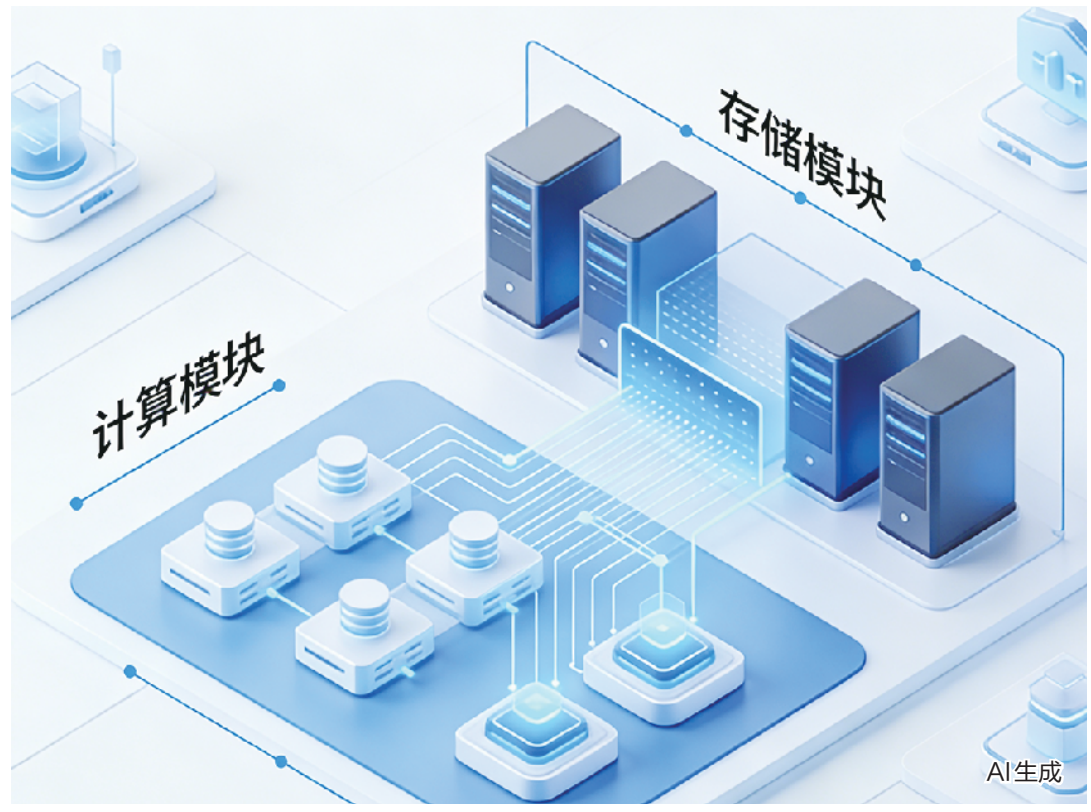
存算分离技术破解“高性能必高成本”痛点

当前,在各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。国际数据公司预计,2025年全球数据量将逼近175泽字节(ZB)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。



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,难以适配拍字节(PB)级、艾字节(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;计算层依托无服务器架构等,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)、计算快速链接(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:高速网络传输,远程直接内存访问(RDMA)、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;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;高可靠冗余机制,通过分布式架构与创新纠删码(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。

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,可以节省整体基础设施成本30%以上。

绿算技术推出的存算分离架构平台GP7000系列产品采用以太网闪存簇设计,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(每秒进行读写操作的次数)、300吉字节每秒(GB/s)带宽与20微秒级延迟,性能较传统存储服务器提升17倍。整机功耗小于900瓦,1GB/s带宽功耗仅为3.1瓦,满足AI工厂的5倍能效目标,通过相关协议实现图形处理器直连。

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AI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”。其基于自研的数据功能虚拟化分布式存储,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,计算节点仅保留中央处理器与内存,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%。

存算分离技术前景光明

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。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年至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。 据《中国电子报》 作者:许子皓

眨眼即发电,眼动操控轮椅更便捷

一名渐冻症患者该如何突破身体的桎梏,与外界重建联系?眼动追踪技术成为重要的辅助工具。然而,现实中,当患者试图通过传统眼动追踪设备控制轮椅移动时,沉重的头戴装置、缠绕的电源线缆与频繁的低电量提醒,往往在他们与自主行动之间竖起了一道“高墙”。

如何让眼球运动成为更便捷、更自由的交互方式?近日,青岛大学物理科学学院教授龙云泽团队与合作者,成功研发出全球首套轻量级自供能眼球追踪系统。该系统运行电力完全源于使用者眨眼时隐形眼镜与眼球摩擦产生的微弱电能。佩戴该系统,渐冻症患者可直接通过眼部动作控制轮椅等外部设备,无需再依赖外接电源。相关成果近日发表于《细胞报告物理科学》。

传统设备存在弊端

当前主流眼动追踪技术主要分为两类。

一类是基于图像识别的光学方案。其精度较高,但设备需集成红外光源与多摄像头,导致设备笨重、佩戴不适,且持续红外照射存在潜在生物安全风险,续航与运行稳定性也面临挑战。

另一类是基于生物电信号的传感方案。该方案通过在使用者眼周皮肤贴附电极,来检测眼球转动时产生的角膜-视网膜电位差变化。然而,该技术难以精确解析眼球的运动角度,且信号易受面部肌肉活动或环境电磁干扰,可靠性与精度有限。

由于现有成熟产品需外接电源,供能问题也成为传统眼动追踪设备使用的一大阻碍。

基于此,龙云泽团队提出了全新思路:让眼睛自己发电。他们开发出一种基于摩擦电纳米发电机单电极模式下的自供电眼动追踪系统。

这套系统具备诸多优势,它极致轻便,佩戴感与普通眼镜无异,且系统运行所需电力完全来自眼球运动,摆脱了电池束缚,实现“能量自给”。与此同时,它还支持高精度检测,可精确识别2度的最小眼偏角度,眼球运动方向检测准确率达99%。

眼睛里建“微型电站”

龙云泽介绍,这套系统的研发灵感,萌芽于一次课题组讨论。

“当时我们从一些前沿研究中了解到,摩擦纳米发电机可应用于监测眼球运动。一位常佩戴隐形眼镜的学生提出‘为什么不能直接把装置做到眼睛里面去’,这个大胆的设想一下子点亮了团队的思路。”龙云泽说,如同冬天脱毛衣时会产生“噼啪”静电,眨眼时眼球与隐形眼镜的摩擦也会产生微弱电荷。

基于摩擦纳米发电机原理,研究团队创新设计“隐形眼镜+框架眼镜”的双层协同系统。该系统通过摩擦起电与静电感应机制,可将眨眼时隐形眼镜与眼球摩擦产生的微弱机械能直接转化为电能,在简化结构的同时提升信号灵敏度。

“其工作机制犹如在眼睛里搭建一座‘微型电站’。”龙云泽说,这座“电站”内部同时集成了“发电机组”和“信号发射台”。其中,聚二甲基硅氧烷(PDMS)材料如同隐形眼镜贴附于使用者眼球表面,是一台“微型摩擦发电机”,每当使用者眨眼或转动眼球,眼球与PDMS材料之间便通过摩擦持续产生电荷。与此同时,一副镜片四周嵌有透明氧化铟锡(ITO)电极的眼镜则扮演着“信号发射台”的角色。透明电极通过静电感应精准捕捉电荷分布和变化,并将其实时转化成可识别的电信号,再经由控制电路传导到外部设备,最终实现精准操控。

要使这一系统稳定运行,材料选择尤为关键。摩擦层PDMS膜需兼具高透光率、高强度、生物相容性和疏水性,以确保在湿润的眼部环境中稳定工作并耐受摩擦。ITO电极则必须在保持高透光率的同时,具备优异的导电性以灵敏捕捉微弱的电场变化。

研发之路并非坦途。这一设想最初即便在团队内部也遭遇了质疑:依赖人体自身微弱生物能的设计,能否安全稳定运转?通过自主设计并迭代改进实验装置,经过反复测试,团队最终以扎实数据验证了路径的可行性。

拓展辅助交互新场景

该项目汇聚物理科学、材料科学、生物医学工程、微电子与控制工程等多领域专家的智慧,并得到香港科技大学教授范智勇团队的支持。“这是一个典型的学科交叉创新项目,没有跨学科的紧密协作和学校的支持,这个构想难以落地。”龙云泽说。

目前,该技术虽尚处于实验室阶段,但在医疗康复、消费电子等多个领域拥有广阔应用前景。这项技术不仅有望为行动受限群体打开一扇无障碍沟通的新窗口,更将为人类与智能设备交互的方式书写全新底层逻辑。

龙云泽介绍,在医疗康复领域,它有望为渐冻症等运动功能障碍患者带来突破性辅助解决方案,患者仅通过自然的眼球转动,即可控制轮椅,操作电脑等智能设备,从而显著提升生活自主性与尊严感。在消费电子领域,该技术与虚拟现实或增强现实设备的结合将催生真正“解放双手”的沉浸式交互体验。在太空作业、智能驾驶等对操作安全性、精确性要求极高的领域,它同样具备应用前景。

同时,该研究为“自供能生物电子学”这一前沿方向提供了技术支撑。未来,此类技术不仅有望让机器人获得更接近人类的感知能力,也可能帮助人类通过自供能电子器件来增强或修复身体功能,开辟人机融合的新路径。

这项技术要从实验室走向广泛应用,还需跨越产业化的一系列挑战,包括系统长期可靠性验证、医疗器械合规及与产业生态的融合等。“我们正积极与相关企业对接,探索合作路径,并积极推进产业化进程。”研究团队核心成员、青岛大学物理科学学院特聘教授张俊说。

据《科技日报》 作者:宋迎迎

▶ 科工前沿

我国星地激光通信架起“多车道高速桥”

近日,中国科学院空天信息创新研究院成功开展超100G星地激光通信业务化应用实验,通信速率达到120吉比特每秒(Gbps)。实验结果表明,通信链路稳定、下传数据质量优良。这标志着我国星地激光通信业务化应用能力迈上一个新台阶。

本次实验利用空天院塔县激光地面站自主研制500毫米口径星地激光通信系统与中科卫星科技集团有限公司研制的AIRSAT-02卫星,在卫星硬件无变化的情况下,通过卫星在轨软件重构,充分挖掘利用激光通信载荷的硬件潜能,将卫星激光通信载荷的能力从60Gbps提升至120Gbps。这不仅刷新了国内星地激光通信传输速率纪录,还破解了超高速星地激光通信链路难以快速建立、长时间稳定维持和高效可靠传输的难题。实验期间,星地之间成功实现了秒级捕获建链,建链成功率超过93%,最大连续通信时长108秒,获取数据量达12.656Tb,并成功处理出高质量遥感影像。

团队技术负责人、空天院高级工程师李亚林说:“若将星地激光通信比作在湍急的河流上架桥,10Gbps传输相当于铺设单车道桥梁,结构相对简单;而120Gbps传输则相当于建设多车道高速大桥,不仅要求架设速度快以实现快速连接,更要在多车道并行下保障极高的通行效率,其工程实现难度呈几何级增长。”

为实现海量数据星地超高速传输,团队聚力开展技术攻关,成功突破一系列核心技术瓶颈。让信号“收得稳”,通过优化强实时光学畸变校正算法,精准抑制大气湍流引起的高频扰动,确保了微弱激光信号的稳定跟踪与高效耦合;让信号“收得对”,深度应用信号损伤补偿技术,在数字域精准消除高速率信号的非线性畸变,同时通过抗大气湍流高效率激光通信空口协议,确保了数据的低误码率;让信号“收得快”,改进非稳态大气信道自适应传输控制策略,有效解决了快衰落信道下的数据吞吐瓶颈,最大化利用信道容量,保障了超高速率传输效率。 据《中国科学报》 作者:高雅丽