附件2

|  |  |
| --- | --- |
| 项目名称 | 新型神经元忆阻器件与全忆阻器神经网络的研究 |
| 项目类别 | □C类√D类 |
| 技术领域 | √信息 □环保 □健康 □旅游 □时尚 □金融 □高端装备制造□文化 □海洋经济 □生物技术 □新能源 □新材料 □其他 |
| 所在平台 | □省级留学人员创业园：                   □省重点企业研究院：                   □省级产业集聚区：                    √其他 |
| 是否回国来浙从事博士后研究工作 | □是，海外博士授予学校： 回国时间： 博士后编号：  |
| √否 |

浙江省“钱江人才计划”C、D类

项 目 申 请 表

姓 名 黄科杰

单 位 浙江大学

部门（地区） 信息与电子工程学院

浙江省人力资源和社会保障厅

**填表说明**

1．技术领域：请在以下相应产业领域栏目打“√”：信息、环保、健康、旅游、时尚、金融、高端装备制造、文化、海洋经济、生物技术、新能源、新材料；不属上述产业领域的，请在“其他”栏打“√”。

2．所在平台：若项目属省级留学人员创业园、省重点企业研究院、省级产业集聚区的，请在相应栏打“√”，并填写相应名称。不属于上述内容的，请在“其他”栏打“√”。

3．是否回国来浙从事博士后研究工作：在“是”或“否”前打“√”，若“是”的，填写相应栏目。

4．表内各栏目填写内容的起讫时间均为最近5年，2017年申请的，各栏目起讫时间为2012年1月至今。

一、申请人基本信息

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 | 黄科杰 | 工作单位 | 浙江大学 |
| 职 务 | 研究员 | 从事专业 | 半导体器件，集成电路，机器学习 |
| 联系地址 | 浙江省杭州市浙大路38号玉泉校区 | 邮 编 | 310027 |
| 单位电话 | 87951754 | 手 机 | 13185000712 | E-mail | huangkejie@zju.edu.cn |
| 留学国别 | 新加坡 | 出国时间 | 2007/12 | 回国时间 | 2016/10 |
| 留学机构名称 | 新加坡国立大学、新加坡科杰研究局、新加坡科技设计大学、新加坡赛灵思（Xilinx）公司 |
| 留学性质 | □公派√自费 | 学习性质 | □大学   □硕士  √博士  √博士后□普访   □高访   □其他\_\_\_\_\_\_\_\_\_\_ |
| 主 要学 习工 作经 历 | 教育经历2011/8–2014/7， 新加坡国立大学，电子与计算机工程，博士，导师：连勇 2004/9–2006/6， 浙江大学， 电路与系统， 硕士， 导师：沈海滨1999/9–2003/6， 浙江大学， 电子信息工程， 学士工作经历2016/10-至今，浙江大学，信息工程与电子学院，研究员2013/7-2016/10，新加坡科技大学（SUTD），Engineering Product Department，博士后研究员2011/2-2013/6，新加坡科技研究局（ASTAR），Data Storage Institute，研究工程师2008/1-2011/2，赛灵思（XILINX），设计部门，工程师2006/12-2007/12，格科微电子，设计部门，工程师2006/7-2006/11，三星半导体，设计部门，工程师 |
| 从事专业工作情况 | （概述本人的专业研究领域、方向和主要业绩）申请人过去几年的研究工作，主要是从事**新型非易失性存储器，低功耗电路**和**神经形态芯片的开发研究工作**。近五年来，申请人已发**20篇**国际期刊论文和会议论文，其中**8篇一作和1篇共同一作**论文发表于顶级期刊，如**IEEE TCAS，TVLSI，TNANO等**。总影响因子超过**40**。申请人的研究成果被许多欧美日著名学者所认可和引用。另外多篇高质量论文正在一些顶级期刊评审（或即将投稿），如**Nature Nanotechnology**, **ACS Nano, IEEE TCAS, Elsevier JPDC**等等。同时也是许多国际顶级期刊和知名会议的审稿人，如**IEEE TCAS，TVLSI, TED, TC，EDL，ISCAS**等等。此外在非易失性存储器电路设计领域还有**4个美国授权专利**，**3个美国专利申请公开**。2011年主持工业界横向项目一个（约**400万人民币**），开发一块16kbit的RRAM阵列。负责过研究项目若干。申请人在2016年以“百人计划”研究员身份加入浙江大学。加入浙江大学之后，积极展开合作，与美国华为研究所，杭州海康威视，上海微系统所，清华脑科学中心，成都电子科技大学等展开密切的合作与交流。目前还在通过新加坡科技设计大学申请新加坡TLAB的一个神经形态芯片设计的项目。申请人目前的研究领域包括非易失性存储阵列和逻辑电路设计，非易失性神经形态芯片设计，机器学习等。目前完成的研究工作主要如下：1. **基于忆阻器的神经形态芯片**

因为机器计算能力的大幅度增强，机器学习特别是深度学习在人工智能领域得到了前所未有的重视和发展。其应用包括图像识别，语音识别，医疗，人机交互等各个方面。但是目前机器学习主要依赖于服务器计算能力，在一些实时或者低功耗等应用方面受到极大的限制。另一方面，传统计算机主要基于冯\*诺依曼（Von-Neumann）架构，在神经网络方面的计算，如卷积，仍有很多不足之处。于是，各种优化的神经网络加速器开始出现，如google的TPU，MIT和NVIDIA合作的用于卷积神经网络（CNN）的EYERISS（局部存储器和权重值重用），中科院开发的寒武纪（乘法器阵列和加法器阵列）。虽然这些新的设计一定程度上优化了这些处理器在神经网路方面的处理能力，但是仍然有很多不足之处。因为他们仍然需要传统的乘法器和加法器（或者ALU）来实现矩阵的乘机和求和。在这里有三个主要的问题，一是**计算单元使用乘法器和加法器**，二是**存储单元使用缓存（SRAM）-主存（DRAM）-硬盘（如FLASH）的层次结构**，三是**存储单元与计算单元数据通讯**问题。存储单元和计算单元分开，往往导致他们之间的连线过长，连线功耗过高，并且极大的影响系统的速度。特别是深度学习的应用，一般需要非常大的网络规模数据，导致**数据在处理器和主存之间频繁地移动**。计算单元使用加法器和乘法器会需要较大的面积和较高的功耗来完成复杂的卷积计算。传统的存储器和存储器架构也同样需要较大的面积和较高的读写和漏电功耗。IBM的TrueNorth使用了神经形态的架构，解决了运算单元和存储器分离的问题，并且简化了神经网络运算。而**忆阻器（Memristor）用来模拟神经突触（Synapse）又给神经形态芯片的设计带来了新的契机，可以很好的解决存储器的问题**。申请人在神经形态芯片的设计上，主要包括了下面三个方向：一是开发新型器件，二是优化学习算法和电路，三是设计系统架构与芯片。新型器件开发，主要是用于神经突触和神经元的忆阻器。申请人在**世界上首次**报道了可同时用于STDP和SRDP的忆阻器。STDP和SRDP是脉冲神经网络的**两种最重要的学习算法**。这一成果以**共同第一作者**发表在**Scientific Reports**上。另外神经网络的学习所需要的时间往往远多于处理所需要的时间。所以为了提高神经形态芯片的学习速度，在**Advanced Material**上发表了纳秒级高速相变存储器用来作为神经突触。申请人并不满足用于神经突触的存储器的开发，同时也致力于用于神经元的存储器的开发。我们正在开发超低功耗的**具有leaky特性的神经元，这一工作将给非易失性神经形态芯片带来革命性的改变**，不仅将彻底改变神经形态芯片的架构，而且可以大幅度提高神经形态芯片的密度降低计算功耗。在芯片开发上，申请人独立设计完成了一款256x256的神经形态芯片，其中神经突触由相变存储器实现，其它由CMOS实现。这是世界上**第二款**用相变存储器设计的神经形态芯片。芯片将于下个月完成加工，开始进行测试。与传统CNN加速芯片不同的是，这里将存储单元（相变存储器）和运算单元（leaky-integrate-and-fire neuron）紧密结合在一起，大幅度提高了密度，并减少了在存储器和运算单元传递数据所消耗的资源。相比国际同行，HPL是4x4的阵列，UCSB是12x12的阵列，IBM是256x256的阵列，目前**我们的芯片在规模上处在国际前列。**IBM开发的基于相变存储器的神经形态芯片（同样也是256x256）使用了2T1R的结构，以及COBA神经网络架构，不利于网络拓展和可重构应用。申请人针对低功耗高密度使用了1T1R结构和CUBA神经网络架构。1T1R结构将有助于**减少40%的面积**。CUBA结构适合**可重构**以及**片外学习**的神经网络架构。同时我们**对神经元使用了时分复用以及二级神经元的方法来提高密度。**芯片的制造在中心国际完成，后端相变存储器工艺与上海微系统所合作。该工作准备以**第一作者**投集成电路顶级会议**ISSCC**。在学习算法方面，将紧密结合硬件架构和器件特性，提出了多种有效的算法。比如对之前基于相变存储器的神经形态芯片设计，首次提出了scheduled STDP片上训练算法，对IBM的相变存储器神经形态芯片的学习算法做了改进，减少了不必要的学习过程，**功耗将大幅度降低**。并且可**适用于多种神经编码**，如rate coding或者spike coding。该工作准备以**第一作者**投**IEEE TCAS**。1. **非易失性逻辑电路**

申请人设计了多种非易失性逻辑电路来解决以下几个问题：一、计算单元和存储单元分离的问题，二、高休眠能耗问题，三、非易失可重构电路设计问题，四、非易失性存储阵列相关设计。因为芯片面积的增大，导致了处理器和存储器之间布线的增长，以致大幅增加数据传输的动态功耗和延时。计算单元和存储单元分离的问题主要靠存储器内计算（In Memory Computing）的方法来解决。申请人提出的存储器内计算的方案，不仅解决了连线的问题，同时**世界上首次**提出用**时分复用**的方法来**共享存储器**解决信号传递功耗过高的问题。这些成果，体现在**TCASI（2016）**，**NVMTS**和**NANOARCH**几篇文章中。高休眠功耗主要是因为晶体管的特征尺寸的缩小和芯片面积增大的结果。**TNANO（2013）**一文，提出了用**two-step-write**的方法，有效的降低了系统休眠时保存状态所需要的功耗，整体功耗可降低**20%**。而**TCASI（2014）**一文，用**区域存储阵列**的方法，结合**“four-phase pipelined saving”**, **“two-sigma-write”**等方法来进一步降低状态存储所需要的功耗。相比TNANO（2013），整体功耗可进一步降低**20%**。这些方案，已经获得了**美国授权专利**。**TCASI(2013)**和**TVLSI（2014）**两文，分别解决了相变存储器和阻变存储器在FPGA应用上的可重构问题和可靠性问题。前者的可靠性问题是由于相变存储器非零偏置的原因，会造成读干扰（amorphous态的相变存储器会变成crystalline态）。申请人提出了用零偏置的非易失性配置存储器来提高可靠性超过15000倍，并降低功耗174倍。同时还提出了多态非易失性可重构存储器用于实时可重构（Real-Time-Reconfiguration）。TVLSI（2014）主要解决阻变存储器作为FPGA可重构开关写入的问题。因为crossbar这种结构会有严重的写串扰和高漏电功耗的问题，特别是在FPGA可重构这个应用上。申请人提出用2D1R这种结构，提高写可靠性1e8，提高速度53%，并降低功耗40.5%。非易失性逻辑电路的研究为神经形态芯片的开发奠定了非常良好的基础。存储器内计算，存储器阵列设计等等都是神经形态芯片设计需要解决的问题。 |

二、五年来主要成果

|  |
| --- |
| 1、参与过的主要项目 |
| 项目名称 | 起止时间 | 项目性质和来源 | 经费总额 | 参与人数、本人排名和任务 |
| 浙江大学百人计划 | 2016-2019 | 中央国库、浙江大学 | 200万 | 1，1，主持 |
| Development of a 16 Kb Resistive Random Access Memory (RRAM) Prototype and Memory Controller | 2011-2013 | 工业界、4DS | 80万新币 | 7，1，主持 |
| Crossbar Memristor based Neuromorphic System for Learning in Spiking Neural Network (SNN) | 2014-2017 | ZJU-SUTD IDC | 40万新币 | 7，1，主持 |
| Non-Volatile Memory based Low Power Field Programmable Gate Array for Smart Electronics Applications | 2014-2017 | MIT-SUTD International Design Center | 25万新币 | 7，1，主持 |
| Bio-plausible Neuron and Bio-plausible Low Power Synapse for Bio-inspired Computer | 2014-2017 | MOE T2 | 50万新币 | 7，1，主持 |
| Artificial cognitive memory | 2011-2013 | A\*STAR SERC TSRP | 360万新币 | 7，1，主持 |
| PCRAM/RRAM integrating with advanced logic devices | 2011-2013 | A\*STAR SERC TSRP | 400万新币 | 7，1，主持 |

|  |
| --- |
| 2、代表性论文、著作（不超过20项） |
| 论文、著作名称 | 发表/出版时间 | 发表/出版载体 | 论文索引情况 | 本人排名 |
| A Low Active Leakage and High Reliability Phase Change Memory (PCM) Based Non-Volatile FPGA Storage Element | 2014.9 | IEEE Transactions ON Circuits and Systems       I-Regular Papers | 27 | 1 |
| A Low Power Localized 2T1R STT-MRAM Array With Pipelined Quad-Phase Saving Scheme for Zero Sleep Power Systems | 2014.9 | IEEE Transactions ON Circuits and Systems       I-Regular Papers | 4 | 1 |
| Magnetic Domain-Wall Racetrack Memory-Based Nonvolatile Logic for Low-Power Computing and Fast Run-Time-Reconfiguration | 2016.01 | IEEE Transactions on Very Large Scale Integration       (VLSI) Systems | 0 | 1 |
| High-Density and High-Reliability Nonvolatile Field-Programmable Gate Array with Stacked 1D2R RRAM Array | 2016.01 | IEEE Transactions on Very Large Scale Integration       (VLSI) Systems | 4 | 1 |
| A Low Power and High Sensing Margin Non-Volatile Full Adder Using Racetrack Memory | 2015.04 | IEEE Transactions ON Circuits and Systems       I-Regular Papers | 6 | 1 |
| A Low-Power Low-VDD Nonvolatile Latch Using Spin Transfer Torque MRAM | 2013.11 | IEEE Transactions ON       Nanotechnology | 15 | 1 |
| Optimization Scheme to Minimize Reference Resistance Distribution of Spin-Transfer-Torque MRAM | 2014.05 | IEEE Transactions on Very Large Scale Integration       (VLSI) Systems | 11 | 1 |
| Racetrack Memory based Non-volatile Storage Elements for Multi-context FPGAs | 2015.09 | IEEE Transactions on Very Large Scale Integration       (VLSI) Systems | 2 | 1 |
| Enabling an Integrated Rate-temporal Learning Scheme on Memristor | 2014.04 | Scientific       Reports | 26 | 共1 |
| Tailoring transient-amorphous states: Towards fast and power-efficient phase-change memory and neuromorphic computing | 2014.01 | Advanced       Materials | 7 | 3 |
| A Novel Operation Scheme Enabling Easy Integration of Selector and Memory | 2017.01 | IEEE       Electron Device Letters | 0 | 4 |
| Low power computing using STT-MRAM | 2014.10 | NVMTS | 3 | 1 |
| STT-MRAM based low power synchronous non-volatile logic with timing demultiplexing | 2014.08 | NANOARCH | 5 | 1 |
| Artificial neuron with somatic and axonal computation units: Mathematical and neuromorphic models of persistent firing neurons | 2011.11 | International Joint Conference on Neural       Networks (IJCNN) | 4 | 2 |
| Axonal slow Integration induced persistent firing neuron model | 2012.06 | International Conference on       Neural Information Processing | 3 | 3 |
| Design and Optimization of Inductive Power Link for Biomedical       Applications | 2011 | InTech | 0 | 1 |

|  |
| --- |
| 3、专利 |
| 专利名称 | 专利类别 | 批准时间 | 授权国家 | 是否投产 |
| Current Writing Circuit for a Resistive Memory Cell Arrangement | 发明 | 2012 | 美国 | 否 |
| Writing Circuit for a Resistive Memory Cell Arrangement and a Memory Cell Arrangement | 发明 | 2013 | 美国 | 否 |
| Reading circuit for a resistive memory cell | 发明 | 2014 | 美国 | 否 |
| Latch circuit and data processing system | 发明 | 2014 | 美国 | 否 |
| Writing circuit for a magnetoresistive memory cell, memory cell arrangement and method of writing into a magnetoresistive memory cell of a memory cell arrangement | 发明 | 2014 | 美国 | 否 |
| Circuit arrangement and a method of writing states to a memory cell | 发明 | 2015 | 美国 | 否 |
| Write Control Circuits And Write Control Methods | 发明 | 2016 | 美国 | 否 |

4、产品（如有产品，说明目前的产业化程度）

 无

5、其他（包括获得的重要奖项、在国际学术会议作重要报告等情况）

Low Power Computing Using Resistive Non-volatile Memories. In 14th Non-volatile Memory Technology Symposium (NVMTS), 27-29 October, 2014. (Invited)

STT-MRAM based Low Power Synchronous non-volatile Logic with Timing Demultiplexing. In 10th ACM/IEEE International Symposium on Nanoscale Architectures (NANOARCH), 8-10 July, 2014.

Domain Wall Racetrack Memory for In Memory Computing. In 50th IEEE International Symposium on Circuits and Systems, 28-30 May, 2017

三、项目可行性说明

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1、立项背景（说明项目意义、国内外研究现状和发展趋势）。**(1) 研究意义**人类对自然界的不断探索和认识是推动人类科学发展的重要手段。几十年来，通过对人脑和神经网络的不断认知，人工智能已有引人瞩目的发展，机器的智能已经在某些方面超越人类。如Alphago于2016年3月以4:1的成绩战胜了韩国围棋职业高手李世乭[1]。虽然机器在某些特定的领域表现出超越人类智慧的能力，但现有计算系统仍然面临两个严重的发展瓶颈：一是对于人脑能轻松胜任的认知任务处理能力不足，难以支撑高水平的智能。二是系统能耗过高，如Alphago需要消耗一百万瓦的功耗才能战胜消耗二十瓦功耗的人脑[2]。这主要是因为现有人工智能都是依靠冯诺依曼架构来实现。因为存储器和处理器分开，**用冯氏计算机来模拟人工神经网络非常低效**，大量数据需要频繁的在处理器和存储器之间移动，导致功耗过高，运算能力不足，难以模拟大型网络结构。为了提高人工神经网络的效率，必须研究适合人工神经网络的硬件。**神经形态芯片采用了仿生的架构，将存储器和处理器紧密的结合在一起，能以非常低的功耗传递数据[3]。**在生物脑中，数据本身则是分布式的存储于网络的各个节点（比如由神经元内的离子浓度表征）以及节点之间的连接（比如由突触的强弱表征）上，运算和存储在结构上高度一体化[4]。这样，用少量甚至单个电子器件模仿单个神经元的功能，而将数量巨大的电子“神经元”以类脑的方式形成大规模并行处理的网络，就成为了非常有吸引力的方向。图1（a）是一个生物上神经元的示意图，包括活体脑细胞（SOMA），轴突（Axon）和树突(Dendrite)[5]。轴突（Axon）用来发送神经信号，树突（Dendrite）接收来自抑制（Inhibitory）神经突触或者刺激（Excitatory）神经突触的信号，增加或者减少神经元膜（Membrane）的电压。神经元被激发（Fire）之后以脉冲的形式被轴突发送，并经过神经突触到达目标神经元。神经突触用来记忆活动历史记录，并调制信号强弱。如图1（b）所示，整个大脑就是由大量的神经元通过神经突触连接而组成，通过脉冲序列来进行通讯，被称为神经网络。神经网络很快被抽象出来模拟成人工神经网络用到人工智能上，如图1（c）所示。在人工神经网络，每一层由若干个神经元组成（如图1（c）中的圆圈），而每两次之间由神经突触来连接（如图1（c）中的线条）。脉冲神经网络（SNN）的神经形态芯片由于其更加仿生，得到了大量的关注与发展[6]。脉冲神经网络不仅可以用于解决非脉冲神经网络所能解决的问题，而且在处理一些时空性相关的任务的时候效率更高[7]。许多研究机构在开发基于脉冲神经网络的神经形态芯片来模拟人类大脑，如曼彻斯特大学的spinnaker[8]，斯坦福大学的neurogrid[9]，苏黎世大学的ROLLS[10]等等。另一方面，研究者也希望能把人脑的认知服务于人类，如IBM开发出了TrueNorth[11]，一种简化了的神经形态芯片架构，但是更适合做智能运算。C:\Users\lenovo\AppData\Local\Microsoft\Windows\INetCacheContent.Word\integ2.gifhttp://mag.uchicago.edu/sites/default/files/1508_Kelly_Neural-network.jpghttps://www.extremetech.com/wp-content/uploads/2015/07/NeuralNetwork.png1. (b) (c)

图1 （a）神经元结构示意图，（b）大脑中的神经网络，（c）人工神经网络虽然神经形态芯片是更接近人脑的一种硬件架构，相比冯氏计算机具有更低的功耗和更高效的处理速度，但是要处理更复杂的场景需要更大规模的网络。**神经形态芯片开发的最大局限性在于芯片的密度。**虽然工业界和学术界在传统CMOS工艺做了各种努力，来提高人工智能的运算能力。但是，传统CMOS工艺的局限性决定了芯片和系统很难提高片上存储的容量来满足现有的深度学习的要求，更远远无法实现具有1011个神经元和1015个神经键的人脑规模，并实现近似人脑的功耗。在芯片或者系统级实现如此规模的神经突触需要高密度的存储器。而忆阻器（Memristor）的出现又给神经形态芯片带来了新的契机。常见的忆阻器有相变存储器（PCM），阻变存储器（RRAM）等，一般都是三明治结构，两端是金属电极，中间是电阻转换材料。忆阻器由于其高密度、低功耗、非易失、后端友好工艺，而且许多特性可以很好的模拟神经突触，所以**忆阻器被认为是实现神经形态芯片的最佳存储器件**，从而解决密度的问题[12]。用于神经突触的高密度忆阻器已经多有报道。但是用于神经元的忆阻器还鲜有报道。神经元是神经形态芯片的一个重要组成部分，**高密度的神经元器件和高密度神经突触是高密度神经形态芯片的两个重要组成部分**。**我们在研究忆阻器件的时候发现忆阻器件具有积分的效应，在模拟神经元方面有着非常大的潜力**。本课题拟研究基于忆阻器的神经元器件，需要完成的任务主要有两点：* 高密度低功耗的神经元忆阻器件。该器件不仅可以做积分运算（Integrate），而且带泄漏（Leaky）特性以及自动重置（Auto Reset）特性。
* 基于全忆阻器的神经形态芯片架构和电路的研究。本课题拟研究基于忆阻器的芯片架构，它将具有这几个特点：高密度、高处理速度和低功耗。这也是所有硬件系统所最求的目标。

**对忆阻器神经元、电路和架构做一定研究之后，下一步的目标是开发并制造全忆阻器神经形态芯片。全忆阻器神经形态芯片将大幅度提高人工智能，使之更接近人类的智能，对人工智能的研究具有革命性的意义。而且在规模达到一定之后，就可以拿来模拟和帮助研究人类的大脑，了解机理，并用来服务于人类。****（2）国内外研究现状及分析**目前国内外还未有人研究基于全忆阻器的神经网络，主要是因为忆阻器是新型器件，能研发的团队不是很多，这些团队目前主要专注于用于神经突触的忆阻器器件，而鲜有人涉及用于神经元的忆阻器器件。目前忆阻器神经元器件只有IBM以“Stochastic phase-change neurons”为题于2016年发表在Nature Nanotechnology上。1. **各国发展项目概况**

**因为神经形态芯片旨在模拟人类复杂而神秘的大脑，并将其用于国防，军事，航天航空，机器人，医疗等各个领域，因此各个国家和地区投入了大量的人力和物力。**2013年4月美国总统奥巴马批准实施“创新型神经技术推动脑科学研究”（BRAIN）计划[13]。为了大力推进实现BRAIN计划的目标，工作组建议未来5年（2016～2020财年）NIH的投资达到每年4亿美元，随后5年（2021～2025财年）为每年5亿美元。“欧盟人脑计划”于2013年入选了欧盟的未来旗舰技术项目，也成为了全球范围内最重要的人类大脑研究项目[14]。该项目旨在建立一套基于神经科学的全新的、革命性的 信息通信技术，建造一种模拟神经元功能的芯片，然后将这种芯片用于建造超级计算机系统。该计划将持续十年，整体投资11.9亿欧元。日本于2014年正式启动庞大的“脑科学与教育”研究项目，其总体目标是：将脑科学研究作为国家教育发展的一项战略任务，将脑科学与教育紧密结合，进行面向教育理论和实际的应用研究[15]。随着欧、美、日相继启动各种人脑计划，中国也将全面启动自己的脑科学计划。科技部正在做面向2030年的科学研究计划，制定中国的脑科学和类脑研究方案，工程和生物研究将交叉融合。“中国脑计划”已获国务院批示，并被列为“事关我国未来发展的重大科技项目”之一，将从认识脑、保护脑和模拟脑三个方向全面启动。1. **基于传统CMOS工艺的类脑芯片研究现状**

神经形态芯片的初衷是制造VLSI芯片来模拟人的大脑进行一系列的仿真，进一步了解人脑的工作机理，加深对人脑的认知。后续的发展是，借鉴人脑的架构和学习规则来建立高密度低功耗计算芯片，来处理人工智能上的各种问题。由于其借鉴生物上的人脑，一般都是脉冲神经网络，而且计算单元（神经元）和存储单元（神经突触）往往紧密的结合在一起。IBM公司2014年8月研制出TrueNorth芯片。该芯片采用28纳米硅工艺制作，包括54亿个晶体管和4096个处理核，相当于100万个可编程神经元，以及2.56亿个可编程突触。每秒可执行460亿次突触运算，总功耗仅为70 mW，每平方厘米功耗20 mW [11]。2014年4月底，美国斯坦福大学的研究人员研制出一款基于人脑构造的高速节能微型芯片Neurogrid， 其速度为普通电脑的9000倍，而所需的能量低于普通电脑[9]。该原型产品由16个定制芯片组成，能够模拟100万个大脑神经元以及几十亿个突触连接。在欧盟人脑计划支持下，海德堡大学现已研制出包含20万个神经元和5000万个神经元突触的神经形态芯片，并将尝试将20个这样的芯片链接起来[16]。与此同时，欧盟人脑计划为英国曼切斯特大学主持的SpiNNaker（脉冲神经网络架构）计划提供了资助，该计划利用两万个ARM芯片搭建大规模并行计算网络，每个芯片代表1000个神经元[8]。欧盟人脑计划的资助将帮助SpiNNaker实现上百万颗ARM处理器的集成。 1. **基于忆阻器的神经网络研究现状**

以上的神经形态芯片，他们的共同的问题仍然是密度的问题。表1中列出了一些使用了片上神经突触的神经形态芯片参数。在神经突触密度一栏，ROLLS和HICANN都做了标准化处理。IBM TrueNorth密度比较高，但是只实现了1-bit的神经突触，所以在模拟人脑的研究上会有很大欠缺。ROLLS和HICANN使用CMOS电路来模拟复杂的神经突触，密度大幅度下降。所以用现有CMOS工艺来模拟并研究人脑，很难做到人脑的规模。忆阻器由于其高密度、低功耗、非易失、后端友好工艺，而且许多特性可以很好的模拟神经突触，所以被认为是模拟神经突触的最理想器件。目前基于忆阻器的芯片主要有以下几个：* HPL的DPE，一个4x4的忆阻器网络[17]
* UCSB的12x12忆阻器阵列[18]
* IBM的256x256带片上学习功能的相变存储器阵列[19]

表1 神经形态芯片以及人脑的比较

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chip | SNN？ | Technology (nm) | Neuron | Synapse/Memory | Clock | Power (W) | Performance (GSOPS) | Normalized (GSOPJ/GFLOPJ) | Area (cm2) | Synapse density (M/cm2) |
| TrueNorth | Y | 28 | Digital (I&F) | 256 Mb 1-b SRAM | 1 KHz | 0.07 | 400 | **5700** | 4.3 | 59.5 |
| ROLLS | Y | 180 | Analog (I&F) | 128 Kb LTP + STP | 100 HZ | 0.004 | 0.00655 | 1.6375 | 0.514 | **10.3** |
| HICANN | Y | 180 | Analog (I&F) | 128 Kb 4b-DAC | 1 MHz | 1.3 | 164 | 126 | 0.5 | 10.6 |
| Human Brain | Y | - | Biological  | 1 Petabit | 10 Hz | 20 | ~ 1e9 | **5e7** | 1260 cm3 | 794 G/cm3 |

1. **基于忆阻器的神经网络研究现状**

以上的设计，一个共同的问题仍然是密度的问题。表1中列出了一些使用了片上神经突触的神经形态芯片参数。在神经突触密度一栏，ROLLS和HICANN都做了标准化处理。可以看出，他们的密度还是远远低于人脑的密度。所以需要用忆阻器来提高芯片的密度。目前基于忆阻器的芯片设计主要有以下几个：* HPL的DPE，一个4x4的忆阻器网络[28]
* UCSB的12x12忆阻器阵列[29]
* IBM的256x256带片上学习功能的相变存储器阵列[30]

因为忆阻器制造的困难，以及网络架构，学习算法和电路的不成熟，目前都还处于探索阶段。如HPL和UCSB分别只做到4x4和12x12。IBM使用了相对成熟的相变存储器工艺，做到了256x256。HPL和UCSB使用了模拟矩阵相乘的方法，主要用于传统的神经网络，但是会有非常高的静态功耗，而且神经元需要用到ADC/DAC[20]，造成面积过大功耗过高。另外，UCSB使用了无选通器件的纵横阵列，潜通路（Sneak Path）上会造成非常高的漏电流和写错误。IBM的相变存储器阵列使用了片上学习电路，用于脉冲神经网络。他们的芯片同样有两个问题，一是使用2T1R造成芯片密度降低；二是神经元使用电容来做积分，有面积过大的问题。所以上述三种神经形态芯片，都需要通过减小神经元的面积来提高整个芯片的密度。1. **神经元电路和器件研究**

在脉冲神经元模型中，最常见的有Leaky Integrate and Fire (I&F) 神经元模型[21]，Izhikevich (IZ)神经元模型[22]，以及Hodgkin-Huxley (HH)神经元模型[23]。**而Leaky I&F神经元模型已被报导可以用于基于忆阻器的神经突触，但是都需要使用电容来做积分。电容的缺点主要是面积过大。**为了解决这个问题，IBM提出了用相变存储器（PCM）的积分特性来模拟神经元[24]，然后检测PCM的阻值，若超过阈值电阻，就输出脉冲信号，并重置神经元（PCM阻值）。这样的神经元忆阻器有几个问题，一是改变PCM的电阻态需要比较高的电流，特别在低阻态的时候。需要的电流一般在几至几十微安（uA）这个数量级。而重置（Reset）也需要更高的能量，一般电流在几十至几百微安。而且，用PCM实现的神经元还需要额外的读脉冲去检测PCM电阻是否已经超过阈值电阻。这样不仅会带来额外的面积和功耗，而且会中断神经元积分的过程。目前关于神经元忆阻器及其相关电路和网络架构的研究基本上还处于萌芽阶段，要最终形成对神经形态芯片突破性的发展，还需要做更多探索性的研究。在国内对神经形态芯片及其应用研究日益重视的背景下，适时启动一些诸如用于神经元器件、电路和基于全忆阻器的神经网络的研究课题是非常必要和迫切的。**参考文献**[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche*, et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature,* vol. 529, pp. 484-489, 2016.[2] R. A. Nawrocki, R. M. Voyles, and S. E. Shaheen, "A Mini Review of Neuromorphic Architectures and Implementations," *IEEE Transactions on Electron Devices,* vol. 63, pp. 3819-3829, 2016.[3] F. C. Morabito, A. G. Andreou, and E. Chicca, "Neuromorphic Engineering: From Neural Systems to Brain-Like Engineered Systems," *Neural Networks,* vol. 45, pp. 1-3, 2013.[4] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth, *Principles of neural science* vol. 4: McGraw-hill New York, 2000.[5] R. Ge, H. Qian, N. Chen, and J. H. Wang, "Input-dependent subcellular localization of spike initiation between soma and axon at cortical pyramidal neurons," *Molecular Brain,* vol. 7, p. 26, 2014.[6] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural networks,* vol. 10, pp. 1659-1671, 1997.[7] R. Brette, M. Rudolph, T. Carnevale, M. Hines, D. Beeman, J. M. Bower*, et al.*, "Simulation of networks of spiking neurons: a review of tools and strategies," *Journal of computational neuroscience,* vol. 23, pp. 349-398, 2007.[8] S. B. Furber, D. R. Lester, L. A. Plana, J. D. Garside, E. Painkras, S. Temple*, et al.*, "Overview of the spinnaker system architecture," *IEEE Transactions on Computers,* vol. 62, pp. 2454-2467, 2013.[9] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat*, et al.*, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proceedings of the IEEE,* vol. 102, pp. 699-716, 2014.[10] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska*, et al.*, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses," *Frontiers in neuroscience,* vol. 9, p. 141, 2015.[11] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan*, et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science,* vol. 345, pp. 668-673, 2014.[12] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters,* vol. 10, pp. 1297-1301, 2010.[13] T. R. Insel, S. C. Landis, and F. S. Collins, "The NIH brain initiative," *Science,* vol. 340, pp. 687-688, 2013.[14] H. Markram, "The human brain project," *Scientific American,* vol. 306, pp. 50-55, 2012.[15] H. Okano, A. Miyawaki, and K. Kasai, "Brain/MINDS: brain-mapping project in Japan," *Phil. Trans. R. Soc. B,* vol. 370, p. 20140310, 2015.[16] J. Schemmel, D. Briiderle, A. Griibl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on*, 2010, pp. 1947-1950.[17] M. Hu, E. P. Date, and I. P. Date, "Dot-Product Engine for Deeping Learning."[18] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature,* vol. 521, pp. 61-64, 2015.[19] S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim*, 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 *Electron Devices Meeting (IEDM), 2015 IEEE International*, 2015, pp. 17.1. 1-17.1. 4.[20] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves*, et al.*, "Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Design Automation Conference (DAC), 2016 53nd ACM/EDAC/IEEE*, 2016, pp. 1-6.[21] A. N. Burkitt, "A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input," *Biological cybernetics,* vol. 95, pp. 1-19, 2006.[22] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on neural networks,* vol. 14, pp. 1569-1572, 2003.[23] L. Abbott and T. B. Kepler, "Model neurons: from hodgkin-huxley to hopfield," in *Statistical mechanics of neural networks*, ed: Springer, 1990, pp. 5-18.[24] T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Stochastic phase-change neurons," *Nature nanotechnology,* vol. 11, pp. 693-699, 2016.[25] G. Van Meer, D. R. Voelker, and G. W. Feigenson, "Membrane lipids: where they are and how they behave," *Nature reviews Molecular cell biology,* vol. 9, pp. 112-124, 2008.[26] M. S. Bretscher, "Asymmetrical lipid bilayer structure for biological membranes," *Nature,* vol. 236, pp. 11-12, 1972.[27] R. Hall and J. H. Racette, "Diffusion and solubility of copper in extrinsic and intrinsic germanium, silicon, and gallium arsenide," *Journal of Applied Physics,* vol. 35, pp. 379-397, 1964.[28] P. M. Fahey, P. Griffin, and J. Plummer, "Point defects and dopant diffusion in silicon," *Reviews of modern physics,* vol. 61, p. 289, 1989.[29] Y. LeCun, "LeNet-5, convolutional neural networks," *URL:* <http://yann>*. lecun. com/exdb/lenet,* 2015.[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.[31] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in computational neuroscience,* vol. 9, 2015.[32] P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, 2015, pp. 1-8.[33] W. He, K. Huang, N. Ning, K. Ramanathan, G. Li, Y. Jiang*, et al.*, "Enabling an integrated rate-temporal learning scheme on memristor," *Scientific reports,* vol. 4, p. 4755, 2014.[34] C. Wang, W. He, Y. Tong, and R. Zhao, "Investigation and manipulation of different analog behaviors of memristor as electronic synapse for neuromorphic applications," *Scientific reports,* vol. 6, 2016.[35] H. Yang, M. Li, W. He, Y. Jiang, K. G. Lim, W. Song*, et al.*, "Novel selector for high density non-volatile memory with ultra-low holding voltage and 10 7 on/off ratio," in *VLSI Technology (VLSI Technology), 2015 Symposium on*, 2015, pp. T130-T131.[36] W. He, H. Yang, L. Song, K. Huang, and R. Zhao, "A Novel Operation Scheme Enabling Easy Integration of Selector and Memory," *IEEE Electron Device Letters,* 2016.[37] K. Huang, Y. Ha, R. Zhao, A. Kumar, and Y. Lian, "A low active leakage and high reliability phase change memory (PCM) based non-volatile FPGA storage element," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 61, pp. 2605-2613, 2014.[38] K. Huang, R. Zhao, W. He, and Y. Lian, "High-Density and High-Reliability Nonvolatile Field-Programmable Gate Array With Stacked 1D2R RRAM Array," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 24, pp. 139-150, 2016. |
| 2、主要内容和预期成果（说明研究开发的主要内容，技术关键（难点）以及最终成果形式和对经济社会发展产生的效益）。(1) 研究内容* **用于神经形态芯片神经元的高密度低功耗忆阻器件的研究。**

研究忆阻器件在模拟神经元的可行性，研究不同的材料和架构，使该器件具有多种神经元特性，包括**积分（Integrate）**，**泄漏（Leaky）**，**自我重置（Auto Reset）**。同时对器件的物理机制进行研究，并建立相应神经元器件模型。该模型不仅反映器件物理特性，而且与神经元模型结合起来，可以用于spice仿真，再进一步建立高层次模型用于基于该器件的脉冲神经网络架构和学习算法的研究。* **基于全忆阻器的神经元电路的研究。**

基于全忆阻器的神经元电路的研究主要有两个目的，一是为了测试神经元忆阻器件并验证其功能。第二个是做基于全忆阻器神经元电路的探索性研究，研究忆阻器神经元在不同神经元模型下的可行性，如Adaptive Exponential I&F，或者Hodgkin-Huxley (HH)神经元模型。同时也将探索神经元忆阻器在传统神经网络上的应用，如Sigmoid神经元。* **基于全忆阻器的脉冲神经网络架构的研究。**

本项目会根据新的神经元忆阻器特性（**高密度和低功耗**）来研究新型神经网络架构。如用基于STDP学习法则的非监督神经网络来分析全忆阻器架构对整个系统性能带来的提升。另外，本项目也会探索神经元忆阻器的非线性在神经网络上的应用，如结合特殊的神经编码来快速传递新型并提高密度。(2) 拟解决的关键科学问题**在神经形态芯片模拟人脑的过程中，最大的难题是芯片的密度。**现有的CMOS半导体技术无法实现1015个神经突触已经1011个神经元规模的网络，特别是神经元和神经突触需要模拟多种生物特性。用忆阻器来模拟神经突触已经有很多研究工作，但是用单个器件去模拟神经元还处于萌芽阶段。本项目试图通过多种手段去解决神经元密度的问题，或者给其他研究者启发，从而共同完成这个具有极大挑战的难题。在具体的研究中，关键的问题主要有下面三点：* 用单个忆阻器件来模拟神经元的主要功能，如积分功能，泄漏功能，以及自动重置功能，以及分析内部的物理原理。这需要通过研究材料，新型器件结构，并经过大量的测试来实现。
* 使用该器件来实现多种神经元模型，如Adaptive Exponential I&F模型，sigmoid神经元模型等。这需要通过结合CMOS电路来验证各种可行性。
* 基于新型神经元忆阻器件的神经网络架构的研究，以及器件的性能与参数对整个神经网络的影响。因为器件的性能未知，于是就给我留下很多值得探索的空间，并且如果利用好这些未知因素，也实现高效神经形态芯片的关键。可以预见的因素有高密度和非线性。前者可以用于新型卷积神经网络架构，后者准备结合神经编码来研究高效神经网络。

(3) 最终成果形式和对经济社会发展产生的效益因为是前沿性的工作和研究，所以最终成功主要以论文和专利来体现。具体的成果列举如下：* 在国外重要学术期刊和国际会议上发表10-15篇高质量论文，其中，顶级期刊和会议论文不少于5篇（如Nano Letter，IEEE Trans，EDL，IEDM，VLSI，DAC，ICCAD等）。
* 申请国内专利2-3个。
* 参加国际合作与交流2-3次，在本项目的基础上申请国际合作项目1-2个。
* 组织学术交流活动或者学术会议1次。

产生的社会效益不会直接体现，需要再做进一步的研究和开发工作。该项目是为开发高密度神经形态芯片做准备，包括器件开发，架构和电路设计等。神经元忆阻器是实现高密度低功耗神经形态芯片必不可少的一部分。最终的神经形态芯片，因其高密度，高性能功耗比，有非常大的潜力抢占人工智能芯片市场。而人工智能的芯片市场正处于井喷式增长的前期，预计10年后将有400亿美元的市场份额。所以，本项目的成功将带来非常巨大的市场前景。 |
| 3、项目实施方案和计划进度安排。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 时间年度研究计划 | 第一年 | 第二年 | 第三年 | 第四年 |
| 春 | 夏 | 秋 | 冬 | 春 | 夏 | 秋 | 冬 | 春 | 夏 | 秋 | 冬 | 春 | 夏 | 秋 | 冬 |
| 神经元忆阻器件制造  | × | × | × |  |  |  | × | × | × |  |  | × | × | × |  |  |
| 建立试验台测试器件特性 |  | × | × | × |  |  |  | × | × |  |  |  | × | × |  |  |
| 忆阻器件神经元特性及器件物理机理的研究 |  |  | × | × | × |  |  | × | × | × | × |  |  | × | × |  |
| Leaky I&F神经元电路设计与仿真 |  |  | × | × |  |  |  | × | × |  |  |  |  |  |  |  |
| Leaky I&F神经元电路制造 |  |  |  |  | × |  |  |  | × |  |  |  |  |  |  |  |
| Leaky I&F神经元电路测试 |  |  |  |  |  | × | × |  |  | × | × |  |  | × | × |  |
| Adaptive Exponential I&F神经元的研究 |  |  |  |  |  |  |  |  | × | × | × |  |  |  |  |  |
| Sigmoid神经元的研究 |  |  |  |  |  |  |  |  |  |  |  | × | × | × |  |  |
| STDP脉冲神经网络建模与仿真 | × | × | × | × | × | × | × | × |  |  |  |  |  |  |  |  |
| 卷积神经网络的建模与仿真 |  |  |  |  |  |  |  |  | × | × | × | × | × | × | × | × |

 |
| 4、现有工作基础和条件（包括配套经费、人员配备等情况）。浙江大学是国内首批成立微电子学院的学校。微电子学院下面包括两个研究所：微纳电子研究所和超大规模集成电路研究所。微纳电子研究所主要从事先进CMOS器件、微纳电子器件、石墨烯电子器件、微纳机电系统、集成光电子学、柔性生物传感系统等研究。超大规模集成电路研究所主要从事嵌入式CPU核，系统芯片设计及平台，模拟和混合信号接口、安全等IP模块，集成电路可制造性设计和软件,智能硬件与智能信息处理等方面的研究。微电子学院为类脑芯片设计和器件开发的研究方向提供的非常良好的平台, 课题申请单位拥有完成本课题所需的全部设备。微纳电子研究所为研究人员提供了微纳超净平台，可以从事半导体工艺研究，器件开发与测试等各项工作。超大规模集成电路研究所为SOC开发提供了良好的平台，包括服务器，集成电路开发软件，以及各种电路测试设备和环境等。配套经费：浙江大学“百人计划”科研启动经费，200万，经费来源：浙江大学，起止年月：2016/12-2019/12, 项目负责人。人员配备：每年1个博士生和1个硕士生的招生资格。目前已经硕士研究生1名，本科生5名，下半年会再增加1名博士生和1名硕士生。同时在积极招聘博士后等研究人员。 |
| 本人声明 | 我保证以上材料属实，如有不实之处，愿承担一切责任。  申请人（签名）：年    月    日 |
| 所在单位审核 | 申请人以上材料经与原件核对，情况属实。表格所填报内容均已在单位内部进行全信息公示，没有异议。 申请人单位（盖章）年    月    日 |
| 所在平台意见 | （不在省级留学人员创业园、省重点企业研究院、省级产业集聚区的，无需填写此栏；若是，请注明平台名称：                       ） （盖章）年    月    日 |
| 市人力社保局或归口管理部门审查意见 |    负责人签字：             年    月    日单位（盖章） |
| “钱江人才计划”管理办公室审查意见 |      负责人签字：             年    月    日单位（盖章） |