iResearch
艾 瑞 咨 询

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）

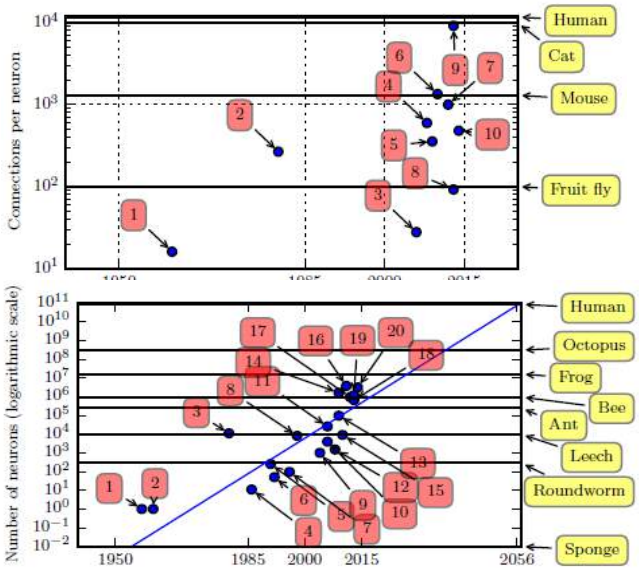
AI芯片：基于矩阵运算、面向AI应用的芯片设计方案

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。

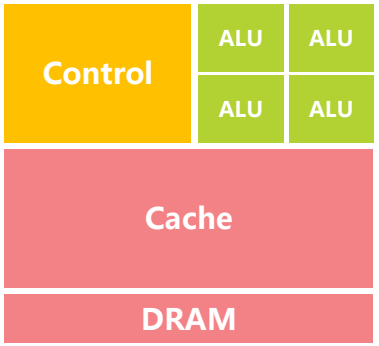
2、当前AI运算指以“深度学习” 为代表的神经网络算法，需要系统能够高效处理大量非结构化数据（文本、视频、图像、语音等）。这需要硬件具有高效的线性代数运算能力，计算任务具有：单位计算任务简单，逻辑控制难度要求低，但并行运算量大、参数多的特点。对于芯片的多核并行运算、片上存储、带宽、低延时的访存等提出了较高的需求。

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。

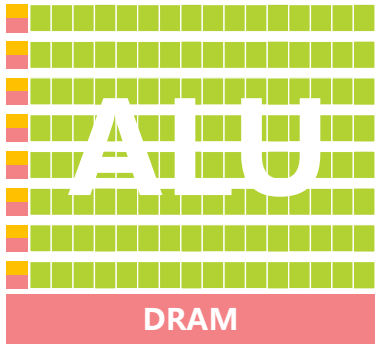
深度学习模型复杂度及规模对芯片算力需求激增



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。

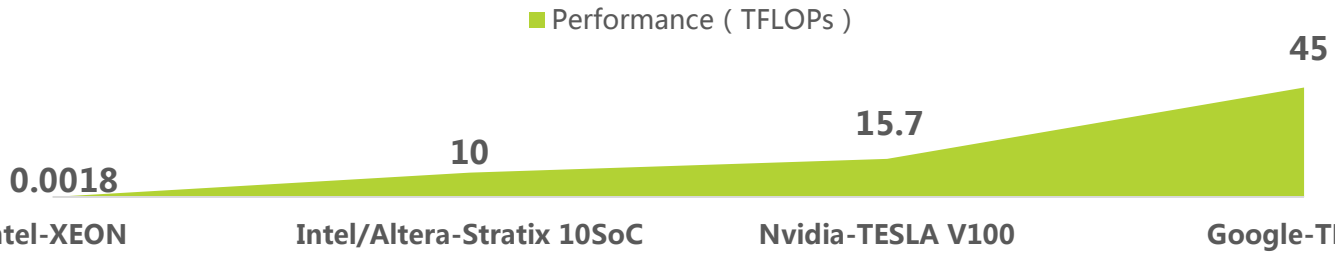
AI芯片实现算力提升

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。

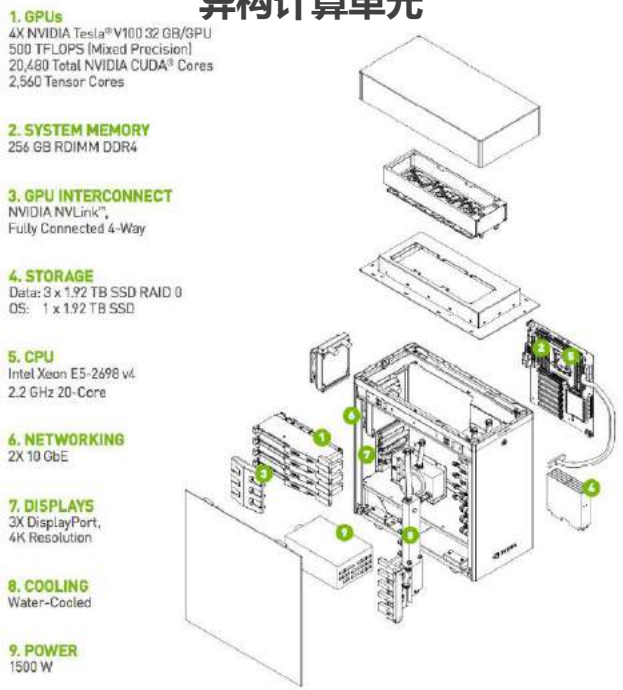
注释：PPA：POWER、PERFORMANCE、AREA，指芯片的算力、功耗和面积。
来源：Intel官网；英伟达官网；公开网络数据；艾瑞研究院。

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



通过SoC封装与CPU组成异构计算单元

SoC	
CPU	GPU
XPU	DSP
Modem	Camera ISP
4K Video	Hifi Audio
LPDDR 4X	UFS
Sensor Processor	Security Engine

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

云端

云端负责承载智能数据分析、模型训练任务和部分对传输带宽要求不高的推理任务。

云数据机房智能服务器：GPU/FPGA/ASIC

算法下发 计算下沉

数据回传 模型训练/推理

边缘 + 终端

基于对数据传输带宽压力、数据安全等因素的考量，边缘和终端侧承载了需要本地实时响应的推理任务。需要独立完成任务涵盖：数据收集、环境感知、人机交互以及部分的推理决策控制任务。

自动驾驶

断网下独立本地处理

智慧安防

超大数据量本地处理

移动互联网

低功耗、数据安全

其他物联网场景

家庭

工业/农业等

数据安全、异构、实时响应

边缘智能服务器：GPU/FPGA/ASIC

终端 SoC+IP：DSP/GPU/ASIC

来源：艾瑞研究院自主绘制。

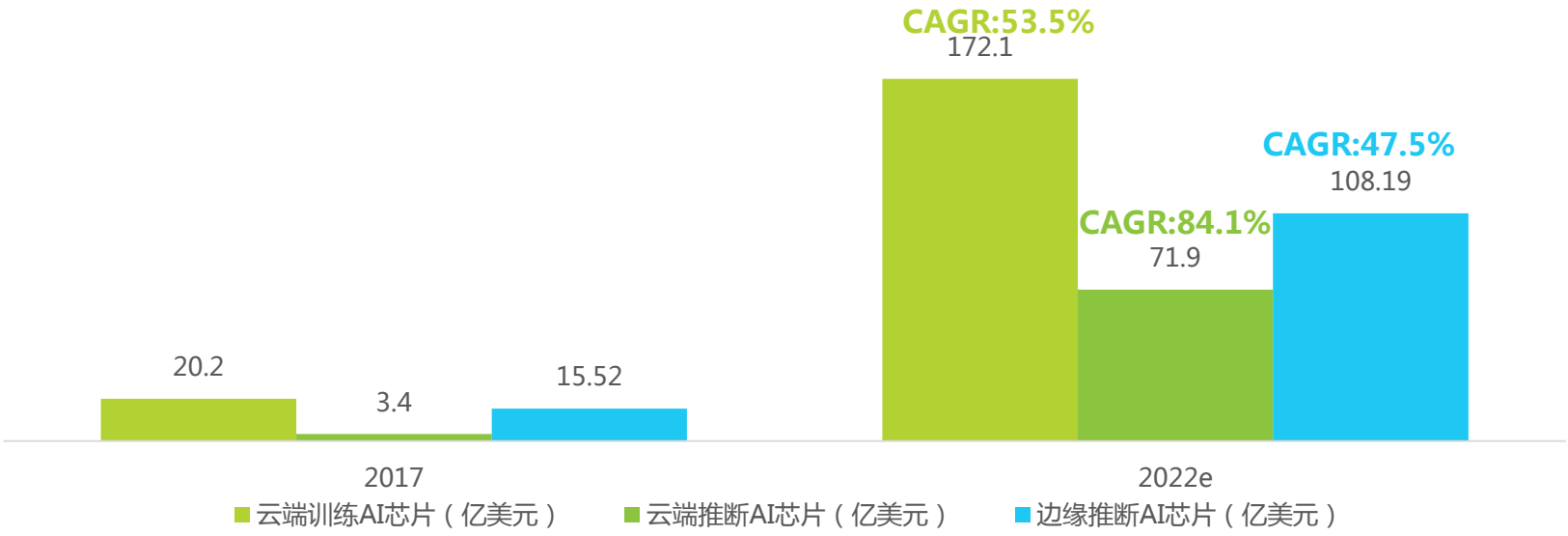
AI芯片整体市场规模预测

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芯片细分市场规模预测



注释：AI芯片细分市场规模单位：亿美元。
来源：《AI芯片：应用落地推动产品多样化》——中金公司研究部

应用场景：云计算

云计算：共享规模化经济效益有效降低边际成本投入

云计算是一种按使用计费的IT服务模型，实现对高可靠、可配置的计算资源池（服务器、存储、网络、应用程序和服务）的方便快捷的访问，资源可通过最少的管理工作快速的配置和发布。云计算具有：资源池、广泛的网络访问、按需自助服务、快速弹性膨胀、测量服务等5个基本特征。云计算服务模式主要包括：IaaS、PaaS、SaaS：

1. IAAS-提供基本的计算（虚拟或专用硬件）、存储、网络资源，使用者在资源中部署运行任意应用程序和操作系统；
2. PAAS-提供部署在云基础设施上的编程语言、库、服务和支持工具，为开发人员提供了一个自助服务门户而无需管理底层基础设施；
3. SAAS-提供在云基础设施上运行的应用程序，程序运行管理皆由服务提供商负责。

相比起传统IT模式，云计算模式可实现：降低用户初始IT投资成本及IDC机房维护费用并实现资本效益配置最大化、IT资源快速弹性扩展、数据价值的有效挖掘以及业务的快速上线部署等。

云计算服务模型

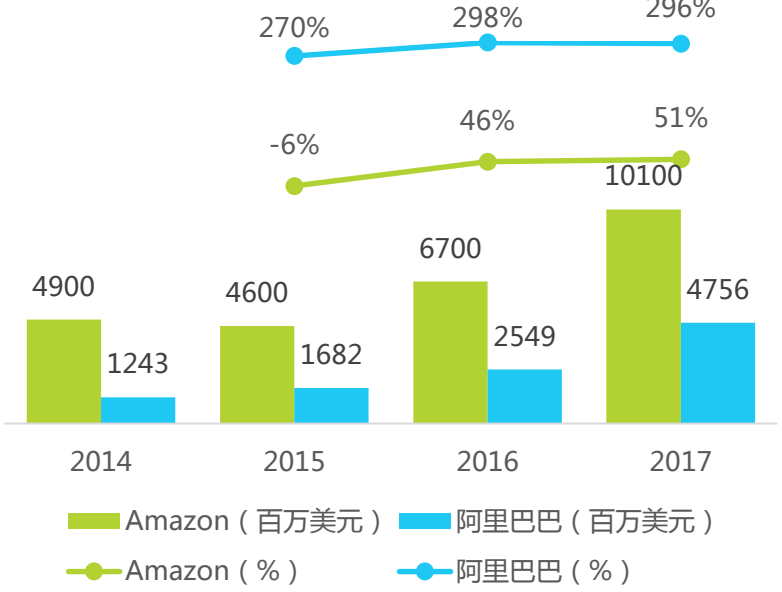
来源：IBM。

云计算中心服务器及硬件市场规模

云计算发展带动上游硬件市场需求

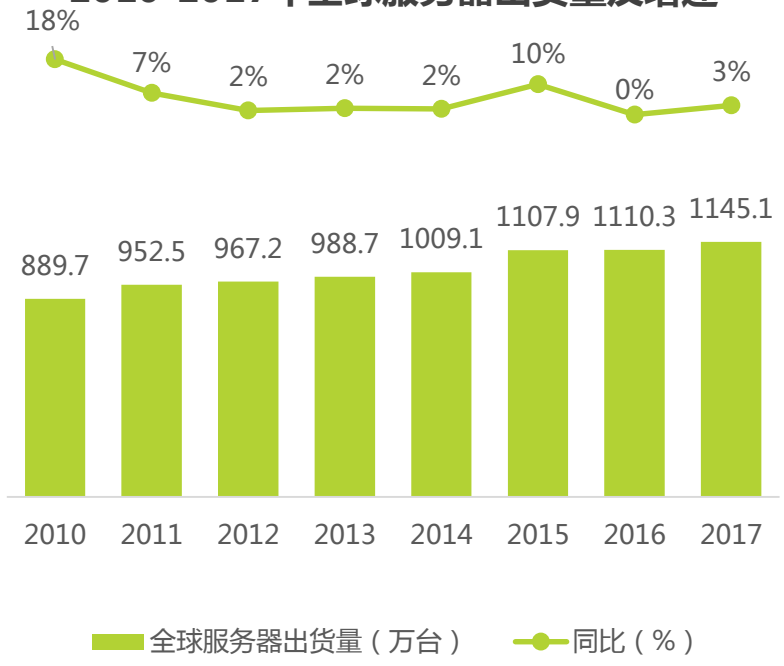
近些年云厂商巨头快速扩张带动了上游数据机房建设热潮，虽然这会抵消部分企业自建机房及采购服务器的需求，但是这部分需求并没有消失而是转移到了对云数据中心IT资源的需求上。中国云产业比美国发展晚2~4年左右，在全球云计算市场中，美国占比达到54.1%，中国仅为5%。2017年中国IT支出为2.4万亿人民币，仅占全球IT总支出金额的9.92%，对应中国GDP水平仍有较大的提升空间。中国虽然起步较晚但发展迅猛，对比亚马逊及阿里巴巴云计算资本支出数据，可以看到阿里巴巴在云计算领域支出总金额绝对值虽然少于亚马逊，但增速却显著超过了亚马逊的资本支出增速，这显示了近年来国内加大对于计算资源基础设施的建设力度，将极大受益于服务器厂商及上游芯片厂商业绩。

2014-2017年 Amazon&阿里巴巴云计算建设资本支出



来源：AMAZON、阿里巴巴公司年报披露数据。

2010-2017年全球服务器出货量及增速



来源：Wind。

云计算与人工智能服务

云计算为AI开发部署提供多元化服务支持

云计算服务供应商可以向客户提供包括计算、存储、数据库、分析、移动、物联网和企业应用等一系列多层级服务。在人工智能应用方面，由于人工智能（以深度学习为代表）的开发及应用对于算力、数据有较大的需求，而云计算服务可以为开发者提供AI计算芯片以及基于其开发的智能服务器集群等强大算力设施的租用，同时也可以为开发者提供PaaS级的开发平台或是直接提供已训练好的人工智能功能模块等产品。通过多元化的服务模式，可以降低开发者的开发成本和产品开发周期，为客户进行方便快捷的AI赋能。

云计算核心产业链及云端人工智能服务模式

注释：AI服务器：指搭载了AI芯片、计算板卡并用于人工智能相关运算的服务器。
来源：AWS官网。

云计算与人工智能

云计算服务模式可显著降低AI应用开发、部署成本

云计算发展自互联网厂商提升服务器使用效率（考虑为应付黑色星期五、双11等特殊日期访问量激增而添置的巨量服务器资源）而逐渐开始的服务器Web租赁服务。在云计算产业链中，云厂商负责基础设施和云组织架构的搭建，并为客户提供PaaS、SaaS服务，具有极高的资本和技术门槛，在产业链中享有极大的话语权。如前文所述，在AI开发中，由于深度学习模型开发及部署需要强大算力支持，需要专用的芯片及服务器支持。开发者如选择自购AI服务器成本过高。通过云服务模式，采取按需租用超算中心计算资源可极大降低项目期初资本投入同时也省却了项目开发期间的硬件运维费用，实现资本配置效率的最大化提升。

部分云端AI芯片售价

芯片	功能	售价：¥
V100（GPU）	训练	10万元，包含硬件+软件（驱动、许可、保修）
P4（GPU）	推断	2万元，包含硬件+软件
FPGA	推断	4~5000元

部分云端AI智能服务器售价

服务器	AI芯片	CPU	售价：$
DGX-2	16*Tesla V100	2*intel Xeon Platinum 8168	$399,000
DGX-1 (Volta)	8*Tesla V100	2*intel Xeon E5-2698 v4	$149,000
DGX-1 (Pascal)	8*Tesla P100	2*intel Xeon E5-2698 v3	$129,000

云计算AI服务收费（AWS）

任务	硬件	收费：$/h
构建	标准实例	0.05~6.45
	GPU实例	1.26~34.27
训练	标准实例	0.13~6.45
	GPU实例	1.26~34.27
模型部署	标准实例	0.07~6.45
	GPU实例	1.26~34.27

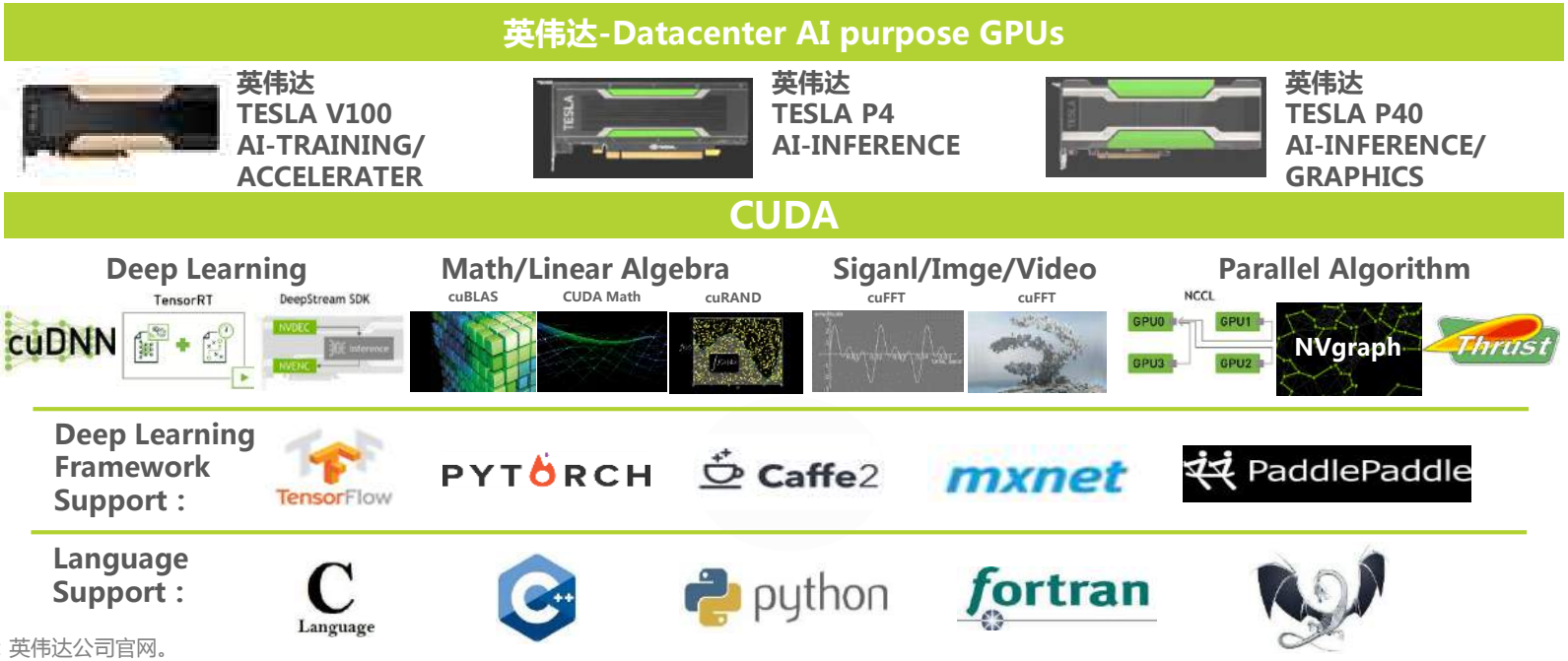
来源：英伟达官网、亚马逊官网、公开网络数据、专家访问。

云计算与AI芯片

GPGPU+CUDA方案提供丰富的AI开发SDK及广泛适用性

英伟达除了在传统独立显卡领域有近7成的市场份额，其在云计算智能服务器领域市场份额更是一家独大。英伟达为客户提供了支持AI应用开发的完备的TESLA GPU产品线，相比于传统CPU服务器，在提供相同算力情况下，GPU服务器在成本、空间占用和能耗分别为传统方案的1/8、1/15和1/8。除了优秀的硬件性能外，英伟达开发了基于GPU的“CUDA”开发平台，为开发者提供了丰富的开发软件站SDK，支持现有的大部分的机器学习、深度学习开发框架 ，开发者可以在CUDA平台上使用自己熟悉的开发语言进行应用开发。公司花费大量时间培养自己的开发生态，包括与高校合作培训专业人才、开展专业竞赛，培养、发展英伟达“GPU+CUDA”的开发者群体，形成了相当可观的产品使用人群，构建了当前英伟达在人工智能领域的霸主地位。

英伟达云端软硬件一体解决方案



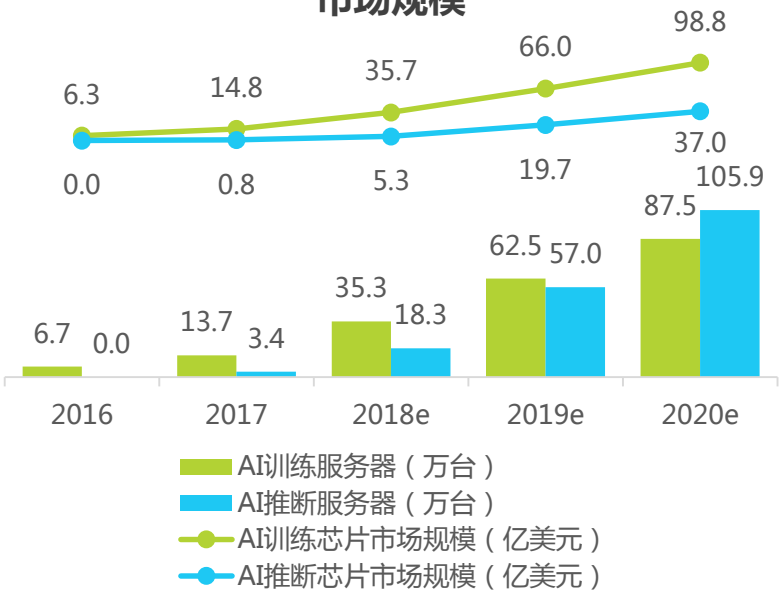
来源：英伟达公司官网。

云计算中心AI芯片市场规模及份额

GPU并非完美，市场期待替代GPU的云端AI芯片解决方案

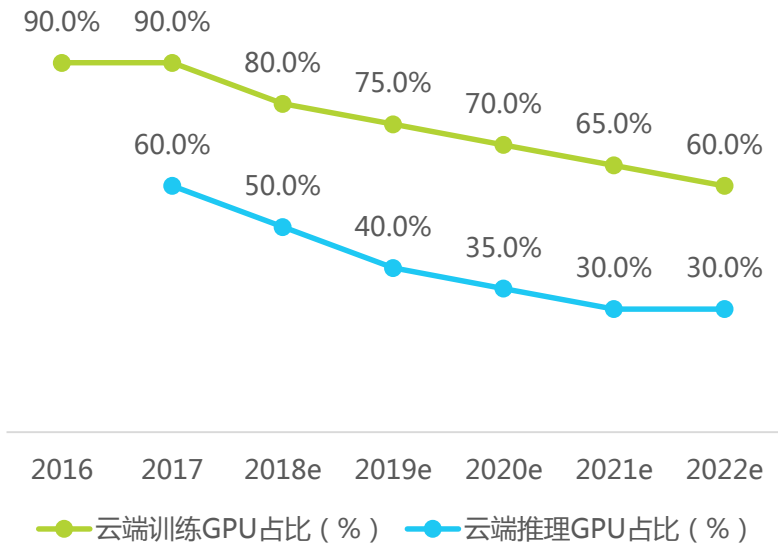
当前全球云计算AI芯片市场英伟达一家独大（尤其是训练端），主要原因是英伟达GPU产品线丰富，编程环境成熟，产品支持市场上主要的开发框架和语言，产品广受AI开发者好评。但同时其产品也存在着功耗偏大、价格昂贵等问题（V100芯片售价达10万元，DGX系列服务器售价过百万元）。基于此，各大云厂商纷纷提出自己的AI芯片开发计划以摆脱上游AI芯片供货商一家独大的垄断市场情况。此外根据数据显示，推断市场未来增速和空间将会高于训练端市场，而GPU芯片并不善于推断任务，因此，在当前智能服务器渗透率尚低，GPU产品并非完美解决方案的情况下，我们认为对于其他AI芯片厂商云计算中心市场依然存在着较大的市场空间可以进入。

2016-2020年全球AI服务器及AI芯片市场规模



来源：IDC， Gartner。

2016-2022年全球云端AI芯片GPU市场份额占比



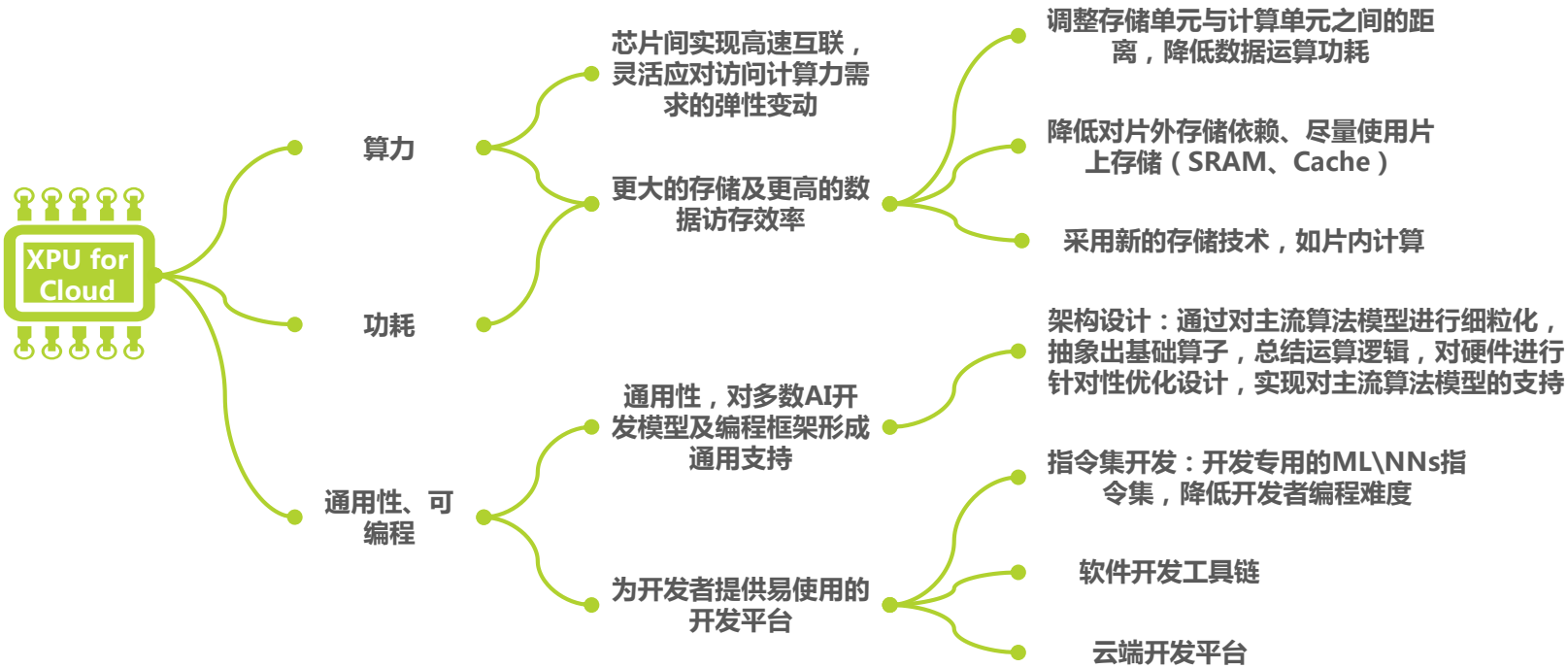
来源：IDC。

云计算AI芯片发展趋势

看好基于AI专用指令集的可编程ASIC芯片及配套开发平台

如前文所述，当前在云端场景下被最广泛应用的AI芯片是英伟达的GPU，主要原因是：强大的并行计算能力（相比CPU）、通用性以及成熟的开发环境。但是GPU也并非是完美无缺的解决方案，明显的缺点如：高能耗以及高昂的价格。目前包括创业公司、科技巨头等都在积极寻找GPU的替代方案，希望实现：既具有GPU通用性、又具有更好的能效和算力表现的通用、可编程产品。当前市场上典型的替代方案包括Google的TPU系列以及寒武纪的MLU系列产品。

云计算场景AI芯片设计思路及发展趋势



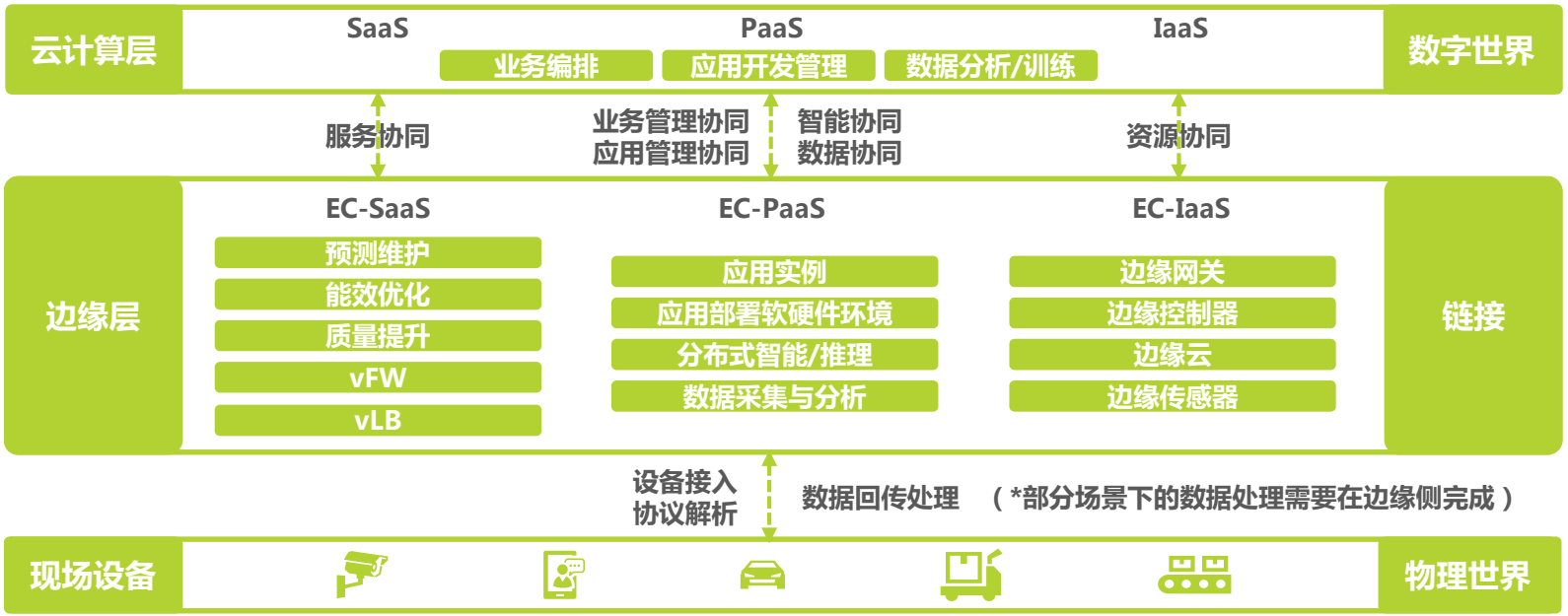
来源：艾瑞研究院自主绘制。

应用场景：边缘计算

边云协同共同实现万物互联时代计算任务需求

边缘计算：在靠近数据源头的网络边缘侧，融合网络/计算/存储/应用等核心能力的分布式开放平台，就近提供边缘智能服务，具有海量联接、实时业务处理、数据优化、应用智能、安全与隐私保护等特点。边缘计算对软硬件系统提出了：1）海量异构联接、2）计算任务在边缘节点实时处理响应、3）硬件功耗/成本/空间/抗干扰等有严格要求、4）分布式资源的动态调度与统一管理、5）支持联接/数据/管理/控制/应用/安全等方面的协同等要求。边云协同放大边缘计算及云计算价值：边缘计算承担数据采集和部分的数据处理任务，支撑云端应用，而云计算通过大数据分析，优化输出的业务规则或模型，下发到边缘侧，为终端提供运行规则/模型。

边云协同组织架构图



来源：《边缘计算与云计算协同产业白皮书》——工业互联网产业联盟

边缘计算场景与人工智能芯片

边缘侧场景繁杂，综合考量AI芯片 “PPACR”

在边缘计算场景，AI芯片主要承担推断任务，通过将终端设备上的传感器（麦克风阵列、摄像头等）收集的数据代入训练好的模型推理得出推断结果。由于边缘侧场景多种多样、各不相同，对于计算硬件的考量也不尽相同，芯片可以是IP in SoC，也可以是边缘服务器，对于算力和能耗等性能需求也有大有小。因此不同于云端场景的“高端、通用”，应用于边缘侧的计算芯片需要针对特殊场景进行针对性设计以实现最优的解决方案。

不同边缘计算场景对AI芯片“PPACR”性能考察要求

		物联网场景	移动互联网	智能安防	自动驾驶
任务描述： 机器视觉 语音识别/ 自然语义处理		1、图像检测 2、视频检测 3、语音识别 4、语义理解	1、照相-场景识别 2、照相-美化 3、AR应用 4、语音助手	1、图像检测 2、视频检测	1、图像语义分割 2、数据融合 3、Slam定位 4、路径规划
性能要求	算力 **Performance**	<1TOPs	1~8TOPs	4~20TOPs	20~4000TOPs（L3~L5）
	能耗 **Power**	接入设备部署现场电源	消费级聚合物锂电池 2,000~5,000mAh	接入设备部署现场电源	动力级硬壳锂电池（组） 200,000~500,000mAh
	面积 **Area**	高（SoC）	极高（SoC）	高~低（SoC、Server）	中（PCIE contains multiple SoC Chips）
	成本控制 **Cost**	高	极高	高~低（IPC，NVR）	中
	可靠性 **Reliability**	高（工业）/中（家用）	中	高	极高
代表厂商		Google NVIDIA Horizon Robotics 云知声 Unisound 思必驰 AISPEECH	ARM cadence 苹果 MEDIATEK SAMSUNG QUALCOMM Cambricon	HISILICON NVIDIA 安霸科技 BITMAIN 君正 Ingenic intelliFusion Horizon Robotics	NVIDIA Horizon Robotics NXP intel MOBILEYE

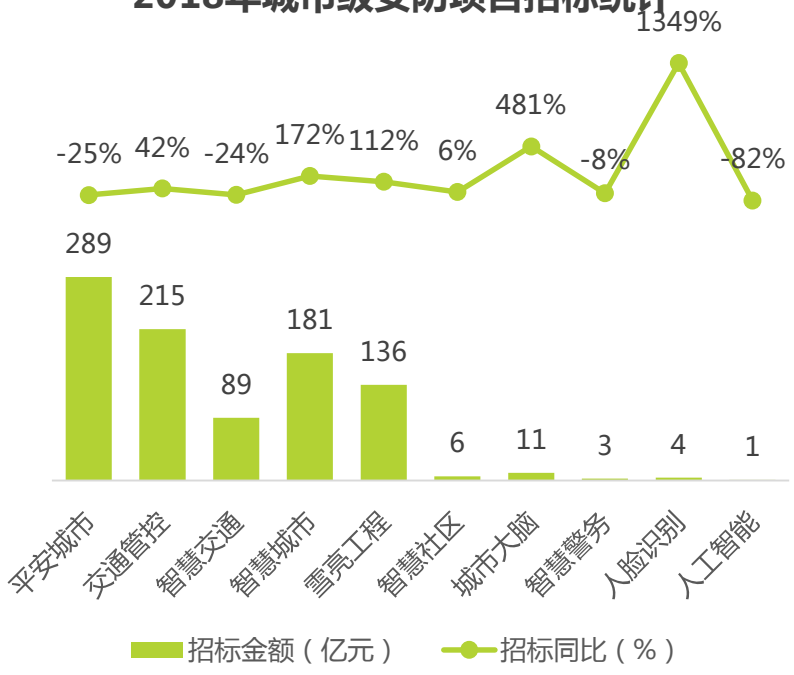
来源：网络公开数据、艾瑞研究院自主绘制。

应用场景1：智慧安防市场

当前国内基于G端需求最明确的边缘计算AI应用场景

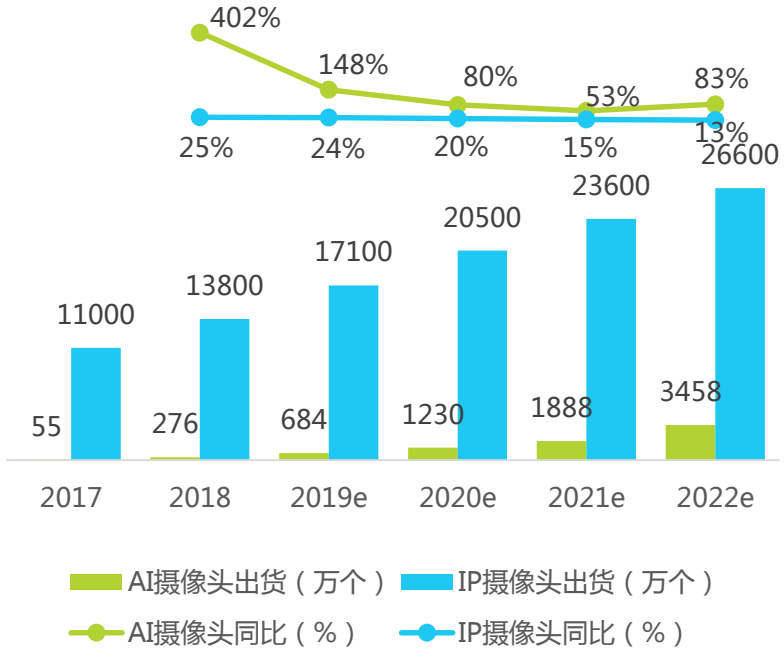
在国内，安防市场是最为确定的边缘侧AI应用场景，主要原因是大量的监控视频数据分析对人力的需求与当前基层人力缺失、人力成本上升之间的显著矛盾。需求端来自于包括政府、大中企业以及个人安防需求，主要市场需求来自于政府，包括平安城市、智慧交通管控、智慧城市、雪亮工程等，其中公安部“雪亮工程” 提出到2020年实现“全域覆盖、全网共享、全时可用、全程可控” ，智能摄像头的应用可以有效解决基层数据传输带宽压力以及基层警力人员缺失等问题，预计“雪亮工程”等政府项目将会对智能安防摄像头市场带来较大的驱动作用。

2018年城市级安防项目招标统计



来源：AI智道，不完全统计。

2017-2022年摄像头AI芯片市场空间



来源：IDC。

©2019.6 iResearch Inc. www.iresearch.com.cn

智慧安防与AI芯片

AI芯片为摄像头提供边缘智能解决方案降低数据回传需求

安防摄像头发展经历了由模拟向数字化、数字化高清到现在的数字化智能方向的发展，最新的智能摄像头除了实现简单的录、存功能外，还可以实现结构化图像数据分析。安防摄像头一天可产生20GB数据，若将全部数据回传到云数据中心将会对网络带宽和数据中心资源造成极大占用。通过在摄像头终端、网络边缘侧加装AI芯片，实现对摄像头数据的本地化实时处理，经过结构化处理、关键信息提取，仅将带有关键信息的数据回传后方，将会大大降低网络传输带宽压力。当前主流解决方案分为：前端摄像头设备内集成AI芯片和在边缘侧采取智能服务器级产品。前端芯片在设计上需要平衡面积、功耗、成本、可靠性等问题，最好采取低功耗、低成本解决方案（如：DSP、ASIC）；边缘侧限制更少，可以采取能够进行更大规模数据处理任务的服务器级产品（如：GPU、ASIC）。

AI芯片在智能安防摄像头中的应用

模拟监控系统

- **前端：ISP**
  核心芯片，对原始图像信号进行降噪、曝光调整，决定最终成像效果的好坏；
- **后端：DVR SoC**
  将模拟音频信号数字化、编码压缩与存储（A/D芯片和视频编解码芯片）。

数字/网络高清

- **前端：IPC SoC**
  集成CPU、ISP、高压缩比视频编解码模块、网络接口、加密模块、内存子系统等；
- **后端：NVR SoC**
  接收摄像机的IP码流进行编解码、存储，适用于环境复杂和分散的大型监控系统。

智能化升级

- **前端：IPC SoC+AI-IP/独立AI芯片**
  -在现有IPC上集成算法实现识别任务；
  -SoC中集成协处理器或增加独立AI芯片进行结构化分析或运行DL算法提升检出率；
- **后端：GPU/ASIC智能服务器**
  将智能推理功能集成在边缘的服务器级产品中，实现更大规模的人工智能应用。如：GPU服务器或最新ASIC服务器方案。

结构化分析摄像机 | **智能网络摄像机** | **深度学习摄像机**

公安 | **交通** | **智能楼宇** | **金融** | **能源** | **司法** | **文教卫**

来源：艾瑞研究院自主绘制。

智慧安防芯片市场

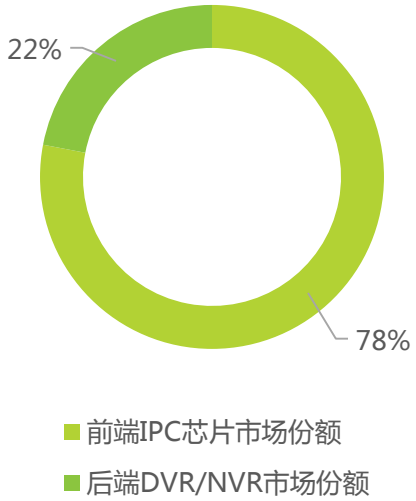
市场考验厂商综合服务能力，海外厂商逐步退出国内市场

摄像头由于任务相对单一，行业内产品同质化程度较高，导致行业竞争激烈。智能安防芯片领域参与者包括国际一线厂商如：英伟达、华为、安霸、TI等，还有创业公司如：比特大陆、地平线、云天励飞等，以及北京君正、国科微等传统安防半导体企业，甚至下游的安防厂商如海康威视、大华股份也开始自研AI芯片产品。英伟达作为AI巨头在安防领域有前、后端完整软硬件解决方案，国内诸多安防厂商均采购其产品；华为推出了Hi3559等智能芯片，虽然算力性能表现暂时不如英伟达，但作为在国内传统IPC市场有近7成市场份额的企业，基于其深厚的市场积累，能够为客户提供完善的、高性价比的解决方案。在面临越来越多的企业涌入安防芯片市场的情况下，行业客户倾向于采购完整的方案，因此除了考察单一芯产品性能以外，更加考察企业的行业积累及整体解决方案设计及交付能力。

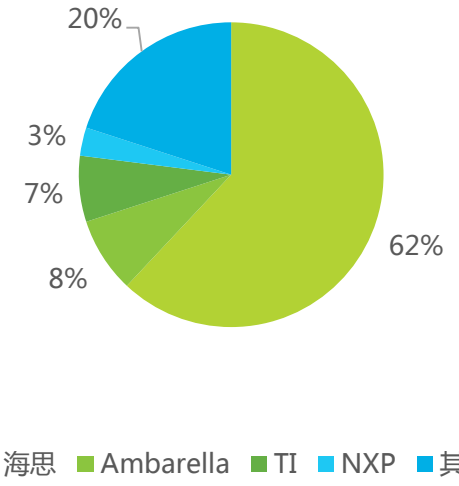
主要安防AI芯片解决方案

	厂商	产品	算力
前端	英伟达	JETSON TX1-GPU	1TOPs
	华为海思	Hi3559AV100	4TOPs
	北京君正	T01-ASIC	
	地平线	旭日	1TOPs
后端	英伟达	TESLA P4	5.5TFlops
	比特大陆	BM1682	3TFlops

2016年安防芯片市场份额



2016年IPC视频编解码芯片市场份额

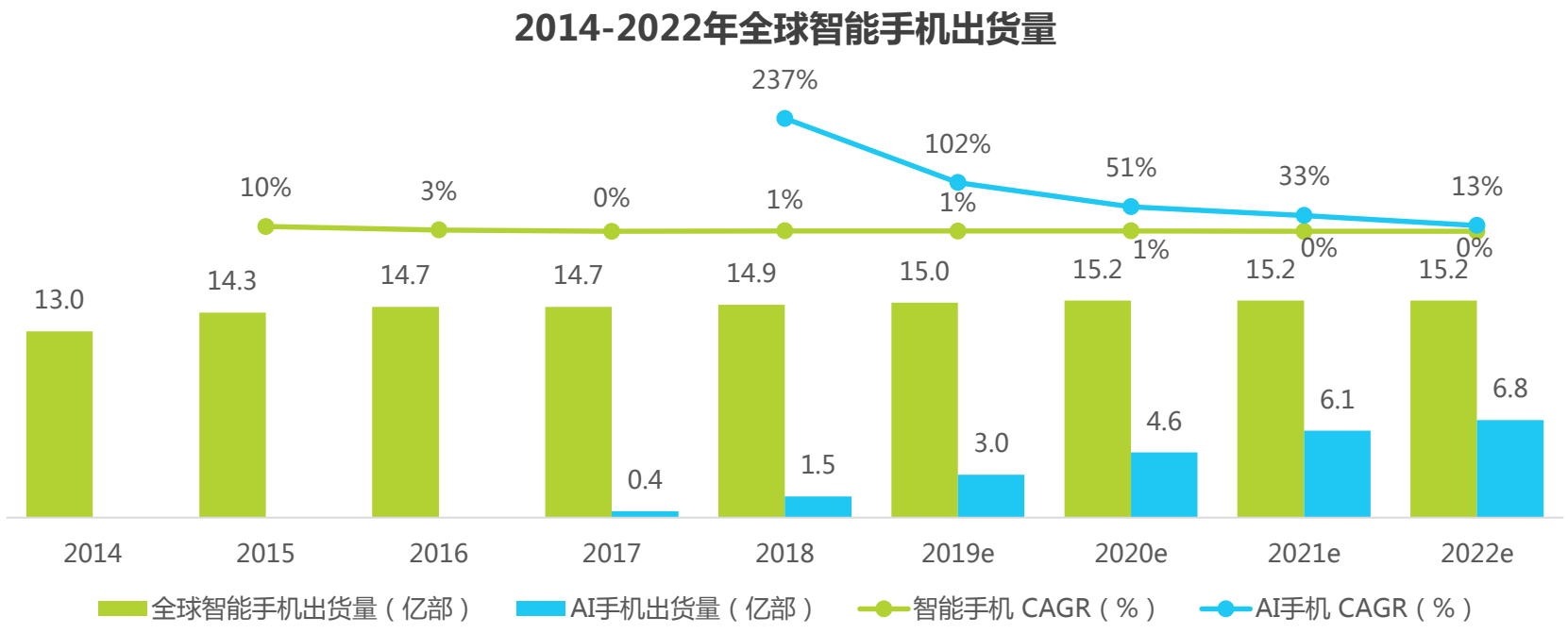


来源：CPS中安网、IHS、公开网络数据。

应用场景2：移动互联网市场

AI芯片为移动互联网消费电子类产品带来新增市场空间

智能手机在经历了近10年的高速增长后，市场已趋于饱和，出货增速趋近于0，行业逐渐转为存量市场。近年来，一批国产厂商在产品质量上逐渐达到了第一梯队的水平，进一步加剧了头部市场的竞争。为实现差异化竞争，各厂商加大手机AI功能的开发，通过在手机SoC芯片中植入AI芯片实现在低功耗情况下AI功能的高效运行。随着未来竞争进一步加剧，以及产量上升所带来的成本下降，预计AI芯片将会进一步渗透进入到中等机型市场，市场空间广阔。移动端AI芯片市场不止于智能手机，潜在市场还包括：智能手环/手表、VR/AR眼镜等市场。AI芯片在图像及语音方面的能力可能会带来未来人机交互方式的改变并进一步提升显示屏、摄像头的能力，有可能在未来改变移动端产品。

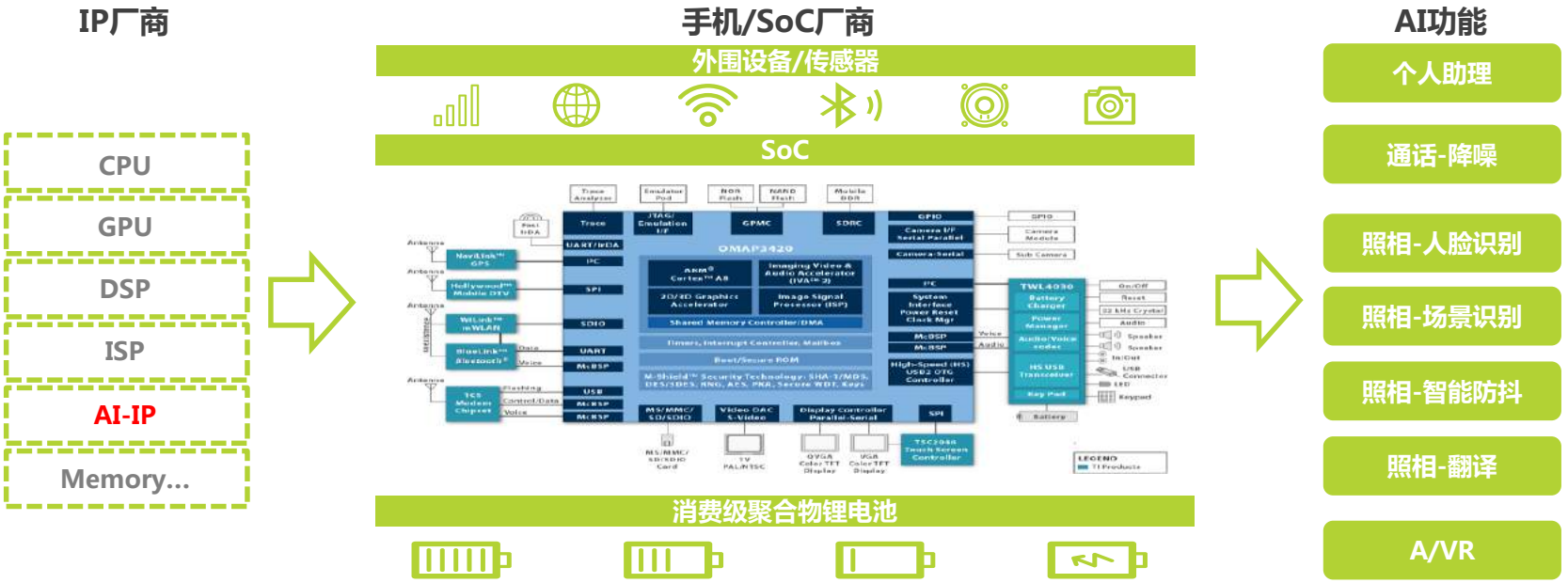


来源：IDC。

移动互联网与AI芯片

智能手机：SoC内嵌AI IP实现手机AI功能的高效执行

通过云数据中心做手机端AI推理任务面临网络带宽延迟瓶颈的问题，严重影响用户使用体验，而CPU适合逻辑运算，但并不适合AI并行运算任务，目前市场上流行在SoC中增加协处理器或专用加速单元来执行AI任务。以智能手机为代表的移动互联网终端是一个多传感器融合的综合数据处理平台，AI芯片需要具备通用性，能够处理多类型任务能力。由于移动终端依靠电池驱动，而受制于电池仓大小和电池能量密度限制，芯片设计在追求算力的同时对功耗有着严格的限制，可以开发专用的ASIC芯片或者是使用功耗较低的DSP作为AI处理单元。



当今手机电池电容量普遍在2000~5000mAh。有限的电量需要被被分配到AP、CP中的射频、CPU、GPU、ISP等诸多电子元器件，用于信号接发、编解码、摄像头、图像处理/渲染等多类型任务，对电子元器件功耗设计提出了极高的要求。通过设计专用的AI加速运算单元并植入在SoC中，在功耗可控的情况下可实现高效的**执行AI运算任务。**

来源：艾瑞研究院自主绘制。

移动互联网芯片市场

传统手机芯片厂商技术实力强劲，行业壁垒较高

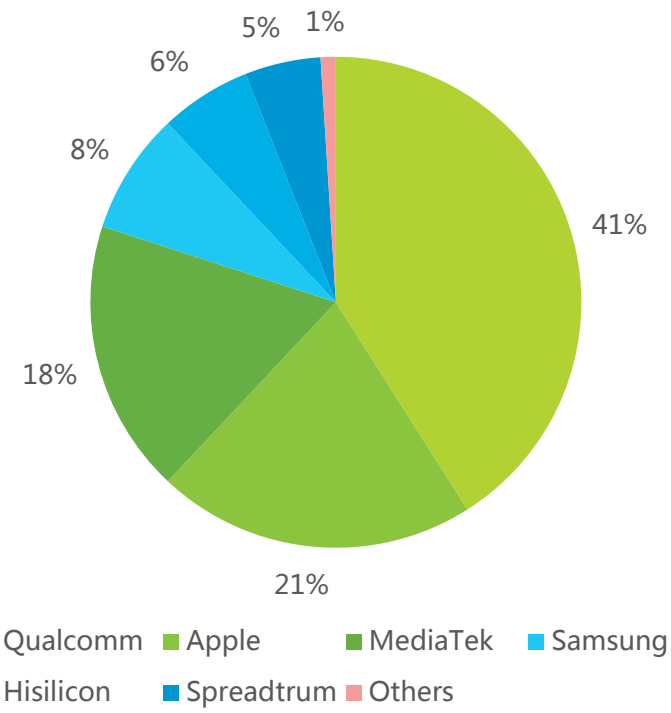
目前手机芯片市场存在以下情况：1）、AI应用场景、功能有限；2）、AI芯片厂商一般向SoC厂提供IP并收取授权费，需要AI-IP与整块SoC进行良好的匹配，而创业公司缺少与SoC厂商合作经验；3）、传统手机SoC厂商和IP厂商都在开发自己的AI加速器，传统IP巨头可以采取IP打包销售的方式推广其AI-IP产品。相比之下新进厂商在成本、功能、产品线、匹配度等都不占优的情况下很难在该领域存活。新进厂商应加强其软件方面优势，并加深与手机厂商合作共同进行手机AI功能开发。

主要手机AI芯片解决方案

厂商		SoC	AI技术	性能
SoC厂商	高通	骁龙845	Hexagon 685 DSP+CPU+GPU	
		骁龙855	Hexagon 690 DSP+CPU+GPU	7TOPs，比上一代提升3倍
	苹果	A11	Neural Engine	0.6TOPs
		A12	Neural Engine	5TOPs
	MTK	P60	APU	0.56TOPs
		P90	APU2.0	2.25TOPs
	华为	Kirin970	NPU 1A	0.512TOPs
		Kirin980	NPU 1H	5TOPs
IP厂商	ARM	N/A	ML/OD Processor	
	Cadence	N/A	P5/P6/Q6/C5	
	Synopsys	N/A	EV5x/6x	
	CEVA	N/A	XM6/XM4	

来源：公开网络数据，艾瑞研究院。

2017年手机SoC芯片市场份额

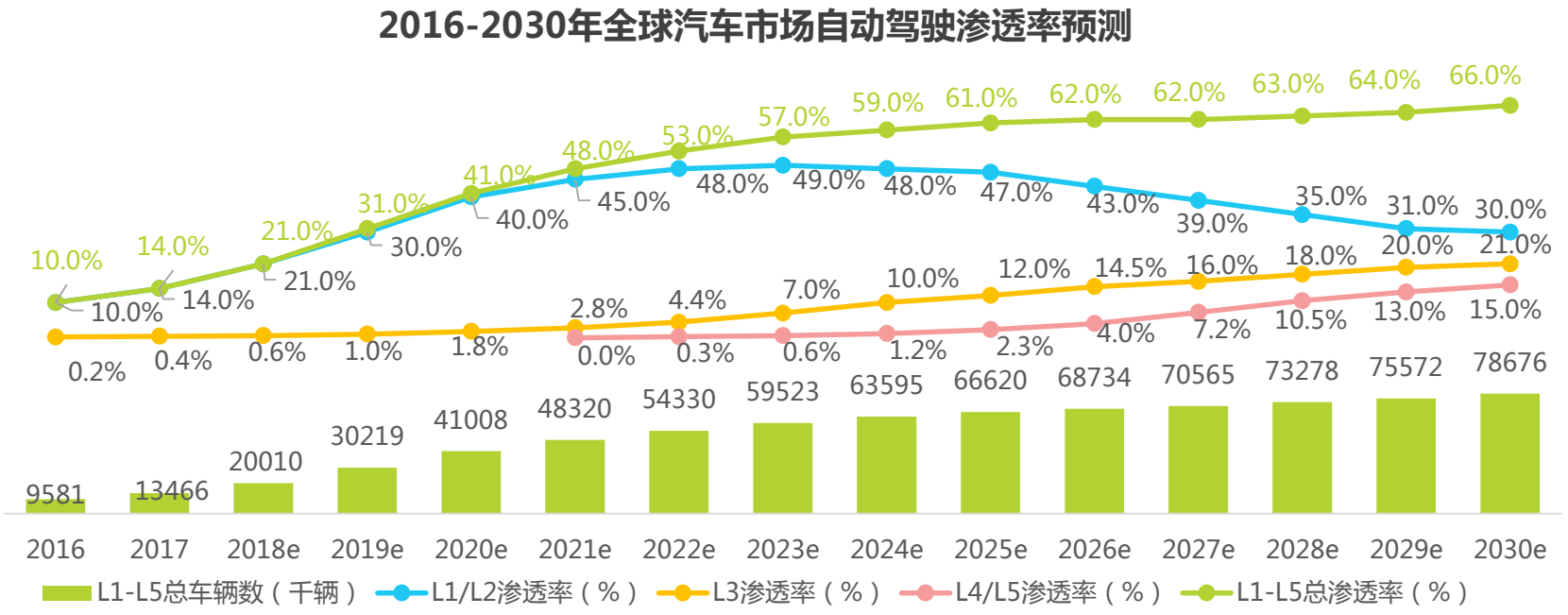


来源：CounterPoint，中金研究部。

应用场景3：自动驾驶

前景广阔，但5~10年内L4~5进入乘用车平台困难较大

根据美国汽车工程师协会（SAE）将自动驾驶按照车辆行驶对于系统依赖程度分为L0~L5六个级别，L0为车辆行驶完全依赖驾驶员操纵，L3级以上系统即可在特定情况下实现驾驶员脱手操作，而L5级则是在全场景下车辆行驶完全实现对系统的依赖。目前商业化乘用车车型中仅有Audi A8、Tesla、凯迪拉克等部分车型可实现L2、3级ADAS。预计在2020年左右，随着传感器、车载处理器等产品的进一步完善，将会有更多的L3级车型出现。而L4、5级自动驾驶预计将会率先在封闭园区中的商用车平台上实现应用落地，更广泛的乘用车平台高级别自动驾驶，需要伴随着技术、政策、基础设施建设的进一步完善，预计至少在2025年~2030年以后才会出现在一般道路上。



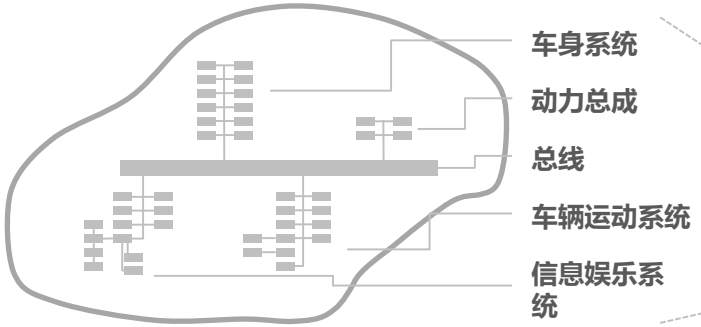
来源：IHS,，天风证券研究所，艾瑞研究院。

自动驾驶与汽车电子发展趋势

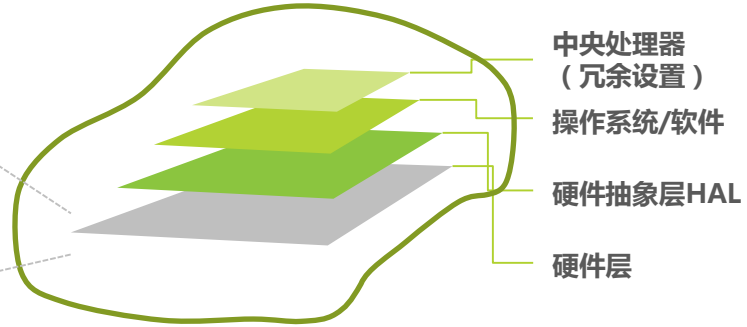
高级ADAS/自动驾驶需要中央“CPU+XPU”异构处理器

目前汽车电子控制系统是分布式ECU架构，不同的信息娱乐、车身、车辆运动和动力总成系统及其细分功能分别由不同独立的ECU单元进行独立控制，部分高档车型上的ECU数量超过100个。未来随着汽车进入L3级以上的高级别自动驾驶时代，随着车载传感器数量及其所产生的数据量剧增，分布式电子系统难以满足对大量、多元的传感器数据进行高效融合处理，并综合所有传感器数据做出车辆控制决策等一系列操作需求。要满足以上功能需求，汽车电子系统由需要向着域控制器（DCU）、多域控制器（MDC）等集中化方向发展，未来，汽车电子操控系统将会进一步向着集中化、软硬件解耦及平台化方向发展，汽车将会由统一的超算平台对传感器数据进行处理、融合、决策最终实现高级别的自动驾驶功能。

传统：分布式ECU汽车电子操纵系统



未来：中央计算单元+操作平台



		Lv1~2	Lv3	Lv4~5
传感器	毫米波	1~3	4~6	6~10
	摄像头	1	2~4	6~8
	激光雷达	N/A	0~1	1~3
算力需求		<1TOPS	10~50TOPS	>50TOPS

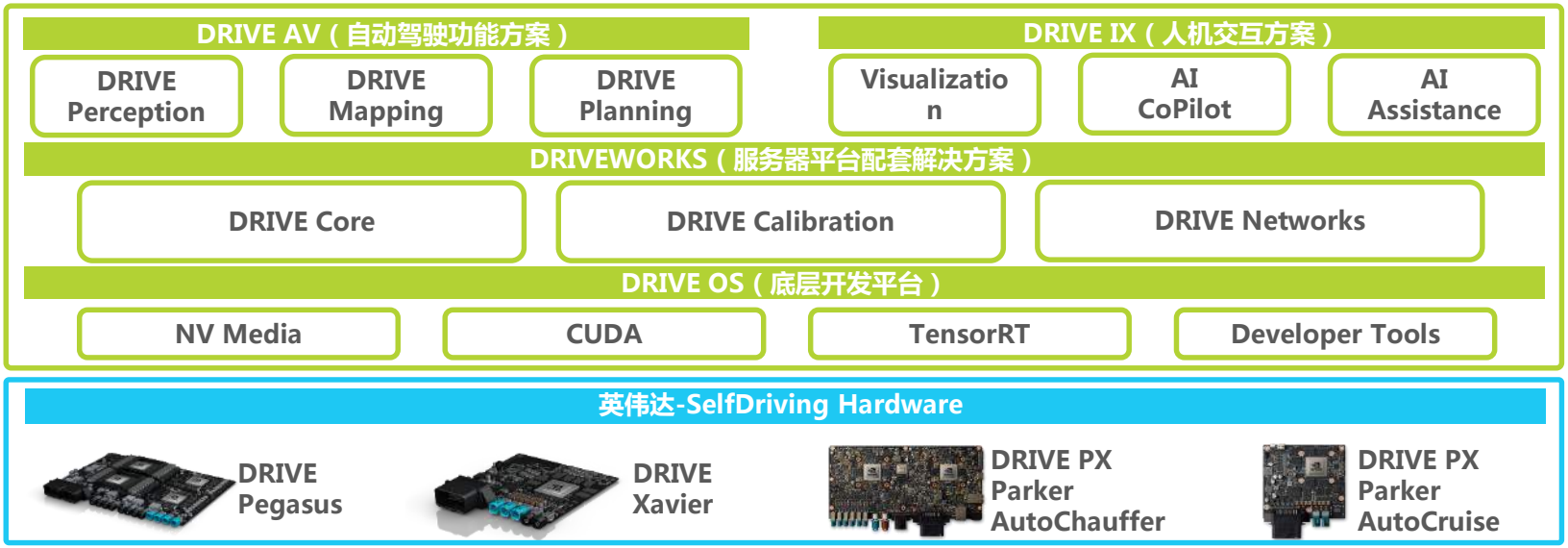
来源：恩智浦、德尔福、公开网络数据，艾瑞研究院自主绘制。

自动驾驶与AI芯片

AI芯片厂商需提供可编译的“硬件+软件”产品解决方案

伴随人工智能技术在视觉领域的应用，基于视觉技术的自动驾驶方案逐渐变为可能，这需要在传统行车电脑平台上添加用于视觉算法处理的AI芯片。自动驾驶汽车计算单元设计需要考虑算力、功耗体积等问题，出于硬件资源最优化应用，往往采取异构计算平台设计方案，及“CPU+XPU”（XPU包括：DSP/GPU/FPGA/ASIC），其中可采取DSP用于图像特征提取任务、GPU/FPGA/ASIC等计算单元用于目标识别、追踪任务等，而CPU则会用于定位、决策等逻辑运算任务。
目前最典型的产品如英伟达的DRIVE PX系列及后续的Xavier、Pegasus等。除硬件计算平台外，英伟达为客户提供配套的软件平台及开放的上层传感器布局和自定义模块使得客户能够根据自身需要进行二次开发，其还为客户提供感知、制图以及行驶策略等解决方案。目前其产品已经被包括ZF、Bosch、Audi、Benz以及Tesla等Tier1s、OEMs厂商及诸多自动驾驶创业公司采用作为其处理器方案所使用。

行业需要完整的硬件+软件整体解决方案（英伟达 GPU+Software Stack）



来源：英伟达官网。

自动驾驶芯片市场

前装市场壁垒高企，企业需有深厚的汽车电子设计经验积累

在全部的边缘计算场景中，用于自动驾驶的计算芯片设计难度最大，这主要体现在：1）算力要求高， L3级以上自动驾驶需要复数种类的传感器实现传感器冗余，包括：6~12颗单目摄像头、3~12台毫米波雷达、5台以内的激光雷达等（不同方案配置侧重不同），因此产生的数据量极大（ 估计L5级一天可产生数据量4000GB），在车辆高速行驶的情况下系统需要能够快速对数据进行处理；2）汽车平台同样是由电池供电，因此对于计算单元功耗有较高的要求，早期计算平台功耗大、产热也较大，对于系统的续航及稳定性都有较大的印象；3）汽车电子需要满足ASIL-D车规级电子产品设计标准，而使自动驾驶所需要的中央处理器达到ASIL-D级设计标准难度更大。

目前自动驾驶市场尚处于发展早期，市场环境不够成熟，但以英伟达、Intel（Mobileye、Altera）等科技巨头为代表的厂商已经投入巨资在该领域开发出了相关的硬件产品及配套软件技术。人工智能芯片创业公司应该加强与OEMs、Tier1或产业联盟合作为其提供AI芯片+软件工具链的全套解决方案。

行业主流自动驾驶核心处理器芯片解决方案

厂商	SoC	结构	性能
英伟达	AutoCruise	2*Denver+4*Cortex A57+2*Pascal GPU 256 CUDA cores	
	AutoChauffeur	4*Denver+8*Cortex A57+4*Pascal GPU 512 CUDA cores+2*dedicated M*M modules	20TFLOPS（FP16）/250W
	Xavier	8* custom ARM64+Volta GPGPU 512 CUDA cores	20 TOPS（INT8 ）/30W
	Pegasus	16* Custom ARM64+2* Volta GPGPU CUDA cores+2*Volta GPU	320TOPS(（INT8）/500W
Mobileye（Intel）	EyeQ3		0.256TFLOPS/2.5W
	EyeQ4	MIPS Warrior CPU+6*VMP（Vector Microcode Processors）+MPC（Multithreaded Processing Cluster）；	2.5TOPS/3W
	EyeQ5	8*CPU+18*Computer Vision Processors	17TOPS/5W

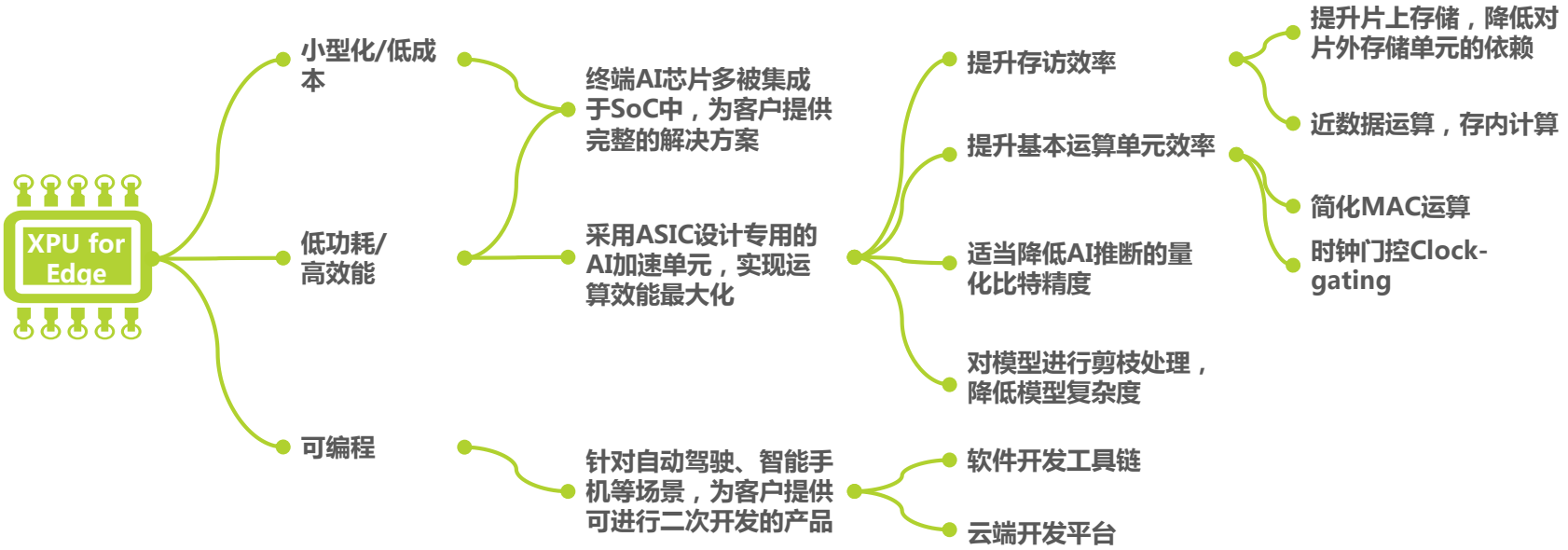
来源：英伟达官网、Mobileye官网。

边缘计算AI芯片技术发展趋势

企业应该具有系统级设计能力，为客户提供完整个解决方案

边缘计算场景呈现多样化分布，除我们提到的安防监控、移动互联网、自动驾驶以外还涉及智慧家具、智能制造、智慧医疗等多样化物联网场景。不同的应用场景基于计算任务、场景限制等，对AI芯片在算力、功耗、成本等方面提出了不同的要求。但总体来看，相对于云计算中心来说，边缘计算场景要求AI芯片在芯片体积、功耗、成本方面做到更经济，由于边缘计算场景主要执行推断任务，芯片算力及计算精度相对于云计算中心可适量下调。由于在功耗、面积、成本方面的限制，AI芯片企业往往需要将AI芯片作为协处理器内置于SoC中，因此对于企业的SoC系统级产品开发能力提出了较强的要求。此外，对于复杂应用场景，如自动驾驶等，芯片企业应为客户提供硬件+软件开发环境的全套解决方案。

边缘计算场景AI芯片设计思路及发展趋势



来源：艾瑞研究院绘制。

AI芯片行业产业链

半导体行业产业链长，具有资本和技术壁垒双高的行业特点

半导体行业产业链从上游到下游大体可分为：设计软件（EDA）、设备、材料（晶圆及耗材）、IC设计、代工、封装等。Fabless与IDM厂商负责芯片设计工作，其中IDM厂商是指集成了设计、制造、封装、销售等全流程的厂商，一般是一些科技巨头公司，Fabless厂商相比IDM规模更小，一般只负责芯片设计工作。

分工模式（Fabless-Foundry）的出现主要是由于芯片制程工艺的不断发展，工艺研发费用及产线投资升级费用大幅上升导致一般芯片厂商难以覆盖成本，而 Foundry厂商则是统一对Fabless和IDM的委外订单进行流片，形成规模化生产优势，保证盈利的同时不断投资研发新的制程工艺，是摩尔定律的主要推动者。当前在半导体产业链中，我国在上游软件、设备、高端原材料以及代工制造与全球一线厂商差距较大，而在封装环节拥有长电、华天、通富微等行业前十企业，今年来在IC设计领域也逐渐涌现了以海思为代表的一批优秀企业。

AI芯片行业产业链梳理

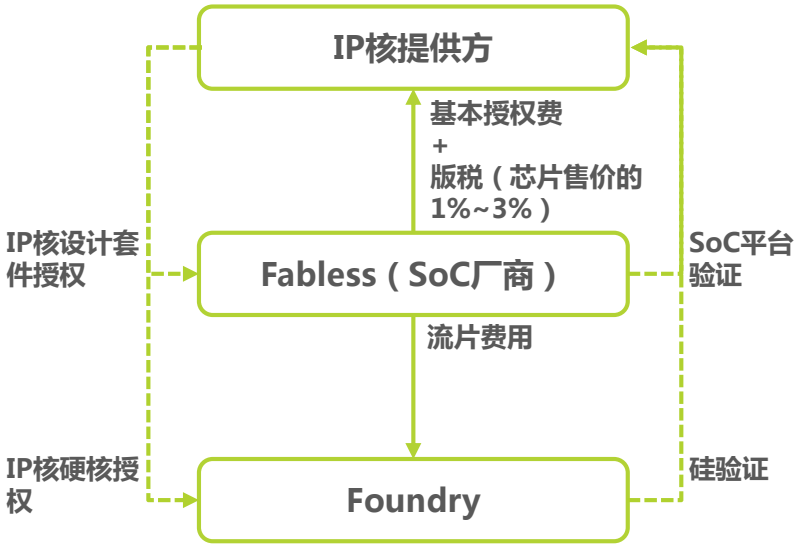
注释： Fabless/Fab-lite指：无晶圆生产模式。
来源：艾瑞研究院。

AI芯片公司商业模式

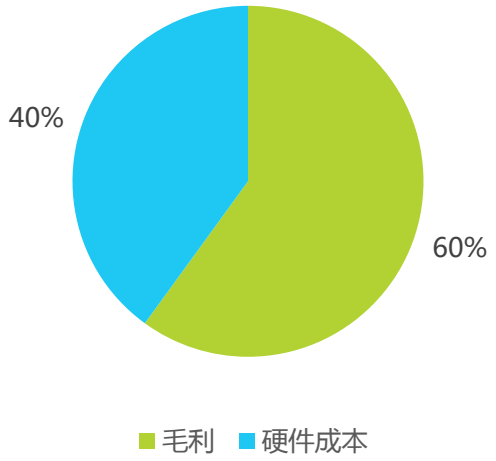
半导体行业商业模式主要可分为：IP授权与流片模式

行业主要存在两种商业模式IP授权和流片模式。其中在IP授权模式中，IP设计公司将自己设计的芯片功能单元，如：CPU、GPU、DSP、NPU等，授权给其他的IC设计公司，如华为海思麒麟970、980芯片获得了寒武纪NPU的IP授权。被授权方将会向授权方支付一笔授权费来获得IP，并在最终芯片产品销售中，以芯片最终售价的1%~3%向授权方支付版税。授权费用实现IP开发成本的覆盖，而版税作为IP设计公司的盈利。但正如手机芯片市场，优质的IP资源往往集中在科技巨头手中，拥有单一或少量IP的创业公司往往因为自身IP竞争力不足、或是难以提供具有综合竞争力的完整解决方案而最终落得被收购或退出市场的境地。流片生产模式虽然前期投入较大，但一款成功的产品将会使公司获得丰厚的利润，一般芯片产品定价采取8:20原则，即硬件成本：最终产品售价=8：20。该比率可能会随厂商对市场话语权不同而上下波动，因此一款成功的芯片销售毛利应在60%以上。但公司是否能够最终实现盈利，还需要在毛利中进一步扣除前期研发费用。

IP授权收费模式



流片模式芯片定价跟随8:20原则



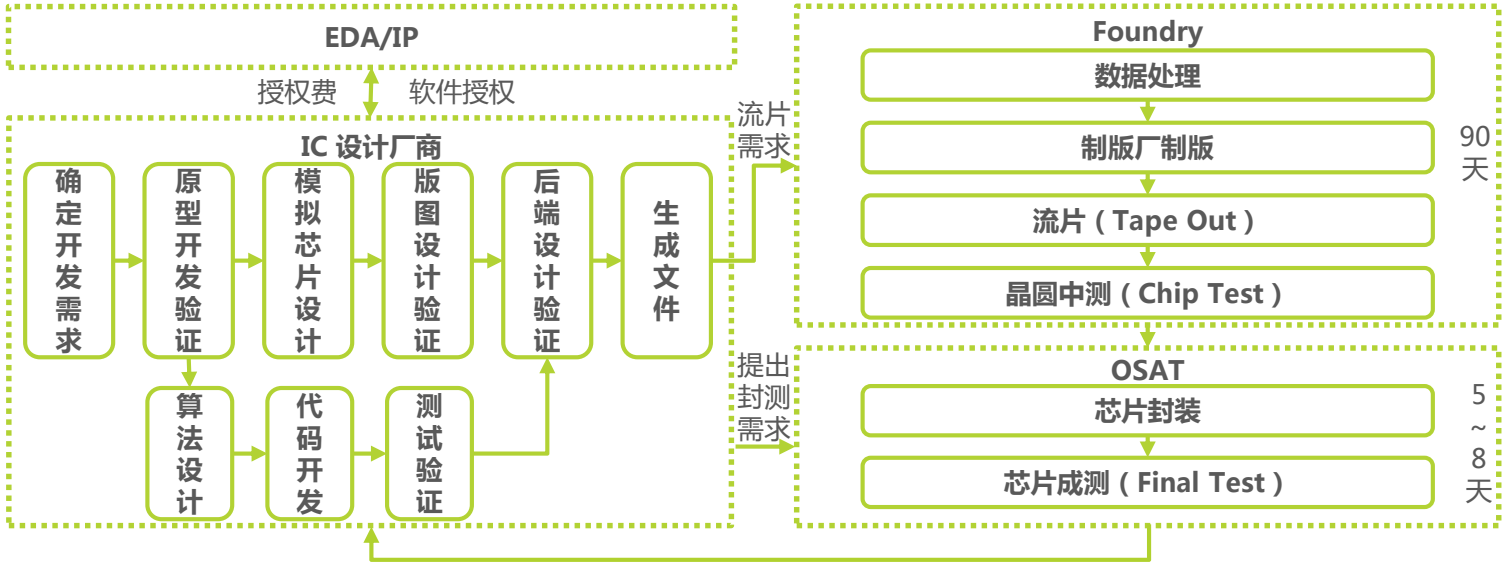
来源：公开网络数据。

AI芯片研发设计投入分析

芯片设计需要厂商承担昂贵的EDA费用及高昂的人力成本

芯片整体设计制造流程大体包括：1）IC设计公司进行芯片架构设计，2）将设计完成的芯片“图纸”文件交由Foundry厂商进行流片，3）裸片将会交由OSAT厂商进行封装，4）产品销售。研发费用主要包括：研发团队人力成本、EDA软件及IP授权费用及其他场地租金、水电费用等。人力成本占研发成本主要部分，项目开发效率与资深工程师数量正相关，国内资深芯片设计工程师年薪一般在50~100万元之间。EDA工具是芯片设计工具，是发展超大型集成电路的基石，EDA工具可有效提升产品良率。目前，该领域被海外厂商高度垄断，CR3大于70%。EDA厂商主要是通过向IC设计公司进行软件授权获取盈利，根据调研，20人的研发团队设计一款芯片所需要的EDA工具采购费用在100万美元/年左右（包括EDA和LPDDR等IP购买成本）。英伟达开发Xavier，动用了2000个工程师，开发费用共计20以美金，Xlinix ACAP动用了1500个工程师，开发费用总共10亿美金。

芯片设计制造流程图



来源：公开网络数据、全志科技招股说明书，艾瑞研究院自主绘制。

AI芯片制造成本投入分析

芯片设计技术积累+市场洞察力=芯片产品市场推广成功与否

在IP授权和流片两大类商业模式中，IP授权由于不涉及芯片制造，仅需要考虑研发费用，资金占用相对小、风险较低。流片除前期的研发投入以外，还需要向代工厂支付巨额的代工费用，对资金占用极大，需要芯片销售达到一定量级才能分摊掉前期巨额投入实现盈利，若期间出现流片失败（即流片未达设计期望性能指标）或者市场推广失利等情况，芯片设计厂商需要承担前期巨额的研发和制造投入、费用损失。芯片单位硬件成本主要包含掩膜、封装、测试和晶圆成本，并受到制程工艺、产量、芯片面积等多因素的影响。我们简要测算16nm制程工艺下，不同产量不同面积的芯片单位成本，可以看出芯片单位硬件成本随芯片面积、产量上升逐渐下降。因此，一款芯片能否获得广大的市场认可，并拥有较长的产品生命周期，实现芯片产品的规模销售和生产显著决定了企业的盈亏情况。

芯片流片单位成本分析（情景分析）

情景分析		假设：12寸晶圆（300mm），4500美元/片；100人研发团队，人均工资50万元/年；研发周期：1.5年；良率：50%								
单位：美元		掩膜（16nm）	封装	测试	晶片成本			芯片硬件成本		
					60mm²	6mm²	3mm²	60mm²	6mm²	3mm²
总成本		4,800,000	大体占总硬件成本的20%							
产量	×10⁵	48.0	6.00	2.00	8.2	0.8	0.4	120.2	112.8	112.4
	×10⁶	4.8	6.00	0.20	8.2	0.8	0.4	30.2	22.8	22.4
	×10⁷	0.5	6.00	0.02	8.2	0.8	0.4	21.2	13.8	13.4
	×10⁸	0.0	6.00	0.00	8.2	0.8	0.4	20.3	12.9	12.5

*注：芯片成本 $= \frac{(\text{掩膜}+\text{封装}+\text{测试})_{\text{单片}}}{\text{成品率}} + \text{晶片成本}$；晶片成本 $= \frac{\text{晶圆成本}}{(\text{每片晶圆晶片数})\times\text{成品率}}$；每片晶圆晶片数 $= \frac{\text{晶圆面积}}{\text{晶片面积}} - \frac{\text{晶圆周长}}{\sqrt{2\times\text{晶片面积}}}$

来源：公开网络数据，艾瑞研究院自行测算。

人工智能产业相关政策梳理

经济转型背景下国内将重点加大对科技领域的政策扶持力度

时间	人工智能产业政策梳理&汇总
2018.12	2018年12月中央经济工作会议召开，会议指出要加快5G商用步伐，加强人工智能、工业互联网、物联网等新型基础设施建设，加大城际交通、物流、市政基础设施等投资力度，补齐农村基础设施和公共服务设施建设短板。
2018.10	习近平在中央政治局第九次学习中提到：加快发展新一代人工智能是赢得全球科技竞争的重要战略抓手，是推动我国科技跨越发展、产业优化升级、生产力整体跃升的重要战略资源。要以关键核心技术为主攻方向，加强基础理论研究，支持科学家勇闯人工智能“无人区”，努力在人工智能发展方向和理论、方法、工具、系统等方面取得突破。
2017.7	**国务院颁发《新一代人工智能发展规划》：**1、构建开放协同的人工智能科技创新体系；2、推进产业智能化升级；3、利用人工智能提升公共安全保障能力；4、加强人工智能领域军民融合；5、构建安全高效的智能化基础设施体系；6、前瞻布局重大科技项目。 1. **2020年**，总体技术和应用与世界先进水平同步，人工智能成为新的经济增长点，**核心产业规模达到1500亿元，带动相关产业规模超过1万亿元**； 2. **2025年**，新一代人工智能技术在智能制造、智能医疗、智慧城市、智能农业、国防建设等领域得到广泛应用，**核心产业规模达到4000亿元、相关产业达到5万亿元**； 3. **2030年，**人工智能理论、技术与应用达到世界领先水平，形成完善高端的产业集群，**核心产业规模达到1万亿元，带动相关产业规模超过10万亿元.**
2017.3	**人工智能首次被写入全国政府工作报告**，李克强指出：“要加快培育壮大新兴产业。全面实施战略性新兴产业发展规划，加快新材料、**人工智能、集成电路**、生物制药、**第五代移动通信等技术研发和转化，做大做强产业集群。**”
2017.2	人工智能首次入选了**《战略性新兴产业重点产品和服务指导目录》**指导目录名单
2016.5	发改委、科技部、工业和信息化部、中央网信办制定了**《“互联网+”人工智能三年行动实施方案》，将支持AI领域芯片、传感器、操作系统、存储系统、高端服务器、关键网络设备、网络安全技术设备、中间件等基础软硬件技术开发，支持开源软硬件平台及生态建设**
2016.3	**人工智能被写进国家“十三五”规划纲要**
2015.7	国务院印发**《关于积极推进“互联网+”行动的指导意见》**中提出，依托互联网平台提供人工智能公共创新服务，加快人工智能核心技术突破并其促进推广应用（智能家居、终端、汽车、机器人等）
2015.6	**《中国制造2025》**首次提及智能制造，推动新一代信息技术与制造技术融合发展，着力发展智能装备和智能产品，推动生产过程智能化。

来源：公开网络数据。

半导体产业相关政策梳理

半导体是发展“安全、可靠、领先”信息科技技术的基石

国内半导体技术发展落后于海外，使我国在发展信息科技产业时，出现上游底层技术严重依赖海外的情况，在当前世界面临去全球化风潮中，不管是从商业还是从国家安全角度考虑，都急需开发自主产品对海外技术和产品实行替代。 国内在CPU领域有海光、兆芯、龙芯等厂商对相关领域进行填补。在布局发展物联网、人工智能等未来产业时，应吸取前车之鉴，提前布局、研发相关领域的通讯、计算芯片等底层技术。

时间	半导体产业政策梳理&汇总
2017.4	科技部《国家高新技术产业开发区“十三五”发展规划》，优化产业结构，推进集成电路及专用装备关键核心技术突破和应用。
2016.12	国务院《 “十三五”国家信息化规划》，大力推进集成电路创新突破，加大面向新型计算、5G、智能制造、工业互联网、物联网的芯片设计研发部书，推动32/28nm、16/14nm工艺生产线建设，加快10/7nm工艺技术研发。
2016.7	国务院《 “十三五”国家科技创新规划》，支持面向集成电路等优势产业领域建设若干科技创新平台；推动我国信息光电子器件技术及集成电路设计达到国际先进水平。
2015.05	国务院《中国制造2025》，将集成电路及专用装备作为“新一代信息技术产业”纳入大力推动突破发展的重点领域；着力提升集成电路设计水平，掌握高密度封装及三维微组装技术，提升封装产业和测试的自主发展能力。形成关键制造装备供货能力。
2014.06	国务院《国家集成电路产业发展推进纲要》，到2015年集成电路产业发展体制机制创新取得明显成效，建立与产业发展规律相适应的融资平台和政策环境，集成电路产业销售规模超过3500亿元；到2020年集成电路水平与国际先进水平的差距逐渐缩小，全行业销售收入年均增速超过20%，企业可持续发展能力大幅增强；基本建成技术先进、安全可靠的集成电路产业体系。
2012.07	国务院《“十二五”国家战略性新兴产业发展规划》，大力提升高性能集成电路产品自主开发能力，突破先进和特色芯片制造工艺技术，先进封装、测试技术以及关键设备、仪器、材料核心技术，加强新一代半导体材料和器件工艺技术研发，培育集成电路产业竞争新优势。
2012.02	工信部《集成电路产业“十二五”发展规划》，到“十二五”末，产业规模再翻一番以上，关键核心技术和产品取得突破性进展，结构调整取得明显成效，产业链进一步完善，形成一批具有国际竞争力的企业，基本建立以企业为主体的产学研用相结合的技术创新体系。
2011.01	国务院《进一步鼓励软件产业和集成电路产业发展的若干政策》，软件产业和集成电路产业是国家战略性新兴产业。是国民经济和社会信息化的重要基础，分别从财税政策、投融资政策、研究开发政策、进出口政策、人才政策、知识产权政策、市场政策七个方面鼓励软件和集成电路发展。

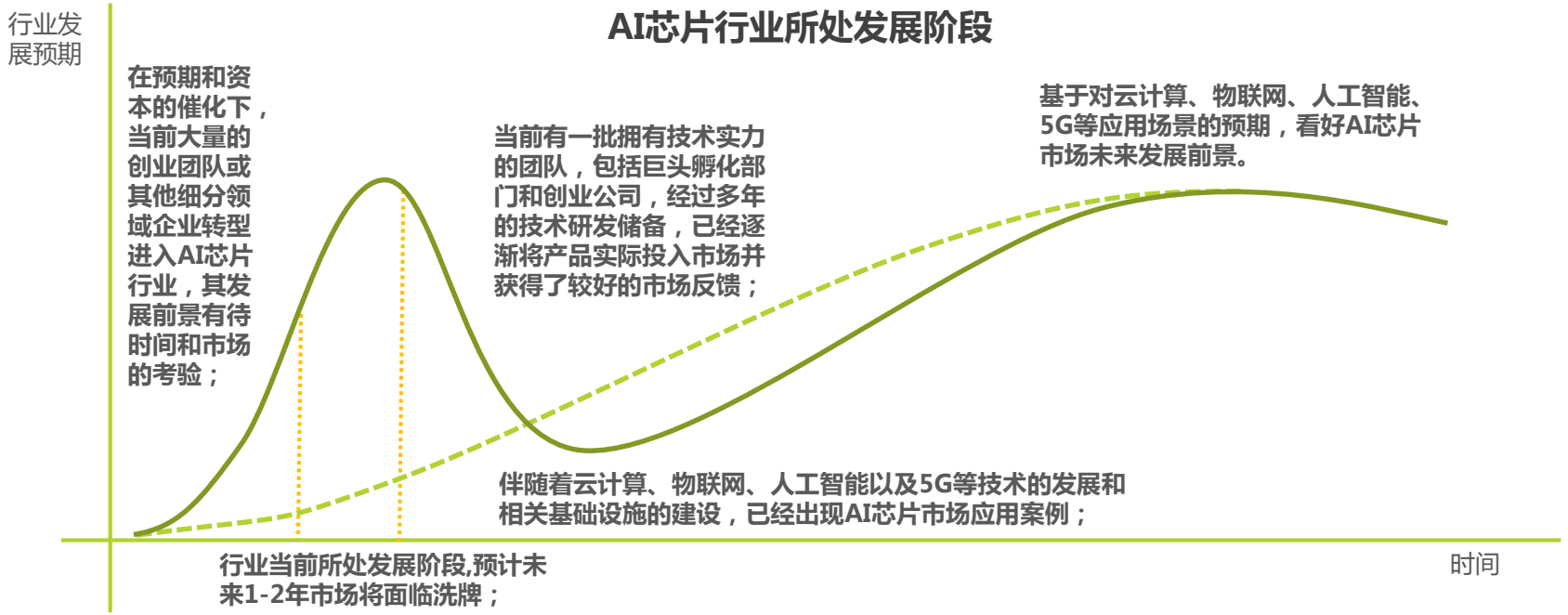
来源：公开网络数据。

AI芯片行业发展阶段

行业存在泡沫，期待有技术实力和业绩的企业的发展前景

国内在CPU等高精尖半导体领域发展长期落后的背景下，在近两年集中出现了大量的AI芯片行业企业，这一方面说明：资本和行业从业人员对AI芯片未来应用前景的认可，另一方面也说明AI芯片的技术门槛并没有CPU高，或者可以说低端AI芯片产品技术门槛并不高。芯片研发、制造成本高昂，对资金需求极大，预计未来1~2年，随着各厂商首批AI芯片产品的面市，市场将会对各厂商的产品和技术进行实际检验，技术不足、产品缺乏竞争力的团队在缺乏后续订单和盈利支撑的情况下将会陆续退出市场，存活下来的企业将会是技术和产品领先、获得市场认可的优秀企业和团队。

当前已经有一批企业在产品研发和市场推广上作出了一定的成绩，其中包括海外和国内的科技巨头和创业公司，如：英伟达、华为海思、寒武纪、比特大陆等，其产品在云端、自动驾驶、智慧安防、移动互联网等场景中获得了较好地应用。



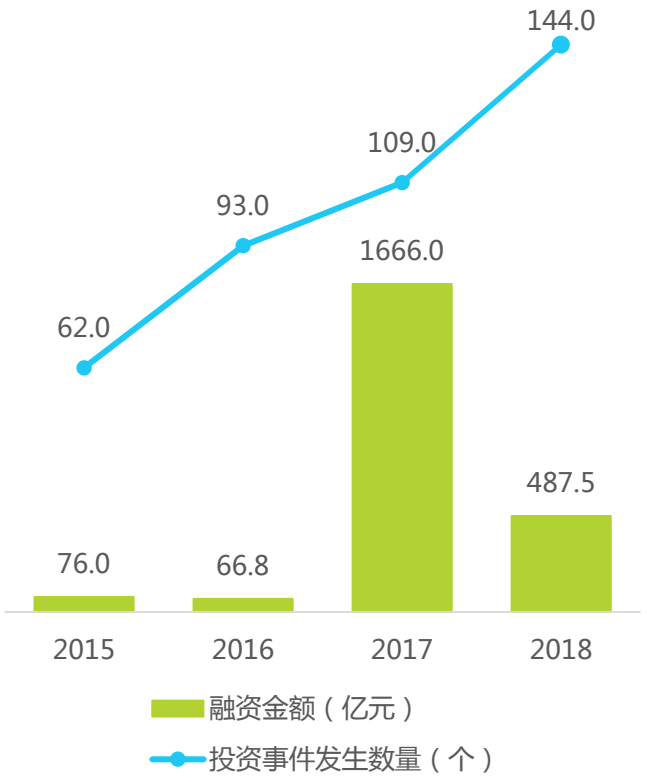
来源：艾瑞研究院自主绘制。

AI芯片行业资本市场热度

AI芯片行业市场预期逐渐趋于理性，创业进入市场检验期

大量AI芯片公司在15~17年成立，18年新增企业数量减少。资本方面，受到宏观经济影响虽然行业内投融资事件相比17年同比增长了32%，但行业整体投融资金额骤减，但头部企业在18年依然持续获得投资人青睐，多家企业创造了估值新高。

2015-2018年AI芯片行业股权投融资事件及金额



目前国内估值排名靠前的AI芯片公司（不考虑上市公司）

公司	成立时间	融资历史及公司估值
比特大陆	2013年10月	18年8月获得4.4亿美元B+轮股权投资，**投后估值甚至达到146.66亿美元。**
地平线	2015年7月	19年1月获得6亿美元的融资，**估值达到30亿美元。**
寒武纪	2016年3月	18年6月获得联想创投、中金资本、阿里巴巴等在数亿美元投资，**估值达到25亿美元。**
肇观电子	2016年5月	18年10月，公司获得中电海康2亿元A轮股权投资，根据注册资本变动推测对应**估值为5.8亿美元**
云知声	2012年6月	18年7月，公司获得6亿元C+轮股权投资，**估值达到4.57亿美元。** 19年4月，公司获得中金、东方证券、清泉资本的D轮投资，对应估值未披露。
思必驰	2007年10月	18年6月，公司获得元禾控股、中民投、深创投、富士康、联发科的5亿元投资，根据注册资本变动比例推测对应**估值达到5亿美元。**
深鉴科技	2016年3月	以**3亿美元**被赛灵思（Xlinix）全资收购。
燧原科技	2018年3月	18年8月，公司获得腾讯、亦合资本、真格基金等3.4亿元Pre-A轮股权投资，**估值达到1.75亿美元。**
熠知电子	2017年2月	17年12月，公司获得红杉资本、云锋基金、高瓴资本、依图科技4.5亿元的A轮投资，对应**估值达到1.55亿美元。**
得一微	2017年11月	18年4月，公司获得TCL资本、清控银杏、华登国际等3亿元A轮股权投资，对应估值推测达到1.05亿美元

注释：估值确认基于市场公开披露融资金额及注册资本变更比例倒推求得，排名仅供参考。
来源：企查查；IT桔子，互联网公开数据。

行业投资布局

关注具有软硬件开发、生态构建、充足资金及产品落地的企业

基于前文对于技术、市场、宏观环境等因素的分析，我们认为提前布局AI芯片领域在未来可能会使投资人享受到较好的投资收益，原因可以总结为：物联网、AI、5G等技术发展塑造了全新的、巨大的应用场景和市场空间，以及传统计算芯片在面临新场景时难以满足其计算需求，这两者共同催生了对于AI芯片的需求及前景可观的市场想象空间。此外，在加大自主替代和科技领域投资的政策背景下，一批拥有优秀创业团队和技术实力的AI芯片公司将会享有较好的市场政策和发展前景。

对于AI芯片行业发展的判断及AI芯片公司评定标准

行业发展判断：

- **AI芯片并行运算处理能力**
  由于AI芯片在面对AI应用时在算力、功耗、成本等方面对传统CPU芯片实现全面赶超，看好其未来的市场应用
- **广阔的市场空间**
  以云计算、智慧安防、智能手机、自动驾驶为代表的IoT+AI应用场景将会产生超百亿美金、10倍增长空间的AI芯片市场
- **自主替代**
  基于国内对半导体技术发展缺失的前车之鉴，在面向IOT、5G、AI等未来广大的应用场景时，应提前布局发展先进、自主可靠地底层硬件技术产品
- **经济转型**
  在经济转型、及外部技术封锁的背景下，国内将会加大提升对科技领域的政策、技术和资金投入和扶持力度

AI芯片公司应具备：

- **充分的芯片开发技术积累**
  优秀的芯片设计团队，完整的ASIC设计能力，AI芯片应具有通用性，能够实现对多种算法开发框架模型的支持；
- **软件开发能力**
  具有优秀的软件开发团队，能够开发配套软件开发包、开发平台，实现对多种开发语言的支持；
- **产业链合作关系**
  具有顺畅的产业链上下游合作关系，主要实现：1、降低开发成本，2、AI芯片产品对于现有硬件生态的适配；
- **市场洞察力**
  具有发现下游应用场景需求的洞察力，能够理解并开发出符合场景需求、生命周期长的AI芯片；
- **构建开发生态**
  对于开发云端芯片厂商来说，能够与云厂商、服务器厂商政府高校立良好的产业合作关系，共同拓展开发生态；
- **政策+资金**
  公司具有较好的产业合作、股东背景，公司能够充分享受政府、产业扶持政策，具有强大的融资能力及资金储备。

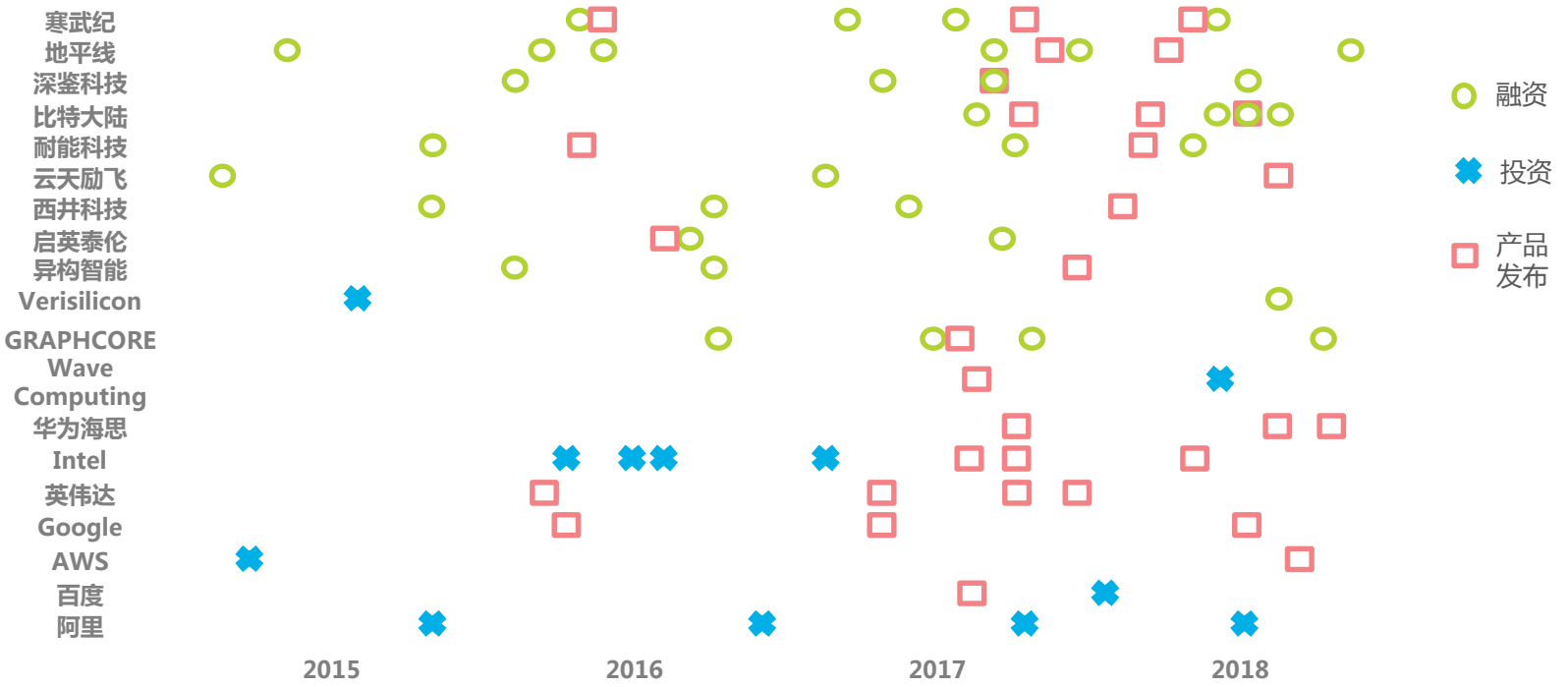
来源：艾瑞研究院自主绘制。

AI芯片行业企业

AI芯片行业市场预期逐渐趋于理性，创业进入市场检验期

从12年开始，英伟达将其GPU产品应用于AI并行运算应用中，人们意识到了AI芯片的巨大潜力，传统半导体行业巨头、科技巨头和众多创业团队纷纷加入到该领域的产品研发中来。国内创业公司多成立于15年以后，从17年开始大量的AI计算芯片产品陆续发布，产品逐步开始实现落地。传统的半导体巨头和科技巨头也在布局AI芯片领域，除自主研发以外，基于资金优势通过对外投资收购优质资产及创业团队等手段加速自身的AI芯片业务发展，典型代表如Intel，大手笔收购了包括Altera、Nervana、Movidius以及Mobileye在内的多家AI芯片企业，阿里巴巴也通过先后投资、收购布局AI芯片的开发。

2015-2019Q1全球部分典型AI芯片企业投融资及AI芯片产品发布事件汇总



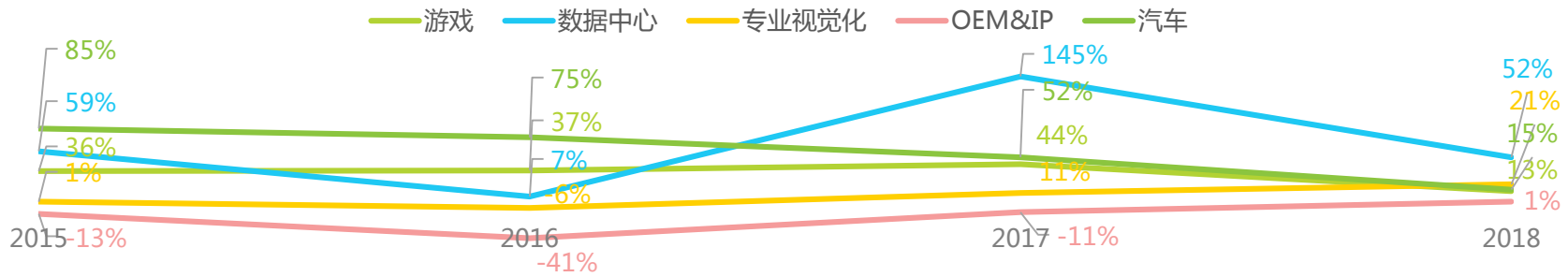
来源：公开网络数据。

典型案例：英伟达

公司业务概述：人工智能芯片第一股

英伟达成立于1993年，最初主要专注于开发图形处理专用处理器，1999年推出了第一款专门用于图像处理的GPU，目前公司在独立显卡领域市场份额超过80%。近10年，公司充分发挥了GPU在并行计算领域的能力，将其应用领域拓展到VR、HPC、AI等热点应用领域，尤其是AI芯片领域，公司计算产品已经拓展至从云到端的各个应用场景，其在AI领域的优秀表现使公司市值在16、17、18H1实现了超10倍的增长。公司业务按市场可划分为：游戏业务、数据中心、专业视觉、汽车和OEM&IP业务,除了传统的游戏业务外，增速最快的业务板块为与AI相关的数据中心和汽车业务。

2015-2018年英伟达各业务板块业绩增速



2015-2018年英伟达二级市场股价（市值）变化情况

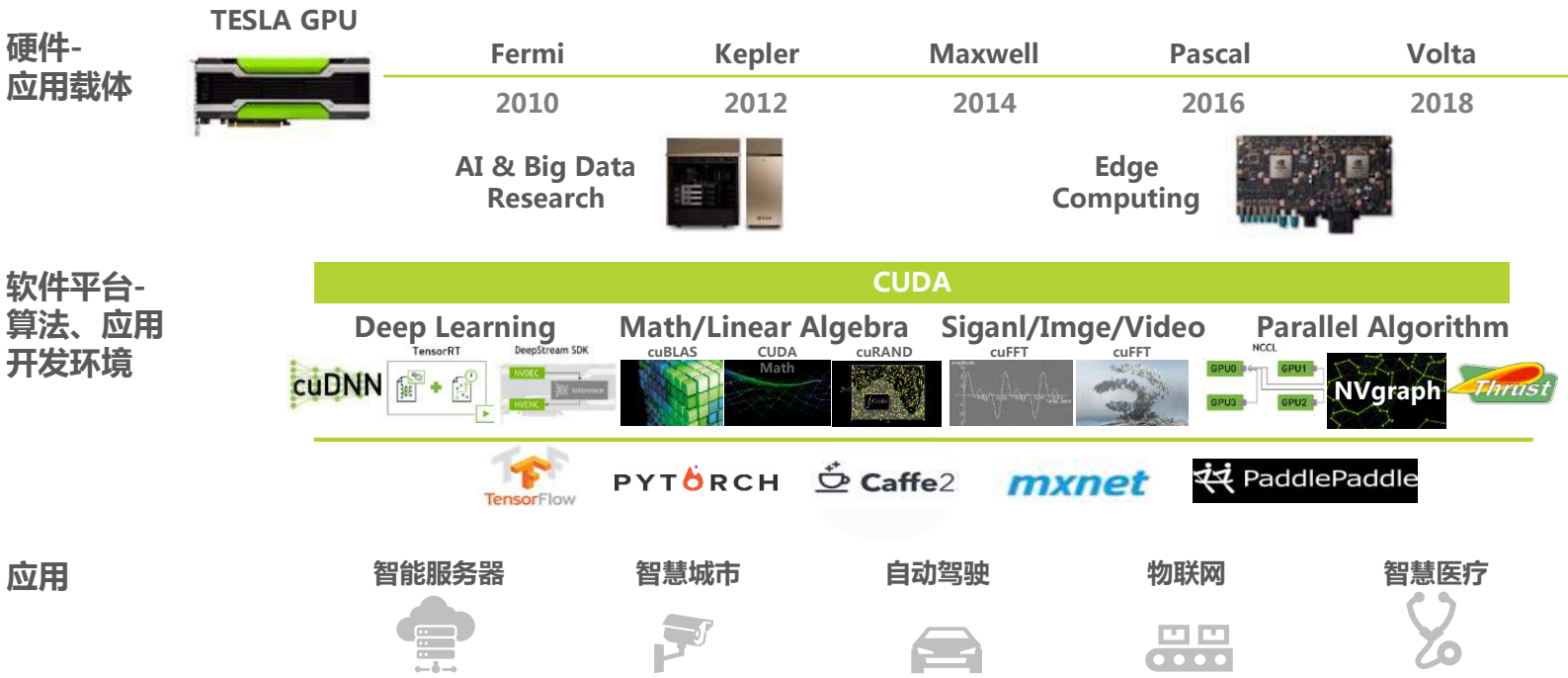
注释：公司在18H2出现的估价下跌主要是由于数字加密货币暴跌导致的矿机需求骤降，直接影响了公司GPU出货，从而影响股价。
来源：Wind。

英伟达AI芯片业务

英伟达为不同细分领域AI开发者提供通用平台级解决方案

英伟达平台发展战略，为各个应用场景的客户提供硬件、软件系统、编程语言、软件开发包以及配套服务等完整解决方案。公司专为专为数据中心、AI、大数据等应用领域开发了Tesla系列GPU，芯片架构每两年实现一次更新，当前最新的产品是Volta架构，代表产品是Tesla V100，单精度浮点运算可达15.7TFLOPS。英伟达还搭建了CUDA开发生态，拥有丰富的开发库和开发套件，对主流编程语言和编程框架形成统一支持，目前已经发展到第10代，在全球拥有超过50万开发者使用。

英伟达AI 产品线及产品应用场景



来源：英伟达官网。

典型案例：寒武纪

全球最早提出神经网络专用芯片架构及通用指令集设计理论

公司创始团队来自与中科院计算所，是全球最早关注并研究AI芯片的团队之一。提出“DianNao”系列微架构实现“小芯片” 解决大规模机器学习算法问题，“DianNao”系列可以在损失一小部分计算精度（半精度计算）下更高效（降低芯片面积和功耗）的完成神经网络计算任务。“PuDianNao”实现了对多种深度学习算法的兼容，“DianNaoYu”（ Cambricon ）提出了神经网络计算芯片指令集架构（ISA），实现对十种代表性神经网络（NN）的兼容，提升了神经网络芯片指令效率，奠定了设计通用型神经网络计算芯片（ASIC）的基础。Google TPU主架构师曾与寒武纪共同开发DianNao架构，其TPU论文大量引用“DianNao”系列论文成果。

DianNao（2014）

ASPLOS2014最佳论文奖（亚洲第一次）

原型神经网络处理器结构，针对CNN\DNN\RNN神经网络计算设计，相比起传统的硬件加速器方案拥有更灵活的可拓展性。其包含一个处理器核，主频为0.98GHz，运算峰值达到0.45TOPs，65nm工艺下能耗仅为0.485W。性能超过主流CPU100倍，但是面积和功耗仅为其1/10，平均性能与GPGPU相当，但功耗仅为其百分之一量级。

DaDianNao（2014）

MICR2O14最佳论文奖（美国以外国家第一次）

是DianNao的多核并行架构版本，揭示神经网络的可分特性使加速器具备可扩展性。包含16个NFU核和更大的片上存储，并支持多芯片间直接高速互连，避免高昂的内存访问开销。主频为606MHz，28nm工艺下功耗约16W。性能超过主流GPU21倍，而功耗仅为其1/330。该架构可实现对深度学习training过程的支持。

ShiDianNao（2015）

ISCA2015

由于DRAM的读写会有相当大的功耗并产生延迟。“ShiDianNao”提出通过加速器与传感器直连而绕过内存，从而降低芯片运算对内存存访的依赖，而CNN算法共享权值存储于SRAM中，避免了对于DRAM的使用。

PuDianNao（2015）

ASPLOS2015

实现了包括k-最近邻、k-均值、朴素贝叶斯、线性回归等7种机器学习算法的兼容。主频为1GHz，峰值性能达每秒1.06TOPs，面积3.51mm^2，功耗为0.596W（65nm工艺下）。PuDianNao运行上述机器学习算法时的平均性能与主流GPGPU相当，但面积和功耗仅为主流GPGPU百分之一量级。

DianNaoYu（2016）

ISCA2016（评分排名位列第一）

全球首个神经网络通用指令集架构，兼容十种代表性的神经网络。针对大规模的神经元计算，单条指令即可完成一次向量或矩阵运算。Cambricon架构下的代码长度分别比GPU\X86\MIPS短6.41、9.86、13.38倍，性能是X86和GPU的91.72倍和3.09倍，而GPU的功耗是Cambricon的130.53倍。

来源：寒武纪“DianNao”系列论文。

寒武纪AI芯片业务

从AI芯片到指令集和软件开发包，构建寒完整开发生态

公司基于其科研成果，开发了从云到端的AI芯片产品，产品主要应用于AI推理环节。公司还开发了软件开发包（SDK）和专用的指令集（ISA），能够对多种经典的深度学习开发框架进行支持，开发者可以使用公司提供的开发环境对寒武纪云端和终端产品进行统一编译。通用指令集及软件开发站为客户提供了友好的开发环境，构建了从硬件到软件的完善生态体系。目前公司产品已经被应用于多个下游应用场景，其中最成功的应用案例是公司Cambricon-1A被应用于华为麒麟970SoC中，而华为海思继续采用了公司Cambricon-1H用于麒麟980CoC。此外公司就其MLU系列云端芯片已经与多家服务器厂商达成了共同开发合作。

寒武纪智能软硬件产品生态

DianNao系列AI芯片

Cloud：MLU100/270

Edge：Cambricon-1A/1H/1M

Cambricon（DianNaoYu）ISA

Cambricon NeuWare SDK

高性能计算库：CN-ML

模型重训练和转换工具：

通用编译器：CN-CC

性能监控及调优：CN-Perf

调试工具：CN-GDB

Deep Learning Framework Supporting

Caffe　Caffe2　mxnet　ONNX

Deployment

HUAWEI　海高　科大讯飞 iFLYTEK　Baidu 百度　中科曙光 Sugon　H3C　星环智能

TAL 好未来　星宸科技 Stardawn　阿里巴巴 Alibaba.com.cn　lenovo 联想　inspur 浪潮

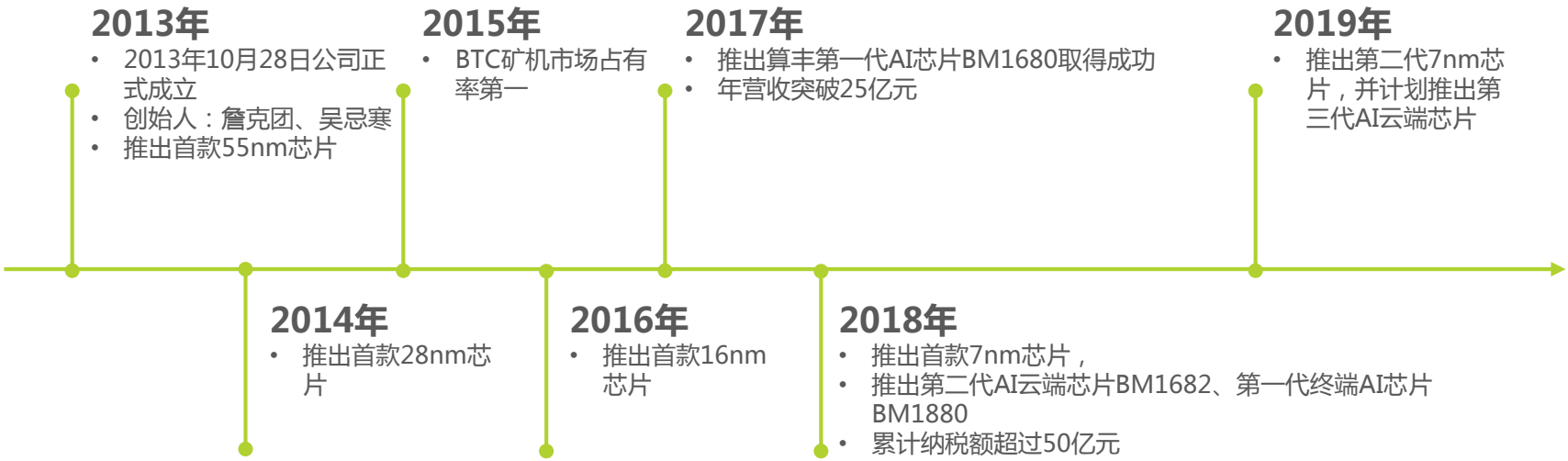
来源：艾瑞研究院。

典型案例：比特大陆

区块链矿机王者进军AI领域

比特大陆是全球领先的算力芯片设计企业，其致力于开发高性能、低功耗、全定制的算力芯片，是全球少数几家掌握最先进7nm制程设计能力并可规模量产7nm芯片的公司之一。目前，比特大陆的产品主要应用于区块链和人工智能两个领域，区块链矿机的市场份额高达74.5%。2017年，比特大陆正式发布了面向人工智能领域的子品牌——“算丰”，并推出了针对深度学习推理的第一代云端AI芯片BM1680。2018年3月，比特大陆快速推出了第二代云端人工智能芯片BM1682，2018年9月份推出了面向终端的AI芯片产品BM1880，并计划于2019年推出其第三代云端AI芯片BM1684。此外，比特大陆基于其芯片，在云端还研发了加速卡、服务器等产品，在终端推出了计算棒、模组、开发板等产品，为不同行业的客户提供适应多种应用场景的产品。

公司发展历史沿革



来源：比特大陆-招股说明书，企查查。

比特大陆AI芯片业务

公司已经开发了可部署于云端和终端的全套软硬件解决方案

芯片

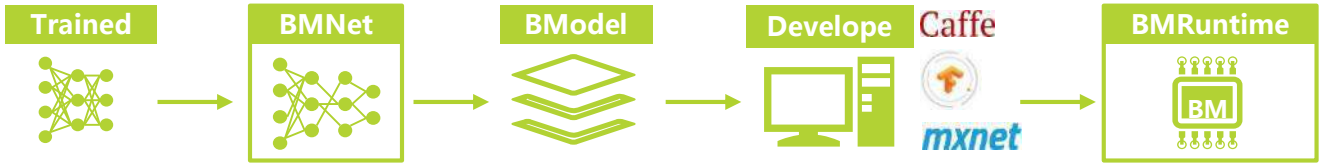
BM1680
- 峰值性能：2TFlops
- 片内存储：32MBytes
- 平均功耗：25W
- 评价：比特大陆第一代AI芯片

BM1682
- 峰值性能：3TFlops
- 片内存储：16MBytes
- 平均功耗：<30W
- 评价：物理算力相比BM1680提升50%，集成视频解码单元产品定位更加专注于图像/视频处理领域

BM1880
- 峰值性能：1TOPS@INT8
- 片内存储：2MBytes
- 平均功耗：2.5W
- 评价：针对于终端应用场景的低功耗、小尺寸封装推理芯片

开发套件

BMNNSDK：
- **BMNet：**对深度神经网络模型进行优化、并转换为比特大陆芯片硬件所支持的bmodel；
- **BMRuntime：**驱动芯片，为上层应用程序提供统一的可编程接口，使程序通过bmodel进行推理。



整体解决产品

BM1682

SA3智能服务器	SE3 迷你机	SC3 加速卡
基于BM1682芯片，面向智能分析领域定制的智能服务器，为安防、视频等行业用户提供更广泛的视频智能应用，具备高性能、低功耗特点	基于BM1682芯片，开发了嵌入式AI迷你机，提供高性价比的硬件设备和人脸识别算法，帮助客户实现定制化AI能力，助力客户提升行业竞争力。	深度学习加速卡产品，标准PCIE3.0接口，采用无风扇设计，适配各种X86服务器，可实现对多种CNN/RNN/DNN神经网络模型的计算加速。

BM1880

人工智能算力棒	人工智能模块	边缘计算开发板
基于BM1880，通过USB，为各种边缘计算场景赋能AI芯片算力	基于BM1880，通过USB接口，可应用于各种人工智能前端产品	集成单颗BM1880、ARM CORTEX A53等，可实现各种强大边缘计算应用的开发

来源：比特大陆公司官网。

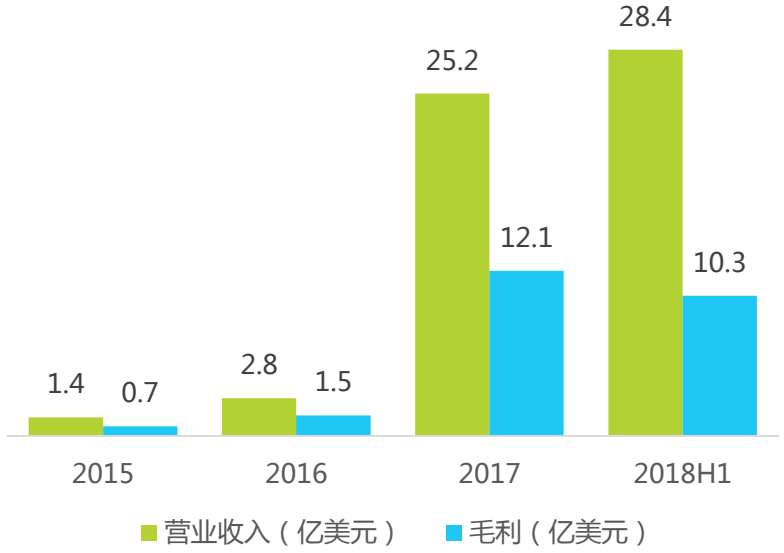
比特大陆AI芯片商业化应用

公司快速实现AI产品商业化变现，商业前景可期

在项目落地方面，比特大陆基于云端AI芯片的人脸闸机助力福建618展会与厦门98投洽会成功举办，累计通行超过30万人次，通道表现稳定可靠，状况良好。在第二届数字中国峰会安保系统中，搭载比特大陆自研芯片的算丰人工智能服务器与海康威视人脸识别算法相融合，全程应用于峰会安全保障工作，3D人脸轨迹系统为日均6万余人次、累计150余万张人脸图片的海量分析提供算力支持。在合作方面，比特大陆与福建当地企业合资成立福建省算域大数据科技有限公司，负责福州城市大脑的投资、建设与运营，为日后福州AI产业发展建设好基础设施。比特大陆还作为首批企业加入海淀城市大脑科技产业联盟，助力海淀“城市大脑”建设，后还与海淀区签署了围绕“智能处理芯片应用场景建设”的重大项目合作意向书，推动算力芯片应用落地。此外，公司还与东亚最大的游戏云平台优必达（ Ubitus ）合作，共同建设公司位于日本、台湾的机房，基于“算丰”芯片，公司协助Ubitus共同开发计算机视觉相关的AI功能。

公司商务合作伙伴

2015-2018H1公司营收情况



来源：比特大陆公司官网。

来源：Wind。

关于艾瑞

在艾瑞　我们相信数据的力量，专注驱动大数据洞察为企业赋能。

在艾瑞　我们提供专业的数据、信息和咨询服务，让您更容易、更快捷的洞察市场、预见未来。

在艾瑞　我们重视人才培养，Keep Learning，坚信只有专业的团队，才能更好的为您服务。

在艾瑞　我们专注创新和变革，打破行业边界，探索更多可能。

在艾瑞　我们秉承汇聚智慧、成就价值理念为您赋能。

- 我们是艾瑞，我们致敬匠心　始终坚信“工匠精神，持之以恒”，致力于成为您专属的商业决策智囊。

扫描二维码
读懂全行业

海量的数据　专业的报告

400-026-2099　ask@iresearch.com.cn

法律声明

版权声明

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，仅供参考。本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。

为商业决策赋能

EMPOWER BUSINESS DECISIONS

艾　瑞　咨　询