CFM MemoryS

2025-2026年全球
存储市场趋势白皮书

Global Memory Market Trend White Paper

2026年3月

CATALOGUE

目 录

第四章 消费类存储产品应用与发展分析
ANALYSIS OF CONSUMER MEMORY PRODUCTS APPLICATION AND DEVELOPMENT

第五章 新兴消费领域存储产品应用与发展
APPLICATION AND DEVELOPMENT OF MEMORY PRODUCTS IN EMERGING CONSUMER SECTORS

第一章 全球 NAND Flash/DRAM 技术发展趋势

GLOBAL NAND FLASH/DRAM TECHNOLOGY DEVELOPMENT TREND

一、高层堆叠与架构创新驱动全球 NAND Flash 技术演进

全球 NAND Flash 向高层堆叠和架构创新的发展持续演进，300 层以上 NAND Flash 加速研发进入量产期，多层 Deck 堆叠、增加 Plane 数量以及混合键合技术迎来更广泛的应用。

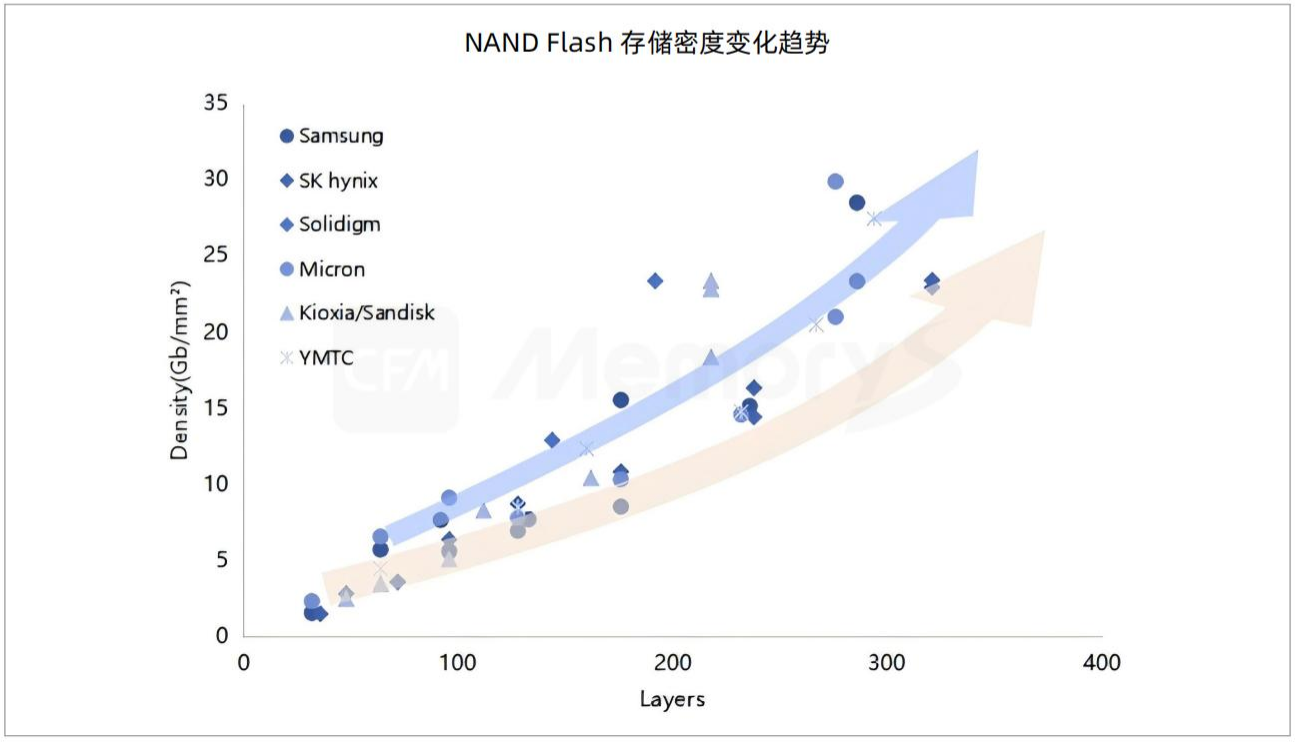
多层 Deck(Multi-Deck) 指在高层 NAND Flash 垂直方向上分段堆叠子存储阵列，通过分段构建以减少单次刻蚀和沉积的难度，再进行垂直贯通以实现总层数的扩展。目前的 NAND Flash 的高深宽比刻蚀控制在可接受范围，随着 NAND Flash 向 400 层发展，需要更多沉积和对准步骤，以及更多的 CMP 和复杂刻蚀，晶圆制造成本有所提升，但每 Bit 成本仍然会随着单 die 存储密度提高而下降。

Plane 作为 NAND Flash 内部细粒度最小的并行操作单元，同一个 die 内的不同 Plane 拥有各自独立的寄存器、页缓冲器、局部行 / 列解码器和部分控制逻辑单元，是可独立执行读写或擦除操作的基本阵列单元。控制器可以同时向不同的 Plane 发送命令，从而分散不同 Plane 中 Block 的管理压力，显著提升 NAND Flash 的吞吐量和并行度，而并行度又是影响 IOPS、QoS、尾延迟、写放大和寿命管理的核心变量。4-Plane NAND Flash 在性能、功耗和复杂度之间取得了较好的平衡，因此目前 4-Plane 仍是最主流的 NAND Flash，而 NAND 原厂也开始探索更高 Plane 数量以提升并行度，譬如美光在 238 层 NAND 中最早引入 6-Plane，铠侠和闪迪也在研究 8-Plane 的 NAND Flash，更高 Plane 数的 NAND Flash 有望逐渐增加。

在 NAND Flash 的架构方面，3D NAND 通过垂直堆叠高层数有效提升存储密度，传统架构通过将存储阵列放置在 CMOS 外围逻辑电路上，节省了芯片面积，但由于存储阵列和逻辑电路仍在同一晶圆中，高层存储单元和 CMOS 逻辑电路难以同时优化。高层 NAND Flash 广泛采用晶圆级别的混合键合技术（Hybrid Bonding），实现存储阵列和逻辑晶圆的解耦制造。混合键合技术本质上是晶圆到晶圆（Wafer-to-Wafer）的直接键合，优势在于可以分别优化存储阵列和 CMOS 外围电路，支持采用更激进的材料和工艺来提升 NAND Flash 的可靠性。这不仅显著提高了存储阵列和外围电路之间的互连程度，提升了 I/O 带宽和访问速度，降低了功耗及芯片面积，并在高层 NAND Flash 量产中，解耦两片晶圆的制造有效提升了生产效率。

国产存储厂商长江存储最早从 64 层 NAND Flash 开始采用混合键合技术，在晶圆对准精度、键合可靠性和高密度互连设计等方面积累了大量专利和工艺经验，引领混合键合技术在 NAND Flash 领域中的广泛应用。铠侠和闪迪已在 218 层的 BiCS8 NAND Flash 中运用键合技术并实现大规模量产，率先用于企业级 SSD 和高性能存储产品中。三星也计划在其约 400 层级的 V10 NAND Flash 中引入混合键合技术。随着 NAND Flash 层数持续向 400 层甚至更高发展，传统单晶圆工艺在刻蚀深孔、沉积层数以及工艺复杂度方面面临越来越大的挑战，而通过键合技术将存储阵列与逻辑电路甚至多个阵列晶圆进行多维集成，可以有效降低制造难度，进一步提升存储密度和性能。

图 1 NAND Flash 存储密度变化趋势



数据来源：CFM 闪存市场

三星（Samsung）

最新量产的 286 层 V9 NAND 存储密度较 V8 NAND 提高约 50%，V9 NAND 通过消除虚通道孔显著减少了存储单元的平面面积，并采用先进的“通道孔刻蚀”技术，可在双层结构中同时钻孔。随着单元层数的增加，穿透更多单元的复杂刻蚀技术能力变得至关重要。V9 NAND 配备 NAND 闪存接口“Toggle 5.1”，将数据输入 / 输出速度提升 33%，最高可到每秒 3.2Gbps，同时，V9 NAND 的功耗较 V8 NAND 降低了 10%。

SK 海力士（SK hynix）

最新量产的 V9 NAND Flash 采用 321 层堆叠结构，是业界最早进入 300 层级量产阶段的 NAND Flash 之一，通过三段堆叠（3-deck）和 Plug 电气连接实现超高层数结构。同时，通过优化通孔填充材料与沉积工艺，降低高深宽比结构带来的形变问题，并引入自动对准矫正技术，提升层间对准精度和制造良率。此外，V9 QLC NAND Flash 将传统 4-Plane 架构扩展至 6-Plane，以提高并行处理能力并缓解 QLC NAND 可能带来的性能下降问题。

美光（Micron）

最新量产的第九代（G9）3D NAND 采用 276 层堆叠结构，在保持高层数扩展的同时，通过 Replacement-Gate（RG）架构结合 CMOS-under-Array（CuA）设计，有效减少外围电路面积并提升单位晶圆存储密度。该架构通过金属控制栅替代传统多晶硅栅极，降低单元间电容耦合并改善电阻特性，从而提升阵列可靠性与写入效

率。在架构层面，美光自 232 层 NAND 开始，引入 6-Plane 并行架构以提高内部并行度和读写带宽，是第一个大规模商用 6-Plane 的 NAND Flash 原厂。

铠侠 / 闪迪（Kioxia/SanDisk）

最新量产的 BiCS8 NAND 为 218 层，并持续向 300 层以上演进，自 BiCS8 NAND 起，引入 CMOS directly Bonded to Array（CBA）键合技术。铠侠与闪迪双方将位于日本四日市和北上工厂的合资协议延长至 2034 年 12 月 31 日，两家公司在设备投资、先进工艺开发以及下一代高层 3D NAND 结构方面保持深度合作，以确保先进 3D NAND 的稳定量产和供应。

长江存储（YMTC）

最新量产的 Xtacking4.0 系列 3D NAND，接口速度达到 3600MT/s。长江存储持续巩固自主研发的混合键合 Xtacking 架构，显著提升 I/O 带宽并缩短信号路径，实现更高的数据吞吐率与存储密度，进一步提高晶圆利用率和生产效率。面对企业级数据量的指数级增长，长江存储凭借 32DP（32-Die）超多堆叠封装技术，推动超大容量 QLC eSSD 步入实用阶段且性能损失几乎与 TLC 4-8 层叠封无异。该系列产品单盘容量最高可达 122.88TB，在显著提升单盘存储容量的同时，突破性地平衡了容量、能效、性能，为 AI 部署带来了更高效的空间利用率、更高的 GPU 使用率以及更低的总拥有成本，实现存算高效协同。

图 2 NAND Flash 技术路线图

公司	~2021	2022	2023	2024	2025	2026~
SAMSUNG	128L V6(133L V6P) 256Gb/512Gb/1Tb TLC	176L V7, 2-deck, COP 512Gb/1Tb TLC/QLC		236L V8, 2-deck, COP 512Gb/1Tb TLC		286L V9, 2-deck, COP 1Tb TLC/QLC
SK hynix	128L V6, 2-deck 512Gb/1Tb TLC	176L V7, 2-deck, PUC 512Gb/1Tb TLC/QLC		236L V8, 2-deck, PUC 512Gb/1Tb TLC		321L V9, 3-deck, PUC 1Tb TLC/QLC
SOLIDIGM 思得	144L, 3-deck, FG 512Gb/1Tb TLC/QLC			192L, FG 1.77Tb QLC		240L QLC
micron	128L Gen4 512Gb/1Tb TLC	176L Gen5, 2-deck 512Gb/1Tb TLC/QLC		232L Gen6, 2-deck 512Gb/1Tb TLC/QLC		276L Gen9 1Tb/2Tb TLC/QLC
KIOXIA SANDISK 闪迪	112L BiCS5 256Gb/512Gb/1Tb TLC		162L BiCS6, CuA	218L BiCS8, CBA 512Gb/1Tb TLC/QLC		BiCS9/ 332L BiCS10
长江存储 YANGTZE MEMORY	64L X1/X2 256Gb/512Gb/1.33Tb TLC/QLC	128L/232L X3 512Gb/1Tb TLC/QLC		X4 NAND 512Gb/1Tb/2Tb TLC/QLC		X5 NAND 2Tb TLC/QLC

数据来源：公开信息，CFM 闪存市场

总体来看，全球 NAND Flash 技术竞争正从单纯的层数提升，逐渐转向层数、架构与制造工艺协同优化的发展阶段。随着 300 层级 NAND Flash 进入规模化量产，Multi-Deck 堆叠、Plane 并行度提升以及晶圆级混合键合等关键技术将成为推动下一代 NAND Flash 性能与密度提升的重要路径。与此同时，各大原厂在刻蚀工艺、材料工程、晶圆键合以及控制器协同设计等方面持续投入研发，以解决超高层结构带来的制造复杂度和成本挑

战。未来随着 NAND Flash 层数向 400 层甚至更高演进，存储阵列与逻辑电路的三维集成程度将进一步提升，NAND Flash 的单位存储密度、I/O 带宽以及能效水平有望持续突破，为数据中心、人工智能及高性能计算等应用提供更高效的存储基础设施。

二、制程升级、应用驱动与架构创新驱动全球 DRAM 技术演进

1、1c nm DRAM 步入量产与出货阶段，全面拥抱 EUV

近十年来，DRAM 制程工艺沿着 1X nm → 1Y nm → 1Z nm → 1a nm → 1b nm → 1c nm 的路径持续演进。在这一微缩进程中，随着制程节点逼近 10nm 左右的物理极限，传统的 ArF 沉浸式光刻已难以满足精度要求，必须引入复杂的多重曝光技术。然而，多重曝光涉及多次对准，不可避免地带来精度损失和良率下降的问题。为突破这一瓶颈，三星、SK 海力士分别在 1Z nm、1a nm 引入更加先进的 EUV 光刻技术，而美光直至在 1γ nm（即 1c nm）才开始采用 EUV 技术。通过 EUV 更精准的光刻，可以在相同面积内实现更优化的电路布局，间接为电容留出更多宝贵的体积空间，或为晶体管设计更有效的结构，从而在一定程度上缓解电容值下降和漏电控制的压力。

图 3 DRAM 技术路线图

公司	~2021	2022	2023	2024	2025	2026~
SAMSUNG	1a nm EUV 2021MP		1b nm EUV 2023MP		1c nm EUV 2026~	
SK hynix	1a nm EUV 2021MP			1b nm EUV 2024MP	1c nm EUV 2025MP	
micron	1a nm 2021MP		1β nm 2022MP		1γ nm EUV 2025MP	
CXMT	G1 2020MP	G3 2022MP		G4 2024MP	G5 2026~	
NANYA	20nm DRAM	1A nm 2022		1B nm 2024	1C nm 2026~	
winbond	25nm 2020		20nm 2023		16nm 2025	

数据来源：CFM 闪存市场

当前世界上最先进的 DRAM 技术为 1c nm DRAM，不过原厂产能主要仍集中在 1a/1b nm 中，且此两种制程覆盖的产品组合更为全面。具体到各产品线，1a nm 主要用于 DDR4、LPDDR4X 等传统 DRAM 产品，同时也承接部分 96GB 及以下容量的 DDR5 与 LPDDR5X 生产，三星还通过该制程生产 HBM3/3E；原厂普遍采用 1b nm 生产 128GB 及以下容量 DDR5、LPDDR5X 以外，SK 海力士、美光还用于生产 HBM3/3E、HBM4。而最新一代制程 1c nm 技术布局也呈现出差异化，三星率先采用该节点启动 HBM4 量产，SK 海力士和美光则侧重于将其应用于高频 LPDDR5X 及大容量 DDR5 产品。

图 4 DRAM 原厂不同制程的产品分布

数据来源：CFM 闪存市场

2、CPU 多核化催生高频 128GB、256GB 等大容量服务器 DDR5 内存条应用需求

随着摩尔定律放缓，单纯提升单核频率遭遇了功耗墙和散热瓶颈，导致继续增加时钟速度变得极其低效且困难。为了在可控的功耗下持续提升性能，CPU 厂商纷纷通过集成更多核心来同时处理多个任务线程。此外，现代应用场景如高清视频渲染、大型 3D 游戏、人工智能训练及多任务处理，对并行计算能力提出了极高要求，这进一步推动了 CPU 核心数跃升至数百个核心数。

经过历代的技术升级，以英特尔、AMD 为代表的服务器 CPU 平台，其核心数从 28 核跃升至最高 288 核。.CPU 核心数和线程数的爆发式增长，也对内存容量、数据传输速率提出了前所未有的要求。随着英特尔 Granite Rapids、Sierra Forest 以及 AMD Turin 等服务器 CPU 平台在大型云服务商的加速导入和验证，进一步拉动高频 24Gb、32Gb DDR5 应用需求增加，96GB、128GB 甚至 256GB 等高容量服务器 DDR5 内存条应用快速增长。

表 1 历代英特尔、AMD 服务器 CPU 核心数、内存类别

	代号	最高核心数	支持的内存类型	通道数	发布时间
intel	Skylake	28（P-Core）	DDR4-2666	6	2017
	Cascade Lake	28（P-Core）	DDR4-2933	6	2019
	Ice Lake	40（P-Core）	DDR4-3200	8	2021
	Sapphire Rapids	60（P-Core）	DDR5-4800	8	2023
	Emerald Rapids	64（P-Core）	DDR5-5600	8	2023
	Granite Rapids	128（P-Core）	DDR5-6400	12	2024
	Sierra Forest	144（E-Core）	DDR5-6400	12	2024
	Diamond Rapids	192（P-Core）	DDR5-7200	16	2026
	Clearwater Forest	288（E-Core）	DDR5-8000	12	2026
AMD	Naples	32（P-Core）	DDR4-2666	8	2017
	Rome	64（P-Core）	DDR4-3200	8	2019
	Milan	64（P-Core）	DDR4-3200	8	2021
	Genoa	96（P-Core）	DDR5-4800	12	2022
	Bergamo	128（E-Core）	DDR5-4800	12	2023
	Turin（Zen5）	128（P-Core）	DDR5-6400	12	2024
	Turin（Zen5c）	192（E-Core）			
	Venice（Zen6）	192（P-Core）	DDR5-8000	16	2026
	Venice（Zen6c）	256（E-Core）			

数据来源：CFM 闪存市场

3、未来 DRAM 技术将向 $6F^2$+High-NA EUV、$4F^2$+VG/VCT、3D DRAM 等方向发展

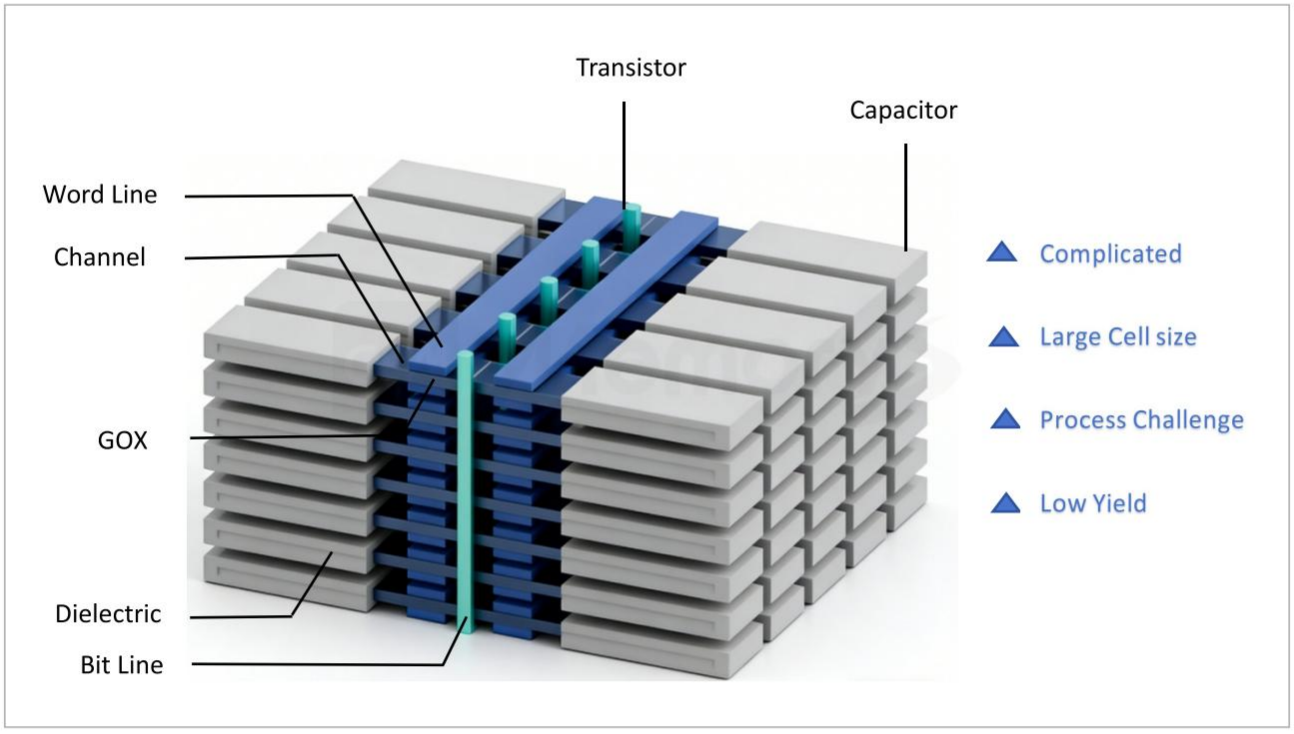
尽管当前原厂 DRAM 技术仍主要采用 $6F^2$ 单元结构，但由于制程持续微缩 10nm 以下，继续沿用平面结构的 $6F^2$ 变得越来越困难，随着单元面积缩小，电容器的体积被迫压缩，易出现电容值大幅下降、漏电流控制难题，对刻蚀和沉积工艺是巨大考验。为应对 10nm 以下极其精密的图形化挑战，确保良率，未来 DRAM 原厂将更广泛地应用高数值孔径极紫外光刻（High-NA EUV）来维持密度和性能的提升。

与此同时，三星、SK 海力等原厂正在探索 $4F^2$ 垂直结构。其中，三星 $4F^2$ VCT 通过改变晶体管沟道的方向，将传统平面结构中的水平沟道变为垂直竖立，形成垂直的电流通道，同时积极开发新型氧化物沟道材料（如

IGZO），以解决垂直沟道的漏电和性能问题。SK 海力士 4F² VG 则通过改变栅极的放置方向。将传统平面结构中的水平栅极变为垂直放置，并被沟道环绕，并且明确将采用混合键合 (Wafer Bonding) 技术来优化芯片结构，以提升单元效率和电气特性。尽管上述 4F² 结构确实大幅提升存储密度，性能与功耗也更优，不过该方案设计与集成复杂度更高，制造工艺难度更大。

值得注意的是，若 DRAM 制程微缩从 1c nm 向 1d、1e、0a、0b、0c 演进，未来走上 3D DRAM 这一技术路径将在所难免。NEO Semiconductor 采用类似 3D NAND 的制造工艺，基于 1T1C 和 3T0C 的 3D X-DRAM 单元并使用 IGZO 材料，旨在为最苛刻的数据应用提供前所未有的密度、功率效率和可扩展性，通过垂直堆栈，其层数可达数百层，DRAM 单 die 高达 128Gb-512Gb。铠侠则开发适用于 3D DRAM 的氧化物半导体通道晶体管（IGZO），采用环绕式栅极（GAA）结构。

图 5 3D DRAM



数据来源: NEO Semiconductor

4、新型存储架构：Groq LPU + SRAM 凭借低延迟、高带宽打破“内存墙”瓶颈

LPU 全称 Language Processing Unit（语言处理单元）由美国 Groq 公司首创，是针对大语言模型推理场景的专用架构，它直击了传统 GPU 在 AI 推理环节的延迟高、成本高、能效比不足等核心痛点。但 LPU 并非替代 GPU，两者之间在推理阶段可形成一种“分层计算”的协同模式（Prefill 由 GPU 负责、Decode 由 LPU 负责）。

2025 年 12 月，英伟达与 Groq 达成非独家许可协议，斥资 200 亿美元获其技术授权，同时，Groq 创始

人 Jonathan Ross、总裁 Sunny Madra 等核心成员将加入英伟达。据悉，英伟达计划在 2026 年 3 月 15 日的 GTC 大会上发布首款原生 LPU 推理专用平台（LPX 机柜），升级后单机柜可搭载 256 颗 LPU 芯片，同时其新一代 GPU 芯片——Feynman 将全球首发台积电 A16（1.6nm）制程，首次集成 LPU 硬件堆栈，形成 GPU+LPU 的异构架构。

表 2 Groq LPU vs NVIDIA B200 GPU

架构	Groq LPU	NVIDIA B200 GPU
内存类型	片上 SRAM	片外 HBM3e
内存容量	230 MB	192 GB
内存带宽	80 TB/s	8 TB/s
推理速度	Llama 2 70B: 300 token/s	8 卡 DGX B200 系统可为单用户提供 >1000 token/s 的生成速度 (Llama 4 Maverick)
	Llama 2 7B: 750 token/s	
	Llama 3 8B: 1300+ token/s	
功耗	约 185W (单卡)	1000W (单 GPU)
延迟特性	确定性执行（亚毫秒级）	动态调度（存在延迟抖动）
应用	AI 推理（特别是 LLM）	训练与推理

数据来源：CFM 闪存市场

传统 GPU 需频繁从片外的 HBM 读写数据，会受制于有限的带宽和较高的延迟，而 LPU 片上 SRAM 的内存带宽高达 80 TB/s 以上，相比 GPU 片外 HBM 的带宽约 8 TB/s，LPU 的速度提升高达 10 倍，追求极致带宽和低延迟。不过 LPU+SRAM 的主要挑战在于容量小、系统级成本高昂。单颗 LPU 仅 230MB 的 SRAM 无法容纳一个大模型，若要要运行一个百亿参数的大模型，则需要数百颗芯片互联，导致硬件采购成本和功耗大幅上升。

LPU 以片上 SRAM 作为模型权重的核心存储介质，直接在芯片设计层面摆脱对 HBM 的依赖，不再受制于 HBM 产能供应不足的窘境，通过自主完成芯片设计后交由晶圆代工厂进行流片和生产，因此，代工厂的工艺产能也变得尤为重要。尽管片上 SRAM 在单芯片容量上无法与 HBM 匹敌，但 LPU 精准锚定“低延迟优先”的推理场景，以极致带宽与确定性响应重构了推理芯片的存储范式。

第二章 2026 年存储市场展望

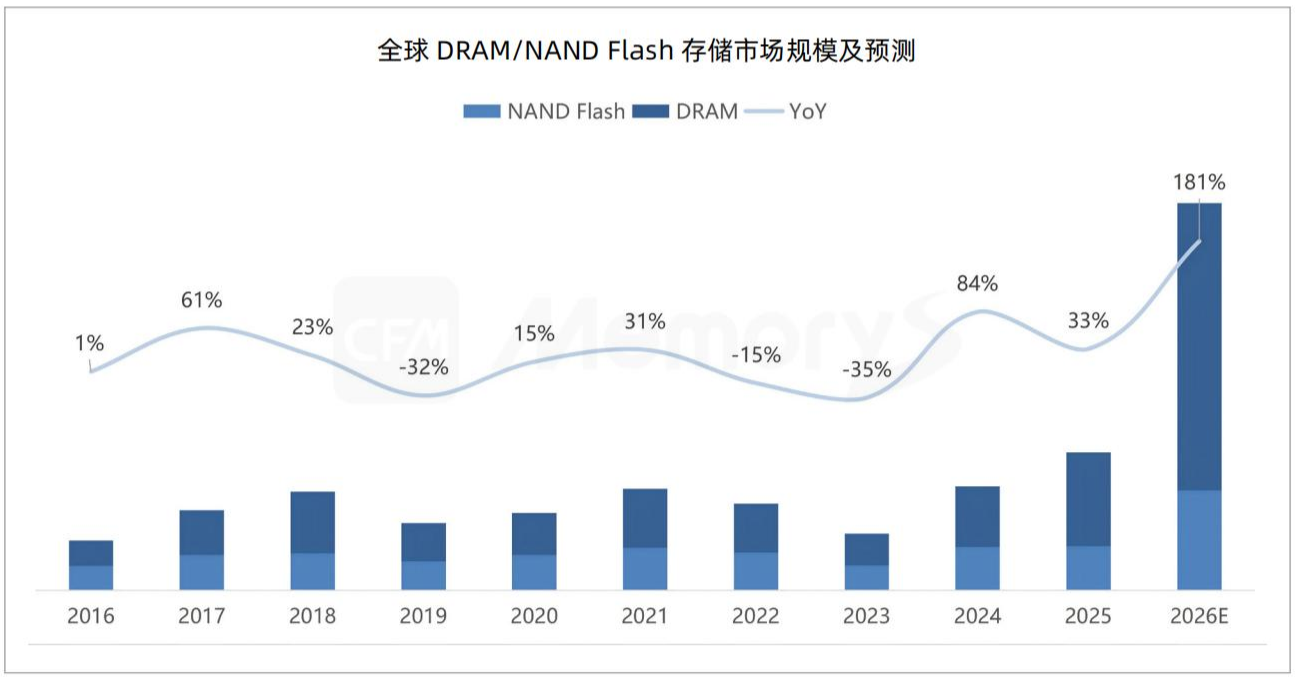
OUTLOOK FOR 2026 MEMORY MARKET

一、存储产业正式进入史诗级的黄金时代

根据 CFM 闪存市场分析与预测显示，2025 年全球 DRAM/NAND Flash 市场规模历史上首次突破 2000 亿，同比增长 33% 至 2216 亿美元；预计到 2026 年这一规模将再度突破 6000 亿美元，全球存储产业进入史诗级的黄金时代。

在 2023 年创下史上最大跌幅后，全球 DRAM/NAND Flash 市场自 2024 年强势反弹，市场规模大幅增长 84%；2025 年在 AI 需求驱动下再度成长 33%，创历史新高。展望 2026 年，尽管以 Mobile、PC 为代表的消费电子需求进入传统周期波动，但 AI 需求与数据中心正迎来结构性扩张，持续拉动全球存储需求的增长。与此同时，全球存储供应增速却滞后于需求端，存储原厂产能难以完全匹配市场需求。多重因素共振下，2026 年存储行业有望实现跨越式增长，预计市场规模将突破历史性的 6000 亿美元。

图 6 全球DRAM/NAND Flash存储市场规模及预测(单位: 亿美元)



数据来源：CFM 闪存市场

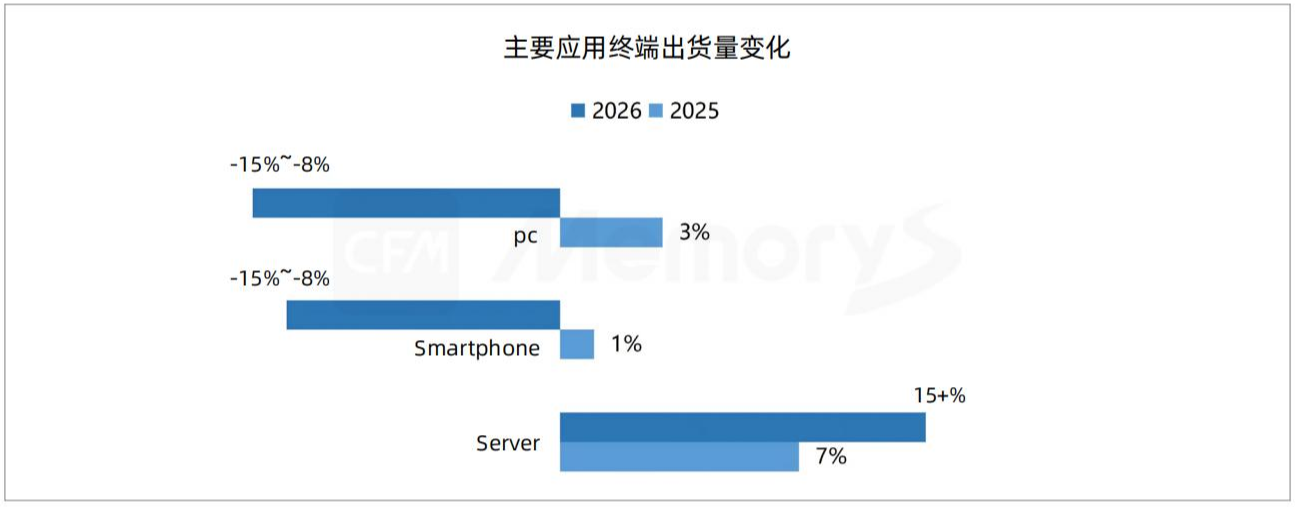
二、AI 正带动存储产业穿越传统周期、释放全新价值

1、2026 年服务器出货量增长，智能手机和 PC 出货量面临下滑

随着 AI-agent 的规模化落地与广泛普及，全球 AI 推理相关工作负载呈现爆发式增长态势，这一趋势不仅直接带动全球 AI 服务器市场需求的激增，同时也对传统通用服务器的应用需求形成显著拉动，推动服务器市场进入新一轮增长周期。CFM 预计 2026 年全球服务器出货量将继续增长至 1610 万台规模。

与此同时，受存储供应紧张、价格上涨影响，手机与 PC 厂商减配降容的成本缓冲空间有限，终端提价已成趋势。高端产品凭借智能化、高端化仍具溢价能力，而低端产品价格敏感度高、提价难度大，受冲击更为显著。CFM 预计 2026 年全球智能手机出货量将下滑 8%-15%，全球 PC 出货量也将下滑 8%-15%。

图 7 2025-2026 年主要应用终端出货量变化



数据来源：CFM 闪存市场

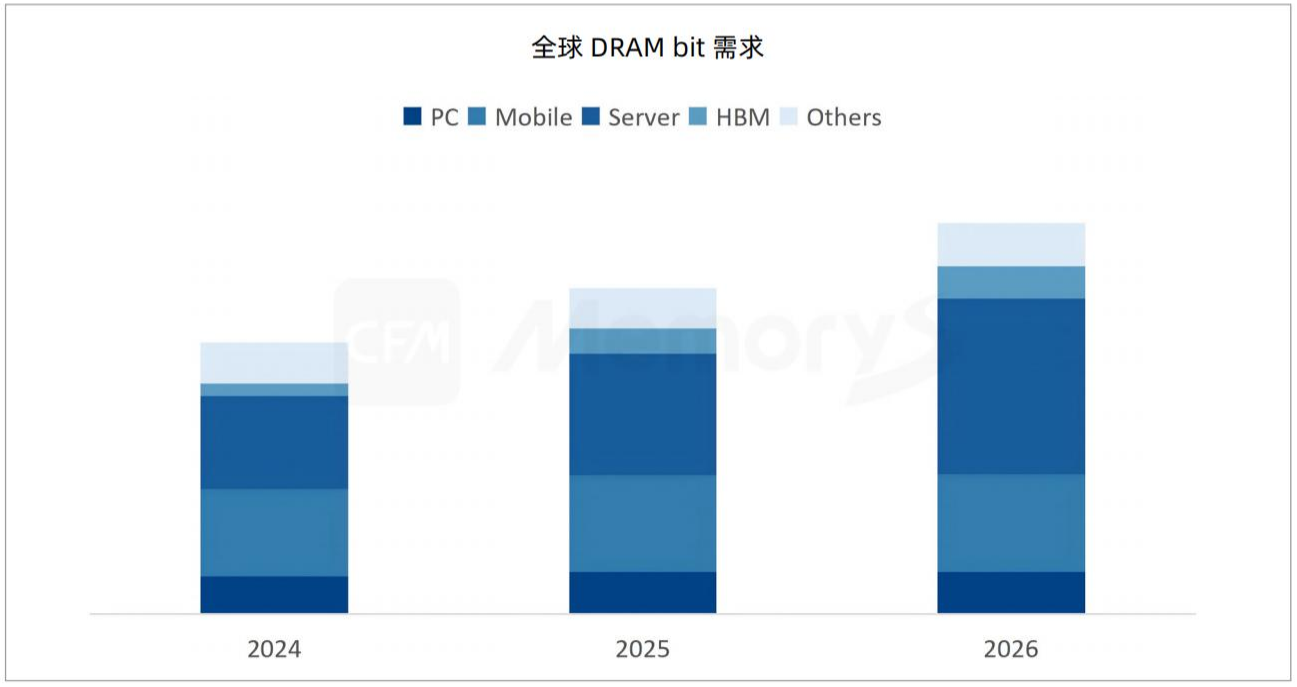
2、AI 推理需求驱动 eSSD 成为 2026 年 NAND 最大应用市场，服务器 DRAM 占比更是超过 50%

尽管以 Mobile、PC 为代表的消费电子需求进入传统周期波动，受 AI 推理需求驱动的超大规模云服务商（Hyperscaler）和大型互联网厂商持续加大服务器采购力度，需求端的强劲复苏直接传导至上游存储芯片领域，带动存储芯片整体采购量出现阶段性激增，成为驱动存储产业增长的核心动力，带动存储产业穿越传统周期、释放全新价值。CFM 预计 2026 年全球 NAND Bit 整体需求同比增长 15%，其中服务器 NAND 需求同比增长逾 60%，服务器 NAND 在所有 NAND 需求中占比约 37%，并首次超过手机成为 NAND 最大应用市场。2026 年全球 DRAM Bit 整体需求同比增长 20%，其中服务器 DRAM Bit 需求同比增长约 45%，服务器 DRAM 含 HBM 在所有 DRAM 需求占比中首次超过 50%。

在 DRAM 领域，市场的结构性增长正由服务器端强势驱动。其中，AI 服务器与高端通用服务器对核心存储组件的需求持续高位运行，导致高密度 DDR5、低功耗 LPDDR5X 及 HBM 普遍呈现供不应求态势。以 LPDDR5X 为例，其应用场景正从智能手机加速向服务器领域拓展，并迎来爆发式增长。英伟达下一代 Rubin 架构将成为关键推手：每颗 Vera CPU 需搭配 8 颗 192GB 的 SOCAMM 模块，单颗 CPU 所需的 LPDDR5X 容量高达 1.5TB，较 Blackwell 平台的 Grace CPU（480GB）激增逾 3 倍。这一趋势深刻揭示了存储需求的结构性变迁。过去，智能手机是 DRAM 消耗的主力；如今，单台 AI 服务器的 DRAM 用量已是传统通用服务器的 2 倍。而训练集群对 HBM 的刚性需求，更使其成为决定算力上限的“战略级资源”。

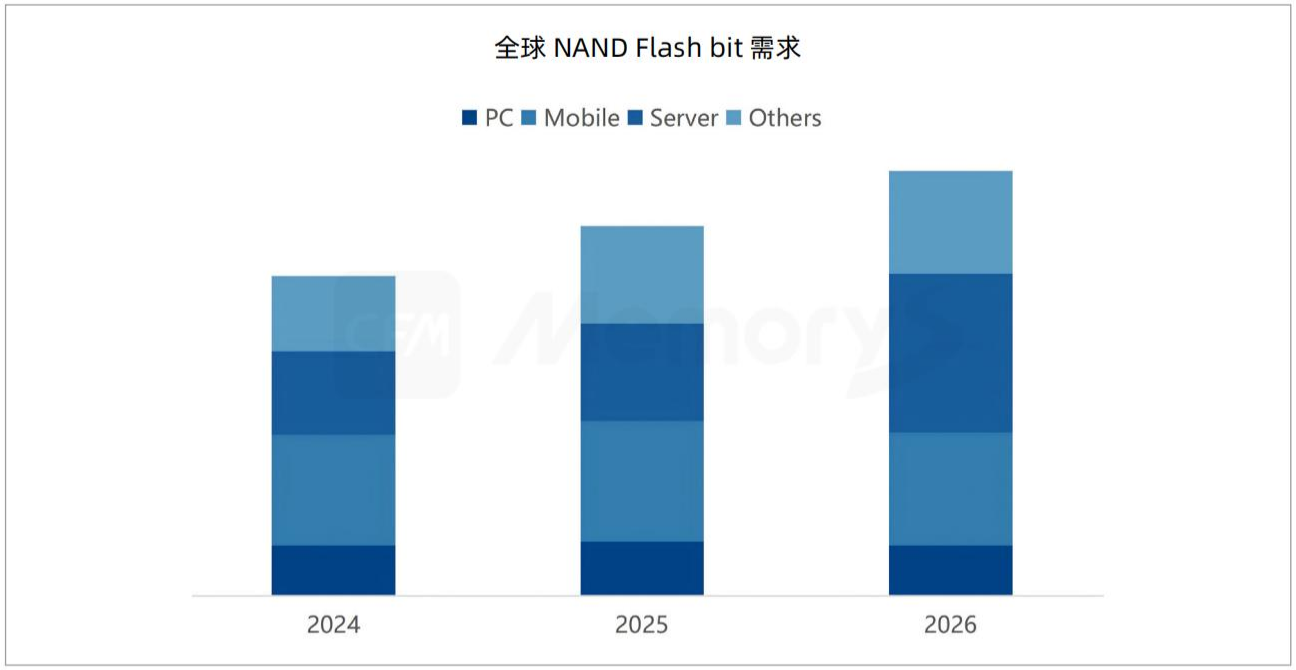
在 NAND 领域也是如此。NAND 市场的增长引擎已完成切换：2026 年，智能手机与 PC 的需求趋于停滞，而服务器端的占比已从过去的 20%+ 飙升至 37%，且仍在扩大。这一结构性转变的背后，是 AI 对存储架构的重塑，针对 AI 推理优化的企业级 SSD 需求正快速攀升。英伟达 Vera Rubin 平台推出的 ICMS 架构，将 KV Cache 从 HBM 分流至 SSD，使单台搭载 72 颗 GPU 的服务器 NAND 用量高达 1.152PB，较传统架构实现指数级跃升。与此同时 NL HDD 市场供应状况持续趋紧，QLC 凭借成本与容量的综合优势，需求呈现稳步上升趋势，进一步拓宽了 NAND 的应用边界。

图 8 全球 DRAM bit 需求（单位：BGb）



数据来源：CFM 闪存市场

图 9 全球 NAND Flash bit 需求（单位：BGB）



数据来源：CFM 闪存市场

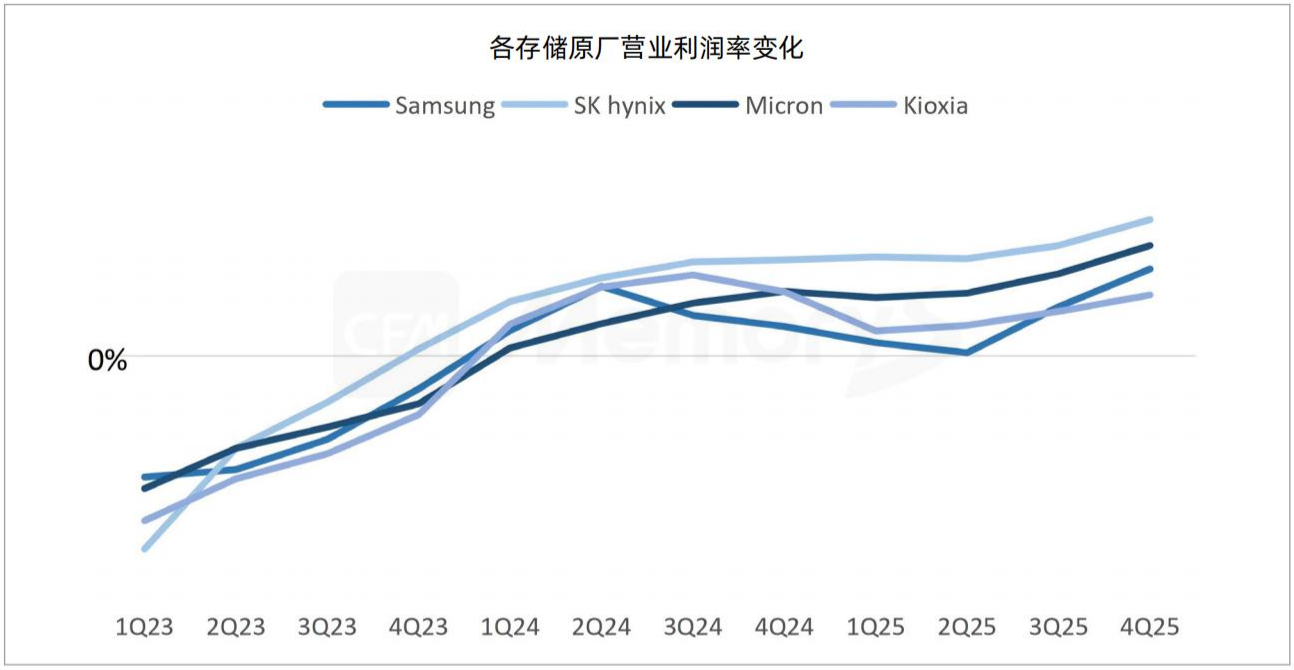
三、稀缺产能重塑定价逻辑：存储市场全面转入“卖方市场”

⊙ 本轮存储产能的稀缺，根源在于理性扩产带来的供应放缓，以及全产业链的库存低位。当井喷的需求骤然来袭，市场已无缓冲地带。在此背景下，涨价不再是简单的市场行为，而成为调节供需、分配稀缺资源的唯一杠杆。

2026 年服务器需求爆发式增长带动全球 DRAM/NAND Flash 需求增长，然而存储原厂产能难以完全匹配市场需求，当前存储行业的供应增速已滞后于下游市场的需求增长节奏，结果导致服务器存储组件的市场缺口持续扩大，甚至制约了服务器整机的出货进度。为应对这一局面，全球主流存储原厂纷纷调整产能分配策略，持续优先保障服务器领域的存储产品出货，这一举措进一步加剧了 Mobile 和 PC 市场的存储供应限制。在此背景下，下游终端厂商为规避因存储组件短缺导致的生产中断风险，积极锁定产能供应，进一步强化了存储市场的供应紧张格局。存储产能已然成为稀缺资源，全行业的严重供应短缺使得供应端定价权大幅提升，存储市场全面转入“卖方市场”，自 2025 年四季度起 DRAM/NAND Flash 价格全面大幅上涨，并预计涨价态势持续 2026 年全年。

全球存储市场正在经历一场深刻的价值转向：未来市场的竞争，将不再是份额的零和博弈，而是如何在高利润区间建立并维持可持续的技术壁垒。随着各存储原厂营业利润率从跌价周期谷底回升至历史高位，维持高利润水平已成为行业共识。经历过巨额亏损的切肤之痛，厂商们在未来竞争中势必更加坚定地聚焦高价值领域，而非重燃价格战。

图 10 各存储原厂营业利润率变化



数据来源：CFM 闪存市场　注：上图三星营业利润率为其 DS 部门（半导体业务）数据

在利润导向的战略下，NAND Flash 行业的增产步伐明显放缓，新工厂建设速度趋于谨慎。展望 2026-2027 年，NAND 行业的供应增速难以重现历史上 30% 以上的高速扩张。与此同时，存储原厂的资本支出正明显向 DRAM 倾斜，以期最大化 DRAM 营收。以 SK 海力士的龙仁工厂与三星的 P5 工厂为例，其新增 DRAM 产能预计分别于 2027 年及 2028 年才能陆续释放。但 DRAM 晶圆产能持续向 HBM、高容量 DDR5、高速 LPDDR5X 等高利润产品倾斜，且 HBM 需求大幅挤占通用 DRAM 产能，使得行业整体 DRAM bit 供给增速受到刚性制约。

存储产业资本开支整体趋于谨慎，叠加 2025 年需求持续消耗，行业库存水位已降至历史低位。全球头部存储厂商均将先进产能优先投向高毛利的 AI 存储产品，成熟制程与消费级存储产能持续被挤压；叠加行业此前去产能、控资本开支的影响，供给端呈现显著结构性错配。在服务器需求持续增长、供应商库存处于低位、行业扩产缓慢的共同作用下，存储市场供需缺口短期内难以缓解，短缺格局已成为行业常态。

表 3 各存储原厂的库存金额(单位: 百万美元)

各存储原厂的库存金额（单位：百万美元）									
厂商	23Q4	24Q1	24Q2	24Q3	24Q4	25Q1	25Q2	25Q3	25Q4
三星	23,441	24,108	23,580	23,276	21246	20,982	20,237	19,555	/
SK 海力士	10,194	10,420	9,740	9,837	9,528	10,038	9,746	9,514	9,863
美光	8,443	8,512	8,875	8,705	8,705	9,007	8,727	8,355	8,205
闪迪	3,216	3,215	3,342	3,384	3,420	2,160	2,079	1,907	1,970

数据来源：CFM 闪存市场　注：上表三星库存为其 DS 部门（半导体业务）数据

AI 需求的爆发与供应紧张，共同开启了存储行业的黄金时代。尽管繁荣之下，智能手机、PC 等传统消费市场正经历着转型阵痛，但这更像是黎明前的必要调整。随着 AI 推理加速向端侧下沉，AI 将以更具想象力的形式重塑每个人的生活。终端应用的创新浪潮，必将对存储芯片提出更多元的要求——这既是黄金时代的下一程，也是存储产业真正的星辰大海。

第三章 AI 时代的存储新需求

NEW MEMORY DEMAND IN AI ERA

一、AI 大模型迈入万亿级时代，重塑 AI 存储新范式

1、打破算力配套定位，存储从 AI 支撑性资源向核心基础设施跨越

AI 大模型的迭代、多模态化和算力跃升正以指数级速度推动存储需求攀升，并对存储从容量、带宽、性能到架构全面提出升级需求，驱动存储行业进入超级周期。

在当前技术条件下，尽管模型架构与量化技术不断优化，大模型的参数量与存储需求仍保持显著正相关。除模型权重静态占用外,运行时的KV缓存和激活值也消耗大量显存; 训练阶段还需存储梯度及优化器状态等信息。AI 大模型的发展遵循缩放定律（Scaling Laws），即在追求最优性能的过程中，模型参数量与训练数据量需同步成比例增长。从国内外主流大模型近年发展情况来看，参数量已从亿级跃升至万亿级，训练数据量实现百倍级增长，上下文窗口则由千级 tokens 扩展至百万级别。伴随大模型持续迭代，上述指标的指数级增长正加速转化为对存储系统容量和性能的需求提升。国内外主流大模型具体参数如下表所示。

表 4 国内外主流大模型具体参数

Google	模型名称	Gemini1.0	Gemini 2.0 Flash	Gemini3.1 PRO
	参数数量	Nano-1：18 亿 Nano-2：32.5 亿	/	激活参数约 1.2 万亿
	上下文窗口	32K tokens	100-tokens	100 万 tokens
OpenAI	模型名称	GPT 1	GPT 3	GPT-5.2
	参数数量	1.17 亿	1750 亿	8000 亿
	训练数据规模	4-5 GB	570GB	/
	上下文窗口	2K tokens	2K tokens	400K tokens
Anthropic	模型名称	Claude 2	Claude 3 Haiku	Claude Opus 4.6
	参数数量	1300 亿	/	7000 亿
	上下文窗口	100K tokens	100K tokens	100 万 tokens（beta 版）
百度	模型名称	文心 1.0	文心 3.0	文心 4.5
	参数数量	3.4 亿	100 亿	4240 亿
	上下文窗口	2K tokens	2K tokens	128K tokens
阿里巴巴	模型名称	Qwen 2.0	Qwen 3-Max-Thinking	Qwen 3.5
	参数数量	72B	1T	122B
	上下文窗口	128K tokens	1M tokens	262K tokens
深度求索	模型名称	DeepSeek LLM	DeepSeek-V3	DeepSeek V4
	参数数量	67B	671B	/
	上下文窗口	4096 tokens	128K tokens	100 万 tokens

数据来源：公开信息

随着 AI 大模型从单一文本处理向文本、图像、音频、视频等多模态融合演进，异构数据爆发式增长，数据体积与复杂度均实现量级提升。为了匹配高异构、高并发、高算力密度下的数据供给效率，多模态模型对存储提出远超单文本模型的更高要求。一方面，原始数据集从文本时代的 TB 级跃升至多形态数据的 PB 级，单样本体积由 KB 增至 MB 甚至 GB 级，导致推理过程中需缓存大量中间特征与键值（KV），使得存储容量需求呈 EB 级弹性增长，并要求其采用分级存储架构。另一方面，多模态模型以图像、视频、音频等高分辨率异构数据为核心，需完成跨模态对齐、拼接与同步计算。这要求存储系统需要在面临大量 GPU 集群并行调度时，提供 TB/s 级的聚合带宽以持续供给数据，具备千万级 IOPS 以应对海量小文件及碎片化数据的并发访问，并保持毫秒级甚至微秒级的低延迟，从而消除数据瓶颈，支撑多模态模型的高效训练与低延迟推理。

近年来，AI 算力呈指数级提升，驱动存储容量和带宽同步增长。从英伟达产品看，Blackwell 引入了第五代张量核心，支持 4 位和 6 位运算的超低精度模式，AI 算力显著提升。GB300 单芯片的 FP4 算力为 15 PFLOPS，而 GB200 为 9 PFLOPS，提升 66.7%；GB300 的容量为 288GB，GB200 提高 50%。从 AMD 产品看，M1355X 峰值算力为 10PFLOPS 较 M1300X 几乎翻倍，容量增长 50%。每一代算力升级均需更大容量、更高带宽的存储支撑，以满足大模型训练与推理中海量数据的高速读写，避免算力受限。同时，FP4、FP6 等超低精度运算的普及，亦要求存储具备细粒度数据调度、高效压缩解压缩及动态精度适配能力。随着算力密度与功耗快速上升，存储加速向高带宽、低延迟、高密度、低功耗方向演进。现阶段，存储已不再是算力的配套资源，而是决定 AI 系统整体性能与上限的核心基础设施。

表 5 英伟达 AI 加速卡的性能数据

架构	型号	FP4 算力	FP8 算力	容量
Hopper	H100	不支持	3.958 PFLOPS	80GB
	H200	不支持	3.958 PFLOPS	141GB
	GH200	不支持	3.958 PFLOPS	96GB/144GB
Blackwell	B200	9 PFLOPS	4.5 PFLOPS	180GB
	B300	14 PFLOPS	4.5 PFLOPS	270GB
	GB200	9 PFLOPS	4.5 PFLOPS	192GB
	GB300	15 PFLOPS	5 PFLOPS	288GB

数据来源：公开信息

表 6 AMD AI 加速卡的性能数据

架构	型号	FP4 算力（峰值）	FP8 算力（峰值）	容量
CDNA 3	MI300X	不支持	5.2 PFLOPS	192GB
CDNA 4	MI350X	18.4 PFLOPS	9.2 PFLOPS	288GB
CDNA 4	MI355X	20 PFLOPS	10 PFLOPS	288GB

数据来源：公开信息

2、AI 大模型全流程驱动：存储从被动承载到主动赋能的需求重构

随着模型规模迅速扩展至万亿级别，在 AI 大模型的全流程运行中，存储的角色已发生深刻转型，从原来存放数据的被动仓库演变成为能够主动提升运算效率的关键加速环节。AI 大模型各阶段对存储提出的新需求如下：

数据输入阶段

将各类原始数据从不同来源接入系统，为后续处理提供“原材料”。当前大模型呈现“多模态、超海量”的特点，此阶段数据来源多元、格式异构，数据量常达 PB 级且接入频率不稳定，多为访问频率较低但需完整留存的原始数据。因此，要求存储支持 PB 级甚至 EB 级海量数据的高效写入与持久化留存，兼容多种存储协议以适配多源数据接入，并基于较低成本保障数据的高可靠性和高吞吐导入能力。

数据准备阶段

对原始数据进行多轮清洗、标注等处理以生成高质量数据集。此阶段存在偶尔读取的原始数据、大量高频读写的中间数据和需长期留存的最终数据，I/O 模式复杂，访问频率差异较大；主要通过采用分布式多节点并行处理保证数据一致性。因此，要求存储具备高 IOPS 与吞吐量以降低读写延迟、适配并行场景，并区分不同类型数据的存储需求，通过分层存储平衡性能和成本。

模型训练开发阶段

基于高质量数据通过分布式训练、参数调优、模型验证等过程，生成具备特定能力的大模型。训练开发过程需频繁读写训练数据、模型参数等，多节点并发访问量极高，且冷数据访问频率显著提升。因此，存储需支持 TB/s 级高带宽传输以满足大文件顺序读需求，具备毫秒级甚至微秒级低延迟与万级以上高并发访问，拥有高可靠性与容错能力以保障长时间训练不中断，并提升冷数据访问速度。

模型推理部署阶段

将训练好的模型部署到实际场景中，接收用户请求并输出结果。此阶段数据访问呈现高频次、低批量、低延迟特点，推理场景多样且存在大量临时推理数据。因此，存储需实现微秒级超低延迟读写以保障响应速度，具备高吞吐并发处理与动态调整能力，支持临时数据的快速写入与自动清理，并具备灵活缓存优化能力以支撑核心技术应用。

归档阶段

长期留存 AI 大模型全流程运作的各类数据。此阶段数据类型多样、访问频率极低但需长期留存，且安全性与可追溯性要求高，部分数据可能被重新调用。因此，要求采用低成本大容量介质支持 PB 级数据长期留存，具备极高可靠性与数据持久化能力，支持数据分类归档、标签管理以实现可追溯与可检索，同时具备低功耗、易维护特性和冷数据快速唤醒能力。

3、跨越存储层级鸿沟，AI 推理场景下不同技术的存储需求解析

检索增强生成（RAG）广泛应用于模型推理部署阶段，主要通过检索外部知识库辅助大模型生成更精准结果，需要卓越的存储效能支持对大型非结构化资料集的快速、无缝存取，实现“检索效率、多源数据协同、数据实时更新”。这既需要为向量索引提供极低延迟的查询性能，又需为原始文档块提供高吞吐量的读取能力，从而对存储系统在高 IOPS、低延迟、高耐用性及分层存储管理方面提出了综合要求。

键值缓存（KV Cache）作为大模型推理阶段的核心优化技术，通过缓存解码过程中的键值状态避免重复计算，提升推理效率。随着上下文窗口向百万 token 级别扩展以及并发请求数量的增加，KV 缓存容量需求成比例增长，要求存储系统支持高容量缓存和弹性扩容。同时，KV 缓存本质上是会话力度的临时状态，在长对话或提示词缓存等场景下需要对其进行持久化以复用，对存储提出可靠性和一致性的要求，保证特定会话上下文中缓存数据的准确恢复。

NV ICMS 通过在本地 SSD（G3）与传统共享式网络存储之间（G4）之间新增一个专为 KV 缓存优化的以太网连接闪存层级（G3.5），将存储层级扩展到 HBM/DRAM/ 本地 SSD 之外。KV 缓存既需要存储级的容量，又需要内存级的响应速度。因此，要求存储系统在有限机柜空间内实现高密度和高能效，提供接近内存级别且可预测的低延迟和智能缓存调度，并在持续负载情况下保证稳定性能。此外，为适配异构存储架构，需具备异构存储协同能力，支持不同存储介质间的数据智能迁移。

二、AI 数据中心驱动存储升级，替代窗口加速与前沿技术破局

1、HDD 供应格局受限，QLC SSD 加速替代“黄金窗口”

HDD 厂商扩产意愿较低且扩容产品量产仍需时间，难以满足 CSP 日益增长的需求。全球云服务商加大投入 AI 专用数据中心，叠加 AI 推理应用的规模化落地，使得云存储容量需求显著增长，HDD 需求随之激增。希捷科技和西部数据两家贡献全球超 80% 的 HDD 产能，HDD 供应商高度集中。同时，受上一轮库存伤害影响，HDD 供应商生产模式趋于保守，经营模式已转为“依订单生产”。即使在当前需求高涨的情况下，主流厂商仍没有明确的扩产计划，HDD 交付周期已延长至两年。为确保数据中心的正常运行，CSP 正加速寻求替代方案。

表 7 HDD 厂商的最新动态

厂商	关于 HDD 的动态
希捷科技	◦产能情况：2026 年 Nearline 硬盘产能已全部售罄，并开始接受 2027 年上半年的订单。 ◦扩产意愿：坚持“按订单生产”的策略，未来将通过单碟容量而非增加单位产量来满足 EB 级增长的需求。 ◦产品进展和规划：2025 年初，基于 Mozaic 3+ 平台的 36TB Exos M 发布。目前，Mozaic 3+ HAMR 硬盘已通过全部主流美国云服务提供商客户的认证，第二代 Mozaic 4TB 单碟产品认证工作也正按计划稳步推进，预计将于下个十年初实现单碟 10TB 的存储密度，推出 100TB 的机械硬盘。

厂商	关于 HDD 的动态
西部数据	◦ 产能情况：2026 年产能已全部售罄，并且已与前七大客户中三家签署长约至 2027-2028 年。 ◦ 扩产意愿：未提出目前的扩产计划，表示将持续聚焦于提升硬盘的面密度。 ◦ 产品进展和规划：基于 ePMR 技术的 UltraSMR 硬盘（容量达当前市面最高 40TB）和 HAMR 硬盘均正在接受两家超大规模客户的认证，分别计划于 2026 年下半年和 2027 年开始量产。西部数据预计将扩展 ePMR 产品的容量至 60TB，并在 2029 年实现 HAMR 产品扩容至 100TB。

数据来源：公开信息

基于容量和性能等优势，QLC SSD 成为客户替代 HDD 的首选方案。随着 AI 产业重心向推理应用迁移且 AI 推理服务器侧重于读取操作，在 HDD 供应短缺之际，CSP 纷纷转向采购 QLC SSD。QLC eSSD 最大容量可达 245.76TB，远超最大容量 40TB 的企业级 HDD，且 HDD 物理尺寸更大。具体性能方面，以 30TB 左右的 QLC eSSD 和 HDD 为例，QLC SSD 顺序读写速度分别可达 12500 MB/s 和 2000 MB/s，远超 HDD 的 250MB/s-290MB/s；随机读写性能更以 170 万 IOPS 和 11 万 IOPS 的绝对优势远超 HDD；QLC eSSD 微秒级的访问延迟表现显著优于 HDD，结合其 PCIe 5.0 接口的先进架构，QLC SSD 在高并发、低延迟及高吞吐等企业级读取密集型或混合型关键应用场景中具备显著优势。

表 8 QLC eSSD 和企业级 HDD 具体规格对比

对比维度	QLC eSSD	企业级 HDD
容量	30.72TB	30TB
接口	PCIe 5.0 x4	SATA 6Gb/s
顺序读取速度	可达 12500MB/s	290MB/s 左右
顺序写入速度	可达 2000MB/s	250MB/s-290MB/s
随机读取速度	170 万 IOPS	170 （4K QD16）
随机写入速度	11 万 IOPS	350（4K QD16）
访问延迟	微秒级	毫秒级

数据来源：公开信息

2、存储技术迭代赋能，助力 AI 训练效能提升

Gen5 SSD 在性能、效率和扩展性的全方位升级，大幅缩短 AI 训练耗时。高性能 SSD 测试结果显示：Gen5 SSD 接口速率较 Gen4 SSD 翻倍至 128GT/s，平均读取延迟降低 50% 以上，推动加速器利用率、训练样本处理吞吐率提升超 10%；单盘配置下，Gen5 SSD 模型训练耗时较 Gen4 SSD 缩短 40% 以上；18 个加速器配置下，Gen5 SSD 较 Gen4 SSD 训练总耗时降低超 10%，年省训练约 40 天以上；Gen5 SSD 随磁盘数量增加

性能持续提升，8 块磁盘配置下 DLRMv2 加速比超 1 倍，而 Gen4 受限于接口带宽，性能提升边际效应更明显。Gen5 SSD 解决了 Gen4 SSD 在 AI 大模型训练中的存储瓶颈，实现了存储性能与 GPU 计算性能的匹配，充分释放计算资源的利用率，最终在模型训练的速度、效率、规模化部署上实现全方位提升。

高容量 DDR5 为处理器性能释放提供核心支撑，提高 AI 训练效率。AI 爆发式的发展对算力提出了海量需求，直接驱动厂商研究发布性能更高的新处理器，而这些处理器又为 AI 应用的普及和效率提升提供了关键的硬件基础。随着处理器的不断升级，最大核心数翻倍增长以及内存通道数的增加使得 CPU 单通道内存容量显著提高（如英特尔从 Emerald Rapids 的 64GB 翻倍增长至 Clearwater Forest 的 192GB，AMD 从 Genoa 的 64GB 增长至 Venice 的 128GB）。高带宽、大容量的 DDR5 内存以高带宽保障数据能高速供给处理器，避免算力因等待数据而闲置，从而缩短训练周期；以大容量支持更大规模的模型参数或数据集直接加载到内存中，减少与硬盘等低速存储的频繁数据交换，提升训练效率。

表 9 主流处理器的具体参数

厂商	处理器	最大核心数	内存通道数	单通道内存容量
英特尔	Emerald Rapids	64	8	64
	Granite Rapids	86	8	86
	Clearwater Forest	288	12	192
	Diamond Rapids	192	16	192
AMD	Genoa	96	12	64
	Bergamo	128	12	85
	Turin	192	12	128
	Venice	256	16	128

数据来源：公开信息，CFM 闪存市场

SOCAMM(LPDDR5X) 以高带宽、低功耗、小体积的优势提升 AI 服务器性能与能效。随着推理类任务负载持续增长，AI 服务器转向持续性在线运行模式，内存能效已成为决定机架级运营成本的核心因素。相较于 DDR5 RDIMM，SOCAMM（LPDDR5X）可提供超 2.5 倍的带宽，功耗和规格尺寸可做到为标准 DDR5 RDIMM 的三分之一，为 AI 基础设施提供了更优的 TCO 与性能效率。从技术迭代看，当前开发重点已经转移到 SOCAMM 2。SOCAMM2 相较于第一代在相同尺寸下实现容量提升，新的容量将实时推理工作负载的首 Token 时延（TTFT）缩短 80% 以上；其数据传输速率从 8533MT/S 提升至 9600 MT/s，在同架构下升级应用可使得带宽提高至 16TB/s，显著增强 AI 工作负载的数据吞吐能力；能效提升可实现超过 20%，进一步优化大型数据中心集群的电源设计以节省更多成本。

HBM“向更高代际、更大容量、更高带宽”的技术演进趋势，提高 AI 服务性能。对比分析主要厂商 AI 加速卡规格的具体表现：技术代际正从当前主流的 HBM3E 加速向下一代 HBM4 迁移，HBM4 将于 2026 年在多款旗舰 AI 加速卡上实现部署；相较于 HBM3E，HBM4 通过将堆叠层数从 12 层增至 16 层，实现单栈容量由 24GB-36GB 提升至 36-48GB；8 堆栈配置下，单卡总容量将从 288GB 提升至 384GB。性能方面，HBM4 接口位宽较 HBM3E 翻倍达 2048 位，传输速度高达 8Gb/s，总带宽跃升至 2TB/s，相比 HBM3E 的 1.2TB/s 提升接近 70%。同时，HBM4 基于低电压硅通孔（TSV）技术和电源分配网络（PDN）优化，功耗效率提升 40%，热阻降低 10%，散热能力提升 30%。HBM 的技术迭代通过容量倍增支持更大模型参数与训练批次，减少通信开销；同时，带宽跃升有效缓解“内存墙”瓶颈，两者协同推动 AI 训练与推理的整体效能提升。以英伟达为例，搭载 288GB HBM4 的 Rubin GPU，其推理与训练性能较采用 HBM3E 的 Blackwell 分别提升 5 倍和 3.5 倍。

表 10 搭载不同 HBM 技术的 AI 加速卡具体规格

厂商	型号	Type of（x）PU	技术	容量（GB）	单堆栈容量（GB）	颗粒密度（Gb）	堆叠高度（# 颗粒）	堆栈数量	发布时间
英伟达	R100	GPU	HBM4	288	36	24	12	8	2026Q1
	B300/GB300	GPU	HBM3E	288	36	24	12	8	2025Q3
	B200/GB200	GPU	HBM3E	192	24	24	8	8	2025Q1
AMD	M1400	GPU	HBM4	384	48	24	16	8	2026Q1
	M1355	GPU	HBM3E	288	36	24	12	8	2025Q4
谷歌	TPU v7e	AI ASIC	HBM4	216	36	24	12	6	2026Q3
	TPU v7p	AI ASIC	HBM4	288	36	24	12	8	2026Q1
	TPU v6p	AI ASIC	HBM3E	192	24	24	8	8	2025Q2
AWS	Trainium 3	AI ASIC	HBM3E	144	36	24	12	4	2025Q4
	Trainium 2.5	AI ASIC	HBM3E	144	36	24	12	4	2025Q2

数据来源：公开信息

3、HBF 借力 HBM 技术红利快速起步，有望在 2028-2030 年爆发

（1）AI 发展、技术演进和成本控制共同驱动 HBF

随着大模型的参数规模持续增长，AI 工作负载对内存容量的需求已远超以往。在处理百万 token 级别的长上下文时，仅靠 HBM 容纳庞大的键值缓存（KV cache）需要部署数十个 GPU 来扩容，不仅大幅增加 HBM 的采购成本，也使得系统功耗和互联复杂性急剧攀升。HBM 的产能限制和有限的单设备容量已成为制约 AI 发展的瓶颈。现阶段，HBM 发展主要面临两大挑战：一是容量不足且扩容受限。即使是 HBM4 单颗可达最高的容量

64GB，仍难以有效满足万亿参数模型实时处理 TB 级数据的需求。同时，受限于 DRAM 的物理特性，当前 HBM 层数堆叠已逼近技术临界点，继续堆叠扩容将导致封装复杂度、良率与功耗问题加剧。二是其成本显著高于传统 DDR 内存，复杂的 3D 堆叠与硅通孔技术推高制造成本，继续堆叠扩容将加剧成本增长。

为了破解 HBM 速度极快但容量见顶的根本性矛盾，新的技术路线 HBF 应运而生。HBF 复用 HBM 成熟的封装和架构技术积累，将存储介质从 DRAM 换成 NAND 闪存，显著缩短研发周期。在占用相同空间的情况下，NAND 容量可达 DRAM 的 10 倍，使得 HBF 能在保持高带宽的同时大幅提高容量，容量密度优势显著。此外，在提供相同容量的情况下，NAND 闪存的成本远低于 DRAM，相对成熟的堆叠工艺和高良率使得 HBF 具备一定成本优势。HBF 的综合优势推动其进入快速发展期。

（2）HBF 技术拆解

NAND 闪存的架构创新，低成本实现高带宽和大容量

HBF 结合 HBM 的先进封装、互连技术与 3D NAND 闪存，实现架构创新。一方面，HBF 采用 HBM 设计理念，高密度垂直堆叠多颗 NAND 芯片以缩短传输路径、实现高带宽互联，搭配并行子阵列架构，将闪存划分为独立子阵列使得各子阵列有独立的读写通道，突破传统 NAND 的同时读写数据通道限制。另一方面，HBF 在结构上采用“堆叠之上的堆叠”方案，芯片内通过 3D NAND 堆叠数百层单元，而模块间则通过 TSV 技术垂直堆叠多颗 NAND 芯片并基于 CBA 技术将逻辑芯片直接键合在 3D NAND 存储阵列下方。通过底层的逻辑芯片和密集的 I/O 引脚，HBF 可以与 GPU 建立超宽的数据通道，且通道的宽度（位数）远超传统 NAND，使得每次传输的数据量呈指数级增长。

HBF 以较低成本同时实现高带宽和大容量，适配 AI 推理“读多写少”的存储需求。具体而言，容量方面，HBF 实现单堆栈容量为 HBM 的 8 倍以上，而传统 NAND 虽整盘容量大，但作为独立芯片的存取路径较长。带宽方面，HBF 目标读取带宽虽略低于 HBM4，但与 HBM3E 处于同一梯队；而传统 NAND 受限于接口协议，带宽存在明显差距。延迟方面，HBF 通过 3D 堆叠缩短物理距离，可实现亚微秒级读取延迟，虽不及 HBM，但远优于传统 SSD。成本方面，由于 NAND 成本显著低于 DRAM，HBF 单位容量成本远低于 HBM，更高成本的组件和复杂工艺推动其成本高于传统 NAND 产品。应用场景方面，高带宽、大容量、低成本的优势使得 HBF 成为以大量读取为主的 AI 推理场景的高性能存储解决方案；HBM 低延迟强写速的优势更适用于实时响应 AI 训练中高频的参数更新与梯度同步；传统 NAND 产品更适用于简单的大容量数据存储。

表 11 传统 NAND、HBM、HBF 的规格差异

对比维度	传统 NAND（以 PCIe 5.0 SSD 为例）	HBM（以 HBM4 为例）	HBF
存储介质	3D NAND 闪存	DRAM	3D NAND 闪存
容量	单颗芯片容量较低，整盘可达数 TB	24GB-36GB，未来最高可达 64GB（单堆栈）	512GB（单堆栈）

对比维度	传统 NAND（以 PCIe 5.0 SSD 为例）	HBM（以 HBM4 为例）	HBF
读取带宽	12GB/s-15GB/s	2TB/s-3.3TB/s	≥ 1638GB/s
访问延迟	微秒级	纳秒级	亚微秒级
单位容量成本	低（基准）	高 （约为传统 NAND 的数十倍）	中 （高于传统 NAND，远低于 HBM）

数据来源：公开信息

HBF 并非直接取代 HBM，而是与 HBM 互补共同解决 AI 存储难题。HBF 与 HBM 虽均采用 3D 堆叠和 TSV 等先进封装技术，但其存储介质与系统定位存在本质差异，决定了二者是互补而非替代关系。HBM 负责处理延迟敏感型的即时运算，HBF 则承接推理、读取密集型的海量数据存储需求。在推理集群中，用 HBF 替代部分 HBM，可以在保证性能的前提下，大幅降低单次推理的硬件成本。因此，HBF 作为大容量扩展层，将与 HBM 协同构建分层存储体系，共同应对 AI 大模型日益增长的带宽与容量需求。

（3）全球存储巨头竞合推进 HBF 进展，预计 2027 年启动商业化进程

HBF 的崛起并非单一厂商的技术突破，而是全球存储巨头集体布局的必然结果。2025 年初，闪迪首次提出 HBF 概念并于 7 月成立了 HBF 技术咨询委员会。三星与 SK 海力士均与闪迪签署谅解备忘录，共同推进 HBF 标准化。2026 年 2 月，闪迪与 SK 海力士联合召开“HBF 规格标准化联盟启动会”，宣布正式启动 HBF 的全球标准化进程，推动 HBF 标准化工作。不同厂商从不同角度验证并展示了 HBF 技术的可行性和潜力。基于各厂商的时间规划，HBF 有望于 2027 年起实现商业化，预计将在 2028-2030 年爆发。

⊙ 具体来看主流厂商的 HBF 进展和规划：

闪迪

通过创建仿真模型，在 4050 亿参数量的 Llama 3.1 模型上，模拟 GPU 搭载 HBM 和 HBF 的性能差异。在不考虑容量差异的情况下，观察推理引擎流程的各阶段推理，发现 HBF 整体性能较 HBM 下降幅度小于 2.2%。闪迪计划于 2026 年下半年交付首批 HBF 内存样品，预计将 2027 年初提供首批采用 HBF 技术的推理设备样品。

SK 海力士

提出新型混合架构 H3，在同一中介层上同步配置 HBM 和 HBF，并直连 GPU，使得只读数据存储在 HBF 中，而动态生成的键值缓存则保存在 HBM 中。仿真测试和模拟场景结果显示，相较于只有 HBM 的配置，HBM、HBF 混合架构配置每瓦性能提升高达 2.69 倍，该系统处理并发查询（批处理大小）的能力提升了高达 18.8 倍。此外，SK 海力士全新的 AIN 产品阵容中，AIN B 系列采用 HBF 高带宽闪存技术，评估与 HBM 协同配置等灵活应用方案。预计 AIN B 的 Alpha 版本将于 2026 年初发布，评估样机将于 2027 年推出，2029-2030 年将聚焦 AIN B 产品。SK 海力士将 2030 年设为 HBF 市场扩张节点。

铠侠

2025 年 8 月成功推出 5TB 大容量、64GB/s 的高带宽闪存模块原型，采用分布式控制器和 PCIe 6.0（64 Gbps，8 通道）作为连接服务器的主机接口，通过 PAM4（四电平脉冲幅度调制）技术，以低功耗实现了 128 Gbps 高带宽。原型测试表明，在功耗低于 40 瓦的情况下，可实现 5TB 容量和 64GB/s 带宽。

三星

已启动 HBF 产品开发的概念设计额其他早期工作，计划利用其在类似技术和产品方面的研发经验，重新定义 HBF 在更广泛的 AI 内存和存储架构重组中的角色。据悉，三星计划最快在 2027 年底或 2028 年初将 HBF 技术应用于英伟达、AMD 和谷歌的实际产品中。

长江存储

明确将 HBF 作为核心战略方向，依托 Xtacking 4.0 架构的晶圆级键合经验，计划通过 TSV 与混合键合技术，将多层 3D NAND 直接与 GPU 或 AI 加速器进行 2.5D/3D 集成，以解决 AI 时代 HBM 成本过高与容量受限的痛点。此外，长江存储提出“HBF+NPU”助力存算协同，为端侧智能设备提供解决思路。

图 11 HBF 规划时间图

企业	2025 年	2026 年	2027 年	2028 年	2029 年	2030 年
SANDISK™ 闪迪	首次提出 HBF 概念	下半年交付首批 HBF 内存样品	年初提供首批采用 HBF 技术的推理设备样品			
SK hynix		年初发布 AIN B 的 Alpha 版本	推出 评估样机			HBF 市场 扩张节点
KIOXIA	8 月推出高带宽闪存模块原型					
SAMSUNG		已启动 HBF 产品开发的概念设计额其他早期工作	2027 年底或 2028 年初将 HBF 技术应用于英伟达、AMD 和谷歌的实际产品中			

数据来源：公开信息

第四章 消费类存储产品应用与发展分析

ANALYSIS OF CONSUMER MEMORY PRODUCTS APPLICATION AND DEVELOPMENT

消费类存储产品的发展正深度贴合智能手机与 PC 两大终端的技术迭代和用户需求升级。QLC 闪存技术在大容量应用上持续发力与端侧 AI 应用的普及形成双轮驱动，推动手机与 PC 存储市场呈现融合发展趋势，大容量、高性能、低延迟成为核心发展方向。同时，存储成本的上涨也导致行业配置分化，高端旗舰机型成为大容量存储的核心落地场景。本章将围绕智能手机与 PC 设备的存储应用现状及未来趋势展开分析，挖掘消费类存储产品的发展逻辑与核心机遇。

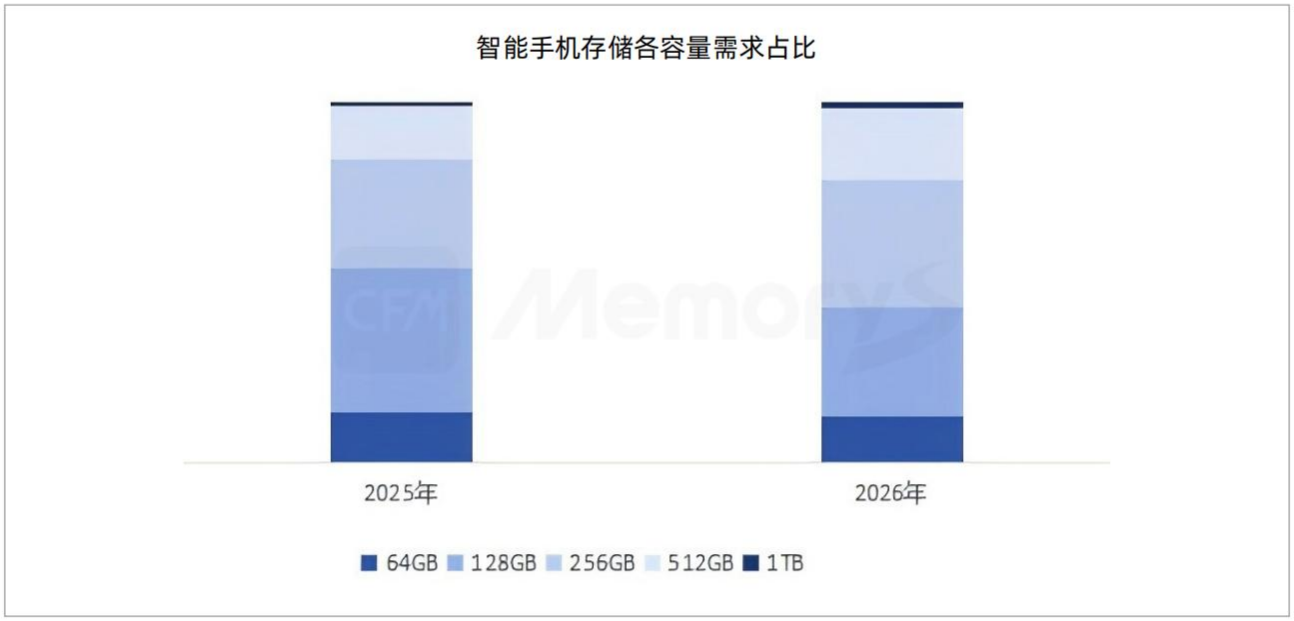
一、智能手机存储：端侧 AI 驱动容量与性能双升级

1、市场现状：QLC 持续推进，512GB 成主流配置

在手机领域，QLC 技术正处于终端评估导入阶段，存储原厂已推出面向智能手机的 QLC UFS 方案，主流手机品牌亦开始规划或评估导入 QLC 闪存。铠侠已于 2026 年 1 月开始向客户提供基于第八代 BiCS FLASH 技术的 QLC UFS 4.1 嵌入式闪存样品，提供 512GB 和 1GB 容量，专为高端智能手机及端侧 AI 设备设计；SK 海力士则已启动 321 层 2Tb QLC NAND 闪存量产，计划是扩展至智能手机的 UFS 产品，为手机端大容量存储铺路。

从市场格局看，2025 年全球智能手机市场在 AI 应用驱动下稳步复苏，用户换机周期延长至 40 个月以上，促使消费者购机时普遍倾向“买大不买小”。据 CFM 闪存市场数据显示，2025 年智能手机存储市场中，512GB 及以上容量机型的占比达约两成。

图 12 智能手机各容量需求占比



数据来源：CFM 闪存市场

2、端侧 AI 重构存储需求：从容量到带宽的全方位升级

端侧 AI 在手机端的加速落地，对存储系统提出了全新要求。AI 手机本地运行大模型不仅需要更大容量，更对内存带宽提出了高要求。例如，vivo 30 亿参数的蓝心端侧大模型内存占用量约为 2GB，OPPO 的 1.5B 小模型经量化后体积也可达 3GB。

异构计算（CPU+GPU+NPU）的普及，使手机在存储架构上与 PC 走向技术趋同——高带宽、低延迟、大容量成为共同诉求。LPDDR5X 正成为横跨两大市场的“通用语言”，在手机端，它是旗舰机型的标配。这也直接推动 UFS 5.0、LPDDR6 等新一代存储标准加速落地。

主流厂商已围绕端侧 AI 全面升级内存配置：三星 Galaxy S26 Ultra 顶配版搭载 16GB 内存，一加与 iQOO 则推出 24GB+1TB 顶配存储组合。在 AI 能力层面，荣耀、OPPO、vivo 等国产厂商也已深度融合 DeepSeek-R1 等大模型，行业正加速迈入以“超大内存 + 系统级模型”为核心的高阶 AI 体验竞争阶段。

3、配置与方案分层：旗舰引领大容量，中低端加速技术下探

2025 年底至 2026 年初发布的旗舰机型，存储配置呈现高度趋同：12GB 内存 +256GB 存储起步，16GB+512GB 成为主流，顶配普遍达 16GB+1TB。安卓旗舰已全面普及 LPDDR5X+UFS 4.0/4.1 组合。从配置对比可见，512GB 已成为旗舰标配，1TB 选项逐渐普及。

表 12 主流手机旗舰标准版硬件参数对比

主流手机旗舰标准版对比				
品牌 / 机型	苹果 17	三星 Galaxy S26	华为 Mate80	小米 17
系统	iOS 26、灵动岛	One UI 8.5	HarmonyOS 6.0	Xiaomi HyperOS 3
处理器	A19 芯片	骁龙 8 Elite Gen5 for Galaxy	麒麟 9020	骁龙 8 Elite Gen5
续航	3692mAh	4300mAh	5750 mAh	7000mAh
充电	40W 有线 +25W 无线	25W 有线 +15W 无线	66W 有线 +50W 无线	100W 有线 +50W 无线
屏幕	6.3 英寸 OLED 屏	6.3 英寸（直角）/6.1 英寸（圆角）	6.75 英寸 OLED 直屏	6.3 英寸 OLED 直屏
影像	48MP 融合式双摄系统（主摄 + 超广角）	50MP 主摄 + 12MP 超广角 + 10MP 长焦，3 倍光变	50MP 超光变 + 40MP 超广角 + 12MP 潜望长焦	50MP 徕卡三摄（主摄 + 超广角 + 浮动长焦）
解锁	Face ID	超声波指纹解锁、人脸识别	侧边指纹解锁、人脸识别	超声波指纹解锁、人脸识别
存储配置	LPDDR5X+NVMe 存储	LPDDR5X+UFS4.0	LPDDR5+UFS3.1	LPDDR5X+UFS4.1
存储容量 / 售价	8+256GB：5999 元 8+512GB：7999 元	12+256GB：6999 元	12+256GB：4699 元 12+512GB：5199 元 16+512GB：5499 元	12+256GB：4499 元 12+512GB：4799 元 16+512GB：4999 元 16+1TB：5299 元

数据来源：公开信息

表 13 主流手机旗舰标准版硬件参数对比(续上表)

主流手机旗舰标准版对比（续上表）					
机型	vivo X300	OPPO Find X9	荣耀 Magic8	一加 15	iQOO 15
系统	OriginOS 6.0、原子岛	ColorOS 16.0	MagicOs 10	ColorOS 16	OriginOS 6.0
处理器	天玑 9500	天玑 9500	骁龙 8 Elite Gen5	骁龙 8 Elite Gen5	骁龙 8 Elite Gen5
续航	6040mAh	7025mAh	7000mAh	7300mAh	7000mAh
充电	90W 有线 +40W 无线	80W 有线 +50W 无线	90W 有线 +80W 无线	120W 有线快充 +50W 无线	100W 有线快充 +40W 无线
屏幕	6.31 英寸直屏	6.59 英寸柔性屏	6.58 英寸 OLED 直屏	6.78 英寸柔性屏	6.85 英寸直屏
影像	2 亿像素蔡司主摄 +5000 万蔡司 APO 长焦 +5000 万蔡司超广角	50MP 主摄 +50MP 超广角 +50MP 潜望长焦，哈苏联名	50MP 主摄 +50MP 超广角 +6400 万超夜神长焦	50MP 主摄 +50MP 超广角 +50MP 潜望长焦	50MP 索尼主摄 +50MP 潜望长焦 +50MP 超广角
解锁	3D超声波单点指纹、人脸识别	屏下超声指纹，面部识别	3D超声波单点指纹、人脸识别	屏下超声指纹，人脸识别	超声波指纹解锁，人脸识别
存储配置	LPDDR5X Ultra+UFS4.1	LPDDR5X+UFS4.1	LPDDR5X+UFS4.1	LPDDR5X+UFS4.1	LPDDR5X Ultra+UFS4.1
存储容量 / 售价	12+256GB：4399 元 16+256GB：4699 元 12+512GB：4999 元 16+512GB：5299 元 16+1TB：5799 元	12+256GB：4399 元 12+512GB：4999 元 16+512GB：5299 元 16+1TB：5799 元	12+256GB：4499 元 12+512GB：4799 元 16+512GB：4999 元 16+1TB：5499 元	12+256GB：3999 元 16+256GB：4299 元 12+512GB：4599 元 16+512GB：4899 元 24+1TB：5399 元	12GB+256GB：4199 元 12GB+512GB：4699 元 16GB+256GB：4499 元 16GB+512GB：4999 元 16GB+1TB：5499 元

数据来源：公开信息

在 1000-3000 元价位的中低端市场，存储配置呈现明显分化。一方面，入门级机型仍以 UFS 2.2+LPDDR4X 为主；另一方面，部分主打性价比的机型开始采用旗舰级存储方案，UFS 4.0/4.1 已开始向 2000 元档机型渗透，LPDDR5/X 正加速下放。

表 14 部分中低端手机及其硬件配置

部分中低端手机及其相关配置						
品牌	时间	存储容量	存储配置	处理器	电池容量	当前价格 / 元
OPPO K13 Turbo Pro	2025.07	12GB + 256GB，12GB + 512GB，16GB + 256GB，16GB + 512GB	LPDDR5X+ UFS4.0	第四代骁龙 8s	7000mAh	1899-2699
一加 Turbo 6	2026.01	12GB+256GB，12GB+512GB，16GB+256GB，16GB+512GB	LPDDR5X +UFS4.1	第四代骁龙 8s 风驰版	9000mAh	2099-2899
iQOO Z11 Turbo	2026.01	12GB+256GB、12GB+512GB、16GB+256GB、16GB+512GB、16GB+1TB	LPDDR5XUltra +UFS4.1	第五代骁龙 8	7600mAh	2699-3999
红米 Note 15	2025.08	8GB+128GB、8GB+256GB、12GB+256GB	LPDDR4X +UFS2.2	第三代骁龙 6	5800mAh	1099-1499
红米 Turbo 5	2026.01	12GB+512GB、12GB+256GB、16GB+512GB、16GB+256GB	LPDDR5X Ultra +UFS 4.1	天玑 8500-Ultra	7560mAh	2299-3099
vivo Y500 Pro	2025.11	12GB+512GB、12GB+256GB、8GB+256GB、8GB+128GB	LPDDR4X +UFS2.2	天玑 7400	7000mAh	1799-2599
荣耀 X70	2025.07	8GB+128GB、8GB+256GB、12GB+256GB、12GB+512GB	LPDDR4X +UFS2.2	第四代骁龙 6	8300mAh	1399-1999
荣耀 500	2025.11	12GB+256GB、12GB+512GB、16GB+512GB	LPDDR5X +UFS 3.1	第四代骁龙 8s	8000mAh	2699-3299
vivo S50	2025.12	12GB+256GB、12GB+512GB、16GB+256GB、16GB+512GB	LPDDR5X +UFS4.1	高通第三代骁龙 8s	6500mAh	2999-3599

数据来源：公开信息

在容量升级的同时，存储方案本身也在持续演进。主流趋势是采用 LPDDR PoP 封装 + 分离式 UFS 方案，该方案在散热和良率上更具优势。集成式存储方案（eMCP/uMCP）则凭借高集成度，在轻薄和中低端机型中仍

占有一席之地。值得注意的是，2026 年存储涨价潮正在催生新的方案探索，部分手机厂商正考虑重新引入外置存储卡设计，以缓解成本压力。

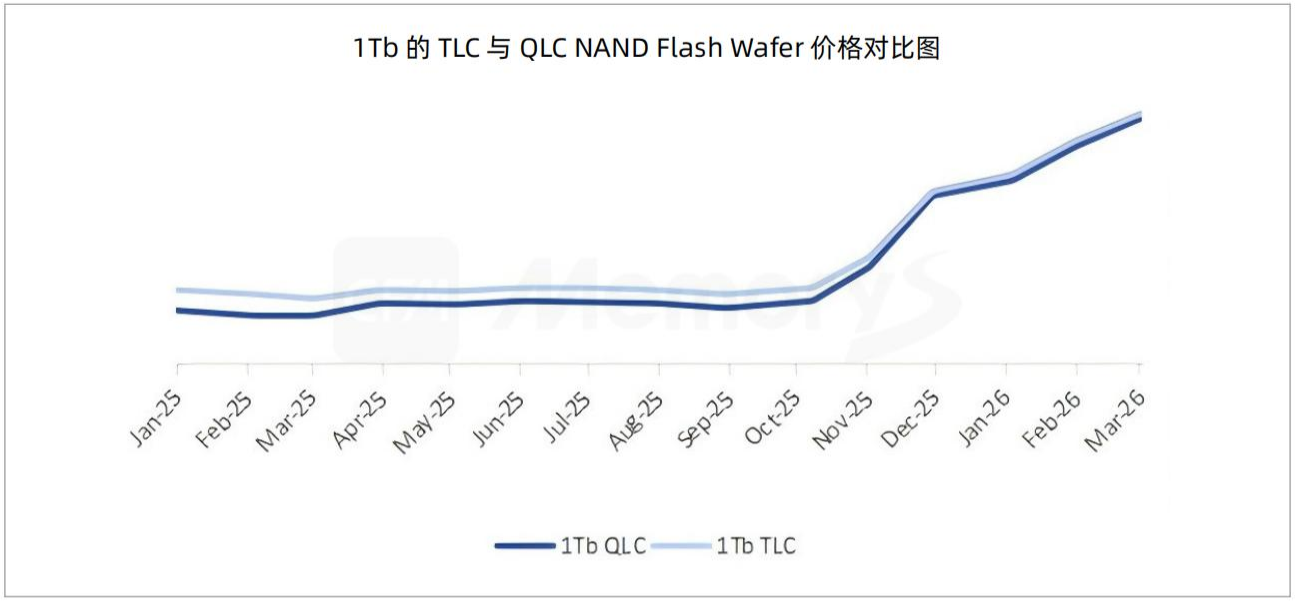
二、PC 设备的存储应用与发展

1、消费级 SSD：QLC 从“可选”走向“主流”

AI PC 的普及推动了市场的温和复苏，但自存储价格大幅上涨后，整机成本显著推高，为应对成本压力，PC 厂商纷纷调整产品策略：一方面在高端市场，通过提升 AI PC 的出货比重获取更高溢价以对冲成本上涨；另一方面在中低端市场，则通过精简配置控制成本、维持利润率。受此影响，PC 整体出货规模收缩，预计 2026 年全球 PC 出货量将下调至约 5%~10%。

在此背景下，QLC 技术凭借同容量下的微弱成本优势，成为终端厂商平衡成本与性能的重要技术路径。QLC 在 PC 领域的普及进程显著领先于手机市场，SK 海力士已量产全球首款 321 层 2Tb QLC NAND，明确将率先应用于 PC SSD。支持 QLC 的主控方案日趋成熟，慧荣科技、群联电子、联芸、英韧科技、点序科技等主控厂商均推出适配 QLC 的解决方案，通过优化固件算法有效缓解缓存饱和后的性能下降问题。长江存储最新推出的 PC42Q QLC SSD，在 Xtacking 4.0 架构的加持下，可靠性、电荷保持能力、存储密度、能耗及速度均实现显著提升，其中 1Tb 版本的耐久度高达 300TBW。

图 13 1Tb 的 TLC 与 QLC NAND Flash Wafer 价格对比图



数据来源：CFM 闪存市场

2、接口升级：PCIe 5.0 与 QLC 结合，开启高性能存储新路径

PCIe 5.0 与 QLC 的结合正成为 PC 存储升级的新方向。这一组合既能满足 AI PC 和高端电竞本对高带宽的

刚性需求，又能通过 QLC 的相对成本优势一定程度上缓解存储价格上涨压力，为终端厂商提供性能与成本兼顾的解决方案。目前，主流原厂已纷纷布局 PCIe 5.0 QLC 产品：美光于 2026 年 1 月推出业界首款面向客户端计算的 PCIe 5.0 QLC SSD——3610 系列；铠侠、三星、Solidigm 等也在积极推进产品商用化。与此同时，封测厂商为 QLC 产品的规模化生产提供了关键支撑，确保产能与良率满足终端需求。

同时，以 TLC 为代表的旗舰性能产品仍在持续迭代，为 PCIe 5.0 树立了性能标杆。长江存储于 2026 年 2 月专为 AI PC 推出首款原厂四通道 PCIe 5.0 TLC SSD——PC550，满载功耗低于 6W，实现了高性能与发热控制的平衡，有效解决了阻碍 PCIe 5.0 SSD 大规模应用于 OEM 设备的关键温控瓶颈。三星、SK 海力士、闪迪等原厂也推出了各自的 TLC 旗舰产品，在顺序读写与随机读写性能上持续突破。部分原厂 PCIe 5.0 消费级 SSD 产品参数对比如下：

表 15 部分原厂 PCIe 5.0 消费级 SSD 产品参数对比

企业	三星	美光	SK 海力士	铠侠	闪迪	长江存储	长江存储
产品名称	9100 Pro	3610	Platinum P51	EXCERIA G3 (VC10)	WD_BLACK SN8100	致态 TiPro 9000	PC550
外形	M.2 2280	M.2 2280	M.2 2280	M.2 2280	M.2 2280	M.2 2280	M.2 2242/2280
主控芯片	自研 5nm Presto 主控	群联 PS5031-31T 无缓主控	SK hynix Alistar 主控	群联 E31T 系列无缓主控	慧荣 SM2508 主控	慧荣 SM2508 主控	国产 4CH, Dram-less
NAND Flash	V8 3D TLC	232 层 3D QLC	238 层 3D TLC NAND	BiCS8 218 层 3D QLC	BiCS8 3D TLC	Xtacking 4.0 3D TLC	Xtacking 4.0 3D TLC
DRAM	有	/	SK hynix LPDDR4	/	有	1GB/2GB LPDDR4	/
容量	1TB-8TB	1TB-4TB	500GB-2TB	1TB-2TB	1TB-4TB	1TB-4TB	512GB-2TB
顺序读取速度	14,800 MB/s	11,000 MB/s	14,700 MB/s	10,000 MB/s	14,900 MB/s	14,000 MB/s	10,500MB/s
顺序写入速度	13,400 MB/s	9,300 MB/s	13,400 MB/s	1TB: 7,900 MB/s2TB: 8,200 MB/s	14,000 MB/s	12,500 MB/s	10,000MB/s
随机读取速度	2,200K IOPS	1,400K IOPS	2,300K IOPS	1,300K IOPS	2,200K IOPS	2,000K IOPS (2TB 版)	1,300K IOPS (1TB 版本)
随机写入速度	2,600K IOPS	1,500K IOPS	2,400K IOPS	1,400K IOPS	2,400K IOPS	1,800K IOPS (2TB 版)	1,300K IOPS (1TB 版本)
功耗	约 5.5W	/	10W	5.3W	约 5.2W	/	Idle < 3W Active < 6W
耐久度	1TB: 600TBW 2TB: 1200TBW	/	500GB: 300TBW 1TB: 600TBW 2TB: 1200TBW	1TB: 600TBW 2TB: 1200TBW	1TB: 600TBW 2TB: 1200TBW	1TB: 600TBW 2TB: 1200TBW	512GB:300TBW 1TB:600TBW 2TB:1200TBW

数据来源：公开信息

从产品布局可见，PCIe 5.0 SSD 已形成清晰的差异化定位：以三星 9100 Pro、SK 海力士 Platinum P51、长江存储 PC550 为代表的 TLC 旗舰产品，凭借极致性能与发热控制的完美平衡抢占高端游戏与创作市场；而以美光 3610、铠侠 EXCERIA G3 为代表的 QLC 产品，则通过性能与成本的平衡，成为 AI PC 及主流市场加速 PCIe 5.0 普及的关键力量。这一分层格局既满足了不同场景对存储带宽的刚性需求，也为终端厂商提供了灵活的成本控制选项，推动 PCIe 5.0 与 QLC 的组合从“技术标杆”向“规模应用”加速落地。

3、内存形态革命：从 SO-DIMM 到 LPCAMM 的演进之路

笔记本电脑内存形态正经历近十年来最重要的变革。传统 SO-DIMM 插槽占用空间大、频率提升受限，而焊接式 LPDDR 内存又导致用户无法升级维修，LPCAMM2 的出现同时解决了这两大痛点，具有以下优势：

体积缩减

外形尺寸较双 SO-DIMM 设计缩减 64%，为超薄本释放内部空间

性能跃升

采用 LPDDR5X 颗粒，速率达 9600MT/s，较 DDR5 SO-DIMM 提升约 1.5 倍

功耗优化

内置 PMIC 电源芯片，待机功耗最高下降 80%

可维护性

可插拔模块化设计，用户可通过螺丝固定自行更换升级

产品进展方面，三星于 2026 年 2 月推出单条 96GB LPCAMM2 模组，速率达 9600MT/s，采用单条双通道架构，仅需安装一根即可实现完整双通道带宽 。美光 Crucial 此前已率先推出 64GB/8533MT/s 模块，适配联想 ThinkPad P1 Gen7、戴尔 Pro Max 系列等机型 。

平台支持层面，英特尔 Panther Lake 平台（酷睿 Ultra 3 系列）中，Ultra X9 388H、X7 368H 等高端型号已明确支持 LPDDR5X-9600 内存 。预计 2026 年，AMD 和 ARM 架构笔记本也将逐步导入 LPCAMM2 标准，未来两年内有望成为高端轻薄本和移动工作站的主流内存形态。

4、AI PC 重构标准：内存带宽成性能新瓶颈

AI PC 的普及正在重新定义 PC 存储的性能基线。大语言模型推理的本质是内存带宽受限而非算力受限——这是当前端侧 AI 面临的核心瓶颈。量化来看，想要流畅运行 70B 级别模型，系统内存至少需要 48GB 以上，内存带宽则需达到 256GB/s 量级。从产品落地来看，厂商正通过提升内存规格突破这一瓶颈。英特尔第二代酷睿 Ultra 处理器 NPU 算力达 48TOPS，总算力 120TOPS；AMD Ryzen AI Max+ 系列将统一内存容量推至 128GB，可本地运行 1200 亿参数模型；微星 AI Edge 主机配备 96GB 可变图形内存，运行高达 1090 亿参数 LLM 时可达 15 Tokens/s 的输出速度。这些进展表明，AI PC 正从“能否运行”迈入“流畅体验”的新阶段。

与此同时，当前市场主流 AIPC 产品的存储配置已形成清晰的分层（见下表）：以联想 ThinkBook 14+、惠普星 Book Pro 16 为代表的商务全能本，普遍搭载 32GB 大容量内存（LPDDR5X 或 DDR5），以满足多任务处理与端侧 AI 推理需求；以 ROG 魔霸新锐、七彩虹隐星 P16 Pro 为代表的游戏与创作本，则凭借高带宽 DDR5 内存和 PCIe 4.0/5.0 接口，为重度渲染、AI 视频处理等场景提供充足性能支撑。

表 16 主流 AIPC 产品存储配置

序号	品牌 / 机型	处理器	内存配置	存储配置	内存类型	存储接口	目标场景
1	Apple 2025 款 MacBook Air 13 英寸	Apple M4（10 核 CPU+8 核 GPU，16 核 NPU）	16GB	256GB SSD	统一内存	板载 PCIe 4.0 NVMe SSD	商务办公、内容创作、AI 轻量推理、移动便携
2	联想 ThinkBook 14+ 2025 锐龙 AI 全能本	AMD 锐龙 7 H260（AI 引擎，16 TOPS）	32GB	1TB SSD	LPDDR5X 7500MT/s	M.2 2280 PCIe 4.0 NVMe	商务办公、数据统计、AI 文档处理、移动生产力
3	华硕 无畏 16 锐龙版 2025	AMD 锐龙 7 H260	16GB	1TB SSD	DDR5 5600MHz	M.2 2280 PCIe 4.0 NVMe	大屏办公、轻度设计、AI 教育应用、学生全能本
4	惠普 星 Book Pro 16 2025	Intel 酷睿 Ultra7 255H	32GB	1TB SSD	DDR5 5600MHz	M.2 2280 PCIe 4.0 NVMe	轻薄创作、AI 办公、影音娱乐、高端商务
5	七彩虹 隐星 P16 Pro 焕新版	Intel 14 代 酷睿 i7-14650HX	32GB	1TB SSD	DDR5 5600MHz	M.2 2280 PCIe 4.0 NVMe	二次元电竞、重度渲染、AI 视频处理、高性能创作
6	神舟 战神 T10 Pro	Intel 14 代 酷睿 i9-14900HX	16GB（支持升级）	1TB SSD	DDR5 5600MHz	M.2 2280 PCIe 4.0 NVMe	高性价比游戏、专业影像、AI 加速计算、学生游戏本
7	宏碁 新暗影骑士·擎 6	Intel 14 代 酷睿 i7-14650HX	16GB	1TB SSD	DDR5 5600MHz	M.2 2280 PCIe 4.0 NVMe	游戏娱乐、电竞、专业影像剪辑、AI 加速创作
8	ROG 魔霸新锐 2025	Intel 酷睿 Ultra7 255HX	16GB	1TB SSD	DDR5 5600MHz	M.2 2280 PCIe 4.0/5.0 NVMe	高端游戏、专业电竞、AI 渲染、重度内容创作

数据来源：公开信息

AI 的爆发与服务器产能的虹吸效应，正将消费级存储推向“量价背离”的十字路口，整个行业正处于涨价的阵痛之中。存储涨价并非只是压力，也可能成为终端形态变革的催化剂。当本地存储回归“操作系统 + 核心应用 + 实时数据”的基本需求，而海量个人内容加速向云端迁移，云手机、云 PC 等新型终端的规模化应用或将真正到来。未来，用户或不再执着于手机本地容量的上限，转而关注跨端数据无缝访问的体验。这意味着，消费类厂商需要重新思考“更轻”的终端策略——云手机、云 PC 的机会，或许真的来了。

第五章 新兴消费领域存储产品应用与发展

APPLICATION AND DEVELOPMENT OF MEMORY PRODUCTS IN EMERGING CONSUMER SECTORS

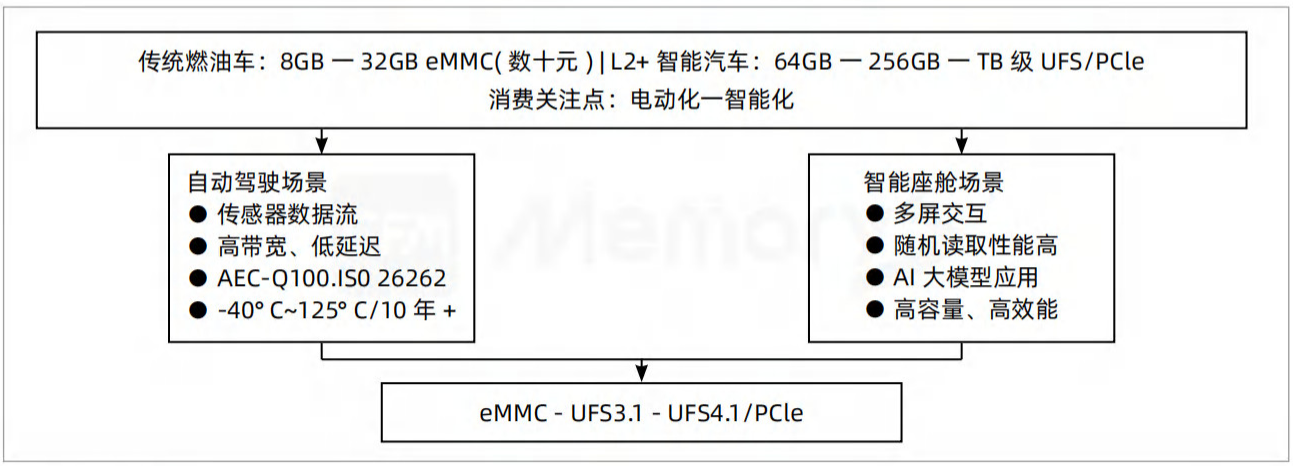
随着端侧 AI 成为消费电子创新的核心引擎，智能汽车与智能穿戴等新兴领域对存储的容量、带宽、功耗提出严苛要求；与此同时，超高清内容创作的全民普及，特别是 8K 视频、高帧率慢动作等高码流录制，对存储的持续写入速度与可靠性形成极限挑战。新兴应用驱动下，存储技术正从通用标准化向场景深度定制转型，成为定义终端体验的关键基石。本章深入剖析智能汽车、智能穿戴、运动影像等新兴消费领域，揭示存储产品如何通过技术创新与产业升级应对多元化需求。

一、智能汽车：车载存储从零部件到智能核心的产业升级

在 AI 定义汽车的浪潮下，智能座舱与自动驾驶正重塑车载存储的价值。传统燃油车服务于信息娱乐系统，典型内置存储容量为 8GB eMMC，中高配可达 16-32GB，而当前 L2+ 级智能汽车为支持高精度地图、多传感器融合及本地 AI 大模型，存储容量已跃升至 64GB-256GB 并向 TB 级迈进。消费者的关注点从“电动化”转向“智能化”，高阶辅助驾驶成为核心卖点，直接拉动了对车载存储的需求。与此同时，不同场景对存储的要求显著分化：自动驾驶以处理海量传感器数据为核心，需高带宽、低延迟及车规级可靠性（符合 AEC-Q100 及 ISO 26262 功能安全）；智能座舱则注重交互体验，对随机读取性能要求极高，并随 AI 大模型渗透，对高容量、高能效存储的需求激增。

技术迭代上，主流车型已从 eMMC 转向 UFS 3.1/4.0，UFS 4.1 及 PCIe 方案正加速落地高端车型。以小鹏汽车为例，其 G6、G9、P7i、X9 等主力车型搭载的 XNGP 智驾系统，需实时运行端到端大模型、处理海量传感器数据，对带宽与容量提出严苛要求，因此率先采用 UFS 4.1 车规存储。值得强调的是，车规级存储与消费级芯片有本质差异：车规级存储需在极端环境下稳定运行 10 年以上，失效率低于百万分之一，并通过功能安全认证。这构成了车规存储的核心壁垒，使其从普通零部件升级为定义智能体验与行驶安全的关键基石。

图 14 车载存储需求分化与技术演进路径



数据来源：公开信息

二、智能穿戴：从“手机配件”到“独立智能体”的存储跃迁

智能穿戴设备正从“手机附属品”向“独立智能体”转型，端侧 AI 与本地软件功能的爆发直接催生了对大容量存储的需求。以 AI 眼镜为例，依托端侧 AI，设备可在弱网甚至无网环境下，独立完成图像识别、实时翻译、会议纪要等复杂任务，真正实现数据即产生、即处理。无论是 AI 模型本身，还是高清拍摄、语音处理等本地应用，都对存储容量形成刚性支撑需求。近期发布的千问 AI 眼镜 G1，其内置 64GB 存储空间，正是为了满足连续拍摄与 AI 功能稳定运行。各类穿戴设备在迈向 “独立智能体” 的过程中，对存储的要求正趋于一致：存储已从被动配件升级为决定端侧智能体验的核心部件。

围绕高性能、小体积、低功耗的极致要求，存储产业链正从多个维度展开系统性创新。在封装集成层面，ePOP 技术将存储与内存立体堆叠于 SoC 上方，可节省占板面积 50%-75%；在超薄定制层面，ATP 推出 6.7mm×6.7mm×0.65mm 的全球最小 e.MMC，较标准尺寸缩小 67%。在协同优化层面，存储方案通过与高通等主流平台认证，以及电源管理固件调优，确保有限散热空间内的稳定运行与能效平衡。

表 17 智能穿戴部分存储产品方案对比

产品系列	存储介质	封装尺寸 / 厚度	应用方向	系统协调 / 其他
江波龙 ePOP5x	eMMC + LPDDR5X 集成；DRAM 传输速率高达 8533Mbps	8.0×9.5mm，厚度最薄 0.52mm	AI 眼镜、高端智能穿戴设备	搭载系统快速启动、低功耗技术、SoC 调优等自研算法
佰维存储 ePoP5X	64GB eMMC 5.1 + LPDDR5X；HS400 高速模式，400MB/s；LPDDR5X 速率 8533Mbps	8.0×9.5mm，厚度 0.54mm	AI 穿戴设备	支持高通平台；集成 RAM 与 ROM，节省 PCB 空间 75%
康盈半导体 ePOP 嵌入式存储	eMMC + LPDDR 集成；最大 64GB+32Gb；读 300MB/s，写 200MB/s	8.0×9.5×0.85mm	智能手表、手环、VR 眼镜、蓝牙耳机	垂直搭载于 SoC，不占用 PCB 板平面空间
时创意 超薄 ePOP	eMMC 5.1 + LPDDR4X；16GB+64GB；读 300MB/s，写 200MB/s；传输速率 4266Mbps	8.0×9.5×0.6mm	AI 眼镜、智能手表	支持 S.M.A.R.T 自监控、LDPC 3.0 ECC 纠错
ATP 电子 6.7mm e.MMC (E700Pc/E600Vc)	eMMC 5.1；64GB TLC (12TBW))；HS400 模式，400MB/s	6.7mm×6.7mm ×0.65mm	智能眼镜、智慧穿戴	电源管理固件调优，功耗节省 70%；已与全球头部穿戴厂商合作验证
铠侠 UFS 4.1 嵌入式闪存	UFS 4.1；256GB-1TB；23.2Gbps/ 通道，总计 46.4Gbps；BiCS FLASH ™ 3D 闪存	9mm×13mm，0.8mm 厚度	智能眼镜、边缘 AI 设备	支持 WriteBooster、HS-LSS（链路启动时间缩短 70%）；可在狭小镜腿内完成 AI 算力布局
三星 Galaxy Watch 系列内置存储	eMMC；Galaxy Watch Ultra：2GB RAM + 64GB；Galaxy Watch8：2GB + 32GB	—	智能手表	与自研 Exynos 处理器协同优化

数据来源：公开信息

三、运动摄影与专业影像：消费级内容创作催生高性能存储市场

随着短视频与社交媒体的深度普及，终端设备视频画质持续升级，4K 已成标配，8K 录制加速渗透。画质的跃升直接推高了对存储持续写入性能的要求，以确保高码流视频录制不掉帧；而户外拍摄等极端环境对可靠性提出军工级标准，现场回放、移动剪辑等即时流程则对读取速度与低延迟提出严苛需求。由此，运动摄影与专业影像的存储方案，正围绕“速度、可靠性、尺寸”三重维度展开极限博弈，具体如下：

表 18 运动摄影与专业影像部分存储方案对比

存储产品	核心架构 / 接口	市场定位	核心优势	典型应用场景
CFexpress	NVMe PCIe 架构、VPG 认证	专业影像旗舰	顶级连续写入性能支持、8K RAW 与高速连拍	专业相机、电影机、高端影像设备
microSD 卡	UHS-I、SD Expres	小型化高可靠存储	体积极小、抗震耐摔接近 SSD 级速度	运动相机、无人机、便携运动摄影机
移动 NVMe SSD	NVMe 高速协议	现场高速工作流	超高速读写、低延迟大文件秒传	素材现场回传、移动直剪、外拍工作流
3.5 英寸 HDD	传统机械硬盘	大容量冷备份	单位容量成本最低、稳定性强	海量素材长期归档、冷备份、数据中心存储

数据来源：公开信息

为适配不同场景的核心需求，各类存储产品的技术路线持续迭代演进，推动了产业价值的重塑：高速 CFexpress 卡单价突破 4000 元，为存储厂商开辟高利润细分赛道；竞争维度从“容量 + 价格”转向“速度 + 耐久性 + 数据安全”；供应链关系深化，联合研发成为主流，从产品定义阶段的全链路定制正取代传统配件采购模式。存储已从通用配件升级为决定设备旗舰性能的核心专供部件，价值链上移趋势愈发明显。

四、其他 AI 消费终端：存储的泛在化与智能化

AI 技术正从核心赛道溢出，渗透至机器人、智能家居等多元场景。不同终端对存储的需求呈现差异化，这些设备虽形态各异，但对存储的需求逻辑高度一致——端侧 AI 能力越强，对存储的性能、功耗、可靠性要求越高。具体如下表所示：

表 19 泛 AI 终端存储需求对比

终端类型	本地 AI 功能	容量需求	性能要求	耐久性要求	安全要求
扫地机器人	高精地图、AI 避障	64GB-128GB	随机读写	高（持续日志写入）	地图数据隐私
AI 摄像头	人形检测、异常报警	64GB-128GB	持续写入速度	极高（7×24h 录制）	视频数据加密
智能健身镜	动作捕捉、实时纠正	32GB-64GB	低延迟读取	中	用户生物信息
AI 学习机	本地知识库、作文批改	128GB-256GB	随机读写、高带宽	中	学习数据隐私

数据来源：公开信息

端侧 AI 的普及正将存储推向前所未有的挑战：其一，本地 AI 模型部署直接推高大容量需求，对成本敏感的消费电子构成压力；其二，终端持续产生的数据流（如 AI 摄像头全天录制）对存储耐久性提出严苛考验，需高耐久介质与智能磨损均衡算法；其三，家庭影像、语音记录等隐私数据要求存储具备硬件级安全能力，从“数据容器”升级为“数据保险箱”。

面对上述挑战，产业已形成清晰的应答路径。技术方案上，eMMC/UFS 凭借高集成度正全面取代 SPI NAND，成为 AIoT 设备的默认选择；QLC NAND 则在大容量、读密集型场景（如智能摄像头本地缓存）中提供性价比方案。产业模式上，存储厂商与主控芯片平台深度合作推出“主控 + 存储”预验证套件以降低开发门槛；针对头部客户，更提供从芯片筛选、定制固件到封装设计的全流程深度定制服务，实现终端产品的全链路协同优化。

五、趋势总结：新兴应用重塑存储技术范式

AI 正在加速渗透并重塑多元终端场景。从智能汽车到可穿戴设备，从运动影像到智能家居等多元场景，所有终端都在向“数据驱动、智能定义”演进。数据、算力、存储——三者构成端侧智能的核心三角，而存储不再是配角，而是决定终端体验的基石。相比云侧算力基础设施的“军备竞赛”，AI 的真正价值需要通过终端应用来“变现”，正如智能手机让互联网走进千家万户，AI 也需要眼镜、汽车等新形态走进每个人的生活。AI 前期的投入，终将依靠终端应用的创新来实现闭环，推动云与端协同发展，而无论云端还是终端，对存储的要求只会更高。

⊙ 回看新兴消费领域，存储产业正沿着三条清晰的路径实现技术范式升级：

一是物理极限的突破

可穿戴设备与 AI 眼镜空间寸土寸金，ePOP 封装将存储与内存立体堆叠于 SoC 上方，节省占板面积高达 75%，厚度向 0.52mm 演进，让方寸之地也能承载完整端侧智能。

二是场景定义的深化

从通用标准件转向深度定制：车载存储强化宽温域与功能安全（ASIL-D），缺陷率控制在 10PPM 以内；穿戴设备主推超薄 ePOP 与低功耗固件调优；运动影像通过 VPG 认证保障 8K 录制不掉帧——场景定义规格，已成新常态。

三是系统协同的升维

存储不再是独立的采购件，而是与主控平台深度协同的定制方案。从芯片筛选、固件优化到封装设计，头部厂商正与存储伙伴实现全链路联合开发，让存储真正融入终端产品的底层逻辑。

展望未来，存储芯片将从“数据的容器”进化为“计算的单元”。在传感器数据的前端处理、端侧 AI 的低延迟推理中，存储将承载更多智能，成为定义终端体验的关键一极。无论是云端还是端侧，数据只会更多、算力只会更强，对存储的要求也只会更高。

版权与免责声明

微信公众号

闪存市场 APP

深圳市闪存市场资讯有限公司

邮箱：Service@ChinaFlashMarket.com