人工智能芯片发展的现状及趋势

尹首一,郭珩,魏少军

清华大学微电子学研究所,北京 100083

摘要　人工智能芯片是人工智能技术的重要组成部分,是实现人工智能算法的硬件基础,也是人工智能时代的战略制高点。分析了现有人工智能芯片的种类、特点、技术路线和市场情况,阐述了当前人工智能芯片面临的机遇与挑战,展望了其未来发展趋势。

关键词　人工智能芯片;深度学习;可重构计算

自1956年达特茅斯会议以来,关于人工智能(artificial intelligence, AI)的研究由于受到智能算法、计算速度、存储水平等多方面因素的影响,经历了两起两落的发展,近年来在语音识别、计算机视觉等领域终于取得了重大突破。究其原因,业界普遍认为有三大要素合力促成了这次突破:丰富的数据资源、深度学习算法和充足的计算力支持。丰富的数据资源取决于互联网的普及和随之产生的海量信息;以深度学习为代表的机器学习算法的精确性和鲁棒性越来越好,适用于不同场景的各类算法不断优化完善,具备了大规模商业化应用的潜力[1];而充足的算力则得益于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。

虽然当前摩尔定律逐渐放缓,但作为推动人工智能技术不断进步的硬件基础,未来10年仍将是人工智能芯片(AI芯片)发展的重要时期,面对不断增长的市场需求,各类专门针对人工智能应用的新颖设计理念和架构创新将不断涌现。

1　AI芯片概述

当前对人工智能芯片的定义并没有一个公认的标准。比较通用的看法是面向AI应用的芯片都可以称为AI芯片,按设计思路主要分为三大类:专用于机器学习尤其是深度神经网络算法的训练和推理用加速芯片[2];受生物脑启发设计的类脑仿生芯片;可高效计算各类人工智能算法的通用AI芯片[3]。

为了支持多样的AI计算任务和性能要求,理想的AI芯片需要具备高度并行的处理能力,能够支持各种数据长度的按位、固定和浮点计算;比当前大几个数量级的存储器带宽,用于存储海量数据;低内存延迟及新颖的架构,以实现计算元件和内存之间灵活而丰富的连接。而且所有这些都需要在极低的功耗和极高的能量效率下完成[4]。

在当前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生产周期,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广泛使用,并且可以通过重新配置适应新的AI算法、架构和任务。

2　AI芯片类型及发展情况

加州理工学院Carver Mead最早开始了AI芯片的研究,在20世纪80年代开始研究神经拟态系统(neuro-

收稿日期:2018-08-17;修回日期:2018-08-27

作者简介:尹首一,副教授,研究方向为可重构计算、神经网络计算芯片、高层次综合、SoC与嵌入式系统设计等,电子信箱:yinsy@tsinghua.edu.cn;魏少军(通信作者),教授,研究方向为超大规模集成电路设计方法学、数字系统高层次综合技术、嵌入式系统设计和可重构计算芯片技术,电子信箱:wsj@mail.tsinghua.edu.cn

引用格式: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; doi: 10.3981/j.issn.1000-7857.2018.17.006

morphic electronic systems)[5],利用模拟电路模仿生物神经系统结构。经过30多年的发展,目前已经诞生了不同特色的各类AI芯片,主要包括图形处理器(graphics processing unit,GPU)、现场可编程门阵列(field-programmable gate array,FPGA)、数字信号处理(digital signal processing,DSP)、专用集成电路(application specific integrated circuits,ASIC)、众核处理器、神经拟态芯片等。近年来基于深度学习的图像识别算法和语音识别算法取得了出色的成绩,引起了学术界和工业界的广泛关注,随着谷歌人工智能围棋程序AlphaGo先后战胜李世乭和柯洁,更是把人工智能的热度推向全社会。谷歌这一成绩离不开背后AI加速芯片的贡献,从初代AlphaGo采用CPU+GPU的搭建方案[6],到最新一代AlphaGo Zero采用专用高性能处理器(tensor processing unit,TPU),芯片的变化带来了计算速度的巨大提升和功耗的大幅下降。由此可见针对不同的计算任务,不同类型的AI芯片往往各具优势。

2.1 AI加速芯片

简单地说,AI加速芯片是指以现有芯片架构为基础,对某类特定算法或者场景进行加速,从而实现在这一特定场景下的计算速度、功耗和成本等方面的优化。通常包括基于深度神经网络的各类算法,以及图像识别、视频检索、语音识别、声纹检测、搜索引擎优化、自动驾驶等任务。AI加速芯片的设计主要有两种思路:利用已有的GPU、FPGA、DSP、众核处理器等芯片以异构计算的方式来实现[7];设计专用的ASIC芯片。

2.1.1 GPU

GPU,即图形处理器,是一种由大量核心组成的大规模并行计算架构,专为同时处理多重任务而设计,原本的功能是帮助CPU处理图形显示的任务,尤其是3D图形显示。为了执行复杂的并行计算,快速进行图形渲染,GPU的核数远超CPU,但每个核拥有的缓存相对较小,数字逻辑运算单元也更简单,更适合计算密集型的任务[8]。Intel的GPU主要做为集成显卡使用,应用于Intel的主板和CPU,而Nvidia和AMD则在独立显卡领域更具优势。

深度神经网络的训练过程中计算量极大,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,因此最先被引入运行深度学习算法,成为高性能计算领域的主力芯片之一。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。

2.1.2 FPGA

FPGA是在PAL、GAL、CPLD等可编程逻辑器件的基础上进一步发展的产物。它作为专用集成电路领域中的一种半定制电路出现,既解决了定制电路灵活性上的不足,又克服了原有可编程器件门电路数量有限的缺点。FPGA利用门电路直接运算,速度快,而用户可以自由定义这些门电路和存储器之间的布线,改变执行方案,以期得到最佳效果。FPGA可以采用OpenCL等更高效的编程语言,降低了硬件编程的难度,还可以集成重要的控制功能,整合系统模块,提高了应用的灵活性,与GPU相比,FPGA具备更强的计算能力和更低的功耗[9]。

目前,FPGA的主要厂商Xilinx和被Intel收购的Altera都推出了专门针对AI加速的FPGA硬件和软件工具。而各个主要的云服务厂商,比如亚马逊、微软、阿里云等都推出了专门的云端FPGA实例来支持AI应用。中国刚刚被Xilinx收购的北京深鉴科技有限公司也是基于FPGA来设计深度学习的加速器架构,可以灵活扩展用于服务器端和嵌入式端。

2.1.3 DSP

DSP是一种由大规模集成电路芯片组成的用来完成某种信号处理任务的处理器。DSP善于测量、计算、过滤或压缩连续的真实模拟信号,广泛应用于通信与信息系统、信号与信息处理、自动控制、雷达、航空航天、医疗、家用电器等领域。针对滤波、矩阵运算、FFT(fast Fourier transformation)等需要大量乘加法运算的特点,DSP内部配有独立的乘法器和加法器,从而大大提高了运算速率。

DSP种类繁多,目前应用于AI领域的DSP主要用于处理视觉系统如图像、视频等方面的任务,在自动驾驶、安防监控、无人机和移动终端等领域最为常见。这些DSP中加入了专为深度神经网络定制的加速部件,如矩阵乘和累加器、全连接的激活层和池化层等。由于DSP具有高速、灵活、体积小、低功耗、可编程的特点,非常适合被用在终端设备中,例如手机和摄像头。

2.1.4 众核处理器

众核处理器采用将多个处理核心整合在一起的处理器架构,主要面向高性能计算领域,作为CPU的协处理器存在。众核处理器适合处理并行程度高的计算密

集型任务,如基因测序、气象模拟等。比起GPU,众核处理器支持的计算任务的控制逻辑和数据类型要更加复杂。2000年后,该领域的芯片研究一直很活跃,例如IBM CELL[10]和Kalray MPPA[11]。Intel的至强融核处理器(Xeon Phi)是典型的众核处理器,其中2017年发布的KNL[12]代表了众核处理器的领先水平。

众核处理器的结构能有效地利用现代网络和服务器等应用中较高的线程并行度,虽然芯片面积和功耗会随着内核数量的增加而增加,但性能也随之有效地增加。而增加运算部件和指令发射宽度等技术在增大芯片面积的同时,会拉长信号传输线路,显著增加线延迟,因此众核处理器更适用于数据中心部署的各类AI训练和推理任务。

2.1.5　ASIC

ASIC是一种为专用目的设计的,面向特定用户需求的定制芯片,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、可靠性更高等优点。ASIC分为全定制和半定制。全定制设计需要设计者完成所有电路的设计,因此需要大量人力物力,灵活性好,但开发效率低下,时间成本高昂。如果设计较为理想,全定制能够比半定制的ASIC芯片运行速度更快。半定制使用库中标准逻辑单元,设计时可以从标准逻辑单元库中选择门电路、加法器、比较器、数据通路、存储器甚至系统级模块和IP核,这些逻辑单元已经布局完毕,而且设计得较为可靠,设计者可以较方便地完成系统设计。

近年来越来越多的公司开始采用ASIC芯片进行深度学习算法加速,其中表现最为突出的是Google的TPU。TPU的主要模块包括24 MB的局部内存、6 MB的累加器内存、256×256个矩阵乘法单元、非线性神经元计算单元,以及用于归一化和池化的计算单元[13]。TPU比同时期的GPU或CPU平均提速15~30倍,能效比提升30~80倍。中国的北京寒武纪科技有限公司、北京比特大陆科技有限公司、北京地平线信息技术有限公司等公司也都推出了用于深度神经网络加速的ASIC芯片。目前基于DNN的算法还没有统一标准,而且算法还在不断快速演进,所以ASIC的设计需要保持一定的可编程性,采取软硬件协同设计。

2.2　类脑仿生芯片

当今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,即“仿生电子脑”。神经拟态主要指用包括模拟、数字或模数混合超大规模集成电路VLSI(也包括神经元或者神经突触模型的新型材料或者电子元器件研究)和软件系统实现神经网络模型,并在此之上构建智能系统的研究。神经拟态工程发展成为一个囊括神经生物学、物理学、数学、计算机科学和电子工程的交叉学科。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,如美国的脑计划、欧洲的人脑项目,以及中国的类脑计算计划等。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、时空联合等特点。

目前神经拟态芯片的设计方法主要分为非硅和硅技术。非硅主要指采用忆阻器等新型材料和器件搭建的神经形态芯片,还处于研究阶段。模拟集成电路的代表是瑞士苏黎世联邦理工学院的ROLLS芯片和海德堡大学的BrainScales芯片。数字集成电路又分为异步同步混合和纯同步两种。其中异步(无全局时钟)数字电路的代表是IBM的TrueNorth,纯同步的数字电路代表是清华大学的天机系列芯片。另外,对于片上自学习能力,最近Intel推出了Loihi芯片,带有自主片上学习能力,通过脉冲或尖峰传递信息,并自动调节突触强度,能够通过环境中的各种反馈信息进行自主学习。中国的上海西井信息科技有限公司也成功制备了带有片上学习能力的芯片[14]。

2.3　通用AI芯片

现今的AI芯片在某些具体任务上可以大幅超越人的能力,但究其通用性与适应性,与人类智能相比差距甚远,大多处于对特定算法的加速阶段。而AI芯片的最终成果将是通用AI芯片,并且最好是淡化人工干预的自学习、自适应芯片。因此未来通用AI芯片应包含以下特征[15]。

1)可编程性:适应算法的演进和应用的多样性。

2)架构的动态可变性:能适应不同的算法,实现高效计算。

3)高效的架构重构能力或自学习能力。

4)高计算效率:避免使用指令这类低效率的架构。

5)高能量效率:能耗比大于5 Tops/W(即每瓦特进行5×10^{12}次运算)。

6)低成本低功耗:能够进入物联网设备及消费类电子中。

7）体积小:能够加载在移动终端上。

8）应用开发简便:不需要用户具备芯片设计方面的知识。

目前尚没有真正意义上的通用AI芯片诞生,而基于可重构计算架构的软件定义芯片(software defined chip)或许是通用AI芯片的出路。软件定义芯片顾名思义就是让芯片根据软件进行适应与调整,简单来说就是将软件通过不同的管道输送到硬件中来执行功能,使芯片能够实时地根据软件、产品、应用场景的需求改变架构和功能,实现更加灵活的芯片设计[16]。沿用这种架构设计出来的芯片,可以让芯片的计算能力按照软件的需求来调整适应,而不是沿用传统芯片设计的刚性架构,让应用适应架构。

可重构计算技术允许硬件架构和功能随软件变化而变化,兼具处理器的通用性和ASIC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[17]。清华大学微电子学研究所设计的AI芯片Thinker,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[18]。Thinker芯片通过以下3个层面的可重构计算技术,实现软件定义芯片。

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 bit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 bit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 bit)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芯片的设计当中。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芯片带来极高的灵活度和适用范围。Thinker团队最新推出的Thinker 2人脸识别芯片,能够做到6 ms人脸识别(iPhone X为10 ms),准确率超过98%[19];以及Thinker S语音识别芯片,不仅功耗只有200 μW,只需要节7号AAA电池就运行1年,而且可以进行声纹识别[20]。《MIT Technology Review》2018年初在一篇专稿中评论了Thinker团队的工作,认为这是中国取得的顶级成就[21]。

3 AI芯片市场现状

2018年全球AI芯片市场规模预计将超过20亿美元,随着包括谷歌、Facebook、微软、亚马逊以及百度、阿里、腾讯在内的互联网巨头相继入局,预计到2020年全球市场规模将超过100亿美元,其中中国的市场规模近25亿美元[22],增长非常迅猛,发展空间巨大。目前全球各大芯片公司都在积极进行AI芯片的布局。在云端,Nvidia的系列GPU芯片被广泛应用于深度神经网络的训练和推理。Google TPU通过云服务Cloud TPU的形式把TPU开放商用,处理能力达到180 Tflop,提供64 GB的HBM内存,2400 Gbit/s的存储带宽[23]。老牌芯片巨头Intel推出了Nervana™ Neural Network Processors (NNP),该系列架构还可以优化32 GB HBM2,1 Tbit/s带宽和8 Tbit/s访问速度的神经网络计算[24]。而初创公司如Graph core、Cerebras、Wave computing、寒武纪、比特大陆等也加入了竞争的行列,陆续推出了针对AI的芯片和硬件系统。

然而对于某些应用,由于网络延迟、带宽和隐私问题等各类原因,必须在边缘节点上执行推断。例如,自动驾驶汽车的推断,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[25];大型城市动辄百万的高清摄像头,其人脸识别如果全部交由云端完成,高清录像的数据传输会让通信网络不堪重负。未来相当一部分人工智能应用场景中,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。业界需要专门设计的AI芯片,赋予设备足够的能力去应对未来越发增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终端设备,包括三星、苹果、华为、高通、联发科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,比如北京中科寒武纪科技有限公司的1A处理器[26]、北京地平线信息技术有限公司的旭日处理器[27]、北京深鉴科技有限公司的DPU[28]等。传统的IP厂商,包括ARM、Synopsys、Cadence等公司也都为包括手机、平板电脑、智能摄像头、无人机、工业和服务机器人、智能音箱等边缘计算设备开发专用IP产品。此外在终端应用中还蕴藏着智慧物联网这一金矿,AI芯片只有实现从云端走向终端,才能真正赋予"万物智能"。

4　AI芯片未来趋势

在AI芯片领域,目前还没有出现一款CPU类的通用AI芯片,人工智能想要像移动支付那样深入人心,改变社会,可能还差一个"杀手"级别的应用。无论是图像识别、语音识别、机器翻译、安防监控、交通规划、自动驾驶、智能陪伴、智慧物联网等,AI涵盖了人们生产生活的方方面面,然而距离AI应用落地和大规模商业化还有很长的路要走。而对于芯片从业者来讲,当务之急是研究芯片架构问题。软件是实现智能的核心,芯片是支撑智能的基础。当前AI芯片发展,短期内以异构计算为主来加速各类应用算法的落地;中期要发展自重构、自学习、自适应的芯片来支持算法的演进和类人的自然智能;长期则朝通用AI芯片的方向发展。

4.1　通用AI计算

AI的通用性实际包括2个层级:第一个层级是可以处理任意问题;第二个层级是同一时间处理任意问题。第一层级的目标是让AI的算法可以通过不同的设计、数据和训练方法来处理不同的问题。例如,利用现在流行的深度学习方法训练AI下棋、图像识别、语音识别、行为识别、运动导航等。但是,不同的任务使用不同的数据集来独立训练,模型一旦训练完成,只适用于这种任务,而不能用于处理其他任务。所以,可以说这种AI的算法和训练方法是通用的,而它训练出来用于执行某个任务的模型是不通用的。第二层级的目标是让训练出来的模型可以同时处理多种任务,就像人一样可以既会下棋,又会翻译,还会驾驶汽车和做饭。这个目标更加困难,目前还没有哪一个算法可以如此全能[29]。

4.2　通用AI芯片

通用AI芯片就是能够支持和加速通用AI计算的芯片。关于通用AI的研究希望通过一个通用的数学模型,能够最大限度概括智能的本质。目前比较主流的看法是系统能够具有通用效用最大化能力:即系统拥有通用归纳能力,能够逼近任意可逼近的模式,并能利用所识别到的模式取得一个效用函数的最大化效益。这是很学术化的语言,如果通俗地说,就是让系统通过学习和训练,能够准确高效地处理任意智能主体能够处理的任务。通用AI的难点主要有2个:通用性,包括算法和架构;实现复杂程度。当前,摩尔定律的逐渐失效和冯·诺伊曼架构的瓶颈这2个巨大的技术挑战也是通用AI芯片需要考虑的问题。想要解决这2个问题仅通过芯片的设计理念和架构创新是行不通的,还需要取决于更先进的制程工艺、新型半导体材料、新型存储器件以及人类对于自身大脑更进一步的认知[30]。

5　AI芯片面临的机遇与挑战

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,AI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其作用,因此,确定应用领域就成为发展AI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,人工智能的"杀手"级应用还未出现,已经存在的一些应用对于消费者的日常生活来说也非刚需,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AI芯片面临的一个不可回避的课题。需要回答一个重要问题:是否会出现像通用CPU那样独立存在的AI处理器?如果存在的话,它的架构是怎样的?如果不存在,目前以满足特定应用为主要目标的AI芯片就一定只能以IP核的方式存在,最终被各种各样的SoC(system-on-a-chip)所集成。这无疑带来了新的问题,芯片的体积和功耗是必须要考虑的重要因素,传统芯片公司在SoC的设计优化和工程实现上无疑比

以算法起家的初创AI芯片公司更具经验。

从芯片发展的大趋势来看,现在还是AI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AI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。未来2年之内AI芯片产业将持续火热,公司扎堆进入,但是到了2020年前后,则将会出现一批出局者,行业洗牌开始,最终的成功与否则将取决于各家公司技术路径的选择和产品落地的速度[31]。

参考文献(References)

[1] 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)[2018-07-01]. http://jxw.beijing.gov.cn/docs/2018-07/20180704102639512942.pdf.
Research Group on the Development of Artificial Intelligence Industry. Beijing artificial intelligence industry development white paper (2018) [R/OL]. (2018-06-30)[2018-07-01]. http://jxw.beijing.gov.cn/docs/2018- 07/20180704102639512942.pdf.

[2] 朱晶. 对国内人工智能芯片产业格局的观察[EB/OL]. [2018-03- 05]. https://mp.weixin.qq.com/s/f_4NZB7XwoGlNtsJOT4HQQ?tdsourcetag=s_pctim_aiomsg.
Zhu Jing. Observation on the structure of AI chip industry in China[EB/OL]. [2018-03-05]. https://mp.weixin.qq.com/s/f_4NZB7XwoGlNtsJOT4HQQ?tdsourcetag=s_pctim_aiomsg.

[3] 中国电子技术标准化研究院. 人工智能标准化白皮书(2018版)[R/OL]. [2018-03-31]. http://www.cesi.ac.cn/images/editor/20180124/20180124135528742.pdf.
China Electronics Standardization Institute. Artificial intelligence standardization white paper (2018 Edition)[R/OL]. [2018-03-31]. http://www.cesi.ac.cn/images/editor/20180124/20180124135528742.pdf.

[4] 吴军宁. AI芯片格局最全分析[EB/OL]. (2018-04-01)[2018-07-30]. http://www.sohu.com/a/226935100_132567.
Wu Junning. The most complete analysis of AI chip pattern[EB/OL]. (2018-04-01)[2018-07-30]. http://www.sohu.com/a/226935100_132567.

[5] Mead C. Neuromorphic electronic systems[J]. Proceedings of the IEEE, 1990, 78(10): 1629-1636.

[6] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[7] Gaster B, Howes L, Kaeli D R, et al. Heterogeneous computing with OpenCL: Revised OpenCL 1.2 edition[M]. San Francisco: Morgan Kaufmann Publishers Inc., 2012.

[8] 韩俊刚, 刘有耀, 张晓. 图形处理器的历史现状和发展趋势[J]. 西安邮电大学学报, 2011, 16(3): 61-64.
Han Jungang, Liu Youyao, Zhang Xiao. GPU: The past, present and future[J]. Journal of Xi'an University of Posts and Telecommunications, 2011, 16(3): 61-64.

[9] Jeff Dorsch. 现场可编程门阵列FPGA芯片及其应用[J]. 集成电路应用, 2018(1): 77-79.
Jeff Dorsch. FPGAs for all seasons[J]. Applications of IC, 2018 (1): 77-79.

[10] Gschwind M, Hofstee H P, Flachs B, et al. Synergistic processing in cell's multicore architecture[J]. IEEE Micro, 2006, 26(2): 10-24.

[11] De Dinechin B D, Ayrignac R, Beaucamps P E, et al. A clustered manycore processor architecture for embedded and accelerated applications[C]//High Performance Extreme Computing Conference. Piscataway NJ: IEEE, 2013, doi: 10.1109/HPEC.2013.6670342.

[12] Sodani A. Knights landing (KNL): 2nd Generation Intel® Xeon Phi processor[C]//2015 IEEE Hot Chips 27 Symposium (HCS). Piscataway NJ: IEEE, 2015, doi: 10.1109/HOTCHIPS.2015.7477467.

[13] Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit[C]//2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA). Piscataway NJ: IEEE, 2017: 1-12.

[14] 中国类脑强人工智能初创公司——西井科技实现全球首次"片上学习"[EB/OL]. [2018-03-31]. http://www.sh.chinanews.com.cn/kjws/2016-09-17/11052.shtml.
The westwell realize the first "on-chip learning" in the world [EB/OL]. [2018-03-31]. http://www.sh.chinanews.com.cn/kjws/2016-09-17/11052.shtml.

[15] 魏少军. 从IA到AI,我们还要走多远[C]. 2018全球人工智能与机器人峰会, 深圳, 2018-07-01.
Wei Shaojun. How far do we need to go from IA to AI[C]. 2018 Global Artificial Intelligence and Robotics Summit, Shenzhen, July 1, 2018.

[16] Halfhill T R. XMOS重新定义晶圆-软件定义芯片挑战ASIC、ASSP以及FPGA[J]. 电子产品世界, 2007(10): 80-84.
Halfhill T R. XMOS redefines siliconsoftware-defined chips attack ASICs, ASSPs, FPGAs[J]. Electronic Engineering & Product World, 2007(10): 80-84.

[17] Hartenstein R. A decade of reconfigurable computing: A visionary retrospective[C]//Proceedings Design, Automation and Test in Europe. Conference and Exhibition 2001. 2001: 642-649.

[18] Yin S, Ouyang P, Tang S, 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.

[19] Yin S Y, Ouyang P, Tang S, 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.

[20] Yin S Y, Zheng S X, Song D D. A 141 μW, 2.46 pJ/neuron binarized convolutional neural network based self-learning speech recognition processor in 28 nm CMOS[C]. 2018 Symposia on VLSI Technology and Circuits, Honolulu,June 18-22, 2018.

[21] Sun Y T. China wants to make the chips that will add AI to any gadget[EB/OL]. (2018-01-24)[2018-03-31]. https://www.technologyreview.com/s/609954/china-wants-to-make-the-chips-that-will-add-ai-to-any-gadget/.

[22] 中国电子学会. 新一代人工智能发展白皮书(2017年)[R/OL]. (2018-03-05)[2018-06-30]. http://www.199it.com/archives/694966.html.
Chinese Society of Electronics. White paper on the development of the new generation of artificial intelligence (2017)[R/OL]. (2018-03-05)[2018-06-30]. http://www.199it.com/archives/694966.html.

[23] Barrus J. Cloud TPU machine learning accelerators now available in beta[EB/OL]. (2018-02-12)[2018-03-31]. https://chinagdg.org/2018/02/cloud-tpu-machine-learning-accelerators-now-available-in-beta/.

[24] Naveen R. Intel® Nervana™ neural network processors (NNP) redefine AI silicon[EB/OL]. (2017-10-17)[2018-03-31]. https://ai.intel.com/intel-nervana-neural-network-processors-nnp-redefine-ai-silicon/.

[25] Chua R. 2017 innovations in edge computing and MEC report [R]. Denver: SDxCentral, 2017.

[26] Liu S, Du Z, Tao J, et al. Cambricon: An instruction set architecture for neural networks[C]//International Symposium on Computer Architecture. Piscataway NJ: IEEE, 2016: 393-405.

[27] 孙永杰.地平线: 架构创新BPU 算法+芯片+云一体化[J]. 通信世界, 2018(13): 31.
Sun Yongjie. Horizon Robotics: Architecture innovation BPU algorithm + chip + cloud integration[J]. Communications World, 2018(13): 31.

[28] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. Fiber, 2015, 56(4): 3-7.

[29] 宋继强, 魏少军. AI芯片: 从历史看未来[EB/OL]. (2018-07-20)[2018-07-31]. http://36kr.com/p/5144249.html.
Song Jiqiang, Wei Shaojun. AI chip: Looking at the future from the perspective of history[EB/OL]. (2018-07-20)[2018-07-31]. http://36kr.com/p/5144249.html.

[30] Luker P A, Rothermel D. The philosophy of artificial intelligence[J]. ACM Sigcse Bulletin, 1990, 26(1): 41-45.

[31] AI芯片终极难题被清华大学IC男神解决了[EB/OL]. (2018-02-12)[2018-03-31]. http://zhidx.com/p/109515.html.
The ultimate problem of AI chip is solved[EB/OL]. (2018-02-12)[2018-03-31]. http://zhidx.com/p/109515.html.

Present situation and future trend of artificial intelligence chips

YIN Shouyi, GUO Heng, WEI Shaojun

Institute of Microelectronics, Tsinghua University, Beijing 100083, China

Abstract The artificial intelligence chips (AI) chip is an important part of artificial intelligence technology. It is the hardware foundation of AI algorithm and the essential of the AI era. This paper analyzes the state of the art, characteristics, potential technical trends and marketing of AI chips, and forecasts the opportunities, challenges and future trends faced by AI chips.

Keywords artificial intelligence chips chip; deep learning; reconfigurable computing

(责任编辑　刘志远)