

非易失性突触存储阵列及神经元电路的设计

叶 勇¹, 亢 勇², 景蔚亮², 杜 源¹, 宋志棠¹, 陈邦明^{1,2}

(1 中国科学院 上海微系统与信息技术研究所, 信息功能材料国家重点实验室, 上海 200050;

2 上海新储集成电路有限公司, 上海 200122)

摘 要: 传统的神经形态芯片一般采用 SRAM 阵列来存储突触权重, 掉电后数据会丢失, 且只能通过单一地址译码进行存取, 不利于突触权重的更新. 为此, 本文基于 40 nm 先进工艺并结合嵌入式相变存储器设计了一种非易失性突触存储阵列及神经元电路, 为神经元的突触权重存储和更新提供了一种有效、高速和低功耗的解决方案.

关键词: 神经元; 突触; 非易失性; 相变存储器; 人工神经网络

中图分类号: TN18

文献标识码: A

文章编号: 1000-7180(2017)11-0001-05

Design of Non-Volatile Synapse Array and Neuron Circuits

YE Yong¹, KANG Yong², JING Wei-liang², DU Yuan¹,

SONG Zhi-tang¹, CHEN Bomy^{1,2}

(1 State Key Laboratory of Functional Materials for Informatics, Shanghai Institute of Micro-system and Information Technology, The Chinese Academy of Sciences, Shanghai 200050, China;

2 Shanghai Xinchu IC corp. Shanghai 200122, China)

Abstract: Traditional neuromorphic chips generally use SRAM array to store the synaptic weights, which would lose data once power down. And the SRAM array can only be accessed through a single address decoding, which is not suitable for synaptic weights update. In this paper, a non-volatile synapse array and neuronal circuits based on 40 nm process combining with embedded phase change memory (ePCM) are proposed. This work would provide an effective, high-speed and low-power solution to store and update the neuron synaptic weights.

Key words: neuron; synapse; non-volatile; phase-change memory; artificial neural network

1 引言

人脑是一个高度复杂的、非线性的和并行的信息处理系统,其基本组成单元称之为神经细胞或者神经元(Neuron),每个神经元主要由细胞体、轴突(Axon)与树突(Dendrite)组成^[1]. 树突的作用是接收其他神经元的信号,轴突的作用是传递和输出信息,神经元之间通过轴突末梢的突触(Synapse)实现相互连接并相互传递信息. 如果与计算机类比的话,

每个大脑神经元相当于处理单元,每个突触就相当于一个存储单元. 正是由于上百亿个神经元细胞和千万亿个突触的互连进而组成的复杂脑神经网络系统才使得大脑具备了学习、认知及信息存储功能.

为了实现对大型神经网络系统的仿真,在硬件上主要有两种方式,一种是基于通用计算机集群的软件实现方法,另一种更有效的实现方式就是采用专用的神经形态芯片来实现^[2]. 随着集成电路工艺的迅猛发展,神经形态芯片已有众多突破性进

收稿日期: 2017-02-20; 修回日期: 2017-03-14

基金项目: 国家自然科学基金(61076121, 61176122, 61106001, 61261160500, 61376006); 中国科学院战略性先导科技专项(XDA09020402); 国家重点基础研究发展计划(2013CBA01904, 2013CBA01900, 2010CB934300, 2011CBA00607, 2011CB932804); 国家集成电路重大专项(2009ZX02023-003)

展^[3-4],并可应用于感知交互、模式识别等实时任务。我们可以看出,用硬件电路直接实现的人工神经网络系统真正能体现生物神经网络的并行信息处理能力和实时性,对实际应用也具有重要意义。

由于神经网络的大规模性,使得神经形态芯片必须采用先进工艺来实现以降低芯片面积和成本,而在先进工艺下芯片代工厂很少会提供嵌入式非易失性存储器(eNVM)解决方案。本文基于 40 nm 先进工艺并结合嵌入式相变存储器设计了一种非易失性突触存储阵列,并以此设计了一款神经元测试芯片,为神经元的突触权重存储和更新提供了一种有效、高速和低功耗的解决方案。

2 相变存储器背景介绍

相变存储器(Phase change memory, PCM)是一种以硫系化合物(典型材料 $\text{Ge}_2\text{Sb}_2\text{Te}_5$, GST)为基础的易失性存储器^[5],通过写电流产生焦耳热使相变材料发生可逆的相变过程,即呈现高阻值的非晶状态和呈现低阻值的结晶状态。本设计所实现的测试芯片是基于 40 nm CMOS 工艺制备,相变存储单元结构为 1T1R(一个选通晶体管和一個相变电阻),其电路示意图及 TEM 图如图 1 所示。该存储单元为蘑菇型结构,相变材料(GST)通过物理气相沉积(PVD)工艺形成于底电极(BEC)之上,厚度约 100 nm,上电极(TEC)和相变材料之间长一层 TiN 以作为二者之间的粘层。底电极通过钨通孔连接到选通晶体管,上电极与金属 1 层相连作为该存储单元的选通位线(BL)。当晶体管选通后,通过给 BL 施加不同程度的电流或电压脉冲就能够将 PCM 编程到特定的状态。

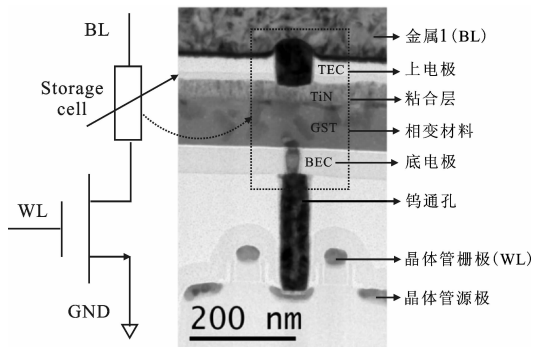


图 1 40 nm 工艺下相变存储单元结构

示意图及 TEM 图

从图 1 也可以看出,相变存储单元的制备过程

位于后道工序(Back-end of line, BEOL),与标准 CMOS 工艺完全兼容,这种良好的兼容性也使其在作为嵌入式存储器上具有非常大的优势。本文正是结合了相变存储器的特点设计了一种适用于神经网络的非易失性突触存储阵列。

3 非易失性突触存储阵列及神经元电路设计

3.1 时间依赖的学习法则

突触是神经元与神经元之间一种特化的连接,它是神经元之间相互联系和进行信息传递及处理的关键性结构。在一个学习或者记忆过程中,神经元之间的突触强度会随着突触前后神经元的活动而变化,正是这种通过突触权重的可塑性才使得神经网络产生了学习效果和自适应能力。这种突触可塑性可以通过时间依赖的学习法则来支配,例如脉冲时间依赖的突触可塑性(STDP)^[6]。对某一个神经元来说,当其膜电位超出阈值并发出尖峰脉冲时,其所有连入的树突及连出的轴突的突触权重都会被更新。假设由 N 个神经元构成的系统,那么如图 2 所示的 $N \times N$ 交叉阵列就能有效代表它们之间所存在的 N^2 种可能的突触连接方式。其中,每一行可以看作是一个神经元的轴突,而每一列可以当作一个神经元的树突,每一行和每一列的相交点即是突触权重值。对神经元 N 来说,当其需要进行电势计算时,需要对所有树突上从前一级神经元传递的信息进行加权求和,也就是说需要读取图 2 中第 N 列上的突触权重值。计算完成后需要更新所有与之相连的树突和轴突上的突触权重值,也就是说需要更新图 2 中第 N 行及第 N 列的突触权重。

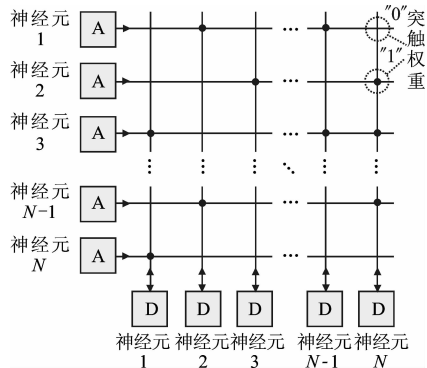


图 2 $N \times N$ 突触存储阵列

上述突触存储阵列对存储器的设计提出了新的要求。首先,为了更有效的实现时间依赖的学习法

则,突触前神经元的电势计算与突触后神经元的突触权重更新能够并行执行,这就意味存储器要同时支持行存取操作和列存取操作,而传统存储器一般仅支持行地址译码的存取,如果需要读取某一列的数据就需要连续不断的进行行存取操作,效率低且浪费功耗.其次,传统的片上存储器设计一般基于SRAM单元,不具备非易失性,即掉电后数据会丢失,因而会使得神经元之间通过长时间学习而存储的突触权重信息会丢失.对超大规模的神经网络系统来说,如果采用将其导出到外部非易失性存储器来避免数据丢失会耗费大量时间和功耗.基于以上分析,本文结合嵌入式相变存储器技术设计了一种非易失性的突触存储阵列.

3.2 非易失性突触存储阵列

本文设计的非易失性突触存储阵列是基于6T SRAM单元改进的,如图3(a)所示.每个存储单元包含10个晶体管和2个相变存储单元(10T2R结构).该存储单元可以分为两个部分,8T双端口SRAM单元和2T2R非易失性存储单元.其中,8T SRAM单元包含两种字线选择,分别命名为轴突字线(AWL)和树突字线(DWL),对应于图2中神经元的轴突行和树突列.因此,对该SRAM存储单元的读写操作既可以通过选通轴突字线来实现,也可以通过选通树突字线来实现.突触权重的非易失性存储是通过将SRAM的值编程到两个互补的相变存储单元中来实现的.基于上述存储单元的突触存储阵列设计如图3(b)所示.从图中可以看出,对该存储阵列($N \times N$)的存取既可以通过AWL地址译码而对某一行进行读写,也可以通过DWL地址译码对某一列进行读写,这两种读写方式拥有各自的译码电路和读写驱动电路,因而可以实现并行同步执行.可见,本设计解决了传统存储器在实现突触存储面临的诸多问题,因而非常适用于神经网络系统中的突触权重存储.

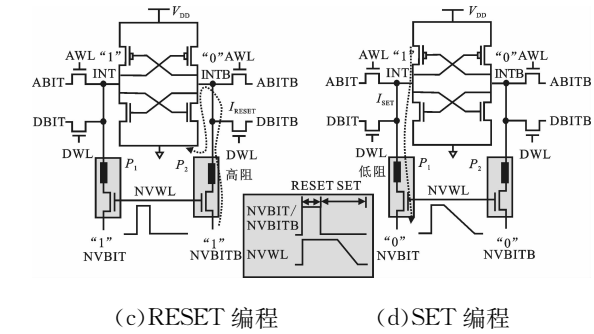
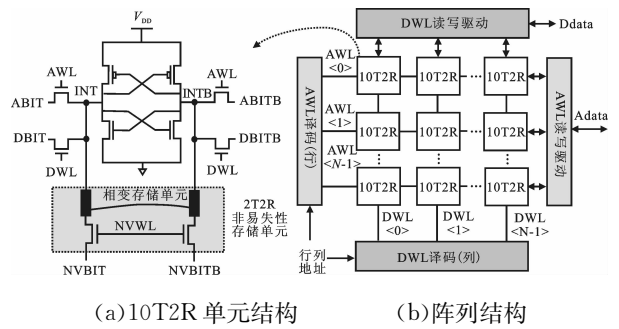


图3 非易失性突触存储阵列

对上述10T2R存储单元操作包括常规SRAM读和写操作,以及非易失性存储(store)和上电恢复(recall)操作.假设INT和INTB节点分别为高电平和低电平.对该10T2R存储单元的store操作分为两个子过程,即相变存储单元 P_2 的RESET过程和 P_1 的SET过程.在RESET过程中,NVBIT和NVBITB被拉高,然后给NVWL施加一个幅度较大、持续时间较短的脉冲,此时,电流 I_{RESET} 会流经 P_2 将其编程为高阻值的非晶态,如图3(c)所示,此时 P_1 无电流流过.在SET过程中,NVBIT和NVBITB被置于低电位,NVWL会被施加一个幅度适当、持续时间较长的脉冲,如图3(d)所示,此时,电流 I_{SET} 流经 P_1 将其编程到低阻值的多晶态.在store操作完成后,两个相变存储单元会呈现不同的阻值,且编程过程复用了SRAM单元,减小了额外的面积开销.由于相变材料的电阻率在晶态和非晶态时存在巨大的差异($R_{reset}/R_{set} \geq 100^{[7]}$),因此利用SRAM单元充当灵敏放大器可以实现相变存储单元的读取操作.当上电进行recall操作时,INTB节点的RC延迟要远远大于INT节点的RC延迟,在SRAM单元的自锁和反馈作用下,INT节点会被逐渐拉高,INTB节点会逐渐拉低,直至达到一个稳定的状态,此时即完成了上电恢复的功能.因此,本文设计的10T2R存储单元不仅具备SRAM的高速存取性能,也能够获得掉电后的数据保持能力,而无需额外的面积开销.

3.3 神经元电路设计

对于每一个神经元来说,它都应该能够执行整合放电(Integrate-and-Fire, IF)功能^[8],即当神经元对所有输入加权求和后得到该神经元的膜电位,如果该膜电位超过预先给定的阈值 V_{th} 时,该神经元就会发出一个spike,并且自身膜电位立刻恢复到一个固定值 V_{reset} .图4所示CMOS电路简单实现了上述

神经元的 IF 功能. 输入信号 AData 为前一级神经元经轴突传递来的信号, 当其激活时(为高电平)时, 由一个较慢的时钟信号 CLK 来记录其脉冲的保持时间, 即由一个 4 位计数器来表征. 由于输入信号和时钟信号是经过“与”的关系, 因此采用频率较低的时钟 CLK 可以过滤掉输入信号端的毛刺. CNT2VDPM 模块完成对脉冲保持时间(计数值)到模拟电压的转换, 即不同的计数值会对应得到不同的模拟电势 VDPM, 该电势就相当于神经元的膜电位. 当模拟电势 VDPM 超过产生放电(fire)的阈值电压 V_{ref_Fire} , 神经元的输出会产生一个尖峰脉冲 NU_FIRE, 并且电势 VDPM 会立即回到一个初始值. 比较放大器的输出端 NU_FIRE 就是该神经元的输出. 为使本设计的神经元测试芯片具备扩展性, 每个神经元还应当可以接收来自外部的激励信号, 如图 4 中的外部输入 EXR_INPUT, 它通过一个二选一多路选择器进入 IF 模块并由 EXT_IINPUT_EN 信号控制.

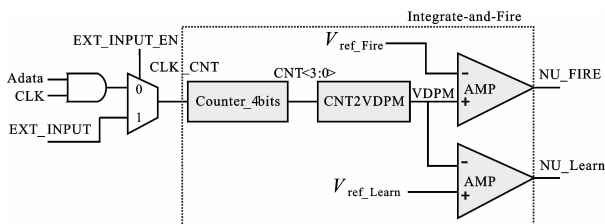


图 4 神经元电路结构示意图

神经元的另一个主要功能就是通过学习更新突触权重, 本设计采用一个模拟参考电压 V_{ref_Learn} 来实现, 通过比较该参考电压与模拟电势 VDPM 进行比较进而产生比较结果 NU_Learn, 外部控制器就会根据该结果对相应的突触权重进行更新. 每一个神经元的学习参考电压都可能是不一样的, 这同样可由外部控制器来对其进行控制.

4 电路仿真

本测试芯片采用 SMIC 先进的 40 nm CMOS 逻辑工艺实现, 结合嵌入式相变存储器技术, 实现了 8×8 非易失性存突触存储阵列和 8 个独立的神经元电路. 测试芯片的版图如图 5 所示.

图 6(a)仿真结果显示了对非易失性突触阵列中相变存储单元的编程过程. NVEN 为编程的使能信号, 脉冲时间为整个 RESET 和 SET 时间之和.

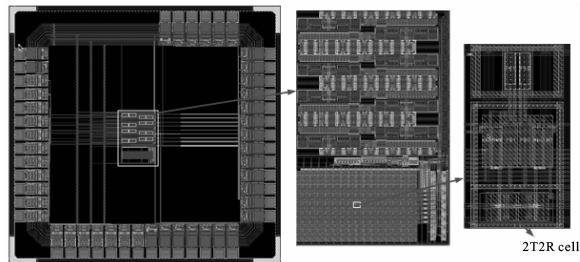


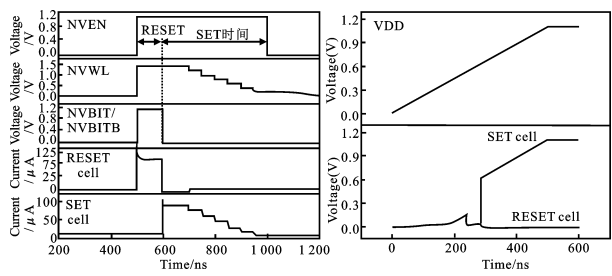
图 5 测试芯片版图

NVBIT 和 NVBITB 会被同时施加一个脉冲信号, 脉冲时间长度为 RESET 过程所需时间. 在 RESET 过程中, NVWL 会保持在一个较高的电压水平以保证足够的 RESET 电流, 而在 SET 过程中, NVWL 电压会逐步降低, 使得相变材料能充分结晶. 为提高编程的可靠性, 本设计中 RESET 时间和 SET 时间长度都是可调的, 且 NVWL 上所施加的电压也是可调的. 编程过程中两个相变存储单元的编程电流如图中所示. 掉电后, 从相变存储单元中的数据恢复过程的仿真结果如图 6(b)所示. 假设重新上电后电源电压在 500 ns 内从 0 V 增加到 1.1 V, INT 节点和 INTB 节点分别与高阻态(RESET)和低阻态(SET)的相变单元连接. VDD 上升时会通过存储单元中的两个 PMOS 晶体管对 BIT 和 BITB 节点进行充电. 显然与高阻连接的节点充电要缓慢, 因此会慢慢建立其 $V_{INTB} > V_{INT}$, 在 SRAM 单元的自锁和反馈作用下, INT 会升高到高电平, 而 INTB 会保持在低电平, 这也和我们预期所保存的值一致.

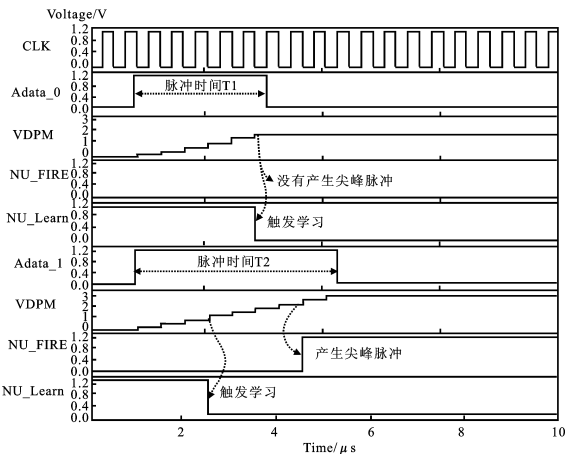
图 6(c)所示的是神经元电路的仿真结果, 图中显示了两个不同脉冲的轴突数据脉冲 Adata_0 和 Adata_1, 该数据脉冲均来自于突触存储阵列, 不同神经元对不同输入的反应和学习结果是不同的. 对 Adata_0 来说, 其保持时间 T_1 不足以产生 spike, 却能触发学习, Adata_1 的保持时间 T_2 足够长因此产生了 spike, 并且也触发了学习. 外部控制器就可以根据不同的神经元对不同轴突输入的结果输出进而更新突触权重并达到学习效果.

5 结束语

本文介绍了一种非易失性突触存储阵列及神经元电路的设计方法. 与传统突触存储阵列相比, 本文设计的突触存储阵列具备非易失性, 且可以同时通过轴突和树突字线选中的方式对突触权重进行更新. 在



(a)非易失性突触阵列的 store 过程 (b)recall 过程



(c)神经元电路的仿真结果

图 6 非易失性突触阵列及神经元电路的仿真结果

完成该芯片的流片和测试工作之后,我们将更多神经元和突触存储阵列以全局的、分布式的片上或片外的连接方式来构建更复杂的网络来模拟大型神经网络,最终目标是实现认知、学习和识别功能。

参考文献:

[1] Haykin S. Neural networks: a comprehensive foundation[M]. 3rd ed. US, Englewood Cliffs, NJ,Prentice-Hall, Inc. , 2007.

[2] Mead C. Analog VLSI and neural systems[M]. Boston, MA, US, Addison-Wesley, 1989:257-278.

[3] Merolla P A, Arthur J V, Alvarez-icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface [J]. Science, 2014, 345(6197): 668-673.

[4] Chu M, Kim B, Park S, et al, Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron[J]. Transactions on Industrial Electronics, 2015, 62(4):2410-2419.

[5] Song Z, Zhan Y P, Cai D, et al. A phase change memory chip based on tisbte alloy in 40 nm standard CMOS technology[J]. Nano-Micro Lett, 2015, 7(2): 172-176.

[6] Dan Y, Poo M M. Spike timing-dependent plasticity: from synapse to perception[J]. Physiol. Rev. , 2006, 86(3):1033-1048.

[7] Wang Y, Cai D, Chen Y, et al. Optimizing set performance for phase change memory with dual pulses set method[J]. ECS Solid State Letters, 2015, 4(7): 32-35.

[8] Indiveri G, Etienne-cummings R, Schemmel J, et al. Neuromorphic silicon neuron circuits[J]. Frontiers in Neuroscience, 2011, 5(5): 1-23.

作者简介:

叶 勇 男,(1991-),硕士研究生. 研究方向为嵌入式存储器研究. E-mail:yeyong1991@mail. sim. ac. cn.

亢 勇 男,(1978-),博士. 研究方向为相变存储器电路设计与测试.

景蔚亮 男,(1980-),硕士. 研究方向为电路设计与测试.

杜 源 男,(1990-),硕士研究生. 研究方向为相变存储器电路设计.

宋志棠 男,(1964-),研究员. 研究方向为纳电子材料与器件研究.

陈邦明 男,(1961-),研究员. 研究方向为存储器开发.