东吴证券 SOOCHOW SECURITIES

格局落定，价值归真：从周期波动走向技术溢价

证券分析师：张良卫
执业证书编号：S0600516070001
联系电话：021-60197988

二零二六年二月六日

摘要：从传统周期到AI驱动的技术溢价与高成长性

- **增长逻辑重构：AI 带来的不是反弹，而是指数级增长的新需求，存储不再仅仅跟随宏观经济进行简单的库存周期波动，而是成为了AI算力的决定性瓶颈（存力决定算力的效率）。**
- **训练端（Scaling Laws 驱动）：** 模型参数量呈线性增长（如GPT-3到GPT-4增幅超9倍），直接拉动HBM（容纳参数）和SSD（存Checkpoint）的容量需求呈指数级爆发 。
- **推理端（商业化落地驱动）：** 随着RAG（检索增强生成）和超长上下文（Long Context）的应用，KV Cache（键值缓存）对显存的占用或持续增加，且用户越多、交互越深，对存储的带宽和容量需求就越大 。
- **估值体系重塑：量价齐升，高业绩持续兑现，行业正处于“营收高增”与“净利率扩张”的共振期 。**

量： AI训练与推理规模的扩大带来绝对出货量的增长。

价： HBM、企业级SSD等高技术含量产品占比提升，拉高了平均售价（ASP）。

利： 存储是重资产行业，营收覆盖固定成本后，新增的涨价红利将大比例转化为净利润，导致净利增速远超营收增速（高经营杠杆） 。

建议关注：闪迪（SNDK.O）、SK 海力士（000660.KS）、三星电子（005930.KS）、美光科技（MU.O）。

风险提示：下游AI资本开支（CAPEX）不及预期风险； 存储原厂产能扩张失控导致供需格局恶化风险；新技术研发进度或良率爬坡不及预期风险。

目录



1、AI价值链梳理

2、存储需求分级与存储硬件架构

3、大语言模型的存储需求

4、训练与推理端技术发展拉动存储需求增长

5、美光科技（MU.O）战略聚焦数据中心，HBM3E 技术突破

6、闪迪（SNDK.O）拆分后的价值重估，AI存储技术持续迭代

7、SK海力士享AI存储先发优势，巨头三星奋力追赶

8、估值情况分析

风险提示

1、AI价值链梳理

1.1 AI产业链梳理——硬件环节率先受益

东吴证券 SOOCHOW SECURITIES

图：AI产业链梳理

算力芯片：GPU（英伟达）、CPU（AMD）、ASIC（Google TPU） **存储（Memory/Storage）**：DRAM/SSD&HDD（SK 海力士、三星、Micron、Sandisk、西部数据、希捷科技） **网络通讯**：光模块（Coherent, Lumentum, 中际旭创）；交换机（Arista, Cisco, Nvidia）；连接芯片/技术（Broadcom、Marvell） **能源与配套**：液冷/散热（Vertiv, CoolIT）；电力/储能（Vertiv） **服务器整机**：Foxconn, Quanta, Supermicro	**核心硬件基础设施**	**卖铲子的角色，持续受益**
云厂商：Microsoft Azure（深度绑定Open AI）、AWS (Amazon)、Google Cloud (GCP)	**云算力平台**	**大规模投资，抢占卡位优势**
大模型：Open AI、Gemini（Google）、Claude（Anthropic）、Llama（Meta）、Qwen（阿里） **中间件及工具链**：Vector DB、LangChain、Hugging Face	**大模型与基础软件**	**持续探索商业化路径**
实际应用场景：**To B**：Microsoft Copilot (Office)、医疗 (药物研发), 法律 (合同审查), 编程 (GitHub Copilot), 客服 (自动应答) **To C**：ChatGPT, Claude、智能眼镜 (Meta Ray-Ban), 机器人 (Robotaxi, 人形机器人)	**商业化场景**	**应用盈利，反哺产业链上游投入**

1.2 竞争格局：存储呈寡头垄断特征，高壁垒支撑高盈利

东吴证券 SOOCHOW SECURITIES

图：AI硬件价值链竞争格局

算力芯片：GPU（英伟达）、CPU（AMD）、ASIC（Google TPU）
存储（Memory/Storage）：DRAM/SSD&HDD（SK 海力士、三星、Micron、Sandisk、希捷科技、西部数据）
网络通讯：光模块（Coherent, Lumentum, 中际旭创）；交换机（Arista, Cisco, Nvidia）；连接芯片/技术（Broadcom、Marvell）
能源与配套：液冷/散热（Vertiv, CoolIT）；电力/储能（Vertiv）
服务器整机：Foxconn, Quanta, Supermicro

核心硬件基础设施

卖铲子的角色，持续受益

算力芯片

毛利率：73%
净利率：50%
绝对寡头

芯片+DRAM价值量占比约55%

内存

硬盘

毛利率：51%
净利率：38%
细分垄断

硬盘价值量占比约2%
（加配占比则提升）

光模块及交换机

连接芯片/技术

光模块
毛利率：37%
净利率：24%
双寡头

交换机
毛利率：64%
净利率：40%
头部集中

连接技术
毛利率：65%
净利率：20%
头部集中

液冷/散热

电力/储能

毛利率：36%
净利率：17%
整合中

合计价值量占比约17%

服务器整机

毛利率：7%
净利率：4%
充分竞争

合计价值量占比约26%

注：（1）毛利率、净利率代表公司为英伟达（算力芯片）；SK 海力士（内存+硬盘）；中际旭创（光模块）；Arista（交换机）；Vertiv（液冷+电力）；工业富联（服务器整机）。选取2024-2025Q3分季度平均毛利率、净利率数据（部分公司财季划分不一致）（2）价值量拆分以机柜、电力及散热总量为基准，不含其他基础设施建设。

1.3 AI算力建设带动高端存储需求

东吴证券 SOOCHOW SECURITIES

图：AI硬件产业链——存储格局梳理

算力芯片：GPU（英伟达）、CPU（AMD）、ASIC（Google TPU）
存储（Memory/Storage）：DRAM/SSD&HDD（SK 海力士、三星、Micron、Sandisk、希捷科技、西部数据）
网络通讯：光模块（Coherent, Lumentum, 中际旭创）；交换机（Arista, Cisco, Nvidia）；连接芯片/技术（Broadcom、Marvell）
能源与配套：液冷/散热（Vertiv, CoolIT）；电力/储能（变压器、UPS电源、备用发电机）
服务器整机：Foxconn, Quanta, Supermicro

毛利率：51%
净利率：38%
细分垄断

注：毛利率、净利率代表公司为SK海力士（内存+硬盘）选取2024-2025Q3季度平均数据

内存

GDDR
一般对应GPU
（显存）

HBM（3、3e/4）高带宽内存、价格高主要用于AIDC
GDDR（6/7）标准显存，便宜成熟主要用于推理卡、边缘AI设备

DDR
一般对应CPU
（“内存”）

DDR5 (高性能/AI标配/高价值)
LPDDR5（边缘计算/汽车电子）
DDR4 (性价比/存量/正在被替代)
DDR3 (网络设备/存量市场)

25Q3：市占率
SK 海力士 33%
三星 33%
美光 26%

25Q3：营收环比增速
SK 海力士 +12%
三星追赶 +30%
美光稳定 +53%
DRAM整体 +31%

AIDC明显拉动

硬盘

SSD
固态硬盘
（电信号）

SLC:速度快，寿命长，容量小，贵
MLC:性能均衡，被TLC取代
TLC:速度寿命适中，成本可控
QLC:容量大、寿命短、写入慢，便宜

25Q3：市占率
三星 35%
SK 海力士 27%
美光 14%
凯侠 14% 闪迪4%

25Q3：营收环比增速
SK 海力士 +27%
三星 +29%
美光 +26%
凯侠 +30% 闪迪+26%
Top 5 +28%

HDD
机械硬盘
（磁信号）

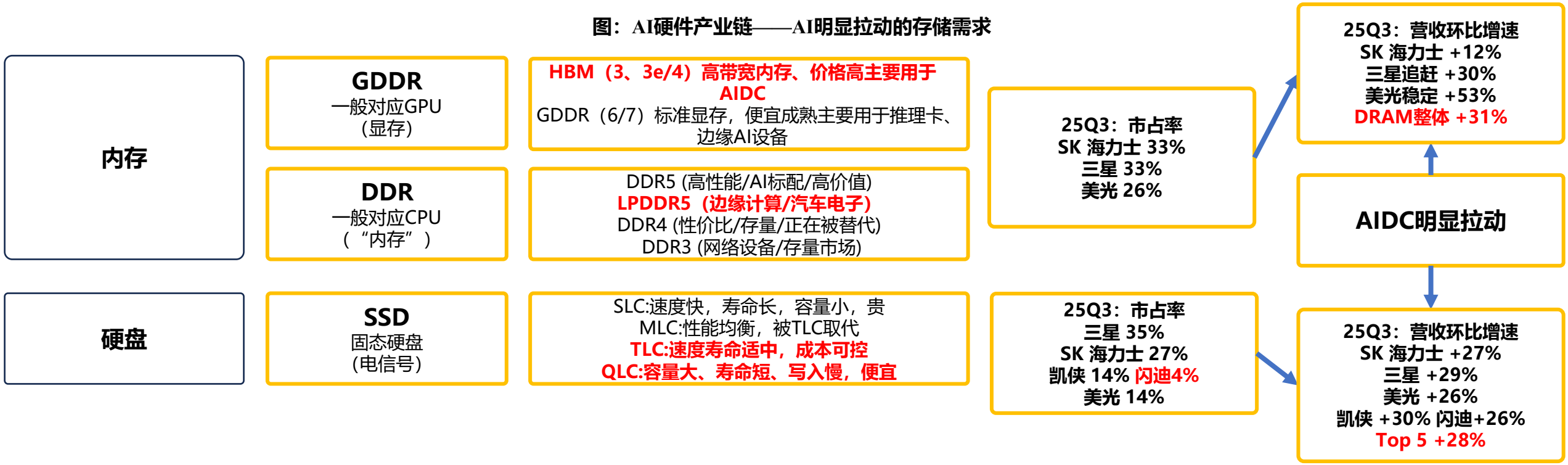
相较于SSD有成本优势，存储寿命长

资料来源：TrendForce，evertiq，东吴证券研究所制作

1.4 架构分级：依数据热度分层，确立三级存储架构模式

- **AI对存储的需求基本可以分为三大维度：冷数据、温数据、热数据。**
- **Storage: SSD+HDD**
  - 冷数据：对传输速度要求较低，用柜外硬盘存储，例如HDD/SSD-QLC，容量大，价格较低。
  - 温数据：对传输速度要求适中，用柜内硬盘存储，例如SSD-TLC，速度价格适中。
- **Memory: DRAM(HBM+DDR)**
  - 热数据：对传输速度要求极高，用柜内内存存储，例如HBM3e、LPDDR5X，速度快价格高。

图：AI硬件产业链——AI明显拉动的存储需求

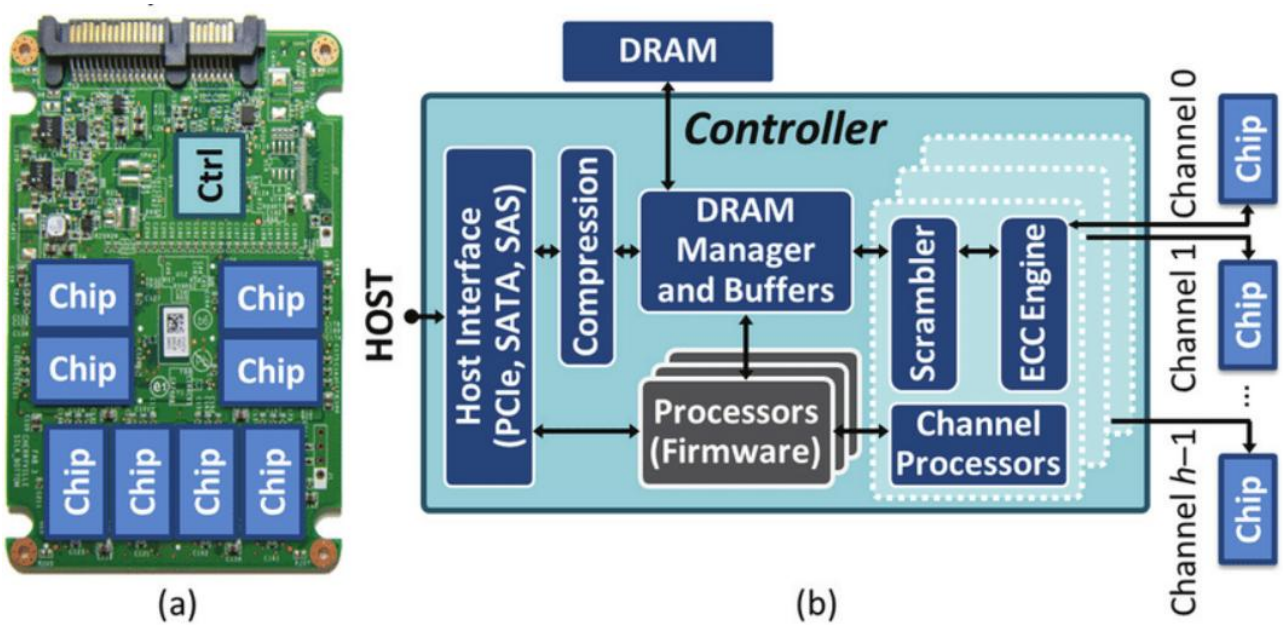


资料来源：TrendForce，evertiq，东吴证券研究所制作

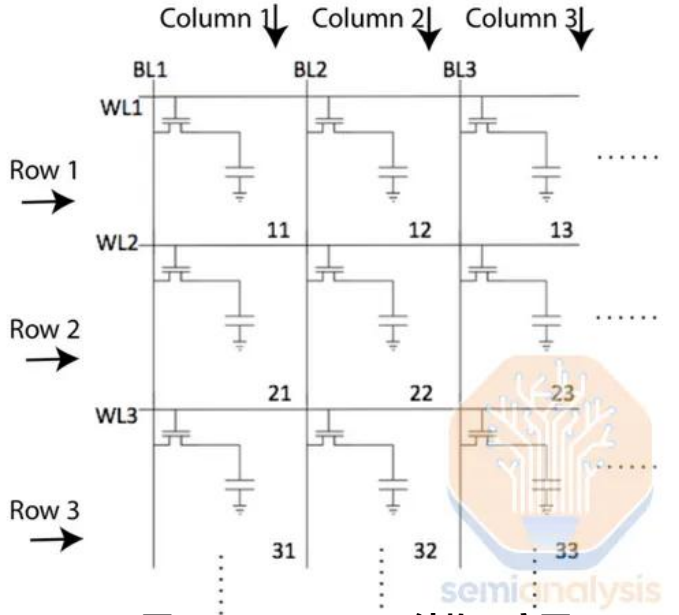
1.5技术路径：内存侧重快读写，持久存储聚焦大容量低成本

- **Memory和Storage的区别：**
- **Memory：**
  - 以DRAM（HBM、LPDDR)等为代表，读写快、断电后不保存数据。
  - 用晶体管+电容（1T1C结构）存储数据，每个单元1bit数据。
- **Storage：**
  - 以SSD(NAND FLASH)、HDD为代表，读写较慢，断电后保存数据。
  - 用浮栅晶体管存储数据，每个单元（1-4bit数据）

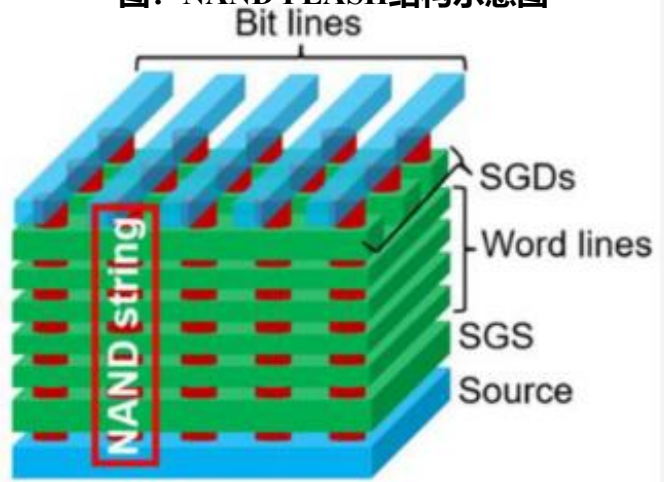
图：SSD系统图示



图：DRAM结构示意图



图：NAND FLASH结构示意图

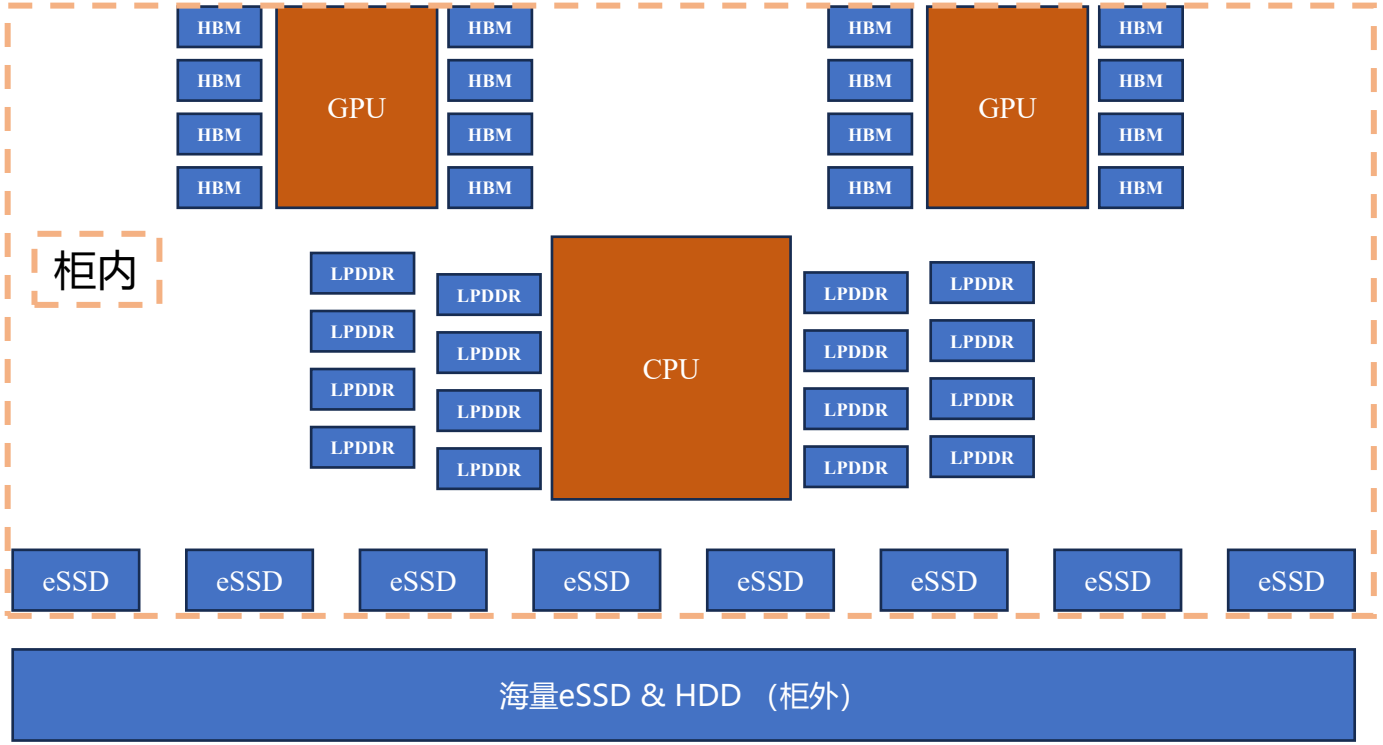


2、存储需求分级与存储硬件架构

2.1 模型适配：数据流转明确硬件分工，全流程拉动存储需求

- **AI对存储的需求本质上可以分为三部分：容纳参数；存放过程；数据注入。**
- **存储物理分层可以分为四部分：HBM、LPDDR、SSD、柜外SSD&HDD。**
  - 容纳参数（存放模型本身权重Weights）：HBM。
  - 存放过程数据(KV cache等)。HBM+LPDDR+SSD，KV cache主要存放HBM+LPDDR内，训练过程中的check point会保存至SSD。
  - 数据注入（存放及预热训练需要的原始数据，注入给HBM或LPDDR）：SSD+柜外SSD&HDD。

图：HBM、LPDDR、eSSD、eSSD&HDD部件图示（以NVIDIA GB200为例）



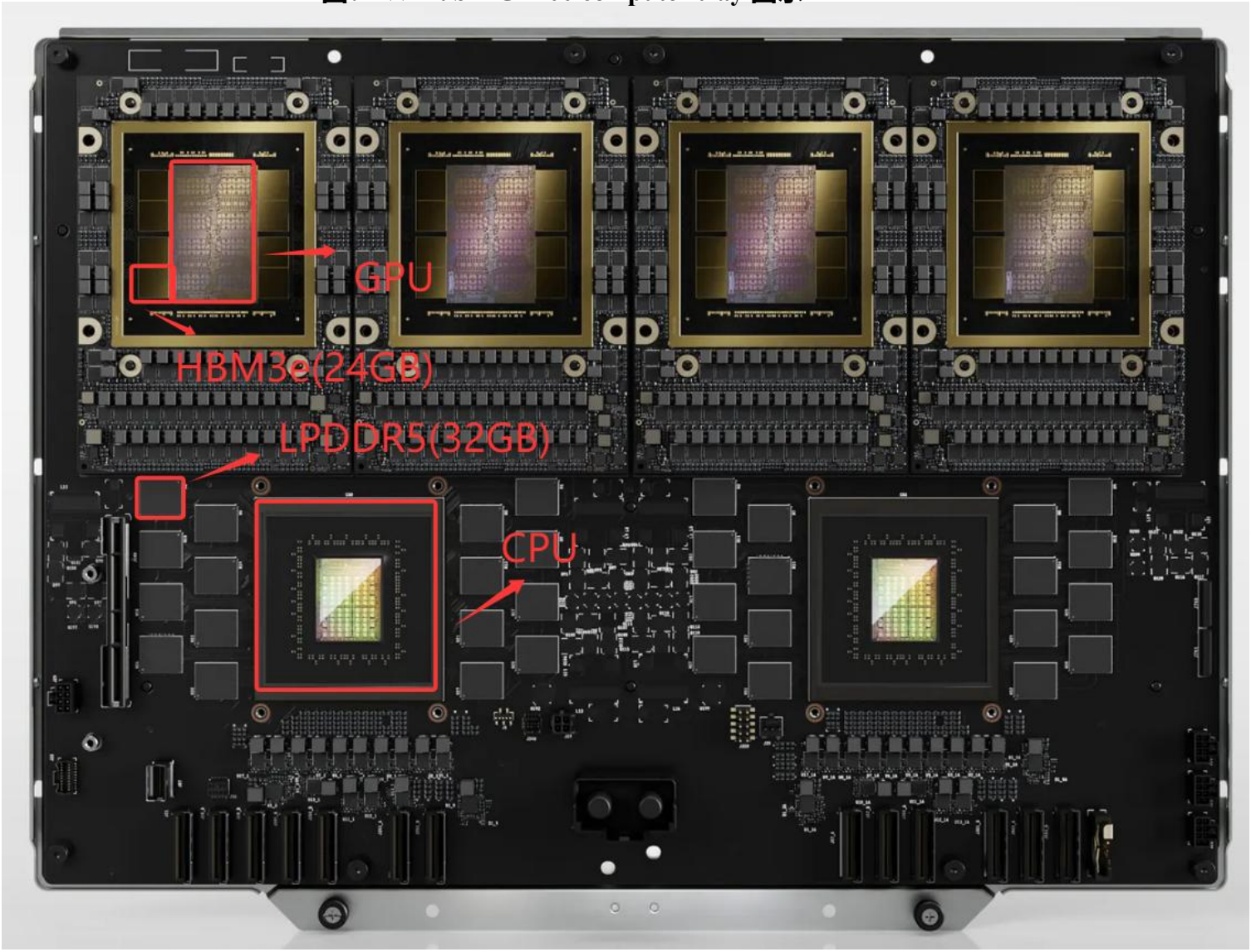
表：HBM、DDR、SSD、HDD&SSD四级存储在训练、推理阶段作用

硬件类型	核心定位	训练阶段作用	推理阶段作用
HBM	计算层	存放模型参数、梯度、优化器状态、中间激活值	存放模型权重、KV Cache
DDR	缓存层	CPU 数据预处理、ZeRO-Offload 卸载区	超大模型无法全入显存时的后备存储，或 CPU 推理内存
SSD	快存	快速保存断点（Checkpoint）、高速向内存输送训练数据	冷启动（Cold Start）时的模型加载、动态加载 LoRA 适配器
HDD&SSD	冷存	海量原始数据集（Raw Data）存储、历史模型归档	推理日志记录、长尾数据归档

2.2 NVIDIA GB200异构存储设计，实现容量与带宽的性能互补

- 在NVIDIA GB200 NVL72/36中，单个计算托盘包含2个GB200 Superchips，每个GB200Superchip包含1个Grace CPU和2个Blackwell GPU。还构建了“HBM + LPDDR + SSD”的三级存储阶梯，实现了带宽高低搭配与容量的互补。
- **HBM3e：**每个GPU配备8颗HBM3e（24GB）内存，单GPU容量达192GB，单计算托盘总容量768GB
- **LPDDR5X：**每个CPU配备16颗LPDDR5X（32GB），单CPU容量达512GB（480GB可用，32GB为备用），单计算托盘总容量1024GB。
- **SSD：**每个计算托盘标配4 * 3.84TB E1.S NVMe SSD，可扩展至8个接口，单盘可扩张至7.68TB，单计算托盘总容量15.36TB/61.44TB（单GPU对应3.84TB/15.36TB）。

图：NV Rubin GB200 computer tray 图示



2.3 LPDDR5X作为高性价比显存补充，平衡带宽成本与容量

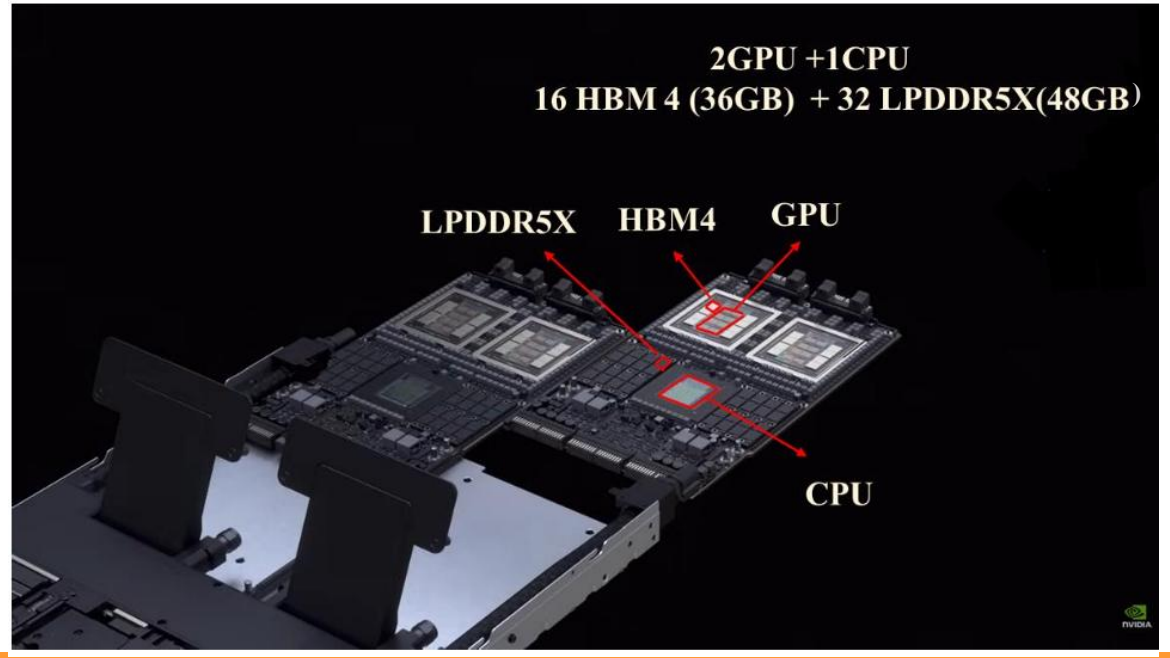
- **LPDDR5X与HBM核心差异在于带宽。**
- LPDDR5X服务于CPU，用于运行操作系统、驱动程序、数据加载器；GPU还可通过NVLink-C2C直接访问这部分内存，存放模型或context。
- HBM3e服务于GPU，高带宽的特性，为GPU提供充足数据，用于存储正在计算的权重（Weights）、KV cache和激活值（activations）。
- 在NVLink 72 GB200中，有18个计算托盘，其中每个托盘有32颗LPDDR5X，32颗HBM3e；单托盘LPDDR5X带宽1092.2GB/s，HBM3e带宽32768GB/s。NVLink 72 GB200机柜中LPDDR5X带宽19.2TB/s，HBM3e带宽576TB/s。
- **所以在实际计算中，LPDDR作为HBM的备用空间，HBM将一些低频KV cache下放至LPDDR内，NV Rubin还会将更低频的KV cache下放至柜内SSD，以拓展存储空间。**

表：LPDDR5X与HBM3E参数特性对比

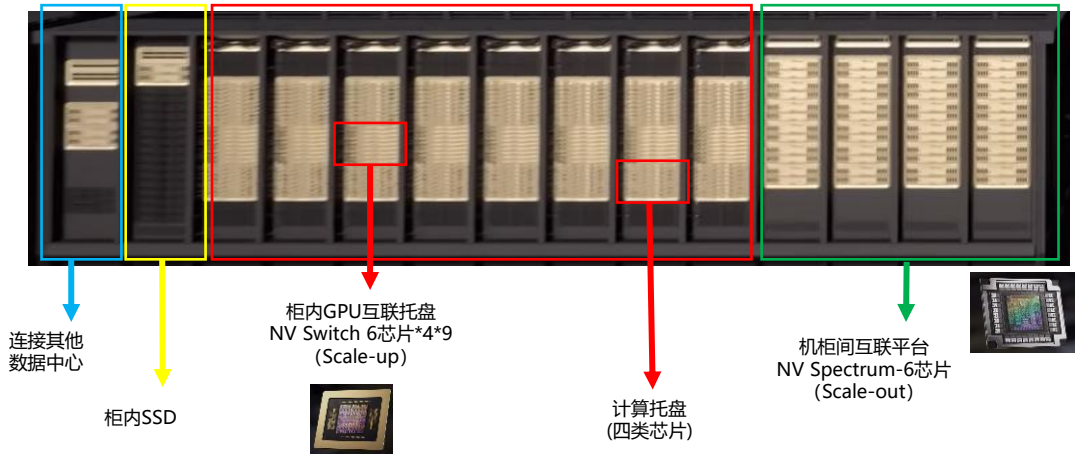
特性	LPDDR5X	HBM3e
核心优势	低功耗、低成本、大容量	大带宽、高密度
单颗带宽	(位宽*等效速率)/8(换算为Byte)*2通道 =**34.132GB/s**	(位宽*等效速率)/8(换算为Byte)=**1024GB/s**
等效速率	8533MT/s	8000MT/s
位宽	16-bit(双通道)	1024-bit
封装方式	平面封装，焊接在PCB板上	3D堆叠 + CoWoS封装
成本	相对较低	价格明显较高

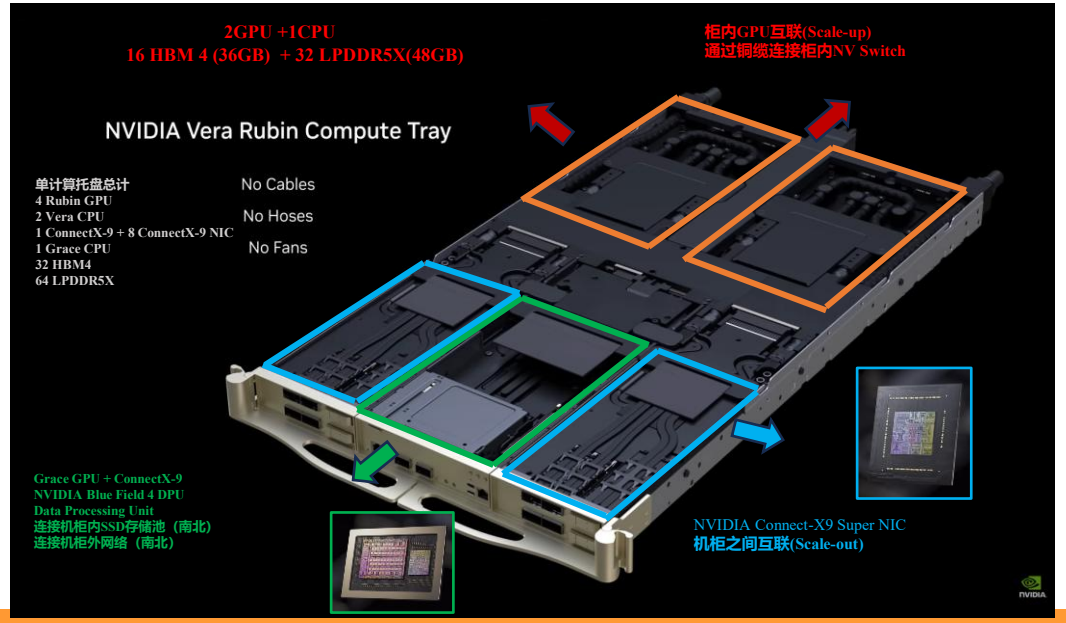
2.4 Rubin架构下互联升级，存储密度稳步提升

- **在Rubin架构中，柜内存储可分为三级**
- 第一级HBM4，HBM与GPU通过CoWos封装，物理距离最近，且HBM带宽最宽。单GPU容量达288GB，单计算托盘容量达1152GB；
- 第二级LPDDR5X，焊接在CPU旁的PCB板上；供CPU及GPU调度使用，带宽中等，相较于HBM差距明显，单CPU容量达1536GB，单计算托盘容量达3072GB；
- 第三级为柜内SSD，通过BlueField 4 DPU调度互联，为CPU、GPU较低速传输数据。单计算托盘容量达64TB（2026CES发布会口径）。



图：NV Rubin POD及Computer tray图解
（基于2026CES发布会推测，最终规格以官方发布为准）





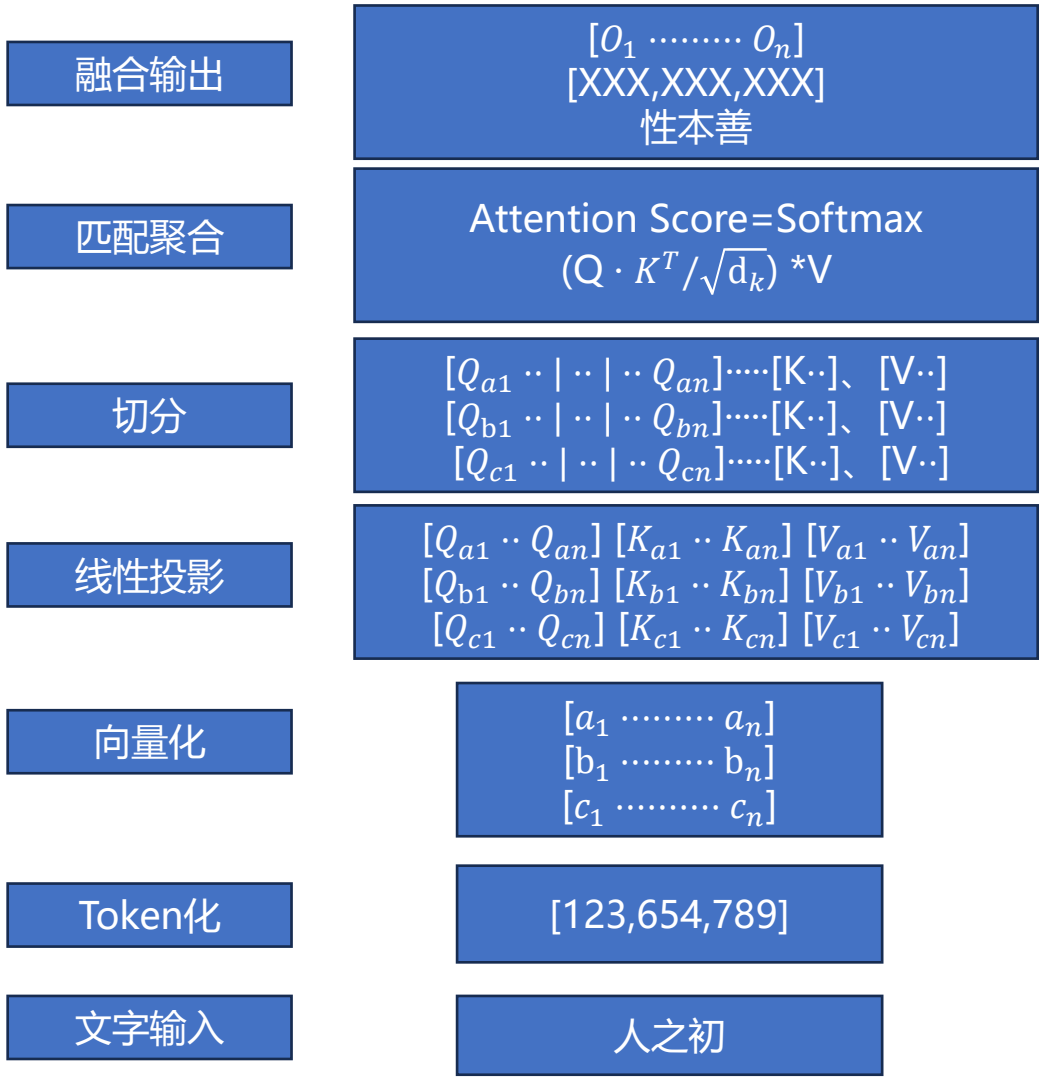
3、大语言模型的存储需求

3.1 transformer架构数据流转频繁——存力决定算力效率

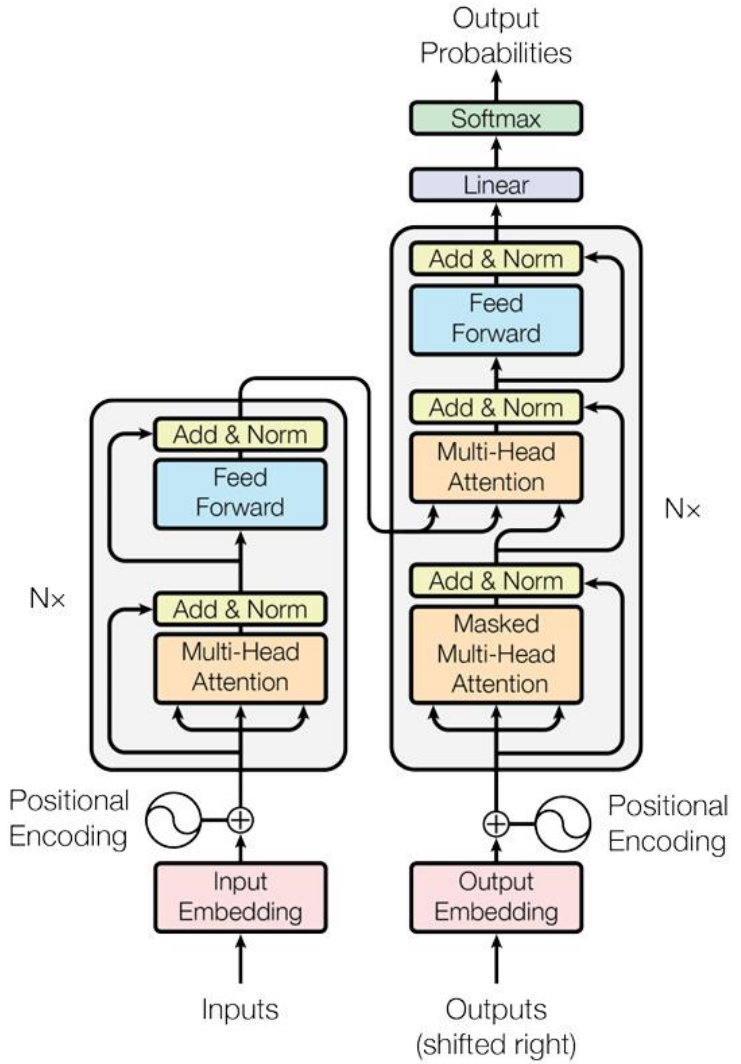
东吴证券 SOOCHOW SECURITIES

- **Transformer 数据流转：**
- 模型无法直接理解文字（如“人之初”），必须经过 Token 化（转数字 ID）和向量化（Embedding），将离散的文字转化为计算机可运算的连续高维向量；
- 通过线性投影，将输入向量转化为 Query（查询）、Key（键）、Value（值）三个矩阵；
- 再把大维度的向量拆分成多个子空间，让模型能同时捕捉句子的不同特征（如语法结构、语义关联等）；
- 最后通过计算注意力分数（Attention Score），衡量词与词之间的相关性，再加权融合 Context Vector，最终经过 Softmax 输出预测结果（“性本善”）

图：基于transformer大语言模型数据处理流程示意图（参数仅为示意，不代表具体含义）



图：transformer 模型图示

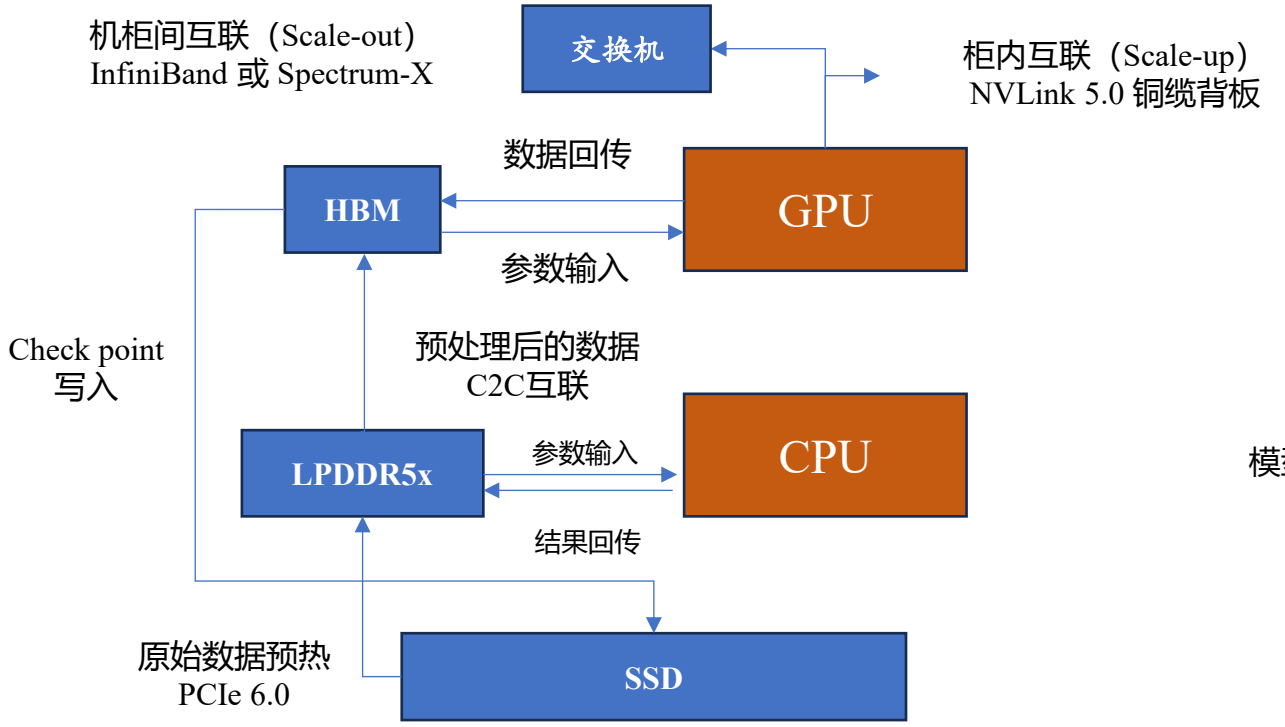


资料来源：《Attention is all you need》，东吴证券研究所制作

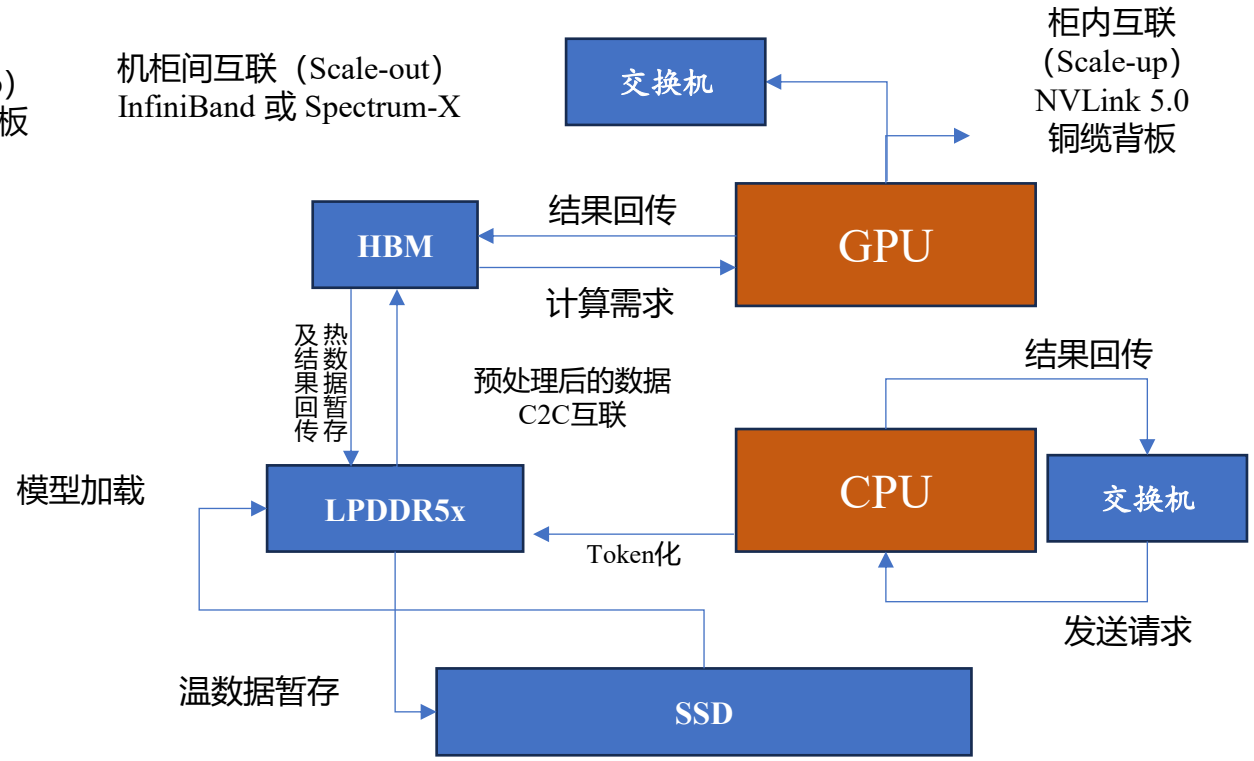
3.2 大语言模型分训练、推理数据流程——存力决定算力效率

- **模型训练时：**模型由SSD经过PCIe 6.0流向LPDDR5x，由CPU进行Token化，再回传给LPDDR5X，通过C2C互联传给HBM作为参数输入给GPU，GPU训练完成后回传结果给HBM，再写入SSD保存模型或周期性保存checkpoint。
- **模型推理时：**模型由SSD经过PCIe 6.0流向LPDDR5x，用户从网络端发送请求至交换机，由CPU进行Token化传输给LPDDR5X后，通过C2C互联连接HBM向GPU发送计算请求，GPU计算完成后传回HBM，HBM再传回LPDDR5X，再通过CPU回传给交换机，再将结果给到用户。

图：大语言模型训练阶段数据流示意图



图：大语言模型推理阶段数据流示意图



3.3 容量测算：解构模型显存消耗或推动需求呈指数级增长

- **AI模型的生命周期主要分为训练（Training）和推理（Inference）两个阶段。这两个阶段对显存有着截然不同的要求。**
- **“16字节法则”混合精度模型训练下显存消耗：**
  - **静态显存需求：**参数（FP16）2Bytes；梯度（FP16） 2Bytes；Adam优化器（FP32）需要维护参数的备份(4B) + 动量(4B) + 方差(4B)=12Bytes；故万亿参数的模型需要16TB级别显存才可存储其静态存储；
  - **动态显存需求：**模型状态也需占用显存空间，包括激活值及临时缓冲区，其中激活值大小与模型规模（层数 L、隐藏层维度 d_model ）、序列长度 (S) 和 训练批次大小 (B) 呈线性正相关。
- **推理显存消耗：**
  - **静态显存需求：**参数（FP16）2Bytes；
  - **动态显存需求：**推理中生成的过程数据KV cache需占用大量显存空间，目前主流大模型采用GQA算法，将多个Query头分成一组，每组共享一对Key/Value头，以Llama-3-70B为例，采用8：1的GQA分组，将KV cache的显存占用降低了8倍。KV cache=2 * L * N_{KV} * d_{head} * P_{bytes} * Batch Size * Context Length(2是Key Value两个矩阵，L为模型层数，N_{KV}是KV头数，d_{head}是模型注意力头维度，P_{bytes}是数据精度，Batch Size是并发请求数量，Context Length是序列长度)。GQA算法降低N_{KV}，使得显存需求得到一定缓解，但由于实际计算核心SRAM的空间极小，每Token计算需重新读取模型，大规模推理任务下，显存带宽仍是核心的限制因素。

表：INT8、FP16、FP32精度下存储比特结构

特性	INT8	FP16	FP32
十进制有效数字	仅整数	3~4位	7~8位
占用位宽	8-bits	16/bits	32bits
比特结构	$b_7b_6b_5b_4b_3b_2b_1b_0$	$b_s\|b_{E5}\cdots b_{E1}\|b_{F10}\cdots b_{F1}$	$b_s\|b_{E8}\cdots b_{E1}\|b_{F23}\cdots b_{F1}$
结构	1位符号、7位指数	1位符号、5位指数、10位尾数	1位符号、8位指数、23位尾数
计算公式 （注：其中V为十进制数值；i为从右至左的序号，且从0开始计；F为尾数计算结果；E为指数计算结果）	$V=-128\cdot b_7+\sum_{i=0}^{i} b_i\cdot 2^i$	$V=(-1)^s\cdot(1+F)*2^{E-15}$ 其中$F=\sum_{i=1}^{10} b_{10-i}\cdot 2^{-i}$	$V=(-1)^s\cdot(1+F)*2^{E-127}$ 其中$F=\sum_{i=1}^{23} b_{23-i}\cdot 2^{-i}$
示例	0 1111111=127	0 10001 1010000000 = 6.5	1 10000011 10100110000000000000000 = - 26.375

3.4 Scaling Laws下参数量持续膨胀，拉高存储容量需求下限

- **模型大小是怎么计算的？**
- **1）先从公司自身算力预算出发，计算出用于训练的算力。**
- 算力计算方法：总算力（FLOPs）=显卡数量*单卡峰值算力*训练时长*有效利用率。
- **2）根据Chinchilla Scaling Laws，C≈6*N*D**（C为算力，N为参数量，D为数据量，6为每个Token在训练过程中通常需要6次浮点运算，前向传播 2 次 + 反向传播 4 次），D≈20*N，C≈120*N^2。
- **3）再根据计算得到的参数量来计算出层数L和维度 d_model （N≈L*12* d_model ²）**
- **4）由模型总维度d_model可以推算出*N*_heads 数量**（d_model= *N*_heads ∗d_(heads_model)）（其中d_(heads_model)为注意力头维度,目前标准维度为128）
- 以GPT-3为例，1750亿参数，模型总维度为12288(模型注意力头为96个，注意力维度为128维)，模型层数为96层。

表：参数量计算拆分（基于1750亿参数GPT-3 ）

	计算公式	实际数值	参数量
Embedding	d_embed * n_vocab	12288 * 50257	617,558,016
Key	d_query * d_embed * n_heads * n_layers	128 * 12288 * 96 * 96	14,495,514,624
Query	d_query * d_embed * n_heads * n_layers	128 * 12288 * 96 * 96	14,495,514,624
Value	d_value * d_embed * n_heads * n_layers	128 * 12288 * 96 * 96	14,495,514,624
Output	d_embed * d_embed * n_heads * n_layers	12288 * 128 * 96 * 96	14,495,514,624
Up-projection	n_neurons * d_embed * n_layers	49152 * 12288 * 96	57,982,058,496
Down-projection	d_embed * n_neurons* n_layers	12288 * 49152 * 96	57,982,058,496
Unembedding	n_vocab * d_embed	50257 * 12288	617,558,016

资料来源：《Language Models are Few-Shot Learners》，《Training Compute-Optimal Large Language Models》，《Scaling Laws for Neural Language Models》，东吴证券研究所

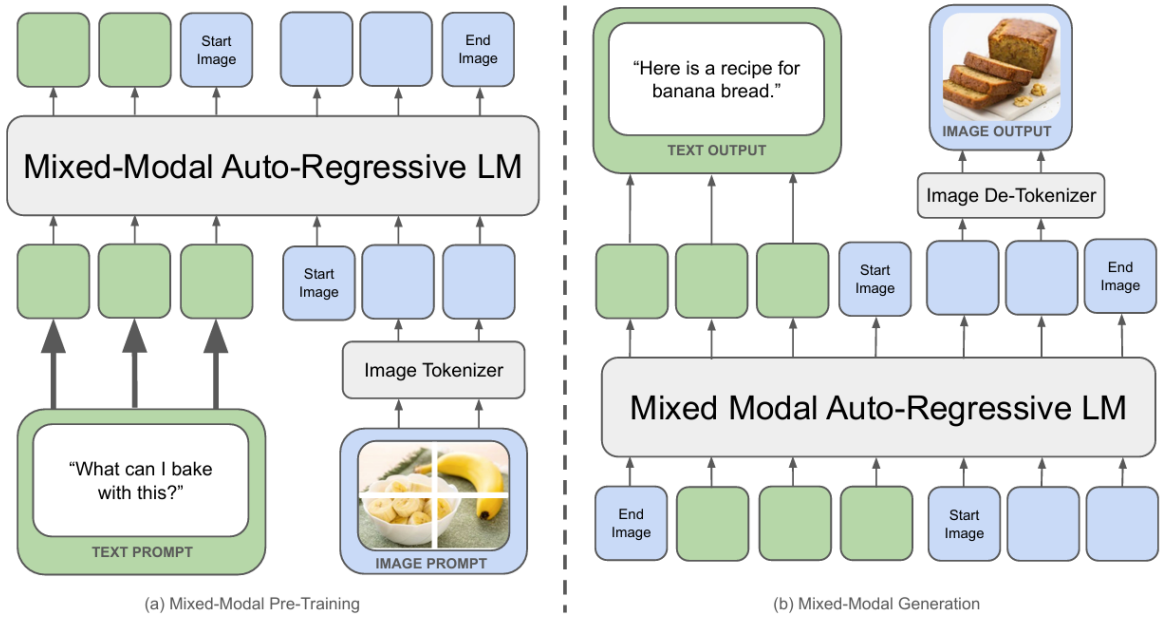
4、训练与推理端技术发展拉动存储需求增长

4.1训练端：模型升级驱动 HBM 及 SSD 的容量扩张

- **大模型在训练阶段对存储的拉动：**
- **更大模型——更大的显存及SSD空间需求**
- 模型参数量的持续扩张是大语言模型发展的主线。这种线性地增长是提升模型能力的核心路径，2020年，GPT-3参数量为1750亿，2023年3月，GPT-4参数量约为1.8万亿。增幅超9倍。在模型参数增长的背景下，为存放下更多参数，对存储（NAND+DRAM）的需求也持续增长。
- **MoE架构——更大显存空间需求**
- 在MoE架构下通过稀疏激活机制，实现了“大模型的容量，小模型的算力”，但模型需全量存储在显存中，本质上是用存储空间换取训练速度，对显存需求拉动明显。
- **长上下文——更大的带宽需求**
- 2020年到2026年GPT-3的上下文长度为2,000 Token，GPT-5.2上下文长度约为40万Token，相差近200倍。更长的上下文长度主要冲击的是显存中的动态部分（激活值），激活值的大小和上下文长度成正比，对显存空间及带宽提出更高要求。

- **多模态模型——更大的SSD空间需求**
- 原生多模态大模型训练，是指从底层架构设计和预训练阶段开始，就将多种模态（文本、图像、视频、音频）作为统一的数据流进行处理的模型。多模态数据被转化为特征向量，模型序列长度激增，对HBM的需求明显增加；多模态数据本身占用空间就较大，预处理后的向量所占空间会进一步加大，从而拉动SSD的容量需求。

图：原生大语言模型训练、推理示意图

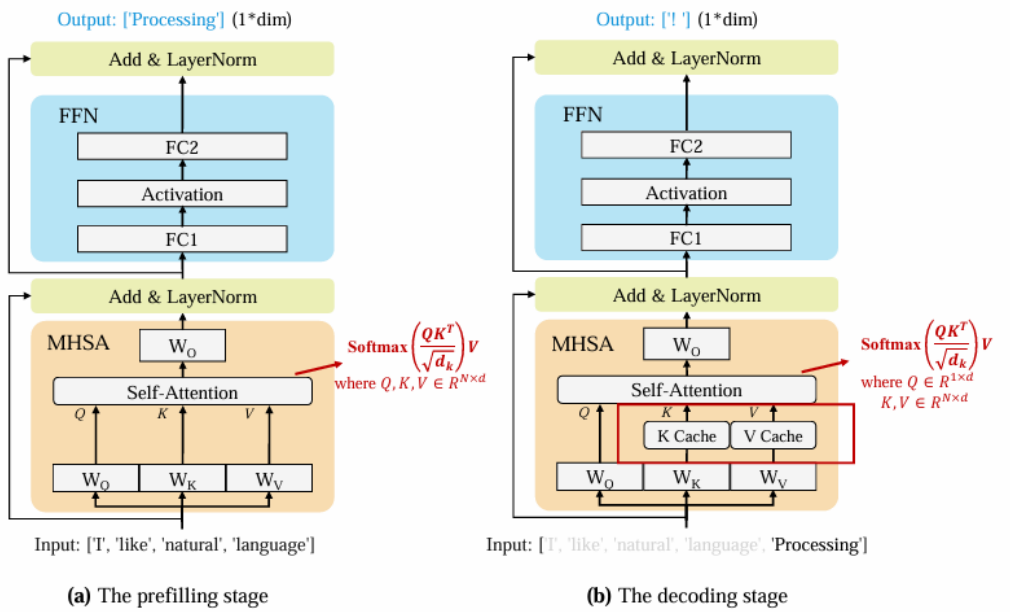


资料来源：《Chameleon: Mixed-Modal Early-Fusion Foundation Models》，《Language Models are Few-Shot Learners》，东吴证券研究所

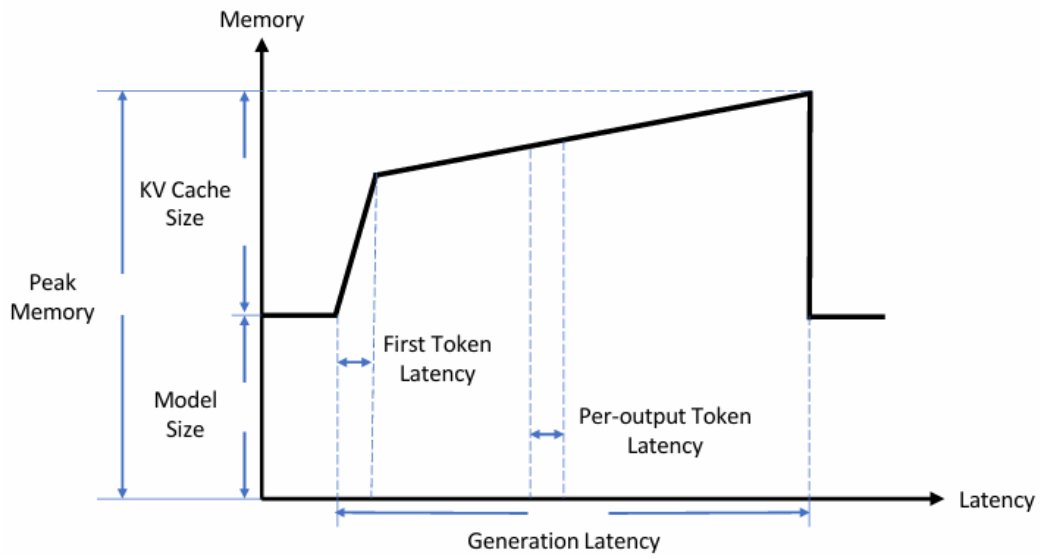
4.2推理端：交互深化与长文本，推升显存容量、带宽需求

- **大模型在推理阶段对存储的拉动：**
- 大模型推理阶段，显存空间主要为存储模型参数及KV cache，且主要受三大技术方向拉动。
  - 一是检索增强生成（RAG），此技术是将海量外部非结构化数据转化为高维向量并注入 Prompt，大量Token被一次性读入，瞬间生成的大量 KV 向量迅速挤占显存；
  - 二是超长的思维链（CoT）会生成大量的中间推理Token，同样占用大量的KV cache空间，对显存空间提出更严格的要求；
  - 三是超长的上下文长度产生的KV cache其规模或超过模型权重本身，大规模的KV cache对显存容量、带宽的要求都是极高的。
- 且推理直接面向消费者，厂商需根据峰值储备算力及存力（即在用户最集中的高峰期也能保证服务的提供），随着大语言模型的用户数、使用频率的持续增长，这同样对存储的需求有巨大的拉动效应。

图：预填充阶段 (a) 与解码阶段（b）示意图



图：一次生成过程中内存随时间（延迟）变化的示意图



5、美光科技（MU.O）战略聚焦数据中心，HBM3E 技术突破

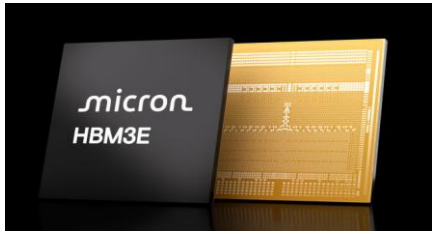
5. 美光科技（MU.O）战略聚焦数据中心，HBM3E 技术突破

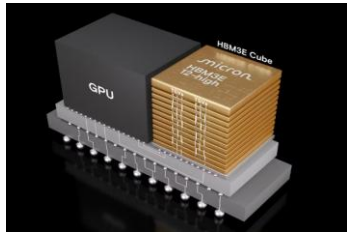
- **财务方面：**美光在2025财年实现了**373.8亿美元**的营收，**同比增长49%**，毛利率（39.8%）、净利率（22.8%）大幅改善，且增长势头延续，2026财年Q1，单季营收达到136.4亿美元，同比增长57%，**单季净利润达52.4亿美元。**
- **核心策略：剥离To C业务，专注AI拉动的HBM市场。**2025年12月，美光正式宣布将于2026年2月以后停止零售渠道销售，将有限的产能释放出来，转向生产高附加值的HBM和企业级SSD。将资源集中服务更大、更具战略意义的客户（即云服务商Hyperscalers和OEM巨头），以确立公司在AI供应链的核心地位。
- **HBM进展及指引：**公司核心产品HBM3E 12Hi（36GB）带宽超1.2 TB/s，引脚速度超过9.2 Gb/s。且HBM3E产品已成功进入NVIDIA Blackwell平台（如HGX B200）及AMD Instinct MI350系列，成为AI训练芯片的关键内存供应商 。公司2026年HBM产能已全部售罄，市场需求高度旺盛，供不应求格局持续性强。25年2月，公司通过利用 EUV（极紫外）光刻技术、设计优化和工艺创新，1γ 节点相比前代产品实现了超过 30% 的单晶圆位元产出增长，能够更高效地扩展内存的供应。
- **SSD产品：**公司核心产品6500 IRON，采用232层 3D TLC NAND，容量达30.72TB，相较于竞品随机写入速度与耐用性有明显优势。

图：美光产品与同业对比图

Product Attribute	Micron 6500 ION SSD	Solidigm D5-P5316	The ION Difference
Sequential 128KB reads	6,800 MB/s	7,000 MB/s	Similar
Sequential 128KB writes	5,757 MB/s	3,643 MB/s	58% better
Random 4KB reads	440,901 IOPS	272,826 IOPS	62% better
Random 4KB read latency	70µs average	106µs	34% better
Random 4KB writes	254,244 IOPS	7,777 IOPS	Over 30X more
Mixed 4KB 70/30 random reads/writes	400,000 IOPS	Not noted on public documents	Significantly better per higher base specs
Mixed 4KB 70/30 random read latency	70µs average	167µs average	More than 2X better
Endurance (100% random writes)	0.3 RDWPD (9 TB/day)	0.026 RDWPD (<1 TB/day)	More than 10X better

图：美光产品（HBM3E及与GPU封装图示）

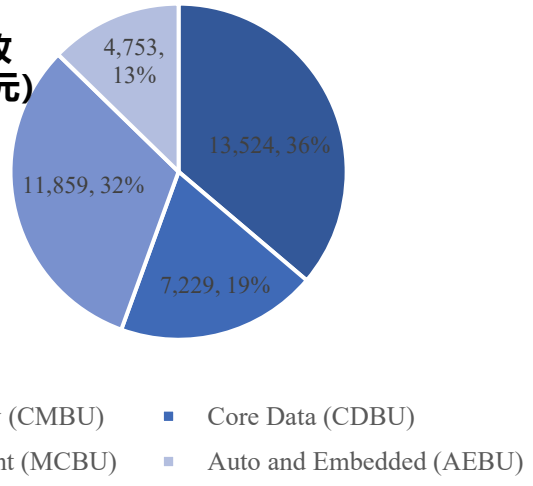




表：美光业绩情况

百万美元	2018F	2019F	2020F	2021F	2022F	2023F	2024F	2025F
财年划分截至	2018/8/30	2019/8/29	2020/9/3	2021/9/2	2022/9/1	2023/8/31	2024/8/29	2025/8/28
营收	30,391	23,406	21,435	27,705	30,758	15,540	25,111	37,378
DRAM	22,625	16,841	14,510	20,039	22,386	10,978	17,603	28,578
NAND	6,510	5,355	6,131	7,007	7,811	4,206	7,227	8,503
Other	1,256	1,210	794	659	561	356	281	297
毛利率（%）	58.9	45.7	30.6	37.6	45.2	-9.1	22.4	39.8
净利率（%）	46.5	27.0	12.5	21.2	28.2	-37.5	3.1	22.8
资本支出	-8,879	-9,780	-8,223	-10,030	-12,067	-7,676	-8,386	-15,857

图：美光分业务营收（2025财年,百万美元）

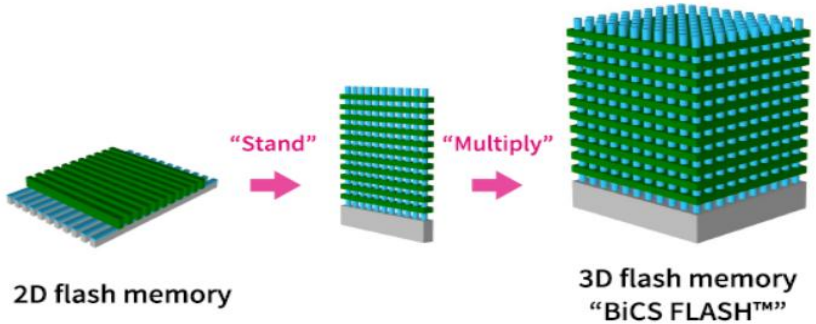


6、闪迪（SNDK.O）拆分后的价值重估，AI存储技术持续迭代

6. 闪迪（SNDK.O）拆分后的价值重估，AI存储技术持续迭代

- **独立历程：**2025年2月24日，西部数据正式完成分拆，闪迪公司（SNDK）独立在纳斯达克上市 。
- **财务方面：**闪迪在2025财年实现了**73.6亿美元**的营收，毛利率30%，2026财年Q2，实现营收30.25亿美元，环比增长 31%，毛利率51.1%，环比+21.2pct，经营利润10.65亿美元，净利润8.03亿美元，环比增长617%，受NAND供应紧张及高价值产品组合（企业级eSSD）占比提升驱动。
- **BiCS技术持续迭代：**公司目前BiCS8技术采用218层堆叠技术，截至2026财年Q1已占其总位元出货量的15%，公司预计在2026财年结束时BiCS8将成为主流规格。公司下一代BiCS9将会堆叠超300层，将会用于制造1TB TLC Die，实现TLC主打性能，QLC主打容量的产品结构。
- **核心技术革命HBF：**面对“内存墙”问题，闪迪创新性地提出HBF方案，提供比HBM高出8到16倍的容量，虽然延迟高于DRAM，但其读取带宽或接近HBM，公司计划于2026年下半年提供HBF样品，并于2027 年推出搭载HBF的AI推理设备。此外闪迪已与SK海力士签署谅解备忘录（MoU），并组建了包含RISC架构发明人David Patterson和前AMD高管Raja Koduri的技术顾问委员会，共建HBF生态及技术标准。

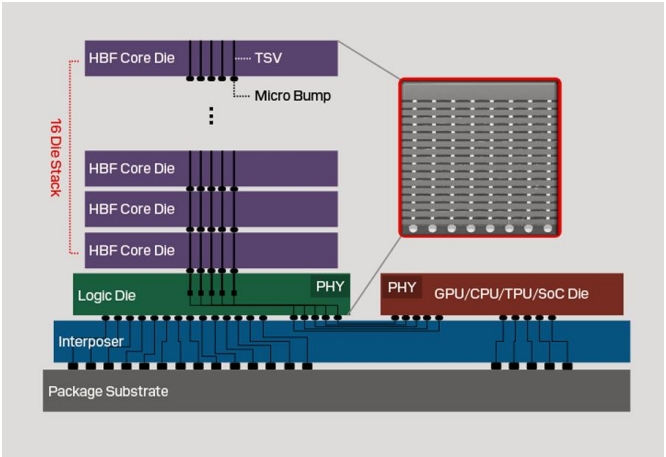
图：2D 3D NAND FLASH示意图



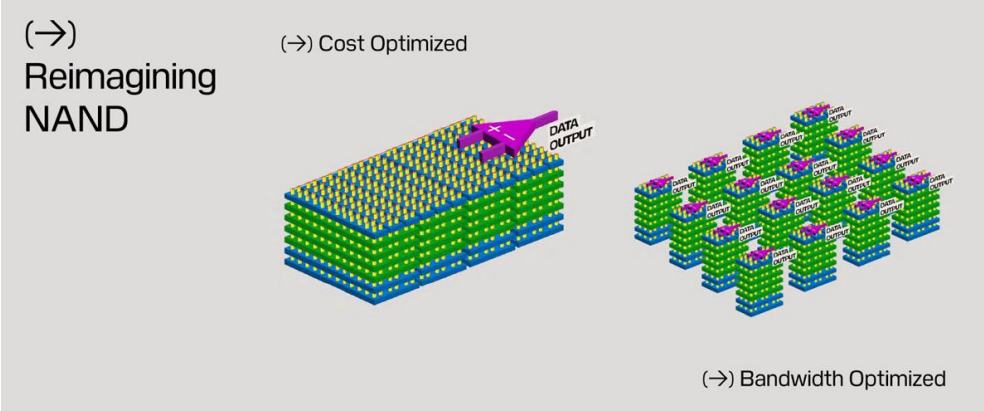
表：闪迪业绩情况

百万美元	2025F	2025Q1	2026Q1	2026Q2
划分截至	2025/6/27	2024/9/27	2025/10/3	2026/1/2
营收	7,355	1883	2308	3025
Cloud	960	300	269	440
Client	4,127.0	1069	1387	1678
Consumer	2,268	514	652	907
毛利率(%)	30	/	29.9	51.1
资本支出	-204	-67	-50	/

图：闪迪HBF 结构示意图



图：闪迪SSD与HBF差异示意图



7、SK 海力士享AI存储先发优势，巨头三星奋力追赶

7. SK海力士受益AI存储先发优势，巨头三星奋力追赶

- **财务复盘：**2021年受数字化需求驱动，三星电子及SK 海力士的营收处于高位；进入2022年受宏观经济及过度库存的影响，行业景气度下行，利润率下行；2023年芯片价格崩盘，行业“减产保价”营收下滑，利润率因资产减值及稼动率下滑而承压；2024年随着减产策略与AI需求的共振，行业快速复苏，SK海力士凭借高溢价的HBM产品拉动，营收、利润共创新高，营业利润率强势复苏至35.5%。25年随AI需求强势驱动下，业绩进入高增通道。
- **DRAM技术路线：**SK海力士采用MR-MUF (Mass Reflow Molded Underfill)工艺，散热性能好、生产效率高，良率行业领先；三星电子或因坚持TC-NCF (Thermal Compression Non-Conductive Film)传统工艺，早期验证较慢，或在HBM4中采用Hybrid Bonding 改进工艺，加速追赶。
- **AI供应链卡位优势：**在HBM3代际中，SK海力士是英伟达一供，在HBM3E阶段仍占英伟达主要份额；三星持续追赶，从“验证”逐步走向“二供”。
- **NAND：**海力士通过Solidigm继承Intel闪存技术，在AI存储需求旺盛中充分受益；三星在NAND领域积淀深厚，但在AI专用QLC SSD细分领域进展慢于海力士，正加速追赶。
- 海力士25Q4实现营收32.83万亿韩元，环比+34%；实现营业利润19.17万亿韩元，环比+68%；实现净利润15.25万亿韩元，环比增长21%。

图：SK海力士分业务营收、研发费用、营业利润率、资本开支

百万美元	2019	2020	2021	2022	2023	2024	2025
营收	23,169	27,085	37,570	34,635	25,090	48,545	68371
DRAM	17,419	19,135	26,737	21,865	15,903	32,806	/
NAND	4,412	6,344	9,200	11,010	7,392	14,135	/
研发费用	2,451	2,642	3,103	3,473	2,872	3,254	/
营业利润率（%）	10.1	15.7	28.9	15.3	-23.6	35.5	48.6
资本开支	-11,949	-8,549	-10,911	-14,756	-6,375	-11,694	/

图：三星电子DS（Device Solutions）分业务营收、营业利润及营业利润率

百万美元	2021	2022	2023	2024	2025
DS营收	83,347	76,421	50,994	81,479	91,563
Memory	63,436	53,193	33,792	61,971	73,265
Non-Memory	19,911	23,228	17,202	19,508	18,299
DS营业利润	25,505	18,489	-11,394	11,074	17,542
DS营业利润率（%）	30.6	24.2	-22.3	13.6	19.1

注：三星电子、SK海力士财年划分与自然年一致。

图：三星、SK海力士、美光HBM验证与量产节奏示意图



注：★ C/S MP 分别代表Engineering Sample(工程样品)送样至英伟达、Commercial Sample（商业样品）、Mass Production（大规模生产）

8、估值情况分析

8. 估值情况分析——AI 驱动下的“量价齐升”与盈利兑现

- **从供需角度看，AI的刚需下存储价格进入上行通道。**在行业供不应求的背景下，叠加原厂对产能的有序控制，确立了存储产品价格持续上涨的趋势。
- **从经营状况看，产品结构持续升级，量价齐升。**企业通过技术升级（例如转产高价值量HBM），在提升产出的同时还拉升产品的ASP，有望充分收益于行业上行红利，兑现公司业绩。
- **从财务视角看，经营杠杆释放，净利弹性显现。**存储行业系重资产投入的行业，随营收快速增长覆盖固定成本开支，涨价带来的收入将大比例转化为净利润，使得存储公司的净利润增速快于营收增速，具备高业绩弹性。
- **业绩兑现支撑估值中枢上移。**行业正处于“营收高增”与“净利率扩张”共振的阶段。具备技术与产能优势的龙头企业，其业绩兑现的确定性强，应享受更高的估值溢价。

表：美光、闪迪、SK海力士、三星电子估值情况

百万美元		Q2 FY25 - Q1 FY26	Q2 FY26E - Q1 FY27E			截至2026-1-23	
股票简称	股票代码	营业收入	彭博一致预期营业收入	彭博一致预期净利润	Forward PE	市值	PE（TTM）
美光科技	**MU.O**	42,312	85,996	44,745	10.05	4498亿美元	37.8
闪迪	**SNDK.O**	7,780	12,073	3,030	22.90	694亿美元	-40
百万美元		**Q4 24-Q3 25**	**Q4 25E - Q3 26E**			**截至2026-1-23**	
股票简称	股票代码	营业收入	彭博一致预期营业收入	彭博一致预期净利润	Forward PE	市值	PE（TTM）
SK海力士	**000660.KS**	59,805	102,178	48,169	7.87	3792亿美元	15.6
三星电子	**005930.KS**	224,011	281,781	58,550	10.44	6115亿美元	27.7

注：其中三星电子2025Q4彭博一致预期净利润基于25Q3及26Q1E平均值

资料来源：Bloomberg，东吴证券研究所

风险提示

风险提示

- **下游AI资本开支（CAPEX）不及预期风险：** 虽然当前CSP（云服务商）对AI服务器投入巨大，但如果未来AI应用商业化落地受阻或大模型迭代速度放缓，可能导致科技巨头削减资本开支。这将直接冲击HBM、企业级SSD等高毛利产品的需求，导致行业估值中枢下移。
- **存储原厂产能扩张失控导致供需格局恶化风险：** 存储行业具有典型的周期性。随着价格回暖和利润修复，若各大原厂（如三星、SK海力士、美光、闪迪）激进释放产能，可能导致供过于求的局面重演，进而引发存储芯片价格（ASP）大幅下跌。
- **新技术研发进度或良率爬坡不及预期风险：** HBF（高带宽闪存）、HBM3E/HBM4以及超高层数3D NAND（如300层+）均属于高难度工艺。如果相关技术（如混合键合CBA、TSV硅通孔）的良率无法如期提升，将影响企业的成本控制能力和新产品的上市时间，导致市场份额流失。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“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。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。 未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

东吴证券投资评级标准

资评级基于分析师对报告发布日后6至12个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证50指数），具体如下：

公司投资评级：

买入：预期未来6个月个股涨跌幅相对基准在15%以上；

增持：预期未来6个月个股涨跌幅相对基准介于5%与15%之间；

中性：预期未来 6个月个股涨跌幅相对基准介于-5%与5%之间；

减持：预期未来 6个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来6个月内，行业指数相对强于基准5%以上；

中性： 预期未来6个月内，行业指数相对基准-5%与5%；

减持： 预期未来6个月内，行业指数相对弱于基准5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所

苏州工业园区星阳街5号

邮政编码：215021

传真：（0512）62938527

公司网址： http://www.dwzq.com.cn

东吴证券 SOOCHOW SECURITIES

东吴证券 财富家园