中信建投证券 CHINA SECURITIES

证券研究报告•GPU行业深度研究

AI大模型浪潮风起，GPU芯片再立潮头

分析师：阎贵成
yanguicheng@csc.com.cn
SAC编号：S1440518040002
SFC 中央编号：BNS315

分析师：金戈
jinge@csc.com.cn
SAC编号：S1440517110001
SFC 中央编号：BPD352

分析师：于芳博
yufangbo@csc.com.cn
SAC编号：S1440522030001

发布日期：2023年3月26日

核心观点

- **核心观点：** GPU具备图形渲染和并行计算两大核心功能，其应用场景主要包括个人电脑、服务器、自动驾驶、移动端。全球GPU市场保持良好成长性，AI服务器成为市场增长的核心支撑，随着生成式AI大模型进入到辅助生产力阶段，服务器GPU市场需求更为旺盛。英伟达凭借其数据中心GPU的核心技术优势，成为全球人工智能芯片的引领者。AMD作为全球领先的芯片设计厂商，在GPU市场中与英伟达互相角逐。国内GPU市场空间广阔，涌现出一批优秀的GPU设计和制造厂商。

- **GPU具备图形渲染和并行计算两大核心功能。**GPU具有数量众多的运算单元，适合计算密集、易于并行的程序，一般作为协处理器负责图形渲染和并行计算。GPU微架构由流处理器、纹理映射单元、光栅化处理单元、光线追踪核心、张量核心、缓存等部件共同组成，微架构的设计对GPU性能的提升发挥着至关重要的作用，也是GPU研发过程中最关键的技术壁垒。GPU应用程序接口（API）帮助GPU高效实现渲染功能，在并行计算方面，CUDA（统一计算设备架构）的诞生大幅降低GPGPU并行计算的编程难度，实现GPU的通用化，“个人计算机”变成可以实现并行运算的“超级计算机”。

- **全球GPU市场保持良好成长性，AI服务器成为市场增长的核心支撑。**2023年GPU全球市场规模预计为595亿美元，行业保持高速增长，CAGR为32.9%。GPU的市场主体可以分为个人电脑GPU、服务器GPU、自动驾驶GPU、移动端GPU。过去的几个季度里，个人电脑GPU市场遭受巨大冲击，出货量显著下滑。核心原因有三点：一、个人电脑市场处于下行周期；二、虚拟货币挖矿退潮对独立GPU出货造成巨大冲击；三、下游板卡厂商开启降库存周期。近期，各类不利因素正在逐渐消融，个人电脑GPU市场迎来曙光。服务器GPU主要用于AI和高性能计算，人工智能行业的高速发展带来了旺盛的AI算力需求，AI服务器成为GPU市场增长的核心支撑。以ChatGPT为代表的自然语言大模型展现出高度智能，生成式AI能力不断突破进入到辅助生产力阶段，AI模型算力需求迈上新台阶，对服务器GPU市场带来显著拉动效应。自动驾驶GPU在高等级自动驾驶中具备显著技术优势，随着高等级自动驾驶渗透率逐步提升，自动驾驶GPU市场也进入高速成长阶段。

- **英伟达凭借其数据中心GPU的核心技术优势，成为全球人工智能芯片的领导者。**英伟达过去专注于GPU芯片设计，目前已经转型成为计算平台企业，成为人工智能芯片的领导者。其主营业务包含游戏&娱乐、数据中心、专业可视化、汽车业务。过去的两个季度中，伴随着个人电脑GPU整体市场需求疲软，英伟达游戏&娱乐业务营收大幅下滑。随着虚拟货币挖矿退潮对GPU独立显卡带来的冲击逐渐下降，公司22Q4游戏&

核心观点

娱乐业务再次环比提升，我们认为公司个人电脑GPU业务正逐步恢复到正常成长阶段。2022年英伟达数据中心业务营收超过游戏&娱乐业务，成为第一大收入来源，公司GPGPU具备核心技术优势，在AI芯片市场中占据主导地位，其数据中心业务将为公司的高质量成长贡献长期动力。在自动驾驶业务方面，英伟达提供全栈式的自动驾驶解决方案，硬件层面上，其Orin和Thor自动驾驶芯片提供大幅算力，同时DLA模块和PVA模块实现AI算法加速；在软件层面上提供完整的开发者套件，其自动驾驶业务的平台化优势保证了英伟达在高等级自动驾驶中的领先地位。

- **AMD作为全球领先的芯片设计厂商，在GPU市场中与英伟达互相角逐。**AMD的数据中心业务和嵌入式业务展现良好的增长趋势，公司同时提供个人电脑GPU和数据中心GPU。公司的集成GPU主要被运用在台式机和笔记本的APU产品，相比独立GPU更具性价比优势。Radeon系列独立GPU构建于RDNA 3架构之上，采用5nm工艺和chiplet设计，实现了性能的整体提升。AMD推出用于数据中心的Radeon Instinct GPU加速芯片，Instinct系列基于CDNA架构。最新的CDNA 2架构实现计算能力和互联能力的显著提升，采用CDNA 2架构的计算芯片MI250X与英伟达的先进计算芯片性能指标不分伯仲。AMD ROCm对标英伟达CUDA，其计算生态也在不断的丰富过程当中。

- **移动端的主要玩家包括高通、ARM、Imagination。**移动端GPU在设计过程中受到能耗和体积方面的限制，都是以集成的SOC芯片的形式出现在移动端。高通在旗舰Android智能手机SoC市场中保持领先地位，ARM是领先的GPU IP公司，Imagination的PowerVR架构在移动芯片领域得到市场的广泛认可，随后陆续提出PowerVR的升级版本IMG系列架构。

- **国内GPU市场空间广阔，涌现出一批优秀的GPU设计和制造厂商。**根据Verified Market Research数据，2020年中国大陆GPU市场规模为47.39亿美元，预计2023年中国GPU市场规模将达到111亿美元。伴随着近期宏观经济回暖以及国内互联网企业纷纷加大AI算力布局，PC和服务器的需求上升有望为国内GPU市场带来整体拉动效应。国内涌现出一批优秀的GPU设计和制造厂商，诸如海光信息和景嘉微。海光信息DCU的产品性能均达到了国际上同类型主流高端处理器水平，在国内处于领先地位，同时海光信息DCU协处理器全面兼容ROCm GPU计算生态。景嘉微GPU研发历史悠久，技术积淀深厚，其GPU性能优越，芯片业务整体展现良好增长势头。

- **风险提示：个人电脑出货不及预期、AI技术进展不及预期、互联网厂商资本开支不及预期、自动驾驶进展不及预期、国产替代进程不及预期、参与厂商众多导致竞争格局恶化。**

目 录

第一章

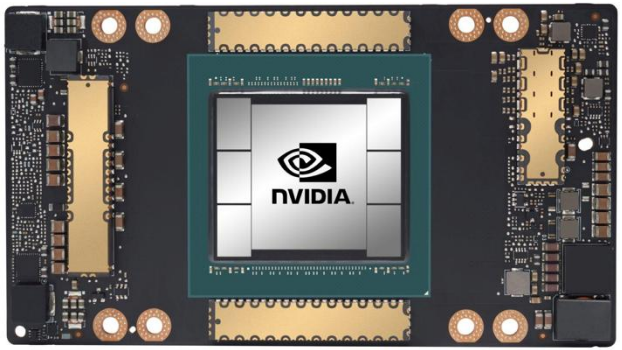
GPU芯片概述 05

中信建投证券 CHINA SECURITIES

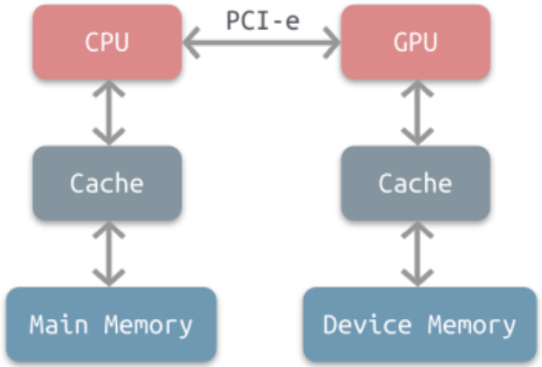
1.1 GPU定义和主要组成

- **GPU（Graphics Processing Unit）：一般称为图形处理器，被广泛用于个人电脑、工作站、移动设备、游戏机、嵌入式系统中做图像和图形相关运算工作。**
- **GPU结构：**GPU是一个异构的多核处理器芯片，针对图形图像处理优化。通常包括运算单元、L0/L1/L2缓存、Warp调度器、存取单元、分配单元、寄存器堆、PCIe总线接口、显卡互联单元等组件。
- **GPU工作方式：**GPU并不是一个独立运行的计算平台，需要通过PCIe总线与CPU连接在一起来协同工作，可以看作CPU的协处理器。

图：英伟达A100 GPU



图： CPU-GPU异构架构

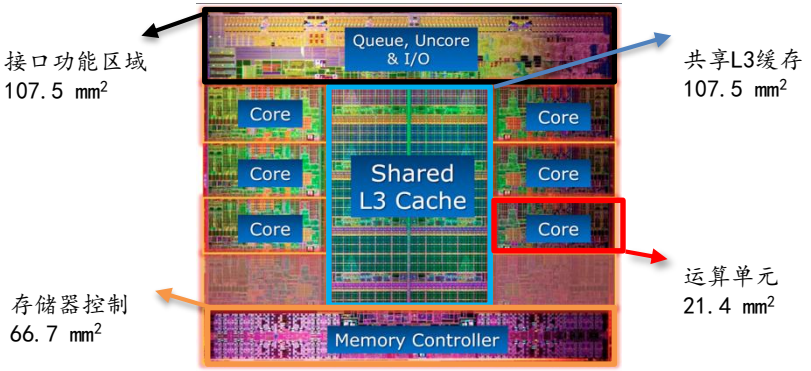


资料来源：英伟达，CSDN，中信建投

1.2 GPU相较于CPU并行计算能力更强（一）

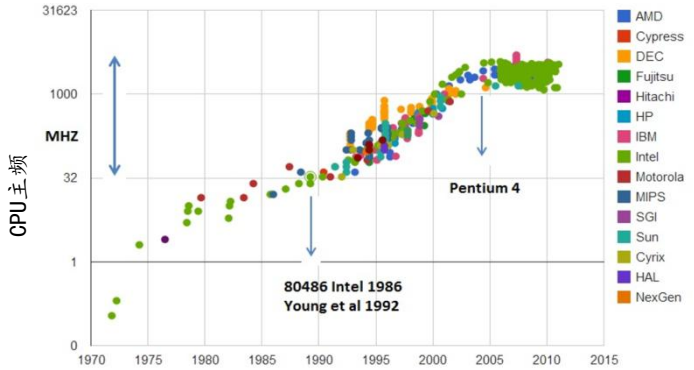
- **CPU当中运算单元占据面积相对较小。**CPU硬件设计过程中为了实现低延迟增加了存储单元和控制单元的复杂度，运算单元在GPU中占据面积相对较小，以Intel Core i7 3960X为例，其运算单元面积（6*21.4 mm²）大致占总芯片面积（435 mm²）的30%。
- **CPU的并行计算能力相对较弱。**CPU通过指令级并行、数据级并行也可以提升其并行计算能力，但是带来的提升也是有限的。
- **单核CPU性能逐步逼近物理极限。**由于CPU受到“能耗墙”的限制，CPU主频难以持续提升，**单核CPU性能逐步逼近物理极限**，采用多核CPU的策略一定程度缓解了CPU性能提升的制约，当前大数据和人工智能带来了海量的数据，CPU已经无法跟上多源异构数据的爆炸性增长。

图：CPU的功能区域分布



Intel Core i7 3960X　22.7亿晶体管 面积435mm²

图：CPU的主频受到“能耗墙”限制

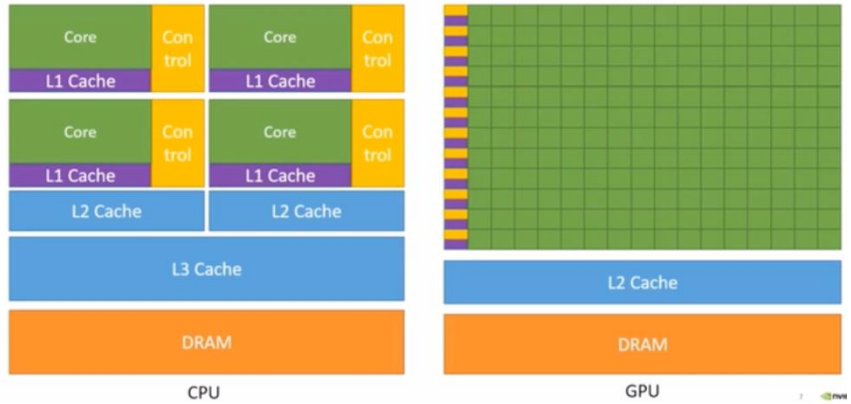


资料来源：Intel，Wikipedia，中信建投

1.2 GPU相较于CPU并行计算能力更强（二）

- **GPU具有数量众多的运算单元，采用极简的流水线进行设计，适合计算密集、易于并行的程序。CPU的运算单元数目相对较少，单一运算核心的运算能力更强，采用分支预测、寄存器重命名、乱序执行等复杂的处理器设计，适合相对复杂的串行运算。**
- **GPU设计过程中侧重吞吐优化，具备强大的内存访问带宽。CPU设计过程中侧重时延优化，包含复杂的多级缓存（L1/L2/L3）和逻辑控制单元。**
- CPU承担运算核心和控制中心的地位，GPU一般作为协处理器负责图形渲染和并行计算。

图：CPU和GPU的架构比对



图表： GPU和CPU的区别

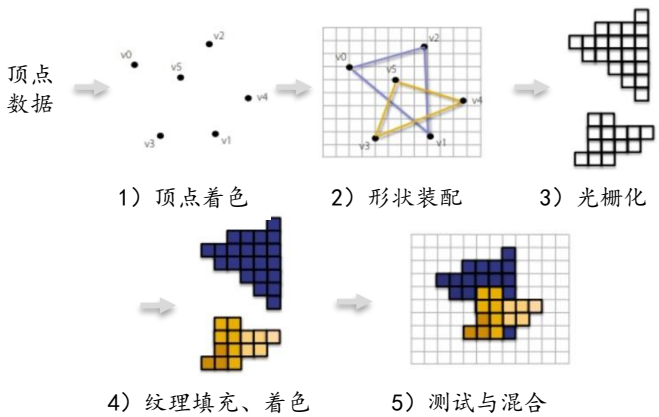
维度	GPU	CPU
核心数量	数千个加速核心	几十个核心
产品特点	简单的逻辑控制	复杂的逻辑控制单元
	多线程以到达超大并行吞吐量	通过多级缓存降低延迟
	高效众多的运算单元（ALU）	少量强大的运算单元（ALU）
适用场景	计算密集、易于并行的程序	逻辑控制、串行运算的程序

资料来源：英伟达，中信建投

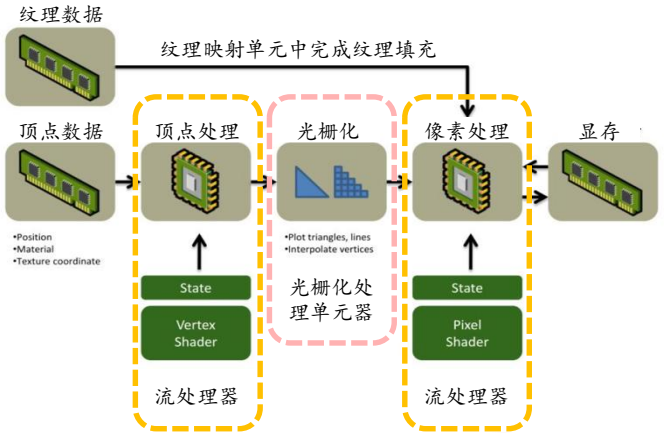
1.3 GPU的核心功能一：图形渲染

- GPU凭借其较强的并行计算能力，已经成为个人电脑中图像渲染的专用处理器。
- 图形渲染具体实现要通过五阶段：顶点着色、形状装配、光栅化、纹理填充着色、测试与混合。
- **GPU渲染流程：**三维图像信息输入GPU后，读取3D图形外观的顶点数据后，1）在**流处理器**中构建3D图形的的整体骨架，即顶点处理；2）由**光栅化处理单元**把矢量图形转化为一系列像素点，即光栅化操作；3）在**纹理映射单元**实现纹理填充；4）在**流处理器**中完成对像素的计算和处理，即着色处理；5）在**光栅化处理单元**中实现测试与混合任务。至此，实现一个完整的GPU渲染流程。

图：渲染流程操作



图：GPU硬件架构下的渲染流程

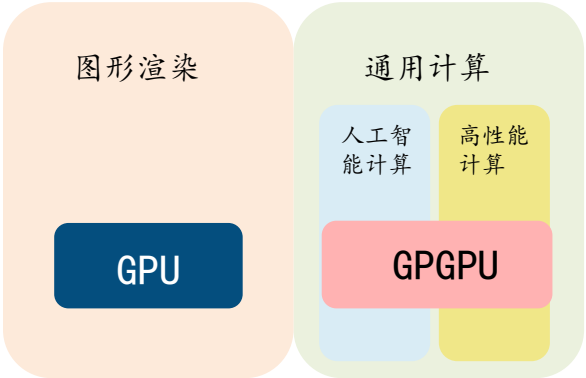


资料来源：OpenGenus，Vmware，中信建投

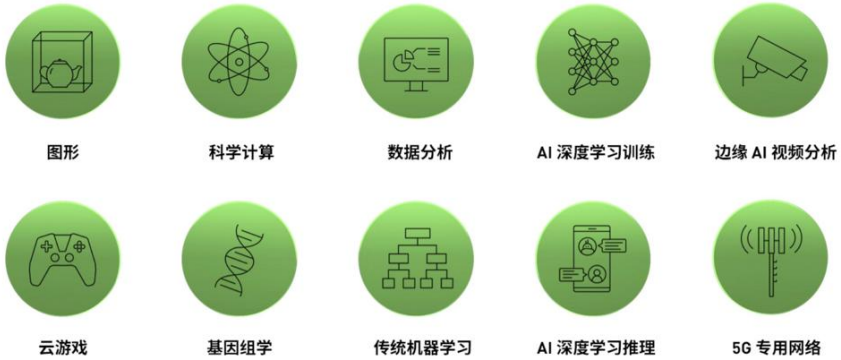
1.4 GPU的核心功能二：通用计算

- 2003年，**GPGPU(General Purpose computing on GPU，基于GPU的通用计算)**的概念首次被提出，意指利用GPU的计算能力在非图形处理领域进行更通用、更广泛的科学计算。GPGPU概念的提出，为GPU更为广泛的应用开拓了思路，GPGPU在传统GPU的基础上进行了优化设计，部分GPGPU会去除GPU中负责图形处理加速的硬件组成，使之更适合高性能并行计算。
- **GPGPU在数据中心被广泛地应用在人工智能和高性能计算、数据分析等领域。**GPGPU的并行处理结构非常适合人工智能计算，人工智能计算精度需求往往不高，INT8、FP16、FP32往往可以满足大部分人工智能计算。GPGPU同时可以提供FP64的高精度计算，使得GPGPU适合信号处理、三维医学成像、雷达成像等高性能计算场景。

图：GPU与GPGPU的差异



图：GPGPU在数据中心中承担的工作任务



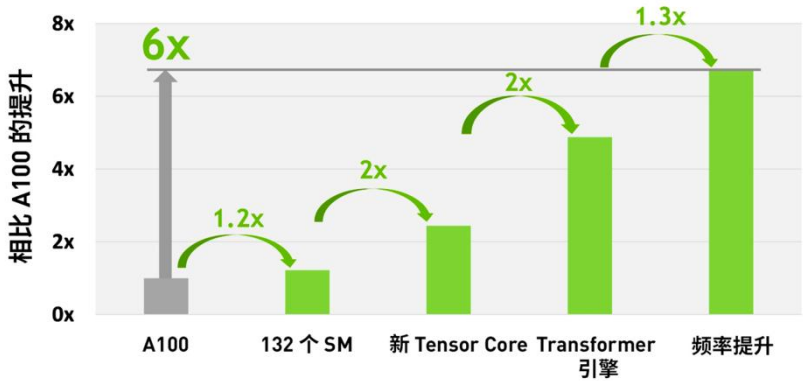
资料来源：CSDN，英伟达官网，中信建投

1.5 微架构设计是GPU性能提升的关键所在

- **GPU微架构（Micro Architecture）是兼容特定指令集的物理电路构成，由流处理器、纹理映射单元、光栅化处理单元、光线追踪核心、张量核心、缓存等部件共同组成。**图形渲染过程中的图形函数主要用于绘制各种图形及像素、实现光影处理、3D坐标变换等过程，期间涉及大量同类型数据（如图像矩阵）的密集、独立的数值计算，而GPU结构中众多重复的计算单元就是为适应于此类特点的数据运算而设计的。
- **微架构的设计对GPU性能的提升发挥着至关重要的作用，也是GPU研发过程中最关键的技术壁垒。**微架构设计影响到芯片的最高频率、一定频率下的运算能力、一定工艺下的能耗水平，是芯片设计的灵魂所在。英伟达H100相比于A100，1.2倍的性能提升来自于核心数目的提升，5.2倍的性能提升来自于微架构的设计。

图：英伟达Ada AD102 GPU架构

图：英伟达H100相比于A100的性能提升

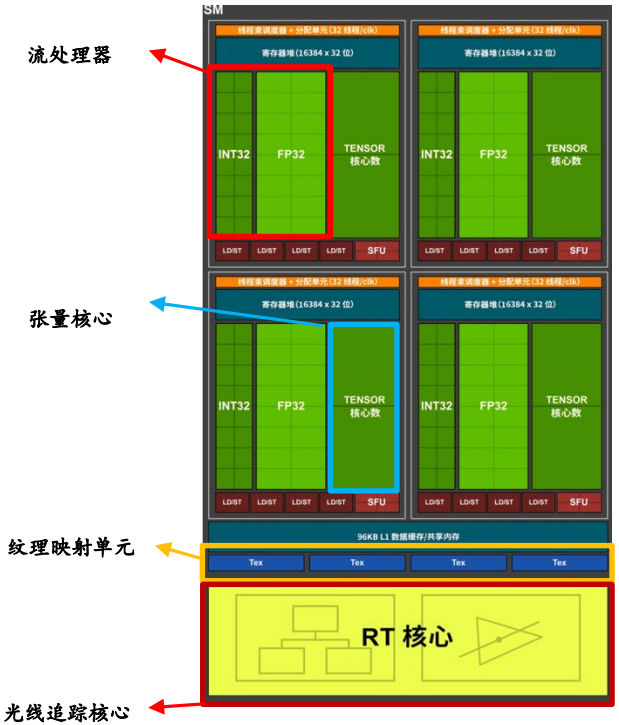


资料来源：英伟达官网，中信建投

1.6 GPU微架构的硬件构成（一）

- **流处理器（Stream Processor）：**是GPU内基本运算单元，通常由整点运算部分和浮点运算部分共同组成，称为SP单元，从编程角度出发，也将其称为CUDA核心。流处理器是DirectX10后引入的一种统一渲染架构，综合了顶点处理（Vertex Pipelines）和像素处理（Pixel Pipelines）的渲染任务，流处理器的数量和显卡性能密切相关。
- **纹理映射单元（Texture Mapping Unit，TMU）：**作为GPU中的独立部件，能够旋转、调整和扭曲位图图像（执行纹理采样），将纹理信息填充在给定3D模型上。
- **光栅化处理单元（Raster Operations Units，ROPs）：**依照透视关系，将整个可视空间从三维立体形态压到二维平面内。流处理器和纹理映射单元分别把渲染好的像素信息和剪裁好的纹理材质递交给处于GPU后端的光栅化处理单元，将二者混合填充为最终画面输出，此外游戏中雾化、景深、动态模糊和抗锯齿等后处理特效也是由光栅化处理单元完成的。

图：英伟达Turing的微架构单元

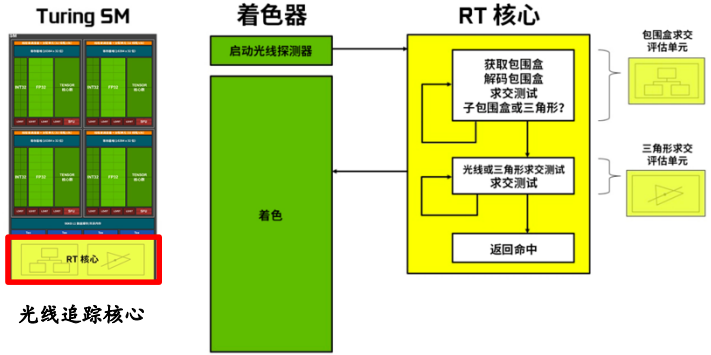


资料来源：英伟达官网，中信建投

1.6 GPU微架构的硬件构成（二）

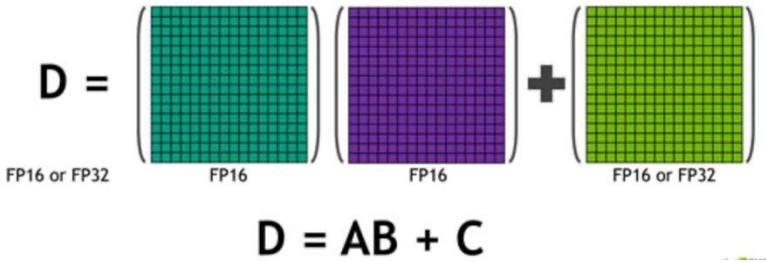
- **光线追踪核心（RT Core）：**是一种补充性的渲染技术，主要通过计算光和渲染物体之间的反应得到正确的反射、折射、阴影即全局照明等结果，渲染出逼真的模拟场景和场景内对象的光照情况。通过采样BVH算法，用来计算射线（光线、声波）与物体三角形求交，与传统硬件相比，RT Core可以实现几何数量级的BVH计算效率提升，让实时光线追踪成为可能。
- **张量核心（Tensor Core）：张量核心可以提升GPU的渲染效果同时增强AI计算能力。**张量核心通过深度学习超级采样（DLSS）提高渲染的清晰度、分辨率和游戏帧速率，同时对渲染画面进行降噪处理以实时清理和校正光线追踪核心渲染的画面，提升整体渲染效果。同时张量核心通过低精度混合运算，极大加速了AI运算速度，让计算机视觉、自然语言处理、语言识别和文字转化、个性化推荐等过去CPU难以实现的功能也得以高速完成。

图：英伟达图灵光线追踪核心



光线追踪核心

图：Tensor Core通过混合精度运算实现AI运算加速

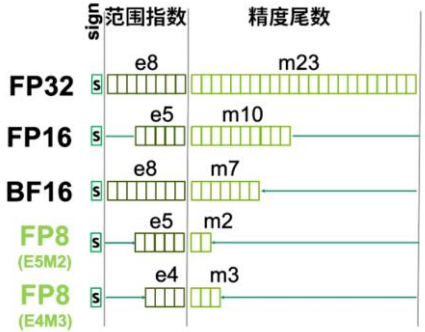


资料来源：英伟达官网，中信建投

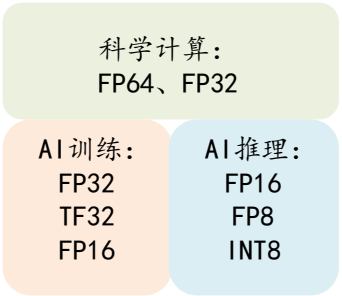
1.7 GPU中常见的数据格式和应用场景

- **计算机中常用的数据格式包括定点表示和浮点表示。**
- **定点表示中小数点位置固定不变，数值范围相对有限，GPU中常用的定点表示有INT8和INT16，多用于深度学习的推理过程。**
- **浮点表示中包括符号位、阶码部分、尾数部分。**符号位决定数值正负，阶码部分决定数值表示范围，尾数部分决定数值表示精度。FP64（双精度）、FP32（单精度）、FP16（半精度）的数值表示范围和表示精度依次下降，运算效率依次提升。
- 除此以外还有TF32、BF16等其他浮点表示，保留了阶码部分但是截断了尾数部分，牺牲数值精度换取较大的数值表示范围，同时获得运算效率的提升，在深度学习中得到广泛应用。

图：不同的浮点表示

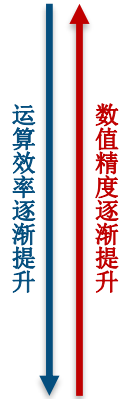


图：不同应用场景的常见数据格式



图表：不同数据格式的构成和用途

数据格式	构成	用途
FP64	1位符号、11位指数、52位尾数	常用于对精度要求高的科学计算
FP32	1位符号、8位指数、23位尾数	深度学习模型训练的常见格式
TF32	1位符号、8位指数、10位尾数	替代FP32数据格式实现深度学习和HPC计算加速
FP16	1位符号、5位指数、10位尾数	深度学习越来越偏向使用FP16
BF16	1位符号、8位指数、7位尾数	提升AI模型的推理速度和部署灵活性
INT8	8个bit表示一个数字	INT8精度相对较低，常用于AI模型的端侧推理

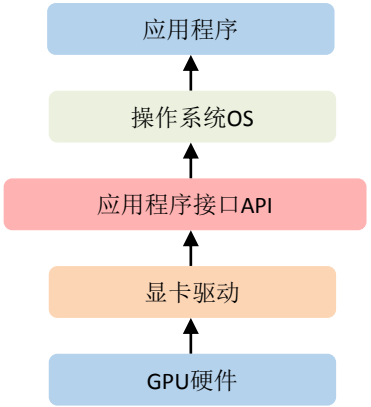


资料来源：英伟达，中信建投

1.8 应用程序接口是GPU和应用软件的连接桥梁

- **GPU应用程序接口（Application Programming Interface，API）：API是连接GPU硬件与应用程序的编程接口，有利于高效执行图形的顶点处理、像素着色等渲染功能。**早期由于缺乏通用接口标准，只能针对特定平台的特定硬件编程，工作量极大。随着API的诞生以及系统优化的深入，GPU的API可以直接统筹管理高级语言、显卡驱动及底层的汇编语言，提高开发过程的效率和灵活性。
- **GPU应用程序接口主要涵盖两大阵营，分别是Microsoft DirectX和Khronos Group技术标准。**DirectX提供一整套多媒体解决方案，3D渲染表现突出，但是只能用于windows系统。OpenGL的硬件匹配范围更广，同时在CAD、游戏开发、虚拟现实等高端绘图领域得到广泛应用。此外还包括苹果的Metal API等。

图：应用程序接口连接GPU硬件与应用程序



图表：GPU应用程序接口主要标准

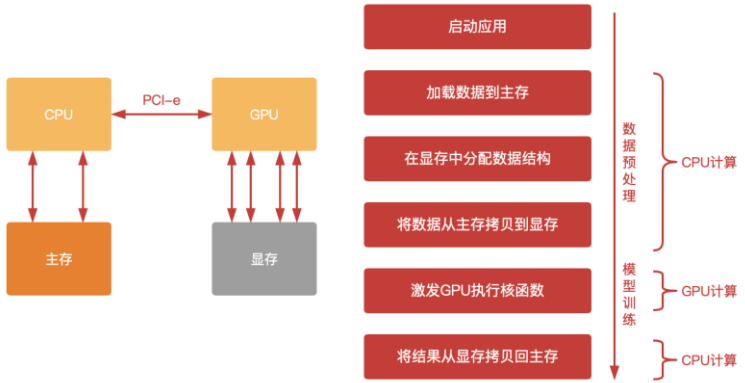
图形API		平台特征
Direct3D		Windows
OpenGL系列	OpenGL	Windows、类Unix、Linux、MacOS
	Vulkan	Windows、Android、Linux
	OpenGL ES	IOS、Android
	WebGL	跨平台
Metal		APPLE

资料来源：CSDN，中信建投

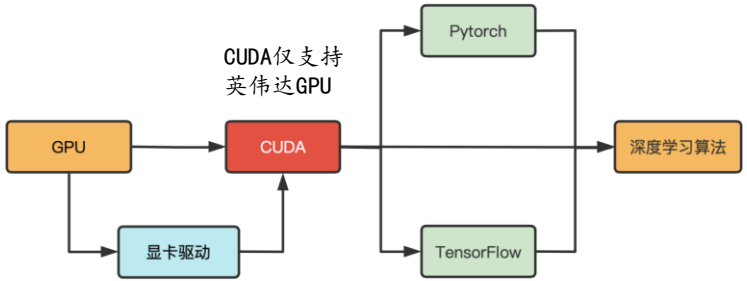
1.9 CUDA架构实现了GPU并行计算的通用化（一）

- GPGPU相比于CPU，其并行计算能力更强，但是通用灵活性相对较差，编程难度相对较高。在CUDA出现之前，需要将并行计算映射到图形API中从而在GPU中完成计算。
- **CUDA大幅降低GPGPU并行计算的编程难度，实现GPU的通用化。**CUDA是英伟达2007年推出的适用于并行计算的**统一计算设备架构**，该架构可以利用GPU来解决商业、工业以及科学方面的复杂计算问题。CUDA架构的里程碑意义在于，GPU的功能不止局限于图形渲染，实现了GPU并行计算的通用化，把“个人计算机”变成可以并行运算的“超级计算机”。英伟达在推出了CUDA以后，相当于把复杂的显卡编程包装成了一个简单的接口，可以利用CUDA直观地编写GPU核心程序，使得编程效率大幅提升。现在主流的深度学习框架基本都是基于CUDA加速GPU并行计算。

图：GPU中并行计算过程



图：CUDA连接GPU与深度学习框架



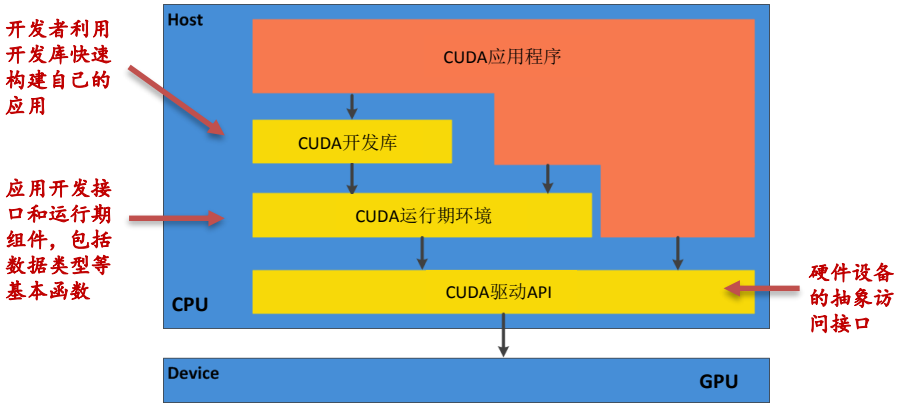
CUDA兼容各类深度学习框架

资料来源：CSDN，中信建投

1.9 CUDA架构实现了GPU并行计算的通用化（二）

- **CUDA：**CUDA采用了一种全新的计算体系结构来调动GPU提供的硬件资源，本质上是应用程序和GPU硬件资源之间的接口。CUDA程序组成包括CUDA库、应用程序编程接口（API）及运行库(Runtime)、高级别的通用数学库。
- **CUDA经过多年优化，形成了独特软硬件配合的生态系统。**其中包括诸多编程语言的开发环境，各种API的第三方工具链，自带的应用于代数运算和图形处理的CUDA库、庞大的应用程序库，从而实现轻松高效的编写、调试优化过程。
- CUDA提供了对其它编程语言的支持，如C/C++，Python，Fortran等语言。
- CUDA支持Windows、Linux、Mac各类操作系统。

图：CUDA程序结构



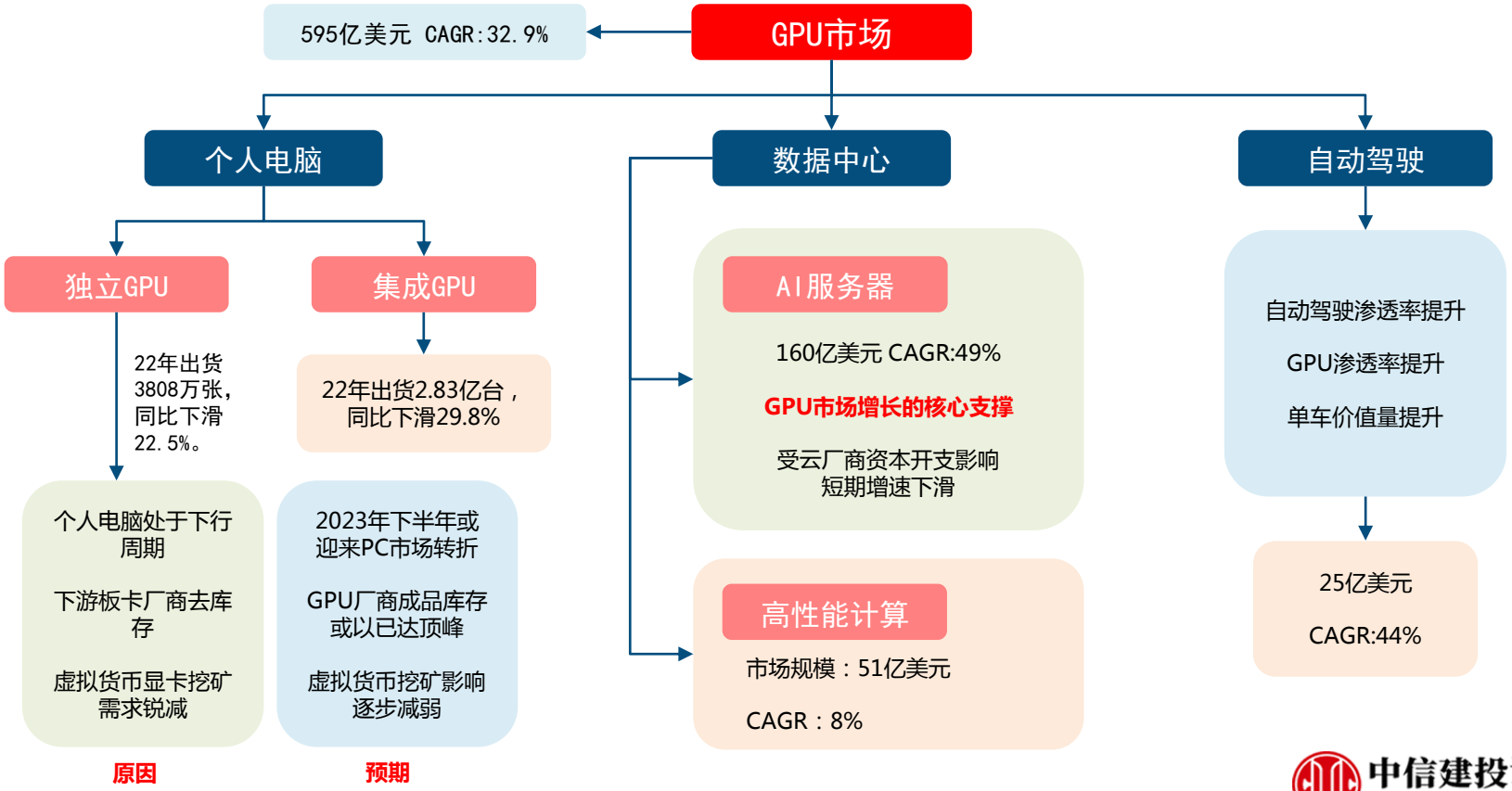
图表：CUDA部分内置函数

函数	功能
cuFFT	利用CUDA进行快速傅里叶变换的函数库
cuBLAS	线性代数方面的CUDA库
cuDNN	利用CUDA进行深度卷积神经网络计算
Thrust	实现众多并行算法的C++模板库
cuSolver	线性代数方面的CUDA库。
cuRAND	随机数生成有关的库

资料来源：CSDN，中信建投

第二章 GPU市场概述 18

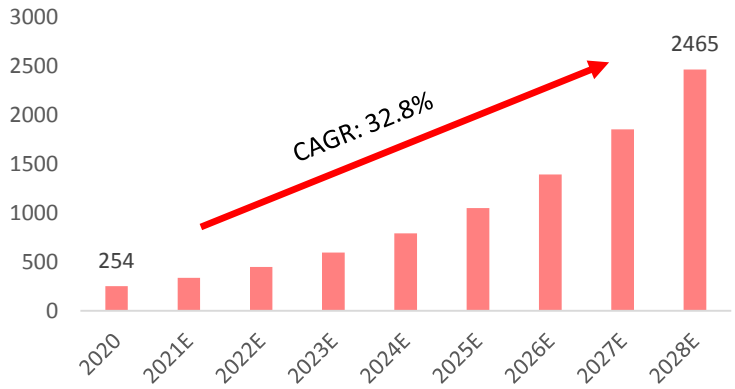
GPU市场空间测算



2.1 GPU市场规模与细分

- 根据Verified Market Research的预测，2020年GPU全球市场规模为254亿美金，预计到2028年将达到2465亿美金，行业保持高速增长，CAGR为32.9%，2023年GPU全球市场规模预计为595亿美元。
- GPU按应用端划分为PC GPU、服务器GPU、智能驾驶GPU、移动端GPU。
- PC GPU可以进一步划分为独立显卡和集成显卡。独立显卡主要用作图形设计和游戏，对性能的要求比较高，主要的厂商包括英伟达和AMD；集成显卡通常用在对图形处理性能需求不高的办公领域，主要产商包括Intel和AMD。
- 服务器GPU通常应用在深度学习、科学计算、视频编解码等多种场景，主要的厂商包括英伟达和AMD，英伟达占主导地位。
- 在自动驾驶领域， GPU通常用于自动驾驶算法的车端AI推理，英伟达占据主导地位。

图：GPU整体市场规模（亿美金）



图表：GPU的构成分类和生产厂商

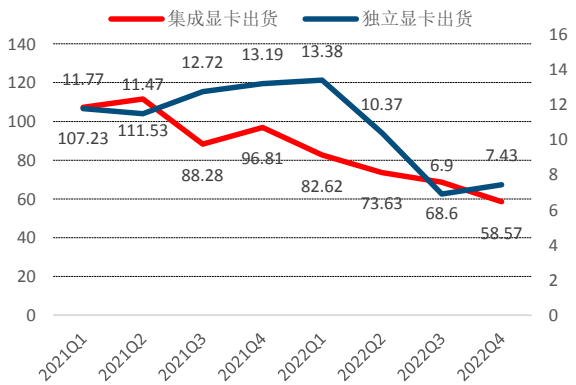
类别		主要领域	主要厂商
PC GPU	独立显卡	图形设计/游戏	NVIDIA、AMD
	集成显卡	办公	Intel、AMD
服务器GPU		AI训练/AI推理/HPC计算	NVIDIA、AMD
智能驾驶GPU		AI推理	NVIDIA
移动端GPU		图形显示	Imagination、高通、ARM

资料来源：Verified Market Research，中信建投

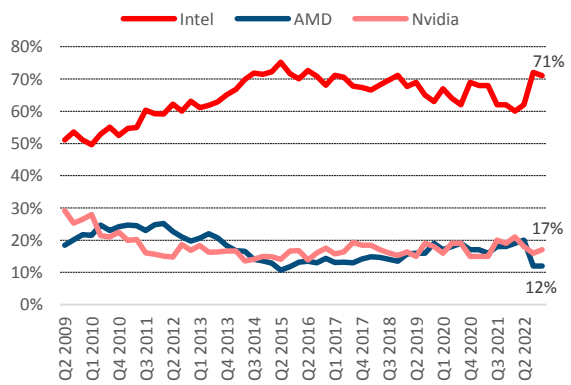
2.2 PC显卡市场迎来至暗时刻后的光明

- **独立显卡市场开始逐渐回暖。**根据Jon Peddie Research的数据，2022年独立GPU出货量下降至3808万台，同比下降22.5%，22Q3单季度出货690万台，同比下降45.7%，是十年以来最大的一次下滑，独立显卡出货情况22Q4开始逐渐转暖。
- **集成显卡出货情况仍然不容乐观。**2022年集成GPU出货量为2.83亿台，同比下滑29.8%。疫情期间的居家办公需求带动了笔记本电脑的消费增长，集成显卡的购买激增一定程度上过早消耗了市场需求，后疫情时代，笔记本电脑端需求减弱叠加供应商的过剩库存导致集成显卡出货不断走低。
- **我们认为2022年独立显卡出货遭遇巨大下滑的原因有三点：一、受宏观经济影响，个人电脑市场处于下行周期；二、部分独立GPU参与虚拟货币挖矿，以太坊合并对独立GPU出货造成巨大冲击；三、下游板卡厂商开启降库存周期。**

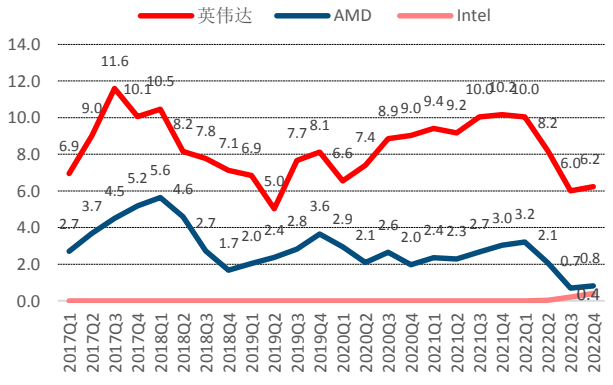
图：PC端不同类型显卡出货量情况(百万台)



图：PC显卡市场市场份额变动（按出货量）



图：独立显卡厂商的出货量情况（百万台）

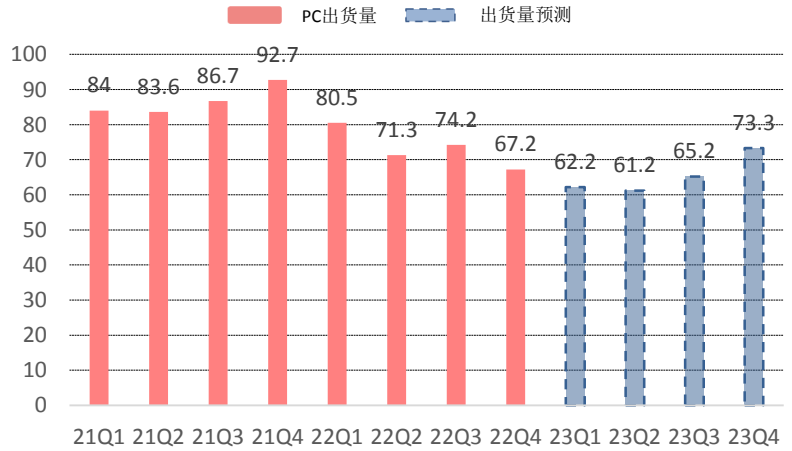


资料来源：Jon Peddie Research，中信建投

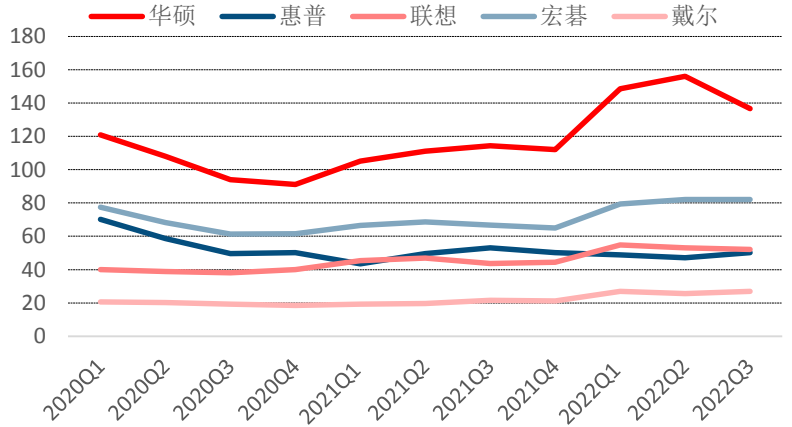
2.3 因素一：个人电脑市场依旧处于下行周期

- **个人电脑市场保持疲软状态。**根据IDC数据，2022年全年PC出货量为2.92亿台，同比下降15.5%，2022Q4全球PC出货量仅为6720万台，同比下降28.1%。IDC预测2023年个人电脑市场全年出货2.608亿台，全年同比下降10.7%。按照2023年的整体出货量情况，我们对四个季度的出货情况做了进一步预测，预计2023Q2-2023Q3后个人电脑出货将迎来逐季度好转。
- **下游PC厂商库存情况得到改善。**当前个人电脑市场正处在PC厂商去库存周期，根据PC厂商的财报披露，华硕和联想的库存天数已经开始减少，其余三家（惠普、戴尔、宏碁）的库存天数并未显著降低，由于所有厂商都在积极采取行动减少产量，预计下游PC厂商库存情况会进一步改善，2023Q3可能恢复到正常库存情况。

图：个人电脑出货情况及预期



图：PC厂商存货周转天数



资料来源：IDC，wind，中信建投
备注：2023年分季度PC出货量为中信建投预测

2.4 因素二：显卡挖矿市场出现转折，以太坊转向权益证明

- **以太坊ETH占据显卡挖矿主要市场。**根据MESSARI数据，在采用GPU挖矿的前7名虚拟货币中，以太坊ETH挖矿收入占GPU矿工总收入的97%。比特币、莱特币等虚拟货币多采用功耗更低的ASIC矿机。
- **2022年9月15日，以太坊运行机制全面升级，从以太坊1.0的工作量证明机制(PoW)转向以太坊2.0的权益证明机制(PoS)，**在工作量证明机制中，需要通过累积显卡提升计算能力，计算能力越强获得记账收益的概率越大；在权益证明机制中，只需通过质押虚拟货币获得收益，质押的虚拟货币数量越大获得记账收益的概率越高。以太坊全面合并后不再需要购入大量显卡、投入计算资源用于挖矿，是显卡挖矿市场的重要转折点。

图：以太坊占据97%的GPU挖矿市场收益

ETH Makes up 97% of Miner Revenue for GPU-Mineable Coins

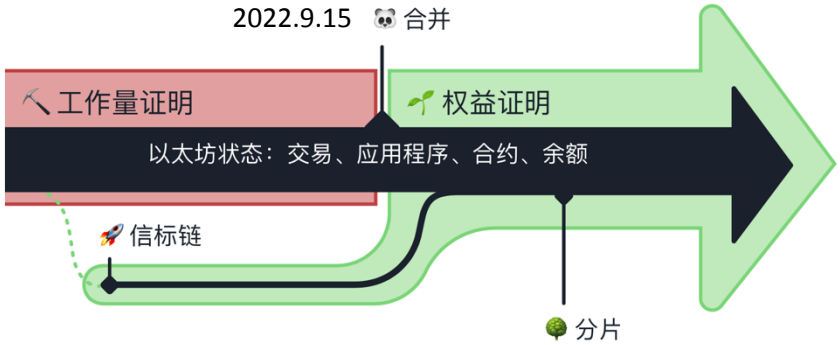
If ETH miners switch over to mining alternative PoW coins, it is likely the coins become unprofitable overnight for a majority of miners due the mining difficulty increasing, unless the coin prices increase several orders of magnitude.

Coin	Current Price	Circulating Market Cap	Block Reward	Daily Blocks	Daily Miner Revenue	% of Total Miner Revenue
ETH	$ 1,787	$ 216,719,304,104	2.01	5760	$ 20,689,171	97%
ETC	$ 21.26	$ 2,872,512,530	3.1	6000	$ 395,436	1.9%
ERG	$ 2.45	$ 78,516,097	66	720	$ 116,424	0.5%
RVN	$ 0.03	$ 284,475,041	2500	1440	$ 97,200	0.5%
CFX	$ 0.08	$ 152,643,787	2	172800	$ 26,957	0.1%
FLUX	$ 0.65	$ 152,947,846	37.5	720	$ 17,550	0.1%
FIRO	$ 1.80	$ 23,949,255	6.25	480	$ 5,400	0.0%

Source: Messari, Bitsbetrippin, Bitpro

MESSARI | Data as of June 9, 2022

图：以太坊由工作量证明机制转向权益证明机制



资料来源：MESSARI，以太坊，中信建投

中信建投证券 CHINA SECURITIES

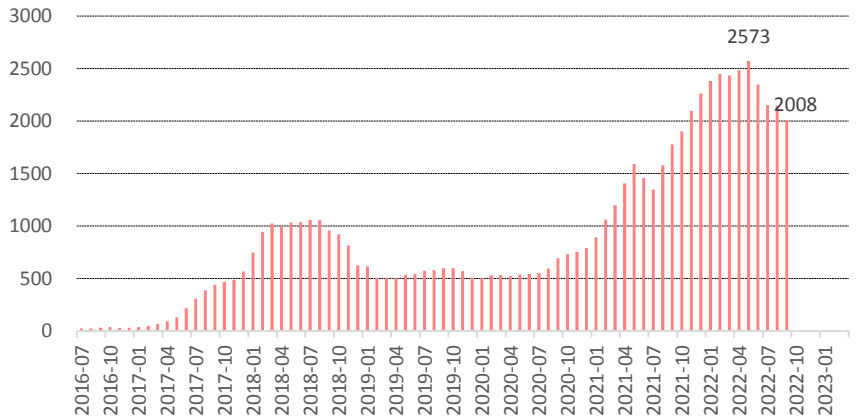
2.5 因素二：以太坊合并预计约500万张二手显卡流入市场

- **挖矿用显卡平均哈希率为46Mh/s。**根据Hive OS矿池数据，通过不同型号显卡的哈希率和占比情况统计，估算得到衡量显卡挖矿能力的平均哈希率为46Mh/s。
- **以太坊合并后显卡需求降至零。**根据以太坊全网算力，测算得到用于以太坊挖矿的GPU数量在2022年5月达到巅峰，大概为2573万张，2022年9月降至2008万张，在以太坊合并之后，显卡需求降至零。
- **如果按照20%回收比例测算，约500万张存量显卡将流入二手市场，预计带来的不利影响在2022Q4-2023Q1之间结束。**

图表：虚拟货币挖矿用显卡统计

型号	哈希率Mh/s	占比
Radeon RX 580	30	10.1%
NVIDIA RTX 3070	62.00	9.4%
NVIDIA GTX 1660 SUPER	32.00	7.9%
NVIDIA RTX 3060 Ti LHR	60.00	5.8%
NVIDIA RTX 2060 SUPER	43.00	4.3%
NVIDIA RTX 3060 Ti	62.00	4.3%
Radeon RX 570	30	3.9%
NVIDIA RTX 3080	100.00	3.6%
Radeon RX 5700 XT	52	3.5%
Radeon RX 6600	30	2.3%
其他	45	44.8%
平均	46	

图：用于以太坊挖矿的显卡数量测算（万张）

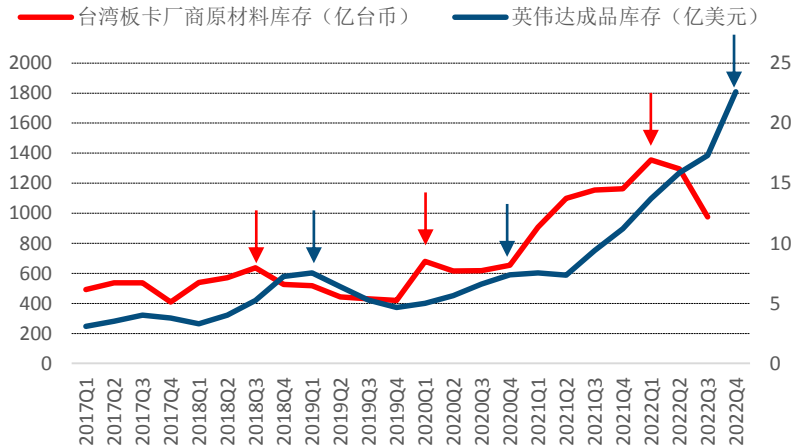


资料来源：Hive OS，etherscan，中信建投

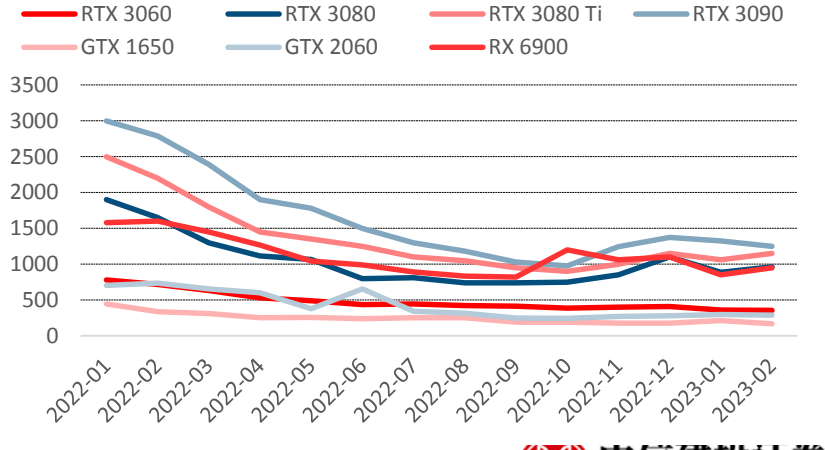
2.6 因素三：GPU厂商库存迎来好转，高端显卡价格企稳回升

- **GPU厂商库存情况即将迎来好转。**根据Bloomberg数据，GPU下游四家台湾板卡厂商（华硕、技嘉、微星、华擎）自2022年一季度原材料库存达到历史高位以后，连续两个季度库存环比降低，当前原材料库存相比最高峰下降28%。复盘历史可见，GPU厂商成本库存高峰多于台湾板卡厂商原材料库存2-3季度后到来，**我们预计GPU厂商的成品库存将于2022Q4到达顶峰。**
- **高端显卡价格开始企稳回升。**根据Amazon上的显卡价格跟踪，英伟达和AMD的高端显卡在2022年10月以后均实现了不同程度的价格回升，例如RTX 3080价格上涨30%，RTX 3090价格上涨28%，显卡价格的回升意味着渠道商库存正逐步回归到正常水平，高端显卡受挖矿市场冲击更为剧烈，高端显卡价格上涨从侧面也能观察到挖矿市场带来的不利影响正在逐渐消失。

图：台湾板卡厂商原材料库存与GPU厂商成品库存情况



图：Amazon显卡价格跟踪（美元）

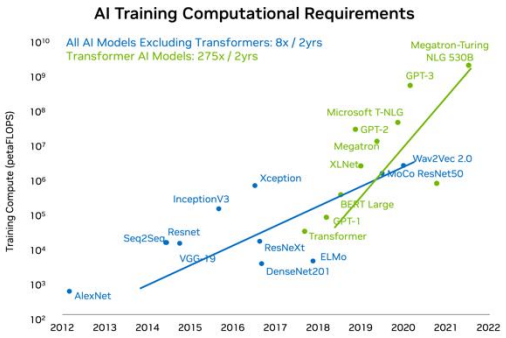


资料来源：Bloomberg，Amazon，中信建投

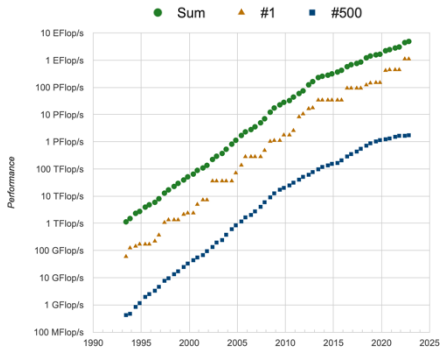
2.7 GPU在数据中心的应用蕴藏巨大潜力

- **在数据中心，GPU被广泛应用于人工智能的训练、推理、高性能计算（HPC）等领域。**
- **预训练大模型带来的算力需求驱动人工智能服务器市场快速增长。**巨量化是人工智能近年来发展的重要趋势，巨量化的核心特点是模型参数多，训练数据量大。Transformer模型的提出开启了预训练大模型的时代，大模型的算力需求提升速度显著高于其他AI模型，为人工智能服务器的市场增长注入了强劲的驱动力。根据Omdia数据，人工智能服务器是服务器行业中增速最快的细分市场，CAGR为49%。
- **战略需求推动GPU在高性能计算领域稳定增长。**高性能计算（HPC）提供了强大的超高浮点计算能力，可满足计算密集型、海量数据处理等业务的计算需求，如科学研究、气象预报、计算模拟、军事研究、生物制药、基因测序等，极大缩短了海量计算所用的时间，高性能计算已成为促进科技创新和经济发展的重要手段。

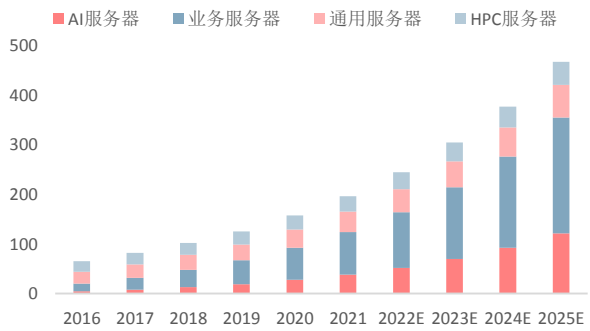
图：大模型时代人工智能算力需求显著提升



图：Top500超级计算机算力总和保持指数级上升



图：中国各类服务器的市场份额（亿）

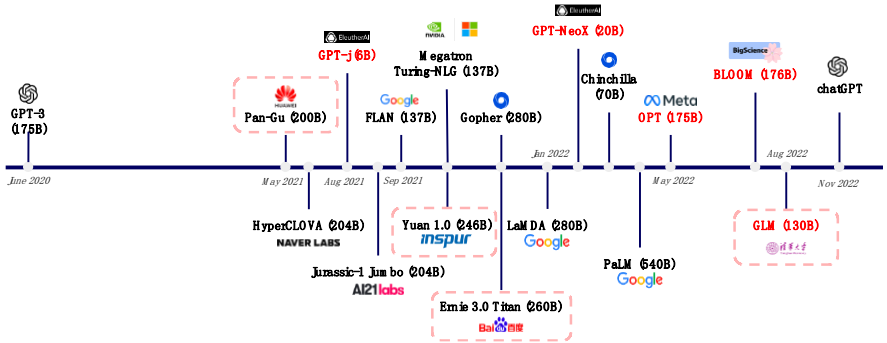


资料来源：英伟达，Top500.org，Frost&Sullivan，中信建投

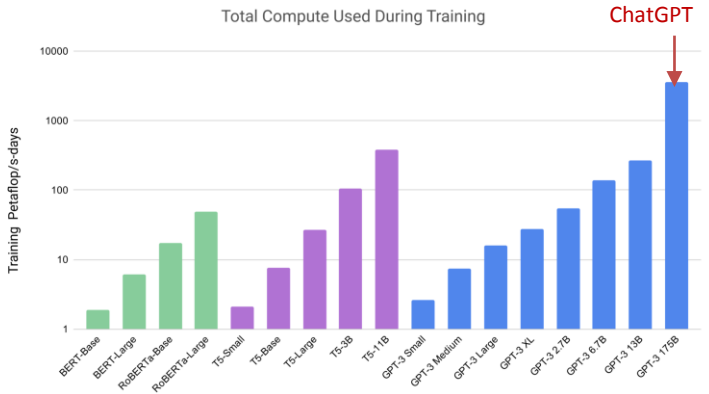
2.8 大模型带来人工智能算力的旺盛需求

- **自然语言大模型参数巨量化是行业发展趋势所向。**以ChatGPT为代表的人工智能模型表现出高度的智能化和拟人化，背后的因素在于自然语言大模型表现出来的涌现能力和泛化能力，模型参数到达千亿量级后，可能呈现性能的跨越式提升，称之为涌现能力；在零样本或者少样品学习情景下，模型仍表现较强的迁移学习能力，称之为泛化能力。两种能力都与模型参数量密切相关，人工智能模型参数巨量化是重要的行业发展趋势。
- **预训练大模型进入千亿参数时代，模型训练算力需求迈上新台阶。**自GPT-3模型之后，大规模的自然语言模型进入了千亿参数时代，2021年之后涌现出诸多千亿规模的自然语言模型，模型的训练算力显著增加。ChatGPT模型参数量为1750亿，训练算力需求为$3.14*10^{23}$ flops，当前各种预训练语言模型还在快速的更新迭代，不断刷新自然语言处理任务的表现记录，单一模型的训练算力需求也不断突破新高。

图：超大规模自然语言模型的发展历程



图：预训练自然语言大模型的算力需求



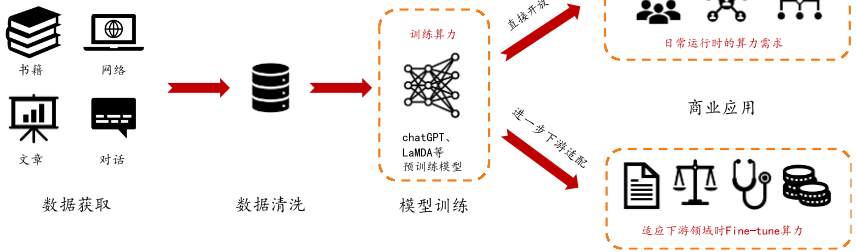
资料来源：stateof AI 2022，Language Models are Few-Shot Learners，中信建投

2.9 大模型带来AI芯片需求的显著拉动

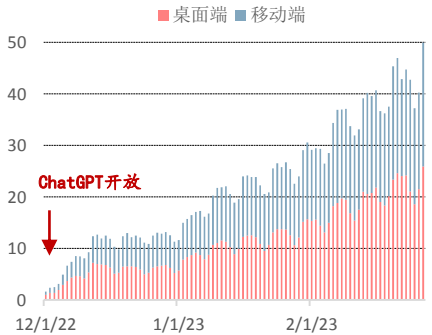
■ **大模型的算力需求主要来自于三个环节：**

- **预训练得到大模型的训练环节。**该环节中，算力呈现海量需求且集中训练的特点，大模型通常在数天到数周内在云端完成训练。模型的训练算力与模型参数量、训练数据量有关，以ChatGPT的训练为例，单次模型训练需要2000张英伟达A100显卡不间断训练27天。
- **适应下游领域时进一步fine-tune环节。**算力需求取决于模型的泛化能力以及下游任务的难度情况。
- **大模型日常运行时的推理环节。**大模型的日常运行中每一次用户调用都需要一定的算力和带宽作为支撑，单次推理的计算量为2N（N为模型参数量），例如1750亿参数的ChatGPT模型1k tokens的推理运算量为$2*1750*10^8*10^3=3.5*10^{14}$ flops=350 Tflops。近期ChatGPT官网吸引的每日访客数量接近5000万，每小时平均访问人数约210万人，假定高峰时期同时在线人数450万人，一小时内每人问8个问题，每个问题回答200字，测算需要14000块英伟达A100芯片做日常的算力支撑。大模型在融入搜索引擎或以app形式提供其他商业化服务过程中，其AI芯片需求将得到进一步的显著拉动。

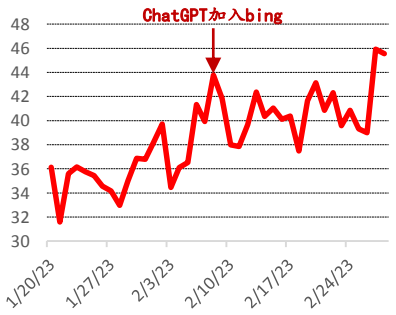
图：大模型的算力需求



图：OpenAI官网每日访问量（百万人）



图：bing搜索每日访问量（百万人）

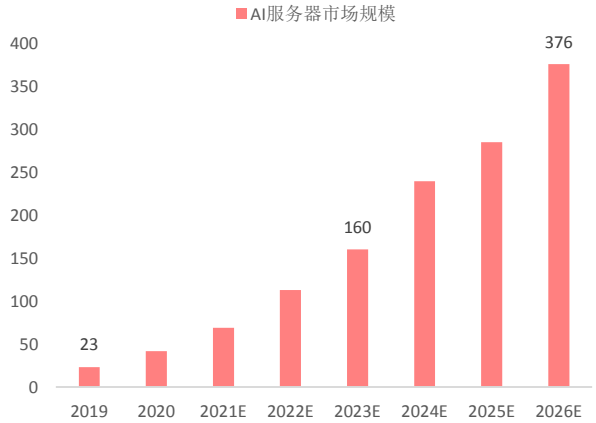


资料来源：similarweb，中信建投

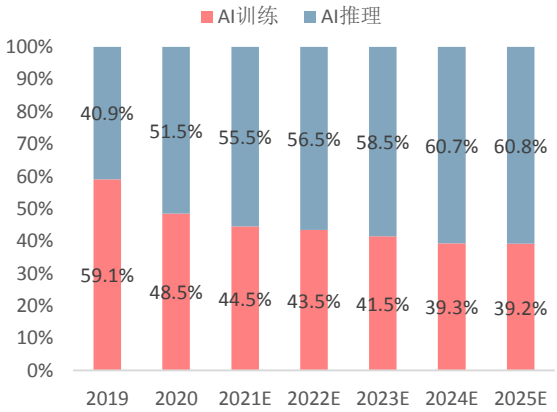
2.10 AI服务器是GPU市场规模增长的重要支撑

- **根据Omdia数据，2019年全球人工智能服务器市场规模为23亿美金，2026年将达到376亿美金，CAGR为49%。**根据IDC数据，2020年中国数据中心用于AI推理的芯片的市场份额已经超过50%，预计到2025年，用于AI推理的工作负载的芯片将达到60.8%。
- 人工智能服务器通常选用CPU与加速芯片组合来满足高算力要求，常用的加速芯片有GPU、现场可编程门阵列（FPGA）、专用集成电路（ASIC）、神经拟态芯片（NPU）等。**GPU凭借其强大的并行运算能力、深度学习能力、极强的通用性和成熟的软件生态，成为数据中心加速的首选，90%左右的AI服务器采用GPU作为加速芯片。**

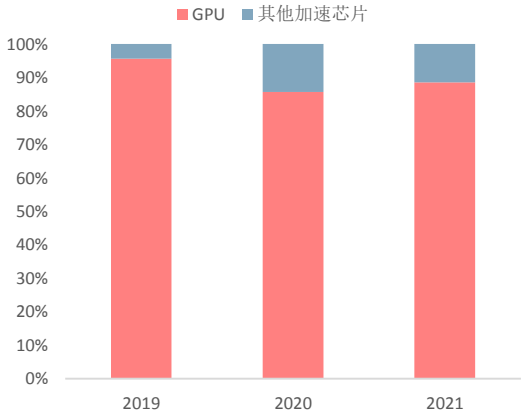
图：全球人工智能芯片市场规模（亿美金）



图：人工智能服务器工作负载预测



图：人工智能服务器加速芯片类型

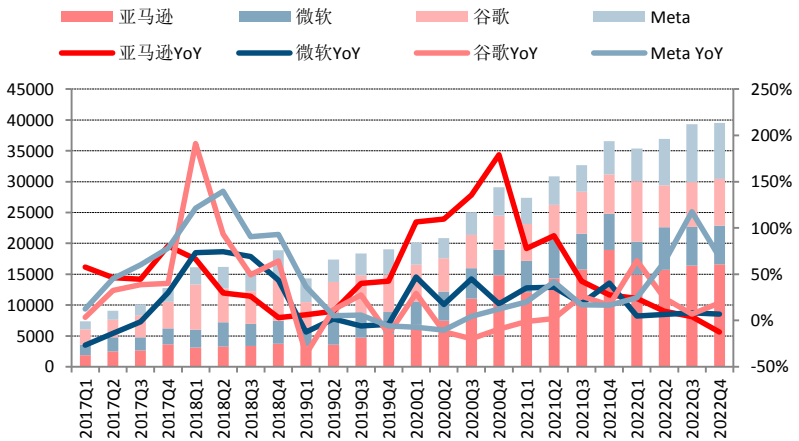


资料来源：Omdia，IDC，中信建投

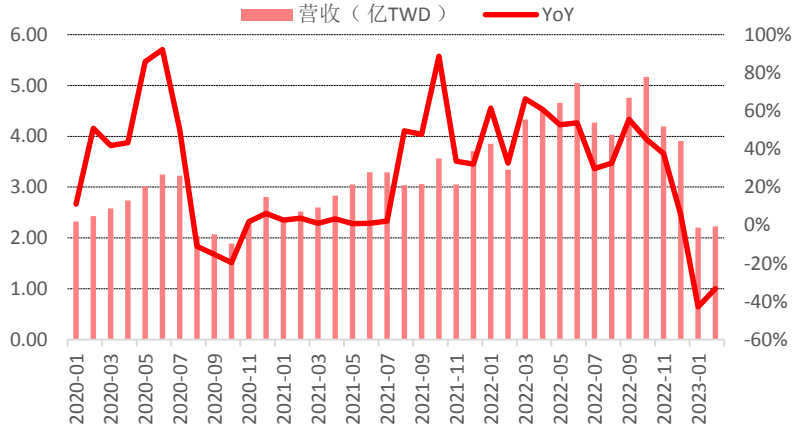
2.11 受云厂商资本开支影响AI服务器市场或将短期增速放缓

- **北美云厂商资本开支有所放缓。**人工智能服务器多采取公有云、私有云加本地部署的混合架构，我们以北美四家云厂商资本开支情况来跟踪人工智能服务器市场需求变动，2022年四家云厂商资本开支合计1511亿美元，同比增长18.5%。Meta预计2023年资本开支的指引为300-330亿美元之前，与2022年基本持平，低于此前22Q3预计的340亿到390亿美元；谷歌预计2023年资本开支将于2022年基本持平，但是会加大AI及云服务的建设投资。
- **信骅科技短期营收下滑有所缓解。**作为全球最大的BMC芯片企业，信骅科技（Aspeed）的营收变化情况一般领先云厂商资本开支一个季度，其月度营收数据可以作为云厂商资本开支的前瞻指标，信骅科技近期营收下滑有所缓解。

图：北美四家云厂商资本开支情况（百万美元）



图：ASPEED营收及增速情况

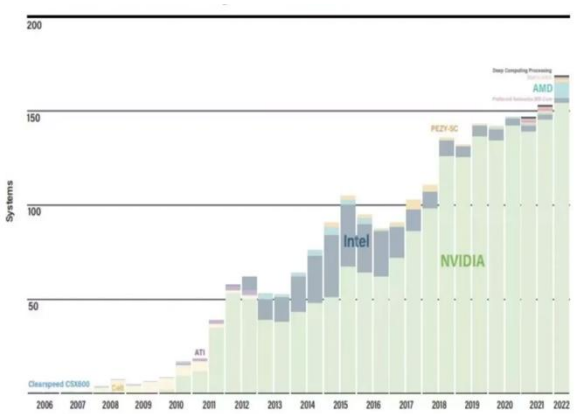


资料来源：Bloomberg，ASPEED，中信建投

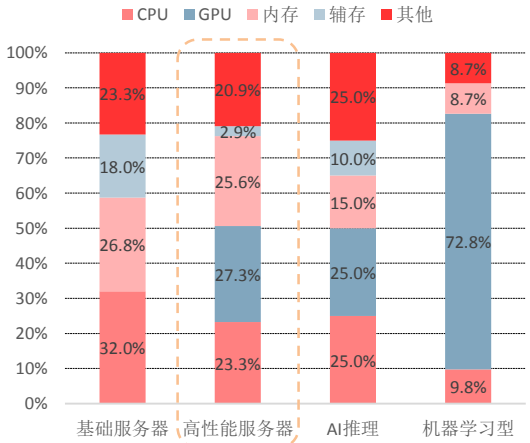
2.12 GPU在超算服务器中的市场规模保持稳定增长

- **GPGPU在高性能计算领域渗透率不断提升。**在高性能计算领域，CPU+GPU异构协同计算架构得到越来越多的应用，全球算力前500的超级计算机中，有170套系统采用了异构协同计算架构，其中超过90%以上的加速芯片选择了英伟达的GPGPU芯片。
- **GPU在超算服务器中的市场规模保持稳定增长。**根据Hyperion Research数据，全球超算服务器的市场规模将从2020年的135亿美金上升到2025年的199亿美金，按照GPU在超算服务器中成本占比为27.3%核算，GPU在超算服务器中的市场规模将从2020年的37亿上升至2025年的54亿美金，CAGR为8%。

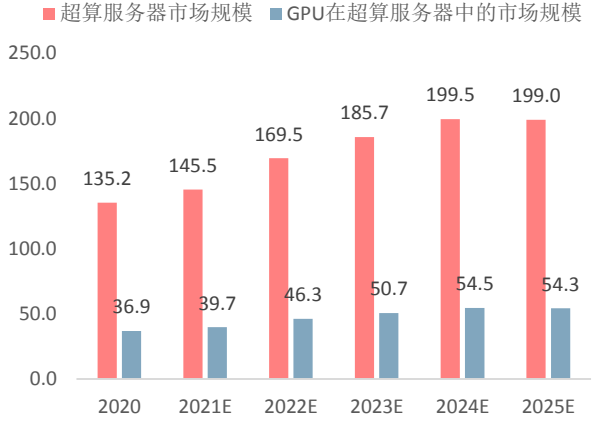
图：Top500超算服务器中加速芯片使用情况



图：不同类型服务器的成本占比



图：GPU在超算中的市场规模（亿美元）

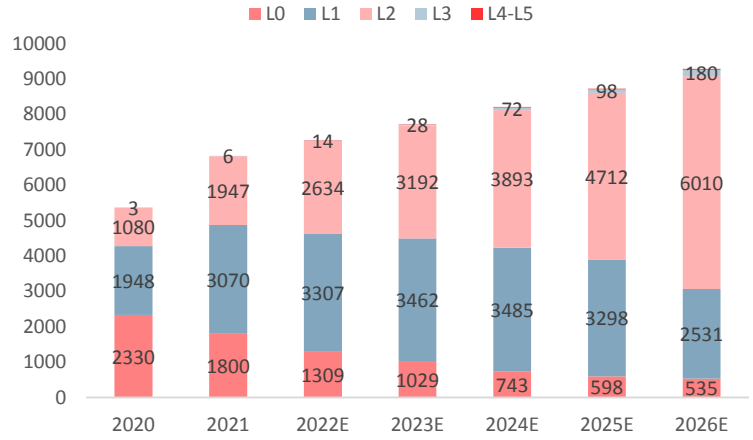


资料来源：top500.org，IDC，中信建投

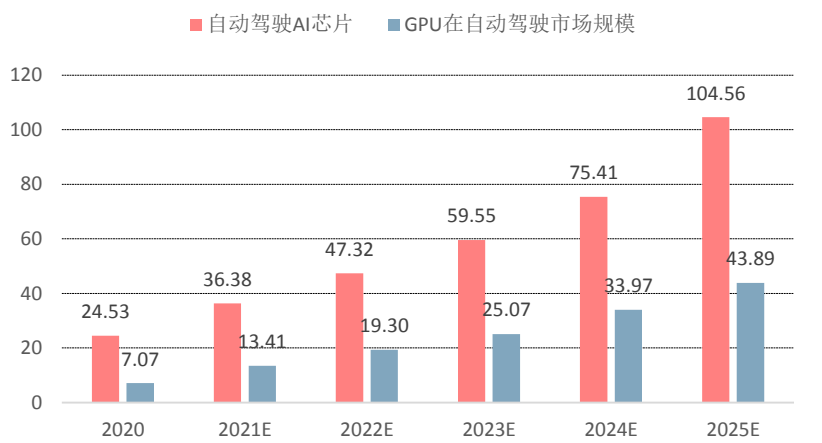
2.13 自动驾驶领域GPU市场保持高成长性

- **在自动驾驶领域，各类自动驾驶芯片得到广泛的应用。**根据Yole数据，全球自动驾驶市场2025年将达到780亿美金，其中用于自动驾驶的AI芯片超过100亿美元。
- **自动驾驶GPU市场保持较高高成长性。**我们根据ICV Tank的自动驾驶渗透数据，假设GPU在L2中渗透率15%，在L3-L5中渗透率50%，估算得到GPU在自动驾驶领域的市场规模，整体规模将从2020年的7.1亿美元上升至2025年的44亿美金，CAGR为44%。

图：自动驾驶渗透率逐步提升



图：GPU在自动驾驶领域的市场规模（亿美元）

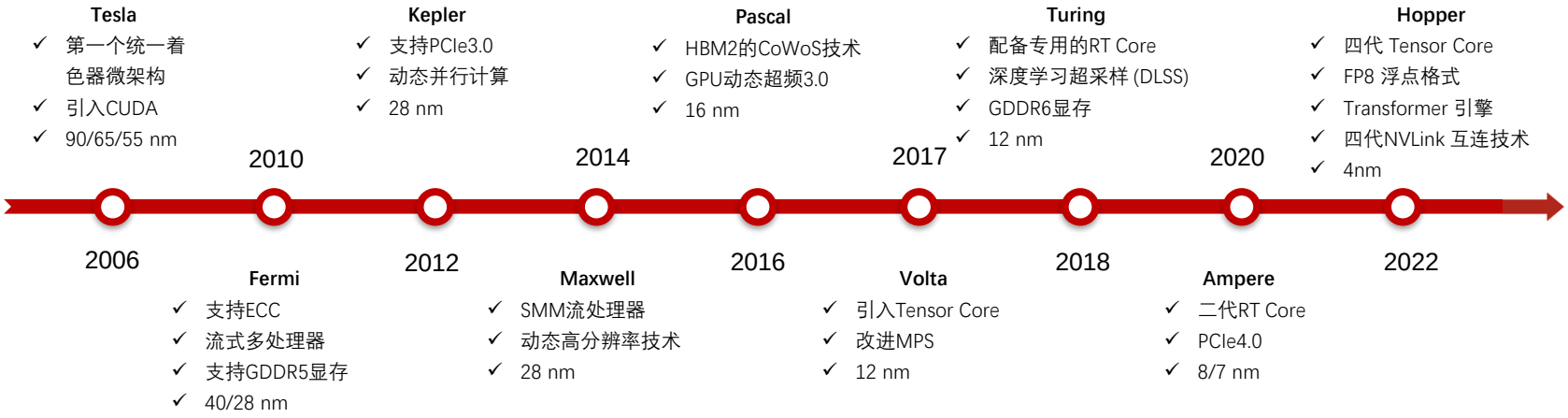


资料来源：ICV Tank，Yole，中信建投

3.1 GPU领域龙头英伟达发展史

- **英伟达（NVIDIA）创立于1993年，是一家专注于智能芯片设计和图形处理技术的半导体公司。**公司产品应用领域包括游戏、数据中心、专业可视化、自动驾驶等，针对具体场景特点，英伟达推出了一系列特定优化的芯片和服务器，同时积极打造相应的软件生态，成为GPU领域的龙头企业。公司当前不仅满足于芯片设计厂商的定位，在芯片、服务器等硬件设施之上，开发CUDA、DOCA等基础软件架构，不断丰富其软件生态，形成了软件业务的全栈式解决方案，最终在应用层面上提供AI计算、高性能计算、自动驾驶、云游戏、元宇宙等众多计算服务，**公司已从一家GPU公司成功转型计算平台企业。**

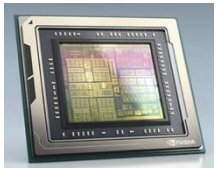
图：英伟达GPU微架构演进历程



资料来源：CSDN，中信建投

3.2 英伟达四大业务下的主要产品体系

图：英伟达主要产品体系

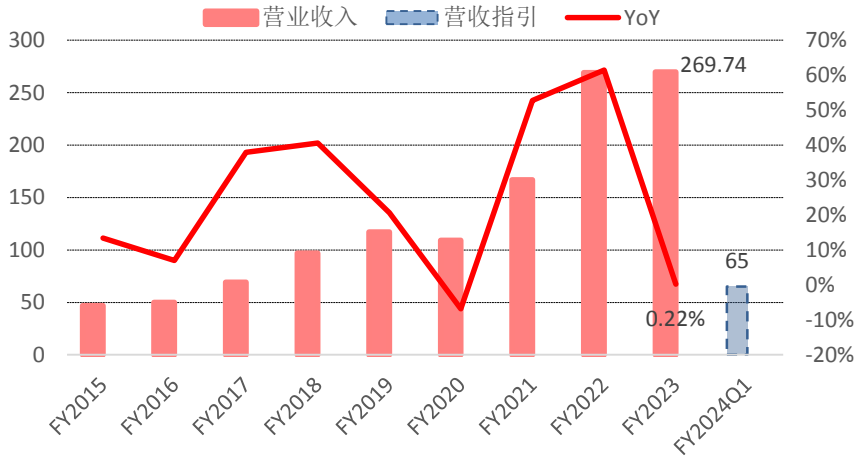
	游戏&娱乐 占FY2023营收33.6%	数据中心 占FY2023营收55.6%	专业可视化 占FY2023营收5.7%	汽车业务 占FY2023营收3.4%
硬件产品	 GeForce系列	 Tesla系列  云端处理器	 Quadro系列	 自动驾驶芯片
软件技术	GEFORCE RTX Studio RTX studio驱动 NVIDIA MAX-Q MAX-Q技术	NVIDIA CUDA CUDA架构 NVIDIA HPC SDK HPC软件开发包SDK NVIDIA NGC NGC软件优化中心	NVIDIA MAXINE MAXINE 视频增强 NVIDIA BROADCAST AI增强会议软件	DRIVE OS DRIVE AV DRIVE IX DRIVE Map DRIVE Works DRIVE Chauffeur DRIVE Concierge
开发平台	GEFORCE NOW GeForce Now 云游戏平台	NVIDIA VIRTUALIZATION 虚拟GPU 平台 Omniverse 设计平台	NVIDIA CLOUDXR 云XR平台 NVIDIA STUDIO Sudio 平台	DRIVE Hyperion 开发平台 DRIVE SDK 模块式开放平台 DRIVE Sim 模拟仿真平台 DRIVE DGX DNN训练平台

资料来源：英伟达官网，中信建投

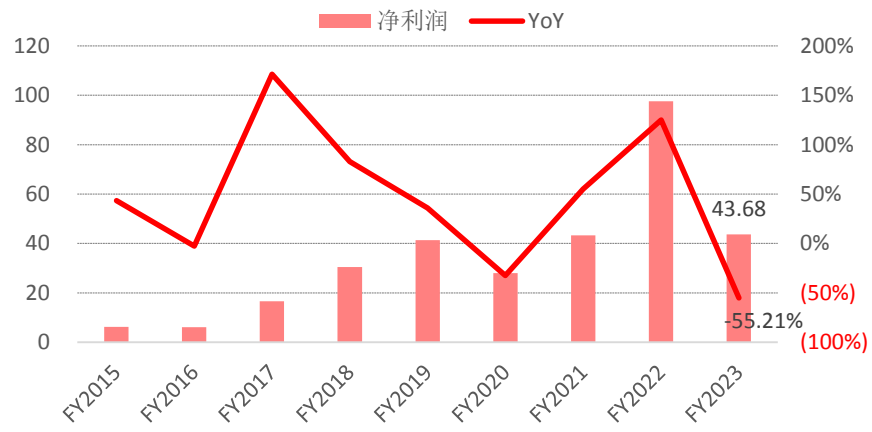
3.3 公司盈利能力历史表现优异

- 公司FY2023年实现营业收入269.74亿美元，与FY2022年同比基本持平。数据中心业务保持快速增长趋势，游戏业务、专业可视化业务营收相对下滑。FY23Q4营业收入为60.5亿美元，同比下降21%，但是环比提升2%，收入业绩的恢复性增长主要得益于游戏业务的快速复苏。公司FY24Q1营收指引为65亿，整体业务重回环比正增长阶段。
- FY2023年GAAP净利润43.68亿美元，同比下降55.21%。第四季度GAAP净利润6.8亿美元，同比下降72%。FY2023财年游戏显卡以及数据中心计算芯片的需求相对疲软，供大于求带来了较高的库存水平，导致了大额的资产减值损失，净利润水平有所下滑。

图：英伟达营业收入及增速（亿美元）



图：英伟达净利润及增速（亿美元）



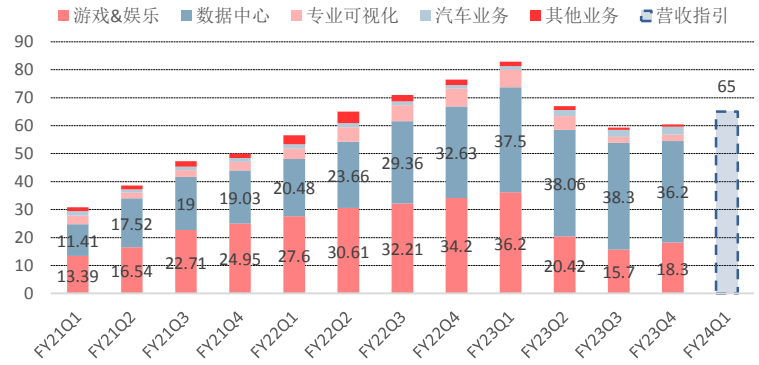
资料来源：英伟达年报，中信建投
备注：英伟达财年为上年1月31日至当年1月30日

中信建投证券 CHINA SECURITIES

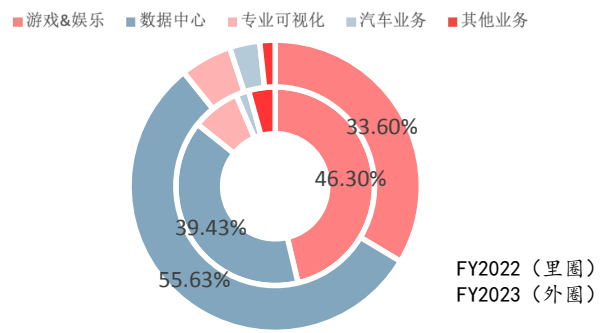
3.4 2022年公司营收结构发生较大变化

- 公司FY2023营收结构发生较大变化，数据中心业务成为主要收入来源，占比55.63%，游戏业务占比下滑。
- FY2023数据中心业务营收达150亿美金，同比增长55.6%，该业务是公司的未来成长引擎，得益于人工智能算力的需求高增，业务保持中长期良好增长态势，FY23Q4受云厂商资本开支影响，以及中国市场需求相对疲软，营收略有下滑。
- FY2023游戏业务营收为90.6亿美金，同比下滑27.3%，营收占比为33.6%。FY23Q2后，受显卡市场冲击，游戏业务营收连续两个季度下滑，FY23Q4得到恢复性增长。
- FY2023专业可视化业务营收达15.44亿美金，同比下滑26.7%。
- FY2023汽车业务营收达到9.03亿美元，同比增长59.5%，主要受益于自动驾驶解决方案的销售增长，营收占比从2021年的2.1%上升到3.35%。

图：英伟达主营业务收入构成（亿美元）



图：英伟达主营业务营收占比情况

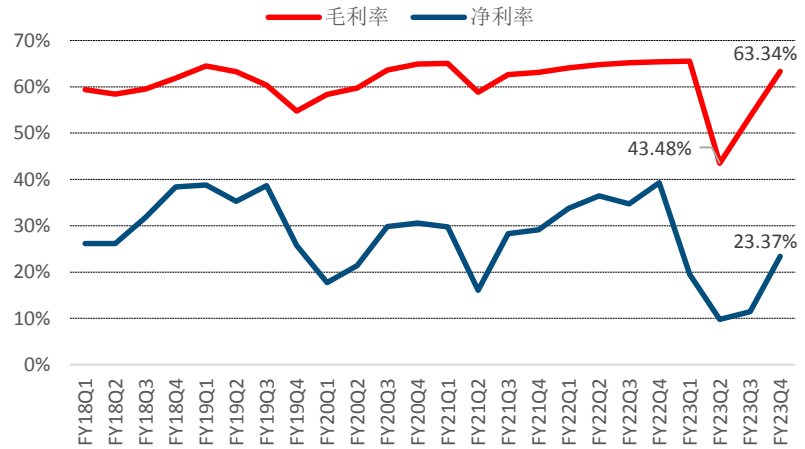


资料来源：英伟达年报，中信建投

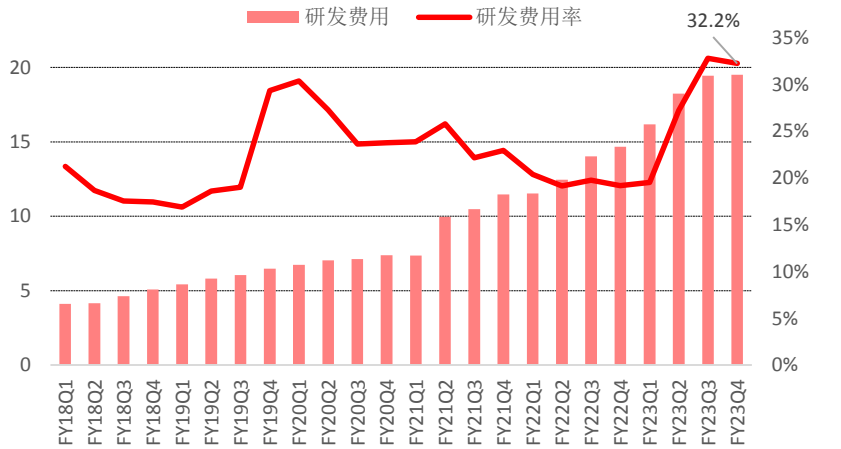
3.5 公司盈利能力水平恢复

- **近期公司整体毛利率、净利率总体恢复到良好水平。**英伟达成立之初，公司的毛利率只有30%+，规模效应促使毛利率在2011年达到50%以上。随着数据中心业务占比不断提升，单价近万美元的Tesla系列加速卡的规模化出货又进一步提升毛利率。FY2022年毛利率提升至64.93%。FY23Q1-Q3公司毛利率水平有所下降，Q3毛利率为43.48%，主要原因是库存和相关储备导致较大的资产减值损失。得益于RTX 40系列显卡的推出，FY23Q4重回63.34%的良好水平。
- **公司研发支出不断增长，研发费用率基本保持在18%以上，以提高在AI领域中的竞争优势。**公司研发投入处于行业较高水平，FY2023年研发费用率为25.75%，保持较高研发投入。

图：公司毛利率、净利率



图：英伟达研发投入情况（亿美元）

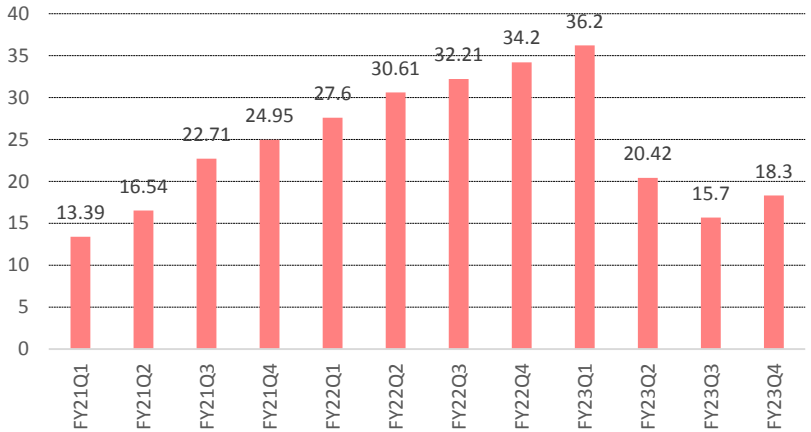


资料来源：Bloomberg，中信建投

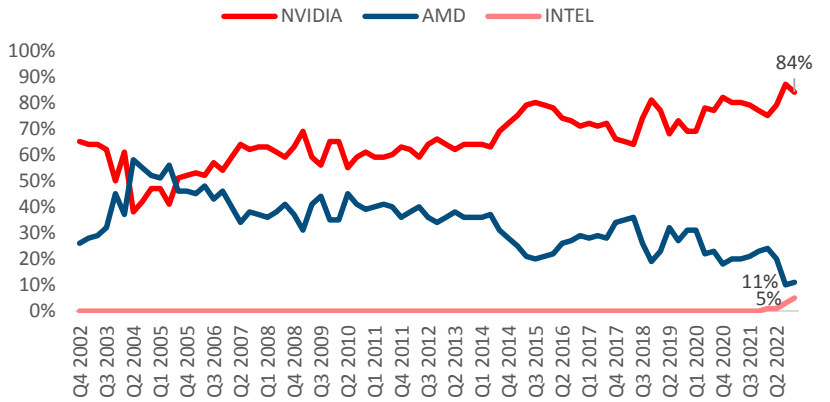
3.6 公司游戏业务简介

- **公司在游戏领域的产品主要包括：GPU芯片和硬件产品、GeForce Now云游戏平台等。**
- **GPU硬件产品主要包括GeForce RTX系列显卡和GeForce GTX系列显卡。**GeForce RTX公司2019年推出的新一代具备先进的光线追踪和AI技术的游戏显卡，采用深度学习采样DLSS及NVIDIA Broadcast等全新前沿AI技术。GeForce GTX系列最早在2007年推出，不含DLSS和光线追踪技术，性价比相对较高，在市场上仍占有相当重要的地位。
- **市场寒冬过后，英伟达游戏业务开始逐步回暖。**受显卡市场影响，FY23Q2以来，英伟达游戏业务收入大幅下滑，FY23Q3单季度同比下滑51.2%，FY23Q4游戏业务开始回暖。

图：英伟达游戏业务单季度营收（亿美元）



图：独立GPU市场份额占比



资料来源：英伟达年报，中信建投

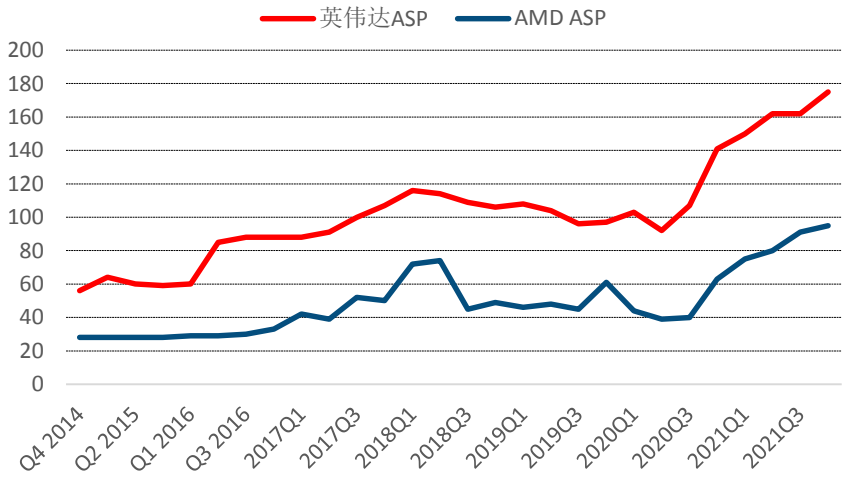
3.7 GPU性能增强带动ASP持续提升

- **英伟达不断推出性能更高的新产品。**英伟达每年都会发布多款GPU产品，新产品的晶体管数目、制程、CUDA核心数、显存容量、渲染和运算能力等方面均有提升。
- **单个GPU的ASP不断提升。**GeForce RTX价值量提升明显，RTX 4090显卡比2020年3090首发时贵了8%，而4080比3080Ti贵了6%。

图表：英伟达RTX系列产品单价及性能

产品型号	单价（元）	CUDA Core 核心数	显存容量
GeForce RTX 4090	12999	16384	24 GB
GeForce RTX 4080	9499	9728	16 GB
GeForce RTX 3090	11999	10496	24 GB
GeForce RTX 3080 Ti	8999	10240	12 GB
GeForce RTX 3080	5499	8960 / 8704	12 GB / 10 GB
GeForce RTX 3070 Ti	4499	6144	8 GB
GeForce RTX 3070	3899	5888	8 GB
GeForce RTX 3060 Ti	2999	4864	8 GB
GeForce RTX 3060	2499	3584	12 GB
GeForce RTX 3050	1899	2560	8 GB

图：英伟达产品ASP持续上升（美元）

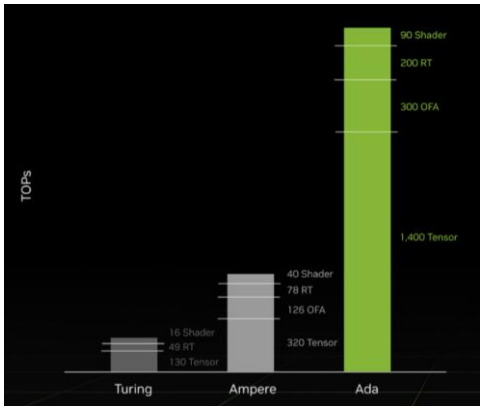


资料来源：英伟达官网，bloomberg，中信建投

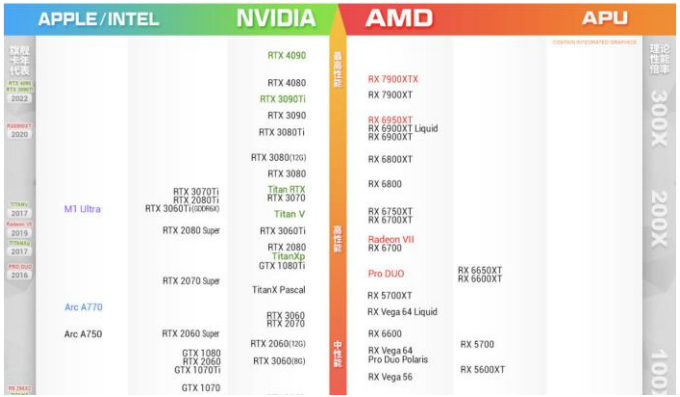
3.8 公司游戏GPU具有显著的技术优势

- **GeForce RTX 40系列显卡实现游戏性能的大幅提升。**GeForce RTX 40系列显卡采用英伟达Ada Lovelace架构，采用第三代RT Core技术实现全景光追性能提升至4倍，DLSS 3技术让渲染帧率成倍增加，配合着色器执行重排序技术、Nvidia Reflex等技术使其性能相较于Ampere架构提升至两倍以上。

图：Ada架构实现性能的显著提升



图：显卡天梯图



图：steam用户显卡统计

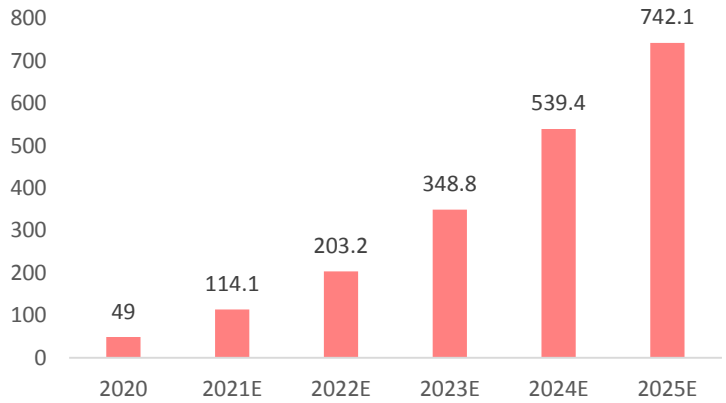
所有显卡	SEP	OCT	NOV	
NVIDIA GeForce GTX 1650	6.32%	5.61%	6.27%	+0.66%
NVIDIA GeForce GTX 1060	6.94%	7.62%	5.77%	-1.85%
NVIDIA GeForce RTX 2060	5.19%	6.10%	4.64%	-1.46%
NVIDIA GeForce RTX 3060 Laptop GPU	4.00%	3.39%	4.63%	+1.24%
NVIDIA GeForce GTX 1050 Ti	4.91%	4.39%	4.60%	+0.21%
NVIDIA GeForce RTX 3060	3.52%	5.47%	3.41%	-2.06%
NVIDIA GeForce GTX 1660 Ti	2.57%	2.28%	2.46%	+0.18%
NVIDIA GeForce RTX 3070	2.33%	2.76%	2.44%	-0.32%
NVIDIA GeForce GTX 1660 SUPER	2.52%	2.34%	2.41%	+0.07%
NVIDIA GeForce GTX 1050	2.45%	2.18%	2.40%	+0.22%
NVIDIA GeForce RTX 3060 Ti	2.10%	2.53%	2.27%	-0.26%
NVIDIA GeForce RTX 3050	1.94%	1.83%	2.26%	+0.43%
AMD Radeon Graphics	1.74%	1.63%	1.93%	+0.30%
NVIDIA GeForce GTX 1070	1.97%	1.88%	1.89%	+0.01%
NVIDIA GeForce RTX 3080	1.69%	1.82%	1.84%	+0.02%

资料来源：英伟达，中信建投

3.9 云游戏业务有望成为未来游戏业务支柱

- **全球云游戏市场增长迅速。**根据IDC的数据，2020年全球云游戏市场规模为49亿元，2025年全球云游戏市场规模将达到742.1亿元，预期年均复合增速72%。谷歌、微软、索尼、Facebook、NVIDIA、Valve、腾讯，以及各大游戏厂商纷纷布局云游戏业务。
- **英伟达GeForce Now云游戏有望成为未来游戏业务支柱。**GeForce Now目前已支持1500余款游戏，支持Steam、Epic Games、GOG.com等游戏启动器，覆盖75个国家和地区。
- 2023年CES大会上，英伟达宣布GeForce NOW云游戏服务登陆汽车平台，首批支持汽车品牌包括比亚迪、现代、起亚、捷尼赛思以及Polestar极星。

图：全球云游戏市场规模(亿元)



图：GeForce Now云游戏登录汽车平台

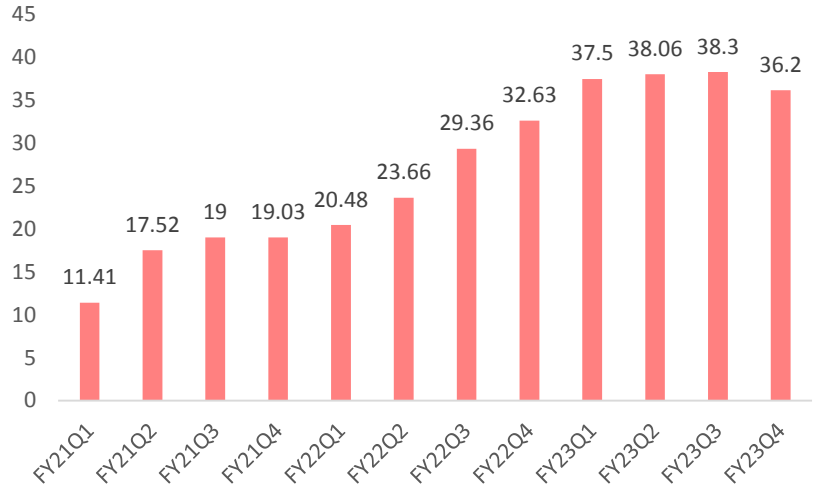
资料来源：IDC，英伟达，中信建投

3.10 公司逐步成为全球AI芯片领域的主导者

- **英伟达的通用计算芯片具备优秀的硬件设计，通过CUDA架构等全栈式软件布局，深度挖掘芯片硬件的性能极限，在各类下游应用领域中，均推出了高性能的软硬件组合，逐步成为全球AI芯片领域的主导者。**
- 早期英伟达在数据中心的产品布局主要为GPU加速服务器。通过不同型号的GPU加速器与CPU、DPU等其他硬件产品组合以及软件的开发，英伟达还推出了面向高性能计算(HPC)、人工智能(DGX)、边缘计算(EGX)等领域中的硬件产品。

图：公司数据中心主要产品

图：公司数据中心业务单季度营收（亿美元）



资料来源：英伟达官网，中信建投

3.11 数据中心业务核心优势一：AI计算能力不断提升

- **从英伟达GPGPU芯片发展历程来看，通过不断提升计算单元数量和引入张量核心，实现了计算能力的提升。**每一代新型架构下的GPGPU均实现了各种数据格式下计算能力的提升，同时通过张量核心的引入，大幅提升高性能计算和AI计算能力。
- **在人工智能领域，公司Transformer引擎技术实现Transformer模型的加速运行。**Transformer模型是当前自然语言处理的主流模型，并且越来越多应用在计算机视觉等其他深度学习领域。公司Transformer引擎是一种定制Tensor Core技术，针对Transformer模型的每一层参数进行分析，灵活使用混合精度从而显著提升模型运行速度。
- **公司与云服务供应商加强合作，实现AI算力云化。**2023春季GTC大会上，英伟达宣布与谷歌云、微软Azure、甲骨文云联手推出DGX云服务，为中小型企业提供了更加便捷的AI算力获取方式。

图表：英伟达GPGPU性能指标

型号	H100	A100	A800	V100
FP64(TFlops)	34	9.7	9.7	7.8
FP32(TFlops)	67	19.5	19.5	15.7
FP16(TFlops)	133.8	78	78	-
INT8 Tensor(Tops)	3958	1248	1248	-
GPU显存（GB）	80	80	80	32
显存带宽（GB/s）	3350	2039	2039	900
互连（GB/s）	900	600	400	300
功耗（W）	700	400	400	300
发布时间	2022.03	2020.03	2022.11	2017.5

图:不同型号GPGPU的性能比对

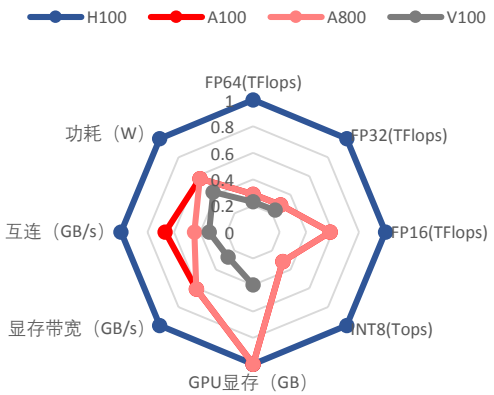
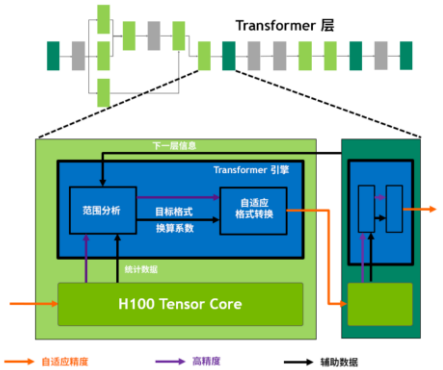


图:英伟达GPU实现Transformer模型加速



资料来源：英伟达官网，中信建投

3.12 数据中心业务核心优势二：芯片互联能力不断提升

- **人工智能领域进入千亿参数的大模型时代，AI算力需求不断增长，在这种趋势下，对数据中心的协同计算能力要求越来越高，对于能够在GPU之间实现无缝高速通信的多节点、多GPU系统的需求也在与日俱增。**
- **NVIDIA NVLink技术最大化地提升GPU吞吐量。**借助NVIDIA NVLink技术，单个NVIDIA H100 GPU通过18路NVLink连接实现900 GB/s总带宽，是PCIe 5.0带宽的7倍。
- **NVIDIA NVSwitch芯片可为计算密集型工作负载提供更高带宽和更低延迟。**每个NVSwitch包含64个NVLink端口，实现8 GPU的高速互联，可以提供无缝、高带宽的多节点GPU集群。
- **双层NVSwitch最多实现256个GPU的高速互联。**通过在服务器外部添加第二层NVSwitch，NVLink网络可以连接多达256个GPU，并提供57.6TB/s的多对多带宽，从而快速完成大型AI作业。

图表：NVLink互联

NVLink	第二代	第三代	第四代
总带宽	300GB/s	600GB/s	900GB/s
单GPU最大链路数	6	12	18
架构支持	Volta	Ampere	Hopper

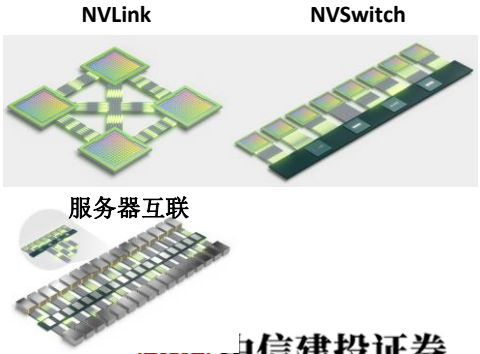
图表:NVSwitch互联

NVSwitch	第二代	第三代	第四代
直连或节点中GPU数量	最多8个	最多8个	最多8个
NVSwitch GPU之间带宽	300GB/s	600GB/s	900GB/s
聚合总带宽	2.4TB/s	4.8TB/s	7.2TB/s
架构支持	Volta	Ampere	Hopper

图表:服务器互联

	服务器互联
直连GPU数量	多达256个
NVSwitch GPU之间带宽	900GB/s
聚合总带宽	57.6TB/s
架构支持	Hopper

图:互联技术示例

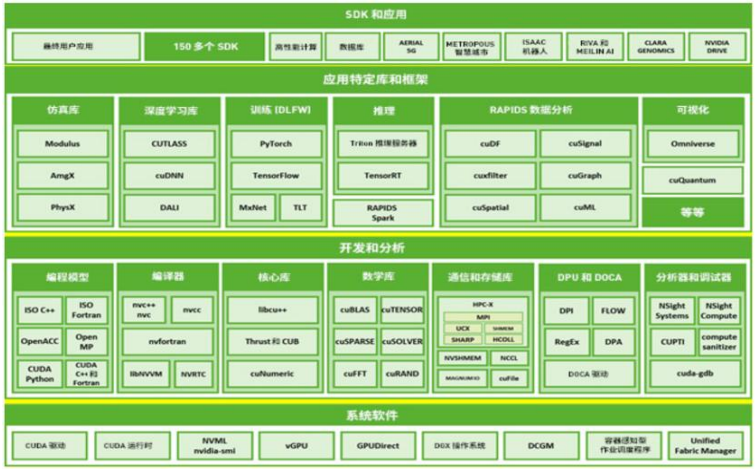


资料来源：英伟达官网，中信建投

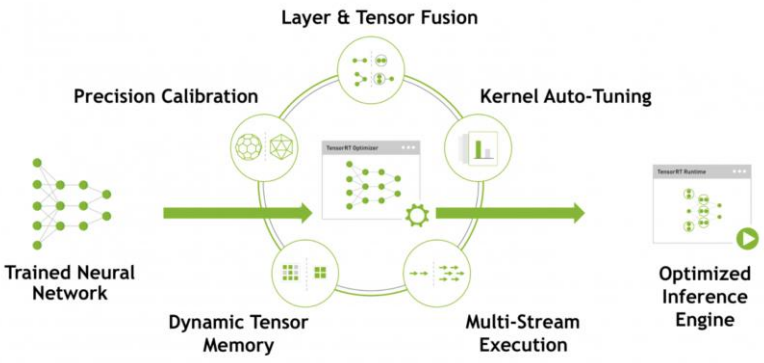
3.13 数据中心业务核心优势三：软硬件深度绑定

- 2006年，英伟达看到了人工智能的兴起及GPU在并行计算方面的优势后，开始斥巨资研发CUDA指令集架构和GPU内部的并行计算引擎。2007年英伟达正式推出CUDA 1.0版本，并使旗下所有GPU芯片都适应CUDA架构，如今**“英伟达GPU+CUDA系统”已成为极具行业壁垒的软硬件生态系统。**CUDA已经迭代至CUDA 11版本，得到开发者的广泛青睐，用户数量不断提升。
- **英伟达开发了用于深度学习的TensorRT推理引擎。**TensorRT基于CUDA并行编程模型，是一个高性能的深度学习计算平台，TensorRT针对深度学习推理提供INT8和FP16优化，深度神经网络的执行速度可比CPU平台快40倍。TensorRT支持Tensorflow、Pytorch、Caffe等深度学习主流框架。

图：英伟达构建的CUDA生态



图：TensorRT实现深度学习推理加速



资料来源：英伟达官网，中信建投

3.14 数据中心业务核心优势四：产品组合不断丰富

- **公司数据中心芯片产品组合已扩展至GPU、CPU、DPU等。**
- **2019年，英伟达以69亿美元并购了Mellanox，推出BlueField系列DPU。**DPU的智能网卡将成为云数据中心设备中的核心网络部件，逐渐承担原先需要CPU来执行的网络数据处理、分发的重任，从而从根本上实现软件定义网络(SDN)和网络功能虚拟化(NFV)的诸多优势，有效降低云计算的性能损失，释放CPU算力，降低功耗的同时大大减少云数据中心的运营成本。最新的BlueField-3芯片能够以400Gbps的速率对网络流量进行保护、卸载和加速。
- **英伟达推出自研CPU Grace，产品组合不断丰富。**在2021GTC大会上，英伟达推出了Grace CPU并计划在2023年量产，这款CPU是英伟达第一次推出的CPU产品，采用了ARM v9指令集，该指令集主要是增强面向矢量、机器学习和数字信号处理器的相关内容，这款CPU的主要应用场景将是在数据中心领域。

图：英伟达不断丰富数据中心产品组合

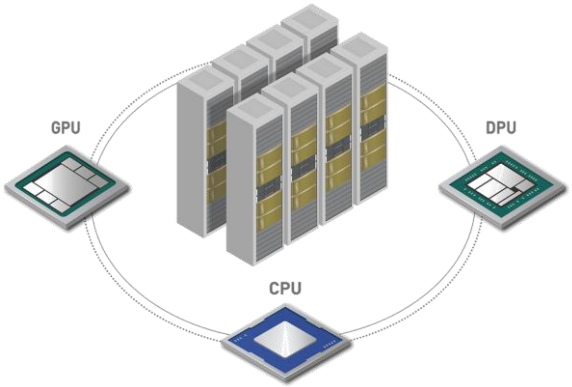
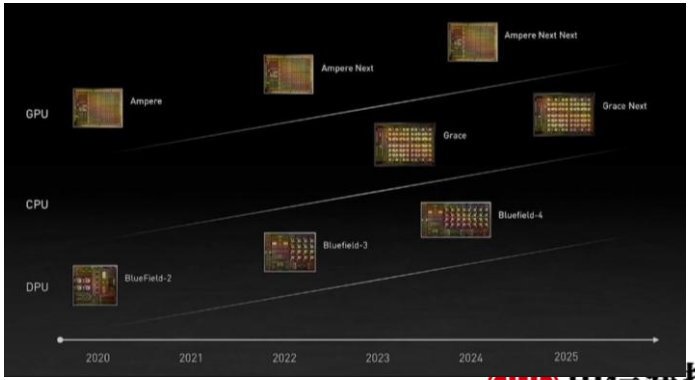


图:数据中心芯片发展路线图

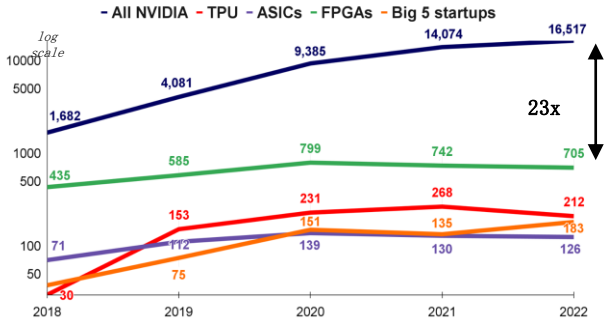


资料来源：英伟达官网，中信建投

3.15 英伟达在AI芯片市场中占据主导地位

- **英伟达凭借优异的硬件性能、不断提升的网络互联能力、CUDA的软硬件协同、以及产品组合的全自研，在AI数据中心和HPC超算中心占据领导者地位。**
- **在学术界，英伟达GPU作为AI芯片的出现频率远超其他类型芯片。**根据stateof.AI 2022报告，英伟达芯片在AI学术论文中的出现频次远超其他类型的AI芯片，是学术界最常用的人工智能加速芯片。
- **在数据中心中，英伟达GPU占据主导地位。**根据LIFTR INSIGHTS数据，在大型数据中心的AI加速芯片中，英伟达的GPU占据了超过80%的AI加速芯片市场份额，在Oracle以及腾讯云中，几乎全部采用英伟达的GPU作为计算加速芯片。在整体数据中心加速芯片市场中，英伟达市场份额为82%，占据主导地位。

图：英伟达芯片在AI学术论文中的出现频次



图：2022年人工智能加速芯片在云上部署情况

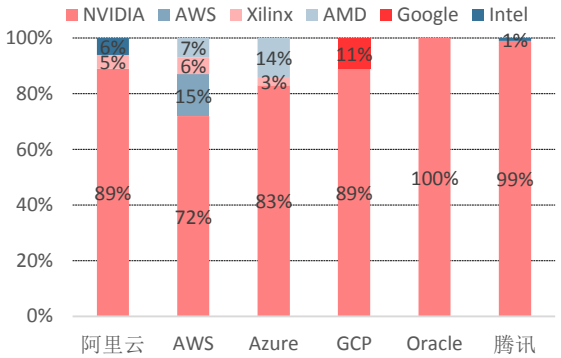
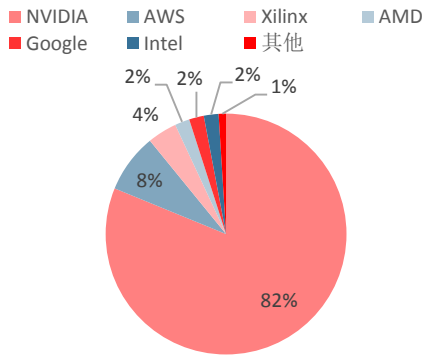


图:2022年人工智能加速芯片市场份额



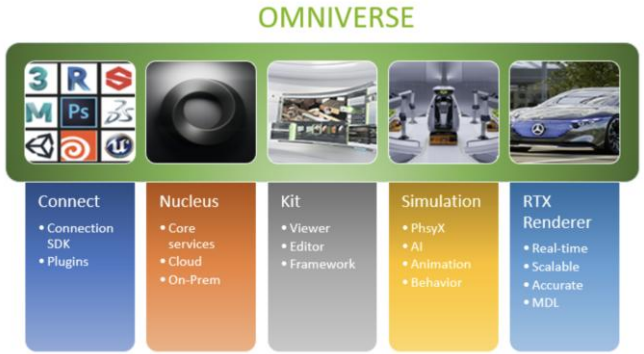
资料来源：stateof AI 2022，LIFTR INSIGHTS，中信建投

中信建投证券 CHINA SECURITIES

3.16 实时协作模拟平台Omniverse

- **Omniverse是一个计算机图形与仿真模拟平台，主要用处是让企业在实际建设工厂、生产产品前，通过数字化模拟“预览”实际的成品。**Omniverse可以应用于媒体娱乐、建筑工程、制造业、自动驾驶等多个领域，利用Omniverse能够将全局照明、实时光线追踪、AI、计算和工程 Simulation 等技术整合到日常工作流程中,提高行业工作流程的灵活性和可扩展性。
- 在汽车制造行业中，沃尔沃和通用汽车使用Omniverse统一产品设计流程，丰田汽车则用来创建数字孪生工厂，奔驰使用这款软件建立和优化新车的生产线，宝马计划在2025年投产的新电动车工厂已经在软件中成功运作。
- **Omniverse平台支持元宇宙的应用程序开发，面向用户提供生成式AI扩展工具。**创作者可使用Audio2Face，基于音频文件生成面部表情；使用Audio2Emotion，生成从快乐和兴奋到悲伤和遗憾的逼真情绪；使用 Audio2Gesture，实现逼真的上半身动作。

图： Omniverse关键模块和功能



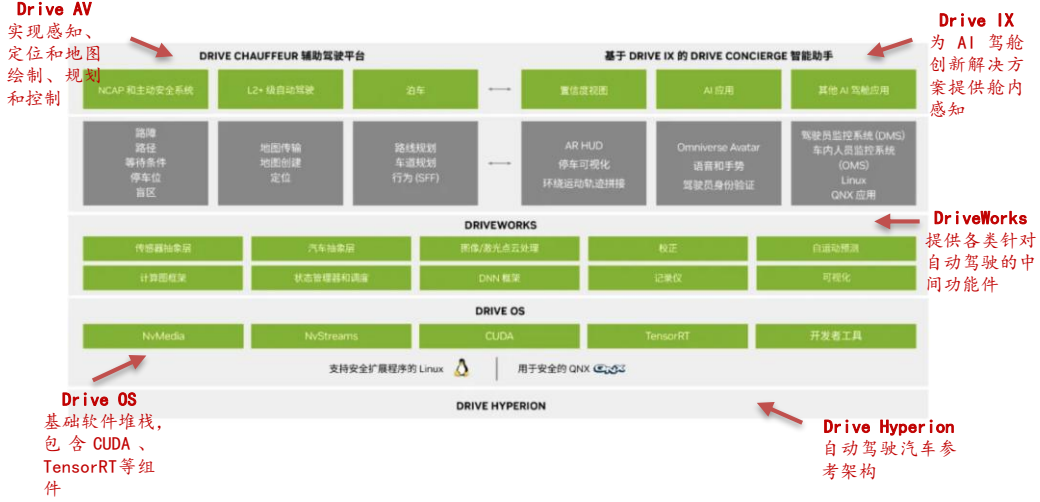
图：基于Omniverse打造的宝马数字工厂

资料来源：英伟达官网，中信建投

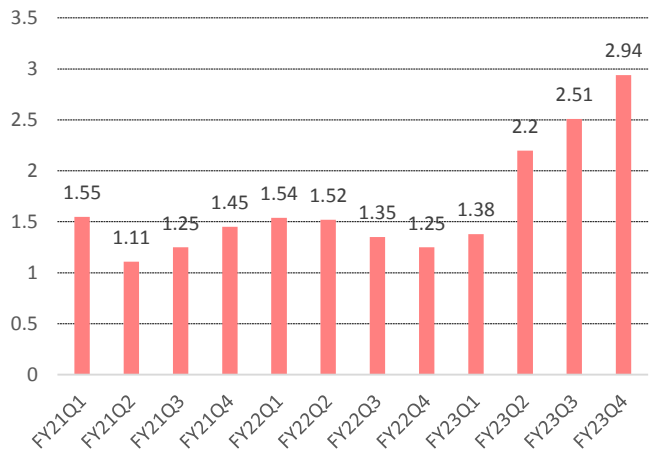
3.17 英伟达提供全栈式的自动驾驶产品解决方案

- **在自动驾驶领域，英伟达提供平台化芯片以及算法开发工具链，已经形成了全栈式的自动驾驶产品解决方案。**在硬件层面，公司推出Xavier、Orin、Thor的高等级自动驾驶芯片。在软件层面，公司推出了自动驾驶配套的底层开发平台Drive OS、模块化定制软件DriveWorks、自动驾驶软件栈Drive AV和AI辅助驾驶平台Drive IX等自动驾驶汽车软件，实现感知、定位和地图绘制、规划和控制、驾驶员监控和自然语言处理等主要功能。通过“硬件+软件”的一体化解决方案，实现L2-L5的自动驾驶应用场景全覆盖，助力下游客户进行自动驾驶技术的测试与开发。

图：公司自动驾驶产品提供“硬件+软件”的整体解决方案



图：公司汽车业务单季度营收（亿美元）

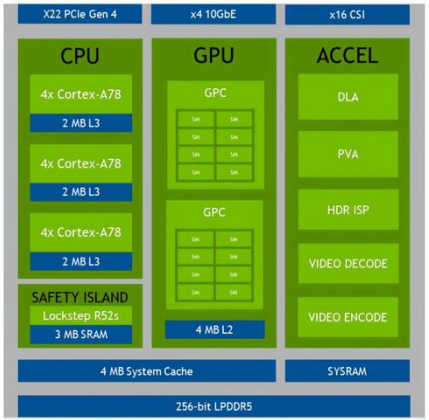


资料来源：英伟达官网，中信建投

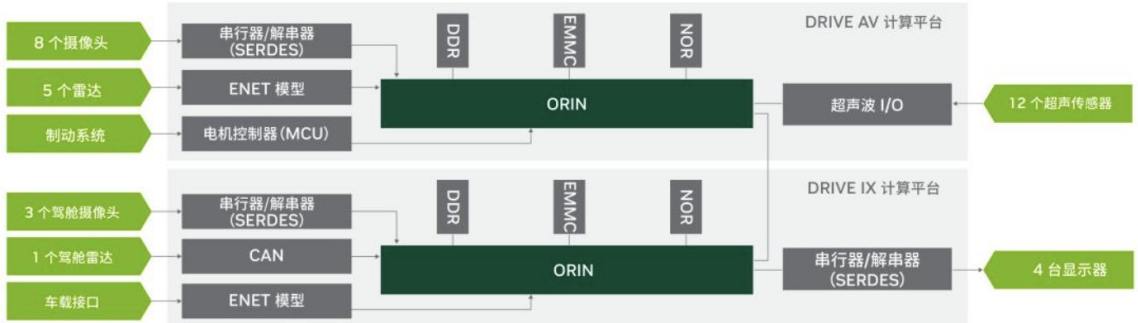
3.18 自动驾驶Orin芯片

- 2021年12月，英伟达正式推出采用Orin芯片，相比前一代Xavier的算力提升7倍，从30 TOPS提升到了254 TOPS。 Orin硬件架构可以简单分为5部分，存储、外围、CPU、GPU和加速器，集成了采用12核的ARM Cortex-A78 CPU，新一代Ampere架构GPU以及全新深度学习加速器DLA和计算机视觉加速器PVA。
- Orin可以覆盖L2-L5的自动驾驶计算需求。单个Orin芯片最高提供6个CSI摄像头接口，通过虚拟通道增加到16个，可以承载4个800万摄像头。
- NVIDIA DRIVE L2+解决方案由两个NVIDIA Orin系统级芯片提供支持，一个用于主动安全、自动驾驶和停车应用，另一个用于AI座舱功能。双Orin芯片可以承载8个800万像素摄像头，5个激光雷达，12个超声波雷达，实现360度场景感知。

图：Orin硬件架构



图：双Orin架构的自动驾驶解决方案

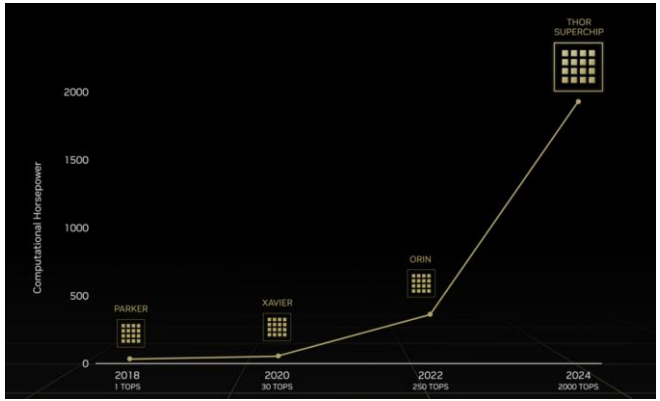


资料来源：英伟达官网，中信建投

3.19 Thor芯片：算力大幅提升，提供全栈式解决方案

- **Thor芯片AI算力大幅提升。**2022年9月，英伟达正式推出采用Thor芯片，单颗芯片算力达到2000 TOPS，性能约是目前主流的英伟达Orin芯片的8倍，单颗FSD芯片的28倍，预计于2025年量产。Thor采用了面向高性能计算HPC的Grace CPU，GPU部分采用RTX 40系列的Ada Lovelace架构和针对Transformer深度神经网络模型优化的Hopper架构，同时采用NVLink互联技术。Hopper架构兼容的FP8精度格式，从而实现神经网络的模型加速。
- **Thor芯片提供全栈式智能汽车解决方案。**Thor芯片提供的极高算力可以同时将包括自动驾驶和辅助驾驶、泊车、驾乘人员监控、数字仪表板、车载信息娱乐等智能功能，统一整合到单个架构中，提供一套包括车身控制、娱乐等在内的全栈式解决方案，降低系统的运行能耗、提升效率，同时降低智能汽车的研发难度。

图：Thor芯片实现计算能力的显著提升



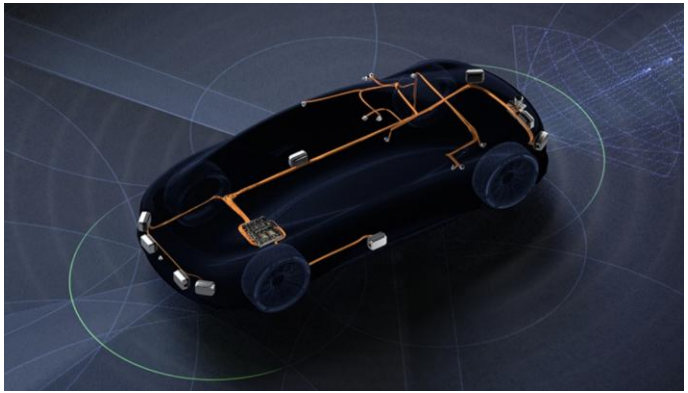
图：单一Thor芯片提供全栈式智能汽车解决方案

资料来源：英伟达官网，中信建投

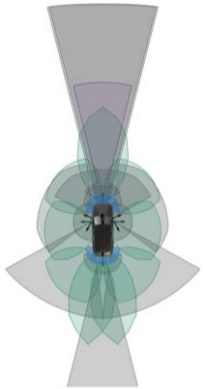
3.20 DRIVE Hyperion自动驾驶平台

- **英伟达DRIVE Hyperion是自动驾驶汽车开发平台和参考架构，用于开发高等级的自动驾驶解决方案。**DRIVE Hyperion构建于Orin芯片基础之上，还包含适用于自动驾驶的开发者软件套件与完整传感器套件（12个外部摄像头、3个内部摄像头、9个雷达、12个超声波、1个前置激光雷达）。通过准确的传感器校正、精确的时间同步、集成的实用程序实现了算法的开发加速。DRIVE Hyperion同时支持无线更新，能够在车辆的完整生命周期内添加新的特性和功能，实现跨代兼容。
- **DRIVE Hyperion得到新能源汽车制造商的广泛青睐。**智己汽车、理想汽车、蔚来汽车、飞凡汽车和小鹏汽车等许多新能源汽车制造商采用DRIVE Hyperion作为平台来开发智能的车型。英伟达在GTC 2023表示，从2023年上半年起，比亚迪将在部分新车上搭载英伟达DRIVE Hyperion平台，实现车辆智能驾驶和智能泊车。

图：DRIVE Hyperion示例



图：Hyperion 8.1传感器规格



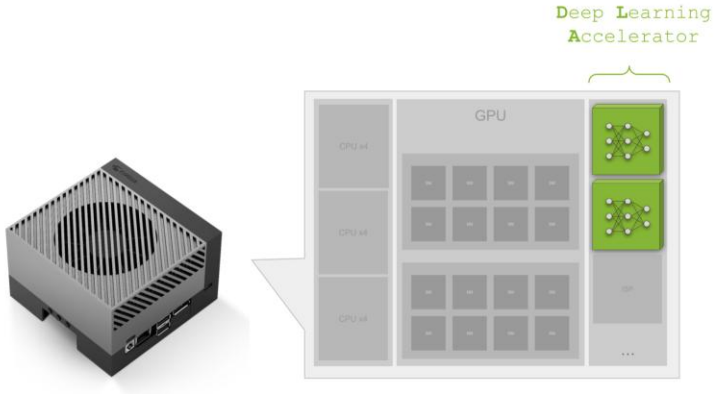
传感器	功能
8个外部摄像头	广域和远距视野
4个外部摄像头	鱼眼近距视野
6个雷达	角落和侧面感知
3个雷达	前后感知
1个激光雷达	前面冗余感知
3个内部摄像头	驾驶域乘客监控
其他传感器（IMU、GPS、GNSS等）	其他信息采集

资料来源：英伟达官网，中信建投

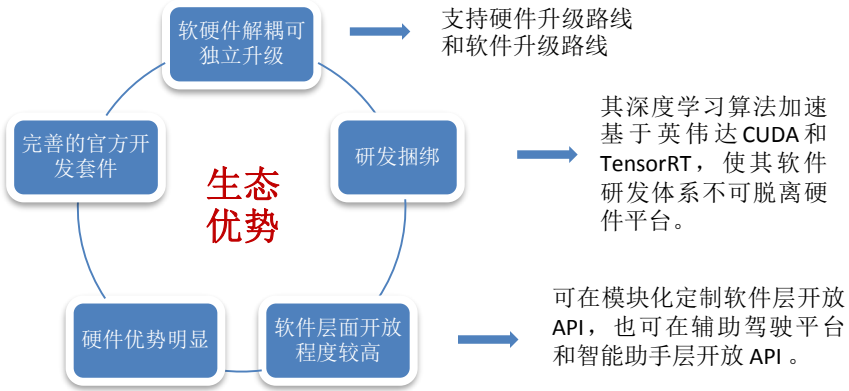
3.21 公司自动驾驶芯片技术领先，同时具备平台化优势

- **公司自动驾驶芯片通过DLA模块和PVA模块实现AI算法加速。**DLA（Deep Learning Accelerator）是一种专用于AI推理的深度学习加速器，英伟达DLA模块由MAC（乘积累加运算）阵列组成，能够有效地执行深度学习的固定推理操作。可编程视觉加速器（PVA），专注于视觉相关的处理，能够比GPU或者DLA模块更快、更好地处理对象检测等视觉处理中的基本任务。
- **公司自动驾驶架构具备灵活、可快速迭代的优势。**公司布局了完整的软件堆栈，围绕着车端、桌面端、云端构建了开发者平台，其上包含各类中间件和成熟的算法模块，形成完整的工具链和丰富的软件生态。客户可以在任何一层买入英伟达的服务，搭建自己的算法或者应用。配合英伟达的高算力自动驾驶芯片，实现自动驾驶算法的开发加速。

图：自动驾驶芯片中的DLA模块实现AI加速



图：公司自动驾驶业务优势

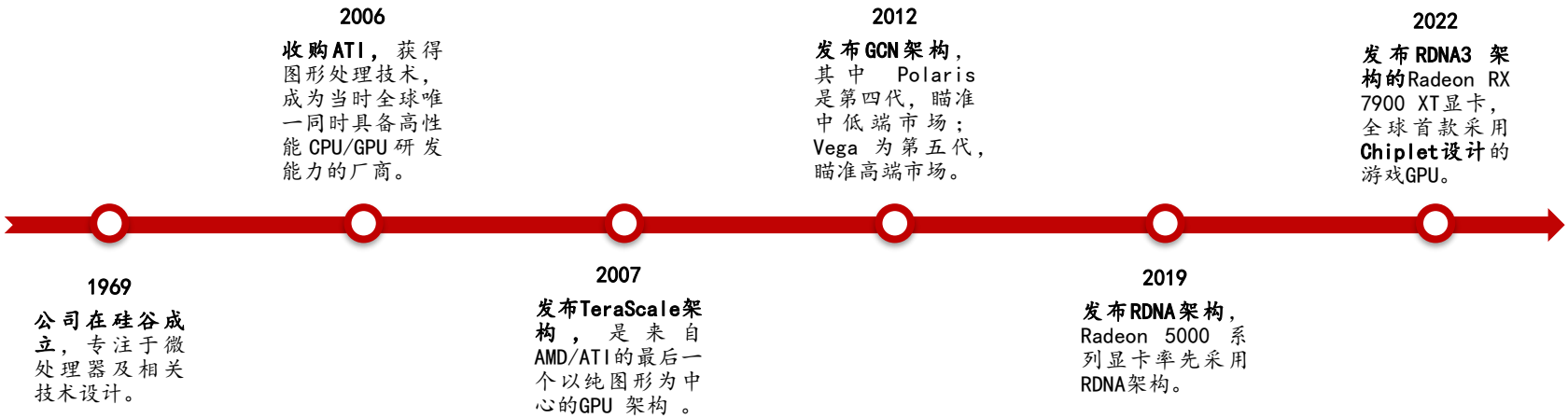


资料来源：英伟达官网，中信建投

4.1 AMD简介

- 美国超威半导体公司(Advanced Micro Devices，AMD)创立于1969年，专门为计算机、通信和消费电子行业提供各类**微处理器以及提供闪存和低功率处理器方案**，公司是**全球领先的CPU、GPU、APU和FPGA设计厂商**，掌握中央处理器、图形处理器、闪存、芯片组以及其他半导体技术，具体业务包括数据中心、客户端、游戏、嵌入式四大部分。公司采用Fabless研发模式，聚焦于芯片设计环节，制造和封测环节则委托给全球专业的代工厂处理。目前全球CPU市场呈Intel和AMD寡头垄断格局，Intel占主导地位。在独立GPU市场中，主要是英伟达（NVIDIA）、AMD进行角逐，Intel目前凭借其锐炬Xe MAX产品也逐步进入独立GPU市场。

图：AMD GPU业务发展史

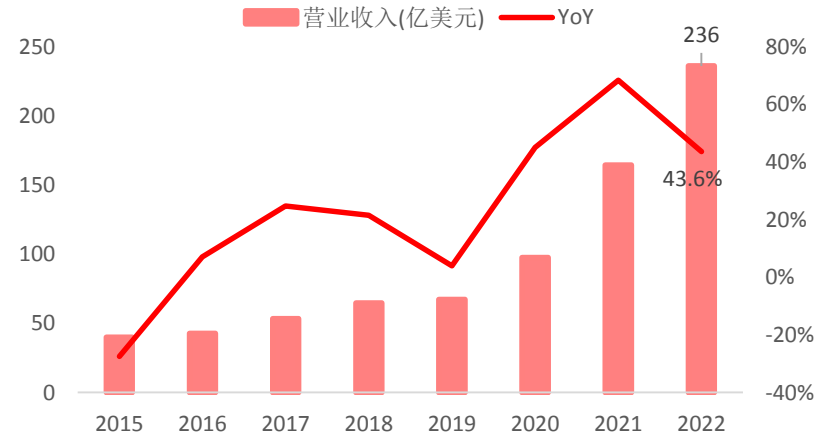


资料来源：AMD官网，中信建投

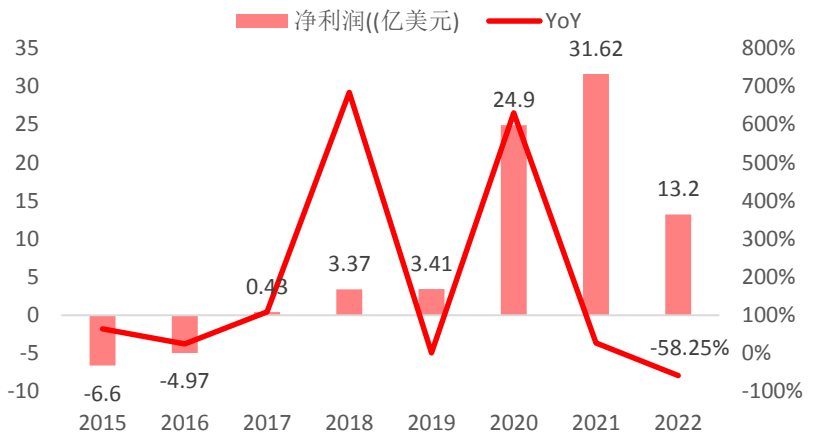
4.2 AMD保持良好的增长态势

- **得益于公司数据中心、嵌入式业务的快速增长，公司营收和净利润实现规模提升。**2022年公司营业收入236亿美元，同比增长43.6%；2022Q4公司营收55.99亿美元，同比增长16%。
- 2022年公司净利润13.2亿美元，同比下降58.25%；2022Q4净利润0.21亿美元，同比下降98%，主要原因系收购赛灵思的无形资产摊销导致净利润下滑。
- 公司预期2023Q1营收53亿美元，同比下滑10%。客户和游戏的细分市场预计会同比下降，部分被嵌入式和数据中心细分市场增长所抵消。

图：AMD营收及增速



图：公司净利润及增速

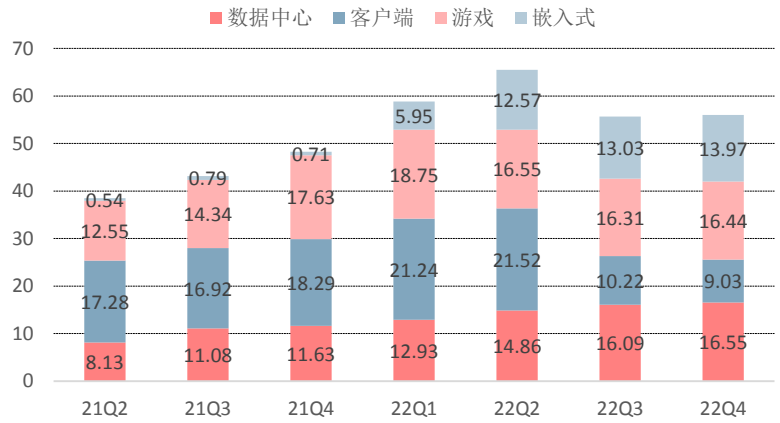


资料来源：AMD，中信建投

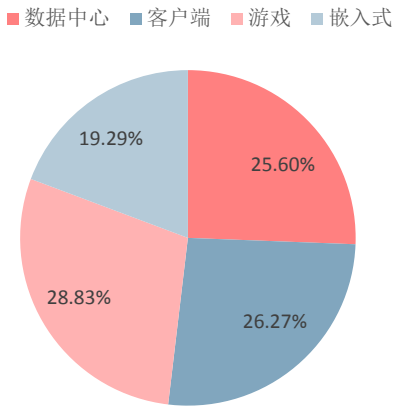
4.3 AMD分业务营收情况

- **公司营收主要包括四部分。**数据中心业务主要包括用于数据中心服务器的各类芯片产品；客户端业务主要包括用于PC的各类处理器芯片；游戏业务主要包括独立GPU及其他游戏产品开发服务；嵌入式业务主要包括适用于边缘计算的各类嵌入式计算芯片。
- **公司数据中心、嵌入式业务的营收增长较快。**2022年，公司数据中心业务收入60.43亿美元，营收占比25.60%；客户端业务收入62.01亿美元，营收占比26.27%；游戏业务收入68.05亿美元，营收占比28.83%；嵌入式业务收入45.52亿美元，营收占比19.29%。

图：分业务营收情况



图：2022年分业务营收占比情况

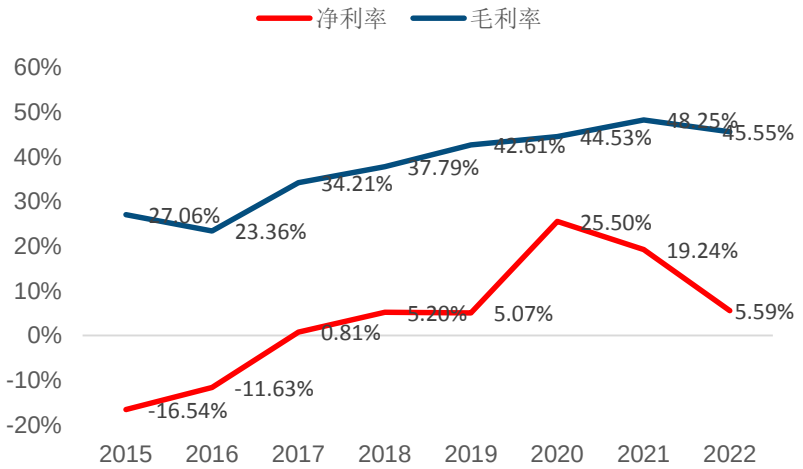


资料来源：AMD，中信建投

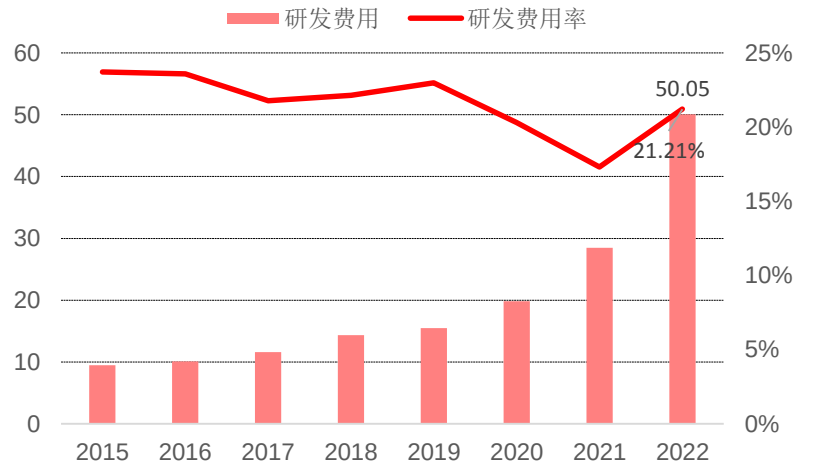
4.4 AMD盈利情况

- 公司2022年毛利率为45.55%，同比下降2.7pt；净利率为5.59%，同比下降13.65pt，主要由于赛灵思收购相关的无形资产摊销以及研发投入的增加。
- 近年来公司不断增加研发投入，2022年研发费用50.05亿美元，同比上升75.9%；研发费用率为21.21%，上升3.9pt，2022年实现了研发费用的大幅提升。截止2022年底，AMD全球员工总数达25000人，相比2021年年底的15500人显著提升。

图：公司毛利率和净利率



图：公司研发费用率



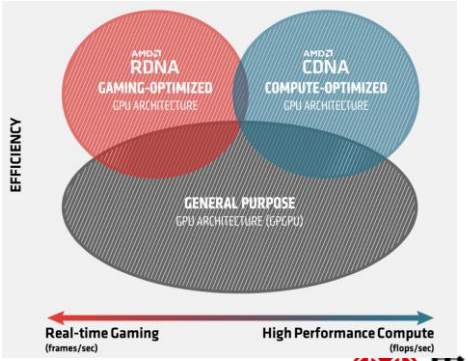
资料来源：AMD，中信建投

4.5 AMD提供集成GPU和独立GPU

- **AMD可以提供集成GPU和独立GPU两类PC GPU。**
- 集成GPU主要被运用在台式机和笔记本的APU产品、嵌入式等产品中，主要用于游戏、移动设备、服务器等应用。APU带有集成的板载GPU，CPU和GPU的高度融合在一起协同计算、彼此加速，相比于独立GPU更具性价比优势。
- **独立GPU为Radeon系列。**AMD的Radeon系列独立GPU按推出时间先后顺序可以分为RX500系列、Radeon 7、RX5000系列、RX6000系列、RX7000系列。Radeon系列显卡具备一定的性价比优势，市场份额有进一步上升的空间。
- **RDNA 3架构采用5nm工艺和chiplet设计**，比RDNA 2架构有54%每瓦性能提升，包括2.7倍AI吞吐量、1.8倍第二代光线追踪技术，5.3 TB/s的峰值带宽、4K 480Hz和8K 165HZ的刷新率等。AMD预计2024年推出RDNA 4架构，将采用更为先进的工艺制造。

图：AMD游戏GPU产品硬件架构

图：AMD不同领域的架构选择

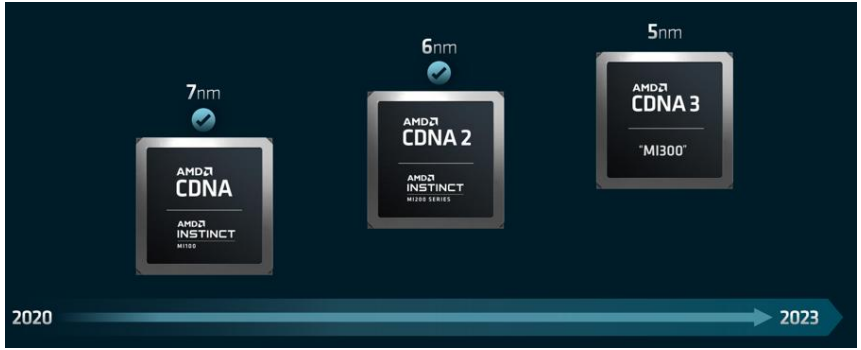


资料来源：AMD官网，中信建投

4.6 CDNA 2架构带来计算性能的大幅提升

- **2018年，AMD推出用于数据中心的Radeon Instinct GPU加速芯片，Instinct系列基于CDNA架构。**
- 在通用计算领域，最新的CDNA 2架构相比CDNA 1架构，实现计算能力和互联能力的显著提升，MI250X采用CDNA 2架构。
- 在向量计算方面，CDNA 2对向量流水线进行了优化，FP64的工作频率与FP32相同，具备同样的向量计算能力。在矩阵计算方面，CDNA 2引入了新的矩阵乘指令级，特别适用于FP64精度，此外Mattrix Core还支持FP32、FP16（BF16）和INT8的计算精度。
- 在互联方面，通过AMD infinity fabric接口实现加速器之间的P2P或者I/O通信，提供800GB/s的总理论带宽，相比上一代提升了235%。

图：AMD数据中心GPU产品架构



图表：AMD和英伟达数据中心GPU产品性能比对

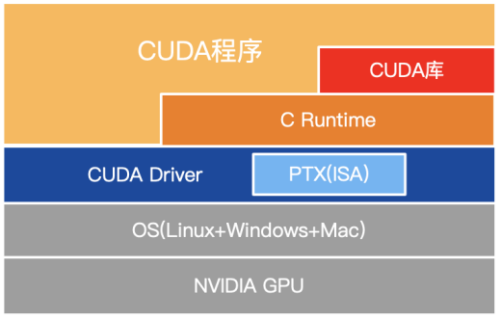
型号	AMD MI250X	英伟达H100	英伟达A100
FP64(TFlops)	47.9	34	9.7
FP32(TFlops)	47.9	67	19.5
FP16(TFlops)	383	1979	624
INT8(Tops)	383	3958	1248
GPU显存(GB)	128	80	80
显存带宽(GB/s)	3277	3350	2039
互连(GB/s)	800	900	600
功耗(W)	560	700	400
发布时间	2021.11	2022.03	2020.03

资料来源：AMD官网，中信建投

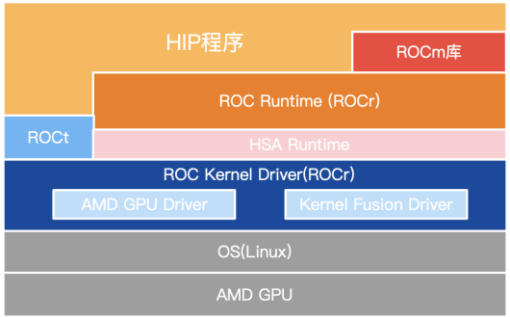
4.7 AMD ROCm计算生态

- **AMD ROCm是Radeon Open Compute (platform)的缩写，是2015年AMD公司为了对标CUDA生态而开发的一套用于HPC和超大规模GPU计算提供的开源软件开发平台。ROCm之于AMD GPU相当于CUDA之于英伟达GPU。**
- **ROCm是一个完整的GPGPU生态系统，在源码级别上实现CUDA程序支持。**ROCm在整体架构上与CUDA类似，实现了主要模块的对齐，封装层次较CUDA更为复杂。ROCm由以下组件组成：HIP程序、ROC运行库、ROCm库、ROCm核心驱动，ROCm支持各类主流的深度学习框架，例如Tensorflow 、PyTorch、Caffe等。

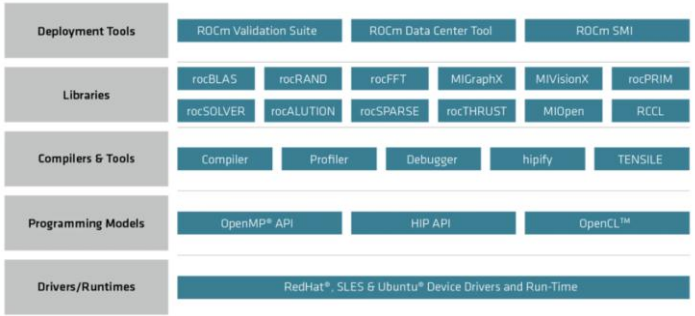
图：NVIDIA的CUDA架构



图：AMD的ROCm架构



图：AMD的ROCm生态组成

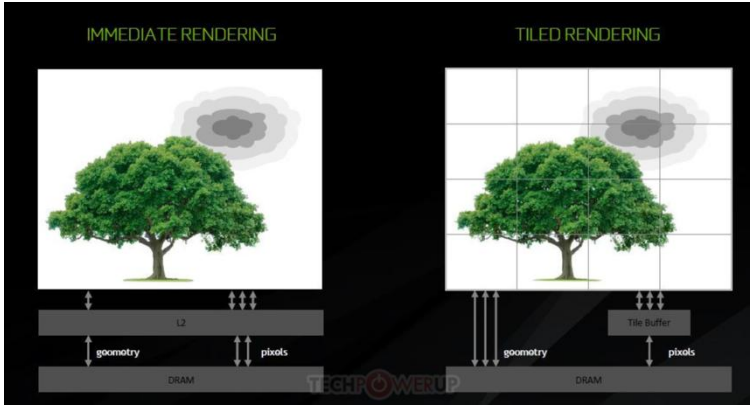


资料来源：CSDN，AMD，中信建投

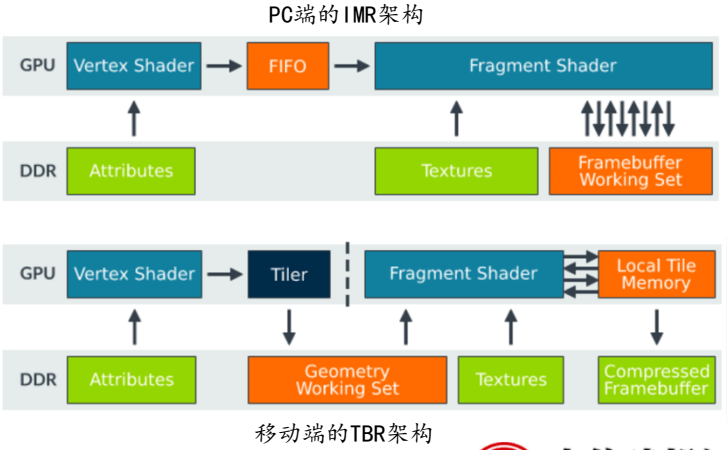
5.1 移动端GPU采用不同的架构设计

- **移动端GPU在设计过程中受到能耗和体积方面的限制，都是以集成的SOC芯片形式出现在移动端，被广泛应用在手机、平板电脑、VR、AR设备、物联网设备当中。**
- SOC芯片中，CPU、GPU共享有限的内存带宽，频繁使用内存带宽会造成较大的能耗，通过采用分块渲染架构（Tile-Based Rendering，TBR）可以有效减少带宽消耗，其核心思想是：将帧缓冲分割为一小块一小块，然后在片上高速内存逐块进行渲染，与PC端采用的及时渲染架构（IMR）相比，极大的减少了DRAM的访问次数，从而降低了整体能耗。
- 分块延迟渲染架构（TBDR）采用影藏面消除（HSR），不会渲染被遮挡的物体表面片，渲染效率进一步提升。

图：分块架构架构



图：即时渲染架构（IMR）和分块渲染架构（TBR）的差异

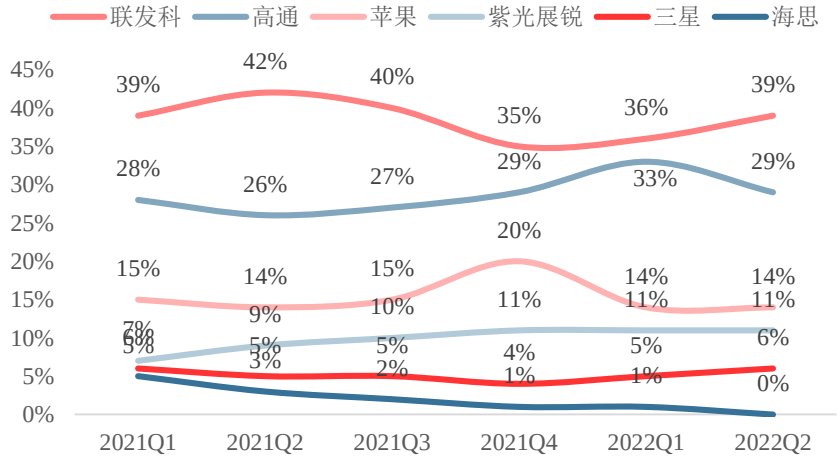


资料来源：TechPowerUp，CSDN，中信建投

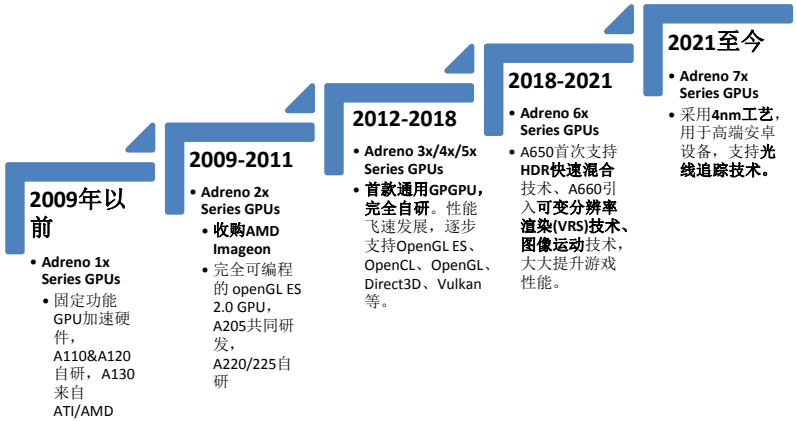
5.2 高通在旗舰Android智能手机SoC市场中保持领先

- 高通自研GPU源自2009年收购于AMD的移动GPU Imageon系列，后改名为Adreno，并集成到自家骁龙SoC中，发展至今已到“Adreno-7”系列，在全球旗舰Android智能手机SoC市场中保持领先。
- 据IDC报告显示，2022Q3全球手机市场出货量下滑8%，高通手机业务营收仍实现40%增长；Counterpoint Research研究显示公司在AP/SoC芯片市场的份额从过往的25%左右提升至30%左右，稳占高端安卓市场。采用骁龙8+的OEM厂商和品牌包括华硕ROG、黑鲨、荣耀、联想、Motorola、努比亚、一加、OPPO、OSOM、realme、红魔、Redmi、vivo、小米和中兴等。

图：全球手机AP/SoC芯片份额



图：高通Adreno GPU发展历史



资料来源：counterpoint，高通，中信建投

5.3 高通移动GPU性能不断提升迭代

- 2018年骁龙855携Adreno 640进入5G时代，2019年高通发布搭载Adreno 660的骁龙888，该GPU是高性能和低功耗的代表产品。Adreno 660采用5nm制程，首次引入可变速率着色(VRS)，为移动设备带来全新桌面级功能，游戏性能提升明显；桌面正向渲染技术以超逼真的细节提升画面从电影景深、运动模糊到动态灯光、阴影多个场景的质感；使用HDR Fast Blend，运行HDR游戏的速度最高可提高2倍，可加速多层的混合。
- 2022年11月，公司发布全新4nm级GPU Adreno 740，将搭载于骁龙8 Gen2，是首个和唯一支持全部HDR格式的移动GPU，支持光线追踪技术和游戏后处理加速器技术。在Notebookcheck的GFXBench 3.0 1080p曼哈顿离屏测试中，分数优于苹果A15。

图表：部分高端移动GPU（智能手机和笔记本电脑）测试分数

GPU型号	像素着色器个数	制程(nm)	性能评分	GFXBench 3.0 1080p曼哈顿离屏（fps）
Apple M2（PC）	8	5	100	483.7
Apple M1（PC）	8	5	70.7	345
Adreno 740		4	54.6	253
Apple A16	5	4	57.3	290.5
Apple A15	5	5	49.6	233
Adreno 730		4	44	206
ARM Mali-G710	10	4	49.25	238
Apple A14 Bionic	2	5	36.8	172.85
Adreno 660		7	28.8	134.27

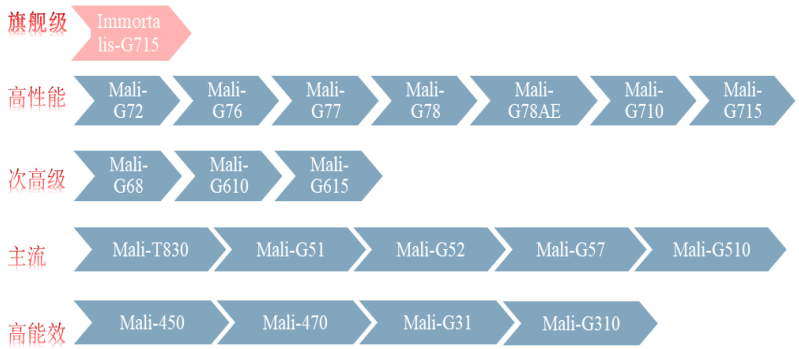
资料来源：Notebookcheck，中信建投

5.4 ARM—全球领先的半导体IP公司

- ARM是全球领先的半导体IP公司，成立于1990年。公司主要产品有CPU、GPU和NPU等处理器IP、安全性IP、系统性IP和相关软件及开发工具。公司通过IP授权向下游厂商收取许可费用和使用费用，客户包含芯片设计、芯片生产等电子行业所有重要公司。
- 公司GPU产品为Mali系列，使用场景有智能手机、平板电脑、笔记本电脑、可穿戴设备、VR/AR、自动驾驶汽车芯片等。据Strategy Analytics报告，ARM智能手机和平板电脑的GPU份额在2016年达到顶峰，2017年开始受苹果iPhone和iPad的GPU出货量增长等因素影响，ARM的GPU市场份额逐步下降到2020年的39%。据ARM官网，截止2022年6月，Mali GPU累计出货量超80亿颗，为全球出货量最高的移动GPU。

图： ARM整体计算解决方案包含的产品及服务

图：ARM mali GPU路线图

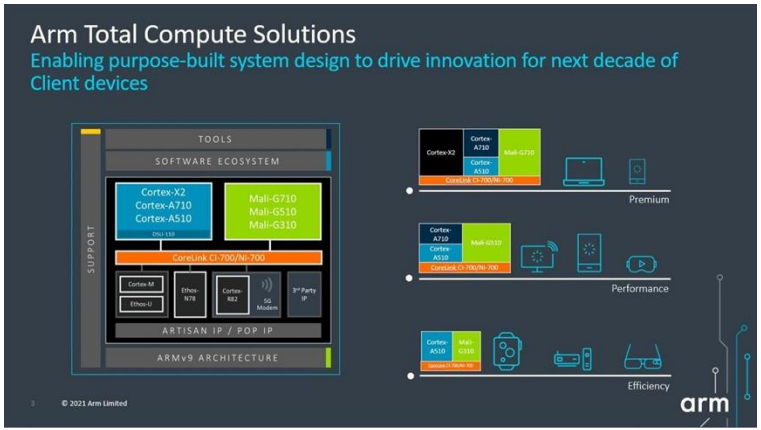


资料来源：Arm，中信建投

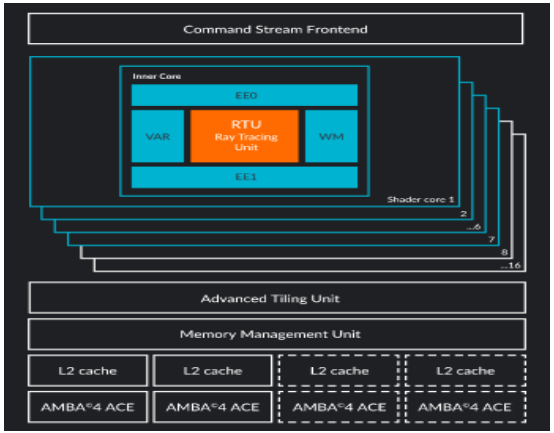
5.5 ARM GPU领跑安卓阵营

- 公司GPU架构为第四代Vallhall，Mali-G7系列定位高端市场，其中Mali-G710在FPS/W峰值和持续工作负载方面表现出色；新出的Mali-G6系列采取G7系列相同架构但使用更少的核心，Mali-G5和Mali-G3定位中端市场。
- 旗舰款Immortails-G715 GPU采用10个及以上内核，支持硬件级光线追踪技术，效能提升15%，机器学习能力强化两倍。Immortails-G715 GPU已搭配Cortex-X3 CPU搭载于联发科新款4nm级旗舰芯片天玑9200。在安兔兔的跑分中，天玑9200相比天玑9000性能提升25%，GPU性能提升32%，功耗降低41%，刷新安卓阵营历史新高；在更侧重GPU的GFX Bench测试中，Immortails-G715比苹果A16帧数高出26fps。

图：ARM整体SoC设计方法



图：Immortails-G715 GPU的架构

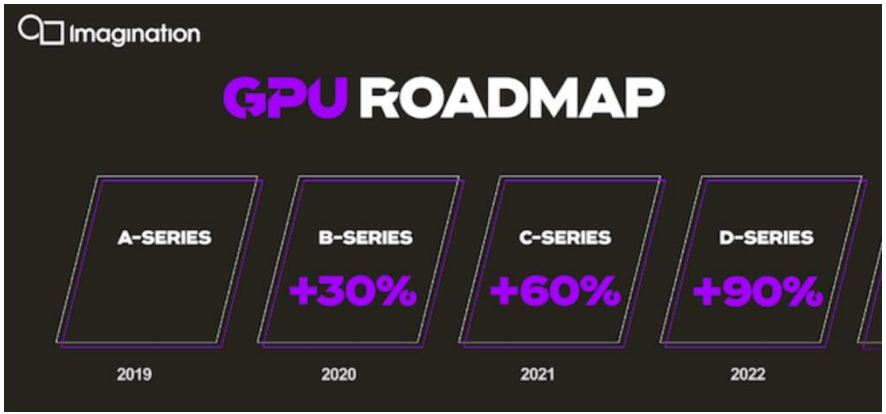


资料来源：Arm，中信建投

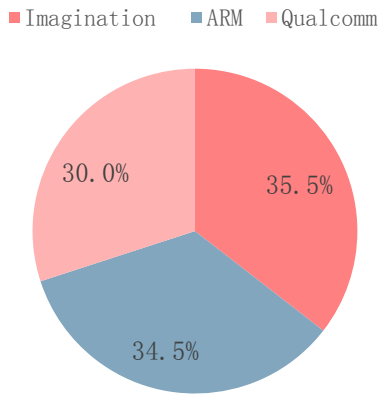
5.6 Imagination提供优秀的移动GPU芯片设计方案

- Imagination是移动GPU芯片设计的领军企业，成立于1985年。近些年，公司不断扩展产品领域，在CPU、人工智能芯片、以太网数据包处理器领域持续发力，产品覆盖汽车电子、AIot、桌面级应用、移动设备、机顶盒、服务器等诸多领域。
- 公司的PowerVR架构在移动芯片领域得到市场的广泛认可，为Intel、LG、德州仪器、三星、索尼、苹果、紫光展锐、海思等诸多公司提供授权服务。
- 面向移动设备，2019年开始公司陆续提出PowerVR的升级版本IMG A系列、IMG B系列、具备光线追踪能力的IMG CXT多层次产品。

图：Imagination产品路线图



图：2019年手机GPU IP市场占有率



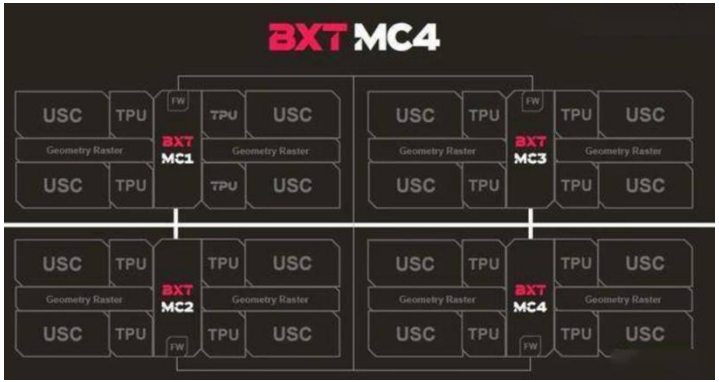
资料来源：Imagination，TSR，中信建投

5.7 IMG B系列采取去中心化的多核架构

- 2020年10月，公司发布IMG B系列高性能GPU IP，这款多核架构GPU IP包括BXE、BXM、BXT、BXS 4个系列33种配置，IMG BXE面向高清显示应用，IMG BXM主打图形处理体验，IMG BXT面向高性能应用，IMG BXS面向汽车应用。
- **IMG B系列采用去中心化的多核架构。**在一组内核中，采用主核、次核的拓展模式，其中一个作为主GPU带有一个控制固件处理器用来分割任务（渲染帧），并将这些渲染帧分割成不同的模块，其他的GPU就将这些分割的任务在自己的硬件上执行。可以利用其HyperLane（超线程）技术，进行多任务并行处理。
- 2021年11月，Imagination推出最新GPU产品IMG CXT实现了4级RTLS硬件光线追踪，首次在移动端实现了桌面级质量的光线追踪效果。

图：IMG B系列产品

图：BXT的多核架构



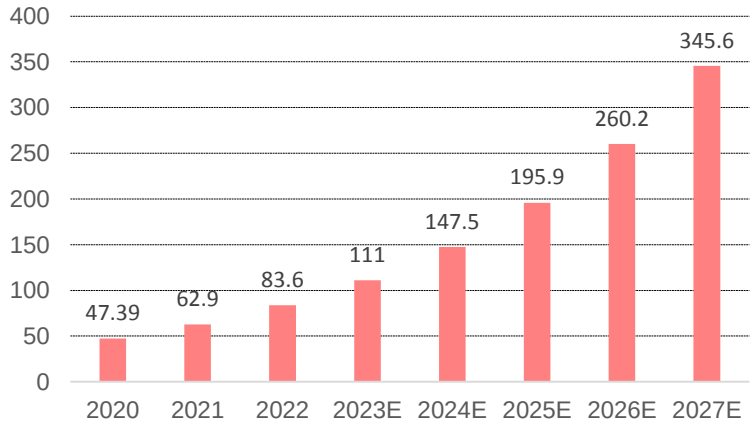
资料来源：Imagination，中信建投

第六章

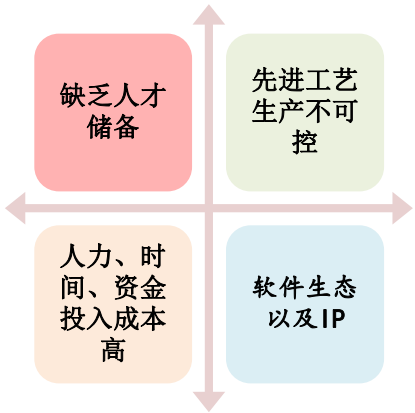
6.1 国内GPU市场空间广阔

- **国内市场空间广阔，PC、服务器拉动GPU需求。**根据Verified Market Research数据，2020年中国GPU市场规模为47.39亿美元，预计2023年中国GPU市场规模将达到111亿美元。中国数字化经济转型持续推进，催生大量对GPU的市场需求，给GPU带来广阔的市场空间。伴随着近期宏观经济回暖以及国内互联网企业纷纷加大AI算力布局，PC和服务器的需求上升有望为国内GPU市场带来整体拉动效应。
- **GPU的国产替代过程中也需要克服诸多困难，例如：软件生态以及IP、先进工艺的生产不可控，缺乏人才储备，人力、时间、资金投入成本较高等。**

图：中国GPU市场规模（亿美元）



图： GPU国产替代过程中需要克服的困难



资料来源：Verified Market Research，中信建投

6.2 海光信息提供服务器、工作站中的高端处理器芯片

- 海光信息成立于2014年，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。产品包括海光通用处理器(CPU)和海光协处理器(DCU)，目前已经研发出多款新能达到国际同类主流产品的高端CPU和DCU产品。2018年10月，公司启动深算一号DCU产品设计，目前海光DCU系列深算一号已经实现商业化应用，2020年1月，公司启动了第二代DCU深算二号的产品研发工作。

图：公司产品矩阵

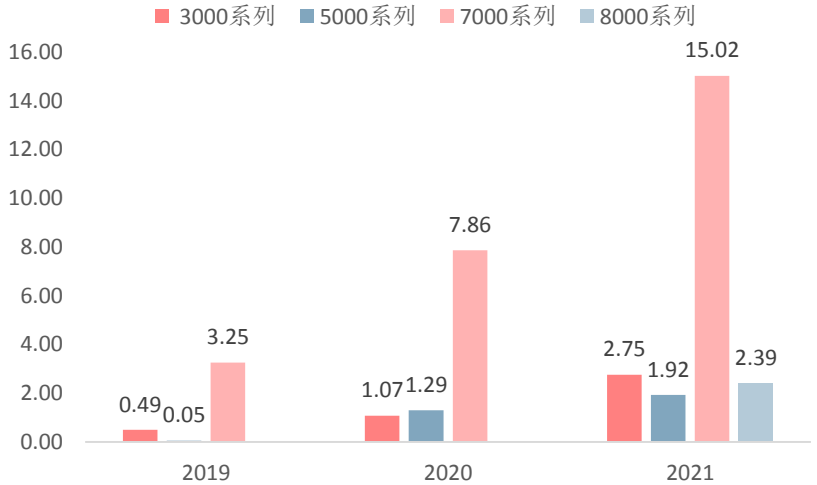
海光3200 CPU

海光5200 CPU

海光7200 CPU

海光8100 DCU

图：公司营收细分



资料来源：海光信息招股说明书，中信建投

中信建投证券 CHINA SECURITIES

6.3 海光DCU详解

- **海光DCU属于GPGPU的一种，海光DCU的构成与CPU类似，其结构逻辑相CPU简单，但计算单元数量较多。**海光DCU的主要功能模块包括计算单元(CU)、片上网络、高速缓存、各类接口控制器等。
- **深度计算处理器(Deep-learning Computing Unit，DCU)。**公司基于通用的GPGPU架构，设计、发布的适合计算密集型和运算加速领域的一类协处理器，定义为深度计算处理器DCU。兼容通用的“类 CUDA”环境以及国际主流商业计算软件和人工智能软件，软硬件生态丰富，可广泛应用于大数据处理、人工智能、商业计算等应用领域。

图：DCU架构示意图

图表：海光DCU 8100性能指标

	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	60-64个计算单元（最多4096个计算核心） 支持FP64、 FP32、 FP16、 INT8、 INT4
内存	4个HBM2内存通道、最高内存带宽为1TB/s 、最大内存容量为32GB
I/O	16 Lane PCIe Gen4 、DCU芯片之间高速互连

资料来源：海光信息招股说明书，中信建投

6.4 海光信息DCU提供高性能算力

- 海光8100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，在国内处于领先地位。2021年下半年DCU正式实现商业化应用，当年贡献2.38亿营收，该业务毛利率为34.84%，产品平均单价为19285元。

图表：深算一号与主流GPGPU性能比对

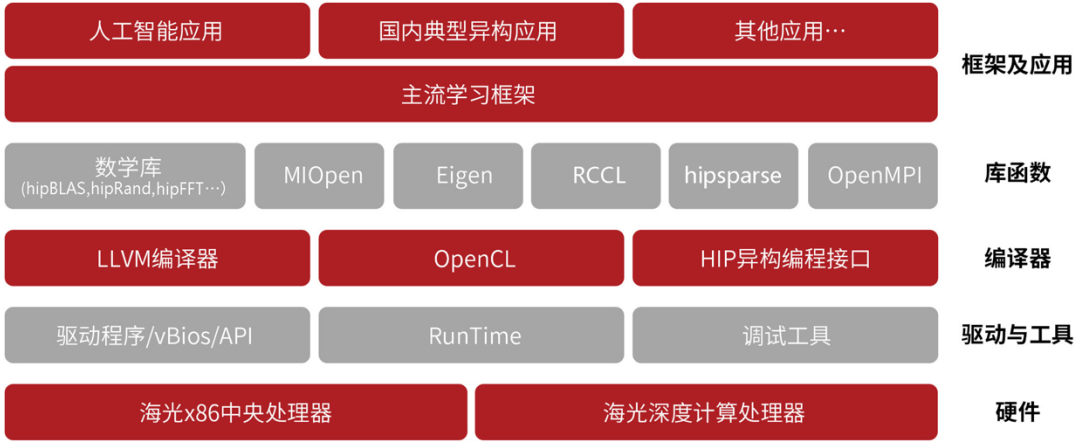
项目	海光信息	NVIDIA	AMD
品牌	深算一号	Ampere 100	MI100
生产工艺	7nm FinFET	7nm FinFET	7nm FinFET
核心数量	4096 (64 CUs)	2560 CUDA processors 640 Tensor processors	120 CUs
内核频率	Up to 1.5GHz (FP64) Up to 1.7Ghz (FP32)	Up to 1.53Ghz	Up to 1.5GHz (FP64) Up to 1.7Ghz (FP32)
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存位宽	4096 bit	5120 bit	4096 bit
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350 W	400 W	300 W
CPU to GPU互联	PCIe Gen4 x 16	PCIe Gen4 x 16	PCIe Gen4 x 16
GPU to GPU互联	xGMI x 2, Up to 184 GB/s	NVLink up to 600 GB/s	Infinity Fabric x3, up to 276 GB/s

资料来源：海光信息招股说明书，中信建投

6.5 ROCm GPU计算生态

- **海光信息DCU协处理器全面兼容ROCm GPU计算生态，由于ROCm和CUDA在生态、编程环境等方面具有高度的相似性，CUDA用户可以以较低代价快速迁移至ROCm平台，因此ROCm也被称为“类CUDA”。**因此，海光DCU协处理器能够较好地适配、适应国际主流商业计算软件和人工智能软件，软硬件生态丰富，可广泛应用于大数据处理、人工智能、商业计算等计算密集类应用领域，主要部署在服务器集群或数据中心，为应用程序提供高性能、高能效比的算力，支撑高复杂度和高吞吐量的数据处理任务。

图：ROCm GPU计算生态

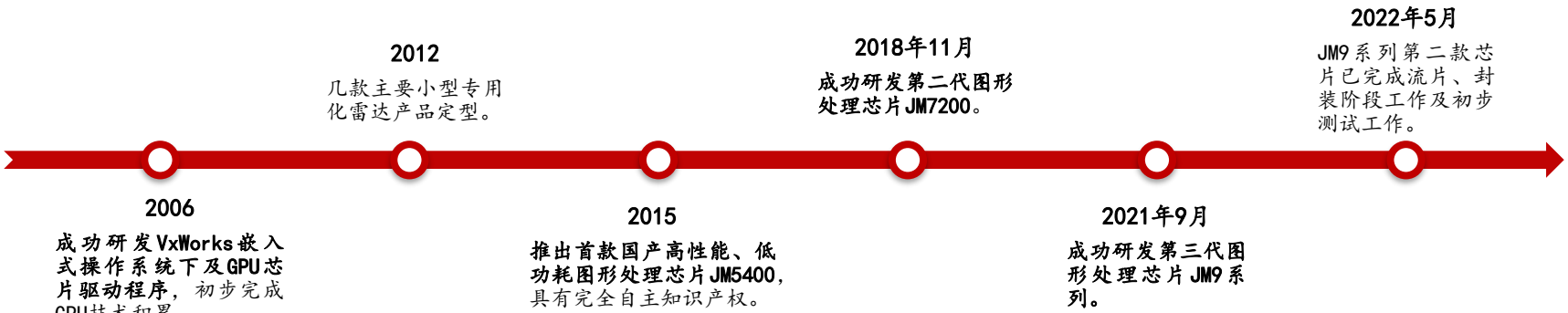


资料来源：海光信息招股说明书，公司官网，中信建投

6.6 景嘉微简介

- **长沙景嘉微电子股份有限公司成立于2006年，2015年推出首款国产GPU，是国内首家成功研制具有完全自主知识产权的GPU芯片并实现工程应用的企业**，2016年在深交创业板成功上市。公司业务布局图形显示、图形处理芯片和小型专用化雷达领域，产品涵盖集成电路设计、图形图像处理、计算与存储产品、小型雷达系统等方向。
- **公司GPU研发历史悠久，技术积淀深厚。**公司成立之初承接神舟八号图形加速任务，为图形处理器设计打下坚实基础；公司2007年自主研发成功VxWorks嵌入式操作系统下M9芯片驱动程序，并解决了该系统下的3D图形处理难题和汉字显示瓶颈，具备了从底层上驾驭图形显控产品的能力。2015年具有完全自主知识产权的GPU芯片JM5400问世，具备高性能、低功耗的特点；此后公司不断缩短研发周期，JM7200在设计和性能上有较大进步，由专用市场走向通用市场；JM9系列定位中高端市场，是一款能满足高端显示和计算需求的的通用型芯片。

图：景嘉微发展历史

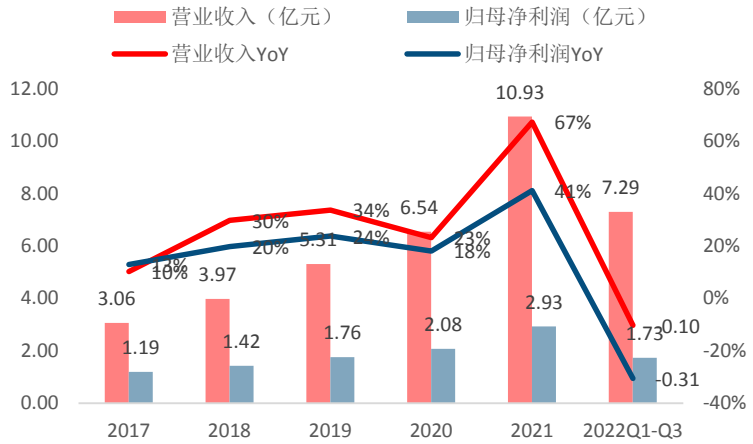


资料来源：景嘉微招股说明书，wind，中信建投

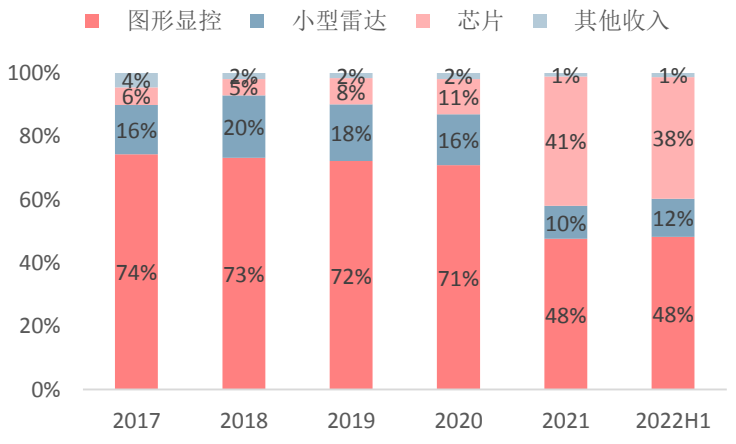
6.7 公司芯片业务展现良好增长势头

- 2022Q1-Q3，公司实现营收7.29亿元，同比下滑10.35%；归母净利润1.73亿元，同比下滑30.60%。
- 近年来，公司收入保持高速增长，受行业景气度旺盛和国产替代加速影响，分别在JM5400和JM9231研发成功时，公司营收增速均实现较大增长。
- 分领域来看，图形显控领域产品销售收入为公司核心收入，2021年芯片业务的快速发展，芯片收入占比提升到38%。2022H1，图形显控领域产品销售收入2.63亿元，芯片业务收入2.09亿元。

图：公司营业收入、净利润及增速



图：公司营收占比情况

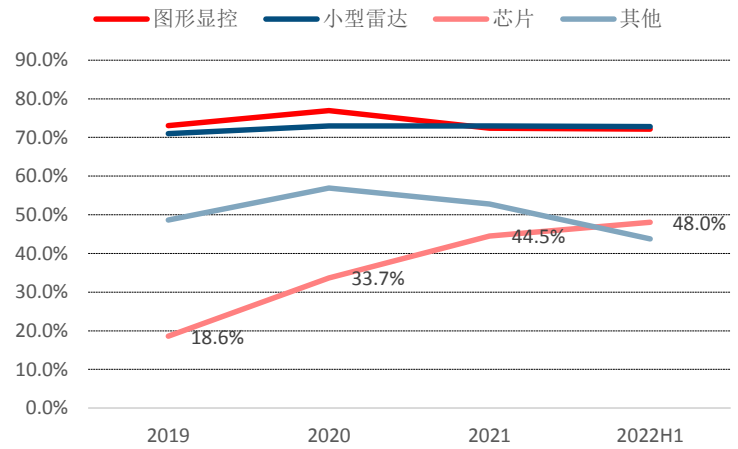


资料来源：wind，中信建投

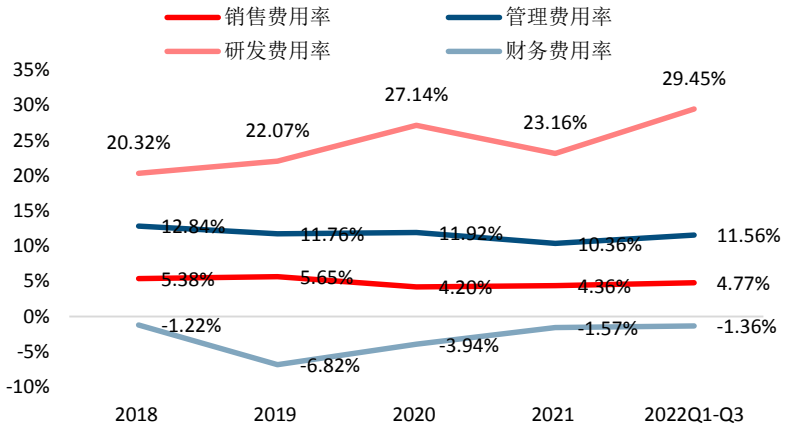
6.8 芯片业务盈利能力不断提升

- **公司芯片领域产品2022H1毛利率48.04%，实现快速增长。**由于公司采购芯片原材料的规模化效应、工艺控制水平的提高降低了芯片产品成本，导致毛利率上升。
- **公司坚持自主研发，研发投入不断加大。**2022Q3公司整体营收有所下滑的背景下，研发费用为8027万元，同比增长51.50%，前三季度合计研发费用2.15亿，研发费用率为29.45%。公司研发人员占比不断提高，2022H1公司有研发人员865名，占比达69.26%。
- 公司管理费用、销售费用和财务费用相对平稳，2022年前三季度分别为11.56%、4.77%和-1.36%。

图：分业务毛利率水平



图：公司期间费用率情况



资料来源：wind，中信建投

6.9 公司GPU性能优越

- JM7200采用28nm CMOS工艺，内核时钟频率最大1300MHz，存储器内存为4GB，支持OpenGL1.5/2.0，能够高效完成2D、3D图形加速功能，支持PCIe2.0主机接口，适配国产CPU和国产操作系统平台，可应用于个人办公电脑显示系统以及高可靠性嵌入式显示系统。
- JM9系列面向中高端通用市场，可以满足地理信息系统、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。2022年5月，JM9系列第二款芯片已完成初步测试工作。

图表：景嘉微9系产品与英伟达GTX系列性能对比

	JM9系列型号一	JM9系列型号二	GTX 1050	GTX 1080
内核性能	1 GHz （支持动态调频）	1.5 GHz （支持动态调频）	1.455GHz	1.6GHz
显存带宽	25.6GB/S	128GB/S	112GB/S	320GB/S
显存容量	8GB	8GB	2GB	8GB
视频解码	H.265/4K	H.265/4K	H.265/4K	H.265/4K
总线接口	PCIe 4.0 X8	PCIe 4.0 X8	PCIe 3.0	PCIe 3.0 X16
像素填充率	8 GPixels/s	32 GPixels/s	46.56G Pixel/s	128G Pixel/s
FP32运算性能	512 GFlops	1.5 TFlops	1.862 TFlops	9 TFlops
输出接口	HDMI 2.0	HDMI 2.0	HDMI 2.0，DisplayPort1.3	HDMI 2.0，DisplayPort1.4
支持平台：支持X86、ARM、MIPS处理器和Linux、中标麒麟、银河麒麟、统信软件等操作系统				

资料来源：wind，中信建投

6.10 国产GPU性能横向比较

- 国产GPU的典型厂商包括景嘉微、芯动科技、摩尔线程等。

图表：国产厂商渲染GPU典型产品性能比对

厂商	英伟达	英伟达	景嘉微	芯动科技	芯动科技	摩尔线程
型号	GeForce RTX 4090	GTX1080	JM9系列	风华一号	风华二号	MTT S80
制程	4nm	16nm	NA	12nm	NA	NA
核心数目	16384	2560	NA	NA	NA	4096个MUSA
时钟频率	2.23-2.52GHz	1.61-1.73GHz	1.5GHz	NA	NA	1.8GHz
显存容量	24GB	8GB	8GB	4GB / 8GB / 16GB	2/4/8GB	16GB
显存类型	GDDR6X	GDDR5X	NA	GDDR6 / GDDR6X	NA	GDDR6
FP32 运算性能	82.58 TFLOPS	8.873 TFLOPS	1.5 TFlops	5 TFLOPS/10 TFlops	1.5 TFLOPS	14.4 TFLOPS
总线接口	PCIe 4.0 x16	PCIE 3.0 X16	PCIE 4.0 X8	PCIe 4.0 x16	PCIe 3.0 x8	PCIe Gen5 x16

资料来源：各公司官网，中信建投

6.11 国产GPGPU性能横向比较

- 国产GPGPU的典型厂商包括海光信息、摩尔线程、壁仞科技、天数智芯等。

图表：国产厂商GPGPU典型产品性能比对

厂商	英伟达	海光信息	摩尔线程	壁仞科技	天数智芯
型号	A100	深算一号	MTT S3000	壁砺100P	天垓100
制程	7nm	7nm FinFET	NA	7nm	7nm
核心数目	6912	4096	4096	NA	NA
时钟频率	0.77-1.41GHz	1.5-1.7GHz	1.9GHz	NA	NA
显存容量	40GB/80GB	32GB	32GB	64GB	32GB
显存类型	HBM2e	HBM2	GDDR6	HBM2E	DRAM HBM2
FP32 运算性能	19.5 TFLOPS	NA	15.2 TFLOPS	240 TFLOPS（峰值）	37 TFLOPS
总线接口	PCIe 4.0 x16	PCIe Gen4 x 16	PCIe Gen5 x16	PCIe 5.0 X16	PCIe Gen4.0 x 16
TDP	250W	350W	≤35W	450-550W	250W

资料来源：各公司官网，中信建投

风险提示

- **个人电脑出货不及预期。**个人电脑出货受宏观经济影响比较大，个人电脑出货不及预期可能对PC端显卡出货造成影响。
- **AI技术进展不及预期。**当前AI技术的快速进步带动了巨大的AI算力需求，如果AI技术进展不及预期，可能对GPU市场的整体需求产生不利影响。
- **互联网厂商资本开支不及预期。**互联网厂商是AI算力和GPGPU的重要采购方和使用方，如果互联网厂商资本开支不及预期，可能会对GPGPU的需求情况产生不利影响。
- **自动驾驶进展不及预期。**GPU在高等级的自动驾驶中渗透率相对较高，如果自动驾驶技术进步不及预期，可能会对GPU在自动驾驶中的应用产生不利影响。
- **国产替代进程不及预期。**GPU的国产替代过程中面临诸多困难，国产替代进程可能不及预期。
- **参与厂商众多导致竞争格局恶化。**在GPU需求旺盛的背景下，国内外涌现出诸多GPU行业的新兴玩家，众多参与厂商可能导致整体竞争格局恶化。

感谢樊文辉、庞佳军对本报告的贡献。

分析师介绍

阎贵成：中信建投证券通信&计算机行业首席分析师，北京大学学士、硕士，专注于云计算、物联网、信息安全、信创与5G等领域研究。近8年中国移动工作经验，6年多证券研究经验。系2019-2021年《新财富》、《水晶球》通信行业最佳分析师第一名，2017-2018年《新财富》、《水晶球》通信行业最佳分析师第一名团队核心成员。

金戈：中信建投证券研究发展部计算机行业联席首席分析师，帝国理工学院工科硕士，擅长云计算、金融科技、人工智能等领域。

于芳博：中信建投人工智能组首席分析师，北京大学空间物理学学士、硕士，2019年7月加入中信建投，主要覆盖人工智能等方向，下游重点包括智能汽车、CPU/GPU/FPGA/ASIC、EDA和工业软件等方向。

评级说明

投资评级标准		评级	说明
报告中投资建议涉及的评级标准为报告发布日后6个月内的相对市场表现，也即报告发布日后的6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数作为基准；新三板市场以三板成指为基准；香港市场以恒生指数作为基准；美国市场以标普500 指数为基准。	股票评级	买入	相对涨幅15%以上
		增持	相对涨幅5%—15%
		中性	相对涨幅-5%—5%之间
		减持	相对跌幅5%—15%
		卖出	相对跌幅15%以上
	行业评级	强于大市	相对涨幅10%以上
		中性	相对涨幅-10-10%之间
		弱于大市	相对跌幅10%以上

分析师声明

本报告署名分析师在此声明：（i）以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，结论不受任何第三方的授意或影响。（ii）本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

法律主体说明

本报告由中信建投证券股份有限公司及/或其附属机构（以下合称“中信建投”）制作，由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。中信建投证券股份有限公司具有中国证监会许可的投资咨询业务资格，本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页。

在遵守适用的法律法规情况下，本报告亦可能由中信建投（国际）证券有限公司在香港提供。本报告作者所持香港证监会牌照的中央编号已披露在报告首页。

一般性声明

中信建投证券研究发展部

北京
东城区朝内大街2号凯恒中心B座12层
电话：（8610）8513-0588
联系人：李祉瑶
邮箱：lizhiyao@csc.com.cn

上海
浦东新区浦东南路528号南塔2106室
电话：（8621）6882-1612
联系人：翁起帆
邮箱：wengqifan@csc.com.cn

深圳
福田区益田路6003号荣超商务中心B座22层
电话：（86755）8252-1369
联系人：曹莹
邮箱：caoying@csc.com.cn

中信建投（国际）

香港
中环交易广场2期18楼
电话：（852）3465-5600
联系人：刘泓麟
邮箱：charleneliu@csci.hk

中信建投证券 CHINA SECURITIES