doi:10.11887/j.cn.202301001 http://journal.nudt.edu.cn

忆阻器类脑计算芯片研究现状综述*

陈长林,骆畅航,刘 森,刘海军
(国防科技大学 电子科学学院,湖南 长沙 410073)

摘 要:为把握忆阻类脑芯片发展现状并总结其发展趋势,对现有忆阻类脑计算芯片与架构进行了调研,对芯片中所采用的忆阻器阵列结构和集成工艺、前神经元电路、后神经元电路、多阵列互连拓扑结构与数据传输策略,以及芯片设计过程中所采用的系统仿真和评估方法进行了对比分析。总结出当前忆阻类脑计算芯片电路设计仍需解决忆阻器可用阻态少、器件参数波动性大、阵列外围电路复杂、集成规模小等问题,并指出了该类芯片走向实际应用仍然面临着忆阻器生产工艺提升、完善开发工具支持、专用指令集开发、确定典型牵引性应用等挑战。

关键词:忆阻器;类脑计算;存算一体;加速芯片;低功耗
中图分类号:TP35 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2023)01-001-14

Review on the memristor based neuromorphic chips

CHEN Changlin, LUO Changhang, LIU Sen, LIU Haijun
(College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: In order to master the current development status and development trends of memristor based neuromorphic chips, the existing memristor based neuromorphic chips and architectures were investigated. The memristor array structure and integration process, anterior and posterior neuron circuits, multi-array interconnection topology and data transmission strategy used in the chip, as well as the system simulation and evaluation methods used in the chip design process were compared and analyzed. It is concluded that the current circuit design of memristor based neuromorphic chips still need to solve the problems of limited resistance states, large device parameter fluctuation, complex array peripheral circuits, small integration scale, etc. It is pointed out that the actual application of this type of chip still faces challenges such as the improvement of memristor production process, improvement of development tool support, special instruction set development, and determination of typical traction applications.

Keywords: memristor; neuromorphic computing; processing-in-memory; acceleration chip; low power consumption

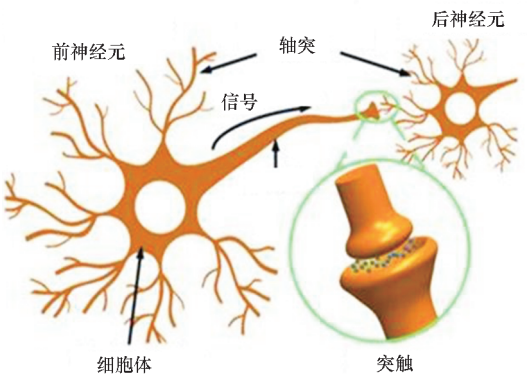
随着以卷积神经网络(convolutional neural networks, CNN)、深度神经网络(deep neural networks, DNN)、递归神经网络(recurrent neural networks, RNN)等为代表的神经网络算法[1]的不断成熟,人工智能(artificial intelligence, AI)在自动驾驶、语音与图像识别、知识搜索、语义理解等众多应用领域获得了广泛应用[2]。为了提高神经网络推理计算的效率和能效,涌现了TPU[3]、DaDianNao[4]、Tianjic[5]、Thinker[6]等多款采用定制架构的AI加速芯片。然而,上述AI芯片本质上仍然采用了存算分离的冯·诺伊曼架构,在计算时需要将数据在存储单元与运算单元之间进行频繁搬移,因而仍然具有较高的延迟和能耗[7]。相较而言,人脑中由于神经突触具有网络权值存储和计算功能(见图1(a))而具有极高的能效。因此,借鉴人脑结构和信息处理机制,开发类脑芯片,突破冯·诺伊曼架构瓶颈实现高能效智能计算[8]成为当前研究的热点。

神经网络由多层神经元以及神经元之间的突触连接组成,其中神经突触数量远远超过神经元数量。例如在人脑中,神经突触数量(10^{14}~10^{15})约为神经元数量的10^3倍。同时,神经网络推理计算过程主要由神经元输入向量与突触权重矩阵的卷积(卷积层)或相乘(全连接层)组成,占所有计算的90%以上[4],因此开发高性能的人工神经突触是类脑芯片设计的关键所在。而忆阻器是

* 收稿日期:2021-03-29
基金项目:国家自然科学基金资助项目(61974164,62074166,61804181,61704191,62004219,62004220)
作者简介:陈长林(1986—),男,山东泰安人,副研究员,博士,硕士生导师,E-mail:chenchanglin@nudt.edu.cn;
刘海军(通信作者),男,山东德州人,副教授,博士,硕士生导师,E-mail:liuhaijun@nudt.edu.cn

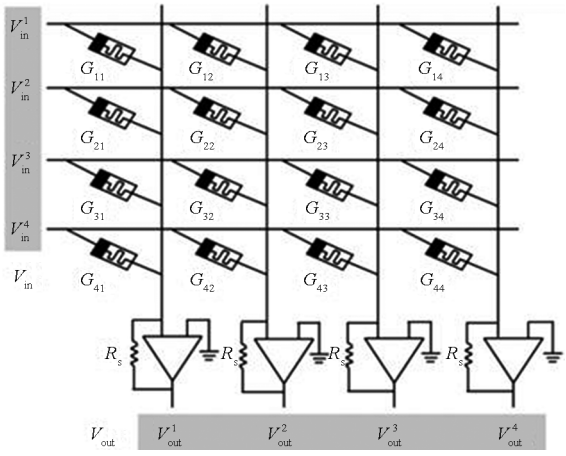
实现人工神经突触的最佳基础器件选择之一:①忆阻器阻值可以在外部信号作用下进行调整,因此能够模拟神经突触的可塑性;②忆阻器是非易失器件,可以用于储存突触连接权重;③流经忆阻器的电流大小由其电导值与电压值相乘得到,对应突触权重与神经元激励的加权相乘运算。同时忆阻器具有结构简单、功耗低、尺寸小、可三维集成、与CMOS工艺兼容等优点,因此成为大规模类脑计算基础器件的最佳选择,获得了大量关注与研究[9]。

在多层神经网络中,相邻两层神经元的突触连接可采用如图1(b)所示忆阻器交叉阵列实现,阵列中每一个交叉点处的忆阻器存储了两个神经元之间的突触连接强度,多个阵列组合可以实现深度神经网络。输入信号以电压的形式加载到忆阻器阵列每一行,与忆阻器作用产生的电流在列线上相累加并被感知电路读取,在单个周期内得到乘累加计算结果,即 $I_j = \sum V_i G_{ij}$。忆阻器交叉阵列同时实现了权值存储和乘累加运算的功能,其中乘法运算的并行度与忆阻器数量相同,累加运算的并行度与阵列的行或列数(取决于阵列计算单元结构)相同。在外围反馈电路配合下,还可通过训练实现忆阻器阻值亦即突触连接强度的在线调整,实现在线学习能力。



(a) 神经元与神经突触

(a) Neurons and synapses



(b) 忆阻器交叉阵列

(b) Memristor crossbar array

图1　生物神经元连接与忆阻器交叉阵列

Fig. 1　Neuron connections and memristor crossbar array

近年来,将忆阻器阵列与CMOS电路混合集成实现类脑芯片已成为研究的热点并取得了诸多研究成果。为把握相关研究进展并总结发展趋势,本文对现有设计进行调研分析。对比分析忆阻类脑芯片中各基础功能模块的实现方案,总结当前研究仍需解决的问题,并指出该类芯片走向实际应用所面临的挑战。

1　现有忆阻类脑计算芯片与架构

基于忆阻器的类脑芯片研究大致可以分为忆阻器阵列研究、忆阻类脑芯片架构研究以及全功能忆阻类脑芯片研究三个方面。其中,忆阻器阵列研究的目的在于优化阵列生产工艺、提取阵列性能参数并搭建板级系统展示其类脑计算应用可行性,忆阻类脑芯片架构研究根据应用需求及阵列性能设计新型体系结构并进行仿真验证,全功能忆阻类脑芯片研究开发完整功能芯片并推广应用。三个方面的研究相互检验、相互促进。

1.1　忆阻器阵列研究

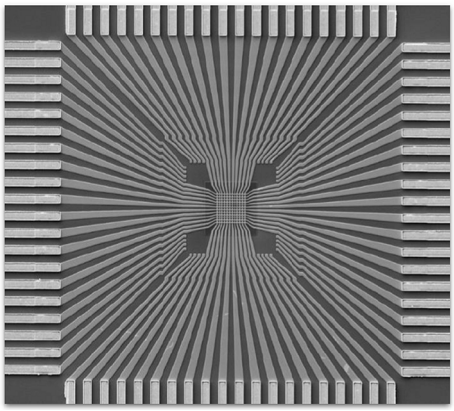
大量忆阻类脑芯片研究以忆阻器阵列芯片+板级电路形式展开,主要进行阵列中器件性能测试并展示基于忆阻器实现多种类脑计算应用的可能性。

利用忆阻器的权值存储和模拟计算能力,忆阻器阵列可用于实现非脉冲型神经网络。如Alibart等[10]于2013年基于2×10规模忆阻器阵列采用离线和在线训练两种方式实现了字母"X"和"T"的分辨,随后该团队Prezioso等[11]于2015年在更大规模(12×12)1R忆阻器阵列上(见图2(a)),通过在线学习方式实现了对3种3×3黑白字母图像的分类,首次通过实验验证了基于忆阻器阵列运行单层感知机网络实现模式识别的可行性。

充分利用忆阻器的阻值可调性能,忆阻器阵列还可用于实现脉冲型神经网络[12]。如Covi等[13]利用脉冲时间依赖可塑性(spike-timing dependent plasticity, STDP)机制,设计实现了包含25个前神经元、5个后神经元和125个神经突触的脉冲神经网络,经过无监督学习训练后的网络可以识别5×5黑白像素的5个字母。Choi等[14]利用赫布学习法则基于9×2忆阻器阵列通过无

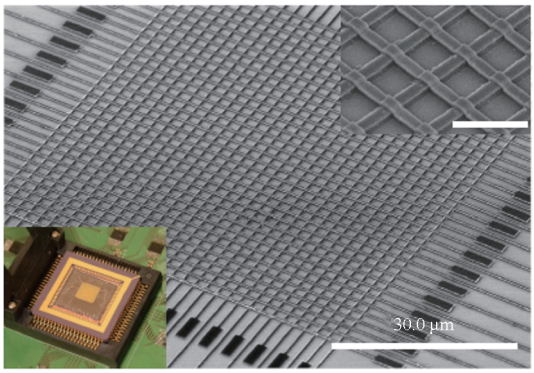
监督学习实现了主成分分析能力。

随着忆阻器阵列生产工艺的不断优化,阵列规模逐渐增大,所能实现的计算也日益复杂。如 Sheridan 等[15]基于 32×32 忆阻器阵列实现了灰度图像稀疏编码功能(见图 2(b))。Li 等[16]基于 128×64 规模的可重构忆阻器交叉阵列(见图 2(c))实现了信号处理、图像压缩、卷积滤波与增强学习[17]等处理。Yao 等[18]基于 1T1R 结构的 128×8 规模阵列实现了 320 像素灰白人脸图像识别功能,其功耗与在英特尔 Xeon Phi 处理器上实现相比降低为原来的 1/1 000。



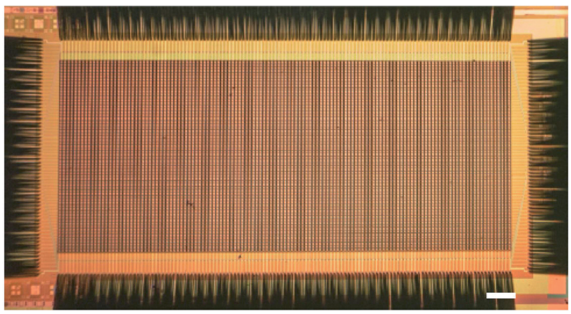
(a) 12×12 阵列芯片[11]

(a) 12×12 sized crossbar[11]



(b) 32×32 阵列芯片[15]

(b) 32×32 sized crossbar[15]



(c) 128×64 阵列芯片[16]

(c) 128×64 sized crossbar[16]

图 2　部分纯忆阻器阵列芯片

Fig. 2　Some pure memristor crossbar chip

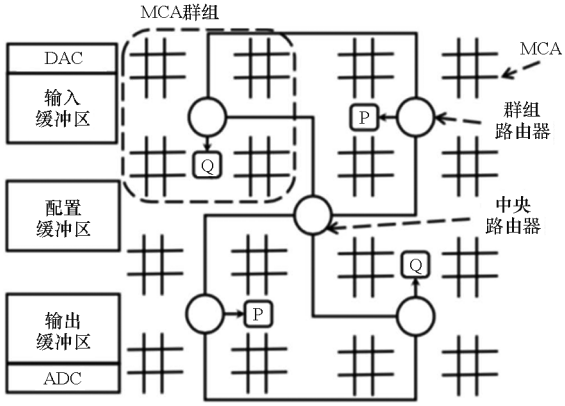
上述工作中,忆阻器阵列芯片只实现了突触阵列的功能,权值读写、待处理信号加载与计算结果读取等功能单元电路仍需在板卡上实现。但相关工作已完整验证了忆阻器类脑计算架构和处理流程,展现了忆阻器类脑计算在图像处理、稀疏编码、信号处理等复杂的非结构化数据处理中,相比传统处理器件具有巨大的能效和面积(集成度)优势,为全功能类脑芯片的设计提供了架构、电路设计依据。

1.2　忆阻类脑计算芯片架构研究

在忆阻器阵列研究成果基础上,根据神经网络计算需求,多个研究团队提出了新型忆阻类脑计算芯片体系架构并进行了仿真验证。

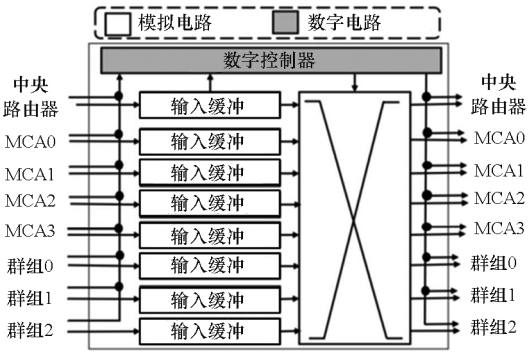
美国匹兹堡大学研究团队在 2015 年提出一款名为“RENO”[19]的忆阻类脑计算架构。如图 3所示,各忆阻器阵列之间采用星型拓扑结构进行互联,并采用混合信号进行传输,其中数字信号用于路由路径的选择,模拟信号用于传输阵列计算结果,以降低 CPU 与各计算单元之间的通信延迟。该结构第一次探讨了阵列间片上互联策略以及神经网络加速单元与控制处理器之间的协作方式。然而阵列之间的模拟通信方式决定了神经网络必须以全展开的方式运行,无法通过权重和中间结果复用等[2]方式降低硬件实现代价。

谢源团队联合清华大学和惠普实验室于 2016 年提出一种名为“PRIME”[20]的处理架构,如图 4 所示,以期打破未来计算系统的“存储墙”限制。其中的忆阻器阵列分为存储单元和全功能单元。全功能单元包含了解码与驱动电路、读出电路、缓存单元等,使得其既可用于数据存储,又



(a) RENO 系统架构

(a) RENO system architecture



（b）混合信号路由结构

（b）Mix signal router structure

图3　RENO系统体系结构[19]

Fig. 3　RENO system architecture[19]

可配置为神经网络计算加速模块。与多款基于数字逻辑电路的神经网络加速系统相比，具有巨大的能效提升。

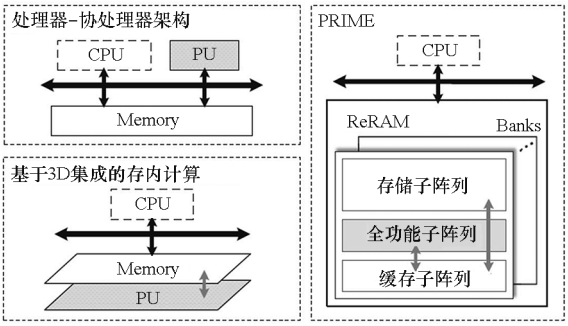


图4　PRIME结构框图[20]

Fig. 4　PRIME system architecture diagram[20]

Shafiee等[21]在2016年提出了“ISAAC”类脑芯片架构，如图5所示。该系统具有多个Tile，每个Tile集成多个原位乘累加单元（in-situ multiply-accumulate，IMA），每个IMA中包含多个忆阻器阵列，用于存储突触权重并执行点积计算。ISAAC的多核结构使得神经网络各卷积层可分配到不同的忆阻器阵列上采用并行流水方式工作，降低了对数据缓冲空间的要求，增加了吞吐量。然而该结构并未充分挖掘神经网络中的数据和权重的复用特性以进一步提高计算效率。

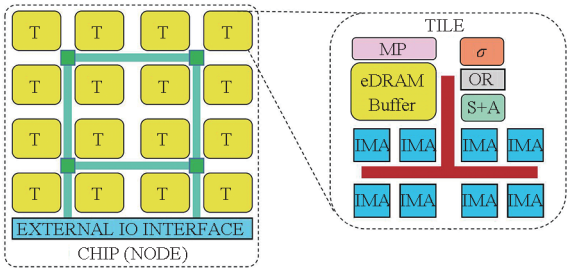


图5　ISAAC类脑芯片架构[21]

Fig. 5　ISAAC neuromorphic chip architecture[21]

在采用与ISAAC类似的层次化多阵列架构基础上，MAX²[22]采用了脉动阵列架构以及网络权重复制和输入特征图像复用等策略（见图6），提高了数据复用率，大大减少了处理单元之间的数据移动需求，提高了时间、能量和面积利用效率。通过在NeuroSim平台上运行VGGNet、ResNet和AlexNet网络评估显示，MAX²的计算能效相比ISAAC提高了5.2倍。

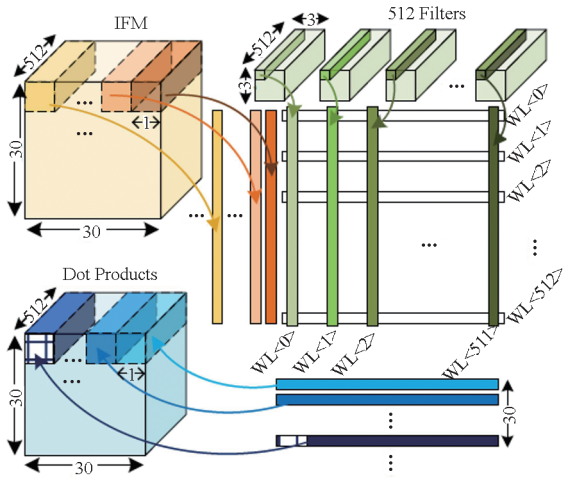


图6　MAX² 权重映射策略

Fig. 6　MAX² weight mapping strategy

南洋理工大学研究团队在2018年提出了一种基于3D CMOS-ReRAM的张量神经网络加速芯片[23]。如图7所示，它们的设计有三层：第一层为忆阻器交叉阵列层，阵列之间采用H-树拓扑结构进行互联，用于存储网络权重或待处理数据；第二层同为忆阻器交叉阵列层，通过硅通孔（through silicon via，TSV）从第一层接收张量核的数据实现乘累加运算；第三层为CMOS电路层，用于协调整个神经网络的工作和执行非线性映射。由于3D CMOS-ReRAM的这种方法没有进行权值复用，且结构相对固定，只能实现一些简单的神经网络结构。

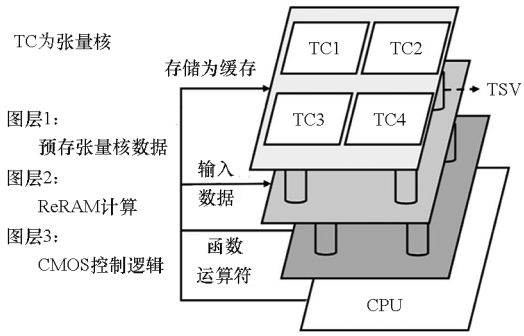


图7　3D CMOS-ReRAM加速芯片主体架构[23]

Fig. 7　3D CMOS-ReRAM accelerator architecture[23]

上述研究对基于多个忆阻器阵列实现复杂类脑计算所应采用的系统体系结构进行了有益的探索,其仿真结果证明了忆阻类脑计算相比于传统数字 AI 芯片在能效上的优势,为忆阻类脑芯片的发展提供了理论指导。然而,上述架构均采用了理想化的阵列模型,并假设忆阻器阵列与 CMOS 外围电路的混合集成工艺以及阵列间的互联通信结构均已具有了完善解决方案。而实际上,这些基础模块实现方案仍需通过流片进行验证。

1.3　全功能忆阻类脑芯片研究

随着忆阻器阵列生产工艺和计算架构的逐渐完善,大量研究开始尝试将忆阻器阵列与 CMOS 外围电路在单个芯片上进行集成以实现完整计算功能,并通过增加阵列数量来实现复杂神经网络。

清华大学吴华强团队基于 130 nm 工艺设计实现了如图 8 所示芯片[24]。该芯片可实现三层感知网络并具备在线学习能力。芯片中集成 2 个 784 × 100 规模和 2 个 100 × 10 规模 1T1R 忆阻阵列,分别用于实现输入神经元和隐藏神经元、隐藏神经元和输出神经元之间的突触连接。感知网络中的每个权重均由一对差分单元实现,隐藏神经元由可配置比较器实现。基于该芯片运行手写体识别网络与全精度网络相比仅有不到 5% 的精度损失。然而也可以看到,该芯片所采用结构主要针对手写体识别网络,无法用于其他类型网络。同时,除隐藏层外,原始待处理信号的加载与最终处理结果的读出仍需片外电路实现。2020 年,该团队将 2 个 1T1R 阵列优化为 1 个 2T2R 阵列,并改进了读出电路设计,将手写体识别网络精度损失降到了 2%[25]。

为通过实验验证多个忆阻器阵列实现多层卷积神经网络的能力,清华大学和马萨诸塞大学联合开发了如图 9 所示多阵列板级系统[26]。该系统搭载了 8 个基于忆阻器的处理单元(processing element,PE)芯片、ARM 核、DRAM 以及其他功能单元。每个 PE 芯片中集成 1 个 128 × 16 规模 1T1R 忆阻器阵列以及输入寄存器、输出寄存器、ADC 采样与移位相加等逻辑单元。每个忆阻器均可实现 32 种阻态。基于该系统运行 5 层 CNN 网络进行 MNIST 手写体识别可达到 97% 以上识别率,与 Tesla V100 型号 GPU 相比实现了 110 倍能效提升和 30 倍性能密度提升。

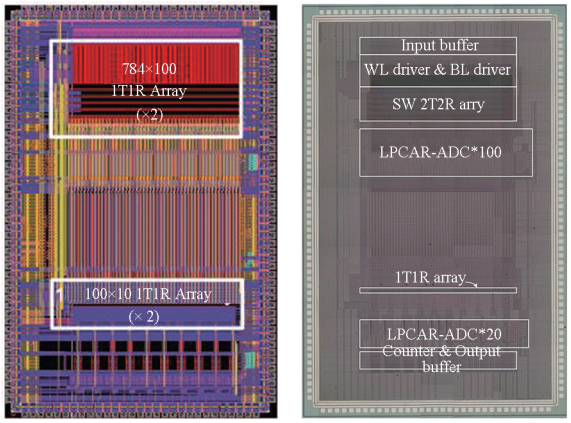
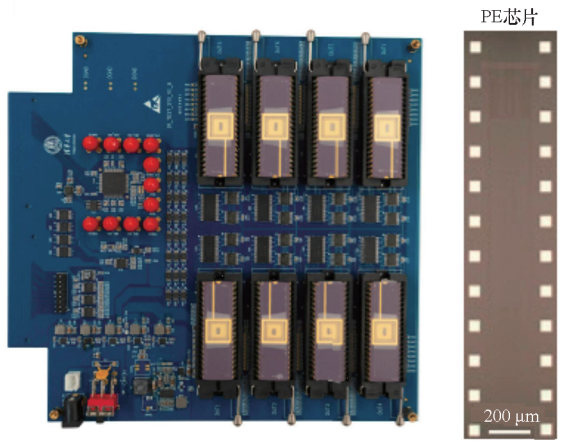


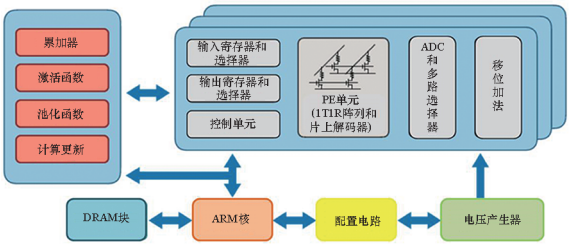
图 8　清华大学设计的两款忆阻类脑芯片[24-25]

Fig. 8　Neuromorphic chip designed by THU[24-25]



(a) 板级系统与芯片版图

(a) Board level system and chip layout



(b) 系统结构框图

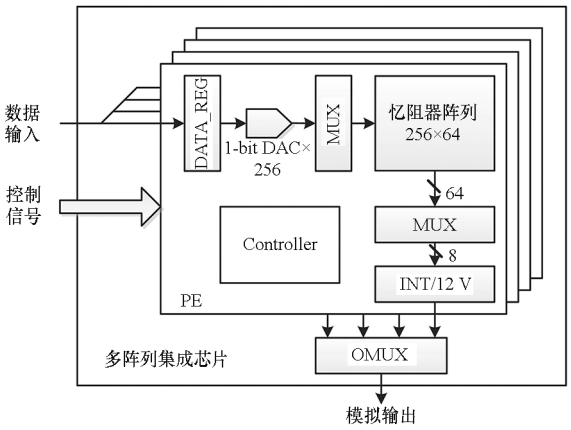
(b) System architecture

图 9　清华大学与马萨诸塞大学联合设计多阵列板级系统[26]

Fig. 9　Multi-crossbar system designed by THU and UMass[26]

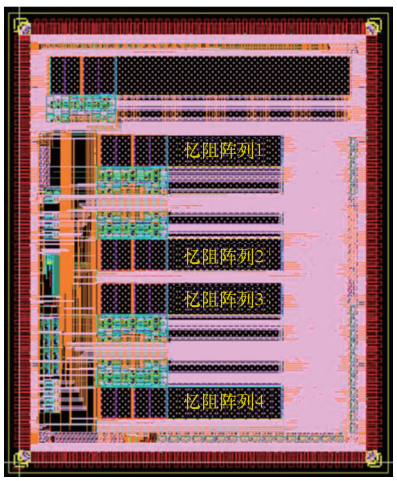
与上述工作不同,国防科技大学研究团队通过将多个阵列集成到单个芯片中进行多阵列协同工作验证。如图 10 所示,芯片基于 180 nm 工艺设计,集成了 4 个计算单元(PE)。其中每个 PE 由 256 × 64 规模忆阻器阵列、输入数据寄存器、1-bit DAC、多路选择器、控制逻辑、积分或电流电压转换等模块组成。待处理数据采用数字形式输入,在芯片内通过 1-bit DAC 转换为模拟脉冲后加载到忆阻器阵列上完成乘累加运算。阵列列线累积电流可在片内进行积分或者直接转换为电压。各 PE 计算结果在输出多路选择器(output multiplexer,OMUX)控制下轮流输出片外并通过

高精度 ADC 进行采样。芯片模拟部分电路设计时钟为 2 MHz,整个芯片具备 16.384 GOPS 计算能力。基于该芯片可实现多层神经网络手写体识别、灰度图像目标检测等应用。



(a) 芯片架构

(a) Chip architecture



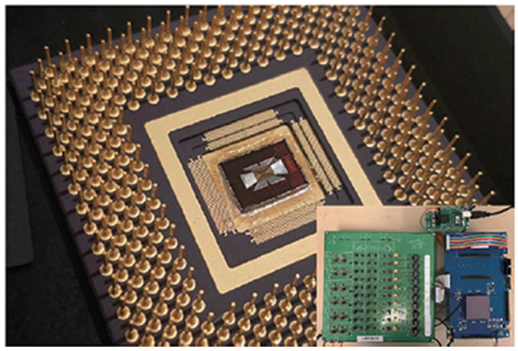
(b) 芯片版图

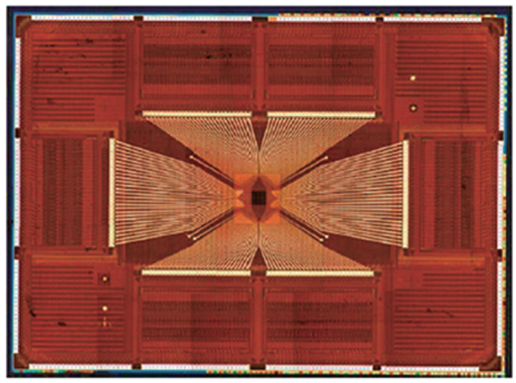
(b) Chip layout

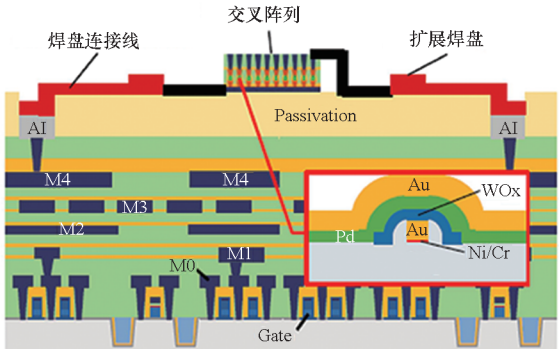
图 10　国防科技大学开发的多阵列集成芯片

Fig. 10　Multi memristor crossbar based neuromorphic chip designed by NUDT

密歇根大学 Cai 等[27]设计实现了如图 11 所示全功能芯片。该芯片集成了 54×108 规模 1R 忆阻器阵列、必要接口电路、数字总线与 OpenRISC 处理器等单元。从图中可以看到,芯片 CMOS 电路部分和忆阻器阵列部分通过一层钝化层相隔离,CMOS 电路管脚与忆阻器阵列行列线之间通过外部引线互联。因此,实际上是通过特殊工艺将 2 个芯片堆叠到了同一衬底上。阵列的每个行线和列线上均配备了 ADC 和 DAC 模块,使得每个行线或列线均可配置为输入或输出接口,增大了阵列使用的灵活度。OpenRISC 处理器用于控制实现忆阻器权重的读写以及待处理信号的加载和处理结果的读出。该芯片最高运行速率为 148 MHz,具有 57.5 GOPS 的峰值计算能力,可实现感知网络、稀疏编码、基于多层网络的主成分分析等应用。然而,由于芯片中使用了大量的 ADC 和 DAC,由此带来的较大面积代价降低了该结构的可扩展性。







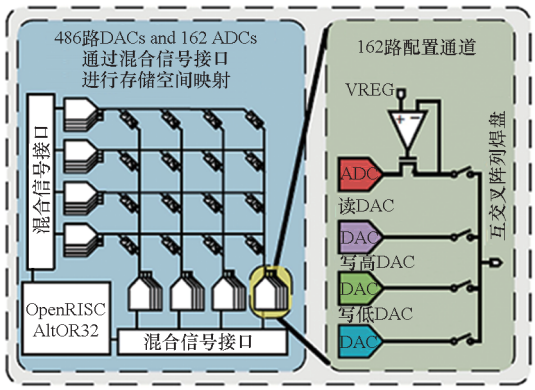


图 11　密歇根大学所设计的全功能忆阻器芯片[27]

Fig. 11　Full function memristor based neuromorphic chip designed by UMich[27]

上述研究通过具体的电路设计和流片工作对忆阻器阵列和 CMOS 外围电路的混合集成工艺进行了验证,对用于待处理数据加载的前神经元电路和用于计算结果读出的后神经元电路,以及其他控制电路的实际性能进行了测试,具备了基本的类脑计算功能,为实现集成更多忆阻器阵列和更复杂系统体系结构的类脑芯片奠定了基础。

2　关键技术与实现方案

通过上述工作可以看到,忆阻类脑计算单元至少需包含图 12 所示各功能模块。

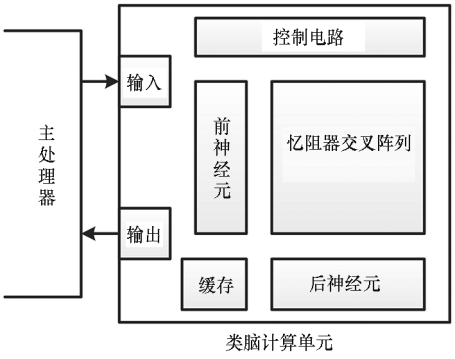


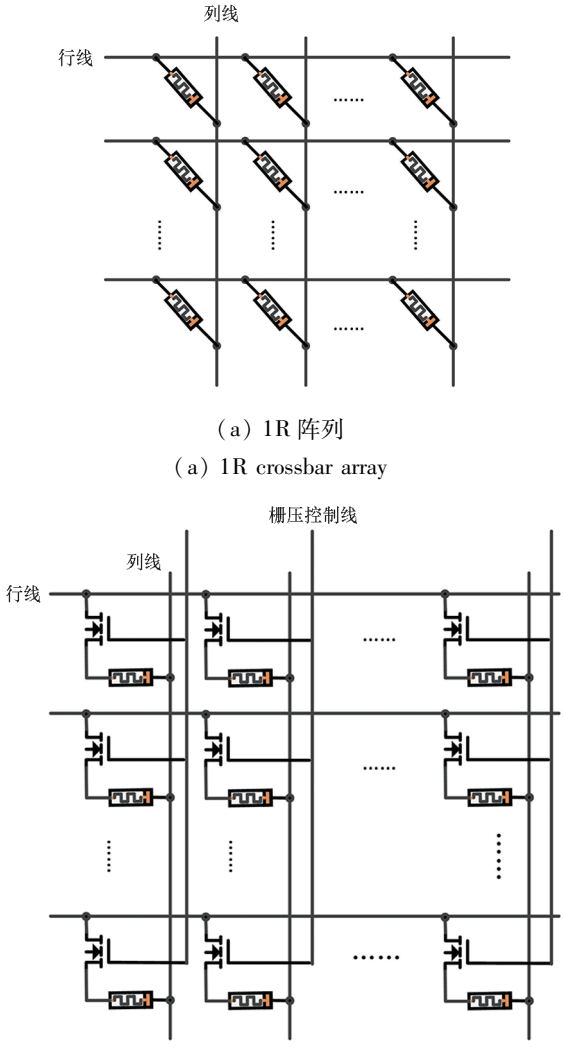
图 12　基于忆阻器的类脑芯片基本组成结构

Fig. 12　Basic structure of a memristor based neuromorphic chip

其中忆阻器阵列实现模拟乘累加计算,前神经元电路实现待处理数据的加载,后神经元电路实现阵列计算结果的读取和激活运算,缓存模块实现输入输出数据的暂存,输入或输出接口电路完成与主处理器的数据和指令通信,控制电路将主处理器指令转换为各功能模块的具体动作。实际上,为实现多层神经网络,还需多个类脑计算单元以一定的架构互联并协同工作。本节对各模块典型实现方案进行总结归纳。

2.1　忆阻器阵列结构与集成工艺

目前常用的忆阻器阵列有 1R(见图 13(a))和 1T1R(见图 13(b))2 种结构[28]。其中 1R 阵列中只有忆阻器,因此可以实现极高的集成密度。例如在密歇根大学所设计的忆阻类脑芯片[27]总面积为 61.64 mm^2,而 54 × 108 规模忆阻器阵列面积仅为 0.14 mm^2。然而对 1R 阵列中某个忆阻器进行阻值调整时所需的电压配置方案比较复杂,进行计算时各行之间存在旁路电流[29],因此使用较为不便。在 1T1R 阵列中,每个忆阻器与 1 个选通 CMOS 晶体管(T)进行串联,通过控制 T 的栅压实现对忆阻器偏压和电流的控制,可有效抑制旁路电流,并能够简化忆阻器阻值调整时所用的电压偏置方案,因此使用灵活,获得了广泛应用。其缺点在于进行忆阻器阻值调整时往往需要通过较大的电流,因此需要 T 具有较大的尺寸,导致阵列尺寸变大。



(a) 1R 阵列

(a) 1R crossbar array

(b) 1T1R 阵列

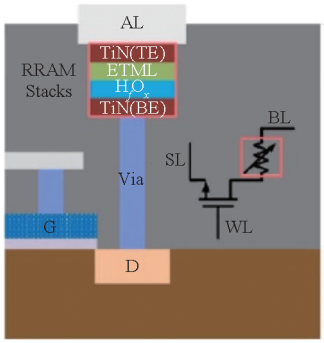
(b) 1T1R crossbar array

图 13　1R 阵列与 1T1R 阵列结构

Fig. 13　1R and 1T1R memristor crossbar array structure

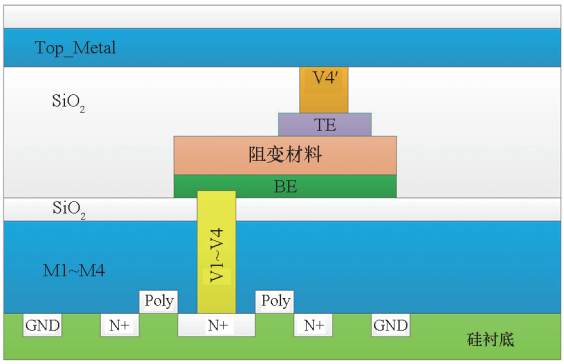
忆阻器为三层结构,由底电极、阻变层和顶电极组成。其中顶电极与底电极相重叠部分为忆阻器件。在 1T1R 忆阻器阵列芯片中,通常将忆阻器夹在两层金属层中间,忆阻器的两端分别通过垂直过孔与选通 CMOS 晶体管的源、漏极和阵列列、行线相连[28]。根据忆阻器与连接底电极的过孔的位置关系不同,可以采用图 14 所示 2 种生产工艺。在图 14 (a)中[30],忆阻器与连接底电极的过孔相重叠,且三层结构尺寸相同,在制作工艺中,使用 1 套光刻掩模版即可完成忆阻器三层结构的制作。然而,连接底电极的钨栓塞过孔通常会由于不完全填充在中间形成孔洞(见图 15),进

而导致忆阻器三明治结构损伤，器件生产良率下降[31-32]。为解决这一问题，国防科技大学采用了图14(b)所示工艺，即将忆阻器底电极向侧方引出一部分，然后在平整位置生长阻变层和顶电极以保证忆阻器底部的平整，进而提高器件良率。



(a) 忆阻器与连接底电极过孔位置重叠[30]

(a) Memristors overlay with the BE via[30]



(b) 忆阻器与连接底电极过孔位置偏移

(b) Memristor position shifts away from the BE via

图14　1T1R阵列2种制作工艺

Fig. 14　Two types of memristor crossbar fabricate process

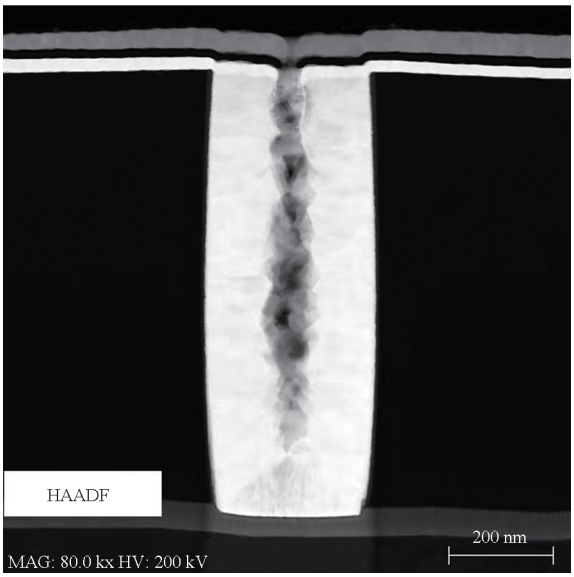


图15　钨栓塞过孔缺陷

Fig. 15　Defects in the via

2.2　前神经元电路

忆阻器阵列采用模拟方式完成乘累加计算，而待处理数据通常为数字形式，因此前神经元电路需要将数字信息转换为模拟脉冲后加载到忆阻器阵列行线上。常用的转换方式如下。

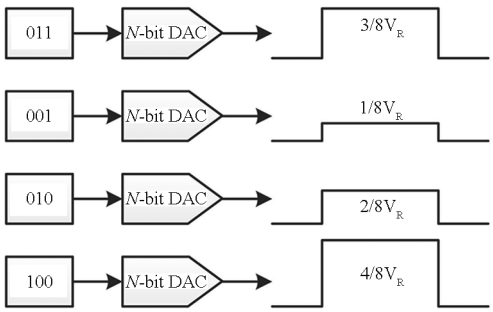
多位精度DAC直接转换：如图16(a)所示，通过多位精度的DAC将数字信息转换成具有一定幅度的模拟电平，是应用最广泛的转换方式[19-20,27,33]之一。其优点在于转换速度快，且与板级验证系统实现方式(DAC芯片转换)一致；缺点在于该方案对DAC精度要求高，且在芯片中集成大量DAC模块会引入较大的面积和功耗代价。

数值-时间转换：如图16(b)所示，通过数值-时间转换器将待处理数据转换成具有固定幅度和对应宽度的模拟脉冲。其优点是转换电路简单，仅需一个模拟传输门和一个计数器即可实现；缺点是随着待处理数据精度的提高，所转换成的模拟脉冲宽度呈指数增长。

1-bit DAC逐位转换：如图16(c)所示，此方案[21]模拟了二进制乘法实现方式，即从待处理数据最低或高位开始逐位读取，根据该位数据为“1”或“0”来确定是否输出具有固定幅度和固定脉宽的模拟脉冲，将计算结果缓存，并与下一位数据的计算结果进行移位相加。通过多次迭代完成对一组输入数据的处理。此方案的优点在于所采用的电压幅度单一，电路实现简单且精度高；缺点在于计算时间与待处理数据位数成正比，与多位精度DAC直接转换方案相比处理延时较长。

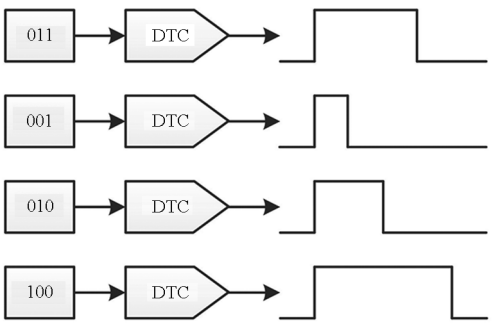
脉冲数量编码：如图16(d)所示，此方案将待处理数据转换为一定数量的脉冲作为阵列的输入[24,36]，每个脉冲的幅度和宽度固定，适宜于脉冲神经网络应用。由于所需脉冲数量随待处理数据位数呈指数增长，因此计算延迟较高。

位片式编码：将待处理数据的所有比特位分为多组，然后将每组数值转换成对应的脉冲数量或脉冲宽度[37]。实际上是将1-bit DAC逐位转换



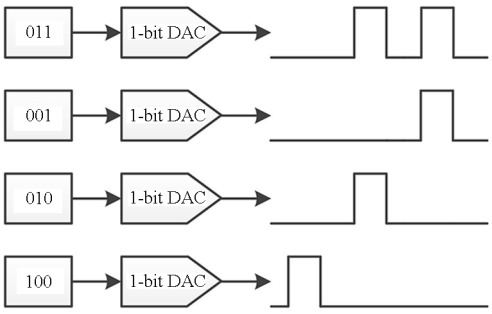
(a) N-bit DAC直接转换

(a) N-bit DAC direct converting



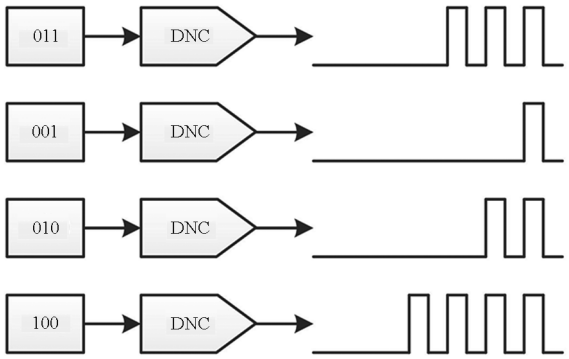
(b) 数值 - 时间转换

(b) Data-time converting



(c) 1-bit DAC 逐位转换

(c) 1-bit DAC converting by bit



(d) 数值转换为脉冲数量

(d) Convert data value to spike number

图 16　前神经元电路的实现方式

Fig. 16　Pre-synapse neuron circuits

与脉冲数量编码结合的一种方式。

上述转换方式中,以 1-bit DAC 逐位转换实现方式最为简单,仅需一个参考电平和一个传输门即可实现,所需的面积和功耗代价均较小,有望在后续忆阻类脑芯片设计中获得广泛应用。需要说明的是,前神经元电路具体实现方式通常根据后神经元电路实现方式确定,以便实现所需的计算结果读取精度和读取速度。

2.3　后神经元电路

后神经元电路用于将忆阻器阵列列线上的累加电流读出并转为数字结果,经缓存、激活等处理后传输至其他阵列进行后续处理,也可以模拟形式直接传输至其他阵列。常用后神经元电路实现方案如下。

ADC 转换:即通过 ADC 模块将忆阻器阵列输出的模拟信号转换为数字信号,存储在输出缓存中作为下一级的输入[19-21,27,33]。与 DAC 转换模块类似,ADC 转换方案具有读取速度快,与板级验证系统实现方案一致的优点。然而由于 ADC 所需的面积与功耗均较大,因此实际应用中通常会采用多列共用一个 ADC 模块的解决方案。此时需要在每列上增加采样保持或积分模块,以便于采用时分复用的方式将各列计算结果读出,导致计算延迟增加。当交叉点阵行数较多时,为保证各列计算结果均能准确保持,还需要在积分保持电路中采用容值大和线性度好的电容,导致芯片面积进一步增加。

电流比较转换:即通过电流镜将乘累加运算所获得的列电流与多个参考电流相比较,以获得列电流分布区间,并对其进行数字编码[34-36]。这一方法由于需要多个参考电流,难以实现高分辨率的电流读出。因此,通常用于小规模卷积运算,且需要与位片式编码前神经元电路相配合使用。

直接传输:即前级阵列的模拟输出信号直接作为下一级阵列信号的输入,阵列与阵列之间无数据缓存器,这种设计极大地增加了数据的吞吐量,提高了计算的效率。然而同时由于中间结果无法缓存,整个神经网络必须全部展开,因此直接传输方案仅能应用于小规模神经网络。当网络规模较大时,由于每一神经元输出均需传输至大量下层神经元,会导致所需连线数量呈指数增长,传输协议复杂度增高。

积分点火:为减少 ADC 模块引入的面积和功耗代价,Liu 等[37]提出了积分点火实现方案,其电路实现方案如图 16(c)所示。其原理为,将列线上的电流送入一个小电容中进行积分,当电容上的电压超过参考阈值时比较器输出高电平,同时将电容放电,当电容电压降到阈值以下时比较器输出低电平。该过程周而复始,在计算时间内根据列线上的电流大小转换成相应数量的脉冲,再通过计数器获得脉冲数量。由于在电容放电期间无法进行积分,导致读取精度和速度下降,Jiang 等[38]提出一种循环感知电路,采用 2 套积分点火电路对同一列线电流进行读取,当其中一个电容在放电时,另一电容进行积分,提高了读取精度和速度。由于积分电路仅用于产生脉冲,因此所需要的电容较小。与 ADC 转换模块相比,积分点火电路结构简单,实现更为容易,可扩展性较好。

通过上述分析可以看到,与直接传输方式相比,积分点火方式便于实现数据的缓存和转发,灵活度高;与 ADC 转换方式相比,积分点火电路结构简单,所需电容较小,且便于与数字电路结合。需要说明的是,不管采用上述何种列电流读取方案,均需将列线电压进行稳定钳位以保证该列各忆阻器上的电流尽可能相互独立。同时读取电路还需考虑阵列中线电阻对计算精度的影响[39]。

2.4 多阵列互联与数据传输

现有基于忆阻器的类脑芯片中大多仅集成了单个忆阻器阵列,然而单个阵列仅能实现两层神经元之间的连接,多层神经网络须由多个忆阻器阵列协同工作才能实现。

在清华大学所设计芯片中[24]集成了 4 个忆阻器阵列(如图 8 所示),其中每 2 个阵列为 1 组实现一层神经网络,2 组阵列之间采用了直接互联的实现方案,第一个阵列的计算结果经处理后直接输入第二个阵列进行后续处理。当阵列数量继续增加时,国防科技大学 Sun 等[39]提供了一种“M – N – P”的直接互联方法,可以将多个基于单个阵列的卷积神经网络互联成高性能的卷积神经网络,以提高神经网络的性能。然而多个阵列之间采用直接互联通常可扩展性较差,使得布线难度提高。

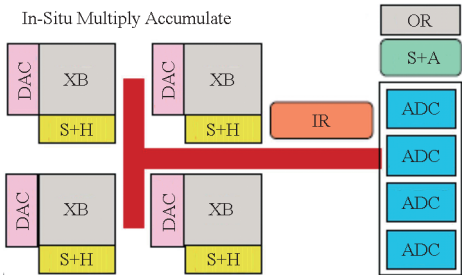
为提高可扩展性,RENO[19]采用了图 3 所示的多层星型拓扑结构,其中包含 4 组忆阻器阵列,通过 1 个模拟路由实现各组之间以及 RENO 与 CPU 之间的数据交换。每组包含 4 个阵列,通过 1 个模拟路由实现组内阵列之间以及组内阵列与顶层路由之间的数据交换。数据传输过程中,数字信号用于实现路由路径的选择与控制,各阵列计算结果仍采用模拟方式进行传输。星型拓扑能够较好地满足神经网络中组播和广播传输需求,然而由于在 RENO 中各阵列计算结果无法进行缓存,因此在实现较大规模神经网络时无法通过权重复用等策略减少对忆阻器数量的需求,应用范围仍然受到较大限制。

在 PRIME[20]中,全功能模块既可用于数据存储又可用于计算,各模块之间通过总线进行互联。各模块计算结果通过可重构读取电路转换成数字形式后在输出寄存器中缓存,然后根据需要通过总线传输至其他模块进行后续处理。由于数字结果便于缓存和转发,因此该结构相对 RENO 具有更高的灵活度。

然而当阵列数量较多时,星型拓扑互联结构会导致顶层路由数据传输压力增大,总线互联方式也会由于总线仲裁代价导致传输效率下降,因此其可扩展性仍然较差。相比之下,ISAAC[21]所采用的层次化混合拓扑结构(见图 5)具有更好的可扩展性和灵活性。ISAAC 由一定数量的 Tile 组成,采用集中式网状(concentrated-mesh, c-mesh)片上网络进行互联。在 Tile 内部采用总线互联结构便于在相邻阵列之间进行高效数据广播通信,在 Tile 之间采用片上网络互联可以实现较高的数据传输并行度和可靠性,当所需 Tile 增加时可方便地进行横向或纵向扩展。

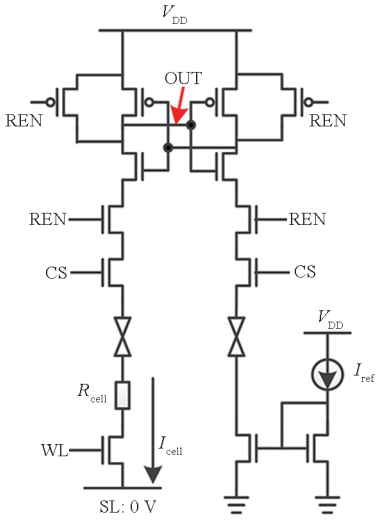
PipeLayer[40]、MAX^2[22]等工作,采用了与图 17类似架构,不同之处在于每个 Tile 内部阵列组织和使用方式以及数据流控制方式。基于以上研究成果,文献[39]指出集成多个忆阻器阵列的类脑芯片典型结构应当如图 18 所示。

当阵列互联架构确定之后,需对网络权重向忆阻阵列的映射方式以及数据加载方式进行优化,以提高权重和输入数据的复用率,进而提高计算效率。相关策略可参考数字 AI 芯片并根据忆阻器阵列特点进行优化,本文不深入展开讨论。需要说明的是,由于在当前生产工艺下,忆阻器耐久性仍然较差,因此在使用中应当避免频繁修改



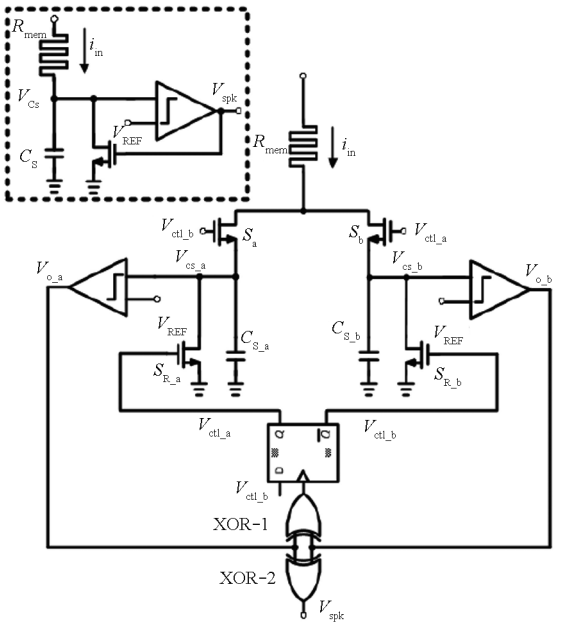
(a) 共享 ADC 转换[21]

(a) Shared ADC sampling[21]



(b) 电流比较转换

(b) AD Converting with current comparison



（c）积分点火转换[38]

（c）Integrate and fire sampling[38]

图 17　后神经元电路实现方式

Fig. 17　Post synapse neuron circuits

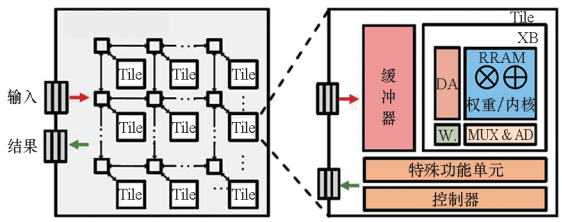


图 18　典型忆阻类脑芯片架构

Fig. 18　Typical memristor based neuromorphic chip architecture

忆阻器阻态。当基于忆阻类脑芯片进行神经网络加速时,需要将所有网络权重在忆阻器阵列中进行一一映射,甚至多次映射,以提高计算并行度。因此,忆阻类脑芯片采用何种阵列互联结构以及数据传输方式应当根据神经网络自身特点来确定。

3　系统仿真与评估

在类脑芯片设计过程中,需要进行系统仿真对芯片电路设计的正确性、计算性能、面积、功耗等进行评估。

由于忆阻器为新原理器件,在常用 EDA 设计工具中尚未得到有效的支持,因此与忆阻器相关的参数,如面积、阻值、延迟等参数需要通过对实际生产的纯忆阻器阵列芯片进行实际测量获得,进而可用于全功能类脑芯片的评估。需要说明的是,在 1T1R 阵列中,由于选通管尺寸远大于忆阻器,因此其面积主要由选通管决定。

对 CMOS 电路部分的性能仿真和评估可采用传统工具实现。例如 ISAAC[21] 使用 CACTI 6.5 的 32 nm 工艺来评估芯片中缓存器与片上互联架构的面积与功耗。移位累加(shift-and-add)电路、最大池化(max-pool)电路和 sigmoid 激活函数电路的功耗及面积评估参数参考 DaDianNao[4]。ADC 与 DAC 模块的面积和功耗参数均来自相关成熟设计。PRIME 使用 Synopsys 公司的 Design Compiler 和 PrimeTime 以及 TSMC 的 65 nm CMOS 工艺库对系统进行建模评估。RENO 则使用了 Cadence 公司的 Virtuoso 开发环境提取系统的面积和功耗参数。

在系统整体处理能力评估方面,现有设计通常选用典型神经网络,如 VGG、MSRA、MLP 等,以及 MNIST 和 ImageNet 等数据集作为仿真测试基准,并与通用 CPU 和 DaDianNao 等基于数字逻辑的计算平台性能进行对比[19-21]。

为在早期即获得所设计系统的面积、延迟、功耗等参数,亚利桑那州立大学 Yu 团队于 2018 年提出一种电路级的宏模型“NeuroSim”[41],用于对基于主流器件或新型器件的类脑计算架构进行评估。如图 19 所示,NeuroSim 在电路级和器件级均提供了灵活的接口和多种设计选择,因此,可以作为芯片设计电路级性能仿真的支持工具。基于 NeuroSim,可以实现包含了器件级、电路级以及算法级的层次化集成仿真框架,对类脑计算系统的识别精度进行指令精度的评估,并可对系统在线学习能力进行衡量。由于 NeuroSim 仅支持少量的器件模型和神经网络结构,该研究团队又在其基础上进行功能完善得到了 NeuroSim +[42] 仿真框架。然而 NeuroSim 所提供的仿真均针对较高描述层次,对电路实现细节关注较少,无法真正反映基于忆阻器实现类脑芯片时所遇到的电路设计困难。

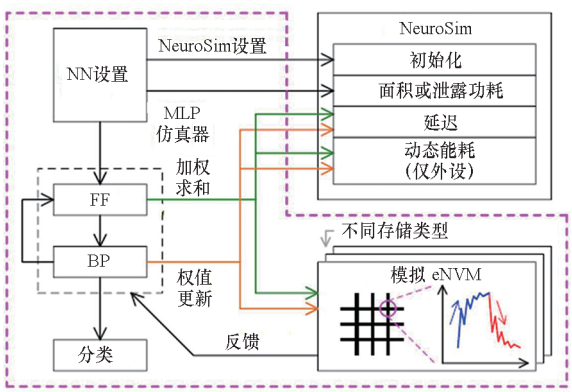


图 19　NeuroSim 仿真平台结构

Fig. 19　NeuroSim simulating platform structure

4 问题与挑战

通过以上分析可以看到,忆阻类脑芯片相比传统基于 CMOS 数字逻辑的 AI 芯片有着诸多优势,然而现有忆阻类脑芯片仍存在如下问题需要通过架构和电路设计优化加以解决:

1)量产阵列中忆阻器可用阻态较少。受限于器件制备工艺,目前连续可调的忆阻器件阻值状态还不稳定,只有二值器件基本可用,导致阵列计算能力大大下降。

2)忆阻器阵列的合理规模尚未确定。现有设计中所采用的忆阻器阵列规模各有不同。增大忆阻器阵列规模有利于提高计算并行度,但也会带来计算精度下降、忆阻器利用率下降、外围电路设计复杂度增高等问题。在现有工艺下,如何在计算并行度、计算精度以及实现代价等因素约束下,确定忆阻器阵列的合理规模是仍需解决的问题之一。

3)阵列外围电路复杂。类脑芯片与仿真架构中忆阻器阵列外围较多地采用 ADC 和 DAC 模块,导致忆阻器阻值调整、待处理数据加载以及处理结果的读出方式复杂,芯片面积和功耗增加,部分抵消了忆阻器阵列本身高集成度和低功耗的优势。

4)现有电路设计对忆阻器非理想特性的考虑不足。在芯片设计中,为保证基本功能能够实现,通常增大电路的适应范围以应对器件参数的波动。而在架构仿真过程中,仿真平台所使用的器件模型则偏于理想化,对器件的不均一性、耐久性等问题考虑不足。如何更加充分地利用忆阻器优良特性,同时针对忆阻器非理想特性在电路层面进行折中设计是提高忆阻类脑芯片性能和可靠性的关键。

5)多阵列类脑计算架构中阵列之间的数据传输、任务分配与协同等详细电路级解决方案仍不成熟。基于多忆阻器阵列的类脑计算系统研究仍处于起步阶段,对阵列之间的数据传输需求、神经网络算法与阵列互联拓扑结构的映射关系研究较少,因此现有设计仍需手工确定静态数据路由传输和任务分配方案。

6)现有类脑芯片中所集成的忆阻器阵列数量仍然较少。单个忆阻器阵列仅能实现简单神经网络,深度神经网络仍需多个忆阻阵列协同工作实现。然而受流片成本与芯片设计难度限制,集成多忆阻器阵列的类脑计算架构仍处于理论研究与高层次仿真阶段,缺乏实际验证。在完善单阵列芯片电路实现方案和计算系统性能的基础上,设计多阵列集成芯片并进行流片验证是忆阻类脑芯片走向实际应用的必经之路。

上述问题的解决将大大提升忆阻类脑芯片的处理能力和可靠性,使其能够用于神经网络加速并获得更高计算能效。然而,在电路设计之外,忆阻类脑芯片真正走向实际应用仍然面临着以下挑战:

1)改进忆阻器生产工艺,提升阵列良率、均一性和耐久性。高质量的忆阻器阵列是发挥忆阻器优良特性,实现高性能类脑芯片的基础。忆阻器生产工艺的提升需要多方合作进行,且投入成本较高,因此实现难度较大。目前已有多项研究对忆阻器应具备的特性参数进行了分析研究,发明能够满足相应质量要求的忆阻器阵列的生产工艺并实现量产是忆阻类脑芯片研发首要面临的挑战。

2)忆阻器模型完善与开发工具支持。当前阶段存在众多类型的忆阻器与相对应的物理或电气模型,然而相关模型尚未得到主流芯片开发工具的支持,导致芯片开发过程中仿真验证困难。完善忆阻器模型并加入主流芯片开发工具的模型库中,对提高芯片开发效率、降低出错概率和流片成本具有重要意义。

3)开发面向忆阻类脑计算的指令集。基于忆阻器阵列进行计算加速过程中,忆阻器阻值调整、待处理信号的加载、计算结果的读出等电路动作均需在程序控制下进行,基于通用处理器指令实现需要大量代码,导致计算延迟增加,效率降低。根据电路工作需求,开发面向忆阻类脑计算的指令集,将大大降低阵列使用难度,提高处理性能。

4)确定适合忆阻类脑计算的典型应用类型。受忆阻器阻变机理限制,忆阻器的耐久性与成熟的 CMOS 器件相比有较大的差距,且阻值调整速度较慢,因此忆阻类脑芯片在计算过程中无法对网络权重进行实时加载,而必须将所有网络权重分配到忆阻器阵列中。当权重数量较大时会导致较高的面积和功耗代价。这一因素导致忆阻类脑芯片不宜用于大规模神经网络的推理加速,而更适于网络规模较小的边缘计算应用[43]。确定忆阻类脑计算系统具有明显优势的典型应用类型,对明确系统架构、电路设计、器件参数与生产工艺等研究具有重要指导意义。在解决上述问题和挑战的基础上,根据应用需求合理设计系统体系架构,优化外围驱动电路,充分利用忆阻器优良特性

并加入对器件非理想因素的容错机制,将使得忆阻类脑芯片各项优势得到切实发挥,并最终走向实际应用。

5　结论

忆阻器的阻变特性、非易失特性,以及高集成度、低功耗、高计算并行度等特点使得其非常适用于进行类脑计算加速。当前忆阻类脑芯片的研究热点已经由基于忆阻器阵列芯片进行板级电路验证转向了集成外围驱动电路的全功能忆阻类脑芯片设计研究,芯片中集成的阵列数量也在逐步增多以实现复杂神经网络加速功能。本文对基于忆阻器的类脑芯片研究现状进行了调研和分析,对比分析了芯片中各基础功能模块的实现方案,总结了当前研究仍需解决的问题,并指出了忆阻类脑芯片走向实际应用所面临的挑战。

参考文献(References)

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521: 436 – 444.

[2] SZE V, CHEN Y H, YANG T J, et al. Efficient processing of deep neural networks: a tutorial and survey [J]. Proceedings of the IEEE, 2017, 105(12): 2295 – 2329.

[3] JOUPPI N P, YOUNG C, PATIL N, et al. In-datacenter performance analysis of a tensor processing unit [C]// Proceedings of the 44th Annual International Symposium on Computer Architecture(ISCA), 2017.

[4] LUO T, LIU S L, LI L, et al. DaDianNao: a neural network supercomputer[J]. IEEE Transactions on Computers, 2017, 66(1): 73 – 88.

[5] PEI J, DENG L, SONG S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture[J]. Nature, 2019, 572: 106 – 111.

[6] YIN S Y, OUYANG P, YANG J X, et al. An energy-efficient reconfigurable processor for binary-and ternary-weight neural networks with flexible data bit width[J]. IEEE Journal of Solid-State Circuits, 2019, 54(4): 1120 – 1136.

[7] SCHULLER I K, STEVENS R, PINO R, et al. Neuromorphic computing: from materials research to systems architecture roundtable[R]. Gaithersburg: U.S. Department of Energy, 2015.

[8] FURBER S. To build a brain[J]. IEEE Spectrum, 2012, 49(8): 44 – 49.

[9] SCHUMAN C D, POTOK T E, PATTON R M, et al. A survey of neuromorphic computing and neural networks in hardware[EB/OL]. (2017 – 05 – 19)[2021 – 01 – 05]. https://arxiv.org/abs/1705.06963.

[10] ALIBART F, ZAMANIDOOST E, STRUKOV D B. Pattern classification by memristive crossbar circuits using ex situ and in situ training[J]. Nature Communications, 2013, 4: 2072.

[11] 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: 61 – 64.

[12] MAASS W. Networks of spiking neurons: the third generation of neural network models [J]. Neural Networks, 1997, 10(9): 1659 – 1671.

[13] COVI E, BRIVIO S, SERB A, et al. Analog memristive synapse in spiking networks implementing unsupervised learning[J]. Frontiers in Neuroscience, 2016, 10: 482.

[14] 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.

[15] SHERIDAN P M, CAI F, DU C, et al. Sparse coding with memristor networks[J]. Nature Nanotechnology, 2017, 12: 784 – 789.

[16] LI C, HU M, LI Y N, et al. Analogue signal and image processing with large memristor crossbars [J]. Nature Electronics, 2018, 1: 52 – 59.

[17] WANG Z R, LI C, SONG W H, et al. Reinforcement learning with analogue memristor arrays [J]. Nature Electronics, 2019, 2(3): 115 – 124.

[18] YAO P, WU H Q, GAO B, et al. Face classification using electronic synapses [J]. Nature Communications, 2017, 8: 15199.

[19] LIU X X, MAO M J, LIU B Y, et al. RENO: a high-efficient reconfigurable neuromorphic computing accelerator design [C]//Proceedings of the 52nd Annual Design Automation Conference, 2015.

[20] CHI P, LI S C, XU C, et al. PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory [C]//Proceedings of the 43rd Annual International Symposium on Computer Architecture, 2016.

[21] SHAFIEE A, NAG A, MURALIMANOHA R, et al. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars [J]. ACM SIGARCH Computer Architecture News, 2016, 44(3): 14 – 26.

[22] MAO M Q, PENG X C, LIU R, et al. MAX^2: an ReRAM-based neural network accelerator that maximizes data reuse and area utilization [J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, 9(2): 398 – 410.

[23] HUANG H T, NI L B, WANG K W, et al. A highly parallel and energy efficient three-dimensional multilayer CMOS-RRAM accelerator for tensorized neural network [J]. IEEE Transactions on Nanotechnology, 2018, 17(4): 645 – 656.

[24] WU H Q, YAO P, GAO B, et al. Device and circuit optimization of RRAM for neuromorphic computing [C]// Proceedings of IEEE International Electron Devices Meeting, 2017.

[25] LIU Q, GAO B, YAO P, et al. A fully integrated analog ReRAM based 78.4 TOPS/W compute-in-memory chip with fully parallel MAC computing [C]//Proceedings of IEEE International Solid-State Circuits Conference, 2020.

[26] YAO P, WU H Q, GAO B, et al. Fully hardware-implemented memristor convolutional neural network [J]. Nature, 2020, 577: 641 – 646.

[27] 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: 290 – 299.

[28] 刘明. 新型阻变存储技术[M]. 北京: 科学出版社, 2014.

LIU M. Emerging resistive memory technology[M]. Beijing: Science Press, 2014. (in Chinese)

[29] ZIDAN M A, FAHMY H A H, HUSSAIN M M, et al. Memristor-based memory: the sneak paths problem and solutions[J]. Microelectronics Journal, 2013, 44(2): 176 – 183.

[30] YAO P, ZHANG W Q, ZHAO Q R, et al. Intelligent computing with RRAM [C]//Proceedings of the 11th International Memory Workshop (IMW), 2019.

[31] 张冠群. 铜互连技术中接触孔钨填充工艺研究[D]. 上海: 复旦大学, 2013.
ZHANG G Q. A study on contact tungsten filling process for copper interconnect technology [D]. Shanghai: Fudan University, 2013. (in Chinese)

[32] HU M, GRAVES C E, LI C, et al. Memristor-based analog computation and neural network classification with a dot product engine[J]. Advanced Materials, 2018, 30(9): 1705914.

[33] ZHANG W Q, PENG X C, WU H Q, et al. Design guidelines of RRAM based neural-processing-unit: a joint device-circuit-algorithm analysis [C]//Proceedings of the 56th Design Automation Conference, 2019.

[34] XUE C X, CHEN W H, LIU J S, et al. 24.1 A 1 Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge processors[C]//Proceedings of IEEE International Solid-State Circuits Conference, 2019.

[35] XUE C X, HUANG T Y, LIU J S, et al. 15.4 A 22 nm 2 Mb ReRAM compute-in-memory macro with 121 – 28 TOPS/W for multibit MAC computing for tiny AI edge devices[C]//Proceedings of IEEE International Solid-State Circuits Conference, 2020.

[36] XUE C X, HUNG J M, KAO H Y, et al. 16.1 A 22 nm 4 Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7 TOPS/W for tiny AI edge devices[C]//Proceedings of IEEE International Solid-State Circuits Conference, 2021.

[37] LIU C C, YAN B N, SONG L H, et al. A spiking neuromorphic design with resistive crossbar[C]//Proceedings of the 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), 2015.

[38] JIANG H, ZHU W J, LUO F, et al. Cyclical sensing integrate-and-fire circuit for memristor array based neuromorphic computing [C]//Proceedings of IEEE International Symposium on Circuits and Systems, 2016.

[39] SUN S Y, XU H, LI J W, et al. Cascaded architecture for memristor crossbar array based larger-scale neuromorphic computing[J]. IEEE Access, 2019, 7: 61679 – 61688.

[40] SONG L H, QIAN X H, LI H, et al. PipeLayer: a pipelined ReRAM-based accelerator for deep learning[C]//Proceedings of IEEE International Symposium on High Performance Computer Architecture(HPCA), 2017.

[41] CHEN P Y, PENG X C, YU S M. NeuroSim: a circuit-level macro model for benchmarking neuro-inspired architectures in online learning[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 37(12): 3067 – 3080.

[42] CHEN P Y, PENG X C, YU S M. NeuroSim: an integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures[C]//Proceedings of IEEE International Electron Devices Meeting (IEDM), 2017.

[43] KRESTINSKAYA O, JAMES A P, CHUA L O. Neuromemristive circuits for edge computing: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(1): 4 – 23.