中国科学: 信息科学 2022 年 第 52 卷 第 4 期: 612–638
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
纪念清华大学电子工程系成立 70 周年专刊 · 评述

存算一体电路与跨层次协同设计优化: 从 SRAM 到铁电晶体管

尹勋钊[1†], 岳金山[2†], 黄庆荣[1], 李超[1], 蔡嘉豪[1], 杨泽禹[1], 卓成[1*], 刘明[2*]

1. 浙江大学信息与电子工程学院, 杭州 310012
2. 中国科学院微电子研究所, 北京 100029
* 通信作者. E-mail: czhuo@zju.edu.cn, liuming@ime.ac.cn
† 同等贡献

收稿日期: 2021–12–17; 接受日期: 2022–01–27; 网络出版日期: 2022–03–29

国家自然科学基金 (批准号: 62104213, 62034007, 62141404) 和北京市科技新星计划 (批准号: Z211100002121125) 资助项目

摘要 人工智能与物联网时代, 大数据模型驱动的应用场景和计算任务层出不穷, 极大促进了国家数字化发展. 然而, 传统冯 · 诺依曼 (John von Neumann) 体系架构的硬件系统由于存算分离的结构特点导致存储墙瓶颈, 在数据密集型应用中消耗了大量的数据搬运成本, 抑制了能效性能提升. 存算一体技术是后摩尔 (Moore) 时代背离传统架构系统的新型计算范式, 利用存储单元器件、电路内在特性, 将基本的计算逻辑任务融入存储单元之中, 从而消除数据搬运开销, 有望实现智能计算硬件平台能效性能的显著提升. 本文以契合存算一体技术的存储器件电路为切入点, 概述基于传统互补金属氧化物半导体 (complementary metal oxide semiconductor, CMOS) 和新型非易失存储器件代表铁电晶体管的存算一体电路, 并从器件、架构芯片、算法应用等层次讨论存算一体电路的跨层次协同设计优化方法.

关键词 存内计算, 静态随机访问存储器, 铁电晶体管, 交叉阵列, 内容寻址存储器

1 引言

人工智能和物联网时代来临, 数据密集型计算模型任务如神经网络、数据搜索等对硬件中的芯片系统在能效、性能、尺寸成本等方面提出了更高要求. 然而, 传统基于冯 · 诺依曼 (John von Neumann) 架构的计算芯片系统由于其固有的存储计算分离的结构特点, 在处理数据密集型模型算法时大量能耗和信号延迟损失在频繁的数据搬运和内存访问上, 导致硬件性能能效表现不佳, 即所谓存储墙瓶颈 [1]. 按照神经网络算法对于存储带宽的需求和实际带宽需求之间的差距 [2], 传统冯 · 诺依曼体系架构芯片

引用格式: 尹勋钊, 岳金山, 黄庆荣, 等. 存算一体电路与跨层次协同设计优化: 从 SRAM 到铁电晶体管. 中国科学: 信息科学, 2022, 52: 612–638, doi: 10.1360/SSI-2021-0420
Yin X Z, Yue J S, Huang Q R, et al. Computing-in-memory circuits and cross-layer integrated design and optimization: from SRAM to FeFET (in Chinese). Sci Sin Inform, 2022, 52: 612–638, doi: 10.1360/SSI-2021-0420

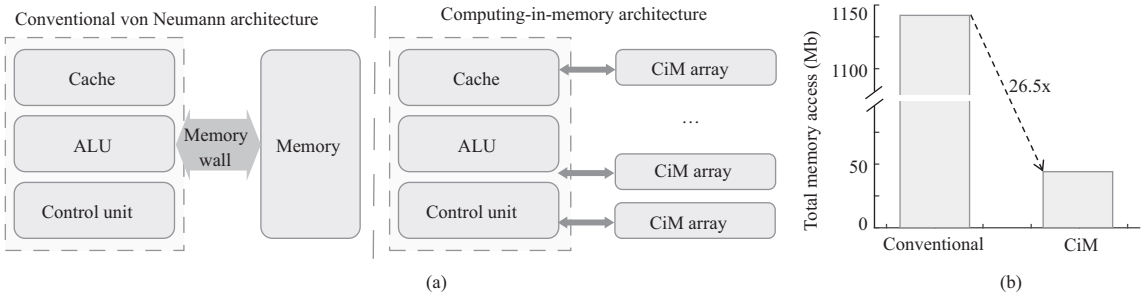

图 1 (a) 冯 · 诺依曼架构和存算一体架构比较示意图以及 (b) 两者在访存操作数量上的对比 [6]

Figure 1 (a) Comparison between conventional von Neumann and computing-in-memory architectures and (b) performance comparison between conventional and CiM architectures

将面临需要把 90% 以上的运算资源都消耗在数据迁移上的窘境. 由于存储墙是冯 · 诺依曼体系架构的固有问题, 在此架构基础上的优化技术可以缓解甚至改善这个问题 [3,4], 但是需要付出功耗、成本、设计复杂度等一系列代价, 不能从根本上加以解决. 这将阻碍传统计算芯片系统在高能效边缘端及终端智能场景中的应用, 严重影响人工智能和物联网硬件设备的发展.

针对人工智能和物联网时代数据密集型应用的高能效、高性能硬件需求, 业界从新的计算范式和新器件两方面创新, 提出了一系列旨在克服传统架构存储墙瓶颈的电路芯片解决方案. 神经形态计算 [5] 通过模拟生物大脑中神经元的结构、全方位互联状态和工作机理, 将负责数据存储和处理的单元整合到同一模块中, 大幅提升了计算系统的信息处理能力和智能计算学习能力. 另一方面, 针对机器学习、模式识别搜索、数字信号处理等应用固有的容错性质, 近似计算 [6] 通过有选择地控制、放松计算操作的精度, 引入误差, 使得近似的计算结果在仍满足上述应用任务的准确度要求的同时, 实现系统能效和性能的大幅提升. 除了上述新型计算范式, 针对传统架构中的存储墙问题, 既然数据迁移成了制约智能计算芯片能效的瓶颈, 我们不妨让计算和存储都发生在同一个地方以彻底避免数据迁移的问题. 这种在存储里加上计算功能的新型范式被称为存内计算或存算一体 (computing-in-memory, CiM 或 process-in-memory, PiM) [7,8]. 如图 1(a) 所示, 存内计算与传统冯 · 诺依曼架构的本质不同在于, 它将计算单元集成在存储器内, 执行计算时不需要数据迁移, 直接在访问数据时就同时完成运算, 因此从根本上解决了存储墙问题. 从图 1(b) [6] 可以看到, 存算一体在解决访存带来的存储墙问题上有很大的优势. 存内计算的概念可回溯到 20 世纪 90 年代, 受限于内存工艺发展未获得广泛应用 [9,10]. 随着神经网络和数据搜索应用的兴起, 存内计算因其高并行性和高能效特点重获关注. 与传统 CPU/GPU 运算不同, 神经网络模型中涉及大量乘累加运算的权重是固定的, 只有输入是实时产生的. 存内计算不仅可利用其并行性加速计算, 还能将权重存储在内存中减少数据搬运, 极大提高了硬件整体系统能效. 近几年工作进展 [11~14] 验证了存内计算相较于传统计算架构, 能以更高能效算力支持逻辑运算、并行搜索和神经网络乘加运算.

新型非易失存储 (non-volatile memory, NVM) 技术的兴起成熟进一步加速了存算一体技术的发展和落地 [15~18]. 新型 NVM 器件, 特别是以自旋转移力矩磁随机存储器 (spin-transfer torque magnetic random access memory, STT-MRAM)、相变存储器 (phase change memory, PCM)、阻变存储器 (resistive RAM, ReRAM) 和铁电晶体管 (ferroelectric FET, FeFET) 等器件为代表的一大批新型工艺器件 [19~21], 可将神经网络模型中的权重或数据搜索中的特征数据向量编码成器件本身的物理状态, 并作为操作数利用物理定律如欧姆定律 (Ohm's law)、基尔霍夫定律 (Kirchhoff's law) 参与到运算中形成紧凑高效的存算电路单元及阵列. 如何利用新型 NVM 器件设计优化支持神经网络或数据搜索关键功能的存算

电路以满足人工智能和物联网应用场景需求, 已成为存算一体技术在器件、电路芯片层面的主要发展方向. 现在已有多种基于先进工艺节点的上述新型 NVM 器件基本原型, 包括 22 nm 和 28 nm 节点上的 STT-MRAM [22, 23], 22 nm RRAM [24] 和 40 nm PCM [25]. 文献 [20, 21] 总结介绍了各类新型存储器件的关键参数、性能指标, 以及相应的存内计算电路实现形式等. 其中, STT-MRAM 和 PCM 是目前最为成熟的新型 NVM 器件, 前者利用自旋极化电流翻转磁层方向以存储器件状态, 需要大尺寸存取晶体管提供高写入电流, 面积效率需进一步提升; 后者则利用电流脉冲实现高阻非晶态和低阻晶态间的可逆相变, 其高低电阻比高于 STT-MRAM, 可实现更高读取容限. RRAM 通过电流写入实现非易失存储的高低电阻状态, 其写入电压与电路电源电压兼容 [26], 相对克服了 PCM 的高写入能耗和延迟弱点, 但由于固有变异性会在大规模存内计算阵列中产生计算误差. 上述 NVM 器件均以可变电阻的形式构建存内计算电路, 其阻值状态对应存储逻辑值. 然而, 这类电流驱动的器件受制于较高的写入能耗, 以及较低的高低电阻比带来的额外电路设计开销, 导致更多功耗. 与之相对, FeFET 因与互补金属氧化物半导体 (complementary metal oxide semiconductor, CMOS) 先进工艺节点的兼容性 [27~29]、独特的三端口结构、可比 CMOS 晶体管的开关电流比和电压驱动的读写机制, 其在构建存内计算电路上相比前述 NVM 实现了设计面积开销、读写能效性能等多方面的优势 [30].

综上, 在之前存算一体技术综述基础上 [20, 31], 本文将论述存算一体技术基于传统 CMOS 工艺静态随机访问存储器 (static random access memory, SRAM) 单元的电路架构芯片协同设计优化工作, 同时以 FeFET 为代表, 介绍利用新型 NVM 器件特性实现高能效低面积存算一体关键功能电路及其应用的一系列工作.

2 基于 SRAM 的存内计算电路

静态随机存储器 SRAM 的工艺比较成熟, 易与其他逻辑器件集成, 因此较多研究工作基于 SRAM 开展了存内计算阵列和包含外围电路的芯片系统设计. 本节将从存内计算阵列、单元电路、架构和算法/电路联合优化等方面介绍基于 SRAM 的存内计算电路设计.

2.1 SRAM 存内计算阵列与单元电路

基于 SRAM 的存内计算阵列的实现方式主要有电流式、电荷式和全数字 3 种. 这 3 种方式的主要区别在于单个存内计算单元产生结果的形式不同. 电流式存内计算 [12, 32~34] 在每一个计算单元将乘法结果表示为一个电流, 然后通过位线将同一列上多个计算单元的电流汇聚到 ADC 输出累加结果. 电荷式存内计算 [35~39] 在每一个存储单元以电容电荷的形式表示乘法结果, 然后通过位线的电荷重分配实现同一列上多个电容电荷的求和平均, 最后送到 ADC 输出累加结果. 一般而言, 电流式存内计算结构简单, 所需晶体管数量较少; 电荷式存内计算结构中, 电容匹配误差小于电流式存内计算结构中产生电流的晶体管阈值电压误差, 因此在相同准确率要求下实现了较高计算并行度. 全数字存内计算 [40~42] 在每一个计算单元将乘法结果表示为 0/1 的数字信号, 并通过传统数字电路加法器树的形式, 在存储器内部完成累加后输出. 全数字存内计算结构没有模拟电路的误差因素, 计算结果能够与传统数字电路完全一致, 因此能够实现高准确率, 但加法器树的数字逻辑电路会占据较大面积.

2.1.1 电流式存内计算单元电路

从具体的电路实现形式看, SRAM 的电流式存内计算单元经历了从 6T, 8T 到分组 6T 的转变, 如图 2(a)~(c) 所示. 尽管采用标准 6T 的 SRAM 单元能够实现最高的存储密度, 但直接在位线上进行计

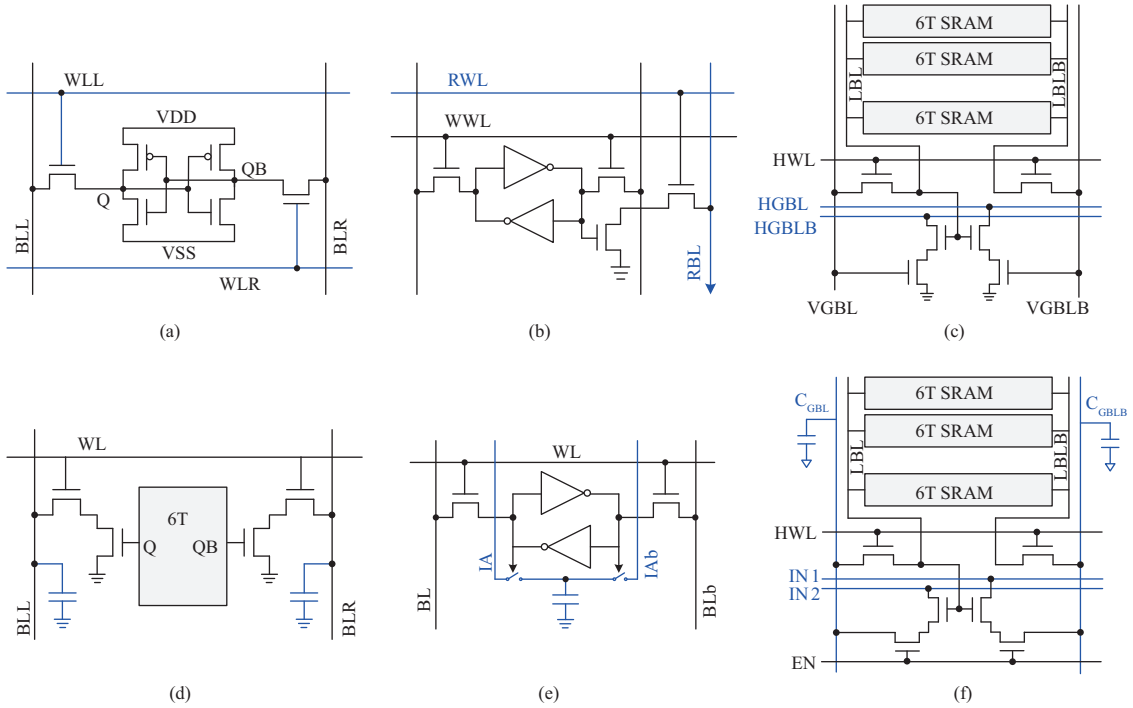

图 2 (网络版彩图) 电流式 (a)~(c) 和电荷式 (d)~(f) SRAM 存内计算单元电路 [12, 33~36, 39]

Figure 2 (Color online) Current-based (a)~(c) and charge-based (d)~(f) SRAM CiM cell circuits. (a) 6T CiM cell; (b) 8T CiM cell; (c) block-wise 6T CiM cell; (d) 10T CiM cell; (e) 8T CiM cell; (f) block-wise 6T CiM cell

算, 其允许的电压范围较小, 否则会引起写扰动的问题, 且 ADC 的感应裕度较小. 相关工作 [12] 采用了两条分离的字线, 每次仅使用其中一条字线进行存内计算操作, 避免写扰动的问题, 但感应裕度仍然较小. 为进一步提升感应裕度, 相关工作 [32,33] 采用了 8T 的结构, 增加两个额外的晶体管, 权重数据连接到存内计算晶体管的栅极, 从而避免写扰动的问题, 并提升感应裕度. 通过进一步节省与输入数据相连的栅极所对应的晶体管, 直接将输入数据连接到权重数据控制的晶体管的源/漏极, 可以实现 7T 的存内计算单元 [43]. 为降低 8T 结构的较大面积开销, 相关工作 [34] 提出了分组 6T 的结构, 由多个 6T 的 SRAM 存储单元共享存内计算电路, 每次仅打开一个 SRAM 用于存内计算, 其面积开销取决于单个分组中 6T SRAM 单元的数量. 在每个分组中包含 16 个 SRAM 单元和 1 个存内计算单元的情况下, 其单个存储单元的等效面积开销大约是标准 6T SRAM 单元的 1.34 倍.

2.1.2 电荷式存内计算单元电路

与电流式存内计算类似, 电荷式 SRAM 存内计算单元电路也包含 10T, 8T 和分组 6T 等多种形式, 每一个存储单元含有一个或两个电容, 如图 2(d)~(f) 所示. 10T 的存内计算单元 [35] 在 6T SRAM 单元两侧各增加两个晶体管和一个电容用于采样存储值 “0” 和 “1” 对应的计算结果, 并通过电荷分配的方式, 在 ADC 结果中进行采样. 8T 的存内计算单元可以有多种实现形式 [36~38], 其中一种 [36] 通过两个开关晶体管将 SRAM 单元的存储值与电容连接, 实现预充、计算与电荷重分配、放电等功能. 分组 6T 的电容式存内计算单元 [39] 与分组 6T 的电流式存内计算单元类似, 两者区别在于存内计算电路部分由电流累加替换为电荷重分配. 在每个分组中包含 32 个 SRAM 单元和 1 个存内计算单元的情况下, 其单个存储单元的等效面积开销大约是标准 6T SRAM 单元的 1.28 倍.

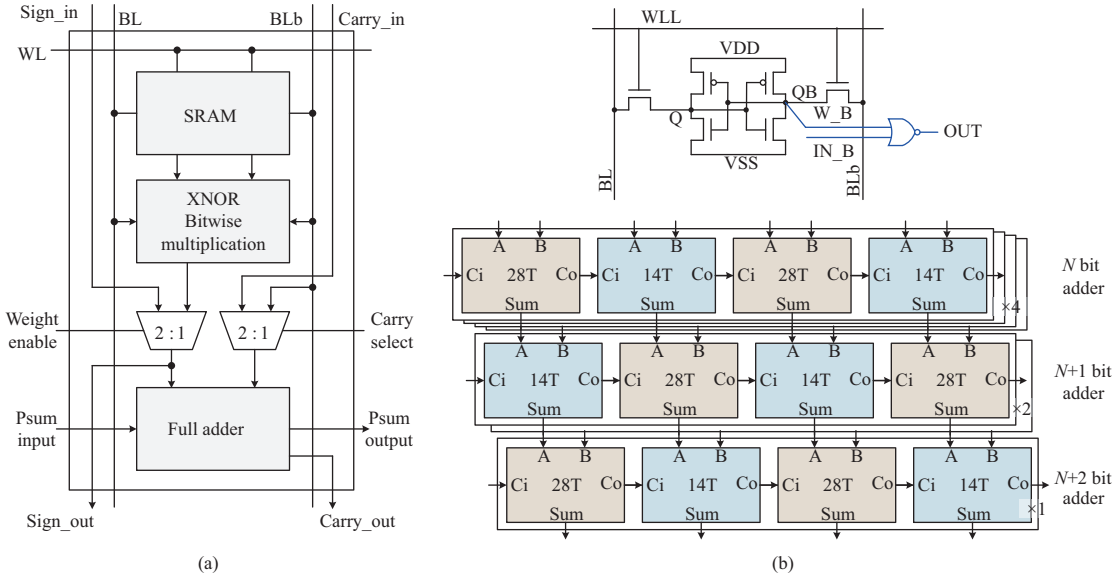

图 3 (网络版彩图) 全数字存内计算单元电路. (a) 集成一位全加器的 SRAM 单元 [40]; (b) 集成 1 bit 乘法电路的 SRAM 单元与加法器树 [41]

Figure 3 (Color online) All-digital CiM circuits. (a) SRAM cell with 1-bit full adder; (b) SRAM cell with 1-bit multiplier and adder tree

2.1.3 全数字存内计算单元电路

存内计算中的模拟电路引入的误差会使计算结果存在不确定性, 导致神经网络计算准确率下降. 而基于全数字电路的存内计算单元 [40~42] 实现了完全没有精度损失的存内计算电路. 其中一种实现方式 [40] 如图 3(a) 所示, 在每一个 SRAM 存储单元旁边放置一个一位全加器, 每一个 SRAM 存内计算单元可以执行一位全加器运算, 并把加法结果和进位传递给下一个 SRAM 单元的一位全加器. 通过不同的配置组合, 该电路可以实现对于 16 行 1 bit 权重或 5 行 16 bit 权重的计算, 且可以实现 1~16 bit 输入激活值的计算. 韩国科学技术院 (Korea Advanced Institute of Science and Technology) 设计的一种电路 [42] 采用了较低的计算并行度, 每次打开存储阵列中的一行, 利用输入的稀疏性, 跳过零值输入的对应行, 并采用流水线的读取操作提升性能. 如图 3(b) 所示, TSMC 设计的一种电路 [41] 通过存内计算阵列中的一位乘法单元和加法器树实现了对 256 行 4 bit 权重与 1 bit 输入数据的乘法和累加操作, 在加法器树结构中, 采用 14T 和 28T 交替的全加器结构, 实现了功耗面积和延时的折中, 提升了 30% 的能量效率. 该工作对于 4 bit 权重和 4 bit 激活值的计算实现了 89 TOPS/W 的能量效率.

2.1.4 其他形式/功能的存内计算单元电路

大部分存内计算阵列采用从行或者列中的一个方向进行输入的方式. 对于神经网络训练过程中的反向传播过程等需求, 需要对转置的权重矩阵进行乘法操作. 基于 SRAM 的可转置存内计算阵列, 可按照需求配置工作模式, 允许从行/列两个方向中任选一个方向进行输入 [44]. 此外, 对于二值化同或 (not exclusive or, XNOR) 神经网络, 其单元电路为实现同或操作, 相比用作乘法的单元结构会做相应修改. 文献 [45] 采用一种字线分离式的 6T 电路结构, 两条字线互为反相输入值, 与 SRAM 存储值 Q 和 $\bar{Q}$ 进行运算, 从而实现异或/同或运算.

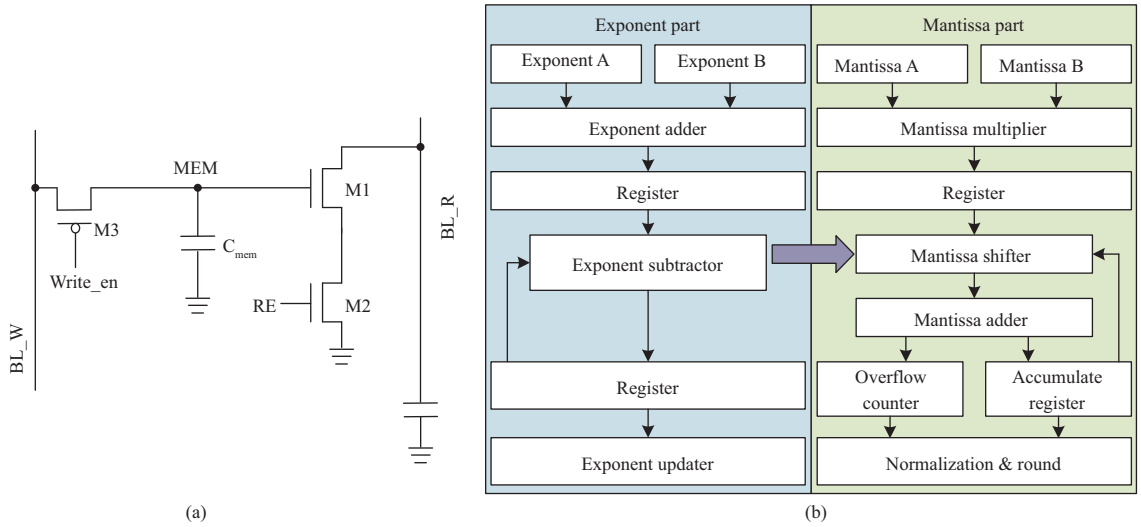

图 4 (网络版彩图) 新型存内计算单元电路结构与功能. (a) 3T1C 动态模拟存储器单元电路 [46]; (b) 存内计算/数字电路混合的浮点存内计算电路 [47]

Figure 4 (Color online) Novel CiM cell circuits and function. (a) 3T1C DARAM CiM cell; (b) CiM-digital-mixed floating-point circuits

此外, 针对存内计算操作还有其他的电路形式. 如图 4(a) 所示, 动态模拟存储器 (dynamic analog RAM, DARAM) [46] 采用 3T-1C 的结构, 其功能介于 SRAM 和 DRAM 之间. 其中 1T-1C 实现与 DRAM 类似的存储功能, 另外的 2T 实现存内计算的操作. 其存储密度介于 SRAM 和 DRAM 之间, 可实现类似于 SRAM 的较高计算并行度, 但依然需要动态刷新操作来保持电容的电荷状态.

浮点存内计算是一种新型存内计算电路功能. 由于浮点数包含符号位、底数位和指数位, 存内计算在并行运算时难以实现基于指数位的移位操作, 因此大多数存内计算工作均以定点数计算为主. 如图 4(b) 所示, 文献 [47] 探索了在存内计算架构中实现浮点操作的电路. 设计采用存内计算和数字电路紧耦合的方式, 由传统数字电路实现底数位对应的加法和乘法, 而存内计算电路实现指数位对应的加法和更新操作, 并通过底数的预归一化操作优化了存内计算与数字电路间的流水线时序. 该工作相比标准数字电路浮点运算单元能够节省 14.4% 的功耗和 11.7% 的面积. 由于该芯片的存内计算电路部分采用纯数字电路方式实现, 因此不会产生模拟误差导致的准确率下降问题.

2.2 SRAM 存内计算架构

2.2.1 存内计算位宽的影响因素

存内计算阵列的计算位宽是一个关键参数, 包含单次运算的输入位宽、存储位宽和输出位宽. 其中, 单次输入位宽由 DAC 控制, 单次输出位宽由 ADC 控制, 而单次运算的存储位宽取决于存算阵列. 对于 SRAM 器件, 一个 6T 或 8T 存储单元保存 1 bit 数据, 实现多比特存储需要多个 6T 或 8T 的存储单元. 若多个存储单元同时打开计算, 则可实现多比特计算.

早期阵列结构设计包括二值的存内计算阵列 [12], 其输入、存储和输出均为 1 bit, 对于 MNIST 等简单任务可实现较高能量效率, 但对于需要多比特计算实现高准确率的较复杂任务如 Cifar-10, ImageNet 等, 无法较好支持. 随着相关研究深入, 多比特的存内计算电路逐渐增多 [32,33,48]. 阵列内部的多比特存储并行计算有多种实现形式. 一种是对高位和低位采用不同宽长比的晶体管 (或电容), 从而产生 2

的幂次倍数的电流 (或电荷); 另一种是采用相同的晶体管, 在结果采样电路中采用 2 的幂次倍数的电容, 用于给不同比特位置的结果设置不同的权值. 这两种方式也可以以一定形式混合使用, 文献 [32] 采用一种孪生 8T (twin-8T, T8T) 的结构, 在单元电路部分用 2 个存储单元实现 2 bit 计算, 同时结合输出采样电路中 2 的幂次的电容设计, 实现了 5 bit 权重数据的存内计算操作.

通过外围数字电路的支持可以灵活实现不同比特位宽 [34, 49]. 即使存内计算阵列仅支持单次 1 bit 输入和 1 bit 存储数据的计算, 也可通过后续数字电路的多周期累加运算实现多比特存内计算. 对于 M bit 的输入数据和 N bit 的权重数据, 需要 $M \times N$ 个周期的存内计算和累加操作才能得到完整的多比特矩阵向量乘法结果. 该方法的优势在于可灵活支持不同的比特位宽, 且灵活支持负数权重的符号位计算.

存内计算阵列输出位宽的选取也非常关键. 若单次计算的输入位宽为 m bit, 存储位宽为 n bit, 打开的行数为 P, 则计算结果包含 $((2^m-1)\times(2^n-1)\times P)$ 种状态, 为实现准确采样, ADC 至少需要 $(\log_2((2^m-1)\times(2^n-1)\times P))$ bit. 考虑 ADC 的误差和噪声影响, 所需采样精度会更高. 高精度 ADC 会带来严重的面积和功耗开销, 为避免以上问题, 需要选择合适的输出比例 (output ratio) [34], 即 Q bit ADC 输出的状态数 (2^Q) 与理想计算结果包含的所有状态数之间的比例. 输出比例为 1 时能实现较高的准确率, 随着输出比例提高, 存内计算阵列的功耗降低, 能效增加, 但准确率逐渐下降. 一般而言, 为了避免过大的 ADC 功耗/面积开销, 现有绝大多数存内计算芯片的输出比例均小于或等于 1.

2.2.2 存内计算阵列的并行度

存内计算芯片由于所采用的器件、电路结构等因素不同, 其并行度存在较大差异. 神经网络算法容忍误差的鲁棒性、实际应用对于准确率的需求等, 也会影响芯片并行度的设计选择. 对于行 (输入) 方向的并行度, 并行打开的行数增多会影响计算准确率. 一方面计算单元的误差会累积在最终的模拟结果 (电压/电流) 中, 导致 ADC 无法准确区分, 另一方面并行度提升也会导致累加的模拟电流/电压范围增大, 对结果采样电路的线性度设计产生较大压力. 对于列方向的并行度, 需要考虑实际物理版图中 ADC 单元与存储阵列的面积匹配, 高分辨率/高速 ADC 所需的面积较大, 对每一列计算单元都匹配一个 ADC 会产生较大的面积开销. 多个 ADC 之间可能存在的串扰问题也是列并行度的考虑因素.

对于 SRAM 器件而言, 电流式存内计算阵列实现了 8/16/32 行等典型的行并行度 [32, 34, 48], 部分对 ADC 精度或网络准确率要求较低的阵列可以并行打开 64 行及以上 [12]. 电容式存内计算实现了 32/64/1152 行等典型的行并行度 [35, 36, 39], 其并行度相比电流式存内计算的提升有一部分来源于单元电路误差的降低和结果采样电路线性度的改善. 全数字存内计算由于无准确率损失, 其并行度主要考虑面积、功耗和性能的平衡, 相关工作的并行度包括 1/16/256 行等 [40~42].

2.2.3 存内计算阵列的映射与数据复用

在面向神经网络算法的专用硬件架构中, 数据复用是一类有效的低功耗设计方法 [50]. 存内计算架构将权重保持在存储器中固定不变, 本身即是一种 “权重固定型” 的数据复用方法. 文献 [36, 48] 考虑了面向输入激活值、存储权重和输出累加结果的多种数据复用策略.

一般而言, 存内计算阵列可以通过直接映射支持矩阵向量乘法, 其中矩阵数据存储在二维的存储阵列中, 而一维向量作为存内计算阵列的输入数据. 对于神经网络中的全连接层运算, 其主要操作为矩阵向量乘法, 当矩阵大小超过存内计算阵列尺寸时, 一般将其切分为不同的子矩阵映射到存内计算阵列中. 对于卷积等广义矩阵向量乘法操作, 一般展开为普通矩阵向量乘法后, 再将其权重映射到存

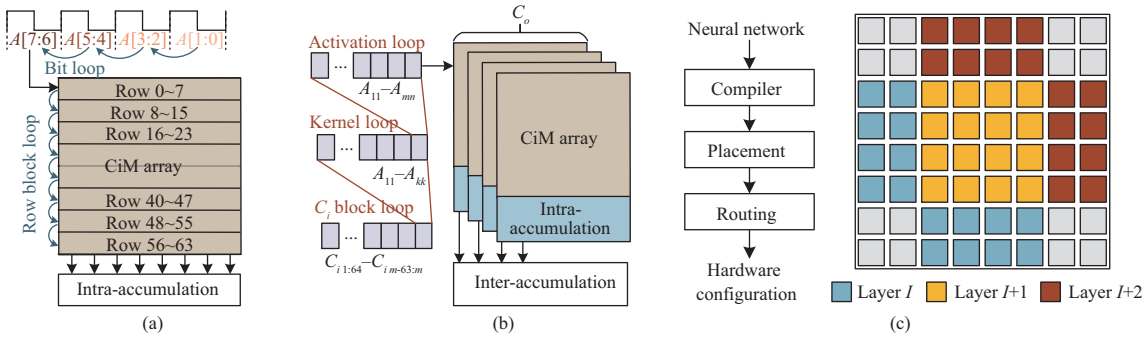

图 5　(网络版彩图) 存内计算数据复用与网络映射. (a) 单阵列与 (b) 多阵列数据复用 [48]; (c) 多层网络映射与并行计算 [36]

Figure 5　(Color online) The data-reuse and network mapping strategies on the CiM architecture. (a) Single CiM array's data-reuse; (b) multiple CiM arrays' data-reuse; and (c) the multi-layer mapping and parallel computing

内计算阵列中 [51].

由于神经网络模型各层的参数规模不同, 在存内计算阵列上的映射和计算效率也会产生差异. 通常, 卷积神经网络的第 1 层通道数较少, 例如 ImageNet 数据集 [52] 的 VGG16 网络 [53] 的第 1 层仅有 3 个输入通道, 卷积核为 3×3, 每一个输出像素位置对应的乘加操作为 27 次, 通常无法占满一个存内计算阵列的所有行, 利用率较低. 文献 [48] 将 64 行的存内计算阵列分为两部分, 分别存储行并行度较小的两块权重, 实现了 2 倍的利用效率.

在卷积神经网络中, 4 维 $[K,K,C_{\rm in},C_{\rm out}]$ 的权重数据会映射到存内计算阵列. 输出通道 $C_{\rm out}$ 对应的维度会被映射到不同的列, 而卷积核尺寸 K 与输入通道 $C_{\rm in}$ 对应的维度会被映射到不同的行, 以卷积核顺序或者以输入通道顺序来映射, 可以区分为通道优先和卷积核优先的两种映射方式 [48]. 由于输入通道维度的数值较为规整, 通常为 2 的幂次或 2 的幂次的整数倍, 如 64/128/512 等, 能够较好地与 2 的幂次的存储阵列结构相匹配, 因此大多数网络层采用通道优先的映射方式实现较高的利用率. 对于输入通道不是 2 的幂次的情况, 如第 1 层卷积, 也可以采用沿卷积核映射的方式, 且能够利用卷积操作在输入数据之间的复用特性来减少外部存储访问. 文献 [48] 通过可调的卷积核映射/通道映射方式, 在 VGG16 网络的多个网络层上实现了 2.8%~11% 的利用率提升.

除了存内计算本身的权重数据复用, 输入数据复用和输出结果原位累加也被用于降低功耗 [48]. 如图 5(a) 所示, 对于每一个存内计算阵列, 其输出结果往往需要多次的原位累加, 例如对于不同比特位数据的存内计算结果, 以及对于不同位置的存储区域的存内计算结果进行累加. 如图 5(b) 所示, 对于多个存内计算阵列共同执行一个复杂的矩阵向量操作时, 根据其行列并行度不同, 可以选择 (1) 对多个存内计算阵列复用输入数据, 产生不同的计算结果; (2) 对多个存内计算阵列提供不同的输入数据, 其输出结果对应位置相同, 进行原位累加. 两种方法可分别降低对于输入数据和输出数据的存储访问. 文献 [48] 的原位累加技术相比完全不采用数据复用的情况实现了 7.9 倍的存储访问下降, 采用多个阵列间的数据复用技术进一步实现了 17.6 倍的存储访问下降.

多个存内计算阵列不仅可用于计算同一层神经网络任务, 也可用于同时计算多层神经网络任务, 如图 5(c) 所示. 针对不同的算法, 在编译器层次需要设计合理的调度策略, 而在电路层次需要设计高速低功耗的互联拓扑和片上网络. 文献 [36] 实现了 16 个高性能高能效的存内计算阵列, 以 4×4 的结构排列, 每一组 2×2 的 4 个阵列通过高效的片上网络传递数据. 该片上网络能重构地实现多个阵列之间同时的双向数据传输, 进而支持多种多样的多层神经网络映射与调度策略. 文献 [36] 通过增加额

外的单指令多数据流 (SIMD) 计算阵列, 实现了对非矩阵向量乘法操作的支持. 此外, 部分工作 [54] 将存内计算阵列与中央处理器 CPU 等模块集成在一颗芯片中, 其中 CPU 可通过标准的处理器 – 存储器总线协议访问存内计算阵列.

2.2.4 存储容量与更新策略

受限于 SRAM 单元的存储密度, 尽管 SRAM 存内计算单阵列容量已从 4 kbit [12] 提升到 384 kbit [39], 多阵列容量提升到 4.5 Mbit [36], 但相比现有神经网络模型参数量, SRAM 存内计算阵列的存储容量仍然较小.

提升 SRAM 存内计算阵列的容量有多种技术. 对于单个计算阵列, 将存内计算阵列的容量提升与计算并行度提升解耦. 除了对于输入向量的字线译码和输出结果的位线译码电路进行多选一之外, 如文献 [34,39], 在存内计算单元中实现了多选一. 通过分组 6T 的单元结构, 每次仅激活 16/32 个 6T 存储单元中的一个, 从而在不提升计算并行度的情况下提升容量. 对于多个计算阵列, 采用合适的拓扑结构实现多核心并行计算, 从而进一步提升存内计算阵列容量. 文献 [36,48,49] 分别实现了 4 核心和 16 核心的多阵列存内计算芯片.

现有 SRAM 存内计算阵列无法一次性存储大型神经网络模型, 而是需要在一部分计算操作完成后, 重新写入新的权重数据. 然而写入权重的过程无法进行存内计算操作, 这导致实际系统效率下降. 文献 [49] 提出了并行的存内计算/更新技术, 其整体单元电路包含存储区域 A、存储区域 B 和存内计算单元 (local computing cell, LCC). 相比传统存内计算单元电路, 其主要区别在于多出一块额外的存储区域 B 和一对位线 B_GBL/B_GBLB. 存储区域 A/B 结构对称, 在其中一块存储区域通过 LCC 进行计算操作的同时, 另一块存储区域可通过 B_GBL/B_BGLB 进行权重更新操作. 当存内计算和权重更新操作完成后, 这两部分存储单元的功能可以立即互换, 实现存内计算操作的高效执行, 避免权重写入操作造成存内计算操作的阻塞. 该技术 [49] 在 Cifar-10 数据集、ResNet18 模型的不同网络层上实现了最高 94% 的性能提升, 整体性能提升效果为 26%.

2.3 SRAM 存内计算算法/电路联合优化

2.3.1 稀疏存内计算

稀疏压缩技术 [55~58] 是传统数字神经网络芯片中常用的优化技术, 通过增加网络中的零值数据降低计算量和存储量, 但存内计算的并行计算特性与规则的阵列结构使得存内计算难以和稀疏技术很好地兼容. 稀疏压缩技术按照权重稀疏模式的粒度可分为细粒度单值稀疏、向量稀疏、卷积核稀疏和通道稀疏 [58]. 细粒度稀疏的压缩率较高, 常用于可灵活设计的数字专用芯片. 而在存内计算阵列中, 由于并行计算的特性, 离散分布的零值权重无法节省计算时间和功耗, 只要存在一个非零数据, 就需要打开整行整列进行运算. 如果采用通道稀疏等粗粒度稀疏模式, 网络压缩率较低, 提升较小.

基于存内计算阵列的特点, 文献 [48] 提出了分块结构化稀疏策略, 如图 6(a) 所示. 以存内计算的行并行度 N 作为稀疏粒度, 每 N 个数据被训练为全零或不全为零的块, 通过 1 bit 的索引标记每一个数据块是全零或不全为零的块. 对于全零的块, 可预先知道计算结果为零, 因此可直接关断 ADC 来节省功耗, 在 Cifar-10 数据集、VGG16 模型的不同卷积层实现了 2.4~13.6 倍的 ADC 功耗节省. 而对于稀疏激活值数据的稀疏, 文献 [48] 采用动态稀疏的策略, 通过统计输入数据中零值的数量, 将输入数据划分为稠密、稀疏和全零 3 种状态. 在稀疏的情况下, 打开两倍的行数同时进行计算, 实现两倍性能提升; 在全零的情况下, 直接跳过对应的计算操作, 缩短计算时间. 通过该方法在 Cifar-10 数据集、VGG16 模型的不同卷积层实现了 1.25~2.72 倍的性能提升.

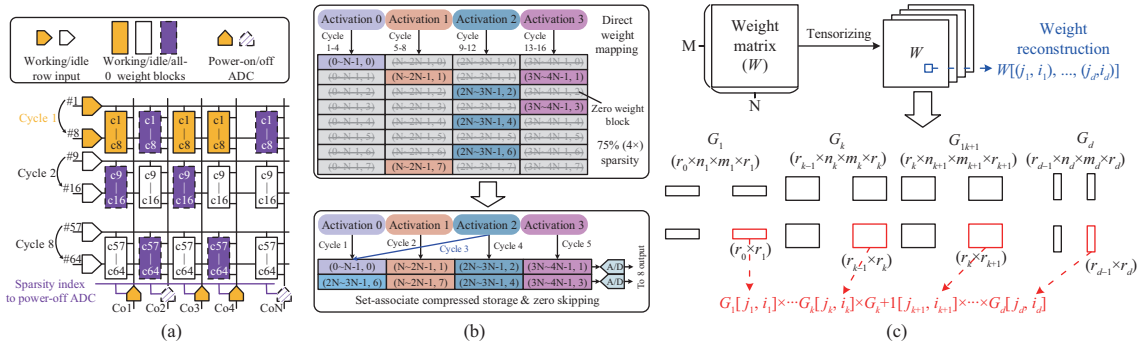

图 6　(网络版彩图) 算法/电路联合优化技术. (a) 分块结构化稀疏 [48]; (b) 分块组相联跳零稀疏 [49]; (c) 张量火车压缩技术 [59]

Figure 6　(Color online) The algorithm-circuit co-optimization. (a) The block-wise sparse CiM architecture; (b) the block-wise set-associate zero-skipping sparse CiM architecture; (c) the tensor-train compression technique

以上基于权重分块结构化稀疏的工作实现了功耗节省, 但无法节省零值权重块的存储空间, 也无法有效利用关断的 ADC. 文献 [49] 进一步提出了分块组相联稀疏跳零策略, 如图 6(b) 所示. 将 M 行 N 列的稀疏权重块作为一组, 控制每一组内的稀疏度保持一致, 并压缩存储在实际的存内计算阵列中. 每一个稀疏权重块通过 7 bit 的索引标记原始的行列位置以及是否为全零块等信息. 在进行存内计算时, 通过索引检测电路, 控制输入的位置与 ADC 的开关, 打开正确的行列位置进行计算, 并将计算结果送到正确的组相联累加寄存器中. 基于以上稀疏策略可同时实现存储压缩、功耗节省和性能提升.

分块结构化稀疏的策略能有效利用稀疏技术, 但需要一定的额外硬件开销. 文献 [59] 提出了比特层次的稀疏技术来节省功耗, 虽不能节省计算时间和存储空间, 但其不需要额外的硬件电路. 在电流式存内计算结构中, 对应 “1” 的权重会在存内计算电路中汇聚电流产生功耗, 因此降低权重中 “1” 的数量可明显降低存内计算阵列的功耗. 该工作通过算法训练在比特层次实现了更高的稀疏度, 从而有效降低了功耗, 该芯片的峰值能量效率可达 691.1 TOPS/W.

利用 ADC 采样结果的稀疏特性, 也可实现功耗节省. 文献 [46] 注意到大部分 ADC 的采样结果值较小, 因此可对 2 次或多次存内计算阵列的计算结果进行模拟域累加后, 再打开 ADC 进行一次采样操作, 通过节省 ADC 的采样次数降低功耗.

2.3.2　其他神经网络压缩算法

除了稀疏压缩技术之外, 其他压缩方法也被用于存内计算. 如图 6(c) 所示, 文献 [59] 注意到神经网络模型的权重规模过大无法一次性放在存内计算阵列中, 因此采用了计算换存储的策略. 该工作采用了张量火车 (tensor-train) 的压缩方法, 将原始的全连接层或卷积层转化为级联的多层小型矩阵乘法, 其网络模型可压缩 2.3∼4954 倍. 该工作通过计算重排和存内计算阵列匹配实现了多层级联的小规模张量乘法操作在存内计算阵列上的高效执行.

2.4　SRAM 存内计算的代表性芯片对比

为了更好地展示 SRAM 存内计算芯片当前的技术水平, 表 1 列出了目前在存储容量、能效、准确率等方面具有代表性的多款芯片的具体参数 [34, 36, 39, 41, 47, 49]. 电流式、电荷式和全数字等多种存内计算架构仍在不断发展, 各有其技术优势. 多款存内计算芯片 [36, 47, 49] 展示了基于存内计算阵列核心搭建片上系统, 从而实现完整神经网络运算的能力. 通过分组 6T 的存内计算单元电路, 目前已经可

表 1 SRAM 存内计算芯片对比

Table 1 Comparison of the state-of-the-art SRAM-based CiM chips

	Current-based		Charge-based		All-digital	
	ISSCC'20 [34]	ISSCC'21 [49]	ISSCC'21 [36]	ISSCC'21 [39]	ISSCC'21 [41]	VLSI'21 [47]
Technology	28 nm	65 nm	16 nm	28 nm	22 nm	28 nm
Area (mm^2)	0.017[a1)]	12[a3)]	25[a3)]	0.096[a1)]	0.202[a2)]	5.832[a3)]
Activation	4/8 bit	2/4/6/8 bit	1∼8 bit	4/8 bit	1∼8 bit	BFloat16
Weight	4/8 bit	1∼8 bit	1∼8 bit	4/8 bit	4/8/12/16 bit	BFloat16
ADC	5 bit	0/2/4 bit	8 bit	5 bit	–	–
Output ratio	1	1	1/4.5	1	1	1
CiM storage	64 kbit	64 kbit	4.5 Mbit	384 kbit	64 kbit	1.28 Mbit
Energy efficiency (TOPS/W)[b)]	68.4	46.3	121	94.3	89	13.7
Accuracy	92.02%[c1)]	66.13%[c2)]	73.04%[c2)]	–	–[c3)]	–[c3)]

a1) CiM bitcell area. a2) CiM macro area. a3) Chip area. b) CiM macro peak energy efficiency. Scaled to 4-bit activation, 4-bit weight, except for [47]. c1) On Cifar-10 dataset. c2) On ImageNet dataset. c3) The same as digital circuits.

以实现 384 kbit 容量的单个存内计算核心, 而多核心设计进一步将片上存内计算核心的存储容量提升到 4.5 Mbit. 存内计算阵列普遍能够支持 4/8 bit 或更加丰富的计算位宽, 并且在 4 bit 位宽下实现了 121 TOPS/W 的峰值能量效率, 在 16 位浮点运算格式 BFloat16 情况下实现了 13.7 TOPS/W 的峰值能量效率. 在采用较高的输出比例情况下, 基于电流式和电荷式架构的存内计算芯片实现了对 ImageNet 等较大规模数据集的较高准确率, 而全数字存内计算阵列则可以实现与传统数字电路完全等价的准确率. 需要说明的是, 尽管表 1 中将各款芯片的能量效率归一化到 4 bit 位宽, 但由于各款芯片的工艺节点、输出比例、选用测试样例稀疏度等多种因素存在差异, 直接以表 1 中数值进行对比并不绝对公平. 同样地, 表 1 中的准确率数据也受限于所采用的量化位宽/网络模型等因素. 因此该表仅作参考, 并不能直接评判各款芯片的能量效率优劣.

3 基于铁电晶体管的存内计算电路

3.1 铁电晶体管

本小节回顾铁电晶体管 FeFET 作为 NVM 器件的基本结构、工作原理、开关机制及其建模存储应用. 在晶体管结构中利用铁电性质材料以调整半导体沟道导电性能的方法技术最早可追溯到 20 世纪 50∼70 年代 [60], FeFET 结构上包含 MOSFET 晶体管及其栅极上叠加的铁电材料层. 由于铁电体固有的晶格极化, 可通过施加电场对其进行切换, 表现出极化特性、存储电荷对于施加电场强度的滞回曲线和非易失性存储特性. 铁电层的极化决定了 FeFET 器件的阈值电压, 因此其阈值电压可通过栅极的电压脉冲波形控制, 如图 7(a) 所示. 铁电层不同的极化程度使得 FeFET 的存储状态在多个对应阈值电压的开关曲线间转换, 因此其电压驱动的写入机制将 FeFET 使能为多值存储器件 (如图 7(c) [61]), 提升了基于 FeFET 的电路信息存储密度和能效. FeFET 通过给栅极施加低阈值电压和高阈值电压之间的电压值实现电流读取操作 (如图 7(b)). 相比于其他如 ReRAM 一类需要电流驱动写入机制因而写入能耗高的二端口 NVM 器件, FeFET 在存算电路设计方面展现出低功耗的优势. 建

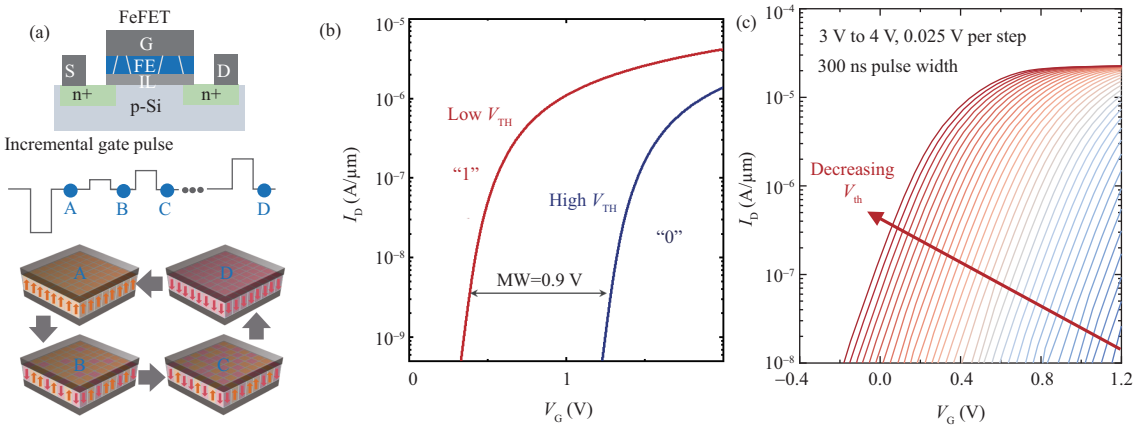

图 7　(网络版彩图) (a) FeFET 器件结构及其栅极写入策略; (b) FeFET 的二值存储特性; (c) FeFET 的多值存储 (multi-level cell, MLC) 特性 [61]

Figure 7　(Color online) (a) FeFET structure and associated gate write scheme; (b) binary cell characteristics of FeFET; (c) multi-level cell characteristics of FeFET

模方面, 考虑铁电层具有不同电荷 – 电场滞回曲线响应的多铁电畴结构以及铁电层下的 MOSFET 结构, 文献 [62] 建立了 FeFET Preisach 多铁电畴模型并与制备器件数据校准, 形成了完备可用于存内计算电路设计的 FeFET SPICE 模型. 近年来氧化铪基等与现代半导体制程工艺兼容, 体现出铁电特性的材料极大扩展了 FeFET 在先进半导体、集成电路设计中的应用 [63~69]. 此外, 氧化铪能够在很薄的厚度仍然保持其铁电性 [70], 这使得基于氧化铪的 FeFET 能集成在各种器件结构上面, 包括平面晶体管、鳍式场效应管 FinFET 和纳米管 nanowire, 极大地增大了 FeFET 的应用场景和器件密度. 基于氧化铪的 FeFET 由于其三端口结构、高开关电流比、编程可调的阈值电压、与 CMOS 工艺的兼容性及其单晶体管实现非易失存储逻辑功能的特性, 有望克服传统计算芯片系统的存储墙瓶颈, 在存算一体技术方面有着广阔的应用潜力, 为人工智能和物联网中的数据密集型任务应用场景从下而上提供现代 CMOS 计算平台更为高效的器件、电路硬件创新动力. 以下我们主要介绍存内计算单元及其交叉阵列、内容寻址存储器 (content addressable memory, CAM) 及其他形式功能的存内计算电路设计优化方法, 验证 FeFET 在未来以数据为中心的应用场景上可扮演的关键角色.

3.2　基于铁电晶体管的存内计算阵列

基于 FeFET 的存内计算交叉阵列根据 FeFET 存储状态个数可分为二值化存算单元阵列以及模拟/多值存算单元阵列. 二值化存算单元 [71~78] 利用 FeFET 存储 "0" 或 "1" 的状态以及单晶体管与门逻辑特性 (如图 7(b)), 在每个单元内将单比特的乘法结果输出, 然后通过位线将同一列上多个单元输出汇聚到 ADC 输出乘累加结果. 模拟/多值存算单元 [79~87] 利用 FeFET 编程可调的阈值电压特性对应的模拟多值存储状态 (如图 7(c)), 在每一个存算单元以输出电流的形式表示栅极电压输入和所存储的权值对应的电导或跨导的乘积, 通过位线累加同一列上多个单元电流并输出. 一般而言, 二值化存算单元的高低存储阈值电压间隔大, 输入电压及输出状态易区分, 结构简单; 模拟/多值存算单元随着存储状态数目增加, 感应容限减少, 需较多外围电路区分输出.

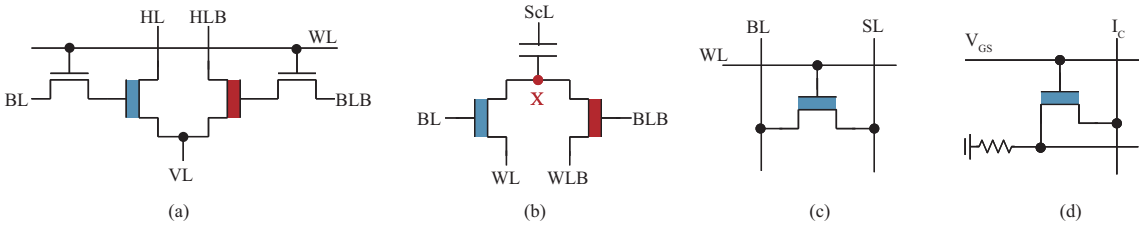

图 8 (网络版彩图) 二值化存算单元. (a) 2FeFET-2T 单元 [71]; (b) 2FeFET-1C 单元 [72]; (c) 1FeFET 单元 [73,74]; (d) 1FeFET-1R 单元 [75,76]

Figure 8 (Color online) Binary CiM cell designs. (a) 2FeFET-2T cell; (b) 2FeFET-1C cell; (c) 1FeFET cell; (d) 1FeFET-1R cell

3.2.1 二值化存内计算单元

FeFET 既可用作 NVM, 也可用作开关. 与 ReRAM 相比, 这是一个独特的优势, 因为三端子结构的读写分离路径可分别控制写入和读取功率. 文献 [71] 第 1 个针对二值卷积神经网络提出基于 FeFET 的交叉阵列电路设计, 其采用 2FeFET-2T 结构, 如图 8(a) 所示, 二值权重数据以高低阈值电压形式存储在 FeFET 中, 2T 作为存取晶体管避免读写干扰. 该单元结构实现了输入电压与存储权值的单比特同或逻辑, 同列各单元输出电流累加, 实现二值卷积操作, 提高了二值卷积神经网络推理效率, 相比两种基于 ReRAM 的交叉阵列 [71], 写入功率提高了 5600 倍和 395 倍, 读取功率提高了 4.1 倍和 3.1 倍. 文献 [72] 提出, 如图 8(b) 所示, 基于 2FeFET-1C 的单元结构, 以电荷分配机制实现输入电压与存储权重的同或逻辑运算, 由于电容器截断输出路径, 仅将输出电压传递至位线, 该结构面向二值神经网络的交叉阵列没有漏泄功率, 相比基于 SRAM 的电流模阵列 [38] 和电荷模阵列 [37], 分别降低了 60% 能耗和 47% 能耗. 根据图 7(b), 单个 FeFET 可实现输入电压和存储状态的与门逻辑, 其输出为电流. 相关工作 [73,74] 利用 FeFET 这一特性提出了数字内存矢量矩阵乘法阵列引擎, 每个存算单元如图 8(c) 所示为单 FeFET, 输出二值乘法电流, 并按列累加电流得到运算结果. 为抑制 FeFET 器件间导通电流和阈值电压差异, 进一步实现大规模并行矩阵向量乘法, 文献 [75,76] 均利用高阻值电阻器设计如图 8(d) 所示 1FeFET-1R 存算单元结构, 配合 ADC 电路实现低功耗、低面积的神经网络交叉阵列. 不同之处在于, 文献 [75] 输入为二值电压, 其矩阵向量乘法输出表示为电流, 而文献 [76] 则通过多比特外围电路将数字输入转换成脉冲宽度编码信号, 以基于电容充放电的电压摆幅生成差分信号作为输出. 两者均以高阻值电阻串联 FeFET 抑制 FeFET 导通电流差异性, 同时减少电流输出, 降低功耗, 分别实现 13714 和 13700 TOPS/W 的超高峰值性能. FeFET 也可用于三维结构及相关工艺以获得阵列超高信息密度, 如文献 [77] 提出了一种用于片上训练加速器的 8 bit 三维与非串联 (NAND) 结构. 由于 FeFET 器件的低运行电压和三维 NAND 结构的超高密度, 外围电路开销减少, 训练过程中可在芯片上存储和计算所有中间数据. 该工作提出了一种自定义设计的 108-Gb、59.91-mm^2 面积、45% 阵列效率的芯片, 能效达 7.76 TOPS/W.

3.2.2 模拟/多值存内计算单元

密集模拟/多值突触 (synapse) 交叉阵列可在存储阵列内执行原位向量矩阵乘法和权重更新以减轻数据移动, 在神经形态硬件加速器上有广阔应用. 然而, 许多模拟/多值权重存储单元具有长延迟或低动态范围的瓶颈, 限制了可达到的性能. FeFET 由于其电压驱动、低延迟的写入机制, 可通过如图 9(a) 和 (b) 所示的增量脉冲写入方式实现高线性度的大范围动态权重存储, 克服上述模拟/多值突

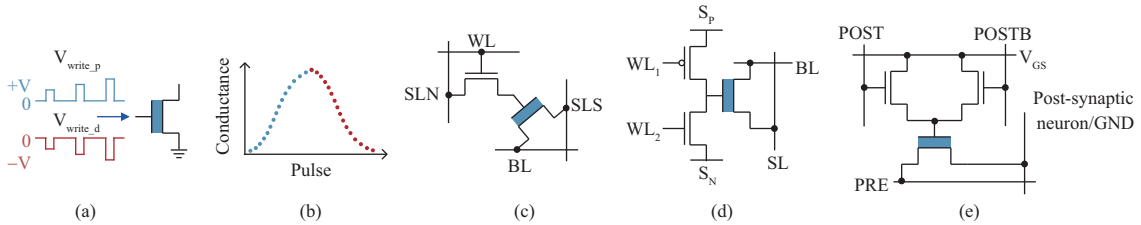

图 9 (网络版彩图) 模拟/多值存算单元. (a) FeFET 增量脉冲写入方案 [79]; (b) FeFET 存储状态对应的可调线性电导 [79]; (c) 1FeFET-1T 单元 [79,82~85]; (d) 一种 1FeFET-2T 单元 [80]; (e) 另一种 1FeFET-2T 单元 [81]

Figure 9 (Color online) Analog/multi-bit CiM design. (a) FeFET incremental amplitude pulse scheme; (b) tunable linear conductance states of FeFET; (c) 1FeFET-1T cell; (d) a 1FeFET-2T cell; (e) another 1FeFET-2T cell

触交叉阵列的困难. 多项工作围绕基于 FeFET 的模拟/多值突触单元设计, 从器件、电路、读写编码方法、架构算法等角度提出优化方法, 验证基于 FeFET 在神经网络加速等应用任务中的潜力. 文献 [79] 提出了如图 9(c) 所示的多比特 1FeFET-1T 突触单元结构, 展示了 FeFET 的压控偏极化切换动态, 验证了 5 bit 权值存储和基于 50 ns 低延迟编程脉冲的 67 倍电导可调范围, 并在交叉阵列中表现出不同电导的高度对称性, 极大扩展了多值 FeFET 在深度神经网络训练推理任务上的应用. 该单元构成的交叉阵列运行 MNIST 数据集准确率达 90%, 与 ReRAM 交叉阵列 [88] 相比将 1 M 图像训练速度提高了 1000 倍, 与 6 bit SRAM 存算阵列相比面积效率超过 10 倍. 上述 FeFET 的可编程动态电导调制范围通过增量脉冲幅度的写入方案实现, 然而该方法会增加交叉阵列写入电路复杂度. 文献 [80,81] 均提出 1FeFET-2T 突触单元结构 (如图 9(d) 和 (e) 所示), 及其支持 FeFET 可调电导特性的电路算法级优化方法. 文献 [80] 利用神经网络训练和推理的混合精度, 克服了非线性和非对称权值更新的挑战, 在算法级别实现了接近软件可比的训练精度. 文献 [81] 与之前增量脉冲写入方案不同, 采用了在单编程脉冲激励下利用 FeFET 器件本身固有的可塑性特性来进行学习算法公式化的替代途径, 在训练数据有限的边缘设备中实现无监督的本地学习.

通过对 1FeFET-1T 存算单元结构器件的替换或优化, 文献 [82,83] 分别首次展示了 Ge FE NWFET 和 FeMFET 两类器件在 1FeFET-1T 突触结构及其交叉阵列在神经网络硬件上的实验和仿真应用. 其中文献 [82] 应用在神经网络在线学习任务上的交叉阵列在训练超过 125 个 epoch 的 100 万 MNIST 图像后学习精度达 88%. 而文献 [83] 提出的 2 bit FeMFET 权重突触单元具有高线性度, 可将编程/擦除电压降低到 <1.8 V 的逻辑兼容水平, 写延迟改善到 <100 ns, 保持时间提高到 10^{10} 次以上, 有望在边缘设备实现低功耗、高性能推理任务. 文献 [78] 首次展示了单片三维后道工序集成的 22 nm, 2 bit 1FeFET-1T 突触单元结构及其交叉阵列, 实现了区域折叠, 占用更小的存储阵列面积, 相比 7 nm SRAM 的存内计算加速器在 CIFAR-10 图像数据集推理任务上实现了 3 倍能效提升. 另一工作 [84] 以器件材料为切入点研究了 HZO/β-Ga2O3 FeFET 在高温下的突触行为应用, 实现了在高温环境下 MNIST 数据集上运行两层多层感知器 (multi-layer perceptron) 网络, 片上学习准确度达 94%. 文献 [85] 利用 1FeFET-1T 突触结构实现了一种高效的用于直接反馈校准的深度神经网络训练加速器 PUFFIN, 克服了后向传播所要求的远程数据依赖的局限性. 除此之外, 该工作还提出了一种基于统计开关的 FeFET 器件随机数发生器和一种超低功耗基于 FeFET 器件的 ADC 电路. 与基于 ReRAM 的训练加速器 [51] 相比, PUFFIN 实现了 1.3 倍的加速和 2.5 倍的能效提升.

上述 1FeFET-1T 突触结构多利用 FeFET 可调动态电导特性, 运算结果多以电流信号表达, 由 ADC 感知. 除电流模感测方法外, 文献 [86] 提出一种基于多值 FeFET 的存算电路单元及其基于电荷

分配的感测方法. 不同于一般性 Id-Vg 测量, 该单元通过源极跟随器 (source follower, SF) 读取阈值电压的方式, 实现基于 FeFET 的 3 bit 权值读取. 基于 SF 的电压读取将神经网络输入和存储在 FeFET 阈值电压中的权值进行局部乘法, 再利用电荷共享检测方式累加乘积得到运算结果. 该设计具有超 10 年的数据保留, 支持 32×1024 的高并行 MAC 操作, 可实现 66 TOPS/W, 相比传统 32×16 并行度的电流检测方案 [89], 提升 64 倍能效. 另一工作 [87] 则在 1FeFET 单元构成的多比特存储阵列上提出一种并行积和运算方法, 运用时钟脉冲数表示输入与 FeFET 存储权值的乘积值并相加得到乘累加输出. 该方法比传统电流加法器 [90] 运算功耗降低 98%, 然而整体阵列的并行度被牺牲.

3.3 基于铁电晶体管的内容寻址存储器

新型 NVM 器件也可应用到高能效低面积的 CAM 电路架构中支持高速并行搜索. 基于 FeFET 的 CAM 单元可分为二值/三态 CAM (binary CAM, BCAM/ternary CAM, TCAM) 和模拟/多比特 CAM (analog/multi-bit CAM). BCAM/TCAM [91~95] 利用 FeFET 二值存储和开关特性, 对两个 FeFET 的存储状态进行编码实现比特存储, 并按位比较输入向量和存储向量得到单元匹配结果, 进而在匹配线上读出 CAM 阵列匹配结果. 模拟/多比特 CAM [61, 96~98] 利用 FeFET 可调阈值电压, 对输入电压与 FeFET 模拟/多值阈值电压的一对一匹配对应关系进行编码配置, 实现 FeFET 单元内存储的任意阈值电压范围都有其唯一匹配的输入电压范围. 由于 FeFET 的单晶体管存储开关特性, 二值/三态 CAM 设计匹配线电流电压感测容限大、易区分, 对感测放大器及外围电路要求不高; 模拟/多比特 CAM 随着存储状态数目增加, 信息密度增加、感测容限减少, 相应感测放大器及外围驱动电路随之变得复杂.

3.3.1 二值/三态 CAM

BCAM 存储和搜索值均只包含 “0”/“1” 两种状态. 一种经典基于 CMOS 的 10T BCAM 单元如图 10(a) 所示, 包含 1 个 SRAM 单元和 4 个比较晶体管, 仅当输入对应的搜索线电压和存储状态一致时才在匹配线生成输出; TCAM 则在二值状态之外加上一个通配符状态 “X”, (例如, 在路由器寻址场景下, IP 地址中被子网掩码掩去的部分), 存储 “X” 状态下的 TCAM 单元不论输入何值均给出匹配的结果. 由于需存储第 3 种状态, TCAM 单元通常较 BCAM 复杂, 图 10(b) 展示了一种常见的电荷式 16T CMOS TCAM, 包括两个 SRAM 单元和 4 个比较晶体管, 搜索结果由匹配线上的电荷表达, 当搜索状态与存储状态失配时, 匹配线与地之间的两条路径至少导通一条, 否则两条路径均关闭. 特别地, 两个 SRAM 单元存储 “0” 来表示通配符 “X”.

基于 FeFET 的 CAM 单元面积比基于 SRAM 的 CAM 小, 能效比 SRAM CAM 高. 例如, 文献 [91] 提出如图 10(c) 所示的 4T-2FeFET TCAM 设计, 该电路面积比 16T CMOS TCAM 减小 42%. 文献 [92] 进一步利用 FeFET 读写路径分离的三端结构和单晶体管与门逻辑, 设计了如图 10(d) 所示的 2FeFET TCAM 单元, 面积仅为 16T CMOS TCAM 的 13%, 阵列搜索速度比 16T CMOS TCAM 快 1.7 倍, 能耗降低 58%. 同样基于 2FeFET 单元结构, 文献 [93] 提出通过检测单元电路匹配线的电流大小而非电荷量判断匹配/失配状态, 如图 10(e) 所示, 该设计实现了 0.54 V 存储窗、~5 V/50 ns 写入波形幅度/宽度, 然而其耐久度仅为 $10^2 \sim 10^3$ 个周期. 文献 [94] 对文献 [92] 的 TCAM 设计流片验证表明, 耐久度超过 10^5 个周期. 2FeFET TCAM 单元在面积方面相较于基于 SRAM 的 CAM 取得较大优势, 面向边缘终端、自供能终端等场景, 在能耗方面仍有提升空间. 为进一步提高能效, 文献 [95] 设计了如图 10(f) 所示的 2FeFET-1T TCAM 单元及阵列, 采用与文献 [93] 类似的 2FeFET 结构, 仅以单个晶体管连接匹配线, 从而减少了匹配线预充电电容大小, 提高了能效; 此外, 文献 [95] 进一步提出一种如图 10(g) 所示的 2FeFET-2T TCAM 单元设计, 利用与非型 (NAND-type) 匹配线连接方式取

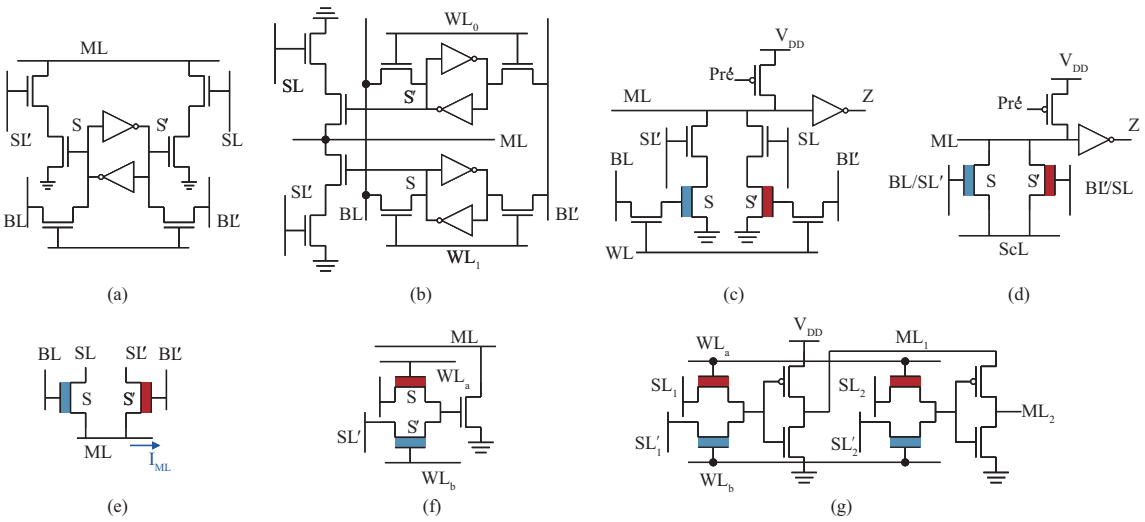

图 10 (网络版彩图) 基于 FeFET 的或非型 (a)∼(d)、(f) 电荷式 [61,91,92,95] 和 (e) 电流式 [93] TCAM 单元, 以及 (g) 与非型电荷式 TCAM 单元 [95]

Figure 10 (Color online) FeFET based NOR-type (a)∼(d), (f) charge domain and (e) current domain TCAM cells, and (g) NAND-type charge domain TCAM cell

表 2 各种 TCAM 设计的指标对比 a)

Table 2 Metric comparison of TCAM designs

Transistors per cell	Matchline type	Cell size (μm^2)	Search style	Search delay (ps)	Search energy (fJ/bit)
16T [99]	NOR	1.12	P	582	0.59
2FeFET-4T [91]	NOR	0.65	P	1022	0.455
2FeFET-2T [95]	NAND	0.44	PF	1430	0.073
2FeFET-1T [95]	NOR	0.36	P	253	0.195
2FeFET [92]	NOR	0.15	P	341	0.35

a) P = 需要预充电, PF = 无需预充电.

消了或非型 (NOR-type) CAM 阵列搜索操作的预充电环节, 大幅减少了充电能耗. 表 2 [99] 总结了上述各项 CAM 设计的性能数据对比.

3.3.2 模拟/多比特 CAM

前述 BCAM/TCAM 通常只利用 FeEFT 器件的高、低阈值电压, 每单元只存储 1 bit 数据. 事实上, FeEFT 的可调阈值电压对应多值存储特性 (MLC), 对 FeFET 的多值存储状态和输入电压对应关系编码, 并辅以驱动、感测电路设计, CAM 电路便可以每单元多比特的形式存储和搜索数据形成模拟/多值 CAM 设计, 大幅提高信息密度.

早在 20 世纪末, 文献 [96] 就提出利用金属 – 铁电 – 半导体 FET (MFSFET, FeFET 的早期形式之一) 的多级特性, 将搜索和存储数据之间的比较转化为二者数位 (digit) 间与 (AND)、或 (OR) 操作的函数, 并结合与非型 (NAND-type) 和或非型 (NOR-type) 匹配线结构各自特点设计了一种混合型 CAM 阵列结构, 如图 11(a) 所示, 实现了每单元 2 bit 数据存储和搜索. 随着铪基 FeFET 与先进高 K 金属栅极集成工艺的进步, FeFET 与 CMOS 工艺的兼容性和可扩展性进一步使 FeFET 多值存储的特性在集成电路领域得到运用. 如图 11(b) 所示, 文献 [92] 提出 FeCAM, 在基于图 10(d) 所示的 2FeFET

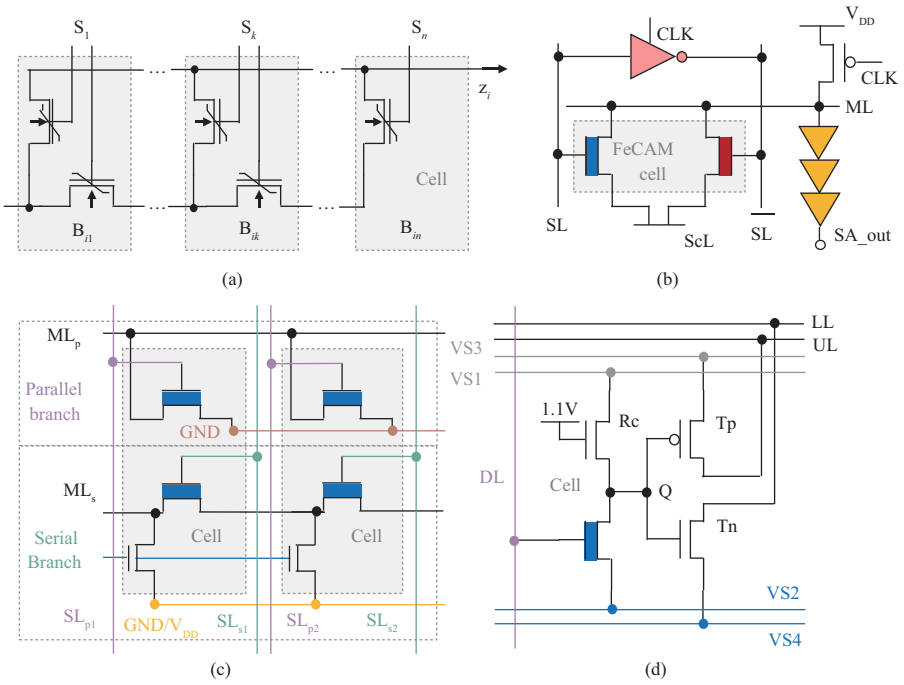

图 11　(网络版彩图) 模拟/多比特 FeFET CAM 设计. (a) 2-MFSFET 2 bit CAM 单元设计与互连方式 [96]; (b) 2-FeFET 3 bit CAM 单元与部分外围电路设计 [61]; (c) 2FeFET-1T 2-bit CAM 单元与阵列结构 [97]; (d) 1FeFET-3T 3 bit CAM 单元设计 [98]

Figure 11　(Color online) Analog/multi-bit FeFET CAM designs. (a) 2-MFSFET 2-bit CAM cell and array design, (b) 2-FeFET 3-bit cell and some peripheral circuit components, (c) 2FeFET-1T 2-bit CAM cell and array design, and (d) 1FeFET-3T 3-bit CAM cell structure

TCAM 基础上输入输出端分别引入反相器和阵列感测放大器形成模拟 CAM 单元电路, 输入电压只有落在 FeCAM 单元中两个 FeFET 存储的阈值电压之间才构成匹配条件, 从而形成模拟连续匹配范围. 该设计在 FeFET 可调阈值电压范围内可分隔出 8 个互不重叠的匹配范围, 支持每单元 3 bit 数据的存储和搜索, 每比特面积占用仅为文献 [92] 中 2FeFET CAM 的 1/3, 16T CMOS TCAM 的 4.5%, 每比特搜索能耗优于 16T CMOS TCAM 8.6 倍. FeCAM 因需要具有高线性度转移特性的模拟反相器, 电路设计成本增加. 而文献 [97] 提出基于 2FeFET-1T 的 CAM 电路单元, 同时引入如图 11(c) 所示的与非型和或非型匹配线结构, 通过编码 FeFET 的多值阈值电压及其对应搜索电压, 可实现 2 bit 数据存储搜索功能, 相比基于 SRAM 的 CAM, 单元面积优化 22.6 倍. 该设计去掉了模拟反相器, 编码方式更加简洁, 易于实际电路调节和拓展. 最后, 为减少 FeFET 在 CAM 单元中的数目以减轻设计负担, 文献 [98] 提出了一种 1FeFET-3T 的 CAM 设计, 可存储搜索 3 bit 数据, 如图 11(d) 所示. 该设计延时较低, 具有较好的抗噪能力.

3.3.3　基于 CAM 的数据密集型应用

上述 CAM 设计不仅利用 FeFET 特性提升了硬件信息密度和能效, 还运用 CAM 高并行度的特点进一步将能效性能优势推进到算法应用层面. 以下概括了基于 FeFET 的 CAM 设计在数据密集型应用中取得的硬件能效、性能、面积效率提升. 例如, 基于 FeFET 的 TCAM [92] 被用于加速基因测序应用 [100], 相比 CPU 的基因片段测序加速 400 多倍. 3 bit 的 FeCAM [61] 在路由器相同 IP 地址查找

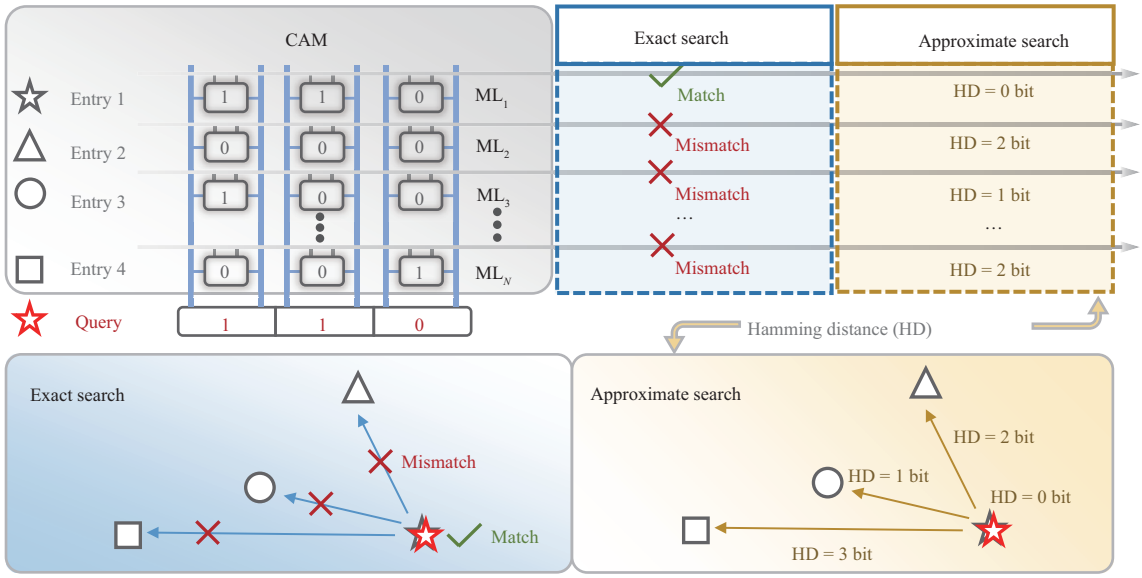

图 12 (网络版彩图) 基于 CAM 的精确搜索与近似搜索

Figure 12 (Color online) CAM based exact matching search and approximate matching search

表通量条件下, 相比基于 CMOS 的 TCAM 面积成本和能耗分别减小 60.5 倍和 23.1 倍. 文献 [97] 提出的 3 bit CAM 设计作为搜索核应用在数据库查询任务中, 相比基于 CMOS 的 TCAM 实现超过 16 倍加速和 29 倍的能量延时积提升, 相比文献 [92] 中的 2FeFET TCAM 在搜索速度和能量延时积上也分别提升 1.6 倍和 2.7 倍. CAM 也被作为联合存储器 (associative memory, AM) 用于加速 GPU 计算 [101,102], 通过存储数据集中被频繁计算的操作数及其相应运算结果减少 GPU 中浮点运算单元的运行, 从而减少 GPU 架构整体能耗. 文献 [95] 中提出的两种 TCAM 设计在加速 GPU 的应用上可分别降低 45.2% 和 51.5% GPU 能耗.

前述 CAM 设计支持精确搜索操作, 即输出与输入向量完全匹配方可生成匹配结果. 然而, 随着数据密集型应用场景的爆发式增长, 支持精确搜索但规模增长缓慢的 CAM 阵列无法满足大数据的匹配需求, 反而由于 CAM 消耗额外能耗导致硬件能效不再提升. 因此, 一系列工作 [94,103] 通过放松输入向量与 CAM 存储向量间的匹配程度约束, 使得存储向量与输入向量间有少量不匹配比特也被视作匹配. 这一近似搜索形式以牺牲可接受的搜索准确率增加 CAM 的匹配命中率, 从而继续提升系统能效性能, 如图 12 所示. 在 2FeFET TCAM 基础上, 文献 [94] 提出通过检测 TCAM 阵列匹配线电压下降速率定量分析搜索输入与存储向量间的匹配度, 在面向少样本学习应用的记忆增强神经网络架构 (memory augmented neural networks, MANNs) 增强存储模块实现基于汉明码距离的近似搜索, 相比传统基于 GPU 的 MANN 架构, 能效提高 50 倍, 延时缩减 2700 倍. 进一步地, 文献 [103] 提出的 2FeFET 多比特 CAM 设计支持多比特向量的存储和最近邻搜索, 相比仅支持二值向量搜索的近似搜索 TCAM 设计, 在面向少样本学习的推理任务上准确率提高 12%.

3.4 其他形式/功能存内计算电路

除了支持神经网络中的并行乘累加及 CAM 中的并行搜索操作, 基于 FeFET 的存储阵列也被应用于支持其他的关键计算功能. 根据 FeFET 存储阵列的读写访问模式, 存在一类运算时激活一到多行阵列、按位进行通用逻辑运算的存内逻辑阵列设计, 支持存储访问频繁的数据密集型应用如数据库、

数据加密等. 文献 [17] 首次提出基于 FeFET 存储单元的通用存内逻辑运算阵列架构, 可支持内存随机访问、布尔逻辑 (与非/与、或非/或、异或/同或、反相等) 及加法操作, 利用 FeFET 三端结构及高开关电流比实现了低面积、低功耗存内逻辑运算. 在此基础上, 基于 FeFET 的存内逻辑阵列功能不断扩展, 涵盖了数据回写与复制 [104, 105]、查找表 [106, 107] 以及存内逻辑、二值卷积和内容寻址存储三合一多功能 [108] 等, 并被应用到多个数据密集型应用中, 如同态数据加密 [105]、少样本学习和元学习 [109, 110] 等. 文献 [111] 通过架构系统的基准测试展示了基于 FeFET 的存内逻辑阵列相比基于其他 NVM 或 SRAM 的存内逻辑阵列的性能能效优势.

除了以基于 FeFET 的存算单元构成阵列来支持并行逻辑运算之外, FeFET 器件本身便可实现单晶体管与门/与非门逻辑功能, 配合 FeFET 的非易失存储特性, 使得 FeFET 可被应用到一系列关键逻辑门设计中, 极大扩展了 FeFET 在物联网的应用场景如自供能能量采集装置、硬件安全设备中的应用前景 [112, 113]. 图 13 介绍了基于 FeFET 的布尔逻辑门存内计算电路. 文献 [114] 首次提出一系列基于 FeFET 存算特性的逻辑门电路, 包括两种实现基本布尔逻辑运算 (与非/与、或非/或、反相等) 的非易失存内逻辑门电路通用设计形式. 文献 [102, 115] 在此基础上进一步扩展, 利用 FeFET 构建了非易失低功耗累加器和乘法器. 文献 [116] 则利用两个 FeFET 器件提出了紧凑的 4 输入逻辑门电路设计方法, 可实现 2 存储值与 2 输入值之间的析取逻辑表达式, 并展示了异或/同或的逻辑功能. 文献 [117, 118] 则利用 FeFET 的可调阈值电压特性提出了可重构的单 FeFET 与非/或非逻辑门. 该逻辑门以 FeFET 存储的 “0” 或者 “1” 作为一个输入, 栅极电压作为第 2 个输入实现存内逻辑运算, 并通过对 FeFET 施加特定源极/衬底偏置电压或写入脉冲, 演示了所提出逻辑门在与非和或非逻辑功能之间的可重构性. FeFET 也被应用于高能效小面积的非易失时序电路设计上. 如图 14 所示, 一系列工作 [119~123] 聚焦于基于 FeFET 的锁存器和触发器设计, 构成了基于 FeFET 的时序电路基础, 也与上述非易失组合逻辑门电路一起构建了面向物联网边缘端芯片的电路单元库.

4 总结与展望

本文在之前存算一体技术的综述文章基础上, 以 SRAM 和 FeFET 作为成熟流片工艺和后摩尔新型工艺器件的代表, 主要介绍了存算一体关键功能电路及其与架构、应用等层面协同优化的一系列工作. 针对 SRAM 存算一体技术, 本文介绍了电流式/电荷式/全数字的 SRAM 存内计算单元与阵列结构, 分析了位宽、计算并行度、映射方式、存储容量等设计因素, 并介绍了浮点存内计算、数据复用、稀疏/张量火车压缩等设计优化方法. 针对 FeFET 存算一体技术, 本文以基于 FeFET 的存内计算突触单元及其交叉阵列结构、基于 FeFET 的 CAM 单元及其阵列结构, 以及基于 FeFET 的存内逻辑门电路等作为 FeFET 存内计算的代表应用, 介绍了新型 NVM 器件与存算一体电路间协同设计优化的多种实现方式. 无论是现阶段较成熟可商业化的 SRAM, 还是后摩尔时代新型非易失存储器件的代表 FeFET, 两者与存算一体技术的结合仍然各自有其继续提升的空间和潜力. 在 SRAM 成熟流片工艺方面, 如何进一步扩大存算阵列规模以更好映射规模越来越大、连接越来越复杂的神经网络模型, 以及优化存算一体加速器芯片中的混合模拟接口以保持存算一体技术的增益, 仍然是一个不小的挑战; 面向存内计算的器件/电路/架构/算法联合优化越来越重要, 但综合考虑准确率、面积、能量效率和通用性的设计范式尚未建立; 以 FeFET 为代表的新型 NVM 器件在器件工艺、电路架构层面仍然需要更多的优化, 如 FeFET 多值模拟状态的线性、可编程性, 以及器件本身极化状态的耐久性等. 无论是 SRAM 等成熟器件还是 FeFET 等新型非易失存储器件, 其存内计算操作仍需要存算阵列之外的输入/输出数据存储, 尚不具备将所有数据存储和计算操作在同一个存储器内完成的能力, 仍面临一定的

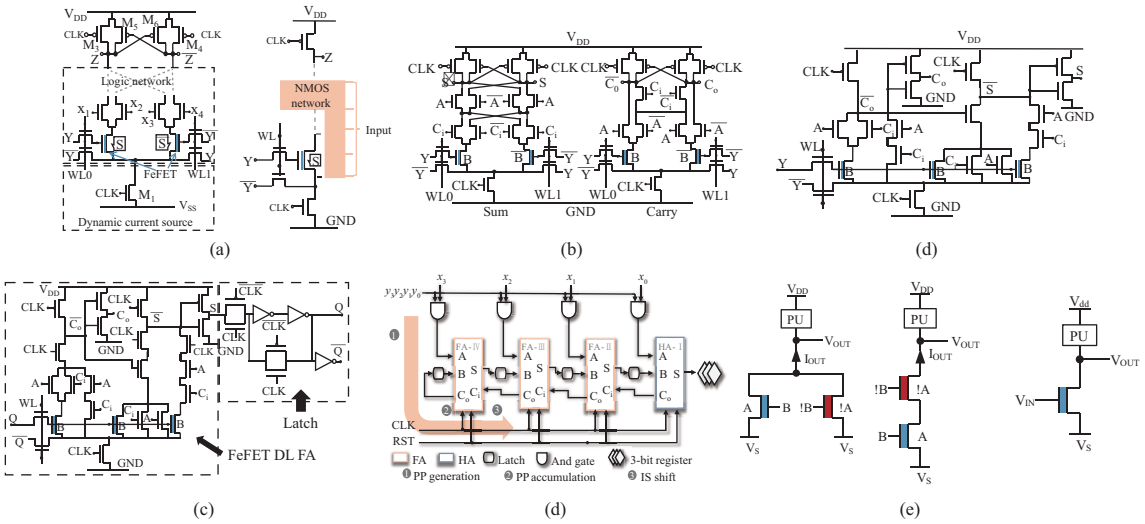

图 13 (网络版彩图) 基于 **FeFET** 的布尔逻辑门电路. **(a)** 两种实现基本布尔逻辑运算的非易失存内逻辑门电路通用形式 [114]; **(b)** 动态逻辑 (**dynamic logic, DL**) **1 bit** 全加器电路 [102]; **(c)** 动态电流模逻辑 (**dynamic current mode logic, DyCML**) **1** 位全加器电路 [102]; **(d)** 累加器电路 [102]; **(e)** 乘法器 [115]; **(f) 4** 输入可重构异或和同或逻辑门电路 [116]; **(g)** 可重构逻辑与非/或非门结构 [117, 118]

Figure 13 (Color online) FeFET based Boolean logic-in-memory (LiM) gate circuits. (a) Two generic non-volatile LiM design styles for basic Boolean logic; (b) dynamic logic 1-bit full adder; (c) dynamic current mode logic full adder; (d) accumulator; (e) multiplier; (f) 4-input reconfigurable gates for XOR/XNOR; (g) reconfigurable NAND/NOR gates

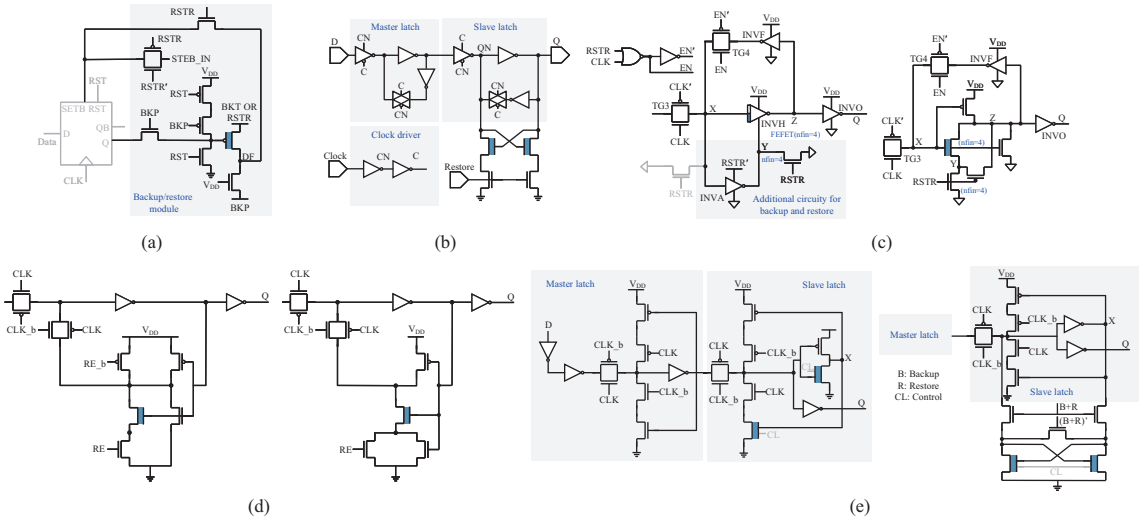

图 14 (网络版彩图) 基于 **FeFET** 的存储逻辑电路. **(a)** 一种非易失触发器电路结构 [119]; **(b)** 一种非易失 **D** 类触发器电路结构 [120]; **(c)** 两种非易失触发器 (**NVFF-1** 和 **NVFF-2**) 的从级锁存电路结构 [121]; **(d)** 另外两种非易失触发器 (**FeFET-out NVFF** 和 **FeFET-in NVFF**) 的从级锁存电路结构 [122]; **(e)** 两种基于双模态 **FeFET** 的非易失触发器电路结构 (**DNVFF-1** 和 **DNVFF-2**) [123]

Figure 14 (Color online) FeFET based storage gates. (a) An NV flip flop (FF) circuit structure; (b) an NV DFF circuit structure; (c) slave latch circuit structures of two NV FF (NVFF-1 and NVFF-2); (d) slave latch circuit structures of another two NV FF (FeFET-out NVFF and FeFET-in NVFF); (e) two NV FF circuit structures (DNVFF-1 and DNVFF-2) based on dual mode FeFET

数据传输瓶颈. 本文回顾论述的基于 SRAM 和 FeFET 的存算一体电路等工作部分解决了上述挑战问题, 因此, 我们相信随着工艺集成、器件、电路、架构算法的跨层次协同研究发展, 存算一体技术终将在边缘端、桌面端和服务器端, 覆盖从超低功耗到高性能的多样化应用需求, 全面支持人工智能物联网时代的高能效高性能智能计算系统.

参考文献

1 Wulf W A, McKee S A. Hitting the memory wall: implications of the obvious. SIGARCH Comput Archit News, 1995, 23: 20–24
2 Deng Z, Xu C, Cai Q, et al. Reduced-Precision Memory Value Approximation for Deep Learning. Hewlett Packard Labs Technical Report HPL-2015-100. 2015
3 Du Z, Fasthuber R, Chen T, et al. ShiDianNao: shifting vision processing closer to the sensor. In: Proceedings of the 42nd Annual International Symposium on Computer Architecture, 2015. 92–104
4 Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit. In: Proceedings of the 44th Annual International Symposium on Computer Architecture, 2017. 1–12
5 Zhang W, Gao B, Tang J, et al. Neuro-inspired computing chips. Nat Electron, 2020, 3: 371–382
6 Liu W, Lombardi F, Shulte M. A retrospective and prospective view of approximate computing. Proc IEEE, 2020, 108: 394–399
7 Zhang D, Jayasena N, Lyashevsky A, et al. TOP-PIM: throughput-oriented programmable processing in memory. In: Proceedings of the 23rd International Symposium on High-Performance Parallel and Distributed Computing, 2014. 85–98
8 Ahn J, Hong S, Yoo S, et al. A scalable processing-in-memory accelerator for parallel graph processing. In: Proceedings of the 42nd Annual International Symposium on Computer Architecture, 2015. 105–117
9 Gokhale M, Holmes B, Iobst K. Processing in memory: the Terasys massively parallel PIM array. Computer, 1995, 28: 23–31
10 Oskin M, Chong F T, Sherwood T. Active pages: a computation model for intelligent memory. In: Proceedings of the 25th Annual International Symposium on Computer Architecture, 1998. 192–203
11 Yin S, Jiang Z, Seo J S, et al. XNOR-SRAM: in-memory computing SRAM macro for binary/ternary deep neural networks. IEEE J Solid-State Circ, 2020, 55: 1733–1743
12 Khwa W S, Chen J J, Li J F, et al. A 65 nm 4 kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3 ns and 55.8 TOPS/W fully parallel product-sum operation for binary DNN edge processors. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2018. 496–498
13 Si X, 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 processors. IEEE Trans Circ Syst I, 2019, 66: 4172–4185
14 Jhang C J, Xue C X, Hung J M, et al. Challenges and trends of SRAM-based computing-in-memory for AI edge devices. IEEE Trans Circ Syst I, 2021, 68: 1773–1786
15 Chi P, Li S, Xu C, et al. PRIME: a novel processing-in-memory architecture for neural network computation in reram-based main memory. SIGARCH Comput Archit News, 2016, 44: 27–39
16 Jain S, Ranjan A, Roy K, et al. Computing in memory with spin-transfer torque magnetic RAM. IEEE Trans VLSI Syst, 2018, 26: 470–483
17 Reis D, Niemier M, Hu X S. Computing in memory with FeFETs. In: Proceedings of the International Symposium on Low Power Electronics and Design, 2018. 1–6
18 Wang Z, Joshi S, Savel'ev S, et al. Fully memristive neural networks for pattern classification with unsupervised learning. Nat Electron, 2018, 1: 137–145
19 Lanza M, Wong H S P, Pop E, et al. Recommended methods to study resistive switching devices. Adv Electron Mater, 2019, 5: 1800143
20 Khan A I, Keshavarzi A, Datta S. The future of ferroelectric field-effect transistor technology. Nat Electron, 2020, 3: 588–597
21 Salahuddin S, Ni K, Datta S. The era of hyper-scaling in electronics. Nat Electron, 2018, 1: 442–450

22 Wei L, Alzate J G, Arslan U, et al. 13.3 A 7 Mb STT-MRAM in 22FFL FinFET technology with 4 ns read sensing time at 0.9 V using write-verify-write scheme and offset-cancellation sensing technique. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2019. 214–216

23 Lee K, Bak J, Kim Y, et al. 1 Gbit high density embedded STT-MRAM in 28 nm FDSOI technology. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2019. 2

24 Jain P, Arslan U, Sekhar M, et al. 13.2 A 3.6 Mb 10.1 Mb/mm^2 embedded non-volatile reram macro in 22 nm finfet technology with adaptive forming/set/reset schemes yielding down to 0.5 V with sensing time of 5 ns at 0.7 V. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2019. 212–214

25 Wu J, Chen Y, Khwa W, et al. A 40 nm low-power logic compatible phase change memory technology. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018

26 Lin C C, Hung J Y, Lin W Z, et al. 7.4 A 256b-wordlength ReRAM-based TCAM with 1 ns search-time and 14× improvement in wordlength-energyefficiency-density product using 2.5 T1R cell. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2016. 136–137

27 Dünkel S, Trentzsch M, Richter R, et al. A FeFET based super-low-power ultra-fast embedded NVM technology for 22 nm FDSOI and beyond. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2017

28 Trentzsch M, Flachowsky S, Richter R, et al. A 28 nm HKMG super low power embedded NVM technology based on ferroelectric FETs. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2016

29 De S, Lu D D, Le H H, et al. Ultra-low power robust 3 bit/cell Hf 0.5 Zr 0.5 O 2 ferroelectric FinFET with high endurance for advanced computing-in-memory technology. In: Proceedings of Symposium on VLSI Technology, 2021. 1–2

30 Lyu X, Si M, Shrestha P, et al. First direct measurement of sub-nanosecond polarization switching in ferroelectric hafnium zirconium oxide. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2019

31 Cheng C D, Tiw P J, Cai Y M, et al. In-memory computing with emerging nonvolatile memory devices. Sci China Inf Sci, 2021, 64: 221402

32 Si X, Chen J J, Tu Y N, et al. 24.5 A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2019. 396–398

33 Yang J, Kong Y, Wang Z, et al. 24.4 sandwich-RAM: an energy-efficient in-memory BWN architecture with pulse-width modulation. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2019. 394–396

34 Si X, Tu Y N, Huanq W H, et al. 15.5 A 28 nm 64 Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2020. 246–248

35 Biswas A, Chandrakasan A P. Conv-RAM: an energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2018. 488–490

36 Jia H, Ozatay M, Tang Y, et al. 15.1 A programmable neural-network inference accelerator based on scalable in-memory computing. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 236–238

37 Valavi H, Ramadge P J, Nestler E, et al. A 64-tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute. IEEE J Solid-State Circ, 2019, 54: 1789–1799

38 Jiang Z, Yin S, Seo J S, et al. C3SRAM: an in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism. IEEE J Solid-State Circ, 2020, 55: 1888–1897

39 Su J W, Chou Y C, Liu R, et al. 16.3 A 28 nm 384 kb 6T-SRAM computation-in-memory macro with 8b precision for AI edge chips. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 250–252

40 Kim H, Chen Q, Yoo T, et al. A 1-16b precision reconfigurable digital in-memory computing macro featuring column-MAC architecture and bit-serial computation. In: Proceedings of the 45th European Solid State Circuits Conference (ESSCIRC), 2019. 345–348

41 Chih Y D, Lee P H, Fujiwara H, et al. An 89TOPS/W and 16.3 TOPS/mm^2 all-digital SRAM-based full-precision compute-in memory macro in 22 nm for machine-learning edge applications. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 252–254

42 Kim J H, Lee J, Lee J, et al. Z-PIM: an energy-efficient sparsity aware processing-in-memory architecture with fully-variable weight precision. In: Proceedings of IEEE Symposium on VLSI Circuits, 2020. 1–2

43 Jiang H, Peng X, Huang S, et al. CIMAT: a compute-in-memory architecture for on-chip training based on transpose SRAM arrays. IEEE Trans Comput, 2020, 69: 944–954

44 Su J W, Si X, Chou Y C, et al. 15.2 A 28 nm 64 kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2020. 240–242

45 Kim J, Koo J, Kim T, et al. Area-efficient and variation-tolerant in-memory BNN computing using 6T SRAM array. In: Proceedings of Symposium on VLSI Circuits, 2019. C118–C119

46 Chen Z, Chen X, Gu J. 15.3 A 65 nm 3T dynamic analog RAM-based computing-in-memory macro and CNN accelerator with retention enhancement, adaptive analog sparsity and 44 TOPS/W system energy efficiency. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 240–242

47 Lee J, Kim J, Jo W, et al. A 13.7 TFLOPS/W floating-point DNN processor using heterogeneous computing architecture with exponent-computing-in-memory. In: Proceedings of Symposium on VLSI Circuits, 2021. 1–2

48 Yue J, Yuan Z, Feng X, et al. 14.3 A 65 nm computing-in-memory-based CNN processor with 2.9-to-35.8 TOPS/W system energy efficiency using dynamic-sparsity performance-scaling architecture and energy-efficient inter/intra-macro data reuse. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2020. 234–236

49 Yue J, Feng X, He Y, et al. A 2.75-to-75.9 TOPS/W computing-in-memory nn processor supporting set-associate block-wise zero skipping and ping-pong CIM with simultaneous computation and weight updating. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 238–240

50 Chen Y H, Krishna T, Emer J, et al. 14.5 Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2016. 262–264

51 Song L, Qian X, Li H, et al. Pipelayer: a pipelined ReRAM-based accelerator for deep learning. In: Proceedings of IEEE International Symposium on High Performance Computer Architecture (HPCA), 2017. 541–552

52 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255

53 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556

54 Jia H, Tang Y, Valavi H, et al. A microprocessor implemented in 65 nm CMOS with configurable and bit-scalable accelerator for programmable in-memory computing. 2018. ArXiv:1811.04047

55 Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network. In: Proceedings of Advances in Neural Information Processing Systems, 2015. 1135–1143

56 Wen W, Wu C, Wang Y, et al. Learning structured sparsity in deep neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 29: 2074–2082

57 Zhang T, Ye S, Zhang K, et al. A systematic DNN weight pruning framework using alternating direction method of multipliers. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 184–199

58 Mao H, Han S, Pool J, et al. Exploring the granularity of sparsity in convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017. 13–20

59 Guo R, Yue Z, Si X, et al. 15.4 A 5.99-to-691.1 TOPS/W tensor-train in-memory-computing processor using bit-level-sparsity-based optimization and variable-precision quantization. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 242–244

60 Sugibuchi K, Kurogi Y, Endo N. Ferroelectric field-effect memory device using $Bi_4Ti_3O_{12}$ film. J Appl Phys, 1975, 46: 2877–2881

61 Yin X, Li C, Huang Q, et al. FeCAM: a universal compact digital and analog content addressable memory using ferroelectric. IEEE Trans Electron Dev, 2020, 67: 2785–2792

62 Ni K, Jerry M, Smith J A, et al. A circuit compatible accurate compact model for ferroelectric-fets. In: Proceedings of IEEE Symposium on VLSI Technology, 2018. 131–132

63 Zheng S, Zhou J, Agarwal H, et al. Proposal of ferroelectric based electrostatic doping for nanoscale devices. IEEE Electron Device Lett, 2021, 42: 605–608

64 Yang G, Niu J, Lu C, et al. Scaling MOS_2 NCFET to 83 nm with record-low ratio of SS ave/SS Ref.= 0.177 and minimum 20 mV hysteresis. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2020

65 Liu H, Wang C, Han G, et al. ZrO_2 ferroelectric FET for non-volatile memory application. IEEE Electron Device Lett, 2019, 40: 1419–1422

66 Liu T, Luo J, Wei X, et al. A novel leaky-fefet based true random number generator with ultralow hardware cost for neuromorphic application. In: Proceedings of the 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM), 2021. 1–3
67 Chen C, Yang M, Liu S, et al. Bio-inspired neurons based on novel leaky-FeFET with ultra-low hardware cost and advanced functionality for all-ferroelectric neural network. In: Proceedings of Symposium on VLSI Technology, 2019. T136–T137
68 Zhao L T, Liu C H, Ren Q H, et al. Research progress of artificial synaptic devices based on ferroelectric materials. J Functional Mater Dev, 2020, 3: 160–168 [赵兰天, 刘晨鹤, 任青华, 等. 基于铁电材料的人工突触器件的研究进展. 功能材料与器件学报, 2020, 3: 160–168]
69 Chen J D, Han W H, Yang C, et al. Recent research progress of ferroelectric negative capacitance field effect transistors. Acta Phys Sin, 2020, 13: 224–252 [陈俊东, 韩伟华, 杨冲, 等. 铁电负电容场效应晶体管研究进展. 物理学报, 2020, 13: 224–252]
70 Cheema S S, Kwon D, Shanker N, et al. Enhanced ferroelectricity in ultrathin films grown directly on silicon. Nature, 2020, 580: 478–482
71 Chen X, Yin X, Niemier M, et al. Design and optimization of FeFET-based crossbars for binary convolution neural networks. In: Proceedings of Design, Automation Test in Europe Conference Exhibition (DATE), 2018. 1205–1210
72 Yin G, Cai Y, Wu J, et al. Enabling lower-power charge-domain nonvolatile in-memory computing with ferroelectric FETs. IEEE Trans Circ Syst II, 2021, 68: 2262–2266
73 Long Y, Kim D, Lee E, et al. A ferroelectric FET-based processing-in-memory architecture for DNN acceleration. IEEE J Explor Solid-State Comput Dev Circ, 2019, 5: 113–122
74 Miller N E, Wang Z, Dash S, et al. Characterization of drain current variations in FeFETs for PIM-based DNN accelerators. In: Proceedings of IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), 2021. 1–4
75 Soliman T, Müller F, Kirchner T, et al. Ultra-low power flexible precision FeFET based analog in-memory computing. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2020. 1–4
76 Saito D, Kobayashi T, Koga H, et al. Analog in-memory computing in FeFET-based 1T1R array for edge AI applications. In: Proceedings of Symposium on VLSI Circuits, 2021. 1–2
77 Shim W, Yu S. Ferroelectric field-effect transistor-based 3-D NAND architecture for energy-efficient on-chip training accelerator. IEEE J Explor Solid-State Comput Dev Circ, 2021, 7: 1–9
78 Dutta S, Ye H, Chakraborty W, et al. Monolithic 3D integration of high endurance multi-bit ferroelectric FET for accelerating compute-in-memory. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2020. 1–4
79 Jerry M, Chen P Y, Zhang J, et al. Ferroelectric FET analog synapse for acceleration of deep neural network training. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2017
80 Luo Y, Wang P, Peng X, et al. Benchmark of ferroelectric transistor-based hybrid precision synapse for neural network accelerator. IEEE J Explor Solid-State Comput Dev Circ, 2019, 5: 142–150
81 Saha A, Islam A N M N, Zhao Z, et al. Intrinsic synaptic plasticity of ferroelectric field effect transistors for online learning. Appl Phys Lett, 2021, 119: 133701
82 Chung W, Si M, Peide D Y. First demonstration of Ge ferroelectric nanowire FET as synaptic device for online learning in neural network with high number of conductance state and G_{max}/G_{min}. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018
83 Ni K, Smith J, Grisafe B, et al. SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2018
84 Noh J, Bae H, Li J, et al. First experimental demonstration of robust HZO/β-Ga_2O-3 ferroelectric field-effect transistors as synaptic devices for artificial intelligence applications in a high-temperature environment. IEEE Trans Electron Dev, 2021, 68: 2515–2521
85 Chen F. PUFFIN: an efficient DNN training accelerator for direct feedback alignment in FeFET. In: Proceedings of IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2021. 1–6
86 Matsui C, Toprasertpong K, Takagi S, et al. Energy-efficient reliable HZO FeFET computation-in-memory with local multiply & global accumulate array for source-follower & charge-sharing voltage sensing. In: Proceedings of

Symposium on VLSI Technology, 2021. 1–2
87 Kamimura K, Nohmi S, Suzuki K, et al. Parallel product-sum operation neuromorphic systems with 4-bit ferroelectric FET synapses. In: Proceedings of the 49th European Solid-State Device Research Conference (ESSDERC), 2019. 178–181
88 Woo J, Moon K, Song J, et al. Improved synaptic behavior under identical pulses using AlO_x/HfO_2Bilayer RRAM array for neuromorphic systems. IEEE Electron Device Lett, 2016, 37: 994–997
89 Yu S. Neuro-inspired computing with emerging nonvolatile memorys. Proc IEEE, 2018, 106: 260–285
90 Marukame T, Nomura K, Matusmoto M, et al. Proposal, analysis and demonstration of analog/digital-mixed neural networks based on memristive device arrays. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2018. 1–5
91 Yin X, Niemier M, Hu X S. Design and benchmarking of ferroelectric FET based TCAM. In: Proceedings of Design, Automation Test in Europe Conference Exhibition (DATE), 2017. 1444–1449
92 Yin X, Ni K, Reis D, et al. An ultra-dense 2FeFET TCAM design based on a multi-domain FeFET model. IEEE Trans Circ Syst II, 2019, 66: 1577–1581
93 Tan A J, Chatterjee K, Zhou J, et al. Experimental demonstration of a ferroelectric HfO_2-based content addressable memory cell. IEEE Electron Device Lett, 2020, 41: 240–243
94 Ni K, Yin X, Laguna A F, et al. Ferroelectric ternary content-addressable memory for one-shot learning. Nat Electron, 2019, 2: 521–529
95 Qian Y, Fan Z, Wang H, et al. Energy-aware designs of ferroelectric ternary content addressable memory. In: Proceedings of Design, Automation Test in Europe Conference Exhibition (DATE), 2021. 1090–1095
96 Hanyu T, Kimura H, Kameyama M. Multiple-valued content-addressable memory using metal-ferroelectric-semiconductor FETs. In: Proceedings of the 29th IEEE International Symposium on Multiple-Valued Logic, 1999. 30–35
97 Li C, Müller F, Ali T, et al. A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2020. 1–4
98 Rajaei R, Sharifi M M, Kazemi A, et al. Compact single-phase-search multistate content-addressable memory design using one FeFET/Cell. IEEE Trans Electron Dev, 2021, 68: 109–117
99 Pagiamtzis K, Sheikholeslami A. Content-addressable memory (CAM) circuits and architectures: a tutorial and survey. IEEE J Solid-State Circ, 2006, 41: 712–727
100 Laguna A F, Gamaarachchi H, Yin X, et al. Seed-and-vote based in-memory accelerator for DNA read mapping. In: Proceedings of IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2020. 1–9
101 Rahimi A, Ghofrani A, Cheng K T, et al. Approximate associative memristive memory for energy-efficient GPUs. In: Proceedings of Design, Automation Test in Europe Conference Exhibition (DATE), 2015. 1497–1502
102 Yin X, Chen X, Niemier M, et al. Ferroelectric FETs-based nonvolatile logic-in-memory circuits. IEEE Trans VLSI Syst, 2019, 27: 159–172
103 Kazemi A, Sharifi M M, Laguna A F, et al. In-memory nearest neighbor search with FeFET multi-bit content-addressable memories. In: Proceedings of 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2021. 1084–1089
104 Lee M, Tang W, Xue B, et al. FeFET-based low-power bitwise logic-in-memory with direct write-back and data-adaptive dynamic sensing interface. In: Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design, 2020. 127–132
105 Reis D, Niemier M T, Hu X S. A computing-in-memory engine for searching on homomorphically encrypted data. IEEE J Explor Solid-State Comput Dev Circ, 2019, 5: 123–131
106 Breyer E T, Mulaosmanovic H, Trommer J, et al. Compact FeFET circuit building blocks for fast and efficient nonvolatile logic-in-memory. IEEE J Electron Dev Soc, 2020, 8: 748–756
107 Chen X, Niemier M, Hu X S. Nonvolatile lookup table design based on ferroelectric field-effect transistors. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2018. 1–5
108 Zhang X, Chen X, Han Y. FeMAT: exploring in-memory processing in multifunctional FeFET-based memory array. In: Proceedings of IEEE 37th International Conference on Computer Design (ICCD), 2019. 541–549
109 Laguna A F, Yin X, Reis D, et al. Ferroelectric FET based in-memory computing for few-shot learning. In: Pro-

ceedings of the 2019 on Great Lakes Symposium on VLSI, 2019. 373–378
110 Reis D, Laguna A F, Niemier M, et al. Exploiting FeFETs via cross-layer design from in-memory computing circuits to meta-learning applications. In: Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE), 2021. 306–311
111 Reis D, Gao D, Angizi S, et al. Modeling and benchmarking computing-in-memory for design space exploration. In: Proceedings of the 2020 on Great Lakes Symposium on VLSI, 2020. 39–44
112 Ma K, Zheng Y, Li S, et al. Architecture exploration for ambient energy harvesting nonvolatile processors. In: Proceedings of IEEE 21st International Symposium on High Performance Computer Architecture (HPCA), 2015. 526–537
113 Liu Y, Li Z, Li H, et al. Ambient energy harvesting nonvolatile processors: from circuit to system. In: Proceedings of the 52nd Annual Design Automation Conference, 2015. 1–6
114 Yin X, Aziz A, Nahas J, et al. Exploiting ferroelectric FETs for low-power non-volatile logic-in-memory circuits. In: Proceedings of IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2016. 1–8
115 Li M, Yin X, Hu X S, et al. Nonvolatile and energy-efficient FeFET-based multiplier for energy-harvesting devices. In: Proceedings of the 25th Asia and South Pacific Design Automation Conference (ASP-DAC), 2020. 562–567
116 Breyer E T, Mulaosmanovic H, Slesazeck S, et al. Demonstration of versatile nonvolatile logic gates in 28 nm HKMG FeFET technology. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2018. 1–5
117 Breyer E, Mulaosmanovic H, Mikolajick T, et al. Reconfigurable NAND/NOR logic gates in 28 nm HKMG and 22 nm FD-SOI FeFET technology. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2017
118 You W X, Huang B K, Su P. An alternative way for reconfigurable logic-in-memory with ferroelectric FET. IEEE Trans Electron Dev, 2022, 69: 444–446
119 Wang D, George S, Aziz A, et al. Ferroelectric transistor based non-volatile flip-flop. In: Proceedings of the 2016 International Symposium on Low Power Electronics and Design, 2016. 10–15
120 Li X, George S, Liang Y, et al. Lowering area overheads for FeFET-based energy-efficient nonvolatile flip-flops. IEEE Trans Electron Dev, 2018, 65: 2670–2674
121 Saki A A, Lin S H, Alam M, et al. A family of compact non-volatile flip-flops with ferroelectric FET. IEEE Trans Circ Syst I, 2019, 66: 4219–4229
122 Kim S K, Oh T W, Lim S, et al. High-performance and area-efficient ferroelectric FET-based nonvolatile flip-flops. IEEE Access, 2021, 9: 35549–35561
123 Thirumala S K, Raha A, Jayakumar H, et al. Dual mode ferroelectric transistor based non-volatile flip-flops for intermittently-powered systems. In: Proceedings of the International Symposium on Low Power Electronics and Design, 2018. 1–6

Computing-in-memory circuits and cross-layer integrated design and optimization: from SRAM to FeFET

Xunzhao YIN[1†], Jinshan YUE[2†], Qingrong HUANG[1], Chao LI[1], Jiahao CAI[1], Zeyu YANG[1], Cheng ZHUO[1*] & Ming LIU[2*]

1. *College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou* 310012*, China*;
2. *Institute of Microelectronics of the Chinese Academy of Sciences, Beijing* 100029*, China*
* Corresponding author. E-mail: czhuo@zju.edu.cn, liuming@ime.ac.cn
† Equal contribution

Abstract In the era of artificial intelligence (AI) and the Internet of Things (IoT), the emerging data-driven applications and tasks have significantly promoted the development of national digitization. However, due to the separation of storage and computing in traditional von Neumann hardware, the resulted memory wall issue leads to heavy data transfer costs in data-intensive applications, which inhibits the improvements of energy efficiency and performance. As a novel computing paradigm deviating from the traditional architecture in the post-Moore era, the computing-in-memory (CiM) technique integrates computing logic into storage by leveraging the characteristics of memory devices and circuits to eliminate the data transfer overhead. Such a promising approach can significantly improve the energy efficiency and performance of intelligent hardware platforms. Based on the traditional CMOS and the emerging non-volatile memory device ferroelectric FET, this review summarizes the key CiM circuit designs, and discusses the integrated design and optimization approaches across the device, architecture, chip, algorithm, and application layers.

Keywords computing-in-memory, static random access memory, ferroelectric field effect transistor, crossbar, content addressable memory