

# 华西电子团队一走进“芯”时代系列深度之四十九“AI芯片”

## AI领强算力时代，GPU启新场景落地

孙远峰/熊军/王海维/王臣复/刘奕司

SAC NO:S1120519080005

SAC NO:S1120519120001

2021年11月28日

## 核心观点

### 1、在人工智能时代，终端AI芯片针对特定场景实现优化方案

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。由于终端场景多种多样各不相同，对于算力和能耗等性能需求也有大有小，应用于终端芯片需要针对特殊场景进行针对性设计以实现最优解方案，最终实现有时间关联度的三维处理能力，这将实现更深层次的产业链升级，是设计、制造、封测和设备材料，以及软件环境的全产业链协同升级过程。

### 2、GPU服务器渗透率提升，自动驾驶等级提升推动GPU算力需求

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。人工智能服务器是AI算力基础设施的主要角色，在服务器中渗透率不断提升。L3自动驾驶算力需求为30-60TOPS，L4需求100TOPS以上，L5需求甚至达1,000TOPS，GPU算力需求提升明显，芯片主要向着大算力、低功耗和高制程三个方向发展。

### 3、软硬件生态构筑Nvidia核心优势，国内AI企业加速发展

2006年Nvidia推出CUDA计算平台，让GPU支持CUDA，在优化硬件性能的同时，赋予良好的兼容性，构筑通用计算平台系统和核心优势。国内AI企业产品如景嘉微J9系列，寒武纪思元370、燧原邃思2.0、地平线征途5、黑芝麻华山二号A1000由低算力逐渐向高算力发展，由终端侧往云端和车载方向发展，加快软硬件和开发者生态系统建设。

#### 核心标的：

【重点推荐】：终端侧AI芯片：瑞芯微、全志科技、晶晨股份、富瀚微；

【受益标的】：景嘉微（国产图显GPU）、寒武纪（云边端车GPU）、云天励飞（神经网络处理器芯片），芯原股份（GPU IP供应商）；

【产业链重点标的】：沐曦、壁仞科技、燧原科技、地平线、黑芝麻；【海外标的】：NVIDIA、AMD、CAN.O；

风险提示：产能不及预期风险；国产替代不及预期；国产GPU生态不及预期



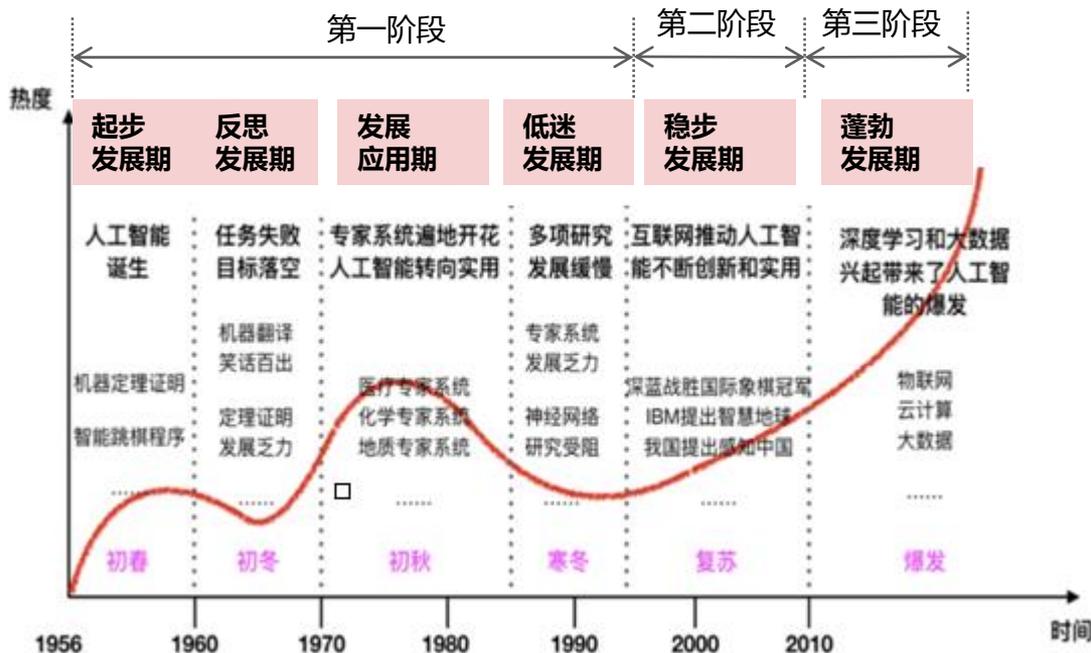
## 目录

- 1 算力时代，GPU开拓新场景
- 2 GPU 下游三大应用市场
- 3 海外GPU巨头Nvidia
- 4 国产GPU赛道掀起投资热潮
- 5 重点投资机会
- 6 风险提示

# 1.1 人工智能经历过三阶段，迎来爆发式增长

- ✓ 广义上讲只要能够运行人工智能算法的芯片都叫作 **AI 芯片**。但是通常意义上的 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。
- ✓ AI芯片也被称为AI加速器或计算卡，即专门用于处理人工智能应用中的大量计算任务的模块（其他非计算任务仍由CPU负责）。

图：AI芯片算力发展阶段



**第一阶段：** 因为芯片算力不足，所以神经网络没有受到重视

**第二阶段：** 通用芯片CPU的算力大幅提升，但仍然无法满足神经网络的需求

**第三阶段：** GPU和和新架构的AI芯片推进人工智能落地

事件1：2014年李天石博士“DianNao”系列论文让科学界看到，在冯诺依曼架构下也可以实现AI专用芯片

事件2：Google 推出的TPU运算架构的 AlphaGo，接连打败李世石和柯洁，看到了专用芯片的商业价值。

## 1.2 推荐式系统模型参数复杂度大幅增加

- ✓ GPT-3模型目前已入选了《麻省理工科技评论》2021年“十大突破性技术”。GPT-3的模型使用的最大数据集在处理前容量达到了45TB。根据OpenAI的算力统计单位petaflops/s-days，训练AlphaGoZero需要1800-2000pfs-day，而GPT-3用了3640pfs-day。

表 递归卷积神经网络-视觉模式/图像识别

### 递归卷积神经网络-视觉模式/图像识别

LeNet-5	6 万个参数 ( 1998 )
AlexNet	6000 万个参数(2012 )
VGG-16	1.38亿参数(2014)
ResNet-50	25000 万个参数 ( 2015)

表 自然语言模型/会话式AI平台

自然语言模型/会话式AI平台	
Open AI GPT-1	1.5亿个参数 ( 2018 )
Google BERT-Large	3.4亿个参数 ( 2018 )
Microsof MT-DNN	3.3亿个参数 (2019)
Open AI GPT-2	15.42亿个参数 ( 2019)
Alibaba PERSEUS-BERT	1.1亿个参数(2019)
NVIDIA Pmojert Megatron	83亿个参数(2019)
Facebook RoBERTa	3.35亿个参数(2019)
Facbook XILM	6.65亿个参数(2019)
NVIDIA Megaltron-Scaled Version of OpenAI GPT-2	83亿个参数(2019)
Microsoft-NLG	172亿个参数 ( 2020 )
Open A GPT-3	多达1750亿个参数(2020)

## 1.3 深度学习模型复杂度对芯片算力需求激增

- ✓ 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**）。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。
- ✓ 自2012年以来，人工智能训练任务所需求的算力每 3.43 个月就会翻倍，大大超越了芯片产业长期存在的摩尔定律（每 18 个月芯片的性能翻一倍）。
- ✓ 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、体积及价格等需求。

从AlexNet到GPT-3，算力增长迅速



资料来源：Nvidia官网，华西证券研究所

## 1.4 部署位置对AI芯片性能要求差异

- ✓ 根据机器学习算法步骤，可分为训练（training）芯片和推断（inference）芯片。训练芯片主要是指通过大量的数据输入，构建复杂的深度神经网络模型的一种AI芯片，运算能力较强。推断芯片主要是指利用训练出来的模型加载数据，计算“推理”出各种结论的一种AI芯片，侧重考虑单位能耗算力、时延、成本等性能。
- ✓ 从部署的位置来看，AI芯片可分为**云端（服务器端）**、**终端（移动端）**两大类。云端芯片，是指部署在公有云、私有云或混合云上的AI芯片，不仅可用于训练，还可用于推断，算力强劲。终端芯片，是指应用于手机等嵌入式、移动终端等领域的AI芯片，此类芯片一般体积小、耗电低、性能无需特别强大。

表 不同部署位置的AI芯片算力要求

应用场景	芯片需求	典型计算能力	典型功耗	典型应用领域
终端	低功耗、高效能、推理任务为主、成本敏感、硬件产品形态众多	<8TOPS	<5瓦	各类消费类电子、物联网
云端	高性能、高计算密度、兼有推理和训练任务、单价高、硬件产品形态少	>30TOPS	>50瓦	云计算数据中心、企业私有云等
边缘端	功耗、性能、尺寸的要求常介于终端与云端之间、推理任务为主、多用于嵌入式设备、硬件产品形态相对较少	5TOPS至30TOPS	4瓦至15瓦	智能制造、智能家居、智能零售、智慧交通、智慧金融、智慧医疗、智能驾驶等众多应用领域

不同部署位置的AI芯片比较



## 1.5 GPU 是较为成熟的通用型人工智能芯片

- ✓ 从技术架构来看，AI芯片主要分为图形处理器（GPU）、现场可编程门阵列（FPGA）、专用集成电路（ASIC）、类脑芯片四大类。其中，GPU是较为成熟的通用型人工智能芯片，FPGA和ASIC则是针对人工智能需求特征的半定制和全定制芯片，类脑芯片颠覆传统冯诺依曼架构，是一种模拟人脑神经元结构的芯片，类脑芯片的发展尚处于起步阶段。

表 三种技术架构AI芯片类型比较

	GPU	FPGA（半定制化）	ASIC（全定制化）
定制化程度	通用性	半定制化	定制化
灵活度	好	好	不好
成本	高	较高	低
编程语言/架构	CUDA、OpenCL等	Verilog/VHDL等硬件描述语言，OpenCL、HLS	/
功耗	大	较大	小
主要优点	峰值计算能力强、产品成熟	平均性能较高、功耗较低、灵活性强	平均性能很强、功耗很低、体积小
主要缺点	效率不高、不可编辑、功耗高	量产单价高、峰值计算能力较低、编程语言难度大	前期投入成本高、不可编辑、研发成本长、技术风险大
主要应用场景	云端训练、云端推断	云端推断、终端推断	云端训练、云端推断、终端推断
代表企业芯片	英伟达Tesla、高通Adreno等	赛灵思Versal、英特尔Arria、百度XPU等	谷歌TPU、寒武纪Cambricon

资料来源：赛迪顾问，华西证券研究所整理

## 1.6 GPU设计之初用于显示图像使用

✓ GPU 图形渲染流水线的具体实现可分为六个阶段，如右图所示。

顶点着色器 (Vertex Shader)

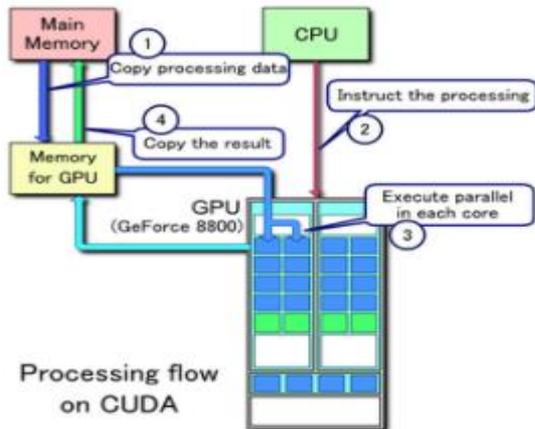
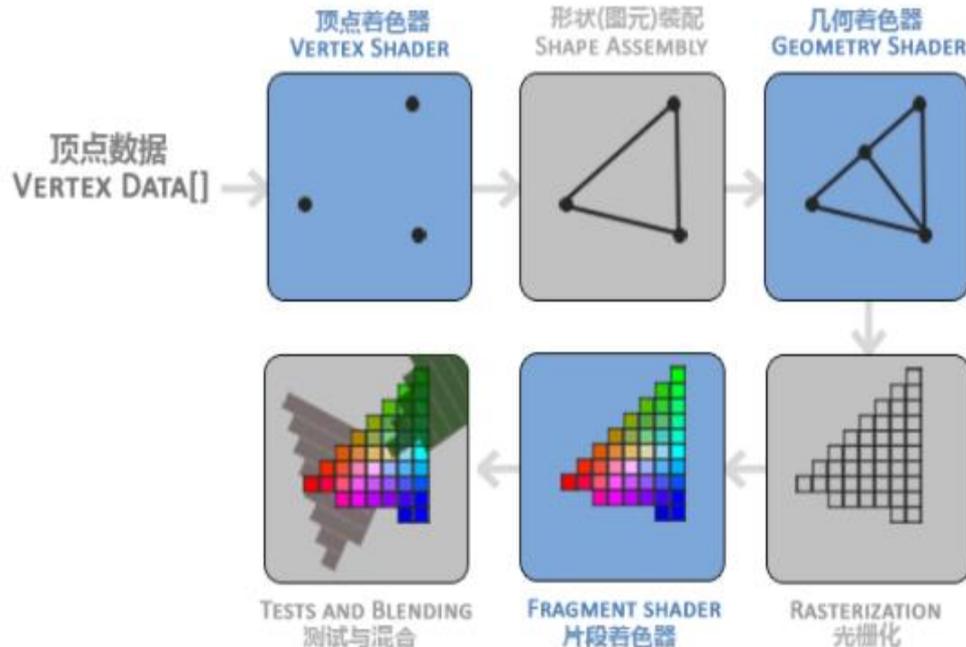
形状装配 (Shape Assembly), 又称 图元装配

几何着色器 (Geometry Shader)

光栅化 (Rasterization)

片段着色器 (Fragment Shader)

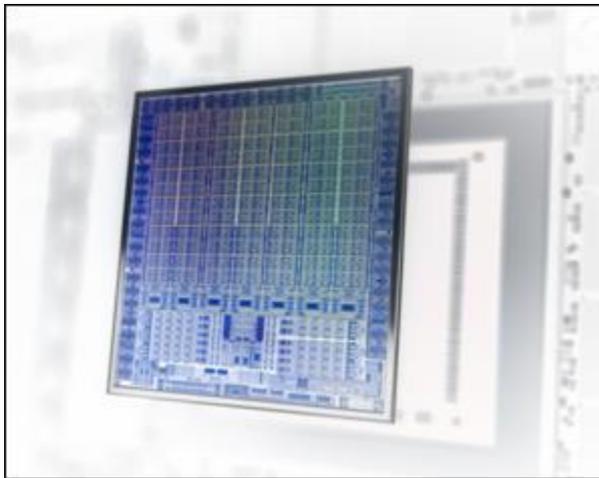
测试与混合 (Tests and Blending)



## 1.7 GPU适合并行运行大量计算

- ✓ GPU (图形处理器) 又称显示核心、显卡、视觉处理器、显示芯片或绘图芯片, 是一种专门在个人电脑、工作站、游戏机和一些移动设备 (如平板电脑、智能手机等) 上运行绘图运算工作的微处理器。
- ✓ GPU使显卡减少对CPU的依赖, 并分担部分原本是由CPU所担当的工作, 尤其是在进行三维绘图运算时, 功效更加明显。图形处理器所采用的核心技术有硬件坐标转换与光源、立体环境材质贴图 and 顶点混合、纹理压缩和凹凸映射贴图、双重纹理四像素256位渲染引擎等。
- ✓ GPU是一种特殊类型的处理器, 具有数百或数千个内核, 经过优化, 可并行运行大量计算。虽然GPU在游戏中以3D渲染而闻名, 但它们对运行分析、深度学习和机器学习算法尤其有用。

英伟达GeForce RTX30系列



AMD Radeon RX6900 XT



## 1.8 GPU vs CPU

- ✓ CPU和GPU 相比，是由于其设计目标的不同，它们分别针对了两种不同的应用场景。CPU需要很强的通用性来处理各种不同的数据类型，同时又要逻辑判断又会引入大量的分支跳转和中断的处理。这些都使得CPU的内部结构异常复杂。而GPU面对的则是类型高度统一的、相互无依赖的大规模数据和不需要被打断的纯净的计算环境。

表 CPU与GPU构架对比

CPU与GPU构架对比



	CPU	GPU
定义与组成	<p>CPU 由数百万个晶体管组成，可以有多个处理内核，通常被称为计算机的大脑。它是所有现代计算机系统必不可少的组成部分，因为它执行计算机和操作系统所需的命令和流程。</p>	<p>GPU 是由许多更小、更专业的内核组成的处理器。在多个内核之间划分并执行一项处理任务时，通过协同工作，这些内核可以提供强大的性能。</p>
微构架	<p>CPU的功能模块多，擅长分支预测等复杂的运算环境，大部分晶体管用在控制电路和Cache上，少部分晶体管用来完成运算工作。</p>	<p>GPU的控制相对简单，且不需要很大的Cache，大部分晶体管可被用于各类专用电路和流水线，GPU的计算速度因此大增，拥有强大的浮点运算能力。</p>
适用领域	<p>CPU 适用于一系列广泛的工作负载，特别是那些对于延迟和单位内核性能要求较高的工作负载。作为强大的执行引擎，CPU 将它数量相对较少的内核集中用于处理单个任务，并快速将其完成。这使它尤其适合用于处理从串行计算到数据库运行等类型的工作。</p>	<p>GPU 最初是作为专门用于加速特定 3D 渲染任务的 ASIC 开发而成的。随着时间的推移，这些功能固定的引擎变得更加可编程化、更加灵活。尽管图形处理和当下视觉效果越来越真实的顶级游戏仍是 GPU 的主要功能，但同时，其他开发人员也开始利用GPU的功能来显着加速高性能计算（HPC），深度学习等领域中的其他工作负载。</p>

## 1.9 AI开源平台众多

- ✓ 广义上AI 开源平台定义为一个深度学习的工具箱，用户可以通过开放平台，基于底层计算芯片运行算法系统。因此AI芯片厂商只有具备支持主流开源平台，才能形成稳定客户群。
- ✓ 两大开源平台Tensorflow与Caffe，Caffe2 和mxnet 也逐步兴起，AI开源平台众多。

表 主流AI 开源平台简介

公司/高校	时间	名称	简介
Stanford	2013.10	Caffe	自2013年底以来第一款主流的工业级深度学习工具包，计算机视觉界最流行的工具包之一
DMLC	2015.09	mxnet	MxNet是一个提供多种API的机器学习框架，主要面向RPython和Julia等语言，目前已被亚马逊云服务采用。
Google	2015.11	Tensorflow	谷歌第二代联机版人工智能深度学习系统，能同时支持多台服务器。
Facebook	2015.12	Torch	深度学习函数库 Torch 的框架，旨在鼓励程序代码再利用及模块化编程。
Microsoft	2016.01	CNTK	CNTK支持RNN和CNN类型的各类网络模型。
Facebook	2017.04	Caffe2	Caffe2 延续了 Caffe 对视觉类问题的支持，且增加了对RNN 和LSTM支持，功能更加完善

## 1.10 未来平台生态之争是AI芯片占领市场关键

- ✓ 广义AI硬件开发环境是专门针对AI硬件推出的适应于硬件计算的开发环境，用户能利用C、C++等软件语言更方便的基于AI芯片进行顶层应用开发，同时可以起到硬件加速的效果。目前比较流行的AI开发环境是Nvidia的CUDA。
- ✓ Xilinx 推出了SDAccel 开发环境，Altera 推出了 OpenCL SDK 开发环境。

表 部分AI硬件平台开源生态支持情况

公司	平台名称	类别	支持的开源生态平台
NVIDIA	Tesla V100	GPU	Caffe, Caffe2, TensorFlow, Torch, CNTK等
Xilinx	reVISION	FPGA	Caffe
Google	TPU2	ASIC	TensorFlow
Qualcomm	NPESDK	ASIC	TensorFlow、Caffe、Caffe2
华为	麒麟970	ASIC	TensorFlow、Caffe、Caffe2

资料来源：Nvidia, Xilinx, 华西证券研究所

## 1.11 中国AI芯片市场规模有望持续快速增长

- ✓ 2019年全球人工智能芯片市场规模为110亿美元。随着人工智能技术日趋成熟，数字化基础设施不断完善，人工智能商业化应用将加速落地，推动AI芯片市场高速增长，预计2025年全球人工智能芯片市场规模将达到726亿美元。
- ✓ 2019年中国AI芯片市场规模约为115.5亿元。5G商用的普及将催生AI芯片在军用、民用等多个领域的应用需求。在政策、市场、技术等合力作用下，中国AI芯片行业将快速发展，在2023年市场规模将突破千亿元。

2019-2025年全球人工智能芯片市场规模及预测（亿美元）



2018-2023年中国人工智能芯片市场规模及预测（亿元）





## 目录

- 1 算力时代，GPU开拓新场景
- 2 GPU 下游三大应用市场
- 3 海外GPU巨头Nvidia
- 4 国产GPU赛道掀起投资热潮
- 5 重点投资机会
- 6 风险提示

## 2.1 GPU微架构

- ✓ GPU其实是由硬件实现的一组图形函数的集合，这些函数主要用于绘制各种图形所需要的运算。这些和像素，光影处理，3D坐标变换等相关的运算由GPU硬件加速来实现。图形运算的特点是大量同类型数据的密集运算——如图形数据的矩阵运算，**GPU的微架构就是面向适合于矩阵类型的数值计算而设计的**，大量重复设计的计算单元，这类计算可以分成众多独立的数值计算——大量数值运算的线程，而且数据之间没有像程序执行的那种逻辑关联性。
- ✓ **GPU微架构的设计研发是非常重要的，先进优秀的微架构对GPU实际性能的提升是至关重要的。**目前市面上有非常丰富GPU微架构，比如Pascal、Volta、Turing（图灵）、Ampere（安培），分别发布于2016年、2017年、2018年和2020年，代表着英伟达 GPU 的最高工艺水平。

Volta构架



灵构架构



## 2.2 GPU架构

- ✓ GPU微架构的运算部份由流处理器(Stream Processor, SP)、纹理单元(Texture mapping unit, TMU)、张量单元(Tensor Core)、光线追踪单元(RT Cores)、光栅化处理单元(ROPs)组成。这些运算单元中,张量单元,光线追踪单元由NVIDIA在伏特/图灵微架构引入。
- ✓ GPU的微架构还包含L0/L1操作缓存、Warp调度器、分配单元(Dispatch Unit)、寄存器堆(register file)、特殊功能单元(Special function unit, SFU)、存取单元、显卡互联单元(NV Link)、PCIe总线接口、L2缓存、二代高位宽显存(HBM2)等接口。
- ✓ 总体布局比较中正,八个GPC与L2 Cache坐落于核心地段,左右为外部存储接口,12道显存控制器负责与6块HBM2存储器数据交互,顶部为PCIe 4.0控制器负责与主机通信,底部又有12条高速NVLink通道与其他GPU连为一体。



## 2.3 GPU架构

- ✓ 每个 SM 包括 4 个区块，每个区块有独立的 L0 指令缓存、Warp 调度器、分发单元，以及 16384 个 32 位寄存器，这使得每个 SM 可以并行执行 4 组不同指令序列。4 个区块共享 L1 指令缓存和数据缓存、shared memory、纹理单元。
- ✓ 每个 SM 除了 INT32、FP32、FP64 计算单元之外，还有额外 4 个身宽体胖的 Tensor Core，这是加速 Deep Learning 计算的重磅武器，已发展到第三代，每个时钟周期可做 1024 次 FP16 乘加运算，与 Volta 和 Turing 相比，每个 SM 的吞吐翻倍，支持的数据类型也更为丰富，包括 FP64、TF32、FP16、BF16、INT8、INT4、INT1(另外还有 BF16)。

表 GPU架构介绍

NVIDIA Architecture	CUDA Cores				Tensor Cores					
	FP64	FP32	FP16	INT8	FP64	TF32	FP16	INT8	INT4	INT1
Volta	32	64	128	256			512			
Turing	2	64	128	256			521	1024	2048	8192
Ampere ( A100 )	32	64	256	256	64	512	1024	2048	4096	16384
Ampere, sparse						1024	2048	4096	8192	

资料来源: Nvidia官网, 艾瑞咨询, 华西证券研究所

表 GA100(完整版)与A100(阉割版)性能对比

	GA100(完整版)	A100 GPU ( 阉割版 )
GPC	8	7
TPC	64	56
SM	128	108
CUDA Core	8192	6912
Tensor Core	512	432
HBM2	6	5
Mem Controller	12	10
Mem Bits	6144	5120

资料来源: Nvidia官网, 艾瑞咨询, 华西证券研究所

## 2.4 GPU API 接口 两大阵营

- ✓ **GPU的API ( Application Programming Interface ) 应用程序接口**发挥着连接应用程序和显卡驱动的桥梁作用。目前GPU API可以分为2大阵营和若干其他类。2大阵营分别是微软的DirectX标准和KhronosGroup标准，其他类包括苹果的Metal API、AMD的Mantle ( 地幔 ) API、英特尔的One API等。
- ✓ DirectX可以让以Windows为平台的游戏或多媒体程序获得更高的执行效率，加强3D图形和声音效果，并提供设计人员一个共同的硬件驱动标准，让游戏开发者不必为每一品牌的硬件来写不同的驱动程序，也降低用户安装及设置硬件的复杂度。DirectX已被广泛使用于Windows操作系统和Xbox主机的电子游戏开发
- ✓ OpenGL是Open Graphics Library的简称，是用于渲染2D、3D矢量图形的跨语言、跨平台的应用程序编程接口 ( API )，相比DirectX更加开放。这个接口由近350个不同的函数调用组成，用来绘制从简单的二维图形到复杂的三维景象。OpenGL常用于CAD、虚拟现实、科学可视化程序和电子游戏开发

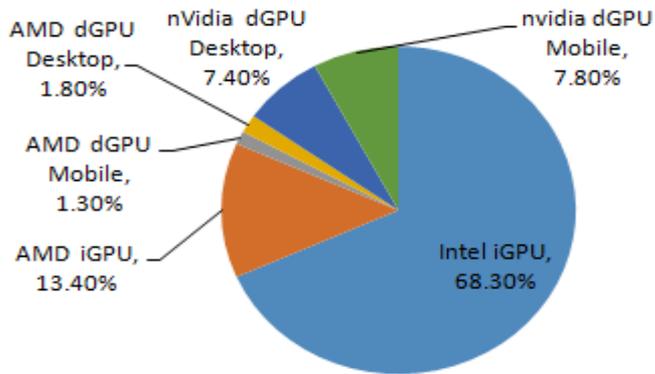
表 GPU API接口及应用领域

厂家	标准	应用领域
微软DirectX	Direct3D	3D图形
	Direct2D	2D图形
	DirectCompute	通用计算
	XAudio	音频
	XInput	Xbox手柄
Khronos Group	OpenGL	图形
	Vulkan	图形
	OpenGL ES	移动图形
	WebGL	网页图形
	OpenCL	通用计算

## 2.5 智能驾驶及云计算发展对高性能GPU带来增量需求

- 按GPU下游的不同应用，可分为终端GPU，服务器GPU，智能驾驶GPU以及军用显控等其他应用领域GPU。在终端GPU中分为集成GPU（集显）与独立GPU（独显），前者注重轻薄，后者注重性能输出。服务器等高性能需求场景下**GPU以独立为主**。
- 从下游应用来看，GPU于手机及PC端渗透率基本见顶，根据中国社科院数据，2011-2018年全球主要国家PC每百人渗透率呈下降趋势，智能手机对PC具有一定替代性。而云计算与智能驾驶及AI的兴起对高算力产生新需求，将带来高性能GPU市场快速增长。

图 2021年二季度GPU市场份额



资料来源：JPR、华西证券研究所

表 目前主要 AI 场景性能需求

场景	芯片需求	典型计算能力	典型功耗
终端	低功耗、低成本	<8TOPS	<5W
云端	高性能	>30TOPS	>50W
边缘/终端	中高性能	5-30TOPS	4-15W

资料来源：架构师技术联盟、华西证券研究所

表 GPU 分类与主要厂商

主要应用端	主要功用	性能要求	主要厂商(产品)
个人终端 (PC/主机/ 手机)	独显	图形设计/3A游戏	高 NVIDIA ( GeForce系列 ) , AMD ( Radeon系列 )
	集显	轻度办公	低 Intel ( HD系列 ) ,AMD ( APU系列 ) , Imagination ( PowerVR系列 ) ,高通骁龙 ( Adreno系列 ) , 苹果
服务器	AI训练/AI推断/科学计算/图形图像处理/视频编解码	高	NVIDIA ( TESLA),AMD(Instinct)
智能驾驶	AI推断	高	NVIDIA ( Orin )

资料来源：架构师技术联盟、华西证券研究所

## 2.6 算力新时代，GPU生态领先，AI芯片三剑客将互补共享市场

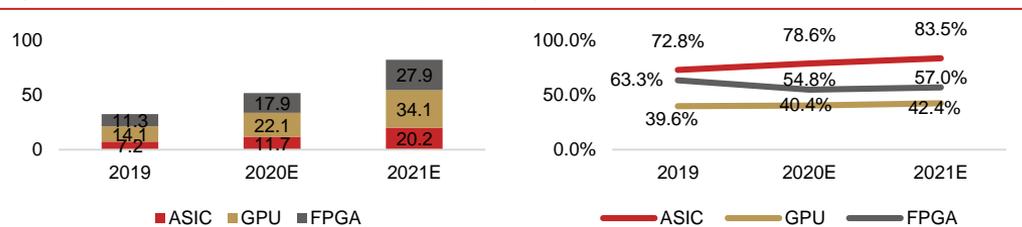
- **AI芯片 (GPU/FPGA/ASIC) 在云端同时承担人工智能“训练”和“推断”过程，在终端主要承担“推断”过程，从性能与成本来看ASIC最优。**ASIC作为专用芯片，算力与功耗在通用芯片GPU具有绝对优势，但开发周期较长，落地较慢，需一定规模后才能体现成本优势。FPGA可以看做从GPU到ASIC重点过渡方案。相对于GPU可深入到硬件级优化，相比ASIC在算法不断迭代演进情况下更具灵活性，且开发时间更短。
- **从生态与落地来看，GPU占据绝对优势，英伟达处垄断地位。**开发者能通过英伟达CUDA平台使用软件语言很方便地开发英伟达GPU实现运算加速，已被广泛认可和普及，积累了良好的编程环境。以TPU为代表的ASIC目前主要运用在巨头的闭环生态，FPGA在数据中心业务中发展较快。

表 不同芯片的主要企业及竞争

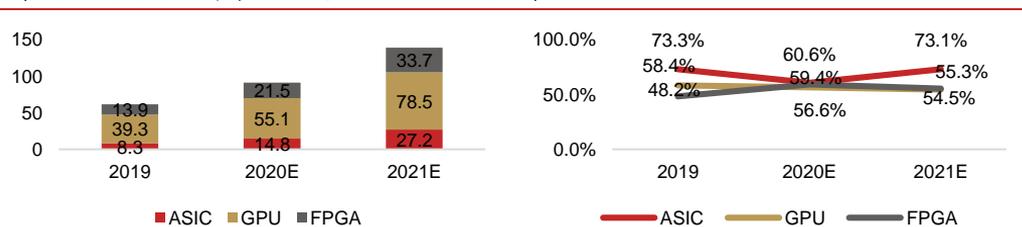
芯片种类	特征	主要企业	竞争
GPU	图像处理、密集型并行运算	英伟达、AMD	集中
FPGA	可定制编程反复烧写	Xilinx、英特尔	集中
ASIC	可根据算法进行定制	谷歌、寒武纪、地平线	分散

资料来源：赛迪顾问、前瞻产业研究院、华西证券研究所

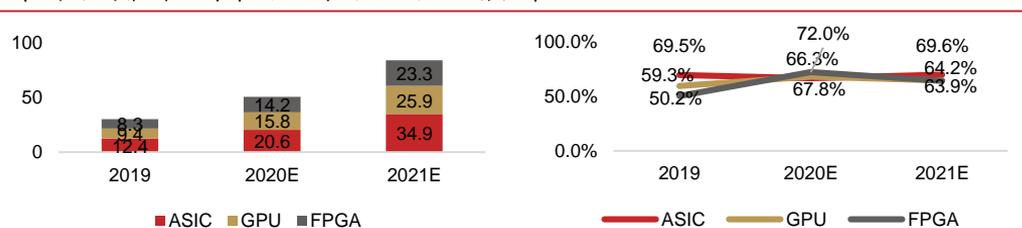
中国云端推断芯片市场结构 (亿元) 及增长率



中国云端训练芯片市场结构 (亿元) 及增长率



中国终端推断芯片市场结构 (亿元) 及增长率

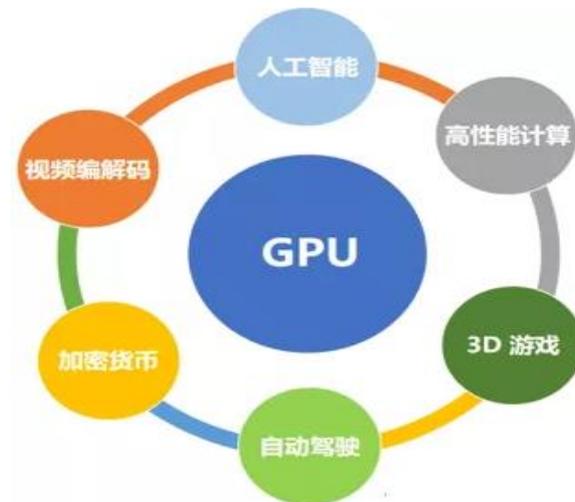
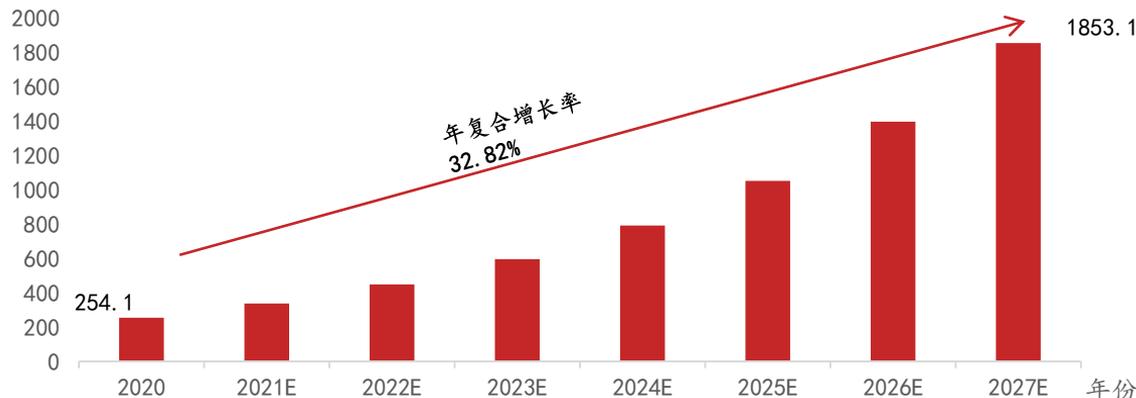


## 2.7 GPU 市场规模与细分市场概述

- ✓ 2020年GPU市场规模为254.1亿美元，预计到2027年将达到1853.1亿美元，从2021年到2027年的复合年增长率为32.82%。
- ✓ GPU市场分为独立，集成和混合市场。2019年集成占据了GPU市场份额的主导地位，但由于混合处理器同时具有集成和独立GPU的能力，因此未来混合细分市场预计将实现最高复合年增长率。
- ✓ 市场分为计算机，平板电脑，智能手机，游戏机，电视等。在2019年，智能手机市场占据了全球GPU市场份额的主导地位，预计在预测期内将继续保持这一趋势。但是，由于对医疗设备等其他设备中对小型GPU的需求不断增长，预计其他领域在未来的复合年增长率最高。由于在设计和工程应用中图形处理器的广泛使用，预计汽车应用细分市场将在预测期内以最高的复合年增长率增长。

GPU 应用场景

GPU全球市场规模预测（单位：亿美元）



## 2.8 GPU三大应用场景——游戏 & AI & 自动驾驶

		
		
<p><b>Gaming – \$100B Market</b></p>	<p><b>AI – Every Internet Query, Enterprise Infused with Intelligence</b></p>	 <p><b>Autonomous Vehicles – \$10T Transportation Industry</b></p>

## 2.9 GPU三大应用场景1——游戏渲染场景对GPU需求旺盛

### 全球游戏PC和显示器的出货量强劲

- ✓ IDC数据显示，2020年游戏PC和显示器的出货量同比增长26.8%，达到5500万台。游戏笔记本电脑在2020年增长了创纪录的26.9%。与PC并行，游戏显示器在2020年也达到了新的高度，与2019年相比增长了77%以上，出货量达到了1430万台。
- ✓ IDC预计2021年游戏显示器的销量将首次超过游戏台式机。即使游戏台式机逐渐受到青睐，游戏笔记本电脑的显示器连接率不断提高也意味着游戏显示器市场的五年复合年增长率预计将超过10%。IDC预计2025年全球销量达到7290万，复合年增长率为5.8%。

表：2020-2025年 游戏 PC预测增长情况

Product Category	2020 Shipments	2020 Average Selling Price	2021 Shipments*	2021 Average Selling Price*	2025 Shipments*	2025 Average Selling Price*	2020-2025 CAGR
Desktop	16.7	\$762	17.5	\$750	15.7	\$721	-1.2%
Monitor	14.3	\$338	19.7	\$352	23.5	\$343	10.5%
Notebook	24.0	\$1041	27.9	\$1055	33.7	\$1030	7.0%
TOTAL	55.0		65.1		72.9		5.8%

### GPU的画质的实时渲染效果增强游戏用户体验

- ✓ 游戏绘图是GPU的传统应用领域，为游戏开发者提供电影级画质的实时渲染
- ✓ GPU 并行计算的基础结构、可以执行海量数据计算
- ✓ GPU 访存速度快
- ✓ GPU拥有更高的浮点运算能力，对图形与媒体加工速度快。

图：英伟达GPU游戏渲染效果



RTX的光追为水面镜面带来逼真的光影效果

## 2.10.1 GPU三大应用场景2——为移动互联网消费电子类产品带来新增市场空间

✓ 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。

2017-2021H1中国AI音箱市场销量及增速（单位：万台、%）

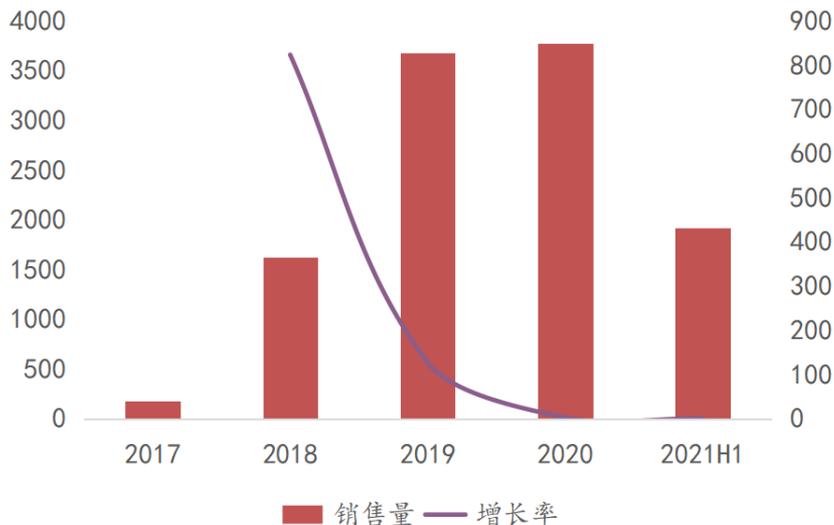


表 主要手机AI芯片解决方案

厂商		SoC	AI技术	性能
SoC 厂商	高通	骁龙845	Hexagon 685DPS+CPU+GPU	
		骁龙855	Hexagon 690DPS+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 1N	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	

## 2.10.2 GPU三大应用场景2——边缘侧场景复杂，综合考量芯片性能

- ✓ 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，对于算力和能耗等性能需求也有大有小。因此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。

表 不同边缘计算场景对AI芯片性能要求

	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
算力 Performance	<1TOPs	1-8TOPs	4-20TOPs	20-4000TOPs (L3-L5)
性能 要求	接入设备部署现场电源	消费级聚合物锂电池 2000-5000mAh	接入设备部署现场电源	动力级硬壳锂电池200000- 500000mAh
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	极高
可靠性 Reliability	高 (工业) / 中 (家用)	中	高	极高
代表厂商	谷歌、英伟达、云知声、 思必驰	苹果、三星、ARM、 cadence、寒武纪	北京君正、安霸科技、云 天励飞	英伟达、英特尔、恩智浦

## 2.10.3 GPU三大应用场景2——智慧安防

- ✓ 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

表 主要安防AI 解决方案

#### 数字/网络高清

##### • 前端：IPC SoC

集成CPU、ISP、高压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

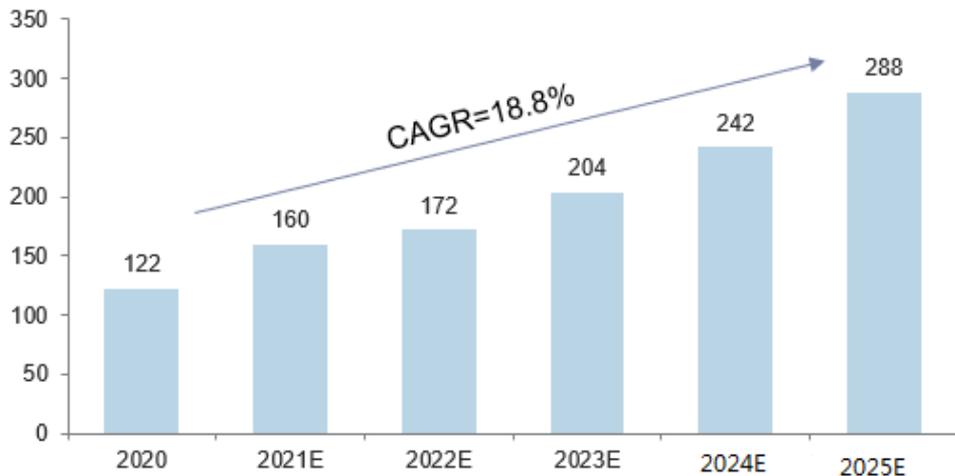
文教卫

	厂商	产品	算力
前端	英伟达	JETSON TX-GPU	1TOPs
	华为海思	Hi3559 AV100	4TOPs
	北京君正	T01-ASIC	
后端	地平线	旭日	1TOPs
	英伟达	TESLA P4	5.5Flo ps
	比特大陆	BM1682	3Flo ps

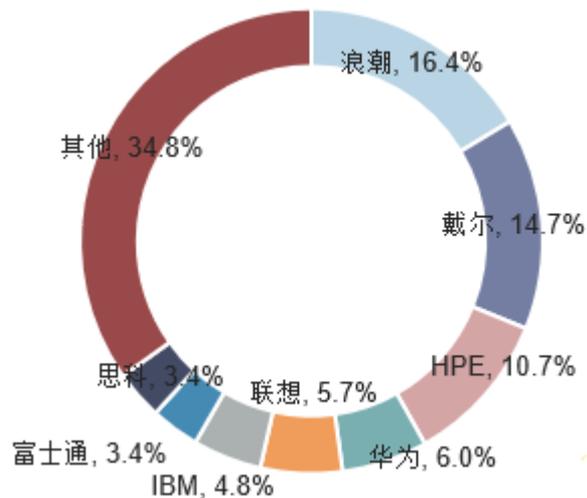
## 2.11.1 GPU三大应用场景2——云端GPU使用

- ✓ 人工智能服务器通常搭载GPU、FPGA、ASIC等加速芯片，利用CPU与加速芯片的组合可以满足高吞吐量互联的需求，为自然语言处理、计算机视觉、语音交互等人工智能应用场景提供强大的算力支持，已经成为人工智能发展的重要支撑力量相比于传统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。
- ✓ 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。2020年全球AI服务器市场规模为122亿美元，预计到2025年全球AI智能服务器市场将达到288亿美元，5年CAGR达到18.8%。

2020-2025年全球AI服务器行业市场规模及增速（单位：亿美元）



2020年上半年全球AI服务器行业市场竞争格局



## 2.11.2 GPU三大应用场景2——云计算服务模式可显著降低AI开发部署成本

- ✓ 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。

**表：部分云端AI芯片售价**

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

**表：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	\$399,00
DGX-1	8*Tesla V100	2*intel Xeon E5- 2698v4	\$149,00
DGX-1(Pascal)	8*Tesla P100	2*intel Xeon E5-2698v3	\$129,00

**表：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
部署模型	标准实例	0.07-6.45
	GPU实例	1.26-34.27

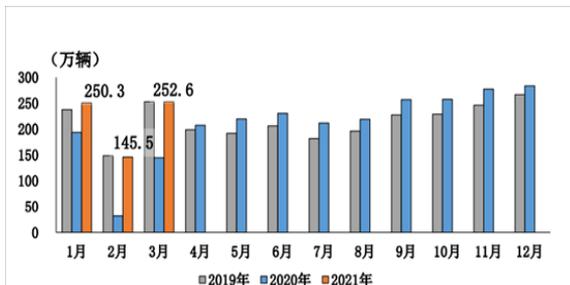
## 2.12.1 GPU三大应用场景3——自动驾驶不断推进

- ✓ 全球自动驾驶迈入商用阶段，未来可期。IDC最新发布的《全球自动驾驶汽车预测报告（2020-2024）》数据显示，2024年全球L1-L5级自动驾驶汽车出货量预计将达到约5425万辆，2020至2024年的年均复合增长率（CAGR）达到18.3%；L1和L2级自动驾驶在2024年的市场份额预计分别为64.4%和34.0%。尽管目前L3-L5级自动驾驶技术的应用具有开拓性意义，L1-L2级自动驾驶将依然是未来5年内带动全球自动驾驶汽车出货量增长的最大细分市场。
- ✓ 我国汽车市场规模不断增长，自动驾驶由L2向L3过渡。中汽协数据显示，2021年1-3月，中国品牌乘用车共销售210.8万辆，同比增长81.5%，占乘用车销售总量的41.5%，占有率比上年同期提升1.4个百分点。2020年1月份至9月份，L2级智能网联乘用车销售量达196万辆，占乘用车总销量的14.7%。更有部分企业加速研发L3级自动驾驶车型，多地开展自动泊车、自动驾驶公交车、无人智能重卡等方面的示范应用。到2025年，我国PA（部分自动驾驶）、CA（有条件自动驾驶）级智能网联汽车销量占当年汽车总销量比例超过50%，C-V2X（以蜂窝通信为基础的移动车联网）终端新车装配率达50%。

2020-2024全球自动驾驶汽车出货量及增长率预测



2021年度1-3月份汽车月度销量



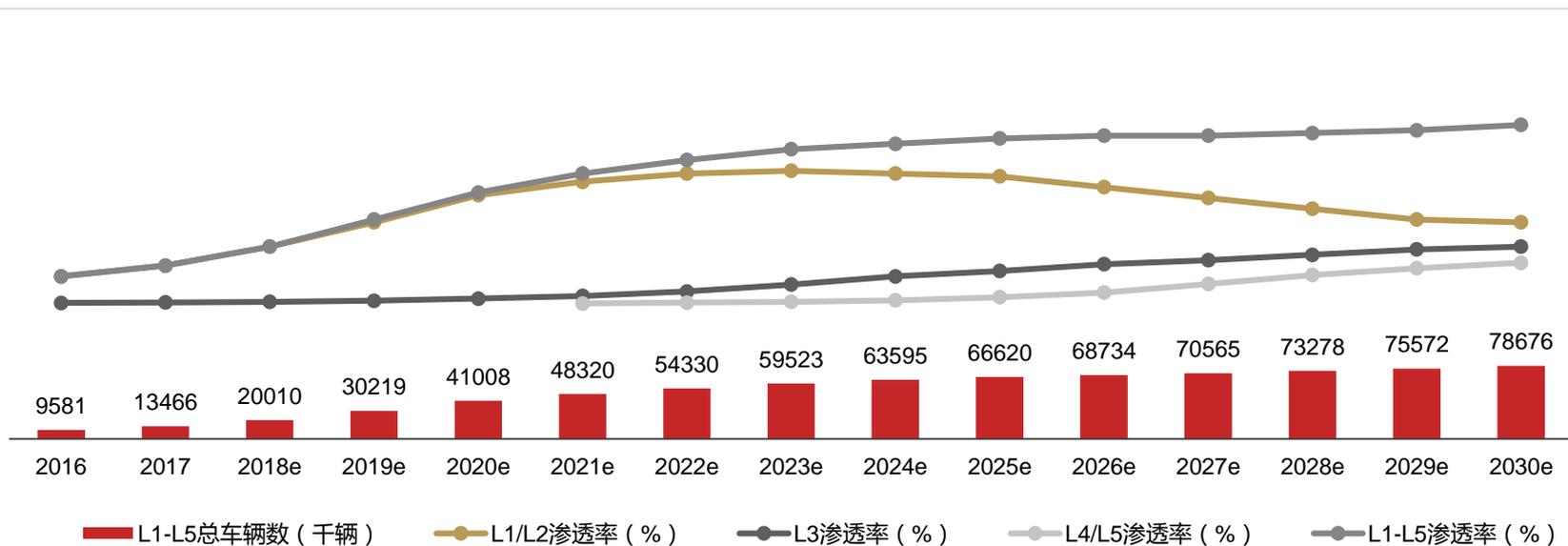
全球汽车智能驾驶市场渗透率预测



## 2.12.2 GPU三大应用场景3——自动驾驶不断推进

- ✓ 随着传感器、车载处理器等产品的进一步完善，将会有更多L3级车型出现。而L4、L5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。

### 2016-2030年全球汽车市场自动驾驶渗透率预测



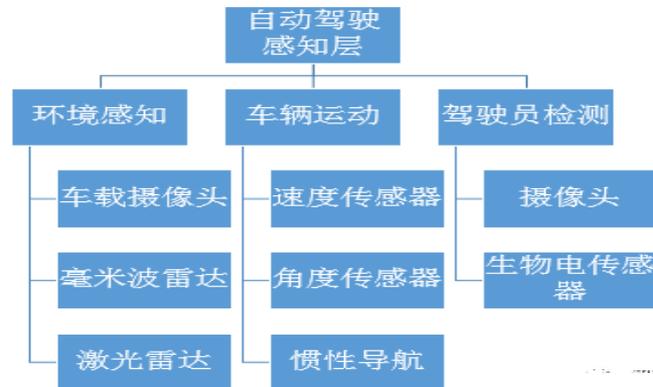
## 2.12.3 GPU三大应用场景3——GPU为自动驾驶主流方案

- ✓ 感知路境，短时处理海量数据。行车过程中依赖雷达等传感器对道路信息进行采集后，处理器每秒需实时数据解析几G量级数据，每秒可以产生超过 1G 的数据。对处理器的计算量要求较高。
- ✓ 自动规划，瞬时反应保障安全。处理分析实时数据后，需要在毫秒的时间精度下对行车路径、车速进行规划，保障行车过程安全，对处理器的计算速度要求较高。
- ✓ 兼具技术成本优势，GPU为自动驾驶领域主流。

图：自动驾驶常见场景



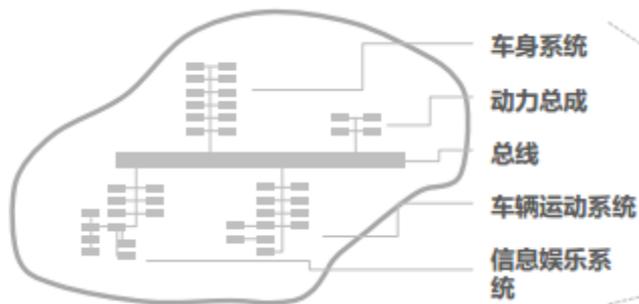
图：自动驾驶感知层



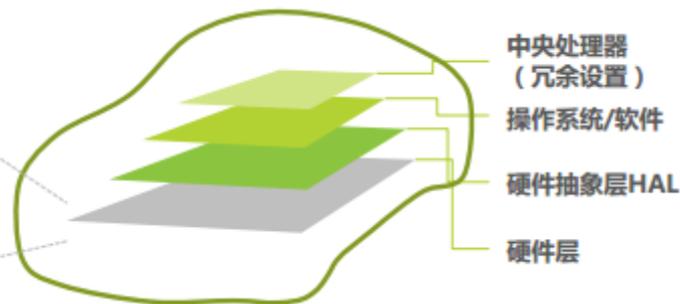
## 2.12.4 GPU三大应用场景3——GPU为自动驾驶主流方案

- ✓ 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，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
算力要求		<1TOPs	10-50TOPs	>50TOPs



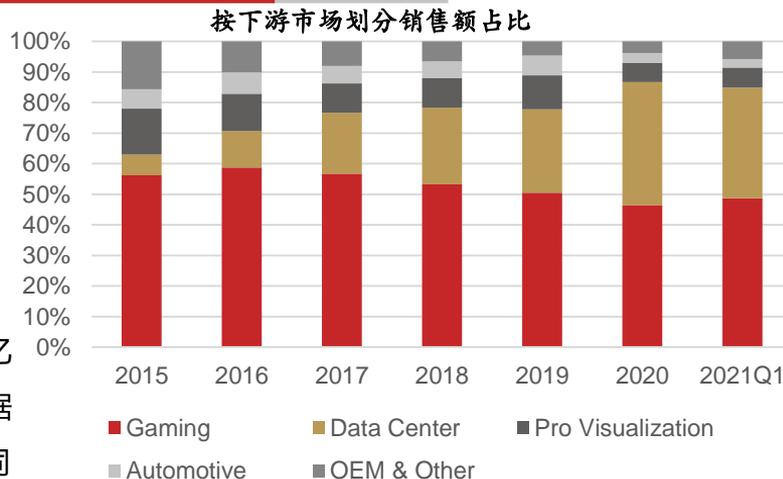
## 目录

- 1 算力时代，GPU开拓新场景
- 2 GPU 下游三大应用市场
- 3 海外GPU巨头Nvidia
- 4 国产GPU赛道掀起投资热潮
- 5 重点投资机会
- 6 风险提示

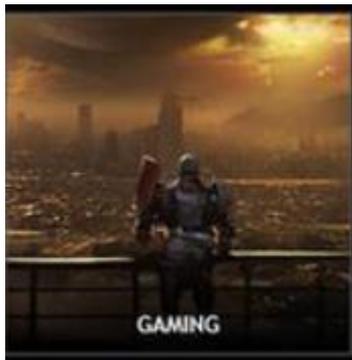
### 3.1 Nvidia : 四大产品线平台 , 游戏业务和数据中心业务快速增长

- ✓ 1) **游戏业务** : GeForce GPU、云游戏平台及相关产品设施
- ✓ 2) **数据中心** : 包括Telsa系列的GPU产品、EGX&HGX&DGX 终端产品为数据中心加速器、边缘计算、AI计算 数据中心**服务器 Grace CPU**;  
DPU : 收购Mellanox , 积极布局数据中心存储、传输 **DPU**;
- ✓ 3) **专业显示** : Quadro 专业绘图工作站、RTX图形处理器、vGPU软件
- ✓ 4) **汽车** : DRIVE AGX 为自动驾驶工作平台

**Q3 FY2022** 英伟达第三季度营收达71.03亿美元, **同比增长50%**。净利润为24.64亿美元, **同比增长84%**。其中游戏业务收入达32.2亿美元, 较去年同期增长42%; 数据中心业务收入为29.4亿美元, 同比增长55%。专业可视化业务收入达5.77亿美元, 同比增长144%; 自动驾驶业务收入达1.35亿美元, 同比增长8%。



资料来源: nvidia 年报, 华西证券研究所

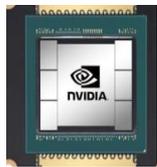


## 3.2 Nvidia : GPU 架构每两年升级一次

2010年

### 首个完整GPU架构

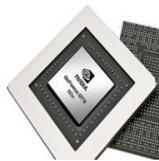
Fermi是第一个可支持与共享存储结合纯cache层次的GPU架构，包含512个accelerator cores，16个SM



Tesla



Fermi



Kepler



Maxwell



Pascal



Volta



Turing



Ampere

2008年

### CUDA架构问世

这是一种用于通用GPU计算的革命性架构。借助CUDA，能够利用GPU的并行处理能力来应对最为复杂的计算挑战。

2012年

首次在GPU中引入了动态并行技术

2014年

Maxwell架构可解决视觉计算领域中最复杂的光照和图形难题，优化功耗，支持微软DX12图形加速接口

2017年

### Volta架构:

首次引入Tensor（张量）运算单元。

2016年

### 适用于大数据工作

Pascal采用了HBM2的CoWoS技术，在深度学习方面，与当代GPU架构相比训练性能提高12倍。首次引入了3D内存及NVLink高速互联总线

2020年

### 人工智能高效运算

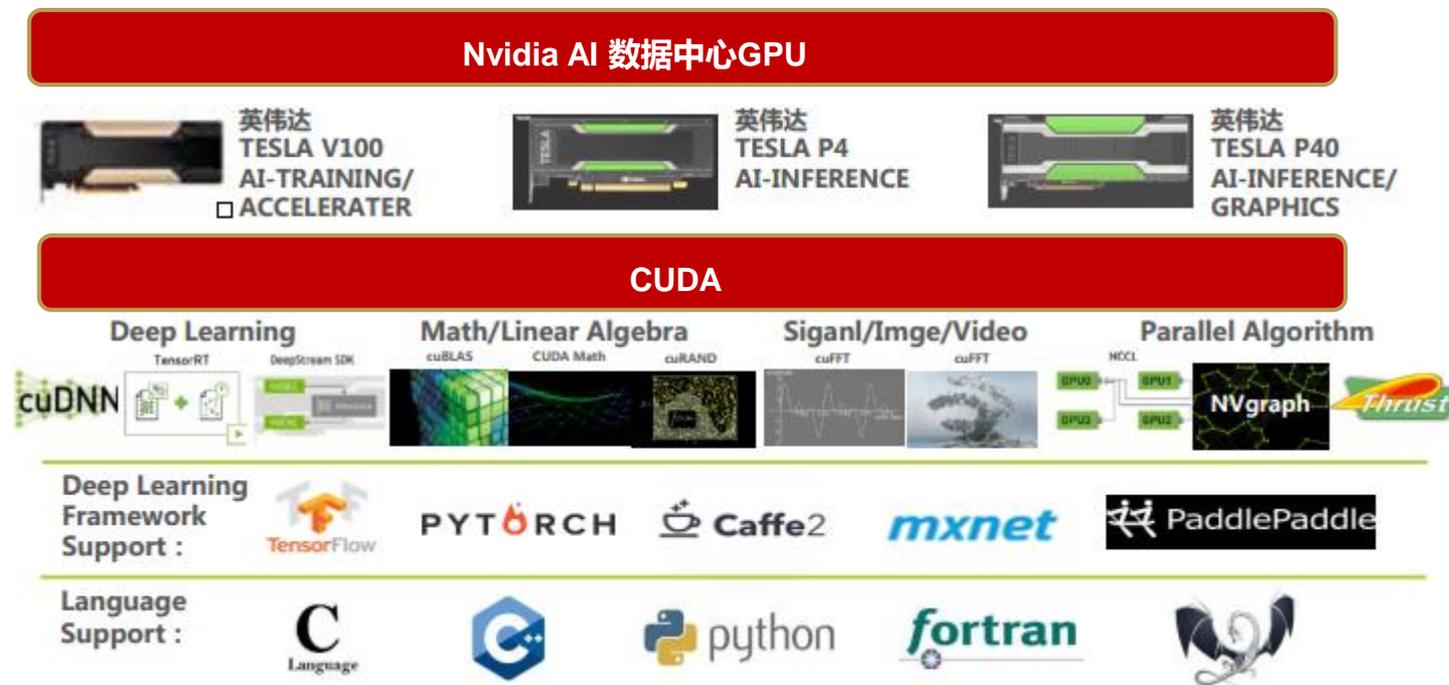
Ampere架构包含540亿个晶体管，是有史以来最大的7nm芯片，通过人工智能和高效能运算解决了全球最重要的科学、产业与巨量资料。

2017年

Turing架构：架构最大的变革，引入了RTX追光技术总线

### 3.3 Nvidia : 推出CUDA 软件堆栈，建立软硬件一体化生态

- ✓ GPU通用计算方面的标准目前有OpenCL、CUDA、AMD APP、DirectCompute。其中OpenCL、DirectCompute、AMD APP(基于开放型标准OpenCL开发)是开放标准，CUDA是私有标准。
- ✓ 2006年，公司推出CUDA 软件推展，推动GPU 向通用计算转变，之后不断强化通用系统生态构建。为开发者提供了丰富的开发软件站SDK、支持现有的大部分机器学习、深度学习开发框架。推出的cuDNN、TensorRT、DeepStream 等优化的软件也为 GPU 通用计算提供加速。



### 3.4 Nvidia : 强劲的游戏业务提供稳定现金流

- ✓ NVIDIA Geforce系列主要面向家庭和企业的娱乐应用，分为面向性能的GTX系列,面向主流市场的GTS和GT系列，已经具有高性价比的GS系列。
- ✓ 衡量显卡性能的参数可包括：1、核心数目；2、显存带宽；3、峰值单精度浮点计算能力；4、峰值双精度浮点计算能力；5、时钟频率；6、架构版本

表 英伟达RXT系列显卡性能对比

发布时间	架构名	版本号	制造工艺	代表作品
2006	Tesla ( G80 )	1.0	TSMC 90nm	GeForce 8800
2008	Tesla 2.0 ( GT200 )	1.3	TSMC 65nm	GeForce GTX 280
2010	Fermi	2.0	TSMC 40nm	GeForce GTX 480
2012	Kepler	3.0、3.5、3.7	TSMC 28nm	GeForce GTX 680 , Tesla K20/K40/K80
2014	Maxwell	5.0、5.2	TSMC 28nm	GeForce GTX 750/980 , Tesla M40
2016	Pascal	6.0、6.1、6.2	TSMC 16nm	GeForce GTX 1080/1070/1060,Tesla P100/P40/P4
2017	Volta	7.0	TSMC 12nm	Tesla v100 , Titan V
2018	Turning	7.5	TSMC 12nm	RTX 2080Ti , Quadro RTX 6000/8000 , Tesla T4
2020	Ampere	8.0	TSMC 7nm	Tesla A100
			Samsung 8nm	RTX 3080/3070/3090

## 3.5 Nvidia : 数据中心提供硬件、软件和整体方案

- ✓ 数据中心主要产品为Tesla 系列。通过将不同类型的GPU加速器与CPU、DDR等硬件产品组合以及软件开发，推出面向高性能计算HPC、人工智能DGX、边缘计算EGX等硬件产品。
- ✓ 软件产品NGC用于GPU加速的云平台服务，加速使用深度学习框架；虚拟GPU主要用于在虚拟机上运营AI、深度学习和高性能计算。

表 不同产品具体介绍

产品分类	产品名称	发布时间	主要参数及性能	售价(万美元)
GPU Tesla 系列 ( 计算显卡 )	<b>P100</b>	2016	Pascal架构, 3584个CUDA cores, 单精度10.6T, 显存16GB	0.75
	P4	2016	Pascal架构, 3584个CUDA cores, 单精度8T, 显存8GB	0.25
	<b>V100</b>	2017	Volta架构, 5120个CUDA , 单精度15.7T, 显存32GB或16GB	1.15
	T4	2018	Turing架构, 2560个CUDA cores, 单精度8.1T, 显存6GB	0.25
	<b>A100</b>	2020	Ampepre架构, 6912个CUDA cores, 单精度19.5T, 显存40GB	
DGX ( 主要用于AI )	DGX-1	2017	8个Tesla V100 GPU, 512Gb DDR4, 2个Inter E5-2698 CPU	14.9
	DGX-2	2018	16个Tesla V100 GPU, 1.5TB内存, 2个Inter 8168 CPU	39.9
	DGX-A100	2020	8个Tesla A100 GPU, 1TB内存, 2个AMD 7742 CPU	19.9
HGX ( 用于AI和超大型数据中心加速器 )	HGX-1	2017	8个Tesla V100 GPU, 256GB显存, AI算力达1petaFLOPS	14.9
	HGX-2	2018	16个Tesla V100 GPU, 512GB显存, AI算力达2petaFLOPS	39.9
	HGX-3	2020	共三个版本, 分别搭载4/8/16个Tesla A100 , 显存为160/320/640 GB, AI算力为达2.5/5/10 petaFLOPS	

资料来源：英伟达官网，华西证券研究所

## 3.6 Nvidia : 数据中心业务持续快速增长

- ✓ 2020年4月，Nvidia 70亿美元完成对Mellanox的收购。形成Nvidia GPU+Mellanox RDMA+Nvidia CUDA 整体解决方案，提供从人工智能计算到网络的端到端技术全堆栈产品。
- ✓ 受益于市场对公司新一代安培架构产品的需求，同时对话式AI、推荐系统等下游场景的训练模型部署的增多。22财年Q3，公司数据中心业务收入为29.4亿美元，同比增长55%。

公司数据中心营业收入（单位：百万美元）

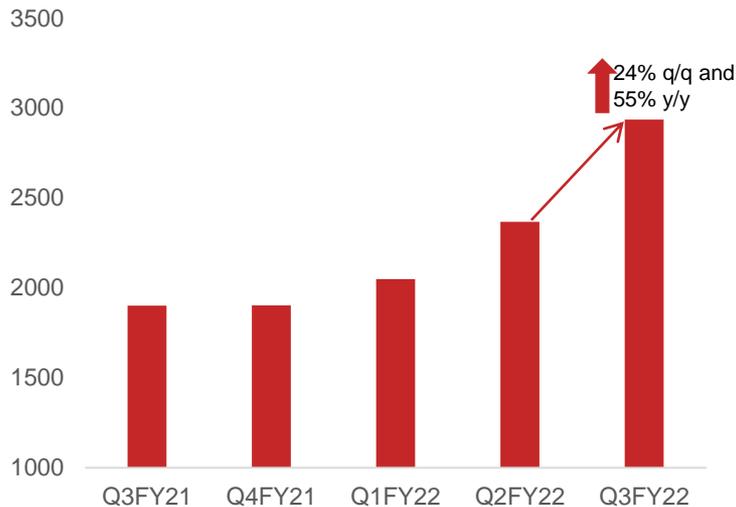


表 英伟达主流推理和训练芯片性能参数

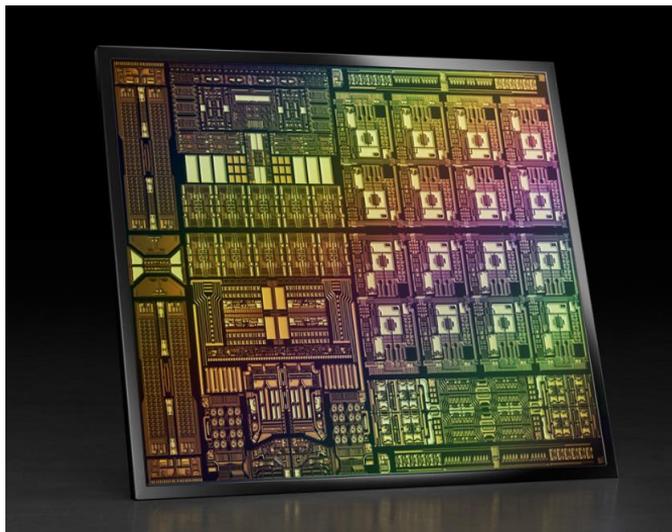
型号	架构	制程	算力	功耗	显存带宽	芯片面积
P100	Pascal	16nm	21TOPS(FP16) 11TOPS(FP32)	300W	732GB/s	610
V100	Volta	112nm	125TOPS(FP16) 16TOPS(FP32)	300W	900GB/s	815
A100	Amper e	7nm	624TOPS(INT8) 312TOPS(FP16) 20TOPS(FP32)	400W	1555GB/ s	826

资料来源：英伟达官网，华西证券研究所

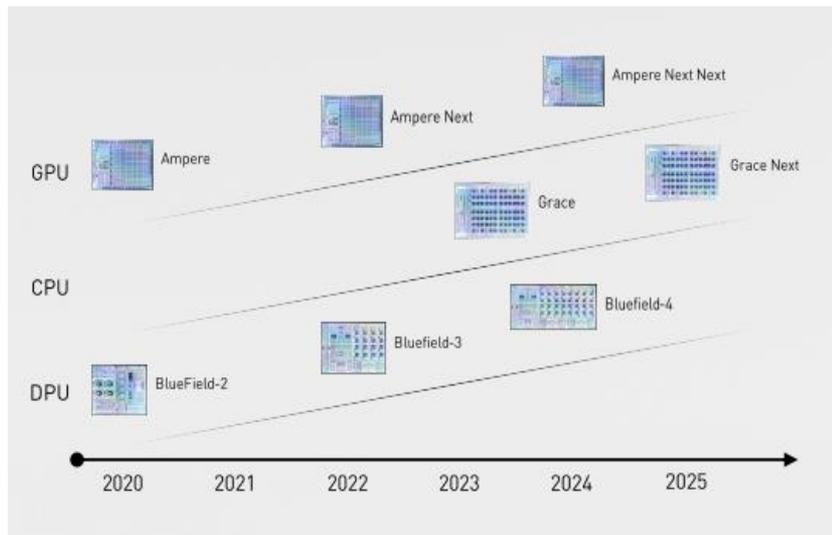
### 3.7 Nvidia : 提供 CPU+GPU+DPU整体方案

- ✓ 融合了Mellanox 的计算推了DPU的产品。BlueField DPU 通过分流、加速和隔离各种高级网络、存储和安全服务，为云、数据中心或边缘等环境中的各种工作负载提供安全的加速基础设施。BlueField DPU 将计算能力、数据中心基础功能的可编程性及高性能网络相结合，可实现非常高的工作负载。
- ✓ GPC 2021年推出了基于ARM 架构的面向服务器市场的CPU，用于大型计算中心或者超级计算机等场景中，通过Nvlink 实现 CPU、GPU 之间的大带宽链接和交互。未来数据中心将具备 GPU+CPU+DPU 整体解决方案。

图：NVIDIA BlueField-3 DPU



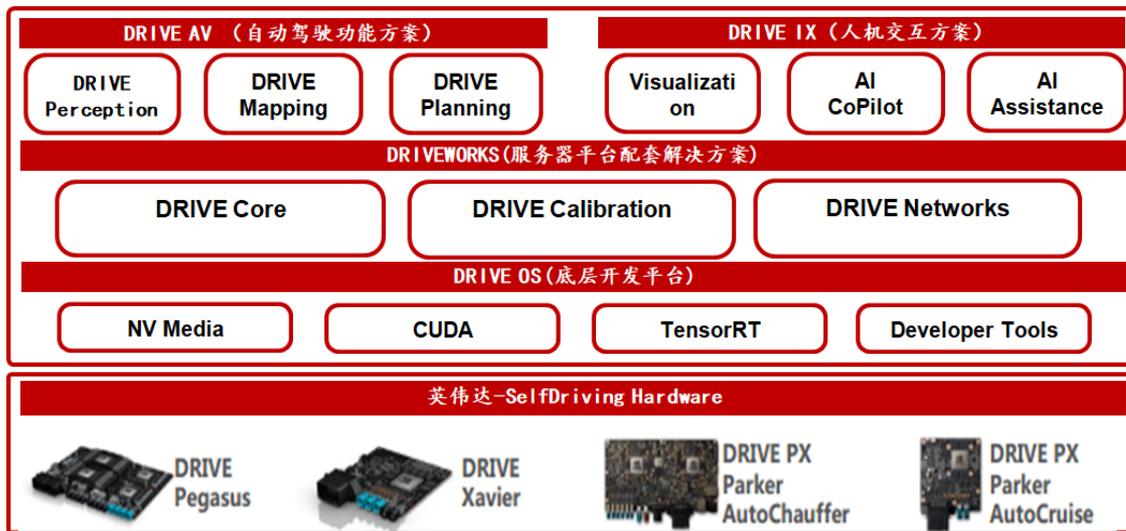
图：英伟达的数据中心产品线布局



## 3.8 Nvidia : 自动驾驶业务步入收获期

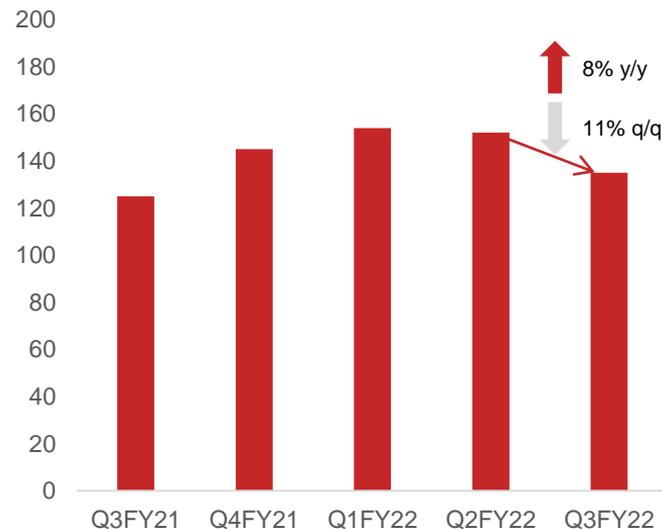
- ✓ 公司形成了分层解耦、全栈式的自动驾驶方案。同时推出了自动驾驶配套的仿真系统、底层开发平台、服务平台解决方案、自动驾驶功能方案和人机交互方案等，覆盖了硬件+软件的一体化解决方案，加快下游客户自动驾驶计算的测试与开发。
- ✓ Q3FY22汽车业务收入为1.35亿美元，同比增长8%，环比下降11%。环比下降主要是受到汽车供应链紧张限制。

公司提供自动驾驶整体解决方案



资料来源：艾瑞咨询，华西证券研究所

公司自动驾驶业务营业收入（单位：百万美元）



资料来源：英伟达官网，华西证券研究所

### 3.9 Nvidia : 覆盖L2/5自动驾驶应用场景

- ✓ 公司累计发布了7款自动驾驶芯片：Tegra X1、Tegra Parker、Tegra Xavier、Driver Xavier、Drive AGX Orin、Orin、Atlan 等。
- ✓ 公司发布8款自动驾驶计算平台，包括Driver PX、Driver PX2、Driver PX Xavier、Driver PX Pegasus、Driver AGX **Xavier**、Driver AGX Pegasus、Drive **Orin**、Pegasus Robotaxi。随着Xavier 和Orin 自动驾驶计算平台的发布，公司新增覆盖L2/3级别的自动驾驶应用场景，目前已经覆盖L2-L5级的自动驾驶的应用场景。

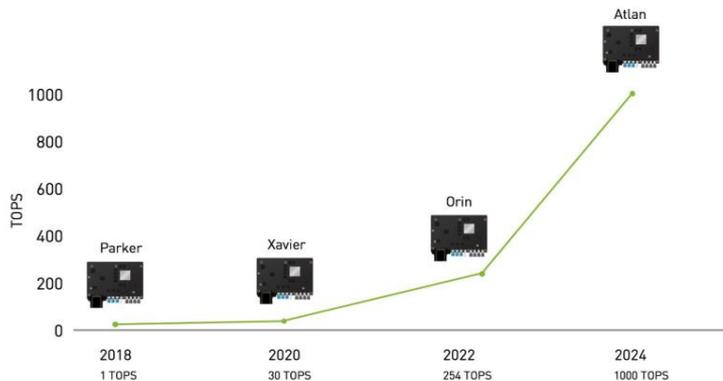


表 英伟达、特斯拉和Mobileye自动驾驶芯片性能对比

公司	型号	算力	功耗	制程	发布时间	量产时间
英伟达	Tegra X1	2TOPS	10W	20nm	2015	
	Tegra Parker	8TOPS	15W	16nm	2016	
	Tegra Xavier	20TOPS	20W	16nm	2017	
	Drive Xavier	30TOPS	30W	12nm	2018	
	Drive AGX Orin	254TOPS	45W	7nm	2019	2022
	Atlan	1000TOPS	N/A	N/A	2021	2023
特斯拉	Tesla FSD	72TOPS	72W	14nm	2019	2020
	Tesla D1	22.6TFLOPS	TDP 400W	7nm	2021	2023
	EyeQ	0.26TOPS	2.5W	40nm		2014
Mobileye	EyeQ4	2.5TOPS	3W	28nm		2018
	EyeQ5	12TOPS	5W	7nm		2020

资料来源：英伟达官网，特斯拉官网，英特尔官网，华西证券研究所

### 3.10 Nvidia : 大算力芯片+计算平台+基础软件构建自动驾驶平台

- ✓ GTC2019 推出 NVIDIA DRIVE AGX Orin，集成了170亿个晶体管并使用新的NVIDIA GPU和12核ARM CPU，性能是上一代Xavier的7倍，将于2022年投入量产。DRIVE AGX Orin可以为从L2到L5的全自动驾驶开发提供兼容的架构平台。2021年4月13日发布最新一代超算力芯片 Atlan，单芯片算力达1,000TOPS，可满足L5需求，预计2023年提供样品。
- ✓ 拥有完备软件开发生态（操作系统Drive OS+中间件Drive Works+软件堆栈Drive AV），工具链稳定，开放程度高。

表 英伟达自动驾驶SOC 芯片

SOC	SOP 时间	适用 场景	内核	算力 (TOPS)	功耗 (W)	制程
Parker	2016	L2-L3	NVIDIA Denver *2 ARM Cortex-A57 *4 Parker Pascal iGPU *1	--	15	16nm
Xavier	2020	L2-L5	ARM Carmel CPU *8 Xavier Volta iGPU *1	30	30	12nm
Orin	2021E	L2-L5	ARM Hercules CPU *12 Next-generation iGPU *1	200	45	--
Atlan	2023E	L4-L5	New ARM Grace CPU Next-generation iGPU	1,000	--	--

表 英伟达自动驾驶计算平台

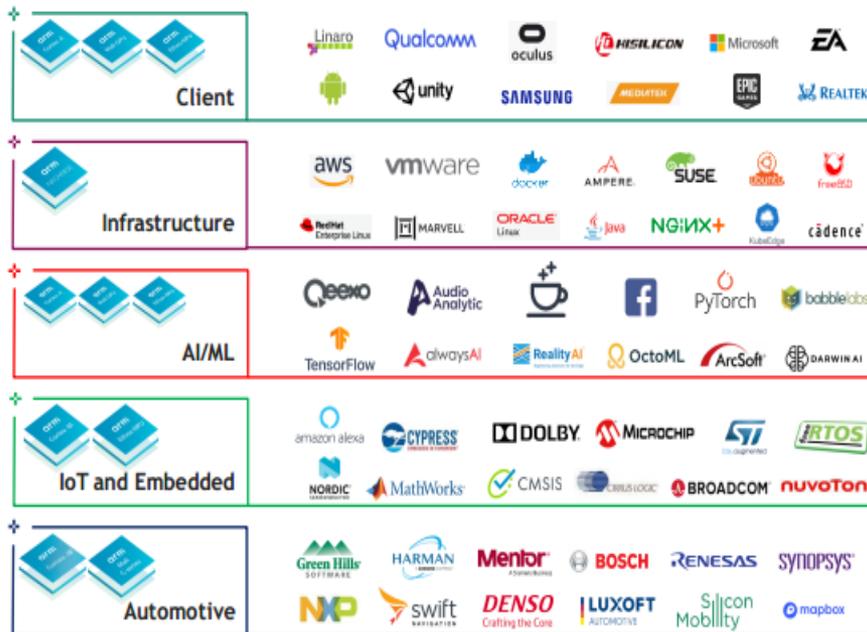
SOC	SOP 时间	适用 场景	内核	算力(TOPS)	功耗 (W)	制程
Parker	2016	L2-L3	NVIDIA Denver *2 ARM Cortex-A57 *4 Parker Pascal iGPU *1	--	15	16nm
Xavier	2020	L2-L5	ARM Carmel CPU *8 Xavier Volta iGPU *1	30	30	12nm
Orin	2021E	L2-L5	ARM Hercules CPU *12 Next-generation iGPU *1	200	45	--
Atlan	2023E	L4-L5	New ARM Grace CPU Next-generation iGPU	1,000	--	--

资料来源：英伟达官网，华西证券研究所

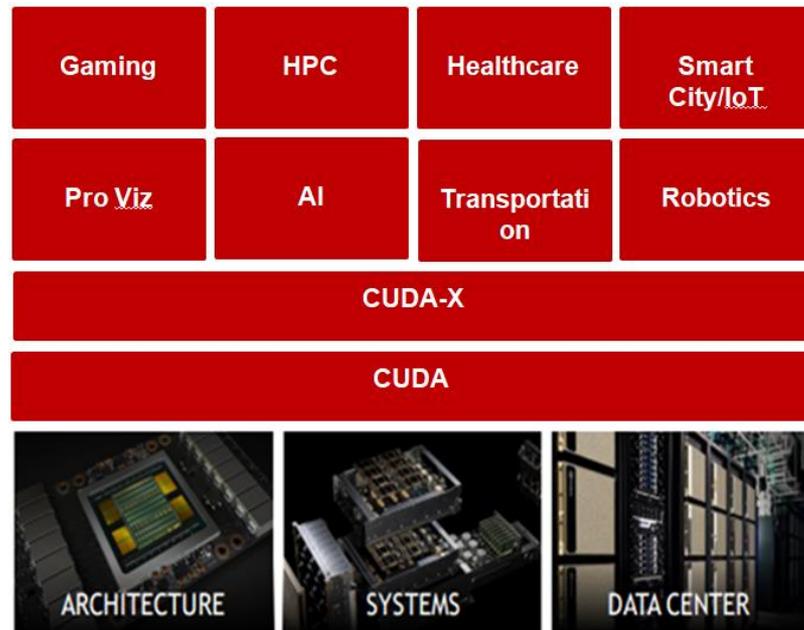
## 3.11 Nvidia : 欲收购Arm创新服务器芯片生态

- ✓ Nvidia的GPU 在数据中心和 workstation 中渲染图像和视频，Arm凭借 Cortex系列产品为CPU设计了蓝图，拥有Mali的GPU系列。
- ✓ 一旦收购整合成功Arm后，Nvidia提供将Arm的CPU与其GPU技术集成的机会，创建一个生态系统的服务器芯片，除了数据中心还可以加速进入消费设备和汽车领域，将在2023年提供2500亿美元的目标市场。

Arm CPU 生态系统



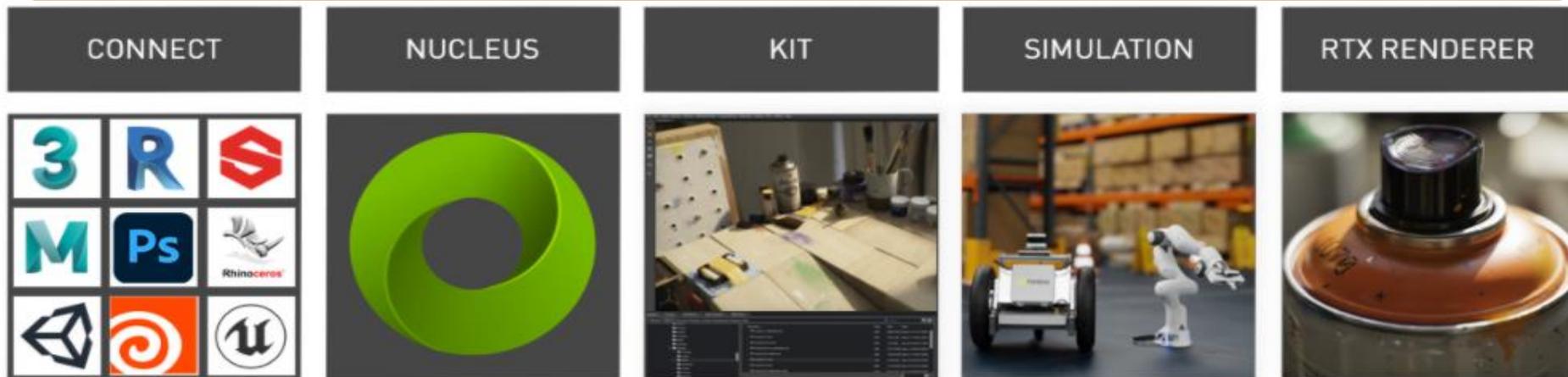
Nvidia 加速计算平台



## 3.12 Nvidia : Omniverse , 软硬件技术的集大成者

- ✓ Omniverse 包含 5 个重要组件，分别是 **Omniverse Connect**、**Nucleus**、**Kit**、**Simulation** 和 **RTX**。这些组件连同所连接的第三方数字内容创作 (DCC) 工具，以及所连接的其他 Omniverse 微服务，共同组成整个 Omniverse 生态系统。
- ✓ **Omniverse Nucleus**：Nucleus 是连接不同位路的用户，实现 3D 资产交互和场景描的数据库引擎。连接以后，负责建模、布局、着色、动画、照明、特效或渲染工作的设计师，可以协作创建场景。Connect 被作为插件分发，使客户端应用程序可以连接到 Nucleus。Kit 是一个用于构建原生 Omniverse 应用和微服务的工具包。

### Omniverse





## 目录

- 1 算力时代，GPU开拓新场景
- 2 GPU 下游三大应用市场
- 3 海外GPU巨头Nvidia
- 4 国产GPU赛道掀起投资热潮
- 5 重点投资机会
- 6 风险提示

## 4.1 AI芯片行业成为融资的热门赛道

- ✓ 2020年国内AI芯片行业投融资金额同比增长了52.8%，2021年1月至4月的投融资事件和金额均已超过去年全年，资本对国内半导体、集成电路领域投资高涨。
- ✓ 从热门领域来看，人工智能领域是2020年资本青睐度较高的细分赛道之一。2020年资本投资的主要是相对成熟且已获得1-2轮甚至2轮以上融资的AI芯片企业。

表 AI芯片行业公司成立时间、融资历史及估值

公司	成立时间	融资历史及公司估值
熠知电子	2017年2月	17年12月获得4.5亿元人民币A轮融资。
深鉴科技	2016年3月	18年7月以3亿美元被赛灵思 (Xilinx) 全资收购。
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。
黑芝麻智能	2017年1月	19年4月获得1亿美元B轮融资，对应估值未披露。
寒武纪	2016年3月	19年9月获得15.5亿元人民币C轮融资。
肇观电子	2016年5月	20年3月获得3亿元人民币B轮融资。
思必驰	2007年10月	20年4月获得4.1亿元人民币E轮融资。
得一微	2017年11月	20年10月获得1亿元人民币B+轮融资。
燧原科技	2018年3月	21年1月获得18亿元人民币C轮融资。
嘉楠科技	2013年4月	21年3月完成1.7美元机构定增。
天数智芯	2015年2月	21年3月完成12亿元人民币C轮融资。
瀚博半导体	2018年12月	21年4月获得5亿元人民币A+轮融资。
云知声	2012年6月	21年6月获得1亿美元D+轮融资。
地平线	2015年7月	21年6月完成15亿美元大C轮融资。
芯驰科技	2018年6月	21年7月获得10亿人民币B轮融资。

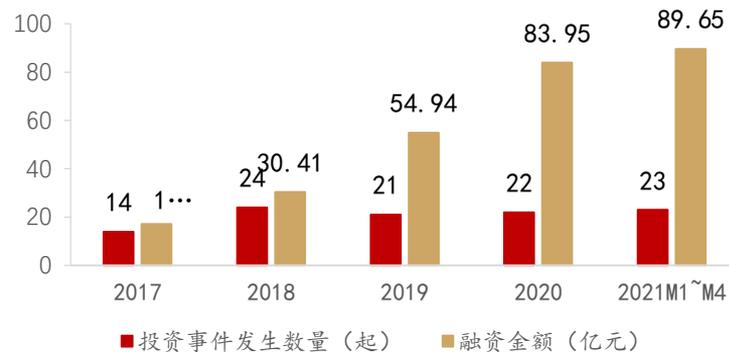
## 4.2 AI芯片行业市场预期逐渐趋于理性，进入市场检验期

- ✓ 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。
- ✓ 未来1-2年，市场将会对各厂商的产品和技术进行实际检验。市场期待更高算力、更低功耗、成本更低的AI芯片。

表 不同公司的芯片介绍

公司名称	芯片名称	芯片类型	芯片算力
燧原科技	邃思2.0	云端AI训练芯片	INT8:320 TOPS TF32:160 TFLOPS
瀚博半导体	博SV100系列	云端AI推理芯片	INT8:200 TOPS
嘉楠科技	勘智K510	云端AI推理芯片	3 TOPS
寒武纪	思元270	云端AI推理芯片	INT8:128 TOPS INT4:256 TOPS
天数智芯	BI	云端AI训练芯片	INT8:295 TOPS
黑芝麻智能	华山二号 A1000 Pro	自动驾驶计算芯片	INT8:106 TOPS INT4:196 TOPS

2017年-2021年AI芯片行业股权投融资事件及金额



资料来源：公开信息，华西证券研究所

## 4.3 沐曦集成电路：多场景高性能GPU

- ✓ 沐曦集成电路专注于设计具有完全自主知识产权，针对异构计算等各类应用的高性能通用GPU芯片。
- ✓ 公司致力于打造国内最强商用GPU芯片，产品主要应用方向包含传统GPU及移动应用，人工智能、云计算、数据中心等高性能异构计算领域，是今后面向社会各个方面通用信息产业提升算力水平的重要基础产品。
- ✓ 拟采用业界最先进的5nm工艺技术，专注研发全兼容CUDA及ROCm生态的国产高性能GPU芯片，满足HPC、数据中心及AI等方面的计算需求。致力于研发生产拥有自主知识产权的、安全可靠的高性能GPU芯片，服务数据中心、云游戏、人工智能等需要高算力的诸多重要领域。



表：沐曦集成电路融资情况

融资轮次	天使轮	pre-A轮	preA+轮
融资时间	2020/11/26	2021/1/18	2021/3/8
融资金额	近亿人民币	数亿人民币	数亿人民币
投资方	和利资本、 泰达科投	红杉资本、 和利资本、 真格基金、 泰达科投	光速中国、 经纬中国、 和利资本、 红杉资本、 真格基金



资料来源：企查查等公开资料整理，华西证券研究所

## 4.4 壁仞科技：推出云端AI芯片

- ✓ 壁仞科技创立于2019年，公司在GPU和DSA（专用加速器）等领域具备丰富的技术储备，聚焦于云端通用智能计算，逐步在AI训练和推理、图形渲染、高性能通用计算等多个领域赶超现有解决方案，以实现国产高端通用智能计算芯片的突破。

图：壁仞科技投资方

部分投资阵容（排名不分先后）



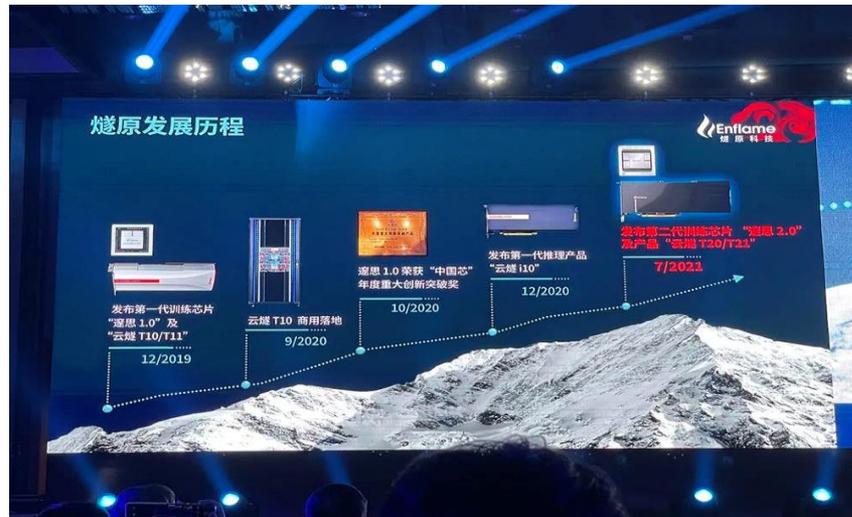
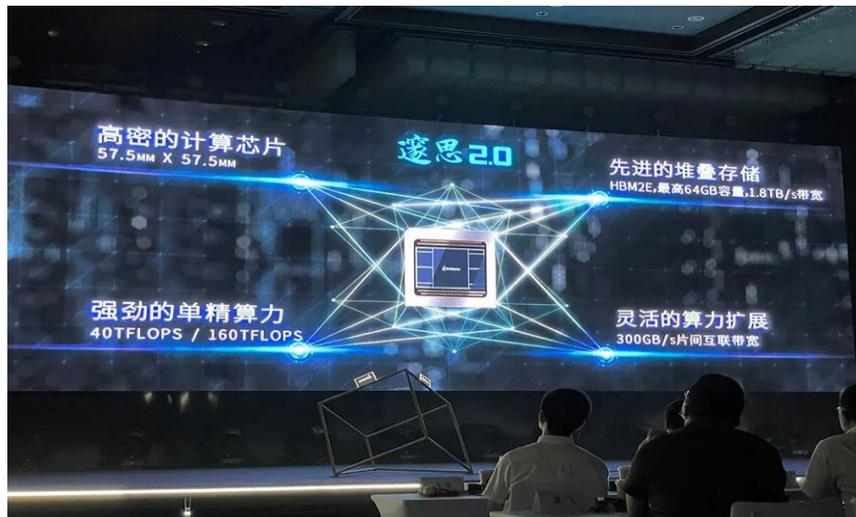
资料来源：华西证券研究所；壁仞科技官网

图：壁仞科技发展历程



## 4.5 燧原科技：推中国最大AI计算芯片

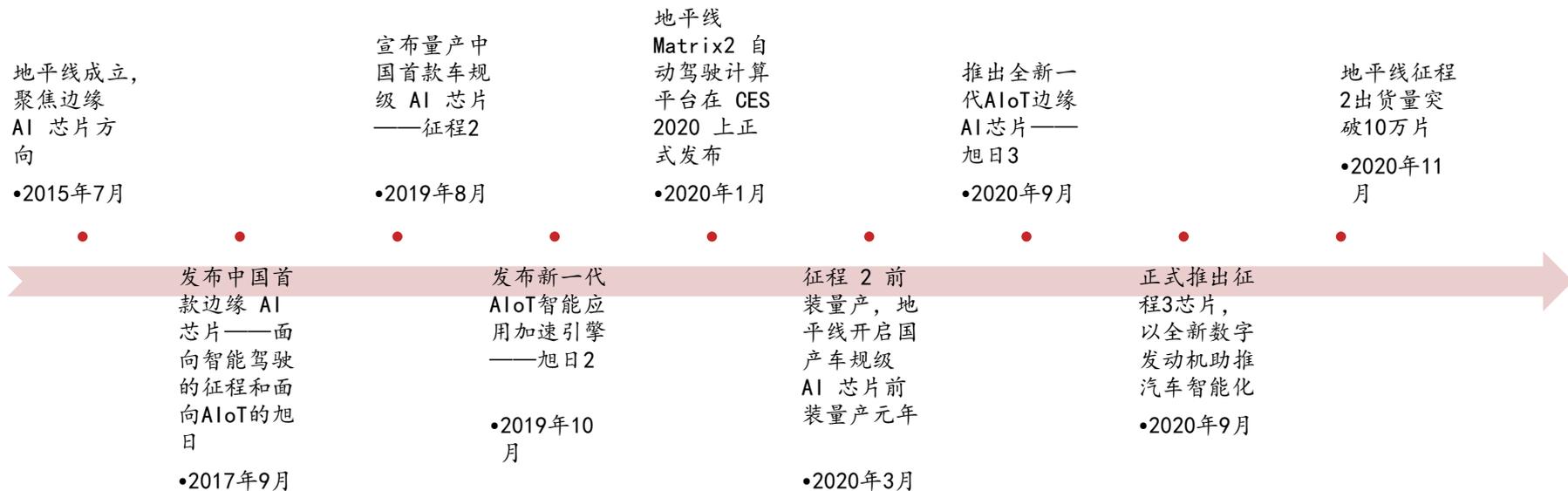
- ✓ 在2021世界人工智能大会期间，上海燧原科技推出第二代云端AI训练芯片邃思2.0及训练产品云燧T20/T21，以及全新升级的驭算Topsrider 2.0软件平台。
- ✓ 邃思2.0是迄今中国最大的AI计算芯片，采用日月光2.5D封装的极限，在国内率先支持TF32精度，单精度张量TF32算力可达160TFLOPS。同时，邃思2.0也是首个支持最先进内存HBM2E的产品。公司主要服务为面向消费电子、汽车电子、计算机及周边、工业、数据处理、物联网等广泛应用市场所提供的一站式芯片定制服务和半导体IP 授权服务。
- ✓ 燧原科技成立于2018年03月19日，成立至今连续获得过5轮融资，累计融资额近32亿元人民币。其最新一笔融资为今年1月完成的18亿元C轮融资，由中信产业基金、中金资本旗下基金、春华资本领投。



## 4.6 地平线：智能驾驶及 AI 应用领域服务

- ✓ 基于创新的人工智能专用计算架构 BPU，地平线已成功流片量产了中国首款边缘人工智能芯片——专注于智能驾驶的征程1 和专注于 AIoT 的旭日1；2019 年，地平线又推出了中国首款车规级 AI 芯片征程 2 和新一代 AIoT 智能应用加速引擎旭日2；2020 年，地平线进一步加速 AI 芯片迭代，推出新一代高效能汽车智能芯片征程 3 和全新一代 AIoT 边缘 AI 芯片平台旭日 3。

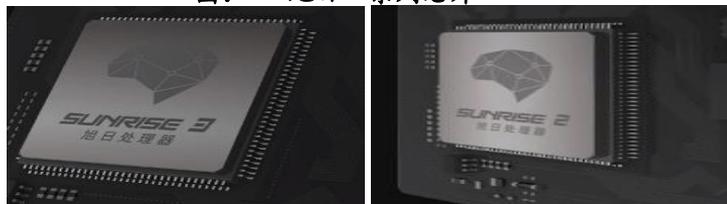
图：地平线发展历程



## 4.7 地平线：AI场景“旭日”系列芯片

- ✓ 智能物联网需求将使云端计算的负荷成倍增长。智能物联网是未来的趋势所向，海量的碎片化场景与计算旭日处理器强大的边缘计算能力，帮助设备高效处理本地数据。
- ✓ 面向AIoT，地平线推出旭日系列边缘 AI 芯片。旭日2采用 BPU 伯努利1.0 架构，可提供 4TOPS 等效算力，旭日3采用伯努利2.0，可提供 5TOPS 的等效算力。

图：“旭日”系列芯片



表：地平线旭日3芯片参数

	旭日3M	旭日3E
CPU	4xA53@1.2G	<a href="#">2xA53@1.2G</a>
BPU	2xBPU@5TOPS	2xBPU@3TOPS
DDR	3200Mbps	2666Mbps
Camera	12M	5M

- ✓ “天工开物”是地平线针对边缘AI产品研发周期长、投入大等特点，专门打造的，从模型训练到芯片部署皆包含在内的“端到端”AI软件解决方案。

图：“天工开物”AI开发平台



## 4.8 地平线：自动驾驶场景“征程”系列芯片

- ✓ 地平线已成为唯一覆盖 L2 到 L4 的全场景整车智能芯片方案提供商。从 2019 年量产中国首款车规级 AI 芯片征程 2，到 2020 年推出第二代车规级芯片征程3。目前，征程 2、征程 3 已在长安、长城、东风岚图、广汽、江淮、理想、奇瑞、上汽等多家自主品牌车企的多款主力爆款车型上实现前装量产。

- ✓ 地平线 Matrix由征程2 架构加速的车规级计算平台，结合深度学习感知技术，为高级别自动驾驶提供了稳定可靠的高性能感知系统。

表：地平线征程系列芯片

	征程2	征程3	征程5
典型算力	4 Tops	5 Tops	96 Tops/128 Tops
典型功耗	2 W	2.5 W	20W/35W
可接入摄像头数量	1 路	6 路	16路
芯片制式	台积电 28 + nm HPC+	16 nm	

现状 芯片已经量产装车，主要负责处理图像数据，支持的功能有两个：DMS 驾驶员监测和 ADAS 辅助驾驶视觉感知信息处理。

支持高级别辅助驾驶、智能座舱、自动泊车辅助、高级别自动驾驶及众包高精地图定位等多种应用场景。

集成自动驾驶和智能交互于一体的全场景整车智能中央计算芯片，征程 5 系列单颗芯片 AI 算力最高可达 128 TOPS。

图：地平线Matrix应用场景



无人的士/小巴



无人物流小车



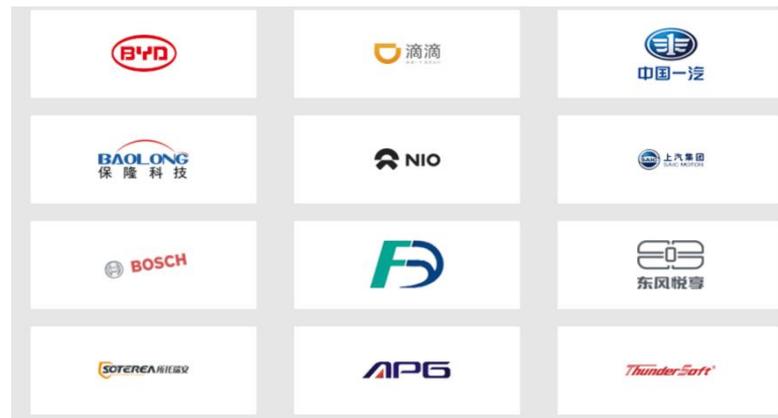
无人卡车

## 4.9 黑芝麻：智能驾驶系统解决方案

- ✓ 黑芝麻智能科技是一家专注于视觉感知技术与自主IP芯片开发的企业。公司主攻领域为嵌入式图像和计算机视觉，提供基于光控技术、图像处理、计算图像以及人工智能的嵌入式视觉感知芯片计算平台，为ADAS及自动驾驶提供完整的商业落地方案。
- ✓ 基于华山二号 A1000 芯片,黑芝麻提供了四种智能驾驶解决方案。单颗 A1000L 芯片适用于 ADAS 辅助驾驶;单颗 A1000 芯片适用于 L2+ 自动驾驶;双 A1000 芯片互联可达 140TOPS 算力,支持 L3 等级自动驾驶;四颗 A1000 芯片则可以支持 L4 甚至以上的自动驾驶需求。另外,黑芝麻还可以根据不同的客户需求,提供定制化服务。
- ✓ 黑芝麻智能首款芯片与上汽的合作已实现量产，第二款芯片A1000正在量产过程中，预计今年下半年在商用车领域实现10万片量级以上的量产，明年将在乘用车领域量产落地。黑芝麻智能已与一汽、蔚来、上汽、比亚迪、博世、滴滴、中科创达、亚太机电等企业在L2、L3级自动驾驶感知系统解决方案上均有合作。

图：黑芝麻智能驾驶解决方案

图：黑芝麻汽车合作商



资料来源：华西证券研究所；中国汽车报等公开资料整理

## 4.10 黑芝麻：自动驾驶场景华山系列芯片

- ✓ 黑芝麻智能科技最新的华山二号(A1000)芯片具备 40-70TOPS 的强大算力、小于 8W 的功耗及优越的算力利用率,工艺制程 16nm,符合 AEC Q-100、单芯片 ASIL B、系统 ASIL D 汽车功能安全要求,是目前能支持 L3 及以上级别自动驾驶的唯一国产芯片。为了应对不同的市场需求,黑芝麻同步发布了华山二号 A1000L。

图：黑芝麻华山系列芯片



表：黑芝麻华山系列芯片

产品名称	发布时间	产品特点
华山一号 A500芯片	2019年8月	国内首款车规级智能驾驶芯片；基于摄像头的汽车传感和人工智能计算为核心，如 Level-2ADAS或 DMS系统。
华山二号 A1000系列芯片	2020年6月	基于摄像头的传感和AI计算为核心，用于 Level-3级自动驾驶系统。
A2000芯片	2021年下半年	预期发布超200TOPS算力的驾驶芯片

表：黑芝麻最新产品A1000系列参数对比

参数	A1000	A1000L
CPU核心数	8	6
车规等级	AECQ-100, ASIL B, ASIL-D安全岛, 系统ASIL D	AECQ-100, ASIL-B
CV算力	DSP + HW	DSP + HW
NN算力	40~70TOPS	16TOPS
功耗	<8w	<5w
传感器输入	GPS, IMU, Camera 4K, 1080PMMW Radar, uSR, Lidar	GPS, IMU, Camera 4K, 1080P, MMW Radar, uSR, Lidar
以太网	Auto GigE	Auto GigE
工艺制程	TSMC 16nm FinFET	TSMC 16nm FinFET

## 4.11 国产GPU最新进展

表 国产GPU最新进展 (2021.11月)

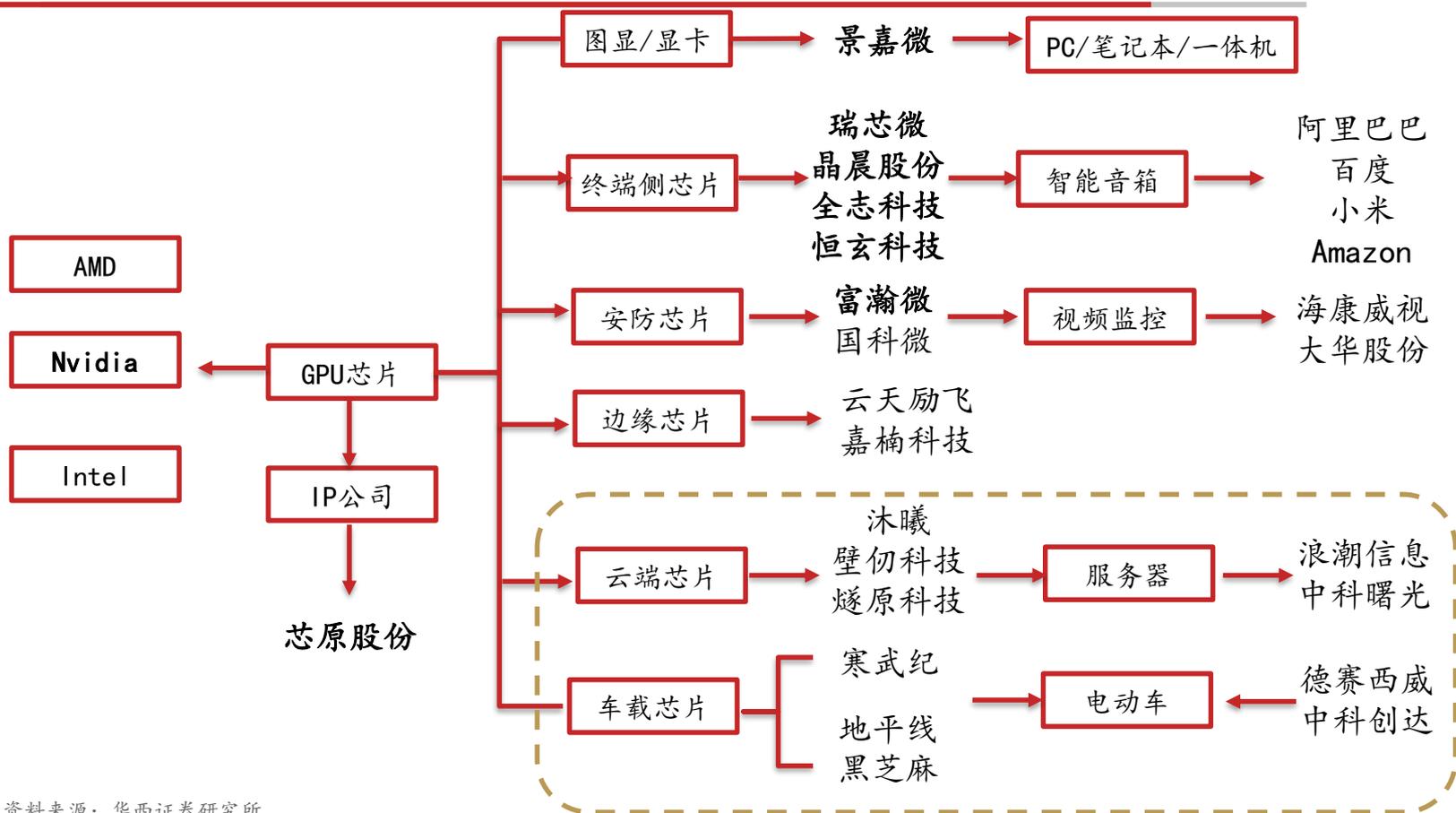
公司名称	时间	产品	描述	融资
壁仞科技	2021.10	首款通用GPU芯片BR100正式交付台积电生产,预计 明年面向市场发布	人工智能(AI)训练和推理、通用运算等众多计算应用场景,可广泛 应用于包括智慧城市、公有云、大数据分析、自动驾驶、医疗健康、生命科学、云游戏等领域。	成立两年, 累计融资50亿元
天数智芯	2021.10	云端7纳米GPGPU产品卡"天垓100"已正式进入量产环节	应用于数据中心、服务器等领域	2021年3月获C轮12亿元融资
景嘉微	2021.11.16	JM9系列已经完成流片、封装、初步测试工作	地理信息系统、媒体处理、CAD辅助设计、游戏、虚拟化等高性能显示需求和人工智能计算需求	
登临科技	2021. 11	Goldwasser系列产品商业化落地。2020年6月Goldwasser系列产品在台积电12nm工艺上 Full Mask 流片成功。	互联网、智慧安防等应用	11月获得新一轮融资
芯动科技	2021.11.17	首款国产高性能服务器级GPU"风华1号"测试成功	搭载全球顶尖的GDDR6X和chiplet 技术,应用于5G数据中心、云游戏、元宇宙等领域	
摩尔线程	2021.11.25	首款国产全功能GPU	内置自主研发的3D图形计算核芯、AI训练与推理计算核芯、高性能并行计算核芯、超高清视频编解码计算等核芯	一年内完成三轮融资, 累计超30 亿元



## 目录

- 1 算力时代，GPU开拓新场景
- 2 GPU 下游三大应用市场
- 3 海外GPU巨头Nvidia
- 4 国产GPU赛道掀起投资热潮
- 5 **重点投资机会**
- 6 风险提示

## 5.1 AI 芯片产业链全图



资料来源：华西证券研究所

## 5.2 国内AI企业聚焦在终端侧推理芯片

表： AI 芯片相关公司及其产品介绍

证券代码	公司	主要产品
688099.SH	晶晨股份	AI 音视频系统终端芯片、智能电视芯片、智能机顶盒芯片
603893.SH	瑞芯微	电源管理芯片RK8XX系列、其他芯片Nano系列、其他芯片RK6XX、RK9XX、RK10XX系列、智能应用处理器芯片RK110X系列、智能应用处理器芯片RK3036系列RK16XX系列、
300458.SZ	全志科技	智能电源管理芯片、智能终端应用处理器芯片
300613.SZ	富瀚微	视频监控多媒体处理芯片、数字接口模块、专业技术服务
688521.SH	芯原股份-U	半导体IP授权服务、一站式芯片定制服务
300474.SZ	景嘉微	图形显控、小型专用化雷达领域核心模块
688256.SH	寒武纪	Cambricon Neuware软件开发平台、寒武纪1A处理器、寒武纪1H处理器、寒武纪1M处理器、思元100 (MLU100) 芯片及云端智能加速卡、思元220 (MLU220) 芯片及边缘智能加速卡、思元270 (MLU270) 芯片及云端智能加速卡、思元290 (MLU290) 芯片及云端智能加速卡
A20645	云天励飞	城市治理AI解决方案、突发疫情防控AI解决方案、云端产品、智慧安防AI解决方案、智慧泛商业AI解决方案、智慧社区解决方案、智慧园区AI解决方案、终端及边缘端产品
CAN.0	嘉楠科技	边缘AI芯片领域



资料来源：Wind，华西证券研究所

## 5.3 建议关注终端+图显+云端+IP 类设计标的

表：重点公司盈利预测

细分行业	证券代码	证券名称	市值 (亿元)	收盘价 (元)	EPS(元/股)				PE			
					2020A	2021E	2022E	2023E	2020A	2021E	2022E	2023E
终端	688099.SH	晶晨股份	491.29	119.5	0.28	1.73	2.46	3.23	427.9	69.1	48.6	37.0
	603893.SH	瑞芯微	626.75	150.37	0.77	1.46	2.19	3.05	195.4	103.0	68.7	49.3
	300458.SZ	全志科技	239.98	72.5	0.62	1.56	2.12	2.71	117.2	46.5	34.2	26.8
	300613.SZ	富瀚微	201.85	167.9	1.10	3.27	4.35	5.56	153.2	51.3	38.6	30.2
云端	688256.SH	寒武纪-U	390.42	97.58	-1.09	-1.95	-2.02	-1.52	-89.9	-50.1	-48.4	-64.3
IP	688521.SH	芯原股份-U	385.84	77.92	-0.05	0.06	0.26	0.51	-1481.4	1393.9	297.7	153.4

资料来源：Wind，华西证券研究所

注：表中公司股价为2021年11月26日收盘价，晶晨股份/瑞芯微/全志科技/富瀚微 盈利预测华西证券研究所预测。

注：寒武纪/芯原股份 盈利预测为Wind的一致预测。

## 5.4 晶晨股份：音视频SOC芯片龙头，加速AIOT市场拓展

- ◆ 公司是智能机顶盒芯片的领导者、智能电视芯片的引领者和 AI 音视频系统终端芯片的开拓者。
- ✓ AI音视频系统终端芯片是晶晨大力发展的产品，2017-2020年营收从0.39亿元增长至3.42亿元，营收占比从2.3%**提升至12.48%**，年复合增长率高达106.2%，具备高成长性。
- ✓ 晶晨股份的AI音视频系统终端芯片主要有智能视频和智能音频两大类，其中智能视频两款产品A311D和A311X基于台积电12nm工艺，而且采用了六核内置神经网络的设计。面向智能音频的A113X/D、S905D2和T962E用于各自终端，其中用于谷歌Home Hub的S905D2 基于台积电12nm工艺，支持远场语音识别。在JBL、Harman和百度等产品上公司的AI音视频系统终端芯片也有应用。
- ✓ A331D: 集成了四核arm Cortex-A73 CPU集群和双核Cortex-A53集群和统一的L2缓存支持INT 8推理的NPU支持所有流行的深度学习框架，包括TensorFlow和Caffe。

公司 AI 音视频系统芯片主要产品

产品系列	产品类别	产品系列	产品描述
A音视频系统终端主要芯片	智能视频 sman vision	A311D	12nm超高性能六核人工智能显示芯片, 内置神经网络处理器
		A311X	12mm超高性能六核人工智能摄像头芯片,内置神经网络处理器, 支持人脸。物体等的实时识别
	智能音频 smant audio	A113X、A113D	高性能四核人工智能语音音箱芯片,支持远场语音识别
		s905D2	12nm智能显示芯片解决方案.支持远场语音识别
		T962E	高性能四核人工智能语音方案形音箱芯片,支持远场语音以及杜比视界,杜比全景声

公司营收和归母净利润及增速 (百万元, %)



## 5.5 瑞芯微：国产AI芯片的先行者，旗舰芯表现亮眼

- ◆ 公司持续深耕AI市场，根据市场研究公司Compass Intelligence 2018年发布的全球人工智能芯片企业排名榜单显示，在国内人工智能芯片领域，瑞芯微已跃升第二。
- ✓ RK3399Pro：瑞芯微首款采用CPU+GPU+NPU硬件结构设计的AI芯片，应用在智能驾驶、图像识别、安防监控、无人机、语音识别。NPU运算性能高达3.0Tops，相较同类NPU芯片性能领先150%；相较GPU作为AI运算单元的大型芯片方案，功耗不到其所需的1%。支持8bit与16bit运算，能够兼容各类AI软件框架。AI接口支持OpenVX及TensorFlowLite/AndroidNN API，AI软件工具支持对Caffe/TensorFlow模型的导入及映射、优化。
- ✓ RK1808：采用22nm FD-SOI工艺，相比主流28nm工艺产品可降低30%，应用于安防、教育、清扫、车载、穿戴、家电、存储。内置2MB系统级SRAM，可实现always-on设备无DDR运行；具有硬件VAD功能，支持低功耗侦听远场唤醒。
- ◆ 发布 RKNN- Toolkit、人工智能软件开发工具包Rock-X SDK，加速落地AI应用。

公司 旗舰产品RK3399 Pro 性能参数



### RK3399Pro

- Dual-core Cortex-A72 up to 1.8GHz  
Quad-core Cortex-A53 up to 1.4GHz
- NPU up to 3.0TOPS
- Mali-T860MP4 GPU
- Dual-channel DDR3/DDR3L/LPDDR3/LPDDR4
- 4K UHD H265/H264/VP9
- HDR10/HLG
- H264 encoder
- Dual MIPI CSI and ISP
- USB Type-C

公司营收和归母净利润及增速（百万元，%）



## 5.6 全志科技：推出多款AI芯片，智能SoC步入高速增长期

- ◆ 公司提供系统级的超大规模数模混合SoC、智能电源管理芯片、无线互联芯片以及相关软硬件的研究与应用技术开发
- ✓ 在智能音箱市场，在原有产品基础上，公司推出高算力新平台，持续为一线互联网标杆客户提供多元化产品合作
- ✓ 在智能扫地机市场，围绕激光、视觉、TOF、双目等传感器品类的扫地机产品应用需求，与客户紧密协同，持续推动新技术的落地应用，保持重要市场地位及市场份额。
- ✓ 在智能家电市场，在家电智能语音方案、智能屏显方案以及IOT方案方面，全面实现国产化替代，进一步巩固了公司在家电智能化的重要市场地位。
- ✓ R329: 2020年3月18日推出，首款搭载Arm中国全新AI处理单元（AIPU）的高算力、低功耗AI语音专用芯片
- ✓ R818: 采用四核 ARM CortexTM-A53，CPU 最高频率 1.6GHz，具有低功耗 CoolFlexTM 电源管理架构

表 公司智能终端应用处理器芯片

产品系列	主要型号产品	产品应用领域
R系列	R16、R329、R329.、R818、MR813	主要市场产品为智能音箱、扫地机器人等
T系列	T3、T7、T5	主要市场产品为行车记录仪、智能后视镜、智能仪表盘等
H系列	H2、H3、H6、H313、H616	主要市场为一体化机顶盒、4K高清机顶盒等
V系列	V3、V526、V533、V536、V831、V833	主要市场产品为泛安防、运动相机等
A系列	A33、A50、A64、A100、A133	主要市场为平板电脑、电子书等
F系列	FIC100S、FIC200S、FIC500S、FIC600、F133	主要市场产品为行车记录仪、智能音箱、看戏机等
其他	VR9、B288、B300、DI	主要市场产品为电子书、无人机、视频一体机等

公司营收和归母净利润及增速（百万元，%）



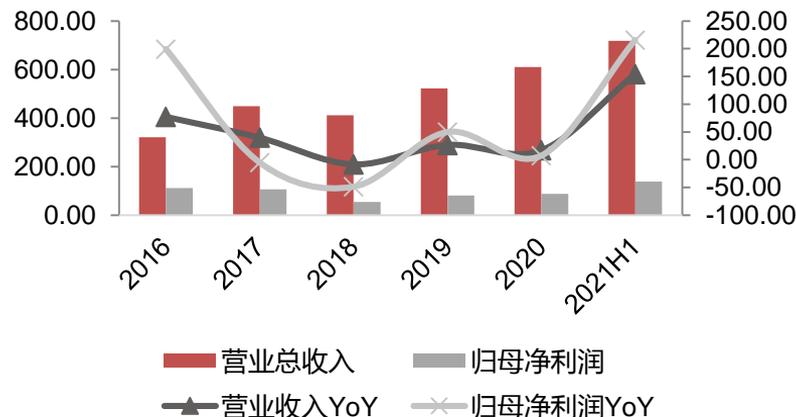
## 5.7 富瀚微：国内安防监控芯片龙头，持续拓展下游应用市场

- ◆ 公司专注于以视频为核心的智慧安防、智慧物联、智能驾驶领域芯片的设计开发，为客户提供高性能视频编解码SoC芯片、图像信号处理器ISP芯片及产品解决方案，以及提供技术开发、IC设计等专业技术服务。
- ✓ ISP芯片：公司开发了基于CMOS传感器的图像信号处理（ISP）类芯片，成为了模拟摄像机市场的主流方案。公司于2018年推出FH8310芯片，最大可支持200万像素，是国内首款百万像素的车规级ISP芯片
- ✓ IPC产品：公司深耕安防芯片多年，是安防行业网络摄像机SoC芯片主要供应商之一，公司IPC产品视频编码技术、像素、存储均领先于市场
- ✓ NVR：公司控股子公司眸芯科技是国内领先的后端SoC厂商，在超大规模SoC芯片设计，高速外设接口及模拟IP设计，低功耗设计等方面经验丰富，市占率高，未来发展前景广阔。

表 公司IPC SoC芯片主要产品及应用

IPC SoC		Camera ISP	
产品型号	应用	产品型号	应用
FH8658	5M高性能智能网络摄像机	FH8536HV200	2M同轴高清摄像机
FH8656	4M高性能智能网络摄像机	FH8322	2M高性能车载前装
FH8652	3M高性能智能网络摄像机	FH8320	960P高性能车载前装
FH8858V200	8M高性能网络摄像机	FH8323	2M高性能车载
FH8856V200	5M高性能网络摄像机	FH8538E	5M同轴高清摄像机
FH8852V200	2M高性能网络摄像机	FH8535	1M同轴高清摄像机
FH8626V100	1080P网络摄像机	FH8310	2M车规级车载
FH8856	5M高性能网络摄像机	FH8556	4K同轴高清摄像机
FH8632	1080P高性能网络摄像机	FH8536H	2M/3M同轴高清摄像机
FH8630M	3M高性能网络摄像机	FH8538M	4M/5M同轴高清摄像机
FH8852	2M高性能网络摄像机	FH8553	3M同轴高清摄像机
FH8830	2M/3M高性能网络摄像机	FH8550M	1080P同轴高清摄像机
FH8610	高性能无线图像信号处理/视频编解码芯片	FH8536E	1080P同轴高清摄像机

公司营收和归母净利润及增速（百万元，%）



资料来源：Wind，华西证券研究所

资料来源：公司官网，华西证券研究所

## 5.8 寒武纪：云、边、端三大场景全面布局

- ◆ 公司的主营业务是应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的研发、设计和销售，为客户提供丰富的芯片产品与系统软件解决方案。公司的主要产品包括**终端智能处理器 IP**、**云端智能芯片及加速卡**、**边缘智能芯片及加速卡**以及与上述产品配套的基础系统软件平台。
- ◆ 经过不断的研发积累，公司产品在行业内赢得高度认可，广泛应用于消费电子、数据中心、云计算等诸多场景。云端智能芯片及加速卡也已应用到国内主流服务器厂商的产品中，并已实现量产出货；边缘智能芯片及加速卡的发布标志着公司已形成全面覆盖云端、边缘端和终端场景的系列化智能芯片产品布局。

表 公司云、边、端三大场景主要芯片

产品类型	寒武纪主要产品	推出时间
终端智能处理器 IP	寒武纪 1A 处理器	2016 年
	寒武纪 1H 处理器	2017 年
	寒武纪 1H 处理器	2018 年
云端智能芯片及加速卡	思元 100 ( MLU100 ) 芯片及云端智能加速卡	2018 年
	思元 270 ( MLU270 ) 芯片及云端智能加速卡	2019 年
	思元 290 ( MLU290 ) 芯片及云端智能加速卡	芯片样品测试中
边缘智能芯片及加速卡	思元 220 ( MLU220 ) 芯片及边缘智能加速卡	2019 年
基础系统软件平台	台 Cambricon Neuware 软件开发平台 ( 适用于公司所有芯片与处理器产品 )	持续研发和升级, 以适配新的芯片

## 5.9 寒武纪：切入自动驾驶芯片、单颗算力200TOPS

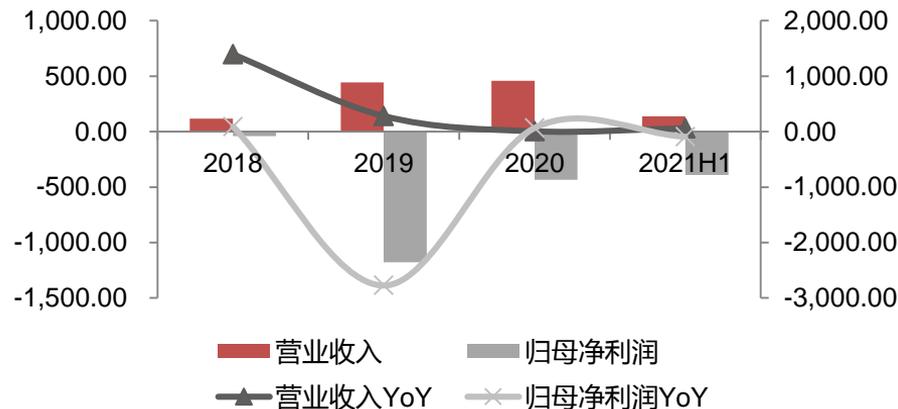
- 在2021世界人工智能大会上，寒武纪创始人兼CEO陈天石透露正在设计一款算力超200TOPS智能驾驶芯片，该芯片继承寒武纪一体化、统一、成熟的软件工具链，采用7nm制程，拥有独立安全岛，符合车规级标准，其定位为“高等级自动驾驶芯片”。
- 在2020年，寒武纪推出了思元290训练芯片和玄思1000智能加速器，补足人工智能训练产品线，标志着寒武纪已初步建立“云边端一体、软硬件协同、训练推理融合”的新生态。向行歌科技增资并引入投资者综合考量了寒武纪中长期发展的战略需求。此次超200TOPS智能驾驶芯片或许就是寒武纪正式切入自动驾驶的开始。

### 公司切入自动驾驶芯片领域



资料来源：腾讯网、华西证券研究所

公司营收和归母净利润及增速（百万元，%）



资料来源：Wind，华西证券研究所

## 5.10 寒武纪：新产品思元 370 性能可比英伟达主流产品

- ◆ 2021年11月3日，寒武纪发布第三代云端AI芯片思元370、基于思元370的两款加速卡MLU370-S4和MLU370-X4、全新升级的Cambricon Neuware软件栈。基于7nm制程工艺，思元370是寒武纪首款采用chiplet（芯粒）技术的AI芯片，集成了390亿个晶体管，最大算力高达256TOPS(INT8)，是寒武纪第二代产品思元270算力的2倍。凭借寒武纪最新智能芯片架构MLUarch03，相较于峰值算力的提升，思元370实测性能表现更为优秀：以ResNet-50为例，MLU370-S4加速卡（半高半长）实测性能为同尺寸主流GPU的2倍；MLU370-X4加速卡（全高全长）实测性能与同尺寸主流GPU相当，能效则大幅领先。

表 寒武纪MLU370系列加速卡与英伟达T4、A10性能对比

	MLU370-S4	MLU370-X4	英伟达 T4	英伟达 A10
时间	2021	2021	2018	2020
制程	7nm	7nm	12nm	7nm
功耗	75W	150W	75W	150W
最大算力	192TOPS	256TOPS	130TOPS ( int8 )	250TOPS ( Int8 )
架构	MLUarch03	MLUarch03	Turing ( 2018 )	Ampere ( 2020 )
内存接口	LPDDR5	LPDDR5	GDDR6 , 320GB/s	GDDR6 , 600GB/s
支持数据类型	INT8、INT16、FP16、BF16、FP32	INT8、INT16、FP16、BF16、FP32	INT4、INT8、INT16、FP32、FP16	INT8、INT16、FP16、BF16、FP32

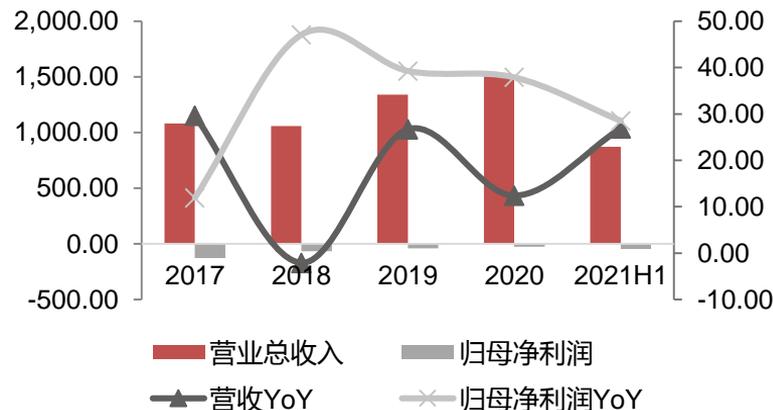
## 5.11 芯原股份：Vivante NPU已被百余款人工智能芯片采用

- ◆ 公司Vivante NPU采用可编程、可扩展、低功耗架构设计，单卷积运算核性能在0.5 TOPS到20 TOPS，进行多卷积运算核扩展后最高算力可达500 TOPS，因此小到可穿戴设备、大到数据中心的计算需求均可满足，还可针对不同芯片尺寸及功耗设定作定制化设计。
- ✓ 图形处理器技术：支持业界主流的嵌入式图形加速标准Vulkan1.0、OpenGL3.2、OpenCL1.2 EP/FP 和OpenVX1.2 等，具有自主可控的指令集及专用编译器，支持每秒2.5 千亿次的浮点运算能力及128 个并行着色器处理单元。
- ✓ 神经网络处理器技术：包括自主可控的卷积神经网络加速、可编程的浮点运算加速、指令集和可编程的浮点运算专用编译器、优化器等工具设计，支持国际标准OpenVX1.2 和OpenCL1.2 EP/FP；支持最大32 位浮点精度数据处理和张量处理的硬件加速；支持0.5TOPs 到6TOPs 性能的单卷积运算核的可扩展架构设计，多卷积运算核扩展后的NPU IP 运算能力可达10TOPs。

公司主要服务内容及应用解决方案



公司营收和归母净利润及增速 (百万元, %)



## 5.12 景嘉微（受益公司）：国产GPU领军企业

- ✓ JM5400打破了国外技术垄断，实现了国产GPU从无到有的突破，并且在性能上做到领先，但其无法满足民用需求，公司在此芯片基础上进一步开发，积极布局民用领域。2018年9月，JM7200完成流片、封装，测试结果符合要求；
- ✓ JM7200的图形处理芯片是在JM5400的基础上升级的，采用了更加先进的28nm工艺制造，增加了片内显存容量，提升了硬件解码能力，支持更高的OpenGL版本，支持更高速的总线接口。图形处理能力和数据传输能力显著提高。JM7200不仅可以满足更高性能的嵌入式系统的要求，还可以用于台式计算机、笔记本电脑等桌面系统的显示要求；
- ✓ JM 7200性能初步满足民用需求，以Nvidia新一代Kepler架构28nm的低端产品GT640作为比较，其核心频率和显存容量稍有优势，但与市场上主流GPU性能差距依然较大。

表 公司产品型号及产品简介

	产品型号	产品简介
国产图形处理器（GPU）芯片	JM5400	JM5400是景嘉微推出的国内首款具有完全自主知识产权的高可靠图形处理芯片，采用全新的架构设计，于2014年4月流片成功。广泛应用于有高可靠性要求的图形生成及显示等领域，满足机载、舰载、车载环境
	JM7200	JM7200采用28nm CMOS工艺，支持4K超高清显示，支持4路独立显示输出，支持十屏同时输出，提供多种丰富的外设接口，可高效完成2D、3D图形加速；支持H.264、VC-1、VP8、MPEG2和MPEG4等格式高清视频硬件解码，将CPU资源占用降至最低；支持对图像及四路多格式的外输入视频进行缩放、旋转、裁剪、叠加、去隔行及融合处理；提供符合OpenGL规范的驱动程序。
	JM7201	采用28nm CMOS工艺，支持4K超高清显示，支持4路独立显示输出，支持多屏同时输出，提供多种丰富的外设接口，可高效完成2D、3D图形加速；支持H.264、VC-1、VP8、MPEG2和MPEG4等格式高清视频硬件解码，运行桌面系统时将CPU资源占用降至最低；提供符合OpenGL规范的驱动程序。
国产高性能显卡	标准MXM显卡	基于JM7201高性能图形处理器的标准MXM结构的显卡，适用于对尺寸高度有严格要求的笔记本电脑、一体机、移动工作站、刀片式主板等场合，具有较高的可靠性，可用于加固领域。
	标准PCIE显卡	基于JM7201高性能图形处理器的标准半高半长尺寸的PCIE接口显卡，能安装在台式机、一体机、服务器等对尺寸高度有较高要求的终端结构内。可用于桌面办公、工业控制等领域。

## 5.13 景嘉微（受益公司）：GPU性能详解

表 景嘉微三款GPU对比

	JM5400	JM7000	JM7200	GT640(DDR3)
DX/OpenGL	OpenGL 1.3	OpenGL 1.5	OpenGL 1.5	OpenGL 4.3 DX12
核心频率	550MHZ	1200MHZ	1200MHZ	900MHZ
主机接口	PCI 2.3	PCI-E 2.0×16	PCI-E 2.0×16	PCI-E 3.0×16
显存位宽	64bit	-	-	128bit
显存频率	DDR3-1600	DDR3-2133	DDR3-2133	DDR3-1800
显存容量	1GB	2-4GB	2-4GB	2GB
渲染能力	2.2GPixels/s	6GPixels/s	-	7.2GPixels/s
浮点性能/GFLOPS	160	500	500	692
工作温度	-55到125℃	-55到125℃	-55到125℃	-
存储温度	-56到125℃	-56到125℃	-56到125℃	-
封装	FCBGA1331	FCBGA1521	FCBGA1521	-
功耗	<6W	<10W	-	50W
工艺	65nm	28nm	28nm	TSMC28nm

- ✓ 已完成与龙芯、飞腾、麒麟软件、统信软件、道、天脉等国内主要的CPU和操作系统厂商的适配工作。
- ✓ 与中国长城、超越电子等十余家国内主要计算机整机厂商建立合作关系并进行产品测试。
- ✓ 与麒麟、长城、苍穹、宝德、超图、昆仑、中科方德、中科可控、宁美等多家软硬件厂商进行互相认证。

公司营收和归母净利润及增速（百万元，%）



## 5.14 云天励飞（受益公司）：专注于计算机视觉领域，具有自主知识产权的神经网络处理器芯片

- ◆ 2017年，公司第一代具有自主知识产权的神经网络处理器采用FPGA实现，依托“深目”系统，已经在云天励飞DeepEye200 PCIe FPGA加速卡上以及IFBOX边缘计算盒子上应用，主要用于目标识别特征提取。
- ◆ 2018年，公司第二代具有自主知识产权的神经网络处理器芯片采用**22nm工艺投片**，该芯片主要面向**嵌入式前端和边缘计算应用**，具备高性能、低功耗、低成本的优点，通过算法芯片化能力，定义了高效、灵活的神经网络指令集、处理器计算架构和工具链，并兼顾了业界算法的通用性。该芯片**搭载于终端和边缘端设备，可实现视频数据的结构化处理**，从而降低网络带宽要求，提高数据处理的质量和效率，减少后台处理成本。该芯片获得2018年度中国行业信息化融合发展论坛“中国芯片最佳技术创新奖”等奖项。
- ✓ 从2018年至2020年，公司分别实现营业收入1.33亿元、2.3亿元和4.26亿元，2020年营收同比2018年增长220.3%。
- ✓ 从2018年至2020年，公司的净利润分别为-1.99亿元、-5.1亿元和-9.42亿元。研发费用分别为1.45亿元、2亿元和2.19亿元。

表 公司的第二代芯片对比其他类型芯片所具备的优势

对比芯片	优势
相对GPU/CPU	具备同等的指令集灵活性，执行效率、能量效率方面相比CPU、GPU有102个数量级的优势
相对DSP	具有执行效率高、功耗低的特点
相对ASIC	具有后向算法可编程的灵活性，更适合深度学习AI算法的演进和迭代部署
相对FPGA	具有高性能、低成本的优势

表 公司的第二代芯片主要应用方向

应用方向	细分领域
视觉端智能以及边缘智能应用	单芯片实现8路视频结构化，多芯片级联实现64路视频结构化分析引擎
	智能相机、便携式警务终端、智能商显、机器人、无人机、无人车、工业智能检测、新零售等领域

## 5.15 嘉楠科技（受益公司）：推出边缘AI芯片领域

- ◆ **2018年9月6日**，嘉楠科技推出智能计算芯片勘智K210。嘉楠成为行业内第一个交付基于RISC-V架构和自研神经网络加速器KPU的商用边缘智能芯片的公司。该芯片依托于完全的自主研发的AI神经网络加速器KPU，具备自主IP、视听兼具与可编程能力大三特点，能够充分适配多个业务场景的需求。算力方面：K210可在0.3W的条件下提供1TOP的算力支持，充分适配多数业务场景中低功耗约束下的算力需求；芯片集成度方面：K210采用视听一体化设计。
- ◆ **2021年7月8日**，嘉楠科技正式发布AI芯片勘智K510。该芯片定位于中高端边缘推理市场，在总线架构、IP核心与视频子系统推出全新设计，相比第一代芯片算力提升3倍。IP核上：K510搭载自主研发IP核的升级版KPU2.0，集中突破AI芯片设计中的“存储墙”和“性能墙”的问题；外设上：K510在外设方面进行了升级，视觉能力大幅提升；算法上：K510的编译器和软件上联合优化带宽需求，对算法进行量化压缩；软件上：K510支持CNN、RNN等各类网络模型算子，支持裸机、嵌入式RTOS和Linux，并移植了向量计算和数据处理操作，支持TensorFlow、PyTorch、ONNX和TVM等主流深度学习框架。

嘉楠科技勘智AI系列芯片



### 勘智K210

勘智K210采用RISC-V处理器架构，具备视听一体、自主IP核与可编程能力强三大特点，支持机器视觉与机器听觉多模态识别，可广泛应用于智能家居、智能园区、智能能耗和智能农业等场景。



#### 处理器架构

RISC-V双核64位CPU



#### 算力

1TOPS



#### 人脸检测

60帧/秒



#### 功耗

300mW



### 勘智K510

勘智K510是一款定位于中高端边缘测应用市场的推理芯片，为AI应用提供高性能图像及语音处理能力。搭载自主研发的第二代神经网络处理器KPU2.0，采用独创计算数据流技术，算力提升3倍同时降低功耗。



#### 开源架构

双核RISC-V 64位CPU



#### 算力

2.5TFLOPS



#### 内存强悍

支持最高LPDDR3/4 x32bit  
容量16Gbits



#### 开发友好

支持TensorFlow多  
数算子

## 6 风险提示

- ◆ **产能不及预期的风险**：GPU生产需要14nm 及其以下的制程，如果产能达不到要求，产品流片存在一定风险；
- ◆ **国产替代不及预期**：目前国产GPU正处于验证和优化之中，存在国产替代不及预期的风险；
- ◆ **国产GPU生态建设不及预期**：国产GPU的生态落后于海外公司，如果生态建设不及预期，会导致产品差距拉大。

- 1、芯时代之一\_半导体重磅深度《新兴技术共振进口替代，迎来全产业链投资机会》
- 2、芯时代之二\_深度纪要《国产芯投资机会暨权威专家电话会》
- 3、芯时代之三\_深度纪要《半导体分析和投资策略电话会》
- 4、芯时代之四\_市场首篇模拟IC深度《下游应用增量不断，模拟IC加速发展》
- 5、芯时代之五\_存储器《存储产业链战略升级，开启国产替代“芯”篇章》
- 6、芯时代之六\_功率半导体《功率半导体处黄金赛道，迎进口替代良机》
- 7、芯时代之七\_半导体材料《铸行业发展基石，迎进口替代契机》
- 8、芯时代之八\_深度纪要《功率半导体重磅专家交流电话会》
- 9、芯时代之九\_半导体设备深度《进口替代促景气度提升，设备长期发展明朗》
- 10、芯时代之十\_3D/新器件《先进封装和新器件，续写集成电路新篇章》
- 11、芯时代之十一\_IC载板和SLP《IC载板及SLP，集成提升的板级贡献》
- 12、芯时代之十二\_智能处理器《人工智能助力，国产芯有望“换”道超车》
- 13、芯时代之十三\_封测《先进封装大势所趋，国家战略助推成长》
- 14、芯时代之十四\_大硅片《供需缺口持续，国产化蓄势待发》
- 15、芯时代之十五\_化合物《下一代半导体材料，5G助力市场成长》
- 16、芯时代之十六\_制造《国产替代加速，拉动全产业链发展》
- 17、芯时代之十七\_北方华创《双结构化持建机遇，由大做强倍显张力》
- 18、芯时代之十八\_斯达半导《铸IGBT功率基石，创多领域市场契机》
- 19、芯时代之十九\_功率半导体②《产业链逐步成熟，功率器件迎黄金发展期》
- 20、芯时代之二十\_汇顶科技《光电传感创新领跑，多维布局引领未来》

- 21、芯时代之二十一\_华润微《功率半导专芯致志，特色工艺术业专攻》
- 22、芯时代之二十二\_大硅片\*重磅深度《半导体材料第一蓝海，硅片融合工艺创新》
- 23、芯时代之二十三\_卓胜微《5G赛道射频芯片龙头，国产替代正当时》
- 24、芯时代之二十四\_沪硅产业《硅片“芯”材蓄势待发，商用量产空间广阔》
- 25、芯时代之二十五\_韦尔股份《光电传感稳创领先，系统方案展创宏图》
- 26、芯时代之二十六\_中环股份《半导硅片厚积薄发，特有赛道独树一帜》
- 27、芯时代之二十七\_射频芯片《射频芯片千亿空间，国产替代曙光乍现》
- 28、芯时代之二十八\_中芯国际《代工龙头创领升级，产业联动芯火燎原》
- 29、芯时代之二十九\_寒武纪《AI芯片国内龙头，高研发投入前景可期》
- 30、芯时代之三十\_芯朋微《国产电源IC十年磨一剑，铸就国内升级替代》
- 31、芯时代之三十一\_射频PA《射频PA革新不止，万物互联广袤无限》
- 32、芯时代之三十二\_中微公司《国内半导刻蚀巨头，迈内生&外延平台化》
- 33、芯时代之三十三\_芯原股份《国内IP龙头厂商，推动SiPaaS模式发展》
- 34、芯时代之三十四\_模拟IC深度PPT《模拟IC黄金赛道，本土配套渐入佳境》
- 35、芯时代之三十五\_芯海科技《高精度测量ADC+MCU+AI,切入蓝海赛道超芯星》
- 36、芯时代之三十六\_功率&化合物深度《扩容&替代提速，化合物布局长远》
- 37、芯时代之三十七\_恒玄科技《专注智能音频SoC芯片，迎行业风口快速发展》
- 38、芯时代之三十八\_和而泰《从高端到更高端，芯平台创新格局》
- 39、芯时代之三十九\_家电芯深度PPT《家电芯配套渐完善,增存量机遇筑蓝海》
- 40、芯时代之四十\_前道设备PPT深度《2021年国产前道设备，再迎新黄金时代》

- 41、芯时代之四十一\_力芯微《专注电源管理芯片，内生外延拓展产品线》
- 42、芯时代之四十二\_复旦微电《国产FPGA领先企业，高技术壁垒铸就护城河》
- 43、芯时代之四十三\_显示驱动芯片深度PPT《显示驱动芯—面板国产化最后1公里》
- 44、芯时代之四十四\_艾为电子《数模混合设计专家，持续迭代拓展产品线》
- 45、芯时代之四十五\_紫光国微《特种与安全两翼齐飞，公司步入快速发展阶段》
- 46、芯时代之四十六\_新能源芯\*PPT深度《乘碳中和之风，基础元件腾飞》
- 47、芯时代之四十七\_AIoT \*PPT深度《AIoT大时代，SoC厂商加速发展》
- 48、芯时代之四十八\_铂科新材《双碳助力公司发展，GPU等新应用构建二次成长曲线》
- 49、芯时代之四十九\_AI芯片《AI领强算力时代，GPU启新场景落地》

## 分析师简介

**孙远峰：**华西证券研究所副所长&电子行业首席分析师，哈尔滨工业大学工学学士，清华大学工学博士，近3年电子实业工作经验；2018年新财富上榜分析师（第3名），2017年新财富入围/水晶球上榜分析师，2016年新财富上榜分析师（第5名），2013~2015年新财富上榜分析师团队核心成员；多次获得保险资管IAMAC、水晶球、金牛等奖项最佳分析师；清华大学校友总会电子系分会理事会副秘书长；2019年6月加入华西证券研究所

**熊军：**华西证券研究所电子行业分析师，东南大学集成电路专业硕士，英伟达、赛迪顾问等实业工作经验；2019年中国证券分析师金翼奖通信行业第一名，2019年11月加入华西证券研究所。

### 分析师承诺

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

### 评级说明

公司评级标准	投资评级	说明
以报告发布日后的6个月内公司股价相对上证指数的涨跌幅为基准。	买入	分析师预测在此期间股价相对强于上证指数达到或超过15%
	增持	分析师预测在此期间股价相对强于上证指数在5%—15%之间
	中性	分析师预测在此期间股价相对上证指数在-5%—5%之间
	减持	分析师预测在此期间股价相对弱于上证指数5%—15%之间
	卖出	分析师预测在此期间股价相对弱于上证指数达到或超过15%

### 行业评级标准

以报告发布日后的6个月内行业指数的涨跌幅为基准。	推荐	分析师预测在此期间行业指数相对强于上证指数达到或超过10%
	中性	分析师预测在此期间行业指数相对上证指数在-10%—10%之间
	回避	分析师预测在此期间行业指数相对弱于上证指数达到或超过10%

### 华西证券研究所：

地址：北京市西城区太平桥大街丰汇园11号丰汇时代大厦南座5层

网址：<http://www.hx168.com.cn/hxzq/hxindex.html>

华西证券股份有限公司（以下简称“本公司”）具备证券投资咨询业务资格。本报告仅供本公司签约客户使用。本公司不会因接收人收到或者经由其他渠道转发收到本报告而直接视其为本公司客户。

本报告基于本公司研究所及其研究人员认为的已经公开的资料或者研究人员的实地调研资料，但本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载资料、意见以及推测仅于本报告发布当日的判断，且这种判断受到研究方法、研究依据等多方面的制约。在不同时期，本公司可发出与本报告所载资料、意见及预测不一致的报告。本公司不保证本报告所含信息始终保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者需自行关注相应更新或修改。

在任何情况下，本报告仅提供给签约客户参考使用，任何信息或所表述的意见绝不构成对任何人的投资建议。

市场有风险，投资需谨慎。投资者不应将本报告视为做出投资决策的惟一参考因素，亦不应认为本报告可以取代自己的判断。在任何情况下，本报告均未考虑到个别客户的特殊投资目标、财务状况或需求，不能作为客户进行客户买卖、认购证券或者其他金融工具的保证或邀请。在任何情况下，本公司、本公司员工或者其他关联方均不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告而导致的任何可能损失负有任何责任。投资者因使用本公司研究报告做出的任何投资决策均是独立行为，与本公司、本公司员工及其他关联方无关。

本公司建立起信息隔离墙制度、跨墙制度来规范管理跨部门、跨关联机构之间的信息流动。务请投资者注意，在法律许可的前提下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的前提下，本公司的董事、高级职员或员工可能担任本报告所提到的公司的董事。

所有报告版权均归本公司所有。未经本公司事先书面授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容，如需引用、刊发或转载本报告，需注明出处为华西证券研究所，且不得对本报告进行任何有悖原意的引用、删节和修改。