存算一体技术产业发展研究

黄璜　张乾

（中国信息通信研究院信息化与工业化融合研究所，北京 100191）

摘要：基于存算一体技术产业发展实际情况，结合人工智能算力快速发展的背景，从基础硬件、计算架构、技术挑战等维度分析存算一体技术发展现状和趋势，研究存算一体产业结构、主要应用、产业发展面临的机遇和挑战，最后根据我国算力技术产业发展实际情况，提出存算一体发展策略。

关键词：内存计算；存算一体；非易失性存储器件；人工智能

中图分类号：TN60；TP333　　**文献标志码**：A

引用格式：黄璜，张乾．存算一体技术产业发展研究[J]．信息通信技术与政策，2023，49（6）：30-39.

DOI：10.12267/j.issn.2096-5931.2023.06.005

0　引言

随着人工智能技术产业的演进和向云端、边缘侧的深入，多种依托人工智能算力的新应用、新业态不断涌现。其中，以 ChatGPT 等大模型训练推理为代表的一系列高算力人工智能应用掀起了算力竞赛浪潮，使得突破经典冯·诺依曼架构，探索新算力再次成为计算技术突破的重大议题。存算一体技术具备高能效比、可快速进行矩阵运算等特点，是实现人工智能算力提升的重要候选架构。笔者重点对存算一体技术的产生背景、发展历程、核心技术发展态势、产业和应用发展态势等方面进行分析和研究，以期为我国存算一体技术产业发展提出建设性意见。

1　存算一体技术背景及发展历程

1.1　存算一体技术背景

1.1.1　“冯·诺依曼瓶颈”问题

在冯·诺依曼架构中，数据从存储单元外的存储器获取，处理完毕后再写回存储器，计算核心与存储器之间有限的总带宽直接限制了交换数据的速度，计算核心处理速度和访问存储器速度的差异进一步减缓处理速度，即“冯·诺依曼瓶颈”[1-2]。

一方面，处理器和存储器二者的需求、工艺不同，性能差距也就越来越大。存储器数据访问速度远低于中央处理器（Central Processing Unit，CPU）的数据处理速度，即“存储墙”问题。另一方面，数据搬运的能耗比浮点计算高 1～2 个数量级[3]。芯片内一级缓存功耗达 25 pJ/bit，动态随机存取内存（Dynamic Random Access Memory，DRAM）访问功耗达 1.3～2.6 nJ/bit[4]，是芯片内缓存功耗的 50～100 倍，进一步增加了数据访问能耗。数据访问和存储已成为算力使用的最大能耗，即“功耗墙”问题。

此外，摩尔定律放缓，工艺尺寸微缩变得越来越困难，甚至趋近极限；传统架构提升使得性能增长速度也在变缓，人们试图寻找一种新的计算范式来取代现有计算范式以跳出冯·诺依曼架构和摩尔定律的围墙，并进行多种路径尝试。

1.1.2　高算力需求的挑战

当前，算力需求快速增长与算力提升放缓形成尖锐矛盾。以人工智能为例，从 1960 年到 2010 年算力需求每两年提升一倍，而从 2012 年 Alexnet 使用图形处理器（Graphics Processing Unit，GPU）进行训练开

始，算力每 3~4 个月提升一倍[5]。谷歌 AlphaGo 在与李世石对弈中仅需要使用 1 920 个 CPU 和 280 个 GPU[6]；而谷歌 GPT-3 开源人工智能模型有 1 746 亿个参数，按照训练 10 天估算，需要 3 000~5 000 块英伟达 A100 GPU；GPT-3.5 训练显卡数量进一步增至 2 万块；预计 GPT-4 训练参数在万亿的数量级[7]，是 GPT-3 的 6 倍以上，运行成本和算力需求将大幅高于 GPT-3.5。

1.2 存算一体技术解决方案

1.2.1 高带宽数据通信

高带宽数据通信主要包括光互联技术和 2.5D/3D 堆叠技术。其中光互联技术具有高带宽、长距离、低损耗、无串扰和电磁兼容等优势，但是光互联器件难以在芯片内布设，且光交换重新连接开销和延迟较大，实用化成本较高，难以大规模应用。

2.5D/3D 堆叠技术通过增大并行带宽或利用串行传输提升存储带宽，简化系统存储控制设计难度，具有高集成度、高带宽、高能效等性能优势。但是目前 2.5D/3D 堆叠技术仅对分立器件或芯片内部进行优化设计，“存”和“算”从本质上依然是分离的，难以弥合“存—算”之间的鸿沟。

1.2.2 缓解访存延迟和功耗的内存计算

为了逾越“存—算”之间的巨大鸿沟，内存计算的概念应运而生。内存计算有两种技术类型，一种是横向扩展（Scale-out），主要是分布式内存计算，典型代表有 Spark 架构，是一种软件的方案；另一种是纵向扩展（Scale-up），又分为两种，一种是近数据端处理（Near Data Processing，NDP），包括近存储计算和近内存计算，另一种是存算一体，依赖经典存储器件或新型的存算器件，如图 1 所示。

分布式内存计算是较早前诞生的基于软件的内存计算方案。2003 年谷歌公司提出的 MapReduce 计算框架，能够处理 TB 级数据量，是一种“分而治之再规约”的计算模型，用多个计算节点来计算。但缺点是在反复迭代计算过程中，数据要落盘，从而影响数据计算速度。2010 年，美国加州大学伯克利分校 AMP 实验室提出的分布式计算框架 Spark，能够充分利用内存高速的数据传输速率，同时某些数据集已经能全部放在内存中进行计算，数据尽量留存在内存中，从而避免落盘，随着内存容量持续增长，Spark 依然活跃在工业界。

近数据端处理又分为两种，一种是近存储计算（In-Storage Computing，ISC），即在非易失存储模块中（固态硬盘等）加入现场可编程逻辑门阵列（Field Programmable Gate Array，FPGA）、ARM 处理器核等计算单元。三星在 2019 年展示产品 Smart SSD（PM1725），集成了数字数据处理器（Numeric Data Processor，NDP），可以通过一些编程模型、库和编译器进行程序编译后在硬盘内计算。近数据端计算的另一种方式是近内存计算（In-Memory Computing，IMC），数据直接在内存中计算后返回，通过将存储层和逻辑层堆叠实现大通道计算，目前业界有三星、英伟达、

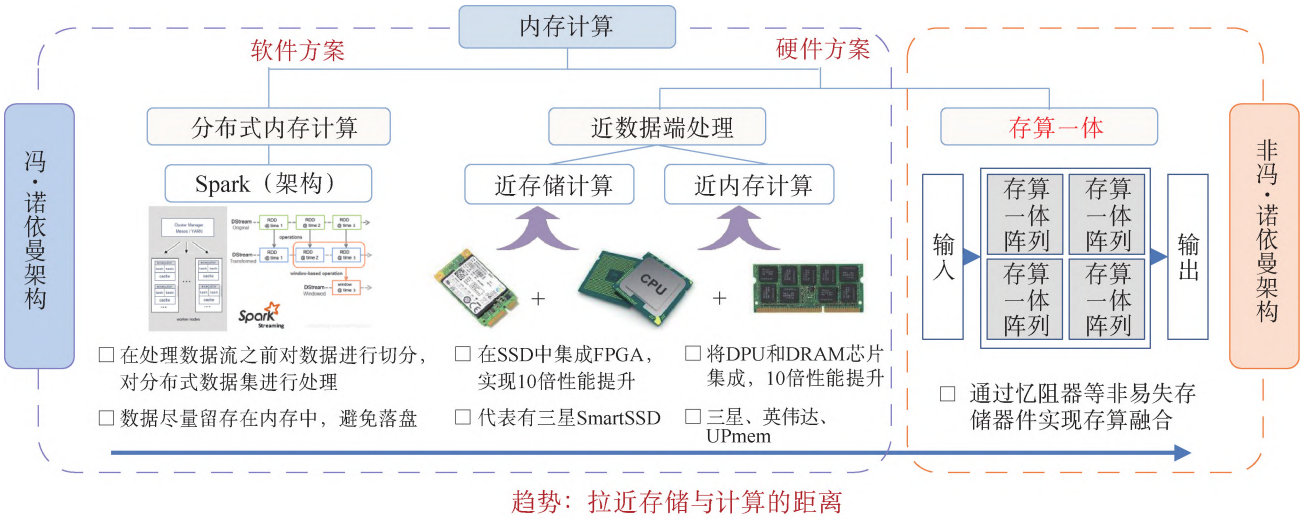


图 1 内存计算体系

UPMem 等企业跟进。

以上基于软件的分布式内存计算和拉近存储与计算距离的近数据端处理，依然保留了经典冯·诺依曼架构的数据处理特点，而基于器件层面实现的存算一体是真正打破了存算分离架构壁垒的非冯·诺依曼架构。一方面，存算一体将计算和访存融合，在存储单元内实现计算，从体系结构上消除了访存操作，从而避免了访存延迟和访存功耗，解决了“冯·诺依曼瓶颈”。另一方面，存算一体恰好能满足人工智能算法的访存密集、规则运算、低精度特性。因此，存算一体是解决“存储墙”“功耗墙”问题的有效方案之一。

2 存算一体核心技术发展态势

存算一体技术体系包含基础理论、基础硬件、计算架构、软件算法和应用五部分。其中基础理论包含近存储计算、计算型存储、欧姆定律、基尔霍夫定律等；基础硬件又包含非易失性存储和易失性存储两大类，非易失性存储又包含基于传统浮栅器件/闪存的存算一体和基于新型非易失性存储器件（Non-Volatile Memory，NVM），包括基于相变存储器（Phase-Change Memory，PCM）的存算一体、基于阻变存储器（Resistive Random Access Memory，ReRAM）的存算一体和基于自旋转移矩磁存储器（Spin-Transfer Torque Magnetoresistence Random Access Menory，STT-MRAM，简称“MRAM”）的存算一体；易失性存储计算则主要基于静态随机存取存储器（Static Random-Access Memory，SRAM）和 DRAM 两类器件。计算架构方面包括逻辑计算、模拟计算、搜索计算三大类型；软件算法包括 TensorFlow、卷积神经网络框架（Convolutional Architecture for Fast Feature Embeddin，Caffe）、卷积神经网络（Convolutional Neural Networks，CNN）、深度神经网络（Deep-Learning Neural Network，DNN）、长短期记忆（Long Short-Term Memory，LSTM）等人工智能相关软件和算法；应用主要包括人工智能、智能物联网（Artificial Intelligence & Internet of Things，AIoT）、图计算、感存算一体等（如图 2 所示）。

2.1 存算一体基础硬件

2.1.1 易失性存储器件：运算较快，但难以实现大规模扩展

存算一体器件与一般 MOSFET 器件的区别在于能“存”，“存”又包括易失性存储和非易失性存储，其中易失性存储的 SRAM 和 DRAM 成为人们优先尝试的对象。

SRAM 二值 MAC 运算可以把网络权重存储于 SRAM 单元中，利用外围电路可以快速实现异或非（XNOR）累加运算，且能够实现二值神经网络运算[8]。DRAM 则利用单元之间的电荷共享机制来实现存算一体，实现较快的运算速度，但是计算对数据具有破坏

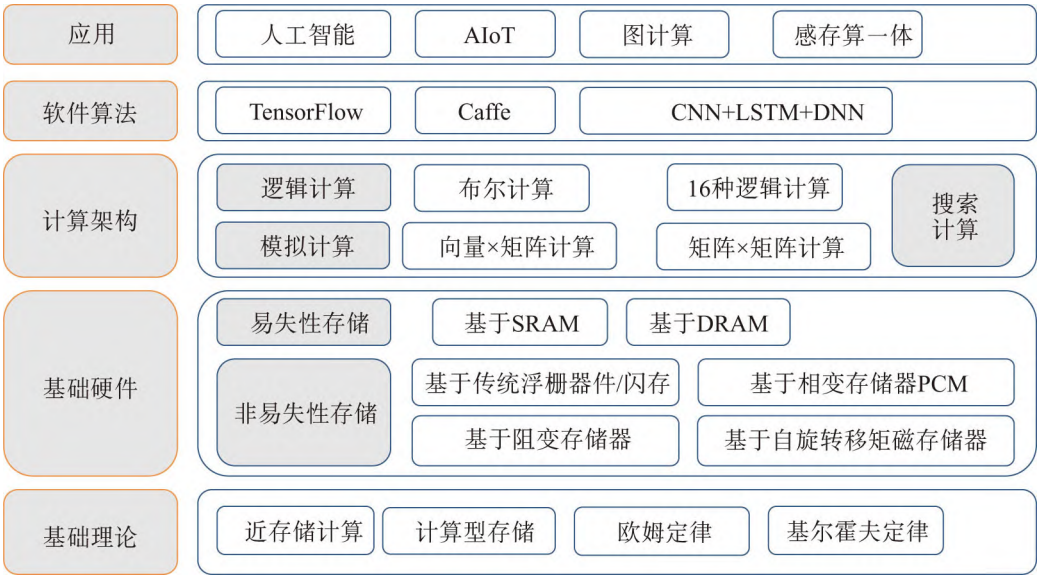


图 2 存算一体技术体系

性，且功耗较大，以上两种存算一体架构均难以在实现大阵列运算的同时保证计算精度。

总的来说，基于易失性存储器件 SRAM 或者 DRAM 存储器的存算一体架构可以实现较快的运算速度，但是难以实现大阵列扩展运算。此外，基于 DRAM 存储器的存算一体架构对数据具有破坏性，并带来显著的功耗问题。

2.1.2 浮栅器件/闪存：工艺成熟，率先应用于存算一体芯片

浮栅器件工艺成熟，编程时间 10～1 000 ns，可编程次数达 10^5次，存储阵列大，可实现量产，运算精度高、密度大、效率高、成本低[9]。NAND Flash 用于存算一体最大的难点是地址和命令只能在 I/O 上传递，不能直接使用，需要十分复杂的技术才能实现模拟计算的功能。因此目前主要使用 Nor Flash 来制造存算一体芯片。

2.1.3 相变存储器：成本及功耗高，已应用于存储级内存中

相变存储器是基于硫属化物玻璃材料，施加合适电流将介质从晶态变为非晶态并再变回晶态，基于材料导电性差异存储数据，如图 3 所示。非晶态相变材料电阻率高、阻值大；多晶态相变材料的电阻率低、阻值小。通过控制脉冲电压幅度产生热量可以实现非晶体和多晶态间转换，从而控制阻值大小，实现存储（阻值态）和计算。优点是高读写速度、寿命长、工艺简单、可以进行多态存储和多层存储；缺点主要是单 bit 成本高、发热量大功耗高、电路设计不完善[10]。

2.1.4 阻变存储器：契合存算一体对器件的需求

ReRAM 是“三明治”结构，包含了上下金属电极和中间的阻变绝缘体层，初始状态为高阻态，需要在两端施加大的电压脉冲“激活”，通过正向/反向电压“击穿”金属氧化层形成导电细丝/氧原子复位，完成在低阻态与高阻态间的转换（如图 4 所示）。优点主要包括可高速读写编程、寿命长、具备多位存储能力、与 CMOS 工艺兼容、功耗低、可 3D 集成；缺点主要有丝状电阻扩展难、相邻单元串扰和器件微缩能力难以兼顾。在商业化上，Crossbar、昕原半导体、松下、Adesto、Elpida、东芝、索尼、海力士、富士通等厂商都在开展 ReRAM 的研究和生产。

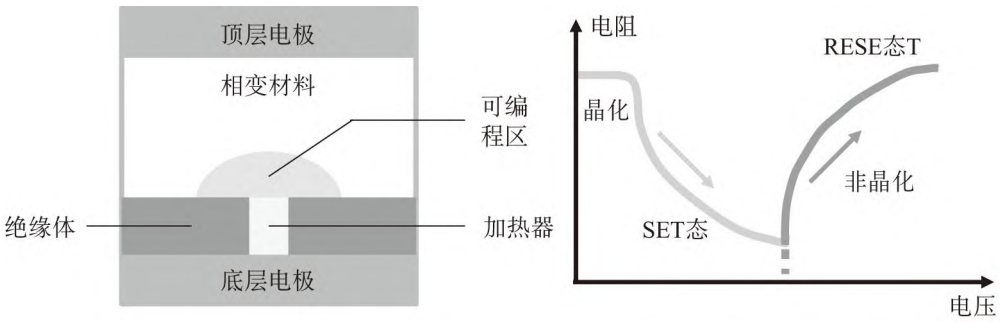


图 3 PCM 器件结构和 R-V 特性

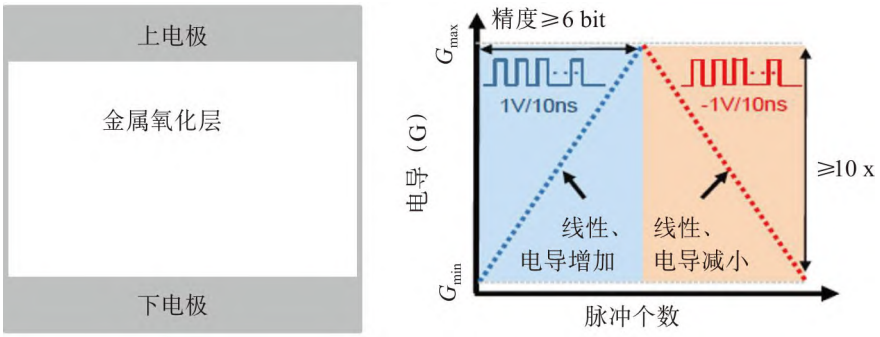


图 4 ReRAM 器件结构和脉冲响应特性

2.1.5　自旋转移矩磁存储器：容量提升有待进一步突破

MRAM 基本结构包含三层，其中底层磁化的方向不变，称为参考层；顶层磁化方向可被编程发生变化，称为自由层；中间层称为隧道层。由于隧道磁阻效应，参考层和自由层的相对磁化方向决定了磁效应忆阻器的阻值大小。参考层和自由层的磁化方向一致时（P 态），磁效应忆阻器的阻值最小；如果磁化方向不一致时（AP 态），磁效应忆阻器的阻值最大（如图 5 所示）。优点主要是读写高速、寿命长，和逻辑芯片整合度高、功耗低；缺点包括临近存储单元之间存在磁场叠加，互相干扰严重。

2.1.6　小结

Nor Flash 工艺成熟，已率先应用于存算一体芯片。SRAM 制作工艺、研发工具都更加成熟稳定，具有耐久性强且操作速度快的特点，可以实时在存算单元中刷新计算数据，具备大算力场景应用潜力。ReRAM 工艺可以与互补金属氧化物半导体（Complementary Metal-Oxide-Semiconductor，CMOS）兼容，具有高速读出、寿命长、功耗低、可 3D 集成等优点，初具产业化潜力，其相关性能如表 1 所示。台积电正开展 MRAM 攻关，未来有望实现突破。但是新型非易失存储器在存算一体技术的应用还存在诸多问题，从实验室到产业化还有一定差距。

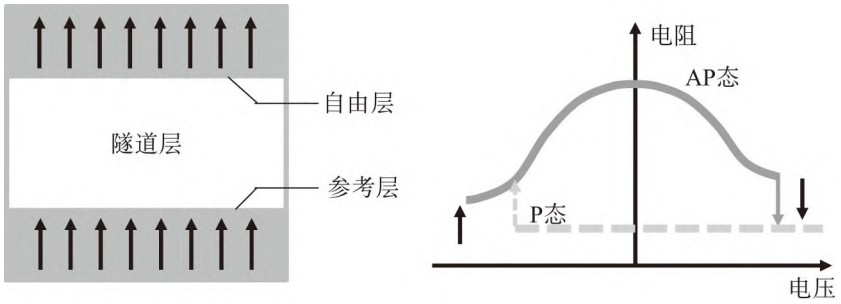


图 5　MRAM 器件结构和 R-V 特性

表 1　存储器件相关性能总结

分类	传统存储器件				新型存储器件		
名称	SRAM	DRAM	NAND Flash	Nor Flash	MRAM	PCM	ReRAM
非易失	否	否	是	是	是	是	是
读取时间（ns）	1～100	30	50	20～50	2～20	20～50	10～50
写入时间（ns）	1～100	15	1 ms	5 s	2～10	50/120	$1\sim10^4$
寿命（重复擦写次数）	10^{16}	10^{16}	10^5	10^5	10^{15}	10^8	10^8
优点	快速	快速	高速读写	工艺成熟 不需要 I/O	寿命长 功耗低	工艺简单	与 CMOS 工艺兼容、功耗低
缺点	难以扩展	破坏性 难以扩展	需要 I/O 难以模拟计算	寿命较低 写入慢	干扰严重	昂贵 功耗高	串扰严重 难以微缩
已有产品容量	MB 级	GB 级	TB 级	/	GB 级	GB 级	/
代表公司	英特尔	三星、 SK 海力士	三星、 SK 海力士	华邦、旺宏 兆易创新、 知存科技	Everspin	英特尔	Crossbar
主要商业应用	高速缓存	内存	外存	驱动存储	嵌入式	混合固态盘、 持久内存	/

2.2 存算一体技术计算架构

2.2.1 逻辑计算：二值忆阻器可以实现完备的布尔逻辑

基于新型忆阻器的存算一体技术架构可实现完备的布尔逻辑计算。如图 6 所示，在 R-R 逻辑运算中，基于欧姆定律和基尔霍夫电压电流定律，根据输入将两个忆阻器件写到对应高低阻态，分别施加电压，输出结果存在 X_2。在 V-R 逻辑运算中，输入是通过施加在单个忆阻器两端的电压幅值 X_1、X_2来表示，而逻辑输出 Y 则由高低阻态来表示。在 V-V 逻辑运算中，根据欧姆定律，输入和输出通过电压幅值低高来分别表示逻辑 0 和 1，需要额外的比较器设计，构成与、或、非 3 类逻辑[10]。

破坏性是指是否会擦除输入的初态。如表 2 所示，只有 R-R 因为输入输出都是忆阻器的阻值，所以输出后原阻值会被擦除，所以具有破坏性；但是电路简单且易级联。V-R 电路具有非破坏性的优点，但是需要额外比较电路，电路复杂度上升。V-V 电路复杂度最高。综合考虑级联性、电路复杂性、破坏性等特性，目前 R-R 和 V-R 更具实用价值。

表 2 R-R、V-R 和 V-V 三种逻辑运算电路的比较

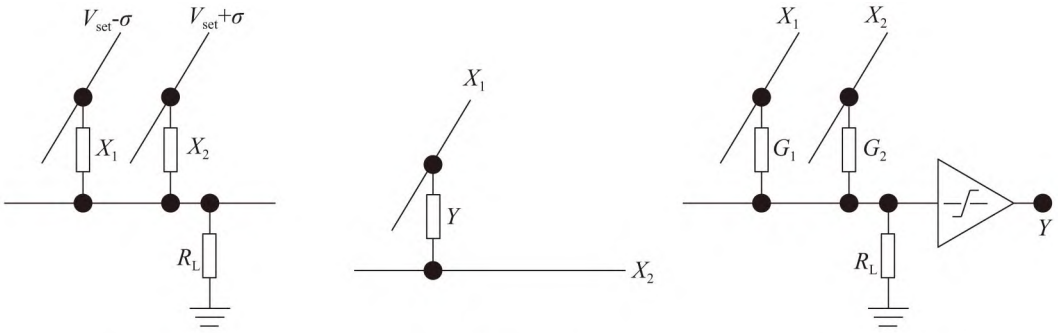
逻辑运算类型	R-R	V-R	V-V
易级联	√	×	√
需要额外比较电路	×	√	√
破坏性	√	×	×
电路复杂度	低	中	高

2.2.2 模拟计算：行列式与矩阵乘运算

基于新型忆阻器的存算一体技术架构，利用欧姆定律和基尔霍夫定律，通过网络阵列可进行矩阵向量乘法运算，如图 7 所示。单个存储单元即可完成 8 bit 乘加法运算（原需 2 500 个晶体管），可并行完成整个矩阵的运算，效率提高 50～100 倍。适用于人工智能训练（超过 90%的运算为矩阵运算）等大数据、低精度、简单乘加运算等场景[1]。

2.2.3 搜索计算：特殊搜索问题具有较高的效能

清华大学的 SQL-PIM 是基于存算一体技术的搜索计算。SQL-PIM 能在不改变结构化存储的前提下支持增、删、改、查操作。针对数据量大的数据库表，SQL-PIM 利用一种特殊的关联分割方法，将大表存



(a) R-R逻辑运算　(b) V-R逻辑运算　(c)V-V逻辑运算

来源：基于新型忆阻器的存内计算[10]

图 6 R-R、V-R 和 V-V 三种逻辑运算电路

$$[x_1\ x_2\ \dots\ x_m]\begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{mn} \end{bmatrix} = [I_1\ I_2\ \dots\ I_n]$$

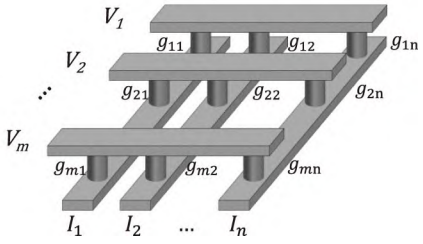


图 7 基于新型忆阻器的向量矩阵乘法

储在多个存内计算阵列中,同时减少每个计算阵列之间的相互通信。与传统的数据库相比,SQL-PIM 能节约 4~6 个数量级的能耗[11]。但是整体而言,存算一体技术应用于搜索运算还停留在实验室阶段,尚未实现产业化或商业化应用。

2.3 存算一体技术挑战

2.3.1 器件特性难以满足全部需求

存算一体技术功能器件纷繁多样,然而目前尚未有一种器件的性能能满足全部应用需求。器件存在均一性差、循环耐久性差、器件状态漂移等问题,目前已有一些优化和解决的方法,但尚未根本解决上述问题。

2.3.2 阵列存在泄露路径、写串扰以及寄生电容电阻问题

存算一体芯片网格阵列面临泄露路径、写串扰以及寄生电容电阻三大问题。在读取器件阻值时,泄露路径的存在引入了并联的电流通路,可能造成错误的读取结果。泄露路径还会带来额外的功耗,并随着阵列规模的扩大而变得更加严重。由于阵列高度并行性带来的写串扰问题会使未被选中器件的阻值受到一定影响。寄生电容、电阻会使电路延迟增加,使远端器件工作异常[12]。

2.3.3 现有集成电路设计与集成技术难以满足需求

控制辅助电路面积和功耗占比太高,外围的器件比存算的部分大很多,外围功耗也会减少存算一体的收益。设计方面,CMOS 走在前沿,与存储存在工艺差距,而统一制程将增加硬件开销,独立制程又将增加系统复杂度。3D 异质集成是可行的路径。

2.3.4 架构设计与开发工具有待标准化

计算的多样性与计算定制性之间存在矛盾。不同计算网络需要定制化的存算一体架构,而全定制又不利于推广。软件和开发工具方面,缺少标准化的异构编程框架;数据映射、数据流配置缺少工具;模拟计算的“模糊/随机性”还需要进行图灵完备性的检验。

3 存算一体技术产业和应用发展态势

3.1 产业发展现状

3.1.1 科研巨头加速布局

IBM 公司重点布局 PCM。2018 年 IBM 公司通过 PCM 实现在数据存储的位置执行计算来加速全连接神经网络训练,该芯片的能效比是传统 GPU 的 280 倍,单位面积算力是传统 GPU 的 100 倍[13]。

三星集团重点布局 DRAM 和 MRAM。2017 年,三星电子存储部门联合加州大学圣巴巴拉分校推出 DRISA 架构,实现了卷积神经网络的计算功能,在提供大规模片上存储的同时也具备较高的计算性能。2022 年初,三星电子在《Nature》上发表了首个基于 MRAM 的存算一体芯片,三星电子采用 28 nm CMOS 工艺重新构建 MRAM 阵列结构,以“电阻总和”(Resistance Sum)的存内计算结构代替了传统的“电流总和”(Current Sum),或电荷共享式的存内计算架构,通过测试分类识别等算法,得到 98%的准确率[14]。

英特尔公司重点布局 SRAM。英特尔公司联合美国密歇根州立大学从 2016 年开始展开基于 SRAM 的计算型存储/存算一体技术研究。2016 年,基于 SRAM 实现了支持逻辑操作的存储器,并在此基础上实现了支持无进位乘法运算的计算型缓存[15]。2018 年英特尔公司发布了面向深度学习算法的神经缓存,可以实现加法、乘法和减法操作[16]。

3.1.2 初创企业涌现,投融资进入活跃期,迎来产业化转折点

存算一体初创公司蓬勃发展,在北美和我国先后涌现多家初创公司。较早成立的初创公司倾向于采用较为成熟的 Nor Flash 器件,知存科技等多家企业在 2021 年实现 Nor Flash 存算一体芯片量产,2021 年成为存算一体产业化元年。近几年,初创企业加快布局 SRAM 领域,但是 ReRAM 等新型非易失存储器件还只在初创企业的蓝图中,尚未实现流片量产。

存算一体技术近年来受到资本市场高度关注,在中美两国涌现的初创企业均获得投融资机会。从 2021 年开始,在我国半导体产业政策和基金双重助力下,存算一体领域投融资尤为活跃,多家初创企业获得上亿元融资。

3.1.3 存算一体技术与类脑计算具有深度关联

存算一体技术是大脑最主要的特征之一,也是实现高算力、高能效计算的一项关键技术。以清华大学为代表的涉及忆阻器领域的科研院校同时进行存算一体技术和类脑计算研究,在材料、器件研发、芯片设计、性能测试等方面深度关联。

存算一体技术和类脑计算具有相同点和不同点。相同点是器件方面均采用忆阻器作为核心器件;应用

都主要面向人工智能。不同点是类脑计算的神经形态器件更复杂，而存算一体器件较为基础；类脑芯片主要采用脉冲神经网络的架构，具有专用性，存算一体技术主要是矩阵结构，具有通用性。

3.2 存算一体技术应用

3.2.1 AI 训练和推理：图像识别、大模型训练推理

2017 年，清华大学团队制备了 128×8 的多值忆阻器阵列，对包含 320(20×16)个像素点的人脸图像进行训练和识别。单幅图像识别耗能可低达 61.16 nJ，识别速度可高达 34.8 ms，识别率超过 85%[17]。

2023 年 3 月，南京大学王欣然教授团队与清华大学吴华强教授团队合作，提出基于二维半导体铁电晶体管的新型存内计算器件架构，通过调节铁电势阱，实现了同时满足 AI 训练和推理需求的底层器件，并展现了高达 103 TOPS/W 级别的能效潜力。该成果突破了边缘端人工智能硬件的关键瓶颈之一[18]。

由于 GPT 等大模型训练中占比 80%~85%的线性计算（Linear）、前馈计算（Feed Forward）、归一化（Layer Norm）以及参数变量乘积等计算流程在进行分解后都可以通过存算一体技术完成，因此存算一体技术在大模型训练方面有望取得应用突破。

与此同时，存算一体计算精度会受到模拟计算低信噪比的影响，通常精度上限在 8 bit 左右，难以实现精准的浮点数计算。现阶段 GPT 大模型训练也主要依赖 H100/A100 等英伟达 GPU 的绝对算力，短期内对能效比等因素不敏感。产业界目前使用的 Nor Flash、SRAM 为主导的存算一体芯片仅在能效比方面拥有优势，在绝对算力方面难以满足智能计算算力需求，难以应用于智能计算中心。

3.2.2 AIoT：终端应用、无人驾驶

随着 AIoT 的快速发展，针对时延、带宽、功耗、隐私/安全性等特殊应用需求，驱动边缘侧和端侧智能应用场景爆发。借助边缘端/终端有限的处理能力，可以过滤掉大部分无用数据，从而大幅度提高用户体验。存算一体技术具有低功耗和适用于低精度 AI 的特性，能够作为协处理器应用于智能终端等 AIoT 场景。

AIoT 是存算一体技术目前布局的重点领域。知存科技重点布局语言唤醒语音活动检测（Voice Activity Detection，VAD）、语音识别、通话降噪、声纹识别等，可以应用在很多嵌入式领域中，包括健康监测以及较低功耗（毫安级）的视觉识别；九天睿芯产品主要用于语音唤醒，或者时间序列传感器信号计算处理；定位推广可穿戴及超低功耗 IoT 设备；后摩智能相关芯片应用于无人车边缘端以及云端推理和培训等场景，2022 年 5 月，后摩智能自主研发的存算一体技术大算力 AI 芯片跑通智能驾驶算法模型。

存算一体技术在向边缘侧延伸过程中面临专用集成电路（Application Specific Integrated Circuit，ASIC）、微控制单元（Microcontroller Unit，MCU）以及边缘计算中心的竞争压力，尚未成为低功耗场景的唯一方案。在语音唤醒等场景中，MCU 足以满足低功耗需求，存算一体芯片不具备优势。随着 5G 等技术的发展，数据处理不再拘泥于本地，边缘计算中心成为端侧智能计算的新路径，存算一体技术面临新的竞争。随着无线充电等新技术的崛起，依赖极低功耗的高续航已经不再是刚需，存算一体芯片低功耗优势场景面临进一步压缩。

3.2.3 感存算一体：在科研领域已取得诸多进展

感存算一体包括触觉/压力感存算一体、视觉/光学感存算一体和嗅觉/气体感存算一体三大类。触觉/压力感存算一体方面，2016 年，新加坡南洋理工大学将阻变压力传感器和阻变存储器串联起来形成触觉记忆单元。视觉/光学感存算一体方面，2019 年，中国香港理工大学提出的 Pd/MoOx/ITO 双端光电阻存储器件（ORRAM），不仅可以进行图像感知和记忆，而且实现了降低图像背景噪声等图像预处理功能。嗅觉/气体感存算一体方面，2017 年，美国斯坦福大学团队将 100 多万个忆阻器与 200 多万个碳纳米管晶体管（Carbon-Nanotube Field-Effect Transistors，CNTFET）集成感知周围气体，并转化为电信号存储在 ReRAM 中。与之前训练学习的气体数据进行对比，从而识别出所检测的气体种类[19]。

3.2.4 矩阵与搜索：图计算和基因工程

图计算中大量操作都可以转换成矩阵乘的形式，因此可以用存算一体技术来处理，在预处理、稀疏矩阵的分隔和映射、硬件控制和数据流设计等环节能够实现超过传统计算的能效比。生物数据的暴增给诸如基因序列查找/匹配的应用带来了很大的挑战。基于存算一体技术的搜索计算能级能够提供高硬件并行度，适用于大规模生物数据处理[11]。

4 结束语

存算一体技术应作为我国先进计算产业发展的重点之一，需保持长期关注，要做好中长期路线制定，在支持现有 Nor Flash 的基础上加强对 ReRAM 等新型非易失存储的研究，并对存算一体相关基础材料、设计工具等加强研发。但是也要明确其短期内难以为我国基础算力技术产业发展发挥巨大作用，因此要向领先国家和企业吸取相关经验教训，避免超前投入。此外，要加快推进存算一体应用融合，在未来 3~5 年内通过自主创新开发专门的存算一体芯片设计工具等基础性产品，提升综合性能，加强“器件—芯片—算法—应用”跨层协同，构建存算一体芯片的产业化应用与生态。

参考文献

[1] WULF W, MCKEE E S. Hitting the memory wall: implications of the obvious [J]. ACM Computer Architechture News, 1994, 23(1): 20-24.

[2] ZIDAN M A, STRACHAN J P, LU W D. The future of electronics based on memristive systems [J]. Nature Electronics, 2018, 1(1): 22-29.

[3] BILL D. The Path to Exascale Computing[R], 2015.

[4] HOROWITZ M. Computing's energy problem (and what we can do about it)[C]// IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014:10-14.

[5] MEHONIC A, KENYON A J. Brain-inspired computing needs a master plan[J]. Nature, 2022, 604: 255-260.

[6] WIKIPEDIA. CHARLES M, NEO J, et al. AlphaGo [EB/OL]. (2023-03-29)[2023-04-12]. https://en.m.wikipedia.org/wiki/AlphaGo#.

[7] MATTHIAS B. GPT-4 has a trillion parameters[EB/OL]. (2023-03-25)[2023-04-12]. https://the-decoder.com/gpt-4-has-a-trillion-parameters/.

[8] 郭昕婕，王绍迪. 端侧智能存算一体芯片概述[J].微纳电子与智能制造，2019，1(2):72-82.

[9] 蒋明峰. 用于神经网络的浮栅单元向量-矩阵乘法器的研究与设计[D].合肥：中国科学技术大学，2020.

[10] 林钰登，高滨，王小虎，等. 基于新型忆阻器的存内计算[J].微纳电子与智能制造，2019，1(2):35-46.

[11] 毛海宇，舒继武，李飞，等. 内存计算研究进展[J].中国科学:信息科学，2021，51(2):173-205.

[12] 徐丽莹，杨玉超，黄如. 基于忆阻器的非易失逻辑研究前沿[J].中国基础科学，2019，21(2):1-11+27+63.

[13] AMBROGIO S, NARAYANAN P, TSAI H, et al. Equivalent-accuracy accelerated neural-network training using analogue memory[J]. Nature, 2018, 558: 60-67.

[14] JUNG S, LEE H, MYUNG S, et al. A crossbar array of magnetoresistive memory devices for inmemory computing[J]. Nature, 2022, 601(7892): 211-216.

[15] YAO P, WU H, GAO B, et al. Face classification using electronic synapses[J]. Nature Communications, 2017, 8(1):15199.

[16] NING H, YU Z, ZHANG Q, et al. An in-memory computing architecture based on a duplex two-dimensional material structure for in situ machine learning[J]. Nature Nanotechnol, 2023(18):493-500.

[17] AGA S, JELOKA S, SUBRAMANIYAN A, et al. Compute caches[C]//IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2017:481-492.

[18] ECKERT C, WANG X, WANG J, et al. Neural cache: bit-serial in-cache acceleration of deep neural networks: 10.1109/ISCA.2018.00040[P], 2018.

[19] 李锟，曹荣荣，孙毅，等. 基于忆阻器的感存算一体技术研究进展[J].微纳电子与智能制造，2019，1(4):87-102.

作者简介：

黄璜 中国信息通信研究院信息化与工业化融合研究所工程师，博士，主要从事先进计算、集成电路、算力基础设施等方面的研究工作

张乾 中国信息通信研究院信息化与工业化融合研究所工程师，博士，主要从事先进计算、人工智能等方面的研究工作

Research on development of compute-in-memory technology and industry

HUANG Huang, ZHANG Qian

(Informatization and Industrialization Integration Research Institute, China Academy of Information and Communications Technology, Beijing 100191, China)

Abstract: Based on the actual situation of the development of compute-in-memory technology and industry, against the backdrop of rapid development of artificial intelligence computing power, this paper first analyzes the current status and trends of compute-in-memory technology development in the dimensions of basic hardware, computing architecture, and technological challenges. Then, it studies the industrial structure, main applications, and opportunities and competitive challenges faced by the compute-in-memory industry. Finally, based on the actual development of China's computing technology industry, a strategy for compute-in-memory development is proposed.

Keywords: processing-in-memory; compute-in-memory; non-volatile memory; artificial intelligence

(收稿日期:2023-05-11)