

华中科技大学

本科生科研实习报告

院 系 计算机科学与技术

专业班级 ACM1701

姓 名 李蓉

学 号 U201714703

指导教师 胡迪青

2020 年 11 月

目录

本科生科研实习报告.....	1
一、 实习概况.....	1
1.1 研究背景.....	1
1.2 相关工作.....	3
1.3 机遇挑战.....	5
1.4 研究内容.....	6
二、 参加的主要活动.....	7
2.1 研究过程.....	7
2.2 研究结论.....	19
2.3 展望工作.....	20
三、 实习总结.....	20

一、实习概况

1.1 研究背景

AI 时代摩尔定律的失效和存储墙问题的日益突出

(1) 过去几十年，IC 集成电路，尤其是晶体管按照摩尔定律的预言飞速发展，电子计算机芯片性能得益于场效应晶体管尺寸的缩小而拔升。但早在上世纪末 1995 年，摩尔本人在《经济学家》杂志写道：“令我感到最为担心的是成本的增加”，这说明随着尺寸减小，器件成本和工艺难度不断增加，芯片性能提升越发困难，更重要的是，2011 年来硅晶体管接近了原子等级，达到了物理极限，成为算力提升的瓶颈之一，摩尔定律时代面临被“终结”的命运，这从近几年 Intel “挤牙膏”般发布 CPU 也可以窥见一二。总而言之，在摩尔定律即将失效的今天，要想提升计算性能不能再从芯片工艺上下手了，在现有的工艺基础上，唯有彻底改变制造硅材料或者体系结构等途径可以带来突破口，从优化架构入手可能是提高芯片性能最重要的手段。

(2) 随着云计算与物联网、5G 等电子信息重要应用领域的快速发展，现在已步入信息爆炸的大数据时代，海量非结构化数据的深度分析，像 NLP 语义理解和图像识别等 AI 应用需要超高速、高带宽、低功耗和低成本的超高性能计算。传统的计算机中一直是冯·诺依曼架构在主导，源于其在工程设计模块化方面的巨大优势，但其计算和存储分离。

(3) 处理器和存储器受关注度不同，需求不同，工艺不同，封装不同，导致二者之间的性能已有几个数量级的差距，如下图所示，从 1980 年到 2000 年，两者的速度失配以每年 50% 的速率增加，总线速度的提升也十分有限，运算性能受到数据存储和传输速度而不再是 CPU 速度的限制。具体来说，CPU 计算速度小于 1ns，但内存 DRAM 在百纳秒左右，存储速度已然比计算速度小 2 个数量级。在能耗上，仅以 45nm 的工艺为例，加减乘小于 1pJ，但 32 位 DRAM 的读要高达 640pJ，这也是百倍的差距，而且写的功耗会更高。随着半导体工艺的进步，虽然总体功耗下降，但是数据搬运的功耗占比越来越大。根据 Intel 的研究显示，

工艺到了 7nm 时代，访存功耗达到 25pJ/bit (45.5%)，通信功耗达到 10pJ/bit (18.2%)，相加得到的数据传输和访问功耗占比达到了 63.7%。

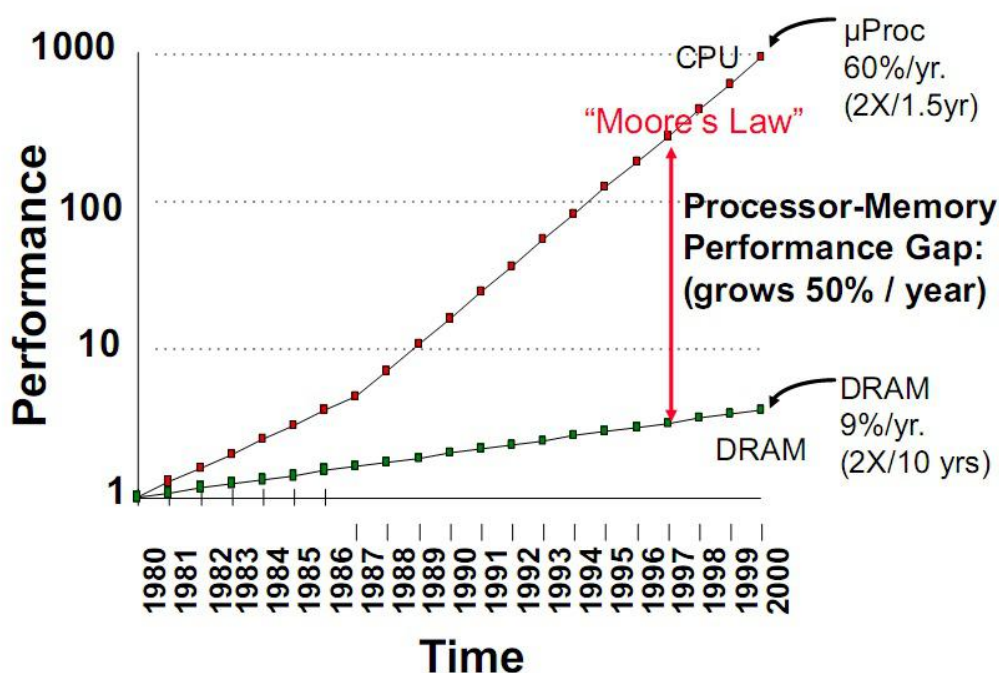


图 1 存储墙剪刀差时间图

(4) 深度学习算法计算需要处理流式数据，大量数据会在计算和存储单元之间流动，基于冯·诺依曼架构的硬件系统在处理相关任务时，整个系统的性能受到数据存储和传输速度的限制，功耗也因为存储读写而很大，冯氏架构的这一现象正如计算机系统结构课程上一样，被称为“存储墙”、“冯·诺依曼瓶颈”或“内存瓶颈”，成为算力提升的第二个桎梏。冯·诺依曼体系结构下存储墙剪刀叉不断增大，访存功耗墙问题也日益突出，工学两界开始从聚焦计算转到聚焦存储，开始从工艺、体系结构等方面尝试缓解这个问题，寻求新的高能效计算技术。

(5) 人工智能的三个浪潮和硬件算力都有关系，70 年代第一次低谷是有了理论模型但无算力，第二波浪潮随着专家系统而起，又因摩尔定律引起的通用计算机性能上扬而下去，第三波浪潮，是从过去十年深度神经网络的提出开始，靠的是 GPU 或者针对数据流设计的专用加速芯片，如 TPU 等硬件加速提供的强大算力，但其存储和计算单元在空间上依旧是分离的，导致深度神经网络的发展也进入瓶颈期。我们仍处于弱人工智能时代，比方说，AlphaGo Zero 需要 5 千个 TPU 训练 40 天才成为地表最强的围棋选手，这耗时还是很大的，远不及强人工智能的

要求。人工智能的广泛应用需要算法提升或硬件革新，但是芯片能提供的算力和人工智能的高需求是矛盾的，芯片至多每 18 个月翻一番集成度来提升算力，但是目前 AI 的需求是每 3 到 4 个月翻一番，因此需要寻找新方法提供算力。

1.2 相关工作

缓解或解决存储墙问题——存算一体

(1) 早在上世纪九十年代存储墙和访存功耗问题暴露之初，研究者就开始寻找解决或者弱化的方法，从最初的多级存储架构，到近存储计算，直到计算型存储/存算一体/存内计算。这些方法大体上可以分为三类：

- 1) 高速带宽数据通信：光互连、2.5D/3D 堆叠。
- 2) 缓解访存延迟和功耗的近数据存储：增加缓存级数、高密度片上存储。
- 3) 缓解访存延迟和功耗的存算堆叠：DRAM 上的逻辑层和存储层的堆叠（类似近数据存储）。
- 4) 软硬件结合的存算融合：SSD 内计算加速器、SSD 内缓存和运行系统综合协作方式，像发布在 MICRO-52 2019 上的 DeepStore 架构。

5) 解决访存延迟和功耗的真正的存算一体（存储器颗粒本身的算法嵌入），可分为基于易失性存储 SRAM 和 DRAM 的存内计算，相变存储器、阻变存储器和浮栅器件（闪存）等非易失性存储器的新型计算型存储。

6) 目前基于闪存的存算一体技术开发者是知存科技公司。它基于 NOR Flash 构建了存算一体芯片，把乘数直接存入存储单元内，再把数值输入到闪存的阵列中，每个单元都进行乘法，最后通过一条路径求和，达到存算一体效果。乘法计算的方式是通过类似模拟电路的电流镜方式，加法的计算方式类似于并联电路电流求和。2016 和 2017 年知存科技曾做出了多个样品，最高峰值运算效率为 40TOPS/W，平均值为 10TOPS/W。

(2) 目前工业界弱化“存储墙”问题的方法是“存储层级结构 (Memory hierarchy)”，也就是计算机组成原理课程上所说的多级 Cache 机制。其核心思想是在微处理器与 DRAM 之间插入几级高速缓冲存储器也叫 Cache（通常是 SRAM 来实现）来缓冲两者之间的速度失配。存储层级结构虽然在一定程度上降低了平均延迟，但它仍是存算分离体系，并未从根本上消除“存储墙”问题。

(3) 要从根本上解决“存储墙”问题，需要从基础器件、电路、架构、系统等多个层面协同和集成创新，发展存算一体的新型计算系统，这也是获益于脑科学的研究和启发。生物层面，大脑存储大量的知识，能够快速提取并访问，而大脑的存储和计算并不是分开的，更多的是存在一定的相容性。未来的计算系统不是基于计算的 memory，而是基于 memory 的计算，需要更多的融合。存算一体技术的提出不仅仅要打破人们对传统存储和计算的认知，它还要解决这些已有的存储墙、带宽墙和功耗墙等“历史遗留”问题，实现存算之间更加低成本的“无缝对接”。存算一体架构最早于 1960 年提出，但未引起人们的重视。一方面是由于过去几十年晶体管的飞速发展使电脑性能有了令人满意的提高，另一方面是过去几十年缺乏能够实现存算一体系统的基础物理器件，即能够做运算的存储器。

(4) 近年来，深度学习这类流式算法的出现，处理的数据量非常大，而计算过程又特别简单，导致传统的冯·诺依曼计算体系彻底崩溃，这时存算一体再次被突出。高访存、高并行、低精度的人工智能和类脑计算等应用的快速出现也驱动了计算型存储/存算一体/存内计算(in-memory computing/process in memory)等非冯体系的发展，这类非冯计算架构将计算和访存融合，从体系结构上消除了访存操作，是一种极具前景的解决方式。AI 算法的访存密集，即大数据需求和计算密集也就是低精度规整运算的特征为存算一体(PIM)的实现提供了有力条件。

(5) 存算融合一方面可以从存储着手，伴随着嵌入了 ARM 核和 DRAM 的 SSD 产品的兴起，能够做运算的存储器已经出现，NAND flash、ARM 和 DRAM、控制器和内部总线实际上构成了一个计算系统，这让存储产品本身就可以做计算任务，因此也为存算一体提供了发展平台。国内有诸多初创公司在探索这个方向，尤其是由于 AI 的引入，各种数据的 Key-Value 只要直接存储在硬盘里，AI 需要的数据就可以自动完成分类，可以显著提升非关系数据库的性能。

(6) 另一方面可以从处理着手，将存储做到计算芯片上，即片上存储。IBM 设计的 Blue Gene Active Storage (BGAS) 结点就是一种“存储上的计算”系统，每个 BGAS 结点包含 32 个处理器，每个处理器通过 PCIe 接口连接 2TB 的 SLC NAND 非易失闪存介质。

(7) 人的大脑就是一个典型的存储计算系统，而仿照人脑的仿生系统也被认为是最有可能颠覆现有技术的终极发展方向。通过将电子元件编程为离散阻值状

态并将不同权重的电子元件相互卷积以建立一个类似突触和神经元的系统，即神经拟态/形态计算，又称为类脑计算。此前，国内清华大学类脑计算团队打造的“天机芯”就是异构融合类脑计算芯片，复旦大学也在单晶体管逻辑架构上有突破性的进展，为存算一体发展奠定了技术基础。美国国防高级研究计划局团队在这项研究上已经涉及了超过 1800 种混合材料，难度之高可想而知，后面架构搭建等也是商用道路上的大坎，因此，类脑的存算一体系统遥不可及。

1.3 机遇挑战

存算一体存在的机遇和挑战

(1) 阿里达摩院发布的 2020 年十大科技趋势，它认为存算一体是突破 AI 算力瓶颈的关键技术。英特尔、Arm、微软等公司均参与到该技术方向的投资，也有多家公司给出了可行的存算一体解决方案，但还没有一家公司的存算一体解决方案受到广泛的市场认可，挑战机遇并存。

(2) 在现有的芯片设计上，存算一体的挑战很多，器件方面，现有的浮栅器件存储就不适合存内计算，在芯片的工艺上，存算一体的设计和流片周期都将会很长，集成度很难上去，甚至连现有的 EDA 工具，目前尚没有支持存算一体设计的。总体来看，存算一体有 IBM、知存科技等数十家大小企业在投入和探索，它们广泛分布在存储、计算等领域里，几大技术方向也都在发展中。但是因可探索的方向很多，且没有人知道哪一种是最适合商用的方向，可以说整个市场还处在早期的百家争鸣状态。

(3) 存算一体技术最大的挑战在软件生态上，毕竟底层硬件发生了翻天覆地的变化，没有配套的软件基本上没有作用。而现实的情况是研究软件的人不愿意围绕硬件特点去设计，研究硬件的人又不了解软件。破局之法需要大公司全力投入和产学研通力合作。其次是存算一体芯片本身的设计制造上，需要器件物理，数集、模集，体系结构，DTCO 的综合层面的集成创新，需要复合型人才，这对于团队的要求就很大。

(4) 存算一体技术还会引入新的技术——模拟计算，乘加运算以模拟计算的方式完成，这与我们现在熟知的 01 数字世界完全不同，同时也是存算一体面临的最大的技术挑战，因为模拟计算本身不可避免的要引入计算误差，还可能会出现不稳定、易受干扰的弊端。

1.4 研究内容

人工智能的另一方向：基于忆阻器的存算一体技术。一旦基于忆阻器的神经形态计算芯片技术成熟，制作类似甚至超越人脑智能和能效的‘超级人工大脑’将变成现实。

本次科研实习旨在探明最有潜力的基于新型忆阻器的存内计算原理、研究和挑战，开拓解决“存储墙”问题，算力提升和 AI 需求速度失配问题的新思路。

(1) 要想破除内存瓶颈，就必须改变存算分离的冯氏架构，近年来随着生物学的发展，受到人类自身脑结构的启发，逐渐出现存算融合、类脑计算的相关研究突破。存算一体技术本身是为了提高能效，算力的提升算是意外收获。人们逐渐发现，与冯·诺依曼计算平台不同，具有大规模并行、自适应、自学习特征的人脑中，信息存储和计算没有明确的界线，都是利用神经元和突触来完成的。为了模拟神经元和突触的特性，出现了许多新原理纳米器件，其中，忆阻器因与突触的特性十分相似而备受青睐。突触可以根据前后神经元的激励来改变其权重，而忆阻器也可通过外加电压的调制来改变其电导值。忆阻器相当于一个“电子突触”，搭建起来的智能芯片具有在线学习能力，可以处理机器系统之前无法胜任的任务，可以实现数据存储的同时进行原位计算，从根本上消除了内存瓶颈。

(2) 忆阻器的阻值由激励历史决定且连续变化，呈现非易失性。它的出现为开发高能效存算一体的新型计算系统提供了新的物理基础。忆阻器具有集成密度高、操作速度快、操作功耗低、非易失等优势，被认为是存算一体基础物理器件的有力竞争者之一，为实现存算一体技术提供了切实可行的解决方案。

(3) 除了存储，忆阻器在神经形态计算芯片领域显示出更加重要的潜力，基于忆阻器的神经形态计算系统为神经网络训练提供了一种快速节能的方法。忆阻器阵列直接使用欧姆定律进行加法运算，使用电路理论所学的基尔霍夫定律进行乘法运算，在一个周期内完成矢量与矩阵的乘积累加运算(MAC)，因而能够实现并行内存内 MAC 运算，实现速度和能效的大幅提升，忆阻器本身的阻值可以用来存储数据，忆阻器可以在极短时间内完成绝大多数的计算任务。这种基于忆阻器阵列乘累加运算的核心单元的计算能效比现有 CMOS 器件提高了两个数量级，这对于具有大量乘累加运算的 AI 任务具有重要意义。忆阻器存算一体借用模拟计算，会存在误差，但是绝大部分 AI 任务都有很好的容错性，能够承受一定的

误差。可以预见，一旦基于忆阻器的神经形态计算芯片技术成熟，制作类似甚至超越人脑智能和能效的“超级人工大脑”将变成现实。

二、参加的主要活动

2.1 研究过程

1. 通过在中国知网旗下的中国学术期刊电子杂志社(CAJEPH)上查阅资料和研读文章，了解存算一体的大致发展思路，形成总的存算一体框架。其中一篇文章是来自中国电子信息产业发展研究院赛迪智库上的互联网经济杂志，作者为李雅琪、温晓君，题目为《存算一体化的发展现状、挑战与对策建议》。

(1) 首先，这篇文章阐述了存内计算的发展现状，主要有三点。

1) 存算一体技术成为缓解低效的存算调用计算的有效解决方案，技术路线分为片外存储和片内存储。片外存储是通过存储器植入计算芯片或者逻辑单元来缩短访存距离进而提高读写速度，片内存储是数模混合的，通过存储器嵌入算法权重的方式，使其具备算法功能，实现了存算一体。

2) 人工智能热潮为存算一体化的应用带来新空间。存算一体上世纪提出后，对其应用的必要性一直缺乏论证，随着基于深度学习技术的 AI 热潮兴起，存算一体化对于 AI 数据访问密集的场景的适用性得以体现，可为 AI 芯片的产业化提供助力。

3) 存算一体芯片市场广阔，国内外企业和科研院所纷纷布局。全球内存计算市场将以每年 22% 的速率持续增长，2020 年底市场体量将达到一百多亿美元。Intel 发布的傲腾 SSD 采用了片外存储技术，实现 CPU 与硬盘直接高速传输数据，平衡内存工作负载应用的性价比。国内的知存科技开发的超低功耗语音识别存算一体芯片采用片外存储，相比专用芯片，算力提升 10 到 50 倍。科研上，加州大学圣芭芭拉分校谢源教授团队设计的 PRIME 架构是通过阻变存储器/忆阻器(ReRAM)实现存算一体。

(2) 其次，该文分析了存算一体化面临的三大挑战。

1) 计算精度有限。浮栅存储器不合适计算，片内存储本质上是使用模拟计算，精度会受到低信噪比的影响，只能做定点计算，难以实现精准的浮点运算，仅适用于嵌入式人工智能等对于能效比高要求而对于精度可以容忍的场景。

2) 应用落地场景边界条件限制较多。存内计算不适合存储需求不高的场景，这会因加上一块大内存而指数级增加成本，且内存容量有上限，性价比不理想导致存内计算无法惠及更多用户和场景。AI 仍在探寻应用场景，存算一体化的落地问题仍需要结合具体应用场景具体分析。

3) 芯片开发生态亟待打造。存算一体芯片的产业化尚处于起步阶段，目前仍面临产业链上游支撑不足、下游应用不匹配等诸多困局。存算一体芯片设计上没有成熟的专用 EDA 工具辅助设计和仿真验证，芯片流片后也没有成熟工具协助测试，落地应用阶段也尚无专用软件与之匹配，这急需产学研多方围绕存算一体打造生态。

(3) 最后，该文结合时代要求给出了两点对策。

1) 把握技术路径转换机遇，加快存算一体技术研发。聚力发展高带宽内存、忆阻器新型非易失性存储器等存算一体技术，着力突破存算一体模拟信号误差、高效模数信号转换等关键技术，需要引导产学研三方合作，成立联合实验室，以加强存算一体芯片的原型设计、验证、流片等技术的工艺储备，加快其产业化进程，推动其在人工智能、物联网等领域的应用，打造我国的核心技术和长项优势。

2) 制定全面且长期的存算一体发展战略。在“十四五”和中长期集成电路研发与产业化布局中，将存算一体作为重点方向，绘制发展路线图。加大存算一体研发支持力度，在国家专项和集成电路、人工智能相关产业化专项中予以引导支持，力争实现我国在芯片领域对美国的弯道超车。强化科研院所高校、芯片设计商、晶圆厂、生产制造企业之间的联动效应和产学研深度融合，构建存算一体产业型生态。

2. 通过第一部分的研究，了解了存算一体的总体发展，其中国内知存科技公司已经计划量产存算一体芯片，为了跟紧该领域的前沿发展，进一步了解其技术方案是有必要的，于是我又在 CAJEPH 查到它们的公开资料。我研读的是北京知存科技公司刊登在微纳电子与智能制造期刊上的文章，作者为郭昕婕、王绍迪，题目为《端侧智能存算一体芯片概述》，下面是文章中的一些细节和我的感悟。

(1) 随着 AI 算法的深入研究和普及, 5G 通信与物联网技术的成熟, 可以预见, 智能万物互联 (artificial intelligent internet of things, AIoT) 时代即将来临。这从国内智能手机厂商小米、华为和 OPPO 都纷纷布局智能家居等 AIoT 领域是可以看出的。未来 AIoT 场景中设备将分为云端、边缘端和终端三类, 边缘和终端将爆发式增长。GPU/TPU 等硬件的产业化解决了云端算力问题, 但是资源受限的边缘终端设备的算力问题仍是缺失一环, 且因其对功耗成本等特殊需求, 将成为 AIoT 大规模产业化应用的核心关键。因此高效能、低成本和长待机的端侧智能芯片是 AIoT 的核心挑战。

(2) 目前应对大数据智能处理, 试图打破冯·诺依曼计算架构瓶颈的方法中, 3D 堆叠计算与增加片上缓存等方法已有很多产品应用, 有谷歌和寒武纪科技等企业在研发与应用, 但这些方法必定带来功耗与成本开销, 难以应用于边缘终端能耗与成本受限的 AIoT 设备, 且其仍然是存算分离结构, 只能缓解而无法从根本上解决冯氏计算架构瓶颈。

(3) 2015 年左右, 随着 AI、IoT 等大数据应用的兴起, 存算一体技术得到国内外学术与产业界的广泛研究与应用。Micro 2017 上, 包括英伟达、英特尔、微软、三星、ETH 与 UCSB 等都推出了存算一体系统原型。非易失性存储器技术适合存算一体芯片的高效实施, 它能实现即时开关机, 不需要额外的片外存储。但是其中基于 PCM、RRAM 与 MRAM 的存算一体芯片尚未实现产业化, 相比较而言, Nor Flash 在技术、工艺成熟度与成本方面在端侧 AIoT 领域更具有优势。

(4) 随着 AIoT 的快速发展, 用户实时延、带宽、隐私/安全性、功耗等特殊应用需求驱动边缘端侧智能应用场景的爆发, 如图 2 所示。考虑到实时产生的数据量、传输带宽以及端侧设备的能耗, 不可能所有运算都依赖云端来完成。通常边缘数据的半衰期都较低, 真正有意义的数据少, 无需全部发送到云端去处理, 大部分数据具有极高的相同模式化特征, 借助边缘终端的处理能力, 可以过滤掉大部分无用数据, 从而大幅度提高用户体验与开销。在无网环境场景下, 边缘终端处理将成为必需。由于端侧智能芯片对成本功耗的重点需求, 依赖器件与架构创新的技术路径越来越受重视。2018 年, 美国 DARPA “电子复兴计划” 研究新的计算拓扑架构用于数据存储与处理, 存算一体技术能提供可行的路径, 带来计算性能的显著提高。

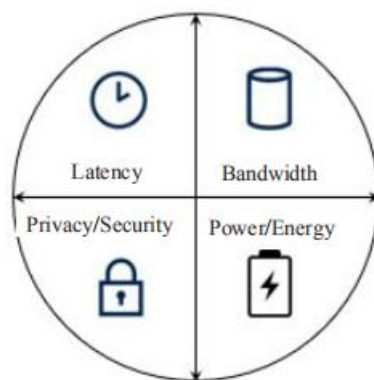


图 2 边缘端侧智能应用场景的需求特征

(5) 目前存算一体芯片研发中，比较成熟的是以 SRAM、MRAM 为代表的通用近存计算架构。如图 3 所示，这种方案采用同构众核架构，MPU 即存算一体单元，包含计算引擎(PE)、缓存、控制和输入输出等。MPU 之间通过片上网络(NoC)连接，每个 MPU 访问各自的 SRAM、MRAM 等缓存，实现高性能并行计算。典型应用是英国的 Graphcore 公司开发的芯片。

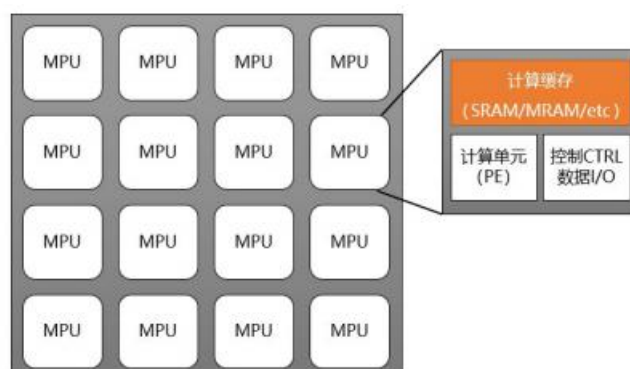


图 3 基于高速缓存的通用近存计算架构

(6) SRAM 是二值存储器，MAC 运算等价于 XNOR 累加运算，适合二值神经网络运算，存算一体的典型设计方案如图 4 所示。核心思想是把网络权重存储在 SRAM 单元中，激励信号从额外字线给入，最终利用外围电路实现 XNOR 累加运算，结果通过计数器或模拟电流输出。但其难点是实现大阵列运算的同时保证运算精度。

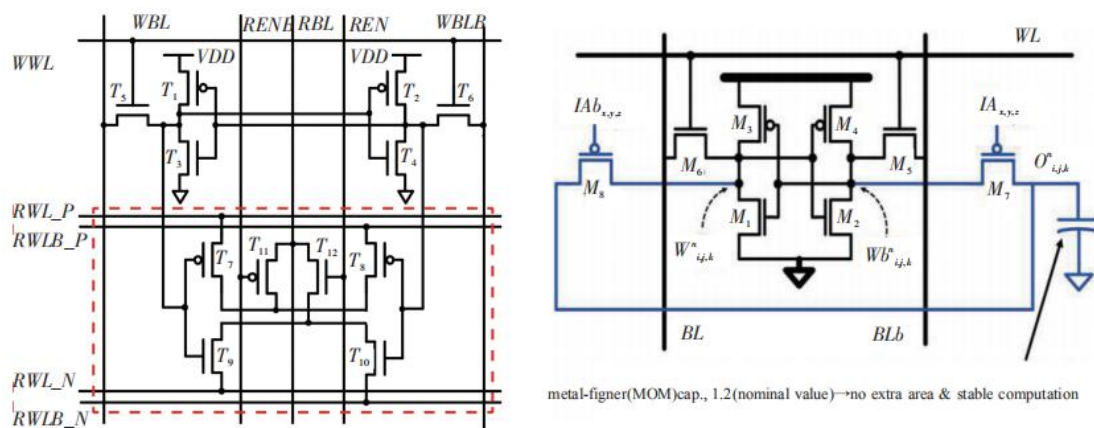


图 4 SRAM 存算一体单元 12 管（左）和 8 管（右）设计

(7) 基于 DRAM 的存算一体设计利用 DRAM 单元之间的电荷共享机制，多行单元同时被选通时，不同单元之间因为存储数据的不同会产生电荷交换共享，从而实现逻辑运算。但其问题是计算操作对数据有破坏性，需要再刷新，带来较大的功耗问题，另一难点也是保证精度。

(8) 基于 RRAM/PCM/Flash 的多值存算一体方案的原理是利用存储单元的多值特性，通过物理电气行为（电路理论所学的基尔霍夫和欧姆定律）来实现多值 MAC 运算，如图 5 所示。它将存储单元看作可变电导或电阻，存储网络权重，无需片外存储器，节约成本；同时，非易失性可实现即时开关机，减少静态功耗，延长待机时间，非常适合功耗受限的边缘终端设备。因为 RRAM/PCM 工艺成熟度不足，尚未实现产业化，但其潜力巨大；基于 Flash 的存算一体技术相对成熟，备受业界关注，有望大规模商用量产。基于 RRAM/PCM/MRAM 的二值存算一体有两种方案，其一是类似 SRAM 存算一体，但可实现存储应用和存算一体应用之间自由切换。另一种是利用存储单元实现布尔逻辑计算。

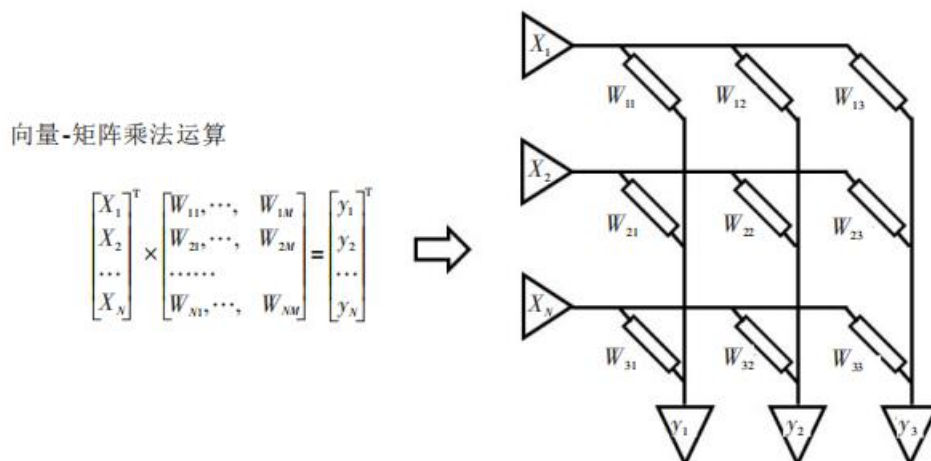


图 5 基于 RRAM/PCM/Flash 的 MAC 运算原理

(9) 存算一体芯片因其功耗成本和算力优势在 AIoT 领域有巨大应用前景，但其大规模产业化有很多挑战，比如技术层面就需要存算一体芯片涉及器件—芯片—算法—应用等多层次的跨层协同，如图 6 所示。

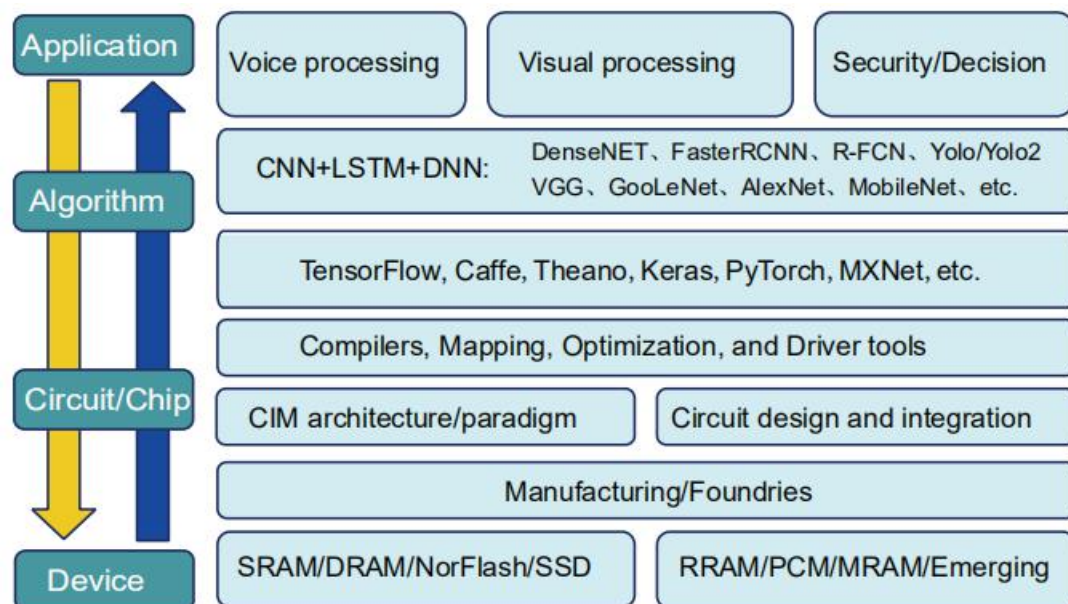


图 6 存算一体器件—芯片—算法—应用的跨层协同示意图

3. 通过第二部分对于存算一体芯片技术路径的学习，可以发现基于忆阻器等非易失性存储器的存算一体技术是最具有潜力的，因其产业布局还不够完善，也是最有市场机遇的，这激起了我对于忆阻器能否解决端侧智能芯片这个通往 AIoT 的核心挑战的强烈兴趣，因此，我在 CAJEPH 查找了忆阻器存算一体技术的相关资料。其中我阅读的一篇是长沙国防科技大学发表在微纳电子与智能制造期刊上的文章，第一作者为李锐，题目为《基于忆阻器的感存算一体技术研究进展》，下面是文中关于存算一体和感存算一体（有望突破高效模数信号转换关键技术的新思路）的介绍，中间会夹杂我的思考。

(1) 使用忆阻器实现存算一体分为数字式和模拟式。其中数字式可分为 3 种：输入输出逻辑均为电压 (V-V 型)、均为电阻 (R-R 型)、输入逻辑为电压输出逻辑为电阻 (V-R 型)。通过二值忆阻器的双极性特征，即上电极施加正电压可使器件变为高阻态 (HRS, 逻辑 0)，上电极施加负电压可使器件转变为低阻态 (LRS, 逻辑 1)，V-R 型逻辑门方案可进行数字逻辑课程中所学的实质蕴涵逻辑 (IMP)，如表 1 所示， $Y \leftarrow X_1 \text{ IMP } X_2$ ，在数字逻辑课程中，我们知道 IMP 在功能上是

完全的，故所有的 16 种布尔逻辑都可以用 V-R 逻辑门组合实现。V-V 逻辑门可以看作是一个简单的感知机网络，输出端是输入电压的加权和，输出端 $V_{com} = R_L \sum G_j (V_j - V_{com})$ ， G_j 为第 j 个忆阻器的电导， R_L 是公共节点与地之间的负载，由于输入和输出都是电压值，可实现级联操作。R-R 逻辑门也可以级联，状态逻辑，可实现“或”逻辑，通过不同结构和电压施加方式可实现其他布尔逻辑，相对于前两种逻辑门更具优势。但数字式方案功耗较高，尺寸较大，单次运算与先进的 CMOS 晶体管电路相比优势不明显，也没有验证实现复杂功能时系统整体的稳定性。可以说，能够展示数字式存算一体实际处理能力的成果几乎为空白。

表 1 V-R 逻辑门的真值表(IMP)

X1	X2	Y
0	0	1
0	1	1
1	0	0
1	1	1

(2) 模拟式存算一体计算核心是利用模拟型忆阻器交叉结构阵列实现乘累加(MAC)运算，其对比晶体管有集成度高、速度快和能效高等优点。STDP 规则是神经系统普遍存在的计算方式，对实现类脑计算，模拟人脑存算一体有重要研究价值。2016 年有研究提出基于忆阻器阵列的脉冲神经网络，使用生物神经网络的 STDP 学习规则进行训练，利用 HfO_2 多值忆阻器实现对 5 种元音字母的非监督学习，抗噪能力达 30%，因忆阻器电学特性类似神经突触，实现 STDP 有明显优势，传统电路实现同等功能需要大量晶体管。但这只是小规模阵列的应用，实现的都是单层感知机，简单数据的线性分类，与现实应用需求有较大差距。

(3) 大规模阵列发展验证了 1T1R 结构抑制泄露电流问题的可行性，为忆阻器走向实用化进一步奠定基础，但外围电路很复杂，需要集成到芯片中，开发与 CMOS 工艺兼容的阻变材料体系。忆阻器状态的波动性使其可用作物理上的随机数发生器，易于实现可靠性高和随机性强的真随机数发生器芯片，在信息安全领域有重要应用。

(4) 使用忆阻器的存算一体架构虽然进展快速,但也有很多问题。其一,忆阻器件参数均一性和可靠性离应用需求差距较大。其二,制约忆阻器规模集成的关键基础问题阵列中的串扰和阵列制备工艺还未解决。最后,从信息采集处理流程考虑,在获取外界信息后,要经过模数采用量化存储,再进行存算一体处理,其中花费的时间和功耗不可忽略,这就引出了感存算一体架构。因为忆阻器本身可以处理模拟信号,传感器采集到的模拟信号可直接送到忆阻器处理单元运算,无需经过 ADC 的转换过程,将极大地提高系统效能。

(5) 随着 AI 的飞速发展,人们希望机器人在拥有智能的同时也具有“感觉”,而触觉是一种与外界交互的基本感觉,因此开发人工触觉记忆单元极具价值。基本的触觉记忆单元是将压力感受器和记忆模块集成起来。有研究者将阻变压力传感器和阻变存储器串联起来形成触觉记忆单元,利用分压原理存储感受器信号,压力传感器利用倒金字塔结构的 AgNWs 作为压力敏感层,对低压强高敏感,阻变存储器采用 MIM 结构, SiO_2 作为阻变层,通过制备触觉记忆单元阵列用于感知并存储字母形压力图案证明了其模拟皮肤的可行性。针对触觉的压力感存算一体技术在仿生传感器、义肢修复和构建更加仿生的 AI 系统等领域具有重要应用前景,但在传感器和人工智能网络上还需要深入研究。

(6) 受人类视觉感知记忆系统的启发,将光探测器与存储器集成起来,实现对光信号的感知记忆过程,可为人类的视觉记忆仿生提供基础。与分立式的光感知记忆系统相比,简单的双端光电阻器件能够直接感应并存储光信息,不仅降低了复杂度而且更利于低功耗与大规模器件集成的实现。其中 Pd/MoO_x /ITO 双端光电阻存储器件 (ORRAM), 可以进行图像感知和记忆。构建的 8×8 简单阵列,实现了增强图像对比度和降低背景噪声的图像预处理功能,如图 7 所示。该器件是以神经信号方式感知信息并传递图形信息,具有光可调和时间依赖可塑性,更易于模拟人类的视觉感知系统。在该工作中, ORRAM 器件进行的是紫外光的感知、记忆与预处理,但是实际上人类视觉感知系统需要处理的光信息却要复杂的多。随着人工视觉技术的发展,能够直接响应光刺激并对视觉信息和感知数据进行存储和实时处理的光阻型随机存储器和光电子突触型器件逐渐成为未来人工视觉研究的热点

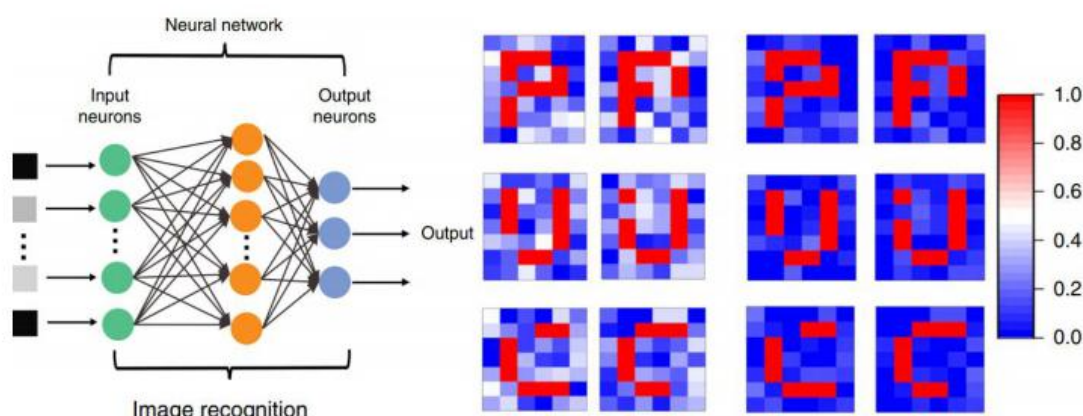


图 7 两端光阻忆阻器人工神经网络的图像预处理功能

(7) 受人类嗅觉感知记忆系统的启发，将气体探测器和存储器集成在一起，可以模拟人类对气体信息的感知和记忆过程，实现人体嗅觉的仿生。有研究将 100 多万忆阻器与 200 多万 CNT 晶体管集成在一个芯片上，集气体感知、存储和计算为一体，构建了 3D 集成纳米系统。通过感知、存储、计算及数据接口四层堆叠构建了 3D 结构，每个单元包括两个 CNT 晶体管、一个 RRAM 以及一个 Si 基晶体管，整个芯片由超过一百万个重复单元构成，实现了 7 种气体氛围的感知、数据存储以及数据原位分类识别。该工作中采用了 CNT 晶体管，有利于节能高密度数据存储的实现，是迄今为止最复杂的纳米电子系统。但是，在此项工作中有多种气体（酒精、白酒、伏特加）响应差距比较小，需要放大器进行信号放大才能使计算层足以进行计算分类识别，这无疑会增加结构的复杂度。

(8) 目前，感存算一体架构的发展处于起步阶段，有很大领域拓展性，器件大多处于单元器件或分立式阵列，尚未互连，功能简单无法发挥集成阵列高效并行运算的优势。但忆阻器存算一体器件相关研究相对成熟，通过解决忆阻器存算一体器件与微纳传感器的高密度三维集成工艺、模拟信号匹配等关键技术后，基于忆阻器的感存算一体技术将进入快速发展阶段。目前的感存算一体器件的研究工作多是基于模拟某一种感官和简单处理，如触觉、视觉、嗅觉等，处理能力十分有限，与人类的感知记忆系统相差甚远，发展多感知融合和多元化处理功能的器件体系，减小与应用需求的差距，是面向未来应用的热门方向。通过感存算一体器件对感知信息进行简单预处理后还需配合搭建系统级架构进行更为复杂的

信息处理,才能够具有接近实际应用的处理能力,这方面的研究工作还比较初级,未来需要深入研究感存算一体信息处理架构、任务调度与分工协作等策略。

4. 通过前面三部分,我已经基本上清楚了基于忆阻器的存算一体甚至是感存算一体技术的多种多样的实现思路,有布尔逻辑也有直接模拟计算的工作原理,为了进一步了解人工智能时代芯片设计尤其是异构非冯架构芯片设计的新思路和新技术,我在知网上查找并阅读了相关文章,主要是想了解是否有类似存算一体而在之前尚未分析到的技术,其中一篇是北京大学深圳研究院发表在微纳电子与智能制造期刊上的文章,第一作者为任源,题目为《人工智能芯片的研究进展》,下面是具体细节。

(1) 自 20 世纪 50 年代,人类对人工智能技术的探索从未停止,尽管有多次寒冬,2006 年, Hinton 首次证明了大规模深度神经网络学习的可能性,至此 AI 芯片开启序幕。2008 年,英伟达推出 Tegra 芯片,是最早的可用于 AI 领域的 GPU,如今成为英伟达最重要的 AI 芯片之一。2010 年, IBM 首次发布类脑芯片原型模拟大脑结构,原型具有感知认知和并行计算能力。2014 年,中科院陈天石博士开启人工智能加速专用芯片(ASIC)的研究领域,前一年, GPU 才广泛应用于 AI 领域。2015 年,国际 FPGA 大会上 Jason Cong 发表了 FPGA 加速 DNN 算法的论文,使得 FPGAs 迅速大火。一年后,谷歌发表 TensorFlow 框架的 TPU 芯片,采用该芯片的 AlphaGo 击败人类世界冠军棋手李世石,同年,寒武纪研发出 DIANNAO, FPGA 芯片在云计算平台得到广泛应用。之后,华为海思的麒麟 970 成为首个手机 AI 芯片;清华大学魏少军教授团队推出算力和能效具有国际水平的 Thinker AI 芯片。

(2) AI 芯片目前有两种发展方向,一种是延续经典的冯氏架构,以加速计算能力为目标,分为 GPU、FPGA、ASIC。另一种是颠覆传统的冯氏架构,采用基于类脑神经结构的神经拟态芯片来提高算力。

(3) GPU 速度快、芯片编程灵活简单,具有高并行结构,相比 CPU 有更多的 ALU,在处理图形数据和复杂算法方面的效率提升 1 到 3 个数量级。GPU 有更强大的浮点运算能力,可以缓解深度学习算法的训练难题,在算法训练上非常高效,但在深度学习算法的推断中对于单项输入进行处理时,并行优势不能完全发挥出来。

(4) 与 GPU 不同, FPGA 同时拥有进行数据并行和任务并行计算的能力, 适用于以硬件流水线方式处理一条数据, 且整数运算性能更高, 因此常用于深度学习算法中的推断阶段。功耗方面, 从体系结构而言, FPGA 具有天生的优势, 每个逻辑单元的功能在重编程(即烧入)时就已经确定, 不需要指令, 无需共享内存, 从而可以极大地降低单位执行的功耗, 提高整体的能耗比。缺点是用硬件的配置实现软件算法, 在实现复杂算法方面有一定的难度。

(5) 由于 GPU、FPGA 通用芯片设计初衷并非专门针对深度学习, 因而天然存在性能、功耗等方面的局限性。FPGA 大量资源用于可配置的片上路由与连线, 计算资源占比较低, 速度和功耗相对 ASIC 存在不小差距, 价格也更贵, 随着 AI 应用规模扩大, ASIC 产业环境成熟, ASIC 逐步体现出优势, 适用于高性能、低功耗的移动应用端, 比如谷歌的 TPU、地平线的 BPU 都属于 ASIC 芯片, 缺点是电路设计需要定制, 功能难以扩展, 开发周期较长。

(6) 神经形态计算近年来成为从根本上克服冯·诺依曼瓶颈的最有吸引力的替代方案, 其最终目标是开发神经形态硬件加速器, 模拟高效生物信息处理, 以弥合网络和真实大脑之间的效率差距, 这被认为是下一代(强)人工智能的主要驱动力。它从结构层面去逼近大脑, 可分为 2 个层次, 一是神经网络层面, 相应的是神经拟态架构和处理器, 如 IBM Truenorth, 这种 AI 芯片将定制化的数字处理内核当作神经元, 内存作为突触, 完全集成在一起, 同时神经元之间可相互通信; 二是神经元与突触层面, 相应的是元器件层面的创新, 比如 IBM 苏黎世制造的全球首个人造纳米尺度的随机相变神经元, 可实现高速无监督学习。目前先进的大规模神经形态系统如图 8 所示, 和人脑规模(10^{10} 个神经元)相差甚远, 为了达到在人脑中规模, 应将多个神经拟态芯片集成在电路板上, 设计目的不再仅仅局限于加速深度学习算法, 还应在基本结构甚至器件层面上改变设计, 开发出新的类脑计算架构, 比如采用忆阻器和 ReRAM 等新器件来提高存储密度。这类芯片技术还未成熟, 意味着有很大的市场机遇, 长期来看类脑芯片可能会带来计算机体系结构的革命。

项目	神经元数量	突触数量	功耗	大小
Neurogrid ^[23]	1 M	8 B	3 W	16-chip board
IBM TrueNorth ^[24]	1 M	256 M	0.07 W	1 chip
IBM TrueNorth NS16e ^[25]	16 M	4 G	8.88 W	16-chip board
SpiNNaker ^[26]	250 K	80 M	48 W	48-chip board
FACETS ^[27]	180 K	40 M	N/A	5 mm ² -chip wafer (200 mm)

图 8 最先进的大规模神经形态系统

(7) AI 芯片核心是神经网络算法的实现，深度学习分为 CNN 和 RNN，但其需要反复训练和大量推理运算，算力要求高，时间长，功耗极大，比如 Alpha Go 功耗 1MW 是人脑能耗(20W)的 5 万倍，差距甚远。但 AI 应用的潜力在物联网和边缘计算这些能耗受限的领域，故脉冲神经网络(SNN)被提出，极具科学意义而被誉为第三代人工神经网络。SNN 具备更多独特的仿脑特性，如信息的脉冲表示，更接近生物神经网络中的学习和记忆机制。下图 9 是世界著名高校与半导体公司研究开发的基于 SNN 的神经拟态电路，可以看出其比基于传统 DNN 的硬件加速器具有更高的能量效率。但是，现有的大多数神经拟态芯片的计算与存储单元在局部依然是分离的，在用于神经元的 CMOS 逻辑电路和用于突触的 SRAM 电路之间依然存在局部的存储壁垒和能量效率问题，所以还不是真正意义上的非冯·诺依曼体系结构。不过最新的具有三维堆叠能力的非易失性存储器(NVM)技术或存内计算技术(in-memory computing)有望解决这一问题。

	Spinnaker ^[28]	Neurogrid ^[29]	TrueNorth ^[30]	Loihi ^[31]	PCRAM Neuromorphic chip ^[32]	Tianji ^[33]
国家	英国	美国	美国	美国	美国	中国
科研单位	纽卡斯尔大学	斯坦福大学	IBM	Inter	IBM	清华大学
硬件实现	基于多核 CPU 实现的 SNN	混合模拟 CMOS 电路实现的 SNN	数字 CMOS 电路实现的 SNN	数字 CMOS 电路实现的 SNN	基于相变存储器的 SNN	数字 CMOS 电路实现的 SNN
电子突触	SDRAM	模拟 CMOS 电路	SRAM	SRAM	PCRAM	未说明
电子神经元	ARM CPU 核	模拟 CMOS 电路	数字 CMOS 电路 (LIF 动态功能)	数字 CMOS 电路 (IF 动态功能)	数字 CMOS 电路 (LIF 动态功能)	数字 CMOS 电路 (LIF 动态功能)
在线学习	无	无	无	STDP 学习规则	STDP 学习规则	未说明
能耗	8 nJ/SynEvent	31.2 nJ/SynEvent	26 nJ/SynEvent	23.6 nJ/SynEvent	0.9 pJ/bit	未说明

图 9 脉冲神经拟态芯片国内外研究现状

通过这四部分的学习，可以发现类脑芯片是最具潜力，也是最有市场机遇，还有可能驱动强人工智能时代到来和实现计算机体系结构革命的人工智能芯片。但是现阶段类脑芯片相比人脑真实的规模至少有 4 个数量级差异，文中因此提出可以采用忆阻器等新器件提高存储密度，在基本结构和器件层面改变设计，开发新的类脑计算体系。最后文中指出，现有的大多数神经拟态芯片仍不是真正意义上的非冯架构，还存在存储壁垒和能量效率问题，并未实现存算一体，但具有三

维堆叠能力的忆阻器等非易失性存储器(NVM)存算一体技术有望解决 AI 芯片问题。这些都说明了忆阻器对于实现类脑存算一体 AI 芯片的重要性。

今年,清华大学钱鹤、吴华强教授团队在 Nature 在线发表了题为“Fully hardware-implemented memristor convolutional neural network”的研究论文,报道了首个完全基于忆阻器阵列芯片卷积网络的完整硬件实现。该成果所研发的基于多个忆阻器阵列的存算一体系统,在处理 CNN 时的能效比 GPU 高两个数量级,大幅提升了计算设备的算力,成功实现了以更小的功耗和更低的硬件成本完成复杂的计算,文中基于忆阻器的五层 CNN 在 MNIST 手写数字识别任务中实现了 96.19%的准确率,为大幅提升 CNN 效率提供了可行的解决方案。

之前没有一个 CNN 是在完全基于忆阻器的硬件系统中实现的,造成这种困境的原因是缺乏针对基于忆阻器的 CNN (mCNN)实现的高效解决方案:首先,mCNN 通常收益低,忆阻器交叉阵列不均匀;其次,由于设备缺陷(如变化、电导率偏移和设备状态锁定),mCNN 很难达到可媲美软件实现结果的性能;第三,CNN 中的核心卷积操作非常耗时,因为它需要沿着不同的输入块滑动,这通常是一个序列过程,会导致忆阻器卷积器和忆阻器阵列之间出现速度不匹配的情况,后者旨在完成全连接向量矩阵乘法。

该研究提出了一种基于忆阻器的灵活计算架构,适用于神经网络,从而克服了忆阻器设备的易变性和不完美特性这两个神经形态计算应用的主要瓶颈。

2.2 研究结论

我沿着解决大数据人工智能时代的算力功耗成本需求(尤其是边缘终端设备)与芯片升级之间的矛盾为目标,试图找到打破传统的冯·诺依曼瓶颈的新的 AI 芯片设计思路,在查阅文献和研究学习的过程中,找到了多种缓解和解决存储墙的技术思路,其中存算一体非冯体系结构设计的芯片既能作为高能效、低成本和长待机的端侧智能芯片解决通往智能万物互联时代的核心挑战,还能成为模拟高效生物信息处理的类脑 AI 芯片,推动强人工智能时代的到来,甚至可能会带来计算机体系结构的革命。而真正的存算一体技术中以基于忆阻器等非易失性存储器(NVM)的(感)存算一体最具潜力,其完全依靠电路理论所学的物理电气行为进行存内计算,省去了许多模数转换和存储过程,大幅提升了能效和算力,已经领

先 GPU 等传统冯氏架构芯片几个数量级，在物联网和边缘计算领域具有广阔的应用前景。

2.3 展望工作

(1) 在明年的未来论坛青创联盟线上学术研讨中，来自清华大学的吴华强教授提出：当思考未来计算的时候，量子计算、光计算是向物理找答案，类脑计算、存算一体是向生物找答案，也就是向大脑找答案。虽然类脑芯片在已经有很多研发成果，但是其相比人脑的规模和能耗都有很大的距离，AlphaGo 用了 176 个 GPU、1202 个 CPU，功耗是 150000W。而我们大脑体积大概 1.2L，有 10^{11} 个神经元， 10^{15} 个突触，思考的时候功耗是 20W。大脑的功耗这么少，还这么聪明，这或许启发着类脑芯片进一步发展的方向，虽然也有类似大脑的脉冲编码方式的基于 SNN 的神经拟态芯片出现，但其在集成度和能耗控制上还可以开展进一步的研究。

(2) 基于忆阻器的感存算一体系统在计算速度、计能效等方面具有非常诱人的前景，但是未来进一步规模产业化还需要基础器件、架构、电路系统、应用的跨层协同创新。根据具体的物联网需求场景开发出基于具体存储器件实现的存算一体端侧智能芯片，并做到性能与成本的最优化，这也是未来需要努力研究并解决的问题。

三、实习总结

(1) 通过这段长时间的科研学习，我基本理清了现阶段各种各样的存算一体技术路线，找到了在摩尔定律几乎失效，AI 算力急剧膨胀的大数据时代，芯片尤其是 AI 芯片设计的一种很有潜力的思路。在数据密集访问而精度要求不高的 AI 算法中，基于忆阻器的存算一体芯片已被验证在性能上超过了 GPU 等传统冯氏架构芯片，且如果工艺成熟，串扰、不稳定等问题被解决，其潜力更加巨大，可能带来整个计算体系结构的革命性进步。


(2) 计算机的“大脑”——芯片设计如何同时满足功耗和算力需求，或许向奇妙无穷的人脑寻找答案，人脑体积小、算力大、能耗低，并且信息存储不分离，人脑作为鬼斧神工的自然进化结晶，何尝不是人类发明的信息处理机器的一种终

极解决方案，而忆阻器或许可以成为一种方案，可能给普通民众生活带来的最大变化就是，如果把用忆阻器技术开发出的人工智能芯片应用在手机中，芯片功耗会大大降低，手机充一次电就可以用两天。

(3) 随着 5G 通信与物联网技术的成熟，未来智能万物互联时代的端侧智能芯片需求也越来越大，未来芯片这个发展方向也是强调功耗和待机时间，算力可以交给云端实现，毕竟扫地机器人等家居物联网产品要是工作一会就没电，消费者指定不会买账，本来就是为了省心做其他事情，结果全用去充电了。

(4) 在阅读论文中的各种存算一体方案的实现过程中，我逐渐意识到集成创新的重要性，比方说，清华大学施路平教授开发的“天机”类脑芯片登陆了顶级学术期刊《nature》杂志封面，搭载该芯片的自动驾驶自行车是可以自动控制平衡、并在操场上对目标人物进行识别、跟随、自动避障，这就是一个综合器件、架构、电路系统、应用的集成创新成果，意义远在点创新之上。未来最需要的人才是复合型人才，无论是大学教授还是企业高管，真正有意义的大成果都来源于集成创新。未来这也是我继续硕士研究的努力方向，这次科研实习为我进行软硬件协同设计的科研奠定了基础，还让我找到了做科研的思路和方法。

华中科技大学本科生生产实习报告评审表

姓名	李蓉	学号	U201714703	指导教师	胡迪青
院（系）专业	计算机科学与技术				
指导教师评语					
1. 学生科研实习表现情况。					
2. 不足及建议。					
<p>李蓉同学刻苦钻研，思维开阔，善于跨层学习与应用。对于新的问题能够积极主动寻找解决思路，基础知识扎实牢固，严谨认真对待研究任务，善于从多个方面思考，有序做好分析总结工作。不足的地方在于编程实践经验不够丰富，这方面的算法储备需要加强</p>					
指导教师（签名）： 					
2020 年 12 月 3 日					