YAVIN 质观

RESEARCH MONOGRAPH · 2026.10

DEEP TECH & CAPITAL NARRATIVE

打破内存墙与重算税

AI存力解耦与超节点互连产业链全景报告

HBM4 · SOCAMM2 · CXL 4.0 · G3.5闪存 · 超节点液冷与PIM/CIM深度研究



Yavin质观

硬科技与资本叙事研究者，长期关注技术、产业、政策与资本化逻辑，
书写技术时代、现代生活与精神处境中的缓慢观察。

扫码关注公众号

EXECUTIVE SUMMARY · 核心执行摘要

六大核心产业判断与结论锚点

Yavin质观 · 旗舰智库专刊

50%	68%	85.95%	<$8万/PB	39.7倍
Agentic推理因KV逐出消耗的重算税	北美五大云厂新增CapEx流向存储与内存	美光FQ1 27指引毛利率超越逻辑代工	MinIO MemKV G3.5闪存层单位容量TCO	CXL-PNM运行DeepSeek-R1硬件降本倍数

01 硬件BOM与重算税

算力账本倒挂：一半算力耗在重算，云厂近七成预算转向存力

长文本与多步Agent使上下文缓存（KV Cache）成倍膨胀，缺乏共享内存池的集群逾 **50%** 算力耗费在"踢出显存、下次再从头重算"的无用功上。外接共享内存与闪存池将有效利用率拉回 **90%+** 后，同等吞吐可少买 **40%–45%** 高价显卡；预计2027年内存与闪存将占去主流云厂 **68%** 的组件采购开支。

02 软件与G3.5闪存减税

软件与闪存省钱：读写分离盘活旧卡，白盒闪存造价直降九成

软件侧通过SGLang前缀树复用与月之暗面 **Mooncake** 读写分离，让存量显卡长文本有效吞吐提升 **5倍以上**。硬件侧在机柜增设 **G3.5高速闪存池**，用BlueField-4 DPU配合MinIO MemKV直写白盒NVMe盘，将64K首字延迟从 **53秒压至0.7秒**，每PB存储造价从传统阵列的百万美元压至 **8万美元以内（降本超90%）**。

03 HBM4与3nm底座代工

存储利润反超：台积电代工底座，3倍晶圆消耗推升86%毛利

HBM4 将位宽翻倍至2048-bit、单栈带宽达 **2.8–3.3 TB/s**，英伟达要求三大原厂为Rubin Ultra定制的HBM统一采用 **台积电3nm工艺** 代工逻辑底座。由于造1GB HBM需消耗普通内存 **3至4倍的晶圆面积**，产能挤出引发全行业存储紧缺，推动美光FQ1 27毛利率指引冲上 **85.95%**，盈利能力反超逻辑芯片设计公司。

04 三级内存与隐形冠军

三级内存改造：告别板载焊死内存，内存就地计算降本40倍

服务器转向"HBM4 + 可插拔主存 + 机柜内存池"三级架构：**SOCAMM2可插拔模组** 解决板载内存无法维修痛点，以传统内存 **1/3体积、功耗和成本** 提供2.5倍带宽，带飞 **嘉泽端子** 与 **优群科技（毛利率48.23%）**；x86转向MRDIMM推动全球份额第一（**36.8%**）的 **澜起科技** 净利预增60%；**CXL近存计算** 就地处理注意力求和，跑DeepSeek-R1将每百万Token硬件成本从 **24.24美元压至0.61美元（降本39.7倍）**。

05 超节点互连与液冷

互连与基建卡点：78层背板推迟新机柜，2千瓦单卡催生液冷

英伟达借 **NVLink Fusion** 收编自研芯片，但其576卡Kyber机柜受困于 **78层PCB背板良率** 推迟至2028年；AMD等开放阵营在专用交换芯片量产前，暂借 **博通Tomahawk 6** 跑通超节点互连。随着新一代芯片单卡功耗升至 **1,400W–2,250W**、单柜达 **155千瓦**，单柜液冷采购升至 **70万–100万元**，**英维克** 打入英伟达全球三大液冷白名单。

06 新型介质与中国突围

新介质与中国突围：无电容3D内存破局，全光超节点换道超车

面对平面内存微缩极限，存储界正演进向 **4F²垂直沟道** 与NEO **无电容3D X-DRAM**（实测<10ns延迟、10倍密度）。国内产业界走通"以存代算、以网补卡"路径：算法端 **DeepSeek MLA** 削减 **93.3%** 显存；系统端 **华为昇腾960** 首发商用 **NPO全光超节点**（**4,096卡** 统一内存池、省电550kW）配 **OceanStor M900记忆存储**（SSD寿命延长 **16倍**）；制造端 **长鑫存储** 首发量产LPDDR6、16nm DDR5良率超 **90%** 并推进 **12层HBM3E** 试产。

TABLE OF CONTENTS · 报告目录

全文八大核心章节与研究导览

19 幅数据图表 · 12 张全景矩阵表 · 297 项溯源信源

01 ECONOMIC INFLECTION & RECOMPUTE TAX

算力经济学的转折点——50%"重算税"账本与云厂资本开支向存力迁移

- 1.1 芯片与机柜BOM成本拆解：计算裸片占比降至四分之一，内存与封装占去64%
- 1.2 长上下文KV缓存膨胀：从"驱逐即重算、跨机即重算"到50%算力空转损耗
- 1.3 云厂商采购账本重构：少买40%–45%旗舰GPU，2027年存储占组件开支68%

02 SOFTWARE SCHEDULING & KV DISAGGREGATION

软件与调度层的第一道防线——前缀缓存、非对称量化与P/D读写分离

- 2.1 推理引擎内生复用：vLLM链式块哈希与SGLang统一基数树（命中率60.8%）对比
- 2.2 生产级调度与显存压缩：PEEK双树遍历、多租户"缓存加盐"隔离与KIVI 2-bit量化
- 2.3 计算与状态物理解耦：月之暗面Mooncake P/D分离架构、LMCache非前缀复用与路由边界

03 G3.5 FLASH CONTEXT MEMORY LAYER

存储金字塔的G3.5层革命——白盒闪存池如何以8万美元/PB重构企业存储?

- 3.1 传统存储层级的G3–G4断层与G3.5推理上下文内存层（ICMS）定位
- 3.2 BlueField-4 DPU + MinIO MemKV零拷贝直连架构与64路并发"显存容量悬崖"实测
- 3.3 企业级存储格局洗牌：不足8万美元/PB白盒闪存方案与VAST Data软件化转型

04 TIER-1 HBM4 & BASE DIE FOUNDRY SHIFT

第一级内存（G1）与晶圆黑洞——HBM4逻辑底座代工与存储厂85.95%毛利率

- 4.1 HBM4 2048-bit宽总线演进、SPHBM4有机基板降本分支与HBM4E 4TB/s路线图
- 4.2 逻辑底座（Base Die）晶圆化：英伟达"台积电3nm底座令"与三星4nm半数产能锁定
- 4.3 3至4倍晶圆面积折损与利润倒挂：美光85.95%毛利率指引与775微米键合设备分工

05 TIER-2 SOCAMM2 & TIER-3 CXL 4.0 POOLING

第二、三级内存（G2/G3）重构与隐形冠军——SOCAMM2、MRDIMM与CXL 4.0

- 5.1 第二级内存模块化：JEDEC SOCAMM2可插拔标准、优群科技连接器与澜起MRDIMM套片倍增
- 5.2 第三级内存池化：CXL 4.0（128 GT/s）亚200ns架构、绑定端口设计与Astera Labs放量
- 5.3 从"注意力在GPU"到"注意力在内存"：CXL近存计算（PNM）跑DeepSeek-R1降本39.7倍

06 SCALE-UP FABRIC & LIQUID COOLING

Scale-Up超节点互连与物理基建卡点——NVLink、UALink、ESUN与155kW液冷

- 6.1 AI后端交换机反超前端网络：NVLink 5/6、NVLink Fusion生态收编与Kyber 78层背板延期
- 6.2 开放协议演进与过渡：UALink 2.0网内计算、AMD Helios借道博通Tomahawk 6与OCP ESUN
- 6.3 机房物理硬约束：单卡突破2千瓦、155kW秒级功率摆动、楼板承重极限与液冷白名单

07 NEXT-GEN MEMORY MEDIA & PIM/CIM

下一代内存介质演进——4F²垂直沟道、无电容3D X-DRAM、非易失内存与存算一体

- 7.1 全球内存IC市场迈向3,155亿美元：DDR6/LPDDR6接口升级与4F²垂直沟道晶体管（VCT）
- 7.2 突破平面1T1C极限：三星16层VS-DRAM与NEO无电容3D X-DRAM硅片级实测验证
- 7.3 28nm以下eNVM替代（MRAM/PCM/ReRAM/FeRAM）与近存/存内计算（PIM/CIM）双轨演进

08 CHINA SYSTEM BREAKTHROUGH & STRATEGY

中国产业链"以存代算、以网补卡"系统突围与2026–2028选型投资框架

- 8.1 算法与软件层减负：DeepSeek MLA削减93.3%显存、腾讯FlashMemory与存量卡盘活
- 8.2 系统工程换道超车：华为昇腾960全球首发商用NPO全光超节点、灵衢统一编址与OceanStor M900
- 8.3 本土制造底座与全局配置策略：长鑫首发LPDDR6与12层HBM3E试产、存算芯片及三阶选型框架

APPENDIX DATA METHODOLOGY & 297 VERIFIED SOURCES

附录：参考文献与核心信源索引

- 涵盖算力经济学与重算税、推理引擎与G3.5闪存、HBM4与三级协同内存、超节点互连与2,250W液冷、新型内存介质与PIM/CIM及中美产业链共 297 项可溯源原始信源

01 CHAPTER 01 · ECONOMIC INFLECTION & RECOMPUTE TAX
算力经济学的转折点——50%“重算税”账本与云厂资本开支向存力迁移

1.1 芯片与机柜BOM成本拆解：计算裸片占比降至四分之一，内存与封装占去64%

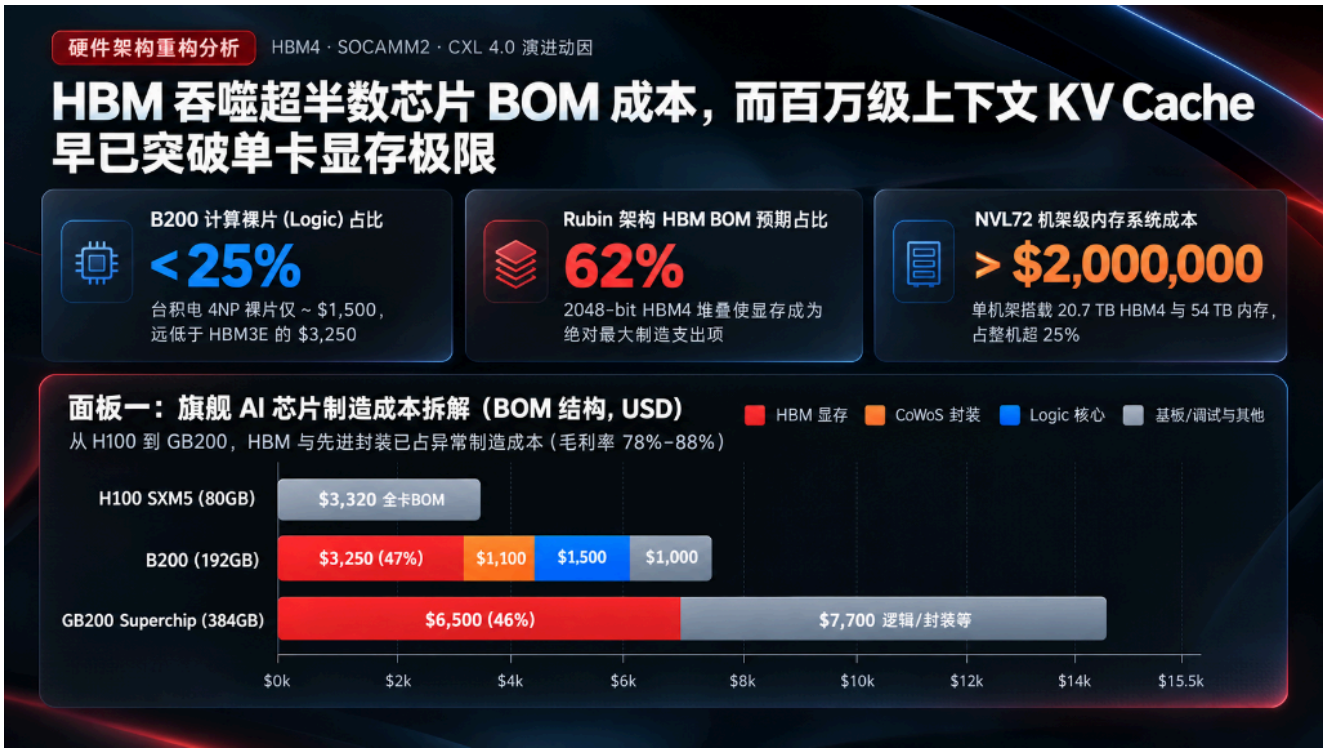
2026年至2028年，全球人工智能基础设施的投入重心正在发生根本性转移。在算力集群迈向十万卡与吉瓦级数据中心的过程中，制约大模型商业化的首要瓶颈已不再是单卡浮点运算峰值（FLOPS），而是内存子系统的容量、带宽与物料清单（BOM）成本。过去三年由预训练主导的“堆叠算力卡”周期，正被长上下文推理、多模态交互与多步智能体工作流驱动的“存力与互连中心化”周期取代。

拆解主流AI加速卡与整机柜的制造BOM账本，硬件价值分配的逆转已经非常清晰：

01 **NVIDIA H100 SXM5 世代（计算裸片主导期）**：单卡总制造成本约为 **3,320美元**，芯片设计厂商享有高达 **88%** 的产品毛利率，逻辑计算芯片在硬件成本中占主导地位。

02 **NVIDIA Blackwell（B200 与 GB200）世代（内存与封装反超期）**：随着HBM合同均价升至约 **17美元/GB**，单颗B200加速卡的总制造成本升至 **6,750至6,850美元**。其中，由台积电4NP工艺代工的计算逻辑裸片（Logic Die）成本仅约 **1,500美元**，占总BOM不足四分之一（约 **22%–23%**）；相比之下，192GB HBM3E显存采购成本高达约 **3,250美元（占47%–52%）**，叠加约 **1,100美元** 的CoWoS-L先进封装成本（占 **17%**）与约1,000美元的基板测试成本（占15%），**内存与先进封装两项合计占去64%的芯片制造成本**。在GB200 Superchip（1颗Grace CPU + 2颗B200 GPU）中，总制造成本升至 **14,200美元**（毛利率收敛至78%），仅372GB HBM3E显存一项便高达 **6,500美元**，接近两颗完整H100芯片的全部制造成本。

03 **NVIDIA Vera Rubin（VR200）世代（存力中心化确立期）**：在单颗售价达 **50,000美元** 的Rubin GPU中，HBM4占芯片级BOM的比重进一步升至 **62%**。从整机柜维度看，满配Vera Rubin NVL72机柜（36颗Vera CPU + 72颗Rubin GPU）总建造成本达 **780万至880万美元**（高于GB200 NVL72的300万美元与GB300 NVL72的500万美元），其中仅内存子系统（20.7 TB HBM4与54 TB LPDDR5X）采购成本即超过 **200万美元（占整柜建造成本的25%以上）**，单柜内存开支较前代增长 **435%至485%**。成本敏感度测算显示，HBM采购均价每波动10%，单颗旗舰GPU即产生约290美元的直接成本增量。



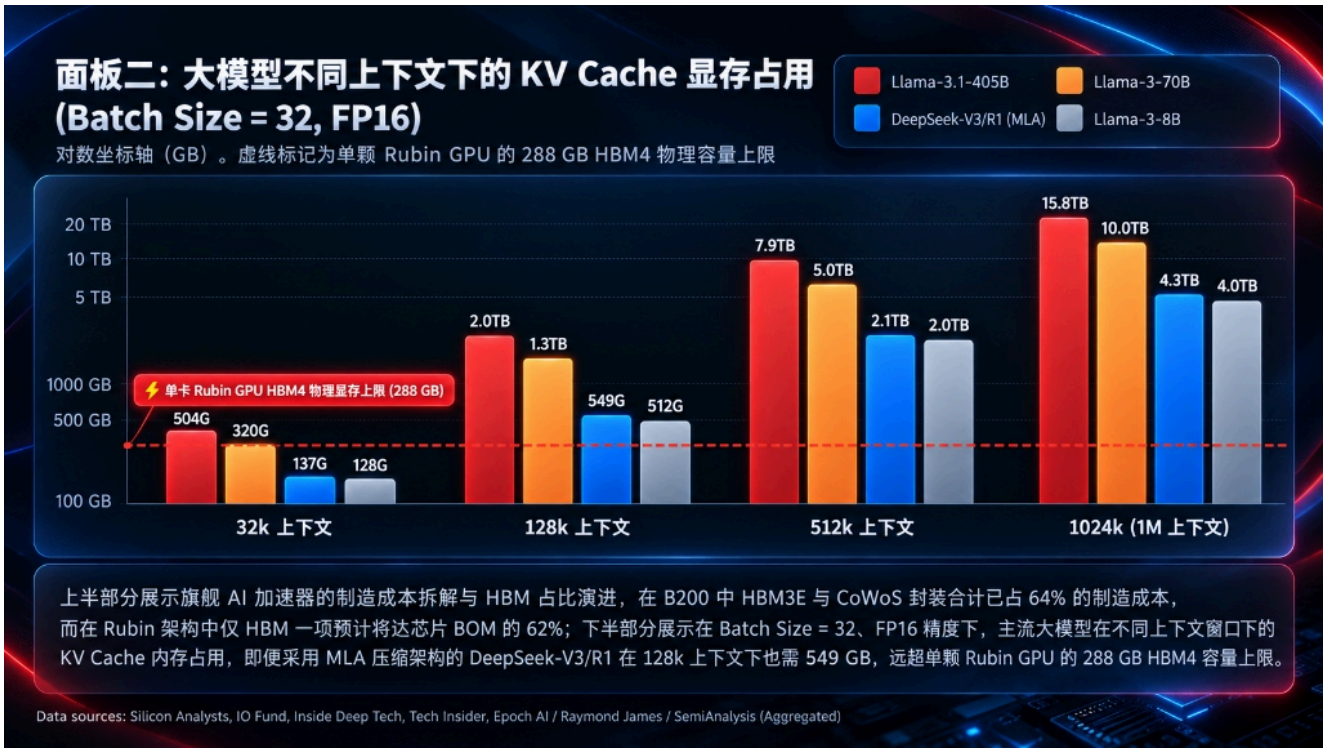
■ 图01：主流AI加速器制造BOM成本结构拆解（B200 vs. GB200 vs. Vera Rubin） — 在B200单卡6,850美元BOM中，台积电4NP计算裸片仅占22%（1,500美元），而HBM3E与CoWoS-L封装合计占64%；Vera Rubin世代HBM4占比进一步升至62%。

1.2 长上下文KV缓存膨胀：从“驱逐即重算、跨机即重算”到50%算力空转损耗

在供给侧HBM成本攀升的同时，需求侧大模型工作负载向长上下文、多小时智能体工作流（Agentic Workflows）与强化学习后训练演进，引发了键值缓存（KV Cache）容量的急剧膨胀。在许多生产级推理机房中，监控面板上维持在 **90%以上** 的GPU利用率掩盖了真实的资源浪费：监控指标只能反映流式多处理器（SM）与张量核心处于繁忙状态，却无法区分算力是在生成新的业务Token，还是在重复执行历史上下文的预填充（Prefill）计算。

这种现象源于Transformer自回归推理的两阶段非对称特征。在预填充（Prefill）阶段，系统将长度为 S 的输入提示词一次性送入模型，执行复杂度为 $O(S^2)$ 的全序列自注意力计算，生成键张量（K）与值张量（V）并驻留于GPU显存（HBM）中；在解码（Decode）阶段，模型逐Token自回归生成输出，每一步均需从HBM中全量读取此前累积的所有历史KV状态。单次推理所需的KV Cache总容量遵循严格线性膨胀关系（容量等于层数、KV头数、头维度、参数字节数、并发批大小 B 与序列长度 S 之积的两倍）。

当上下文窗口从32K扩展至128K、512K乃至1M Token，且维持标准并发批次（如 B=32）以摊薄权重读取开销时，KV Cache体积迅速突破单卡乃至整节点的HBM容量上限。以32路并发为例，在128K上下文下，即便经过MLA低秩压缩的DeepSeek-V3/R1，其KV Cache仍达 **549 GB**，接近单颗Rubin GPU（288GB）显存容量的两倍；而在1M上下文下更达到 **4,392 GB**，采用GQA架构的Llama-3.1-405B则高达 **16,128 GB**。



■ 图02：不同上下文长度与模型架构下的KV Cache显存占用（B=32标准并发） — 在32路并发下，当上下文从32K扩展至128K与1M Token时，主流大模型的KV Cache从数百GB激增至数万GB，远超单卡288GB HBM物理容量上限。

TABLE 01 · 第一章 · KV缓存膨胀与重算税 Yavin质观 · 核心数据矩阵

主流开源大模型在不同上下文长度与并发批次下的 KV Cache 内存膨胀测算

长上下文与高并发下 KV Cache 内存需求远超静态权重，触发高达 50% 的集群重算税（Recompute Tax）

模型与注意力机制	静态模型权重	单请求 KV 体积 (128K)	32 并发 KV 池需求 (128K)	系统瓶颈与重算税表现
Llama-3.1-70B GQA (8 KV Heads)	**140 GB** (FP16) 70 GB (FP8)	**40.0 GB** (FP16) **20.0 GB** (FP8) 每Token占 320 KB	**1,280 GB** (FP16) **640 GB** (FP8) 达权重体积 9.1 倍	8×H100 (640GB) 装入权重后仅剩500GB；高并发频繁触发缓存驱逐，**重算税达 45%–52%**
Llama-3.1-405B GQA (8 KV Heads)	**810 GB** (FP16) 405 GB (FP8)	**63.0 GB** (FP16) **31.5 GB** (FP8) 126层深度放大	**2,016 GB** (FP16) **1,008 GB** (FP8) 跨双机16卡仍溢出	单节点显存彻底击穿，跨节点流水线并行通信阻塞，**解码带宽利用率仅 15%–22%**
Qwen3-235B-A22B MoE + GQA	**470 GB** (FP16) 235 GB (FP8)	**29.5 GB** (FP16) **14.8 GB** (FP8) 稀疏激活省算不省存	**944 GB** (FP16) **472 GB** (FP8) Agent多轮会话倍增	MoE虽仅激活22B参数，但全量权重常驻+长上下文KV暴涨，**急需CXL/G3.5分层卸载**

模型与注意力机制	静态模型权重	单请求 KV 体积 (128K)	32 并发 KV 池需求 (128K)	系统瓶颈与重算税表现
DeepSeek-V3 / R1 671B MoE + MLA	**671 GB** (原生FP8) 激活 37B	**17.2 GB** (BF16) **8.6 GB** (FP8) 每Token仅 70 KB	**549 GB** (BF16) **275 GB** (FP8) 较GQA压缩 93.3%	低秩联合压缩大幅缓解显存压力，但百万Token与高并发下仍达TB级，**配合3FS/Mooncake满血释放**

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

在真实生产环境中，触发KV Cache丢失并导致“重算税（Recompute Tax）”的机制主要有两类：

01 **容量溢出驱逐（驱逐即重算）**：在Agent交互与多轮对话中，单次会话上下文常累积至124K Token以上，且包含约 **70%的高频智能体交互步（平均间隔约2.5秒）** 与 **30%的人类阅读或外部工具调用静默期（平均间隔约40秒）**。在数十秒的静默期内，新涌入的并发请求会迅速占满HBM空间，迫使本地内存管理器按LRU策略将旧会话的KV状态清除。当用户或工具在40秒后返回下一轮消息时，系统只能从头重新执行全量Prefill计算。

02 **无状态路由漂移（跨机即重算）**：分布式推理集群普遍采用多副本负载均衡。当同一会话的下一轮请求被无状态调度器分发至集群内另一台物理服务器时，由于原有的KV Cache保存在上一节点的本地显存或本地SSD（G3层）中，新接管的GPU无法直接读取历史状态，只能从第0个Token开始重新计算。

这两类重算机制直接拖累了服务性能与财务账本：在64K长文本下，频繁重算使首字延迟（TTFT）恶化至 **53秒**，128K高并发下更易触发显存耗尽（OOM）报错；在一座拥有 **128张高端GPU** 的中型推理集群中，若缺乏跨节点共享上下文池，缓存未命中率普遍高于 **50%**——按云端高端GPU每小时2至4美元的租赁成本计算，**仅重复计算历史前缀一项，每年就白白消耗约200万美元的算力租赁与电力开支**。

1.3 云厂商采购账本重构：少买40%–45%旗舰GPU，2027年存储占组件开支68%

2026年下半年，北美超大规模云服务商（Google、AWS、Meta、Microsoft）及头部AI实验室的资本开支继续保持高位：2026年头部云厂年度总CapEx处在 **7,250亿至8,500亿美元** 区间（Alphabet单年CapEx指引上调至1,950亿–2,050亿美元），Citi 2026全球TMT峰会预测 **2027年全球超大规模云厂CapEx将突破1万亿美元**；Anthropic在2026年3月的年化经常性收入（ARR）已逼近 **200亿美元**，全美AI数据中心电力消耗亦从2023年的3 GW升至2026年的 **28 GW**。

但在万亿美元级CapEx总额内部，组件采购结构正在发生深刻调整。通过引入 **Prefill/Decode（P/D）读写分离路由** 与 **三级内存/G3.5上下文卸载**，集群的有效GPU利用率（Useful GPU Utilization）从约 **50%提升至90%以上**。按算力守恒测算（1 – 50%/90% = 44.4%），**在维持**

同等有效Token吞吐量的前提下，云厂商对单价3万–5万美元旗舰通用GPU的冗余采购需求直接减少了40%至45%，单Token综合推理成本随之下降 **40%至60%**。Meta在生产环境中的实测同样证实，引入CXL内存池化回收闲置DRAM后，特定推理业务所需的服务器整机采购量减少了**25%**。

TABLE 02 · 第一章 · 算力经济账重构 Yavin质观 · 核心数据矩阵

2025–2027 年主流 AI 超节点机柜系统级 BOM 结构与 TCO 对比

从“堆砌昂贵GPU”向“三级内存+G3.5上下文闪存池”跃迁，同等吞吐少买 40%–45% 旗舰算力卡

对比维度	GB200 NVL72 (2025)	GB300 NVL72 (2026)	VR200 NVL72 (2026Q4)	VR200 + G3.5/CXL 解耦机柜
算力与显存底座	72×Blackwell B200 **13.8 TB HBM3E** 17.0 TB LP5X (板载焊死)	72×Blackwell Ultra **20.7 TB HBM3E (12H)** 18.0 TB SOCAMM1	72×Rubin GPU + 36×Vera **20.7 TB HBM4** **54.0 TB SOCAMM2**	72×Rubin + **54TB SOCAMM2** + **16TB CXL 4.0池** + **500TB BlueField-4 G3.5**
单机柜系统造价	**约 300 万美元** (初期基准配置)	**约 500 万美元** (HBM3E 12H与液冷升级)	**780万–880 万美元** (单柜内存超200万美元)	**仅增配 <8 万美元/PB 闪存** (增幅不足整柜 2%)
存力占BOM比重	**38%–42%** 逻辑GPU仍占主导	**48%–52%** 与逻辑芯片平分秋色	**60%–64%** HBM4+SOCAMM2 反超	**65%–68%** 确立“存力中心化”结构
64K长文本重算税	**48%–55%** 频繁驱逐，算力空转	**35%–42%** SOCAMM缓解单机溢出	**28%–35%** 跨机调度仍触发重算	**< 8% (有效利用率 >90%)** PB级全局KV微秒级直取
等效吞吐TCO降幅	基准线 (1.00×) 64K首字延迟高达 53 秒	单Token成本 **-22%** 单机内存容量红利显现	单Token成本 **-45%** 单卡带宽升至 19.2–22 TB/s	单Token成本 **-40% 至 -60%** **等效少买 40%–45% GPU**

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

被释放出来的40%–45%“重算冗余GPU预算”，正沿着三条明确的路径重新分配：

01 **流向广义存力（DRAM + NAND将在2027年占68%组件CapEx）**：2026年仅内存一项已占超大规模云厂总CapEx的 **30%**，数据中心消耗了全球约70%的高端内存产出。据TrendForce产业模型预测，**到2027年，DRAM与NAND闪存合计将占主流云服务商组件级资本开支（Component CapEx）的68%**。瑞银（UBS）拆解模型亦显示，2027年全球AI服务器平均DRAM搭载容量将达 **14,754 GB（约14.75 TB，同比增长64.7%）**，为传统企业级服务器（520 GB）的 **28.4倍**。

02 **流向P/D分离催化的定制化推理ASIC（出货增速达82%）**：由于P/D分离将推理拆解为计算受限的Prefill与带宽受限的Decode，云厂商得以分别配置高性价比硬件（如搭载128GB GDDR7的NVIDIA Rubin CPX或自研ASIC）。2026年AI服务器中定制ASIC出货增速达 **82%**——OpenAI联合博通基于台积电N3P（3nm）与HBM4开发的 **700W专用推理芯片“Jalapeño”**，在运行DeepSeek-R1等模型时的每千瓦Token吞吐达到NVIDIA Blackwell的 **1.5至1.9倍**，推动博通AI业务单季营收突破 **167亿美元**，Marvell定制XPU业务迈向20亿–40亿美元。

03 **流向机柜级互连与DPU存储控制面（NVIDIA的平台化对冲）**：面对单卡冗余需求被压缩，NVIDIA一方面将Vera Rubin NVL72整柜售价提升至 **780万–880万美元**（高交互200–250 tok/s下每百万Token成本反比GB200降3倍）；另一方面推出CMX平台、BlueField-4 STX DPU与cuObject v1.0接管G3.5存储入口（其网络业务TTM营收已达 **257.4亿美元**），并于2026年9月联手联发科（35亿美元合作）推出 **NVLink Fusion**，允许云厂自研ASIC接入NVL72背板，通过NVSwitch机柜互连与DPU控制面维持生态主导权。

02 CHAPTER 02 · SOFTWARE SCHEDULING & KV DISAGGREGATION
软件与调度层的第一道防线——前缀缓存、非对称量化与P/D读写分离

2.1 推理引擎内生复用：vLLM链式块哈希与SGLang统一基数树（命中率60.8%）对比

消除高达50%的“重算税”是一项跨层系统工程，涵盖模型注意力压缩、KV量化、引擎级前缀复用、分布式P/D分离以及G3.5闪存池化五个核心技术维度（见表2-1）。

TABLE 03 · 第二章 · 去重算税工程全景 Yavin质观 · 核心数据矩阵

消除 AI 推理“重算税”的五大技术维度与工程边界矩阵

从模型算法源头压缩、引擎基数树复用到 P/D 分离与 G3.5 闪存卸载的全栈协同减税方案

技术减税维度	代表架构与开源方案	核心量化降本与提速实绩	工程适用边界与安全防线
模型注意力机制 (源头缩减体积)	**DeepSeek MLA** (低秩压缩) **NSA** (原生稀疏注意力) **TransMLA / MLAfication**	• MLA+量化将KV体积 **削减 93.3%** • 后训练无损迁移实现 **5.35×–10.6× 吞吐提升**	原生MLA需在预训练固化；存量GQA模型需经正交旋转与校准转换
KV Cache 量化 (精度换取空间)	**FP8 KV** (主流生产标配) **KIVI 2-bit 非对称量化** **GrowPage 动态预算分配**	• KIVI 将峰值显存削减 **2.6×**、吞吐提升 **2.35×–3.47×** • GrowPage 每瓦特吞吐高 **2.09×–6.13×**	Key按通道、Value按Token非对称量化；缺乏融合反量化内核会使解码变慢
推理引擎调度 (块级前缀复用)	**vLLM APC** (链式块哈希) **SGLang 统一基数树** **PEEK 协同双树遍历调度**	• SGLang v0.5.20 缓存命中率升至 **60.8%** • PEEK 带动 TTFT 降低 **7.1×–7.9×**、吞吐增 **3.6×–4.5×**	多租户共享面临 **前缀计时侧信道攻击**；必须强制启用 **缓存加盐 (Cache Salting)**
P/D 分离与池化 (算存解耦调度)	**Mooncake** (月之暗面) **LMCache** (位置无关分块) **分阶段功率控制**	• Mooncake 长文本吞吐提升 **525%** (少用43% GPU) • LMCache 多轮RAG问答吞吐提升 **15×**、省电 **32.3%**	短提示词 (<500 Token) 跨网开销大于收益；亲和性路由需防范副本负载倾斜
G3.5 闪存卸载 (PB级共享池)	**MinIO MemKV** + BlueField-4 **VAST Data CNode on DPU** **华为 OceanStor M900**	• MemKV 将 64K 首字延迟从 **53秒压至 703毫秒** • 华为 M900 令 SSD 擦写寿命延长 **16 倍**	摒弃 4KB 随机 I/O 与传统 POSIX 协议；采用 **2–16MB 大块与 800G RDMA 零拷贝**

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

在KV数据离开GPU显存之前，软件层的第一道防线是推理服务引擎（Serving Engine）内部的前缀缓存（Prefix Caching）。前缀缓存本质上是一项针对Prefill阶段的算力跳过机制——它将已计算完成的KV张量转化为可跨请求检索的持久索引，直接跳过命中前缀的Transformer前向计算。围绕如何检索并复用PagedAttention物理块，主流开源引擎形成了两条技术路线：

- **vLLM自动前缀缓存（APC，链式块哈希路线）**：将逻辑KV Cache切分为固定大小的物理块（如每块16个Token），每个块的密码学哈希值由"当前块Token序列 + 前置所有块哈希值"共同决定并存入全局哈希表。其局限在于 **块边界对齐刚性（Block-Hash Misalignment）**——未填满16个Token的尾部或跨块边界发生分叉的局部前缀均无法命中。
- **SGLang RadixAttention（统一基数树/压缩前缀树路线）**：源于斯坦福与伯克利LMSYS团队的SGLang放弃定长哈希表，采用 **基数树（Radix Tree）** 在树节点中按任意长度Token序列索引KV张量。新请求进入时以 O(k) 时间复杂度执行单次下探寻找最长前缀匹配（LPM），仅计算分叉点之后的增量后缀，在多轮对话、共享系统提示词的Agent及思维树（Tree-of-Thoughts）搜索中优势明显。**2026年9月发布的SGLang v0.5.20升级了统一基数树（Unified Radix Tree）架构**，支持直接缓存分支点状态，在 **DeepSeek-V4-Flash** 模型上将实测Token缓存命中率从43.8%提升至 **60.8%**，并原生支持Mamba 1/2状态空间模型。

2.2 生产级调度与显存压缩：PEEK双树遍历、多租户"缓存加盐"隔离与KIVI 2-bit量化

在生产集群中规模化落地引擎级前缀缓存，必须同步解决多租户安全隔离、批次调度驱逐与显存体积膨胀三大工程问题：

01 **多租户"缓存加盐（Cache Salting）"阻断计时侧信道攻击**：前缀缓存依赖底层Token ID序列完全一致，若业务层将动态时间戳置于系统提示词前端，会导致首Token立即分叉、命中率归零。更关键的安全隐患在于 **前缀缓存计时侧信道攻击（Prefix Cache Side-Channel）**——在多租户共享GPU集群中，攻击者可通过构造特定提示词并精确测量返回的TTFT（命中缓存时响应显著加快），反推其他客户近期是否输入过特定商业机密或私有代码。因此，生产级云平台必须强制启用 **缓存加盐（Cache Salting）** 机制，将租户唯一标识符（Tenant ID）注入哈希键或基数树根节点命名空间，从逻辑上阻断跨租户命中与信息泄露。

02 **PEEK协同双树遍历调度器消除批次拆散**：传统先来先服务（FCFS）调度器按到达顺序组批，常将共享同一前缀的请求拆散到不同批次，导致前序批次生成的KV块在后续请求到达前已被LRU策略驱逐。2026年提出的 **PEEK框架** 将待处理请求队列构建为挂起树（Pending Tree），与引擎基数缓存树执行单次同步下探（Dual-Walk Co-descent），配合分组优先调度与队列感知保护性驱逐，在4×H100集群上将SGLang与vLLM的缓存命中率提升 **3.0倍与2.6倍**，带动TTFT下降 **7.9倍与7.1倍**，端到端吞吐提升 **3.6倍与4.5倍**。

03 **KIVI 2-bit非对称量化与融合算子**：观测KV张量数值分布可见，Key缓存在少数固定通道（Channel）存在显著离群值，而Value缓存无明显通道离群但跨Token量化会污染全局输出。

KIVI算法 保留最近生成的残差窗口（如128个Token）为FP16，对历史主体缓存执行 **“Key按通道（Per-Channel）、Value按Token（Per-Token）”的2-bit非对称量化**，配合GPU融合反量化算子，在精度波动小于2%的前提下将峰值显存削减 **2.6倍**、批大小扩大 **4倍**、实际吞吐提升 **2.35至3.47倍**（需注意：若缺乏定制融合反量化内核，单纯开启量化反而会因解压开销拖慢解码速度）。此外，**GrowPage** 动态KV预算分配亦实现了较纯GPU基准高出 **2.09至6.13倍** 的每瓦特吞吐效率。

2.3 计算与状态物理解耦：月之暗面Mooncake P/D分离架构、LMCache非前缀复用与路由边界

当单机显存潜力被充分利用后，以 **Mooncake**（月之暗面Kimi生产推理平台，获USENIX FAST 2025最佳论文并于2026年2月加入PyTorch生态）、DistServe、Splitwise及 **LMCache** 为代表的系统，从分布式计算调度与集群资源池化维度重构了推理架构。

在传统耦合式推理中，计算密集型的Prefill与显存带宽密集型的Decode混跑在同一张GPU上，长提示词Prefill突发会严重阻塞Decode流式生成。**Mooncake** 在物理分离Prefill节点池与Decode节点池的基础上，构建了 **“以KV Cache为中心（KVCache-Centric）”** 的解耦体系：

01 **全局三级缓存池与Conductor智能调度**：Mooncake Store将集群内未被充分利用的CPU、主机DRAM、本地SSD与多张RDMA网卡聚合为全局三级缓存池（L1 GPU HBM、L2主机DRAM、L3本地/远端NVMe），采用LRU与软/硬钉住（Pinned）混合策略管理KV对象。全局调度器 **Conductor** 实时评估“跨网迁移已有KV块的传输耗时”与“直接重算耗时”的收益比，调度Prefill节点仅计算增量后缀，并在流量洪峰期启动 **早期拒绝（Early Rejection）** 策略防止算力沉没；同时，在P/D分离下仅对Decode通道执行分阶段功率封顶（Phase-Decoupled Power Control），即可在不损失吞吐的前提下节省 **32.3%** 的电力消耗。

02 **零拷贝RDMA传输引擎与少用43% GPU的生产实绩**：Mooncake的拓扑感知RDMA传输引擎在搬运40GB KV数据（Llama3-70B 128K上下文）时，于4×200Gbps和8×400Gbps RoCE网络下分别跑出 **87 GB/s与190 GB/s** 实测带宽（达TCP的2.4至4.6倍）。在Kimi生产环境中，Mooncake使A800/H800集群请求承载量提升 **75%–115%**（模拟长文本吞吐提升 **525%**），**相当于用少43%的GPU硬件承载同等业务流量**；在128块H200的Kimi-K2部署中实现224K tok/s预填充与288K tok/s解码吞吐，并将万亿参数模型权重热更新时间从53秒缩短至 **7.2秒**。

03 **LMCache位置无关分块复用与亲和性路由边界**：针对RAG多文档问答中召回片段顺序变化导致标准前缀缓存失效的痛点，**LMCache** 将文本按固定Chunk（如256 Token）切分并独立哈希，结合独立CUDA Stream实现“当前层注意力计算与下一层KV块I/O预取”流水线重叠，在RAG场景下缩短TTFT **4.5倍**、多轮问答吞吐提升 **15倍**。但工程落地需把握两条边界：一是短提示词（<500 Token）下跨网传输开销大于P/D分离收益；二是生产实证表明，盲目开启基于入口亲和性（Ingress Affinity）的缓存路由可能引发副本间残余负载倾斜（Residual Load Skew），热点排队延迟会抵消缓存命中收益，上线前需经影子流量回放（Shadow Replay）严格验证。

CHAPTER 03 · G3.5 FLASH CONTEXT MEMORY LAYER

03 存储金字塔的G3.5层革命——白盒闪存池如何以8万美元/PB重构企业存储?

3.1 传统存储层级的G3–G4断层与G3.5推理上下文内存层（ICMS）定位

只要KV缓存被限制在单机HBM（G1层）或单机主机DRAM（G2层）之内，高并发工作集一旦越过本地容量上限，驱逐就仍会退化为全量重算。审视传统AI数据中心的存储金字塔，在 **单机本地直连NVMe SSD（G3层）** 与 **外部企业级网络存储（G4层）** 之间，存在着明显的架构断层：

- **本机G3层（本地NVMe SSD）的“上下文孤岛”困境**：将KV Cache卸载至本机SSD虽能缓解单节点显存溢出，但数据被锁死在单台服务器内；一旦下一轮请求被负载均衡器路由至其他服务器，数据无法跨机共享，依然触发跨机重算。
- **传统G4层（外部网络存储）的“协议栈高延迟”困境**：若直接将KV状态写入传统通用分布式文件系统或对象存储，必须经过POSIX或HTTP/S3协议转换、操作系统内核网络栈中断、复杂纠删码、ACID事务锁及4KB细粒度随机I/O，数毫秒至数秒的拉取延迟会导致GPU计算核心长时间停顿。



■ 图03：AI推理五级内存与存储拓扑：G3.5共享上下文层填补容量与延迟鸿沟 — 位于本地NVMe（G3）与传统外部网络存储（G4）之间的G3.5共享上下文层，以PB级容量、约130微秒延迟与800GbE RDMA全Pod共享，打通跨机KV缓存复用。

为填补这一断层，结合NVIDIA在GTC确立的 **STX（Storage Technology Exchange）** 参考架构与CMX平台，产业界正式定义了 **G3.5层（在华为体系中称为L3.5层）——推理上下文内存存储层（Inference Context Memory Storage, ICMS）**。其核心设计准则在于：**不再将KV Cache视为需要永久归档的“存储文件”，而是将其作为跨计算Pod共享、高吞吐、微秒级响应的“临时内存状态（Ephemeral Memory State）”**。

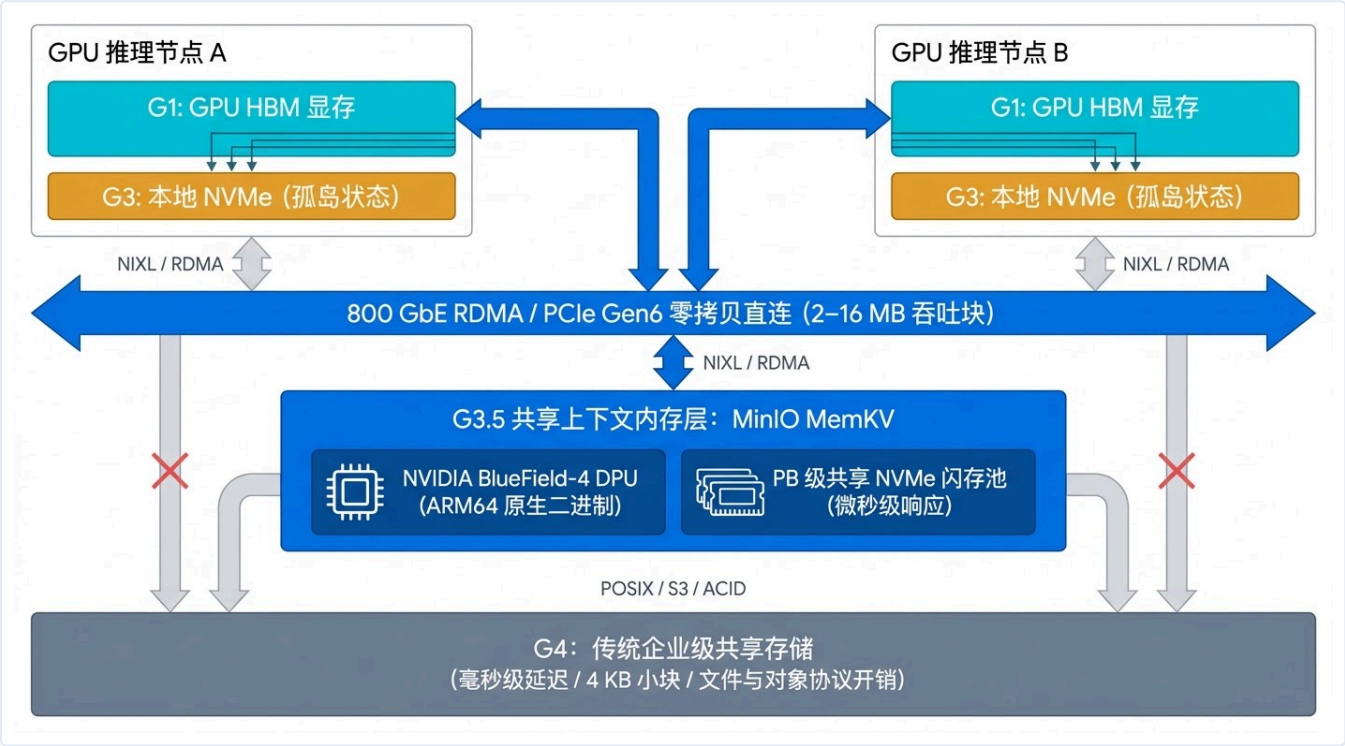
3.2 BlueField-4 DPU + MinIO MemKV零拷贝直连架构与64路并发“显存容量悬崖”实测

作为G3.5层的代表性工程实现，MinIO于2026年5月发布的 **MemKV** 配合 **NVIDIA Vera BlueField-4 STX DPU**，对传统存储实施了“去协议栈与去宿主机CPU化”改造（见图04）：

01 **DPU原生单二进制运行（消除宿主机CPU与OS内核开销）**：BlueField-4 DPU集成了Vera ARM64核心、128GB LPDDR5板载内存（为BlueField-3的4倍）与6倍计算性能。MemKV被编译为一个体积 **不足200 MB的ARM64原生静态单二进制文件**，直接内嵌运行在BlueField-4

DPU上，利用ARM SVE向量指令集加速纠删码与Highway Hash校验，省去了专用x86存储服务器机头与操作系统内核栈开销。

02 **2–16 MB大块条带化与800GbE RDMA零拷贝直传**：摒弃传统存储的4KB细粒度分页，MemKV采用专为GPU连续张量摄取调优的 **2 MB至16 MB大尺寸块（Extent-Based Blocks）**，跨多块裸NVMe盘执行无共享（Shared-Nothing）并行条带化I/O；网络层通过Spectrum-X 800GbE、PCIe Gen6、基于NVMe/RDMA的NIXL协议及GPUDirect Storage（GDS），将存储介质访问延迟从传统的15微秒压降至 **2微秒（<2 µs）以内**（端到端大块拉取约130 µs），减少GPU主机45%的CPU占用，实现 **200+ GB/s** 持续吞吐，并通过vLLM上游原生 `OffloadingConnector API` 无缝挂载。



■ 图04：G3.5推理上下文内存层（MinIO MemKV + BlueField-4 DPU）零拷贝直连架构图 — 计算Pod内的GPU通过vLLM OffloadingConnector与NIXL/GPUDirect RDMA（800GbE），直接跨网读写运行在BlueField-4 DPU上的MemKV单二进制引擎与白盒NVMe闪存池。

MinIO在搭载 **8块NVIDIA H200 GPU + 2台存储节点（48块裸NVMe盘，双路400GbE）** 的测试平台上，基于NVIDIA AIPerf智能体代码基准（1,000个会话、上下文达124K Token），对8B至405B的8款模型进行了198个稳态窗口实测，验证了关键的 **“显存容量悬崖（Capacity Cliff）”收益反转规律**：

- **跨越容量悬崖后TTFT提速75倍、吞吐提升196%**：在Llama 3.1 70B模型64K上下文测试中，无MemKV系统因频繁驱逐重算，TTFT长达 **53秒**；引入MemKV后TTFT降至 **703毫秒（提速约75倍）**。在128K高并发下，传统架构直接触发OOM中断，而MemKV支撑数百路并发平稳运行，Llama 3.1 70B（TP=4, 并发C=64）单卡有效吞吐提升 **+196%**（完成轮次达3.42倍），Llama

3.3 70B FP8（C=96）吞吐增长 **+187%**，Qwen3-32B（C=128）P90 TTFT从51.7秒缩至11.5秒（提升4.5倍）。

- **悬崖之下与悬崖之上的性能分界**：在低并发区间（如 C=1 至 C=32），当活跃KV工作集尚未塞满本地HBM或主机DRAM时，原生vLLM在 C=32 可达6,281 tok/s/GPU，而开启MemKV直写为5,162 tok/s/GPU（在Qwen3-0.6B等极小模型上，全量Prefill仅0.35秒，与0.2–0.4秒网络传输相当），此时外部传输带来轻微开销。但 **一旦并发越过 C=48 至 C=64 的“显存容量悬崖”**，原生vLLM因重算风暴骤降至4,092 tok/s/GPU，纯DRAM缓存在 C=64 耗尽内存后亦跌至5,017 tok/s/GPU；唯有MemKV在跨越悬崖后继续攀升，在 C=48 达到 **7,673 tok/s/GPU** 峰值，并在C=64 保持 **6,833 tok/s/GPU（高出原生基准67.0%）**。

3.3 企业级存储格局洗牌：不足8万美元/PB白盒闪存方案与VAST Data软件化转型

由于G3.5层的KV Cache属于“可重建的临时状态”（极端断电最坏仅退回Prefill重算，无需传统金融级双控冗余），G3.5白盒闪存池与传统企业级全闪存阵列（AFA）及商业并行文件系统的TCO拉开了 **10倍至50倍** 的差距（见表3-1）：

- **G3.5白盒裸闪存方案（MinIO MemKV + BlueField-4）**：利用标准高密度NVMe驱动器构建 **1 PB共享上下文内存池的硬件总成本不到80,000美元（<8万美元/PB，折合<$80/TB）**——不到两张旗舰GPU的采购价，却能为128卡集群每年挽回 **200万美元** 的重算浪费。
- **传统全闪存阵列与商业并行文件系统**：部署1 PB可用容量的Pure Storage FlashBlade//E（五年订阅模式）总合同成本接近 **100万美元/PB**；而采购1 PB规模的 **IBM Storage Scale（原GPFS）** 高性能一体机粗略报价更高达 **400万美元/PB**。

TABLE 04 · 第三章 · G3.5 上下文存储对比 Yavin质观 · 核心数据矩阵

AI 推理上下文存储（G3.5 层）四大技术路线架构与 1 PB 规模 TCO 对比

MinIO MemKV + BlueField-4 DPU 白盒方案以 <$8万/PB 的硬件成本，重构传统企业级存储定价体系

对比维度	MinIO MemKV + BF-4 白盒	传统企业级全闪存阵列	商业并行文件系统 (GPFS等)	华为 OceanStor M900 / 3FS
软件栈与元数据	**<200MB ARM64单二进制** 2–16MB大块 + 无共享条带化	重量级专有存储控制器OS 双控/多控强一致性事务锁	POSIX并行文件系统客户端 复杂分布式锁与元数据集群	**KV感知自适应记忆引擎** Retention生命周期冷热预测
数据路径与硬件	**内嵌 BlueField-4 DPU** 800G RDMA + NIXL零拷贝直传	专用存储机头与光纤/RoCE网 绑定高溢价原厂认证介质	需额外部署高配x86存储节点与专属网卡，内核栈开销高	**灵衢总线单跳直连** 消除协议转换与CPU中断转发

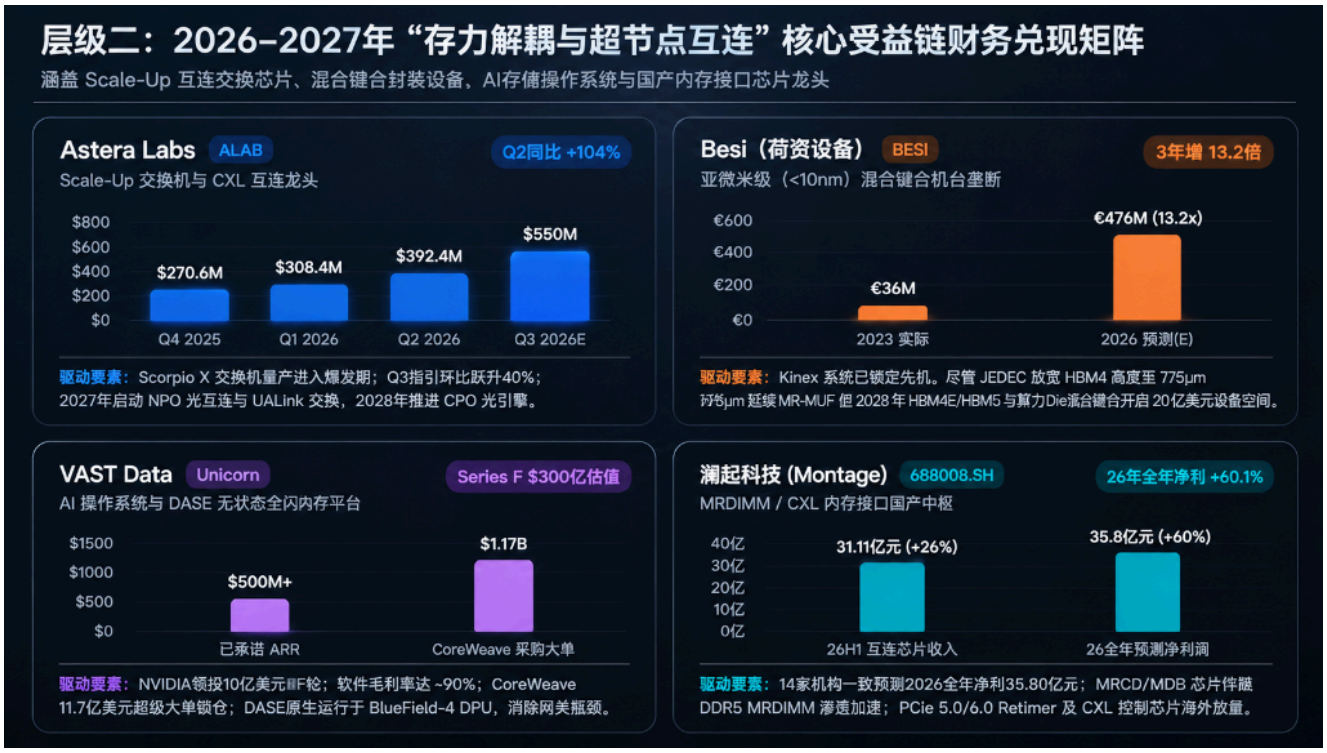
对比维度	MinIO MemKV + BF-4 白盒	传统企业级全闪存阵列	商业并行文件系统 (GPFS等)	华为 OceanStor M900 / 3FS
64K长文本延迟	**TTFT 703 毫秒 (提速75倍)** 存储访问 <2μs / 端到端 ~130μs	毫秒级协议开销叠加队列等待 高并发下易触发GPU空转	4KB细粒度I/O寻址抖动明显 难以满足百路并发流式拉取	**推理访问延迟低至60μs** 单套提供 **64 PB** 共享空间
1 PB 规模 TCO	**< 8 万美元 / PB (<$80/TB)** 每年挽回 200 万美元重算税	**约 100 万美元 / PB** (如 FlashBlade//E 五年订阅)	**约 400 万美元 / PB** (如 IBM Storage Scale 一体机)	**24 DWPD 寿命延长 16 倍** 承诺三年免换盘，折旧大幅下降
核心适用客群	超大规模云厂、独立AI推理集群、开源vLLM/SGLang生产机队	传统金融、政企核心数据库与强持久性合规存储负载	超算中心HPC科学计算与大规模预训练Checkpoint归档	昇腾超节点集群、国内大模型实验室及运营商智算中心

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

面对白盒方案的价格冲击，以 **VAST Data**、WEKA、Pure Storage等为代表的外部存储厂商迅速向“DPU原生AI操作系统”转型：

01 **VAST Data剥离硬件绑定，维持90%软件毛利率并获300亿美元估值**：VAST于2026年4月完成由NVIDIA领投的10亿美元F轮融资，估值升至 **300亿美元**；截至2026年6月累计软件预订额突破 **40亿美元**，承诺ARR超 **5亿美元**，**软件毛利率维持在约90%**，净收入留存率（NRR）超300%，并签下GPU云厂商 **CoreWeave 11.7亿美元** 商业合同。其核心策略是将无状态计算节点（CNode）直接移植进BlueField-4 DPU原生运行（卸载KV Cache同样实现20倍TTFT提速），推出“VAST Amplify”全局去重方案（实现6倍有效容量以对冲SSD涨价），并在DASE架构上整合对象、文件、向量数据库与支持MCP协议的 **AgentEngine**。

02 **BlueField-4 STX的DOCA安全机制与企业级壁垒**：NVIDIA在BlueField-4 STX中内置了三大硬件安全库——`DOCA Vault`（审查文件流量）、`DOCA Argus`（监控主机内存）与 `DOCA Flow`（硅片级拦截恶意Agent越权网络请求）。由于 `DOCA Vault` 底层依赖 `virtio-fs`，目前仅支持文件协议而不支持标准S3对象协议；VAST通过将PolicyEngine委托给BlueField-4并与CrowdStrike集成，在企业级多租户AI云市场建立了差异化安全壁垒。



■ 图05：“存力解耦与超节点互连”二阶受益链核心财务兑现矩阵 — 在存力与互连解耦浪潮中，Astera Labs（Q2营收+104%）、Besi（混合键合营收增12倍）、VAST Data（软件毛利率~90%、估值300亿美元）与澜起科技（全球份额36.8%、净利预增60%）率先兑现业绩增长。

04 CHAPTER 04 · TIER-1 HBM4 & BASE DIE FOUNDRY SHIFT
第一级内存（G1）与晶圆黑洞——HBM4逻辑底座代工与存储厂85.95%毛利率

4.1 HBM4 2048-bit宽总线演进、SPHBM4有机基板降本分支与HBM4E 4TB/s路线图

回到内存金字塔顶端的GPU片上显存（G1层），JEDEC于2025年4月发布的 **HBM4标准（JESD270-4）** 对3D堆叠DRAM实施了代际重构。在HBM3与HBM3E世代（8层24GB至12层36GB），行业主要通过拉高单引脚频率（从6.4 Gbps推至9.6 Gbps）在1024-bit接口下获取1.2–1.33 TB/s带宽，但高频引发的电压抬升与信号抖动已逼近物理极限。



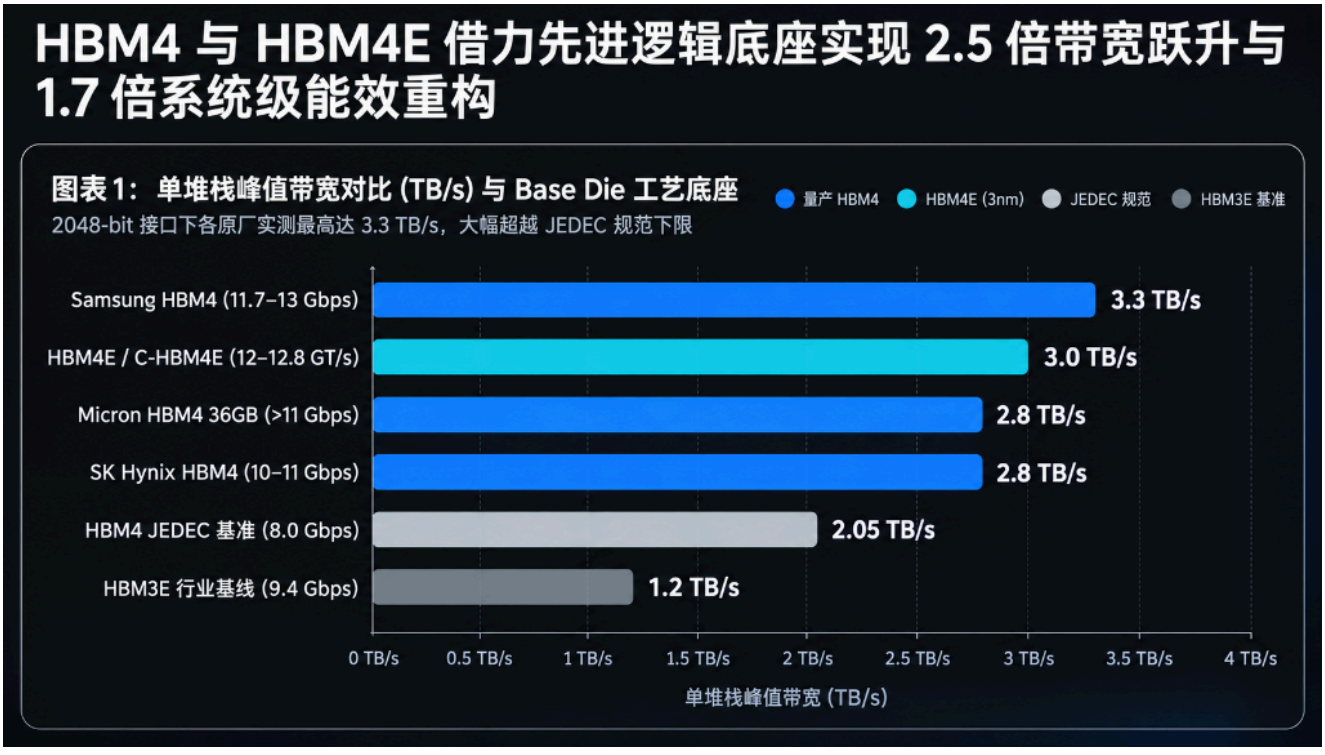
■ 图06：HBM1至HBM4E单堆栈峰值带宽跨代跃进（TB/s） — 从初代HBM1（0.13 TB/s）到HBM3E（1.2 TB/s），再到2048-bit总线的HBM4（2.0–3.3 TB/s）与HBM4E（4.0 TB/s），单堆栈带宽实现30倍跨越。

HBM4重回“以位宽换吞吐、以空间换能效”的设计路线（见表4-1）：

01 **2048-bit接口翻倍与64伪通道高并发**：HBM4将单堆栈物理接口位宽从沿用十年的1024-bit翻倍至 **2048-bit**，独立通道数从16个翻倍至 **32个（含64个伪通道）**，数据预取位宽从128-bit扩展至256-bit，并引入命令/数据总线分离与定向刷新管理（DRFM）以抵御行锤（Rowhammer）干扰。在高度并行的MoE路由与批次推理中，翻倍的独立访问路径显著降低了存储体冲突（Bank Conflicts）。

02 **三大原厂量产实测超越JEDEC基准**：JEDEC定义的8.0 Gbps基准速率对应2.048 TB/s单栈带宽，但在2026年商业化交付中三大原厂均大幅超标——**SK海力士**（2026Q2规模出货）速率突破10–11 Gbps，单栈带宽达 **2.5–2.8 TB/s**，能效提升超40%；**美光**（2026Q1量产供货Vera Rubin 36GB 12-Hi）实测速率超11 Gbps、单栈带宽 **>2.8 TB/s**（较HBM3E提升2.3倍），并送样48GB 16-Hi；**三星**（2026年2月量产商出）基于第六代10nm级（1c）DRAM晶圆实现 **11.7 Gbps稳定速率与13 Gbps峰值速率**，将单栈最大带宽推升至 **3.3 TB/s**。在搭载8颗36GB HBM4（共288GB）的NVIDIA Vera Rubin GPU上，单卡聚合显存带宽达 **19.2至22 TB/s**（单柜NVL72总带宽达1.4–1.6 PB/s）；而AMD MI430X/MI455X更搭载12颗36GB HBM4实现 **432GB** 单卡容量。

03 **SPHBM4有机基板分支与HBM4E 4TB/s演进**：针对CoWoS硅中介层成本高昂的痛点，JEDEC于2026年7月发布 **SPHBM4标准（JESD330-4）**，保留HBM4核心DRAM堆栈，但通过专用Base Die执行4:1高频串行化转换，将外部接口收敛至 **512-bit**，使2.0 TB/s带宽首次能够直接部署在常规有机封装基板上。而在旗舰侧，2026年5–6月送样、2027年主力出货的 **HBM4E** 将单引脚速率推升至 **14–16 Gbps**，单栈带宽达 **3.6 TB/s（三星）至4.0 TB/s（SK海力士）**，单栈容量升至12层48GB与16层64GB（随搭载1,024 GB HBM4E的Rubin Ultra规模部署）。



■ 图07：HBM4与HBM4E单堆栈峰值带宽对比（按存储原厂及逻辑底座代工划分） — 三大存储原厂HBM4实测带宽（2.8–3.3 TB/s）均超越JEDEC 2.0 TB/s基准，向台积电3nm/三星4nm逻辑底座演进的HBM4E进一步升至3.6–4.0 TB/s。

TABLE 05 · 第四章 · HBM4/4E 微架构演进 Yavin质观 · 核心数据矩阵

HBM3E、HBM4、SPHBM4 与 HBM4E 跨代核心技术规范与微架构演进矩阵

物理接口翻倍至 2048-bit，逻辑底座（Base Die）全面转向台积电 3nm 与三星 4nm 先进逻辑工艺

技术指标	HBM3E (8H / 12H)	HBM4 (12H / 16H)	SPHBM4 (JESD330-4)	HBM4E / 定制 cHBM
接口位宽与通道	**1024-bit** 16 通道 (32 伪通道)	**2048-bit (翻倍)** 32 通道 (64 伪通道)	**512-bit (4:1串行化)** 兼容常规有机封装基板	**2048-bit + 自定义 C2C** 支持近存计算逻辑下沉
单引脚速率与带宽	8.0–9.6 Gbps **1.0–1.23 TB/s**	基准 8.0 / **实测 10–13 Gbps** **2.05–3.30 TB/s**	高频串行化引脚传输 **单栈峰值 2.0 TB/s**	**14.0–16.0 Gbps** **单栈 3.6–4.0 TB/s**
单栈容量与Bank数	24 GB / **36 GB** 64 Banks (128-bit 预取)	**36 GB / 48 GB** **256 Banks (256-bit预取)**	**36 GB / 48 GB** 复用 HBM4 核心 DRAM 颗粒	**48 GB (12H) / 64 GB (16H)** Rubin Ultra 单卡达 **1,024 GB**
Base Die 制造工艺	存储原厂自产 **DRAM 周边工艺 (1.1V)**	**Foundry 先进逻辑工艺** TSMC N12 / **N3P** / 三星 4nm	专用高速串行化逻辑底座 免除昂贵 CoWoS 硅中介层	**英伟达强制要求台积电 3nm** VDDQ 降至 0.68V，底座能效翻倍
主力搭载算力平台	NVIDIA H200 / B200 / GB300 AMD MI325X / MI355X	**NVIDIA Vera Rubin (288GB)** **AMD MI455X (432GB)**	中高端推理卡与定制 ASIC 缓解 CoWoS 封装产能瓶颈	**NVIDIA Rubin Ultra (2027)** 新一代云厂自研旗舰 XPU

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

4.2 逻辑底座（Base Die）晶圆化：英伟达“台积电3nm底座令”与三星4nm半数产能锁定

HBM4最具结构性影响的变化，是位于堆栈最底层的 **逻辑基础裸片（Logic Base Die）** 脱离存储原厂自有DRAM产线，全面转向晶圆代工厂的先进逻辑工艺（12nm/5nm/4nm/3nm）。

在HBM3E及更早代际中，Base Die由存储原厂采用DRAM周边工艺制造，仅负责被动信号转发。但在HBM4下，翻倍的2048根I/O通道使宿主机侧HBM4 PHY面积从HBM3E的11 mm²增至 **15 mm²（增加36.4%）**，单栈功耗升至120–160W，高电容、低晶体管密度的DRAM工艺已无法承载。据台积电与创意电子（GUC，已于TSMC N3P 3nm节点流片HBM4 PHY IP并完成硅验证）披露的数据，将Base Die迁移至 **3nm先进逻辑工艺** 后，工作电压 VDDQ 从HBM3E的1.1V降至 **0.679V–0.963V**，使Base Die自身能效提升 **2倍**；演进至12.8 GT/s的HBM4E/cHBM4E相比HBM3E更实现 **2.5倍带宽跃升、1.7倍系统级能效提升与1.8倍面积效率改善**。同时，3nm逻辑底座允许云厂与AI芯片商集成上亿个逻辑门，将内建自测试（BIST）、ECC纠错、私有串行接口及轻量级近存计算算子下沉至显存底部（如NVIDIA推进的定制NVHBM较通用HBM4E提速30%并降耗15%）。

TABLE 06 · 第四章 · 存储三巨头博弈 Yavin质观 · 核心数据矩阵

三大存储原厂 HBM4 与 HBM4E 逻辑底座代工策略、封装路线与财务表现对比

英伟达下达 Rubin Ultra"台积电 3nm 底座令"，3–4 倍晶圆折损推升美光 FQ1 27 毛利率指引至 85.95%

对比维度	SK 海力士 (SK hynix)	三星电子 (Samsung)	美光科技 (Micron)
DRAM 核心晶圆	成熟高良率 **1b nm (第5代 10nm)** HBM4E 全面导入 **1c nm**	激进跨代采用 **1c nm (第6代 10nm)** 实测稳定 11.7 Gbps / 峰值 13 Gbps	主力采用 **1β (1-beta) / 1γ nm** 收购力积电 P5 厂扩充 HBM4 产能
HBM4 Base Die	**台积电 12nm** 标准版 + **台积电 3nm (N3P)** 旗舰版双轨	标准版采用自家 **4nm 代工** (内部需求锁死三星半数4nm产能)	初期自研 CMOS 底座，高端定制版协同转向 **台积电 3nm**
Rubin Ultra 定制版	**遵循英伟达指令采用台积电 3nm** 单栈 HBM4E 冲刺 **4.0 TB/s**	**罕见妥协：改用台积电 3nm 底座** HBM4E 已于 10 月初通过英伟达验证	**统一采用台积电 3nm 逻辑底座** 深度绑定北美大厂定制化接口
775μm 键合路线	**Advanced MR-MUF** (良率 ~70%) 受益 775μm 放宽延续至 16-Hi	12H 采用 TC-NCF；16H HBM4E 率先推进 **无凸块混合键合**	**TCB 热压键合** 主力量产 利好韩美半导体与 ASMPT 设备链
份额与财务实绩	2026 HBM 营收预计 **546 亿美元 (+58%)** 预警供应短缺持续至 2028–2030 年	HBM 营收份额从 2025Q2 **15% 反弹至 2026Q2 的 38%**；2027 产能 80% 锁空	关停消费品牌 Crucial；**FQ1 27 单季营收指引 615 亿美元、GAAP 毛利率 85.95%**

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

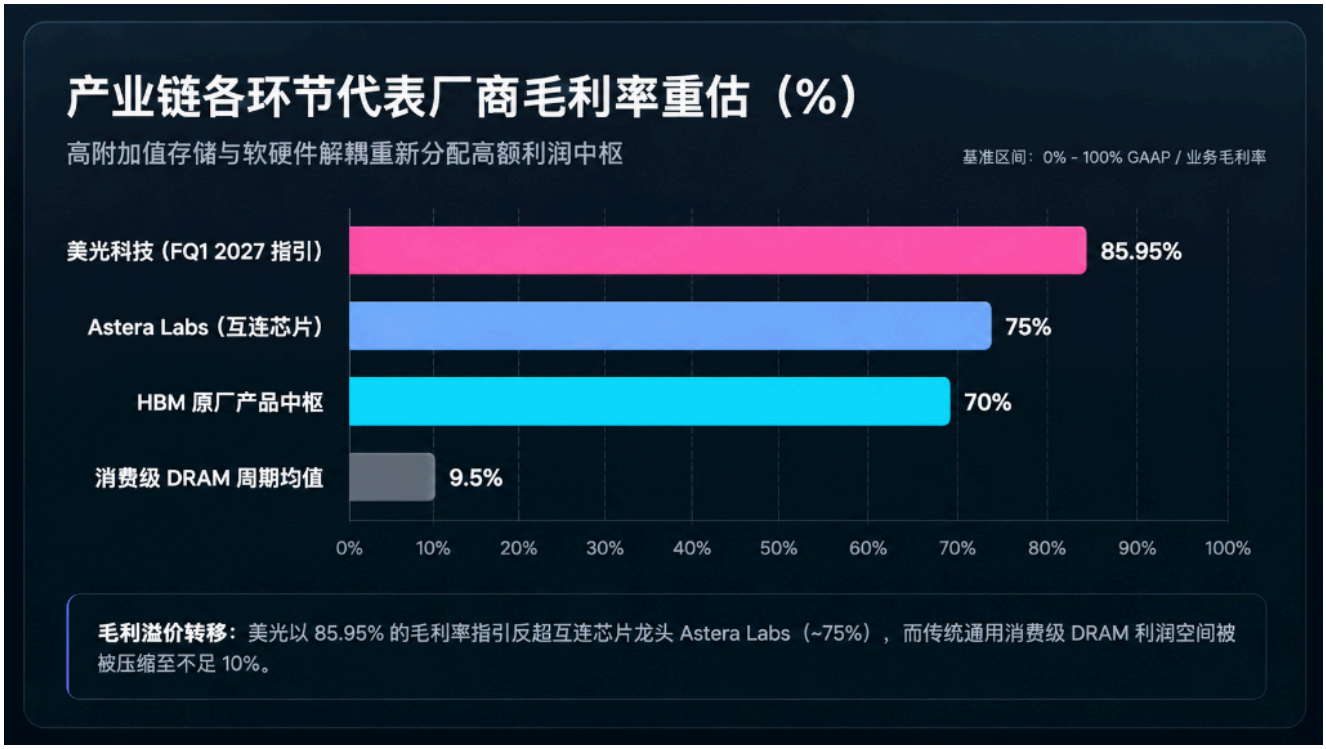
围绕逻辑底座代工权，2026年10月供应链出现了两项标志性转折：

01 **英伟达下达"台积电3nm底座统一令"，三星罕见妥协**：据2026年10月供应链消息，**NVIDIA已向SK海力士、三星、美光发出统一要求，规定三家原厂在制造供应Rubin Ultra等下一代平台的定制化HBM（NVHBM）时，必须全部采用台积电3nm工艺生产Base Die**。长期坚持IDM垂直整合的三星在评估后做出让步，同意针对NVIDIA定制HBM采用台积电3nm底座以确保供货份额（三星HBM4E已于10月初通过Rubin Ultra验证）。

02 **三星"双轨策略"锁死自家50%的4nm代工产能**：为支撑自身代工业务并服务AMD及北美云厂ASIC（如斩获OpenAI/博通Jalapeño芯片HBM4独家订单），三星在标准版HBM4上采用自家4nm底座。到2026年下半年，**三星内存部门的内部Base Die需求直接占去了Samsung Foundry 50%的4nm总产能**。外部Fabless客户被迫与三星内存部门争抢晶圆，叠加台积电5nm以下节点涨价5%–10%，三星顺势将外部4nm代工报价上调 **15%**。借此双轨策略，三星在全球HBM市场的营收份额从2025Q2的 **15%** 回升至2026Q2的 **38%**。

4.3 3至4倍晶圆面积折损与利润倒挂：美光85.95%毛利率指引与775微米键合设备分工

随着HBM4与先进封装占据加速卡逾六成BOM成本，全球半导体行业的毛利率格局在2026–2027年出现了历史性倒挂：**重资产IDM存储原厂的毛利率首次反超了顶层轻资产Fabless逻辑芯片设计公司。**



■ 图08：AI硬件产业链代表厂商毛利率历史性倒挂（存储原厂反超逻辑芯片设计厂） — 受HBM与代工成本上升影响，逻辑与互连芯片厂毛利率收敛至70%–75%区间，而美光FQ1 2027单季GAAP毛利率指引升至85.95%，改写半导体行业四十年利润分配格局。

美光科技（Micron）于2026年9–10月披露的SEC文件给出了确切财务印证：在截至2026年9月3日的2026财年10-K年报中，美光实现总营收 **1,331.88亿美元**（其中DRAM营收1,006.8亿美元，占75.6%；NAND营收317.9亿美元，占23.9%）；而在2026年9月30日提交的8-K公告中，**美光将2027财年第一季度（FQ1 2027）单季营收指引上调至615亿美元，单季GAAP毛利率指引达到85.95%**（单季稀释EPS指引达37.84美元）。同期，互连芯片龙头Astera Labs非GAAP毛利率为74%–75%，NVIDIA等逻辑芯片厂硬件毛利率亦从85%–88%高点向70%–75%区间收敛。

驱动存储原厂毛利率突破85%的物理根源，在于HBM的 **“3至4倍晶圆面积折损（Wafer Penalty）”** 及随之引发的全行业产能挤出效应：

01 **256个Bank与两万根TSV造成的“三换一”产能折损**：在Hot Chips 2026大会上，美光披露一颗标准DDR5裸片仅有32个Bank，而一颗HBM4裸片内部集成了 **256个Bank** 与超过 **20,000根深硅刻蚀硅通孔（TSV）**。庞大的TSV阵列与电源走线使单颗HBM裸片面积显著大于同容量DDR5，再叠加12–16层堆叠良率损耗（成熟MR-MUF良率约70%），**生产1 GB容量**

HBM3E/HBM4消耗的晶圆面积是标准服务器DDR5的3倍以上，是手机LPDDR的4倍。TrendForce统计显示，2026年HBM消耗三大原厂 **22%–23%** 的DRAM晶圆投片量，却仅产出全行业 **9%** 的比特容量；**到2027年，HBM将消耗30%的DRAM晶圆投片量，而比特产出仅占13%。**

02 **2027年产能提前锁空与全品类存储涨价**：每将一片晶圆转产HBM，市场上即减少三片晶圆的常规内存供给。2026年10月19日，北美科技巨头通过长期协议（LTA）提前锁定了三星2027年全年 **80%的DRAM、NAND与HBM总产能**；SK海力士预计2026年仅HBM营收即达 **546亿美元（+58%）** 并预警短缺将持续至2028–2030年；美光则在2025年12月关停消费品牌 **Crucial（英睿达）**，并于2026年1月斥资 **18亿美元收购台湾力积电（PSMC）P5晶圆厂** 扩建1γ/1δ HBM4产线（2027下半年投产）。产能挤出引发全线涨价：2027年全球HBM均价（ASP）预计再涨 **121%**；常规DRAM合约价在2026Q1环比上涨 **90%–95%**、Q2再涨 **58%–63%**、Q3续涨13%–18%（32GB DDR5 RDIMM现货升至2,100美元，1GB DDR4裸片亦从12美元升至46.11美元）；企业级NAND端受G3.5卸载拉动，**30TB级企业级NVMe SSD采购价上涨约257%**。

03 **775微米高度放宽：TC键合延续与混合键合顺延至2028**：在后道封装设备端，JEDEC于2026年1月将HBM4封装总高度上限从720微米放宽至 **775微米**（配合台积电5微米减半微凸块）。由于无凸块铜对铜混合键合（Hybrid Bonding）初期良率仅约10%（商业化盈亏线需60%），775微米的宽限使SK海力士与美光得以在12-Hi及16-Hi HBM4上继续沿用成熟的 **MR-MUF与热压键合（TCB）**，直接利好TCB设备龙头 **韩美半导体（Hanmi）** 与 **ASMPT（占全球TCB份额26%）**。混合键合的大规模商业化则顺延至三星16-Hi HBM4E及2027–2028年后的HBM5/zHBM世代（2028年混合键合设备TAM将突破 **20亿美元**，龙头 **Besi** 的混合键合营收已从2023年3,600万欧元增长12倍至2026年的 **4.76亿欧元**）。

05 CHAPTER 05 · TIER-2 SOCAMM2 & TIER-3 CXL 4.0 POOLING

第二、三级内存（G2/G3）重构与隐形冠军——SOCAMM2、MRDIMM与CXL 4.0

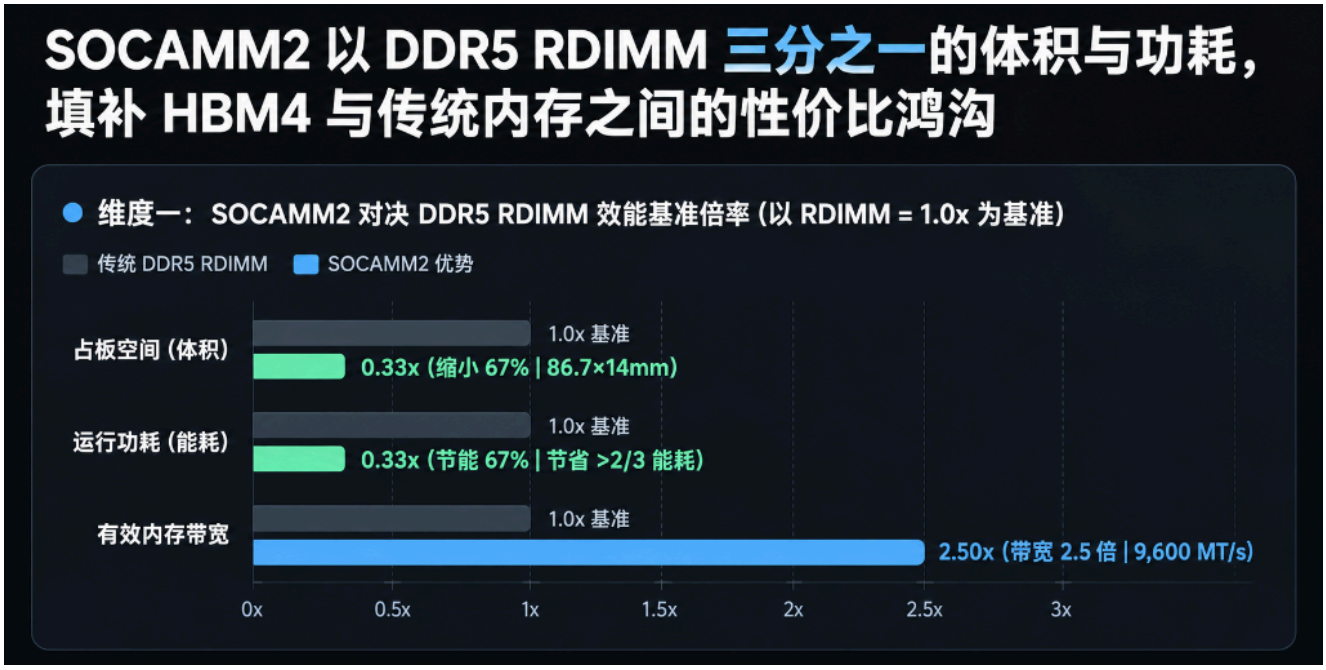
5.1 第二级内存模块化：JEDEC SOCAMM2可插拔标准、优群科技连接器与澜起MRDIMM套片倍增

当GPU侧的HBM4专注于服务极热的活跃权重与当前解码步时，承担温KV Cache卸载、Agent强化学习（RL）沙盒环境及数据预处理的宿主机CPU内存层（第二级内存），正同时在ARM超节点与x86通用服务器两端完成架构升级：

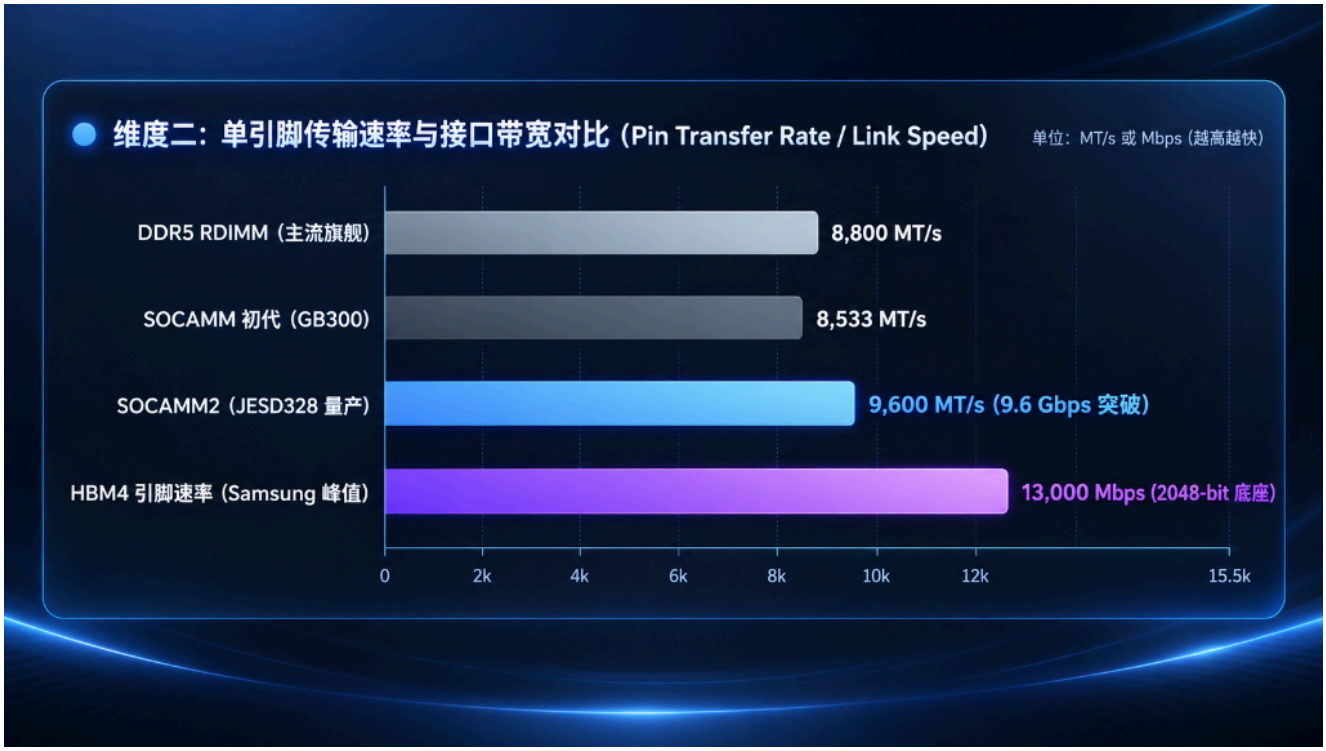
01 **ARM超节点侧：JEDEC JESD328（SOCAMM2）终结板载焊死，拉动优群科技48.23%毛利率。**在NVIDIA Grace Hopper与GB200世代中，为获取高于DDR5的能效，移动端LPDDR5X被直接表面贴装焊接在计算母板上——一旦单颗DRAM颗粒失效，云厂无法现场更换，必须将整块价值数万美元的CPU+GPU计算托盘返厂甚至报废。为解决可维护性与液冷空间矛盾，美光与NVIDIA在GB300中引入初代SOCAMM，随后联合三星、SK海力士移交标准组织，于 **2026年6月1日正式发布JEDEC JESD328标准（SOCAMM2 Version 1.00）**。SOCAMM2采用 **694引脚** 高密度平面弹簧压缩连接器（尺寸86.7×14×1.0 mm，支持25次拆装循环），水平平铺锁固于主板，契合全液冷冷板覆盖；单模组集成4个16-Hi堆叠（共64颗IC），提供128-bit双通道、1.05V低压与 **9.6 Gbps（9600 MT/s）** 速率（超越DDR5 RDIMM的8800 MT/s），192GB与256GB版本已由美光、海力士、三星于2026年量产。相比同容量DDR5 RDIMM，**SOCAMM2体积与功耗均仅为1/3，有效带宽达2至2.5倍，而单位GB成本仅约为HBM4的1/3（约5–8美元/GB）**（见图09、图10）。在Vera Rubin平台中，88核 **Vera CPU** 借助SOCAMM2支持 **1.5–2 TB本地内存（整柜54 TB）与1.2 TB/s带宽**，通过 **1.8 TB/s一致性NVLink-C2C** 与Rubin GPU直连回填KV Cache（TTFT缩短80%以上），并在256-CPU纯Vera机架中承载 **22,500个并发Agent强化学习沙盒**；**AMD亦确认将于2027年发布的下一代EPYC“Verano”处理器中原生支持SOCAMM2**。伴随SOCAMM2放量，占全球笔电SO-DIMM与M.2连接器 **50.5%份额** 的 **优群科技（Argosy, 3217.TWO）** 切入AI服务器供应链（联想高端机种月出货破5万颗），推动2026上半年综合毛利率升至 **48.23%**（净利率22.6%），并提前卡位2027年DDR6连接器。

02 **x86服务器侧：DDR5 MRDIMM“1+10”套片翻倍，推动全球第一澜起科技净利预增60%。**在x86通用AI服务器（如Intel 18A-P Diamond Rapids与AMD新一代EPYC）中，随着单机平均DRAM奔向14.75 TB，第二子代 **DDR5 MRDIMM（多路复用秩内存模组）** 全面普及。它通过同时访问两个Rank并多路复用到单通道，将传输速率翻倍至 **12,800 MT/s**。更重要的是其 **“1+10”套片价值量倍增机制**：传统RDIMM每条仅需“1颗RCD + 4/8颗DB（甚至无DB）”，而MRDIMM升级为 **“1颗MRCD（多路复用寄存时钟驱动器）+ 10颗MDB（多路复用数据缓冲器）”**，单条内存接口芯片颗数翻倍且均价抬升。全球内存接口芯片前三强占据超93%份额，其中中国龙头 **澜起科技（688008.SH）以36.8%的市占率超越瑞萨与Rambus位居全球第一**

——其2026Q1归母净利润增长 **61.3%**，上半年互连芯片营收达 **31.11亿元（+26.4%）**，机构一致预测其2026全年净利润达 **35.80亿元（同比增长60.14%）**。



■ 图09：SOCAMM2对比传统DDR5 RDIMM核心效能优势（相对百分比） — JEDEC JESD328 SOCAMM2可插拔压缩模组在物理占板体积与运行功耗上仅为传统DDR5 RDIMM的33%（1/3），而有效聚合带宽提升至250%（2.5倍）。



■ 图10：主流服务器内存模组单引脚传输速率演进（Gbps / MT/s） — 从DDR5 RDIMM（8.8 Gbps）到SOCAMM2 LPDDR5X（9.6 Gbps）、第二子代MRDIMM（12.8 Gbps）、LPDDR6（14.4 Gbps）与DDR6（17.6 Gbps），主机内存接口速率持续倍增。

5.2 第三级内存池化：CXL 4.0（128 GT/s）亚200ns架构、绑定端口设计与Astera Labs放量

为消除单节点内存孤岛（Stranded Memory）并在机架尺度构建突破 **100 TB** 的动态共享内存池，CXL联盟发布了 **CXL 4.0规范**，在物理层与端口微架构上实现了两项关键跨越：

01 **PCIe 7.0（128 GT/s）物理层与亚200纳秒零延迟加密**：CXL 4.0将单通道速率从CXL 3.x的64 GT/s翻倍至 **128 GT/s**（PAM4调制，单x16链路双向聚合带宽达 **1.536 TB/s**）。借助256B延迟优化型Flit，Panmnesia实测链路控制器往返延迟 **<100 ns**，Synopsys（2026年8月首发完整CXL 4.0 IP）实测端到端 `CXL.mem` 加载到使用（Load-to-Use）延迟控制在 **200纳秒以内（<200 ns，比SSD快100倍）**，且IDE硬件AES-GCM实时加密在Skid模式下实现 **零周期延迟开销**。同时，规范新增对单链路 **至多4个Retimers** 与原生x2宽度的支持，打通了跨机架光电扩展路径。

02 **绑定端口（Bundled Ports, BPD）与精简端口（Streamlined Ports）微架构**：CXL 4.0首次允许将主机与加速器之间的多条物理链路聚合为单一逻辑管道（BLD）。为降低硅面积与功耗，每个绑定组仅需保留 **1个标准全功能端口**（负责68B Flit兼容、全局复位、IDE加密并共享ATS地址转换结果），其余扩展通道全部配置为剥离控制面的 **精简端口**（仅支持256B Flit与UIO数据流）。运行时通过流量喷洒（Traffic Spraying，以独立256B Flit为单位分担负载），**将4条x16 CXL 4.0链路绑定即可提供>2 TB/s带宽，8条x16绑定更突破4 TB/s**，媲美乃至超越单颗HBM4堆栈。配合主机发起的封装后修复（PPR）以及Marvell等推动利用回收DDR4/降频DDR5构建CXL池，**第三级内存的单位容量成本被压降至HBM的1/10**（见图11与表5-1）。



■ 图11：AI三层内存架构单位容量成本指数对比（以HBM4 = 100为基准） — 按数据温度分层：第一级HBM4成本指数为100（~$17+/GB），第二级SOCAMM2降至33（约1/3），第三级CXL 4.0共享内存池降至10（约1/10），形成阶梯式成本结构。

TABLE 07 · 第五章 · 三级协同内存架构 Yavin质观 · 核心数据矩阵

AI 数据中心三级协同内存架构（HBM4 + SOCAMM2 + CXL 4.0）全方位工程对标

按数据温度分层：1/3 成本的 SOCAMM2 承接温数据，1/10 成本的 CXL 4.0 近存计算实现 39.7 倍硬件降本

对比维度	第一级：HBM4 / HBM4E (极热层)	第二级：SOCAMM2 / MRDIMM (温层)	第三级：CXL 4.0 / PNM (百TB池化层)
物理形态与标准	CoWoS 封装内 **3D TSV 堆叠** JEDEC JESD270-4 (2048-bit)	**86.7×14×1.0mm 可插拔压缩模组** JEDEC **JESD328** (694引脚 LP5X)	PCIe 7.0 物理层 (**128 GT/s PAM4**) 绑定端口 (BPD) + 精简端口架构
容量与带宽量级	单卡 **288 GB–1,024 GB** 单卡带宽 **19.2–32 TB/s**	单 Vera CPU **1.5–2 TB (整柜54 TB)** 速率 **9.6 Gbps** (MRDIMM 达 12.8 Gbps)	单机架池化 **100 TB+** 4××16 绑定 **>2 TB/s**，延迟 **<200 ns**
体积、功耗与成本	成本指数 **100** (~$17+/GB) 单栈功耗升至 120–160W	成本指数 **33 (约HBM的1/3)** 较RDIMM：**1/3体积、1/3功耗、2.5倍带宽**	成本指数 **10 (约HBM的1/10)** 支持回收 DDR4/降频 DDR5 混插建池
核心承载负载	当前激活的模型权重、极热注意力窗口与当前步解码状态	温 KV Cache 卸载 (TTFT 缩短 80%)、**22,500 个并发 RL 代码沙盒**	跨节点无网跳数 KV 共享池、**CXL-PNM 内存端原位注意力计算**

对比维度	第一级：HBM4 / HBM4E (极热层)	第二级：SOCAMM2 / MRDIMM (温层)	第三级：CXL 4.0 / PNM (百TB池化层)
代表产品与降本	海力士/三星/美光 HBM4 长鑫存储 12H HBM3E (2027量产)	NVIDIA **Vera Rubin**、AMD **EPYC Verano** **优群科技**连接器、**澜起** 1+10 套片	**Astera Labs** Leo CXL、**UVA KV Server** 跑满血 R1 **每百万Token成本降 39.7 倍**

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

在这条高速互连赛道上，**Astera Labs（NASDAQ: ALAB）** 兑现了高速业绩增长：2026Q1营收增长93%至3.084亿美元，**Q2 2026营收同比增长104%至3.924亿美元**，Q3单季营收指引达**5.4亿–5.6亿美元（环比增长40%）**，非GAAP毛利率稳定在 **74%–75%**，面向未来五年 **250亿美元TAM**。其 **Leo CXL智能内存控制器** 已斩获北美云厂专为“AI推理KV Cache卸载”定制的Design Win（2027年向两家北美云厂批量出货）；其 **Scorpio X交换芯片** 已于2026Q3超越PCIe Retimer成为第一大收入来源，配合收购aiXscale Photonics布局NPO/CPO光学引擎，将单颗XPU配套互连价值量推升至数千美元。

5.3 从“注意力在GPU”到“注意力在内存”：CXL近存计算（PNM）跑DeepSeek-R1降本39.7倍

在机架级分布式推理中，即便使用RDMA网络，跨机复用KV Cache仍需经历“源GPU → 源DRAM → 源网卡 → 交换机 → 目的网卡与DRAM → 目的GPU”的四跳序列化搬运。CXL 3.0/4.0字节寻址内存池消除了网络栈跳数，并在2025–2026年演进出三类代表性系统：

01 **无网跳数CXL共享池（TraCT / Photonic-CXL）**：弗吉尼亚理工大学 **TraCT** 系统通过软件两级锁在非一致性CXL裸内存上实现vLLM无网KV共享，较RDMA将平均TTFT降低 **2.6倍**、P99延迟降低 **3.0倍**、有效吞吐提升 **1.9倍**；引入硅光互连的 **Photonic-CXL** 进一步削减50%以上延迟，将TTFT提升 **6.6倍**。

02 **消除GC开销的语义感知CXL-SSD（HyMCache / LM-CXD）**：针对多轮对话KV块“只读不改且可重新计算生成”的特性，**LM-CXD** 由LMCache直接下发按层预取计划，并在NAND闪存块擦除时直接丢弃残留活跃块，**省去了SSD内部消耗大量带宽的垃圾回收（GC）机制**，将平均TTFT较标准CXL-SSD降低 **4.03倍**（逼近纯DRAM的1.5倍以内）。

03 **处理器亲和性反转与CXL近存计算（CXL-PNM / KV Cache Server）**：定量建模表明，若将KV卸载至内存而注意力计算留在GPU执行（Attention-on-GPU），每层均需跨总线搬运庞大的Key与Value张量，传输量达 O(S·d)（随序列长度 S 线性增长）；反之，**若将注意力计算下沉至CXL内存侧执行（Attention-on-Near-Memory）**，在内存端直接完成注意力加权求和，仅将固定维度的输出向量传回GPU执行FFN，则每层传输量恒定为 O(d)，与序列长度 S 脱钩。基于这一原理，集成轻量级近存计算单元的 **CXL-PNM** 在百万Token推理中实现了 **21.9倍吞吐提升与60倍单Token能耗下降**；弗吉尼亚大学（UVA）提出的 **Memory-Centric KV Cache Server** 在

运行满血版 **DeepSeek-R1-671B** 稳态推理时，将每小时硬件成本从纯GPU集群的59.23美元降至 **3.53美元（降低16.8倍）**，机柜功耗压至18.6 kW，**每百万Token生成成本从24.24美元直接降至0.61美元，降幅达39.7倍。**

TABLE 08 · 第五章 · 存力互连隐形冠军 Yavin质观 · 核心数据矩阵

存力解耦与超节点互连细分硬件赛道“隐形冠军”份额、ASP 增量与财务表现矩阵

在存储原厂之外，掌握内存接口、高频连接器、CXL/Scale-Up 交换与液冷组件的龙头兑现业绩高增

隐形冠军企业	核心赛道与全球市占率	2026 核心财务与业务兑现	单机柜 ASP 弹性与技术壁垒
澜起科技 (688008.SH)	**全球内存接口芯片第一** 全球份额 **36.8%** (超越瑞萨与Rambus)	• **2026Q1 归母净利大增 61.3%** • 上半年互连芯片营收 **31.11 亿元 (+26.4%)** • 2026 全年净利预计 **35.80 亿元 (+60.1%)**	x86 转向 **MRDIMM (12,800 MT/s)** 使单条芯片从 1 颗升级为 **“1 MRCD + 10 MDB”**，价值量成倍扩张
Astera Labs (NASDAQ: ALAB)	**CXL 控制器与 Scale-Up 交换** 北美云厂核心互连底座	• **2026Q2 营收同比暴增 104% 达 3.924 亿美元** • Q3 单季营收指引 **5.4–5.6 亿美元 (+40% QoQ)** • 非 GAAP 毛利率稳居 **74%–75%**	**Leo CXL** 斩获北美云厂 KV 卸载大单；**Scorpio X** 超越 Retimer 成第一支柱，单 XPU 配套冲数千美元
优群科技 (Argosy) (3217.TWO)	**SOCAMM2 压缩连接器龙头** 占全球 SO-DIMM/M.2 **50.5%**	• **2026H1 综合毛利率高达 48.23%** • 营业利益率 **28.1%**、净利率 **22.6%** • 高端 AI 服务器连接器月出货破 5 万颗	攻克 1.0mm 超薄空间内 **694 引脚 9.6 Gbps** 平面弹簧压缩工艺，并提前卡位 2027 年 DDR6 连接器
Besi & 英维克 (BESI.AS / 002837.SZ)	**混合键合设备霸主 (Besi)** **液冷三大白名单 (英维克)**	• Besi 混合键合营收飙升 **12倍 达 4.76 亿欧元** • **英维克** 打入英伟达 CDU/快接头/Manifold 白名单并切入交换机冷板	2028 混合键合设备 TAM 破 20 亿美元；单芯片独立冷板推升单柜液冷 BOM 至 **70万–100万元**

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

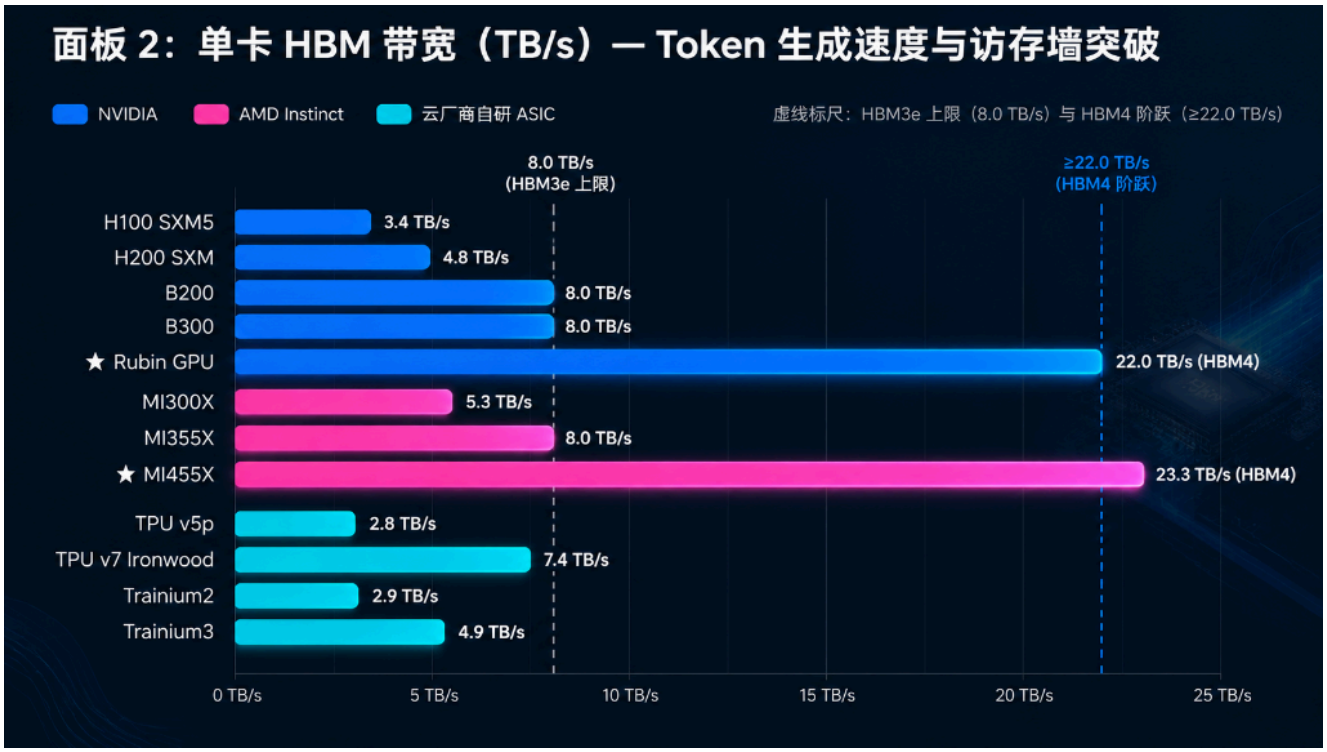
06 CHAPTER 06 · SCALE-UP FABRIC & LIQUID COOLING
Scale-Up超节点互连与物理基建卡点——NVLink、UALink、ESUN与155kW液冷

6.1 AI后端交换机反超前端网络：NVLink 5/6、NVLink Fusion生态收编与Kyber 78层背板延期

当单卡HBM容量（见图12）与单卡显存带宽（见图13）无法独立装下万亿参数MoE模型与百万Token并发时，数据中心网络被划分为机柜内内存语义域 **Scale-Up（纵向扩展）**、跨机柜RDMA网络 **Scale-Out（横向扩展）** 与跨园区 **Scale-Across**。Dell'Oro Group于2026年9月发布的数据显示：**2026年第二季度（2Q 2026），全球AI后端网络交换机（AI Back-End Switch）销售额在历史上首次超越了传统前端数据中心网络**（Celestica位居第一、NVIDIA第二、Arista第三、Cisco份额增速最快），预计2030年AI后端交换机市场将突破 **1,000亿美元**（广义AI网络迈向2,000亿美元）。



■ 图12：主流AI加速芯片单卡HBM显存容量演进（从80GB迈向1,024GB） — 从H100（80GB）到B200（192GB）、B300/Rubin（288GB）、AMD MI455X（432GB）及2027年Rubin Ultra（1,024GB），单卡显存容量四年增长逾12倍。



■ 图13：主流AI加速芯片单卡HBM总带宽演进（突破20 TB/s大关） — 随着HBM4 2048-bit总线导入，Vera Rubin与MI455X单卡显存带宽双双跨过19.2–22 TB/s门槛，Rubin Ultra更将冲刺32 TB/s。

在Scale-Up超节点互连领域，**NVIDIA** 凭借 **NVLink 5（单卡1.8 TB/s，224G PAM4 SerDes）与NVLink 6（Rubin单卡3.6 TB/s，整柜260 TB/s）** 建立了先发优势，但也面临物理工程挑战：

01 **NVL72跨卡权重分片将并发KV容量拉高26%**：在GB300 NVL72（18个计算托盘+9个NVSwitch交换托盘，130 TB/s全互连带宽，汇聚21TB HBM3E + 17TB LPDDR5X = 38TB统一内存池）中，利用NVLink域将模型权重跨72卡分片共享，整柜可同时驻留的长上下文KV会话数从TP=1下的293/585路升至 **370路（BF16）与740路（FP8）**，并发容量提升 **26%以上**，在MLPerf v6.0 DeepSeek-R1测试中跑出单柜 **673,936 tokens/s** 吞吐。

02 **NVLink Fusion生态收编与Kyber机柜78层PCB背板延期**：为应对开放联盟竞争，NVIDIA于2026年9月联手联发科（35亿美元合作）及Astera Labs推出 **NVLink Fusion**，允许云厂自研ASIC直接接入NVL72/MGX机柜背板。但在制造端，SemiAnalysis于2026年7月披露，NVIDIA原定推出的下一代超高密度 **Kyber机柜（支持NVL576及1,152卡全光互连）**，因其核心计算PCB背板高达 **78层**，在压合良率与信号完整性上遭遇瓶颈，**量产时间推迟至2028年**（过渡型NVL72x2双拼机柜亦取消），2026–2027年主力仍由采用无缆化（Cableless）计算托盘设计的Vera Rubin Oberon NVL72承担。

6.2 开放协议演进与过渡：UALink 2.0网内计算、AMD Helios借道博通Tomahawk 6与OCP ESUN

为打破NVLink闭源垄断，非NVIDIA阵营开辟了 **UALink** 与 **Scale-Up Ethernet（OCP ESUN / UEC）** 两条开放路线（见表6-1）：

01 **UALink 1.0/2.0微架构升级与2026“借道过渡”**：由AMD、阿里、Apple、AWS、Google、Meta、微软等70余家企业组建的UALink联盟，于2025年4月发布UALink 200G 1.0，并在**2026年4月发布UALink 2.0四大规范簇**。其基于IEEE P802.3dj 200Gbps PHY，采用64B事务层Flit → 640B链路层Flit → 680B FEC块帧结构，支持单Pod跨1–4个机柜（线缆<4米）全互连**1,024颗加速器**、往返延迟 **<1微秒（Sub-μs）**，并在2.0中引入 **网内计算（In-Network Compute）** 与 **UCIe 3.0芯粒级集成**。然而，由于Astera Labs与Marvell的原生UALink交换芯片需待2027年放量，**AMD定于2026下半年出货的72卡旗舰机柜“Helios”，需先采用将UALink协议封装在博通Tomahawk 6以太网交换芯片中进行“隧道传输（Tunneling）”的过渡方案**。

02 **博通102.4T Tomahawk 6与Scale-Up Ethernet（ESUN / UEC 1.0）**：博通依托全球首款量产**102.4 Tbps交换芯片Tomahawk 6**（支持64端口1.6T及CPO光电共封，两层拓扑即可构建12.8万张GPU集群）、250ns低延迟Tomahawk Ultra及800G网卡Thor Ultra，联合OCP推出**ESUN（纵向扩展以太网，吸纳UEC 1.0标准并定制精简Scale-Up报头）**，使云厂能用单一以太网硅片同时管理Scale-Up与Scale-Out网络。在欧美光互连节奏上，2026–2027年主力为1.6T/800G可插拔与铜缆，**2027下半年启动NPO（近封装光学）初始商用，2028年步入CPO（光电共封）规模部署**。

TABLE 09 · 第六章 · 超节点互连协议战 Yavin质观 · 核心数据矩阵

三大 Scale-Up 超节点互连协议阵营（NVLink vs. UALink vs. OCP ESUN）架构与生态对比

2026Q2 全球 AI 后端交换机销售额历史性反超前端网络；576 卡 Kyber 受困 78 层背板延期，全光 NPO 加速

对比维度	NVIDIA NVLink 5/6 & Fusion	开放联盟 UALink 1.0 / 2.0	OCP ESUN & 华为灵衢 UB+NPO
协议定位与阵营	**英伟达闭源护城河 + 统战收编** 联手联发科(35亿美元)推 **NVLink Fusion**	**AMD/阿里/苹果/AWS/谷歌/Meta等70+** 2026.04 发布 **UALink 2.0 四大规范簇**	**博通/OCP ESUN**：吸纳 UEC 1.0 精简报头 **华为灵衢 UB**：开放 600 页规范统一编址
单卡带宽与规模	NVLink 5 (1.8TB/s) → **NVLink 6 (3.6TB/s)** NVL72 整柜带宽 **260 TB/s** (KV容量+26%)	基于 200Gbps PHY，支持跨 1–4 机柜 **1,024 颗加速器** 全互连，延迟 **<1μs**	博通 **Tomahawk 6 (102.4 Tbps)** 撑 12.8 万卡 华为 **昇腾 960 支持 4,096 卡 (2μs延迟)**

对比维度	NVIDIA NVLink 5/6 & Fusion	开放联盟 UALink 1.0 / 2.0	OCP ESUN & 华为灵衢 UB+NPO
交换芯片与时间表	NVSwitch 大规模量产，允许云厂自研 ASIC 接入 NVL72/MGX 机柜背板	2.0 引入 **网内计算** 与 UCIe 3.0；但原生交换芯片待 2027 年，**AMD Helios 暂借 TH6 隧道**	博通 TH6 2026 量产；**华为昇腾 960 全球首发商用 NPO** (欧美 NPO 待 2027H2)
物理基建卡脖子点	**576 卡 Kyber 机柜受困于 78 层 PCB 背板良率推迟至 2028 年**；主力仍为 Oberon NVL72	线缆长度受限 (<4米)，协议规范与专用交换硅片存在约 12 个月量产时差	华为以 **5,500 个 Hi-ONE 光引擎替代 4.8 万颗光模块**，整柜省电 **>550kW**、可用度 99.8%

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

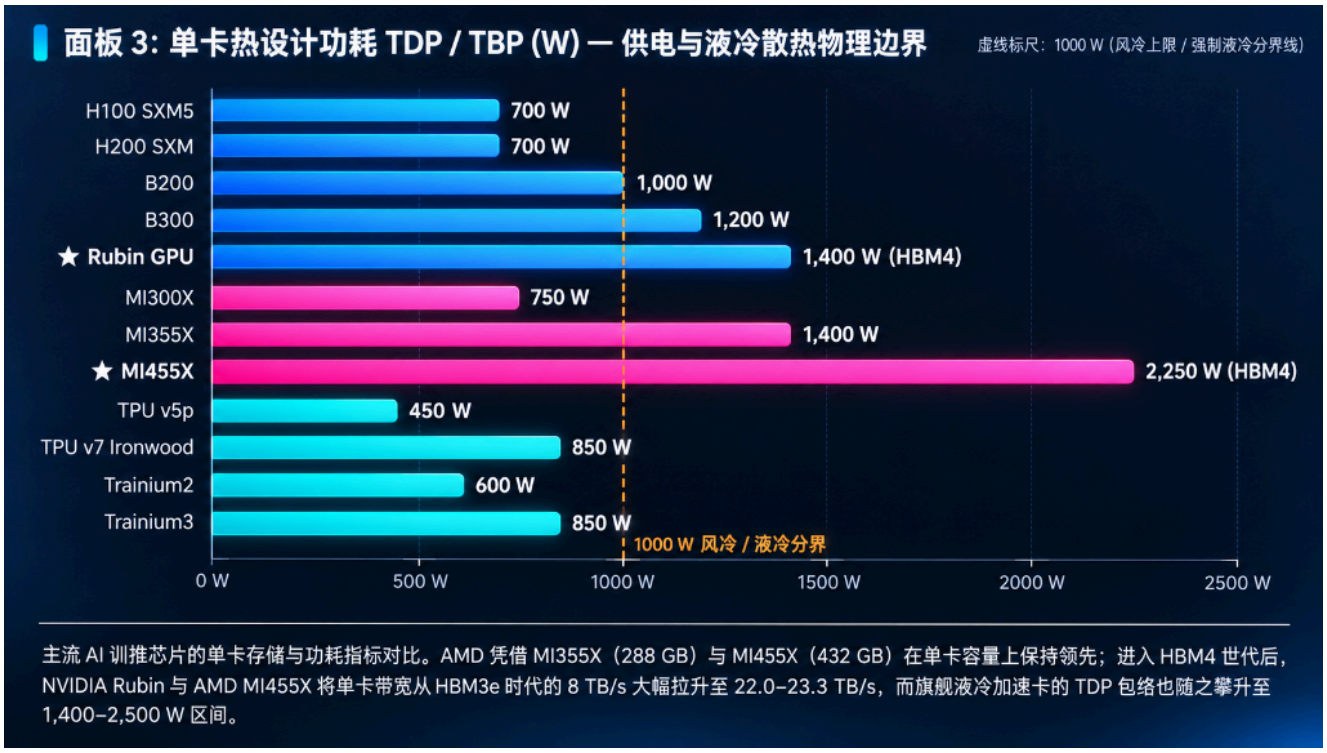
6.3 机房物理硬约束：单卡突破2千瓦、155kW秒级功率摆动、楼板承重极限与液冷白名单

当72颗高功耗GPU、18颗交换芯片与数十TB内存被集装进单台48U机柜时，电力、承重与散热成为决定超节点交付的物理门槛：

01 **单卡突破2,000W与秒级功率摆动（Power Swing）**：进入HBM4世代，NVIDIA Vera Rubin单卡功耗升至 **1,400W**，AMD Instinct MI455X更达到 **2,250W**，全面越过 **1,000W风冷散热上限**（见图14）。GB300与Rubin NVL72单机柜标称功耗达 **132–142 kW**，峰值瞬态功耗（EDPp）升至 **155 kW**。由于AI负载存在1–2秒内从60 kW跳变至150+ kW的“秒级功率摆动”，机柜需在每颗GPU侧集成65焦耳电解电容削峰模块（削减30%电网峰值拉取），并向2027年的800 VDC高压直流母排演进。

02 **2,400 kg/m²楼板承重极限**：满配GB300 NVL72机柜净重达 **1,580公斤**（单位重量价值密度达 **$3,165/kg**），在0.7平方米占地面积上产生 **2,200–2,470 kg/m²** 的地面压强，达到TIA-942标准架空地板承重上限（1,224 kg/m²）的 **1.6至2.4倍**，部署前必须对机房楼板进行结构加固。

03 **单柜液冷BOM升至70万–100万元与英维克白名单突围**：按冷板式液冷3–5元/W（约5,000元/kW）造价核算，GB200 NVL72单柜液冷价值量约55万–60万元；而在GB300与Rubin NVL72中，由于取消计算托盘风扇、将大串联冷板改为 **单芯片独立冷板**，柜内冷板与快接头（UQD/NVQD）用量增加 **2至3倍**，推动 **单机柜液冷总价值量提升30%–50%至70万–100万元人民币**（二次侧内部：冷板与CDU各占33%、快接头占25%、Manifold占10%，仅CLS交换机托盘液冷板单柜即达8万元）。在维谛（Vertiv）与台系奇鋐（AVC）、双鸿（Auras）之外，中国大陆液冷龙头 **英维克（002837.SZ）已正式打入NVIDIA的CDU、NVQD快接头、Manifold三大核心组件白名单**，并切入CLS交换机液冷板供应链。

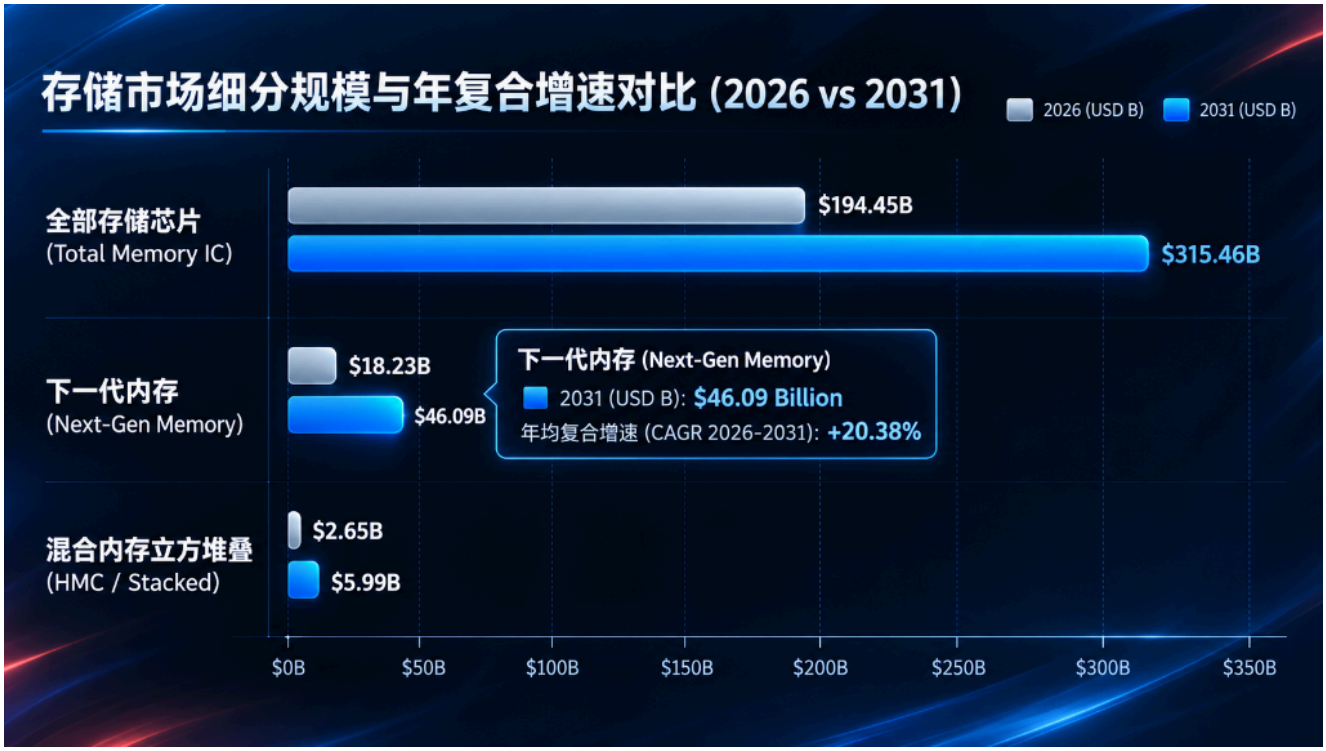


■ 图14：主流AI加速芯片单卡热设计功耗（TDP）演进与1000W液冷分界线 — 从昇腾910C（310W）、H100/Jalapeño（700W）到B200（1,000W风冷极限），再到B300/Rubin（1,400W）与AMD MI455X（2,250W），超节点全面进入全液冷时代。

07 CHAPTER 07 · NEXT-GEN MEMORY MEDIA & PIM/CIM
下一代内存介质演进——4F²垂直沟道、无电容3D X-DRAM、非易失内存与存算一体

7.1 全球内存IC市场迈向3,155亿美元：DDR6/LPDDR6接口升级与4F²垂直沟道晶体管（VCT）

从全行业周期看，全球内存集成电路（Memory IC）市场规模在2026年达到 **1,944.5亿美元**，预计将以 **10.17%的复合年增长率（CAGR）** 于2031年增至 **3,154.6亿美元**（见图15）。其中，DRAM占55.70%份额，NAND Flash受AI推理KV卸载拉动以11.35% CAGR增长；而涵盖新型非易失内存与三维堆叠架构的 **“下一代内存”细分市场，正从2026年的182.3亿美元以20.38%的CAGR扩张至2031年的460.9亿美元**（中东主权AI基建以21.65% CAGR领跑区域增速）。



■ 图15：全球内存IC与下一代内存市场规模跃迁（2026年 vs. 2031年预测，十亿美元） — 全球内存IC市场将从2026年的1,944.5亿美元增至2031年的3,154.6亿美元（CAGR 10.17%），其中下一代内存细分赛道以20.38%的CAGR扩张至460.9亿美元。

在主流接口标准与底层晶圆微缩层面，DRAM正推进两项代际演进：

01 **DDR6（17.6 Gbps）标准与长鑫全球首发量产LPDDR6（14.4 Gbps）**：JEDEC DDR6标准摒弃DDR5双32-bit设计，重构为 **四24-bit（Quad 24-bit）子通道架构**，起步速率8.8 Gbps、最高达 **17.6 Gbps（超频路线图达21 Gbps）**，配合CAMM2/LPCAMM2模组于2026年启动平台认证。而在低功耗领域，速率达 **14.4 Gbps** 的LPDDR6全球首发量产由 **中国长鑫存储**

（CXMT）于2026年8月底率先完成，并由小米旗舰智能手机首发搭载（依托长鑫第五代11.95nm平台与45:1深宽比电容，单晶圆产出提升50%以上）。

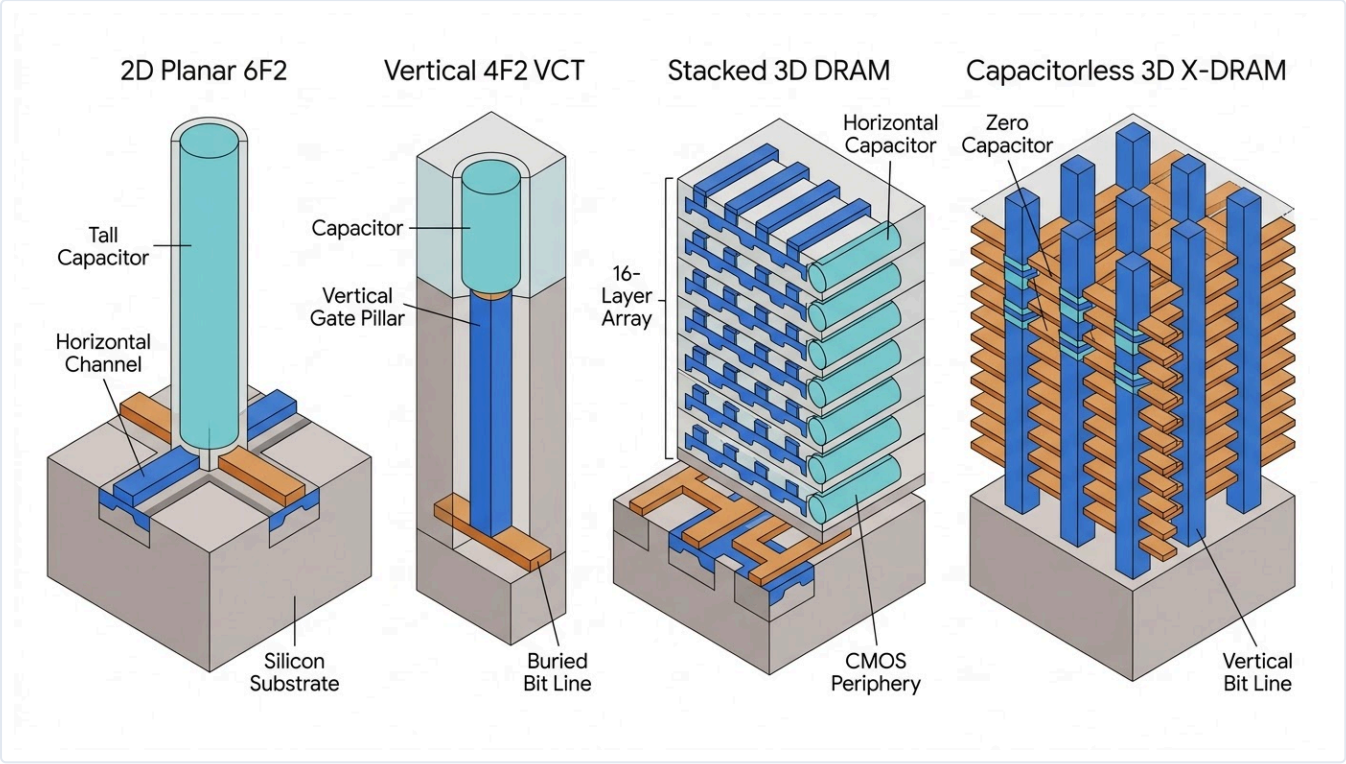
02 **4F²垂直沟道晶体管（VCT / VG-DRAM）突破平面微缩极限**：当平面DRAM逼近10nm以下（D0a/D0b）节点时，传统1T1C **6F²单元架构** 面临高深宽比（>45:1）电容倾倒漏电与水平晶体管短沟道效应两大物理限制。破局路径是将存取晶体管旋转90度直立，采用环绕栅极（GAA）垂直沟道设计（位线埋底、字线环中、电容置顶），**使单元面积直接缩减33%至理论极限的4F²**（导通电流 >10 μA/cell），SK海力士与长鑫存储均已将其列为下一代核心量产路线。

7.2 突破平面1T1C极限：三星16层VS-DRAM与NEO无电容3D X-DRAM硅片级实测验证

为彻底摆脱二维平面微缩的束缚，存储工业正沿着两条原生三维路径重构DRAM单元（见图16）：

01 **垂直堆叠电容式3D DRAM（VS-DRAM）**：借鉴3D NAND工艺，将晶体管与水平深槽电容在垂直方向多层堆叠，并利用晶圆键合将外围CMOS置于阵列下方或上方（CBA）。三星已展示 **16层垂直堆叠VS-DRAM原型**，目标在2030年前后实现规模商业化。

02 **无电容（1T0C / 3T0C）氧化物半导体与NEO 3D X-DRAM硅验证**：通过省去占空间最大的电容结构，利用IGZO薄膜晶体管极低关态漏电或硅基浮体效应（Floating-Body Cell）直接存储电荷。**2026年4月，美国NEO Semiconductor宣布其3D X-DRAM概念验证（POC）测试芯片成功通过硅片级电学验证并获战略投资**。3D X-DRAM完全兼容现有 **300层以上3D NAND制造工艺**（无需EUV光刻微缩），采用双栅极浮体1T0C/3T0C单元，POC实测显示：**读写延迟低于10纳秒（<10 ns），85°C高温下数据保持时间超过1秒（优于JEDEC 64ms标准的15倍以上，仿真可达450秒），擦写寿命超 10^{14} 次**。据此测算，单颗DRAM裸片容量有望达到 **512Gb（为传统2D DRAM密度的8至10倍）**，刷新功耗降低85%、激活延迟降低50%。



■ 图16：DRAM存储单元底层架构跨代演进：从2D平面6F²、垂直沟道4F²到无电容3D X-DRAM — DRAM微缩演进四阶段：（1）传统2D平面6F²（水平晶体管+深孔电容）→（2）垂直沟道4F² VCT（GAA直立晶体管，面积缩减33%）→（3）垂直堆叠电容3D DRAM（多层水平电容）→（4）无电容3D X-DRAM（1T0C浮体/IGZO，兼容300层+ 3D NAND工艺实现10倍密度）。

7.3 28nm以下eNVM替代（MRAM/PCM/ReRAM/FeRAM）与近存/存内计算（PIM/CIM）双轨演进

在DRAM与NAND之外，以 **MRAM、PCM、ReRAM与FeRAM** 为代表的新型非易失性内存（eNVM，占下一代内存市场12%份额）正由两大商业需求拉动：一是 **28nm及以下FinFET/FD-SOI先进制程对传统浮栅eFlash的替代**（传统eFlash需高压擦写，与28nm以下超薄栅氧层物理不相容）；二是 **借助CXL总线重塑持久内存层**（见表7-1）：

- **磁阻内存（MRAM，2025年市场规模45亿美元）**：台积电（22ULL及16/12nm）与格罗方德（22FDX）已将 **STT-MRAM** 确立为车规与IoT替代eFlash的主流方案，被宝马、丰田、通用指定用于2026–2027新一代EV平台（预计2028年超60%新建22nm以下车规MCU采用eMRAM）。独立MRAM龙头 **Everspin** 2026Q2营收创1,870万美元新高（+42%），并于 **2026年9月30日演示全球首个基于CXL接口的MRAM持久内存平台**（写速比NAND快100倍）。在更前沿的 **SOT-MRAM（自旋轨道矩，读写路径分离，功耗降至STT的1%）** 领域，2026年8月台积电与UT Austin在300mm晶圆上验证了 **2ns极速切换、2pJ超低能耗** 的SOT-MRAM交叉阵列；**中国致真存储（TMC）发布了全球首颗8Mb SOT-MRAM芯片**；imec分拆的Vertical Compute获6,700万美元融资推进LLM专用高密度MRAM。
- **相变（ePCM）、阻变（ReRAM）与铁电（FeRAM）**：**意法半导体（ST）** 联合三星代工在28nm/18nm FD-SOI上规模量产 **ePCM**（Stellar车规ASIL-D MCU集成21MB ePCM，耐受

+165°C超高温并支持单比特直接覆写零停机OTA；18nm STM32V8能效提升超50%）；**Weebit Nano** 领跑独立ReRAM IP授权（手握1.64亿澳元现金）；**CEA-Leti** 在22nm FD-SOI上验证10nm厚掺锆氧化铪（HZO）铁电FeRAM（耐受 >10^{12} 次循环），德国 **FMC** 亦推进基于氧化铪的DRAM+技术。

TABLE 10 · 第七章 · 下一代内存介质 Yavin质观 · 核心数据矩阵

六大下一代与新型非易失性内存介质（Next-Gen Memory）物理特性与商业化矩阵

全球内存 IC 迈向 3,155 亿美元；下一代内存以 20.38% CAGR 增至 460.9 亿美元，无电容 3D X-DRAM 与 MRAM 领跑

介质技术路线	底层物理存储机制	读写延迟与擦写寿命	核心微缩与能效优势	2026 商业化里程碑与代表厂商
4F² VCT DRAM (垂直沟道晶体管)	晶体管旋转 90° 直立 GAA 位线埋底 + 字线环中 + 电容置顶	**~10 ns** 导通电流 >10 μA/cell	单元面积较传统 6F² **直接缩减 33%** 突破 10nm 以下平面短沟道极限	**SK海力士** 列为 10nm 以下主力路线；**长鑫存储** 提前布局垂直环绕栅极 DDR5
无电容 3D X-DRAM (1T0C 浮体 / IGZO)	砍掉高深宽比电容，用双栅极浮体或 IGZO 沟道存电荷	**POC实测 < 10 ns** 85°C 保持 >1s，寿命 **>10^{14} 次**	兼容 **300层+ 3D NAND 工艺 (免EUV)** 单裸片冲 **512Gb (10倍密度)**，刷新功耗降 85%	**NEO Semiconductor** 2026.04 通过硅片级 POC 验证；三星同步推进 16 层 VS-DRAM
STT / SOT-MRAM (自旋矩磁阻内存)	磁性隧道结 (MTJ) 自旋翻转 SOT 实现读写路径物理分离	STT: 3–10 ns **SOT: 2 ns 极速 / 2 pJ 能耗**	28nm 以下全面替代高压 eFlash SOT 写入功耗降至 STT 的 **1%**	**Everspin** 演示全球首个 **CXL-MRAM**；**台积电** 300mm 晶圆验证 SOT；**致真存储** 首发 8Mb SOT
嵌入式相变 ePCM (硫系相变内存)	硫系化合物晶态与非晶态受控可逆热相变	纳秒级读取 / 单比特直接覆写 耐受 **+165°C Grade 0 超高温**	无需像 Flash 先擦除整个扇区 支持车规 MCU **零停机 OTA 升级**	**意法半导体 (ST) + 三星 28/18nm FD-SOI** 量产 21MB Stellar 车规 MCU 与 STM32V8 (能效+50%)
氧化物阻变 ReRAM (導電细丝阻变)	金属氧化物介质层导电细丝生灭调控电阻	读 **5–20 ns** / 多比特电导态 寿命 10^6–10^9 次	低电压、后端工艺 (BEOL) 易集成 天然契合模拟/数字存算一体阵列	**Weebit Nano** 领跑独立 IP 授权 (手握1.64亿澳元现金)；**亿铸科技** 推进大算力 ReRAM CIM
HZO 铁电 FeRAM (掺锆氧化铪铁电)	10nm 超薄氧化铪铁电晶畴电场极化翻转	**5–20 ns** 擦写寿命突破 **>10^{12} 次**	完全兼容标准 CMOS 产线 极低写入能耗，兼具 DRAM 速度与非易失性	**CEA-Leti** 在 22nm FD-SOI 验证 HZO 铁电电容；德国 **FMC** 推进 DRAM+ 与 3D 铁电存储

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

与此同时，旨在消除"冯·诺依曼数据搬运税"的 **存算一体（PIM / PNM / CIM）** 形成了"国际存储巨头主攻近存处理（Near-Bank PIM）、中国初创深耕存内计算（CIM）"的双轨格局（见表7-2）：三星 **HBM-PIM**（单栈1.2 TFLOPS，GEMV系统提速2–2.5倍、降耗60%–70%）、SK海力士 **GDDR6-AiM/AiMX**（LLM解码提速13–16倍）与法国 **UPMEM**（DDR4 DIMM DPU，基因组比对提速6–31倍）率先落地，并通过新一代 **DCC数据中心级PIM编译器** 解决跨Bank分块难题（在AttAcc PIM上实现单算子13.17倍加速、端到端LLM推理提速4.52倍）。

TABLE 11 · 第七章 · 存算一体芯片矩阵 Yavin质观 · 核心数据矩阵

全球与中国存算一体（PIM / PNM / CIM）核心代表企业、介质路线与实测能效对比矩阵

国际存储巨头主攻近存处理（HBM-PIM / AiM），中国初创军团深耕成熟制程存内计算（CIM）实现 7–10 倍能效跃升

代表企业 / 机构	技术路线与介质底座	核心产品架构与实测能效 / 加速指标	商业化落地定位与最新进展
三星电子 (Samsung)	**HBM-PIM / LPDDR-PIM** 近存储体处理 (Near-Bank)	• 在 DRAM Bank 旁集成 16-bit 浮点可编程 PCU • 单栈提供 **1.2 TFLOPS**，系统端吞吐提升 **2–2.5倍**、能耗降 **60%–70%**	无需修改硬件接口即可替换标准 HBM，正融入 **HBM4E 定制 3nm 逻辑底座**
SK海力士 & UPMEM (AiM / DIMM-PIM)	**GDDR6-AiM / AiMX** UPMEM **DDR4 DIMM DPU**	• 海力士 AiM 在 16Gb 颗粒内嵌处理单元，**LLM 解码提速 13–16 倍** • UPMEM 单芯片 8 个 DPU，基因组比对提速 **6–31 倍**	新一代 **DCC 数据中心级编译器** 攻克跨 Bank 分块，端到端 LLM 推理再提速 **4.52 倍**
知存科技 (Witmem) (中国模拟 Flash CIM)	**浮栅 / NOR Flash** 模拟存内计算 (Analog CIM)	• **WTM2101** 成为全球首颗百万级量产商用存算芯片 • **WTM-8** 以传统方案 **5% 的功耗** 提供 **>24 TOPS** 边缘算力	稳居全球端侧穿戴与 AI 眼镜低功耗视觉/语音 CIM 商用出货榜首
后摩智能 (Houmo) (中国数字 SRAM CIM)	**12nm 全数字 SRAM 存算** 彻底消除 AD/DA 转换损耗	• **鸿途 H30 / H50**：物理算力 **256 TOPS**、功耗仅 **35W** • 达成 **15 TOPS/W 超高能效比（超过传统架构 7 倍）**	攻克大算力精度无损难题，规模切入高阶智能驾驶、Robotaxi 与泛机器人市场
亿铸科技 & 千芯科技 (中国 ReRAM / 可重构CIM)	亿铸：**ReRAM 全数字超异构** 千芯：**可重构存算一体**	• **亿铸** 基于新型忆阻器 POC 芯片实测能效比超传统 AI 芯片 **10 倍以上** • **千芯** 芯片吞吐量提升 **10–40 倍**	千芯已通过国内头部互联网大厂内测；在 **12–28nm 成熟制程** 开辟非 EUV 大算力底座

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

08 CHAPTER 08 · CHINA SYSTEM BREAKTHROUGH & STRATEGY
中国产业链“以存代算、以网补卡”系统突围与2026–2028选型投资框架

8.1 算法与软件层减负：DeepSeek MLA削减93.3%显存、腾讯FlashMemory与存量卡盘活

在海外顶尖3nm晶圆代工与HBM4受出口管制、单卡显存容量与带宽存在客观代差的背景下，中国AI产业链在2025–2026年走出了一条 **“算法级极限压缩 + 系统级以存代算 + 超节点以网补卡”** 的系统突围路径（见表8-1）。

在算法与分布式推理框架层，国内团队从源头降低了对进口HBM容量的依赖：

01 **DeepSeek MLA + 混合量化削减93.3% KV显存**：通过低秩键值联合压缩（仅缓存压缩后的潜在向量），DeepSeek-V2单靠MLA架构即较传统稠密GQA节省71.88%的KV空间；叠加网络层数精简（60/95）与6-bit量化（6/16），**综合KV Cache显存占用较基线下降93.3%**（计算式：1 – 0.2812 × (60/95) × (6/16) = 93.3%）。

02 **腾讯FlashMemory省90%显存反升准确率**：在2026年腾讯云披露的面向DeepSeek-V4的FlashMemory与Memory Indexer生产实测中，平均KV Cache显存开销从基线的0.93 GB压降至 **0.10 GB（压缩约90%）**，且在节省九成显存的同时，模型平均推理准确率微升 **0.6个百分点**。

03 **Mooncake + SGLang HiCache盘活存量H20与A800/H800集群**：国内大厂已在算力受限的NVIDIA H20及存量A800/H800生产集群上，完整跑通搭载 **DeepSeek-V4与Qwen3-235B FP8** 的SGLang HiCache + Mooncake P/D分离与分层缓存方案，利用国产大容量DDR5主存与PCIe Gen5 NVMe盘承接历史KV卸载，在不依赖禁运新卡的前提下将存量集群有效推理吞吐提升 **2倍以上**。

8.2 系统工程换道超车：华为昇腾960全球首发商用NPO全光超节点、灵衢统一编址与OceanStor M900

在硬件系统工程层，华为推出的 **“昇腾960全光超节点 + 灵衢（UnifiedBus）网络 + OceanStor M900 AI记忆存储”** 三位一体架构，构建了“以网补卡、以存代算”的本土标杆：

01 **灵衢（UnifiedBus）总线与4,096卡全局内存统一编址**：为消除PCIe、RoCE、CXL、HCCS十余种协议转换带来的通信损耗，华为于2025年8月开放600余页灵衢技术规范。灵衢将节点间互连带宽推升至TB/s级、端到端延迟从7微秒压至 **2微秒（2 μs）**，并实现了 **跨4,096张卡的全局内存统一编址**——超节点内所有NPU的片上显存与主机内存被池化为单一共享大内存池，任意

计算卡可直接按内存地址读写远端KV Cache；通过二层CLOS四平面组网更可扩展至51.2万乃至100万张卡。

02 **昇腾960全球首发商用NPO（近封装光学）超节点**：在欧美厂商NPO方案尚需等到2027下半年商用之际，华为于 **2026年9月17日全联接大会（HC 2026）发布全球首个商用NPO超节点——昇腾960**。单台超节点支持 **4,096张卡互连**（提供8 EFLOPS FP8或16 EFLOPS FP4算力），内部通过部署 **5,500个自研Hi-ONE高密光引擎，直接替代了原本需要的48,000颗传统可插拔光模块**，使整机总功耗降低 **超过550 kW**，系统可用度提升至 **99.8%**。

03 **OceanStor M900 AI记忆存储（64 PB共享容量、16倍SSD寿命）**：同步发布的L3.5层OceanStor M900集群，通过灵衢总线"单跳直连（One-hop direct connection）"消除协议转换与CPU中断转发，提供 **64 PB全共享内存空间、TB/s级带宽与低至60微秒（60 µs）的推理访问延迟**；同时引入 **KV感知自适应存储与Retention生命周期预测算法**，支持每日24次全盘写入（**24 DWPD**），**将SSD读写寿命延长16倍，承诺三年免换盘稳定运行**，解决了高频KV卸载磨损闪存颗粒的折旧痛点。

8.3 本土制造底座与全局配置策略：长鑫首发LPDDR6与12层HBM3E试产、存算芯片及三阶选型框架

在底层芯片制造端，中国本土存力与互连供应链在2026–2027年取得了实质性进展：

- **长鑫存储（CXMT）首发量产LPDDR6、推进12层HBM3E试产并进军NAND**：截至2026Q1，长鑫已占据全球DRAM市场约 **8%** 份额（年底总月产能升至 **30万–35万片晶圆/月**，规划2031年达80万片/月，科创板IPO预计估值超 **2,000亿元**）。其16nm 16Gb DDR5颗粒良率于2026年8月突破 **90%**，直至2027年的DRAM产能已被国内客户100%锁定，并于8月底率先实现 **LPDDR6** 全球首发量产。在HBM赛道，长鑫划拨 **5万片/月（50K WPM）** 晶圆产能，上海HBM后段TSV封装基地于2026年底投产；继16nm HBM3交付样品（2026年出货约1.22亿Gb）后，**2026年9月长鑫已启动12层HBM3E风险试产（初期良率约25%），计划于2027年实现12层HBM3E规模量产，将与海外三巨头的技术代差缩短至2至3年**。同年9月，长鑫在北京设线跨界进军NAND闪存，与 **长江存储（YMTC，Xtacking 4.0全球NAND份额达14%居第三）** 共同构建国产全栈存力底座。
- **国产互连芯片与成熟制程存算一体（CIM）协同演进**：除全球内存接口第一的 **澜起科技（36.8%份额）** 外，**盛科通信（688702.SH，Arctic系列400G/800G RoCE交换芯片）**、中兴通讯及星云智联、云豹智能等DPU厂商正承接国内超节点互连需求。而在成熟制程（12nm–28nm）存算一体赛道，**知存科技**（WTM2101全球首颗百万级量产NOR Flash模拟存算芯片，WTM-8以5%功耗提供>24 TOPS算力）、**后摩智能**（12nm全数字SRAM存算智驾芯片鸿途H30/H50，256 TOPS、15 TOPS/W超传统7倍能效）、**亿铸科技**（基于ReRAM忆阻器的全数字超异构大算力POC芯片，实测能效超传统AI芯片 **10倍以上**）与 **千芯科技**（可重构存算芯片吞吐提升10–40倍并通过互联网大厂内测），开辟了不依赖EUV光刻机的国产算力补充路径。

TABLE 12 · 第八章 · 中美存力与互连全景对标 Yavin质观 · 核心数据矩阵

北美前沿方案与中国本土“以存代算、以网补卡”产业链换道超车全景对照表

在 3nm 代工与 HBM4 受限下，中国以“MLA 算法减负 + 昇腾 960 全光 NPO + M900 记忆存储 + 长鑫 HBM3E”系统破局

产业链核心环节	北美及日韩前沿标杆方案	中国本土突围主体与代表产品	2026–2027 本土突围深度与阿尔法评估
算法与推理引擎层 (源头削减重算税)	**vLLM / SGLang / LMCache** KIVI 2-bit 量化与 PEEK 调度	**DeepSeek + 月之暗面 + 腾讯云** · **DeepSeek MLA+量化削减 93.3% KV 显存** · **腾讯 FlashMemory** 省 90% 显存反升准确率 · **Mooncake P/D分离** 长文本吞吐 **+525%**	全球领跑 · 盘活存量 在不依赖禁运新卡下，将国内存量 H20/A800 及国产卡集群有效推理吞吐提升 **2 倍以上**
超节点光互连网络 (Scale-Up 统一内存域)	**NVIDIA NVLink 6 / Kyber** UALink 2.0 / 博通 Tomahawk 6	**华为昇腾 960 + 灵衢总线 + 盛科通信** · **全球首发商用 NPO 超节点 (4,096卡)** · **5,500 个光引擎替代 4.8 万颗光模块** (省电550kW) · **灵衢 UnifiedBus** 2μs 延迟全局内存统一编址	换道超车 · 以网补卡 欧美 NPO 需待 2027H2，华为率先商用全光 NPO 绕开 78 层铜背板瓶颈，可用度达 **99.8%**
G3.5 / L3.5 记忆存储 (PB级上下文闪存池)	**MinIO MemKV + BlueField-4** VAST Data CNode (<$8 万/PB)	**华为 OceanStor M900 + DeepSeek 3FS** · **M900** 单跳直连提供 **64 PB** 共享空间、**60μs** 延迟 · **24 DWPD 将 SSD 寿命延长 16 倍** (三年免换盘)	架构独创 · 以存代算 首创 KV 感知生命周期预测，彻底解决高频 KV 卸载烧穿闪存颗粒的折旧痛点
DRAM、HBM 与闪存 (底层存储晶圆制造)	**SK海力士 / 三星 / 美光** HBM4 (3nm底座) + 256GB SOCAMM2	**长鑫存储 (CXMT) + 长江存储 (YMTC)** · 长鑫 **全球首发量产 LPDDR6 (14.4Gbps)** · 16nm DDR5 良率 **>90%**，月产能达 **30–35万片** · **5万片/月产能强攻 12层 HBM3E** (2027量产)	代差收窄至 2–3 年 长鑫 2027 年 DRAM 产能已被国内大厂 100% 锁空，并跨界进军 NAND 与长存构建全栈底座
互连芯片、液冷与CIM (二阶高弹性硬件链)	**Astera Labs / 维谛 (Vertiv)** Besi 混合键合 / 奇鋐双鸿	**澜起科技 + 英维克 + 本土存算四小龙** · **澜起科技** 内存接口占 **36.8%** 居全球第一 · **英维克** 打入英伟达三大液冷白名单 · **知存/后摩/亿铸/千芯** 成熟制程存算提效 **7–10倍**	全球龙头 + 非EUV后路 澜起与英维克同步收割海内外红利；成熟制程 CIM 为极端光刻封锁备好大算力后路

资料来源：上市公司财报/招股书、JEDEC、PCI-SIG、SemiAnalysis 及 Yavin质观 整理（2026.10）

综合全文八大章节的工程实测与供应链数据，面向2026–2028年的企业架构选型与产业投资可遵循三大核心准则：

01 **企业级“去重算税”三阶架构选型准则**：第一阶段（单节点或并发 C<32 的容量悬崖之下），无需盲目采购外置存储，优先启用 **SGLang统一基数树 + PEEK双树调度 + KIVI 2-bit量化 + 多租户缓存加盐**；第二阶段（64K–128K长文本、跨过 C=48 容量悬崖的多机集群），部署 **Mooncake P/D分离 + LMCache + G3.5白盒闪存池（MinIO MemKV + BlueField-4，<$80K/PB）或华为OceanStor M900**，用微秒级闪存命中替代50%重算；第三阶段（万亿参数MoE与百万Token智能体工厂），导入 **Scale-Up统一内存域（NVL72或昇腾960 NPO）+ SOCAMM2/MRDIMM + CXL 4.0近存计算（CXL-PNM）**，并同步完成150kW+液冷与2,400 kg/m²楼板加固。

02 **把握“上游存储涨价周期”与“中游解耦隐形冠军”两条确定性主线**：HBM4的3–4倍晶圆面积折损决定了全球DRAM与企业级eSSD的供需紧平衡将持续至2028年，存储原厂（SK海力士、三星、美光及本土长鑫存储、长江存储）处于高景气周期；而在存算解耦的物理连接处，**澜起科技（MRDIMM套片翻倍+CXL）、Astera Labs（Scorpio X + Leo CXL + NPO）、优群科技（SOCAMM2/DDR6连接器）、VAST Data（DPU原生数据OS）、英维克（液冷三大白名单）以及Besi/ASMPT/韩美半导体（TCB扩产与2028混合键合拐点）** 享有明确的单机柜ASP扩容红利。

03 **重估中国“超节点光互连 + 记忆存储”的系统工程价值**：在单卡制程受限背景下，中国算力基建的核心突破口在于以 **华为昇腾960 NPO全光超节点、灵衢统一内存总线、OceanStor M900记忆存储、长鑫HBM3E/DDR5与国产RoCE/CXL互连链** 为代表的全栈系统输出。AI算力竞赛的下半场，比拼的不再是单纯堆砌GPU数量，而是谁能构建更高效率的存力与互连底座，让每一张计算卡少做无用功。

APPENDIX & METHODOLOGY · 资料口径与全量信源

附录：四级信源说明与297项参考资料索引

DATA & REFERENCES

资料口径与四级信源分级说明： （1）**上市公司财报与招股书（最高优先级）**：SK海力士、三星电子、美光科技（Micron FQ4 26）、澜起科技、嘉泽端子（BizLink/Argosy）、Astera Labs、Marvell、博通、英维克、中际旭创、长鑫科技、盛合晶微等官方财报与招股书； （2）**国际标准组织与权威机构**：JEDEC（JESD270-4 HBM4、JESD328 SOCAMM2、MRDIMM）、PCI-SIG（PCIe 7.0）、CXL Consortium（CXL 4.0）、UALink Consortium（UALink 2.0）、SemiAnalysis、TrendForce、IDC、Yole Group等； （3）**顶会论文与开源工程**：OSDI / SOSP / ASPLOS / ICML / ISCA / IEEE JSSC / ISSCC / IEDM（vLLM PagedAttention、SGLang RadixAttention、Mooncake、KIVI、LMCache、MinIO MemKV、H2O、StreamingLLM、Neo 3D X-DRAM等）； （4）**供应链交叉验证**：未上市主体或未公开产线良率、晶圆惩罚系数（Trade Ratio）及机柜级BOM拆解均经多源交叉核验。

一、AI算力BOM重构、HBM4/4E与逻辑底座核心信源

[01] HBM3e vs HBM4: 2026 Specs, Performance & Supply Guide - Kynix (kynix.com/Blog/hbm3e-vs-hbm4-...)

[02] AI Accelerator Manufacturing Cost Estimates (2026) - Silicon Analysts (siliconanalysts.com/data/ai-chip-costs)

[03] Marvell Expands AI Memory and Storage Infrastructure with (650group.com/blog/marvell-expand...)

[04] NVIDIA B100 Cost Breakdown: BOM and Yield Economics (siliconanalysts.com/analysis/b100-cost-...)

[05] Why Nvidia's Next AI Battle Is About Tokens per Watt - I/O Fund (io-fund.com/ai-stocks/nvidia-ai...)

[06] NVIDIA GB200 Supply Chain: The Global Ecosystem Explained (intuitionlabs.ai/articles/nvidia-gb2...)

[07] Nvidia's memory costs soar 485%, latest AI systems now cost $7.8 (tomshardware.com/tech-industry/artif...)

[08] Nvidia Vera CPU vs Intel, AMD - Tech Insider (tech-insider.org/nvidia-vera-cpu-int...)

[09] NVIDIA Vera Rubin: Full Guide to Rubin GPU, Vera CPU and NVL72 (insidedeeptech.com/nvidia-vera-rubin-n...)

[10] KV Cache Architecture and Local I/O Assessment for Multi-Hour (chierhu.medium.com/kv-cache-architectu...)

[11] NVIDIA Vera CPU 为AI 工厂提供高性能、高带宽和高效率 (developer.nvidia.cn/blog/nvidia-vera-cp...)

[12] NVIDIA Vera CPU: The Amdahl Argument for a Purpose-Built AI (atlaspeakresearch.com/report/1683d7)

[13] HBM4 Explained: The AI Memory Bottleneck - TechDirectory (techdirectory.sg/kb/hbm4-explained)

[14] SOCAMM2 | Micron Technology Inc. - 美光科技 (micron.cn/products/memory/lpd...)

[15] Sandisk HBF: The Missing Memory Tier to Unlock HBM-Powered AI (ersaelectronics.com/blog/sandisk-hbfs)

[16] Synopsys CXL 4.0 IP Hits 128 GT/s, Claims 3-6x KV Cache Offload (storagereview.com/news/synopsys-cxl-4...)

[17] High Bandwidth Memory - Wikipedia (en.wikipedia.org/wiki/High_Bandwidth...)

[18] HBM undergoes major architectural shakeup as TSMC and GUC (tomshardware.com/pc-components/dram/...)

[19] HBM4 and the Shift to Customized AI Memory - Utmel (utmel.com/blog/news/semiconductu...)

[20] 美光專為NVIDIA Vera Rubin 打造的HBM4、SOCAMM2 (tw.micron.com/about/press/news/mi...)

[21] HBM is undergoing major changes (eu.36kr.com/en/p/3969913513341185)

[22] Samsung Foundry Pivots Half of 4nm Capacity to HBM4 - Tech Times (techtimes.com/articles/326959/202...)

[23] Sk Hynix's AI Strategy: Analysis of Dominance in Semiconductor (klover.ai/sk-hynix-ai-strateg...)

[24] AI Accelerators Usher In New Era For IC Test (semiengineering.com/ai-accelerators-ush...)

[25] Samsung introduces SOCAMM2 LPDDR5X memory module for AI (tomshardware.com/tech-industry/samsu...)

[26] SK Hynix to Use TSMC 3nm for HBM4, Samsung May Follow - SmBom (smbom.com/news/14496)

[27] Hanmi Semiconductor Cements TC Bonder Leadership at (techtimes.com/articles/329060/202...)

[28] From UCLA's Semiconductor Hub to 16-Hi HBM: The Thermal and (ninescrolls.com/insights/from-ucla-...)

[29] HBM Innovation Outpaces Standards Development - EE Times (eetimes.com/hbm-innovation-outp...)

[30] Why Hybrid Bonding is Replacing Solder Bumps in Advanced 3D (utmel.com/blog/categories/tec...)

[31] Samsung admits that its cutting-edge process is not competitive (en.eeworld.com.cn/news/manufacture/ei...)

[32] The current state of Hybrid Bonding in 2026 — TSMC sits at 6 (tomshardware.com/tech-industry/semic...)

[33] TSMC Asks Suppliers for 5 µm HBM Microbumps — Halving HBM4's (ninescrolls.com/news/tsmc-asks-supp...)

[34] Hybrid Bonding in 2026: TSMC Ships 6 µm While Its Newest (ninescrolls.com/news/hybrid-bonding...)

[35] DDR5, CAMM2, LPCAMM2, CUDIMM, MRDIMM, DDR6 - TIPTECH (tiptech.sk/en/blog/nove-format...)

[36] Samsung Electronics advances memory modules (electronics-journal.com/news/104572-samsung...)

[37] DDR4/5 Memory Connector | SO-DIMM | JEDEC SO-024 and (amphenol-cs.com/ddr.html)

[38] Introducing LPDDR5 SOCAMM2 Connector - Argosy Research Inc. (argosytw.com/en/news/introducing...)

[39] NVIDIA Moves to SOCAMM2, Phases Out Initial SOCAMM Design (techpowerup.com/341002/nvidia-moves...)

[40] Micron Delivers Industry's Highest Capacity SOCAMM2 for Low (investors.micron.com/news/press-release/...)

[41] News Posts matching 'SOCAMM2' - TechPowerUp (techpowerup.com/news-tags/SOCAMM2)

[42] “如今是低功耗之战”.....继HBM之后,'SOCAMM2'量产竞争升温- MK (mk.co.kr/cn/business/12022062)

[43] JEDEC JESD328 - Accuris Standards Store (store.accuristech.com/standards/jedec-jes...)

[44] SOCAMM: A Full Guide to Detachable LPDDR5X When Soldered (insidedeeptech.com/socamm-lpddr5x-ai-s...)

[45] CAMM Connectors - Amphenol Communications Solutions (amphenol-cs.com/connectors/camm-con...)

[46] new(ish), screwable (replaceable, non soldered) LPDDR5X RAM (reddit.com/r/LocalLLaMA/commen...)

[47] CAMM (memory module) - Wikipedia (en.wikipedia.org/wiki/CAMM_(memory_m...)

[48] 产业深度研究与技术分析报告 (storagereview.com/zh-CN/news/micron-1...)

[49] Introducing Samsung's SOCAMM2: New LPDDR Memory Module (semiconductor.samsung.com/news-events/tech-bl...)

[50] Nvidia's Vera Rubin platform in depth - Tom's Hardware (tomshardware.com/pc-components/gpus/...)

[51] Nvidia Offers First Look at Vera Rubin Superchip - PCMag (pcmag.com/news/nvidia-offers-...)

[52] JEDEC nearing completion of SOCAMM 2 LPDDR5X memory standard (videocardz.com/newz/jedec-nearing-...)

[53] AMD confirms EPYC SOCAMM2 support starts with Verano in 2027 (videocardz.com/newz/amd-confirms-e...)

[54] Server Memory - AMD (amd.com/en/products/process...)

[55] AMD Previews Preliminary EPYC Venice Performance Projections (reddit.com/r/hardware/comments...)

[56] “Introducing the CXL 4.0 Specification” Webinar Q&A Recap (computeexpresslink.org/blog/introducing-th...)

[57] Powering the Future of Memory-Centric Computing with CXL 4.0 VIP (community.cadence.com/cadence_blogs_8/b/f...)

[58] Compute Express Link® (CXL®) (computeexpresslink.org/wp-content/uploads/...)

[59] Synopsys Rolls Out First Complete CXL 4.0 IP for AI Memory (allaboutcircuits.com/news/synopsys-rolls...)

[60] CXL 4.0 Verification for AI & HPC Systems - SmartDV (smartdvtech.com/cxl-4-0-verificatio...)

[61] Introducing Compute Express Link® (CXL®) 4.0 (computeexpresslink.org/wp-content/uploads/...)

[62] Introducing the CXL 4.0 Specification - Compute Express Link (computeexpresslink.org/wp-content/uploads/...)

[63] About CXL® - Compute Express Link (computeexpresslink.org/about-cxl)

[64] CXL 4.0 IP Solution for AI Memory Connectivity | Synopsys (synopsys.com/blogs/chip-design/c...)

[65] Panmnesia, Meta Researchers Outline Cross-Rack CXL Fabric (supercomputing.news/emerging/panmnesia-...)

[66] Controller IP | Panmnesia (panmnesia.com/product/ctrl)

[67] Compute Express Link (CXL) Explained: Protocols, Device Types (ic-components.com/blog/Compute-Expres...)

[68] CXL Adds Port Bundling to Quench AI Thirst - EE Times (eetimes.com/cxl-adds-port-bundl...)

[69] Breaking Through the KV Cache Memory Wall with DPU ... - Marvell (marvell.com/blogs/breaking-thro...)

[70] Inside the NVIDIA Vera Rubin Platform: Six New Chips, One AI (developer.nvidia.com/blog/inside-the-nvi...)

[71] Reimagining LLM Inference Infrastructure with Memory-Centric KV (cs.virginia.edu/~skadron/Papers/KVs...)

[72] LIQID Memory and GPU Pooling for Neoclouds (liqid.com/solutions/neocloud)

[73] Proxics: an efficient programming model for far memory accelerators (arxiv.org/pdf/2604.18120)

二、AI推理“去重算税”、引擎优化与G3.5上下文闪存核心信源

[01] From Cache Hits to Production SLAs | Part 3 of 3 - MinIO (min.io/blog/from-cache-hit...)

[02] Introducing MinIO MemKV: (min.io/blog/introducing-mi...)

[03] MinIO MemKV: Purpose-Built AI Inference Cache Storage (efficientlyconnected.com/minio-memkv-ai-infe...)

[04] MinIO Announces MemKV, Purpose-Built Context Memory Store for (min.io/press/minio-announc...)

[05] KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache - arXiv (arxiv.org/html/2402.02750v2)

[06] Disaggregated LLM Serving with CXL Shared Memory KV Cache at (vtechworks.lib.vt.edu/server/api/core/bit...)

[07] What Is Prefix Caching? Why Repeated Prompts Get Cheaper (spheron.network/blog/prefix-caching...)

[08] KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache (raw.githubusercontent.com/mlresearch/v235/mai...)

[09] Mooncake is the serving platform for Kimi, a leading LLM ... - GitHub (github.com/kvcache-ai/Mooncake)

[10] Mooncake: The KVCache-Centric Architecture Powering Kimi at Scale (themenonlab.blog/blog/mooncake-kvcac...)

[11] Prompt Caching: Stop Paying GPUs to Read the Same Prompt Twice (min.io/blog/prompt-caching...)

[12] NVIDIA GB300 NVL72: Specs, Power, Capacity | ModulEdge (moduledge.com/blog/gb300-nvl72)

[13] Inference Performance · MemKV (docs.min.io/memkv/internals/inf...)

[14] Total Recall: MinIO MemKV Sells GPUs a Petabyte of Shared Context (autonomymag.com/2026/05/12/total-re...)

[15] MinIO MemKV | LLMS3 (llms3.com/node/minio-memkv)

[16] Mooncake: A KVCache-centric Disaggregated Architecture for LLM (zhongzhuzhou.org/blog/2026-05-28-moo...)

[17] Predictive Queue-Informed KV Cache Management for LLM Serving (arxiv.org/html/2607.02525v2)

[18] KV Cache Management: PagedAttention & RadixAttention (analyticsvidhya.com/blog/2026/08/pageda...)

[19] A KVCache-centric Disaggregated Architecture for LLM Serving (researchgate.net/publication/3976459...)

[20] SnapKV - arXiv (arxiv.org/pdf/2404.14469)

[21] KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache - arXiv (arxiv.org/abs/2402.02750)

[22] KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache (alphaxiv.org/abs/2402.02750)

[23] KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache #5492 (github.com/ggml-org/llama.cpp/...)

[24] MinIO Introduces MemKV for Petabyte-Scale AI Inference Memory (storagereview.com/news/minio-introduc...)

[25] mini-llmcache is a pedagogical LMCache reimplementation ... - GitHub (github.com/jianzhnie/mini-llmc...)

[26] Prefill-Decode Disaggregation Meaning | Bitcoin Mining Glossary (d-central.tech/mining-glossary/pre...)

[27] LMCache: Accelerating LLM Inference - Emergent Mind (emergentmind.com/topics/lmcache)

[28] Beyond Prefix Caching! How LMCache Speeds Up RAG by 4.5x By (blog.lmcache.ai/en/2024/10/09/beyon...)

[29] Mooncake: A KVCache-centric Disaggregated Architecture for LLM (arxiv.org/abs/2407.00079)

[30] A KVCache-centric Architecture for Serving LLM Chatbot - USENIX (usenix.org/system/files/fast25...)

[31] Daily Papers - Hugging Face (huggingface.co/papers)

[32] Introducing UALink 200G 1.0 Specification (ualinkconsortium.org/wp-content/uploads/...)

[33] CXL-Enabled KV-Cache Management Beyond GPU Limits - arXiv (arxiv.org/html/2511.00321v1)

[34] Bridging LLM Serving and CXL-SSDs with Chunk-Aware KV Cache (arxiv.org/html/2609.26828)

[35] NVIDIA GB200 Grace Blackwell Superchip — Knowledge Base (yobitel.com/knowledge-base/nvid...)

[36] RadixAttention: AI Inference Definition | InferenceX by SemiAnalysis (inferencex.semianalysis.com/glossary/radix-atte...)

[37] What Is SGLang? Architecture, Performance, and When to Use It (yottalabs.ai/post/what-is-sglang...)

[38] SGLang - High-Throughput LLM Serving Engine - AI/TLDR (ai-tldr.dev/tools/sglang)

[39] SGLang RadixAttention vs vLLM on One H100 - Effloow (effloow.com/articles/sglang-rad...)

[40] KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache (reddit.com/r/LocalLLaMA/commen...)

[41] SnapKV: Cache compression technique for faster LLM generation (medium.com/@techsachin/snapkv-...)

[42] MinIO bổ sung bộ nhớ đệm MemKV quy mô petabyte cho suy luận (csc-jsc.com/blog/minio-adds-pet...)

[43] vLLM direct KV offloading + MemKV · MemKV (docs.min.io/memkv/integrate/vll...)

[44] A KVCache-centric Architecture for Serving LLM Chatbot - Ruoyu Qin (qinruoyu.com/static/media/slides...)

[45] A Photonic-CXL Memory Appliance for Scalable KV Cache ... - arXiv (arxiv.org/abs/2607.27187)

[46] NVIDIA Vera Rubin NVL72 - Gigabyte (gigabyte.com/FileUpload/Global/W...)

[47] Supermicro NVIDIA GB300 NVL72 Rack System | Arc Compute (arccompute.io/solutions/gpu-serve...)

[48] Disaggregated KV Cache System for Sparse Attention LLMs with CXL (arxiv.org/html/2606.19746v1)

[49] A KV Cache Frameworkfor Multi-Turn LLM Serving with CXL-Hybrid (arxiv.org/html/2607.18141v2)

[50] Efficient LLM Generative Inference via CPU-GPU Hybrid Computing (arxiv.org/html/2604.18529v1)

[51] Scale-up Networking for AI: Multi-Protocol Interconnect and Optics (marvell.com/ai-data-center/scal...)

[52] Not All Prefills Are Equal: PPD Disaggregation for Multi-turn LLM (openreview.net/forum)

[53] Open Standards for AI Networking: How UEC, UALink, and OCP (cornelis.com/blog/open-standards...)

[54] NVIDIA GB200 Interconnect Architecture Analysis: NVLink (naddod.com/blog/nvidia-gb200-i...)

[55] Cisco AI Rack for NVIDIA GB300 NVL72 (cisco.com/c/dam/en/us/solutio...)

[56] Inside NVIDIA Blackwell Ultra: The Chip Powering the AI Factory Era (developer.nvidia.com/blog/inside-nvidia-...)

[57] Specifications - UALink Consortium (ualinkconsortium.org/specification)

[58] About UALink (ualinkconsortium.org/about-ualink)

[59] Ultra Accelerator Link™ (UALink™) Consortium Publishes Four (businesswire.com/news/home/202604076...)

[60] Ethernet Standards for Scale-Up AI: An Overview of ESUN, SUE (synopsys.com/articles/ethernet-s...)

[61] Ethernet Standards | AI Clusters | AI Infrastructure - Broadcom Inc. (broadcom.com/topics/what-are-the...)

[62] Scale-Up Ethernet Ecosystem Ramps Up: Cadence ESUN (community.cadence.com/cadence_blogs_8/b/i...)

[63] Ultra Ethernet (UEC) - LLMS3.com (llms3.com/node/ultra-ethernet...)

[64] Intersect360 Research White Paper: - UEC 1.0 (ultraethernet.org/wp-content/uploads/...)

[65] Ultra Ethernet: A Full Guide to AI Scale-Out Fabric After InfiniBand (insidedeeptech.com/ultra-ethernet-uec-...)

[66] ESUN IP Solution for AI Scale-Up Networks | Synopsys (synopsys.com/designware-ip/inter...)

三、下一代内存介质（3D DRAM/eNVM）与存算一体（PIM/CIM）核心信源

[01] Hybrid Memory Cube Market - Size & Forecast - Mordor Intelligence (mordorintelligence.com/industry-reports/hy...)

[02] The Global Memory and Storage Technology Market 2027-2037 (futuremarketsinc.com/memory-and-storage-...)

[03] HBM4 vs HBM3 vs HBM3E: Architecture, Performance, and Real (nevsemi.com/blog/hbm4-vs-hbm3-v...)

[04] China's CXMT beats Western chipmakers to announcement of (tomshardware.com/pc-components/dram/...)

[05] Samsung to adopt hybrid bonding for HBM4 memory | Tom's Hardware (tomshardware.com/pc-components/dram/...)

[06] HBM roadmaps for Micron, Samsung, and SK hynix: To HBM4 and (tomshardware.com/tech-industry/semic...)

[07] JEDEC SPHBM4 Standard Enables HBM4-Class Bandwidth on (techpowerup.com/350725/jedec-sphbm4...)

[08] HBM4E: what Samsung and SK hynix's new AI memory does (rcrtech.com/semiconductor-news/...)

[09] Memory IC Market - Size, Share & Growth - Mordor Intelligence (mordorintelligence.com/industry-reports/me...)

[10] DDR6 Memory Officially Announced: 17,600 MT/s Speed ... - Oscoo (oscoo.com/news/ddr6-memory-of...)

[11] Next Generation Memory Market Size & Growth 2031 (mordorintelligence.com/industry-reports/ne...)

[12] HBM4 compared to HBM4E - Ersa Electronics (ersaelectronics.com/blog/hbm4-hbm4e)

[13] A Comparative Analysis of Samsung Electronics and SK Hynix (k-brainstorm.com/en/hbm4-transition-...)

[14] DDR6 Up To 17.6 Gbps & LPDDR6 Up To 14.4 Gbps (design-reuse.com/news/16147-jedec-co...)

[15] Advances in DRAM and 3D NAND Technologies | PDF | Flash Memory (scribd.com/document/862847433/...)

[16] Beyond NAND, Part 4: What Could Follow DRAM (eyesonflash.com/beyond-nand-part-4)

[17] Resources - Neo Semiconductor | X-Nand (neosemic.com/resources)

[18] HBM3e and HBM4: IC design guide for next-generation high (blogs.sw.siemens.com/semiconductor-packa...)

[19] HBM4 Explained: High Bandwidth Memory for AI Servers (server-hardware.com/knowledge/hbm4-expl...)

[20] JEDEC Officially Releases HBM4 Memory Standard - News (allaboutcircuits.com/news/jedec-official...)

[21] JEDEC publishes SPHBM4 standard with HBM4 bandwidth on (videocardz.com/newz/jedec-publishe...)

[22] JEDEC releases new SPHBM4 standard to slash AI memory costs (tomshardware.com/pc-components/dram/...)

[23] Samsung Electronics Begins Shipment of Industry-First HBM4E (semiconductor.samsung.com/news-events/news/sa...)

[24] (News) The HBM4 Battle Begins! Memory Stacking Challenges (trendforce.com/news/2024/06/07/new...)

[25] AI Semiconductor Packaging Materials & HBM Technologies (usdanalytics.com/our-insights/chemic...)

[26] DDR6 RAM standard to launch in CAMM2 form-factor with speeds (notebookcheck.net/DDR6-RAM-standard-t...)

[27] JEDEC Confirms CAMM2 Memory For Desktop PCs: DDR6 Up To (wccftech.com/jedec-camm2-memory-...)

[28] 中國記憶體里程碑式突破！長鑫不聲不響甩出一張王牌| 科技| 鉅亨號 (hao.cnyes.com/post/269061)

[29] 搜索 (m.expreview.com/exp_new/exp4_search...)

[30] MICRON TECHNOLOGY Aktie kaufen oder nicht? MU Analyse (aktien.guide/aktien/Micron-Techn...)

[31] Born in Finland, Born to ALD: memory - BALD Engineering (blog.baldengineering.com/search/label/memory)

[32] 2025 Technical Program - Squarespace (static1.squarespace.com/static/67a3eee4385d...)

[33] Lam Research Corporation - Facebook (facebook.com/LamResearchCorporat...)

[34] 3D X-DRAM Hits Proof-of-Concept, HBM-Replacement with 10x (wccftech.com/3d-x-dram-hits-proo...)

[35] NEO Semiconductor Demonstrates 3D X-DRAM Proof-of-Concept (prnewswire.com/il/news-releases/ne...)

[36] NEO Semiconductor's revolutionary 3D X-DRAM for AI processors (tomshardware.com/tech-industry/artif...)

[37] X-DRAM TM - Neo Semiconductor (neosemic.com/x-dram)

[38] Home - Neo Semiconductor | X-Nand (neosemic.com)

[39] Next-Generation Memory Technologies: MRAM, RRAM, and PCM (ersaelectronics.com/blog/next-generatio...)

[40] Magnetoresistive RAM Market Size, Share & Forecast 2034 (dimensionmarketresearch.com/report/magnetoresis...)

[41] Embedded MRAM Market Size, Share & Forecast Report 2025 to 2034 (dataintelo.com/report/embedded-mra...)

[42] PCM - STMicroelectronics (st.com/content/st_com/en/a...)

[43] MRAM-Info | MRAM Industry Portal (mram-info.com)

[44] 国产SLC芯片技术突破与应用解析 (bbs.csdn.net/weixin_32533659/art...)

[45] MRAM memory makers (mram-info.com/companies/mram-memo...)

[46] 存算一体芯片，实打实的火了 - 36氪 (m.36kr.com/p/2724653364454664)

[47] ST rolls a new FD-SOI microcontroller with phase-change memory (eejournal.com/article/st-rolls-a-...)

[48] STMicroelectronics FX62A 28 nm FD-SOI ePCRAM Advanced (techinsights.com/blog/stmicroelectro...)

[49] Automotive MCUs in 28nm FD-SOI with ePCM NVM (st.com/content/dam/AME/201...)

[50] ASIL-D automotive-grade microcontroller in 28nm FD-SOI with full (ieeexplore.ieee.org/iel7/10185199/10185...)

[51] FD-SOI - STMicroelectronics (st.com/content/st_com/en/a...)

[52] Resistive RAM Innovations in the Global Market Context (eureka.patsnap.com/report-resistive-ra...)

[53] Advanced ReRAM Market: 28.7% CAGR to 2034? (datainsightsreports.com/reports/advanced-re...)

[54] Earnings call transcript: Weebit Nano Q1 2026 shares fall as cash (investing.com/news/transcripts/ea...)

[55] Weebit Nano (weebit-nano.com/wp-content/uploads/...)

[56] Ferroelectric RAM (FeRAM) Market Size, Share & Growth Report 2035 (snsinsider.com/reports/ferroelectr...)

[57] Progress of emerging non-volatile memory technologies in industry (pmc.ncbi.nlm.nih.gov/articles/PMC11618178)

[58] Magnetoresistive RAM (MRAM) Market Size & Share 2026–2035 (gminsights.com/industry-analysis/m...)

[59] CXL Device & Fabric Buyer's Guide: GA Memory, Switches (stevescargall.com/blog/2025/06/cxl-de...)

[60] Press - Objective Analysis (objective-analysis.com/press)

[61] Practical Memory Disaggregation - SymbioticLab (symbioticlab.org/publications/files/...)

[62] CXL Memory Expansion Explained: 2026 Sourcing - Cosolvic (cosolvic.com/blog/cxl-memory-exp...)

[63] 存算一体是啥新趋势？值得教授学者纷纷下海造芯| 附报告下载 - 量子位 (qbitai.com/2022/08/37183.html)

[64] Processing-in-Memory for Genomics Workloads (computer.org/csdl/magazine/mi/20...)

[65] accelerating long sequence alignment using processing-in-memory (academic.oup.com/bioinformatics/arti...)

[66] 一文看懂"存算一体" - 36氪 (m.36kr.com/p/3423575387360644)

[67] A Survey on AI Datacenter Hardware Architecture - arXiv (arxiv.org/html/2609.26829v1)

[68] DCC: Data-Centric Compilation of Machine Learning Kernels (pure.mpg.de/rest/items/item_373...)
[69] Processing-in-Memory: A System-Level Analysis of DRAM ... - Uplatz (uplatz.com/blog/processing-in-...)
[70] A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2 (researchgate.net/publication/3501722...)
[71] Accelerating Long Sequence Alignment Using Processing-In-Memory (upcommons.upc.edu/bitstreams/d0986303...)
[72] Evaluating Machine Learning Workloads on Memory-Centric (people.inf.ethz.ch/omutlu/pub/MLonUPME...)
[73] Scalability Limitations of Processing-in-Memory using Real System (bu-icsg.github.io/publications/2024/p...)

四、2026–2028全球及中国存力、超节点互连与液冷供应链核心信源

[01] KV Cache-Aware Routing and P/D Disaggregation on Kubernetes (ai.engineer/talks/YXowceUKYJI-k...)
[02] AI Memory Infrastructure | LLMS3 (llms3.com/node/ai-memory-infr...)
[03] Samsung secures long-term agreements for bulk of 2027 DRAM (sammyfans.com/2026/10/19/samsung-...)
[04] 澜起科技-688008-机构调研汇总 - 众问 (appletree.fund/688008-sh)
[05] CXL Memory: Complete Guide to Composable Infrastructure (blog.eduonix.com/2026/08/cxl-memory-...)
[06] DeepSeek V4 Flash Distributed KV Cache: From P/D Direct Transfer (aik8s.run/en/ai-k8s/practices...)
[07] Supporters | InferenceX by SemiAnalysis (inferencex.semianalysis.com/quotes)
[08] OceanStor M900 为华为SuperPoDs 引入 PB 级上下文内存 - Unite.AI (unite.ai/zh-cn/oceanstor-m90...)
[09] Dylan Patel on AI Compute 2026: Claims vs. Evidence - Luminix AI (useluminix.com/reports/industry-an...)
[10] Chip News — Semiconductor, AI Chip & Memory Pricing Wire (memoryindex.io/chip-news)
[11] DRAM_ETF - Reddit (reddit.com/r/DRAM_ETF/hot)
[12] Agentic NetOps — September 6, 2026 - CyberSecurity Institute (cybersecurityinstitute.com/blog)
[13] Astera Labs at Citi's 2026 Global TMT Conference: AI build-out (investing.com/news/transcripts/as...)
[14] Korea's HBM Monopoly, $43T Downstream Exposure, and the Allied (forcedalpha.com/news/korea-hbm-supp...)
[15] 存储半导体月报：担忧之墙是否合理？CXMT技术进展与HBM需求展望 (moomoo.com/hans/news/post/1001...)
[16] 上游为零部件，包括冷板、冷却液、冷却分配单元CDU、接头 (hansenfluid.com/en/data-center/Basi...)
[17] Vera Rubin NVL72 vs GB200 NVL72? Inference TCO & Architecture (inferencex.semianalysis.com/blog/vera-rubin-nvl...)
[18] 天风美女首席孙潇雅已封神？ - 新浪财经 (finance.sina.com.cn/stock/roll/2025-08-...)
[19] Samsung Foundry Allocates Half of 4 nm Capacity for HBM4 Base (techpowerup.com/352432/samsung-foun...)
[20] 優群科技(3217) 深度剖析：DDR5 浪潮與AI伺服器驅動的連接器先鋒 (uc913.com/stock-3217)
[21] 優群2026-08-12 法說會 - BigGo 財經 (finance.biggo.com.tw/quote/3217.TW/earni...)
[22] FMS 2025 CXL Demos - Compute Express Link (computeexpresslink.org/fms-2025-cxl-demos)
[23] The Economics of Exabyte Storage - Solidigm (solidigm.com/products/technology...)
[24] GPU + Object Storage Convergence | LLMS3 (llms3.com/node/gpu-object-sto...)
[25] Scality's AI Inferencing Factory accelerates inferencing (m.storagesserver.com/news/scality-s-ai-i...)
[26] 1.6T Optical For Ai Data Centers: The Complete Technical Guide (vitextech.com/blogs/blog/1-6t-opt...)
[27] AI Data Center Networking - NextGenInfra Insights (nextgeninfra.io/topics/ai-data-cent...)
[28] NVLink, NVSwitch, and all that - Doubleword (blog.doubleword.ai/nvlink-scale-up)
[29] Earnings Call Highlights: Astera Labs (ALAB) | Q3 FY25 (simpletechtrend.com/en/post/astera-labs...)
[30] 华为发布全球首个采用NPO的超节点——昇腾960超节点 - 开源中国 (oschina.net/news/502544)
[31] KV Cache Overview: From Attention Principles to Kubernetes (aik8s.run/en/ai-k8s/inference...)
[32] Multi-Machine and Disaggregated LLM Inference (aik8s.run/en/ai-k8s/inference...)
[33] RBG Multi-Role Inference Orchestration: From CPU Control Plane to (aik8s.run/en/ai-k8s/practices...)
[34] 华为推出OceanStor M900 AI记忆存储——与昇腾超节点匹配 (chnfund.com/article/ARaa969a54-...)
[35] VAST Data - The S3 & Object Storage Ecosystem | LLMS3 (llms3.com/node/vast-data)
[36] NVIDIA Vera BlueField-4 STX - Glenn K. Lockwood (glennklockwood.com/garden/vera-bluefie...)
[37] Apple Weighs Return to Commercial Servers as Accelerator (finance.biggo.com/news/33fb0418-4cde-...)
[38] What Is the "Zero-Sum Game" of Memory Chips and Why Does (exceldisc.com/blog/what-is-the-ze...)
[39] October 2026 IT Hardware Market Update: DDR5 Memory Crunch (servermonkey.com/blog/october-2026-e...)
[40] Micron Secures AI Future with $1.8 Billion Acquisition of PSMC's P5 (markets.financialcontent.com/stocks/article/toke...)
[41] Samsung said to use TSMC tech for Nvidia's next-generation (tradingview.com/news/seekingalpha:7...)
[42] Nvidia reportedly wants all its HBM suppliers to use TSMC's 3nm (thenextweb.com/news/nvidia-nvhbm-s...)
[43] (News) NVIDIA Reportedly Pushes TSMC 3nm NVHBM Base Dies (trendforce.com/news/2026/10/19/new...)
[44] Scaling Up - MonolithIC 3D Inc. (monolithic3d.com/uploads/6/0/5/5/605...)
[45] Astera Labs targets $25B AI connectivity TAM as Scorpio and (seekingalpha.com/news/4549928-astera...)
[46] Micron says HBM uses about three times the wafer supply of DDR5 (mlq.ai/news/micron-says-hb...)
[47] CAMM2 Connector (Compression Attached Memory Module) (youtube.com/watch)
[48] 【優群法說會重點內容備忘錄】未來展望趨勢 20260812 - 富果直送 (blog.fugle.tw/post/earnings-call-...)
[49] 優群(3217)可以買嗎? 股價分析：EPS、配息、法人籌碼 - FinLab (finlab.finance/stocks/3217)
[50] 澜起科技** 缘何暴涨？ - 财富号 (caifuhao.eastmoney.com/news/20260922095649...)
[51] 澜起科技，全球第一！ - 搜狐时间线 (timeline.sohu.com/news/cQD3xouvGB)
[52] 澜起科技的“惊魂三日”-钛媒体官方网站 (tmtpost.com/8083567.html)
[53] 公司公告_澜起科技：2026年半年度报告新浪财经 (money.finance.sina.com.cn/corp/view/vCB_AllBu...)
[54] 澜起科技(688008)2026年一季报点评：2026Q1归母净利润同比+61.3 (zx.upchina.com/report/details/7000...)
[55] Montage Technology Co., Ltd. 澜起科技股份有限公司截至2026年6月 (hkexnews.hk/listedco/listconews...)
[56] 澜起科技(688008) 盈利预测_F10_同花顺金融服务网 (basic.10jqka.com.cn/688008/worth.html)
[57] Astera Labs (ALAB) Earnings Call Transcripts - Stock Analysis (stockanalysis.com/stocks/alab/transcr...)
[58] Astera Labs, Inc. Common Stock - Perplexity (perplexity.ai/finance/ALAB/earnings)
[59] Astera Labs (ALAB) Q2 2026 Earnings Call Transcript (fool.com/earnings/call-trans...)
[60] Q3 Revenue Guidance of $550M Implies 40% Sequential Leap (finance.biggo.com/news/US_ALAB_2026-0...)
[61] Exploring CXL Use Cases with CXL 3.X Switches for AI Applications (computeexpresslink.org/blog/exploring-cxl-...)
[62] News Posts matching 'hybrid bonding' | TechPowerUp (techpowerup.com/news-tags/hybrid%20...)
[63] Hybrid Bonding: Has the Inflection Point Arrived? - 36氪 (eu.36kr.com/en/p/3756077063209481)
[64] Decisive Battle in Hybrid Bonding: Key Insights and Trends - 36氪 (eu.36kr.com/en/p/3407854283296132)
[65] Petabyte+ storage server recommendations - Reddit (reddit.com/r/storage/comments/...)
[66] 华为推出“OceanStor M900” SSD读写寿命提升16倍 - 凤凰网 (i.ifeng.com/c/8wUYW9u2T71)
[67] Technology | LLMS3 (llms3.com/category/technology)
[68] Why GPU Networking Is AI's Next Bottleneck | EM360Tech (em360tech.com/tech-articles/ai-in...)
[69] In the AI Era, Ethernet Set to Surge in Scale-Out and Ramp (650group.com/blog/in-the-ai-era-...)
[70] AI Back-End Switch Market Will Push Past $100 Billion by 2030 (delloro.com/news/ai-back-end-sw...)
[71] NVLink 5 vs NVLink 6: Bandwidth, Domains and Cluster Design (gpusmith.com/articles/en/nvlink-...)

[72] The Complete Guide to the Next Speed Tier for AI Data Centers (thenetworkdna.com/2026/10/16t-etherne...)

[73] Marvell's CPO Story Unfolds at its Industry Analyst Meeting (650group.com/blog/marvells-cpo-s...)

[74] 结合Deepseek代码探讨MLA的改进及收益 - Hugging Face (huggingface.co/blog/Junrulu/mla-co...)

[75] 美媒稱中國長鑫存儲在先進存儲芯片領域取得突破 - 香港新聞網 (hkcna.hk/hk/content/s6aa2892...)

[76] 面向DeepSeek-V4 的FlashMemory：长上下文KV Cache 如何压到约 (developer.cloud.tencent.com/article/2689117)

[77] Article timeline - AI/LLM on Kubernetes Infrastructure (aik8s.run/en/timeline.html)

[78] 华为追赶英伟达：把单卡之争变成一场集群战 - 通信人家园 (txrjy.com/thread-1439687-1-1....)

[79] 聊聊华为的Atlas 950超节点 - 鲜枣课堂 (xzclass.com)

[80] 华为推出OceanStor M900 AI记忆存储 - 凤凰网 (i.ifeng.com/c/8wUYTL4pw3e)

[81] 华为副董事长汪涛：发布全球首个NPO技术超节点昇腾960 - 凤凰网科技 (tech.ifeng.com/c/8wUIUiGtYaK)

[82] 614期2026年度长鑫存储（CXMT）深度战略调研与投资采购指南 (bilibili.com/opus/1185068566074 8...)

[83] 长鑫HBM3样品通过验证,2027年大规模投产 (huizhan086.com.cn/zhxw/282.html)

[84] 长鑫存储拟2027年量产12层HBM3E 与全球存储巨头技术差距或缩短 (i.ifeng.com/c/8tUXW1lKuWo)

[85] 不只搞内存，长鑫存储被曝将进军闪存 - 观察者网 (guancha.cn/CaiJing/2026_09_18_...)

ABOUT YAVIN 质观

Yavin质观

硬科技与资本叙事研究者，长期关注技术、产业、政策与资本化逻辑，
书写技术时代、现代生活与精神处境中的缓慢观察。

微信扫码关注「Yavin质观」公众号

获取更多硬科技、产业、政策与资本化逻辑深度研究报告

研究声明与免责条款：本报告由「Yavin质观」独立研究编制，报告所引用的财务数据、市场出货量、工程技术参数及供应链信息均来源于公开财报、招股说明书、国际标准组织（JEDEC、PCI-SIG、CXL、UALink）、第三方研究机构（TrendForce、IDC、SemiAnalysis、Yole等）、学术论文及行业公开信息。针对非上市企业或未公开披露之产线良率、成本模型与情景推演，均属于研究假设与工程估算，仅供参考，在任何情况下均不构成对任何证券、芯片产品或投资行为的建议。未经许可，不得用于商业篡改或二次售卖。

© 2026 Yavin质观 · 保留所有权利