半导体

证券研究报告
2025 年 09 月 26 日

AI 存储革命已至，“以存代算”开启存储新纪元

“以存代算”发展背景：AI 推理成价值核心，HBM 瓶颈凸显产业痛点，“以存代算”应运而生。当前，AI 推理已成为衡量大模型商业化价值的关键标尺，但在实际应用中仍面临“推不动、推得慢、推得贵”的严峻挑战。为突破算力瓶颈与“存储墙”制约，“以存代算”作为一种颠覆性技术范式应运而生。该技术通过将 AI 推理过程中的矢量数据（如 KV Cache）从昂贵的 DRAM 和 HBM 显存迁移至大容量、高性价比的 SSD 介质，实现存储层从内存向 SSD 的战略扩展，而非简单替代。其核心价值在于显著降低首 Token 时延、提升推理吞吐量，并大幅优化端到端的推理成本，为 AI 大规模落地提供可行路径。

“以存代算”核心技术：“以存代算”CachedAttention 技术是一种通过将 AI 推理中历史对话的 KV Cache 缓存到 HBM+DRAM+SSD 等外部存储介质。在该系统中，HBM 作为 GPU 本地高速存储，负责存储当前活跃会话的 KV Cache，支撑 LLM 推理计算；DRAM 作为中间缓存层，承接 HBM 的异步写入与 SSD 的预加载，平衡速度与容量；SSD 则作为长期存储层，提供大容量持久化存储，承载非活跃历史数据。“以存代算”CachedAttention 将首 Token 时延（TTFT）显著缩短了 87%，并提升了 Prefill 阶段 7.8 倍的吞吐量，从而将端到端推理成本降低了 70%。

“以存代算”硬件突破：在“以存代算”技术范式下，SSD 不再是单纯的数据存储载体，而是深度参与 AI 推理的核心组件，其需承接从 HBM、DRAM 卸载的 KV Cache，因此被赋予大容量、高吞吐、低延迟的新要求，以缓解对高成本 HBM 的依赖。同时，SSD 主控芯片作为“控制大脑”，需通过先进算法优化数据寻址调度，支撑 AI 推理中数据高效流转。在此背景下，AI SSD 技术将沿三大方向发展：颗粒上，向 QLC 颗粒演进，凭借技术升级实现高性能与大容量兼顾，满足 AI 大模型数据存储调用需求；接口协议上，以 PCIe 5.0/6.0 接口搭配 NVMe 协议为基础，未来融入 CXL 技术，进一步提升带宽与降低延迟；功能上，向智能化升级，如铠侠计划推出软件让 SSD 自主处理 AI 检索任务，Solidigm 探索液冷方案优化散热，实现存储与 AI 推理的深度协同。

“以存代算”企业布局：“以存代算”的核心实践已获产业龙头积极布局。华为 UCM 作为“以存代算”产品化关键载体，构建智能分级缓存，数据可根据记忆热度在 HBM、DRAM、SSD（固态硬盘）等存储介质中实现按需流动；同时融合多种稀疏注意力算法，实现存算深度协同。除了以 HBM+DRAM+SSD 构建的多级缓存体系外，还存以 KVCache 缓存技术为核心的多元实践。**浪潮存储 AS3000G7** 优化存储架构与缓存管理机制，智能调度 KVCache 数据，具备高扩展性，能快速处理热数据，为 AI 推理等应用提供高效稳定的存储算力。**焱融 YRCloudFile** KVCache 依托自研分布式文件系统，实现 KVCache 数据在分布式环境下的高效存取与智能负载均衡，兼容性强，提升数据与计算协同效率。**国际层面，铠侠、美光、Solidigm** 等巨头正积极推动 AI SSD 的技术迭代与产品创新。我们认为，QLC+PCIe/NVMe+CXL 有望构筑下一代 AI SSD 基座，推动 SSD 从单纯存储介质，升级为 AI 推理“长期记忆”载体。

投资建议：AI 存储革命已至，“以存代算”催生核心机遇，显著节省算力消耗加速 AI 推理，带动 SSD 需求增速高于传统曲线。建议关注：
存储模组厂商：江波龙（天风计算机联合覆盖）、德明利、佰维存储、朗科科技、联芸科技、万润科技等；**存储芯片：**兆易创新、普冉股份、北京君正、东芯股份、恒烁股份、澜起科技、聚辰股份等；**存储分销与封测：**香农芯创、深科技、太极实业、中电港等

风险提示：地缘政治带来的不可预测风险，需求复苏不及预期，技术迭代不及预期，产业政策变化风险

投资评级

行业评级	强于大市(维持评级)
上次评级	强于大市

作者

唐海清 分析师
SAC 执业证书编号：S1110517030002
tanghaiqing@tfzq.com
李泓依 分析师
SAC 执业证书编号：S1110524040006
lihongyi@tfzq.com

行业走势图

资料来源：聚源数据

相关报告

1《半导体-行业研究周报:25Q2 半导体业绩总结及展望：AI 驱动与国产替代共筑成长主线》 2025-09-02
2《半导体-行业研究周报:3 季度半导体景气度展望乐观，持续重点关注国产算力及自主可控方向》 2025-08-25
3《半导体-行业研究周报:“以存代算”开启存储新纪元，7 月半导体行情延续景气》 2025-08-19

内容目录

图表目录

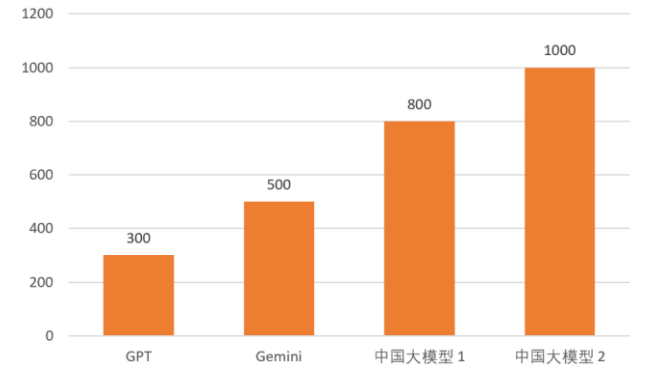
1. 发展背景：AI 推理成价值核心，HBM 瓶颈凸显产业痛点

1.1. AI 大模型推理中存在推不动、推得慢、推得贵三大挑战

当前，人工智能已步入发展深水区，AI 推理正成为下一个增长的关键阶段，推理体验和推理成本成为了衡量模型价值的黄金标尺。华为公司副总裁、数据存储产品线总裁周越峰指出，AI 时代，模型训练、推理效率与体验的量纲都以 Token 数为表征，Token 经济已经到来。ChatGPT 的访问量呈现线性增长，最新访问量达到 4 亿，受益于中国 AI 大模型 DeepSeek 的快速发展，日均调用量也在快速上升，2025 年 1 月开始，中国 AI 推理的需求增长 20 倍，未来三年算力需求有望快速增长。IDC 表示，2024 年算力需求 60%是训练，40%是推理，到 2027 年中国用于推理的算力需求——工作负载将达到 72.6%。

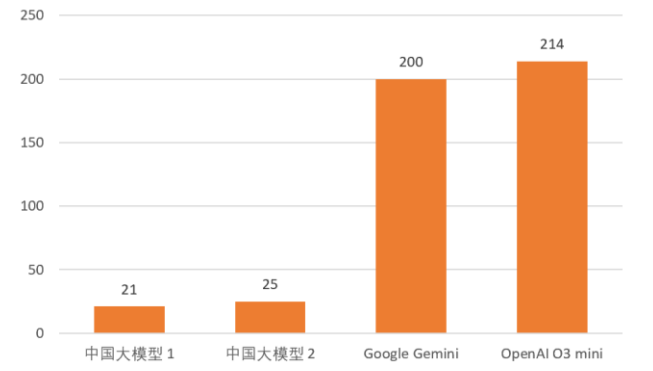
据电子发烧友网，当下，AI 大模型推理应用落地中，遇到推不动、推得慢和推得贵的三大挑战。首先，长文本越来越多，输入超过模型上下文窗口的内容，推理窗口小就推不动；其次，由于中美在 AI 基础设施的差距，中国互联网大模型首 Token 时延普遍慢于美国头部厂商的首 Token 时延，时延长度为后者的两倍；推得贵，美国大模型的推理吞吐率为中国大模型推理吞吐率的 10 倍。

图 1：美国大模型推理首 Token 时延=1/2 中国大模型（TTFT 毫秒）



资料来源：华为官网，天风证券研究所

图 2：美国大模型推理吞吐率=10 倍中国大模型（token/秒）



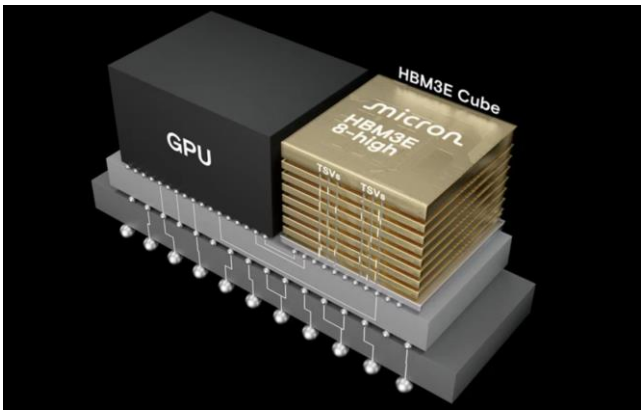
资料来源：华为官网，天风证券研究所

1.2. HBM 突破存储墙，海外垄断下技术难度和成本高企成最大障碍

上述 AI 推理中所遇到的挑战，主要受制于传统 DRAM 面临“存储墙”瓶颈，内存的存取速度严重滞后于处理器的计算速度，严重制约了 AI 模型的训练和推理速度。直到 HBM 的出现，彻底改变了传统 DRAM 的布局模式。

HBM 已经成为 AI 革命的核心，是对传统内存瓶颈的有效突破。HBM（高带宽内存）是一种专用内存技术，用于 AI 处理器、GPU 和 HPC 系统。HBM3 每堆栈可提供高达 819 GB/s 的传输速度，对于支持大型语言模型 (LLM)、神经网络训练和推理工作负载至关重要。与传统内存芯片相比，HBM 芯片最大特点在于采用了先进的 3D 堆叠技术，通过硅通孔（TSV）将多个 DRAM 芯片垂直堆叠在一起，并与 GPU 或 CPU 等处理器封装在同一模块中，实现了大容量、高位宽的 DDR 组合阵列，能有效解决“存储墙”问题。

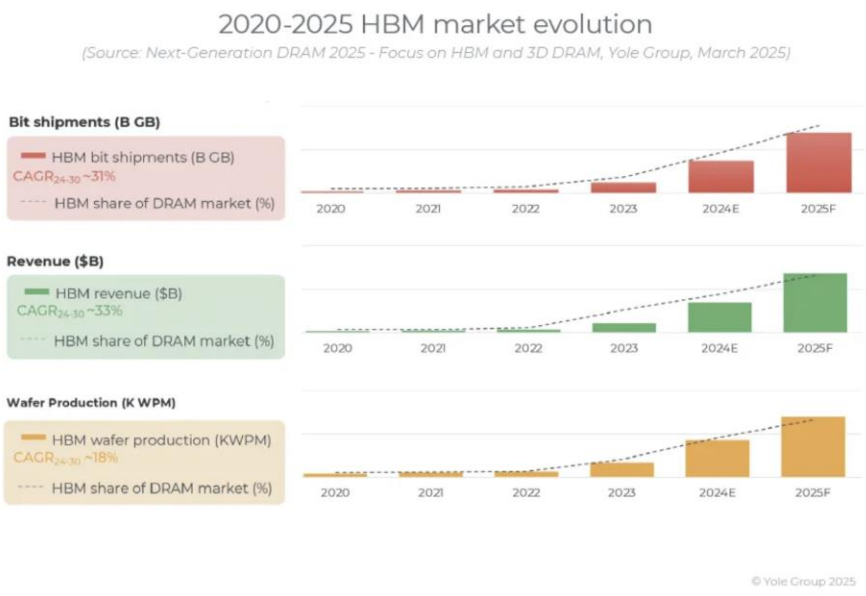
图 3：美光的 HBM3E 产品结构图



资料来源：美光官网，中国电子报公众号，天风证券研究所

根据 Yole Group 的数据，HBM 市场未来几年都呈现出高速增长的态势。全球 HBM 收入预计将从 2024 年的 170 亿美元增长至 2030 年的 980 亿美元，复合年增长率达 33%。HBM 在 DRAM 市场中的收益份额预计将从 2024 年的 18% 扩大到 2030 年的 50%。从位出货量来看，HBM 从 2023 年的 1.5 B GB，到 2024 年的 2.8 B GB。到 2030 年，预测将达到 7.6 B GB。

图 4：2020-2025 HBM 市场发展



资料来源：Yole Group，半导体行业观察公众号，天风证券研究所

当前 AI 算力生态高度依赖 HBM 硬件升级，HBM 市场呈现寡头竞争格局。然而极高的技术难度和高昂成本成为了制约其大规模应用的一大障碍。在 AI 服务器中，HBM 的成本占比约为 20%—30%，仅次于用于计算的 AI 芯片。当前全球 HBM 市场由三星、SK 海力士等主导，且受到美国出口政策的影响。根据 2024 年 12 月 2 日发布的新规，美国禁止向中国出口 HBM2E（第二代 HBM 的增强版）及以上级别的 HBM 芯片。不仅美国本土生产的 HBM 芯片受到限制，任何在海外生产但使用了美国技术的 HBM 芯片也受到出口管制。该禁令于 2025 年 1 月 2 日正式生效。目前，国产厂商 HBM 的突破还在推进中。

2. 核心技术：以存代算元年开启，打破 HBM 限制重构 AI 存储生态

2.1. 产业背景：AI 推理成本高企与 HBM 瓶颈催生“以存代算”新范式

当前，人工智能技术的蓬勃发展推动大模型训练走向规模化，但真正创造持续商业价值的核心环节在于推理过程。观察者网表示，AI 推理算力需求正迅速超越训练，成为成本与性能的关键瓶颈。

在这一背景下，键值缓存（KV Cache）技术应运而生，成为提升推理效率的关键机制。其原理是将已生成 Token 对应的 Key（表征历史输入特征）和 Value（用于输出计算的参考信息）临时存储起来，使模型在生成新 Token 时可直接复用缓存结果，避免重复计算，从而显著降低计算负载、提升响应速度。

然而，KV Cache 的高度依赖也带来了新的问题：它需占用大量 GPU 显存资源，尤其是价格昂贵的高带宽内存（HBM）。随着生成文本长度与对话轮次的增加，缓存数据量急剧膨胀，极易导致 HBM 与 DRAM 资源耗尽。更为关键的是，面对大模型 PB 级的天量参数和持续增长的序列长度，传统推理架构对 HBM 的过度依赖已日益成为系统扩展的瓶颈。

尽管 HBM 在性能上表现卓越，但其高昂的成本与有限的供应极大地限制了大规模部署的经济性。另一方面，虽然 SSD 具备成本低、容量大的优势，但其传输速率与延迟尚无法满足高频实时推理的要求。正是在性能、容量与成本构成的“不可能三角”困境中，“以存代算”作为一种突破性的技术范式逐渐走向成熟。

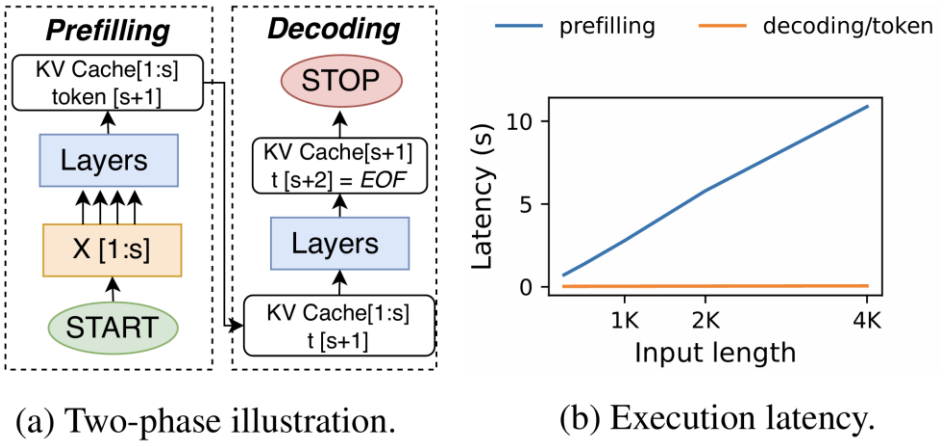
以存代算 CachedAttention 用 GPU 外部低成本的存储介质（HBM+DRAM+SSD）来缓存历史对话的 KV Cache，然后在后面轮次对话到来时，重新加载缓存的历史对话 KV Cache 并重用，那么在新一轮对话中只用 Prefilling 阶段的新 Token。

2.2. 技术背景：KV Cache 重复计算成推理效率关键制约

Transformer 算法是生成式 AI 模型的基石。Transformer 模型由多个 Transformer 层组成，每层都包含两个模块：自注意力（Self-Attention）和前馈网络（FFN）。对于输入的 Token，每层都会对每个 Token 计算生成 Query（Q）、Key（K）和 Value（V）。Key（K）和 Value（V）通常缓存在 GPU 中，称为 KV Cache（KV 缓存），他们占用空间很大。

在 LLM 的推理过程中，包含两个阶段：**预填充阶段（Prefilling Phase）和解码阶段（Decoding Phase）**。预填充阶段并行地处理所有输入 Prompt 的 Token，生成 KV Cache。解码阶段利用预填充阶段生成的 KV Cache，迭代地生成输出 Token，每次迭代输出一个 Token。

图 5：LLM 推理中包含预填充阶段（Prefilling Phase）和解码阶段（Decoding Phase）



资料来源：《Cost-Efficient Large Language Model Serving for Multi-turn Conversations with CachedAttention》Pengfei Zuo 等，天风证券研究所

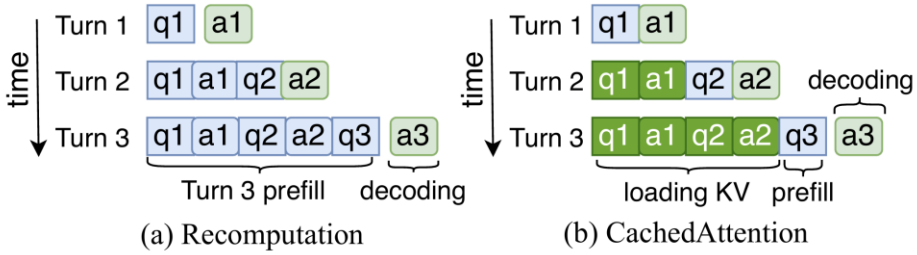
让人类参与多轮对话是 LLM 的一个基本特征，多轮对话会话由一系列连续对话组成。根据对开源大模型对话数据集 ShareGPT 的统计（ShareGPT 是通过收集不同人与 ChatGPT 的真实对话形成的数据集），发现 **ShareGPT 中有超过 73% 的对话都是多轮的。**

由于在多个对话轮次中重复计算 KV Cache，因此 LLM 服务引擎在执行多轮对话中效率低下，产生高昂成本。在单轮对话中，LLM 将 KV Cache 存储在 GPU 上有限的高带宽内

存 （HBM）中。当对话结束时，LLM 会丢弃与该会话关联的 KV Cache，以释放 HBM 中的空间供其他活动会话使用。当用户在对话中发送下一条消息时，LLM 会再次计算整个 KV Cache，这导致重复计算相同的 KV Cache 浪费宝贵的 GPU 计算资源。

如图 6（a）所示，在第一轮对话中，LLM 生成 a1 的 q1 KV Cache。完成第 1 轮后，LLM 会丢弃 KV Cache 以回收 HBM 空间。在第二、三轮对话中，LLM 重新生成 a1 的 q1 KV Cache。**随着对话轮数的增加，新一轮对话的输入 Token 中历史 Token 的比例急剧增加，到后面的轮次中历史 Token 的比例会超过 99%。**

图 6：重复计算和以存代算的对比



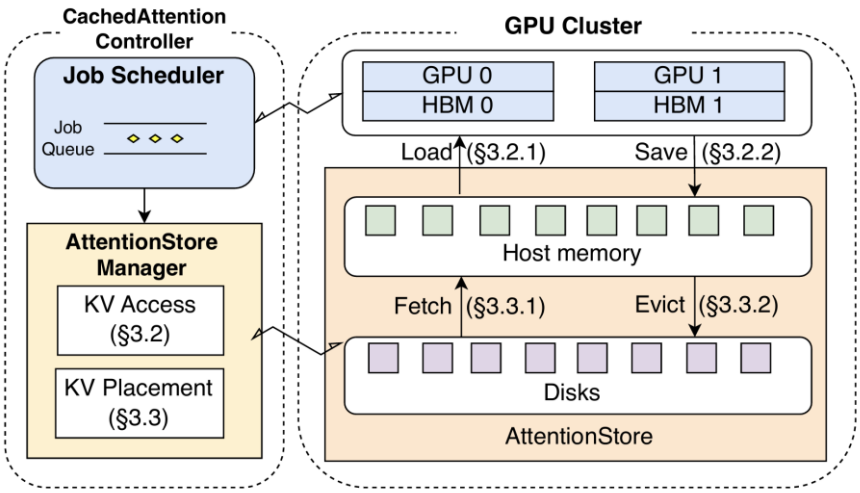
资料来源：《Cost-Efficient Large Language Model Serving for Multi-turn Conversations with CachedAttention》Pengfei Zuo 等，天风证券研究所

2.3. 技术机制：以存储换计算，实现 KV Cache 持久化与多级缓存

以存代算 CachedAttention 是一种新的 Attention 技术，用 GPU 外部低成本的存储介质（ AttentionStore ）来缓存历史对话的 KV Cache，然后在后面轮次对话到来时，重新加载缓存的历史对话 KV Cache 并重用，那么在新一轮对话中只用 Prefilling 阶段的新 Token。

具体而言，当关联的对话会话处于非活动状态时，CachedAttention 会将 KV Cache 保存在名为 AttentionStore 的 KV Cache 系统中，而不是像传统注意力机制那样将其删除。如果将来激活了同一对话，则会从 AttentionStore 获取其 KV Cache 并重复用于推理。通过这样做，CachedAttention 仅执行部分 prefilling 阶段的 tokens，即在新的对话回合中输入的新 tokens，而不是预填充所有的 tokens。如图 6（b），在执行第三轮的推理时， 使用 q1 q2 a1 a2 的 KV Cache，只需要输入 q3 即可。CachedAttention 有效消除了历史 token 的重复计算，从而降低了 prefilling 成本。

图 7：CachedAttention 的系统架构



资料来源：《Cost-Efficient Large Language Model Serving for Multi-turn Conversations with CachedAttention》Pengfei Zuo 等，天风证券研究所

与 CachedAttention 相互配合的是一个涉及 HBM+DRAM+SSD 的多级 KV Cache 缓存

系统，他们三者的作用及关系如下：

1. 高带宽内存（HBM）

- **核心定位：**GPU 本地高速存储，用于实时支撑 LLM 推理计算，是 KV 缓存访问的 “第一梯队”。
- **关键作用：**
  - 存储当前活跃会话的 KV Cache，直接供 GPU 的自注意力计算（Attention）和前馈网络（FFN）调用，避免计算阻塞。
  - 预留读写缓冲区（read buffer/write buffer）：读缓冲区用于提前加载 DRAM 中的 KV 缓存，与 GPU 计算重叠；写缓冲区用于暂存未完成保存的 KV 缓存，避免阻塞下一个推理任务。

2. 主机内存（DRAM）

- **核心定位：**HBM 与 SSD 之间的 “中间缓存层”，平衡存储容量与访问速度。
- **关键作用：**
  - 存储近期可能被访问的非活跃会话 KV Cache，作为 HBM 的扩展，避免频繁从低速 SSD 加载。
  - 承接 HBM 的异步 KV 缓存写入：在推理计算（如解码阶段）同时，将 HBM 中已完成计算的 KV 缓存异步保存到 DRAM，减少 HBM 占用。
  - 作为预取目标：通过调度感知预取（scheduler-aware fetching），将 SSD 中即将被访问的 KV 缓存提前加载到 DRAM，确保 GPU 访问时能命中高速存储。

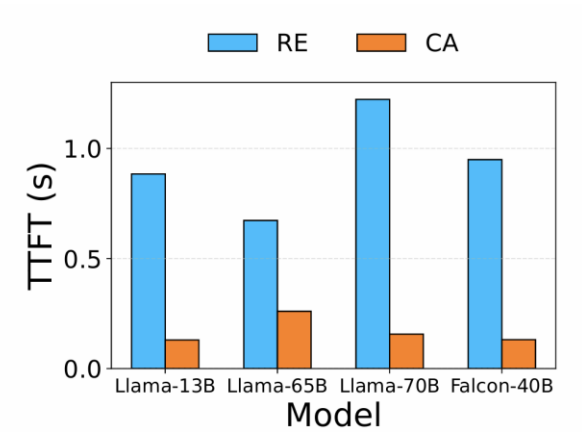
3. 固态硬盘（SSD）

- **核心定位：**海量 KV 缓存的 “长期存储池”，解决 HBM 和 DRAM 容量不足的问题。
- **关键作用：**
  - 存储海量非活跃会话的 KV Cache，提供 TB 级容量，避免因 HBM/DRAM 容量限制导致 KV 缓存被丢弃。
  - 配合调度感知驱逐（scheduler-aware eviction）：当 DRAM 空间不足时，将长期不被访问的 KV Cache 从 DRAM 迁移到 SSD；当 SSD 空间不足时，驱逐最不可能被访问的会话 KV Cache。

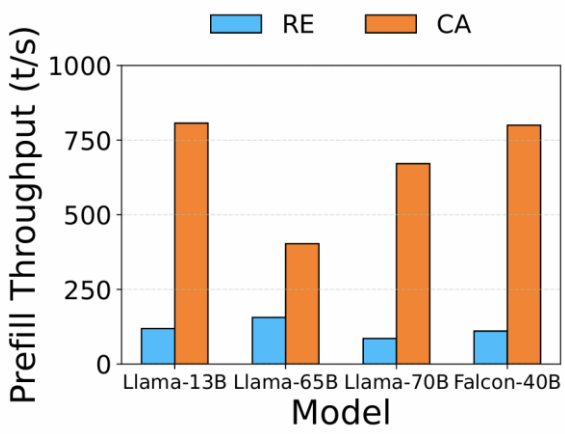
大量的实验结果表明，CachedAttention 将首 Token 时延（TTFT）显著缩短了 87%，并提升了 Prefill 阶段 **7.8 倍**的吞吐量，从而将端到端推理成本降低了 70%。

图 8：首 Token 时延（TTFT）显著缩短

图 9：Prefill 阶段吞吐量提升



资料来源：《Cost-Efficient Large Language Model Serving for Multi-turn Conversations with CachedAttention》Pengfei Zuo 等，天风证券研究所



资料来源：《Cost-Efficient Large Language Model Serving for Multi-turn Conversations with CachedAttention》Pengfei Zuo 等，天风证券研究所

3. 企业布局：华为领衔重点推广，浪潮、焱融乘风而上

3.1. 华为 UCM：以存代算产品化载体，软件定义突破 HBM 资源枷锁

华为推出的 UCM（推理记忆数据管理器），是一款以 KV Cache（键值缓存）和记忆管理为核心的推理加速套件，包括对接不同引擎与算力的推理引擎插件（Connector）、支持多级 KV Cache 管理及加速算法的功能库（Accelerator）、高性能 KV Cache 存取适配器（Adapter）三大组件。

华为方面介绍称，依托 UCM 层级化自适应的全局前缀缓存技术，系统能直接调用 KV 缓存数据，避免重复计算，使首 Token 时延最大降低 90%。同时，UCM 将超长序列 Cache 分层卸载至外置专业存储，通过算法创新突破模型和资源限制，实现推理上下文窗口 10 倍级扩展，满足长文本处理需求。

图 10：UCM 以 KVCache 和记忆管理为中心提供全场景系列化推理加速能力

资料来源：华为官网，天风证券研究所

UCM 具备智能分级缓存能力，可根据记忆热度在 HBM、DRAM、SSD（固态硬盘）等存储介质中实现按需流动；同时融合多种稀疏注意力算法，实现存算深度协同，使长序列场景下 TPS（每秒处理 token 数）提升 2-22 倍，显著降低每 Token 推理成本，为企业减负增效。**UCM 的智能分级缓存能力，可将 AI 推理所需数据从 DRAM 内存迁移至 SSD 闪存介质，以此优化计算效率。**其核心价值在于降低对 HBM 和 GPU 的过度依赖，并实现“存算一体”系统创新。该技术的本质是存储层的扩展，而非替代 DRAM。

UCM 通过创新架构设计和存储优化，突破了 HBM 容量限制，提升了国内 AI 大模型推理性能，完善了中国 AI 推理生态的关键环节。

图 11：记忆数据在 HBM、DRAM、SSD 中按需流动

资料来源：华为官网，天风证券研究所

3.2. 以存代算多元实践：KVCache 缓存技术成为大模型系统架构“标配”

在“以存代算”的技术范式下，除了以 HBM+DRAM+SSD 构建的多级缓存体系外，还存在其他形式的 KV Cache 缓存技术，它们同样以存储资源换取计算效率，显著提升大模型推理性能。尽管在具体实现上可能不严格遵循 HBM+DRAM+SSD 的层级结构，但其核心思想一致——通过外部存储介质缓存历史 KV Cache，避免重复计算，从而降低延迟、提升吞吐、节约算力。

3.2.1. 浪潮存储 AS3000G7：存储托管 KV Cache，实现“以存代算”

浪潮存储 AS3000G7作为**国内首款推理加速存储**，可存储所有 KV Cache 及多轮对话结果。其创新架构通过将 KV Cache 从 GPU 写入本机内存，再经高速网络缓存至 **AS3000G7**，下轮对话时按需拉取缓存无需重新计算，彻底实现“以存代算”，显著节省算力消耗并提升资源利用率。

图 12：浪潮存储 AS3000G7

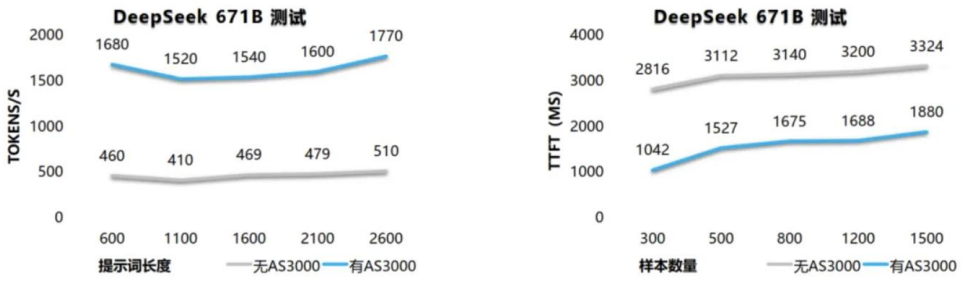
资料来源：浪潮数据存储公众号，天风证券研究所

作为**国内首款推理加速存储**，AS3000G7 以**四大核心优势**重塑推理效率：**降低响应延迟：**将历史 Token 缓存至 **AS3000G7** 存储层，下轮对话从 NVMe SSD 硬盘中拉取历史 token 的 KV Cache，减少 GPU 重复计算带来的资源消耗，TTFT 降低 90%；**承载更多并发：**TTFT 在 400ms 以内的前提下，系统可支持的吞吐量（Token/s）可达原方案 5 倍，单位 GPU 资源可承载更多推理请求；**降低 GPU 功耗：**TTFT 的降低与并发的提升，单 Token 平均功耗下降 60%，在承载同等规模 token 负载时，GPU 服务器整机功耗降低。**生态兼容适配：**广泛兼容国产与海外芯片的异构算力平台，深度适配 vLLM 框架下的 deepseek 等主流大模型，优化推理体验。在某头部客户联合测试中，采用 1 台 GPU 服务器搭配 1 台 **AS3000G7** 推

理加速存储的组合方案实现：

- 稳定支撑 500+并发对话，TTFT 降低 90%，响应速度大幅提升
- 同硬件配置下吞吐量（Tokens/s）提升 5 倍，在不增加 GPU 资源的情况下，实现更高并发的推理请求
- 单 token 功耗降低 70%，单位算力成本降低 60%，推理性价比提升

图 13：浪潮某头部客户联合测试



资料来源：浪潮数据存储公众号，天风证券研究所

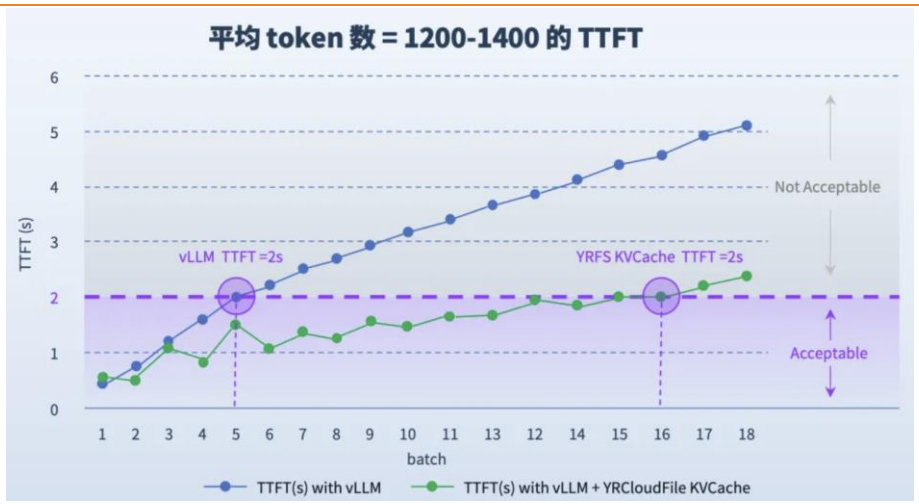
随着大模型推理需求的持续攀升，**AS3000G7** 的推出恰逢其时。其通过“**以存代算**”的技术创新突破 KV Cache 重计算瓶颈，为 AI 规模化应用筑牢存储根基。未来，随着多模态与实时交互场景的普及以及存储与计算的协同优化，KV Cache “以存代算”将成为降本增效的核心竞争力，为智能时代的推理存储构建新基准。

3.2.2. 焱融 YRCloudFile KVCache：以存代算显著提升 AI 推理性价比

焱融存储 YRCloudFile KVCache 方案通过“HBM + YRCloudFile 高性能分布式文件存储”的组合方式，将 KVCache 从显存扩展至高性能共享存储，不仅缓解了显存压力，还有效避免因缓存占用过高导致的推理卡顿或任务中断，在保障推理效率与响应速度的同时，实现更精准、高性价比的大模型推理。

YRCloudFile KVCache 在实际客户的 AI 推理系统中展现出显著性能优势。推理系统的响应延迟与并发处理能力是衡量用户体验的关键指标。经客户实测，YRCloudFile 在这两个核心维度上均实现了明显优化，显著提升了整体推理效率与稳定性。

图 14：YRCloudFile KVCache 在实际客户的 AI 推理系统中展现出显著性能优势



资料来源：焱融科技公众号，天风证券研究所

- **测试内容**

在相同的 NVIDIA H20 显卡配置下，选用 DeepSeek-R1-Distil-Lama-70B 模型，基于

evalscope （使用 longalpaca 数据集，设定不同 --max-prompt-length 参数），对原生 vLLM 与 vLLM + YRCloudFile KVCache 两种方案在并发数递增时的 TTFT 表现进行对比测试。

- **测试结论**

首先，需要强调的是，TTFT 是衡量推理体验的关键指标。理想情况下，TTFT 应稳定在 2 秒以内，这表明用户体验良好。一旦 TTFT 超过 2 秒，用户体验将显著下降。

- 在并发数仅达到 5 时，原生 vLLM 的 TTFT 就已突破 2 秒阈值；
- 搭载 YRCloudFile KVCache 后，在 TTFT 稳定保持在 2 秒以内的前提下，系统可支持的并发数大幅提升至 16，相比原生方案提高了 3.2 倍。

这表明 YRCloudFile KVCache 不仅能够显著降低响应延迟，还能在单位 GPU 资源下承载更多推理请求，大幅提升系统吞吐量和性价比，全面优化大语言模型的推理体验。

在大语言模型推理规模化应用的关键阶段，YRCloudFile KVCache 不仅有效解决了显存瓶颈问题，还能够应对上下文长度不断增长的压力，为大模型推理提供更具弹性与性能优势和性价比的底层支撑。最新的用户实测数据也进一步证明了其在高并发场景下的出色表现。无论是高并发的智能客服场景，还是复杂的多轮对话应用，YRCloudFile KVCache 都能为快速、流畅的推理响应提供坚实保障，助力企业在 AI 推理时代抢占先机。

4. 硬件突破：SSD 需求有望超越传统曲线，SSD 主控帮助寻址调度

1）AI 推理带动的 SSD 需求将持续超越传统存储增长曲线

“以存代算”技术范式下，SSD 不再仅是数据存储载体，而是深度参与 AI 推理流程的核心组件。通过承接从 HBM 和 DRAM 中卸载的 KV Cache、历史对话记录及 RAG 知识库等温数据，**SSD 有效缓解了对高成本 HBM 的依赖，显著降低系统总拥有成本（TCO）。**同时，其对大容量、高吞吐、低延迟的刚性需求，也反向驱动 SSD 在接口协议、颗粒类型、智能管理等方面持续迭代，推动产业向高性能、高可靠性、大容量方向升级，重塑了 SSD 在 AI 基础设施中的地位。

2）SSD 主控芯片起数据寻址调度作用

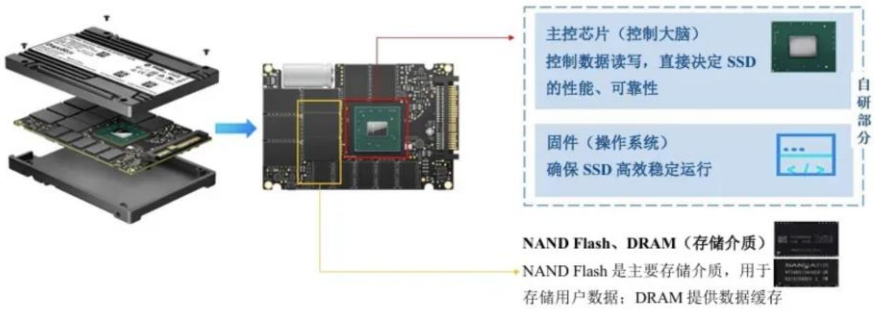
目前来看，QLC 闪存优化、PCIe 接口代际升级、AI 算法融合等正成为厂商的几大突破方向。华为采用海思自研的 Hi1812/Hi1822 系列主控芯片，技术突破点在于维持长期读写速度（从物理层限制转为数学优化），并通过均衡磨损算法延长 SSD 寿命。

4.1. 企业级 SSD：AI 浪潮中崛起，性能、可靠与 AI 应用适配的深度融合

企业级 SSD 由固态电子存储芯片阵列制成，核心部件包括主控芯片、固件和存储介质（NAND Flash、DRAM），其中主控芯片和固件直接决定企业级 SSD 的性能和可靠性等产品表现。主控芯片（控制大脑）控制数据读写，直接决定 SSD 的性能、可靠性固件（操作系统）确保 SSD 高效稳定运行 NAND Flash、DRAM（存储介质） NAND Flash 是主要存储介质，用于存储用户数据；DRAM 提供数据缓存 。

图 15：企业级 SSD 的核心部件示意图

企业级 SSD 的核心部件示意图



资料来源：电子发烧友网公众号，天风证券研究所

总线是计算机不同功能部件之间交互数据的通路，对于 SSD 而言，总线就 是数据自 SSD 到 CPU 所走的路。总线承载能力具有一定上限，其位宽、传输 频率和通道数共同决定了数据理论传输速度。SSD 的总线类型可分为 SATA 总线、SAS 总线、PCIe 总线三类，对比如下：

图 16：SSD 总线类型

总线类型	版本代际	带宽
SATA	3.2	6Gb/s
SAS	4.0	24Gb/s
PCIe	3.0	32Gb/s
	4.0	64Gb/s
	5.0	128Gb/s
	6.0	256Gb/s

注：PCIe 带宽基于 PCIe x4（四通道）计算。

资料来源：电子发烧友网公众号，天风证券研究所

SATA 总线最初为个人电脑和消费级市场设计，更注重存储容量而非传输速 度。SAS 总线相对于 SATA 总线提供更高的数据传输速率和容错能力，主要面 向早期企业级场景应用。PCIe 总线被实际应用后，PCIe SSD 相较于其他类型产品提供了更高数据传输速度和更低延迟，在高性能、高可靠性要求的企业级应 用场景中表现突出，已成为目前最主流的企业级 SSD。PCIe 总线自 2003 年首 次推出以来持续迭代，传输速度显著提升。目前企业级 PCIe SSD 市场产品以 PCIe 4.0 为主，PCIe 5.0 产品已逐步推向市场。

SSD 根据应用场景不同，主要分为企业级 SSD 和消费级 SSD。企业级 SSD 主要应用于 AI、云计算、大数据等数据中心应用场景，消费级 SSD 广泛应用于电脑、手机、移动硬盘等消费电子场景。与消费级 SSD 相比，企业级 SSD 在产品性能、可靠性、耐用性等方面表现更为突出，主要指标对比情况如下：

图 17：企业级 SSD 和消费级 SSD 对比

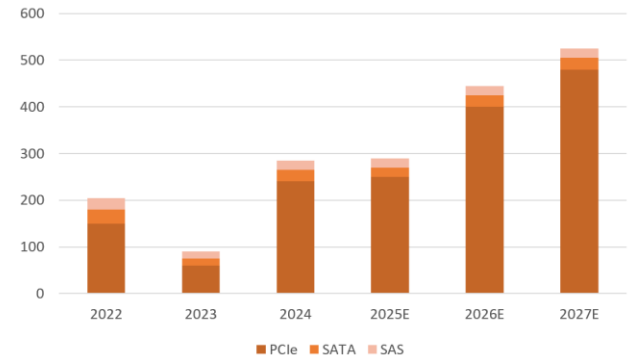
对比维度	具体指标	企业级 SSD	消费级 SSD
应用领域		AI、云计算、大数据等数据中心应用场景	电脑、手机、移动硬盘等消费电子场景
主控芯片价格		25~100+美元	<5 美元
容量		2TB-128TB	64GB - 4TB
性能	并行性	高 / 并行访问	一般 / 单进程访问
	延迟	以最少的延迟量访问存储设备，要求极低延迟	对于一般用户而言，可接受的延迟相对较长
可靠性	数据保护能力	高：断电保护、端到端数据保护	低：基本数据保护，一般无断电保护
	数据安全	硬件加密，符合企业安全标准	基本数据安全功能
	UBER（不可纠误码率）	$\leq 10^{-18}$	$\leq 10^{-15}$
	MTBF（平均无故障时间）	200-300 万小时	100-150 万小时
耐用性	工作负载频度	24 小时×365 天	大部分时间处于空闲状态

注：UBER 指在应用错误纠正机制后，每比特读取操作中仍无法修复的错误数量占总读取量的比率，用于量化数据损坏风险。企业级 SSD 的 UBER 要求严苛，通常需低于 10^{-17}（即每读取 10^{17} 比特数据，最多出现 1 次不可修复错误），比消费级 SSD 的标准高两个数量级。

资料来源：电子发烧友网公众号，天风证券研究所

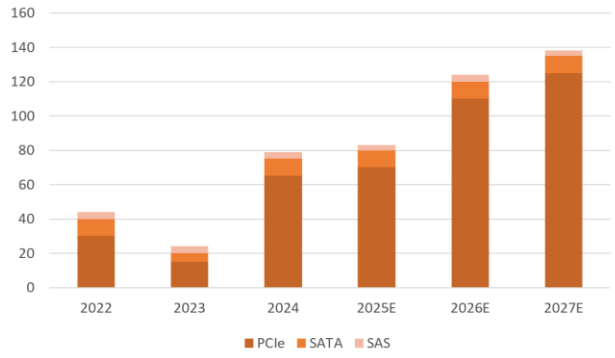
根据 Forward Insights 统计，2022 年，全球企业级 SSD 市场规模为 204.54 亿美元，并将随着存储行业需求提振不断增长，预计 2027 年市场规模将达到 514.18 亿美元，年复合增长率达到 20.25%，其中，PCIe 接口的企业级 SSD 占主导且占比持续上升，其在终端数据中心等场景的应用覆盖率不断增加。**2022 年，中国企业级 SSD 市场规模为 44.71 亿美元，**预计中国企业级固态硬盘市场规模将保持增长，2027 年将达到 135.09 亿美元，年复合增长率为 24.75%。

图 18：全球企业级 SSD 市场规模（单位：亿美元）



资料来源：Forward Insights，电子发烧友网公众号，天风证券研究所

图 19：中国企业级 SSD 市场规模（单位：亿美元）



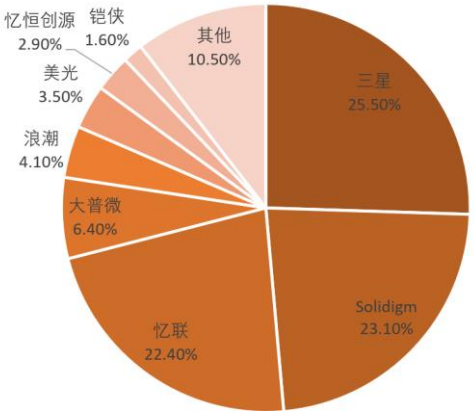
资料来源：Forward Insights，电子发烧友网公众号，天风证券研究所

企业级 SSD 行业具有研发难度高、技术迭代快、客户培育周期长、资金投 入大等特点。国外龙头企业起步较早，在生产技术、产品性能、品牌知名度等 方面具有较强竞争优势，在市场中处于主导地位，全球范围内呈现韩国（三星 和 SK 海力士）优势显著，美国、日本紧随其后，中国奋起直追的局面。我国企业级 SSD 行业起步相对较晚，市场份额小，整体生产技术与国际先进水平相比存在一定差距，本土企业有较大发展空间以及较长国产化替代过程。

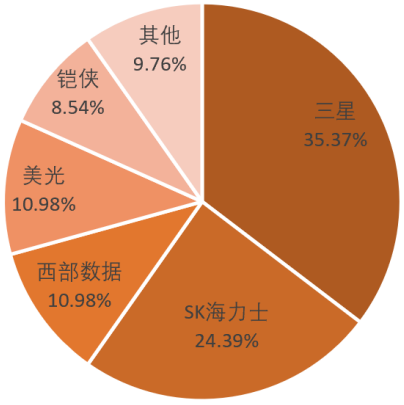
全球企业级 SSD 市场目前集中度较高，5 家龙头企业三星、SK 海力士、西 部数据、美光和铠侠占据了全球 90%以上的市场份额，这些公司在半导体存储领域拥有深厚的技术积累、广泛的产品布局和强大的研发能力。AI、云计算、大数据等新一代信息技术领域的快速发展带动了企业级 SSD 的市场需求和技术进步，同时也为其他新兴存储厂商提供了提高市场份额的宝贵机会，推动了整个行业的技术创新和市场多元化。

图 20：2023 年中国企业级固态硬盘市场份额

图 21：2023 年全球企业级 SSD 市场份额情况



资料来源：IDC，电子发烧友网公众号，天风证券研究所

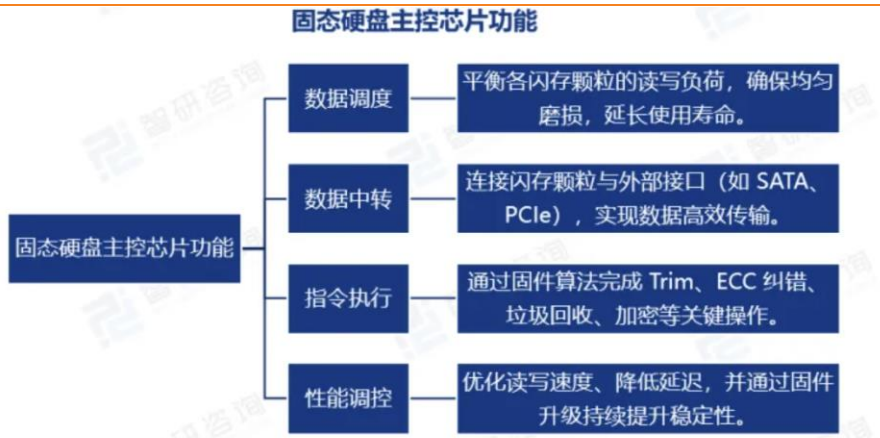


资料来源：上市公司年报，集邦咨询，电子发烧友网公众号，天风证券研究所

4.2. SSD 主控：AI SSD 智能化演进的核心驱动力

固态硬盘主控芯片是固态硬盘的 “大脑”，主要用于管理存储颗粒中数据的写入、读取与擦除，并与系统厂商推出的各类外部计算机或电子设备 CPU 进行通信和数据交换。主控芯片能够通过数据存储管理和数据纠错算法，有效降低存储颗粒中存储单元因器件特性导致数据挥发等因素而引起数据错误的概率。

图 22：SSD 主控芯片功能

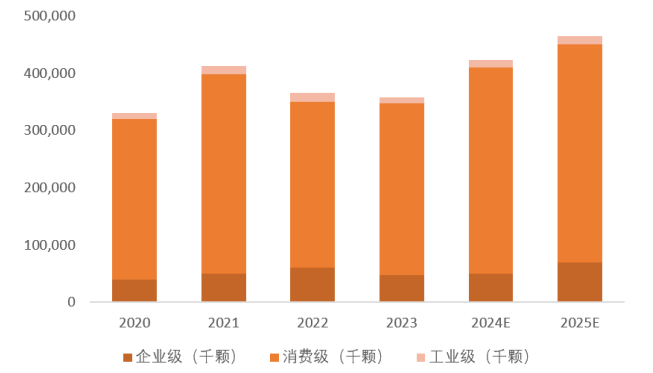


资料来源：智研科信公众号，天风证券研究所

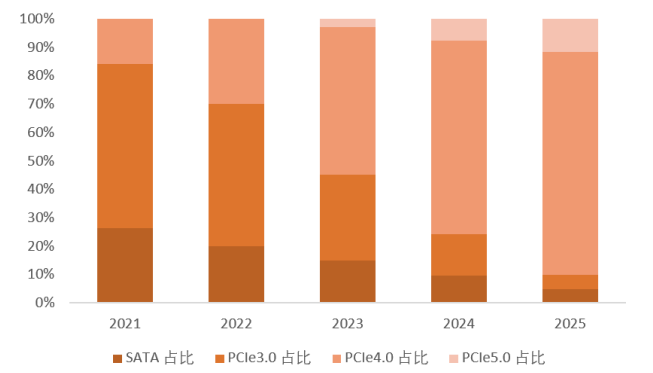
SSD 主控芯片出货呈上升趋势。2024 年，随着生成式 AI 浪潮席卷全球，企业级数据中心市场迎来需求复苏，PC 及消费类市场尽管相对温和，但同样处于上扬态势。根据 CFM 闪存市场数据，2024 年全球 SSD 主控市场共出货约 3.885 亿颗，相较 2023 年增长 8%。按照 SSD 接口渗透率变化看，2024 年全球 PC 前装市场搭载的 SSD 模组主力依旧为 PCIe4.0，相较 2023 年，PCIe5.0 接口 SSD 产品已在 PC 前装市场获得一定数量的商用，CFM 闪存市场预计 2025 年这一比例将快速提升。

图 23：2020 年-2025 年全球 SSD 主控芯片出货量情况

图 24：SSD 接口渗透率变化



资料来源：中国闪存市场，联芸科技招股说明书，天风证券研究所

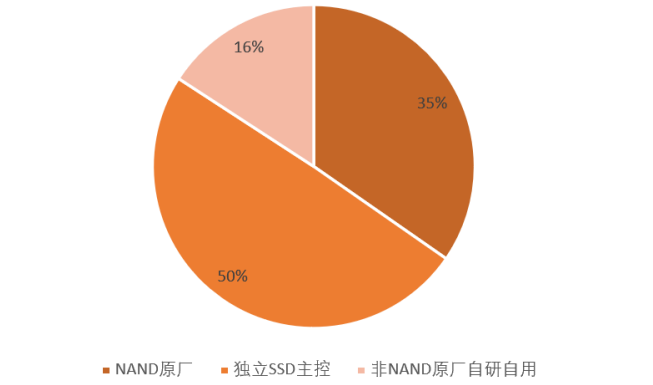


资料来源：中国闪存市场，联芸科技公司年报，天风证券研究所

全球 SSD 主控芯片厂商主要分为三类：第一类是 NAND 原厂的自研自用厂商（三星、海力士、美光、Solidigm、铠侠、西部数据等），其主控芯片仅用于自有 NAND 颗粒的模组生产而不单独出售；第二类是非原厂厂商，通过外采 NAND 颗粒搭配自研主控芯片生产模组，同时也会对外销售部分主控芯片；第三类是独立主控芯片厂商（慧荣科技、联芸科技、瑞昱、得一微等），专注于向市场直接销售主控芯片产品。这三类厂商共同构成了 SSD 主控芯片市场的完整产业格局。

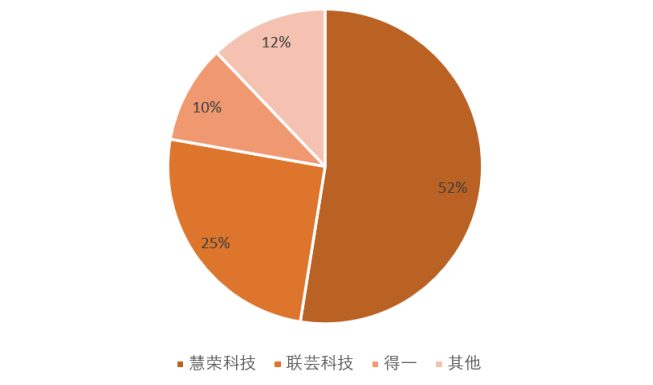
独立第三方主控厂商来看，联芸科技出货量占比 25%。CFM 闪存市场数据显示，2024 年慧荣科技 SSD 主控出货量约 1 亿颗，占比约 52%；联芸 SSD 主控出货量约 4900 万，占比约 25%；得一微电子 SSD 主控出货量约 2000 万，占比约 10%；其他厂商合计占比约 12%。

图 25：2024 年全球 SSD 主控芯片各类型厂商市场占有率



资料来源：CFM 闪存市场，天风证券研究所

图 26：2024 年独立第三方主控厂 SSD 主控出货份额



资料来源：CFM 闪存市场，天风证券研究所

4.3. SSD 技术趋势：AI 推理驱动 SSD 角色升维，QLC+PCIe 5.0/6.0 构筑未来趋势

在 AI 发展进程中，训练与推理两大核心环节对存储的特殊需求，直接推动了 AI SSD 的快速崛起。与标准固态硬盘不同，AI SSD 专为处理深度学习、神经网络训练和实时数据分析等人工智能应用的巨大数据吞吐量、低延迟和高 IOPS（每秒输入/输出操作数）需求而设计。

AI 推理环节中，SSD 可在推理过程中协助调整、优化 AI 模型，尤其 SSD 可以实时更新数据，以便微调推理模型结果。AI 推理主要提供检索增强生成（RAG，Retrieval-Augmented Generation）和大型语言模型（LLM，Large Language Model）服务，而 SSD 可以储存 RAG 和 LLM 参考的相关文档和知识库，以生成含有更丰富信息的响应。目前 TLC/QLC 16TB 以上等大容量 SSD 便成为 AI 推理主要采用的产品。

AI 对存储“高性能、大容量、高能效”的三重刚需，让 SSD 成为 AI 场景下的最优解。TrendForce 数据显示，全球范围内，2024 年 AI 相关的 SSD 采购容量将超过 45EB，未来几年，AI 服务器有望推动 SSD 需求年增率平均超过 60%，而 AI SSD 需求在整个 NAND Flash（闪存）的占比有机会自 2024 年的 5%，上升至 2025 年的 9%。

图 27：不同存储各指标对比

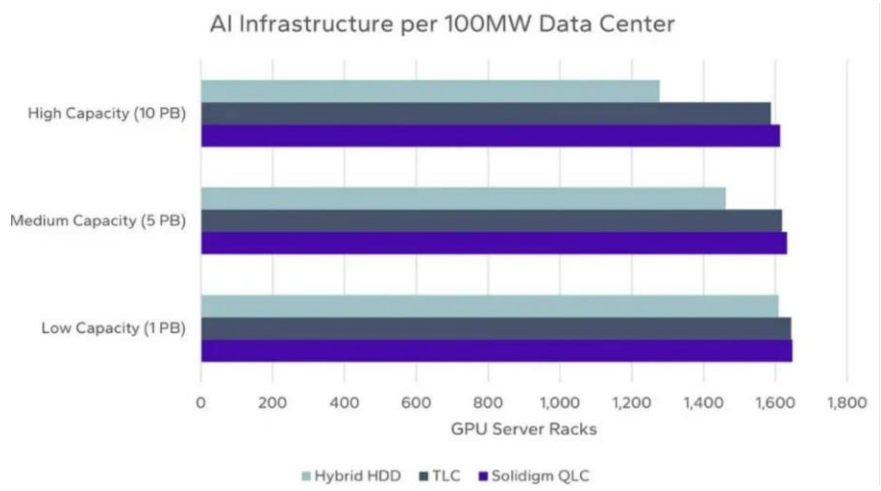
资料来源：铠侠，半导体产业纵横公众号，天风证券研究所

充分挖掘 SSD 的容量潜力以加速推理服务，成为亟待突破的关键点。以存代算中，多轮对话产生的 KV Cache 需在 HBM、DRAM、SSD 间动态流转（如华为 UCM 的分级缓存机制），其中 SSD 承担 “长期记忆数据” 存储角色，负责保存历史对话、RAG 知识库等热温数据（容量需求达 TB 至 PB 级）。相较于 HBM 的高成本和 DRAM 的容量限制，SSD 以较高的性价比提供了大容量存储支持，成为缓解 HBM 依赖的核心介质。

4.3.1. 技术趋势总览：存储与计算无缝配合，QLC+PCIe/NVMe+CXL 构筑下一代 AI SSD 基座

从颗粒的选择上，AI SSD 会朝着 QLC 颗粒方向走。铠侠 CEO 柳茂知也表示，QLC SSD 是 AI 行业最好的选择。尽管从 SLC 到 MLC，再到 TLC，最终到 QLC，SSD 的性能一直在下降，但随着技术的演变，2025 年 QLC SSD 的速度已经比 2017 年的 TLC SSD 快很多了。**如今 QLC SSD 的顺序读写速度可达 7000MB/s 左右，性能十分强大，能够满足 AI 大模型数据存储和调用的要求。**

图 28：基于 100 兆瓦数据中心的 AI 基础设施规模



资料来源：Solidigm，半导体产业纵横公众号，天风证券研究所

从传输接口与协议层面来看，采用 PCIe 接口并支持 NVMe 协议，未来大概率会成为 AI SSD 的标准配置。PCIe 接口凭借不断升级的带宽能力，从 PCIe 3.0 发展到如今的 PCIe 5.0，目前业内已经推进到了 PCIe 7.0。

此外，NVMe 协议专门针对闪存存储进行优化，为 SSD 提供了极高的 I/O 吞吐量和低延迟，这对于减少数据访问瓶颈非常重要。在 PCIe 接口之上构建了高效的数据访问机制，极大地降低了延迟，提升了 IOPS 性能，能充分发挥闪存的快速读写特性。随着技术发展，PCIe 接口和 NVMe 协议还会持续演进，融入如 CXL 等新兴技术。

4.3.2. 铠侠：高性能与高容量双轨并行，从硬件创新迈向软件定义智能

铠侠重点围绕 AI 驱动的存储技术创新、SSD 业务拓展及资本效率优化，以巩固其在 NAND

闪存市场的竞争力。

凯侠 AI SSD 有两类产品线：第一类是高性能 SSD。铠侠的 CM9 系列，专为 AI 系统设计，搭载针对数据中心优化的 PCIe 5.0，最大限度地发挥需要高性能和高可靠性的 GPU 功能。**第二类是容量型 SSD。**铠侠的 LC9 系列，适用于推理中使用的大型数据库等用例，当时容量为 122.88 TB，未来计划推出更大容量产品。

图 29：铠侠 LC9 系列 SSD

资料来源：KIOXIA 铠侠中国社公众号，天风证券研究所

对于未来的 AI SSD，铠侠也提出了自己的设想，主要从两个方面突破。第一是速度更快。现在的 SSD 每秒能处理 200 万-300 万次小文件读写，多采用 TLC 和 QLC ，而新产品将采用 XL-FLASH 的 SLC 闪存，速度提升到每秒 1000 万次以上，特别适合 AI 需要频繁读取零碎数据的场景。**第二是更智能。**目前 AI 检索数据要依赖内存，2026 年铠侠将推出 AiSAQ 软件，让 SSD 自己就能处理 AI 的检索任务。这样不仅能减轻内存负担，还能让 AI 应用运行更高效，尤其适合智能终端和边缘计算设备。

4.3.3. 美光：引领接口速率与存储密度，以性能与性价比重塑市场标杆

美光最新发布三款 AI SSD。

第一款是美光 9650 SSD，全球首款 PCIe 6.0 的 SSD，主要用在数据中心领域。能够提供 28 GB/s 的性能。据美光测试，相较于 PCIe 5.0 SSD，9650 SSD 的随机写入与随机读取的存储能效分别提升高达 25%和 67%。

第二款是美光 6600 ION SSD，单盘容量最高达 245TB，主要应用在超大规模部署与企业级数据中心整合服务器基础设施、构建大型 AI 数据湖。相较于竞品，该产品的存储密度提升高达 67%，单机架存储容量突破 88PB，大幅降低总体拥有成本（TCO）。

第三款是美光 7600 SSD，主要用于 AI 推理与混合工作负载。据称，能够在高度复杂的 RocksDB 工作负载下实现业界领先的亚毫秒级延迟。

4.3.4. Solidigm：聚焦 QLC 技术与液冷方案，以场景化存储优化 AI 效率

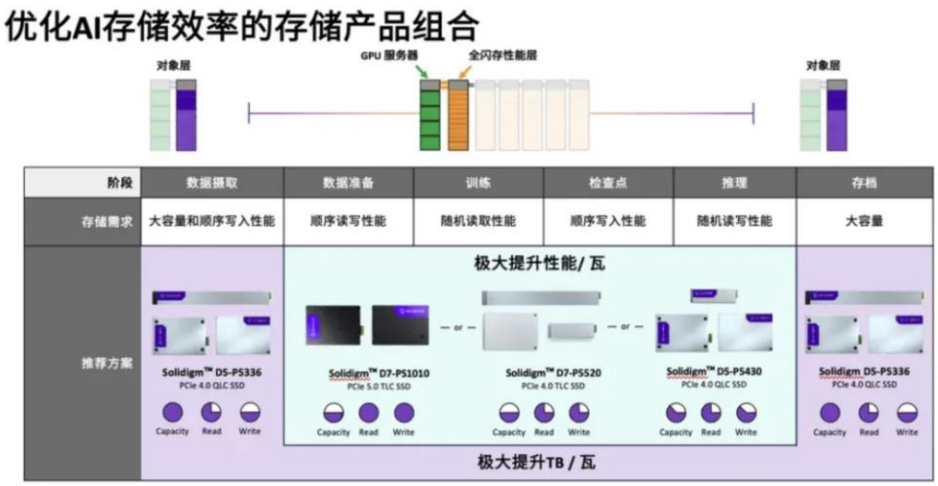
Solidigm 将 AI 存储方案大致分为两类。

一类是，直连式存储（DAS），针对训练等对性能极度敏感的场景，它更关注单位功耗下的 IOPS；另一类是网络存储（包括 NAS 文件/对象存储），针对数据摄取、归档和 RAG（检索增强生成）等大容量场景，对读性能要求较高，同时也追求最低成本存储海量数据。

Solidigm 在 AI SSD 中的另一个亮点是 QLC SSD。自 2018 年发布首款 QLC SSD 以来，Solidigm 已累计出货超过 100EB 的 QLC，并为全球 70%的领先 OEM AI 解决方案提供商提供服务。

Solidigm 不仅推动 QLC 技术的普及与应用，还在液冷 SSD 技术领域进行大胆尝试。2025 年 3 月，Solidigm 展示其首款采用 SolidigmD7-PS1010 E1.S 9.5mm 外形规格的冷板液冷 SSD，该方案显著提升了散热效率。

图 30：Solidigm 优化 AI 存储效率的存储产品组合



资料来源：Solidigm，半导体产业纵横公众号，天风证券研究所

5. 投资建议：关注国产“以存代算”相关芯片公司机遇

“以存代算”正从技术理念加速迈向产业现实，其通过存储资源扩展替代重复计算的范式，将深刻重构 AI 推理基础设施，并为存储产业链带来前所未有的增长机遇。 华为 CachedAttention、UCM 等技术方案的落地及全球巨头的快速跟进，已验证了其经济性与必要性。

建议关注：

存储模组厂商：江波龙（天风计算机联合覆盖）、德明利、佰维存储、朗科科技、联芸科技、万润科技等

存储芯片：兆易创新、普冉股份、北京君正、东芯股份、恒烁股份、澜起科技、聚辰股份等

存储分销与封测：香农芯创、深科技、太极实业、中电港等

5.1. 存储模组厂商

5.1.1. 江波龙：企业级 SSD 产品组合+自研主控芯片的双轮驱动

江波龙专注于半导体存储领域，为客户提供从产品设计、存储芯片、主控芯片设计及固件开发，到封装、测试、制造等全方位的存储定制服务。

国产替代龙头，企业级存储实现放量增长。2025H1 公司企业级存储业务规模增长明显，企业级存储业务收入达到 6.93 亿元，同比增长 138.66%。公司是国内少数具备“eSSD+RDIMM”产品设计、组合以及规模供应能力企业，在 eSSD 与 RIDMM 产品组合基础上，已成功点亮 SOCAMM 产品，结合 MRDIMM、CXL2.0 内存拓展模块构建了全面的企业级产品体系。公司的 eSSD 与 RIDMM 产品已成功完成鲲鹏、海光、飞腾等多个国产 CPU 平台服务器的兼容性适配，DDR5 RDIMM 产品也通过了 AMD 旗下 Threadripper PRO 9000WX 系列工作站 CPU 认证，为在主流平台上的广泛应用提供了坚实的技术基础。公司企业级业务进入快速增长阶段，客户涵盖运营商、大型及中型互联网企业、服务器企业等，产品已在通信、互联网、金融等行业历经多次严苛考验并成功交付。

主控自研修筑技术壁垒，TCM 模式不断突破，加码长期高毛利高壁垒。截至 2025 年 7 月

底，公司主控芯片全系列（含 UFS\eMMC\SD\高端 U 盘等场景，下同）产品累计实现超过 8000 万颗的批量部署。公司设计并成功流片了历史上首批 UFS 自研主控芯片，搭载公司自研主控的 UFS4.1 产品的整体性能超越市场同类产品。基于公司 UFS 主控芯片的技术实力优势，公司已与闪迪达成战略合作，共同面向移动及 IOT 市场推出定制化的高品质 UFS 产品及解决方案。

图 31：江波龙企业级 SSD

资料来源：江波龙公众号，天风证券研究所

5.1.2. 佰维存储：产品布局与 AI 战略深度融合，SSD 技术领先

佰维存储专注于半导体存储器的研发设计、封装测试、生产和销售，核心产品及服务涵盖半导体存储器和先进封测服务，具体可分为六大产品线：嵌入式存储、PC 存储、工车规存储、企业级存储、移动存储和先进封测服务。

公司企业级存储有 4 大类别，分别为 SATASSD、PCIeSSD、CXL 内存及 RDIMM 内存条，主要应用于数据中心、通用服务器、AI/ML 服务器、云计算、大数据等场景。公司 SS 系列企业级 2.5"SATASSD 产品包含 SS811、SS821 等系列，支持异常掉电保护、端到端的数据保护、Thermal Throttling、动态和静态磨损平衡、支持电源动态管理、S.M.A.R.T、垃圾回收和 TRIM、固件备份、InternalRAID 等特性。公司 SP 系列企业级 PCIe SSD 产品，包含 Gen4 和 Gen5 两类产品。用创新架构，可实现超低且一致的读写延迟，具备优秀的能效比表现，可为客户提供业界领先的 KIOPS/Watt 综合性能。

图 32：佰维存储企业级 SSD

资料来源：佰维 BIWIN 公众号，天风证券研究所

5.1.3. 德明利：“芯片 + 算法 + 场景”全链条发展

德明利是一家专注于国产存储主控芯片研发及存储模组方案的集成电路企业，以“芯片+算法+场景”全链条技术能力，为智能终端、数据中心、工业控制等高价值场景提供高可靠性存储解决方案。公司产品线涵盖固态硬盘类、嵌入式存储类、内存条类及移动存储类四大系列，已广泛应用于数据中心、手机、车载电子、平板、安防监控等多元应用场景。

德明利 ES1020 系列工业级 SSD，作为行业主流存储模组厂商率先搭载自研主控芯片的工业级产品，实现从芯片到模组的全链路国产化设计制造。其自研 TW6501 芯片是国内首颗支持 ONFI 5.0 的 SATA SSD 主控，采用 RISC-V 架构，满足宽温要求。产品支持 SATA III 协议，提供 64GB-4TB 容量及 M.2/2.5 寸/mSATA 全形态，具备 200 万小时 MTBF、超 3K 次擦写寿命及 7×24 小时稳定写入性能。依托灵活可扩展的自研固件平台，为工业控制、安防监控、通信、电力等关键行业端侧 AI 硬件提供自主安全、稳定可靠的存储解决方案。

“2+3”自研主控拓展提速，加快高性能领域研发创新。公司闪存模组以自研主控芯片为核心，随着研发实力不断增强，芯片研发不断拓展提速，同时推动两颗主控量产导入，三颗主控芯片立项研发。公司新一代自研 SD6.0 存储卡主控芯片成功量产，产品适配工作进展顺利，目前已有各类搭载该款主控的存储卡模组送样，待客户验证通过后即可实现批量导入；公司自研 SATASSD 主控芯片成功量产，已经完成产品适配与测试，并实现批量销售。新一代 SD6.0 存储卡主控芯片主要基于 40nm 工艺，提升读写性能，基于 multi-voltage 多电源域的低功耗设计方法，SATASSD 主控芯片为国内率先采用 RISC-V 指令集打造的无缓存高性能控制芯片，支持最新的 ONFI5.0 接口，二者均采用目前业界领先的 4KLDPC 纠错技术，可以灵活适配 3DTLC/QLC 等不同类型的闪存颗粒。

图 33：德明利 ES1020 系列工业级 SSD

资料来源：芯师爷公众号，天风证券研究所

5.2. 存储芯片设计

5.2.1. 兆易创新：利基存储格局优化，端侧 AI 推动定制化需求增长。

公司专用型存储芯片包括 NORFlash、SLCNANDFlash 和利基型 DRAM 三条产品线，形成了丰富的产品矩阵，满足客户在不同应用中对容量、电压以及封装形式的多元需求，已在消费电子、工业、通讯、汽车电子等领域实现了全品类覆盖。

NOR Flash 方面，公司产品覆盖 512Kb 到 2Gb 的容量范围，支持 1.2V、1.8V、3V、1.65~3.6V 以及 1.8VVCC&1.2VVIO 等多种供电类型，并针对不同市场应用需求分别提供高性能、低功耗、高可靠性、小封装等多个产品系列，可满足客户在不同应用领域多种产品应用中对容量、电压以及封装形式的需求。2025 年上半年，公司推出了专为 1.2VSoC 应用打造的双电压供电 SPINOR Flash 产品，进一步强化公司在双电压供电闪存解决方案领域的战略布局，为市场提供先进嵌入式存储解决方案，可应用于智能可穿戴设备、医疗健康、物联网、数据中心及边缘人工智能等新兴领域。2025 年，公司为率先实现 45nm 节点 SPINORFlash 大规模量产的公司之一，存储密度得到显著改善，持续保持技术和市场的领先。

SLC NAND Flash 方面，公司产品容量覆盖 1Gb~8Gb，采用 3V/1.8V 两种电压供电，具有高速、高可靠性、低功耗的特点，其中 SPINANDFlash 在消费电子、工业、汽车电子等领域已经实现了全品类的产品覆盖。2025 年上半年，公司推出了兼备更快读取速度和坏块管理功能的高速 QSPINANDFlash 产品，可应用于工业、IoT 等快速启动应用场景。

公司利基型 DRAM 产品广泛应用于网络通信、电视、机顶盒、智能家居、工业等领域。2025 年上半年，公司 8Gb 容量 DDR4 产品市场推广顺利进行，营收稳步增长；LPDDR4 产品开始贡献营收。公司控股子公司青耘科技开展的定制化存储业务正有序推进中，业务进展顺利。

5.2.2. 联芸科技：高壁垒“存储大脑”主控赛道龙头，AIoT 芯片带动第二增长曲线

联芸科技是国内领先的数据存储主控芯片及 AIoT 信号处理芯片设计企业，业务覆盖消费电子、工业控制、智能物联等领域。

公司主控芯片核心产品出货量全球领先。在 SSD 主控领域，公司已实现从 SATA 到 PCIe 5.0 的全协议覆盖，构建了消费级、企业级和工业级的全场景产品矩阵。**消费级 SSD：**公司全面布局 SATA、PCIe 3.0 及 PCIe 4.0 主控芯片产品线，凭借高性能、低功耗和优异的兼容性，出货量实现稳定增长，并在头部笔电前装市场实现大规模商用。**企业级 SSD：**高性能 SATA 主控芯片已获得主流服务器和系统等客户的认可，并实现大规模商用，为下一代企业级 PCIe SSD 主控芯片的研发和市场推广奠定坚实基础。**工业级 SSD：**公司已完成 SATA、PCIe 3.0 及 PCIe 4.0 主控芯片的全平台布局。

图 34：截至 2024 年联芸科技目前已成熟量产的主控芯片

公司目前已经成熟量产的主控芯片产品

产品系列	推出时间	接口类型	应用领域	顺序读写性能
MK6XX 系列	2015 年	SATA	工业级	400MB/s、400MB/s；50K IOPS、30K IOPS
MK8XX 系列	2016 年	SATA	工业级	500MB/s、450MB/s；90K IOPS、70K IOPS
MAS090X 系列	2017 年	SATA	企业级、消费级 / 工业级	560MB/s、5300MB/s；100K IOPS、80K - 90K IOPS
MAP100X 系列	2019 年	PCIe	消费级 / 工业级	2,600 - 3,500MB/s、2,400 - 3,000MB/s；350K - 800K IOPS、300K - 600K IOPS
MAS110X 系列	2020 年	SATA	消费级 / 工业级	560MB/s、530MB/s；100K IOPS、80K IOPS
	2021 年	SATA	企业级	560MB/s、530MB/s；100K IOPS、90K IOPS
MAP120X 系列	2021 年	PCIe	消费级 / 工业级	3,600MB/s、3,200MB/s；600K - 800K IOPS、500K - 600K IOPS
MAP160X 系列	2022 年	PCIe	消费级 / 工业级	7,400MB/s、6,500MB/s；1,000K - 1,500K IOPS、1,000K IOPS

资料来源：联芸科技年报，天风证券研究所

5.3. 其他重点公司

表 1：公司估值表（截至 2025 年 9 月 26 日）

公司/估值信息	总市值	EPS	PE

		（亿元）	2025E	2026E	2027E	2025E	2026E	2027E
存储模组主控	江波龙	590.99	1.81	2.91	3.75	78.04	48.65	37.73
	佰维存储	420.60	0.84	1.51	2.03	108.25	60.29	44.64
	德明利	390.93	2.61	3.76	4.87	66.17	45.92	35.44
	联芸科技	252.54	0.27	0.37	0.52	203.56	147.28	106.27
	万润科技	136.52	--	--	--	--	--	--
	朗科科技	54.63	-0.15	0.73	1.08	-176.22	37.42	25.17
存储芯片	澜起科技	1,524.20	2.04	2.71	3.49	65.41	49.11	38.09
	兆易创新	1,267.40	2.36	3.11	3.91	80.97	61.36	48.81
	东芯股份	463.83	-0.06	0.15	0.36	-1,927.15	691.46	289.46
	北京君正	394.77	1.05	1.34	1.66	77.87	61.23	49.47
	聚辰股份	224.86	2.81	3.88	4.99	50.60	36.68	28.52
	普冉股份	146.57	1.80	2.43	3.08	54.96	40.82	32.15
	恒烁股份	39.47	--	--	--	--	--	--
存储分销封测	深科技	383.36	0.80	0.93	1.11	30.57	26.55	22.15
	香农芯创	361.05	1.31	1.90	2.48	59.53	41.13	31.42
	太极实业	162.09	0.30	0.31	0.32	26.02	25.05	23.73
	中电港	162.09	--	--	--	--	--	--

资料来源：iFinD，天风证券研究所　　注：表格按公司总市值排序，EPS 预测数据来自 iFinD 一致预期

6. 风险提示

地缘政治带来的不可预测风险：随着地缘政治冲突加剧，美国等国家/地区相继收紧针对半导体行业的出口管制政策，国际出口管制态势趋严，经济全球化受到较大挑战，对全球半导体市场和芯片供应链稳定带来不确定风险。未来如美国或其他国家/地区与中国的贸易摩擦升级，限制进出口及投资，提高关税或设置其他贸易壁垒，半导体行业相关公司还可能面临相关受管制设备、原材料、零备件、软件及服务支持等生产资料供应紧张、融资受限的风险等，进而对行业内公司的研发、生产、经营、业务造成不利影响。

需求复苏不及预期：受到全球宏观经济的波动、行业景气度等因素影响，集成电路行业存在一定的周期性，与宏观经济整体发展亦密切相关。如果宏观经济波动较大或长期处于低谷，集成电路行业的市场需求也将随之受到影响。另外，下游市场需求的波动和低迷亦会导致集成电路产品的需求下降，或由于半导体行业出现投资过热、重复建设的情况进而导致产能供应在景气度较低时超过市场需求。

技术迭代不及预期：集成电路行业属于技术密集型行业，集成电路涉及数十种科学技术及工程领域学科知识的综合应用，具有工艺技术迭代快、资金投入大、研发周期长等特点。多年来，集成电路行业公司坚持自主研发的道路并进一步巩固自主化核心知识产权。如果行业内公司未来技术研发的投入不足，不能支撑技术升级的需要，可能导致公司技术被赶超或替代，进而对公司的持续竞争力产生不利影响。

产业政策变化风险：集成电路产业作为信息产业的基础和核心，是国民经济和社会发展的战略性产业。国家陆续出台了包括《国务院关于印发进一步鼓励软件产业和集成电路产业发展若干政策的通知》（国发[2011]4 号）在内的一系列政策，从财税、投融资、研究开发、进出口、人才、知识产权、市场应用、国际合作等方面为集成电路企业提供了更多的支持。未来如果国家相关产业政策出现重大不利变化，将对行业发展产生一定不利影响。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

特别声明

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	海口	上海	深圳
北京市西城区德胜国际中心 B 座 11 层	海南省海口市美兰区国兴大道 3 号互联网金融大厦 A 栋 23 层 2301 房	上海市虹口区北外滩国际客运中心 6 号楼 4 层	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100088	邮编：570102	邮编：200086	邮编：518000
邮箱：research@tfzq.com	电话：(0898)-65365390	电话：(8621)-65055515	电话：(86755)-23915663
	邮箱：research@tfzq.com	传真：(8621)-61069806	传真：(86755)-82571995
		邮箱：research@tfzq.com	邮箱：research@tfzq.com