忆阻器及其存算一体应用研究进展

江之行,席悦,唐建石*,高滨,钱鹤,吴华强*

清华大学集成电路学院,集成电路高精尖创新中心,北京 100084

摘要　深度学习的飞速发展带来了巨大的算力需求,然而基于存算分离的“冯·诺依曼架构”的传统硅基芯片面临着“存储墙”等问题,芯片算力增长逐渐陷入瓶颈。为了解决这个矛盾,研究人员从生物大脑的工作模式得到启发,提出了基于忆阻器的存算一体架构。这种全新的架构在处理神经网络等任务时在能效和速度上较“冯·诺依曼架构”有望实现几个数量级的提升,是实现超低功耗、超高算力计算芯片的最有潜力的技术路线之一。本文综述了各种类型忆阻器的工作机理与最新进展,对比了国内外研究团队的器件研究进展;综述了基于忆阻器的存算一体芯片在神经网络、信号处理和机器学习等方向的应用演示的研究进展;总结了基于忆阻器的存算一体芯片目前面临的挑战,并提出中国在该领域进一步发展的建议。

关键词　忆阻器;类脑计算;存算一体;神经网络;信号处理

自20世纪中叶,受大脑神经元结构与特性启发,研究人员先后提出了人工神经网络(artificial neural network, ANN)[1]和脉冲神经网络(spiking neural network, SNN)[2]等多种算法。近10年来,深度学习(deep learning, DL)的相关研究成果出现了爆炸式增长[3],实际应用也逐步落地,已经彻底改变了人类的日常生活。与此同时,深度学习的迅速发展也给芯片带来巨大的算力需求,这种需求平均每3~4个月就会翻一番[4],远远超过了摩尔定律[5]的发展速度。更具挑战的是,近年来芯片算力增长正逐渐陷入瓶颈[6],算力的需求与芯片所能提供的算力之间出现了尖锐的矛盾。

芯片算力增长逐渐变缓主要有以下几个原因:在器件角度,漏电流的影响使得晶体管尺寸微缩变得愈发困难[7-8];在制造角度,由于逐渐逼近物理极限,先进制程芯片生产的成本越来越难以控制[9]。这使得单位面积晶体管密度增长放缓[10]。此外,在架构角度,目前芯片多采用存算分离的“冯·诺依曼架构”,这使得在执行以大数据为核心的计算任务时,数据会在计算单元和存储单元之间来回搬移,

收稿日期:2022-09-02;修回日期:2023-01-12

基金项目:科技部重大项目(2021ZD0201205,2022ZD0210200);国家自然科学基金委重点项目(92264201,92064001)

作者简介:江之行,博士研究生,研究方向为新型存储器,电子信箱:jiangzx22@mails.tsing.edu,cn;唐建石(通信作者),副教授,研究方向为新型存储器与类脑计算,电子信箱:jtang@tsinghua.edu.cn;吴华强(共同通信作者),教授,研究方向为忆阻器与存算一体技术,电子信箱:wuhq@tsinghua.edu.cn

引用格式:江之行, 席悦, 唐建石, 等. 忆阻器及其存算一体应用研究进展[J]. 科技导报, 2024, 42(2): 31-49; doi: 10.3981/j.issn.1000-7857.2024.02.004

由此导致了一些“冯·诺依曼瓶颈”问题[11]:首先,访问存储单元的速度远低于计算单元的运算速度,并且差距正在越来越大,即存在“存储墙”问题[12];其次,目前计算机体系中的存储器通常具有由不同存储介质组成的层级结构[13],存储容量越大的层级访问延时也大,这就导致了数据跨层级传输时存在巨大的时间损耗。

对比目前各种计算芯片动辄成百上千瓦的功耗,人类大脑只需要约20 W就可以实现灵活高性能的计算[14],这启发人们重新关注大脑的结构与特性。与计算机传统的冯·诺依曼架构中存算分离的模式不同,实际生物大脑中的神经元既参与计算又参与存储[15],不存在上述的各种存算分离局限,受此启发,研究人员设计出了存算一体架构。这种全新的架构在处理特定任务时可以在能效和速度上较“冯·诺依曼架构”实现几个数量级的提升[16],是实现超低功耗、超高算力计算芯片的最具潜力的路径之一[10]。

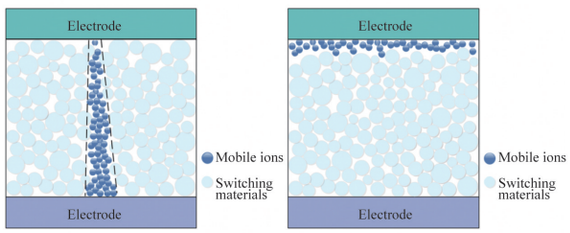
存算一体架构需要一种既可以作为计算单元又可以作为存储单元的器件,忆阻器[17-18]的特性刚好与之契合。忆阻器是基于器件阻值来存储信息的,并且其电阻可以通过外加的激励实现连续、可逆的调节,在去掉激励后仍可以保持当前电阻状态[19],这种特性与生物神经元突触非常相似,因此忆阻器也被称为“电子突触器件”[20]。忆阻器的阵列集成通常以交叉阵列(crossbar)的形式实现[21],如果将矩阵中的元素一一映射到对应忆阻器电导值,输入输出信号分别穿过交叉阵列行列中间的节点,那么仅基于欧姆定律、基尔霍夫电流定律等物理定律,就可以自然地完成向量矩阵乘运算。这一特性使得基于忆阻器的存算一体架构既可以运行各类神经网络,也可以运行包含向量矩阵乘法的其他算法,这为存算一体芯片未来的广泛应用打下了基础。此外,除了作为“电子突触器件”,近些年的研究也拓展了忆阻器在“电子神经元器件”“电子树突器件”等仿生器件中的应用[22-25]。受篇幅影响,本文仅聚焦于目前忆阻器最为主要的“电子突触器件”应用的研究现状。

1 忆阻器器件研究进展

忆阻器通常由金属-介质-金属(metal-insulator-metal, MIM)的夹层结构组成,包含2层电极和1层忆阻功能层,其电学特性往往与电极层和功能层密切相关。依据不同的工作机理,广义上的忆阻器主要包含以下几种类型:阻变随机存储器(resistive random access memory, RRAM),相变存储器(phase-change memory, PCM),磁随机存储器(magnetic random access memory, MRAM),铁电存储器(ferroelectric memory)。本节将依此分类对这4类器件的结构、特性与研究进展展开讨论,并在最后一节着重对国内外的研究现状进行了对比。

1.1 阻变随机存储器

阻变随机存储器依照其阻变机理主要可分为导电细丝型和非导电细丝型(界面型)2类(图1)[26]:导电细丝型器件依赖于阻变介质中导电通道的形成与断裂,其主要包括氧空位通道型阻变器件和金属通道型阻变器件;非导电细丝型器件则依赖于阻变介质的体效应,或者阻变介质与电极之间的界面效应。前者往往具备更快的操作速度及良好的非易失性[27],因此相较于后者其更适合于作为需要数据保持的“电子突触器件”。后者通常具备天然的模拟阻变特性,但由于保持特性较差[19],其往往作为动态忆阻器应用于例如“电子树突器件”等领域。下面重点讨论2类导电细丝型器件的阻变机理与研究进展。



(a) 导电细丝型 (b) 非导电细丝型

图1 2种不同阻变机理示意

1.1.1 氧空位通道型阻变器件

氧空位通道型阻变器件又被称为金属氧化物

阻变存储器(metal-oxide RRAM, OxRRAM),或价变存储器(valence change memory, VCM)。在其MIM结构中,通常中间层介质为绝缘的过渡族金属氧化物。现阶段业内关于OxRRAM的阻变机理仍存在争议,较为普遍的看法是其阻变过程依赖于介质层内形成的氧空位导电细丝[28-32]。当导电细丝连接时器件呈现低阻态(low-resistance state, LRS),而当细丝断开时器件处于高阻态(high-resistance state,HRS)。由于其阻变特性与氧离子迁移密切相关,因此也被称为阴离子型阻变器件[26]。图2[33]为OxRRAM器件阻变过程示意。

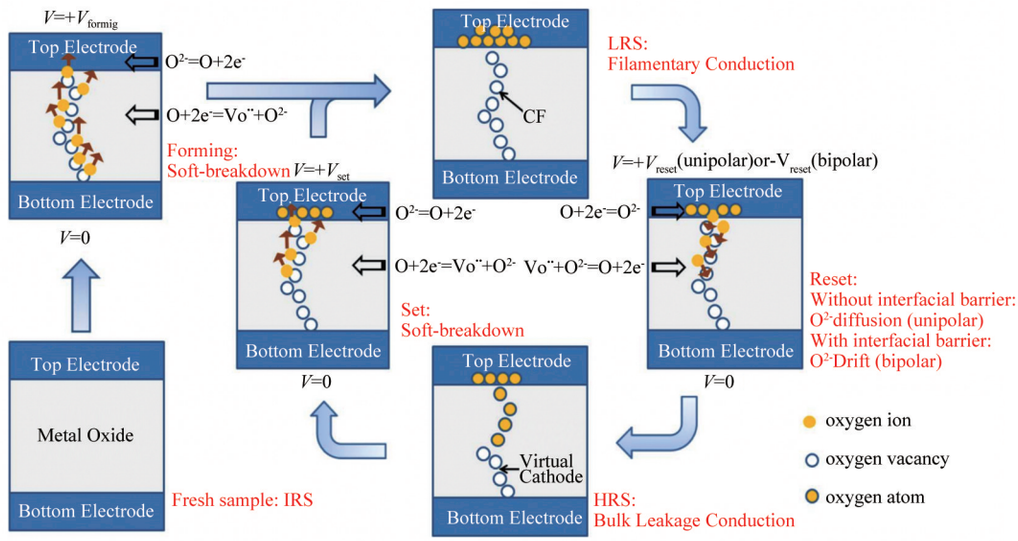


图2　OxRRAM器件阻变过程示意

为了在神经网络中作为记忆权重(weight)的突触,OxRRAM器件需要具有在多个阻态间切换的能力,即实现多比特单元(multi-bit cell, MLC)。在较早期的研究中,韩国浦项科技大学Hyunsang Hwang团队[34]基于Ta/N-TaO_x/Pt体系的器件,通过在TaO_x中掺杂N元素消除多余的导电路径,将导电细丝限制在局部区域,改善了器件的多比特特性。大多数的O_xRRAM在设置(set)时会出现电导态的突变,这对各种深度学习任务都有不利影响。清华大学吴华强团队[35]通过在介质层内插入热增强层来增加阻变时的温度,使得阻变层产生更多条弱导电细丝,而不是一条强导电细丝。这种方法有效改善了器件的模拟特性以及多级数据保持能力。目前有多种不同材料体系的OxRRAM有了相关的展示,包括TaO_x/HfO_x[36]、WO_x[37]和TiO_2[38]等,不同材料体系的器件在保持性(retention)、可靠性(reliability)、耐擦写性(endurance)等方面各有优劣。密歇根大学Wei团队[39]研究了含有6种过渡金属(Zr、Hf、Nb、Ta、Mo、W)构成的高熵氧化物(high-entropy oxides, HEO)作为阻变层,利用高熵材料的"鸡尾酒"效应,有效结合不同材料体系的O_xRRAM技术的优点,改善了忆阻器的整体特性。

1.1.2　金属通道型阻变器件

金属通道型阻变器件,又称导电桥型随机存储器(conductive bridging random access memory, CBRAM),或电化学型存储器(electrochemical metallization memory, ECM)。其结构通常包含活泼电极,阻变介质层和惰性电极3个部分,其中的活泼电极通常为Ag或Cu,阻变介质层可以是固态电解质,也可以是氧化物材料[40]。金属通道型阻变器件的阻变机理如下:在正压激励下,活泼电极发生电化学反应,产生金属阳离子。这些阳离子在电场作用下漂移通过阻变介质层,在惰性电极附近还原为金属原子并逐渐堆积,直至形成连接两端电极的金

属桥,器件被设置到低阻态。而在相反电压激励下,金属桥发生电化学溶解,器件重置(reset)为高阻态。由于金属通道型阻变存储器是由金属阳离子构成导电细丝,因此也被称为阳离子型阻变器件[26]。其*I*–*V*特性曲线与微观机理如图3[41]所示。

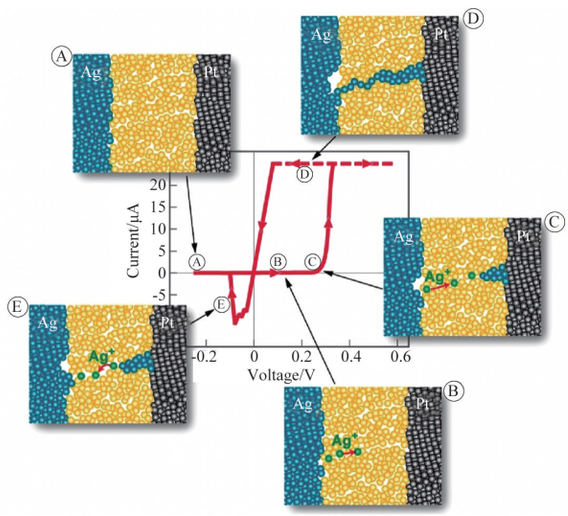
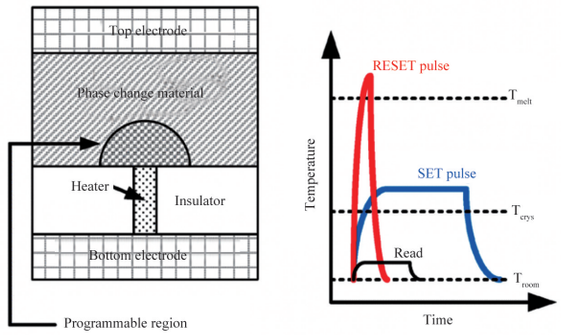


图3 CBRAM器件的阻变*I*–*V*特性曲线与微观机理示意

由于CBRAM具有大开关比和高驱动电流等特性,适合作为选通管(selector)与忆阻器串联组成1S–1R(1–Selector–1–RRAM)结构,以有效缓解交叉阵列中的潜行路径电流问题[42]。华中科技大学缪向水团队[43]设计了一种基于CuS/GeSe的选通管,这种选通管具有超高开关比(1.25×10^9)、高驱动电流(600 μA)以及超低关断电流(~100 fA)等特性。与O_xRRAM相同,CBRAM也可以用于非易失性存储应用,但由于CBRAM中的金属阳离子具有较强的扩散能力,使得CBRAM器件的保持特性与耐擦写特性提升困难。为了限制金属离子的扩散,研究人员研究了各种双层材料。比利时微电子研究中心Belmonte等[44]对比了Cu/Al_2O_3、Cu/La_2O_3、Cu/Ta/GeSe 3种不同的双层/多层堆叠结构并对比分析了不同体系器件的开关比,保持特性和耐擦写特性的差异区别。当用于神经网络应用时,CBRAM需要良好的线性特性。台湾交通大学Tseng团队[45]基于Te/MgO/HfO_x/TiN CBRAM器件研究了HfO_2先沉积后退火的改进方法。退火后在HfO_x薄膜中产生了更多的氧空位,这种过量的氧空位调节了金属碲细丝的形状,通过最窄部分的连接与断开实现电阻的连续切换,从而改善了线性度。

1.2 相变存储器

相变存储器(PCM)是相对成熟的非易失性存储技术之一,其工作机理主要依靠如$Ge_2Sb_2Te_5$(GST)等相变材料[46]。这类材料在晶态和非晶态之间的电阻有很大差异,并且可以通过施加特定电压脉冲产生的焦耳热实现二者之间的切换(图4[47])。当对PCM器件施加较短脉宽的大电压脉冲时,相变材料迅速升温,并在脉冲结束后又快速降温,这使得相变材料经历了淬火过程而转变为非晶态,器件因此展现出高阻态。而当对PCM器件使用较长的小电压脉冲时,相变材料保持在结晶温度以上并缓慢冷却,这使得相变材料会逐渐转变为晶态,器件因此展现出低阻态[13]。



(a) 器件横截面示意

(b) 通过施加不同强度和不同脉宽的脉冲改变温度实现编程和读操作

图4 相变存储器结构及工作原理

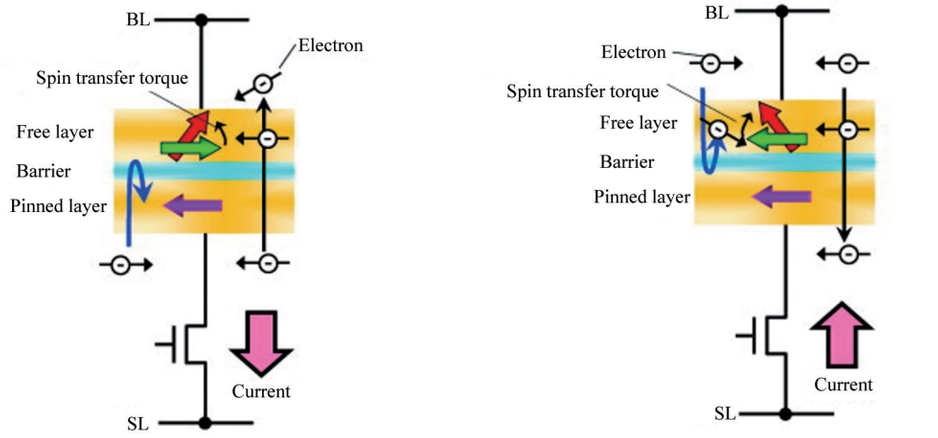
尽管目前PCM研究已经较为成熟,但其目前仍然面临许多挑战。首先,PCM器件存在高编程电流的问题,尤其是reset操作,单个器件操作电流甚至会高于100 μA[48],这会导致较大的编程功耗,严重阻碍其大规模集成。针对这个问题,斯坦福大学Wong团队[49]在相变材料和底部电极中间插入了热阻挡层,将产生的热量限制在PCM器件内,从而降低了驱动电流。意大利米兰理工大学Lacaita

等[50]研究发现,通过减小加热材料与相变材料接触面积,可以显著减小编程电流。其次,相变材料的结晶相通常是稳定的,但非晶相往往是亚稳态[51],这会导致PCM器件电阻随时间变化而稳定增加,这种现象也被称为电阻漂移。深圳大学丁科元等[52]设计了用由交替沉积的限制材料和相变层组成的多层异质结构代替了相变层,降低了成分变化和相分离的可能性,从而有效降低了电阻漂移的影响。另外,在芯片大规模集成时,器件的耐擦写特性同样非常重要[51],由于不同相之间密度不同,在反复擦写后可能会在相变材料与电极的边界处形成空隙,导致PCM卡在高电阻状态。韩国延世大学的Ko团队[53]通过优化沉积条件,抑制了空隙的产生,大大提高了耐擦写特性。

1.3　磁随机存储器

磁随机存储器(MRAM)同样是一种比较成熟的非易失性存储器。磁隧穿结(magnetic tunnel junction,MTJ)是MRAM中的关键元件。MTJ由2层铁磁材料构成,并且这2层中间有一个薄(1~2 nm)的绝缘层用于电子隧穿。这2个铁磁层中,一个是参考磁层,具有固定的磁化方向,另一个是自由层,可以在2种方向之前切换。自旋转移扭矩MRAM(spin-transfer torque MRAM,STT-MRAM)是目前的最成功的一类MRAM[54],其基本结构是1T-1MTJ,即由一个晶体管和一个MTJ组成。STT-MRAM主要利用自旋转移力矩效应改变自由层的磁化方向,进而实现器件状态的改变(图5[55])。当电子从参考层流向自由层时,电子自旋会在参考层中发生极化,随后自旋角动量转移到自由层,使得自由层与参考层磁化方向相同,此时MTJ处于低阻状态,也称为"0"状态。相反则自旋方向与自由层相反的电子会被固定层反射回自由层,从而使得自由层与参考层磁化方向相反,此时MTJ处于高阻状态,也称为"1"状态[56]。

相比于其他忆阻器,STT-MRAM的独特优势在于低操作电压,使用CMOS逻辑电路中的电压就足以进行写入操作,因此可以与处理器逻辑电路做在同一个芯片里[54]。另外,STT-MRAM的耐擦写次数非常高,适合于实现对读/写循环有大量需求的在线训练神经网络。但STT-MRAM的缺点在于其开关比较低,典型值仅为2左右,这导致其数据存储的可靠性降低,通常只能存储二进制数据,不利于进行模拟计算。北京航空航天大学赵巍胜团队[57]将MTJ和肖特基二极管集成制造了整流隧道磁阻器件(rectified tunnel magnetoresistance, R-TMR),通过调整交流电(AC)和直流电(DC)的比



(a) 从自旋方向相反向自旋方向相同切换　(b) 从自旋方向相同向自旋方向相反切换

图5　磁随机存储器状态转换

例,实现了高开关比(>100)。此外,最近几年MRAM研究领域出现了一些新技术,包括自旋轨道矩磁随机存储器(spin orbit torque MRAM, SOT-MRAM)和电压调控磁各向异性磁随机存储器(voltage-controlled magnetic anisotropy MRAM, VC-MA-MRAM)等。SOT-MRAM使用基于三端MTJ的概念来隔离读取和写入路径,提高了器件的耐擦写特性与读取稳定性[58]。VCMA-MRAM利用VC-MA效应[59],能进一步降低写入能量,减小MTJ的面积[60]。

1.4 铁电存储器

铁电存储器(ferroelectric memory)主要依靠外加电场改变铁电材料的极化状态,从而引起器件电阻值的变化[61]。铁电材料状态的读取可以通过3种方法来实现(图6[62]),从而产生了3种不同的研究方向。第一种方法是在铁电材料两侧施加一个大电场,根据铁电材料是否极化会有大或者小的电流流过铁电材料。由于这种基于电容的读取方法是一种破坏性操作,因此需要读取后进行回写。这种方案用于铁电随机存储器(ferroelectric random-access memory,FeRAM)。第二种方法是将铁电材料集成到场效应晶体管的栅极中,不同极化态决定了不同的阈值电压,基于此的器件被称为铁电场效应晶体管(ferroelectric field-effect transistor,FeFET)。第三种方法是在2个电极中间夹着一个非常薄(通常只有几纳米)的铁电材料,通过该器件的隧穿电流取决于铁电极化方向,这种器件被称为铁电隧道结(ferroelectric tunnel junction,FTJ)。由于FTJ的读出电流很小,有利于允许大规模并行操作[63],因此适合作为类脑计算系统的突触结构。

1921年,Valasek在罗谢尔盐(Rochelle salt)中首次观察到铁电特性[64],此后人们逐渐发现了各种铁电材料。2011年,研究人员偶然在氧化铪(HfO_2)中发现了铁电性[65],这一发现点燃了人们对铁电存储器的研究热情,因为氧化铪在前端工艺和后端工艺中都是一种成熟且完全兼容CMOS的材料。近年来,许多研究尝试使用FTJ构建交叉阵列,但FTJ较低的开关比,限制了其模拟计算的应用。为解决该问题,东芝Fujii等[66]和德累斯顿工业大学Max等[67]分别通过将铁电层与非常薄的隧道层结合成双层堆叠结构,通过改变势垒高度和隧穿长度产生不同的隧穿电阻,同时实现了大开关比,长保持时间以及低运行电压。除了开关比,FTJ的导通电流过低致外围电路无法读取也是一个巨大的挑战,20 nm技术节点的FTJ单器件Ion的典型值小于1 fA,即使一列有1024个器件总电流也小于1 pA[68],而理想状况下至少要达到1 μA[69]。美国佐治亚理工学院Yu团队[70]从DRAM的结构获得启发,设计了一种圆柱结构的FTJ,增加了器件的有效面积,提升了单个器件的电流,仿真证明了一列100个器件即可使得电流高于1μA。

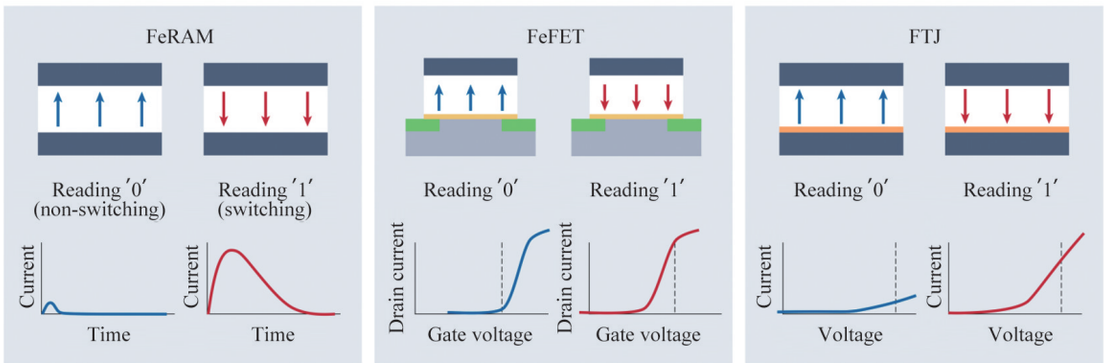


图6 3种读取铁电材料状态的方式(图中虚线表示读取电压)

1.5　忆阻器器件研究进展

表1总结了国内外一些研究团队的器件指标,以及IEEE国际设备与系统路线图(IRDS)[71]中提出的指标要求。尽管许多器件在部分指标上已经达到了要求,但截至目前,尚没有一种能够满足所有指标要求的器件。研制综合特性优异的忆阻器器件,仍是未来攻克的重点。

表1　国内外忆阻器器件研究进展

器件种类	研究地区	材料体系	研究进度	模拟电导态数	开关比	开关速度/ns	操作电压(set/rst)/V	保持特性	耐擦写特性
目标	IRDS[71]	—	—	>16	$>10^6$	<10	<1/1	>10 a(@85℃)	$>10^9$
金属氧化物阻变存储器	中国[35,72]	$TiN/TaO_x/HfO_2/TiN$	阵列	128	10	<50	1.6/1.5	>10 min(@125℃)	$>10^7$
	中国[73]	TaO_x	阵列	—	4.6/752	<5	0.8/1.2	>200s	—
	美国[74]	$TiN/Ti/HfO_2/TiN$	阵列	5	20	50	1.3/2.5	>10 a(基于动态地址重映射)	—
	美国[75]	$Ta/HfO_2/Pt$	阵列	32	20	100	0.65/1.1	$>10^6$ s(@25℃)	—
	美国[76]	$Ta/TaO_x/Ir$(or Pd)	阵列	—	10	—	—	>100 a(@85℃)	$>10^4$
	中国台湾[77]	$Ti/HfO_2/TiN$	阵列	—	—	100	1.5/1.5	>100 a(@85℃)	$>10^6$
导电桥型随机存储器	中国[78]	$Ag/ZrO_2/WS_2/Pt$	器件	—	>3000	10	0.2/0.1	$>4\times10^4$ s	$>10^9$
	中国[79-80]	Ag/GeSe/TiN	器件	16	>10	$<4\times10^4$	0.6/1	3×10^3 s	—
	中国[81]	$Cu/Ta_2O_5/TaO_x/W$	器件	—	10^4	<15	3/2	$>5\times10^4$ s	$>10^6$
	美国[82-83]	Cu/MoS_2 double layer/Au	阵列	4	5	—	0.1~0.2	1.8×10^4 s	>20
	韩国[84]	$Cu/HfO_2/Ta/Cu_2S/W$	器件	8	>20	—	0.5/0.5	—	—
相变存储器	中国[85]	$Pt/Ge_2Sb_2Te_5/Pt$	器件	4	>140	10~40	<5	$>10^4$ s	>300
	中国[86]	Al/TiN/GeS/W	阵列	—	$>10^6$	10~100	3.2/3.2	—	10^8
	韩国[53]	$C_{15}(Ge_{21}Sb_{36}Te_{43})$	器件	—	>200	<700	<5	—	$>10^8$
	法国[87]	$Ge_2Sb_2Te_5$	阵列	—	>25	30/200	1	—	$>10^7$
磁随机存储器	中国[88]	SOT-MRAM	器件	—	<3	0.3	—	—	—
	中国[89]	$Pt/Co/AlO_x$	器件	—	—	<5	10/30	—	—
	韩国[90-91]	MgO/CoFeB	阵列	2	<3	<50	<1.5	>10 a@85℃	$>10^{10}$
	中国台湾[92]	—	阵列	—	—	2.75	0.85	—	—
铁电存储器	中国[93]	$Pt/BaTiO_3/Nb:SrTiO_3$	器件	—	6×10^6	—	1/1	$>5\times10^4$ s	2×10^4
	中国[94]	$Hf_{0.5}Zr_{0.5}O_2$	阵列	16	5	500	4/4	—	$>10^6$
	韩国[95]	$HfZrO_x$ FeFET	器件	5	>32	10^5	3.7/3.2	—	—
	美国[96]	$Hf_{0.5}Zr_{0.5}O_2$ FeFET	器件	200	>256	50	5/5	—	$>10^7$
	法国[97]	$Co/BiFeO_3/Ca_{0.96}Ce_{0.04}MnO_3$	阵列	1000	—	100	1.5/1.4	>10 a	4×10^6

2　应用演示研究进展

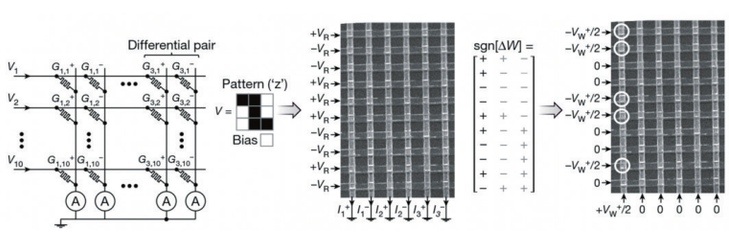
近年来,得益于忆阻器器件各种特性的进步,忆阻器的硬件集成与应用演示也有了巨大突破。

2.1　神经网络

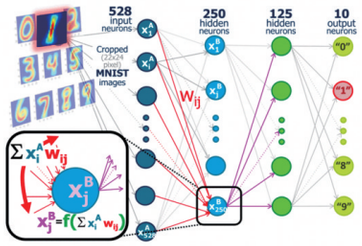
受生物大脑神经网络启发的神经网络算法在图像处理和语音识别等任务中表现出与人类近似甚至超过人类的强大性能,引发了人们极大的研究热情[3]。由于神经网络中存在大量的向量矩阵乘法运算,因此非常适合使用忆阻器阵列实现加速。目前,基于忆阻器存算一体架构的神经网络已被证明在速度和能效方面有望较基于冯·诺依曼架构的芯片高出 1000倍以上[16],对于需要高能效的边缘计算应用等非常有吸引力。因此,近年来研究人员针对神经网络算法的忆阻器阵列实现展开了探索。

前馈神经网络(feedforward neural network,FNN)是应用最广泛的神经网络算法之一[98]。其由2层或多层神经元排列组成,每个神经元仅接收上

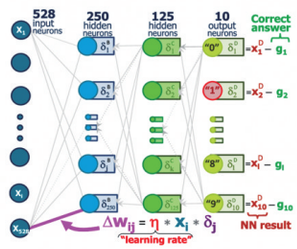
一层的输入并把结果输出到下一层,各层之间不存在反馈。其中,感知机(perceptron)是最简单的一类前馈神经网络,单层感知机(single-layer perceptron,SLP)只有输入层和输出层,多层感知机(multiple-layer perceptron,MLP)在输入层和输出层之间插入了隐藏层。2015年美国加州大学圣巴巴拉分校Strukov团队[21]率先报道了基于12×12规模的RRAM阵列构成的单层感知机网络,实现3个字母"Z""V"和"N"的分类任务(图7(a)[21]),这是存算一体架构向着神经网络应用迈出的重要一步。2015年IBM的Burr等[99]在一个具有164885个突触单元的大规模PCM阵列上搭建了用于手写数字识别的3层感知机网络,并且实现了网络的反向传播训练(图7(b)~(c)[99])。卷积神经网络(convolutional neural network, CNN)是另一种典型的前馈神经网络,卷积核及池化概念的引入,使其较感知机网络能够精简训练参数、降低过拟合、实现更高维度的信息提取。清华大学吴华强团队[36]实现了一个全硬件的5层卷积神经网络(图7(d)[36]),并提出了混合训练的方式以补偿器件存在的部分非理想特性,实现推理准确率接近软件计算的准确率。另外,该团队通过并行卷积技术和相同核的复制,大大提高了并行度,在能效比和性能密度方面相比于Tesla V100图形处理器(GPU)提高了约2个数量级。



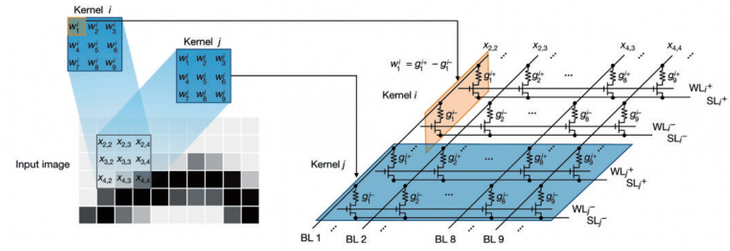
(a) 使用忆阻器构建的单层感知机的电路结构、工作模式和权重调整方法[21]



(b) 多层感知机的前项推理结构示意[99]



(c) 多层感知机的反向传播结构示意[99]

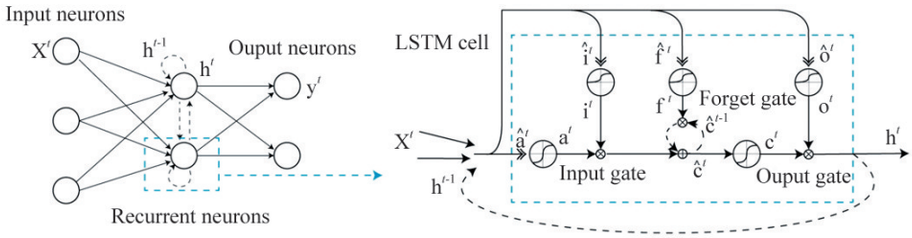


(d) 使用忆阻器阵列实现卷积运算[36]

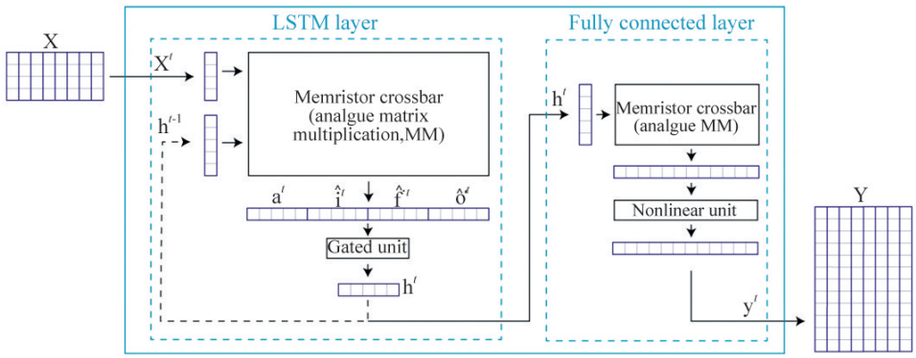
图7 忆阻器阵列在前馈神经网络方向的验证工作

循环神经网络(recurrent neural networks, RNN)是另一类常见的神经网络算法,通常用于处理序列信号。与前馈神经网络最大的不同在于,其层间存在反馈信号的连接与信息传递(图8(a)[100])。传统的RNN算法在处理长序列信号时存在长期依赖性问题,为了解决这个问题,研究人员提出了长短期记忆网络(long short-term memory,LSTM),通过引入3个门和单元状态实现有选择性的记忆和遗忘(图8(b)[100])。美国马萨诸塞大学的Yang团队构建了包含忆阻器LSTM层和忆阻器全连接层的多层网络,并在回归和分类实验中成功演示了LSTM的推理与原位训练,相比全数字系统显著降低了延迟与功耗(图8(c)[100])。储备池计算(reservoir computing,RC)[101]是另一种受生物大脑启发的循环神经网络。储备池计算由输入层、储备池和输出层3部分组成,其中储备池是随机生成的稀疏连接的递归神经元网络,模拟了生物大脑神经元之间的复杂稀疏连接。由于仅需要训练输出层的权重

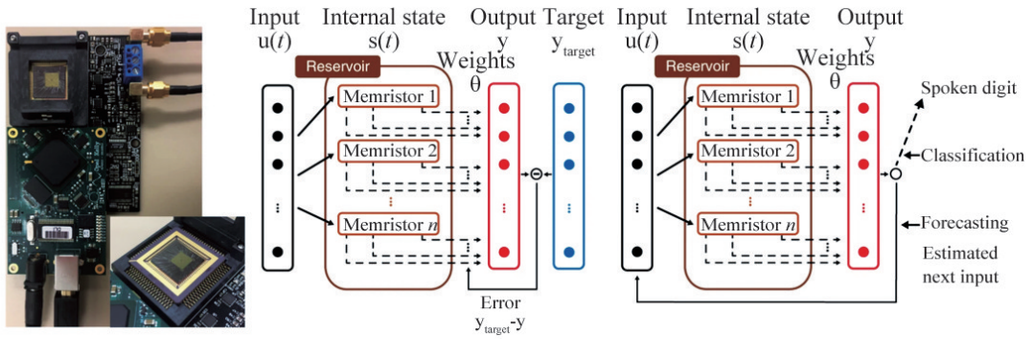
参数,因此具有低训练成本和低硬件开销等特点。美国密歇根大学的Lu团队基于32×32的RRAM阵列构建了储备池计算网络(图8(d)[102]),实现了语音数字识别和混沌序列预测,相比数字系统有效降低了功耗,同时也说明了忆阻器器件之间的差异让权重分布更加随机,使得对于输出的响应更加多样化,从而有利于提高储备池计算系统性能。



(a) 多层RNN网络示意[100]　(b) LSTM单元结构[100]



(c) 使用忆阻器阵列实现LSTM的数据流示意[100]



(d) 基于忆阻器的储备池计算系统训练和测试阶段示意[102]

图8　忆阻器阵列在循环神经网络方向的验证工作

此外,一些其他神经网络算法同样在基于忆阻器的存算一体架构上得到广泛的研究与演示:生成对抗网络(generative adversarial networks,GAN)[103]是一种近年来研究非常火热的网络,该网络由两部分组成,包括生成器(generator)和判别器(discriminator)。生成器负责生成逼真样本,判别器则负责

判别该样本的真假,两者交替训练,最终使生成器可以输出非常逼真的图像或数据。清华大学吴华强团队[104]首次基于RRAM阵列实验演示了使用生成对抗网络的手写数字图案生成,并证明了RRAM的固有随机噪声可用于生成对抗网络的输入,具备生成数字多样性高的优势。

有别于上述人工神经网络,脉冲神经网络(SNN)[2]是一种更贴合生物神经元工作方式的网络,其内部神经元传递的是时间间隔不一的脉冲信号,而非前馈神经网络和循环神经网络中采用的连续值。由于算法中增加了时间这个维度,脉冲神经网络非常适合处理基于时空事件的信息,但目前主要受限于没有有效的训练算法[42,105],无法搭建更深更复杂的网络,因此仍需算法层面的突破。近年来,许多研究已经基于包括RRAM、PCM、MRAM和FeFET等在内的各种器件实现了脉冲神经网络算法的仿真或硬件演示[106-109],验证了基于忆阻器阵列实现SNN算法的高能效和高扩展性等特点。

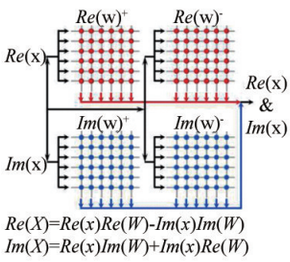
2.2 信号处理与机器学习

随着物联网(Internet of Things,IoT)的发展,为了减小通信带宽和延迟,许多应用要求网络边缘端能够快速且低功耗地预处理传感器的原始数据,并仅向云端传输重要的分析结果[110]。基于忆阻器阵列的存算一体架构有望实现较传统CMOS系统更高效的信号处理,使其成为了一种有吸引力的解决方案。与神经网络类似,一些信号处理与传统机器学习的算法核心也是向量矩阵乘法,因此同样可以使用忆阻器阵列来运行,获取更高的速度和能效。更重要的是,忆阻器阵列能够直接接收并处理从传感器中获取的模拟信号,从而大幅降低输入电路的复杂度。在近些年的研究中,研究人员对信号变换、信号编码和传统机器学习等算法的基于忆阻器的实现进行了前沿探索[111]。

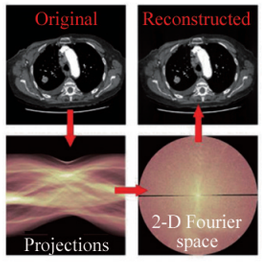
傅里叶变换(Fourier transformation)是用途最广泛、最重要的信号处理算法之一,可以将信号从时域转换为频域。由于计算机是离散系统,实际工程中常用离散傅里叶变换(discrete Fourier transform,DFT)来近似傅里叶变换。清华大学吴华强团队[112]基于RRAM阵列首次实现了离散傅里叶变换,并演示了高保真度的医学图像重建功能(图9(a)~(c)[112]),能效较CPU提高了128倍。此外,为实现高精度傅里叶变换处理,团队提出了一种准模拟映射(QAM)方案,该方法较传统量化映射(QM)方案提升了映射精度,并具有更高的器件读噪声鲁棒性。此外,离散余弦变换(discrete cosine transform, DCT)是较离散傅里叶变换具有更高压缩率的算法,广泛用于数字信号处理和图像压缩。美国马萨诸塞大学的Yang团队[110]在128×64的RRAM阵列上实现了基于二维离散余弦变换(2D-DCT)的图像压缩(图9(d)~(f)[110]),在保留了包含频谱幅度前15%的频率(即压缩比为20∶3)后使用软件进行还原仍可以得到相似的重构图像,但还原质量不佳。美国中佛罗里达大学的Zhang等[113]在此基础上通过将2D-DCT重构为线性变化并进行了频谱优化,在提高图像生成质量的同时,降低了延迟、功耗和芯片面积(图9(g)~(h)[113])。

上述的频域变换压缩信号的思路是先对信号进行采样,再通过频域变化挖掘稀疏性,最后通过压缩算法实现压缩。而在一些信号处理的任务中,数据采集与压缩是顺序进行的,如果采用上述的压缩方法会导致采样采集的大部分数据都是无效冗余的数据,造成资源浪费。为了解决这个问题,人们提出了压缩感知(compressed sensing,CS),压缩感知的基本思想是同时完成采样与压缩,从高维信号中获取少量采样测量值,最后准确地恢复该信号。IBM的Le Gallo等[114]基于256 Kb的PCM阵列实现了压缩感知的编码和解码的实验演示,功耗仅为现场可编程门阵列(FPGA)系统功耗的2%。另外该团队还指出由于压缩感知解码的过程仅仅执行读操作,因此观测矩阵仅需编程一次,能够有效避免反复编程带来的运行速度下降、功耗增加和器件可靠性降低等问题。

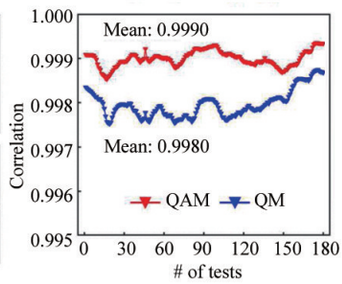
另外,基于忆阻器阵列的传统机器学习算法目前也已经得到了深入研究。主成分分析(principal component analysis,PCA)是目前使用最广泛的数据降维算法之一,也是许多机器学习算法的重要预处理步骤。美国密歇根大学的Lu团队[37,115-116]基于9×2的忆阻器阵列使用无监督的在线学习(图10



（a）基于RRAM阵列的一维DFT实现示意[112]

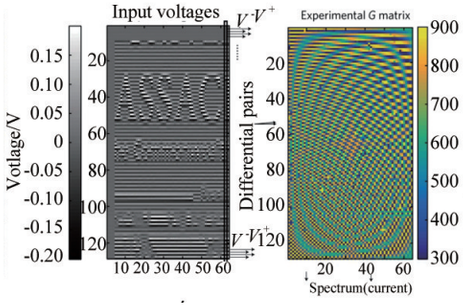


（b）CT重建任务的原始图像、中间结果和重建图像[112]

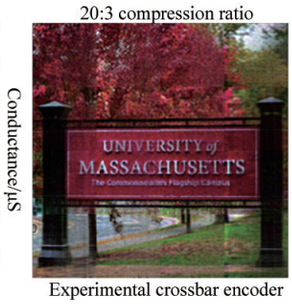


（c）180种不同测试输入信号下基于QAM、基于QM和基于软件的DFT结果之间的相关性[112]

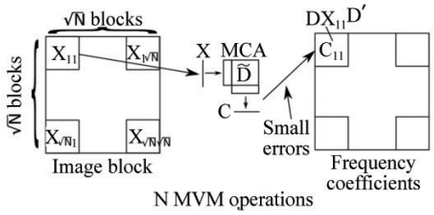
（d）用于压缩的原始图像按照白色箭头顺序逐块输入到2D-DCT[110]



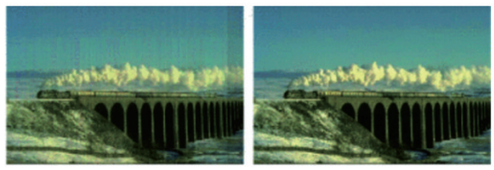
（e）输入信号与忆阻器阵列上的权重的像素图[110]



（f）从2D-DCT解码的图像[110]



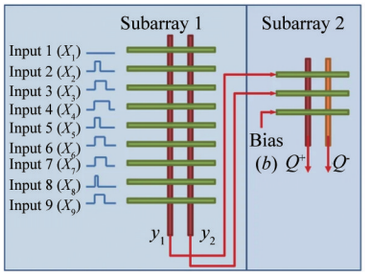
（g）使用重构的DCT矩阵实现2D-DCT[113]



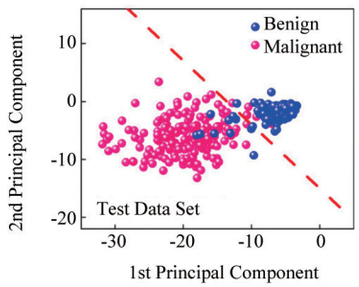
（h）优化前后的重建图像对比[113]

图9　忆阻器阵列在信号变换方向的验证工作

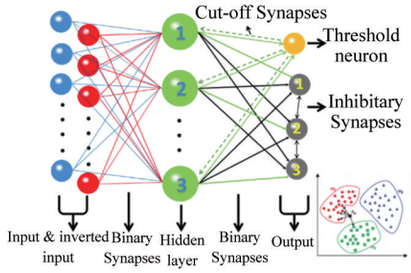
(a)[37]),以乳腺癌测量数据库中部分数据作为训练集学习主成分,将9维的原始数据降维到2维的主成分数据,有效地将测试集的数据分成簇(图10(b)[116]),最后使用监督学习将肿瘤分类,实现了97.1%的准确率,十分接近于软件准确率(97.6%)。k近邻算法(k-nearest neighbors,kNN)是另一种在模式识别方面取得巨大成功的机器学习算法。算法核心是计算两点在欧几里得空间中的距离,可以使用忆阻器阵列实现并行计算加速。北京大学康晋锋团队[117-118]设计了kNN算法的有监督训练方案(图10(c)[118]),并通过仿真验证了基于kNN的MNIST手写数字识别,实现了90%以上的准确率。该团队还通过实验证明即使在器件电导变化高达60%或输入噪声高达40%的情况下,分类精度也不会出现明显下降,证明该算法非常适合RRAM器件(图10(d)[118])。



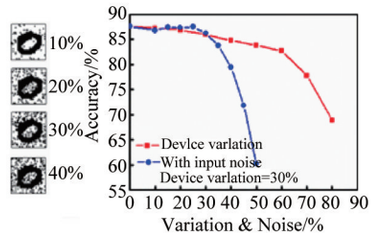
（a）基于忆阻器阵列实现PCA算法结构示意图[37]



（b）使用监督学习过程得到的决策边界(红色虚线)可以有效将来自忆阻器PCA分析后的测试数据分开[116]



（c）kNN算法的网络结构示意[118]



（d）不同器件变化和输入噪声对准确率的影响[118]

图10 忆阻器阵列在机器学习领域的验证工作

2.3 应用领域研究进展总结

国内外忆阻器应用领域的研究进展如表2所示。目前,国内外的研究团队已经完成了许多神经网络、信号处理和机器学习等算法的硬件阵列或仿

表2 国内外忆阻器应用领域研究进展

	分类	算法	研究地区	研究进度	阵列规模	展示应用
神经网络	前馈神经网络	感知机	美国[21]	硬件阵列	12×12	字母三分类
		感知机	中国[105]	硬件阵列	128×8	人脸三分类
		多层感知机	美国[99]	硬件阵列	500×661	手写数字识别
		卷积神经网络	中国[36]	硬件阵列	16 Kb	手写数字识别
	循环神经网络	长短期记忆网络	美国[100]	硬件阵列	128×64	回归预测、步态分类
		储备池计算	美国[102]	硬件阵列	32×32	语音数字识别、混沌序列预测
		储备池计算	中国[119]	硬件阵列	64×5	混沌序列预测、手写字母识别
	其他神经网络算法	生成对抗网络	中国[104]	硬件阵列	128×8	手写数字生成
		脉冲神经网络	美国[107]	硬件阵列	1.4 Mb	手写数字识别
		脉冲神经网络	日本[106]	硬件阵列	1.48 Mb	手写数字识别
		脉冲神经网络	美国[108]	软件仿真	—	—
		脉冲神经网络	中国[120]	软件仿真	—	手写数字识别
信号处理与机器学习	信号变换	离散傅里叶变换	中国[112]	硬件阵列	16 Kb	CT医学影像重建
		离散余弦变换	美国[110]	硬件阵列	128×64	图像压缩
		离散余弦变换	美国[110]	软件仿真	128×64	图像压缩
	信号编码	压缩感知	瑞士[114]	硬件阵列	>1 Mb	图像压缩与恢复
	机器学习	主成分分析	美国[37,115-116]	硬件阵列	9×2	肿瘤数据降维
		k近邻算法	中国[117-118]	软件仿真	—	手写数字识别

真验证演示,到目前为止,使用硬件阵列演示的只是一些较为简单的算法,对于更多真正需要高算力的复杂网络算法仍处于仿真验证阶段。另外,现有的报道通常仅展示相比CMOS系统在能耗比上的优势,尚未有能与目前顶尖GPU或FPGA相抗衡的峰值算力。

3　结论

国际上对基于忆阻器的存算一体架构的研究仍处在起步状态。在器件优化和材料工程方面,不论是哪种工作机制的器件,都仍存在一些非理想效应,尚不能完全满足应用的需求[71],离真正产业化还有一定距离。在硬件集成和应用演示方面,目前完成的应用演示仅局限在较为简单的算法,对于复杂但实用的深度学习算法(如ResNet50、yolo等)仍处于仿真验证阶段。

就目前在器件优化和应用演示的研究进展来看,中国在基于忆阻器的类脑计算领域在国际上已经占有一席之地,在领域内的进一步发展可以从以下4个角度开展。

1)在器件优化方面,综合特性优异的器件,仍是需要攻克的重点。基于现有的忆阻器器件深入探索工作机理,综合优化器件的线性度、对称性、开关比、保持特性和耐擦写特性等参数。此外,要探索忆阻器的三维集成技术,从而进一步提高集成密度和功能多样性,为未来探索复杂神经网络应用打下基础。

2)在电路设计方面,交叉阵列结构和外围电路等还有很大优化空间。首先,串联线阻分压问题严重限制了单个阵列的大小,尤其是在先进工艺节点下。如2T2R结构[72]等全新的交叉阵列设计有助于缓解IR drop问题构建更大的交叉阵列。其次,目前外围电路(ADC和DAC等电路)的面积和能耗远超过忆阻器阵列,严重降低了存算一体架构本身的速度和能耗优势。因此需要进行器件、电路和架构的协同设计,包括高精度低功耗的ADC设计和高吞吐量的数据流方案[121]。另外,如何发挥忆阻器本身模拟计算的优势,实现感存算一体也是一种非常有吸引力的发展方向。

3)在芯片集成方面,为了进一步降低功耗、面积和延迟等,需要将包括数字和模拟转换电路、缓冲存储器、数字处理器等在内的所有功能模块与忆阻器阵列集成到一块芯片。除此之外,还需要设计灵活的架构与调度方案,在保证高能耗的前提下实现高算力与高灵活性,使芯片能够支持各种不同深度学习和复杂信号处理等任务。

4)在产业发展方面,参考GPU和FPGA等已经成熟技术的发展道路,基于忆阻器的类脑计算作为一种仍处于起步阶段的新技术,要实现产业化落地,需要芯片制造厂、芯片设计厂和高校等加强合作,共同研发、制定目标、应用和推广。另外,要避免与成熟产品直接竞争,利用存算一体技术自身的优势,开拓新应用、打造新场景、吸引新用户,创造新市场。

可以预见,通过各个领域的密切合作,基于忆阻器的类脑计算芯片领域将会实现连续不断的突破,而中国的学者和机构也将在其中扮演越来越重要的角色。

参考文献(References)

[1] Abiodun O I, Jantan A, Omolara A E, et al. State-of-the-art in artificial neural network applications: A survey[J]. Heliyon, 2018, 4(11): e00938.

[2] Tavanaei A, Ghodrati M, Kheradpisheh S R, et al. Deep learning in spiking neural networks[J]. Neural Networks, 2019, 111: 47-63.

[3] Le Cun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[4] Dario A, Danny H. AI and Compute[EB/OL]. (2018-05-16) [2022-04-28]. https://openai. com/blog/ai-and-compute/.

[5] Moore G E. Cramming more components onto integrated circuits[J]. Proceedings of the IEEE, 1998, 86(1): 82-85.

[6] Patterson D. 50 Years of computer architecture: From the mainframe CPU to the domain-specific tpu and the open RISC-V instruction set[C]//2018 IEEE International Solid-State Circuits Conference(ISSCC). Piscataway, NJ:

IEEE, 2018: 27-31.
[7] Lundstrom M. Moore's Law forever[J]. Science, 2003, 299 (5604): 210-211.
[8] Chau R, Doyle B, Datta S, et al. Integrated nanoelectronics for the future[J]. Nature Materials, 2007, 6(11): 810-812.
[9] Leiserson C E, Thompson N C, Emer J S, et al. There's plenty of room at the Top: What will drive computer performance after Moore's law[J]. Science, 2020, 368(6495): eaam9744.
[10] Salahuddin S, Ni K, Datta S. The era of hyper-scaling in electronics[J]. Nature Electronics, 2018, 1(8): 442-450.
[11] McKee S A. Reflections on the memory wall[C]//Proceedings of the 1st conference on Computing frontiers. New York: Association for Computing Machinery, 2004: 162.
[12] Wong H-S P, Salahuddin S. Memory leads the way to better computing[J]. Nature Nanotechnology, 2015, 10 (3): 191-194.
[13] Qureshi M K, Gurumurthi S, Rajendran B. Phase change memory: From devices to systems[J]. Synthesis Lectures on Computer Architecture, 2011, 6(4): 1-134.
[14] Furber S. Large-scale neuromorphic computing systems [J]. Journal of Neural Engineering, 2016, 13(5): 051001.
[15] Indiveri G, Liu S C. Memory and information processing in neuromorphic systems[J]. Proceedings of the IEEE, 2015, 103(8): 1379-1397.
[16] Hu M, Strachan J P, Li Z, et al. Dot-product engine as computing memory to accelerate machine learning algorithms[C]//2016 17th International Symposium on Quality Electronic Design (ISQED). Piscataway, NJ: IEEE, 2016: 374-379.
[17] Chua L. Memristor-The missing circuit element[J]. IEEE Transactions on Circuit Theory, 1971, 18(5): 507-519.
[18] Strukov D B, Snider G S, Stewart D R, et al. The missing memristor found[J]. Nature, 2008, 453(7191): 80-83.
[19] Zhao M R, Wu H Q, Gao B, et al. Investigation of statistical retention of filamentary analog RRAM for neuromophic computing[C]//2017 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2017: 39.4.1-39.4.4.
[20] Wu Y, Yu S M, Guan X M, et al. Recent progress of resistive switching random access memory (RRAM) [C]// 2012 IEEE Silicon Nanoelectronics Workshop (SNW). Piscataway, NJ: IEEE, 2012: 1-4.
[21] Prezioso M, Merrikh-Bayat F, Hoskins B D, et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors[J]. Nature, 2015, 521 (7550): 61-64.
[22] Ielmini D, Wang Z, Liu Y. Brain-inspired computing via memory device physics[J]. APL Materials, 2021, 9(5): 050702.
[23] Kumar S, Wang X, Strachan J P, et al. Dynamical memristors for higher-complexity neuromorphic computing [J]. Nature Reviews Materials, 2022, 7(7): 575-591.
[24] Li X Y, Zhong Y N, Chen H, et al. A memristors-based dendritic neuron for high-efficiency spatial-temporal information processing[J]. Advanced Materials, 2022, doi: 10.1002/adma.202203684.
[25] Li X Y, Tang J S, Zhang Q T, et al. Power-efficient neural network with artificial dendrites[J]. Nature Nanotechnology, 2020, 15(9): 776-782.
[26] Sun W, Gao B, Chi M F, et al. Understanding memristive switching via in situ characterization and device modeling[J]. Nature Communications, 2019, 10(1): 3453.
[27] Xu X X, Lv H B, Liu H T, et al. Superior retention of low-resistance state in conductive bridge random access memory with single filament formation[J]. IEEE Electron Device Letters, 2015, 36(2): 129-131.
[28] Bersuker G, Veksler D, Nminibapiel D M, et al. Toward reliable RRAM performance: Macro-and micro-analysis of operation processes[J]. Journal of Computational Electronics, 2017, 16(4): 1085-1094.
[29] Degraeve R, Fantini A, Gorine G, et al. Quantitative model for post-program instabilities in filamentary RRAM[C]//2016 IEEE International Reliability Physics Symposium (IRPS). Piscataway, NJ: IEEE, 2016: 6C-1-1-6C-1-7.
[30] Gao B, Wu H Q, Wu W, et al. Modeling disorder effect of the oxygen vacancy distribution in filamentary analog RRAM for neuromorphic computing[C]//2017 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2017: 4.4.1-4.4.4.
[31] Zhao Y D, Huang P, Chen Z, et al. Modeling and optimization of bilayered TaOx RRAM based on defect evolution and phase transition effects[J]. IEEE Transactions on Electron Devices, 2016, 63(4): 1524-1532.
[32] Huang P, Liu X Y, Li W H, et al. A physical based analytic model of RRAM operation for circuit simulation[C]// 2012 International Electron Devices Meeting. Piscat-

away, NJ: IEEE, 2012: 26.6.1-26.6.4.

[33] Wong H-S P, Lee H Y, Yu S, et al. Metal-oxide RRAM [J]. Proceedings of the IEEE, 2012, 100(6): 1951-1970.

[34] Misha S H, Tamanna N, Woo J, et al. Effect of nitrogen doping on variability of TaOx -RRAM for low-power 3-bit MLC applications[J]. ECS Solid State Letters, 2015, 4 (3): 25.

[35] Wu W, Wu H Q, Gao B, et al. Improving analog switching in HfO_2-based resistive memory with a thermal enhanced layer[J]. IEEE Electron Device Letters, 2017, 38 (8): 1019-1022.

[36] Yao P, Wu H Q, Gao B, et al. Fully hardware-implemented memristor convolutional neural network[J]. Nature, 2020, 577(7792): 641-646.

[37] Cai F, Correll J M, Lee S H, et al. A fully integrated reprogrammable memristor—CMOS system for efficient multiply-accumulate operations[J]. Nature Electronics, 2019, 2(7): 290-299.

[38] Serb A, Bill J, Khiat A, et al. Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses[J]. Nature Communications, 2016, 7(1): 12611.

[39] Ahn M, Park Y, Lee S H, et al. Memristors based on (Zr, Hf, Nb, Ta, Mo, W) high-entropy oxides[J]. Advanced Electronic Materials, 2021, 7(5): 2001258.

[40] Islam R, Li H, Chen P Y, et al. Device and materials requirements for neuromorphic computing[J]. Journal of Physics D: Applied Physics, 2019, 52(11): 113001.

[41] Valov I, Waser R, Jameson J R, et al. Electrochemical metallization memories—fundamentals, applications, prospects[J]. Nanotechnology, 2011, 22(25): 254003.

[42] Xia Q F, Yang J J. Memristive crossbar arrays for brain-inspired computing[J]. Nature Materials, 2019, 18(4): 309-323.

[43] Lin Q, Li Y, Xu M, et al. Dual-Layer selector with excellent performance for cross-point memory applications [J]. IEEE Electron Device Letters, 2018, 39(4): 496-499.

[44] Belmonte A, Reale G, Fantini A, et al. Effect of the switching layer on CBRAM reliability and benchmarking against O_xRAM devices[J]. Solid-State Electronics, 2021, 184: 108058.

[45] Hsu C-L, Saleem A, Singh A, et al. Enhanced linearity in CBRAM synapse by post oxide deposition annealing for neuromorphic computing applications[J]. IEEE Transactions on Electron Devices, 2021, 68(11): 5578-5584.

[46] Raoux S, Xiong F, Wuttig M, et al. Phase change materials and phase change memory[J]. MRS Bulletin, 2014, 39(8): 703-710.

[47] Wong H-S P, Raoux S, Kim S, et al. Phase change memory[J]. Proceedings of the IEEE, 2010, 98(12): 2201-2227.

[48] Li Y B, Wang Z R, Midya R, et al. Review of memristor devices in neuromorphic computing: Materials sciences and device challenges[J]. Journal of Physics D: Applied Physics, 2018, 51(50): 503002.

[49] Ahn C, Fong S W, Kim Y, et al. Energy-efficient phase-change memory with graphene as a thermal barrier[J]. Nano Letters, 2015, 15(10): 6809-6814.

[50] Lacaita A L, Redaelli A. The race of phase change memories to nanoscale storage and applications[J]. Microelectronic Engineering, 2013, 109: 351-356.

[51] Boniardi M, Ielmini D, Lavizzari S, et al. Statistics of resistance drift due to structural relaxation in phase-change memory arrays[J]. IEEE Transactions on Electron Devices, 2010, 57(10): 2690-2696.

[52] Ding K Y, Wang J J, Zhou Y X, et al. Phase-change heterostructure enables ultralow noise and drift for memory operation[J]. Science, 2019, 366(6462): 210-215.

[53] Park J H, Kim S W, Kim J H, et al. Enhancement of a cyclic endurance of phase change memory by application of a high-density $C1_5(Ge_{21}Sb_{36}Te_{43})$ film[J]. AIP Advances, 2016, 6(2): 025013.

[54] Na T, Kang S H, Jung S O. STT-MRAM sensing: A review[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2021, 68(1): 12-18.

[55] Kawahara T, Ito K, Takemura R, et al. Spin-transfer torque RAM technology: Review and prospect[J]. Microelectronics Reliability, 2012, 52(4): 613-627.

[56] Worledge D C. Spin-Transfer-Torque MRAM: The next revolution in memory[C]//2022 IEEE International Memory Workshop (IMW). Piscataway, NJ: IEEE, 2022: 1-4.

[57] Zhang K, Cao K H, Zhang Y, et al. Rectified tunnel magnetoresistance device with high on/off ratio for in-memory computing[J]. IEEE Electron Device Letters, 2020, 41(6): 928-931.

[58] Garello K, Yasin F, Kar G S. Spin-Orbit torque MRAM for ultrafast embedded memories: From fundamentals to large scale technology integration[C]//2019 IEEE 11th International Memory Workshop (IMW). Piscataway, NJ: IEEE, 2019: 1-4.

[59] Weisheit M, Fähler S, Marty A, et al. Electric field-induced modification of magnetism in thin-film ferromagnets[J]. Science, 2007, 315(5810): 349-351.

[60] Nozaki T, Yamamoto T, Miwa S, et al. Recent progress in the voltage-controlled magnetic anisotropy effect and the challenges faced in developing voltage-torque MRAM[J]. Micromachines, 2019, 10(5): 327.

[61] Chanthbouala A, Garcia V, Cherifi R O, et al. A ferroelectric memristor[J]. Nature Materials, 2012, 11(10): 860-864.

[62] Schroeder U, Park M H, Mikolajick T, et al. The fundamentals and applications of ferroelectric HfO_2[J]. Nature Reviews Materials, 2022, 7(3): 653-669.

[63] Fuller E J, Keene S T, Melianas A, et al. Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing[J]. Science, 2019, 364 (6440): 570-574.

[64] Valasek J. Piezo-Electric and allied phenomena in rochelle salt[J]. Physical Review, 1921, 17(4): 475-481.

[65] Böscke T S, Müller J, Bräuhaus D, et al. Ferroelectricity in hafnium oxide thin films[J]. Applied Physics Letters, 2011, 99(10): 102903.

[66] Fujii S, Kamimuta Y, Ino T, et al. First demonstration and performance improvement of ferroelectric HfO_2-based resistive switch with low operation current and intrinsic diode property[C]//2016 IEEE Symposium on VLSI Technology. Piscataway, NJ: IEEE, 2016: 1-2.

[67] Max B, Hoffmann M, Slesazeck S, et al. Ferroelectric tunnel junctions based on ferroelectric-dielectric $Hf_{0.5}Zr_{0.5}O_2/A1_2O_3$ capacitor stacks[C]//2018 48th European Solid-State Device Research Conference (ESSDERC). Piscataway, NJ: IEEE, 2018: 142-145.

[68] Yu S, Hur J, Luo Y C, et al. Ferroelectric HfO_2-based synaptic devices: Recent trends and prospects[J]. Semiconductor Science and Technology, 2021, 36(10): 104001.

[69] Chang M F, Shen S J, Liu C C, et al. An offset-tolerant fast-random-read current-sampling-based sense amplifier for small-cell-current nonvolatile memory[J]. IEEE Journal of Solid-State Circuits, 2013, 48(3): 864-877.

[70] Luo Y C, Hur J, Yu S M. Ferroelectric tunnel junction based crossbar array design for neuro-inspired computing[J]. IEEE Transactions on Nanotechnology, 2021, 20: 243-247.

[71] IEEE international roadmap for devices and systems[EB/OL]. [2022-10-13]. https://irds.ieee.org/.

[72] Liu Q, Gao B, Yao P, et al. A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing[C]//2020 IEEE International Solid-State Circuits Conference. Piscataway, NJ: IEEE, 2020: 500-502.

[73] Wang L F, Ye W, Lai J R, et al. A 14 nm 100 Kb 2T1R transpose RRAM with >150X resistance ratio enhancement and 27.95% reduction on energy-latency product using low-power near threshold read operation and fast data-line current stabling scheme[C]//2021 Symposium on VLSI Technology. Piscataway, NJ: IEEE, 2021: 1-2.

[74] Wu T F, Le B Q, Radway R, et al. A 43pJ/cycle nonvolatile microcontroller with 4.7 μs shutdown/wake-up integrating 2.3-bit/cell resistive RAM and resilience techniques[C]//2019 IEEE International Solid- State Circuits Conference. Piscataway, NJ: IEEE, 2019: 226-228.

[75] Jiang H, Han L L, Lin P, et al. Sub-10 nm Ta channel responsible for superior performance of a HfO_2 memristor [J]. Scientific Reports, 2016, 6(1): 28525.

[76] Golonzka O, Arslan U, Bai P, et al. Non-Volatile RRAM embedded into 22FFL FinFET technology[C]// 2019 Symposium on VLSI Technology. Piscataway, NJ: IEEE, 2019: T230-T231.

[77] Ho C, Chang S-C, Huang C-Y, et al. Integrated HfO_2-RRAM to achieve highly reliable, greener, faster, cost-effective, and scaled devices[C]//2017 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2017: 2.6.1-2.6.4.

[78] Yan X B, Qin C Y, Lu C, et al. Robust $Ag/ZrO_2/WS_2/Pt$ memristor for neuromorphic computing[J]. ACS Applied Materials & Interfaces, 2019, 11(51): 48029-48038.

[79] Sun Y, Xu H, Liu S, et al. Short-Term and long-term plasticity mimicked in low-voltage Ag/GeSe/TiN electronic synapse[J]. IEEE Electron Device Letters, 2018, 39(4): 492-495.

[80] Li J, Xu H, Sun S Y, et al. In situ learning in hardware compatible multilayer memristive spiking neural network [J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(2): 448-461.

[81] Yu J, Xu X X, Gong T C, et al. Suppression of filament overgrowth in conductive bridge random access memory by Ta_2O_5/TaO_x bi-layer structure[J]. Nanoscale Research Letters, 2019, 14(1): 111.

[82] Xu R, Jang H, Lee M-H, et al. Vertical MoS_2 double-

layer memristor with electrochemical metallization as an atomic-scale synapse with switching thresholds approaching 100 mV[J]. Nano Letters, 2019, 19(4): 2411-2417.

[83] Jang H, Liu C, Hinton H, et al. An atomically thin optoelectronic machine vision processor[J]. Advanced Materials, 2020, 32(36): 2002431.

[84] Lim S, Sung C, Kim H, et al. Improved synapse device with MLC and conductance linearity using quantized conduction for neuromorphic systems[J]. IEEE Electron Device Letters, 2018, 39(2): 312-315.

[85] He M Z, He D, Qian H, et al. Ultra-Low program current and multilevel phase change memory for high-density storage achieved by a low-current SET pre-operation[J]. IEEE Electron Device Letters, 2019, 40(10): 1595-1598.

[86] Jia S, Li H, Gotoh T, et al. Ultrahigh drive current and large selectivity in GeS selector[J]. Nature Communications, 2020, 11(1): 4636.

[87] Navarro G, Sabbione C, Bernard M, et al. Highly Sb-rich Ge-Sb-Te engineering in 4 Kb phase-change memory for high speed and high material stability under cycling[C]//2019 IEEE 11th International Memory Workshop (IMW). Piscataway, NJ: IEEE, 2019: 1-4.

[88] Jiang Y H, Zhou H Y, Zhu D Q, et al. Computational study for spin-orbit torque magnetic random access memory[C]//2021 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2021: 8.2.1-8.2.4.

[89] Yang M Y, Deng Y C, Wu Z H, et al. Spin logic devices via electric field controlled magnetization reversal by spin-orbit torque[J]. IEEE Electron Device Letters, 2019, 40(9): 1554-1557.

[90] Song Y J, Lee J H, Han S H, et al. Demonstration of highly manufacturable STT-MRAM embedded in 28 nm logic[C]//2018 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2018: 18.2.1-18.2.4.

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

[92] Chang T-C, Chiu Y-C, Lee C-Y, et al. A 22 nm 1 Mb 1024 b-read and near-memory-computing dual-mode STT-MRAM macro with 42.6 GB/s read bandwidth for security-aware mobile devices[C]//2020 IEEE International Solid- State Circuits Conference. Piscataway, NJ: IEEE, 2020: 224-226.

[93] Xi Z N, Ruan J J, Li C, et al. Giant tunnelling electroresistance in metal/ferroelectric/semiconductor tunnel junctions by engineering the Schottky barrier[J]. Nature Communications, 2017, 8(1): 15217.

[94] Yu J, Li Y, Sun W X, et al. Energy efficient and robust reservoir computing system using ultrathin (3.5 nm) ferroelectric tunneling junctions for temporal data learning [C]//2021 Symposium on VLSI Technology. Piscataway, NJ: IEEE, 2021: 1-2.

[95] Seo M, Kang M-H, Jeon S-B, et al. First demonstration of a logic-process compatible junctionless ferroelectric FinFET synapse for neuromorphic applications[J]. IEEE Electron Device Letters, 2018, 39(9): 1445-1448.

[96] Chung W, Si M, Ye P D. First demonstration of Ge ferroelectric nanowire FET as synaptic device for online learning in neural network with high number of conductance state and Gmax/Gmin[C]//2018 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2018: 15.2.1-15.2.4.

[97] Boyn S, Girod S, Garcia V, et al. High-performance ferroelectric memory based on fully patterned tunnel junctions[J]. Applied Physics Letters, 2014, 104(5): 052909.

[98] Ojha V K, Abraham A, Snášel V. Metaheuristic design of feedforward neural networks: A review of two decades of research[J]. Engineering Applications of Artificial Intelligence, 2017, 60: 97-116.

[99] Burr G W, Shelby R M, Sidler S, et al. Experimental demonstration and tolerancing of a large-scale neural network (165000 synapses) using phase-change memory as the synaptic weight element[J]. IEEE Transactions on Electron Devices, 2015, 62(11): 3498-3507.

[100] Li C, Wang Z, Rao M, et al. Long short-term memory networks in memristor crossbar arrays[J]. Nature Machine Intelligence, 2019, 1(1): 49-57.

[101] Tanaka G, Yamane T, Héroux J B, et al. Recent advances in physical reservoir computing: A review[J]. Neural Networks, 2019, 115: 100-123.

[102] Moon J, Ma W, Shin J H, et al. Temporal data classification and forecasting using a memristor-based reservoir computing system[J]. Nature Electronics, 2019, 2 (10): 480-487.

[103] Gui J, Sun Z N, Wen Y G, et al. A review on generative adversarial networks: Algorithms, theory, and ap-

plications[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, DOI: 10.1109/TKDE.2021.3130191)

[104] Lin Y D, Wu H Q, Gao B, et al. Demonstration of generative adversarial network by intrinsic random noises of analog RRAM devices[C]//2018 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2018: 3.4.1-3.4.4.

[105] Yao P, Wu H Q, Gao B, et al. Face classification using electronic synapses[J]. Nature Communications, 2017, 8(1): 1-8.

[106] Mochida R, Kouno K, Hayata Y, et al. A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture[C]//2018 IEEE Symposium on VLSI Technology. Piscataway, NJ: IEEE, 2018: 175-176.

[107] Ishii M, Kim S, Lewis S, et al. On-chip trainable 1.4M 6T2R PCM synaptic array with 1.6K stochastic LIF neurons for spiking RBM[C]//2019 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2019: 14.2.1-14.2.4.

[108] Singh S, Sarma A, Jao N, et al. NEBULA: A neuromorphic spin-based ultra-low power architecture for SNNs and ANNs[C]//2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA). Piscataway, NJ: IEEE, 2020: 363-376.

[109] Laguna A F, Yin X, Reis D, et al. Ferroelectric FET based in-memory computing for few-shot learning[C]// Proceedings of the 2019 on Great Lakes Symposium on VLSI. New York: Association for Computing Machinery, 2019: 373-378.

[110] Li C, Hu M, Li Y N, et al. Analogue signal and image processing with large memristor crossbars[J]. Nature Electronics, 2018, 1(1): 52-59.

[111] Zhao H, Liu W Z, Tang J S, et al. Memristor-based signal processing for edge computing[J]. Tsinghua Science and Technology, 2022, 27(3): 455-471.

[112] Zhao H, Liu Z W, Tang J S, et al. Implementation of discrete Fourier transform using RRAM arrays with quasi-analog mapping for high-fidelity medical image reconstruction[C]//2021 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2021: 12.4.1-12.4.4.

[113] Zhang B, Uysal N, Ewetz R. Computational restructuring: Rethinking image processing using memristor crossbar arrays[C]//2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). Piscataway, NJ: IEEE, 2020: 1594-1597.

[114] Le Gallo M, Sebastian A, Cherubini G, et al. Compressed sensing with approximate message passing using in-memory computing[J]. IEEE Transactions on Electron Devices, 2018, 65(10): 4304-4312.

[115] Choi S, Sheridan P, Lu W D. Data clustering using memristor networks[J]. Scientific Reports, 2015, 5(1): 10492.

[116] Choi S, Shin J H, Lee J, et al. Experimental demonstration of feature extraction and dimensionality reduction using memristor networks[J]. Nano Letters, 2017, 17(5): 3113-3118.

[117] Jiang Y N, Kang J F, Wang X N. RRAM-based parallel computing architecture using k-nearest neighbor classification for pattern recognition[J]. Scientific Reports, 2017, 7(1): 45233.

[118] Liu C, Han R Z, Zhang S, et al. A high accuracy and robust machine learning network for pattern recognition based on binary RRAM devices[C]//2017 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA). Piscataway, NJ: IEEE, 2017: 1-2.

[119] Liang X P, Zhong Y N, Tang J S, et al. Rotating neurons for all-analog implementation of cyclic reservoir computing[J]. Nature Communications, 2022, 13(1): 1549.

[120] Fu Y Y, Zhou Y, Huang X D, et al. Forming-free and annealing-free V/VOx/HfWOx/Pt device exhibiting reconfigurable threshold and resistive switching with high speed (<30 ns) and high endurance (>1012/>1010) [C]//2021 IEEE International Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2021: 12.6.1-12.6.4.

[121] Christensen D V, Dittmann R, Linares-Barranco B, et al. 2022 roadmap on neuromorphic computing and engineering[J]. Neuromorphic Computing and Engineering, 2022, 2(2): 022501.

Review of recent research on memristors and computing-in-memory applications

JIANG Zhixing, XI Yue, TANG Jianshi*, GAO Bin, QIAN He, WU Huaqiang*

School of Integrated Circuits, Beijing Advanced Innovation Center for Integrated Circuits, Tsinghua University, Beijing 100084, China

Abstract The rapid development of deep learning raises a massive demand for computing power. However, traditional silicon-based chips based on the von Neumann architecture with physically separated memory and computing units, are facing critical issues such as the "memory wall", and hence the increase of chip computing power is gradually hitting a bottleneck. To address this problem, researchers have been inspired by the working mechanism of biological brain and proposed a computing-in-memory architecture based on memristors. This novel architecture is expected to achieve several orders of magnitude improvement in energy efficiency and speed over the von Neumann architecture for tasks such as artificial neural networks. It is one of the most promising technologies to achieve ultra-low power consumption and ultra-high computing power. This article first reviews the working mechanisms of various types of memristors, and summarizes the latest device research internationally. Then, the progress on application demonstrations of memristor-based computing-in-memory chips such as neural networks, signal processing, and machine learning are reviewed. The current challenges in this field and further research directions are concluded in the end.

Keywords memristor; brain-inspired computing; computing-in-memory; neural networks; signal processing

（责任编辑　王志敏）