



【中泰电子】存储： AI推理带来需求爆发、驱动范式升级， 周期能见度大幅拉长

分析师：

王芳 S0740521120002，杨旭 S0740521120001，康丽侠S0740525040001

中泰证券研究所
专业 | 领先 | 深度 | 诚信

- **存储是AI推理的核心瓶颈，驱动存储需求爆发、存储范式改进。**
- LLM 推理的解码阶段本质是memory-bound，核心存储负载包括：模型权重、KV Cache、激活值、RAG 向量库等。相较于模型权重等静态张量数据，KV Cache是随上下文长度和并发数动态膨胀的张量数据，推理性能（TTFT / TPS）高度依赖对KV Cache的保存和对KV Cache的管理效率。在传统冯·诺依曼架构下，大模型推理时的大量高维张量数据、Transformer的自注意力机制均加剧了内存墙问题，数据量巨大、搬运成本高，严重拖慢推理效率，存储使访存带宽与延迟逐步成为制约系统吞吐与响应性能的核心瓶颈，因此黄仁勋一直在说：“GPU 大部分时间都在等数据，而不是在计算”，“计算能力增长远快于内存带宽，GPU 经常处于饥饿状态（starving for data）。”而提升存储带宽和容量可以显著增强推理性能、降低推理成本，“以存代算”是必然趋势。
- 随着模型越来越大、上下文越来越长、使用人数增加等，AI推理带来HBM、DRAM、SSD、HDD的需求全面爆发，同时面对大模型推理的访存受限问题，产业界也在推进存储器性能升级和存储层级优化，存储从单一层级向高带宽+大容量+分级管理的协同架构演进，存储与计算的关系也由传统解耦逐步走向协同优化。1) 高带宽存储器解决方案：包括HBM、WOW 3D堆叠DRAM、HBF。2) 优化存储分级管理系统：CXL内存池化技术，Prefill和Decode阶段的分机柜部署（英伟达GTC2026推出的最新方案）。
- **在前期价格涨幅大、行业看接下来2年供需紧张的情况下，对合约价的跟踪会从“从价格还能涨多少”转向“价格高位保持多久”，预计合约价26年全年上涨，27年至少保持高位，客户与原厂签多年长协拉长周期。**
- 现货价占市场10%左右：近期部分现货市场价格有回调10-20%左右，主要系存储模组的现货价与合约价价差大，贸易商获利了结心理强。
- 合约价占市场90%左右：目前北美CSP因担心拿不到产能，陆续与原厂签订3-5年长约，国内模组厂也有签，客户与原厂基本达成价格共识，周期能见度拉长。此前周期中合约价上涨3-4个季度就会回落，主要系由消费电子库存周期主导，本轮周期由AI需求驱动，服务器占存储的敞口达到50%-60%，行业供需高度紧张，预计26年全年价格上涨，但是逐季价格收敛，预计27年价格也保持高位。
- **存储是AI硬件板块中短期业绩确定性最强，供需在可见的2年内持续紧张，同时估值中枢有提升的潜在可能性。**
- 服务器占存储敞口提升到50%-60%，存储的容量和性能是AI推理的核心瓶颈环节，同时是AI硬件业绩确定性中短期最强、估值最低的方向。
- **投资建议：**建议关注：1) 弹性模组及主控：德明利、江波龙、佰维存储、大普微、联芸科技等；2) 兆易创新、普冉股份、东芯股份、北京君正、澜起科技、聚辰股份、恒烁股份等；3) 设备：微导纳米、拓荆科技、中微公司、精智达、华海清科、中科飞测、京仪装备、骄成超声、百傲化学、北方华创等。4) 光刻机产业链：茂莱光学、汇成真空、波长光电、阿石创、联合化学、富创精密、永新光学等。
- **风险提示：**1) 长鑫长存产能释放加剧竞争的风险。2) AI CAPEX不及预期的风险。3) 数据更新不及时，模型测算偏差风险。

目录

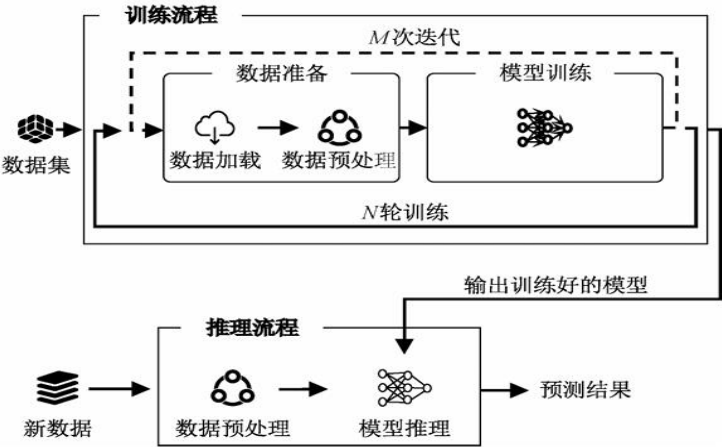
一、AI推理带来存储需求爆发和存储范式的改进

二、看未来2年供需持续紧张，原厂与客户签订长协

三、存储是AI硬件估值最低、业绩确定性最强的方向

- **大模型训练和推理对存储需求存在区别：**训练：基于提前备好的海量静态数据集，数据总量可控、规则、可预测，偏向一次性学习过程，但由于数据集规模大、计算密集，单任务维度下显存消耗、整体存储容量需求远大于推理；推理：数据实时输入、请求粒度小、并发高、上下文长度差异大、延迟要求严格，偏向持续性的应用过程，为避免重复计算需保留大量KV cache，其在每一次token生成都会访问、对延迟高度敏感、会随序列长度动态扩展，是显存占用和带宽消耗的核心因素，因此在智能体AI普及带来用户爆发式增长、参数规模扩大、应用复杂度提升的背景下，数据会快速动态膨胀，预计远期在数据中心中，推理存储需求占比（2030年预计70%+）远超训练。
- **大模型训练：**1) 运行机制：计算密集型工作，系统需反复读写、写入巨量数据，数据流动频率高、负载持续、IO密度高，但是训练阶段数据集通常是固定规模，不随时间线性增长。2) 存储介质：训练样本通常存在HDD/SSD，模型参数、激活值等核心计算内容的加载与处理在HBM/GDDR（GPU 显存），DRAM用于扩展内存、保存部分模型权重，SSD作为补充设备，用于保存中间文件（例如检查点文件、超出HBM/DRAM的数据、不活跃Token）。
- **大模型推理：**1) 运行机制：存储需求由“规模扩张”与“动态波动”共同驱动。随大模型参数规模扩大以及应用复杂度提升，推理侧存储需求呈现显著非线性增长特征。2) 存储介质：训练好的模型首先存储在SSD，推理时从SSD加载至DRAM，再从DRAM加载至HBM（用户输入query及生成token相关计算），HBM的KV cache亦持续更新支持实时推理，若上下文过长导致DRAM无法容纳，继续缓存至SSD，推理结束后，完整的Session数据、用户日志、输入输出等数据在HDD/SSD长期储存。

图表：大模型训练和推理流程



图表：大模型不同阶段对存储介质的需求

存储层级	存储介质	典型带宽 / 延迟	单节点容量范围	典型用途	训练使用	推理使用	经典应用场景	部署位置
热层	HBM / GDDR (GPU 显存)	2-4 TB/s / <1μs	80-192GB/GPU	模型参数、KV Cache、激活值	核心使用	核心使用	Transformer 层计算、注意力矩阵缓存	GPU 板卡内
热层	DDR5 / MRDIMM (系统内存)	200-800 GB/s / ~100ns	512GB-4TB / 节点	中间状态、微批数据、缓存热权重	使用	使用	CPU 预处理、KV Cache 扩展	GPU 主机或 CPU 节点内
温层	NVMe SSD (PCIe 4/5)	5-14 GB/s / 10-100μs	8-64TB / 节点	模型权重加载、Embedding 索引、Session 缓存	高频使用	高频使用	模型权重存取、推理上下文缓存	AI 推理 / 训练服务器本地
冷层	HDD (SATA/SAS)	200 MB/s / 5-10ms	10-20TB / 盘	日志、归档、语料原始数据、历史模型；仅用于离线阶段 / 冷备份	仅用于离线阶段	不使用	训练语料原始存储、日志归档、备份	独立存储服务或对象存储集群 (NAS/OSS/S3 冷层)

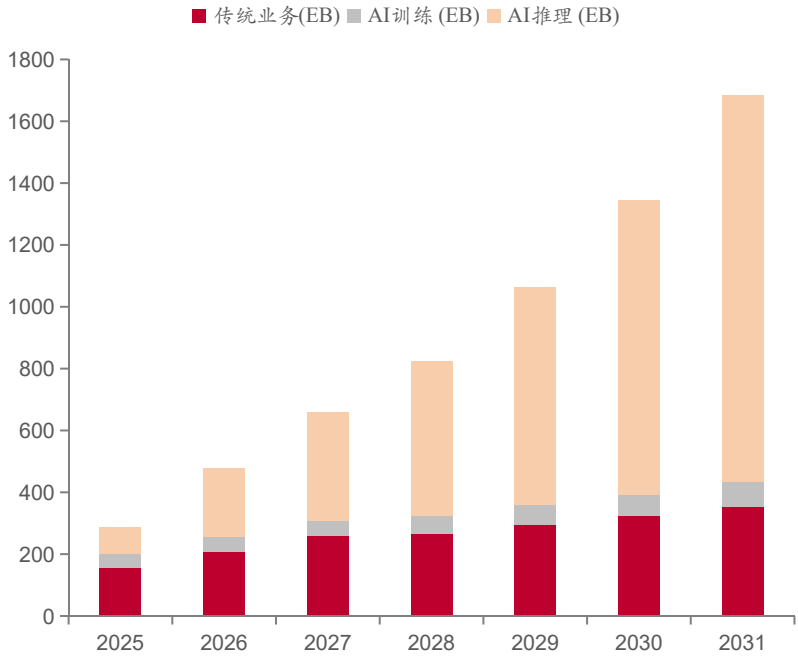
1.1 AI推理带来存储需求爆发

- 从单模型/单任务维度对比，训练对存储容量与带宽需求远高于推理：训练的计算需求、显存消耗需求、存储容量消耗需求、带宽需求均高于推理，尤其是存储容量需求约推理100-1000倍，原因是训练需要存参数、梯度、优化器状态、激活值、Checkpoint、数据集，而推理通常只需要存参数、KV Cache、中间缓存等。
- 但从AI平台维度对比，推理数据将会动态膨胀，在数据中心的存储需求占比预计快速提升，未来预计远超训练：AI从“生成式”向“智能体AI”迈进，不仅拓宽应用场景、提升普及度，使用户数爆发式增长，还对上下文记忆能力、自主性、规划能力、持续学习能力都提出更高要求，此背景下KV Cache规模快速膨胀，多并发请求放大实时显存占用，同时模型权重、向量数据库及推理过程中的中间数据（如生成token缓存）推动整体存储容量需求持续攀升，因此推理对存储的需求正急剧增长，未来预计远超训练。

图表：大模型训练和推理对存储需求的区别

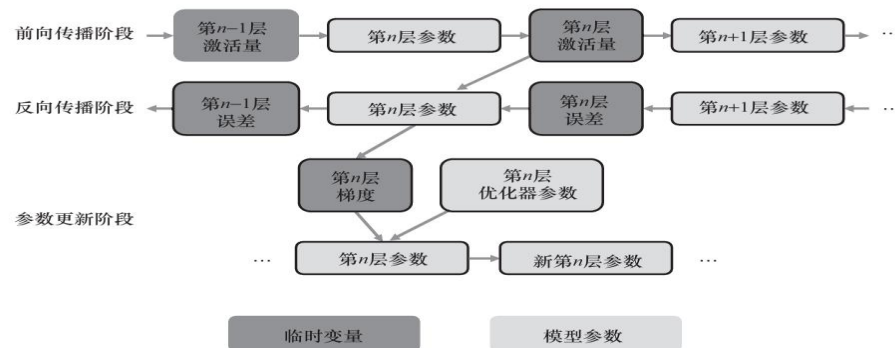
	训练	推理
特征	计算密集型，计算量约推理2-3倍	延迟敏感型
计算流程	前向传播、反向传播	仅前向传播
数据存储类型	基础训练数据、模型参数、激活值、梯度数据、优化器、检查点等	模型权重、KV cache、激活值、RAG向量数据库、Token的长期落盘等
显存需求消耗	消耗大：模型参数、梯度、优化器、激活值都与显存消耗，优化器、激活值是消耗核心	消耗相对训练小：KV Cache是显存消耗的核心因素，压力取决于Batch Size（模型一次性处理的样本数量）、上下文长度等
存储容量需求	极大，训练对存储总需求约推理100-1000倍：需保存海量训练数据、大量检查点，通常需TB-PB级训练数据、TB级Checkpoint。	中等：通常需要GB-TB级存储（涉及大量RAG/上传自有文档等场景除外），仅模型参数+KV Cache。
带宽需求	极高：因为训练计算密集，数据搬运量大	中高
延迟需求	容忍较高延迟	追求低延迟，因为通常是实时或近实时任务

图表：AI重心已从部署训练转向推理，未来数据中心的推理存储需求预计超过训练



- 大模型训练需要存储海量数据集、大量权重、激活值、梯度、状态等，其中激活值、优化器内存消耗最大。
- 一、前向传播阶段：训练数据依次通过模型各层计算并产生激活值。
 - 基础训练数据（原始输入）：即用于模型学习的原始训练语料（文本、图像等多模态数据），是前向传播的核心输入，属于静态数据，总量固定，需提前存储在本地或网络存储（HDD/SSD）中，供模型读取计算。
 - 模型参数（初始权重）：模型中可学习参数的总数量，属于固定基础数据，是前向计算的核心依据，近年参数量爆发式增长，由亿级增长至万亿级。
 - 激活值：前向传播过程中，模型各层运算经过激活函数处理后生成该层输出，即激活值，需临时保存，用于后续反向传播计算梯度。激活值显存消耗大，其显存占用通常与批量大小（batch size）、序列长度（sequence length）、模型层数等因素密切相关。
- 二、反向传播阶段：逐层反向计算误差梯度，梯度计算完毕后激活值被释放，梯度数据保留用于参数更新。
 - 误差数据：前向传播结束后，计算结果与训练标签的偏差值（损失值），用于指导梯度计算的方向，但需临时存储至梯度计算完成。
 - 梯度数据：误差以与前向传播相反方向，与各层参数、前向传播的激活值进行计算，用于衡量参数调整的方向和幅度，存储体量与模型参数规模正相关。
- 三、参数更新阶段：优化器根据保存梯度及自身参数更新模型参数，更新完毕后上版本模型参数、梯度数据、优化器参数释放。
 - 优化器：根据保存的梯度及自身参数（如学习率、动量等）更新模型的参数。为了让模型参数更新时能尽可能逼近最优值，许多不同的优化器被提出，如SGD和Adam，这些优化器需要保存额外的信息。
 - 更新后的模型参数（新权重）：替换原有的初始权重，成为后续训练迭代或推理的基础，需存储至模型下一次迭代更新。
- 四、全流程阶段。
 - 检查点（checkpoint）：训练中某个特定时点保存的模型快照，包括模型参数、优化器状态、梯度数据、训练元数据等，保障训练意外中断后能恢复进度。检查点数据主要保存在SSD中，避免丢失，DRAM负责加载检查点数据至内存，再迁入HBM，缩短数据恢复时间。

图表：大模型训练中各阶段的数据依赖



1.1 AI推理带来存储需求爆发

图表：大模型训练时的主要数据

存储对象	数据类型和定义	训练时对应的主要阶段	物理存储介质
基础训练数据（原始输入）	前向传播的核心输入，属于静态数据。	前向传播	提前存储在HDD/SSD，根据需求加载至CPU的DRAM中，转换为Batches，再由GPU显存计算。
模型参数（初始权重）	模型中可学习参数的总数量，属于固定基础数据。	前向传播	训练时加载至HBM/GDDR（GPU 显存）处理，超出显存容量时借助DRAM扩展，训练中断/检查点文件/超出DRAM容量限制的数据存储至NVMe SSD存储。
激活值	模型各层运算经过激活函数处理后生成输出值。	前向传播	训练时加载至HBM/GDDR（GPU 显存）处理，超出显存容量时借助DRAM扩展，训练中断/检查点文件/超出DRAM容量限制的数据存储至NVMe SSD存储。
梯度数据	误差以与前向传播相反方向，与各层参数、前向传播的激活值进行计算，用于衡量参数调整的方向和幅度。	反向传播	训练时加载至HBM/GDDR（GPU 显存）处理，超出显存容量时借助DRAM扩展，训练中断/检查点文件/超出DRAM容量限制的数据存储至NVMe SSD存储。
优化器	根据保存的梯度及自身参数（如学习率、动量等）更新模型的参数。	参数更新	训练时加载至HBM/GDDR（GPU 显存）处理，超出显存容量时借助DRAM扩展，训练中断/检查点文件/超出DRAM容量限制的数据存储至NVMe SSD存储。
检查点（checkpoint）	训练中某个特定时点保存的模型快照，包括模型参数、优化器状态、梯度数据、训练元数据等，保障训练意外中断后能恢复进度。	全阶段	SSD负责保存训练中断的 checkpoint（含模型参数、优化器状态等数据），支持断点续训；DRAM负责快速加载 checkpoint 到内存，再迁入 HBM，缩短恢复时间。

- 在推理过程中，占据主要存储资源的数据类型包括：模型权重、KV Cache、激活值、RAG向量数据库以及Token长期落盘数据。其中，前四类主要以高维张量形式存在，构成推理阶段的核心存储负载，并主要驻留于HBM与DRAM等高速存储层级。
- 1) 模型权重（静态张量数据）：模型权重为预训练及微调阶段生成的参数矩阵集合，其规模由参数量与量化精度共同决定。例如，INT8精度下7B模型约占用7GB存储空间。推理过程中，权重通常常驻于HBM以满足高带宽访问需求，在显存受限场景下，部分权重可分层卸载至DRAM或通过分页/流式加载机制调度。
- 2) KV Cache（动态张量数据，核心增量负载）：KV Cache用于缓存历史Token的Key与Value，以避免自回归解码过程中重复计算注意力，从而显著降低算力开销，其本质是随序列长度动态扩展的高维张量，KV Cache是推理侧显存占用与带宽消耗的核心决定因素。
 - 每个Token的KV Cache占用空间约为： $2 \times \text{模型层数} \times \text{隐藏层维度} \times \text{精度字节数}$ ，总KV Cache规模随Batch Size、上下文长度（Sequence Length）及并发请求数线性增长，并在系统层面呈现出显著的动态膨胀的特点。
 - 在Prefill阶段：以批量写入为主，快速构建KV Cache；在Decode阶段：以高频读取为主，并随生成过程持续追加写入。
 - （注：Batch size，模型一次性处理的样本数量（如句子或文档数），增大Batch Size可利用GPU并行能力提升速度，但会占用更多显存，且过大会导致收敛变慢）
- 3) 激活值：（张量数据，瞬时数据）：激活值为前向计算过程中产生的中间结果，生命周期极短，主要在SRAM与HBM之间流转。
- 4) RAG向量数据库（外部数据库）：RAG引入外部知识库，通常以向量形式存储于SSD/HDD。在高并发低延迟场景下，其索引结构（如ANN Index）通常需常驻于CPU DRAM，以降低检索延迟；部分热点数据亦可能被加载至HBM参与加速。
- 5) Token的长期落盘（文字符）：用户输入与模型输出的Token通常以日志或数据流形式持久化至SSD/HDD。其底层表现为离散整数序列或字符串，相较于高维张量数据，其存储体量与成本占比极低，对整体存储架构影响有限。

图表：Transformer大模型推理时的主要数据

存储对象	状态与读写模式	物理存储介质	底层数据结构	生命周期与持久性	数据量级	架构侧核心影响
模型权重	在训练时生成，在推理时属于静态数据，初始化加载后只读不写	HBM为主 (注：在极限单卡或多租户海量 LoRA 场景下，不活跃权重会降级至 CPU DRAM 或 NVMe SSD)	矩阵：高维度的密集浮点张量矩阵 ((如 FP16, FP8, 或 INT4 量化格式)	静态数据，全局持久，只要模型服务 (Serving) 在线，即常驻内存或显存阵列中。	十 GB 到 千 GB 级 (受参数量和数据精度影响)	决定了推理集群即最低显存容量底线。
KV Cache	动态膨胀，随生成步数高频追加写入并伴随极高频读取	分层级联调度： 1. 活跃态：锁定在 HBM。 2. 排队/降级态：卸载到 DRAM。 3. 全局/休眠态：压缩后写入企业级 NVMe SSD (用于 Agent 长期休眠或全局共享存储池)。	矩阵：键值对浮点张量 ((Key/Value Tensors, 常通过 PagedAttention 机制切分为非连续的 Blocks, 并辅以 8-bit/4-bit 量化)	动态数据，会升级到全局级 (普通请求随会话结束释放; Agent 记忆或公共系统 Prompt 可持续保留数小时至数天)	几 GB 到 甚至 PB 级 与并发数和序列长度正相关	决定了最大并发吞吐量上限”与“最大上下文窗口长度”。
激活值	极度动态与瞬态，随前向传播极高频生成与擦除	GPU 片上缓存 (SRAM) 与 GPU 显存 (HBM) 之间高速流转	矩阵：瞬时特征张量矩阵 / 隐藏状态计算中间值 (Hidden States)	瞬时数据，微秒/毫秒级瞬间，仅在数据穿透当前网络层的一刹那存在，计算完毕立即覆写或释放。	几 MB 到 瞬时几 GB ((Decode阶段占用极微; 但在长文本 Prefill 阶段会产生较大的显存峰值)	是导致长文本处理时发生“突发性显存溢出 (OOM 崩溃)”的最核心因素。
RAG 向量数据库	缓变静态为主，支持低频的增删改查与批量异步更新	独立存储集群，依托于高性能 NVMe SSD、HDD 等，部分存储在 DRAM	矩阵：稠密数学向量 (Dense Vectors, 如 1024 维) + 纯文本切片 (Text Chunks) + 元数据标签 (Metadata JSON)。	外部数据库，永久持续，作为企业的独立外部知识库，跨越单次模型推理的生命周期。	TB 到 PB 级 (体量由企业私有域文档语料库的总量决定)	决定了外部知识库的检索延迟与网络 I/O 瓶颈，是突破模型“静态知识盲区”的外置数据库。
Token 长期落盘	高频顺序追加	HDD、SSD	文字符/字符串：离散且极简的整数数组 (Token IDs) 或反序列化后的纯文本字符串日志。	永久持续，离线持久化存储，供后续数据挖掘或训练使用。	单次请求仅需 几 KB, 但是随系统长期运转, 日积月累可达 GB/TB 级流水数据	

■ AI推理驱动存储需求指数级爆发。

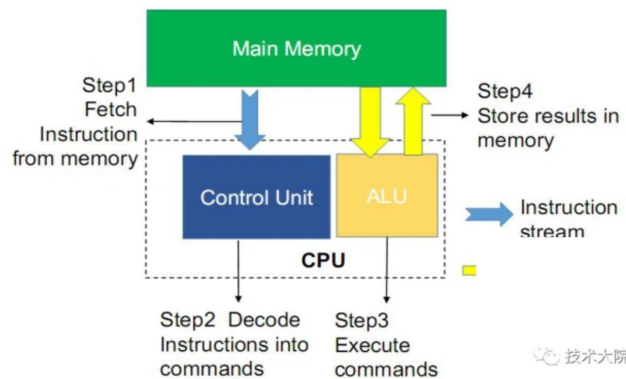
- 1) KV Cache多层缓存成为推理系统“标配”，带来存储需求全面爆发。
- 定位分工：HBM 成本高、延迟低，承担热点计算与高频访问；DRAM承接层级缓存与中等频度访问；SSD作为成本/容量折中层，承接冷数据与长周期缓存/索引。
- 工程逻辑：在大体量查询与长上下文背景下，系统优先复用 Prefill 阶段的 KV cache，以降低 Decode 计算与端到端时延。当再次遇到相似问题时，可直接调用已缓存 KV，无需重复计算，整体算力成本更优。
- 随着“缓存保留时长”与“并发度”提升，热数据上收至 HBM、冷数据下沉至 DRAM/SSD 的比重上升，带动DRAM 与 SSD配置同步放大。当前海外大型互联网公司已在基础设施侧普遍采用 HBM+DRAM+SSD的KV Cache多层缓存方案。
- 2) 对话范式升级：从模型自答到思维链展开、与外部工具/Agent 联动，Token消耗量明显提升。
- 范式切换：2024 年以前，主流对话以模型自答为主，外部检索与数据库调用有限；2025 年起，链式推理（CoT）与工具调用/多 Agent 协作渗透率提升，token 用量与外部数据访问显著增加。
- 量纲变化：模型在理解与拆解问题后，还需跨检索/地图/支付/本地生态等多环节交互，产生二次与多次数据读写。单次复杂任务的 token 消耗从千级提升至万级，存在10 倍量级的上行空间。
- 多环节协作引致的中间态与历史态需要更长时间的可追溯与更低成本的快速载入，强化了 DRAM/SSD 对中低温数据的承接作用。
- 3) 媒介升级：从文本到多模态，存储需求进一步提升。
- 过去以文本为主，当前多模态（图、音、视频）快速普及，视频生成/理解成为重点方向。
- 工程结论：多模态（尤其视频）在推理端的时空 token 密度更高，需要更大的活跃窗口与更频繁的分页/换入。由此带来 HBM 的峰值压力与DRAM/SSD 的持续扩容；SSD 在承接冷段缓存、检索索引与长周期知识库方面弹性最强。

图表：存储的分层

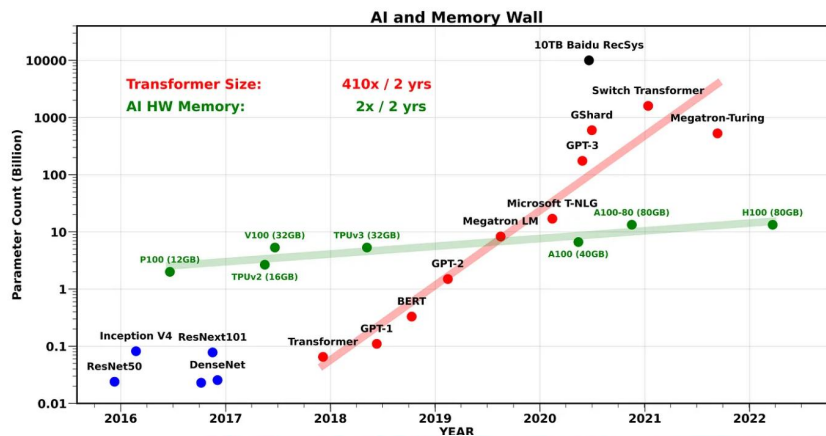
存储层级	在大语言模型（LLM）推理中存储什么？	在AI视频生成推理中存储什么？
HBM(GPU显存)	模型权重 + 活跃的KV Cache。这是推理的主战场，所有计算都在此进行。KV Cache用于避免重复计算注意力，是支持流式输出的关键。当KV Cache超过HBM容量时，非活跃部分会被卸载至DRAM或SSD。	正在执行计算的模型参数 + 当前帧的中间特征张量（激活值）。例如，生成视频时，当前正在渲染的帧及其相关神经网络层的输出会驻留在HBM中。部分模型参数（如不活跃层）可根据策略卸载。
DRAM(主机内存)	非活跃或溢出的KV Cache（如用户会话上下文）、准备被GPU加载的模型分块。作为HBM的溢出池，通过类似LLM或英伟达Dynamo引擎的技术，存放暂时不用但很快可能需要的中间数据，避免重新计算。	待处理的视频帧数据。从HBM卸载的部分模型参数或中间特征。当HBM容量不足时一些计算量较小的组件会被换出到此，充当高速缓存。
SSD(高速本地盘)	所有模型权重的本地副本、被换出的KV Cache（当DRAM也不足时）。作为“温存储”，保障模型和部分中间数据能快速加载至DRAM和HBM。也用于存储短期的对话日志和推理结果NVMe-of等高速访问协议。	视频素材库、模型检查点（Checkpoint）、近期生成的视频成品。从DRAM进一步溢出的中间数据。充当高速缓存，存放生成过程中所需的热门素材和组件，支持
HDD(硬盘/对象存储)	模型归档副本、长期日志/审计数据、训练数据集。作为持久化、低成本的“冷存储”或“数据湖”，用于备份和归档，几乎不会参与日常高频推理。通常不存储KV Cache。	海量的原始视频素材、长期归档的视频成品、训练用的数据集。存放访问频率极低但需要长期保留的巨量冷数据。

- 内存墙问题在大模型时代更为严重，在数据规模快速增长的同时，系统瓶颈正由单纯的“容量不足”转向“带宽与数据搬运效率受限”，高频的数据在计算单元与多层存储之间迁移，使内存墙问题凸显。
- 在传统计算机体系中，基于冯·诺依曼架构，计算单元与存储系统在物理上分离，数据在处理器与存储之间频繁搬运，形成典型的“内存墙”瓶颈。在大模型训练与推理场景中，以高维矩阵运算为主的数据密集型负载显著放大了这一问题，使访存带宽与延迟逐步成为制约系统吞吐与响应性能的核心瓶颈，因此黄仁勋一直在说：“GPU 大部分时间都在等数据，而不是在计算。”、“计算能力增长远快于内存带宽，GPU 经常处于饥饿状态（starving for data）。”
- 从需求结构来看，训练端与推理端对存储体系的压力呈现出显著差异：
- 1) 训练侧：随着模型参数规模的持续提升，单卡HBM容量及集群级总显存需求同步增长，同时还需承载优化器状态（如Adam的动量与方差项）及中间激活值（activation）。整体来看，其存储需求增长路径相对静态、线性且可预测，主要依赖于模型规模与并行策略（如数据并行、张量并行、流水并行）的设计。
- 2) 推理侧：存储需求呈现出更强的动态性与非线性特征。一方面，随着模型规模持续扩大，模型权重本身已占据显著存储空间，在推理过程中需在计算单元与显存之间频繁调度，带来较高的数据搬运开销；另一方面，超长上下文（如百万token级）显著放大KV Cache占用，自回归生成过程中KV Cache持续累积，叠加多并发请求下的动态调度，使显存压力呈现出明显的随机波动。此外，RAG（Retrieval-Augmented Generation）引入外部知识库访问，进一步增加了对存储容量与访存带宽的要求。

图表：存算分离架构



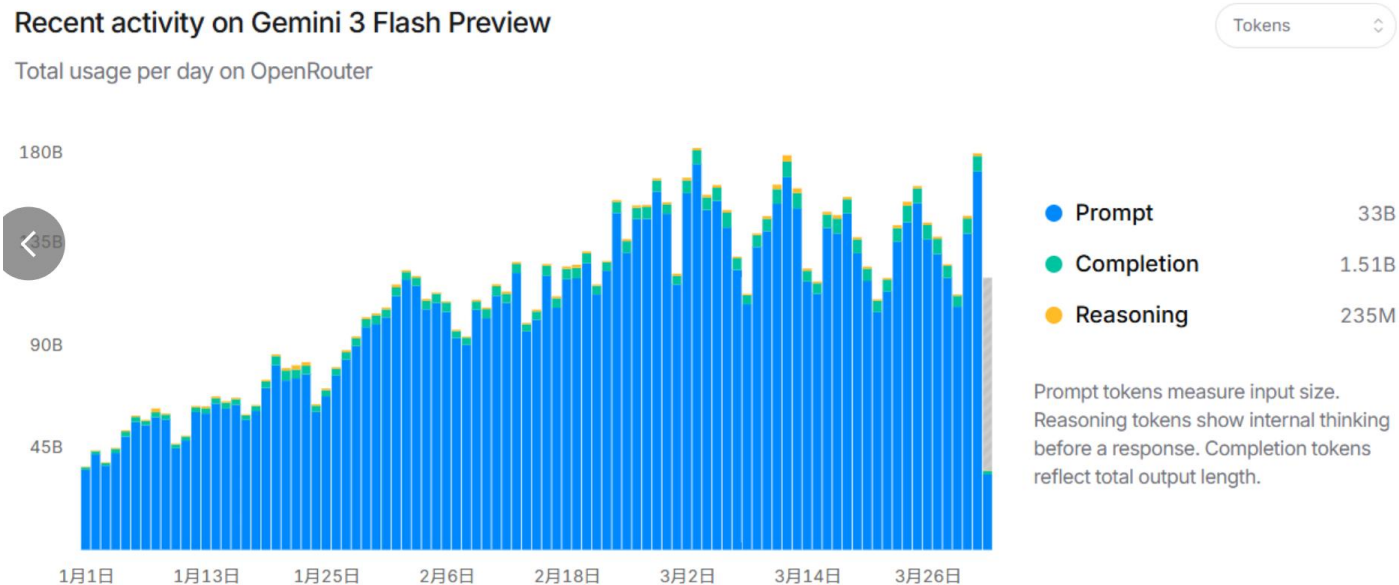
图表：模型参数量增长趋势（红线）VS 单GPU内存扩展趋势（绿线）



1.2 AI推理带来存储范式改进

- 在长上下文推理场景下，大模型性能瓶颈正从传统的算力约束（compute-bound）转向显著的内存约束（memory-bound）。
 - OpenRouter 平台上 Gemini 1.5 Flash 的每日总 Token 使用量：Prompt (33B)：对应 Input Tokens。这包括用户发送的问题、上传的文档以及维持对话连贯性的历史上下文；Completion (1.51B)：对应 Output Tokens。即模型生成的回答内容；Reasoning (235M)：对应思考 Token。这是模型在输出最终答案前进行的内部推理逻辑（类似于 OpenAI o1 或 DeepSeek R1 的思维链）。
 - 在全口径 Token 消耗中，95% 的 Prompt 占比意味着模型在运行过程中，绝大部分能耗和延迟并非产生于 GPU 的逻辑运算，而是产生于海量上下文在存储单元与计算单元之间的频繁搬运。随着 Gemini 等长文本模型（Long-Context）成为主流，KV Cache（键值缓存）存储容量及带宽的挤占已达到临界点。

图表：Gemini 1.5 Flash 的每日总 Token 使用量

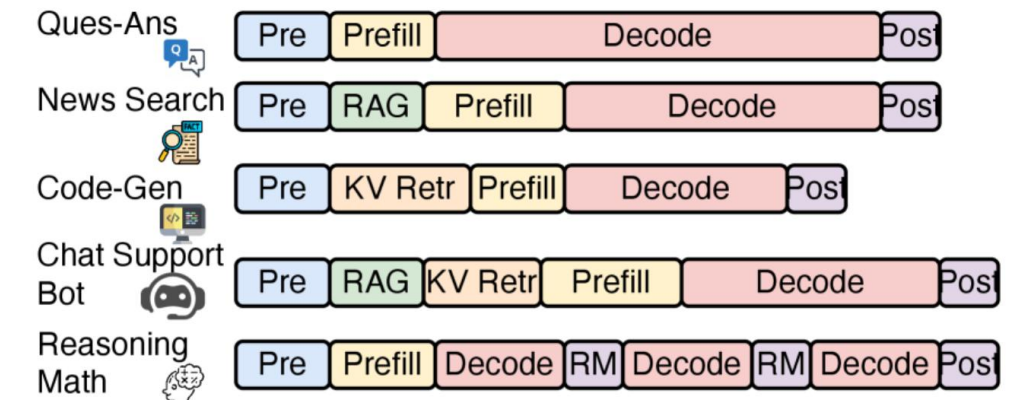


- 大语言模型推理分为两个阶段：计算密集型的prefill（处理输入提示）和带宽密集型的decode（生成输出token），decode阶段受限于存储。
- 一次带有 RAG 向量数据库的推理的过程：
 - 1、请求到达：用户输入一个 Prompt（例如 20 个字）。
 - 2、RAG检索：系统在唤醒大模型推理之前，将Prompt向量数据库（通常存在高性能 NVMe SSD 中）进行高频并发检索，找回相关的背景知识。
 - 3、Prompt 拼接：将找回的背景知识与用户的原始问题拼接，组装成一个可能长达数千甚至上万 Token 的“超级 Prompt”。
 - 4、预填充阶段 (Prefill – GPU计算密集)：
 - 4.1 Prompt 被切成 Token 序列。随后，Token 通过模型底层的词嵌入层（Embedding Layer）进行“查表寻址”，批量转化为高维初始特征向量矩阵。
 - 4.2 始向量矩阵会一次性、并发地穿过模型内部的数十个隐藏层（每层包含 Attention 和 FFN 模块）：在 Attention 模块中，模型对输入序列进行全局上下文的关联计算，为每一个 Token 算出对应的 Key 和 Value 张量，并将这些各层独立、互不干扰的上下文记忆（KV Cache）批量写入 HBM 中，为后续的生成做准备。在 FFN 模块中：矩阵数据通过庞大的前馈神经网络映射矩阵，提取出复杂的深层逻辑特征。
 - 4.3 经过所有隐藏层的层层非线性变换，初始矩阵最终演化为一个蕴含了全量上下文意图的特征矩阵，该矩阵进入输出端（LM Head）计算全词表概率，最终顺势生成并吐出全句的第一个回复词（First Token）。
 - Prefill阶段是计算密集型（Compute-Bound），是将几千个词组成的巨大矩阵一次性送入网络，进行大量的矩阵乘矩阵（GEMM）运算。
 - 5、解码阶段 (Decode – 存储带宽紧张)：
 - 5.1 上一轮刚刚生成的1个 Token，转化为初始向量。
 - 5.2 Token 会层层穿过模型，在每一层中经历Attention和FNN两大核心组件：Attention 模块（上下文聚合与记忆写入），该 Token 会去 GPU 的高带宽显存（HBM）中读取之前所有历史 Token 积攒的 KV Cache进行匹配计算；同时，计算出自身的 Key 和 Value 向量，追加存入 HBM 中，完成记忆的更新；FFN 模块（深度知识提取），该向量与当前层庞大的固化模型权重进行矩阵相乘，提取出深层逻辑特征。**为了处理1个Token，GPU从HBM 中把历史 KV Cache和静态权重矩阵全部搬运到计算核心，繁重的数据搬运导致算力闲置，带来“访存受限（Memory-bound）”难题。**
 - 5.3 当该 Token 历经几十层的非线性变换，从最后一层穿出时，已演化为一个浓缩了全部上下文意图的“终极特征向量”。该向量会进入输出映射头（LM Head），与系统内包含数万个词汇的“标准词典矩阵”进行点积运算，最终生成一张庞大的“全词汇概率表（Logits）”。
 - 5.4 系统根据设定的采样参数，从这张概率表中挑选出最合适的下一个词。这个刚出炉的新词，将立刻化身为下一轮的输入，结合 HBM 中的 KV Cache，重新启动步骤 1 到 3，周而复始，直至模型吐出停止符（EOS）。

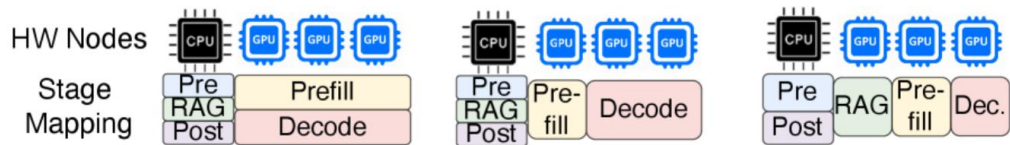
图表：Transformer大模型推理两个阶段的对比

推理阶段	资源需求特性	计算模式	显存占有模式
Prefill	计算密集型	高并行度，矩阵乘法为主	低，固定规模
Decode	内存密集型	顺序生成，访存为主	随着上下文增长指数级上升

图表：Transformer大模型在不同场景下的推理步骤

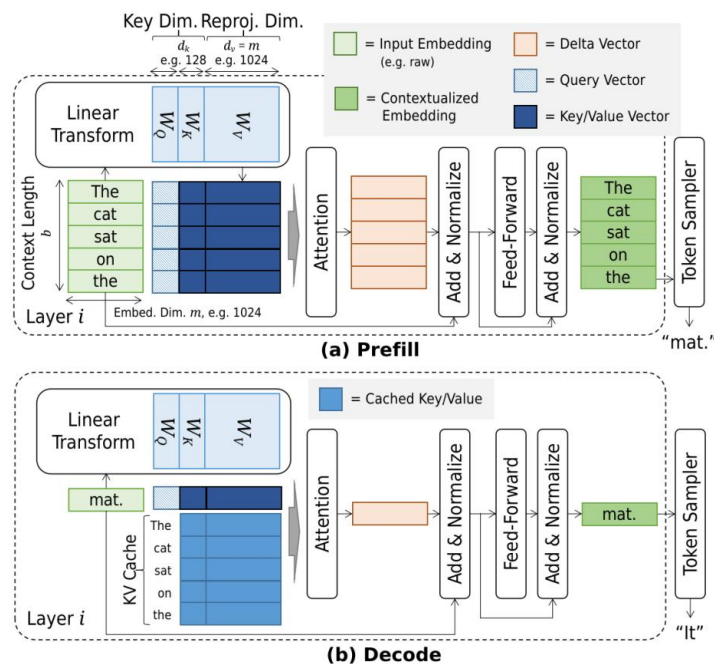


(A) Various LLM inference pipelines for different applications.



(B) Example mappings for News Search request on 1 CPU + 3 GPUs

图表：推理核心是Prefill和Decode阶段



■ “存储带宽与访存延迟”是大模型推理的核心瓶颈。

- LLM推理的关键性能指标包括首字延迟（TTFT，Time to First Token）与生成速度（TPS，Token Per Second），分别对应系统的响应能力与吞吐能力。其中，推理性能的本质约束并非算力不足，而是数据在存储与计算单元之间的传输效率。
- 1) Prefill阶段（决定TTFT）：延迟受数据规模与前处理开销主导。
- 首字生成前，系统需完成模型权重加载、输入序列处理及注意力计算，其延迟主要受以下因素影响：长上下文与RAG引入：超长输入或外部检索数据（RAG）显著增加数据读取与计算规模，拉长首字生成时间推理策略复杂化（如思维链）：部分模型在输出首字前已完成多步隐式生成，进一步放大TTFT。
- 2) Decode阶段（决定TPS）：带宽与延迟成为主导瓶颈。
- 在自回归解码过程中，系统逐Token生成输出。单步计算量有限，但需反复访问模型权重与历史KV Cache，使性能高度依赖存储系统：带宽约束（决定吞吐）：每次生成Token均需高频访问权重与KV Cache，计算呈“低算力密度、高数据搬运”特征，存储带宽直接决定TPS上限延迟约束（决定稳定性）：KV Cache随序列长度动态增长，访存延迟在逐Token生成过程中持续累积，导致长文本场景下推理速度逐步下降。

图表：大语言模型推理的主要硬件瓶颈及应对研究方向总结

	大语言模型解码（Decode）的特征与趋势 + 有前景的研究方向	内存容量	内存带宽	互连延迟	计算
硬件改进的驱动因素	传统 Transformer 解码		✓	✓	
	MoE	✓	✓	✓	
	推理模型	✓	✓	?	
	多模态	✓	✓	?	
	长上下文	✓	✓	?	
	检索增强生成	✓	✓	?	
	扩散				✓
助力前行之路	① 高带宽闪存	✓		↑	
	② 近存计算		✓	↑	
	③ 3D 计算-逻辑堆叠		✓	↑	
	④ 低延迟互连			✓	

■ 提升存储带宽和容量可以显著增强推理性能、降低推理成本，“以存代算”是必然趋势。

➤ 英伟达H100和H200的计算能力相同，H200的升级主要体现在HBM容量和带宽的提升，HBM容量提升76%，带宽提升43%。

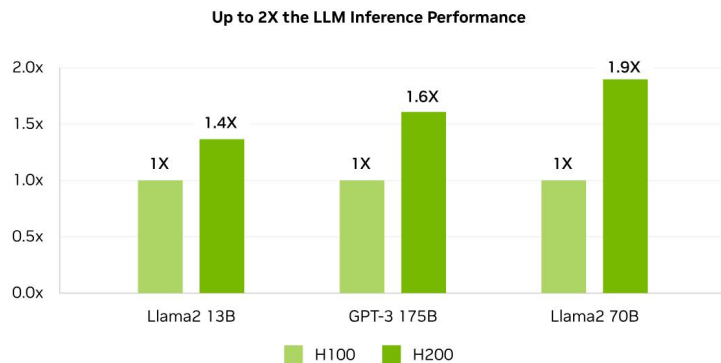
➤ 推理性能翻倍，吞吐量翻倍，延迟降低50%：在处理 Llama 2 等大语言模型时，H200 的推理速度比 H100 提高了接近 2 倍。Llama-2-70B 如果采用 FP8 量化，模型权重本身大约占 70GB。根据 NVIDIA 官方评测指标及硅谷硬件机构 Uvation 等的实测，在 Llama-2-70B 的推理中，受限于 80GB 容量，模型权重占据了70GB，剩余的10GB容纳不了高并发用户的KV Cache，H100 的 Batch Size 在 32 左右即触及显存瓶颈，计算核心绝大部分时间都在空转等待；而 H200 凭借 141GB 的大容量，模型权重后剩余 70GB用于KV Cache保存，Batch Size可以提升到 128，实现了吞吐量翻倍，同时因为更高的带宽，数据搬移效率大幅提升，推理延迟降低50%。

➤ 总拥有成本的降低：在总拥有成本和功耗层面，H200均较H100有50%的下降。H200 前期成本比 H100 高出约 20%，但是由于吞吐量翻倍且能效更高，每次推理的成本降低了 68%。

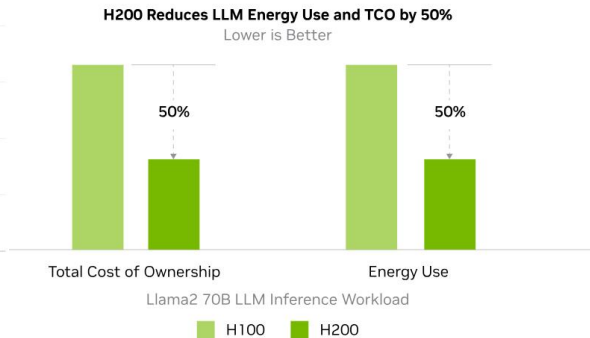
图表：H100 H200的性能对比

核心规格维度	NVIDIA H100 (SXM)	NVIDIA H200 (SXM)	对比
算力 (FP8)	1,979 TFLOPS (稠密), 3,958 TFLOPS (稀疏)	1,979 TFLOPS (稠密), 3,958 TFLOPS (稀疏)	计算能力一样
HBM容量	80 GB	141 GB	提升76%
显存带宽	3.35 TB/s	4.8 TB/s	提升 43%
显存介质类型	HBM3	HBM3e	

图表：H200推理性能大幅提升



图表：H200的功耗和成本降低50%



图表：H100和H200性能对比

指标	H100	H200 提升
吞吐量	~60 tokens/秒	120+ tokens/秒 (约2倍)
延迟	基准水平	约降低50%
最大稳定上下文	48K tokens	128K+ tokens
批处理扩展	峰值在32	线性扩展至128

领域	H200 优势	业务结果
成本	每 token 成本降低 68%	AI 投资回报更快
响应性	延迟降低 50%	提供更高端的用户体验
基础设施	可替代 3 个 H100 集群	服务器成本降低 60%

■ 面对大模型推理的访存受限问题，产业界从模型优化和硬件优化双线推进。存储体系正从单一层级向高带宽+大容量+分级管理的协同架构演进，存储与计算的关系也由传统解耦逐步走向协同优化。

■ 1、模型层面优化：

➤ 1.1算法层优化：核心思路是将数据变得更轻、更少。

- 权重量化与 KV Cache 量化：采用 AWQ、GPTQ、FP8 等低比特量化技术，将传统的 16 位浮点数 (FP16) 强行降维至 8 位 (INT8) 甚至 4 位 (INT4)，像是给数据“降分辨率”，在微弱的精度损耗下，直接将模型权重与上下文历史记忆的物理体积压缩 50% 至 75%，是目前主流商业化部署的绝对标配。

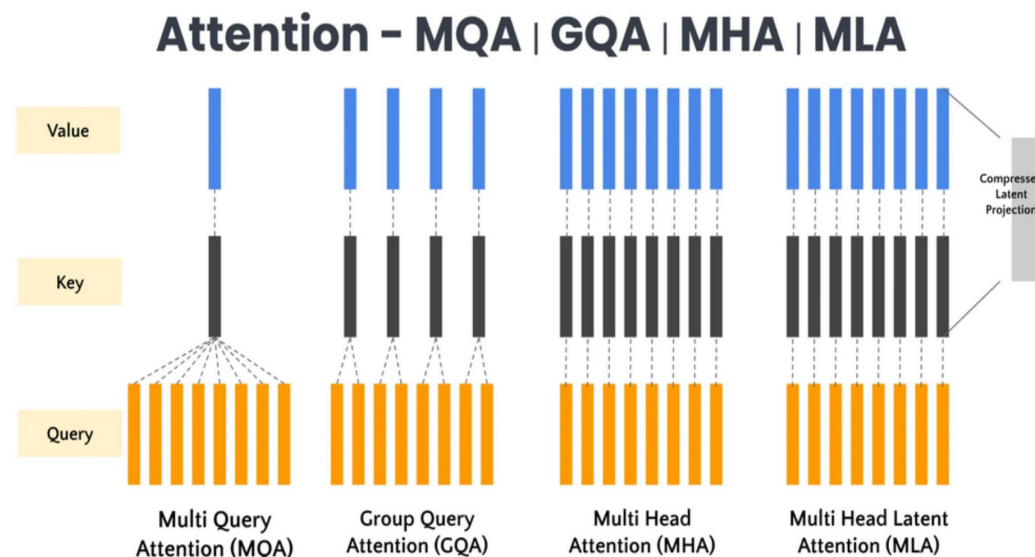
- Attention 注意力机制的变化：针对 KV Cache 随注意力头数 (Heads) 暴增的特点，引入分组查询注意力 (GQA) 或多查询注意力 (MQA)。通过让多个查询头 (Q) 强制共享同一组键值头 (K 和 V)，降低 KV Cache 的显存占用。以 Llama-2/3-70B 为例，采用 GQA 后其 KV Cache 体积暴降 8 倍。

➤ 1.2框架层优化：不改变数据大小，充分利用存储空间。

- PagedAttention (显存分页调度)：打破了传统推理框架必须为用户预留“连续物理显存”的刚性制约。借鉴操作系统的虚拟内存分页技术，将 KV Cache 切分为极小的离散数据块 (Blocks) 并按需动态分配。该技术将显存碎片浪费率从 50% 降至 4% 以下，直接拉动单台服务器的并发服务人数 (吞吐量) 跃升 2 到 4 倍。

- Chunked Prefill (块状预填充削峰)：针对长文本输入时瞬间产生的海量激活值 (Activations) 洪峰，采用“化整为零”的策略。将单次十万字的超长序列切分为多个小块 (Chunks) 串行计算，算完即刻释放中间激活值。该方案以极小的算力调度开销为代价，实现了显存占用的“削峰填谷”，清除了长文本瞬时爆显存 (OOM) 的系统隐患。

图表： Attention注意力机制的变化



■ 面对大模型推理的访存受限问题，产业界从模型优化和硬件优化双线推进。

➤ 2、硬件及架构优化：

➤ 2.1提升存储器的带宽

• HBM：通过 2.5D 先进封装将 DRAM 颗粒与计算核心绑定。未来演进方向是 HBM4 及更高层数的堆叠（12层至16层），持续提升带宽。

• WOW 3D堆叠DRAM：通过DRAM芯片堆叠在计算芯片上，使用混合键合技术，提高带宽。

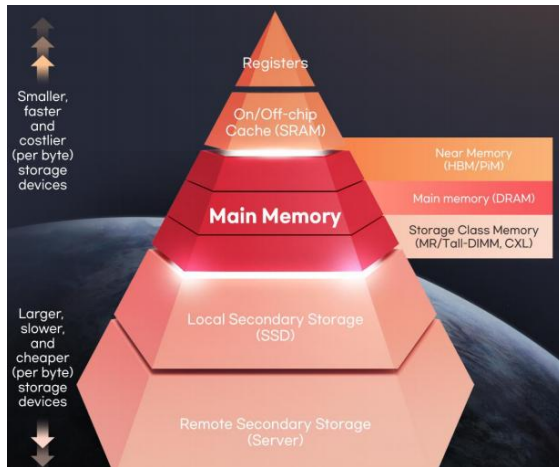
• HBF：通过堆叠NAND闪存而制成，通过在HBF中存储KV Cache，GPU和HBM可以减轻存储KV Cache的负担。

➤ 2.2存储和计算的解耦与重构

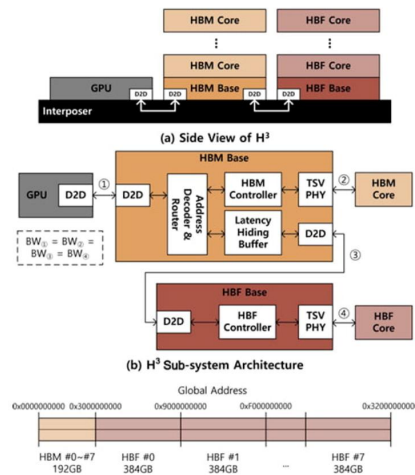
• CXL内存池化：通过 PCIe 总线，将整个数据中心的海量 DRAM 组成一个“全局共享内存池”，GPU 可以在瞬间向池子“动态借用”几百 GB 内存，实现显存容量的液态化按需分配，极大地降低了数据中心的硬件闲置成本。

• Prefill（预填充）与 Decode（解码）的物理分离部署：由于预填充（计算密集）和解码（访存密集）对硬件资源的要求完全冲突，顶级框架正在将它们拆分到不同的物理集群。A 集群全用高算力 GPU 专门负责“读长文”（Prefill），读完后将生成的庞大 KV Cache 通过高速网络瞬间传输给 B 集群；B 集群全用大显存/大带宽 GPU，专门负责“逐字吐出回复”（Decode）。

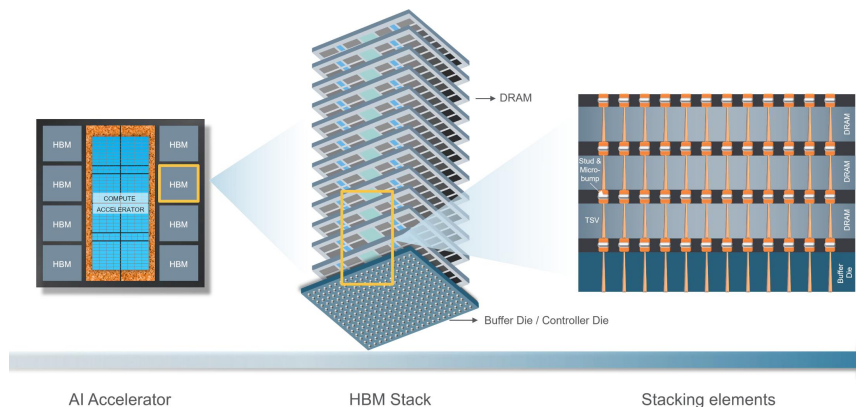
图表：存储分级



图表：HBF



图表：HBM



- HBM预计向更高带宽、更大容量、更高I/O密度升级，从而解决内存墙的带宽、容量问题。
- HBM使用TSV、Microbump实现3D堆叠结构，并采用2.5D封装技术与GPU直接封装在一起，在不占用面积的前提下，实现容量拓展、高带宽和降低功耗。
- HBM4目前已进入量产阶段，带宽在2TB/s以上（较HBM3E提升67%）、整体存储容量36-64GB（最大容量较HBM3E提升70%），HBM4E尚在技术展示与研发阶段，有望于2027年量产。
- HBM未来提升点：1）带宽：同步提升单通道速率、通道数，速率的提升主要源于制程升级，通道数的增长主要通过改进键合工艺（Microbump → Bump-less Cu-Cu Hybrid Bonding）等方式缩小pitch，HBM4-HBM8带宽预计由2TB/s提升至64TB/s，增长32倍；2）容量：同步提升单Die容量、堆叠层数，HBM4-HBM8单Die容量预计由24-32Gb左右提升至80Gb，堆叠层数预计由8-16Hi提升至20-24Hi，整体存储容量预计由36-64GB提升至200-240GB，增长4-5倍。3）定制化logic die：HBM承担更多的计算功能。

图表：HBM技术演进

技术世代	HBM/HBM2	HBM2E	HBM3	HBM3E	HBM4	HBM4E	HBM5
主要供应商							
发布时间	2013年/2016年	2018年	2020年	2023年	2025年	2027年(预计)	2030年(预计)
DRAM Die堆叠	4Hi/4-8Hi	4-8Hi	8-12Hi	8-12Hi	8-16Hi	8-16Hi	16Hi- >20Hi
单Die密度	2Gb/8-16Gb	16Gb	16Gb	24Gb	24-32Gb	32Gb	待定
存储容量	1GB/4-8GB	8-16GB	16-24GB	24-36GB	36-64GB	36-64GB	待定
I/O速率	1 Gbps/2-2.4 Gbps	3.2-3.6 Gbps	5.6-6.4 Gbps	8.0-9.8 Gbps	>9 Gbps	>12 Gbps	待定
带宽	128 GB/s/205-307 GB/s	460 GB/s	819 GB/s	1.2 TB/s	>2 TB/s	>3 TB/s	待定
典型应用	AMD Fiji GPU/NVIDIA P100	NVIDIA V100	NVIDIA A100	NVIDIA H200	NVIDIA Rubin GPU	NVIDIA Rubin Ultra	待定
制程节点	SK 海力士	2x nm/2y/2z nm	1y nm	1z nm	1b nm	1c nm	待定
	三星		1y nm	1z nm	1a nm	1c nm	待定
	美光		1z nm		1β nm	1γ nm	待定
堆叠技术	SK 海力士		MR-MUF	MR-MUF	MR-MUF	MR-MUF	HCB
	三星		TC-NCF	TC-NCF	TC-NCF	TC-NCF	HCB
	美光		TC-NCF	TC-NCF	TC-NCF	待定	HCB

图表：HBM未来技术发展路线

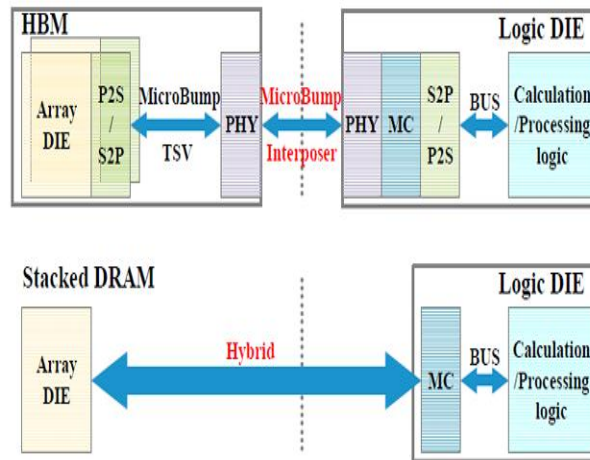
技术参数	HBM5 (2030)	HBM6 (2032)	HBM7 (2035)	HBM8 (2038)
单通道速率	8 Gbps	16 Gbps	24 Gbps	32 Gbps
I/O通道数	4096	4096	8192	16384
带宽	4 TB/s	8 TB/s	24 TB/s	64 TB/s
单die容量	40 Gb	48 Gb	64 Gb	80 Gb
堆叠层数	16-Hi	16/20-Hi	20/24-Hi	20/24-Hi
存储容量	80 GB	96/120 GB	160/192 GB	200/240 GB

- **WoW 3D堆叠DRAM**：定制化DRAM（容量、层数均定制），原理是在1片逻辑芯片上堆叠多层DRAM芯片，通常使用TSV硅通孔技术、Wafer on Wafer的混合键合工艺（Hybrid Bonding）实现多层芯片之间的电气连接。其对比HBM带宽更高、功耗更低，大模型推理算力需求激增下，显存带宽成为瓶颈之一，WoW 3D DRAM有望显著提升AI推理效率。
- 1) 高带宽：3D DRAM采取WoW混合键合工艺，可使通孔间距（Pitch）显著缩小，从而构建更多I/O数量，带宽=位宽*数据的传输速率，因此I/O数量增加通过提升位宽提高了带宽，以紫光国芯23年论文中的2层SeDRAM方案为例，通过将Mini-TSV的Pitch缩小至1.5um，构建了131072个IO数量，平均每Gb的IO数量达2048个（是192Gb HBM3的386倍），最终每Gb带宽为135GB/s（是HBM3的29倍）。
- 2) 低功耗：主要源于①数据的传输路径短：3D堆叠DRAM与逻辑芯片是3D结构，属于近存计算；②移除了传统HBM互联结构中耗时耗能的PHY区域（物理接口模块）；③IO速度慢：紫光国芯2层DRAM方案IO速度541Mbps（是HBM3的75%）；④混合键合功耗低：铜对铜的互连电阻更小，而HBM的2.5D封装中芯片常通过Micro Bump互连，微凸点通常由铜柱/焊锡材料构成。
- 3) 但是3D DRAM容量拓展性不及HBM，主要系3D封装下一颗计算芯片仅能配一颗WoW 3D DRAM，但是HBM采用2.5D封装，1颗计算芯片可用多颗HBM。

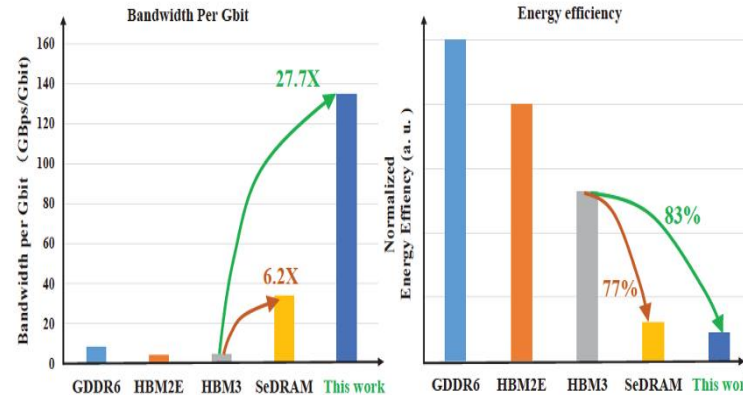
图：SeDRAM性能对比

	传统DRAM	HBM			WOW 3D堆叠DRAM		WOW 3D堆叠DRAM的特点
类型	GDDR6 ISSCC2018	HBM2E ISSCC2020	HBM3 ISSCC2022	SeDRAM IEDM2020	SeDRAM（2层）		
连接方式	-	ubump, TSV	ubump, TSV	Hybrid bonding	Hybrid bonding, Mini-TSV		
ubump / TSV pitch(um²)	-	48*55	96*110	-	1.5 * 1.5	Pitch更小	
HB pitch	-	-	-	3	3		
是否需要PHY	-	Yes	Yes	No	No	去掉PHY，降低功耗	
DRAM堆叠层数	-	8	8	1	2	堆叠层数少于HBM	
每层DRAM容量（GB）	-	2	3	0.5	4		
存储容量（GB）	1	16	24	0.5	8		
IO数量	32	1024	1024	4096	131072	IO数量更多	
IO速度(Mbps)	16384	4096	7168	266	541	IO速度慢，功耗低	
总带宽(GBps)	64GBps	512GBps	896GBps	136GBps	8656GBps		
耗电量（相对值）	100%	80%	53%	12%	9%	低功耗性能显著	
每Gb带宽（GBps/Gb）	8	4	4.7	34	135	单位容量的带宽更高	

图表：3D堆叠DRAM中逻辑-DRAM的接口（对比HBM）



图：带宽和功耗对比

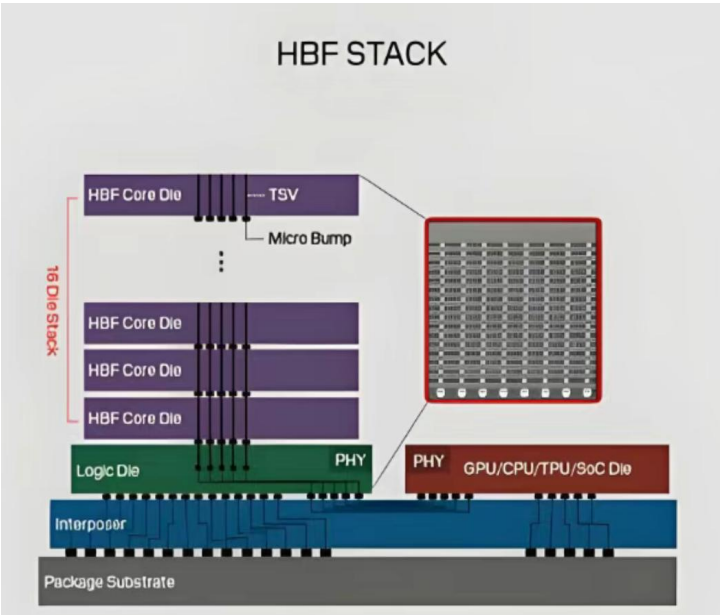


- **HBF**：通过堆叠NAND闪存显著提升容量，定位是HBM的补充，通过在HBF中存储KV Cache，GPU和HBM可以减轻存储KV Cache的负担。
- 1) 巨大容量兼具成本优势：同一物理空间内单个HBF堆叠容量高达512GB（比HBM容量高出10倍以上），8堆栈可至4TB，可支撑万亿参数模型本地加载。但是其基于单位成本更低的NAND技术，可显著降低总体拥有成本。
- 2) 高带宽、低功耗：通过并行架构，带宽接近HBM水平，HBF带宽可高达1.6TB/s，是传统PCIe4.0 SSD的200倍以上，基本达到HBM3带宽水平，但低于HBM4，可以满足AI推理等快速数据读取的需求。同时，静态功耗远低于需要不断刷新的DRAM。
- 综合看，HBF的定位是解决HBM容量不足和SSD速度太慢的存储产品，适用于储存模型权重、长文本、特征库等“温/冷数据”。25年底-26年初，研究指出HBF可将AI推理性能/瓦特提升至纯HBM配置的2.69倍，并在Llama 3.1 405B等模型上仅损失2.2%性能。
- HBF有望于27年小规模部署，导入谷歌、英伟达、AMD等AI芯片，2030年或大规模普及，2038年可能超越HBM成为AI存储的主力。

图表：HBF海外主要厂商进度

公司名称	进度
闪迪	HBF由闪迪发明，公司于2025年与SK海力士签订谅解备忘录，合作开发并制定规范。闪迪的目标是在2026年下半年交付其HBF闪存的第一批样品，首款集成该技术的AI推理硬件预计将于2027年初推出。
SK海力士	拥有HBM及至关重要的TSV技术的量产记录，并且对NAND闪存非常了解，因此成为闪迪理想的合作伙伴。在产品进度方面，SK海力士计划26H2发布一款HBF原型产品。
三星	三星计划2027年底 - 2028年初将 HBF 技术导入英伟达 / AMD / 谷歌产品。
铠侠	铠侠于2025年8月推出新一代采用菊花链连接的HBF模块原型，该原型产品将拥有5TB的大容量与64GB/s的高带宽。

图表：HBF堆栈架构

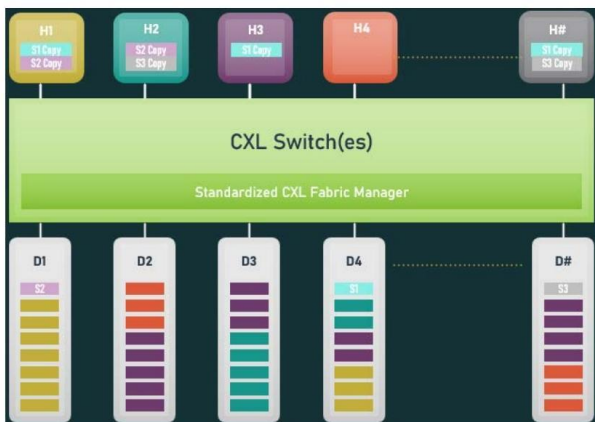


图表：HBF成本与性能介于HBM和传统SSD之间

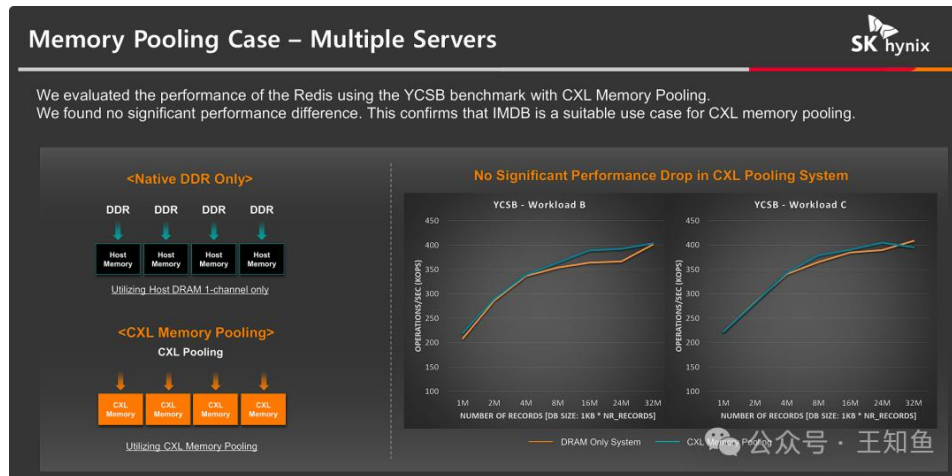
特性	HBF	HBM	传统 NAND SSD
核心优势	高容量、高带宽、低成本	极高带宽、超低延迟	高容量、极低成本
单堆叠 / 单芯片典型容量	最高 512GB	约 24-48GB	1TB-2TB
带宽水平	非常高，接近 HBM	极致高	相对较低
单位成本	相对较低	极高	极低
最佳应用场景	AI 推理、读密集型任务	AI 训练、高性能计算	数据存储、归档

- **CXL内存池化方案可实现内存资源的整合与统一调度。**CXL内存池化方案支持跨CPU、GPU及其他计算加速器的内存资源进行统一寻址、统一调度与透明访问，在具体硬件实现上，企业级服务器可通过PCIe 5.0/CXL链路连接到 CXL Switch，进而与池化的内存资源相连。这一方案可实现内存资源的整合与统一调度，从而支撑更大规模、更高并发的大模型训练与推理任务。
- **CXL方案有助于存储分层，优化存储架构。**CXL内存池的引入，在DRAM与SSD之间创造了一个新的性能层级，有望作为一种高性能、可扩展的容量层，优化存储架构。CXL有助于突破“内存墙”，让多核CPU不再需要因为等待数据而空转。在系统处于极限高并发状态时，CXL方案由于能减少排队时间，反而能使延迟下降。
- **迈向“以内存为中心”的计算架构：**通过CXL实现多服务器共享内存，标志着数据中心正从传统的“单机独占内存”向“池化共享资源”转变。这不仅能显著提高内存利用率，还能让多台服务器协同处理超大规模数据集，而无需昂贵的数据搬迁。虽然CXL内存的访问延迟理论上高于本地内存，但在多线程并发访问的情况下，系统的整体吞吐量并未出现显著下滑。这意味着CXL能够作为有效的“扩展内存层”，完美平衡容量需求与性能表现。
- **CXL技术渗透率有望快速提升，支持CXL功能将成为服务器标配。**CXL在服务器DRAM中的总份额预计将从2024年的近乎为零，快速增长至2030年的约15%。支持CXL功能的服务器占比将提升至2026年的68%，并在2030年达到99%，这意味着届时几乎所有新出货服务器在硬件层面都将具备CXL能力。

图表：CXL内存池化方案

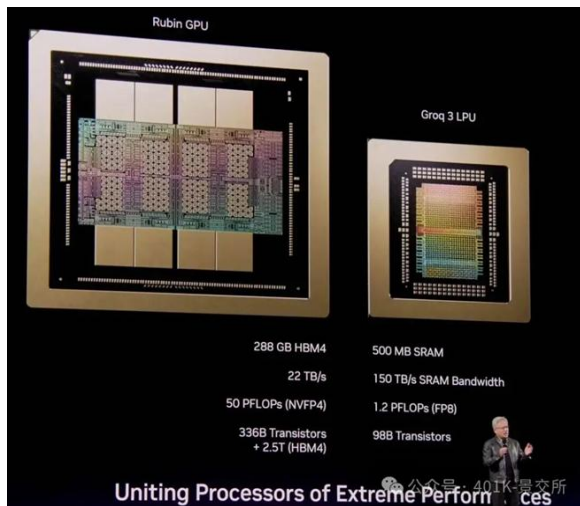


图表：CXL性能可以媲美本地内存



- NVIDIA Dynamo实现Prefill（预填充）与 Decode（解码）的物理分离部署，是软硬结构的终极推理架构。
- NVIDIA Dynamo是2025年3月GTC大会推出的开源分布式推理服务框架，旨在解决大语言模型（LLM）推理过程中的资源不匹配和低效问题。Dynamo通过PD分离式部署技术，将LLM推理拆分为两个阶段：计算密集型的预填充（Prefill）和内存密集型的解码（Decode），并分别部署到不同计算芯片上运行，从而实现资源的最优利用和推理效率的大幅提升。**Dynamo软件使两者协同，prefill环节交给Vera Rubin，对延迟敏感的decode环节给Groq，可将GPU每瓦Token处理性能提升35x。**
- 1) Vera Rubin机柜：搭载288GB HBM4，显存带宽22TB/s（较前代提升175%），针对Prefill阶段，主攻高吞吐量，针对海量上下文。
- 2) Groq3 LPX机柜：机架搭载32×8=256个LPU处理器，提供128GB（单颗500MB，存储带宽150TB/s，是HBM4的7倍）片上SRAM，AI推理算力达到315PFLOPS，Memory带宽40PB/s，Scale-up带宽640TB/s，针对Decode环节，主攻低延迟，生成token。
- GTC发布CMX+STX存储与计算架构：针对AI长上下文推理设计的新存储架构，NVMe SSD适合上下文热路径，有望大量使用，突破AI token膨胀下HBM的容量与成本瓶颈。具体而言，CMX是GPU外部的上下文内存存储平台，是STX架构的核心硬件，STX是基于BlueField-4构建的存储参考架构，基于Vera Rubin，整合CMX、Spectrum-X、DOCA 等构建近计算、高吞吐、低延迟的AI存储系统。

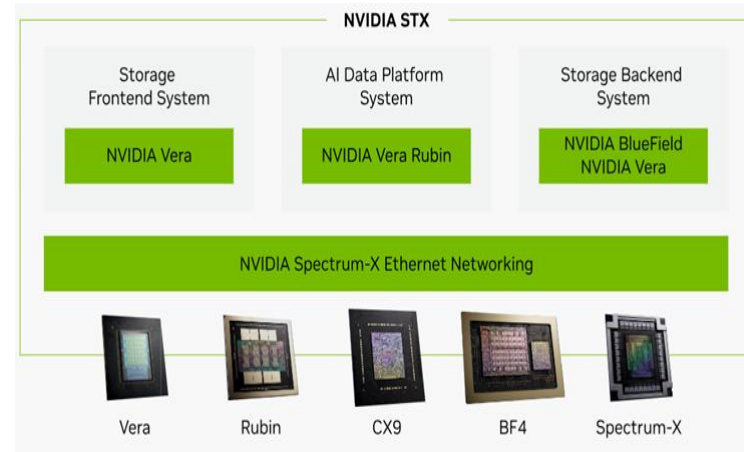
图表：Rubin GPU、Groq3 LPX参数



图表：Groq3 LPX机柜



图表：STX存储架构



目录

一、AI推理带来存储需求爆发和存储范式的改进

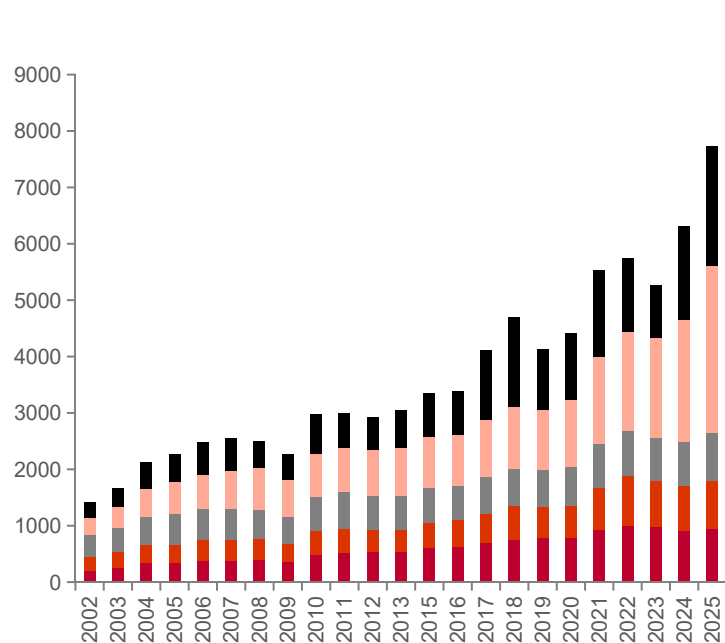
二、看未来2年供需持续紧张，原厂与客户签订长协

三、存储是AI硬件估值最低、业绩确定性最强的方向

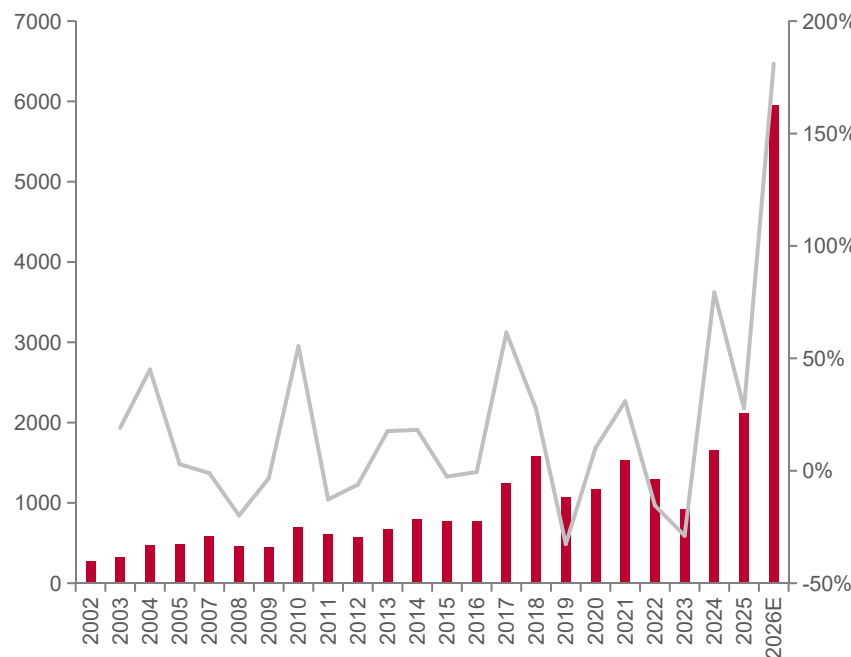
- 2025年全球半导体市场规模7722亿美金，其中存储器市场规模2116亿美金，创历史新高，yoy+28%，占比27%。
- 存储三大下游需求，手机、PC和服务器。此前PC+手机占存储需求的50%左右，存储周期受消费电子的库存周期影响大，服务器仅占比20-30%。26年存储器下游需求中服务器占比有望提升至40%左右，DRAM下游需求中服务器占比有望提升到50%-60%（含HBM），手机+PC需求占比<40%。

图表：全球半导体市场规模（亿美金） 图表：全球存储器市场规模（亿美金）

■ 分立、光电器件、传感器 ■ 模拟芯片 ■ 微型元件 ■ 逻辑芯片 ■ 存储器

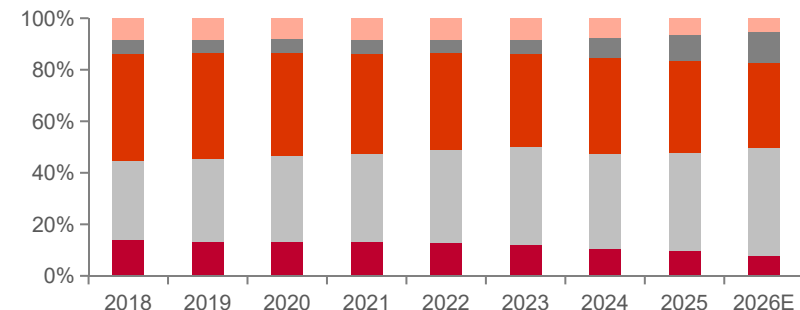


■ 存储器 — yoy

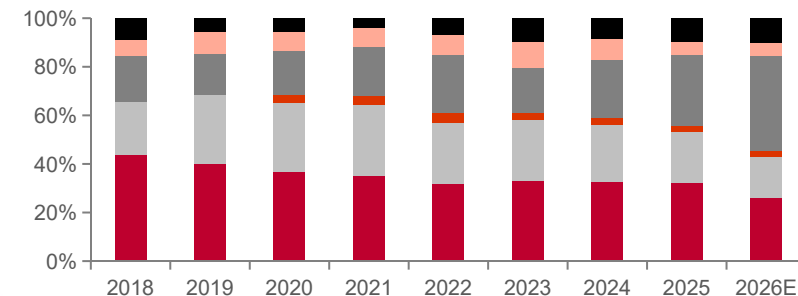


图表：全球DRAM（上方）/NAND（下方）下游需求

■ PC ■ Server ■ Mobile ■ Graphics ■ Others



■ Handsets ■ PC SSD ■ Game Console ■ Enterprise SSD ■ UFD+Memory Card ■ Others



- 受益AI数据中心建设需求，服务器是近几年存储需求的最强驱动力，26年有望成为DRAM、NAND最大下游需求。
- 手机、PC和服务对DRAM、NAND存储需求直观体现在出货量和单机容量两个维度。
- 1) 手机和PC的出货量已达到相对稳态天花板状态，对存储的需求更多体现在单机容量逐步提升。2021-2026E，手机出货量 CAGR-4%，PC 出货量 CAGR-7%；同期手机单机DRAM 容量CAGR +17%、NAND容量CAGR+14%；同期 PC DRAM 容量 CAGR+10%、NAND容量 CAGR+12%。
- 2) 服务器出货量和单机容量均在快速提升。2021-2026E 服务器出货量CAGR+5%，同期服务器单机DRAM容量 CAGR+14%、NAND容量 CAGR+27%。

图表：DRAM、NAND三大下游的出货量和单机容量趋势

	2018	2019	2020	2021	2022	2023	2024	2025	2026E	2021-2026E CAGR
出货量（百万台）										
手机	1,402	1,373	1,281	1,360	1,206	1,164	1,236	1,260	1,097	-4%
PC	265	273	312	361	301	260	263	285	253	-7%
服务器	12	12	13	14	15	12	14	17	17	5%
YOY										
手机		-2%	-7%	6%	-11%	-3%	6%	2%	-13%	
PC		3%	14%	16%	-17%	-14%	1%	8%	-11%	
服务器		0%	7%	7%	10%	-17%	17%	15%	4%	
DRAM单机容量（GB/台）										
手机	4	5	6	6	8	8	9	10	13	17%
PC	7	8	8	8	10	12	12	12	13	10%
服务器	353	420	476	545	585	750	743	828	1,040	14%
YOY										
手机		15%	19%	11%	22%	5%	9%	21%	29%	
PC		5%	-1%	6%	29%	17%	3%	-3%	7%	
服务器		19%	13%	15%	7%	28%	-1%	11%	26%	
NAND单机容量（GB/台）										
手机	72	92	117	145	179	229	231	258	283	14%
PC	192	323	367	451	564	729	770	733	783	12%
服务器	3,660	4,463	5,893	8,236	10,752	18,326	14,395	17,799	26,682	27%
YOY										
手机		28%	27%	24%	23%	28%	1%	12%	10%	
PC		69%	14%	23%	25%	29%	6%	-5%	7%	
服务器		22%	32%	40%	31%	70%	-21%	24%	50%	

注意：DRAM单机容量没有考虑HBM

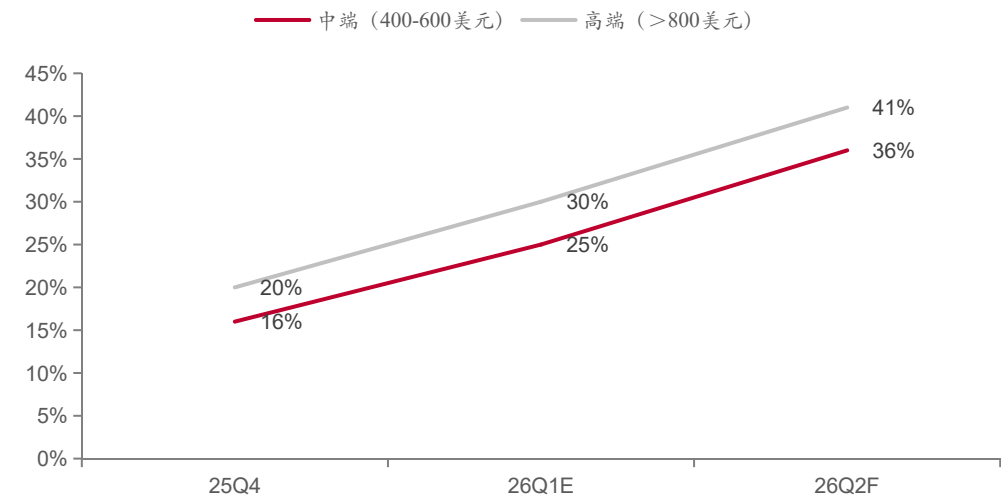
来源：IDC、Trendforce，中泰证券研究所

- 存储涨价大周期下，手机、PC、服务器存储成本不断提升。
- 手机：
 - 1) 低端机 (<\$200)：低端机常采用6GB LPDDR4X+128GB eMMC存储配置，假设低端机价格200美元，25年中至今RAM、ROM的平均容量不变，25年中至2026年6GB LPDDR4X现货价从14美元增长至80美元，128GB eMMC现货价从7.2美元增长至31美元，25年中至26年4/7日存储总成本占比从11%提升至55%。
 - 2) 中端机 (\$400-600)：据Counterpoint，中端机型（400-600美元，假设存储配置为8GB LPDDR5X+ 256GB UFS 4.0）的DRAM和NAND占比在2026年Q1将分别提升至14%和11%，并预计于Q2攀升至20%和16%，25Q4-26Q2F存储成本占手机总价值预计由16%提升至36%；
 - 3) 旗舰机 (>\$800)：据Counterpoint，旗舰型（>800美元，假设存储配置为16GB LPDDR5X HKMG+512GB UFS 4.1）的DRAM和NAND占比在2026年Q1将分别提升至17%和13%，并预计于Q2攀升至23%和18%，25Q4-26Q2F存储成本占手机总价值预计由20%提升至41%。

表：低端机 (<\$200) 的存储BOM测算

成本拆分		25H1	2026 (26.4.7)
RAM成本	手机RAM存储容量 (GB)	5.6	5.6 (假设不变)
	假设使用6GB LPDDR4X (美金)	14.0	80.0
	ASP (美元/GB)	2.3	13.3
	RAM成本 (美元)	13.0	74.4
	RAM 占总价 (假设200 \$) 比例	6.5%	37.2%
ROM成本	手机ROM存储容量 (GB)	144.1	144.1 (假设不变)
	假设使用128GB eMMC (美金)	7.2	31.0
	ASP (美元/GB)	0.1	0.2
	ROM成本 (美元)	8.1	34.9
	ROM 占总价 (假设200 \$) 比例	4.1%	17.5%
存储总成本	RAM+ROM成本 (美元)	21.1	109.3
	RAM+ROM成本比例	10.6%	54.6%

图：中端、高端手机存储成本占比



- 存储涨价大周期下，手机、PC、服务器存储成本不断提升。
- PC：
 - 1) AI轻薄本（\$800）：常采用32GB LPDDR5/5X+1TB PCIe4.0存储配置，假设价格为800美元，25年10月至26年1月，32GB LPDDR5现货价由15美元涨至32美元，25年10月至26年3月10日1TB PCIe4.0现货价由54美元涨至200美元，25年10月至今存储总成本占比由9%提升至29%；
 - 2) 游戏本（\$1100）：常采用16GB DDR5+1TB PCIe4.0存储配置，25年10月至26年1月，16GB DDR5现货价由57美元涨至122美元，25年10月至26年3月10日1TB PCIe4.0现货价由54美元涨至200美元，25年10月至今存储总成本占比由10%提升至29%。

表：笔电（游戏本）存储BOM测算

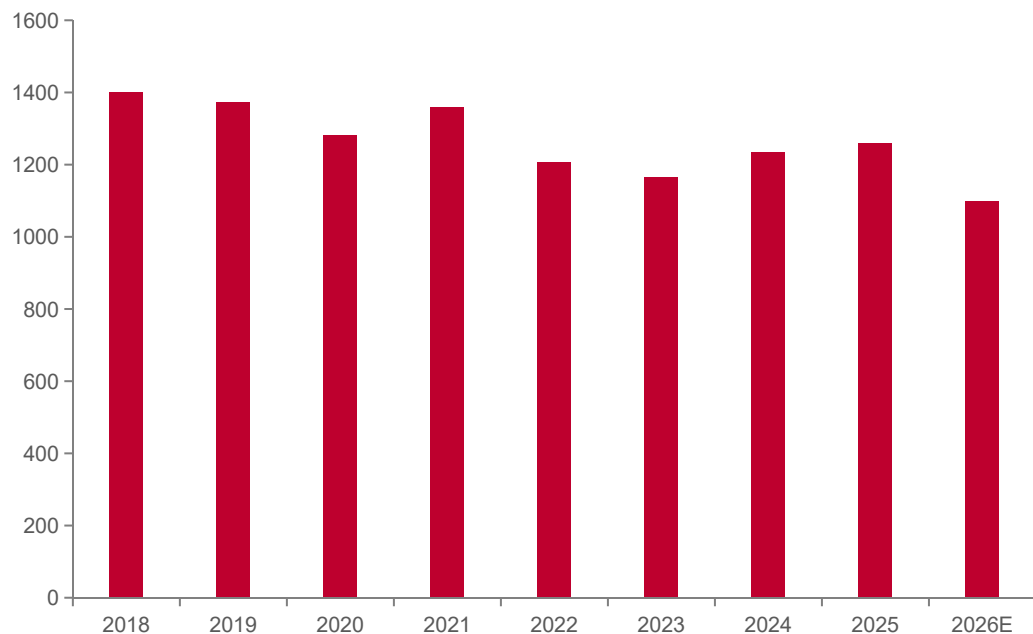
成本拆分		25年10月	26年
RAM成本（假设使用32GB LPDDR5）	电脑 RAM存储容量（GB）	32	32
	RAM成本（美元）	14.5	32（1月价格）
	RAM占比（假设800\$）	2%	4%
ROM成本（假设使用1TB PCIe4.0）	电脑ROM存储容量（TB）	1	1
	ROM成本（美元）	54	200（4/7价格）
	ROM占比（假设800\$）	7%	21%
存储总成本	RAM+ROM成本（美元）	68.5	232
	RAM+ROM成本占比	9%	29%

表：商用台式机（中低端机）存储BOM测算

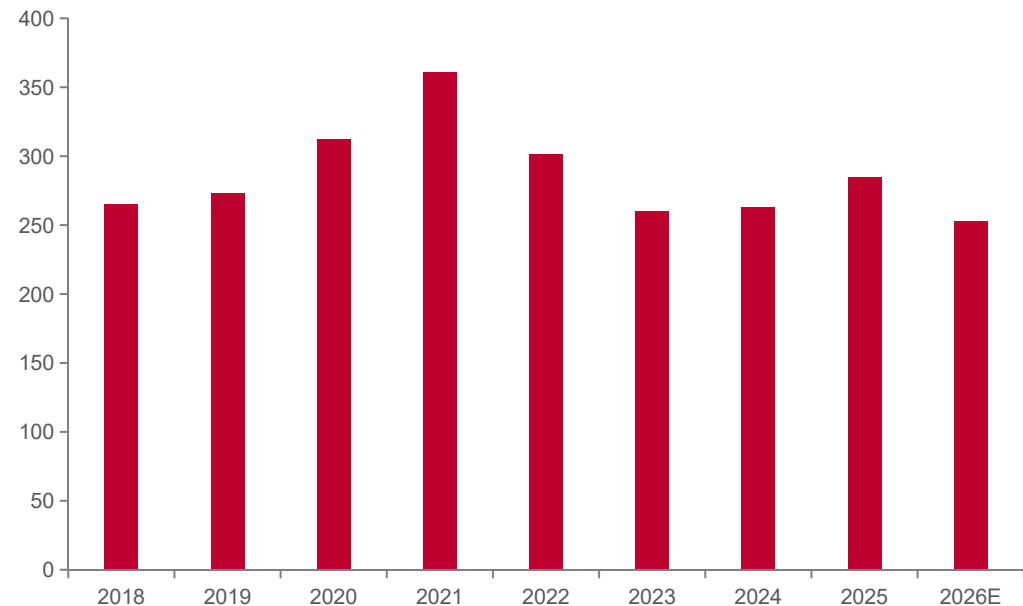
成本拆分		25年10月	26年
RAM成本（假设使用16GB DDR5）	电脑 RAM存储容量（GB）	16	16
	RAM成本（美元）	57	122（1月价格）
	RAM占总价（假设1100\$）比例	5%	11%
ROM成本（假设使用1TB PCIe4.0）	电脑ROM存储容量（TB）	1	1
	ROM成本（美元）	54	200（4/7价格）
	ROM占总价（假设1100\$）比例	5%	15%
存储总成本	RAM+ROM成本（美元）	110.75	322
	RAM+ROM成本比例	10%	29%

- 在存储芯片涨价大周期背景下，产业链成本端压力持续传导，消费电子终端出货承压明显，行业整体出货预期同步下修。
- 1) 手机市场表现尤为疲软，全球智能手机2025年出货量12.6亿部，2026年预计同比下滑12.9%至11亿部。
- 2) PC市场同样承压，2026年全球PC出货量由前期预期的2.85亿台下修至2.53亿台，同比降幅达11.3%。

图：全球手机出货量预测（百万部）



图：全球PC出货量预测（百万部）



- 普通服务器存储配置：以联想 ThinkSystem SR650 V4服务器为例，存储成本占比由25Q2的10%提升至26Q2的32%。
- DRAM方面配置8个32GB TruDDR5 6400MHz(1Rx4) RDIMM，NAND方面配置4个NVMe PCIe 5.0 x4 HS SSD，26年Q2服务器存储成本占比为32%，参照DDR5 RDIMM 32GB、PCIe 4.0 SSD 2TB 25Q2-26Q2的价格涨幅分别倒推25Q2 DRAM、NAND存储成本，并且假设其他组件成本不变，计算后25Q2整体存储成本占比为10%。

表：联想 ThinkSystem SR650 V4服务器存储成本占比

部件名称	组件	数量	单价\$	26Q2价格			25Q2价格		
				总价\$	占比	涨幅	单价\$	总价\$	占比
DRAM	DRAM: ThinkSystem 32GB TruDDR5 6400MHz(1Rx4) RDIMM	8	1906.1	15248.7	22%	371.01%	404.7	3237.4	6%
NAND	NAND:ThinkSystem 2.5"U.2 VA 3.2TB Mixed Use NVMe PCIe 5.0 x4 HS SSD	4	1840.5	7362.0	10%	251.19%	524.1	2096.3	4%
存储成本				22610.7	32%	324%	5333.7		10%
算力	Intel Xeon 6520P 24C 210W 2.4GHz Processor	2	1844.6	3689.2	5%	0%	1844.6	3689.2	7%
电源	ThinkSystem 1300W 230V/115V Platinum CRPS Hot-Swap Power Supply v2.4	2	167.7	335.4	0%	0%	167.7	335.4	1%
散热	ThinkSystem V4 2U Standard Heatsink	2	36.5	73.0	0%	0%	36.5	73.0	0%
机箱	Chasis:ThinkSystem SR650 V4 24x2.5" Chassis	1	516.2	516.2	1%	0%	516.2	516.2	1%
其他				43511.6	62%	0%		43511.6	81%
组件成本				48125.3	68%	0%	48125.3		90%
总价				70736.0			53459.0		

■ AI服务器存储配置：我们详细梳理了英伟达AI服务器，NVL72 GB200、GB300中，HBM、DRAM、NAND三类存储的价值量占机架售价的比例在8-17%左右。2026年，英伟达NVL72、NVL144机架对HBM需求2.1EB，NAND需求范围在49-99EB。

图表：英伟达AI服务器的存储配置梳理

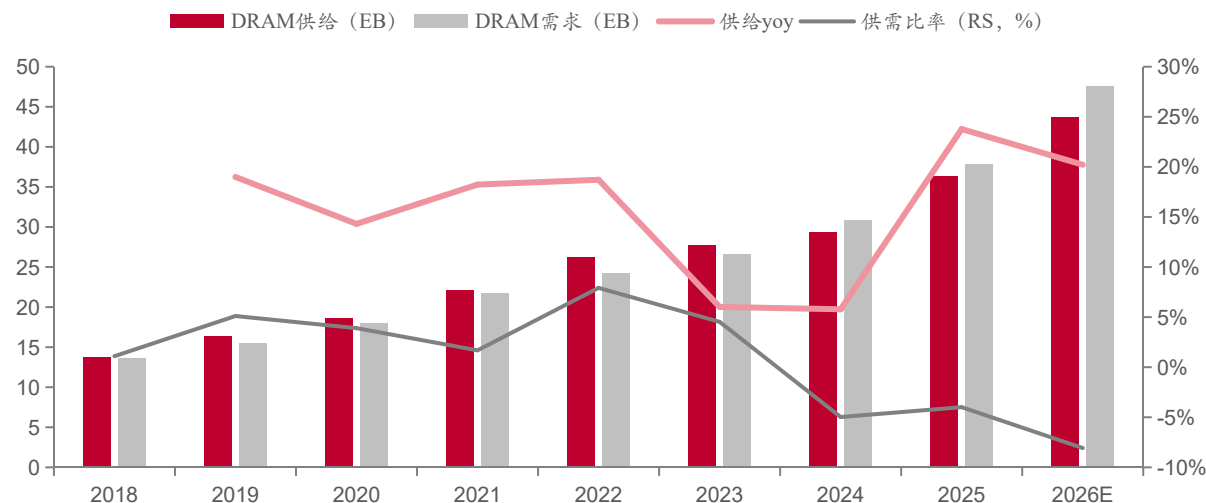
	DGXA100	DGXH100	DGXH200	GB200 NVL72	GB300 NVL72	NVL144 (Rubin)
类型	8卡服务器	8卡服务器	8卡服务器	机架	机架	机架
GPU	8颗A100	8颗 H100	8颗H200	72颗B200	72颗B300	72颗Rubin, 144 Rubin (按 dual-reticle 计数)
CPU	2颗CPU	2颗intel 8480C	2颗intel 8480C	36颗Grace CPU	36颗Grace CPU	36颗Vera CPU
HBM	640GB (单颗GPU配置6颗16GB HBM2e, 但实际调用5颗)	640GB (单颗GPU配置6颗16GB HBM3, 但实际调用5颗)	1128GB (单颗GPU使用6颗24GB HBM3E)	13.5 TB HBM3e (单颗GPU使用8颗24GB HBM, 全柜8*24*72)	20TB HBM3E (单颗GPU配置8颗36GB, 全柜36*8*72)	20TB HBM4 (单颗GPU配置8颗36GB, 全部就是36*8*72)
CPU配置的DRAM	2TB DDR4	2TB DDR5 (32根64GB内存条)	2TB DDR5 (32根64GB内存条)	17 TB LPDDR5X (单颗CPU配置16颗30GB LPDDR5=480GB)	17 TB LPDDR5X (单颗CPU配置16颗30GB LPDDR5=480GB)	使用SOCAMM2
SSD配置	数据缓存盘: 8×3.84TB NVMe企业级SSD, 合计30TB (DGX A100整机标配) 123。 另有OS盘为2×1.92TB NVMe SSD (部分配置), 用于系统和管理软件, 容量约3.84TB	1、2×1.92 TB NVMe M.2 SSD (each) in RAID 1 array) 【可用容量1.92TB, 因为做备份: 装操作系统、驱动、管理软件】 2、8×3.84 TB NVMe U.2 SED in RAID 0 array)	1、2×1.92 TB NVMe M.2 SSD (each) in RAID 1 array) 【可用容量1.92TB, 因为做备份: 装操作系统、驱动、管理软件】 2、8×3.84 TB NVMe U.2 SED in RAID 0 array)	一柜 18 个 compute tray, 每个 tray 8× E1.S Gen5 NVMe, 合计就是 18 × 8 = 144 块 E1.S 3.84 TB 盘: 144 × 3.84 =0.55 PB 7.68 TB 盘: 144 × 7.68 =1.11 PB。	最多 144 个 E1.S PCIe 5.0 容量取决于选配: 3.84 TB 盘: 144 × 3.84 =0.55 PB 7.68 TB 盘: 144 × 7.68 =1.11 PB。 (可选) 引导盘: 部分 OEM 在 tray 上另配 M.2 NVMe 作为 OS/Boot	最多 144 个 E1.S PCIe 5.0 容量取决于选配: 3.84 TB 盘: 144 × 3.84 ≈ 0.55 PB 7.68 TB 盘: 144 × 7.68 ≈ 1.11 PB。 (可选) 引导盘: 部分 OEM 在 tray 上另配 M.2 NVMe 作为 OS/Boot

图表：英伟达AI服务器的存储价值量占比情况

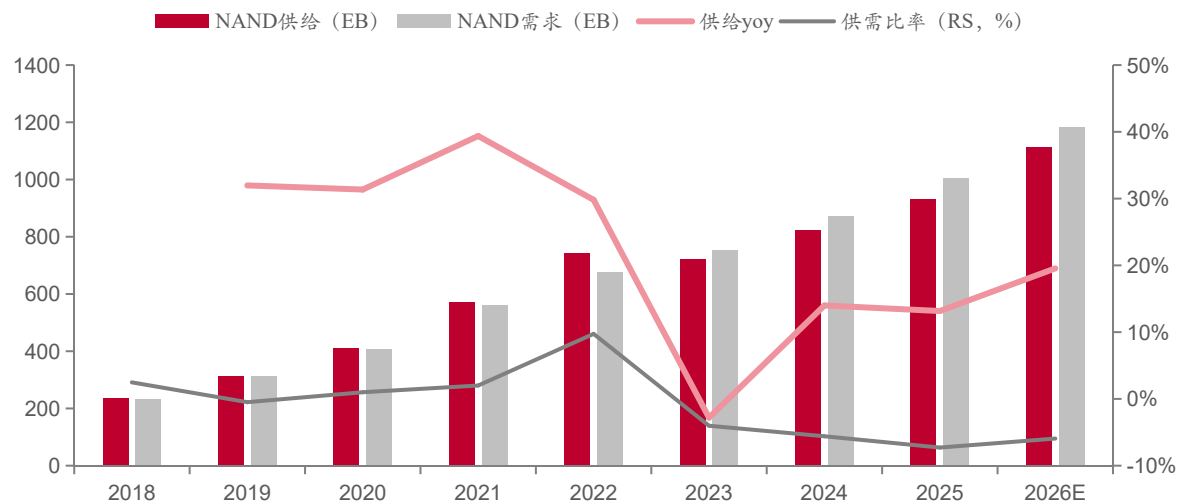
	DGXA100	DGXH100	DGXH200	GB200 NVL72	GB300 NVL72	NVL144 (Rubin)
美金/GB (2025年均价)						
HBM	14	14	14	14	14	14
DRAM	2.5	2.5	4.2	4.2	4.2	4.2
NAND Flash	0.07	0.07	0.08	0.08	0.08	0.08
机架配置的容量 (TB)						
HBM	0.6	0.6	1.1	13.5	20.0	20.0
DRAM	2.0	2.0	2.0	17.0	17.0	17.0
NAND Flash	33.8	33.8	33.8	553.0	553.0	553.0
机架配置的存储价值量 (美金)						
HBM	9,267	9,267	16,333	200,172	296,550	296,550
DRAM	5,120	5,120	8,602	73,114	73,114	73,114
NAND Flash	2,426	2,426	2,599	42,467	42,467	42,467
合计	16,813	16,813	27,534	315,752	412,131	412,131
不同存储的价值量占比						
HBM	55%	55%	59%	63%	72%	72%
DRAM	30%	30%	31%	23%	18%	18%
NAND Flash	14%	14%	9%	13%	10%	10%
机架的售价 (万美金)	10	20	30	320	350	360
AI服务器中的存储价值量占比	17%	8%	9%	10%	12%	11%

- 存储周期复盘：上轮周期供给收缩驱动反弹；本轮周期是AI需求爆发引领的成长，供不应求的局面预计维持时间更久（目前预期至少至28年），价格有望持续保持高位。
- 上轮周期（23Q3-24Q3）：原厂主动削减产能带动DRAM与NAND供给同比收缩，叠加库存低位，推动市场实现反弹。2022年，消费电子需求疲软叠加库存高企，存储市场供过于求，价格持续探底，主要原厂自2022年底开始下调产能利用率并削减资本开支，2023供给主动收缩叠加行业逐步去库存至低位，使得2023年供小于求，并驱动存储价格在2023年下半年触底反弹。
- 本轮周期（25Q2至今）：AI需求爆发开启大周期。2024年以来NAND、DRAM一直处于供不应求的状态，2026年DRAM、NAND需求供给缺口在4EB、70EB，需求缺口率8%、6%：1) 供给端看，2024Q4以来，三星等存储原厂陆续启动NAND减产，通过技术升级推动容量提升；DRAM产能虽维持满载但仍无法满足需求，原厂加速将产能从转向HBM和1c DRAM等高附加值产品。2) 需求端看：AI带来的爆发式需求是驱动本轮周期增长的核心因素，一方面服务器出货量提升且单台价值量显著增长，一方面服务器产能爆满导致其他存储产能被挤占，供应量缩减，驱动存储价格提升。

图表：2018-2026E年DRAM供需及产能利用率

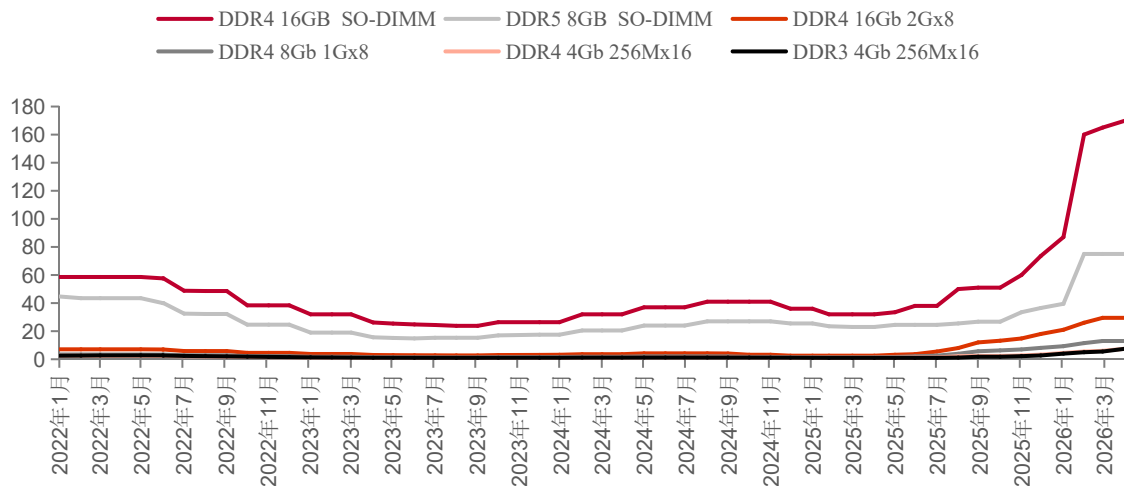


图表：2018-2026E年NAND供需及产能利用率

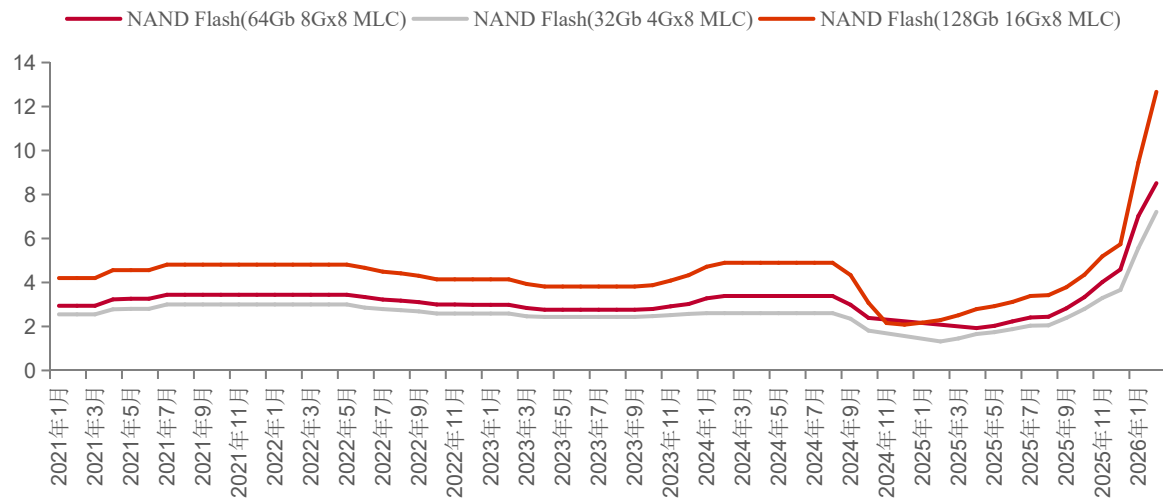


- 价格复盘：上一轮周期价格上涨4个季度后开始回落，本轮存储周期25Q2开始，目前仍处于涨价通道，周期明显拉长。
- 1) 21Q4-23Q2下行周期合约价格：DRAM、NAND价格自20Q4、21Q4开始下降，价格下降6个季度，价格下降到23Q2，价格跌幅约40%-55%、约20%。
- 2) 23Q3-24Q3上行周期合约价格：DRAM、NAND价格自23Q3开始全面上涨，价格上涨4—5个季度，到24Q3，DRAM、NAND价格自低点平均涨幅50-60%。此次是大容量NAND、DRAM(DDR5 LPDDR5)领涨。
- 3) 24Q3-25Q1下行周期合约价格：24Q3手机存储价格率先承压，24Q4 PC存储价格开始下降，25Q1 企业级存储价格开始下降，消费类价格跌幅30-40%左右，企业级存储价格相对坚挺。
- 4) 25Q2-至今新的上涨周期：DDR4、LPDDR4原厂减产，25Q2率先开启涨价，涨价幅度远超DDR5、LPDDR5和NAND，NAND中主要是MLC、SLC NAND涨价；25Q3开始存储全线涨价且涨幅扩大，除主流DRAM、NAND价格上涨外，利基DRAM、2D NAND、NOR Flash全面涨价；25Q3大型CSP厂商对26年Capex进行规划、制定关键器件的采购计划，开始与存储原厂谈26年长期供货协议（LTA），26年CSP存储需求超过供给，存储现货价开始飙升，25Q4存储合约价价格涨幅开始超预期。

图表：DRAM合约价变化（美元）



图表：NAND合约价变化（美元）



- 自25Q4合约价涨幅开始超预期，26Q1合约价涨幅最大，预计26年全年价格上涨，但涨幅逐季收窄。在行业供需紧张看到至少28年的背景下，对价格的跟踪焦点需要从“能涨多少”转为“高位预计维持多久”
- 25Q3大型CSP厂商对26年Capex进行规划、制定关键器件的采购计划，开始与存储原厂谈26年长期供货协议（LTA），26年CSP存储需求超过供给，存储现货价开始飙升，25Q4存储合约价价格涨幅开始超预期。26Q1是合约价涨幅最大阶段，DRAM、NAND平均环比价格翻倍，26Q2环比涨幅58%-75%，预计26Q3合约价高个位数涨幅，26Q4收窄。
- 合约价Q1：超预期。PC DRAM混合平均售价:从+50%~55%上调至+110%~115%；服务器DRAM混合平均售价:从+60%~65%上调至+93%~98%；移动DRAM混合平均售价:从+45~50%上调至+58~63%；整体DRAM混合平均售价:从+55~60%上调至+93~98%；eSSD价格:从+33~38%上调至+75~80%；TLC/QLC NAND价格:从+30~35%上调至+75~80%；整体NAND混合平均售价:从+33~38%上调至+85~90%。
- 合约价Q2：PC DRAM价格:从+10~15%上调至+40~45%；服务器DRAM价格:从+10~15%上调至+43~48%；移动DRAM(LP5X)价格:从+13~18%上调至+58~63%；eSSD价格:从+15~20%上调至+68~73%。NAND整体平均售价:从+18~23%上调至+70~75%，TLC/QLCNAND价格:从+15~20%上调至+60~65%。
- 合约价Q3：预计常规型DRAM涨幅+3%~+8%；NAND涨幅+5%~10%。但是由于原厂大幅缩减NAND供应、AI需求加剧NAND用量，预计26H2 NAND缺货状况加剧，价格有望超预期增长。
- 合约价Q4：预计常规型DRAM、NAND涨幅+0%~5%。

图表：NAND 2026价格预测

	26Q1F	26Q2F	26Q3F	26Q4F
eMMC/UFS	+55%~60%	+20%~25%	+5%~10%	持平
Enterprise SSD	+53%~58%	+15%~20%	+5%~10%	+0%~5%
Client SSD	+68%~73%	+20%~25%	+5%~10%	+持平
3D NAND wafer	+50%~55%	+15%~20%	+5%~10%	+0%~5%
TLC/QLC NAND	+75%~80%	+60%~65%	/	/
eSSD	+75%~80%	+68%~73%	/	/
Total NAND	+85%~90%	+70%~75%	+5%~10%	+0%~5%

图表：DRAM 2026价格预测

	26Q1F	26Q2F	26Q3F	26Q4F
PC DRAM	整体:+110%~115%	整体: +40%~45%	+3~8%	+0~5%
Server DRAM	整体:+93%~98%	整体: +43%~48%	+3~8%	+0~5%
Mobile DRAM	整体:+58%~63%	LPDDR5X:+58%~63%	LPDDR4X 持平; LPDDR5X +5%~10%	LPDDR4X 持平; LPDDR5X +3%~8%
Graphics DRAM	GDDR6:+45%~50%; GDDR7: +45%~50%	GDDR6:+0%~5%; GDDR7: +8%~13%	GDDR6:+0%~5%; GDDR7: +3%~8%	GDDR6:持平; GDDR7: +0%~5%
Consumer DRAM	整体+75%~80%	整体+45%~50%	DDR3 +0%~5%; DDR4 +0%~5%	DDR3 持平; DDR4 持平
Total DRAM	常规型+93%~98%	常规型+58%~+63%;	常规型+3%~8%;	常规型+0%~5%;

图表：25Q4、26Q1 DRAM、NAND合约价涨幅持续超预期

	25Q4					26Q1F					
发布日期	2025/9/25	2025/10/15	2025/11/8	2025/11/28	2025/12/31	2025/11/8	2025/11/8	2025/11/28	2025/12/31	2026/1/21	2026/3/3
eMMC/UFS		+20%~25%	+20%~25%	+20%~25%	+20%~25%	+15%~20%	+20%~25% (+5pct)	+20%~25%	+30%~35% (+10pct)	+55%~60% (+25pct)	
Enterprise SSD		+20%~25%	25%~30% (+5pct)	25%~30%	+25%~30%	+15%~20%	+20%~25% (+5pct)	+20%~25%	+33%~38% (+13pct)	+53%~58%(+20pct)	
Client SSD		+20%~25%	+20%~25%	+20%~25%	+20%~25%	+15%~20%	+20%~25% (+5pct)	+20%~25%	+40%~45% (+20pct)	+68%~73%(+28pct)	
3D NAND wafer		+35%~40%	+65%~70% (+30pct)	95%~100% (+30pct)	+145%~150% (+50pct)	+25%~30%	+15%~20% (-10pct)	+15%~20%	+30%~35% (+15pct)	+50%~55%(+20pct)	
Total NAND	+5%~10%	+20%~25%	+25%~30% (+5pct)	+25%~30%	+33%~38% (+8pct)	+18%~23%	+20%~25% (+2pct)	+20%~25%	+33%~38% (+13pct)	+55%~60%(+22pct)	+85%~90%(+30pct)

	25Q4						26Q1F				
	2025/9/24	2025/10/29	2025/11/8	2025/11/28	2025/12/31	2026/2/2	2025/11/8	2025/11/8	2025/11/28	2025/12/31	2026/2/2
PC DRAM		DDR4:+23%~28%; DDR5:+18%~23%	DDR4:+25%~30% (+2pct); DDR5:+25%~30% (+7pct)	DDR4:+43%~48% (+18pct); DDR5:+38%~43% (+13pct)	DDR4:+43%~48%; DDR5:+38%~43%	DDR4:+38%~43%(-5pct); DDR5:+38%~43%	DDR4:+10%~15%; DDR5:+10%~15%	DDR4:+18%~23% (+8pct); DDR5:+18%~23% (+8pct)	DDR4:+18%~23%; DDR5:+18%~23%	DDR4:+65%~70% (+47pct); DDR5:+50%~55% (+32pct)	DDR4:+105%~110% (+40pct); DDR5:+105%~110% (+55pct)
Server DRAM		DDR4:+20%~25%; DDR5:+15%~20%	DDR4:+30%~35% (+10pct); DDR5:+28%~33% (+13pct)	DDR4:+60%~65% (+30pct); DDR5:+53%~58% (+25pct)	DDR4:+60%~65%; DDR5:+53%~58%	DDR4:+53%~58%(-7pct); DDR5:+53%~58%	DDR4:+10%~15%; DDR5:+10%~15%	DDR4:+25%~30% (+15pct); DDR5:+23%~28% (+13pct)	DDR4:+18%~23% (-7pct); DDR5:+15%~20% (-8pct)	DDR4:+65%~70% (+47pct); DDR5:+60%~65% (+45pct)	DDR4:+88%~93% (+23pct); DDR5:+88%~93% (+28pct)
Mobile DRAM		LPDDR4X +18%~23%; LPDDR5X +20%~25%	LPDDR4X +38%~43% (+20pct); LPDDR5X +38%~43% (+18pct)	LPDDR4X +38%~43%; LPDDR5X +38%~43%	LPDDR4X +48%~53% (+10pct); LPDDR5X +43%~48% (+5pct)	LPDDR4X +48%~53%; LPDDR5X +43%~48%	LPDDR4X +8%~13%; LPDDR5X +8%~13%	LPDDR4X +15%~20% (+7pct); LPDDR5X +15%~20% (+7pct)	LPDDR4X +20%~25% (+5pct); LPDDR5X +20%~25% (+5pct)	LPDDR4X:+45%~50% (+25pct); LPDDR5X:+45%~50% (+25pct)	LPDDR4X:+88%~93% (+43pct); LPDDR5X:+88%~93% (+43pct)
Graphics DRAM		GDDR6:+13%~18%; GDDR7: +28%~33%	GDDR6:+28%~33% (+15pct); GDDR7: +28%~33%	GDDR6:+25%~30% (-3pct); GDDR7: +30%~35% (+2pct)	GDDR6:+25%~30%; GDDR7: +30%~35%	GDDR6:+25%~30%; GDDR7: +30%~35%	GDDR6:+3%~8%; GDDR7: +13%~18%	GDDR6:+20%~25% (+7pct); GDDR7: +20%~25% (+7pct)	GDDR6:+15%~20% (-5pct);	GDDR6:+45%~50% (+30pct); GDDR7: +45%~50% (+25pct)	GDDR6:+45%~50%; GDDR7: +45%~50%
Consumer DRAM		DDR3+18%~23%; DDR4 +20%~25%	DDR3 +20%~25% (+2pct); DDR4 +25%~30% (+5pct)	DDR3 +55%~60% (+35pct); DDR4 +45%~50% (+20pct)	DDR3 +55%~60%; DDR4 +45%~50%	DDR3 +55%~60%; DDR4 +45%~50%	DDR3+8%~13%; DDR4 +10%~15%	DDR3 +15%~20% (+7pct); DDR4 +15%~20% (+5pct)	DDR3 +15%~20%; DDR4 +15%~20%	DDR3 +45%~50% (+30pct); DDR4 +45%~50% (+30pct)	DDR3 +45%~50%; DDR4 +45%~50%
Total DRAM	常规型 +8~13%; HBM Blended +13~18%	常规型 +18~23% (+10pct); HBM Blended +23~28% (+10pct)	常规型 +30~35% (+12pct); HBM Blended +35~40% (+12pct)	常规型 +45~50% (+15pct); HBM Blended +50~55% (+15pct)	常规型 +45~50%; HBM Blended +50~55%	常规型 +45~50%; HBM Blended +50~55%	常规型 +8~13%;	常规型 +20~25% (+12pct);	常规型 +15~20% (-5pct);	常规型+55~60% (+40pct); HBM Blended +50~55%	常规型+90~95% (+35pct); HBM Blended +80~85%(+30pct)

■ DRAM现货价与合约价仍维持较大价差。

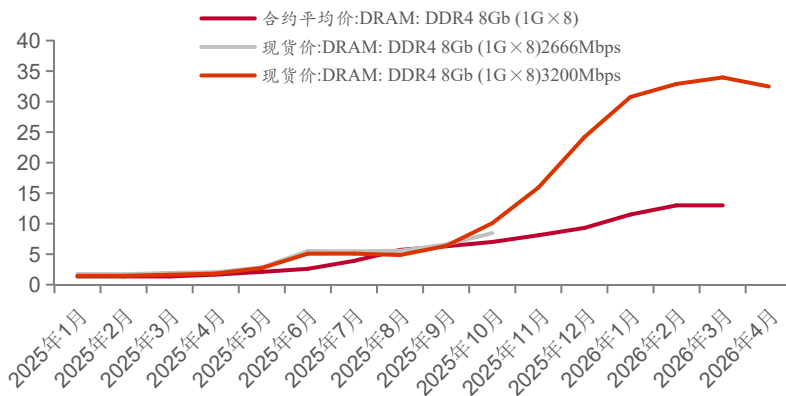
- 合约价以大客户为主，按周期议价并锁价，不受协议期间价格波动影响，占比较大；现货价反映市场即时供需情况，价格波动更大、占比较小。
- 现货价：25/6/2至26/5/12，DDR3 4Gb涨幅668%，DDR4 8Gb涨幅847%~1060%，DDR4 16Gb涨幅846%~1031%，DDR5 16Gb（2Gx8）涨幅624%。
- 合约价：25/6月至26/3月，DDR3 4Gb（256Mx16）涨幅650%，DDR4 4Gb（256Mx16）涨幅500%，DDR4 8Gb（1Gx8）涨幅400%，DDR4 16Gb（2Gx8）涨幅436%。
- 当前阶段，现货价加速上行，大幅推动合约价跟涨，但合约价一般滞后于现货价，两者仍维持较大价差。25年6月至26年3月DDR4 8Gb（1Gx8）、DDR4 16Gb（2Gx8）现货价涨幅1120%、1335%，合约价涨幅400%、436%，目前26年3月现货价与合约价价差分别约19、45美元。

注：价差为现货价减去合约价；现货价为每月月末价格。

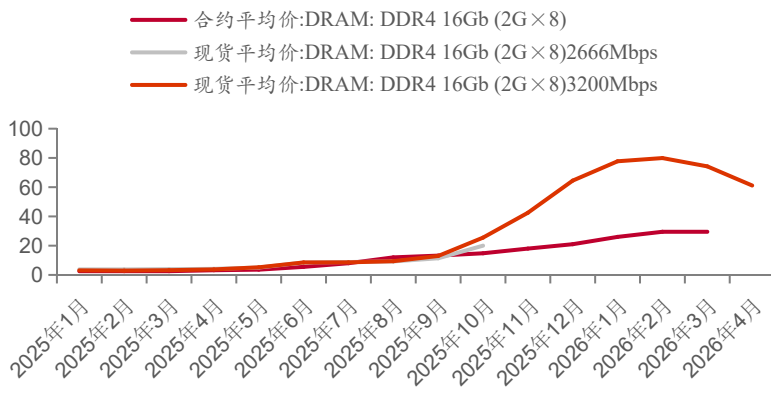
图表：DRAM合约价与现货价月度变化（美元）

时间	DDR4 8Gb 1Gx8				DDR4 16Gb 2Gx8					DDR5 16Gb 2Gx8				
	合约价	现货价	qoq	价差	合约价	qoq	现货价	qoq	价差	合约价	qoq	现货价	qoq	价差
		(3200Mbps)					(3200Mbps)					(4800/5600Mbps)		
2025年1月	1.35	1.46	0%	0.11	2.55	0%	3.11	-3%	0.56	3.75	-4%	4.69	-4%	0.94
2025年2月	1.35	1.45	-1%	0.1	2.55	0%	2.95	-5%	0.4	3.8	1%	4.89	1%	1.09
2025年3月	1.35	1.63	12%	0.28	2.55	0%	3.24	10%	0.69	4.25	12%	5.32	12%	1.07
2025年4月	1.65	1.82	12%	0.17	3.2	25%	3.76	16%	0.56	4.6	8%	5.52	8%	0.92
2025年5月	2.1	2.73	50%	0.63	3.6	13%	5.2	38%	1.6	4.8	4%	5.56	4%	0.76
2025年6月	2.6	5.08	86%	2.48	5.5	53%	8.53	64%	3.03	5.1	6%	6.08	6%	0.98
2025年7月	3.9	5.09	0%	1.19	8	45%	8.59	1%	0.59	5.25	3%	6.14	3%	0.89
2025年8月	5.7	4.86	-5%	-0.84	12	50%	9.38	9%	-2.62	5.25	0%	6.02	0%	0.77
2025年9月	6.3	6.36	31%	0.06	13.2	10%	12.96	38%	-0.24	6.1	16%	7.68	16%	1.58
2025年10月	7	10.08	58%	3.08	14.8	12%	25.5	97%	10.7	8.7	43%	15.49	43%	6.79
2025年11月	8.1	15.9	58%	7.8	18	22%	42.5	52%	24.5			27.17	75%	
2025年12月	9.3	24.18	52%	14.88	21	17%	64.49	52%	43.49			29.13	7%	
2026年1月	11.5	30.76	27%	19.26	26	24%	77.71	20%	51.71			37.17	28%	
2026年2月	13	32.9	7%	19.9	29.5	13%	79.91	3%	50.41			39.5	6%	
2026年3月	13	33.96	3%	19.2	29.5	0%	74.23	-7%	44.73			37.43		

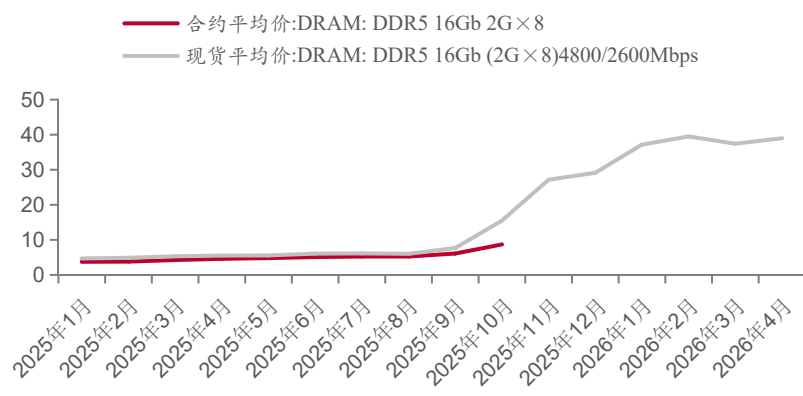
图表：DDR4 8GB 合约价与现货价（美元）



图表：DDR4 16GB 合约价与现货价（美元）



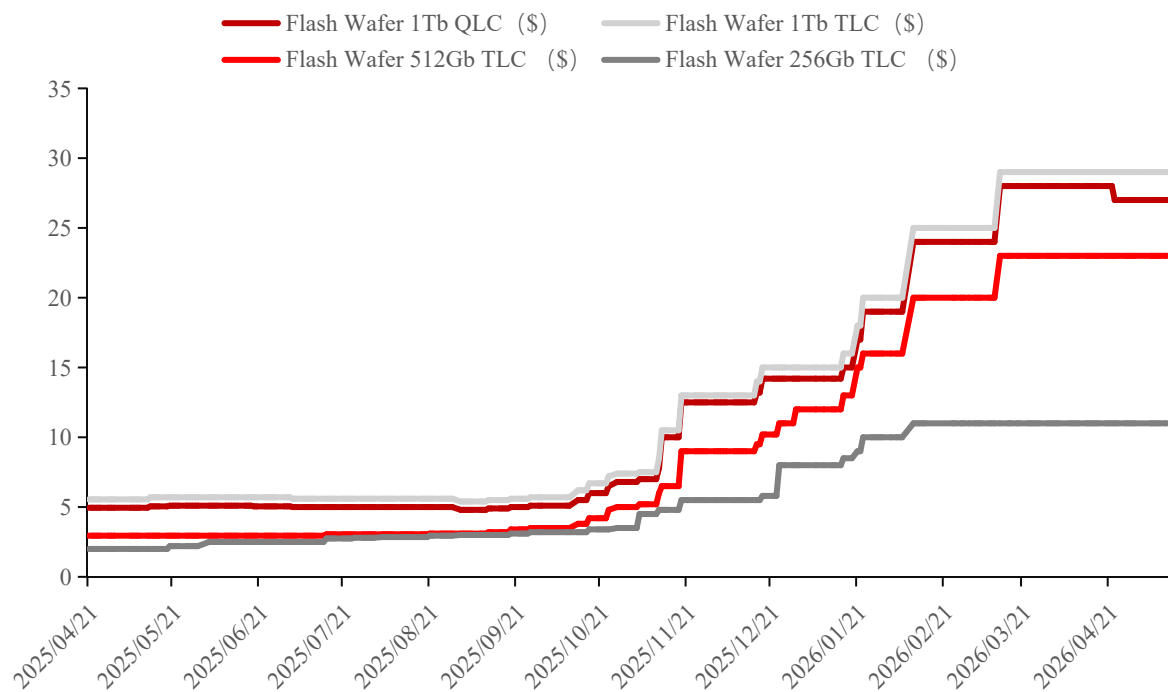
图表：DDR5 16GB 合约价与现货价（美元）



■ NAND价格25Q3开始进入快速上行通道，在供给受限+需求结构升级背景下，26Q1涨幅显著。

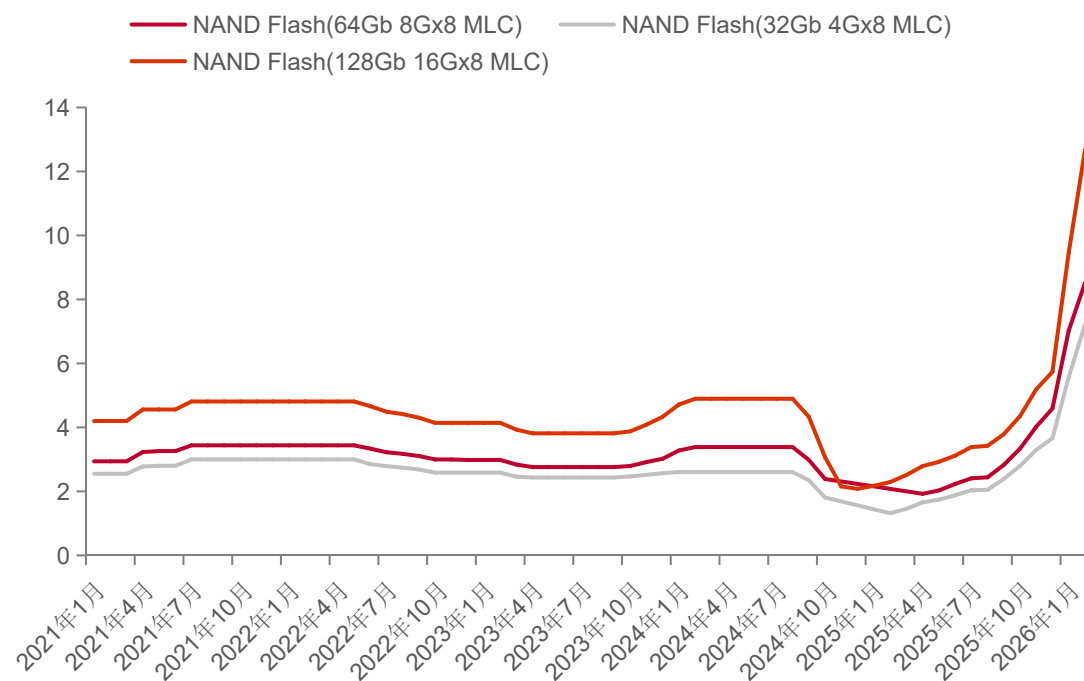
- 1) 现货价：25年9月初至26年5/12日，1Tb QLC、1Tb TLC、512Gb TLC、256Gb TLC现货价涨幅463%、437%、642%、267%，由于数据中心成为核心需求源，NAND成为算力体系的一部分，另外供给端头部大厂NAND产能收缩，偏向于通过升级增大位元供给，资金和产能转向高价值的通用DRAM、HBM，目前供需逐步转向全面结构型紧缺，26年1月至5/12日涨幅显著，分别达90%、93%、92%、38%。
- 2) 合约价：25年9月至26年2月，64Gb 8Gx8 MLC、32Gb 4Gx8 MLC、128Gb 16Gx8 MLC合约价涨幅201%、202%、235%。

图表：NAND现货价变化



来源：cfm

图表：NAND合约价变化（单位：美元）



来源：trendforce

- 存储的焦点已经从“看谁更便宜”，转向“看谁能拿到货”，因此客户与原厂在签订长协以锁定产能、保障供应。
- 目前虽然CSP客户仍在争取最优条款，但已经开始陆续签署3-5年长协。
 - 美光已经于Q1与某个CSP签订首个5年期战略供应协议，其他客户在洽谈中。
 - 三星已决定今年起将供货政策全面转向LTA体系，正与谷歌、微软深度谈判签订长协，微软拟向三星支付超100亿美元的巨额预付款，若采购方在协议期内未完成约定采购量，差额将从预付款中扣除。
 - SK海力士近期也正在推动与谷歌签署一份为期5年的通用DRAM长期供货协议，超出市场普遍预期的3年。
- 原厂与北美CSP厂商长协时间拉长至3-5年，锁定高底价。
 - 1) 合约期延长：过去长协通常是1-2年，目前合约期限延长至3-5年。
 - 2) LTA锁定高底价：闪迪力求将已上涨80%-100%的26Q2价格定为未来3年绝对“底价”。即便Q2暴涨后的价格无法完全作为底价，最终基准价也大概率落在Q1-Q2之间。
 - 3) 厂商将高额预付款作为分配产能的前提：部分厂商要求客户一次性支付一年（或更久）供货额的预付款，该资金构成刚性约束，若客户违约或不提货，预付款将直接作为违约金没收。
- LTA锁定未来3到5年盈利确定性：过去的LTA多为1年期“锁量不锁价”，需求走弱时客户大量违约，而存储厂商顾忌后续合作，缺乏实质惩罚机制。当前LTA谈判中议价权完全逆转，为存储厂商提供了前所未有的保护与利润保障。
 - 1) LTA提供关键资金来源。来自大型科技公司等客户的预付款将成为制造商未来进行投资、扩大产能的关键资金来源。
 - 2) LTA减少需求收缩冲击。订单断崖式下滑带来的风险被显著压缩，即便未来需求收缩，厂商依然可以依靠约定的价格和供货量来保障业绩。
 - 3) LTA提升资本开支效率。厂商避免无序扩产竞争，转向更精确测算后的设备投资。随着LTA占比持续提升，内存厂商的利润率未来有望维持在比过去更高、也更稳定的水平。
 - 4) LTA成为厂商筛选长期合作伙伴的工具。产能分配直接与客户历史履约记录挂钩。上一轮周期违约的客户当前面临严重供货限制，而信守合约的客户则获得全力保供支持。

图表：LTA的转变

维度	消费电子周期	AI基础设施周期
期限	1年	3-5年
价格机制	锁量不锁价，季度重新谈判	地板价锁定（26Q1-26Q2的平均价格）
违约成本	低	高额预付款
议价权	客户主导	供应商主导

目录

一、AI推理带来存储需求爆发和存储范式的改进

二、看未来2年供需持续紧张，原厂与客户签订长协

三、存储是AI硬件估值最低、业绩确定性最强的方向

存储是AI硬件中短期业绩确定性最强、估值最低板块

图表：AI核心硬件的盈利预测

所属板块	股票代码	公司名称	市值 (亿\$)	25年预期 EPS (\$)	26E预期 EPS (\$)	25PE	26E PE	26E利润增速
AI芯片	NVDA. O	英伟达	57,097	2.9	4.7	81.1	50.3	60%
	AVGO. O	博通	20,823	6.7	11.3	65.2	39.0	65%
	AMD. O	超威半导体	7,333	4.0	7.3	113.0	61.8	75%
	MRVL. O	迈威尔科技	1,638	1.6	2.8	116.9	64.4	81%
	MU. O	美光科技	8,751	8.1	58.8	96.2	13.2	609%
存储	000660. KS	SK海力士	8,777	39.8	185.7	32.3	6.7	346%
	005930. KS	三星电子	10,655	4.4	27.9	43.2	6.6	458%
	SNDK. O	闪迪	2,048	2.7	64.7	512.3	21.4	2063%
	285A. T	KIOXIA	1,530	3.5	6.1	89.1	46.8	97%
	STX. O	希捷科技	1,805	7.9	14.8	101.3	54.4	83%
互连网络	WDC. O	西部数据	1,686	5.3	9.9	93.1	49.3	101%
	ALAB. O	ASTERA LABS	392	1.8	3.0	128.3	76.3	63%
	CRDO. O	CREDO TECH	340	0.6	3.1	292.0	58.6	350%
	CSCO. O	思科	4,563	3.8	4.2	30.5	27.2	11%
	ANET. N	ARISTA网络	1,861	2.9	3.6	51.4	40.7	22%
液冷	688205. SH	德科立	62	0.2	0.3	186.0	144.6	309%
	VRT. N	VERTIV	1,445	4.1	6.4	91.4	58.6	53%
	002837. SZ	英维克	144	0.1	0.2	159.0	88.5	110%
电源	2308. TW	台达电	1,709	0.8	1.3	86.9	50.3	78%
	300870. SZ	欧陆通	65	0.4	0.6	134.5	98.5	78%
PCB	2383. TW	台光电子	524	1.3	2.8	110.3	51.6	115%
	002463. SZ	沪电股份	291	0.3	0.4	52.8	36.2	43%
光模块	COHR. N	COHERENT	792	3.5	5.5	115.5	74.2	55%
	300502. SZ	新易盛	891	1.4	2.6	63.9	34.5	84%
	300308. SZ	中际旭创	1,717	1.4	3.4	110.9	45.5	135%

备注：截至5月15日

■ 存储超级大周期，建议关注：

- 1) 弹性模组及主控：德明利、江波龙、佰维存储、大普微、联芸科技等。
- 2) 上游设计端：兆易创新、普冉股份、东芯股份、北京君正、澜起科技、聚辰股份、恒烁股份等。
- 3) 设备：微导纳米、拓荆科技、中微公司、精智达、华海清科、中科飞测、京仪装备、骄成超声、百傲化学、北方华创等。
- 4) 光刻机产业链：茂莱光学、汇成真空、波长光电、阿石创、联合化学、富创精密、永新光学等。

- 长鑫长存产能释放加剧竞争的风险。
- AI CAPEX不及预期的风险。
- 数据更新不及时，模型测算偏差风险。

重要声明

- 中泰证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。
- 本报告基于本公司及其研究人员认为可信的公开资料或实地调研资料，反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。本公司力求但不保证这些信息的准确性和完整性，且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，可能会随时调整。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。
- 市场有风险，投资需谨慎。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。
- 投资者应注意，在法律允许的情况下，本公司及其本公司的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。本公司及其本公司的关联机构或个人可能在本报告公开发布之前已经使用或了解其中的信息。
- 本报告版权归“中泰证券股份有限公司”所有。事先未经本公司书面授权，任何机构和个人，不得对本报告进行任何形式的翻版、发布、复制、转载、刊登、篡改，且不得对本报告进行有悖原意的删节或修改