文章编号:1007 - 4252(2021)06 - 0006 - 525

用于存算一体的磁性随机存储器概述

姚佳伦[1],杨雨梦[1],陈昊瑜[2,*]
(1. 上海科技大学 信息科学与技术学院,上海 201210; 2. 上海华力微电子有限公司,上海 201203)

摘要:磁性随机存储器是一种基于自旋电子学的新型信息存储器件,其主要结构单元是一个由磁性层和隧穿层组成的磁性隧道结,通过铁磁材料相对的磁化方向表现出高低两种阻值状态,以此实现信息的非易失存储。它具有极快的开关速度、近乎为零的泄露功耗、极高的可靠性等显著优点,是实现存算一体化技术的理想器件之一。本综述论述了磁性随机存储器在存算一体领域的研究进展,包括器件的基本结构和相应控制方法,着重对其在算术逻辑和神经网络的计算研究现状做了阐述。最后,对磁性随机存储器在存算一体中的应用做了相应总结和展望。

关键词:铁磁材料;存算一体;磁场随机存储器;非易失性存储器;计算机架构
中图分类号:TN40 **文献标志码**:A

An Overview of In-memory Computing Based on Magnetic Random Access Memory

YAO Jia-lun[1], YANG Yu-meng[1], CHEN Hao-yu[2,*]
(1. School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China; 2. Shanghai HuaLi Microelectronics Corporation, Shanghai 201203, China)

Abstract: Magnetic random access memory (MRAM) is a novel information storage device based on spintronics. Its building block typically consists of a magnetic tunnel junction, which is a sandwich structure with two ferromagnetic layer and one insulating barrier. The relative directions of the ferromagnets' magnetizations lead to high and low resistance states, which is then employed to store the data in a non-volatile fashion. MRAM has extremely fast switching speed, nearly zero leakage power consumption and extremely high reliability, all of which stands it out as one of the ideal devices for in-memory computing applications. This review discusses the research progresses on MRAM based in-memory computing, including the basic device structures and corresponding control methods. We will then mainly focus on the description of current status of MRAM based in-memory computing using arithmetic logic and neural network. Finally, some challenges and perspectives on the field are given.

收稿日期:2021-11-29; **修订日期**:2021-12-05
基金项目:上海科技大学教师启动经费.
作者简介:姚佳伦(1996-),男,硕士研究生,主要研究方向为存储器工艺研发 (E-mail:yaojl1@shanghaitech.edu.cn).
杨雨梦(1989-),男,博士,助理教授,主要研究方向为自旋电子器件 (E-mail:yangym1@shanghaitech.edu.cn).
通信作者:陈昊瑜(1979-),男,总监,主要研究方向为存储器工艺研发 (E-mail:chenhaoyu@hlmc.cn).

Key words: Ferromagnet; In-memory computing; Magnetic field random access memory; Non-volatile memory; Computer architecture

0 引言

在传统的基于冯·诺依曼架构的计算机系统中,处理器和存储器在物理空间上是分离的,两者之间通过数据总线进行信息的交互[1-4]。为匹配处理器的高速性能,存储器引入了"分层架构",由靠近处理器端到远离处理器端依次为寄存器、多级高速缓存、主存和硬盘。靠近处理器的寄存器速度快,但存储密度低;而远离处理器的硬盘速度慢很多,但存储密度高。以个人计算机为例,三级高速缓存一般在几十 MB 的量级,存储介质是静态随机存储器(Statistic Random Access Memory, SRAM);主存采用动态随机存储器(Dynamic Random Access Memory, DRAM),容量在几十 GB 量级;速度更慢但是容量更大的硬盘则在几个 TB 量级。在速度上,SRAM 要比 DRAM 快上 100 倍左右,DRAM 更是比硬盘快上 1 000 倍。不难发现,计算系统的性能不仅取决于处理器的性能,同时也受到连接存储器的总线带宽的约束。此外,数据在处理器和存储器之间的频繁迁移也导致了比计算本身更大的能耗。这一挑战被称为"冯诺依曼架构瓶颈",在最近高性能计算需求应用场景中越发明显。

近年来蓬勃发展的人工智能(Artificial Intelligence, AI)正是这样一个需要对海量数据进行计算的场景。庞大而复杂的人工神经网络每一层都需要大量的数据读取、运算和写回,这就需要处理器对存储器中的大量数据进行频繁访问[6-8],因此带来巨量的运行功耗。如表 1 所示,即使是图像识别这样的任务,传统计算机也需要 kW 级的功耗;而与之不同的是,在人脑中这样的任务仅需要消耗 20 W 的功耗。因此,模仿人脑高度并行且互连的神经元结构和突触,进行原位计算和存储的"存算一体"相关研究受到热烈关注。与传统的冯·诺依曼架构不同,原位进行数据存储与计算能大幅降低反复读写数据以及数据传输的功耗。不难理解,这其中的关键元件单元就是存储器。

表 1 基于"存算一体"的大脑与传统冯·诺依曼架构计算机的对比[5]

Table 1 Comparison of "in-memory computing" based human brain and conventional CMOS based computer[5]

	"存算一体"的大脑	传统计算机
运行速度	1 ms	1 ns
尺寸	1 ~ 10 μm	10 ~ 100 nm
可靠性	80%	>99.9999%
功耗	~20 W	$>>10^3$ W
架构	未知	冯·诺依曼
器件类型	神经元	晶体管
器件数量	10^{12}	10^{10}
信号类型	脉冲	数字

到目前为止,多种存储器介质被研究用于构建存算一体系统,包括基于电荷存储原理的传统存储器和基于电阻存储原理的新型存储器。传统存储器主要包括 SRAM[9-13]、DRAM[13-16]和 Flash[17-20]。其中 SRAM 和 DRAM 是易失性器件,频繁的刷新并不利于降低功耗。而 Flash 虽然是非易失性的,但是随着读写次数增加,浮栅氧化层会逐渐失效,反复读写可靠性很低。因此,各种基于电阻改变的新型存储器是实现存算一体的有效载体[21]。

这其中主要包括相变存储器(Phase Change Memory, PCM)[22-23]、电阻记忆存储器(Resistive Random Access Memory, RRAM)[24-27]和磁性随机存储器(Magnetic Random Access Memory, MRAM)[28-32]。PCM 和 RRAM 基于原子层级重构来改变阻值,优点是有较大的阻值窗口,而缺点则是读写速度和读写可靠性要劣于 MRAM[3-4]。MRAM 则是基于对电子"自旋"的控制,可以达到理论上的零静态功耗,同时具有高速和非易失性以及近乎无限的写入次数。表 2 对比了各种存储器件的性能参数,可以看出 MRAM 在速度、耐久性、功耗这些方面具有不可替代的优越性。因此,MRAM 是实现存算一体的理想存储器之一。

表 2　各种存储器性能参数对比[21]

Table 2 Comparison of various memory devices[21]

器件	尺寸	读延迟	写延迟	耐久性	静态功耗
HDD	N/A	5 ms	5 ms	$>10^{15}$	1 W
SLC Flash	4 ~ 6 F^2	25 μs	500 μs	10^4 ~ 10^5	0
DRAM	6 ~ 10 F^2	50 ns	50 ns	$>10^{15}$	刷新功耗
PCM	4 ~ 12 F^2	50 ns	500 ns	10^8 ~ 10^9	0
STT-MRAM	6 ~ 50 F^2	10 ns	50 ns	$>10^{15}$	0
ReRAM	4 ~ 10 F^2	10 ns	50 ns	$>10^{11}$	0

本综述将对 MRAM 的基本原理进行简要介绍,然后重点介绍如何利用它实现存算一体,最后总结并展望了 MRAM 在存算一体应用中的发展前景。

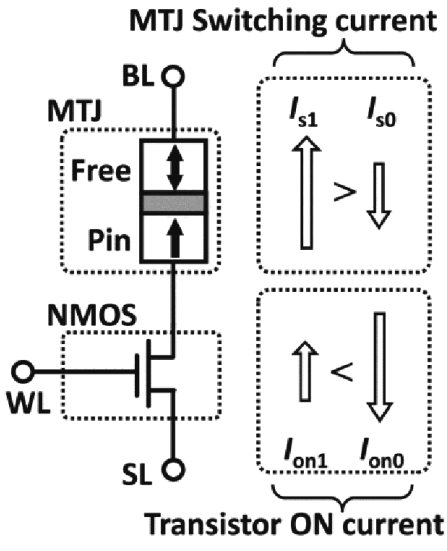


图 1　MRAM 存储单元结构示意图[28]

Fig. 1 Schematics of an MRAM bit-cell structure[28]

1　MRAM 器件结构概述

1.1　MRAM 的基本结构

如图 1 所示,MRAM 的核心结构由一个磁性隧道结(Magnetic Tunnel Junction, MTJ)和一个访问晶体管构成。MTJ 呈现“三明治”结构,两层磁性固定层和自由层之间夹着一层隧穿层。这其中,铁磁层材料一般使用 CoFeB,隧穿层材料则为 MgO。固定层的磁化方向是不变的,而自由层的磁化方向可以被改变。当固定层和自由层磁化方向一致时,称为“平行状态”,MTJ 的隧道磁阻(Tunnel Magnetoresistance, TMR)为低;当磁化方向不一致时,称为“反平行状态”,TMR 为高。数据的写入通过切换自由层的磁化方向实现,读取则通过使电流通过结来测量磁阻大小实现[28,31]。

访问晶体管的栅极与字线相连,形成“1T1M”的结构来实现存储单元的选择。由于 MTJ 翻转电流的不对称性,晶体管的写入驱动电流也有不对称性。

1.2　MRAM 的基本类型

早期的 MRAM 采用外部磁场来控制存储单元[32],在这种技术下读写易受干扰而且单元结构难以继续缩小尺寸。现代 MRAM 发展出了自旋转移扭矩 MRAM(Spin Transfer Torque-MRAM, STT-MRAM)、自旋轨道扭矩 MRAM(Spin Orbit Torque-MRAM, SOT-MRAM)和电压控制各向异性 MRAM(Voltage Controlled Magnetic Anisotropy-MRAM, VCMA-MRAM)几种类型。它们各自有不同的特点和应用场景,也是如今存算一体中使用的 MRAM 的基本类型。下面将简要介绍这几种主要的 MRAM。

1.2.1 STT-MRAM

STT-MRAM 的出现在推动 MRAM 商业化的进程中具有重要意义。如图 2(a)所示,STT 是通过将流过固定磁性层的自旋极化电流引导到自由磁性层,将自旋流中的角动量转移给自由层,实现其磁化取向的翻转。相比线圈产生的磁场,自旋流具有极高的电磁转换效率。此外,局域化产生的自旋流也被证明更适合存储单元的结构微缩。

截止目前,STT-MRAM 已经被验证能够取代三级缓存,而且相同容量下 STT-MRAM 的面积是 SRAM 的 43% 左右。更重要的是,STT-MRAM 的非易失性也消除了 SRAM 不工作时的泄露功耗,因此具有很高的能效比。需要指出的是,STT-MRAM 的速度不能进一步满足一级和二级缓存的要求。另外,STT 写入时会有很高密度的电流流过隧穿层 MgO,其产生的电压力会使得器件耐用性不可避免地降低。

1.2.2 SOT-MRAM

为了克服 STT 的速度和可靠性问题,SOT 技术应运而生。如图 2(b)所示,SOT-MRAM 是将 MTJ

堆叠在一层重金属衬底上,电流在重金属层平面注入,通过自旋霍尔效应或 Rashba 效应相互作用,实现自旋流中角动量从重金属层到磁性自由层的转移[33-35]。因此,这个结构的读写电流路径是分离的,这大大提高了器件的耐久性、速度和能效比。

显而易见地,作为一种三端器件,SOT-MRAM 的缺点是面积开销较大,这限制了其在高密度场景中的应用。

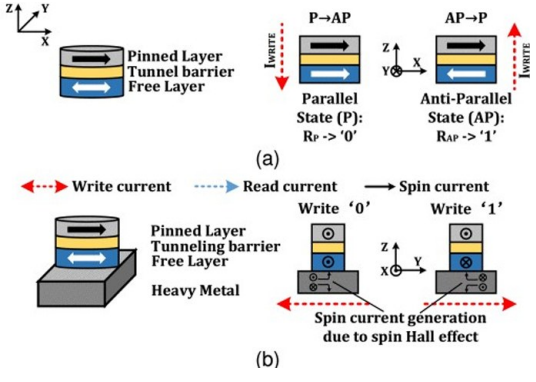


图 2 (a) 平行和反平行状态下的 MTJ 传统结构和 STT 磁环翻转示意图;(b) 堆叠重金属衬底的 MTJ 结构和 SOT 磁化翻转示意图[33]

Fig. 2 (a) Device structure of conventional Magnetic Tunnel Junction (MTJ) in parallel and anti-parallel states, with Spin-Transfer Torque (STT) switching scheme; (b) The stacking device structure of MTJ and heavy metal substrate, which uses spin-orbit torque induced magnetization switching scheme[33]

1.2.3 VCMA-MRAM

如图 3 所示,降低功耗的另一种策略是通过 VCMA 在自由层上施加电压,降低自由层磁化翻转所需克服的能量势垒,从而实现数据的写入操作[36]。由于在写入操作之前,存储单元要"预读取"其状态以产生对应的单极(P-AP 或 AP-P)写脉冲状态,因此其写入速度相对较慢。

另外,结合了 SOT 和 VCMA 的优点的电压控制型 SOT-MRAM 也在探索中,即通过 SOT 效应来切换自由层,VCMA 来选择 MTJ。通过两种技术的有机结合,能够提高 SOT 技术的密度,同时还有较高的翻转速度。

2 MRAM 存算一体应用研究现状

2.1 算术逻辑计算

算术逻辑计算包括布尔运算和全加运算,是其

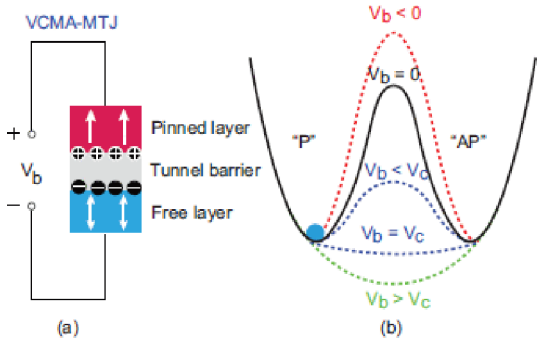


图 3 (a) VCMA-MTJ 器件示意图;(b) 不同电压对 MTJ 的"能垒"影响[36]

Fig. 3 (a) Schematics of a VCMA-MTJ device; (b) Illustration of the impacts of various bias voltages on the energy barrier of a MTJ device[36]

他计算形式的基础,因而也是基于 MRAM 的存算一体系统中研究最为广泛和深入的。算术逻辑的实现方法一般是基于外围读取电路做数字或模拟计算,也可以先对输入进行计算,将结果写入 MRAM 单元,实现计算和存值的同步[37-40]。

RIMPA[42]是一种基于自旋霍尔器件的双工模式存算一体架构,自旋存储器阵列的一部分可以作为非易失性的存储模块,其余部分可以配置为运算模块。这样一种可重构的架构[35]在增加不多的面积开销的情况下大大地减少了能耗。Fan 课题组[41]还提出了在增加额外逻辑电路的情况下,通过修改存储阵列的外围电路实现存算一体。如图 4 所示,计算结果通过单元读取操作读出,具体策略是将位线的电流与参考电流的比较转换为电压的比较,然后通过灵敏放大器读出。在此基础之上,该存算一体架构被应用在图像边缘提取场景中,结果表明,计算过程的能耗明显降低[33]。Fan 等人进一步提出了 IMCS2[43]和 HieIM[44]两种存算一体架构,两种架构都可以灵活配置在存储和计算双工模式下。

如图 5(a)所示,Zhang 等人[45]在 2018 年提出了一种基于 STT-MRAM 的"对偶参考"的存算一体电路架构。仿真结果在操作错误率、灵敏噪声容限、操作延迟和动态功耗方面验证了设计的鲁棒性和性能。在此基础上,该团队同年提出了另一种"互补参考"的电路架构[46],如图 5(b)所示,同样基于 STT-MRAM,仿真结果显示,对比"对偶参考"架构,该设计降低了 67.1% 的操作错误率,提高了 57.4% 的噪声容限,降低了 20.8% 的延迟以及 23.4% 的平

均动态功耗和 65.6% 的平均静态功耗。因此,“互补参考”架构进一步提高了存算一体系统的可靠性和低功耗性能。

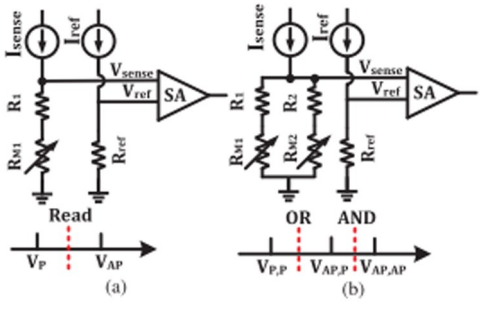
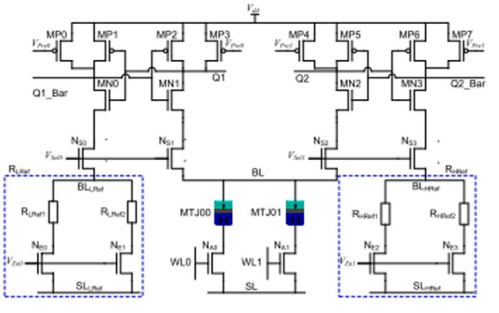
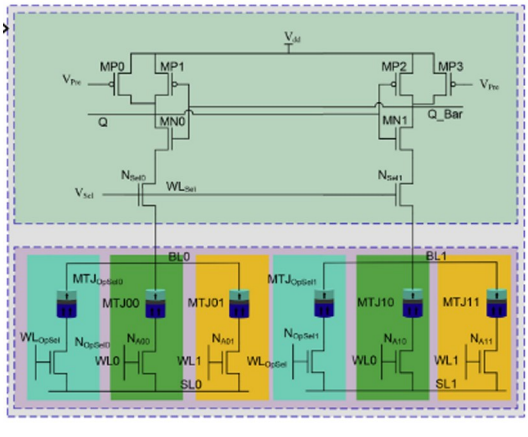


图 4　V_{sense} 与 V_{ref} 的比较(a) 存储器读取;(b) 存内逻辑操作[41]

Fig. 4 Comparison between V_{sense} and V_{ref} for (a) memory read; (b) in-memory logic operation[41]



(a)



(b)

图 5　(a)“对偶参考”电路图[45];(b)“互补参考”电路图[46]

Fig. 5 (a) Schematic of DualRef CIM cell[45]; (b) Schematic of ComRef CIM cell[46]

在传统的存算一体电路架构中,通常采用一个固定电流作为参考值,而在不同数量字线选中下的噪声容限一般不同,容易导致输出错误。Chen 等人[47]设计了一个随噪声容限自动匹配参考电流的方案,结果显示,该方案支持最多 16 条字线同时选中。Kim 等人[48]则尝试以尽可能少的外围电路实现高能效比的基于 SOT-MRAM 的存算一体系统。通过将同时选中的字线降到一条实现了能耗的降低。Louis 等人[49]提出了基于忆阻逻辑的 STT-MRAM 存算一体阵列,通过施加合适电压得到使 MRAM 改变存储数据的电流。

Cai 等人[50]于 2020 年提出一种结合了 SOT 和 VCMA 两种控制方式混合的 MRAM 存算一体结构,该设计采用先计算然后写入存储单元的“只写”模式实现存算一体目标。图像处理应用的实验结果显示该混合设计比采用单种控制方式有更高的能效比,同时减少了外围电路面积。

Raghunathan 等人[51]于 2020 年提出一种谷间耦合的自旋霍尔存储器(VSH-MRAM),该设计有两个主要特点,一个是应用垂直磁化各向异性翻转磁化方向,另一个是应用调制电荷电流与自旋电流比的集成门,实现了一个紧凑高能效的存算一体架构。仿真实验结果表明,该设计不仅降低了读写能耗,也改善了读写延迟。

如图 6 所示,北航的 Zhao 团队于 2020 年提出一种具有时序特性的高速低功耗 TST-MRAM 存算一体方案[52],通过在时序上反映的位线电压变化实现逻辑操作,因此 D 触发器被用来缓存数据。研究通过一个四位加法器验证显示了该方案在速度、能耗和性能方面的优势。该团队在 2021 年继续在 MRAM 存算一体方面做了两项研究。一种是基于 STT-MRAM 的存算一体架构,该设计[53]结合了读存算一体(RLM)和写存算一体(WLM)模式,既发挥了 RLM 的高速特性又发挥了 WLM 的完整特性,在简单布尔逻辑运算的基础上实现了全加运算操作,仿真结果显示了其高速和低功耗优势。另一项工作同样基于 STT-MRAM,在较少的电路修改的条件下,实现了并行进位和加法的多位全加器操作[54]。另外,课题组也对基于 STT-MRAM 的通用计算架构做了相关研究,期望建立像 CPU 一样的通用计算平台[55]。

来自印度的几个团队也做了 MRAM 存算一体相关工作。Shreya 等人[56]研究了基于电压控制的

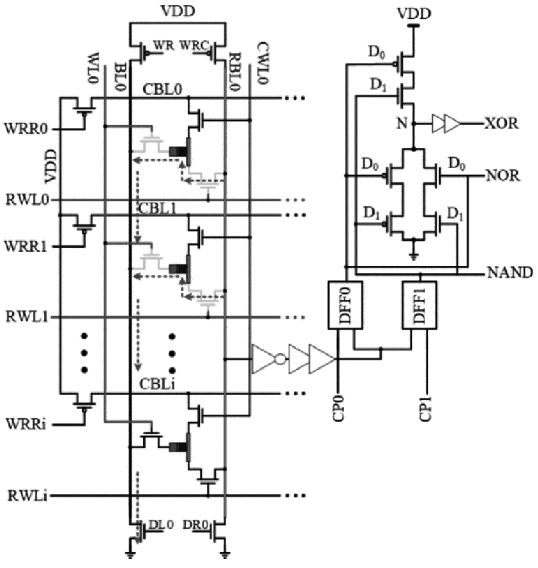


图 6 基于时序的存内计算结构[52]

Fig. 6 TIMC structure[52]

SOT-MARM 存算一体应用,设计了基于此器件的全加器,在逻辑运算和数据搬运的能耗上分别实现了 53.98% 和 2.7% 的降低,同时该工作研究了最佳性能时的电压脉冲值在 1.2 V,脉冲宽度在 1 ns。最后,还对寄生变量进行了蒙特卡洛分析。Nehra 等人[57]则设计了一种基于 STT 和 SOT 混合机理的串行三阶单元结构,即实现一个单元存储 3 bit 的数据,这样可以大大提高存储器密度。在该设计中,写入数据最多需要 2 步,大多数情况下写入数据只需 1 步完成,读取也只需 1 步可实现超快读取。基于此器件的布尔逻辑运算电路和全加器电路显示在相同功耗性能的条件下,该设计需要的晶体管数量比基于自旋霍尔效应器件的全加器要少 33%。除此之外,Monga 等人[58]研究了双工模式下的存算一体架构。在 International Solid-State Circuits Conference (ISSCC20)会议上,Chang 等人[59]提出了一种针对移动安全应用的高带宽 STT-MRAM 宏设计,该设计读取带宽达到了 42.6 GB/s。

表 3 总结了近几年的研究工作的对比,可知在实现“与(AND)”操作且速度增加 1 倍时,单次运算操作的功耗降低了 37.5%。对于稍复杂的“异或(XOR)”和全加计算(Full Adder)来说,在保持纳秒级的速度的条件下,功耗在几十至几百微瓦的水平。可以看到,为了进一步提高速度降低功耗以及系统集成度,多种控制方式结合是一个趋势,例如,STT 与 SOT 的结合[57],使得 MRAM 在读写电流减小的同时,单元密度得到提升。同时,MRAM 存算一体系统也能实现哈希计算[59],可进一步应用在数据加密场景中。

表 3 几项工作的对比

Table 3 Comparison of several works

	时间	类型	速度(ns)	功耗(μW)	操作
文献[42]	2017	STT	2.0	9.40	AND
文献[43]	2018	STT	1.0	14.6	AND
文献[44]	2018	STT	4.0	5.90	AND
文献[48]	2019	SOT	N/A	5.10×10^6	Alexnet
文献[59]	2020	STT	2.9	79.0	Hash
文献[55]	2020	STT	31	205	XOR
文献[58]	2021	STT+SOT	6.0	128	SUM
文献[57]	2021	STT+SOT	1.228	49.0	Full Adder

2.2 神经网络计算

MRAM 的高速、低功耗、可靠性高的优点,吸引了研究者对其在人工神经网络计算中的应用实现进行研究。由于 MRAM 具有高阻和低阻两种状态,适合存储一个比特的数据,基于 MRAM 的二进制神经网络得到了广泛关注。

Fan 等人[60]于 2017 年用基于 SOT-MRAM 的存算一体架构探索了一种二进制卷积神经网络(Binary convolutional neural network)加速器,特点在于使用了一个简单的数字处理单元。在 2020 年,团队还提出了 MRIMA[61],一种基于 MRAM 的存内计算加速器(如图 7 所示),能够在一个时钟周期内实现布尔逻辑函数。该设计应用在二进制权重和低位宽卷积神经网络,相比于 ASIC 芯片表现出 1.7 倍的能效比和 11.2 倍的速度提升,相比于 DRAM 解决方案节能 1.8 倍,速度快 2.4 倍。所以该加速器非常适用于低功耗数据加密场景。

Wei 与 Zhao 等人[62]基于多阶 STT-MRAM 的存算一体架构实现了 BCNN 的计算以降低功耗。Xie 与 Zhao 等人提出一种基于非易失性存储器的 DASM[63],研究了同时在读取和写入时提高数据流的并行性。同年,他们还提出了一种 PXNOR-BNN,

即通过在 SOT-MRAM 预置异或运算来加速二进制神经网络计算[64]。实验结果显示,该方案的性能与基于读取的 SOT-MRAM 方案相当。Zhao 团队还研究了一种基于 SOT-MRAM 的同时支持读取写入的存算一体架构[65]。因为 SOT-MRAM 出色的写入性能,该方案的控制便基于写入操作,以此提高数据密集型的卷积神经网络的并行计算性。在最近的 2021 年,Zhao 团队提出了一种电压控制型的 SOT-MARM 器件来为 BNN 构建存算一体系统[66],旨在实现并行编程和计算。该器件将多个 MTJ 堆在一个重金属层上共享一个 SOT 写入电流,计算时就像正常读写操作一样(图 8)。在 40 nm 节点下,该 BNN 网络实现了每比特 4 fJ 的功耗,写入时间在 2 ns左右,读取时间在 0.36 ~ 1.5 ns 之间。

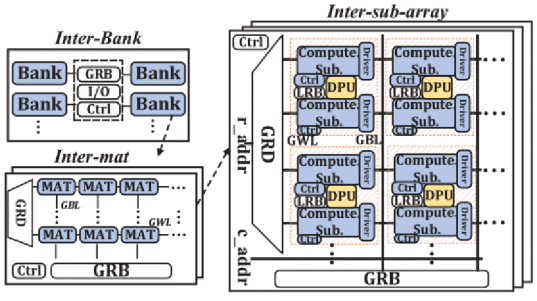


图 7　MRIMA 结构[61]

Fig. 7 MRIMA architecture[61]

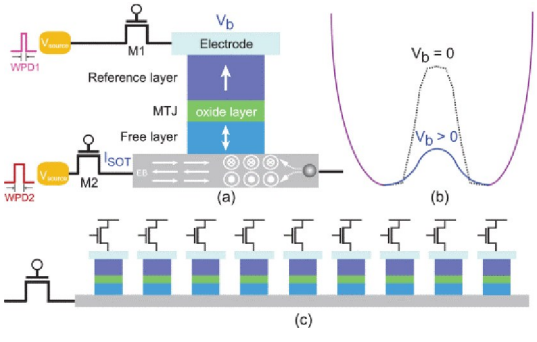


图 8　(a) VC-SOT MTJ 结构;(b) 电压偏置下能量变化调制原理;(c) 共享重金属层的多个 MTJ 堆叠结构[66]

Fig. 8 (a) VC-SOT MTJ structure; (b) VC-SOT principle for MTJ switching by modulating the energy barrier with a bias voltage; (c) Multiple MTJs are stacked on a heavy metal [66]

Kang 等人[67]探索了更深节点下(28 nm 和 7 nm)的基于 STT-MRAM 的存算一体系统,评估了当工艺节点从 65 nm 过渡到 7 nm 时 CNN 推理计算在基于 MRAM 和基于 SRAM 的存算一体架构中面积、漏功耗、能耗和延迟等性能参数。实验结果表明,基于 STT-MRAM 的存算一体系统在 65 nm 节点缩小到 7 nm 时仍能表现出比 SRAM 系统小 5 倍的面积、20 倍的漏功耗和 7 倍的能耗,毫无疑问,STT-MRAM 是为物联网和神经网络应用提高能效比和高密度的良好解决方案。

Wood 等人[68]研究了在有限玻尔兹曼机中,MRAM 工艺变量对其电压特定的影响,提出了基于 Python 脚本的仿真环境框架,分析了在器件尺度工艺参数的变化对机器学习应用的精确度的影响,验证了其仿真框架的可行性。

Kim 等人[69]研究了如何提高多行字线选中的并行能力。字线选中并行能力受到 MRAM 器件的较小的 I_H/I_L 的限制,结果使得 BNN 更容易受到器件变量的影响。他们改进算法和硬件架构,应用一个输入管理方案,以减少训练时输入数字“1”的数目,同时采用重训练方案提高了 BNN 对 MRAM 器件变量的鲁棒性。实验结果显示,多条字线并行选中的能力提高,从而提高了带宽和能效。

来自复旦大学的团队研究了一种脉冲神经网络[70],在网络中使用了多个二进制 MRAM 单元来存储多位定点权重值,因为多值 MRAM 仍具有可靠性和电路复杂性方面的问题。

综上所述,MRAM 的存算一体系统将神经网络计算在存储单元直接进行,避免了数据的频繁迁移。存内计算本身具有的高并行性也给神经网络计算带来了便利。

MRAM 由于本身具有高阻和低阻两个阻态,所以特别适合用于二进制神经网络计算。其次,MRAM 的低功耗和无限写入次数也适合于非常消耗计算资源的神经网络计算任务。因此,MRAM 的存算一体系统更加适合于神经网络计算场景。在 AI 推理芯片中,MRAM 的存算一体方案具有广阔的前景。

2.3　问题与挑战

MRAM 的存算一体化研究已经取得了长足的进步,但是挑战和问题是仍然存在的[39,71]。

1)MRAM 有限的隧道磁阻比限制了位线模拟信号计算结果的精确性。因此,差值明显的两个阻值态对于存算一体实现是非常必要的。但是这个差

值易受工艺、电压、温度等变量影响。

2)MRAM 读写失败的可能性在先进工艺节点下有所增加。工艺过程、电压温度等等因素都会影响 MTJ 物理参数的一致性,因此邻近单元之间的噪声容限易受影响,限制了计算阵列的规模。

3)对于大规模的存算一体阵列,位线和外围电路的面积和能耗将不能被忽略。当大量字线单元被激活,计算形成的位线电流密度是相当大的,需要更宽的金属导线。因此,仔细设计外围电路显得非常重要,而且 CMOS 磁性混合电路的优化需要综合考虑。

3 结论

新兴器件的快速发展为后摩尔时代的高速低功耗计算提供了可能的解决方案。其中,MRAM 作为一种自旋电子器件,以其非易失性、高速低功耗特性、高可靠性,在非冯·诺依曼架构的存算一体系统中有着巨大的应用潜力。MRAM 的存算一体可以实现完全的算术逻辑计算,并且可以工作在存储和计算两种模式下,具备非常大的灵活性。存算一体化的 MRAM 具有高并行性和高能效比,尤其适合于神经网络计算,特别是二进制神经网络,为高速低功耗的 AI 应用提供了有力的支撑。

参考文献:

[1]李雅琪, 温晓君. 存算一体化的发展现状挑战与对策建议 [J]. 互联网经济, 2020, 04: 15-17.

[2]郭昕婕, 王绍迪. 端侧智能存算一体芯片概述 [J]. 微纳电子与智能制造, 2019, 1(02): 72-82.

[3]Yin L, Cheng R, Wen Y, et al. Emerging 2D Memory Devices for In-Memory Computing [J]. Advanced Materials, 2021, 33(29): e2007081.

[4]Sebastian A, Le Gallo M, Khaddam-Aljameh R, et al. Memory devices and applications for in-memory computing [J]. Nature Nanotechnology, 2020, 15(7): 529-544.

[5]Del Valle J, Ramírez J G, Rozenberg M J, et al. Challenges in materials and devices for resistive-switching-based neuromorphic computing [J]. Journal of Applied Physics, 2018, 124(21): 211101.

[6]Chang L, Li C, Zhao X, et al. Trend of Emerging Non-Volatile Memory for AI Processor [C]. // 2021 18th International SoC Design Conference (ISOCC). 2021: 223-224.

[7]Zhao Y, Fan Z, Du Z, et al. Machine Learning Computers With Fractal von Neumann Architecture [J]. IEEE Transactions on Computers, 2020, 69(7): 998-1014.

[8]Zanotti T, Puglisi F M, Pavan P. Smart Logic-in-Memory Architecture for Low-Power Non-Von Neumann Computing [J]. IEEE Journal of the Electron Devices Society, 2020, 8: 757-764.

[9]曾剑敏, 张 章, 虞志益, 等. 基于 SRAM 的通用存算一体架构平台在物联网中的应用 [J]. 电子与信息学报, 2021, 43(06): 1574-1586.

[10]Hsu Y T, Yao C Y, Wu T Y, et al. A High-Throughput Energy-Area-Efficient Computing-in-Memory SRAM Using Unified Charge-Processing Network [J]. IEEE Solid-State Circuits Letters, 2021, 4: 146-149.

[11]Si X, Tu Y N, Huang W H, et al. A Local Computing Cell and 6T SRAM-Based Computing-in-Memory Macro With 8-b MAC Operation for Edge AI Chips [J]. IEEE Journal of Solid-State Circuits, 2021, 56(9): 2817-2831.

[12]Su J W, Si X, 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, 2021.

[13]Zhang X, Mohan V, Basu A. CRAM: Collocated SRAM and DRAM With In-Memory Computing-Based Denoising and Filling for Neuromorphic Vision Sensors in 65 nm CMOS [J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2020, 67(5): 816-820.

[14]Choi H, Lee Y, Kim J J, et al. A Novel In-DRAM Accelerator Architecture for Binary Neural Network [C]. Proceedings for IEEE Cool Chips. 2020.

[15]Xie S, Ni C, Sayal A, et al. eDRAM-CIM: Compute-In-Memory Design with Reconfigurable Embedded-Dynamic-Memory Array Realizing Adaptive Data Converters and Charge-Domain Computing [C]. // 2021 IEEE International Solid- State Circuits Conference (ISSCC). 2021, 64: 248-+.

[16]Yu C, Yoo T, Kim H, et al. A Logic-Compatible eDRAM Compute-In-Memory With Embedded ADCs for Processing Neural Networks [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(2): 667-679.

[17]丁士鹏, 黄 鲁. 基于 NOR Flash 的存算一体模拟乘加电路设计 [J]. 信息技术与网络安全, 2021, 40(06): 69-74.

[18]徐伟民, 黄 鲁, 蒋明峰. 基于 NOR Flash 的卷积计算单元的设计 [J]. 信息技术与网络安全, 2020, 39

(05): 63-68.

[19] Feng Y, Wang F, Zhan X, et al. Flash memory based computing-in-memory system to solve partial differential equations [J]. Science China Information Sciences, 2020, 64(6): 169401.

[20] Lue H T, Hu H W, Hsu T H, et al. Design of Computing-in-Memory (CIM) with Vertical Split-Gate Flash Memory for Deep Neural Network (DNN) Inference Accelerator [C]. // 2021 IEEE International Symposium on Circuits and Systems (ISCAS). 2021: 1-4.

[21] Mittal S, Vetter J S. A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2016, 27(5): 1537-1550.

[22] Zhang H, Huang B, Zhang Z, et al. On-Chip Photonic Synapses Based on Slot-Ridge Waveguides With PCMs For In-Memory Computing [J]. IEEE Photonics Journal, 2021, 13(2): 1-13.

[23] Fong S W, Neumann C M, Wong H S P. Phase-Change Memory—Towards a Storage-Class Memory [J]. IEEE Transactions on Electron Devices, 2017, 64(11): 4374-4385.

[24] 徐丽莹, 杨玉超, 黄　如. 基于忆阻器的非易失逻辑研究前沿 [J]. 中国基础科学, 2019, 21(02): 1-11+27+63.

[25] 张　章, 李　超, 韩婷婷, 等. 基于忆阻器的感存算一体技术综述 [J]. 电子与信息学报, 2021, 43(06): 1498-1509.

[26] He W, Yin S, Kim Y, et al. 2-Bit-per-Cell RRAM based In-Memory Computing for Area-/Energy-Efficient Deep Learning [J]. IEEE Solid-State Circuits Letters, 2020, 3: 194-197.

[27] Wei W C, Jhang C J, Chen Y R, et al. A Relaxed Quantization Training Method for Hardware Limitations of Resistive Random Access Memory (ReRAM)-Based Computing-in-Memory [J]. IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 2020, 6(1): 45-52.

[28] Koike H, Tanigawa T, Watanabe T, et al. Review of STT-MRAM circuit design strategies, and a 40-nm 1T-1MTJ 128Mb STT-MRAM design practice [C]. // 2020 IEEE 31st Magnetic Recording Conference (TMRC). 2020: 1-2.

[29] Shadman A, Zhu J G. High-speed STT MRAM incorporating antiferromagnetic layer [J]. Applied Physics Letters, 2019, 114(2): 022403.

[30] Senni S, Delobelle T, Coi O, et al. Embedded Systems to High Performance Computing using STT-MRAM [C]. Design Automation and Test in Europe Conference and Exhibition. 2017: 536-541.

[31] Sbiaa R, Piramanayagam S N. Recent Developments in Spin Transfer Torque MRAM [J]. Physica Status Solidi-Rapid Research Letters, 2017, 11(12): 1700163.

[32] Andre T, Alam S M, Gogl D, et al. ST-MRAM Fundamentals, Challenges, and Outlook [C]. IEEE International Memory Workshop. 2017: 161-164.

[33] He Z, Zhang Y, Angizi S, et al. Exploring a SOT-MRAM Based In-Memory Computing for Data Processing [J]. IEEE Transactions on Multi-Scale Computing Systems, 2018, 4(4): 676-685.

[34] Rizk M, Diguet J P, Onizawa N, et al. NoC-MRAM Architecture for Memory-Based Computing: database-search case study [C]. IEEE International New Circuits and Systems Conference. 2017: 309-312.

[35] Parveen F, Angizi S, He Z Z, et al. Low Power In-Memory Computing based on Dual-Mode SOT-MRAM [C]. // 2017 IEEE/Acm International Symposium on Low Power Electronics and Design (Islped). 2017.

[36] Kang W, Chang L, Zhang Y G, et al. Voltage-Controlled MRAM for Working Memory: Perspectives and Challenges [C]. Design Automation and Test in Europe Conference and Exhibition. 2017: 542-547.

[37] Cai H, Chen J, Zhou Y, et al. Toward Energy-Efficient STT-MRAM Design With Multi-Modes Reconfiguration [J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2021, 68(7): 2633-2639.

[38] Sethuraman S, Tavva V K, Srinivas M B. Techniques to Improve Write and Retention Reliability of STT-MRAM Memory Subsystem [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2021.

[39] Cai H, Liu B, Chen J, et al. A survey of in-spin transfer torque MRAM computing [J]. Science China Information Sciences, 2021, 64(6): 160402.

[40] Zhang H, Kang W, Wu B, et al. Spintronic Processing Unit Within Voltage-Gated Spin Hall Effect MRAMs [J]. IEEE Transactions on Nanotechnology, 2019, 18: 473-483.

[41] He Z Z, Angizi S, Parveen F, et al. High Performance and Energy-Efficient In-Memory Computing Architecture based on SOT-MRAM [C]. Proceedings of the IEEE/Acm International Symposium on Nanoscale Architectures (Nanoarch 2017). 2017: 97-102.

[42] Angizi S, He Z, Parveen F, et al. RIMPA: A New

Reconfigurable Dual-Mode In-Memory Processing Architecture with Spin Hall Effect-Driven Domain Wall Motion Device [C]. // 2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). 2017: 45-50.

[43] Parveen F, Angizi S, He Z, et al. IMCS2: Novel Device-to-Architecture Co-Design for Low-Power In-Memory Computing Platform Using Coterminous Spin Switch [J]. IEEE Transactions on Magnetics, 2018, 54(7): 1-14.

[44] Parveen F, He Z, Angizi S, et al. HielM: Highly flexible in-memory computing using STT MRAM [C]. Asia and South Pacific Design Automation Conference Proceedings. 2018: 361-366.

[45] Zhang L, Kang W, Cai H, et al. A Robust Dual Reference Computing-in-Memory Implementation and Design Space Exploration Within STT-MRAM [C]. // 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). 2018: 275-280.

[46] Zhang L, Deng E, Cai H, et al. A high-reliability and low-power computing-in-memory implementation within STT-MRAM [J]. Microelectronics Journal, 2018, 81: 69-75.

[47] Chen T L, Tseng W T. Automatic Reference Current Architecture in Computing in Memory by MRAM [C]. Proceedings of the 2019 IEEE Eurasia Conference on IoT, Communication and Engineering (Ecice). 2019: 86-88.

[48] Kim K, Shin H, Sim J, et al. An Energy-efficient Processing-in-memory Architecture for Long Short Term Memory in Spin Orbit Torque MRAM [C]. // 2019 IEEE/Acm International Conference on Computer-Aided Design (Iccad). 2019.

[49] Louis J, Hoffer B, Kvatinsky S. Performing Memristor-Aided Logic (MAGIC) using STT-MRAM [C]. // 2019 26th IEEE International Conference on Electronics, Circuits and Systems (Icecs). 2019: 787-790.

[50] Cai H, Jiang H, Zhou Y, et al. Interplay Bitwise Operation in Emerging MRAM for Efficient In-memory Computing [J]. CCF Transactions on High Performance Computing, 2020, 2(3): 282-296.

[51] Thirumala S K, Hung Y T, Jain S, et al. Valley-Coupled-Spintronic Non-Volatile Memories With Compute-In-Memory Support [J]. IEEE Transactions on Nanotechnology, 2020, 19: 635-647.

[52] Wang J K, Zhang Y, Lian C Y, et al. Efficient Time-Domain In-Memory Computing Based on TST-MRAM [C]. IEEE International Symposium on Circuits and Systems. 2020.

[53] Wang C, Wang Z, Zhang Y, et al. Computing-in-Memory Paradigm Based on STT-MRAM with Synergetic Read/Write-Like Modes [C]. // 2021 IEEE International Symposium on Circuits and Systems (ISCAS). 2021: 1-5.

[54] Wang C, Wang Z, Wang G, 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.

[55] Pan Y, Jia X, Cheng Z, et al. An STT-MRAM based reconfigurable computing-in-memory architecture for general purpose computing [J]. CCF Transactions on High Performance Computing, 2020, 2(3): 272-281.

[56] Shreya S, Jain A, Kaushik B K. Computing-in-memory using voltage-controlled spin-orbit torque based MRAM array [J]. Microelectronics Journal, 2021, 109: 104943.

[57] 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.

[58] Monga K, Chaturvedi N, Gurunarayanan S. A Dual-Mode In-Memory Computing Unit Using Spin Hall-Assisted MRAM for Data-Intensive Applications [J]. IEEE Transactions on Magnetics, 2021, 57(4): 1-10.

[59] Chang T C, Chiu Y C, Lee C Y, et al. A 22nm 1Mb 1024b-Read and Near-Memory-Computing Dual-Mode STT-MRAM Macro with 42.6GB/s Read Bandwidth for Security-Aware Mobile Devices [C]. IEEE International Solid State Circuits Conference. 2020: 224-+.

[60] Fan D, Angizi S. Energy Efficient In-Memory Binary Deep Neural Network Accelerator with Dual-Mode SOT-MRAM [C]. // 2017 IEEE International Conference on Computer Design (ICCD). 2017: 609-612.

[61] 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.

[62] Pan Y, Ouyang P, Zhao Y, et al. A Multilevel Cell STT-MRAM-Based Computing In-Memory Accelerator for Binary Convolutional Neural Network [J]. IEEE Transactions on Magnetics, 2018, 54(11): 1-5.

[63] Chang L, Ma X, Wang Z, et al. DASM: Data-Streaming-Based Computing in Nonvolatile Memory Architecture for Embedded System [J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2019, 27(9): 2046-2059.

[64] Chang L, Ma X, Wang Z, et al. PXNOR-BNN: In/With

Spin-Orbit Torque MRAM Preset-XNOR Operation-Based Binary Neural Networks [J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2019, 27(11): 2668-2679.

[65] Chang L, Wang Z H, Zhang Y G, et al. Ultra-fast and Energy-efficient Write-Computing Operation for Neuromorphic Computing [C]. International SoC Design Conference. 2019: 140-141.

[66] Wang H, Kang W, Pan B, et al. Spintronic Computing-in-Memory Architecture Based on Voltage-Controlled Spin-Orbit Torque Devices for Binary Neural Networks [J]. IEEE Transactions on Electron Devices, 2021, 68(10): 4944-4950.

[67] Shi Y, Oh S, Huang Z, et al. Performance Prospects of Deeply Scaled Spin-Transfer Torque Magnetic Random-Access Memory for In-Memory Computing [J]. IEEE Electron Device Letters, 2020, 41(7): 1126-1129.

[68] Wood P, Pourmeidani H, Demara R F. Modular Simulation Framework for Process Variation Analysis of MRAM-based Deep Belief Networks [C]. // 2020 Southeast Conference. 2020: 1-2.

[69] Ahn D, Oh H, Kim H, et al. Maximizing Parallel Activation of Word-Lines in MRAM-Based Binary Neural Network Accelerators [J]. IEEE Access, 2021, 9: 141961-141969.

[70] Wang Y, Wu D, Wang Y, et al. A Low-Cost Hardware-Friendly Spiking Neural Network Based on Binary MRAM Synapses, Accelerated Using In-Memory Computing [J]. Electronics, 2021, 10(19).

[71] Shao Q, Wang Z, Zhou Y, et al. Spintronic memristors for computing [J]. arXiv pre-print server, 2021.