AI芯片行业研究报告

2019年

摘要

AI芯片主要适用于包括训练、推理在内的AI应用，擅长并行计算。主要应用于云端、边缘及物联网设备终端。市场空间在2022年有望超过500亿美元；

AI芯片在云端主要为数据分析、模型开发（训练）及部分AI应用（推理）等提供算力支持。英伟达基于其完备的GPU+CUDA生态主导云端AI芯片市场，但其产品售价高昂，GPU计算效能及功耗不如FPGA及ASIC芯片，市场寻求潜在替代方案；

边缘侧和终端对于AI芯片需求更加分散，不同场景需要综合考虑芯片的PPACR。AI芯片作为协处理器难以单独实现应用功能，对厂家软件及系统开发交付能力同样有很高的考量。不同的应用场景中，拥有较高的固有行业壁垒，这需要AI芯片厂商能够加强与产业固有主体的合作，融入现有产业结构；

芯片行业具有资本和技术壁垒双高的特点，高昂的研发费用需要广大的市场进行支撑，对于AI芯片厂商来说除了核心软硬件技术开发实力外，市场洞察及成本控制亦是不可或缺的能力；

行业当前接近Gartner技术曲线泡沫顶端，未来1~2年将会面临市场对于产品的检验，只有通过市场检验和筛选的优质团队才能够继续获得产业、政策和资本的青睐和继续支持。

来源：艾瑞研究院自主绘制。

关于人工智能芯片（AI芯片）

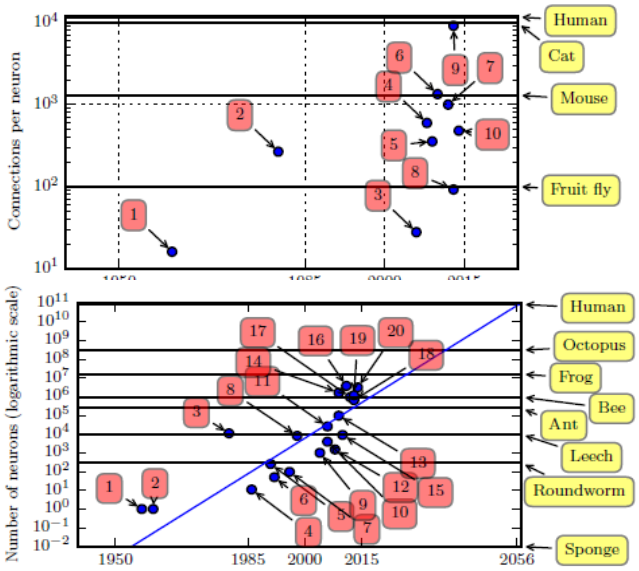
AI芯片：基于矩阵运算、面向AI应用的芯片设计方案

1、定义：当前AI芯片设计方案繁多，包括但不限于GPU\FPGA\ASIC\DSP等。目前市场上的对于AI芯片并无明确统一的定义，广义上所有面向人工智能（Artificial Intelligence，AI）应用的芯片都可以被称为AI芯片。

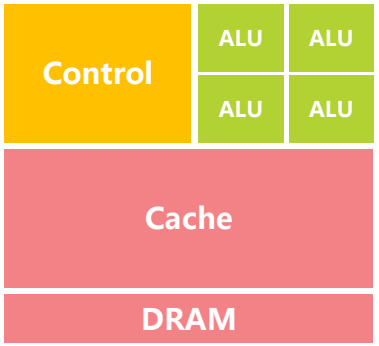
2、当前AI运算指以“深度学习” 为代表的神经网络算法，需要系统能够高效处理大量非结构化数据（文本、视频、图像、语音等）。这需要硬件具有高效的线性代数运算能力，计算任务具有：单位计算任务简单，逻辑控制难度要求低，但并行运算量大、参数多的特点。对于芯片的多核并行运算、片上存储、带宽、低延时的访存等提出了较高的需求。

3、针对不同应用场景，AI芯片还应满足：对主流AI算法框架兼容、可编程、可拓展、低功耗、体积及造价等需求。

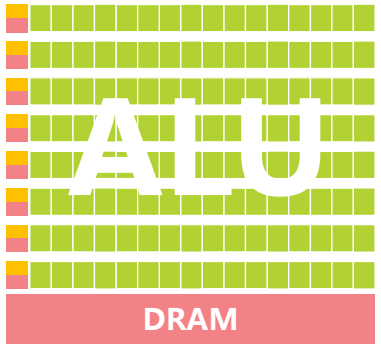
深度学习模型复杂度及规模对芯片算力需求激增



通过架构设计AI芯片跨越工艺限制，算力效能对CPU实现大幅超越



- 芯片工艺制程逼近物理极限；
- CPU芯片中大量晶体管用于构建逻辑控制和存储单元，用于构建计算单元的晶体管占比极小；
- 为了保证兼容性，CPU构架演进发展受限。



- 工艺提升缓慢，面对大规模并行运算需求，需要对芯片架构进行重新设计；
- GPU：开发即面向图像处理等大规模运算需求；
- FPGA/ASIC：对缓存、计算单元、连接进行针对性优化设计。

注释：DL：Deep Learning，指深度学习。

来源：《Deep Learning》——Ian Goodfellow、Yoshua Bengio、Aaron Courville；英伟达官网。

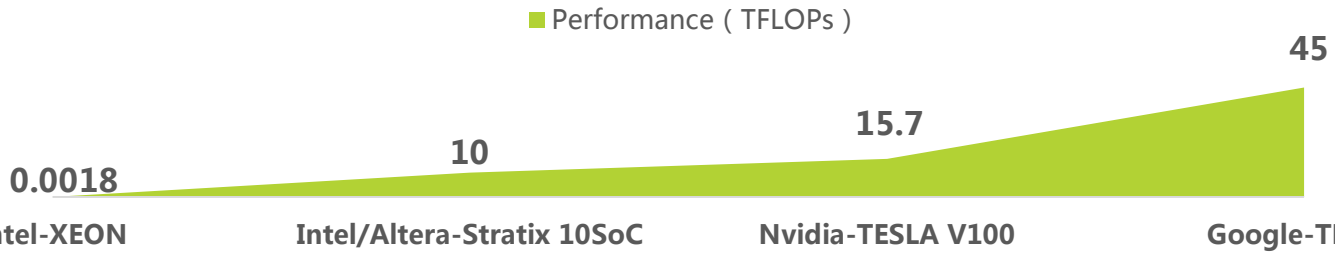
AI芯片实现算力提升

AI芯片满足AI应用所需的“暴力计算”需求

早在上世纪80年代，学术界已经提出了相当完善的人工智能算法模型，但直到近些年，模型的内在价值也没有被真正的实现过。这主要是受限于硬件技术发展水平，难以提供可以支撑深度神经网络训练/推断过程所需要的算力。直到近年来GPU\FPGA\ASIC等异构计算芯片被投入应用到AI应用相关领域，解决了算力不足的问题。

下图以云计算场景为例，通过对全球几大科技巨头的代表性云端芯片产品计算性能对比，我们可以发现ASIC芯片相比起其他几种芯片，在计算效能、大小、成本等方面都有着极大优势，未来随着通用AI指令集架构的开发，预计会出现最优配置的AI计算芯片。

典型的云端计算芯片算力表现比较



Intel-XEON	Intel/Altera-Stratix 10SoC	Nvidia-TESLA V100	Google-TPUv2
CPU	FPGA	GPU	ASIC
优点：通用性好，串行运算能力强；适用于逻辑运算； **缺点**：开发难度最大，大量的晶体管用于构建控制电路和高速缓冲存储器，运算单元占比少，架构限制了算力的进一步提升。	**优点**：可对芯片硬件层进行灵活编译，且功耗远远小于CPU和GPU **缺点**：硬件编程语言难以掌握，单个单元的计算能力比较弱，硬件成本较高，电子管冗余，功耗可进一步压缩。	**优点**：相比CPU，由于其多线程结构，拥有较强的并行运算能力，与FPGA和ASIC相比通用性更强； **缺点**：价格、功耗等不如FPGA和ASIC，并行运算能力在推理端无法完全发挥。	**优点**：针对专门的任务进行架构层的优化设计，可实现PPA最优化设计、量产后成本最低； **缺点**：初始设计投入大，可编程架构设计难度较大，针对性设计会限制芯片通用性。

注释：PPA：POWER、PERFORMANCE、AREA，指芯片的算力、功耗和面积。
来源：Intel官网；英伟达官网；公开网络数据；艾瑞研究院。