

华金证券电子团队一走进“芯”时代系列深度之六十“AI算力GPU”

AI产业化再加速，智能大时代已开启

——GPU行业深度报告

孙远峰/王臣复/王海维

SAC执业证书编号：

S0910522120001/S0910523020006/S0910523020005

2023年3月26日



本报告仅供华金证券客户中的专业投资者参考
请仔细阅读在本报告尾部的重要法律声明

- 在芯片算力快速提升、日趋庞大的数据量共同支撑下，AI算法迭代升级加速。AI的发展经历了很长时间的积累，其能不断跨越科学与应用之间的鸿沟主要得益于技术突破、行业落地、产业协作等多方面的推动，而技术突破是其中最为关键的要素。从起步阶段发展到当下深度学习阶段，算法、数据和算力构成了AI三大基本要素，并共同推动AI向更高层次的感知和认知发展。算法方面，目前深度学习仍然是AI技术发展的主导路线，但是早期所使用的有监督学习方式由于受限于对大量标注数据依赖与理解能力缺乏，而且模型通用性较差，正逐步被新的技术所取代，在芯片算力的快速提升、日益庞大的数据量这两者的支撑下，新算法正处于加速迭代升级过程中。
- 自监督学习的算法模型快速发展，“预训练+精调”的开发范式迈向成熟，新一轮AI技术产业化之路开启。谷歌、脸书等多家企业先后发布使用自监督学习的算法模型，通过挖掘无标注数据的监督信息，减少人为干预。现阶段自监督学习本质上仍依赖规范化、标签化的数据，主要借助预训练模型构筑并学习数据特征。“预训练”的做法一般是将大量低成本收集的训练数据放在一起，经过某种预训方法去学习其中的共性，然后将其中的共性“移植”到特定任务的模型中，再使用相关特定领域的少量标注数据进行“微调”，这样的话，模型只需要从“共性”出发，去“学习”该特定任务的“特殊”部分即可。预训练模型成功的关键是自监督学习与Transformer的结合。预训练大模型在海量数据的学习训练后具有良好的通用性和泛化性，用户基于大模型通过零样本、小样本学习即可获得领先的效果，同时“预训练+精调”等开发范式，让研发过程更加标准化，显著降低了人工智能应用门槛。整体上来看，关于本轮AI技术突破所带来的产业化变局，我们三个核心观点：1、基于GPT为代表的大模型AI的通用能力，未来几年大模型AI的渗透广度、深度和速度有可能会超预期；2、ChatGPT采用的是闭源模型，其加速的产业落地会刺激更多的厂商加大大模型AI的研发投入，进而推动AI产业化发展；3、大模型AI通用能力的提升，带动的将不仅仅是云计算市场的增长，伴随着多种技术与商业化路径的逐步成熟，云、边缘、端的增量市场空间均有望渐次打开。

核心观点（2）

- **云端计算进入高性能计算时代，大模型训练仍以GPU为主。**虽然AI芯片目前看有GPU、ASIC、CPU、FPGA等几大类，但是基于几点原因，我们判断GPU仍将是训练模型的主流硬件：1、Transformer架构是最近几年的主流，该架构最大的特点之一就是能够利用分布式GPU进行并行训练，提升模型训练效率；2、ASIC的算力与功耗虽然看似有优势，但考虑到AI算法还是处于一个不断发展演进的过程，用专用芯片部署会面临着未来算法更迭导致芯片不适配的巨大风险；3、英伟达强大的芯片支撑、生态、算法开源支持。
- **模型小型化技术逐步成熟，从训练走向推理，云、边、端全维度发展。**我们认为至少有四大投资主线应持续关注：1、GPU方面，在英伟达的推动下，其从最初的显卡发展到如今的高性能并行计算，海外大厂已经具备了超过20年的技术、资本、生态、人才等储备，形成了大量的核心技术专利，而且也能充分享有全球半导体产业链的支撑，这都或是目前国内厂商所缺失的。近几年在资本的推动下，国内涌现出数十家GPU厂商，各自或都具备一定的發展基础，但整体经营时间较短，无论从技术积淀、产品料号布局、高端料号性能来说，与国外大厂仍具备较大差距。但国产化势在必行，国内相关产业链重点环节也积极对上游芯片原厂进行扶持，国产算力芯片需要不断迭代以实现性能的向上提升，后续持续关注相关厂商料号升级、生态建设和客户突破；2、AI在端侧设备应用普及是大势所趋，目前，知识蒸馏、剪枝、量化等模型小型化技术在逐步成熟，AI在云、边、端全方位发展的时代已至。除了更加广泛的应用带来需求量的提升外，更复杂算法带来更大算力的需求也将从另一个维度推动市场扩容；3、数据的高吞吐量需要大带宽的传输支持，光通信技术作为算力产业发展的支撑底座，具备长期投资价值；4、Chiplet技术可以突破单一芯片的性能和良率等瓶颈，降低芯片设计的复杂度和成本。基于向Chiplet模式的设计转型，已经是大型芯片厂商的共识，相关产业链具备长期投资价值。
- **建议关注：瑞芯微、晶晨股份、星宸科技（待上市）、全志科技、北京君正、中科蓝讯、富瀚微、恒玄科技**
- **风险提示：技术创新风险、宏观经济和行业波动风险、国际贸易摩擦风险。**

- 01 由专用走向通用，GPU赛道壁垒高筑
- 02 产业化路径显现，全球AI竞赛再加速
- 03 全维智能化大时代，国产算力行则必至
- 04 建议关注
- 05 产业相关
- 06 风险提示

由专用走向通用，GPU赛道壁垒高筑

- 1.1 什么是GPU
- 1.2 始于图形处理设备
- 1.3 浮点计算能力与可编程性结合
- 1.4 GPU发展三大方向
- 1.5 英伟达显卡发展历程
- 1.6 GeForce RTX 40系列，时代最强
- 1.7 英特尔的核显
- 1.8 核显与独显性能对比
- 1.9 图形流水线是GPU工作的通用模型
- 1.10 统一渲染架构的推出开启了通用计算大时代
- 1.11 从简单到越来越复杂的流水线
- 1.12 光线追踪时代开启
- 1.13 光线追踪算法要求的计算量巨大
- 1.14 走向新场景的GPGPU
- 1.15 GPU与GPGPU的对比
- 1.16 GPGPU与CPU的对比
- 1.17 并行计算发展的核心
- 1.18 SIMT，主流GPU的系统架构核心
- 1.19 GPGPU架构，以A100为例
- 1.20 Fermi是第一个完整的GPU计算架构
- 1.21 通用算力提升是英伟达GPU架构演进的重点之一
- 1.22 多方面构建的高壁垒
- 1.23 人才与研发投入，以英伟达为例
- 1.24 国外厂商多年间构筑了庞大的专利池
- 1.25 英伟达全栈布局构筑强大生态
- 1.26 走向异构，海外厂商横向布局不断

产业化路径显现，全球AI竞赛再加速

- 2.1 AI技术赋能实体经济面临的瓶颈
- 2.2 ChatGPT的破圈
- 2.3 ChatGPT的成功离不开预训练大模型
- 2.4 预训练模型的发展历程
- 2.5 Transformer架构成主流
- 2.6 自监督学习与Transformer的结合
- 2.7 大模型的突现能力
- 2.8 参数量爆发式增长的ChatGPT
- 2.9 预训练大模型，第三波AI发展的重大拐点
- 2.10 生成式AI、边缘AI技术即将步入成熟期
- 2.11 大模型是大算力和强算法结合的产物
- 2.12 AI芯片三剑客
- 2.13 训练端GPU担纲
- 2.14 数据中心迈入“高算力”时代，兵家必争
- 2.15 英伟达数据中心业务快速增长
- 2.16 自动驾驶研发两大商业路线
- 2.17 自动驾驶实现的两种技术路线
- 2.18 单车智能化推动算力升级加速
- 2.19 自动驾驶具备广阔市场前景

全维智能化大时代，国产算力行则必至

- 3.1 全球数据中心负载任务量快速增长
- 3.2 全球计算产业投资空间巨大
- 3.3 预训练大模型对于GPU的需求
- 3.4 国内市场需求将保持高增长
- 3.5 云计算及云部署方式
- 3.6 不同云部署方式的市场占比
- 3.7 企业上云持续向细分行业渗透
- 3.8 从“资源上云”迈入“深度用云”
- 3.9 信创从试点走向推广
- 3.10 公有云主要参与厂商
- 3.11 云计算产业链
- 3.12 集成显卡与独立显卡市场份额
- 3.13 独立显卡英伟达一家独大
- 3.14 性能强大的H100
- 3.15 国产厂商两条发展路径：GPU和GPGPU
- 3.16 先求有，再求好
- 3.17 生态先兼容主流，未来将走向自建
- 3.18 国产之路已开启，部分国产GPU设计厂商列表
- 3.19 GPU发展离不开全球产业链的支撑
- 3.20 制程升级对于算力芯片性能提升具有较高贡献度
- 3.21 摩尔定律发展趋缓
- 3.22 Chiplet技术潜力大
- 3.23 Chiplet技术发展历程
- 3.24 行业巨头推动，产业加速落地
- 3.25 采用Chiplet技术的产品不断出现
- 3.26 算力两大演进方向：更大算力&更多样化应用
- 3.27 存量替代与增量成长并存
- 3.28 高吞吐量离不开高速传输
- 3.29 光通信前景可期

分目录（4）

04 建议关注

- 4.1 瑞芯微
- 4.2 晶晨股份
- 4.3 星宸科技（待上市）
- 4.4 全志科技
- 4.5 北京君正
- 4.6 中科蓝讯
- 4.7 富瀚微
- 4.8 恒玄科技

06 风险提示

- 技术创新风险
- 宏观经济和行业波动风险
- 国际贸易摩擦风险

05 产业相关

- 5.1 海光信息
- 5.2 龙芯中科
- 5.3 景嘉微
- 5.4 寒武纪-U
- 5.5 中芯国际
- 5.6 芯原股份-U
- 5.7 华大九天
- 5.8 概伦电子
- 5.9 长电科技
- 5.10 华天科技
- 5.11 通富微电
- 5.12 炬芯科技
- 5.13 源杰科技
- 5.14 光迅科技
- 5.15 摩尔线程（未上市）

01

由专用走向通用，GPU赛道壁垒高筑

- 1.1 什么是GPU
- 1.2 始于图形处理设备
- 1.3 浮点计算能力与可编程性结合
- 1.4 GPU发展三大方向
- 1.5 英传达显卡发展历程
- 1.6 GeForce RTX 40系列，时代最强
- 1.7 英特尔的核显
- 1.8 核显与独显性能对比
- 1.9 图形流水线是GPU工作的通用模型
- 1.10 统一渲染架构的推出开启了通用计算大时代
- 1.11 从简单到越来越复杂的流水线
- 1.12 光线追踪时代开启
- 1.13 光线追踪算法要求的计算量巨大
- 1.14 走向新场景的GPGPU
- 1.15 GPU与GPGPU的对比
- 1.16 GPGPU与CPU的对比
- 1.17 并行计算发展的核心
- 1.18 SIMT，主流GPU的系统架构核心
- 1.19 GPGPU架构，以A100为例
- 1.20 Fermi是第一个完整的GPU计算架构
- 1.21 通用算力提升是英伟达GPU架构演进的重点之一
- 1.22 多方面构建的高壁垒
- 1.23 人才与研发投入，以英伟达为例
- 1.24 国外厂商多年间构筑了庞大的专利池
- 1.25 英伟达全栈布局构筑强大生态
- 1.26 走向异构，海外厂商横向布局不断

02

产业化路径显现，全球AI竞赛再加速

03

全维智能化大时代，国产算力行则必至

04

建议关注

05

产业相关

06

风险提示

1. 由专用走向通用，GPU赛道壁垒高筑

1.1 什么是GPU

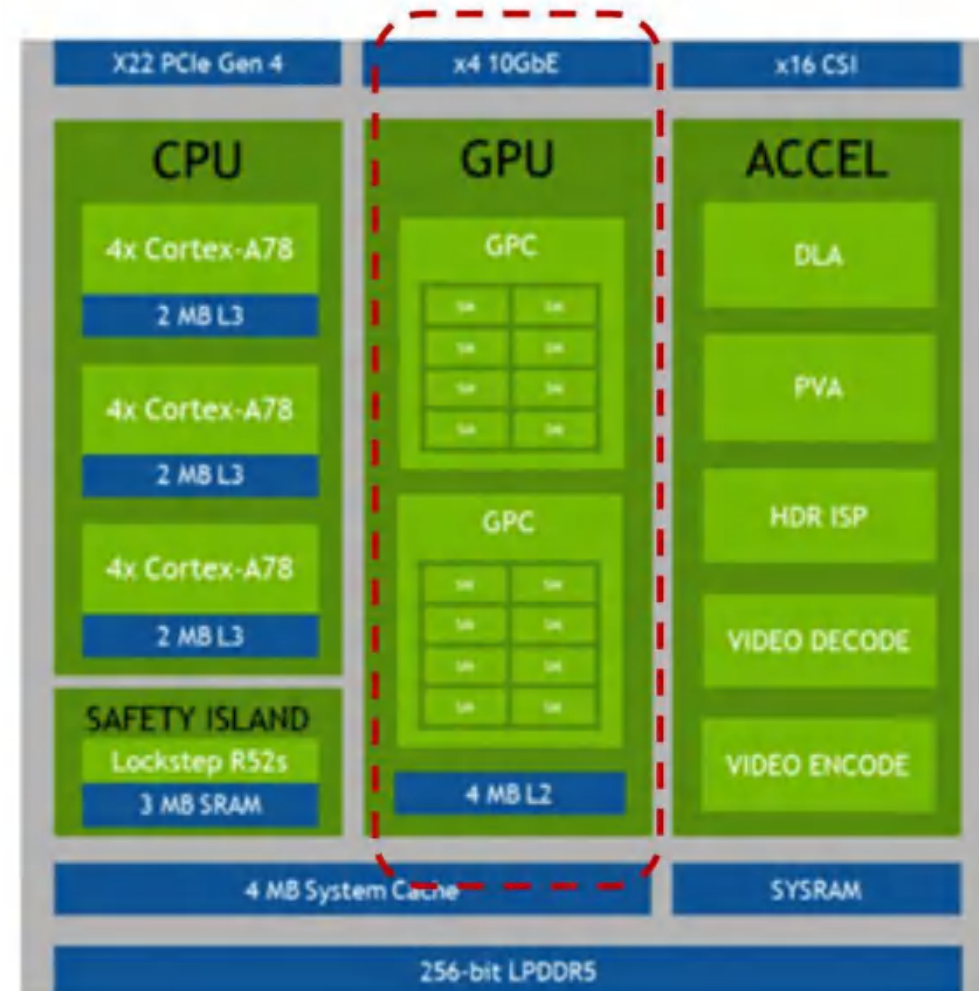
- 图形处理器（graphics processing unit，缩写：GPU），又称显示核心、视觉处理器、显示芯片，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像和图形相关运算工作的微处理器。
- NVIDIA公司在1999年发布GeForce 256图形处理芯片时首先提出GPU的概念。从此NVIDIA显卡的芯片就用这个新名字GPU来称呼。GPU使显卡削减了对CPU的依赖，并执行部分原本CPU的工作，尤其是在3D图形处理时。

GPU与显卡



资料来源：痞客邦，华金证券研究所

SOC中的GPU模块



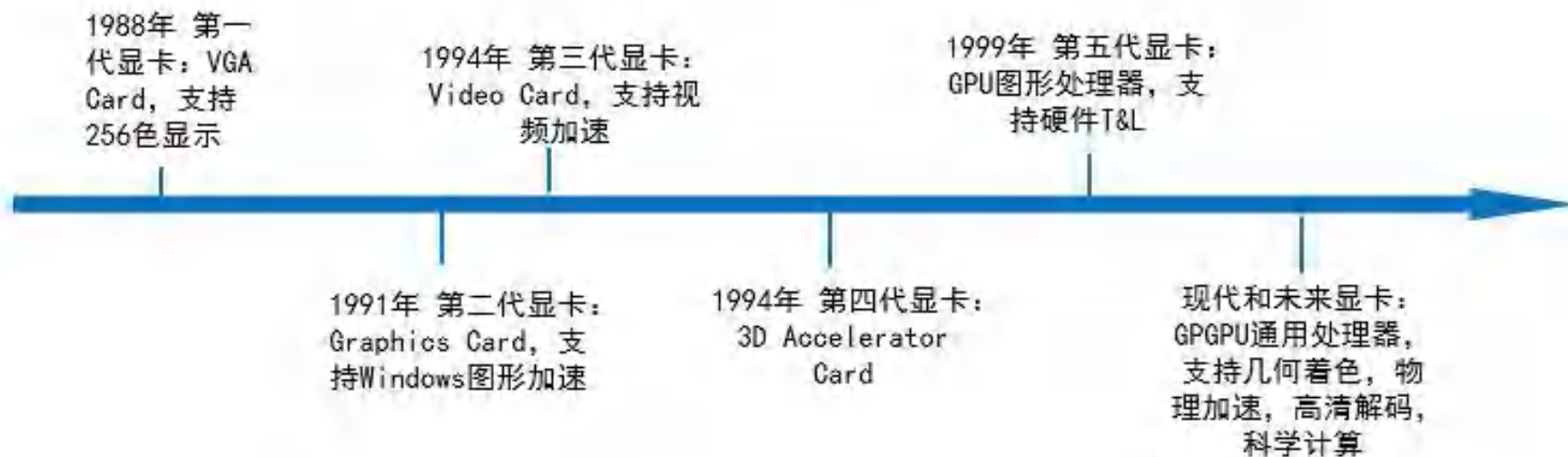
资料来源：痞客邦，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.2 始于图形处理设备

- 最早计算机是黑白显示的时代，机器对于显示的要求极低，随着计算机的普及和软件的多样化，使用者对于显示的要求越来越高。VGA（Video Graphics Array，视频图形阵列）是一种标准的显示接口，是IBM于1987年提出的一个使用模拟信号的电脑显示标准。VGA标准由于可以呈现的彩色显示能力大大加强，因此迅速成为了显示设备的标准，也推动了VGA Card也即是显卡的诞生。早期的VGA Card的唯一功能是输出图像，图形运算全部依赖CPU，当微软Windows操作系统出现后，大量的图形运算占据了CPU的大量资源，如果没有专门的芯片来处理图形运算，Windows界面运作会大受影响而变得卡顿，因此出现专门处理图形运算的芯片成为必然趋势。
- 1993年1月，英伟达创立，1999年，英伟达发布了划时代的产品GeForce 256，首次推出了所谓图形处理器（GPU，Graphic Processing Unit）的概念，它带来了3D图形性能的一次革命。

图：显卡发展历程

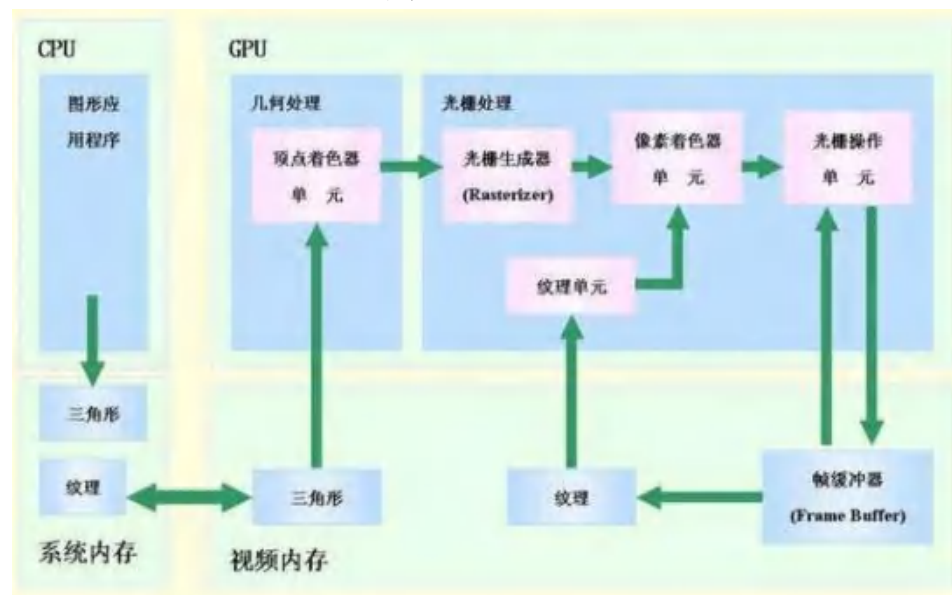


1. 由专用走向通用，GPU赛道壁垒高筑

1.3 浮点计算能力与可编程性结合

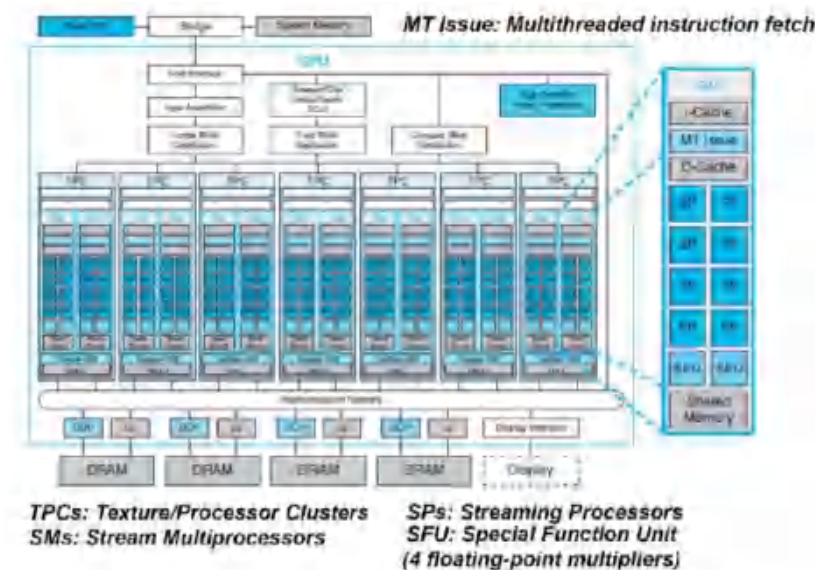
- GeForce 256 是一款用于实时图形处理的专用处理器，GeForce图形处理器的发布，实现了顶点的矩阵变换和光照计算，图形实时处理应用需要高内存带宽和大量的浮点计算能力。2001年英伟达发布了第三代显示核心GeForce3，GeForce3不仅集成了来自之前GeForce 256和GeForce2芯片的“静态”坐标转换和照明引擎，更增加了称为“顶点着色单元”的可编程顶点处理器功能。游戏开发者可借由加上顶点程序，让游戏产生令人惊艳的全新效果。
- 可编程性与浮点计算能力相结合，基于GPU的通用计算也开始出现，GPU朝着通用计算的方向持续演进。2006年，英伟达 CUDA (Compute Unified Device Architecture, 统一计算设备架构)，及对应工业标准的OpenCL的出现，让GPU实现更广泛的通用计算功能，GPGPU的概念落地。

GPU的图形（处理）流水线



资料来源：搜狐网，华金证券研究所

NVidia Tesla架构



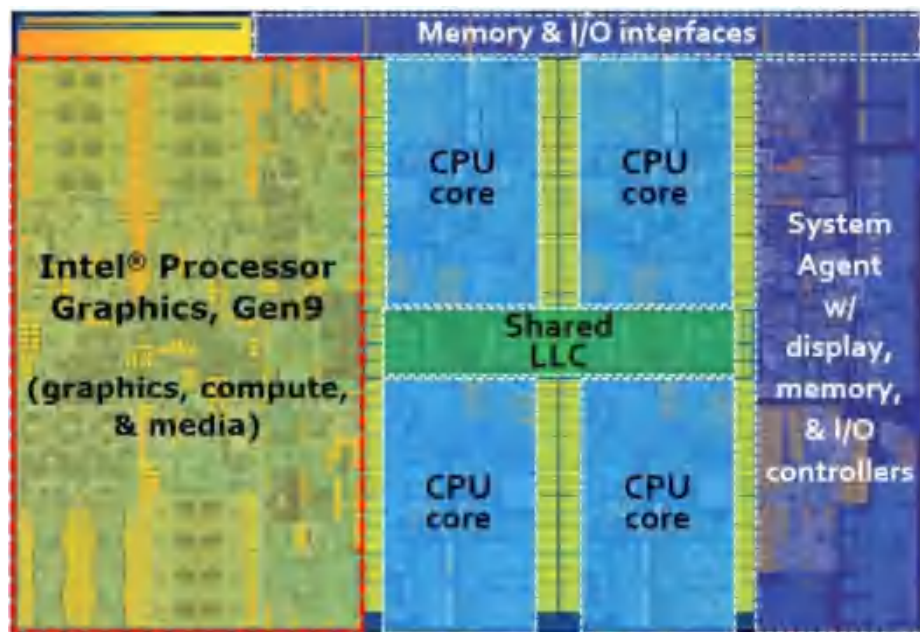
资料来源：《深入GPU硬件架构及运行机制》博客园，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.4 GPU发展三大方向

- GPU最初用在PC和移动端上运行绘图运算工作的微处理器，与CPU集成以集成显卡（核显）的形态发挥功能。NVIDIA于2007年率先推出独立GPU（独显），使其作为“协处理器”在PC和服务器端负责加速计算，承接CPU计算密集部分的工作负载，同时由CPU继续运行其余程序代码。
- 2019年NVIDIA的中国GTC大会设置了两大主题：AI和图形。从大会的关注重点可以看出，GPU未来趋势主要是3个：大规模扩展计算能力的高性能计算（GPGPU）、人工智能计算（AIGPU）、更加逼真的图形展现（光线追踪Ray Tracing GPU）。

四核心Intel处理器的die shot框图（带有Gen9核显）



英伟达三大产品系列



资料来源：CSDN，华金证券研究所

资料来源：英伟达，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.5 英伟达显卡发展历程

时间	发布型号	制程	亮点
1995	STG-2000X	500nm	采用第一代NV1核心，核心频率12MHz，同时支持2D、3D处理能力
1998	RIVA 128	350nm	第一款成功的显示核心。第一款支持微软Direct3D加速的图形芯片，也是第一个提供硬件三角形引擎的128 bit图形芯片，加入了对OpenGL技术的支持
1999	Riva TNT2	250nm	奠定英伟达显卡王朝的基石，核心频率和显存容量都有了极大的提升，从这一代开始，英伟达开始产品进行了市场化细分
1999	GeForce 256	220nm	首次推出了所谓图形处理器（GPU）的概念，增加了Pixel Shader流水线的数目，支持硬件T&L引擎，第一款硬件支持T&L的显卡，亦支援MPEG-2硬件视频加速。Quadro也是以GeForce256为基础开始研发。
2001	GeForce 3	180nm	英伟达首款支持DirectX 8.0的产品，并支持可编程的T&L引擎
2002	GeForce 4 Ti 4200	150nm	新一代的T&L引擎，并支持高效率的反锯齿技术
2004	GeForce 6800	130nm	渲染管线首次突破性增长到16条，采用GDDR3显存，频率达到了1.1GHz。同年，英伟达SLI（可扩展的链接接口）技术问世，单台PC的图形处理能力大大提升。
2006	GeForce 8800 GTX	90nm	世界上第一块支持DirectX 10的PC桌面显卡。GeForce 8采用统一流水线结构，传统显示核心的架构分为顶点着色引擎和像素着色引擎。所谓统一渲染，即GPU中不再有单独的顶端渲染单元和像素渲染单元，而是由一个通用的渲染单元同时完成顶点和像素渲染任务。统一渲染架构具有硬件利用效率高以及编程灵活的优点，进一步提升了GPU内部运算单元的可编程性，让GPU运行高密度的通用计算任务就成为可能
2010	GeForce GTX 480	40nm	采用英伟达推出全新一代的Fermi架构，Fermi架构GPU产品在保持图形性能的前提下，将通用计算的重要性提升到前所未有的高度，大规模GPU计算从之开始。30亿个晶体管的大芯片全局ECC设计、可读写缓存、更大的shared memory、甚至出现了分支预测概念。Fermi是英伟达最后一款在游戏显卡上保留强悍双精度的微架构
2013	GeForce GTX Titan	28nm	采用Kepler架构，与前一代的Fermi架构相比，Kepler架构不仅仅是性能的提升，功耗和温度上也得到了极大的改善。Fermi架构中英伟达主要专注于提升计算与曲面细分的性能。然而在Kepler架构中，英伟达转向了提升效率、可编程性与性能，效率的提升来自采用了统一的GPU时钟、简化的静态指令调度和更加优化的每瓦性能。专用的双精度CUDA核心被用来弥补Kepler CUDA核心为了节省芯片面积而放弃的双精度计算能力
2014	GeForce GTX 970	28nm	采用英伟达第四代GPU架构Maxwell架构，Kepler的改进版架构。最明显的变化是在SMX单元和GPC单元上，Maxwell的SMM（之前叫SMX）单元从之前Kepler的包含192个CUDA Core下降到128个，但发射器从之前的每SMX一个变为了每SMM四个，目的是降低每个SMM单元的运算压力提升效率，增加了两个寄存器，然后L1缓存翻倍，GPC单元的L2缓存增加到了2M。Maxwell将具备以下三大特性：提升图形性能，降低编译难度（这应该归功于ARMv8核心和统一内存寻址增强技术）和提高能耗比。
2016	GeForce GTX 1080	16nm	这一代显卡的工艺和架构全面升级。架构方面，采用了Pascal架构，Pascal是Maxwell的接替者，增强了异步计算功能实现硬件层了对DirectX API的更高版本（DirectX 12 Feature Level 12_1）的支持，高端产品还配备带宽更高的HBM2显存，性能和能耗比都有了很大提升
2018	GeForce RTX 2080	12nm	第一代GeForce RTX系列，支持光线/路径追踪硬件加速，使实时光线追踪成为可能。新GeForce显卡最大的亮点就是集成了光线追踪核心的Turing GPU，从技术上拉开了与上代显卡的差距，NVIDIA宣布图灵架构的时候表示新一代显卡的光线追踪性能是现有Pascal显卡的6倍之多
2020	GeForce RTX 3090	三星 8nm	采用了全新的Ampere安培架构，相比RTX20系的图灵架构是革命性的提升，Ampere集成了第二代RT光线追踪核心、第三代Tensor张量核心，并支持PCIe4.0、DisplayPort1.4a、HDMI2.1
2022	GeForce RTX 40系列	4nm	采用最新的Ada Lovelace架构，较上一代 Ampere 晶体管和 CUDA 核心数量提升 70%，着色器、光追、深度学习性能均实现重大飞跃。Ada Lovelace架构的创新大体上可以分为三个板块，分别是带来了新的全景光线追踪、着色器执行重排序（SER）和DLSS 3

1. 由专用走向通用，GPU赛道壁垒高筑

1.6 GeForce RTX 40系列，时代最强

- 2022秋季GTC大会上，英伟达发布GeForce RTX® 40系列GPU，旨在为游戏玩家和创作者提供革命性性能，其中新旗舰产品RTX 4090 GPU的性能相较上一代提升最高可达4倍。作为全球首款基于全新NVIDIA® Ada Lovelace架构的GPU，RTX 40系列在性能和效率上都实现了巨大的代际飞跃，根据NVIDIA创始人兼首席执行官黄仁勋的介绍，RTX光线追踪和神经网络渲染的时代正在全面展开。
- RTX 40系列GPU具有一系列新的技术创新：包括流式多处理器具有高达83 TFLOPS的着色器能力、第三代RT Cores的有效光线追踪计算能力达到191 TFLOPS、第四代Tensor Cores具有高达1.32 Petaflops的FP8张量处理性能、着色器执行重排序（SER）通过即时重新安排着色器负载来提高执行效率、Ada光流加速器带来2倍的性能提升、架构上改进来实现与TSMC 4N定制工艺技术紧密结合等。

GPU 规格参数对比			
	RTX 4090	RTX 4080	RTX 3080 Ti
核心架构	Ada Lovelace	Ada Lovelace	Ampere
制造工艺	TSMC 4N	TSMC 4N	Samsung 8N
晶体管数量	763 亿	459 亿	283 亿
图形处理器集群 (GPC)	11	7	7
纹理处理集群 (TPC)	64	38	40
流处理单元数量 (SM)	128	76	80
CUDA 核心数量	16384	9728	10240
张量单元 (Tensor Cores)	512 (第四代)	304 (第四代)	320 (第三代)
光追单元 (RT Cores)	128 (第三代)	76 (第三代)	80 (第二代)
纹理单元 (TUs)	512	304	320
光栅单元 (ROPs)	176	112	112
加速频率	2520 MHz	2505 MHz	1665 MHz
显存传输率	21 Gbps	22.4 Gbps	19 Gbps
二级缓存	73728 KB	65536 KB	6144 KB
显存容量	24GB GDDR6X	16 GB GDDR6X	12 GB GDDR6X
显存位宽	384 bit	256 bit	384 bit
显存带宽	1008 GB/s	716.8 GB/s	912 GB/s
整板功率 (TGP)	450 W	320 W	350 W

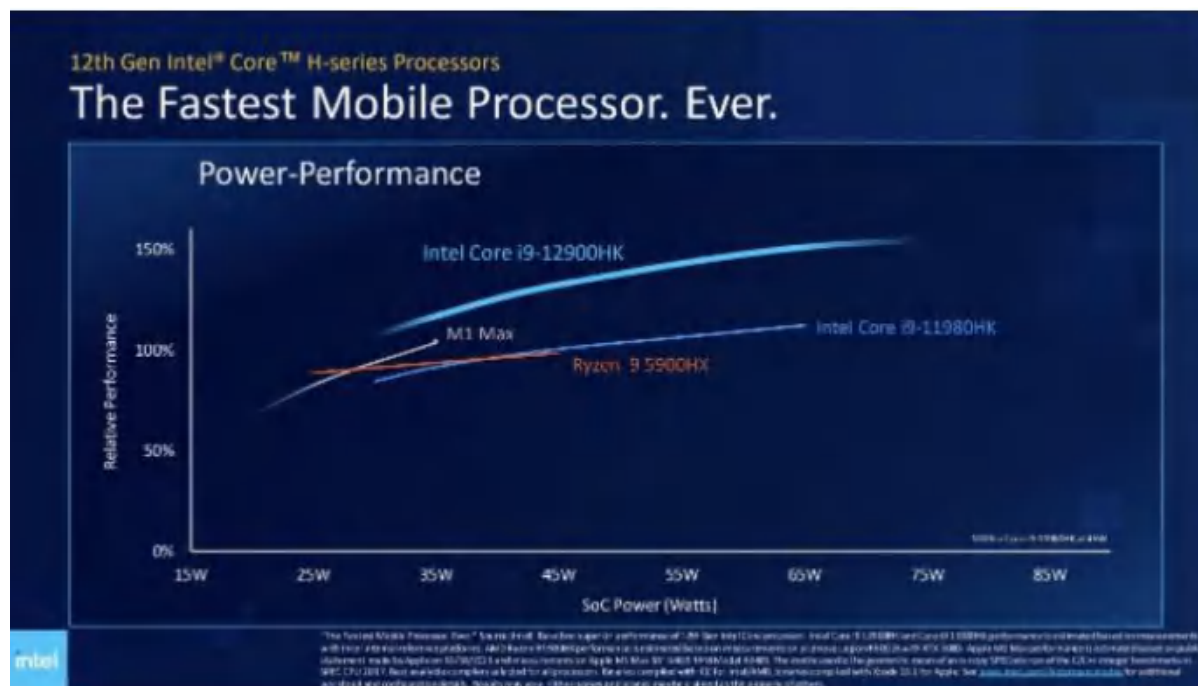
资料来源：电脑评测网，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.8 核显与独显性能对比

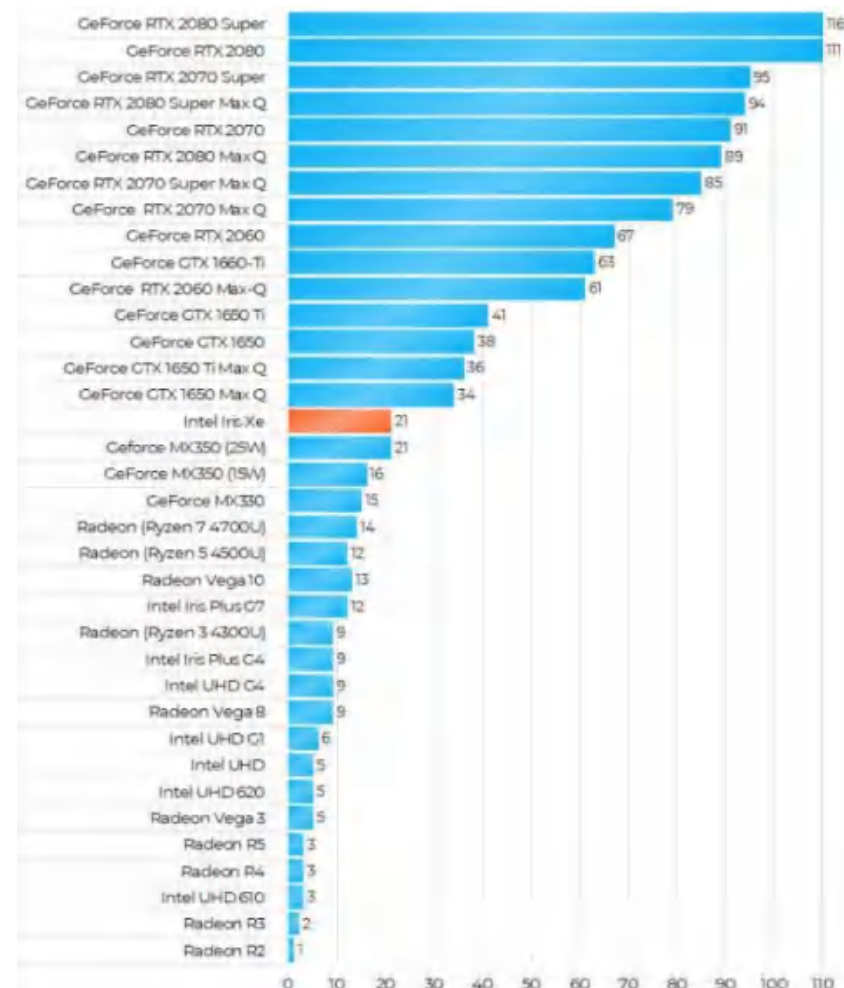
- 2022年1月25日，搭载第12代酷睿Alder Lake-H处理器的笔记本正式上市，采用最新一代Intel 7制程工艺，内置Iris XE GPU，拥有48组EU单元，加速频率高达1450MHz。

Intel第12代酷睿性能图



资料来源：量子位，华金证券研究所

英特尔Iris XE GPU的跑分



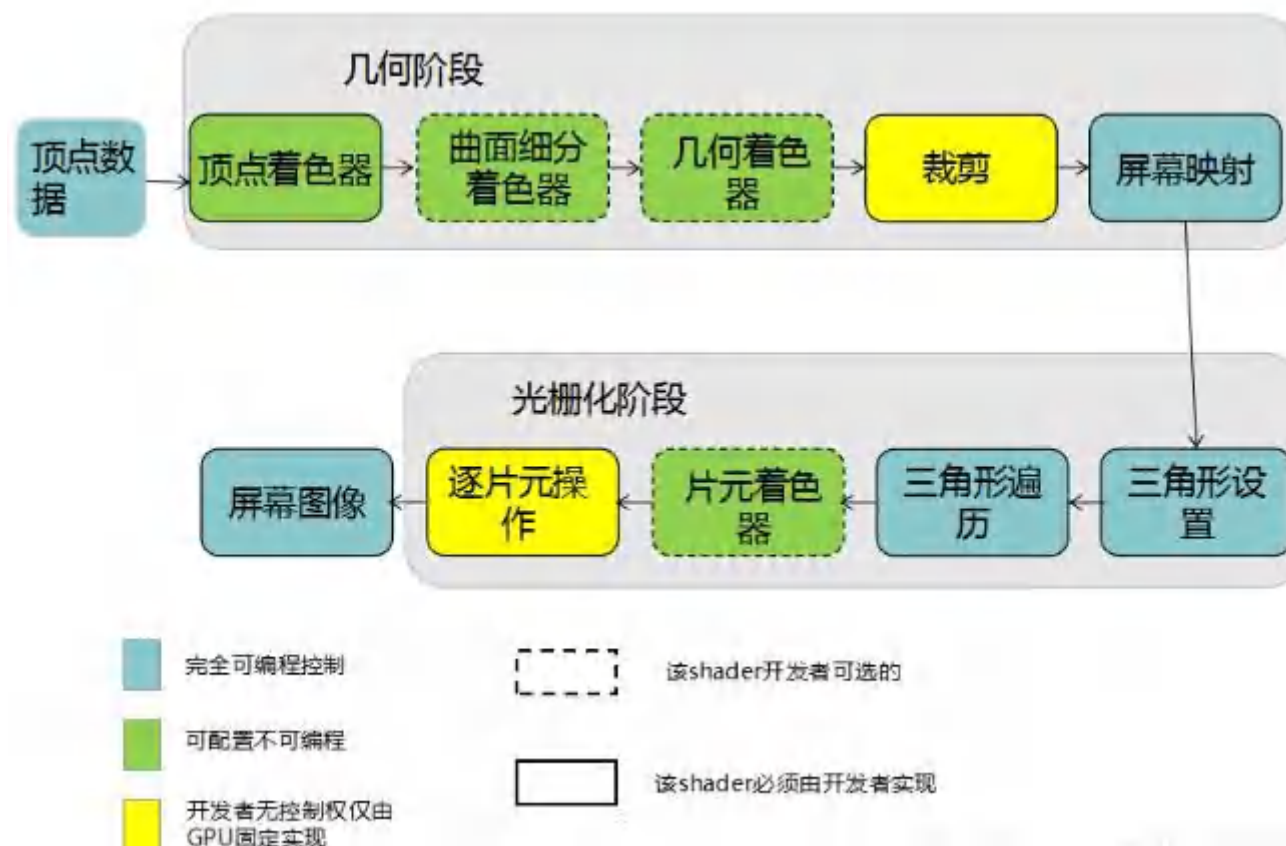
资料来源：zmmoo，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.9 图形流水线是GPU工作的通用模型

- 图形流水线（graphics pipeline），也叫图形管线，指的是一连串的图形处理任务，这一系列的工作先后有序、不可颠倒，因此得以有这个形象的称呼。图形流水线是GPU工作的通用模型，它以某种形式表示的三维场景为输入，输出二维的光栅图形到显示器。

图：图形流水线

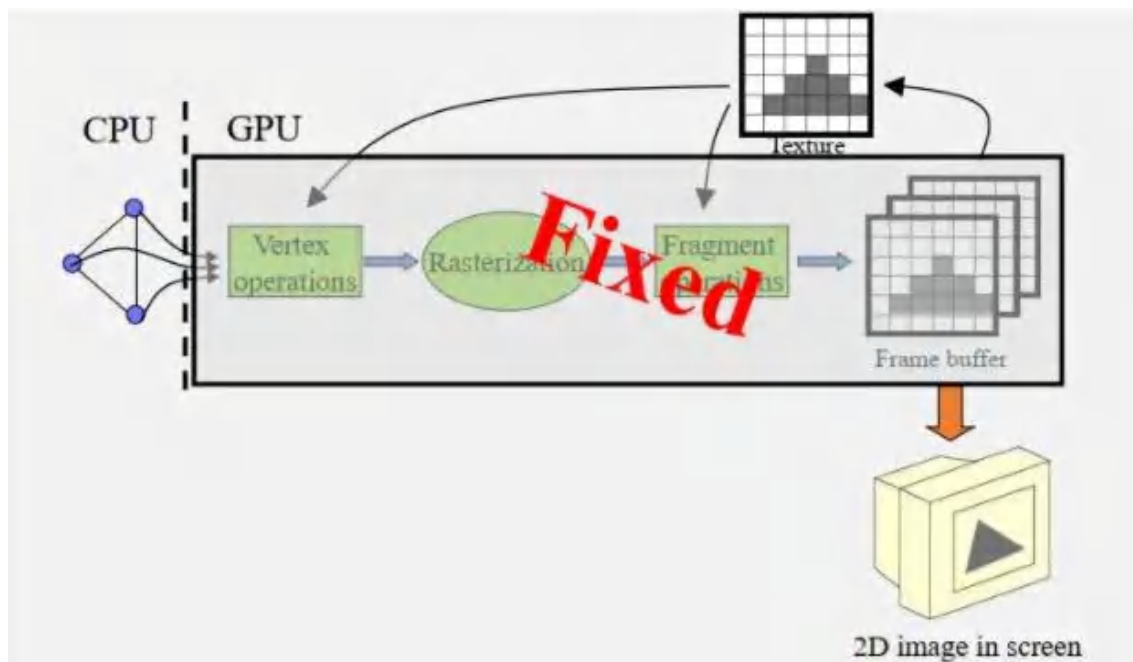


1. 由专用走向通用，GPU赛道壁垒高筑

1.10 统一渲染架构的推出开启了通用计算大时代

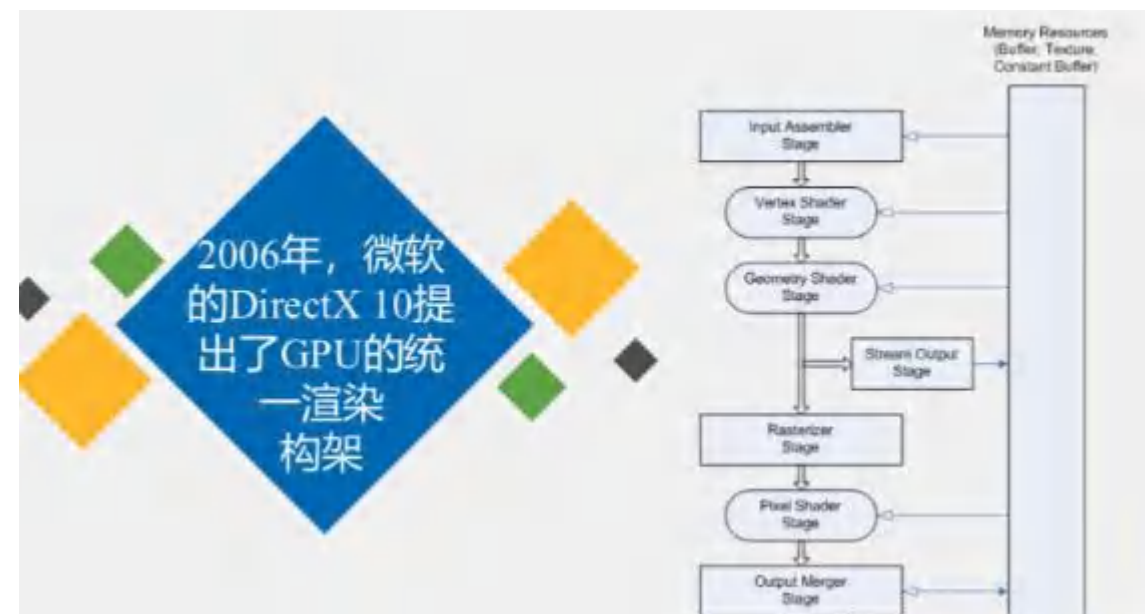
- GPU的硬件结构从固定功能流水线架构发展为大规模并行的统一染色器架构。所谓统一渲染，即GPU中不再有单独的顶端渲染单元和像素渲染单元，而是由一个通用的渲染单元同时完成顶点和像素渲染任务。为了实现这一点，图形指令必须先经过一个通用的解码器、将顶点和像素指令翻译成统一渲染单元可直接执行的渲染微指令，而统一渲染单元其实就是一个高性能的浮点和矢量计算逻辑，它具有通用和可编程属性。在统一渲染架构的GPU中，Vertex Shader和Pixel Shader概念都将废除同时代之以ALU。ALU是个完整的图形处理体系，它既能够执行对顶点操作的指令（代替VS），又能够执行对象素操作的指令（代替PS）。基于统一渲染架构，Shader Core被挖掘出了更多的使用方法，比如通用计算。

早期的GPU只支持固定管线



资料来源：CSDN，华金证券研究所

统一渲染架构



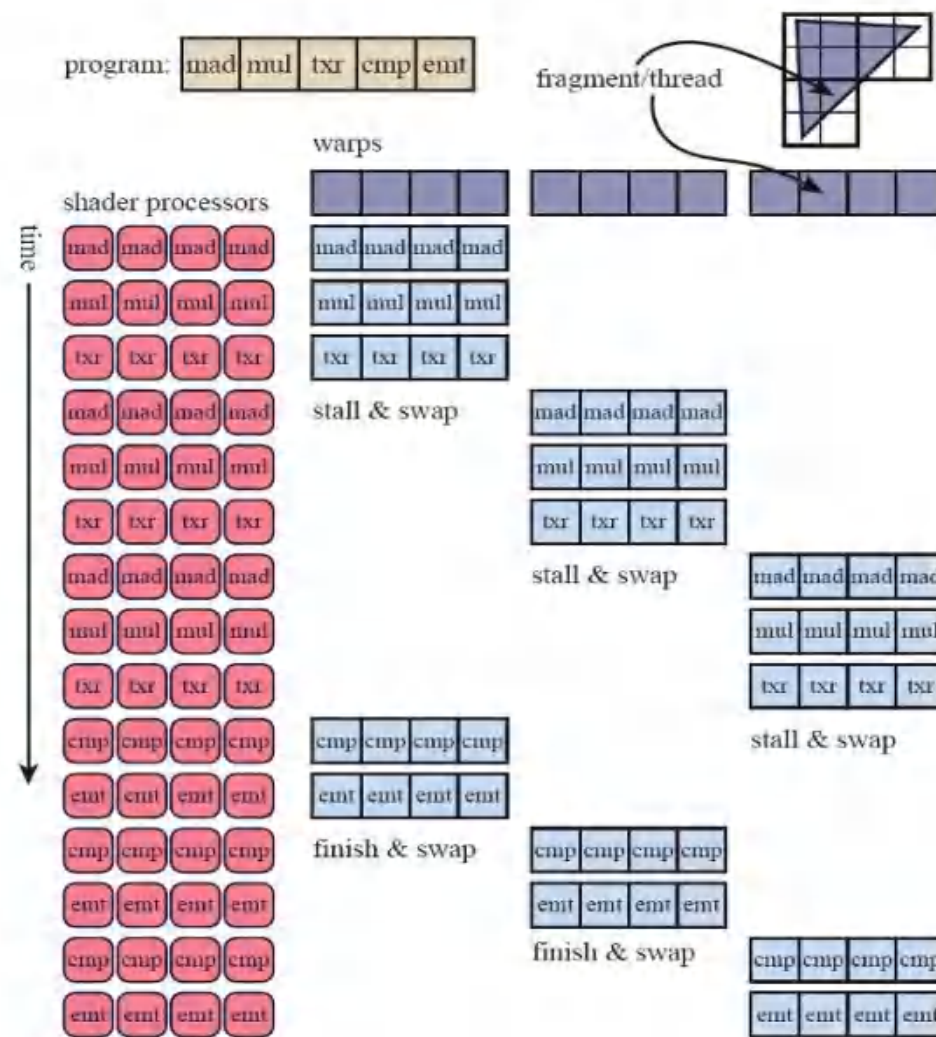
资料来源：CSDN，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.11 从简单到越来越复杂的流水线

- 以前GPU只支持固定管线，并且不支持编程，2002年，GPU在Vertex Operations和Fragment Operations这两个模块中具有了可编程功能，2006年GPU流水线中增加了一种新的模块，Geometry Shader（几何元着色器），使得图形程序开发者在可编程渲染管道（programmable render pipeline）下能够更大的发挥自由度。再之后，Tessellation（细分曲面技术）、Mesh着色器等等功能的加入，GPU的流水线变得越来越复杂。
- GPU要实现对二维屏幕上每一个像素点的输出，需要很多个并行工作的着色处理器shader processor同步工作，示意图中将硬件中的四个小处理器连为一组，软件层面将各类渲染任务按4个thread打成一个卷warp发给硬件，同时加入了多warp切换的机制，保证了GPU任务执行的高效性。

当代GPU渲染管线示意图

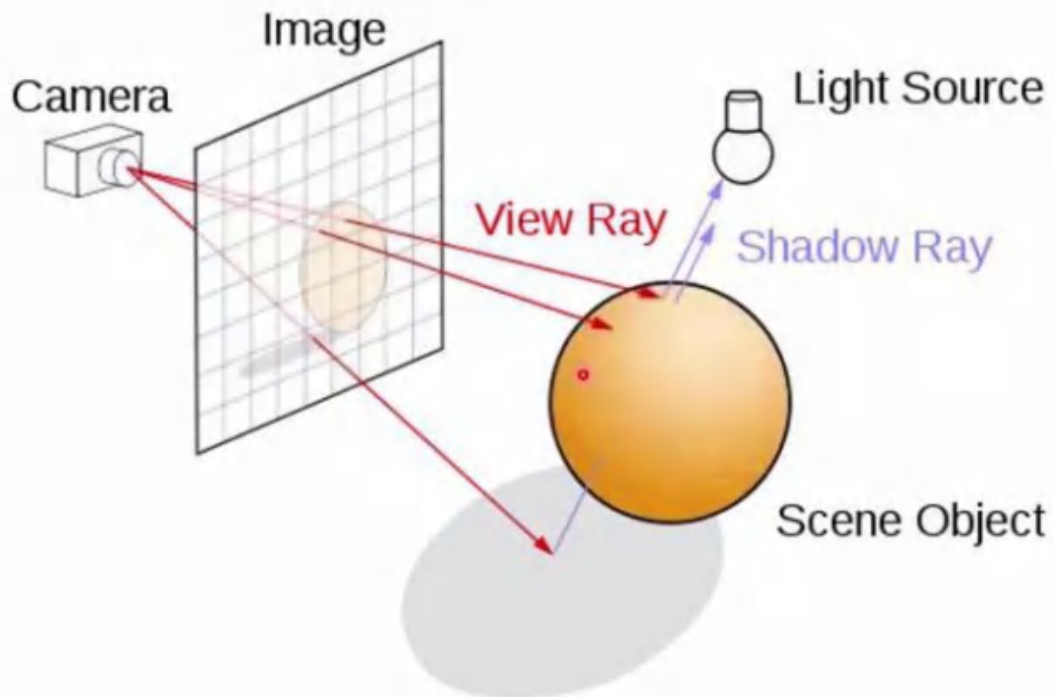


1. 由专用走向通用，GPU赛道壁垒高筑

1.12 光线追踪时代开启

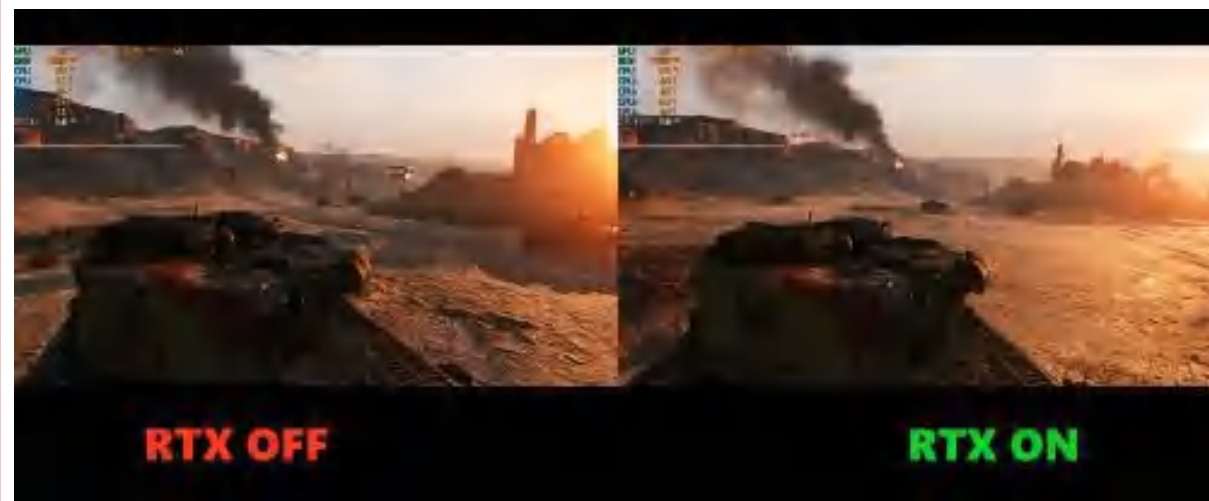
- 光线跟踪是一种真实地显示物体的方法，该方法由Appel在1968年提出。光线跟踪方法沿着到达视点的光线的反方向跟踪，经过屏幕上每一个像素，找出与视线相交的物体表面点P0，并继续跟踪，找出影响P0点光强的所有光源，从而算出P0点上精确的光线强度，在材质编辑中经常用来表现镜面效果。光线跟踪或称光迹追踪是计算机图形学的核心算法之一。在算法中，光线从光源被抛射出来，当他们经过物体表面的时候，对他们应用种种符合物理光学定律的变换。最终，光线进入虚拟的摄像机底片中，图片被生成出来。

光线追踪原理图



资料来源：CSDN，华金证券研究所

光线追踪对比图

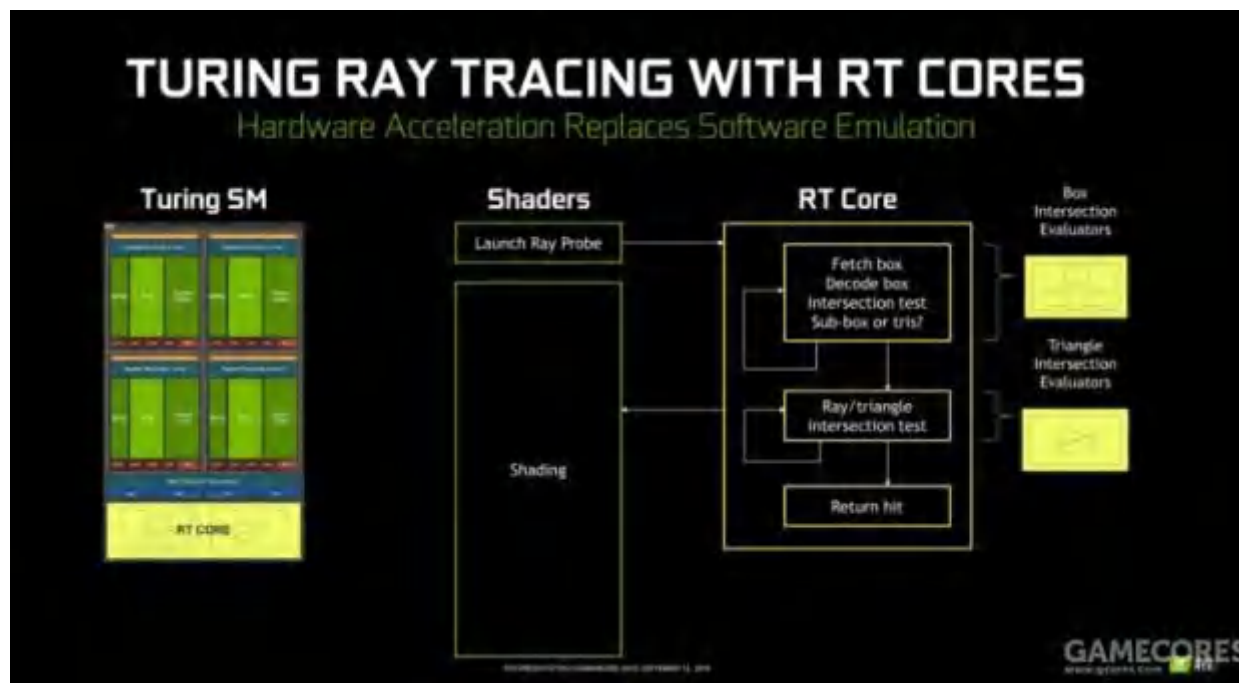


资料来源：新浪，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.13 光线追踪算法要求的计算量巨大

- 光线追踪与光栅化的实现原理不同。光栅化渲染管线是传统的渲染管线流程，是以一个三角形为单元，将三角形变成像素的过程；光线追踪渲染管线则是以一根光线为单元，描述光线与物体的求交和求交后计算的过程。和光栅化线性管线不同的是，光线追踪的管线是可以通过递归调用来衍生出另一根光线，并且执行另一个管线实例。光线追踪最大难点在于对算力要求极高，计算量非常庞大。
- 2018年NVIDIA发布的RTX 2080 GPU，采用Turing架构，在GPU中集成了68个独立的 RT(ray tracing) Core（专门为光线追踪服务的，实质上它是一条特异化的专用流水线），用于光线追踪，光线处理能力达到了10 Giga/S，1080P@60Hz需要处理的光线约为6Giga/S，光线追踪对于反射和阴影有着更逼真的处理效果，尽管目前仍然是采用光线追踪和传统光栅图形处理相结合的方式来进行图形渲染，但其效果已经远超传统光栅图形处理。

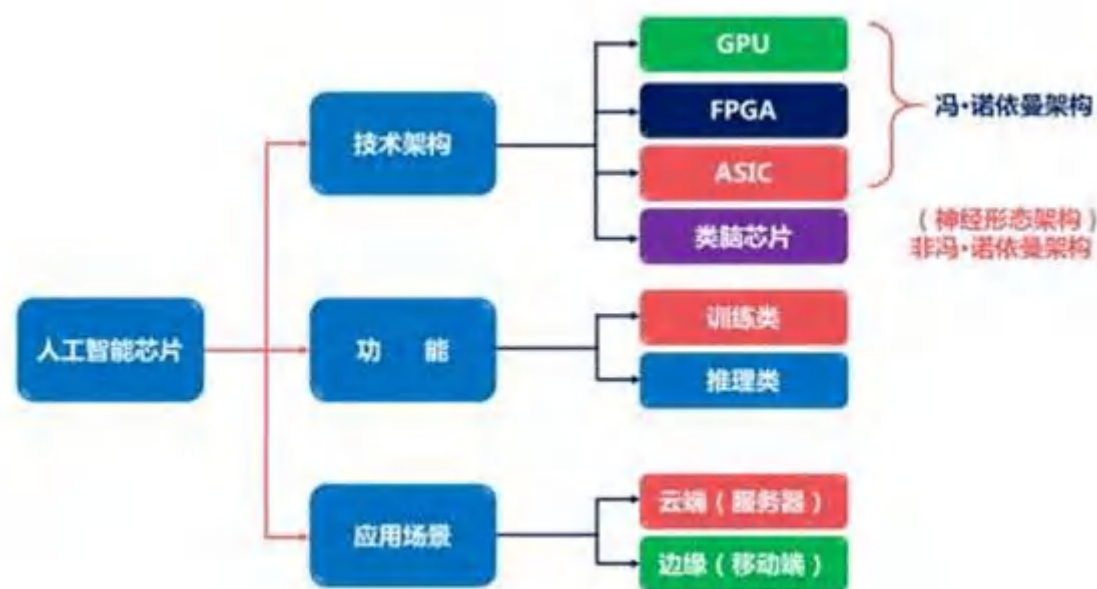


1. 由专用走向通用，GPU赛道壁垒高筑

1.14 走向新场景的GPGPU

- 对GPU通用计算进行深入研究从2003年开始，并提出了GPGPU概念，前一个GP则表示通用目的（General Purpose），所以GPGPU一般也被称为通用图形处理器或通用GPU。伴随着GPU Shader单元计算能力的不断增长，GPU也在向通用计算开始扩张边界。GPU从由若干专用的固定功能单元（Fixed Function Unit）组成的专用并行处理器，进化为了以通用计算资源为主，固定功能单元为辅的架构，这一架构的出现奠定了GPGPU的发展基础。
- GPGPU由于其高并发性、高吞吐量以及不断提升的可编程能力，目前的应用已经扩展到科学计算、区块链、大数据处理、工程计算、金融、基因等方面。

AI芯片的分类



计算是未来科学和工程突破的关键



一些典型科学与工程计算应用的计算需求估计

二维非混合流计算:	10^{16} 次/秒;
二维分子动力学计算:	10^{14} 次/秒;
三维多介质(氢)计算:	10^{15} 次/秒;
输运计算:	10^{15} 次/秒;
全球海洋计算:	10^{15} 次/秒;
三维重结构计算:	10^{15} 次/秒;
QCD, 聚冷格子计算:	10^{16} 次/秒;
三维计算:	10^{16} 次/秒;
计算流体力学:	10^{16} 次/秒;
原子内电子能级计算:	10^{17} 次/秒;

1. 由专用走向通用，GPU赛道壁垒高筑

1.15 GPU与GPGPU的对比

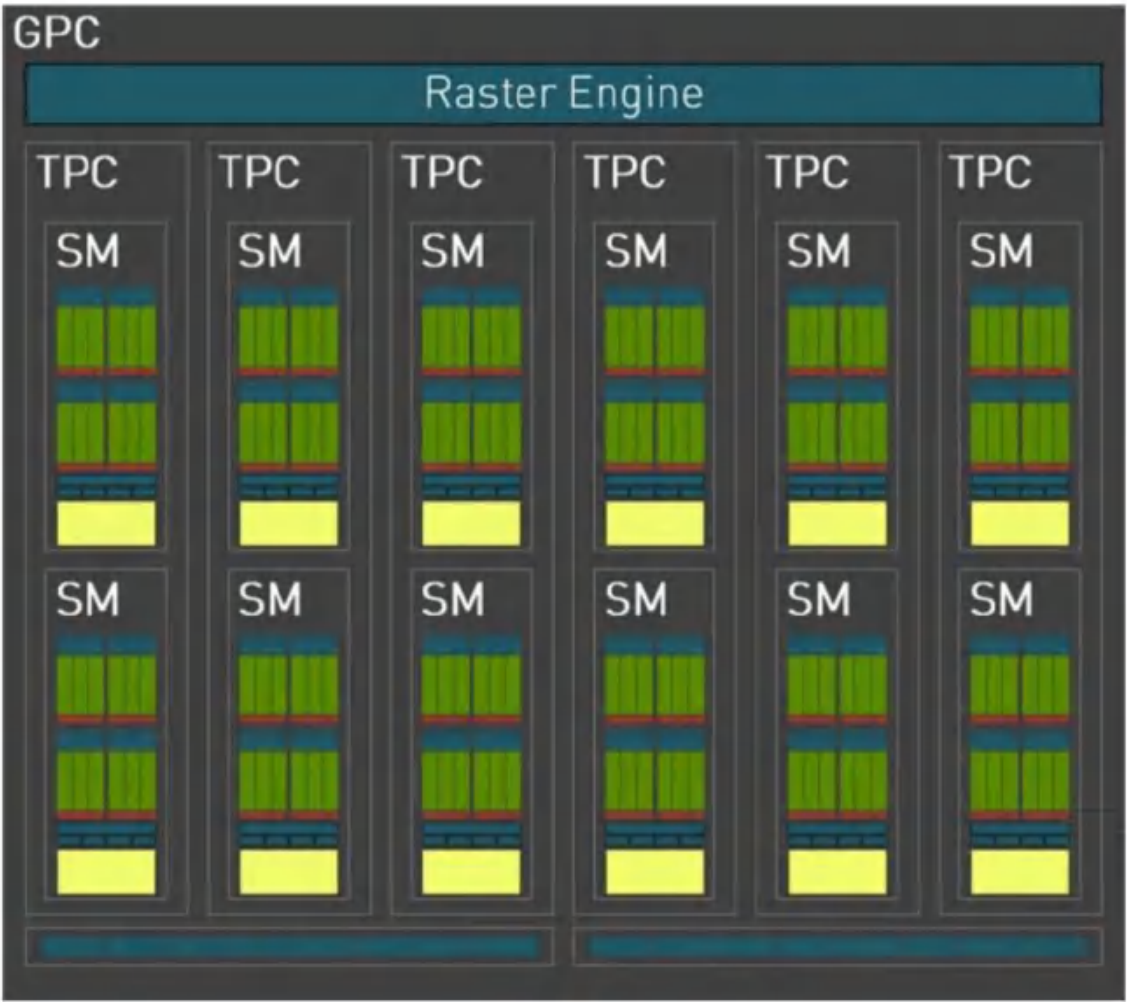
➤ GPU的核心价值体现在图形图像渲染，GPGPU的重点在于算力，虽然都是由GPU的架构演进而来，但所关注的重点有明显区别。GPGPU架构设计时，去掉了GPU为了图形处理而设计的加速硬件单元，保留了GPU的SIMT架构和通用计算单元，使之更适合高性能并行计算，并能使用更高级别的编程语言，在性能、易用性和通用性上更加强大。

GPU与GPGPU对比

技术门槛：GPU的软件系统设计复杂度超越AI芯片100倍以上		
	GPU	GPGPU&AI
指令集	千条量级	百条以内
算力利用率	基于硬件层的任务管理和智能调度能力要求高 让芯片从硬件层即提高算力的利用率	多依赖于软件层的调度实现，但： <ul style="list-style-type: none">增加了软件开发的复杂度降低了硬件算力的利用率减缓了软件栈迭代更新的速度
开发生态	<ul style="list-style-type: none">生态系统复杂涉及行业应用广泛行业标准成熟(OpenCL, OpenGL, Vulkan, DirectX)	<ul style="list-style-type: none">新生态面临客户冗长的适配期和巨大的软件投入产品“快”，应用“慢”

资料来源：新浪网，华金证券研究所

NVIDIA GeForce RTX 40的GPC单元



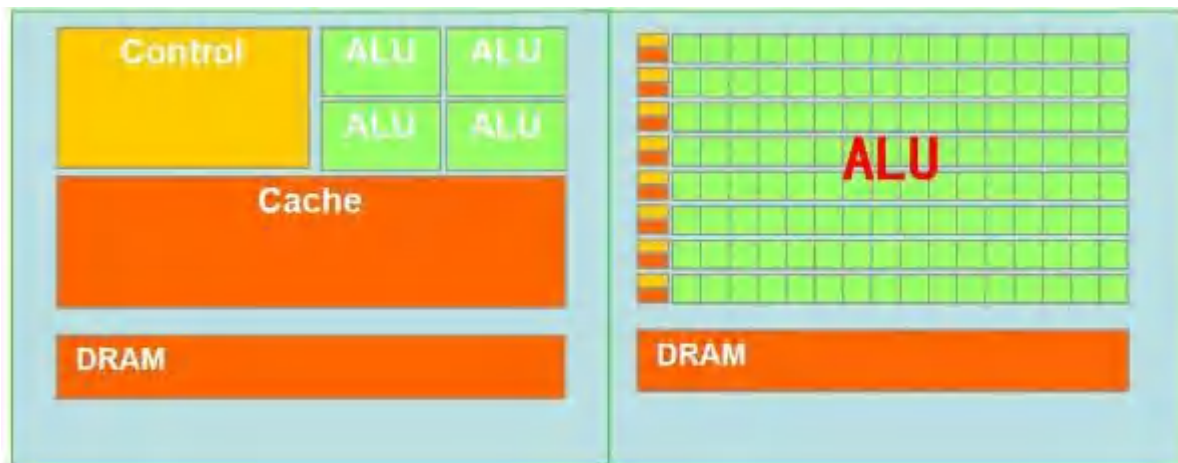
资料来源：英伟达，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.16 GPGPU与CPU的对比

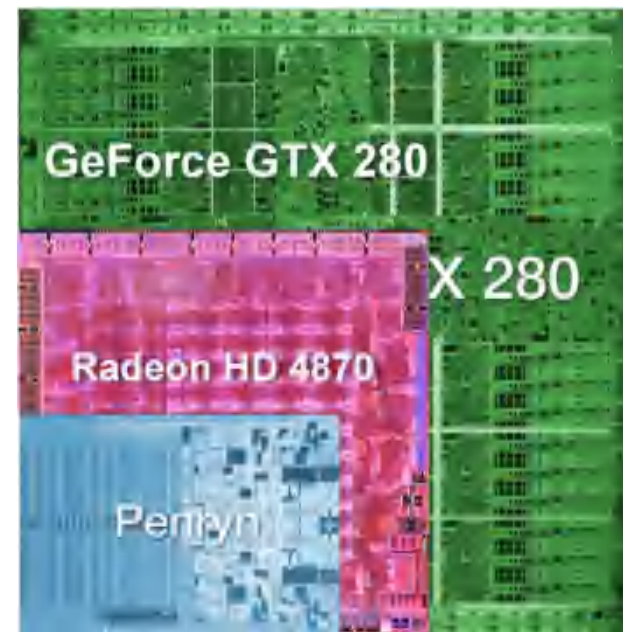
- CPU作为计算机系统的运算和控制核心，是信息处理、程序运行的最终执行单元。CPU内部主要由运算器、控制器和寄存器组成，运算器执行数值计算，寄存器储存数据。CPU是程序的调用者和运行者，计算机的每一条指令都要经过CPU的解析和执行。GPU无法单独工作，必须由CPU进行控制调用才能工作。CPU可单独作用，处理复杂的逻辑运算和不同的数据类型，但当需要大量的处理类型统一的数据时，则可调用GPU进行并行计算。
- CPU与GPU从设计之初就是为了实现不同的目标，GPU的构成相对简单，有数量众多的计算单元和超长的流水线，特别适合处理大量的类型统一的数据。GPU为并行而设计，更重视整体数据吞吐量（Throughput）；CPU为串行而设计，更看重任务间的时延（Latency）。与超标量乱序CPU相比，通过减少用于控制逻辑的面积并增加算术逻辑单元的面积，GPU可以在高度并行的工作负载上获得更好的单位面积性能。

CPU与GPGPU架构对比（ALU用于计算的晶体管）



资料来源：imagination，华金证券研究所

CPU与GPU芯片面积对比



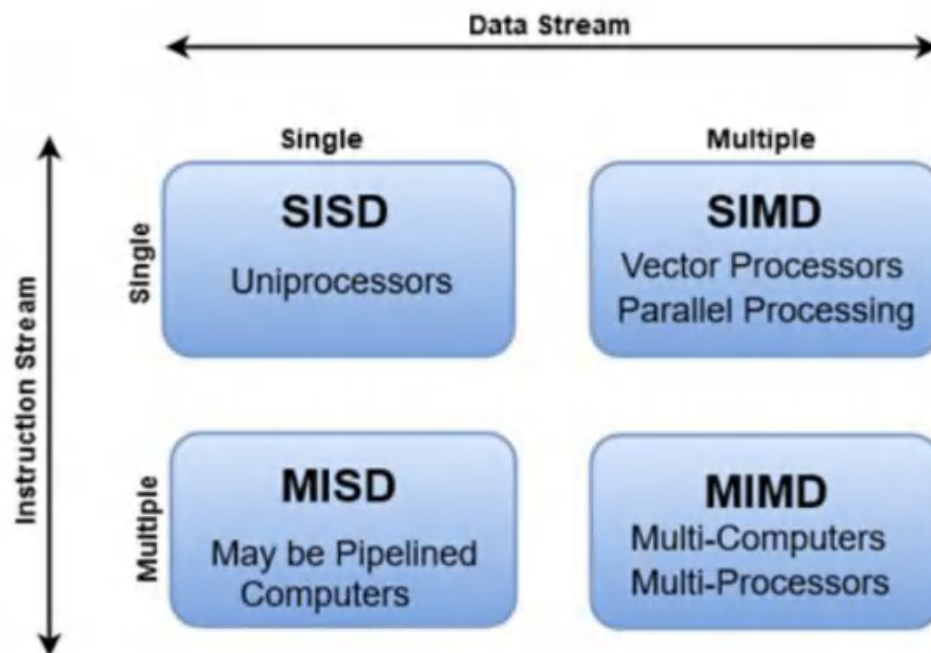
资料来源：anandtech，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.17 并行计算发展的核心

- 现代计算机发展经历了串行计算时代、并行计算时代，并行计算机是由一组处理单元组成的，这组处理单元通过互相之间的通信与协作，以更快的速度共同完成一项大规模的计算任务。并行计算机体系结构的发展主要体现在计算节点性能的提高及节点间通信技术的改进两方面。
- 弗林分类法，根据指令流和数据流的不同组织方式把计算机体系的结构分为四类：单指令流单数据流（SISD）、单指令流多数据流（SIMD）、多指令流多单数据流（MISD）、多指令流多数据流（MIMD）。指令流指的是机器执行的指令序列；数据流指指令流调用的数据序列，包括输入数据和中间结果。SIMD是一种执行模型，这意味着处理器将其用于在管道中将相似的数据集排队并并行执行的方法，是现代CPU和GPU使用的最受欢迎的EM之一。

弗林分类法

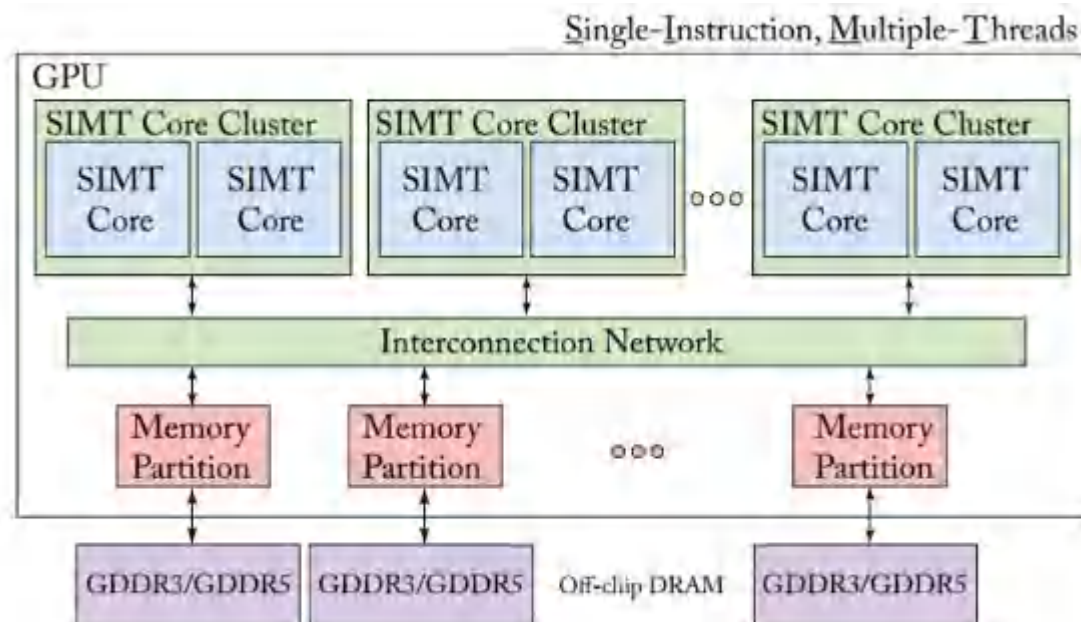


1. 由专用走向通用，GPU赛道壁垒高筑

1.18 SIMT，主流GPU的系统架构核心

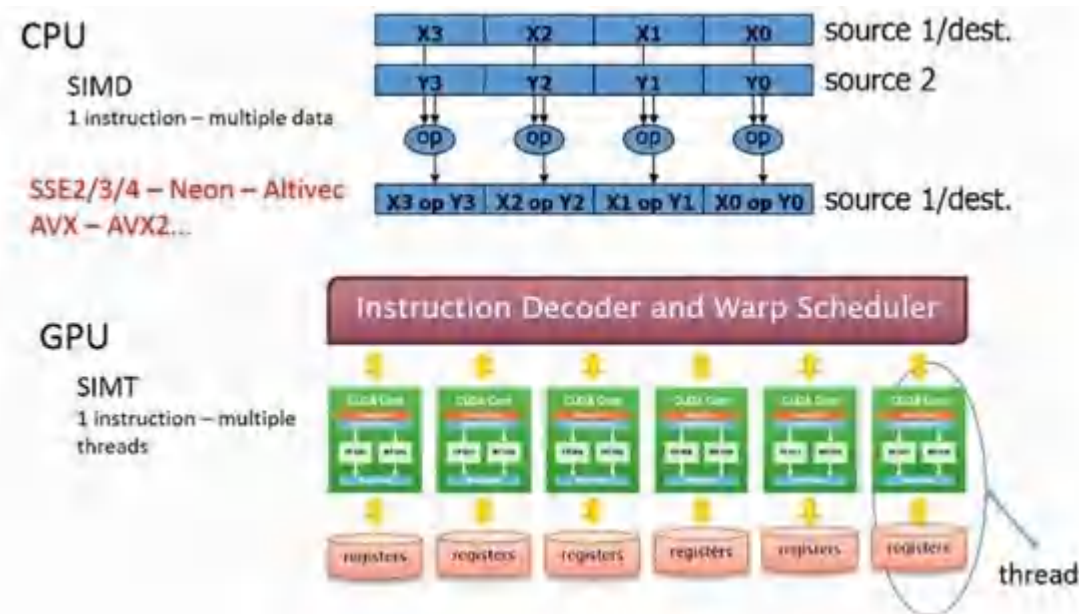
- 现代的GPU架构中，每个GPU会包含很多的core，英伟达称之为流多处理器(streaming multiprocessors, SM)。每个核都在执行单指令多线程的程序(single-instruction multiple-thread, SIMT)。在单个核上执行的线程可以通过暂存内存（有点像阻塞操作，保存现场）进行通信，并使用快速barrier操作进行同步。
- SIMT与SIMD（同一条指令多个数据）的共同点是同一条指令。SIMT是SIMD的线程等价物，不同之处在于，SIMD使用执行单元或矢量单元，而SIMT将其扩展为利用线程。SIMT的好处是无需开发者费力把数据凑成合适的矢量长度，并且SIMT允许每个线程有不同的分支。SIMT的主要优点是它减少了指令预取带来的等待时间。

现代GPU简单架构示意图



资料来源：CSDN，华金证券研究所

SIMD与SIMT对比

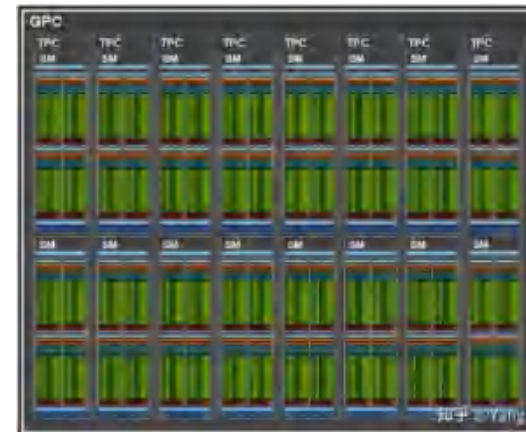


资料来源：新浪VR，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.19 GPGPU架构，以A100为例

- A100是NVIDIA2020年5月14日发布的采用新一代Ampere架构的计算卡，使用了GA100核心。Ampere架构仍然沿用了成熟的GPC-TPC-SM多级架构，GA100内部包含8组图形处理集群（Graphics Processing Cluster，GPC），每组GPC包含8组纹理处理集群（Texture Processing Cluster，TPC），每组TPC又包含8组流式多处理器（Streaming Multiprocessor，SM），另外还有内存控制器组成。



1. 由专用走向通用，GPU赛道壁垒高筑

1.20 Fermi是第一个完整的GPU计算架构

- 英伟达的Fermi是第一个完整的GPU计算架构，该架构在保持图形性能的前提下，将通用计算的重要性提升到前所未有的高度，大规模GPU计算从此开始。
- 要做通用计算，需要更强大的线程管理能力，更强大的仲裁机制，丰富的共享cache和寄存器资源以及充足的发射端等。全新Fermi架构，是以处理器为目标进行设计，因此Fermi架构新增了以前GPU上从来没有的东西，包括更多的指令双发射、统一的L2全局缓存、64KB的可配置式L1或者Shared Memory、大量的原子操作单元等等。

GF100费米架构核心示意图



资料来源：快懂百科，华金证券研究所

1. 由专用走向通用，GPU赛道壁垒高筑

1.21 通用算力提升是英伟达GPU架构演进的重点之一

- 2016年3月英伟达推出Pascal架构,采用16nm和14nm的工艺。该架构建立在五大技术突破之上,启用了全新的计算平台,打破了从书桌端到数据中心的传统思维。Pascal彻底采用全新设计,为深度学习和其他计算工作负载提供更好的性能。该架构利用全新的混合精度指令,可为深度学习提供每秒超过20万亿次浮点运算的性能峰值。

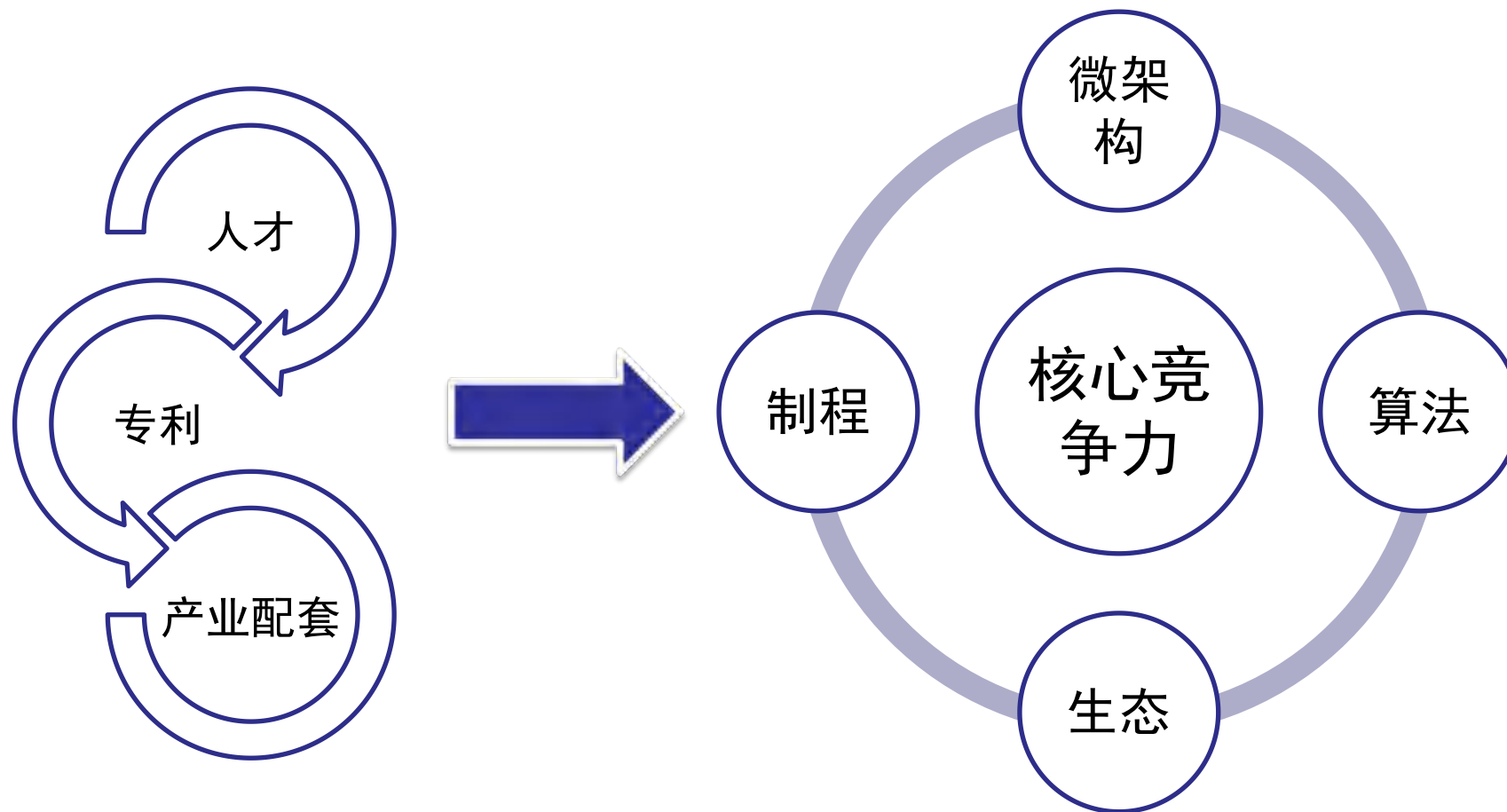
英伟达架构两年升级一次



1. 由专用走向通用，GPU赛道壁垒高筑

1.22 多方面构建的高壁垒

➤ GPU的体系结构与算法是各个公司的核心机密。

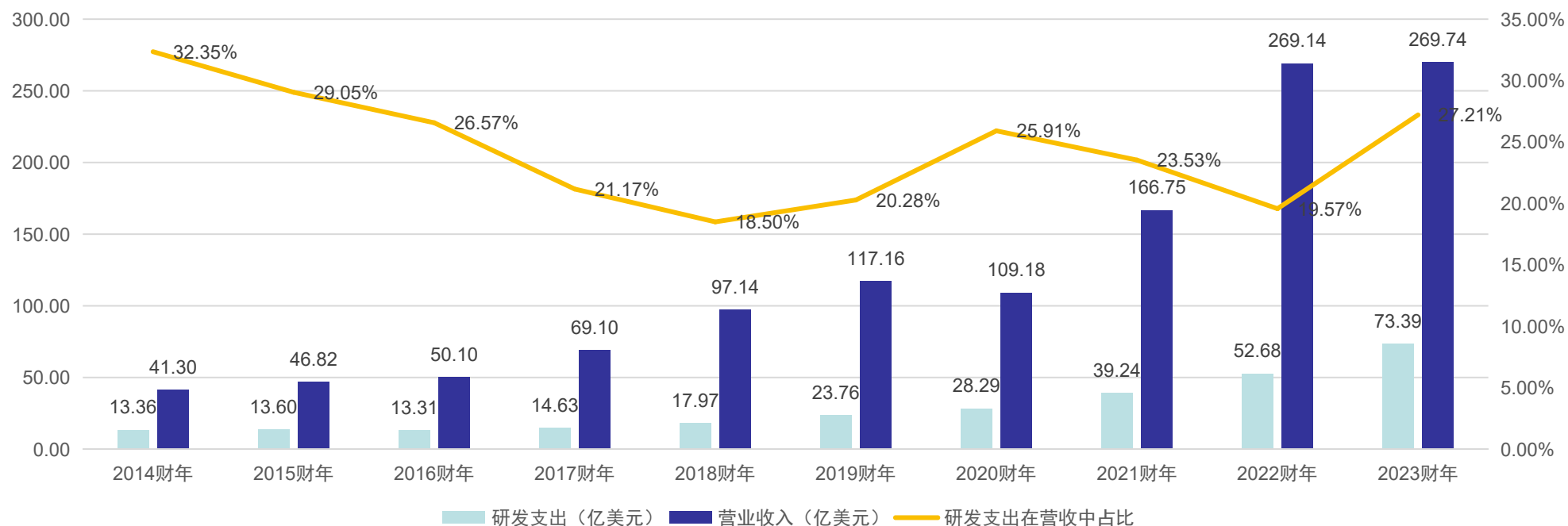


1. 由专用走向通用，GPU赛道壁垒高筑

1.23 人才与研发投入，以英伟达为例

- 根据英伟达官网报告显示，公司共有22,500名员工；根据公司最新财年的年报显示，公司职员中有80%属于技术人员，有50%的具备高等学历。
- 根据英伟达最新的公告显示，整个2023财年，英伟达总收入269.7亿美元，与前一个财年几乎持平，研发支出高达73.39亿美元，研发支出在营收中占比高达27.21%。截至2023财年，公司十年间共计研发支出高达290.23亿美元。

图：近十个财年英伟达营收（亿美元）、研发支出（亿美元）、研发支出在营收中占比



1. 由专用走向通用，GPU赛道壁垒高筑

1.24 国外厂商多年间构筑了庞大的专利池

- 根据万雪佼、徐步陆在2017年发布的《图形处理器（GPU）专利态势研究》的内容显示，全球GPU专利呈现以下几大特点：
- 1、从全球专利公开国看，GPU专利全球布局重心在美国。其中超过总数80%的5459个专利家族有美国专利，剩余世界五大大专专利局的中日欧韩分布也排名靠前，均有超过10%专利家族有该国专利布局。从各国公开趋势来看，在美国、中国、韩国专利布局比重呈逐年上升趋势；
- 2、从专利权人分布看，全球GPU技术领域专利数量排名前20的公司占有全球70%的GPU专利，GPU专利技术相对集中。排名靠前的公司以美国居多，其次是英国（ARM和Imagination Tech）。日本游戏公司索尼电脑娱乐公司和任天堂公司也有少量GPU专利。除台湾VIA公司外，排名前100的没有中国专利权人。GPU技术领域全球专利家族持有数量排名前三的分别是NVIDIA、Intel和AMD。其中NVIDIA持有专利数量占全球总量的近20%。
- 3、我国原生GPU企业，历史短，专利数量极少且布局仅在国内。

1. 由专用走向通用，GPU赛道壁垒高筑

1.25 英伟达全栈布局构筑强大生态

- 2006年，NVIDIA推出CUDA，这是一种用于通用GPU计算的革命性架构。CUDA的存在使得开发者使用GPU进行通用计算的难度大幅降低，使得开发者可以相对简单有效地对英伟达GPU芯片进行编程，使科学家和研究人员能够利用GPU的并行处理能力来应对最复杂的计算挑战。
- 芯片是算力基础，但要充分发挥其性能，必须构建完备的系统软件底层库，英伟达构建了从底层系统软件、驱动软件、平台到上层的应用框架。此外，英伟达提供全面的算法库，几乎全部开源。

图：英伟达提供全堆栈的AI、HPC软件

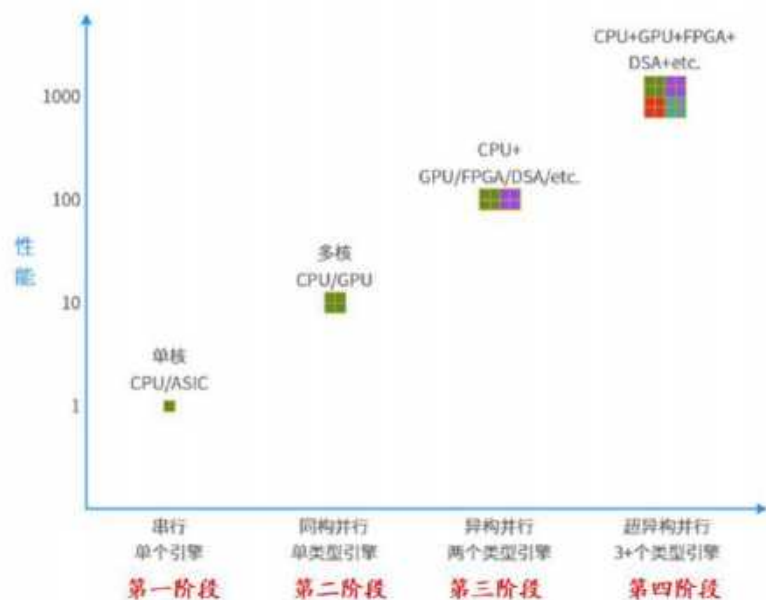


1. 由专用走向通用，GPU赛道壁垒高筑

1.26 走向异构，海外厂商横向布局不断

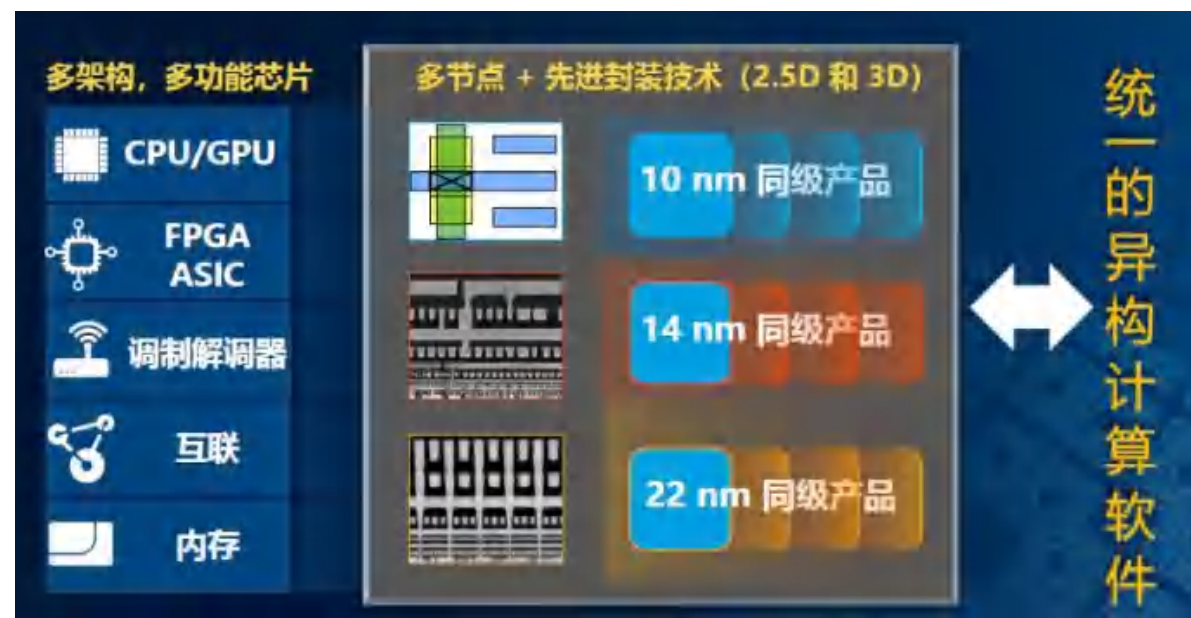
- 异构计算主要是指使用不同类型指令集和体系架构的计算单元组成系统的计算方式。异构计算近年来得到更多关注，主要是因为通过提升CPU时钟频率和内核数量而提高计算能力的传统方式遇到了散热和能耗瓶颈。而与此同时，GPU等专用计算单元虽然工作频率较低，具有更多的内核数和并行计算能力，总体性能-芯片面积比和性能-功耗比都很高，却远远没有得到充分利用。云和边缘计算的数据中心、自动驾驶等超级终端领域都是典型的复杂计算场景，这类型场景的计算平台都采用了大算力芯片，也是异构计算最重要的落地场景。2015年12月29日，英特尔公司宣布完成对Altera公司的收购，Altera公司是FPGA(可编程逻辑阵列)技术的领先提供商。2022年2月14日，AMD宣布以全股份交易(all-stock transaction)方式完成对赛灵思(Xilinx)的收购。英伟达自研CPU，在2022 GTC大会上，NVIDIA宣布推出首款面向AI基础设施和高性能计算的基于Arm Neoverse架构的数据中心专属CPU——Grace CPU超级芯片。面向未来，海外大厂横向布局不断。

大算力芯片走向异构



资料来源：极术社区，华金证券研究所绘制

超异构的三大要素



资料来源：《AI计算迈入超异构时代》宋继强，华金证券研究所绘制

01 由专用走向通用，GPU赛道壁垒高筑

02 产业化路径显现，全球AI竞赛再加速

- 2.1 AI技术赋能实体经济面临的瓶颈
- 2.2 ChatGPT的破圈
- 2.3 ChatGPT的成功离不开预训练大模型
- 2.4 预训练模型的发展历程
- 2.5 Transformer架构成主流
- 2.6 自监督学习与Transformer的结合
- 2.7 大模型的突现能力
- 2.8 参数量爆发式增长的ChatGPT
- 2.9 预训练大模型，第三波AI发展的重大拐点
- 2.10 生成式AI、边缘AI技术即将步入成熟期
- 2.11 大模型是大算力和强算法结合的产物
- 2.12 AI芯片三剑客
- 2.13 训练端GPU担纲
- 2.14 数据中心迈入“高算力”时代，兵家必争
- 2.15 英伟达数据中心业务快速增长
- 2.16 自动驾驶研发两大商业路线
- 2.17 自动驾驶实现的两种技术路线
- 2.18 单车智能化推动算力升级加速
- 2.19 自动驾驶具备广阔市场前景

03 全维智能化大时代，国产算力行则必至

04 建议关注

05 产业相关

06 风险提示

2. 产业化路径显现，全球AI竞赛再加速

2.1 AI技术赋能实体经济面临的瓶颈

- 过去，绝大部分人工智能企业和研究机构遵循算法、算力和数据三位一体的研究范式，即以一定的算力和数据为基础，使用开源算法框架训练智能模型。而这也导致了当前大部分人工智能处于“手工作坊式”阶段，面对各类行业的下游应用，AI逐渐展现出碎片化、多样化的特点，也出现了模型通用性不高的缺陷。这不仅是AI技术面临的挑战，也限制了AI的产业化进程。随着人工智能赋能实体经济进入深水区，企业通常面临数据资源有限、算力投资难度大、模型泛化能力差、高水平人才稀缺的发展瓶颈。

人工智能发展的瓶颈问题

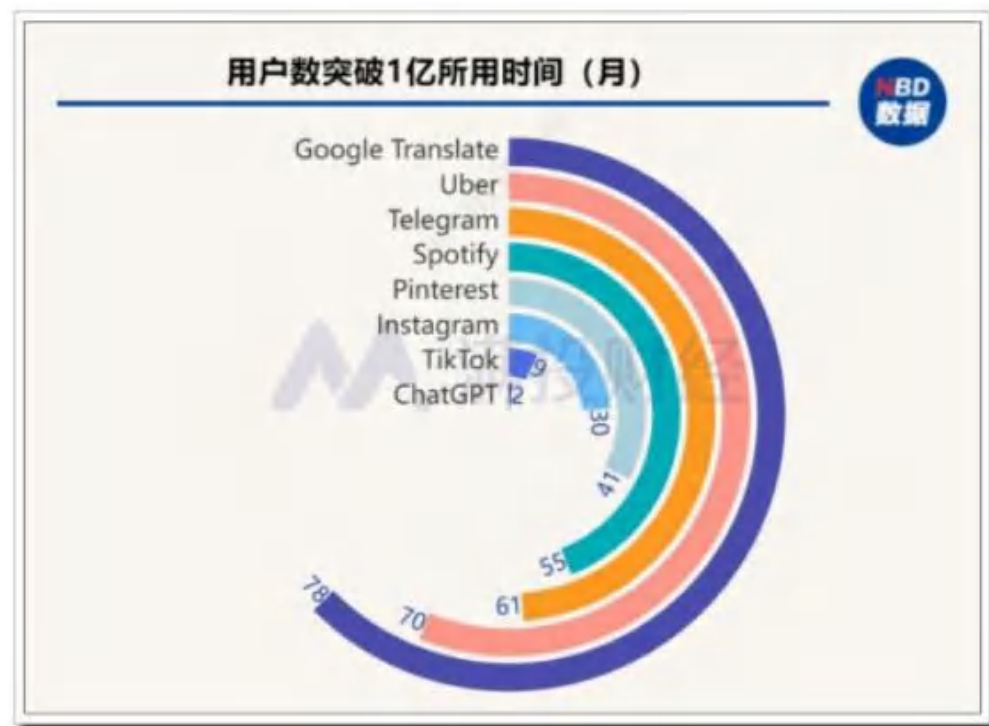


2. 产业化路径显现，全球AI竞赛再加速

2.2 ChatGPT的破圈

- 聊天生成型预训练变换模型（Chat Generative Pre-trained Transformer）简称ChatGPT，是OpenAI开发的人工智能聊天机器人程序，于2022年11月推出，上线两个月后已有上亿用户。
- ChatGPT目前仍以文字方式互动，而除了可以用人类自然对话方式来互动，还可以用于甚为复杂的语言工作，包括自动生成文本、自动问答、自动摘要等多种任务。

ChatGPT突破1亿用户数所需时间对比



资料来源：满投财经，华金证券研究所

ChatGPT介绍



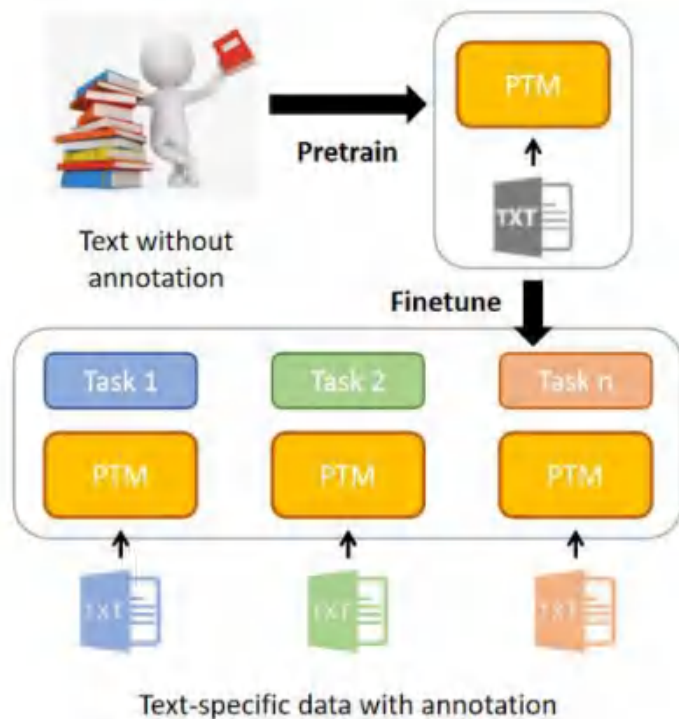
资料来源：cnbeta，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.3 ChatGPT的成功离不开预训练大模型

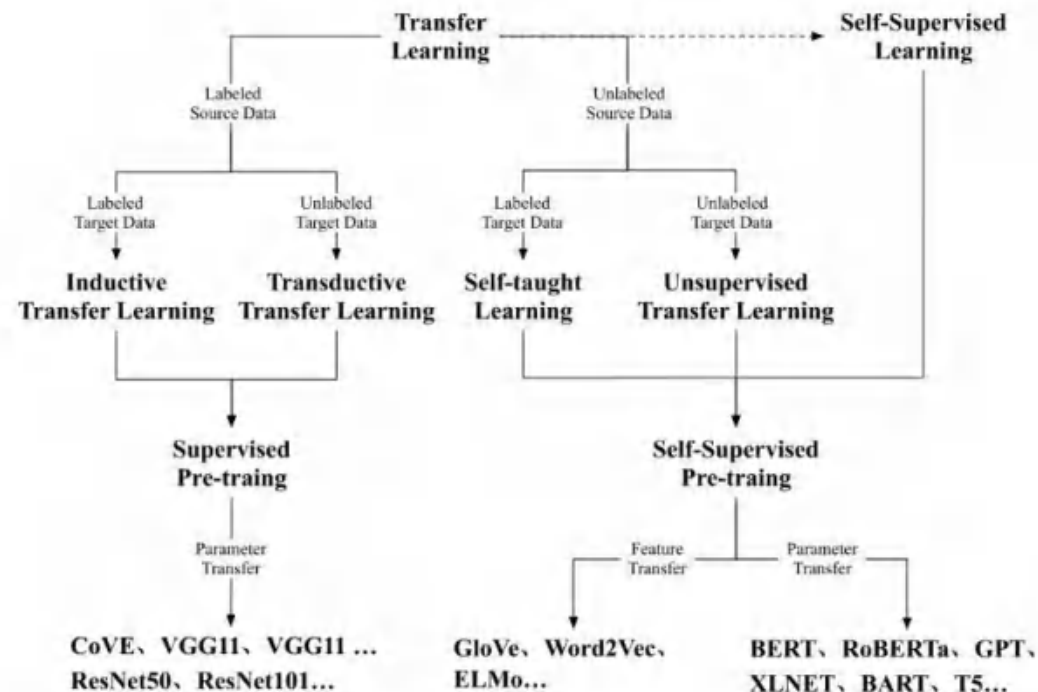
- 人工智能需要用大量的数据对其进行训练，理论上讲，投喂数据越多、数据质量越高，模型效果就会越好。而预训练（Pre-trained Models, PTMs），就是预先训练好的模型，可以帮助人们降低模型创建和训练的成本。预训练大模型需要深度学习的算法，也需要大的数据、大的算力，做自监督学习（模型直接从无标签数据中自行学习，无需标注数据），再面向不同的任务、在不同的应用场景里做少量任务数据进行迁移学习，进而应用于很多场景。
- ChatGPT能够实现当前的交互，离不开OpenAI在AI预训练大模型领域的积累。

NLP模型开发领域的标准范式“pretrain+finetune”



资料来源：Datawhale，华金证券研究所

预训练的起源与发展



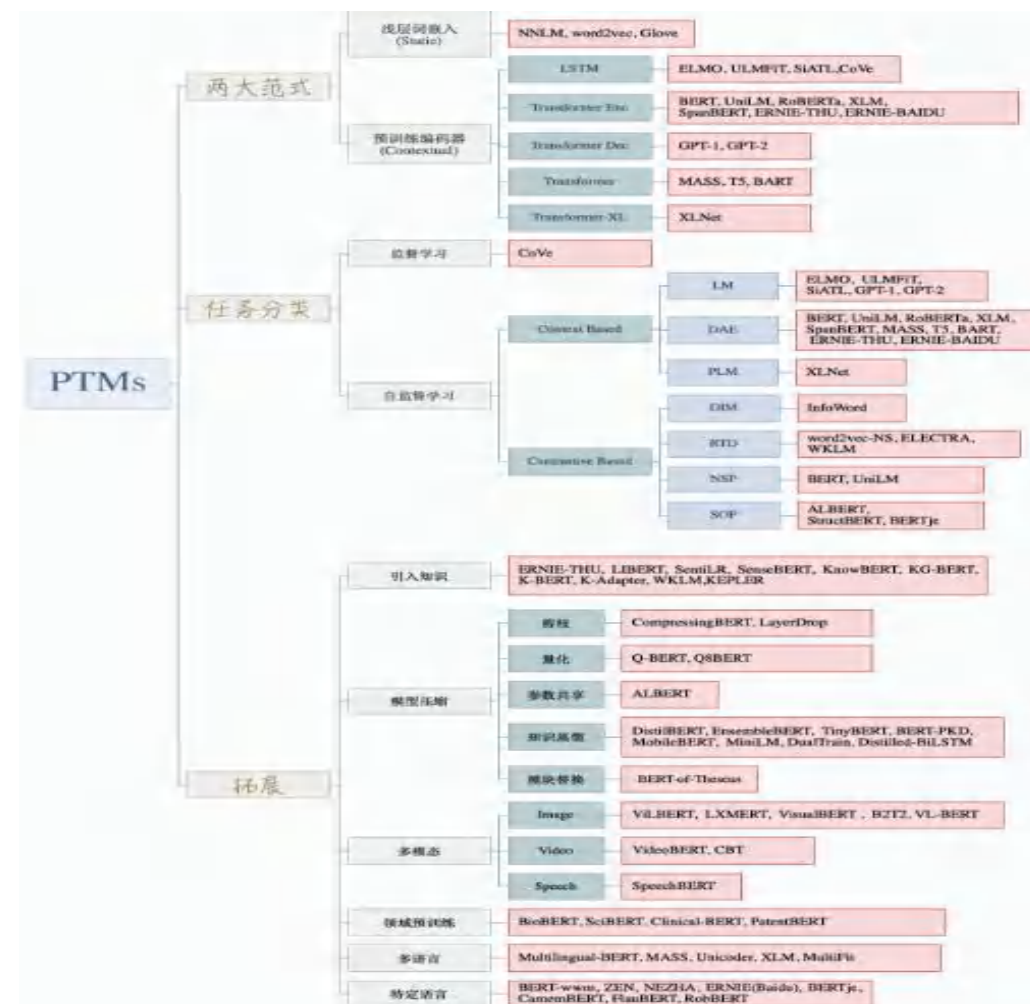
资料来源：阿里云开发者社区，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.4 预训练模型的发展历程

- 预训练的研究最早起源于迁移学习。迁移学习的核心思想，即运用已有的知识来学习新的知识，通俗来说就是将一个预训练的模型被重新用在另一个任务中。早期的预训练模型主要基于有标签数据。而在NLP领域，由于下游任务的多样性以及数据标注的复杂性，导致无法获得一个像ImageNet这样大规模的有标签数据，所以NLP领域尝试使用自监督学习的方法来获取预训练模型，自监督学习的主要思想就是利用文本间的内在联系为监督信号。
- 2017年出现的Transformer结构，给NLP领域预训练模型的发展带来了绝大的突破。Transformer的成功，也诱使CV领域加入了自监督预训练模型的赛道。如今，自监督预训练已经成为当前人工智能研究的重点，几乎所有的最新的PTM都是采用类Transformer结构与自监督学习的方法。

预训练模型的分类



资料来源：CSDN，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.5 Transformer架构成主流

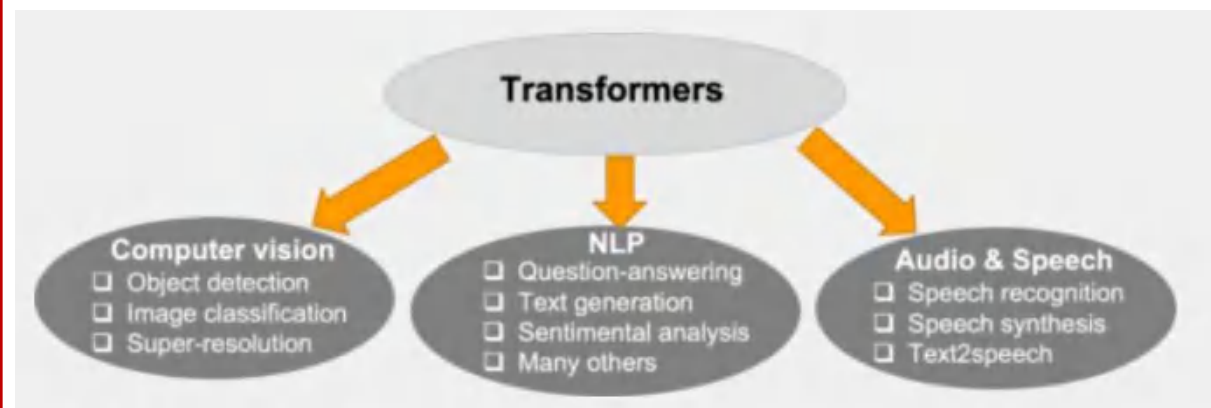
- 2017年，谷歌团队首先提出Transformer模型。该团队将Transformer概括为一句话：“Attention is All You Need.” 目前Transformer已经成为自然语言处理领域的主流模型，基于Transformer的预训练语言模型更是成为主流。除了NLP之外，Transformer也逐渐成为很多基于序列的语音应用的主流AI模型，在很多场景中已取代RNN/LSTM，比如自动语音识别、语音合成等等
- Transformer受欢迎的主要原因是其架构引入了并行化，它利用了强大的TPU和并行训练，从而减少了训练时间。

基于 Transformer 架构的 NLP 模型规模



资料来源：新浪，华金证券研究所

基于 Transformer 架构的应用

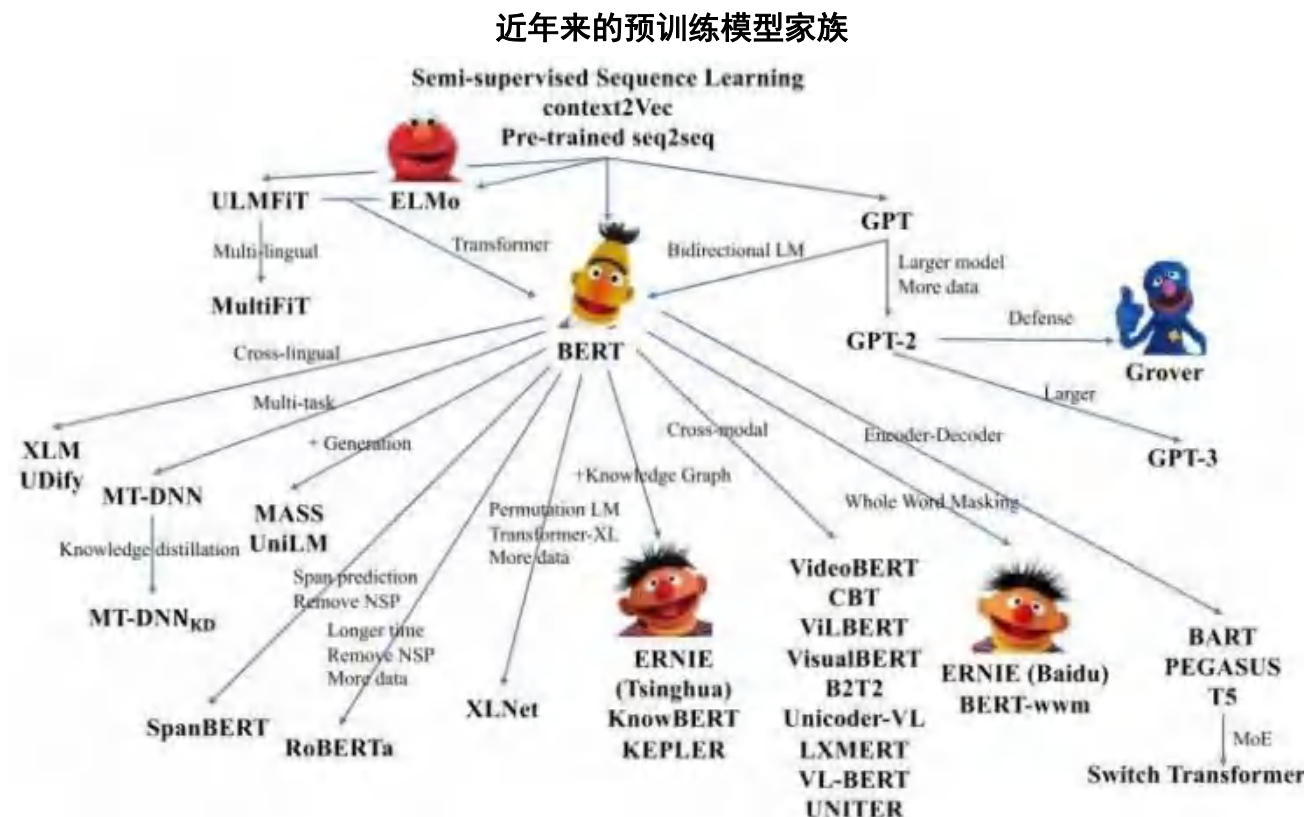


资料来源：新浪，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.6 自监督学习与Transformer的结合

- 自监督学习是从无标注数据中提取知识的一种手段，它能够利用数据本身的隐藏信息作为监督，和无监督有非常相似的设置。由于自然语言很难标注且又存在大量未标注的句子，所以NLP领域的预训练模型主要致力于自监督学习，进而大大促进了NLP领域的发展。
- 预训练模型成功的关键是自监督学习与Transformer的结合，具有代表性的工作是GPT和BERT系列模型。后续的其他预训练模型都是这两个经典模型的变体。

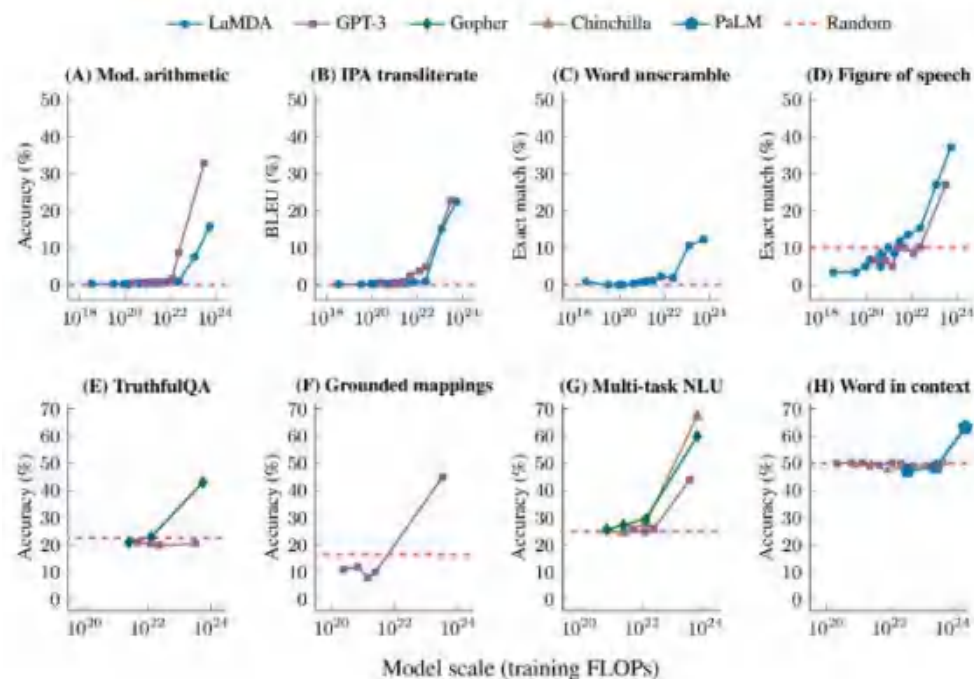


2. 产业化路径显现，全球AI竞赛再加速

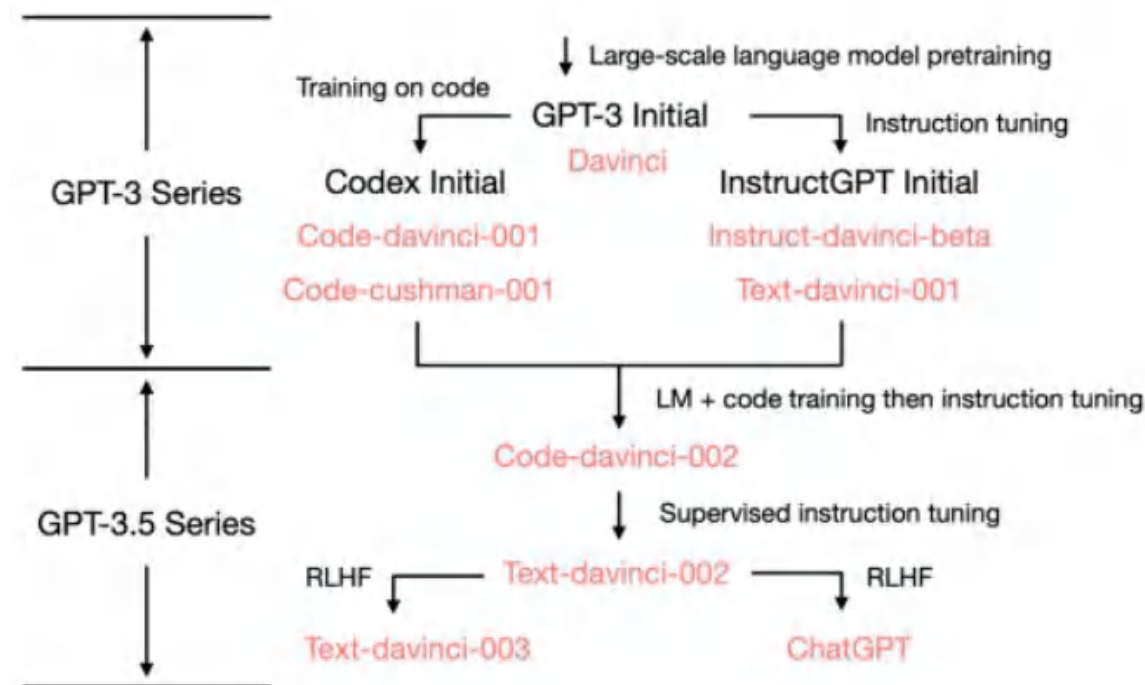
2.7 大模型的突现能力

- 当扩展大型语言模型时，偶尔会出现一些较小模型没有的新能力，这种类似于「创造力」的属性被称作「突现」能力。GPT-3的论文表明，语言模型执行多位数加法的能力对于从100M到13B参数的模型具有平坦的缩放曲线，近似随机，但会在一个节点造成性能的飞升。
- 初代GPT-3展示了三个重要能力：语言生成、上下文学习、世界知识。基本上三种能力都来自于大规模预训练：在有3000亿单词的语料上预训练拥有1750亿参数的模型。

大模型的「突现」能力



GPT-3.5 的进化树



资料来源：《Emergent Abilities of Large Language Models》Jeff Dean等，华金证券研究所

资料来源：《拆解追溯GPT-3.5各项能力的起源》符尧，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.8 参数量爆发式增长的ChatGPT

- GPT模型的训练需要超大的训练语料，超多的模型参数以及超强的计算资源。2018年，OpenAI发布了生成式预训练语言模型GPT，可用于生成文章、代码、机器翻译、问答等各类内容。GPT的参数量1.17亿，预训练数据量约5GB；2019年2月份发布的GPT-2的参数量15亿，预训练数据量40GB；2020年5月发布的GPT-3的参数量高达1,750亿，预训练数据量高达45TB。

图：ChatGPT与GPT 1-3的技术对比

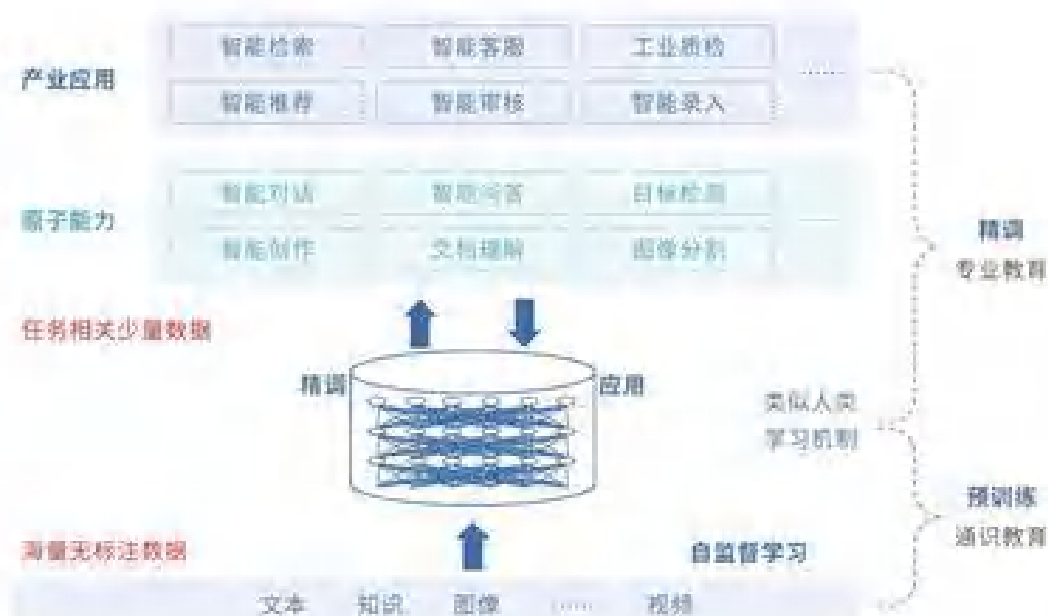


2. 产业化路径显现，全球AI竞赛再加速

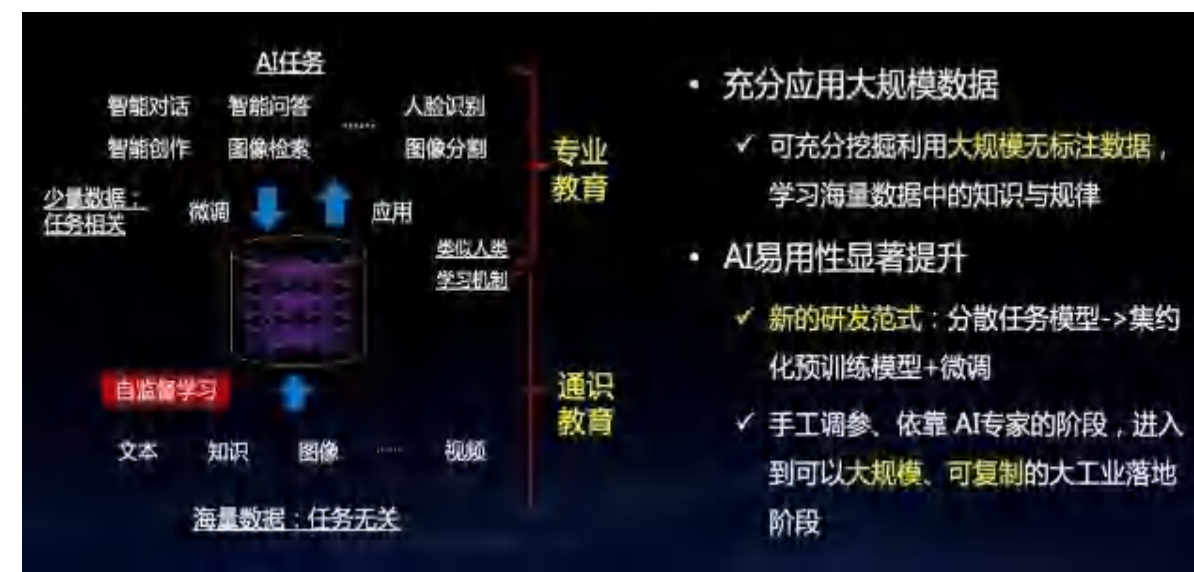
2.9 预训练大模型，第三波AI发展的重大拐点

- 深度学习时代，为了充分训练深层模型参数并防止过拟合，通常需要更多标注数据喂养。在NLP领域，标注数据更是一个昂贵资源。预训练从大量无标注数据中进行预训练使许多NLP任务获得显著的性能提升。
- 大模型通常是在大规模无标注数据上进行训练，学习出一种特征和规则。基于AI大模型进行应用开发时，将大模型进行微调（在下游特定任务上的小规模有标注数据进行二次训练）或者不进行微调，就可以完成多个应用场景的任务，实现通用的智能能力。预训练大模型在海量数据的学习训练后具有良好的通用性和泛化性，用户基于大模型通过零样本、小样本学习即可获得领先的效果，同时“预训练+精调”等开发范式，让研发过程更加标准化，显著降低了人工智能应用门槛，成为AI走向工程化应用落地的重要手段。

训练大模型“预训练+精调”模式



预训练大模型的基本原理

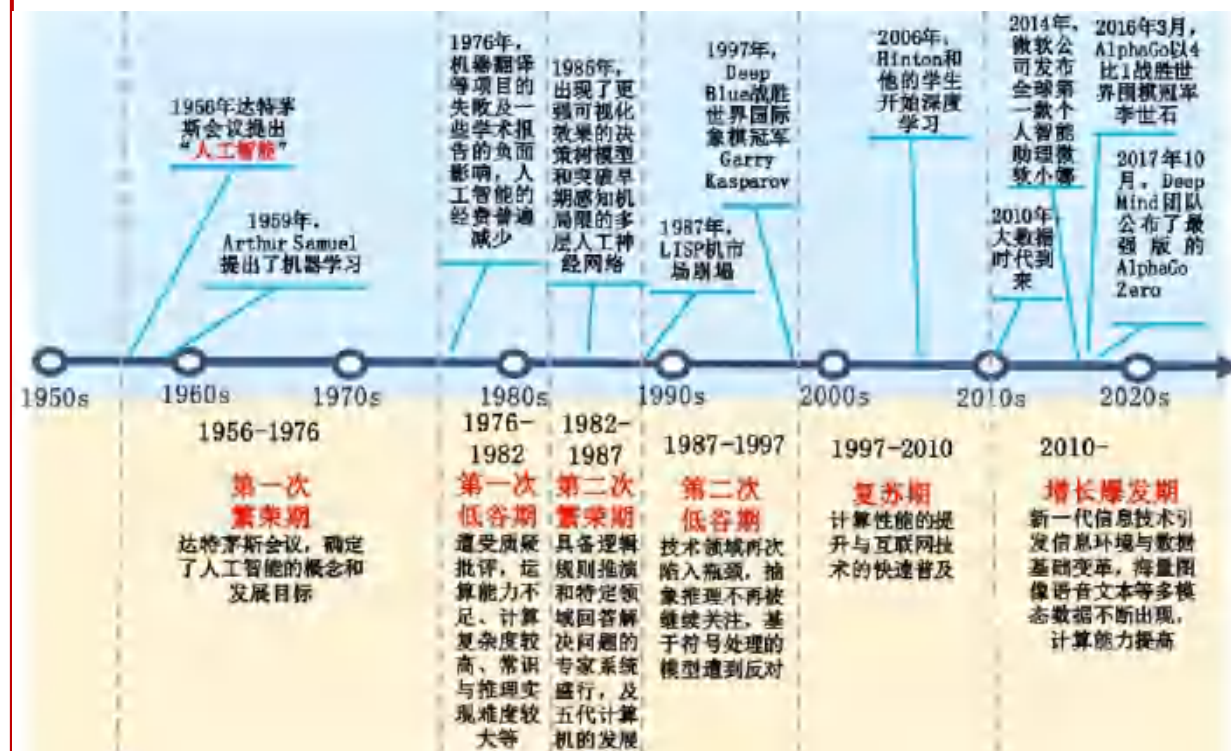


2. 产业化路径显现，全球AI竞赛再加速

2.10 生成式AI、边缘AI技术即将步入成熟期

- 根据Gartner发布的2022年Gartner人工智能（AI）技术成熟度曲线（Hype Cycle™）显示，在多项人工智能技术中，生成式AI、合成数据、边缘AI等当下均处于期望膨胀期，预计2-5年达到高峰期。

人工智能发展历程



资料来源：《人工智能标准化白皮书》，华金证券研究所

人工智能技术成熟度曲线

2022 年 Gartner 人工智能技术成熟度曲线

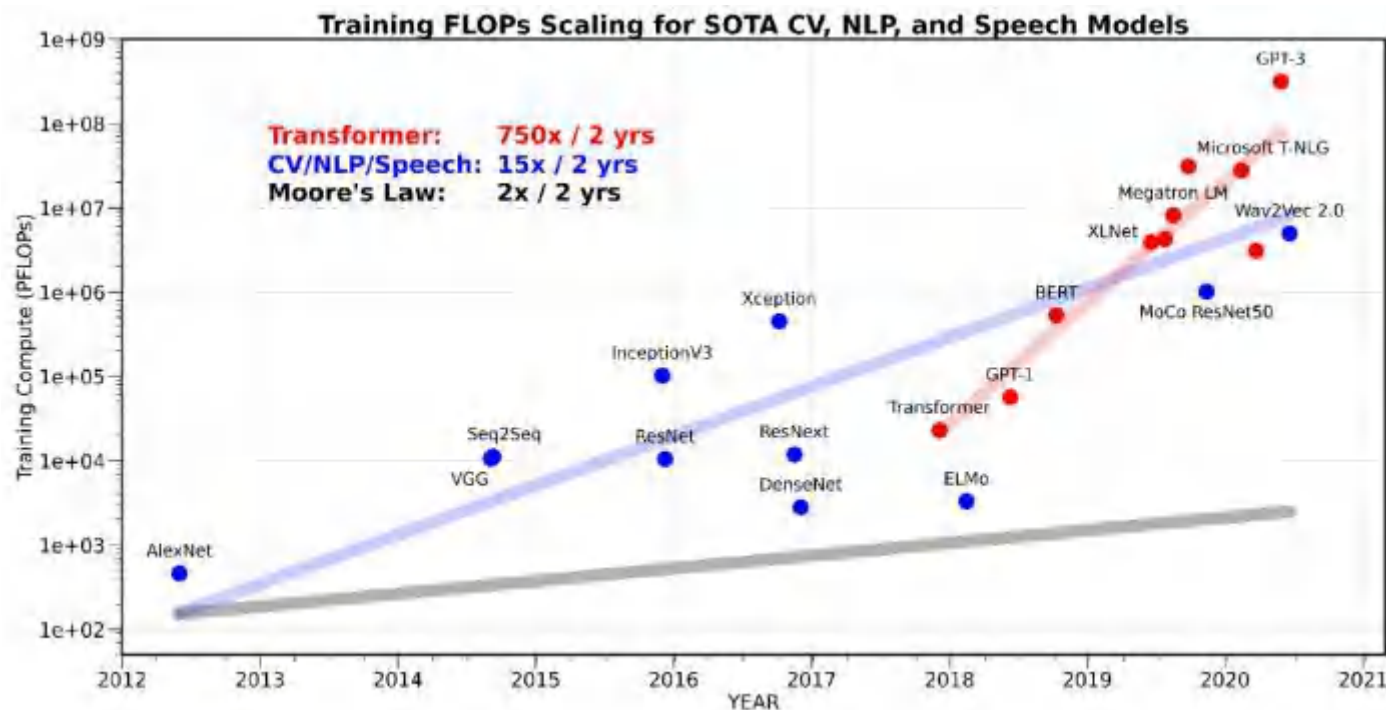


资料来源：Gartner，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.11 大模型是大算力和强算法结合的产物

- ChatGPT等AI应用需要基于大量模型训练，以GPT-3模型为例，其存储知识的能力来源于1750亿参数，训练所需的算力高达3650PFLOPS-day。据Lambda实验室测算，如果采用英伟达V100 GPU和当时最便宜的云服务进行计算，GPT-3训练一次需要355个GPU年（一块GPU运行355年的运算量）、花费460万美元。
- 美国市场研究机构TrendForce在2023年3月1日的报告中测算称，处理1800亿个参数的GPT-3.5大模型，需要的GPU芯片数量高达2万枚。未来GPT大模型商业化所需的GPU芯片数量甚至超过3万枚。在2022年11月，英伟达在官网公告中提到，微软Azure上部署了数万枚A100/H100高性能芯片。这是第一个采用英伟达高端GPU构建的大规模AI算力集群。



资料来源：腾讯云，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.12 AI芯片三剑客

➤ 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。

图：AI芯片三剑客

	GPU	FPGA	ASIC
特性	图形处理器，图像和图形相关运算工作的微处理器	现场可编程门阵列，可以重构电路的芯片，一种硬件可重构的体系结构	专用集成电路，应特定用户要求和特定电子系统需要而设计制造的集成电路
性能	较高	较低	高
灵活性	较低	较高	低
成本	较高	低	高
功耗	高	较低	低
同构性	较低	较高	低
优点	可以支撑大量数据的并行计算，适合对数据密集型的应用进行计算和处理	可无限次编程，延时性比较低，同时拥有流水线并行和数据并行、灵活性高	功耗低，适合量产
缺点	功耗高，管理控制能力弱，不具备可编程性	开发难度大、只适合定点运算、价格比较昂贵	研发成本高昂，开发周期长，灵活性低

资料来源：华金证券研究所整理

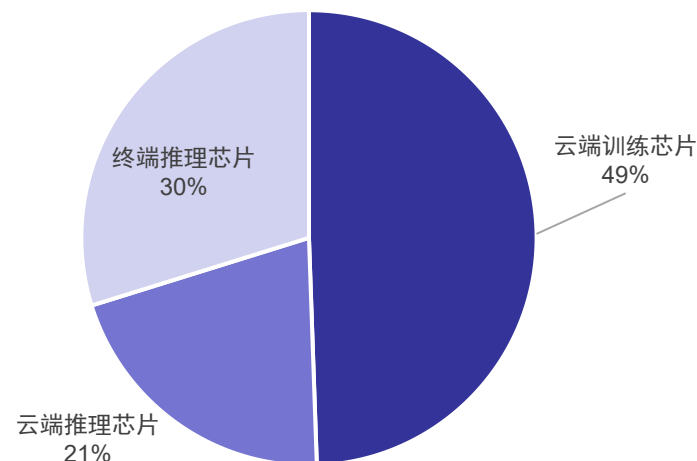
请仔细阅读在本报告尾部的重要法律声明

2. 产业化路径显现，全球AI竞赛再加速

2.13 训练端GPU担纲

- 虽然AI芯片目前看有三大类，但是基于几点原因，我们判断GPU仍将是主流：1、Transformer架构是最近几年的主流，该架构最大的特点之一就是能够利用分布式GPU进行并行训练，提升模型训练效率；2、ASIC的算力与功耗虽然看似有优势，但考虑到AI算法还是处于一个不断发展演进的过程，用专用芯片部署会面临着未来算法更迭导致芯片不适配的巨大风险；3、英伟达强大的芯片支撑、生态、算法开源支持。

2018年全球AI芯片市场结构

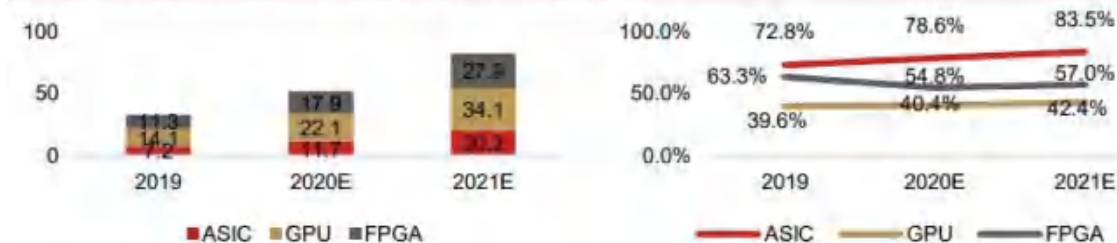


资料来源：赛迪顾问，华金证券研究所

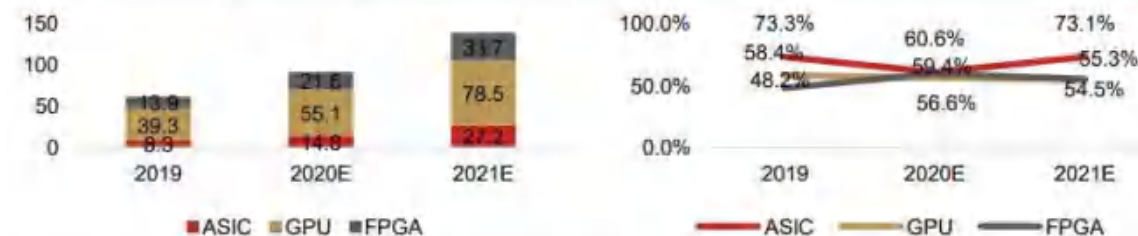
请仔细阅读在本报告尾部的重要法律声明

不同场景对于不同类型AI芯片的占比预测

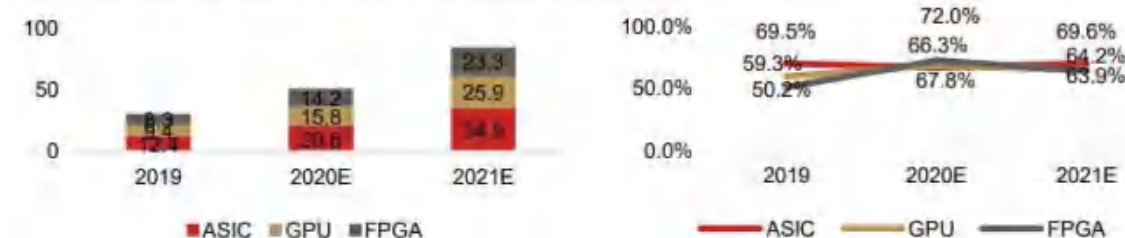
中国云端推断芯片市场结构（亿元）及增长率



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



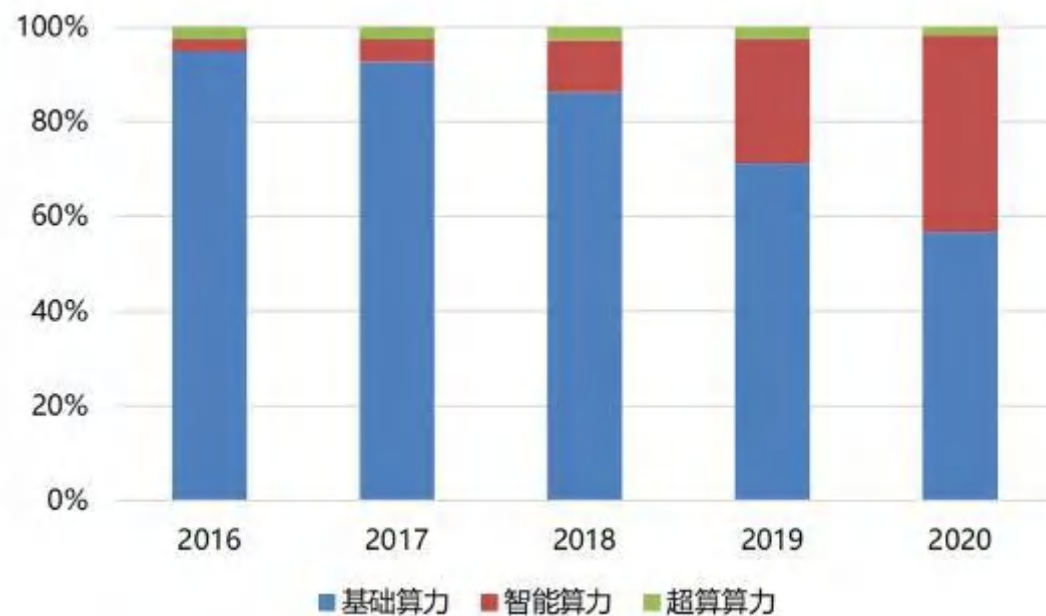
资料来源：赛迪顾问，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.14 数据中心迈入“高算力”时代，兵家必争

- 工信部发布的《新型数据中心发展三年行动计划（2021-2023年）》明确了算力内涵并引入测算指标FLOPS，对数据中心发展质量进行评价，指出到2023年底，总算力规模将超过200 EFLOPS，高性能算力占比将达到10%，到2025年，总算力规模将超过300 EFLOPS。
- 由于GPU比CPU更适合处理企业数据中心和超大规模网络中AI和机器学习所需的许多计算，数据中心对GPU的需求是一个不断增长的机会。

2016-2020中国算力结构变化



资料来源：信通院，华金证券研究所

2020-2025年全球AI服务器行业市场规模及增速（单位：亿美元）



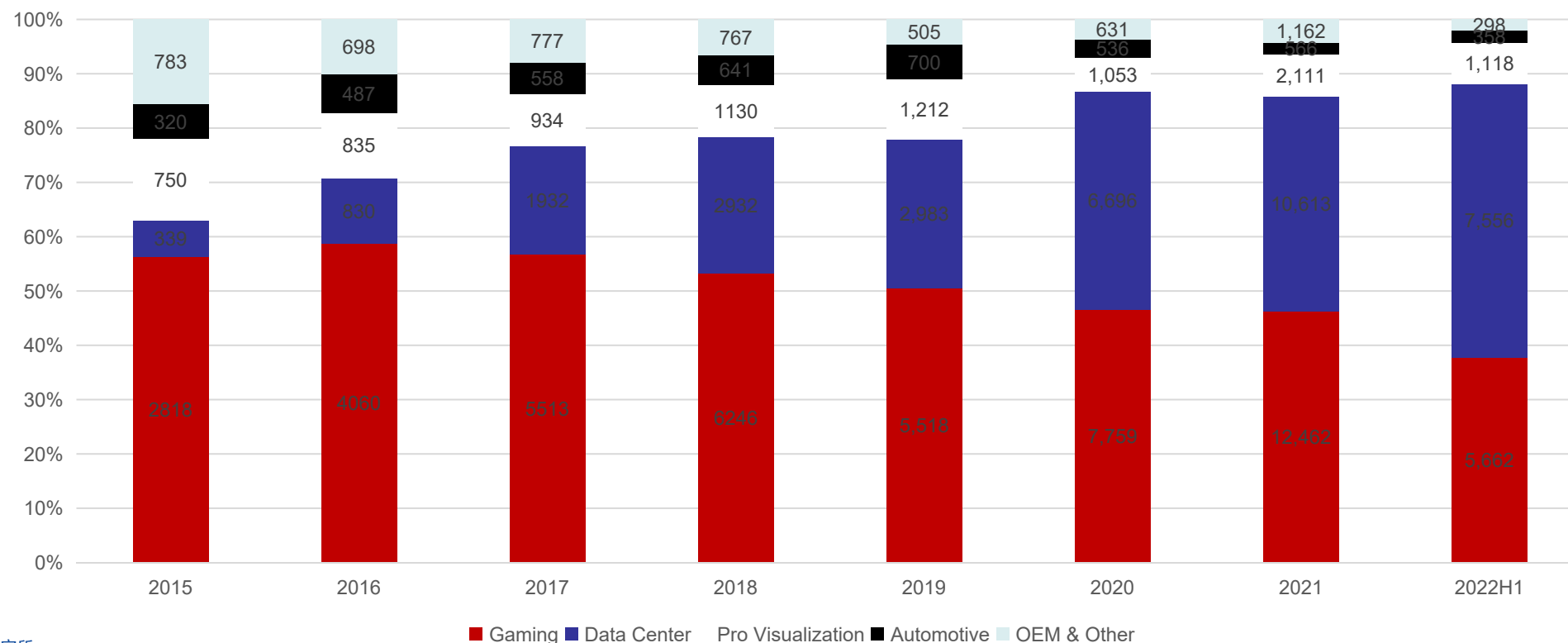
资料来源：华经产业研究院，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.15 英伟达数据中心业务快速增长

- 英伟达有四大产品线平台，包括游戏业务、数据中心、专业显示和汽车业务。2023财年第一季度，英伟达游戏业务收入较上年同比增长31%，环比增长6%；数据中心收入同比增长83%，环比增长15%，主要是由用于训练和推理的GPU销售所驱动的；专业显示的收入同比增长67%，环比下降3%；汽车收入同比下降10%，环比增长10%，同比下降由于汽车制造商供应限制等因素导致。

图：英伟达按下游市场划分销售占比（百万美元）



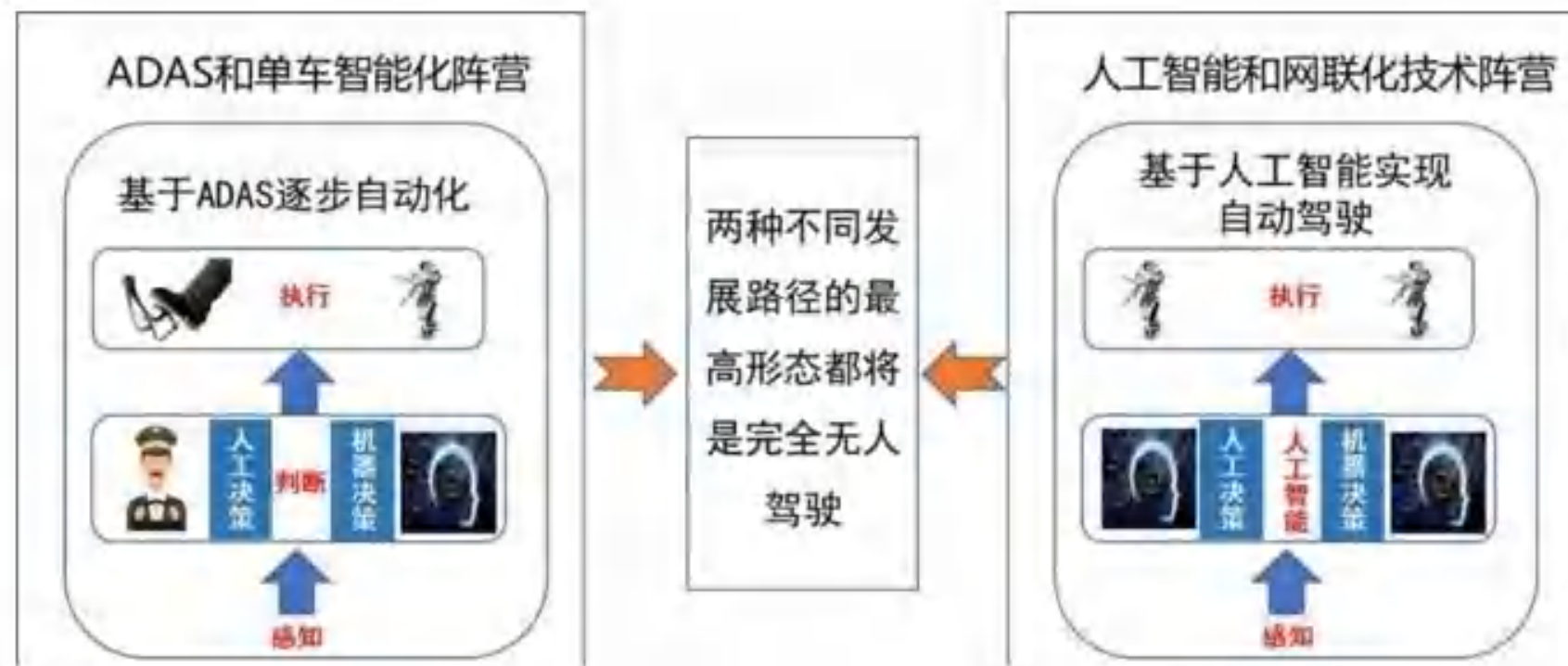
资料来源：wind，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.16 自动驾驶研发两大商业路线

- 自动驾驶研发有两大路线：以传统车企为代表的渐进式路线，从L1逐步升级到L5；以科技公司为代表的跨越式路线，跳过驾驶辅助系统，直接从高度自动驾驶L4系统切入，首先会在一些相对较易的商用场景率先落地。

汽车制造商和互联网企业的自动驾驶技术发展路径

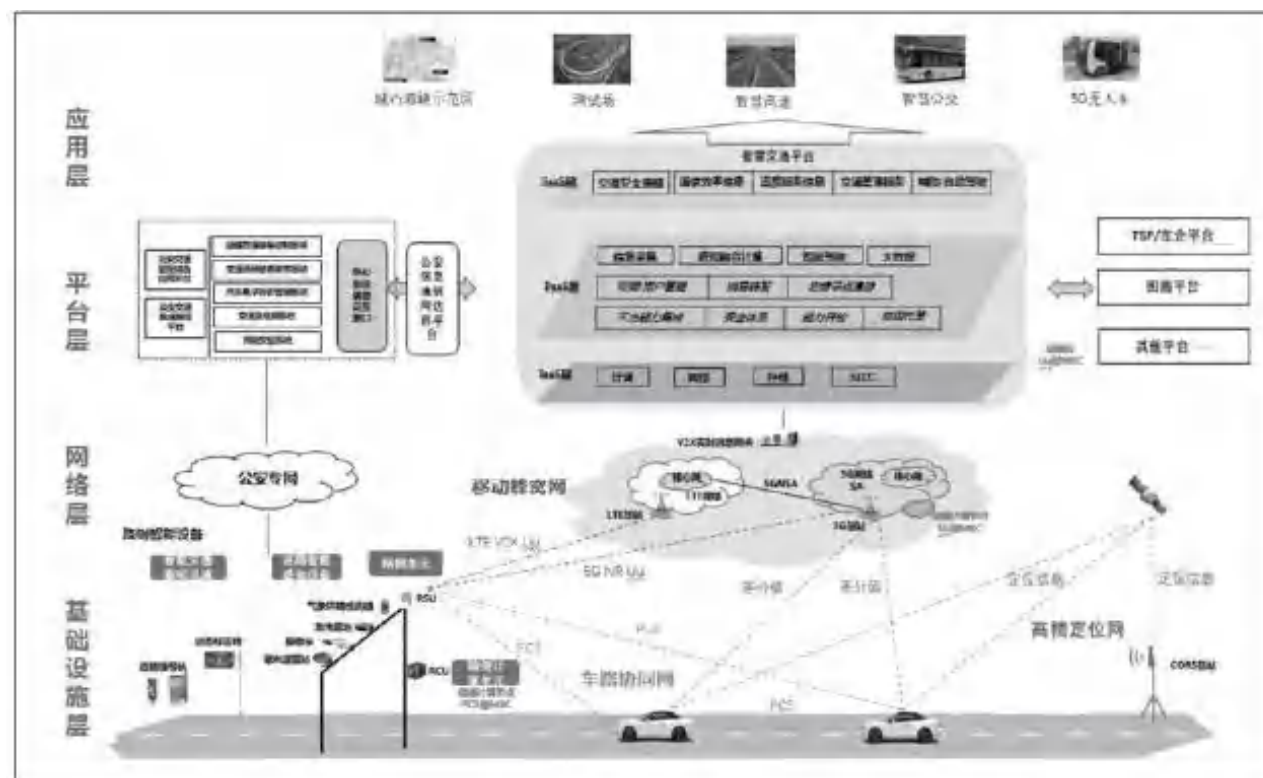


2. 产业化路径显现，全球AI竞赛再加速

2.17 自动驾驶实现的两种技术路线

- 从商业场景来看，实现的自动驾驶的路径主要有两条，一是单车智能，即通过摄像头、雷达等传感器以及高效准确的算法，赋予车辆自动驾驶的能力；二是车路协同，即主要通过5G、高精地图，来感知路况从而具备无人驾驶功能。
- 从当下技术角度来看，无论单车智能还是车路协同都存在不足之处，两者结合可以提升自动驾驶安全。但是从商业角度，车路协同需要大量的、长期的基础设施建设，车企目前主要还是选择单车智能的技术路线，而且这样也能满足对于自动驾驶技术的自主可控。

车路协同系统架构



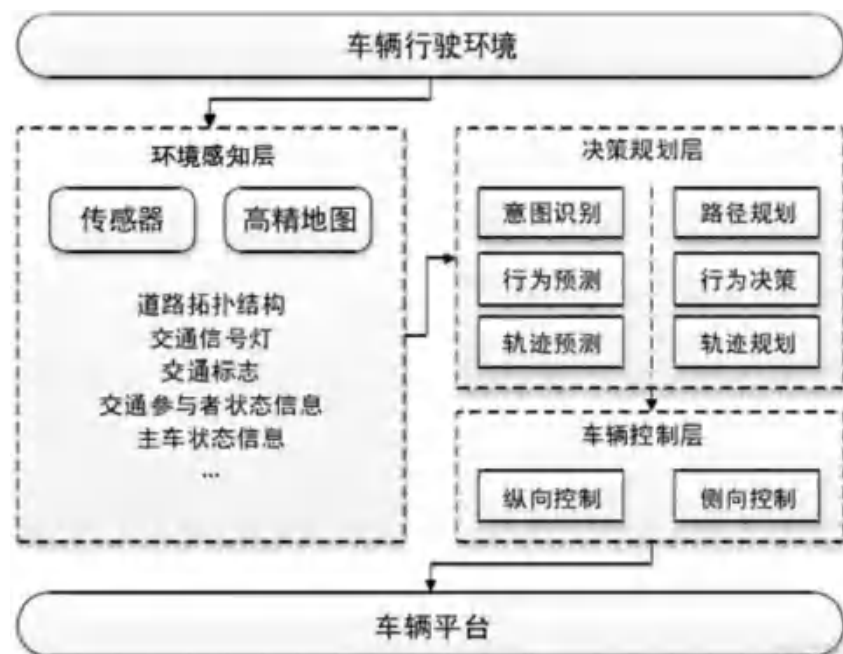
资料来源：中国新通信，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.18 单车智能化推动算力升级加速

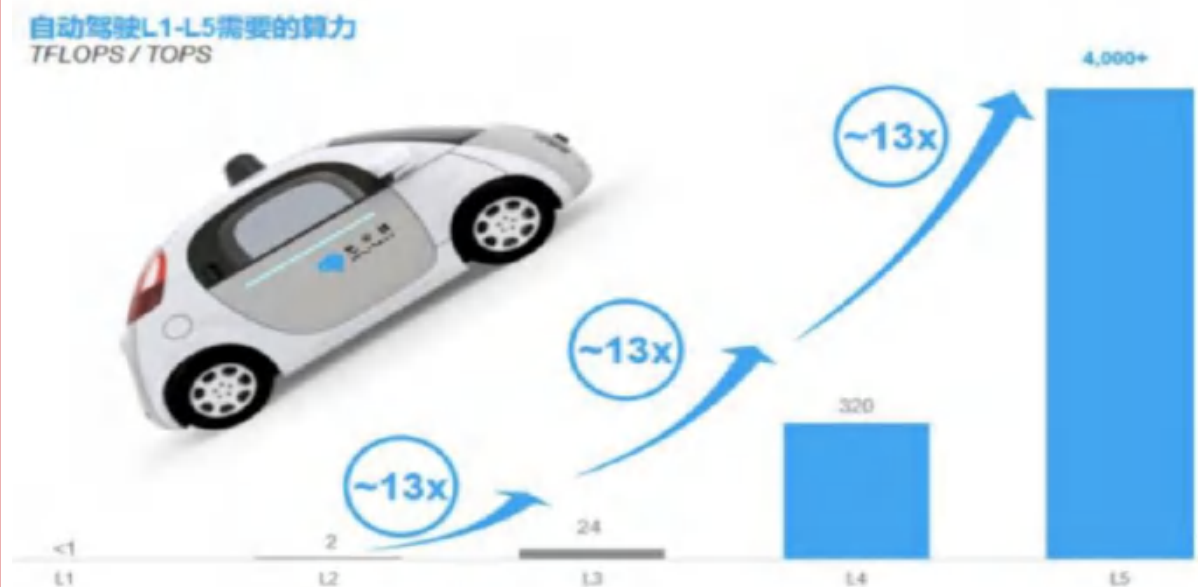
- 自动驾驶的完整流程包括感知、决策、控制，自动驾驶域算法一般也被划分感知算法、融合算法和执行算法三种。随着车辆自动驾驶等级的提升，对于车辆的主动性要求也大幅度提升，自动驾驶算法的难度就在于在所面对场景的多样性和复杂性。
- 由于不依赖人工智能算法实现基于机器的环境感知和规划决策，L1-L2级传统汽车不需要太大的车载算力，因此多采用小算力、微控制器的解决方案。从L2级开始，尤其是L3级以上的自动驾驶汽车需要装备大算力芯片支撑感知、决策算法的高效运行。根据地平线公司的预测，自动驾驶每提高一级，算力就增加一个数量级。L2级别大概需要2个TOPS的算力，L3需要24个TOPS，L4为320TOPS，L5为4000+TOPS。

自动驾驶核心技术



资料来源：51CTO，华金证券研究所

不同等级自动驾驶对于算力的需求



资料来源：地平线，华金证券研究所

2. 产业化路径显现，全球AI竞赛再加速

2.19 自动驾驶具备广阔市场前景

- 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%。中国仍将是全球汽车工业的主要市场，ICV的报告预计，到2026年中国汽车销售市场约占到全球的40.12%。

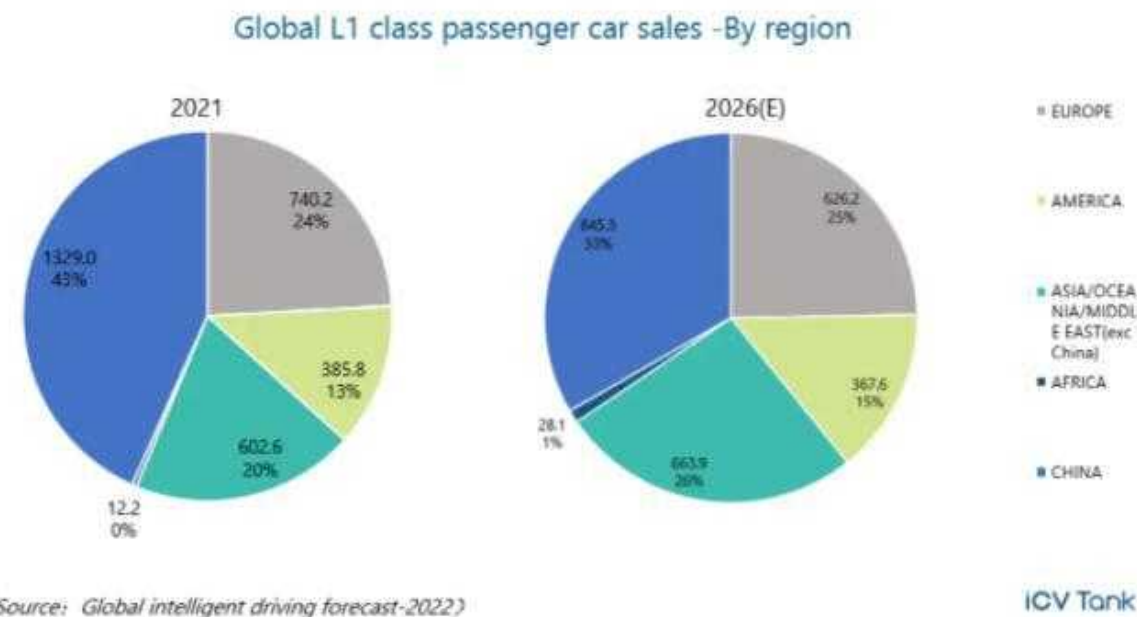
全球自动驾驶汽车出货量及增长率预测（2020-2024）



来源：IDC, 2020

资料来源：IDC，华金证券研究所

全球自动驾驶汽车出货量及增长率预测（2020-2024）



资料来源：ICV，华金证券研究所

01

由专用走向通用，GPU赛道壁垒高筑

02

产业化路径显现，全球AI竞赛再加速

03

全维智能化大时代，国产算力行则必至

- 3.1 全球数据中心负载任务量快速增长
- 3.2 全球计算产业投资空间巨大
- 3.3 预训练大模型对于GPU的需求
- 3.4 国内市场需求将保持高增长
- 3.5 云计算及云部署方式
- 3.6 不同云部署方式的市场占比
- 3.7 企业上云持续向细分行业渗透
- 3.8 从“资源上云”迈入“深度用云”
- 3.9 信创从试点走向推广
- 3.10 公有云主要参与厂商
- 3.11 云计算产业链
- 3.12 集成显卡与独立显卡市场份额
- 3.13 独立显卡英伟达一家独大
- 3.14 性能强大的H100
- 3.15 国产厂商两条发展路径：GPU和GPGPU
- 3.16 先求有，再求好
- 3.17 生态先兼容主流，未来将走向自建
- 3.18 国产之路已开启，部分国产GPU设计厂商列表
- 3.19 GPU发展离不开全球产业链的支撑
- 3.20 制程升级对于算力芯片性能提升具有较高贡献度
- 3.21 摩尔定律发展趋缓
- 3.22 Chiplet技术潜力大
- 3.23 Chiplet技术发展历程
- 3.24 行业巨头推动，产业加速落地
- 3.25 采用Chiplet技术的产品不断出现
- 3.26 算力两大演进方向：更大算力&更多样化应用
- 3.27 存量替代与增量成长并存
- 3.28 高吞吐量离不开高速传输
- 3.29 光通信前景可期

04

建议关注

05

产业相关

06

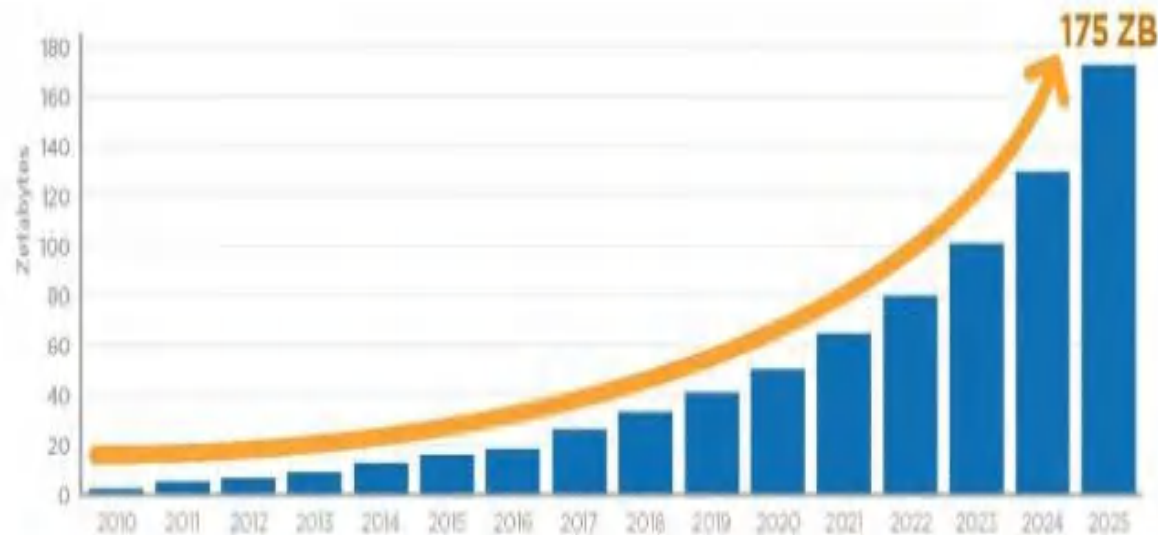
风险提示

3. 全维智能化大时代，国产算力行则必至

3.1 全球数据中心负载任务量快速增长

- 大规模张量运算、矩阵运算是人工智能在计算层面的突出需求，高并行度的深度学习算法在视觉、语音和自然语言处理等领域上的广泛应用使得计算能力需求呈现指数级增长。根据IDC的预测，从2018年至2025年，全球的数据增长量达到5倍以上，将从2018年的32ZB增至2025年的175ZB。中国将在2025年以48.6ZB的数据量及27.8%的占比成为全球最大的数据汇集地。
- 根据Cisco的预计，2021年全球数据中心负载任务量将超过2016年的两倍，从2016年的不到250万个负载任务量增长到2021年的近570万个负载任务量。

2010年至2025年全球数据量增长情况



资料来源：IDC，华金证券研究所

2016年-2021年数据中心负载任务量变化



资料来源：Cisco Global Cloud Index，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

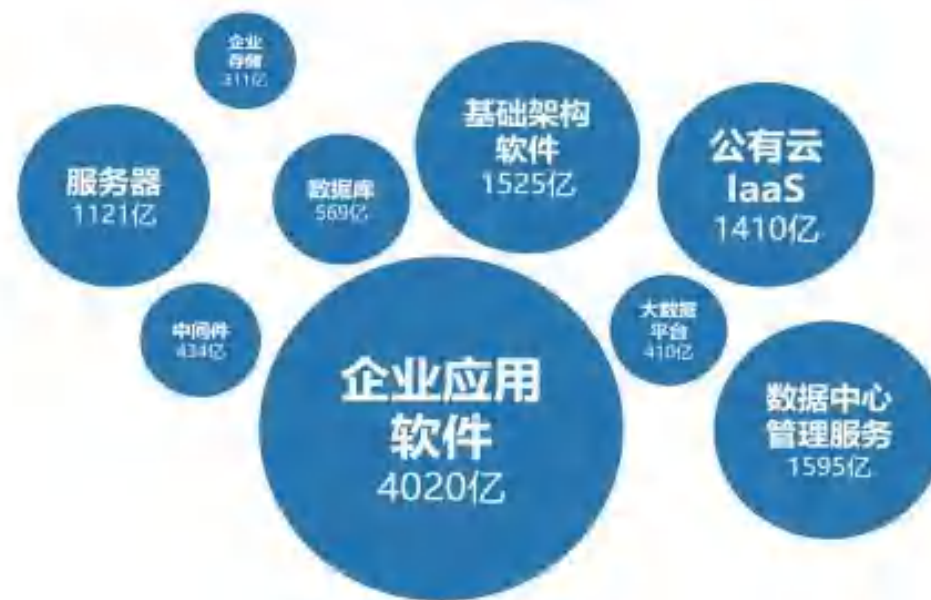
3.2 全球计算产业投资空间巨大

- 根据《鲲鹏计算产业发展白皮书》内容显示，数字化浪潮正重塑世界经济格局，数字经济正在成为全球可持续增长的引擎。IDC预测，到2023年数字经济产值将占到全球GDP的62%，全球进入数字经济时代。新的计算产业链将推动全球计算产业快速发展，带动全球数字经济走向繁荣。
- IDC预测，到2023年，全球计算产业投资空间1.14万亿美元。中国计算产业投资空间1043亿美元，接近全球的10%，是全球计算产业发展的主要推动力和增长引擎。

鲲鹏计算产业定义



2023年全球计算产业投资额（美元）







3. 全维智能化大时代，国产算力行则必至

3.3 预训练大模型对于GPU的需求

- 根据TrendForce的估计，2020年，GPT模型处理训练数据所需的GPU数量达到了20000左右。展望未来，GPT模型（或ChatGPT）商业化所需的GPU数量预计将达到30000个以上。这些均使用英伟达的A100 GPU作为计算基础。
- 根据中关村在线的新闻显示，目前英伟达A100显卡的售价在1.00~1.50万美元之间。英伟达还将A100作为DGX A100系统的一部分进行销售，该系统具有八块A100，两块AMD Rome 7742 CPU，售价高达199,000美元。

英伟达数据中心GPU对比

NVIDIA Data-Center GPUs Specifications				
VideoCardz.com	NVIDIA H100	NVIDIA A100	NVIDIA Tesla V100	NVIDIA Tesla P100
Picture				
GPU	GH100	GA100	GV100	GP100
Transistors	80B	54.2B	21.1B	15.3B
Die Size	814 mm²	828 mm²	815 mm²	610 mm²
Architecture	Hopper	Ampere	Volta	Pascal
Fabrication Node	TSMC N4	TSMC N7	12nm FFN	16nm FinFET+
GPU Clusters	132/114*	108	80	56
CUDA Cores	16896/14592*	6912	5120	3584
L2 Cache	50MB	40MB	6MB	4MB
Tensor Cores	528/456*	432	320	-
Memory Bus	5120-bit	5120-bit	4096-bit	4096-bit
Memory Size	80 GB HBM3/HBM2e*	40/60GB HBM2e	16/32 HBM2	16GB HBM2
TDP	700W/350W*	250W/300W/400W	250W/300W/450W	250W/300W
Interface	SXM5/PCIe Gen5	SXM4/PCIe Gen4	SXM2/PCIe Gen3	SXM/PCIe Gen3
Launch Year	2022	2020	2017	2016

资料来源：cnbeta，华金证券研究所

DGX A100组件



资料来源：foresine，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.4 国内市场需求将保持高增长

- 人工智能领域的应用目前处于技术和需求融合的高速发展阶段，在运算加速方面逐渐形成了以GPGPU解决方案为主的局面。根据前瞻产业研究院的数据，未来几年内，中国人工智能芯片市场规模将保持年均40%至50%的增长速度，到2024年，市场规模将达到785亿元。
- 聚集强大人工智能算力的智算中心是中国数字经济高速发展的产物，是一种新型的公共基础设施。国家已经出台了相关政策，并把智算中心列为“新基建”。

中国人工智能芯片市场规模（亿元）



资料来源：海光信息招股书，华金证券研究所

东数西算枢纽节点区域特点及布局思路

枢纽节点	集群	要求	重点/优先支持场景
京津冀	张家口数据中心集群	端到端单向时延：≤20ms PUE<1.25 上架率≥65%	海量规模数据的集中处理，工业互联网、金融证券、灾害预警、远程医疗、视频通话、人工智能推理等低时延、高频实时交互型的业务需求
长三角	长三角生态绿色一体化发展示范区数据中心集群		
粤港澳大湾区	深圳数据中心集群		
成渝	重庆数据中心集群		
贵州	贵安数据中心集群	PUE<1.2 上架率≥65%	后台加工、离线分析、存储备份等对实时性要求
内蒙古	和林格尔数据中心集群		
甘肃	庆阳数据中心集群		
宁夏	中卫数据中心集群	端到端单向时延：≤10ms	金融市场高频交易、VR/AR、超高清视频、车联网、联网无人机、智慧电力、智能工厂、智能安防等实时性要求高的业务需求
城市内部数据中心			

资料来源：前瞻产业研究院，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.5 云计算及云部署方式

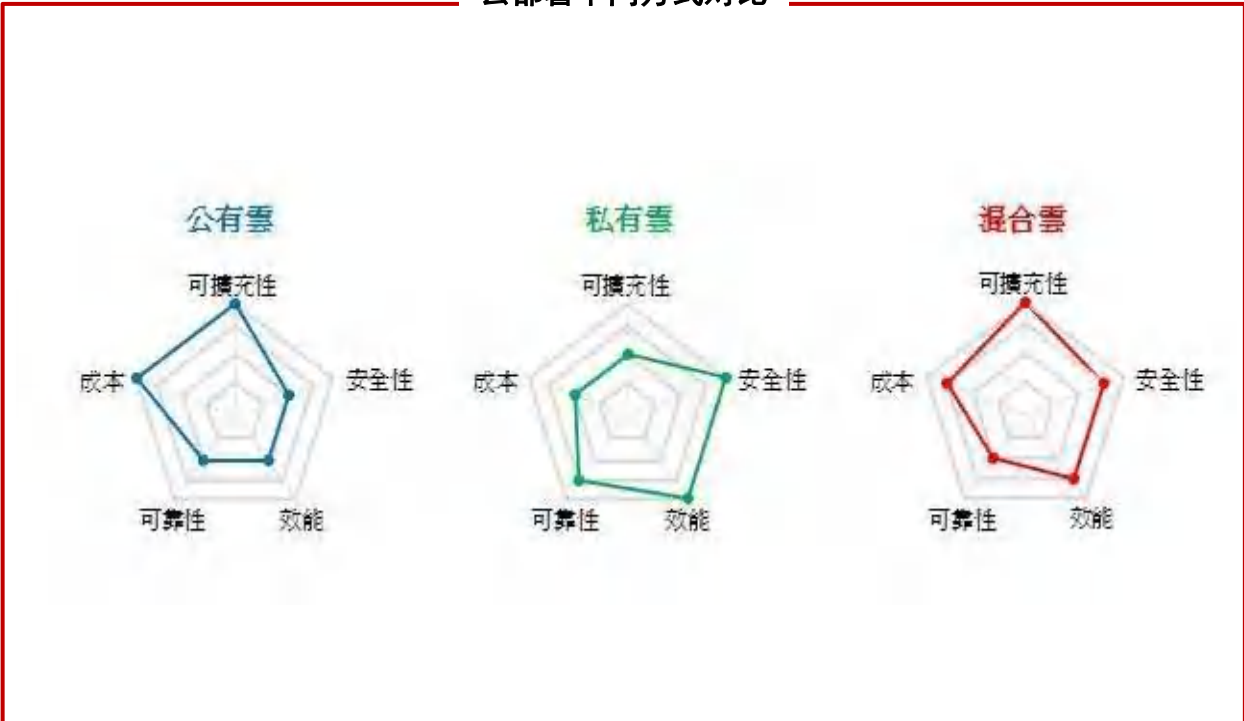
- 云计算广义的来说是厂商通过建立网络服务器集群，向各种不同类型客户提供在线软件服务、硬件租借、数据存储、计算分析等不同类型的服务。
- 云计算按后台位置主要分为公有云、私有云（含政务云）、混合云三种形态。目前国内主流公有云如阿里云、华为云、腾讯云等。私有云如政务云、金融云、工业云、物流云等。

云部署方式分类

云之家部署类别	公有云	私有云	混合云
部署方式	在线服务，开箱即用	私有化安装部署 数据私密性好	混合部署
扩展方式	扩展性和可靠性高	自主可控性好	兼具平台扩展能力和自主可控性
成本投入	无需额外的硬件投入	需要额外的硬件资源投入	软硬件资源投入适中
运维成本	无部署运维成本	有软硬件部署运维成本	部分软硬件部署运维成本

资料来源：搜狐，华金证券研究所

云部署不同方式对比



资料来源：可道云，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.6 不同云部署方式的市场占比

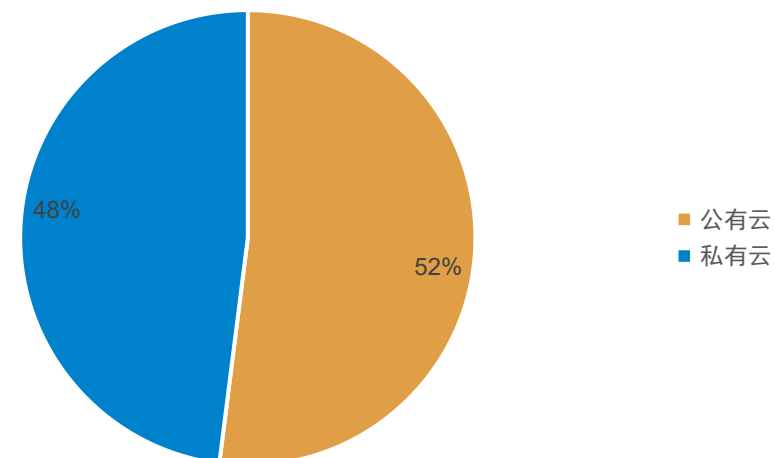
- 公有云是第三方提供的服务，而私有云是企业内建的供企业自身使用的云服务。根据中商情报网的数据，目前中国云计算市场中，公有云与私有云的市场规模相差较小，公有云占比较多，达52%；私有云占比48%。
- 从规模上来看，根据IDC的数据，2021年中国公有云市场达1853亿元。IDC预测，未来5年，中国公有云市场会以复合增长率30.9%继续高速增长，预计到2026年，市场规模将达到1057.6亿美元，中国公有云服务市场的全球占比将从2021年的6.7%提升为9.9%。

2中国公有云市场规模预测（亿美元）



资料来源：IDC，华金证券研究所

中国云计算市场规模占比情况



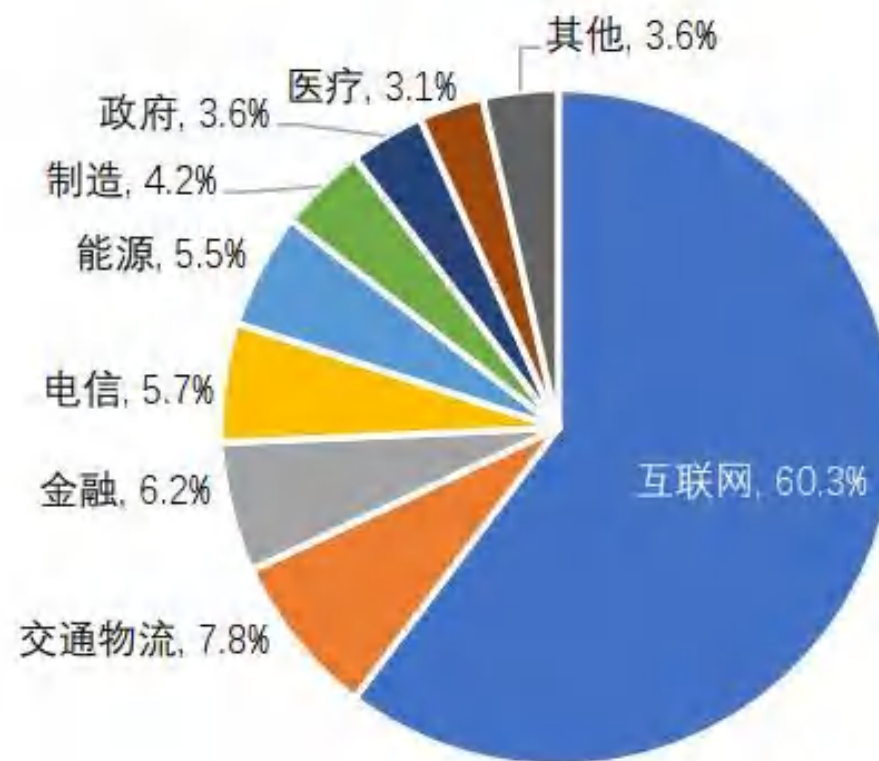
资料来源：中商情报网，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.7 企业上云持续向细分行业渗透

- 据Gartner公司测算，2015-2021年，全球政府和企业的云计算市场渗透率逐年上升，由4.3%上升至15.3%。云计算用户已经遍及互联网、政务、金融、教育、制造等各个行业。在中国，互联网行业是云计算产业的主流应用行业，占比约为1/3；在政策驱动下，中国政务云近年来实现高增长，政务云占比约为29%；交通物流、金融、制造等行业领域的云计算应用水平正在快速提高，占据了更重要的市场地位。

2018年中国云计算产业结构（按行业）



3. 全维智能化大时代，国产算力行则必至

3.8 从“资源上云”迈入“深度用云”

- 基础云计算服务将向新一代算力服务演进。作为云服务的升级，人工智能、区块链、大数据、扩展现实等算力服务不断成熟，并呈现出泛在化、普惠化、标准化的特点。新一代智能算力服务形成数字经济的核心生产力，成为加速行业数字化及经济社会发展的重要引擎。

2022年中国云计算市场十大预测

IDC 2022年中国云计算市场十大预测

IDC FutureScape



3. 全维智能化大时代，国产算力行则必至

3.9 信创从试点走向推广

- 根据经济参考报的分析，我国信息国产化的成长大致分为四个阶段：自1999年-2013年从无到有的核高基时代；2013年-2016年的“棱镜门”事件与去IOE；2016年-2018年的芯片半导体安全可靠时代；2018年底至今的大安全可靠，从党政军试点，关键行业金融、电信（运营商）、教育信、能源、医疗、交通、航空航天、建筑逐步推广，统称为“2+8”，中国一步一步的走出自己的新创产业。

信创产业体系全景图



资料来源：《中国信创产业发展白皮书（2021）》，华金证券研究所

2023 年中国与全球计算机产业市场空间预测（亿美元）

领域	产品	全球		中国		全球占比
		市场空间	5 年 CAGR	市场空间	5 年 CAGR	
硬件	服务器	1121	3.7%	340	12.4%	30%
	企业存储	311	1.0%	60	6.9%	19%
软件	基础架构软件	1525	5.3%	29.2	198%	2%
	数据库	569	7.5%	40	26.9%	7%
	中间件	434	10.3%	14	15.7%	3%
	大数据平台	410	15.6%	27	44.7%	6%
	企业应用软件	4020	8.2%	156	1170%	4%
云计算	公有云	1410	31.4%	289	51%	20%
	其中 SaaS	296	44.8%			

资料来源：《中国信创产业发展白皮书（2021）》，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.10 公有云主要参与厂商

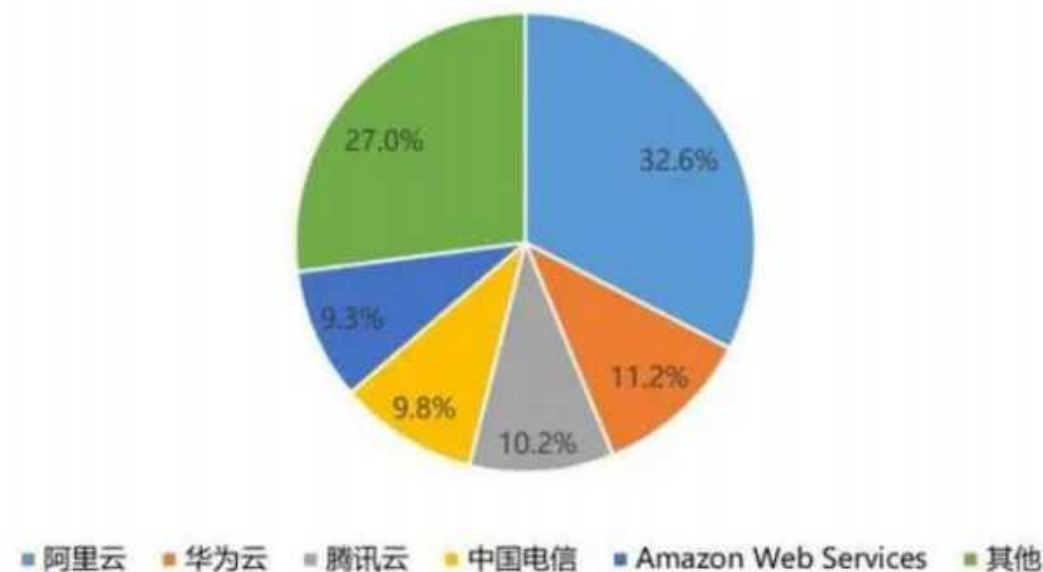
- 国际权威行业研究机构IDC最新发布《中国公有云服务市场（2022H1& 2022Q2）跟踪》报告。报告数据显示：2022年第二季度，阿里云以33.8%排第一，然后是华为云、中国电信、腾讯云、Amazon Web Services，分别占11.7%、11.5%、9.9%、8.3%。
- 华为云在中国Top3云厂商中保持增速最快，位列中国公有云市场（IaaS+PaaS）第二，其中IaaS市场份额上升至11.7%，IaaS+PaaS市场份额上升至11.2%。

中国Top5公有云IaaS厂商份额占比（2022Q2）



资料来源：IDC，华金证券研究所

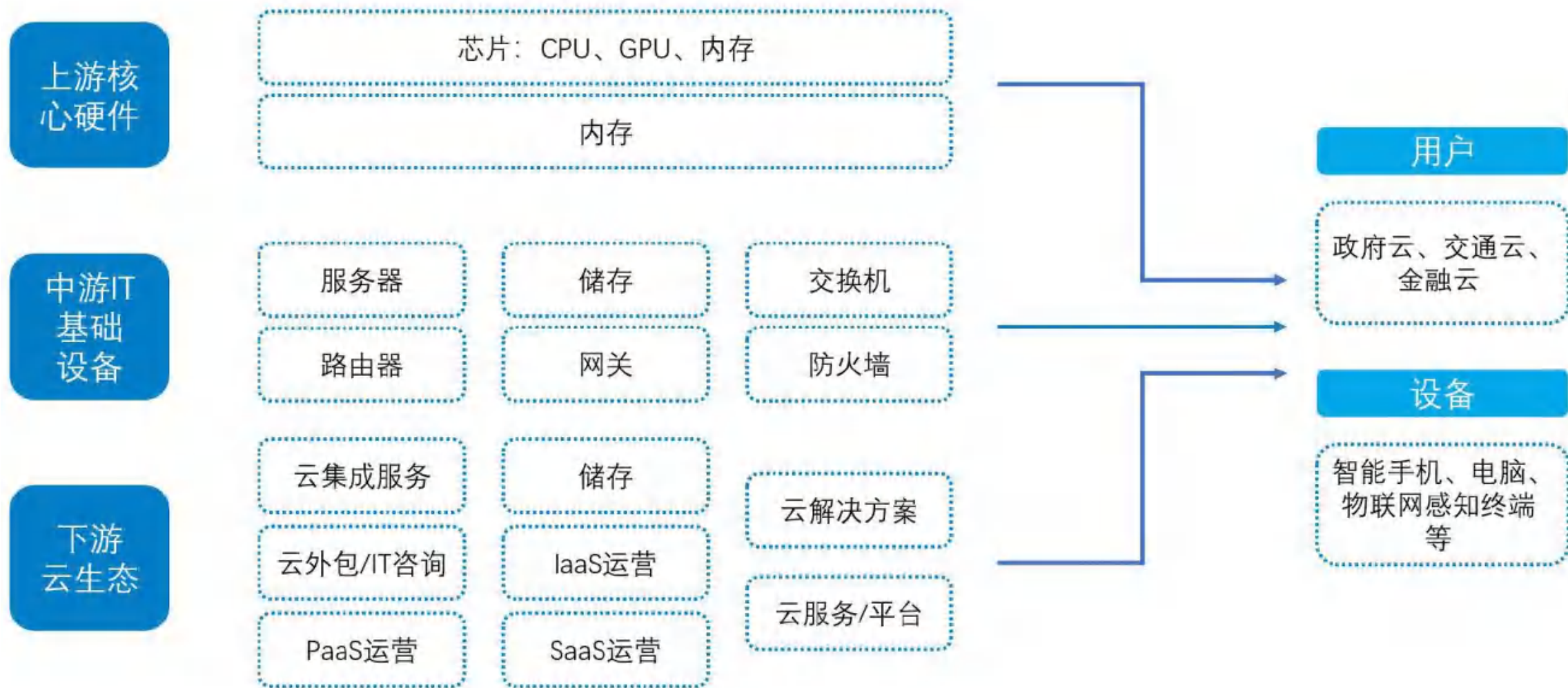
中国Top5公有云IaaS+PaaS厂商份额占比（2022Q2）



资料来源：IDC，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.11 云计算产业链



3. 全维智能化大时代，国产算力行则必至

3.12 集成显卡与独立显卡市场份额

- 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。
- 从量级上来看，集成显卡的出货量级最大，根据GPU行业调研机构JPR的公布的数据，2021年Q2，全球GPU出货量高达1.23亿，其中英特尔占据了68.3%的份额，AMD和英伟达分别维16.5%、15.2%。英特尔的高份额主要来源于其CPU和GPU的捆绑销售，即作为集成显卡的形式运行在PC当中。独立显卡则是英伟达一家独大。

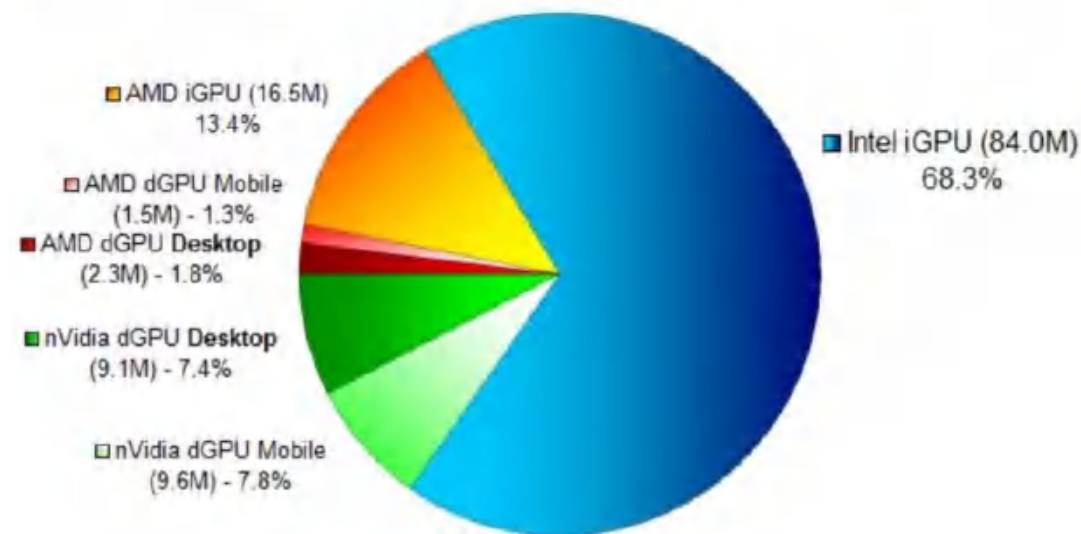
GPU的分类

区别	集成显卡	独立显卡
与CPU的关系	集成在CPU里面的图像处理单位，构成CPU的一部分	单独插在主板上的图像处理单位，其接口是PCIe接口，是一个单独的电脑组件
价格	低	高
兼容性	较好	较差
性能	较差	较好
升级成本	低	高
功耗	低	高
是否占用电脑内存	是	否
主要生产商与产品	Intel(HD系列)、AMD (APU系列)	AMD (Radeon系列)、NVIDIA (GeForce系列)
主要应用领域	移动计算市场，如笔记本和智能手机	高性能游戏电脑，VR/AR，人工智能

2021年Q2各厂商GPU出货量市场份额占比

Aufteilung PC-GPU Absätze - Breakdown PC GPU Sales - Q2'21

Summe aller PC-GPUs - Sum of all PC GPUs: 123M



资料来源：JPR，华金证券研究所

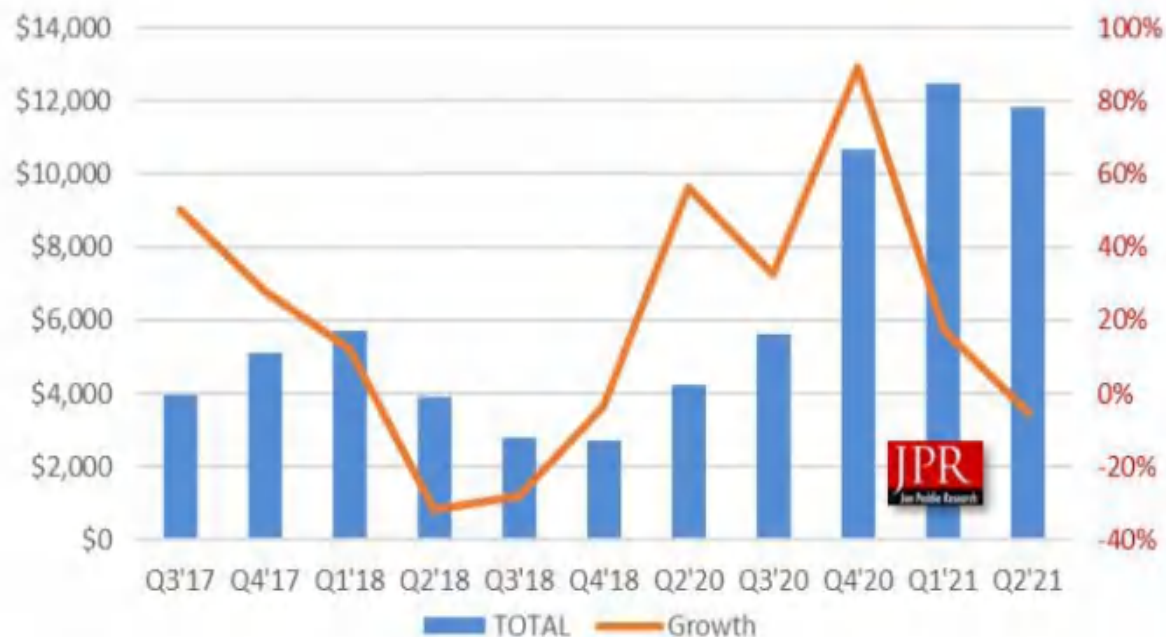
资料来源：华经情报网，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.13 独立显卡英伟达一家独大

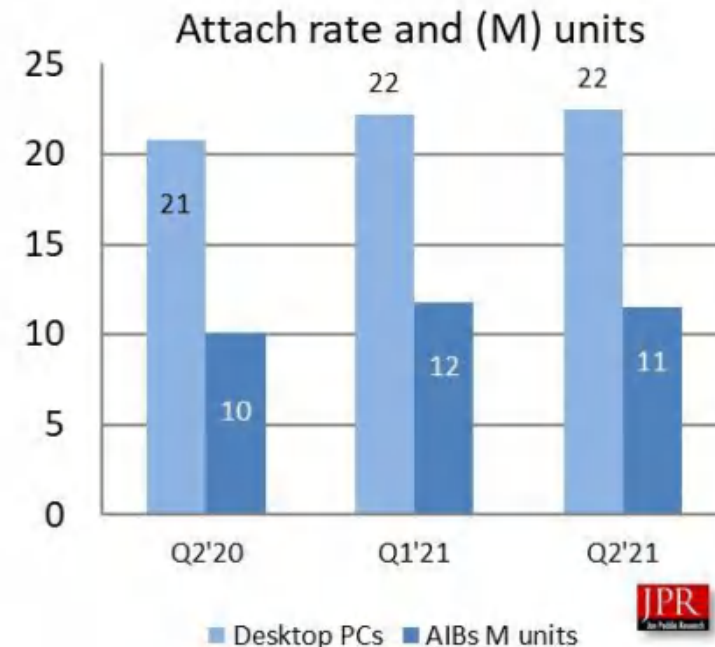
- 根据GPU行业调研机构Jon Peddie Research (JPR) 发布的报告显示，2021年第二季度，全球独立显卡市场销售额达118亿美元，同比增幅达到了179%，预计到2023年，整个市场将达441亿美元。2021年二季度独立显卡的出货量约为1100万块，比第一季度的1200万块减少了2.9%（台式机市场同期增长1.2%），但对比2020年同期的1000万块大幅增加了13.4%，而同期台式机市场只增长了8.0%。2021年第二季度的厂商份额方面，NVIDIA达到了80%，其次是AMD。

独立显卡市场规模（百万美元）



资料来源：JPR，华金证券研究所

独立显卡出货量（百万块）



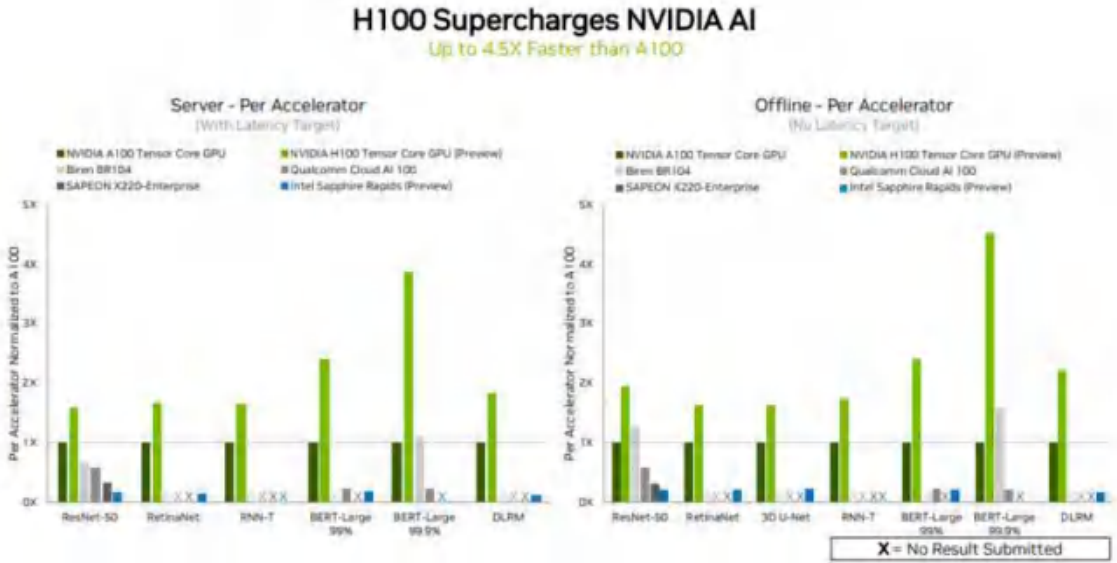
资料来源：JPR，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.14 性能强大的H100

- 英伟达于2022年3月份发布基于新架构NVIDIA Hopper的H100 GPU，H100采用Hopper架构，800亿晶体管，812的面积，台积电N4工艺，由800亿个晶体管组成，并采用了众多开创性的技术，包括强大的全新Transformer引擎和NVIDIA NVLink®互连技术，以加速最大规模的AI模型。根据MLCommons社区发布了最新的MLPerf 2.1基准测试结果，以NVIDIA A100相比，H100在MLPerf模型规模最大且对性能要求最高的模型之一——用于自然语言处理的BERT模型中表现出4.5倍的性能提升
- 根据新浪科技新闻，H100售价约24万人民币。

H100性能对比图



资料来源：雷锋网，华金证券研究所

东数西算枢纽节点区域特点及布局思路

	NVIDIA H100 SXM	NVIDIA H100 PCIe
FP64	30 TFLOPS	24 TFLOPS
FP64 Tensor Core	60 TFLOPS	48 TFLOPS
FP32	60 TFLOPS	48 TFLOPS
TF32 Tensor Core	1,000 TFLOPS*	800 TFLOPS*
BFLOAT16 Tensor Core	2,000 TFLOPS*	1600 TFLOPS*
FP8 Tensor Core	4,000 TFLOPS*	3200 TFLOPS*
INT8 Tensor Core	4,000 TOPS*	3200 TOPS*
GPU memory	80GB	80GB
GPU memory bandwidth	3TB/s	2TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max thermal design power (TDP)	700W	350W
Multi-Instance GPUs	Up to 7 MIGs @ 10GB each	
Form factor	SXM	PCIe dual-slot air-cooled
Interconnect	NVLink: 900GB/s PCIe Gen5: 128GB/s	NVLink: 600GB/s PCIe Gen5: 128GB/s
Server options	NVIDIA HGX™ H100 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX™ H100 with 8 GPUs	Partner and NVIDIA-Certified Systems with 1-8 GPUs

资料来源：CSDN，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.15 国产厂商两条发展路径：GPU和GPGPU

➤ GPU技术上面向渲染和AI计算两大方向，国内厂商基本上也各自选择了两者中的不同方向进行突破，小部分厂商同时进行两个方向的技术布局。GPGPU方向的代表性厂商像海光、壁仞、沐曦、登临、天数智芯等，渲染路线的代表性厂商像景嘉微、摩尔线程（两个方向均有布局）、芯动科技等。

壁仞™ 104P	
制程	7nm
系统接口、带宽、互连协议	PCIe5.0 X16, 128GB/s, 支持CXL
FP32 TFLOPS (峰值)	112
TF32+ TFLOPS (峰值)	224
BF16 TFLOPS (峰值)	448
INT8 TOPS (峰值)	896
内存容量、接口带宽、带宽	32GB HBM2E; 2,048bit, 819GB/s
互连	192GB/s BLink™, 支持3个端口, 最高可实现4卡全互连
安全虚拟实例	最高4份
视频编解码 (FHD@30fps)	32路HEVC/H.264编码、256路HEVC/H.264解码
TDP	300W
产品形态	全高全长, 双槽位PCIe卡

	海光 8100
典型功耗	260-350W
典型运算类型	双精度、单精度、半精度浮点数据和各种常见整型数据
计算	① 60-64 个计算单元（最多 4096 个计算核心） ② 支持 FP64、FP32、FP16、INT8、INT4
内存	① 4 个 HBM2 内存通道 ② 最高内存带宽为 1TB/s ③ 最大内存容量为 32GB
I/O	① 16 Lane PCIe Gen4 ② DCU 芯片之间高速互连

MTT S3000

满处理单元: 4596 MUSA核心

张量计算单元: 128核心

GPU核心频率: 1.0GHz

FP32算力: 52 TFLOPS

显存容量: 32GB

显存类型: GDDR6

显存位宽: 256bit

显存带宽: 448GB/s

总线接口: PCIe Gen5 x16

显示接口: 2 × DisplayPort 1.4a

散热方式: 被动散热

尺寸: 268.0mm × 110.90mm × 38.7mm

安全: MUSA安全引擎1.0, 支持TEE

虚拟化: GPU弹性切分, SR-IOV隔离



昆仑芯第二代芯片



FEATURES

增强的通用计算能力

XPU-R架构为CLUSTER的算力提升2-3倍, 进一步扩展通用AI计算能力

高性能分布式AI系统

芯片间K-Link互联支持3D数据和推理中模型并行, 在数据并行推理的场景下

支持硬件虚拟化

计算单元和存储单元物理隔离, 优化了加速芯片的利用率, 在保证低延迟吞吐量的情况下, 支持推理和训练等混合工作负载。

7NM

先进工艺

GDDR6

高性能显存

PCI-e4.0x16

高强度接口

150w

功耗

XFP16: 128 TOPS

高度集成了ARM CPU算力 Cortex A55 × 8

资料来源：各公司官网，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.16 先求有，再求好

- 国产GPU厂商真正的大量出现是最近几年的事情，从当前来看，大量企业属于初创期，无论是产品能力还是市场规模都无法和国外大厂相提并论，但是我们认为国内厂商当下是从0到1的阶段，先求有产品，能够通过产品逐步打开一定的市场，再求快速迭代，拉近与国外大厂的差距，并形成自身的竞争力。

景嘉微已量产GPU产品与英伟达 GT640性能对比

	JM5400	JM7200	JM7201	GT640(DDR3) (Nvidia)
流片/发布时间	2014.4	2018.8	2019	2012.6
工艺	65nm CMOS	28nm CMOS	28nm CMOS	28nm
内核时钟频率	最大550MHz	最大1300MHz	最大1200MHz	900MHz
主机接口	PCI 2.3	PCI-E 2.0×16	PCI-E 2.0×16	PCI-E 3.0×16
显存带宽	12.8GB/s	17GB/s	-	-
存储容器量	1GB DDR3	4GB DDR3	4GB DDR3	2GB DDR3
像素填充率	2.2Gpixels/s	5.2Gpixels/s	4.8Gpixels/s	7.2Gpixels/s
浮点性能/GFLOPS	160	500	-	692
功耗	不超过6W	桌面小于20W、嵌入式小于10W	桌面10W-15W	50W

资料来源：华经情报网，华金证券研究所

2021年上半年部分AI芯片企业融资情况一览

融资时间	企业名称	融资轮次	融资金额	芯片应用领域
2021/1/1	地平线	C2轮	4亿美元	边缘，自动驾驶
2021/2/1	地平线	C3轮	3.5亿美元	边缘，自动驾驶
2021/2/2	燧原科技	C轮	18亿人民币	云端
2021/2/3	登临科技	A-轮	未披露	云端、GPU+
2021/2/4	摩尔线程	Pre-A轮	数十亿	云端，GPU
2021/3/1	百度（昆仑）	战略融资	20亿	云端
2021/3/1	地平线	C4轮	未披露	边缘，自动驾驶
2021/3/1	后摩智能	天使轮	数千万美元	云端，边缘
2021/3/22	千芯半导体	战略融资	数千万人民币	云端，边缘
2021/3/26	墨芯	Pre-A轮	1亿	云端，终端
2021/3/30	壁仞科技	B轮	未披露	云端，GPU
2021/4/1	超博半导体	A-轮	5亿人民币	云端推理
2021/4/1	地平线	C5轮	未披露	边缘，自动驾驶
2021年4月	紫元展锐	战略融资	53.5亿人民币	终端
2021/4/1	SambaNova	D轮	6.76亿美元	云端
2021/4/3	Groq	D1轮	3亿美元	云端
2021/5/1	地平线	C6轮	超3亿美元	边缘，自动驾驶
2021/5/7	启英泰伦	B轮	数千万人民币	终端
2021/5/26	视海芯图	Pre-A轮	数千万人民币	终端
2021/6/1	地平线	C7轮	15亿	边缘，自动驾驶
2021/6/5	九天壹芯	A轮	亿元级	终端，神经视觉
2021/6/6	芯聚能	A轮	未披露	边缘，汽车
2021/6/30	沐曦集成电路	股权融资	未披露	云端，GPU
2021/6/30	杭州国芯科技	战略融资	未披露	终端
2021/7/1	智微芯半导体	B轮	数亿元	边缘，终端AI视觉
2021/7/1	锐思智芯	Pre-A轮	近亿元人民币	终端，机器视觉
2021/7/16	埃瓦科技	A轮	亿元级人民币	终端，3D AI视觉

资料来源：ebrun，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.17 生态先兼容主流，未来将走向自建

- 对于GPU用户，产品性能是一个门槛，另一个门槛是易用性，这涉及到生态问题。国外大厂经过多年发展，已经形成了比较强的生态，国内GPU厂商在发展初期一般需要先兼容主流生态来尽可能提升自身芯片的易用性。
- 面向AI应用场景时，AI对于生态要求比较高，涉及框架、应用、模型的适配等，以英伟达为例，作为通用的DSA，拥有着广泛使用的编程生态，虽然在硬件的计算能效上会低于一些专用芯片，但是其较高的易用性大幅度降低了下游开发者的应用门槛。现在主流的深度学习框架基本都是基于CUDA进行GPU并行加速。

地平线天工开物开发平台



摩尔线程开发的CUDA ON MUSA兼容方案



资料来源：ofweek， 华金证券研究所

资料来源：eetop， 华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.18 国产之路已开启，部分国产GPU设计厂商列表

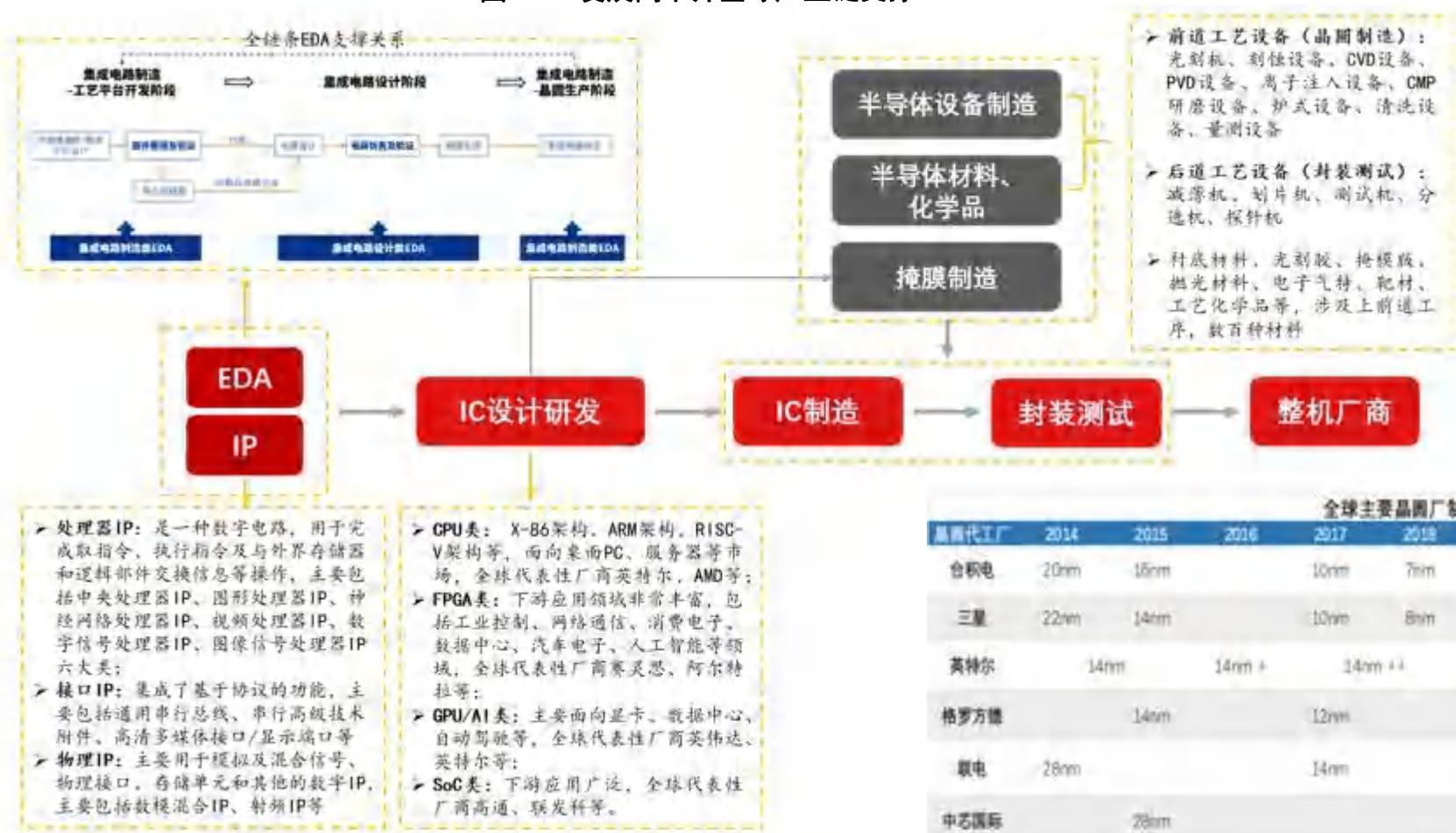
公司	地区	主业	代表性产品
景嘉微电子	长沙	产品主要涉及图形显控领域、小型专用化雷达领域、芯片领域和其他	JM5系列、JM7系列和JM9系列为代表的系列图形处理芯片。JM9系列图形处理芯片已完成流片、封装阶段工作及初步测试工作
芯动科技	珠海	一站式IP和芯片定制，聚焦计算、存储、连接等三大赛道	“风华1号”4K级多路服务器GPU：渲染能力160GPixel/s，单精度浮点性能5TFLOPS，AI性能25TOPS；“风华2号”4K级三屏桌面GPU：浮点算力1.5TFLOPS（FP32），AI性能12.5TOPS（INT8）
航锦科技	辽宁	化工板块，军工产品涵盖存储芯片、总线接口芯片、模拟芯片、图形处理芯片、特种FPGA、多芯片组件等	高性能图形处理芯片SG6931：国产化自主可控PCI-E 图形处理器，支持ows、Linux、VxWorks、DOS 等通用操作系统以及多种国产处理器反熔丝FPGA：国内独家供应商，用于军用战斗机
兆芯	上海	中央处理器、图形处理器、芯片组三大核心技术，具备相关IP自主设计研发的能力。	开先KX-6000 X86处理器：代号“陆家嘴”，16nm制程工艺，8核心设计，主频3.0GHz，集成8MB L3缓存，双通道DDR4-3200内存
昆仑芯	北京	深耕AI加速领域	昆仑芯2代芯片：2021年量产，第二代云端通用人工智能计算处理器，7nm先进工艺，GDDR6高性能显存，256 TOPS@INT8，128 TFLOPS@FP16，最大功耗120W
凌久微电子	武汉	图形处理GPU/SOC的研发以及配套软件生态构建	凌久GP201：统一渲染架构的自主高性能图形处理芯片，主频1GHz，单精度浮点1TFlops，最大支持32GB DDR4/LPDDR4显存
天数智芯	上海	自研云端通用计算GPGPU及基础软件	天垓100：国内唯一量产的通用GPU，基于SIMT架构，7 nm制程、240亿晶体管，2.5D CoWoS封装，1.2TB/s带宽和32GB容量内存
壁仞科技	上海	GPU、DSA和计算机体系结构领域	BR100：7nm制程工艺，770亿晶体管，原创架构下以OAM模组形态部署，能在通用UBB主板上形成8卡点对点全互连拓扑
沐曦集成电路	上海	完全自主研发的GPU IP，异构计算和各种高性能GPU	MXN系列GPU（曦思）用于AI推理，MXC系列GPU（曦云）用于科学计算及AI训练，以及MXG系列GPU（曦彩）用于图形渲染
登临科技	上海	高性能通用计算平台的芯片研发与技术创新	Goldwasser-XL：INT8算力 512TOPS，GPU+较国际主流产品有更小的功耗和更高的性能
摩尔线程	北京	研发设计全功能GPU芯片及相关产品	MTT S80：消费级游戏显卡，“春晓”芯片核心，内置 4096 个 MUSA 流处理核心，16GB GDDR6 高速显存，1.8GHz 的主频，14.4TFLOPS 的单精度浮点算力 MTT S3000：面向服务器，“春晓”芯片核心，4096 个 MUSA核心，128 个专用张量计算核心，晶体管 220 亿，运行频率为 1.9GHz，32GB GDDR6 显存；FP32 算力 15.2TFLOPS
瀚博半导体	上海	高性能通用加速芯片和智能视觉芯片设计	SG100：7nm制程工艺，SR-IOV虚拟化，集渲染、AI、视频于一体 载天VA10：INT8算力400TFLOPS，适用于实时AI应用
燧原科技	上海	人工智能领域云端算力产品	邃思2.0：12nm制程，最高达160TFLOPS(TF32)，64GB容量，1.8TB/s带宽
芯瞳半导体	西安	GPU 芯片设计、异构计算平台、嵌入式显示和GPU 应用	GenBu01：40nm制程，国内第一款统一渲染架构芯片，1GB显存
龙芯中科	北京	先进制程芯片和高性能通用图形处理器芯片	龙芯中科于 2017 年开始研发 GPU,第一款 GPU IP 核已经在龙芯 7A2000 桥片样片中流片成功
海光信息	天津	海光通用处理器 (CPU) 和海光协处理器 (DCU)	海光8000系列。海光8100，典型运算类型包括双精度、单精度、半精度浮点数据和各种常见整型数据，采用先进的FinFET工艺，典型应用场景下性能指标可以达到国际同类型高端产品的同期水平
寒武纪-U	北京	注于人工智能芯片产品的研发与技术创新	思元220、思元290、思元370等

资料来源：各公司官网，华金证券研究所

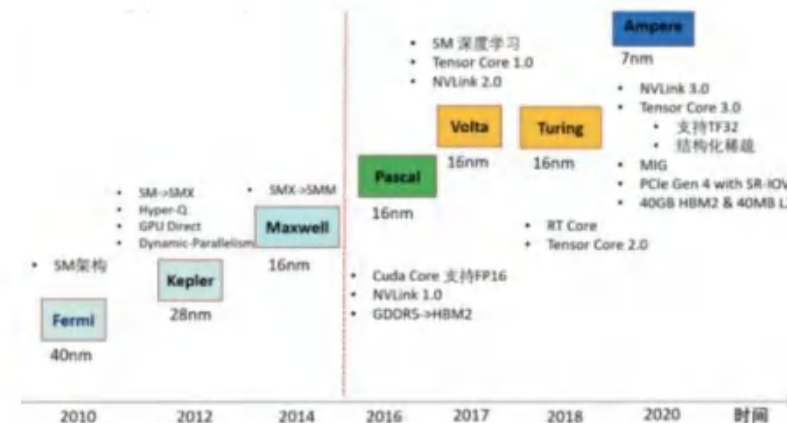
3. 全维智能化大时代，国产算力行则必至

3.19 GPU发展离不开全球产业链的支撑

图：GPU发展离不开全球产业链支撑



图：英伟达产品升级离不开制程升级



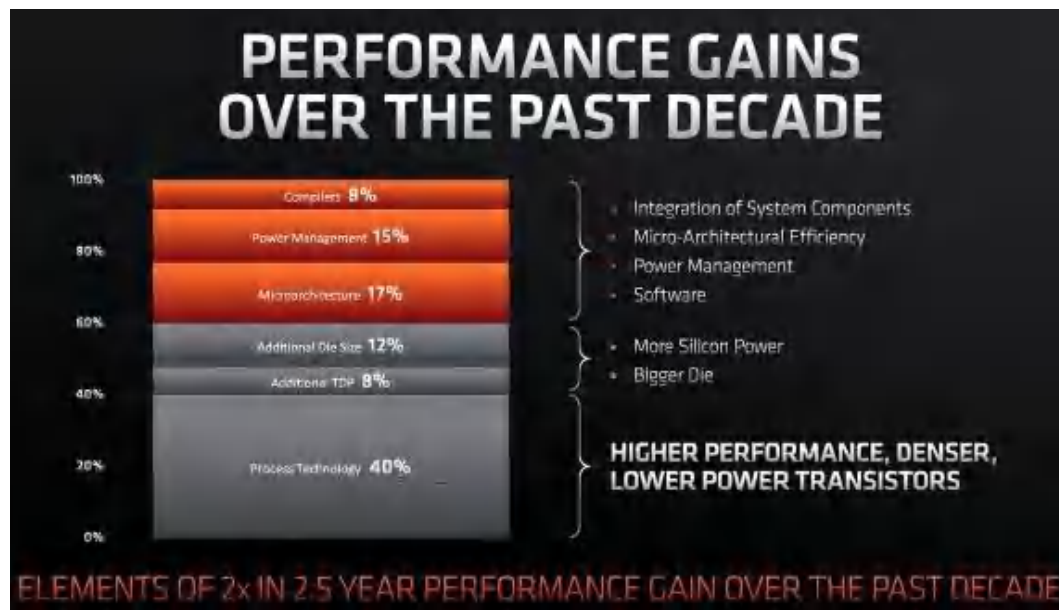
全球主要晶圆厂制程节点技术路线图											
晶圆代工厂	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
台积电	20nm	16nm		10nm	7nm	7nm+	5nm	5nm	3nm		2nm
三星	22nm	14nm		10nm	8nm	7nm EUV	5nm	3nm			
英特尔		14nm	14nm+	14nm++		10nm	10nm+	7nm	7nm++	7nm++	
格罗方德		14nm		12nm							
联电	28nm			14nm							
中芯国际		28nm				14nm					

3. 全维智能化大时代，国产算力行则必至

3.20 制程升级对于算力芯片性能提升具有较高贡献度

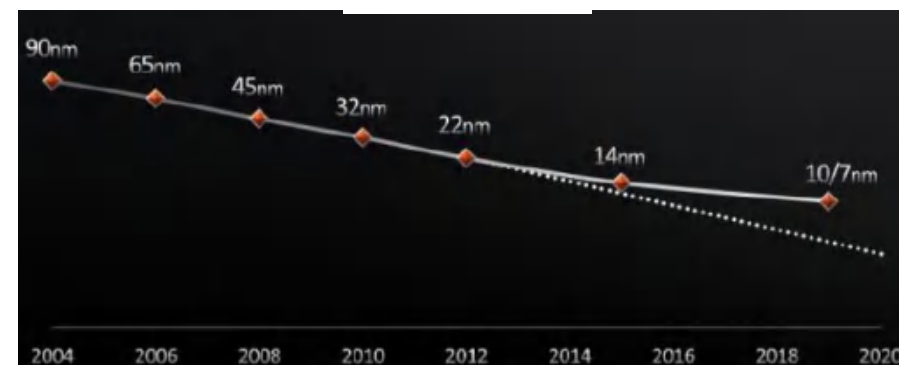
- 以过去十年CPU的性能大幅度提升为例，根据AMD的分析，在性能提升的贡献上面，制程工艺技术占40%，裸片尺寸和额外的TDP占另外20%，微架构、电源管理和编译器构成了图片的其余部分。

过去十年芯片性能提升的贡献分析

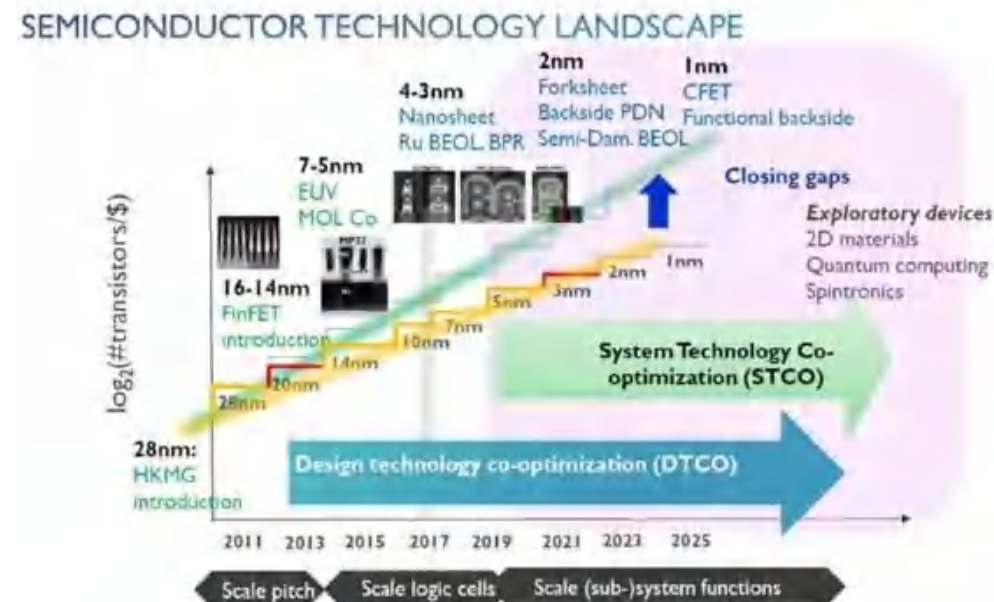


资料来源: ednchina, chinaflashmarket, 电子发烧友, 华金证券研究所

摩尔定律放缓



半导体工艺技术路线图

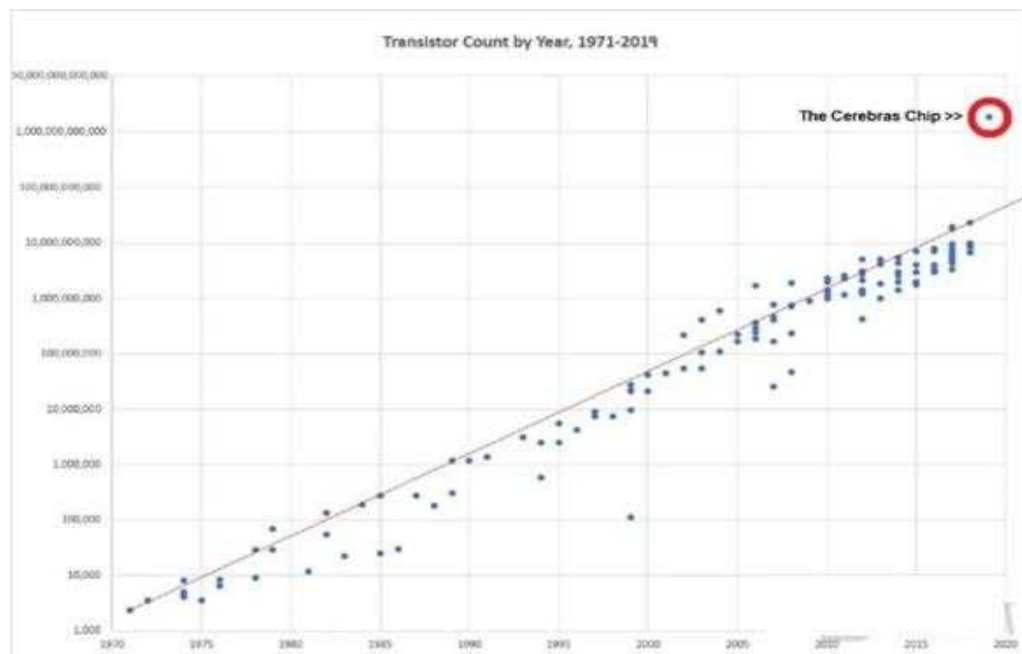


3. 全维智能化大时代，国产算力行则必至

3.21 摩尔定律发展趋缓

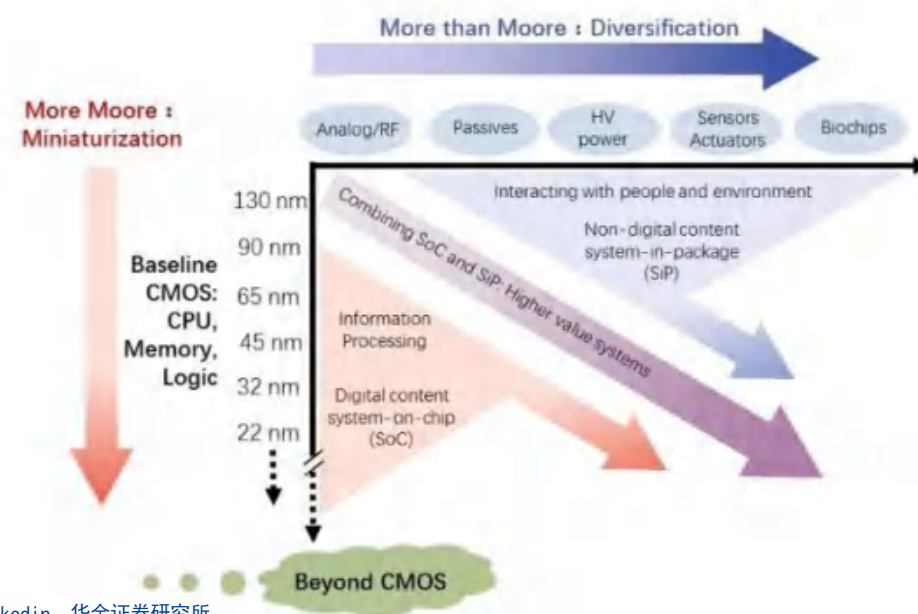
- 2000年之前，每一代芯片的性能提升来自两个方面：一是按照Denard（登纳德）微缩效应，每代芯片的频率提升带来了40%的改进；二是每代芯片晶体管密度提升带来的体系结构的改进符合波拉克法则，即平方根级别的提升，达41%。将这两方面的性能提升叠加，最终得到1.97倍，于是每代会有差不多一倍的提升，而且，芯片晶体管密度的“摩尔定律”可换算成性能的“摩尔定律”。如今，“摩尔定律”已经越来越偏离最早的预测。
- 英伟达（NVIDIA）创始人兼CEO黄仁勋表示，以类似成本实现两倍业绩预期对于芯片行业来说已成为过去，“摩尔定律已经死了。”

2019年Cerebras芯片偏离了摩尔定律发展



资料来源：钛媒体，华金证券研究所

半导体制造工艺的两种演进路线图



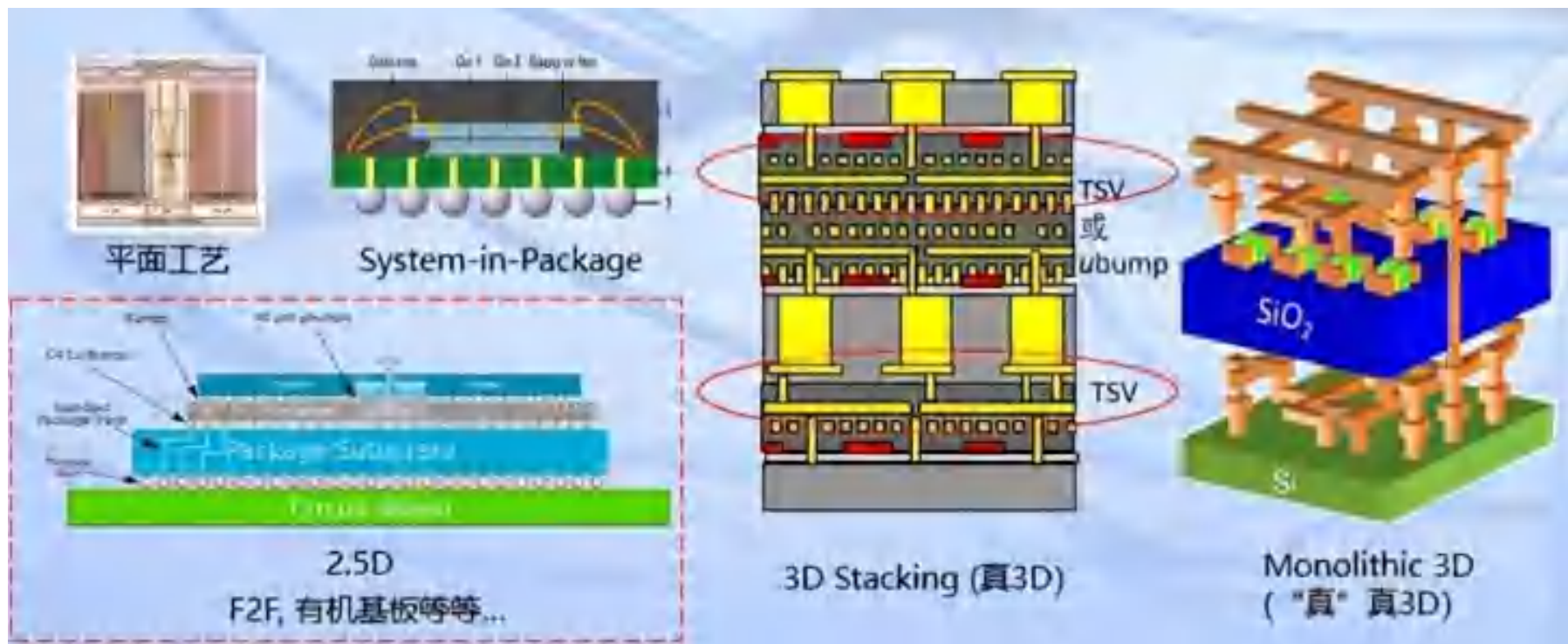
资料来源：LinkedIn，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.22 Chiplet技术潜力大

- Chiplet俗称芯粒，也叫小芯片，是将一类满足特定功能的die（裸片）通过die-to-die内部互联技术将多个模块芯片与底层基础芯片封装在一起，Chiplet是一种类似打乐高积木的方法，能将采用不同制造商、不同制程工艺的各种功能芯片进行组装，从而实现更高良率、更低成本。
- Chiplet技术可以突破单一芯片的性能和良率等瓶颈，降低芯片设计的复杂度和成本。基于向Chiplet模式的设计转型，已经是大型芯片厂商的共识。

Chiplet是芯片生产和集成技术发展趋势

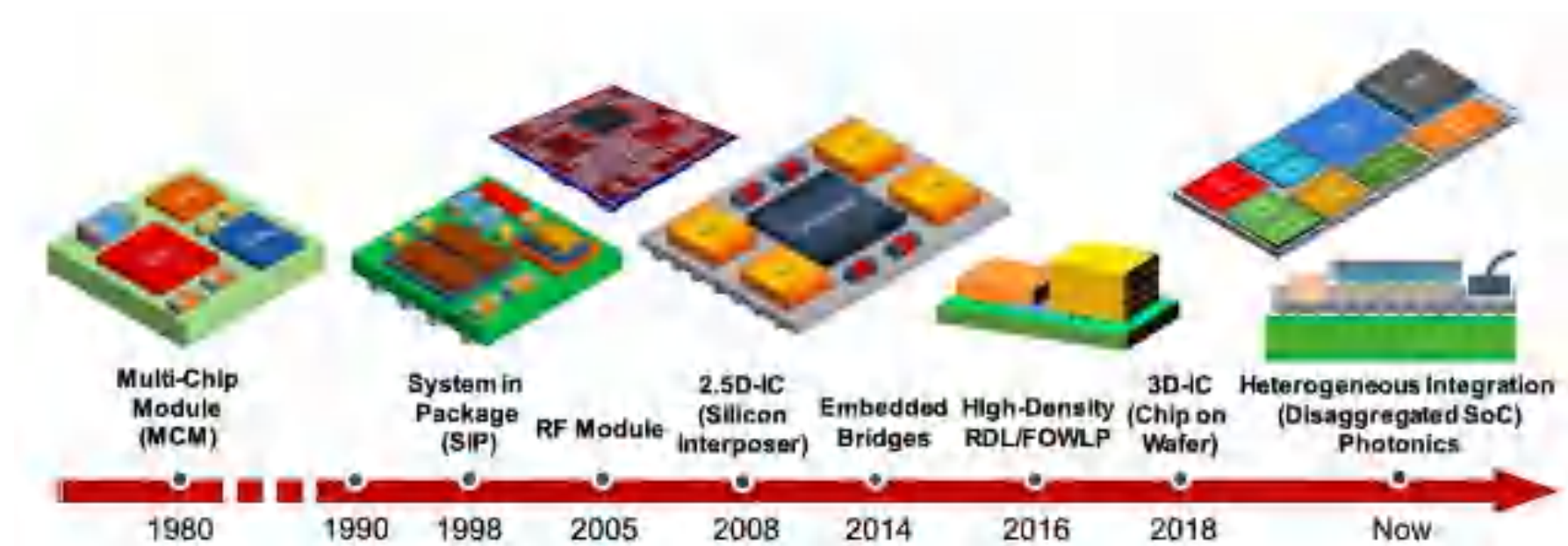


3. 全维智能化大时代，国产算力行则必至

3.23 Chiplet技术发展历程

- 先进封装工艺包括倒装焊（FlipChip）、晶圆级封装（WLP）、2.5D封装（Interposer）、3D封装（TSV）、Chiplet等。
- 根据Yole数据，2021年全球封装市场规模约达777亿美元。其中，先进封装全球市场规模约达777亿美元。其中，先进封装全球市场规模约350亿美元。预计到2025年，先进封装的全球市场规模将达到420亿美元，2019-2025年，全球封装市场的CAGR约8%。

芯片整合已演进至2.5D/3D及Chiplet封装



3. 全维智能化大时代，国产算力行则必至

3.24 行业巨头推动，产业加速落地

- 英特尔、AMD、Arm、高通、台积电、三星、日月光、google云、Meta(facebook)、微软等十大行业巨头成立了Chiplet标准联盟，正式推出了通用Chiplet（芯粒）的高速互联标准“Universal Chiplet Interconnect Express”，简称“UCIe”，旨在定义一个开放的、可互操作的标准，用于将多个硅芯片（或芯粒）通过先进封装的形式组合到一个封装中。
- 中国计算机互连技术联盟（CCITA）秘书长、中科院计算所研究员郝沁汾接受媒体采访时表示，中国可以采用28nm成熟工艺的芯片，通过Chiplet封装方式，使其性能和功能接近16nm甚至7nm工艺的芯片性能。

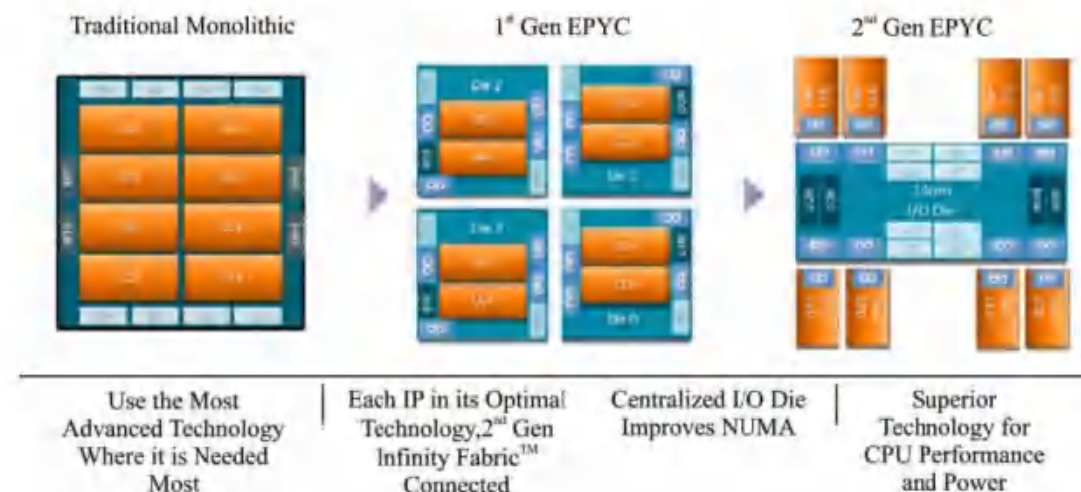


3.25 采用Chiplet技术的产品不断出现

英特尔公司基于AIB的多样化的Chiplet生态系统



AMD EPYC芯片演化



请仔细阅读在本报告尾部的重要法律声明

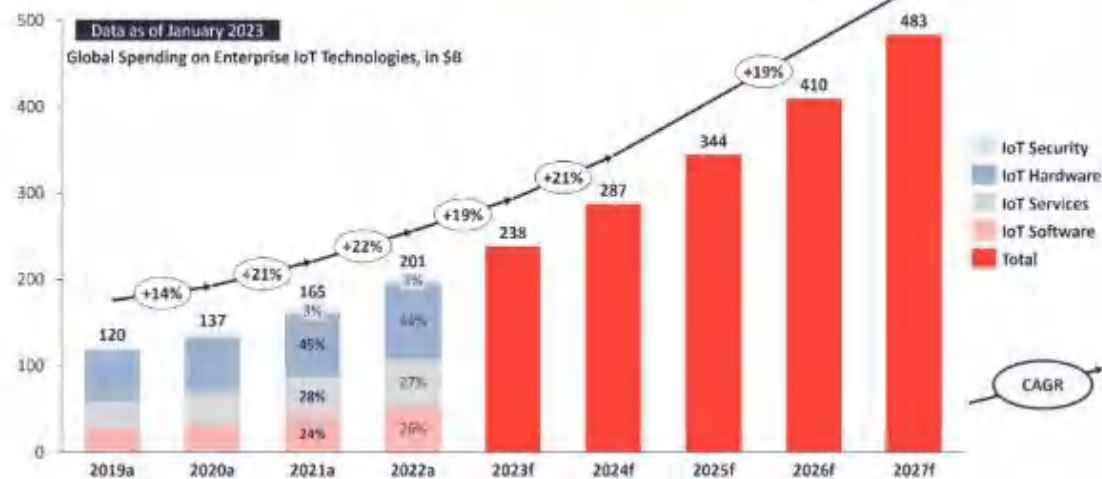
3. 全维智能化大时代，国产算力行则必至

3.26 算力两大演进方向：更大算力&更多样化应用

- 据IDC预测数据，伴随万物感知、万物互联以及万物智能时代的开启，2025年全球物联网设备数将超过400亿台，产生数据量接近80ZB。预估未来五年全球算力规模将以超过50%的速度增长，到2025年整体规模将达到3300EFlops。超过一半的数据需要依赖终端或者边缘的计算能力进行处理。
- 市场研究机构IoT Analytics发布了全球企业物联网支出的跟踪研究报告，报告显示，过去的2022年，全球各行业企业在物联网方面的支出2010亿美元，同比增长了21.5%。在2022年全球经济低迷、疫情打断正常生产的影响下，能实现这一增长实属不易。预计到2027年，企业对物联网的支出将达到4830亿美元

IoT市场预测

Enterprise IoT market 2019–2027



资料来源：IoT Analytics，华金证券研究所

多样化是算力发展的另一个维度



资料来源：华为，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.27 存量替代与增量成长并存

➤ 整体来看，处理器芯片下游应用广泛，既包括消费电子如手机、平板、扫地机器人、无人机等，又包括各种类型的行业应用如安防、商显、工业等。对于处理器芯片需求的增长有两个维度，一方面是出货量增长带来的，一方面是性能持续升级带来的。而对于国内芯片厂商来说出货量的增长又可细分为两个维度，一方面是国产替代带来的，另一方面是下游需求增量带动的。

终端应用场景主要市场出货量和供应情况概览

	国内	全球	核心技术点	核心厂商
智能手机	3.51	13.50	需要BP配套，竞争激烈，先进制程跟进	高通、联发科、三星、苹果、海思、紫光展锐等
平板电脑	0.28	1.68	需要BP配套，竞争激烈，先进制程跟进	高通、联发科、三星、苹果、海思、瑞芯微、全志等
PC	0.57	3.49	X86架构	英特尔、高通、AMD
智能电视	0.39	2.15	全格式视频编解码、图像显示处理、数字电视解码	联发科、晶晨、瑞芯微、全志、海思
机顶盒	0.72	3.05	全格式视频编解码、运营商准入	晶晨、博通、海思、国科微、瑞芯微、全志
安防	4.7	8	图像处理、视频编解码	海思、TI、富瀚微、厦门星宸、北京君正、联咏、瑞芯微、全志等
商显	0.09	-	图像显示处理、视频编解码	瑞芯微、晶晨、全志、MSTAR等
智能座舱	0.137	-	先进制程、图像显示处理、人机交互、车规认证、车规操作系统	高通、联发科、三星、海思、瑞萨、恩智浦、瑞芯微、晶晨、全志等
智能手表	0.40	1.28	图像显示处理、人机交互、操作系统、功耗控制	高通、恒玄等
VR/AR	-	0.11	图像显示处理、人机交互、视频解码、功耗控制	高通、瑞芯微、全志等

资料来源：IDC、Counterpoint，华金证券研究所

3. 全维智能化大时代，国产算力行则必至

3.28 高吞吐量离不开高速传输

- 中国联通于2019年在行业内首次发布算力网络白皮书，率先倡导算力网络概念，提出算力网络是云网融合新发展阶段。从网络角度看，算力网络是面向计算和智能服务的新型网络体系，“IPv6+”和“全光底座”是算力网络的技术基石；从算力和服务角度看，算力网络是网络化的算力基础设施，是依托网络构建的多样化算力资源调度和服务体系，是数字基础设施服务的新形态。
- 光纤通信网络是信息基础设施重要组成和关键承载底座，国家“十四五”规划、信息通信业“十四五”规划等对于光纤通信网络的未来发展高度重视，千兆光网发展、骨干网演进和服务能力升级成为重点内容。

图：全光算力网络是实现高品质、高安全、高可靠的算力业务传送的关键所在

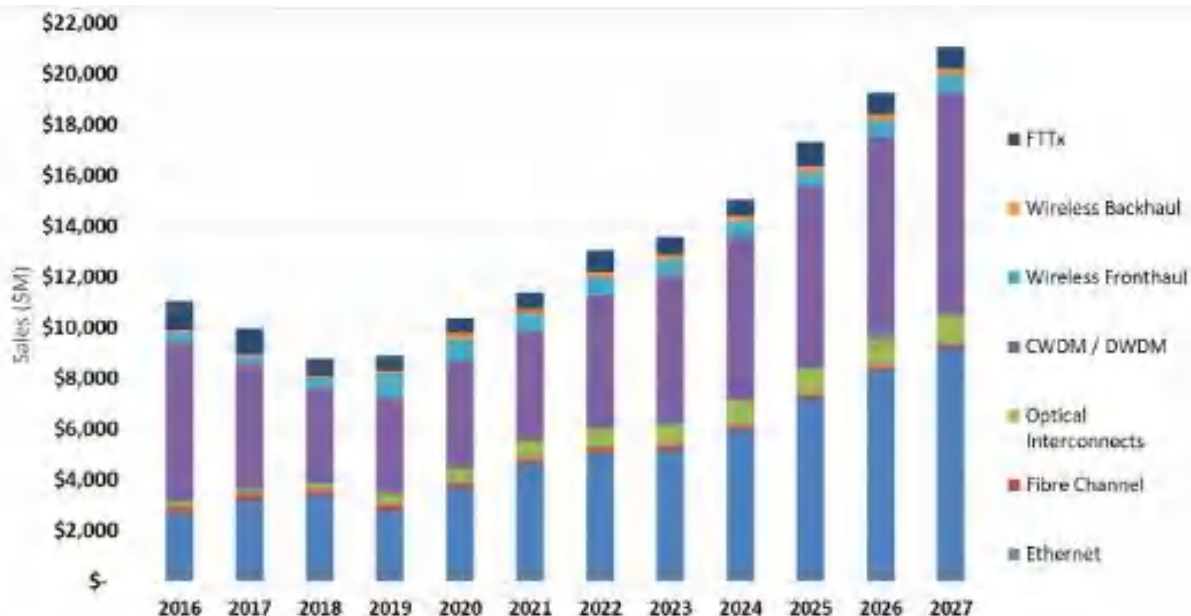


3. 全维智能化大时代，国产算力行则必至

3.29 光通信前景可期

- 光取代电作为信息载体是未来的趋势，从光通信到光芯片，光将渗透到人们生活的方方面面，这一进程正在加速；面临海量的数据需求的场景下，电的进化在变弱，光的技术在变强。光 and 电，都是信号承载的载体，相对于电通信，光通信的优点包括传输速率高、能耗效率高、无信号间干扰等。
- LightCounting认为，光模块市场在2020年和2021年分别增长17%和10%之后，2022年有望再次实现收入的强劲增长，预期14%。不过，预计在2023年将放缓至4%，然后在2024-2025年恢复。

光通信行业市场



资料来源：LightCounting，华金证券研究所

光通信产业链



资料来源：联赢激光官网，华金证券研究所

- 01 由专用走向通用，GPU赛道壁垒高筑
- 02 产业化路径显现，全球AI竞赛再加速
- 03 全维智能化大时代，国产算力行则必至
- 04 建议关注
- 05 产业相关
- 06 风险提示

4. 建议关注

4.1 瑞芯微

- 瑞芯微主要致力于大规模集成电路及应用方案的设计、开发和销售，在大规模SoC芯片设计、数模混合芯片设计、影像处理、高清视频编解码、人工智能及系统软件开发上具有丰富的经验和技術储备，形成了多层次、多平台、多场景的专业解决方案，赋能智能硬件、机器视觉、行业应用、消费电子、汽车电子等多元领域。
- 公司产品涵盖智能应用处理器芯片、数模混合芯片、接口转换芯片、无线连接芯片及与自研芯片相关的模组产品等，并为客户提供技术服务。

营业总收入及增长率（亿元）



资料来源: wind, 华金证券研究所

归属母公司股东的净利润及增长率（亿元）



资料来源: wind, 华金证券研究所

4. 建议关注

4.2 晶晨股份

- 晶晨半导体是全球布局、国内领先的无晶圆半导体系统设计厂商，为智能机顶盒、智能电视、音视频系统终端、无线连接及车载信息娱乐系统等多个产品领域提供多媒体SoC芯片和系统级解决方案，业务覆盖全球主要经济区域，积累了全球知名的客户群。产品技术先进性和市场覆盖率位居行业前列，为智能机顶盒芯片的领导者、智能电视芯片的引领者和音视频系统终端芯片的开拓者。晶晨半导体拥有丰富的SoC全流程设计经验，坚持超高清多媒体编解码和显示处理、内容安全保护、系统IP等核心软硬件技术开发，整合业界领先的CPU/GPU技术和先进制程工艺，实现前所未有的成本、性能和功耗优化，提供基于多种开放平台的完整系统解决方案，帮助全球顶级运营商、OEM、ODM等客户快速部署市场。

公司部分料号



资料来源：公司官网，华金证券研究所

4. 建议关注

4.3 星辰科技（待上市）

- 公司为全球领先的视频监控芯片企业，主营业务为视频监控芯片的研发及销售，产品主要应用于智能安防、视频对讲、智能车载等领域。公司在芯片设计全流程具有丰富经验，可支撑大型先进工艺下的SoC设计。公司自研全套AI技术，包含AI处理器指令集、AI处理器IP及其编译器、仿真器等全套AI处理器工具链。公司拥有大量核心IP资源，包含图像IP、视频IP、高速模拟IP和音频IP等。公司在视频监控领域持续研发创新，在图像信号处理、音视频编解码、显示处理等领域具有领先优势，并积极投入AI等新领域的芯片研发。公司拥有ISP技术、AI处理器技术、多模视频编码技术、高速高精度模拟电路技术、先进制程SoC芯片设计技术等多项核心技术，公司拥有已授权专利92项，其中境内发明专利11项，境外专利81项；在申请中专利154项，其中境内发明专利63项，境外专利91项。

公司产品线

产品领域	产品类型	产品型号	终端应用场景
智能安防	IPC SoC	SSC369G/SSC359G/SSC339G/SSC338Q/SSC357G/SSC30KQ/SSC336Q/SSC337DE/SSC335/SSC333E等	专业安防、民用安防、家庭消费类监控摄像头
	NVR/XVR SoC	SSR950G/SSR931G/SSR920G/SSR910Q/SSR650G/SSR621Q等	
视频对讲	视频对讲芯片	SSU9481G/SSC9381G/SSC9351Q/SSC9341/SSC9211等	视频对讲系统
智能车载	CAR DVR	SSC8838G/SSC8629G/SAC8542/SSC8836Q/SSC8629Q/SSC8336N/SSC8339DN/SSC333/SAC8539/SSC8826Q等	行车记录仪、其他车载摄像头、运动相机

4. 建议关注

4.4 全志科技

- 公司是领先的智能应用处理器SoC、高性能模拟器件和无线互联芯片设计厂商。公司目前的主营业务为系智能应用处理器SoC、高性能模拟器件和无线互联芯片的研发与设计。主要产品为智能应用处理器SoC、高性能模拟器件和无线互联芯片，产品广泛适用于智能硬件、平板电脑、智能家电、车联网、机器人、虚拟现实、网络机顶盒以及电源模拟器件、无线通信模组、智能物联网等多个产品领域。公司以客户为中心，凝聚卓越团队和坚持核心技术长期投入，在超高清视频编解码、高性能CPU/GPU/AI多核整合、先进工艺的高集成度、超低功耗、全栈集成平台等方面提供具有市场突出竞争力的系统解决方案和贴心服务。

公司SoC产品包



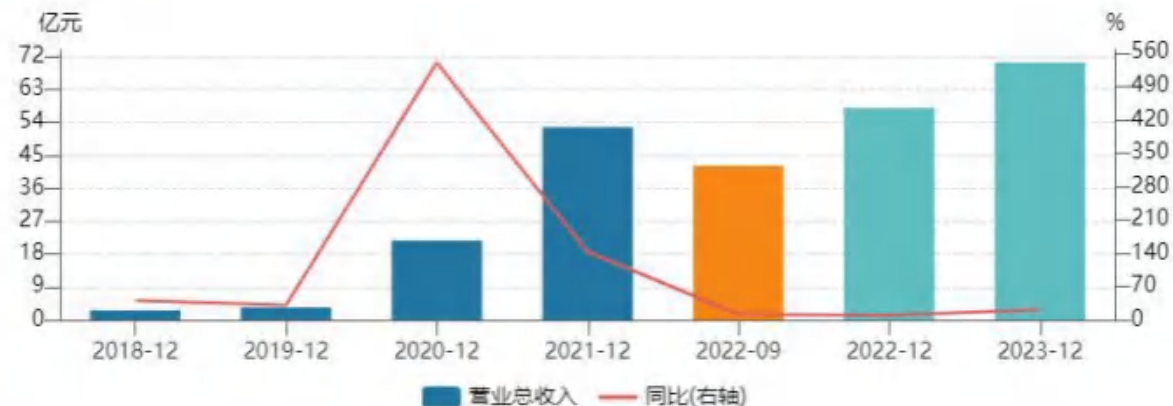
资料来源：公司年报，华金证券研究所

4. 建议关注

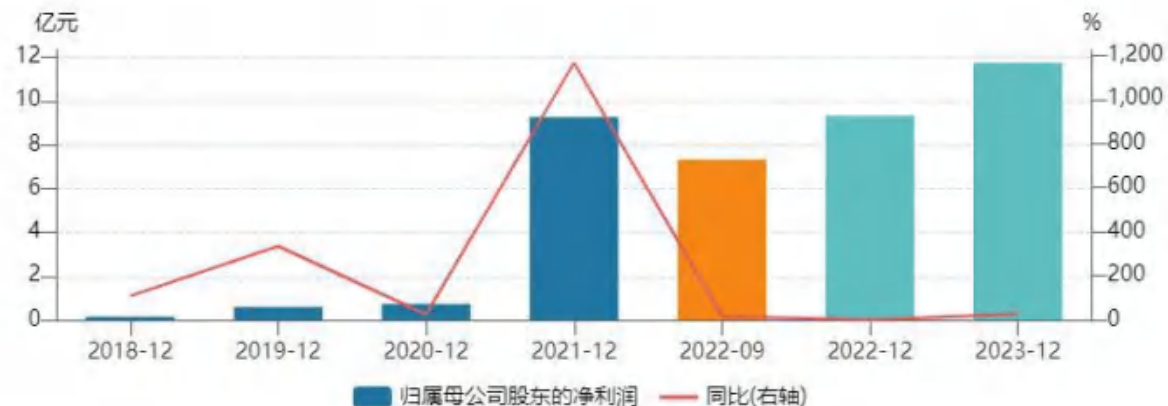
4.5 北京君正

- 公司是一家集成电路设计企业，拥有全球领先的32位嵌入式CPU技术和低功耗技术。公司主营业务为微处理器芯片、智能视频芯片等ASIC芯片产品及整体解决方案的研发和销售。公司拥有较强的自主创新能力，多年来在自主创新CPU技术、视频编解码技术、图像和声音信号处理技术、SoC芯片技术、软件平台技术等多个领域形成多项核心技术。公司已形成可持续发展的梯队化产品布局，基于自主创新的XBurst CPU和视频编解码等核心技术，公司推出了一系列具有高性价比的微处理器芯片产品和智能视频芯片产品，各类别的芯片产品分别面向不同的市场领域。公司积极培养复合型人才，形成合理的人才梯队，不断加强团队凝聚力，已累计获得多项专利证书，是国内外领先的嵌入式CPU芯片及解决方案提供商之一。

营业总收入及增长率



归属母公司股东的净利润及增长率



资料来源: wind , 华金证券研究所

4. 建议关注

4.6 中科蓝讯

- 公司是专注于研发、设计与销售无线音频SoC芯片的高科技公司。公司主营业务为无线音频SoC芯片的研发、设计与销售，主要产品包括TWS蓝牙耳机芯片、非TWS蓝牙耳机芯片、蓝牙音箱芯片等，产品可广泛运用于TWS蓝牙耳机、颈挂式耳机、头戴式耳机、商务单边蓝牙耳机、蓝牙音箱、车载蓝牙音响、电视音响等无线音频终端。自设立以来，公司始终专注于设计研发低功耗、高性能无线音频SoC芯片，产品已应用于知名手机品牌传音，进入飞利浦、联想、铁三角、创维、纽曼、山水、惠威、摩托罗拉等专业音频厂商，同时在夏新、Aukey、网易、唱吧、360、爱奇艺、QCY、天猫精灵等电商及互联网客户中占据重要地位，市场认可度高。在深耕无线音频芯片领域的基础上，公司持续推动技术升级、优化产品结构、拓展产品应用范围。通过持续的技术研发和市场开拓，目前公司部分芯片产品已应用至智能手表、智能车载支架等物联网产品中，丰富了公司产品的应用场景。2020年12月，公司获得“2020中国物联网技术创新奖”。

公司产品主要应用场景



4. 建议关注

4.7 富瀚微

- 公司成立于2004年4月，专注于视频监控芯片及解决方案，满足高速增长的数字视频监控市场对视频编解码和图像信号处理的芯片需求。公司提供高性能视频编解码SoC和图像信号处理器芯片，以及基于这些芯片的视频监控产品方案。公司致力于与国内外设备制造商、解决方案提供商建立紧密合作关系，共同把握市场契机，为客户提供高性价比的产品和服务，持续创造价值。公司是国家集成电路设计企业、上海市高新技术企业、上海市科技小巨人企业、上海市企业技术中心，公司先后承担多项国家、市级研发和产业化类项目，公司研发项目连续多年被认定为上海市高新技术成果转化百佳项目，并获中国半导体创新产品和技术奖、上海市科学技术奖等荣誉。

归属母公司股东的净利润及增长率（亿元）

	2021年		2020年		同比增减
	金额	占营业收入比重	金额	占营业收入比重	
营业收入合计	1,717,003,045.46	100%	610,247,904.24	100%	181.36%
分行业					
集成电路设计	1,717,003,045.46	100.00%	610,247,904.24	100.00%	181.36%
分产品					
专业安防产品	1,250,729,740.62	72.84%	286,235,032.20	46.91%	336.96%
智能硬件产品	278,498,561.02	16.22%	135,253,885.49	22.16%	105.91%
汽车电子产品	175,078,805.93	10.20%	72,993,400.01	11.96%	139.86%
技术服务	2,226,243.18	0.13%	69,625,894.98	11.41%	-96.80%
其他	10,469,694.71	0.61%	46,139,691.56	7.56%	-77.31%
分地区					
境内	1,478,065,575.75	86.08%	467,273,628.59	76.57%	216.32%
境外	238,937,469.71	13.92%	142,974,275.65	23.43%	67.12%
分销售模式					
直销	1,235,617,609.62	71.96%	394,348,394.89	64.62%	213.33%
分销	481,385,435.84	28.04%	215,899,509.35	35.38%	122.97%




资料来源：公司年报，华金证券研究所

4. 建议关注

4.8 恒玄科技

- 公司主营业务为智能音视频SoC芯片的研发、设计与销售，为客户提供AIoT场景下具有语音交互能力的边缘智能主控平台芯片，产品广泛应用于智能蓝牙耳机、WiFi智能音箱、智能手表等低功耗智能音视频终端产品。公司产品已经进入三星、华为、OPPO、小米等全球主流安卓手机品牌，同时也进入包括哈曼、SONY、Skullcandy、漫步者、万魔等专业音频厂商，并在谷歌、阿里、百度等互联网公司的智能音频产品中得到应用。品牌客户的深度及广度是公司重要的竞争优势和商业壁垒。公司主要产品为蓝牙音频芯片、WiFi SoC芯片，并逐步拓展到智能手表芯片。公司智能音视频SoC芯片能够集成多核异构CPU、WiFi/蓝牙基带和射频、音频CODEC、电源管理、存储、嵌入式语音AI和主动降噪等多个功能模块，是智能音视频设备的主控平台芯片。

公司产品系列

产品系列一	 普通蓝牙 音频芯片	BES2000 系列 BES2000 series	<ul style="list-style-type: none">• 700MHz• 低功耗• 集成式芯片• 蓝牙音频 <ul style="list-style-type: none">• TWS earbuds• Headset earphones• Headphones• Bluetooth speakers	 蓝牙耳机
产品系列二	 智能蓝牙 音频芯片	BES2300 系列 BES2300 series	<ul style="list-style-type: none">• 智能TWS耳机• 语音唤醒功能• 集成式芯片• 智能音箱 <ul style="list-style-type: none">• Smart TWS earbuds• Smart neckband earphones• Smart headphones• Smart speakers	 智能耳机
产品系列三	 Type-C 音频芯片	BES3000 系列 BES3000 series	<ul style="list-style-type: none">• Type-C耳机• Type-C蓝牙音箱 <ul style="list-style-type: none">• Type-C headphones• Type-C audio speakers	 Type-C耳机

资料来源：公司官网，华金证券研究所

- 01 由专用走向通用，GPU赛道壁垒高筑
- 02 产业化路径显现，全球AI竞赛再加速
- 03 全维智能化大时代，国产算力行则必至
- 04 建议关注
- 05 产业相关
- 06 风险提示

5. 产业相关

5.1 海光信息

➤ 公司的主营业务是研发、设计和销售应用于服务器、工作站等计算、存储设备中的高端处理器。公司的产品包括海光通用处理器（CPU）和海光协处理器（DCU）。根据我国信息产业发展的实际需要，公司研发出了多款性能达到国际同类型主流高端处理器水平的产品。公司专注于高端处理器的研发、设计与技术创新，掌握了高端处理器核心微结构设计、高端处理器SoC架构设计、处理器安全、处理器验证、高主频与低功耗处理器实现、高端芯片IP设计、先进工艺物理设计、先进封装设计、基础软件等关键技术。秉承“销售一代、验证一代、研发一代”的产品研发策略，公司建立了完善的高端处理器的研发环境和流程，产品性能逐代提升，功能不断丰富，已经研发出可广泛应用于服务器、工作站的高端处理器产品。

产品类 型	处理器种 类	指令集	主要产品	产品特征	典型应用场景
海光 CPU	通用处理 器	兼容x86指 令集	海光3000系列	内置多个处理器核心，集成通用的高性能外 设接口，拥有完善的软硬件生态环境和完备 的系统安全机制，适用于数据计算和事务处 理等通用型应用	云计算、物联 网、信息服务等
			海光5000系列		
			海光7000系列		
海光 DCU	协处理器	兼容“类 CUDA”环 境	海光8000系列	内置大量运算核心，具有较强的并行计算能 力和较高的能效比，适用于向量计算和矩阵 计算等计算密集型应用	大数据处理、人 工智能、商业计 算等

资料来源：公司招股书，华金证券研究所

5. 产业相关

5.2 龙芯中科

- 公司主营业务为处理器及配套芯片的研制、销售及服务，主要产品与服务包括处理器及配套芯片产品与基础软硬件解决方案业务。目前，龙芯中科基于信息系统和工控系统两条主线开展产业生态建设，面向网络安全、办公与业务信息化、工控及物联网等领域与合作伙伴保持全面的市场合作，系列产品在电子政务、能源、交通、金融、电信、教育等行业领域已获得广泛应用。
- 龙芯中科研制的芯片包括龙芯1号、龙芯2号、龙芯3号三大系列处理器芯片及桥片等配套芯片。
- 公司发布2022年业绩预告，预计2022年年度实现营业收入为78,000万元到82,000万元，与上年同期相比，将减少42,125万元到38,125万元，同比减少35%到32%。预计2022年年度实现归属于母公司所有者的净利润为5,000万元到7,000万元，与上年同期相比，将减少18,680万元到16,680万元，同比减少79%到70%。

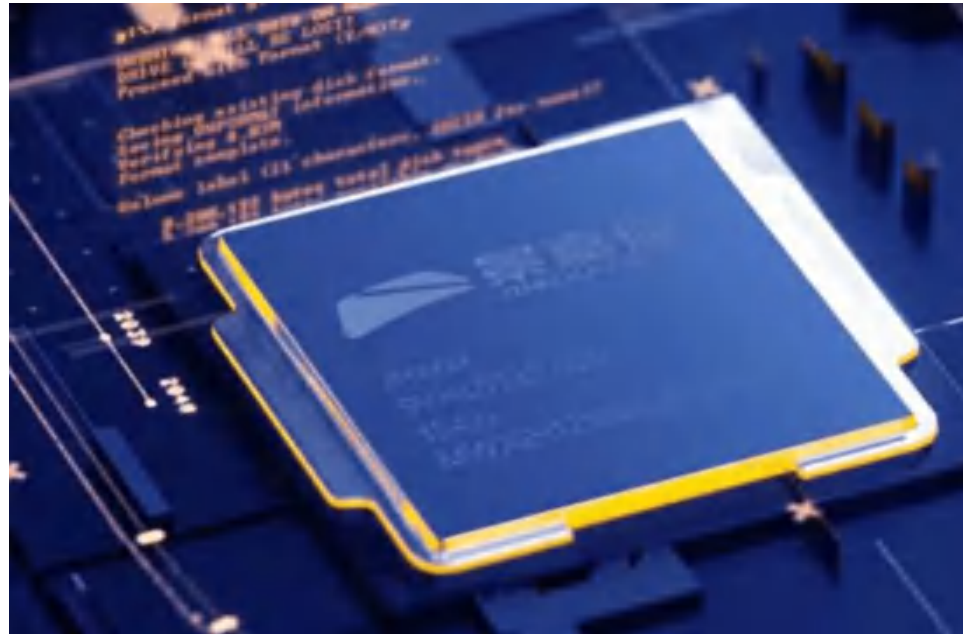
序号	产业领域	系列	型号	推出时间	简介	主要应用场景
1	工控类	龙芯1号	龙芯1B	2012年	面向数据采集和网络设备应用的SoC芯片，集成DDR2、LCD、USB2.0、MAC等接口	远程数据采集，以太网交换机，小型通信终端机，电表集中器等
2			龙芯1C300(龙芯1C)	2014年	面向工控和物联网应用的SoC芯片，集成SDRAM、LCD、OTG、MAC、NFC等接口	打印机、地理信息探测仪等
3			龙芯1C101	2018年	面向智能门锁等应用的MCU芯片，集成Flash、TSensor、VPM、ADC等功能模块	门锁应用等
4		龙芯2号	龙芯2K1000	2018年	64位双核SoC芯片，主频1.0GHz，集成DDR2/3、PCIe2.0、SATA2.0、USB2.0、DVO等接口	交换机、边缘网关、工业防火墙、工业平板、智能变电站、柱号自动机等
5			龙芯2K1000LA	2022年	64位双核SoC芯片，主频1.0GHz，基于LA264处理器核，集成DDR2/3、PCIe2.0、SATA2.0、USB2.0、DVO等接口	交换机、边缘网关、工业防火墙、工业平板、智能变电站、柱号自动机等
6			龙芯2K0500	2022年	64位单核SoC芯片，主频500MHz，集成DDR3、2D GPU、DVO、PCIe2.0、SATA2.0、USB2.0、USB3.0、GMAC、PCI、彩色黑白打印接口、HDA及其他常用接口	工控互联网应用、打印终端、HDC等

序号	产业领域	系列	型号	推出时间	简介	主要应用场景
7	信息化和工控类	龙芯3号	龙芯3A3000	2017年	64位四核处理器，主频1.2~1.3GHz，集成双通道DDR3-1600和HT3.0接口	桌面与终端类应用
8			龙芯3A4000	2019年	64位四核处理器，主频1.8~2.0GHz，集成双通道DDR4-2400和HT3.0接口	桌面与终端类应用
9			龙芯3A5000	2021年	64位四核处理器，主频2.1~2.5GHz，采用全新的LoongArch指令系统，集成双通道DDR4-3200和HT3.0接口	桌面与终端类应用
10		龙芯3号	龙芯3C5000	2021年	64位十六核处理器，主频2.0~2.2GHz，采用全新的LoongArch指令系统，通过MCM封装，集成四个3A5000芯片，集成四通道DDR4-3200和HT3.0接口，最高支持四路互联	服务器类应用
11			龙芯3C5000	2022年	64位十六核处理器，主频2.0~2.2GHz，采用全新的LoongArch指令系统，片上集成16个高性能LA364处理器核，集成四通道DDR4-3200和HT3.0接口，最高支持十六路互联	服务器类应用
12		配套芯片	龙芯7A1000	2018年	龙芯3号处理器的配套桥片，通过HT3.0接口与处理器相连，外接接口包括PCIe2.0、GMAC、SATA2.0、USB2.0和其他低速接口	与龙芯3号系列配套使用
13			龙芯7A2000	2022年	第二代龙芯3号系列处理器配套桥片，通过HT3.0接口与处理器相连，外接接口包括PCIe2.0、USB3.0、SATA6.0，显示接口支持为三路DP和一路DVI，可直连显示器；内置一个网络PHY，直接提供网络接口输出；片内集成了门控寄存器，采用统一编程架构，搭配32位DDR4内存接口，最大支持16GB显存容量	与龙芯3号系列配套使用

5. 产业相关

5.3 景嘉微

- 公司致力于信息探测、信息处理和信息传递领域的技术和综合应用，为客户提供高可靠、高品质的解决方案、产品和配套服务，是国内成功自主研发国产化图形处理芯片(GPU)并产业化的企业。公司主要从事高可靠电子产品的研发、生产和销售，产品主要涉及图形显控、小型专用化雷达和其他三大领域。图形显控是公司现有核心业务，也是传统优势业务，小型专用化雷达和芯片是公司未来大力发展的业务方向。公司在图形显控领域拥有图形显控模块、图形处理芯片、加固显示器、加固存储和加固计算机等五类产品，其中图形显控模块是公司最为核心的产品。公司以ITR流程为指导，建立了现场服务工程师、公司售后服务部、产品生命周期管理团队三级技术服务保障体系。



5. 产业相关

5.4 寒武纪-U

➤ 公司自成立以来一直专注于人工智能芯片产品的研发与技术创新，致力于打造人工智能领域的核心处理器芯片，让机器更好地理解和服务人类。公司核心人员在处理器芯片和人工智能领域深耕十余年，带领公司研发了智能处理器指令集与微架构等一系列自主创新关键技术。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。采用公司终端智能处理器IP的终端设备已出货过亿台；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。

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 1M 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	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

资料来源：公司招股书，华金证券研究所

5. 产业相关

5.5 中芯国际

- 公司是全球领先的集成电路晶圆代工企业之一，也是中国大陆技术最先进、规模最大、配套服务最完善、跨国经营的专业晶圆代工企业，主要为客户提供0.35微米至14纳米多种技术节点、不同工艺平台的集成电路晶圆代工及配套服务，在逻辑工艺领域，中芯国际是中国大陆第一家实现14纳米FinFET量产的晶圆代工企业，代表中国大陆自主研发集成电路制造技术的最先进水平；在特色工艺领域，中芯国际陆续推出中国大陆最先进的24纳米NAND、40纳米高性能图像传感器等特色工艺，与各领域的龙头公司合作，实现在特殊存储器、高性能图像传感器等细分市场的持续增长，除集成电路晶圆代工业务外，中芯国际亦致力于打造平台式的生态服务模式，为客户提供设计服务与IP支持、光掩模制造、凸块加工及测试等一站式配套服务，并促进集成电路产业链的上下游合作，与产业链各环节的合作伙伴一同为客户提供全方位的集成电路解决方案。

一站式服务



资料来源：公司官网，华金证券研究所

5. 产业相关

5.6 芯原股份-U

- 芯原是一家依托自主半导体IP，为客户提供平台化、全方位、一站式芯片定制服务和半导体IP授权服务的企业。公司至今已拥有高清视频、高清音频及语音、车载娱乐系统处理器、视频监控、物联网连接、数据中心等多种一站式芯片定制解决方案，以及自主可控的图形处理器IP、神经网络处理器IP、视频处理器IP、数字信号处理器IP和图像信号处理器IP五类处理器IP、1,400多个数模混合IP和射频IP。主营业务的应用领域广泛包括消费电子、汽车电子、计算机及周边、工业、数据处理、物联网等，主要客户包括IDM、芯片设计公司，以及系统厂商、大型互联网公司。
- 公司目前拥有GPU、NPU、VPU、DSP和ISP五类处理器IP、1,400多个数模混合IP和射频IP。



5. 产业相关

5.7 华大九天

- 公司主要从事EDA工具软件的开发、销售及相关服务。EDA工具是集成电路领域的上游基础工具，应用于集成电路设计、制造、封装、测试等产业链各个环节，是集成电路产业的战略基础支柱之一。公司主要产品包括模拟电路设计全流程EDA工具系统、数字电路设计EDA工具、平板显示电路设计全流程EDA工具系统和晶圆制造EDA工具等EDA工具软件，并围绕相关领域提供技术开发服务。公司相关产品和服务主要应用于集成电路设计及制造领域。公司曾荣获“第二届集成电路产业技术创新奖(成果产业化奖)”、“中国半导体创新产品和技术奖”、“第八届中国电子信息博览会创新奖”等多项荣誉。

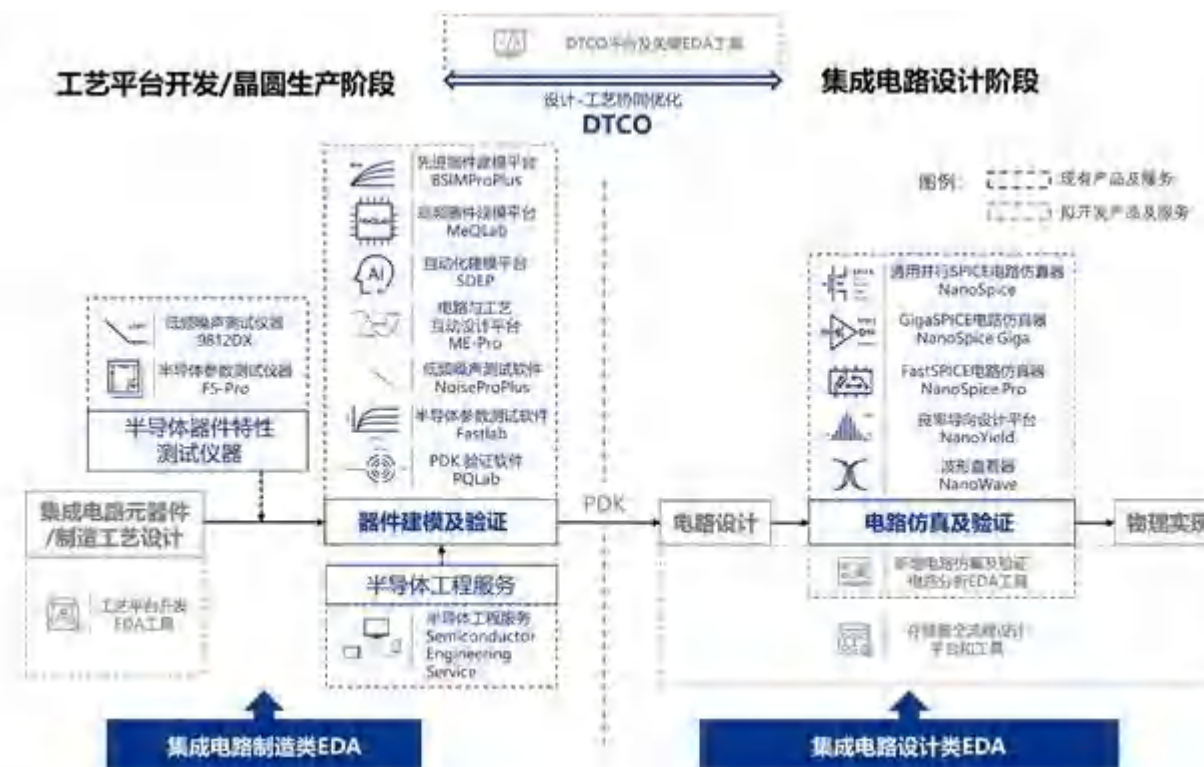


5. 产业相关

5.8 概伦电子

- 公司是一家具备国际市场竞争力的EDA企业，拥有领先的EDA关键核心技术，致力于提高集成电路行业的整体技术水平和市场价值，提供专业高效的EDA流程和工具支撑。公司的主营业务为向客户提供被全球领先集成电路设计和制造企业长期广泛验证和使用的EDA产品及解决方案，主要产品及服务包括制造类EDA工具、设计类EDA工具、半导体器件特性测试仪器和半导体工程服务等，公司通过EDA方法学创新，推动集成电路设计和制造的深度联动，加快工艺开发和芯片设计进程，提高集成电路产品的良率和性能，增强集成电路企业整体市场竞争力。

公司主要产品和服务



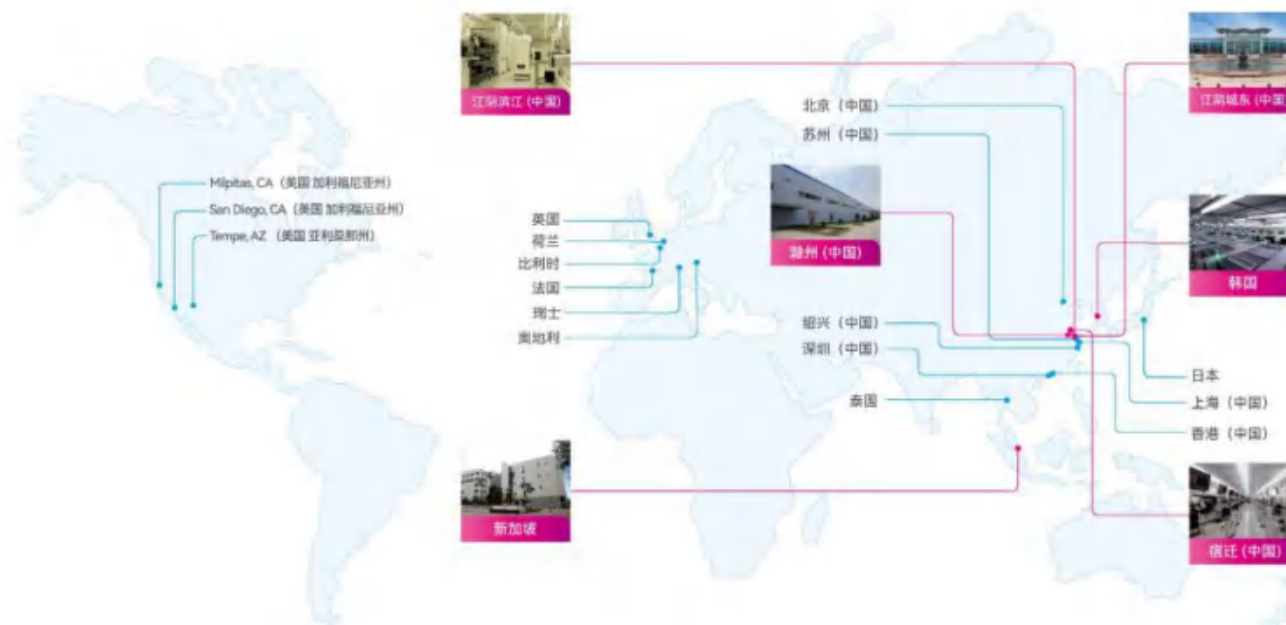
资料来源：公司招股书，华金证券研究所

5. 产业相关

5.9 长电科技

- 公司全球知名的集成电路封装测试企业。公司面向全球提供封装设计、产品开发及认证，以及从芯片中测、封装到成品测试及出货的全套专业生产服务。长电科技致力于可持续发展战略，崇尚员工、企业、客户、股东和社会和谐发展，合作共赢之理念，先后被评定为国家重点高新技术企业，中国电子百强企业，集成电路封装技术创新战略联盟理事长单位，中国出口产品质量示范企业等，拥有国内高密度集成电路国家工程实验室、国家级企业技术中心、博士后科研工作站等。公司生产、研发和销售网络已覆盖全球主要半导体市场。公司具有广泛的技术积累和产品解决方案，包括有自主知识产权的Fan-out eWLB、WLCSP、Bump、PoP、fcBGA、SiP、PA等封装技术，另外引线框封装及自主品牌的分立器件也深受客户褒奖。

长电科技全球布局



资料来源：公司官网，华金证券研究所

5. 产业相关

5.10 华天科技

- 公司主要从事半导体集成电路、MEMS传感器、半导体元器件的封装测试业务。目前公司集成电路封装产品主要有DIP/SDIP、SOT、SOP、SSOP、TSSOP/ETSSOP、QFP/LQFP/TQFP、QFN/DFN、BGA/LGA、FC、MCM(MCP)、SiP、WLP、TSV、Bumping、MEMS等多个系列，产品主要应用于计算机、网络通讯、消费电子及智能移动终端、物联网、工业自动化控制、汽车电子等电子整机和智能化领域。公司不断加强先进封装技术和产品的研发力度，加大研发投入，完善以华天西安为主体的研发仿真平台建设，依托国家级企业技术中心、甘肃省微电子工程技术研究中心、甘肃省微电子工程实验室等研发验证平台，通过实施国家科技重大专项02专项等科技创新项目以及新产品、新技术、新工艺的不断研究开发，自主研发出FC、Bumping、MEMS、MCM(MCP)、WLP、SiP、TSV、Fan-Out等多项集成电路先进封装技术和产品，随着公司进一步加大技术创新力度，公司的技术竞争优势将不断提升。

公司厂区



资料来源：公司官网，华金证券研究所

5. 产业相关

5.11 通富微电

- 公司是由南通华达微电子有限公司和富士通(中国)有限公司共同投资、由中方控股的中外合资股份制企业，专业从事集成电路封装测试。公司目前的封装技术包括Bumping、WLCSP、FC、BGA、SiP等先进封测技术，QFN、QFP、SO等传统封装技术以及汽车电子产品、MEMS等封装技术；测试技术包括圆片测试、系统测试等。公司拥有国家认定企业技术中心、国家博士后科研工作站、江苏省企业院士工作站、省级工程技术研究中心和企业研究院等高层次研发平台，拥有2000多人的技术管理团队。在行业内率先通过ISO9001、ISO/TS16949等质量体系。采用SAP、MES、设备自动化、EDI等信息系统，可按照客户个性化的规范自动控制生产过程，实时和客户进行信息交互。

公司七大基地



资料来源：公司官网，华金证券研究所

5. 产业相关

5.12 炬芯科技

- 公司是中国领先的低功耗系统级芯片设计厂商，主营业务为中高端智能音频SoC芯片的研发、设计及销售，专注于为无线音频、智能穿戴及智能交互等智慧物联网领域提供专业集成芯片。公司的主要产品为蓝牙音频SoC芯片系列、便携式音视频SoC芯片系列、智能语音交互SoC芯片系列等，广泛应用于蓝牙音箱、蓝牙耳机、蓝牙语音遥控器、蓝牙收发一体器、智能教育、智能办公、智能家居等领域。公司的智能音频SoC芯片产品占据我国市场重要地位，已成为和音频相关的低功耗无线物联网领域的主流供应商，并已逐步实现相关芯片领域的国产替代，产品已进入的主要终端品牌包括华为、哈曼、SONY、安克创新、罗技、OPPO、小米、传音、飞利浦、漫步者、联想、纽曼、魅族等，并在阿里巴巴、网易和酷我等互联网公司的音频产品中得到应用。

蓝牙音箱下游终端品牌名单		
年份	首次进入终端品牌	说明
2021年1-6月	Vizio	持续扩展电视厂商客户群体
2020年	ION、绿联	进入电视厂商及其它更广泛细分市场终端品牌的供应链
2019年	华为、荣耀、天猫精灵、小鸟听、沃尔玛、OPPO、铁三角、飞利浦	首次进入华为、天猫精灵、沃尔玛的供应链，分别进入手机厂商、互联网公司和商超渠道
2018年	联想、京东、夏普、漫步者、昂思(Oontz)	终端品牌逐渐扩大到专业音频厂商之外，如联想、京东、夏普等
2017年	哈曼、SONY、安克创新、魅族	首次进入哈曼、SONY等国际一线终端品牌的供应链，并进入安克创新等跨境电商品牌的供应链
2016年	小米、Creative	首次进入小米供应链，并逐渐进入国际二线品牌的供应链
2015年	朗琴、不见不散、Doss	以国内终端品牌为主

资料来源：公司招股书，华金证券研究所

5. 产业相关

5.13 源杰科技

➤ 公司聚焦于光芯片行业，主营业务为光芯片的研发、设计、生产与销售，主要产品包括2.5G、10G和25G及更高速率激光器芯片系列产品等，目前主要应用于光纤接入、4G/5G移动通信网络和数据中心等领域。公司采取以直销为主、经销为辅的销售模式，设立市场与销售部负责开发客户、产品推广以及维护客户关系。市场与销售部根据客户需求情况制定销售计划，将接到的订单需求反馈给生产与运营部，协调产品研发、生产、交付、质量等服务工作，同时承担跟单、售后、技术支持等工作。新产品及客户导入方面，由于光芯片产品设计参数、性能指标多，公司市场与销售部根据客户需求先与其进行深度技术交流，研发部在此基础上进行产品设计、材料选型、样品生产等工作，然后在厂内进行样品性能测试、可靠性测试，并将样品送至客户处进行综合测试。测试通过后，客户会小批量下单采购，并在多批次生产合格后，转入批量采购公司的成熟产品主要通过展会、现有客户推荐、销售经理开发等方式寻求新客户。

公司主要产品类型

产品速率	产品类型	应用领域
2.5G	1310nm DFB激光器芯片	光纤接入 PON (GPON) 光纤接入 10G-PON (XG-PON) 光纤接入 40km/80km
	1490nm DFB激光器芯片	
	1270nm DFB激光器芯片	
	1550nm DFB激光器芯片	
10G	1270nm DFB激光器芯片	光纤接入 10G-PON (XGS-PON) 4G移动通信网络 4G/5G移动通信网络
	1310nm FP激光器芯片	
	1310nm DFB激光器芯片	
	CWDM 6波段 DFB激光器芯片	
25G	CWDM 6波段 DFB激光器芯片	5G移动通信网络
	LWDM 12波段 DFB激光器芯片	
	MWDM 12波段 DFB激光器芯片	
	CWDM 4波段 DFB激光器芯片	
50G	LWDM 4波段 DFB激光器芯片	数据中心 100G 数据中心 200G
	PAM4 CWDM 4波段 DFB激光器芯片	
	1270/1290/1310/1330nm大功率25/50/70mW激光器芯片	
	硅光直流感光源	
		数据中心 100G/200G/400G

资料来源：公司招股书，华金证券研究所

5. 产业相关

5.14 光迅科技

- 公司是中国最大光通信器件供货商，是目前中国唯一一家有能力对光电子器件进行系统性，战略性研究开发的高科技企业，是中国光电子器件行业最具影响的实体之一。国家高技术研究发展计划成果产业基地武汉光通信与光传感材料及器件成果产业化基地的主要建设单位之一，并被国家科学技术部火炬高技术产业开发中心认定为“国家火炬计划重点高新技术企业”，主要从事光通信领域内光电子器件的研究、开发、制造和技术服务。先后承担国家“863”、“973”、国家科技攻关等项目数十余项。中兴通讯、华为技术、烽火通信为代表的国内通信系统设备厂商已成为公司稳定的客户。

公司研发方向（2021年年报）

主要研发项目名称	项目目的	项目进展	拟达到的目标	预计对公司未来发展的影响
高速激光器芯片	推进200/400/800G光通信用激光器芯片产品化开发	按计划进行	形成供应能力	提升公司产品竞争力和供应链连续性
高速探测器芯片	推进200/400/800G光通信用探测器及阵列芯片产品化开发	按计划进行	形成供应能力	提升公司产品竞争力和供应链连续性
相干光器件	开发核心相干用光芯片及器件	按计划进行	形成供应能力	形成新的市场突破点
超光谱光放大器	扩大光纤传输带宽应用	按计划进行	向L++波段继续扩展	形成新的市场突破点
数据中心用高端光模块	提升封装密度和光路耦合效率，降低单位bit数据传输能耗。推进CPO光电合封技术开发	按计划进行	保持技术演进方向，继续研发高速率产品	形成技术领先优势，抓住市场热点，扩大市场份额
波长选择开关	实现智能化ROADM系统应用	按计划进行	形成供应能力	形成新的市场突破点
50GPON器件	开发下一代高速光接入芯片及器件	按计划进行	形成供应能力	形成新的市场突破点
窄线宽激光器	开发低噪声窄线宽激光器，应用传感及新型通信领域	按计划进行	形成供应能力	扩展新业务领域
激光雷达用器件	开发激光光源芯片及器件	按计划进行	达到车规级可靠性的要求	扩展新业务领域

资料来源：公司年报，华金证券研究所

5.15 摩尔线程（未上市）

➤ 摩尔线程成立于2020年10月，由英伟达前全球副总裁、中国区总经理张建中创立。公司专注于研发设计全功能GPU芯片及相关产品，支持3D图形渲染、AI训练与推理加速、超高清视频编解码、物理仿真与科学计算等多种组合工作负载，兼顾算力与算效，能够为中国科技生态合作伙伴提供强大的计算加速能力，广泛赋能数字经济多个领域。2022年3月31日，摩尔线程公布了首批显卡产品：面向电脑和工作站的MTT S60以及面向服务器的MTT S2000。两张显卡都采用了第一代MUSA架构（Moore Threads Unified System Architecture，摩尔线程统一系统架构，中文名为“苏提”）；2022年11月，摩尔线程公布了第二批显卡产品：面向电脑和工作站的显卡MTT S80以及面向服务器的MTT S3000；采用了新一代MUSA架构“春晓”，并使用了PCIe Gen5插槽。

公司主要产品型号

	MUSA MTT S60	MUSA MTT S80
公布时间	2022年3月	2022年11月
制程工艺	12nm	7nm
MUSA GPU核心数	2,048	4,096
性能	6 TFLOPS (192 GPix/s fill rate)	14.4 TFLOPS (460 GPix/s fill rate)
显存	8GB LPDDR4X	16GB GDDR6
API支持	DirectX, Vulkan, OpenGL, OpenGL ES	DirectX, Vulkan, OpenGL, OpenGL ES
系统支持	X86/ARM/LoongArch; Windows/Linux	X86/ARM/LoongArch; Windows/Linux
输出接口	DisplayPort 1.4, up to 8K	DisplayPort 1.4a X3, HDMI 2.1 X1, up to 8K

资料来源：维基百科，华金证券研究所

- 01 由专用走向通用，GPU赛道壁垒高筑
- 02 产业化路径显现，全球AI竞赛再加速
- 03 全维智能化大时代，国产算力行则必至
- 04 建议关注
- 05 产业相关
- 06 风险提示

6. 风险提示

技术创新风险：随着下游市场对产品的性能需求不断提升，集成电路设计行业技术升级和产品更新换代速度较快，企业需紧跟市场发展步伐，及时对现有产品及技术进行升级换代，以维持其市场地位。同时，集成电路产品的发展方向具有一定不确定性，因此企业需要对主流技术迭代趋势保持较高的敏感度，根据市场需求变动和工艺水平发展制定动态的技术发展战略。未来若公司技术研发水平落后于行业升级换代水平，或公司技术研发方向与市场发展趋势偏离，将导致公司研发资源浪费并错失市场发展机会，对公司产生不利影响。

宏观经济和行业波动风险：集成电路行业是面临全球化的竞争与合作并得到国家政策大力支持的行业，受到国内外宏观经济、行业法规和贸易政策等宏观环境因素的影响。近年来，全球宏观经济表现平稳，国内经济稳中有升。未来，如果国内外宏观环境因素发生不利变化，可能会对公司经营带来不利影响。

国际贸易摩擦风险：伴随全球产业格局的深度调整，国际贸易摩擦不断，集成电路产业成为贸易冲突的重点领域，也对我国相关产业的发展造成了客观不利影响。2022年8月以来，美国推出了多项贸易管制政策通过限制产品、设备以及技术等项目的出口以限制中国半导体行业的发展。

华金电子-走进“芯”时代系列深度报告

- 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_存储器深度《存储产业链战略升级，开启国产替代“芯”篇章》
- 6、芯时代之六_功率半导体深度《功率半导体处黄金赛道，迎进口替代良机》
- 7、芯时代之七_半导体材料深度《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》

华金电子-走进“芯”时代系列深度报告

- 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_功率半导体深度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《半导体硅片厚积薄发，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》

华金电子-走进“芯”时代系列深度报告

- 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《双碳助力发展, GPU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》
- 50、芯时代之五十_江海股份《乘“碳中和”之风,老牌企业三大电容全面发力》
- 51、芯时代之五十一_智能电动车1000页PPT(多行业协同)《智能电动车★投研大全》

华金电子-走进“芯”时代系列深度报告

- 52、芯时代之五十二_瑞芯微PPT深度《迈入全球准一线梯队，新硬件十年前景可期》
- 53、芯时代之五十三_峰岹科技《专注BLDC电机驱动控制芯片，三大核心技术引领成长》
- 54、芯时代之五十四_纳芯微《专注高端模拟IC，致力国内领先车规级半导体供应商》
- 55、芯时代之五十五_晶晨股份《核心技术为躯，全球开拓为翼》
- 56、芯时代之五十六_国微&复微《紫光国微与复旦微的全面对比分析》
- 57、芯时代之五十七_国产算力SoC《算力大时代，处理器SoC厂商综合对比》
- 58、芯时代之五十八_高能模拟芯《高性能模拟替代渐入深水区，工业汽车重点突破》
- 59、芯时代之五十九_南芯科技《电荷泵翘楚拓矩阵蓝图，通用产品力屡复制成功》
- 60、芯时代之六十_AI算力GPU《AI产业化再加速，智能大时代已开启》

华金证券研究所电子团队简介

孙远峰：华金证券总裁助理&研究所所长&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、金牛奖等奖项最佳分析师；2019年开始未参加任何个人评比，其骨干团队专注于创新&创业型研究所的一线具体创收&创誉工作，以“产业资源赋能深度研究”为导向，构建研究&销售合伙人队伍，积累了健全的成熟团队自驱机制和年轻团队培养机制，充分获得市场验证；清华校友总会电子工程系分会副秘书长；

王臣复：电子行业高级分析师，北京航空航天大学工学学士和管理学硕士，2年半导体产业一级股权投资经历，曾就职于华西证券研究所、欧菲光集团投资部、融通资本、平安基金等，2023年2月加入华金证券研究所；

王海维：电子行业高级分析师，华东师范大学硕士，电子&金融复合背景，主要覆盖半导体板块，善于个股深度研究，2018年新财富上榜分析师（第3名）核心成员，先后任职于安信证券/华西证券研究所，2023年2月入职华金证券研究所；

行业评级体系

收益评级：

领先大市 — 未来6个月的投资收益率领先沪深300指数10%以上；

同步大市 — 未来6个月的投资收益率与沪深300指数的变动幅度相差-10%至10%；

落后大市 — 未来6个月的投资收益率落后沪深300指数10%以上；

风险评级：

A — 正常风险，未来6个月投资收益率的波动小于等于沪深300指数波动；

B — 较高风险，未来6个月投资收益率的波动大于沪深300指数波动。

分析师声明

孙远峰/王臣复/王海维声明，本人具有中国证券业协会授予的证券投资咨询执业资格，勤勉尽责、诚实守信。本人对本报告的内容和观点负责，保证信息来源合法合规、研究方法专业审慎、研究观点独立公正、分析结论具有合理依据，特此声明。

本公司具备证券投资咨询业务资格的说明

华金证券股份有限公司（以下简称“本公司”）经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司及其投资咨询人员可以为证券投资人或客户提供证券投资分析、预测或者建议等直接或间接的有偿咨询服务。发布证券研究报告，是证券投资咨询业务的一种基本形式，本公司可以对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向本公司的客户发布。

法律声明

免责声明：

本报告仅供华金证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因为任何机构或个人接收到本报告而视其为本公司的当然客户。

本报告基于已公开的资料或信息撰写，但本公司不保证该等信息及资料的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映本公司于本报告发布当日的判断，本报告中的证券或投资标的价格、价值及投资带来的收入可能会波动。在不同时期，本公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，本公司将随时补充、更新和修订有关信息及资料，但不保证及时公开发布。同时，本公司有权对本报告所含信息在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以本公司向客户发布的本报告完整版本为准。

在法律许可的情况下，本公司及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，提请客户充分注意。客户不应将本报告为作出其投资决策的惟一参考因素，亦不应认为本报告可以取代客户自身的投资判断与决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，无论是否已经明示或暗示，本报告不能作为道义的、责任的和法律的依据或者凭证。在任何情况下，本公司亦不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告版权仅为本公司所有，未经事先书面许可，任何机构和个人不得以任何形式翻版、复制、发表、转发、篡改或引用本报告的任何部分。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“华金证券股份有限公司研究所”，且不得对本报告进行任何有悖原意的引用、删节和修改。

华金证券股份有限公司对本声明条款具有惟一修改权和最终解释权。

法律声明

风险提示：

报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价或询价。投资者对其投资行为负完全责任，我公司及其雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。

华金证券股份有限公司

办公地址：

上海市浦东新区杨高南路759号陆家嘴世纪金融广场30层

北京市朝阳区建国路108号横琴人寿大厦17层

深圳市福田区益田路6001号太平金融大厦10楼05单元

电话：021-20655588

网址：www.huajinsc.cn

致谢



欢迎关注“远峰电子”公众号