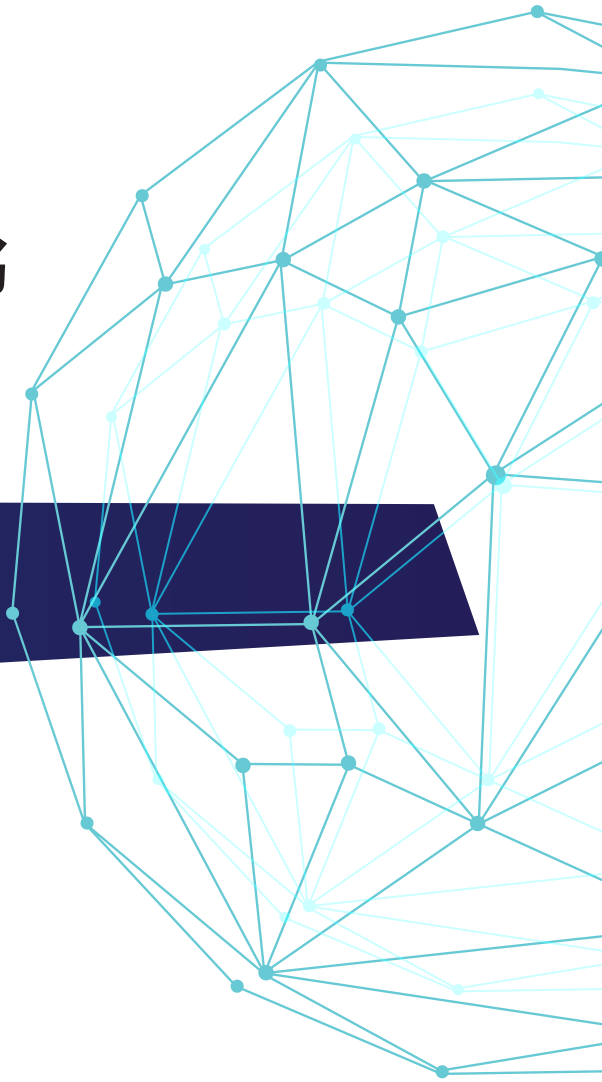
人工智能芯片技术白皮书
（2018）

White Paper on AI Chip Technologies



清華大學 Tsinghua University

北京未来芯片技术高精尖创新中心
BEIJING INNOVATION CENTER FOR FUTURE CHIPS

目录

北京未来芯片技术高精尖创新中心

人工智能芯片技术白皮书（2018）

编写委员会主席

尤　政	中国工程院院士	清华大学
魏少军	IEEE Fellow	清华大学

编写委员会副主席

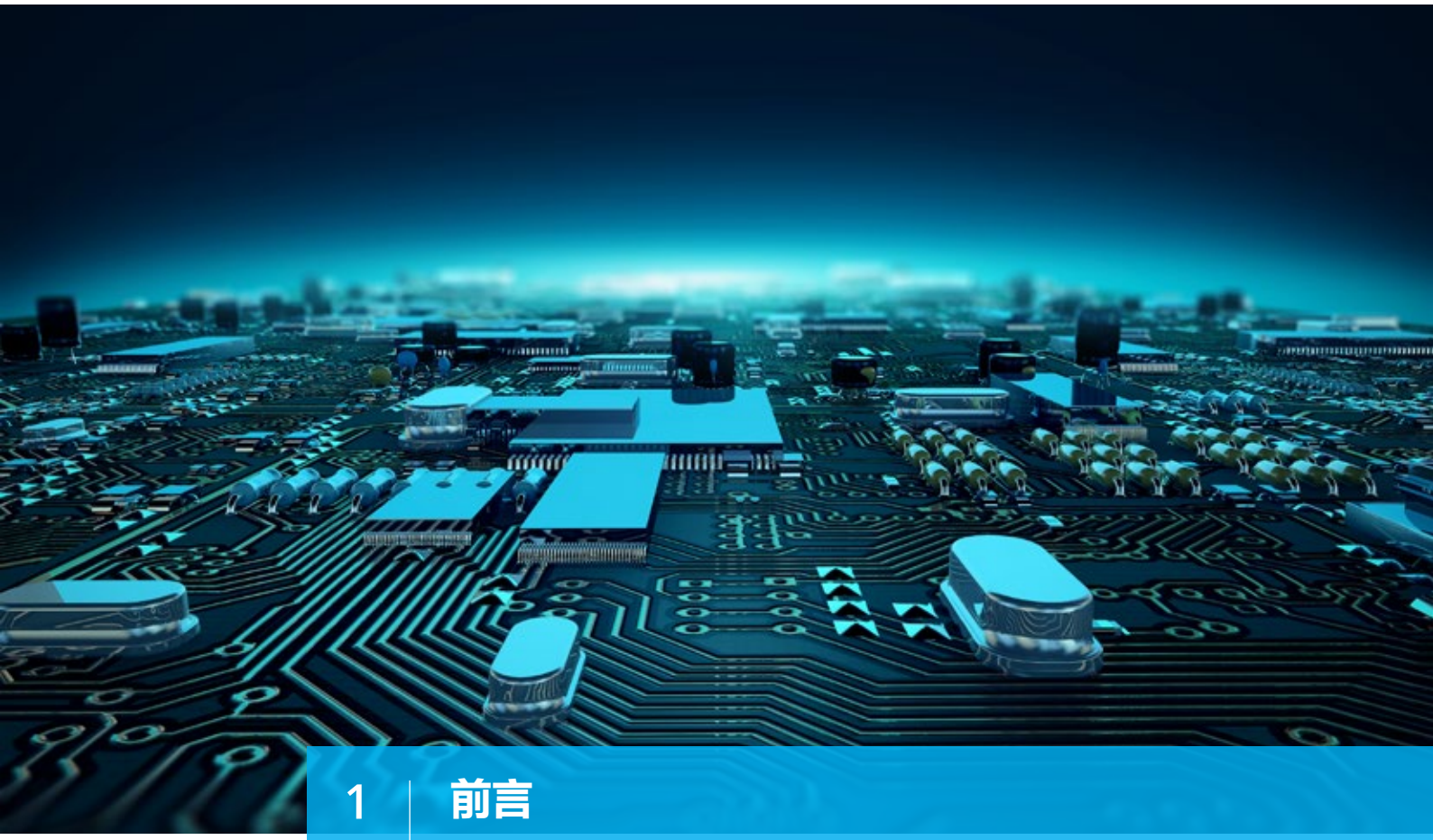
吴华强		清华大学
邓　宁		清华大学

编写委员会成员（按姓氏笔划排序）

尹首一		清华大学
王　玲		清华大学
朱　晶		北京半导体行业协会
刘勇攀		清华大学
杨建华		马萨诸塞大学
杨美基	IEEE Fellow	香港应用科技研究院
吴臻志		清华大学
汪　玉		清华大学
张孟凡	IEEE Fellow	台湾新竹清华大学
陈　安		半导体研究联盟
陈怡然	IEEE Fellow	杜克大学
郑光廷	IEEE Fellow	香港科技大学
胡晓波	IEEE Fellow	圣母大学
唐　杉		新思科技
黄汉森	IEEE Fellow	斯坦福大学
凡德·斯皮格尔	IEEE Fellow	宾夕法尼亚大学
谢　源	IEEE Fellow	加利福尼亚大学圣巴巴拉分校

中心介绍

北京未来芯片技术高精尖创新中心成立于 2015 年 10 月，是北京市教委首批认定的“北京高等学校高精尖创新中心”之一。中心充分发挥清华大学的学科、科研和人才优势，联合校内多个院系资源，组建了微电子、光电子及柔性集成、微系统、类脑计算、基础前沿、综合应用六个分中心以及微纳技术支撑平台。中心主任由清华大学副校长尤政院士担任。中心以服务国家创新驱动发展战略和北京市全国科技创新中心建设为出发点，致力于打造国家高层次人才梯队、全球开放型微纳技术支撑平台，聚焦具有颠覆性创新的关键器件、芯片及微系统技术，推动未来芯片产业实现跨越式发展。



1 前言

1.1 背景与意义

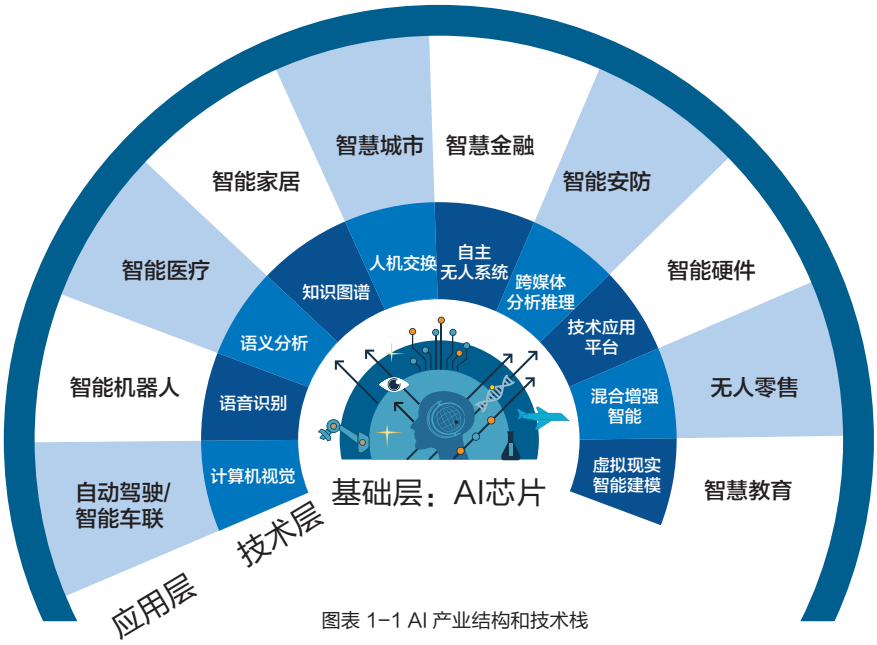
人工智能 (Artificial Intelligence，英文缩写为 AI)，是研究、开发用于模拟、延伸和扩展人类智能的理论、方法、技术及应用系统的一门科学技术。人工智能的本质是对人类思维过程的模拟。从 1956 年正式提出“人工智能”概念算起，在半个多世纪的发展历程中，人们一直在这一领域进行长期的科学探索和技术攻坚，试图了解智能的实质。和任何曾经处于发展过程中的新兴学科一样，人工智能早期发展并非一帆风顺，它曾受到多方质疑，不断经历起伏。近些年，大数据的积聚、理论算法的革新、计算能力的提升及网络设施的演进，使得持续积累了半个多世纪的人工智能产业又一次迎来革命性的进步，人工智能的研究和应用进入全新的发展阶段。

当前，人工智能正逐渐发展为新一代通用技术，加快与经济社会各领域渗透融合，已在医疗、金融、安防、教育、交通、物流等多个领域实现新业态、新模式和新产品的突破式应用，带动生产流程、产品、信息消费和服务业的智能化、高附加值转型发展。人工智能已处于新科技革命和产业变革的核心前沿，成为推动经济社会发展的新引擎。

实际上，人工智能产业得以快速发展，无论是算法的实现、海量数据的获取和存储还是计算能力的体现都离不开目前唯一的物理基础——芯片。可以说，“无芯片不 AI”，能否开发出具有超高运算能力、符合市场需求的芯片，已成为人工智能领域可持续发展的重要因素。

尽管全球人工智能产业还处于初期发展阶段，但随着政府和产业界的积极推动，人工智能技术在大规模产业化应用方面突飞猛进，在算法和芯片等人工智能基础技术层面积累了强大的技术创新，这些成果未必能即时商业化，但对未来科技的影响深远。

为了更好地厘清当前 AI 芯片领域的发展态势，进一步明确 AI 芯片在新技术形势下的路线框架、关键环节及应用前景，北京未来芯片技术高精尖创新中心根据学术界和工业界的最新实践，邀请国内外 AI 芯片领域的顶尖研究力量，共同开展《人工智能芯片技术白皮书》的编制工作。

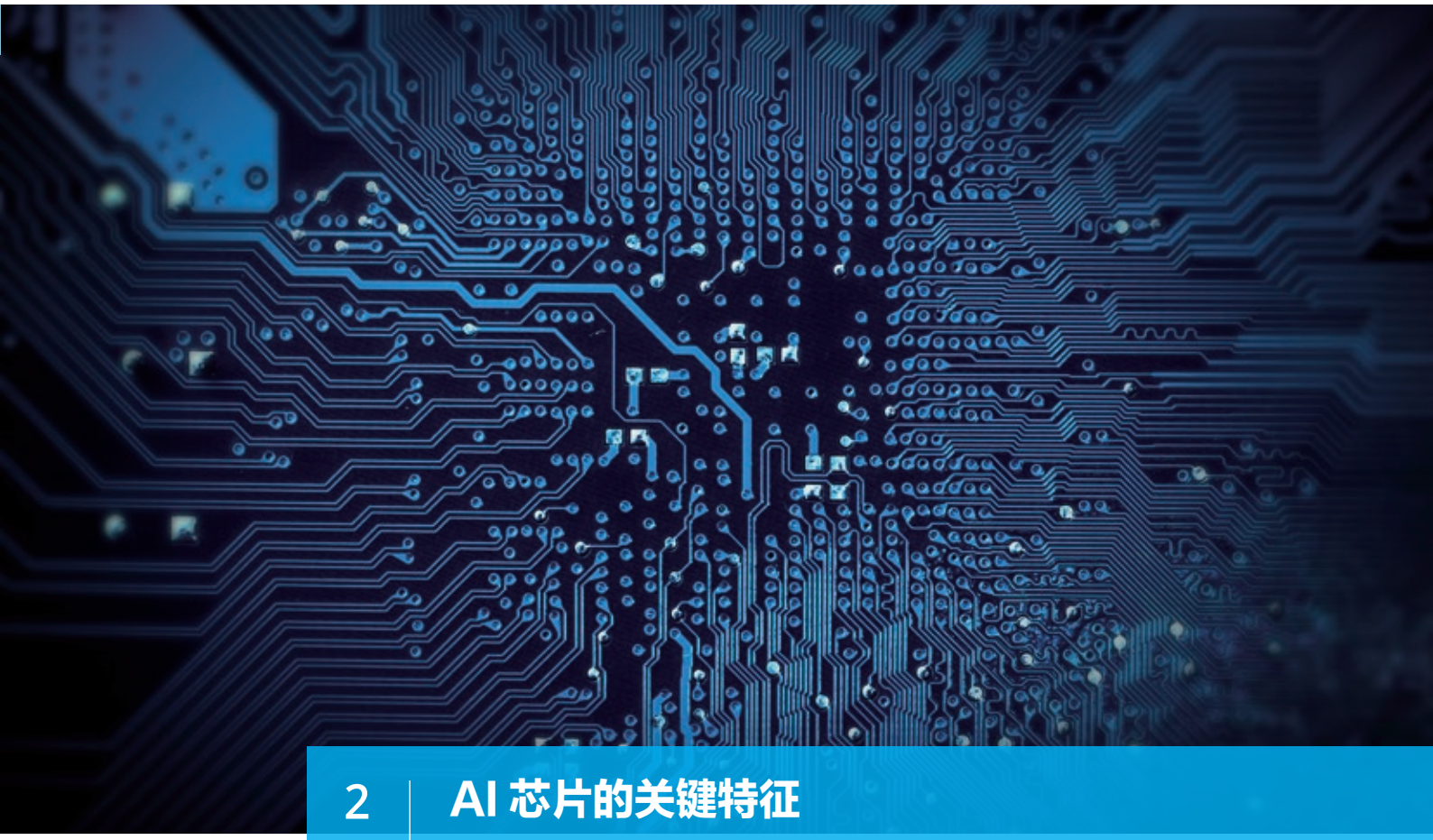


图表 1-1 AI 产业结构和技术栈

1.2 内容与目的

本文主要包括九方面内容：第 1 章为发展 AI 芯片产业的战略意义以及白皮书基本内容概述。第 2 章综述了 AI 芯片的技术背景，从多个维度提出了满足不同场景条件下 AI 芯片和硬件平台的关键特征。第 3 章介绍近几年的 AI 芯片在云侧、边缘和终端设备等不同场景中的发展状况，总结了云侧和边缘设备需要解决的不同问题，以及云侧和边缘设备如何协作支撑 AI 应用。第 4 章在 CMOS 工艺特征尺寸逐渐逼近极限的大背景下，结合 AI 芯片面临的架构挑战，分析 AI 芯片的技术趋势。第 5 章讨论了建立在当前 CMOS 技术集成上的云端和终端 AI 芯片架构创新。第 6 章主要介绍对 AI 芯片至关重要的存储技术，包括传统存储技术的改进和基于新兴非易失存储（NVM）的存储器解决方案。第 7 章重点讨论在工艺、器件、电路和存储器方面的前沿研究工作，和以此为基础的存内计算、生物神经网络等新技术趋势。第 8 章介绍神经形态计算技术和芯片的算法、模型以及关键技术特征，并分析该技术面临的机遇和挑战。第 9 章主要讨论 AI 芯片的基准测试和技术路线图的相关问题。第 10 章展望 AI 芯片的未来。

在人工智能热潮面前，本文一方面希望与全球学术和工业界分享 AI 芯片领域的创新成果；另一方面希望通过对 AI 芯片的技术认知和需求的深入洞察，帮助相关人士更加清晰地了解 AI 芯片所处的产业地位、发展机遇与需求现状，通过对 AI 芯片产业现状及各种技术路线的梳理，增进对未来风险的预判。目前人工智能技术整体发展仍处于初级阶段，未来还有很多技术和商业层面的挑战。我们要去除在产业发展过程中一窝蜂“逐热而上”的虚火，在充满信心、怀抱希望的同时，保持冷静和客观，推动 AI 芯片产业可持续发展。



2 AI 芯片的关键特征

2.1 技术总述

目前，关于 AI 芯片的定义并没有一个严格和公认的标准。比较宽泛的看法是，面向人工智能应用的芯片都可以称为 AI 芯片。时下，一些基于传统计算架构的芯片和各种软硬件加速方案相结合，在一些 AI 应用场景下取得了巨大成功。但由于需求的多样性，很难有任何单一的设计和方法能够很好地适用于各类情况。因此，学界和业界涌现出多种专门针对人工智能应用的新颖设计和方法，覆盖了从半导体材料、器件、电路到体系结构的各个层次。

本文探讨的 AI 芯片主要包括三类，一是经过软硬件优化可以高效支持 AI 应用的通用芯片，例如 GPU；二是侧重加速机器学习（尤其是神经网络、深度学习）算法的芯片，这也是目前 AI 芯片中最多的形式；三是受生物脑启发设计的神经形态计算芯片。

AI 技术是多层面的，贯穿了应用、算法机理、芯片、工具链、器件、工艺和材料等技术层级。各个层级环环紧扣形成 AI 的技术链，如图表 2-1 所示。AI 芯片本身处于整个链条的中部，向上为应用和算法提供高效支持，向下对器件和电路、工艺和材料提出需求。一方面，应用和算法的快速发展，尤其是深度学习、卷积神经网络对 AI 芯片提出了 2-3 个数量级的性能优化需求，引发了近年来 AI 芯片研发的热潮。另一方

图表 2-1 AI 芯片相关技术概览

面，新型材料、工艺和器件的迅速发展，例如 3D 堆叠内存，工艺演进等也为 AI 芯片提供了显著提升性能和降低功耗的可行性，这个推动力来源于基础研究的突破。总体而言，这两类动力共同促进了 AI 芯片技术近年来的快速发展。

2.2 新型计算范式

AI 计算既不脱离传统计算，也具有新的计算特质，包括：

1. 处理的内容往往是非结构化数据，例如视频、图像及语音等，这类数据很难通过预编程的方法得到满意的结果。因此，需要通过样本训练、拟合及环境交互等方式，利用大量数据来训练模型，再用训练好的模型处理数据。

2. 处理的过程通常需要很大的计算量，基本的计算主要是线性代数运算，典型的如张量处理，而控制流程则相对简单。对于这类运算，大规模并行计算硬件较传统通用处理器更为适合。

3. 处理的过程参数量大，需要巨大的存储容量，高带宽、低延时的访存能力，以及计算单元和存储器件间丰富且灵活的连接。数据本地化特征较强，适合数据复用和近内存计算。

2.3 训练和推断

AI 系统通常涉及训练（Training）和推断（Inference）过程。简单来说，训练过程是指在已有数据中学习，获得某些能力的过程；而推断过程则是指对新的数据，使用这些能力完成特定任务（比如分类、识别等）。对神经网络而言，训练过程就是通过不断更新网络参数，使推断（或者预测）误差最小化的过程；推断过程则是直接将数据输入神经网络并评估结果的正向计算过程。虽然训练和推断有很多类似的基本运算，都需要具有大量的并行处理，高内存带宽和低延迟操作，但是两者在计算和存储资源方面的需求方面存在显著的差异。

训练：首先，对于训练来说，计算精度非常重要，因为它直接影响推断的准确度。支持训练的硬件必须支持具有较长字长的浮点数或定点数。其次，训练中通常同时包括正向和反向的计算过程，需要多次迭代，计算量要求非常高。这就需要支持训练的芯片不仅要具有强大的单芯片计算能力，还要具备很好的扩展性，可以通过多芯片系统提供更强大的计算能力。再次，训练过程，特别是离线训练，必须处理大量的数据（高达 10^{15} 到 10^{18} 字节），因此，它对内存数量、访问内存的带宽和内存管理方法的要求都非常高。第四，由于训练需要更新（写入）和使用（读取）神经网络中的参数（权重），因而需要更复杂的数据同步技术。最后，重要参数的频繁写入也要求存储器能支持更快速的写入（特别是对于在线训练），这对于一些存储器技术来说是很大的挑战。

推断：对推断来说，运算和存储的需求都远远低于训练。但由于推断的应用场景多种多样，部署在从云到端的各种设备，如数据中心、自动驾驶汽车、智慧家庭和 IoT 设备等，其需求和约束呈现出多样化的特点。对于多数应用来说，速度、能效、安全和硬件成本等是最重要的考虑因素，而模型的准确度和数据精度则可以依具体情况适当降低。

虽然目前大部分机器学习方法都可以比较清晰地划分为训练和推断的过程，但还有一些领域，比如增强学习（Reinforcement Learning）和在线学习（On-line Learning）则处于持续学习和改进模型的进程中。因此，在未来的 AI 应用当中，训练（学习）和推断在更多场景下会是交织在一起的。

2.4 大数据处理能力

人工智能的发展高度依赖海量的数据。满足高效能机器学习的数据处理要求是 AI 芯片需要考虑的最重要因素。一个无法回避的现实是，运算单元与内存之间的性能差距越来越大，内存子系统成为芯片整体处理能力提高的障碍，也就是通常所说的“内存墙”。人工智能工作负载多是数据密集型，需要大量的存储和各层次存储器间的数据搬移，导致“内存墙”问题更加突出。为了弥补计算单元和存储器之间的差距，学术界和工业界正在两个方向上进行探索：（1）富内存的处理单元。增加片上存储器的容量并使其更靠近计算单元，使得数据计算单元和内存之间的数据移动成本（时间和功耗）大大减少。（2）具备计算能力的新型存储器。直接在存储器内部（或更近）实现计算。这种方法也被称为存内计算（Process-in-Memory, PIM）或近数据计算（Near Data Computing, NDC）。

2.5 数据精度

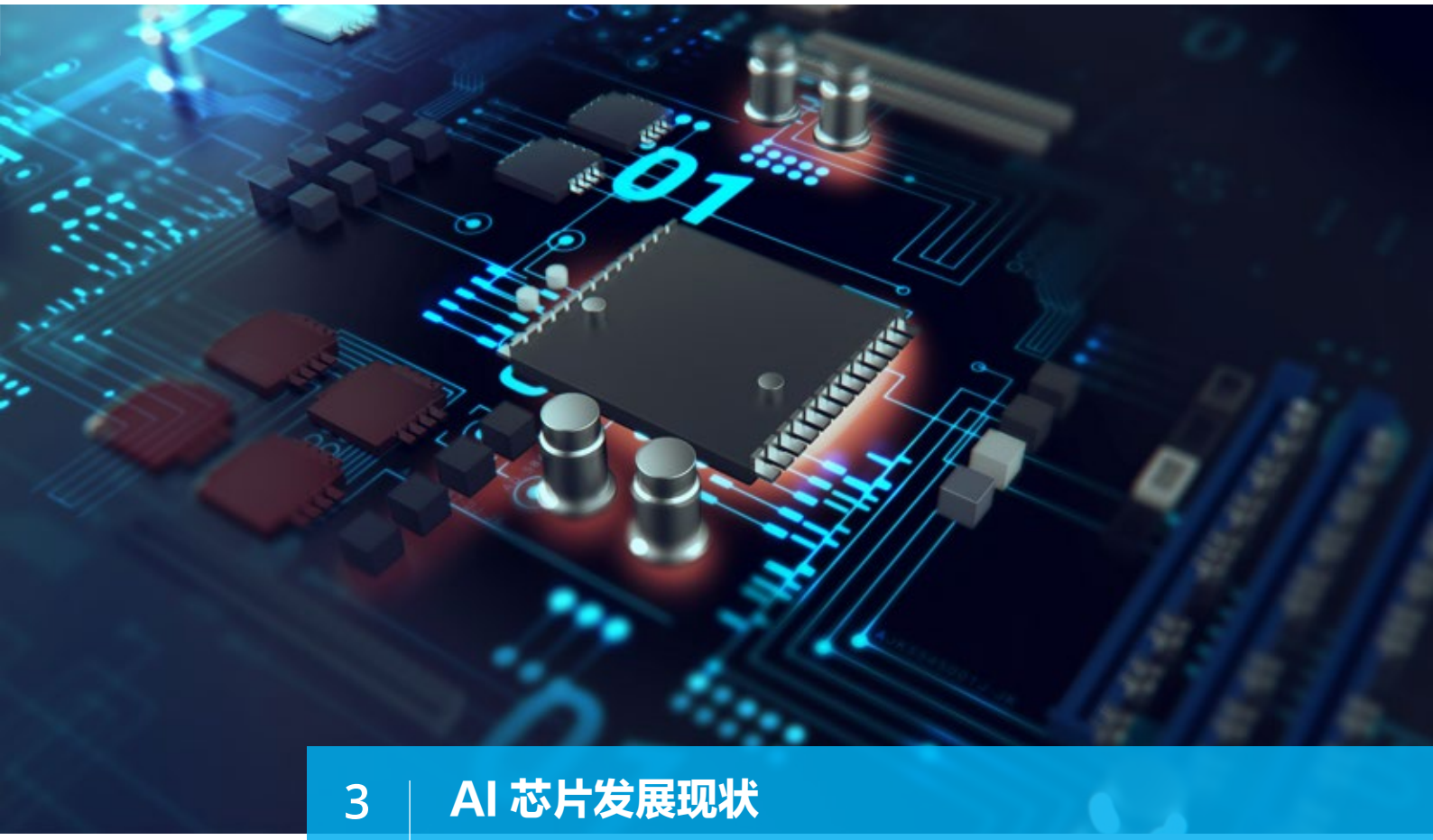
低精度设计是 AI 芯片的一个趋势，在针对推断的芯片中更加明显。对一些应用来说，降低精度的设计不仅加速了机器学习算法的推断（也可能是训练），甚至可能更符合神经形态计算的特征。近期已经证明，对于学习算法和神经网络的某些部分，使用尽可能低的精度（例如二进制数据）就足以达到预期效果，同时可以节省大量内存和降低能量消耗。通过对数据上下文数据精度的分析和对精度的舍入误差敏感性，来动态地进行精度的设置和调整，将是 AI 芯片设计优化的必要策略。

2.6 可重构能力

人工智能各领域的算法和应用还处在高速发展和快速迭代的阶段，考虑到芯片的研发成本和周期，针对特定应用、算法或场景的定制化设计很难适应变化。针对特定领域（包括具有类似需求的多种应用）而不针对特定应用的设计，将是 AI 芯片设计的一个指导原则，具有可重构能力的 AI 芯片可以在更多应用中大显身手，并且可以通过重新配置，适应新的 AI 算法、架构和任务。

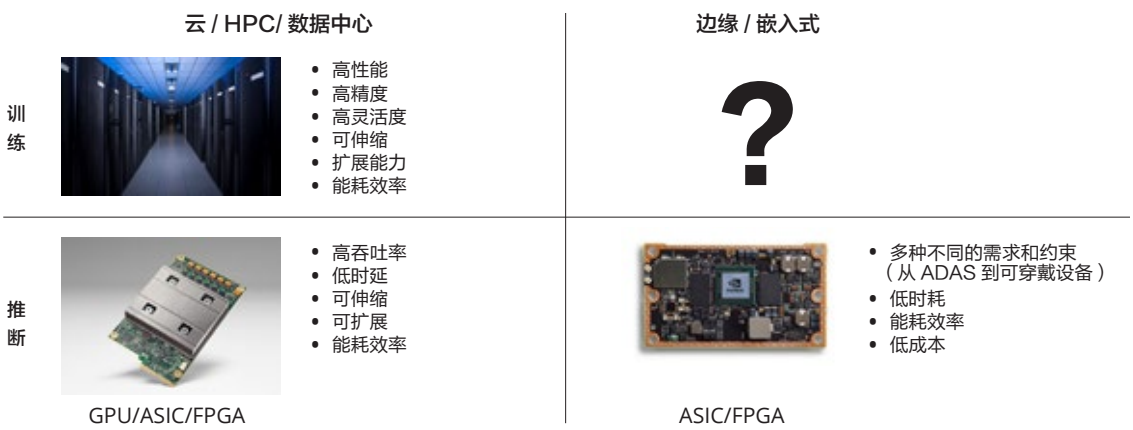
2.7 软件工具

就像传统的 CPU 需要编译工具的支持，AI 芯片也需要软件工具链的支持，才能将不同的机器学习任务和神经网络转换为可以在 AI 芯片上高效执行的指令代码，如 NVIDIA GPU 通过 CUDA 工具获得成功。基本处理、内存访问及任务的正确分配和调度将是工具链中需要重点考虑的因素。当工具链将神经网络或其它模型映射到 AI 芯片上时，也有很多优化代码的机会，比如神经网络剪枝、权重压缩和动态量化等。目前，AI 算法开发框架如 TensorFlow、Caffe 和 PyTorch 等，在 AI 应用研发中已经起到了至关重要的作用。对 AI 芯片来说，构建一个集成化的流程，将 AI 模型的开发和训练，硬件无关和硬件相关的代码优化，自动化指令翻译等功能无缝的结合在一起，将是成功部署的关键要求。



3 AI 芯片发展现状

从 2015 年开始，AI 芯片的相关研发逐渐成为学术界和工业界研发的热点。到目前为止，在云端和终端已经有很多专门为 AI 应用设计的芯片和硬件系统。同时，针对目标应用是“训练”还是“推断”，我们可以把 AI 芯片的目标领域分成 4 个象限，如图表 3-1 所示。其中，在边缘 / 嵌入设备中以推断应用为主，训练的需求还不是很明确。有些高性能的边缘设备虽然也会进行训练，但从硬件本身来说，它们更类似于云端设备。未来的边缘和嵌入设备可能都需要具备一定的学习能力，以支持在线学习功能。其他几个象限都有自身实现的需求和约束，目前也都有针对性的芯片和硬件系统。

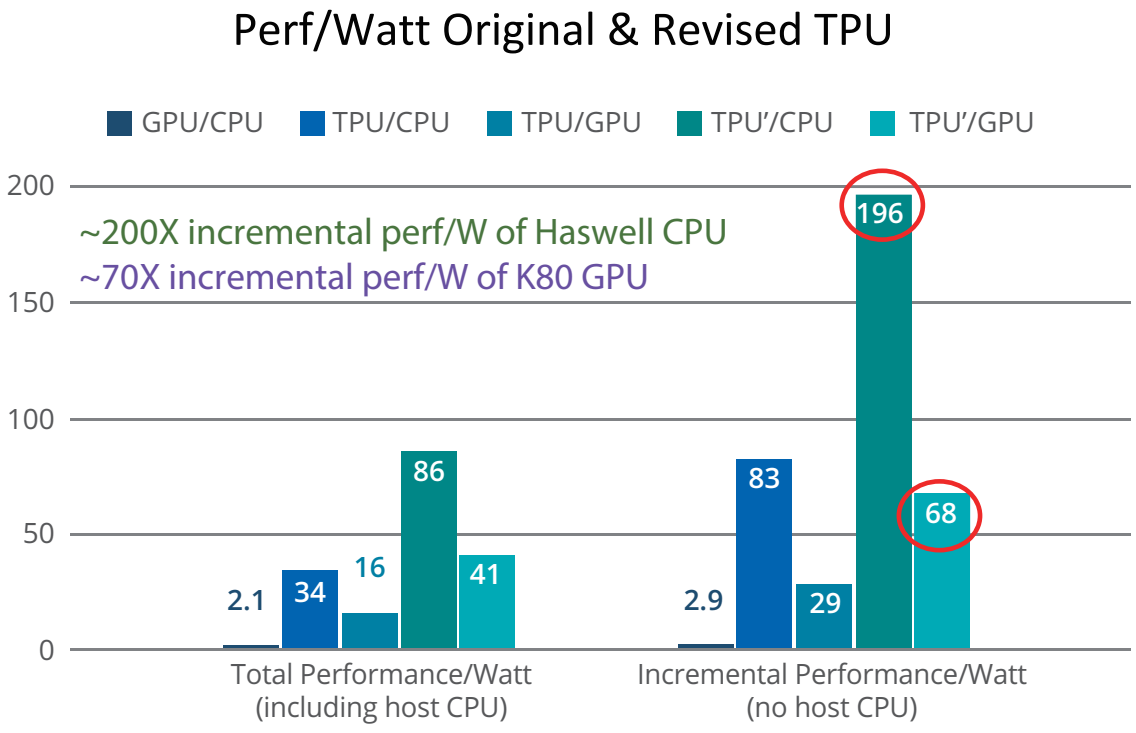


图表 3-1 AI 芯片的目标领域

3.1 云端 AI 计算

在云端，通用 GPU，特别是 NVIDIA 系列 GPU 芯片，被广泛应用于深度神经网络训练和推理。与 CPU 相比，拥有数千个计算内核的 GPU 可以实现 10-100 倍的吞吐量。其最新的 Tesla V100 除了 GPU 核之外，还专门针对深度学习设计了张量核（Tensor Cores），能够提供 120 TFLOPS（每秒 120 万亿次浮点指令）的处理能力。同时，NVIDIA GPU 还有比较完善的软件开发环境，是目前 AI 训练领域使用最广泛的平台。

面向云端 AI 应用，很多公司开始尝试设计专用芯片以达到更高的效率，其中最著名的例子是 Google TPU，可以支持搜索查询、翻译等应用，也是 AlphaGo 的幕后英雄。由于使用了专用架构，TPU 实现了比同时期 CPU 和 GPU 更高的效率（如图表 3-2 所示）。第一代的 TPU 仅能用于推断，面对目前被 NVIDIA GPU 赚得盆满钵满的深度学习训练市场，Google 随后又发布了第二版 TPU（TPU2），除了推断以外，还能高效支持训练环节的加速。Google 最近还通过云服务把 TPU 开放商用，处理能力达到 180 TFLOP，提供 64GB 的高带宽内存（HBM），2400GB/s 的存储带宽。



图表 3-2 Google TPU 性能（Hot Chips 2017）

针对云端的训练和推断市场，从芯片巨头到初创公司都高度重视。英特尔宣布推出 Nervana™ 神经网络处理器 (NNP)，该系列架构还可以优化 32GB HBM2、1TB/s 带宽和 8Tb/s 访问速度的神经网络计算。初创公司，如 Graphcore、Cerebras、Wave Computing、寒武纪及比特大陆等也加入了竞争的行列。

此外，FPGA 在云端的推断也逐渐在应用中占有一席之地。一方面，FPGA 可以支持大规模并行的硬件设计，和 GPU 相比可以降低推断的延时和功耗。微软的 Brainwave 项目和百度 XPU 都显示，在处理批量小的情况下，FPGA 具有出色的推断性能。另一方面，FPGA 可以很好地支持不同的数值精度，非常适合低精度推断的实现。进一步地，FPGA 的可编程能力也使它可以相对更快地支持新的算法和应用。目前，FPGA 的主要厂商如 Xilinx、Intel 都推出了专门针对 AI 应用的 FPGA 硬件（支持更高的存储带宽）和软件工具；主要的云服务厂商，比如亚马逊、微软及阿里云等推出了专门的云端 FPGA 实例来支持 AI 应用。一些初创公司，比如深鉴科技等也在开发专门支持 FPGA 的 AI 开发工具。

3.2 边缘 AI 计算

随着人工智能应用生态的爆发，越来越多的 AI 应用开始在端设备上开发和部署。对于某些应用，由于各种原因（如延迟，带宽和隐私问题），必须在边缘节点上执行推断。比如，自动驾驶汽车的推断就不能交由云端完成，否则如果出现网络延时，则会发生灾难性后果。再比如，大型城市动辄百万的高清摄像头，其人脸识别如果全交由云端完成，高清录像的数据传输会让通信网络不堪重负。

边缘设备实际上覆盖了一个很大的范围，其应用场景也五花八门。比如自动驾驶汽车可能就需要一个很强的计算设备，而在可穿戴领域，则要在严格的功耗和成本约束下实现一定的智能 。在未来相当一部分人工智能应用场景中，边缘设备主要执行推断计算，这就要求边缘处的终端设备本身具备足够的推断计算能力。而目前边缘处理器芯片的计算能力并不能满足在本地实现深度神经网络推断的需求。因此，业界需要专门设计的 AI 芯片，赋予设备足够的能力去应对越来越多的人工智能应用场景。除了计算性能的要求之外，功耗和成本也是在边缘节点工作的 AI 芯片必须面对的重要约束。

智能手机是目前应用最为广泛的边缘计算设备，包括苹果、华为、高通、联发科和三星在内的手机芯片厂商纷纷推出或者正在研发专门适应 AI 应用的芯片产品。另外，也有很多初创公司加入这个领域，为边缘计算设备提供芯片和系统方案，比如地平线机器人、寒武纪、深鉴科技、元鼎音讯等。传统的 IP 厂商，包括 ARM、Synopsys 等公司也都为包括手机、智能摄像头、无人机、工业和服务机器人、智能音箱以及各种物联网设备等边缘计算设备开发专用 IP 产品。

自动驾驶是未来边缘 AI 计算的最重要应用之一，MobileEye SOC 和 NVIDIA Drive PX 系列提供神经网络的处理能力可以支持半自动驾驶和完全自动驾驶，处理来自多路视频摄像头、雷达、激光雷达以及超声传感器的输入，并将这些数据相融合以确定汽车所处的精确位置，判断汽车周围的环境，并为安全行驶计算最佳路径和操作。

3.3 云和端的配合

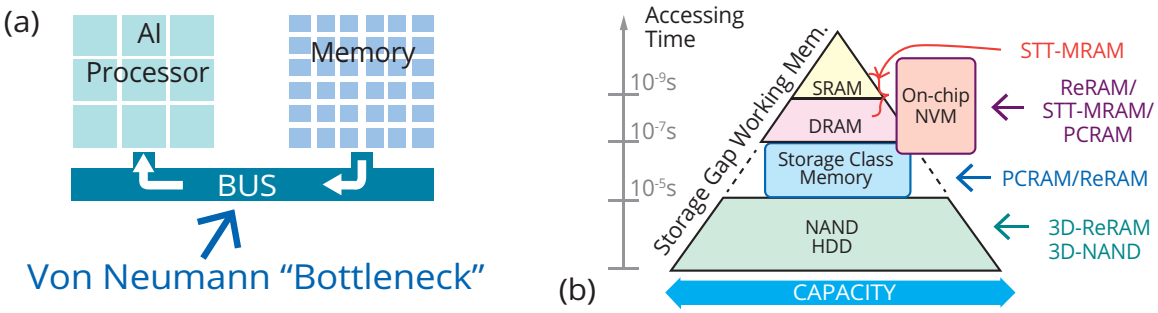
总的来说，云侧 AI 处理主要强调精度、处理能力、内存容量和带宽，同时追求低延时和低功耗；边缘设备中的 AI 处理则主要关注功耗、响应时间、体积、成本和隐私安全等问题。

目前云和边缘设备在各种 AI 应用中往往是配合工作。最普遍的方式是在云端训练神经网络，然后在云端（由边缘设备采集数据）或者边缘设备进行推断。随着边缘设备能力的不断增强，越来越多的计算工作负载将在边缘设备上执行，甚至可能会有训练或者学习的功能在边缘设备上执行。另一方面，云的边界也逐渐向数据的源头推进，未来很可能在传统的终端设备和云端设备直接出现更多的边缘设备，它们会把 AI 处理分布在各种网络设备（比如 5G 的基站）中，让数据尽量实现本地处理。从这个角度看，未来云和边缘设备以及连接他们的网络可能会构成一个巨大的 AI 处理网络，它们之间的协作训练和推断也是一个有待探索的方向。

4 AI 芯片的技术挑战

当我们讨论一个新的技术趋势时，首先需要清楚它背后的成因。很多大的技术创新都是需求驱动和技术瓶颈同时存在的情况下产生的。AI 芯片和它背后的各种技术也不例外。首先是巨大需求，一方面体现在 AI 应用的需求，也体现在 AI 特别是深度学习所要求的新的计算范式（这一点在上文已有介绍）。需求的驱动要求能够更加高效地处理 AI 运算的硬件，而在目前的技术框架下我们也遇到一些瓶颈问题，特别是冯·诺伊曼瓶颈和 CMOS 工艺和器件瓶颈。在详细介绍各种 AI 芯片的技术创新和未来的发展趋势之前，本节先简单介绍讨论一下这两个问题。

4.1 冯 · 诺伊曼瓶颈



图表 4-1 (a)AI 芯片中的冯 · 诺伊曼“瓶颈”(b) 内存层级结构

如前所述，提高 AI 芯片性能和能效的关键之一在于支持高效的数据访问。如图表 4-1 所示，在传统冯·诺伊曼体系结构中，数据从处理单元外的存储器提取，处理完之后再写回存储器。在 AI 芯片实现中，基于冯·诺伊曼体系结构，提供运算能力相对是比较简单易行的，但由于运算部件和存储部件存在速度差异，当运算能力达到一定程度，由于访问存储器的速度无法跟上运算部件消耗数据的速度，再增加运算部件也无法得到充分利用，即形成所谓的冯 · 诺伊曼“瓶颈”，或“内存墙”问题，是长期困扰计算机体系结构的难题。目前常见的方法是利用高速缓存（Cache）等层次化存储技术尽量缓解运算和存储的速度差异。

性能指标	AlexNet	VGG 16	GoogLeNet V1	ResNet 50
Top-5 错误率	16.4	7.4	6.7	5.3
卷积层数量	5	13	57	53
权重值数量	2.3M	14.7M	6.0M	23.5M
MAC 数量	666M	15.3G	1.43G	3.86G
全连接层数量	3	3	1	1
权重值数量	58.6M	124M	1M	2M
MAC 数量	58.6M	124M	1M	2M
总权重值数量	61M	138M	7M	25.5M
总 MAC 数量	724M	15.5G	1.43G	3.9G

图表 4-2 常见神经网络的基本参数（source：[Vivienne17]）

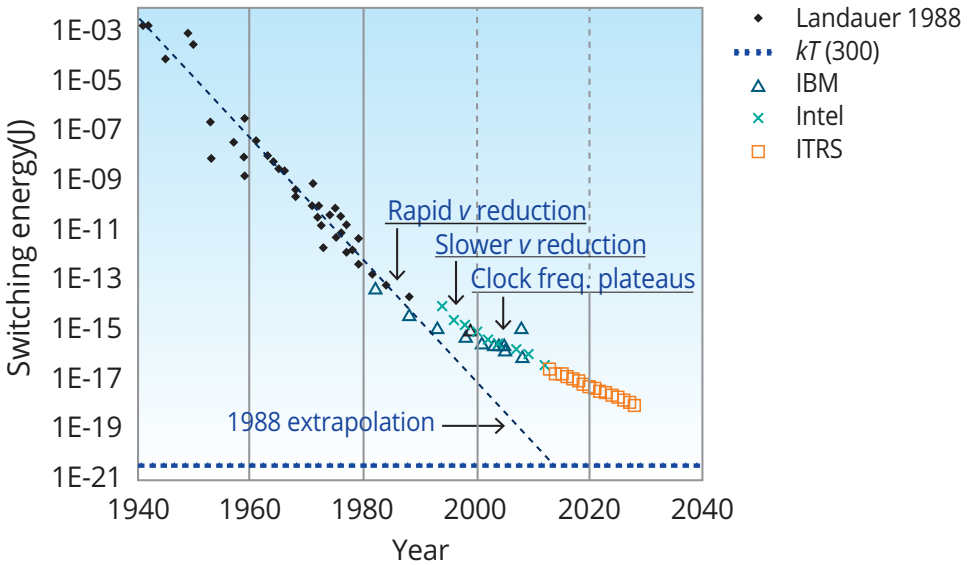
然而，AI 芯片中需要存储和处理的数据量远远大于之前常见的应用。图表 4-2 列出了一些常见的深度神经网络的主要参数，其中 VGG16 网络需要 138M 个权重参数，一次推断过程需要 15.5G 次乘累加运算。这都使得冯 · 诺伊曼“瓶颈”问题在 AI 应用中愈发严重。可以不夸张地说，大部分针对 AI，特别是加速神经网络处理而提出的硬件架构创新都是在和这个问题做斗争。概括来说，在架构层面解决这一问题的基本思路有二：1）减少访问存储器的数量，比如减少神经网络的存储需求（参数数量，数据精度，中间结果）、数据压缩和以运算换存储等；2）降低访问存储器的代价，尽量拉近存储设备和运算单元的“距离”，甚至直接在存储设备中进行运算。

4.2 CMOS 工艺和器件瓶颈

现今的计算机可以达到 10^{15} FLOPS 的处理速度。这些系统在先进科技的研究（生物学，气候分析，基因组学，脑基因组学，材料开发等）中扮演着重要角色。在许多方面，计算能力驱动着现代社会的发展。人工智能，特别是机器学习的发展将需要更加强有力的，超过 10^{18}FLOPS 运算能力的计算系统。

当前，构建这些系统的基础是 CMOS 技术的芯片，而 CMOS 工艺能够不断提高系统性能主要得益于集成尺寸的缩小。过去 30 年，摩尔定律很好地预测了这种计算进步。2018 年，10 纳米工艺的芯片已经大规模量产，7 纳米开始量产，5 纳米节点的技术定义已经完成。然而，由于基础物理原理限制和经济的原因，持续提高集成密度将变得越来越困难 [Theis16]。目前，CMOS 器件的横向尺寸接近几纳米，层厚度只有几个原子层，这会导致显著的电流泄漏，降低工艺尺寸缩小的效果。此外，这些纳米级晶体管的能量消耗非常高，很难实现密集封装。

另外，物联网（IoT）、社交媒体和安全设备产生了大量的数据，存储、交换和处理这些数据都需要大



图表 4-4 逻辑器件的最小翻转能耗趋势（Source: [Theis16]）

量的存储器。目前 DRAM 技术已经接近极限，DRAM 性能和存储器容量的进步主要通过封装技术来实现，即堆叠多个 DRAM 芯片并通过硅通孔（TSV）将其连接到存储器控制器芯片上。增加数据带宽则是通过更宽的数据总线来实现的。非易失存储技术的主力是 NAND 闪存，最先进的 3D NAND 具有多达 64 层和 256 Gb 的容量，预计于 2018 年进入市场。由于 DRAM 和 NAND 闪存都是独立于计算核心的，和计算核心进行数据交换的代价（包括时间和能耗）非常大，而目前能够和计算核心紧耦合的片上存储器的唯一方案是 SRAM，其容量为兆级。即使采用最小的 SRAM 单元填充 1 平方厘米芯片面积的一半，也只有约 128 兆的片上存储容量。因此，我们有充足的理由开发提供大量存储空间的片上存储器技术，并探索利用片上存储器去构建未来的智能芯片架构 [Aly15]。

在计算架构和器件层面，我们也需要一种新的方法。大脑可以看作一个可以处理大量（通常是模糊的）信息，且具有超高密度、错误恢复能力和高能效的计算模型。神经元和大脑突触的能量消耗比最先进的 CMOS 器件还低几个数量级。另外，大脑可以处理模式识别和特征提取的问题，这对传统架构来说非常困难，甚至不可能实时实现。理想情况下，我们需要具有生物系统优点而规避速度慢等缺点的器件和材料。近年来，可以存储模拟数值的非易失性存储器发展迅猛，它可以同时具有存储和处理数据能力，可以破解传统计算体系结构的一些基本限制，有望实现类脑突触功能。

5 AI 芯片架构设计趋势

5.1 云端训练和推断：大存储、高性能、可伸缩

之前我们分析了云端训练和推断的基本需求，虽然训练和推断在数据精度、架构灵活和实时性要求上有一定的差别，但它们在处理能力（吞吐率）、可伸缩可扩展能力以及功耗效率上具有类似的需求。因此，针对云端的训练和推断而开发的专用芯片和技术创新，基本都是围绕这几个需求。

NVIDIA 的 V100 GPU 和 Google 包括四颗芯片的 Cloud TPU [Google]，是目前云端商用 AI 芯片的标杆。在深度学习计算的处理能力方面，V100 达到 120TFLOPS，Cloud TPU 则达到 180TFLOPS。值得一提的是，这种处理能力都是由专门针对深度学习需求而设计的运算单元提供。在存储和访存能力上，V100 有 16 GB HBM2 存储器，支持 900 GB/s 的带宽；而 Cloud TPU 单颗芯片有 16GB HBM 存储器，支持 600GB/s 的带宽。另外，它们共同的特点是支持多芯片的扩展能力，V100 支持 NVIDIA 的 NvLink 互连方式，可以扩展到 8 芯片的系统；而 Cloud TPU 也支持高速的芯片间互连接口和板级互连接口，非常适合在云端和数据中心部署。图表 5-1 是 Cloud TPU 的机柜，包括 64 个 TPU2，能够为机器学习的训练任务提供 11.5 PFLOPS 的处理能力和 4 TB 的 HBM 存储器。同时，这些运算资源还可以灵活地分配和伸缩，能够有效支持不同的应用需求。

图表 5-1 Google Cloud TPU Pod (Hot Chips 2017)

从 NVIDIA 和 Goolge 的设计实践我们可以看出云端 AI 芯片在架构层面，技术发展的几个特点和趋势：

1. 存储的需求（容量和访问速度）越来越高。一方面，由于处理大量数据的要求，需要更大容量的存储器。另一方面，限制运算能力提高的主要因素是访问存储器的速度，因此，未来云端 AI 芯片会有越来越多的片上存储器（比如 Graphcore 公司就在芯片上实现的 300MB 的 SRAM）和能够提供高带宽的片外存储器（HBM2 和其它新型封装形式）。

2. 处理能力推向每秒千万亿次（PetaFLOPS），并支持灵活伸缩和部署。对云端 AI 芯片来说，单芯片的处理能力可能会达到 PetaFLOPS 的水平。实现这一目标除了要依靠 CMOS 工艺的进步，也需要靠架构的创新。比如在 Google 第一代 TPU 中，使用了脉动阵列（Systolic Array）架构，而在 NVIDIA 的 V100GPU 中，专门增加了张量核来处理矩阵运算。为了将 GPU 扩展为更大的系统，NVIDIA 专门开发了的 NVSwitch 交换芯片，可以为多个 GPU 提供高带宽互连。在最新发布的 DGX-2 系统中，16 颗 V100 GPU 连接在一起，提供 2PFPLOS 的处理能力，可以实现大规模神经网络的并行训练。除此之外，我们还看到一些更为“极端”的架构设计。比如晶圆级集成技术，即用整个晶圆制成一个“超级芯片”；再比如在运算单元中使用无时钟电路实现更高的速度和更低的功耗。此外，还有一条路径是通过多芯片、多板卡互连来实现更强的运算和存储能力，而不是单纯追求单芯片的处理能力。未来应该可以看到越来越多的产品，以系统（或者云服务）而非单芯片的形式，提供可伸缩和配置的处理能力。这种强大处理能力的灵活性还体现在训练和推断任务的部署上，比如在白天将更多的硬件用于推断任务，满足应用需求，而晚上则把更多的资源分配给训练任务。

3. 专门针对推断需求的 FPGA 和 ASIC。随着 AI 应用的爆发，对推断计算的需求会越来越多，一个训练好的算法会不断复用。推断和训练相比有其特殊性，更强调吞吐率、能效和实时性，未来在云端很可能会有专门针对推断的 ASIC 芯片（Google 的第一代 TPU 也是很好的例子），提供更好的能耗效率并实现更低的延时。另外，FPGA 在这个方向也有独特优势，从微软提出的 BrainWave 架构就可以看出端倪。

5.2 边缘设备：把效率推向极致

相对云端应用，边缘设备的应用需求和场景约束要复杂很多，针对不同的情况可能需要专门的架构设计。抛开需求的复杂性，目前的边缘设备主要是执行“推断”。在这个目标下，AI 芯片最重要的就是提高“推断”效率。目前，衡量 AI 芯片实现效率的一个重要指标是能耗效率——TOPs/W，这也成为很多技术创新竞争的焦点。在 ISSCC2018 会议上，就出现了单比特能效达到 772 TOPs/W 的惊人数据 [Bankman18]。

在提高推断效率和推断准确率允许范围内的各种方法中，降低推断的量化比特精度是最有效的方法。它既可以大大降低运算单元的精度，又可以减少存储容量需求和存储器的读写。但是，降低比特精度也意味着推断准确度的降低，这在一些应用中是无法接受的。由此，基本运算单元的设计趋势是支持可变比特精度，比如 [Lee18] 的 BitMAC 就能支持从 1 比特到 16 比特的权重精度。

除了降低精度以外，提升基本运算单元（MAC）的效率还可以结合一些数据结构转换来减少运算量，比如通过快速傅里叶变换（FFT）变换来减少矩阵运算中的乘法（参考 [Vivienne17]）；还可以通过查表的方法来简化 MAC 的实现等。

对于使用修正线性单元（ReLU）作为激活函数的神经网络，激活值为零的情况很多；而在对神经网络进行的剪枝操作后，权重值也会有很多为零。基于这样的稀疏性特征，一方面可以使用专门的硬件架构，比如 [Parashar17] 中提出的 SCNN 加速器，提高 MAC 的使用效率，另一方面可以对权重和激活值数据进行压缩（参考 [Vivienne17]）。

另一个重要的方向是减少对存储器的访问，这也是缓解冯 · 诺伊曼“瓶颈”问题的基本方法。利用这样的稀疏性特性，再有就是拉近运算和存储的距离，即“近数据计算”的概念，比如把神经网络运算放在传感器或者存储器中。对于前者，已经有很多工作试图把计算放在传感器的模拟部分，从而避免模 / 数转换（ADC）以及数据搬移的代价。另一个趋势是先对传感器数据进行简单处理，以减少需要存储和移动的数据。比如，先利用简单神经网络，基于图像传感器得到的数据初步定位目标物体，再把只包括目标物体的部分存储，并传输给复杂的神经网络进行物体的识别。后者，即存内计算，我们会在后面的新型存储技术中详细讨论。

此外，在边缘设备的 AI 芯片中，也可以应用各种低功耗设计方法来进一步降低整体功耗。比如，当权重或者中间结果的值为零的时候，对 MAC 进行时钟门控（Clock-gating）。而 [Bert17] 提出的动态电压精度频率调整，则是在传统芯片动态功耗调整技术中增加了对于推断精度的考量。再者，目前一些运算单元采用异步设计（或者无时钟设计）来降低功耗，这也是一个值得探索的方向。

未来，越来越多的边缘设备将需要具备一定的“学习”能力，能够根据收集到的新数据在本地训练、优化和更新模型。这也会对边缘设备以及整个 AI 实现系统提出一些新的要求。

最后，在边缘设备中的 AI 芯片往往是 SoC 形式的产品，AI 部分只是实现功能的一个环节，而最终要通过完整的芯片功能来体现硬件的效率。这种情况下，需要从整个系统的角度考虑架构的优化。因此，终端设备 AI 芯片往往呈现为一个异构系统，专门的 AI 加速器和 CPU，GPU，ISP，DSP 等其它部件协同工作以达到最佳的效率。

5.3 软件定义芯片

在 AI 计算中，芯片是承载计算功能的基础部件，软件是实现 AI 的核心。这里的软件即是为了实现不同目标的 AI 任务，所需要的 AI 算法。对于复杂的 AI 任务，甚至需要将多种不同类型的 AI 算法组合在一起。即使是同一类型的 AI 算法，也会因为具体任务的计算精度、性能和能效等需求不同，具有不同计算参数。因此，AI 芯片必须具备一个重要特性：能够实时动态改变功能，满足软件不断变化的计算需求，即“软件定义芯片”。

通用处理器如 CPU、GPU，缺乏针对 AI 算法的专用计算、存储单元设计，功耗过大，能效较低；专用芯片（ASIC）功能单一，难以适应灵活多样的 AI 任务；现场可编程门阵列（FPGA）尽管可以通过编程重构为不同电路结构，但是重构的时间开销过大，而且过多的冗余逻辑导致其功耗过高。以上传统芯片都难以实现 AI 芯片需要的“软件定义芯片”这一特性。

可重构计算技术允许硬件架构和功能随软件变化而变化，具备处理器的灵活性和专用集成电路的高性能和低功耗，是实现“软件定义芯片”的核心，被公认为是突破性的下一代集成电路技术。清华大学微电子所设计的 AI 芯片（代号 Thinker [Shouyi17, Shouyi18]），采用可重构计算架构，能够支持卷积神经网络、全连接神经网络和递归神经网络等多种 AI 算法。Thinker 芯片通过三个层面的可重构计算技术，来实现“软件定义芯片”，最高能量效率达到了 5.09TOPS/W：

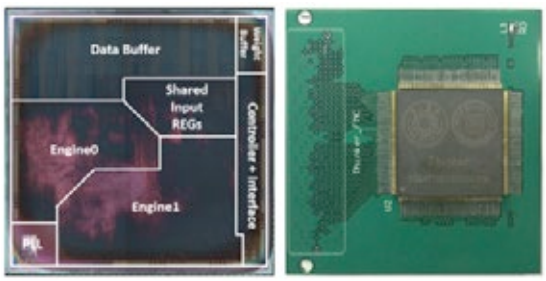
1. 计算阵列重构：Thinker 芯片的计算阵列由多个并行计算单元互连而成。每个计算单元可以根据算法所需要的基本算子不同而进行功能重构。此外，在复杂 AI 任务中，多种 AI 算法的计算资源需求不同，因此 Thinker 芯片支持计算阵列的按需资源划分以提高资源利用率和能量效率。

2. 存储带宽重构：Thinker 芯片的片上存储带宽能够根据 AI 算法的不同而进行重构。存储内的数据分布会随着带宽的改变而调整，以提高数据复用性和计算并行度，提高了计算吞吐和能量效率。

3. 数据位宽重构：16 比特数据位宽足以满足绝大多数应用的精度需求，对于一些精度要求不高的场景，甚至 8 比特数据位宽就已经足够。为了满足 AI 算法多样的精度需求，Thinker 芯片的计算单元支持高低（16/8 比特）两种数据位宽重构。高比特模式下计算精度提升，低比特模式下计算单元吞吐量提升进而提高性能。

可重构计算技术作为实现“软件定义芯片”的重要技术，非常适合应用于 AI 芯片设计当中。采用可重构计算技术之后，软件定义的层面不仅仅局限于功能这一层面。算法的计算精度、性能和能效等都可以纳入软件定义的范畴。可重构计算技术借助自身实时动态配置的特点，实现软硬件协同设计，为 AI 芯片带来了极高的灵活度和适用范围。

Technology	TSMC 65nm LP
Supply	0.67V~1.2V
Area	4.4mm*4.4mm
SRAM	348KB
Frequency	10~200MHz
Peak performance	409.6GOPS
Power	4mW~447mW
Energy efficiency	1.6TOSP/W~5.09TOPS/W

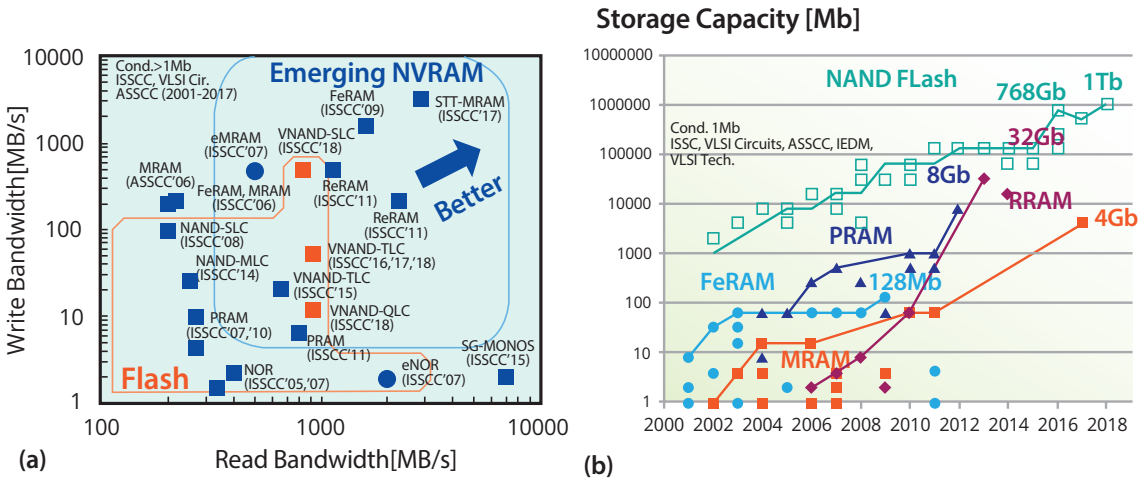


图表 5-2 清华大学 Thinker 芯片

6 AI 芯片中的存储技术

AI 芯片由于在广泛的识别和分类任务中显示出其优越的能力，成为物联网（IoT）系统和大数据处理的关键促成因素之一。如前所述，提高 AI 芯片的性能和能效的关键之一在于数据访问。而在传统的冯·诺伊曼体系结构中，数据从存储器串行提取并写入到工作内存，导致相当长的延迟和能量开销。从器件到体系结构的全面创新预计将赋予 AI 芯片更强的能力。近期，面向数字神经网络的加速器（GPU、FPGA 和 ASIC）迫切需要 AI 友好型存储器；中期，基于存内计算的神经网络可以为规避冯·诺依曼瓶颈问题提供有效的解决方案；最后，基于忆阻器的神经形态计算可以模拟人类的大脑，是 AI 芯片远期解决方案的候选之一。

6.1 AI 友好型存储器

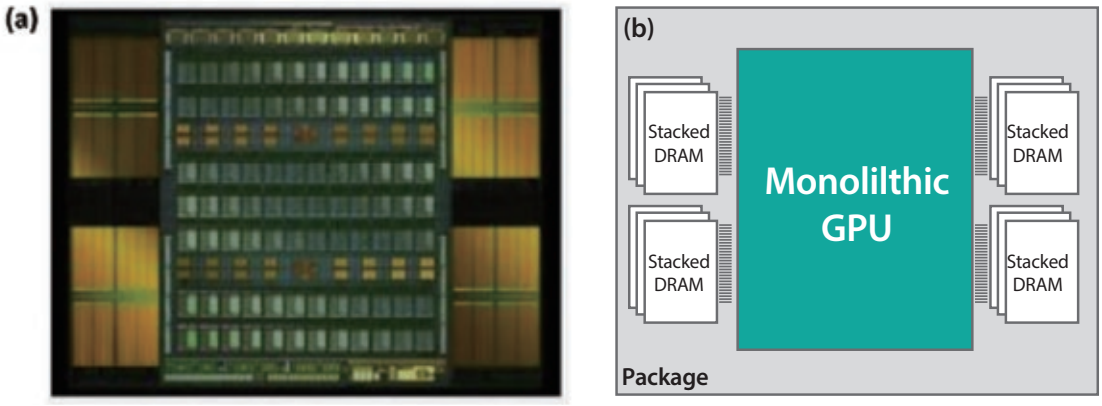


图表 6-1 新兴内存的 (a) 带宽和 (b) 存储容量 [ISSCC Trend]

考虑到并行访问大量数据的需求，人工智能和大数据处理需要高带宽、大存储容量的内存。图表 6-1 显示了当前主要存储技术中带宽和容量的快速增长。考虑到在不断尺寸缩放的情况下，传统 NVM 所面临的困难越来越多，新兴的 NVM 由于其相对较大的带宽和迅速增长的容量，可以在 AI 芯片的存储技术中发挥至关重要的作用。

6.2 片外存储器

DRAM 和 NAND 闪存由于高密度的单元结构特点，通常被用作具有相对较大容量的片外存储器。3D 集成已经被证明是增加商业存储器的带宽和容量的有效策略，其可以通过使用从底部到顶部的硅通孔（TSV）技术，堆叠多个管芯或者单片制造的方法来完成。DRAM 的代表作品包括 HBM [Lee14] 和混合存储器立方体（HMC）[Jeddeloh12]。图表 6-2 显示了 NVIDIA 的 GPU 产品与 HBM 集成的 AI 应用程序 [NVIDIA]。对于 NAND 闪存，3D NAND 正在深入研究。最近，三星已经开发出 96 层 3D NAND。



图表 6-2 (a) 硅片照片和（b）用于数据中心应用的具有高带宽存储器（HBM）的 NVIDIA GPU 概念图 [NVIDIA]

6.3 片上（嵌入型）存储器

由于能够连接逻辑和存储器电路，并且与逻辑器件完全兼容，SRAM 是不可或缺的片上存储器，其性能和密度不断受益于CMOS的尺寸缩放。然而，其易失性使得芯片上或芯片外的非易失性存储器成为必须。虽然目前 NOR 闪存被广泛用作片上 NVM，但由于其存取时间较短且写入能量较大，限制了系统的性能。

器件指标	SRAM	DRAM	NAND	NOR	PCM	STT-MRAM	ReRAM
写能耗	低	低	高	高	中等	中等	中等
写延时	~1ns	~5ns	> 100μs	10μs~1ms	100~150ns	2~200ns	10~100ns
读延时	~1ns	20~80ns	5~200μs	~ 50ns	~ 50ns	1.3~25ns	3~200ns
编程窗口	好	好	好	好	可变	小	可变
擦写次数	无限制	无限制	10^4–10^5	10^4–10^5	10^8–10^9	~10^{15}	10^5–10^{10}
单元尺寸	~100F^2	~7F^2	~4F^2	~10F^2	~4F^2	~12F^2	~4–6F^2

图表 6-3 当前主要和新兴存储器的器件指标

6.4 新兴的存储器

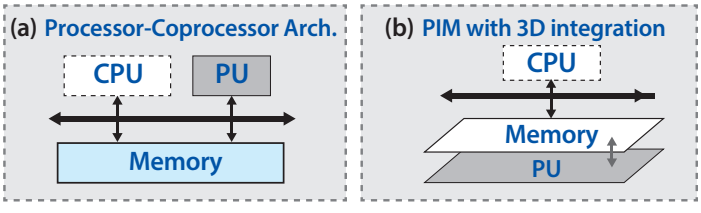
新兴的 NVM 技术可以显著改善用于商业和嵌入式应用的 AI 友好型存储器。对于商业存储器，新兴的 NVM 因为速度较为匹配，可以用作存储级内存（SCM）来弥补工作内存和存储之间的访问时间差别。因为可以高密度集成，相变存储器（PCM）和阻变存储器（ReRAM）是 SCM 的主要候选者。此外，自旋力矩传输存储器（STT-MRAM）由于其高耐久性和高速度被认为是 DRAM 的替代品。对于嵌入式应用，基于新兴 NVM 的片上存储器也可以提供比传统 NVM 更好的存取速度和低功耗，可在非常有限的功率下工作，这对于物联网边缘设备上的 AI 芯片特别具有吸引力。

7 新兴计算技术

新兴计算技术已经被提出并被研究，以减轻或避免当前计算技术中的冯 · 诺依曼体系结构的“瓶颈”。主要的新兴计算技术包括近内存计算、存内计算，以及基于新型存储器的人工神经网络和生物神经网络。虽然成熟的 CMOS 器件已被用于实现这些新的计算范例，但是新兴器件有望在未来进一步显著提高系统性能并降低电路复杂性。

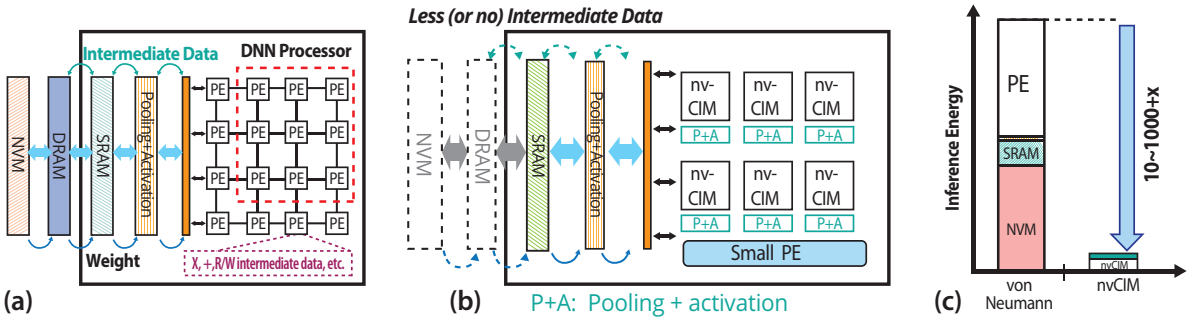
7.1 近内存计算

除了将逻辑电路或处理单元（PU）放置在存储器附近，将它们通过宽总线连接以最小化由数据传输引起的延迟和功率损耗以及增加带宽等方法外，近存储器计算可以通过将存储器层置于逻辑层顶部而进一步实现高性能并行计算。 新兴的 NVM 也适用于这种方法，因为它可以通过 CMOS 的后道工序（Back-end-of-line, BEOL）与逻辑器件集成。



图表 7-1（a）传统的冯 · 诺依曼架构和（b）近存储器计算 [Chi16]

7.2 存内计算（In-memory Computing）

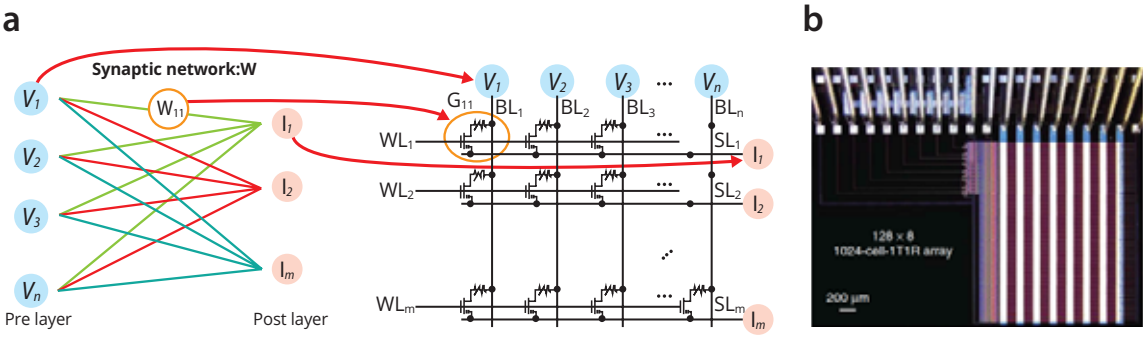


图表 7-2 AI 芯片基于（a）von Neumann（b）内存计算和（c）功耗比较 [Chen18]

存内计算与传统的冯·诺依曼体系结构有着本质不同，该体系结构直接在存储器内执行计算而不需要数据传输。这个领域的最新进展已经证明了存内计算具有逻辑运算和神经网络处理的能力 [Chen17 & 18]。图表 7-2 的展示了基于冯·诺依曼和存内计算架构的 AI 芯片的概念图。利用存内计算模块后，功耗和延迟可以显著降低。

7.3 基于新型存储器的人工神经网络

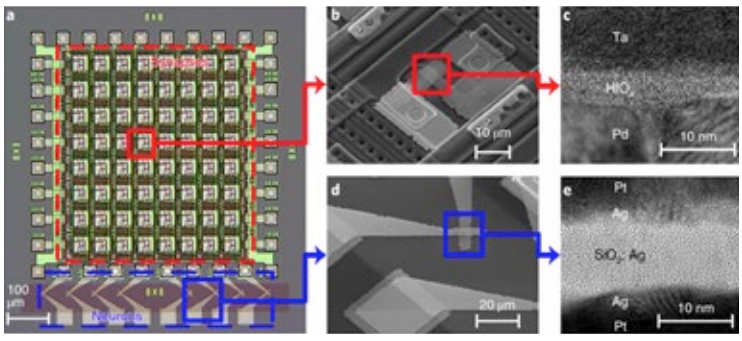
基于新兴非易失性存储器件的人工神经网络计算最近引起了人们的极大关注 [Yang13]。这些器件包括铁电存储器（FeRAM）、磁隧道结存储器（MRAM）、相变存储器（PCM）和阻变存储器（RRAM）等，它们可用于构建待机功耗极低的存储器阵列。更重要的是，它们都可能成为模拟存内计算（Analog In-memory Computing）的基础技术，实现数据存储功能的同时参与数据处理。这些器件一般都以交叉阵列（Crossbar）的形态实现，其输入 / 输出信号穿过构成行列的节点。图表 7-3 就是一个 RRAM 交叉阵列的例子，其中矩阵权重被表示为电导。交叉阵列非常自然地实现了向量和矩阵乘法，这对于各种基于 AI 的应用具有重要的意义。使用图中集成 1024 单元的阵列进行并行在线训练，清华大学吴华强课题组在国际上首次成功实现了灰度人脸分类。与 Intel 至强处理器（使用片外存储）相比，每次迭代模拟突触内的能量消耗低 1,000 倍，而测试集的准确度与 CPU 计算结果相近 [Yao17]。另外，在 [Li18] 中，基于 128×64 的 ReRAM 交叉阵列，输出精度为 5-8 比特的精确模拟信号处理和图像处理也得到了实验演示。和传统的 CMOS 电路相比，模拟存内计算可以用非常低的功耗实现信号的并行处理，从而提供很高的数据吞吐率 [Li18]。由于存储元件的状态可以映射为突触，这种交叉开关阵列实际上实现了一个全连接的硬件神经网络 [Prezioso15] [Hu18] 的物理实例。



图表 7-3 一个 ReRAM 交叉阵列用于人脸识别的例子（a）在 1T1R 阵列上映射一层神经网络，'T' 代表晶体管，'R' 代表 ReRAM（b）使用 CMOS 兼容工艺制造的 1024-cell-1T1R 阵列的显微照片。[Yao17]

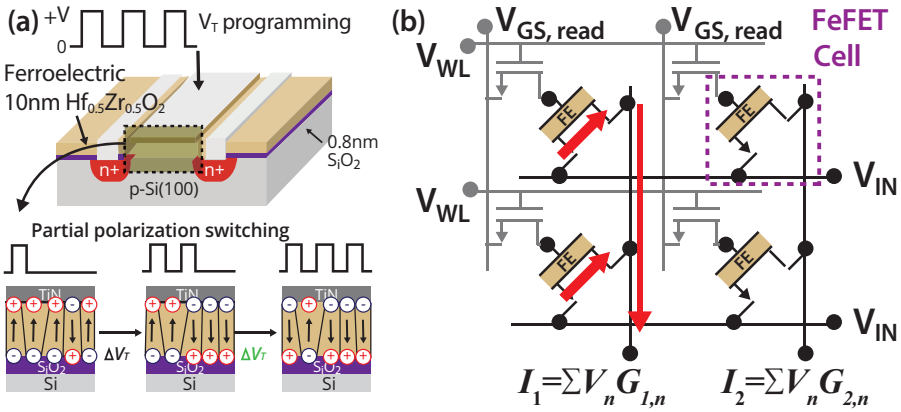
7.4 生物神经网络

上述人工神经网络本质上是存储和计算并行。另一种更具生物启发性的方法是采用脉冲神经网络等，更严格地模拟大脑的信息处理机制。IBM TrueNorth 和最近宣布的英特尔 Loihi 展示了使用 CMOS 器件的仿生脉冲神经网络硬件实现。前者包括 106 个神经元和 2.56×108 SRAM 突触 [Merolla14]，后者拥有 1.3×105 个神经元和 1.3×108 突触 [Intel17]。脉冲神经网络方法需要有与生物神经元的动力学相似的人工突触和神经元。然而，由于 CMOS 器件需要用多个晶体管来模拟一个突触或神经元，我们需要新的具有生物突触和神经元内在相似性的紧凑型物理结构，用于复制生物神经网络行为。实际上，对模拟突触功能至关重要的人工突触已经被简单的两终端忆阻器实现 [Wang17]。最近，带有积分泄漏和发放功能的人工神经元也被单一的忆阻器器件实现 [Wang18]。第一个基于忆阻器的人工神经元和突触集成的神经网络示意见图表 7-4。图中展示了采用忆阻神经网络进行无监督学习的模式分类 [Wang18]。实验证明虽然神经形态计算仍然是其技术成熟度的早期阶段，但是它代表了 AI 芯片的一个很有前景的长期方向。



图表 7-4 用于无监督学习的模式分类的第一个完全集成的基于忆阻器的神经网络。a, 集成了忆阻器的神经网络的光学显微照片，由 8×8 记忆突触和 8 个忆阻人工神经元组成。b, 单个人工突触的扫描电子显微照片。c, 突触的透射电镜图像。d, 单个人工神经元的扫描电子显微照片。e, 神经元截面的高分辨率透射电子显微照片。[Wang18]。

除了两终端器件外，新兴的三端晶体管也可以用来构建神经网络。例如，将铁电电容器集成到晶体管的栅极所构成的 FeFET，铁电电容器的极化程度与通道的跨导相关。FeFET 可以提供快速的编程速度，低功耗和平滑的对称模拟编程。利用上述交叉阵列结构，FeFET 也可以自然地实现向量 – 矩阵乘法（见图表 7-5 ）。交叉开关阵列中 FeFET 的模拟状态能够表示全连接神经网络中突触的权重 [Jerry17]。此外，铁电体层的晶格极化动力学还可以实现脉冲神经网络（SNN）的时间学习规则，如脉冲时序而定的可塑性（Spiking-timing-dependent Plasticity, STDP ）[Boyn17]。



图表 7–5（a）FeFET 显示的模拟计算能力（b）由 FeFET 伪交叉网络实现模拟的存内计算 [Jerry17]

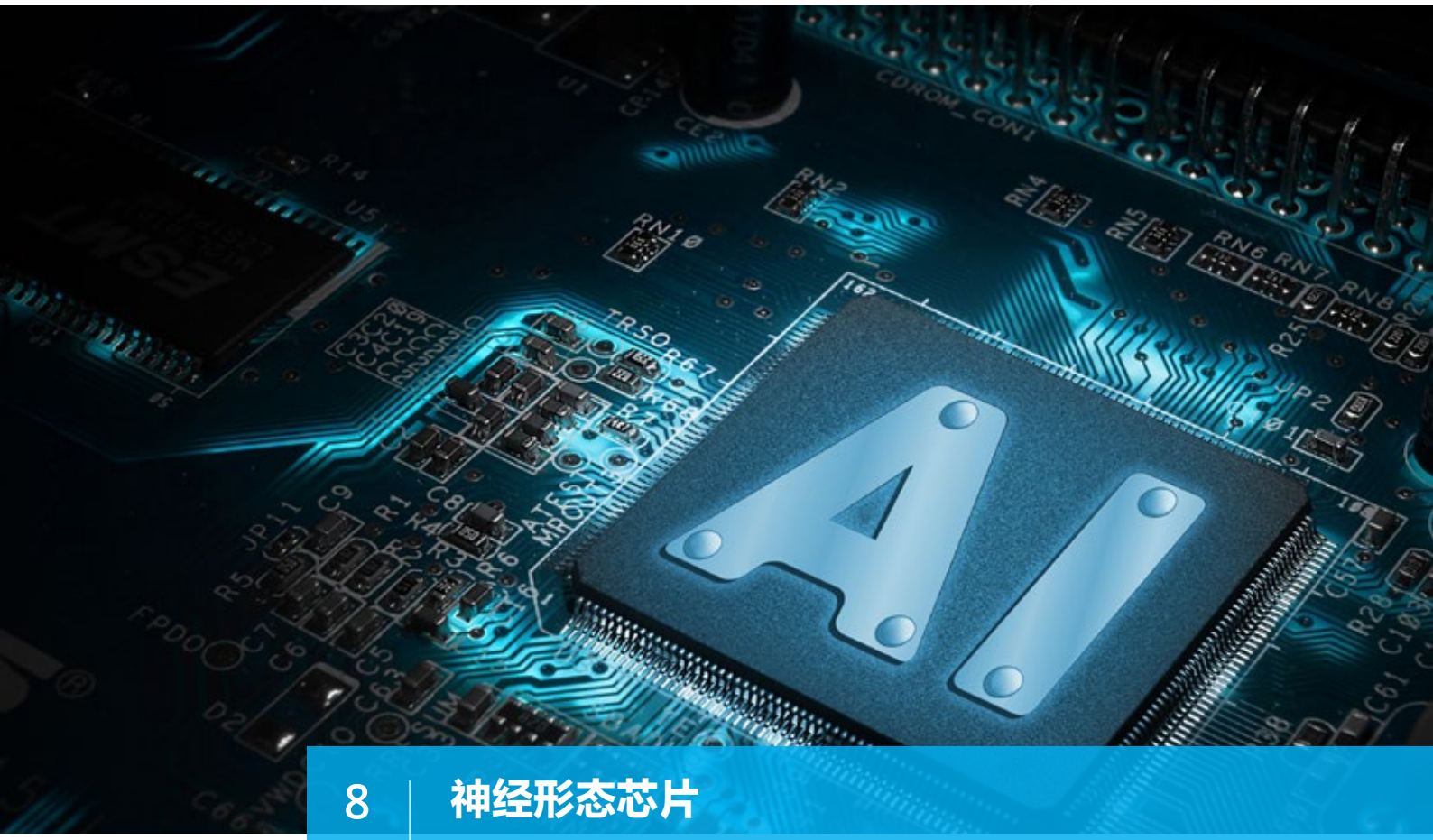
虽然新兴的计算可以使用当前的 CMOS 器件实现，但新兴的内存技术仍然是支撑新型计算技术和 AI 技术蓬勃发展的重中之重。AI 友好型存储器在近期也迫切期望用于解决缓解冯 · 诺依曼“瓶颈”难题。近内存计算、存内计算和基于忆阻器的神经形态计算在超越冯·诺依曼计算方面都具重要性，可服务于 AI 技术的持续快速发展。

7.5 对电路设计的影响

模拟存内计算有很大的潜力实现比数字乘累加单元更快的速度和更高的能效。然而，模拟电路操作也给外围电路的设计带来了新的挑战。

与数字方法不同，由于每个矩阵元素的误差在求和过程中会被累积并影响输出，用模拟量来表示神经网络的权重要求对存储元件进行高精度编程，另外，对于某些新兴的存储器件来说，模拟编程过程可能是相对随机的。因此，实现高精度模拟状态编程可能需要多个周期才能完成，对于需要频繁重新编程的应用而言，这可能非常耗时且能效很低。对于这些应用，编程电路和算法的优化至关重要。

另一方面，在存储器件制造过程中，面积（密度）是第一考量因素。在晶体管尺寸能够保持功能的前提下，以尺寸更小为优化方向。加上工艺本身的各种偏移，RAM 单元在实际制造过程中具有极高的不匹配性，为了弥补不同 RAM 单元特性不匹配的短板，在完成多比特计算时往往需要额外的抗失调与失配补偿电路，或者通过在线学习更新神经网络参数的方法来进行补偿。此外，为了接收来自传统的数字电路的信号并将结果传回数字系统，系统需要快速和高能效的信号转换电路（包括 DAC 和 ADC）。对于基于欧姆定律和基尔霍夫定律的矢量矩阵乘法，输入通常采用电压信号形式，而输出结果则是电流信号。在很大的测量范围内精确测量电流值，也是一个需要解决的问题。



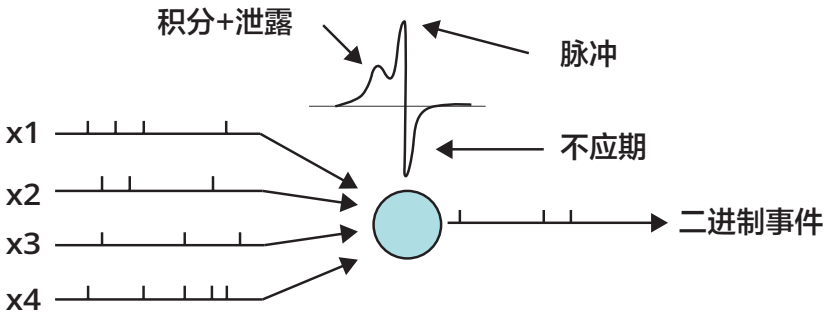
8 神经形态芯片

神经形态芯片（Neuromorphic Chip）采用电子技术模拟已经被证明了的生物脑的运作规则，从而构建类似于生物脑的电子芯片，即“仿生电脑”。其与神经形态工程含义类似。神经形态工程（Neuromorphic Engineering）在 1980 年代晚期由加州理工学院教授卡弗 · 米德 (Carver Mead) 提出，指利用具有模拟电路 (Analog Circuits) 的超大规模集成电路 (Very-large-scale Integration, VLSI) 来模拟生物的神经系统结构。近些年神经形态计算也用来指采用模拟、数字、数模混合 VLSI 以及软件系统实现的神经系统模型。受到脑结构研究的成果启发，研制出的神经形态芯片具有低功耗、低延迟、高速处理、时空联合等特点。

8.1 神经形态芯片的算法模型

广义上来讲，神经形态计算的算法模型可以大致分为人工神经网络 (Artificial Neural Network, ANN)、脉冲神经网络 (Spiking Neural Network, SNN)，以及其他延伸出的具有特殊数据处理功能的模型。其中 ANN 是目前机器学习特别是深度学习使用的主要模型，是本文其它部分讨论的主要内容。因此，本节主要讨论 SNN 算法。这类算法具有几个特征：第一，其神经元的输出是具有时间维度编码的脉冲序列；第二，通过膜电位表征时间维度，即膜电位代表了历史上收到的脉冲能量，且发放会导致膜电位的变化，有时序记忆特性。因此，多个神经元可以实现时 – 空二维空间的表达能力。

神经元动态行为的模拟算法较多，一般采用微分动力方程表述，有较好的仿生能力，但不利于硬件实现。因此，基于简化算法，尤其是基于泄漏积分发放模型（Leaky Integrate and Fire，LIF）的方法广受关注。其原理是将连接到本神经元的所有轴突的脉冲，根据突触强度进行加权求和，得到神经元的积分电位，再与之前的膜电位相加并更新，得到新的膜电位。如果膜电位超过设定阈值则发放脉冲，否则不发放。如图表 8–1 所示。 可见，LIF 算法具有将时间和空间信息联合表达的特点。



图表 8–1 积分 – 发放模型工作示意图

SNN 最常见的学习基础算法是依赖神经元脉冲时序而定的可塑性（ Spike-timing-dependent Plasticity, STDP ）。STDP 是已经在生物脑中得到验证、较为有效的训练算法。作为一种局部训练的、非监督的非反向传播算法，它不能保证一定能够训练出性能优异的网络。另外，由于 SNN 网络训练比较困难，因此在实际应用中，也会采用反向传播算法进行训练。SNN 的权重精度一般不高，而且涉及大量的乘累加操作，低精度权重和低精度激活值的 ANN 网络比较适合于移植到 SNN 中。一个较为粗糙的对比是，对于 MLP 或 CNN 等前馈型网络拓扑结构，基于 LIF 模型的 SNN 网络与采用二值激活神经元和 ReLU 激活函数的 ANN 网络类似。换言之，在没有历史记忆情形下，SNN 与 ANN 具有一定程度的等价性。SNN 与 ANN 的对比见图表 8–2 所示。

类别	ANN（人工神经网络、深度学习算法）	SNN（脉冲神经网络算法）
神经元激活值	不带时间轴的多值（定点或浮点数）	脉冲串（带有时间的二值）
时序表达方法	RNN 等网络中的回环结构	膜电位和网络回环
空间表达方法	通常是较规则互联的神经元阵列，处理图像通常采用滑窗卷积操作	可非规则互联的神经元，一般没有滑窗过程（需要并行化展开）
激活函数	多使用非线性激活函数	没有激活函数
推理	卷积、池化、多层感知器模型（MLP）等	泄露积分发放模型（LIF）等
训练	反向传播较流行	STDP，Hebb 定律，反向传播
归一化方法	批归一化等	赢者通吃（Winner takes all）
激活负数表示	负数神经元激活值	抑制型神经元
典型传感器	数字相机，麦克风	DVS 相机
理论来源	数理推导	脑启发
共同点	积分过程，MLP 拓扑结构	

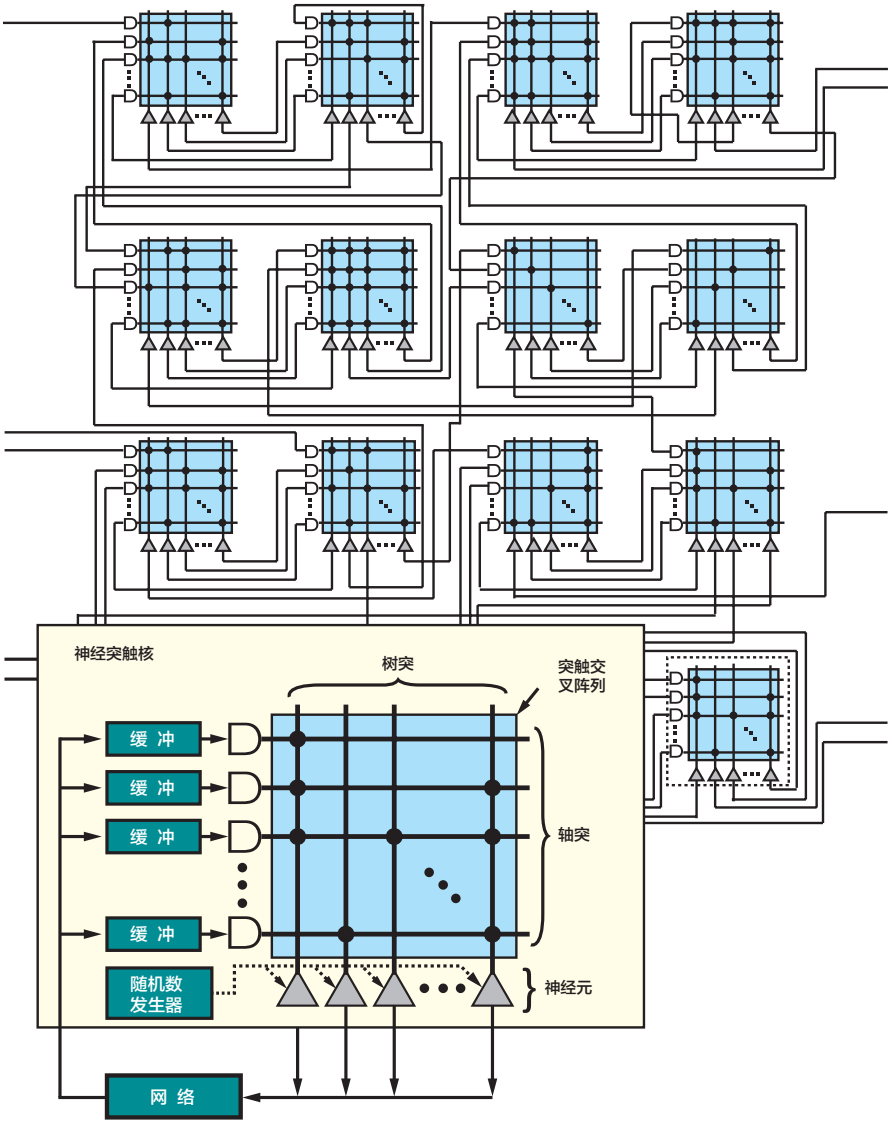
图表 8-2 脉冲神经网络与人工神经网络的对比

8.2 神经形态芯片的特性

8.2.1 可缩放、高并行的神经网络互联

借鉴生物脑的互联结构，神经形态芯片可以实现任意神经元间的互联。即在指定规模的仿生神经网络下，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。如此强大的细粒度互联能力是其他神经网络 / 深度学习芯片目前还无法做到的。

为了实现复杂互联，神经形态芯片通过横纵交叉矩阵（Crossbar）、片上网络（NoC）、芯片外部高互联链路的多层级方案实现。最简单的方案是横纵交叉矩阵，如图表 8-1 所示，在矩阵上交叉处有连接点代表有突触连接，连接的强度用多值权重值表示。但其规模受限，为了扩大规模，可以采用高速、共享的物理链路互联，例如芯片内部的 2D Mesh 网络和芯片与芯片（外部）互联的 SerDes 接口 / 光纤接口。这些互联方案是多神经元分时共享的，而不像生物脑那样，传输链路彼此独立，因此就需要传输的数据携带有目标地址信息，打成数据包在这个共享的链路上传输。神经元级别的传输具有地址的任意性，因此每个神经元目的地址是不同的，发射的数据包采用神经元级别的小包，一般为几十个比特，因此适合采用存储 - 转发方式，如 2D-mesh 网络和高扇入扇出总线。这样的传输网络没有绝对的位置，目的地址采用相对于发射端的偏移位置，因此可以传到任意大的网络，如果路径较远可以用中继核转发。值得注意的是，普通计算机采用内存进行数据交换也可以达到同样的效果，但是需要串行传输，而这种片上网络可以支持成百上千级别的并行传输。



图表 8-3 神经形态芯片体系结构，芯片由大量的神经突触核组成，每个核都具有横纵交叉阵列（Synaptic crossbar）（来自 TrueNorth）

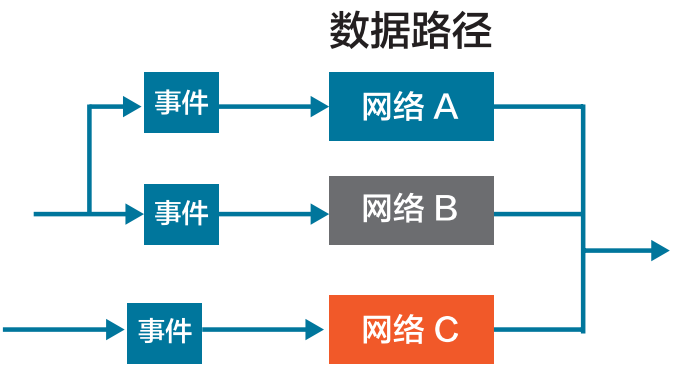
8.2.2 众核结构

由于生物脑中神经元是分簇的，簇内互联较多，簇间较少，因此神经形态芯片中采用分簇结构是可行的，即簇内包括一组神经元，且采用上文所述的交叉阵列与输入的信号（其他神经元的输出轴突）互联，簇间采用例如 2D Mesh 这样的片上网络。簇就成了芯片的功能基础单元，有时也称为核（Core），这个结构也称为众核。该单元主要包括了输入输出路由、突触交叉阵列（实际上是存储器实现的）、积分运算部分和发

放运算部分等。其中输入输出数据可以认为是与外界交互的存储区域，突触存储是簇内私有的，别的簇无法访问。这些簇在地位上是平等的，没有主从关系，整个系统是去中心化并且可缩放规模的。从整体上看来，突触存储可能占据了一半甚至更多的内存，且是散状分布在各个簇内的，跟计算单元离得很近，因此这个结构具有近计算存储或者说近存储计算的特性。其理论源于生物脑的计算 – 存储一体化，优点是可以解决传统冯 · 诺依曼计算机中的“存储墙”问题。

8.2.3 事件驱动

神经芯片的事件驱动也是受脑启发的：生物体神经元接收脉冲，自身膜电位相应改变。没有输入的脉冲，就没有膜电位对应改变的必要（但会有电位的泄漏）。因此电路是否工作取决于是否有包（事件）触发。数字电路的特点往往是后选式的，像处理器中算术逻辑单元（Arithmetic and Logic Unit, ALU），先把所有可能的支路都进行运算，然后选择需要的支路。而事件驱动是前驱式的，一个模块具有很多输入，当输入端收到信息（如脉冲数据包）时，模块就启动计算。因此路由是否传输数据，神经元模块是否运算，取决于是否有信息。这在很大程度上降低了大规模芯片的动态功耗。



图表 8–4 采用事件驱动的前驱式结构

8.2.4 数据流计算

数据流核心观点较多，本文主要参考的观点是将运算和传输用一个有向图（Graph）表示，这个有向图中节点代表运算，边代表传输关系。数据流图就像一个具有各种分支，合并复杂工序加工流水线，让数据通过这个系统得到输出。每一个节点是否运算，取决于其前继节点的运算结果是否已经就位，后继节点的输入线路是否有能力接受新的数据。如果满足，则进行运算。它是无中心、动态众核计算的较快方式，具有众核计算的友好性。数据流计算的不足是控制逻辑简单，不容易表示循环递归之类的结构，然而在神经网络这个范畴它反而显得很合适。因此，在 Tensorflow 等神经网络框架中，默认采用数据流的方式表达神经网络的连接关系。

8.3 机遇与挑战

目前神经形态芯片的设计方法主要分为基于传统 CMOS 技术的神经形态计算电路和基于新型纳米器件的神经形态计算电路。传统 CMOS 技术发展相对比较成熟，如在第 7 章中提到的 IBM TrueNorth 芯片是异步 – 同步混合（无全局时钟）数字电路的代表作，清华大学的天机系列芯片实现了纯同步数字电路的神经形态芯片；瑞士苏黎世联邦理工学院的 ROLLS 芯片和海德堡大学的 BrainScaleS 则是模拟集成电路的代表作品。而基于新型纳米器件的神经形态计算电路目前最受关注的方向是利用忆阻器等器件搭建的神经形态芯片（详见第 7 章）。

神经形态芯片在智能城市、自动驾驶的实时信息处理、人脸深度识别等领域都有出色的应用。如 IBM TrueNorth 芯片可以用于检测图像中的行人、车辆等物体，且功耗极低（65mW）。它也可被用于语音、图像数据集识别等任务，准确性不逊于 CNN 加速器芯片。此外，在线学习能力也是神经形态芯片的一大亮点。研究人员已证明，与其他典型的 SNN 网络相比，在解决 MNIST 数字体识别问题上，英特尔 Loihi 芯片将学习速度提高了 100 万倍 [Davies18]。

在传统 CMOS 工艺下，神经形态芯片的物理结构较为成熟，但对于可以仿真大规模神经网络而言（如大于人脑 1% 规模的系统而言），仍存在很多挑战，包括（1）散热问题将导致单芯片规模无法继续增长，片上存储和积分计算单元的密度不够，导致集成的突触和神经元数量无法继续提升，功耗居高不下。（2）由于其阵列众核的特性，在片上、跨芯片、跨板、多机等尺度下的互联和同步问题突出。（3）为了提升密度，大多 ASIC 芯片可模拟的神经形态算法过于单一或简化，缺乏灵活性和模仿真实生物神经元的能力。

对基于忆阻器交叉阵列的神经形态芯片而言，在很长一段时间内，都需要针对具体的应用需求并结合能效目标，对交叉阵列的规模、突触的连接方式和漏电流的控制等不断进行优化。

除了上述物理结构问题，神经形态芯片在软件，特别是算法上，同样面临巨大的挑战。几年前，深度学习、CNN 技术未得到有效发展时，ANN 也面临类似的问题，但随着基于反向传播算法的深度卷积神经网络的不断进步，这个问题得到了很大程度的缓解。而目前，SNN 相关算法大都处于研究阶段。主要根源在于人们对生物脑的运作机理挖掘还不充分，受启发有限。值得欣慰的是，近几年脑图谱技术发展迅速，绘制详细的动物全脑神经元级别静态连接结构已逐渐成为现实。动态运行的脑机制探测，如动物的低级视觉皮层信息解译也取得了长足的进步。相信这些脑科学的实验证据将大大助力神经形态计算取得突破。

9 AI 芯片基准测试和发展路线图

目前，学术界和工业界都有很多团队研究和开发针对 AI 应用的芯片，我们无疑会看到越来越多的 AI 芯片出现。在这种 AI 芯片的开发热潮中，两个重要的工作是必不可少的：客观地评估和比较不同的芯片（即基准测试，Benchmark），以及可靠地预测 AI 芯片的发展路径（即路线图，Roadmap）。

基准测试旨在提供统一的方法学来评估和比较不同 AI 应用的芯片。如前面章节所讨论的，目前已经出现了各种架构和技术来实现各种形式的神经形态计算和机器学习加速。一方面，从 CMOS 工艺、电路到架构的各个层面都可以在定制芯片或专用加速器中被优化，以执行神经网络的数学计算和近似（例如，快速矩阵运算），其在计算和能耗效率方面可以显著超越通用 CPU。然而，就效率和能力而言，这些系统仍然比生物大脑差几个数量级。另一方面，神经网络的直接实现可能有赖于能够模仿神经元和突触行为的材料和器件。例如，突触行为已经在金属 – 绝缘体 – 转变器件、相变存储器、基于细丝形成或氧迁移的氧化物电阻开关、自旋转移力矩器件，以及铁电隧道结等中得到了展示。针对这些材料、器件和体系结构，清楚定义一组性能要求和量化参数，对于进行基准测试和指导研究方向，是非常重要的。

技术的进步通常从具有适度功能的较小系统开始，逐步扩展到能够解决复杂问题的大系统。一个基于技术、设计或应用的共性明确路线图不仅可以提供衡量技术发展的指标，还有助于确定研究差距和关键挑战。在 CMOS 技术的基准测试和发展蓝图中，技术选项（比如场效应晶体管）和通用性（比如晶体管特征尺寸）是有达成一致的明确定义的。与此不同，AI 各种各样的应用、算法、体系结构、电路和器件对确定基准和路线图的共同基础提出了巨大的挑战。

基于通用芯片设计领域的经验，我们普遍认为不太可能找到普适的“最佳”器件、架构或算法。例如，在冯 · 诺依曼架构下，CMOS 器件似乎很难被新兴器件击败 [Nikonov15, Nikonov13]，而一些新兴器件（例如自旋电子器件）可能在非布尔架构中表现得更好，如用于细胞神经网络 [Pan16]。另一个使得设计 AI 芯片的基准测试更有挑战性的原因是，除了需要考虑与计算本身相关的能量、性能和准确度之外，还必须考虑其他操作，比如，由于诸如输入、输出和存储器访问等而导致的性能和能量开销。这对于非冯 · 诺依曼硬件来说尤其困难，因为真正的公平对比必须同时考虑最先进的冯 · 诺依曼硬件平台（CPU、GPU）和相关的算法。此外，随着工艺进步和创新引入未来的芯片，上述平台的性能也将发生变化。最后，在 AI 领域，不论是理论研究还是应用需求，都在不断引入新的算法，如神经网络的结构和计算模型，对基准测试和路线图的探讨必须考虑到所有这些因素。

目前，我们还没有看到任何公开、全面的针对 AI 芯片的基准测试工作。业界对于 AI 芯片的评估主要靠运行一些常见的神经网络以及其中使用较多的基本运算来进行，比如由百度提出 Deepbench[Baidu]。美国自然科学基金委（NSF）、美国国防高级研究计划局（DARPA）和半导体研究联盟（SRC）资助的几个研究计划也认识到这项工作的重要性，EXCEL 中心的研究人员（由美国 NSF 和 SRC 资助）正在积极研究非冯 · 诺依曼硬件的基准测试方法学，比如针对 MNIST 数据集 [EXCEL] 的任务。

为了应对面向 AI 应用的硬件基准测试的相关挑战，我们需要收集一组架构级功能单元，确定定量和定性的优值（Figures of Merits, FoM）并开发测量 FoM 的统一方法。

神经形态计算的材料和器件需要具备：1）多态行为，能够根据过去的历史决定当前状态值；2）低能耗，能以很低的能耗从一种状态切换到另一种状态；3）非易失性：无需刷新就可以保持状态的属性；4）阈值行为：受到重复激励后可以剧烈地改变某些属性；5）容错性。

判断一颗基于某种特定器件工艺、电路形式和体系结构而实现的芯片好坏，在很大程度上取决于它针对的具体应用和算法 / 模型。为了对各种各样的器件进行基准测试，有必要明确目标应用、适用的算法和模型以及电路设计等信息。只有提供足够详细的信息，才可以既不限制选择范围，又同时明确器件需求。NRI 纳电子研究计划（Nanoelectronics Research Initiative）进行的超越 CMOS 器件的基准测试研究使用反相器、NAND 门和加法器作为标准布尔逻辑电路模块来比较所有器件 [Nikonov15，Nikonov13]。AI 芯片的基准测试还需要定义具有可量化参数的通用功能单元。许多神经网络使用的卷积、池化和激活函数等功能，可能是合适的功能单元，随着新的算法和计算模型的出现，基准测试中也可能会引入其他的附加功能单元。在架构级别上，操作 / 秒 / 瓦特（operations/second/watt）和吞吐量可以是两个对系统性能互补的测量。如何确定其他神经形态计算模型，如脉冲神经网络，也必须进行研究，因为它们经常带有与其他标量神经网络不同的计算形式。

目前，有一些神经形态器件的定量参数被提出和评估，包括调制精确度（如阻抗水平）和范围（如开关率）、线性度、不对称性及变异性等。它们对于对神经网络性能表征都非常关键。一些通常用于布尔逻辑的器件参数对于神经形态计算也很重要，包括尺寸、速度、操作电压 / 电流、功率 / 能量、可靠性、耐久性和良率等，这些参数之间的权衡需要仔细评估。算法准确度是人工智能应用的一个关键考虑因素，也应该包含在 FoM 中。

为 AI 芯片开发统一的基准测试方法，可以利用 NRI 赞助研究工作所获得的知识和经验。虽然这项工作主要集中在布尔逻辑单元（NAND 门和 ALU），但它为更高级的架构级基准奠定了基础。例如，在 [Perricone17] 中，作者扩展了 [Nikonov15] 的工作，针对执行并行工作负载的多核处理器，在架构层面对几种新型器件进行基准测试。他们提出的分析框架基于新型设备来预测多核处理器的性能和能耗，扩展这个框架来处理非冯 · 诺依曼体系结构，将需要新的、基于适当功能单元的性能和能耗模型。AI 芯片的测试基准和路线图必须超越器件和电路（即加法器，乘积累积单元）的层次，以量化各种因素同时发挥作用将如何提高计算原操作（例如凸优化），以及应用级任务的能量 / 性能 FoM（例如图像分析）。这项工作需要算法研究人员、架构师、电路设计人员和器件专家共同努力才能很好地完成。

10 展望未来

在目前这个时间点，人工智能芯片还处在它的“婴儿期”，未来充满了不确定性。我们唯一可以肯定的是，它是整个人工智能技术发展的基础，也将是推动整个半导体领域技术进步的最重要的力量之一。如本文所介绍，在全球范围，从科技巨头到初创公司，投入了巨大资源，在包括架构、电路、器件、工艺和材料等各个层面展开探索。这些努力很有可能把半导体技术提升到新的高度，惠及科学技术的整体发展。但在 AI 芯片大热的背景下，我们更应该清醒地认识未来面临的挑战并积极应对。

首先，人工智能技术的整体发展还处在初级阶段，整个人工智能芯片行业的发展也随之面临极大的不确定性。目前，人工智能领域取得的主要进展是基于深度神经网络的机器学习，更擅长解决的是感知问题，虽然它已经在一些特定任务上取得了超过人类的成绩，但在认知问题上还处于摸索阶段，离我们追求的、潜在社会期待的所谓通用人工智能还有巨大差距，那需要计算能力和计算系统的能源效率比现在至少提高几个数量级。更准确地说，我们甚至不知道目前的主流技术路径，是否是通向通用人工智能的正确道路。“未来能否有一个终极算法来实现通用人工智能？”这个问题还没有人能给出肯定的答案。芯片是人工智能算法的物理基础，它与算法唇齿相依。如果能有统一的终极算法出现，那么我们很可能会看到一个终极芯片出现。但在未来很长一段时期，不同的应用仍然需要不同的算法（也包括传统算法），因此我们还必须探索不同的架构，探索新的器件甚至材料。随着底层芯片技术的进步，人工智能算法也将获得更好的支持和更快的发展。

一些新的器件和材料的出现，也许会让更多新的算法和模型（比如类脑计算）变得更加成熟和高效，也让我们有更多的机会和路径去探索通用人工智能。而在这一过程中，人工智能本身也很有可能被用于研发新的芯片技术，形成算法和芯片相互促进的良性循环局面。可以说，目前人工智能技术的不确定性，给各种技术创新提供了一个巨大的舞台，在最终收敛之前，我们可以期待在这个舞台上看到前所未有的精彩表演。

其次，随着人工智能和物联网的持续快速发展，越来越多的应用需求和应用场景将不断涌现。需求驱动的 AI 芯片技术创新将促进创新链与产业链更加紧密结合，推动开放合作、共享共赢的产业生态形成。CMOS 技术与新兴信息技术的交叉融合，开源软件到开源硬件的潮流渐显，预示着我们将迎来一个前所未有的协同创新机遇期。

总而言之，在终点还非常模糊的情况下，AI 芯片的探索过程将是充满挑战也是令人兴奋的，需要从政策制定者到每一个从业者的共同努力和长期坚持。在保持清醒认识的前提下，促进政策、资金、市场、应用和技术形成一个良性发展的产业环境，我们完全可以对 AI 芯片的发展前景充满信心，也可以期待从技术进步到应用落地所带来的各种红利。

参考文献

[Vivienne17] Sze, Vivienne., et al. "Efficient processing of deep neural networks: A tutorial and survey." Proceedings of the IEEE 2017,105(12): 2295-2329. DOI: 10.1109/JPROC.2017.2761740

[Theis16] T.N. Theis., et al. "The End of Moore's Law: A New Beginning for Information Technology," Invited paper, IEEE Computing in Science & Engineering, 2017,19 (2): 41-50. DOI: 10.1109/MCSE.2017.29

[Aly15] M. M. Sabry Aly., et al. "Energy-Efficient Abundant-Data Computing: The N3XT 1,000X," IEEE Computer, 2015, 48(12): 24 – 33. DOI: 10.1109/MC.2015.376

[Google] N. P. Joupp., et al. "In-datacenter performance analysis of a tensor processing unit". In Proceedings of the 44th Annual International Symposium on Computer Architecture(ISCA) 2017, 1-12. DOI: 10.1145/3079856.3080246

[Bankman18] D. Bankman., et. al. "An Always-On 3.8μJ/86% CIFAR-10 Mixed-Signal Binary CNN Processor with All Memory on Chip in 28nm CMOS," Solid-State Circuits Conference (ISSCC), 2018 IEEE International. IEEE, 2018, 222 – 224. DOI: 10.1109/ISSCC.2018.8310264

[Lee18] J, Lee., et al. "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision" Solid-State Circuits Conference (ISSCC), 2018 IEEE International. IEEE, 2018, 218 – 220. DOI: 10.1109/ISSCC.2018.8310262

[Parashar17] A. Parashar., et al. "SCNN: An accelerator for compressed-sparse convolutional neural networks." 2017 ACM/IEEE ISCA, DOI: 10.1145/3079856.3080254

[Bert17] M, Bert., et al. "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltageaccuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI." Solid-State Circuits Conference (ISSCC), 2017 IEEE International. IEEE, 2017, 246 – 247. DOI: 10.1109/ISSCC.2017.7870353

[Shouyi17] S. Y. Yin., et al. "A 1.06-to-5.09 TOPS/W reconfigurable hybrid-neural-network processor for deep learning applications." VLSI Circuits, 2017 Symposium on. IEEE, 2017, C26 - C27. DOI: 10.23919/VLSIC.2017.8008534

[Shouyi18] S. Y. Yin., et al. "A High Energy Efficient Reconfigurable Hybrid Neural Network Processor for Deep Learning Applications." IEEE Journal of Solid-State Circuits 2018 53(4): 968-982. DOI: 10.1109/JSSC.2017.2778281

[ISSCCTrends2017] http://isscc.org/wp-content/uploads/2017/05/ISSCC2017_TechTrends.pdf

[Lee14] D. U. Lee., et al., "A 1.2V 8Gb 8-Channel 128GB/s High-Bandwidth Memory (HBM) stacked with effective Microbump I/O Test Methods Using 29nm Process and TSV", ISSSCC, 2014, 432–433. DOI: 10.1109/ISSCC.2014.6757501

[Jeddeloh12] J. Jeddeloh., et al., “Hybrid Memory Cube: New DRAM Architecture Increases Density and Performance”, VLSI, 2012, 87 – 88. DOI: 10.1109/VLSIT.2012.6242474

[Nvidia] http://www.nvidia.com/object/what-is-gpu-computing.html

[Chi16] P. Chi., et al., “PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory,” in Proc. ISCA, 2016, 27-39. DOI: 10.1109/ISCA.2016.13

[Chen17] Y. H. Chen., et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." IEEE Journal of Solid-State Circuits 2017,52 (1):127-138. DOI: 10.1109/JSSC.2016.2616357

[Chen18] W. -H. Chen., et al., “A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processor”, ISSCC, 2018, 494-496 DOI: 10.1109/ISSCC.2018.8310400:

[Yang13] J. J. Yang., et al., “Memristive devices for computing”, Nature Nanotechnology, 2013,8, 13-24. DOI: https://doi.org/10.1038/nnano.2012.240

[Yao17] Yao, Peng., et al. "Face classification using electronic synapses." Nature communications 2017,8 15199.

[Li18] C. Li., et al. "Analogue signal and image processing with large memristor crossbars." Nature Electronics 2018,1: 52-59. DOI: 10.1038/s41928-017-0002-z

[Prezioso15] M. F. Prezioso., et al., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors." Nature 2015,521 (7550):61-4. DOI: 10.1038/nature14441

[Hu18] M. Hu, et al., "Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine." Advanced Materirals 2018, 30(9): 1705914. DOI: 10.1002/adma.201705914. (2018)

[Merolla14] P. A. Merolla., et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface." Science 2014,345 (6197):668-673. DOI: 10.1126/science.1254642

[Intel17] https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerateartificial-intelligence/

[Wang17] Z. Wang., et al., "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing." Nature Materials 2017,16 (1):101-108. DOI: 10.1038/nmat4756.

[Wang18] Z. Wang., et al, "Fully memristive neural networks for pattern classification with unsupervised learning", Nature Electronics 2018,1, 137–145. DOI: https://doi.org/10.1038/s41928-018-0023-2

[Jerry17] M. Jerry., et al., "Ferroelectric FET analog synapse for acceleration of deep neural network training." IEEE International Electron Devices Meeting (IEDM), San Francisco, USA. (2017) DOI: 10.1109/IEDM.2017.8268338

[Boyn17] Boyn, S., et al. "Learning through ferroelectric domain dynamics in solid-state synapses." Nature Communications 2017,8:14736. DOI: 10.1038/ncomms14736 (2017).

[Davies18] M. Davies., et al., "Loihi: a Neuromorphic Manycore Processor with On-Chip Learnin", IEEE Micro, 2018, 38(1): 82 – 99. DOI: 10.1109/MM.2018.112130359

[Nikonov15] D. Nikonov., et al. “Benchmarking of beyond-cmos exploratory devices for logic integrated circuits,” Exploratory Solid-State Computational Devices and Circuits, IEEE Journal on, 2015, 1, 3–11. DOI: 10.1109/JXCDC.2015.2418033

[Nikonov13] D. Nikonov., et al. “Overview of beyond-cmos devices and a uniform methodology for their benchmarking,” Proceedings of the IEEE, 2013, 101(12): 2498–2533. DOI: 10.1109/JPROC.2013.2252317

[Pan16] C. Pan., et al. "Non-Boolean Computing Benchmarking for Beyond-CMOS Devices Based on Cellular Neural Network," in IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 2016, 2, 36-43. DOI: 10.1109/JXCDC.2016.2633251

[Baidu] “DeepBench",https://github.com/baidu-research/DeepBench

[EXCEL] https://collectivecomputing.nd.edu/about/

[Perricone17] R. Perricone., et al. “Can Beyond CMOS Devices Illuminate Dark Silicon,” Conference on Design , 2016:13-18.

索引

White Paper on AI Chip Technologies

地址：北京市海淀区荷清路 3 号院 A 座
电话：010-62799552
网站：http://www.icfc.tsinghua.edu.cn
微信公众号：未来芯片技术高精尖创新中心

关注微信公众号