文章编号: 1007-4252(2025)05-0405-06 DOI: 10.20027/j.gncq.2025.0047 CSTR: 32180.14.j.gncq.2025.0047

面向高精度推理的 PLRAM 存算一体化芯片设计与实现

朱赛克[1,2], 金健孜[3], 鲁辞莽[4,*]

(1. 浙江大学信息与电子工程学院, 杭州 310027; 2. 西南大学人工智能学院, 重庆 400715; 3. 华东师范大学集成电路科学与工程学院, 上海 200241; 4. 上海闪易半导体有限公司, 上海 200120)

摘要: 新型存算一体架构可在存储阵列内直接执行计算操作, 有望突破冯·诺伊曼架构的能效瓶颈。文章聚焦于提升可编程线性随机存取存储器(programmable linear random-access memory, PLRAM)在神经网络推理中的计算精度, 通过优化器件结构, 显著拓宽晶体管 I_d-V_d 输出特性的线性范围。实验数据显示, 与传统闪存芯片相比, 改进型闪存芯片在相近的能耗下实现了更集中的推理误差分布, 且各层芯片的推理精度均优于前代芯片, 最终推理准确率达到 94.6%, 较传统闪存芯片提升了约 4.5%。结果表明, 扩展晶体管的线性工作区能够有效抑制前向推理过程中的非线性误差累积, 从而显著提高模拟计算的可靠性与芯片推理精度, 为后续高精度、低功耗神经网络加速器的设计提供新的器件优化方向。

关键词: 存算一体; 模拟存储器; 可编程线性随机存取存储器; 线性工作区; 推理精度

中图分类号: TN40 **文献标志码:** A

Enhancing inference accuracy through linear-region optimization in the PLRAM computing-in-memory chip

ZHU Saike[1,2], JIN Jianzi[3], LU Cimang[4,*]

(1. College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, China; 2. College of Artificial Intelligence, Southwest University, Chongqing 400715, China; 3. School of Integrated Circuits Science and Engineering, East China Normal University, Shanghai 200241, China; 4. Shanghai Flashbillion Semiconductor Co., Ltd., Shanghai 200120, China)

Abstract: Novel computing-in-memory architectures, which execute computations directly within memory arrays, present a promising solution to overcoming the von Neumann bottleneck. This study focuses on enhancing the computational precision of programmable linear random-access memory (PLRAM) for neural network inference. By refining the device structure, the linear region of the transistor I_d-V_d output characteristics has been substantially extended. Experimental evaluations from neural network

收稿日期: 2025−09−21; **修订日期:** 2025−10−17

基金项目: 新一代人工智能国家科技重大专项 (No. 2020AAA0109001).

作者简介: 朱赛克(1994—), 男, 博士, 讲师, 主要研究方向为基于新型存储的存算一体化芯片(E-mail: saikezhu@gmail.com).

通信作者: 鲁辞莽 (1986—), 男, 博士, 主要研究方向为先进集成电路技术研究(E-mail: lucimang@126.com).

tasks indicate that, under comparable energy consumption, the improved flash chip achieves a more concentrated inference error distribution and consistently superior inference accuracy across all network layers. The overall inference accuracy attains 94.6%, marking an enhancement of approximately 4.5% over conventional flash chips. These findings demonstrate that extending the linear operating region effectively mitigates nonlinear error accumulation during forward inference, substantially enhances the reliability of analog computation, and improves inference accuracy. This research offers a novel device-level optimization approach for the development of future high-accuracy and low-power neural network accelerators.

Key words: computing-in-memory (CIM); analog memory; programmable linear random-access memory; linear region; inference accuracy

0 引言

为突破冯·诺伊曼架构在人工智能计算中的能效瓶颈，存算一体(computing-in-memory, CIM)架构近年来受到广泛关注。该架构通过在存储阵列内部直接执行向量-矩阵乘积等运算，有效避免了频繁的数据搬移，从而实现了高并行度和低能耗的神经网络计算[1-7]。在器件层面，利用模拟存储器件的新型工作区来实现多比特权重存储，有助于满足边缘计算场景对高集成密度和低功耗的要求。

目前，多种新型非易失存储器(non-volatile memory, NVM)已被用于构建人工突触，包括阻变存储器(resistive random-access memory, RRAM)、相变存储器(phase-change memory, PCM)[8-10]、磁阻随机存储器(magnetoresistive random-access memory, MRAM)[11-12]等。这些器件具有结构简单、功耗低以及与 CMOS 工艺兼容等优势，但其实际应用仍受限于器件可靠性不足、阻值变异与漂移以及导通电导调节的非对称性等固有缺陷，这些问题严重影响了神经网络的训练和推理精度。因此，亟须开发兼具多态模拟存储能力和高可靠性的存算一体器件。

闪存(flash memory)作为一种典型的电荷型非易失性存储器，核心优势在于断电后仍能保持数据的完整性。闪存的工作原理基于浮栅晶体管的电荷调控机制，通过向存储单元注入或释放电子来调控浮栅电位，从而实现数据的存储与读取。凭借低成本、高存储密度、良好的工艺兼容性及非易失性等特性，闪存在嵌入式系统和大规模数据存储领域得到了广泛应用。然而，传统闪存器件依赖宽阈值电压窗口进行二值存储，难以支持高线性度和细粒度的电导调节。此外，编程过程中采用的热载流子注入机制会对栅氧化层造成较大冲击，易引发界面陷阱和击穿问题，从而制约了器件的可靠性与使用寿命。

为突破传统闪存在线性度与可靠性方面的瓶颈，本研究团队基于 55 nm 工艺平台提出了一种新型闪存单元——可编程线性随机存取存储器(programmable linear random-access memory, PLRAM)[13]，如图 1 所示。该器件采用解耦控制栅与隧穿氧化层的设计，并引入双向 Fowler-Nordheim隧穿机制来实现电子的注入与抽取，有效缓解了传统器件中

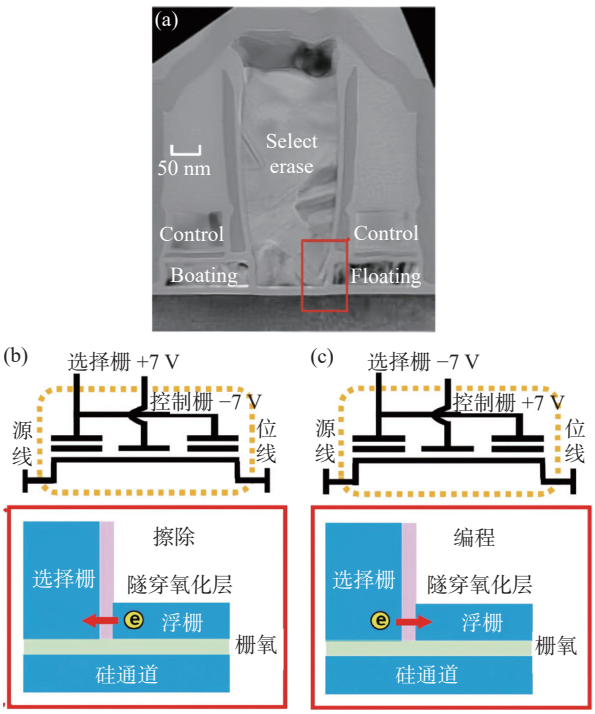

图 1 (a)PLRAM 的单元结构；(b)擦除与(c)编程机制示意图[13]

Fig. 1 (a) Structure of the PLRAM cell; (b) schematic diagram of the erase and (c) program mechanism[13]

热载流子注入对栅氧化层的损伤, 显著提升了器件的可靠性与使用寿命。得益于这种对称隧穿结构, PLRAM 在编程与擦除过程中能够实现高精度、细粒度的电导调节, 满足高位宽模拟存储对电导可控性的需求[13-18]。

尽管 PLRAM 器件具备优异的单步电导调控能力, 但在开环编程/擦除模式下, 仍易受到器件涨落和过程波动的影响, 难以实现高精度电导控制。为此, 本团队提出了一种自校准编程/擦除方案(self-calibrated program/erase scheme)[13], 通过在每次编程前实时测量器件的电导状态并施加多档控制电压, 以实现不同速率的电导更新, 从而实现对电导状态的精准调节。该方法显著提升了编程精度, 使单个 PLRAM 单元能够支持高达 7 bit 的模拟存储精度。

基于 PLRAM 的系统级芯片(system-on-chip, SoC)架构如图 2 所示, 其主要由 PLRAM 存算阵列、片上静态随机存取存储器(Static Random-Access Memory, SRAM)、计算与控制单元、数模转换模块(digital-to-analog converter, DAC)、模数转换模块(analog-to-digital converter, ADC)及 I/O 接口组成。该芯片采用 PLRAM 阵列存储网络各层的权重矩阵, 并利用其线性欧姆特性来执行向量-矩阵乘法运算。输入神经元电路负责将上一层的数字输出转换为模拟电压, 以驱动下一层 PLRAM 阵列的字线; 输出神经元电路则读取位线上的模拟累加结果, 并通过模数转换得到数字结果。

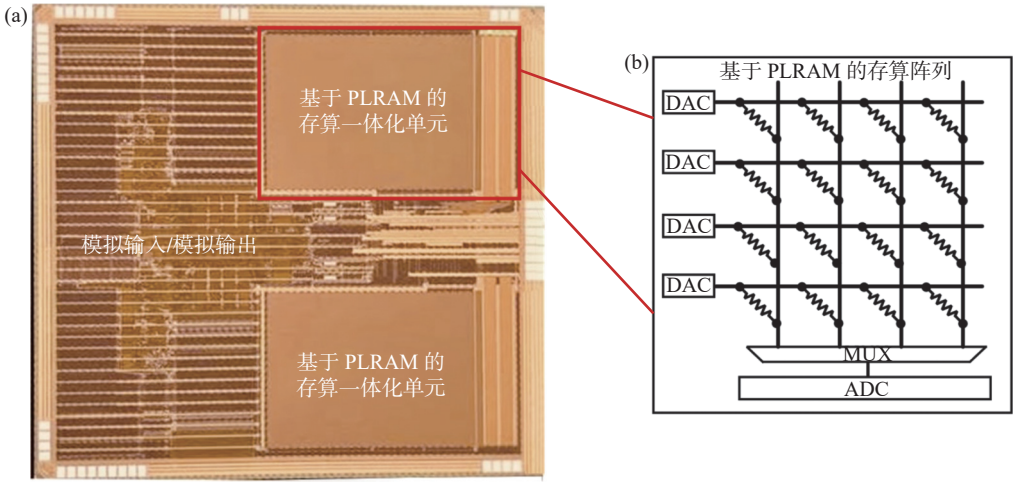

图 2　(a) 改进型闪存裸芯片(die)照片及 (b) 存算阵列结构示意图

Fig. 2　(a) Die photograph of the enhanced flash chip and (b) schematic diagram of its computing-in-memory array architecture

本文提出一种基于新一代 PLRAM 的存内计算芯片, 其核心创新在于通过对工艺与设计参数的优化, 可显著拓展存储单元的线性工作区, 从而提升模拟计算的线性度与稳定性。在典型神经网络推理任务中, 该芯片与前代产品进行了系统性对比。实验结果表明, 在能耗相近的情况下, 改进型闪存芯片显著降低了推理误差, 并提升了整体计算精度。

1　PLRAM 器件线性工作区的改进

在存算一体芯片中, PLRAM 单元依靠其漏极电流-漏极电压(I_d-V_d)线性导通特性来实现多比特权值存储功能。当其工作在线性区时, 漏极电流 I_d 近似与输入电压成比例, 其电导值由浮栅电荷状态决定, 并以连续可调的方式映射至神经网络权重。然而, 随着漏源电压的增加, 存储晶体管逐渐进入饱和区, 导致电流-电压关系偏离线性, 从而引入推理误差。器件层面波动的累积会进一步引发阵列失配与偏差, 这种阵列级误差在网络前向传播过程中逐层累积, 最终导致输出误差, 如图 3 所示。

为缓解 PLRAM 器件的非线性问题, 采用沟道长度更短的浮栅晶体管, 以增强漏致势垒降低效应(drain-induced barrier lowering, DIBL)。假设在零漏极偏置下的过驱动电压为 $V_{OV} \triangleq mV_{GS} - V_{TH0}$(其中, V_{GS} 为栅源电压, V_{TH0} 为漏源电压 V_{DS} 趋于 0 时的等效阈值电压), 考虑到 DIBL 效应会使阈值电压随 V_{DS} 下降, 有效阈值电压 V_{TH}^{eff} 可表示为:

$$V_{TH}^{eff}(V_{DS}) = V_{TH0} - \eta V_{DS}, \quad 0 \leqslant \eta < 1 \tag{1}$$

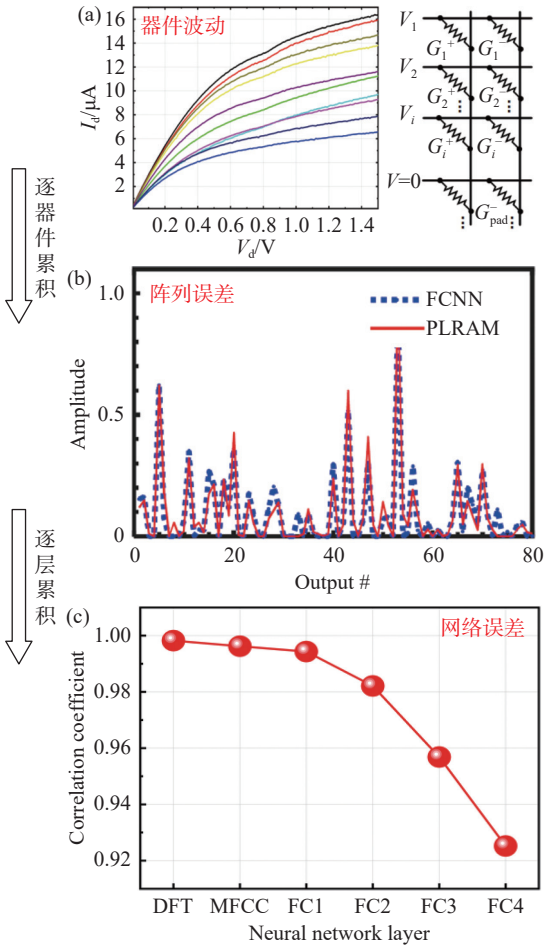

图 3 PLRAM 存算一体芯片中“器件—阵列—网络”误差传递机制。(a) 单器件特性图; (b) 阵列特性图; (c) 网络特性图

Fig. 3 The "Device-Array-Network" error transmission mechanism in PLRAM computing-in-memory chips. (a) Single-device characteristic diagram; (b) array characteristic diagram; (c) network characteristic diagram

式中, η 为 DIBL 系数, 其数值随沟道长度 L 的减小而增大。根据线性工作区的判定条件 $V_{DS} \leqslant V_{GS}-V_{TH}^{eff}$, 可推导出进入饱和区的临界电压 V_{DS}^{in} 为:

$$V_{DS} \leqslant V_{OV}+\eta V_{DS} \Rightarrow V_{DS}^{in} \leqslant \frac{V_{OV}}{1-\eta} \tag{2}$$

由式 (2) 可知, 通过缩短沟道使 η 增大, 可进而提高 V_{DS}^{in}。这使得器件在更高的 V_{DS} 下仍能保持近似线性导通。在线性工作区内, 工作电流 I_D 可近似用式 (3) 表示:

$$I_D \approx k\left(V_{OV}-\frac{1}{2}V_{DS}\right)V_{DS} \tag{3}$$

$$k=\mu C_{ox}\frac{W}{L} \tag{4}$$

式中, k 为跨导系数, μ 为沟道载流子迁移率, C_{ox} 为单位面积栅氧电容, W 为有效沟道宽度。可见, 通过减小沟道长度 L 可使跨导系数 k 增大, 进而显著扩展 I_d-V_d 曲线的线性工作区, 使得晶体管在更高漏极偏压下仍能保持近似线性导通, 如图 4 所示。这一改进有效提升了权值映射的精确性, 降低了模拟矩阵计算中的非线性累积误差, 为实现多比特高精度计算提供了更可靠的器件基础。

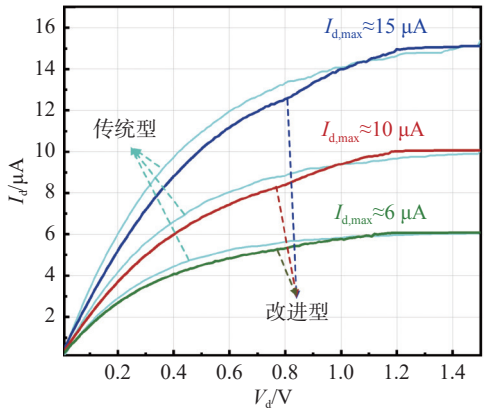

图 4 改进型闪存芯片与传统闪存芯片中 PLRAM 器件的 I_d-V_d 特性曲线对比

Fig. 4 Comparison of I_d-V_d characteristic curves of PLRAM devices in an enhanced flash chip and a conventional flash chip

此外, 为进一步降低器件非线性引起的累积误差, 改进型闪存芯片在阵列与电路层面引入了“列归一化+权重缩放”机制, 如图 5 所示。具体实现方式为: (1)采用小权重提升策略对每列权重所映射的电导 G_{ij} 进行比例缩放, 使所有单元工作点均远离晶体管饱和区, 从而提高器件的线性度; (2)将列电导总和对齐至预设目标值 G_{sum}; (3)利用平移不

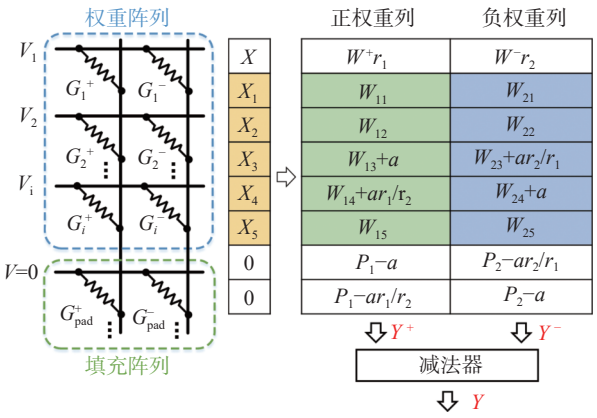

图 5 基于填充阵列与权重缩放的列归一化机制

Fig. 5 Column normalization mechanism based on dummy arrays and weight scaling

变性在差分正/负列上同步引入小偏置, 并在每列底部增设填充单元以实现权重归一化, 确保各列总导通水平一致, 以降低列间失配引起的系统性偏差。列电导的和满足:

$$\sum_i G_i + \sum G_{\text{pad}} = G_{\text{sum}} \tag{5}$$

式中, G_{pad}、G_{sum} 分别为填充阵列的电导以及列电导和。差分输出电压 V_{out} 可表示为:

$$V_{\text{out}} = \sum_i V_i \frac{(G_i^+ - G_i^-)}{G_{\text{sum}}} \tag{6}$$

式中, V_i 为第 i 行的输入电压, G_i^+和G_i^-分别为第 i 行主权重阵列的正、负电导值。

上述架构与电路层面的设计优化, 共同保障了改进型闪存芯片在执行大规模模拟矩阵计算时的稳定性和精度。

2　PLRAM 芯片推理精度验证

为系统评估改进型闪存芯片的推理性能, 构建一个 3 层全连接神经网络, 并与前代 Flash 芯片进行对比实验。实验采用相同的输入数据和网络结构, 以确保结果的可比性。采用相关系数 C_{C} 作为评价指标, 用于衡量芯片计算结果与理想模型输出之间的一致性, C_{C} 值越大, 表示计算精度越高。

如图 6 所示, 相比于传统闪存芯片, 改进型闪存芯片的推理误差分布在零点附近收敛得更为集中, 表明线性工作区扩展与电路优化有效抑制了非线性误差的累积。同时, 从图 7 可以看出, 改进型闪存芯片在各层的推理输出均表现为更高的相关系数: FC1、FC2 与 FC3 层分别达到 0.996 6、0.994 2 和 0.977 7, 相较于传统闪存芯片的 0.995 0、0.991 9 和 0.968 3 均有提升。最终推理准确率达到 94.6%, 相较于传统闪存芯片提升 4.5%。这一结果表明, 在保持相近能耗水平的前提下, 改进型闪存芯片能够显著提升网络整体的推理精度。

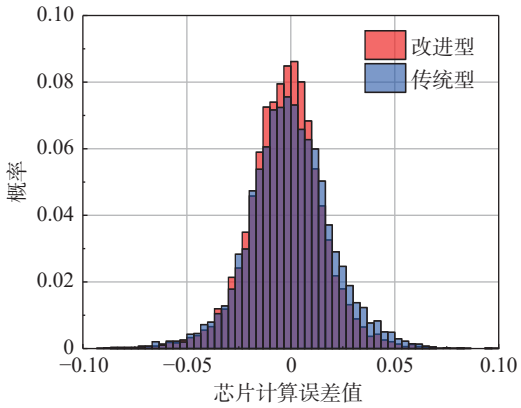

图 6　改进型闪存芯片与传统闪存芯片推理误差概率分布的对比

Fig. 6 Comparison of inference error probability distribution between enhanced flash chip and conventional flash chip

表 1 对比了本文改进型闪存芯片与前代芯片及 RRAM 存算芯片的关键性能指标。可以看到, 改进型闪存芯片在 55 nm 工艺下实现了 8 MB 的存储容量, 阵列规模扩展至 4 096×1 024。得益于更高的并行度与优化的线性工作区特性, 芯片能效由 11 TOPS·W^{-1} 提升至 25 TOPS·W^{-1}, 面积为 12.24 mm^2。其中, TOPS(tera operations per second)表示每秒万亿次操作。

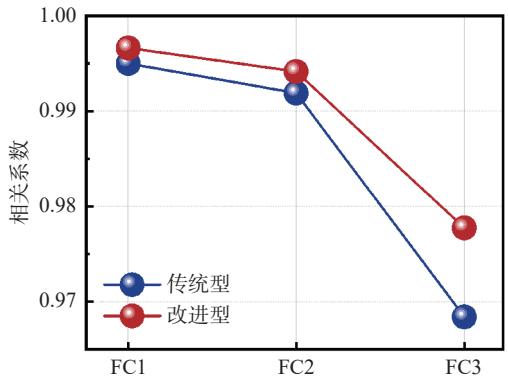

图 7　改进型闪存芯片与传统闪存芯片推理相关系数的对比

Fig. 7 Comparison of inference result correlation coefficients between the enhanced flash chip and the conventional flash chip

表 1　基于 PLRAM 的存算一体化架构芯片性能对比

Table 1 Performance comparison of computing-in-memory chips based on PLRAM architecture

对比项	IDEM 2019[16]	RRAM[19]	本文
工艺节点/nm	90	22	55
存算单元	PLRAM	RRAM	PLRAM
权重容量/MB	0.36	4	8
计算精度	Int8	Int8	Int8
能效/(TOPS·W^{-1})	11	51.4	25
面积/mm^2	N/A	24.48	12.24
阵列规模	128×128	512×1 024	4 096×1 024×2

可见，改进型闪存芯片在存储容量、能效和集成度方面均取得了显著进展，并成功实现了语音识别、目标检测等边缘场景应用，验证了该存算方案在资源受限以及对能效和时延敏感场景下的应用前景。

3 结论

本文介绍了一种基于改进型 PLRAM 实现的存算芯片，通过缩短沟道长度和优化工艺，显著扩展了器件 I_d-V_d 的线性工作区，同时结合列归一化与权重缩放机制，有效抑制了非线性误差的累积。与前代闪存芯片相比，改进型闪存芯片大幅降低了推理误差，使 3 层全连接网络的整体推理精度提升了 4.5%。此外，在存储容量、阵列规模以及能效方面均取得显著提升，为大规模神经网络的硬件部署奠定了重要基础。

参考文献：

[1] LI J, ABBAS H, ANG D S, et al. Emerging memristive artificial neuron and synapse devices for the neuromorphic electronics era[J]. Nanoscale Horizons, 2023, 8(11): 1456-1484.

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

[3] MARKOVIĆ D, MIZRAHI A, QUERLIOZ D, et al. Physics for neuromorphic computing[J]. Nature Reviews Physics, 2020, 2(9): 499-510.

[4] BURR G W, SHELBY R M, SEBASTIAN A, et al. Neuromorphic computing using non-volatile memory[J]. Advances in Physics: X, 2017, 2(1): 89-124.

[5] JOSHI V, LE GALLO M, HAEFELI S, et al. Accurate deep neural network inference using computational phase-change memory[J]. Nature Communications, 2020, 11: 2473.

[6] 朱赛克, 赵毅. 面向神经形态系统的突触器件及芯片综述与展望 [J]. 功能材料与器件学报, 2024, 30(6): 287-299.

[7] 陈华, 曲益明, 吴文豪, 等. 低功耗人工智能计算系统研究进展综述 [J]. 功能材料与器件学报, 2024, 30(6): 300-309.

[8] WUTTIG M, BHASKARAN H, TAUBNER T. Phase-change materials for non-volatile photonic applications[J]. Nature Photonics, 2017, 11(8): 465-476.

[9] FONG S W, NEUMANN C M, PHILIP WONG H S. Phase-change memory: Towards a storage-class memory[J]. IEEE Transactions on Electron Devices, 2017, 64(11): 4374-4385.

[10] PARK S O, HONG S, SUNG S J, et al. Phase-change memory via a phase-changeable self-confined nano-filament[J]. Nature, 2024, 628(8007): 293-298.

[11] 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.

[12] KENT A D, WORLEDGE D C. A new spin on magnetic memories[J]. Nature Nanotechnology, 2015, 10(3): 187-191.

[13] GAO S, YANG G, QIU X, et al. Programmable linear RAM: A new flash memory-based memristor for artificial synapses and its application to speech recognition system[C]// Proceedings of the 2019 IEEE International Electron Devices Meeting. New York: IEEE, 2019: 14.1. 1-14.1. 4.

[14] GAO S, CONG Y, ZHANG Z, et al. Superior data retention of programmable linear RAM (PLRAM) for compute-in-memory application[C]//Proceedings of the 2020 IEEE International Reliability Physics Symposium (IRPS). New York: IEEE, 2020: 1-5.

[15] 赵亮, 高世凡, 张圣波, 等. 基于闪存的可编程线性忆阻器及其存算一体化电路与系统 [J]. 中国科学: 信息科学, 2022, 52(1): 176-186.

[16] 高世凡. 基于片上非易失存储器的模拟计算架构研究 [D]. 杭州: 浙江大学, 2023.

[17] JIN J, GAO S, LU C, et al. Device nonideality-aware compute-in-memory array architecting: direct voltage sensing, *I-V* symmetric bitcell, and padding array[J]. IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 2025, 11: 19-24.

[18] ZHU S, LU C, QIU X, et al. Mitigating RC-delay induced accuracy loss in analog in-memory computing: a non-compromising approach[J]. IEEE Transactions on Consumer Electronics, 2024, 70(4): 7544-7550.

[19] HUANG W H, WEN T H, HUNG J M, et al. A nonvolatile AI-edge processor with 4 MB SLC-MLC hybrid-mode ReRAM compute-in-memory macro and 51.4-251 TOPS/W [C]//Proceedings of the 2023 IEEE International Solid-State Circuits Conference. New York: IEEE, 2023: 15-17.