

引用格式: 龚龙庆, 徐伟栋, 娄冕. SRAM 存内计算技术综述[J]. 微电子学与计算机, 2021, 38(9): 1-7. [GONG L Q, XU W D, LOU M. An overview of SRAM in-memory computing[J]. Microelectronics & Computer, 2021, 38(9): 1-7.]

SRAM 存内计算技术综述

龚龙庆, 徐伟栋, 娄 冕

(西安微电子技术研究所, 陕西 西安 710054)

摘 要: 在处理深度神经网络这类数据密集型应用的过程中, 处理器和存储器间大量数据的频繁传输会造成严重的性能损耗和能量消耗, 也是当前冯·诺伊曼架构最大的瓶颈. 针对传统冯·诺伊曼体系架构的局限性, 基于 SRAM 的存内计算技术将运算单元集成到内存中, 支持数据的即存即算, 彻底突破了冯·诺伊曼瓶颈, 有望成为新一代智能计算架构. 本文从体系结构的角度阐明了冯·诺伊曼架构所引起的“功耗墙”和“存储墙”问题, 并给出了存内计算技术的兴起原因. 文章围绕近几年国内外关于 SRAM 存内计算架构的研究, 以其中几种经典架构为例描述了各类 SRAM 存内计算的工作机理、优缺点及意义, 并从器件级、电路级和架构级的角度分别概述了目前关于 SRAM 存内计算技术的关键影响因素. SRAM 存内计算技术潜力巨大, 用途广泛, 将会给机器学习应用, 图计算应用和基因工程提供高效低能耗的系统结构支持, 最后展望了未来几年内 SRAM 存内计算技术在器件、电路和架构方面的发展情况.

关键词: 数据密集型应用; 冯·诺伊曼架构; SRAM; 存内计算

中图分类号: TP3

文献标识码: A

文章编号: 1000-7180(2021)09-0001-07

An overview of SRAM in-memory computing

GONG Longqin, XU Weidong, LOU Mian

(Xi'an Microelectronics Technology Institute, Xi'an 710054, China)

Abstract: In the process of processing data-intensive applications such as deep neural networks, the frequent transfer of large amounts of data between the processor and the memory causes severe performance loss and energy consumption, which is the biggest bottleneck of the current von Neumann architecture. In view of the limitations of the traditional von Neumann architecture, the SRAM-based in-memory computing technology integrates the computing unit into the memory to support data storage and calculation, which completely breaks through the von Neumann bottleneck and is expected to become a new generation Intelligent computing architecture. This paper clarifies the problems of "power wall" and "storage wall" caused by the von Neumann architecture from the perspective of architecture, and gives the reasons for the rise of in-memory computing. The paper focuses on the research of SRAM-based in-memory computing architectures in recent years, and describes the working mechanism, advantages and disadvantages and significance of various SRAM-based in-memory computing architectures by taking several classical architectures as examples. And from the perspective of device level, circuit level and architecture level, the key factors of current SRAM-based in-memory computing technology are summarized respectively. The SRAM-based in-memory computing technology is a promising and versatile technology that will provide efficient and low-energy system architectures for machine learning applications, graph computing applications and genetic engineering. The paper looks forward to the development of SRAM-based in-memory computing technology in devices, circuits and architectures in the coming year

Key words: data-intensive applications; Von Neumann architecture; SRAM; in-memory computing

1 引言

近年来,互联网、云计算、人工智能和物联网等高新技术的发展,推动了数据量的急剧增长.全球每天会创建约 5 亿条推文、3 200 亿封电子邮件,在 Facebook 上产生 4 PB 数据,在可穿戴设备上生成 28 PB 的数据^[1].根据国际数据公司 IDC(International Data Corporation)发布的《数据时代 2025》^[2],全球数据将由 2018 年的 33 ZB 增长到 2025 年的 175 ZB,约每 40 个月翻一番,平均每天会生成约 491 EB 数据.海量的数据开创了无人驾驶、人脸识别、智能手环、专家服务等新的市场,也为日常生活提供了更多便利的选择.但如何有效处理这些大数据,从中提取关键信息同样也对计算机系统提出了更高的挑战.

传统的计算机结构是基于冯·诺伊曼架构,如图 1 所示,由运算器、控制器、存储器以及输入输出设备组成,各个模块之间通过总线互联.这种架构逻辑清晰、结构简单、功能完备、以计算为主,能兼容各种类型的算法,适用于那些计算密集型应用.在处理像深度神经网络、图像处理程序这类数据密集型应用时,大量数据在总线频繁迁移所产生的高延时和大功耗会严重制约整个系统的整体性能.一方面,运算器和存储器之间的性能差距正在逐步拉大.近几十年来,运算器的计算速度以每年 55% 的速率增长,而存储器的存取速度则以每年 10% 的速率增长.运算器每次完成一组计算后都需要等待从存储器中取出的下一组数据,运算性能无法得到充分发挥,形成所谓的“存储墙”.同时,数据的有限带宽进一步限制了系统性能.如图 1 所示,CPU 速率为 92 TB/s,而 Cache 速率和总线带宽分别为 1 TB/s 和 167 GB/s^[3],存储速率和总线带宽严重制约了系统速率.另一方面,频繁的数据迁移还会带来严重的功耗问题.例如,在进行一次 32 位加法运算时,数据运算仅需要 0.2 pJ,而搬运数据的功耗是 42 pJ,是运算功耗的两百多倍,大部分能量都被用于搬运数据,造成了“功耗墙”问题.在电子信息领域,“存储墙”与“功耗墙”共同造成了冯·诺伊曼瓶颈.

为了打破冯·诺伊曼瓶颈,在智能大数据背景下构建更高效的硬件设施与计算架构,研究者提出了存内计算技术.该技术将计算和存储融合,支持数据在存储模块中直接运算,并将最终结果反馈回处

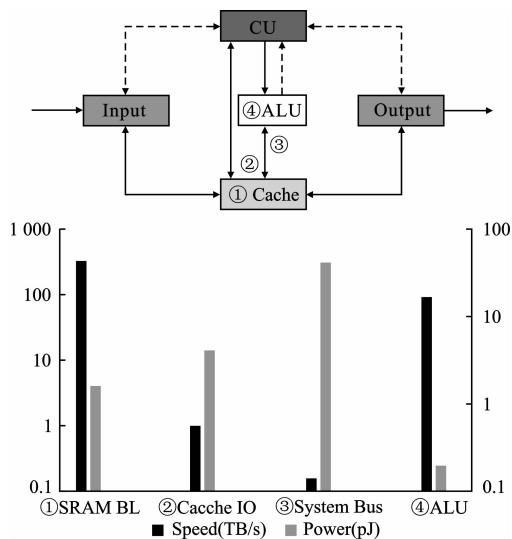


图 1 冯·诺伊曼架构和冯·诺伊曼瓶颈

理器,从而大幅缩减了数据在总线传输的用时和能耗. Cache 作为离 CPU 最近的存储器,也是目前最常见的存储器,目前大都采用 SRAM. 基于 SRAM 的存内计算技术可以直接在 Cache 内进行数据的读写和计算,无需再经过总线,具有读写速度快、功耗低、可靠性强、可扩展性高等优点,引起了人们的普遍关注,有望成为新一代主流智能计算架构.

2 基于 SRAM 的存内计算技术

近几年关于存内计算技术的研究大都集中于模拟计算,能够实现多位数据的运算,具有更高的计算效率和系统能效,但要实现足够灵活可靠的模拟计算仍然是现有存内计算架构面临的关键挑战. 目前模拟域的 SRAM 存内计算主要可分为基于电压域的计算和基于时域的计算两种.

2.1 基于电压域的 SRAM 存内计算技术

在电压域上实现的 SRAM 存内计算技术,通常先使用 DAC 单元将数字量先转化为线性的电压值,然后利用电荷共享的方式实现计算,最后再用 ADC 单元将模拟的计算结果转换为相应的数字信号. 通过将数据量转换成电压值的方式实现多位算法,简单易行,也是当下存内计算的最主要实现形式. 如果按计算模块和存储模块之间的距离划分,存内计算技术又可细分为近内存计算、全局计算、阵列中计算和单元中计算四种. 下面分别举例描述了这四种不同 SRAM 存内计算.

(1) 基于位串性技术的 8T SRAM 通用近内存计算^[3]. Wang 等人提出了一种混合近内存计算阵

列,可用于需要高效、高灵活性和高可编程性的一般用途的应用. 该方案支持逻辑操作、整数/浮点算术运算和比较运算. 如图 2 所示,权重存储于 SRAM 单元中,输入被应用于计算字线 CWLs,模拟结果则通过对计算位线 CBL 或反计算位线 CBLB 上的电压进行平均来获得. 该近内存计算电路使用 ADC 和逻辑门来产生数字输出,并行度大,相同列多个行上的单元可同时进行同一运算,但计算面积占比大,每一行都需要两个增益放大器和一个运算电路.

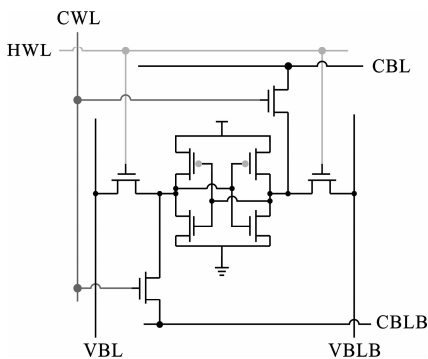


图 2 8T 转置单元

(2)基于分裂式字线的 6TSRAM 单比特全局计算^[4]. 如图 3 所示,当输入为 1 时, WLL 将被设置成高电平, WLR 接地;当输入为 -1 时,则相反. 当权重为 1 时,储值 Q 为 1;当权重为 -1 时, Q 为 0. 此时, BL/BLB 上表示的是输入和权重同或的结果. 如果结果为 1,则在 BL 或 BLB 上形成一个充电电流;如果结果为 0,则会形成一个放电电流最终通过全局读出电路进行识别. 该架构结构简单、面积小、功耗低,但只能用于简单的逻辑操作,不支持复杂算术运算.

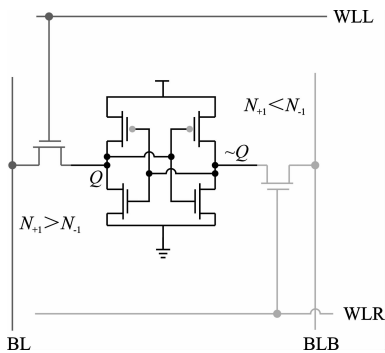


图 3 分裂式字线 6T 单元

(3)基于局部计算单元和 6TSRAM 的多比特阵列中计算^[5]. 如图 4 所示,分比特乘法计算单元由

16 个存储用的 6TSRAM 和局部计算单元构成. 在读写模式下, HWL 置为高电平, 系统通过 VGBL 和 VGBLB 对阵列中相应的 SRAM 单元进行正常读写. 计算模式下, HWL 接地, LBL 上会反映存储的权重值. 4 bit 输入 $IN[3:0]$ 会被拆分成两组分别送到对应的垂直方向上的两条全局位线上, 其中 $VGBL$ 代表 $IN[3:2]$, $VGBLB$ 代表 $IN[1:0]$, 并用 4 种电压电平来分别表示这两位两比特输入数据. 最后, 水平方向上的全局位线上 HGBL 就会表现为 $IN[3:2]$ 跟权重 W 的乘法电流. 对应地, HGBLB 上会呈现出 $IN[1:0]$ 跟权重 W 的乘法电流. 基于局部计算单元和 6TSRAM 的多比特阵列中计算面积小、容量大, 可进行多位数据的 MAC 运算, 能适配不同的神经网络, 灵活度高, 同时具有很高的能耗效率, 但这种阵列中计算对参考电压设计要求很高, 计算周期长.

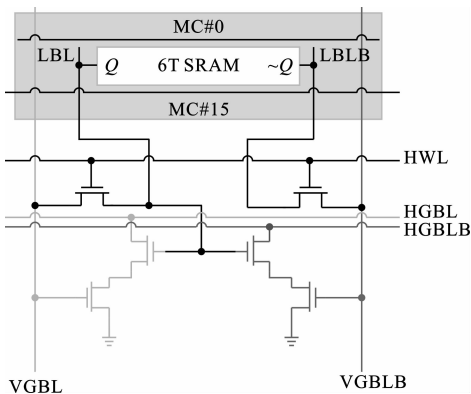


图 4 分比特乘法计算单元

(4)基于双生 8TSRAM 的多比特单元中计算^[6]. 如图 5 所示, 每个基本单元由高位 8T 单元和低位 8T 单元构成. 其中, 高位 8T 单元读端口上两个晶体管的宽长比是低位 8T 单元的两倍, 从而流过高位 8T 单元的电流刚好是低位 8T 单元的两倍, 因此该双生 8TSRAM 正好能存储 2 bit 的权重. 两位输入模式下, 会在读字线上输入四种不同的电压, 产生不同的电流. 每一个 T8T SRAM 单元可以实现 2 bit 输入和 2 bit 权重的乘法操作, 其对应到 RBL 上的电流 ($I_{MC-T8T} = I_{MC-M8T} + I_{MC-L8T}$) 正比于输入 IN 和权 W 的乘法结果. 由于双生 8T SRAM 单元的引入, 该架构能很好地支持多比特的深度神经网络, 同时获得很高的识别准确率, 但这类单元中面积较大, 对参考电压设计的要求高, 完成一次运算时间长.

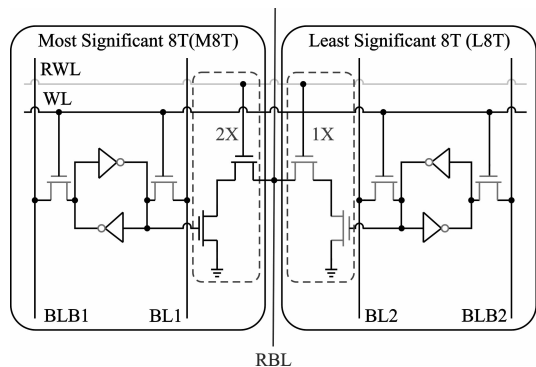


图 5 双生 8T SRAM 单元

尽管电压域 SRAM 存内计算技术原理简单、易于操作、计算效率高,但电路的工作电压本身存在上限,也难以被多次细分,只能代表有限的位数.因此,电压域 SRAM 存内计算技术本身存在着天然的瓶颈.

2.2 基于时域的 SRAM 存内计算技术

基于时域的 SRAM 存算技术通常会使用线性的脉冲宽度或路径延时来表示多位数字.由于时间量不像电压量那样存在上限,可以被多次细分,具有更大的优势.根据数字量表示方式的不同,目前时域 SRAM 存内计算技术又可被分为两类:基于路径延时的计算和基于脉冲宽度调制的计算.

(1) 基于路径延时的 TDNN 存内计算^[7]. 每个计算单元都会产生一定的延时,将计算路径上所有计算单元串联在一起,就能完成信号的累加.如图 6 所示,框内为一个计算单元的具体电路形式,其中, t 为两个脉冲 V_p 和 V_n 的时间差,权重值 w_i 控制电阻 R_p 和 R_n 来产生不同的延时,特征值 x_i 用来选择 V_p 和 V_n 的传播方向,最终延时 t_i 为权重和特征的乘累加结果.通过模拟延时链的方式,该方法获得了 48.2 TOPS/W 的高能效和 76.5 GE/PE 的硬件有效性,给众多研究者提供了新的研究思路,但只能通过比较脉冲先后来实现 BNN,无法完成更高精度的神经网络.

(2) 基于脉冲宽度调制的三明治 8T-RAM 存内计算^[8]. Yang 等人设计了一款基于 8TSRAM 三明治 RAM 设备,能够实现特征值 8 bit,权重 1bit 的 BWN 网络.如图 7 所示,该方案将特征值和权重值的存储单元分别存储在脉冲宽度调制单元 PWMU 附近.在计算模式下,两位输入会经 2—4 译码器置换成 4 种电压水平,用于调整脉冲宽度.权重作用在选择器上,如果权重为 1,脉冲宽度会对应扩展,反

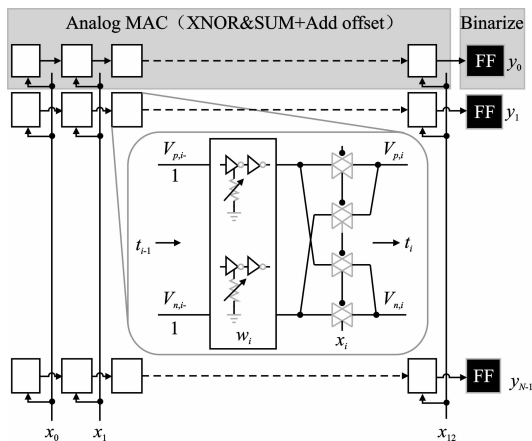


图 6 时域神经网络协处理器

之则缩短,最终经脉冲量化器量化得到最终结果.该设计通过特制的 PWMU 实现了时间域的乘加计算,能耗效率可达到 119.7 TOPS/W,但是 8T SRAM 和 PWMU 单元都会造成大面积开销.同时,模拟域的 SRAM 存内计算对 PVT 以及版图走线比较敏感,相应的计算精度、量化误差问题仍亟需进一步优化改进.

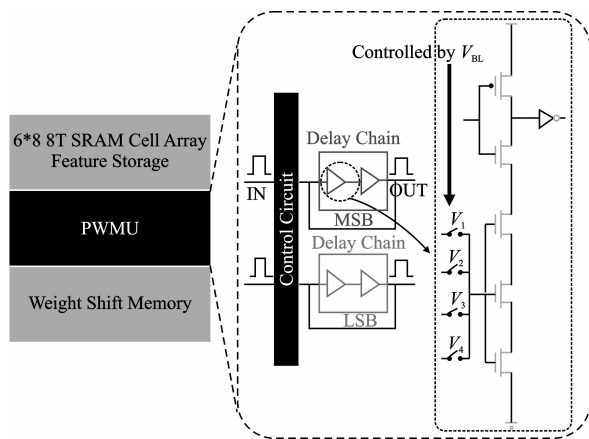


图 7 三明治式 SRAM 架构

3 基于 SRAM 存内计算的关键影响因素

3.1 器件特性

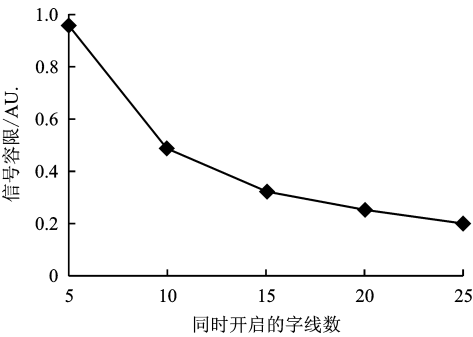
(1) 单元结构. 部分 SRAM 存内计算会采用传统的 6TSRAM^[5, 7], 面积相对较小,但读取不稳定,应用范围受限.出于功能需要,8T^[3, 6, 8]、10T^[9-10]、12T^[11]、或者特定的 6TSRAM^[4, 12] 都有在相应的存内计算技术中被用到.这些新单元构成的存储阵列往往会占据芯片的绝大部分面积,往往单元管子数目越多、阵列规模越大,芯片面积越大.目前多数 SRAM 存内计算技术都会重新修改设计规则,使用

全定制的单元结构来实现计算功能,并缩减阵列面积开销。

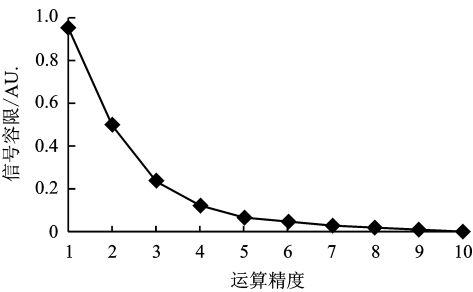
(2)工艺变动. SRAM 单元中的工艺变动会导致位线电压的波动,进而影响检测域值和最终结果,传统 6TSRAM 所受影响最大,检测阈值最低. 目前,常用的方法是采用更大的晶体管^[6, 13],更低的信号裕度^[8, 14],或模拟-数字混合的存内计算架构^[5, 8, 10, 15].

3.2 电路选择

(1)字线开启数目. 往往开启的字线数越多,计算并行度越高、系统能效越高,但位线上的电压下降越厉害,当 V_{BL} 低于某个 SRAM 的写容限时,该 SRAM 上的储值会由原来的“1”下拉到“0”,从而造成“读破坏”问题. 为了避免这种现象发生,保证单元储值不受计算的影响,通常要求 V_{BL} 保持在 SRAM 写容限之上^[14, 16],或者采用非 6T 传统架构的 SRAM^[5, 6, 10]. 同时,开启字线越多,同一条位线上所见到的 MAC 值增加,而位线上的最大电压摆幅是固定的,从而造成位线上的信号容限缩减. 当信号容限低于增益放大器的输入失调电压时,位线上的值无法再被准确读出,计算结果读出错误,如图 8 (a)所示。



(a)信号容限和字线开启数目



(b)信号容限和运算精度

图 8 有限的信号容限

(2)计算精度. 在深度神经网络的 MAC 操作

中,提高输入和权重的精度,以及累积的数量都能够提高全输出精度,促使推断结果更准确,但这需要字线/位线上所容纳的值更多,信号裕度也会呈指数级下降,如图 8(b)所示. 同时,更高精度的输入/输出需要更长的时间和更多的 ADC 单元将其转换. 因此,设计 SRAM 存内计算电路需要在准确度和性能、面积之间做合适的折中。

(3)周边电路. 在存储模式下,位线和反位线分别为开启 SRAM 单元的正反两项,可直接用增益放大器读取单元储值. 但在模拟 SRAM 存内计算电路中,位线和反位线上的电压可能是多位数据的计算结果,不再反向,需要分别和参考电压比较来获取计算结果. 在之前的工作中,为了减少 PVT 的影响,这些参考电压大都是通过冗余的 SRAM 列来产生^[4-6],但会消耗大量面积,造成存储密度下降,参考电压产生电路设计和控制也比较复杂. 同时,在某些 SRAM 存内计算电路^[3,15]中,需要使用到逻辑门来完成计算功能,这进一步加剧了电路的面积开销和能量损耗。

3.3 系统架构

(1)硬件配置与应用需求. 存内计算中逻辑模块通常会使用 FPGA、CGRA 和 ASIC 处理器等非通用处理器,而不同的应用对存内计算结构要求并不同,这就要求在设计存内计算结构时需要同时考虑硬件配置与应用需求,而该问题又与电路设计和器件选择强相关。

(2)数据存储方式. 传统的数据存储方式并不适用于神经网络的高效实施,会形成计算资源的分配不均匀,计算资源利用率低. 该问题的产生通常又与编译器设计和内存控制器设计相关. 内存中数据的存储方式需要考虑到计算资源的利用率,尽量使所有计算资源能够时分复用,管道串行,避免因计算资源集中在某一块而导致长时间的数据等待. 对于特定的存内计算架构,需要重新设计合适的数据映射策略来提高计算资源的利用率。

(3)系统软件的调度方法. 运行在存内计算系统中的应用要在中央处理器和存内计算端来回切换. 此时,存内计算中的多任务调度、存内计算任务和传统访存任务的调度都需要具体设计. 另外,存内计算中多个计算模块之间会共享数据,包括并发地对数据进行增删改查等操作. 因此,存内计算系统急需灵活的系统支持. 目前的存内计算都是针对特定的

硬件结构或特定应用提供的上层软件支持,灵活性不够。

4 基于 SRAM 存内计算的发展趋势

随着神经网络等加速算法向更深的层次和更复杂的结构发展,人工智能硬件化也在朝着更大的并行性、更高的能源效率、更大的吞吐量和更好的兼容性发展。SRAM 存内计算技术具有高算力、低功耗、低成本等优势,未来在人工智能领域会有非常大的应用前景。

(1)单元结构. 基于传统 6TSRAM 的存内计算技术会存在读破坏、功能单一的缺点,为了实现更复杂的运算,研究者提出了各种不同结构的 SRAM 单元,如分列式字线的 6TSRAM^[4,12],用作转置单元的 8TSRAM^[3,17],能存储 2 bit 权重的双 8TSRAM^[6]。出于计算的需求,大多数存内计算架构并不再采用传统的 6TSRAM。具备读稳定、抗 PVT 扰动的新型 SRAM 单元将成为存内计算技术的主流。

(2)电路形式. 传统的 SRAM 阵列只能用于存储数据,并无计算功能。随着人工智能的蓬勃发展,从近数据计算^[8,14,17]到全局计算^[15,16,18],再到阵列中计算^[5,19],最后到单元中计算^[6,11],未来 SRAM 存内计算中计算部件将会越来越接近 SRAM 存储阵列,存储和计算的距离越来越短,所需的数据搬运越来越少,系统能效越来越高。

(3)系统结构. 目前 SRAM 存内计算芯片主要有两种用途,一种是作为 AI 算法的加速器,专门用来处理 AI 算法,另一种是计算机协处理器,辅助计算机处理那种算子单一但数据量大的计算任务。未来 SRAM 存内计算的大规模普及,需要产业生态的建设,芯片厂商、软件工具厂商和应用集成厂商等的协同研发,实现性能与应用场景结合。

5 结束语

大数据时代的到来,人工智能应用场景中的深度神经网络模型不断发展,对计算架构提出了更高的要求。本文讨论了存内计算技术的结构、优势以及基于不同器件的存内计算技术,并对基于 SRAM 的存内计算技术展开了详细介绍,给出了目前 SRAM 存内计算的研究现状、关键影响因素和未来发展趋势。

基于 SRAM 的存内计算可以打破冯·诺伊曼瓶颈,并用作具有高能效的终端 AI 芯片。目前,基于 SRAM 存内计算技术总体还处于发展初期,尤其是当前很有潜力的模拟存内计算,存在着一些硬件和软件支持上的问题。其中,低功耗读取电路,如灵敏放大器的设计仍然是实现低功耗高能效存内计算的关键,可以研究具有更好性能的读取电路设计来进一步降低其能量消耗。

参考文献:

- [1] RACONTEUR. A day in data[EB/OL]. [2020-06-04]. <https://www.raconteur.net/infographics/a-day-in-data/>.
- [2] REINSEL D, GANTZ J, RYDNING J. Data age 2025: The digitization of the world from edge to core[R]. IDC White Paper- #US44413318.
- [3] WANG J C, WANG X W, ECKERT C, et al. A 28-nm compute SRAM with Bit-Serial logic/arithmetic operations for programmable in-memory vector computing[J]. IEEE Journal of Solid-State Circuits, 2020, 55(1): 76-86. DOI: 10.1109/JSSC.2019.2939682.
- [4] SI X, KHWAW 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 DNNedge processors[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2019, 66(11): 4172-4185. DOI: 10.1109/TCSI.2019.2928043.
- [5] SI X, TU Y N, HUANG W H, et al. 15.5 A 28nm 64Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips[C]//2020 IEEE International Solid-State Circuits Conference (ISSCC). San Francisco, CA, USA: IEEE, 2020: 246-248. DOI: 10.1109/ISSCC19947.2020.9062995.
- [6] SI X, CHEN J J, TUI Y N, et al. 24.5A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning[C]//2019 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA: IEEE, 2019: 396-398. DOI: 10.1109/ISSCC.2019.8662392.
- [7] MIYASHITA D, KOUSAI S, SUZUKI T, et al. A Neuromorphic chip optimized for deep learning and CMOS technology with time-domain analog and digital mixed-signal processing[J]. IEEE Journal of Solid-State Circuits, 2017, 52(10): 2679-2689. DOI: 10.

- 1109/JSSC. 2017. 2712626.
- [8] YANG J, KONG Y Y, WANG Z, et al. 24.4 sandwich-RAM: An energy-efficient in-memory BWN architecture with pulse-width modulation [C]//2019 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA: IEEE, 2019: 394-396. DOI: 10.1109/ISSCC.2019.8662435.
- [9] NGUYENV T, KIM J S, LEE J W. 10T SRAM computing-in-memory macros for binary and multibit MAC operation of DNN edge processors[J]. IEEE Access, 2021(9): 71262-71276. DOI: 10.1109/ACCESS.2021.3079425.
- [10] ZHANG Y Q, XU L, DONG Q, et al. Recryptor: A reconfigurable cryptographic cortex-M0 processor with in-memory and near-memory computing for IoT-security[J]. IEEE Journal of Solid-State Circuits, 2018, 53(4): 995-1005. DOI: 10.1109/JSSC.2017.2776302.
- [11] JIANG Z W, YIN S H, SEOK M, et al. XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks [C]//2018 IEEE Symposium on VLSI Technology. Honolulu, HI, USA: IEEE, 2018: 173-174. DOI: 10.1109/VLSIT.2018.8510687.
- [12] DONG Q, JELOKA S, SALIGANEM, et al. A 0.3 V VDDmin 4+2T SRAM for searching and in-memory computing using 55 nm DDC technology[C]//2017 Symposium on VLSI Circuits. Kyoto, Japan, 2017: C160-C161. DOI: 10.23919/VLSIC.2017.8008465.
- [13] BISWAS A, CHANDRAKASANA P. Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications[C]//2018 IEEE International Solid-State Circuits Conference (ISSCC). San Francisco, CA, USA: IEEE, 2018: 488-490. DOI: 10.1109/ISSCC.2018.8310397.
- [14] GONUGONDLA S K, KANG M G, SHANBHAG N. A 42 pJ/decision 3.12 TOPS/W robust in-memory machine learning classifier with on-chip training[C]//2018 IEEE International Solid-State Circuits Conference (ISSCC). San Francisco, CA, USA: IEEE, 2018: 490-492. DOI: 10.1109/ISSCC.2018.8310398.
- [15] ZHANG Z X, CHEN J J, SIX, et al. A 55nm 1-to-8 bit configurable 6T SRAM based computing-in-memory unit-macro for CNN-based AI edge processors [C]//2019 IEEE Asian Solid-State Circuits Conference (A-SSCC). Macau, Macao, China: IEEE, 2019: 217-218. DOI: 10.1109/A-SSCC47793.2019.9056933.
- [16] GUO R Q, LIU Y G, ZHENG S Y, et al. A 5.1pJ/Neuron 127.3 us/inference RNN-based speech recognition processor using 16 computing-in-memory SRAM macros in 65nm CMOS[C]//2019 Symposium on VLSI Circuits. Kyoto, Japan: IEEE, 2019: C120-C121. DOI: 10.23919/VLSIC.2019.8778028.
- [17] ECKERT C, WANG X W, WANG J C, et al. Neural cache: bit-serial in-cache acceleration of deep neural networks[C]//2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA). Los Angeles, CA, USA: IEEE, 2018: 383-396. DOI: 10.1109/ISCA.2018.00040.
- [18] YUC S, YOOT, KIM T T, et al. A 16K current-based 8T SRAM compute-in-memory macro with decoupled read/write and 1-5bit column ADC[C]//2020 IEEE Custom Integrated Circuits Conference (CICC). Boston, MA, USA: IEEE, 2020: 1-4. DOI: 10.1109/CICC48029.2020.9075883.
- [19] XUE C X, CHEN W H, LIU Y S, et al. A 1Mb Multibit ReRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN based AI edge processors[C]//2019 IEEE International Solid-State Circuits Conference (ISSCC). San Francisco, CA, USA: IEEE, 2019: 388-390. DOI: 10.1109/ISSCC.2019.8662395.

作者简介:

龚龙庆 男,(1962-),研究员,博士生导师.研究方向为空间嵌入式计算机设计与应用技术.

E-mail: glq771@sina.com.

徐伟栋 男,(1993-),博士研究生.研究方向为存内计算技术.

娄冕 男,(1987-),博士,高级工程师.研究方向为IC设计技术.