

AI芯片行业研究报告

2019年





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。

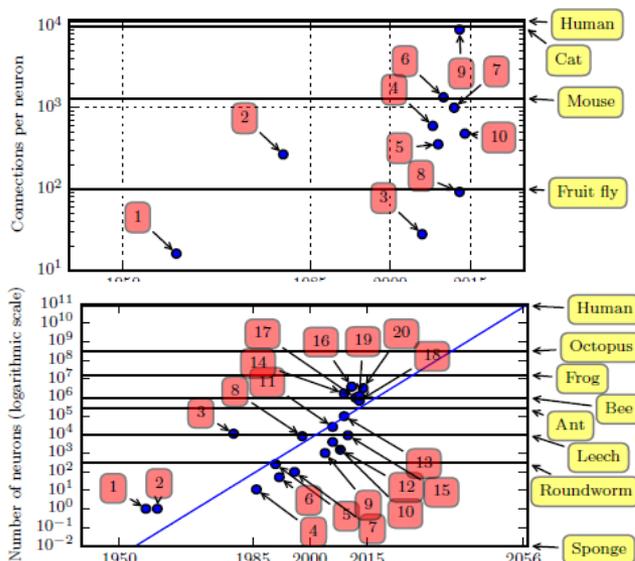
AI芯片行业概述	1
AI芯片应用场景及市场需求分析	2
AI芯片行业产业链及商业模式分析	3
AI芯片行业发展展望	4
企业推荐	5

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

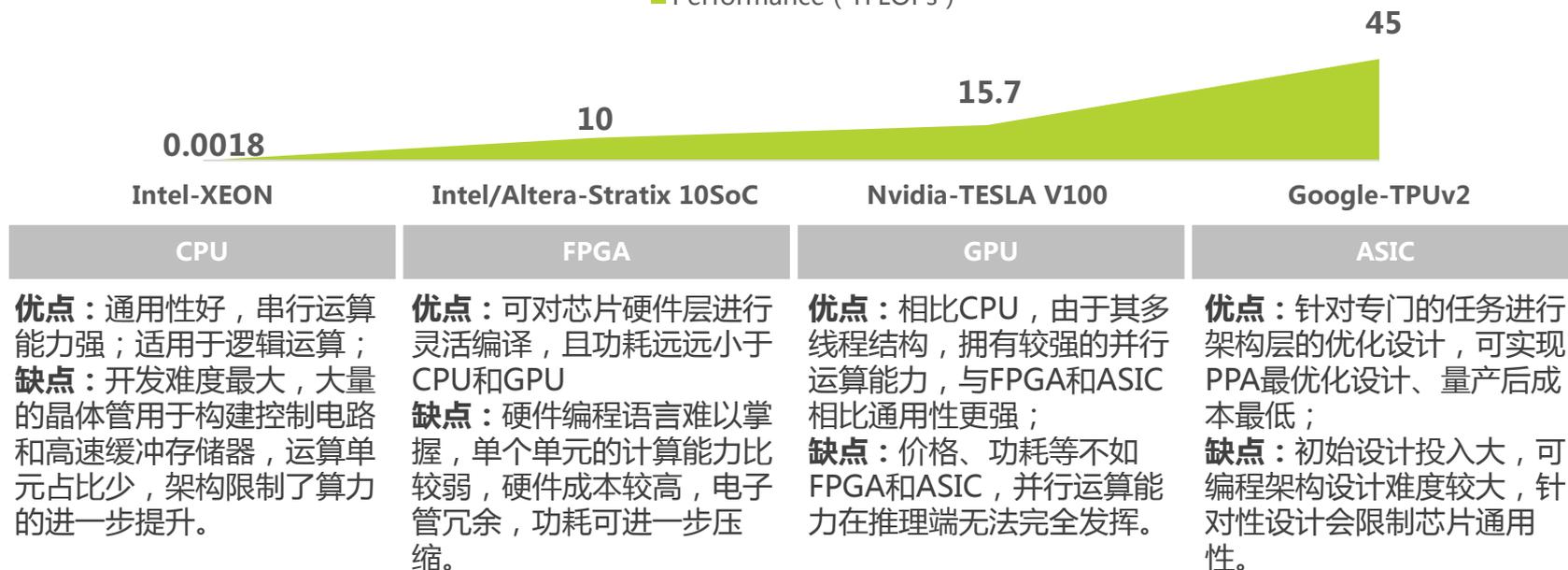
AI芯片满足AI应用所需的“暴力计算”需求

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

■ Performance (TFLOPs)



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。

AI芯片产品定位

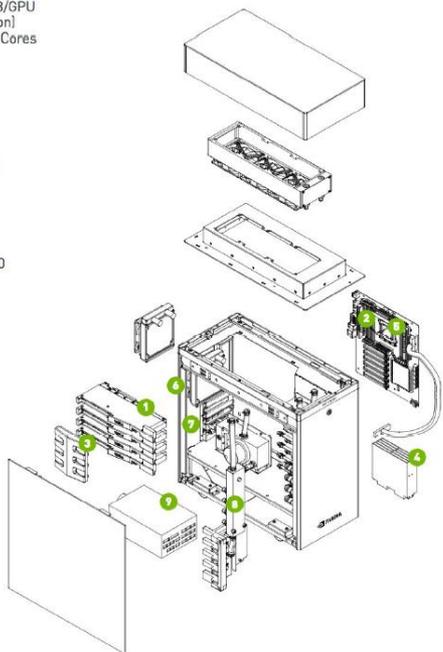
AI芯片对CPU并非替代，与CPU共同满足新时代计算需求

目前来看，AI芯片并不能取代CPU的位置，正如GPU作为专用图像处理器与CPU的共生关系，AI芯片将会作为CPU的AI运算协处理器，专门处理AI应用所需要的大并行矩阵计算需求，而CPU作为核心逻辑处理器，统一进行任务调度。

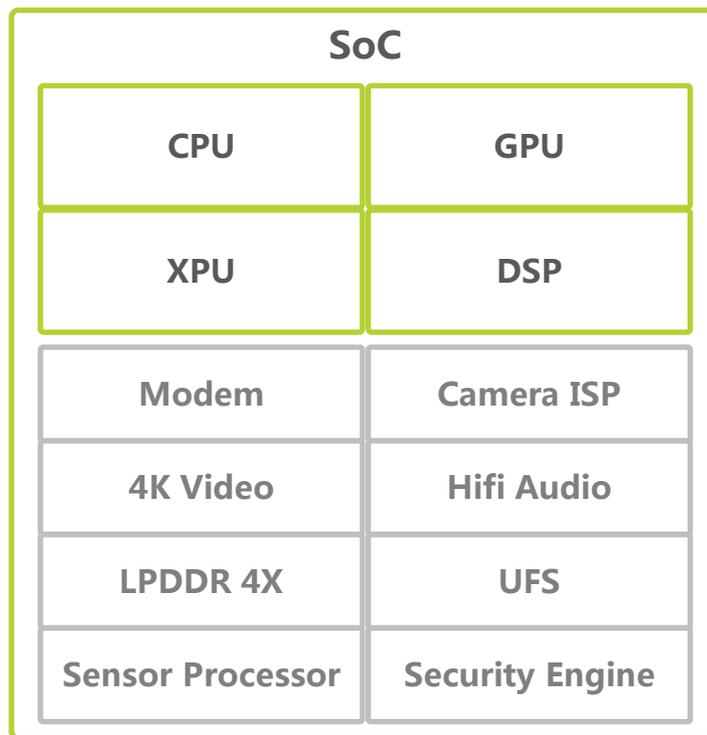
在服务器产品中，AI芯片被设计成计算板卡，通过主板上的PCIE接口与CPU相连；而在终端设备中，由于面积、功耗成本等条件限制，AI芯片需要以IP形式被整合进SoC系统级芯片，主要实现终端对计算力要求较低的AI推断任务。

服务器级产品中通过PCB上PCIE接口与CPU组成异构计算单元

- 1. GPUS**
4X NVIDIA Tesla® V100 32 GB/GPU
500 TFLOPS (Mixed Precision)
20,480 Total NVIDIA CUDA® Cores
2,560 Tensor Cores
- 2. SYSTEM MEMORY**
256 GB RDIMM DDR4
- 3. GPU INTERCONNECT**
NVIDIA NVLink™,
Fully Connected 4-Way
- 4. STORAGE**
Data: 3 x 1.92 TB SSD RAID 0
OS: 1 x 1.92 TB SSD
- 5. CPU**
Intel Xeon E5-2698 v4
2.2 GHz 20-Core
- 6. NETWORKING**
2X 10 GbE
- 7. DISPLAYS**
3X DisplayPort,
4K Resolution
- 8. COOLING**
Water-Cooled
- 9. POWER**
1500 W



通过SoC封装与CPU组成异构计算单元



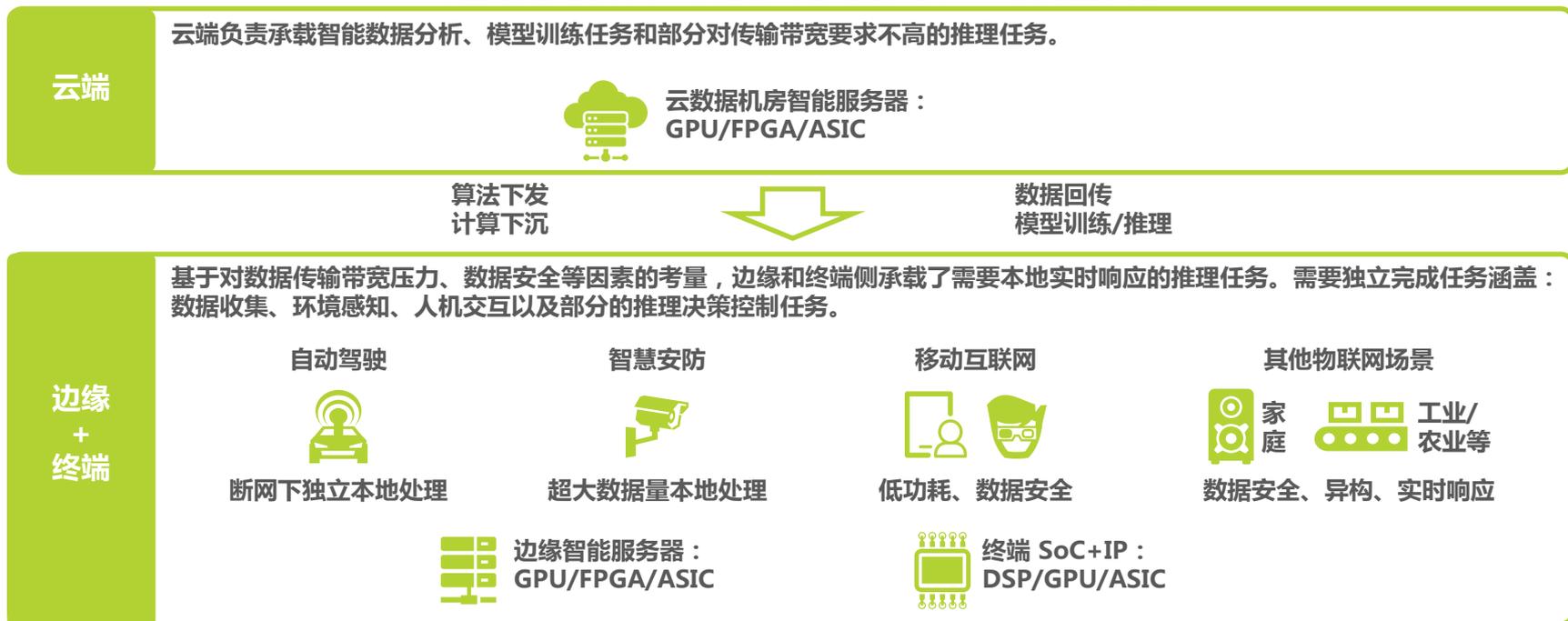
来源：英伟达官网，公开网络数据。

AI芯片应用场景

AI芯片为AI应用落地提供了商业化可行的算力解决方案

在人工智能发展初期，算法已经通过数据中心（云端）在大数据分析、精准营销、商业决策等方面实现了成功地应用落地。而未来，智能化将会逐渐渗透进入能源、交通、农业、公共事业等更多行业的商业应用场景中，除了部署在云端进行数据分析等工作，人工智能还需要下沉到摄像头、交通工具、移动设备终端、工业设备终端中，与云计算中心协同实现本地化的、低延时的人工智能应用。考虑到任务算力需求，以及传输带宽、数据安全、功耗、延时等客观条件限制，现有云端计算解决方案难以独自满足人工智能本地应用落地计算需求，终端、边缘场景同样需要专用的AI计算单元。

云端与边缘侧人工智能应用场景对于AI芯片的需求



来源：艾瑞研究院自主绘制。

AI芯片整体市场规模预测

AI芯片市场规模近5年增长有望接近10倍

1、市场根据AI芯片功能及部署场景将AI芯片分为：训练/推断、云端/边缘两个维度进行划分。训练端由于需要对大量原始数据进行运算处理，因此对于硬件的算力、计算精度，以及数据存储和带宽等都有较高要求，此外在云端的训练芯片应该有较好的通用性和可编程能力。推理端对于硬件性能要求没有推断端高，实证证明一定范围的低精度运算可达到同等推理效果，但同时这要求模型训练精度要达到较高水平。

2、根据中金公司研究部数据显示，2017年，整体AI芯片市场规模达到39.1亿美元，其中云端训练AI芯片20.2亿美元，云端推理芯片3.4亿美元，边缘计算AI芯片15.5亿美元；到2022年，整体AI芯片市场规模将会达到352.17亿美元，CAGR55%，其中云端训练AI芯片172.1亿美元，CAGR 54%，云端推断芯片71.9亿美元，CAGR 84%，边缘计算AI芯片108.2亿美元，CAGR 48%。

2017-2022年AI芯片细分市场预测

CAGR:53.5%
172.1



CAGR:47.5%
108.19

CAGR:84.1%



预览已结束，完整报告链接和二维码如下：

https://www.yunbaogao.cn/report/index/report?reportId=1_21105

