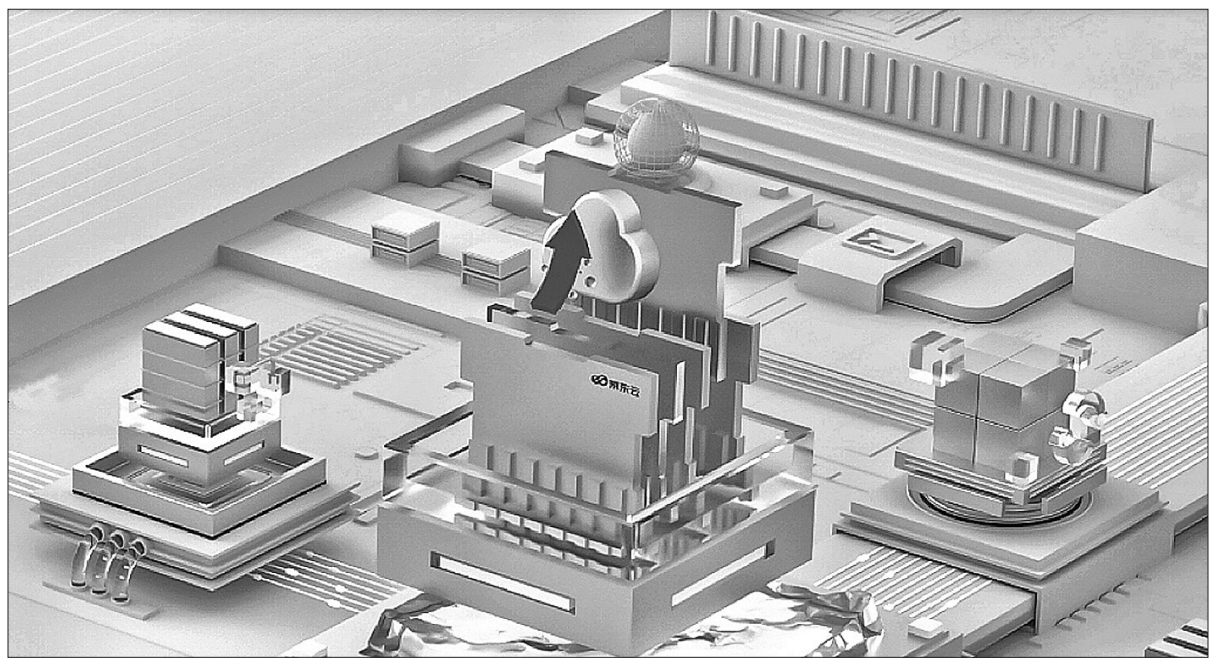
“存算分离”按下加速键

本报记者 许子皓



当前，在各大AI大模型激烈竞争的浪潮下，大模型参数正在呈指数级激增、上下文长度直指百万Token。IDC预计，2025年全球数据量将逼近175ZB大关。庞大的数据量让传统存算一体架构“紧耦合”的固有瓶颈日益凸显，数据存储与计算资源捆绑配置，要么“大马拉小车”造成资源闲置，要么难以应对峰值负载，成为了企业数字化转型的核心难题。

在此背景下，存算分离技术迎来产业化与规模化的双重爆发，不仅破解了困扰行业多年的“内存墙”难题，更重构了算力基础设施的配置逻辑。

打破“捆绑” 重构算力配置逻辑

“过去我们的算力资源配置，就像买手机必须捆绑固定套餐，不管用不用得上，都得全额付费。”业内人士向记者表示，在传统存算一体架构下，数据存储与计算单元紧密绑定，企业为应对业务峰值，往往需要按最高负载配置硬件，导致非峰值时段资源利用率不足，运维成本居高不下。

存算分离的核心正是打破这种捆绑关系，实现存储与计算的“物理解耦、逻辑协同”，也就是将数据存储功能与计算功能从物理设备层面拆分，通过高速网络实现存储资源池与计算资源池的独立部署、弹性调度，改变传统紧耦合架构中存储与计算绑定扩容的固有模式。

这种架构革新的核心价值，在于破解传统架构下三大核心痛点：一是资源浪费，避免算力闲置而存储不足或存储冗余而算力短缺的失衡问题；二是扩展受限，传统紧耦合架构扩容需整体升级设备，难以适配PB级、EB级数据增长需求；三是安全隐患，数据与算力绑定存储，易导致数据泄露、故障传导等风险。

分离之后的存储层可根据数据量按需扩容，轻松承载EB级海量数据；计算层依托Serverless等架构，随业务负载弹性伸缩，用完即释放，避免资源浪费；再借助智能IP广域网（AI WAN）、CXL等技术，保障跨节点数据传输的低延迟与高可靠。

从技术逻辑来看，存算分离的落地需三大核心支撑：一是高速网络传输，RDMA（远程直接内存访问）、硅光互连等技术的成熟应用，将存储与计算节点间的传输延迟压缩至微秒级，为资源解耦提供基础；二是弹性资源调度，软件定义存储技术的普及，实现存储资源的动态分配与按需扩容，适配不同场景的算力需求波动；三是高可靠冗余机制，通过分布式架构与创新EC冗余技术，在降低存储冗余成本的同时，保障数据可靠性。这三大支撑技术在2025年的全面成熟，推动了存算分离从高端场景向通用领域渗透。

中国电子企业协会电子信息产融合作工作委员会成员绿算技术高级副总裁黄飞表示，存算分离并非要取代此前火爆的存算一体技术，而是形成互补共生的格局。存算分离聚焦数据中心级、广域级的大规模资源调度，适合AI大模型训推、大数据分析等场景；存算一体则侧重端侧、边缘侧的本地化高效计算，比如车载终端、智能摄像头等设备，两者共同构成“端云协同”的算力基础设施体系。

破解行业长期痛点 持续拓展应用领域

当前，存算分离技术在核心介质、网络传输、调度算法等领域实现多项关键突破，推动技术从实验室走向规模化商用。

在技术层面，存算分离领域最显著的突破是摆脱对专用硬件的依赖，通过全栈软件优化实现通用硬件的高性能适配，破解长期以来“高性能必高成本”的行业痛点。

例如，京东云发布的云海AI存储解决方案，通过软件栈深度调优、支持超低冗余EC存储、支持全场景统一存储和存算分离技术，而这项存算分离技术可以将计算和存储解耦独立，存算资源独立调度，提高资源利用率和系统可靠性的同时，降低存储成本。

据了解，云海AI存储的存算分离技术架构，可以实现低至1.1x副本的冗余EC存储，节省整体基础设施成本30%以上。

绿算技术推出为G3级（Nvidia ICMS）定制的存算分离架构平台GP7000系列产品，该系列产品采用以太网闪存簇（EBOF）设计，单系统配备24个PCIe 5.0 NVMe U.2盘位，通过双主控板实现冗余。单机提供7000万IOPS、300GB/s带宽与20μs级延迟，性能较传统存储服务器提升17倍。整机功耗<900W，每GB/s带宽功耗仅3.1W，满足AI工厂的5倍能效目标，支持BlueField-3/4 DPU与Spectrum-X交换机，通过NVMe-oF/RoCEv2/GDS协议实现GPU直连。

高速网络传输技术的优化，是存算分离低延迟落地的核心保障。RDMA（远程直接内存访问）网络技术的深度优化，消除了数据在计算节点与存储节点间的搬运延迟，京东云、华为等企业的方案均已实现该技术的成熟应用。

在人工智能与大模型训推领域，生成式AI与大模型的规模化应用对算力与数据访问效率提出更高要求，存算分离技术通过“数据就地计算、算力动态调度”的核心特性，有效解决了传统架构下数据频繁搬运导致的训推效率低、成本高的问题，成为AI基础设施的核心支撑技术。

华为近期发布的TaurusDB是其新一代云原生数据库，主打“商业数据库的性能与可靠性，开源数据库的灵活与开放”。其基于自研的DFV分布式存储，采用计算与存储分离架构，完全兼容MySQL生态，让客户应用平滑迁移，同时通过计算存储分离机制，显著减少资源冗余，提升整体效率。

阿里达摩院基于分布式智能存储系统构建大模型训练数据中心，可支撑千亿参数大模型的高效训练。其核心逻辑正是存算分离，通过存储与计算资源的弹性调度，避免了传统架构下的资源浪费与数据搬运延迟，成为大模型研发过程中的重要技术支撑。

在金融科技领域，金融行业对数据安全性、业务连续性及实时性要求更高，存算分离技术凭借其高可用、弹性扩展及合规适配特性，在银行、证券等细分领域得到广泛落地，有效解决了传统架构下资源利用率低、节点重建效率低、业务抖动等痛点。

微众银行作为国内首家数字银行，2025年基于TDSQL数据库推出存算分离“Diskless架构”，应对数据规模从不到10PB激增至110PB以上、服务器数量增至2万台的业务挑战。该架构通过服务器去本地盘化、计算无状态化改造，将存储资源集中为远程存储池，计算节点仅保留CPU与内存，实现资源弹性分配。

京东云表示，某股份制银行通过部署京东云云海分布式存储系统，快速打通存力卡点，存储资源利用率提升3倍，综合成本降低50%。

行业发展面临挑战 “存算分离”前景光明

尽管存算分离在2025年取得显著进展，但行业发展仍面临不少挑战。记者采访了解到，技术层面，超远距离存算拉远场景下的算效优化、多协议兼容与异构资源调度的复杂度等问题，仍增加了企业迁移与运维成本；产业层面，行业标准不统一导致方案碎片化，跨厂商协同难度较大，产业链上下游技术适配成本偏高；安全层面，多节点协同场景下的全链路防护仍需加强，跨区域、跨行业数据传输的合规管控难度不小。

不过，行业对存算分离的未来充满信心。绿算技术预测，2026—2030年，存算分离将进入技术深度融合、产业生态成熟、应用场景泛化的新阶段。技术上，存算分离将与存算一体、云边协同等技术深度融合，CXL、AI WAN等技术的持续迭代将进一步优化远距离存算协同效能；产业上，行业标准将逐步统一，跨厂商协同成本将显著降低；应用上，存算分离将从互联网、金融向医疗、教育、工业制造等传统行业深度渗透；安全上，AI驱动的智能防护技术将广泛应用，推动数据要素安全流通。

随着技术创新的持续加码与生态体系的不断完善，存算分离将成为未来数字基础设施的核心架构模式，为全球数字经济高质量发展注入新动力，推动人工智能、大数据等新兴技术规模化应用。

消费“挖潜” 发展“提速”

（上接第1版）

在优化实施机制方面，优化资金分配方式，完善全链条实施细则，严厉打击骗补套补和“先涨后补”等违法违规行为。同时，落实全国统一大市场建设要求，对汽车报废更新、汽车置换更新、6类家电以旧换新、4类数码和智能产品购新，明确在全国范围内执行统一的补贴标准。

国务院发展研究中心市场经济研究所副所长魏际刚表示，2026年“两新”政策有两方面变化，一是政策支持范围体现了“消费提质”的鲜明导向，在家电领域“做减法”，在数码领域“做加法”，降低新型智能终端的渗透门槛；二是补贴标准从“定额补贴”转向“精准化、差异化设计”，比如汽车调整为按车价比例补贴，实现了“价高补多、价低补少”的精准支持，增强了政策的科学性与公平性。“‘两新’政策不仅将延续对扩大内需的支撑作用，更将在推动绿色转型、建设全国统一大市场等方面产生深远影响。”魏际刚表示。

加快新技术新模式 创新应用

自2025年以来，在政策支持、市场需求和技术进步的共同推动下，我国人工智能产业蓬勃发展，数百款AI终端新品密集发布。宇树科技人形机器人惊艳春晚，长虹发布首款治愈系AI TV，海尔推出搭载“AI之眼”智能感知系统的AI家电矩阵，联想秀出全球首款卷轴屏AI PC，小米亮出首款AI眼镜，科大讯飞录音笔和翻译机AI能力大升级，阿里巴巴首款AI眼镜开启预售，华为发布搭载鸿蒙AI的Mate 80手机，“豆包”手机引发热议……这些产品不再靠“黑科技”标签吸引尝鲜者，而是以“实用体验”赢得市场认可，不断释放多样化、差异化消费潜力，成为经济增长新亮点。

挖潜扩大有效需求，需要加强供需适配，扩大优质商品和服务供给。2025年11月，工业和信息化部等多部门发布《关于增强消费品供需适配性进一步促进消费的实施方案》（以下简称《实施方案》），鼓励开发家庭服务机器人、智能家电和人工智能手机、电脑、玩具、眼镜、脑机接口等人工智能终端，以智能产品为载体提供娱乐、健康、陪护等生活服务。

国家信息中心大数据发展部人工智能处副处长易成岐表示，强化人工智能融合赋能，从供给侧看，一是要持续强化技术供给，进一步提升人工智能的情感认知力、自主决策力、逻辑推理力、行动协作力、创造涌现力和超长记忆力。二是要丰富产品和服务供给，推动人工智能与低空飞行、增材制造、脑机接口等前沿技术深度融合，推动智能终端产品不断推陈出新。同时，也要充分发挥人工智能个性化和情感化的独特作用，丰富拓展个性消费、认知消费、情感消费等新型服务消费方式。

在中国企业管理研究会副理事长赵永辉看来，未来“人工智能+消费”将呈现三大趋势：一是消费全链条智能化，AI将全面嵌入设计、生产、流通、营销和服务环节，实现需求预测更精准、供给响应更高效；二是个性化与柔性化成为主流，基于AI的柔性制造和智能供应链将推动大规模定制走向商业化常态，消费者需求将更快速地转化为产品方案；三是场景融合不断深化，AI将推动线上线下深度融合，沉浸式、体验式、智能化的消费场景加速扩展，消费金融、数字服务等衍生业态同步壮大。

伴随数字技术的不断涌现，技术本身即成为消费的热点领域。未来一段时期，虚拟现实、数字孪生、6G、人形机器人等数字化的核心技术将孕育更为广阔的市场空间。

开拓新型消费 催生新机遇

近年来，我国消费市场繁荣兴旺，离不开“文旅有温度、银发有尊严、绿色有动力”的消费新模式新场景不断创新融合。新型消费，一方面成为激活内需潜能、培育经济增长新动能的关键抓手，另一方面在技术赋能、场景融合与需求深挖中获得了广阔发展空间。

文旅融合点燃消费新活力。2025年，我国文艺创作与表演销售收入同比增长17.3%，沉浸式、场景式业态创新推动文旅消费提质升级。以游戏、动漫等为代表的数字文化消费展现出较大潜力，数字文化服务销售收入同比增长16.6%。2026年，我国文旅消费将积极推广“文旅+科技”，打造“云游景区”“虚拟演艺”“数字文物展”等线上产品，同时线下落地“沉浸式剧本杀景区”“光影秀主题公园”等，让“文旅+”深度融合进一步发展释放文旅消费乘数效应。

中国文化产业协会沉浸式文旅产业专业委员会主任委员卜希霆表示，当前文化和旅游领域体现出“政策赋能、科技加持、场景落地、消费下沉、业态提质”五大核心特点，未来文旅发展需聚焦科技与文化深度融合、全域场景价值挖掘、数据要素价值释放。

当前，中国60岁以上人口将突破3亿，在“生存型养老”转向“品质型享老”的过程中，银发消费需求上升，智能监测设备、旅居养老等健康服务类消费规模正在不断拓展。2025年，老年人养护消费同比增长24.9%。《实施方案》明确指出，优化适老化产品供给。加强适老化产品研发设计，重点开发应用养老服务机器人、多功能护理床、健康监测设备等急需产品开展优质老年用品惠老助企行动，推进居家适老化改造，发布适老化产品和服务推广目录，鼓励电商平台、商超等设立银发消费版面或专区。

中国老龄产业协会专职驻会副秘书长王永春表示，发展银发经济，要优先围绕康养旅游、老年文体、智慧健康等形成可供给、可复制的产品与服务，依托基本养老服务网络，把“医养+文旅”“适老化改造+家居消费”等业态更紧密地嵌入社区与家庭生活场景。

随着“双碳”目标关键期的到来，开展绿色消费成为顺应消费升级趋势、培育消费新增长点的有力举措。数据显示，2024年至2025年我国实现汽车以旧换新1830万辆，其中新能源汽车占比近60%。实现家电以旧换新1.92亿台，其中一级能效（水效）占比达到了90%以上。《实施方案》指出，要促进绿色产品扩容迭代，适应绿色化低碳化消费趋势，提高消费品能效、水效限定值标准；鼓励新能源汽车、高能效家电、绿色建材家装等领域绿色低碳消费；制定智能家居互联互通国家标准，支持骨干企业联合开发全屋智能化绿色化解决方案。

中国社会科学院数量经济与技术经济研究所研究员周勇表示，绿色低碳生活方式具体表现为低碳出行、节约资源、绿色选购、循环利用等一系列行为模式，这需要一套由政府主导、市场协同、社会参与的复合型规则体系，涵盖经济、市场、法治和社会激励等方面，形成推广绿色低碳生活方式的“组合拳”。

这些由需求迭代与供给创新共振催生的新形态，不仅激活了市场潜力，更推动制造业向“按需智造”跃迁、服务业向“体验增值”深耕，更为经济高质量发展注入“需求牵引供给、供给创造需求”的强劲动能。

业内人士表示，拓展新型消费，还需在供给端加快新一代信息技术基础设施建设步伐，推进新型城市基础设施建设，优化数字技术应用基础环境；加快建设数字化社区，提升社区数字化管理水平；加快新技术新装备应用，推进数字技术等在医疗、教育、养老、旅游等领域应用。

阿里自研AI芯片 浮出水面

本报讯 记者**姬晓婷**报道：近日，阿里巴巴旗下平头哥官网悄然上线一款名为“真武810E”的高端AI芯片。至此，自2025年起备受关注的阿里自研AI芯片PPU（AI专用并行处理器）正式面世。据悉，该处理器已在阿里云实现多个万卡集群部署，服务了国家电网、中国科学院、小鹏汽车、新浪微博等400多家客户。

据平头哥官网介绍，“真武”PPU采用自研并行计算架构和片间互联技术，配合全栈自研软件栈，实现软硬件全自研。其内存为96G HBM2e，片间互联带宽达到700 GB/s，可应用于AI训练、AI推理和自动驾驶。阿里巴巴已将“真武”PPU大规模用于千问大模型的训练和推理，并结合阿里云完整的AI软件栈进行深度优化，为客户提供一体化产品和服务。

记者从阿里巴巴了解到，平头哥早在2020年就秘密启动了通用GPU芯片“真武810”的研发，并于2022年年底、2023年年初，完成了研发和场景验证。但在此次正式官网上线前，这款芯片的研发和验证几乎都是处在“只对内部开放”的状态。

2025年，市面上逐渐出现关于阿里自研PPU的讨论。有消息称，平头哥首代PPU的整体性能可对标英伟达H20，升级版本甚至在部分指标上超过A100；随后，央视《新闻联播》的画面中短暂出现了这颗芯片的关键规格：96GB HBM2e显存、700GB/s的片间互联带宽、PCIe5.0×16接口，以及400W的功耗水平。

这款芯片面世，体现了阿里巴巴通义实验室、阿里云、平头哥之间的通力合作。阿里巴巴正在将“通云哥”打造成一台AI超级计算机，力图基于平头哥全栈自研芯片、阿里云大规模云服务能力、“千问”开源模型三方面技术实力，在芯片架构、云平台架构和模型架构上协同创新，实现在阿里云上训练和调用大模型时的最高效率。

闪迪2026财年第二季度净利润 同比激增672%

本报讯 记者**许子皓**报道：近日，全球存储芯片领军企业闪迪（SanDisk）公布了2026财年第二季度财报。财报显示，公司本季度营收达30.3亿美元，同比增长61%；净利润为8.03亿美元，同比激增672%；经调整后每股收益（EPS）达6.20美元，同比暴涨404%。

据了解，数据中心业务成为闪迪本次业绩增长的核心引擎，AI需求成关键驱动力。本季度，闪迪数据中心业务销售额达4.4亿美元，环比增长64%，同比增长76%。

闪迪董事长兼首席执行官戴维·戈克勒在财报电话会议中表示，公司正处于AI基础设施广泛扩张的核心位置，随着AI工作负载的持续扩展，企业级SSD需求在整个生态系统中加速增长，尤其是推理环节显著推动了每次部署中NAND含量的大幅增加。目前，云超大规模企业、边缘和企业数据中心、OEM厂商等各类AI基础设施建设者均在采用闪迪产品，其技术已成为优化AI基础设施的关键支撑。全场景需求共振，产品与合作布局持续优化。

除数据中心业务外，边缘计算与消费级市场同样表现稳健。在边缘计算领域，PC、移动设备等终端因AI功能升级推动存储配置高端化，需求显著超过供应，业务营收16.78亿美元，环比增长21%，同比增长63%；消费级市场则向高价值产品倾斜，公司推出的Extreme Fit系列USB-C闪存盘等创新产品备受市场青睐，同时与绘儿乐、国际足联等知名品牌的授权合作持续落地，进一步巩固了细分市场的领导地位，营收9.07亿美元，环比增长39%，同比增长52%。

对于2026财年第三季度的业绩展望，闪迪预计非GAAP毛利率将达到65%至67%，经调整EPS指引区间为12美元至14美元，营收中位数预计达46亿美元，较市场预期高出50%以上。