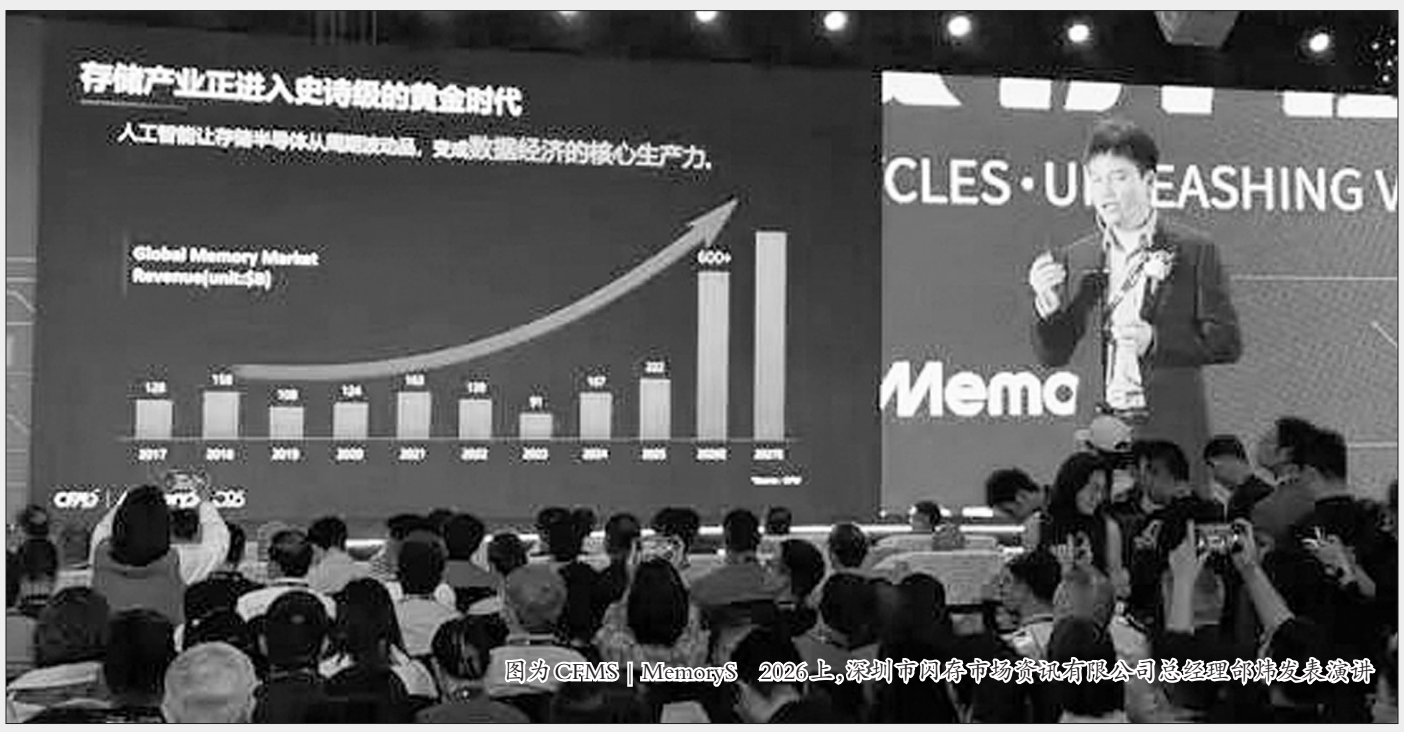
存储价格上涨是一次长周期范式转移



图为CFMS | MemoryS 2026上，深圳市闪存市场资讯有限公司总经理邰炜发表演讲

本报记者 张心怡

全球性的存储芯片短缺，不仅体现在价格跳涨、产能争夺和下游终端厂商的成本压力，甚至影响到了普通人的日常决策。在大模型、智能体等AI范式的部署竞速期，数据产生的方式、速度、规模已经远远超出了人们的想象，对存储芯片的需求呈现线性暴涨态势。

在3月27日举办的CFMS | MemoryS 2026上，深圳市闪存市场资讯有限公司总经理邰炜表示，存储行业的焦点已经从“看谁更便宜”，转向“看谁能拿到货”。本次存储的价格上涨完全不同于以往的周期性价格反弹，而是一次长周期的范式转移。

哪些规格的存储产品特别缺？

过去几十年，存储一直被视为行业周期的晴雨表。但如今，存储迎来了真正意义上的“时代拐点”，从BOM表的成本项上升为AI计算的战略资源，也从周期性产品变成数字经济的核心竞争力。

而存储重要性的直线上升，本质上是由于AI时代数据生成方式的变化。相比曾经由用户写文档、拍照、拍视频产生的数据，当前及未来由海量大模型的多模态参数所不间断产生的数据规模，带来了远超想象的存储空间需求。

在数据中心服务器这一AI算力基础设施上，存储容量正大幅提升。单个通用服务器的存储需求通常包括512GB~1TB的DDR5、4TB及以下的eSSD；而AI服务器的DDR5需求在1.5TB~4TB，eSSD需求为4/8TB并逐步向8/16TB提升，还需要HBM3E或HBM4。

邰炜指出，大模型的训练、推理、微调以及多模型应用等每一个环节都把存储的带宽和容量拉到极致。HBM从小众高端产品一举变成AI时代的石油；大容量DDR5内存成为AI服务器标配；企业级SSD不仅是存储的载体，更成为整个算力架构突破性能瓶颈的关键。

从具体规格来看，哪些存储产品的需求会格外紧俏？

DRAM方面，DDR5-6400（6400频率的DDR5）需求将在2026年上涨，DDR5-7200和DDR5-8000将在2027年加速落地；PCIe6.0 SSD在服务器的渗透开始加速；美光科技与英伟达联合开发的SOCAMM、基于LPDDR的MRDIMM等也开始大量应用；LPDDR不再是手机和超级本的专属存储，服务器、汽车等领域都将抢夺LPDDR产能。

NAND方面，随着AI短上下文演变为长上下文，KV Cache占用的存储空间会随着Token数量和并发请求量而线性暴涨。HBM无法承受如此量级的存储，因此KV Cache的需求开始大规模转向NVMe SSD，驱动eSSD成为2026年NAND最大的应用市场。

价格和产能走势如何？

由于AI服务器需求的爆发性增长，存储供应端的结构性错配和短缺已成常态。面对这种情况，全球头部原厂做出了同一个选择：将最先进的产能优先投向AI存储产品。这种产能分配，导致成熟制程、消费级产能被持续挤压——行业库存水位降至历史安全线以下。邰炜预计，2026—2027年存储产能增量有限。

“在这一阶段，有些产品刚从原厂下线就被立刻插入客户的服务器中，这在以往是不可想象的。尽管存储企业开始增加资本支出以扩张产能，但存储产能扩张周期往往长达18至24个月，最早要到2027年才有新产能释放。而新释放的产能也满足不了市场需求，供应问题在短期内难以缓解。”邰炜指出，“所以我们判断，在2026年，全球没有任何一款主流的AI产品能够做到供需完全平衡。”

同样值得注意的是，全球头部存储原厂在经过几轮的“巨亏周期”以后，不再盲目扩产，而是“纪律性增产”，优先提供高技术、高价值、高壁垒的产品。这也导致供应弹性下降，未来波动收窄，但利好整个行业的长期景气。

从先进产品来看，HBM凭借3D堆叠结构和TSV技术，已经实现了远超传统DDR的内存带宽，定义了当前技术的天花板。在HBM3e和HBM4中，混合键合技术正逐渐取代传统的微凸块，成为推动HBM性能升级的新核心；2.5D和3D封装也在系统集成中起到关键支撑作用。

此外，Base Die将是HBM的下一个决胜点。为了让计算和存储靠得更近，很多计算能力会下放给Base Die。

NAND的发展也将更具备价值导向，加速高附加值的企业级SSD和下一代存储技术渗透。随着300层以上NAND大规模量产，成本将持续下探，高可靠性、低延迟、高使用寿命将成为新的定价标尺。

在价格方面，存储价格从2025年第四季度起迎来“史诗级上涨”。合约价、现货价同步跳涨，并向渠道、模组、整机全线传导。存储的焦点已经从“看谁更便宜”，转向“看谁能拿到货”。

“很多人问我们价格什么时候会下跌，我们认为这完全不同于以往的周期性价格反弹，而是一次长周期的范式转移。”邰炜说道。不过，存储产品价格在经历连续三个季度大涨以后，预计从今年第三季度开始，涨价趋势放缓并逐渐收敛，但在具体产品线存在分化。

面向高涨的景气和供需的错配，邰炜认为“要在繁荣之下保持足够的冷静”，一方面要警惕AI投资过热带来的阶段性波动，资本涌入、项目扎堆，都可能导致短期需求透支；另一方面，要警惕价格过快上涨对下游需求的抑制，存储涨价最终要由终端承担，过度透支需求会反噬整个生态。

邰炜还向存储供给端和需求端提出建议。对于供给端，要坚持理性扩产，平衡AI与消费、高端与基础，用长协、共赢、稳定供给替代短期博弈。对于需求端，一要提前规划、多元供货，意识到锁产能比谈价格更重要；二要拥抱新技术架构，从被动买存储转向主动优化存储；三要与上游联合定义产品，共同应对市场变化。

“2026年存储相关领域的投资规模将远超其他所有芯片类型，真正影响AI未来竞争格局的产品形态是存储。其中市场需求增长最快的是企业级存储，在AI带动下的存力发展最直接的体现就在eSSD的容量上。”在3月27日举办的CFMS | MemoryS 2026上，长江存储科技有限责任公司固态硬盘事业部负责人谭弘表示。他代表长江存储分享了AI与企业级存储的未来，以及如何使用eSSD技术实现存算协同，突破AI时代的存力瓶颈。

AI时代 存力是决定生产效率的“炼油设备”

本报记者 连晓东

存力不是“油桶” 而是“炼油设备”

存力在不同历史时期扮演着不同角色。在PC时代，存力主要担当外存设备角色，主要产品形态是软盘、光盘；进入网络时代，开始出现数据中心概念的存力，成为信息基础设施；进入到移动互联网时代，用户对实时交互要求越来越高，短视频、直播等促进了闪存的发展，替代了一部分HDD成为主流。“现在我们进入到AI时代，GPU引领存储架构的升级，以QLC、HBM、高带宽闪存为代表的新技术新产品形态不断涌现。随着AI在云和端的持续渗透，存力将真正成为GPU时代决定生产效率的‘炼油设备’，而不是‘油桶’。”谭弘表示。

业界有观点认为AI竞赛正在进入下半场，谭弘表示，AI的上半场主要集中在训练，重在“厚积”，“这好比修炼内功，通过海量的数据来为系统筑基”，而真正要让AI发挥作用，关键在于推理侧，“AI的推理犹如（武术中的）招式，在多种多样化的应用场景中，拳、掌、腿等不同招式各适其用，经过训练不同的模型和数据，将适配于各种推理场景，为最终用户释放最大的价值。”谭弘表示，推理侧重在“薄发”，考验的是灵活运用，“一招制敌（解决问题）”。

存储带宽瓶颈 严重限制算力有效利用

随着全球各大训练模型的成熟，推理需求迎来全面爆发，算力和存力进行系统层面的深入整合将会是未来一个重大的发展趋势。“然而在这一整合实践中，由于存储墙的存在，AI在训练和推理中时刻面临着存储带宽的瓶颈。实际上，当前的算力增长已超过了存储带宽所带来的支撑限度，这就意味着海量的模型参数和上下文需要通过一个相对狭窄的漏斗口，即存储的带宽，进行传输，导致算力无法充分释放。”谭弘表示。

谭弘援引IEEE一篇文章中提出的论点说，当前AI革命的关键，已不仅仅在于算力，“真正限制我们大规模语言模型的瓶颈不是数学——而是存储”，并从训练和推理两个不同阶段展开说明。

“从训练阶段来看，随着模型变大，故障发生间隔也从之前的小时级别缩短到分钟级别，导致训练失败频次加剧，造成GPU的等待。”谭弘表示，“当前规模算力集群可用度在50%左右。”

“这时Checkpoint机制的重要性就体现出来。这就像我们打一个3A游戏，很多3A游戏不止有一种结局。一旦剧情发展不如人意，我们可以随时退回到过去某一个存档（Checkpoint），从而经过不断的尝试，最终达到我们的目标。Checkpoint可以使我们提高训练推进的效率，不用每次都回到起始点重来，能显著节省训练成本。”他说道。

从推理阶段看，一大痛点则是模型参数量的急剧上升。“当下，主流模型的参数量规模都在以指数级向上增长，而GPU所配的存储容量的增长是线性的，两者的差距会越来越大。为了实现更长的上下文推理，连续的记忆/防止幻觉，降低每Token的成本，GPU需要把KV Cache下放到eSSD，这需要更大容量、更高性能的eSSD，以支撑海量Token的吞吐。”

总结而言，谭弘认为，在系统层面，存储带宽限制了算力的有效利用，存储和不同GPU之间仍然存在壁垒。

eSSD可有效突破AI训练和推理瓶颈

尽管存储业界在持续提升带宽能力，如从SSD到更快的DDR再到HBM，带宽正在不断地拓宽。但此外还有怎样的解决办法和方向，来突破AI训练和推理的瓶颈呢？

“在训练阶段可以使用大容量的单盘的QLC eSSD来存放Checkpoint，可显著提升GPU的利用率，减少等待时间，降低训练成本。”谭弘表示，QLC eSSD规模部署已趋于成熟，在特定场景下的写入性能和写入耐久性已经非常接近TLC eSSD。

从推理场景看，谭弘表示，AI推理已经引发了存储的进一步分层，业界已经推出了Token的缓存层和性能的缓冲层。首先把KV Cache下放到eSSD作为一个缓冲。其次在性能缓冲层上，实现在多用户、多模型切换场景下对数据进行预读等，从而提升I/O的速度，减少等待。“至此，企业级的eSSD已经承担起上下文状态的管理、查阅知识的工作，eSSD将不仅是数据仓库，也将成为存算协同的数据引擎。”他说道。

eSSD又如何通过持续的技术创新突破存储瓶颈，释放算力潜能？谭弘表示，在长文本推理和KV Cache方面，需要极高的读取性能，eSSD的接口不断升级，用更低延迟的控制器以减少CPU、GPU的等待。据了解，目前PCIe 5.0已经全面商用，PCIe 6.0预计2027-2028年进入企业级市场，PCIe 7.0产品的研发也已经在路上。此外，RAG知识库模型的加载与热切换需要eSSD同时具备超大容量和更高的性能，与XPU直连，在eSSD和XPU之间直接传输数据，提高效率。“启动和多模态推理时，需要更加稳定的峰值读取性能，对eSSD来讲，在接口、性能、容量、生态协同、品质等方面的要求在不断的提升。”他说道。

最后，谭弘表示，长江存储作为国内一家3D NAND研发和制造的半导体企业，经过多年的发展，已经从存储颗粒晶圆供应商成长为一家全面提供存储方案的制造公司，有能力提供全场景的存储解决方案。在产品和技术方面，长存将持续加大投入的力度，聚焦企业级存储核心的需求，通过创新和工艺优化，不断提升产品在可靠性、容量及性能上的表现；在供应链方面，则将始终与合作伙伴合作，提供更加稳定、可持续的供应链支持。