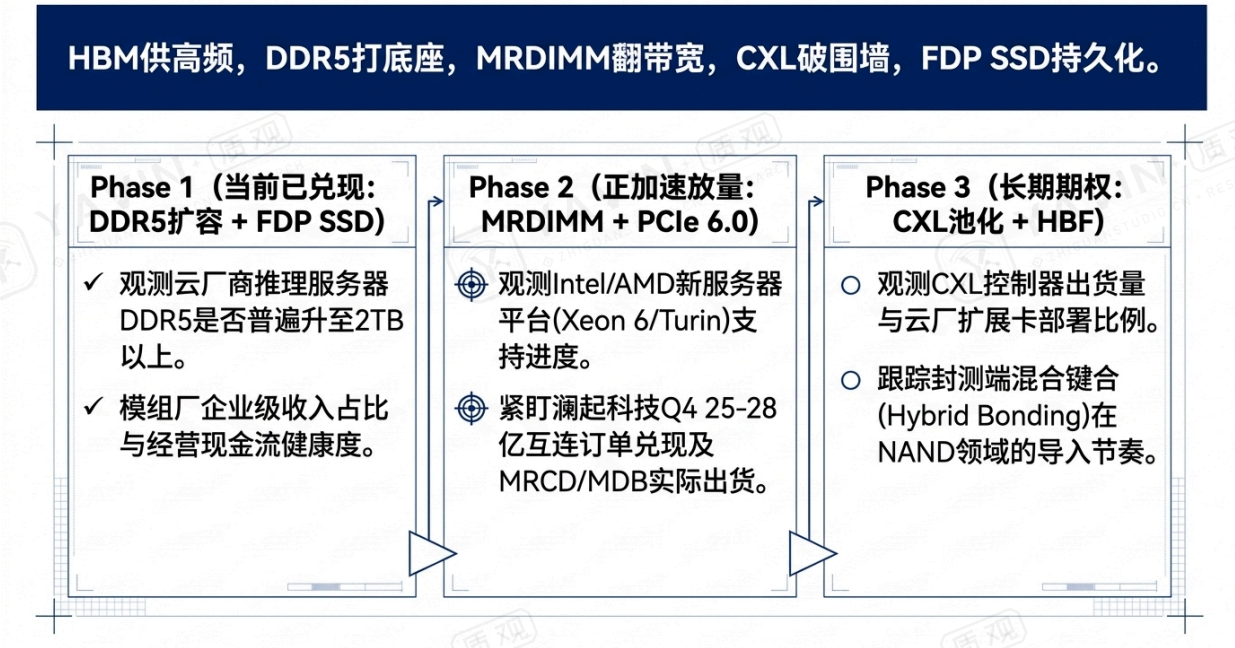
公司名称	核心产品与当前主要收入来源	截至 2026Q3 的商业化与财务证据	下一阶段核心验证指标（区分现实与期权）
澜起科技 (688008.SH)	DDR5 RCD、MRCD/MDB 套片、PCIe Retimer、CXL MXC	Q3 归母净利 **12.40 亿元（+162%）**，互连毛利率 **68.4%**；Q4 预计交付互连在手订单 **25.0–27.9 亿元**（当前主要由 DDR5 子代 RCD 驱动）	**MRDIMM“1+10”套片在互连总收入中的实际占比拉升节奏**；CXL 控制器商用出货量
聚辰股份 (688123.SH)	DDR5 SPD 集线器芯片、工业级与车规级 EEPROM	Q3 单季扣非净利 **1.32 亿元（+56%）**，毛利率 >55%，随服务器 DDR5 渗透率提升创单季新高	AI 服务器单机内存插槽满配率提升与新一代 SPD 芯片配套份额
Rambus (RMBS.US)	DDR5 RCD、MRDIMM 接口芯片组、HBM4/CXL 控制器 IP	Q3 产品收入创历史新高（**1.143 亿美元**），产品毛利率高达 **78.3%**	北美云厂商新一代 x86 服务器平台 MRDIMM 接口芯片导入放量斜率
Astera Labs (ALAB.US)	PCIe 5.0/6.0 Retimer、Scorpio 交换芯片、Leo CXL 控制器	Q3 营收指引 **5.40–5.60 亿美元（同比 +138%）**，Retimer 与交换芯片在 AI 机柜大规模放量	**Leo CXL 内存控制器在超大规模云厂商的实际规模化部署比例**
Marvell (MRVL.US)	数据中心光互连 DSP、定制 ASIC、Structera CXL 控制器	数据中心收入占比超 70%，推出内置硬件 LZ4 压缩并支持旧 DDR4 复用的 CXL 控制器	云厂商采用 CXL 扩展卡复用退役 DDR4 构建大容量缓存池的实际采购量

公司名称	核心产品与当前主要收入来源	截至 2026Q3 的商业化与财务证据	下一阶段核心验证指标（区分现实与期权）
大普微 / 联芸科技 (301666 / 688449)	企业级 PCIe 5.0 SSD 主控芯片、百 TB 级 QLC 企业级 SSD	联芸 Q3 营收同比 **+65.5%**；大普微自研主控与 61TB/122TB QLC SSD 导入头部互联网云厂	**头部云厂商的重复采购订单规模**、PCIe 5.0 / FDP 企业级主控收入确认占比
江波龙 / 香农芯创 (301308 / 300475)	企业级存储组合（eSSD + DDR5 RDIMM）、分销与自研封测	江波龙 H1 企业级收入 **+209%**，**Q3 单季经营现金流转正（+8.77 亿元）**；香农前三季营收破千亿	**企业级存储业务的持续毛利率水平**、自研主控导入比例与大客户粘性
佰维存储 (688525.SH)	嵌入式与消费级存储模组、企业级存储、晶圆级先进封测	Q3 受消费级波动影响净利环比回落 62%，正累计斥资 **75.9 亿元** 投建松山湖晶圆级先进封测基地	账面 240 亿元存货消化与现金流回流进度、**晶圆级先进封测产线良率与外部客户导入**
长鑫科技 / 东芯股份 (688825 / 688110)	服务器与移动端 DRAM 晶圆（长鑫）；SLC NAND 利基存储（东芯）	长鑫 Q3 毛利率 **89.53%**、单季净利 **544 亿元**、预收锁量 **467 亿元**；东芯 Q3 净利 **8.30 亿元（环比 +59%）**	服务器 DDR5/LPDDR5X 产能释放节奏与 **2027Q4–2028H1 全球新晶圆厂投产后的价格拐点**

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

5.2 围绕降低静态 HBM 依赖的三阶段商业化兑现节奏

EXHIBIT 10 **AI 服务器存力重构三阶段演进路标：当期规模兑现、平台爬坡与中长期期权**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

跳出单一芯片型号与局部财务指标，围绕 AI 服务器存力展开的跨层演进最终指向三项核心产业判断。第一，分层卸载削减的是长文本与高并发负载对 HBM 的静态驻留容量依赖，并未消除核心矩阵计算与每步高频索引检索对极高带宽显存的刚性需求，HBM 及其配套的极限晶圆减薄、TSV 与先进键合工艺仍长期

立于硬件价值金字塔顶端。第二，外溢的 TB 级温冷缓存正在系统性抬升主机内存、高速互连与企业级闪存的单机配额，但不同技术路线的商业兑现存在明确时序阶梯：大容量 DDR5 RDIMM、PCIe 5.0 Retimer 与搭载 FDP 的企业级 QLC SSD 已进入规模业绩兑现期；MRDIMM“1+10”接口套片与 CXL 单机内存扩展正随新一代服务器平台启动快速爬坡；CXL 跨服务器内存池化与 HBF 高带宽闪存则属于更偏中长期的技术期权。第三，存储产业的竞争边界已由单一晶圆介质的周期博弈，扩展为涵盖高速混合信号接口、缓存一致性协议控制器、4nm 企业级 SSD 主控、晶圆级先进封测与上层异步缓存调度的完整系统工程。

主要技术文献与核验来源

本报告工程实测基准与协议架构核验自 vLLM 与 SGLang 团队发布的 GLM-5.3 百万上下文 Hybrid HiSparse 技术报告、月之暗面与清华大学 Mooncake 架构论文、华为推理记忆管理器 UCM 金融长文本实测白皮书、DeepSeek 动态稀疏注意力与 3FS 文件系统架构报告、Meta 与 Google 联合提交的 NVMe TP4146 FDP 生产评估文献、SanDisk 与 SK 海力士 HBF 技术报告，以及 JEDEC JESD328 MRDIMM、HBM4 与 CXL 3.1/3.2 标准规范；境内外各上市公司营收、利润率、存货、合同负债及在手订单数据均取自 2026 年第三季度法定报告与公开披露公告。

06 CHAPTER 06 · TAKE-OR-PAY LTA SHIELD, $450 DDR5 & NEOCLOUD DSCR

存储周期的护盾与裂缝：LTA 长单锁量、终端挤出与算力信用传导

2026 年 10 月韩国半导体出口额升至 564.4 亿美元、同比增长 259%，头部存储原厂单季毛利率逼近八至九成，然而全球存储与算力板块在二级市场却频现剧烈震荡。穿透表观行情分歧，必须在资产负债表与信用链条上严格区分上游“一级云厂 LTA 长协防火墙”与下游“消费终端物料挤出及高杠杆 Neocloud 信用裂缝”。

6.1 HBM 三倍晶圆消耗与 2027–2028 照付不议长协的周期重塑

EXHIBIT 11 存储周期定价机制重塑：照付不议 LTA 长协护盾与现货合约双轨分化

CY2027 产能锁定	CY2028 产能覆盖	4Q26 整体 NAND 合约价	10 月消费现货单周杂音
100% 售罄	>50% LTA	+15%—20%	–9.19%
长单提前包圆核心 eSSD 产出	1–5 年长约 lock-in 持续延伸	eSSD 需求强势拉动逆势上涨	512Gb TLC 散货跌幅未穿击合约盘

铠侠 (Kioxia) LTA“照付不议 (Take-or-Pay)”防火墙与刚性价格底线机制 13,000—15,000 日元刚性底线

定价与履约维度	消费级现货市场表象（短线情绪波动）	AI 企业级 LTA 防火墙（法定现金流护城河）
价格传导隔离	10 月下旬 512Gb TLC 消费级现货单周下跌 –9.19%，引发市场误判周期见顶。	受企业级 eSSD 刚需拉动，4Q26 整体 NAND 合约价逆势上涨 +15%—20%。
照付不议与价格底线	无合约保护，原厂在淡季被迫低于现金成本降价倾销。	LTA 条款强制绑定 Take-or-Pay + Price Floors，在 13,000—15,000 日元区间筑起刚性利润底。
资产负债表蜕变	纯现货“周期股”高 Beta 暴涨暴跌，估值遭大幅折价。	CY27 100% 售罄 + CY28 逾半锁定，现金流能见度跨过传统猪周期。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

与过去三十年原厂盲目抢建产能、任由现货价格大幅起落的传统存储周期不同，本轮 AI 存力周期的供给弹性与定价底座已被双重锁定。在物理供给端，每生产 1 GB HBM3E/HBM4 因裸片尺寸扩大、TSV 通孔面积开销及多层堆叠良率折损，消耗的等效晶圆面积达标准 DDR5 的 3 倍以上；三大原厂 2026–2027 年新增洁净室产能几乎被 HBM4 与大容量企业级 RDIMM/MRDIMM 悉数吸收，从晶圆源头限制了通用颗粒的无序扩产。在商业契约端，面向北美四大云厂商与中国头部互联网企业，上游原厂普遍签下了直抵 2027–2028 年的刚性照付不议 LTA 长协，配套上下限价格走廊与高额预付保证金——例如美光已锁定千亿美元级五年期长协与逾 200 亿美元客户押金，长鑫科技截至 2026 年三季度末账面合同负债 70.36 亿元叠加其他非流动负债长期锁量押金 396.98 亿元，合计构筑起 467.34 亿元的预收缓冲垫。即便零售现货出现阶段性回踩，核心数据中心合约价仍在价格走廊保护下保持韧性。

EXHIBIT 12 核心存储原厂 2027–2028 年长单约束条款、预付保证金与扩产纪律对标

全球四大存储原厂 LTA 长协锁量、预收担保金比例与资本纪律全景矩阵 HBM + eSSD 预付款重构现金流

存储原厂	远期产能 LTA 锁定率	合约期与预收担保金比例	资本开支纪律与产业信号
铠侠 (Kioxia)	CY27 100% 售罄 CY28 >50% 覆盖	1–5 年期 LTA，内嵌刚性 Price Floors（价格底线）保护条约	拒绝无合同裸扩产，锁量优先供给高毛利大容量 eSSD。
闪迪 (SanDisk)	FY27 约 1/2 锁定 FY28 约 2/3 锁定	>4 年期 Take-or-Pay；客户预付担保金占底价收入 17.6%，单季现金流入 $4.65 亿	下游巨头直接以真金白银为原厂晶圆扩产垫付风险保证金。
美光 (Micron)	长协谈判直指 2028 年持续供不应求	客户预付担保金占底价收入 ~21%；强力托举 FY26 经营现金流冲高至 $440 亿	HBM3E/4 与企业级 SSD 双轮兑现，自由现金流历史性跃升。
SK 海力士	HBM 与大容量 eSSD 提前全额售罄	平均 ~5 年期长约 深度绑定预付定金与年度配额协议	主动搁置 Solidigm 上市，严控行业重复建设与资本结构纪律。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

6.2 450 美元内存的消费端挤出与高杠杆 Neocloud 偿债覆盖率测试

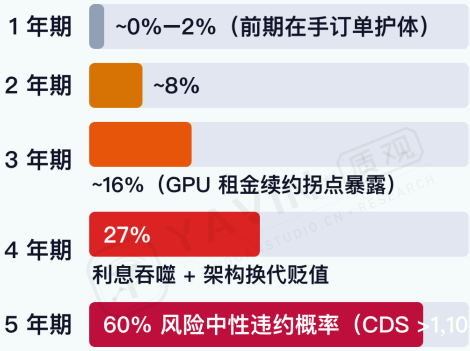
然而，上游数据中心存力的高毛利正沿着两条薄弱环节向下游传导成本挤压。第一重摩擦发生在缺乏顺价能力的传统消费电子整机端：2026 年秋末，单条 32GB DDR5 内存模组渠道成本升至约 450 美元，令主流轻薄本与中低端智能手机的内存加闪存物料占比从常规的 10%–12% 抬升至 22%–28% 以上。难以向终端消费者转嫁涨价的整机厂被迫推迟非刚性备货并下调单机内存配置指引，进而导致消费级现货价与数据中心企业级长协价出现双轨分化。

更具财务穿透力的第二重风险，潜伏在依赖高息债务抢购算力与存储集群的新兴算力租赁商 Neocloud 资产负债表上。不同于每年拥有数千亿美元经营性现金流的超大规模云厂商，二三线算力租赁商普遍通过特殊目的实体，以采购的 GPU 与高配存储服务器作为抵押品，借入年化利率高达 12%–15% 的私募信贷或高收益债。随着单机 HBM 与企业级大容量存储涨价推高初始资本开支、同质化裸卡时租费率却边际回落，2026 年 11 月部分头部 Neocloud 的 5 年期 CDS 信用违约互换利差一度突破 1,100bp。

EXHIBIT 13 **产业链信用传导裂缝：高杠杆 Neocloud 债务敞口与 CDS 信用利差走阔机制**

Neocloud 净债务 / EBITDA	2026E 资产负债率	CoreWeave 5Y CDS 利差	行业新增融资缺口
7.5× 杠杆	84%	>1,100 bp	$270 亿
大型科技公司仅 0.1×（极端分化）	大型科技仅 ~45%，风险堆于自身表内	隐含 5 年累计风险中性违约率 ~60%	2031–2032 年将迎 $150B 到期洪峰

① CoreWeave CDS 隐含累计违约概率曲线 次贷级信用警报



② 信用收缩向股票与上游硬件链条的传染 资产负债表冲击

股市先行定价折价：单月内 CoreWeave 估值大幅回撤 **–35%**，IREN 同步重挫 **–46%**，股权融资窗口收窄。

期限错配与再融资悬崖：用 8–14% 高息资产抵押贷款采购 3–4 年即过时的 GPU，2031–2032 迎来 **$150 亿集中到期高峰**。

上游传导链：一旦 Neocloud 高杠杆算力买家主动收紧新单，最先承压的不是自研巨头，而是高溢价裸卡与边缘存储采购。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

设单座万卡算力与存力中心总资本开支为 C_0，债务融资比例为 λ_D（普遍达 70%–80%），加权债务利率为 r_d，债务偿还期限为 T_{debt}（硬件真实技术迭代周期仅 3–4 年，即便会计折旧拉长至 6 年亦无法改变物理残值衰减），集群有效出租率为 u，单卡时租净费率为 p_h，扣除电力与运维开支后，项目偿债覆盖率 DSCR 受下式严格约束：

EQUATION · CORE QUANTITATIVE & ECONOMIC MECHANISM

$$\mathrm{DSCR} = \frac{N_{\mathrm{GPU}} \cdot 8760 \cdot u \cdot p_{\mathrm{h}} - \mathrm{OpEx}_{\mathrm{power+maint}}}{C_0 \cdot \lambda_{\mathrm{D}} \cdot \left(r_{\mathrm{d}} + \frac{1}{T_{\mathrm{debt}}} \right)}$$

当存储涨价推高初始资本开支且私募债务融资利率达 12%–15% 时，一旦集群出租率或单卡时租下滑，项目净现金流将迅速跌破 1.0x 偿债覆盖底线。

EXHIBIT 14 算力基础设施信用重定价压力测试：15% 融资成本对五年期项目现金流的冲击

模型基准规模与 CapEx	盈利与无杠杆回报基线	债务杠杆与追索结构	高息击穿临界点
100 MW	75% EBITDA	80% 负债	15% 利率
单位 CapEx $3,000 万/MW · 基准年收 $1,200 万/MW	无杠杆税前 IRR ~15% · 5 年期 GPU 周期	90% 有息债具备母公司追索权 (Recourse)	DSCR 由 1.44× 压穿至 1.00× 破产红线

15% 借款成本如何击穿 5 年期 GPU 算力中心现金流底线（100MW 财务压力模型）

银行预警线 1.20× · 违约红线 1.00×

偿债覆盖率方程： $DSCR = \frac{\text{NOI（算力租金净营业收入）}}{\text{Principal + Interest（刚性本息偿付）}} \leq 1.00\times$ ⇒ 触发母公司追索与 GPU 抵押物踩踏清算

基准情景（10% 贷款利率）	利率飙升（15% 二级债成本）	双重打击（年收 <$1,000 万/MW）
累计 1.44×	DSCR → 1.00×	DSCR < 1.00×
首年 DSCR 仅 1.25×，勉强立于银行 1.20× 预警线之上，安全垫极薄。	新增借款成本升至 15%，全部 EBITDA 仅够勉强偿还高息本息，零自由缓冲。	租金跌破 $1,000 万/MW 即直接击穿 1.0×，引爆 90% 母公司追索权连带违约。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

压力测试情景	集群出租率 u	裸金属等效时租 p_h	融资利率 r_d	项目偿债覆盖率 (DSCR)	对上游存储与算力供应链的实质传导影响
基准景气情景（主权长约锁定）	92%	$2.20 / GPU·h	8.5%	**1.68x（安全）**	按期履约提取企业级 RDIMM / QLC eSSD 长单
中度挤压情景（租金回落 + 利差走阔）	82%	$1.55 / GPU·h	12.5%	**1.06x（紧平衡）**	推迟非绑定扩产期权，优先采用 MRDIMM/eSSD 降本
极端信用冻结（无长约裸租 + 15% 高息）	<70%	≤$1.30 / GPU·h	**15.0%**	**0.74x（击穿违约线）**	二手服务器折价抛售冲击现货，但**四大云厂资产负债表防火墙阻断系统性违约蔓延**

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

上述三档压力情景清晰揭示了信用传导的边界：当无长约保护的高杠杆裸租项目在 15% 借贷成本与租金回落双重挤压下击穿 0.74x 偿债底线时，局部违约将引发二手服务器折价出清并冲击渠道现货；但北美四大云商与中国头部互联网大厂稳健的资产负债表及照付不议长协，有效阻断了局部信用风险向上游晶圆原厂的系统性蔓延。本轮存力重构中具备持续现金流兑现能力的主体集中在三类节点：向上掌握先进晶圆产能与长协定价走廊的原厂，向中卡住 MRDIMM、CXL 及 NVMe FDP 主控咽喉的芯片企业，以及向下压降每百万 Token 综合存力摊销成本的系统架构定义者。

STATUTORY FILINGS × SYSTEM ARCHITECTURE × SUPPLY CHAIN VERIFICATION

研究方法论、数据核验口径与核心产业判断

S1 · JEDEC / OCP 标准与开源推理框架实证（HARDWARE & SOFTWARE CO-VERIFICATION）

四层存力参数严格核验自 JEDEC DDR5 MRDIMM、CXL 2.0/3.1、OCP NVMe FDP 技术规范及 vLLM / Mooncake / Dynamo 生产级缓存卸载实测。

S2 · 2026Q3 法定财报与跨国财年对齐（STATUTORY Q3 2026 ALIGNMENT）

SK 海力士、三星、美光（FY26Q4）、铠侠（FY26Q2）及长鑫科技、澜起科技、佰维存储、江波龙、兆易创新均对齐至自然年 2026Q3 法定披露口径。

S3 · 资产负债表与信用链条穿透（LTA CORRIDOR × DSCR STRESS TEST）

拒绝将现货渠道短期波动混同于数据中心长协崩塌，通过合同负债锁量比与 SPV 算力项目偿债覆盖率（DSCR）方程定量界定风险传导边界。

S4 · 公开非商业研究声明（NON-COMMERCIAL PUBLIC SHARING）

本报告彻底剔除一切未公开公开交流记录、卖方机构评级与商业推广信息，仅供存储架构工程与产业研究人员公开交流。

非商业免费分享与免责声明：本报告由「Yavin 质观」基于公开技术文献、学术顶会论文、国际通用行业统计及上市公司法定披露文件独立整理，仅作个人公开研究免费交流分享，非商业用途。报告内容在任何情况下均不构成任何证券投资建议、个股评级、目标价预测或商业采购依据，读者应当独立判断并自行承担风险。