人工智能芯片产业发展现状及展望

文 _ 赵荣杰 房超 许蔓舒★

内容提要：

人工智能是新一轮科技革命和产业变革的重要驱动力，已成为全球科技竞争的焦点领域。人工智能芯片（AI 芯片）是人工智能技术的核心硬件基础，对人工智能的发展影响巨大。本文对人工智能与集成电路产业的深度融合点，对 AI 芯片的技术发展与产业趋势进行研究，展望中国 AI 芯片产业发展的巨大潜力。

关键词：人工智能；人工智能芯片；集成电路

自 2016 年 AlphaGo 击败李世石赢得人机围棋大战后，人工智能在全球范围引发高度关注，成为新的投资风口。全球各大企业加快布局，各国政府也纷纷出台相关战略，以促进人工智能技术的发展。从硬件来看，若没有高性能的芯片提供可靠算力，根本无法承载日益完备的机器学习模型和大规模的基础数据库。离开适配芯片的人工智能，只能是存在于理论研究而无法落地的人工智能。

广义而言，AI 芯片指的是专门用于处理人工智能应用中大量计算任务的模块，即面向人工智能领域的芯片均被称为 AI 芯片。狭义的 AI 芯片指的是针对人工智能算法做了特殊加速设计的芯片。“无芯片不 AI”，以 AI 芯片为载体实现的算力是人工智能发展水平的重要衡量标准，发展更注重超速运算能力的 AI 芯片成为推动人工智能产业爆发的关键核心要素之一。

AI 芯片的发展历经磨难，但产业发展现状良好

AI 芯片在曲折中前进

2006 年以前，尚未出现突破性的人工智能算法，而且能够获取的数据资源也较为有限。传统 CPU 已经能够完全满足当时的计算需要，学界和产业界对 AI 芯

★ 赵荣杰，启元实验室，博士；房超，启元实验室，清华大学高技术实验室，研究员；许蔓舒，中信改革发展研究基金会金融实验室，咨询专家。

AlphaGo 击败李世石后世界掀起人工智能浪潮

片没有特殊需求，因此，AI 芯片产业的发展一直较为缓慢。

2006-2010 年期间，游戏、高清视频等行业快速发展，同时也助推了 GPU 产品的迭代升级。2006 年，NVIDIA 发布了通用并行计算架构 CUDA，使 GPU 具备了可编程性，即令 GPU 既能做游戏和渲染，也能做并行度很高的通用计算。统一计算设备架构推出后，GPU 编程更加易用便捷。研究人员发现，GPU 所具有的并行计算特性比通用 CPU 的计算效率更高，更加适用于深度学习等人工智能先进算法所需的“暴力计算”场景。在 GPU 的助力下，人工智能算法的运算效率可以提高几十倍，由此，研究人员开始大规模使用 GPU 开展人工智能领域的研究和应用。

2010 年之后，以大数据、云计算等为代表的新一代信息技术高速发展，并逐渐开始普及。研究人员在云端采用“CPU+GPU”混合计算模式，使得开展人工智能所需的大规模计算更加便捷高效，进一步推动了人工智能算法的演进和 AI 芯片的广泛使用，同时也促进了各种类型的 AI 芯片的研究与应用。

2016 年，谷歌旗下 Deep Mind 公司研发的人工智能系统 AlphaGo 击败了韩国棋手李世石，使得以深度学习为核心的人工智能技术引发了全球热潮。此后，业界对于人工智能算力的要求越来越高，而 GPU 价格昂贵、功耗高的缺点限制了其在场景各异的应用环境中的使用，因此，研究人员开始研发针对人工智能算法进行特殊加速的定制化芯片。大量 AI 芯片领域的初创公司与传统互联网巨头纷纷入局争夺市场，专用 AI 芯片呈现出百花齐放的格局，在应用领域、算力、能耗比等方面都有了极大的提升[①]。

AI 芯片呈现专用化、多样化发展态势

集成电路的发展、芯片的升级换代一直是依靠工艺、架构和应用三方面来拉动的。随着摩尔定律接近极限，工艺改进已经难以降低成本，人工智能的密集计算型需求已成为当前芯片技术的主要驱动力之一。通用处理器的架构无法适应人工智能算法的高需求，各种新的架构已然成为当前处理器芯片性能提升的关键手段[②]。

处理器芯片面向人工智能硬件优化升级，目前有两种发展路径：一种是延续传统计算架构，主要以三种类型的芯片为代表，即 GPU、FPGA、ASIC，但 CPU 依旧发挥着不可替代的作用；另一种是颠覆经典的冯诺依曼计算架构，采用神经拟态工程，利用电子技术模拟已经被证明的生物脑的运作规则，从而构建神经拟态芯片。为满足不同场景的应用需求，AI 芯片的发展逐渐呈现出专用化、多样化的特点。

① 商惠敏 . 人工智能芯片产业技术发展研究 . 全球科技经济瞭望 . 2021;36(12):24-30.
② 施羽暇 . 人工智能芯片技术体系研究综述 . 电信科学 . 2019;35(04):114-9.

AI 芯片为集成电路产业发展提供新的方向

AI 芯片作为芯片的一个分支，有其专用性也有其普遍性，专用于人工智能领域，同时和其他芯片一样，与集成电路产业本身的发展密不可分。AI 芯片的发展受到集成电路产业发展水平的制约，同时又为集成电路产业的发展提供新的方向。

人工智能，特别是深度学习，这几年呈爆发性的发展，很大程度上得益于集成电路技术多年的积累。如果不是集成电路技术已经发展到了一定的高度，能够给大规模的机器学习提供足够的处理能力，就没有战胜人类顶尖棋手的 AlphaGo。过去十几年驱动芯片技术发展的主要是通信，即多媒体和智能手机这些应用。而随着这些应用增长放缓，芯片技术的发展已经逐步转向了 AI 领域，AI 的驱动效应将在芯片技术上会有更明显的体现。

IT 巨头纷纷介入芯片领域，加速 AI 芯片研发

AI 芯片产业发展持续发力，竞争激烈

传统芯片企业优势地位明显

高通、英伟达、英特尔、AMD 等传统芯片厂商凭借在芯片领域多年的领先地位，迅速切入人工智能领域，积极布局，目前处于引领产业发展的地位，在 GPU 和 FPGA 方面则基本位于垄断地位。英伟达推出了 Tesla 系列 GPU 芯片，专门用于深度学习算法加速。AMD 于 2018 年推出了 Radeon Instinct 系列 GPU，主要应用在超算、数据中心等人工智能算力基础设施上，用于深度学习算法加速。当前，GPU 作为业界使用最为广泛、人工智能计算最成熟的通用型芯片，成为数据中心、超算等大型算力设施的首选，在效率和场景应用要求大幅提升和变化之前，GPU 仍将是 AI 芯片领域的主要领导者。

IT 巨头纷纷加大 AI 芯片研发定制力度

2015 年以来，谷歌、微软、IBM、Meta、苹果、亚马逊等国际互联网及 IT 巨头纷纷跨界开展 AI 芯片研发，力图突破算力瓶颈，并把核心部件掌握在自己手中。如谷歌于 2016 年发布了专门针对开源框架 TensorFlow 开发的芯片 TPU，并帮助 AlphaGo 击败李世石；近年，谷歌还推出了可在 Google Cloud Platform 中使用的云端芯片 Cloud TPU 以及用于边缘端推理的 Edge TPU，打造闭环生态。微软于 2017 年发布了基于 FPGA 芯片组建的 Project Brainwave 低时延深度学习系统，让微软的各种服务可以更迅速地支持人工智能功能。2018 年，亚马逊发布了高性能推理芯片 AWS Inferentia，支持 TensorFlow、Caffe2 等主流框架。

类脑芯片领域呈现异军突起之势

IBM 公司率先在类脑芯片领域取得突破，推出了 True North 类脑芯片，其采用 28nm 技术，整合 54 亿个晶体管和 4096 个处理核，相当于 100 万个可编程神经元，以及 2.56 亿个可编程突触，而功耗仅为 65 毫瓦，该研究成果被《科学》杂志刊登。由于神经突触要求可变与有记忆功能，

谷歌的无人驾驶汽车

IBM 采用 CMOS 工艺兼容的相变非挥发存储器（PCM）的技术实现，加快了商业化进程。2019 年，清华大学施路平教授团队发布了类脑芯片“天机芯”，使用 28nm 工艺流片，包含约 40000 个神经元和 1000 万个突触，支持同时运行卷积神经网络、循环神经网络以及神经模态脉冲神经网络等多种神经网络，是全球首款既能支持脉冲神经网络又可以支持人工神经网络的异构融合类脑计算芯片[①]。西井科技发布的 Deep South 芯片，核心是用 FPGA 模拟神经元以实现脉冲神经网络的工作方式，包含约 5000 万个神经元和高达 50 多亿个神经突触，可以直接在芯片上完成计算，并在“无网络”情况下使用，处理相同计算任务时，Deep South 芯片的功耗仅为传统芯片的几十至几百分之一。浙江大学与杭州电子科技大学共同研发了“达尔文”芯片，集成了 500 万个晶体管，包含 2048 个硅材质的仿生神经元和约 400 万个神经突触，可从外界接受并累积刺激，产生脉冲信号，处理和传递信息。

AI 芯片产业发展未来可期

不断涌现的新场景应用需求将催生超低功耗 AI 芯片

随着以 5G、物联网、人工智能等为核心的新一代信息技术的高速发展，涌现出越来越多新的应用场景和需求。物联网领域将需要体积更小、功耗更低、能效比更高的 AI 芯片。边缘端芯片如手机中的 AI 芯片，其功耗一般在几百毫瓦至几瓦，云端训练芯片的功耗通常要达到数百瓦，而超低功耗 AI 芯片的工作功耗一般是几十毫瓦甚至更低。如在以智能手表为代表的智能可穿戴设备领域，此类设备需要具备语音识别、心率检测等智能生物信号处理功能，电池容量因设备尺寸等原因受到极大限制，因此需要集成体积小且能效比超高的人工智能加速芯片，降低对电池的消耗；在智能家居等领域，智能门锁需要具备人脸识别、指纹识别等功能，而且不能经常更换电池，这就对门锁中的智能模块提出了极高的能效比要求。除此之外，制造业等工业应用场景中也需要使用超低功耗 AI 芯片，如安装在机械臂、管道中的智能传感器须由电池供电，使用超低功耗 AI 芯片可以有效减少电池消耗，大幅降低此类设备的维护成本。

开源芯片的普及将提升行业整体的发展水平

随着摩尔定律接近极限，通用芯片的性能提升陷入瓶颈，通用处理器架构无法适应不同场景人工智能算法的高需求，对新型架构 AI 芯片的需求日益增长，这为许多初创型中小企业带来新的市场机遇。然而，芯片领域过高的技术门槛和知识产权限

① 李钢，李繁荣，程健．应用场景需求：驱动人工智能芯片设计发展．前沿科学．2018;12(4):4.

制，严重阻碍了 AI 芯片的进一步技术创新和市场响应速度。如果开源芯片能够普及，首先可以节省 IP 模块方面的费用，降低研发成本。同时，由于开源的设计可以由社区持续地改进，所有人都能享受到最新、最优化的成果，这样便可以提高行业整体的发展水平。2014 年，美国加州大学伯克利分校的研究团队正式发布了“RISC-V”开源精简指令集架构，具有灵活简洁、模块化、扩展性强、易实现等优点，可以较好地适应高性能计算设备、专用硬件设备、低功耗嵌入式设备等众多应用领域的需求，而且“RISC-V”完全免费，因此，“RISC-V”也成为目前推广度、普及度最高的开源芯片项目，并已逐渐成为芯片设计领域的主流指令集之一。

国际技术进出口交易会上智能机器人“争奇斗艳”

AI 芯片将从特定场景的加速芯片向通用智能芯片发展

目前，人工智能技术在图像处理、语音识别、自动驾驶等应用领域取得巨大进展，但是要从单点突破走向全面开花，需要人工智能领域产生像 CPU 一样可适用于任意人工智能应用场景的通用 AI 芯片。总体来看，短期内 AI 芯片仍将以“CPU+GPU+AI 加速芯片”的异构计算模式为主，中期会重点发展可自重构、自学习、自适应的 AI 芯片，未来将会走向通用的 AI 芯片。通用 AI 芯片发展的主要难点在于通用性和实现的复杂度，同时还面临着传统冯诺伊曼架构的技术瓶颈以及摩尔定律接近物理极限这两大挑战。未来，随着新型半导体材料和物理器件以及芯片的制程工艺等出现新突破，以及人类对于大脑和智能本身形成更深层次的认知，将有望最终实现真正意义上的通用 AI 芯片。但是，专用芯片与通用芯片永远都不是互相替代的关系，二者必须协同工作才能发挥出最大的价值[①]。

类脑芯片将持续扮演通用人工智能“探路者”角色

目前，类脑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来模拟已经被证明的生物脑的运作规则，从而实现类脑的学习、决策、认知等数据处理和分析功能。神经拟态计算通过模拟大脑的运行机制实现存算一体化，在算法以及芯片的设计上，可以实现以低于 1000 倍的功耗去完成同样效果的模型训练。因此，神经拟态芯片是一种环境友好型的芯片，其体积小、功耗低的特点，符合生物进化最本质的优势。2020 年 6 月，Gartner 发布报告预测，到 2025 年神经拟态芯片有望取代 GPU，成为用于人工智能系统的主要芯片之一。

① 张蔚敏，蒋阿芳，纪学毅 . 人工智能芯片产业现状 . 电信网技术 . 2018(02):67-71.

第十七届“中国芯”集成电路产业促进大会

中国 AI 芯片产业发展潜力巨大

当前，全球 AI 芯片的发展还处于起步阶段，中国 AI 芯片产业同样也还处在起步阶段。未来五年，中国 AI 芯片产业将会迎来飞速发展，产业增速也将处于世界顶尖。

产业发展环境不断优化，产业规模初显

中国从顶层高度重视人工智能产业和芯片产业发展，相继发布一系列产业支持政策，优化产业发展环境。如国务院《新一代人工智能发展规划》提出研发神经网络处理器以及高能效、可重构类脑计算芯片等；财政部和税务总局《关于集成电路设计和软件产业所得税政策的公告》明确指出，对已成立且符合条件的集成电路设计企业和软件企业实行税收优惠减免政策；国务院《新时期促进集成电路产业和软件产业高质量发展若干政策》明确了集成电路和软件相关企业或项目的税收优惠政策。

目前，中国 AI 芯片产业规模初显。全球市场洞察公司最新报告显示，全球 AI 芯片市场规模预计到 2026 年增长至 700 亿美元，复合年增长率（CAGR）将达到 35 % 左右。iiMedia Research 数据显示，2020 年中国 AI 芯片市场规模达 183.8 亿元，预计 2023 年将突破千亿元，复合增长率高于全球。

产业链上企业全面发展，产业应用水平处于世界前列

AI 芯片的产业链上游为原材料和生产设备（晶圆材料和设备、封装材料和设备等），中游是集成电路（设计、制造和封测），下游是行业应用（数据中心、通信设备、IoT 等）。

目前，中国 AI 芯片产业链上发展最为迅猛、技术含量最高、融资案例最多的是芯片设计类企业。上游的晶圆材料和封装材料等传统原材料企业，主要集中在江苏和广州，国产供给率逐年提高。国内的晶圆制造设备和封装设备，中低端可以自给自足，高端设备仍处于快速研发阶段；中游晶圆加工、制造和封测的企业主要是第三方代工厂，比较知名的有台积电和中芯国际。AI 芯片设计类企业有百度、阿里、腾讯等互联网公司，也有华为、中兴等通信企业，还有寒武纪、燧原科技、地平线等创业公司；下游包括浪潮、联想等服务器提供商，IoT 终端生产企业及边缘计算解决方案提供商。

中国 AI 芯片产业的整体水平处于世界中等水平。其中，上游和中游属于中下等水平，下游应用处于世界前列。产业链上游设备、材料环节，市场头部效应明显，进入壁垒非常高，技术突破难度大。国内华为及科研院所已开始高端光刻机的研发，但技术水平有待验证。产业链中游制造环节，进入壁垒较高，国产化水平较低，第三方代工厂如中芯国际，对 AI 芯片架构的理解和 IP 核的丰富程度不足；AI 芯片设计企业需要兼具芯片流片经验、AI 算法和 AI 框架理解，中国已陆续涌现一批技术型企业，产品初步成型，但技术水平有待继续迭代验证，国产

化水平中等。封测环节技术壁垒最低，毛利率最低，国产化水平高，但总体环节价值较低。

服务于制造、装配等领域的 ABB 单臂 YuMi 机器人

行业应用市场巨大，未来将多领域开花

从细分市场结构分类，AI 芯片可分为云端芯片和终端芯片，云端芯片又分为云端训练和云端推理芯片，终端芯片一般指终端推理芯片。云端芯片一般部署在公有云、私有云、混合云或数据中心、超算等计算基础设施领域，主要用于深度神经网络模型的训练和推理，处理语音、视频、图像等海量数据。终端芯片承担推理任务，需要独立完成数据收集、环境感知、人机交互及部分推理决策控制任务。目前，云端训练芯片的比例仍然最大，但增速最慢，云端推理芯片与终端推理芯片市场在未来几年都将保持快速增长。未来 2~3 年，随着区域性大规模数据中心的陆续建设完成，云端训练芯片增长速度将放缓；而随着 5G 和 IoT 等新兴 AI 芯片市场需求的释放，云端推理芯片、终端推理芯片市场增长速度将持续呈上升趋势。

中国具有全球最大的 AI 芯片应用市场，下游的行业应用中国处于世界前列[①]。从行业应用角度来看，中国 AI 芯片行业场景发展不平衡。云计算和安防行业是我国 Top2AI 芯片应用行业，交通、金融和工业等其他行业占比较低，但增速高，未来占比会显著提升。未来数据中心需求不减，云计算依旧会是我国最大的 AI 芯片应用行业；安防行业也会给 AI 芯片提供较大增量。此外，零售、医疗等行业对 AI 芯片的应用程度将持续提升，具备较高增速。

结语

核心芯片行业处于人工智能产业的最上游，是人工智能产业发展的基础和先锋。当 CPU 和传统计算架构无法满足基于深度学习模型的算法对大规模并行计算能力的需求时，突破专门用于处理人工智能应用中大量计算任务的核心芯片势在必行。同时在 PC 时代和移动互联网时代分别处于霸主地位的 X86 架构和 ARM 架构的发展历程表明，核心芯片将决定一个新的计算平台的基础架构和发展生态。从产业链本身的各个环节来看，中国 AI 芯片产业发展面临包括产业链关键环节存在短板、顶尖人工智能 + 芯片的复合人才缺失，AI 芯片的高能耗对经济和环境有影响等诸多挑战，但基于海量的数据要素资源、优秀的人工智能算法基础以及良好的政策支持等利好因素叠加，中国 AI 芯片产业将拥有广阔的发展前景。

（编辑 季节）

① 胡滨雨，郭敏杰．中国人工智能芯片期待突破．中国电信业．2021(04):36-9.