中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

基于忆阻器的神经形态计算及系统应用研究

作者姓名：张续猛

指导教师：刘明　研究员　中国科学院微电子研究所

刘琦　研究员　中国科学院微电子研究所

学位类别：工学博士

学科专业：微电子学与固体电子学

培养单位：中国科学院微电子研究所

2020 年 5 月

Research on the neuromorphic computing and system applications with memristors

A dissertation submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Microelectronics and Solid-State Electronics

By

Zhang Xumeng

Supervisor: Professor Liu Ming

Professor Liu Qi

Institute of Microelectronics, Chinese Academy of Sciences

May 2020

中国科学院大学

研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：张续航

日　　期：2020.5.24

中国科学院大学

学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：张续航　　导师签名：刘明

日　　期：2020.5.24　　日　　期：2020.5.25.

摘 要

受人脑的启发，构建智能高效的神经形态机器一直以来是科学家们孜孜不倦的追求。尤其在当今大数据时代，随着数据量的急剧增加，基于存算分离体系架构的传统冯·诺依曼计算机在实时高效的数据处理问题上越来显得捉襟见肘。因此，借鉴人脑结构构建神经形态机器的研究被推向了历史的浪潮顶尖。在人脑中，神经元和突触是基本的计算单元，那么设计实现神经元和突触电路便是构建神经形态机器的关键。然而，由于传统 CMOS 器件缺乏内在的动态特性而导致神经元和突触电路构成复杂，很难集成到人脑的规模，且 CMOS 器件即将达到其物理瓶颈。因此，开发具有内在动态特性和可微缩的新原理器件来构建高效的神经形态机器受到产业界和学术界的广泛关注。忆阻器，具有结构简单、内在动态特性丰富、功耗低、可微缩性好、易于三维集成、与现有 CMOS 工艺兼容等优点，被认为是实现低功耗、高密度神经形态机器的理想硬件单元。本文围绕如何利用忆阻器电致阻变行为的动态特性实现脉冲神经元和人工突触进而构建高效的神经形态机器，开展了器件优化，电路设计，行为模拟，系统验证等方面的工作。取得了如下创新性的成果：

(1) 多值阻变忆阻器件及神经突触功能仿生

a. 在 Cu/a-Si/Pt 结构的忆阻器中，我们发现通过控制注入 a-Si 层中的 Cu 离子量，器件表现出易失性和非易失性两种转变行为。利用该独特特性，生动地模拟了生物突触的短时程和长时程可塑性。此外，在重复刺激下，还观察到该器件可以实现从短时程记忆到长时程记忆的转变，对应于生物体的训练过程。

b. 为实现突触器件的多值可控及线性电导调制，我们进一步通过双层堆叠工艺设计了一种缓变特性良好的 $Pd/HfO_2/WO_x/W$ 突触器件。在脉冲编程下，可实现极宽的电导变化范围（>300）和高耐久周期（$>10^6$）。然后利用变脉幅和变脉宽两种编程方案分别得到了近似线性的电导变化行为，在 MNIST 手写数字数据集上的系统仿真识别率达到 95.3%，接近理想情况。

(2) 基于易失性忆阻器的神经元电路及系统验证

a. 首次将忆阻器中离子导电桥的自发断裂过程引入到神经元电路的实现中，将 $Ag/SiO_2/Au$ 器件作为漏电积分发射神经元的阈值开关。实现了神经元动作电位的四个关键特征：动作电位的全或无、阈值发射、不应期和强度调制的频率响应。系统仿真证明所设计的神经元可作为用于模式识别突触后神经元。

b. 为提高神经元的可集成度，解决单忆阻器神经元在脉冲串刺激下不能正常的连续放电以及缺乏对忆阻器突触在线训练的能力等问题。我们通过将 Ag/SiO_2:Ag/Au 忆阻器与简单的数字电路相结合构建了混合神经元电路，并设计了专门的侧向抑制忆阻器阵列。进一步构建了一个由混合神经元和忆阻器突触组成的全硬件多层脉冲神经网络并实现了在线训练。

c. 鉴于脉冲神经网络（SNN）的训练算法不成熟，基于转换方法的 SNN 具有更高的识别精度。为实现转换 SNN，利用具有金属-绝缘态转变（MIT）性质的 NbO_x 器件构建了一种 1T1R 结构的神经元电路，输出脉冲发放频率与输入电压关系可以匹配人工神经网络（ANN）中的 ReLU 激活函数。并进一步利用该神经元电路和忆阻器突触阵列构建了 320×10 的网络，实现了 ANN 到 SNN 的转换。在 MNIST 手写体数据库上得到了与 ANN 相当的识别率。

(3) 基于 NbO_x 器件的脉冲传入神经电路及脉冲机械感受系统

首次提出了一种基于 NbO_x 忆阻器的人工脉冲传入神经电路作为传感器与脉冲神经网络的紧凑接口。通过 TEM 测试验证了 NbO_x 器件的转变机制，并系统研究了传入神经放电频率与外界刺激强度的关系。利用该传入神经，我们进一步实现了一个以无源压电元件为触觉传感器的零静态功耗脉冲机械感受系统。实验证明，该传入神经电路可以用来构建具有自我感知意识的高级神经形态机器。

关键词：忆阻器，人工突触，脉冲神经元，脉冲神经网络，人工传入神经

Abstract

Inspired by the human brain, constructing a high-efficient neuromorphic machine has always been the pursuit of scientists. Especially, with the data boosting in the big data era, the conventional von Neumann computer with separated computation and memory units is limited to efficiently process data in real-time. Thus, emulating the brain architecture to build a neuromorphic machine is promising to alleviate this challenge. We know the neurons and the synapses are the basic units in the human brain, thus, implementing neuron circuits and synaptic circuits is the basic task for constructing neuromorphic machines. However, due to the conventional CMOS devices lack of inherent dynamics, the constructed neurons and synaptic circuits are complex, which makes it difficult to reach the size of the human brain. Moreover, CMOS devices are about to reach its physical bottleneck. Therefore, developing new principle devices with inherent dynamics and high scalability to construct efficient neuromorphic machines has attracted great attention of industry and academia. Memristor, is considered to be an ideal hardware unit to realize high-efficient and high-dense neuromorphic machines, because of its simple structure, rich intrinsic dynamics, low power consumption, good scalability, friendly three-dimensional integration, and compatibility with CMOS technology et al. In this thesis, we focus on using the resistive switching dynamics of memristors to implement spiking neurons and plastic synapses for building an efficient neuromorphic machine. In details, we carried out the study about the device optimization, circuit design, behavior emulation, and system demonstration. The following achievements have been implemented:

(1) Multi-level switching memristors and synaptic functions demonstration

a. In the Cu/a-Si/Pt memristor, we found that the device exhibits both the volatile and non-volatile switching behavior by controlling the amount of Cu ions injected into the a-Si layer. Using this unique feature, we vividly emulated the short-term and long-term plasticity features of biological synapses. In addition, under repeated stimuli, the device

could switch from short-term memory to long-term memory mechanisms, which corresponds to the training process of organisms.

b. To implement the controllable multi-levels and linear conductance modulation of synaptic devices, we further designed a $Pd/HfO_2/WO_x/W$ synaptic device with excellent analog switching behavior by stack technology. Under the pulse programming, a wide conductance range (> 300) and a high endurance period (> 10^6) are achieved. Using both the amplitude-modulation and the width-modulation programming schemes, we achieved a sub-linear conductance updating curve of the device. Based on the updating curve, 95.3% accuracy is achieved on MNIST handwritten datasets through system simulation.

(2) Neuron circuits and system demonstration based on volatile memristors

a. For the first time, the spontaneous breaking process of the ion filament in the memristor is introduced into the neuron circuit realization. A $Ag/SiO_2/Au$ memristor is used as the threshold switch of the leaky integration-and-fire neuron. This neuron displays four critical features of biological neurons: the all-or-nothing spiking of an action potential, threshold-driven spiking, a refractory period, and a strength-modulated frequency response. As post-neurons, the designed neuron is demonstrated to be applicable to digit recognition.

b. To improve the integration and solve the problem that the single-memristor neurons cannot continuously fire decently under pulse train stimuli as well as lacking the ability for on-line training of memristor synapse. By combining Ag/SiO_2:Ag/Au memristor with simple digital circuits, we build a hybrid spiking neuron and specially design a lateral inhibition memristor array. Then, we further demonstrate an entire hardware multi-layer spiking neural network consists of the hybrid neurons and memristive synapses for the first time, and implement the online training.

c. Given the training algorithms of spiking neural network (SNN) is immature, the conversion based SNN features a higher recognition accuracy. For implementing the conversion based SNN and matching the rectified linear unit (ReLU) activation function in analog neural network (ANN), a Mott neuron with a 1T1R structure is proposed. Using

the 1T1R Mott neurons and memristive synapses, we experimentally demonstrated a one-layer (320×10) conversion-based SNN for the first time. The recognition accuracy on the MNIST database is equal to that of ANN.

(3) A spiking afferent nerve and spiking mechanoreceptor system based on NbO_x device

For the first time, an artificial spiking afferent nerve based on highly reliable NbO_x Mott memristors is proposed as a compact interface between sensor and spiking neural networks. We demonstrated the switching mechanisms of the NbO_x by TEM and systematically studied the relationship between the afferent nerve spiking rate and the external stimulation intensity. Using this afferent nerve, we further build a power-free spiking mechanoreceptor system with a passive piezoelectric device as the tactile sensor. The results show that the proposed afferent nerve is promising for constructing a high-level neuromorphic machine with self-awareness.

Key Words: memristor, artificial synapse, spiking neuron, spiking neural network, artificial afferent nerve

目 录

图目录

表目录

第1章 绪论

1.1 前言

纵观人工智能发展历史，创造像人类大脑一样工作的机器一直是科学技术发展不竭的动力和创新源泉[1]。近年来，随着脑科学、大数据和深度学习算法的不断发展和成功应用，当前的计算机在图像分类、语音识别、自然语言处理、任务决策、智能驾驶等领域显示出优越的能力[2,3]，在某些情况下甚至超过了人类专家[4,5]。例如，2016年计算机程序 AlphaGo 首次在围棋战略游戏中击败了人类顶级选手[4]，这一事件引起了国际上的广泛关注，开启了人工智能的新纪元。随后，2019 年，AlphaStar II 又在实时战略游戏中打败了职业星际玩家，代表了当前人工智能的最高水平[5]。背后的推手之一是信息处理系统性能的不断提升，在过去的 40 多年里，随着集成电路工艺和设计技术的进步，信息处理系统的计算性能提升了近千亿倍[6]。

尽管计算性能的提升所带来的好处令人印象深刻，但从能效上来讲，现有计算系统与人脑相比仍有着近 5 个数量级的差距[7]，如在 AlghaGo 与人类选手的对战中，据估 AlphaGo 的功耗达 2 兆瓦，而人脑的功耗仅为 20 瓦。导致这种差距的重要原因之一是现有冯·诺依曼架构采用计算与存储分离的形式（图 1.1（a）），计算单元与存储单元之间的数据通信消耗了大量的能量和时间，造成了“冯 •诺依曼瓶颈”问题。此外，存储与计算单元之间的性能不匹配造成了数据读取和存储过程中相当大的延时（称为“存储墙”），这两个问题的存在大大地降低了计算机信息处理的效率[8]。为了提高数据处理的效率，人们提出了一些改进的技术。例如，利用具有多核和高通量互连的图形处理器（GPU）提高计算并行性[9]。然而，当 GPU 用于神经网络计算时，突触的权值仍然存储在分立的单元中，例如静态随机存储器（SRAM），工作时需要对这些单元频繁地访问以获取数据，并不断地充电以存储信息。张量处理器（TPU）是另一种专用集成电路，通过在大容量下使用低精度计算提高计算效率，但延迟问题仍然存在，还是无法达到人脑的效率[10]。因此，借鉴人脑的结构和信息处理机制构建更高效的智能计算系统是必然趋势。

虽然大脑的结构和功能还没有被充分探索清楚，但根据当前神经科学的已有研究

结果，我们大致可以将大脑的高效率归因于三个方面：感存算一体的计算模式，高度互联的功能性拓扑结构，以及时间依赖的神经元和突触功能[1]。在人类大脑中有数十亿个神经元，每个神经元通过成千上万个突触与其它神经元相连，构成复杂的神经网络[11]。神经元产生离散的动作电位（脉冲信号）并直接通过连接的突触交换或传递信息，是基本的计算单元；在外界信息刺激下，连接突触的权重可以动态调整，从而完成学习和记忆的功能[12]。基于动作电位的时序信息处理过程是大脑实现稀疏高效的信息传输的重要原因[13]。神经形态计算借鉴了人脑的结构和机理，包括神经元以及它们之间的突触连接，通过学习确定突触权重，通过推理完成认知计算，具有并行计算、模拟计算和存内计算的特点[14]。图 1.1（b）给出了一种神经形态计算系统的架构图，通常采用多核设计的形式，神经元和突触直接通信，旨在实现低功耗、低延时、具有时空信息处理能力的高效智能机器[15]。

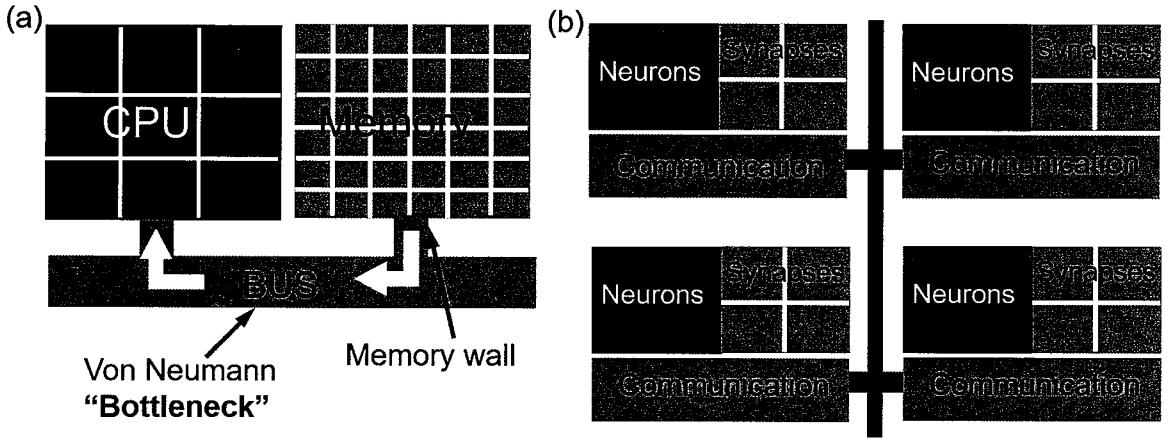


图 1.1 （a）冯·诺依曼架构和（b）神经形态架构对比[16]

1.2 生物神经网络和网络模型

“神经形态”的概念最初是在 20 世纪 90 年代由加州理工学院教授 Carver Mead 提出的[17]，其具体指利用模拟电路模拟生物神经元和突触的相关功能，从而构建类似于大脑的信息处理系统。近些年，神经形态计算技术也指采用模拟、数字、数模混合电路以及软件算法对神经网络的实现。前面提到，生物神经网络的基本构成单元是神经元和突触[11]，图 1.2（a）给出了一个典型的神经元结构，通常包括多个树突，这些树突和胞体一起接收和整合输入信号[18]；神经元产生的信号会通过轴突到达轴突末

梢然后传递给其它多个神经元。轴突末梢与其它神经元连接的位置在生物学上又叫做突触，突触的强度决定了前神经元可以向后神经元传递信号的大小[19]，人工神经网络便是由此得来。图 1.2（b）给出了一个典型的网络原理图，包括输入层（input layer）、隐藏层（hidden layer）和输出层（output layer）三层。输入层接收数据输入并与隐藏层连接的突触进行加权求和后作为隐藏层的输入（$\sum W_H * X$），隐藏层的神经元对输入的数据进行非线性变换后输出（$y = f(\sum W_H * X)$）。其中，隐藏层可以扩展至多层。隐藏层的输出再与输出层连接的突触进行加权求和作为输出层的输入（$\sum W_O * Y$），输出层的神经元经过再次非线性变换后（$y = f(\sum W_O * X)$）得到最终的输出结果。因此，神经网络中实现计算的核心是向量与矩阵的点积运算[2]。

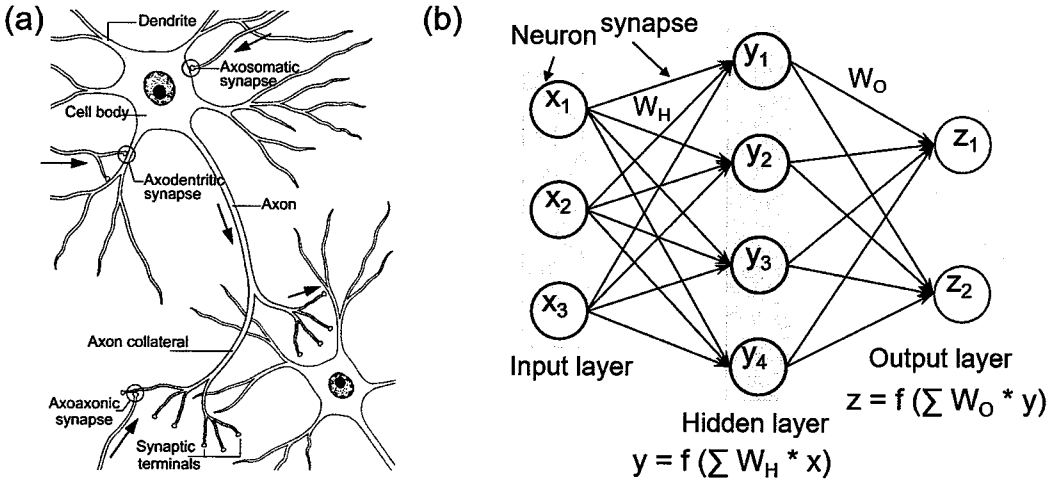


图 1.2 （a）生物神经网络原理图[11]；（b）典型的三层神经网络原理图，包括输入层、隐藏层和输出层

广义上来讲，根据所用神经元模型的不同，神经网络的算法模型大致可以分为人工神经网络（artificial neural network，ANN）和脉冲神经网络（spiking neural network，SNN）[20]。其中，ANN 是目前深度学习使用的主要模型[2, 8]，该模型主要模拟了生物神经网络中的拓扑互联结构，利用非线性函数代表神经元，在一定程度上实现生物神经网络的认知功能。相比于 ANN，SNN 采用脉冲神经元作为网络计算的基本单元[21]，并在计算过程中引入了时间参量，具有事件驱动和稀疏编码的特点，更忠实地模拟了大脑的功能，具有更高的计算能效。下面将对以上两种算法模型进行论述：

1.3 神经网络的算法模型

1.3.1 人工神经网络

在 ANN 中，神经元模型包含输入，对输入的求和单元以及非线性变换函数，如图 1.3（a）所示。在这里，非线性变换函数的形式包括 sigmoid 函数、tanh 函数以及 ReLU 函数等（图 1.3（b））。这些非线性函数可以对输入的模拟值进行变换给出相应的模拟输出值，与生物神经元接收输入信号产生动作电位信号的形式不同。正因如此，采用这些非线性函数作为神经元进行非线性变换的网络被称为 ANN。为了对网络进行训练使得网络具有认知功能，当前普遍使用且有效的算法是由 Rumelmat 等人于 1985 年提出的反向传播算法（BP 算法）[22]。利用 BP 算法对网络的训练过程主要有两个过程：前向推理过程和反向传播过程。图 1.3（c）给出了训练过程原理图，在前向推理过程中，输入信息通过输入层经隐藏层，逐层计算得到输出层的结果。如果输出层结果与期望值相差较大，则把输出与期望之差的平方和作为目标函数进行反向传播，逐层求出目标函数对相关权值的偏导数，最终得到每一层对应权重的改变量，使得网络的输出值逐渐逼近期望值。训练完成后，网络便具有了一定的事物认知能力。

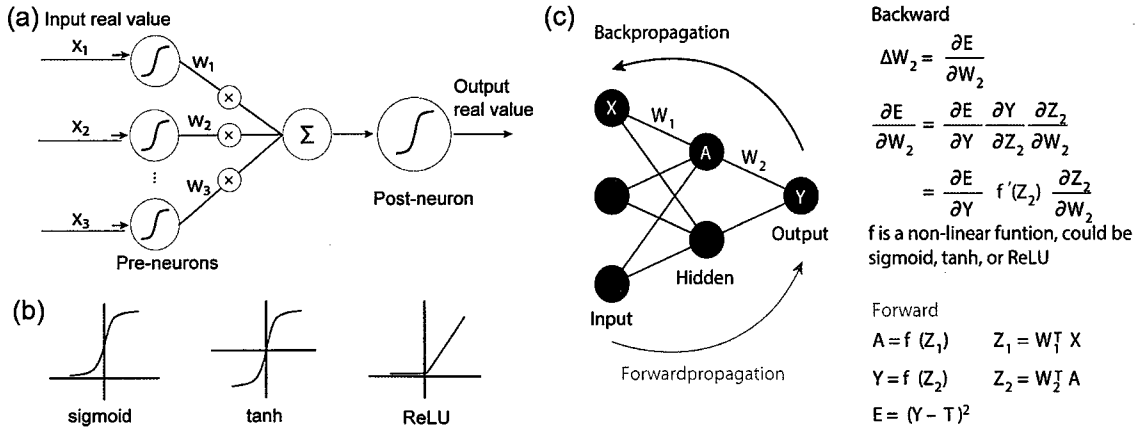


图 1.3 （a）ANN 工作原理图；（b）常用非线性激活函数；（c）基于 BP 算法的训练过程[1]

1.3.2 脉冲神经网络

相应的，SNN 采用了脉冲神经元作为基本的计算单元，是一种更仿生的神经形态计算模型，它是在 20 世纪 90 年代由奥地利格拉茨大学教授 Wolfgang Maass 提出的[21]。SNN 与 ANN 之间最重要的区别在于信息处理的方式。如前所述，ANN 使用模拟的真实值进行计算（例如，信号的大小），而 SNN 则是使用脉冲信号来处理信息。

脉冲信号的发生过程本质上是二进制事件，可以是 0，也可以是 1。如图 1.4（a）所示，SNN 中的神经元只有在接收或发射脉冲信号时才会被激活，与生物神经元一致，因此它是事件驱动的，这使得它在推理时间段内相比于 ANN 具有更高的能量效率。没有输出脉冲信号的神经元基本上处于静默状态，不会对其它神经元有任何的输出，这与 ANN 不同。在 ANN 中，不管神经元是不是有输入或者输出，其神经元始终处于工作状态。此外，SNN 中神经元以二值动作电位信号传递信息，使得网络中庞大的数学点积运算$\Sigma_i V_i \times W_i$的计算密集度更低，从而降低计算量，能耗更低。对于脉冲神经元的实现，到目前已经有很多神经元模型被提出用来模拟动作电位的产生过程，如 leaky integrate-and-fire（LIF）神经元模型[23]（图 1.4（b））、Hodgkin - Huxley 模型[24]、Izhikevich 模型[25]等。在实际应用过程中，考虑到计算复杂度，LIF 和 Izhikevich 为常用的神经元模型[26]。相比于 ANN，脉冲时间依赖可塑性（spike-timing-dependent plasticity，STDP）是 SNN 中主要的学习算法之一（图 1.4（c））[12,27]，该算法受启发于生物神经网络，已经在生物网络中得到验证。在该算法中，突触权重的改变量依赖于突触前后神经元脉冲发放的相对时间，是一种局部训练的非监督非反向传播算法。

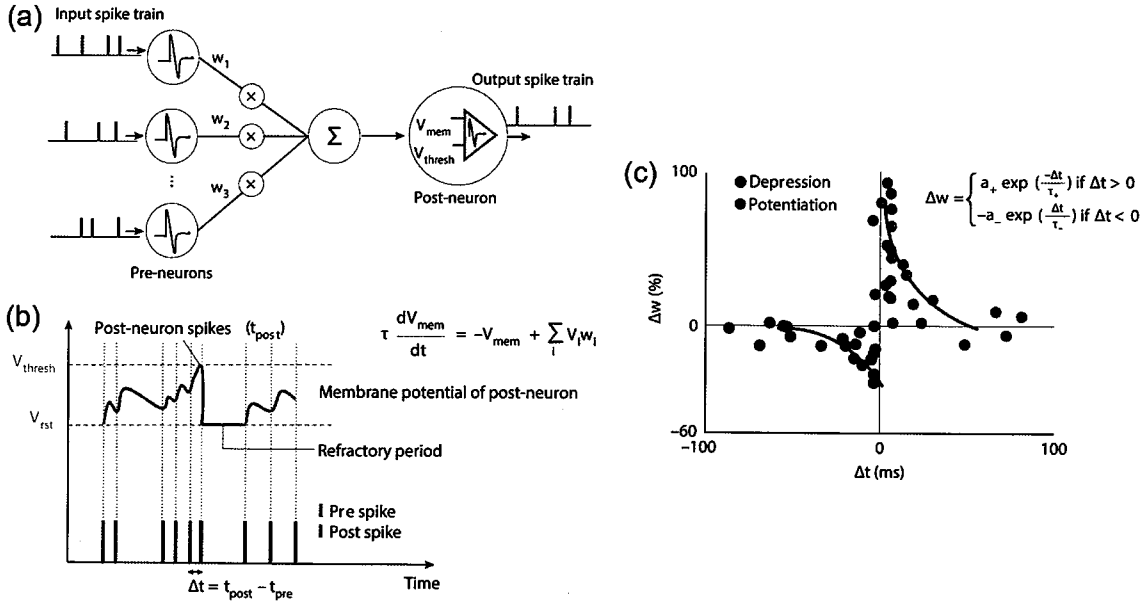


图 1.4 （a）SNN 工作原理图；（b）LIF 神经元模型；（c）STDP 突触权值更新曲线[1]

1.3.3 ANN 和 SNN 的对比

为清楚的展示两种网络模型的区别，表 1.1 给出了两种模型之间的对比。相对于

ANN 来说，虽然 SNN 的训练算法还不成熟，无法达到 ANN 所能实现的精度，但由于 SNN 在计算过程中引入了时间参量而且使用与生物神经元工作原理相似的脉冲神经元完成计算过程，因此具有更高的生物合理性[28, 29]。另外，由于 SNN 神经元的二进制脉冲工作模式，具有事件驱动的特点[13]，因此具有更高的能效而且更适合时空信息问题的处理。国际上认为 SNN 是一种更高阶的神经形态计算模型，是实现高效神经形态计算技术的算法之一，也是当前学术界和产业界研究的重点。

表 1.1 ANN 和 SNN 的对比

模型类别	人工神经网络（ANN）	脉冲神经网络（SNN）
生物合理性	生物合理性不足	具有生物合理性
编码方式	模拟的实值（定点或浮点数）	脉冲串（含事件信息的二值信号）
激活函数	多使用非线性激活函数	脉冲神经元
时空信息处理	不适合时空信息处理	本质上适合时空信息处理
计算速度	较慢，串行同步计算	较快，并行异步计算
计算能力	相比较于 ANN，SNN 在计算过程中引入了时间参量，理论上具有更强大的计算能力	
训练算法	较成熟，主要基于数学计算的反向传播算法	相对不成熟，有仿生的 STDP，SRDP 和基于数学计算的时间反向传播等
功耗	所有神经元激活，功耗较高	事件驱动，功耗较低
应用	ANN 比较成熟，在一些实际任务，如图像、视频、语音等都有应用；SNN 更适合于动态信号的实时处理，像语音、视频、机器人控制等	

1.4 脉冲神经网络的硬件实现

目前，神经形态计算的硬件实现主要分为基于传统 CMOS 技术的神经网络硬件实现和基于新型纳米器件的神经网络硬件实现，其重点是构建基本的神经元电路和神经突触电路。传统 CMOS 技术发展相对比较成熟，当前已经取得了一定的成果。基于新型纳米器件的神经形态计算技术正处于起步阶段，目前最受关注的方向是利用忆阻器构建神经形态芯片[20]。狭义上，神经形态芯片也特指对于脉冲神经网络的硬件实现，下面将主要阐述基于这两种技术的脉冲神经网络的硬件实现。

1.4.1 CMOS 基脉冲神经网络

在传统 CMOS 工艺下，首先需要构建基本的神经元和突触电路，然后借鉴大脑的互联结构，实现任意神经元之间通过突触的连接。在特定规模的神经网络下，任意一个神经元都可以把信息传递给指定的另一个或多个神经元，并且通过连接的突触进行原位计算而不是从内存中读取后再计算。为了实现复杂的互联，神经形态芯片通常通过横纵交叉矩阵（X-bar）、片上网络（Network on Chip, NoC）、多核互联的多层级方案实现，如图 1.5 所示[30-33]。在矩阵交叉点处有静态随机存储器（SRAM）或其它类型存储器存储权值表示突触连接，连接的强度表示权重大小，突触单元负责连接前后神经元电路。

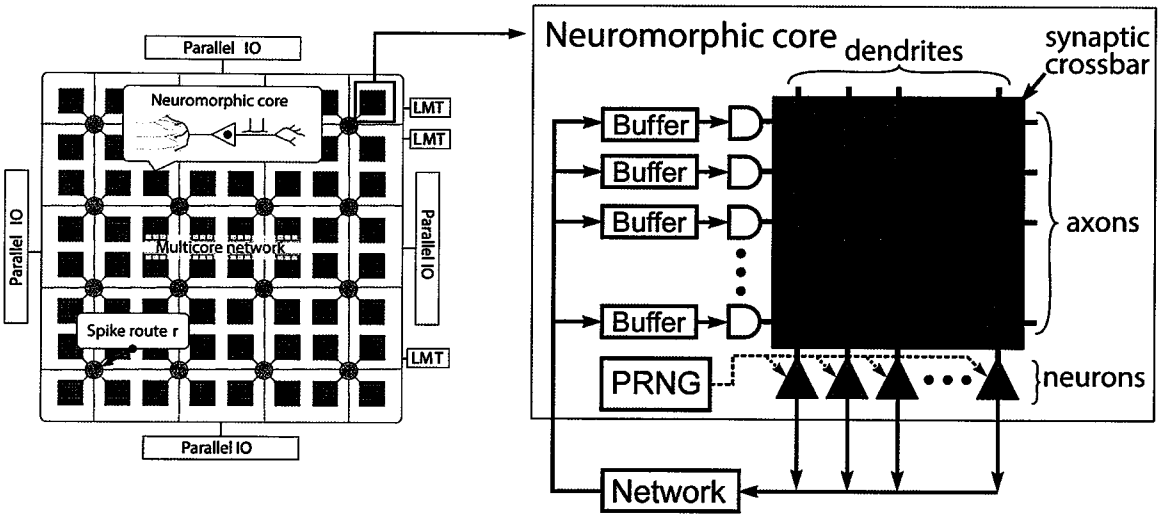


图 1.5 CMOS 基神经形态芯片原理图[31, 33]

基于此架构形式，已经有成功的芯片产品陆续面世并在智能城市、自动驾驶的实时信息处理、人脸深度识别等领域展现出应用潜力。例如 IBM 的 TrueNorth[34]、Intel 的 Loihi[30]、斯坦福大学的 Neurogrid[35]、以及海森堡大学的 BrainScaleS[36]等。其中，IBM 的 TrueNorth 是全局异步局部同步混合数字电路的代表作，该芯片不具有在线学习能力，主要对预训练网络进行推理，可以用于检测图像中的行人、车辆等物体，具有极低的功耗（65 mW）。它也可用于语音、图像数据集识别等任务，识别精度不逊于 CNN（convolutional neural network，卷积神经网络）加速器芯片。在线学习能力赋予神经形态芯片更智能的信息处理方式，英特尔的 Loihi 基于 14 nm FinFET 技术，实现了在线学习，包括 128 个神经形态核，支持多达 12.8 万个神经元和 1.28 亿个神经

突触。与其它典型的 SNN 芯片相比，Loihi 芯片在解决 MNIST 手写体识别问题上将学习速度提高了 100 万倍[20]。Neurogrid 是一种数字-模拟混合的可编程神经形态芯片，提供了一百万个神经元及其突触连接的多尺度大脑模型的实时模拟。该系统由核心神经芯片组成，具有 10 个 256×256 的脉冲神经元矩阵。海森堡大学的 BrainScaleS 芯片也采用了模拟和数字的混合设计，工作速度比传统计算机快 1000 至 10000 倍。第二代的 BrainScaleS 更是具备了可自由编程的线上学习能力，对实现实时学习过程具有重要意义。然而，考虑到晶体管并不是专门为实现神经形态计算而发明和优化的，缺乏内在的神经元和突触特性，因而通常需要复杂的电路来实现神经元和突触的相关功能[37,38]，这限制了单个芯片的集成规模。而且晶体管三维堆叠相当困难，难以形成人脑的三维拓扑结构。其次，随着晶体管尺寸已经接近物理极限，现有 CMOS 工艺集成度难以进一步提升，导致其在长远发展上看难以达到人脑的规模。与生物智能系统中 10 μm^2 的神经元面积、0.001 μm^2 的突触面积和~2fJ 的突触操作能耗相比，当前 CMOS 基神经形态系统分别是前者的 20 倍、400 倍和 2000 倍[32]，如表 1.2 所示。因此，发展新原理器件实现更高效的神经形态计算是未来发展类脑智能的必然需求。

表 1.2 生物系统和 CMOS 基神经形态系统的对比[32]

Types	Biological system	Silicon	Ratio
Neuron density	100 k/mm^2	5 k/mm^2	20 ×
Synaptic area	0.001μm^2	0.4μm^2	400 ×
Synaptic Op energy	~2 fJ	~4 pJ	2000 ×
Max firing rate	100 Hz	1 GHz	10,000,000 ×
Synaptic error rate	75%	0%	∞

1.4.2 忆阻器基脉冲神经网络

基于忆阻器的神经形态计算是一种基于基础器件变革的新兴计算技术[39, 40]。与现有 CMOS 晶体管相比，忆阻器具有结构简单、功耗低、可微缩性好、动力学机制丰富、易于三维集成等优点[15,41-43]，且忆阻器与生物神经元和突触有着类似的离子动力学过程，可以更忠实地模拟突触和神经元的相关功能[8]。对于突触而言，突触权重由忆阻器电导表示，结合忆阻器电导的连续可变性，可实现生物突触效率和突触可塑性这两个最重要特征的模拟。根据欧姆定律和基尔霍夫定律，其阵列形式可以自然实

现神经网络中的点积操作（$\Sigma_i V_i \times W_i$）[44]，如图 1.6 所示，并且同时具备模拟计算、并行计算、存内计算的特点，这对于神经形态计算的实现具有重要意义。所有前神经元的脉冲信号输入到阵列中，流过忆阻器突触器件的电流作为加权后的值在后神经元的输入端积累求和。当前基于多种转变机制的忆阻器，例如氧化还原反应基忆阻器[45-49]、相变基忆阻器[40, 50-52]、铁电基忆阻器[53, 54]和磁阻隧穿基忆阻器[55, 56]等，已被成功用于原位点积运算和基于 STDP 规则的突触学习。忆阻器的交叉阵列也可以使用 CMOS 神经形态芯片的架构以事件驱动的方式连接，以构建具有原位计算和学习功能的高密度、大规模神经形态芯片。此外，忆阻器内在的阈值转变和积分特性也可用于模拟脉冲神经元的相关功能，已在上述的多种忆阻器[57-65]中得到了验证。这里，神经元电路有时也用 CMOS 技术来实现，构成忆阻器-CMOS 混合神经形态芯片[66-68]。鉴于忆阻器在构建神经形态芯片上的潜在优势，被认为是构建低功耗、高密度神经形态芯片的理想硬件单元，当前已引起了学术界和工业界的广泛关注。

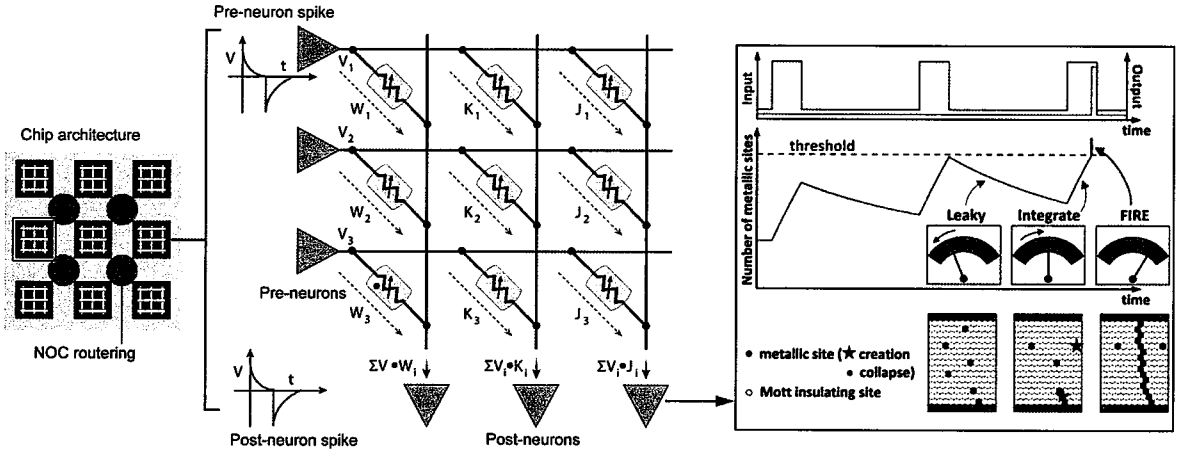


图 1.6 忆阻器基神经形态系统硬件原理图[1, 61]

1.5 选题意义和研究内容

随着物联网、人工智能、云计算等新兴技术的发展，人类社会正由信息化向智能化迈进。国际上，Google、Facebook、Microsoft、IBM 等国际巨头公司都在布局人工智能技术。发展人工智能技术有助于提高我国社会、经济和国防等领域的智能化信息水平，助推传统产业升级。也是《“十三五”国家科技创新规划》等国家科技发展战略的重要内容。

人工智能技术发展的核心是芯片，然而，随着半导体工艺趋向物理极限，摩尔定律脚步放缓，基于现有CMOS技术的计算芯片性能提升速度已经不足以应对智能计算任务的爆炸性增长。忆阻器的出现为发展人工智能计算芯片提供了变革型的技术路线，在国际上引起了广泛地关注。

相对于国外半导体产业，我国在半导体核心计算芯片领域方面的研究起步较晚，一直以来是行业的痛点。然而忆阻器作为一种新兴器件，有望从基础器件层面实现芯片技术的变革。基于忆阻器的新一代神经形态计算技术正处于研究初期，相关体系和理论架构并不成熟，尚未形成垄断优势，这为我国在智能化时代神经形态芯片领域的发展提供了新机遇。但也正因为相关研究刚刚起步，还有许多关键科学和技术问题仍未解决，其中包括：（1）忆阻器突触器件的性能优化与集成；（2）忆阻器基神经元电路功能的稳定和丰富；（3）一定规模系统的构建与验证。

本论文针对如何利用忆阻器实现神经元和神经突触功能，从而构建高效的神经形态系统开展了系统和深入的研究，论文共分 7 章，每章的内容如下：

第一章，介绍本论文的研究背景和研究意义。在人工智能的大背景下对神经形态计算技术进行了概述；比较分析了神经形态硬件实现的两种主要方案；最后指出了本论文的研究意义。

第二章，对忆阻器基神经形态计算技术进行了概述。从器件的基本分类、工作机制、神经元突触的仿生特性、神经元电路的模拟、到系统实现等方面进行了阐述。

第三章，开展了忆阻器在神经突触功能实现方面的研究。验证了生物突触的长时程可塑性、短时程可塑性以及反复训练过程中短时程可塑性到长时程可塑性的转变。并且根据系统应用进一步优化了器件设计，实现了缓变性能良好的突触器件，探讨了不同脉冲编程方案下突触电导更新的非线性对系统性能的影响。

第四章，基于忆阻器离子动力学机制开展了神经元电路的研究。首先利用易失性阈值转变器件作为阈值开关实现了一种 LIF 神经元电路，验证了神经元电路的四个关键功能：动作电位的全或无、阈值驱动放电、不应期和输入强度调制的频率响应。另外，为进一步提高神经元的集成度，丰富神经元电路的功能，我们对器件进行了工艺优化并提出了忆阻器-CMOS 混合设计的神经元电路方案，最后结合专门设计的侧向抑制阵列首次实验验证了一个全硬件多层脉冲神经网络系统并实现了在线训练。

第五章，开展了转换 SNN 神经元电路的研究，以解决 SNN 识别精度比较低的问题。利用电路内部固有的寄生电容，构建了 NbO_x 基的 1T1R 神经元，其输出脉冲发放频率与输入电压关系可以匹配 ANN 中的 ReLU 激活函数。结合该神经元电路进一步构建了 320 × 10 的网络，实现了 ANN 到 SNN 的转换，摒弃了 ANN 中 ADC 的使用，在 MNIST 数据库上实现了与 ANN 相当的识别率。

第六章，首次基于忆阻器开展了传入神经电路的研究。为实现神经形态机器与环境之间交互的紧凑接口，提出了一种基于 NbO_x 忆阻器的传入神经电路。该传入神经电路可以将模拟信号转换成动态的频率信号。系统研究了该传入神经在不同输入信号形式下的工作模式，并进一步将传入神经与压电传感器相连，构建了一个不需任何外部电源的人工机械感受系统。

第七章，对本文的工作进行了总结，并对未来的工作进行了展望。

参考文献

[1] Roy K, Jaiswal A, and Panda P, Towards spike-based machine intelligence with neuromorphic computing [J]. Nature, vol. 575, pp. 607-617, Nov 2019.

[2] LeCun Y, Bengio Y, and Hinton G, Deep learning [J]. Nature, vol. 521, pp. 436-44, May 28 2015.

[3] Chen Y, Luo T, Liu S, et al., DaDianNao: A Machine-Learning Supercomputer [J]. pp. 609-622, 2014.

[4] Silver D, Huang A, Maddison CJ, et al., Mastering the game of Go with deep neural networks and tree search [J]. Nature, vol. 529, pp. 484-9, Jan 28 2016.

[5] Vinyals O, Babuschkin I, Czarnecki WM, et al., Grandmaster level in StarCraft II using multi-agent reinforcement learning [J]. Nature, vol. 575, pp. 350-354, Nov 2019.

[6] International Technology Roadmap for Semiconductors[M]. http://www.itrs2.net/, 2015.

[7] Schuller IK and Stevens R, Neuromorphic Computing:From Materials to Systems Architecture (Report of a Roundtable Convened to Consider Neuromorphic Computing Basic Research Needs)[M]. U.S. Department of Energy, 2015.

[8] Xia Q and Yang JJ, Memristive crossbar arrays for brain-inspired computing [J]. Nature Materials, vol. 18, pp. 309-323, 2019.

[9] NVIDIA Launches the World's First Graphics Processing Unit: GeForce 256 http://www.nvidia.com/object/IO_20020111_5424.html (NVIDIA, accessed 30 July 2018). [J].

[10] Jouppi NPea, In-Datacenter Performance Analysis of a Tensor Processing Unit [J]. In 44th Int. Symp. Computer Architecture (ISCA), pp. 1-12, 2017.

[11] Purves D, Augustine GJ, Fitzpatrick D, et al., Neuroscience, 3rd ed. [M]. Inc. Massachusetts, USA: Sinauer Associates, 2012.

[12] Caporale N and Dan Y, Spike timing-dependent plasticity: a Hebbian learning rule [J]. Annual review of neuroscience, vol. 31, pp. 25-46, 2008.

[13] Perez-Carrasco JA, Bo Z, Serrano C, et al., Mapping from Frame-Driven to Frame-Free Event-Driven Vision Systems by Low-Rate Rate Coding and Coincidence Processing--Application to Feedforward ConvNets [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 2706-2719, 2013.

[14] Yu SM, Neuro-Inspired Computing With Emerging Nonvolatile Memory [J]. Proceedings of the Ieee, vol. 106, pp. 260-285, Feb 2018.

[15] Zidan MA, Strachan JP, and Lu WD, The future of electronics based on memristive systems [J]. Nature Electronics, vol. 1, pp. 22-29, 2018.

[16] Burr GW, Narayanan P, Shelby RM, et al., Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: comparative performance analysis (accuracy, speed, and power), in 2015 Ieee International Electron Devices Meeting, ed New York: Ieee, 2015.

[17] Mead C, Neuromorphic electronic systems [J]. Proceedings of the IEEE, vol. 78, pp. 1629-1636, 1990.

[18] Urbanczik R and Senn W, Learning by the dendritic prediction of somatic spiking [J]. Neuron,

vol. 81, pp. 521-8, Feb 5 2014.

[19] Abbott LF and Regehr WG, Synaptic computation [J]. Nature, vol. 431, pp. 796–803, 2004.

[20] 尤政, 魏少军, 吴华强, et al., 人工智能芯片技术白皮书 [R]. 2018.

[21] Maass W, Networks of spiking neurons: The third generation of neural network models [J]. Neural Networks, vol. 10, pp. 1659-1671, Dec 1997.

[22] Rumelhart DE, Geoffrey E. Hinton GE, and J. Williams RJ, Learning representations by back-propagating errors [J]. Nature, vol. 323, pp. 533–536, 1986.

[23] Zador CFSaAM, Novel integrate-and-fire-like model [J]. Proceedings of the 5th Joint Symposium on Neural Computation, 1998.

[24] Huxley AHaA, A quantitative description of membrane current and its application to conduction and excitation in nerve [J]. The Journal of Physiology, vol. 117, pp. 500-544, 1952.

[25] Izhikevich EM, Simple model of spiking neurons [J]. IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 14, pp. 1569-1572, 2003.

[26] Izhikevich EM, Which Model to Use for Cortical Spiking Neurons? [J]. IEEE Transactions on Neural Networks, vol. 15, pp. 1063–1070, 2004.

[27] Froemke RC and Dan Y, Spike-timing-dependent synaptic modification induced by natural spike trains [J]. Nature, vol. 416, pp. 433-438, Mar 2002.

[28] Pfeiffer M and Pfeil T, Deep Learning With Spiking Neurons: Opportunities and Challenges [J]. Front Neurosci, vol. 12, p. 774, 2018.

[29] Rueckauer B, Lungu IA, Hu Y, et al., Conversion of Continuous-Valued Deep Networks to Efficient Event-Driven Networks for Image Classification [J]. Front Neurosci, vol. 11, p. 682, 2017.

[30] Davies M, Srinivasa N, Lin T-H, et al., Loihi: A Neuromorphic Manycore Processor with On-Chip Learning [J]. Ieee Micro, pp. 82-99, 2018.

[31] Imam N and Cleland TA, Rapid online learning and robust recall in a neuromorphic olfactory circuit [J]. Nature Machine Intelligence, vol. 2, pp. 181-191, 2020.

[32] Davies M, Putting the 'learning' in machine learning processors: an introduction to the Loihi neuromorphic research chip [R]. Zenodo2018.

[33] Cassidy AS, Merolla P, Arthur JV, et al., Cognitive Computing Building Block: A Versatile and Efficient Digital Neuron Model for Neurosynaptic Cores, presented at the 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 2013.

[34] Merolla PA, Arthur JV, Alvarez-Icaza R, et al., A million spiking-neuron integrated circuit with a scalable communication network and interface [J]. Science, vol. 345, pp. 668-673, 2014.

[35] Benjamin BV, Gao P, McQuinn E, et al., Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations [J]. Proceedings of the Ieee, vol. 102, pp. 699-716, May 2014.

[36] Davison AP, Müller E, Schmitt S, et al., HBP Neuromorphic Computing Platform Guidebook[M]. https://www.humanbrainproject.eu/en/silicon-brains/how-we-work/hardware/, 2020.

[37] Indiveri G, Linares-Barranco B, Hamilton TJ, et al., Neuromorphic silicon neuron circuits [J]. Front Neurosci, vol. 5, p. 73, 2011.

[38] Douglas MMaR, A silicon neuron [J]. Nature, vol. 354, pp. 515-518, 1991.

[39] Jo SH, Chang T, Ebong I, et al., Nanoscale memristor device as synapse in neuromorphic systems [J]. Nano Lett, vol. 10, pp. 1297-301, Apr 14 2010.

[40] Kim S, Ishii M, Lewis S, et al., NVM Neuromorphic Core with 64k-cell (256-by-256) Phase Change Memory Synaptic Array with On-Chip Neuron Circuits for Continuous In-Situ Learning, in 2015 Ieee International Electron Devices Meeting, ed New York: Ieee, 2015.

[41] Yang JJ, Strukov DB, and Stewart DR, Memristive devices for computing [J]. Nat Nanotechnol, vol. 8, pp. 13-24, Jan 2013.

[42] Jain S, Ankit A, Chakraborty I, et al., Neural network accelerator design with resistive crossbars: Opportunities and challenges [J]. IBM Journal of Research and Development, vol. 63, pp. 10:1-10:13, 2019.

[43] Narayanan P, Fumarola A, Sanches LL, et al., Towards on-chip acceleration of the backpropagation algorithm using non-volatile memory [J]. IBM Journal of Research and Development, vol. 61, 2017.

[44] Li C, Hu M, Li Y, et al., Analogue signal and image processing with large memristor crossbars [J]. Nature Electronics, 2017.

[45] Prezioso M, Merrikh-Bayat F, Hoskins BD, et al., Training and operation of an integrated neuromorphic network based on metal-oxide memristors [J]. Nature, vol. 521, pp. 61-64, 2015.

[46] Yao P, Wu H, Gao B, et al., Face classification using electronic synapses [J]. Nat Commun, vol. 8, p. 15199, May 12 2017.

[47] Sheridan PM, Cai F, Du C, et al., Sparse coding with memristor networks [J]. Nat Nanotechnol, May 22 2017.

[48] Cai F, Correll JM, Lee SH, et al., A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations [J]. Nature Electronics, vol. 2, pp. 290-299, 2019.

[49] Prezioso M, Mahmoodi MR, Bayat FM, et al., Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits [J]. Nat Commun, vol. 9, p. 5311, Dec 14 2018.

[50] Ambrogio S, Ciocchini N, Laudato M, et al., Unsupervised Learning by Spike Timing Dependent Plasticity in Phase Change Memory (PCM) Synapses [J]. Front Neurosci, vol. 10, p. 56, 2016.

[51] Pritish Narayanan GWB, Stefano Ambrogio, Robert M. Shelby, Neuromorphic Technologies for Next-Generation Cognitive Computing, presented at the 2017 IEEE International Memory Workshop (IMW) Monterey, CA, USA, 2017.

[52] Ambrogio S, Narayanan P, Tsai H, et al., Equivalent-accuracy accelerated neural-network training using analogue memory [J]. Nature, vol. 558, pp. 60-67, Jun 2018.

[53] Jerry M, Chen P-Y, Zhang J, et al., Ferroelectric FET Analog Synapse for Acceleration of Deep Neural Network Training, in 2017 Ieee International Electron Devices Meeting, ed New York: IEEE, 2017.

[54] Kim MK and Lee JS, Ferroelectric Analog Synaptic Transistors [J]. Nano Lett, vol. 19, pp. 2044-2050, Mar 13 2019.

[55] Mondal A and Srivastava A, Energy-efficient Design of MTJ-based Neural Networks with Stochastic Computing [J]. ACM Journal on Emerging Technologies in Computing Systems, vol. 16, pp. 1-27, 2020.

[56] Song KM, Jeong J-S, Pan B, et al., Skyrmion-based artificial synapses for neuromorphic computing [J]. Nature Electronics, vol. 3, pp. 148-155, 2020.

[57] Wang Z, Joshi S, Savel'ev S, et al., Fully memristive neural networks for pattern classification

with unsupervised learning [J]. Nature Electronics, vol. 1, pp. 137-145, 2018.

[58] Wang JJ, Hu SG, Zhan XT, et al., Handwritten-Digit Recognition by Hybrid Convolutional Neural Network based on HfO_2 Memristive Spiking-Neuron [J]. Sci Rep, vol. 8, p. 12546, Aug 22 2018.

[59] Zhang Y, He W, Wu Y, et al., Highly Compact Artificial Memristive Neuron with Low Energy Consumption [J]. Small, p. e1802188, Nov 14 2018.

[60] Tuma T, Pantazi A, Le Gallo M, et al., Stochastic phase-change neurons [J]. Nat Nanotechnol, May 16 2016.

[61] Stoliar P, Tranchant J, Corraze B, et al., A Leaky-Integrate-and-Fire Neuron Analog Realized with a Mott Insulator [J]. Advanced Functional Materials, p. 1604740, 2017.

[62] Mulaosmanovic H, Chicca E, Bertele M, et al., Mimicking biological neurons with a nanoscale ferroelectric transistor [J]. Nanoscale, vol. 10, pp. 21755-21763, Dec 2018.

[63] Chen C, Yang M, Liu S, et al., Bio-Inspired Neurons Based on Novel Leaky-FeFET with Ultra-Low Hardware[M]. New York, 2019.

[64] Dutta S, Saha A, Panda P, et al., Biologically Plausible Ferroelectric Quasi-Leaky Integrate and Fire Neuron[M]. New York: Ieee, 2019.

[65] Wu MH, Hong MC, Chang C-C, et al., Extremely Compact Integrate-and-Fire STT-MRAM Neuron: A Pathway toward All-Spin Artificial Deep Neural Network[M]. New York: Ieee, 2019.

[66] 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, in 2019 Ieee International Electron Devices Meeting, ed New York: Ieee, 2019.

[67] Valentian A, Rummens F, Vianello E, et al., Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses, in 2019 Ieee International Electron Devices Meeting, ed New York: Ieee, 2019.

[68] Yan B, Yang Q, Chen W-H, et al., RRAM-based Spiking Nonvolatile Computing-In-Memory Processing Engine with Precision-Configurable In Situ Nonlinear Activation[M]. vol. T86-T87. New York: Ieee, 2019.

第2章 忆阻器基神经形态计算技术概述

2.1 忆阻器概述

忆阻器，也称为电阻开关，其电阻值连续可调，状态取决于内部所流过的电流或外部所施加的电压的历史状态，被认为是除电阻、电容、电感之外的第四种基本的电子元件。关于忆阻器的理论可以追溯到上世纪 70 年代，加利福尼亚大学伯克利分校的蔡少棠（Leon Chua）教授最先从电路对称的角度出发，推导出了忆阻器的存在，并且证明了忆阻器阻值和流经的电荷之间的关系[1]。在其后的很长一段时间内，忆阻器并没有给出实验的证明。直到 2008 年，惠普实验室的 D. B. Strukov 等人利用掺杂和非掺杂的双层 TiO_2 薄膜器件首次将忆阻器理论和实验现象联系到了一起[2]。

2.1.1 忆阻器分类

忆阻器通常为“三明治”结构，即金属/电介质/金属（metal/dielectric/metal，MIM），金属一般是导电性良好的导电材料，电介质可以是金属氧化物材料、相变材料、铁磁材料、铁电材料等。根据介质层材料的不同，所形成的忆阻器通常具有不同的转变机制[3]，如图 2.1 所示。例如：（1）氧化还原反应机制（redox）：电解质层材料中通常伴随有阳离子或阴离子的移动以及移动过程中的氧化还原反应；（2）相变机制（phase change）：对应电介质材料在热激励下从非晶到晶态的转变，材料的晶态和非静态具有不同的电阻状态；（3）铁电隧穿结效应（ferroelectric tunnel junction（FTJ）effect）：介电层内铁电极子取向导致隧穿电阻的变化；（4）磁阻效应（magnetoresistive effect）：磁极子翻转导致介电层隧穿电阻的变化。这些转变机制所导致的电阻转变通常具有非易失特性，因而已被广泛用于新型非易失性存储器的实现，例如阻变存储器[4]，相变存储器[5]，铁电存储器[6]和磁存储器[7]等。这些新型存储器具有低功耗、低操作电压、高密度、高操作速度，高可靠性以及可微缩性好等优点，与传统的基于电荷的存储器相比更具有竞争力[8]。此外，这种具有电阻记忆特性的器件可以通过内在的物理机制处理信息，大大减小了硬件面积并降低系统能耗，例如神经形态计算[9]和硬件安全[10]等，其中忆阻器在神经形态计算方面的应用是本论文研究的重点。

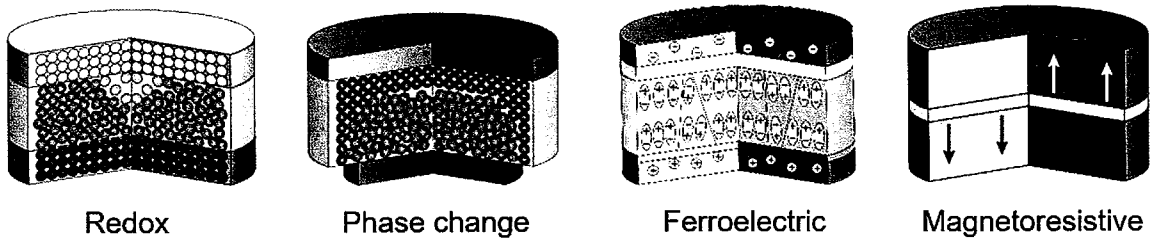


图 2.1 不同转变机制类型忆阻器的器件结构原理图[3]

2.1.2 忆阻器性能参数及应用需求

虽然不同转变机制的介电层形成不同类型的忆阻器，但就其应用而言这些器件具有一些共性，如保持特性（retention）、耐久性（endurance）、转变速度（switching speed）、转变能耗（switching energy）、可微缩性（scalability）、随机性（stochasticity）、可分辨的状态数（distinguishable states）、良率（yield）等。Z. Wang 等人在最近的综述文章中给出了当前所报道的不同类型忆阻器之间最优性能的对比[3]（表 2.1）。对于不同的应用来说，其特性的需求是这些基本特性的集合并且根据具体应用而略有不同。

表 2.1 不同类型忆阻器的器件特性[3]

Types	MiS (%)	MaDS	MiSE (fJ)	MaSS (ps)	MaE (cycles)	MaR@RT (years)	MiS (nm)
Redox	~9.65	64 levels	<10[11]	85	10^{12}	>1000	~2
Phase change	~9.62	16 levels	1000	700	10^{11}	>1000	~5
FTJ	~24.5	~32 levels[12]	100	600[12]	4×10^{6}	>100[12]	~20
Magnetoresistive	~0.29	~20 levels[13]	10	200	10^{12}	10	~10

MiS: minimum stochasticity; MaDS: maximum number of distinguishable states; MiSE: minimum switching energy; MaSS: maximum switching speed; MaE: maximum endurance; MaR@RT: maximum retention at room temperature; MiS: minimum scalability

下面将对这些参数及对应的应用需求进行简单的介绍。

（1）保持特性（retention）：又可以称为器件的状态稳定性，指器件被编程到某一个状态后该状态可以稳定存在的时间长短。对于需要作长时间存储的应用来讲，例如存储器或者用作边缘推理的突触阵列，一般要求器件可以在 85°C 保持十年以上。然而，

作为神经突触只是为实现网络训练过程的加速时，对器件的保持特性便没有这么高的要求。甚至，如果器件的衰退速度比较快的话（< ms），可以用来实现神经元电路或者突触的短时程可塑性机制。

（2）耐久性（endurance）：又称抗疲劳特性，指器件可以进行不同阻态之间来回转换的最多可编程次数。对于几乎所有的计算和存储应用类型来讲，其耐久性越高越好。据报道，redox 基的忆阻器[4]和 magnetoresistive 忆阻器[14]的耐久性均可以高达 10^{12} 次以上。比较而言，器件在完成神经元模拟时所需要的耐久性要高于用作存储时的情况，因为神经元的每一次脉冲发放都对应一个转变周期，而器件在用作存储或者突触时并不是每次都需要发生状态的转变。

（3）转变速度（switching speed）：指器件从一个状态编程到另一个状态所需要的最短时间。不管是用作存储、计算还是神经元的实现，一般来讲转变速度越快越好。快的转变速度通常对应更快的计算速度或更快的写入时间，可以大大降低操作的延时。据报道，redox 忆阻器的最快转变速度已经可以达到 ~85 ps[15]。

（4）转变能耗（switching energy）：器件从一个状态转变到另一个状态所需要消耗的能量。和耐久性一样，几乎对于所有的应用类型来讲，该能耗越低越好。转变能耗越低对于系统应用来讲所需要的能量就越少，这在当前大数据时代对于低功耗的终端产品来说尤为重要。据报道 magnetresistive 器件的转变能耗已经可以低于 10 fJ[16]，接近生物突触的单个脉冲操作能耗（~2 fJ）。

（5）可微缩性（scalability）：指器件在保持有阻变特性的前提下能够达到的最小尺寸。器件的该性能决定了芯片的集成密度，对于所有应用来讲，器件的可微缩性越小越好。当前在 redox 忆阻器中已经可以实现 2 nm×2 nm 的器件尺寸[17]，约为单个生物突触面积（~0.001 μm^2）的 1/250，这表明了忆阻器在用作神经形态计算时具有巨大的面积优势。

（6）随机性（stochasticity）：又可以称为波动性（variation）。在不同周期的操作中，器件在转变过程中的转变电压或者编程状态都会有所不同。一般可分为同一个器件不同周期的波动性和不同器件之间的波动性。对于存储应用来讲，该波动性越小越好，然而对于计算或者神经元的模拟来讲，则要视情况而定。例如，利用忆阻阵列解决组合优化问题时，适当的波动性有助于系统在迭代过程中跳出局部最小值达到全局最优

值。对于精确推理来讲，波动性则越小越好。相应的，对于神经元的模拟来讲，由于生物神经元具有本征的随机性，在信息编码和鲁棒性上都有很好的应用，因此器件具有一定的随机性对于神经元的仿生实现也是有必要的。当前已有报道利用忆阻器的随机性进行族编码和概率神经网络的实现[18, 19]。

（7）可分辨的状态数（distinguishable states）：也可以说是器件的多值存储能力。指的是单个器件可以编程到的状态数目。对于一般应用来讲，该状态数目越多越好，可以在不增加成本的条件下大大提升存储密度和计算能力。当前在 redox 忆阻器中报道的可实现的最多状态数目已经可以达到 64（2^6）个[20]。

（8）良率（yield）：指在大规模阵列制备时，可以正常工作的器件占总器件的比例。该比例的大小与芯片是否可实现系统应用直接相关。与其它可视化的产品一样，器件的良率也是越高越好。

图 2.2 给出了不同器件相应特性的对比，可以更直观的看到不同类型器件在当前工艺条件下所能达到的性能。综合考虑器件的各种特性以及制备工艺，我们主要基于 redox 忆阻器展开了相关的研究工作，后面也将主要针对该类型器件的特性和有关神经形态计算方面的工作展开介绍。

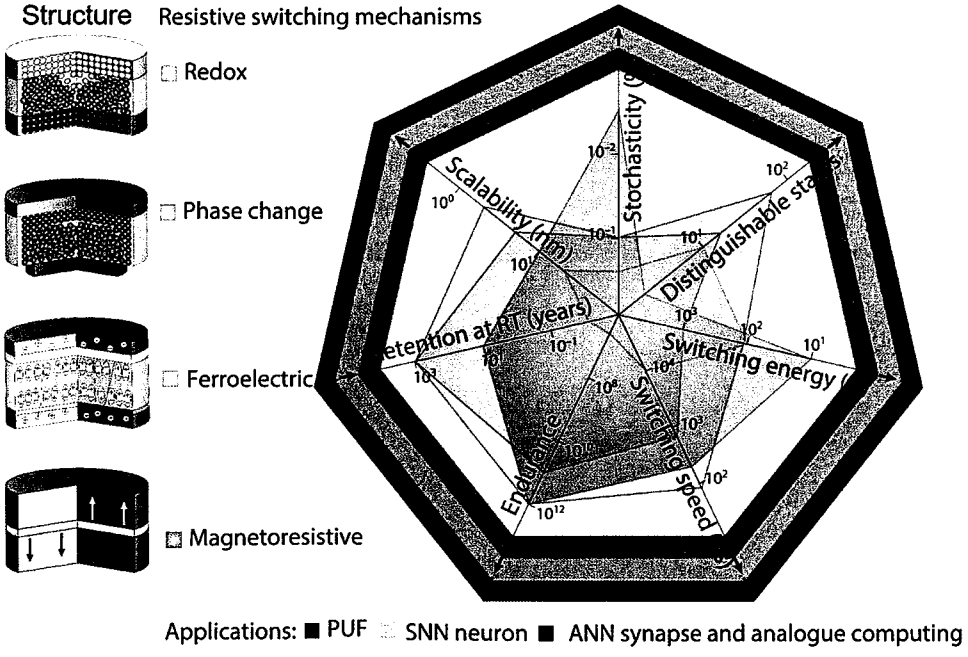


图 2.2 不同转变机制忆阻器的特性对比[3], PUF: physical unclonable function

2.1.3 Redox 忆阻器的转变模式

Redox 忆阻器是由于离子（阳离子或阴离子）迁移引起器件结构发生变化从而导致电阻发生变化的一类器件，该变化过程涉及电极材料或绝缘体材料的氧化还原反应或者两者都有[3,21,22]。根据器件结构的不同，迁移过程可能在几乎整个器件内形成贯穿的局部导电通路，也可能发生在电极和材料的界面处或者均匀分布在整个介电层内，由此可以将忆阻器大致分为两大类：导电细丝型（filament）和非导电细丝型（non-filament）。其中导电细丝型又可以根据细丝成分分为氧空位细丝型忆阻器和金属细丝型忆阻器。下面将对这几种具有不同转变机制的 Redox 忆阻器的相关特性分别进行介绍：

（1） 氧空位细丝型忆阻器：该类型的忆阻器是 redox 忆阻器中一种常见的类型，在电激励下通常在两电极间的介电层内导致元素化学价的变化（例如 Ti^{4+}变成 Ti^{3+}）并且伴随相应的氧化还原反应，最终形成连通上下电极的导电通路，如图 2.3（a）所示。2010 年，D.-H. Kwon 等人[23]用透射电子显微镜（TEM）在单极性转变的 $Pt/TiO_2/Pt$ 器件中观察到锥形柱状的 Ti_4O_7 细丝导电通路，给出了直接的实验证明（图 2.3（b））。大多数的介电层由一层或者多层绝缘体材料组成（例如，金属氧化物 AlO_x，TaO_x，TiO_x，ZrO_x，WO_x，TaO_x，或者氮化物 AlN_x，CuN_x 等），两个金属电极分别作为上下电极形成交叉结构的器件。在器件制备或者 forming 的过程中，介电层材料与活性相对比较高的电极（如 Ti、TiN、Ta 等）之间会首先产生缺陷，形成氧空位或者氮空位。另一个电极一般为惰性电极，负责提供电学接触并在接触面处形成一个比较高的势垒（例如，肖特基势垒）。近年来，随着材料体系及器件性能的不断探索和优化，TaO_x 和 HfO_x 材料已经逐渐成为主流的介电层材料[24]，一方面是因为这两种材料制备的器件具有更高的可靠性，另一方面是因为这两种材料体系有 CMOS 兼容的优势。器件的初始态通常为高阻态，为使器件正常的工作，需要一个 forming 过程对器件进行初始化形成导电通路，随后在正反电压激励下使得器件在高低阻态之间来回切换，表现出双向转变特性。图 2.3（c）给出了一般的直流转变模式曲线，在正向电压扫描下，当达到 forming 电压时，流过器件的电流会突然达到限流（compliance current, CC），器件由高阻态变为低阻态。多数器件 forming 之后会保持在低阻的状态，表现为非易失的特性。若要使器件再次变为低阻态，需要对器件进行反向的 reset 操作，当达到 reset 电压后，流过器件的电流逐渐减小，器件由低阻态再次回到高阻态。随后再次对

器件进行 set 操作，器件由高阻态变为低阻态，这里的 set 电压一般比 forming 电压要小一些。

除了这种非易失的器件转变特性以外，一些器件在 forming 之后无法保持低阻态，当回扫电压低于一定的电压值时，器件会自发的从低阻态变为高阻态，从而表现出易失性的双向阈值转变行为，如图 2.3（d）所示。这种器件特性常见于 NbO_x 或 VO_x 材料体系中，forming 操作后，器件会在介质层内形成 NbO_2 或者 VO_2 导电通道，该通道具有绝缘态-金属态转变（IMT）特性，静息状态时，通道处于高阻态。当施加在器件上的电压达到阈值电压（V_{th}）时，通道会由高阻的绝缘态变为低阻的金属态，从而导致器件由高阻态变为低阻态。当器件上的电压小于一定的保持电压（V_{hold}）时，流过通道的能量不足以维持其金属态，这时通道自发的回到绝缘态，器件再次由低阻态变为高阻态。由于该转变行为是由热量驱动的，所以器件在正负电压下表现出对称的阈值转变行为。

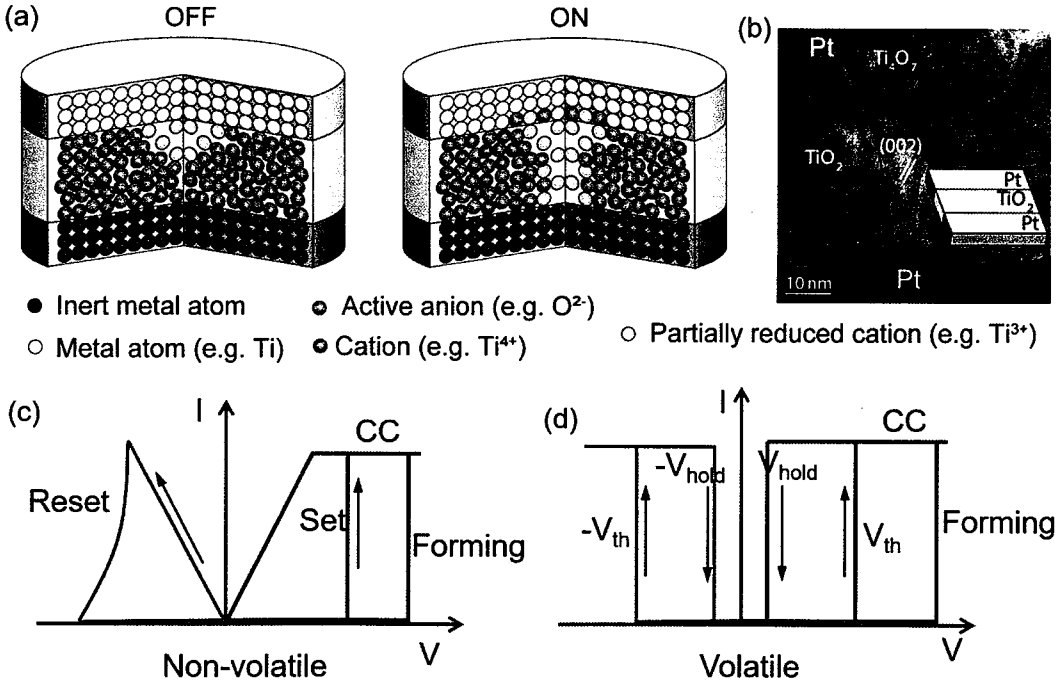


图 2.3 （a）氧空位细丝型忆阻器在高低阻态下的结构图[3]；（b）$Pt/TiO_2/Pt$ 器件在低阻态下的 Ti_4O_7 导电通道[23]；（c）器件的非易失双向转变模式曲线；（d）器件的易失性阈值转变模式曲线

（2） 金属细丝型忆阻器：金属细丝型忆阻器也称为离子导电桥忆阻器。与氧空位细

丝型忆阻器相比，该类型的器件通常需要一个电化学活性金属电极（例如 Ag 或 Cu 等）和一个惰性金属电极（例如 W、Pd、Pt 或者 Au 等）。两个电极之间有一层氧化物、硫化物或者半导体材料作为介质层。该类型器件的阻变过程是由阳离子的氧化还原和电迁移导致的，这涉及到电化学活性金属电极（如 Ag 或 Cu）氧化为阳离子（Ag^{+} 或 Cu^{2+}），随后这些离子在电场的作用下在介质层中移动并被还原成核，随着成核过程的不断进行，介质层中形成金属导电通路（图 2.4（a））。在这里，介质层中离子迁移率和氧化还原速率的差异会导致导电通路的成核位置不同，从而生长模式不同[25]（图 2.4（b））。例如，在高氧化还原速率和高离子迁移率介质层中（如硫化物、碘化物、硒化物，碲化物和三元硫系化合物等良好离子导体），金属细丝一般从惰性电极方向朝着活性电极方向生长；而在低氧化还原速率和低离子迁移率介质层中（如氧化物或氮化物等低离子导体）通常导致活性金属细丝通道由活性电极向惰性电极呈锥形逐渐生长成核；对于低离子迁移率和高氧化还原速率的介质层，细丝会在电介质内部发生成核，同时大量原子会沉积在原子核的阴极侧，导致间隙填充形成通道；相应的，对于高离子迁移率和低氧化还原速率的情况，成核发生在惰性电极一侧，但有限的离子供应意味着还原主要发生在高局域电场的边缘，从而导致通道朝着活性电极方向呈树枝状生长。Y. Hirose 和 H. Hirose 首次在平面结构的 Ag/As_2S_3:Ag/Ag 器件中观察到金属细丝的存在[26]，X. Guo 等人也观测到介质层内树枝状金属细丝的形成[27]。离子通道的形成使得器件由高阻态变为低阻态并且其形成后的保持特性决定了器件在低阻态下的寿命。

与氧空位细丝型忆阻器相似，导电桥忆阻器也有非易失性和易失性两种转变模式。图 2.4（c）给出了该类器件在非易失情况下的转变曲线，器件初始态处于高阻态，为使器件可以正常工作，通常也需要一个 forming 操作对器件进行初始化。初始化后器件可在高低阻态之间进行切换。对于某些器件结构，当金属通道形成后，如果不加电激励或者电激励不足时，金属导电通路和周围介质之间的界面能最小化或者 Thomson-Gibbs 效应倾向于使得导电通路自发破灭，从而使得器件表现出易失的特性，图 2.4（d）给出了器件导电通路破灭比较迅速的转变曲线模式。2016 年，Z. Wang 等人通过电镜在 Au/SiO_x:Ag/Au 器件中首次原位观测到 Ag 细丝的生长和自发断裂过程[28]。

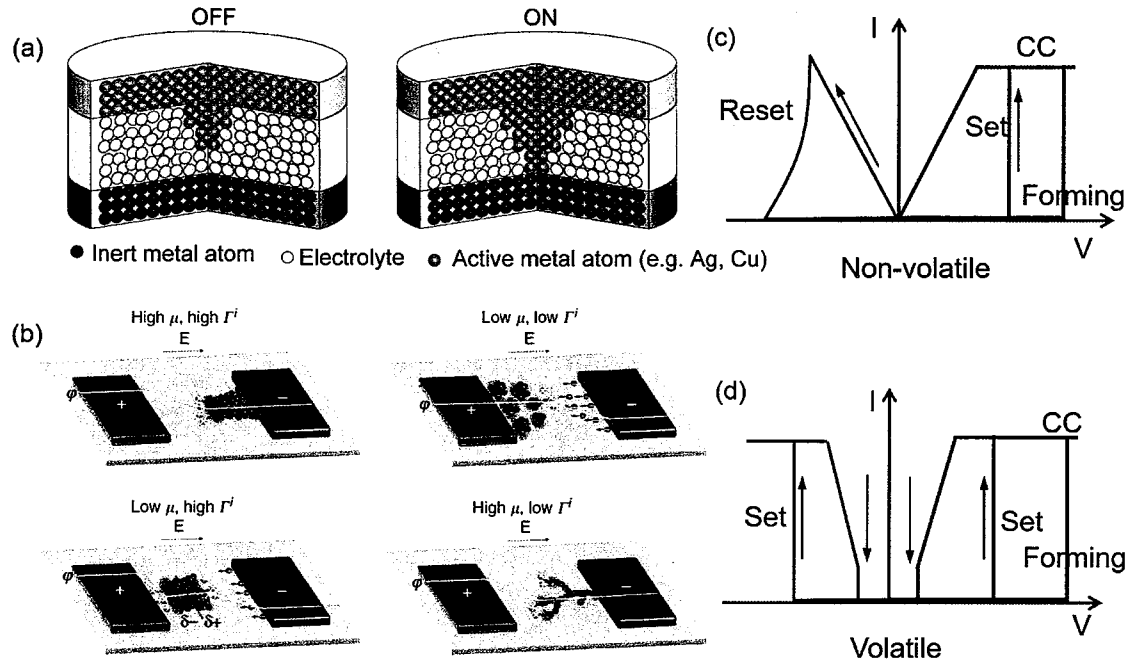


图 2.4 （a）金属细丝型忆阻器在高低阻态下的结构图[3]；（b）不同离子迁移率和氧化还原速率下离子通道的生长模式[25]；（c）器件的非易失双向转变模式曲线；（d）器件的易失性阈值转变模式曲线

（3） 非细丝型忆阻器：对于细丝型忆阻器，通常会在介电层内形成细丝状的导电通路，连接上下电极，使得器件由高阻态转变为低阻态。而对于某些器件来说，在转变过程中并不会出现明显的导电通路，因而又称为非细丝型（non-filament）忆阻器。和细丝型忆阻器类似，非细丝型忆阻器也包括一层或多层介电层，由上下两个电极连接。单层介质层材料通常为 $PrCaMnO_3$（PCMO），$La_{0.8}Sr_{0.2}MnO_3$ 或 $SrTiO_3$ 等具有钙钛矿结构的金属氧化物材料。多层介电层结构的器件通常是 TiO_x、TaO_x、HfO_x、NbO_x 等二元金属氧化物介电层的组合。相比于细丝型忆阻器 set 过程的突变特性，非细丝型忆阻器的 set 过程是缓变的而且均一性比较好。根据介电层材料和器件的结构，非细丝型器件的导电区域通常是有效电极面积覆盖的整个器件区域，或者至少是大部分区域，而不是细丝型器件的局部导电通路。另外，对于非细丝型器件来说，一般不需要 forming 过程，并且 set 过程中不需要外部电路的限流操作。非细丝型忆阻器所表现出的缓变特性也使得该类型器件可以直接在简单的脉冲编程下获得多个电阻状态而不需要像细丝型忆阻器那样精确的控制限流的大小。正是因为这种特性，非细丝型忆阻

器在突触的动态模拟和多值存储上具有很大的应用潜力。然而，非细丝型忆阻器的保持特性往往不如非易失型的细丝型忆阻器的保持特性好，这是由界面层内氧离子的扩散导致的。

非细丝型忆阻器的电阻转变过程可以描述为电场驱动下电极和介电层界面处氧空位的再分配，这种氧空位的再分配伴随着界面层材料的氧化还原反应，从而导致界面处电子传输特性的改变（例如，氧化还原反应导致了界面层肖特基势垒高度和厚度的改变）。由于转变过程伴随着界面隧穿氧化层的变化，因此器件的 set 和 reset 直流曲线都具有很大的非线性，使得该类器件可以实现单 R 的阵列集成而不需要额外的选通单元。2015 年，D. Lee 等人利用 TiN/PCMO/Pt 结构的器件制备了 8 k 的纯 R 阵列并通过 TEM 技术观测到了器件转变过程界面层的变化[29]。2017 年，D. Cooper 等人在 $Pt/SrTiO_3/Nb:SrTiO_3$ 器件中采用原位电镜技术[30]，观测到器件在转变前后 $SrTiO_3$ 层中 Ti^{4+}和 Ti^{3+}比例的变化，证明了器件在 set 和 reset 过程中氧空位的再分布并给出了进一步的原理解释，如图 2.5 所示。当在 Pt 电极上施加正向偏压或者在 $Nb:SrTiO_3$ 电极上施加负向时，$SrTiO_3$ 介电层中的氧离子在电场作用下朝着 Pt 电极移动，导致界面层处氧空位增多并在 Pt 电极处产生氧气，界面处的 Ti^{4+}被还原成 Ti^{3+}，Ti^{3+}含量增多，界面层电导增加，器件电阻降低。该反应过程可以由下面的两个反应公式表示：

$$O_O^x \rightarrow V_O^{\cdot\cdot} + 2e' + \frac{1}{2}O_2 \quad \text{.....(2.1)}$$

$$2Ti^{4+} + 2e^- \rightarrow 2Ti^{3+} \quad \text{.....(2.2)}$$

当在 $Nb:SrTiO_3$ 电极上施加正向偏压或者在 Pt 上施加负向偏压时，$Pt/SrTiO_3$ 界面发生相反的反应过程，界面处的电导降低，器件电阻增加。

除了单层介电层内氧空位的再分布导致的电阻转变，氧空位的再分布也在多层结构的器件中有报道，如 $Ta/TaO_x/TiO_2/Ti$[11, 31, 32]和 $TiN/a\text{-}Si/TiO_2/TiN$[33-35]等。另外，对于非细丝型的器件，有些也是由空间电荷陷阱机制导致的，如 $Pt/Ta_2O_5/HfO_{2-x}/TiN$[36]和 $TiN/NbO_x/TiO_y/NbO_x/Pt$[37]等。值得一提的是，非细丝型的器件是由电场作用下离子的迁移造成的电阻的变化，因而通常具有更低的操作电流，适合于那些需要低功耗但对器件的保持特性没有较高要求的场景。

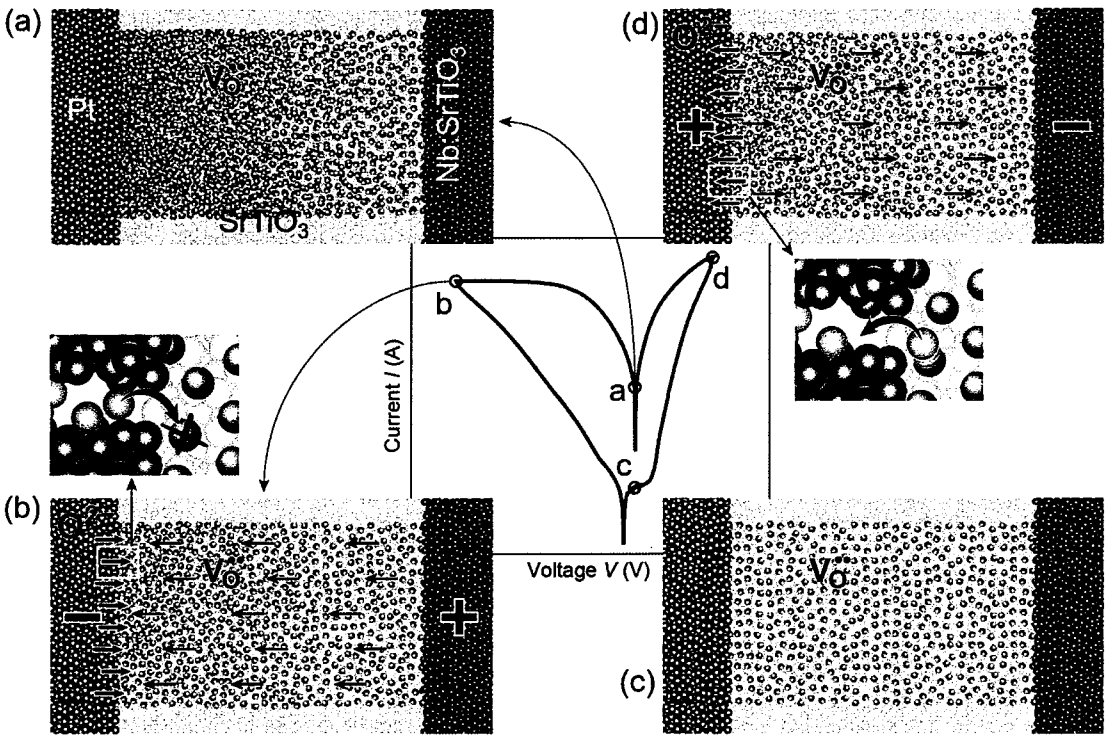


图 2.5 非细丝型 $Pt/SrTiO_3/Nb:SrTiO_3$ 器件在转变过程中 $SrTiO_3$ 介电层内氧空位的再分布情况[30]

2.1.4 忆阻器的阵列集成

在系统应用时，忆阻器通常以交叉阵列的形式进行集成，这样可以将忆阻器的集成密度提高到 $4F^2$（F 是特征尺寸）。另外，采用三维集成技术可以进一步将集成密度提高到 $4F^2/N$（N 是三维集成的层数）。然而，对于单 R 的阵列来说，其读写过程通常伴随有漏电回路，如图 2.6（a）所示。由于忆阻器是一种无源器件，当需要单独对目标器件进行读写操作时，其它器件组成的回路会产生并联的效果从而形成旁路漏电，使得器件不能编程到目标值，或者读出的电流值变大产生误读现象。另外，在执行写操作时，编程电压也可能会对旁路中的器件进行误操作。为了避免这种现象，通常需要给器件串联一个选通器（S），如图 2.6（b）所示。选通器的引入可以缓解旁路漏电流问题，使得在阵列中能够对单个器件进行读写操作。常用的选通器有两种（表 2.2）：一种是使用晶体管（transistor）作为选通器，组成 1T1R 的结构；另一种是使用两端的非线性器件（selector）作为选通器，组成 1S1R 的结构。

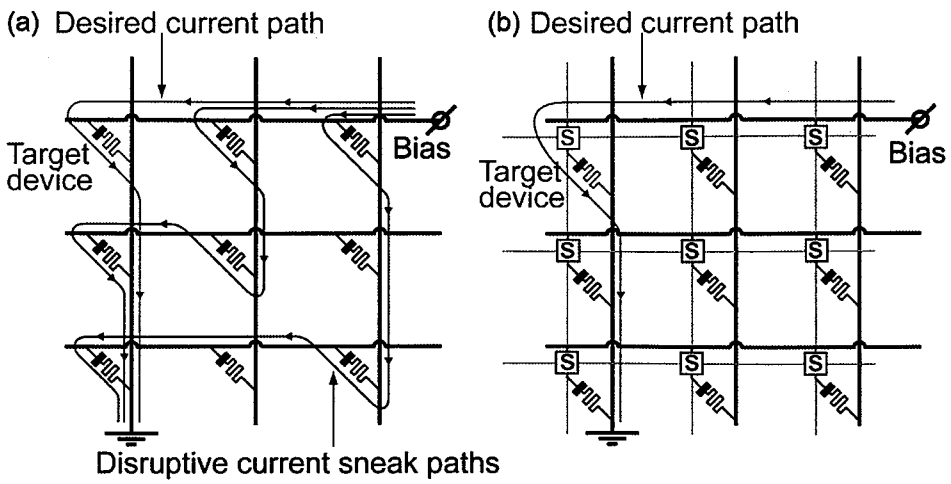
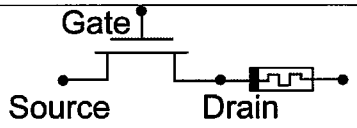
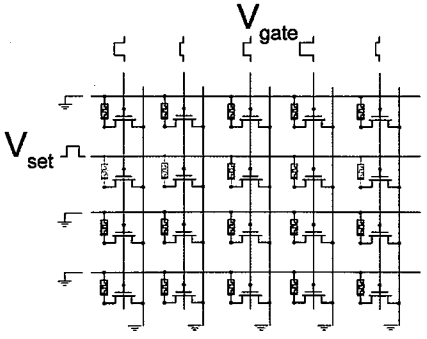
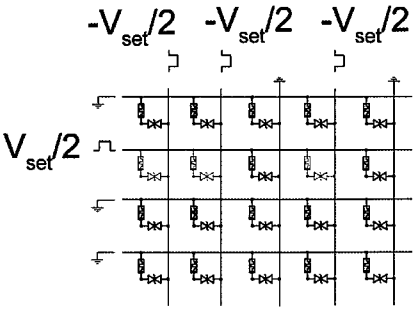


图 2.6 （a）纯忆阻器阵列中的漏电问题；（b）带有选通器的忆阻器阵列原理图[38]

表 2.2 1T1R 和 1S1R 集成方案的对比[9]

Architecture	1T1R	1S1R
Cell diagram		
Cell area	> $8F^2$	$4F^2$
Scalability and 3D stackability	Limited by transistor size; difficult for 3D stacking	Scalable at 2D and stackable for 3D
Role of the series devices	Mitigate sneak path current and half-selected issue. Control current to the memristor for precise resistance tuning	Mitigate sneak path current and half-select issue
Array programming		
Array operation (reading) schemes	Pulse amplitude/width	Pulse width

对于 1T1R 的结构来讲，忆阻器与晶体管的漏极（drain）相连，晶体管的栅极（gate）作为控制端。当晶体管的栅极施加有选通电压时，在器件远离漏极的一端或者在晶体

管的源端（source）加电压对器件进行读写操作。其它没有选通的晶体管处于关闭状态，大大缓解了旁路的漏电流。然而在可集成度上，1T1R 结构的可集成密度受限于晶体管的尺寸（$>8F^2$），这在一定程度上损害了忆阻器在高密度集成方面的优势。另外，晶体管的三维堆叠技术尚不成熟，因而 1T1R 的集成方案不利于进行三维集成。但是由于 CMOS 技术比较成熟，所以当前实现大阵列集成还是多采用 1T1R 的方法，例如美国马萨诸塞大学杨建华老师课题组制备的 8 kB 阵列[20]、清华大学的 128×8 阵列[39]以及中国科学院微电子研究所刘明课题组基于 28 nm 工艺的 1MB 阵列[40]等都是采用 1T1R 的集成方案。

与 1T1R 结构的集成方案相比，1S1R 结构采用两端的非线性器件作为选通器，1S1R 单元保持着两端的结构，因此，在集成度上 1S1R 结构仍旧具有忆阻器集成密度上的优势，并且适合三维集成。忆阻器和选通器串联时，组成的堆叠器件具有非线性特性，在进行读写操作时，通常采用“1/2 偏压”的读写方案。其中在所选需要编程器件单元的上电极所在行（或列）施加 1/2 偏压，在下电极所在列（或行）施加-1/2 偏压，其它端口全部接地。这样就使得选通的器件处于全偏压的而状态而其它未选通的器件处于半偏压的状态，从而实现对选通器件的读写操作而对其它未选通器件没有什么影响。通常，非线性选通器的非线性度越高，其可有效执行的交叉阵列规模越大。例如，当 I-V 关系中的非线性为 10^5 时，假设读出裕度为 10%，可以支持具有超过 10^{10} 个单元的交叉阵列[9]。理想的选通器不要 forming 操作，具有可微缩、可堆叠、高非线性、高耐久性、开关速度快、电流密度大、能耗低、均一性好等特点。选通器的性能要求通常比忆阻器的性能要求要高，部分原因是因为选择器不仅需要在忆阻器编程期间工作，而且要在更频繁的读取操作过程中工作。目前，实验上集成的大规模 1S1R 阵列并没有广泛报道。最近的阵列级仿真表明，加权和的精度取决于选通器的非线性[41]。因此，若要实现大规模阵列的应用，制备非线性大的选通器是必要的。我们在前面提到的细丝型忆阻器中具有易失型阈值转变特性的器件便可以在这里作为选通管使用。

另外，除了给忆阻器串联选通器以外，制备具有自选通特性的忆阻器或者 I-V 非线性比较高的器件进行纯 R 集成也是一种不错的选择，这种方案可以充分保持忆阻器理论上的优势。目前，来自加利福尼亚大学圣塔芭芭拉分校的 D. B. Strukov 等人基

于 $Pt/Al_2O_3/TiO_{2-x}/Ti/Pt$ 结构的自选通器件制备了 4 kB 的纯 R 阵列并实现了小于 4%的编程误差[42]。密歇根大学卢伟等人基于 $Au/WO_x/Pt$ 结构的非线性忆阻器实现了 54 × 108 的阵列集成并验证了阵列执行乘加操作的可行性[43]。中国科学院微电子研究所刘明课题组基于 $TiN/TiO_x/HfO_x/Ru$[44]和 $TiN/HfO_2/CuGeS/W$[45]等结构的自选通器件制备了 1 kB 的纯 R 阵列并进行了三维集成。

2.2 忆阻器基神经突触

从前面的介绍中可以看到，忆阻器的材料体系多种多样，材料体系或者器件结构的不同通常会造成器件不同的转变特性。在生物神经形态系统中，神经突触是数量最多的部件（$\sim10^{15}$）[46]，它负责神经元之间复杂的连接，在生物体的学习和记忆活动中起着至关重要的作用。当前在利用忆阻器实现神经形态计算的工作中，也多是利用忆阻器实现生物突触的相关功能。本小节将从生物突触的分类出发，着重介绍忆阻器在实现生物化学突触的长时程可塑性以及短时程可塑性等突触功能时的相关特性、实现方法及研究现状。

2.2.1 生物突触的分类

虽然生物系统中生物突触的数量非常庞大，但总体来看，生物突触可以分为两大类：一类是电突触，另一类是化学突触。相比于化学突触，电突触的数量比较少，它允许动作电位直接以电耦合的方式从一个神经元直接传到另一个神经元[46]。图 2.7(a)给出了电突触的生物结构原理图，可以看到来自前神经元的突触后膜直接通过离子通道与后神经元的突触前膜相连，形成一个间隙连接结构（gap junction）。Gap junction 在前突触和后突触中形成成对排列的离子通道，因此每对通道形成一个物理孔洞供动作电位的传输。与动作电位相关的离子（例如 Na^+和 K^+）可以简单地在前神经元和后神经元之间扩散。这种连接结构具有信号双向传递的特性，传递的方向取决于连接的两个神经元哪个先产生动作电位信号。电突触的另一种特殊性质是可以实现动作电位的快速传递，这是因为信号在电突触上的传递几乎是同步的，不会发生化学突触中信号传递过程中的延时。正因为这种特性，电突触在保证神经元之间动作信号的同步上起着重要的作用。

在生物神经系统中，大多数的突触是化学突触，因为它们是生物体实现学习和记

忆的基础，图 2.7（b）给出了化学突触的生物结构原理图。在化学突触中，前神经元和后神经元之间形成的间隙通常比电突触要大许多，在生物学中又将该间隙称为突触间隙（synaptic cleft）。化学突触的信号传递是通过前神经元突触后膜向突触间隙释放神经递质（neurotransmitter）从而作用到后神经元的突触前膜上实现的，该过程涉及电信号转换为化学信号和化学信号再转化为电信号两个过程。当前神经元上有动作电位产生并通过轴突传递到突触前膜时，突触前膜上的门控离子通道会打开允许 Ca^{2+} 从细胞外扩散到突触末梢内部造成突触末梢内 Ca^{2+} 的浓度瞬间升高到一定的值。Ca^{2+} 浓度的升高使得突触小泡（synaptic vesicle）与突触前膜融合，突触小泡内的神经递质被释放到突触间隙中。然后神经递质在突触间隙内扩散并与突触后膜上的受体结合。神经递质与受体的结合会导致突触后膜上的离子通道打开（或关闭），从而改变离子（例如 Na^{+} 和 K^{+}）流入（或流出）后突触神经元的能力，导致后突触神经元动作电位的产生或抑制。这样，信息就从前神经元传递到了后神经元。与电突触不同，化学突触信息传递的效率可以根据神经元的活动进行动态调整，与细胞外 Ca^{2+} 的浓度和突触后膜上受体的形态和个数等都有关。正是因为化学突触这种传递效率可以动态调整的特性使得生物体能够不断适应外界环境，是生物体学习和记忆功能的基础。化学突触的传递效率可动态调整的这种行为又称为突触可塑性，根据这种可塑性维持的时间长短又可以分为长时程突触可塑性和短时程突触可塑性等。当前基于忆阻器实现突触仿生的工作基本上是对化学突触的这种可塑性机制的实现，在后面的小节中我们将化学突触统一称为突触并对相关可塑性机制和忆阻器实现展开介绍。

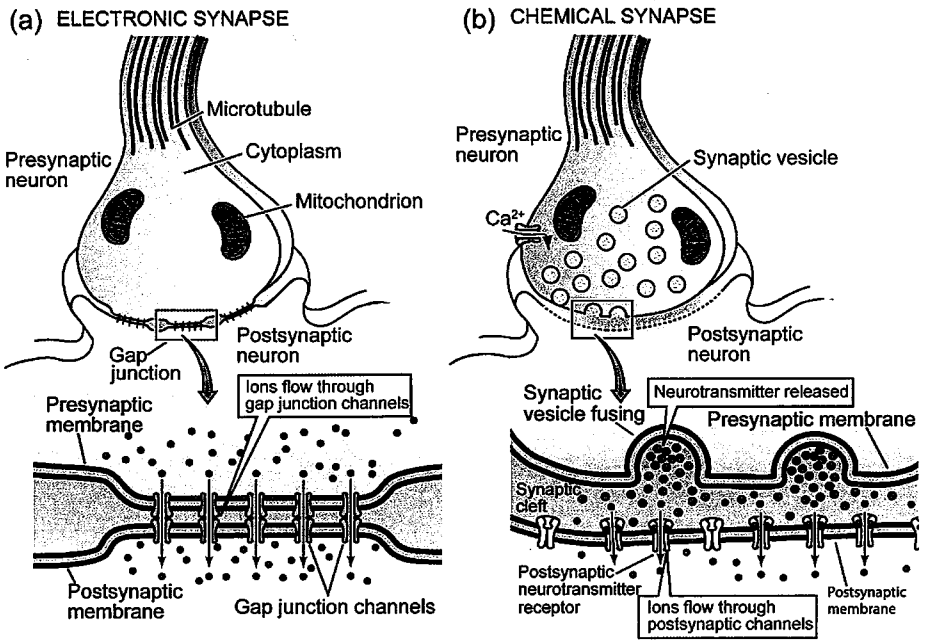


图 2.7 电突触和化学突触的生物结构原理图[46]

2.2.2 短时程可塑性

突触的短时程可塑性一般是由前突触内 Ca^{2+}的浓度和神经递质的数量等决定的，相比于长时程突触可塑性，突触的传递效率可维持的时间比较短（毫秒到分钟量级），在生物系统的计算过程中起着重要的作用[47]。根据突触传递效率改变的特性又可以分为短时程增强（short-term potentiation, STP），短时程抑制（short-term depression，STD）和强直后增强（post-tetanic potentiation, PTP）等[46]。图 2.8 给出了生物突触在不同的短时程可塑性机制下膜电位的响应原理图。对于短时程增强来讲，当前突触接收动作电位刺激时，门控 Ca^{2+}离子通道会打开，胞外的 Ca^{2+}内流使得突触末梢内的 Ca^{2+}浓度增加，前突触通过突触间隙向后突触释放神经递质，造成突触后膜电位的增加。前突触在接收到动作电位后 Ca^{2+}浓度的恢复比较慢，需要一个过程。如果前一个动作电位的刺激导致的 Ca^{2+}浓度的改变还没有恢复到正常状态之前又接收到一个动作电位信号的话，前突触内 Ca^{2+}的浓度会在剩余 Ca^{2+}的基础上继续增加，使得突触末梢内 Ca^{2+}的浓度比前一次刺激后的更高，造成更多神经递质的释放。这时，后突触的膜电位响应比第一次更大，对应 STP 现象。相同的，如果前突触继续接收动作电位刺激，在一定数量内，后突触膜电位的响应会继续增加。因而，生物突触的 STP 现象可以简单的概括为前突触内 Ca^{2+}浓度依赖的突触可塑性行为。另外，前突触内装有神经

递质的突触小泡在神经递质释放后也需要一个恢复过程。当前突触接收到一连串的动作电位刺激后，前突触内突触小泡的数量减少，再接收动作电位时没有足够的神经递质释放，会造成后突触响应信号逐渐降低，对应 STD 行为。因而，生物突触的 STD 现象可以简单的概括为前突触内神经递质含量依赖的可塑性行为。除了基本的 STP 和 STD 之外，当前突触接收到高频的动作电位刺激时，可造成前突触末梢内 Ca^{2+}浓度的持续升高，导致另一种形式的短时程突触可塑性，称为 PTP。通常 PTP 在开始时会有一定的延时，在高频动作电位刺激结束几分钟后会增强神经递质的释放。可以看到，与 STP 相比，PTP 的持续时间要长一些（几分钟）。因此，PTP 也可以看成是由前突触末梢内 Ca^{2+}浓度依赖的突触可塑性行为，可能包括前突触内蛋白激酶的激活，这些酶的激活增强了 Ca^{2+}进入突触小泡并与突触前膜融合的能力，从而增加了神经递质的释放量，使得后突触的响应增加。

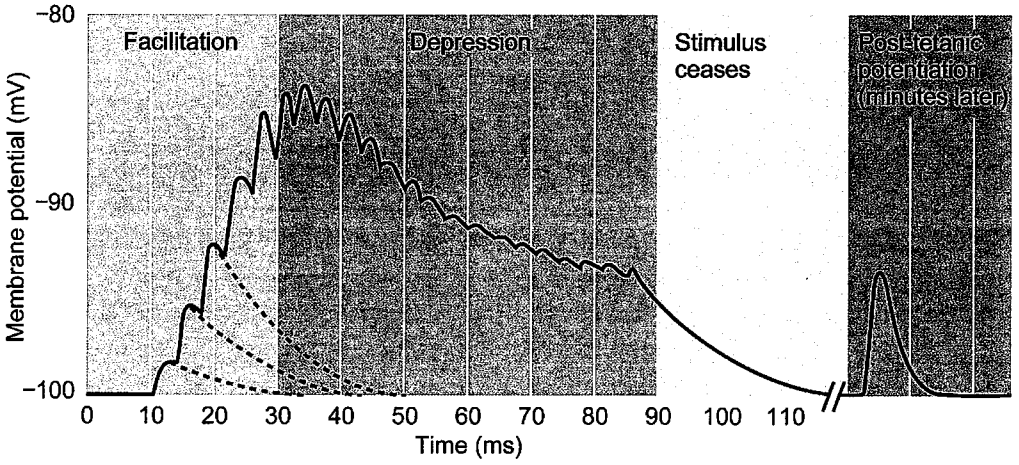


图 2.8 生物突触的短时程增强、短时程抑制和强直后增强动作电位响应原理图[46]

在忆阻器中，对突触短时程可塑性的模拟方案通常分为两种，一种是利用活性金属电极在电场作用下的氧化还原反应和电迁移过程实现的；一种是利用材料中缺陷（如氧空位等）的电迁移过程实现的。需要指出的是，在利用这两种方案进行突触短时程可塑性功能的模拟时，器件通常需要具备一定的易失特性（如前面我们所提到的易失性阈值转变忆阻器）。在进行突触的短时程可塑性机制模拟时，将活性金属离子（例如 Ag^{+}或 Cu^{2+}）或氧空位在介质层中的注入看成是生物突触中 Ca^{2+}向突触末梢内的扩散过程，因此具有非常高的突触离子动力学相似性，如图 2.9 所示。

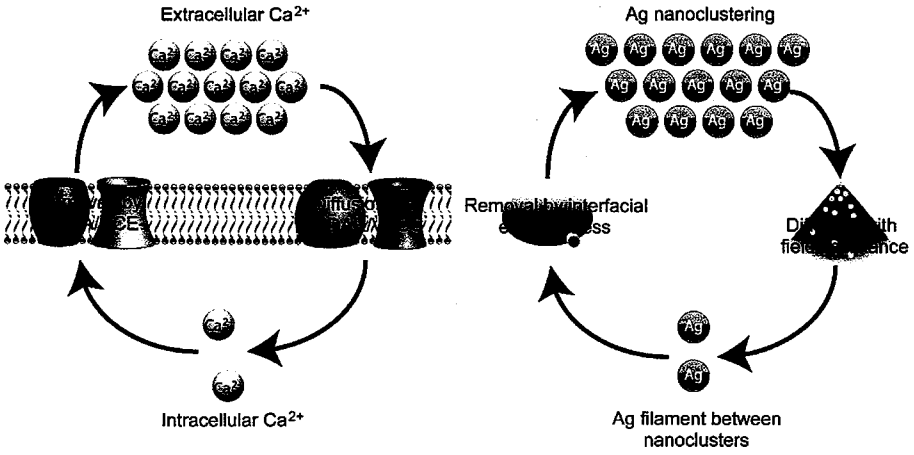


图 2.9 生物突触中 Ca^{2+}的流入流出和忆阻器中活性离子的运动过程对应图[28]

目前，国内外广大学者已在该领域开展了大量的研究工作。2011 年，T. Ohno 等人通过精确调控 Ag/Ag_2S/Pt 原子开关中的量子电导实现了器件在低频刺激下的短时程可塑性突触行为，而且通过反复刺激实现了短时程记忆到长时程记忆的转变[48]。同年，密歇根大学的卢伟课题组利用 Pd/WO_x/W 器件中氧空位的迁移过程模拟了突触的短时程可塑性机制，也通过反复施加刺激实现了突触器件的短时程记忆到长时程记忆的转变[49]。然而，在这些工作中，只实现了突触 STP 行为的模拟，未报道对突触 STD 行为的实现。直到 2016 年，Z. Wang 等人采用掺杂工艺制备了 Au/SiO_xN_y:Ag/Au 忆阻器，将突触的 STP 行为与介电层内 Ag 原子的迁移对应并将 Ag 原子定向移动的耗尽过程与 STD 联系起来，同时实现了 STP 和 STD 突触行为[28]。在文章中，Z. Wang 等人将 STP 和 STD 具化为 PPF（pair-pulse facilitation，对脉冲易化）和 PPD（pair-pulse depression，对脉冲减弱）来进行描述，如图 2.10 所示。到目前为止，利用离子的动态效应实现突触短时程可塑性仿生的工作已经在大量的材料体系中得到验证，例如生物高聚物薄膜[50,51]，二维材料（h-BN）[52]、有机金属化合物[53]、金属氧化物[54]等。在这些工作中，Y. Park 等人基于 Au/lignin/ITO/PET 结构的忆阻器也进行了 PTP 突触短时程效应的仿生实现[50]。另外，我们在前面提到，生物突触的短时程可塑性机制在生物体的计算过程中起着重要作用，利用忆阻器突触的短时程效应实现计算功能的验证还需进一步探索。

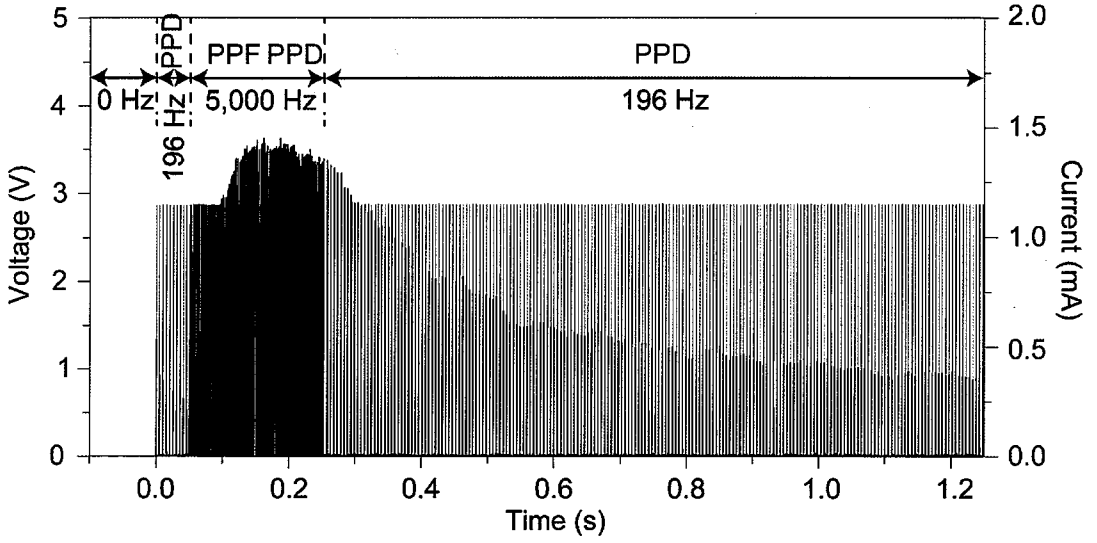


图 2.10 基于忆阻器的 PPF 和 PPD 短时程可塑性突触曲线[28]

2.2.3 长时程可塑性

如上所述，突触的短时程可塑性的效果通常持续的时间比较短（毫秒到分钟量级），可以在计算过程起作用。然而，这种可塑性机制并不能支持生物体的长时间记忆行为，这些记忆需要几周、几个月或者几年的时间长度。因此，如果要在一段很长的时间内都起作用，生物突触还必须支持一种长时间的可塑性机制，称为长时程可塑性。在长时程可塑性机制中又分为两种情况：一种是长时程增强可塑性（long-term potentiation，LTP），它指的是突触的信号传递效率（又称权重）在相关的神经元活动模式下增强并维持很长一段时间；另一种是长时程减弱可塑性（long-term depression，LTD），与 LTP 相反，它指的是突触的权重在相关刺激下减弱并长时间保持。生物突触的 LTP 和 LTD 行为最初是在生物脑的海马体结构中发现的，当突触接收到一个短时间的高频刺激后，突触权重的增强效果会持续几个小时，如图 2.11（a）所示。与之对应，当突触受到长时间低频刺激后（例如刺激频率 1 Hz，时长 10 分钟以上），突触的传递效率会受到抑制达几个小时，如图 2.11（b）所示。

图 2.12 给出了生物突触发生长时程增强的部分原因，机理解释如下：生物突触后膜的谷氨酸（NMDA）受体是一种双离子门控通道，当生物突触受到短时间的高频刺激时，后突触膜电位的极化会使得堵塞 NMDA 受体的 Mg^{2+}移走，然后来自前突触的谷氨酸神经递质与 NMDA 受体结合，Ca^{2+}通道打开。进而使得后突触中的 Ca^{2+}浓度增加，突触的传递效率增强。另外，Ca^{2+}浓度的增加会激活后突触中钙离子依赖的

蛋白激酶（CaMKII）和蛋白激酶 C（PKC），使得后突触中产生新的 AMPA 受体（一种谷氨酸受体），从而增加后突触中 AMPA 受体的数量，增加了后突触对神经递质的敏感度。而突触长时程减弱的部分原因则是长时间的低频刺激使得后突触内 Ca^{2+}浓度降低，后突触中的蛋白磷酸酶激活，将部分 AMPA 受体溶解，AMPA 受体的数量减少，后突触对神经递质的敏感度降低。因而，生物突触的长时程可塑性机制可以简单的概括为后突触内 Ca^{2+}浓度依赖和 AMPA 受体数目依赖的突触可塑性行为。

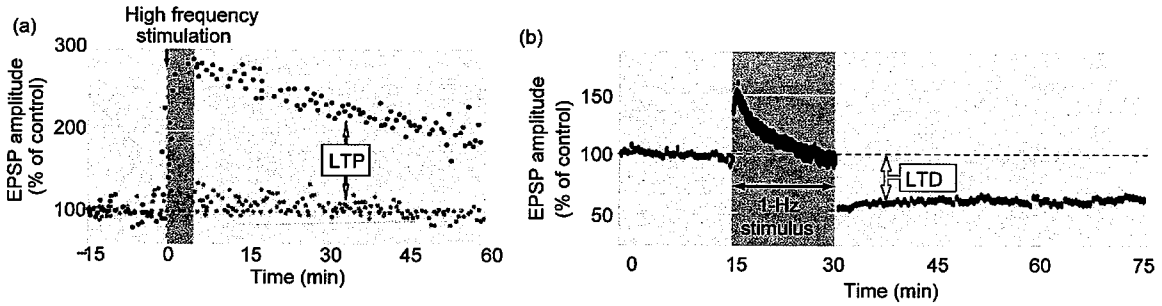


图 2.11 （a）生物突触在短时间高频刺激下的 LTP；（b）生物突触在长时间低频刺激下的 LTD[46]

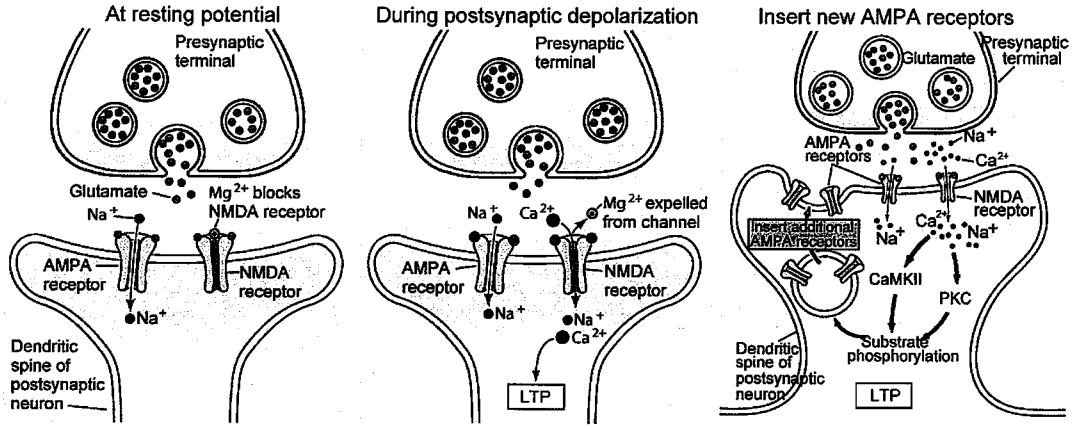


图 2.12 生物突触发生长时程增强可塑性的机理图，包括 NMDA 通道打开和新的 AMPA 受体产生[46]

与短时程可塑性的模拟不同，在用忆阻器实现突触的长时程可塑性模拟时需要器件具备一定的保持特性，也就是我们前面所提到的非易失性忆阻器。在进行长时程突触可塑性的实现时，忆阻器除了保持特性以外还要具备一定的多值特性。理想情况下，非易失的缓变忆阻器在实现上效果最佳。2010 年，密歇根大学的卢伟课题组在国际

上首次利用 W/Si/Si:Ag/Cr/Pt 结构的忆阻器对生物突触的相关长时程可塑性进行模拟[55]，如图 2.13（a）所示。图 2.13（b）给出了该器件在三角波扫描下的 I-V 特性，可以看到良好的缓变特性，并且后一次的扫描曲线会沿着前一次扫描的曲线叠加，说明器件具有非易失特性。这是因为器件的 Si:Ag 层为低阻态，Si 层为高阻态，在正向电压扫描时，Ag 原子会向 Si 层移动使得 Ag 界面向下电极逐渐移动，器件的电导逐渐增加。器件在反向电压扫描时，Ag 界面的移动方向相反，器件的电导降低。因而，器件的这种电导的内在调节可以用来模拟生物突触的可塑性行为。给器件施加连续正向脉冲激励，可以得到流过器件的电流逐渐增加，对应生物突触的 LTP 行为。在正向脉冲刺激后，给器件施加反向的脉冲刺激，流过器件的电流逐渐降低，对应突触的 LTD 行为，如图 2.13（c）所示。该工作说明利用忆阻器可以对生物突触的相关长时程可塑性行为进行模拟，开辟了利用忆阻器模拟神经突触的先河。

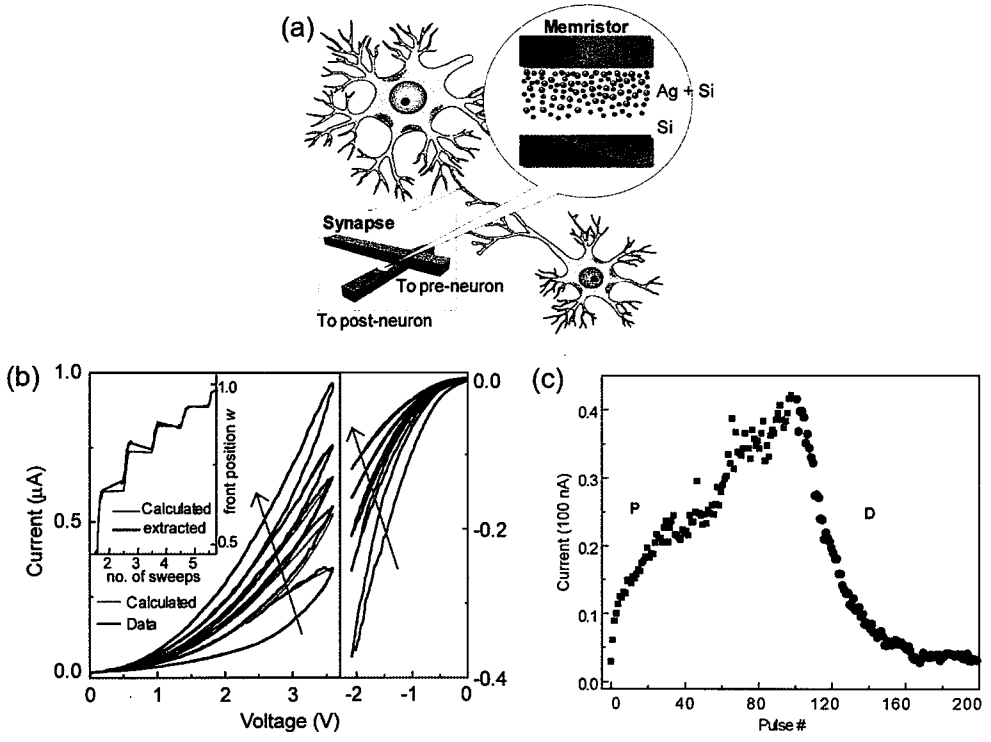


图 2.13 （a）生物突触和 W/Si/Si:Ag/Cr/Pt 忆阻器突触；（b）器件的直流特性；（c）器件在脉冲操作下的 LTP 和 LTD 特性曲线[55]

对应生物神经元发放离散动作电位的特性，在利用忆阻器实现神经突触的长时程可塑性特性仿生时通常会利用连续的脉冲对器件的电导进行连续调制，在每一个编程

脉冲前后都跟随有一个读脉冲，保证可以将器件电导的变化完整读出，读脉冲对应的电流值或者电导值的变化对应相关工作中常见的 LTP/LTD 曲线，如图 2.14（a）所示。利用单个忆阻器进行突触仿生的大部分工作都是采用的这种编程方案，这也是器件可以实现突触应用的电学特性基础。然而，对于大多数器件来说，采用这种编程方案得到的 LTP 和 LTD 曲线通常表现出非线性和非对称性，这会导致器件在用于系统的在线训练时不能通过简单的脉冲编程得到想要的权重改变量，从而影响系统的训练效果。为了优化 LTP/LTD 曲线的线性度和对称度，从脉冲编程的角度提出了两种优化方案，一种是变脉幅编程；另一种是变脉宽编程。所谓变脉幅编程便是在 LTP/LTD 验证过程中逐渐增加编程脉冲的幅度，从而一定程度上提高 LTP/LTD 曲线的线性度和对称度，如图 2.14（b）所示。图 2.14（c）给出了相应的变脉宽编程方案，在对器件电导连续调制过程中，固定脉幅逐渐增加脉冲的宽度。由于变脉幅和变脉宽的方案需要额外的引入相关的脉冲调制电路，因而在一定程度上会增加外围读写电路的复杂度。此外，从器件结构设计出发，增加保热层[56]或者离子限制层[57,58]也在一定程度上可以提高 LTP/LTD 曲线的线性度和对称度。由于突触器件的长时程可塑性是神经形态计算机实现学习和记忆的基础，近几年来国内外学者开展了广泛的研究工作，已经成为忆阻器应用的最重要方向之一。

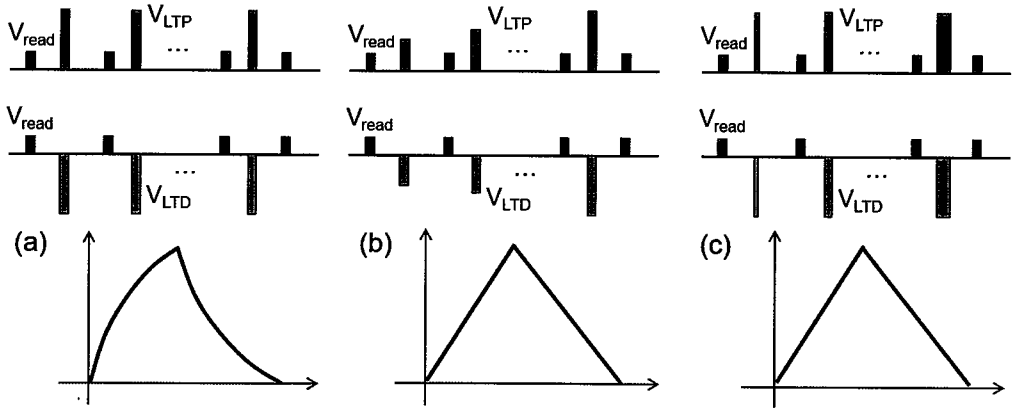


图 2.14 （a）固定脉冲的 LTP/LTD 实现方案及曲线；（b）变脉幅的编程方案及 LTP/LTD 曲线；（c）变脉宽的编程方案及 LTP/LTD 曲线

2.2.4 突触学习规则

在生物系统中，实现对新事物认知和学习过程通常需要遵循一定的规则。简单来讲就是突触实现长时程可塑性行为时，前突触和后突触神经元动作电位发放之间的关

系。根据这种关系，在生物学上又分为时序依赖可塑性（spike-timing-dependent plasticity，STDP）和频率依赖可塑性（spike-rate-dependent plasticity，SRDP）。下面我们将分别对这两种学习规则的生物基础和器件实现进行介绍。

2.2.4.1 时序依赖可塑性

前面我们讲到当前突触受到短时间的高频刺激时会发生 LTP 行为，短时间的低频刺激并不能使得后突触阻塞 NMDA 受体的 Mg^{2+}移开，从而不会发生 LTP 现象。然而，如果前突触接收到低频刺激的同时，后突触也接收到了刺激信号，那么突触后膜的膜电位会发生极化将 Mg^{2+}移开，这时后突触就会和前突触释放的谷氨酸神经递质发生联合作用使得 Ca^{2+}通道打开，突触的强度增强，这就是所谓的赫布学习规则（Hebbian learning rule）[59]。研究进一步发现，前突触接收刺激信号的时间和后突触接收信号的时间之差会导致突触的权重朝着不同的方向改变[60]。当前突触接收刺激信号的时间提前于后突触接收信号的时间时，会造成突触的长时程增强，如图 2.15（a）所示。相反，如果后突触接收信号的时间先于前突触接收信号的时间时，突触的权值会发生长时程减弱（图 2.15（b））。并且突触权重的改变量还和突触前后的时间差量有关，时间差越小，权重的改变量越大，相反，则越小。人们根据这样的特性得到了突触权重的改变量和突触前后刺激时间差之间的关系，这就是当前普遍使用的 STDP 学习规则，如图 2.15（c）所示。因此，时序依赖可塑性可以描述为前后动作电位时间依赖的一种突触可塑性机制。

当前，在利用忆阻器实现生物突触的 STDP 学习规则时，主要有两种方案：一种是前后脉冲重合叠加的方案（overlap），另一种是前后脉冲不重合的方案（non-overlap）。图 2.16 分别给出了两种编程方案下的脉冲形式，器件的一个电极作为突触前膜接收来自前神经元的刺激信号，另一个电极作为突触后膜接收来自后神经元的动作电位信号。对于 overlap 方案来说，加在器件上的刺激信号通常具有正负电压值，单个脉冲的电压值不会改变器件的阻态，当前后脉冲叠加时会在器件上产生较大的有效电压幅度，从而使得器件的阻态发生改变（图 2.16（a））。另外，根据电压的波形设计，前后脉冲之间的间隔不同产生的有效电压幅度不同导致对器件电导的调整量不同，从而得到与前后脉冲时序相关的器件电导调制。而 non-overlap 的实现方案不需要前后脉

冲的叠加（图 2.16（b）），在这种实现方式中引入了器件的内在动态特性，一般在二阶或多阶忆阻器中实现。

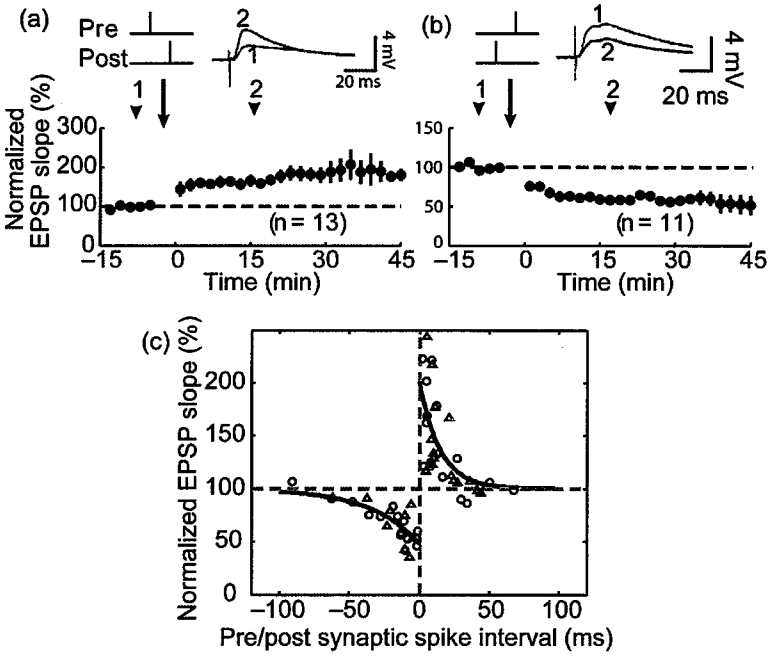


图 2.15 （a）前突触脉冲先于后突触脉冲产生 LTP；（b）后突触脉冲先于前突触脉冲产生 LTD；（c）生物突触的 STDP 曲线[60]

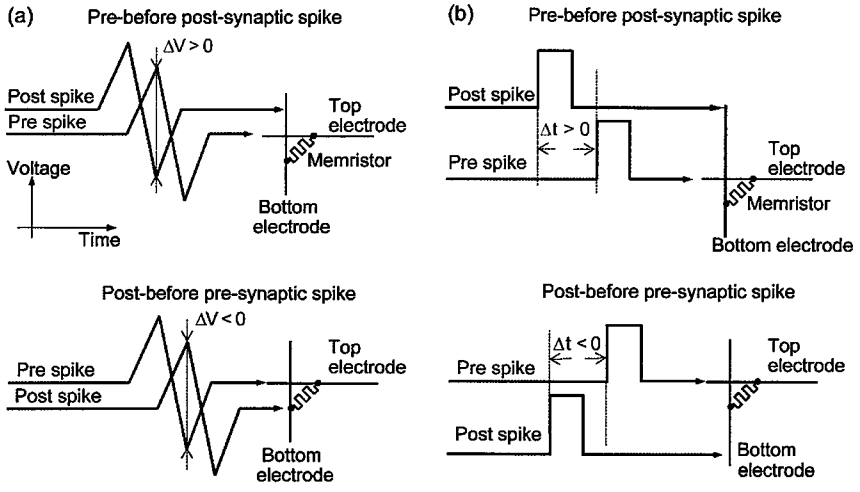


图 2.16 （a）忆阻器实现 STDP 的 overlap 编程方案[61]；（b）忆阻器实现 STDP 的 non-overlap 编程方案

基于以上两种编程方案，近几年来，国内外研究人员开展了大量的工作。例如，

S. Yu 等人[62]、Y. Li 等人[63]以及 M. Prezioso 等人[64]先后采用 overlap 的方式在不同的忆阻器器件中实现了 STDP。图 2.17 给出了 Y. Li 等人的工作中所用到的脉冲形式及对应的正 STDP 曲线。另外，Y. Li 等人以及 M. Prezioso 等人还通过设计脉冲波形在器件中实现了多种形式的 STDP 曲线，进一步丰富了忆阻器在实现相关突触学习规则的可能性。

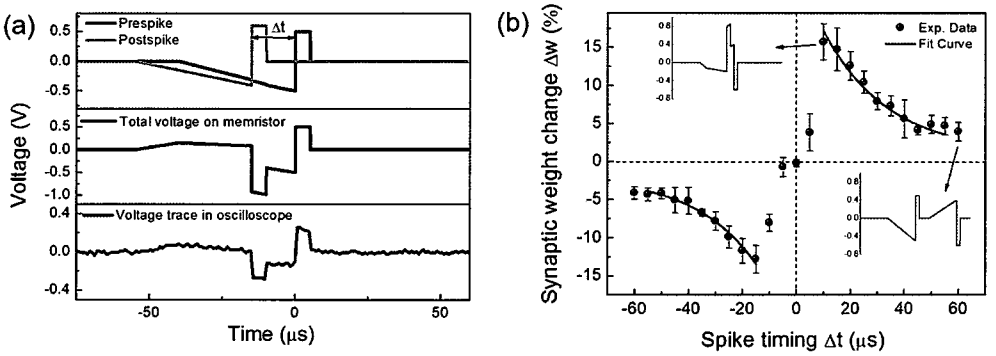


图 2.17 （a）测试 STDP 的前后脉冲形式；（b）忆阻器实现的 STDP 曲线[63]

相比于 overlap 的形式，non-overlap 编程方案具有更高的仿生度。利用 non-overlap 实现 STDP 的方案又可以分为两种：一种是利用二阶（或高阶）忆阻器中的动态特性，例如 C. Du 等人的工作[65]（图 2.18（a））；另一种是将易失性阈值转变忆阻器和非易失性忆阻器串联，其中阈值转变忆阻器作为时间控制单元，如 Z. Wang 等人的工作[28]（图 2.18（b））。

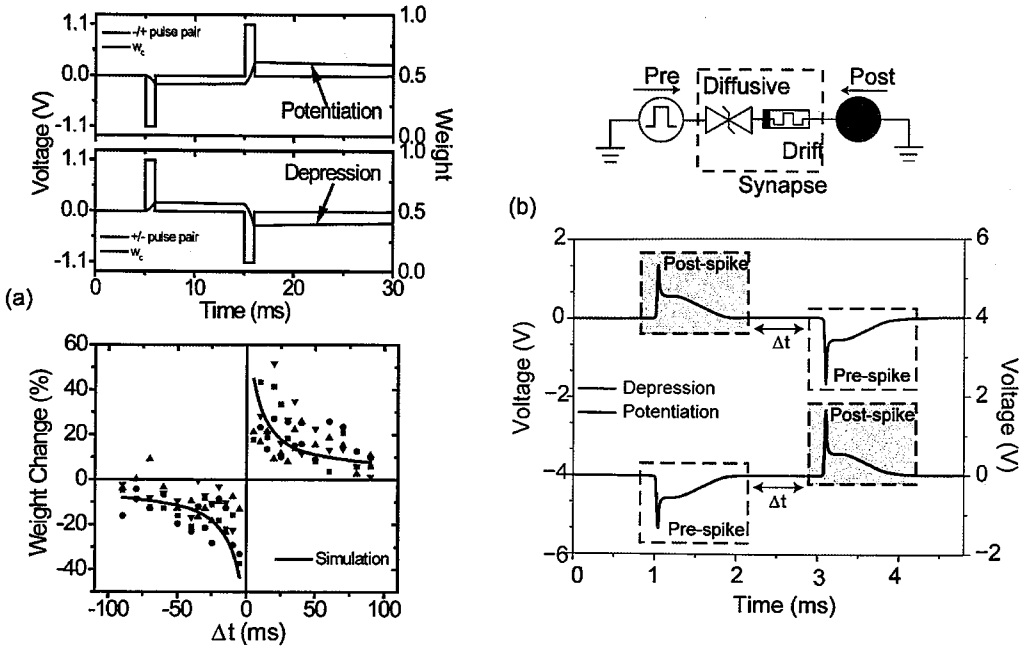


图 2.18 （a）基于二阶忆阻器的 non-overlap 形式的 STDP[65]；（b）易失型忆阻器作为时间控制单元的 non-overlap STDP 实现形式[28]

由于上述 STDP 是由前后神经元脉冲对的时间差决定的，因而又称为对脉冲 STDP（pair-STDP）。除了这种形式以外，还有三脉冲 STDP（triple-STDP）的形式[60]，也在忆阻器基的突触工作中有相关报道[66, 67]。

2.2.4.2 频率依赖可塑性

频率依赖可塑性（SRDP）是另外一种常用的突触可塑性学习规则，它描述的是突触前后动作电位发放频率之间的关系对突触权重进行调整的一种规则。BCM（Bienenstock, Cooper, and Muntro）理论是 SRDP 学习规则的典型代表[67, 68]，BCM 理论提出了 LTP 或 LTD 的阈值漂移现象，并指出突触权重可以根据后神经元动作电位发放的平均频率进行动态调整，具有历史依赖特性，如图 2.19（a）所示。根据 BCM 模型，当突触前神经元激发时，如果突触后神经元发放高频动作电位信号或者或后突触内部 Ca^{2+}浓度比较高，则突触倾向于实现 LTP。如果突触后神经元处于低频放电状态或者后突触内的 Ca^{2+}浓度比较低，则突触倾向于发生 LTD 过程。

在忆阻器中，SRDP 学习规则的实现通常通过给予器件不同频率的脉冲刺激，观测器件在不同频率刺激下的动态响应，如图 2.19（b）所示。可以看到，器件一开始接收到高频刺激（200 Hz）的情况下，响应电流增加，对应 LTP 过程。在高频刺激之

后施加低频刺激（10 Hz 或 1 Hz），器件的响应电流降低，对应 LTD 行为。然而，在 1 Hz 的低频刺激后，再给器件施加 10 Hz 的低频刺激，器件的响应电流又会相应增加，发生 LTP 行为。因此，突触器件的响应具有历史依赖特性。图 2.19（c）给出了不同历史刺激频率下，不同的后刺激频率对器件电导改变量的影响。可以看到，当历史刺激频率比较低时，LTD 和 LTP 行为的转折频率比较低。随着历史刺激频率的增加，转折频率逐渐变大，发生阈值漂移的现象。SRDP 学习规则已在多个课题组的工作中得到验证，例如华中科技大学缪向水课题组[63]和郭新课题组[69]以及清华大学潘峰课题组[70]等。最近，基于忆阻器实现 BCM 学习规则又有了最新进展，Z. Wang 等人利用 $Pt/WO_{3-x}/W$ 结构的忆阻器验证了 BCM 学习规则并构建两层神经网络仿真实现方向选择性功能[67]。

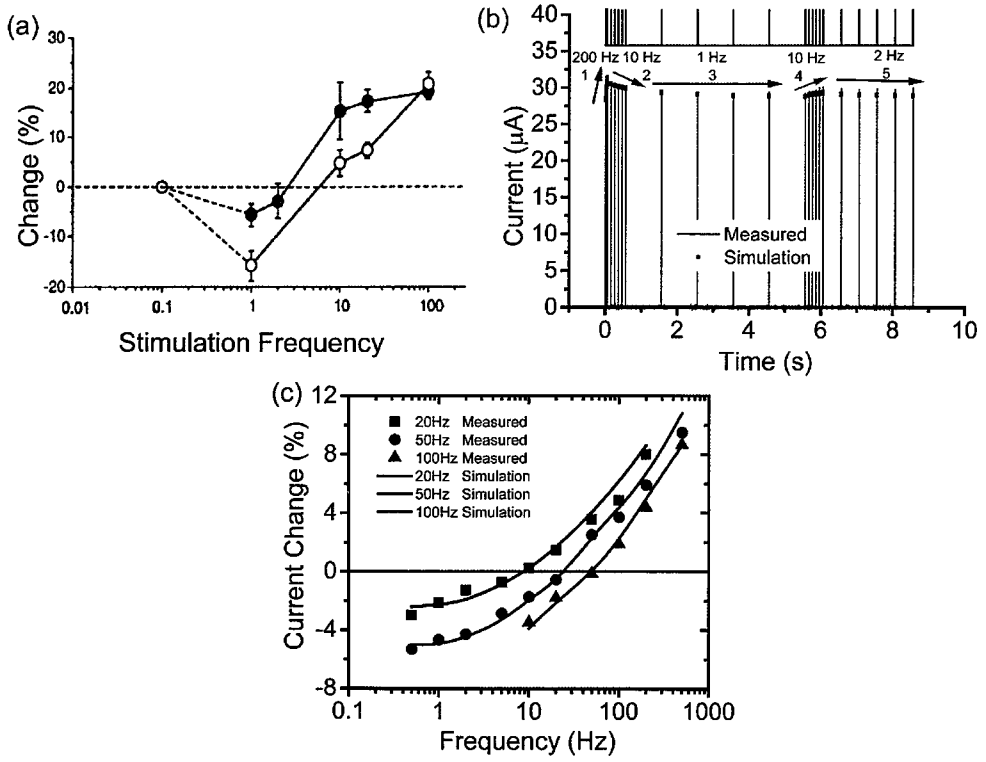


图 2.19 （a）生物突触可塑性的阈值漂移现象[68]；（b）忆阻器在不同频率刺激下的电流响应及相关的历史依赖特性；（c）忆阻器突触器件的可塑性阈值漂移特性[65]

2.3 忆阻器基神经元电路

在生物系统中实现相关功能时，神经元和突触是密不可分的，需要两者的相互配合。神经元接收整合来自突触前神经元的动作电位信号并在达到阈值条件时产生动作

电位向其他神经元传输。生物体中信号以离散的动作电位进行传输是实现高效计算的重要原因。简单来讲，神经元的核心功能是对输入的信号进行积分并产生动作电位信号。在前面小节中我们论述了突触的相关功能和忆阻器实现，本小节将从生物神经元的工作机制出发，论述当前的神经元模型并探讨忆阻器基神经元电路的实现。

2.3.1 生物神经元和模型

2.3.1.1 动作电位的产生机制

如图 2.20 所示，生物神经元中动作电位的产生过程可以分为以下五个步骤：

（1） 当神经元处于静息状态时，细胞膜上大多数的 Na^+和 K^+离子通道处于关闭状态，Na^+和 K^+不能流入流出细胞膜，神经元的膜电位保持基本不变。

（2） 当接收到刺激信号时，膜电位提升，细胞膜发生去极化，Na^+离子通道首先打开。然后细胞膜外的 Na^+在浓度梯度的作用下内流，使得细胞膜进一步去极化，更多的 Na^+离子通道打开，Na^+离子大量内流。

（3） Na^+离子的大量内流使得膜电位增加，一旦达到阈值，膜电位迅速接近膜外 Na^+离子的电势。动作电位的这个阶段叫做上升阶段。

（4） 门控 Na^+离子通道在打开不久后失活，阻止 Na^+继续流入。并且在上升阶段，大多数电压门控 K^+离子通道打开，导致膜内 K^+快速流出，膜电位迅速降低。这个阶段叫做下降阶段。

（5） 在动作电位的最后一个阶段，称为过充或超极化，在该阶段膜电位会比静息状态的膜电位更低。最终，K^+离子通道关闭，在离子泵的作用下，逐渐恢复膜内 Na^+和 K^+的浓度，膜电位恢复到静息状态。

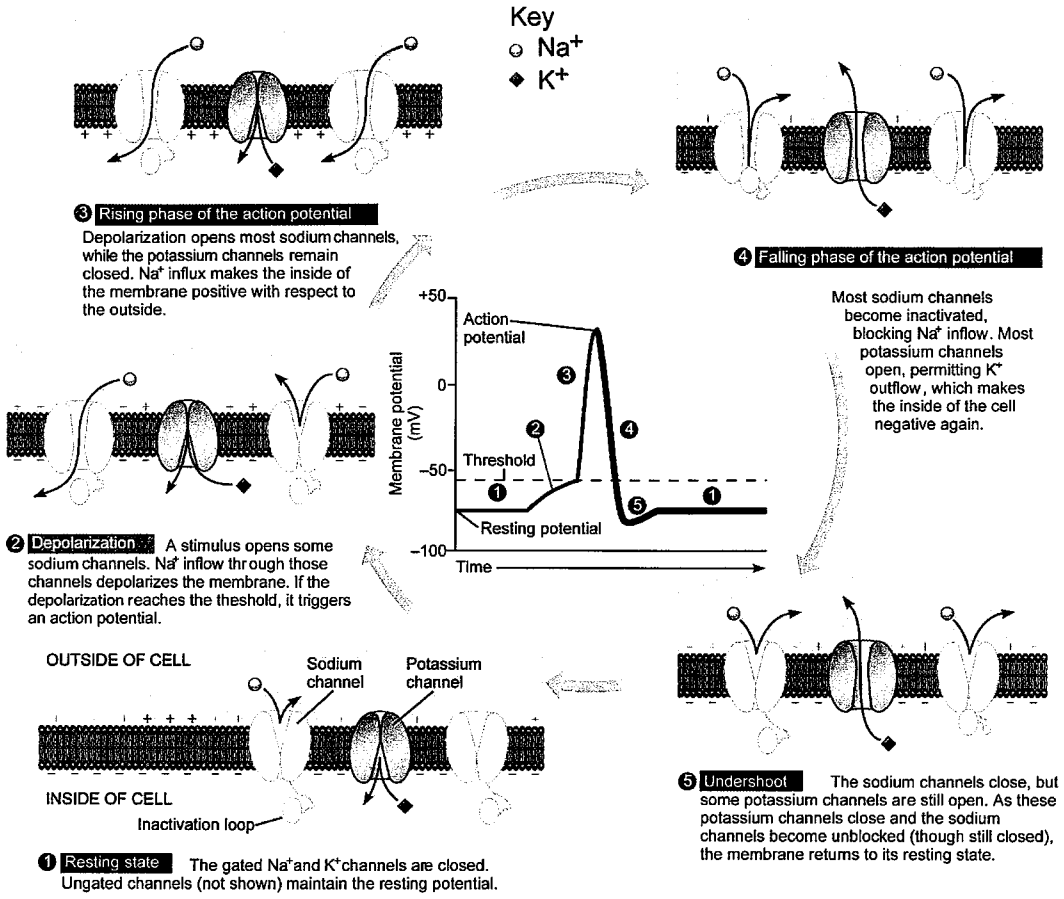


图 2.20 动作电位的产生过程[71]

2.3.1.2 神经元模型

为了在电路上或者数学上描述神经元产生动作电位的过程，人们相继提出了不同的神经元模型。图 2.21 给出了不同神经元模型在计算复杂度和生物可信度之间的对比。这些模型包括：现象学模型，其目标是使用简单的数学抽象（例如，漏电积分激发（LIF）模型[72]）捕捉神经元的输入-输出行为；生物物理模型，其目标是模拟神经膜的电生理状态（例如 Hodgkin-Huxley（H-H）模型[73]）。可以看到不同模型的计算复杂度和生物可信度之间有一个权衡，Izhikevich 神经元模型是为解决这两者之间的矛盾提出来的一种数学模型[74]。考虑到电路实现和算法应用，LIF（又称 integrate-and-fire with adaption）神经元模型和 H-H 模型为常用的两个模型，下面将主要针对这两个神经元模型进行介绍。

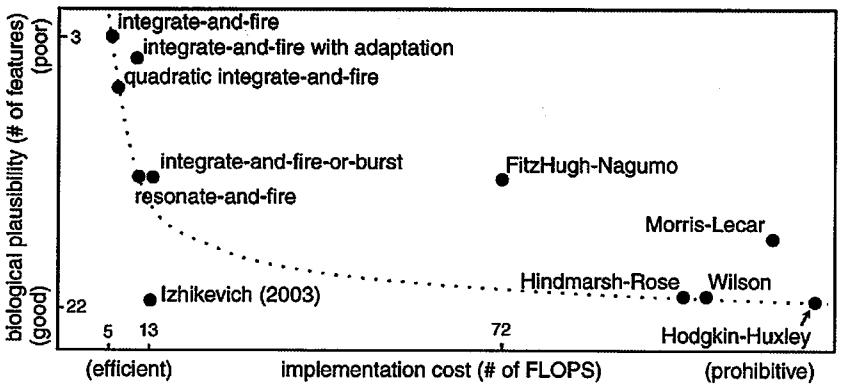


图 2.21 不同神经元模型的计算复杂度和生物可信度比较[74]

图 2.22（a）给出了 H-H 神经元的电路模型原理图，该模型是一种生物物理模型，是由 A. L. Hodgkin 和 A. F. Huxley 两人于 19 世纪 50 年代提出的。在该模型中分别用两个可变电阻（R_{Na} 和 R_K）表示生物神经元的 Na^+和 K^+通道，电容器表示生物膜，另外用一个固定电阻 R_L 表示生物膜的漏电通道。其中 R_{Na} 和 R_K 的开启电压不同，该神经元电路工作时对电容器充电使得膜内电位上升，R_{Na} 先打开，R_K 后打开，根据两个通道打开顺序的不同使得膜内电位变化，产生动作电位。H-H 模型可以实现生物神经元的 23 种发放模式。相比于 H-H 模型，LIF 神经元模型相对来说比较简单，如图 2.22（b）所示。该模型是一种现象模型，同样使用电容器作为生物膜对输入信号进行积分，R_L 作为漏电回路的电阻，阈值开关（或者可变电阻）作为离子通道。LIF 神经元工作时，对电容器充电同时伴随着 R_L 的漏电，当电容器上的电压达到一定值（膜电位的阈值）时，阈值开关打开，膜电位降低至 u_{rest}（静息膜电位）。该神经元电路并不能实现生物神经元的 23 种模式，但由于其数学公式比较简单，计算量比较小，因而是脉冲神经网络算法验证中常用的神经元模型。

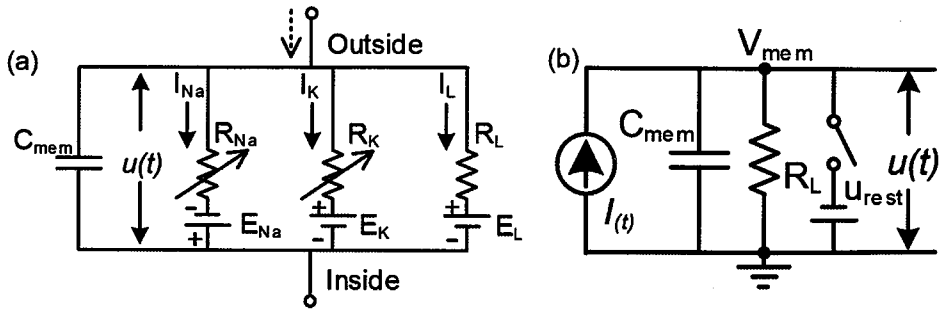


图 2.22 （a）H-H 神经元的电路模型；（b）LIF 神经元的电路模型

2.3.2 忆阻器基神经元电路实现

利用忆阻器实现的神经元电路主要是基于上述两种神经元模型，根据神经元电路是否使用电容器完成积分过程又可以分为有电容神经元电路和无电容神经元电路，下面我们将从这两个方面对当前忆阻器基神经元电路进行概述。

2.3.2.1 有电容神经元电路

所谓有电容神经元电路指的是忆阻器神经元电路使用了电容器完成积分功能，这与 H-H 神经元模型和 LIF 神经元模型相对应。H-H 神经元电路的实现通常要求器件具有易失性阈值转变特性，如我们在本章 2.1 节中谈到的两种易失性忆阻器。目前，基于忆阻器实现 H-H 神经元电路的工作主要有两个，一个是 2013 年 HP 实验室的工作[75]，另一个是 2018 年 HRL 实验室的工作[76]。在 HP 实验室的工作中，M. D. Pickett 等人利用两个 NbO_2 忆阻器分别作为 Na^+离子通道和 K^+离子通道并对原始的 H-H 模型进行了简单调整（图 2.23（a）），然后辅助以简单的电阻耦合实现了神经元的阈值放电、全或无动作电位、无衰减传输、不应期以及周期放电、震颤放电和快速放电神经元行为。图 2.23（b）给出了所用到的 NbO_2 器件的 I-V 曲线，可以看到易失性的阈值转变特性。基于该忆阻器实现的神经元的全或无阈值放电特性如图 2.23（c）所示，可以看到，当神经元接收比较小的刺激时（0.2 V），电容器 C1 上的电压不足以使 M1 打开，没有完整的动作电位信号产生。当输入刺激足够大时（0.3 V），神经元电路输出动作电位，完成一次放电。据我们所知，该工作是利用忆阻器实现脉冲神经元电路的开山之作。为了进一步实现更多的神经元放电模式，KRL 实验的 W. Yi 等人利用 VO_2 忆阻器作为离子沟道，并调整了电路的实现方式，模拟了生物神经元中的 23 种放电模式，该工作为利用忆阻器基神经元构建紧凑的人工大脑实现了可能。

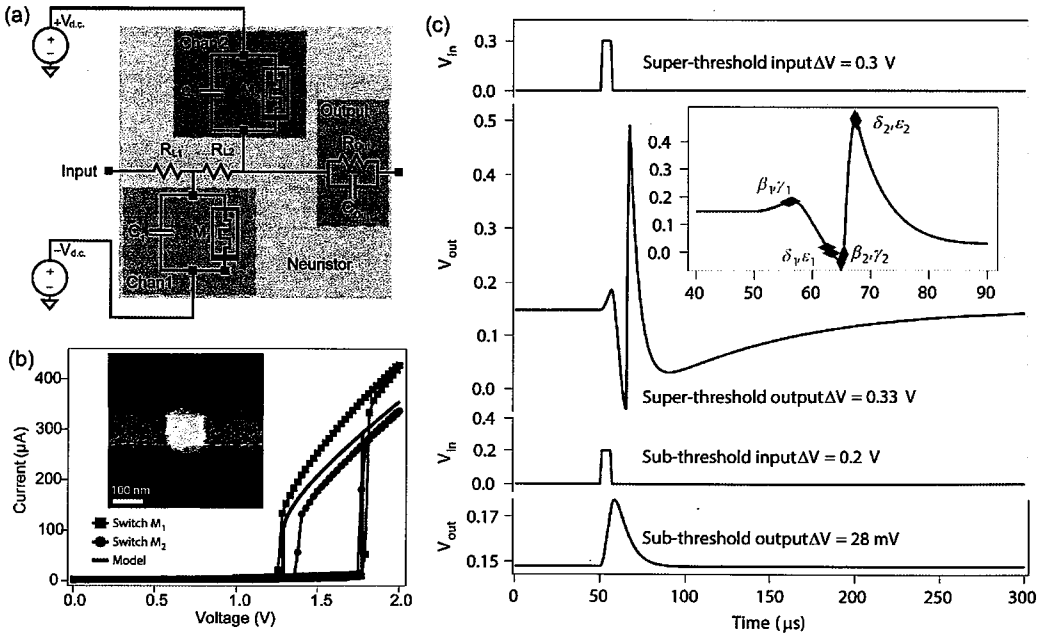


图 2.23 （a）基于 NbO_2 忆阻器的 H-H 神经元电路；（b）NbO_2 忆阻器的 I-V 曲线；（c）H-H 神经元的全或无阈值放电特性[75]

除了实现 H-H 神经元电路以外，LIF 神经元电路的实现则相对比较简单，已报道了较多的研究工作。在 LIF 神经元电路中，电容器作为积分单元，忆阻器作为阈值开关同时作为漏电通道，如图 2.24（a）所示。2016 年，J. Lin 等人首次利用 VO_2 忆阻器实现了 LIF 神经元，实现了基本的漏电积分发射特性[77]。图 2.24（b）给出了所用 VO_2 器件的直流特性，可以看到也具有易失性阈值转变的特性。该神经元电路的工作原理如下，连续的电流脉冲输入对电容器进行充电，完成积分功能，电容器上的电压升高。当电容器上的电压达到器件的阈值时，器件转变为低阻态，然后电容器通过忆阻器回路进行放电，电容器上的电压迅速降低，完成放电过程，如图 2.24（c）所示。需要指出的是，在脉冲输入的过程中，电容器也会通过忆阻器回路漏电，因此，整个过程是漏电积分发射的过程。图 2.24（d）给出了神经元实现积分放电过程中忆阻器电阻的变化，可以看到，器件初始态为高阻态，当电容器上的电压达到阈值电压时器件变为低阻态，这时流过器件的电流迅速增加，形成电流尖峰信号（图 2.24（e））。在两串输入脉冲间隔内，由于电容器上的电压不足以保持器件的低阻态，器件自发回到高阻态，从而可以为下一次的放电做准备。基于这样的基本原理，L. Gao 等人[78]和 M. Jerry 等人[79]也相继利用易失性阈值转变忆阻器实现了 LIF 神经元电路的设计并进

行了系统功能的仿真。特别的是，M. Jerry 的工作中还用到了器件转变阈值的随机性，从而实现了随机玻尔兹曼机的验证。此外，还有许多基于不同结构的易失性阈值转变器件的研究工作报道了 LIF 神经元电路的相关特性。除了利用这种易失性阈值转变忆阻器以外，在 R. A. Cobley 等人[80]和 J. W. Jang 等人[81]的工作中利用非易失性忆阻器 set 过程中的阈值转变特性作为电阻开关并辅助以 reset 操作也实现了 LIF 神经元电路功能的验证。

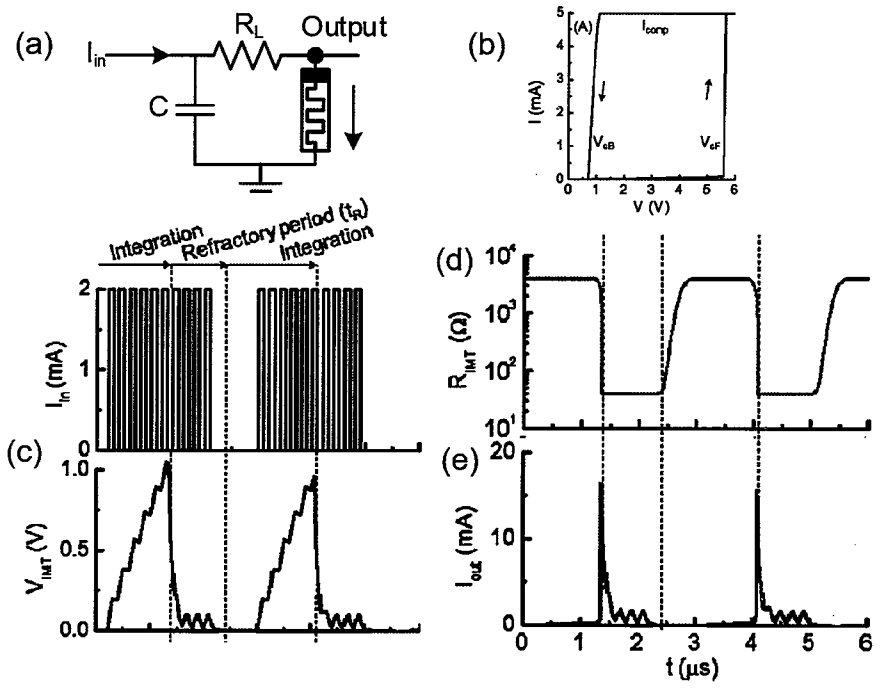


图 2.24 （a）基于 VO_2 忆阻器的 LIF 神经元电路；（b）VO_2 忆阻器的 I-V 曲线；（c）LIF 神经元的漏电积分发射特性；（d）忆阻器在放电周期内电阻的变化；（e）放电过程中流过器件电流的变化[77]

2.3.2.2 无电容神经元电路

如上所述，在有电容器的神经元电路中积分功能是由电容器实现的，忆阻器通常是作为阈值开关使用。前面我们提到忆阻器在实现生物突触功能仿生的过程中可以表现出连续脉冲刺激下电导的积累，该特性也可以看成是器件电导的积分过程。因此可以用来实现神经元中电容器的积分功能，构建无电容器的神经元电路，从而实现更加紧凑的神经元电路。2016 年，IBM 的 T. Tuma 等人[18]利用相变忆阻器实现了这样一个无电容的神经元电路。在该工作中，神经元的膜电位由相变器件的相结构表示。图 2.25（a）给出了设计的神经元电路的原理图，器件初始态为非晶态（高阻态），上电

极接收整合后的前神经元的输入信号，在该信号的刺激下器件的相结构逐渐晶化，器件电导逐渐增加，当电导增加到所设定的阈值时，触发脉冲发生模块（spike event generation）产生脉冲输出同时反馈给突触，并根据 STDP 学习规则调整突触的权值。当器件的电导达到阈值触发输出脉冲后，外围 reset 电路会将器件复位到初始的低电导状态以进行下一次的积分发射操作。图 2.25（b）给出了器件在连续脉冲操作下电导及对应相结构的变化，可以明显的看到器件电导在脉冲下的积分过程。我们知道，器件电导的变化快慢跟输入脉冲的强度（例如脉幅和脉宽）有关，因此不同的输入强度会导致不同的积分速度，从而导致神经元的放电频率不同，如图 2.25（c）所示。随着脉冲输入强度的增加，神经元的放电频率增加，这和生物神经元的强度依赖特性一致。此外，由于器件相结构的变化具有随机性，因而神经元电路发射的频率也具有随机性。T. Tuma 等人进一步利用该随机性验证了生物神经网络中的族群编码功能并实现了时序信号的检测。基于这样的工作原理，S. Lashkare 等人[82]和 J. J. Wang 等人[83]分别在 PCMO 基和 HfO_2 基的非易失性忆阻器中验证了神经元的积分发射功能。

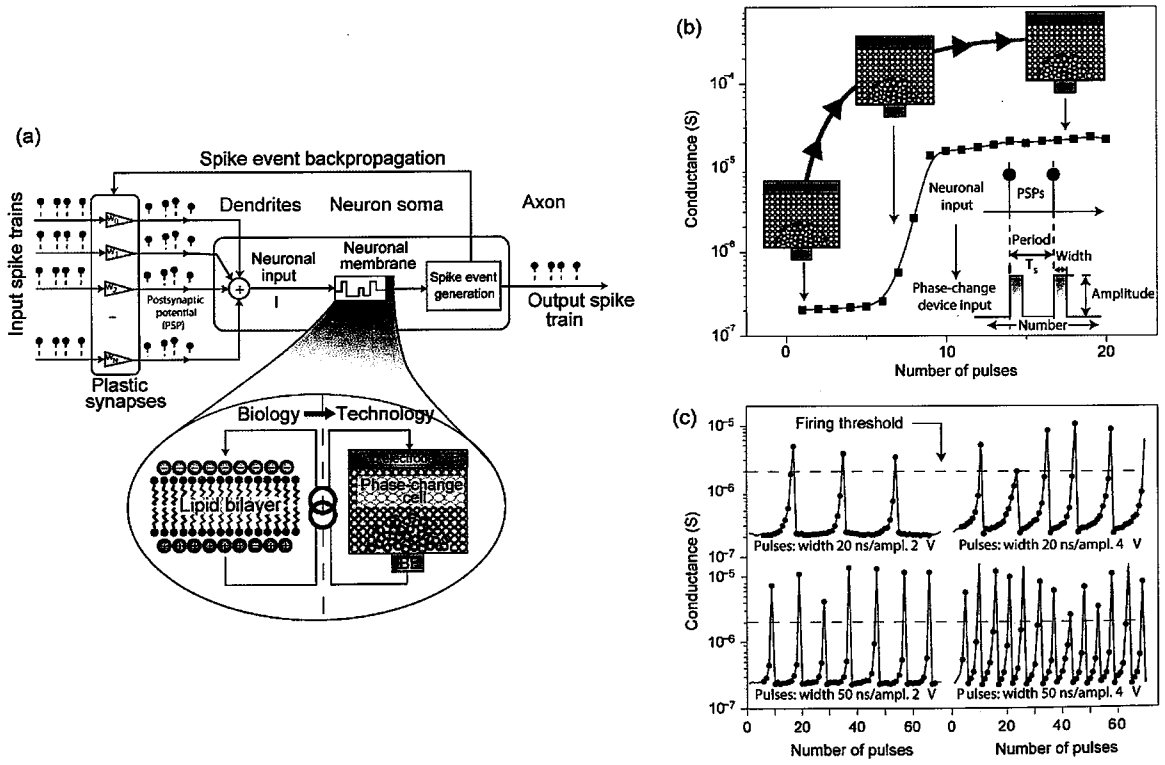


图 2.25 （a）相变忆阻器神经元电路原理图；（b）忆阻器在连续脉冲刺激下电导的积分特性；（c）相变神经元在不同输入脉冲强度下的放电特性[18]

在上述电路的实现过程中，由于器件具有非易失特性，通常需要额外的 reset 电路在每一次放电后将器件复位，这在一定程度上增加了外围电路的复杂度。考虑利用易失性的器件实现神经元电路的设计是一种有效的解决方案。Z. Wang 等人利用 $Pt/SiO_xN_y:Ag/Pt$ 结构的易失性忆阻器进行了神经元电路的仿生工作[84]，如图 2.26（a）所示。在该工作中，由于器件具有阈值转变特性，不需要额外的阈值比较电路，另外器件的易失特性也省略了 reset 电路的设计。图 2.26（b）给出了器件在四组脉冲刺激输入下的积分发射特性，器件的初始态为高阻态，在每组脉冲输入的前几个脉冲下器件进行积分，响应电流很小，当积分脉冲的数目足够多时器件导通，响应电流突然增大，对应发射行为。可以看到，神经元每次放电结束一段时间后再次输入脉冲刺激又会进行新的积分发射过程，这说明器件自发恢复到了初始高阻态。另外在输入脉冲间隔内，器件的积分效果也会有一定程度的泄露，因此，实现的是漏电积分发射的神经元功能。图 2.26（c）给出了神经元器件在多次放电行为下需要的积分脉冲数目的统计分布，符合高斯随机性分布，这是由器件内离子积累的随机性导致的。基于该神经元，Z. Wang 等人进一步验证了全忆阻的脉冲神经网络并实现了非监督学习。

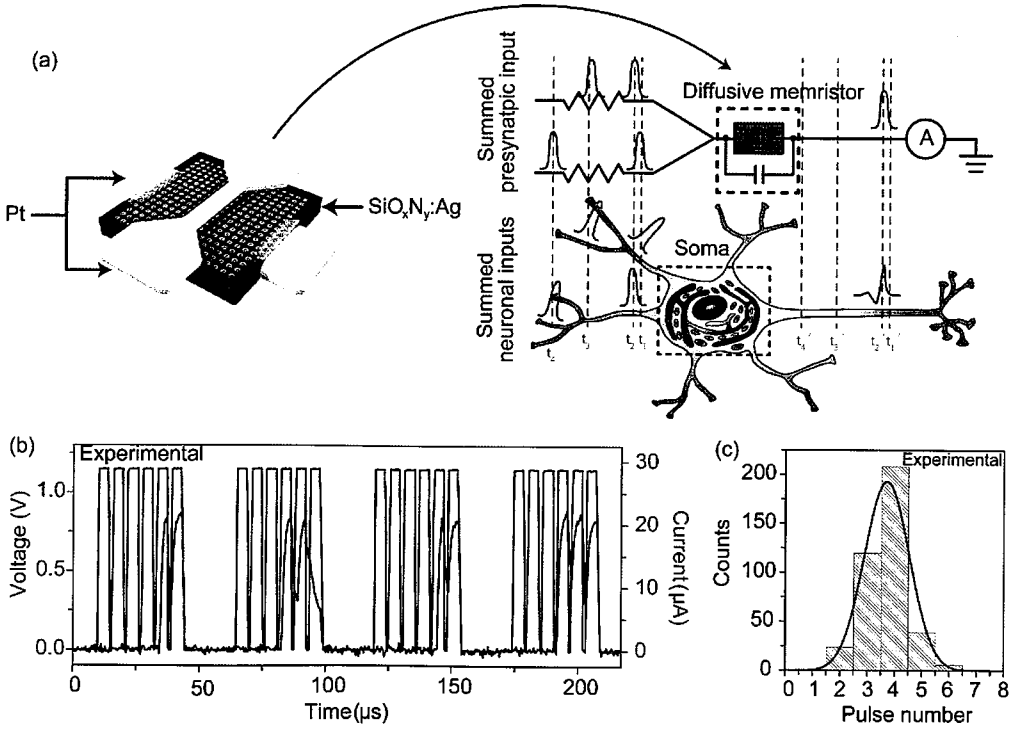


图 2.26 （a）易失性忆阻器神经元电路和生物神经元对应结构；（b）忆阻器在连续脉冲刺激下的积分放电特性；（c）触发放电事件需要的积分脉冲数目的统计结果[84]

除了以上工作外，基于铁电材料、磁材料等实现的无电容神经元电路也相继有报道，为了展示当前神经元工作的进展，我们在表 2.3 中给出了相关总结。值得一提的是，利用忆阻器实现神经元有利于构建紧凑的神经形态机器，然而该领域的研究刚刚起步，神经元的功能还比较单一，尚未达到系统级的应用，需要国内外学者的共同努力。

表 2.3 忆阻器基神经元工作统计

Type	Materials	Switching mode	Mechanisms	Circuit model	Capacitor	Ref
Redox	SiO_2; HfO_2; PCMO; WO_x/PEDOT:PSS	NVM	Channel growth process for integration	LIF & Quasi H-H	Without	[82, 83, 85, 86]
	SiO_xN_y:Ag; Ag/MoS_2; Ag/FeO_x	VM	Channel growth and self-rupture	LIF		[84, 87, 88]
Phase change	GST	NVM	Amorphous-crystalline phase change	IF		[18, 89]
	$GeNb_4Se_8$	VM	Mott switching	LIF		[90]
Ferroelectrics	HfSiO-FET	NVM	Polarization switching	IF		[91]
	HfZrO-FET	VM	Volatile polarization switching	LIF		[92, 93]
Magnetics	MgO/CoFeB/MgO	VM	Switching back magnetization under high current	IF		[94]
Redox	Ag/GeS_2; Cu/Ti/Al_2O_3;	NVM	Abrupt set switching	LIF	With	[81, 95]
	SiO_xN_y:Ag; Ag/MoS_2	VM	Channel growth and self-rupture	LIF		[84, 96]
Phase change	GST	NVM	Abrupt switching phase change	LIF		[80, 97]
	VO_2, NbO_2; B-Te	VM	Mott switching	H-H& LIF		[75-77, 98]

*NVM: non-volatile memristor VM: volatile memristor

2.4 忆阻器基神经网络

至此，我们概述了忆阻器在实现神经元和突触的相关工作，下面将进一步介绍如何利用忆阻器构建神经形态系统以及相关工作进展。在第一章中我们提到神经形态系统的两种算法模型分别是人工神经网络和脉冲神经网络，在这里，我们也将从这两种

算法模型出发进行工作概述。

2.4.1 忆阻器基人工神经网络

当前在利用忆阻器实现系统应用的工作中大多是人工神经网络（ANN）的形式，在 ANN 中，忆阻器阵列主要作为突触权重使用，忆阻器的电导代表权重，每个忆阻器存储一个权重值。图 2.27（a）给出了一个忆阻器基人工神经网络（ANN）的架构原理图，包括 M 行 N 列的忆阻器权重阵列，数模转换器（DAC）和模数转换器（ADC）以及写入电路（Write Circuit），可以实现正向传递、反向传播和权重更新操作。正向传递用于执行 ANN 推理过程，反向传播和权重更新用于执行训练操作。同一行中的忆阻器突触器件共享一个字线（WL），同一列中的突触器件共享一个位线（BL）。在这里忆阻器突触阵列支持三种主要操作：串行编程、并行向量矩阵乘法操作（VMM）和并行权重更新。所谓串行编程操作，即按行对忆阻器突触器件执行写入操作，其中，写入电路可以提供相应的电流或电压将器件编程到目标电导值。在进行推理操作时，VMM 是通过使用 DAC 同时驱动所有 WLs（BLs）和使用 ADC 感应流经每个 BL（WL）的电流来执行的。该过程只需要一步简单的读操作，因此可以实现大规模并行 VMM 操作。前向传递和反向回传过程中实现的 VMM 分别用 $I_{out}=V_{in}\times G$ 和 $I_{out}=V_{in}\times G^{T}$ 来表示，其中 V_{in} 表示输入电压，I_{out} 表示输出电流，G 表示突触器件电导。执行前向传递过程的 VMM 时，DAC 将数字量转换为模拟值输出相应的电压作为输入（V_{in}）作用到 WLs 上， BLs 上获得 VMM 输出电流（I_{out}）。随后，输出模拟电流通过 ADC 电路转换为数字量并通过激活函数（Out Act）得到最后输出。最后，通过沿着 BLs 和 WLs 驱动 DAC 来执行并行权重更新操作。此外，在忆阻器阵列中实现 VMM 操作时也有多种编码方案（包括电压、时间、位串行和随机编码），如图 2.27（b）所示。在电压和时间编码方案中，输入电压幅值（V）或脉冲宽度（T）与数字输入的大小成比例，该调制过程由相应的 DAC 电路完成。而位串行和随机编码方案则是通过将数字输入转换为固定时间宽度内的多个电压脉冲来实现的。在位串行和随机编码方案中，需要对每个电压脉冲执行 VMM 运算并对这些结果累加得到最后输出。需要指出的是，ANN 中的权重有正值和负值两种，然而，忆阻器的电导只有正值。为了在忆阻器突触中实现负值，通常具有两种方案[99]：一种移位方法；另一种是使用两个忆阻器

形成差分电阻对。在移位的方法中，忆阻器突触电导动态范围的一半用来表示正值的，另一半用来表示负值。而在第二种方法中两个忆阻器电导的差具有正值和负值，以此来表示正负权重，该方案通常需要成对输入或成对的输出。相比较而言，这两种方案各有优势；第二种方案具有更高的精度，而第一种方案具有更好的面积效率。

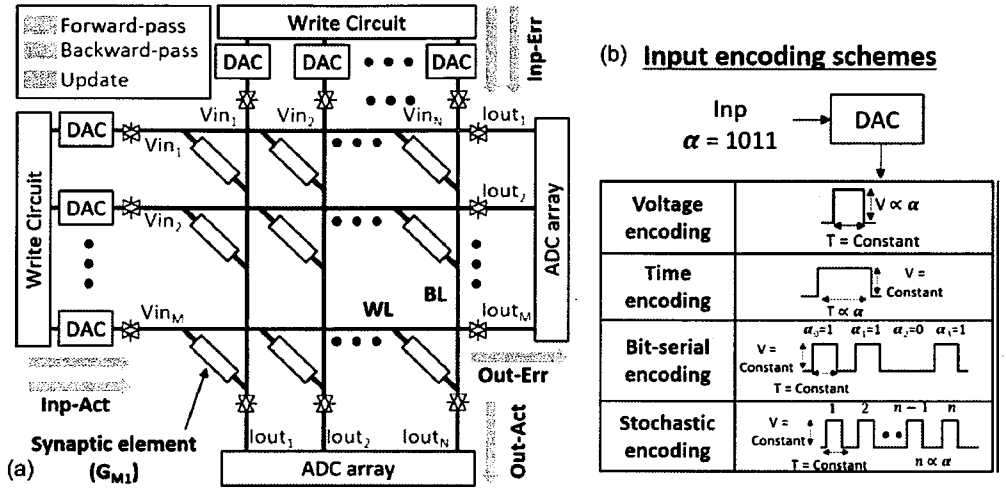


图 2.27 （a）忆阻器基人工神经网络硬件架构原理图；（b）ANN 中常用的编码方案[99]

基于上述工作原理，国内外的研究学者开展了广泛的研究工作。2013 年，加利福尼亚大学圣塔芭芭拉分校的 D. B. Strukov 课题组在国际上首次利用 Pt/TiO_{2-x}/Pt 忆阻器交叉阵列实现了单层感知机的功能，经过训练以后能够对字母“X”和“T”进行识别[100]。2015 年，该课题组又利用 Pt/TiO_{2-x}/Al_2O_3/Pt 忆阻器构建了 12×12 的交叉阵列（图 2.28（a））实现了感知机的功能[101]，并且可以进行非监督学习对字母“z”、“v”、“n”进行分类（图 2.28（b）和（c））。随后，基于忆阻器阵列实现 ANN 的工作相继报道，例如，清华大学吴华强课题组在 128×8 的 1T1R 阵列上实现了人脸识别[39]，密歇根大学卢伟课题组利用 WO_x 基忆阻器实现了稀疏编码[102]，马萨诸塞大学杨建华和夏强飞课题组利用 128×64 的 1T1R 忆阻器阵列（当时最大的模拟忆阻器阵列）实现了图像的压缩和解码[20]、双层神经网络的在线训练并对手写体字符进行识别[103]，长短时程记忆网络（LSTM）的构建[104]，强化学习的实现[105]等。上述工作中 ADC 和 DAC 电路通常是板卡级的实现，为进一步强化应用，近两年，密歇根大学卢伟课题组[43]、台湾国立清华大学张孟凡课题组[106]以及清华大学吴华强课题组[107]又先后推出了忆阻器突触阵列和相关 CMOS 外围电路的集成芯片，推动了该领域进一步的发展。

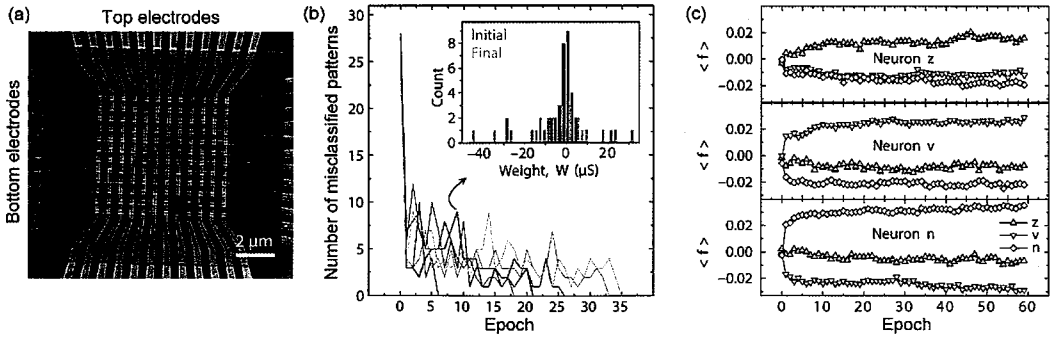


图 2.28 （a）12×12 忆阻器交叉阵列；（b）训练过程网络输出的迭代结果；（c）不同输出神经元在网络训练过程中对不同输入模式的输出结果[101]

从前面的介绍中，我们可以知道，ADC 和 DAC 电路在忆阻器基 ANN 的实现中起着关键的作用。然而，由于 ADC 和 DAC 电路的面积和能耗通常比较大，在一定程度上会降低忆阻器带来的高密度和低能耗优势。图 2.29（a）给出了忆阻器阵列和相关外围电位在构建系统时的面积和能耗分配示例，在这里忆阻器阵列规模为 64×64，ADC 和 DAC 精度分别为 10 比特和 6 比特。可以看到，在这种硬件参数下，ADC 的面积和能耗分别约占芯片总面积和总能耗的 60%和 88%。为降低系统中 ADC 的面积和能耗，ADC 的跨列共享是当前降低 ADC 功耗/面积的最流行方法，然而，这种方案会在一定程度上增加运算的延迟，如图 2.29（b）所示。因此，综合考虑系统的面积、功耗以及延时是实现忆阻器 ANN 硬件设计的关键。

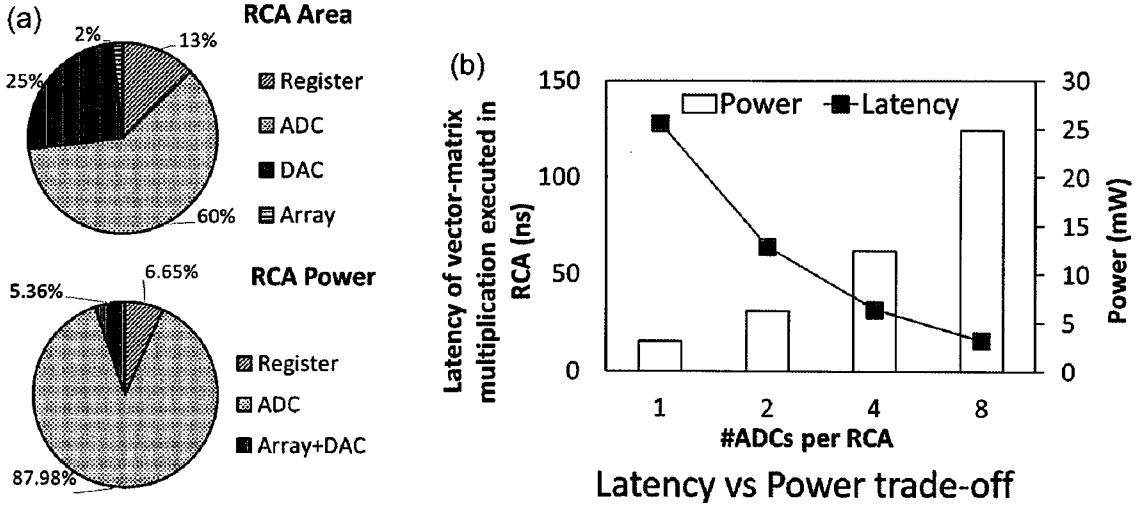


图 2.29 （a）忆阻器基 ANN 硬件中各单元的面积和能耗对比；（b）共享 ADC 方案中忆阻器 ANN 的能耗和延时折中对比[99]

2.4.2 忆阻器基脉冲神经网络

相对于 ANN，忆阻器基脉冲神经网络（SNN）是另一种系统实现方案。图 2.30 给出了一种忆阻器基 SNN 硬件实现的架构原理图，主要包括忆阻器阵列和 LIF 神经元电路，写入电路未在图中给出。在这里，神经元电路可以是 CMOS 基的神经元电路，也可以是忆阻器基的神经元电路。SNN 采用脉冲信号作为输入或者输出，避免了 ANN 中大量 ADC 和 DAC 电路的使用，因而理论上可以得到更低的能耗。在进行推理过程中，携带有编码信息的脉冲串施加到 WL 上，BL 上连接的 LIF 神经元对 VMM 电流进行积分操作并在达到神经元阈值时产生脉冲输出。在 SNN 中，根据算法的不同，输入信号又有不同的编码方式[108]，例如频率编码，时间编码、相位编码、群编码以及稀疏编码等。另外，在 SNN 的训练过程中，常采用 STDP 和 SRDP 等生物学习规则，这也是和 ANN 的不同之处。在这样的学习规则下，忆阻器突触的权重可以直接通过前后神经元动作电位之间的关系直接进行动态调整，因而可能不需要专门的写入电路，这就使得忆阻器基的 SNN 系统可以动态的适应外界环境，从而有利于构建更高效的神经形态系统。

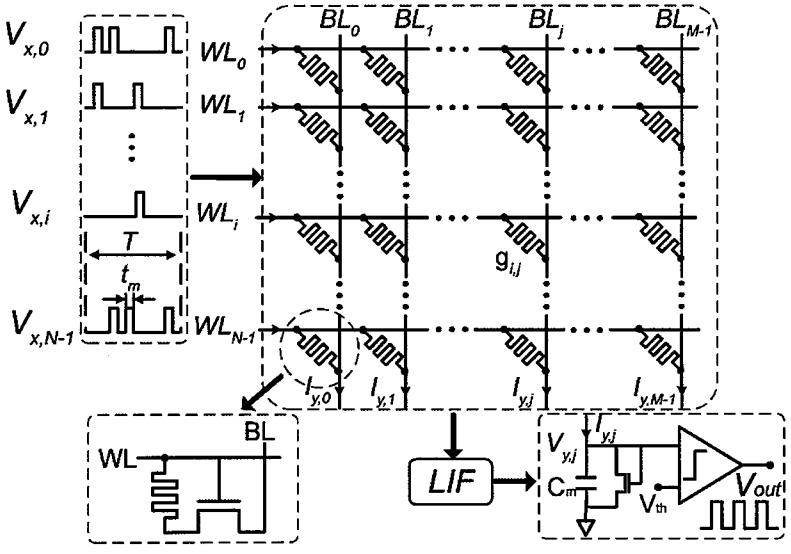


图 2.30 （a）忆阻器基脉冲神经网络硬件架构原理图[109]

鉴于忆阻器基 SNN 硬件构建存在的潜在优势，国内外的研究学者也相继开展了一些初步的研究工作。根据所用神经元电路不同，相关工作主要可以分为两种，一种是利用 CMOS 基脉冲神经元电路，另一种是利用忆阻器基脉冲神经元电路。M. Prezioso 等人在 2018 年的工作是第一种方案的代表[61]。在该工作中，M. Prezioso 等

人实验上利用 20×20 忆阻器突触阵列和 CMOS 神经元验证了 STDP 学习规则并实现了输入信号的相关性探测（图 2.31（a））。图 2.31（b）给出了相关硬件的原理图。所用神经元电路遵循 LIF 神经元电路的实现方式，第一个跨导放大器（TIA）用于读取忆阻器阵列上的 VMM 电流，随后的积分放大器对读取的电流进行积分。当积分放大器的输出电压大于设定的阈值（V_{b2}）时，最后的比较器输出高电平触发脉冲发生器产生具有正负极性的电压脉冲（图 2.31（a）），该形式的电压脉冲可以在忆阻器突触中实现 overlap 的 STDP 学习规则。工作过程中，脉冲发生器的反馈回路和积分放大器的放电回路具有同步开关的特点，分别对应推理过程和训练过程的不同操作。在推理过程中，脉冲发生器的反馈回路和积分放大器的放电回路同时断开；当脉冲发生器触发产生动作电位信号时，反馈回路和放电回路同时闭合。反馈回路的连通用于给突触传递后神经元动作电位信号，放电回路导通用于实现对电容放电。基于此，进一步验证了两种低噪声模式的相关性检测。图 2.31（c）给出了两种模式相继输入下忆阻器突触电导的演化过程。此外，米兰理工大学的 D. Ielmini 课题基于忆阻器突触阵列和 CMOS 神经元电路也做了一些相关的系统仿真工作[110-112]，初步证实了利用忆阻器突触阵列构建 SNN 芯片的可行性。

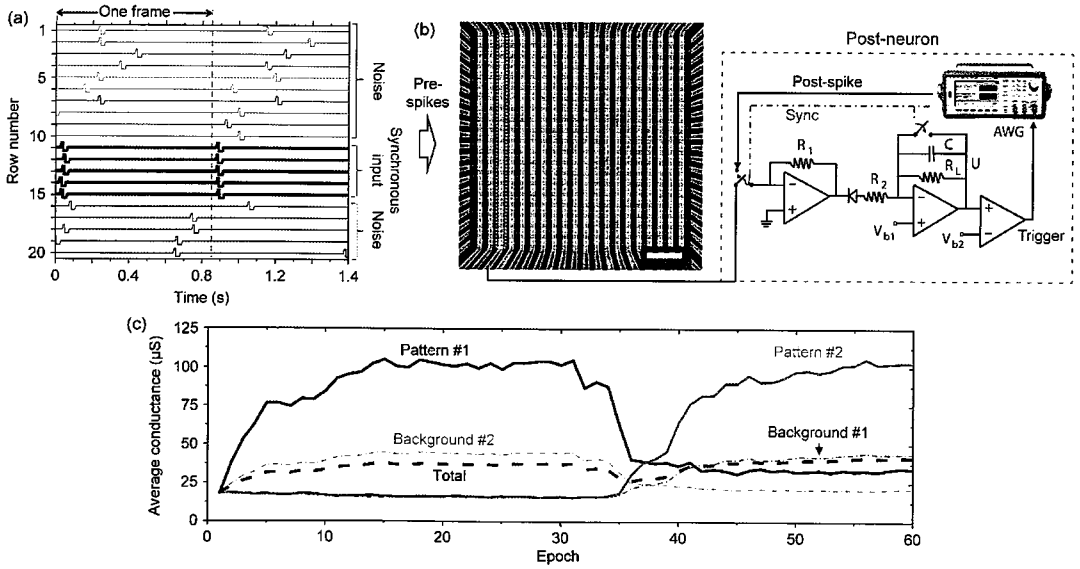


图 2.31 （a）输入信号的模式；（b）忆阻器突触阵列 SEM 图和 CMOS 神经元原理图；（c）两种输入模式下突触电导对迭代次数的演化[61]

由于 CMOS 神经元电路结构复杂，不利于进行大规模集成，利用忆阻器神经元

构建神经形态系统越来越引起人们的广泛关注。2018 年，马萨诸塞大学安姆斯特分校的杨建华课题组在国际上首次利用忆阻器神经元和神经突触构建了 8×8 的全忆阻脉冲神经网络[84]，图 2.32（a）给出了系统硬件示意图。基于该系统，验证了卷积推理运算并实现了输入模式的非监督学习。同年，该课题组又实现了 4×4 的全忆容器耦合脉冲神经网络[113]（图 2.32（b）），并验证了赫布学习规则以及网络的推理结果。

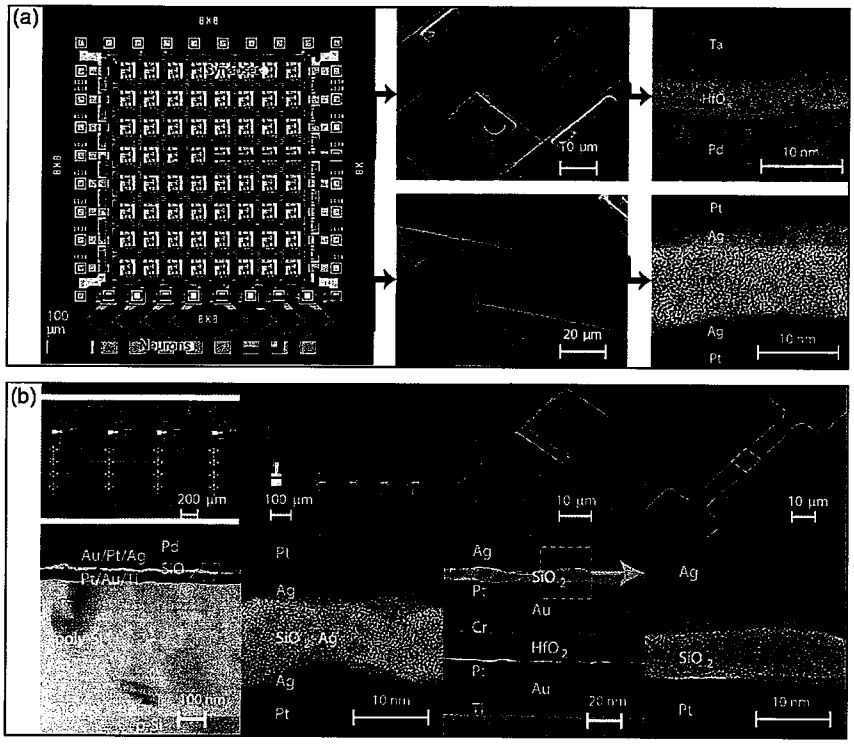


图 2.32 （a）全忆阻脉冲神经网络硬件图[84]；（b）全忆容脉冲神经网络硬件图[113]

总之，基于忆阻器的 SNN 硬件迎合当前大数据时代低功耗的需求，具有广阔的应用前景，然而相关研究工作还处于起步阶段，尚需进一步研究和探索。

2.5 本章小结

本章从忆阻器的分类出发，针对 redox 类型忆阻器的转变机制、电学特性进行了概述。系统阐述了神经元和神经突触的生物工作机制以及在忆阻器中实现的方法和研究进展，最后对利用忆阻器实现神经网络系统的方案和研究现状进行了介绍。

参考文献

[1] L.O.Chua, Memristor-The Missing Circuit Element [J]. IEEE Transactions on Circuit Theory, 1971.

[2] Strukov DB, Snider GS, Stewart DR, et al., The missing memristor found [J]. Nature, vol. 453, pp. 80-3, May 1 2008.

[3] Wang Z, Wu H, Burr GW, et al., Resistive switching materials for information processing [J]. Nature Reviews Materials, 2020.

[4] Lee MJ, Lee CB, Lee D, et al., A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures [J]. Nat Mater, vol. 10, pp. 625-30, Aug 2011.

[5] Zhang W, Mazzarello R, Wuttig M, et al., Designing crystallization in phase-change materials for universal memory and neuro-inspired computing [J]. Nature Reviews Materials, 2019.

[6] Jiang J, Bai ZL, Chen ZH, et al., Temporary formation of highly conducting domain walls for non-destructive read-out of ferroelectric domain-wall resistance switching memories [J]. Nat Mater, vol. 17, pp. 49-56, Jan 2018.

[7] Natterer FD, Yang K, Paul W, et al., Reading and writing single-atom magnets [J]. Nature, vol. 543, pp. 226-228, Mar 8 2017.

[8] Pan F, Gao S, Chen C, et al., Recent progress in resistive random access memories: Materials, switching mechanisms, and performance [J]. Materials Science and Engineering: R: Reports, vol. 83, pp. 1-59, 2014.

[9] Xia Q and Yang JJ, Memristive crossbar arrays for brain-inspired computing [J]. Nature Materials, vol. 18, pp. 309-323, 2019.

[10] Nili H, Adam GC, Hoskins B, et al., Hardware-intrinsic security primitives enabled by analogue state and nonlinear conductance variations in integrated memristors [J]. Nature Electronics, vol. 1, pp. 197-202, 2018.

[11] Wang I-T, Lin Y-C, Wang Y-F, et al., 3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation, in 2014 Ieee International Electron Devices Meeting, ed New York: Ieee, 2014.

[12] Ma C, Luo Z, Huang W, et al., Sub-nanosecond memristor based on ferroelectric tunnel junction [J]. Nat Commun, vol. 11, p. 1439, Mar 18 2020.

[13] Song KM, Jeong J-S, Pan B, et al., Skyrmion-based artificial synapses for neuromorphic computing [J]. Nature Electronics, vol. 3, pp. 148-155, 2020.

[14] Shiokawa Y, Komura E, Ishitani Y, et al., High write endurance up to 10(12) cycles in a spin current-type magnetic memory array [J]. Aip Adv, vol. 9, p. 4, Mar 2019.

[15] Choi BJ, Torrezan AC, Strachan JP, et al., High-Speed and Low-Energy Nitride Memristors [J]. Advanced Functional Materials, vol. 26, pp. 5290-5296, 2016.

[16] Grezes C, Ebrahimi F, Alzate JG, et al., Ultra-low switching energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product [J]. Appl Phys Lett, vol. 108, p. 5, Jan 2016.

[17] Pi S, Li C, Jiang H, et al., Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension [J]. Nat Nanotechnol, Nov 12 2018.

[18] Tuma T, Pantazi A, Le Gallo M, et al., Stochastic phase-change neurons [J]. Nat Nanotechnol, May 16 2016.

[19] Serb A, Bill J, Khiat A, et al., Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses [J]. Nat Commun, vol. 7, p. 12611, Sep 29 2016.

[20] Li C, Hu M, Li Y, et al., Analogue signal and image processing with large memristor crossbars [J]. Nature Electronics, 2017.

[21] Liu Q, Sun J, Lv HB, et al., Real-Time Observation on Dynamic Growth/Dissolution of Conductive Filaments in Oxide-Electrolyte-Based ReRAM [J]. Advanced Materials, vol. 24, pp. 1844-1849, Apr 2012.

[22] Jiang H, Han L, Lin P, et al., Sub-10 nm Ta Channel Responsible for Superior Performance of a HfO2 Memristor [J]. Sci Rep, vol. 6, p. 28525, 2016.

[23] Kwon DH, Kim KM, Jang JH, et al., Atomic structure of conducting nanofilaments in TiO2 resistive switching memory [J]. Nat Nanotechnol, vol. 5, pp. 148-53, Feb 2010.

[24] International Technology Roadmap for Semiconductors[M]. http://www.itrs2.net/, 2015.

[25] Yang Y, Gao P, Li L, et al., Electrochemical dynamics of nanoscale metallic inclusions in dielectrics [J]. Nat Commun, vol. 5, p. 4232, Jun 23 2014.

[26] Hirose Y and Hirose H, Polarity-dependent memory switching and behavior of Ag dendrite in Ag-photodoped amorphous AS_2S_3 films [J]. Journal of Applied Physics, vol. 47, pp. 2767-2772, 1976.

[27] Guo X, Schindler C, Menzel S, et al., Understanding the switching-off mechanism in Ag+ migration based resistively switching model systems [J]. Appl Phys Lett, vol. 91, p. 133513, 2007.

[28] Wang Z, Joshi S, Savel'ev SE, et al., Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing [J]. Nat Mater, Sep 26 2016.

[29] Lee D, Park J, Moon K, et al., Oxide based nanoscale analog synapse device for neural signal recognition system in 2015 Ieee International Electron Devices Meeting, ed New York: IEEE, 2015.

[30] Cooper D, Baeumer C, Bernier N, et al., Anomalous Resistance Hysteresis in Oxide ReRAM: Oxygen Evolution and Reincorporation Revealed by In Situ TEM [J]. Adv Mater, vol. 29, Jun 2017.

[31] Wang YF, Lin YC, Wang IT, et al., Characterization and Modeling of Nonfilamentary $Ta/TaO_x/TiO_2/Ti$ Analog Synaptic Device [J]. Sci Rep, vol. 5, p. 10150, 2015.

[32] Hsu C-W, Wang I-Y, Lo C-L, et al., Self-rectifying bipolar TaO_x/TiO_2 RRAM with superior endurance over 10^{12} cycles for 3D high-density storage-class memory[M]. New York: Ieee, 2013.

[33] Ma J, Chai Z, Zhang W, et al., Identify the critical regions and switching/failure mechanisms in non-filamentary RRAM (a-VMCO) by RTN and CVS techniques for memory window improvement, in 2016 Ieee International Electron Devices Meeting, ed New York: Ieee, 2016.

[34] Govoreanu B, Crotti D, Subhechha S, et al., a-VMCO: a novel forming-free, self-rectifying, analog memory cell with low-current operation, nonfilamentary switching and excellent variability[M]. New York: Ieee, 2015.

[35] Govoreanu B, Piazza LD, Ma J, et al., Advanced a-VMCO resistive switching memory through

inner interface engineering with wide (>10^2) on_off window, tunable μA-range switching current and excellent variability[M]. New York: Ieee, 2016.

[36] Yoon JH, Song SJ, Yoo I-H, et al., Highly Uniform, Electroforming-Free, and Self-Rectifying Resistive Memory in the Pt/Ta_2O_5/HfO_{2-x}/TiN Structure [J]. Advanced Functional Materials, vol. 24, pp. 5086-5095, 2014.

[37] Kim KM, Zhang J, Graves C, et al., Low-Power, Self-Rectifying, and Forming-Free Memristor with an Asymmetric Programing Voltage for a High-Density Crossbar Application [J]. Nano Lett, vol. 16, pp. 6724-6732, Nov 09 2016.

[38] Berdan R, Serb A, Khiat A, et al., A mu-Controller-Based System for Interfacing Selectorless RRAM Crossbar Arrays [J]. Ieee T Electron Dev, vol. 62, pp. 2190-2196, Jul 2015.

[39] Yao P, Wu H, Gao B, et al., Face classification using electronic synapses [J]. Nat Commun, vol. 8, p. 15199, May 12 2017.

[40] Xu XX, Tai L, Gong TC, et al., 40x Retention Improvement by Eliminating Resistance Relaxation with High Temperature Forming in 28 nm RRAM Chip, in 2018 Ieee International Electron Devices Meeting, ed New York: Ieee, 2018.

[41] Woo J, Peng X, and Yu S, Design Considerations of Selector Device in CrossPoint RRAM Array for Neuromorphic Computing [J]. In 2018 IEEE Int. Symp. Circuits and Systems (ISCAS) https://doi.org/10.1109/ISCAS.2018.8351735, 2018.

[42] Kim H, Nili H, Mahmoodi MR, et al., 4K-Memristor Analog-Grade Passive Crossbar Circuit [J]. arXiv, 2019.

[43] Cai F, Correll JM, Lee SH, et al., A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations [J]. Nature Electronics, vol. 2, pp. 290-299, 2019.

[44] Xu XX, Luo Q, Gong TC, et al., Fully CMOS Compatible 3D Vertical RRAM with Self-aligned Self-selective Cell Enabling Sub-5nm Scaling[M]. New York: Ieee, 2016.

[45] Luo Q, Xu XX, Liu HT, et al., Demonstration of 3D Vertical RRAM with Ultra Low-leakage, High-selectivity and Self-compliance Memory Cells[M]. New York: Ieee, 2015.

[46] Purves D, Augustine GJ, Fitzpatrick D, et al., Neuroscience, 3rd ed. [M]. Inc. Massachusetts, USA: Sinauer Associates, 2012.

[47] Abbott LF and Regehr WG, Synaptic computation [J]. Nature, vol. 431, pp. 796–803, 2004.

[48] Ohno T, Hasegawa T, Tsuruoka T, et al., Short-term plasticity and long-term potentiation mimicked in single inorganic synapses [J]. Nat Mater, vol. 10, pp. 591-5, Aug 2011.

[49] Chang T, Jo S-H, and Wei Lu W, Short-Term Memory to Long-Term Memory Transition in a Nanoscale Memristor [J]. ACS NANO, vol. 5, pp. 7669-7676, August 23 2011.

[50] Park Y and Lee JS, Artificial Synapses with Short- and Long-Term Memory for Spiking Neural Networks Based on Renewable Materials [J]. ACS Nano, Aug 28 2017.

[51] Kim MK and Lee JS, Short-Term Plasticity and Long-Term Potentiation in Artificial Biosynapses with Diffusive Dynamics [J]. ACS Nano, vol. 12, pp. 1680-1687, Feb 27 2018.

[52] Shi Y, Liang X, Yuan B, et al., Electronic synapses made of layered two-dimensional materials [J]. Nature Electronics, vol. 1, pp. 458-465, 2018.

[53] Xu W, Cho H, Kim YH, et al., Organometal Halide Perovskite Artificial Synapses [J]. Adv Mater, vol. 28, pp. 5916-22, Jul 2016.

[54] Yan X, Zhou Z, Zhao J, et al., Flexible memristors as electronic synapses for neuro-inspired

computation based on scotch tape-exfoliated mica substrates [J]. Nano Research, 2017.

[55] Jo SH, Chang T, Ebong I, et al., Nanoscale memristor device as synapse in neuromorphic systems [J]. Nano Lett, vol. 10, pp. 1297-301, Apr 14 2010.

[56] Wu W, Wu H, Gao B, et al., Improving Analog Switching in HfOx Based Resistive Memory with Thermal Enhanced Layer [J]. Ieee Electr Device L, pp. 1-1, 2017.

[57] Wang Z, Yin M, Zhang T, et al., Engineering incremental resistive switching in TaOx based memristors for brain-inspired computing [J]. Nanoscale, vol. 8, pp. 14015-22, Aug 7 2016.

[58] Woo J, Moon K, Song J, et al., Improved Synaptic Behavior Under Identical Pulses Using AlO_x/HfO_2 Bilayer RRAM Array for Neuromorphic Systems [J]. Ieee Electr Device L, vol. 37, pp. 994-997, 2016.

[59] Caporale N and Dan Y, Spike timing-dependent plasticity: a Hebbian learning rule [J]. Annual review of neuroscience, vol. 31, pp. 25-46, 2008.

[60] Froemke RC and Dan Y, Spike-timing-dependent synaptic modification induced by natural spike trains [J]. Nature, vol. 416, pp. 433-438, Mar 2002.

[61] Prezioso M, Mahmoodi MR, Bayat FM, et al., Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits [J]. Nat Commun, vol. 9, p. 5311, Dec 14 2018.

[62] Wu Y, Yu S, and Wong H-SP, Alo_x-based Resistive Switching Device With Gradual Resistance Modulation For Neuromorphic Device Application, presented at the 2012 IEEE International Memory Workshop (IMW) New York, 2012.

[63] Li Y, Zhong Y, Zhang J, et al., Activity-dependent synaptic plasticity of a chalcogenide electronic synapse for neuromorphic systems [J]. Sci Rep, vol. 4, p. 4906, May 09 2014.

[64] Prezioso M, Merrikh Bayat F, Hoskins B, et al., Self-Adaptive Spike-Time-Dependent Plasticity of Metal-Oxide Memristors [J]. Sci Rep, vol. 6, p. 21331, 2016.

[65] Du C, Ma W, Chang T, et al., Biorealistic Implementation of Synaptic Functions with Oxide Memristors through Internal Ionic Dynamics [J]. Advanced Functional Materials, vol. 25, pp. 4290-4299, 2015.

[66] Yang R, Huang H-M, Hong Q-H, et al., Synaptic Suppression Triplet-STDP Learning Rule Realized in Second-Order Memristors [J]. Advanced Functional Materials, p. 1704455, 2017.

[67] Wang Z, Zeng T, Ren Y, et al., Toward a generalized Bienenstock-Cooper-Munro rule for spatiotemporal learning via triplet-STDP in memristive devices [J]. Nat Commun, vol. 11, p. 1510, Mar 20 2020.

[68] Kirkwood A, Rioult MG, and Bear MF, Experience-dependent modification of synaptic plasticity in visual cortex [J]. Nature, vol. 381, pp. 526-528, 1996.

[69] Xiong J, Yang R, Shaibo J, et al., Bienenstock, Cooper, and Munro Learning Rules Realized in Second-Order Memristors with Tunable Forgetting Rate [J]. Advanced Functional Materials, vol. 29, p. 1807316, 2019.

[70] Yin J, Zeng F, Wan Q, et al., Adaptive Crystallite Kinetics in Homogenous Bilayer Oxide Memristor for Emulating Diverse Synaptic Plasticity [J]. Advanced Functional Materials, vol. 28, p. 1706927, 2018.

[71] Baars BJ and Gate NM, Cognition, Brain and Consciousness, 2nd ed. [M]. ELSEVIER, 2010.

[72] Zador CFSaAM, Novel integrate-and-fire-like model [J]. Proceedings of the 5th Joint

Symposium on Neural Computation, 1998.

[73] Huxley AHaA, A quantitative description of membrane current and its application to conduction and excitation in nerve [J]. The Journal of Physiology, vol. 117, pp. 500-544, 1952.

[74] Izhikevich EM, Simple model of spiking neurons [J]. IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 14, pp. 1569-1572, 2003.

[75] Pickett MD, Medeiros-Ribeiro G, and Williams RS, A scalable neuristor built with Mott memristors [J]. Nature Materials, vol. 12, pp. 114-117, 2013.

[76] Yi W, Tsang KK, Lam SK, et al., Biological plausibility and stochasticity in scalable VO_2 active memristor neurons [J]. Nat Commun, vol. 9, p. 4661, Nov 7 2018.

[77] Lin J, Annadi A, Sonde S, et al., Low-voltage artificial neuron using feedback engineered insulator-to-metal-transition devices, in 2016 Ieee International Electron Devices Meeting (IEDM), ed New York: Ieee, 2016.

[78] Gao L, Chen P-Y, and Yu S, NbO_x based oscillation neuron for neuromorphic computing [J]. Appl Phys Lett, vol. 111, p. 103503, 2017.

[79] Jerry M, Parihar A, Grisafe B, et al., Ultra-Low Power Probabilistic IMT Neurons for Stochastic Sampling Machines[M]. New York: Ieee, 2017.

[80] Cobley RA, Hayat H, and Wright CD, A self-resetting spiking phase-change neuron [J]. Nanotechnology, vol. 29, p. 195202, May 11 2018.

[81] Jang J-W, Attarimashalkoubeh B, Prakash A, et al., Scalable Neuron Circuit Using Conductive-Bridge RAM for Pattern Reconstructions [J]. Ieee T Electron Dev, vol. 63, pp. 2610-2613, 2016.

[82] Lashkare S, Chouhan S, Chavan T, et al., PCMO RRAM for Integrate-and-Fire Neuron in Spiking Neural Networks [J]. Ieee Electr Device L, vol. 39, pp. 484-487, Apr 2018.

[83] Wang JJ, Hu SG, Zhan XT, et al., Handwritten-Digit Recognition by Hybrid Convolutional Neural Network based on HfO_2 Memristive Spiking-Neuron [J]. Sci Rep, vol. 8, p. 12546, Aug 22 2018.

[84] Wang Z, Joshi S, Savel'ev S, et al., Fully memristive neural networks for pattern classification with unsupervised learning [J]. Nature Electronics, vol. 1, pp. 137-145, 2018.

[85] Mehonic A and Kenyon AJ, Emulating the Electrical Activity of the Neuron Using a Silicon Oxide RRAM Cell [J]. Front Neurosci, vol. 10, p. 57, 2016.

[86] Huang HM, Yang R, Tan ZH, et al., Quasi-Hodgkin-Huxley Neurons with Leaky Integrate-and-Fire Functions Physically Realized with Memristive Devices [J]. Adv Mater, vol. 31, p. e1803849, Jan 2019.

[87] Hao S, Ji X, Zhong S, et al., A Monolayer Leaky Integrate-and-Fire Neuron for 2D Memristive Neuromorphic Networks [J]. Advanced Electronic Materials, p. 1901335, 2020.

[88] Zhang Y, He W, Wu Y, et al., Highly Compact Artificial Memristive Neuron with Low Energy Consumption [J]. Small, p. e1802188, Nov 14 2018.

[89] Pantazi A, Wozniak S, Tuma T, et al., All-memristive neuromorphic computing with level-tuned neurons [J]. Nanotechnology, vol. 27, p. 355205, Sep 2 2016.

[90] Stoliar P, Tranchant J, Corraze B, et al., A Leaky-Integrate-and-Fire Neuron Analog Realized with a Mott Insulator [J]. Advanced Functional Materials, p. 1604740, 2017.

[91] Mulaosmanovic H, Chicca E, Bertele M, et al., Mimicking biological neurons with a nanoscale ferroelectric transistor [J]. Nanoscale, vol. 10, pp. 21755-21763, Dec 2018.

[92] Dutta S, Saha A, Panda P, et al., Biologically Plausible Ferroelectric Quasi-Leaky Integrate and Fire Neuron[M]. New York: Ieee, 2019.

[93] Chen C, Yang M, Liu S, et al., Bio-Inspired Neurons Based on Novel Leaky-FeFET with Ultra-Low Hardware[M]. New York, 2019.

[94] Wu MH, Hong MC, Chang C-C, et al., Extremely Compact Integrate-and-Fire STT-MRAM Neuron: A Pathway toward All-Spin Artificial Deep Neural Network[M]. New York: Ieee, 2019.

[95] Palma G, Suri M, Querlioz D, et al., Stochastic neuron design using Conductive bridge RAM [J]. 2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), vol. 13828470, 2013.

[96] Kalita H, Krishnaprasad A, Choudhary N, et al., Artificial Neuron using Vertical MoS2/Graphene Threshold Switching Memristors [J]. Sci Rep, vol. 9, p. 53, Jan 10 2019.

[97] Wright CD, Hosseini P, and Diosdado JAV, Beyond von-Neumann Computing with Nanoscale Phase-Change Memory Devices [J]. Advanced Functional Materials, vol. 23, pp. 2248-2254, 2013.

[98] Lee D, Kwak M, Moon K, et al., Various Threshold Switching Devices for Integrate and Fire Neuron Applications [J]. Advanced Electronic Materials, p. 1800866, 2019.

[99] Jain S, Ankit A, Chakraborty I, et al., Neural network accelerator design with resistive crossbars: Opportunities and challenges [J]. IBM Journal of Research and Development, vol. 63, pp. 10:1-10:13, 2019.

[100] Alibart F, Zamanidoost E, and Strukov DB, Pattern classification by memristive crossbar circuits using ex situ and in situ training [J]. Nat Commun, vol. 4, p. 2072, 2013.

[101] Prezioso M, Merrikh-Bayat F, Hoskins BD, et al., Training and operation of an integrated neuromorphic network based on metal-oxide memristors [J]. Nature, vol. 521, pp. 61-64, 2015.

[102] Sheridan PM, Cai F, Du C, et al., Sparse coding with memristor networks [J]. Nat Nanotechnol, May 22 2017.

[103] Li C, Belkin D, Li Y, et al., Efficient and self-adaptive in-situ learning in multilayer memristor neural networks [J]. Nat Commun, vol. 9, p. 2385, Jun 19 2018.

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

[105] Wang Z, Li C, Song W, et al., Reinforcement learning with analogue memristor arrays [J]. Nature Electronics, vol. 2, pp. 115-124, 2019.

[106] Chen W-H, Dou C, Li K-X, et al., CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors [J]. Nature Electronics, vol. 2, pp. 420-428, 2019.

[107] Yao P, Wu H, Gao B, et al., Fully hardware-implemented memristor convolutional neural network [J]. Nature, vol. 577, pp. 641-646, Jan 2020.

[108] Pfeiffer M and Pfeil T, Deep Learning With Spiking Neurons: Opportunities and Challenges [J]. Front Neurosci, vol. 12, p. 774, 2018.

[109] Liu C, Jiang H, Yan B, et al., A spiking neuromorphic design with resistive crossbar [J]. pp. 1-6, 2015.

[110] Wang W, Pedretti G, Milo V, et al., Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses [J]. Science Advances, vol. 4, p. 8, Sep 2018.

[111] Milo V, Pedretti G, Carboni R, et al., Demonstration of hybrid CMOS/RRAM neural networks

with spike time/rate-dependent plasticity, in 2016 Ieee International Electron Devices Meeting, ed New York: Ieee, 2016.

[112] Ambrogio S, Balatti S, Milo V, et al., Neuromorphic Learning and Recognition With One-Transistor-One-Resistor Synapses and Bistable Metal Oxide RRAM [J]. Ieee T Electron Dev, vol. 63, pp. 1508-1515, 2016.

[113] Wang Z, Rao M, Han JW, et al., Capacitive neural network with neuro-transistors [J]. Nat Commun, vol. 9, p. 3208, Aug 10 2018.

第3章 忆阻器基神经突触研究

由前面章节的背景介绍我们知道，忆阻器是一种理想的神经形态器件。自从 2010 年 Jo 等人[1]用 a-Si 基的忆阻器在实验上验证了突触的功能以来，神经突触的功能在各种各样的忆阻器器件上陆续得到了验证，包括阻变存储器[2-5]，相变存储器[6-9]，磁存储器[10]，铁电隧穿结[11]或铁电 FET[12]等。其中，离子基阻变忆阻器[3, 13]，由于其制备工艺简单，生物突触相似性高，具有低功耗和高密度三维集成的潜力而被广泛研究用于模拟生物突触的动态特性。根据相关文献报道[14, 15]，在生物突触中，突触的权重由不同的离子种类（例如 Ca^{2+}，Na^{+}和 K^{+}）的浓度调制，激活或抑制神经递质从前神经突触释放到后神经突触。与生物突触相似，离子基忆阻器的电导可以通过在功能层中阳离子（如 Cu^{2+} 和 Ag^{+}）的积累或回流连续调制，从而能生动地模仿生物突触的动力学机制[5, 16, 17]。到目前为止，许多突触功能，例如短时程可塑性[3, 18]（短时程增强，对脉冲易化和抑制），长时程可塑性[4, 19]（长时程增强和减弱）和时序依赖可塑性（STDP）[20-22]等，已经在离子基忆阻器中实现。在生物神经系统中，短时程可塑性被认为在实现计算功能中起着关键的作用，而长时程的可塑性则被认为是学习和记忆功能实现的核心[23]。值得注意的是，在单一的忆阻器中同时实现短时程和长时程可塑性可用于模拟生物突触的短时程记忆到长时程记忆的动态进化过程，这在构建类脑神经形态系统上是至关重要的[5, 16]。另外，在利用忆阻器突触实现系统应用时，我们不仅要考虑对生物突触功能的实现，还要考虑实现相关功能参数的优越性[24]。一个理想突触器件的电学特性，至少需要具备以下几点特性：良好的缓变特性，线性和对称的 LTP/LTD 调节过程，低能耗等[25]。为进行突触功能仿生实现和器件性能优化，我们开展了相关的研究工作，希望能用忆阻器实现生物突触的相关功能及特性验证，为以后的系统实现提供理论指导。

在本章中，我们制备了两种忆阻器来进行神经突触的仿生实现。首先我们以 Cu/a-Si/Pt 结构的离子基忆阻器为基础，成功地模拟了突触的短时程和长时程可塑性行为。在施加单一的小刺激脉冲下，观察到突触的短时程增强现象，并且在两个相同的脉冲刺激下表现出对脉冲易化突触行为。在一系列相同的正向和负向脉冲的作用下，我们

又分别实现了长时程增强（LTP）和长时程减弱（LTD）突触功能。有趣的是，在重复相同的小脉冲刺激下，该突触器件也可以从短时程记忆（short term memory，STM）转变为长时程记忆（long term memory，LTM）。此外，通过调节突触前和突触后脉冲的时间间隔成功实现了与生物学习相关的 STDP 突触功能。其次，为实现突触器件的多值可控及线性电导调制，我们进一步通过双层堆叠工艺，设计了一种具有低工作电压和良好模拟转变特性的 $Pd/HfO_2/WO_x/W$ 叠层界面型转变突触器件。在脉冲编程下，可实现极宽的连续电导窗口（>300）和高耐久周期（$>10^6$）。随后，我们分别探讨了器件在变脉幅和变脉宽编程方案下的电导调制特性，得到了近似线性的电导变化行为，在 MNIST 手写数字数据集上的系统仿真实现了 95.3%的识别率，接近理想效果。下面将基于制备的两种忆阻器分别展开讨论。

3.1 基于 Cu/a-Si/Pt 忆阻器的突触仿生实现

3.1.1 Cu/a-Si/Pt 器件的制备工艺流程

该小节主要讲述本工作中所用到的 Cu/a-Si/Pt 器件的制备过程，具体的制备步骤如下：

(1) 经过第一步光刻后，在 SiO_2/Si 衬底上通过电子束蒸发沉积厚度分别为 40 nm 和 10 nm 的 Pt 和 Ti 作为底电极，然后剥离形成垂直的下电极线，其中 10 nm 的 Ti 作为粘附层。

(2) 进行第二步光刻，磁控溅射淀积 10 nm 的 a-Si 作为功能层，剥离形成中间层图形。

(3) 进行最后一步光刻，然后通过磁控溅射淀积 40 nm 的 Cu 作为上电极。

(4) 剥离获得器件图形。

制备后的 Cu/a-Si/Pt 器件的 SEM 图像如图 3.1 所示。上电极与下电极的交叉部分即为器件单元的面积。除了面积为 5 μm × 5 μm 的器件，我们同时还制备了 1 μm × 1 μm，2 μm × 2 μm，3 μm × 3 μm，和 4 μm × 4 μm 的器件。在本工作中，我们主要基于 5 μm × 5 μm 的器件进行实验操作。所有的电学测量都是在安捷伦 B1500A 上进行的。在测量过程中，在 Cu 电极上施加电压，Pt 电极接地。在突触功能测试中，器件的响应电流作为后突触响应电流。

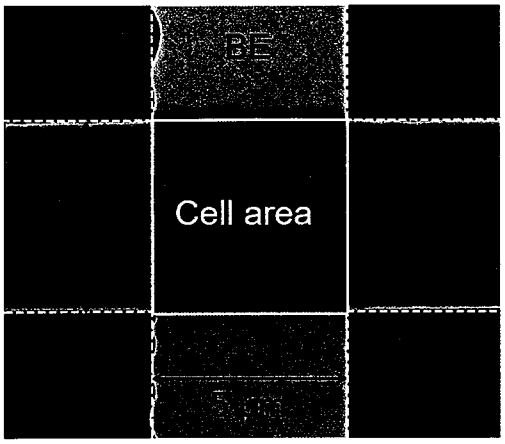


图 3.1 Cu/a-Si/Pt 器件的 SEM 图像（5 μm × 5 μm）

3.1.2 Cu/a-Si/Pt 器件的长时程可塑性

图 3.2 给出了生物突触和 Cu/a-Si/Pt 器件的对应结构原理图。Cu 电极作为前突触，Pt 电极作为后突触。当在 Cu 电极上施加正偏置时，Cu 原子被氧化为 Cu 离子，在电场作用下注入阻变（RS）层，导致器件电导增加。这个过程可以用来模拟 Ca^{2+}在生物突触中的涌入，通过释放更多的神经递质来增强突触的权重的过程。

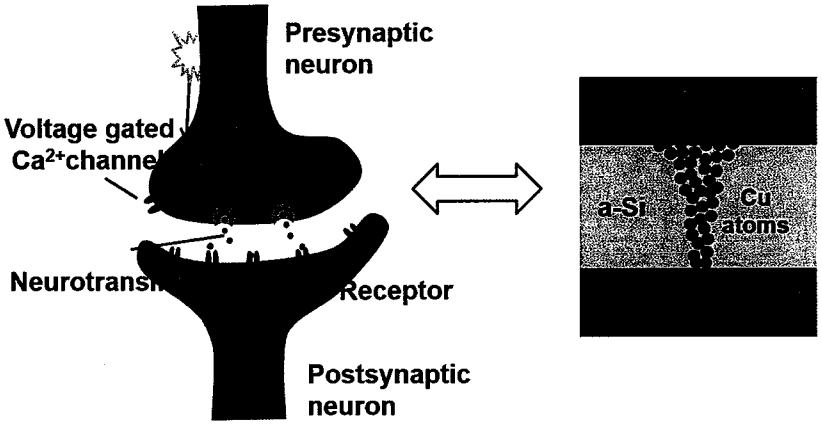


图 3.2 生物突触和 Cu/a-Si/Pt 的对应结构原理图

为了模拟生物突触的长时程可塑性特性，我们首先对器件的多值转变特性进行了研究。图 3.3 给出了器件在不同限流下进行 set 操作的电导增强过程和不同 reset 电压下的电导减弱过程，该结果表明通过调节限流或者 reset 电压可以调控器件中细丝的形态，从而实现多值操作。

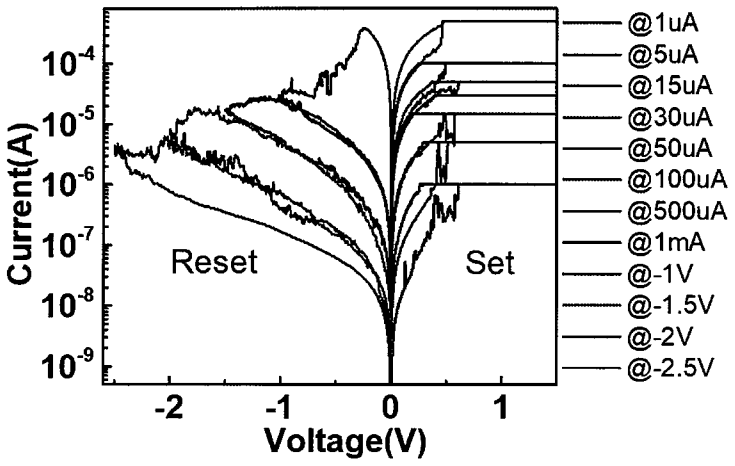


图 3.3 Cu/a-Si/Pt 器件在不同限流和 reset 电压下的直流扫描曲线

为更直观的实现突触的长时程可塑性，我们分别对器件施加一连串间隔时间为 4.5 μs 的正电压脉冲（1.6 V、500 ns）和负电压脉冲（-3 V、500 ns），并通过在每个脉冲后施加读取电压脉冲（0.2 V、1 μs）实时测量器件电导值。测试结果如图 3.4（a）所示。随着正脉冲或负脉冲的数目的增加，可以清楚地看出器件电导的逐渐增加或降低，成功实现了生物的 LTP 和 LTD 过程。基于赫布学习规则，突触权重的增加将增加后神经元的兴奋性[26]。为了避免神经元过度兴奋，突触权重的增加必须有一个饱和值。图 3.4（b）给出了在不同电压幅度的一系列正脉冲下，Cu/a-Si/Pt 器件的电导饱和特性。在前面的几个脉冲中，电导值迅速增加，这是因为采用了较宽（1 μs）的脉冲刺激。随后，器件电导逐渐到达饱和值，并且饱和值随电压振幅升高而增大。结果表明，器件可以内在的实现突触的饱和特性，有利于稳定系统的实现。

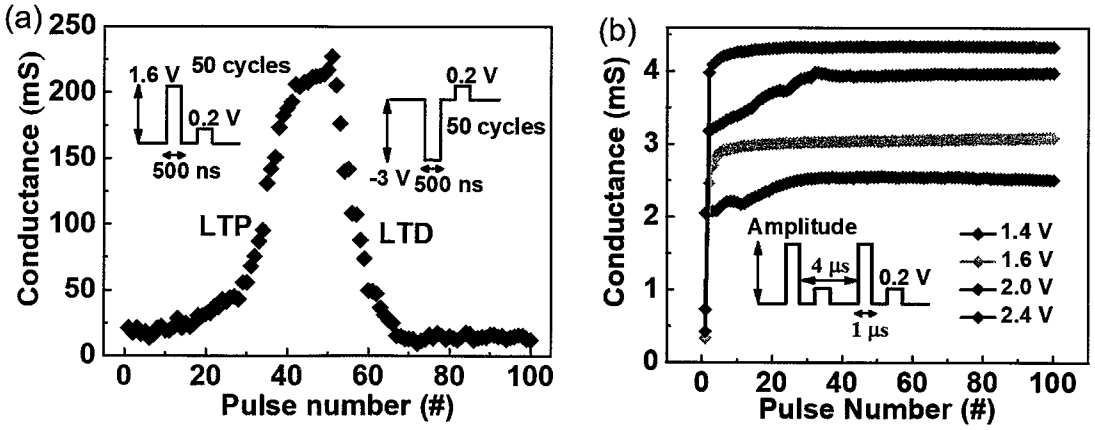


图 3.4 （a）Cu/a-Si/Pt 器件的 LTP 和 LTD 过程；（b）Cu/a-Si/Pt 器件的电导饱和特性

前面的部分表明，突触的长时程可塑性可以通过调节施加脉冲的参数（例如,脉冲幅度和数目）来实现。然而，前后神经元之间活动的时间关系（Δt）也是长时程可塑性的关键决定因素。对于一个特定的突触刺激，如果突触前神经元活动超前于后神经元的活动（Δt＞0），将会导致突触的长时程增强，如果突触后神经元的活动超前于前神经元(Δt<0)，将导致长时程减弱现象。因此，突触权重的变化是前后神经元活动时序 Δt 的一个函数，即所谓的 STDP[27]。为了模拟该功能，我们将两个具有不同间隔时间 Δt 的电压脉冲（突触前脉冲（0.4 V, 10 μs)；突触后脉冲（－0.6 V, 10 μs））分别施加到器件的上电极和下电极。权重变化的百分比可以用以下公式来描述:

$$\text{Weight change (\%)} = \frac{W_1 - W_0}{W_0} \times 100\% \qquad (1)$$

在公式（1）中，W_1 和 W_0 分别是在一定的固定间隔时间内的最终和初始电导。测试结果如图 3.5 所示，当 Δt＞0 时，器件的电导（权重）增加，并且权重随着 Δt 的增加而减小。相反，当 Δt<0 时，器件的权重降低，并且权重随着 Δt 的减少而增加。该结果表明基于 Cu/a-Si/Pt 忆阻器能够成功模拟突触的 STDP 学习规则。

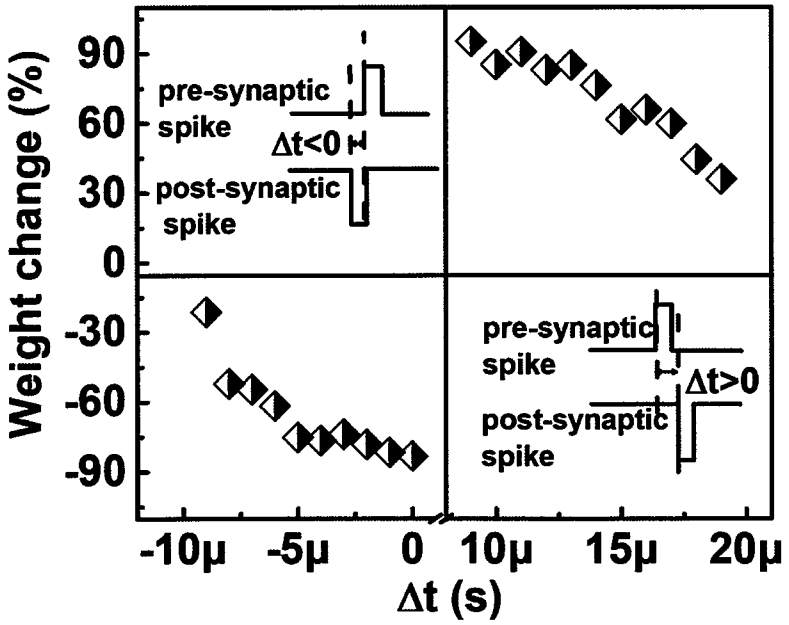


图 3.5 基于 Cu/a-Si/Pt 器件的 STDP 学习规则实现

3.1.3 Cu/a-Si/Pt 器件的短时程可塑性

生物突触的短时程可塑性是神经系统中计算功能实现的关键[23]。在这个过程中，Ca^{2+}进入前神经突触末端触发神经递质的迅速释放，然后导致突触的短时程增强特性。

增强后，Ca^{2+}返回到静息状态的水平，并准备下一次触发。这个过程可以由忆阻器的易失性特征模拟实现。图 3.6（a）展示了在小限流（10 μA）的直流电压扫描下 Cu/a-Si/Pt 器件的易失性阈值转变（TS）行为。当在器件上施加正电压时，当电压达到阈值电压（~ 0.5 V）时，器件转变为高电导状态（HCS），通过器件的电流迅速增加。回扫过程中，当电压下降到一定的值(~ 0.05 V)时，器件自发返回到初始低电导状态（LCS）。该行为是由于界面能最小化而引起的 Cu 导电细丝的自发破裂，该过程已直接通过原位透射电镜观察到[3]。基于这种阈值转变特性，可以用 Cu/a-Si/Pt 器件实现突触的短时程可塑性功能，包括短时程增强和对脉冲易化行为。图 3.6（b）展示了我们的器件在单个电压脉冲下的短时程增强特性。当单个电压脉冲（0.8 V, 30 ms）施加到该器件上时，该器件经过一个短时间延迟（~ 5 ms）后切换到高电导状态，当施加电压脉冲撤销后器件经过~30 ms 的弛豫自发衰变到初始的低电导状态。该衰变现象可以由界面能的最小化导致铜原子的自发团簇来解释[3]。

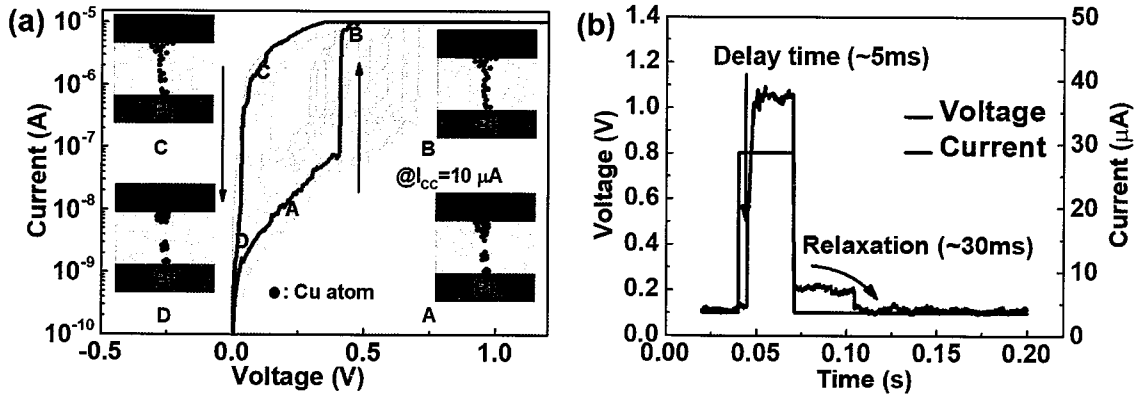


图 3.6 （a）小限流下 Cu/a-Si/Pt 器件的阈值转变特性和相应的器件功能层中 Cu^{2+}的动态变化过程；（b）单脉冲作用下器件响应电流的延迟和弛豫现象

生物突触中的对脉冲易化短时程功能可以描述如下：当刺激间隔时间小于 Ca^{2+}的恢复时间时，第二个突触后响应电流（蓝线）比第一个突触后响应电流（红线）大，如图 3.7（a）中所示[14]。同样的，在我们的器件中，对脉冲易化功能可以用一对具有较短间隔（0.5 ms）的电压脉冲（0.8 V, 1 ms）来模拟，如图 3.7（b）所示。第二刺激脉冲的响应电流大于第一个刺激脉冲的响应电流（图 3.7（b）中的蓝线）。当两个脉冲刺激之间的间隔时间比器件中 Cu 原子的弛豫时间短时，可以通过阻变层中 Cu 原

子的积累来解释，该积累行为导致了器件电导的增加。

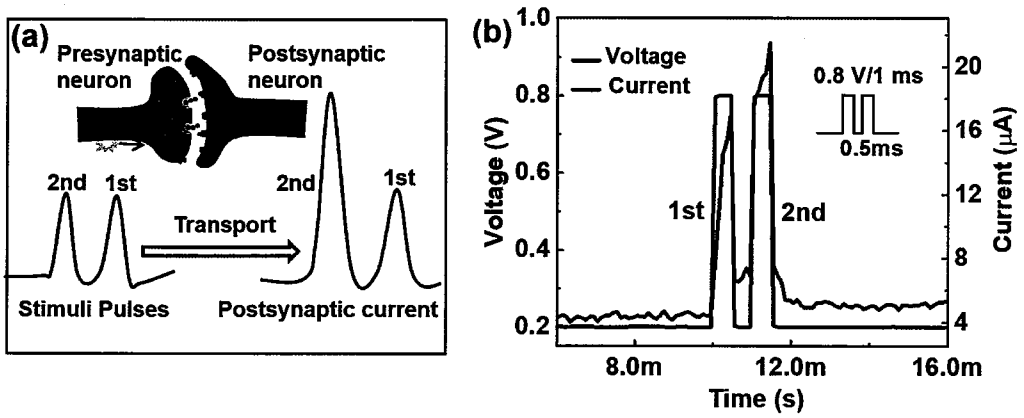


图 3.7 （a）生物突触的对脉冲易化原理图；（b）Cu/a-Si/Pt 器件的对脉冲易化现象

3.1.4 短时程记忆到长时程记忆的转换

在生物系统中，短时程记忆通常比长时程记忆所保持的时间更短[15]。根据大脑的遗忘和记忆行为，短时程记忆可以通过反复训练过程转化为长时程记忆。图 3.8（a）展示了基于 Cu/a-Si/Pt 器件的短时程记忆到长时程记忆转变的示意图。在少量刺激下，Cu^{2+}迁移到功能层的数目很少并且这种情况下 Cu 原子倾向于自发衰变到初始状态。在反复激励下，更多的 Cu^{2+}将迁移到功能层导致器件电导的长时程变化，就像在生物系统的训练过程一样。该过程采用一连串刺激脉冲（0.8 V, 1 ms）和读脉冲（(0.2 V, 10 ms）（图 3.8（b)），并实时测量突触后响应电流（图 3.8（c））。在每一个刺激/读取周期，在 0.8 V 刺激脉冲下，器件响应电流迅速增加，然后衰减到一个剩余低电导状态，对应于生物突触的短时程记忆行为。值得注意的是，突触后响应电流在前 5 个刺激脉冲（图 3.8（d）为放大图）有一个整体的增加。随着刺激数目的增加最终达到一个稳定的高电导状态，就像生物突触的长时程记忆行为一样，如图 3.8（c）所示。

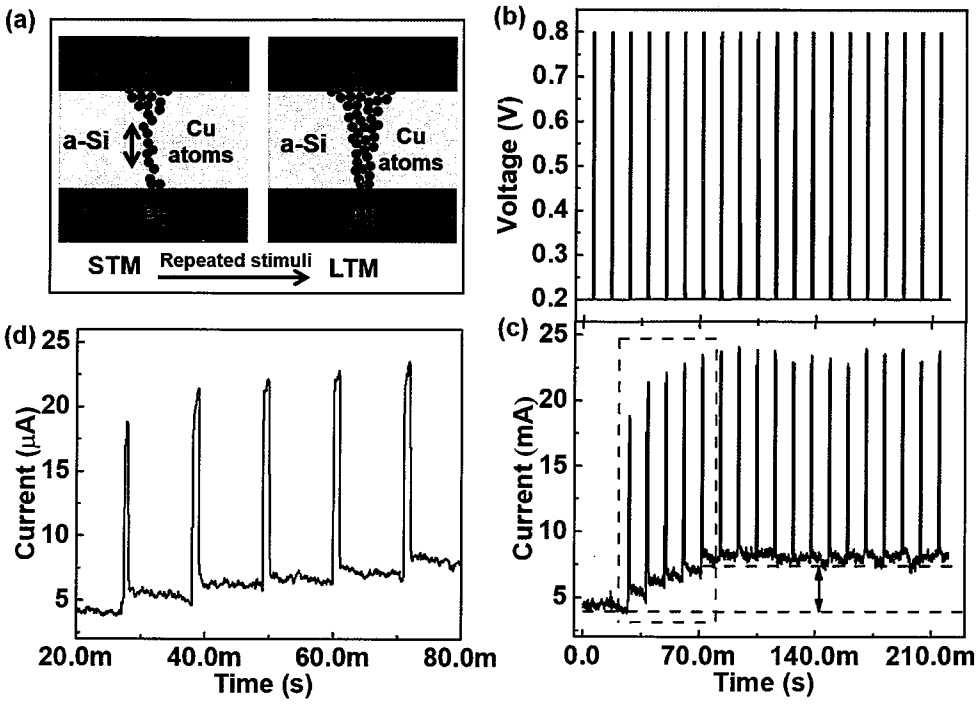


图 3.8 （a）反复刺激下 Cu 离子的动态变化示意图；（b）施加到器件上的 20 个连续脉冲；（c）器件对应的响应电流；（d）是图（c）中红色矩形框内的放大图

3.2 基于 $Pd/HfO_2/WO_x/W$ 忆阻器的突触性能优化

前面提到，为实现突触器件的多值可控及线性电导调制，我们进一步通过双层堆叠工艺，设计了一种具有低工作电压和良好模拟转变特性的 Pd/HfO2/WOx/W 叠层界面型转变突触器件，下面将对该部分工作进行展开陈述。

3.2.1 $Pd/HfO_2/WO_x/W$ 突触器件的制备工艺与电学表征

所设计的 $Pd/HfO_2/WO_x/W$ 叠层器件的制备工艺流程如图 3.9（a）所示，化学机械抛光（chemical mechanical polishing, CMP）后的 W 栓用作底部电极（BE）。然后采用等离子体增强化学气相沉积（PECVD）技术，在氧等离子体中对 W 进行快速热退火 300 s。氧等离子体采用 $2.4\times10^6 cm^3/s$，100W，400 ℃。随后，用原子层沉积法（ALD）分别以 $Hf[N(CH_3)(C_2H_5)]_4$ 和 O_3 为 Hf 的前驱体和氧源在 260℃下沉积了 3 nm HfO_2。最后，采用离子束溅射法沉积了厚度为 70 nm 的钯（Pd）作为顶电极。制备后的 $Pd/HfO_2/WO_x/W$ 器件的 TEM 图像如图 3.9（b）所示。可以清楚的看到 HfO_2/WO_x 双层结构。钨栓的直径为 2 μm，对应器件的有效面积大约为 3.14 μm^2。对器件的直

流特性扫描是在 Agilent B1500A 上进行的，脉冲测试是在 Keithley 4200 上进行的。进行电学测试时，在 Pd 顶电极上施加电压，W 电极接地。在突触功能测试中，器件的响应电流作为后突触响应电流。

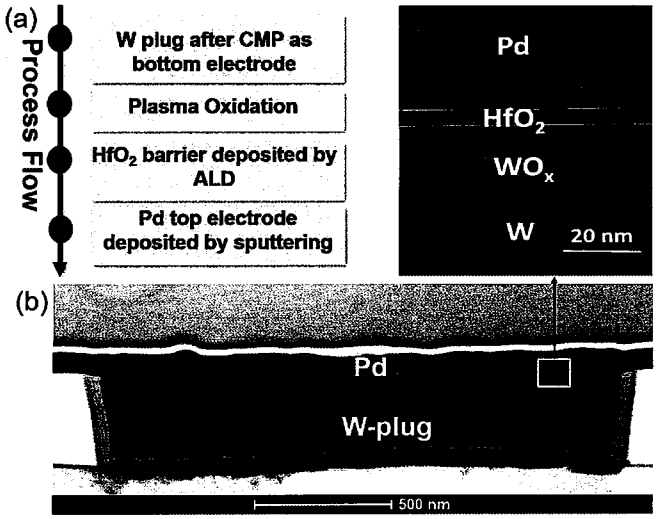


图 3.9 （a）$Pd/HfO_2/WO_x/W$ 器件的制备流程图；（b）器件的 TEM 图像

随后，我们对器件进行了电学性能测试，发现该器件无需进行 forming 操作并且初始态处在高阻态和低阻态之间的中间状态。首先，我们对器件施加从 0 至 2 V 的正扫描电压和 1mA 的限流对器件进行 set 操作，然后施加 0 至-3 V 的负扫描电压进行 reset 操作，紧跟着再一次的 set 操作来展现器件的转变特性，测试结果如图 3.10（a）所示。可以看到器件在 set 和 reset 过程都是缓变的。为进一步体现器件的缓变特性，我们对器件的 set 过程执行不同的限流操作（50 μA、100 μA、200 μA、500 μA 和 1 mA），并且在 reset 过程对应执行具有不同复位电压（-0.8 V、-1.2 V、-1.6 V、-2.0 V、-2.4 V 和- 3.0 V）的直流扫描，如图 3.10（b）所示。可以看到，器件在不同的限流和不同的 reset 电压下均可以将器件编程到不同的状态，表现出优异的缓变特性。这些结果初步表明该器件具有良好的模拟转变行为，适用于人工突触的实现。

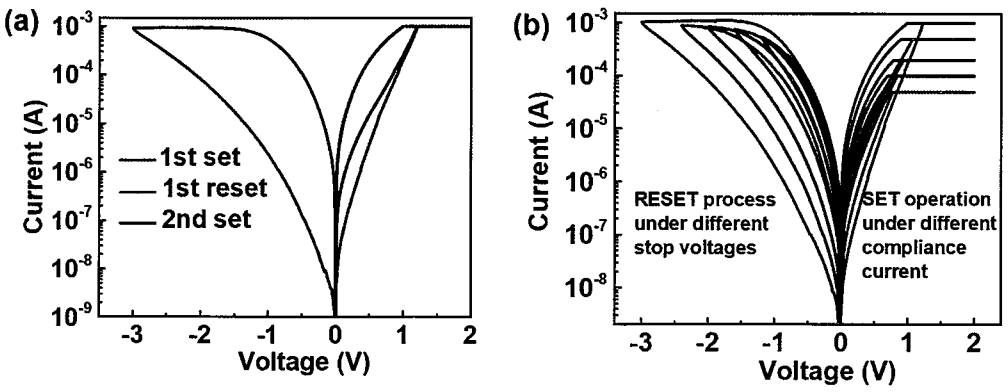


图 3.10 （a）Pd/HfO_2/WO_x/W 器件的第一个周期扫描曲线；（b）器件在不同限流和 reset 电压下的缓变特性

根据器件的 Pd/HfO_2/WO_x/W 结构和基本直流转变特性，我们提出了一个简单的器件模型来解释器件的转变机理。该器件可以等效为一个二极管串联一个可变电阻，如图 3.11（a）所示。这是因为高功函数的 Pd（5.12 eV）顶电极与 n 型 HfO_2 界面形成肖特基势垒，可以看成一个二极管。氧化过程中 WO_x/W 的底电极界面产生高浓度的氧空位，可以看成是准欧姆接触。底电极 W 氧化后形成导电 WOx 层。当 ALD 在 260℃下开始在 WO_x 薄膜上沉积 HfO_2 时，HfO_2 中的氧与 WO_x 表面的 W 发生氧化还原反应。当器件施加电压时，界面处的氧含量会发生变化，从而导致界面处电阻的变化，可以等效为可变电阻。图 3.11（b）给出了器件在初始态下的原理图，包括 HfO_2 势垒层，HfO_2 和 WO_x 界面处的转变区域，以及含有高浓度氧空位的 WO_x 导电层。当在顶部电极给器件施加正偏压时，氧离子会在电场的作用下向 HfO_2 区域移动，导致氧空位在 HfO_2/WO_x 界面堆积，使得界面处的导电性能更好，从而器件将转变到低阻态，如图 3.11（c）所示。图 3.11（d）给出了器件上施加负电压时的转变示意图，在电场作用下，氧离子向界面处移动，氧化界面处的材料，使得界面处的导电性变低，从而器件转变为高阻态。由于界面处的导电性能取决于氧空位的含量而该含量是一个连续的变量，所以器件表现出良好的缓变特性。

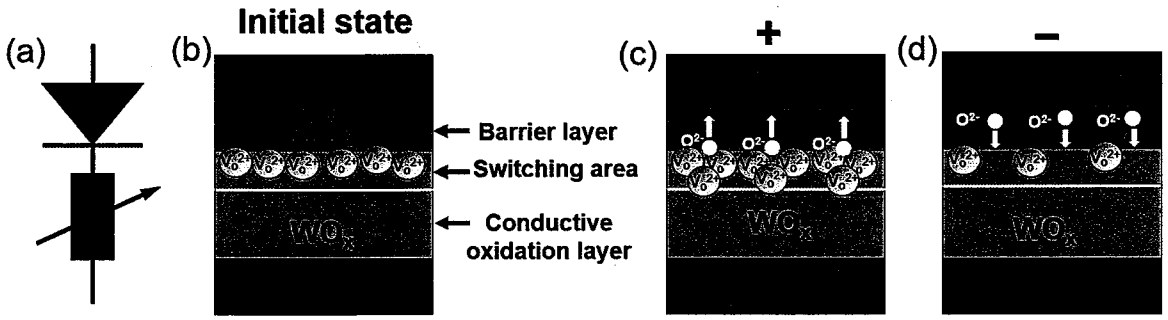


图 3.11 （a）Pd/HfO_2/WO_x/W 器件的等效结构；（b）器件初始态下的原理结构图；（c）器件在正电压阳离子移动示意图；（d）器件在负电压下阳离子移动示意图

为进一步验证器件的可靠性，我们对器件分别进行了两次 1000 个循环的直流扫描。如图 3.12 所示。在第一次 1000 个循环的连续扫描之后，器件表现出轻微的退化，这是由于连续电压扫描导致的器件在应力下的缺陷积累导致的。在进行第二次 1000 个循环的扫描之前，我们让器件静默 12 个小时，结果发现器件的第二次扫描的曲线几乎和第一次的情况一致，这说明器件具有良好的稳定性和自我恢复功能，这和生物突触的疲劳及恢复特性相似[14]。

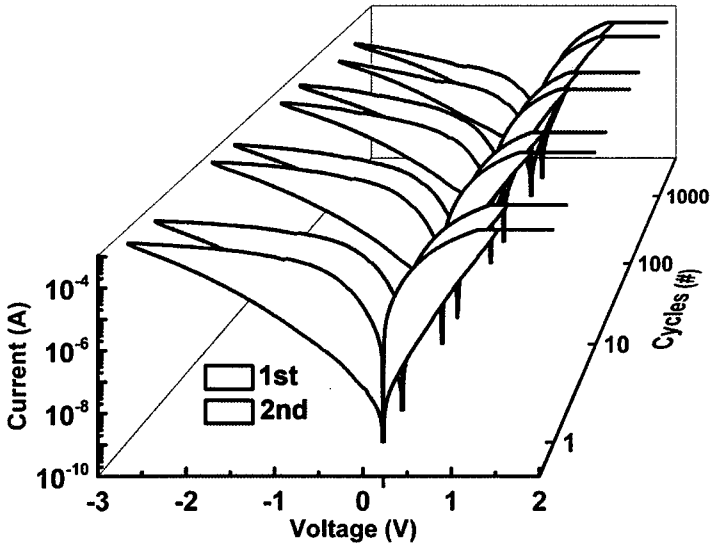


图 3.12 （a）Pd/HfO_2/WO_x/W 在两次 1000 个周期扫描下的曲线

3.2.2 Pd/HfO_2/WO_x/W 突触器件的脉冲编程方案及优化

为了评估器件的突触特性，我们对器件分别施加 500 个幅度为 1.2 V、1.4 V 和 1.6 V，宽度为 1μs 的脉冲，如图 3.13（a）所示。在每个电压下都可以对器件的阻态

进行连续编程，得到了良好的 LTP 曲线。随后，我们又验证了器件的 LTD 特性，分别施加 500 个幅度为-1.6 V、1.8 V 和 2.0 V，宽度为 1μs 的脉冲，测试结果如图 3.13（b）所示。另外，为进一步评估该器件在连续脉冲操作下的模拟转变行为，我们对器件的 LTP/LTD 过程分别施加 1000 个编程脉冲，如图 3.13（c）所示。可以看到，即使在 1000 个脉冲之下，器件的电导状态仍可以被连续地调节，且仍有继续增加的趋势。获得了非常宽的连续电导调制窗口（>300），该性能远远优于当前已报道器件的模拟调制窗口[28-30]，这种具有高度缓变的器件转变行为对于提高神经形态计算系统的容量和精度是非常有益的。通过对器件 LTP 和 LTD 特性的验证，可以发现该器件在脉冲操作下仍旧可以保持良好的缓变特性，是一种良好的突触候选器件。

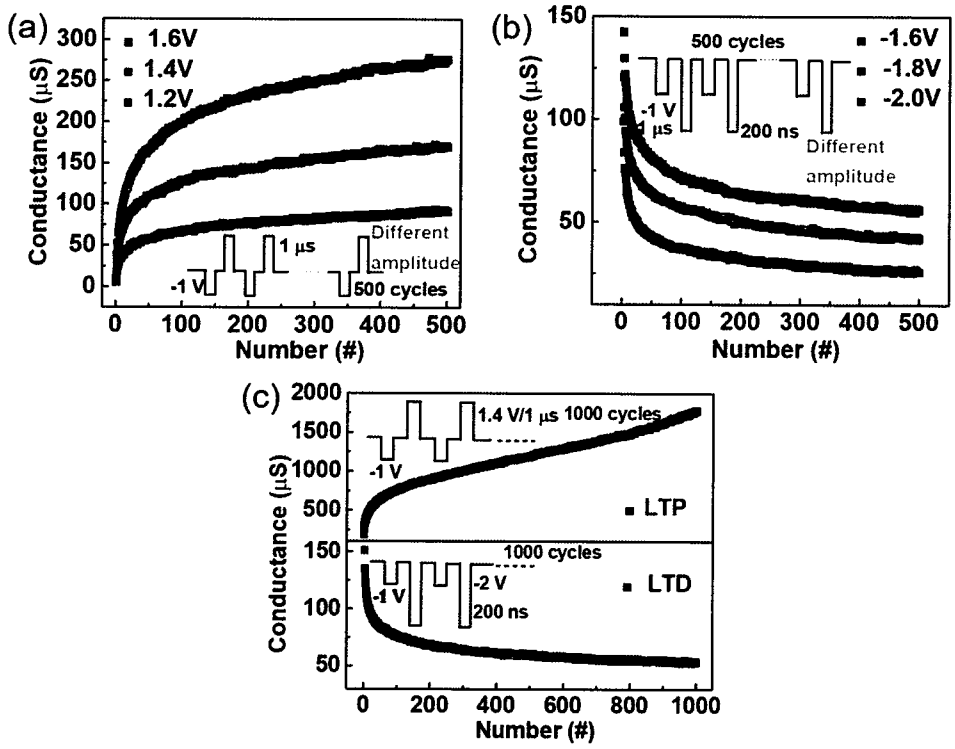


图 3.13 （a）$Pd/HfO_2/WO_x/W$ 器件在不同电压下的 LTP 曲线；（b）$Pd/HfO_2/WO_x/W$ 器件在不同电压下的 LTD 曲线；（c）$Pd/HfO_2/WO_x/W$ 在 1000 个连续脉冲下的 LTP/LTD 曲线

接下来我们又对器件在连续脉冲操作下的耐久性进行了测试，如图 3.14（a）所示。每个 LTP/LTD 周期包括 500 个振幅为 1.4 V（1 μs）的增强训练脉冲和 500 个振幅为-1.8 V（1 μs）的抑制脉冲。这里，由于测试时间的关系，我们只给出了 420 个周期的情况，对应着器件在 4.2×10^5 个连续脉冲下的结果，我们在后面的应用测试中又

进行了 1×10^6 的脉冲操作，可以看到，器件在连续脉冲下可以达到的最大电导呈现出稍微的退化趋势，这和直流扫描下的情况一致，因为连续脉冲的操作使得器件内产生了缺陷的积累。但根据直流的测试结果我们知道这种退化在一段时间的等待后可以恢复。为了清楚的看到器件在脉冲下的转变效果，我们在图 3.14（b）中给出了图 3.14（a）矩形区域的放大图，展示了 10 个 LTP/LTD 周期的情况。可以看到器件在 4×10^5 个连续脉冲操作后仍旧可以表现出良好且均一的 LTP/LTD 行为。

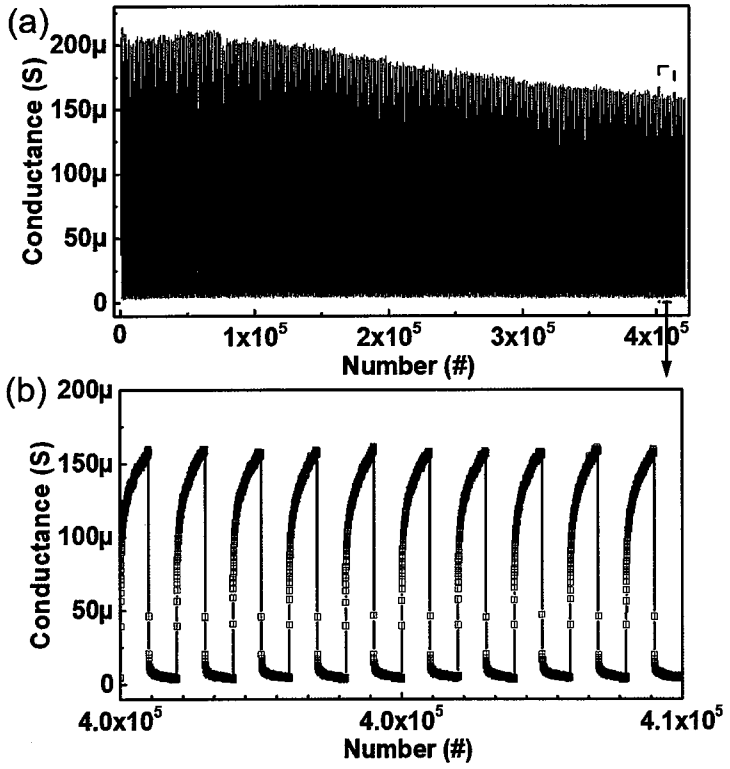


图 3.14 $Pd/HfO_2/WO_x/W$ 器件在连续脉冲操作下的耐久性测试

从图 3.13 中我们知道，器件在不同脉冲电压操作下具有不同的线性度。为了评估该突触器件在系统应用时对模式识别精度的影响，我们使用器件在不同电压操作下的 LTP/LTD 曲线验证了 MNIST 手写数据集在两层（$784 \times 100 \times 10$，图 3.15（a））神经网络上的学习效果。图 3.15（b）给出了网络在器件不同电压下对应不同线性度的 LTP/LTD 曲线的学习效果，可以看出与线性度为 1 的理想状态相比，器件的非线性 LTP/LTD 更新曲线会降低网络的学习性能，且线性度越低，网络的识别结果越差，与已报道文献中权重更新非线性度对网络性能影响的研究结果一致[12,28,31-33]。这是因为由于器件电导更新的非线性使得在网络训练过程中计算得到的权重误差被线性映射

到编程脉冲个数后不能线性调节电导的改变量[34, 35]。然而这对于线下的网络训练并没有什么影响，因为只需要将训练后的权值通过写验证的方式写入权重阵列中就可以了[36]。

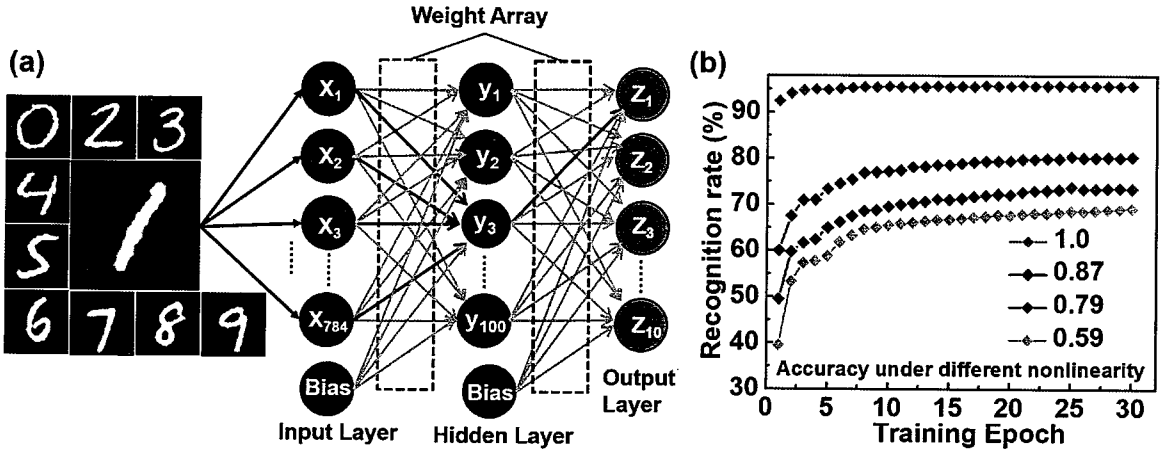


图 3.15 （a）用于学习 MNIST 数据集的两层神经网络原理图；（b）网络在不同线性度权重更新曲线下的识别结果

能够实现网络的在线训练是利用忆阻器实现突触的一大优势[34]，所以提高器件权重更新的线性度还是急需解决的问题。通过脉冲编程的方案提高器件的线性度是一种有效的解决方案，例如：变脉幅编程和变脉宽编程。在这里，为了提高器件电导更新的线性度，我们分别采用了以上两种方案进行了验证。图 3.16（a）和 3.16（b）分别给出了器件的变脉幅编程方案和测试结果图，可以看到器件在 LTD 过程呈现出准线性，而 LTP 过程表现出先慢后快的内凹非线性。这种内凹的非线性会使得网络迭代的过程比较慢，但最终也会编程到想要的电导值。另外，器件在变脉宽编程方案下的结果如图 3.16（c）和 3.16（d）所示，得到了更好的线性度。这说明该器件在系统应用时更适合于变脉宽的编程方案。

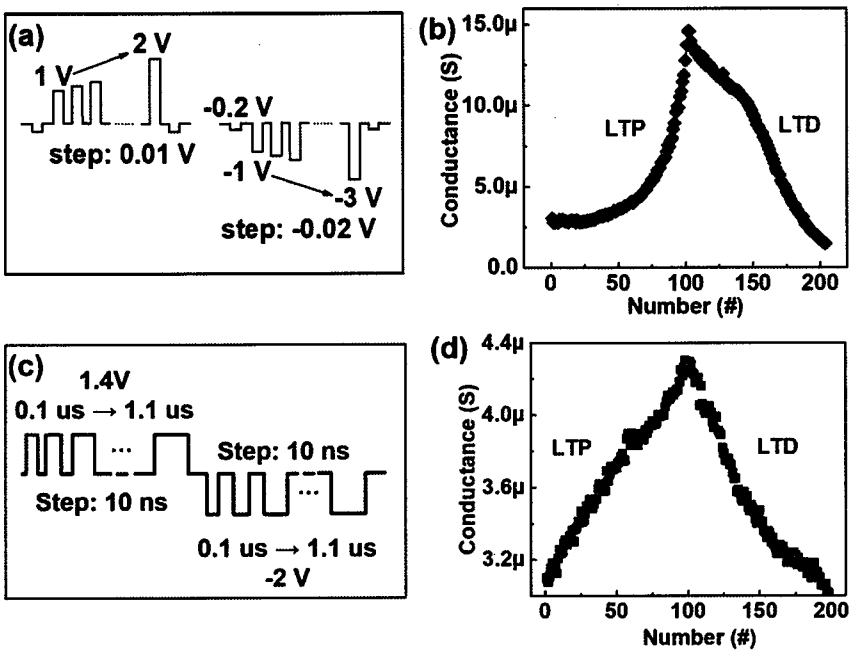


图 3.16 （a）变脉幅编程方案原理图；（b）变脉幅编程方案下 $Pd/HfO_2/WO_x/W$ 器件的 LTP/LTD 曲线；（c）变脉宽编程方案原理图；（d）变脉宽编程方案下 $Pd/HfO_2/WO_x/W$ 器件 LTP/LTD 曲线

接下里，我们分别利用图 3.16（b）和 3.16（d）的 LTP/LTD 曲线进行网络仿真，仿真结果如图 3.17 所示。两种编程方案下均可以得到接近理想情况的识别率，变脉宽的编程方案相对来讲会得到更快的训练效果和略高的识别率(95.3%)。该结果表明，通过改变脉冲编程方案，我们可以在 $Pd/HfO_2/WO_x/W$ 器件中实现接近线性的电导调制曲线，进而适用于网络系统实现。然而，需要指出的是，变脉宽和变脉幅的方案通常需要读取当前电导状态以便根据修改量确定脉冲参数，在一定程度上会增加外围电路的复杂度。因此，需要继续探索更优的编程方案。

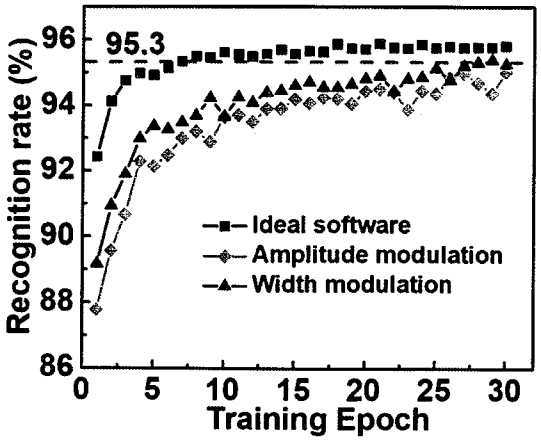


图 3.17 网络在变脉幅和变脉宽编程方案下的识别结果与理想情况的对比

3.3 本章小结

本章围绕着利用忆阻器的特性实现突触功能仿生开展了相关的研究工作，将生物突触的功能与器件的本征机理紧密联系起来。结合系统应用对突触器件的性能要求进行了探讨，进一步优化了器件设计，并通过系统仿真验证了器件电导更新的线性度对系统性能的影响，最后给出了合适的脉冲编程方案。为突触器件的设计和系统实现给出了理论指导。本章取得的成果如下：

1） 证明了在单个的 Cu/a-Si/Pt 忆阻器中，可以通过脉冲调制方案实现器件的易失性和非易失性电阻转变行为。基于该器件的电阻转变特性，我们成功地模拟了生物突触的长时程和短时程可塑性功能，包括：长时程增强、长时程减弱、峰时依赖可塑性、短时程增强、对脉冲易化、和短时程记忆到长时程记忆的转变。这些结果表明，Cu/a-Si/Pt 离子基忆阻器具有丰富的动态特性，适合进行突触仿生以便构建高效的类脑神经形态系统。该部分研究成果发表在 2017 年 9 月份的 IEEE Electron Device Letters 期刊上。

2） 通过叠层方案，制备了一种缓变且 CMOS 兼容的忆阻器用于神经突触的验证。该器件具有工作电压低（1.4V/-3V）、电导缓变窗口大（> 300）、耐久性强（$>10^6$）等优点。为提高电导编程的线性度，我们分别验证了变脉幅和变脉宽的编程方案，并进行了两层网络的系统仿真验证。在MNIST手写体数据集上得到95.3%的识别结果，接近理想值。这些结果表明 $Pd/HfO_2/WO_x/W$ 器件作为神经形态计算中的电子突触具

有巨大潜力。该部分研究内容在 2018 年 6 月份举办于夏威夷的硅纳电子研讨会（2018 Silicon Nanoelectronics Workshop）上做了口头报告。

参考文献

[1] Jo SH, Chang T, Ebong I, et al., Nanoscale memristor device as synapse in neuromorphic systems [J]. Nano Lett, vol. 10, pp. 1297-301, Apr 14 2010.

[2] Du C, Ma W, Chang T, et al., Biorealistic Implementation of Synaptic Functions with Oxide Memristors through Internal Ionic Dynamics [J]. Advanced Functional Materials, vol. 25, pp. 4290-4299, 2015.

[3] Wang Z, Joshi S, Savel'ev SE, et al., Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing [J]. Nat Mater, Sep 26 2016.

[4] Choi S, Tan SH, Li Z, et al., SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations [J]. Nat Mater, vol. 17, pp. 335-340, Apr 2018.

[5] Ohno T, Hasegawa T, Tsuruoka T, et al., Short-term plasticity and long-term potentiation mimicked in single inorganic synapses [J]. Nat Mater, vol. 10, pp. 591-5, Aug 2011.

[6] Kuzum D, Jeyasingh RG, Lee B, et al., Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing [J]. Nano Lett, vol. 12, pp. 2179-86, May 9 2012.

[7] Kim S, Du C, Sheridan P, et al., Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity [J]. Nano Lett, vol. 15, pp. 2203-11, Mar 11 2015.

[8] Burr GW, Narayanan P, Shelby RM, et al., Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: comparative performance analysis (accuracy, speed, and power), in 2015 Ieee International Electron Devices Meeting, ed New York: Ieee, 2015.

[9] Kim S, Ishii M, Lewis S, et al., NVM Neuromorphic Core with 64k-cell (256-by-256) Phase Change Memory Synaptic Array with On-Chip Neuron Circuits for Continuous In-Situ Learning, in 2015 Ieee International Electron Devices Meeting, ed New York: Ieee, 2015.

[10] Song KM, Jeong J-S, Pan B, et al., Skyrmion-based artificial synapses for neuromorphic computing [J]. Nature Electronics, vol. 3, pp. 148-155, 2020.

[11] Boyn S, Grollier J, Lecerf G, et al., Learning through ferroelectric domain dynamics in solid-state synapses [J]. Nat Commun, vol. 8, p. 14736, Apr 03 2017.

[12] Jerry M, Chen P-Y, Zhang J, et al., Ferroelectric FET Analog Synapse for Acceleration of Deep Neural Network Training, in 2017 Ieee International Electron Devices Meeting, ed New York: IEEE, 2017.

[13] Liu Q, Sun J, Lv HB, et al., Real-Time Observation on Dynamic Growth/Dissolution of Conductive Filaments in Oxide-Electrolyte-Based ReRAM [J]. Advanced Materials, vol. 24, pp. 1844-1849, Apr 2012.

[14] Purves D, Augustine GJ, Fitzpatrick D, et al., Neuroscience, 3rd ed. [M]. Inc. Massachusetts, USA: Sinauer Associates, 2012.

[15] Baars BJ and Gate NM, Cognition, Brain and Consciousness, 2nd ed. [M]. ELSEVIER, 2010.

[16] Chang T, Jo S-H, and Wei Lu W, Short-Term Memory to Long-Term Memory Transition in a Nanoscale Memristor [J]. ACS NANO, vol. 5, pp. 7669-7676, August 23 2011.

[17] Kim MK and Lee JS, Short-Term Plasticity and Long-Term Potentiation in Artificial Biosynapses with Diffusive Dynamics [J]. ACS Nano, vol. 12, pp. 1680-1687, Feb 27 2018.

[18] Li B, Liu Y, Wan C, et al., Mediating Short-Term Plasticity in an Artificial Memristive Synapse by the Orientation of Silica Mesopores [J]. Adv Mater, p. e1706395, Mar 15 2018.

[19] Li Y, Zhong Y, Zhang J, et al., Activity-dependent synaptic plasticity of a chalcogenide electronic synapse for neuromorphic systems [J]. Sci Rep, vol. 4, p. 4906, May 09 2014.

[20] Shouval HZ, Wang SS, and Wittenberg GM, Spike timing dependent plasticity: a consequence of more fundamental learning rules [J]. Frontiers in computational neuroscience, vol. 4, 2010.

[21] Prezioso M, Merrikh Bayat F, Hoskins B, et al., Self-Adaptive Spike-Time-Dependent Plasticity of Metal-Oxide Memristors [J]. Sci Rep, vol. 6, p. 21331, 2016.

[22] Yang R, Huang H-M, Hong Q-H, et al., Synaptic Suppression Triplet-STDP Learning Rule Realized in Second-Order Memristors [J]. Advanced Functional Materials, p. 1704455, 2017.

[23] Abbott LF and Regehr WG, Synaptic computation [J]. Nature, vol. 431, pp. 796–803, 2004.

[24] Woo J and Yu SM, Resistive Memory-Based Analog Synapses The pursuit for linear and symmetric weight update [J]. IEEE Nanotechnol. Mag., vol. 12, pp. 36-44, Sep 2018.

[25] Xia Q and Yang JJ, Memristive crossbar arrays for brain-inspired computing [J]. Nature Materials, vol. 18, pp. 309-323, 2019.

[26] Caporale N and Dan Y, Spike timing-dependent plasticity: a Hebbian learning rule [J]. Annual review of neuroscience, vol. 31, pp. 25-46, 2008.

[27] Froemke RC and Dan Y, Spike-timing-dependent synaptic modification induced by natural spike trains [J]. Nature, vol. 416, pp. 433-438, Mar 2002.

[28] Moon K, Cha, E., Lee D, et al., ReRAM-based analog synapse and IMT neuron device for neuromorphic system, presented at the 2016 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA), Hsinchu, 2016.

[29] Wang YF, Lin YC, Wang IT, et al., Characterization and Modeling of Nonfilamentary Ta/TaO_x/TiO_2/Ti Analog Synaptic Device [J]. Sci Rep, vol. 5, p. 10150, 2015.

[30] Govoreanu B, Piazza LD, Ma J, et al., Advanced a-VMCO resistive switching memory through inner interface engineering with wide (>10^2) on_off window, tunable μA-range switching current and excellent variability[M]. New York: Ieee, 2016.

[31] Woo J, Moon K, Song J, et al., Improved Synaptic Behavior Under Identical Pulses Using AlO_x/HfO_2 Bilayer RRAM Array for Neuromorphic Systems [J]. Ieee Electr Device L, vol. 37, pp. 994-997, 2016.

[32] Center IAR, Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element [J]. 2016.

[33] Park S, Sheri A, Kim J, et al., Neuromorphic Speech Systems using Advanced ReRAM-based Synapse, in 2013 Ieee International Electron Devices Meeting, ed New York: Ieee, 2013.

[34] Ambrogio S, Narayanan P, Tsai H, et al., Equivalent-accuracy accelerated neural-network training using analogue memory [J]. Nature, vol. 558, pp. 60-67, Jun 2018.

[35] Li C, Belkin D, Li Y, et al., Efficient and self-adaptive in-situ learning in multilayer memristor neural networks [J]. Nat Commun, vol. 9, p. 2385, Jun 19 2018.

[36] Li C, Hu M, Li Y, et al., Analogue signal and image processing with large memristor crossbars [J]. Nature Electronics, 2017.

第4章 基于忆阻器离子动力学机制的神经元电路研究

第一章论述到，在生物系统中，稀疏和异步的动作电位信号是以大规模并行的方式进行通信和处理的[1, 2]，这是生物神经网络可以高效的实时执行任务的重要原因。脉冲神经网络（SNN）在一定程度上继承了生物神经网络的优点[3]，因此，基于 SNN 的神经形态硬件具有低功耗、快速推理和事件驱动信息处理等特点[4-7]。这使得它们成为在当前大数据时代进行高效边缘信息处理的有力竞争者。为构建高效的神经形态芯片，忆阻器被广泛研究用于神经突触的实现[8-11]，对此我们在第三章给出了相关的研究和探讨。然而，除了神经突触外，脉冲神经元是另外一个至关重要的构成单元，它完成的主要功能是：（1）对通过突触加权的前神经元输入信号进行时空动态积分[12]；（2）以一定模式发放离散的动作电位信号[13]。国际上，利用 CMOS 器件实现脉冲神经元电路已经给了广泛的报道[14-16]。然而，由于 CMOS 器件缺乏与神经元相似的动态特性，因而需要复杂的电路设计来实现神经元的功能，这些 CMOS 神经元电路比忆阻器基的人工突触具有更低的能效和集成密度，故而很难实现和忆阻器突触的无缝集成。

鉴于此，为构建高效的神经元电路，近几年来，国内外学者开展了基于忆阻器神经元的研究工作，希望能利用忆阻器内在的动态特性实现神经元电路的行为特征。相关的工作涵盖了阻变器件[17-19]，铁电器件[20-22]，磁[20-26]和相变器件[27-30]研究领域。目前，基于这些新型器件的神经元工作大致可以分为两大类：（1）利用电容器实现神经元的积分功能，忆阻器作为动态阈值开关[27, 28, 30-32]；（2）对忆阻器器件转变过程进行控制，利用器件内部的积分特性实现神经元的积分功能[18, 19, 29, 33]。这些新型神经元电路的研究为脉冲神经元的实现开辟了一条新的途径。然而，由于该领域的研究处于初步阶段，在系统应用中神经元所能实现的功能还比较有限，因此，还需要相关研究人员进一步的探索和优化。

本章基于忆阻器离子动力学机制开展了神经元创新性的研究工作。首先，我们基于 Ag/SiO_2/Au 结构的易失性阈值转变（threshold switching, TS）忆阻器构建了一种新型漏电积分-发射（leaky integration-and-fire, LIF）神经元。在该神经元电路中，外部

电容器用来实现神经元的积分功能，TS 器件作为神经元的动态阈值开关，并用一个电阻器输出神经元信号。通过这种简单的电路结构，我们实现了脉冲神经元的四个关键功能：动作电位的全或无、阈值驱动放电、不应期和输入强度调制的频率响应。并通过系统仿真验证了基于该神经元的神经网络在数字识别中的可行性。另外，为进一步提高神经元的集成度，丰富神经元电路的功能，我们对 TS 器件进行了优化并提出了忆阻器-CMOS 混合设计的神经元电路。在该神经元电路的实现中，单个 TS 忆阻器作为后神经元的动态积分器，收集前神经元的输入信号，并决定是否产生动作电位输出。两个简单的数字 D 类锁存器检测放电事件并输出固定尖峰信号，同时通过提供不稳定周期信号确保忆阻器在连续脉冲串输入下能够正常工作。为了对忆阻器突触进行原位学习，成功地将增强、抑制和侧抑制信号引入到了神经元电路中，并设计了侧向抑制阵列（lateral inhibition array, LIA）。结合忆阻器突触，我们利用该混合神经元和 LIA，首次实验验证了一个 30×10×10 全硬件多层 SNNs。在该系统中，10 个隐藏层神经元利用 LIA 进行无监督学习，对第一层的输入模式进行预编码，10 个输出神经元在第二层以有监督的方式进行进一步识别。实验结果表明，该混合神经元能够对忆阻器突触进行原位学习，并具有构建高密度神经形态系统的潜力。

4.1 忆阻器中的离子动力学机制

在第二章中，我们从时间尺度上对忆阻器的动力学转变特性进行了分类和总结。由于金属导电细丝型忆阻器具有丰富的短时程离子动力学特性[34-36]，更符合神经元特性模拟的需求，我们选取了金属导电细丝机制的忆阻器作为实现神经元电路的核心单元。Z. Wang 等人[37]在 2016 年对该类器件易失性电阻转变行为的动力学特性进行了原位 TEM 表征并给出了解释[38]。其观测到的 Ag 细丝动态演变过程和响应的电学特性可以由图 4.1 解释。器件在初始状态时为高阻态，不存在完整的导电通路。当在 Ag 电极上施加正向电压激励时，电极中的 Ag 原子会发生氧化反应变成带有正电荷的 Ag 离子。然后，这些 Ag 离子会在电场的驱动作用下向阴极移动，并在移动的过程中被来自阴极的电子所还原成为 Ag 原子，被还原的 Ag 原子将发生沉淀造成银通道的生长[39]，对应图 4.1 中的①→②。随着氧化还原反应的持续进行，最终形成完整的银导电通路，使得器件由高阻态变为低阻态，对应 4.1 中的③。在介电层中，导电通路

的生长对剩余介质层内的电场形成正反馈[40]，器件表现为阈值转变现象。当撤去施加的电压或者施加的电压强度不够时（通常也称使器件不能维持低阻态的电压为保持电压），Ag 导电通路和周围介质之间的相互作用（界面能最小化或 Thomson-Gibbs 效应）导致了 Ag 导电通路溶解，直至破灭，对应 4.1 中的③→④，这种情况下器件自发的由低阻态变为高阻态。最终，导电通路中 Ag 原子团聚成稳定状态，对应 4.1 中的⑤。因此，该类型的器件便表现出易失性阈值转变特性。随后，W. Wang 等人[35]于 2019 年又通过表面限制自扩散机制对这种金属细丝机制忆阻器的寿命进行了系统分析，并指出相比铜导电细丝的忆阻器，银导电细丝的忆阻器寿命更短，因而更容易自发断裂，这是因为铜具有比银更大的表面激活能（铜的表面激活能约为 1.1 eV，银的表面激活能约为 0.52 eV）。基于此，我们在神经元电路的研究中采用了 Ag 导电细丝的忆阻器作为实现神经元功能的核心单元，并分别开展了利用忆阻器易失性阈值转变特性作为神经元电路阈值开关和利用忆阻器内部银通路生长动力过程实现神经元积分功能的研究。下面，我们将分别展开论述。

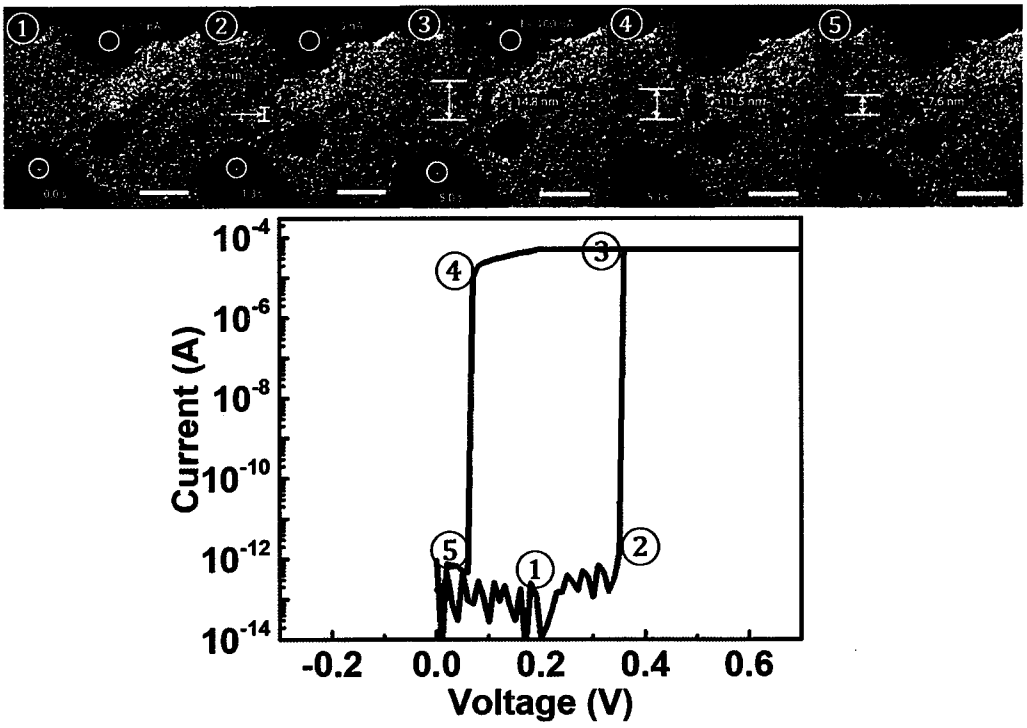


图 4.1 金属细丝型忆阻器中的动力学过程和电学特性[37]

4.2 基于忆阻器动态阈值开关特性的神经元电路研究

4.2.1 神经元电路的设计理念

生物神经元通过突触相互连接，当前神经元有动作电位产生时，通过连接的突触将信号传递到后神经元。然后后神经元通过树突和胞体对通过突触的前神经元信号完成积分功能，并在膜电位达到阈值时在轴丘处触发动作电位，继而产生的动作电位会通过轴突继续向后传播[2]。LIF 神经元模型是依据生物神经元通过细胞膜对电荷进行积累的过程而衍生出来的[41]。根据生物神经元的工作原理和 LIF 神经元模型，我们提出了基于 TS 忆阻器实现 LIF 神经元电路的方法，原理图如图 4.2。在这里，忆阻器阵列作为人工突触连接前后神经元，电容器作为积分单元接收并整合经过突触加权后的前神经元信号（1，2，3，4），TS 忆阻器作为动态阈值开关判断电容器上的电压是否达到阈值并给出动作电位输出。TS 忆阻器的初始态为高阻态，与电容器并联不容易泄露电荷。当电容器上累积的电位（膜电位）达到 TS 器件的阈值时，器件会打开变为低阻态，神经元会激发并输出脉冲。随后，由于电容器通过 TS 器件放电，会使得电容器上的电位降低，当电容器上的电位低于 TS 器件的保持电压时，器件又会自发回到高阻态再次充电。基于这样的工作原理，我们制备了 $Ag/SiO_2/Au$ 器件作为阈值开关并进行实验验证。

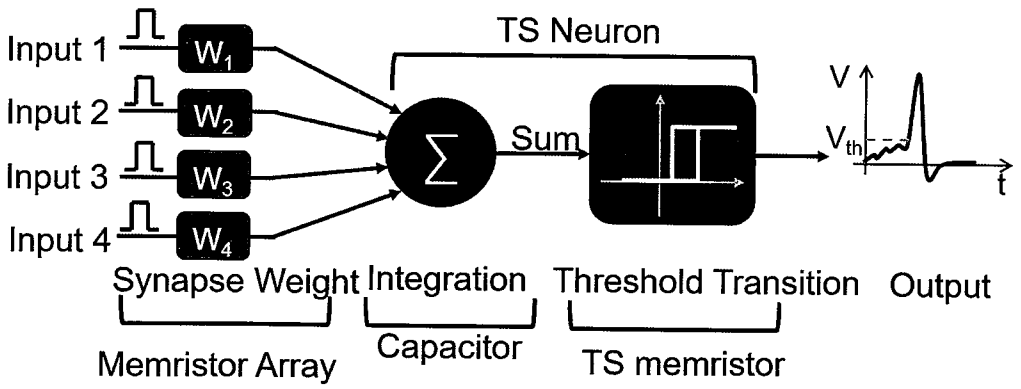


图 4.2 TS 忆阻器作为动态阈值开关的神经元电路原理图

4.2.2 器件的制备工艺和电学性能表征

本小节主要讲述本工作中所用到的 $Ag/SiO_2/Au$ 器件的制备过程，图 4.2 给出了具体的制备步骤：

(1) 经过第一步光刻后，在 SiO_2/Si 衬底上通过电子束蒸发沉积厚度分别为 40 nm 和 10 nm 的 Au 和 Ti 作为底电极，然后剥离形成垂直的下电极线，其中 10 nm 的 Ti 作为粘附层。(图 4.3（a）和（e）)

(2) 进行第二步光刻，磁控溅射淀积 30 nm 的 SiO_2 作为功能层，剥离形成中间层图形。(图 4.3 （b）和（f）)

(3) 进行最后一步光刻，然后通过磁控溅射分别淀积 40 nm 和 10nm 的 Ag/Au 作为上电极，其中 10 nm 的 Au 作为保护层防止 Ag 氧化。

(4) 剥离获得器件图形。(图 4.3（d）和（h）)

制备后的 Ag/SiO_2/Au 器件的 SEM 图像如图 4.4 所示。上电极与下电极的交叉部分即为器件单元的面积。除了面积为 5 μm × 5 μm 的器件，我们同时还制备了 1 μm × 1 μm、2 μm × 2 μm、3 μm × 3 μm、和 4 μm × 4 μm 的器件。在本工作中，我们主要基于 4 μm × 4 μm 和 5 μm × 5 μm 的器件进行实验操作。

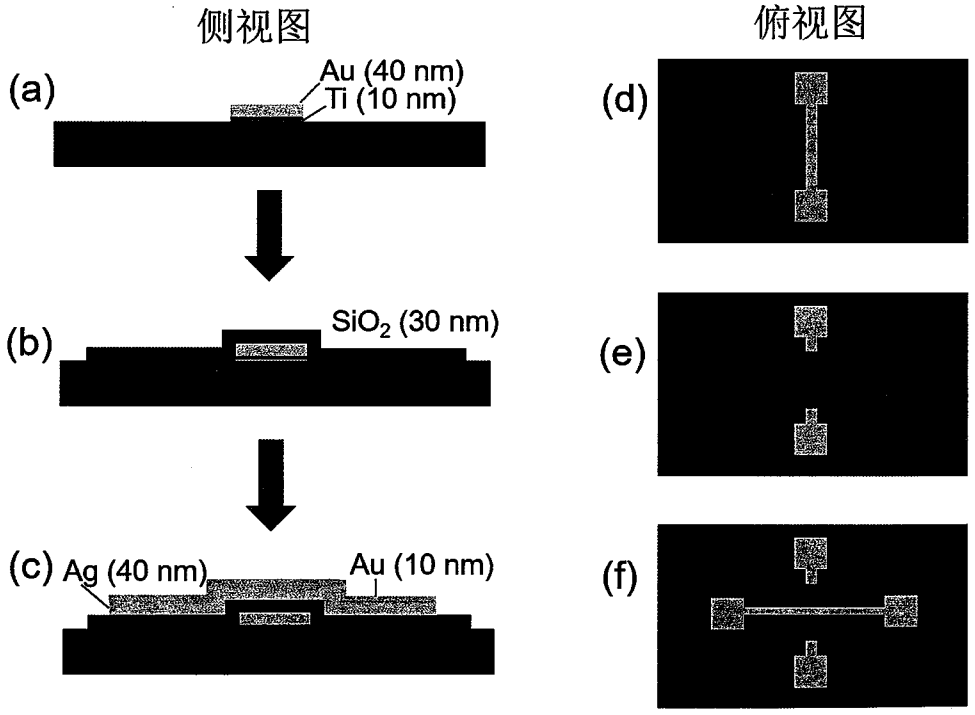


图 4.3 Au/Ag/SiO_2/Au/Ti 器件的制备工艺流程图

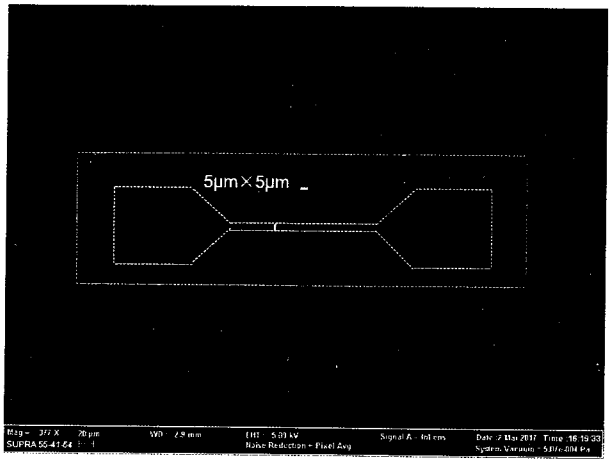


图 4.4 Au/Ag/SiO_2/Au/Ti 器件的 SEM 图像（5 μm × 5 μm）

随后，我们使用 Agilent B1500A 对器件的电学性能进行表征。测试时，电压激励施加在 Ag 上电极上，Au 下电极接地，如图 4.5（a）所示。图 4.5（b）给出了 forming 操作后器件在 100 μA 限流下的 I-V 曲线。可以看到，器件的初始态为高阻态，当施加的电压高于某一阈值电压（V_{th2}）时，通过器件的电流发生跳变达到限流，器件由高阻态（HRS）变为低阻态（LRS）。当施加的电压低于某一保持电压值（V_{th1}）时，该器件自发的由 LRS 恢复到 HRS。器件的这种易失性的阈值转变行为是因为在大电压下会在器件内部形成 Ag 的导电通路[39,42,43]，使得器件表现为低阻。当撤去电压或者器件上的电压足够小时，由于界面能最小化的作用使得器件上的电压不足以维持 Ag 导电通路的细丝形态，从而发生团聚使得细丝断裂[35, 37, 38, 44]，如图 4.6 所示，符合前面分析的 TS 器件的特征，因而可以用来验证所提出的 LIF 神经元电路。

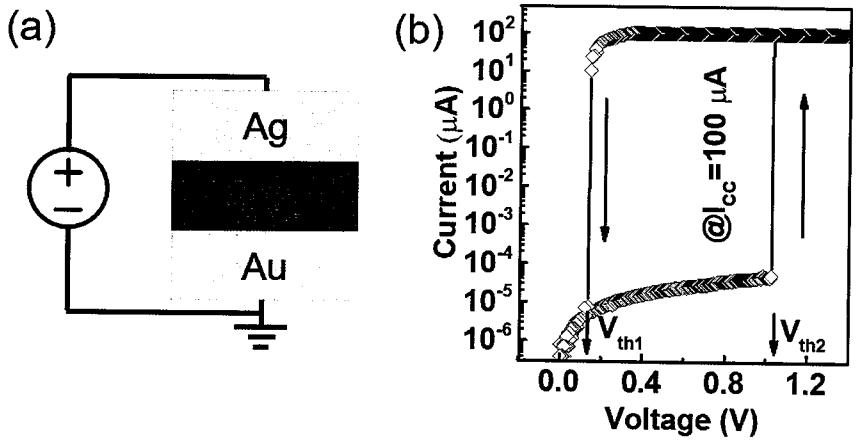


图 4.5 （a）器件的测试原理图；（b）Au/Ag/SiO_2/Au/Ti 器件的直流转变特性

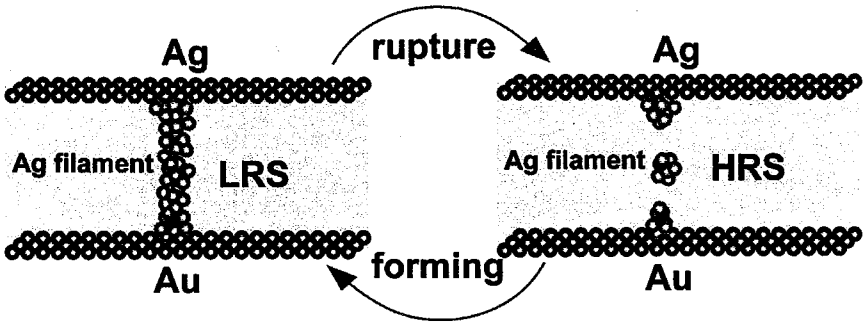


图 4.6 器件的转变机理解释

为进一步研究器件在外部刺激下的暂态响应，我们对器件进行了脉冲操作，如图 4.7 所示。在 1.2 V/5 ms 的刺激脉冲以外施加 0.1 V 的小电压用来实时读取器件的状态变化。可以看到，在脉冲刺激下，器件经过一段时间的延时（delay）由高阻态变为低阻态，在撤掉脉冲刺激后，器件会经过一段弛豫时间（relaxation）后由低阻态自发的回到高阻态。在这里，器件响应的延时是由 Ag 导电通道在 SiO_2 介质层中的生长过程导致的，而弛豫时间对应的是 Ag 导电通路的断裂和 Ag 原子的团簇时间。需要指出的是，器件的响应延时和弛豫时间都跟所加激励的大小有关。图 4.8（a）和（b）给出了器件在 0.8 V 和 1.6 V 脉冲幅度下的响应结果。结果表明，所加激励的脉冲幅度越大，则 Ag 导电通路的生长速度就会越快，对应的响应延时越短。而弛豫时间的变化则和响应延时相反，脉冲幅度越大，器件打开后在脉冲下的保持时间越长并且流过通道的电流越大，形成更加稳定的 Ag 导电通路，从而会导致更长的弛豫时间。不同激励电压下响应延时和弛豫时间的统计如图 4.8（c）和（d）所示。

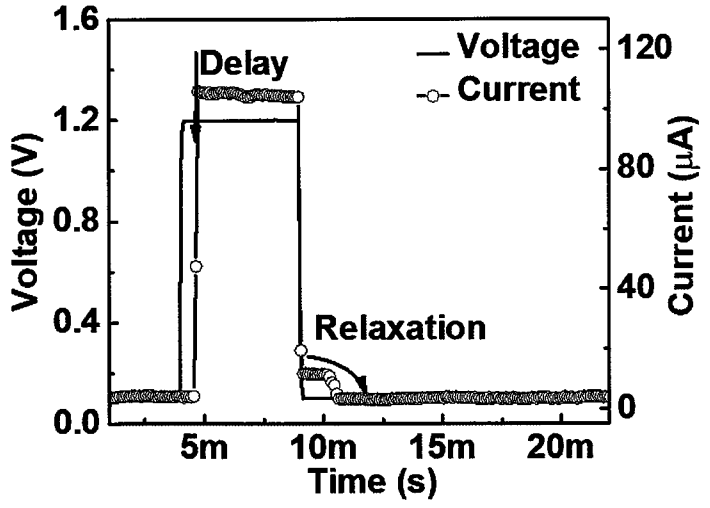


图 4.7 器件在脉冲刺激下的暂态响应

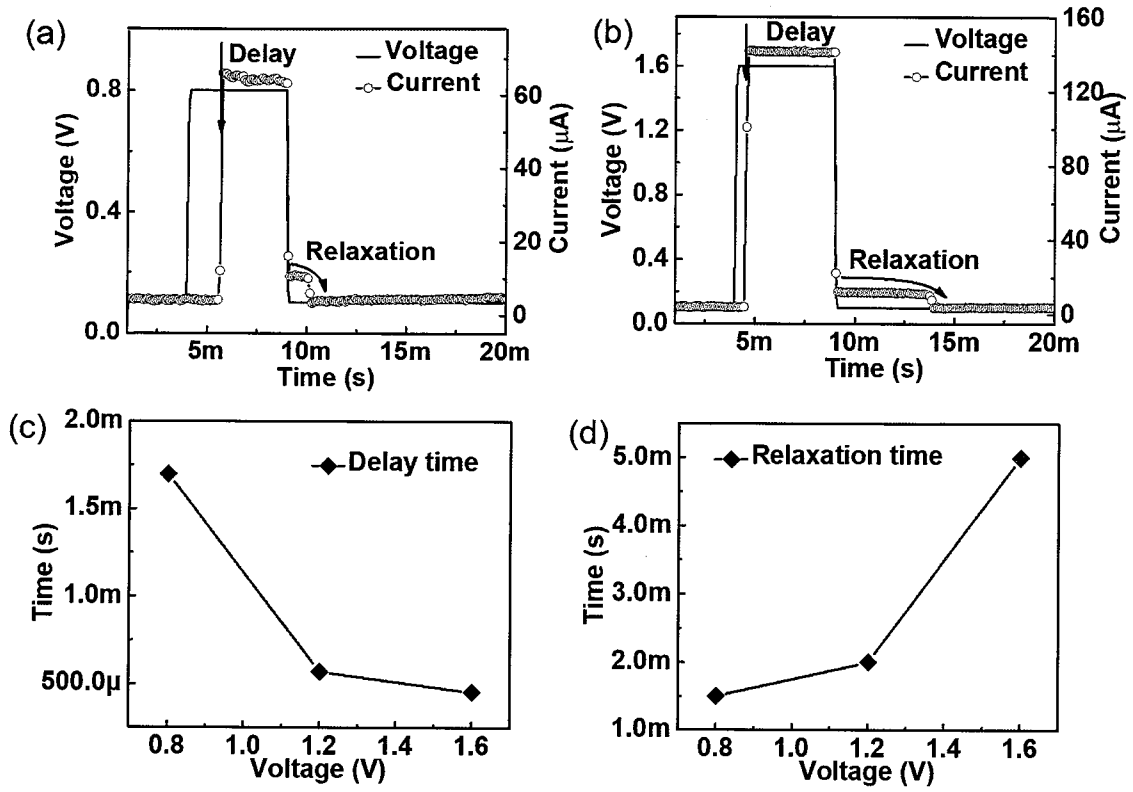


图 4.8 （a）&（b）器件在 0.8 V 和 1.6 V 脉冲刺激下的暂态响应；（c）不同刺激电压的响应延时统计；（d）不同刺激电压下的弛豫时间统计

4.2.3 TS 神经元电路及放电特性研究

基于上述制备的 TS 器件，我们使用器件作为阈值开关验证了前面提出的 LIF 神经元电路，电路细节如图 4.9 所示。我们采用一个固定电阻 R_o（$R_o = 51\ k\Omega$）作为输出电阻，TS 器件与该输出电阻串联，然后用电容器 C（C = 100 nF）与这两个元件并联构成 LIF 神经元电路。最后，神经元与突触电阻 R_S（$R_S = 510\ k\Omega$） 连接。作用于左节点（输入）的脉冲信号发生器作为输入信号，R_o 上的电压作为输出信号。在操作过程中，整个电路可分为充电回路（Charge loop, CL: 1-2-3-1）和放电回路（Discharge loop，DL: 2-4-3-2），该两个回路在图 4.9 中用红色箭头进行了标示。在构建电路时，我们使用探针台将器件的两个电极引出并连到带有其它电路元件的面包板上。在进行实验测试时，我们用泰克 AFG3102 脉冲发生器作为神经元的输入源，另一台泰克 DPO3032 示波器用来实时测量输入输出的信号。

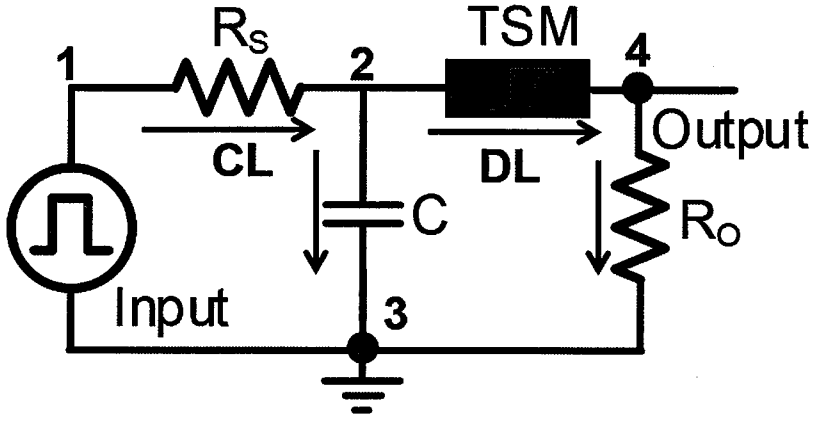


图 4.9 连接有突触电阻的 TS 基神经元电路原理图

随后，我们进行了神经元电路性能的测试。当由脉冲发生器产生的一连串频率为 100 Hz，幅度为 2 V 的电压脉冲（7 ms 脉宽）输入到电路的左侧节点“1”上时，电容器将通过 CL 回路积累电荷并且提升节点“2”的电位，如图 4.10（a）所示。在充电期间, TS 器件保持在高阻态（$R_H \sim 20\ G\Omega$），CL 中的 RC 时间常数（$\tau_C = R_sC$）远远小于 DL 中的时间常数（$\tau_D = (R_H + R_O)\ C$），这意味着电容通过 TS 器件泄露的电荷远小于通过充电回路对电容器充电的电荷量，因此电容器进行充电。当神经元电路节点“2”和“4”之间的电压达到或者超出 V_{th2} 时，TS 器件经过一段时间的弛豫后就会打开，从高阻态转变为低阻态。一旦器件处于低阻态（$R_L \sim 3\ k\Omega$），DL 中的 RC 时间常数（$\tau_D' = (R_L + R_O)\ C$）就会远大于 CL 中的 τ_C。在这种情况下，电容器进行放电，同时神经元将通过输出电阻 R_O 发射动作电位脉冲，如图 4.10（b）所示。由于电容器的放电，节点“2”和“4”之间的电压（即器件两端的电压）将降低。当电压降低到 V_{th1} 以下时，TS 器件将自发地恢复到高阻态，以准备下一次的放电活动。值得注意的是，在神经元放电期间，电容器的净电荷积累几乎为零。这是因为当器件打开时，任何输入电压脉冲通过 CL 对电容器的充电都会通过 DL 回路直接释放掉，这一特性可以模拟生物神经元的不应期。只有当电路节点“2”和“4”之间的电压小于 TS 器件的保持电压 V_{th1} 时，TS 器件才能恢复到高阻态，这时神经元的不应期结束，然后电容器再次进入充电状态为下次的放电行为做准备。图 4.10（b）给出了神经元的不应期和积分阶段，可以看到神经元的输出信号的时间长度大于输入脉冲的宽度，证明了不应期的存在。并且在积分阶段，神经元电路的输出电压基本为零，这说明放电结束后器件确实回到

了高阻态。这些结果表明，该 TS 神经元能够很好地模拟生物神经元的积分-发射功能和不应期，并且动作电位的输出具有“全或无”的特点。需要指出的时，神经元的最大放电次数与 TS 器件的耐久性相对应。根据报道，这种器件有优越的耐久性（$>10^8$ 开关周期)[45]，这意味着该神经元的最大放电次数可以超过 10^8 次。

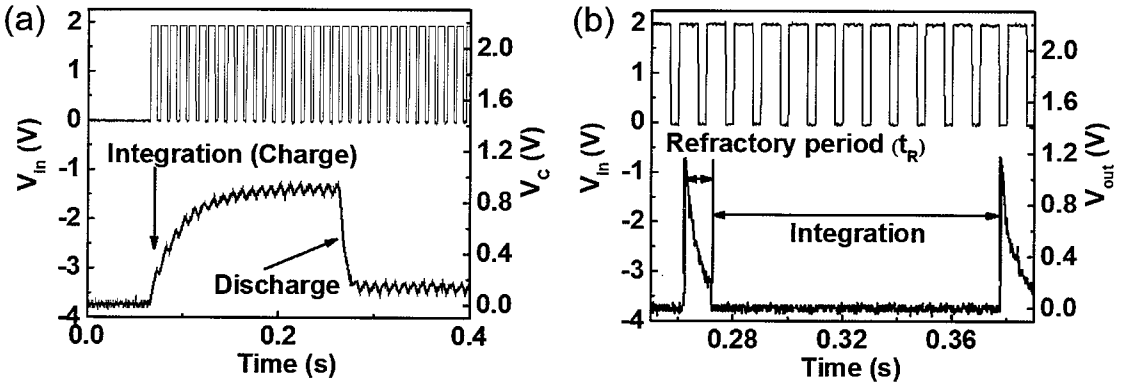


图 4.10 （a）电容器在连续输入脉冲下的积分效应；（b）神经元产生的动作电位信号以及对应的积分阶段和不应期

在相邻的脉冲之间，输入节点不施加电压，电容器主要通过输入回路（2-1-3）放电（CL 的泄漏可以忽略不计），图 4.11（a）给出了电容器在充电过程中的放大图，可以清楚地看出相邻脉冲之间神经元的泄漏效应。此外，当相邻脉冲之间的间隔增加时，打开设备所需的脉冲数增加，因为脉冲间隔时间变长导致泄露增多。当间隔增加到一定值时，电容器的充电和放电达到平衡，电容器上的电压不足以使得 TS 器件打开，也就不会产生动作电位了。图 4.11（b）给出了神经元产生动作电位的放大图，对应的不应期时间约为 10 ms，大于 4.8（d）中单个器件的弛豫时间，这是因为当器件打开后，电容器上的电荷会通过器件释放掉，在释放的过程中器件仍会有大于 V_{th1} 的电压，所以神经电路的不应期相对于单个器件来讲要长一些。

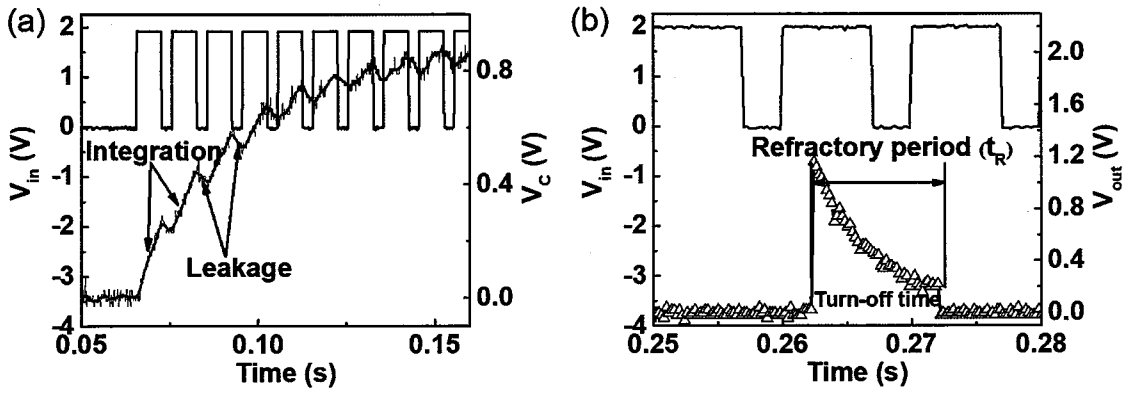


图 4.11 （a）积分过程中电容器的泄露效应；（b）神经元产生的动作电位的放大图

除了输入脉冲间隔会影响电容器对电荷的积累外，输入脉冲的幅度也会对其有影响。图 4.12 给出了电容器在 100 Hz，70%占空比的输入脉冲下不同的脉冲幅度对电容器积分特性的影响。可以看出，脉冲幅度越大，电容器的充电速度越快，从而需要的脉冲数目越少。在脉冲幅度比较小的情况下，电容器充电达到最大电压后还需要一定数量的脉冲输入才能将 TS 器件打开，这是因为在小电压下，TS 器件的延时比较大，在该过程中积分时间的长短很大程度上决定于 TS 器件的延时。随着脉冲幅度的增加，电容器所能充到的最大电压变大，对应能够施加在 TS 器件两端的电压会变大。根据前面的讨论我们知道器件在大电压刺激下延时会变短，因此所需要的输入脉冲个数会变少。并且随着输入脉冲幅度的增加，积分时间的长短逐渐主要由电容器的积分时间决定。

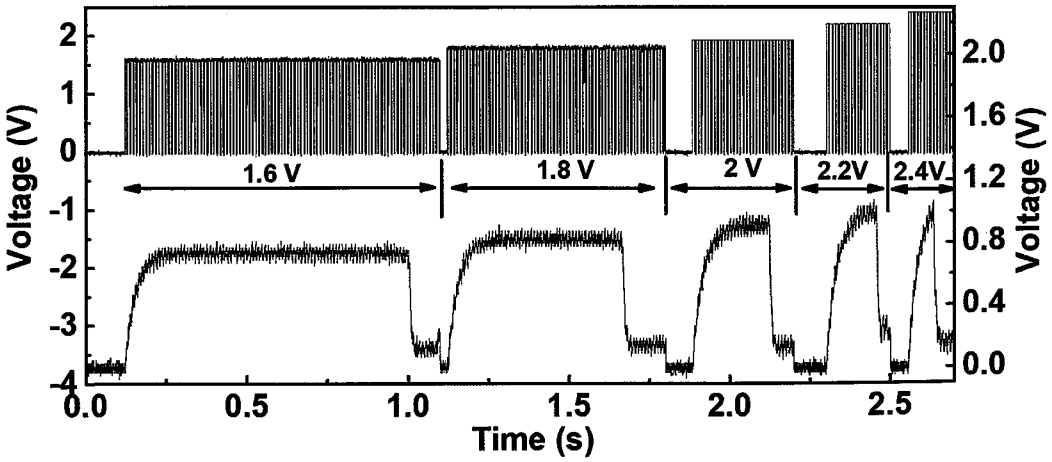


图 4.12 在不同输入脉冲幅度下电容器上的积分特性

在生物神经元中，动作电位的发射频率会随着刺激强度的增加而增加[1,46]。上述积分特性可以间接的体现神经元电路的放电频率。为了更直观的模拟生物神经元的这种强度调制的发射频率特性，具有不同振幅（1.2 V、1.4 V、1.8 V 和 2 V），相同频率（100 Hz）和相同占空比（70%）的一系列脉冲被施加到该神经元的输入端，如图 4.13（a）所示。可以直观的看到，随着输入脉冲振幅的增加，放电频率明显增加。图 4.13（b）和图 4.13（c）分别给出了七种不同输入脉冲下放电频率的统计和所需要脉冲数目的统计结果。这些结果表明，该 TS 神经元电路可以很好的模拟生物神经元在输入强度调制下的输出频率响应特性。

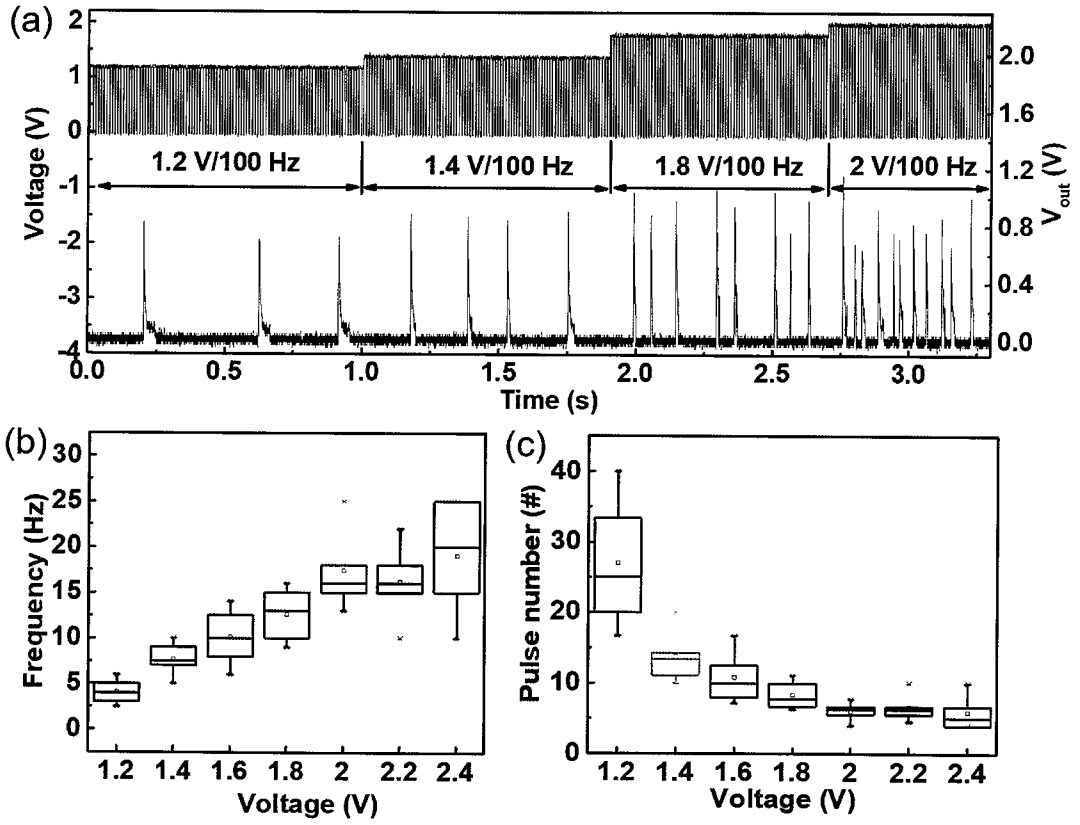


图 4.13 （a）器件在不同输入脉冲幅度下的放电特性；（b）不同电压下神经元放电频率统计；(c)不同电压下神经元放电所需脉冲数目统计

4.2.4 基于 TS 神经元的 SNN 网络仿真验证

为了验证 TS 神经元在脉冲神经网络中应用的可行性，我们提出了一种基于忆阻器突触阵列和 TS 神经元的数字识别系统。图 4.14 给出了这个系统的示意图，该系统为 30 × 10 的单层网络，输入图像具有 5×6 个像素点，输出为 10 个 TS 神经元，并且

在这 10 个 TS 神经元之间引入了侧向抑制功能[47]。黑色和白色像素点分别表示逻辑“1”和“0”，映射到输入电压脉冲分别为 (1 V、10 ms) 和 0 V。该网络只执行推理的过程，突触权重的训练过程在 MATLAB 上实现，训练后将突触的权值映射到忆阻器阵列中。当一个数字图像，例如数字“6”，被对应施加到 30 个输入节点上时，10 个 TS 输出神经元将积累输入的电流，由于侧向抑制功能的存在，最后只有神经元“6”放电，对应生物神经网络中“赢者通吃”的规则[47]。

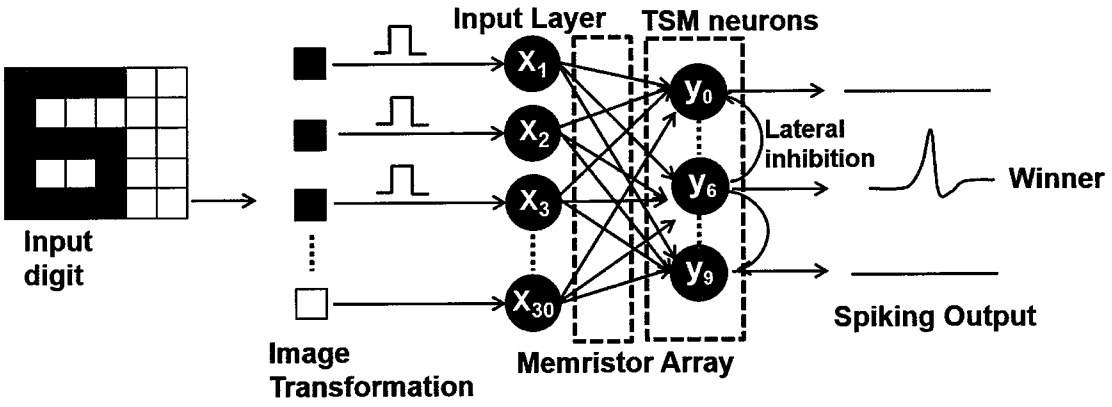


图 4.14 基于 TS 输出神经元的单层 SNN 网络示意图

图 4.15 给出了神经元“1”接收的来自其它神经元侧抑制信号的示意图。9 个 n 型 MOS 晶体管的栅极分别连接到其他神经元的输出端，并作为导通开关连接到神经元“1”中电容器的两个端口上。当神经元“6”首先产生动作电位输出时，该动作电位信号将输出到其他 9 个神经元对应连接的 MOS 管的栅极上，就会将这 9 个神经元的晶体管打开。一旦晶体管导通，就会短路连接到它两端的电容器，从而对电容器放电，将使得短路掉的神经元不会再产生输出脉冲，如图 4.16 所示。这里，开态晶体管用红色表示，而关态晶体管则用黑色表示。值得注意的是，在这里我们只是给出了神经元实现侧向抑制的一个简单的电路设计，若要实现 TS 神经元在输出层的大规模应用，还需进行进一步的优化研究。

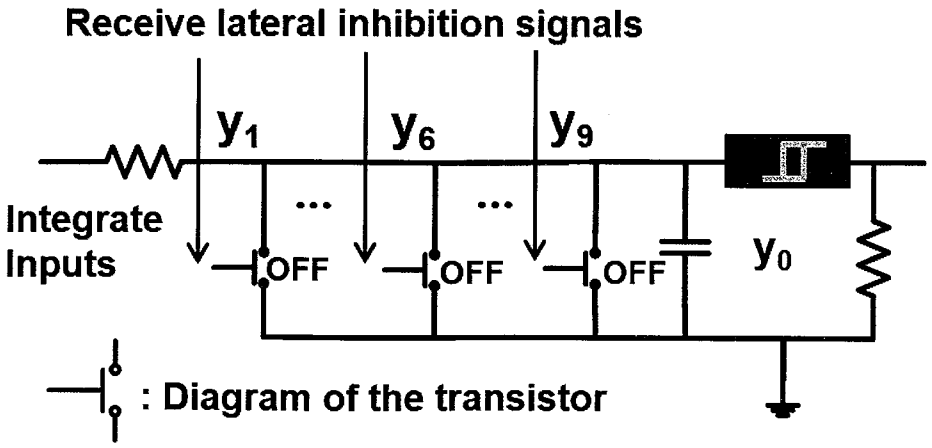


图 4.15 神经元“1”接收的来自其它神经元的侧向抑制信号

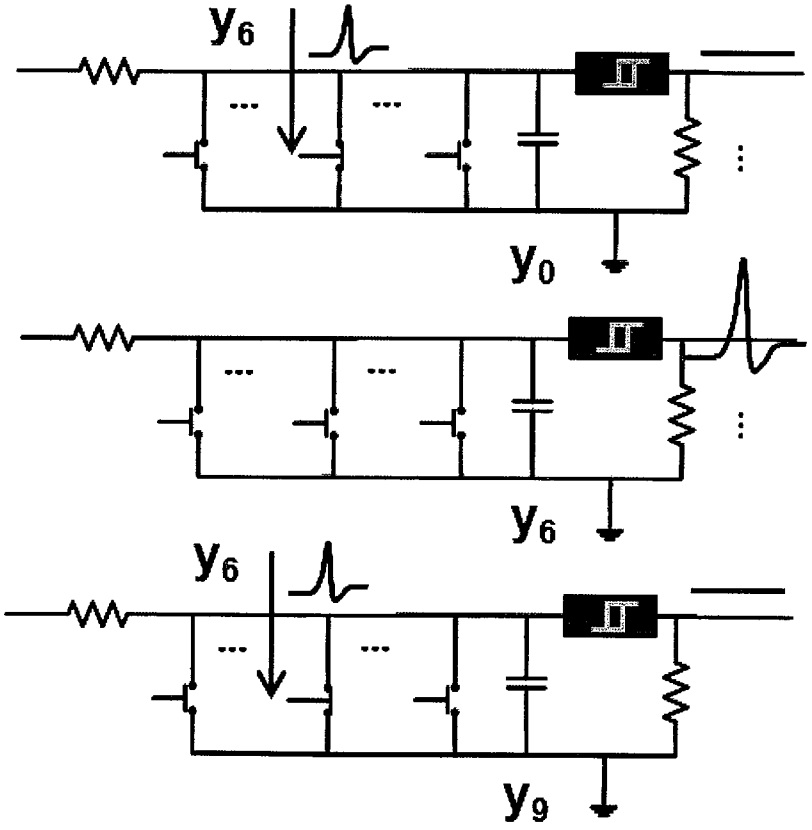


图 4.16 神经元“6”产生的动作电位信号对其它神经元的侧向抑制作用原理图

为进一步验证该网络对输入数字的识别结果，我们将不同的输入信号（从“1”到“9”再到“0”）依次输入到网络的输入层。图 4.17 给出的输出结果表明当数字图像被施加到输入节点时，只有目标神经元产生动作电位输出，而其他的神经元的输出则保持零。这些结果表明，该 TS 神经元作为输出神经元时可以成功地对输入模式进行识别，

从而适合于系统级的应用。

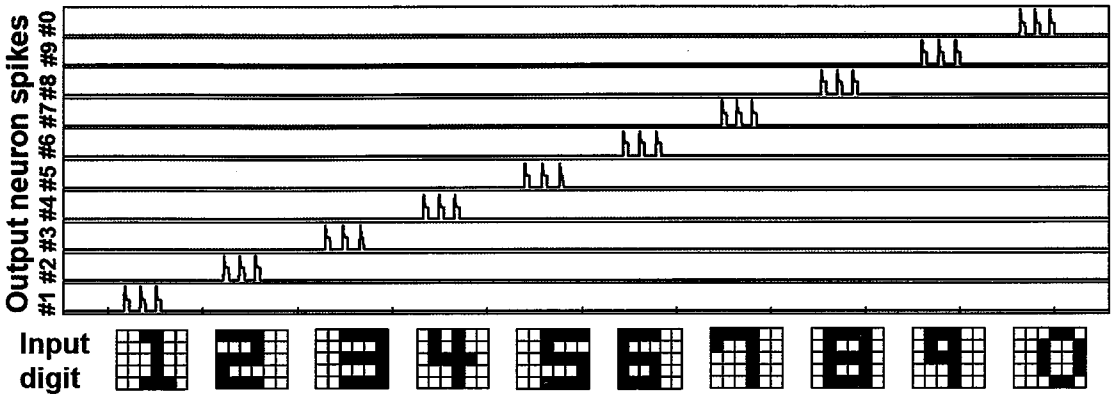


图 4.17 网络中输出层的 TS 神经元在不同输入模式下的输出结果

4.3 忆阻器-CMOS 混合神经元电路设计

基于前面工作的研究发现，器件的延时过程也可以被看成是介质内 Ag 原子在电激励下的积分过程。为进一步提高神经元的集成度并且丰富神经元电路的功能，我们提出利用器件内 Ag 细丝生长的过程来表征神经元电路的积分功能，以此去掉神经元电路中的电容器。在最近的几年里，已有课题组报道采用器件自身状态转变过程中的积分特性来实现神经元的功能。例如：IBM 的 T. Tuma 等人利用相变器件在连续脉冲操作下的电导变化来表示膜电位的变化[29]；Umass 的 Z. Wang 等人利用金属细丝的动力学生长过程来模拟神经元的漏电积分发射特性[19]；以及北京大学 C. Chen 等人利用铁电介质层中铁电极子的连续翻转过程来实现神经元的积分发射特性等[22]。然而，这些神经元电路实现的功能还比较单一，并且对系统应用中重要的神经元特性，例如在脉冲串刺激下正常的连续放电[13]以及可以直接对忆阻器突触执行原位编程操作[48, 49]等功能，还未得到系统的验证。

在这里，受生物神经元工作机制的启发，我们设计了一种忆阻器-CMOS 混合型 LIF 脉冲神经元电路，以弥补当前所报道的单个忆阻器神经元工作中的不足。

4.3.1 混合神经元电路的设计理念

通过第二章节的描述，我们了解到，在生物神经系统中，典型的神经元主要包括胞体（soma），轴突（axon）和许多树突（dendrites）。为进一步理解生物神经元的工作过程以开展后续的工作，在这里我们重申对生物神经元的描述并给出示意图，如

图 4.18（a）所示。树突与胞体一起接收和整合来自前神经元的兴奋或抑制信号以改变膜电位的状态[2]。一旦膜电位超过阈值，轴丘（axon hillock）就会通过打开或关闭压控离子门通道，使 K^+或 Na^+离子进入或流出细胞，产生一个完整的动作电位（具有“全或无”的特点）。然后轴突将产生的动作电位传递给其他相连的后神经元。放电后，细胞膜电位随膜内外 K^+和 Na^+离子浓度的恢复而恢复到静息状态。在放电过程中，神经元有一段不应期，在此期间神经元对继续到来的输入信号没有响应，不会再发出动作电位。前神经元的轴突末梢和后神经元的树突末梢（或胞体）形成突触，来自前神经突触的信号会通过突触间隙传送到后神经突触。突触的强度（可以释放的神经递质的数量）决定了信号从前神经元传递到后神经元的强度。重要的是，该突触强度可以根据突触前和突触后尖峰脉冲到来的相对时间进行原位调整，这就是生物神经系统中所谓的峰时依赖可塑性（STDP）学习规则[48, 50]。需要注意的是，由于生物神经元具有离子噪声、热噪声引起的电荷载流子的混沌运动以及背景噪声等，因而生物神经元在产生动作电位的过程中具有内在的随机性[51]。正是因为生物神经元的这种随机动力学，使得生物系统对外界噪声表现出很强的鲁棒性[52, 53]。

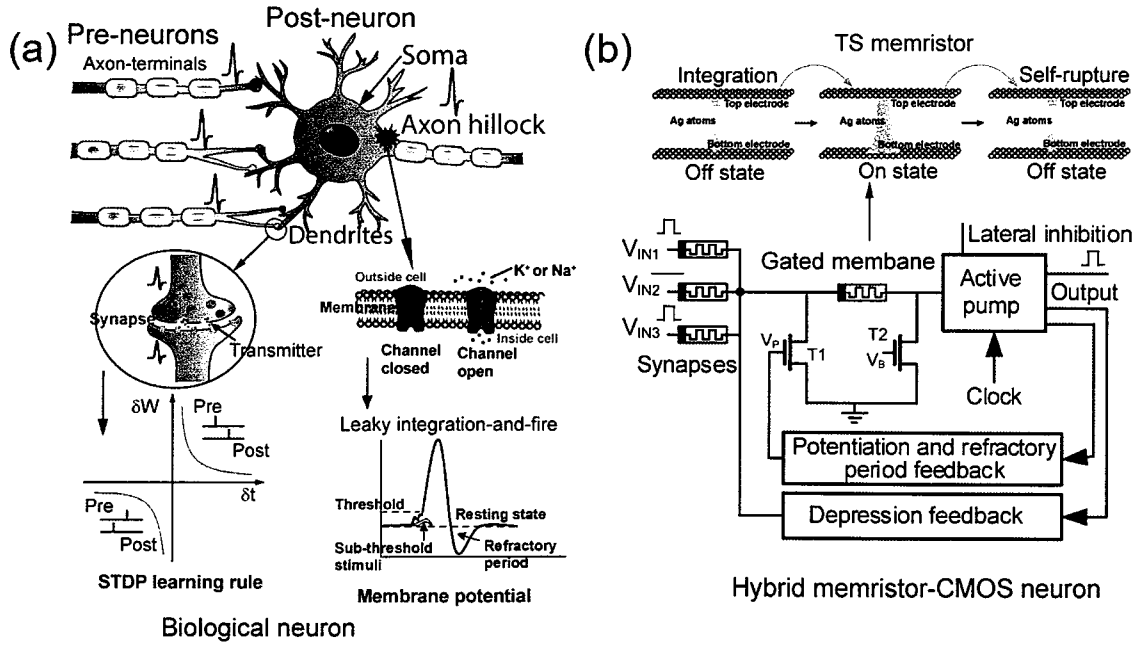


图 4.18 （a）生物神经元和突触原理图；（b）忆阻器-CMOS 混合神经元电路结构

在生物神经元结构及工作机理的启发下，我们设计了一个包含 TS 忆阻器和

CMOS 辅助模块的混合脉冲神经元，电路结构如图 4.18（b）所示。在该电路中，TS 忆阻器作为门控膜，通过 Ag 细丝的生长过程表征对输入信号的动态集成，并由器件内在的 TS 开关特性决定是否发生放电事件（从关闭状态突然切换到打开状态）。考虑到 TS 忆阻器固有的随机转变特性，该混合神经元也能有效地模拟生物神经元中存在的随机神经元动力学。这里，CMOS 辅助模块将 TS 器件产生的动作电位信号整形并给出稳定的输出脉冲，同时提供不应期反馈信号保证 TS 器件在连续脉冲串激励下能够连续正常的放电。在不应期内，由于离子动力学机制，器件可以自发恢复到初始状态，不需要任何复位操作。在 CMOS 辅助模块的帮助下，我们还适当地将增强、抑制和侧向抑制信号引入到了神经元电路中，以实现对忆阻器突触的原位编程操作，我们将在后面的介绍中给出具体的电路细节。

4.3.2 器件的制备工艺和电学性能表征

为更好的利用 TS 忆阻器实现神经元的功能，我们对器件的结构和工艺进行了优化。该小节主要讲述本部分工作中所用到的 Ag 掺杂 SiO_2 器件（Au/Ag/SiO_2:Ag/Au）的制备过程，具体的制备步骤如下：

(1) 经过第一步光刻后，在 SiO_2/Si 衬底上通过电子束蒸发沉积厚度分别为 40 nm 和 10 nm 的 Au 和 Ti 作为底电极，然后剥离形成垂直的下电极线阵列，其中 10 nm 的 Ti 作为粘附层。

(2) 进行第二步光刻，SiO_2 和 Ag 共同溅射 10 nm 作为功能层，剥离形成中间层图形。

(3) 进行第三步光刻,然后通过磁控溅射分别淀积 10 nm 和 40nm 的 Ag/Au 作为上电极阵列，其中 40 nm 的 Au 作为保护层防止 Ag 氧化。

(4) 进行最后一步光刻，然后通过电子束蒸发沉积厚度分别为 50 nm 和 10 nm 的 Au/Ti 作为电极 pad，然后剥离形成 pad 图形，其中 10 nm 的 Ti 作为粘附层。

(5) 剥离获得器件阵列图形。

这里，我们将电极线和 pad 分开制备是为了 bonding 过程中更稳定的连接。制备后的 Au/Ag/SiO_2:Ag/Au 器件的结构原理图如图 4.19（a）所示。为了获得 forming-free 的 TS 器件，我们在器件制备过程中采用了 SiO_2 和 Ag 共溅射的工艺实现掺杂并且将

介电层的厚度降低到 10 nm 以实现更低的操作电压。另外，为方便进行 bonding 以将其应用于系统，我们制备了包含 32 个分立器件的 TS 忆阻器阵列。图 4.19（b）给了阵列中 32 个离散器件的 SEM 图像。上电极与下电极的交叉部分即为器件单元的面积，每个器件的面积相同，为 4 μm × 4 μm。需要指出的是，该部分工作中所用到的突触器件是由美国马萨诸塞大学杨建华老师课题组提供的忆阻器阵列[8]，在此不对工艺细节进行过多的描述。

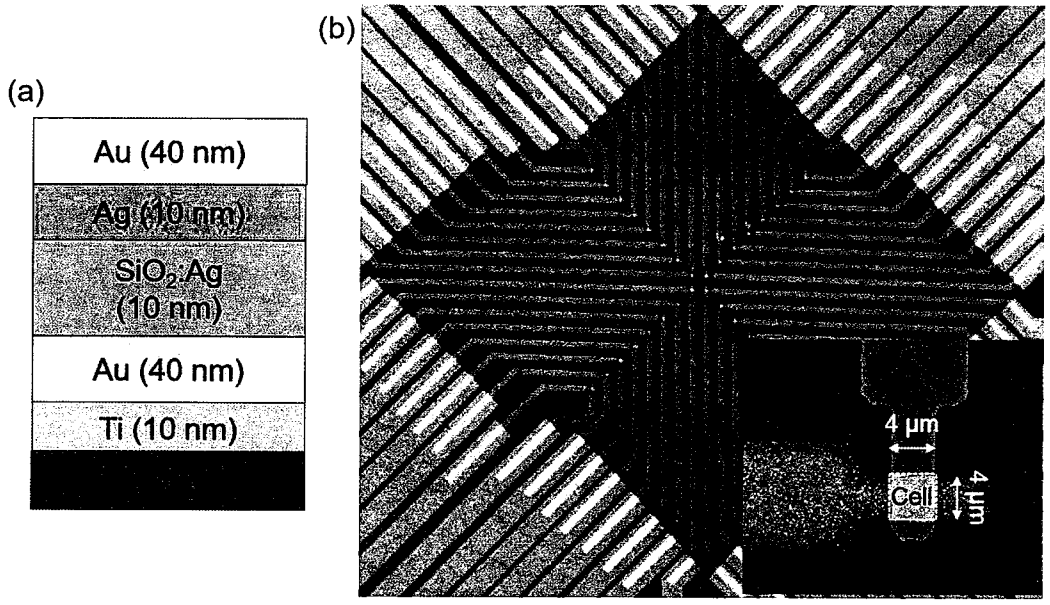


图 4.19 （a）Au/Ag/SiO_2:Ag/Au/Ti 器件结构；（b）离散器件阵列的 SEM 图像及单个器件的放大图(4 μm × 4 μm)

随后，我们用 Agilent B1500A 对单个 Au/Ag/SiO_2:Ag/Au 器件进行直流电学特性扫描。测试时电压激励施加在 Ag 上电极上，Au 下电极接地。图 4.20 给出了器件前 100 次正电压扫描下的易失性阈值转变 I-V 曲线，器件的初始状态处于高阻态(HRS)。可以看到相比于之前非掺杂 Ag/SiO_2/Au 器件，Au/Ag/SiO_2:Ag/Au 器件不需要 forming 操作，并且阈值电压降到了 0.4 V 左右。这是因为掺杂的 Ag 原子会在原始器件的介质层内形成团簇，在电压激励下起到局域电场增强的作用并且作为银通道的构成部分[54]。直流电学特性初步表明该器件转变过程中具有丰富的短时程离子动力学，可用于神经元功能的模拟。

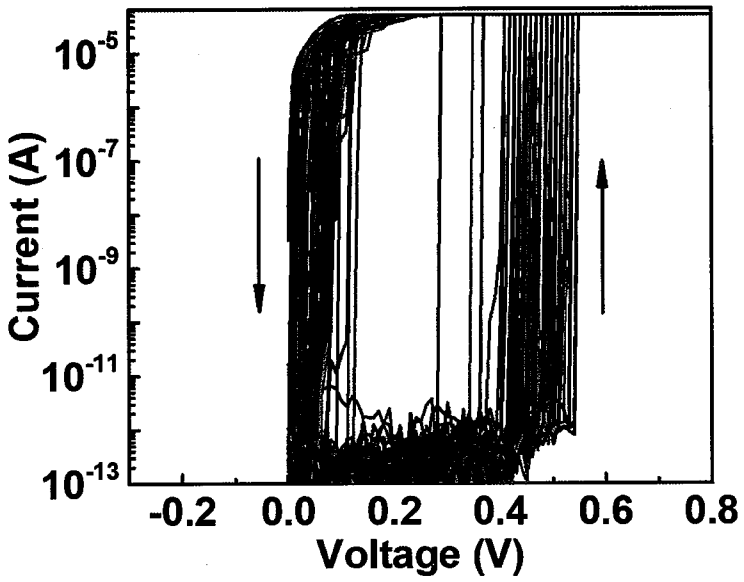


图 4.20 Au/Ag/SiO_2:Ag/Au/Ti 器件直流转变曲线

在生物神经元中，动作电位的发放具有统一的随机神经动力学特性，这在信号编码和组合优化问题中起着关键作用[51]。国际上普遍认为 Ag 细丝的生长和断裂过程也具有随机的内在物理动力学[54, 55]，这导致了器件阈值电压和保持电压分布的随机性，因此每个周期之间的转变电压应遵循某一概率函数分布。为在器件中验证该理论，我们随机选取了 10 个器件进行了直流扫描，每个器件扫描 100 个循环。图 4.21（a1-a10）给出了不同器件的直流扫描曲线和转变电压累积分布情况，可以看到每个器件均具有良好的阈值转变特性，并且器件的阈值转变电压呈现出一定的波动性。随后，我们对这些器件的转变电压进行了统计并与伽马分布曲线进行了对比，如图 4.21（b）所示，结果表明器件的阈值转变电压是随机的，并且遵循统一的伽马分布，这证实了该器件可以内在的实现生物神经元中动作电位的随机动力学行为。接下来，我们在图 4.21（c）中给出了对每个器件的阈值电压分布之间的比较，可以看到每个器件的平均阈值电压分布在 0.43 V 左右，这意味着阵列中的器件之间的均一性良好，适合于系统应用。

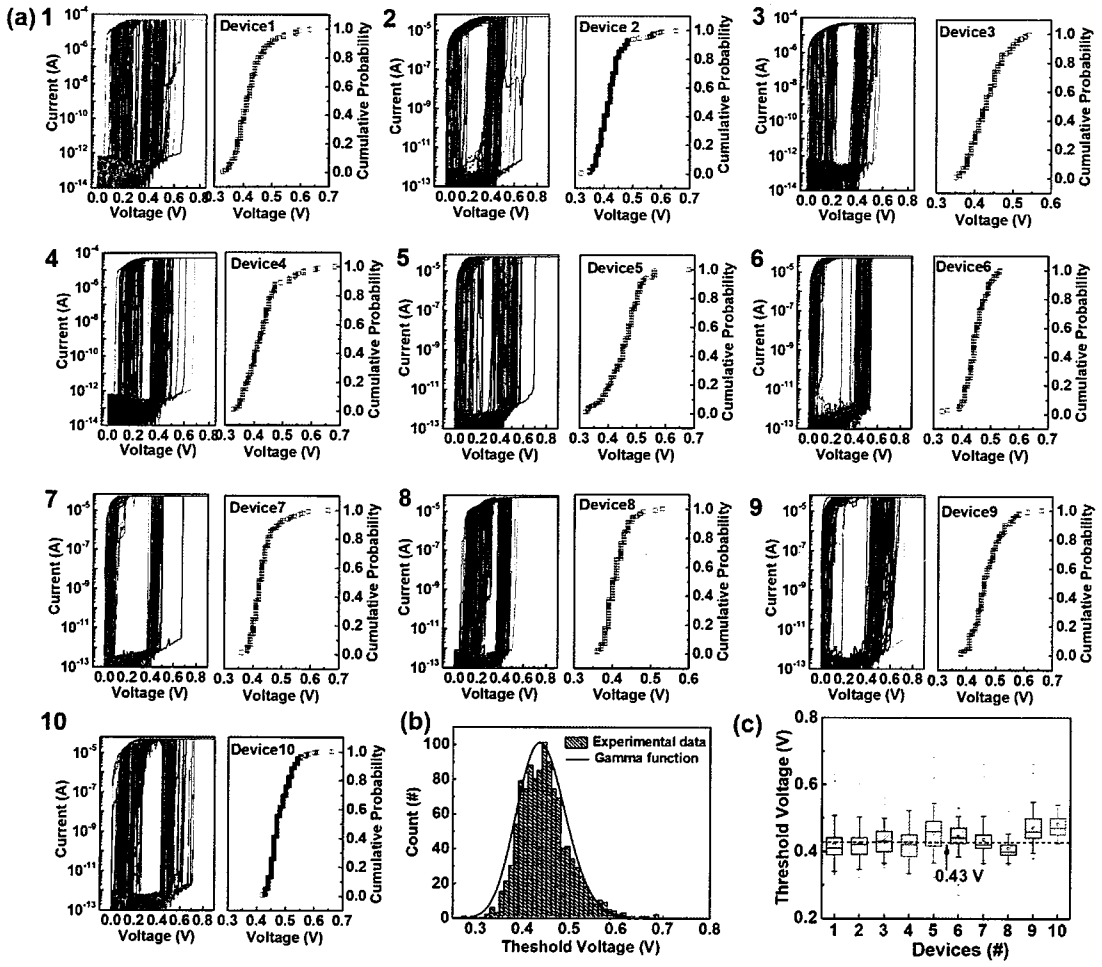


图 4.21 (a) 10 个不同器件的直流曲线和阈值电压累积分布；（b）阈值电压统计与伽马分布曲线的对比；（d）不同器件之间阈值转变电压的比较

为了进一步研究利用该器件内在的积分过程模拟 LIF 神经元特性的可行性，我们利用脉冲对器件进行操作。在测量过程中，将器件与一个 n 型 MOS 管（BS170）相连，该 MOS 管一方面用于匹配所设计的神经元电路中的部件 T2，一方面用于限制电流来保护 TS 器件，提高器件的寿命。测试原理如图 4.22（a）所示。用 Agilent B1500A 的 SMU 模块给晶体管的栅极施加恒定电压（1.95 V），然后通过 Agilent B1500A 的 WGFMU 脉冲模块施加脉冲。图 4.22（b）给出了器件在 1.2 V/1 ms 刺激脉冲和 0.05 V 读取电压下的动态响应结果。在刺激时间范围内，器件经过一段时间的延时后由初始的高阻态变为低阻态，对应响应电流突然增加达到晶体管的限流。这是因为在这段延迟时间内，Ag 原子在氧化还原反应和电场驱动的作用下，在 SiO_2 介质中逐渐迁移，最终形成一个 Ag 通道，使得器件转变为 LRS，表现出神经元的积分和发射特性，该

延时时间对应的是积分时间。当所施加的刺激电压撤去后，器件的响应电流逐渐变小，经过一段弛豫时间后自动恢复到 HRS 状态。器件的这种自挥发电导响应特性与生物神经元膜电位的“泄漏”特性相对应，允许神经元在积分脉冲输入间隔的泄露以及神经元放电后自动恢复到静止状态。

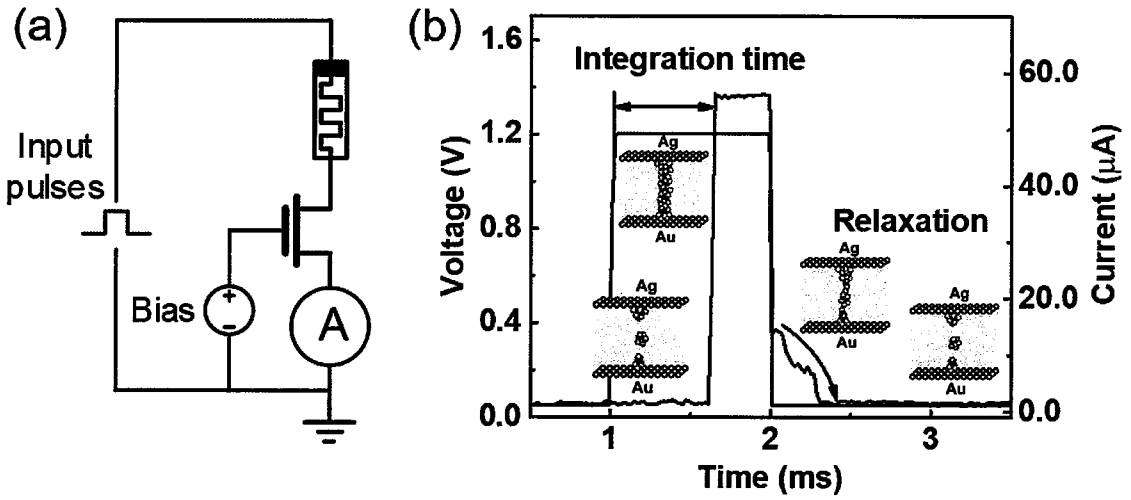


图 4.22 （a）脉冲测试原理图；（b）器件在脉冲刺激下动态的电学响应和通道形貌示意图

器件的延时和弛豫时间与刺激电压的强度相关。为了研究脉冲幅度对积分时间和弛豫时间的影响，我们在器件上施加具有固定宽度（1 ms）和不同幅度的脉冲，并对不同电压下的延时和弛豫时间进行了统计，统计结果如图 4.23 所示。结果表明，当脉冲幅度从 1.0 V 增大到 1.4 V 时，积分时间的平均值逐渐减小，弛豫时间则相反。换句话说，当流过突触后的前神经元的动作电位强度较高时，激发突触后神经元所需要的时间则越短，反之则相反，这与生物神经元中的情况类似。由于 Ag 导电通路生长和断裂过程的随机性，器件在不同幅度脉冲输入下所需的积分时间和弛豫时间均呈现随机概率分布。图 4.24（a）和（b）分别给出了积分时间和弛豫时间与伽马分布函数的对比，这说明器件在脉冲激励下也可以表现出与生物神经随机动力学相关的随机放电行为。

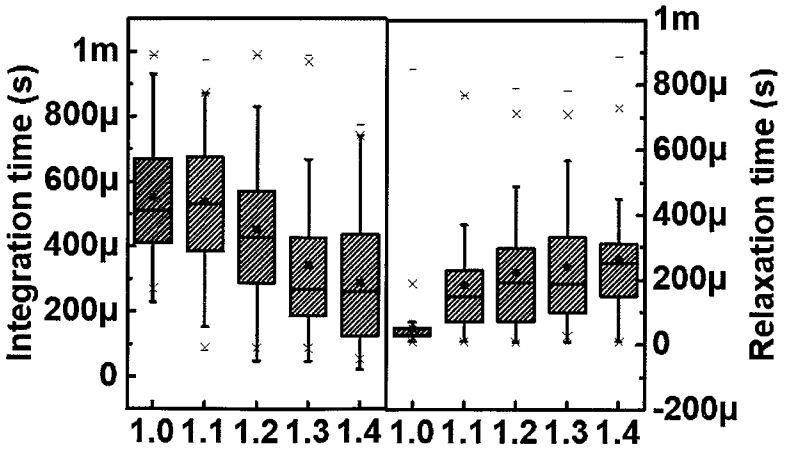


图 4.23 器件在不同脉冲幅度刺激下积分时间和弛豫时间的统计比较

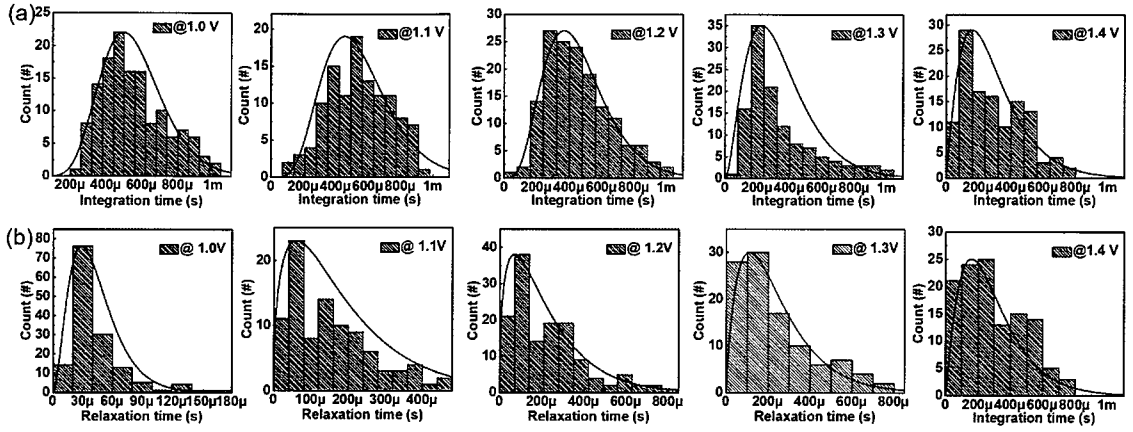


图 4.24 （a）器件积分时间在不同幅度脉冲下的分布情况；（b）器件弛豫时间在不同幅度脉冲下的分布情况

接下来，为了验证器件在连续脉冲操作下的 LIF 行为，我们以较短宽度（250 μs）和时间间隔（250 μs）的脉冲作为输入信号，如图 4.25（a）所示。可以看到，需要四个脉冲的连续刺激才可以使得器件发生第一次放电事件，这体现了多个脉冲刺激的 LIF 过程。图 4.25（b）给出了不同幅度的脉冲刺激下发生第一次放电事件的统计结果，在较高脉冲幅度刺激下需要的积分脉冲数较少。据此，我们可以得出结论，神经元的放电频率可以通过突触后动作电位的强度来调节，其中突触后的电位强度取决于连接的突触权重。需要指出的是，刺激脉冲之间的间隔时间必须短于器件的弛豫时间，这样才会体现出积分过程，不然前一个脉冲刺激产生的积累效果会在后一个脉冲到来之前全部泄露掉，器件就不会产生放电行为。然而，在满足脉冲间隔时间（250 μs）

小于器件的弛豫时间的情况下，当器件发生放电事件后，在下一个输入脉冲到来之前，器件不足以恢复到其初始高阻态，产生亚阈值放电行为，如图 4.25（a）所示。这表明简单的 TS 器件在脉冲串激励下不能进行连续且独立的 LIF 行为，这是利用单个忆阻器积分特性实现神经元电路的工作中普遍存在的问题。若要克服这个问题，便需要给器件提供一个不应期，使器件能够在不应期内自发恢复到其高阻态。因此，在这项工作中，我们在混合神经元电路中引入不应期反馈信号来解决这个问题（更多细节将在图 4.26 中给出）。

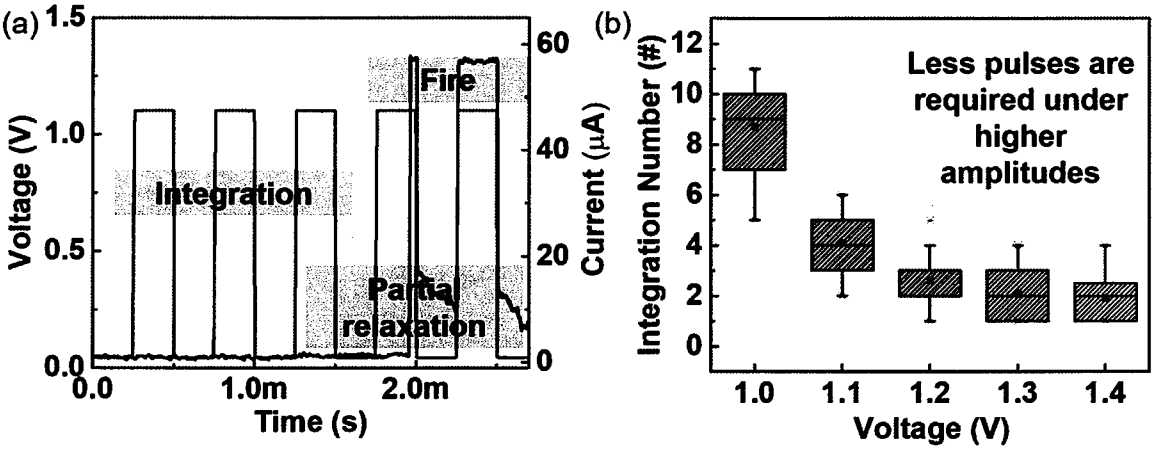


图 4.25 （a）器件在短脉冲输入下的积分放电行为；（b）不同脉冲幅度激励下器件发生第一个放电行为所需要的脉冲个数统计

4.3.3 混合神经元电路原理分析

图 4.26 给出了本部分工作中提出的混合神经元电路的原理图。CMOS 辅助模块包括有源泵和抑制操作模块。其中有源泵包括两个 D 型锁存器（L1 和 L2）、一个与门（G1）和一个或门（G3）。G1 产生输出尖峰信号，G3 产生侧向抑制信号。L2 的输出同时作为增强和不应期信号来控制晶体管 T1。L1 的输出触发由一个与门（G2）、一个驱动器和一个开关晶体管（T3）组成的抑制操作反馈电路。在初始状态时，晶体管 T1 处于关闭状态，可视为开路，节点 1 的电位是突触后动作电位。当 TS 器件在输入刺激下打开时，会触发 L1 产生高电平输出作为 L2 的输入，然后 L2 产生高电平信号使 T1 处于打开状态，这时 T1 可以被视为短路状态。T1 打开后，节点 1 近似接地，对连接有输入脉冲的突触器件进行 SET 操作，使得突触器件的电导增大，对应突触权重增强。同时，节点 1 的接地会使得器件两端的电压在接地的时间段内几乎为

零，提供了神经元的不应期，让 TS 器件在该期间弛豫回初始的高阻态 HRS。最初，由于在 T3 的栅极上的电压为零，抑制操作模块也可以被视为开路。只有当 TS 器件打开后，才会触发 L1 产生输出信号使得抑制操作模块激活，并抬高节点 1 的电位。节点 1 电位的抬高会对输入为零的突触器件进行 RESET 操作，使得突触器件的电导减小，对应突触权重的减弱。需要指出的是，T1 和 T3 的打开时间发生在两个不同的时钟周期内，因此突触器件的增强和抑制操作不会相互影响。

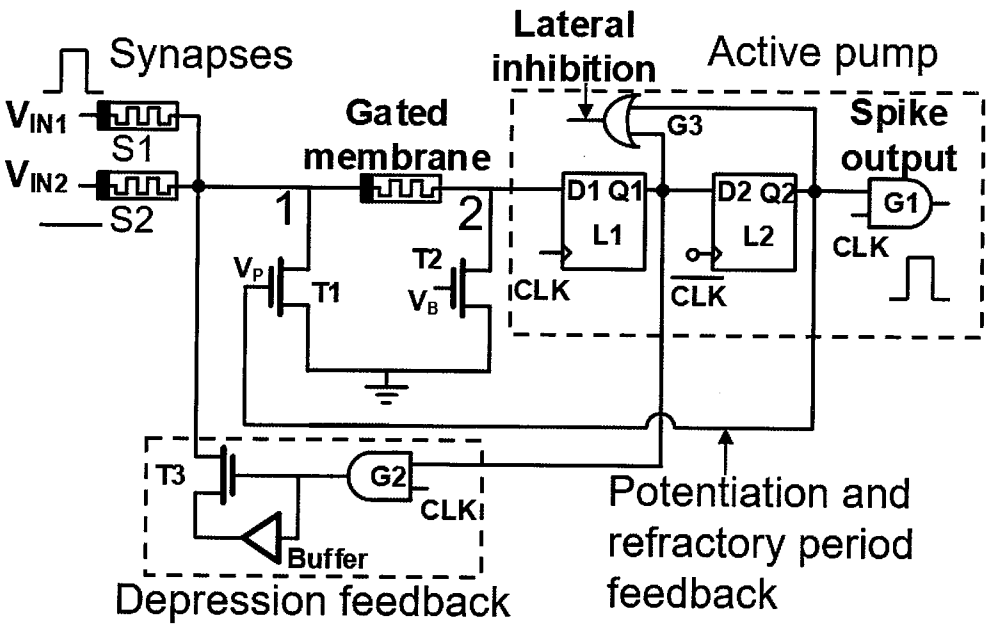


图 4.26 忆阻器-CMOS 混合神经元电路原理图

为了实现该神经元电路的验证，我们首先制作了 PCB 电路板。TS 忆阻器进行封装并连接到 PCB 电路板上。所用 CMOS 数字芯片的型号和供电电源列于表 4.1 中。然后，我们以 Keysight 81160A 脉冲发生器产生输入信号，并用 Keysight Infinivision MSO-X 3104T 示波器测量不同节点的信号变化。

表 4.1 混合神经元电路的数字电路型号和所用电源参数

Name	Model	Sources	Name	Model	Sources
L1	SN74AUC16373	1.5 V	T1	BS170	2.5 V
L2	SN74AUC16373	2.5 V	T2	BS170	1.95 V
G1	SN74LVC08A-Q1	2.0 V	T3	BS170	2.5 V

G2	SN74LVC08A-Q1	2.5 V	Buffer	SN74LVC2G126-EP	2.0 V
G3	SN74AUP1G32	1.5 V	Inverter	SN74AUC1G14-EP	2.0 V

图 4.27 给出了在两个连续放电周期下神经元电路中五个关键节点的输出时序图。这里，为了清楚地显示关键节点上的电压演变，在测量过程中禁用了抑制操作模块，并且用两个固定电阻作为突触电阻（S1＝10 kΩ，S2＝40 kΩ），V_{IN1} 产生输入脉冲信号，V_{IN2} 接地。可以看到，在第一个放电周期内，输入第五个脉冲时，TS 器件打开，导致节点 2 上的电压突然增加。然后，节点 2 上的电压用作 L1 的输入，并在 CLK 信号的控制下出发 L1 输出高电平信号。随后，L1 的输出（L2 的输入）激活 L2 以输出使 T1 接通的高电平电压。当 T1 打开时，节点 1 上的电压几乎为零，这为 TS 忆阻器提供了足够长的时间（不应期（500 μs）+ 脉冲间隔（250 μs））来衰减到其初始 HRS 状态，并为下一次放电事件做准备。在此期间，G1 通过对 L2 的输出和 CLK 信号执行“与”逻辑操作来产生固定的输出脉冲信号。需要指出的是，在这里 CLK 信号是一个全局时钟信号，由信号发生器产生，参数为 2 V，2 kHz 和 50%占空比。所有的输出脉冲是一样的，因为输出脉冲是 L2 的输出和 CLK 信号的“与”操作的结果。因此，该电路可以输出固定的脉冲信号，模拟了生物神经元中动作电位的“全或无”特征[2, 56]。

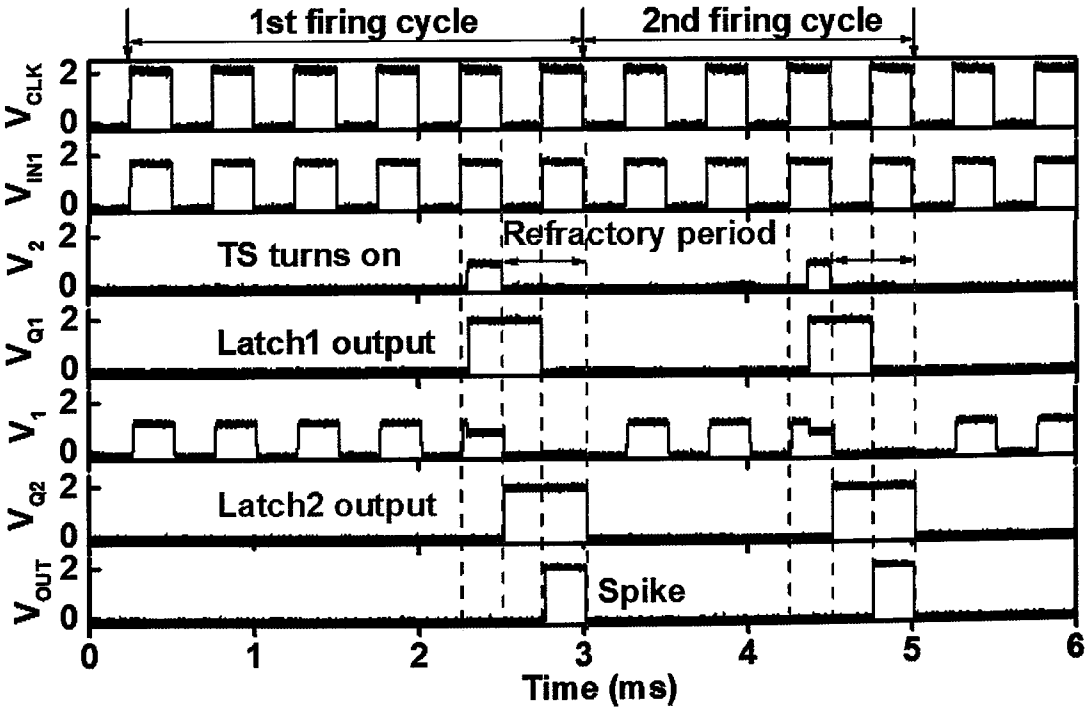


图 4.27 混合神经元电路中关键节点的电压在输入刺激下的动态变化

为了进一步验证该神经元电路在不同输入强度下的放电特性，我们分别测量了不同幅度（1.4 V，1.6 V，1.8 V 和 2.0 V，2 kHz，50%占空比）的脉冲串加在 S1 输入端下的放电行为，这相当于在相同前神经元输入脉冲幅度不同突触权重下对应放电行为，实验结果如图 4.28 所示。可以直观的看到放电频率随着输入脉冲幅度的增加而增加，这说明神经元可以通过输出不同的脉冲频率来对不同的刺激强度进行分类，对应频率编码脉冲神经网络的应用。图 4.29 给出了不同输入脉冲幅度下神经元输出的脉冲信号的放大图，可以观察到所有输出脉冲具有相同的形式（2.0 V，250μs）。这说明该神经元电路的输出脉冲不受输入信号的影响，可以产生固定的动作电位信号。

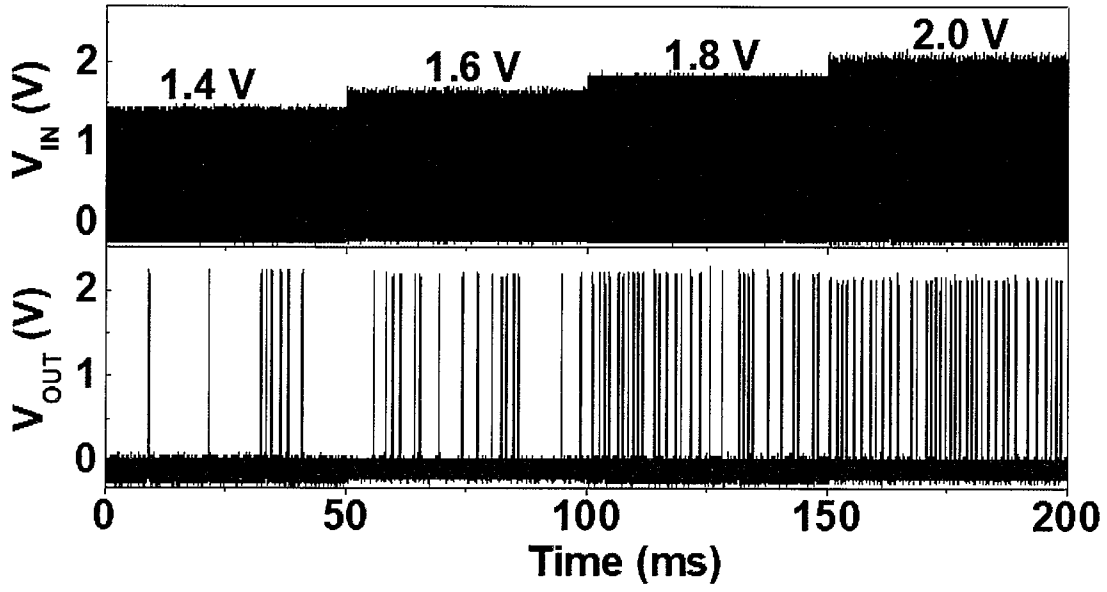


图 4.28 在不同输入脉冲幅度下混合神经元的放电情况

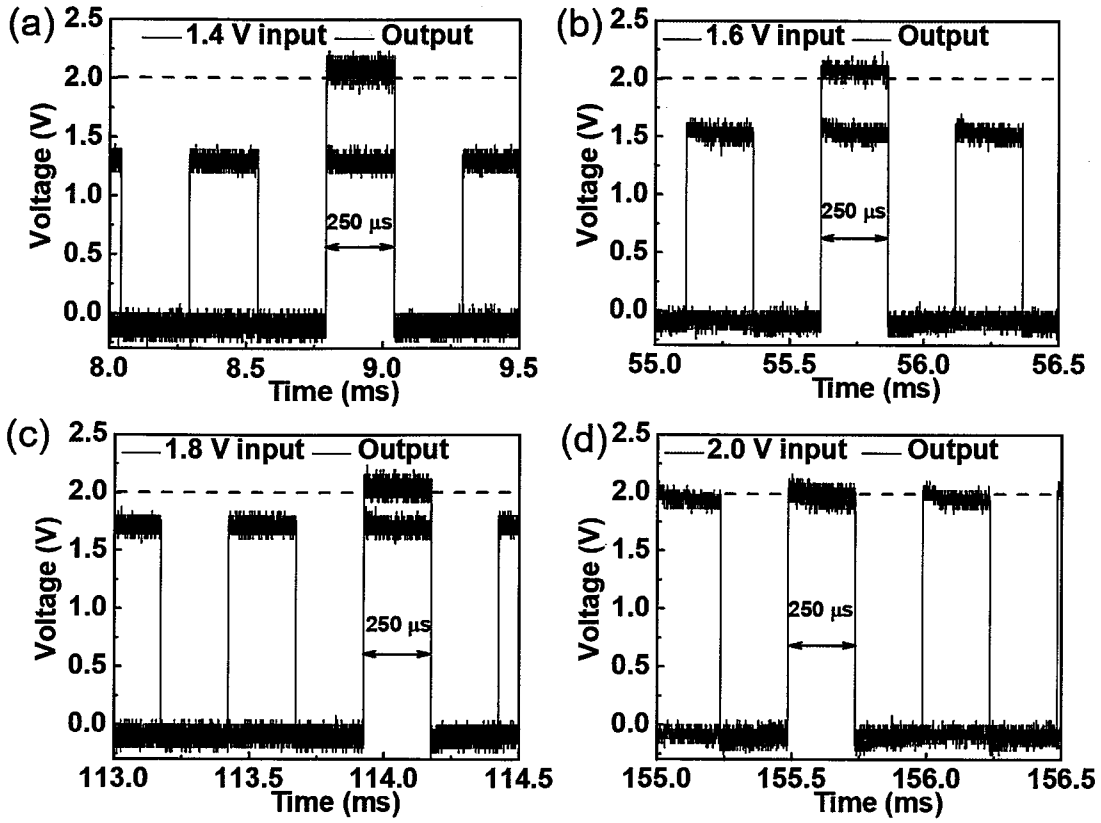


图 4.29 在不同输入脉冲幅度下混合神经元输出脉冲的放大图

图 4.30 进一步给出了神经元在不同输入脉冲强度下放电频率的统计结果。更清楚的证明了该神经元电路输入强度调制的放电频率特性。

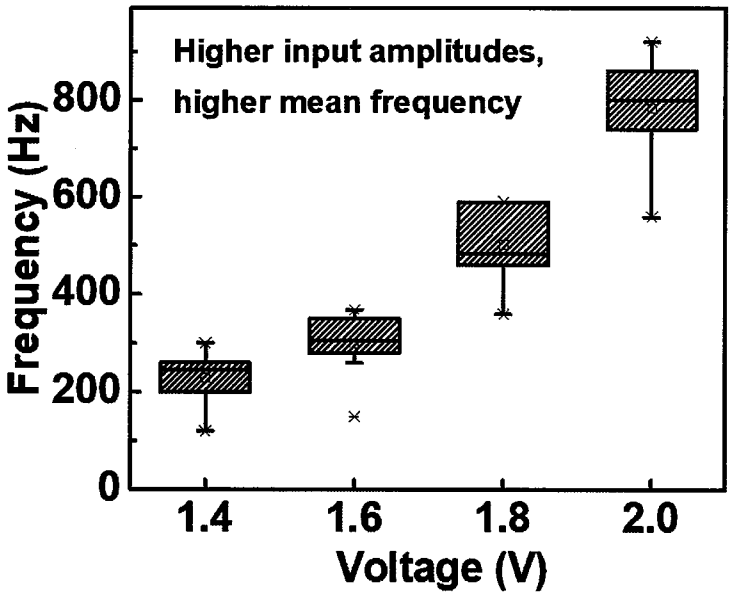


图 4.30 不同输入脉冲幅度下混合神经元放电频率的统计结果

此外，由于该神经元电路的最后输出是由 CMOS 辅助电路产生的，所以该神经元可以看成是一个有源的神经元，具备一定的驱动能力。为了验证该功能，我们将两个神经元通过 5 kΩ 电阻直接相连来测试其放电特性，如图 4.31 所示。脉冲宽度为 1 ms 的脉冲作为输入（黑色），这是为了使得 TS 器件可以在一个脉冲内打开。这两个神经元共享与输入信号相同的时钟信号。红色曲线显示了第一个神经元的输出，它作为第二个神经元的输入。蓝色曲线给出了第二个神经元的脉冲输出。第二个神经元的输出滞后于第一个神经元的输出一个时钟周期，这是由锁存器脉冲触发的特性所导致的。结果表明，在第一个神经元和第二个神经元的输出端均可以观察到完整的输出脉冲，这表明所提出的神经元电路可以通过连接的突触直接驱动相邻的神经元，从而该混合神经元电路可以通过忆阻器突触在多层网络中传播脉冲信号。

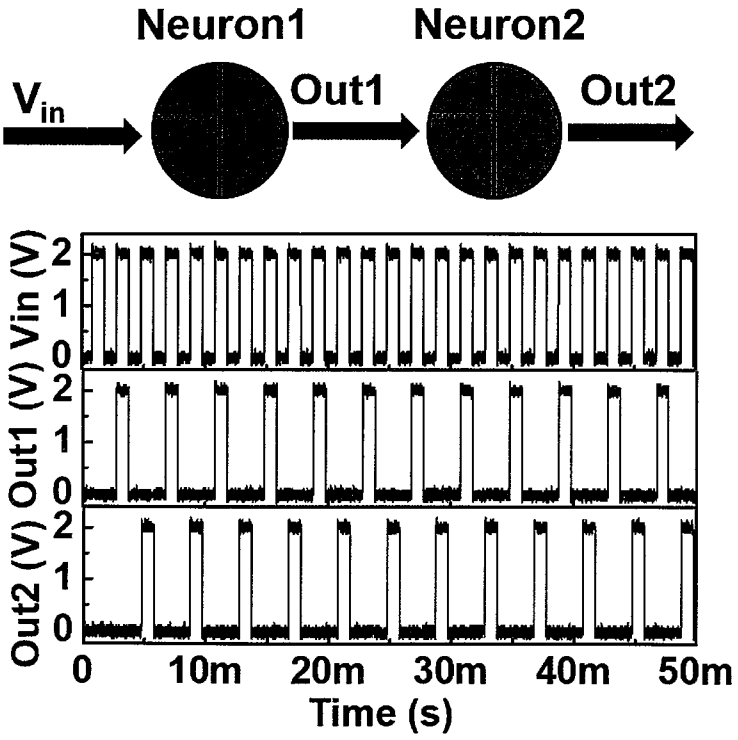


图 4.31 混合神经元的驱动能力展示

我们在前面提到，该神经元电路可以对所连接的突触进行原位编程。接下来，我们将对该工作过程进行分析。如图 4.32 所示，在 TS 忆阻器开启的时钟信号内，L1 首先输出抑制信号，该信号将会激活神经元电路中的抑制操作模块来提升节点 1 的电位。由于 S2 的输入接地，节点 1 电位的提升将对突触器件 S2 执行 RESET 操作，但

由于 S1 具有输入信号，因此对突触器件 S1 的状态没有任何影响。在下一个时钟信号期间，L2 生成用于增强的反馈信号。该反馈信号使节点 1 接地，由于 S1 的输入端具有输入脉冲信号而使得 S1 发生 SET 操作。然而，由于 S2 此时的输入为零，所以 S2 的状态不会发生变化。这种增强和抑制的突触操作过程可以看作是一个优化的赫布学习规则[57]：前后神经元同时具有动作电位输出时，其所连接的突触强度增强，反之，则抑制。

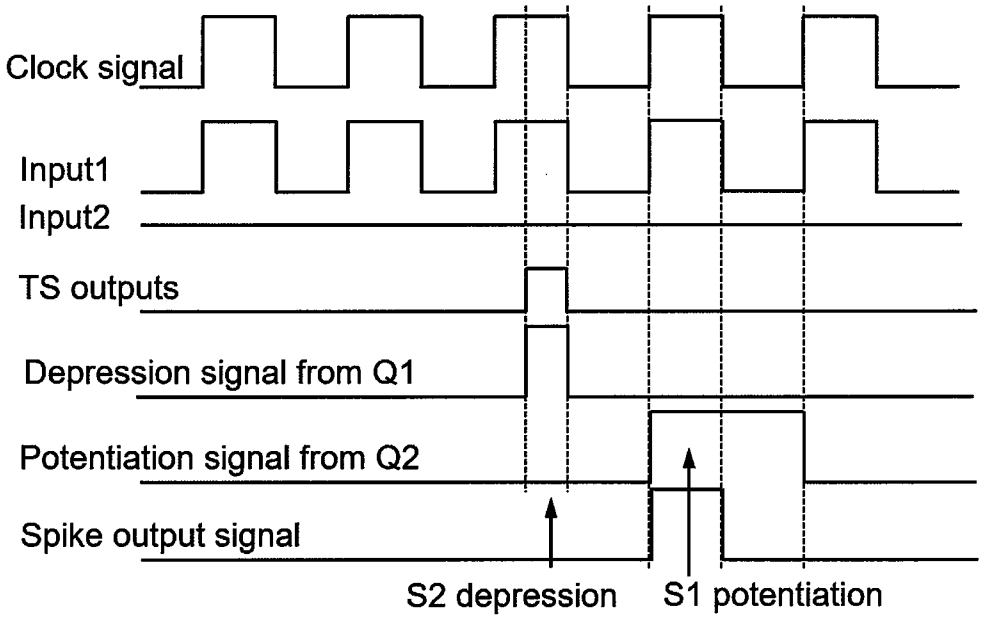


图 4.32 神经元对突触器件进行原位操作的原理图解释

为了实验证明该神经元用于原位学习的可行性，我们将两个忆阻器突触器件（Ta/HfO_2/Pd）连接到神经元电路。在进行操作之前，突触器件 S1 和 S2 被编程到中等电导状态（~ 400 μS @ 0.2 V）。然后在输入端 V_{IN1} 上施加一系列脉冲，V_{IN2} 接地。这里与 Z. Wang 等人[19]的工作相比，突触的增强和抑制操作都是在神经元内进行的，避免了外部抑制电路的使用，减少了硬件开销，更好地服从生物系统中的赫布学习过程。图 4.33 给出了抑制操作模块激活时节点 1 和节点 2 的电压变化，可以看到 TS 器件打开后节点 1 电位的增加，这是因为 G2 的输出同时激活了 T3 和驱动器，该提升的电位将会首先对 S2 进行 RESET 操作。输入信号时间与不应期（RP）的重叠对 S1 进行 SET 操作。为了避免 V_{IN1} 输入信号直接降低 S2，这里使用了低于突触器件复位电压（~1.7 V）且高于 SET 电压（~1.2 V）的输入脉冲幅度（1.6 V）。学习过程结束

后，突触 S1 变成高电导状态（~980 μS @ 0.2 V），突触 S2 降低到低电导状态（~42 μS @ 0.2 V）。

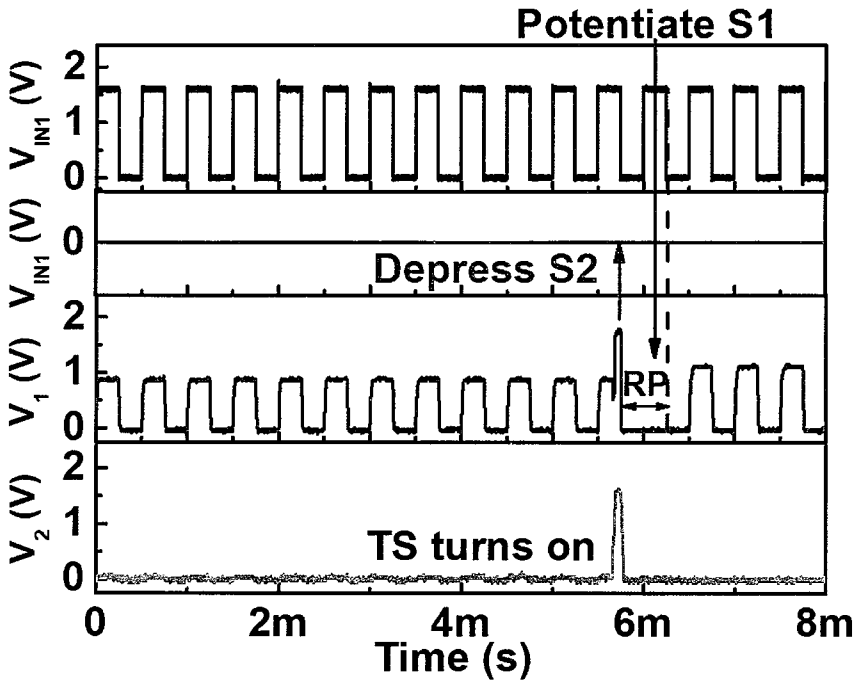


图 4.33 神经元抑制模块激活时节点 1（V_1）和节点 2（V_2）的电压变化

图 4.34 给出了学习过程中神经元的放电行为，可以看到第一次放电行为需要更多的脉冲来触发，这是因为开始时突触后膜电位（节点 1 的电位）比较低。神经元放电后就会对连接的突触器件进行原位编程操作，导致突触后膜电位（节点 1 的电位）升高，从而更少的脉冲便可以触发神经元产生放电信号。随着输入脉冲计数的增加，输出脉冲频率增加，证明了混合神经元的原位学习能力。

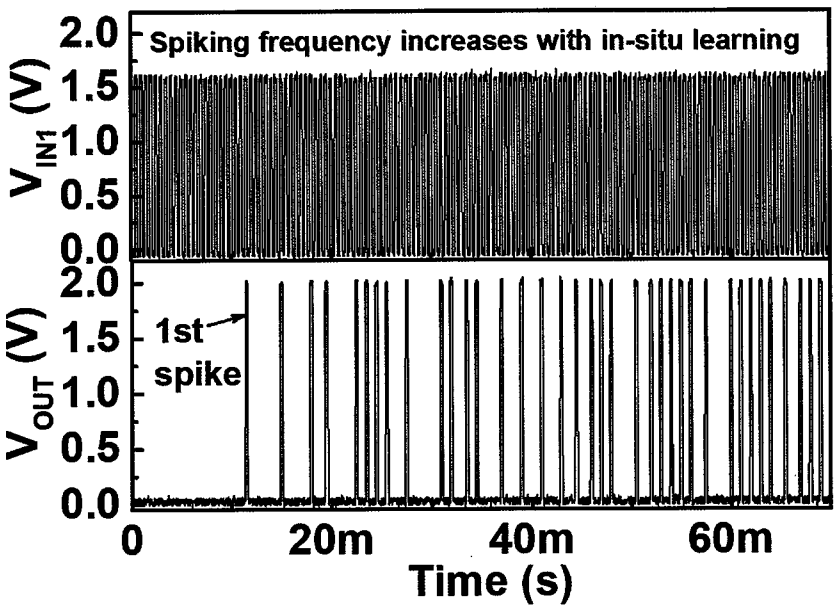


图 4.34 混合神经元在学习过程中放电频率的变化

4.3.4 侧向抑制电路的设计

侧向抑制操作是脉冲神经网络进行非监督学习的关键特征[47,58]，它可以支持赢者通吃（winner-take-all, WTA）学习规则的实现，WTA 表明一旦获胜神经元放电，其它神经元就会受到抑制。为了使用所提出的神经元电路执行侧抑制操作，我们设计了一个侧抑制阵列（Lateral inhibition array, LIA），由忆阻器阵列和比较器构成，如图 4.35 所示。比较器的输出作用于 1T1R 突触的共享栅极。为了成功地进行侧抑制操作，LIA 应该具有两个特点：首先，当没有神经元放电时，来自神经元的所有侧抑制信号（V_{L1}-V_{L10}）都是“0”，这时，LIA 的所有输出（L_{G1}-L_{G10}）都应该是“1”，以激活所有突触进行推理操作。其次，当获胜神经元放电时（如 N1），N1 的侧抑制信号为“1”（$V_{L1}=1$），其它神经元的侧抑制信号（V_{L2}-V_{L10}）均为“0”。在这种情况下，LIA 输出的 L_{G1} 应为“1”，其他的输出（L_{G2}-L_{G10}）为“0”。因此，只有获胜神经元的感受野活跃，其他神经元的感受野受到抑制，然后可以对获胜者神经元的感受野进行原位学习操作。为使侧向抑制忆阻器阵列权重有效，我们在 LIA 中引入了偏置输入 V_{BIAS}（后面将给出相关的数学分析）。这里，比较器的正输入端用作参考端，负输入端接收来自忆阻器阵列的信号，这种方法中避免了使用差分电阻对负权重值的表示[59]，可以大大节省硬件开销。

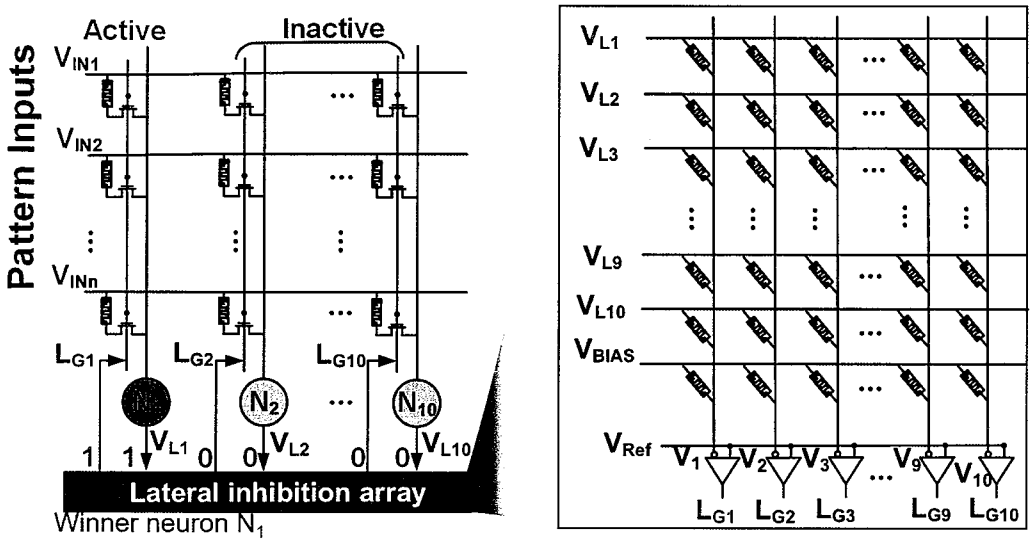


图 4.35 侧向抑制电路原理图

下面将给出 LIA 忆阻器权值的计算方法：

(1) 当所有后神经元处于静息状态时，由于需要并行推理操作，LI 电路应激活所有与后神经元相连的突触。因为开始时所有后神经元来自 G3 的 LI 输出信号都是零，这里我们记为“0”，如输入矩阵 X_0 的第一列所示。对应的 LI 电路的输出应全部为“1”，如输出矩阵 Y 的第一行所示。

(2) 在发生侧向抑制过程中，只有一个神经元放电，其他神经元受到抑制。这意味着，如果后神经元“1”首先放电，那么它有 LI 输出信号，对应为“1”。同时，其他后神经元不应该再放电，也就没有 LI 信号，对应为“0”，如输入矩阵 X_0 的第二列所示。在这种情况下，来自 LIA 电路的输出信号应该是输出矩阵 Y 的第二行中显示的值。对于其他情况，例如后神经元“2”首先激发，LIA 电路输入显示在输入矩阵 X_0 的第三列中， LIA 相应的输出显示在输出矩阵 Y 的第三行中，以此类推。

LIA Input Matrix X_0 (11 × 10)

$$\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

LIA Output Matrix Y (11 × 10)

$$\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}$$

LIA Input Matrix X (11 × 11)

$$\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}$$

LIA Output Matrix Y_{inv} (11 × 10)

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0
\end{bmatrix}$$

LIA Weight Matrix W (11 × 10)

$$\begin{bmatrix}
0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

由于 X_0 是 11×10 的结构，Y 是 11×10 的结构，矩阵等式 $X^T \times W = Y$ 中的 W（权重）无解。因此，我们在 X_0 中增加一行作为 LIA 的偏置输入形成 11×11 的结构，如输入矩阵 X 所示。在这种情况下，$X^T \times W = Y$ 中的 W 具有唯一解。但是，该解同时包含正值和负值。负值在用忆阻器实现时通常需要一个差分电阻对，这会增加突触器件的数量和外围电路的复杂性。为了解决这个问题，我们重新定义了 LIA 电路的输出

矩阵 Y，并对其进行了取反，如 LIA 输出矩阵 Y_{inv} 所示。此时，$X^T \times W = Y$ 中的 W 具有唯一解并且所有的值都是正的，如 LIA 权重矩阵 W 所示。在用忆阻器实现时，我们将“0”映射到器件的高阻态 HRS，将“1”映射为忆阻器的低阻态 LRS。然后，将 Y_{inv} 作为比较器负端的输入，正端作为参考端，比较器的输出作用于神经元感受野中突触器件的共享栅极。图 4.36 给出了计算得出的 LIA 权值矩阵映射为器件电导并写入阵列中的电导谱图。

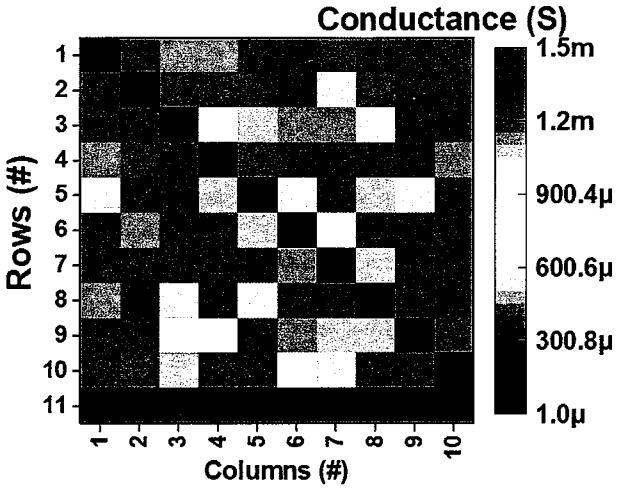


图 4.36 侧向抑制阵列的电导谱图

在验证该 LIA 电路的功能时，突触阵列通过探针卡连接到 PCB 板上。利用 MATLAB 控制 81160A 脉冲发生器产生输入信号。用 PicoScope 4000 系列示波器和 Keysight Infinivision MSO-X 3104T 示波器同时测量了 10 路输出信号。

接下来，我们分别在两种输入条件下对 LIA 进行了测试：所有神经元的 V_L 为“0”（0 V）和只有获胜神经元的 V_L 为“1”（1.5 V），如图 4.37（a）所示。具体来讲，对于所有神经元的 V_L 为“0”的输入条件，一开始没有神经元放电，侧向抑制输出（V_{L1}-VL_{10}）为“0”（0 V）。因此，除了偏置输入是“1”（1.5 V）（如图 4.37（a）的左侧）之外，LIA 的输入都是“0”（0 V）。在这种情况下，所有的 LIA 输出（L_{G1}-L_{G10}）都是“1”（3 V），用于激活所有突触，如图 4.37（b）的左侧所示。当只有获胜神经元的 V_L 的输入是“1”时，这对应于只有获胜神经元的侧抑制信号是“1”（1.5 V），其他的神经元是静默的，如图 4.37（a）的右侧所示。在这种情况下，只有获胜神经元的 L_G 为“1”（3 V），而对应于其它神经元的 LIA 输出都为“0”（0 V），如图

4.37（b）的右侧所示。因此，只能对获胜神经元的感受野进行编程操作。

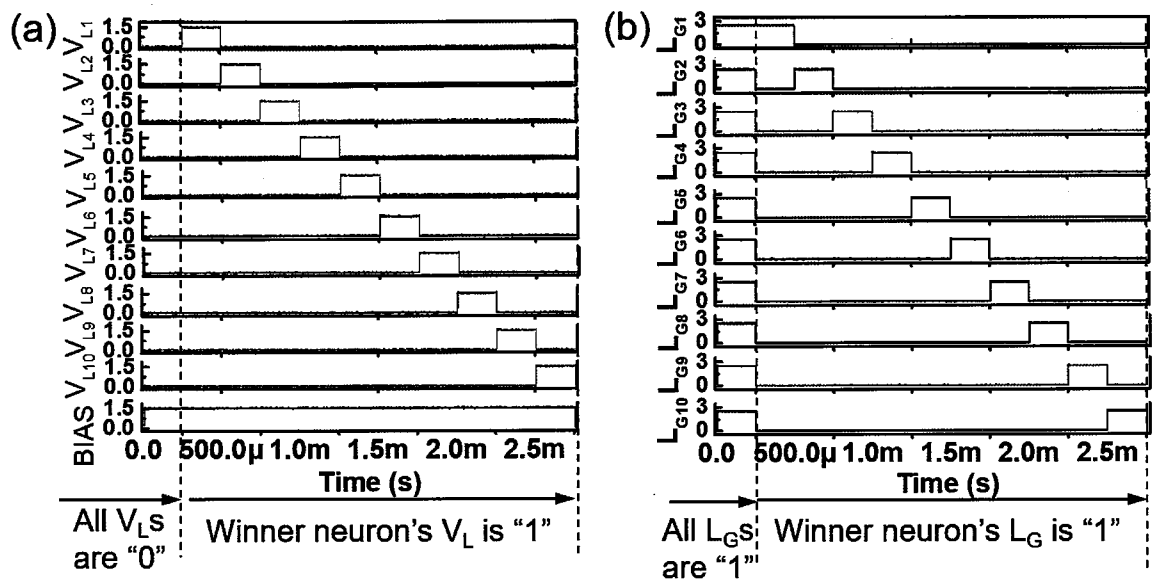


图 4.37 （a）LIA 电路的测试输入；（b）LIA 电路的测试输出

需要注意的是，获胜神经元的侧抑制信号发生在 TS 器件打开时刻，而且侧向抑制持续时间总是比不应期要长，说明侧向抑制信号可以被及时触发，如图 4.38 所示。这些实验结果表明，该 LIA 电路可以实现上述理论分析的功能并支持所提出的混合神经元的侧抑制操作，进而结合 WTA 学习规则进行非监督学习。

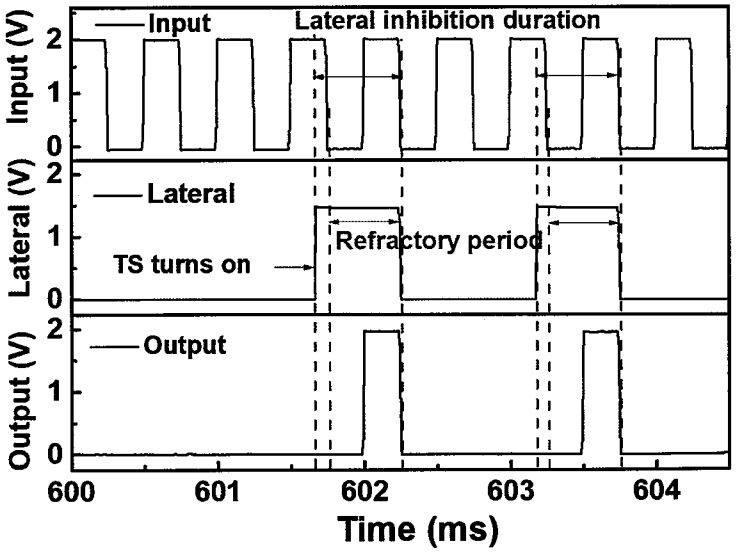


图 4.38 混合神经元侧向抑制信号的输出时刻表示

4.3.5 全硬件脉冲神经网络实现

基于所提出的混合神经元和侧抑制电路，我们进一步证明了一个全硬件的多层SNN，在第一层执行非监督学习用于对输入模式进行预编码，在第二层执行有监督学习用于输入模式的识别。所构建的网络为 30 × 10 × 10 两层结构，框架图如图 4.39 所示，10 个隐藏层神经元从第一层中的 30 个输入神经元接收输入信号并且借助 LIA 电路以非监督学习的方式编程 30 × 10 忆阻器突触阵列。输出层 10 个神经元接收隐藏层神经元的输出并以监督学习的方式编程 10 × 10 忆阻器突触阵列。

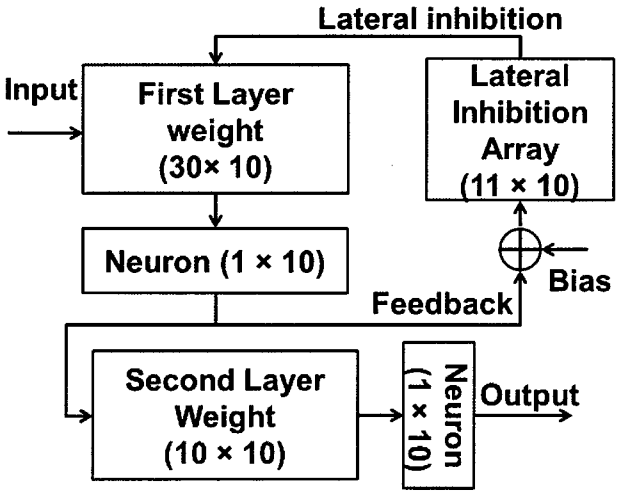


图 4.39 两层 SNN 网络的框架图

图 4.40 进一步给出了 SNN 的详细电路原理图以及相关的硬件平台图像。

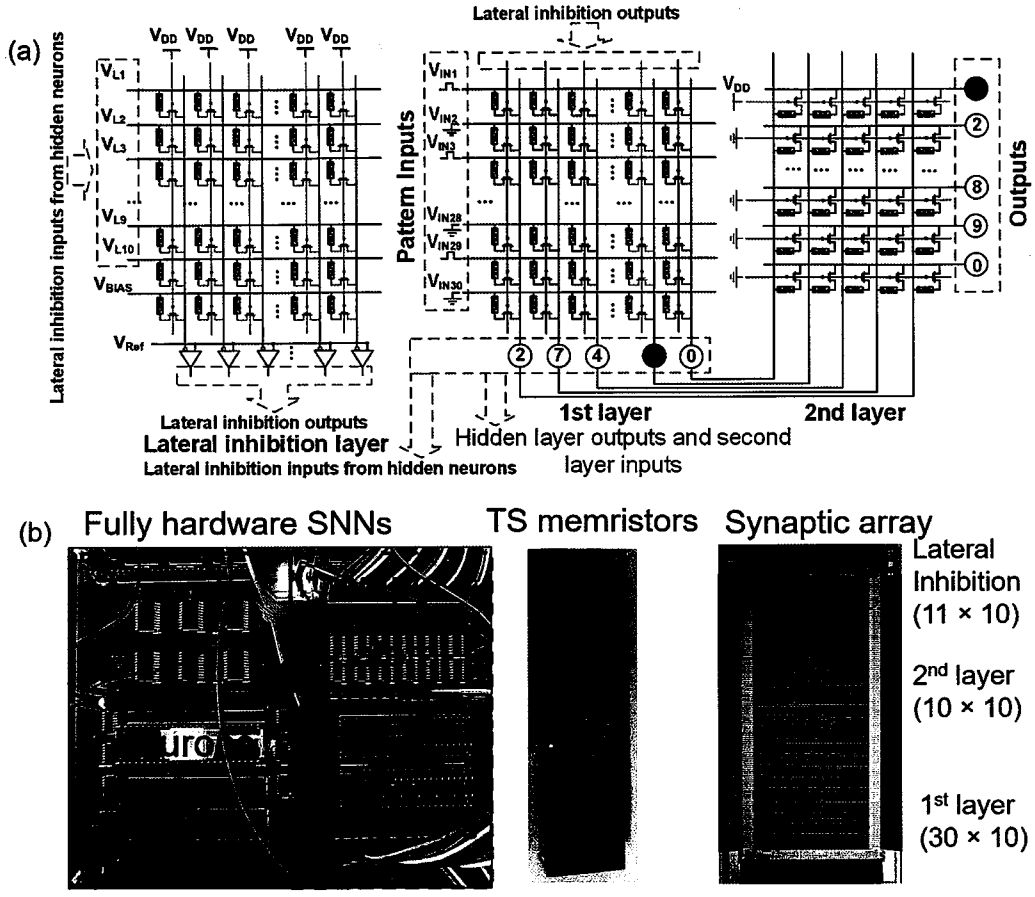


图 4.40 （a）两层 SNN 网络示意图以及（b）硬件平台图像

实验过程包括两个过程：学习过程和推理过程。在学习过程中，由于同时测量通道（10 个通道）的数量有限，我们对两层网络进行了逐层训练。图 4.41 上半部分给出了用于学习的数字模式，下半部分给出了推理过程中加了噪声点的数字模式，每个模式包括 30 个像素点（6×5）。在实际操作中，黑色像素点被视为“1”，然后被映射为 1.6 V 的输入脉冲（2 kHz，250 μs 脉冲宽度）。白色像素被看成“0”，因此与这些像素对应的输入接地。在训练之前，为了清楚地演示突触器件的变化，将第一层的电导初始化为中间值（约 400 μS）。然后逐渐增加对应像素点为“1”的输入端上的输入脉冲，直到观察到有神经元放电，读出阵列的电导值并进行下一次循环，如图 4.42（a）所示，每个输入模式操作 30 个循环。

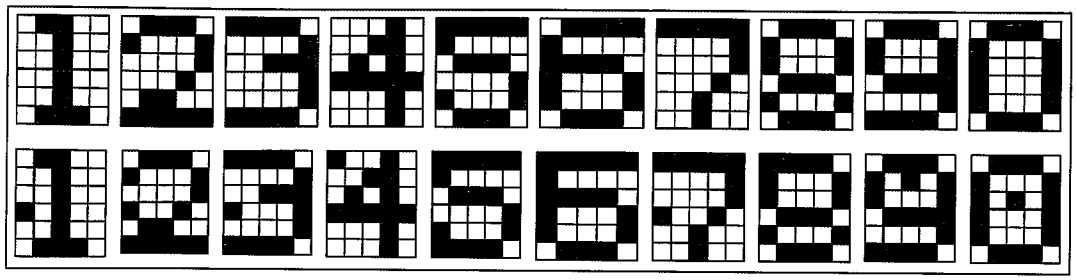


图 4.41 用于学习和推理的数字模式：上半部分为学习的模式，下半部分为用于推理的模式

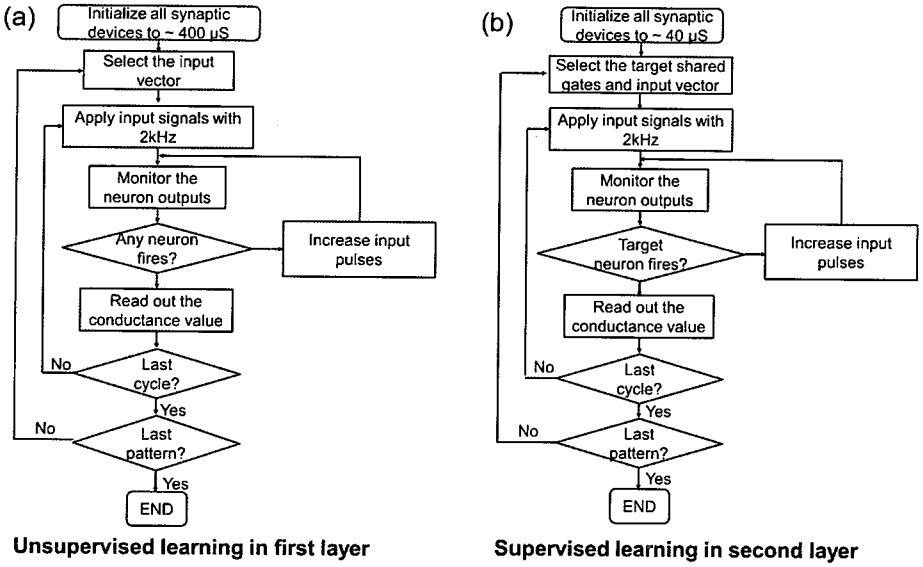


图 4.42 （a）网络第一层非监督训练流程图；（b）网络第二层监督训练流程图

在训练第二层时，我们采用监督学习的方法，突触器件被初始化为低电导值（约 40 μS）。3.0 V 电压施加在目标神经元突触感受野的共享栅极上，其它神经元突触感受野的栅极接地。然后，逐渐增加施加于所选输入端的输入脉冲数，直到观察到放电事件，读出阵列的电导值并进行下一次循环，如图 4.42（b）所示，每个输入目标神经元操作 30 个循环。

接下来，我们根据前面的训练方法进行实验操作。首先，在第一层执行非监督学习，以实现对输入模式的预编码。4.43（a）给出了第一层忆阻器突触阵列初始化后的电导值谱图。在训练期间，权重调制（电导变化）遵循优化的赫布学习规则。4.43（b）给出了当数字“1”作为输入时，在 30 次放电事件过程中神经元“1”的感受野的变化。可以清楚的看到突触器件的电导随着放电事件的进行而被明显的改变，这说明神

经元“1”完成了学习功能。对于其它数字模式作为输入时对应的放电神经元的感受野也表现出同样的趋势，在此不再一一给出。对 10 个输入模式的训练完成后，观察到清楚地看到突触阵列电导地变化，如图 4.43（c）所示。然后，用训练好的权值网络进行推理验证，在此过程中，图 4.41 中带有噪声像素点的模式作为推理的输入模式。图 4.43（d）给出了具有噪声像素点的不同输入模式下的神经元响应的放电频率。结果表明，当输入模式为数字“5”（或数字“6”）时，输出神经元“4”和输出神经元“5”都会有放电事件的发生；当输入模式为数字“8”（或数字“9”）时，输出神经元“7”和输出神经元“3”也都会有放电事件的发生。这是因为这些模式之间的相似度很高，这导致几乎相同的突触后膜电位，如图 4.44 所示。然而不同输入模式下还是分别有最大输出频率的输出神经元，因此，在频率编码的脉冲神经网络中并没有什么影响。实验结果表明，该神经元电路借助 LIA 电路能够成功地在忆阻器突触阵列中进行非监督学习，实现对未标记输入模式的预编码。

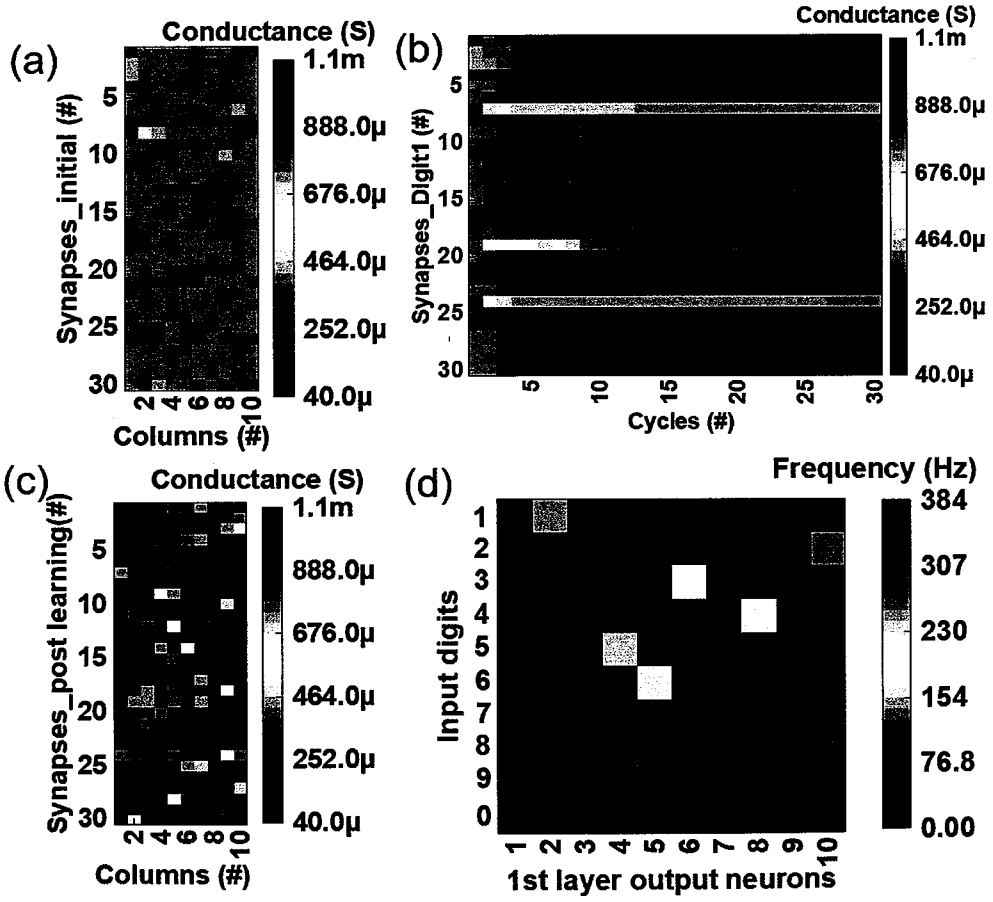


图 4.43 （a）网络第一层突触电导的初始值；（b）数字“1”作为输入时对应隐藏层放电神

经元的感受野变化；（c）训练后第一层突触的电导谱图；（d）不同输入模式下对应的不同隐藏层神经元的放电频率

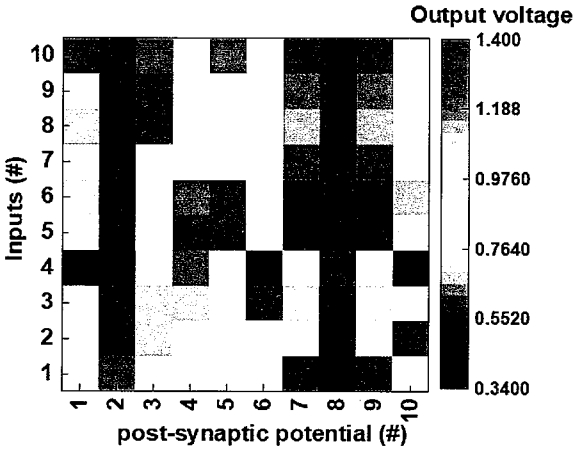


图 4.44 推理过程中不同输入模式下对应的隐藏层神经元的突触后膜电位

值得注意的是，由于神经元的随机性，在学习过程中成功地避免了静默神经元的出现，这对于以非监督的学习方式执行 WTA 学习规则是至关重要的[47]。为了进行比较，我们用不具有随机性的神经元进行了仿真模拟，图 4.45（a）给出了仿真过程中的突触电导变化，可以看到训练后只有一列电导被更新，这说明所有的输入模式均被该列对应的神经元所捕获。图 4.45（b）给出了该列器件在训练过程中的变化，可以看到大多数器件在训练的过程中表现出振荡的演变，这是因为每个不同的输入模式都会造成这一列突触器件的更新。而其它神经元由于没有放电事件的发生，所以对应的突触感受野没有任何变化，如图 4.45（c）所示。该仿真结果进一步验证了神经元随机性的重要性，不具有随机性的神经元不能对输入模式进行正常的预编码，或者会导致预编码效果很差。

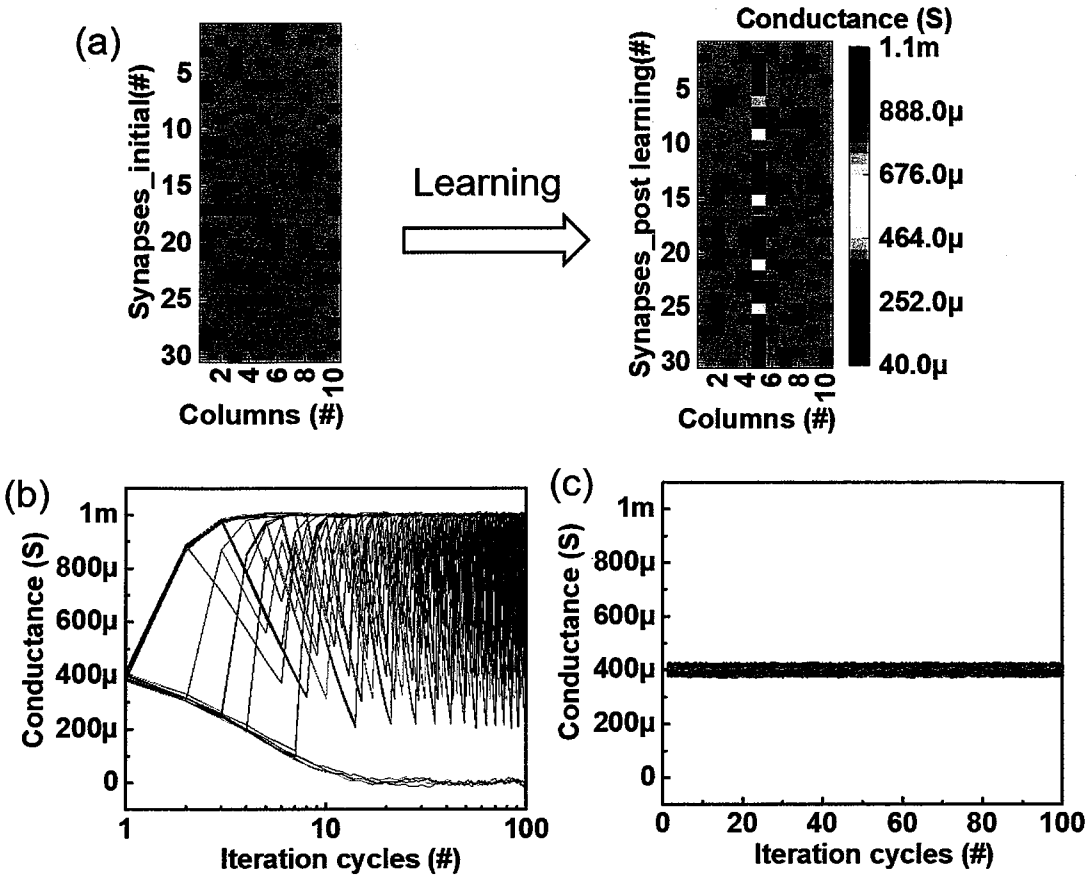


图 4.45 无随机性的神经元用作非监督学习的仿真结果：（a）训练前后权值谱图的变化；（b）获胜神经元的感受野在训练过程中的演变过程；（c）无放电事件神经元的感受野在训练过程中的演变过程

然后，我们以有监督的方式进行第二层的训练。图 4.46（a）给出了该层突触阵列电导初始化后的谱图（低电导状态）。训练时，我们在目标神经元感受野的共享栅极上施加恒定电压（3 V），其它神经元感受野的共享栅极接地。当数字“1”作为输入时，隐藏层中第二个神经元有输出信号。因此，我们在第二层输入的相应输入端口施加输入信号，其它输入端口接地。图 4.46（b）给出了数字“1”在前 30 次训练迭代中对应的突触电导演化。我们注意到，目标感受野的增强操作在一次迭代周期内基本完成，这是因为输入脉冲足够强，可以直接 SET 突触器件。这可能会限制可以学习的模式数量，但可以通过使用较弱的脉冲来实现模拟增强过程以及引入突触的随机性来缓解这种情况。由于其他突触器件的输入端接地，电导值几乎没有变化，因为初始化的低电导无法进一步编程。其他输入模式下电导演化与此类似，也不一一赘述。训

练完成后，第二层的权重图如图 4.46（c）所示。可以看到，在每个神经元的感受野中，只有一个相关的突触被编程。被编程突触的位置正好对应了隐藏层中频率最高的神经元的位置。图 4.46（d）进一步给出了第二层神经元在隐藏层神经元不同编码模式下的放电频率，观察到清楚的识别结果。该结果表明，基于隐藏层神经元预编码的结果，输出层神经元可以进一步实现有监督的学习并完成模式识别。因此，利用该混合神经元电路有希望构建一个具有在线学习能力的高密度脉冲神经形态机器。

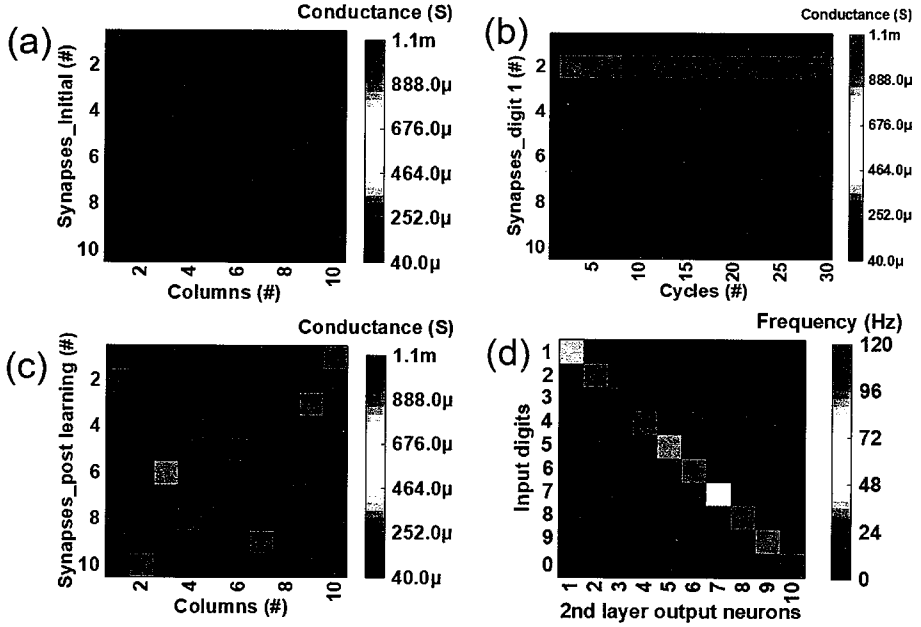


图 4.46 （a）网络第二层突触电导的初始值；（b）数字“1”作为输入时对应输出目标神经元的感受野变化；（c）训练后第二层突触的电导谱图；（d）不同输入模式下对应的不同输出神经元的放电频率

4.3.6 结果讨论

本部分工作所使用的 Ag 掺杂 SiO_2 TS 忆阻器具有均一性好、无需 forming 操作的特点，有利于未来大规模集成的实现。重要的是，在信号编码和组合优化中起着关键作用的随机神经元动力学和“遗忘”行为可以在具有内在随机性和泄漏性的 TS 忆阻器中很好地实现。近年来，基于相变材料的随机神经元已经被成功实现，并且证实在实现种群编码和时间相关性检测中具有重要意义[29]。本研究中使用的脉冲参数只是一个示例，根据所报道的 TS 忆阻器的最新实验数据，作为输入的脉冲宽度可以缩放到 μs 甚至 ns 量级[44, 45]，从而实现更快的计算。此外，所提出的混合设计理念还可扩

展到其他基于不同金属氧化物（如 NbO_2 和 VO_2）的忆阻器中，这些金属氧化物已被证实在模拟脉冲神经元中具有丰富的物理动力学。

为了利用混合神经元实现 WTA 学习规则，设计了基于忆阻器阵列的 LIA 电路。在这项工作中，我们验证了 10 个神经元之间的 LIA 电路，通过相应地增加阵列大小（n×（n+1），n 是 WTA 神经元数目），该 LIA 电路在 n 个 WTA 神经元之间仍然可行。此外，由于所使用的 CMOS 辅助模块使得该神经元电路具有一定的驱动能力，因此可以扩展到更多层的网络结构。

4.4 本章小结

脉冲神经元电路是实现高效脉冲神经网络硬件的关键单元，本章工作围绕着如何利用忆阻器内在的离子动力学机制实现神经元电路展开了相应的理论探讨和实验验证。初步提出了利用离子基忆阻器实现神经元电路的方法，为优化神经元电路的功能又进一步提出了忆阻器-CMOS 混合设计的方案，实现了稳定可靠的神经元电路并进行了系统级的实验验证。本章工作取得的成果主要如下：

1) 制备了具有短时程离子动力学的 TS 忆阻器并基于此提出一种构建神经元的方法，该神经元具有 LIF 神经元模型的基本特性，成功实现了生物神经元的四个基本特征：动作电位的全或无，阈值驱动放电，不应期，和输入强度调制的频率响应。此外，基于 TS 神经元，我们还模拟了单层的脉冲神经网络，成功实现了对输入数字模式的识别。该部分研究成果发表在 2018 年 2 月份的 IEEE Electron Device Letters 期刊上，并申请了一个中国专利。

2) 为进一步提高神经元的集成度并丰富神经元的功能，提出了一种能够同时实现随机 LIF 神经元功能和在多层网络中对忆阻器突触进行原位学习的忆阻器-CMOS 混合神经元。此外，还特别设计了一种侧向抑制电路用于非监督学习。最后，结合提出的神经元和 LIA 电路，实验验证了一个全硬件的两层脉冲神经网络并进行了原位学习操作，进一步推动了全忆阻器基神经形态计算系统的研究进程。该部分研究内容已申请了一个国际专利，正在准备投稿。

参考文献

[1] Baars BJ and Gate NM, Cognition, Brain and Consciousness, 2nd ed. [M]. ELSEVIER, 2010.

[2] Purves D, Augustine GJ, Fitzpatrick D, et al., Neuroscience, 3rd ed. [M]. Inc. Massachusetts, USA: Sinauer Associates, 2012.

[3] Maass W, Networks of spiking neurons: The third generation of neural network models [J]. Neural Networks, vol. 10, pp. 1659-1671, Dec 1997.

[4] Merolla PA, Arthur JV, Alvarez-Icaza R, et al., A million spiking-neuron integrated circuit with a scalable communication network and interface [J]. Science, vol. 345, pp. 668-673, 2014.

[5] Davies M, Srinivasa N, Lin T-H, et al., Loihi: A Neuromorphic Manycore Processor with On-Chip Learning [J]. Ieee Micro, pp. 82-99, 2018.

[6] Pei J, Deng L, Song S, et al., Towards artificial general intelligence with hybrid Tianjic chip architecture [J]. Nature, vol. 572, pp. 106-111, Aug 2019.

[7] Imam N and Cleland TA, Rapid online learning and robust recall in a neuromorphic olfactory circuit [J]. Nature Machine Intelligence, vol. 2, pp. 181-191, 2020.

[8] Wang Z, Li C, Song W, et al., Reinforcement learning with analogue memristor arrays [J]. Nature Electronics, vol. 2, pp. 115-124, 2019.

[9] Yao P, Wu H, Gao B, et al., Fully hardware-implemented memristor convolutional neural network [J]. Nature, vol. 577, pp. 641-646, Jan 2020.

[10] Prezioso M, Mahmoodi MR, Bayat FM, et al., Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits [J]. Nat Commun, vol. 9, p. 5311, Dec 14 2018.

[11] Song KM, Jeong J-S, Pan B, et al., Skyrmion-based artificial synapses for neuromorphic computing [J]. Nature Electronics, vol. 3, pp. 148-155, 2020.

[12] Izhikevich EM, Which Model to Use for Cortical Spiking Neurons? [J]. IEEE Transactions on Neural Networks, vol. 15, pp. 1063–1070, 2004.

[13] Izhikevich EM, Simple model of spiking neurons [J]. IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 14, pp. 1569-1572, 2003.

[14] Douglas MMaR, A silicon neuron [J]. Nature, vol. 354, pp. 515-518, 1991.

[15] Indiveri G, Linares-Barranco B, Hamilton TJ, et al., Neuromorphic silicon neuron circuits [J]. Front Neurosci, vol. 5, p. 73, 2011.

[16] Beck ME, Shylendra A, Sangwan VK, et al., Spiking neurons from tunable Gaussian heterojunction transistors [J]. Nat Commun, vol. 11, p. 1565, Mar 26 2020.

[17] Mehonic A and Kenyon AJ, Emulating the Electrical Activity of the Neuron Using a Silicon Oxide RRAM Cell [J]. Front Neurosci, vol. 10, p. 57, 2016.

[18] Lashkare S, Chouhan S, Chavan T, et al., PCMO RRAM for Integrate-and-Fire Neuron in Spiking Neural Networks [J]. Ieee Electr Device L, vol. 39, pp. 484-487, Apr 2018.

[19] Wang Z, Joshi S, Savel'ev S, et al., Fully memristive neural networks for pattern classification with unsupervised learning [J]. Nature Electronics, vol. 1, pp. 137-145, 2018.

[20] Mulaosmanovic H, Chicca E, Bertele M, et al., Mimicking biological neurons with a nanoscale ferroelectric transistor [J]. Nanoscale, vol. 10, pp. 21755-21763, Dec 2018.

[21] Dutta S, Saha A, Panda P, et al., Biologically Plausible Ferroelectric Quasi-Leaky Integrate and Fire Neuron[M]. New York: Ieee, 2019.

[22] Chen C, Yang M, Liu S, et al., Bio-Inspired Neurons Based on Novel Leaky-FeFET with Ultra-Low Hardware[M]. New York, 2019.

[23] Sengupta A, Panda P, Wijesinghe P, et al., Magnetic Tunnel Junction Mimics Stochastic Cortical Spiking Neurons [J]. Sci Rep, vol. 6, p. 30039, Jul 21 2016.

[24] Akhilesh Jaiswal SR, Gopalakrishnan Srinivasan, and Kaushik Roy, Proposal for a Leaky-Integrate-Fire Spiking Neuron Based on Magnetoelectric Switching of Ferromagnets [J]. Ieee T Electron Dev, vol. 64, 2017.

[25] Romera M, Talatchian P, Tsunegi S, et al., Vowel recognition with four coupled spin-torque nano-oscillators [J]. Nature, vol. 563, pp. 230-234, Nov 2018.

[26] Wu MH, Hong MC, Chang C-C, et al., Extremely Compact Integrate-and-Fire STT-MRAM Neuron: A Pathway toward All-Spin Artificial Deep Neural Network[M]. New York: Ieee, 2019.

[27] Pickett MD, Medeiros-Ribeiro G, and Williams RS, A scalable neuristor built with Mott memristors [J]. Nature Materials, vol. 12, pp. 114-117, Feb 2013.

[28] Lin J, Annadi A, Sonde S, et al., Low-voltage artificial neuron using feedback engineered insulator-to-metal-transition devices, in 2016 Ieee International Electron Devices Meeting (IEDM), ed New York: Ieee, 2016.

[29] Tuma T, Pantazi A, Le Gallo M, et al., Stochastic phase-change neurons [J]. Nat Nanotechnol, May 16 2016.

[30] Yi W, Tsang KK, Lam SK, et al., Biological plausibility and stochasticity in scalable VO_2 active memristor neurons [J]. Nat Commun, vol. 9, p. 4661, Nov 7 2018.

[31] Wang Z, Rao M, Han JW, et al., Capacitive neural network with neuro-transistors [J]. Nat Commun, vol. 9, p. 3208, Aug 10 2018.

[32] Jerry M, Parihar A, Grisafe B, et al., Ultra-Low Power Probabilistic IMT Neurons for Stochastic Sampling Machines[M]. New York: Ieee, 2017.

[33] Stoliar P, Tranchant J, Corraze B, et al., A Leaky-Integrate-and-Fire Neuron Analog Realized with a Mott Insulator [J]. Advanced Functional Materials, p. 1604740, 2017.

[34] Yang Y, Gao P, Li L, et al., Electrochemical dynamics of nanoscale metallic inclusions in dielectrics [J]. Nat Commun, vol. 5, p. 4232, Jun 23 2014.

[35] Wang W, Wang M, Ambrosi E, et al., Surface diffusion-limited lifetime of silver and copper nanofilaments in resistive switching devices [J]. Nat Commun, vol. 10, p. 81, Jan 8 2019.

[36] Xia Q and Yang JJ, Memristive crossbar arrays for brain-inspired computing [J]. Nature Materials, vol. 18, pp. 309-323, 2019.

[37] Wang Z, Joshi S, Savel'ev SE, et al., Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing [J]. Nat Mater, Sep 26 2016.

[38] Valov I, Linn E, Tappertzhofen S, et al., Nanobatteries in redox-based resistive switches require extension of memristor theory [J]. Nat Commun, vol. 4, p. 1771, 2013.

[39] Liu Q, Sun J, Lv HB, et al., Real-Time Observation on Dynamic Growth/Dissolution of Conductive Filaments in Oxide-Electrolyte-Based ReRAM [J]. Advanced Materials, vol. 24, pp.

1844-1849, Apr 2012.

[40] Shrestha PR, Nminibapiel DM, Campbell JP, et al., Analysis and Control of RRAM Overshoot Current [J]. Ieee T Electron Dev, vol. 65, pp. 108-114, Jan 2018.

[41] Abbott LF, Lapicque's introduction of the integrate-and-fire model neuron (1907) [J]. Brain Research Bulletin, vol. 50, pp. 303-304, 1999.

[42] Sun H, Liu Q, Li C, et al., Direct Observation of Conversion Between Threshold Switching and Memory Switching Induced by Conductive Filament Morphology [J]. Advanced Functional Materials, vol. 24, pp. 5679-5686, 2014.

[43] Yuan F, Zhang Z, Liu C, et al., Real-Time Observation of the Electrode-Size-Dependent Evolution Dynamics of the Conducting Filaments in a SiO_2 Layer [J]. ACS Nano, vol. 11, pp. 4097-4104, Apr 25 2017.

[44] Zhao XL, Ma J, Xiao XH, et al., Breaking the Current-Retention Dilemma in Cation-Based Resistive Switching Devices Utilizing Graphene with Controlled Defects [J]. Advanced Materials, vol. 30, p. 9, Apr 2018.

[45] Midya R, Wang Z, Zhang J, et al., Anatomy of Ag/Hafnia-Based Selectors with 10^{10} Nonlinearity [J]. Adv Mater, Jan 30 2017.

[46] Sivaramakrishnan S, Sterbing-D'Angelo SJ, Filipovic B, et al., GABA(A) synapses shape neuronal responses to sound intensity in the inferior colliculus [J]. J. Neurosci., vol. 24, pp. 5031-5043, May 2004.

[47] Diehl PU and Cook M, Unsupervised learning of digit recognition using spike-timing-dependent plasticity [J]. Frontiers in computational neuroscience, vol. 9, p. 99, 2015.

[48] Froemke RC and Dan Y, Spike-timing-dependent synaptic modification induced by natural spike trains [J]. Nature, vol. 416, pp. 433-438, Mar 2002.

[49] Shouval HZ, Wang SS, and Wittenberg GM, Spike timing dependent plasticity: a consequence of more fundamental learning rules [J]. Frontiers in computational neuroscience, vol. 4, 2010.

[50] Sen Song KDMaLFA, Competitive Hebbian learning through spike-timing-dependent synaptic plasticity [J]. Nature Neuroscience, vol. 3, pp. 919-926, 2000.

[51] Averbeck BB, Latham PE, and Pouget A, Neural correlations, population coding and computation [J]. Nature reviews. Neuroscience, vol. 7, pp. 358-66, May 2006.

[52] Maass W, Noise as a Resource for Computation and Learning in Networks of Spiking Neurons [J]. Proceedings of the Ieee, vol. 102, pp. 860-880, May 2014.

[53] Stromatias E, Neil D, Pfeiffer M, et al., Robustness of spiking Deep Belief Networks to noise and reduced bit precision of neuro-inspired hardware platforms [J]. Front Neurosci, vol. 9, p. 222, 2015.

[54] Pan F, Gao S, Chen C, et al., Recent progress in resistive random access memories: Materials, switching mechanisms, and performance [J]. Materials Science and Engineering: R: Reports, vol. 83, pp. 1-59, 2014.

[55] Hasegawa T, Terabe K, Tsuruoka T, et al., Atomic switch: atom/ion movement controlled devices for beyond von-neumann computers [J]. Adv Mater, vol. 24, pp. 252-67, Jan 10 2012.

[56] Tang J, Yuan F, Shen X, et al., Bridging Biological and Artificial Neural Networks with Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges [J]. Adv Mater, p. e1902761, Sep 24 2019.

[57] Caporale N and Dan Y, Spike timing-dependent plasticity: a Hebbian learning rule [J]. Annual review of neuroscience, vol. 31, pp. 25-46, 2008.

[58] Pfeiffer M and Pfeil T, Deep Learning With Spiking Neurons: Opportunities and Challenges [J]. Front Neurosci, vol. 12, p. 774, 2018.

[59] Alibart F, Zamanidoost E, and Strukov DB, Pattern classification by memristive crossbar circuits using ex situ and in situ training [J]. Nat Commun, vol. 4, p. 2072, 2013.

第5章 用于转换 SNN 的 1T1R 神经元电路设计及系统验证

在第一章中论述到脉冲神经网络（spiking neural network, SNN）与传统的人工神经网络（artificial neural network, ANN）相比理论上具有更强的计算能力和更低的能耗[1-4]。这是因为 SNN 在进行数据处理时采用了可以产生离散脉冲信号的脉冲神经元而不是传统 ANN 中的连续非线性函数，并且还将时间参量引入到了计算过程中，具有异步通信，稀疏编码，事件驱动等优点[5, 6]。然而，到目前为止，SNN 在常用数据库上还不能达到 ANN 所能实现的精度，这大大限制了脉冲神经网络在实际工作中的应用[7, 8]。限制 SNN 发展的因素主要有两个[5, 9]：一个是缺少有效的脉冲训练算法；另一个是缺少基于脉冲形式统计的数据集。当前常用数据集多是面向 ANN 应用的，因而，在用于 SNN 的验证时需要将数据集里表征的模拟量转换为相应频率的脉冲串[10]。这个额外的转换过程会造成精度的损失从而降低了 SNN 数据处理过程中的效率。

为了缓解应用 SNN 处理数据的困境，通过调整权值和神经元参数将 ANN 转换为 SNN 是一种有效的方法[7, 8, 11]。在一定程度上，这种基于转换方法的 SNN 可以同时具有 SNN 的高能效和 ANN 的高精度[12]。实现这种转换的关键是利用脉冲神经元匹配 ANN 中的非线性激活函数。然而，基于 CMOS 电路实现的神经元通常由电容和几十晶体管组成[13]，这大大限制了其可扩展性和大规模的使用。尽管基于新型材料的忆阻器（如 Redox 忆阻器[14-19]，相变忆阻器[20-22]，铁电忆阻器[23-25]，磁忆阻器（STT-MRAMs）[26-30]和 Mott 材料[31-34]等）已被广泛探索用于模拟脉冲神经元的实现，但针对转换 SNN 中神经元电路需求还没有专门设计并且缺少有力的系统实验验证。

基于此，在本工作中，我们提出了一种 1T1R 结构的 Mott 神经元，以满足基于转换方法的 SNN 的需求。该神经元电路的脉冲发放频率依赖于输入电压值的大小并且可以内在的实现人工神经网络中的整流线性单元（ReLU）激活函数。然后利用该神经元，我们首次实验验证了由 640 × 10 的 RRAM 突触阵列和 10 个 1T1R 输出神经元组成的单层 SNN（320×10）。图 5.1 给出了网络及所用神经元函数的原理图。在对 MNIST 手写体数据集的推理过程中，神经元的转换准确率达到 95.7%，并且基于 1T1R 神经元的整流线性电压-频率关系及其固有的随机性，实现了 85.7%的识别准确率。最

后，为了实现并行多任务和更好的系统集成，我们又提出了神经元的X-bar集成结构。这些结果表明，1T1R 神经元在构建大规模转换 SNN 以高效的完成数据处理问题上具有很大的应用潜力。

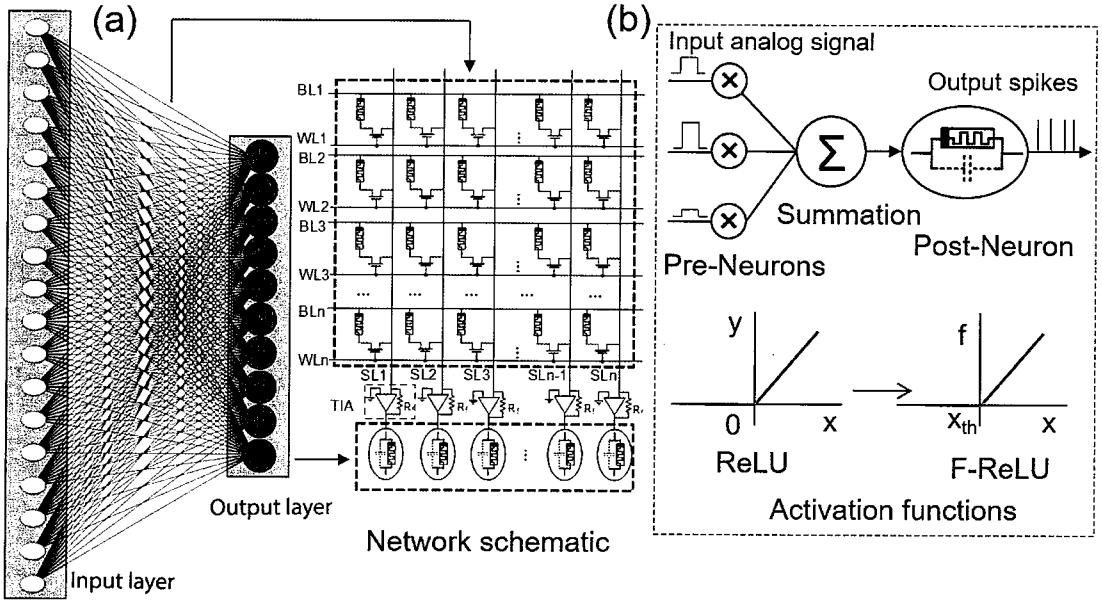


图 5.1 （a）单层转换 SNN 及其相应硬件部分的原理图，模拟输入脉冲输出；（b）ANN 中的 ReLU 函数和 SNN 中的 F-ReLU 激活函数

5.1 基于转换方法的 SNN

2013 年，Perez-Carrasco 等人[8]为了用卷积神经网络处理来自基于事件的传感器的数据首次提出 ANN 到 SNN 转换的方法，把卷积神经网络中的数据处理单元转化为具有漏电和不应期的脉冲神经元。他们的转换方法和其他的转换方法一样，几乎都遵循频率编码的思想，这样模拟神经元的激活函数就转化为脉冲神经元的发射速率。基于此，转换方法便可以充分利用深度学习的工具包，这意味着很多用于任务处理的最新深度学习网络可以直接转换为 SNN，甚至转换的 SNN 可以达到和 ANN 媲美的精度[12]。然而，从 ANN 到 SNN 的转换也有其缺陷：首先，并不是所有的 ANN 都能很容易地转换成 SNN[5]。一个主要的原因是，在 ANN 中，激活函数可以根据输入的正负来判定输出是正值还是负值，而在 SNN 中，神经元的发放频率始终是正的。图 5.2 给出了 ANN 中常用激活函数的原理图，例如 sigmoid 函数，tanh 函数和 ReLU 函数。2015 年，Cao 等人[4]提出了脉冲神经元与传递函数之间的联系，即输入电流和输

出频率与目前 ANN 中单元的标准模型整流线性单元（ReLU）激活函数相匹配。因为 ReLU 激活函数只有在输入为正值的时候才有效并且输出也为正值，其它的输入的情况下输出皆为零[35]。因此，构建满足 ReLU 激活函数的神经元电路是实现转换 SNN 的关键。

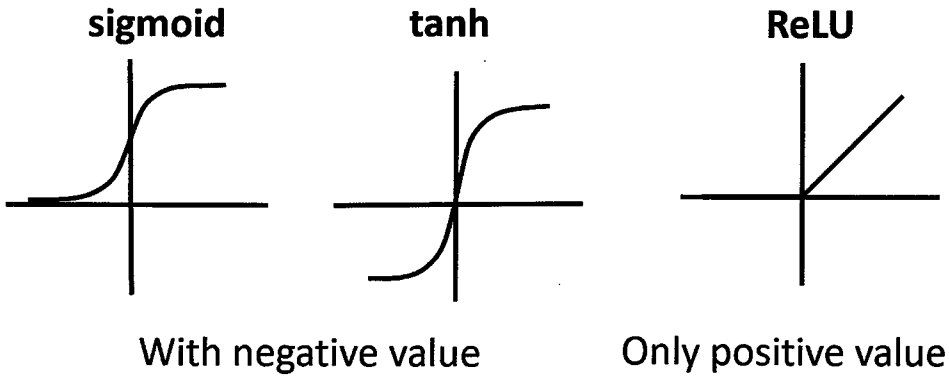


图 5.2 ANN 中常用的非线性激活函数

此外，基于事件的数据集比较稀少，因此传统基于帧的图像数据库，如 MNIST 或 CIFAR 还是常被用来评估转换后 SNN 的准确性[7]。许多工作通常将 ANN 中开始输入的模拟值（例如灰度或 RGB 值）转换为泊松分布的脉冲发放频率[5, 10]，但这会在网络推理的开始引入波动性并损害网络的性能。一个简单的替代方法是在第一个输入层中仍旧使用模拟输入值，而在随后的神经元中采取频率输出[36]，如图 5.3 所示。需要指出的是，基于转换方法的 SNN 只是用在推理的过程，训练的过程主要还是采用 ANN 训练的方法。然而，由于该转换的 SNN 读取的是最后神经元输出的平均频率，因此，推理的精度跟推理的时间相关，这在一定程度上会造成推理的延时，从而损害 SNN 的低能耗性能。尽管如此，与 ANN 相比，在硬件实现上脉冲运算还是比模拟值矩阵乘法高效的多[5]。根据以上转换 SNN 的优点和神经元电路的需求，在本工作中，我们利用 NbO_x 器件实现了符合 ReLU 激活函数的脉冲神经元，并进行了转换 SNN 系统推理的验证。

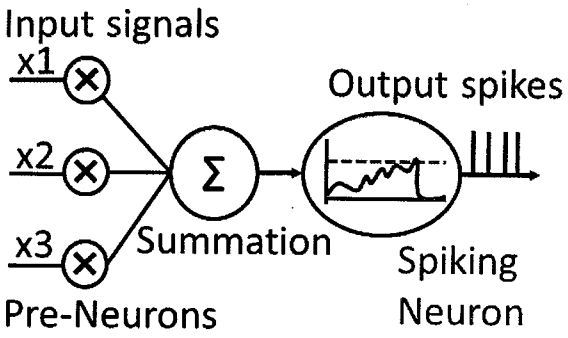


图 5.3 模拟输入脉冲输出的转换 SNN

5.2 NbO_x 器件和 1T1R 神经元电路

5.2.1 NbO_x 器件制备工艺和电学性能表征

该小节主要讲述本工作中所用到的 poly-Si/NbO_x/TiN 器件的制备过程，图 5.4 给出了具体的制备步骤：

(1) 在硅衬底上分别用物理气相沉积法（Physical Vapor Deposition, PVD）和等离子增强化学气相沉积法（Plasma Enhanced Chemical Vapor Deposition, PECVD）沉积一层 SiO_2（150nm）/poly-Si（60nm）/SiO_2（150nm）多层膜。(图 5.4（a）和（e）)

(2) 光刻进行图形化，然后采用一步刻蚀的方法暴露出侧壁光滑的 poly-Si 作为底电极并去胶。(图 5.4 （b）和（f）)

(3) 在室温下用磁控溅射法在侧壁上沉积 50 nm NbO_x 功能层，溅射的过程中使用 NbO_2 靶材并在腔体内通 0.8 SCCM 的氧气以优化薄膜的质量。(图 5.4（c）和（g）)

(4) 保持腔体的真空状态，不对样品进行操作，直接在室温下用磁控溅射法继续沉积 40 nm TiN 作为上电极。

(5) 剥离获得器件图形。(图 5.4（d）和（h）)

这里，为了获得较低的器件转变电流，降低器件能耗，我们采用了热导率较低的 poly-Si 作为下电极。由于 NbO_x 功能层在侧壁上沉积，器件的下电极尺寸为 poly-Si 电极的厚度（~60 nm），上电极的尺寸为 TiN 电极的宽度，在这里分别制备了宽度为 5 μm, 10 μm 和 20 μm 的上电极。对应器件的大小分别为 60 nm × 5 μm, 60 nm × 10 μm 和 60 nm × 20 μm。本工作中使用器件的大小为 60 nm × 20 μm。

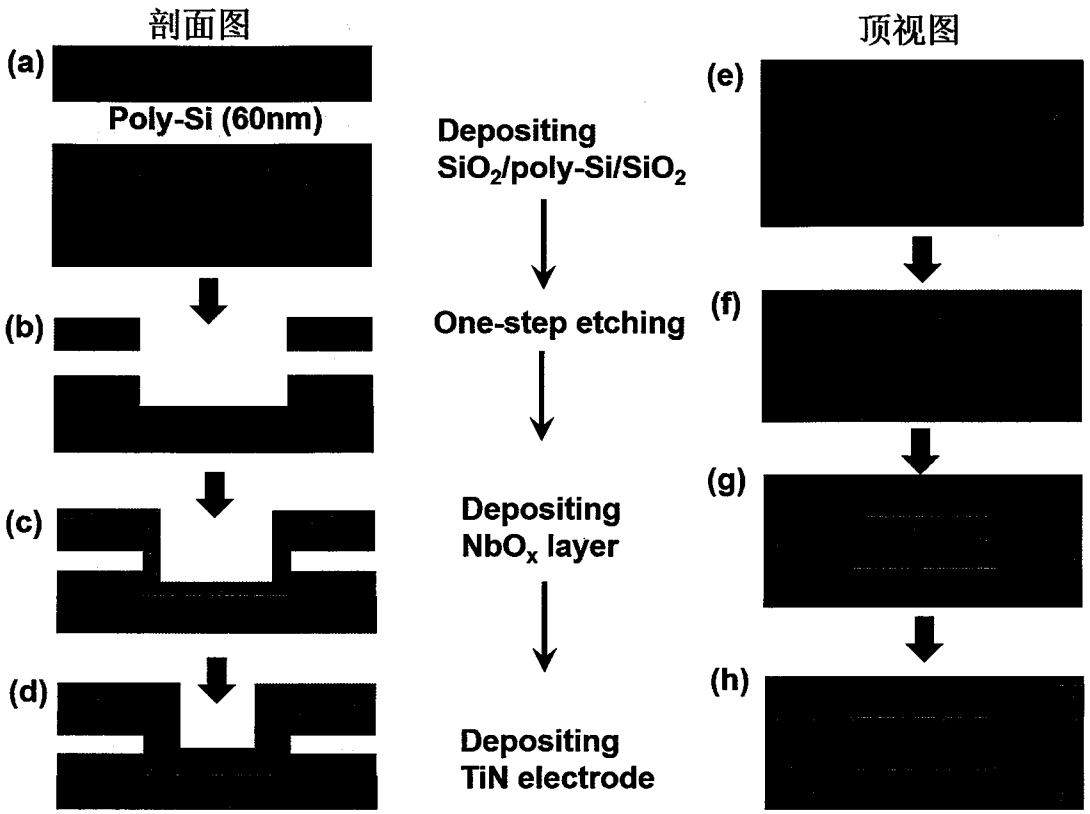


图 5.4 NbO_x 器件的制备流程

由于 NbO_x 薄膜为非晶态，所以器件的初始态为高阻状态。为使器件可以正常工作，需要先对器件进行 forming 操作。Forming 的过程中在 TiN 电极上施加正电压，在电场和热的作用下会在 NbO_x 层内形成 NbO_2 晶体通道，如图 5.5 所示。该 NbO_2 通道决定了器件的转变特性。我们知道 NbO_2 是典型的 Mott 绝缘体金属转变材料，它在大的电压刺激下会从绝缘体状态（HRS）转换到金属状态（LRS），当电压低于一定值时，可以自发地从金属状态回到绝缘体状态。因此，该器件呈现出典型的双向易失性阈值转变行为。

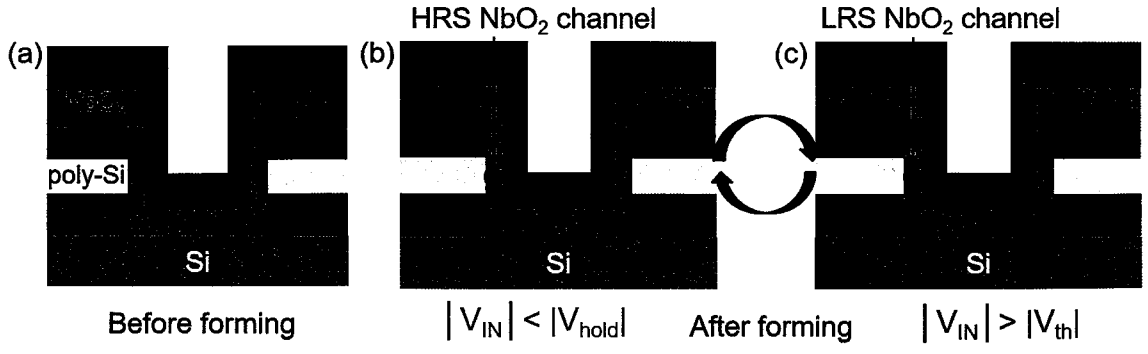


图 5.5 （a）Forming 之前器件初始结构；（b-c）Forming 之后形成 NbO_2 通道，并可以在高低阻态之间切换

图 5.6（a）中给出了器件在 100 μA 限流下正负电压双向扫描的 100 个循环的 I-V 曲线，可以看到对称的双向阈值转变特性。在扫描过程中，扫描电压施加在 TiN 电极上，poly-Si 电极接地。器件在平衡状态时为高阻态，从 0 V→ 3 V 扫描器件，当 TiN 电极上的电压大于 V_{Th} 时，器件由高阻态跳变为低阻态，对应着 NbO_2 通道由绝缘态变为金属态。回扫的过程中，即 3 V→ 0 V，当 TiN 电极上的电压小于保持电压 V_H 时，器件由低阻态自发的回到高阻态，对应着 NbO_2 通道由金属态变为绝缘态。负向电压扫描时也表现出相同的转变特性。在这里 V_{Th} 和 V_H 之间的窗口称为滞回窗口，该窗口的大小决定了神经元电路振荡的幅度大小。图 5.6(b)中显示了阈值电压（V_{Th}）和保持电压（V_H）在正负电压扫描方向上的累积分布。两个参数的紧密分布保证了所构成的神经元电路放电的稳定性。

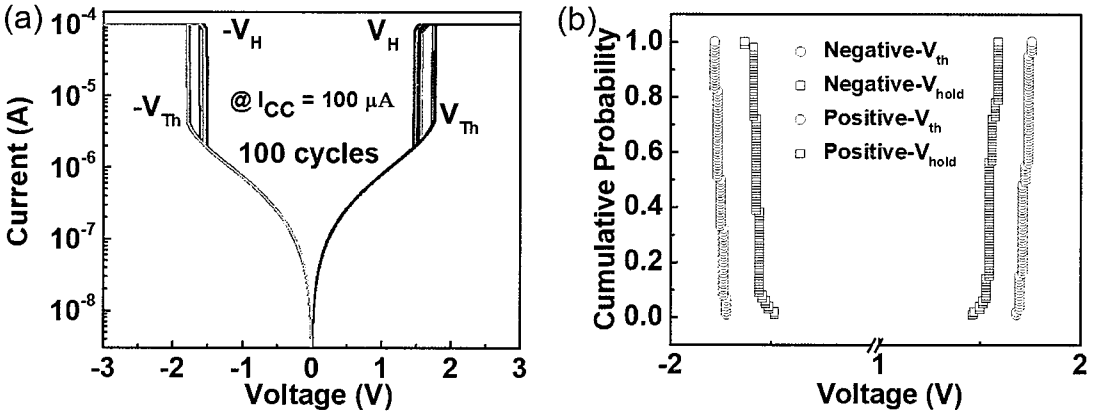


图 5.6 （a）器件的双向易失性阈值转变曲线；（b）V_{Th} 和 V_H 在正负电压扫描下的累积分布

为进一步证明器件的 V_{Th} 和 V_H 在不同扫描周期内的独立性，我们统计了 V_{Th} 和 V_H 在两个方向上的周期-周期波动，如图 5.7（a）所示。可以看出在不同扫描周期内，器件的 V_{Th} 和 V_H 均具有一定的波动性，该波动性的存在是神经元电路具有随机性的内在原因。此外，尽管具有波动性，我们可以看出，每个周期的波动是独立的，即器件的每个周期的 V_{Th} 和 V_H 并不会相互影响。这保证了神经元的每次放电事件的独立

性。为了证明器件在系统应用中的合理性，我们统计了 10 个不同器件的 V_{Th} 和 V_H 的分布情况，如图 5.7（b）所示。统计结果显示不同器件之间的 V_{Th} 和 V_H 在微小波动下的分布几乎相同，这表明了所制备器件的均一性，可以用于系统验证。因为神经元在实际工作过程中只用到了正向电压扫描下的转变，所以我们这里只给出了正向 V_{Th} 和 V_H 的分布情况。

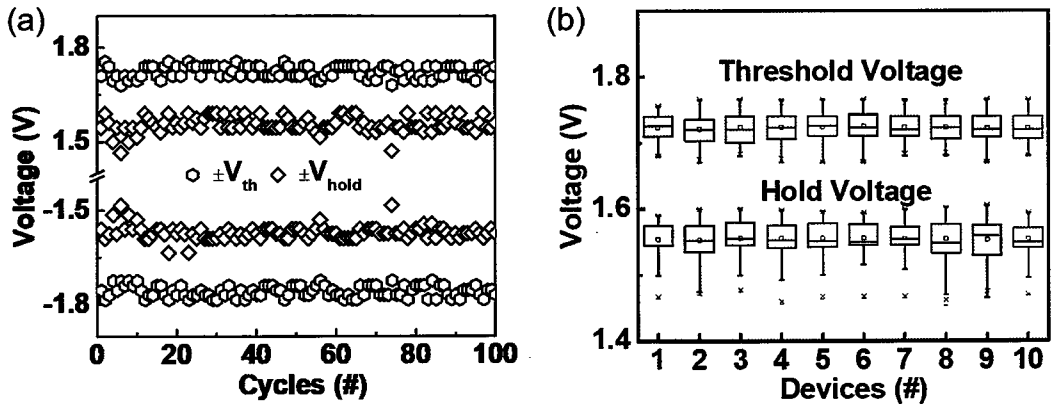


图 5.7 （a）不同转变周期内器件阈值电压和保持电压的统计；（b）10 个不同器件的阈值电压和保持电压的统计比较

5.2.2 1T1R 神经元电路设计及性能表征

考虑到 NbO_x 器件的上述特点以及转换 SNN 神经元电路的需求，我们设计了一种新型的神经元电路，如图 5.8 所示。NbO_x 器件串联到晶体管的漏极上形成 1T1R 的结构，晶体管的栅极作为输入端，NbO_x 器件与晶体管漏极相连的端口作为输出。在这项工作中，我们使用电路中固有寄生电容进行电荷积分，从而避免了外部电容的使用。为了匹配 ReLU 函数，使神经元在负输入下没有脉冲输出，我们使用了 N 型晶体管。图 5.8 右侧给出了神经元的等效电路，其中晶体管的沟道电阻（$R_{channel}$）等效为在不同输入电压下可调的集成电阻，那么，神经元电路工作时就会通过沟道电阻进行充电，通过器件的导通通路进行放电。对于该电路，晶体管最初处于断开状态，因此电压主要落在晶体管上。当我们在栅极上施加电压时，电路将通过晶体管通道对寄生电容充电。一旦 NbO_x 器件上的电压降高于其阈值电压 V_{Th}，器件将打开，然后寄生电容通过 NbO_x 通道放电，直到电压低于保持电压 V_H。因此，该神经元电路可以表现出漏电积分发射神经元的基本特性。由于 $R_{channel}$ 在不同的输入电压下是可调的，因此可以产生不同的积分时间常数，从而产生不同的神经元发射速率。

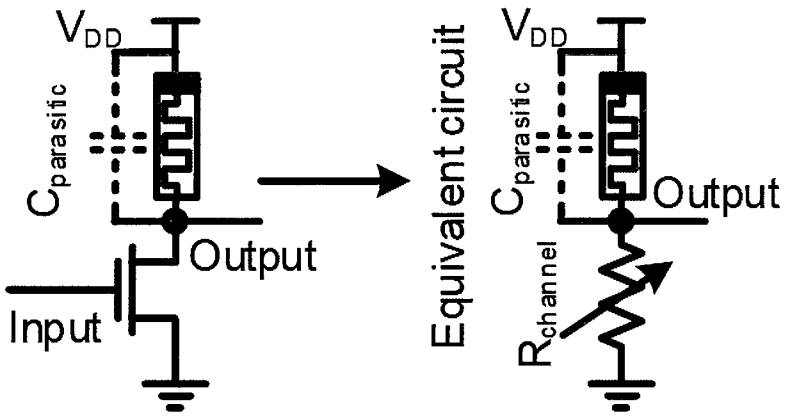
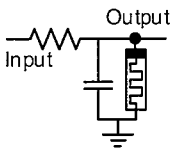
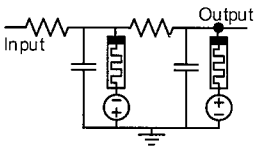
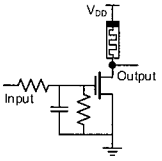
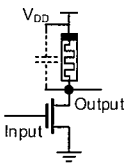


图 5.8 1T1R 神经元电路及其等效电路

为直观展示该神经元的电路结构与其它常用电路结构的不同，我们在表 5.1 给出了该工作与其它已报道 Mott 神经元工作的比较。从表格中可以看出本工作中所使用的神经元电路没有外部电容器的引入，这大大提高了该神经元电路的集成密度。另外，该神经元电路实现的神经元模型是在 SNN 算法中最常见的漏电积分发射（LIF）模型，而且该神经元既可以工作在模拟量输入的模式下又可以工作在脉冲输入的模式下，这拓展了神经元电路的应用范围。从神经元电路所能匹配的激活函数上可以看到，该神经元电路可以实现对转换 SNN 友好的 ReLU 函数，这提供了利用该神经元电路实现转换 SNN 验证的可能性。值得一提的是，由于该神经元电路引入了晶体管，因此在阵列操作中可以作为选通的作用，这使得该电路可以工作在 X-bar 集成阵列中，简化了神经元电路的集成方案。后面我们会进一步在实验上给出系统实现的证明，这也是国际上首次利用新型神经元和突触器件实现可观的系统级效果。

表 5.1 常见 Mott 神经元电路工作的比较

Spiking Neurons Based on Mott Devices				
References	[34, 37]	[31, 32]	[38]	This work
Circuit schematic				
Model	LIF	H-H	LIF	LIF
Input signal	Analog/spikes	Analog/spikes	Spikes	Analog/spikes

Function	#	#	Sigmoid	ReLU
Integration	Low	Low	Low	High (X-bar)
System	Simulation	#	Simulation	Experiment

图 5.9 给出了 1T1R 神经元在不同栅极电压下的输出曲线。操作过程中，在晶体管栅极上施加固定电压，NbO_x 器件 TiN 电极上施加从 0 V 到 3 V 的扫描电压，栅极电压从 1.6 V 变化到 2.2 V。在较低的栅极电压（从 1.6 V 到 1.8 V）下，由于沟道电阻比较大并且和晶体管的限流作用，NbO_x 器件不会发生转变。当栅极电压高于 1.8 V 时，器件开始发生转变，并且滞回窗口随栅极电压的增加而增大。结果显示 1T1R 神经元的总电流比单个晶体管的总电流略低，这是由 NbO_x 器件的电阻导致的。

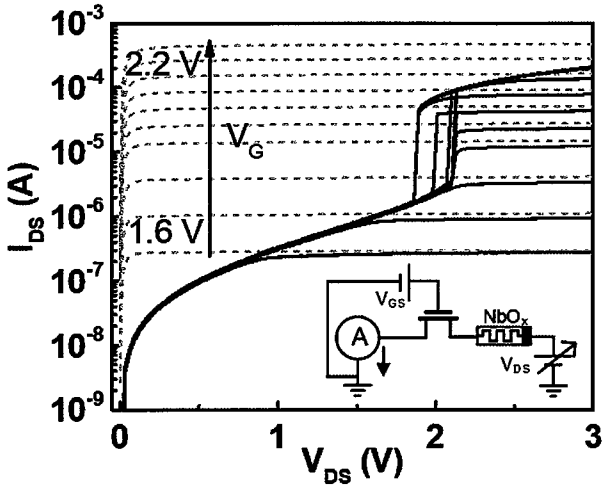


图 5.9 1T1R 神经元电路的输出特性曲线及测试电路原理图

为了与模拟输入脉冲输出的转换 SNN 的操作模式相匹配，我们在栅极上施加模拟电压，并用 Keysight InfiniiVision MSO-X 3104T 示波器测量晶体管漏端的输出电压。测量过程中，NbO_x 的上电极施加 2.5 V 固定的驱动电压。图 5.10 给出了 1.84 V 栅极电压输入下的输出结果，可以看到连续的积分发射行为。

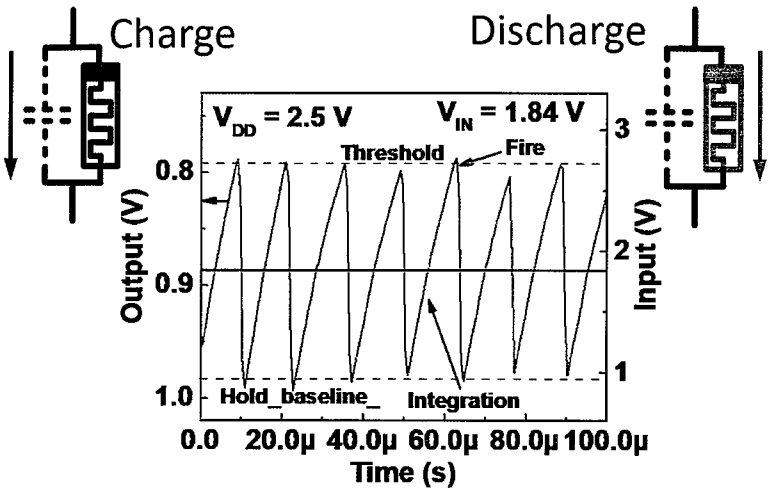


图 5.10 1T1R 神经元电路的积分发射输出

进一步研究该神经元电路的频率响应以面对系统应用，我们在晶体管栅极上施加不同的输入电压得到不同的输出频率以验证神经元在做推理过程中的可行性。图 5.11（a）给出了从 10 个不同神经元之中提取的统计放电频率与输入电压之间的关系，可以看到输出频率随着输入电压的增加线性增加。此外，我们观察到约为 500 kHz 的线性频率范围，在 100 μs 的推理时间内可以得到 50 个 level（> 5 bits）的神经元精度，这对于一般的系统应用来说已经足够了。如果要实现更高的神经元精度可以通过增加推理时间来实现。值得注意的是，正因为使用了 N 型晶体管的 gate 端作为输入端，消除了在负值输入下而产生放电的情况。因为 NbO_x 器件可以同时工作在正电压和负电压下，所以如果只是简单的以固定电阻作为充电电阻的话无法避免负输入下仍旧有脉冲输出的情况，这样就会损害网络的性能[34, 39, 40]。随后，为了匹配 ReLU 激活函数，我们将原点移动到放电起始电压，如图 5.11（b）所示。

为了对比，我们还探究了该神经元电路在 1.9 V 固定栅极电压，不同驱动电压下的频率输出特性，测试原理图如图 5.12（a）所示。Agilent B1500A 的 SMU 通道输出恒定电压加到晶体管的栅极，同时 WGFMU 模块产生不同脉幅的脉冲作为输入并测量流过晶体管沟道的电流。图 5.12（b）给出了不同脉冲输入下的放电输出特性，并在图 5.12（c）中做了统计。可以看到输出频率随输入电压的增加而增加，然而会在较高的电压下趋于饱和，并且线性频率范围（从 2.6 V 到 2.9 V）只有 30 kHz（在 100 μs 推理时间内不到 2 比特），远低于以栅极作为输入，漏端作为驱动端的工作模式。

另外，这种情况下工作电压高于前面的操作模式，增加了系统能量损耗。在前面所使用的神经元的工作模式下，由于晶体管栅极的高输入阻抗，使用栅极作为输入还可以减低前级电路的驱动负载。总而言之，以栅极作为输入，漏端为驱动的工作模式具有更好的神经元特性，更适用于系统功能的验证。

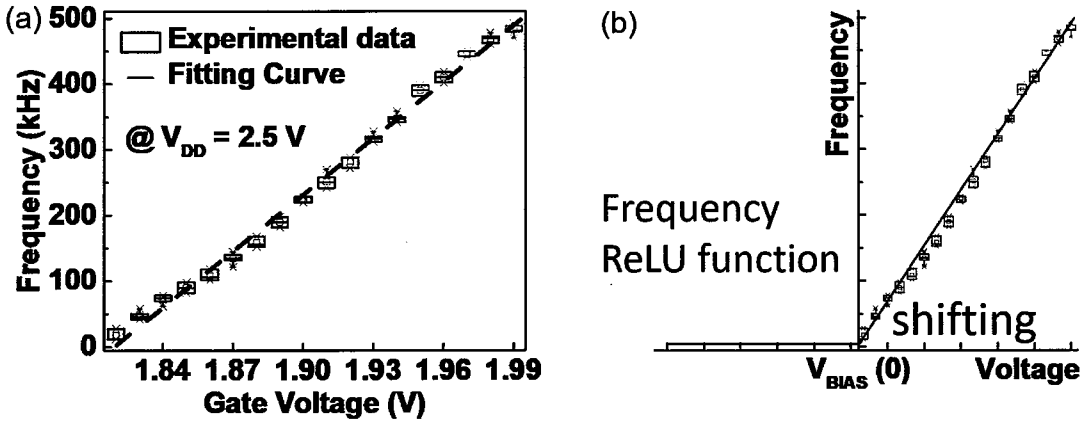


图 5.11 （a）1T1R 神经元电路的输出频率输入电压关系；（b）匹配的频率 ReLU 函数

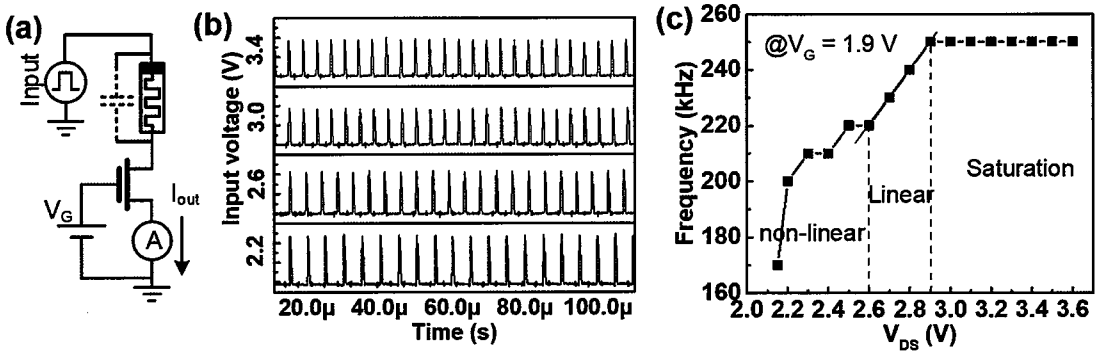


图 5.12 （a）栅极电压固定，漏极作为输入端的原理图；（b）不同输入电压下的输出；（c）输出频率随输入电压的演变关系

5.3 HfO_2 突触器件的电学性能表征

系统的构建除了需要基本的神经元电路以外，还需要突触器件作为连接神经元与存储权值的单元。在本工作中，突触器件是由美国马萨诸塞大学提供的 128×64 的 1T1R 忆阻器阵列，详细制备工艺可以参考 Jiang 和 Wang 等人的工作[41-44]。图 5.13（a）给出了 1T1R 结构的突触器件的原理图。为了在该阵列上验证 ANN 常用的反向

传播等算法，需要突触器件的电导调制过程是线性和对称的，关于该需求已经在很多工作中给出了报道和系统的探讨[45-47]。为了增加突触器件的电导，我们从驱动电路板向忆阻器顶部电极和串联晶体管的栅极施加同步的正电压脉冲[45]（图 5.13（b））。栅极电压用来指定限流的大小，从而决定由此产生的忆阻器电导。逐步增加的栅极电压导致了电导的线性增加。在电导减少的调制过程中，我们先在忆阻器的底电极上施加一个足够的正脉冲来初始化突触器件（即将器件初始化到高阻态），然后使用使电导增加的方案将忆阻器设置到所需的电导水平，这里栅极电压逐渐降低从而逐渐降低了器件电导，编程方案如图 5.13（c）所示。利用该方案，我们实现了电导调制过程中的线性和对称性，且尽可能的降低了周期到周期和器件与器件之间的波动性[45]。

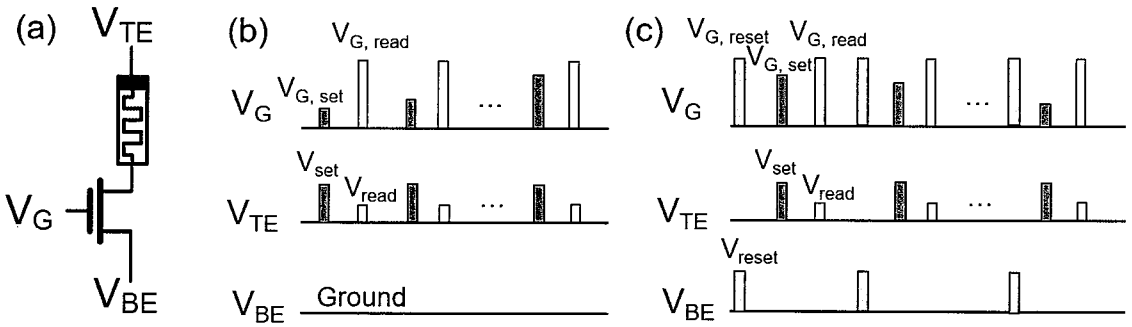


图 5.13 （a）1T1R 突触器件原理图；（b）电导增加调节方案；（c）电导降低调节方案

在实际操作中，根据网络大小的需求，我们使用了 128×50 子阵列作为突触阵列。图 5.14 给出了阵列中所有器件在 40 个脉冲下的长时程增强和长时程抑制的电导演化及对应的器件与器件之间的波动性。可以看到该突触器件在前述脉冲编程方案下具有很好的线性和对称的权值更新行为，支持在 ANN 模式下的训练过程中快速准确的迭代。

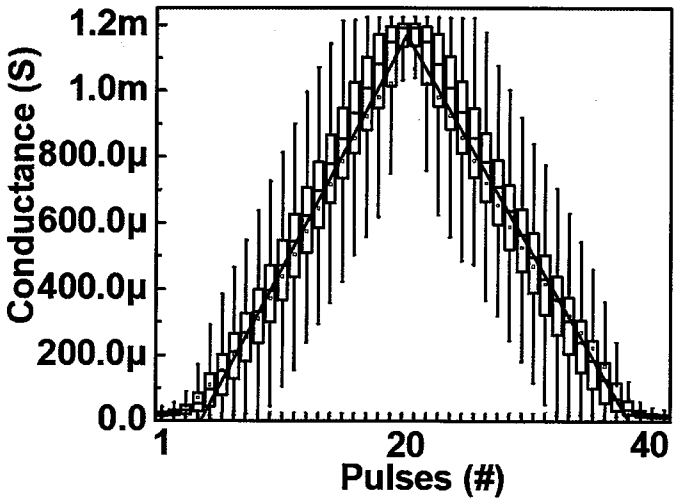


图 5.14 阵列中 128×64 个突触器件（Ta/HfO_2/Pd）在脉冲编程下电导变化曲线

5.4 转换 SNN 的训练和推理过程

5.4.1 基于 1T1R 突触阵列 ANN 网络的训练

为验证本工作提出的 1T1R 神经元的推理功能，我们首先基于上述 1T1R 突触阵列训练了结构为 320 × 10 的单层全连接前向神经网络。使用的数据集为降维并二值化的 MNIST 手写体数据库（每个图像有 20×16 个二值化像素），如图 5.15（a）所示。在测试期间，读取振幅为-0.2/0.2V（差分对电压）的脉冲表示像素“1”，0/0V 表示像素“0”。图 5.15（b）给出了训练流程图。这里，阵列中 128×50 的忆阻器子阵列用来存储权值。在实际操作中，为了表示负权值，我们用一个差分电阻对来表示一个权值，因此每个输入数字模式需要 640 个输入。鉴于阵列规模的限制，每个输入模式被分成 5 个子模式，每个子模式由 64 个像素按顺序输入到阵列中。每个子模式分别在一个 128 × 10 的差分电阻对子阵列上执行。在这种情况下，每个输入数字具有五个读出电流信号。网络训练过程中采用 ReLU 激活函数并执行有监督的 BP 学习规则。计算机充当 ReLU 神经元，计算 delta G 并控制权重阵列的更新。图 5.16 给出了训练后的权值谱图。

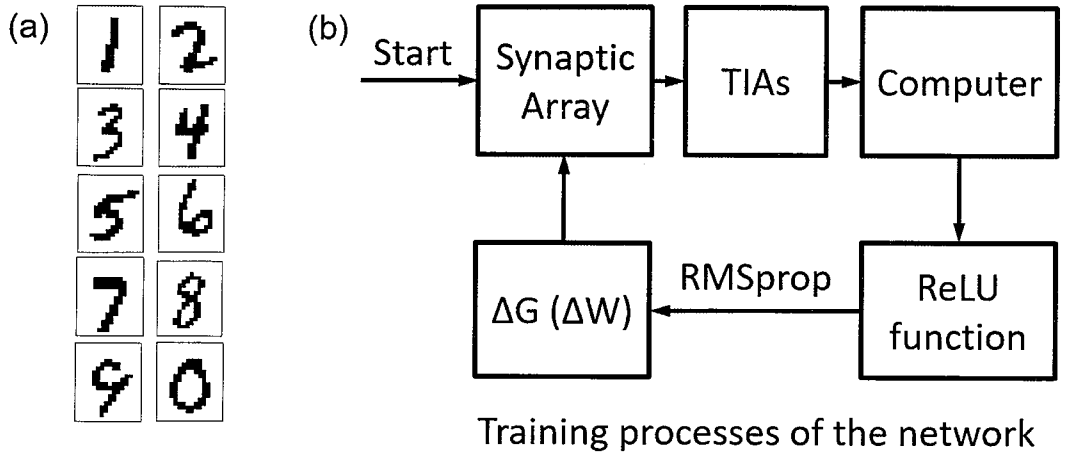


图 5.15 基于 1T1R 突触阵列的 ANN 训练流程

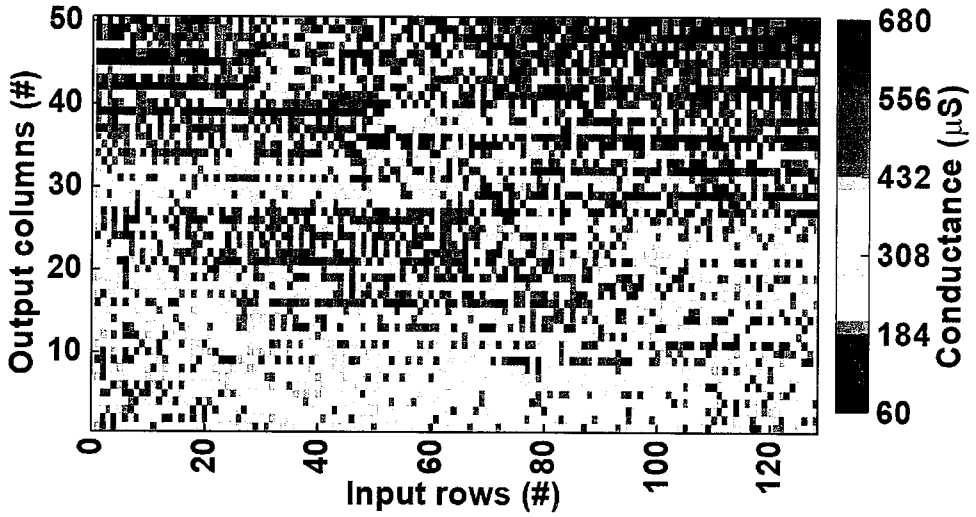


图 5.16 训练后的权值谱图

图 5.17 给出了识别精度与训练周期的关系，可以看到，训练结束后，识别率可以达到 86%左右，比理想情况仅低几个百分点。识别率的降低主要是由于 MNIST 数据集采取了降维的处理方式，这影响了初始输入信号的可信度；另外，忆阻器突触阵列的非理想特性也会影响输出结果。

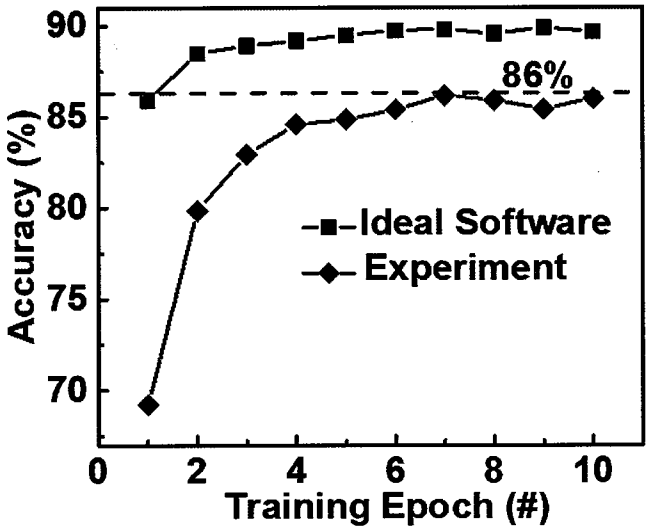


图 5.17 ANN 网络在 MNIST 数据集上的训练结果

5.4.2 1T1R 神经元的推理验证

接下来，我们使用 1T1R 神经元验证推理功能，将神经元与前面训练好的突触阵列通过跨导放大器（TIA）相连，以识别降维后的 MNIST 数据集。此外，为了将神经元的频率输出曲线的初始放电电压作为零点以便与 ReLU 神经元函数相匹配，我们在栅极端施加了 1.81 V 的恒定电压作为偏置电压。图 5.18 给出了推理过程中转换 SNN 的硬件图。每个输入数字（20×16）分成五部分分别输入到阵列中。差分电阻对用于表示负重量，因此每个输入数字模式需要 640 个输入。然后将每个输入数字的 5 个输出（I_1 - I_5）相加，经 TIA 转换成电压作为 1T1R 神经元的输入。

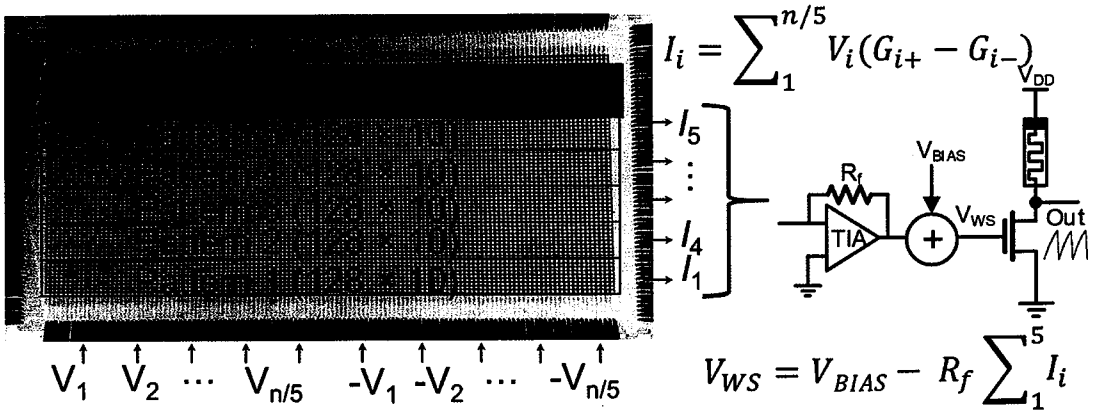


图 5.18 转换 SNN 的硬件原理图

图 5.19 给出了在前 30 个测试数字作为输入情况下神经元“1”的输出结果。在测

试期间，脉宽为 100 μs 脉幅为−0.2/0.2 V（差分对电压）的读脉冲表示像素“1”，0/0 V 表示像素“0”。从放电输出行为可以大致观察到输出频率随输入数字模式的变化而变化。

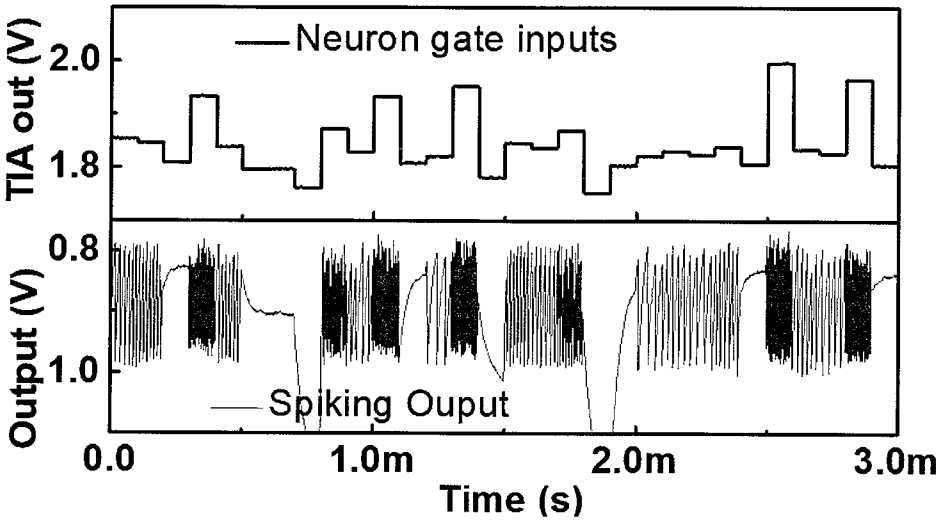


图 5.19 神经元“1”在前 30 个输入数字下的输出结果

图 5.20 显示了在 10 个测试图像下提取的 10 个神经元的峰值频率。每个输入测试图像有 10 个不同的峰值频率，对应于 10 个神经元的输出。对于每个输入模式，观察到 10 个输出神经元具有不同频率的输出，这意味着振荡神经元可以对输入模式进行分类。图中标记为不同输入模式下对应的具有最大输出频率的神经元，最后将输入图像分类为该输出频率最大的神经元。图 5.21 进一步给出了十个输出神经元在 10000 个测试图像下的推理结果。

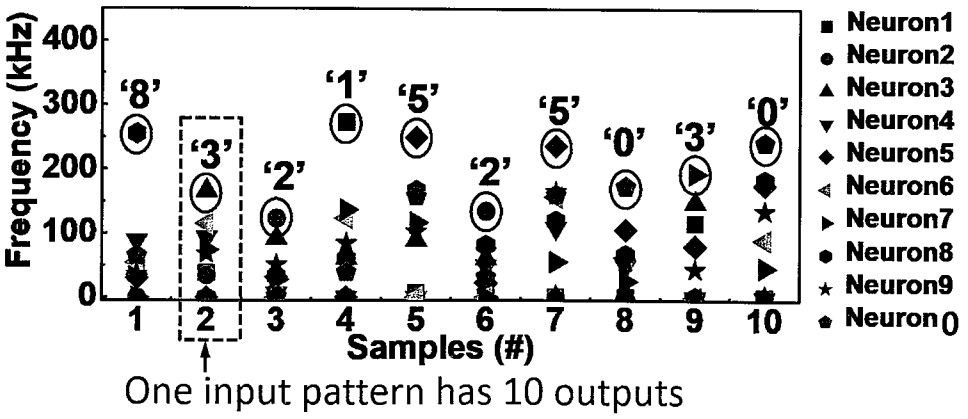


图 5.20 10 个神经元在 10 个输入图像下对应的输出频率

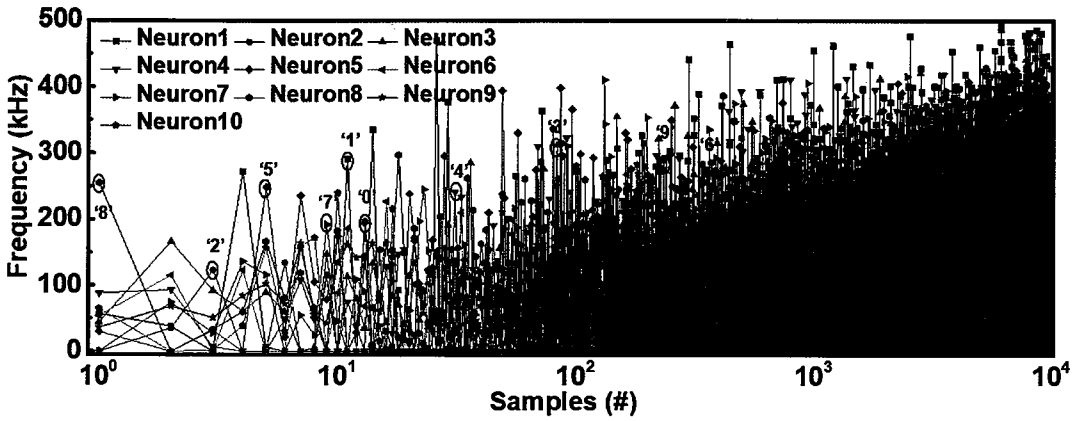


图 5.21 10 个神经元在 10000 个测试图像下对应的输出频率

为了判断推理过程的转换精度和正确率，我们进一步统计了与测试标签相对应的软件神经元的分类错误标签、与软件神经元相对应的 1T1R 神经元的转换错误标签以及与测试标签相对应的 1T1R 神经元的分类错误标签。与 10000 个测试数据中的正确标签相比，ReLU 软件神经元有 1404 个错误标签，识别率达到 86%（图 5.22 的顶部面板）。图 5.22 的中间面板显示了 1T1R 神经元与软件神经元的转换错误标签，其中 247 个标签与软件神经元不匹配，对应转换准确率为 97.5%。然而，在这 247 个错误标签中，有 85 个标签（红色）与测试标签相比是正确的。这是因为 1T1R 神经元的随机性有一定的概率降低错误神经元的输出频率，提高正确神经元的输出频率。为了验证这一假设，我们提取了 1T1R 神经元校正的十个软件神经元的输出结果（表 5.2）。可以看到，错误神经元的输出值确实略高于实际神经元的输出值，而后者本应该具有最大输出值。图 5.22 的底部面板给出了 1T1R 神经元与测试标签相比的错误标签（1426 个）。在 10000 幅测试图像中，1T1R 神经元获得了 85.7%的准确率，与软件神经元获得的结果基本一致。值得注意的是，利用多层神经网络或卷积神经网络可以进一步优化识别精度。这些结果表明，本文中提出的 1T1R 神经元能够很好地将模拟信号转换成频率信号，并能利用器件固有的随机特性优化识别精度，为实现基于新型神经元硬件的转换 SNN 奠定了基础。

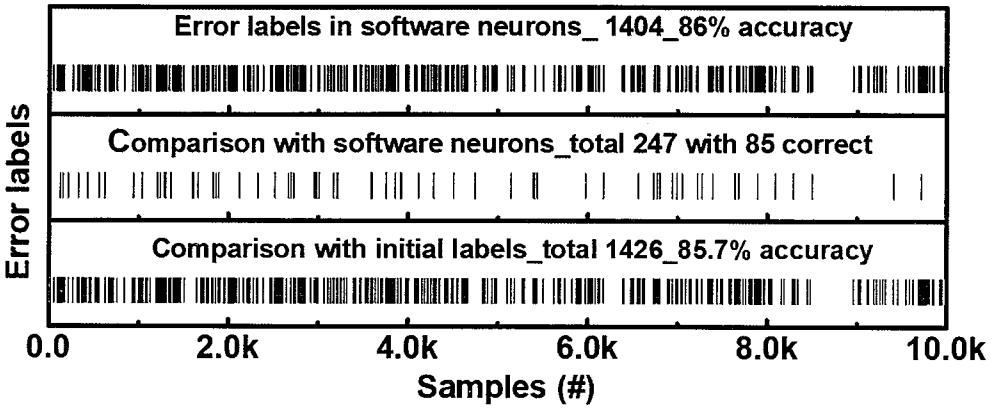


图 5.22 软件神经元和 1T1R 神经元在推理过程中错误标签的对比

表 5.2 1T1R 神经元校正的十个输入模式对应的软件神经元的输出结果：红色表示最大神经元输出值，蓝色表示次值神经元输出

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
N1	0.132	0.089	0	0.006	0.133	0	0.283	0	0	0.283
N2	0.276	0	0	0	0.229	0	0	0.052	0.218	0
N3	0.268	0	0.481	0.218	0.129	0.225	0.346	0	0.385	0.346
N4	0.262	0.247	0.479	0	0.267	0	0	0.239	0.388	0
N5	0	0.176	0	0.439	0	0.369	0.015	0.386	0	0.015
N6	0.214	0.349	0.165	0	0.266	0.133	0.119	0.161	0.117	0.119
N7	0.046	0.047	0.005	0.34	0	0.373	0.347	0.14	0.014	0.347
N8	0	0.237	0.208	0.243	0.165	0.171	0	0.358	0.131	0
N9	0.241	0.352	0	0.345	0.2	0.27	0.097	0.176	0.4	0.097
N10	0	0.318	0	0.436	0.094	0.155	0	0.381	0.044	0

*N→Neuron；P→Pattern

5.5 1T1R 神经元的 X-bar 集成架构

X-bar 结构通常用于快速数据传输网络[48]，基于忆阻器 X-bar 结构的突触阵列在执行复杂任务中具有强大的并行计算能力。理想情况下，执行不同任务的神经网络需要集成到一个芯片中，用于神经机器人的多功能实现[49]。在这里，为了实现神经元的友好集成，我们基于 1T1R 神经元提出了神经元的 X-bar 结构，如图 5.23 所示。该 X-bar 结构由 1T1R 神经元构成，X-bar 中的每个 1T1R 单元都充当一个单独的神经元。多路复用器（MUX）和驱动模块用于激活工作列。在执行推理期间，每行用于执行一

个任务。共享读出电路用于并行读出每一列的输出频率信号。

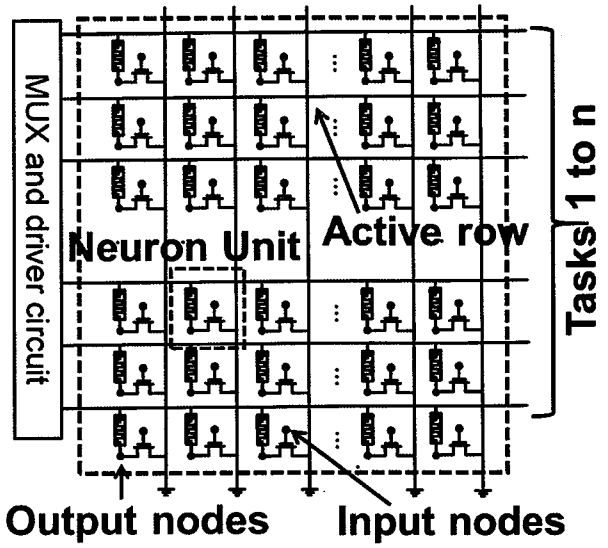


图 5.23 1T1R 神经元的 X-bar 集成方案原理图

为了验证这个设计，我们进行了基于 10×10 神经元 X-bar 的仿真。首先，随机生成一个电压模式来表示 10 个任务的 TIA 输出，如图 5.24（a）所示。每个列都是一个任务的 TIA 输出。然后将生成的图形并行地应用于神经元 X-bar 的栅上端。输出结果如图 5.24（b）所示，其中输出峰值强度与 TIA 输出类似。这些结果表明，所提出的神经元 X-bar 结构可以并行执行多个任务。

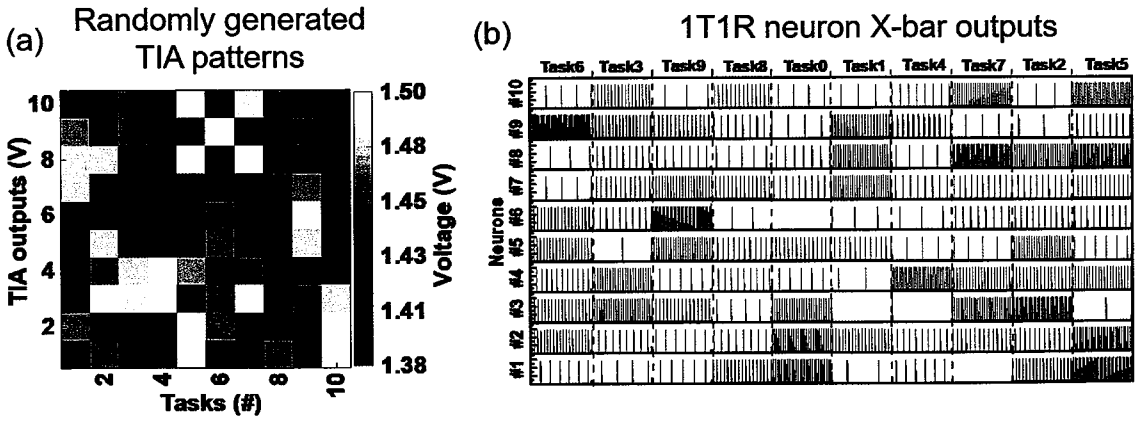


图 5.24 （a）随机生成的 TIA 的输出模式；（b）1T1R 神经元的并行输出仿真结果

5.6 本章小结

基于转换方法的 SNN 可以同时实现 SNN 的高能效和 ANN 的高精度，就目前形

势来讲，是 SNN 推向实际应用的有效方案之一。为面向转换 SNN 的实现，我们开展了以下工作：

1) 基于 NbO_x-Mott 忆阻器提出了一种 1T1R 结构的脉冲神经元来匹配 ANN 中的 ReLU 函数，该神经元利用了电路内部固有的寄生电容，便于高密度的集成；

2) 在这项工作中，我们首次实验验证了基于 Mott 型 1T1R 神经元和忆阻器突触的单层 SNN（320×10）。网络第一层中保留模拟输入，以缓解基于脉冲的数据集的不足。实验结果表明，虽然神经元的转换误差为 4.7%，但神经元固有的随机性有利于优化推理结果，使其准确性（85.7%）与软件神经元（86%）接近；

3) 最后，为了实现多任务并行和更好的系统集成，提出了 1T1R 神经元的 X-bar 结构。

这些结果表明，本文提出的 1T1R 神经元有希望在未来构建大规模的转换 SNN 网络芯片以高效的执行边缘计算的任务。

本章的研究内容在 2019 年 12 月份于旧金山举办的国际电子器件会议上（IEDM 2019）做了口头报告并在线发表。

参考文献

[1] Maass W, Networks of spiking neurons: The third generation of neural network models [J]. Neural Networks, vol. 10, pp. 1659-1671, Dec 1997.

[2] Merolla PA, Arthur JV, Alvarez-Icaza R, et al., A million spiking-neuron integrated circuit with a scalable communication network and interface [J]. Science, vol. 345, pp. 668-673, 2014.

[3] Imam N and Cleland TA, Rapid online learning and robust recall in a neuromorphic olfactory circuit [J]. Nature Machine Intelligence, vol. 2, pp. 181-191, 2020.

[4] Cao Y, Chen Y, and Khosla D, Spiking Deep Convolutional Neural Networks for Energy-Efficient Object Recognition [J]. International Journal of Computer Vision, vol. 113, pp. 54-66, 2014.

[5] Pfeiffer M and Pfeil T, Deep Learning With Spiking Neurons: Opportunities and Challenges [J]. Front Neurosci, vol. 12, p. 774, 2018.

[6] Davies M, Srinivasa N, Lin T-H, et al., Loihi: A Neuromorphic Manycore Processor with On-Chip Learning [J]. Ieee Micro, pp. 82-99, 2018.

[7] Rueckauer B, Lungu IA, Hu Y, et al., Conversion of Continuous-Valued Deep Networks to Efficient Event-Driven Networks for Image Classification [J]. Front Neurosci, vol. 11, p. 682, 2017.

[8] Perez-Carrasco JA, Bo Z, Serrano C, et al., Mapping from Frame-Driven to Frame-Free Event-Driven Vision Systems by Low-Rate Rate Coding and Coincidence Processing--Application to Feedforward ConvNets [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 2706-2719, 2013.

[9] Roy K, Jaiswal A, and Panda P, Towards spike-based machine intelligence with neuromorphic computing [J]. Nature, vol. 575, pp. 607-617, Nov 2019.

[10] Diehl PU and Cook M, Unsupervised learning of digit recognition using spike-timing-dependent plasticity [J]. Frontiers in computational neuroscience, vol. 9, p. 99, 2015.

[11] Pei J, Deng L, Song S, et al., Towards artificial general intelligence with hybrid Tianjic chip architecture [J]. Nature, vol. 572, pp. 106-111, Aug 2019.

[12] Sengupta A, Ye Y, Wang R, et al., Going Deeper in Spiking Neural Networks: VGG and Residual Architectures [J]. Front Neurosci, vol. 13, p. 95, 2019.

[13] Indiveri G, Linares-Barranco B, Hamilton TJ, et al., Neuromorphic silicon neuron circuits [J]. Front Neurosci, vol. 5, p. 73, 2011.

[14] Jang J-W, Attarimashalkoubeh B, Prakash A, et al., Scalable Neuron Circuit Using Conductive-Bridge RAM for Pattern Reconstructions [J]. Ieee T Electron Dev, vol. 63, pp. 2610-2613, 2016.

[15] Mehonic A and Kenyon AJ, Emulating the Electrical Activity of the Neuron Using a Silicon Oxide RRAM Cell [J]. Front Neurosci, vol. 10, p. 57, 2016.

[16] Lashkare S, Chouhan S, Chavan T, et al., PCMO RRAM for Integrate-and-Fire Neuron in Spiking Neural Networks [J]. Ieee Electr Device L, vol. 39, pp. 484-487, Apr 2018.

[17] Wang Z, Joshi S, Savel'ev S, et al., Fully memristive neural networks for pattern classification

with unsupervised learning [J]. Nature Electronics, vol. 1, pp. 137-145, 2018.

[18] Wang Z, Rao M, Han JW, et al., Capacitive neural network with neuro-transistors [J]. Nat Commun, vol. 9, p. 3208, Aug 10 2018.

[19] Zhang Y, He W, Wu Y, et al., Highly Compact Artificial Memristive Neuron with Low Energy Consumption [J]. Small, p. e1802188, Nov 14 2018.

[20] Tuma T, Pantazi A, Le Gallo M, et al., Stochastic phase-change neurons [J]. Nat Nanotechnol, May 16 2016.

[21] Pantazi A, Wozniak S, Tuma T, et al., All-memristive neuromorphic computing with level-tuned neurons [J]. Nanotechnology, vol. 27, p. 355205, Sep 2 2016.

[22] Wright CD, Hosseini P, and Diosdado JAV, Beyond von-Neumann Computing with Nanoscale Phase-Change Memory Devices [J]. Advanced Functional Materials, vol. 23, pp. 2248-2254, 2013.

[23] Chen C, Yang M, Liu S, et al., Bio-Inspired Neurons Based on Novel Leaky-FeFET with Ultra-Low Hardware[M]. New York, 2019.

[24] Dutta S, Saha A, Panda P, et al., Biologically Plausible Ferroelectric Quasi-Leaky Integrate and Fire Neuron[M]. New York: Ieee, 2019.

[25] Mulaosmanovic H, Chicca E, Bertele M, et al., Mimicking biological neurons with a nanoscale ferroelectric transistor [J]. Nanoscale, vol. 10, pp. 21755-21763, Dec 2018.

[26] Akhilesh Jaiswal SR, Gopalakrishnan Srinivasan, and Kaushik Roy, Proposal for a Leaky-Integrate-Fire Spiking Neuron Based on Magnetoelectric Switching of Ferromagnets [J]. Ieee T Electron Dev, vol. 64, 2017.

[27] Li S, Kang W, Huang Y, et al., Magnetic skyrmion-based artificial neuron device [J]. Nanotechnology, Jun 22 2017.

[28] Wu MH, Hong MC, Chang C-C, et al., Extremely Compact Integrate-and-Fire STT-MRAM Neuron: A Pathway toward All-Spin Artificial Deep Neural Network[M]. New York: Ieee, 2019.

[29] Romera M, Talatchian P, Tsunegi S, et al., Vowel recognition with four coupled spin-torque nano-oscillators [J]. Nature, vol. 563, pp. 230-234, Nov 2018.

[30] Sengupta A, Panda P, Wijesinghe P, et al., Magnetic Tunnel Junction Mimics Stochastic Cortical Spiking Neurons [J]. Sci Rep, vol. 6, p. 30039, Jul 21 2016.

[31] Pickett MD, Medeiros-Ribeiro G, and Williams RS, A scalable neuristor built with Mott memristors [J]. Nature Materials, vol. 12, pp. 114-117, 2013.

[32] Yi W, Tsang KK, Lam SK, et al., Biological plausibility and stochasticity in scalable VO_2 active memristor neurons [J]. Nat Commun, vol. 9, p. 4661, Nov 7 2018.

[33] Stoliar P, Tranchant J, Corraze B, et al., A Leaky-Integrate-and-Fire Neuron Analog Realized with a Mott Insulator [J]. Advanced Functional Materials, p. 1604740, 2017.

[34] Gao L, Chen P-Y, and Yu S, NbO_x based oscillation neuron for neuromorphic computing [J]. Appl Phys Lett, vol. 111, p. 103503, 2017.

[35] Li Y and Yuan Y, Convergence Analysis of Two-layer Neural Networks with ReLU Activation, presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., 2017.

[36] Zambrano D and Bohte SM, Fast and Efficient Asynchronous Neural Computation with Adapting Spiking Neural Networks [J]. arXiv preprint arXiv:1609.02053, 2016.

[37] Lin J, Annadi A, Sonde S, et al., Low-voltage artificial neuron using feedback engineered insulator-to-metal-transition devices, in 2016 Ieee International Electron Devices Meeting (IEDM), ed New York: Ieee, 2016.

[38] Jerry M, Parihar A, Grisafe B, et al., Ultra-Low Power Probabilistic IMT Neurons for Stochastic Sampling Machines[M]. New York: Ieee, 2017.

[39] Midya R, Wang Z, Asapu S, et al., Artificial Neural Network (ANN) to Spiking Neural Network (SNN) Converters Based on Diffusive Memristors [J]. Advanced Electronic Materials, p. 1900060, 2019.

[40] Chen P-Y, Seo J-S, Cao Y, et al., Compact oscillation neuron exploiting metal-insulator-transition for neuromorphic computing [J]. pp. 1-6, 2016.

[41] Jiang H, Han L, Lin P, et al., Sub-10 nm Ta Channel Responsible for Superior Performance of a HfO2 Memristor [J]. Sci Rep, vol. 6, p. 28525, 2016.

[42] Wang Z, Li C, Song W, et al., Reinforcement learning with analogue memristor arrays [J]. Nature Electronics, vol. 2, pp. 115-124, 2019.

[43] Wang Z, Li C, Lin P, et al., In situ training of feed-forward and recurrent convolutional memristor networks [J]. Nature Machine Intelligence, vol. 1, pp. 434-442, 2019.

[44] Li C, Hu M, Li Y, et al., Analogue signal and image processing with large memristor crossbars [J]. Nature Electronics, 2017.

[45] Li C, Belkin D, Li Y, et al., Efficient and self-adaptive in-situ learning in multilayer memristor neural networks [J]. Nat Commun, vol. 9, p. 2385, Jun 19 2018.

[46] Wang IT, Chang CC, Chiu LW, et al., 3D Ta/TaO x /TiO2/Ti synaptic array and linearity tuning of weight update for hardware neural network applications [J]. Nanotechnology, vol. 27, p. 365204, Sep 9 2016.

[47] Woo J and Yu SM, Resistive Memory-Based Analog Synapses The pursuit for linear and symmetric weight update [J]. IEEE Nanotechnol. Mag., vol. 12, pp. 36-44, Sep 2018.

[48] Jeffrey J. Nelson LKN and Jessop A, High performance crossbar switch, United States Patent, 2000.

[49] Shulaker MM, Hills G, Park RS, et al., Three-dimensional integration of nanotechnologies for computing and data storage on a single chip [J]. Nature, vol. 547, pp. 74-78, Jul 05 2017.

第6章 人工传入神经及机械感受系统实现

在前面章节中，我们分别探讨了如何利用忆阻器实现人工突触和神经元，进而构建高效的脉冲神经网络。至此，构建的该脉冲神经网络主要用于对已经接收到的信号的处理进而完成任务决策。然而，传感器从环境中采集的信号通常为模拟信号，如果想要获得进一步的决策结果，首先需要将传感器感知到的模拟信号转换为脉冲神经网络可以处理的脉冲信号，然后再将该具有感知信息的脉冲信号输入到网络中进行进一步的处理[1-3]。在生物神经系统中，传入神经完成这样的一个功能，它将接收来自感受器的信号转换为动作电位信号，传递给中枢神经系统和大脑进行进一步的处理[4]。受生物神经系统的启发，为实现集感存算于一体的智能处理系统，需要构建一种类似于生物体中传入神经的特殊单元来实现传感器的模拟信号到神经网络可以处理的脉冲信号的转换。基于传统的CMOS电路构建的传入神经通常都是用多个反相器构成锁相环电路实现模拟信号到脉冲频率信号的转变。例如，斯坦福大学鲍哲南课题组于2015年和2019年分别在Science上发表工作利用锁相环电路实现人工传入神经，其输出频率与生物传入神经的动作电位频率相匹配[5, 6]。并在2019年的工作中利用锁相环结合压力感受器和突触晶体管构成了完整的传入神经环路来控制生物运动神经。然而，由于传统CMOS结构的器件缺乏内在的动态特性导致传入神经电路复杂，不利于微缩和大规模集成，且CMOS器件即将达到其物理瓶颈，因此研究如何利用忆阻器内在的动态特性构建传入神经电路是一种有效的解决方案。NbO_x忆阻器是一种高集成度的双端器件，它具有负微分电阻（Negative Differential Resistance, NDR）行为和丰富的器件物理动力学[7-11]，能够用来模拟生物神经元和进行模拟计算。[12-15]

基于此，在这项工作中，我们基于制备的NbO_x忆阻器报道了一种紧凑的人工脉冲传入神经（Artificial Spiking Afferent Nerve, ASAN）电路。为了构造该ASAN电路，首先验证了一个由NbO_x器件和电阻组成的紧凑NbO_x振荡器，它可以将模拟输入信号转

换成相关的脉冲振荡输出频率。在该ASAN电路中，输入刺激与传感器产生的电压相关，振荡频率与神经元的脉冲频率相关，且该脉冲频率又取决于传感器产生的电压强度。ASAN电路在适当的刺激下显示了输入强度和峰值频率之间的准线性关系，并且在遭遇有害刺激时倾向于降低放电频率，这生动地模拟了生物传入神经的功能[16, 17]。为了充分验证该ASAN的放电特性，将三种不同类型的输入脉冲（方波、三角波和正弦波）分别作为ASAN的输入刺激信号，得到了一致性的结果。在此基础上，我们又进一步提出了一种基于压电传感器的零静态功耗机械感受系统，并进行了实验验证。

6.1 生物和人工脉冲机械感受系统

该工作的设计实现受启发于生物神经系统，图 6.1 展示了我们的机械脉冲感受系统的原理图以及其对应的生物机械感受系统。在生物体中，传入神经接收皮肤下触觉感受器的信号并将信号的强弱转换为对应动作电位的发射频率，进而传送到大脑皮层的中枢神经系统进行进一步的处理（图 6.1（a））。然后中枢神经系统将处理后的信号通过传出神经传递给效应器，以对外部或内部环境做出响应[18]。我们的人工脉冲机械感受系统（图 6.1（b））由一个两端传感器和一个紧凑的振荡器组成，其中紧凑的振荡器作为人工传入神经 （artificial spiking afferent nerve, ASAN）。该 ASAN 主要包含两个无源元件：一个电阻器和一个 NbO_x 忆阻器。在生物系统中，当输入刺激强度超过传入神经阈值时，传入神经的动作电位发放频率会随着刺激强度的增加而增加。然而，当刺激强度过高时，神经元细胞为防止受到伤害而表现出保护性抑制行为，神经元的活性降低，从而放电频率会降低[17]。因此，生物传入神经的放电频率与刺激强度呈现出先增加后减少的关系。在我们的人工机械感受系统中，传感器产生模拟电压信号作为输入，该电压信号的强弱与传感器接收到的刺激强度呈正相关。NbO_x 基 ASAN 可以将传感器产生的电压信号的大小转换成相应的脉冲发放频率，然后将产生的脉冲频率信号传输到高阶脉冲神经网络进行进一步处理。值得注意的是，在正常刺激强度下，ASAN 的脉冲发放频率与刺激强度成准线性关系。一旦传感器产生的电压强度过高，ASAN 的放电频率开始降低，直到停止发放脉冲信号，和生物传入神经的行为一致。ASAN 的放电周期主要由积分时间和弛豫时间决定，在正常输入刺激强度

下的准线性频率响应是因为该过程中 ASAN 的积分时间占主导地位，随着输入强度的增加积分时间逐渐减少，从而发放频率逐渐增加。而在较高的输入强度下，NbO_x 忆阻器打开后的弛豫时间变长，最终放电周期主要由弛豫时间决定，从而导致频率在高刺激强度下逐渐下降。在后文中，我们仔细讨论了导致弛豫时间增加的原因，主要有两个：第一，输入强度的增加导致在放电后流过忆阻器的总电流增加，这需要更长的弛豫时间。第二，总电流的增加导致在弛豫过程中忆阻器中产生更大的焦耳热，这使得器件在其导通状态上停留的时间更长[11]。ASAN 在充电时间和弛豫时间相等时到达最大发放频率，一旦弛豫时间超过积分时间，脉冲发放频率开始下降。最终，过高强度刺激下器件保持其导通状态时，ASAN 停止发放脉冲。

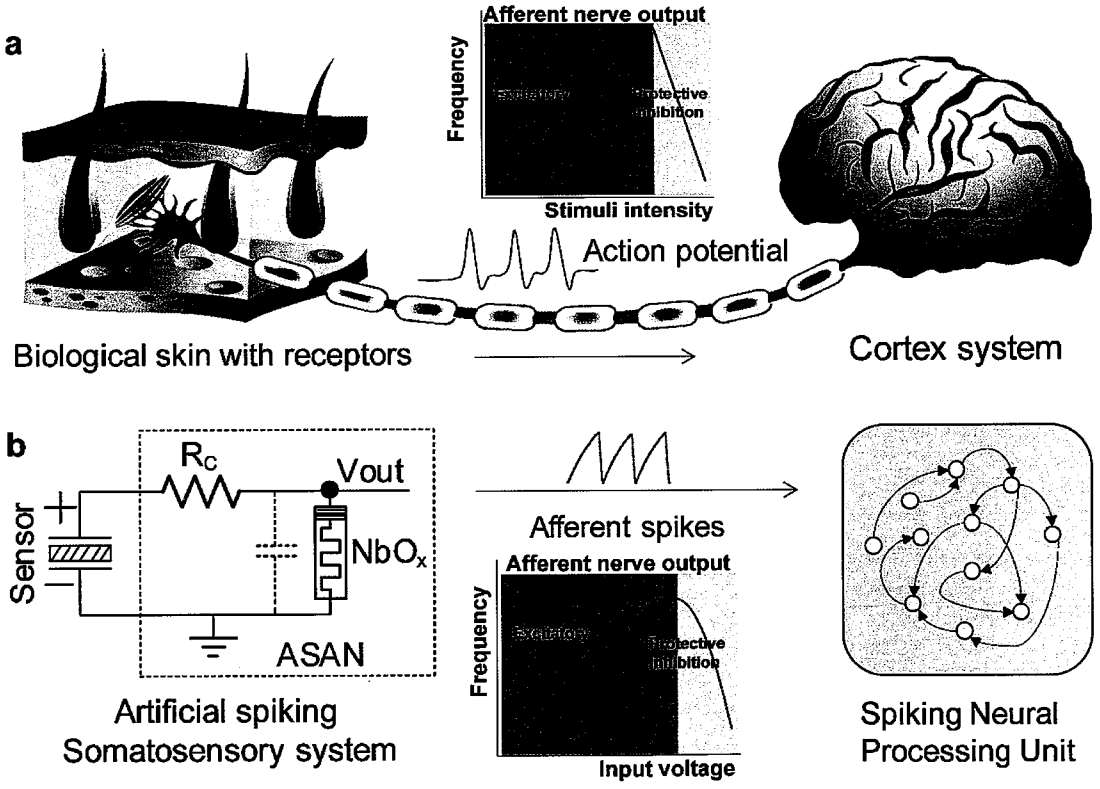


图 6.1 （a）生物机械感受系统原理图；（b）人工脉冲机械感受系统（artificial spiking somasensory system），由传感器（sensor）和人工传入神经电路（artificial spiking afferent nerve, ASAN）构成

6.2 NbO_x 器件特性和模型

6.2.1 NbO_x 器件的制备及表征

本工作中所用器件的制备工艺流程与第五章中 NbO_x 器件的制备流程一致，在此不再赘述，此外，在该工作中使用器件的大小为 60 nm × 20 μm。

为进一步表征器件的材料体系，在这里，我们对器件进行透射电镜（transmission electron micrograph, TEM）测试。图 6.2（a）给出了器件的高分辨率截面 TEM 图像。可以清楚地看到 poly-Si/NbO_x/TiN 的三明治结构。由于溅射薄膜的过程是垂直于样品表面，而 NbO_x 层是在侧面沉积的，所以最后沉积的 NbO_x 的有效厚度大约为 25 nm，小于理论上的淀积厚度（50 nm）。那么，器件的有效工作厚度也就为 25 nm。图 6.2（b）到图 6.2（f）是对器件内各元素成分的元素谱分析，可以清楚的显示出各元素在器件中的分布情况。

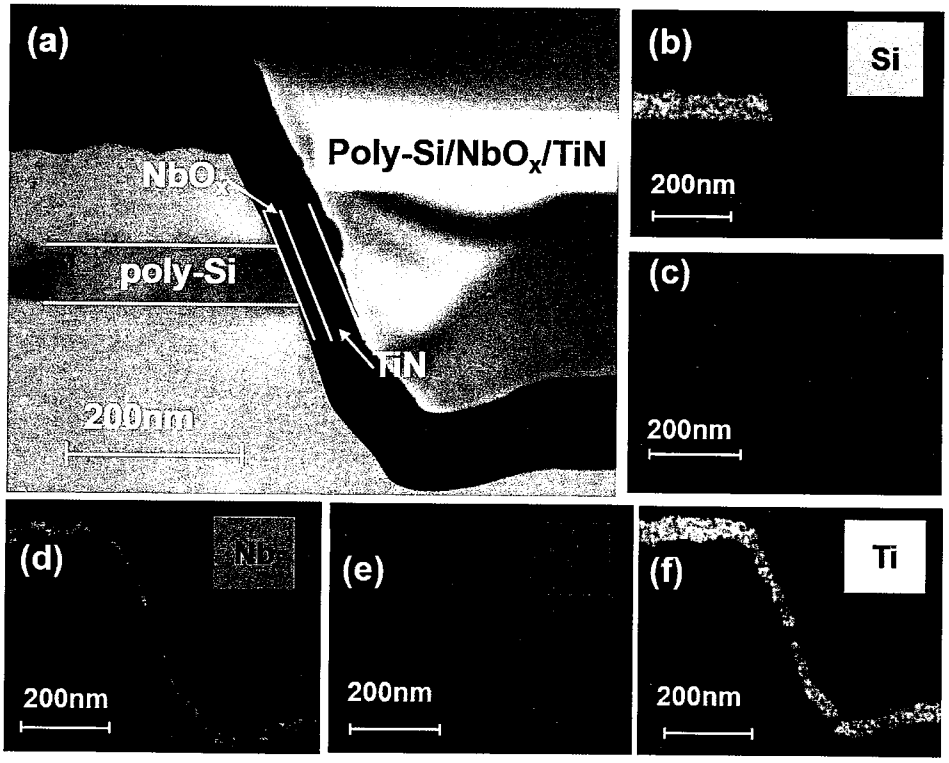


图 6.2 （a）NbO_x 器件的 TEM 截面图；（b-f）器件内各个元素的分布谱图

由于淀积的 NbO_x 为非晶结构，器件的初始电学状态表现为高阻态。为实现器件的正常阈值转变特性，需要对器件初始化，即执行 forming 操作。这里，我们使用 Agilent B1500A 对器件从 0 V 到 5 V 进行直流扫描，扫描过程中，器件的 TiN 上电极加电压 poly-Si 下电极接地。器件在 4.5 V 左右发生 forming 操作，导致器件从 56 GΩ（@1 V）的初始状态变为 35 MΩ（@1 V）相对低的电阻状态，如图 6.3（a）所示。

这是因为在 forming 过后，NbO_x 功能层内将生成 NbO_2 晶体通道（图 6.3（b））[10, 11, 19]，该晶体通道的高阻态决定了器件正常转变过程中的高阻态。当 TiN 电极上施加的正或负电压超过阈值转变电压（V_{TH}）时，NbO_2 通道转变到低电阻状态（low resistance state, LRS）。然后，当施加的电压低于保持电压值（V_H）时，通道状态自发恢复到高阻状态（high resistance state, HRS）。在转变过程中，观察到约 2 μA 的阈值电流，远低于先前报告的 NbO_2/TiN 结构的器件[20, 21]。这归因于多晶硅底电极的低导热性。正负电压扫描下，可以得到阈值转变电压和保持电压绝对值几乎相同的双向易失性转变行为，这与先前研究中报告的一致[7, 21, 22]。

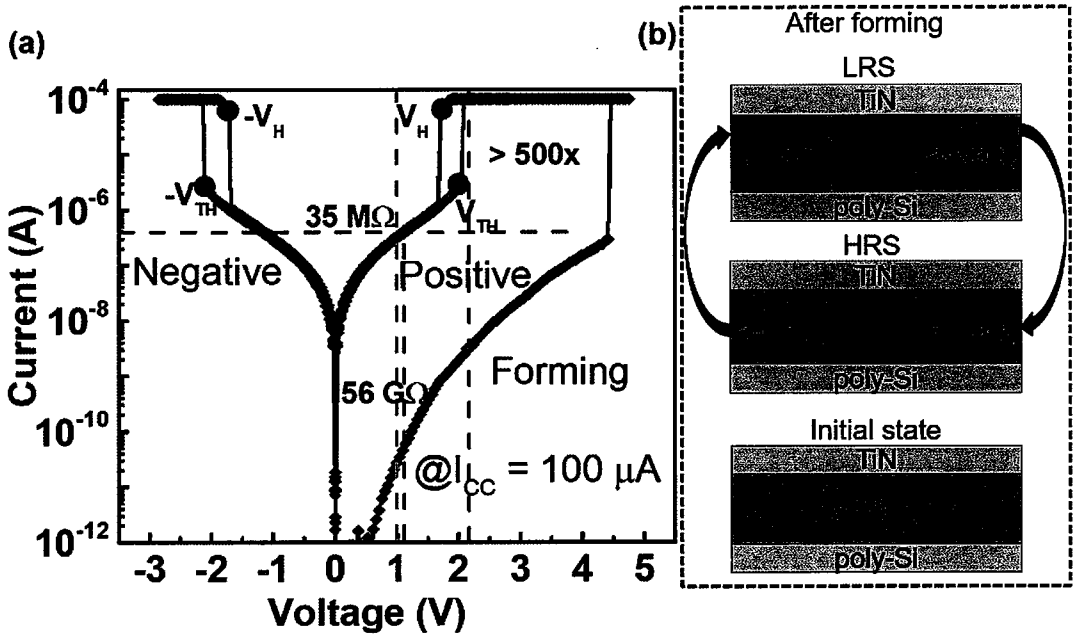


图 6.3 （a）NbO_x 器件的 forming 前后的电压扫描曲线；（b）器件 forming 前后的原理图

为分析导致这种转变行为的原因，我们又对 NbO_x 器件进行直流电流扫描操作，观察到 NDR-1 和 NDR-2 两个负微分电阻区域（negative differential resistance, NDR）[19]，如图 6.4 所示。惠普实验室的 Kumar 等人研究发现 NDR-1 是由热驱动非线性 Poole-Frenkel（PF）输运机制的不稳定性导致的，NDR-2 是由温度控制的 Mott 金属-绝缘体转变（temperature controlled Mott Metal-insulator-transition, TC Mott MIT）引起的[10, 13]。这两种 NDR 行为都可以归结为焦耳热导致的转变行为。电压模式下的滞回窗口包括 NDR-1 和 NDR-2 两个区域。因此，在这项工作中，我们认为焦耳热驱动的 NDR 机制是导致 NbO_x 器件电压扫描下的易失性阈值转变行为的原因。

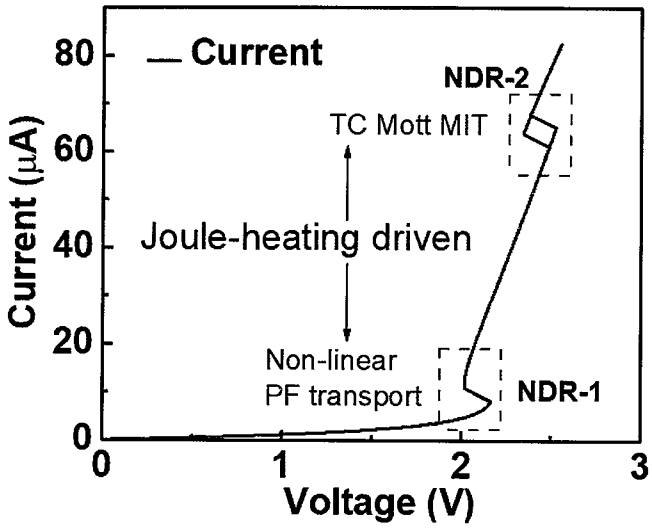


图 6.4 NbO_x 器件 forming 后的电流扫描曲线

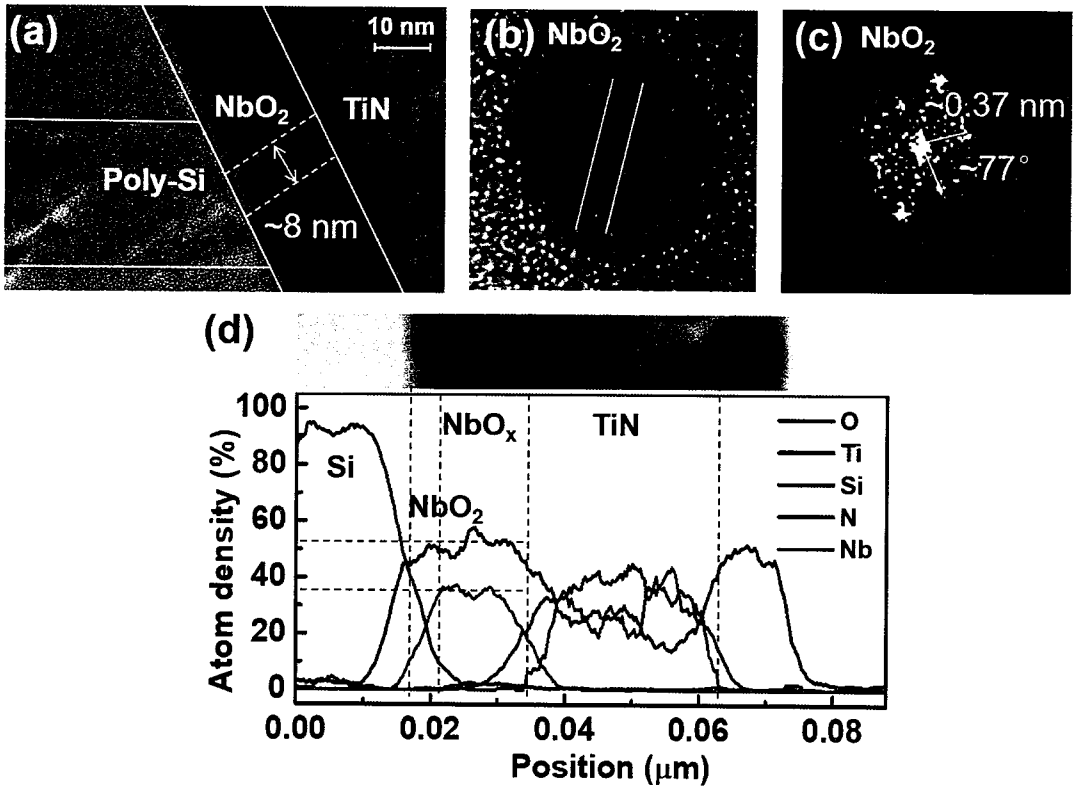


图 6.5 （a-c）器件 NbO_2 通道的放大图及表征分析；（d）通道线扫描的能量色散谱（EDS）

为进一步证明 forming 过后形成了 NbO_2 通道，我们对 TEM 截面的交叉器件区域进行了观察，观察到直径约 8 nm 的 NbO_2 晶体的圆形区域（图 6.5（a））。我们认为

圆形区域是 NbO_2 通道末梢的横截面，由于 TiN 顶部电极宽（约 20μm），在 TEM 样品切割过程中可能会遗漏完整的晶体 NbO_2 通道。图 6.5（b）为图 6.5（a）中 NbO_2 晶体区域的放大图。清晰的晶格条纹和相应的快速傅立叶变换（FFT）图像（图 6.5（c））表明纳米颗粒是高度结晶的 NbO_2。相对晶面之间的测量角度约为 77°，这与其他研究中的相关报道一致[10, 19]。图 6.5（c）给出了通道的线扫描能量色散谱（energy disperse spectroscopy, EDS），从沿着沟道（从 poly-Si 到 TiN）的元素分布我们可以看到 Nb:O 原子比约为 2。这些结果表明在 forming 过程中形成了 NbO_2 晶体通道。在 forming 操作后，器件的阈值转变行为发生在 NbO_2 晶体通道中。

6.2.2 NbO_x 器件 SPICE 模型

在本研究中，我们使用了惠普实验室 Pickett 等人提出的双相忆阻器模型[22]。在这个模型中，对 forming 后的 NbO_2 通道有四个假设：（1）通道形状为对称的圆柱体；（2）金属态和绝缘态之间的转变温度在通道内是固定的，（3）通道外部的环境温度为室温；（4）焦耳热的流动沿着二维的径向方向。器件总电阻由相位分数的函数描述（公式 1）：

$$R_{ch}(u)=\frac{\rho_{ins}L}{\pi r_{ch}^2}[1+(\frac{\rho_{ins}}{\rho_{met}}-1)u^2]^{-1} \quad \text{.....（6.1）}$$

其中 R_{ch} 为沟道电阻，ρ_{ins} 和 ρ_{met} 分别为金属相和绝缘相电阻率，r_{ch} 为通道半径，L 为沟道长度，$u=r_{met}/r_{ch}$ 为径向坐标表示的金属相分数。u 随时间 t 的动态状态演化关系如公式 6.2 所示：

$$\frac{du}{dt}=(\frac{d\Delta H}{du})^{-1}(R_{ch}(u)i^2-\Gamma_{th}(u)\Delta T) \quad \text{.....（6.2）}$$

在公式 2 中，ΔH 和 Γ_{th} 分别为系统焓和绝缘壳的导热系数，ΔT 为焦耳热引起的温度变化。系统焓和导热系数可分别用公式 6.3 和公式 6.4 来表示：

$$\Delta H=\pi L r_{ch}{}^2[c_p\Delta T\frac{u^2-1}{2\ln u}+\Delta h_{tr}u^2] \quad \text{.....（6.3）}$$

$$\Gamma_{th}(u)=2\pi L\gamma(\ln\frac{1}{u})^{-1} \quad \text{.....（6.4）}$$

因此，焓的变化如公式 6.5 所示:

$$\frac{d\Delta H}{du}=\pi L r_{ch}{}^{2}[c_{p}\Delta T\frac{1-u^{2}+2u^{2}\ln u}{2u(\ln u)^{2}}+2\Delta h_{tr}u] \quad (6.5)$$

根据上述五个方程，我们编写了 LTspice 模型并进行了仿真。图 6.6 给出了器件在两个三角形电压扫描周期下的模拟结果并完成了和实验结果的匹配，显示出双向易失性阈值转变行为。同时，该模型也作为我们后面传入神经电路的仿真验证。

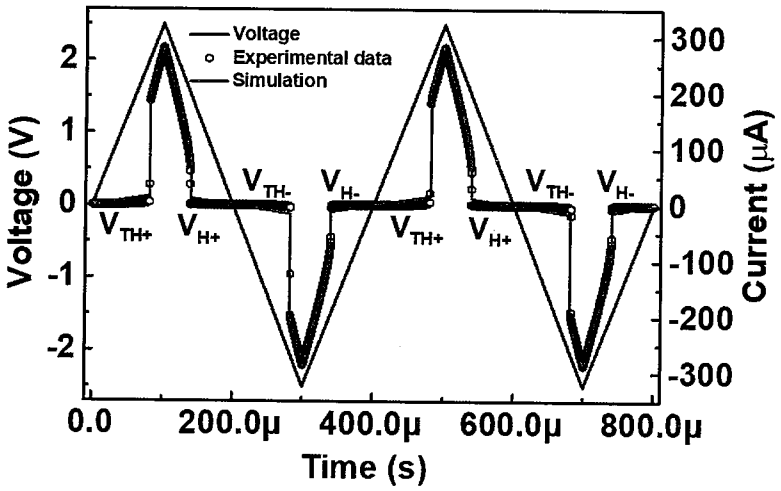


图 6.6 NbO_x 器件在三角波扫描下的仿真和实验数据

6.3 人工传入神经的工作原理

6.3.1 无外接电容器的传入神经电路

考虑到生物传入神经的行为和 NbO_x-Mott 忆阻器的特点，我们设计了一种紧凑的人工脉冲传入神经（图 6.7（a））。如前面所述，该传入神经由一个固定电阻（R_c）和一个具有固有寄生电容的 NbO_x 忆阻器构成。应注意的是，由于器件的纳米级尺寸，本征寄生电容小于 1 pF，与测试电路中数十 pF 的外部寄生电容相比，本征寄生电容可以忽略不计，图中的电容表示总寄生电容，大约为 20 pF。R_c 的一个节点作为输入节点，另一个节点与 NbO_x 忆阻器的上电极连接，NbO_x 忆阻器的下电极接地。这里使用的 R_c 为 75 kΩ，远小于 NbO_x 忆阻器的高阻态电阻 R_{HRS} 且远大于其低阻态电阻 R_{LRS}（$R_{HRS}>>R_c>>R_{LRS}$）。当在输入节点上施加电压时，因为 $R_{HRS}C_{parasitic} >> R_cC_{parasitic}$，寄生电容首先通过充电回路（$C_L$）充电[23]。一旦电容器上的电压超过 NbO_x 忆阻器的阈值电压 V_{TH}，由于焦耳热驱动的 NDR 机制（$V^2/R_{HRS}\times\Delta t$），忆阻器转变到其 LRS，由于 R_{LRS} 远小于 R_c 导致 $R_{LRS}C_{parasitic} << R_cC_{parasitic}$，电容器通过放电回路（DL）放

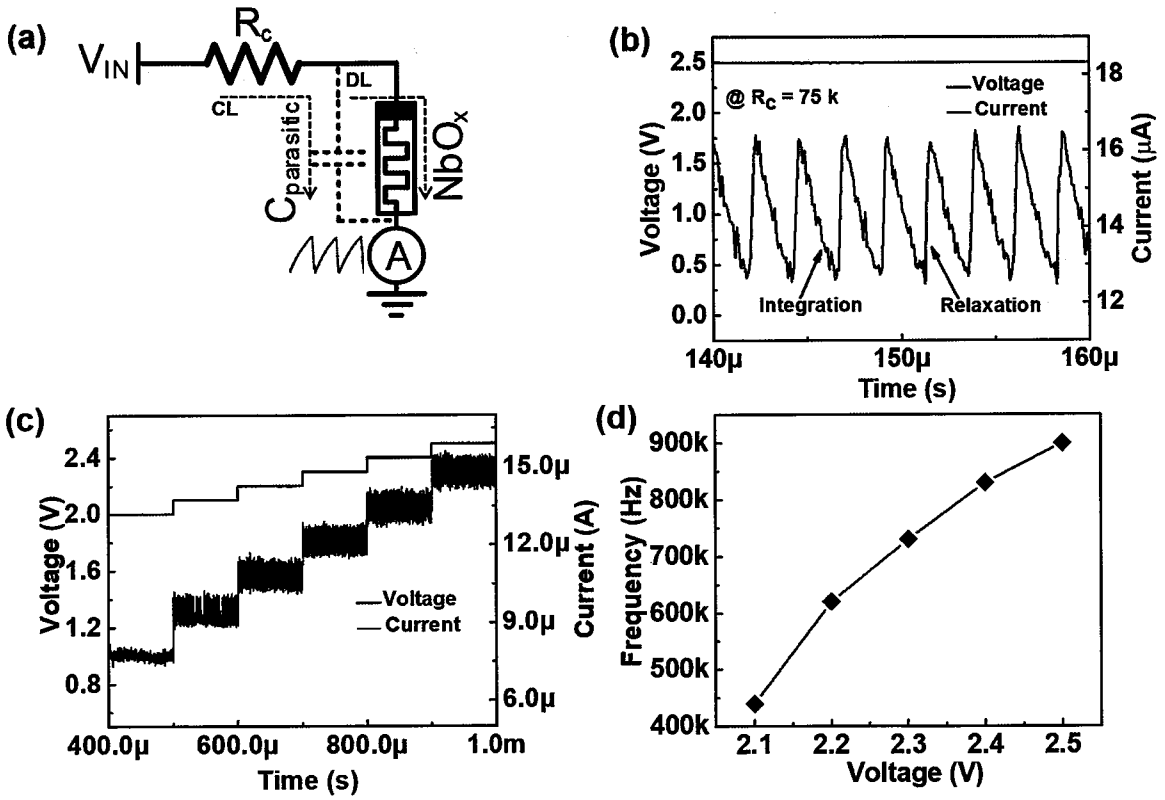


图 6.7 （a）传入神经电路原理图；（b）恒定输入下的振荡输出电流；（c）不同电压下的振荡输出电流；（d）输出频率-输入电压的准线性关系

电[23]。当器件打开后，放电过程占主导地位，电容器上存储的净电荷量减少，电容器上的电压降低。当电容器上的电压降到 V_H 以下时，器件上产生的焦耳热不足以再保持 NbO_2 的金属状态，器件返回到 HRS，电容器再次开始充电。在连续输入下，由于电容器的作用，忆阻器在 HRS 和 LRS 之间不断转换，观察到振荡行为，每一次振荡对应一次传入神经动作电位的发放，如图 6.7（b）所示。该振荡电流为测量到的流过电路的总电流。这里，为了便于理解，我们将 V_H 到 V_{TH} 的充电时间定义为积分时间，将 V_{TH} 到 V_H 的放电时间定义为弛豫时间。为了估计输入强度和输出频率之间的关系，我们在输入节点上施加了不同的电压（如图 6.7（c））。当输入电压超过 2 V 时，传入神经开始工作，并且振荡频率随输入电压的增大而增大。图 6.7（d）给出了振荡频率与输入电压的准线性关系，每个数据点是由如图 6.7（c）中每一电压下计算出的频率平均值，可以看出振荡频率随电压的增加而明显增加。

为了进一步证明该传入神经电路在持续变化电压刺激下的频率变化，我们采用具

有 1 V/ ms 斜率的三角波脉冲（从 1.9 V 到 2.5 V）作为输入刺激信号，如图 6.8（a）所示。第三个面板显示了相应的频率变化。振荡频率随输入刺激电压的升高而增大，随刺激电压的降低而减小，表现出和输入电压相同的变化趋势，这说明传入神经电路也可以在三角形脉冲下正常工作。图 6.8（b）是从图 6.8（a）中提取的频率-电压关系曲线，包括电压升高和电压下降两个过程，得到输出频率-输入电压的准线性关系。

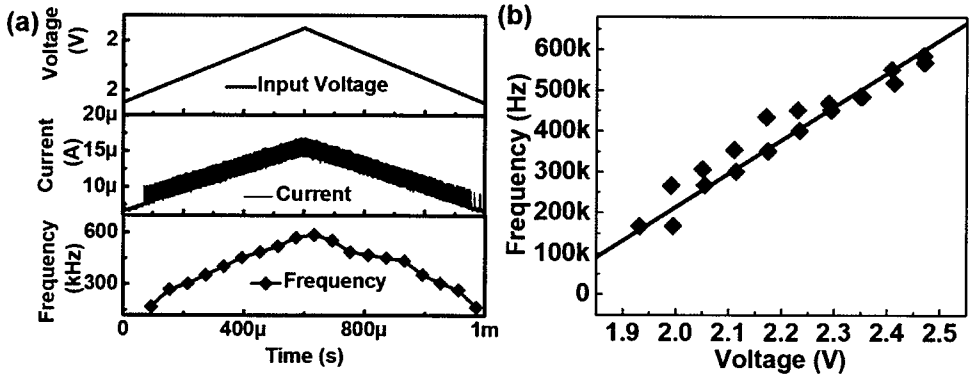


图 6.8 （a）传入神经在三角波脉冲输入下的振荡输出及频率变化；（b）输出频率-输入电压准线性关系

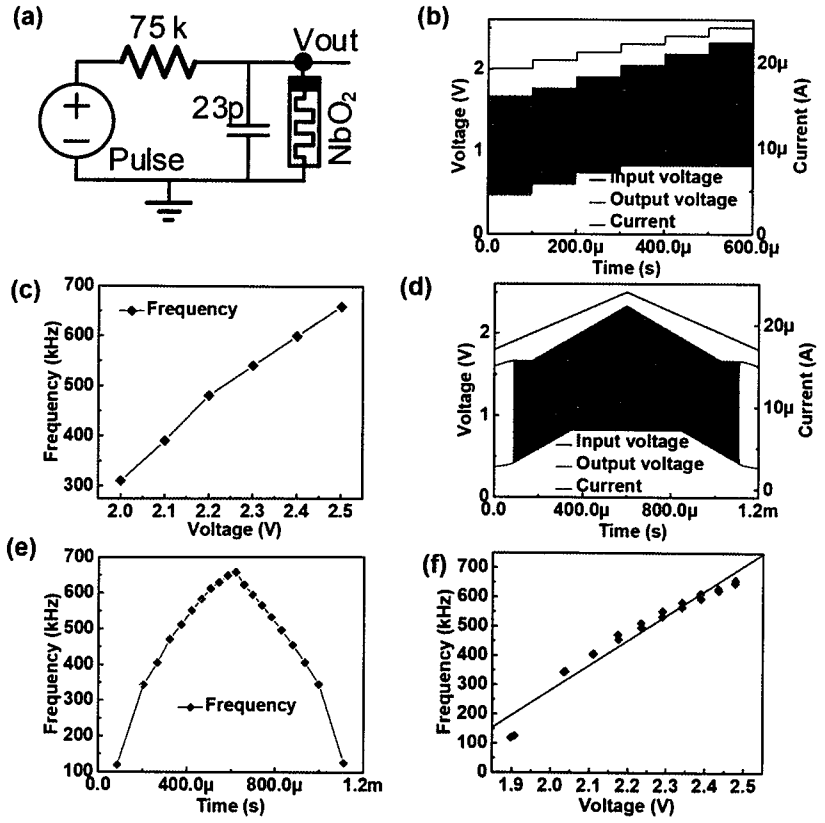


图 6.9 （a）传入神经的仿真电路；（b-f）仿真结果

接下来，我们利用 6.3.2 中所述的 SPICE 模型进行了仿真，验证该传入传经电路工作的合理性，仿真电路如图 6.9（a）所示。在仿真中，使用了 23 pF 的电容器作为电路中的总电容，分别观测了振荡过程中流过电路的总电流和电容器上的输出电压变化，图 6.9（b）-（f）给出了仿真结果。结果表明振荡电流在不同输入电压下的频率变化和图 6.8 中的实验数据表现出的变化趋势一致，证明了图 6.8 中实验操作的合理性。电容器上的输出电压在不同输入电压下的振荡具有稳定的振荡窗口，这是由器件模型的阈值电压 V_{TH} 和保持电压 V_H 决定的。

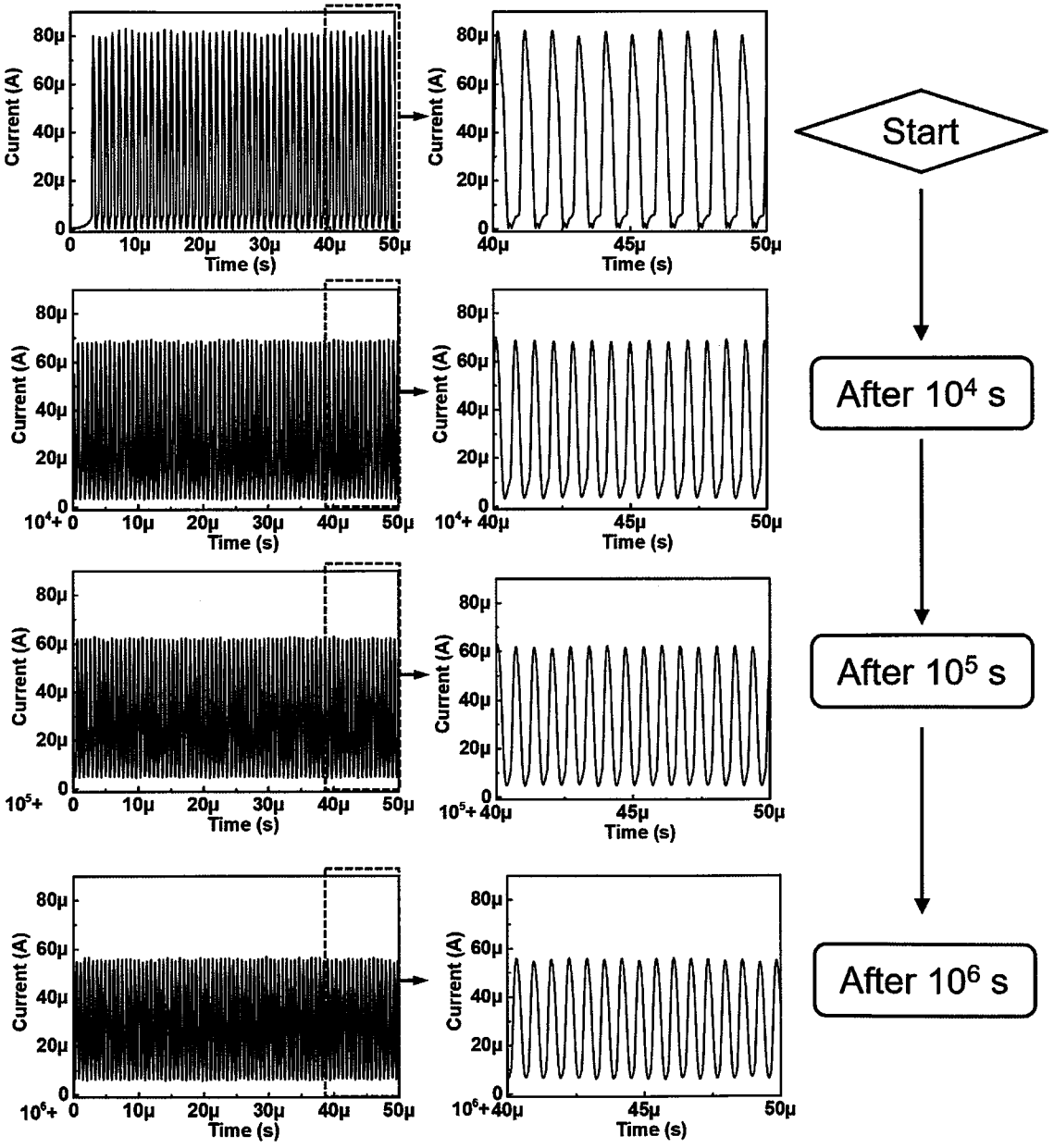


图 6.10 传入神经的耐久性测试

需要注意的是，忆阻器的耐久性对实际应用至关重要，传入神经的每一次振荡对应 NbO_x 器件的一个转变周期（从打开到恢复）。器件的转变周期决定了传入神经的寿命。因此，我们在传入神经电路上施加长时间的电压刺激来测试 NbO_x 器件的耐久性。图 6.10 给出了不同时间段内的输出结果，在测试过程中，我们测量的是流过 NbO_x 器件的振荡输出电流。振荡周期小于 1 μs，在连续运行 10^6 s 之后，传入神经仍旧可以保持振荡输出，产生了大于 10^{12} 的耐久值。然而，随着施加刺激的时间增加，可以看到器件的输出振荡峰值电流逐渐降低。这是由于长时间对器件施加激励使得 NbO_2 通道外周形成氧空位缺陷从而导致通道的直径变小，变小的通道直径在导通瞬间的电阻具有更大的阻值，造成了峰值电流变小。另外，我们可以观察到，随着刺激时间的增加，器件的漏电增加，说明器件的高阻态变低，这是 NbO_2 通道外周形成的氧空位缺陷导致的。除了峰值电流和漏电电流的变化，我们还可以看到传入神经的放电频率逐渐增加。这是因为随着放电次数的增加，器件的阈值电压和保持电压发生漂移，导致器件的滞回窗口变小，如图 6.11 所示。滞回窗口的变小造成了传入神经的充电时间和放电时间变短，从而导致振荡周期变短，频率增加。在实际工程应用，我们需要考虑到器件的这种退化特性加补偿电路，或者优化器件的制备工艺使得形成的晶体通道更加稳定[24]。甚至，更理想的情况下，由于生物体内神经元的发射具有自稳态，也就是说神经元的平均发放次数基本相同，那么不同神经元内器件的退化过程一致，就会相互补偿，仍旧可以实现工程的应用。

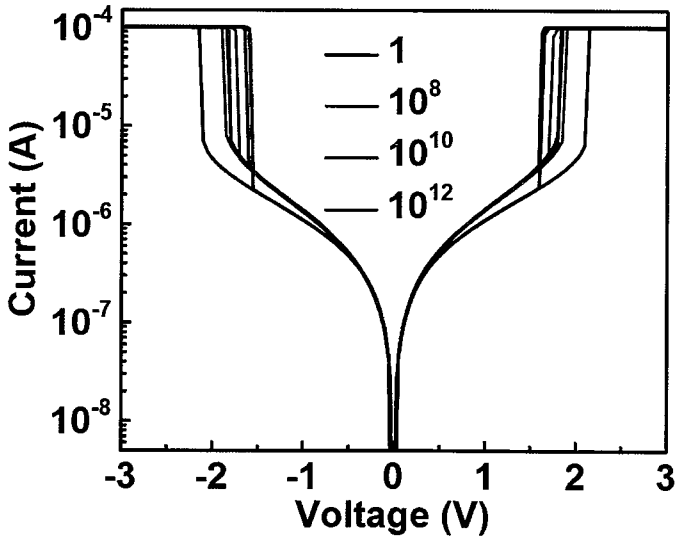


图 6.11 （a）NbO_x 器件在不同放电周期后直流转变曲线的变化

高能效被认为是生物系统的一个重要优点。最近的工作报道，人工突触在神经脉冲刺激下的能量消耗为~pJ 级甚至是~fJ [25-28]。为了验证该传入神经适用于对接高能效的脉冲神经形态机器，根据图 6.8（c）的实验数据，我们进一步计算了它的能耗（如图 6.12 所示）。瞬态功率由输入电压和输出电流的乘积计算得到，总能量消耗是功率对时间的积分。每个振荡峰的能量消耗是通过将总能量消耗除以一段时间内的振荡次数来确定的。如图 6.12（f）给出了不同振荡频率下的每次振荡事件的能量消耗。最低能量消耗低至 38 pJ/事件。我们认为，使用阈值电压更低、V_H-V_{TH} 滞回窗口更小的 NbO_x 器件和寄生电容更小的测试电路[9, 22]，可以进一步降低能量消耗。

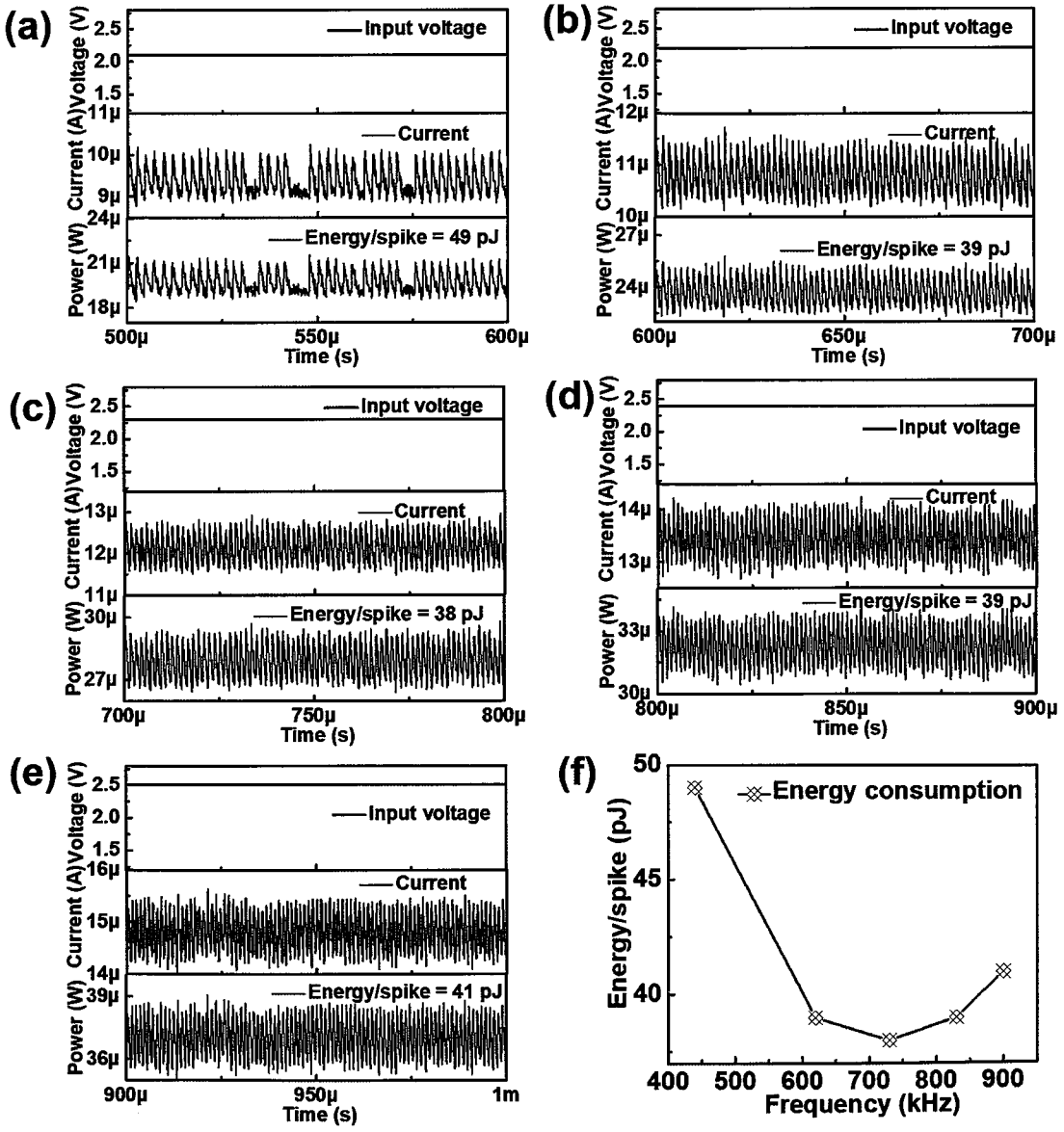


图 6.12 传入神经在不同频率下振荡一次的能量消耗

6.3.2 外接电容器的传入神经电路

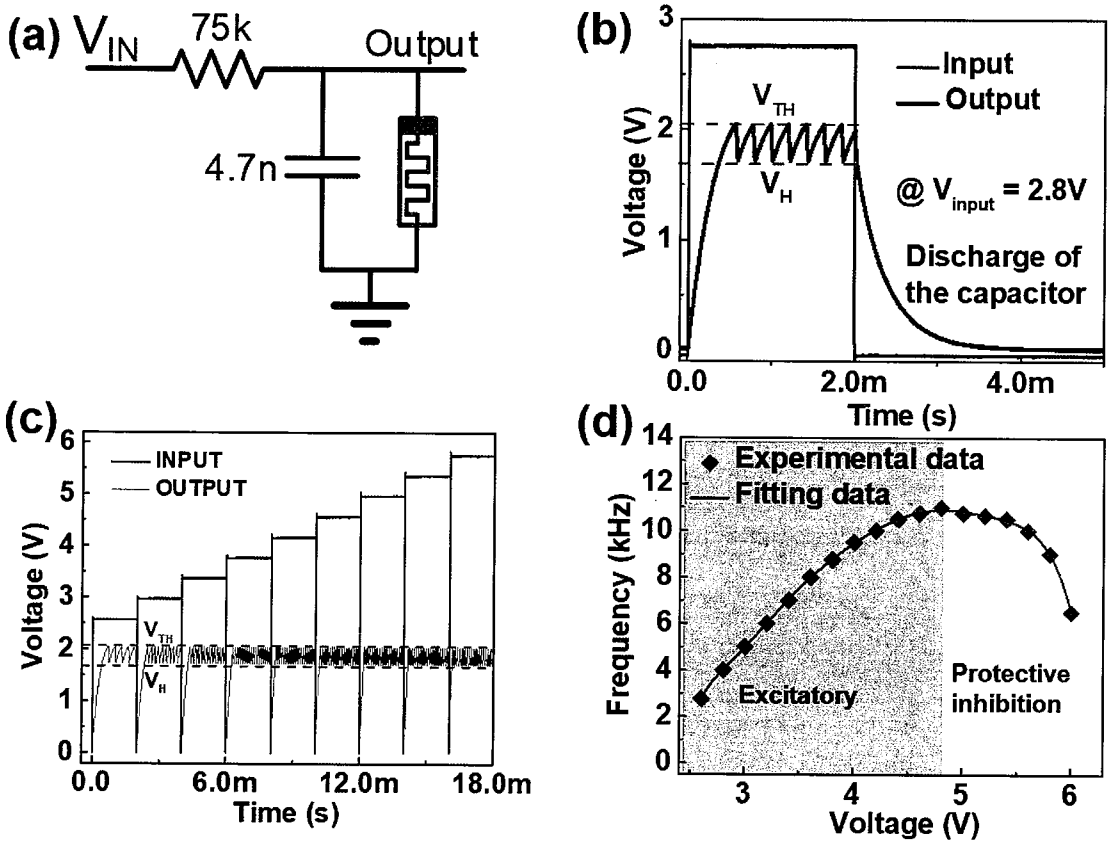


图 6.13 （a）外接电容器的传入神经电路；（b-c）电路的输出结果；（d）输出频率和输入电压的关系曲线

为了演示传入神经在连续输入的模拟传感信号下的振荡行为，我们用一个 4.7 nF 的外接电容器并联在 NbO_x 器件两端，如图 6.13（a）所示。这是因为，传入神经寄生电容大约为 20 pF，由此产生的 RC 时间常数太小，不易于操作以完成我们对人工机械感受器系统的验证。这里，电容器上的电压作为输出信号。在这里，我们用 Keysight 81160A 脉冲发生器作为信号源，Keysight InfiniiVision MSO-X 3104T 示波器采集输出信号。图 6.13（b）给出了带有并联电容器的传入神经的实验结果。在输入信号的开始，可以观察到电容器明显的积分过程，当电容器上的电压达到器件的阈值电压 V_{TH} 后，器件转变为低阻态，电容器开始放电。当电容器上放电的电压低于器件的保持电压 V_H 后，器件恢复到高阻态，电容器又开始充电。如此反复，形成振荡的输出。图 6.13（c）给出了不同输入刺激电压下传入神经的输出。由于不同输入电压下的输出振荡频率是用外围示波器分别单独进行测试的，所以在每个输入信号的开始，可以观察

到电容器明显的积分过程。然后依次施加递增的输入电压，输出电压始终保持在 NbO_x 器件的 V_{TH} 和 V_H 之间振荡。其输出频率-输入电压关系统计结果如图 6.13（d）所示。当输入电压从 2.4 V 增加到 4.8 V 时，振荡频率呈准线性增加，振荡频率在输入刺激为 4.8 V 时达到最大值。然后，随着输入电压的进一步增加，振荡频率开始降低，最终在 6.2 V 时停止振荡。上述这种频率响应与生物传入神经的行为一致。

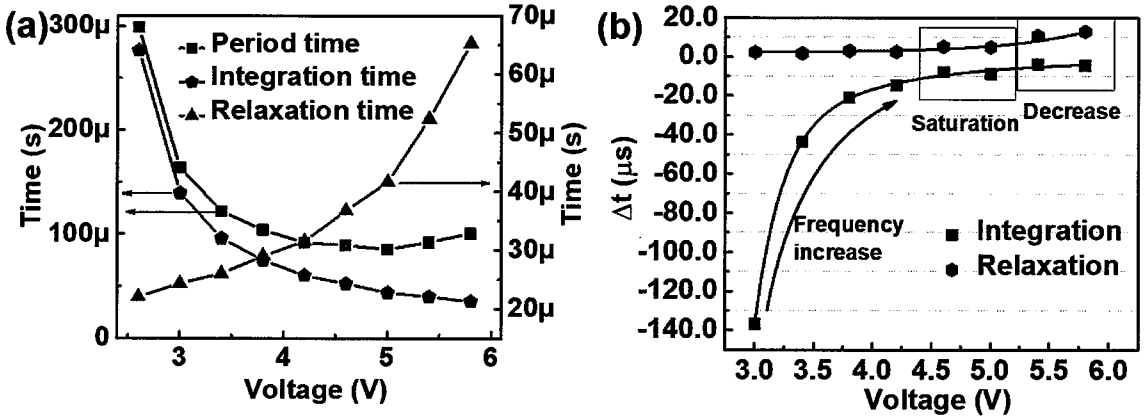


图 6.14 （a）不同输入电压下的积分时间和弛豫时间统计；（b）不同输入电压下的积分时间和弛豫时间的变化量的统计

到目前为止，已有文献系统地论证了频率随输入强度的增加而增加的行为[9, 12]，而频率随输入强度的进一步增加而减少的现象还没有具体报道。为系统解释该现象，我们对不同输入电压下的积分时间和弛豫时间进行了统计，如图 6.14（a）所示。可以看到积分时间随输入电压的增大而减小，弛豫时间随输入电压的增大而增大。振荡周期大小随着输入电压先减小后增大。在正常输入强度下（< 4.8 V），输出频率和输入强度的准线性关系是因为传入神经积分时间在整个周期内占主导地位，随着输入强度的增加积分时间逐渐减小，从而导致频率逐渐升高（振荡周期 = 积分时间 + 弛豫时间）。当输入电压达到 4.8 V 左右时，积分时间和弛豫时间相等，得到最大输出频率。当输入电压继续增大时，弛豫时间开始大于积分时间并占主导，振荡周期变长，频率降低。为更清楚的看出积分时间和弛豫时间的变化量，我们在图 6.14（b）将 6.14（a）中的每个数据点与先前的相邻数据点进行比较，给出了不同输入电压下积分时间和弛豫时间的变化量。结果表明积分时间的变化量逐渐减小，弛豫时间的变化量逐渐增大。积分时间和弛豫时间的变化过程分为三个阶段：（1）频率增加阶段，较低的

输入电压强度下，积分变化值急剧减小，弛豫时间的变化量基本不变。在这个阶段，积分时间的变化主导频率的变化；（2）频率饱和阶段，积分时间和弛豫时间的变化量基本不变，输出频率达到饱和；（3）频率降低阶段，积分时间的变化量基本不变，弛豫时间变化量开始增大，弛豫时间的变化主导频率的变化。

图 6.15 中给出了不同电压下积分时间和弛豫时间的倒数。两个值都与输入电压呈线性关系，可以更清楚的看到振荡周期在积分时间和弛豫时间相等时达到最小值。由图中可以看出弛豫时间的斜率绝对值大于积分时间的斜率绝对值，说明弛豫时间的相对变化率大于积分时间的相对变化率，导致了输出频率上先增后减的效果。

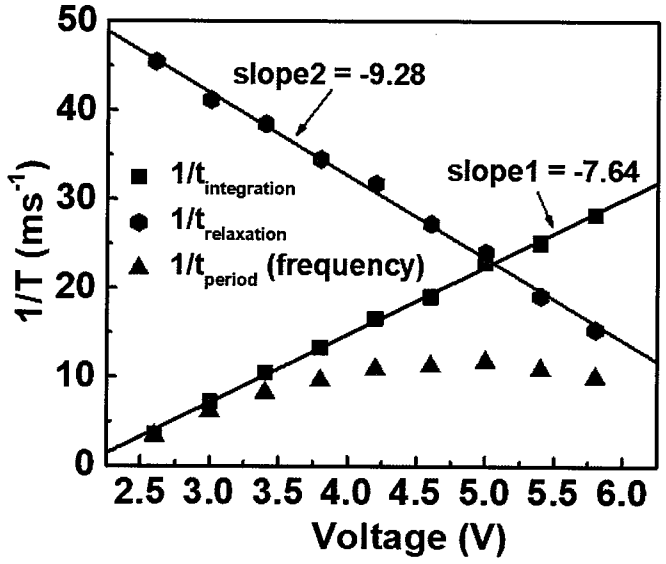


图 6.15 不同输入电压下的积分时间和弛豫时间的倒数统计

此外，为了进一步了解该传入神经电路停止振荡的原因，我们对传入神经振荡过程中器件上的保持能量功率进行了分析。在 6.2 V 输入电压的情况下，根据 Kirchoff 电压定律，器件低阻态 $R_{on} = V_{out}*R_c/（V_{in}-V_{out}）$=1.72 V * 75 KΩ/（6.2 V-1.72 V）= 28.8 KΩ。根据直流特性曲线，估计器件高阻态值约为 1 MΩ。根据估计出的高低阻态值，计算出在 5.6 V 输入电压下，传入神经有振荡输出，流过器件的能量保持功率约为 100.6 μW，如图 6.16（a）所示。这说明，器件上的能量功率大于该值时，器件就会保持在低阻导通状态。实际上器件的保持功率应该小于 100.6 μW，这是因为 NbO_x 器件具有非线性 I-V 特性。而在 6.2 V 输入电压下，当在第一次震荡过后，器件上的能量功率大约为 106 μW（图 6.16（b）），该功率足以使器件保持在其低阻导通状态，

从而使得传入神经电路不再有振荡输出。

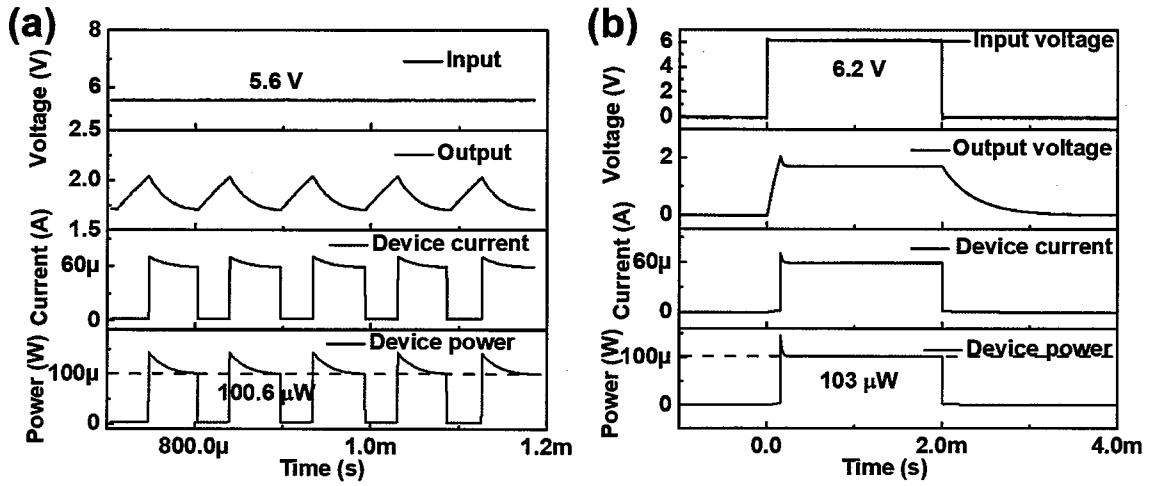


图 6.16 （a）NbO_x 器件的正常振荡过程中的保持功率；（b）振荡停止状态下器件上的能量消耗

接下来，我们分别对 5.6 V、6.0 V 和 6.2 V 输入电压下传入神经的振荡行为进行比较，如图 6.17 所示。在 6.2 V 输入电压下，传入的输出电压始终大于 NbO_x 器件的保持电压 V_H，导致器件始终处于“导通”状态，不能自发回到高阻态进行再一次的充电过程，这也说明了传入神经停止振荡的原因。

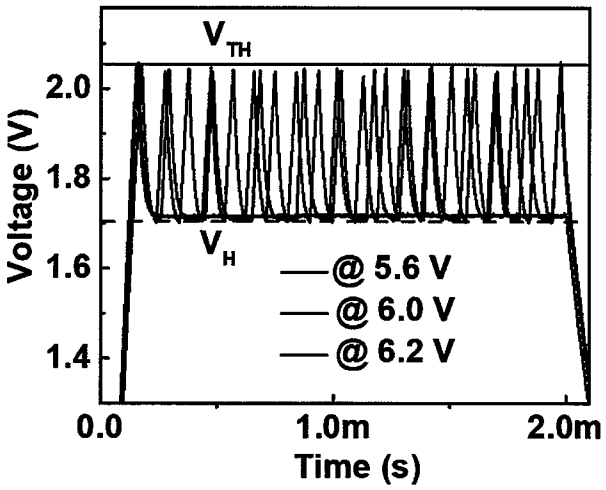


图 6.17 传入神经在不同输入电压下输出电压的比较

在自然界中，各种信息通常以模拟信号的形式进行传输，并且从传感器产生的信号也是连续多变的[29]。为了模拟这些信号，我们在输入节点上分别施加有直流偏置和

没有直流偏置的正弦信号，如图 6.18 所示。首先施加具有直流偏置的正弦信号（即使输入信号仅具有正电压）作为传入神经电路的输入（图 6.18（a））。测量输出信号并统计出响应的频率变化，电压输出显示在第二个面板中，输出频率显示在第三个面板中。我们可以看到频率随时间的变化也呈现出正弦曲线的形式。图 6.18（b）给出了无直流偏置（即同时具有正电压和负电压）正弦输入信号下的输出特性。结果显示正负电压下传入神经就有相同的振荡行为，这是因为该 NbO_x 器件具有对称的双向阈值转变行为。输出频率也表现出正弦曲线状，可以很好的反映输入电压强度的变化。图 6.18（c）给出了在大的刺激电压下的输出尖峰行为。观察到与生物神经元细胞相似的保护性抑制行为，当刺激电压正常时，传入神经可恢复到正常状态，说明了基于 NbO_x 器件的传入神经电路的稳定性。

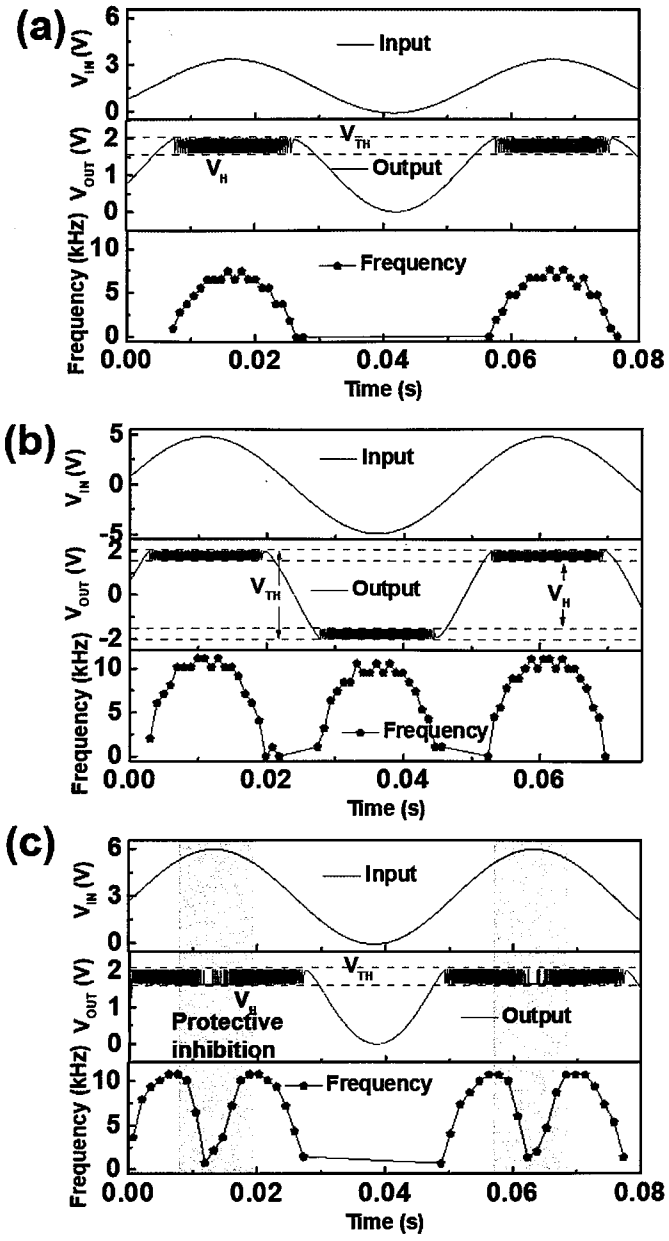


图 6.18 （a）传入神经在有偏置电压正弦波输入下的输出；（b）传入神经在无偏置电压正弦波输入下的输出；（c）传入神经在大振幅正弦波输入下的输出

这里，我们注意到，图 6.18（c）中的振荡停止电压约为 5.9 V，略小于图 6.13 中的振荡停止电压（6.2 V）。这种差异是由 NbO_x 器件的保持电压 V_H 的波动造成的。对图 6.18（c）进行测试时器件的保持电压 V_H（1.62 V）略低于测试图 6.13 中数据时器件的保持电压 V_H（1.71 V），因此观察到图 6.18（c）中较低的振荡停止电压。图 6.19 给出了器件在连续直流扫描下 V_{TH} 和 V_H 的变化演示，该图只是用来说明 V_H 的变化，

并不是图 6.18 对应的转变电压。

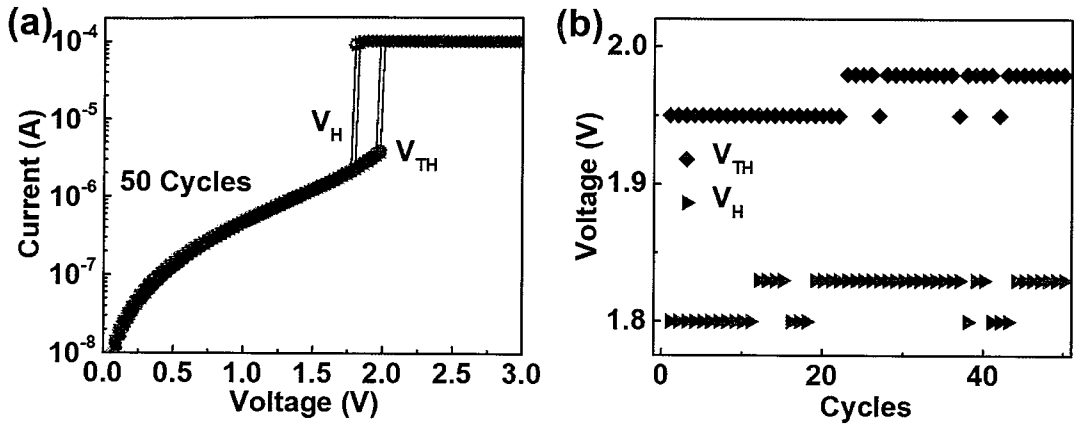


图 6.19 （a）NbO_x 器件在 50 个循环下的直流曲线；（b）V_{TH} 和 V_H 随着扫描次数的变化

此外，为了拓展该传入神经电路的应用范围，我们使用了一个 47 nF 的电容器作为外接电容来测试其在不同输入点下的振荡输出特性。图 6.20（a）给出了测试原理图，在这里，我们使用 Agilent B1500 输出脉冲并采集流过器件的电流，Keysight InfiniiVision MSO-X 3104T 示波器读取电容器上的输出电压。测试结果如图 6.20（b）和图 6.20（c）所示。随后，我们对不同电压下的输出频率进行了统计，结果如图 6.20（d）所示。输出频率随着电压的变化关系同样满足先增后减的趋势，实现了从 0 Hz 到 1100 Hz 的较低频率范围，该范围与人类神经系统相匹配（从 1 Hz 到 1000 Hz）。以上结果表明，我们的传入神经也可以在多变的模拟信号下工作，且根据实际应用场景调节外部电容器的值得到不同的频率输出范围，甚至适合于人机接口。

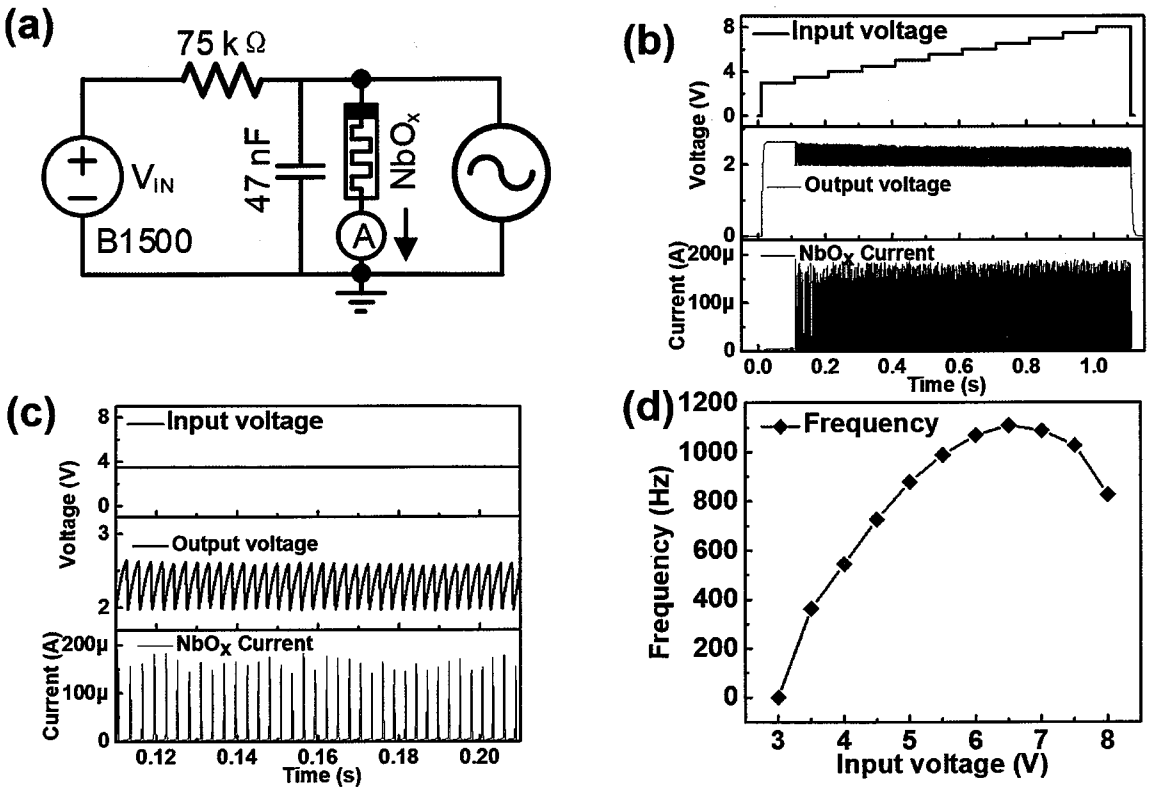


图 6.20 （a）传入神经外接 47 nF 电容器的测试电路原理图；（b-c）测试结果；（d）输出频率和输入电压的关系曲线

6.4 脉冲机械感受系统

在生物体中，机械感受系统是检测机械刺激（如压力、触摸、拉伸和振动）并通过传入神经将感知到的刺激信息发送到大脑进行进一步处理以产生对外部或内部环境做出适当反应的主要感觉结构。在无损害刺激下，机械感受器可以将刺激强度编码为成比例的动作电位发放频率[5]，但由于生物神经元的保护性抑制机制[17]，在有害的刺激下倾向于停止发放动作电位。为了验证该功能，我们将压电传感器与传入神经电路相连，我们演示了一种人工脉冲机械感受器系统（artificial spiking mechanoreceptor system, ASMS）。图 6.21 给出了该系统的电路原理图。当压力作用在压电器件上时，上电极产生正电压，当压力升高时产生负电压[30-32]，这是传感器内部原子结构变形的结果[33]。需要注意的是，我们的人工机械感受系统的输入电压信号是由压电器件产生的，所以系统不需要任何外部电源。压电器件在施加压力时由于内部原子结构变形在其上下极板之间产生电压差，随后由于电荷的泄漏导致电压降低。当压力被瞬间撤去

后，产生了相反的电压，然后通过电荷的泄漏最后达到静息状态，如图 6.22 所示。

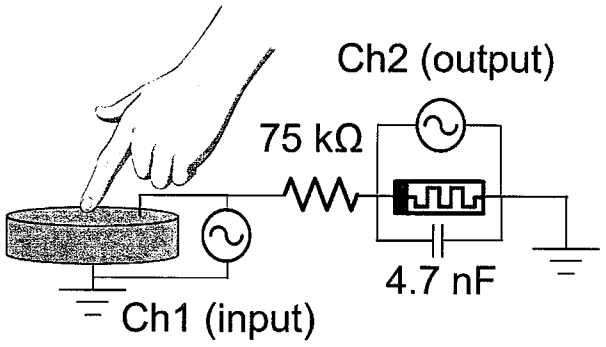


图 6.21 人工脉冲机械感受系统原理图

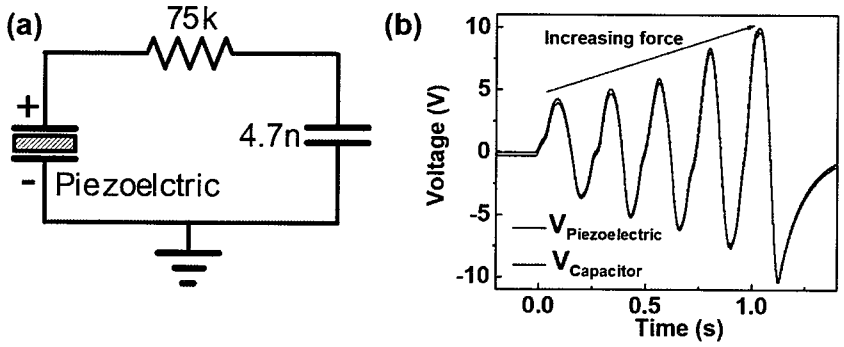


图 6.22 （a-b）压电器件的测试原理图和连续刺激下的电压输出特性

图 6.23（a）给出了人工机械感受器系统的测试结果。在压电传感器上施加一个力，产生一个类似正弦波的电压信号，然后由传入神经将该信号转换成振荡的频率信号。可以看出，输出结果与外接模拟电源电压作用于传入神经时表现出相同的行为。当施加的压力过大时，产生高峰值电压，使得传入神经停止振荡，这对应于传入神经的保护性抑制行为。从图中我们可以清楚地观察到在压电传感器动态输入下的频率响应，其响应趋势与产生的电压一致。图 6.23（b）到图 6.23（e）是对图 6.23（a）的放大。

为了进一步说明我们的人工机械感受系统在不同压力下的频率响应，我们在压电器件上施加不同强度的力，如图 6.24 所示。当压电器件上的压力很小时，产生的电压不足以驱动器件，则无法获得动态振荡输出。当施加的压力足够大时，压电器件产生的电压就可以驱动电容器使得器件打开，然后产生振荡输出。随着施加压力的增加，压电器件产生的峰值电压增加，从而传入神经的峰值频率增加。实验结果表明，结合

压电传感器，我们成功实现了一个零静态功耗的人工机械感受系统，其传入神经可用于将压电传感器产生的模拟信号转换为动态振荡频率输出。这些结果表明，我们的传入神经可以用于构建具有自我感知意识的神经形态机器，从而具有很大的应用潜力。

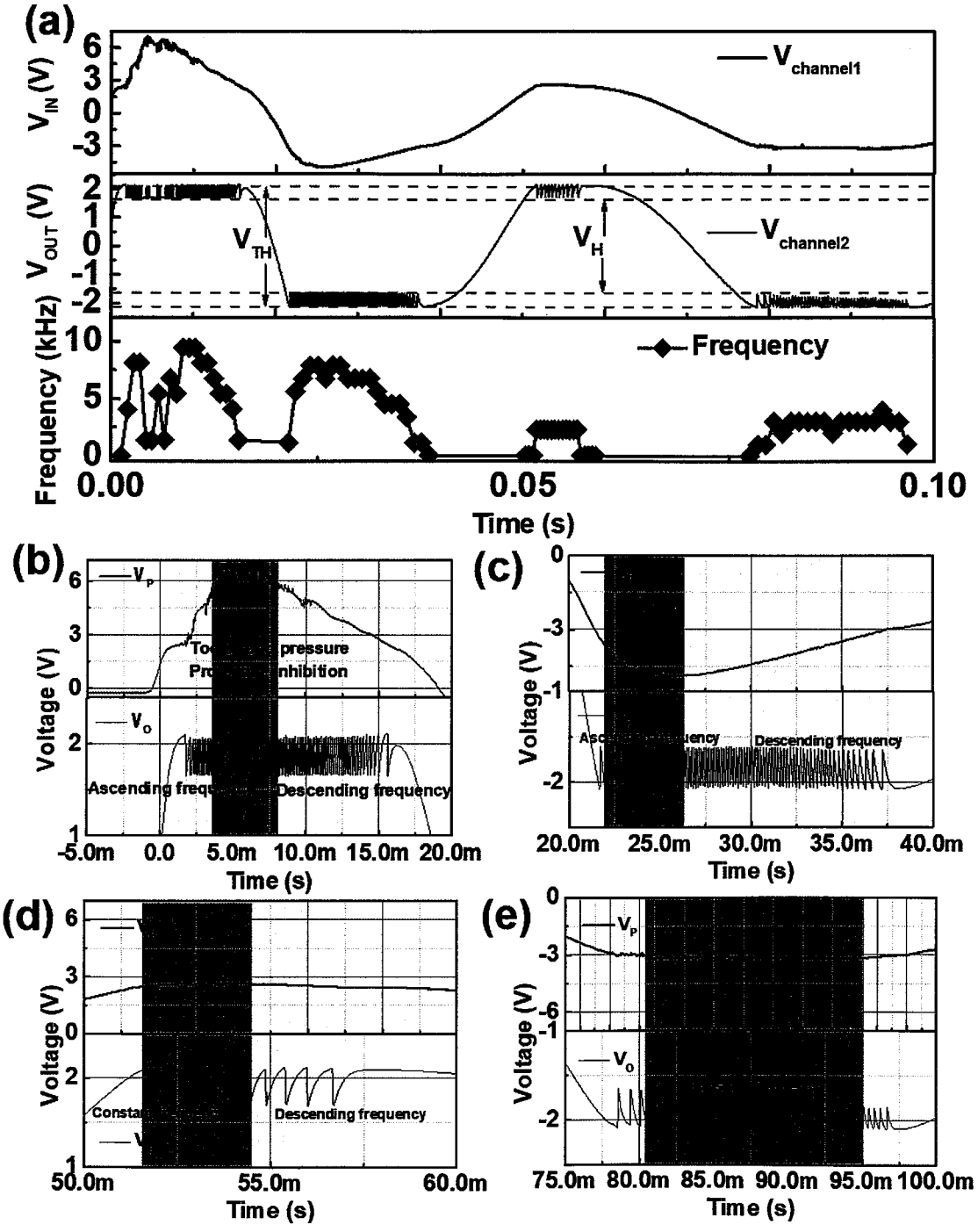


图 6.23 （a）人工机械感受系统的输出结果；（b-e）输出结果在各时间段的放大图

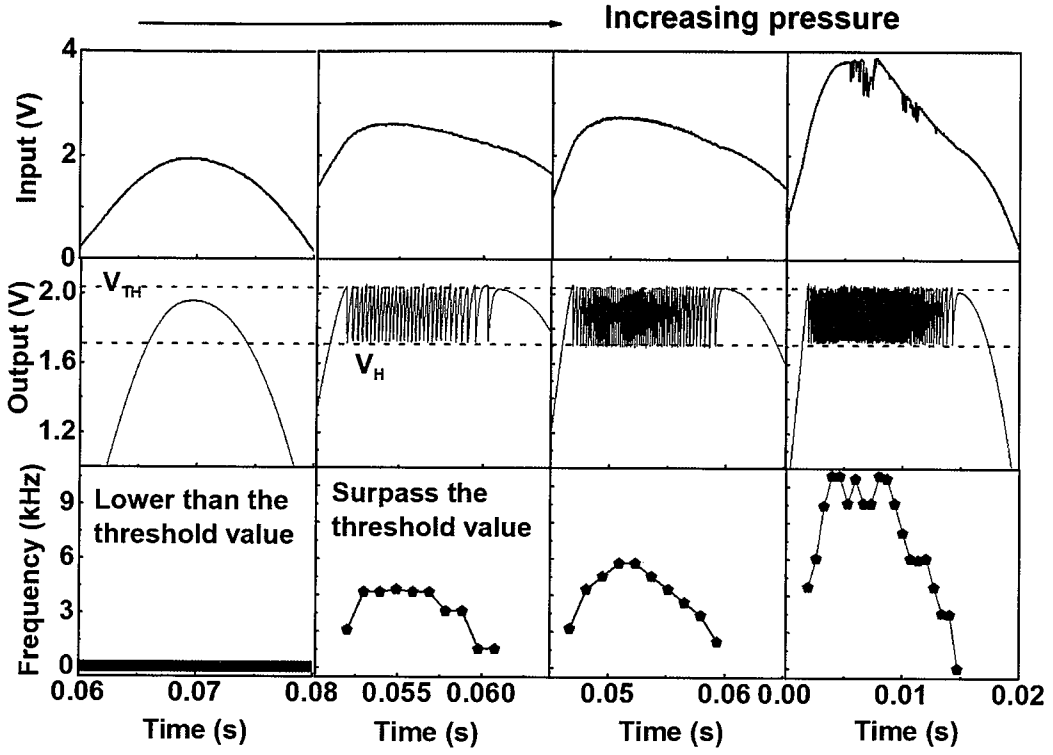


图 6.24 人工机械感受系统在不同压力强度下的输出结果

6.5 本章小结

传入神经与传感器的结合对于神经形态机器与环境的交互是至关重要的，它可以将环境中的模拟信号转换成脉冲信号，然后由神经形态机器内部系统进一步处理以对外界环境做出响应。基于 CMOS 电路的环形振荡器被报道可以用来实现传入神经的功能，它可以将模拟信息转换成振荡输出，其中振荡频率由振荡器的延迟控制。考虑到晶体管的物理局限性和缺乏理想的动态特性，忆阻器由于其高集成密度、低功耗和固有的动态特性等而成为一种很有前途的候选器件，因此，可以用来构建高效的传入神经电路。近年来，基于忆阻器对大脑皮层神经元的模拟受到了广泛的关注[14, 15, 34-39]，并进行了系统验证。本工作所提出的传入神经是对忆阻器基神经元的功能扩展，可以用来进一步构建神经形态机器与环境之间交互的接口，从而有望在不久的将来构建具有自我意识的新一代智能神经形态机器。

综上所述，在本工作中，我们首次提出并实验验证了一种 NbO_x-Mott 忆阻器的传入神经电路。该传入神经电路可以将模拟信号转换成动态的振荡频率。在一定的输入信号强度范围内，频率与输入电压呈准线性关系，当输入信号的强度过大时，传入神

经的输出频率降低并最终停止产生振荡信号，这一行为与生物神经元非常相似。传入神经在方波、三角形和正弦波等不同输入信号下的动态振荡行为得到了系统地研究。此外，我们进一步将传入神经与压电传感器相连，构建了一个不需任何外部电源的人工机械感受系统。该感受系统能够对压力信号做出响应，并将压力强度转换为相应的振荡频率。传入神经还可以很容易地扩展到处理来自其他传感器的感知信号，例如味觉、视觉、听觉、温度、磁场和湿度等。我们的传入神经单元除了可以构建多种形式的感觉系统外，还具有神经元的漏电积分发射特性，适合应用于脉冲神经元的构建[12, 40]，或具有输入强度依赖的耦合振子神经网络[41-43]。因此，该传入神经电路可以进一步用于构造复杂的神经网络来处理中枢神经信息，从而构建高效的神经形态机器。

本章的研究成果发表在 2020 年 1 月份的 Nature Communications 期刊上，并申请了一个国际专利。

参考文献

[1] Shulaker MM, Hills G, Park RS, et al., Three-dimensional integration of nanotechnologies for computing and data storage on a single chip [J]. Nature, vol. 547, pp. 74-78, Jul 05 2017.

[2] Tan H, Tao Q, Pande I, et al., Tactile sensory coding and learning with bio-inspired optoelectronic spiking afferent nerves [J]. Nat Commun, vol. 11, p. 1369, Mar 13 2020.

[3] Jung YH, Park B, Kim JU, et al., Bioinspired Electronics for Artificial Sensory Systems [J]. Adv Mater, vol. 31, p. e1803637, Aug 2019.

[4] Wall PD and Gutnick M, Properties of afferent nerve impulses originating from a neuroma [J]. Nature, vol. 248, pp. 740-743, 1974.

[5] Tee BCK, Chortos A, Berndt A, et al., A skin-inspired organic digital mechanoreceptor [J]. Science, vol. 350, pp. 313-+, Oct 2015.

[6] Kim Y, Chortos A, Xu WT, et al., A bioinspired flexible organic artificial afferent nerve [J]. Science, vol. 360, pp. 998-+, Jun 2018.

[7] Cha E, Park J, Woo J, et al., Comprehensive scaling study of NbO_2 insulator-metal-transition selector for cross point array application [J]. Appl Phys Lett, vol. 108, p. 3, Apr 2016.

[8] Gibson GA, Musunuru S, Zhang JM, et al., An accurate locally active memristor model for S-type negative differential resistance in NbO_x [J]. Appl Phys Lett, vol. 108, p. 5, Jan 2016.

[9] Liu XJ, Li S, Nandi SK, et al., Threshold switching and electrical self-oscillation in niobium oxide films [J]. Journal of Applied Physics, vol. 120, p. 10, Sep 2016.

[10] Kumar S, Wang Z, Davila N, et al., Physical origins of current and temperature controlled negative differential resistances in NbO_2 [J]. Nat Commun, vol. 8, p. 658, Sep 22 2017.

[11] Del Valle J, Salev P, Tesler F, et al., Subthreshold firing in Mott nanodevices [J]. Nature, vol. 569, pp. 388-392, May 2019.

[12] Gao L, Chen P-Y, and Yu S, NbO_x based oscillation neuron for neuromorphic computing [J]. Appl Phys Lett, vol. 111, p. 103503, 2017.

[13] Kumar S, Strachan JP, and Williams RS, Chaotic dynamics in nanoscale NbO_2 Mott memristors for analogue computing [J]. Nature, vol. 548, pp. 318-321, Aug 17 2017.

[14] Pickett MD, Medeiros-Ribeiro G, and Williams RS, A scalable neuristor built with Mott memristors [J]. Nature Materials, vol. 12, pp. 114-117, 2013.

[15] Yi W, Tsang KK, Lam SK, et al., Biological plausibility and stochasticity in scalable VO_2 active memristor neurons [J]. Nat Commun, vol. 9, p. 4661, Nov 7 2018.

[16] Sivaramakrishnan S, Sterbing-D'Angelo SJ, Filipovic B, et al., GABA(A) synapses shape neuronal responses to sound intensity in the inferior colliculus [J]. J. Neurosci., vol. 24, pp. 5031-5043, May 2004.

[17] Stetler RA, Gao Y, Signore AP, et al., HSP27: Mechanisms of Cellular Protection Against Neuronal Injury [J]. Curr. Mol. Med., vol. 9, pp. 863-872, Sep 2009.

[18] Purves D, Augustine GJ, Fitzpatrick D, et al., Neuroscience, 3rd ed. [M]. Inc. Massachusetts, USA: Sinauer Associates, 2012.

[19] Park J, Cha E, Karpov I, et al., Dynamics of electroforming and electrically driven insulator-metal transition in NbOx selector [J]. Appl Phys Lett, vol. 108, p. 5, Jun 2016.

[20] Luo Q, Zhang X, Gong T, et al., Memory Switching and Threshold Switching in a 3D Nanoscaled NbO_X System [J]. Ieee Electr Device L, vol. 40, pp. 718 - 721, May 2019.

[21] Cha E, Woo J, Lee D, et al., Nanoscale (~10nm) 3D vertical ReRAM and NbO_2 threshold selector with TiN electrode, in 2013 Ieee International Electron Devices Meeting, ed New York: Ieee, 2013.

[22] Pickett MD and Williams RS, Sub-100 fJ and sub-nanosecond thermally driven threshold switching in niobium oxide crosspoint nanodevices [J]. Nanotechnology, vol. 23, p. 215202, Jun 1 2012.

[23] Midya R, Wang Z, Asapu S, et al., Artificial Neural Network (ANN) to Spiking Neural Network (SNN) Converters Based on Diffusive Memristors [J]. Advanced Electronic Materials, p. 1900060, 2019.

[24] Rao F, Ding K, Zhou Y, et al., Reducing the stochasticity of crystal nucleation to enable subnanosecond memory writing.full. [J]. Science, vol. 358, pp. 1423-1427, 2017.

[25] Yu H, Gong J, Wei H, et al., Mixed-halide perovskite for ultrasensitive two-terminal artificial synaptic devices [J]. Materials Chemistry Frontiers, vol. 3, pp. 941-947, 2019.

[26] Chen Y, Yu H, Gong J, et al., Artificial synapses based on nanomaterials [J]. Nanotechnology, vol. 30, p. 012001, Jan 4 2019.

[27] Wang I-T, Lin Y-C, Wang Y-F, et al., 3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation, in 2014 Ieee International Electron Devices Meeting, ed New York: Ieee, 2014.

[28] Yu S, Gao B, Fang Z, et al., A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation [J]. Adv Mater, vol. 25, pp. 1774-9, Mar 25 2013.

[29] Struzik M, Garbayo I, Pfenninger R, et al., A Simple and Fast Electrochemical CO_2 Sensor Based on $Li_7La_3Zr_2O_{12}$ for Environmental Monitoring [J]. Adv Mater, vol. 30, p. e1804098, Nov 2018.

[30] Zhang G, Liao Q, Zhang Z, et al., Novel Piezoelectric Paper-Based Flexible Nanogenerators Composed of $BaTiO_3$ Nanoparticles and Bacterial Cellulose [J]. Advanced science, vol. 3, p. 1500257, Feb 2016.

[31] Vivekananthan V, Alluri NR, Purusothaman Y, et al., A flexible, planar energy harvesting device for scavenging road side waste mechanical energy via the synergistic piezoelectric response of $K_{0.5}Na_{0.5}NbO_3$-$BaTiO_3$/PVDF composite films [J]. Nanoscale, vol. 9, pp. 15122-15130, Oct 2017.

[32] Wu C, Kim TW, Park JH, et al., Self-Powered Tactile Sensor with Learning and Memory [J]. ACS Nano, vol. 14, pp. 1390-1398, Feb 25 2020.

[33] Wang L, Liu S, Gao G, et al., Ultrathin Piezotronic Transistors with 2 nm Channel Lengths [J]. ACS Nano, vol. 12, pp. 4903-4908, May 22 2018.

[34] Lin J, Annadi A, Sonde S, et al., Low-voltage artificial neuron using feedback engineered insulator-to-metal-transition devices, in 2016 Ieee International Electron Devices Meeting (IEDM), ed New York: Ieee, 2016.

[35] Stoliar P, Tranchant J, Corraze B, et al., A Leaky-Integrate-and-Fire Neuron Analog Realized with a Mott Insulator [J]. Advanced Functional Materials, p. 1604740, 2017.

[36] Lee D, Kwak M, Moon K, et al., Various Threshold Switching Devices for Integrate and Fire Neuron Applications [J]. Advanced Electronic Materials, p. 1800866, 2019.

[37] Jerry M, Parihar A, Grisafe B, et al., Ultra-Low Power Probabilistic IMT Neurons for Stochastic Sampling Machines[M]. New York: Ieee, 2017.

[38] Tuma T, Pantazi A, Le Gallo M, et al., Stochastic phase-change neurons [J]. Nat Nanotechnol, May 16 2016.

[39] Wang Z, Rao M, Han JW, et al., Capacitive neural network with neuro-transistors [J]. Nat Commun, vol. 9, p. 3208, Aug 10 2018.

[40] Chen P-Y, Seo J-S, Cao Y, et al., Compact oscillation neuron exploiting metal-insulator-transition for neuromorphic computing [J]. pp. 1-6, 2016.

[41] Romera M, Talatchian P, Tsunegi S, et al., Vowel recognition with four coupled spin-torque nano-oscillators [J]. Nature, vol. 563, pp. 230-234, Nov 2018.

[42] Sharma AA, Bain JA, and Weldon JA, Phase Coupling and Control of Oxide-Based Oscillators for Neuromorphic Computing [J]. IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, vol. 1, pp. 58-66, 2015.

[43] Shukla N, Parihar A, Cotter M, et al., Pairwise Coupled Hybrid Vanadium Dioxide-MOSFET (HVFET) Oscillators for Non-Boolean Associative Computing, in 2014 Ieee International Electron Devices Meeting, ed New York: Ieee, 2014.

第7章 总结与展望

7.1 论文工作的总结

神经形态芯片具有低功耗，低延时，高鲁棒性，事件驱动，时空信息高速处理等特点，在当前人工智能时代智能城市的构建，自动驾驶，声音识别等领域都有着广阔的应用前景。脉冲神经元和可塑性的神经突触是构建神经形态芯片的基础单元。在传统CMOS工艺下，由于晶体管并不是面向神经形态计算而提出的，因而缺乏与神经元和突触相似的动力学特性，在实现神经元和突触的功能时通常需要复杂的电路设计，这大大限制了在单个芯片上可集成的神经元和突触的数量。另外，当前晶体管尺寸已经接近物理极限，难以进一步缩放并且很难进行三维集成。忆阻器具有结构简单、功耗低、可微缩性好、易于三维集成、内在动力学机制丰富等优点，可以忠实地模拟神经元和突触地相关功能，被认为是实现低功耗、高密度类脑神经网络的理想硬件单元，在国际上引起了广泛地关注。但由于基于忆阻器的神经形态芯片地研究刚处于起步阶段，还存在一些关键问题需要解决：如，忆阻器突触器件可控的低电流缓变特性优化，忆阻器基神经元电路功能的稳定和丰富，以及一定规模系统的构建与验证。

本文围绕神经形态计算领域中存在的上述问题，在忆阻器的制备、神经突触功能的实现、神经元电路的优化设计以及系统验证等方面展开了以下工作：

（1）利用忆阻器实现神经突触的相关功能。首先以 Cu/a-Si/Pt 结构的离子基忆阻器为基础，生动模拟了生物突触的短时程和长时程可塑性行为。其次，为实现突触器件的多值可控及线性电导调制，我们进一步通过双层堆叠工艺，实现了一种具有低工作电压控、缓变特性可控的 $Pd/HfO_2/WO_x/W$ 叠层突触器件，并分别探讨了器件在不同脉冲编程方案下电导调制的非线性对系统性能的影响。为突触器件的设计和系统功能实现提供了理论基础。

（2）围绕如何利用忆阻器内在的离子动力学机制实现神经元电路展开了相应的理论探讨和实验验证。提出了利用离子基忆阻器实现神经元电路的两种方法：首先，利用 TS 忆阻器作为阈值开关实现了一种新型漏电积分-发射神经元。验证了脉冲神经元的四个关键功能：动作电位的全或无、阈值驱动放电、不应期和输入强度调制的频率响

应。并通过系统仿真验证了该神经元在数字识别中的可行性。另外，为进一步提高神经元的集成度，丰富神经元电路的功能，我们对 TS 器件进行了优化并提出了忆阻器-CMOS 混合设计的神经元电路，可以实现对忆阻器突触的原位操作。并结合专门设计的侧向抑制阵列在国际上首次实验验证了一个全硬件多层脉冲神经网络计算系统，实现了非监督学习和监督学习两种学习算法，演示了带噪声的数字识别功能。这个工作进一步推动了全忆阻器基神经形态计算系统的研究进程。

（3）针对脉冲神经网络训练算法不成熟的问题，引入了转换脉冲神经网络的理念。利用电路内部固有的寄生电容，构建了 NbO_x 基的 1T1R 神经元来匹配人工神经网络中的 ReLU 激活函数，并实验验证了基于 1T1R 神经元和忆阻器突触阵列的单层转换脉冲神经网络（320×10），在 MNIST 数据集上实现了 85.7%的识别率。最后，为了实现并行多任务和更好的系统集成，提出了 1T1R 神经元的 X-bar 集成结构。该工作为利用忆阻器构建高效的神经形态芯片提出了一个新的技术途径。

（4）为构建感存算融合的神经形态感知系统，发展了神经形态计算机与环境之间交互的神经接口电路。首次提出并实验验证了一种基于 NbOx-Mott 忆阻器的传入神经电路，该传入神经电路可以将模拟信号转换成动态的频率信号。为验证其功能，我们进一步将传入神经与压电传感器相连，构建了一个不需任何外部电源的人工机械感受系统。该感受系统能够对压力信号做出响应，并将压力强度转换为相应的振荡频率。该传入神经还可以很容易地扩展到处理来自其他传感器的感知信号，例如味觉、视觉、听觉、温度、磁场和湿度等。该工作进一步扩展了忆阻器基神经元的功能，展示了利用忆阻器构建感存算一体的智能系统的可行性。

7.2 未来工作的展望

针对当前基于忆阻器的神经形态计算的发展状况和研究热点，并结合本论文在忆阻器基神经形态器件及系统方面开展的研究工作，对忆阻器基神经形态智能系统的构建提出了以下几点展望：

（1）当前，忆阻器基神经形态计算的绝大多数工作都是集中在对神经突触的功能实现上，在系统应用中对器件的性能需求也已有系统探讨。然而，当前报道的突触阵列规模还比较小，不足以支撑产业应用。一方面是因为当前制备工艺还不足够成熟，需

要产业界和学术界进一步的努力；另一方面是因为突触阵列在用作计算时通常是进行并行读取，大阵列的读取过程会产生大的电流容易损坏外围电路且产热比较严重；因此，制备低电流的突触器件是实现大阵列应用的必然趋势。

（2）对于忆阻器基神经元电路的研究来说，目前正处于起步阶段，神经元所能实现的功能比较单一，尚缺少系统层次的探讨。若要使用忆阻器神经元进行一定规模的演示，还需要从器件功能验证到外围辅助电路的优化设计等方面展开更加系统和深入的研究。

（3）高效的脉冲神经网络算法是神经形态芯片实现智能化的内在灵魂，而现有的算法还不够成熟而且并不是针对忆阻器硬件平台提出的，因此，针对忆阻器的内在特性实现硬件和算法的共同设计是利用忆阻器构建神经形态系统的关键。

（4）人脑的高计算效率来源于其感存算一体的天然结构，这使得她们可以实时地获取外界信息并进行处理，然后及时地作出反应。国际上利用忆阻器感知信息的工作已有报道，接下来考虑如何进一步与深层神经形态信息处理系统结合是未来构建感存算一体人工智能系统的必经之路。

（5）3D 神经形态芯片是实现高密度神经形态计算所追求的目标，忆阻器器件具有易三维集成的优势，结合先进的传感器和逻辑器件在不同层制备具有不同功能的器件分别用于感知、数据处理和信息存储已经成为行业研究的热点和重要的技术路线之一。