人工智能芯片概述

尹首一
（清华大学微电子学研究所 北京 100084）

摘 要：作为人工智能时代的硬件载体，人工智能芯片的重要性不言而喻。介绍了现有人工智能芯片的种类和各自技术路线，阐述了人工智能芯片的各个发展阶段及其特点，最后展望了未来发展趋势。

关键词：人工智能；可重构计算架构；人工智能芯片
中图分类号：TP391.4 **文献标识码：**A **国家标准学科分类代码：**510.40

Artificial intelligence chips review

YIN Shouyi
（*Institute of Microelectronics, Tsinghua University, Beijing 100084, China*）

Abstract: As the underlying computing hardware of the artificial intelligence era, the importance of artificial intelligence chips is self-evident. The concept and key technologies of artificial intelligence chips are introduced, the development stages and characteristics of artificial intelligence chips are expounded, and the future development trend is forecasted at last.

Keywords: AI; reconfigurable computing architecture; AI chip

0 引 言

人工智能（aritificial intelligence，AI）是一门融合了数学、计算机科学、统计学、脑神经学和社会科学的前沿综合性技术。它的目标是希望计算机可以像人一样思考，替代人类完成识别、分类和决策等多种功能。在2016年AlphaGo击败李世石赢得人机围棋大战后，人工智能引发了全球热潮。与此同时，Google、FaceBook、Amazon、Intel等巨头纷纷成立AI团队，促进人工智能技术的进一步发展。在国内，国务院发布了人工智能发展规划，从国家层面对人工智能加以支持[1]，各类互联网公司和初创公司纷纷投入到人工智能产业。今天，海量数据的形成、深度学习算法的革新、硬件技术的变革、互联网生态的完善助力人工智能产业呈现爆发式发展，而其中以核心人工智能芯片为基础的强大计算力发挥着至关重要的作用[2]。

1 人工智能芯片概况

当前人工智能的主流技术深度神经网络概念早在20世纪40年代就已经被提出，然而几经起落，甚至被90年代中期出现的支持向量机所全面压制。主要原因就是当时没有可以用于大规模并行计算的诸如图形处理器（graphics processing unit，GPU）等芯片

尹首一（通信作者），长聘教授，研究方向为可重构计算、低功耗设计、人工智能芯片设计。E-mail：yinsy@tsinghua.edu.cn

的硬件条件，神经网络的训练仍然耗时太久，训练成本过于高昂。随着摩尔定律的不断演进发展，高性能芯片大幅降低了深度学习算法所需的计算时间和成本，人工智能技术终于在语音识别、计算机视觉等领域取得了重大突破。然而，深度神经网络的计算量在不断膨胀，读写的数据量日趋庞大，网络结构也越来越多样化，这就要求作为硬件基础的人工智能芯片必须不断进行相应的发展，以应对性能、功耗、灵活性这3个方面的挑战。

当前实现人工智能计算的技术路线可概括为3类：第1类是基于冯·诺依曼体系结构的通用处理器，诸如大家所知的CPU、GPU、DSP等都属于这一类型。它以算术逻辑单元为计算核心，由于其通用性需要应对包括分支跳转、中断等复杂的指令处理，需要消耗很多片上资源。因此CPU的并行计算处理能力并不高，此外处理器本身频繁的读取操作会带来大量的访存功耗问题；第2类则是专用集成电路（application specific integrated circuit，ASIC）。它针对特定的计算网络结构采用了硬件电路实现的方式，能够在很低的功耗下实现非常高的能效比。在网络模型算法和应用需求固定的情况下，ASIC是一个不错的选择。但ASIC本身研发的周期很长，通常在1～2年，这就使得ASIC本身存在对算法迭代跟进的风险性问题；第3类是基于可重构架构实现的处理器，该技术是将计算部分设计为可配置的处理单元，并且通过相应的配置信息来改变存储器与处理单元之间的连接，从而达到硬件结构的动态配置目标。深度神经网络因为计算量大、数据量大、结构特点多样，基于冯·诺依曼结构的通用处理器以及专用处理器很难在这样的算法上同时展现出灵活性和高能效，可重构处理器在通用处理器和专用处理器之间做了一定的折中和权衡，可以兼顾智能应用算法中的高性能、低功耗、高灵活度的特点。

2 人工智能芯片发展阶段

近几年来，人工智能技术的热潮如火如荼，随着人工智能产品的大规模落地应用，面向不同场景的各类算法纷纷涌现，计算数据呈爆炸式增长，芯片作为人工智能技术的硬件基础和产业落地的必然载体，吸引了众多巨头和初创公司纷纷入局，各类人工智能芯片陆续面世。针对不同应用场景，不同芯片的处理速度、能耗、支持的算法也各有优势。根据人工智能产业的发展状况和技术成熟度划分，可以分为4个阶段[3]。

2.1 人工智能芯片初级阶段

第一个阶段，人工智能芯片从2016年开始爆发，到目前在架构设计上已经比较稳定，相关的编译器的技术越来越成熟，整个产业格局基本成型。可以说，目前的人工智能芯片软硬件技术已经为大规模商用做好了准备。这类芯片主要采用现有的以CPU、GPU、DSP、FPGA为代表的传统芯片架构来运行深度学习算法，主要部署在云端。

在云端训练环节，深度神经网络的计算量极大，而且数据和运算是可以高度并行的，GPU具备进行海量数据并行运算的能力，并且为浮点矢量运算配备了大量计算资源，与深度学习的需求不谋而合，成为云端训练的主力芯片，以70%以上的市场占有率傲视群雄。但由于GPU不能支持复杂程序逻辑控制，仍然需要使用高性能CPU配合来构成完整的计算系统。

在云端推理环节，计算量相比训练环节少，但仍然涉及大量的矩阵运算。虽然GPU仍有应用，但并不是最优选择，更多的是采用异构计算架构来完成云端推理任务。FPGA提高了芯片应用的灵活性和可编程性，与GPU相比具备更强的计算能力和更低的功耗，在云端加速领域优势明显。在产业应用没有大规模兴起之时，使用这类已有的通用芯片可以避免专门研发ASIC的高投入和高风险，但是，由于这类通用芯片的设计初衷并非专门针对深度学习任务，因而天然存在性能、功耗等方面的瓶颈，随着人工智能应用规模的扩大，这类问题日益突出[4]。

2.2 人工智能芯片发展阶段

新的计算模式往往会催生出新的专用计算芯

片,面对人工智能时代对算力的强大需求,学术界和产业界纷纷提出了自己的解决方案,谷歌(Google)的TPU、麻省理工学院(MIT)的Eyeriss、韩国科学技术院(KAIST)的UNPU和寒武纪的1A则是其中具有代表性的芯片,这类芯片在大规模量产的情况下具备性能更强、体积更小、功耗更低、成本更低等优点。目前一部分通过采用语音识别、图像识别、自动驾驶等算法切入人工智能领域的公司,也希望通过打造匹配算法的定制芯片和产品来实现盈利。

当前深度学习部署呈现出从云到端,赋能边缘的趋势,但应用于云端的人工智能芯片普遍存在功耗高、实时性低、带宽不足、数据传输延迟等问题,难以满足边缘计算的需求。在边缘端进行推理的应用场景较之云端更为多样化,智能手机、可穿戴设备、ADAS、智能摄像头、语音交互、VR/AR、智能制造等边缘智能设备需求各异,需要更为定制化、低功耗、低成本的嵌入式解决方案,这就给了初创公司更多机会,针对不同的细分市场来设计差异化产品。就未来整体市场规模来说,边缘计算芯片在智能终端的带动下将是云端数据中心芯片市场的5倍以上。未来几年,我们应该可以看到“无芯片不AI”的景象,随着人工智能应用场景的逐渐落地,底层技术和硬件方向也更加清晰,随之而来的是各类芯片公司的白热化竞争[5]。

2.3 人工智能芯片进阶阶段

在这一阶段,随着深度学习算法的不断演进,当前的芯片架构难以满足越来越高的算力支持、越来越低的功耗需求和层出不穷的各类算法,架构创新是人工智能芯片的必由之路,而可重构计算架构则是其中最具代表性的技术之一。可重构计算架构是一种介于通用处理芯片和专用集成电路之间的、利用可配置的硬件资源,根据不同的应用需求灵活重构自身的新型体系结构,同时具备通用计算芯片兼容性和专用集成电路高效性的优点,被《国际半导体技术路线图》(2015版)评为“后摩尔”时代最具发展前景的未来通用计算架构技术。该技术也被美国国防部推动的“电子复兴计划”(ERI)列为未来芯片的核心支柱性体系结构技术之一。可重构计算架构天然契合各类人工智能算法对专用计算芯片的需求,同时也能保证算法和硬件的持续演进性,非常适合应用于人工智能芯片的设计当中。采用可重构计算架构之后,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,为人工智能芯片带来了极高的灵活度和适用范围。

美国Wave Computing公司推出的DPU芯片[6]和清华大学微电子学研究所设计的Thinker系列芯片[7]是采用可重构计算架构的代表性工作,相比传统架构,它们具备较强的灵活性和计算能效,同时也具备处理器的通用性和ASIC的高性能和低能耗。

2.4 人工智能芯片未来阶段

在更远的未来,随着算法演进,应用落地,会不断给人工智能芯片提出新的要求,加上底层半导体技术的进步,我们可以期待在3~5年内看到第二次人工智能芯片技术创新的高潮,诸如存内计算芯片、类脑仿生芯片、光子芯片等前沿技术将会从实验室走向产业应用[8]。

现有的人工智能芯片主要采用“存、算分离”的计算架构,即内存访问和计算是分开的,而神经网络同时具有计算密集和访存密集的特点,内存访问的功耗和延迟等问题突出,因此内存成为了处理器性能和功耗的瓶颈。为了解决“存储墙”问题,不少学者提出了存内计算的概念,在内存内直接采用模拟电路实现模拟计算,从而不再需要在处理器和内存之间耗费大量时间和能量移动数据。相比传统的数字电路人工智能芯片,使用存内计算加模拟计算的电路能效比将大幅提高。

类脑仿生芯片的主流理念是采用神经拟态工程设计的神经拟态芯片。神经拟态芯片采用电子技术模拟已经被证明的生物脑的运作规则,从而构建类似于生物脑的电子芯片。神经拟态研究陆续在全世界范围内开展,并且受到了各国政府的重视和支持,

美国的脑计划、欧洲的人脑项目，以及最近中国提出的类脑计算计划等。受到脑结构研究的成果启发，复杂神经网络在计算上具有低功耗、低延迟、高速处理以及时空联合等特点[9]。

硅光子技术目前在数据中心和5G的高速数据传输中获得了越来越多的应用。除此之外，硅光子还可以用来以超低功耗直接加速深度学习计算，把深度学习的两个输入调制到两束光上面，然后让两束光在光子芯片的器件上完成SVD分解和干涉相乘，最后再把光信号转化为数字信号读出结果。最后，这些光器件都可以集成到同一块硅光子芯片上，从而实现高性能光计算模组。

3 人工智能芯片未来趋势

目前全球人工智能产业还处在高速变化发展中，广泛的行业分布为人工智能的应用提供了广阔的市场前景，快速迭代的算法推动人工智能技术快速走向商用，人工智能芯片是算法实现的硬件基础，也是未来人工智能时代的战略制高点，但由于目前的AI算法往往都各具优劣，只有给它们设定一个合适的场景才能最好地发挥它们的作用，因此，确定应用领域就成为发展人工智能芯片的重要前提。但遗憾的是，当前尚不存在适应多种应用的通用算法，因此哪家芯片公司能够抓住市场痛点，最先实现应用落地，就可以在人工智能芯片的赛道上取得较大优势。

架构创新是人工智能芯片面临的一个不可回避的课题。从芯片发展的大趋势来看，现在还是人工智能芯片的初级阶段。无论是科研还是产业应用都有巨大的创新空间。从确定算法、应用场景的人工智能加速芯片向具备更高灵活性、适应性的通用智能芯片发展是技术发展的必然方向，弱监督、自我监督、多任务学习、对大型神经网络表现更好的智慧型芯片将成为学术界和产业界研究的重要目标。计算架构的高度并行和动态可变性，适应算法演进和应用多样性的可编程性，更高效的大卷积解构与复用，更少的神经网络参数计算位宽，更多样的分布式存储器定制设计，更稀疏的大规模向量实现，复杂异构环境下更高的计算效率，更小的体积和更高的能量效率，计算和存储一体化将成为未来人工智能芯片的主要特征[10]。

站在2019年的起点，人工智能芯片的架构创新除了关注神经网络计算，更要关注全芯片的架构创新。以安防智能芯片为例，这是一个典型的系统级问题，除了需要解决神经网络加速问题，还需要处理曝光、白平衡、视频编解码等，并不仅仅是做好一个神经网络加速器就能解决的问题。除了神经网络计算还需要很多计算密集型的模块，这些模块采用什么计算架构，也是整个智能芯片的核心问题。因此，人工智能芯片的架构创新就不能只是神经网络计算架构创新，传统计算架构也必须创新，这将是人工智能芯片架构创新的真正内涵。

参考文献

[1] 中华人民共和国国务院. 新一代人工智能发展规划[Z]. 2017-07-20.
The State Council of the PRC. A new generation of artificial intelligence development planning[Z]. 2017-07-20.

[2] 人工智能产业发展研究课题组. 北京人工智能产业发展白皮书（2018年）[R/OL]. (2018-06-30)[2019-02-28]. http://jxj.beijing.gov.cn/docs/2018-07/20180704102639512942.pdf.
Artificial Intelligence Industry Development Research Group. Beijing artificial intelligence industry development white paper（2018）[R/OL]. (2018-06-30)[2019-02-28]. http://jxj.beijing.gov.cn/docs/2018-07/20180704102639512942.pdf.

[3] YANN L C. Deep learning hardware: past, present, and future[C]// 2019 IEEE International Solid- State Circuits Conference - (ISSCC). IEEE, 2019:12-19.

[4] 朱海鹏. 深度学习硬件:FPGA vs GPU vs ASIC[EB/OL]. (2017- 11- 07) [2019- 02- 30]. https://www.jianshu.com/p/74792ad68a2a.
ZHU H P. Deep learning hardware: FPGA vs GPU vs ASIC[EB/OL]. (2017-11-07)[2019-02-30]. https://www.

jianshu.com/p/74792ad68a2a.

[5] 魏少军. AI芯片发展需要应用和架构创新双轮驱动[C]. GTIC 2018全球AI芯片创新峰会. 上海, 2018.
WEI S J. AI chip developing requirement application and architecture innovation two-wheel drive[C]. GTIC 2018 Global AI Chip Innovation Summit. Shanghai, 2018.

[6] HEMSOTH N. First in-depth view of wave computing's DPU architecture, systems[EB/OL]. (2017-08-23)[2019-03-12]. https://www.nextplatform.com/2017/08/23/first-depth-view-wave-computings-dpu-architecture-systems/.

[7] YIN S Y, YANG P O, TANG S B, et al. A high energy efficient reconfigurable hybrid neural network processor for deep learning applications[J]. IEEE Journal of Solid-State Circuits, 2018, 53(4): 968-982.

[8] 唐杉. AI芯片0.5与2.0[EB/OL]. (2019-02-25) [2019-03-12]. https://mp.weixin.qq.com/s/jpgTCY3cC_AQhBxKznLaOw.
TANG S. AI chip 0.5 and 2.0[EB/OL]. (2019-02-25)[2019-03-12]. https://mp.weixin.qq.com/s/jpgTCY3cC_AQhBxKznLaOw.

[9] 清华大学, 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[R]. 北京:北京未来芯片技术高精尖创新中心, 2018.
Tsinghua University, Beijing Innovation Center for Future Chips. Artificial intelligence chip technology white paper[R]. Beijing: Beijing Innovation Center for Future Chips, 2018.

[10] 尹首一, 郭珩, 魏少军. 人工智能芯片发展的现状及趋势[J]. 科技导报, 2018, 36(17):45-51.
YIN S Y, GUO H, WEI S J. Present situation and future trend of artificial intelligence chips[J]. Science and Technology Guide, 2018, 36(17): 45-51.