仅供内部参考，请勿外传！

计算机

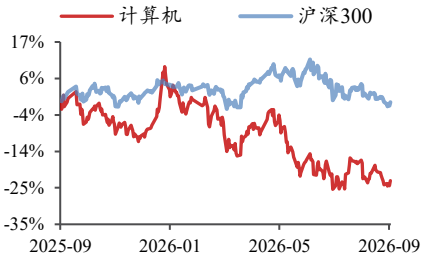
2026 年 09 月 20 日

AI 需求驱动存储景气上行，Agent 应用加速落地

——行业周报

投资评级：看好（维持）

行业走势图



数据来源：聚源

相关研究报告

《大模型迭代拓展能力边界，应用落地兑现产业价值—行业深度报告》-2026.9.18

《Rubin 升级驱动硬件价值扩散，AI 应用持续深化—行业周报》-2026.9.13

《OpenAI 发布 GPT-6 Astra，大模型迈向执行主体—行业点评报告》-2026.9.7

张越（分析师）
zhangyue1@kysec.cn
证书编号：S0790524090003

叶进新（联系人）
yejinxin@kysec.cn
证书编号：S0790126060053

● **AI 算力重塑存储需求，技术升级与供需共振拓展成长空间**

数据规模持续增长与 AI 算力需求释放，推动存储由配套支撑环节升级为影响算力效率的关键基础设施。截至 2026 年 6 月底，全国先进存力规模超过 640EB，政策提出到 2030 年提升至 1700EB，存储产业成长空间持续打开。技术层面，SRAM、DRAM 与 NAND 围绕速度、容量及成本形成分层协同，DDR5 提升主存性能，HBM 通过高位宽与垂直堆叠满足 AI 高带宽需求，3D NAND 支撑海量数据存储。供需层面，AI 服务器与高容量内存需求扩张，HBM 迭代带动单芯片容量及晶圆投入提升，产能向高附加值产品集中；叠加新增产能释放周期较长、云厂商通过长约锁定供应，存储价格与供应商议价能力获得支撑。AI 需求正成为存储产业重要增长引擎，技术升级与供给结构优化有望推动产业价值提升。

● **Agent 向专业场景持续渗透，国产算力与软硬件生态协同升级**

本周 AI 产业围绕复杂任务执行与专业场景交付持续演进。应用端，Kimi 推出金融行业解决方案，整合权威数据源与专业技能，提升建模、研究及报告交付效率；谷歌发布 Gemini 3.8 Live，支持实时对话与后台工具调用并行；Anthropic 整合 Claude 对话与办公能力，拓展文档、演示及持续任务执行场景。模型端，小米推进 MiMo-V2.6 强化学习训练，智谱通过 Infra Agent 参与推理系统优化，在国产算力集群上实现部署与吞吐效率提升。算力端，华为发布昇腾 960 超节点与灵衢互联方案，推动大规模集群及 NPO 光互连发展；联发科天玑 9600 Pro 强化端侧大模型运行能力，台积电持续推进先进制程扩产。应用深化与基础设施升级相互促进，有望加快 AI 能力向产业生产力转化。

● **计算机板块稳步上涨，AI 算力与应用共同呈现活跃行情**

9 月 14 日至 9 月 18 日，计算机指数上涨 1.58%，分别跑赢沪深 300、上证指数及创业板指 1.64、0.97 和 0.06 个百分点，在 31 个一级行业中排名第 11 位。AI 细分方向同步上涨，AI 算力指数累计上涨约 4.22%，AI 应用指数累计上涨约 2.40%，算力方向相对领先约 1.82 个百分点。个股层面，富瀚微、科泰电源、沐曦股份-U 分别上涨 19.12%、17.53%和 13.46%，均录得两位数涨幅；AI 应用方向涨幅前十个股周涨幅均超过 10%，中岩大地上涨 28.53%。结合存储需求扩张、国产超节点升级与 Agent 专业场景落地，产业催化持续丰富，可关注算力基础设施与 AI 应用协同发展带来的细分领域机会。

● **受益标的**

国产算力：寒武纪、海光信息、禾盛新材、沐曦股份、浪潮信息、工业富联等；**海外算力：**东山精密、胜宏科技、联特科技、德邦科技等；**算力租赁：**宏景科技、利通电子、协创数据等；**大模型：**智谱、MiniMax、阿里巴巴、腾讯控股；**液冷：**盾安环境、同飞股份、恒铭达等。

● **风险提示：**大模型技术迭代不及预期风险、AI 技术应用不及预期风险、相关政策不及预期风险、AI Infra 建设不及预期风险等。

仅供内部参考，请勿外传！

目　录

1、 算力需求重构存储定位，供给刚性改写周期形态 3
1.1、 存储从配套走向瓶颈，存储层级沿带宽重构 3
1.2、 存储周期属性弱化，供给刚性成为定价主轴 10
2、 计算机产业动态更新 11
2.1、 国内 CSP 厂商与大模型厂商最新动向 11
2.2、 海外 CSP 厂商与大模型厂商最新动向 13
2.3、 国产算力产业最新进展 14
2.4、 海外算力产业最新进展 15
3、 上周市场表现回顾（2026/9/14-2026/9/18） 15
4、 风险提示 17

图表目录

图 1： 半导体存储器根据存储原理分为随机存储器（RAM）和只读存储器（ROM） 3
图 2： 六晶体管（6T）SRAM 存储单元为应用最广泛的 SRAM 位单元架构 4
图 3： 为兼顾速度、容量与成本，CPU 缓存采用分级设计 4
图 4： 除 CPU 外，GPU 等 AI 芯片也采用 SRAM 阵列来支撑并行计算 5
图 5： DRAM 相同面积可容纳的比特数远多于 SRAM，因电荷泄漏而须周期性刷新 5
图 6： DDR 需要在容量、成本与兼容性间取得平衡，目前逐代演进至第五代 7
图 7： HBM 以硅通孔将多层 DRAM 裸片垂直堆叠，经硅中介层与 GPU 并排封装 8
图 8： 在 BiCS FLASH（铠侠自研 3D NAND FLASH）存储单元中，电子在穿过柱体中心的电极（图中灰色部分所示）和电荷存储膜（粉色部分）之间进行交换 9
图 9： 2D NAND 存容量扩展受限，3D NAND 依靠垂直堆叠扩容，已成为主流技术 9
图 10： 根据存储方式的不同，NAND Flash 可分为 SLC、MLC、TLC 和 QLC，分别适配不同的应用场景 10
图 11： 2026 年全球半导体市场增长 90%至 1.51 万亿美元，存储为主要驱动因素 10
图 12： 2027 年 HBM 迭代升级、需求增长，对传统 DRAM 产能挤出效应持续放大 11
图 13： 在本周 TMT 板块行情共振向上的背景下，计算机板块相对表现偏温和 16
图 14： 本周 AI 细分方向呈现一定分化，AI 算力指数整体表现明显优于 AI 应用指数 16
图 15： AI 算力指数整体小幅上涨，但板块内部仍存在较为明显的结构性行情 16
图 16： AI 应用方向指数整体上涨，龙头个股赚钱效应更为突出 17

表 1： 主流的同步 DRAM (SDRAM) 分为四类，针对不同应用场景优化，各有优劣 6
表 2： DDR 代际迭代呈现清晰演化特征：工作电压持续下探，单颗粒存储容量扩容，接口数据传输带宽不断提升 7
表 3： HBM 的迭代沿两条主线推进：接口位宽逐代翻倍，堆叠层数与单层容量同步抬升 8

仅供内部参考，请勿外传！

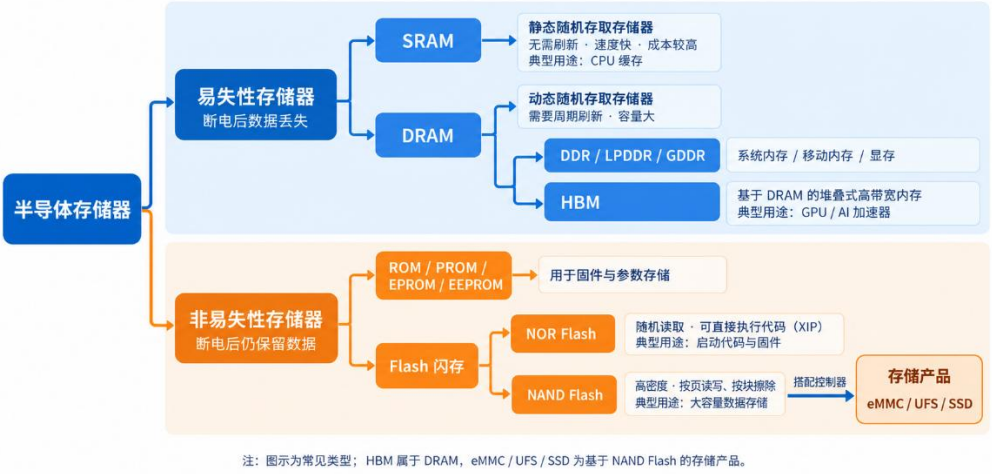
1、 算力需求重构存储定位，供给刚性改写周期形态

1.1、 存储从配套走向瓶颈，存储层级沿带宽重构

数据量级的持续跃升与算力需求的快速释放，推动存储从配套支撑环节升级为数字基建的核心支柱。在 5G、AI、AIOT 等新兴战略产业的快速发展带动下，信息数据呈现爆发式增长，数据规模从原来的 GB、TB、PB 跃升至 EB、ZB 级，存储随之从数据承载环节上升为制约算力释放的关键基础设施。中国信息通信研究院的数据显示，截至 2026 年 6 月底，全国数据存力总规模达 2021EB，同比增长 11%，先进存力占比提升至 32%，规模超过 640EB；工业和信息化部提出，到 2030 年先进存力规模提升至 1700EB。半导体存储是当前存力的主要载体，历经半个多世纪演进已形成以 DRAM 与 NAND 为主体的产品格局。

半导体存储也称为存储芯片，根据数据存储原理的不同，半导体存储器可以分为随机存储器（RAM）和只读存储器（ROM）。RAM 可直接与 CPU 交换数据，读写速度快，属易失性存储器，通常用作运行中程序的临时数据介质；按单元结构，RAM 进一步分为静态随机存储器（SRAM）与动态随机存储器（DRAM），前者以触发器保存数据、无需刷新，后者靠电容电荷保存、需定期刷新。ROM 只能读取事先写入的数据，属非易失性存储器，用于存放固定程序与数据；按写入方式可分为掩膜 ROM、PROM、EPROM、EEPROM 与闪存（Flash），其中 Flash 为当前主流，并进一步划分为 NOR Flash 与 NAND Flash。

图1：半导体存储器根据存储原理分为随机存储器（RAM）和只读存储器（ROM）



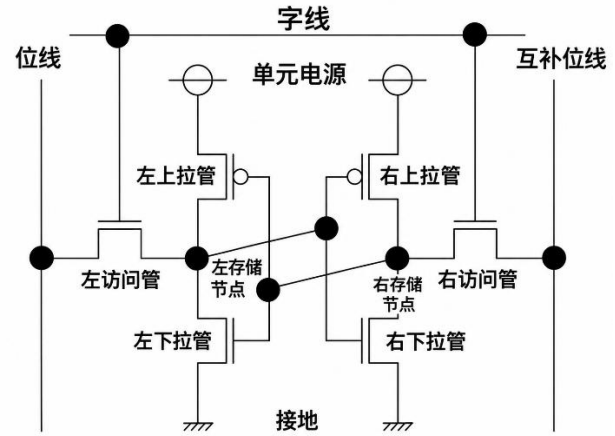
资料来源：芯存社、开源证券研究所

（一）SRAM

SRAM 即静态随机存储器，又称片上缓存，其数据由电路稳态维持，只要供电便无需像 DRAM 那样周期性刷新。存储单元采用六晶体管结构：其中四只构成两个交叉耦合的反相器，即两只 PMOS 上拉管与两只 NMOS 下拉管首尾相接形成锁存器，另两只 NMOS 为访问管，受字线控制。交叉耦合使两个反相器相互强化、形成正反馈，两个存储节点必然维持一高一低，分别代表 1 与 0。之所以必须用有源电路而非电容保存状态，是因为电荷会泄漏，状态必须由电路持续锁定。代价是每比特占用六个晶体管，而 DRAM 每比特仅需一只晶体管加一个电容，单元面积显著更大、密度更低。

仅供内部参考，请勿外传！

图2：六晶体管（6T）SRAM 存储单元为应用最广泛的 SRAM 位单元架构

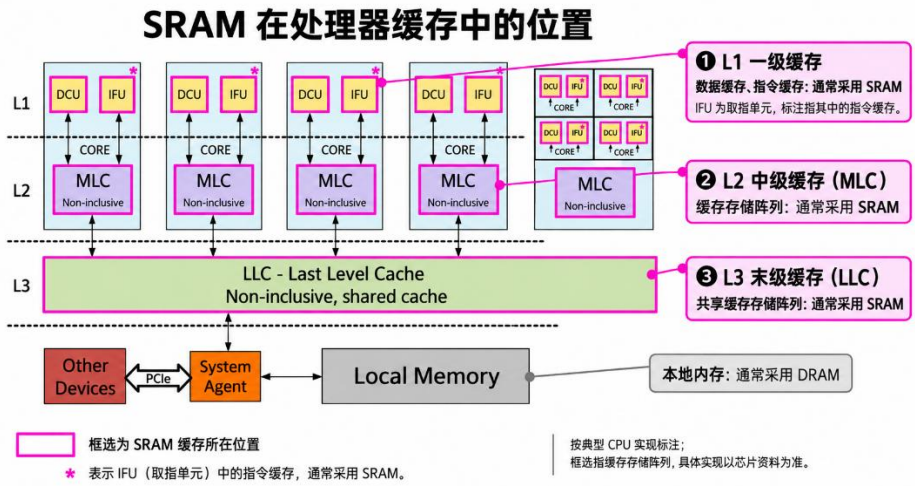


资料来源：Changhwan Shin 等《Advanced MOSFET Designs and Implications for SRAM Scaling》、开源证券研究所

SRAM 的优势与代价同源。其无需刷新即可保持数据，读写时序可预测，访问延迟落在纳秒量级，先进节点下可低于 1 纳秒，快于 DRAM，且省去刷新带来的周期性功耗；与 CMOS 逻辑工艺兼容，可同片集成，因而能紧贴运算单元构成片上缓存与寄存器堆。代价在于面积与微缩难度：每比特六个晶体管使单元面积远大于 DRAM 的一晶体管一电容结构，单位成本高、集成度低，难以达到主存容量；单元须兼顾读稳定与可写入，单元比例须大于 1.0 才能避免读扰动，工艺微缩后电压下降与阈值波动使静态噪声容限退化，待机漏电逐代上升，低功耗优势仅在动态层面。

SRAM 最主要的应用是作为 CPU 的高速缓存，现代 CPU 的性能在很大程度上取决于缓存系统的设计。为兼顾速度、容量与成本，CPU 缓存采用分级设计：L1 最靠近 CPU 核心，速度最快、容量最小，通常为几十 KB，分为指令缓存与数据缓存；L2 速度稍慢、容量更大，通常为几百 KB 至数 MB，作为 L1 与下一级间的缓冲；L3 由所有 CPU 核心共享，可达几十 MB，是访问主存前的最后一道防线。该分级使多数访问能在最快层级命中，平均访问延迟因而远低于直接访问主存。除 CPU 缓存外，在 GPU 和 TPU 等专用 AI 芯片中，巨大的 SRAM 阵列被用来高效缓存输入数据、权重和中间计算结果，以满足海量数据并行计算的需求。

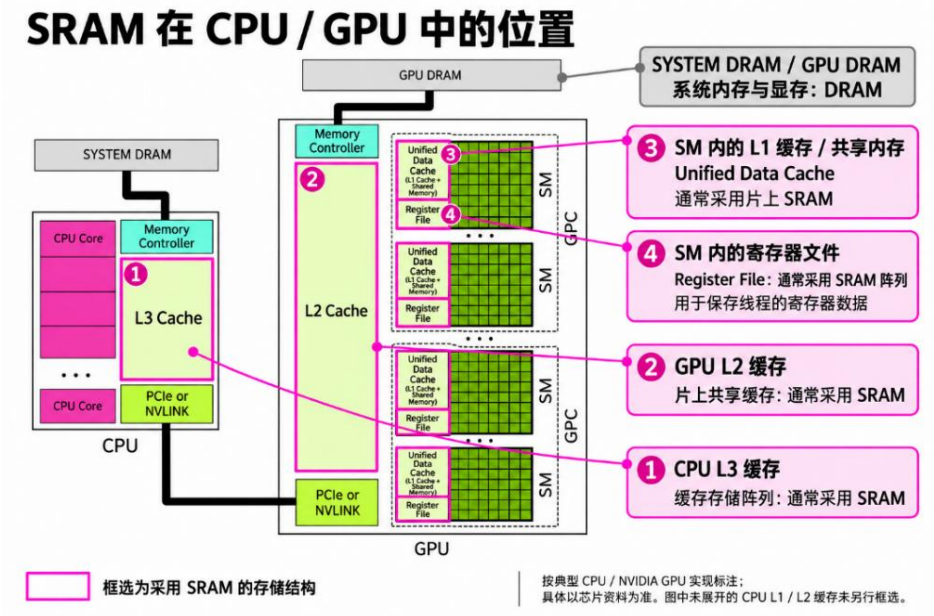
图3：为兼顾速度、容量与成本，CPU 缓存采用分级设计



资料来源：Intel、FPGA 设计论坛、开源证券研究所

仅供内部参考，请勿外传！

图4：除 CPU 外，GPU 等 AI 芯片也采用 SRAM 阵列来支撑并行计算

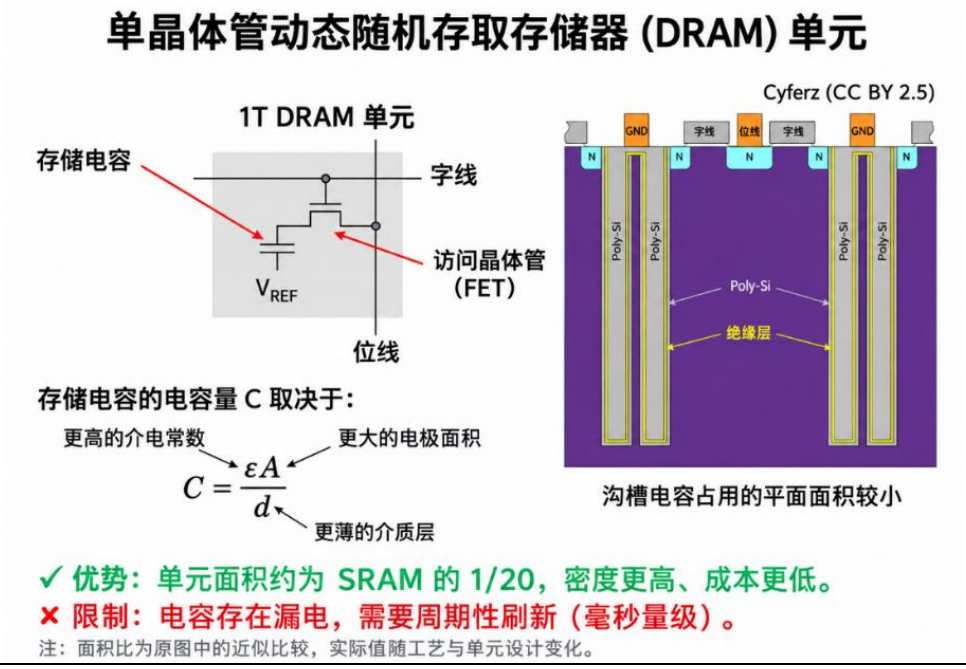


资料来源：NVIDIA、开源证券研究所

（二）DRAM

DRAM 即动态随机存储器，又称系统主存，其数据以电容电荷存储，因电荷泄漏而须周期性刷新。其存储单元仅由一个晶体管与一个电容构成，以电荷有无表示 1 与 0，是半导体存储器中结构最简的形态。简化带来的是密度与成本优势，相同面积可容纳的比特数远多于 SRAM，单位容量成本也低得多，由此承担系统主存职能，成为处理器直接寻址的工作内存。代价则源于电容的物理特性，介质并非理想绝缘，电荷会随时间泄漏，单元必须由控制器周期性读取并重写；刷新既占用带宽，也使其访问延迟显著高于 SRAM，存储层级中容量与速度的取舍由此落定。

图5：DRAM 相同面积可容纳的比特数远多于 SRAM，因电荷泄漏而须周期性刷新



资料来源：MIT OpenCourseWare、开源证券研究所

仅供内部参考，请勿外传！

DRAM 的产品谱系沿两条主线分化：通用主存序列以容量与成本为先，带宽优先序列以位宽为先。前者的差异在功耗约束。DDR 面向计算机与服务器，采用同步接口与双倍数据率；LPDDR 面向移动终端，把低功耗与小体积置于首位。后者共享高带宽目标而集成路径不同。GDDR 在常规封装下堆高单引脚速率，以容忍较高延迟换取带宽，服务于图形与显示；HBM 以硅通孔垂直堆叠多颗裸片、与 GPU 同封装，以位宽而非频率取胜。四类共享同一单元结构，差异集中于接口标准、工作电压与封装形态，而非存储机制本身，这也决定了它们之间不存在直接替代关系。

表1：主流的同步 DRAM (SDRAM) 分为四类，针对不同应用场景优化，各有优劣

对比维度	DDR（双倍数据速率内存）	LPDDR（低功耗 DDR）	GDDR（图形 DDR）	HBM（高带宽内存）
核心定位	通用主存，CPU 协同标准配置	低功耗场景通用内存	图形渲染专用高带宽内存	云端训练极致带宽方案
带宽表现	适中，64 位数据总线	以功耗优先，带宽低于 DDR	显著优于 DDR，低于 HBM	四类最高，3D 堆叠+超宽总线
延迟表现	四类中最低	中等	高于 DDR	较高
成本水平	最低，规模效应最强	较低，消费电子规模支撑	中等，低于 HBM、高于 DDR	四类最高
功耗水平	常规，无深度节能优化	四类最低，支持调压、变频、分级断电	较高	四类最高
典型应用	服务器 CPU 主存、通用计算、电网供电型推理设备	移动终端、汽车电子、边缘计算、消费级 AI、数据中心节能补充	GPU 显存、图形渲染、生成式视觉 AI	数据中心 AI 训练、HPC 高性能计算、高端近内存架构
封装形态	DIMM 内存条（RDIMM/MRDIMM 等）	BGA 封装，直接焊接主板	面向 GPU 显存场景	3D 垂直堆叠，配套 2.5D 封装
核心优势	部署最广、延迟最低、成熟可靠、性价比高	功耗最优、场景适配性强、成本适中	带宽优于 DDR/LPDDR，成本低于 HBM	带宽极致，从根源缓解 AI 内存瓶颈
核心短板	非针对 AI 访存优化，带宽上限低	缺少数据中心级 RAS（ECC、故障隔离）	容量受限，AI 场景定位次优，无强制应用场景	成本极高、散热难度大，长期局限于数据中心

资料来源：中关村集成电路产业联盟、开源证券研究所

DDR 即双倍速率同步动态随机存储器，是 DRAM 面向通用计算的主存接口标准，其在时钟的上升沿与下降沿各传输一次数据，同频下速率翻倍。它与 DRAM 并非并列关系：DRAM 界定存储介质的单元结构与工作方式，DDR 则是在此之上由 JEDEC 制定的接口规范，覆盖引脚、时序、电压与命令协议。其中，存储数据的仍是 DRAM 的电容单元阵列，DDR 规定的是这颗芯片与内存控制器间的通信方式。多颗 DDR 颗粒焊装于模组板，并配以电压调节与系统管理元件，构成可插拔的内存条。除双沿采样外，同步接口使读写与系统时钟对齐、可流水线化，进一步抬高了有效带宽。面向个人计算机与服务器，它须在容量、成本与兼容性间取得平衡，目前沿此约束逐代演进至第五代。

DDR 代际迭代呈现清晰演化特征：工作电压持续下探，单颗粒存储容量扩容，接口数据传输带宽不断提升。JEDEC 于 2020 年 7 月发布 JESD79-5，即 DDR5 规范，标志 DDR 系列迭代至第五代。DDR5 主要作为处理器系统主存，内存条插接于主板插槽，经由内存控制器与 CPU 互联。DDR4 工作电压 1.2V，器件密度 2Gb 至 16Gb，速率区间 1600 至 3200MT/s；DDR5 电压降至 1.1V，器件密度升至 8Gb 至 64Gb，速率区间 3200 至 6400MT/s，带宽较 DDR4 翻倍，并引入片上 ECC，同时把电压调节器移至模组，以放宽电压容差、改善良率。刷新机制同步细化，细粒度刷新与同 bank 自刷新使部分 bank 刷新时其余 bank 仍可访问。片上 ECC 及其他制程微缩配套技术

仅供内部参考，请勿外传！

被纳入 DDR 标准，用于应对 DRAM 先进制程节点的工艺制造限制。

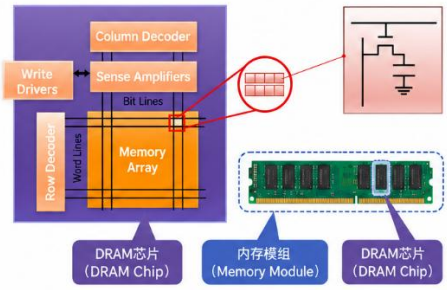
图6：DDR 需要在容量、成本与兼容性间取得平衡，目前逐代演进至第五代

	DDR1	DDR2	DDR3	DDR4	DDR5
封装引脚数	184（TSOP）	240（FBGA）	240（FBGA）	288（FBGA）	288（FBGA+PMIC）
Bank 分组	无	无	2 组	4 组（Bank Group）	8 组（双通道独立控制）
散热设计	被动散热	被动散热	被动散热	部分带散热片	集成 PMIC 散热优化
3D 堆叠技术	无	无	无	3DS 堆叠（部分）	3DS TSV 堆叠（支持 128GB + 模组）

- **制程微缩**：DDR5 采用 14nm 以下工艺（如三星 1β、美光 1z），晶体管密度较 DDR1 提升 25 倍。
- **密度提升**：DDR5 单颗粒密度达 64Gb，是 DDR1 的 64 倍，通过 3DS TSV 堆叠实现。
- **能效优化**：DDR5 功耗较 DDR1 降低 87%（2.5V→1.1V），结合 PMIC 动态调节，能效比提升了 7 倍。
- **晶体管密度**：DDR5 每颗粒晶体管超 5000 万，支持更高带宽和并发处理。

典型应用场景

- DDR1/DDR2：早期手机、工业控制（如 PLC）、机顶盒。
- DDR3：消费级 PC、游戏主机（如 PS4）、车载娱乐系统。
- DDR4：高性能 PC、数据中心（如 AWS 服务器）、AI 工作站。
- DDR5：AI 服务器（如英伟达 DGX）、高端游戏 PC（如 Intel Core i9 平台）、边缘计算设备



DRAM 芯片内部结构

资料来源：芯存社、开源证券研究所

表2：DDR 代际迭代呈现清晰演化特征：工作电压持续下探，单颗粒存储容量扩容，接口数据传输带宽不断提升

对比维度	DDR1	DDR2	DDR3	DDR4	DDR5
制程节点	180-130nm	90-65nm	65-40nm	28-16nm	14-10nm（1z/1β 工艺）
单颗粒密度	128Mb-1Gb	256Mb-2Gb	512Mb-4Gb	1Gb-16Gb	16Gb-64Gb
封装类型	TSOP/FBGA	FBGA	FBGA	FBGA/3DS	FBGA（集成 PMIC）/3DS TSV
预取位数	2 位	4 位	8 位	8 位	16 位（双通道）
Bank 数量	4	4-8	8-16	16	32（8 组×4 Banks）
信号传输技术	单端（SSTL_2）	单端（SSTL_18）	单端（SSTL_15）	伪差分（POD）	差分信号+均衡技术
工作电压	2.5V	1.8V	1.5V（标准） 1.35V（低电压）	1.2V（标准） 1.05V（低电压）	集成 PMIC(5V 输入-1.1V 输出）
纠错机制	无	无	无	可选外部 ECC	片内 ECC（单比特纠错）
晶体管数量	~200 万	~500 万	~1200 万	~2500 万	~5000 万（14nm 工艺）
能效比（mw/GB）	12-1519	8-1019	5-719	3-419	1.5-219

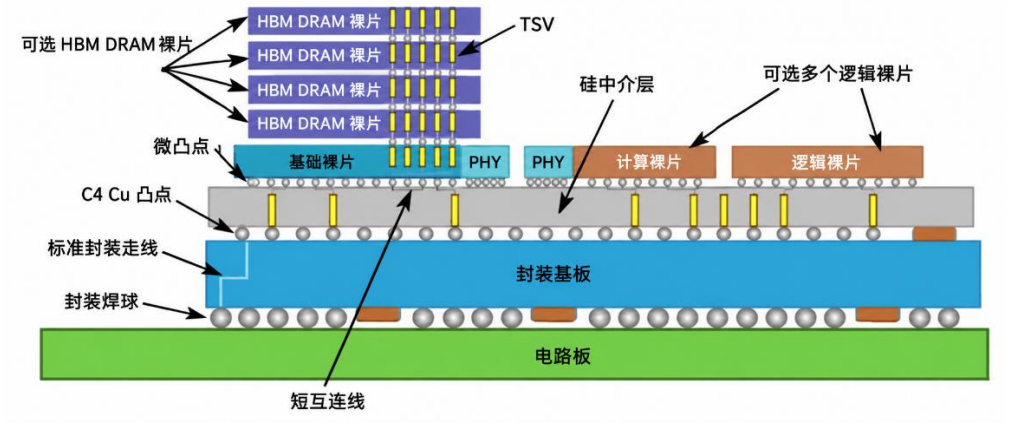
数据来源：芯存社、开源证券研究所

CPU 主存与 GPU 内存的取向不同，前者以延迟与容量为先，后者以带宽为先，带宽缺口最终把 GPU 内存从 GDDR 推向 HBM。DDR 以 DIMM 插槽插在主板，经内存控制器与 PCB 走线接入处理器，链路长达十几厘米，单通道 64bit 位宽、4800MT/s 下带宽约 38GB/s，对 CPU 侧已属充裕。GPU 以万级计算核心并行换吞吐，训练大模型时每秒需搬入数 TB 数据，靠远端低速链路供数将令核心大面积空转；处理器与内存的增速分化持续拉大，即所谓内存墙瓶颈。GDDR 是 GPU 内存的第一代解决方案，通过把颗粒紧贴 GPU 排布，缩短链路并提升引脚速率，GDDR6X 在 384bit 位宽、19Gbps 下带宽约 912GB/s，足以应对图形渲染和中小规模 AI 模型；但当模型参数规模突破千亿级别时，单次训练迭代需要的中间激活值和梯度数据量高达数百 GB，该带宽仍显不足，GPU 内存由此转向 HBM。

仅供内部参考，请勿外传！

HBM（High Bandwidth Memory，高带宽内存）是 3D DRAM 的代表性产品，凭借高位宽特性突破单封装内存带宽瓶颈。HBM 以硅通孔将多层 DRAM 裸片垂直堆叠，经硅中介层与 GPU 并排封装，以位宽而非频率换取带宽。实现依赖三项工艺：其一为 TSV，在每层裸片刻出数千个微孔并填充铜，形成穿透硅片的垂直互连；其二为微凸点，在 TSV 露头处以铜柱凸点完成层间键合与电连接；其三为硅中介层，GPU 与 HBM 倒装焊接于同一硅片，经其表面的高密度铜互连交换信号。中介层与芯片同为硅材质，热膨胀系数一致，可避免温度循环下的应力开裂，有机基板则难以匹配该特性。台积电的 CoWoS 是 HBM 封装的事实标准，其产能在一定程度上直接决定了全球 HBM 和 AI GPU 的供应量。

图7：HBM 以硅通孔将多层 DRAM 裸片垂直堆叠，经硅中介层与 GPU 并排封装



资料来源：马千里、马永辉等《2.5D 封装关键技术的研究进展》、开源证券研究所

HBM 的迭代沿两条主线推进：接口位宽逐代翻倍，堆叠层数与单层容量同步抬升。位宽在 HBM4 翻倍至 2048bit，引脚速率由 6.4Gbps 经 9.6Gbps 升至 10Gbps 以上（SK 海力士增强版），单层容量由 HBM3 的 16Gb 增至 HBM3E 的 24Gb，堆叠层数由 8 层扩至 12 层并向 16 层延伸。三者叠加使单堆栈带宽从 819GB/s、1.23TB/s 抬升至 2TB/s 以上。位宽与层数继续加码使凸点间距成为瓶颈，微凸点现处于 40 至 55μm 量级。混合键合取消焊球、改以铜与铜直接键合后，可将间距压至 10μm 级、IO 密度提升 4 至 16 倍，是 16 层以上堆叠的候选路径。然而混合键合难点在于晶圆需在纳米级平整度下室温贴合再退火，层数越多缺陷累积概率越高。

表3：HBM 的迭代沿两条主线推进：接口位宽逐代翻倍，堆叠层数与单层容量同步抬升

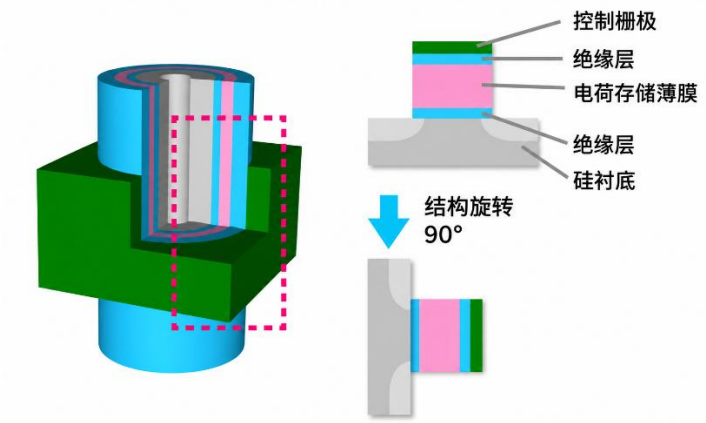
对比维度	HBM1	HBM2	HBM2E	HBM3	HBM3E	HBM4
发布时间	2013 年	2016 年	2018 年	2020 年	2023 年	2025 年（JEDEC 标准）
制程节点	28nm	17nm	17nm	14nm	10nm	10nm（逻辑芯片 4nm/3nm）
单堆栈容量	1GB（4 层）	4GB（8 层）	8GB（8 层）	16GB（8 层）	24GB（12 层）	48GB（16 层）
带宽	125GB/s	256 GB/s	307 GB/s	819 GB/s	1.2 TB/s	2.0 TB/s
引脚速率	1.0Gbps	1.6 Gbps	3.6 Gbps	6.4 Gbps	9.2 Gbps	8.0 Gbps（JEDEC 标准）
工作电压	1.5V	1.2V	1.2V	1.35V	0.8-1.05V	0.7-0.9V
TSV 接口数量	1024	1024	1024	1024	1024	2048
封装技术	TSV 堆叠	TSV 堆叠	TSV 堆叠	TSV 堆叠	TSV+MR-MUF	TSV+混合键合（部分厂商）
能效比（mW/GB）	15-20	10-12	8-10	5-7	3-4	2-3
典型应用	AMD Fiji GPU	NVIDIA P100	NVIDIA V100	NVIDIA A100	NVIDIA H200	NVIDIA Rubin GPU

数据来源：芯存社、开源证券研究所

仅供内部参考，请勿外传！

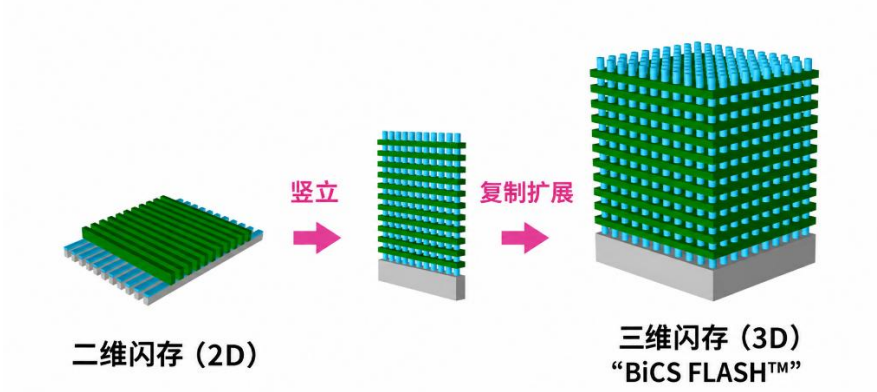
NAND 闪存是闪存的两大分支之一，其单元以串为单位串联共享位线，放弃随机访问以换取单元面积与密度，成为海量数据存储的主导介质。每个单元为带浮栅的场效应管，电子经隧穿进入浮栅改变阈值电压，断电仍可保持数年，属非易失性存储。读写以页为单位、擦除以块为单位，写前须先擦，故需控制器承担磨损均衡与坏块管理。主流产品已转为垂直堆叠：控制栅与绝缘层交替叠起后一次打孔，孔内填电荷陷阱膜与柱状电极，密度靠层数而非缩小线宽提升。3D NAND 凭借大容量、低成本优势，广泛应用于固态硬盘、存储卡及手机内置存储；在数据中心，NAND SSD 正在逐步替代传统机械硬盘，NOR Flash 依靠快速随机读取能力，在嵌入式固件存储领域维持稳固的市场地位。

图8：在 BiCS FLASH（铠侠自研 3D NAND FLASH）存储单元中，电子在穿过柱体中心的电极（图中灰色部分所示）和电荷存储膜（粉色部分）之间进行交换



资料来源：KIOXIA、开源证券研究所

图9：2D NAND 存容量扩展受限，3D NAND 依靠垂直堆叠扩容，已成为主流技术



资料来源：KIOXIA、开源证券研究所

NAND 的存储密度靠在一个单元内切分多个阈值档位换取，档位越多，读取窗口越窄、耐久越低。单层单元（SLC）以两个阈值态存 1bit，多层单元（MLC）扩至四个态存 2bit，三层单元（TLC）与四层单元（QLC）进一步扩至八个态与十六个态，分别存 3bit 与 4bit。每增加一位，单位面积密度与每 GB 成本同步改善，耐久与读取可靠性则同向下滑。相邻档位的电压间隔随之收窄，须以更强的纠错码补偿，可擦写次数由 SLC 约十万次降至 QLC 约千次。用途随之分层，企业级盘为保耐久选 SLC 或 MLC，消费级盘与 U 盘多用 TLC 或 QLC，数据中心读密集场景也开始引入 QLC

仅供内部参考，请勿外传！

以摊薄单位容量成本。

图10：根据存储方式的不同，NAND Flash 可分为 SLC、MLC、TLC 和 QLC，分别适配不同的应用场景

SLC、MLC、TLC 和 QLC 性能参数对比

	SLC	MLC	TLC	QLC
存储位数 / 单元	1bit	2bit	3bit	4bit
结构类型	2D / 3D NAND	2D / 3D NAND	2D / 3D NAND	3D NAND（主流）
制程工艺	14-19nm（2D）/96-128层（3D）	19-40nm（2D）/128-238层（3D）	19-40nm（2D）/238-400层（3D）	19-40nm（2D）/238-400层（3D）
P/E 次数（消费级）	10,000-100,000次	3,000-5,000次	1,000-3,000次	300-1,000次
P/E 次数（企业级）	100,000 + 次	5,000-10,000次	5,000-10,000次	1,000-3,000次
顺序读取速度	200-500MB/s（2D）、1.5-3GB/s（3D）	200-500MB/s（2D）、2-5GB/s（3D）	200-500MB/s（2D）、3-7GB/s（3D）	150-400MB/s（2D）、2-6GB/s（3D）
顺序写入速度	100-300MB/s（2D）、1-2GB/s（3D）	100-250MB/s（2D）、1-4GB/s（3D）	80-200MB/s（2D）、1-5GB/s（3D）	50-150MB/s（2D）、0.8-3GB/s（3D）
随机读写性能（消费级）	50-100K IOPS（2D）、200-500K IOPS（3D）	30-80K IOPS（2D）、100-300K IOPS（3D）	20-60K IOPS（2D）、80-200K IOPS（3D）	10-30K IOPS（2D）、50-100K IOPS（3D）
随机读写性能（企业级）	100-200K IOPS（3D）	200-500K IOPS（3D）	300-800K IOPS（3D）	100-300K IOPS（3D）
能效比（3D版本）	高（1.2V IO电压）	中（1.0-1.2V）	低（0.8-1.0V）	极低（0.6-0.8V）
单芯片容量（3D）	最高128Gb（96层）	最高256Gb（128层）	最高512Gb（238层）	最高1Tb（400层）
成本（位成本）	最高	中	低	最低
典型应用场景	工业控制、军工、高耐用性设备	企业级SSD、车载存储	消费级SSD、U盘、手机存储	大容量数据中心、云存储、归档设备

性能优先级：

SLC > MLC > TLC > QLC，但实际体验受主控和缓存设计影响显著。

SLC/MLC：适合高频写入场景（如数据库、AI训练），但MLC已逐步被企业级TLC替代。

TLC：主流选择，平衡速度与成本，支持PCIe 4.0/5.0接口，满足游戏和创意设计需求。

QLC：适合大容量存档（如视频库、备份），但需避免频繁写入。

选购建议：

- 普通用户：TLC性价比最优，QLC适合大容量需求。
- 专业用户：高端TLC（如带独立缓存）或企业级MLC。
- 企业级：3D QLC用于冷存储，SLC/MLC用于高频写入场景

读写速度：SLC > MLC > TLC（缓存内） > QLC（缓存内）

缓外速度：SLC ≈ MLC > TLC > QLC

寿命：SLC > MLC > TLC > QLC（传统）≈ TLC（3D QLC）

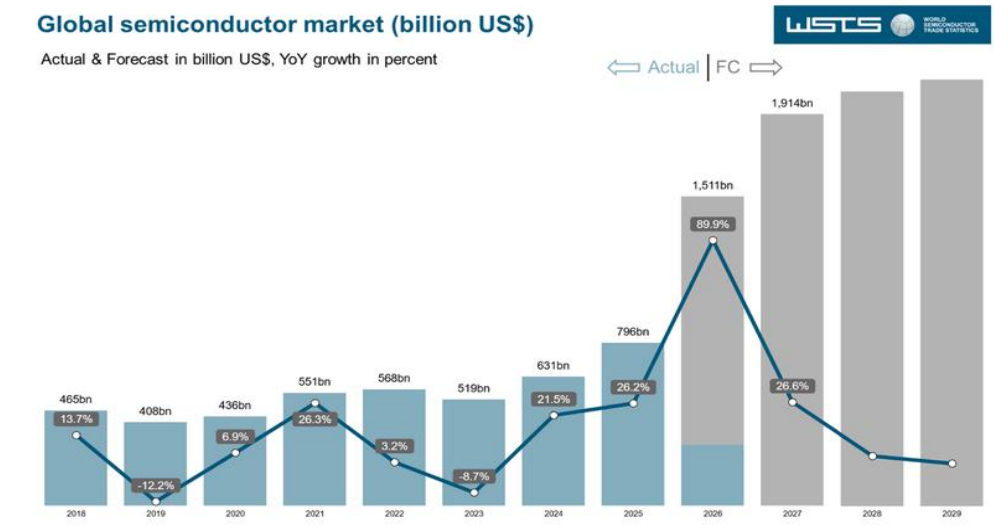
容量/成本比：QLC > TLC > MLC > SLC

数据来源：芯存社、开源证券研究所

1.2、 存储周期属性弱化，供给刚性成为定价主轴

存储是本轮半导体上行周期中弹性最大的品类，驱动变量已由消费电子切换为 AI 算力，需求可见度提高而供给弹性未同步改善。据 WSTS 2026 年春季预测，2026 年全球半导体市场增长 90%至 1.51 万亿美元，存储增长约 250%至超 8000 亿美元，占比过半。该预期一年内被大幅上修：WSTS 在 2025 年春季曾预测 2026 年存储市场为 2148 亿美元、增长 16.2%。据 TrendForce，HBM 与 RDIMM 合计将占 2026 年 DRAM 位元供给的 51%，PC 与手机厂商因成本压力下调排产，云厂商则为锁定供应接受涨价并签署长期协议；而晶圆投片仅温和增长、新厂产能需到 2027 年下半年才爬坡，供需节奏差构成本轮周期的基本形态。

图11：2026 年全球半导体市场增长 90%至 1.51 万亿美元，存储为主要驱动因素



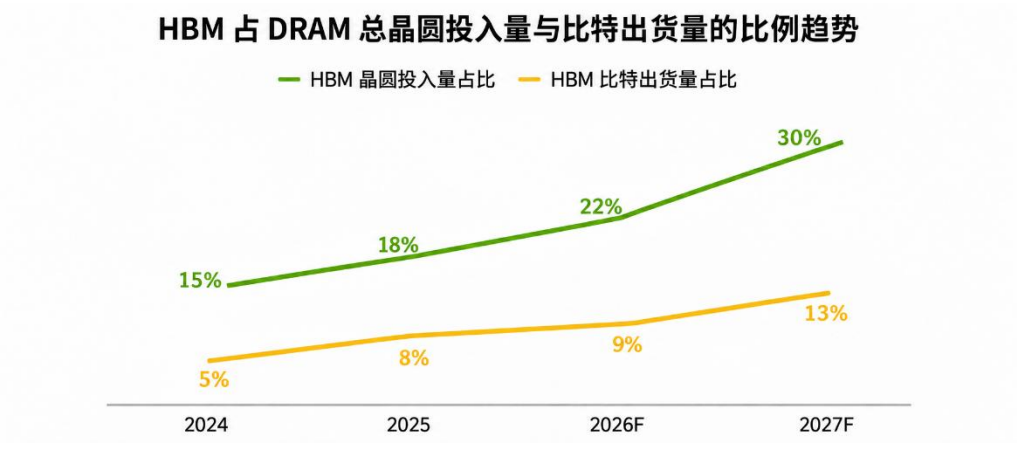
资料来源：WSTS

仅供内部参考，请勿外传！

虽然 DRAM 与 NAND 合约价同期上行，但驱动来源不同：DRAM 源于产能向服务器与 HBM 重新配置，NAND 源于供给端的主动收缩。2026 年 Q2 传统 DRAM 合约价环比上涨 58%至 63%，NAND 上涨 70%至 75%；2026 年 Q3 传统 DRAM 涨幅收敛至 13%至 18%，TrendForce 指出原因是 PC 与手机客户吸收涨价的能力有限、部分需求由高容量 RDIMM 转向低容量 RDIMM；NAND 的涨价则来自供给纪律，企业级 SSD 合约价 2026 年预计累计上涨 235%，供应商以限制对客户级 SSD 供给的方式维持价格。价格机制同时在反向调节产能分配：2026 年 Q1 HBM 每晶圆收入被 DDR5 64GB RDIMM 反超，供应商或将视 HBM 定价结果重新调配产出；产能紧张延续之下，供应商在 2027 年价格谈判中的议价能力预计增强。

存储供给的刚性源于晶圆产能扩张速度与 AI 需求增速之间的错配，该缺口短期内难以通过扩产弥合。单颗 GPU 搭载的 HBM 容量随代际持续抬升，H100 为 80GB HBM3，DGX B200 折合每颗 180GB；TrendForce 预计 2026 年 AI ASIC 单颗容量将由 96GB 或 192GB 提升至 216GB 或 288GB，2027 年 Rubin Ultra 每颗 GPU 进一步升至 384GB。HBM 的单位晶圆产出效率显著低于常规 DRAM：TrendForce 预计至 2026 年底，三大厂 HBM 晶圆投入占 DRAM 晶圆总投入的比例约为 22%，位元供给占比仅约 9%，2027 年两者分别升至 30%与 13%；随着代际演进带来的裸片尺寸扩大与需求同步增长，对常规 DRAM 产能的挤出效应预计进一步加剧。当前供应商库存已处历史低位，新增供给主要配给服务器。

图12：2027 年 HBM 迭代升级、需求增长，对传统 DRAM 产能挤出效应持续放大



数据来源：TrendForce、开源证券研究所

2、 计算机产业动态更新

2.1、 国内 CSP 厂商与大模型厂商最新动向

（一）月之暗面推出 Kimi 金融行业方案，整合 9 项专业技能

9 月 17 日，月之暗面 Kimi 推出金融行业解决方案，整合 10+权威数据源、9 项金融技能、5 项合规安全措施及机构级数据建模与报告交付能力。方案内置 Wind 万得、东方财富、标普全球、财联社、财新数据等数据源，通过 MCP 直接接入研究任务，关键数字可回溯原始出处。在已落地机构合作中，该方案已将过去以天计的资料处理与初稿制作压缩至小时级，让专业人员将更多时间留给分析与判断。

仅供内部参考，请勿外传！

Kimi 与一线机构分析师、基金经理和投行团队合作，将金融工作封装为机构财务建模、机构研究报告、机构 PPT、业绩点评、一致预期地图、组合复盘、持仓早报、HK IPO 透镜等 9 个技能，并已在中金公司、中信建投证券、德意志银行等数十家机构跑通。财务建模人力投入由 5 至 7 人天降至 0.5 至 1 人天，量子科技深度报告周期从 10 至 20 天缩短至约 2 天。

合规方面，Kimi 与中信建投证券共建“风险评估网关”，落地数据分类分级、个人信息保护、数据源与工具访问授权、生成内容核验与人工复核、审计与责任追溯五项措施。该网关已在试点中覆盖 50 余家发行人、60 余份临时受托管理事务报告，单份报告人工制作时间由约 30 分钟缩短至 10 分钟、投入下降 67%，并支持跨场景复用。

（二）小米公开 MiMo-V2.6 RL 训练，押注 Agent 时代

9 月 17 日，MiMo 负责人罗福莉首次公开 MiMo-V2.6 强化学习训练的实时页面，让外界看到这一代模型 RL 阶段的训练状态。页面显示两个版本，MiMo-V2.6-Pro 处于 step 12、累计样本约 301K、训练成本约 80.63 万美元、Token 达 25.1B；MiMo-V2.6-Flash 处于 step 17、样本约 426K、成本约 35.45 万美元、Token 达 40.5B。两个版本累计成本超 116 万美元，平均每小时约 3.09 万美元、折合人民币约 20.72 万元。

罗福莉透露团队正扩大三个方向规模，计算规模上单个 step 约 2B tokens、采用 1568 个 prompts×16 rollouts 的完全异步训练；训练环境引入 multi-task agentic RL，混合视觉、代码、通用与聊天任务；奖励评估引入组内贡献归因并结合测试用例与评分细则。能力方面，DeepSWE v1.1 基准 MiMo-V2.6-Pro 达 63.72、Flash 达 60.77，平均上下文长度接近 10 万 Token，信号指向小米正押注 Agent 时代。

（三）智谱 GLM 自优化推理系统，10 万国产卡两周部署

9 月 17 日，智谱 GLM 首席科学家唐杰公布 GLM-5.3-Flash 推理系统优化研究。该模型在超 10 万颗国产 AI 加速器组成的集群上，从首次运行到承载全部生产流量仅用两周，端到端吞吐能力提升 3.2 倍。完成大量优化工作的除基础设施工程师外，还有由 GLM-5.3 驱动的 Infra Agent，唐杰称其为“帮助优化服务自身的模型”，意味着 AI 开始参与优化承载自身运行的系统。

团队引入“dense feedback”机制，让 Agent 获得接近工程判断的反馈。Agent 定位了 KDA 上下文并行路径中 TF32 舍入误差累积问题，修复 KV Transfer 与 DeepEP 调度重叠问题、将额外开销从超 30%降至 1%以下，并将 Decode Kernel 性能提升 1.71 倍。优化采用 ReplaySSM 以计算换内存、INT8/FP8/BF16 混合精度缓存、EPD 分离式架构等方案，硬件利用率与单 token 成本接近主流 NVIDIA 平台。

该模型曾以匿名名 Ox-Alpha 运行于 OpenCode 与 OpenRouter，六天处理超 62 万亿 token。唐杰认为虽距真正递归自我改进仍远，但“模型优化系统、系统服务模型”的最小循环已出现，工程师角色正从直接解决问题转向设计反馈系统，Agent 完成的工程任务或成为下一代模型的训练数据。

仅供内部参考，请勿外传！

2.2、 海外 CSP 厂商与大模型厂商最新动向

（一）OpenAI 或迎发布周，GPT-6 Sol 疑似开启灰度测试

9 月 16 日，OpenAI 产品负责人 Tibo 在 X 上发文称本周发布的产品与功能数量将达到 DevDay 2025 级别，CEO Altman 亦罕见预告“本周会有一个大版本发布，DevDay 期间还有更多重磅发布”，引发外界对 GPT-6 Sol 或将发布的猜测。多信号叠加指向模型换代。ChatGPT 官方宣布自 10 月 14 日起 GPT-5.5 停止提供，被视为为下一代模型腾位置。产品层面，9 月 8 日 OpenAI 推出 ChatGPT Images 2.5，9 月 10 日一天内连推 Agents API、GPT-Live-1、Data agent 及面向金融行业的 ChatGPT for Financial Services，横跨智能体、实时语音、数据分析与垂直行业。无论最终发布新模型还是产品升级，信号均指向 AI 竞争重点正从“模型更聪明”转向“模型能否完成复杂任务”。

（二）Opus 5.2 灰度上线，Anthropic 押注 RSI 自我改进

9 月 15 日，Claude Opus 5.2 在 Claude Code 中开启灰度测试。开发者通过抓包发现前端名称未变、但模型 slug 已指向 Opus 5.2，Anthropic 疑似跳过 5.1 版本直接端出 5.2。实测反馈显示，Opus 5.2 响应速度大幅提升、输出简洁直切要害，并彻底告别“偷懒病”，可自主进行“严酷循环”不断迭代完善代码直至任务完成，被评价为底层能力蜕变而非简单微调。

开发者用“Tibo”冷知识探针测试是否被灰度路由，验证新旧版本权重不同。据 8 月中旬曝光的内部风险报告，Anthropic 手握三套方案应对 GPT-6 Astra，其中未发布的内部模型 Model 2 在 CoBench v2 榜单打出 62.8 分，较 Mythos 5 高出 12.5 分，已接管公司大部分代码编写；更进一步的 RSI 模型宣称可替代研究团队 85%的工作，性能较 Model 2 再高 22 分。

产业层面，Anthropic 的 ASI 路径清晰，即让 AI 创造更好的 AI 而非仅服务人类。Fable 5.2 家族已整装待发，预计最快 9 月底、最迟 10 月底全面推向市场。面对 OpenAI GPT-6 Astra 的威胁，拥有自主迭代能力的 Opus 5.2 能否在 9 月全面开放，成为 ASI 终局之战的焦点。

（三）谷歌上线 Gemini 3.8 Live，语音可后台调用工具

9 月 16 日，Google DeepMind 发布 Gemini 3.8 Live 与 Gemini 3.8 Live Extended Thinking 两个实时对话模型，官方称其为迄今最先进的实时对话模型。核心突破是“说话、思考、后台干活”三件事互不打断，Extended Thinking 遇到复杂任务会先用“Let me check that…”接住，再边推理边汇报进度，同时可先回应请求、将工具调用与 API 请求放到后台执行。模型支持 97 种语言中途自动切换、近实时视觉输入直接进对话。

成绩单方面，Extended Thinking 在 Artificial Analysis 语音质量榜以 82.6 分位列总榜第一，在 τ-Voice 上取得 68.6%、小幅领先 GPT-Live-1 Astra 的 67.9%，在银行场景 τ-Voice-banking 取得 35.1%。两模型被定位为“可上生产的语音 Agent 基础组件”，标准版主打规模化与成本效率，Extended Thinking 主打高复杂度多步推理。这标志人机配合正从“下指令、等结果”转向“边聊边并行干活”，屏幕首次不再是人机交互的必经之路。

（四）Anthropic 合并 Cowork 与 Chat，Claude 成统一入口

仅供内部参考，请勿外传！

9 月 17 日，Anthropic 宣布 Claude Cowork 与 Claude Chat 合并为统一的 Claude，用户无需提前判断任务应交给哪个工具，Claude 能根据任务需求自动判断所需能力，让 Cowork 与 Design 的功能在任意对话中直接调用，并继承已有上下文、技能与连接器。同时，Claude Docs 与 Claude Slides 正式推出，Claude Design 亦可直接在对话中使用，用户可直接编辑内容、演示或将文件下载为 PowerPoint 或 PDF 格式。

使用上，用户可一句话让 Claude 完成周报并同步生成管理层幻灯片，任务可在关闭电脑后继续执行，甚至设为每周自动执行。这三项功能目前处于 Beta 测试，未来几周首先向 Pro 与 Max 计划用户开放，覆盖网页、桌面与移动端，之后陆续覆盖 Team 与 Free 计划；Enterprise 用户的管理员将在组织设置变化前至少 30 天收到通知。原 Cowork 用户的聊天记录、项目、生成内容、连接器与技能均保留在原位。

2.3、 国产算力产业最新进展

（一）华为发布昇腾 960 超节点，灵衢互联支持百万卡集群

9 月 17 日，华为在 2026 年全联接大会上宣布昇腾 960DT 芯片提前就绪，支持 2 PFLOPS FP8、4 PFLOPS FP4 算力，较上代实现翻倍，预计 2027 年一季度就绪。依托韬定律创新方向，昇腾系列芯片坚持一年一代演进节奏，2028 年、2029 年将陆续推出昇腾 970 与 980 芯片实现算力继续翻倍。发布会核心是灵衢 UnifiedBus 统一互联总线，基于其单集群可支持 100 万颗 AI 处理器，10 万卡集群浮点算力利用率从 20%提升至 35%。

硬件方面，华为推出业界首个量产 NPO 光引擎 Hi-ONE，传输能力达 7.2Tbps 且内置光源。昇腾 960 超节点采用该 NPO 方案，单柜卡规模，提供 8 EFLOPS FP8 算力与 1PB HBM 容量。鲲鹏 950 超节点升级 4096 节点、256TB 统一内存池，十万级沙箱启动提速 30 倍。OceanStor M900 记忆存储将 KV Cache 时延降低超 50%。灵衢加二层 CLOS 组网可构建百万卡超大规模集群。

生态方面，昇腾已跨越生态拐点，CANN 社区月活开发者突破 5000 名，外部开发者占比 61%；鲲鹏生态全球开发者超 416 万，openEuler 装机量超 2000 万套，成为中国服务器操作系统份额第一。端侧方面，华为以“麒麟+昇腾”实现端侧算力自给，以“盘古+三方大模型”提供端侧智能，盘古 MoE 30B 参数大模型已在华为 Mate X2 商用，并将围绕手机、AI PC、车与家庭打造四大端侧算力平台。

（二）联发科推出 2 纳米天玑 9600 Pro，端侧可跑 30B 模型

9 月 16 日，联发科发布 2nm 旗舰芯片天玑 9600 Pro，基于台积电 2nm 制程，内部集成 330 亿晶体管，采用 AI 原生架构设计，支持端侧运行 30B 参数 MoE 模型。双 NPU 整体重构，超能效 NPU 功耗降低 40%，超性能 NPU 每瓦算力提升 50%。生成式 AI 引擎 3.0 将 INT4 运算能力翻倍并加入混合注意力运算，叠加智能内存调控技术，使大模型冷启动提速 27%。跑分方面，天玑 9600 Pro 在苏黎世 AI 性能测试达 25000 分以上，预填充速度提升 60%。

CPU 采用 2+3+3 全大核方案，缓存总容量 34.5MB、二级缓存提升 25%，超大核功耗降 37%、多核降 61%，Geekbench 6.4 单核接近 4300 分、多核超 12600 分。GPU 沿用 AI 算力补传统渲染思路，同功耗性能提升 18%、光追性能提升 18%，APU 5 支持 480P 超分至 1080P，帧生成单帧功耗最多降 22%。影像方面，ISP 与 NPU 组

仅供内部参考，请勿外传！

成联合架构，与虹软合作全景深计算光学，支持 4K 60 帧、4K 慢动作 240 帧。

真机方面，vivo 宣布首款搭载天玑 9600 Pro 的手机 vivo X500 Pro 系列将于下周一发布，OPPO 紧随其后，将于下周二发布同样搭载该芯片的 Find X10 Pro Max。

2.4、 海外算力产业最新进展

（一）台积电曝光 2027 年扩产计划，3nm 月产能冲 21 万片

9 月 14 日，据台媒《经济日报》援引供应链消息，台积电首次曝光 2nm 与 3nm 扩产目标。3nm 月产能 2025 年底达 12 万至 13 万片，今年底扩增至 18 万片，目标 2027 年中超过 21 万片、半年内增加 16%，换算 2025 年底至 2027 年中一年半内增加近 70%。2nm 月产能今年底将达 9 万片，2027 年中进一步扩大至 11 万片、半年内增加 22%。

投入方面，台积电此前表示 2026 年资本支出将介于 600 亿美元至 640 亿美元，其中约 70%至 80%用于先进制程，并正新增三座 3nm 晶圆厂，分别位于中国台湾、美国亚利桑那州及日本。同时，台积电持续推进 A16、A14 等更先进制程，与 ASML 合作推动 High NA EUV 自 6 英寸光罩转向 12 英寸光罩。IMEC 亦宣布携手 ASML 利用 High NA EUV 单次曝光完成 22nm 间距导线图形化，较此前的 28nm 间距再进一步。

（二）英伟达推出 CUDA-Q Logical，容错量子开发提速 7 倍

9 月 15 日，NVIDIA 宣布扩展开源 CUDA-Q 平台，推出 CUDA-Q Logical 编排层，为容错量子计算应用提供可编程且可验证的方法。借助该层，研究人员可设计并编排容错应用所需组件，在不同选项间切换，识别能发挥逻辑量子比特最佳性能的系统配置，且不受量子比特类型与架构限制。英伟达量子计算副总裁称，这将大幅缩短实现实用量子-GPU 超级计算的时间。

成效方面，Fermilab 利用 CUDA-Q Logical 将容错架构开发周期从 5 个月缩短至 3 周，实现 7 倍提速；Iceberg Quantum 针对 Diraq 量子比特建模，仅用 15 万个物理量子比特即创建 1000 个逻辑量子比特，约为 Diraq 此前估算的 1/10。此外，Sandia 国家实验室开发的容错量子硬件就绪度基准 QUOPS 现已在 CUDA-Q 上提供，可评估容错量子硬件在实际应用中的就绪程度。

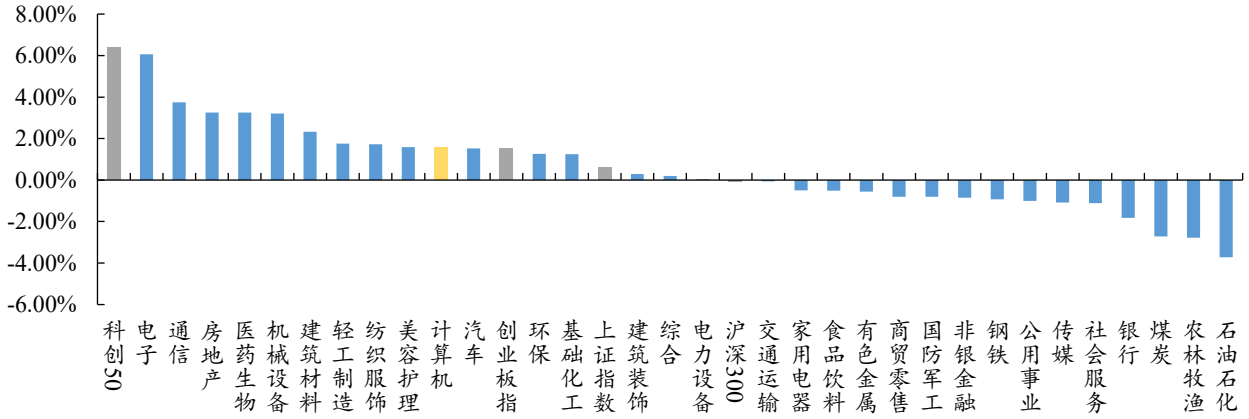
3、 上周市场表现回顾（2026/9/14-2026/9/18）

计算机指数上涨 1.58%，相对主要市场指数表现较强。同期沪深 300 指数下跌 0.06%，上证指数上涨 0.61%，创业板指上涨 1.52%，科创 50 指数上涨 6.39%，计算机指数分别跑赢 1.64、0.97 和 0.06 个百分点，跑输 4.81 个百分点。

横向比较一级行业，本周科创 50、电子、通信表现居前，分别上涨 6.39%、6.06%和 3.75%；计算机板块上涨 1.58%，在 31 个一级行业中排名第 11 位。在本周 TMT 板块行情共振向上的背景下，计算机板块相对表现偏温和。

仅供内部参考，请勿外传！

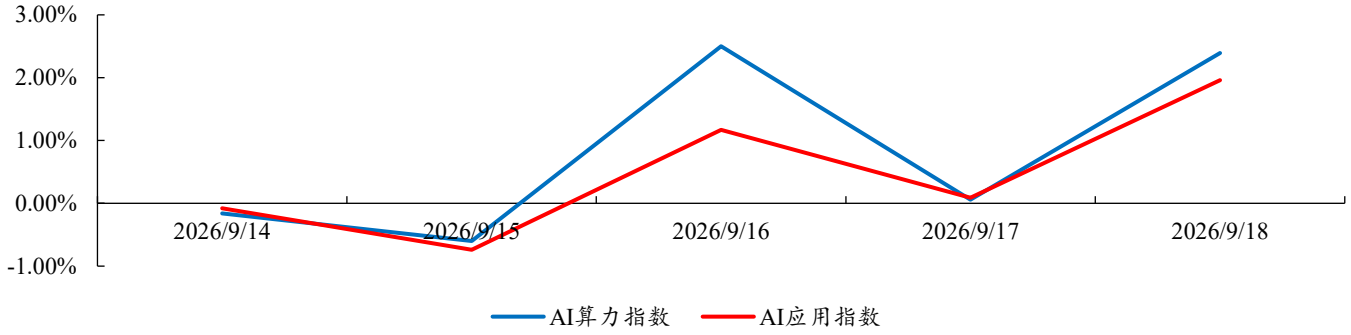
图13：在本周 TMT 板块行情共振向上的背景下，计算机板块相对表现偏温和



数据来源：Wind、开源证券研究所

本周 AI 细分方向呈现一定分化，**AI 算力指数整体表现明显优于 AI 应用指数**。按照日度涨跌幅复合计算，9 月 14 日至 9 月 18 日 AI 算力指数累计上涨约 **4.22%**，AI 应用指数累计上涨约 **2.40%**，AI 算力相对 AI 应用取得约 **1.82** 个百分点的超额收益。

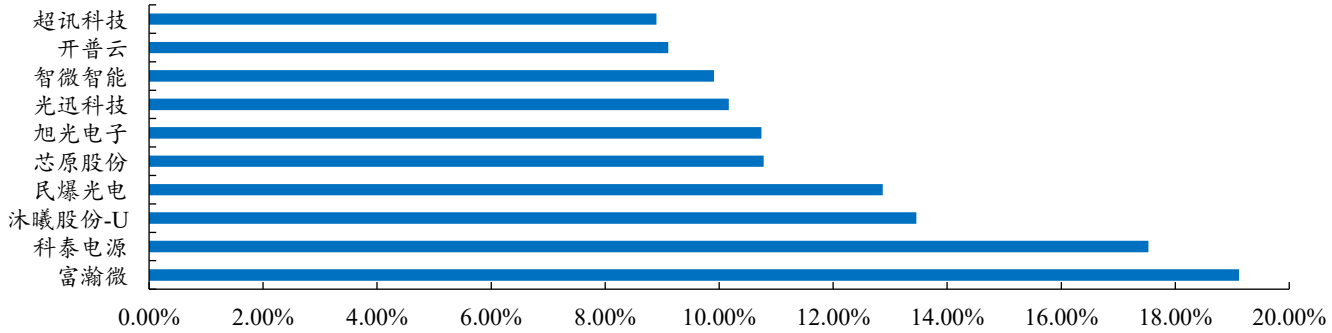
图14：本周 AI 细分方向呈现一定分化，AI 算力指数整体表现明显优于 AI 应用指数



数据来源：Wind、开源证券研究所

AI 算力指数整体小幅上涨，但**板块内部仍存在较为明显的结构性行情**。富瀚微本周上涨 **19.12%**，科泰电源上涨 **17.53%**，沐曦股份-U 上涨 **13.46%**，均录得两位数涨幅。

图15：AI 算力指数整体小幅上涨，但板块内部仍存在较为明显的结构性行情

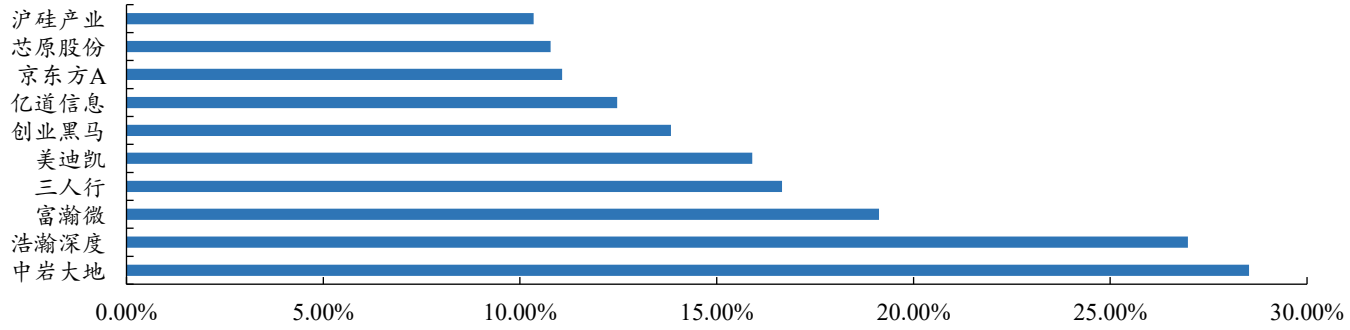


仅供内部参考，请勿外传！

数据来源：Wind、开源证券研究所

AI 应用方向指数整体上涨，龙头个股赚钱效应更为突出。涨幅前十个股周涨幅全部超过 10%，中岩大地本周上涨 **28.53%**，位居 AI 应用板块涨幅首位。

图16：AI 应用方向指数整体上涨，龙头个股赚钱效应更为突出



数据来源：Wind、开源证券研究所

4、 风险提示

大模型技术迭代不及预期风险。大模型技术迭代是推动产业能力边界拓展与下游场景深化的核心驱动力，当前行业正处于多模态感知、复杂推理、智能体执行等前沿技术的攻坚突破期，技术突破节奏、能力落地效果均存在一定不确定性。若后续核心技术突破进度、模型能力提升幅度低于市场预期，无法有效支撑下游高价值场景的深度适配与功能升级，可能导致产业落地节奏整体延后，行业成长进程存在阶段性放缓的可能性。

AI 技术应用不及预期风险。下游场景规模化渗透是大模型产业价值兑现的核心支撑，当前大模型正从技术概念验证期向垂直行业规模化应用期过渡，落地进度受行业数字化基础、客户付费意愿、投入产出比验证等多重因素共同制约。若后续垂直行业适配进度、企业付费转化节奏不及市场预期，产业商业价值的兑现周期可能有所拉长，行业长期成长动能的释放节奏存在一定波动可能。

相关政策不及预期风险。大模型产业发展高度依赖政策环境的引导规范与配套支持，当前人工智能领域处于快速发展阶段，数据安全合规、技术监管标准、产业扶持政策、市场准入规则等相关制度仍处于持续完善过程中。若后续行业监管要求趋严、扶持政策落地进度慢于预期，或合规要求抬升导致企业研发与运营成本增加，可能对大模型技术迭代、商业化落地及行业竞争格局形成阶段性扰动，行业发展节奏存在相应调整的可能性。

AI Infra 建设不及预期风险。基础设施是大模型技术迭代与产业规模化落地的底层算力与工程支撑，算力集群建设、高速互联部署、存储体系配套等基建环节具备投入规模大、建设周期长、产业链协同要求高的特征，落地进度受硬件供给、资金投入、工程能力等多重因素约束。若后续高端算力供给受限、基建项目建设进度慢于市场预期，可能形成算力供给瓶颈，制约大模型训练迭代与应用规模化落地，行业整体发展节奏存在阶段性承压的可能。

仅供内部参考，请勿外传！

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已正式实施。根据上述规定，开源证券评定此研报的风险等级为R3（中风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为专业投资者及风险承受能力为C3、C4、C5的普通投资者。若您并非专业投资者及风险承受能力为C3、C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师声明

本研究报告的署名人员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，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（Outperform）	预计相对强于市场表现 5%～20%；
	中性（Neutral）	预计相对市场表现在－5%～＋5%之间波动；
	减持（Underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（Overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（Underperform）	预计行业弱于整体市场表现。
备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数（北交所基准指数为北证 50 指数）、港股基准指数为恒生指数、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）、美股基准指数为标普 500 或纳斯达克综合指数。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取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。		

仅供内部参考，请勿外传！

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（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。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。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。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。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。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户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。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。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。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，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。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海
地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

深圳
地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

北京
地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

西安
地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn