存内计算芯片研究进展及应用

郭昕婕[①] 王光耀[②] 王绍迪*[①]

[①](北京知存科技有限公司 北京 100191)

[②](北京航空航天大学集成电路科学与工程学院 北京 100191)

摘 要：随着数据快速增长，冯诺依曼架构内存墙成为计算性能进一步提升的关键瓶颈。新型存算一体架构(包括存内计算(IMC)架构与近存计算(NMC)架构)，有望打破冯诺依曼架构瓶颈，大幅提高算力和能效。该文介绍了存算一体芯片的发展历程、研究现状以及基于各类存储器介质(如传统存储器DRAM, SRAM和Flash和新型非易失性存储器ReRAM, PCM, MRAM, FeFET等)的存内计算基本原理、优势与面临的问题。然后，以知存科技WTM2101量产芯片为例，重点介绍了存算一体芯片的电路结构与应用现状。最后，分析了存算一体芯片未来的发展前景与面临的挑战。

关键词：存算一体；存储墙；功耗墙；存内计算；近存计算；冯诺依曼架构瓶颈

中图分类号：TN4 **文献标识码**：A **文章编号**：1009-5896(2023)05-1888-11

DOI: 10.11999/JEIT220420

Technology Developments and Applications of In-memory Computing Processors

GUO Xinjie[①] WANG Guangyao[②] WANG Shaodi[①]

[①](*Beijiing Zhicun* (*Witmem*) *Technology Corporation Limited*, *Beijing* 100191, *China*)

[②](*School of Integrated Circuit Science and Engineering*, *Beihang University*, *Beijing* 100191, *China*)

Abstract: Memory wall has become one of the key challenges in Von Neumann architecture, memory-centric computing architectures, such as In-Memory Computing (IMC) and Near-Memory Computing (NMC) are expected to break the Von-Neumann bottleneck, improving computing performance and energy efficiency. The progress of memory-centric computing technology, as well as the principles, advantages and problems based on a variety of memory media, such as traditional memories (e.g., DRAM, SRAM and Flash) and emerging non-volatile memories (e.g., ReRAM, PCM, MRAM and FeFET) are introduced in this paper. Then, the circuit structure and main applications with IMC chips are highlighted, taking Witmem's product WTM2101 as an example. Finally, the future development prospects and challenges of the all-in-one chip are also analysed.

Key words: Memory-centric computing; Memory wall; Power wall; In-Memory Computing (IMC); Near-Memory Computing (NMC); Von-Neumann bottleneck

1 引言

随着人工智能、物联网、大数据等应用的兴起，数据以爆发式的速度进行增长。自2012年起，全世界每天产生的数据量约为2.5×10^{18} Byte，且该体量仍然以每40个月翻倍的速度在持续增长[1]。海量数据的高效存储、迁移与处理成为当前信息领域的重大挑战。然而，由于冯诺依曼架构的局限性[2]，数据的高效处理遇到了存储墙和功耗墙两大问题。在冯诺依曼架构中，数据存储与处理是分离的，存储器与处理器之间通过数据总线进行数据传输，如图1(a)所示。一方面，存储器的访问速度远远小于处理器的运算速度，系统整体会受到传输带宽的限制，导致处理器的实际算力远低于理论算力，难以满足大数据应用的快、准响应需求，称为存储墙问题。通过增加数据总线带宽或者时钟频率可以在一定程度上提高处理器性能，但必将带来更大的功耗与硬件成本开销，且其扩展性也严重受限。另一方面，冯诺依曼架构的存储与计算分离，数据在存储器与处理器之间的频繁迁移带来巨大的传输功耗，

收稿日期：2022-04-08；改回日期：2022-10-09；网络出版：2022-10-20

*通信作者： 王绍迪 shaodi.wang@witintech.com

基金项目：科技部“科技助力经济2020”重点专项项目(SQ2020YFF0404823)

Foundation Item: The Ministry of Science and Technology's Key Special Project (SQ2020YFF0404823)

称为功耗墙瓶颈。例如，英伟达的研究报告指出，在22 nm工艺节点下，浮点运算所需的数据传输功耗是数据处理功耗的约200倍[3,4]。上述存储墙与功耗墙问题并称为冯诺依曼架构瓶颈。

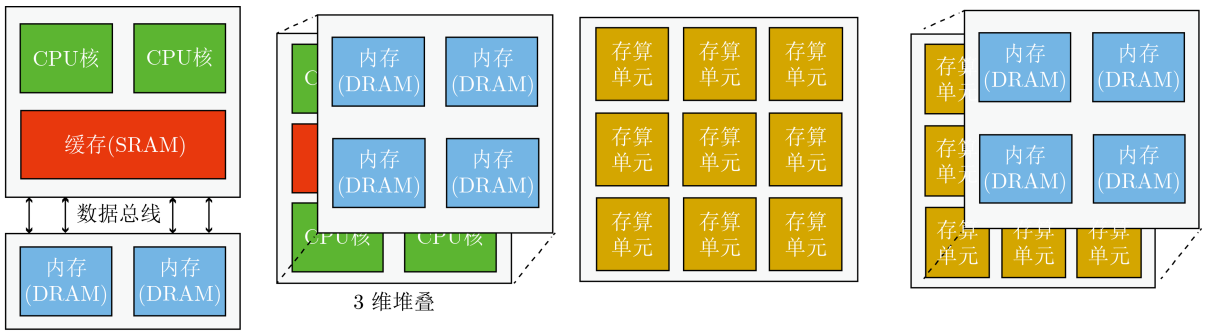
为了缓解冯诺依曼架构瓶颈，目前产业界采用的主流方案是通过高速接口、光互联、3维堆叠、增加片上缓存等方法来提高数据带宽，并把存储器和处理器之间的数据传输距离缩短，以减小功耗。其中，产业界应用较多的是3维堆叠技术与增加片上缓存等方法。以3维堆叠技术[5,6]为例，其基本思想就是把更多的存储器通过在垂直方向上的堆叠，提高数据带宽，并缩短两者之间的距离，这在本质上称为近存计算架构[7,8]，如图1(b)所示。但是，3维堆叠技术并没有改变冯诺依曼架构，只能在一定程度上缓解，但并不能从根本上解决冯诺依曼架构瓶颈。存内计算，作为一种新型计算架构，直接利用存储器本身进行数据处理，从根本上消除数据搬运，实现存储与计算融合一体化，有望突破冯诺依曼架构存储墙与功耗墙瓶颈，成为后摩尔时代集成电路领域的重点研究方向之一，如图1(c)所示。近年来，基于先进2.5D/3D封装技术，结合近存计算和存内计算的新型架构得到业界的重点关注，有望从存储与计算两方面进一步优化性能，如图1(d)所示。

2　存算一体技术发展历程

存算一体包括近存计算与存内计算，其概念最早在1969年被提出[9,10]，后续各国学者在电路、算法、计算架构、操作系统、系统应用等层面开展了一系列相关研究。例如，1997年，文献[11]展示了一种智能内存(Intelligent RAM)方案，其将处理器和DRAM集成在单颗芯片上，算力可达到当时最先进的Cray向量处理器(Cray T-90)的5倍。1999年，文献[12]提出了一种嵌入计算功能的灵活内存(FlexRAM)方案，仿真结果表明该芯片架构可使计算性能提升25～40倍。但是，早期由于缺少大数据处理的应用需求，加之芯片的制造成本昂贵、设计复杂，存算一体技术多年来仅停留在研究阶段。

2015年以来，由于摩尔定律的逐渐失效与冯诺依曼架构的局限性越来越明显，加之大数据应用的驱动，工艺水平的不断提高，存算一体技术重新受到关注，并成为研究热潮。例如，在2017年微处理器顶级年会(Micro2017)上，众多高校和企业都推出了他们的存算一体芯片或系统原型[13–15]，包括苏黎世联邦理工学院、加利福尼亚大学圣巴巴拉分校、英伟达、英特尔、微软、三星等。2019年，文献[16]提出的SRAM存算一体芯片可实现二值权重的神经网络卷积计算。2020年，文献[17]展示了一款ReRAM存算一体芯片，在降低计算延迟的同时大幅提升能效。2021年，文献[18]提出三值DRAM存算一体架构实现神经网络运算加速。2022年，文献[19]提出了多芯粒的存算一体集成芯片。文献[20–24]基于SRAM/ReRAM发表了一系列存算一体器件、芯片与系统相关的研究成果。迄今，基于SRAM, DRAM, Flash, ReRAM, PCM, FeFET, MRAM等各类存储介质，涌现出了一系列相关研究工作[25–38]，存算一体芯片研究百花齐放，如图2所示。特别地，2021-2022年，被誉为芯片领域奥林匹克的顶级国际会议ISSCC收录了存算一体相关论文20余篇，研究单位包括三星、台积电、麻省理工学院、普林斯顿大学、清华大学、北京大学、复旦大学、中国科学院大学等国际顶尖高校和企业。

虽然基于各类存储介质的存算一体芯片研究百花齐放，但是各自在大规模产业化之前都仍然面临一些问题和挑战。更具体地，SRAM工艺成熟，且微缩性好；但是属于易失性存储器(掉电数据丢失)，且单元面积较大，成本较高，难以通过较低成本实现大规模、大算力存内计算芯片。DRAM工艺成熟，且单元面积较小；但同属易失性存储



(a) 冯诺依曼计算架构　(b)基于 3 维堆叠技术的近存计算架构　(c) 存内计算架构　(d) 近存计算+存内计算架构

图 1 计算架构的演变示意图

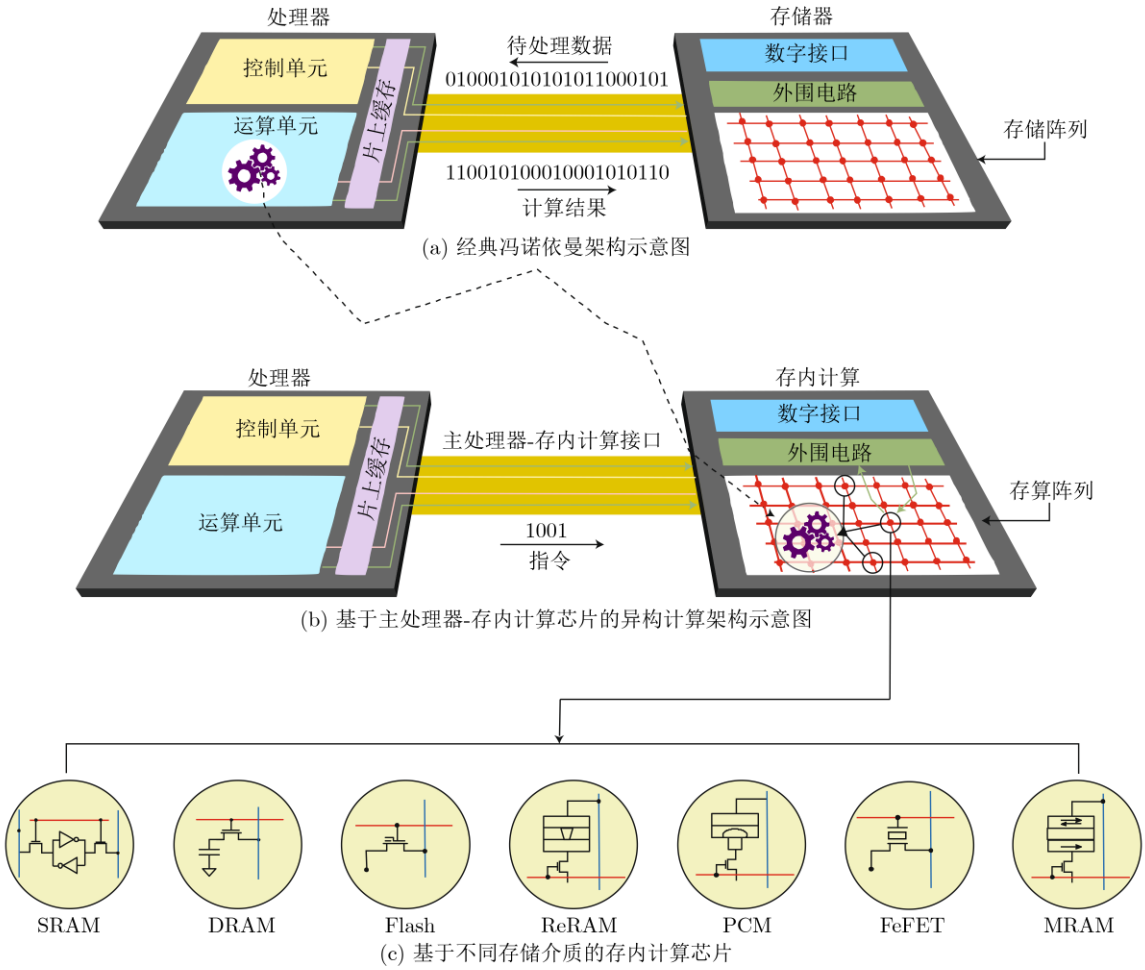


图 2 基于不同存储介质的计算架构演变图[39]

器，需定期刷新，且存在漏电问题，难以实现高精度存内计算芯片，近年来被广泛应用于近存计算。ReRAM属于非易失性存储器，且能够实现大规模交叉点阵列，是未来实现存内计算芯片的潜力介质之一；但是目前的工艺尚不成熟，存储单元的多比特精度较低(低于8 bit)，且一致性/鲁棒性较差。PCM属于非易失性存储器，且能够实现大规模交叉点阵列；但是功耗较大，速度较慢，耐久性较差。FeFET可实现非易失性存储，且能实现交叉点阵列；但是目前的工艺也尚不成熟。MRAM是非易失性存储器，具有高耐久性、高速度、低功耗等优点，工艺相对较成熟，扩展性较好，但是器件的阻值(约几千欧姆)与高低阻值比率(约250%)相对较小，在实现多比特存内计算芯片方面具有一定挑战。Flash是非易失性存储器，掉电数据不丢失，且工艺成熟，成本低，已实现量产芯片(如Mythic的M1076，知存科技的WTM2101)，但在微缩性方面存在一定挑战；幸运的是，随着2.5D/3D先进封装技术的快速发展，可以实现与先进逻辑工艺的兼容集成。综上，基于不同存储器介质的存算一体芯片之间的性能比较如表1所示。

存算一体技术在产业界的进展同样十分迅速，国内外多家企业在积极研发，例如我国台湾的台积电，韩国三星、日本东芝、美国Mythic，国内的知存科技等。但是当前最接近产业化的主要是台积电、Mythic和知存科技。从2019年至今，台积电得益于其强大的工艺能力，已基于SRAM与ReRAM发表了一系列存算一体芯片研究成果[40,41]，具备量产代工能力。Mythic已于2021年推出基于NOR Flash的存内计算量产芯片M1076，可支持80 MB神经网络权重，单个芯片算力达到25 TOPS，主要面向边缘侧智能场景。知存科技于2021年推出基于NOR Flash的存内计算SoC芯片WTM2101，其算力比市场同类芯片高出两个数量级，功耗低于1 mW，主要面向端侧低功耗、低成本应用场景。

3 存内计算芯片研究现状

由于计算范式和存储介质的不同，存内计算芯

片可以有不同的分类方法。根据计算范式的不同，主要分为模拟式和数字式两种。模拟式存内计算是指存储单元内部或阵列周边的信号以模拟信号的方式进行操作，数字式存内计算是指在实际运算过程中，存储单元内部或阵列周边的信号以数字信号的方式进行操作。其中，诸多的研究工作同时包含了模拟和数字两种运算方式。同时，根据存储介质的不同，存内计算芯片可分为基于传统存储器和基于新型非易失性存储器两种。传统存储器包括SRAM, DRAM和Flash等；新型非易失性存储器包括ReRAM, PCM, FeFET, MRAM等。其中，距离产业化较近的是基于NOR Flash和基于SRAM的存内计算芯片。

3.1 SRAM存内计算

基于SRAM的存内计算芯片以典型的6T(6-Transistor)基本单元为基础，如图3(a)所示。由于SRAM是二值存储器，二值乘累加运算等效于同或累加运算，可以用于二值神经网络运算，其核心思想是网络权重存储于SRAM单元中，激励信号从字线给入，最终利用外围电路实现同或累加运算，结果通过计数器或模拟电流/电压输出。如果要实现多比特精度运算，通常需要多个单元进行拼接，这不可避免地会带来面积开销。对6T基本单元的一个简单修改是将字线进行拆分，如图3(b)所示。此外，为了解决读写干扰问题，可以采用8T基本单元，但明显增加了布局面积，如图3(c)所示。基于SRAM的存内计算技术由于其工艺成熟度与良好的微缩性，受到业界的高度关注，近几年的ISSCC会议上连续报道了多篇相关论文。例如2021年，存内计算共有两个分论坛，共收录8篇论文，其中5篇是SRAM存内计算芯片。在2022年的ISSCC中，北京大学提出了一种基于动态逻辑且无模数转换器的SRAM存内计算芯片[42]。SRAM存内计算技术的主要应用难点是在保证运算精度的前提下，实现高算力和小面积。

3.2 DRAM存内计算

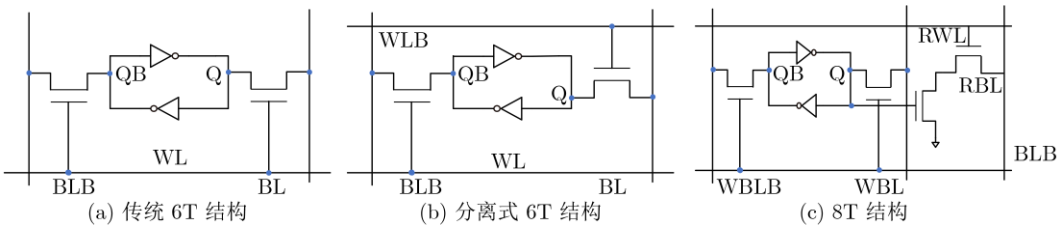
基于DRAM的存内计算芯片层次结构可分为阵列、子阵列和单元，一组阵列由若干子阵列和用于读写操作的相关外围电路组成，而子阵列则包含若干行1T1C(1-Transistor-1-Capacitor)单元、感知放大器和本地解码器。其基本原理是利用DRAM单元之间的电荷共享机制[13,43]。如图4所示为一种典型实现方案[43]，当多行单元同时被选通时，不同单元之间因为存储数据的不同会产生电荷交换共享，从而实现逻辑运算。DRAM存内计算方案的主要难点有二：一是其本身为易失性存储器，计算操作会破坏数据，需要每次运算后进行刷新，带来功耗问题；二是实现大阵列运算时难以保证运算精度。

3.3 ReRAM/PCM存内计算

ReRAM/PCM存内计算的基本原理是利用存储单元的模拟多比特特性，通过基于电流/电压的欧姆定律与基尔霍夫定律进行矩阵乘加运算，主要有1T1R (1-transistor-1-resistance)结构和交叉阵列结构两种实现方案，如图5(a)和图5(b)所示。ReRAM能够实现大规模交叉点阵列，使其成为学术界的热点研究方向。自2008年ReRAM首次实验发现以来，基于ReRAM的存内计算研究就层出不穷。尤

表 1　基于不同存储介质的存内计算芯片性能比较

标准	SRAM	DRAM	Flash	ReRAM	PCM	FeFET	MRAM
非易失性	否	否	是	是	是	是	是
多比特存储能力	否	否	是	是	是	是	否
面积效率	低	一般	高	高	高	高	高
功耗效率	低	低	高	高	高	高	高
工艺微缩性	好	好	较差	好	较好	好	好
成本	高	较高	低	低	较低	低	低
技术成熟度	测试芯片	测试芯片	量产产品	测试芯片	测试芯片	器件	测试芯片



(a) 传统 6T 结构　(b) 分离式 6T 结构　(c) 8T 结构

图 3 基于SRAM的存内计算单元结构

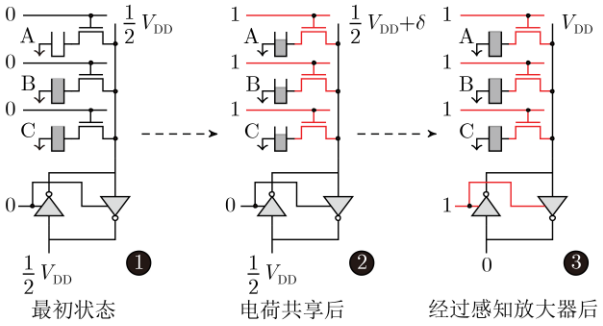


图 4 基于DRAM的存内计算基本原理[43]

其2020年，清华大学研发出基于多个ReRAM阵列的存内计算系统，该系统在手写数字集上的识别准确率达到96.19%，与软件的识别准确率相当，证明了存内计算架构全硬件实现的可行性，其测试芯片如图5(c)所示[24]。ReRAM存内计算技术未来具有非常大的应用潜力，目前的主要难点在于工艺尚不太成熟，多比特精度实现较困难，一致性/鲁棒性较差。

3.4 MRAM存内计算

MRAM存内计算主要有两种技术方案：(1) 基于读/写操作的数字式存内计算；(2) 基于基尔霍夫电流定律和欧姆定律的模拟式存内计算。早期的MRAM存内计算大多基于数字式方案，如2015年日本东北大学提出基于读操作实现多种布尔逻辑并流片验证，获得了48.3%的能效提升[44]；2019年，北京航空航天大学提出基于单次写操作的数字式MRAM存内计算方案，实现计算结果原位存储的同时降低了延时和功耗[45–47]。基于MRAM的模拟式存内计算的难点在于器件的阻值(约几千欧姆)与高低阻值比率(约250%)相对较小，难以实现多比特精度。近年来，得益于计算范式、器件、电路的多层次创新突破，MRAM模拟存内计算发展迅速。2021年，美国普林斯顿大学通过电路级优化，流片验证了第一款基于STT-MRAM的模拟存内计算硬核[48]；2022年，韩国三星公司在Nature期刊上发表了基于电阻累加方案的MRAM模拟存内计算芯片原型，并实现了最高405 TOPS/W的能效比[49]，其阵列的布局图、显微图和结构如图6所示。

3.5 NOR Flash存内计算

基于NOR Flash的存内计算技术原理与ReRAM类似，如图7(a)所示。目前，NOR Flash存内计算芯片技术相对较成熟，已于2021年实现量产。美国的Mythic和国内的知存科技都已推出NOR Flash存内计算芯片产品，其中，Mythic推出了M1076芯片(如图7(b)所示)，知存科技推出了WTM2101量产SoC芯片(如图7(c)所示)。

3.6 基于其他介质的存内计算

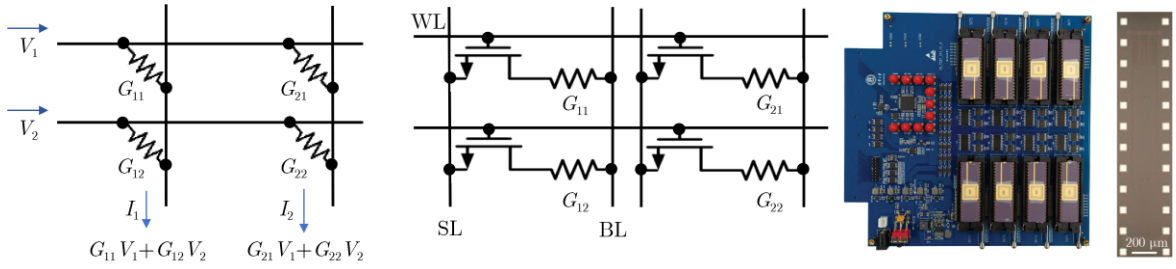
此外，学术界还发表了基于NAND Flash以及新型纳米器件(如FeFET、斯格明子等)的存内计算相关工作，其基本原理与上述方案类似，但是目前仅仅是概念阶段，这里不再详述[50–54]。

4 存内计算芯片应用现状：以WTM2101为例

随着万物互联的不断发展，智能设备主要包括3类：云端、边缘端和终端。云端设备的要求主要是高算力、大吞吐量、高可靠性，当前的存内计算进展还难以满足需求。边缘端设备，如安防、自动驾驶等，对算力、时延、功耗、安全性等具有相对综合的需求；终端设备则主要关注功耗、成本和隐私。目前存内计算芯片应用方面尚处于起步阶段，本节以知存科技推出的量产SoC芯片WTM2101为例，讨论其在边缘端和终端的应用，侧重于语音场景的介绍，同时介绍其核心电路与芯片架构、性能与应用场景。

4.1 核心电路与芯片架构

在NOR Flash存内计算芯片当中，向量-矩阵乘法运算基于电流/电压的跨导与基尔霍夫定律进行物理实现，如图7(a)所示。因此，其核心是设计NOR Flash单元阵列以满足大规模高能效向量-矩阵乘法运算。同时，在核心电路的基础上，根据算法特征设计芯片架构，以充分利用神经网络数据流式的特点来实现芯片的并行化与流水线。在传统



(a) ReRAM存内计算的交叉阵列结构 (b) ReRAM存内计算的 1T1R 阵列结构 (c) ReRAM存内计算的测试芯片[24]

图 5 基于ReRAM的存内计算阵列结构与测试芯片

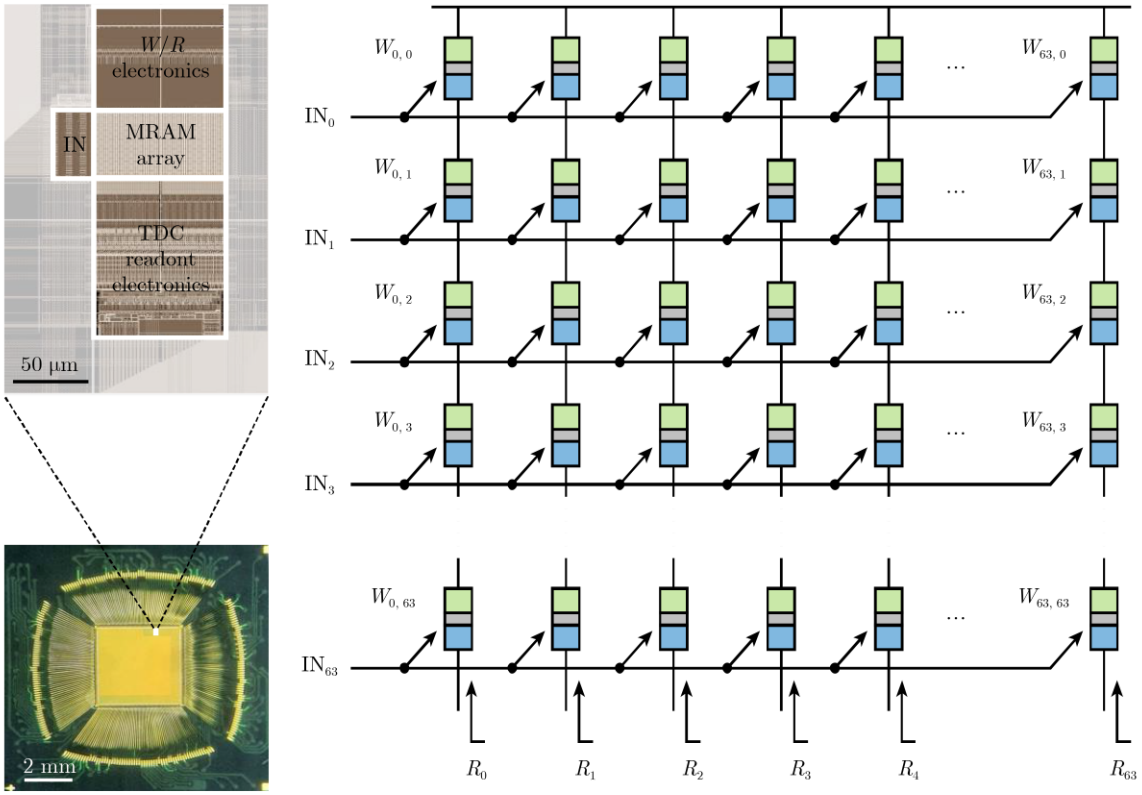
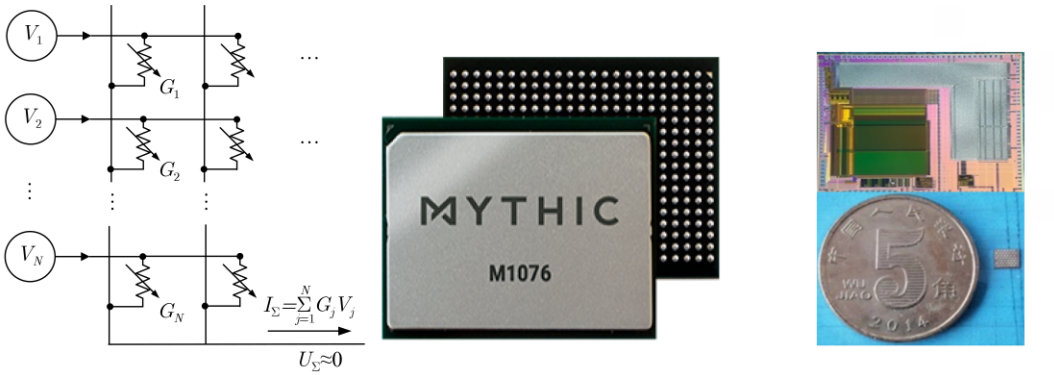


图 6 基于MRAM的存内计算阵列布局图、显微图和结构[49]



(a) NOR Flash存内计算技术原理　(b) Mythic的存内计算产品M1076　(c) 知存科技的存内计算产品WTM2101

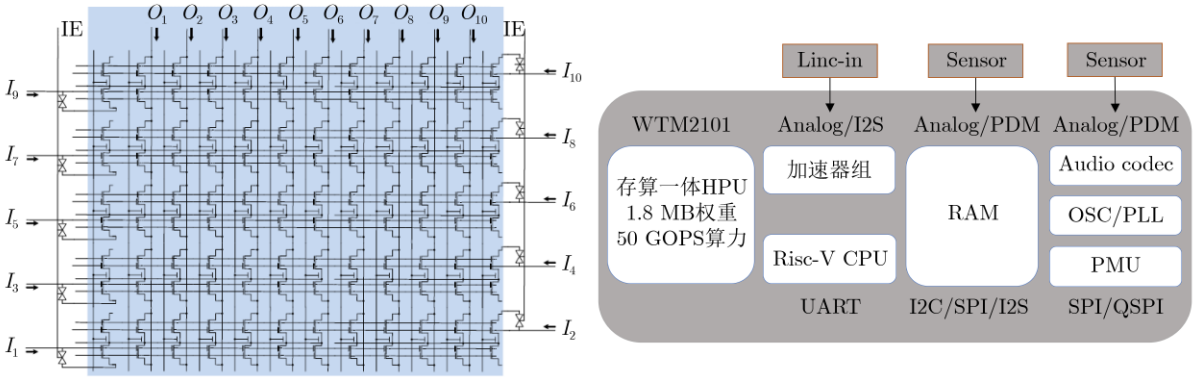
图 7 基于NOR Flash的存内计算技术原理与相关产品

NOR Flash阵列中，对某一个特定器件编程会不可避免地改变同一行上其他器件的状态，称为行干扰。作为存内计算应用，NOR Flash编程需要逐个器件进行单独操作，每个器件存储8 bit(256个量化状态)以上的信息，微小的干扰就将导致状态的变化。因此，需要抗编程干扰阵列结构来消除编程干扰。除此之外，NOR Flash基于浮栅中电子的数量来存储信息，随着时间的增加，电子会泄露，造成阈值电压漂移。作为存储应用的NOR Flash器件通常只保存1～2 bit信息(对应2～4个不同状态)，状态之间的裕量比较大，不用特殊设计即可保存信息10年以上。但在存内计算应用中，NOR Flash器件需要存储8 bit(256个不同状态)以上信息，状态之间的裕量非常小，且通过整个阵列同时工作。因此，阈值电压漂移的影响非常大。WTM2101通过特殊的电路设计抑制阈值电压漂移对计算精度的影响。此外，为了同时实现低功耗计算与低功耗控制，WTM2101结合了RISC-V指令集与NOR Flash存内计算阵列，其阵列结构与芯片架构如图8所示，包括1.8 MB NOR Flash存内计算阵列，一个RISC-V核，一个数字计算加速器组，320 kB RAM以及多种外设接口。

4.2 性能与应用场景

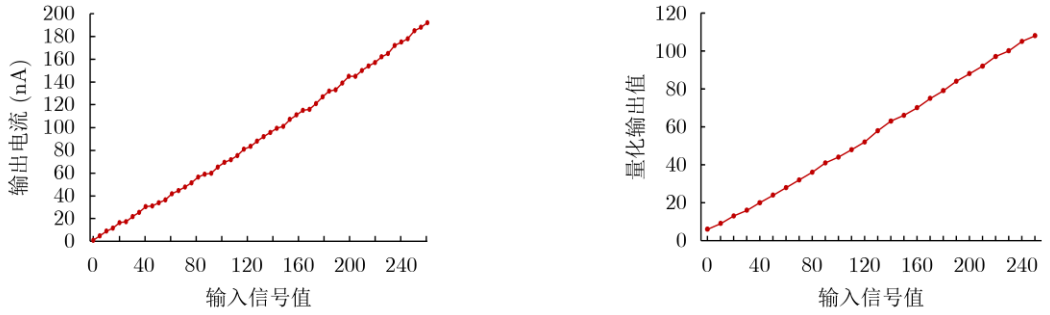
WTM2101基于40 nm工艺进行流片，单个

NOR Flash 器件能够存储8 bit权重，因此可以进行8 bit精度的矩阵乘加运算。如图9所示为输入信号与输出电流之间的关系，单元和芯片均呈现良好的线性关系。WTM2101具有4大优势特点：(1)基于存内计算架构，可高效地实现神经网络语音激活检测和上百条语音命令词识别。(2)以超低功耗实现神经网络环境降噪算法、健康监测与分析算法。(3)典型应用场景下，工作功耗均在微瓦级别。(4)采用极小封装尺寸。基于以上优势特点，WTM2101可应用于智能可穿戴设备、智能家居、安防监控、玩具机器人等；适应多种应用，如语音识别、语音降噪/增强、轻量级视觉识别、健康监测和声纹识别等。如图10所示为搭载WTM2101的耳机产品及其自动化部署流程。如图11所示为基于WTM2101的耳机降噪前后效果的波形和频谱对比。如表2所示为部署在WTM2101的神经网络的各层累计余弦相似度(指存内计算相对于8-bit量化计算的余弦相似度)，可以看到经过8层神经网络计算，余弦相似



(a) WTM2101阵列结构[55]　　(b) WTM2101芯片架构

图 8 WTM2101的阵列结构与芯片架构



(a) 单个单元　　(b) 整个芯片

图 9 WTM2101存内计算芯片8 bit精度运算测试结果

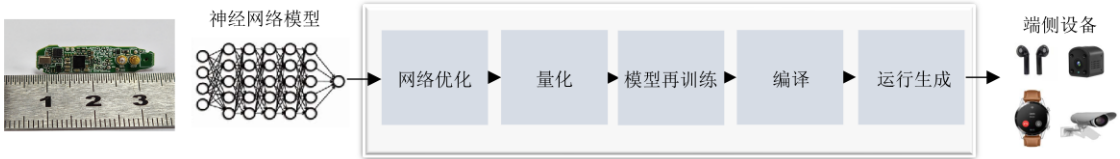
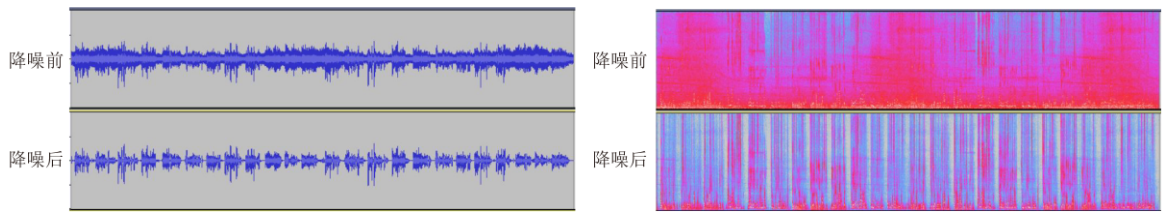


图 10 搭载WTM2101的耳机产品与WTM2101自动化部署流程



(a) 降噪前与降噪后的波形图　　(b) 降噪前与降噪后的频谱图

图 11 WTM2101降噪性能图

度依旧保持在0.99以上。如表3所示为WTM2101在语音激活检测、语音唤醒、命令词识别、环境去噪和声纹识别方面与市场同类产品的对比。

表 2　神经网络的累计余弦相似度

神经网络	累计余弦相似度
第0层	0.993
第1层	0.996
第2层	0.997
第3层	0.998
第4层	0.998
第5层	0.997
第6层	0.994
整个神经网络	0.994

表 3　WTM2101与市场同类产品的性能比较

标准	市场现有同类产品		WTM2101	
	算力复杂度(Mops)	功耗(mA)	算力复杂度(Mops)	功耗(mA)
语音激活检测	0.1	0.1	0.1	0.07
语音唤醒	20	0.6	400	0.4
40命令词识别	30	2	400	0.6
100命令词识别	100	10	400	0.8
环境去噪	150	15	800	1
声纹识别	1500	150	1500	2

5　存内计算芯片的应用前景与挑战

存内计算芯片技术，因其高算力、低功耗、低成本等优势，未来可为物联网、大数据和人工智能等具有海量数据特征的智能应用场景提供高能效硬件解决方案。但要实现大规模产业化仍存在诸多挑战：(1)模拟计算精度提升困难，模拟存内计算的精度受到信噪比的影响，很难做到8 以上。数字存内计算则不受信噪比的影响，但其能效、面积和成本需要综合权衡。近年来，通过数模混合的设计方式，可以在精度、成本与功耗之间得到很好的折中，是存内计算发展的一大重要方向。(2)工具链环节需更加完善以建立良好的生态：存内计算芯片产业化处于起步阶段，目前面临相关工具链支持不足的问题，导致算法/应用厂商移植困难。随着存内计算技术的快速发展，以及企业在这个技术领域持续加大投入，相应的编译、优化等工具链可以快速进步，有望建立初步的应用生态。(3)跨层协同设计需进一步加强：存内计算芯片涉及器件-芯片-工艺-算法-应用等多层次的跨层协同，各层环环相扣，密不可分，需要跨层协同来实现性能(精度、功耗、时延、可靠性等)与成本的最优。

参 考 文 献

[1] CHEN C.L.Philip and ZHANG Chunyang Data-intensive applications, challenges, techniques and technologies: A survey on Big Data[J]. *Information Sciences*, 2014, 275: 314–347. doi: 10.1016/j.ins.2014.01.015.

[2] WULF W A and MCKEE S A. Hitting the memory wall: Implications of the obvious[J]. *ACM SIGARCH Computer Architecture News*, 1995, 23(1): 20–24. doi: 10.1145/216585.216588.

[3] ZIDAN M A, STRACHAN J P, and LU W D. The future of electronics based on memristive systems[J]. *Nature Electronics*, 2018, 1(1): 22–29. doi: 10.1038/s41928-017-0006-8.

[4] 张和. 基于MRAM和SRAM的混合器件存算一体芯片设计[D]. [博士论文], 北京航空航天大学, 2021.
ZHANG He. Computing in memory chip design with hybrid devices based on MRAM and SRAM [D]. [Ph. D. dissertation], Beihang University, 2021.

[5] NAIR R, ANTAO S F, BERTOLLI C, *et al*. Active memory cube: A processing-in-memory architecture for exascale systems[J]. *IBM Journal of Research and Development*, 2015, 59(2/3): 17:1–17:14. doi: 10.1147/JRD.2015.2409732.

[6] AKIN B, FRANCHETTI F, and HOE J C. Data reorganization in memory using 3D-stacked DRAM[J]. *ACM SIGARCH Computer Architecture News*, 2015, 43(3S): 131–143. doi: 10.1145/2872887.2750397.

[7] FARMAHINI-FARAHANI A, AHN J H, MORROW K, *et al*. NDA: Near-DRAM acceleration architecture leveraging commodity DRAM devices and standard memory modules[C]. 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA), Burlingame, USA, 2015: 283–295. doi: 10.1109/HPCA.2015.7056040.

[8] GAO M, AYERS G, and KOZYRAKIS C. Practical near-data processing for in-memory analytics frameworks[C]. 2015 International Conference on Parallel Architecture and Compilation (PACT), San Francisco, USA, 2015: 113–124. doi: 10.1109/PACT.2015.22.

[9] KAUTZ W H. Cellular logic-in-memory arrays[J]. *IEEE Transactions on Computers*, 1969, C-18(8): 719–727. doi: 10.1109/T-C.1969.222754.

[10] STONE H S. A logic-in-memory computer[J]. *IEEE Transactions on Computers*, 1970, C-19(1): 73–78. doi: 10.1109/TC.1970.5008902.

[11] PATTERSON D, ANDERSON T, CARDWELL N, *et al*. Intelligent RAM (IRAM): Chips that remember and compute[C]. 1997 IEEE International Solids-State Circuits Conference. Digest of Technical Papers, San Francisco,

USA, 1997: 224–225. doi: 10.1109/ISSCC.1997.585348.

[12] KANG Yi, HUANG Wei, YOO S M, *et al*. FlexRAM: Toward an advanced intelligent memory system[C]. 1999 IEEE International Conference on Computer Design: VLSI in Computers and Processors, Austin, USA, 1999: 192–201. doi: 10.1109/ICCD.1999.808425.

[13] LI Shuangchen, NIU Dimin, MALLADI K T, *et al*. DRISA: A DRAM-based reconfigurable in-situ accelerator[C]. The 50th Annual IEEE/ACM International Symposium on Microarchitecture, Cambridge, UK, 2017: 288–301. doi: 10.1145/3123939.3123977.

[14] SKARLATOS D, KIM N S, TORRELLAS J. Pageforge: a near-memory content-aware page-merging architecture[C]// Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture. 2017: 302-314.

[15] AGRAWAL S R, IDICULA S, RAGHAVAN A, *et al*. A many-core architecture for in-memory data processing[C]. The 50th Annual IEEE/ACM International Symposium on Microarchitecture, Cambridge, UK, 2017: 245–258. doi: 10.1145/3123939.3123985.

[16] SEBASTIAN A, TUMA T, PAPANDREOU N, *et al*. Temporal correlation detection using computational phase-change memory[J]. *Nature Communications*, 2017, 8(1): 1115. doi: 10.1038/s41467-017-01481-9.

[17] BISWAS A and CHANDRAKASAN A P. CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks[J]. *IEEE Journal of Solid-State Circuits*, 2019, 54(1): 217–230. doi: 10.1109/JSSC.2018.2880918.

[18] LIU Qi, GAO Bin, YAO Peng, *et al*. 33.2 A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing[C]. 2020 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, USA, 2020: 500–502. doi: 10.1109/ISSCC19947.2020.9062953.

[19] ZHU Haozhe, JIAO Bo, ZHANG Jinshan, *et al*. COMB-MCM: Computing-on-memory-boundary NN processor with bipolar bitwise sparsity optimization for scalable multi-chiplet-module edge machine learning[C]. 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, USA, 2022: 1–3. doi: 10.1109/ISSCC42614.2022.9731657.

[20] TAN Fei, WANG Yiming, YANG Yiming, *et al*. A ReRAM-based computing-in-memory convolutional-macro with customized 2T2R bit-cell for AIoT chip IP applications[J]. *IEEE Transactions on Circuits and Systems II:Express Briefs*, 2020, 67(9): 1534–1538. doi: 10.1109/TCSII.2020.3013336.

[21] GUO Ruiqi, LIU Yonggang, ZHENG Shixuan, *et al*. A 5.1pJ/neuron 127.3us/inference RNN-based speech recognition processor using 16 computing-in-memory SRAM macros in 65nm CMOS[C]. 2019 Symposium on VLSI Circuits, Kyoto, Japan, 2019: C120–C121. doi: 10.23919/VLSIC.2019.8778028.

[22] WAN Weier, KUBENDRAN R, GAO Bin, *et al*. A voltage-mode sensing scheme with differential-row weight mapping for energy-efficient RRAM-based in-memory computing[C]. 2020 IEEE Symposium on VLSI Technology, Honolulu, USA, 2020: 1–2. doi: 10.1109/VLSITechnology18217.2020.9265066.

[23] SHEN Wensheng, HUANG Peng, WANG Xiangyu, *et al*. A novel capacitor-based stateful logic operation scheme for in-memory computing in 1T1RRAM array[C]. 2020 4th IEEE Electron Devices Technology & Manufacturing Conference (EDTM), Penang, Malaysia, 2020: 1–4. doi: 10.1109/EDTM47692.2020.9117832.

[24] YAO Peng, WU Huaqiang, GAO Bin, *et al*. Fully hardware-implemented memristor convolutional neural network[J]. *Nature*, 2020, 577(7792): 641–646. doi: 10.1038/s41586-020-1942-4.

[25] MERRIKH-BAYAT F, GUO Xinjie, KLACHKO M, *et al*. High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(10): 4782–4790. doi: 10.1109/TNNLS.2017.2778940.

[26] FICK L, BLAAUW D, SYLVESTER D, *et al*. Analog in-memory subthreshold deep neural network accelerator[C]. 2017 IEEE Custom Integrated Circuits Conference, Austin, USA, 2017: 1–4. doi: 10.1109/CICC.2017.7993629.

[27] MAHMOODI M R and STRUKOV D. An ultra-low energy internally analog, externally digital vector-matrix multiplier based on NOR flash memory technology[C]. The 55th Annual Design Automation Conference, San Francisco, USA, 2018: 33. doi: 10.1145/3195970.3195989.

[28] KANG Mingu, GONUGONDLA S K, PATIL A, *et al*. A multi-functional in-memory inference processor using a standard 6T SRAM array[J]. *IEEE Journal of Solid-State Circuits*, 2018, 53(2): 642–655. doi: 10.1109/JSSC.2017.2782087.

[29] YANG Jun, KONG Yuyao, WANG Zhen, *et al*. 24.4 sandwich-RAM: An energy-efficient in-memory BWN architecture with pulse-width modulation[C]. 2019 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, USA, 2019: 394–396. doi: 10.1109/ISSCC.2019.8662435.

[30] CHIU Y C, ZHANG Zhixiao, CHEN Jiajing, *et al*. A 4-Kb 1-to-8-bit configurable 6T SRAM-based computation-in-memory unit-macro for CNN-based AI edge processors[J]. *IEEE Journal of Solid-State Circuits*, 2020, 55(10):

2790–2801. doi: 10.1109/JSSC.2020.3005754.

[31] JIA Hongyang, VALAVI H, TANG Yinqi, *et al*. A programmable heterogeneous microprocessor based on bit-scalable in-memory computing[J]. *IEEE Journal of Solid-State Circuits*, 2020, 55(9): 2609–2621. doi: 10.1109/JSSC.2020.2987714.

[32] JIANG Zhewei, YIN Shihui, SEO J S, *et al*. C3SRAM: An in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism[J]. *IEEE Journal of Solid-State Circuits*, 2020, 55(7): 1888–1897. doi: 10.1109/JSSC.2020.2992886.

[33] YIN Shihui, JIANG Zhewei, SEO J S, *et al*. XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks[J]. *IEEE Journal of Solid-State Circuits*, 2020, 55(6): 1733–1743. doi: 10.1109/JSSC.2019.2963616.

[34] YAN Bonan, YANG Qing, CHEN Weihao, *et al*. RRAM-based spiking nonvolatile computing-in-memory processing engine with precision-configurable in situ nonlinear activation[C]. 2019 Symposium on VLSI Technology, Kyoto, Japan, 2019: T86-T87. doi: 10.23919/VLSIT.2019.8776485.

[35] XUE Chengxin, CHEN Weihao, LIU J S, *et al*. 24.1 A 1Mb multibit RRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN based AI edge processors[C]. 2019 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, USA, 2019: 388–390. doi: 10.1109/ISSCC.2019.8662395.

[36] XUE Chengxin, CHEN Weihao, LIU J S, *et al*. Embedded 1-Mb ReRAM-based computing-in-memory macro with multibit input and weight for CNN-based AI edge processors[J]. *IEEE Journal of Solid-State Circuits*, 2020, 55(1): 203–215. doi: 10.1109/JSSC.2019.2951363.

[37] ZHA Yue, NOWAK E, and LI Jing. Liquid silicon: A nonvolatile fully programmable processing-in-memory processor with monolithically integrated ReRAM[J]. *IEEE Journal of Solid-State Circuits*, 2020, 55(4): 908–919. doi: 10.1109/JSSC.2019.2963005.

[38] ZHANG H, LIU J, BAI J, *et al*. HD-CIM: Hybrid-Device Computing-In-Memory Structure Based on MRAM and SRAM to Reduce Weight Loading Energy of Neural Networks[J]. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2022, 69(11): 4465–4474.

[39] SEBASTIAN A, LE GALLO M, KHADDAM-ALJAMEH R, *et al*. Memory devices and applications for in-memory computing[J]. *Nature Nanotechnology*, 2020, 15(7): 529–544. doi: 10.1038/s41565-020-0655-z.

[40] DONG Qing, SINANGIL M E, ERBAGCI B, *et al*. 15.3 A 351TOPS/W and 372.4GOPS compute-in-memory SRAM macro in 7nm FinFET CMOS for machine-learning applications[C]. 2020 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, USA, 2020: 242–244. doi: 10.1109/ISSCC19947.2020.9062985.

[41] CHIH Y D, LEE P H, FUJIWARA H, *et al*. 16.4 an 89TOPS/W and 16.3TOPS/mm^2 all-digital SRAM-based full-precision compute-in memory macro in 22nm for machine-learning edge applications[C]. 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, USA, 2021: 252–254. doi: 10.1109/ISSCC42613.2021.9365766.

[42] YAN Bonan, HSU J L, YU Pangcheng, *et al*. A 1.041-Mb/mm^2 27.38-TOPS/W Signed-INT8 dynamic-logic-based ADC-less SRAM compute-in-Memory Macro in 28nm with reconfigurable bitwise operation for AI and embedded applications[C]. 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, USA, 2022: 188–190. doi: 10.1109/ISSCC42614.2022.9731545.

[43] SESHADRI V, LEE D, MULLINS T, *et al*. Ambit: In-memory accelerator for bulk bitwise operations using commodity DRAM technology[C]. The 50th Annual IEEE/ACM International Symposium on Microarchitecture, Cambridge, UK, 2017: 273–287. doi: 10.1145/3123939.3124544.

[44] NATSUI M, SUZUKI D, SAKIMURA N, *et al*. Nonvolatile logic-in-memory LSI using cycle-based power gating and its application to motion-vector prediction[J]. *IEEE Journal of Solid-State Circuits*, 2015, 50(2): 476–489. doi: 10.1109/JSSC.2014.2362853.

[45] ZHANG He, KANG Wang, CAO Kaihua, *et al*. Spintronic processing unit in spin transfer torque magnetic random access memory[J]. *IEEE Transactions on Electron Devices*, 2019, 66(4): 2017–2022. doi: 10.1109/TED.2019.2898391.

[46] ZHANG He, KANG Wang, WANG Lezhi, *et al*. Stateful reconfigurable logic via a single-voltage-gated spin hall-effect driven magnetic tunnel junction in a spintronic memory[J]. *IEEE Transactions on Electron Devices*, 2017, 64(10): 4295–4301. doi: 10.1109/TED.2017.2726544.

[47] WANG Haotian, KANG Wang, PAN Biao, *et al*. Spintronic computing-in-memory architecture based on voltage-controlled spin–orbit torque devices for binary neural networks[J]. *IEEE Transactions on Electron Devices*, 2021, 68(10): 4944–4950. doi: 10.1109/TED.2021.3102896.

[48] DEAVILLE P, ZHANG Bonan, CHEN L Y, *et al*. A maximally row-parallel MRAM in-memory-computing macro addressing readout circuitsensitivity and area[C]. ESSCIRC 2021-IEEE 47th European Solid State Circuits Conference (ESSCIRC), Grenoble, France, 2021: 75–78. doi: 10.1109/ESSCIRC53450.2021.9567807.

[49] JUNG S, LEE H, MYUNG S, *et al*. A crossbar array of

magnetoresistive memory devices for in-memory computing[J]. *Nature*, 2022, 601(7892): 211–216. doi: 10.1038/s41586-021-04196-6.

[50] SONG K M, JEONG J S, PAN Biao, *et al*. Skyrmion-based artificial synapses for neuromorphic computing[J]. *Nature Electronics*, 2020, 3(3): 148–155. doi: 10.1038/s41928-020-0385-0.

[51] CHEN Chao, LIN Tao, NIU Jianteng, *et al*. Surface acoustic wave controlled skyrmion-based synapse devices[J]. *Nanotechnology*, 2022, 33(11): 115205. doi: 10.1088/1361-6528/ac3f14.

[52] HUANG Yangqi, KANG Wang, ZHANG Xichao, *et al*. Magnetic skyrmion-based synaptic devices[J]. *Nanotechnology*, 2017, 28(8): 08LT02. doi: 10.1088/1361-6528/aa5838.

[53] HU Hanwen, WANG Weichen, CHEN C K, *et al*. A 512 Gb in-memory-computing 3D-NAND flash supporting similar-vector-matching operations on edge-AI devices[C]. 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, USA, 2022: 138–140. doi: 10.1109/ISSCC42614.2022.9731775.

[54] KIM M, LIU Muqing, EVERSON L R, *et al*. An embedded NAND flash-based compute-in-memory array demonstrated in a standard logic process[J]. *IEEE Journal of Solid-State Circuits*, 2022, 57(2): 625–638. doi: 10.1109/JSSC.2021.3098671.

郭昕婕：女，博士，研究方向为存算一体芯片设计.

王光耀：男，硕士生，研究方向为存算一体芯片设计.

王绍迪：男，博士，研究方向为存储器及存算一体架构设计.

责任编辑：马秀强