计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330364
61(12): 2937−2951, 2024

基于 SRAM 和 NVM 的存内计算技术综述

张　章[1]　施　刚[1]　王启帆[1]　马永波[1]　刘　钢[2]　钱利波[3]
[1](合肥工业大学微电子学院　合肥　230601)
[2](上海交通大学电子信息与电气工程学院　上海　200240)
[3](西安电子科技大学杭州研究院　杭州　311231)
(zhangzhang@hfut.edu.cn)

Survey of In-Memory Computing Technology Based on SRAM and Non-Volatile Memory

Zhang Zhang[1], Shi Gang[1], Wang Qifan[1], Ma Yongbo[1], Liu Gang[2], and Qian Libo[3]
[1](*School of Microelectronics*, *Hefei University of Technology*, *Hefei* 230601)
[2](*School of Electronic Information and Electrical Engineering*, *Shanghai Jiao Tong University*, *Shanghai* 200240)
[3](*Hangzhou Institute of Technology*, *Xidian University*, *Hangzhou* 311231)

Abstract　The fast and low-power in-memory computing architecture, which integrates memory and calculation, breaks through the traditional von-Neumann system that separates memory and calculation, and solves the problem of "memory wall" that limits the arithmetic power of the processor, which has become a research hotspot of new computing architecture. The basic devices for in-memory computing include fast and mature static random access memory (SRAM), low power, fast response and non-volatile memristor, and high density, low static power and non-volatile magnetic random access memory (MRAM). Up to the present, a great variety of in-memory computing studies have been proposed based on these devices, however, a systematic and comprehensive literature review on these in-memory computing architectures is still missing. In this paper, we firstly introduce the in-memory computing principles of different devices, the current development status of in-memory computing architectures, and the practical application scenarios from the three directions of SRAM, memristor, and MRAM. Next, the existing solutions and the future directions for the problems and challenges of current in-memory computing architectures are given. Finally, we summarize the research focus of in-memory computing based on the above devices, outline the shortcomings of the current research, and look forward to the future development direction.

Key words　non-von Neumann; SRAM; memristor; MRAM; in-memory computing

摘　要　集存储与计算于一身的快速低功耗存内计算架构,突破了存储与计算分离的传统冯·诺依曼体系,解决了限制处理器算力的"内存墙"问题,成为新型计算架构的研究热点.存内计算的基础器件包括高速且工艺成熟的静态随机存取存储器(static RAM, SRAM)、低功耗高响应且非易失的忆阻器(memristor)、高密度低静态功耗非易失的磁性随机存取存储器(magnetic RAM, MRAM).研究者们基于上述器件完成大量存内计算研究,但是关于这些存内计算架构全面且系统总结的文献综述仍然缺失.首先从 SRAM、忆阻器、MRAM 方向出发概述了不同器件的存内计算原理、当前存内计算架构发展状况和实际应用场景

收稿日期: 2023-05-15; **修回日期**: 2023-09-26
基金项目: 国家自然科学基金项目(U19A2053, 62111540271); 浙江省自然科学基金项目(LDT23F04022F04)
This work was supported by the National Natural Science Foundation of China (U19A2053, 62111540271) and the Zhejiang Provincial Natural Science Foundation of China (LDT23F04022F04).
通信作者: 刘钢(gang.liu@sjtu.edu.cn)

等. 然后针对当前存内计算架构存在的各种问题和挑战给出了现有解决方案和未来解决方向. 最后对基于以上器件的存内计算研究重点进行了总结并概述了目前的研究短板、展望未来的发展方向.

关键词 非冯·诺依曼; 静态随机存取存储器; 忆阻器; 磁性随机存取存储器; 存内计算

中图法分类号 TP391

伴随着大数据时代的到来, 基于传统的冯·诺依曼架构的 CPU 在处理海量数据与大规模并行计算面前显得力不从心. 根据摩尔定律, 随着技术的进步, CPU 的性能每年大约会提升 60%[1]; 另一方面, 当前计算机使用的主要存储单元是动态随机存取存储器(dynamic RAM, DRAM). 当 DRAM 长宽比降低到一定程度时, 由于电容性泄漏等原因无法从特征尺寸的缩小中获得更多的优势, 性能平均每年提升 7%[1]. 存储单元的性能改进速度远远不能满足 CPU, 这种内存改进速度和 CPU 速度之间的不匹配被称为“内存墙”. 内存墙存在的原因与冯·诺依曼架构密切相关, 当今冯·诺依曼架构中 CPU 性能达到了 DRAM 的 80 多倍, 导致了数据处理时 DRAM 读写带来了巨大的延迟[2]. 不仅如此, 分离的计算模块和存储模块带来了数据传输功耗[2]. 传统的 CPU 并不适应人工智能(artificial intelligence, AI)时代的要求, 而集存储与计算于一体的存内计算架构则避免了数据传输带来的延迟和功耗, 有着更高的效率. 因此, 新兴的存内计算成为了当前学者们研究的重要方向.

存内计算常应用于各种智能 AI 场景, 包括云计算的边缘任务加速等[3]. 目前常见的存内计算芯片还达不到云计算的大算力和大吞吐量的要求. 而由于乘加计算是存内计算最基础的计算方式, 因此有着大量乘加计算的神经网络非常适合使用存内计算进行加速, 研究者们也重点研究了神经网络和存内计算二者的结合. 因此本文在存内计算应用方面也以神经网络加速为主. 目前已有的存内计算综述往往针对某一种器件进行介绍, 本文则介绍了 3 种器件的存内计算原理以及应用场景和现有架构, 并总结了各自的优缺点.

传统的存储器件例如 DRAM, 由于使用其构建复杂的逻辑电路具有一定的挑战性, 故其适用于近内存计算的实现[4-5]. 静态随机存取存储器(static RAM, SRAM)由于其良好的稳定性、成熟的工艺、较快的读取速度, 已经被广泛用于图形加速器或机器学习[6-7]. 闪存(flash memory)则因为其读写性能较差以及可擦写次数少, 并不适合用于大算力的存内计算架构. 新兴的存储器件例如忆阻器(memristor)和磁性随机存储器(magnetic RAM, MRAM)也逐渐崭露头角. 忆阻器[8]是一种新型的纳米存储器件, 器件的阻值与流经器件的电荷相关. 忆阻器以其非易失性、低延迟、低功耗成为了实现存内计算的重要方向[9-10]. MRAM 则同时具有 SRAM 的读写速度和 DRAM 的高集成度, 且理论上可以无限次地写入数据[11-12]. 因此, MRAM 理所当然地成为了存内计算的重要实现方式. 这几种器件的性能对比如表 1 所示[13].

Table 1 Performance Comparison of Several Popular Memories

表 1 几种常用存储器的性能对比

参数	SRAM	忆阻器	MRAM	DRAM	闪存
尺寸/F^2	120~200	4~10	6~50	6~10	4~6
读延迟/ns	1~8	10	5~10	10~60	2.5×10^4
写延迟/ns	8	10	12	10~60	2×10^5
易失性	是	否	否	是	是
耐久性次数	$>10^{15}$	10^{11}	$>10^{15}$	$>10^{15}$	10^4~10^5

基于以上器件的特点, 本文将从 SRAM、忆阻器、MRAM 这 3 种存储器以及三者的片上集成, 综述存内计算的研究进展以及存在的问题, 并展望未来存内计算的发展方向.

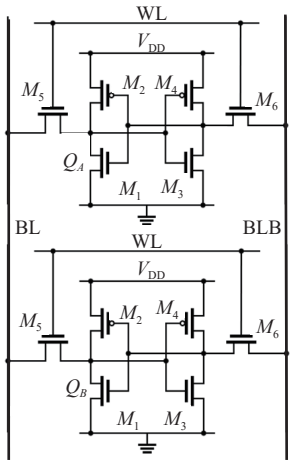
1 基于 SRAM 的存内计算

2016 年, Jeloka 等人[14]提出了基于 SRAM 的存内逻辑计算. 随后, 大量研究基于此原理进行了更深入的研究. 同时, 根据 SRAM 的存内逻辑运算原理, SRAM 又被用于神经网络的硬件加速.

1.1 基于 SRAM 的存内逻辑运算

SRAM 的存内逻辑运算通过激活 SRAM 同一列的 2 个或多个存储单元的字线, 并通过灵敏放大器感测位线电压得到了这些存储单元存储比特的逻辑运算结果[15-17]. 然后在增加额外的逻辑门后, 其可以实现逻辑或非和逻辑与非运算. 其原理如图 1 所示, 将同一列上的 2 个存储单元的字线同时打开(即横杆和纵杆的逻辑电压均为 1), 则位线 BL 上的逻辑值为这 2 个存储单元所存储的值逻辑与运算的结果, 位

线 BLB 上的逻辑值为这 2 个存储的值逻辑或非运算的结果. 与传统的 SRAM 阵列相比, 新的阵列有着更高的密度和更低的功耗. 由此, 更多的研究者基于此原理提出了更多的 SRAM 存内计算架构. Aga 等人[18]在此基础上提出了一种新的存内计算架构, 该架构通过添加解码器和使用单端灵敏放大器实现了逻辑异或运算. Dong 等人[19]提出了一种 4+2T 的 SRAM 单元, 具有比 6T SRAM 单元更好的噪声容限, 其中 T 指晶体管(transistor).



(a) 基于SRAM的存内逻辑计算架构电路

Q_A	Q_B	BL	BLB
0	0	0	1
0	1	0	0
1	0	0	0
1	1	1	0

(b) 基于SRAM的存内逻辑计算逻辑真值表

Fig. 1 Principles of SRAM-based in-memory logic operations

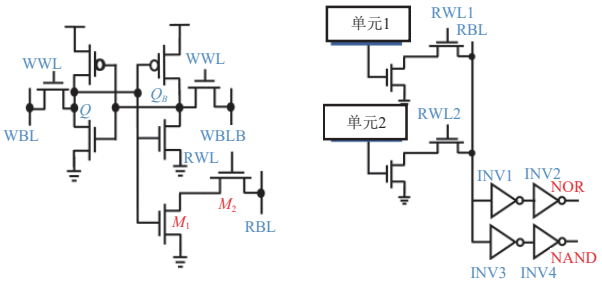
图 1 基于 SRAM 的存内逻辑操作原理

针对传统的 6T SRAM 单元所存在的读写干扰问题和存储内容翻转问题[20], 8T SRAM 和 10 T SRAM 单元被研究者们提出. Agrawal 等人[21]提出了使用 8T SRAM 单元和 8+T SRAM 单元的解耦读写路径来实现存内布尔运算, 如图 2 所示, 成功实现了逻辑与非(NAND)、逻辑或非(NOR)、逻辑异或(XOR)等逻辑运算. 相比于 6T SRAM 单元, 8T SRAM 单元提高了数据的吞吐量和数据的处理速度. Rajput 等人[22]提出一种 8T SRAM 单元与算术电路协同的架构, 具有更高的能量利用率和读取裕量. Chen 等人[23]提出了一种具有额外差分 PMOS 接入晶体管的 8T SRAM 结构, 这种存内计算架构速度更快且更可靠, 而且可以实现更复杂的复合布尔逻辑运算.

除了使用更多晶体管, 使用读写分离的 SRAM 单元来降低读写干扰问题也被证明是可行的[24-27].

1.2 SRAM 应用于神经网络

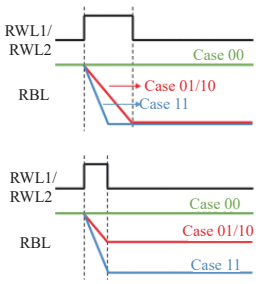
在神经网络算法中, 进行得最多的计算便是乘加计算. 研究者们提出二值化神经网络(binary neural networks, BNN), 将输入和权重进行二值化, 输入和权重只能为 1 或−1, 这样基于 SRAM 的乘法便可以



(a) 8T SRAM单元 (b) NOR和NAND操作

A	B	NOR	NAND	XOR
0	1	1	1	0
0	1	0	1	1
1	1	0	1	1
1	0	0	0	0

(c) 布尔运算真值表



(d) 布尔运算时序图

Fig. 2 Boolean operation of 8T SRAM

图 2 8T SRAM 的布尔运算

视为逻辑同或运算[28-30].

基于 SRAM 的存内乘法计算需要将输入 V 连接到字线上作为操作数, 另一个操作数 W 则存储在 SRAM 单元里, 其结果由位线得出, 其乘法真值表如表 2 所示. 然而, 使用 1 b 权重会带来较大的精度损失. 因此研究者们将目光投向了并行计算, 以此实现多位权重运算[31-32]. 文献 [32] 提出的并行计算如图 3 所示, 其中包括 WL 开关矩阵用于激活多行字线、用于闪存 ADC 的多电平检测器(MLSA)、用于生成基准电压的参考生成器.

Table 2 Multiplication Truth Table

表 2 乘法真值表

V	W	$V\times W$
−1	−1	1
−1	1	−1
1	−1	−1
1	1	1

Si 等人[33]提出了一种基于 6T SRAM 的双拆分结构实现了完全并行的乘积以及累加和计算并流片证明. 测试结果表明, 使用此种架构可以实现全连接层计算时间 2.3 ns. 不仅如此, 其能效最高可达 55.8TOPS/W. 但是这种架构使用了大量的晶体管, 导致面积较大.

Nguyen 等人[34]提出了用于深度神经网络(deep neural network, DNN)存内计算的 10T SRAM 并设计

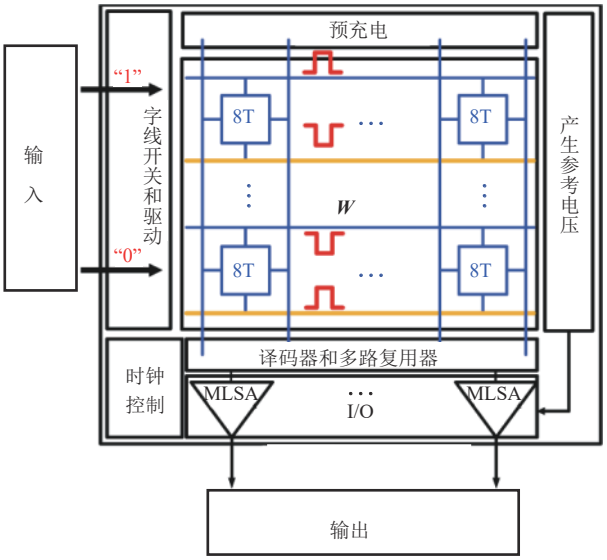


Fig. 3 SRAM array for parallel computing

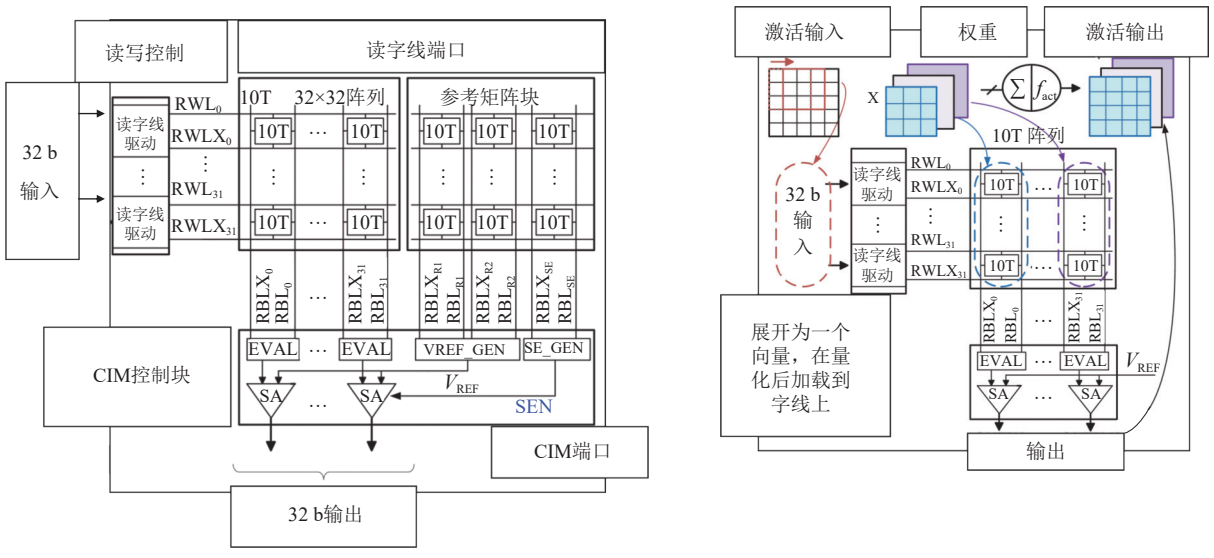
图 3 并行计算的 SRAM 阵列

了整体架构成功映射了 LeNet-5 手写数字识别网络. 该架构支持完全并行的乘加计算并且实现了 4 b 权重、4 b 输入、8 b 输出的乘加操作. 如图 4 所示, 32 个输入被分别输入进一个 32×32 的 10T SRAM 阵列中. 输入向量与权重矩阵做完乘法后, 同一列上的所有 SRAM 单元的位线上的电流都将在 EVAL 单元中求和并转换成模拟电压. 敏感放大器模块将此模拟电压和参考电压对比生成数字信号的输出, 而参考电压则由 3 列 10T SRAM 组成的参考阵列块生成, 同时此参考阵列块还生成检测放大器 SA 的检测使能信号. 如此便完成了矩阵和向量之间的乘加运算. 此架构通过 28 nm CMOS 工艺实现, 并且拥有高能效和高吞吐量的优势.

Su 等人[35] 提出了可用于 2~8 b 运算的双向转置 6T SRAM, 其支持神经网络推理和训练的过程. 通过使用双向转置乘法单元实现了前向传播和反向传播过程, 提升了神经网络的精度. 不仅如此, 在对最大乘加结果进行预测并根据此预测结果来权衡信号裕量和模拟域位线信号的线性后, 增强了乘加运算信号裕量.

在 DNN 训练的过程中, 数据 0 的乘加计算会产生不必要的功耗. 为了进一步降低功耗, 稀疏性处理也被研究者们所青睐. Han 等人[36] 提出了一种基于 8T SRAM, 不仅具有前向传播和后向传播, 还具有稀疏性处理功能的 DNN 加速器. 通过只存储非 0 值及其地址的方式来过滤 0 值, 使得不必要的计算被跳过, 最后实现了功耗上 3.2%~28.7% 的降低.

除了上述工作, 近年来, 国内外众多优秀团队为 SRAM 的存内计算做出了贡献. Nasrin 等人[37] 基于 8T SRAM 提出了一种用于 DNN 内存计算推理的协同设计方法, 使用了一种无乘法函数逼近器以及配套的计算流程. Iqbal 等人[38] 提出了一种 10T1C 结构的 SRAM 单元, 其有着更高的单元线性度和数据吞吐量, 用于 DNN 的加速. Yan 等人[39] 提出了一种无 ADC 的动态 SRAM 架构, 极大地加快了 SRAM 单元的读写速度. Choi 等人[40] 使用 8T1C 的 ARAM 单元实现了完全并行的一步多位计算, 提升了计算速度并且减少了功耗.



(a) 基于10T SRAM 的乘加框架

(b) 基于乘加框架的权重映射

Fig. 4 10T SRAM based DNN inference framework

图 4 基于 10T SRAM 的 DNN 推理框架

2 基于忆阻器的存内计算

忆阻器，早在 1971 年就被蔡少棠教授推测出来，但由于当时并没有找到相对应的器件所以并没有引起研究者的重视. 在 2008 年，惠普公司的研究人员提出了 Pt/TiO_2/Pt 三明治叠层的忆阻器器件，引起了研究者的广泛关注. 忆阻器的高阻低阻状态可表示成 0 和 1 进行数据的保存和逻辑运算，而多值忆阻器的出现则使得忆阻器在存内类脑运算中备受关注.

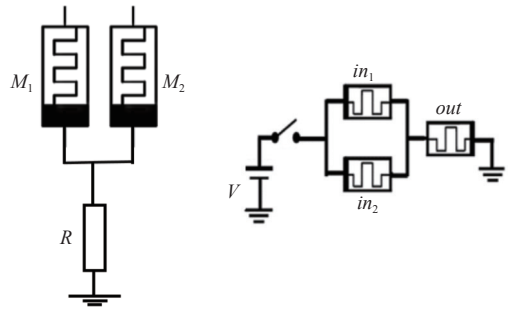
2.1 基于忆阻器的存内逻辑运算

2010 年，Borghetti 等人[41]提出了忆阻器实质蕴含逻辑(material implication, IMP)，并通过使用忆阻器件作为锁存器和逻辑门来启用有状态逻辑. IMP 的真值表如表 3 所示，忆阻器高阻值视为逻辑 0，低阻值视为逻辑 1. 其基础的 IMP 逻辑单元如图 5(a)所示，

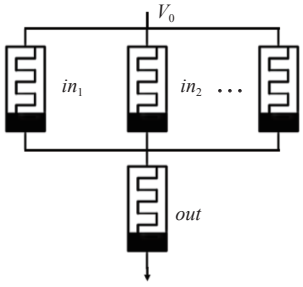
Table 3 Truth Table of IMP

表 3 实质蕴含逻辑真值表

M_1	M_2	输出
0	0	1
0	1	1
1	0	0
1	1	1



(a) 实质蕴含逻辑运算 (b) MAGIC或非门



(c) FELIX输入或非门

Fig. 5 Calculation principle of in-memory logic of memristor

图 5 忆阻器存内逻辑计算原理

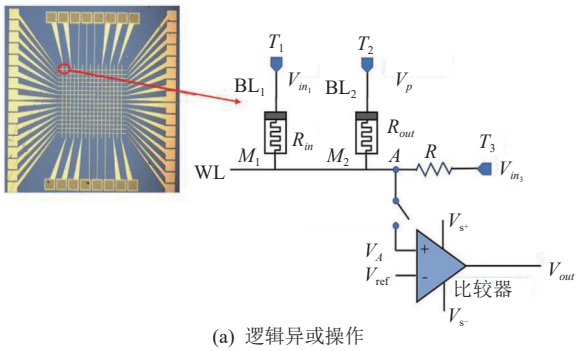
其由 2 个并联的忆阻器与 1 个电阻串联后接地构成. 具体的 IMP 运算过程为：首先根据输入给忆阻器 M_1，M_2 施加 1 个置逻辑 0 的正电压或者置逻辑 1 的负电压，使得 M_1，M_2 处于我们要进行计算的状态. 随后给忆阻器 M_1 的正端施加一个忆阻器状态转变的阈值电压，给忆阻器 M_2 的正端则施加一个相对较小的电压. 则电路分压后，忆阻器 M_2 的状态即为此次 IMP 运算结果. 通过级联的方式，使用这种方法可以实现所有的布尔运算. 但是这种方法导致了一个输入被输出所替换，等于丢失了这个输入，当需要复用输入时会带来很大的不便，降低了逻辑运算的效率.

Kvatinsky 等人[42]提出了一种对忆阻器有状态逻辑的改进设计方法，即忆阻器辅助逻辑(MAGIC). 在 MAGIC 设计方法中，输入单元和输出单元被分开，使得输入不会出现丢失的情况. MAGIC 门的输入是负责输入的忆阻器的初始状态，输出为负责保存输出的忆阻器的状态. 以 MAGIC 逻辑或非门举例来说，如图 5(b)所示，由 2 个忆阻器并联作为 2 个输入与作为输出的忆阻器串联. MAGIC 门的操作分为 2 个步骤：首先是输出忆阻器初始化为一个已知的逻辑状态，然后施加电压 V 给输入忆阻器，输出将保存为输出忆阻器的状态.

基于 MAGIC 方法，Jang 等人[43]提出了 MAGIC 逻辑非门和 MAGIC 逻辑或非门并成功地在 1S1M 阵列中实现了并行的逻辑运算. Gupta 等人[44]提出了一种更高效的存内逻辑计算方法，即 FELIX 方法. 如图 5(c)所示，其操作方法与 MAGIC 方法类似. 以逻辑或非操作举例，输出忆阻器初始化为低阻，输入电压输入到忆阻器的一端，其输出连接到输出忆阻器上，输出忆阻器另一端连接到地. 当逻辑或非的结果为 0 时，改变输出忆阻器的状态. 与 MAGIC 方法相比，FELIX 方法不需要使用额外的辅助函数来完成逻辑运算，且具有更快的速度和更高的效率.

基于以上设计方法，研究者们提出了更多基于忆阻器的逻辑器件. Xu 等人[45]提出了混合忆阻器专用的逻辑异或门、逻辑同或门、逻辑或非门. Liu 等人[46]提出了基于忆阻器和 CMOS 的进位前瞻加法器. Teimoory 等人[47]提出了基于忆阻器的乘法器设计. Rohani 等人[48]提出了基于忆阻器 IMP 原理的半并行全加器. Ali 等人[49]基于 MAGIC 设计方法提出了忆阻器阵列内高效乘法的算法. Li 等人[50]为了提高忆阻器阵列空间维度上的灵活性提出了 SCMOS 逻辑，其对忆阻器阵列中忆阻器的位置有着更低的要求.

目前研究的重要方向还是集中在对忆阻器逻辑计算算法和架构的改进. Song 等人[51]基于 IMP 设计方法提出了一种新的内存逻辑运算的架构 V/R-R, 这种架构融合了 R-R 逻辑架构和 V-R 逻辑架构, 同时具有 2 种架构的优势. 该架构不仅级联性更优, 且需要的忆阻器数量更少、功耗更低. 如图 6 所示, 当其执行逻辑异或操作时分 2 个部分完成. 首先, 默认的忆阻器状态为高阻值状态, 忆阻器 M_1 根据输入改变其状态, 忆阻器 M_2 维持高阻值状态. 然后给端口 T_1, T_2, T_3 施加对应的电压, T_2 上的电压应为一个常量, 运算结果以电压的形式输出. 相比较于文献 [37] 提出的 IMP 方法, 这种架构不会使输入出现丢失的情况, 且在复杂逻辑运算时更快速. 此外, Song 等人还提出了一种基于忆阻器的可行的并行全加法器.



(a) 逻辑异或操作

$p\rightarrow(V_{in_1}, V_{in_3})$	$q\rightarrow R_{in}$	$out\rightarrow(V_{out}, R_{out})$
0	0	0
0	1	1
1	0	1
1	1	0

(b) 逻辑异或真值表

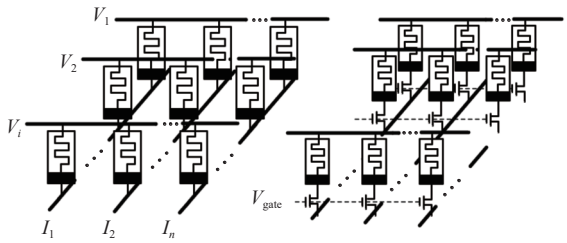
Fig. 6 V/R-R architecture and its truth table of logic XOR operation

图 6 V/R-R 架构及其逻辑异或操作真值表

近年来, 研究者们依然对忆阻器存内计算抱有很大热情并积极推动技术的进步与发展. Pandey 等人[52]提出了基于忆阻器和 CMOS 的比较器电路, 相比较于传统比较器有着更小的功耗和更快的速度. Paramasivam 等人[53]基于忆阻器提出了一种 2 b 的比较器, 其有着更低的功耗和更高的单位面积效率. Biswas 等人[54]针对忆阻器在电路中可能出现的逻辑故障提出了新的验证和测试方法. Yang 等人[55]提出了基于忆阻器和 MOSFET 的三元逻辑电路和单元. Wu 等人[56]使用 ZnO 忆阻器实现了三值加法器电路. Shanmukh 等人[57]提出了基于忆阻器的低功耗 3 b 编码器设计.

2.2 基于忆阻器的存内类脑运算

忆阻器用于神经网络加速时, 往往以阵列的形式将权重保存为器件的阻值, 输入需要量化为电压, 运算结果以电流的形式输出. 常见的忆阻器阵列如图 7 所示. 以交叉阵列举例, 输入电压 V_i 可视为输入向量, 忆阻器阵列电导 g_{jk} 可编程为需要相乘的矩阵, 每列忆阻器的电流会自动累加为电流 I_n, I_n 的值为 $\sum g_{jk}\times V_i$. 这样就完成了矩阵乘加计算.



(a) 交叉阵列 (b) 1T1R阵列

Fig. 7 Common memristor array applied for neural network acceleration

图 7 常见的用于神经网络加速的忆阻器阵列

由于计算的并行度高、功耗低速度快, 使用忆阻器存内计算来加速神经网络很快吸引了大量研究者. Soudry 等人[58]提出了基于忆阻器的多层神经网络可以进行在线梯度下降训练. Chabi 等人[59]提出并模拟了一个具有片上监督学习规则的忆阻器神经网络, 其采用交叉阵列实现. Zheng 等人[60]提出了基于忆阻器交叉阵列的尖峰神经网络(spike neural network, SNN), 其学习方式是尖峰时序依赖可塑性(spike timing dependent plasticity, STDP)规则, 在 Mnist 数据集上测试得到了 97.1% 的准确率. Yakopcic 等人[61]使用忆阻器实现了具有 Sobel 边缘检测功能的多层感知器, 实验表明其能够以 337 mW 的功耗实时处理 4K UHD 视频. Zhong 等人[62]提出了一种基于忆阻器的动态 RC 系统, 通过调整系统参数实现了对时间信号的有效处理.

Hong 等人[63]提出了一种用于人像修复的复值 Hopfield 神经网络的存内计算电路. 实验结果表明, 该网络有着 97% 的高精度和 0.1 ms 的快速恢复时间, 具备抗干扰能力较强、误差容忍度较高等优势, 为人像修复提供了一个可行高效的方案. 不仅如此, 该团队还提出了基于忆阻器的离散余弦变换 2 维电路用于图像压缩[64]. 该工作提出了一种一步式计算方法, 能够在降低功耗的同时大大加快运算速度. 最后的实验结果表明该电路的计算精度达到 99% 以上, 在

图像压缩领域有着快速且良好的效果.

除了上述工作，研究者们还积极推动忆阻器片上集成的发展[65-67]. Yao 等人[68]提出了完全硬件实现忆阻器卷积神经网络，使用 Mnist 数据集测试可以达到 96% 的识别效果. 如图 8(a)所示，该硬件系统片上集成了 8 个 2 048 单元的忆阻器，并使用了一种软硬件混合训练方法来调整硬件上的缺陷和误差. Wan 等人[69]提出了一种基于 RRAM 的存内计算芯片 NeuRRAM，可以为不同的模型架构重新配置 CIM 内核，能源效率与先进芯片相比提高了 2 倍，同时也有着较高的推理精度，如图 8(b)所示. 该芯片在处理 Mnist 数据集识别时精度达到 99%，且在语音识别、图像恢复上都有着优异的表现. 2 种提出的系统对比如表 4 所示.

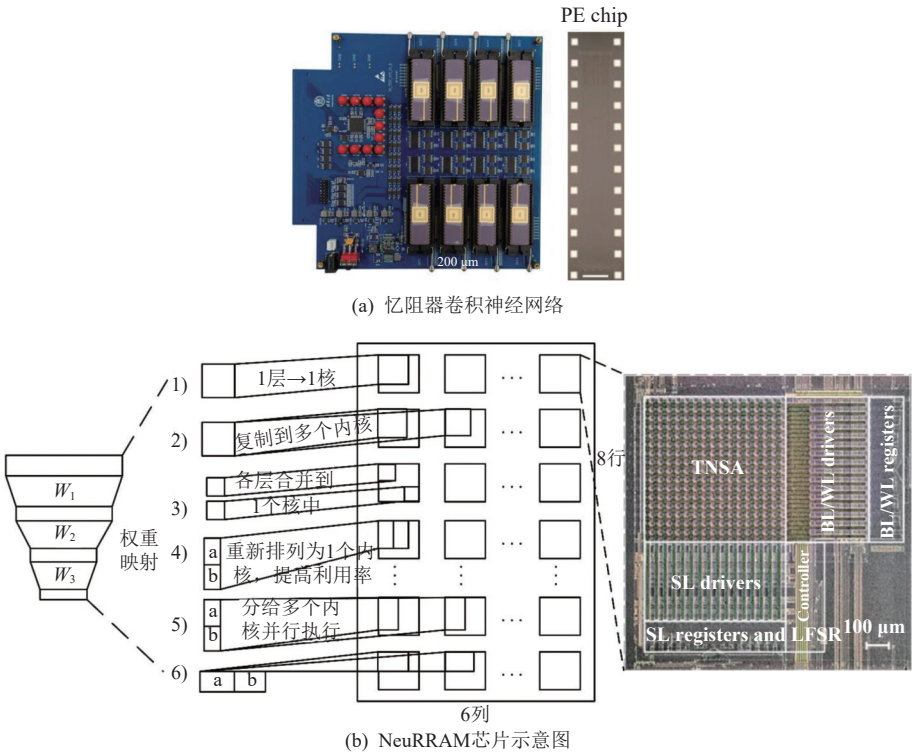


Fig. 8 Hardware neural network based on memristor

图 8 基于忆阻器的硬件神经网络

Table 4 Comparison of Memristor-Based Hardware Systems

表 4 基于忆阻器的硬件系统对比

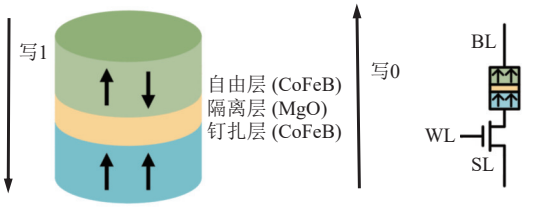
来源	年份	阵列尺寸	权重位数/b	工艺/nm	速度/GOPS
文献 [61]	2019	1 024×512	8	22	34
文献 [62]	2021	256×512	3	55	10
文献 [64]	2023	2 048	4	130	81.92
文献 [65]	2021	48×256×256	4	130	754

在忆阻器的实际使用中，总是存在着一定的偏差，例如权重写入偏差. 这些偏差在忆阻器神经网络中明显降低了网络的精度. 针对这个问题，Gao 等人[70]提出了一个统一的贝叶斯推理框架，将忆阻器硬件的偏差与软件算法联系在一起，很大程度上降低了硬件偏差带来的网络精度损失. Mohan 等人[71]基于一个 4×4 忆阻器交叉阵列的 STDP 学习实现了片上偏移校准. Qin 等人[72]基于二值忆阻器提出了高鲁棒性的 BNN 网络加速器，实验表明即使在高输入摆动和噪声的干扰下，其网络精度也超过 90%. Fu 等人[73]研究了忆阻器电导水平与周期变化的关系，使用的最佳水平数会明显地减少忆阻器周期变化带来的干扰并节省能量和减少延时.

合肥工业大学的 Zhang 等人[74]提出了基于 FPGA 的栅极控制忆阻器模型，可以更好地实现高精度多状态的计算. 该团队成功地使用了这种模型搭建了 4 态量化感知器，经过测试后发现该模型与软件的训练精度相差仅 0.3%. 同时，还研究了基于忆阻器的门控递归单元用于处理时间序列数据[75]. 该单元有效地缓解了 RNN 网络收敛的问题，并减少了外围电路的参与，实现了对时间序列高精度的处理.

3 基于 MRAM 的存算一体

MRAM, 因其成熟的技术和良好的 CMOS 兼容性而被视为实现高速、低功耗存内计算有前途的候选者[76-77]. 最初 MRAM 通过磁场对信息进行写入, 故而其又称为磁场驱动型 MRAM. 此时的 MRAM 还存在着写入问题和信息保存易出错的问题, 因此自旋传递转矩磁随机存取存储器(spin transfer torque magnetic RAM, STT-MRAM)被提出以缓解这些挑战, 其结构如图 9 所示. STT-MRAM 的基本存储元件是垂直磁各向异性(perpendicular magnetic anisotropy, PMA)磁隧道结(magnetic tunnel junction, MTJ), 由自由层、隔离层、钉扎层组成, 电流从位线流向源线时写入 1, 反向流动时写入 0. 虽然 STT-MRAM 工艺较为成熟, 也有着诸如面积小、功耗低的优势, 但其较大的写入延迟和功耗限制了它的进一步应用[78]. 因此近年来, 研究者们也开始研究具有良好写入功能的自旋轨道转矩磁随机存取存储器(spin orbit torque-magnetic RAM, SOT-MRAM)[79].



(a) PMA-MTJ结构 (b) 基本STT-MRAM单元

Fig. 9 Schematic diagram of STT-MRAM

图 9 STT-MRAM 的示意图

3.1 基于 MRAM 的存内逻辑运算

MRAM 可以通过 MAGIC 或者 IMP 等设计方法完成存内逻辑运算, 其 MAGIC 或非门结构与忆阻器的 MAGIC 或门相同. 以使用 MRAM 的 MAGIC 或非门举例, 2 个 MRAM 作为输入, 1 个 MRAM 作为保存结果的输出. 2020 年, Angizi 等人[80] 提出了基于 STT-MRAM 的内存加速器, 并使用卷积神经网络进行了实验. 该方法在单个时钟周期内, 实现了操作数之间完整的布尔逻辑运算. 由于计算的高度并行性, 在加速计算的同时还减少了能量消耗. 2021 年, Wang 等人[81] 研究了基于 MRAM 的高效存内计算. 在逻辑与和逻辑或功能的基础上, 利用多数逻辑实现了全加器运算.

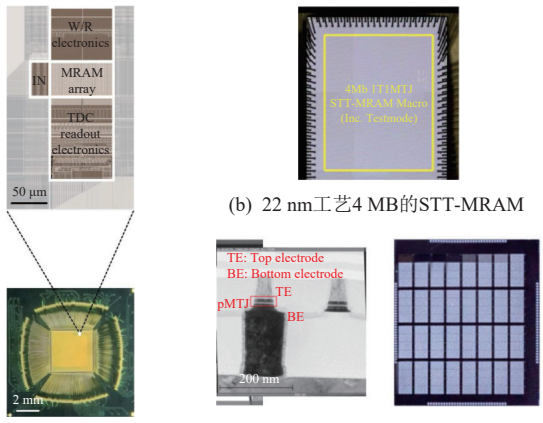
Nehra 等人[82] 研究了 STT/SOT 串联 3 级 MRAM 单元的存内计算架构. 该架构成功地实现了高性能的逻辑与、逻辑或、逻辑异或以及磁性全加器逻辑电路, 且极大地节省了写入能量并实现了数据的快速读取和面积上的减少. Kim 等人[83] 研究了基于 SOT-MRAM 的双写入方案, 有效地降低了 MRAM 存储单元的写入能量. 将相同的数据写入 2 个单元时, 该工作最多可以降低 62.6% 的能耗, 且写入能耗平均可以降低 26.3%.

3.2 MRAM 应用于神经网络

与 SRAM 类似, MRAM 可以通过逻辑同或运算完成神经网络的乘加运算, 这时神经网络一般选择 BNN.

Cai 等人[84] 研究了商用 MRAM 存内计算架构, 提出了一步卷积和无限宽度卷积操作, 有效地降低了延迟和能耗. Pham 等人[85] 研究了用于 BNN 神经网络的 STT-MRAM 的存内计算. 通过允许跨行执行的无限制操作和引入外围电路的校准, 提高了 BNN 计算的准确性. 最后使用了 Mnist 数据集进行识别测试, 精度达到了 98.42%.

Jung 等人[86] 提出了一种克服 MRAM 存内计算低电阻问题的 64×64 MRAM 阵列, 用于图像分类和面部检测. 该芯片显微照片如图 10(a)所示, 读写控制电路位于 MRAM 阵列的上方, 下方则是 TDC 读出电路. 使用该款芯片加速 2 层全连接神经网络后得到 93.23% 的数字识别精度, 在人脸检测方面实现单层网络识别精度 93.4%, 实现 8 层的 VGG-8 神经网络识别精度 98.86%. Chiu 等人[87] 提出了 22 nm 工艺 4 MB 的 STT-MRAM 芯片, 该芯片支持 8 b 输入、8 b 权重、26 b 输出. 该芯片显微镜照片如图 10(b)所示, 芯片面积为 18 mm^2, 实现了 192 Gbps 读写和解密带宽



(a) 64×64的MRAM阵列 (b) 22 nm工艺4 MB的STT-MRAM (c) 28 nm工艺的MRAM阵列

Fig. 10 Three MRAM chips in recent years

图 10 近年 3 款 MRAM 芯片

以及 25.1~55.1 TOPS/W 的 8 b 存内计算. Singh 等人[88]研究了基于 STT-MRAM 的存内逻辑加速器，并采用 28 nm 工艺流片证明实现了高鲁棒性的存内计算，该芯片如图 9(c)所示. 这几款芯片对比如表 5 所示.

Table 5 Comparison of Several MRAM Chips

表 5 几款 MRAM 芯片对比

来源	年份	阵列	工艺/nm	权重位数/b	能效/(TOPS·W^{-1})
文献 [82]	2021	64×64	28	1	262~405
文献 [83]	2021	4 MB	22	2/4/8	25.1~960.2
文献 [84]	2022	128×512	28	1	17.8

Wang 等人[89]研究了基于 TSOT-MRAM 的存内计算架构，其与基于 STT-MRAM 的存内计算架构相比降低了能耗. Kim 等人[90]研究了基于 SOT-MRAM 的数字存内计算架构，并采用 28 nm 的 CMOS 工艺实现，同时还实现了卷积神经网络加速. Lu 等人[91]研究了基于 SOT-MRAM 的贝叶斯神经网络与硬件的协同算法，开发了器件、电路、算法的整个框架，以增加了能耗为代价减少了不确定度并提升了精度. Nisar 等人[92]提出了基于 STT-MRAM/SOT-MRAM 的 4 b 的 MRAM 单元，其减少面积的同时降低了延迟和能耗.

4 3 维存算一体架构

相比较于传统的 2 维架构，3 维架构有着更高的集成度，因此 3 维架构对实现高速低功耗的存内计算芯片有着重要的意义.

2019 年，Srinivasa 等人[93]提出了一种 2 层的 3 维 SRAM 存内计算芯片架构，该架构具有较强的鲁棒性，并有效减少了数据写入的时间与功耗. Hsueh 等人[94]基于 10T 的 SRAM 单元提出了可堆叠的存内计算芯片，实现了数据吞吐量的提高. Li 等人[95]则基于鳍式场效应晶体管提出了 4 层的存内计算架构，其 9T 的 SRAM 单元能够在单个周期内完成逻辑与非、逻辑或、逻辑异或等运算. Kota 等人[96]提出了 96 MB 的电感耦合式 3 维堆叠 SRAM，最终实现了能耗和面积的有效减少，其提出的架构示意图如图 11 所示.

2017 年，Adam[97]开发了 2 层的 10×10 大小忆阻器阵列用于神经形态计算. 其芯片照片如图 12(a)所示，虽然还存在不少问题，但是这项工作为 3 维存内计算芯片打下了基础. Fernando 等人[98]提出了一种基于忆阻器的 3 维多核架构，其能够实现在线学习功能，且 3 维的架构使得芯片面积大为缩小. Veluri 等人[99]提出了交错的 3 维忆阻器阵列，如图 12(b)所示，并且实现了 2 层全连接神经网络和 4 层卷积神经网络. Huo 等人[100]提出了 2 KB 的 3 维忆阻器芯片，其架构如图 12(c)所示. 3 维的架构提升了芯片的集成度并减少了功耗. Sun 等人[101]提出了基于 4 层垂直忆阻器阵列的储备池计算，如图 12(d)所示，通过软硬件协同设计实现了高精度、低功耗、高面积效率. Li 等人[102]提出了 8 层单片集成忆阻器阵列，用于实现神经网络加速并使用卷积神经网络进行验证.

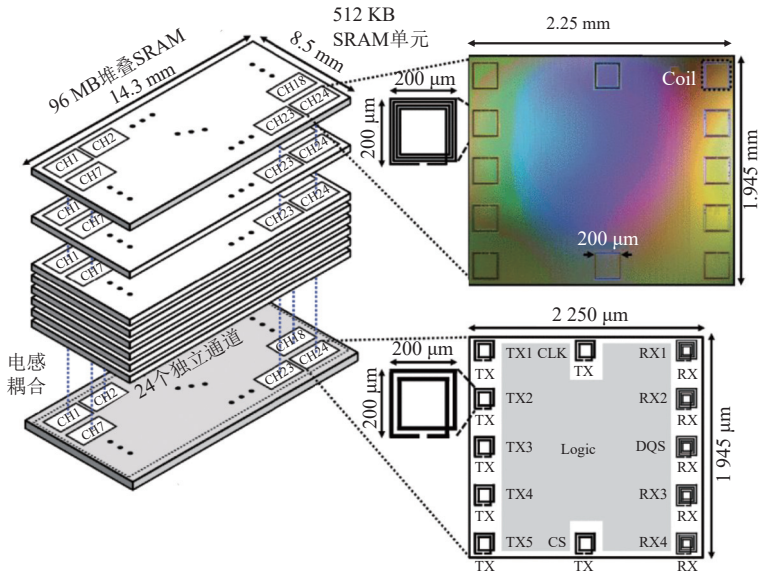
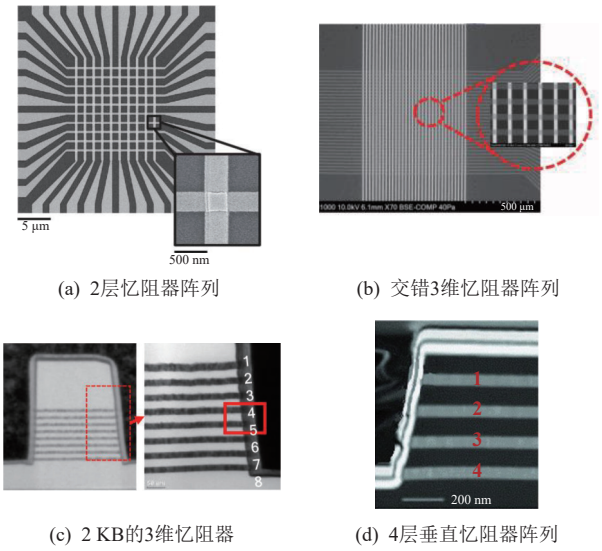


Fig. 11 Eight-layer stack SRAM

图 11 8 层堆叠 SRAM



(a) 2层忆阻器阵列 (b) 交错3维忆阻器阵列

(c) 2 KB的3维忆阻器 (d) 4层垂直忆阻器阵列

Fig. 12 3D memristor memory chip

图 12 3 维忆阻器存算芯片

5 大规模存内计算芯片

随着处理数据量的增加, 存内计算芯片的容量也受到了研究者的关注, 高精度的存内计算需要大内存容量以适应多位输入和多位权重. Xue 等人[103]提出了 4 MB 的忆阻器存内计算架构, 该架构由 8 个 1 024×512 的忆阻器阵列组成, 实现了纳秒级别的延迟和较高的能效. Hung 等人[104]提出了 8 MB 的忆阻器存内计算芯片, 该芯片采用 22 nm 工艺, 支持 8 b 输入、8 b 权重、19 b 输出. 该芯片使用了 32 个 256 KB 大小的子阵列来执行并行存内计算, 其计算延迟为 14.4 ns, 能效为 21.6TOPS/W. Liu 等人[105]提出了一款 32 MB 的 5T2R1C 结构的芯片, 实现了 CNN 的硬件加速. 这项工作在处理数据 0 时关闭 ADC 来减少功耗, 有效地降低了能耗. 该工作支持最高 8 b 精度的计算, 且可以实现 2 490.32 TOPS/W 的峰值能效和 479.37 TOPS/W 的平均能效.

制约大规模存内计算芯片的一个重要挑战是面积. 忆阻器和 MRAM 虽然面积足够小但是工艺不及 SRAM, 需要外围电路的支持, 导致其目前并没有什么面积优势. Gauchi 等人[106]研究了可扩展的存内计算 SARM 芯片之间的互连问题. 研究显示, 芯片的延迟与其阵列间的互联有关. 较小的存内计算芯片应当使用尽可能少的子阵列来减少延迟, 而较大容量的存内计算芯片应当使用适当数量的子阵列来减少延迟. 如果考虑能耗和性能之间的权衡, 以 32 KB 为例, 使用 16 个 2 KB 大小阵列互联的方案性能最佳.

除此之外, 芯片的速度与功耗以及器件的可靠性和非理想性也是对大规模存内计算的一个限制. 较高的电压和温度都可能影响到器件的稳定性, 从而对计算造成影响. 而速度和功耗则是存算一体芯片保持对传统芯片优势的关键点. 同时, 大规模存算宏单元间的同步也是待研究的方向.

6 主要挑战

1)速度和功耗问题. 在存内计算架构中, 负责将阵列的电流和转为数字信号的模数转换器(ADC)消耗了大量的能量, 是一个主要的瓶颈. 文献 [107] 提出了综合考虑能耗与数据吞吐量的平衡, 在不同的网络之间选择闪存 ADC 或者逐次逼近型 ADC[107]. 从器件方面来说, SRAM 虽然有着较快的读写速度, 但是其静态功耗会影响计算的稳定性. 忆阻器和 MRAM 虽然不存在静态功耗问题, 但是由于工艺的问题, 读写性能目前无法与 SRAM 比拟. 在未来, 随着工艺的提升, 忆阻器和 MRAM 有望优化速度与功耗问题.

2)软件工具较少. 存内计算芯片还存在着对软件工具链方面研究较少的问题, 而软件工具的开发与应用可以有效帮助研究者开发高效的存内计算芯片. 例如, 量化工具对输入、权重、网络部分和的量化虽然会导致精度的略微降低, 但是会对硬件性能有一个极大的提升. 而将神经网络部署到硬件上的高效编译映射工具也是不可或缺的一环[108], 在未来存内计算芯片编译器的适配问题也将引起研究人员的重视. 针对器件非理想的表现, 存内计算仿真器的开发势在必行. 在仿真器中可以通过考虑器件的各种非理想表现对权重进行调整以提高硬件识别的精度并评估硬件的性能[109].

3)计算精度有限. 高精度的存内计算需要使用多个存储单元来存储多位权重, 这无疑会增大延时, 计算过程会受到模拟计算低信噪比的影响导致精度降低, 且无法实现高精度的浮点数计算. 目前一般认为 8 b 权重精度的方案比较可靠, 多值忆阻器可用于缓解这个问题. 例如, 一个 8 值忆阻器单元可用来存储一个 3 b 的二进制数.

4)存储单元可靠性问题. 以忆阻器来说, 它的阻值在不同的读写周期可能发生变化从而影响存内计算的精度. 而对于 MRAM 来说, 较高的电压和较高的温度都会影响它的可靠性. 目前 SRAM 的可靠性要高于 MRAM 和忆阻器[110].

7 总结与展望

本文探讨了基于 3 种不同器件的存内计算原理，并介绍了现有的存内计算架构. 由于 SRAM 拥有可靠的稳定性和成熟的工艺等优势，因此在存内计算方向有着不可忽视的地位. 但是 SRAM 存在着集成度低、功耗大、易失的缺点，因此未来 SRAM 存算一体架构在高集成、低功耗方向将是重点研究方向. 目前，国内外对于忆阻器存内计算的研究集中在较小规模阵列上的逻辑运算、矩阵乘加运算以及算法的优化，忆阻器存内计算架构因为其能效高、速度快常用于神经网络加速. 在未来，针对忆阻器阵列的可靠性、忆阻器误差校准以及大规模忆阻器阵列的应用还需要多加研究. MRAM 在存内计算架构上取得了一定的成果，当下 MRAM 的存内计算集中于实现 BNN. 未来，针对 MRAM 的写入功耗过大和读取可靠性还需要继续深入研究.

基于 3 种器件设计的存内计算架构有着不同的特点. MRAM 和 SRAM 使用逻辑异或来完成乘法，故这 2 种器件一般常用于硬件加速 BNN 神经网络和支持 1 b 权重精度. 而忆阻器可使用逻辑运算也可使用模拟计算，常用于 CNN 和 DNN 等网络，一般支持 1~8 b 的权重精度. 而从能耗和面积角度来说，SRAM 架构静态功耗大，面积也较大，忆阻器和 MRAM 更有优势.

总的来说，目前存内计算芯片普遍存在工艺浮动、串扰干扰、噪声干扰等问题. 同时，忆阻器与 MRAM 这类新型器件与工艺、电路、架构和算法的协同设计与优化也值得研究者的探索. 而面对存内计算的精度问题，混合异构架构有望实现大于 8 b 的高精度计算.

作者贡献声明：张章提出论文架构、给出调研方向、参与并指导论文撰写；施刚参与论文撰写和实际调查；王启帆负责文献收集和整理；马永波负责论文图表制作；刘钢和钱利波负责论文审阅.

参 考 文 献

[1] Patterson D, Anderson T, Cardwell N, et al. A case for intelligent ram[J]. IEEE Micro, 1997, 17(2): 34−44

[2] Wolf M. The Physics of Computing [M]. Amsterdam: Elsevier, 2016: 1−265

[3] Tu Fengbin, Wang Yiqi, Wu Zihan, et al. ReDCIM: Reconfigurable digital computing-in-memory professor with unified FP/INT pipeline for cloud AI acceleration[J]. IEEE Journal of Solid-State Circuits, 2023, 58(1): 243−255

[4] Ali M F, Jaiswal A, Roy K. In-memory low-cost bit-serial addition using commodity DRAM technology[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2020, 67(1): 155−165

[5] Sudarshan C, Said H M, Weis C, et al. Optimization of DRAM based PIM architecture for energy-efficient deep neural network training[C]//Proc of the 35th IEEE Int Symp on Circuits and Systems (ISCAS). Piscataway, NJ: IEEE, 2022: 1472−1476

[6] Ali M, Roy S, Saxena U, et al. Compute-in-memory technologies and architectures for deep learning work-loads[J]. IEEE Transactions on Very Large Scale Integration Systems, 2022, 30(11): 1615−1630

[7] Verma N, Jia Hongyang, Valavi H, et al. In-memory computing: Advances and prospects[J]. IEEE Solid-State Circuits Magazine, 2019, 11(3): 43−55

[8] Chua L. Memristor-the missing circuit element[J]. IEEE Transactions on Circuit Theory, 1971, 18(5): 507−519

[9] Duan Shukai, Hu Xiaofang, Dong Zhekang, et al. Memristor-based cellular nonlinear/neural network: Design, analysis, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(6): 1202−1213

[10] Yang Xiaoxuan, Taylor B, Wu Ailong, et al. Research progress on memristor: From synapses to computing systems[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2022, 69(5): 1845−1857

[11] Seo Y, Roy K. High-density SOT-MRAM based on shared bitline structure[J]. IEEE Transactions on Very Large Scale Integration Systems, 2018, 26(8): 1600−1603

[12] Nejat A, Ouattara F, Mohammadinodoushan M, et al. Practical experiments to evaluate quality metrics of MRAM-based physical unclonable functions[J]. IEEE Access, 2020, 8: 176042−176049

[13] Chang T C, Chang K C, Tsai T M, et al. Resistance random access memory[J]. Materials Today, 2016, 19(5): 254−264

[14] Jeloka S, Akesh N B, Sylvester D, et al. A 28 nm configurable memory (TCAM/BCAM/ SRAM) using push-rule 6T bit cell enabling logic-in-memory[J]. IEEE Journal of Solid-State Circuits, 2016, 51(4): 1009−1021

[15] Yueh W, Chatterjee S, Zia M, et al. A memory-based logic block with optimized-for-read SRAM for energy-efficient reconfigurable computing fabric[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2015, 62(6): 593−597

[16] Kang K, Jeong H, Yang Y, et al. Full-swing local bitline SRAM architecture based on the 22 nm FinFET technology for low-voltage operation[J]. IEEE Transactions on Very Large Scale Integration Systems, 2016, 24(4): 1342−1350

[17] Jaiswal A, Agrawal A, Ali M F, et al. I-SRAM: Interleaved wordlines for vector Boolean operations using SRAMs[J]. IEEE Transactions on Circuits and Systems I, 2020, 67(12): 4651−4659

[18] Aga S, Jeloka S, Subramaniyan A, et al. Compute caches[C]//Proc of the 23rd IEEE Int Symp on High Performance Computer Architecture (HPCA). Los Alamitos, CA: IEEE Computer Society,

2017: 481−492

[19] Dong Qing, Jeloka S, Saligane M, et al. A 4+2T SRAM for searching and in-memory computing with 0.3V VDD_{min}[J]. IEEE Journal of Solid-State Circuits, 2017, 53(4): 1006−1015

[20] Lin Zhiting, Li Luanyun, Wu Xiulong, et al. Half-select disturb-free 10T tunnel FET SRAM cell with improved noise margin and low power consumption[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2021, 68(7): 2628−2632

[21] Agrawal A, Jaiswal A, Lee C, et al. X-SRAM: Enabling in-memory Boolean computations in CMOS static random access memories[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2018, 65(12): 4219−4232

[22] Rajput A K, Pattanaik M. Implementation of Boolean and arithmetic functions with 8T SRAM cell for in-memory computation[C/OL]// Proc of the 1st Int Conf for Emerging Technology (INCET). Piscataway, NJ: IEEE, 2020[2023-01-02]. https://ieeexplore.ieee.org/document/9154137

[23] Chen Jian, Zhao Wenfeng, Wang Yuqi, et al. A reliable 8T SRAM for high-speed searching and logic-in-memory operations[J]. IEEE Transactions on Very Large Scale Integration Systems, 2022, 30(6): 769−780

[24] Chaturvedi M, Garg M, Rawat B, et al. A read stability enhanced, temperature tolerant 8T SRAM cell[C/OL]//Proc of the 1st Int Conf on Simulation, Automation & Smart Manu-facturing (SASM). Piscataway, NJ: IEEE, 2021[2023 -01-02]. https://ieeexplore.ieee.org/document/9841199

[25] Wen Liang, Cheng Xu, Zhou Keji, et al. Bit-interleaving-enabled 8T SRAM with shared data-aware write and reference-based sense amplifier[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2016, 63(7): 643−647

[26] Yu C, Yoo T, Chai K T C, et al. A 65 nm 8T SRAM compute-in-memory macro with column ADCs for Professing neural networks[J]. IEEE Journal of Solid-State Circuits, 2022, 57(11): 3466−3476

[27] Agrawal A, Kosta A, Kodge S, et al. CASH-RAM: Enabling in-memory computations for edge inference using charge accumulation and sharing in standard 8T-SRAM arrays[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2020, 10(3): 295−305

[28] Rastegari M, Ordonez V, Redmon J, et al. XNOR-net: ImageNet classification using binary convolutional neural networks[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 525−542

[29] Agrawal A, Jaiswal A, Roy D, et al. Xcel-RAM: Accelerating binary neural networks in high-throughput SRAM compute arrays[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2019, 66(8): 3064−3076

[30] Raman S R S, Nibhanupudi S S T, Kulkarni J P. Enabling in-memory computations in non-volatile SRAM designs[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2022, 12(2): 557−568

[31] Lee E, Han T, Seo D, et al. A charge-domain scalable-weight in-memory computing macro with dual-SRAM architecture for precision-scalable DNN accelerators[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(8): 3305−3316

[32] Jiang Hongwu, Liu Rui, Yu Shimeng. 8T XNOR-SRAM based parallel compute-in-memory for deep neural network accelerator[C]//Proc of the 63rd Int Midwest Symp on Circuits and Systems (MWSCAS). Piscataway, NJ: IEEE, 2020: 257−260

[33] Si Xin, Khwa W S, Chen J J, et al. A dual-split 6T SRAM-based computing-in-memory unit-macro with fully parallel product-sum operation for binarized DNN edge professors[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2019, 66(11): 4172−4185

[34] Nguyen V T, Kim J S, Lee J W. 10T SRAM computing-in-memory macros for binary and multibit MAC operation of DNN edge Professors[J]. IEEE Access, 2021, 9: 71262−71276

[35] Su J W, Si Xin, Chou Y C, et al. Two-way transpose multibit 6T SRAM computing-in-memory macro for inference- training AI Edge chips[J]. IEEE Journal of Solid-State Circuits, 2022, 57(2): 609−624

[36] Han J, Heo J, Kim J, et al. Design of professing-in-memory with triple computational path and sparsity handling for energy-efficient DNN training[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2022, 12(2): 354−366

[37] Nasrin S, Badawi D, Cetin A E, et al. MF-net: Compute-in-memory SRAM for multibit precision inference using memory-immersed data conversion and multiplication-free operators[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(5): 1966−1978

[38] Iqbal B, Grover A, Rawat H. A profess and data variations tolerant capacitive coupled 10T1C SRAM for in-memory compute (IMC) in deep neural network accelerators[C]//Proc of the 4th Int Conf on Artificial Intelligence Circuits and Systems (AICAS). Piscataway, NJ: IEEE, 2022: 459−462

[39] Yan Bonan, Hsu J L, Yu P C, 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 recon-figurable bitwise operation for AI and embedded applications[C]//Proc of the 68th IEEE Int Solid-State Circuits Conf (ISSCC). Piscataway, NJ: IEEE, 2022: 188−190

[40] Choi E, Choi I, Jeon C, et al. A 133.6TOPS/W compute-in-memory SRAM macro with fully parallel one-step multi-bit computation[C/OL]//Proc of the 43rd IEEE Custom Integrated Circuits Conf (CICC). Piscataway, NJ: IEEE, 2022[2023-01-02]. https://ieeexplore.ieee.org/document/ 9772821

[41] Borghetti j, Snider G S, Kuekes P J, et al. ‘Memristive’ switches enable ‘stateful’ logic operations via material implication[J]. Nature, 2010, 464(7290): 873−876

[42] Kvatinsky S, Belousov D, Liman S, et al. Magic—memrist-or-aided logic[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2014, 61(11): 895−899

[43] Jang B C, Nam Y, Koo B J, et al. Memristive logic-in-memory integrated circuits for energy-efficient flexible electronics[J]. Advanced Functional Materials, 2018, 28(2): 1704725

[44] Gupta S, Imani M, Rosing T. FELIX: Fast and energy-efficient logic in memory[C/OL]//Proc of the 31st IEEE/ACM Int Conf on Computer-Aided Design (ICCAD). Piscataway, NJ: IEEE,

2018[2023-01-02].https://ieeexplore. ieee.org/document/8587724

[45] Xu Xiaoyan, Cui Xiaole, Luo Mengying, et al. Design of hybrid memristor-MOS XOR and XNOR logic gates[C/OL]//Proc of the 13th Int Conf on Electron Devices and Solid-State Circuits (EDSSC). Piscataway, NJ: IEEE, 2017[2023-01-02]. https://ieeexplore.ieee.org/document/8126414

[46] Liu Gongzhi, Zheng Lijing, Wang Guangyi, et al. A carry lookahead adder based on hybrid CMOS-memristor logic circuit[J]. IEEE Access, 2019, 7: 43691−43696

[47] Teimoory M, Amirsoleimani A, Ahmadi A, et al. A hybrid memristor-CMOS multiplier design based on memristive universal logic gates[C]//Proc of the 60th Int Mid-west Symp on Circuits and Systems (MWSCAS). Piscataway, NJ: IEEE, 2017: 1422−1425

[48] Rohani S G, Nejad N T, Radakovits D. A semiparallel full-adder in imply logic[J] IEEE Transactions on Very Large Scale Integration Systems, 2020, 28(1): 297−301

[49] Ali A, Ben-Hur R, Wald N, et al. Efficient algorithms for in-memory fixed point multiplication using MAGIC [C/OL]//Proc of the 31st IEEE Int Symp on Circuits and Systems (ISCAS). Piscataway, NJ: IEEE, 2018[2023-01-02]. https://ieeexplore.ieee.org/document/8351561

[50] Li Zhiwei, Zhu Xi, Li Nan, et al. SCMOS: Series-connected memristor-only stateful logic[C/OL]//Proc of the 15th IEEE Int Conf on Solid-State & Integrated Circuit Technology (ICSICT). Piscataway, NJ: IEEE, 2020[2023-01-02]. https://ieeexplore.ieee.org/document/9278249

[51] Song Yujie, Wang Xingsheng, Wu Qiwen, et al. Reconfigurable and efficient implementation of 16 Boolean logics and full-adder functions with memristor crossbar for beyond von Neumann in-memory computing[J]. Advanced Science, 2022, 9(15): 2200036

[52] Pandey N, Verma S, Jeph S, et al. Design of a digital magnitude comparator based on memristor logic circuit[C]//Proc of the 1st Int Mobile and Embedded Technology Conf (MECON). Piscataway, NJ: IEEE, 2022: 430−434

[53] Paramasivam K, Nithya N, Nepolean A. A novel hybrid CMOS-memristor based 2-bit magnitude comparator using memristor ratioed logic universal gate for low power applications[C/OL]//Proc of the 1st Int Conf on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA). Piscataway, NJ: IEEE, 2021[2023-01-02]. https://ieeexplore.ieee.org/docu ment/9675534

[54] Biswas B R, Gupta S. Memristor-specific failures: New verification methods and emerging test problems[C/OL]//Proc of the 40th VLSI Test Symp (VTS). Los Alamitos, CA: IEEE Computer Society, 2022[2023-01-02]. https://ieeexplore.ieee.org/document/9794274

[55] Yang J, Lee H, Jeong J H, et al. Circuit-level exploration of ternary logic using memristors and MOSFETs[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2022, 69(2): 707−720

[56] Wu Zhixin, Zhang Yuejun, Du Shimin, et al. A three-valued adder circuit implemented in ZnO memristor with multi-resistance states[C/OL]//Proc of the 14th Int Conf on ASIC (ASICON). Piscataway, NJ: IEEE, 2021[2023-01-02]. https://ieeexplore.ieee.org/document/9794274

[57] Shanmukh S, RohitKumar S, Hemaprasad P, et al. Low power 3-bit encoder design using memristor[C/OL]//Proc of the 2nd Int Conf on Intelligent Technologies (CONIT). Piscataway, NJ: IEEE, 2022[2023-01-02]. https://ieeexplor e.ieee.org/document/9848019

[58] Soudry D, Castro D Di, Gal A, et al. Memristor-based multilayer neural networks with online gradient descent training[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2408−2421

[59] Chabi D, Wang Zhaohao, Bennett C, et al. Ultrahigh density memristor neural crossbar for on-chip supervised learning[J]. IEEE Transactions on Nanotechnology, 2015, 14(6): 954−962

[60] Zheng Nan, Mazumder P. Learning in memristor crossbar-based spiking neural networks through modulation of weight-dependent spike-timing-dependent plasticity[J]. IEEE Transactions on Nanotechnology, 2018, 17(3): 520−532

[61] Yakopcic C, Fernando B R, Taha T M. Design space evaluation of a memristor crossbar based multilayer perceptron for image Professing[C/OL]//Proc of the 31st Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2019[2023-01-02]. https://ieeexplore.ieee.org/document/8852005

[62] Zhong Yanan, Tang Jianshi, Li Xinyi, et al. Dynamic memristor-based reservoir computing for high-efficiency temporal signal Professing[J]. Nature Communications, 2021, 12(1): 1−9

[63] Hong Qinghui, He Bang, Zhang Zedi, et al. In-memory computing circuit implementation of complex-valued hopfield neural network for efficient portrait restoration [J/OL]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023[2023-03-22]. https://ieeexplore.ieee.org/document/10040238

[64] Hong Qinghui, He Bang, Zhang Zedi, et al. Circuit design and application of discrete cosine transform based on memristor[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2023, 13(2): 502−513

[65] Hung J M, Xue C X, Kao H Y, et al. A four-megabit compute-in-memory macro with eight-bit precision based on CMOS and resistive random-access memory for AI edge devices[J]. Nat Electron, 2021, 4: 921−930

[66] Xue C X, Chen W H, Liu J S, et al. 24.1 A 1 MB multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge professors [C]//Proc of the 65th IEEE Int Solid-State Circuits Conf (ISSCC). Piscataway, NJ: IEEE, 2019: 388−390

[67] Xie Jing, Afshari S, Esqueda I S. Hexagonal boron nitride memristor arrays for analog-based machine learning hardware[J]. NPJ 2D Materials and Applications, 2022, 6((1):): 1−7

[68] Yao Peng, Wu Huaqiang, Gao Bin, et al. Fully hardware implemented memristor convolutional neural network[J]. Nature, 2020, 577(7792): 641−646

[69] Wan W, Kubendran R, Schaefer C, et al. A compute-in-memory chip based on resistive random-access memory[J]. Nature, 2022, 608((7923):): 504−512

[70] Gao Di, Huang Qingrong, Zhang G L, et al. Bayesian inference based

robust computing on memristor crossbar[C]//Proc of the 58th ACM/IEEE Design Automation Conf (DAC). Piscataway, NJ: IEEE, 2021: 121−126

[71] Mohan C, Camuñas-Mesa L A, José M. Neuromorphic low-power inference on memristive crossbars with on-chip offset calibration[J]. IEEE Access, 2021, 9: 38043−38061

[72] Qin Yifan, Kuang Rui, Huang Xiaodi, et al. Design of high robustness BNN inference accelerator based on binary memristors[J]. IEEE Transactions on Electron Devices, 2020, 67(8): 3435−3441

[73] Fu Jingyan, Liao Zhiheng, Wang Jinhui. Level scaling and pulse regulating to mitigate the impact of the cycle-to-cycle variation in memristor-based edge AI system[J]. IEEE Transactions on Electron Devices, 2022, 69(4): 1752−1762

[74] Zhang Zhang, Xu Ao, Li Chao, et al. Gate-controlled memristor FPGA model for quantified neural network[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 69(11): 4583−4587

[75] Zhang Zhang, Chen Qilai, Han Tingting, et al. Memristor-based circuit demonstration of gated recurrent unit for predictable neural network[J]. IEEE Transactions on Electron Devices, 2022, 69(12): 6763−6768

[76] Wang Mengxing, Cai Wenlong, Cao Kaihua, et al. Current induced magnetization switching in atom-thick tungsten engineered perpendicular magnetic tunnel junctions with large tunnel magnetoresistance[J]. Nature Communications, 2018, 9(1): 1−7

[77] Cao Kaihua, Cai Wenlong, Liu Yizheng, et al. In-memory direct professing based on nanoscale perpendicular magnetic tunnel junctions[J]. Nanoscale, 2018, 10(45): 21225−21230

[78] He Yanxiang, Shen Fanfan, Zhang Jun, et al. Cache optimization approaches of emerging non-volatile memory architecture: A survey[J]. Journal of Computer Research and Development, 2015, 52(6): 1225−1241 (in Chinese)
(何炎祥, 沈凡凡, 张军, 等. 新型非易失性存储器架构的缓存优化方法综述[J]. 计算机研究与发展, 2015, 52(6): 1225−1241)

[79] Tsou Y J, Chen W J, Shih H C, et al. Thermally robust perpendicular SOT-MTJ memory cells with STT-assisted field-free switching[J]. IEEE Transactions on Electron Devices, 2021, 68(12): 6623−6628

[80] Angizi S, He Z, Awad A, et al. MRIMA: An MRAM-based in-memory accelerator[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 39(5): 1123−1136

[81] Wang Chao, Wang Zhaohao, Wang Gefei, et al. Design of an area-efficient computing in memory platform based on STT-MRAM[J]. IEEE Transactions on Magnetics, 2021, 57(2): 1−4

[82] Nehra V, Prajapati S, Kumar T N, et al. High-performance computing-in-memory architecture using STT-/SOT-based series triple-level cell MRAM[J]. IEEE Transactions on Magnetics, 2021, 57(8): 1−12

[83] Kim J, Bae K, Park J. Low power SOT-MRAM cell configuration for dual write operation[C/OL]//Proc of the 6th Int Conf on Electronics, Information, and Communication (ICEIC). Piscataway, NJ: IEEE, 2021[2023-01-02]. https://ieeexplore.ieee.org/document/9369790

[84] Cai Hao, Bian Zhongjian, Fan Zhonghua, et al. Commodity bit-cell sponsored MRAM interaction design for binary neural network[J]. IEEE Transactions on Electron Devices, 2022, 69(4): 1721−1726

[85] Pham T N, Trinh Q K, Chang I J, et al. STT-BNN: A novel STT-MRAM in-memory computing macro for binary neural networks[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2022, 12(2): 569−579

[86] Jung S, Lee H, Myung S, et al. A crossbar array of magne-toresistive memory devices for in-memory computing[J]. Nature, 2022, 601(7892): 211−216

[87] Chiu Y C, Yang C S, Teng S H, et al. A 22 nm 4 Mb STT-MRAM data-encrypted near-memory computation macro with a 192 GB/s read-and-decryption bandwidth and 25.1-55.1TOPS/W 8b MAC for AI operations[C]//Proc of the 68th IEEE Int Solid-State Circuits Conf (ISSCC). Piscataway, NJ: IEEE, 2022: 178−18

[88] Singh A, Zahedi M, Shahroodi T, et al. CIM-based robust logic accelerator using 28 nm STT-MRAM characterization chip tapeout[C]//Proc of the 4th Int Conf on Artificial Intelligence Circuits and Systems (AICAS). Piscataway, NJ: IEEE, 2022: 451−454

[89] Wang Jinkai, Bai Yining, Wang Hongyu, et al. Reconfigurable bit-serial operation using toggle SOT-MRAM for high-performance computing in memory architecture[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2022, 69(11): 4535−4545

[90] Kim T, Jang Y, Kang M G, et al. SOT-MRAM digital PIM architecture with extended parallelism in matrix multiplication[J]. IEEE Transactions on Computers, 2022, 71(11): 2816−2828

[91] Lu Anni, Luo Yandong, Yu Shimeng. An algorithm-hardware co-design for Bayesian neural network utilizing SOT-MRAM's inherent stochasticity[J]. IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 2022, 8(1): 27−34

[92] Nisar A, Dhull S, Mittal S, et al. SOT and STT-based 4-bit MRAM cell for high-density memory applications[J]. IEEE Transactions on Electron Devices, 2021, 68(9): 4384−4390

[93] Srinivasa S, Ramanathan A K, Li Xueqing, et al. ROBIN: Monolithic-3D SRAM for enhanced robustness with in-memory computation support[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2019, 66(7): 2533−2545

[94] Hsueh F K, Lee C Y, Xue C X, et al. Monolithic 3D SRAM-CIM macro fabricated with BEOL gate-all-around MOSFETs[C/OL]// Proc of the 65th IEEE Int Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2019 [2023-01-02]. https://ieeexplore.ieee.org/document/8993628

[95] Li K S, Hsueh F K, Shen C H, et al. FinFET-based monolithic 3D+ with RRAM array and computing in memory SRAM for intelligent IoT chip application[C]//Proc of the 28th IEEE SOI-3D-Subthreshold Microelectro-nics Technology Unified Conf (S3S). Piscataway, NJ: IEEE, 2018[2023-01-02]. https://ieeexplore.ieee.org/document/8640186

[96] Kota S, Tatsuo O, Kodai U, et al. A 96-MB 3D-stacked SRAM using inductive coupling with 0.4-V Transmitter, termination scheme and 12: 1 serdes in 40 nm CMOS[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(2): 692−703

[97] Adam G C, Hoskins B D, Prezioso M, et al. 3D memristor crossbars for analog and neuromorphic computing applications[J]. IEEE Transactions on Electron Devices, 2017, 64(1): 312−318

[98] Fernando B R, Qi Yangjie, Yakopcic C, et al. 3D memristor crossbar architecture for a multicore neuromorphic system [C/OL]//Proc of the 32nd Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2020[2023-01-02]. https://ieeexplore.ieee.org/document/9206929

[99] Veluri H, Li Yida, Niu J X, et al. High-throughput, area-efficient, and variation-tolerant 3D in-memory compute system for deep convolutional neural networks[J]. IEEE Internet of Things Journal, 2021, 8(11): 9219−9232

[100] Huo Qiang, Yang Yiming, Wang Yiming, et al. A computing-in-memory macro based on three-dimensional resistive random-access memory[J]. Nature Electronics, 2022, 5(7): 469−477

[101] Sun Wenxuan, Zhang Woyu, Yu Jie, et al. 3D reservoir computing with high area efficiency (5.12TOPS/mm^2) implemented by 3D dynamic memristor array for temporal signal professing[C]//Proc of the 36th IEEE Symp on VLSI Technology and Circuits (VLSI Technology and Circuits). Piscataway, NJ: IEEE, 2022: 222−223

[102] Lin Peng, Li Can, Wang Zhongrui, et al. Three-dimensional memristor circuits as complex neural networks[J]. Nature Electronics, 2020, 3(4): 225−232

[103] Xue C X, Hung J M, Kao H Y, et al. A 22 nm 4 Mb 8 b-precision ReRAM computing-in-memory macro with 11.91 to 195.7 TOPS/W for tiny AI edge devices[C]//Proc of the 67th IEEE Int Solid-State Circuits Conf (ISSCC). Piscataway, NJ: IEEE, 2021: 245−247

[104] Hung J M, Wen T H, Huang Y H, et al. 8 b precision 8 Mb ReRAM compute-in-memory macro using direct-current-free time-domain readout scheme for AI edge devices[J]. IEEE Journal of Solid-State Circuits, 2023, 58(1): 303−315

[105] Liu Dingbang , Zhou Haoxiang, Mao Wei, et al. An energy-efficient mixed-bit CNN accelerator with column parallel readout for ReRAM-based in-memory computing[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2022, 12(4): 821−834

[106] Gauchi R, Kooli M, Vivet P, et al. Memory sizing of a scalable SRAM in-memory computing tile based architecture[C]//Proc of the 27th IFIP/IEEE Int Conf on Very Large Scale Integration (VLSI-SoC). Piscataway, NJ: IEEE, 2019: 166−171

[107] Yu Shimeng, Sun Xiaoyu, Peng Xiaochen, et al. Compute in-memory with emerging nonvolatile-memories: Challenges and prospects [C/OL]//Proc of the 13th IEEE Custom Integrated Circuits Conf (CICC). Piscataway, NJ: IEEE, 2020[2023-01-02]. https://ieeexplore.ieee.org/document/9075887

[108] Lin Zhiting, Zhang Jian, Wu Xiulong, et al. Memory compiler for RRAM in-memory computation[C]//Proc of the 7th Int Conf on Integrated Circuits and Microsystems (ICICM). Piscataway, NJ: IEEE, 2022: 382−385

[109] Staudigl F, Merchant F, Leupers R. A survey of neuromorphic computing-in-memory: Architectures, simulators, and security[J]. IEEE Design & Test, 2022, 39(2): 90−99

[110] Bocquet M, Hirtzlin T, Klein J O, et al. Embracing the unreliability of memory devices for neuromorphic computing[C/OL]//Proc of the 58th IEEE Int Reliability Physics Symp (IRPS). Piscataway, NJ: IEEE, 2020[2023-01-02]. https://ieeexplore.ieee.org/document/9128346



Zhang Zhang, born in 1982. PhD, professor. Senior member of CCF. His main research interests include AI neural morphology chip, sensory memory computing integration and deep learning.

张　章, 1982 年生. 博士, 教授. CCF 高级会员. 主要研究方向为 AI 神经形态芯片、感存算一体、深度学习.



Shi Gang, born in 1999. Master candidate. His main research interests include deep learning and integration of storage and calculation.

施　刚, 1999 年生. 硕士研究生. 主要研究方向为深度学习、存算一体化.



Wang Qifan, born in 1999. Master candidate. His main research interests include deep learning and integration of storage and calculation.

王启帆, 1999 年生. 硕士研究生. 主要研究方向为深度学习、存算一体化.



Ma Yongbo, born in 1999. Master candidate. His main research interests include deep learning, and integration of storage and calculation.

马永波, 1999 年生. 硕士研究生. 主要研究方向为深度学习、存算一体化.



Liu Gang, born in 1982. PhD, professor. The national science fund for excellent young scholars. His main research interest includes organic optoelec-tronic information functional materials and devices.

刘　钢, 1982 年生. 博士, 教授. 国家优秀青年基金获得者. 主要研究方向为有机光电信息功能材料和器件.



Qian Libo, born in 1984. PhD, professor. The national science fund for excellent young scholars, distinguished professor of Huashan scholars. His main research interests include high efficiency power management IC, and high efficiency energy capture interface IC and micro energy system.

钱利波, 1984 年生. 博士, 教授. 国家优秀青年基金获得者, 华山学者特聘教授. 主要研究方向为高效功率管理集成电路、高效俘能接口集成电路与微能源系统.