美国人工智能芯片最新发展

徐晨
国家工业信息安全发展研究中心，北京，100040

摘　要： 近年来，随着全球对人工智能（AI）行业投资的不断增加，人工智能算法和应用处于高速发展和快速迭代时期，作为支持人工智能底层基础的人工智能芯片，呈现出广阔的发展前景。美国作为人工智能芯片的领导者，高度重视人工智能芯片的研发。本文梳理了美国在图形处理器（GPU）技术、现场可编程逻辑门阵列（FPGA）技术和类脑芯片等方面的发展动向，分析了人工智能芯片技术路线的优缺点和发展方向，研判了人工智能芯片在促进信息化装备、军事仿真训练、智能化指挥系统等领域的军事应用。

关键词： 人工智能芯片；类脑芯片；机器学习；神经网络
中图分类号：TP18　　**文献标志码：**A　　**DOI：**10.19772/j.cnki.2096-4455.2022.7.006

0　引言

人工智能芯片目前尚无准确定义，但可以从广义和狭义两个角度来对其进行阐释：首先，从广义角度，只要能够运行人工智能算法的芯片，都可以被视作人工智能芯片；其次，从狭义角度，人工智能芯片指针对人工智能算法做了特殊加速设计的芯片（现阶段的人工智能算法一般以深度学习算法为主，也可以包括其他机器学习算法），这也被视为通常意义上对人工智能芯片的定义[1-3]。通常来讲，只要是面向人工智能应用而设计的芯片都可被称为人工智能芯片。近年来，人工智能成了引领新一轮科技革命和产业变革的战略性技术，已被美国提升至国家战略层面。人工智能芯片作为人工智能技术的核心硬件，为人工智能应用提供了强大的算力支撑，其重要性不言而喻，已引起了英伟达、英特尔、谷歌、IBM等美国科技公司的强烈关注，纷纷加速其布局以抢夺先发优势。

1　人工智能芯片发展动向

美国科技公司积极推动人工智能芯片技术的发展，在晶体管密度、计算内核数、时钟频率、功耗等方面均有较大进步。英伟达继续领跑图形处理器（GPU）技术，新产品性能较之前大幅提升；英特尔积极发展现场可编程逻辑门阵列（FPGA）技术和类脑芯片技术，推出全球密度最高FPGA芯片和类脑神经拟态系统；谷歌利用张量处理器（TPU）技术，推出专用集成电路（ASIC）硬件平台，具备高性能机器学习推理能力；此外，美国初创公司Cerebras Systems推出规模最大的人工智能芯片，专门用于处理人工智能算法问题；初创公司 Myhtic推出了具有足够的存储与大量并行计算单元的AI芯片，可显著减少数据移动能力；麻省理工学院的初创公司Lightmatter推出了加速的光子计算测试芯片，将重新定义AI智能芯片领域的发展。

1.1　图像处理器核心性能持续提升

2019年6月，美国超威半导体公司推出GPU芯片RX5700XT，采用RDNA架构和7nm工艺制造，核心面积251mm^2，单位面积性能较之前提升2.3倍，具有103亿个晶体管、2560个流处理器、40个计算单元、160个纹理单元、加速频率1905MHz、功耗225W；7月，英伟达推出GPU芯片RTX2080Super，采用“图灵”架构和12nm工艺制造，核心面积545mm^2，性能提升了25%左右，具有136亿个晶体管、3072个流处理器、384个张量计算核心、192个纹理单元、加速频率1815MHz、功耗250W。这两款GPU芯片具有强大的人工智能图形处理能力，将大幅提升显控系统的成像质量。

2020年英伟达公布了用于超级计算任务的A100人工智能芯片，这款基于第八代Ampere架构

的芯片所采用的弹性计算技术能将每个芯片分割为多达7个独立实例来执行推理任务，人工智能算力提升20倍以上，被业界认为是史上最大的性能飞跃。这是人类有史以来首次可以在一个平台上实现对横向扩展以及纵向扩展的负载的加速；此外，A100人工智能芯片在提高吞吐量的同时，降低了数据中心的成本[4]；2020年年底，Mythic 推出了其第一代AI芯片M1108 AMP。与很多AI芯片不同，M1108采用更加成熟的模拟计算技术，将足够的存储与大量并行计算单元打包在芯片上，可最大化内存带宽并减少数据移动能力。

1.2 现场可编程逻辑门阵列（FGPA）的逻辑密度越来越高

2019年11月，英特尔发布全球最高密度的FGPA芯片Stratix® 10 GX 10M，采用14nm工艺制造，核心面积1400mm^2，拥有1020万个逻辑单元以及443亿个晶体管。这款高密度的FGPA芯片采用了英特尔先进的嵌入式多芯片互连桥接（EMIB）技术将两块FGPA的逻辑系统连接，形成了多达25920个高带宽连接，内部数据带宽高达6.5TB/s。此外，还具有308兆比特的内存，6912个数据信号处理器（DSP），2304个用户I/O引脚。该芯片将支持ASIC和SoC技术的仿真与原型设计，也将广泛支持测试测量、计算、网络、航空航天和国防等相关应用。

2021年8月，麻省理工学院的初创公司Lightmatter发布了一块AI加速的光子计算测试芯片，该芯片由毫瓦级的激光光源供电，利用硅光子和MEMS技术的处理器，其速度比传统芯片快1000倍，但是功耗却只有普通电子器件的千分之一，采用的是两个层叠的芯片组，面积约为150mm^{2}左右，内部拥有超过十亿FinFET晶体管、数万光子算术单元，这将重新定义AI智能芯片领域的发展。

1.3 专用集成电路在特定领域体现价值

2019年3月，谷歌推出智能化专用集成电路硬件平台Coral，包含完整的本地人工智能工具包，可在设备上创建、培训和运行神经网络。该平台搭载谷歌Edge TPU ASIC芯片，包含可移动模块化系统、USB加速器、500万像素摄像头等组件，最大限度地减少延迟和功耗，使低功耗设备具备高性能的机器学习推理能力。

谷歌的Edge TPU边缘人工智能芯片是专为在边缘运行TensorFlow Lite ML模型而设计的ASIC芯片，可用于越来越多的工业使用场景，如预测性维护、异常检测、机器视觉、机器人学、语音识别等，可以应用于制造、本地部署、医疗保健、零售、智能空间、交通运输等各个领域，具有体型小、功耗低、性能出色的优势，可以在边缘部署高精度人工智能。

2020年 2 月，谷歌发布首个全球人工智能模型平台（Model Play），该平台搭载了Edge TPU人工智能芯片，是一款面向全球用户的人工智能模型资源交流与交易平台，为机器学习与深度学习提供丰富和多样化的功能模型，可兼容多种人工智能芯片，帮助用户快速创建和部署模型，显著提高了模型开发和应用效率，降低了人工智能开发及应用门槛。

1.4 类脑芯片向“模拟大脑”的目标迈近一大步

2019年7月，英特尔在DARPA“电子复兴计划”年度峰会上发布Pohoiki Beach神经拟态系统，该系统由64块Loihi芯片组成，采用14nm工艺，总面积3840mm^2，拥有1320亿个晶体管、800万个神经元，处理速度比传统CPU快1000倍，效率高1万倍，功耗小100倍，将为图像识别、自动驾驶领域带来巨大的技术提升。

2020年8月，美国苹果公司公布其最新 A14仿生芯片，该芯片的CPU性能相比上一代A13仿生芯片提升40%，GPU性能相比上一代仿生芯片提升50%，优于包括英特尔芯片在内的其他芯片；A14仿生芯片还搭载了定制技术，这些技术可以驱动速度更快的神经引擎，实现更强大的机器学习能力。2020年11月，苹果公布A14X 仿生芯片的CPU和GPU性能基准，与A12Z仿生芯片相比，多核测试的性能提高了35%。

1.5 史上规模最大智能芯片显著提升学习速度

2019年8月，美国初创公司Cerebras Systems推出有史以来规模最大的人工智能芯片，专门设

计用于处理人工智能应用问题，显著提升学习速度。该芯片采用台积电16nm工艺制造，面积达46225mm^2、拥有1.2万亿个晶体管、40万个计算内核、18吉比特片上静态随机存储器，已被美国能源部的阿贡国家实验室和劳伦斯·利弗莫尔国家实验室应用于人工智能计算机中。

2 人工智能芯片的特点及发展趋势分析

当前，人工智能多样化的场景应用对人工智能芯片的性能、功耗、延迟以及成本等指标提出不同需求，人工智能芯片呈现出多技术路径并行发展的态势。

2.1 人工智能芯片不同技术路径各有优缺点

根据设计需求，人工智能芯片主要分为：图形处理器（GPU）、现场可编程逻辑门阵列（FGPA）、专用集成电路（ASIC）、类脑芯片、通用智能芯片。通常根据具体应用场景，在性价比、能效比、可靠性之间折中选择。

不同技术路径人工智能芯片的特点为：（1）GPU叠加大量计算单元和高速内存，逻辑控制单元简单、通用性强。但GPU不能独立工作、功耗大、价格成本高，通常用于3D图像处理和密集型并行计算。（2）FPGA具备可重构数字门电路和存储器，硬件配置灵活，能快速适应算法的迭代更新，功耗和速度优于GPU。但FPGA编程门槛高、峰值性能不如GPU，通常用于算法更新频繁的小规模计算领域。（3）ASIC计算能力和计算效率根据算法需要定制，体积小、功耗低、计算性能高，速度比FGPA快5~10倍，功耗远优于GPU，量产后成本也将低于FGPA。但ASIC开发周期长、上市速度慢、面临风险高，常用于需求量较大的专用领域。（4）类脑芯片模拟人类大脑处理信息，以极低功耗对信息进行异步、并行、低速和分布式处理，具备感知、识别和学习等功能，性能强大且通用性强。但类脑芯片开发技术难度大，目前仍处于研发阶段。（5）通用智能芯片具有可编程性、架构动态可变性、架构高效重组、高计算效率、低成本、低功耗等特征，可按照软件的需求来调整芯片计算能力，是人工智能芯片发展的最终目标。但通用智能芯片开发技术难度大，目前还没有真正意义上的通用智能芯片。

2.2 人工智能芯片朝高性能、高密集度、高智能方向发展

通过在计算架构、器件材料、电路设计、制造工艺上的改进和创新，人工智能芯片朝高性能、高密集度、高智能化方向发展，在算力、功耗、成本等方面不断提升。此外，多模异构集成和通用智能芯片也成了人工智能芯片未来的重要发展方向。

2.2.1 制造工艺进步推动芯片性能持续提升

随着制造工艺水平的不断提升，传统架构人工智能芯片的晶体管密度更高，核心数更多、运算速度更快、功耗更低，计算能力持续上升。当前，大多数人工智能芯片都还在采用10nm以上工艺，随着更先进工艺技术的不断被尝试使用，高密度芯片将不断被推出。

2.2.2 多模异构集成，实现优势互补

人工智能技术需要大数据驱动的数据算法，同时也需要小数据、小样本算法应用，单个类型的人工智能芯片都不能将处理效果发挥到最佳。针对多样化的人工智能算法，采用多模异构集成，融合不同人工智能算法，形成优势互补，是人工智能未来发展的重点方向。

2.2.3 通用智能芯片是未来发展的终极目标

人工智能芯片需不断调整架构以适应人工智能多变的算法，新架构的反复开发使成本和技术难度不断提升，通用人工智能芯片根据算法需求自动调整架构，极具灵活性和适应性，是未来技术发展的必然方向。

3 人工智能芯片的军事影响和意义分析

人工智能芯片是实现未来军用人工智能技术的核心和关键，推动人工智能武器、智能电子战、智能作战管理、智能仿真、智能情报分析与图像识别、武器装备自动故障诊断与排除、作战机器人和智能无人机等军用人工智能技术的发展，进

而为未来战争的作战样式带来翻天覆地的改变。

3.1 促进信息化装备显控系统的性能提升

图形处理器作为信息化装备显控系统的“大脑”，是实现“人机对话”的重要元件之一，其先进程度直接对是否能制敌于“千里”之外构成了影响。GPU具备的安全性、稳定性、可靠性以及强大并行计算能力，将增强对战场信息的采集、分析、显示能力，提升信息化装备的快速反应能力，为夺取信息权提供了有力的保证。

3.2 促使军事仿真训练环境更加真实、更贴近实战

人工智能芯片为军事训练提供了一种低成本、高效率、高稳定性的解决方案，可使军事仿真训练的战场地形分辨率更高、地幅更大，训练环境更加真实、更贴近实战，兵力生成精度更高、速度更快，使模拟训练的人员感受到与真实战场相近的压力，从而有效地提高士兵的战场承受能力以及指挥员的临场指挥能力。

3.3 增强雷达数字信号处理能力，提升定位精度

现场可编程逻辑门阵列是雷达不可或缺的核心部件，快速并行处理能力有效增强了雷达的数字信号处理能力，使其在日益复杂的电磁环境中显著提升多目标同时处理的能力。FGPA器件不仅扩充了雷达功能、提升了运算速度，还实现了雷达的编程化和模块化处理，使其更加符合现代化信息战争的需求。

3.4 高密度芯片推动信息装备的小型化、轻量化进程

现场可编程逻辑门阵列是现代化信息装备的必备芯片，90%以上的大型军用电子设备靠其发挥作用，目前信息装备中FGPA芯片可达近千片，价格昂贵、功耗巨大。高密度FGPA具有高集成度，在提升单位面积计算能力的同时降低了功耗，可有效缩减信息装备尺寸，进一步提高机动作战能力，适应现代高科技战争发展趋势。

3.5 为智能化指挥信息系统提供新的技术途径

指挥信息系统面临着复杂多变的战场环境，需要具备小样本学习、抗噪性、通用智能等新能力，类脑芯片的设计架构与人的大脑机构相似，在适应性方面表现出更加类人的特性，为智能化指挥信息系统提供了新的技术途径。类脑芯片的发展有望不断推动军事智能的发展，成为未来指挥信息系统实现更高级智能化水平的有力推手。

4 几点认识

4.1 人工智能芯片迎来巨大的发展机遇

人工智能芯片的发展还处在初级阶段，科研和产业应用都拥有巨大的创新空间，将在材料、架构、设计理念和应用场景等方面迎来巨大的发展机遇。未来几年，人工智能芯片将持续火热，技术创新将不断涌现。

4.2 ASIC成为人工智能芯片的重要分支

ASIC针对特定需求开发，能够更好地根据需求进行性能和功耗的定向优化，其专用的芯片架构与高复杂度的算法相匹配，量产后在性能、功耗、成本等方面均具有较大优势，长期来看，非常适用于人工智能应用。随着仿真与原型设计技术的不断成熟，ASIC有望在今后取代GPU和FPGA，成为人工智能芯片的重要分支。

4.3 架构创新是人工智能芯片发展的重点方向

随着人工智能应用场景的多样化，架构创新使人工智能芯片的智能化水平越来越高。目前，短期发展采用异构集成的方式加速各类应用算法；中期发展着重在自重构、自学习、自适应方面发展，支持算法的演进和类人的自然智能；长期发展朝淡化人为干预的通用型芯片方向发展。

参考文献

[1] James A P. Towards strong AI with analog neural chips[C]. 2020 IEEE International Symposium on Circuits and Systems (ISCAS2020), Seville, Spain, 2020.

[2] 汪鑫.人工智能芯片的概念和应用分析[J].中国新通信,2020,22(20):112-113.

[3] 尹首一.人工智能芯片概述[J].微纳电子与智能制造,2019,1(2):7-11.

[4] 葛悦涛,任彦.2020年人工智能芯片技术发展综述[J].无人系统技术,2021,4(2):14-19.