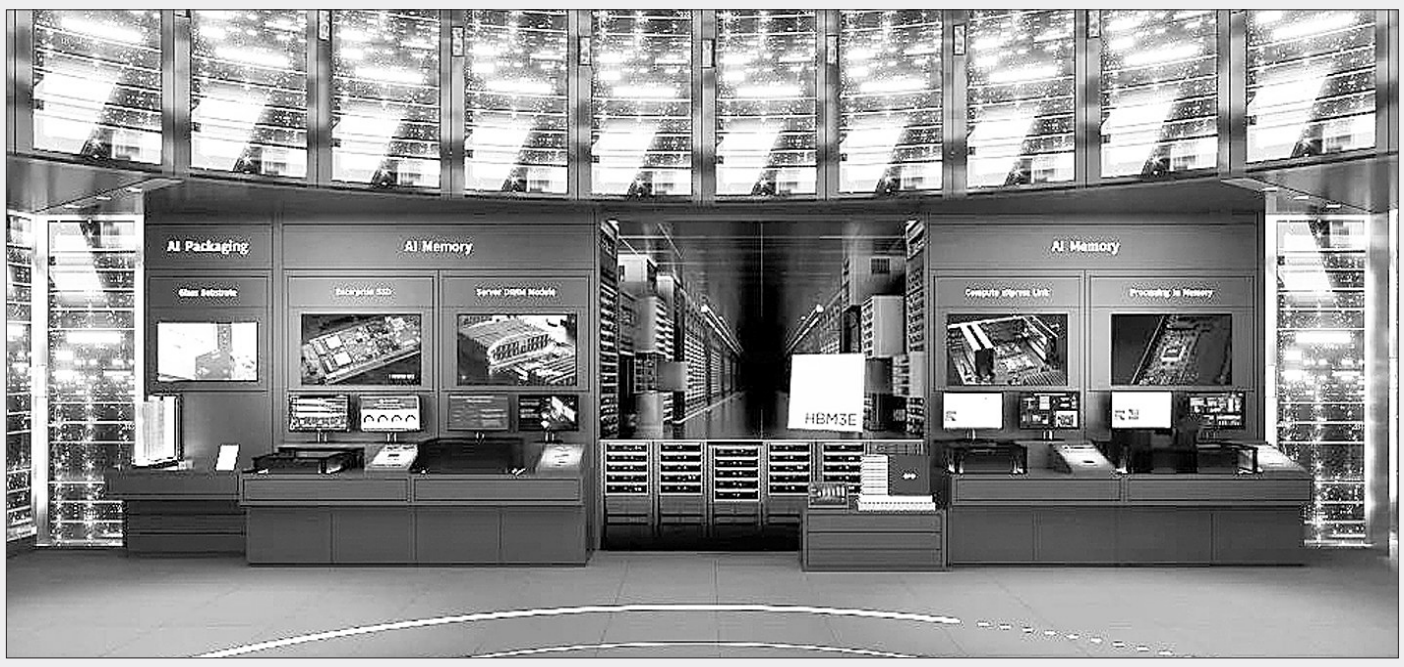
存储芯片市场迎新一轮涨价潮



本报记者 许子皓

近期，存储芯片市场迎来了新一轮的涨价潮，多家存储头部企业纷纷宣布提高部分产品报价。国内存储企业也紧跟步伐，纷纷上调提货价格。在1个月之前，业界还普遍预估存储芯片价格有望在今年6月份或7月份启动上涨，如今上涨提前到来，存储市场是否进入了复苏期？

提价底气源自供需两端双重激励

2025年5月初，三星与主要客户就提高DRAM芯片售价达成一致，并表示半导体行业整体DRAM价格都在上涨，这也是三星近一年来首次上调DRAM价格，具体比例因客户而异，但平均上调率已确定，DDR4将上调20%，DDR5上调约5%；闪迪宣布自4月1日起所有产品提价超过10%；美光不仅表示将针对新订单提高价格，平均涨幅约11%，还在3月25日发布涨价函，预计此次涨价幅度将在10%~15%；SK海力士也表示，其DDR5和eMMC存储芯片现货价格预计将上涨12%左右。

国内存储企业也紧跟步伐，纷纷上调提货价格。长江存储旗下零售品牌致态宣布将于4月起上调提货价格，涨幅可能超过10%。这一轮价格上涨的速度和幅度均超出了业内原先的预期。专家表示，这一轮涨价潮的主要原因在于存储芯片市场供需两端发生了明显变化。供给端的减产策略与需求端的新兴需求崛起，共同推动了市场的复苏与价格的上涨。

目前，各大企业减产策略成效显著。2025年年初，由于智能手机和笔记本电脑等核心消费性电子产品出货量持续低迷，供应商对2025年上半年需求看法并不乐观，为避免进一步削弱利润率，美光、铠侠、闪迪、三星和SK 海力士等NAND Flash大厂纷纷启动减产计划。主要通过降低2025年稼动率和延后制程升级等方式达成减产目的。

以美光为例，其在2024财年净亏损达50.66亿美元，为了应对这一困境，美光从2024年9月开始削减NAND Flash闪存产量，并计划将减产幅度扩大至30%。三星也在2024年11月宣布减产，目标是将NAND Flash芯片产量减少15%~20%。这些减产行动使得市场上NAND Flash的供应量得到有效控制，为产品价格反弹铺垫了基础。随着价格触底、市场开始补库存，模组厂和OEM厂开始加大采购，Enterprise SSD回温更带动高端Wafer产品需求。集邦最新调研数据显示，预计2025年第二季度，NAND Flash价格将比第一季度上涨不超过5%；3D NAND Wafers（多层垂直堆叠闪存晶圆）价格将环比上涨10%~15%；Client SSD（消费级固态硬盘）价格将环比上涨3%~8%。

此外，AI应用的“火力全开”成为驱动存储芯片市场需求快速提升的关键因素。随着AI技术在各领域的深入应用，对存储芯片的需求呈现出爆发式增长。特别是在AI服务器领域，需求增长尤为显著。2024年第二季度全球服务器市场收入达到了454.22亿美元，与去年同期相比增长了35%，AI服务器在整体服务器市场中的占比也持续攀升，现已接近30%。

AI服务器是智算中心提供AI计算能力的核心硬件，专为满足AI计算任务的高性能需求而设计。AI大模型需要用海量的数据进行训练、推理等，对数据进行深度挖掘后，再推送给通用数据中心，数据规模、数据调配工序将呈现爆发式增长，数据中心内部流量传输将更加密集，因此，对存储产品的需求也越来越高。

是德科技大中华区高速数字市场部经理李坚表示：“原来的人工智能做的是小算力、小模型，后来变成了中算力、中模型，而今天行业内做的是真正的大模型、大算力，大算力的一个非常基础的要求就是大带宽、大量的数据交换。存储方面，数据中心一般使用的是DDR4、DDR5系列产品，速率在8.4GT/s。我们预计未来会上升到DDR6或DDR7。此外，我们很可能会使用HBM3、HBM3E或HBM4。和今天的DDR产品相比，会有一个数量级的提升。”

不仅如此，AI在智能手机、个人电脑、智能穿戴等领域的加速渗透，也进一步推动了对更高容量和更高性能存储芯片的需求。目前AI手机的DRAM配置已提升至16GB，AI个人电脑设备的内存容量也普遍达到32GB。此外，智能汽车有望引入开源大模型，也将进一步提升存储需求，这些都成为了存储市场复苏的关键推动力。

存储市场周期性与结构性变化交织

回顾过去几年，存储市场历经从下行周期到复苏的复杂历程。自2021年第三季度起，存储市场遭受重创，DRAM价格下跌57%，NAND价格同期下跌55%。这一时期，存储芯片产业陷入长达近两年的下行周期。市场下行主要源于供应商生产过剩，以及市场需求衰退，使得存储芯片市场供需失衡严重。加上全球经济环境的不确定性、消费电子市场的饱和，进一步冲击终端设备销量。2022年，全球智能手机出货量为12.1亿部，同比下降11.3%；全球PC出货量为2.92亿台，同比下降28.5%，创下自2009年以来的最大年度跌幅。

为扭转局面，各大存储芯片厂商纷纷减产。铠侠及美光率先在2022年第四季度启动减产，三星于2023年第二季度跟进。截至2023年9月，三星削减NAND产量，减产幅度提升到50%，减产领域集中在128层以下制程；SK海力士下半年再减少NAND Flash产量5%~10%；铠侠减少产能50%；美光NAND Flash产能减少30%。

从2023年下半年开始，市场出现积极变化。减产效应逐渐显现，终端需求也逐步回暖，存储芯片市场开始走出下行周期。从2023年10月到2023年年底，现货市场NAND Flash价格指数涨幅达40%。以三星、西部数据以及金士顿为代表的固态硬盘产品，价格显著回升，回升幅度普遍在100元至130元之间。需求端数据同样向好，2023年10月23日至11月3日期间，整体PC市场销量同比上升1.4%，平板市场销量同比增长13.5%，手机市场销量同比增长10.2%。

进入2024年，全球AI浪潮兴起，成为推动存储需求爆发的关键因素，叠加终端需求持续回暖，存储芯片市场加速复苏，行业景气度攀升，价格上涨幅度扩大。2024年第一季度，存储销售额同比增长86%。DRAM芯片合约价格上涨多达20%，NAND Flash上涨多达23%~28%。到2024年第二季度，DRAM合约价季涨幅上修至13%~18%，NAND Flash合约价季涨幅同步上修至约15%~20%，延续强劲势头。行业头部企业业绩大幅增长，三星2024年第一季度营业利润激增931%；SK海力士2024年第一季度盈利超150亿元，而去年同期该公司还亏损136亿元。

但到了2024年下半年，存储芯片市场又发生了动荡，从价格走势来看，在上半年普遍上涨之后，下半年市场走向急转直下，出现了明显的价格回落。自2024年第三季度起，存储价格逐季下滑，进入第四季度，这种跌势越发明显，以消费类存储产品为例，价格下跌幅度高达30%，企业级存储产品价格也出现10%~20%的下滑。三星、美光、SK海力士、西部数据等企业继续使用减产策略应对，使得DRAM和NAND市场价格连续下跌，尤其是成熟制程的DDR4和LPDDR4X产品，价格压力极为突出。但HBM乘着GPU的东风狂涨500%，市场份额不断扩大，LPDDR5x、DDR5等新一代存储芯片的应用也在一定程度上缓冲了整体市场的下行压力。

而长江存储、长鑫存储等国内厂商，凭借本土供应链响应优势以及在部分技术领域的突破，在国内市场份额逐步提升，同时积极拓展海外市场，通过差异化竞争策略，在全球存储芯片市场中分得一杯羹。

再到2025年，存储市场逐渐进入复苏期，专家认为，5年期间，存储市场既有传统周期性波动因素作用，又受到AI等新兴需求的影响，虽短期内市场波动复杂难辨，但从长期来看，存储市场可能不会完全遵循过往典型的周期性规律，而是进入周期性与结构性变化交织的新阶段，周期性特征或有所弱化或变形。

中国存储企业迎来发展新机遇

在存储芯片市场新一轮涨价潮以及行业格局变动的大背景下，中国存储企业迎来了诸多发展机遇，有望在市场份额拓展、技术创新和产业生态建设等方面实现突破，提升在全球存储芯片市场中的竞争力。

市场份额拓展层面，国际存储巨头因减产推高价格，为中国企业打开了价格竞争空间。长江存储凭借Xtacking技术，将存储单元与控制电路分离制造，通过混合键合技术解决信号干扰与散热难题，在提升芯片读写速度的同时，将生产成本降低约30%，这种“高性能+高性价比”组合使其在消费级SSD市场快速渗透，2024年全球NAND Flash市场份额从6%提升至9%。长鑫存储则聚焦DRAM领域，其17nm DDR4产品已通过三星、联想等头部客户认证，2024年DRAM出货量同比增长55%，在全球市场份额中突破5%。此外，国内厂商依托本土供应链响应优势，在华为、小米等终端品牌的本土化配套需求中占据先机，华为2024年发布的Mate70系列手机中，长江存储的UFS 3.1芯片搭载比例达60%，较2023年提升40%。

技术创新领域，兆易创新在Serial NOR Flash市场持续领跑，2024年通过近存计算技术研发，将芯片延迟降低至5ns以下，其车规级产品已进入比亚迪、蔚来等车企的供应链。佰维存储则在先进封测领域发力，投资12亿元建设的12英寸晶圆级封测产线投产，可实现HBM3E级别的256层芯片堆叠，2024年研发投入占比达18%，推动其AI服务器存储解决方案收入同比增长210%。

存储芯片市场的新一轮涨价潮，既是市场供需关系调整的结果，也是行业技术发展和应用拓展的体现，为存储企业带来了拓展市场份额、推动技术创新和完善产业生态的新机遇，同时也促使企业加大研发投入，提升技术创新能力，加强与产业链上下游企业的合作，共同应对市场的不确定性。

大模型藏身小芯片

本报记者 张心怡

“给我生成一份审讯盗窃案件的笔录提纲。”指令输入笔记本电脑之后，DeepSeek 16B（160亿参数版）在毫秒间生成了一份包含基本信息、案件概述、权利告知、事实调查、其他重点事项、笔录确认、注意事项等一级标题，且每个一级标题都包含3~5个二级标题的笔录提纲。这是记者在第12届中国国际警用装备博览会的中星微展台看到的一幕。

今时今日，用大模型生成提纲已经是家常便饭，但这份笔录提纲的特别之处在于：它是在笔记本电脑没有联网的情况下生成。这意味着160亿参数的DeepSeek大模型，完全基于一枚嵌入在只有名片大小处理板的单芯片运行。

虽然联网的大模型能够基于云端的算力资源池实现更强大的功能，但也对计算和存储成本、网络条件有着较高的要求。而在城市感知、智能制造、智慧农业、智能交通等行业场景中，存在大量成本低、硬件配置相对简单却对千行百业的数智化升级起到关键作用的终端、边缘设备，比如摄像头、边缘盒子、车路协同设备等。如果此类设备能够基于嵌入式芯片调用大模型能力，将对企业、行业场景的提质增效起到关键作用。

此外，在机器人等涉及用户个人信息采集的场景中，也需要嵌入式芯片搭配离线语言大模型，在保证机器人与用户交互的同时，保护用户的数据安全。

嵌入式芯片承载离线大模型

“嵌入式芯片和云端芯片的设计思路不太一样。云端芯片追求极致的大算力，而前端嵌入式受到的制约条件非常多，能耗、发热、成本都要考虑到。”中星微智能研发中心总工程师周学武向《中国电子报》记者表示。

当前，嵌入式芯片能够承载的大模型一般在70亿参数规模。本次中星微展示的“星光智能五号”嵌入式AI芯片，能够运行160亿参数版本的DeepSeek大模型。为了让嵌入式芯片以尽可能高的效率处理多模态信息，中星微团队采用了多核异构的芯片架构，包括CPU、GPU、NPU，分别对应标量算力、矢量算力和张量算力。此外还有用于视频编解码的VPU、信息加解密的ECU，以及多核调度单元HCP（异构计算池）。

周学武表示，之所以选择这种架构，是为了模拟大脑兼具形象思维和逻辑思维的特点。

其中，对形象思维的模拟是基于“直觉式”的端到端计算，比如NPU或GPU能够直接输出对图片的识别结果。对于逻辑思维的模拟则主要基于CPU完成的“常识式”计算。

“把两种计算融合在一起，可以实现更高精度的识别。因为CPU的‘常识式’计算能够对可能产生的大模型幻觉进行纠正。”周学武说道。

另一个提升芯片运行大模型能力的设计，在于HCP。这一系统能够调动芯片中的20多个核心，并根据用户需求采用不同的策略调度算力，比如效率优先原则或者算力均衡原则，以寻求在有限的条件下实现最佳的性能。

基于嵌入式芯片，终端可以在不联网的情况下使用离线大模型，以满足部分对信息安全有较高要求的场景，以及机器人等涉及用户语音等个人数据的场景。

“未来5到10年，会有大量的机器人应用嵌入式芯片。目前机器人的发展重点是运动控制，就是机器人怎么走得稳、怎么行动敏捷。但决策、思考能力以及语音对话能力还需要通过网络实现，要先采集用户的语音，通过网络传到云端，云端给解析好了再传回来形成指令。这存在实时性响应和用户隐私的问题。如果基于嵌入式芯片构建机器人大脑，就可以保护用户数据，也能提升交互的实时性。”周学武说道。

算法定制与链路优化不可少

行业场景中，摄像头、工控盒子、车路协同设备等端、边侧终端，具有部署体量大、成本敏感、工作环境适应性较强等特点。要让此类终端用上大模型，既需要轻量化、易部署的芯片，也需要做好算法的定制与数据链路的优化。

在中星微展台，记者看到了一个连接了笔记本电脑的摄像头，在笔记本电脑搜索“戴安全帽的工人”，显示屏立刻出现了两天前展台搭建时的施工画面。这一过程是通过端（摄像头）、边（边缘盒子）、云（云网络和云平台）协同完成。其中，端侧和边侧除了部署中星微的嵌入式芯片，还部署了将视频“切”成图片再打上标签的算法。

具体来说，对于摄像头正在录制或者录好的存量视频，首先抽取关键帧或者关键数据做成图片，再根据矢量算法提取图片的关键点，成为包含索引信息和特征向量的标签。

在这一过程中，1G的视频可以抽取2M的关键帧图片，2M的图片再提取出2KB的标签，在用指令检索时，端侧和边缘侧终端会检索出标签对应的图片，也就是在KB级的数据中搜索，从而显著提升了搜索和解析效率。而搜索结果会发送给云端的大模型进行核实比对。

“搜索到标签之后，能够找到标签对应的图片，由于图片属性包含时间戳和相应摄像机的IP位置，就能对应出是哪路摄像机在哪个时间点拍到了目标，并回溯到视频，从而了解事件的前因后果。这样就达到了高效快速的效果。”周学武说道。他表示，在端边预处理一发到云端比对一回传端边比对结果的过程中，大模型不断训练、不断学习，之后检索结果会越来越准确，更贴近用户多样化的检索需求。

黄仁勋：物理AI是机器人革命的基石

本报讯 5月19日，英伟达创始人兼CEO黄仁勋在Computex2025（台北国际电脑展）登台亮相。在长达近100分钟的主旨演讲中，“AI”和“机器人”成为黄仁勋高频提及的关键词。他指出，物理AI是机器人革命的基石，物理AI与机器人技术将开启新一轮工业革命。

当前，AI正在推动新一轮工业革命——这是一场由AI工厂驱动的革命。“英伟达已不仅是一家科技公司，更是全球基础设施的核心提供者。”黄仁勋表示，历史上没有任何一家科技公司会一次性提前公布长达五年的发展路线图。但若无法提前规划电力、土地、资金等基础设施，将无法去建设AI数据中心。因此，英伟达详细公布了路线图，让全球合作伙伴能提前布局。

这些基础设施与第一次工业革命时的电力网络类似。当时，通用电气、西门子等公司意识到电力是新技术，必须建设配套基础设施。同样，信息时代的基础设施是互联网，而今天的新基础设施是“智能”（Intelligence）。黄仁勋表示，尽管现在说“智能基础设施”可能令人困惑，但十年后AI将融入万物，成为社会基础设施的核心。而这一基础设施需要“AI工厂”。与主要用于存储和处理数据的传统数据中心不同，AI工厂输入能源，输出极具价值的“Tokens”。未来，企业将像汇报工厂产量一样汇报Token生成量。

黄仁勋认为，世界已发生了根本性改变。1993年，人们预估芯片市场规模为3亿美元；如今，数据中心市场规模达万亿美元，而AI工厂和AI基础设施将达数万亿美元。

黄仁勋表示，过去十多年，AI从感知（图像识别、语音识别）发展到生成式AI（文本生成、图像生成）。如今，AI正迈向“推理AI”——具备解决新问题的能力，如链式思考、树状思考等。这种AI能模拟人类思维过程，使用工具并与其他代理式AI（Agentic AI）进行协作，成为“数字员工”。到2030年，全球劳动力短缺或达数千万，代理式AI将填补这一缺口。

“人工智能发展的下一波浪潮是物理AI。”黄仁勋表示，人工智能需理解惯性、摩擦力、因果关系等物理规律，例如预测一个球滚入车底后的运动轨迹。目前，英伟达与DeepMind和迪士尼合作开发了高保真开源物理引擎NVIDIA Newton，以此来训练机器人。通过仿真环境，AI可在虚拟世界中学习技能，再迁移至实体机器人。

黄仁勋认为，机器人技术的未来依赖仿真与AI。就机器人技术而言，人类的演示非常重要，人类可以向机器人示范如何执行任务，但人工示范不具备可扩展性。机器人制造商面临的主要挑战是缺乏大规模、真实和合成的数据来训练模型。借助人工智能，机器人可以从人类示范中进行泛化。这本质上是使用人工智能来扩展、放大人类示范过程中收集的数据量，以训练人工智能模型。

（杨鹏岳）

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号