大数据与决策研究

（政策与技术跟踪专题）

2023 年第 37 期（总第 193 期）

广西壮族自治区信息中心
广西壮族自治区大数据研究院　　2023 年 10 月 9 日

编者按：近年来，以 ChatGPT 为代表的人工智能大模型爆发，随之而来的算力和能耗问题愈发凸显。存算一体将存储与计算有机融合，被视为突破人工智能（AI）算力瓶颈和大数据的关键技术。存算一体技术未来将在自动驾驶、AI 处理、物联网等云边端场景广泛应用，助力提升运算效率、降低系统功耗。本期将介绍存算一体技术相关情况。

本期要目

存算一体的概念与技术分类

一、存算一体技术背景

目前，主流芯片如 CPU、GPU 等均按照冯·诺依曼架构设计。冯氏架构以计算为中心，计算和存储分离，二者配合完成数据的存取与运算。由于处理器的设计以提升计算速度为主，存储则更注重容量提升和成本优化，“存”“算”之间存在性能失配，导致出现访存带宽低、时延长、功耗高等问题，也称为“存储墙”“功耗墙”。访存愈密集，“墙”的问题愈严重，算力提升愈困难。特别是以人工智能为代表的访存密集型应用快速崛起，访存时延和功耗开销无法忽视，计算架构的变革尤为迫切。

存算一体作为一种新型计算架构，有望解决传统冯·诺依曼架构下的“存储墙”“功耗墙”问题，受到国内外广泛关注。其核心是将存储与计算完全融合，有效克服冯·诺依曼架构瓶颈，并结合后摩尔时代先进封装、新型存储器件等技术，实现计算能效的数量级提升。

二、存算一体的概念

存算一体是指计算单元与存储单元融合，在实现数据存储的同时直接进行计算，以消除数据搬移带来的开销，极大提升运算效率，实现计算存储的高效节能。存算一体非常符合高访存、高并行的人工智能场景计算需求。

三、存算一体技术分类

根据存储与计算的距离远近，将广义存算一体的技术方案分为三类，分别是近存计算、存内处理、存内计算。

（一）近存计算（PNM）

近存计算通过芯片封装和板卡组装等方式，将存储单元和计算单元集成，增加访存带宽、减少数据搬移，提升整体计算效率。近存计算仍是存算分离架构，本质上计算操作由位于存储外部、独立的计算单元完成。其技术成熟度较高，主要包括存储上移、计算下移两种方式。其中，**存储上移**采用先进封装技术将存储器向 CPU、GPU 等处理器靠近，增加计算和存储间的链路数量，提供更高访存带宽；**计算下移**采用板卡集成技术将数据处理能力卸载到存储器，由近端处理器进行数据处理，有效减少存储器与远端处理器的数据搬移开销。

（二）存内处理（PIM）

存内处理是在芯片制造的过程中，将存和算集成在同一个晶粒中，使存储器本身具备一定算的能力。存内处理本质上仍是存算分离，相比于近存计算，“存”与“算”距离更近。当前存内处理方案大多在 DRAM（即动态随机存取内存）芯片中实现部分数据处理，可以提供大吞吐低延迟片上处理能力，可应用于语音识别、数据库索引搜索、基因匹配等场景。

（三）存内计算（CIM）

存内计算即狭义的存算一体。在芯片设计过程中，不再区分存储单元和计算单元，真正实现存算融合。存内计算本质是利用不同存储质的物理特性，对存储电路进行重新设计使其同时具备计算和存储能力，直接消除“存”“算”界限，使计算能效达到数量级提升的目标。存内计算包含模拟和数字两种实现方式，模拟存内计算适用于低精度、低功耗计算，如端侧可穿戴设备等；数字存内计算适用于高精度、功耗不敏感计算，如云边 AI 场景。

四、存算一体的优势

（一）更优的性能

存算一体通过减少存储和计算单元之间的数据搬运，可以大幅缩短系统响应时间，提高数据的处理速度，并且存储单元参与逻辑计算意味着可以在面积不变的情况下规模化增加计算核心数。存算一体架构的性能天花板远高于当前的传统方案，在特定领域算力可达 1000TOPS 以上。

（二）更高的能效

存算一体技术可以大幅降低数据传输的能量损耗，提升能效比。研究发现，存算一体芯片每瓦能提供的算力相比传统冯·诺依曼架构下的芯片可以达到 2—3 个数量级（>100 倍）的提升。

（三）更低的成本

解决大算力芯片的内存墙问题常采用 GDDR（即图形用

双倍数据传输率存储）或 HBM（即高带宽存储器）内存方案。但这些方案在冯·诺依曼架构下有明显的性能天花板，成本也较高。而采用存算一体架构做大算力 AI 芯片，则可以将芯片成本降低 50%—70%。

（来源：《达摩院 2023 十大科技趋势报告》《存算一体白皮书（2022 年）》）

存算一体产业情况

一、市场规模

目前存算一体产业仍处于起步阶段，尚未形成完整产业链生态。行业内公司的主要发展方向多集中在容易落地的小算力场景，长期来看，存算一体芯片具有巨大的发展潜力，将从小算力场景逐步扩展到整个 AI 芯片领域。有关机构测算，到 2025 年，基于存算一体技术的小算力芯片市场规模约为 125 亿元；到 2030 年，基于存算一体技术的中小算力芯片市场规模约为 1069 亿元，基于存算一体技术的大算力芯片市场规模约为 67 亿元，总市场规模约为 1136 亿元。

二、产业链分布

产业链上游的存算一体公司往往依赖于全栈自研，除流片与代工厂合作外，需要具备编译工具开发、算法开发等自研能力。在工具链和 EDA（即电子设计自动化）工具方面，国内外存算一体公司目前都选择在已成熟工具链和 EDA 软件上进行改造，专门针对存算一体技术的工具链和 EDA 设计软件还有待开发。而我国存算一体公司的存储技术大多来自对国外 IP 的购买，未来可能面临 IP 授权问题。

产业链中游的存算一体公司主要负责芯片设计，在芯片制造和封测方面需要与相关代工厂合作。

产业链下游主要是存算一体芯片应用，应用场景从麦克风、智能手表/环和 TWS 耳机等拓展到智能安防、移动终端、AR/VR 和自动驾驶。

图 1 存算一体产业链上游分类

图 2 存算一体产业链中游分类

图 3 存算一体产业链下游分类

三、国内代表性企业

存算一体的商业模式主要分为三种：IP 授权，定制/联合开发以及自主 SoC 芯片。国内存算一体初创公司倾向于 SoC 芯片研发，而有存储器背景的存算公司则更倾向于存储技术 IP 授权。目前在存算一体领域实现量产的公司有九天睿芯、智芯科和闪易半导体等，其余头部公司均已完成多次流片。国内存算一体芯片代表性企业见下表。

表 1 存算一体芯片代表性企业

企业	产品定位	产品技术亮点	所在省市
九天睿芯	神经拟态感存算一体架构芯片	采用前端模拟预处理（ASP）+模数转换（ADC）+模拟加速器（ADA）架构，将感、存、算集合为一体。	广东深圳
智芯科	面向边缘计算的大算存内计算 SoC	（1）先进的数据流架构; （2）由 SDK 驱动带来的通用性大算力模拟内存计算。	浙江杭州
闪易半导体	存算一体 AI 芯片	基于双向 Fowler-Nordheim 隧穿进行擦写的闪存忆阻器	上海市
阿里达摩院	基于 DRAM 的 3D 键合堆叠存算一体 AI 芯片	（1）存储芯片：采用异质集成嵌入式 DRAM; （2）计算芯片：流式定制化加速器架构; （3）封装：3D 混合键合技术。	浙江杭州
后摩智能	基于存算一体技术的大算力 AI 芯片	基于数字域存内计算的电路	江苏南京
知存科技	基于存算一体技术的人工智能芯片	使用 Flash 存储器同时完成神经网络的存储和运算	北京市

企业	产品定位	产品技术亮点	所在省市
苹芯科技	存算一体芯片与非冯架构智能算力平台	基于 SRAM 及新型存储器存内计算技术，打造非冯架构计算体系。	北京市
千芯科技	大算力存算一体 AI 芯片	（1）针对大算力的存内逻辑/存内计算创新架构; （2）支持 CUDA 语法。	北京市
恒烁半导体	基于 Nor Flash 技术的存算一体终端推理芯片	通过 Flash 阵列的模拟计算来高度并行化完成矩阵计算	安徽合肥
中科声龙	存算一体高通量算力芯片	基于片内超大规模全相联网络的高通量算力芯片，实现计算核心与数据通路之间的性能平衡。	北京市
新忆科技	新型存储器技术研发	新一代非易失性存储器技术	北京市
亿铸科技	基于 RRAM 的全数字存算一体大算力 AI 芯片	（1）核心 IP 均为自研，软件—架构—芯片—工艺—制造均可实现自主可控和国产化; （2）全数字路线，解决模拟计算精度不高等问题。	江苏苏州

（来源：《存算一体芯片深度产业报告》）

存算一体应用场景

一、人工智能（AI）和大数据计算

存内计算适用于AI的深度神经网络应用和基于AI的大数据技术。通过存算一体技术，可将带 AI 计算的大量乘加计算的权重部分存在存储单元中，从而在读取数据的同时进行数据输入和计算处理，在存储阵列中完成卷积运算等人工智能核心运算。

按照算力大小进行划分，存算一体主要应用场景分为小算力场景和大算力场景，进而可细分为五类场景，其中边缘/端侧小算力场景包括智能可穿戴设备、智能安防、移动终端、增强现实/虚拟现实（AR/VR）等，大算力场景包括自动驾驶等。

（一）智能可穿戴设备

可穿戴设备的特征是总是处于工作、待机或可存储状态，因此对于低功耗需求强烈。存算一体技术能够减少不必要的数据搬运，功耗相较传统的芯片降低 10—20 倍，符合可穿戴设备对低功耗的需求。在极低功耗的基础上，存算一体在人工智能加速上比当前芯片的效率提升几十到几百倍不等。

（二）智能安防

智能安防指基于智能视觉、多维感知、组网协同等技术构建的前端智能体系，属于偏视觉类的垂直场景。存算一体

的高并行计算能力比传统芯片的计算实时性高出很多，满足智能安防需求。

（三）移动终端

移动终端是具备通信功能的微型计算机设备，可支持图像识别等 AI 应用。移动终端受制于电池容量，对芯片的功耗有严格限制。存算一体在视觉信号处理上可以达到端侧产品低功耗要求，可在此类场景下应用。

（四）AR/VR

AR/VR 眼镜中的电池小、散热差，对低功耗有较高的要求。AR/VR 场景中会涉及较多的人工智能交互（如语音识别、手势识别等）。存算一体非常适合嵌入到 AR/VR 芯片中，发挥计算效率和实时性方面的优势，为用户提供更真实流畅的交互体验。

（五）自动驾驶

自动驾驶无需人类操作即能通过雷达、GPS、计算机视觉等技术感测环境和导航，其对芯片散热、实时性及可靠性有高要求。存算一体的低功耗、低延迟的特性能够很好地匹配自动驾驶需求，同时该技术也能在较低的成本下把算力做大。

二、感存算一体

存算一体助力含 AI 存算一体芯片的传感器实现零延时和超低功耗的智能视觉处理能力。集传感、储存和运算为一体的感存算一体架构，在解决冯·诺依曼架构的存储墙瓶颈的同时，与传感结合可以提高整体效率。

三、类脑计算

存算一体是类脑计算的关键技术基石。类脑计算是借鉴生物神经系统信息处理模式和结构的计算理论、体系结构、芯片设计以及应用模型与算法的总称。类脑计算试图借鉴人脑的物理结构和工作特点，让计算机完成特定计算任务，从而高速处理信息，属于大算力高能效领域。存算一体天然是将存储和计算结合在一起的技术，非常适合应用在类脑计算领域。

（来源：《存算一体芯片深度产业报告》）

编辑部地址：南宁市体强路 18 号广西信息中心 1412 号房
联系电话：0771-6113592
电子邮箱：dsjyjs@gxi.gov.cn
网　　址：http：//gxxxzx.gxzf.gov.cn/

扫描二维码获取
更多决策参考信息