



电子行业研究

买入（维持评级）

行业深度研究

证券研究报告

电子组

分析师：樊志远（执业 S1130518070003）

fanzhiyuan@gjzq.com.cn

AI 驱动存储新周期

行业观点

首先，我们复盘了前面几轮半导体周期，根据供需关系按照时间维度分为长中短三种周期：长周期（8-10 年）——为需求周期、中周期（4-6 年）——为产能周期、短周期（3-5 个季度）——为库存周期，其中需求-产能-库存三种周期相互嵌套。存储器是半导体中仅次于逻辑的第二大细分市场，其历史表现与整个半导体周期走势一致，但波动性大于整个行业，大市场与强周期属性并存。通过复盘我们发现每轮存储大周期（08 年、16 年等）的开启都是由新兴技术推动产品升级和创新，进而催生新产品的总量、渗透率和存储器价值量的提升，推动存储器市场规模上升一个台阶，随着 AI 驱动需求提升，当下我们走在新一轮存储大周期的起点。

AI 驱动下，会出现存储大周期吗？1）我们看见大模型类型和机制转变正产生大量数据存储需求，首先是大模型中引入的思维链提示使 LLM 能将复杂问题分解为可操作步骤，复现类似人类的推理过程，显著提升模型的推理能力与问题解决技能，推理时长的增加也提升 Token 的消耗量，2025 年几乎所有主流大模型都已内化思维链机制。同时根据希捷科技的数据，从文本向音视频的切换，背后的存储单位是从 KB 往 TB 乃至 EB 的增长。随着多模态模型渗透率的提升，存储需求有望进一步提升。2）其次，我们看见头部模型差距缩小，推理降本趋势明显。从 ChatGPT3 发布以来，大语言模型的推理的成本以指数级别下降，单美元可以生成的 token 数量持续增长。模型的推理成本快速降低，有望带动应用的爆发，也有望拉动存储需求的增长。3）最后，我们看见 AI 正在重新定义数据存储，KV Cache 或将成为大模型推理优化的关键突破。KV Cache 是 Transformer 架构中显著提升推理效率的关键性能优化机制，KV Cache 显存占用随 Token 数量线性增长，同时 KV Cache 的优化效果也与文本长度正相关。当 KV Cache 成为未来数年提升大模型推理效率的关键一环，存储的需求也有望进一步提升。

AI 驱动存储需求快速攀升，存储原厂资本开支还未进入扩张周期。DRAM 和 NAND Flash 原厂的重心正逐渐转变，从单纯扩产转向制程技术升级、高层数堆栈、混合键合以及 HBM 等高附加价值产品。根据 TrendForce 的数据，2025 年 DRAM 的资本支出预计为 537 亿美元，预计在 2026 年进一步成长至 613 亿美元，同比增长约 14%。2025 年 NAND Flash 的资本支出预计为 211 亿美元，2026 年预计小幅增长至 222 亿美元，同比增长约 5%。当前行业无尘室空间已接近瓶颈，各大 DRAM 厂商中仅三星与 SK hynix 仍具备有限的扩线空间，即使资本开支上修，26 年产能增量亦有限。而在 NAND Flash 领域，Kioxia/SanDisk 相对扩产更为积极，其他原厂则继续专注 HBM 及 DRAM。

投资建议

我们建议持续重点关注存储器库存、价格数据以及 AI 算力提升对存储芯片的需求拉动。

风险提示

需求低于预期，存储器价格下跌，存储器价格上涨挤压其他领域需求。



内容目录

一、周期复盘：半导体与存储周期.....	4
1.1 长周期（8-10 年）：看需求.....	4
1.2 中周期（4-6 年）：看产能.....	5
1.3 短周期（3-5 个季度）：看库存.....	6
1.4 存储周期复盘：大市场与强周期并存.....	7
二、AI 驱动的会是大周期吗？.....	9
2.1 大模型类型和机制转变正产生大量数据存储需求.....	9
2.2 头部模型差距缩小，推理降本趋势明显.....	11
2.3 AI 正在重新定义数据存储，KV Cache 或将成为大模型推理优化的关键突破.....	12
三、AI 驱动存储需求快速攀升，存储原厂资本开支还未进入扩张周期.....	15
四、投资建议.....	19
五、风险提示.....	19

图表目录

图表 1： 产品需求推动全球半导体上行大周期.....	4
图表 2： 2003-2011 年之间全球 PC 出货量快速攀升.....	5
图表 3： 2012-2016 年全球智能手机出货量快速提升.....	5
图表 4： 全球半导体行业每 4-6 年经历一轮完整产能周期.....	5
图表 5： 半导体行业“复苏-扩张-顶峰-衰退”四大阶段.....	6
图表 6： 全球半导体资本开支增速 vs. 半导体销售额增速.....	6
图表 7： 从芯片厂商库存看：全球半导体行业每 3-5 个季度经历一轮库存周期.....	7
图表 8： 艾睿电子存货周转天数变化.....	7
图表 9： 安富利存货周转天数变化.....	7
图表 10： 存储板块通常大市场与强周期并存.....	8
图表 11： 存储价格具有明显周期性，以 DDR3 4Gb 为例（单位：美元）.....	8
图表 12： DRAM 与 NAND 指数具有类似的周期性.....	8
图表 13： 存储周期复盘.....	9
图表 14： 思考链复现类似人类的推理过程.....	10
图表 15： 引入思考链提升了大模型的“智力”与“计算能力”.....	10
图表 16： 引入思维链显著提升大模型的性能表现.....	10
图表 17： 增加思考时间及 Token 消耗进一步提升模型表现.....	10
图表 18： 2025 年大模型线路图.....	11
图表 19： 生成式 AI 对存储的需求持续提升.....	11



图表 20: 不同阶段的科技创新对存储的需求持续提升.....	11
图表 21: 大模型推理成本按照指数下降, 每年降低幅度约 10 倍.....	12
图表 22: Deepseek 依靠推理端算法优化可以实现较好智能程度 (单位: %)	12
图表 23: 西部数据针对不同的存储要求, 将存储设备和架构规划分成了五层.....	13
图表 24: KV Cache 显存占用随 Token 数量线性增长.....	14
图表 25: 不同的模型架构, 存储结构略有差异.....	14
图表 26: KV Cache 技术在不同场景下的性能提升效果.....	15
图表 27: 2024 年全球 DRAM 和 NAND 合计占据 97% 的份额.....	16
图表 28: 主要 NAND Flash 厂商 2025 年主动控产稳价.....	16
图表 29: 3D NAND 堆叠层数突破与 QLC/PLC 演进.....	16
图表 30: 2025 年全球 DRAM 产能增幅有限 (单位: 千片/月)	17
图表 31: 主要 NAND 供应商的 Roadmap 显示出向更高层数和更高存储密度的趋势	17
图表 32: 26 年 DRAM 与 NAND Flash 产业的资本支出和扩产仍未进入扩张周期.....	18
图表 33: 受 AI 需求影响, 各大存储厂商新的产能将优先考虑 HBM 的供应	18



一、周期复盘：半导体与存储周期

半导体行业是一个分工较为精细的行业，产业链自下而上的环节分别为：终端客户-渠道及分销商-IC设计公司（Fabless）/IDM-晶圆代工厂（Fab）/封测厂-半导体设备/材料，这么长的链条导致在终端需求发生变化时，容易发生上游不同环节反应不及时，因此也导致了半导体行业呈较为明显的周期性。但其周期性背后的根本归因是：供需关系。

存储板块作为半导体最大的细分领域之一，存储的周期与半导体板块的周期既有相似的地方，也具有不同的特色，在复盘存储周期之前，我们先复盘前面几轮半导体周期，根据供需关系按照时间维度分为长中短三种周期：长周期（8-10年）——为需求周期、中周期（4-6年）——为产能周期、短周期（3-5个季度）——为库存周期，其中需求-产能-库存三种周期相互嵌套，共同推动半导体行业波浪式前进。

1.1 长周期（8-10年）：看需求

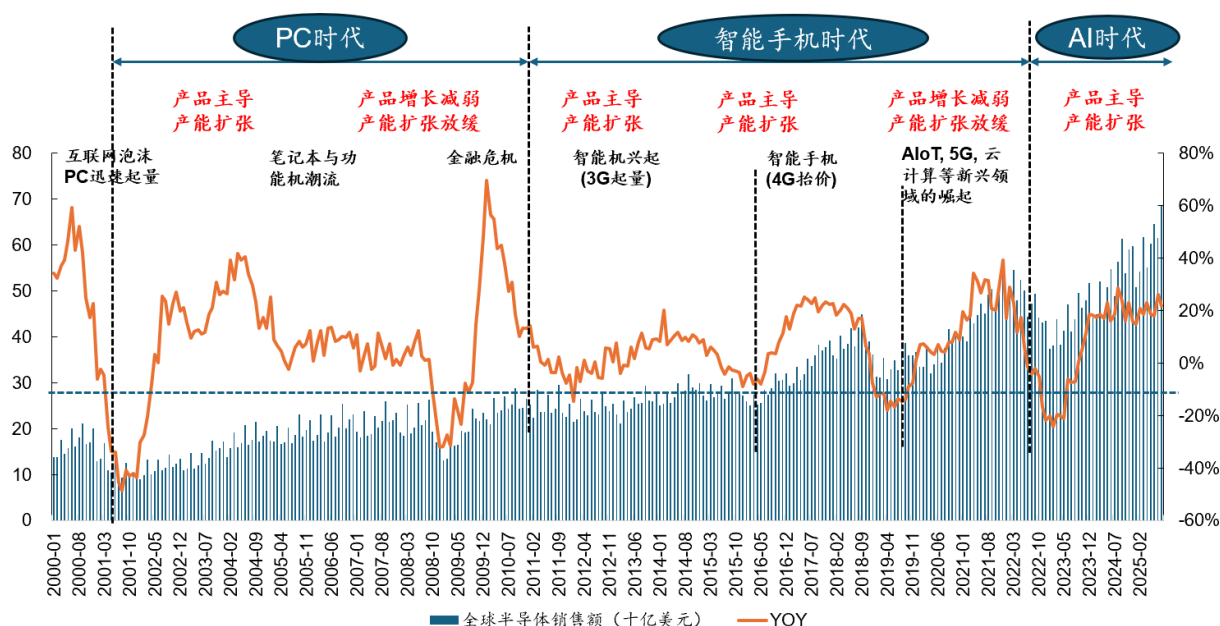
我们复盘过去二十多年以来半导体行业的发展历程发现，每一次大的技术革命都能驱动半导体维持将近十年的景气度，其核心逻辑是：底层的新技术催生新产品的总量、渗透率和单品半导体价值量的提升。例如：

2001-2010年前后：我们称之为PC时代，自互联网泡沫破灭之后，PC终端迅速兴起，随之而来的功能机、笔记本等陆续出现，根据Gartner数据全球PC出货量，2003年单季度为4000万部左右，到2011年前后，单季度出货量达到约9000万部。随着超薄光驱、TFT液晶屏、锂电池以及无线通讯技术等的发展促使笔记本电脑向更轻薄的发展方向，其内置CPU芯片的厚度、面积均开始大幅下降，含硅量持续提升。根据WSTS数据，全球半导体销售额从2001年1390亿美元提升至2010年的2989亿美元，10年复合增速达9%。

2010-2021年，2010年前后，苹果开启了智能手机时代，从3G手机到4G、5G手机，全球智能手机出货量从2012Q3的1.86亿部，到2016年平均每季度超过4亿部。以及由之而来的TWS耳机、智能家居AIOT终端和云计算等新兴领域迅速崛起，也推动了算力CPU、存储器、SoC、电源管理等芯片的需求。全球半导体销售额也从2011年的2995.2亿美元成长到2021年的5558.9亿美元，10年复合增速为7%。

2022年全球半导体行情周期下行，行业处于展望悲观中。随之而来ChatGPT于2023年初横空出世，为半导体行业前行带来曙光。AI驱动基础设施建设，包括算力芯片、存储器、交换机、服务器、PCB等全产业链需求回暖，2023年下半年下游终端开启补库周期，2025年9月全球半导体销售额达610.97亿元，单月销售额创历史新高。

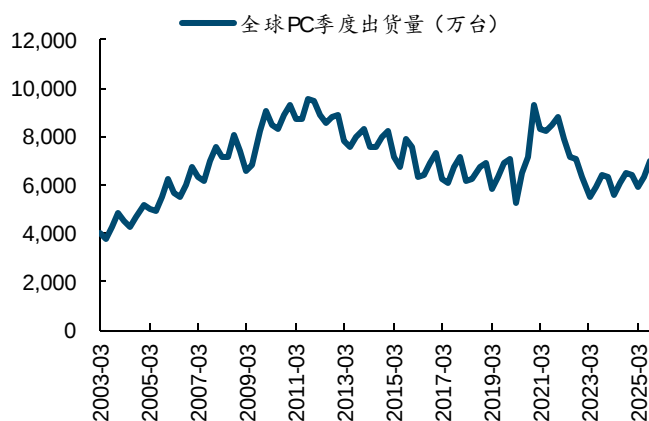
图表1：产品需求推动全球半导体上行大周期



来源：WSTS，国金证券研究所

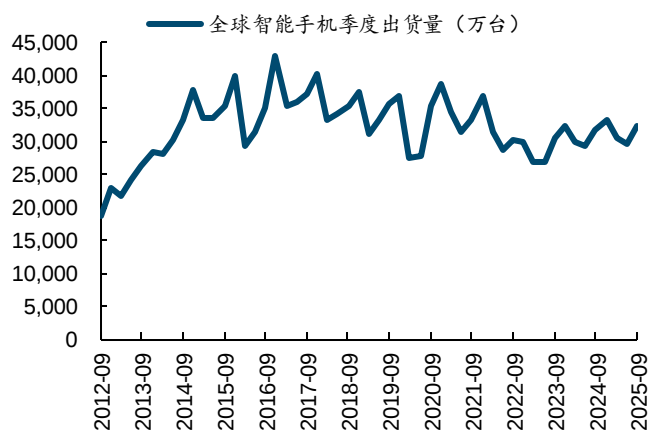


图表2：2003-2011 年之间全球PC出货量快速攀升



来源：Gartner, iFind, 国金证券研究所

图表3：2012-2016 年全球智能手机出货量快速提升

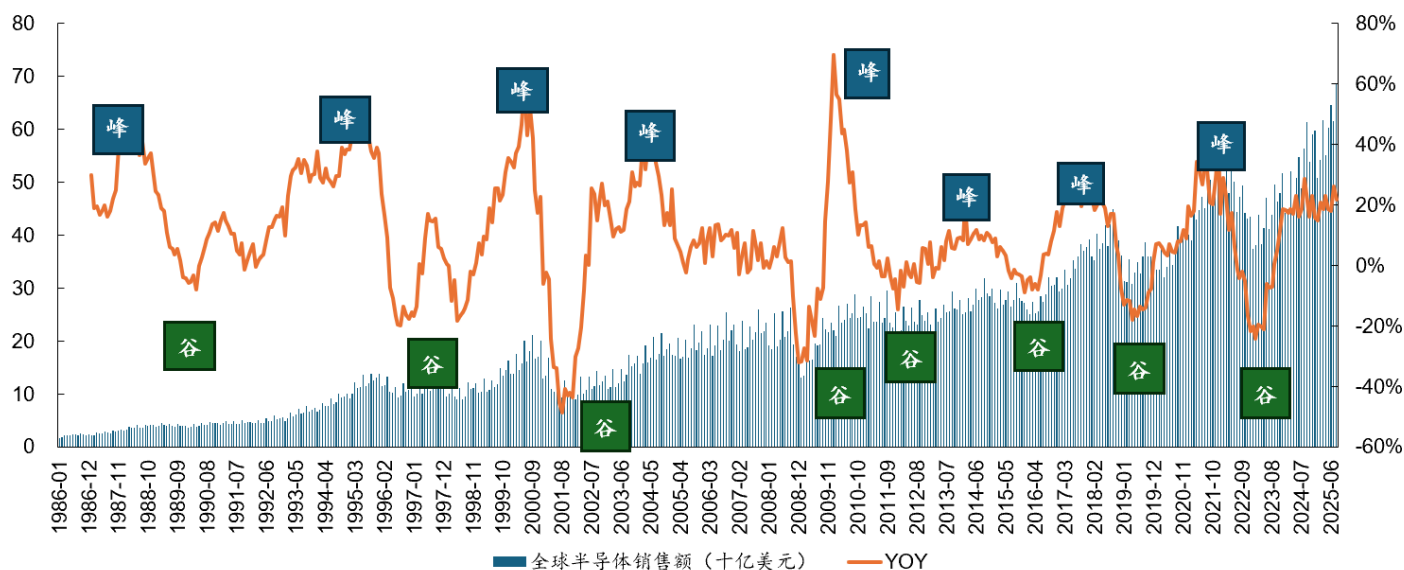


来源：IDC, iFind, 国金证券研究所

1.2 中周期（4-6 年）：看产能

回顾 1986 年以来半导体的发展，可以发现半导体行业每 4-6 年经历一轮完整的波峰-波谷-波峰的轮转。我们认为其背后核心为晶圆厂、封测厂的资本开支与产能扩张进度的驱动，称之为“产能周期”即在供给驱动下的半导体行业重复演绎着复苏—扩张—高峰—衰退—复苏...的过程。

图表4：全球半导体行业每 4-6 年经历一轮完整产能周期



来源：WSTS, 国金证券研究所

我们将产能周期分为四个阶段：复苏-扩张-顶峰-衰退，四个阶段依次交替轮转。

复苏：这个阶段，经历前面的衰退，各环节供应商较为谨慎，表现为全产业链条库存水位较低。但随着下游需求开始改善，终端客户开始拉货，传导到渠道端、芯片设计公司，其订单开始增加，出现被动去库存现象，同时上游晶圆厂小幅增加资本开支。

扩张：随着终端需求持续旺盛，渠道端及 IC 设计公司芯片出现供不应求、芯片交期延长，导致芯片价格上涨，出现主动补库存现象，同时上游晶圆厂稼动率持续满产，开始加大资本开支。

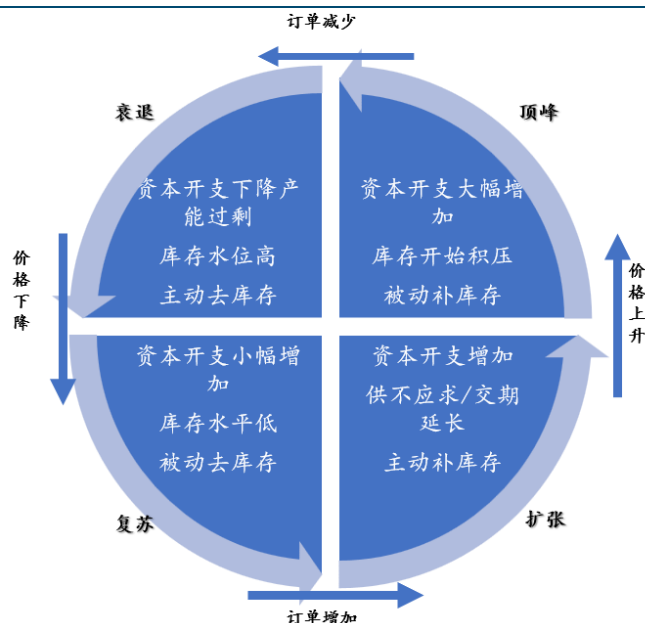
顶峰：芯片价格是一个明确的讯号，由于晶圆厂从计划到投产、量产基本上需要 1-2 年的周期，在产能紧缺的情况下，晶圆厂提高代工价格进一步推动芯片价格持续上涨，导致部分环节厂商开始囤货（以期转差价），层层叠加推高终端产品生产升本，压缩终端厂商利润，导致部分厂商减少订单，随着时间推移，芯片厂商库存开始积压，出现被动补库存现



象。

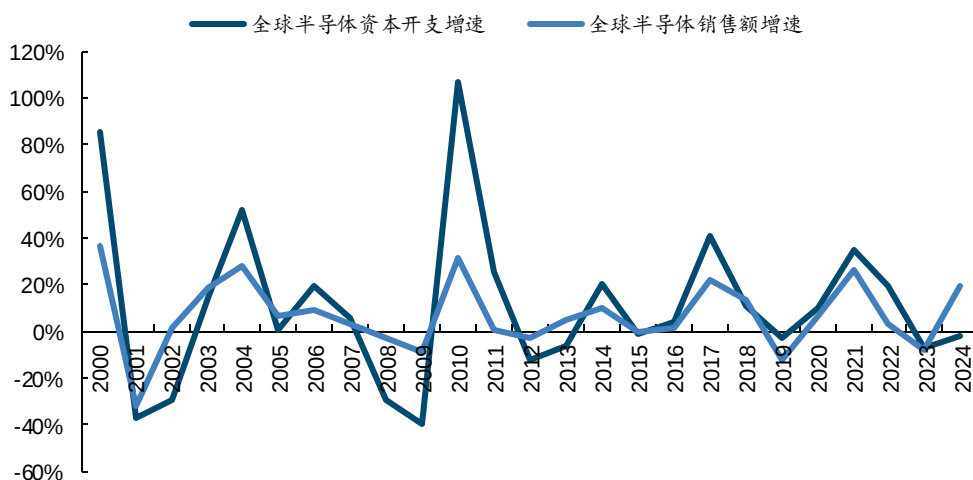
衰退：随着终端订单持续减少，芯片厂商库存水位持续抬升（伴随囤货的厂商甩货），而同时部分晶圆厂新产能开始投产，进一步增加供给，导致芯片价格下杀，行业进入主动去库存阶段，晶圆厂大幅下调资本开支计划。随着行业杀价、甩库、呈现萧条景象，各环节厂商给予谨慎甚至悲观预期，但是新的复苏已经走在路上...

图表5：半导体行业“复苏-扩张-顶峰-衰退”四大阶段



来源：国金证券研究所制作

图表6：全球半导体资本开支增速 vs. 半导体销售额增速



来源：WSTS, iFind, 国金证券研究所

1.3 短周期（3-5个季度）：看库存

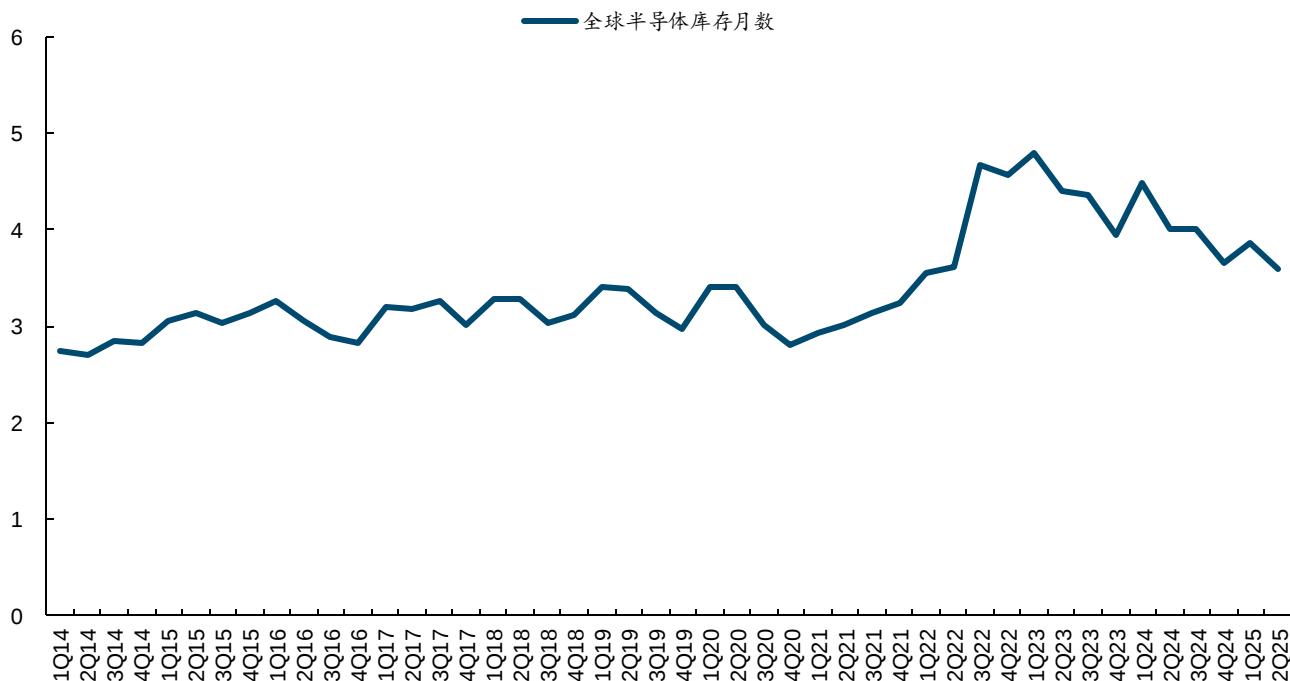
由于半导体产业链条较长，各环节厂商大多专注一部分，很难做到纵观整条产业链的供需变化，因此导致在一轮周期中全行业能保持供需平衡的时间很短，而供需失衡是常态，其最典型的一个指标就是库存水位变化。因此，我们可以通过观察库存水位作为判断半导体库存周期变化的一个前瞻性指标，库存周期主要表现为：被动去库存（复苏）—主动补库存（扩张）—被动补库存（顶峰）—主动去库存（衰退）—被动去库存（复苏）...

我们统计全球半导体平均库存月数可以看到，在 2021 年之前库存周期较为平稳，每 3-5



个季度经历一次波峰和波谷（当然这个过程中也受季节性因素影响），在 2021 年宏观环境变化及产业链供需错配导致行业出现大缺货，扰乱了正常的库存周期，整体库存周期拉长到 3 年，到 2023Q4 完结了这一轮的库存周期。从 24 年以来的数据观察，我们预计目前半导体行业已恢复到正常的库存周期节奏，当然数据还要进一步观察。

图表7：从芯片厂商库存看：全球半导体行业每 3-5 个季度经历一轮库存周期

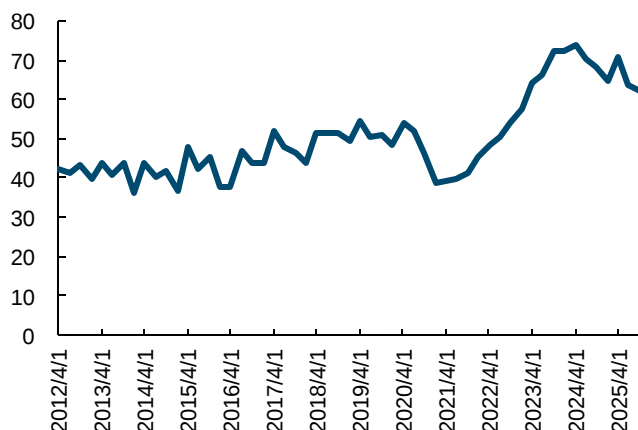


来源：Bloomberg, 国金证券研究所

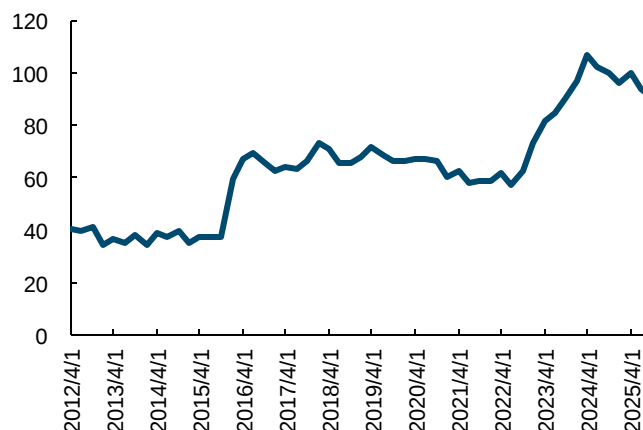
渠道端库存，我们分别观察安富利和艾睿电子等全球头部的半导体分销商的库存情况，库存周转天数变化趋势与芯片厂商库存变化趋势相似，也是每 3-5 个季度经历起伏。

图表8：艾睿电子存货周转天数变化

图表9：安富利存货周转天数变化



来源：Bloomberg, 国金证券研究所



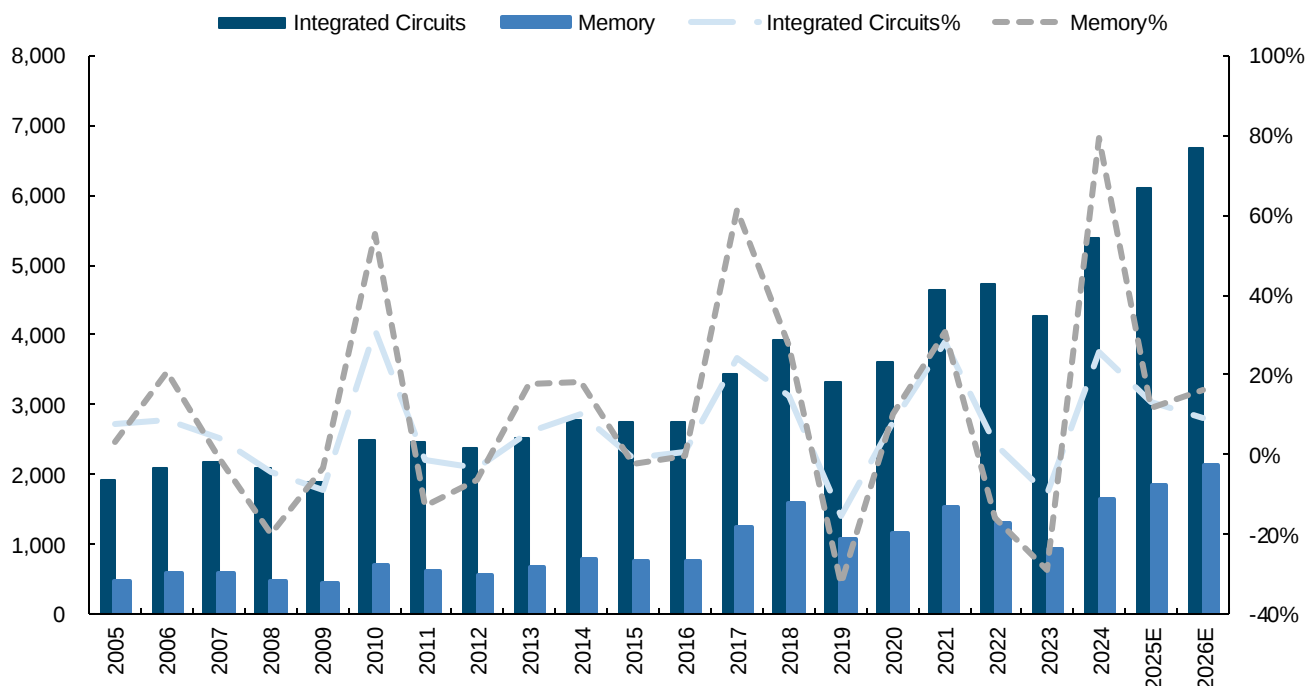
来源：Bloomberg, 国金证券研究所

1.4 存储周期复盘：大市场与强周期并存

存储器作为半导体中仅次于逻辑的第二大细分市场，其历史表现与整个半导体周期走势一致，但波动性大于整个行业。根据世界半导体贸易统计组织（WSTS）的数据，2024 年全球半导体的市场规模为 5395.1 亿美元，同比增长 26%。其中，存储器细分赛道市场规模达 1655.2 亿，同比增加 79%，占比约为 31%。从历史上看，半导体以及存储细分赛道呈现出趋同的周期性，但存储板块在一定程度上放大了行业整体波动性。2000 年以来，全球存储行业已经历多轮周期，每一轮完整周期大概会历时 3-5 年。随着 AI 驱动需求提升，当下我们走在新一轮存储大周期的起点。



图表10：存储板块通常大市场与强周期并存

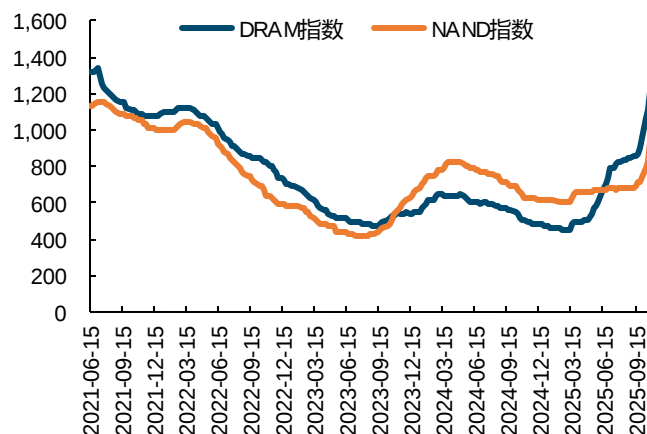
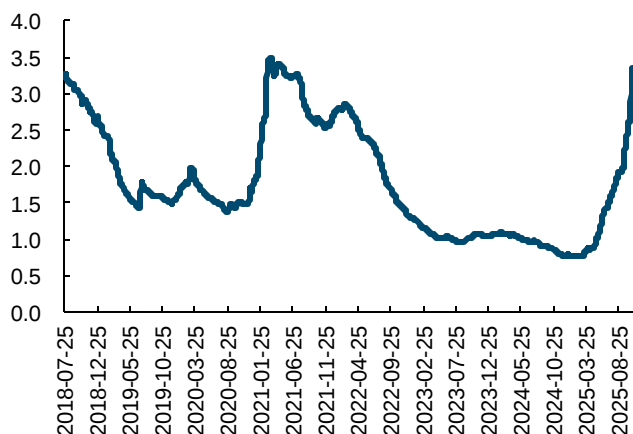


来源：WSTS，国金证券研究所

颗粒价格与存储价格也息息相关。当行业开始上行周期，下游需求走高，行业新增产能有限的情况下，市场格局逐渐从供过于求走向供不应求，原厂会有意识的去推高现货价格，以此抬升公司盈利能力。当新增产能被大幅度开出，行业重新走向供过于求，渠道和终端库存开始走高，现货价格开始下跌。原厂会通过缩减资本开支以及降低稼动率的方式减少晶圆产出，同时等待下游库存去化以及新需求的出现，周而复始形成存储周期。

图表11：存储价格具有明显周期性，以DDR3 4Gb为例
(单位：美元)

图表12：DRAM与NAND指数具有类似的周期性



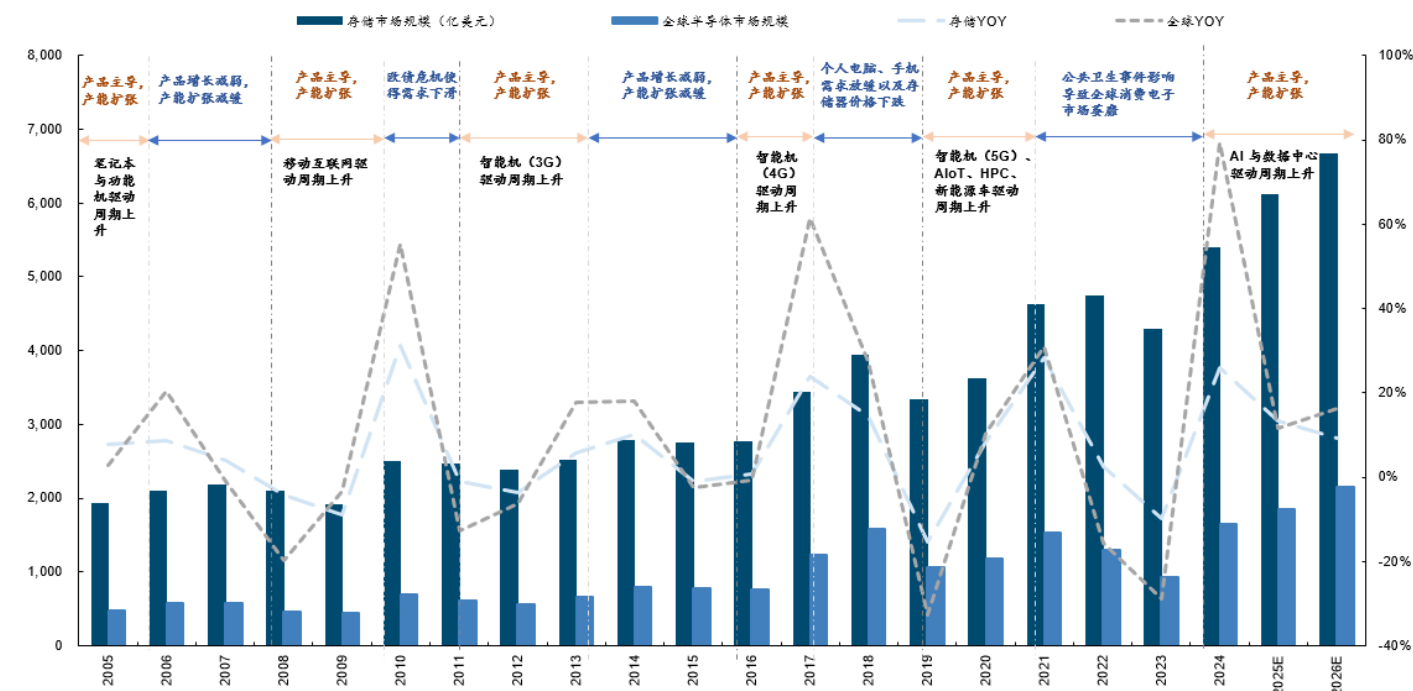
来源：iFind，国金证券研究所

来源：iFind，国金证券研究所

每轮存储周期亦有不同，需求驱动的周期通常持续性更长及弹性更大。2008年移动互联网周期和2016年4G手机周期皆充分印证了这一点。移动互联网爆发催生高清音视频等应用，用户侧对手机存储的需求从最初的16GB，逐步向64GB、128GB乃至更大容量升级。这种增长并非短期技术迭代推动，而是随应用场景深化持续释放持续提升的过程。2023年AI时代开始落地，我们同样看好模型和应用落地后对存储需求长期且大量的拉动。



图表13：存储周期复盘



来源：WSTS，国金证券研究所制图

二、AI 驱动的会是大周期吗？

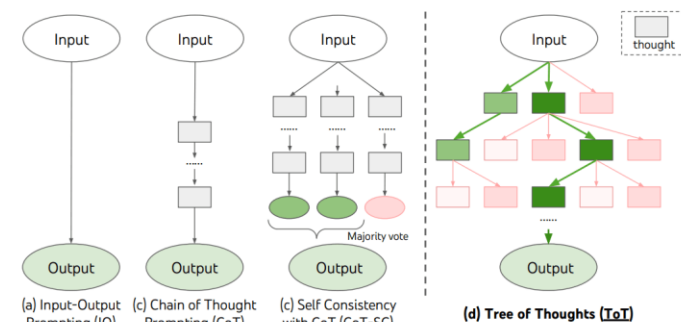
本轮周期的重要驱动因子是 AI，是需求端爆发驱动的大周期。历史上我们看到过 2008 年智能手机时代的存储大周期，但是 AI 作为工业革命后最重要的产业革命，需求的上限有多高可能难以预测，但我们希望通过研究 AI 技术与产品的变化，探讨未来在 AI 驱动下，对存储需求的趋势与确定性。

2.1 大模型类型和机制转变正产生大量数据存储需求

思维链 (CoT) 较早是由 DeepMind 在 2022 年发表的论文《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》中提出的。随后在 2024 年 9 月，OpenAI 发布的 o1 模型首次将思维链作为核心技术亮点推出，思维链提示使 LLM 能将复杂问题分解为可操作步骤，复现类似人类的推理过程。该方法对需多步骤解决的任务 (如数学应用题、符号推理、常识推理) 效果显著。通过引导模型输出中间推理步骤，思维链提示可在过程中及时识别并修正错误，最终提升答案准确性。2025 年所有主流大模型都已内化思维链机制，如 GPT-4 Turbo、Claude 3.7 Sonnet、文心一言 4.0、Gemini 2.5 Pro 等。同时，思维链出现了多种形式，如明示型思维链、树状思维链、反思型思维链等，还衍生出了一致性思维链、自动化思维链、外部验证思维链等高级技巧，进一步提升了模型的推理能力和可靠性。

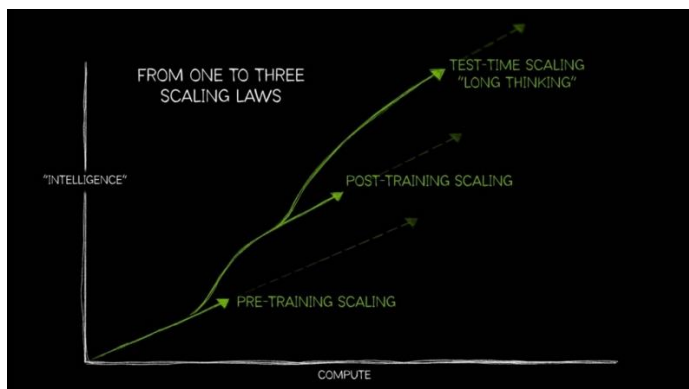


图表14：思考链复现类似人类的推理过程



来源：《Tree of Thoughts: Deliberate Problem Solving with Large Language Models》，国金证券研究所

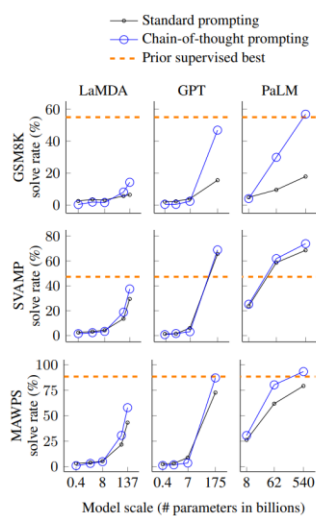
图表15：引入思考链提升了大模型的“智力”与“计算能力”



来源：Nvidia，国金证券研究所

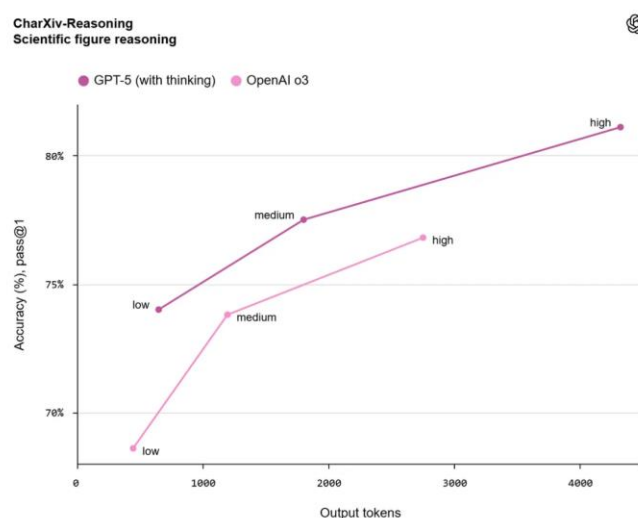
思维链提示通过强化模型的推理能力与问题解决技能，显著提升 LLM 的性能。该技术通过结构化推理流程引导模型生成更精准且可靠的输出，尤其适用于复杂任务。思维链与训练阶段的扩展定律类似，可以通过更长时间的“思考”来持续优化输出质量，呈现指数级增长特性，且似乎不存在性能上限。但是引入思维链以及思考时间的增加，也会同步提升 Tokens 的消耗量。

图表16：引入思维链显著提升大模型的性能表现



来源：《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》，国金证券研究所

图表17：增加思考时间及 Token 消耗进一步提升模型表现

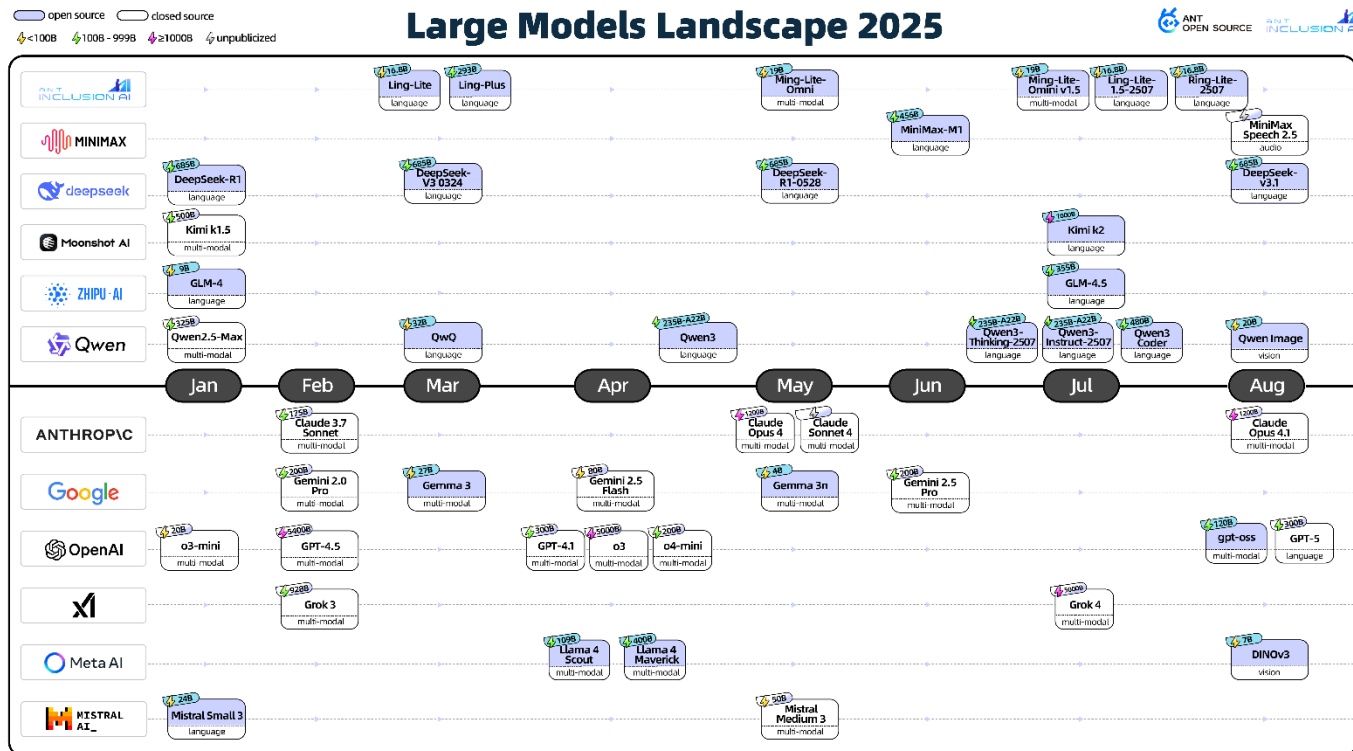


来源：OpenAI，国金证券研究所

模型的变化以及 tokens 消耗量增加，正在产生大量数据的需求。从 23 年以来，大模型本身从专注文字交互的大语言模型，逐渐升级到融合文本、图像、音视频的多模态模型，升级的背后对存储器的容量、带宽、低延迟提出更高要求。同时，随着 Sora2 这类视频生成模型的普及与迭代，存储需求将进一步迎来指数级增长。根据希捷科技的数据，从文本向音视频的切换，背后的存储单位是从 KB 往 TB 乃至 EB 的增长。随着多模态渗透率的提升乃至未来全模态模型的落地，我们看好存储需求进一步提升。

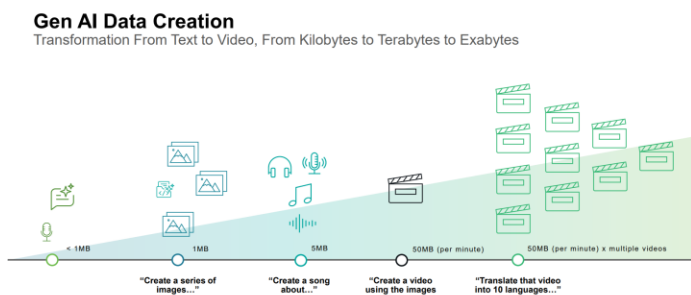


图表18：2025 年大模型线路图



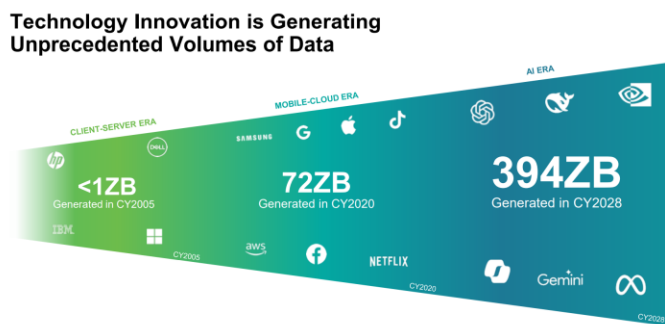
来源：蚂蚁开源，国金证券研究所

图表19：生成式 AI 对存储的需求持续提升



来源：希捷科技，国金证券研究所

图表20：不同阶段的科技创新对存储的需求持续提升



来源：希捷科技，国金证券研究所

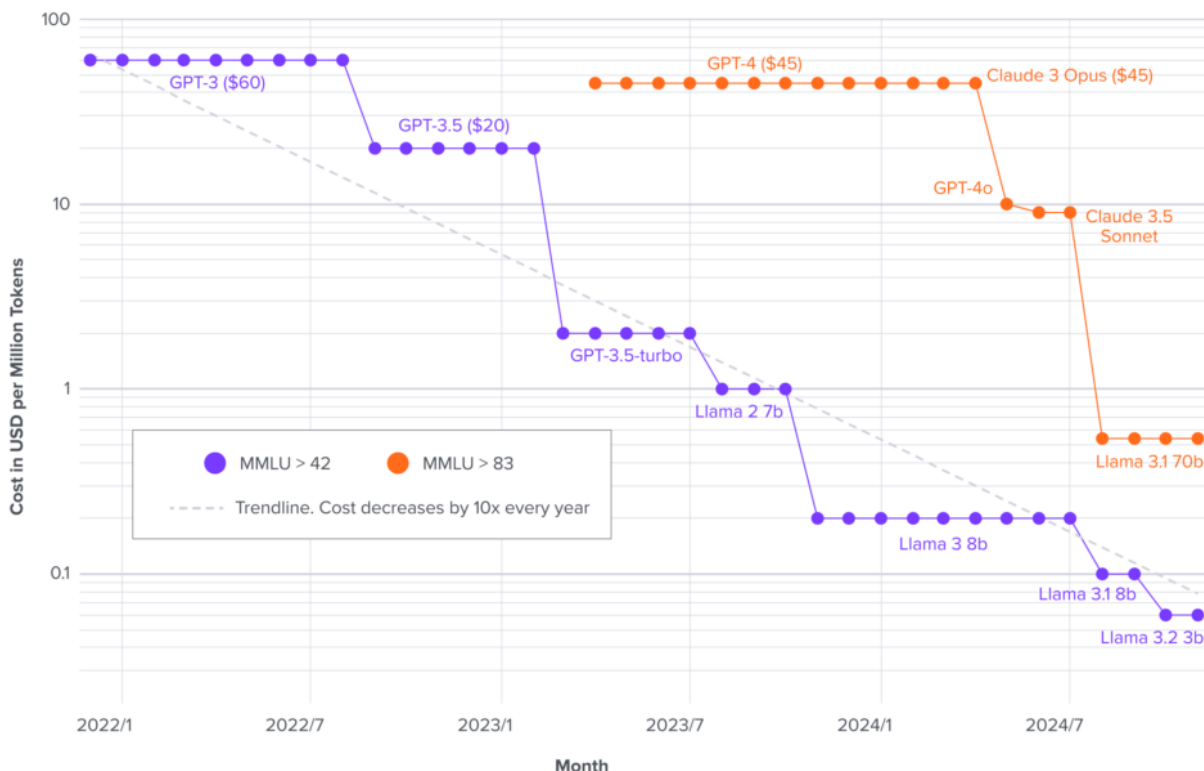
2.2 头部模型差距缩小，推理降本趋势明显

从 chatgpt3 发布以来，大语言模型的推理的成本以指数级别下降，单美元可以生成的 token 数量持续增长。同时模型的能力也持续增加。相较于最早的 chatgpt3，目前的主流模型都可以通过更低的推理成本达到更高的智能水平。在 2021 年 1 月，GPT3 是唯一可以达到 MMLU42 分的大语言模型，当时百万 token 的成本在 60 美元，截至 2024 年 11 月，由 together.ai 提供的 Llama3.2B 可以同样达到 MMLU42 分的水平，但百万 token 的成本以及降低到 0.06 美元。而可以达到 MMLU83 分的大语言模型中，Llama3.170B 截至 2024 年 11 月百万 token 的成本已经小于 1 美元。根据 A16z Infrastructure 测算，推理成本每年降低幅度约 10 倍。



图表21：大模型推理成本按照指数下降，每年降低幅度约10倍

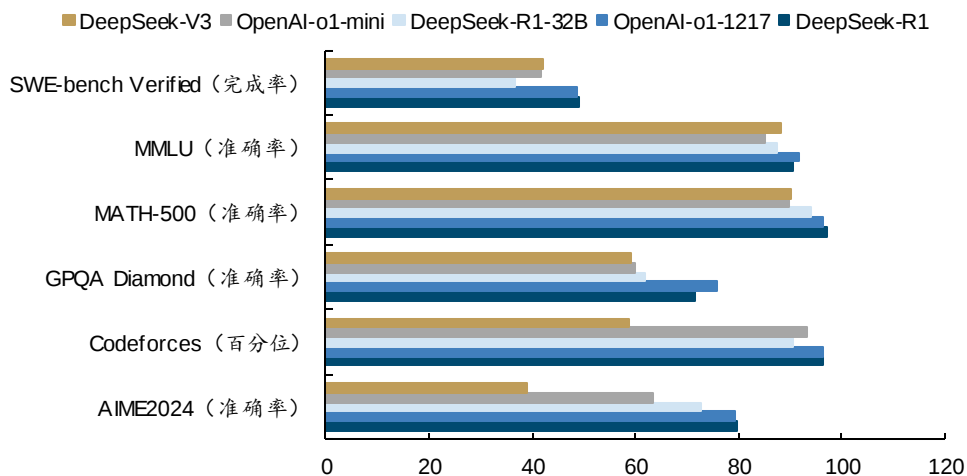
Cost of the Cheapest LLM with a Minimum MMLU Score (Log Scale)



来源：A16z Infrastructure, 国金证券研究所

另外我们注意到，目前模型的能力提升，除了模型的预训练以外，推理的算法升级也不断落地，通过包括强化学习、MOE 等方法提升模型推理能力。我们认为：模型的推理成本快速降低，有望带动应用的爆发，而应用爆发将带动更多的推理算力需求，同时推理算法的迭代也带动更多算力需求。

图表22：Deepseek 依靠推理端算法优化可以实现较好智能程度（单位：%）



来源：《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》，国金证券研究所

2.3 AI 正在重新定义数据存储，KV Cache 或将成为大模型推理优化的关键突破

随着 AI 大模型及高性能计算（HPC）的快速发展，数据规模持续攀升，带动数据存储及处理资源需求同步增长。面向不同业务场景，业内通常按照访问频次与业务重要性将数据划分为“热数据—温数据—冷数据”三个层级，以优化数据管理与资源配置。1) 热数据 (Hot Data)：指访问频率高、对核心业务实时性要求强的数据。此类数据需要在低时延、高性



能介质中存放，典型载体包括 DRAM、NVMe SSD 等；2) 温数据 (Warm Data)：指访问频次适中、业务价值较为重要的数据。相比热数据，对实时性要求略低，但需保证在合理时间内可靠访问，常使用成本相对较低、容量更大的磁盘阵列等设备存储；3) 冷数据 (Cold Data)：指访问频率低、价值时效性弱的数据，多用于长期留存及合规备份。其访问需求不高，可存放于磁带库等更低成本、超大容量的介质中。基于数据“冷热分层”的特性，企业可以构建差异化存储策略，实现对性能、成本与可靠性的最优平衡，从而有效降低整体存储成本并提升数据访问效率。

从核心到边缘到终端，大家不仅面对大数据带来的变化，同时快数据也带来新的变化。新基建需要新存储，新存储的核心自然会聚焦在创新存储架构上，因为基础架构已经成为了用户的应用升级与业务转型的基石。新基建促进产业升级与转型，新存储促进创新架构与时俱进。但是，这一切，都根于企业的应用多元化。不同的应用诞生不同的数据类型，不同的数据类型需要不同的存储来匹配。以往业内一般分为冷存储和热存储，然而用户面临不可阻挡的企业数字化转型趋势，这种粗放式存储设备和架构规划要适应各方面的需求，事实上不可能。西部数据把最中心的存储分成五层：1) 极热存储，需要高频读写，拥有极高 IOPS 和带宽需求。对应用户的典型应用环境如：面向交易领域的 OLTP(On-Line Transaction Processing 联机事务处理)、面向智能视频监控领域的 DSP(Digital Signal Processor 数字信号处理器)、自动驾驶口等。一般采用 DRAM 方案；2) 热存储，需要数据频繁读写，持续低延时，有着高 IOPS、高带宽应用需求。面向用户的数据库领域典型应用可见于 OLTP、DSP、自动驾驶等方面。一般使用高性能 SSD 方案；3) 温存储，需要读取较频繁，有较高性能 IOPS，带宽要求，性能峰值并不特别需要持续维持在高性能的状态。其典型应用如：OLAP(联机分析处理 On-line Analytical Processing)、AI 日训练，以及传统企业应用。一般使用主流 SSD 方案；4) 冷存储，需要少量写入，多次读取，其数据与温数据有相关性。主要面向备份，典型应用如：大数据分析、AI 训练，以及包括一些在线交易的用户的画像分析。一般使用 CMR HDD 方案；5) 极冷存储，需要一次写入，极少读取，适合长期持久归档存储的数据，即一次写入后一年之内可能读取的次数都很少，但又需要长期保存。其典型应用场景包括：金融的交易数据、医疗档案、广电与监控行业的一些视频资料、在线游戏直播的一些视频数据等。一般使用 SMR HDD 方案。

图表23：西部数据针对不同的存储要求，将存储设备和架构规划分成了五层



来源：CSDN，国金证券研究所

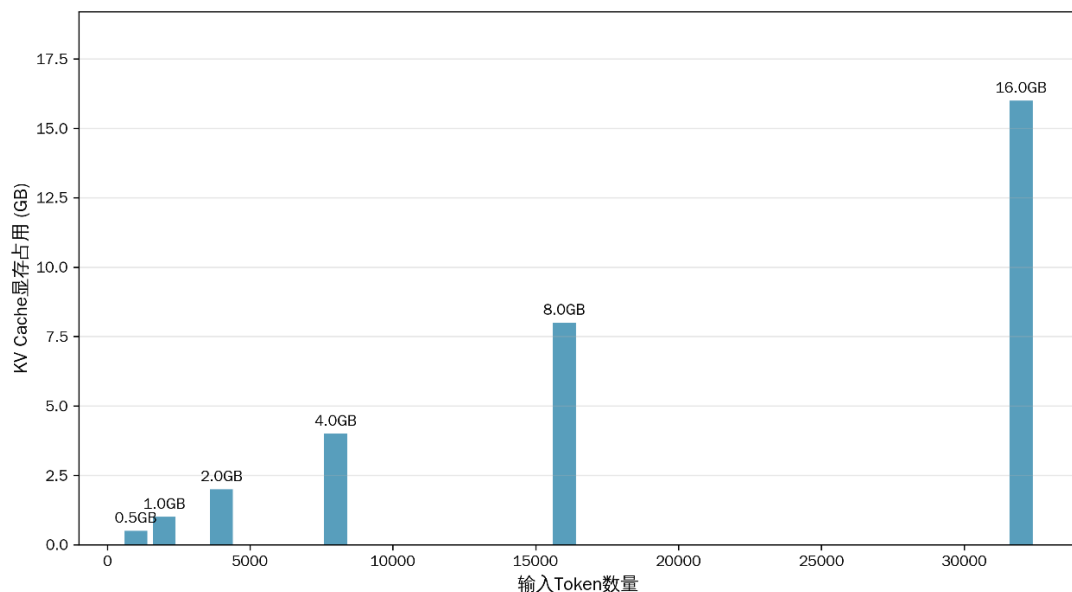
全球数据规模持续呈指数级扩张，行业已全面进入 Zettabyte (ZB) 级时代。在以人工智能 (AI)、深度学习 (DL) 和机器学习 (ML) 为代表的新一代技术体系中，数据已成为驱动模型演进与创新的核心生产要素。然而，不同应用场景对存储能力的要求差异显著，单一架构难以满足 AI 全流程的数据处理需求。具体来看，AI 相关工作负载在容量、吞吐量、延迟及 IOPS 等关键指标上呈现明显分层特征：大规模训练阶段强调超高容量与高吞吐的数据读写；模型微调与特征工程阶段兼顾中等延迟与高并发访问能力；推理与在线服务阶段则更加依赖低延迟与高 IOPS 的性能表现。数据分层管理与多级存储体系成为支撑 AI 工作负载的关键基础设施。

KV Cache，全称 Key-Value Cache (键值缓存)，是 Transformer 架构中一种关键的性能优化机制。从本质上讲，它是一种“空间换时间”的策略，通过缓存已计算的 Key 和 Value 矩阵，避免在自回归生成过程中重复计算，从而显著提升推理效率。根据微软亚洲研究院



的最新研究，在处理 100 万 token 的长文本时，单张 A100GPU 可能需要超过 30 分钟的推理时间，而 KV Cache 技术的引入，能够将这一时间缩短 2-3 倍。

图表24: KV Cache 显存占用随 Token 数量线性增长



来源：火山引擎开发者社区，BetterYeah，国金证券研究所

KV Cache 在内存中的组织方式对大模型推理性能具有关键影响。不同模型架构（如 Transformer、MoE 等）虽在具体实现上存在差异，但总体目标均指向在有限的显存/内存预算下最大化读写效率与访问吞吐。当前主流 KV Cache 实现普遍围绕“存储效率最优化”展开，主要采用以下结构化管理与压缩策略：1）预分配策略提升访问一致性，依据模型的最大序列长度预先分配连续内存区，能够减少运行时的内存碎片，并显著降低动态分配带来的开销，有助于提升推理链路的稳定性；2）动态扩容机制增强灵活性，在推理过程中，当实际序列长度超出初始阈值时，系统可自动扩展缓存容量，以适配长文本或复杂对话场景，提高实用性与鲁棒性；3）内存池管理降低频繁分配成本，通过引入统一的内存池（memory pool），可避免频繁的 malloc/free 操作，减少系统级开销，改善高并发情形下的整体吞吐效率；4）数据压缩技术有效降低存储占用，借助量化（如 FP8、INT8、甚至 INT4）等技术，对缓存精度进行压缩，在保持模型效果可接受的前提下显著降低显存消耗，从而提升多并发推理场景下的部署密度。KV Cache 的组织与优化已成为大模型推理系统的重要工程化环节。其技术路径的持续演进，有望进一步改善算力利用效率，并推动推理成本结构优化。

图表25: 不同的模型架构，存储结构略有差异

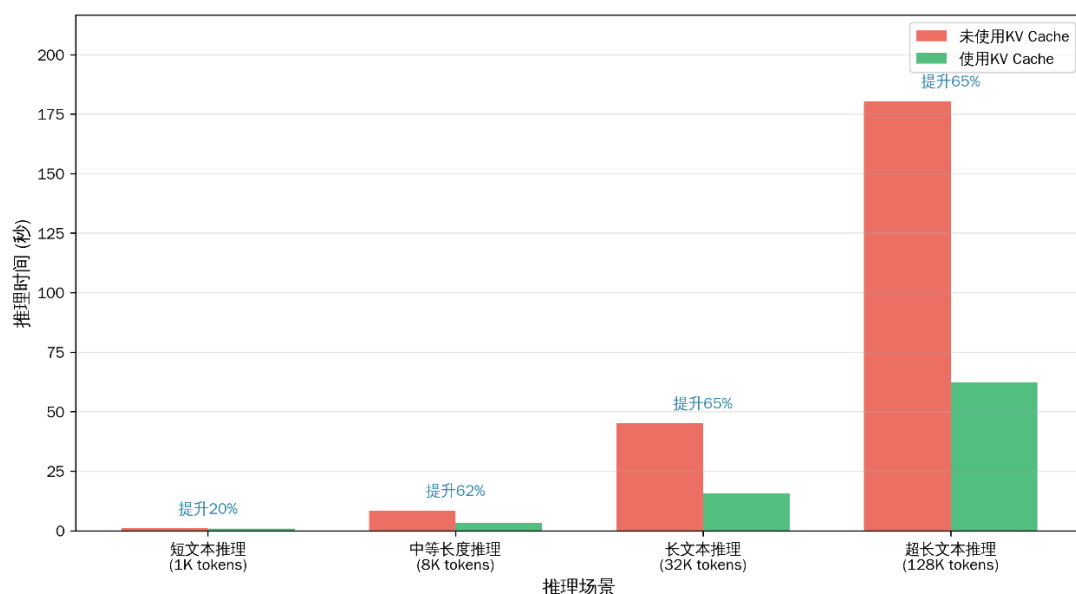
模型架构	存储维度	内存布局	优化策略
GPT 系列	[batch, heads, seq_len, head_dim]	连续存储	内存对齐优化
LLaMA 系列	[batch, seq_len, heads, head_dim]	分层缓存	动态扩容
Transformer 库	可配置	灵活布局	自适应调整

来源：前沿 AI 技术洞察，BetterYeah，国金证券研究所

KV Cache 的效果很大程度上取决于缓存命中率。在不同的应用场景下，命中率表现差异显著。针对短文本的推理，KV Cache 带来 20% 的性能提升，中等长度推理性能提升达到 62%，长文本推理的性能提升高达 65%，而超长文本的推理性能提升则达到了 66%。KV Cache 针对性能提升的规律说明，KVCache 在处理越长的文本时，优化效果越明显。这正好解决了长上下文应用中最迫切的性能需求。



图表26: KV Cache 技术在不同场景下的性能提升效果



来源: InfoQ, BetterYeah, 国金证券研究所

随着大模型技术快速迭代、推理场景持续扩展,支撑其高效运行的 KV Cache 亦呈现加速演进的趋势。未来 KV Cache 技术将向更智能化、自动化方向演进,以适配不同模型规模与业务负载需求,主要体现为:1) 自适应压缩能力增强,系统将基于内容重要性、访问频度与资源占用情况,动态调整压缩策略,以兼顾推理性能与算力成本,实现更精细化的缓存管理;2) 预测性缓存能力提升,依托用户行为模式与上下文语义的学习,系统可实现提前预取(prefetch),在推理链路中进一步降低延迟,提高端到端吞吐效率;3) 跨模态缓存,随着多模态模型成为主流, KV Cache 有望在文本、图像、音频等不同模态间实现统一的缓存编排机制,提高资源利用率并简化模型部署。

在硬件侧, KV Cache 的优化方向亦日趋明确,围绕 KV Cache 的读写、压缩、重排等操作,预计将出现更多专用硬件加速模块,以减少通用 GPU 负载、提升整体能效。针对内存层级结构优化,未来架构将进一步优化 GPU HBM、CPU 内存及 SSD 之间的层级协同,实现更高效的冷/温/热数据的调度,缓解大模型推理的内存瓶颈。

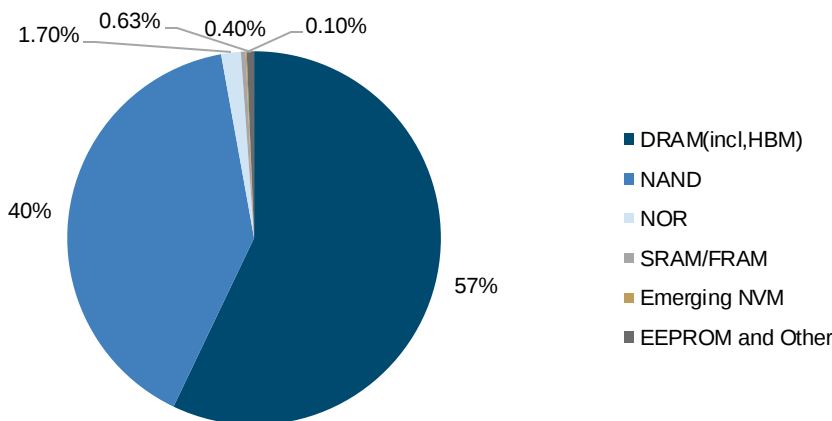
目前随着大模型应用的普及和上下文窗口的不断扩展, KV Cache 技术的重要性将进一步凸显。无论是 ChatGPT 的流畅对话体验,还是 Claude 的长文档处理能力,都需要 KV Cache 的技术支撑。未来 KV Cache 将在智能化、硬件加速、边缘部署等方向持续演进,为更广泛的 AI 应用提供强有力的技术支撑。智能化与硬件化趋势明确, KV Cache 有望在未来数年持续提升大模型推理效率并改善算力成本结构。

三、AI 驱动存储需求快速攀升,存储原厂资本开支还未进入扩张周期

NAND Flash 和 DRAM 为存储器行业的主流产品。根据 YOLE 的数据,从 2024 年全球存储市场按技术类型的收入结构来看, DRAM (含 HBM) 与 NAND 占比分别为 57% 与 40%,合计达到 97%,产业格局高度集中。各类存储技术在成熟度、成本结构与应用定位上差异显著,行业分层格局较为清晰。 DRAM 受益于 AI 推理、HPC 及服务器需求增长,目前仍是高性能计算场景的核心配置; NAND 则广泛应用于移动终端、企业级 SSD 与云数据中心,规模效应明显,成本下降路径清晰。 NOR Flash 应用场景稳定但市场占比仅为 1.7%,增长空间有限。 EEPROM、(NV)SRAM、FRAM 聚焦低功耗与高可靠性的专业化市场,这类产品生命周期长、技术迭代慢,呈现“小而稳”的市场格局。新兴非易失性存储(NVM)技术优势显著但量产受限。 MRAM、RRAM 等新型 NVM 在读写速度、功耗及耐久性方面具备潜在优势,但受制于良率、制造成本等瓶颈,尚未跨越从研发到规模化量产的关键门槛。



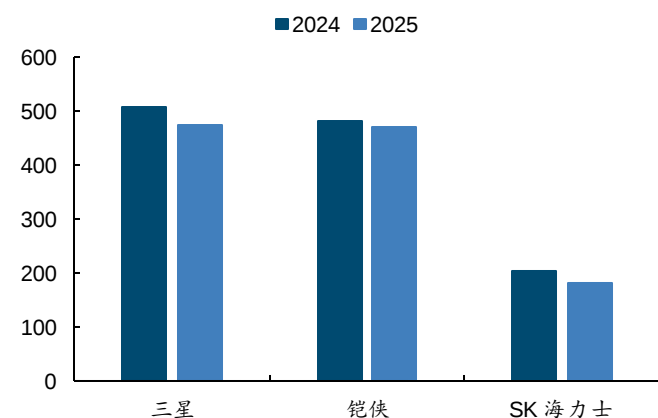
图表27：2024 年全球 DRAM 和 NAND 合计占据 97% 的份额



来源：Yole，国金证券研究所

2025 年，主要 NAND Flash 厂商已开始通过产能调整应对市场价格波动。根据《朝鲜日报》，三星电子、SK 海力士及铠侠在 2024 年价格长期维持在成本水平后，已积极采取措施推动价格上涨。根据 Omdia 的数据，三星电子 2024 年 NAND 晶圆产量为 507 万片，2025 年的产能目标下调至约 472 万片，下降约 7%；铠侠产量由去年的 480 万片调整至 469 万片。Omdia 预计，三星电子及铠侠的减产趋势将持续至 2025 年。SK 海力士及美光亦采取保守产能策略，以期从价格回升中获益。其中，美光的新加坡 Fab7 工厂产量维持在约 30 万片的相对低位，供应策略同样偏保守。主要 NAND Flash 厂商通过控制产量和调整供应节奏，有望在 NAND Flash 市场价格回升周期中实现利润稳定，同时为未来产能布局与价格谈判提供策略空间。

图表28：主要 NAND Flash 厂商 2025 年主动控产稳价



来源：Omdia，国金证券研究所

图表29：3D NAND 堆叠层数突破与 QLC/PLC 演进

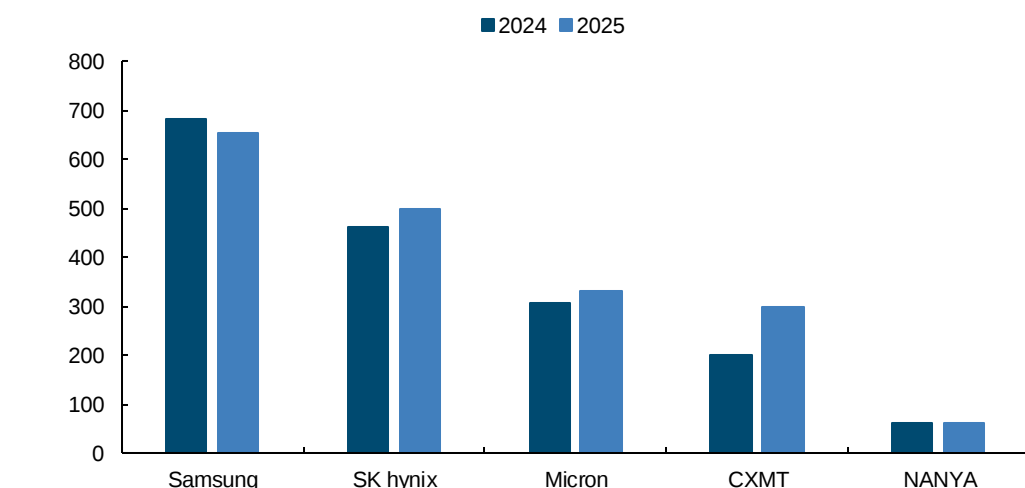
厂商	2025 年技术节点	2026 年规划
三星	V10 NAND：420-430 层，下半年量产	430 层产品，接口速度 5.6GT/s
SK 海力士	321 层 QLC 量产，采用三重堆叠	400 层，单位晶圆 TB 容量增加 30%
美光	340 层 NAND 进入样品阶段	400 层技术开发
铠侠	218 层 BiCS 3D NAND	300 层以上技术
YM	128 层以上高端产品占比超 60%	200 层以上技术

来源：芯存社，国金证券研究所

2025 年 HBM 市场份额竞争将进一步加剧，根据 TrendForce 的测算，全球 DRAM 产能由 2024 年底的约 180 万片/月增长至 2025 年底的约 192 万片/月。其中 TSV 技术在全球 DRAM 产能中的渗透率将由 2024 年底的约 15% 提升至 2025 年底的约 19%，显示高带宽存储需求持续增长。



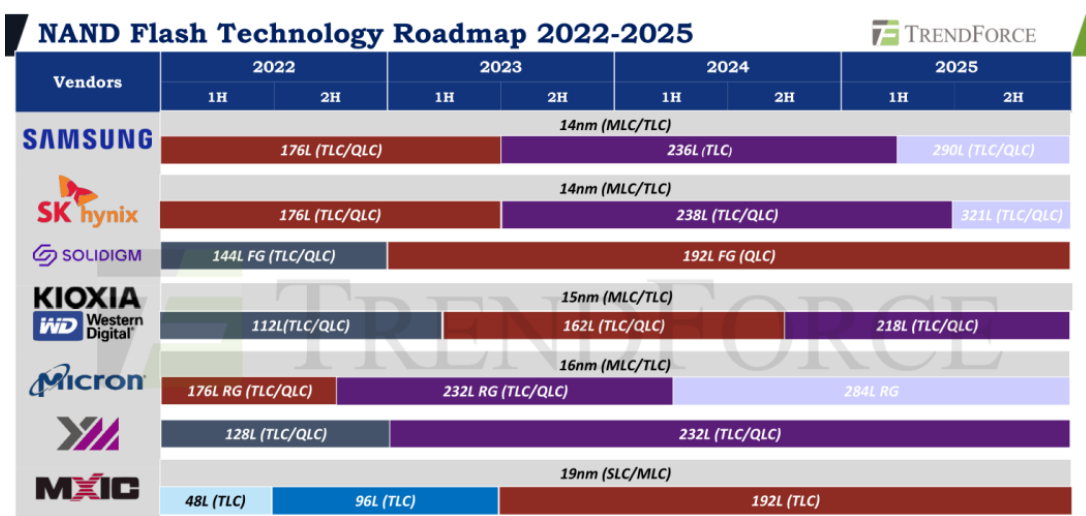
图表30：2025 年全球 DRAM 产能增幅有限（单位：千片/月）



来源：TrendForce，国金证券研究所

自 2022 年以来，全球 NAND 产业的技术演进持续围绕“更高层数 (L)”与“更高存储密度”两大方向加速推进。三星、SK 海力士、美光等国际头部厂商正不断向更高层数的 TLC、QLC 架构演进，通过堆叠层数提升、结构优化与工艺迭代，实现单位面积存储密度的持续提升与成本曲线的进一步下探。铠侠/西部数据及 Solidigm 同步推进先进节点，行业预计其主流产品将在 2024 年进入 218 层及以上的量产阶段。相比之下，国内长江存储的层数进展虽仍低于国际领先水平，但已处于行业中游梯队，技术代际迭代速度明显加快，未来追赶空间明确。

图表31：主要 NAND 供应商的 Roadmap 显示出向更高层数和更高存储密度的趋势



来源：TrendForce，国金证券研究所

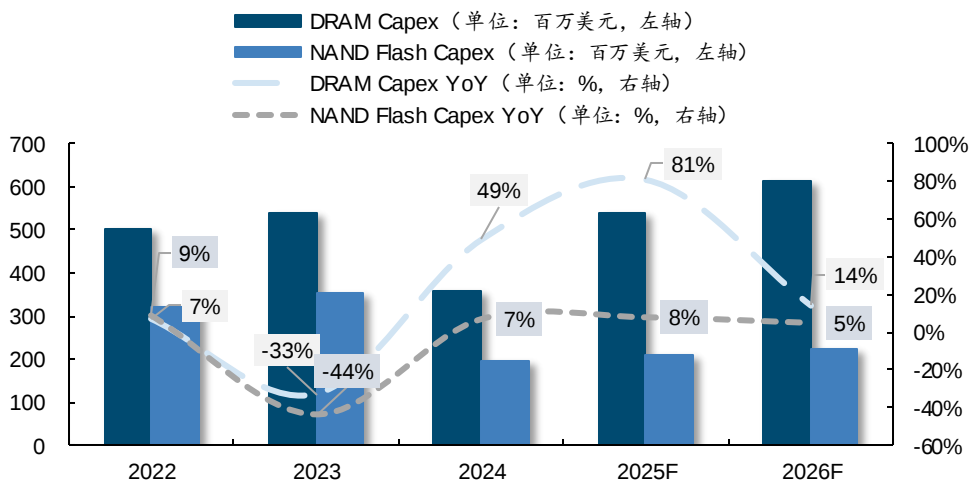
本轮 NAND Flash 行业需求强劲上行，核心驱动力来自 AI 应用对高容量存储需求的快速攀升，叠加 HDD 供应紧缩，促使云端服务供应商 (CSP) 加速向 NAND 体系转单，形成明显的结构性缺口，而非短期行情波动。与此同时，过去数年行业多次经历景气反转，使得主要厂商在资本支出与扩产节奏上更趋审慎。进入 2026 年后，全球 NAND 厂商资本开支重点将转向先进制程迭代与 Hybrid Bonding 技术导入，而非新产能扩张，使得行业位元供应增幅受到显著限制。TrendForce 预测 NAND 市场供给紧缩将延续至 26 年全年。

随着存储器平均销售价格(ASP)持续提升，供应商获利也有所增加，DRAM 与 NAND Flash 后续资本支出将会持续上涨，但对于 2026 年的位元产出成长的助力有限。DRAM 和 NAND Flash 产业的投资重心正逐渐转变，从单纯地扩充产能转向制程技术升级、高层数堆栈、混合键合以及 HBM 等高附加价值产品。根据 TrendForce 的数据，DRAM 产业的资本支出在 2025 年预计将达到 537 亿美元，预计在 2026 年进一步成长至 613 亿美元，同比增长约 14%。NAND Flash 部分，资本支出在 2025 年预计为 211 亿美元，2026 年



预计小幅增长至 222 亿美元，同比增长约 5%。

图表32：26 年 DRAM 与 NAND Flash 产业的资本支出和扩产仍未进入扩张周期



来源：TrendForce，国金证券研究所

从资本开支来看，美光仍是 DRAM 领域扩产最积极的厂商，但重心仍倾向于 HBM。根据 TrendForce 的数据，美光 2026 年资本支出预计达到 135 亿美元 (YoY +23%)，重点投向 1y 制程渗透及 TSV 设备建置；SK hynix 的扩产力度同样显著，2026 年资本支出预计为 205 亿美元 (YoY +17%)，主要用于 M15X 的 HBM4 产能扩张；三星预计投入 200 亿美元 (YoY +11%)，用于 HBM 的 1C 制程导入及 P4L 产能小幅提升。当前行业无尘室空间已接近瓶颈，各大 DRAM 厂商中，仅三星与 SK hynix 仍具备有限的扩线空间；美光则需待 ID1 新厂完工后才有新增产出，最早 2027 年才能贡献产能。因此，即便后续资本支出上修，对 2026 年的产能拉动极为有限。

在 NAND Flash 领域，Kioxia/SanDisk 因无 DRAM 业务，被视为最积极扩产的厂商。两家公司 2026 年资本支出预计 45 亿美元 (YoY +41%)，将加速 BiCS8 的量产，并投入 BiCS9 的研发。美光的 NAND 规划则以小幅扩张产能、推进 G9 制程及强化企业级 SSD 布局为主，2026 年 NAND 资本支出预计同比增长 63%。三星与 SK hynix/Solidigm 则在 NAND 端采取更保守的扩产策略，并将投资重心进一步向 HBM 与 DRAM 转移。

图表33：受 AI 需求影响，各大存储厂商新的产能将优先考虑 HBM 的供应

公司	新 FAB	地点	概况
SAMSUNG	P4L	韩国	前期工作已启动，新的 fab 厂建设预计在 2025 年完成
SK hynix	M15X	韩国	M15X 位于 M15 厂旁，其部分 TSV 设备亦设置于此，将更有利于 HBM 的生产。该厂预计于 2025 年底实现量产
	Yongin	韩国	目前正在进行土地整备，预计明年开始建设晶圆厂，并于 2027 年 6 月开始量产
Micron	A3	中国台湾	工厂的洁净室已准备就绪，将用于维持 26 年的晶圆产能，最大规模约为 4.5 万片。
	A5	中国台湾	位于台中的一块土地已被收购
	Boise Fab	美国	新晶圆厂位于美光研发中心旁边，已于去年动工，预计将于 2026 年开始投片生产
	N.Y.Fab	美国	厂址仍处于准备阶段，首座晶圆厂预计在 2025 年内开工建设，并于 2028 年开始量产
	Hiroshima New Fab	日本	该厂计划于 2026 年开工建设，预计在 2027 年底完工，主要用于 HBM 相关产品的生产
	Singapore Fab	新加坡	该扩产计划仍处于讨论阶段，最早的投产时间将在 2028 年之后
NANYA	5A	中国台湾	该厂于去年动工，新晶圆厂的建设预计将在 2026 年完成，整厂的规划产能为 45K
PSMC	Tongluo	中国台湾	未来计划将逻辑产品相关设备集中至该晶圆厂，从而为其他既有晶圆厂释放出更多用于 DRAM 生产的空间

来源：TrendForce，国金证券研究所



四、投资建议

我们建议持续重点关注存储器库存、价格数据以及 AI 算力提升对存储芯片的需求拉动，建议关注如下三个方向：

- 1) 定制化存储龙头：兆易创新
- 2) 企业级模组厂商：香农芯创、德明利、江波龙、佰维存储
- 3) 存储配套及其他芯片：普冉股份、澜起科技、聚辰股份、东芯股份、北京君正、朗科科技、恒烁股份等

五、风险提示

需求低于预期：目前 AI 发展拉动存储器需求持续提升，但不排除 AI Capex 下滑的风险，从而导致需求下滑。

存储器价格下跌：到目前为止，存储的价格还在继续上涨，虽然我们预计后续上涨趋势有望持续，但也不排除持续下跌的情况。

存储器价格上涨挤压其他领域需求：存储器价格上涨已经显现出对 AI 外其他领域，例如消费电子等市场的需求的挤压。



行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。



特别声明：

国金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。

形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。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，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券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多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海

电话：021-80234211

邮箱：researchsh@gjzq.com.cn

邮编：201204

地址：上海浦东新区芳甸路 1088 号

紫竹国际大厦 5 楼

北京

电话：010-85950438

邮箱：researchbj@gjzq.com.cn

邮编：100005

地址：北京市东城区建国门内大街 26 号

新闻大厦 8 层南侧

深圳

电话：0755-86695353

邮箱：researchsz@gjzq.com.cn

邮编：518000

地址：深圳市福田区金田路 2028 号皇岗商务中心

18 楼 1806



【小程序】
国金证券研究服务



【公众号】
国金证券研究