存储全产业链发力AI时代

本报记者 许子皓

AI带动的存储价格疯涨也反映到了3月27日在深圳举办的CFMS | MemoryS 2026上：现场参会人员规模比去年翻番，会议所在的酒店人满为患不说，连酒店工作人员都在问来宾“存储还会缺多久”，因为工作人员正在考虑拆卖旧电脑的内存条。

随着存储“超级周期”的到来，业界已经从比拼“谁能拿到货”“谁的存货多”，转向“如何用好存货度过周期”以及解决AI核心痛点，存储产业正在新的竞争中重塑，技术创新与场景适配能力成为企业穿越周期的核心竞争力。

图为CFMS | MemoryS 2026现场

从“容器”到AI发展核心“引擎”

随着生成式AI从训练阶段全面转向推理落地，存储产业正经历一场深刻的范式转移，过去作为数据“容器”的存储设备，如今已升级为决定AI生产效率的核心“引擎”，其战略价值被提升至前所未有的高度。

WSTS预计，2026年全球半导体市场规模将增长26.3%，达到9750亿美元，距离万亿美元大关仅一步之遥，其中存储领域投资规模远超其他芯片类型。

深圳市闪存市场资讯有限公司总经理邰炜表示：“我们预计2026年全球存储市场规模将突破6000亿美元，数倍于以往的任何时期。AI服务器在整体服务器出货中占比将突破20%，单台AI训练服务器所需SSD容量已超256GB，推理服务器达70GB~100GB，较2023年增长2.5~3倍。”他进一步提出，这一爆发式增长的核心驱动力，是AI推理场景对存储的结构性需求变革。传统数据生成方式为“用户数量×设备数量×数据精度×使用时间”，而AI时代则转变为“大模型个数×大模型参数×多模态×再训练时间×设备”，海量大模型的多模态参数不间断产生的数据规模，带来了远超想象的存储空间需求。

这种需求变革直接催生了存储产业的结构性调整。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：“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。”当前，GPU集群可用度仅约50%，核心症结就在于存储带宽瓶颈，海量模型参数和上下文需通过狭窄的“存储漏斗”传输，导致算力无法充分释放。

三星电子执行副总裁兼方案平台开发团队负责人张实完则指出，人工智能正经历从“生成式AI”向“物理AI（Physical AI）”的深刻跨越。物理AI高度依赖高分辨率视频、3D点云等连续时间序列数据的庞大吞吐，“高性能存储已不再是可有可无的选项，而是决定系统决策效率与规模的核心基石。”

市场研究数据印证了这一趋势，Fgi预测2024—2029年企业级SSD复合增长率达34.9%；摩根大通指出，2024—2027年AI服务器用企业级SSD复合增长率将高达71%。

联芸科技董事长方小玲进一步补充，随着AI应用从训练迈向推理，存储的核心定位正发生根本性转变：在训练时代，存储是算力的“仓库”；而在推理时代，存储正成为算力的“加速器”。传统存储架构以大文件顺序读写为优化目标，而AI推理场景中，小文件随机读写占比高达60%以上，且对延迟稳定性要求极高，KV Cache的高频访问、模型参数的动态加载等场景，都需要存储系统具备微秒级延迟和百万级IOPS的性能表现。

在AI大模型多模态发展、智能体应用普及的浪潮下，存储的性能、架构与生态体系迎来全方位重构，全球存储产业链正以集体发力的姿态，抢占AI时代的产业新高地。

多维创新破解AI存储痛点

面对AI推理带来的随机读写压力、长上下文存储需求、多模态数据爆发等核心痛点，全球存储企业展开了全方位的技术创新，从接口协议、产品形态到软件算法，构建起多层次的解决方案体系，让存储不再成为AI发展的瓶颈。

接口协议的代际升级成为性能突破的关键抓手。PCIe 6.0已进入规模商用阶段，而PCIe 7.0也已紧锣密鼓地进行研发。

Cadence亚太区IP与生态系统销售群资深总监陈慧新介绍，PCIe作为关键互联技术，在主机与加速器、网络接口、NVMe SSD等环节扮演核心角色。面对带宽与延迟的双重挑战，PCIe 6.0采用PAM4信号与轻量级前向纠错技术，单lane速率达64GT/s，而PCIe 7.0将速率翻倍至128GT/s，为AI存储提供了充足的带宽支撑。

三星在现场带来的PCIe 6.0固态硬盘PM1763，在25W功耗限制下实现了2倍性能提升和1.5倍能效优化；长江存储则发布了PCIe Gen5系列企业级eSSD，其中PE522顺序读达14GB/s，随机读3400K IOPS，写延迟低至5μs，为AI推理提供了高速响应保障。

存储分层架构的重构成为效率优化的核心路径。针对KV Cache在长上下文推理中爆发式增长的存储需求，行业普遍采用“内存+ SSD”的分层缓存策略。

铠侠SSD首席技术执行官福田浩一提出，SSD已形成四大发展方向：KV Cache扩展、NVIDIA Storage-Next适配、高容量QLC方案及HDD替代，为此铠侠推出企业级CM9系列CMX版，以25.6TB容量和3DWPD混合耐久度，成为大规模推理环境的优质选择。

近期Agent应用爆发带来了Token消耗的剧增，数据交互频率呈指数级上升，温数据和热数据的占比也显著提升，具备高密度和低成本优势的QLC NAND成为承载海量AI数据的首选介质，然而其固有的耐用性低和随机写入性能弱等短板，却限制了其规模化应用。

平头哥半导体在存储架构创新上展现出独特优势，其提出的ZNS+QLC技术组合成为破解AI存储成本与性能平衡难题的关键方案。

平头哥半导体产品总监周冠锋在演讲中表示，ZNS（分区命名空间）技术通过优化数据写入策略，可降低写放大、减少垃圾回收频率，显著提升QLC SSD的使用寿命和性能稳定性，解决了QLC介质耐用性短板和性能不可预测的行业痛点。周冠锋指出：“平头哥推出的镇岳510企业级SSD主控芯片与上层存储系统通过‘原生设计、接口规范、软硬协同、软件定义’四大支柱，成功打通QLC主流介质从理论优势到大规模商用的‘最后一公里’，标志着企业级存储正式迈入高性能、大容量、低成本的新时代。”

端云协同与软硬融合成为场景落地的关键支撑。英特尔中国区技术部总经理高宇提出在AI技术飞速发展的当下，智能体凭借强大的复杂任务处理能力成为市场焦点，不仅让用户体验到AI拆解目标、调度工具、输出结果的高效性，更推动其在办公、数据分析等领域广泛应用。但与此同时，智能体的大规模应用也暴露出高算力与Token消耗、云端限流、隐私安全三大核心痛点，动辄数千万甚至数亿的月度Token消耗让用户面临高额成本，云端流量拥堵导致使用体验下降，数据上云也让个人隐私与敏感信息暴露在安全风险中。在此背景下，英特尔提出云与端侧结合的混合部署策略，成为破解智能体发展难题、推动其在AIPC上落地的关键方向。

针对智能体的部署需求，英特尔梳理出四种模式，其中纯云端部署存在成本与隐私短板，纯本地部署则面临大模型长窗口推理需求对计算设备提出的极高挑战，而协同协作（接力方式）、云端与端侧双主力模型智能决策这两种混合部署模式，成为英特尔为AIPC量身打造的科学方案。协同协作模式以云上模型为主力，根据端侧算力与本地AI能力智能下发任务，让端侧完成适配性工作，既节省云端算力与Token，又实现隐私数据本地处理；双主力模型模式则在端侧部署决策系统，结合上下文情景判断任务分发方向，灵活调用云端超强算力与端侧够用算力，实现资源的最优配置。

秉承“一切为了存储”理念的江波龙则强调以“集成存储探索端侧AI”。江波龙董事长、总经理蔡华波认为，目前AI哄抬了存储产品价格，但云端AI的发展和基础建设的落地，也将快速推动端侧AI百花齐放，江波龙更愿意打造“存储产品的Foundry（代工厂）模式”， 结合工程、工艺、技术等综合能力，满足大容量、高速度、低延迟、小尺寸、甚至定制化端侧AI产品需求。

构建全场景智能存储新生态

全场景智能存储生态的构建，绝非单一企业能够独立完成，需要芯片设计、存储制造、终端应用、软件算法等多个产业链各环节的深度协同。

终端应用企业牵引生态场景落地。小鹏汽车嵌入式平台高级总监段志飞表示，车载存储需求已从单一容量要求转向与车型定位、算力方案匹配的梯度化定义，需要存储厂商深度参与平台定义与量产适配。

阿里云千问大模型高级产品解决方案架构师李彬指出，大模型从文本交互迈向全模态交互，对存储的容量、吞吐、延迟提出更高要求，需与存储企业联合优化数据处理流程。江波龙与AMD、紫光展锐联合开发，通过存储智能体与HLC技术，实现大模型本地高效部署，验证了终端与存储企业协同创新的价值。

软件算法企业激活生态效能潜力。腾讯操作系统内核资深技术专家曾敬翔分享，通过UMRD自适应画像、MGLRU冷热探测等算法创新，重构Linux SWAP分配器，使服务器内存利用率显著提升，为云存储生态提供软件优化方案。Solidigm推出的Luceta AI软件套件，利用生成式AI实现质量检测自动化，构建“硬件+软件”的一体化解决方案。

产业链协同成为生态构建的核心支撑。英特尔在魔搭社区上线AI PC专区，开放AI开发工具、参考代码和技能库，涵盖热搜摘要、OCR、语音识别等核心技能，举办开发者大赛鼓励社区共创；长江存储从存储颗粒供应商成长为全方案提供商，形成“芯片-模组-固件-生态”的全栈能力，在供应链上与合作伙伴紧密协作，提供稳定的产能支持；铠侠通过与NVIDIA、VMware等厂商深度合作，确保产品兼容主流AI平台，加速场景落地。平头哥半导体也积极参与生态共建，其镇岳510主控芯片已与多家存储模组厂商完成适配，支持从企业级SSD到消费级存储产品的全场景应用，为产业链提供灵活的底层芯片解决方案。

从芯片设计到终端应用，从硬件创新到软件优化，全场景智能存储生态的构建离不开产业链各环节的深度协同。AI时代的存储生态，需要上游芯片企业的技术突破、中游设备厂商的方案创新、下游应用企业的场景反馈，更需要全产业链的开放协作。

存储企业应与产业链伙伴携手共建AI存储生态，通过技术协同、资源共享与标准共建，共同推动各行业数字化转型，为数字经济发展提供坚实支撑。

本报记者 赵晨

当地时间3月24日，谷歌推出名为TurboQuant的内存压缩算法，称可在不损失准确性的情况下，将大型语言模型运行时的关键部分——KV Cache（键值缓存）的内存占用减少为原有的1/6。消息发布后，存储芯片板块头部企业股价集体跳水，全球主要存储企业市值损失合计超6200亿元。

在3月27日举办的CFMS | MemoryS 2026上，阿里云千问大模型高级产品解决方案架构师李彬肯定了该技术在模型推理过程中的价值，但他同时也指出，考虑到模型上下文长度和模型自身参数等方面的飞速发展，AI对存储需求持续增长的大趋势不会改变。

李彬深入剖析了大模型应用从简单的对话助手向复杂Agent（智能体）转变的路径，并重点阐述了这一过程中存储技术面临的全新挑战与机遇。一个由全模态大模型驱动、7×24小时运行的智能体时代正在到来，存储技术正面临着前所未有的挑战，存储产业或将迎来由大模型驱动的结构性增长。

图为阿里云千问大模型高级产品解决方案架构师李彬发表演讲

应用层演进：智能体全天候运行改写存力峰谷规律

AI应用的发展正经历着质的飞跃。最初，我们仅通过Chatbox（聊天框）与模型进行简单的问答；随后发展为Copilot（副驾驶）模式，辅助人类工作；而当前的最新趋势则是General Agent（通用智能体）的崛起。

作为通用智能体，“龙虾”不仅能进行任务规划，更具备了长期记忆和远程执行能力。

李彬指出，这意味着AI不再局限于人类的“工作时间”，而是能够以7×24小时不间断地运行，无论是白天还是夜晚，都能通过远程调用工具自动完成任务。这种全天候的运行模式，彻底改变了传统AI应用的算力、存力的波峰波谷规律，使算力与存储负载由日间峰值转向全天候均衡分布，对存储系统整体利用率与运维的连续性提出了更高的要求。

技术层突破：参数增长和架构创新呼唤存储吞吐效率升级

在模型技术层面，Qwen大模型从最初的2T训练数据发展到如今的45T数据，参数量级的飞速增长伴随着架构创新，对存储吞吐效率的要求呈指数级上升。

李彬强调，MoE（混合专家模型）架构的广泛应用虽然在一定程度上降低了对算力的需求，但并未减少对存储的依赖，反而因为参数量向TB级甚至10TB级发展，对显存和存储提出了更高要求。为了应对这一挑战，行业正在探索KV Cache压缩技术。谷歌的研究指出，通过缓存压缩技术，推理过程中的KV Cache消耗可降至1/6，这为解决长上下文处理中的高显存占用问题提供了重要思路。

此外，针对端侧推理场景，还可以利用Flash存储辅助显存的技术，解决边缘设备显存有限、大模型无法完整加载的问题，从而在不牺牲推理性能的前提下，显著降低部署成本。该技术进一步打开了大模型在端侧落地的可能性。

全模态融合：视频理解与生成驱动存储需求爆发

李彬分析道，当下，模型已不再局限于纯文本交互，而是向包含图像、语音、视频的全模态方向发展。特别是在自动驾驶和具身智能领域，模型需要具备快速定位物体、理解复杂场景的能力。

例如，模型不仅需要识别图像，还要通过逻辑推理找出图片中的不同之处。这种从“感知”到“认知”的跨越，使得原本沉睡的监控视频、用户相册、自动驾驶影像等历史沉睡数据因模型可解析而重获价值，带来海量非结构化视频数据的存储、索引与加速访问需求。

同时，AI生成的短视频、短剧等内容占比已极高，其生成过程本身产生海量高质量视频数据，后续还需配套存储、二次理解、检索与再编辑，形成“生成-存储-理解-再生成”正向循环，持续放大底层存储系统的容量与带宽压力。