中信证券 CITIC SECURITIES

计算机行业“构筑中国科技基石”系列报告25

GPU：研究框架（100页）

中信证券研究部　计算机团队
杨泽原，丁奇

2023年2月13日

核心观点

- **核心结论：GPU的核心竞争力在于架构等因素决定的性能先进性和计算生态壁垒。国内GPU厂商纷纷大力投入研发快速迭代架构，推动产业开放构建自主生态，加速追赶全球头部企业。国产替代需求持续释放叠加国际局势不确定性加剧，AI&数据中心、智能汽车、游戏等GPU需求有望高增，国产GPU迎来发展黄金期，我们看好国产GPU公司的发展与投资机遇。**

- **理解GPU的核心：性能先进性+生态计算壁垒**

  - **GPU物理性能取决于微架构、制程、流处理器数量、核心频率等，其中微架构是核心点。**我们认为微架构的快速创新迭代是GPU性能领先的前提，其图形渲染单元和通用计算单元设计向着“更多、更专、更智能”的方向优化迭代。根据应用场景来划分，数据中心要求强算力、高并发吞吐量；游戏业务要求浮点运算能力强、访存速度快；图形显示要求图显专业化、精细化等。

  - **生态：GPU生态构筑通用计算极深壁垒，CUDA生态占据大部分市场，类CUDA生态蓬勃发展。**GPU生态由上层算法库，中层接口、驱动、编译器和底层硬件架构三大部分基本构成。GPU研发难度在图形渲染硬件层面和通用计算软件生态层面，在IP、软件栈方面研发门槛较高，需要较长的积累，先发优势明显。CUDA生态从2006年推出至今，经过不断发展完善，几乎已在行业生态内处于垄断地位，目前ROCm等兼容Cuda的类计算生态蓬勃发展并处于快速推广阶段。

- **海外复盘：NVIDIA与AMD（ATI）的竞争贯穿GPU发展历程，架构创新升级和新兴AI等领域前瞻探索是领跑的关键**

  - **NVIDIA长期居于GPU市场领导地位，近年AMD凭借RDNA架构在游戏市场强势崛起。**Verified Market Research数据显示，2022年全球独立GPU市场规模约448.3亿美元，NVIDIA和AMD的市场份额占比约为8:2。根据JPR数据，NVIDIA凭借自身性能领先和CUDA生态优势性始终占有GPU领域超50%的市场份额，数据中心业务更是全面领先，在游戏显卡领域，近年AMD凭借RDNA系列架构强势崛起。

  - **NVIDIA先后与AMD等企业在性能方面竞争博弈，架构创新升级和新兴领域前瞻探索是领跑GPU行业的关键。**NVIDIA凭借性能领先长期占据超五成市场份额，AMD（ATI）也曾因架构出色、性能惊艳实现反超。同时NVIDIA早在2006年前瞻性布局通用计算、构建CUDA生态，为如今AI&数据中心领域的全面领先构筑牢固的壁垒。NVIDIA积极布局异构芯片、汽车、元宇宙等新市场，寻找新的强有力业务增长点。

核心观点

- **国内GPU市场：各应用场景市场广阔，国内厂商大有可为**
  - **需求端1—AI：数据中心和终端场景不断落地对计算芯片提出更多更高需求。**新一轮AI对算力需求远超以往：ChatGPT类语言大模型底层是2017年出现的Transformer架构，该架构相比传统的CNN/RNN为基础的AI模型，参数量达到数千亿，对算力消耗巨大，对算力硬件有大量需求。甲子光年预测，中国AI芯片市场规模2023年达到557亿元。AI芯片可进一步细分为云端和终端，中国云端芯片市场规模较大，甲子光年预计2023年增长至384.6亿元，对应复合年增速到52.8%；终端芯片市场规模甲子光年预计2023年增长至173亿元，对应年复合增长率达62.2%，伴随各AI终端落地预计将保持较快增长速度。
  - **需求端2—汽车：汽车智能化浪潮下域控制器GPU市场前景广阔。** 自动驾驶和智能座舱是智能汽车中具有广阔前景的方向。盖世汽车数据预计，2025年自动驾驶域控制器出货量将达到432万台，每台自动驾驶域控制器配备1-4片高性能计算GPU；智能座舱域控制器出货量达到528万台，绝大多数智能座舱域控制器配备1片GPU。自动驾驶技术不断提高和座舱进一步智能化拉动汽车GPU市场规模快速扩张。
  - **需求端3—游戏：游戏玩家人数持续增长，游戏GPU市场稳中有升。**Newzoo Expert数据显示全球游戏玩家人数在2021年已达到30.57亿人，且预计2020-2025年全球游戏玩家人数复合年增率为4.2%；游戏市场内，游戏机和PC两大主体出货量再创新高，游戏机三大巨头2021年出货量高达4008万台；2021年Q4全球PC GPU出货量（包括集成和独立显卡）高达11000万片。
- **投资建议：**
  - **产业逻辑：**GPU的核心竞争力在于架构先进性能和生态丰富性，国产厂商正持续大力投入研发实现GPU架构创新升级和快速迭代，力争赶超国际领先水平；同时构建与主流适配良好的生态环境，打造自主开放的软硬件生态和信息产业体系。
  - **投资建议：外部不确定因素叠加内部加速自主创新背景下，国产GPU厂商有望加速崛起。伴随政策大力扶持、国际科技贸易政策影响、国产厂商产品性能提升及生态逐步完善，国产GPU龙头正迎来关键发展机遇。1）推荐：海光信息（CPU+GPGPU）。建议关注景嘉微、寒武纪（电子覆盖）。2）一级市场（排名不分先后）：关注壁仞科技、摩尔线程、沐曦、天数智芯、登临科技、燧原科技等。**
- **风险因素：产业链安全风险；市场竞争加剧风险；商业需求不及预期风险；产品研发不及预期风险；国产替代进程不及预期风险；宏观经济环境风险。**

报告亮点与创新之处

- 第一，**我们从性能和生态2个维度构建了GPU完整的研究体系。**1）**性能：**决定GPU是否“高效”，其中微架构/制程是影响GPU性能的核心要素。2）**生态：**CUDA构筑通用计算坚固壁垒。
- 第二，**提出在评估GPU性能的指标的重要性上：微架构、制程、流处理器数量、核心频率对GPU性能影响较大。**我们详细梳理了GPU的微架构、制程、显存容量/位宽/带宽/频率、核心频率等各类性能参数及重要性程度，并利用“核心数*核心频率*2”公式对性能算力进行量化，揭示可用3DMark、MLPerf 等GPU软件跑分进行相关性能测试评估。
- 第三，**详细拆解了NVIDIA Fermi和Hopper两大典型微架构的具体硬件实现，**在顶点处理、光栅化计算、纹理贴图、像素处理的图形渲染流水线上对Fermi架构进行了拆分；在指令接收、调度、分配、计算执行的通用计算流水线上对Hopper架构进行了简单易懂的描述，并指明更多、更专、更智能等未来架构升级迭代的方向。
- 第四，**明晰了生态是构建通用计算壁垒的基石。**提出GPU研发难度在图形渲染硬件和通用计算软件生态层面，在IP、软件栈方面研发门槛较高，需要较长的积累，先发者优势明显。CUDA生态从2006年推出至今，经过不断发展完善，几乎已在行业生态内处于垄断地位。
- 第五，**深度复盘Nvidia/AMD（ATI）的产品迭代和竞争发展史，**通过对NVIDIA长期保持领先和AMD（ATI）反超进行总结得出结论：**架构创新升级和新兴领域前瞻探索是领跑GPU行业的关键。**
- 第六，梳理和测算了国内GPU在AI&数据中心、智能汽车、游戏行业的**市场空间和发展趋势。**

CONTENTS

目录

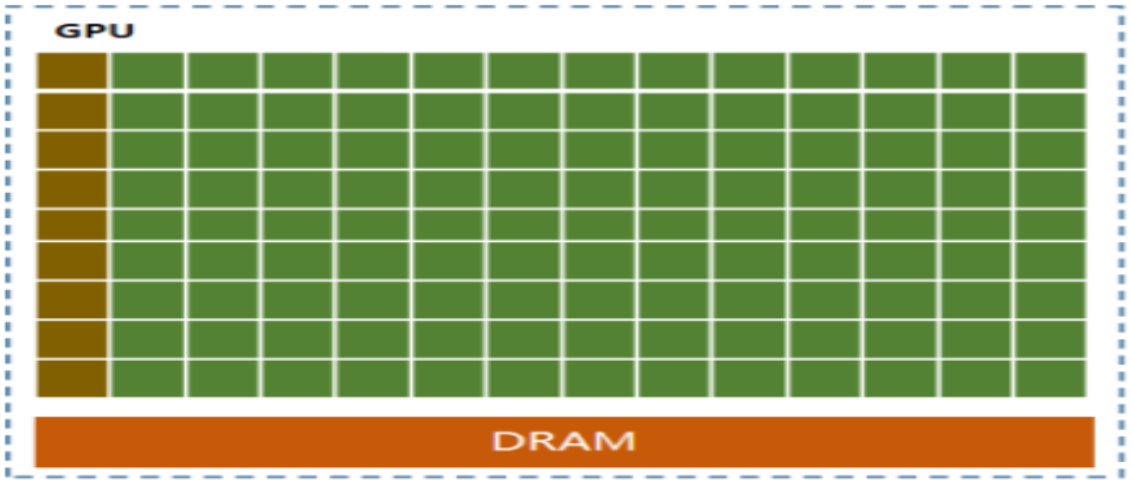
1.理解GPU的核心：性能+生态

I. GPU：计算机图形处理以及并行计算的核心
II. 性能：决定GPU是否“高效”，其中微架构是GPU性能领先的关键
III. 生态：构筑通用计算壁垒

1.1 GPU定位：计算机图形处理以及并行计算的核心

- **GPU全称**是Graphic Processing Unit，即图形处理单元，是计算机显卡的核心。
- **GPU是计算机的图形处理以及并行计算内核。**
  - 它的主要功能可以分为：1）图形图像渲染计算 GPU；2）作为运算协作处理器 GPGPU。
  - GPU的功能主要集中于执行高度线程化、相对简单的并行任务处理。
- **GPU vs GPGPU：**
  - GPGPU全称通用GPU，运用CUDA及对应开放标准的OpenCL实现通用计算功能运算，能够辅助CPU进行非图形相关程序执行。
  - 由GPU性能拓展至计算密集领域，将GPU强大的并行运算能力运用于通用计算领域。多侧重科学计算、AI领域、大数据处理、通用计算、物理计算、加密货币生成等领域。

GPU内部架构



资料来源：NVIDIA官网

GPU与GPGPU对比

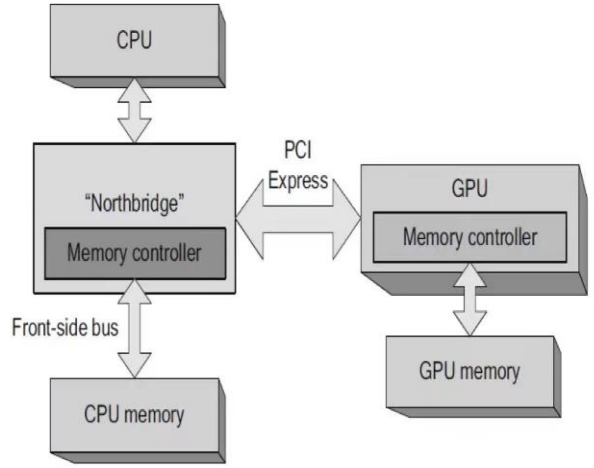
	GPU	GPGPU
主要执行任务	图形渲染	并行计算
功能	图形渲染、图形计算，对于游戏性能有关键影响	多进行AI领域相关计算，科学计算和通用计算
国内主要公司	景嘉微、摩尔线程、象帝先、芯动科技、格兰菲、励算、深流微、芯瞳、绘智微	壁仞、沐曦、登临、天数智芯、红山微电子、瀚博

资料来源：搜狐@十一号组织，中信证券研究部

1.1 GPU分类：应用于PC、服务器、移动端

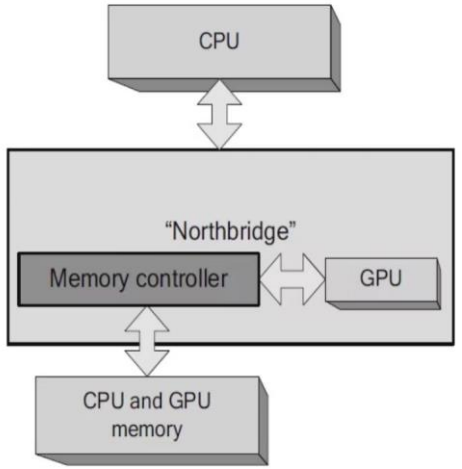
- **依据接入方式不同分为：独立GPU和集成GPU。**
  - 1）独立GPU：大部分封装于独立显卡电路板上，使用PCIE接口和特定显存，不受空间和供电限制，性能相对更好、渲染画质更佳。主要厂商包括AMD（Radeon系列）、NVIDIA（Geforce系列）。2）集成GPU：通常未拥有独立显存，集成于CPU内部，与CPU共同使用Die和系统内存，节省空间占位和制作难度，价格较低、兼容性更佳且供电量少。主要厂商包括Intel（HD系列）、AMD（APU系列）。
- **依据应用端不同分为：PC GPU、服务器GPU和移动GPU。**
  - 1）PC端：集成GPU主要运用于提高轻办公效率，对性能要求较低；独立GPU主要运用于图形设计、提高图片制作清晰度以及3A游戏绘图渲染能力，对性能要求较高。2）服务器端：主要进行专业可视化处理、AI训练、AI推断的深度学习、提高计算运行能力以及视频编解码等功能，以独立GPU为主。3）移动端：提高游戏体验、提升游戏处理性能，应用场景包括AR、桌面、云计算、数据中心等。受移动端功耗和体积限制，一般为集成GPU。

独立GPU



资料来源：微信公众号@GPU and Computing

集成GPU



资料来源：微信公众号@GPU and Computing

主要厂商及产品

	主要厂商	产品系列
PC GPU	NVIDIA、Intel、AMD	Xe LP、 TITAN V
服务器GPU	NVIDIA、AMD	Tesla、FireStream
移动GPU	Imagination、高通、苹果、ARM、三星、华为、联发科	PowerVR系列、Adreno系列、公版Mali系列、Exynos、麒麟

资料来源：NVIDIA官网，CSDN@Finovy Cloud ，中信证券研究部

1.1 GPU产业链：设计→制造→封装

- **GPU产业链主要包括三大环节：设计、制造和封装。**
- **GPU整体商业模式包括三种：IDM和、Fab+Fabless和 Foundry。**
  - IDM模式：指将GPU产业链的三个环节整体化，充分结合自主研发和外部代工，集设计、制造、封装为一体，公司垂直整合GPU整体产业链。
  - Fab+Fabless：充分发挥各企业比较优势，仅负责芯片电路设计，将产业链其他环节外包，分散了GPU研发和生产的风险。
  - Foundry：公司仅负责芯片制造环节，不负责上游设计和下游封装，可以同时为多家上游企业服务。

GPU产业链

资料来源：华经情报网，各公司官网，中信证券研究部

供给模式代表厂商

供给模式	代表国外厂商
IDM	英特尔、三星、TI
Fab+Fabless	NVIDIA、Apple、AMD、ARM、Qualcomm、华为、海思、MTK、Broadcom
Foundry	台积电、SMIC、UMC、Global Foundries

资料来源：IT智库，eefocus，中信证券研究部

1.2 GPU性能：衡量GPU“高效”的指标

- **性能是衡量GPU运行、执行命令高效的指标。**
- **GPU物理性能评估主要在于比较各硬件的物理参数。**
  - 评估GPU物理性能的参数主要包括：微架构、制程、图形处理器数量、流处理器数量、显存容量/位宽/带宽/频率、核心频率。
  - 我们认为，评估GPU性能的指标依次为：微架构/制程>流处理器数量/核心频率>显存带宽/容量>其他。

GPU性能参数

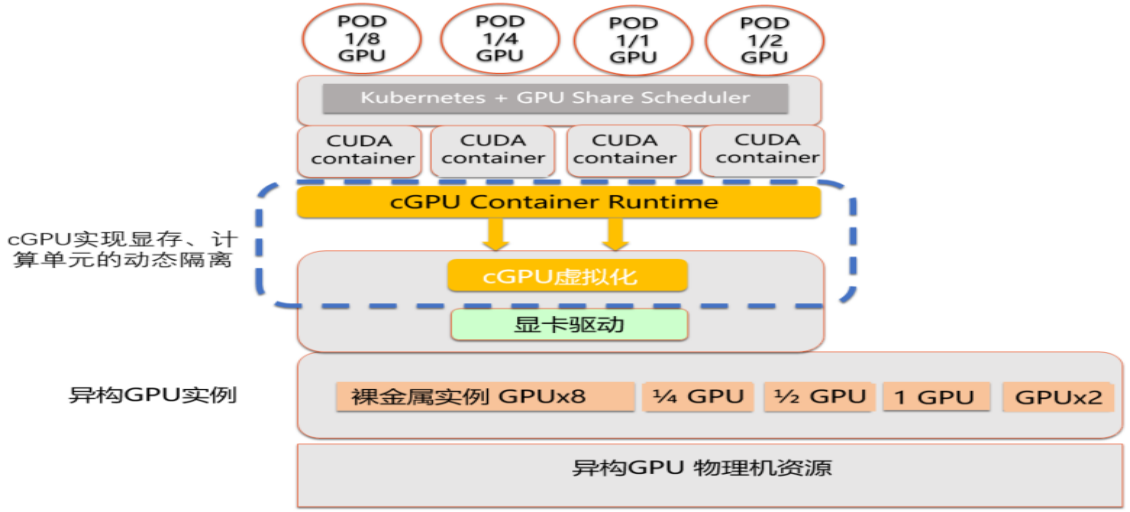
性能指标	含义
微架构	GPU的硬件电路设计构造方式
制程	GPU的制造工艺和设计规则，代表不同电路特性，通常以生产精度nm表示
图形处理器单元数量	包含了光栅单元ROP，纹理单元TMU的数量，数量越多可执行指令越多
CUDA核数	CUDA是执行函数的重要部件，CUDA核数越多，性能运行越好
Tensor核数	指张量处理单元的数量，Tensor Core核数越多，性能越好
核心频率	指显示核心的工作频率，能反映显示核心的性能优良
显存容量	显存容量越大，GPU能够处理的数据量越大
显存位宽	指显存在单位时钟周期内所传送数据的位数，位数越大瞬间传送数据量越大
显存带宽	等于显存频率×显存位宽/8，与显存频率、位宽成正比
显存频率	反映显存速度，以MHz为衡量单位，越高端的显存，频率越高

资料来源：CSDN@Charles Ren，NVIDIA官网，中信证券研究部

1.2 GPU性能影响因素：微架构、制程、核心频率

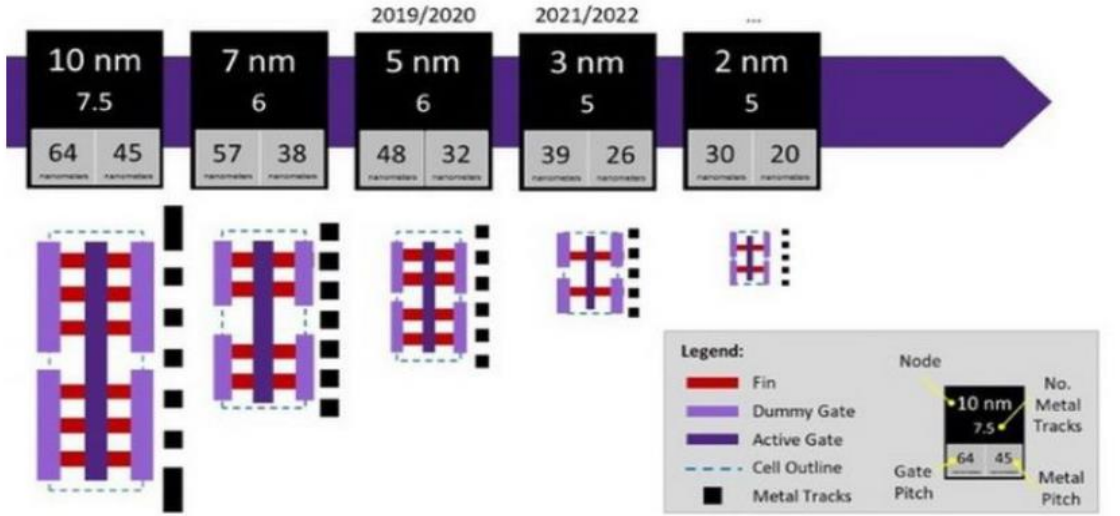
- **微架构：**又称为微处理器体系结构，是硬件电路结构，用以实现指令执行。
- **制程：**指GPU集成电路的密集度。在晶体管硬件数量一定的情况下，更精细的制程能够减少功耗和发热。现阶段GPU主流最先进工艺制程为5nm。
- **核心频率：**代表GPU显示核心处理图像频率大小/工作频率，能够反映显示核心的性能。

GPU微架构



资料来源：阿里云官网

GPU制程

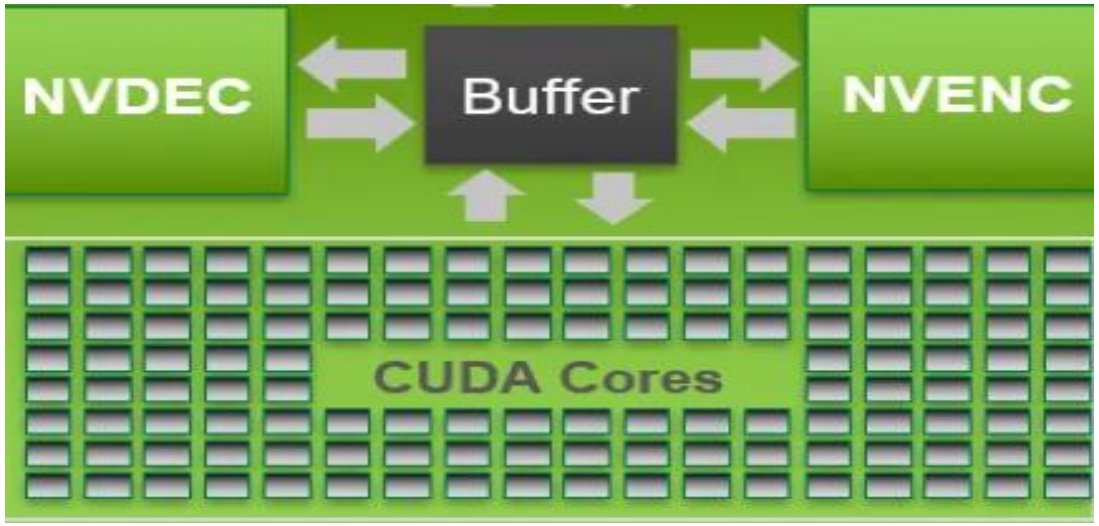


资料来源：半导体行业观察微信公众号

1.2 GPU性能影响因素：图形处理器单元数量、CUDA核数、Tensor核数

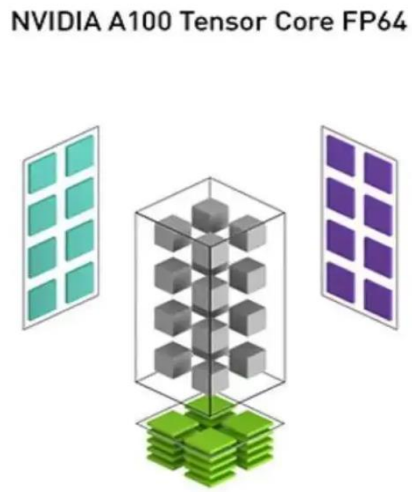
- **图形处理器单元数量：**指GPU内部图形处理单元，涵盖光栅单元（ROP）和纹理单元（TMU）等数量。
  - 光栅单元（ROP）：进行光线、反射计算，负责游戏中高分辨率、高画质的效果生成。
  - 纹理单元（TMU）：能够对二进制的图形进行一系列翻转、缩放变化，再将其纹理传输至3D平面模型中。
- **CUDA核数：**作为GPU内部的流处理器，是主要的计算单元，CUDA核数越多，GPU性能等级越高。
- **Tensor核数：**能够进行张量核加速GEMM计算以及加速卷积和递归神经网络运行，Tensor核数越多，在人工智能、深度学习领域的性能越强。

GPU CUDA Core



资料来源：SHERLOCK

GPU Tensor Core

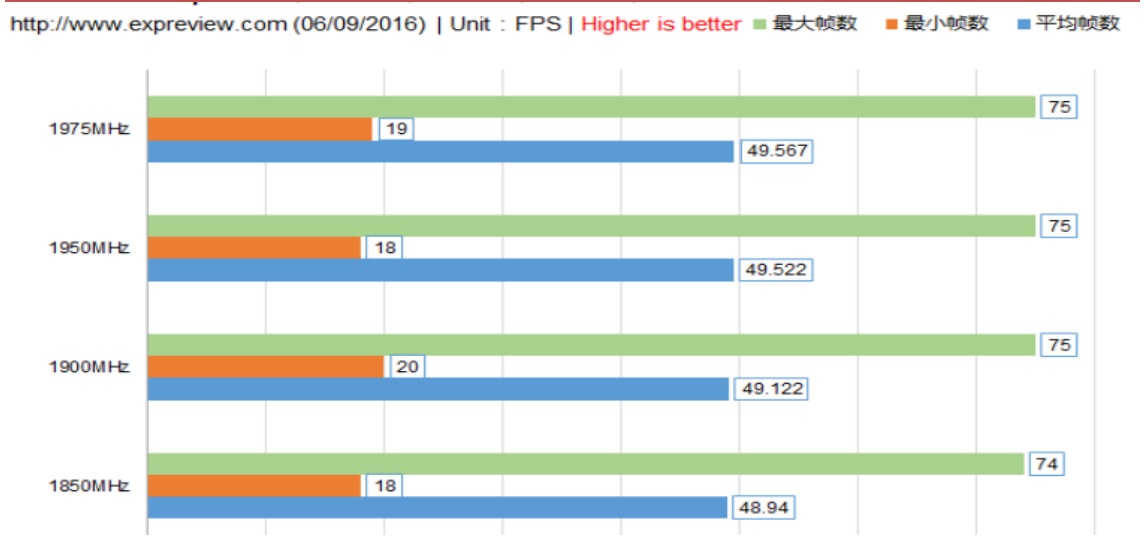


资料来源：NVIDIA A100 Tensor Core GPU Architecture白皮书

1.2 GPU性能影响因素：显存容量、显存位宽、显存频率、显存带宽

- **显存容量：显存作为GPU核心部件，用以临时存储未处理数据。**
  - 显存容量的大小对于GPU存储临时数据的多少起决定性作用，在GPU核心性能能够提供充足支撑前提下，越大的显存容量能够减少数据读取次数，减少延迟出现。
- **显存位宽：**是GPU在单位时钟周期内传送数据的最大位数，位数越大GPU的吞吐量越大。
- **显存频率：**显存数据传输的速度即显存工作频率，通常以MHz为显存频率计数单位。
- **显存带宽：**显存带宽=显存频率X显存位宽/8，为显存与显卡芯片间数据传输量。

显存频率



资料来源：EXPreview

显存带宽

资料来源：NVIDIA官网

1.2 微架构的先进性：GPU性能的抓手

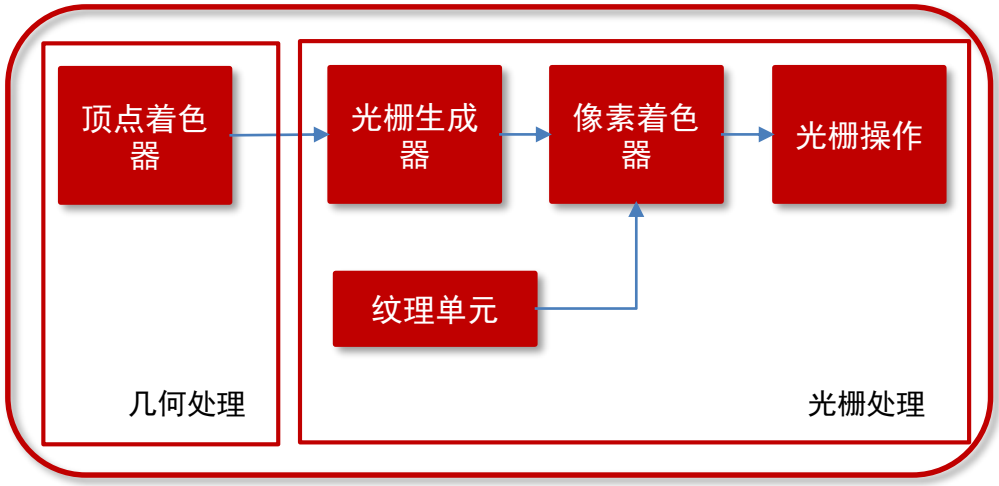
- **微架构（Micro Architecture）：GPU的硬件电路设计构造方式。**
  - 微架构又称为微处理器体系结构，是在图形函数和指令集条件下处理器中的执行方法。某一特定指令集可以在不同微架构中执行，但在运行过程中因设计目的不同而存在技术效果不同。
- **GPU微架构包括流处理器、渲染核、双精度浮点运算单元、特殊运算单元、流式多处理器、纹理处理器、图形处理器、流处理器阵列。**
  - GPU架构工作流程为：Vertex Shader（定点着色器）建立图形骨架，再通过算法转化进行光栅化计算，进而进行纹理映射，再由Pixel Shader（像素着色器）像素处理，最终由ROP（光栅化引擎）输出。
- 不同微架构决定了GPU各方面性能的不同，NVIDIA等国际GPU厂商均加大投入研发新架构作为**提升竞争力的重要抓手**。

微架构中各单元简介

名称	功能
流处理器（SP）	GPU最基本单元
渲染核（shader）	升级版本的流处理器，用于顶点处理、像素处理
双精度浮点运算单元（SFU）	仅用于双精度浮点运算
流式多处理器（SM）	基本计算单元，由SP、DP、SFU等构成
纹理处理器簇（TPC）	由SM控制器、多个SM和L1缓存构成
光栅化处理单元（ROPs）	对3D图形进行几何、设置、纹理和光栅处理
张量单元（Tensor Core）	专门用于矩阵乘积累加的高性能计算核心

资料来源：厦门大学@许少聪，中信证券研究部

微架构工作流程



资料来源：搜狐@爱玩客iVankr，中信证券研究部绘制

1.2 微架构的先进性：以 Fermi架构为例—总览

Fermi核心微架构

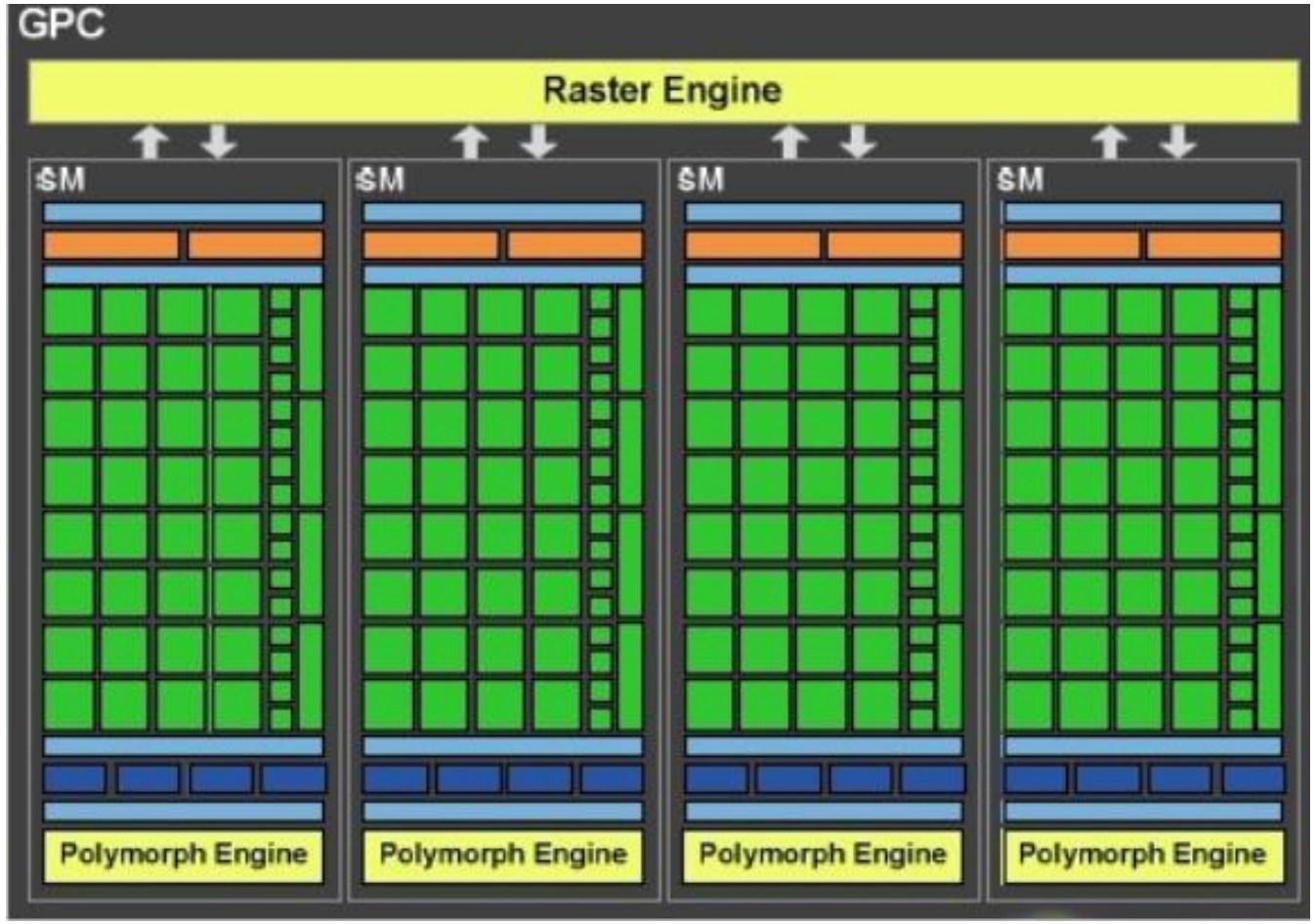
- **Fermi架构共含4个GPC，16个SM，512个CUDA Core。每32个CUDA Core组成1个SM，每个SM为垂直矩形条带。**

- **核心性能：**
  - 晶体管数高达30亿个，引入缓存单元，合计1MB
  - 可同时执行线程指令流24576个
  - 使用并行内核，全局分配逻辑支持与CPU并行传输

资料来源：NVIDIA官网

1.2 微架构的先进性：以 Fermi架构为例—GPC架构拆分

Fermi GPC 核心微架构



资料来源：中关村在线

- **GPC为图形处理团簇，是Fermi架构的组成核心，负责顶点、几何、光栅化、纹理和像素处理。组成部分包括：**
  - 1个光栅引擎Raster Engine（上部黄色部分）
  - 4个SM单元（矩形部分）
- **SM之间彼此独立，可各自调度多个Thread Wraps到内部的图形渲染、计算执行单元上运行。**

1.2 微架构的先进性：以 Fermi架构为例—SM架构拆分

资料来源：NVIDIA官网

- **SM全称Streaming Multiprocessor，Fermi架构下，每个SM具有32个 CUDA Core，组成部分包括：**
  - 2 个 Warp Scheduler/Dispatch Unit（橙色部分）
  - 分别位于两条 lane 上的32 个 CUDA Core（绿色部分）
  - 1个register file-寄存器文件和 L1 cache（浅蓝色部分）
  - 16 个 Load/Store units (LD/ST Unit)，支持各线程同时从Cache/DRAM存取数据
  - 4 个 Special Function Units (SFU)，用于计算sin/cos这类特殊指令

1.2 微架构的先进性：以 Fermi图形渲染流水线为例—指令接收

Fermi核心微架构

Host Interface

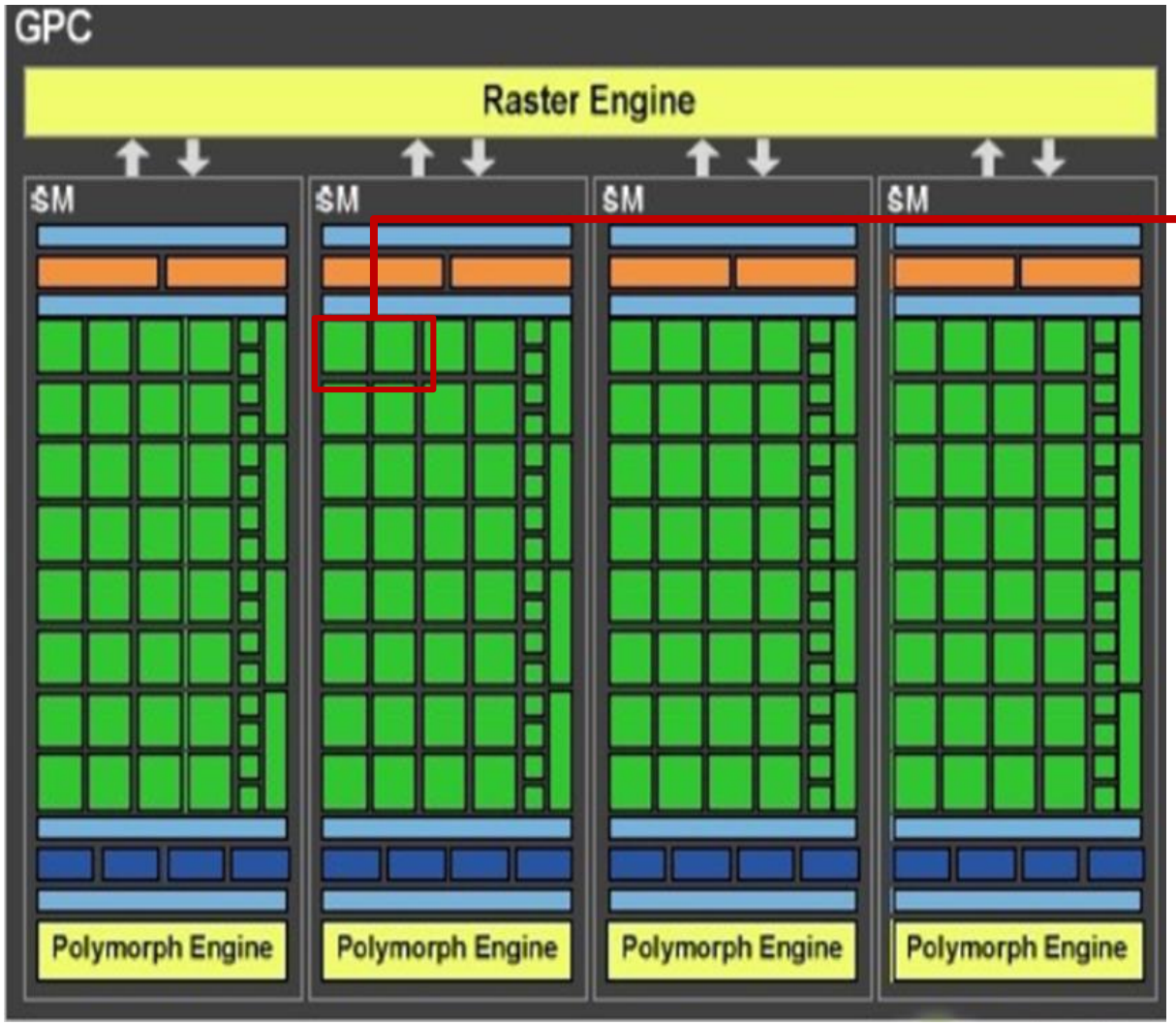
Giga Thread Engine

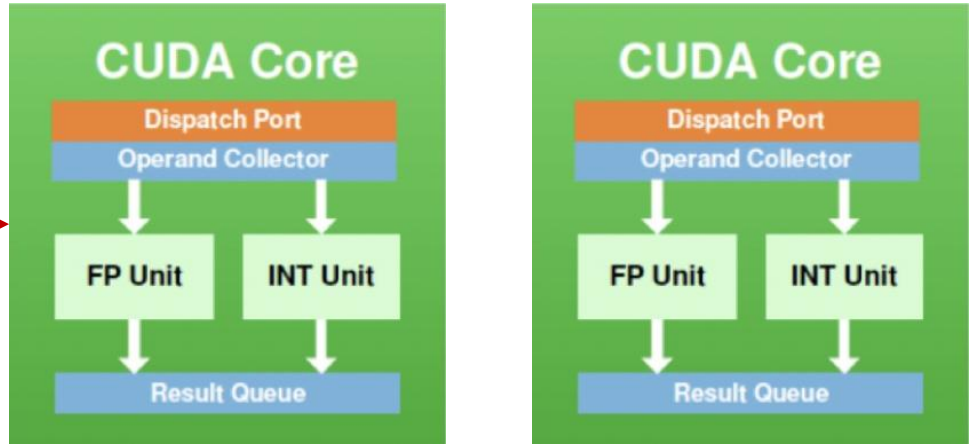
- **Host Interface（黑色部分）**
  - 为主机接口，图形渲染流水线中负责接收指令。通过PCI-Express 将 GPU和CPU相连接，并读取CPU指令。再通过Front End（前端）处理指令。
- **GigaThread Engine（橙色部分）**
  - 为全局调度器，图形渲染流水线中负责将特定的数据从Host Memory中复制到Framebuffer中，创建Thread Blocks（线程块）再分配给各个彼此独立的SM线程调度器。

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构的先进性：以 Fermi图形渲染流水线为例—顶点处理

Fermi 核心微架构



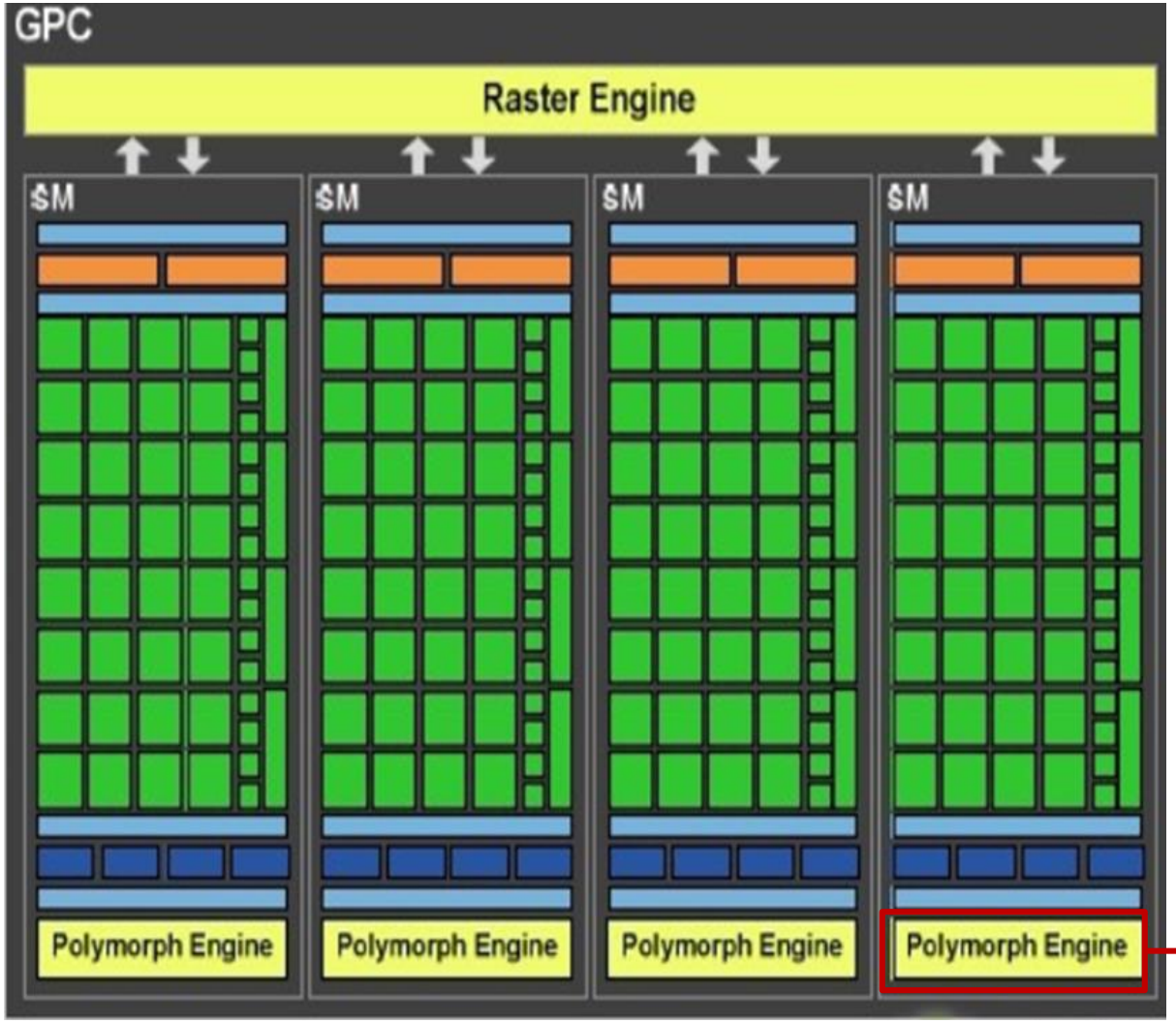


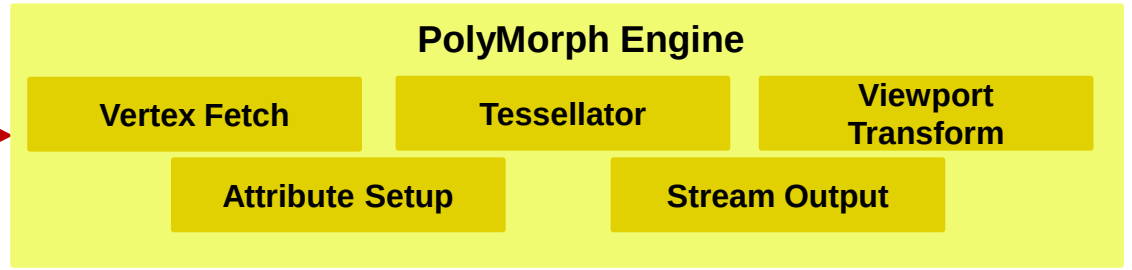
- **单个CUDA Core 组成部分包括：**
  - 1个Dispatch Port和1个Operand Collector、1个FP Unit和1个INT Unit和Result Queue。
- **在图形渲染流水线中：**
  - Vertex-shader执行单元对GPU前端读取的图形信息进行顶点数据确定，通过Vertex-shader 建立3D图形框架。

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构的先进性：以 Fermi图形渲染流水线为例—顶点处理

Fermi 核心微架构



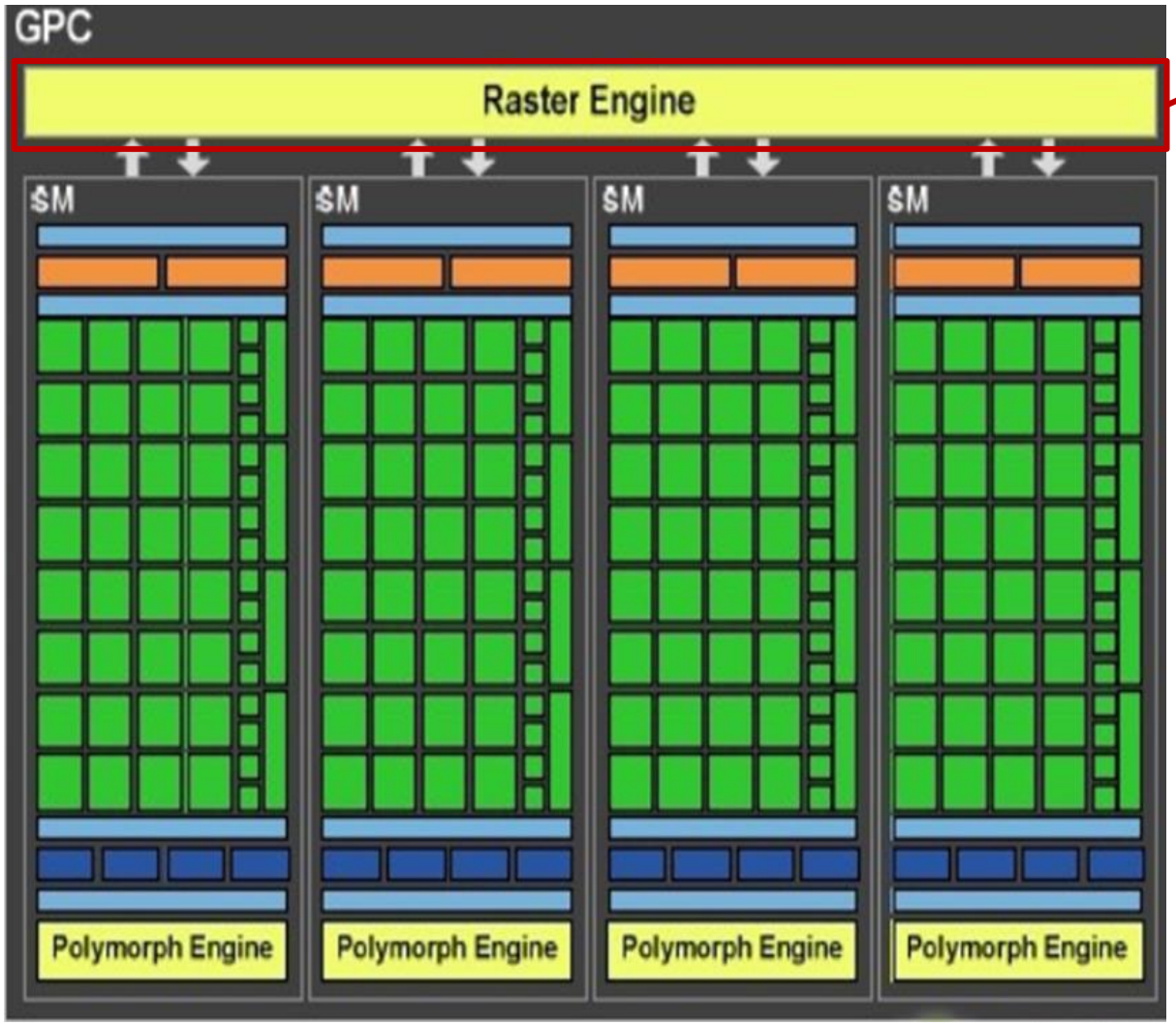


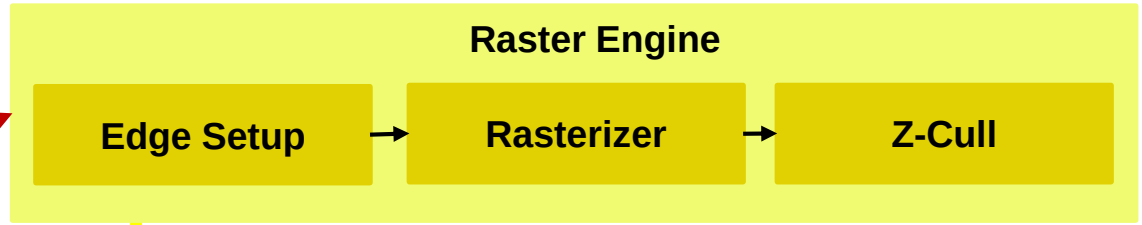
- **PolyMorph Engine（黄色部分，多形体引擎）**
  - 是全球首款实现了可扩展几何学流水线的重要元件。主要负责顶点拾取（Vertex Fetch）、细分曲面（Tessellation）、视口转换（Viewport Transform）、属性设定（Attribute Setup）、流输出（Stream Output）五个方面的处理工作。
- **在图形渲染流水线中：**
  - Vertex Fetch通过三角形索引取出三角形数据。
  - Viewport Transform负责模块处理已完成vertex-shader的所有指令，进行裁剪三角形，准备栅格化。
  - Attribute Setup确保经过插值后的vertex-shader数据在pixel-shader中的可读性。

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构的先进性：以 Fermi图形渲染流水线为例—光栅化计算

Fermi 核心微架构





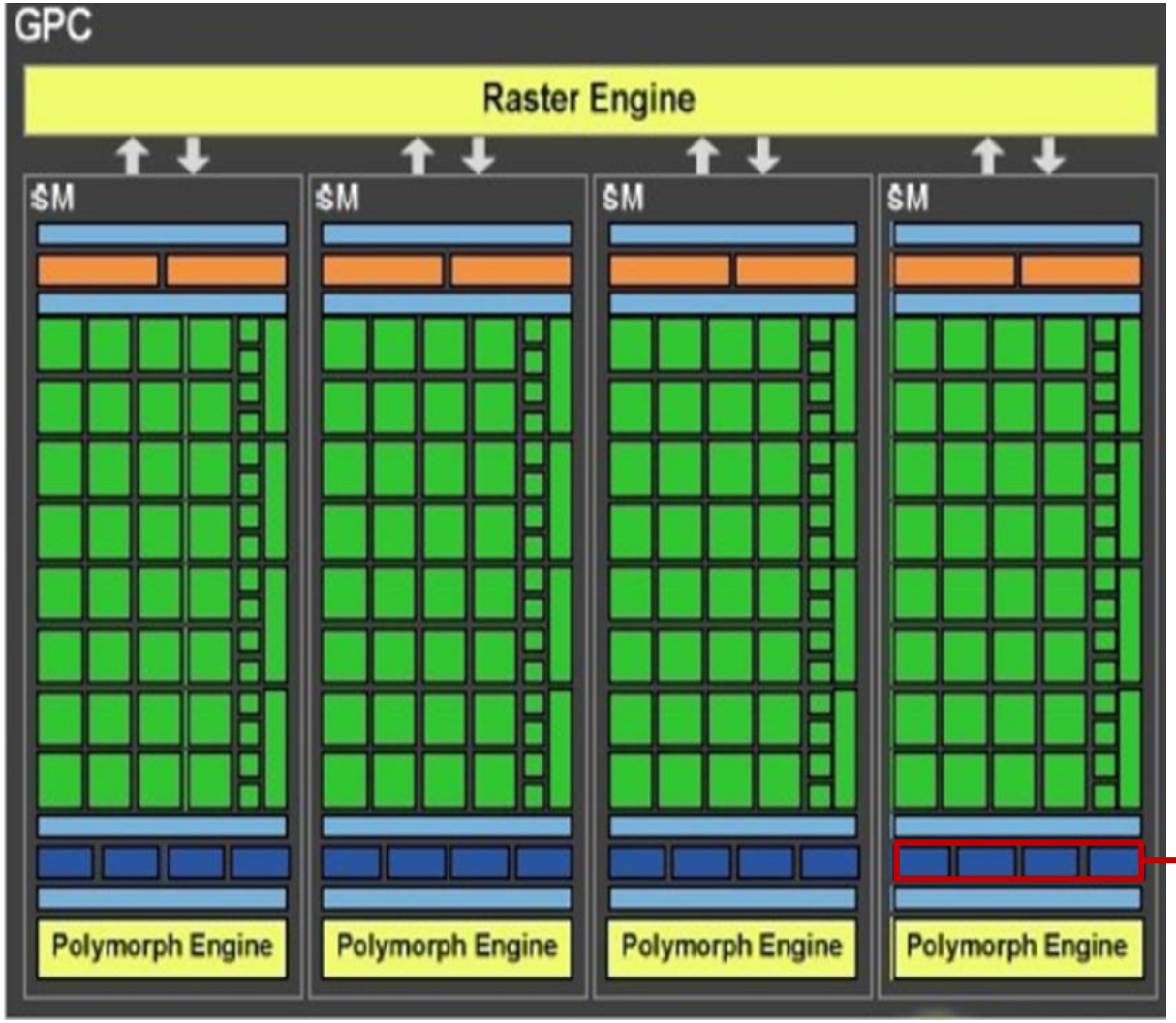
- **Raster Engine（黄色部分）**
  - 为光栅引擎，将光栅化处理硬件单元进行结合，包括Edge/Triangle Setup（边缘/三角形设定）、Rasterization（光栅化）和Z-Culling（Z轴压缩)。以流水线的形式运行指令，每时钟循环周期能够处理8个像素。

- **在图形渲染流水线中：**
  - 将Vertex-shader生成图形上的顶点和线段转化为对应的像素点，光栅化引擎在过程中负责接受三角形的像素信息生成和背面提出、Early-Z剔除。

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构的先进性：以 Fermi图形渲染流水线为例—纹理贴图

Fermi 核心微架构

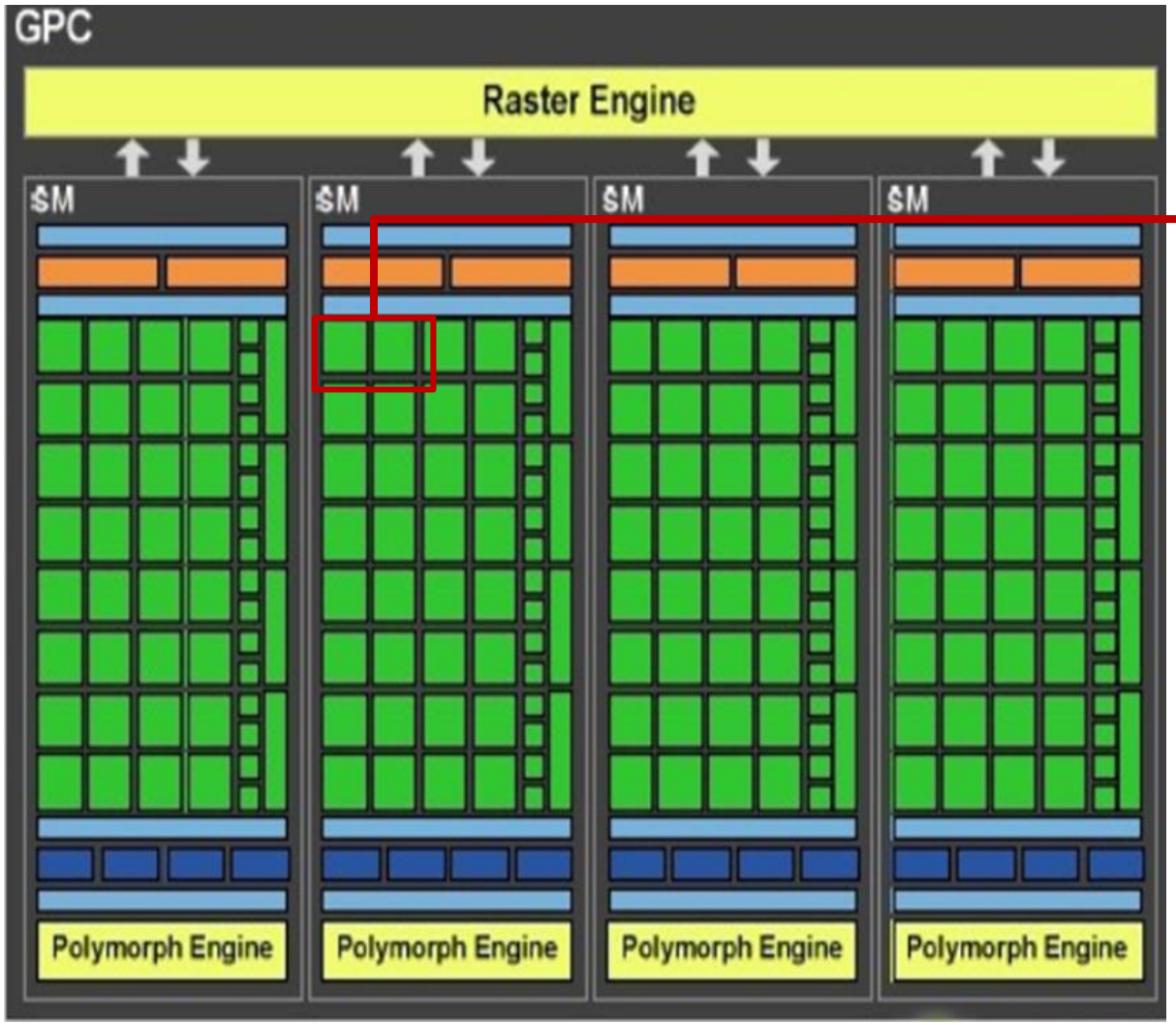


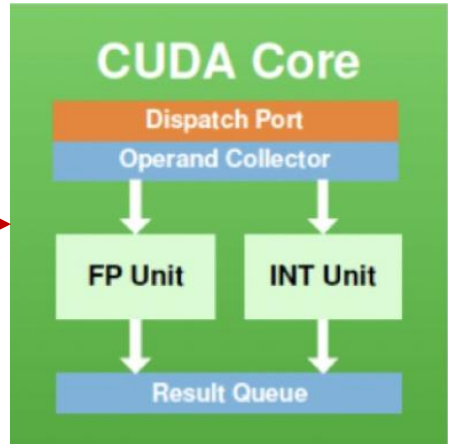
- **Texture Mapping Unit（蓝色部分）**
  - 为纹理映射单元，能够移动、变形、调整图形的大小和位置，主要功能是执行纹理采样。
- **在图形渲染流水线中：**
  - 将图片对应贴至经过顶点处理、光栅化计算后形成的3D多边形骨架的表面上，进一步形成直观的图形。

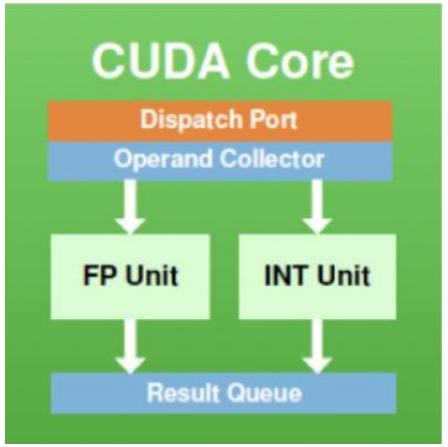
资料来源：NVIDIA官网，中信证券研究部

1.2 微架构的先进性：以 Fermi图形渲染流水线为例—像素处理

Fermi 核心微架构







- **在图形渲染流水线中：**
  - Pixel-shader执行单元对经过光栅化处理的像素点进行计算和处理，进而确定每个像素的最终属性。

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构的先进性：以 Fermi图形渲染流水线为例—最终输出

Fermi核心微架构

64 KB Shared Memory / L1 Cache

L2 Cache

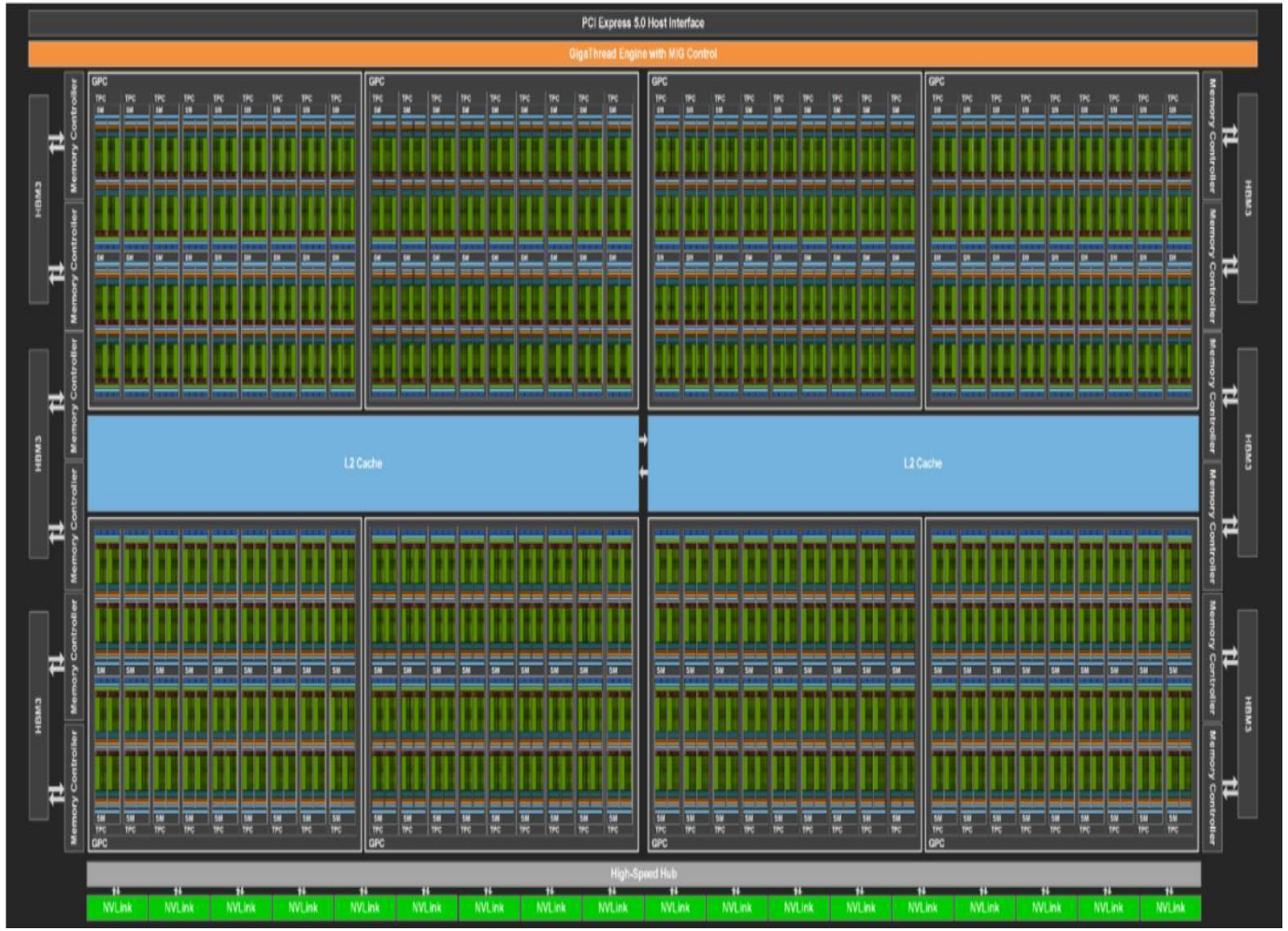
- **L1 Cache**
  - 为SM中的L1缓存，提高临时寄存器的使用效率，大幅降低CUDA运行耗时。在图形渲染流水线中负责处理寄存器溢出、堆栈操作和全局LD/ST，并且作为Vertex-shader和Pixel-shader的数据通信缓存。

- **L2 Cache**
  - 为L2缓存，与内部全部SM均相连通，为SM计算过程中需要读取相同数据的需求（如Vertex-shader和Pixel-shader）提供缓存支持。
  - 在图形渲染流水线中支持最终图形输出数据存放、读取操作，纹理操作，并且提供有效且高频的数据支撑。

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构的先进性：以 Hopper架构为例—总览

Hopper 核心微架构



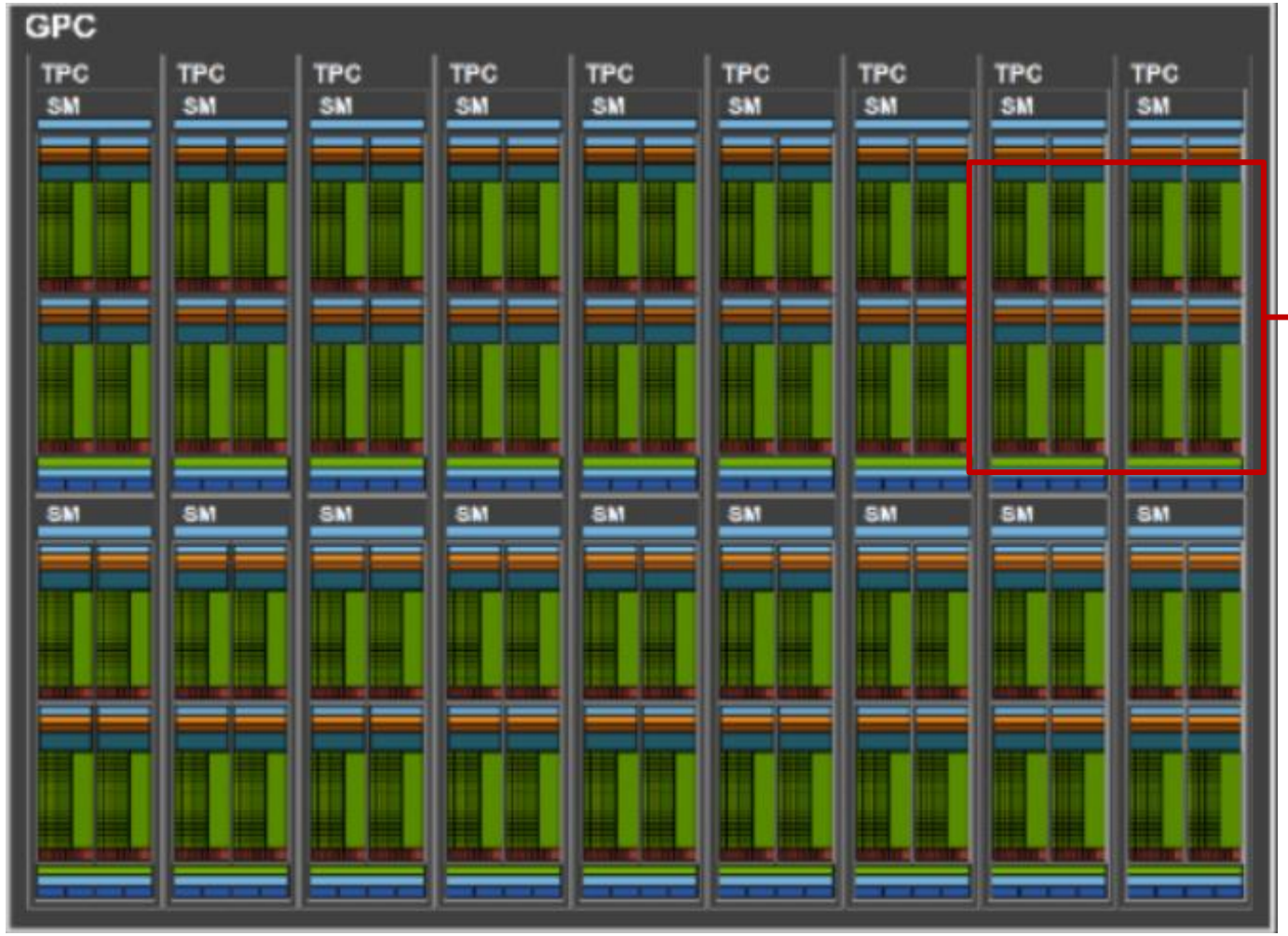
- **完整的GH100 GPU架构包括以下单元:8个 GPC、72个TPC、2个SM/TPC、每个完整 GPU 内含 144 个 SM。**

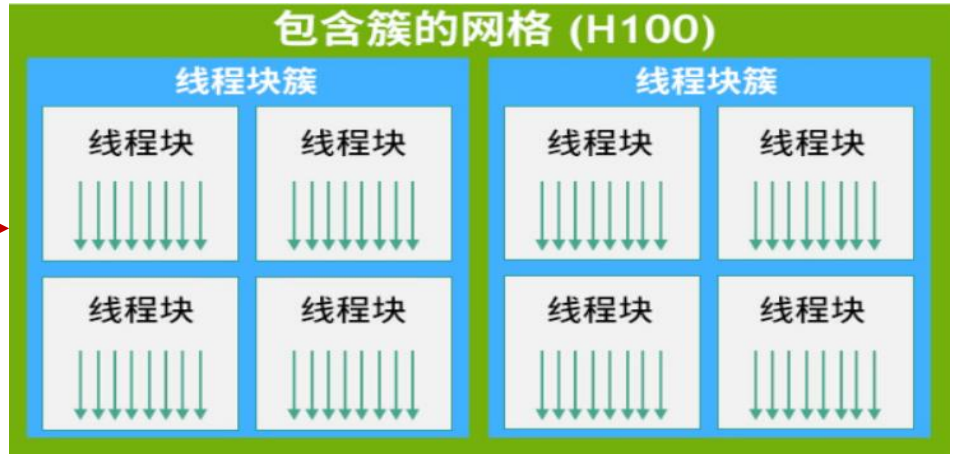
- **核心性能：**
  - 新型流式多处理器（SM），第四代Tensor Core提速6倍，DPX指令最高提速动态编程7倍，IEEE FP64和FP32芯片处理提速3倍。
  - 第二代多实例 GPU (MIG) 技术，扩增计算容量将近3 倍。
  - GPU 实例的显存带宽大幅度扩容近 2 倍，采用50 MB 二级缓存架构，支持大数据量重复访问。
  - 第三代NVSwitch 、PCIe 5.0

资料来源：NVIDIA H100 Tensor Core GPU Architecture白皮书

1.2 微架构的先进性：以 Hopper架构为例—GPC模块拆分

Hopper GPC核心微架构





- **每个GPC由9个TPC即纹理处理集群 (Texture Processor Cluster)组成。每个TPC包括2个SM单元，256个 FP32 CUDA Core 核心，8个 Tensor Core 核心。**

- **GPC线程块簇：**
  - 相比先前架构中的线程块分布，Hopper架构中新增了簇层次结构，该线程块簇在GPC内跨不同SM并发运行，新增了全新的内存访问方式和协作功能，能够实现不同SM之间的数据共享。

资料来源：NVIDIA H100 Tensor Core GPU Architecture白皮书，中信证券研究部

1.2 微架构的先进性：以 Hopper架构为例—SM模块拆分

Hopper SM核心微架构

- **SM全称Streaming Multiprocessor，Hopper架构下，每个SM包含128 个 FP32 CUDA Core 核心和 4 个第四代 Tensor Core 核心，主要组成部分包括：**
  - 1个L1 Instruction Cache，1个 Data Cache Cache 和4个L0 Instruction Cache（浅蓝色部分）
  - 4 个 Warp Scheduler（橙色部分）
  - 4个 Dispatch Unit（红褐色部分）
  - 4个Register file-寄存器文件（青色部分）
  - 128个 FP32 Unit（草绿色部分）
  - 64 个 FP64 Unit (墨绿色部分)
  - 4 个 Special Function Units (SFU/橘红色色部分)
  - 32个 LD/ST Unit（深红色部分）

资料来源：NVIDIA H100 Tensor Core GPU Architecture白皮书

1.2 微架构的先进性：以 Hopper通用计算流水线为例—指令接收

L0 Instruction Cache

- **L0 Instruction Cache：**
  - 全称指令缓冲区，能够存储 GPU用以绘图显示、数据变更、复制资源等指令的存储容器。
- **在通用计算-GPGPU流水线中：**
  - 主要负责检查当前指令（instruction）中的数据是否完备（ready）。分为两种情况处理：数据完备，传入Warp；数据不完备则存储于Instruction Buffer中。
  - Instruction Buffer可以屏蔽掉总线延时。因为GPU流水线上任务是并行处理，互不依赖的。

资料来源：NVIDIA H100 Tensor Core GPU Architecture白皮书，中信证券研究部

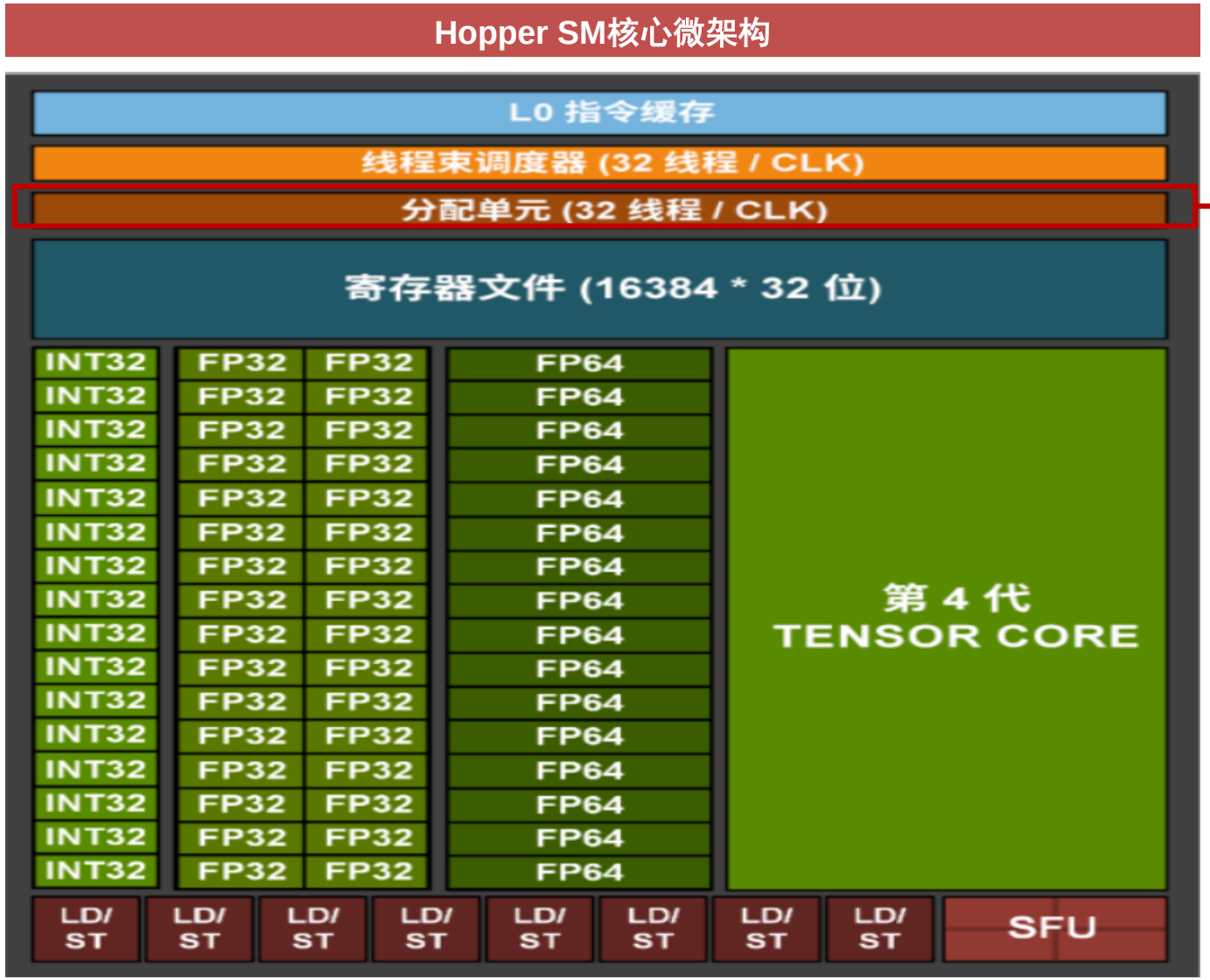
1.2 微架构的先进性：以 Hopper通用计算流水线为例—指令调度

Warp Scheduler

- **Warp Scheduler：**
  - 全称线程束调度器，在CUDA中，每32个线程组成线程束（warp），指令以一个warp为单位执行。
- **在通用计算-GPGPU流水线中：**
  - 主要负责任务调度。Warp Scheduler需要先确认当前Function Unit的状态，再将Instruction Buffer中已完备（ready）的指令调度给下一级的Dispatcher Unit。
  - 在单个时钟周期内可以同时调度两个warp指令。

资料来源： NVIDIA H100 Tensor Core GPU Architecture白皮书，中信证券研究部

1.2 微架构的先进性：以 Hopper通用计算流水线为例—指令分配

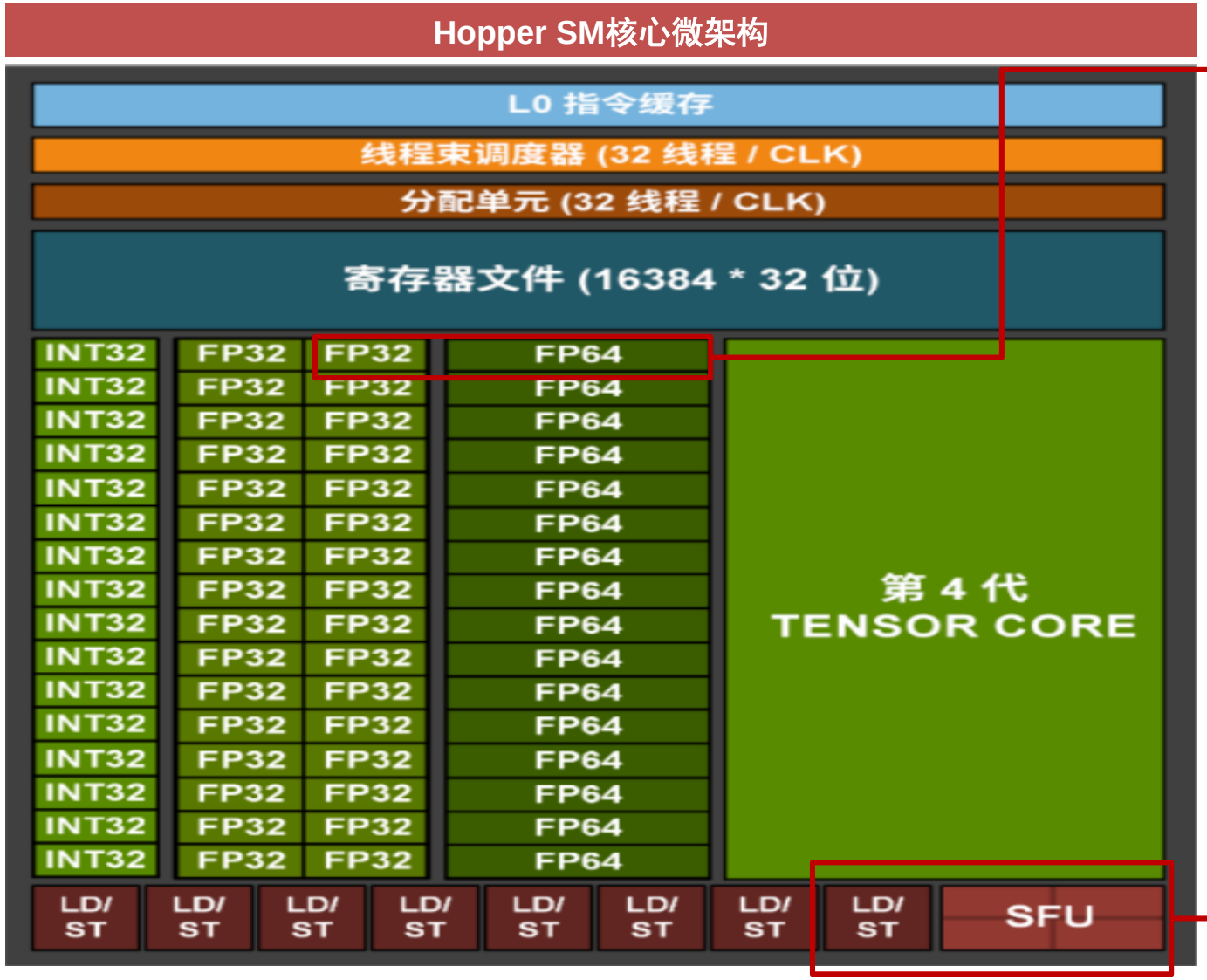


Dispatch Unit

- **Dispatcher Unit：**
  - 全称调度单元，可依据输入的数据、信息决定下一步需要调动的程序模块。
- **在通用计算-GPGPU流水线中：**
  - 主要负责根据指令（instruction）和Threadmask计算出下属各个function unit的instruction和register offset，用其计算结果，将指令传递至处于闲置状态的function unit下运行。

资料来源：NVIDIA H100 Tensor Core GPU Architecture，中信证券研究部

1.2 微架构的先进性：以 Hopper通用计算流水线为例—计算执行

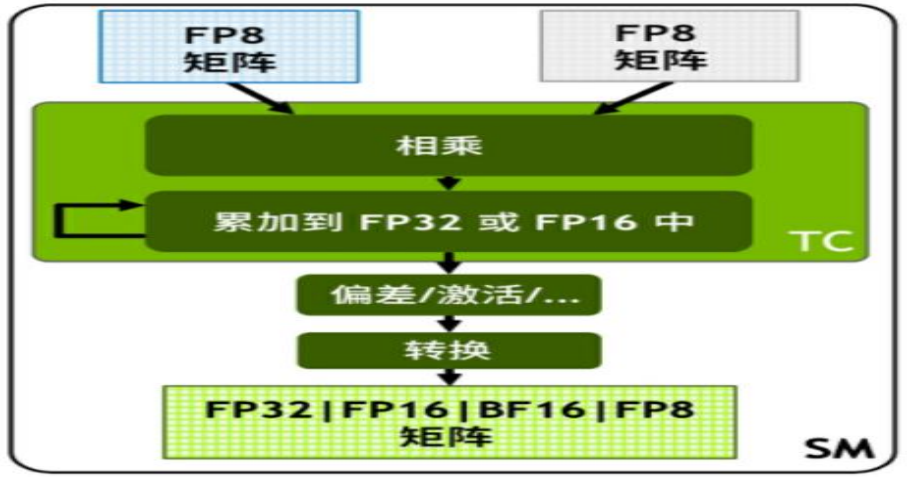


- **Function Unit：**
  - 为SM中的核心组成部件，称作功能单元，包括INT32 Unit、FP32 Unit、FP64 Unit、LD/ST Unit和SFU。
- **在通用计算-GPGPU流水线中：**
  - FP32 Unit和 FP64 Unit分别支持FP16/FP32的低精度计算以及FP64的高精度计算。
  - LD/ST Unit即加载/存储单元负责处理寄存器文件中的读写值
  - SFU负责用于计算正弦函数、余弦函数、指数、对数、倒数等特殊指令。

资料来源： NVIDIA H100 Tensor Core GPU Architecture ， 中信证券研究部

1.2 微架构的先进性：以 Hopper通用计算流水线为例—计算执行

Hopper SM核心微架构

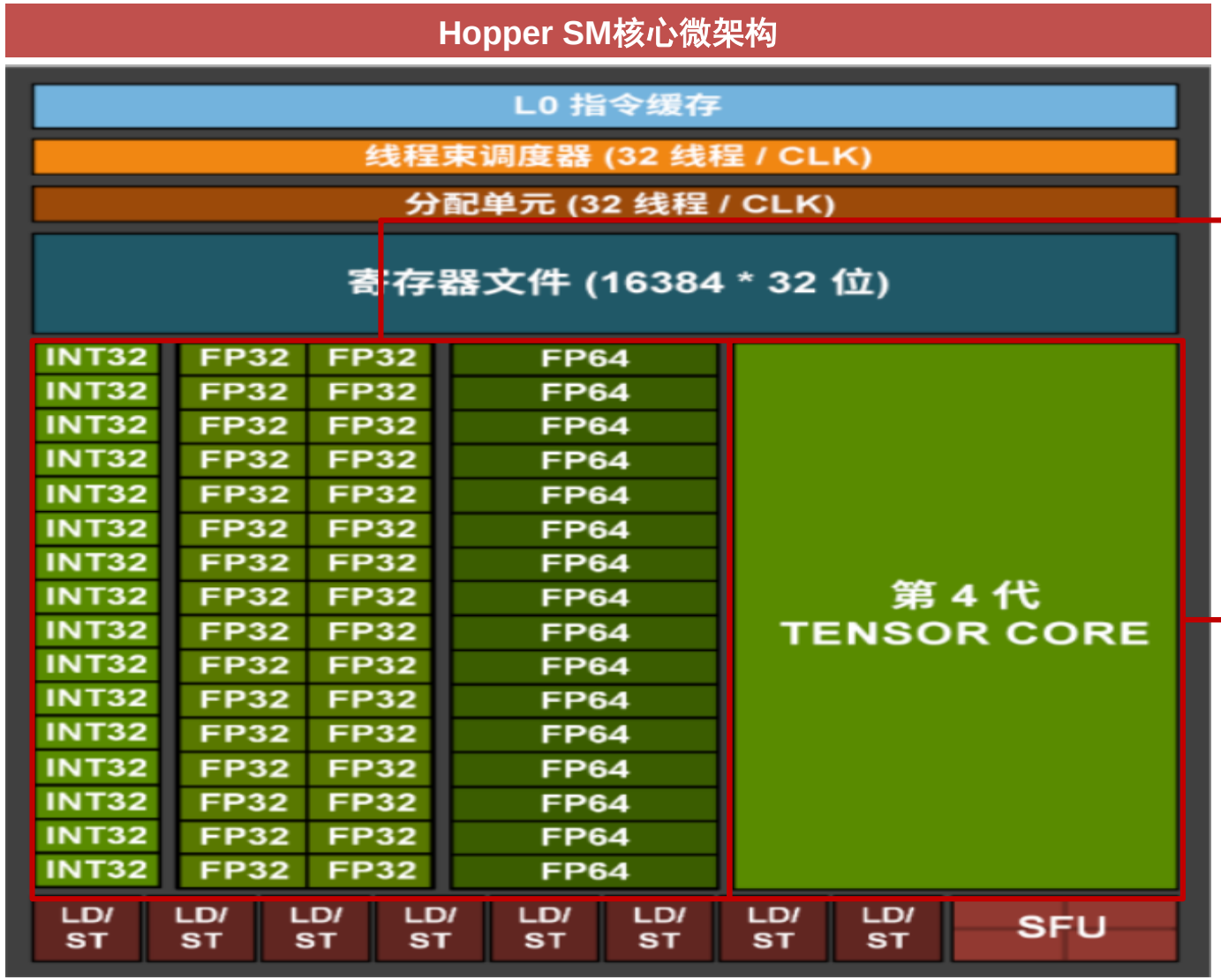


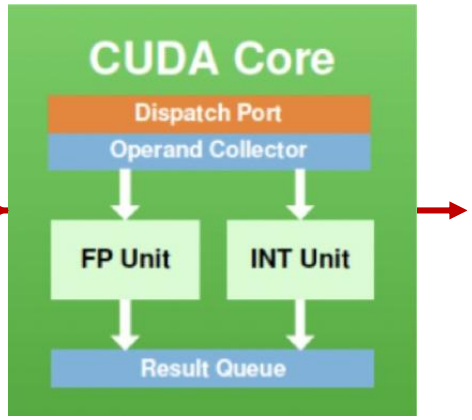
- **Tensor Core：**
  - 专门用于MMA(矩阵乘积累加)的高性能计算核心，可大幅度提升AI和HPC应用的性能。与其他运算相比，能够实现在GPU 内跨 SM 并行运行，并大幅提高吞吐量和效率。

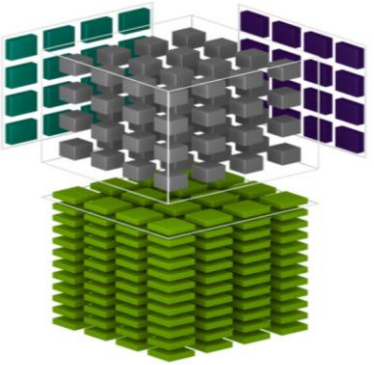
- **在通用计算-GPGPU流水线中：**
  - Tensor Core 专用于矩阵运算执行，对各类型数据高效管理，能够节省30%的操作数传输功耗。

资料来源： NVIDIA H100 Tensor Core GPU Architecture ，中信证券研究部

1.2 微架构的先进性：以 Hopper架构为例—CUDA vs Tensor Core







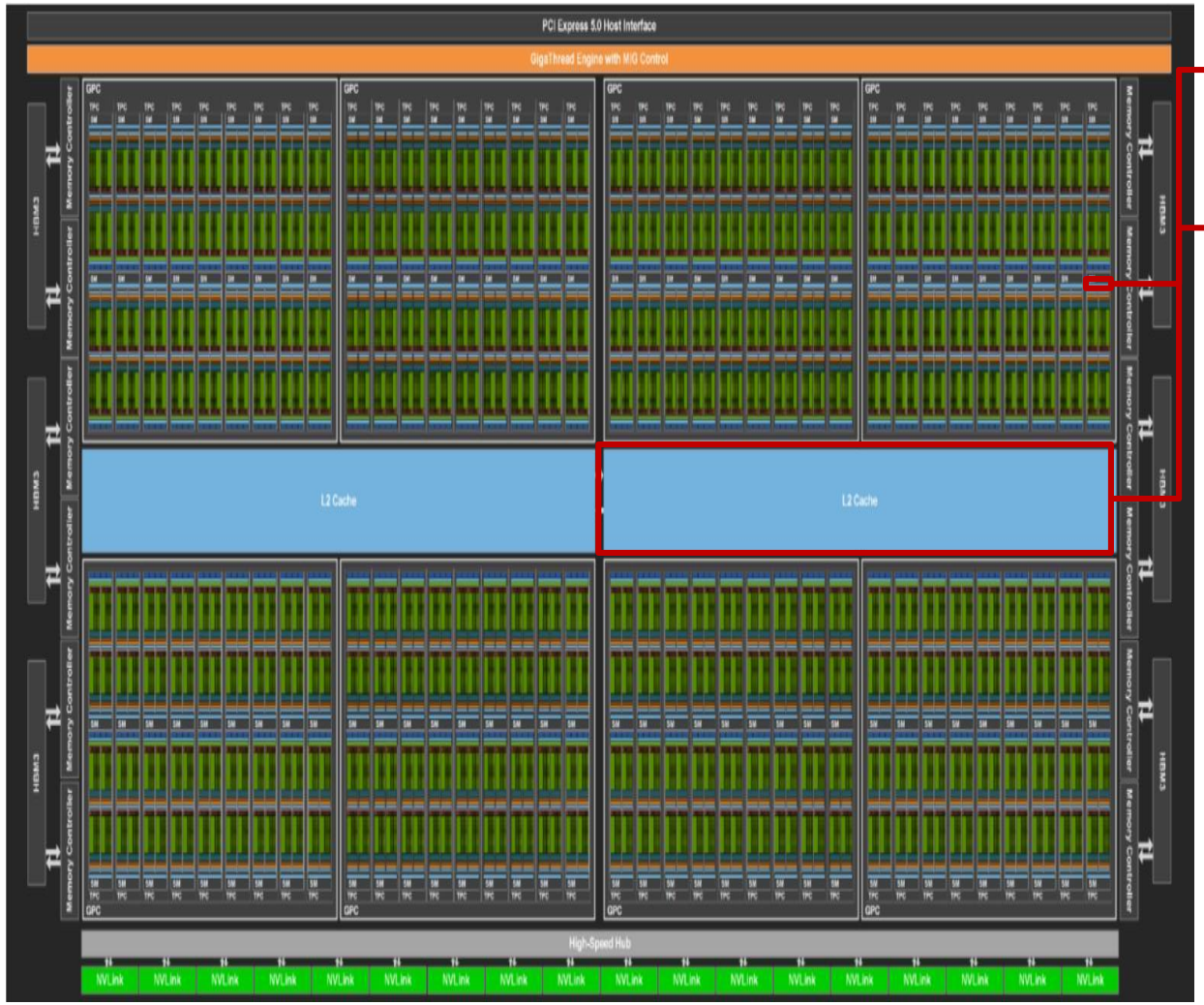
- **全能计算型浮点运算单元CUDA Core：**
  - 架构上划分为不同精度的计算核心支持多种数据类型，包括INT32、FP32、FP64，每次运算执行一次乘法 1 x 1 per GPU clock。

- **张量运算专用执行单元Tensor Core：**
  - 专门为深度学习、神经网络训练和推理运算设计的运算内核，支持 FP8、FP16、BF16、TF32、FP64 和 INT8 MMA 数据类型，每次运算执行一次矩阵乘法 [1 1 1 [1 1 1
    1 1 1 x 1 1 1 per GPU clock
    1 1 1] 1 1 1]

资料来源： NVIDIA H100 Tensor Core GPU Architecture ，中信证券研究部

1.2 微架构的先进性：以 Hopper通用计算流水线为例—结果输出

Hopper 核心微架构



L1 Data Cache/ Shared Memory

L2 Cache

- **L1 Data Cache**
  - 为SM中的L1数据缓存，也称共享缓存，单个L1缓存有256KB的存储容量。在通用计算流水线中L1缓存负责缓存内存地址，作为连续缓存供给warp调度器使用。

- **L2 Cache**
  - 为L2缓存，也称二级缓存，与内部全部SM均相连通，作为公用缓存支持GPU读取操作。在通用计算流水线中作为Global Memory缓存，存储GPU的部分拷贝，容量大，供给整体GPU使用。

资料来源：NVIDIA H100 Tensor Core GPU Architecture，中信证券研究部

1.2 架构的先进性：NVIDIA历代微架构对比

NVIDIA历代微架构对比

架构代号	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper
中文代号	费米	开普勒	麦克斯韦	帕斯卡	伏特	图灵	安培	赫柏
时间	2010	2012	2014	2016	2017	2018	2020	2022
核心参数	16个SM，每个SM包括32 Cuda Cores，共计512 Cuda Cores	15个SMX，每个SMX包括192个单精度+64个双精度的Cuda cores	16个SMM，每个SM包括4个处理块，每个处理块包括32个CUDA内核+8个LD/STUnit+8个SFU	Pascal架构有GP100、GP102 GP100 有60个SM 每个SM包括 64个cuda cores 32个DP cores	80个SM， 每个SM里 32个FP64 64个INT32 64个FP32 8个Tensor core	TU102核心72个SM，SM全新设计，每个SM里64个INT32 64个FP32 8个Tensor core	A100有108SMs 每个SM 64个FP32 64个INT32 32个FP64 4个Tensor core	H100有132 SM 每个SM 128个FP32 64个INT32 64个FP64 4个Tensor core
特点\优势	首个完整GPU计算架构，支持与共享存储结合纯Cache层次的GPU架构，支持ECC的GPU架构	游戏性能大幅提升 首次支持GPU Direct技术	相比Kpler的每组SM单元192个减少到了每组128个，但是每个SMM单元拥有更多的逻组控制电路	NVLink一代，双向互联带宽160GB/s P100有56个SM HBM	Nvlink2.0 Tensor Core 1.0 满足深度学习和AI运算	Tensor Core 2.0 RT Core 1.0	Tensor Core 3.0 RT Core 2.0 Nvlink 3.0 结构稀疏性 MIG1.0	Tensor Core 4.0 Nvlink 4.0 结构稀疏性矩阵 MIG 2.0
纳米制程	40/28nm 30亿晶体管	28nm 71亿晶体管	28nm 80亿品体管	16nm 153亿品体管	12nm 211亿晶体管	12nm 186亿晶体管	7nm 283亿晶体管	5nm 800亿个晶体管
代表型号	Quadro 7000	K80 K40M	M5000 M4000	P100 GTX1080 P6000	V100 TiTan V	T4 2080TI RTX 5000	A100、A30 3090	H100

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构未来方向：更多、更专、更智能

- **更多：硬件上，图形渲染单元和通用计算单元数量增多。**
  - GPU产品迭代发展过程中，包括晶体管、流处理器单元、纹理单元和光栅单元等硬件单元数量上升。
  - 以NVIDIA A100和H100产品架构对比为例，在SM数量、TPC数量、FP32 Core核心数量、FP64Core 核心数量上都具有显著增加，同时也使得H100在性能峰值上得到显著提升。

更多：硬件单元数量增多

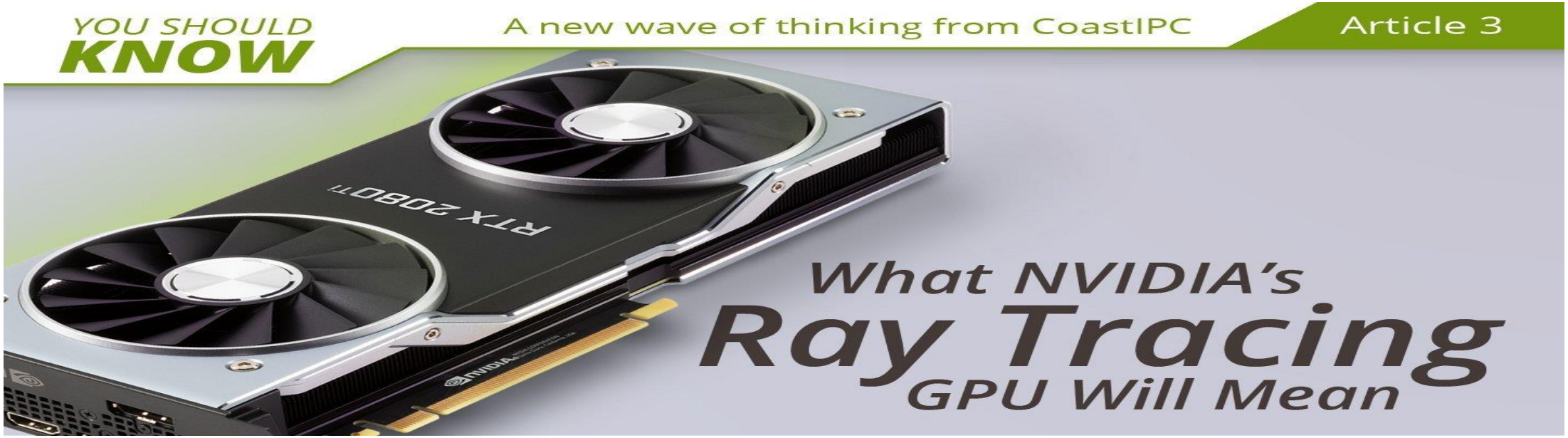
	A100	H100
Tensor Cores	432	528
SM数量	108	132
TPC数量	54	66
FP64 CUDA Cores	3456	8448
FP32 CUDA Cores	6912	16896
Menory Bandwidth	2TB/sec	3TB/sec
使用 FP16 累加的 FP16 Tensor TFLOPS 峰值	312/624[2]	1000/2000[2]
FP32 Vector	19.5TFLOPS	60 TFLOPS
FP64 Vector	9.7 TFLOPS(1/2 FP32 rate)	30 TFLOPS

资料来源：NVIDIA官网，中信证券研究部

1.2 微架构未来方向：更多、更专、更智能

- **更专：图形渲染能力更精细，通用计算能力更高效。**
  - **图形渲染领域：**采用光线追踪技术，相较传统光栅化渲染方式，光线追踪采用基于物理渲染方式，使得所实现的效果更加接近显示，具有更加逼真的图显能力。
  - **通用计算领域：**目前国际各大厂商均推出GPGPU计算解决方案，大规模扩展计算能力的高性能计算。例如：1）ATI Stream：为程序员提供SDK开发工具包以协同进行GPU计算。2）NVIDIA CUDA：推出统一计算架构，由管线分工式设计转变为统一化的处理器设计，学习成本较低，能够通过C、C++编程语言进行程序编写。

更专：图形渲染 Ray Tracing GPU



资料来源：EXPreview@倪嘉声

1.2 微架构未来方向：更多、更专、更智能

- **更智能：GPU AI运算能力上升。**
  - GPU在AI领域得到广泛的应用，包括自动驾驶、医疗影像分析、人工智能计算能力、金融模型建立等领域，如第三代的张量单元相较于上代在吞吐量上提升了1倍。
  - GPU自身结构特点决定其在AI的发展方向：1）多线程，计算单元数量多，并行计算方式能够同时计算大量数据。2）拥有更直接、迅速访问缓存的能力。3）拥有更高精度的浮点算力，能够更佳适配于推理训练、深度学习。

更AI智能

资料来源：NVIDIA官网

1.2 GPU跑分指标：GPU性能的直观体现

- **评估GPU性能的参数主要包括：算力、纹理/像素填充率、功耗、加密性能等。**
  - 算力性能参数的核心指标包括算力（单/双精度浮点性能等）、功耗。GPU算力越强，GPU的综合计算能力和运行性能越强。
  - 图形渲染性能参数的核心指标包括纹理填充率、像素填充率等。
  - 其他性能参数：显存使用率、加密性能等。
- 一般NVIDIA的GPU的F32峰值算力计算方法为：核心数*核心频率*2

GPU性能参数指标

性能指标	含义
算力	也称计算吞吐量，单位为GFLOP/s，表示每秒浮点运算量
纹理填充率	指GPU在单位时间内所能处理的纹理贴图数量，单位为Mtexels/S，计算公式为核心频率×纹理单元数目/1000
像素填充率	等于ROP运行的时钟频率 x ROP的个数 x 每个时钟ROP可以处理的像素个数，表明GPU每秒处理像素数量
功耗	指功率的损耗，即输入与输出功率之差，一般体现于元器件上热能的耗散
加密性能	包括AES-256和SHA-1哈希两项常用加密技术的性能
单精度浮点性能	评估三维图形显示能力，通过“Julia”分形测量，计算公式为核心数*核心频率或FP32 cores ×GPU Boost Clock × 2
双精度浮点性能	用于评估三维图形生成能力，通过“Mandelbrot”分形测量，计算公式为FP64 Cores × GPU Boost Clock × 2

资料来源：CSDN@Charles Ren，CSDN@AliceWanderAI，中信证券研究部

1.2 GPU性能测试跑分

- **GPU的性能指标可以通过GPU综合评分软件进行直观的比较。**
- **常见GPU测试工具包括GPU-Z、Mlperf、3DMark、FurMark、AIDA64 Extreme、GpuTest和Gpu burn等。**
  - 基本信息检测主要通过GPU-Z；主流游戏测评主要通过3DMark；AI 性能基准测试主要通过Mlperf。

常见GPU测试工具

测试工具名称	描述
3DMark	提供 Time Spy、Fire Strike、Tomb Raid等测试
MLperf	由学术界、研究实验室以及与AI领域相关等机构联合发起，针对软硬件的推理、训练性能提供评估
GPU-Z	显示包括显卡型号，显存型号，显卡品牌，基本规格，动态频率，实时温度等重要信息
FurMark	可进行温度压力测试，稳定性测试和OpenGL测试
MSI Kombustor	通过密集模拟和演示进行显卡测试，且提供热性能和稳定性测试
OCCT	测试电压、频率和分辨率等超频参数的稳定性
AIDA64 Extreme	提供GPGPU Benchmark 测试。包括 GPU 在内的系统稳定性测试
GpuTest	基于TessMark的曲面细分测试、几何实例测试，还提供对GPU进行OpenGL基准压力测试
Gpu_burn	提供gpu压力测试

资料来源：CSDN@StarkerRegen，volvootofinans，中信证券研究部

1.2 GPU图形渲染游戏性能测试—3DMark

- **3DMark集成了PC和移动设备内游戏完整的基准测试，能够为不同PC适配不同测试，且通过3DMark分数实现与其他CPU、GPU组合系统比较，提供游戏性能估算。**
  - 3DMark经过近10年更新，现已覆盖十多项基准测试、压力测试和功能测试。

3DMark 测试汇总

测试工具名称	适配设备	描述
Time Spy	游戏型 PC	对windows系统下游戏型 PC 的 DirectX 12 基准测试。是DirectX 12 前期研发的应用程序之一，有助于实现新 API 提供的性能收益。
Port Royal	PC	游戏玩家的实时光线追踪基准测试。同时支持微软 DirectX 显卡的光线追踪性能。
Night Raid	具有集成显卡 的PC	针对windows系统下装配集成显卡和 Arm 处理器的低功耗平台等的小型移动运算设备的 DirectX12 基准测试
Wild Life	PC、智能移动设备	适用于微软、安卓和 iOS 系统的跨平台基准测试。使用 Vulkan 图形 API。
Fire Strike	游戏型 PC	使用于游戏型PC 的 DirectX 11 基准测试产品。Fire Strike 包括显卡测试、物理测试和 CPU 和 GPU 联合测试。
CPU Profile	现代处理器	导入新的 CPU 基准测试方法，共包含六个测试
存储基准测试	游戏玩家SSD	扩展了 3DMark测试范围，用于存储硬件包括SSD在内等的游戏性能。
压力测试	PC	针对组装型 PC在 GPU升级和超频情况下对系统可靠性和稳定性检测，可定位硬件故障

资料来源：3DMark官网，中信证券研究部

1.2 GPU AI性能测试—MLPerf

- **MLPerf 基准测试不仅提供AI训练测试，还提供推理解决方案支持。**
  - 在训练领域，MLPerf 覆盖八大工作负载测试，包括视觉渲染、语言识别、个性化推荐和深度学习等。
  - 在推理领域，MLPerf 在七大不同神经网络进行用例测试，包括计算机视觉领域、推荐系统、语言处理和医学影像场景。

MLPerf 提交分类

测试性能	描述
图像分类	提供离线场景 (Offline Scenario) 性能测试，适用于计算机视觉问题，从一组固定的类别中分配一个标签到一个输入图像
目标检测（轻量级）	在图像或视频中确定实物目标对象的能力，并在每个标的对象周围指定一个边界框
目标检测（重量级）	分层检测图像中出现的重视程度不同的对象能力，并分别鉴别对象像素掩码
生物医学图像分割	基于神经网络，主要对医学影像中复杂图形的识别能力进行测试，通过三维分割技术执行提供医学用例
自动语音识别（ASR）	测试实况识别对话，音频转录能力。
自然语言处理 (NLP)	指能够根据文本上下文对语段进行理解，并且提供回答问题、解释语句的能力。
推荐系统	能够通过了解用户与产品广告的交互及内容，在公开社交网站、商务网址等面向特定客户提供个性化推荐能力
强化学习	评估不同行为的可能性，能够在测试游戏中赢得比赛。

资料来源：MLPerf 官网，中信证券研究部

1.2 应用场景：数据中心、游戏业务、图形显示、OEM、加密货币

- **GPU技术不断发展，GPU的应用场景也随之不断拓宽，不仅包含图形处理，还在AI、边缘计算等新领域发挥重要作用。**
- **图形显示是GPU最基本的功能**
  - GPU的诞生原因就是分担CPU计算量，凭借其处理并行计算的优势承担图像信息的运算工作。在游戏画面显示、图像运算等领域广泛应用。
- **GPGPU被视为AI时代的算力核心**
  - 应用于人工智能场景的服务器通常搭载GPU、FPGA、ASIC等加速芯片。加速芯片和中央处理器的性能结合支撑高吞吐量的运算需求，为图形视觉处理、语音交互等场景提供算力支持，已经成为人工智能发展的重要支撑力量。GPU由于在架构设计上擅长进行大量数据运算，被广泛应用于人工智能计算中。在人工智能的应用和研究、智能安防、边缘计算、无人驾驶等领域发挥作用。

GPU重要应用场景

应用场景	主要特点	具体应用方式	类型
数据中心	作为加速芯片	集中于AI计算领域，覆盖数据中心加速器、边缘计算	GPGPU为主
游戏业务	并行计算结构、浮点运算能力强、访存速度快	游戏绘图、画质渲染、增强用户游戏体验，云游戏平台建设	传统GPU
图形显示	图显专业化、精细化	广泛渗透于Quadro专业绘图、3D渲染、专业设计软件、传输DPU	传统GPU
OEM&IP	高性能、低功耗、产品迭代速度快	服务器厂商核心技术开发、半导体行业IP大厂授权	GPGPU为主
加密货币	核心数量多，适合大量重复的较简单运算	挖矿速度与矿机算力正相关	GPGPU为主

资料来源：电子元器件论坛@华秋商城，中信证券研究部

1.2 应用场景：人工智能芯片GPGPU、FPGA、ASIC的选择

- **GPGPU：**为通用图形处理器，擅长图形处理，“粗粒度并行”技术。特点为拥有高灵活性、运用并行结构、在图形和复杂算法上效率较高；缺点为价格贵且功耗高。
- **FPGA：**为现场可编程逻辑阵列，擅长于算法更新频繁的专用领域。特点为灵活性适中、可以同时进行数据并行和任务并行计算、制作成本低于ASIC、定制化、功耗低。在国内多用于芯片验证。
- **ASIC：**为专用集成电路，应用于市场需求量大的专用领域。指应特定用户要求和特定电子系统的需要而设计、制造的集成电路，特点是灵活性较低、高性能、成本高、可靠性高。缺点是算法相对固定、开发时间成本高。

GPGPU FPGA ASIC对比

	定义	技术	应用场景
GPGPU	通用图形处理器	将GPU与CPU结合，并将一些计算密集型任务从CPU移到到GPU，CUDA编程环境和CTM编程环境的出现，使GPU打破图形语言的局限成为真正的并行数据处理超级加速器。	人工智能计算，多媒体分析、无人驾驶、VR/AR等产业、金融、电信等行业
FPGA	现场可编程逻辑阵列	在 PAL、 GAL、 CPLD 等可编程器件基础上进一步发展，用户可以通过烧入FPGA配置文件定义门电路与存储器连线，使其具有不同功能	人工智能、信号处理、嵌入式处理、原型验证、接口应用与逻辑黏合
ASIC	专用集成电路，指应特定用户要求和特定电子系统的需要而设计、制造的集成电路	采用定制设计，用复杂可编程逻辑器件和FPGA 进行设计，与用户系统密切结合	安全相关产品、人工智能、消费电子、航空航天及其他为特定用途定制的场景

资料来源：爱码网@X-knight，Stack，CSDN@zj xlink，中信证券研究部

1.3 GPU指令集：GPU进行图形渲染和通用计算的指令集合

- **GPU指令集本质是硬件执行功能的机器码。**
  - 指令是计算机运行的最基本工作单位，是GPU功能实现的重要基础，通常包括指令格式、寻址方式和数据形式等。
  - GPU指令集是GPU中用以计算和控制系统的指令集合，指令集的先进与否直接关系到GPU性能的高低。操作系统通过指令集对硬件进行管理和资源分配，并规范程序按认可方式编译运行。
  - GPU指令集分类包括PTX、CUDA SASS指令集等。

CUDA SASS 指令集

```
IMAD R4, R5, R0, -R4 ;
IMAD.HI.U32 R49, R5, c[0x0][0x1d8], RZ ;
IMAD.WIDE.U32 R20, R17, c[0x0][0x168], R20 ;

IMAD.MOV.U32 R31, RZ, RZ, c[0x0][0xc] ;
IMAD.SHL.U32 R0, R2, 0x8, RZ ;
IMAD.IADD R2, R7, 0x1, R11 ;
```

IMAD
- 本身为计算整数d=a*b+c(左图为R4=R5*R0-R4)的指令，单独的IMAD等同于三个32bit数进行乘加。

WIDE
- 在IMAD指令的基础上，表示乘数a和b为32位，加数c为64bit，输出结果d也为64bit。

SHL
- 表示乘数b为2的幂次，不需要完整的乘法器功能，不影响计算结果。

IADD
- 表示乘数b为1，能够让功能单元得到更明确的信息，进而优化计算。

资料来源：CUDA编程方法论@cloudcore，中信证券研究部

1.3 GPU指令集：以 SASS指令集为例

- **指令集相关性质：**
  - 指令集本身在特定架构改变下会表现为指令性能变化，而本身的编码和功能并没有发生改变。
  - 兼容性：经过CUDA C、C++编译完成后，会同时生成与SM单元对应的PTX和SASS代码。
  - 指令执行吞吐是评价GPGPU执行的有效指标，GPU指令吞吐一般用每单位SM在一周期内执行的指令数量计算
- **SASS指令集分类：**主要包括Predicate操作指令、Float指令、Integer指令、格式转化/数据移动/内存操作/跳转分支指令和Uniform DataPath指令

SASS指令集基本分类

指令名称	描述
Predicate操作指令	也称作guard predicate，由4bit编码指定，是控制线程是否执行指令的方式之一
Float指令	基本包含4大类：float64、float32、float16和MUFU指令
Integer指令	基本包含算术指令、移位指令、逻辑操作指令和其他位操作指令
格式转换指令	主要为数值格式的转换，在整型和浮点型间转化
数据移动指令	以MOV指令、PRMT指令为首的数据搬运操作
内存操作指令	指令较为复杂，包含memory的load操作和store操作、Cache control指令、Texture操作指令以及Surface操作指令
跳转和分支指令	是SASS指令集中最频繁随架构变动的指令，包含了定向跳转或条件定向跳转、不定跳转、分支管理操作、跳转目标管理和特殊跳转指令
Uniform DataPath指令	与SM中用于warp公共计算的ALU功能单元相配套，使得每个warp只需要单个执行

资料来源：CUDA编程方法论@cloudcore，中信证券研究部

1.3 GPU指令集：以 SASS中的具体指令为例

- **MOV: 能够完成基本传送指令**
  - MOV指令是编程中最基本的指令，能够将数据从起始源地址传送到目标地址。功能范围覆盖立即数传送、寄存器传送、储存器传送、段寄存器传送。
- **MUFU: 作为SASS指令集中计算超越函数的重要工具。**
  - 超越函数指的是相对有限次加减乘除等组合而言，硬件上无法用多项式表示的函数需要通过该指令进行近似计算，若对精度有进一步要求，还需要调用数学函数库中其他软件。

MOV 立即数传送示例

```
MOV CL,4 ; CL←4, 字节传送

MOV DX,0FFH ; DX←00FFH, 字传送

MOV SI,200H ; SI←0200H, 字传送

MOV BVAR,0AH ; 字节传送 ; 假设BVAR是一个字节变量,定义如下:BVAR DB 0

MOV WVAR,0BH ; 字传送 ; 假设wvar是一个字变量, 定义如下: wvar dw 0
```

资料来源：CSDN@Jock.Liu

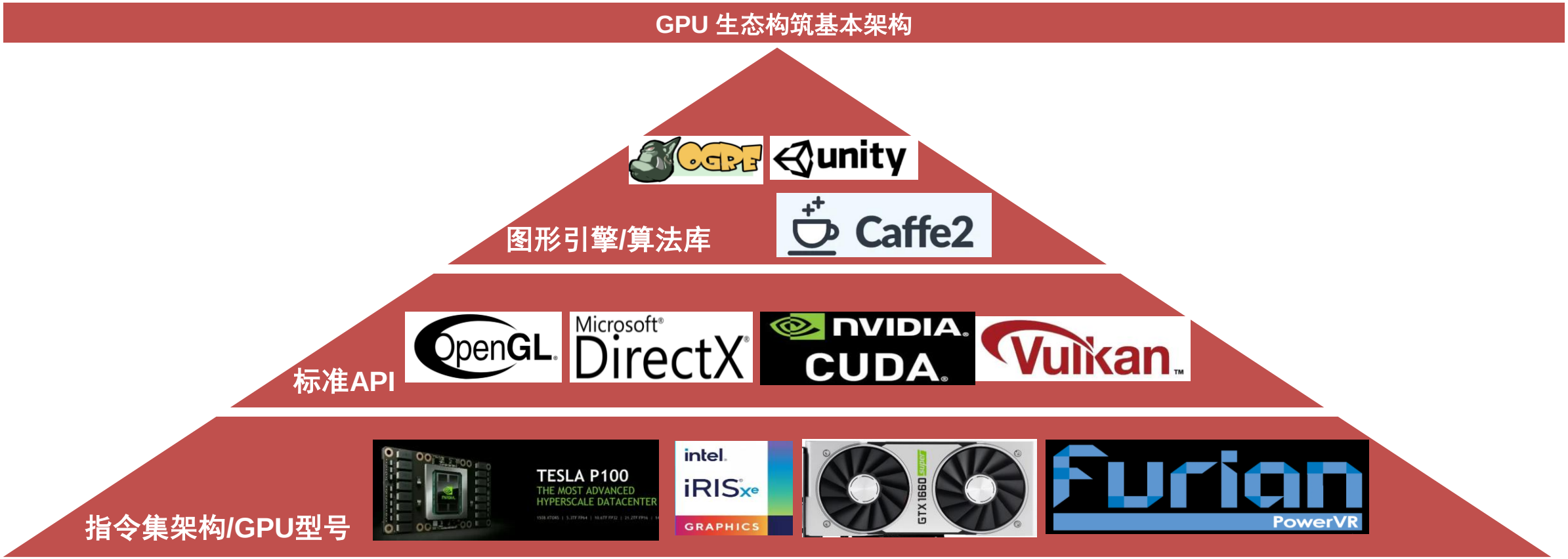
常见的MUFU类指令

```
MUFU.RSQ R5, R10 ;
MUFU.RCP R3, R14 ;
MUFU.EX2 R9, R8 ;
MUFU.LG2 R10, R9 ;
MUFU.COS R9, R19 ;
MUFU.SIN R10, R19 ;
```

资料来源：CUDA编程方法论@cloudcore

1.3 GPU生态体系：构筑通用计算壁垒

- **完善的GPU生态体系能够兼容不同的软件、硬件平台，使得GPU性能得到最佳释放。**
  - GPU生态的由三大部分基本构成：1）上层图形引擎、算法库。2）中层标准API接口适配各类驱动、编译器。3）底层硬件/指令集架构。



资料来源：各公司官网，中信证券研究部

1.3 GPU生态体系：构筑行业壁垒的基石

- **IP研发难度高：**
  - IP研发难度大、需要多年沉淀才能产出稳定性较佳的产品。目前GPU领域中，想要短期内产出需要依赖外部IP授权。市场上大多公司使用Imagination提供的IP，即在购买商用GPU IP之后自行修改迭代。以苹果芯片IP专利为例，苹果在A10之前处理器芯片都是采用Imagination的IP。
- **软件门槛高：**
  - 计算机芯片除了硬件之外，还要求有与之配套的软件体系，而GPU软件体系复杂，涵盖各类图形API、计算接口、基础库、应用对接适配等等。NVIDIA在各类软件驱动测试上已投入大量时间，形成较强的生态效应。

Apple A10芯片

资料来源：ESM China

GPU 软件

资料来源：Techpowerup

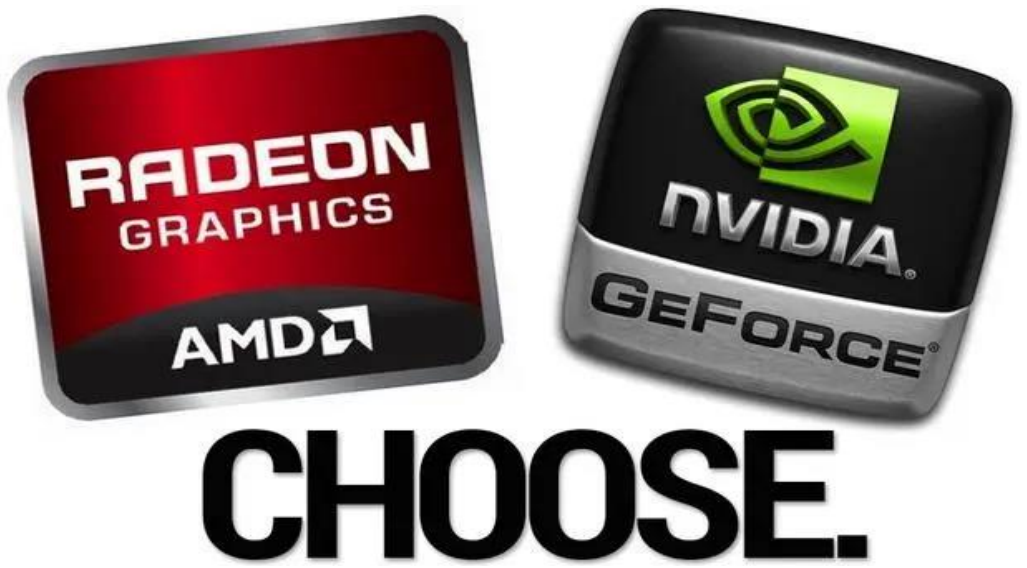
1.3 GPU生态体系：构筑通用计算壁垒

- **规模化商用难：**
  - 要实现规模化商用，就需要厂商实现软硬件技术生态完整部署。由于行业在生态建设上先发优势明显，NVIDIA通过早期与客户企业的平台适配、软件开源合作，较好的用户体验加强了客户粘性，使新的竞争企业难以进行转移。
- **市场认可度：**
  - 市场认可度一方面需要生产商具备生产高品质产品的实力，另一方面需要用户经过一定时间使用形成反馈累积为企业信誉。在GPU市场内，长期以来，AMD和NVIDIA两大产商占据了主要市场份额，也相应形成了较高的市场认可度，客户在选择产品时普遍优先考虑这两大产商的产品；其他厂商的产品，由于未使用过且市场认可度不高，存在较大的不确定性，客户选购意愿较低。

NVIDIA合作伙伴

资料来源：搜狐@计算杂谈

NVIDIA 和 AMD GPU主打产品系列



资料来源：百家号@只说数码科技

1.3 GPU生态体系：以NVIDIA CUDA平台为例

- **CUDA概述：**
  - CUDA是NVIDIA研发的通过利用GPU运算处理的编程、并行计算平台，大幅度提高计算效率。CUDA目前广泛应用于诸多领域，包括CT图像再现、光线追踪、视频处理、计算生物学以及化学等。
- **CUDA平台形成庞大的生态系统几乎占据全部市场**
  - NVIDIA最新生态架构组件包含六大部分：编程语言和API、开发库、分析和调试工具、数据中心工具和集群管理、GPU加速应用程序和GPU与CUDA架构链接。
  - **在通用计算GPU领域的生态几乎是被CUDA生态所占据**。CUDA生态建设难度高、要求复杂。

CUDA生态系统组件

	CUDA工具包/支持应用/GPU	第三方工具链
编程语言和API	PGI 工具包、 C、C++、Fortran、Python	PyCUDA、Altimesh Hybridizer、OpenACC、OpenCL、Alea-GPU
开发库	cuBLAS、cuRAND、cuFFT、cuSPARSE、cuTENSOR、cuSOLVER、nvGRAPH、Thrust、nvJPEG、NPP、光流SDK、NVSHMEM、NCCL、cuDNN、TensorRT、Riva、DALI	OpenCV、FFmpeg、ArrayFire、MAGMA
分析和调试工具	NVIDIA Nsight、CUDA GDB、CUDA-Memcheck	ARM Forge、TotalView Debugger、PAPI CUDA Component、TAU Performance System、VampirTrace
数据中心工具和集群管理	HPC 容器、Kubernetes 、DCGM、NVML API	Bright Cluster、Ganglia、StackIQ 、 Altair PBS Works
GPU与CUDA架构链接	GeForce GPU、Quadro GPU、数据中心 GPU、Tegra	/

资料来源：CSDN@奔跑的小蘑菇，中信证券研究部

1.3 GPU生态体系：以NVIDIA CUDA平台为例

- **CUDA栈组成：**
  - CUDA技术栈由NVIDIA GPU、Operating System、CUDA Driver和CUDA程序组成，其中底层GPU提供硬件支持相关指令运行，操作系统环境和驱动将底层硬件与上层软件（CUDA程序、函数库等）连结。
  - 同时CUDA提供广泛的开发工具和集成：Nsight、Visual Profiler、CUDA MemCheck、CUDA GDB 和 OpenACC等。
- **CUDA软件堆栈主要由CUDA Library、CUDA runtime API和CUDA driver API三层组成**
  - 其核心是CUDA C语言，通过nvcc编译器进行翻译、运行。

CUDA栈组成

主要栈组成	描述
CUDA程序	C Runtime、CUDA库
CUDA Driver	PTX（ISA）、SASS
OS	Linux、Windows、Mac
NVIDIA GPU	GeForce RTX系列、Tesla 系列、Quadro系列、Titan系列

资料来源：CSDN@MoussaTintin，NVIDIA官网，中信证券研究部

CUDA软件堆栈组成

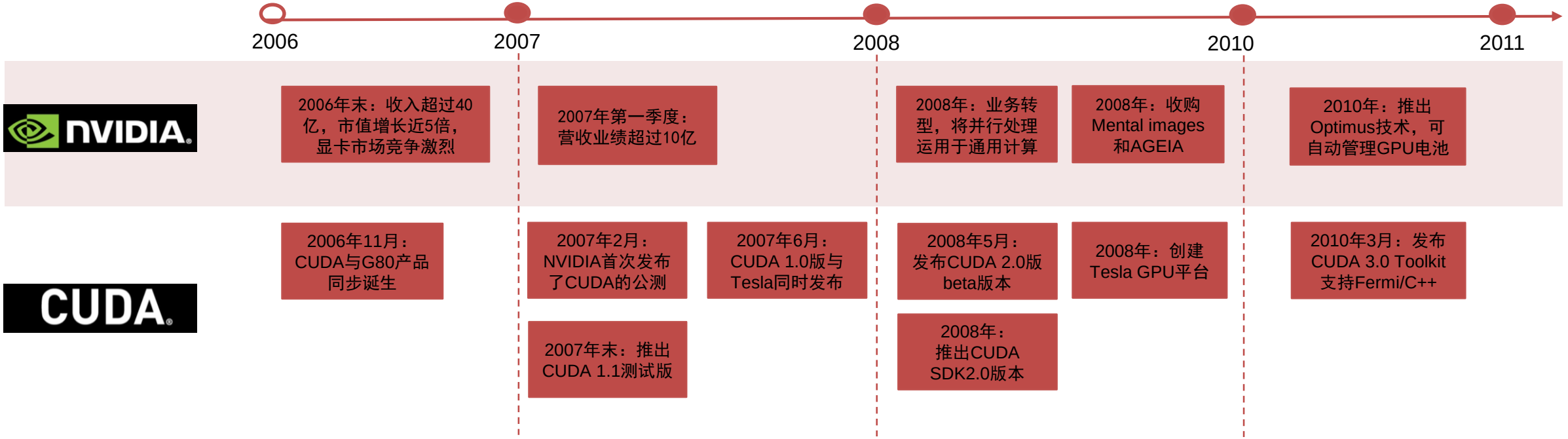
主要软件栈组成	描述
CUDA C语言	C 拓展引入函数类型限符、变量类型限定符和内置变量类型等
Nvcc编译器	作为驱动编译器，能够输出PTX，CUDA二进制序列和标准C
API	包括了运行时（Runtime）API和驱动（Driver）API，实现多种管理、提高互操作性
函数库	提供简单高效的常用函数，包含CUFFT，CUBLAS和CUDPP三个函数库

资料来源：博客园@青竹居士，NVIDIA官网，中信证券研究部

1.3 GPU生态体系：以NVIDIA CUDA平台发展历程为例

- **CUDA萌芽及发展：**
  - NVIDIA CUDA平台最早于2006年初步形成，在2007年正式推出CUDA1.0公测版本。
  - 在2008年-2010年，CUDA平台进一步发展，拓展了新局域的同步指令、扩充全速常量内存并且支持递归，NVIDIA向各软件厂商免费提供开发工具，使得CUDA生态初具规模。
  - 在2008年推出CUDA 2.0版本，2010年推出CUDA 3.0版本。

NVIDIA CUDA平台发展历程



资料来源：NVIDIA官网，中信证券研究部

1.3 GPU生态体系：以NVIDIA CUDA平台发展历程为例

- **CUDA进一步发展与完善：**
  - 随着2011年 CUDA 4.0的推出，标志着NVIDIA在HPC（(High performance computing）即高性能通用计算领域的一大突破。结合GPU Direct2.0技术，实现GPU内部全局统一定址，并且拥有更加完善的C++支持，在性能和协作方面都得到了较大的提升。
  - 在2012-2020年间，NVIDIA基本保持历年推出新一代CUDA平台的频率，对其生态和效率进行完善与升级。
  - 2020年开启CUDA 11.0版本时代，至今已推出至CUDA 11.7版本，能够支持多种并行语言结构，且对CUDA平台内软件进行更新。

NVIDIA CUDA平台发展历程

	2011	2012	2014	2015	2017	2018	2020
NVIDIA	收购ICERA，推出“Project Denver”架构	推出首个NVIDIA GRID虚拟化GPU	发布Tegra K1 SHIELD平板电脑	发布Jetson TX1实现新一代自主机器	推出NVIDIA Isaac助力机器训练	NVIDIA Clara平台赋能智能医疗	在高性能计算、自动驾驶等多领域深耕
CUDA	2011年2月：发布CUDA Tool Kit 4.0版本，支持GPU Direct 2.0技术	2012年7月：发布NVIDIA CUDA 5.0 for MAC	2014年2月：发布 CUDA TOOLKIT V6.0 支持 Windows, Linux, and Mac OS系统	2015年8月：发布 CUDA 7.0 支持 MAC, OS, X 系统	2017年6月：发布CUDA TOOLKIT 8.0.61 支持Windows, Linux, and Mac OS系统	2018年3月：发布CUDA TOOLKIT 9.0版本 2018年9月：开放CUDA Toolkit 10.0 Archive下载	2020年：基于Ampere架构发布CUDA 11.0，支持 Windows, Linux, and Mac OS系统

资料来源：NVIDIA官网，中信证券研究部

1.3 GPU生态体系：以AMD ROCm平台为例

- **ROCm概述：**
  - 全称为Radeon Open Computing platforM，是基于AMD GPU系列开源设计的计算生态，其目标是建立与NVIDIA CUDA生态可替代的平台，构建开放式软件平台，提供出色灵活性和卓越性能，让开源计算语言、编译器、库和工具助力高性能计算和人工智能社区代码开发。
- **ROCm与CUDA对比：**
  - 为实现对CUDA平台的可替代性，ROCm复制了CUDA的技术栈，涵盖HIP程序、库、 Runtime、PTX、OS等。
  - ROCm作为开源平台，提供开发标准支持，并且封装层次相较CUDA更优，对一般开发者不开放。

ROCm平台与CUDA平台模块对比

CUDA	ROCm	备注
CUDAAPI	HIP	C++ 扩展语法
NVCC	HCC	编译器
CUDA函数库	ROC库、HC库	/
Thrust	Parallel STL	HCC 原生支持
Profiler	ROCm Profiler	/
CUDA-GDB	ROCm-GDB	/
DirectGPU RDMA	ROCn RDMA	peer2peer
TensorRT	Tensile	张量计算库
CUDA-Docker	ROCm-Docker	/

资料来源：CSDN@MoussaTintin，AMD官网，中信证券研究部

1.3 GPU生态体系：以AMD ROCm平台为例

- **ROCm优势：**
  - 1）扩大支持和访问范围。支持AMD Instinct MI210 和AMD Radeon Pro W6800的工作站GPU。2）性能优化。FP64矩阵操作能够更好地进行高速缓存处理，以及改善内核启动延迟和运行时间。3）助力开发者研发。提供、预包装的HPC和AI/ML框架，可随时在AMD Infinity Hub上下载。4）易于获取资源。在ROCm信息门户、AMD 社区支持下，能够远程读取AMD加速器云 (AAC)，用于开发、测试和基准测试。
- **ROCm模块：**
  - 面向任意一种工作负载，ROCm堆栈都包括部署和管理工具、优化库以及编程和系统工具。其中，系统工具包括编译、调试、性能分析和系统管理等。

ROCm平台模块介绍

模块	描述
部署和管理工具	部署和管理工具简化了部署和运行HPC和ML代码的过程，包括验证套件以确保设备环境能够承载软件运行。ROCm Data Center Tool有助于收集作业遥测和统计信息。包括与第三方工具进行适配；同时能够监测如温度等环境因素，进一步支持AMD系统管理接口
架构	架构覆盖关键行业和应用，包括对HPC 和ML软件
库	库包括对数学函数、分布式计算的支持，以及容器和扩展通信。
编程模型	编程模型包括OpenMP、HIP和 OpenCL，以及帮助操作人员编译、运行、配置和调试软件的工具。ROCm支持C/C++，并提供可以自动将CUDA软件转换为HIP的转化工具，即可移植的异构计算接口，使其具有通用性。
设备驱动	设备驱动和运行时环境支持Red Hat Enterprise Linux 、SUSE Linux Enterprise Server Distribution和Ubuntu Linux。ROCm的优势是，供应商能够很容易地为他们的加速器创建设备驱动程序，从而扩大了平台的使用范围和多样性。包括工作站和数据中心级的GPU加速器。
GPU支持	GPU支持包括范围广泛的AMD Radeon和Instinct 加速器，同时开放支持第三方GPU和 FPGA设备

资料来源：CSDN@MoussaTintin，AMD官网，中信证券研究部

2.他山之石：Nvidia/AMD竞争启示—架构创新升级和新兴领域前瞻探索是主旋律

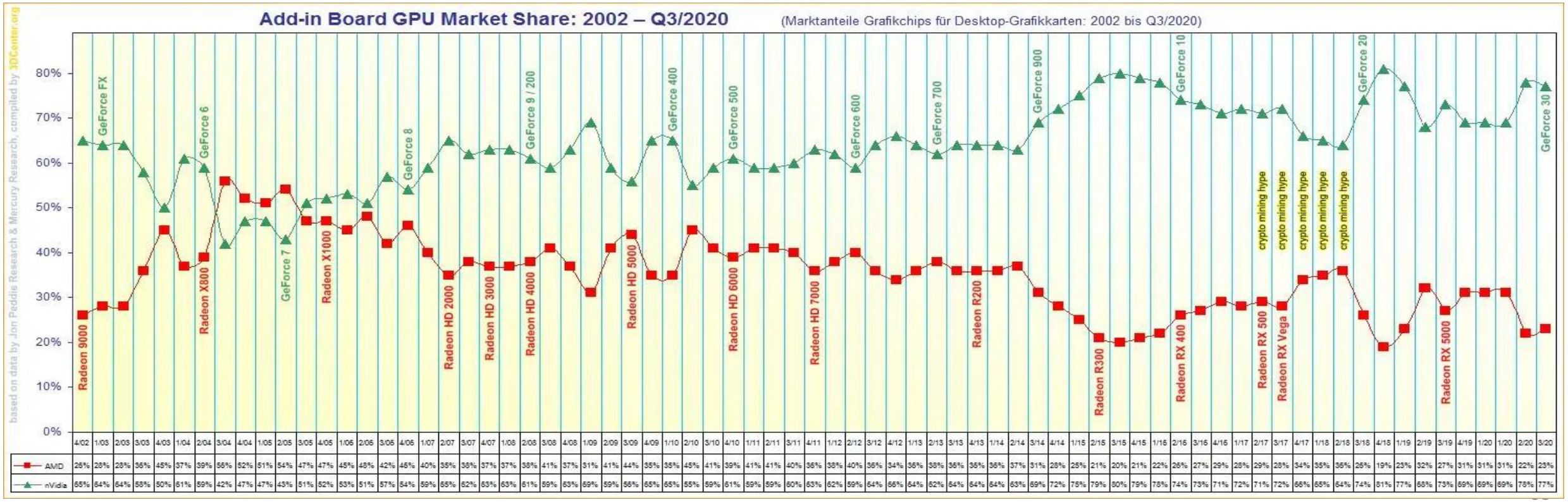
I. NVIDIA、AMD（ATI）的产品迭代一览

II. GPU行业竞争史：架构创新升级和新兴领域前瞻探索是领跑GPU行业的关键

2.1 总览：NVIDIA经历风雨遥遥领先，AMD（ATI）再显峥嵘

- **总体而言，NVIDIA引领GPU行业发展数十年，大多数时期技术和市场份额均领先；AMD（ATI）长期与保持NVIDIA竞争，N/A卡之争愈演愈烈。**
  - NVIDIA市场份额虽有波动，但长期高于50%，与其产品性能优势和生态构建优势密不可分。2006年起，英伟达GPU架构保持约每两年更新一次的节奏。在这一节奏下，英伟达代际之间产品性能提升显著，生态构建完整，Geforce系列产品市占率长期超过Radeon镭龙系列，NVIDIA牢牢把握市场龙头地位。2019年后，AMD凭借RDNA架构再度崛起。

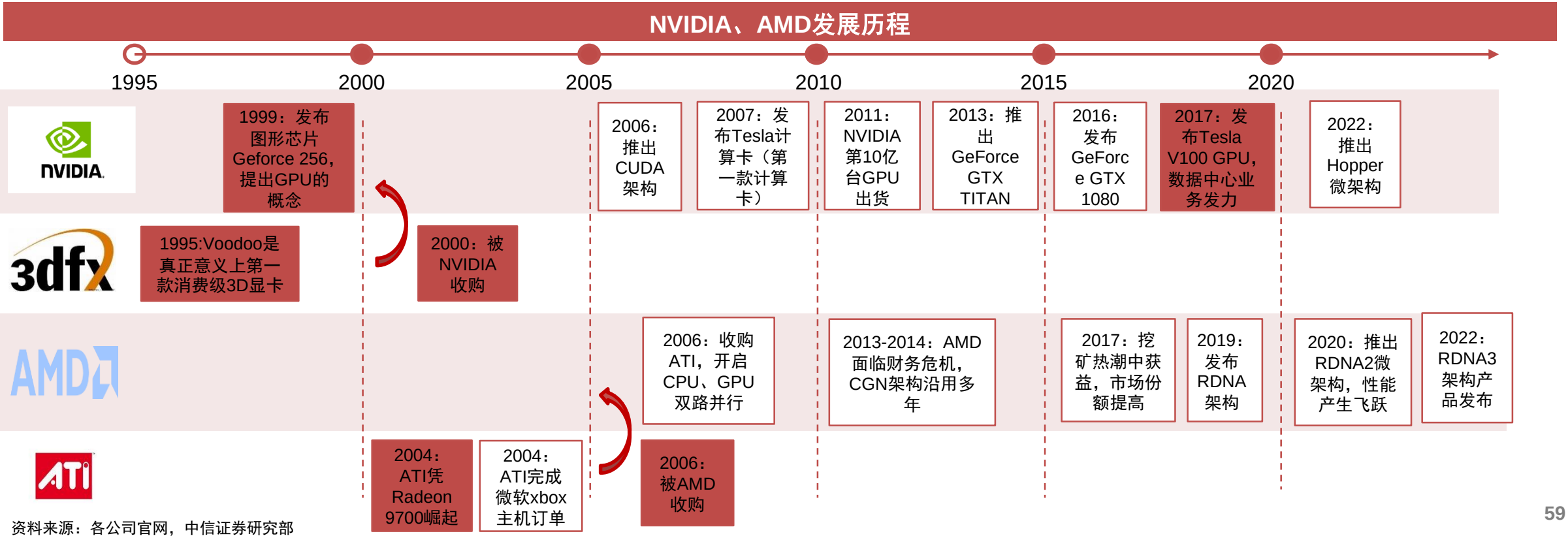
2002-2020年全球GPU市场份额



资料来源：JPR，快科技

2.1 NVIDIA、AMD（ATI）等企业构筑GPU发展主旋律

- **自1999年NVIDIA提出GPU概念，GPU已经有20余年发展历史**
  - 1995年，3Dfx发布第一款消费级3D显卡，拉开图形处理芯片的发展序幕。1999年，NVIDIA提出GPU概念，奠定其GPU行业霸主地位，自此AMD、ATI、3Dfx等企业与NVIDIA合力推动GPU快速发展。
- **NVIDIA率先构筑通用计算的CUDA生态，引领GPU的行业革命；AMD（ATI）CPU、GPU双线并行紧随其后**
  - 如今人工智能高速发展，几乎应用于各行各业，GPU是目前应用最广的AI芯片。NVIDIA把握游戏、数据中心市场机遇；AMD加速提升架构性能紧随其后。



资料来源：各公司官网，中信证券研究部

2.1 NVIDIA：把握图形、数据中心历史机遇，驱动业绩快速增长

1999年至今，NVIDIA GeForce 系列不断更新

- GeForce系列显卡经过二十多年的发展，产品已经涵盖不同价位、不同应用领域的低、中、高端图形显示和通用计算，是NVIDIA主力产品。
- 最新产品代际下NVIDIA已经在2022年9月20日推出GeForce 40系列首款产品。

NVIDIA GPU产品与工艺演进

• GPU重要产品 • 关键技术变化	• GeForce 2系列 • GeForce 3 系列 • 首款可编程GPU	• GeForce 4（NV25）全系列	• GeForce FX系列 • 可编程图形硬件诞生 • SoC产品Tegra移动处理器	• GeForce 6（NV40）全系列 • SLI技术允许多个GPU相连	• GeForce 7 系列	• GeForce 8 系列 • GeForce 8800 GTX先支持统一渲染 • 通用计算CUDA架构	• GeForce 9 系列 • SIMT执行模型 • 第一款计算卡Tesla C870	• GeForce 9 系列 • Iray 渲染软件	• GeForce 100 系列 • GeForce 200 系列	• GeForce 300 系列 • GeForce 400 系列 • 笔记本电脑Optimus技术	• GeForce 400 系列 • 首款双核移动SoC Tegra 2
工艺制程	150nm		130 nm		110nm	90nm		55nm		40nm	
改进战略	优化	架构	工艺	架构	架构	工艺、优化	架构	制程	架构	工艺	架构
核心微架构	Kelvin		Rankine	Curie		Telsa			Fermi		
年份	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011

资料来源：NVIDIA官网，中关村在线，CSDN@凌凌1301，中信证券研究部

2.1 NVIDIA：把握图形、数据中心历史机遇，驱动业绩快速增长

- **2006年，NVIDIA推出CUDA，为GPU通用计算奠定基础；目前其在数据中心领域业务占比已赶超游戏业务**
  - NVIDIA数据中心业务自2017年开始快速扩张，先后发布V100、A100等高性能通用计算显卡，为全球提供顶尖的AI算力。
  - 短短4年时间，其数据中心业务占比已经从2017年的19%增长至2021年的45%，现已超过传统游戏业务占比。

NVIDIA GPU产品与工艺演进

• GPU重要产品 • 关键技术变化	• GeForce 600 系列 • 第一个虚拟化 GPU GRID	• GeForce 700 系列 • 高性能 Geforce GTX TITAN	• GeForce 700 系列 • 嵌入式 AI平台 Jetson	• GeForce 900 系列 • 专为深度神经网络打造的 Geforce GTX TITAN X 发布	• 人工智能车辆计算平台 NVIDIA DRIVEPX 2	• 发布 GV100，加入 Tensor Core	• GeForce 20 系列 • 光线追踪技术突破	• GeForce 16 系列	• GeForce 30 系列 • A100显卡 • 兼容IEEE的 FP64 Tensor Core • DLSS 2 技术	• GeForce 30 系列 • NVIDIA Omniverse	
工艺制程	28nm				16nm	12nm			7nm/三星8nm		4nm
改进战略	工艺	优化	架构	优化	架构&工艺	工艺	优化	架构	工艺	架构	-
核心微架构	Kepler		Maxwell		Pascal	Volta	Turing		Ampere		Hopper / Ada lovelace
年份	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022

资料来源：NVIDIA官网，中关村在线， CSDN@凌凌1301,中信证券研究部

2.1 AMD：ATI时代开端奠定市场基础

- **AMD显卡发展可大致划分为两阶段：第一阶段ATI时代从1985年至2006年，第二阶段从2006年至今为AMD时代—Radeon系列持续迭代更新**
  - 自早期开始，AMD分为两路研发，兼顾高端显卡市场和低端显卡市场，其中，高端产品如Radeon 8500、Radeon X1800 XT等；从高端显卡中衍生出多款低端显卡产品，包括Radeon 9000、9000 Pro、9100、9200以及9250。

AMD GPU产品与工艺演进

• GPU重要产品 • 关键技术变化	• Radeon256 • ATI显卡开端 • Radeon 7200 DDR	• 高端显卡Radeon 7200 • 低显卡的Radeon 7000	• 首款完全支持DX 8.1的显卡：Radeon 8500 • Radeon9000	• Radeon 9800	• Radeon 9500 & Radeon X800	• 全面开启CrossFire的一代：Radeon X1800 XT	• AMD收购ATI • Radeon X 1950 Pro	• Radeon HD 2900 XT • Radeon HD 3870 X2	• Radeon HD 4870	• 最后一代ATI标志的显卡：Radeon HD 5870	• AMD首款显卡Radeon HD 6970
工艺制程	180nm		150 nm		110 nm	90nm	80nm	55nm		40nm	
改进战略	优化	优化	优化、制程	优化	优化、制程	制程、优化	制程、优化	制程、优化	优化	制程	架构
核心代号	R100	RV100	RV250	RV350	R300	RV520	RV515	RV670	RV770	RV870	TeraScale 3
年份	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010

资料来源：超能网@鲲鹏，中信证券研究部

2.1 AMD：AMD时代再续辉煌，架构、制程多点突破

■ 2012年以来，AMD在架构上保持创新态势，制程引领行业先进性

- 2012年发布Radeon HD系列，AMD在架构上实现创新，推出GCN架构，并且是业界第一款采用28纳米工艺制程的GPU图形芯片。在图形渲染和通用计算领域性能均领先市场内竞争对手。
- 2019年，AMD推出RDNA架构，同时兼容原有GCN架构，在性能、功耗、能效等多方面实现超越，正式开启第五代架构革新之路。

AMD GPU产品与工艺演进

• GPU重要产品 • 关键技术变化	• Radeon HD 7970	• Radeon R9 290X	• Radeon R9-295X2	• 首次使用HBM显存 • Radeon R9 Fury X	• Radeon RX 480	• Radeon RX Vega系列 • Radeon RX 500系列	• Radeon Vega 系列 • 第二代Threadripper处理器	• Radeon VII • RX 5000系列 • Radeon RX 5500系列	• Radeon RX 6800系列 • Radeon RX 6000	• Radeon RX 6600 • Radeon RX 6000M系列	• Radeon RX 7000系列
工艺制程	28nm				14nm		7nm			6nm	5nm
改进战略	制程、架构	架构、优化	架构、优化	优化	制程、架构	优化	制程、优化	架构	架构、优化	制程	制程、架构
核心微架构	GCN架构		Hawaii	GCN2.0	Polaris架构	VEGA架构		RDNA1	RDNA2		RDNA3
年份	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022

资料来源：超能网@鲲鹏、@PC物语，中信证券研究部

2.2 1962-1995年：图形处理技术不断发展，3Dfx凭Voodoo一枝独秀

- **1962年起，计算机图形学不断发展，图形处理技术实现从2D到3D的突破**
  - 1962年麻省理工学院博士伊凡•苏泽兰奠定了计算机图形学基础；1984年，SGI公司推出了面向专业领域的高端图形工作站，俗称图形加速器，是首个专门的图形处理硬件。
  - 1994年，3D Labs发布GLINT 300SX， 是PC最早的3D硬件加速图形芯片，从此开启3D显卡时代。
- **1995年，3Dfx发布Voodoo图形芯片组配和Glide API接口，一度统治市场**
  - Glide是3Dfx为Voodoo打造的底层3D API，是第一个在PC游戏领域得到大范围使用的程序接口，使得Voodoo无须硬件厂商额外提供API就可以直接开发游戏，具有易用性和稳定性。NVIDIA同期的riva 128性能与其有差距。
  - 当时的顶级游戏和部分PC游戏基本都支持Glide。因此即使Voodoo的价格远远高于市场上其他产品，也深受消费者追捧。

Voodoo系列显卡

资料来源：太平洋电脑网

图形处理发展史

1962	1984	1987	1994	1995
麻省理工学院博士伊凡•苏泽兰奠定了计算机图形学基础	SGI公司推出首个专门的图形处理硬件	IBM发布第一台VGA(视频图形阵列)，图形处理进入2D时代	1994年，3D Labs发布GLINT 300SX，开启3D显卡时代	3dfx公司发布图形加速卡Voodoo，是真正意义第一款消费级3D显卡

资料来源：百家号@爱集微APP，中信证券研究部

2.2 1996-2000年：Nvidia依靠性能优势击败3Dfx，3Dfx盛极而衰

- **3Dfx Voodoo系列后续产品被NVIDIA反超，开始由盛转衰**
  - 1996年，3Dfx凭借Voodoo成为全球3D显卡和GPU制造领域的垄断者。1997年，NVIDIA推出的NV 3（RIVA 128）有128bit的2D、3D加速图形核心，采用0.35微米工艺，支持微软Direct 3D接口，且性价比高于Voodoo，被OEM厂商广泛使用。
  - 1998-1999年，NVIDIA推出NV4性能击败Voodoo3，随后3Dfx的Voodoo4延迟发布、Voodoo5由于能耗大、散热高败给NVIDIA。
- **1999年8月，NVIDIA公司发布图形芯片Geforce 256，首次提出GPU的概念**
  - Geforce 256采用技术包括硬件变换、“T&L”、立方环境材质贴图和顶点混合、凹凸映射贴图、双重纹理四像素256位渲染引擎、纹理压缩等，兼容Direct X和Open GL，被称为世界上第一款GPU。此前如顶点变换必须在CPU中完成，光栅化后像素有限等，而GPU将这些功能独立出来，使显示核心与CPU并列成为计算机核心，大大减少CPU的运算压力。

3Dfx性能落后Nvidia

资料来源：百家号@南京1号电脑超市

Nvidia Geforce 256

资料来源：搜狐@微型计算机杂志

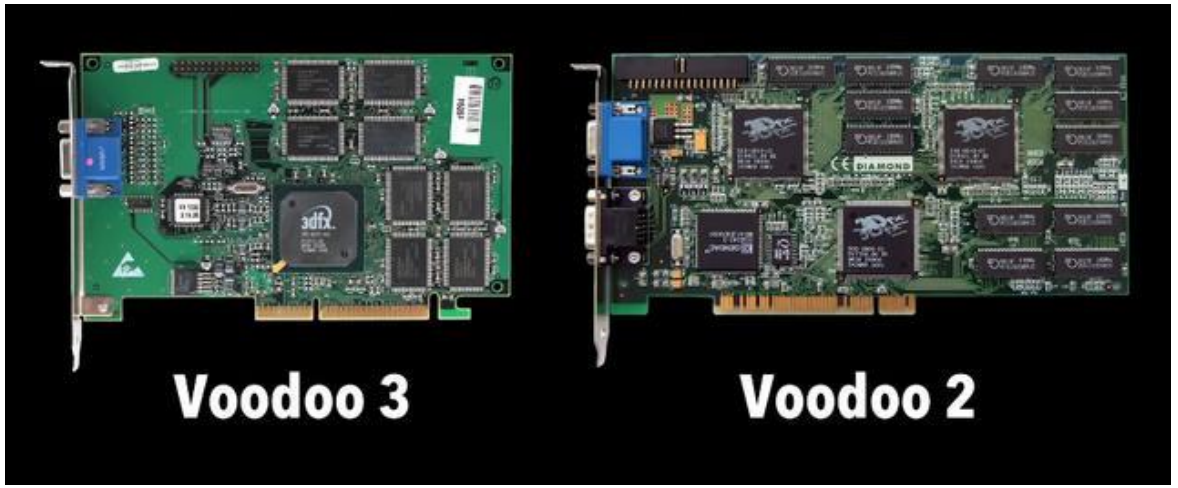
2.2 1996-2000年：Nvidia依靠性能优势击败3Dfx，3Dfx盛极而衰

- **1999年，NVIDIA崛起，击败并收购难以为继的3Dfx**
  - 1999年，NVIDIA的Geforce 256奠定了NVIDIA在GPU市场的领先地位。与此同时的3DFX由于一系列决策失误，由盛转衰。
  - 2000年12月15日，Nvidia低价收购3Dfx图形处理相关所有产业。
- **3Dfx的失败可以归因于战略和产品策略问题**
  - 战略失误：芯片厂商涉足板卡制造领域面临困难。3Dfx收购板卡制造商STB，希望独自生产Voodoo显卡，但二者的合作并没有表现出1+1>2的效果，反而拖慢新品发布进度，令3Dfx丢掉不少市场份额。同时失去原有板卡厂商合作伙伴，DIAMOND、GIGABYTE、CREATIVE、ELSA都加入NVIDIA阵营。
  - 产品策略问题：Voodoo3与Voodoo2相比性能进步很小，只是换了马甲；Voodoo4和Voodoo5不支持硬件转换和TV输出功能，失去了DVD和家庭影院市场；Voodoo4和Voodoo5不支持DDR内存，而自身适配的SDRAM在性价比上输给NVIDIA的DDR内存，再次流失市场份额。

3Dfx与NVIDIA阵营

资料来源：搜狐@南京一号电脑超市

Voodoo2与Voodoo3对比

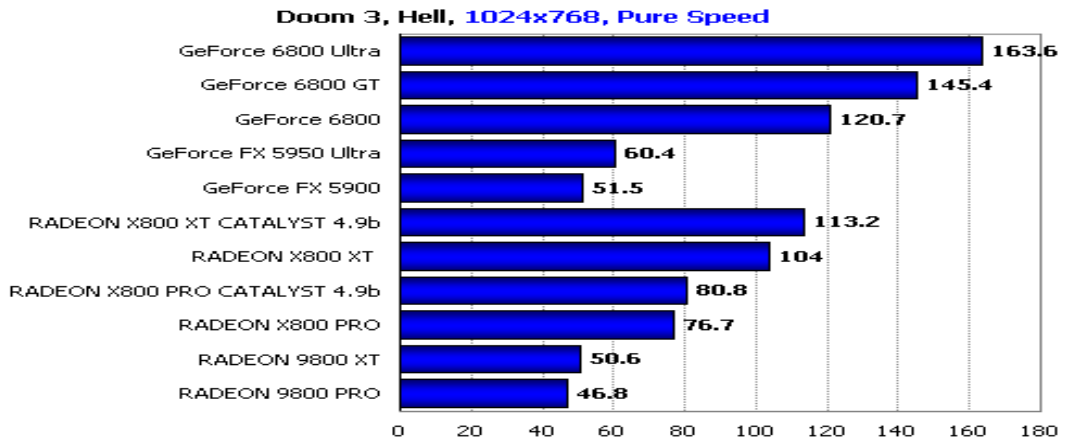


资料来源：百家号@南京1号电脑超市

2.2 2000-2004年：ATI凭借Radeon系列与NVIDIA分庭抗礼

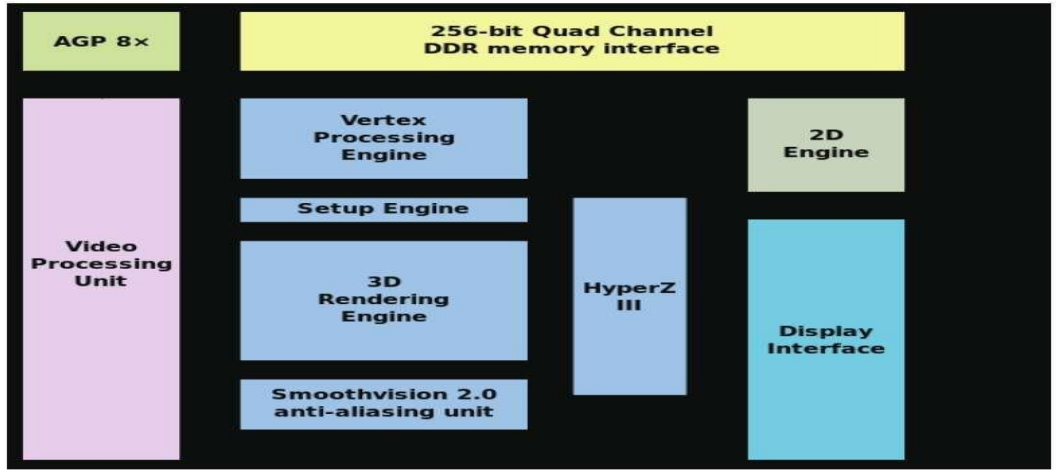
- **2000年显卡市场格局洗牌，ATI凭Radeon 9700强势崛起**
  - 2000年，ATI发布Radeon 256，180nm工艺，内有3000万颗晶体管，具备在当时属于先进技术的几何变形、图像剪切功能、光照效果，性能优于Nvidia同代的Geforce 256。自此，PC端独立显卡市场形成Radeon系列与Geforce系列对峙的局面。
  - 2002年，ATI发布R 300（即Radeon 9700）支持DirectX9.0、4顶点着色器、8像素流水线、256位DDR内存总线；2003年发布Radeon 9800pro，性能均超过Nvidia的Geforce FX5900。ATI逐步站稳脚跟。随后NVIDIA的Geforce FX6800又在性能反超Radeon 9800。在此之后，ATI真正与Nvidia在GPU市场平分秋色，二者产品性能相互追赶。
- **NVIDIA遇强力对手，市场份额跌破50%；ATI获微软XBOX 2主机图形芯片订单**
  - 在ATI Radeon 9700和XBOX 2 订单的帮助下，ATI市场份额最高达到55%，而NVIDIA市场份额跌破50%，为NVIDIA迄今为止最低点。
  - 微软和Nvidia共同研发微软第一代XBOX的图形处理器芯片，而2003年ATI获得第二代XBOX的图形处理器订单，股价因此由上涨1美元到13.2美元，并在2004年顺利完成该订单显卡的开发工作，市场份额有所上涨。

Geforce FX6800性能反超Radeon 9800



资料来源：中关村在线

ATI R300架构

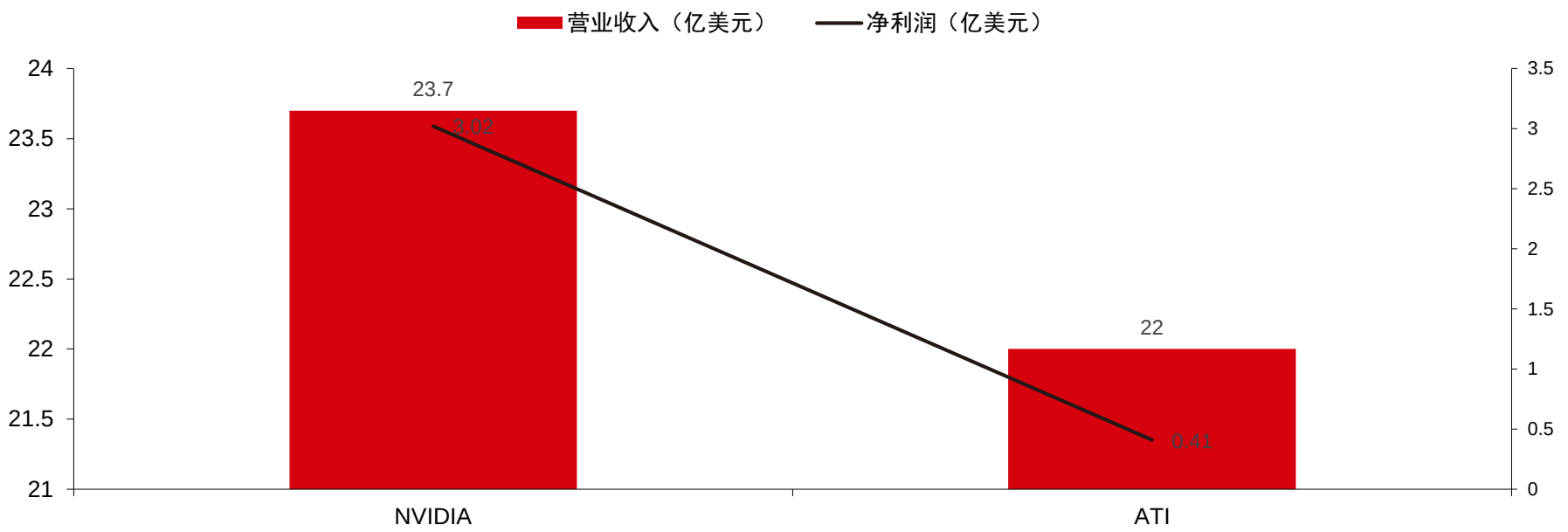


资料来源：IIS7站长之家@钱纲

2.2 2004-2006年：ATI被AMD收购，NVIDIA重回领先位置

- **NVIDIA稳定推新，ATI并未在后续产品上坚守住阵地，同时净利润开始下滑**
  - ATI后面推出的X300、X550、X600、X700、X1600性能落后于同期英伟达产品，竞争处于下风。
  - 2005年，ATI、NVIDIA交替发布新产品，ATI年收入达到22亿美元，但净利润仅不到5000万美元，同期NVIDIA营业收入约23亿美元，净利润超3亿美元。
- **2006年，AMD以54亿美元收购ATI**
  - 2006年，AMD为弥补独立芯片组的欠缺，以54亿美元收购ATI。AMD也因此背负巨额外债。

2005年NVIDIA、ATI收入利润对比

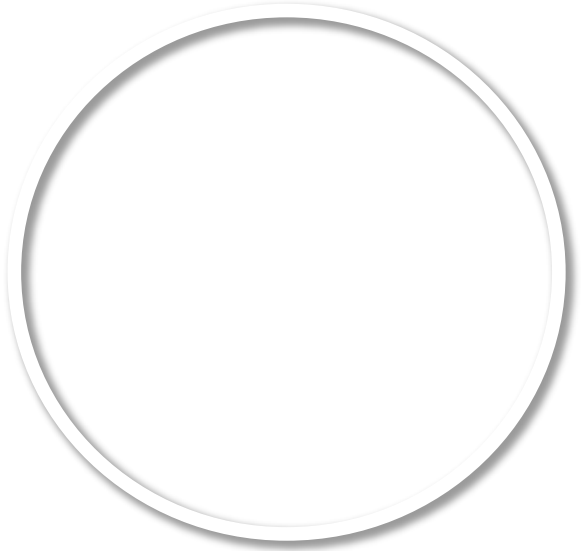


资料来源：新浪网，中信证券研究部

2.2 2006-2012年：Nvidia架构快速更新迭代，开创通用计算生态先河

- **自2006年起，英伟达GPU架构保持约每两年更新一次的节奏，代际之间产品性能提升显著，性能和市场份额均领先AMD。**
- **2006年，英伟达推出了CUDA编程软件，使GPU成为通用并行数据处理加速器，并逐步构筑起CUDA生态。**
  - CUDA让显卡可以用于通用并行计算和其他非图形计算，使得GPU能够承担和CPU一样的计算任务。程序员可以通过CUDA直接对GPU进行编程。为NVIDIA的数据中心业务高速扩张打下基础。
  - CUDA包括硬件平台和软件栈（软件集合）两层含义，加上第三方应用及工具的扩展，形成从开发到应用的CUDA生态。CUDA生态也成为NVIDIA的生态护城河。

CUDA生态

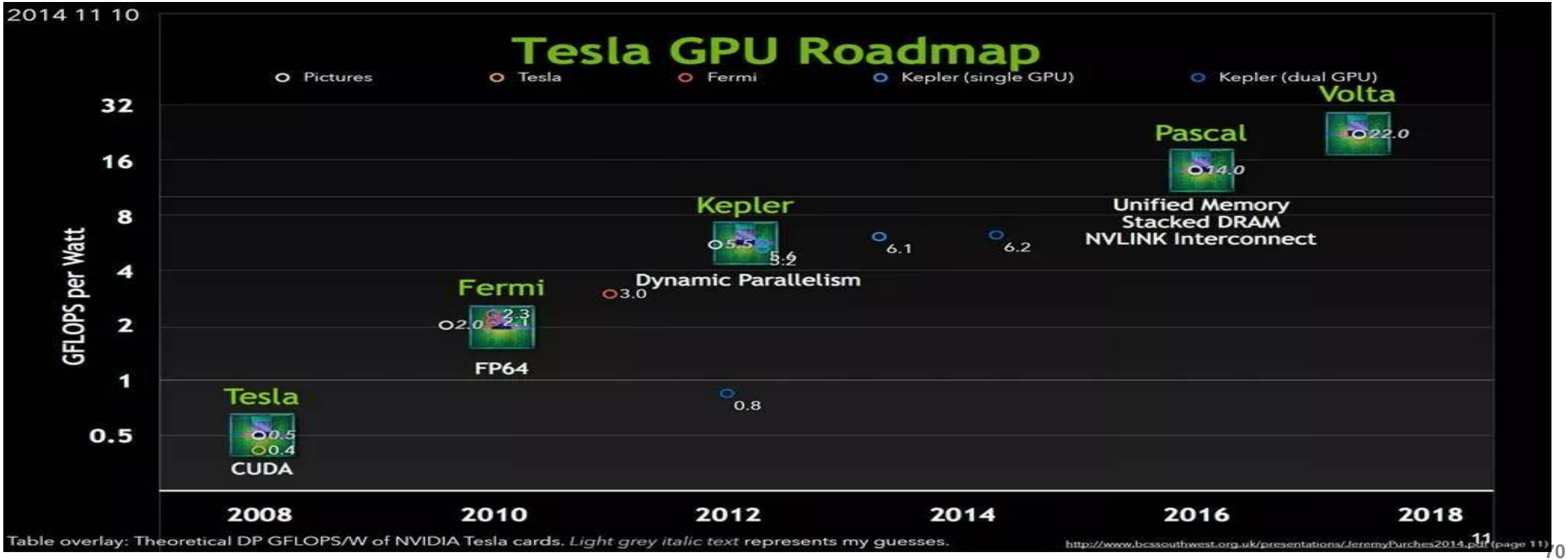


资料来源：CSDN@普通网友，中信证券研究部

2.2 2006-2012年：Nvidia架构快速更新迭代，开创通用计算生态先河

- **2007年，英伟达发布Tesla计算卡，标志用于计算的GPU产品线正式独立；**
  - Tesla架构是第一代真正用于并行运算的GPU架构，今天的并行计算架构中仍有很多该架构硬件设计的影子。随后NVIDIA的通用计算架构仍保持大约两年一代的进度进行升级迭代，2010年发布Fermi架构，2012年发布Kepler架构。这一行为标志着GPU在通用计算和超级计算领域开始逐渐取代CPU成为主角。

Nvidia通用计算架构迭代



资料来源：搜狐@电脑爱好者

2.2 2013-2019年：Nvidia性能遥遥领先，数据中心业务快速增长

- **2013-2014年，AMD（ATI）产品研发进展缓慢，NVIDIA性能持续领先**
  - 2013年，AMD GPU仍然沿用GCN架构，使得R9 290X功耗高于对手。并且AMD在新产品推出上后继乏力。2015年，AMD推出的Fury X功耗比优于NVIDIA kepler系列，但架构仍未升级；同时Fury 2X的延迟发布使得大众对AMD的信心下降。
  - 2014年，NVIDIA推出Maxwell架构，使得GeForce GTX在性能、图形和效率方面取得突破性进展，NVIDIA持续掌握GPU市场的主动权。
- **2014-2016年，NVIDIA GeForce GTX 1080带来市场的全面领先，AMD（ATI）仍在苦苦支撑。**
  - 2016年5月，英伟达推出了采用16纳米FinFET制程的Pascal架构，核心频率与上代相比显著提升，超频突破2GHz。GeForce GTX 1080采用Pascal架构。不到一年以后又推出GeForce GTX 1080Ti，虽然架构没有改变，但拥有3584个流处理器、224个纹理单元，游戏性能与1080相比提升约35%。至此，NVIDIA完成了10系显卡从入门款到旗舰款的全型号覆盖。

NVIDIA、AMD研发投入对比

资料来源：智能电视网

GeForce GTX 1080性能参数对比

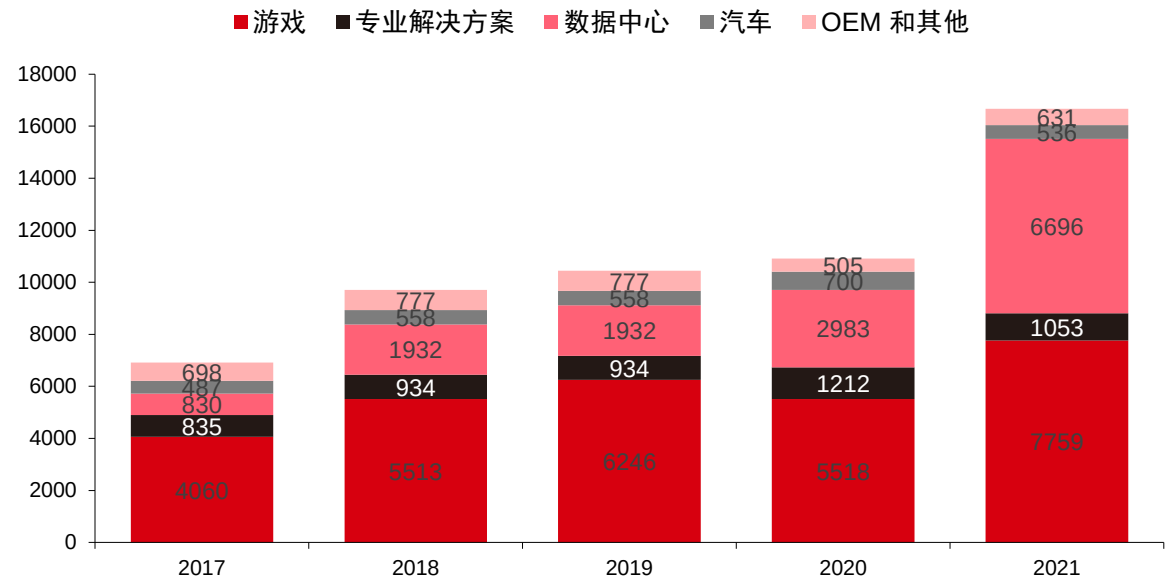
NVIDIA GPU Specification Comparison				
	GTX 1080 Ti	NVIDIA Titan X	GTX 1080	GTX Titan X
CUDA Cores	3584	3584	2560	3072
Texture Units	224	224	160	192
ROPs	88	96	64	96
Core Clock	?	1417MHz	1607MHz	1000MHz
Boost Clock	1600MHz	1531MHz	1733MHz	1075MHz
TFLOPs (FMA)	11.5 TFLOPs	11 TFLOPs	9 TFLOPs	6.6 TFLOPs
Memory Clock	11Gbps GDDR5X	10Gbps GDDR5X	10Gbps GDDR5X	7Gbps GDDR5
Memory Bus Width	352-bit	384-bit	256-bit	384-bit
VRAM	11GB	12GB	8GB	12GB
FP64	1/32	1/32	1/32	1/32
FP16 (Native)	1/64	1/64	1/64	N/A
INT8	4:1	4:1	?	?
TDP	250W	250W	180W	250W
GPU	GP102	GP102	GP104	GM200
Transistor Count	12B	12B	7.2B	8B
Die Size	471mm2	471mm2	314mm2	601mm2
Manufacturing Process	TSMC 16nm	TSMC 16nm	TSMC 16nm	TSMC 28nm
Launch Date	03/2017	08/02/2016	05/27/2016	03/17/2015
Launch Price	$699	$1200	MSRP: $599 Founders $699	$999

资料来源：爱搞机网站

2.2 2013-2019年：Nvidia性能遥遥领先，数据中心业务快速增长

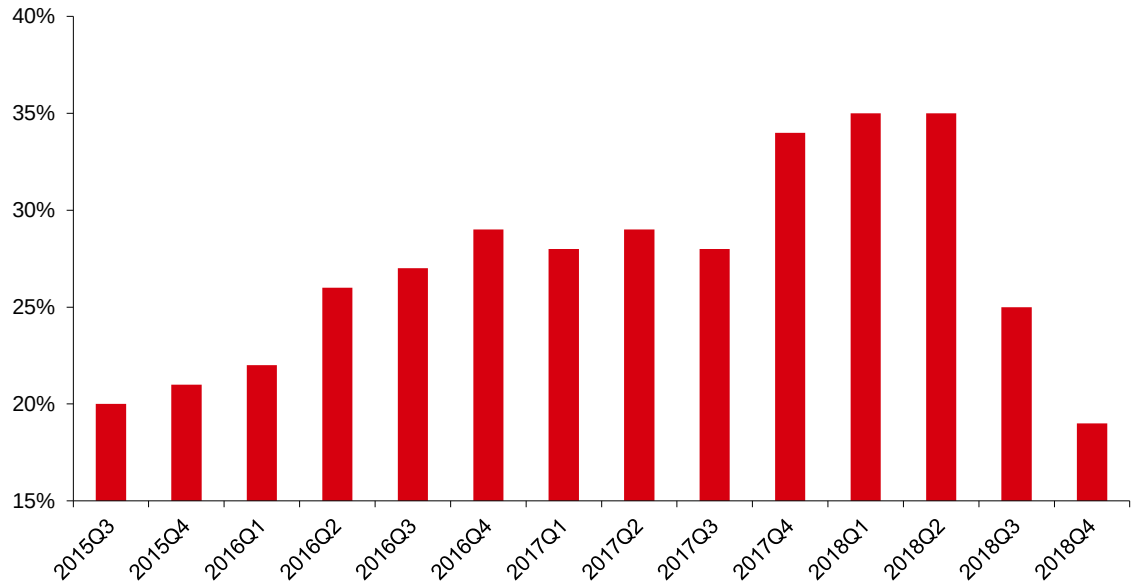
- **2016-2019年，NVIDIA性能始终保持领先并不断拓展业务边界，数据中心业务开始发力**
  - NVIDIA押注AI芯片，2017年发布专为数据中心和高性能计算打造的Tesla V100 GPU，采用Volta架构，有超过210亿个晶体管，是上代Tesla P100的1.37倍。数据中心业务自此开始快速增长，随后成为拉动NVIDIA营收增长的重要力量。
- **2017年挖矿热潮中，由于A卡浮点性能更好，AMD市场份额有所提升**
  - 在2017年左右兴起的挖矿热潮中，AMD显卡由于浮点性能更好而更受青睐。2015年末、2016年末、2017年末，AMD在GPU领域的市场份额逐年回升，分别为21%、29%、34%。2018年虚拟货币价格暴跌，比特币全年跌幅超过70%，二手市场充斥大量低价显卡，GPU的出货量受到一定影响。

NVIDIA 2017-2021年英伟达全球营收（百万美元）



资料来源：Statista，中信证券研究部

AMD 2016-2019年GPU市场份额

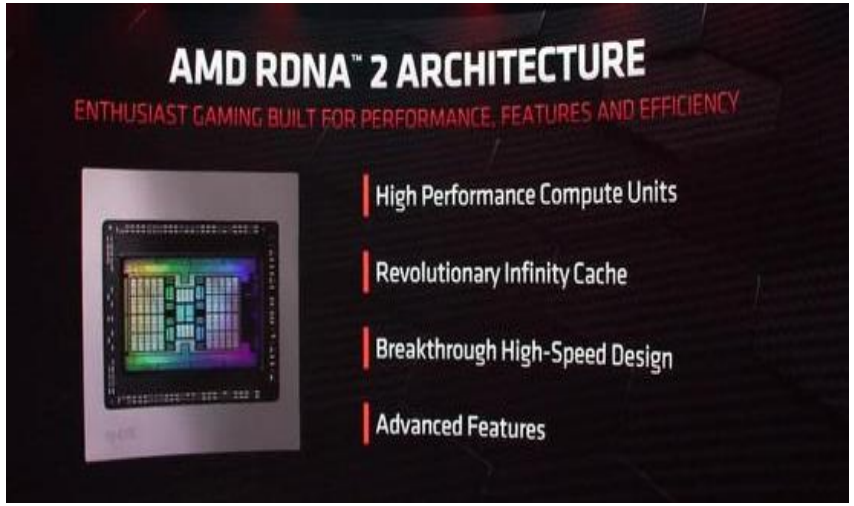


资料来源：3D Center，中信证券研究部

2.2 2019年至今：AMD借助RDNA架构再度崛起，NVIDIA、AMD瓜分GPU市场

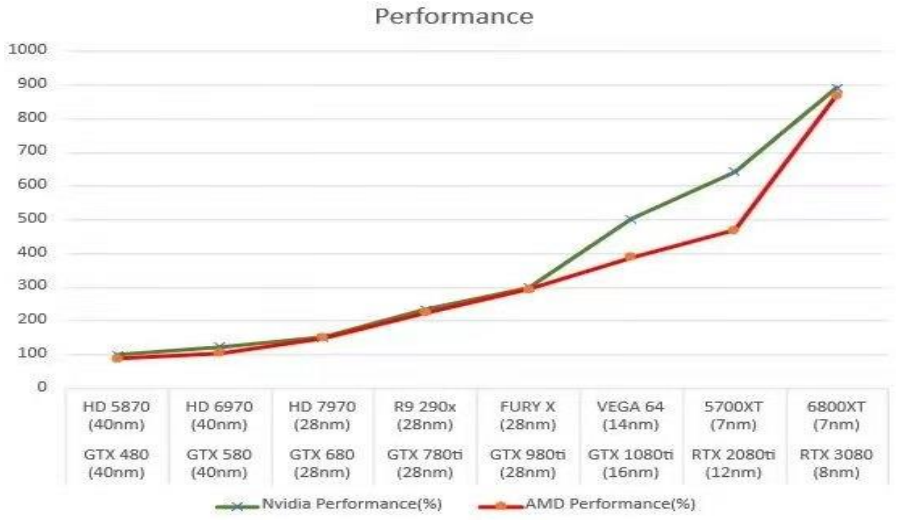
- **2019年发布RDNA架构产品Radeon RX 5700再显峥嵘**
  - AMD产品开始在性能方面追赶NVIDIA。Radeon RX 5700系列采用Radeon DNA架构，即RDNA架构，其完全兼容GCN架构的指令。Radeon RX 5700还采用7nm工艺、GDDR6显存、PCI-e 4.0总线，使得其性能跑分超过NVIDIA的GeForce GTX 1080。
  - 2020年发布的RDNA 2架构又实现性能提升1倍、能效提升至少50%、完整支持DX12U和光线追踪等目标。RDNA 3架构于已于2022年11月推出。
- **NVIDIA、AMD两大巨头瓜分GPU市场，NVIDIA仍保持明显优势**
  - NVIDIA成功拓展AI业务，股价自2015年以来增长超过40倍，AMD一直与其竞争，但短期很难战胜NVIDIA。
  - 3D Center数据显示，2022Q2 NVIDIA在独立GPU的市场份额为79%，AMD则占20%的市场份额，合计99%。Intel凭借在PC端的优势占据剩下1% 的市场份额。

RDNA 2架构升级



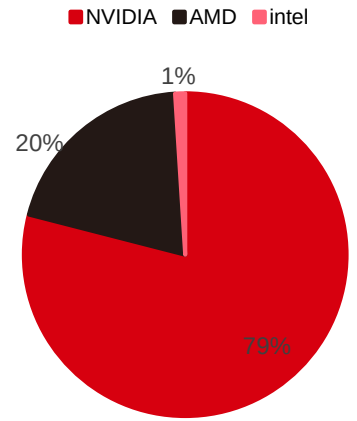
资料来源：超能网

AMD、NVIDIA十年产品性能比较



资料来源：机锋网

2022Q2独立GPU市场份额



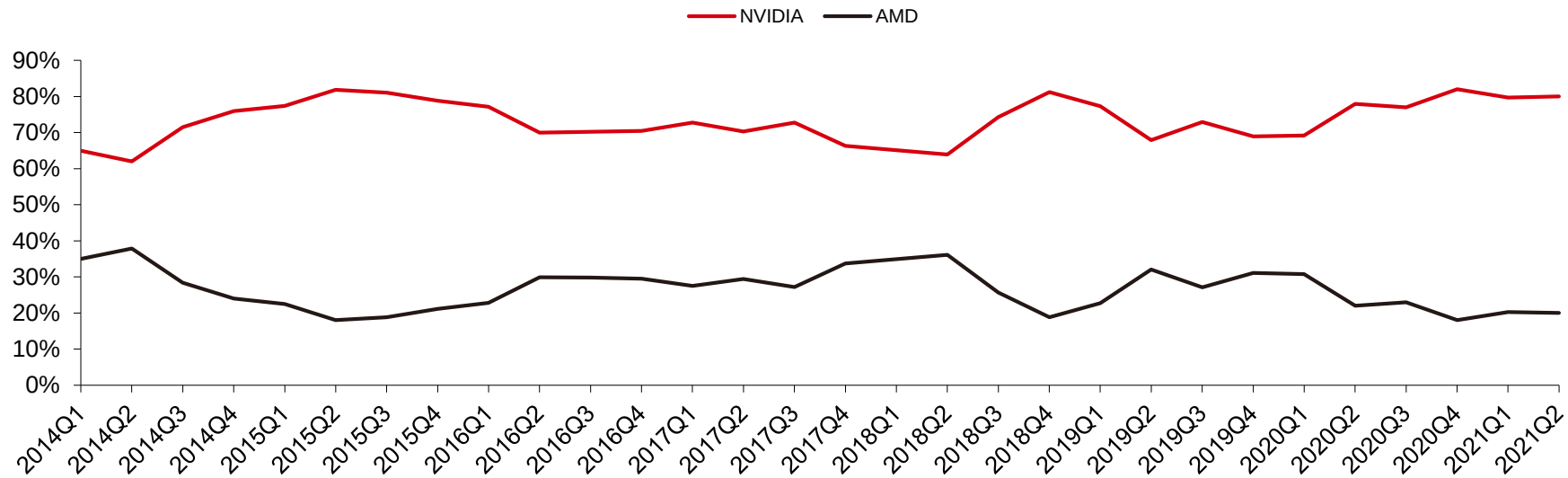
资料来源：3D Center，中信证券研究部

2.2 小结：架构创新升级和新兴领域前瞻探索是领跑GPU行业的关键

架构创新升级和新兴领域前瞻探索是领跑GPU行业的关键

- NVIDIA坚持每两到三年完成一次架构迭代，持续保持领先的图显和计算性能：2001年发布Kelvin，2003年发布Rankine，2004年发布Curie，2006年发布Tesla，2009年发布Fermi，2012年发布Kepler，2014年发布Maxwell，2016年发布Pascal，2017年发布Volta，2018年发布Turing，2020年Ampere，2022年先后发布Hopper和Ada lovelace。架构创新迭代高效，架构之间性能提升显著。而AMD（ATI）也曾凭借Radeon 9700、Radeon 9800强势崛起，近些年的RDNA架构也令其市场份额快速提高。
- NVIDIA前瞻性布局新兴领域数据中心、自动驾驶等领域，推动业绩爆发增长。其自2006年开始构筑CUDA生态并推出Tesla通用计算GPU架构，从以硬件为核心的企业变成以软硬件平台为核心的科技公司，前瞻性的布局使其在计算生态上构筑了极深的壁垒，占据了绝大部分市场。近年公司开始布局元宇宙等领域，持续探索新兴领域以保持GPU行业的龙头地位。

2014-2021年GPU市场份额



资料来源：超能网，中信证券研究部

2.2 未来竞争：NVIDIA新品性能提升飞跃，N卡地位难以动摇

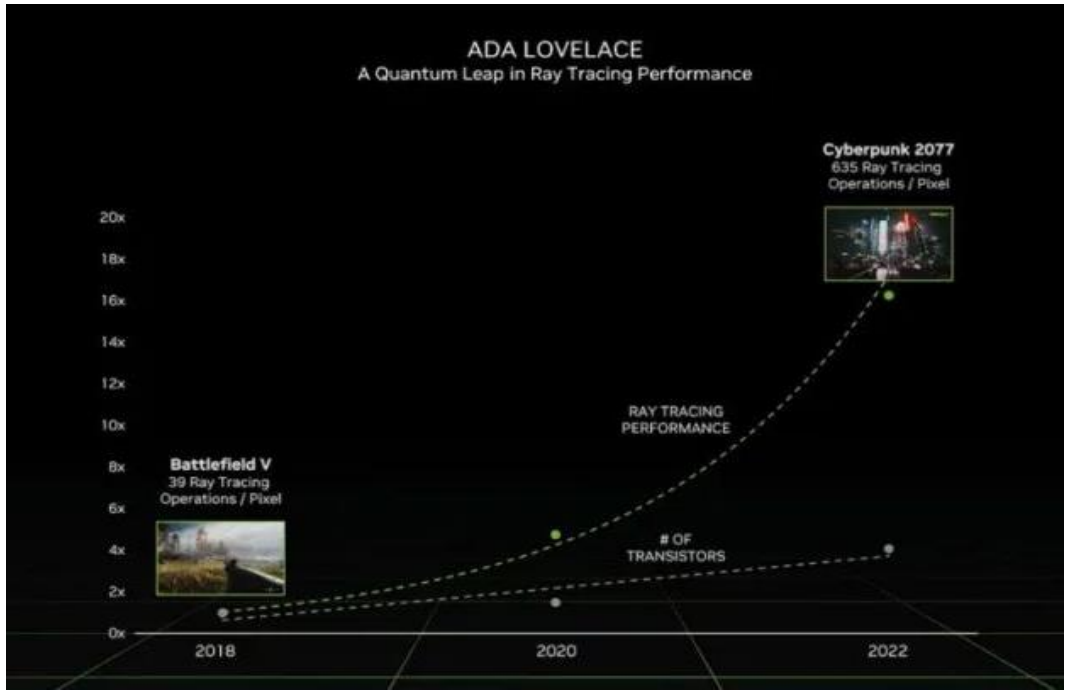
- **近期，NVIDIA公布了GeForce系列新品GeForce RTX 4090，RTX 4080和RTX 4070**
  - 2022年9月20日秋季GTC大会上发布的NVIDIA GeForce RTX40系列代表了目前显卡的性能巅峰，RTX 40系列采用全新的Ada Lovelace架构，台积电5nm级别工艺，拥有760亿晶体管和18000个CUDA核心，与Ampere相比架构核心数量增加约70%，能耗比提升近两倍，可驱动DLSS 3.0技术。性能远超上代产品。
  - Ada Lovelace架构对于RT Core、Tensor Core和SM单元都进行了升级，NVIDIA在SM多单元处理器中引入着色器执行重排序技术，使GPU也拥有CPU的乱序处理能力。

NVIDIA RTX 40系列显卡参数

显卡	RTX4090	RTX4080 16GB	RTX4080 12GB
核心构架	Ada Lovelace(艾达.洛夫莱斯)	Ada Lovelace(艾达.洛夫莱斯)	Ada Lovelace(艾达.洛夫莱斯)
核心代号	AD102-300	AD103-300	AD104-400
核心工艺	台积电5nm	台积电5nm	台积电5nm
CUDA核心数量	16384	9728	7680
Tensor Core	第四代	第四代	第四代
RT Core	第三代	第三代	第三代
GPU频率	2230 MHz	2210MHz	2310MHz
Boost频率	2520 MHz	2505MHz	2610MHz
显存位宽	384bit	256bit	192bit
显存类型	GDDR6X	GDDR6X	GDDR6X
显存容量	24GB	16GB	12GB
宽带	1008GB/s	720GB/s	504GB/s
TGP	450W	340W	285W
PCIe版本	4	4	4
DLSS版本	3	3	3
首发价格	1599美元（12999RMB）	1199美元（9199RMB）	899美元（7199RMB)

资料来源：泡泡网，中信证券研究部

NVIDIA Ada Lovelace架构性能提升

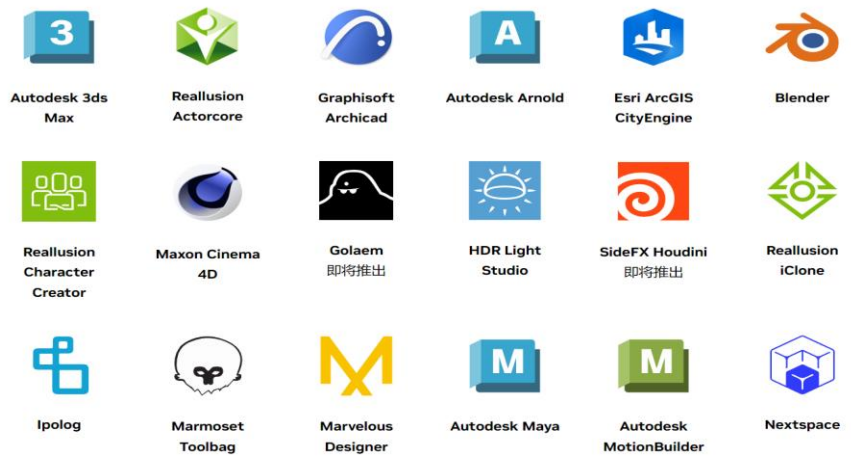


资料来源：NVIDIA 2022.9.20 GTC大会

2.2 未来竞争：NVIDIA维持游戏和数据中心领先地位，同时瞄准元宇宙、智能汽车市场

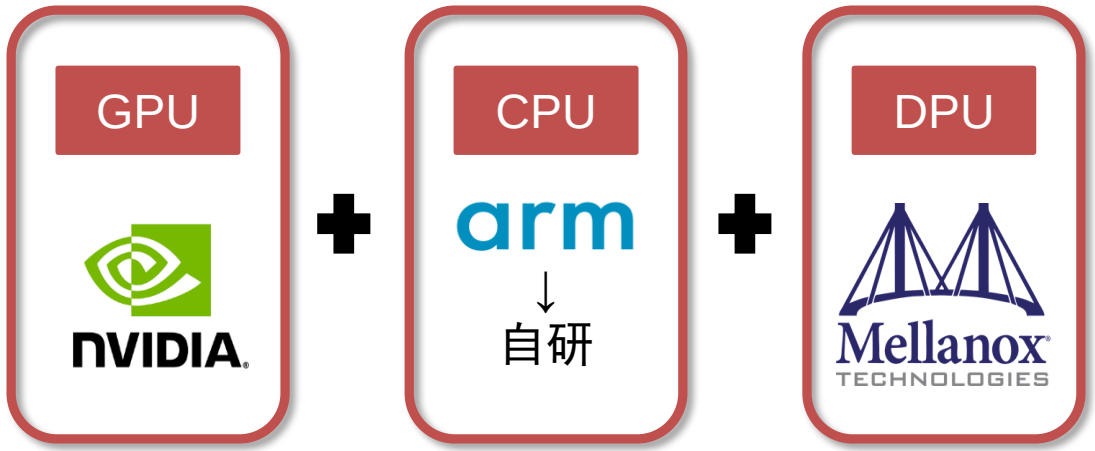
- **NVIDIA各主要业务线持续发力，以技术创新技术保持行业领先，开启元宇宙布局**
  - 游戏：NVIDIA新发布Ada Lovelace架构的4000系列GPU有极强的光追性能，比前代提高1-3倍，性能显著领先AMD，但成本也显著提高。
  - 汽车：在2022 GTC大会上发布的NVIDIA DRIVE Thor SoC系统，算力达到2000TOPS，公司计划在2025年装车。Thor可以将智能汽车的所有功能集成在单个AI计算器上，将显著降低成本，对智能座舱领域将是颠覆性影响。
  - 元宇宙：Omniverse是为元宇宙打造的软硬件方案，彰显其在元宇宙领域提前布局的野心。使用者可以在Omniverse中创建虚拟世界，而他们所创建的虚拟物体也会成为NVIDIA元宇宙生态的重要组成部分。与NVIDIA的GPU、CPU等硬件基础一起，共同构成NVIDIA元宇宙的一站式云服务体系。
- **云端芯片市场格局未定，NVIDIA实行“GPU+CPU+DPU”路线，布局云端异构AI芯片**
  - 2021年的GTC大会NVIDIA推出面向数据中心AI和高性能计算的自研的采用ArmNeoverse架构的Grace芯片。并取得ARM授权协议，可开发ARM架构CPU芯片。2019年，NVIDIA以70亿美元收购Mellanox，2020年推出BlueField-2 DPU，成功布局DPU业务。

Omniverse链接的部分软件



资料来源：NVIDIA官网

NVIDIA异构芯片路线

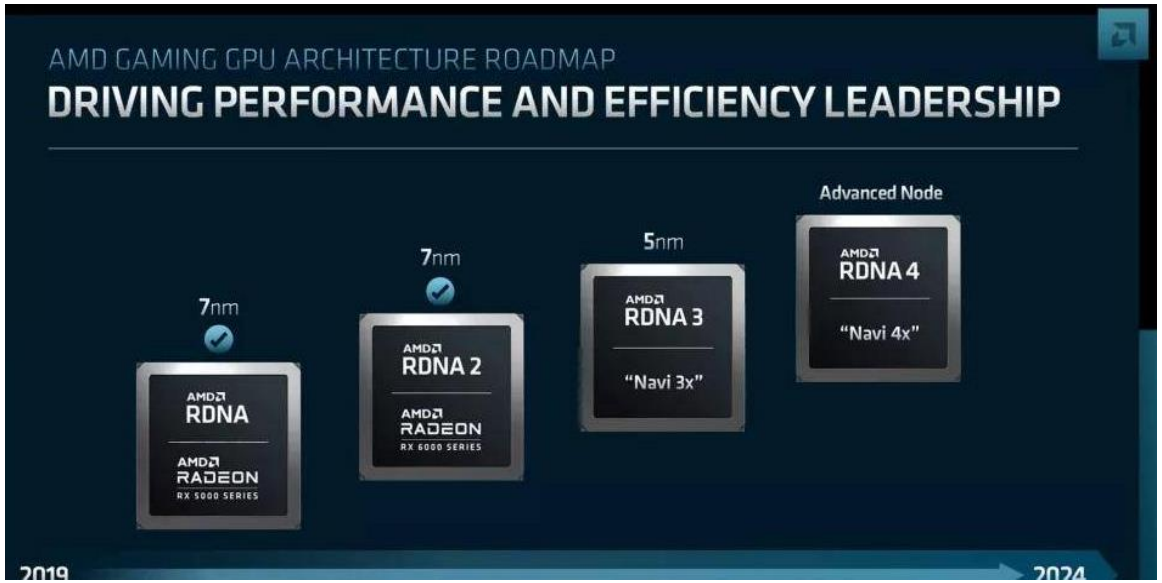


资料来源：NVIDIA官网，百度百科ARM，Mellanox官网，中信证券研究部绘制

2.2 未来竞争：AMD加快RDNA系列架构迭代和性能提升

- **AMD GPU架构稳定升级，不断挑战NVIDIA显卡卡皇地位**
  - RDNA架构迭代路径清晰，代际之前性能提高显著，即将发布的RDNA 3架构相比RDNA 2每瓦性能提高超过50%。预计2024年前RDNA 4 架构可正式发布。
  - AMD在2022年11月4日发布搭载RDNA 3的下代显卡Radeon RX 7000 系列，其采用5纳米制程和小芯片封装工艺，新一代“无限缓存”。据称，旗舰RX 7950 XT显卡有15360个核心，频率达2.5GHz，512MB 3D缓存，搭配256bit GDDR6显存，支持PCI-E 5.0接口。
  - 预计2024年发布Navi 4x系列，采用RDNA 4架构，也将使用更先进的制程工艺。

AMD RDNA架构迭代



资料来源：AMD官网

AMD 下一代GPU渲染图

资料来源：中关村在线

2.2 未来竞争：AMD结合自身CPU优势全方位布局AI芯片

- **2022年6月19日，AMD讲述其未来发展战略，概述为技术和产品组合更新、扩大数据中心解决方案产品组合、加速打造无所不在的 AI 领域领导地位、扩大PC领先、推动图形解决方案发展势头。**
- **AMD结合CPU优势，GPU、FPGA、APU业务多点布局抢占AI芯片行业先机**
  - AMD希望未来将更多AI功能引入CPU的硬件层面中，如AVX-512 VNNI指令集。AMD认为，在CPU中运行大部分推理很重要并会是未来趋势。AMD预计明年发布全球第一个数据中心APU——Instinct MI300（此前该系列为GPU加速卡），面向训练领域，结合使用Zen 4架构的CPU和CDNA 3架构的GPU。APU是将处理器和独显集成到一个晶片上，实现GPU和CPU的融合。
  - AMD收购Xilinx（赛灵思）以更好地开展FPGA业务，补全FPGA领域的短板，扩大自身AI芯片市场。

AMD Instinct系列加速卡

VideoCardz	AMD Radeon Instinct M160	AMD Instinct MI100	AMD Instinct MI250X	AMD Instinct MI300
Architecture & Nodes	7nm GCN5(GFX906)	7nm CDNA1(GFX908)	6nm CDNA2(GFX90A)	5nm CDNA3(GFX940)+6nm(base)
CPU	-	-	-	Zen4(APU Mode)
GPU	Vega 20	Arcturus	Aldebaran(MCM)	?(3D Die Stacking)
Base Chiplets	-	-	-	up to 2
Compute Tiles	1	1	2	up to 8
Compute Units	64	120	220	TBC
GPU Clock Speed	1800 MHz	-1500 MHz	-1700 MHz	TBC
FP16 Compute	29.5 TFLOPS	185 TFLOPS	383 TFLOPS	TBC
FP32 Compute	14.7 TFLOPS	23.1 TFLOPS	47.9 TFLOPS	TBC
FP64 Compute	7.4 TFLOPS	11.5 TFLOPS	47.9 TFLOPS	TBC
VRAM	32 GB HBM2	32GB HBM2	128 GB 8x HBM2e	up to 8x HBM3 stack
Memory Clock	2.0 Gbps	2.4 Gbps	3.2 Gbps	TBC
Memory Bus	4096-bit	4096-bit	8192-bit	up to 8192-bit
Memory Bandwidth	1 TB/s	1.23 TB/s	3.2 TB/s	TBC
Form Factor	Dual Slot.Full Length	Dual Slot.Full Length	OAM	OAM
TDP	300W	300W	560W	up to 600W+

资料来源：快科技，中信证券研究部

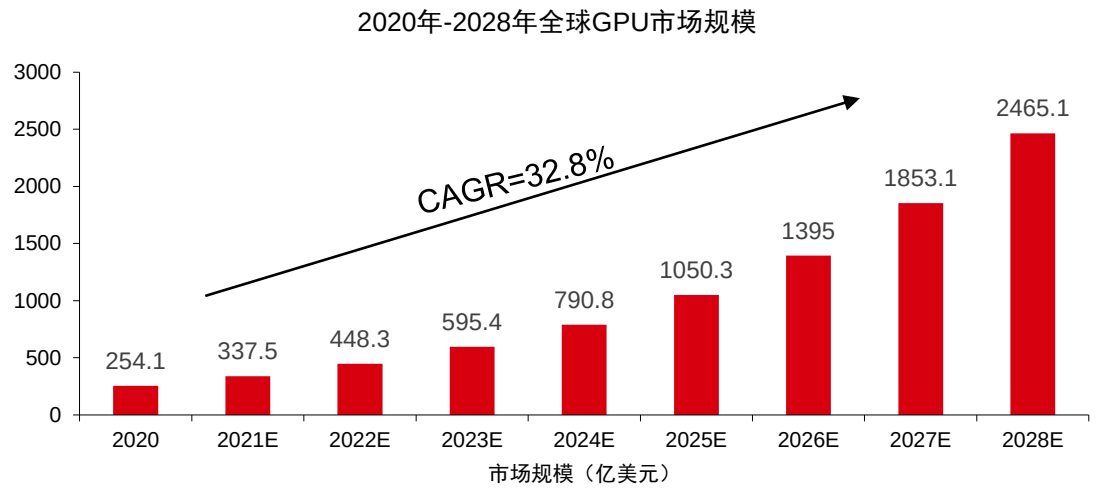
3.国内市场：GPU细分市场前景广阔，国内厂商大有可为

I. 市场概览：国内外GPU市场规模庞大，AI&数据中心、汽车、游戏可重点关注
II. AI&数据中心：数据量级和算力需求的提升拉动数据中心业务与国家超算需求高增
III. 汽车：汽车智能化浪潮下汽车GPU市场前景广阔
IV. 游戏：游戏玩家人数持续增长，释放游戏GPU市场需求

3.1 GPU市场空间广阔，国内企业规模逐步起量

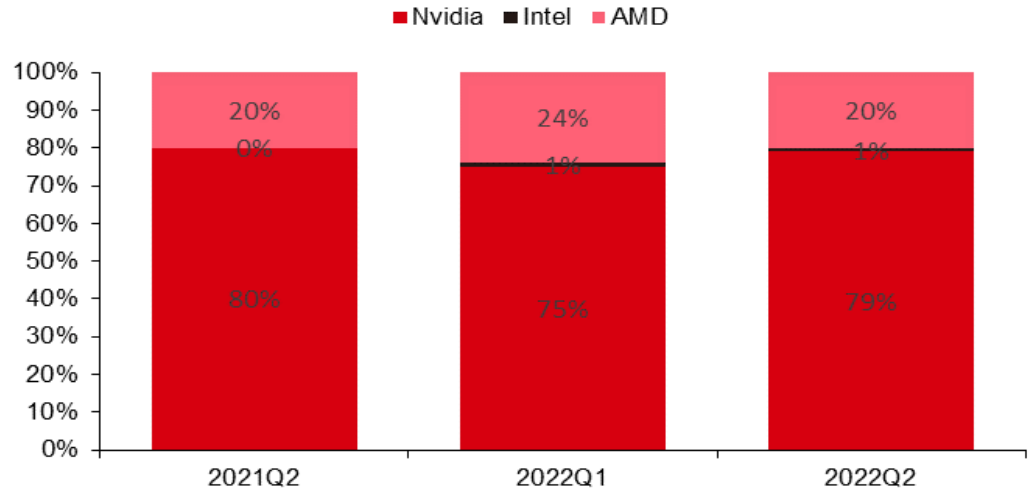
- **2022年全球GPU市场规模达到448.3亿美元，国内外市场空间正高速增长，年复合增长率达到32.8%**
  - Verified Market Research 数据显示，2020年，全球GPU市场规模为254.1亿美元，且该机构预计2028年市场规模将达到2465.1亿美元，对应年复合增长率达32.8%。
- **国际独立GPU市场由Nvidia、AMD八二分成，国内市场中国企业体量快速增长**
  - 国际市场上，英伟达、AMD瓜分市场，Jon Peddie Research数据显示2022Q1英伟达占据79%市场份额，AMD占据21%。英伟达在独立GPU领域一枝独秀，AMD在集成GPU领域可与英伟达竞争。
  - 根据各公司财报，国内GPU龙头企业景嘉微2022年上半年营业收入5.44亿人民币，2021年营业收入10.93亿人民币；2022年上半年海光信息营业收入为25.3亿元，而英伟达2022Q2营收为67亿美元，2021年NVIDIA中国区的营收约为71亿美元。相比之下，国产厂商相对规模暂时较小，未来成长空间广阔。

2020-2028年全球GPU市场规模



资料来源：Verified Market Research（含预测），半导体行业观察，中信证券研究部

2021Q2-2022Q2全球独立GPU市场份额

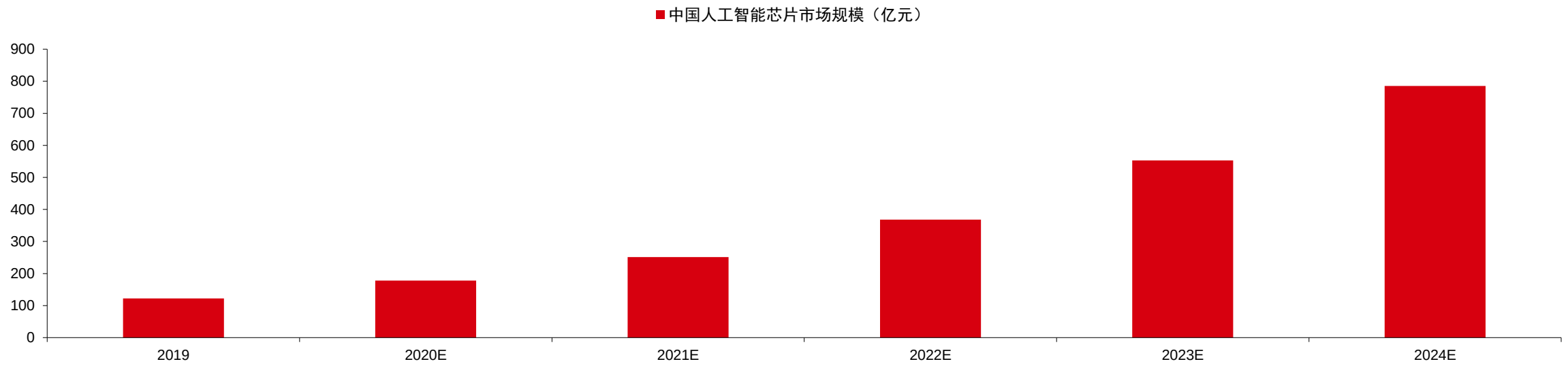


资料来源：芯智讯， Jon Peddie Research ，中信证券研究部

3.1 国内市场：GPU应用市场可划分为—AI&数据中心、智能汽车、游戏

- **GPU应用场景不断扩大拉动GPU市场空间迅猛增长，根据Verified Market Research预测，2027年中国GPU市场规模将会增长至345.57亿美元。**
- **GPU市场主要应用场景可概括为：AI&数据中心、智能汽车、游戏。**
  - AI&数据中心：新一轮AI对算力需求远超以往：ChatGPT类语言大模型底层是2017年出现的Transformer架构，该架构相比传统的CNN/RNN为基础的AI模型，参数量达到数千亿，对算力消耗巨大，对算力硬件有大量需求。随着对商业数据和大数据处理要求算力的不断提高，GPU的通用计算能力正在越来越广泛地被应用与数据中心和国家超算中心的建设。
  - 智能汽车：智能汽车方兴未艾，自动驾驶和智慧座舱是智能汽车发展的主要方向，均需大量使用GPU。
  - 游戏：游戏业务是GPU应用的传统领域，对游戏画面进行3D渲染，英伟达的游戏业务稳中有进。

2019-2024年中国人工智能芯片市场规模及预测

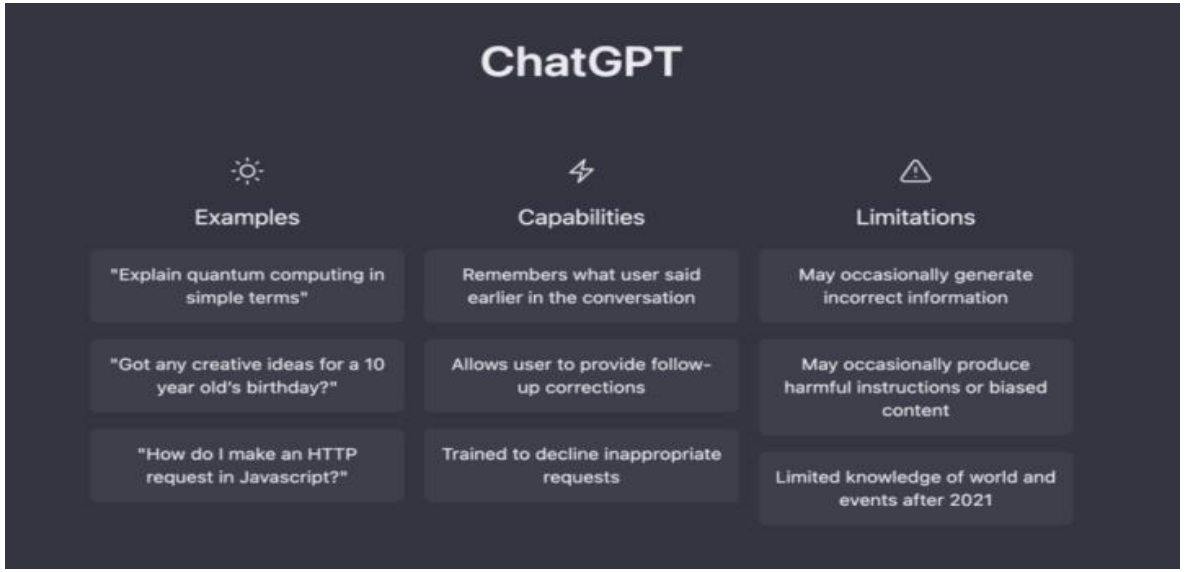


资料来源：前瞻产业研究院（含预测），公司招股说明书，中信证券研究部

3.2 AI：ChatGPT等AI大模型加速对大算力的需求

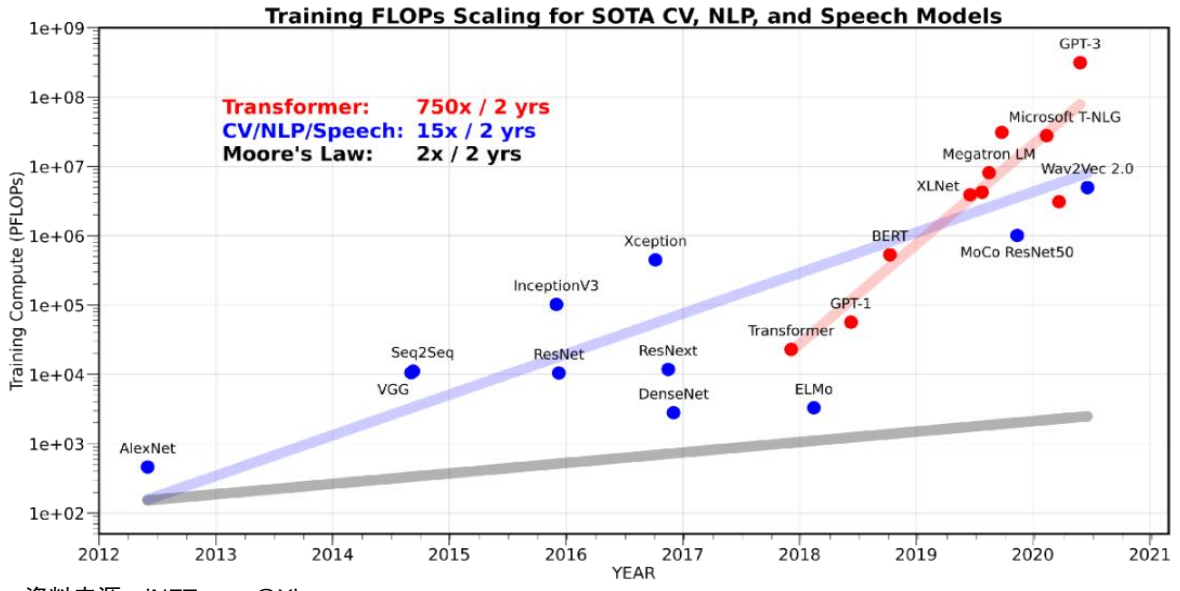
- **ChatGPT 模型引发市场关注，对话类AI效果超大众预期，大模型需要更大的算力。**
  - 2022年11月人工智能实验室 OpenAI 推出了一款AI对话系统—ChatGPT，ChatGPT模型从 GPT-3.5 系列中的一个模型微调而成，并在 Azure AI 超级计算基础设施上进行训练，能够进行有逻辑的对话、撰写代码、撰写剧本、纠正错误、拒绝不正当的请求等，效果超越大众预期。这标志着对话类人工智能可以在大范围、细节问题上给出较合理准确的答案，并根据上下文形成一定像人类一样有逻辑且有创造力的回答。
  - ChatGPT的优化主要来自模型的增大，以及因此带来的算力增加。GPT、GPT-2和GPT-3（当前开放的版本为GPT-3.5）的参数量从1.17亿增加到1750亿，预训练数据量从5GB增加到45TB，其中GPT-3训练单次的成本就高达460万美元。

ChatGPT界面



资料来源：ChatGPT官网

大模型算力

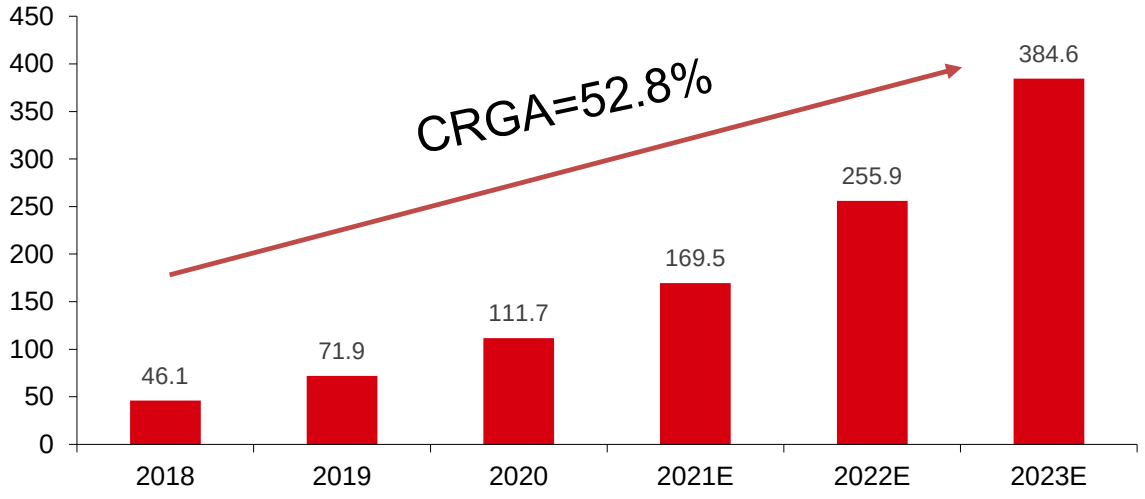


资料来源：iNFTnews@Xiaoz

3.2 AI：数据中心和终端场景不断落地对计算芯片提出更多更高需求

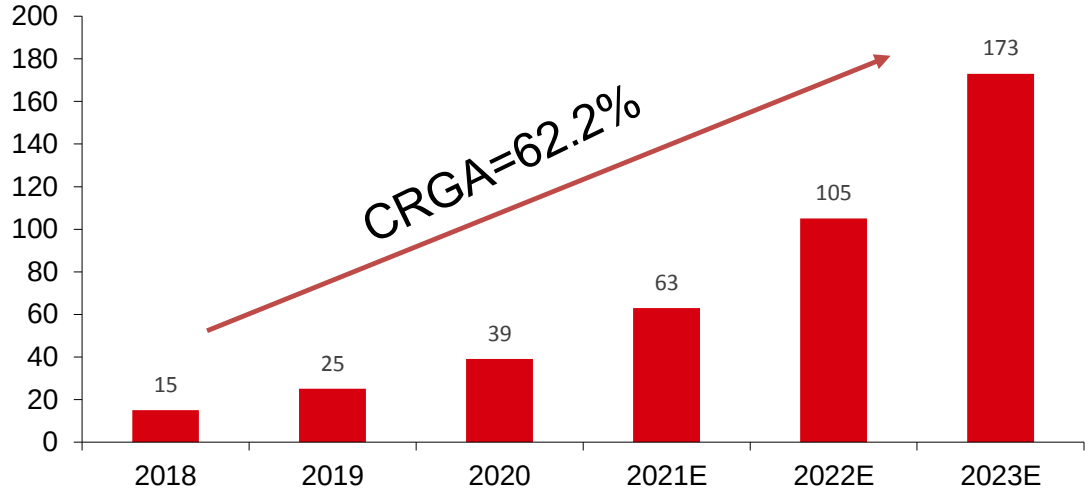
- **依据部署位置划分，AI芯片可以细分为终端芯片和云端芯片，云端芯片市场空间越为终端芯片的2-3倍**
  - 云端芯片：云端芯片应用于云端服务器，可以进一步细分为推理芯片和训练芯片。根据甲子光年数据，2018年中国云端芯片市场约46.1亿元，该机构预计2023年增长至384.6亿元。
  - 终端芯片：应用于嵌入式、移动终端、智能制造、智能家居等领域的AI芯片，终端芯片需要低功耗和更高的能效比，但是对算力的需求也相对较低，主要应用与AI推理。根据甲子光年数据，2018年中国终端芯片市场约15亿元，该机构预计2023年增长至173亿元。
- **AI芯片总市场232亿元，其中云端芯片市场空间更大，预计终端芯片将随着AI在多行业落地将进一步放量**
  - 甲子光年预测，中国AI芯片市场规模将从2021年232亿元增长至2023年的500亿元左右，对应中国云端芯片市场的复合增长率为52.8%；终端芯片市场规模相对较小，但由于人工智能在汽车、安防、智能家居等行业渗透，届时市场规模增长率达到62.2%。

中国云端AI芯片市场规模（亿元）



资料来源：甲子光年（含预测），中信证券研究部

中国终端AI芯片市场规模（亿元）

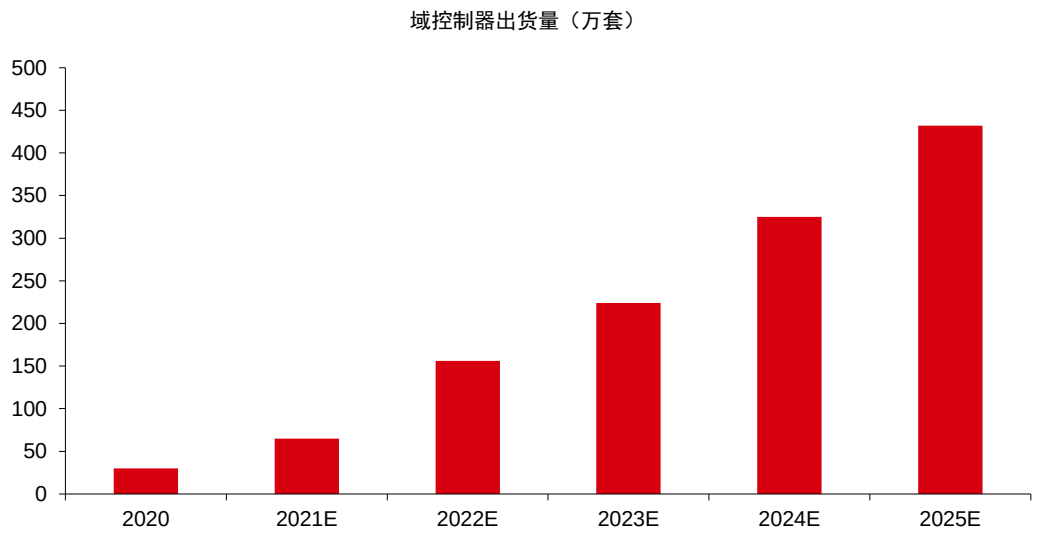


资料来源：甲子光年（含预测），中信证券研究部

3.2 汽车智能化浪潮下汽车GPU市场前景广阔—自动驾驶

- **汽车智能化浪潮下，自动驾驶和智能座舱是最具有发展前景的两个方向，GPU应用于二者的域控制器**
- **GPU两大功能助力自动驾驶**
  - 智能汽车主流的域控制器采用SoC与MCU结合的方案，SoC（片上系统） 由GPU、CPU、AI引擎、DPU等组成。GPU在自动驾驶中的作用表现在图形处理和并行计算，ADAS 平台可以利用 GPU 的并行计算能力实时分析来自激光雷达、雷达和红外摄像头的传感器数据。
- **盖世汽车预计到2025年中国自动驾驶域控制器出货量达到432万台**
  - 自动驾驶域控制器与SoC之比在1:1到1:4之间，市场份额较高的SoC通常搭载一片GPU。

中国自动驾驶域控制器出货量



资料来源：盖世汽车（含预测），中信证券研究部

智能汽车主要产品使用SoC

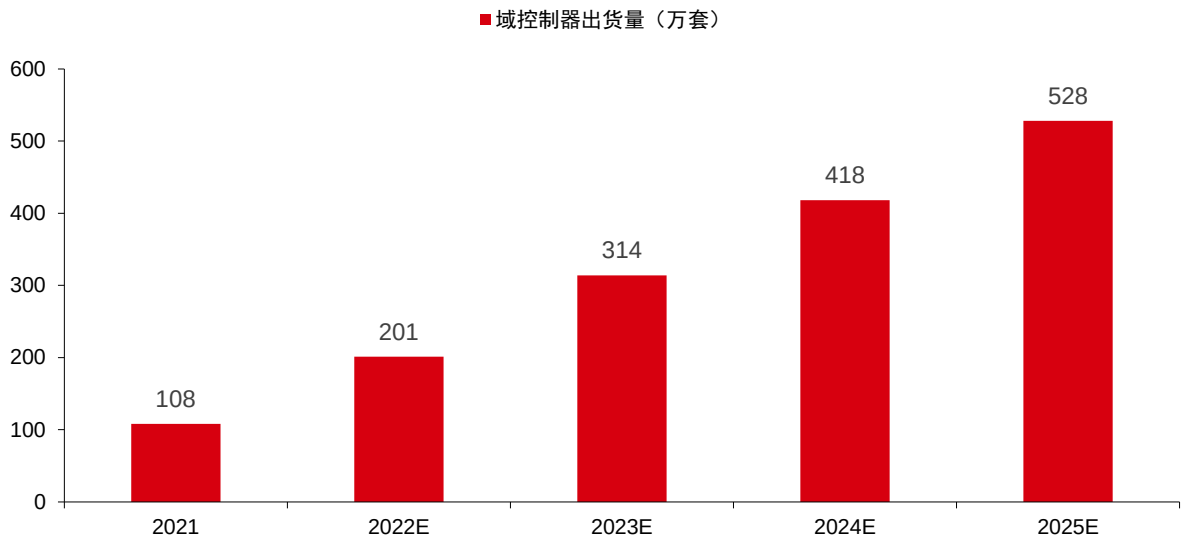
车型	智能域控SoC	AI算力（TOPS）	域控制器供应商	车型上市时间
特斯拉Model3	2颗FSD	144	自研	已量产
小鹏P5/P7	1颗Xavier	30	德赛西威	已量产
小鹏G9	2颗Orin X	508	-	预计2022年第三季度
理想L9	2颗Orin X	508	德赛西威	2022年6月
蔚来ET7	4颗Orin X	1016	自研	已量产
威马M7	4颗Orin X	1016	-	预计2022年下半年
上汽智己L7	2颗-4颗Orin X	500~1000+	创时智驾	2022年4月
上汽非凡R7	2颗-4颗OrinX	500~1000+	德赛西威	预计2022年下半年
哪吒S	1颗昇腾610	200	华为	预计2022年底
长安阿维塔11	2颗昇腾610	400	华为	2022年8月

资料来源：芯八哥微信公众号，中信证券研究部

3.3 汽车智能化浪潮下汽车GPU市场前景广阔—智能座舱

- **GPU虚拟化技术助力智能座舱一芯多屏化发展**
  - 智能座舱芯片主要为SoC（片上系统），SoC由GPU、CPU、AI引擎、DPU等组成。
  - 智能座舱向着一芯多屏的形态发展，这对芯片并行计算的要求不断提高，GPU硬件虚拟化技术在智能座舱中有着无可替代的优势。在智能座舱屏幕、仪表盘、车载与各系统中均需要使用GPU。
- **盖世汽车预计2025年中国智能座舱域控制器出货量达到528万台**
  - 智能座舱一台域控制器内置一个SoC，位于汽车的中央显示屏内，一个SoC通常搭载一片GPU。

中国智能座舱域控制器出货量



资料来源：盖世汽车（含预测），中信证券研究部

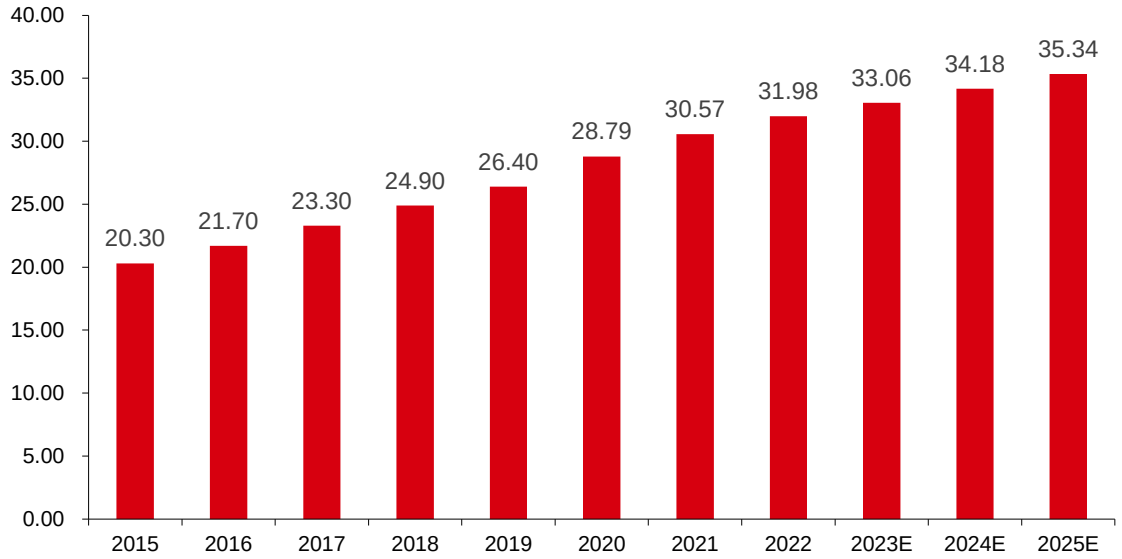
小鹏G9座舱

资料来源：小鹏汽车官网

3.3 游戏玩家人数持续增长，游戏GPU市场规模稳中有升

- **游戏市场是传统意义上GPU最关键的市场**
  - 随着千禧一代游戏需求的进一步提升，全球游戏玩家数量稳增，相应地扩展了游戏GPU市场规模。Newzoo Expert预计2020-2025年全球游戏玩家人数复合年增率为4.2%。
- **游戏GPU的主要分类**
  - 根据现行市场上的主要产品可划分为四类：1）游戏机、2）PC端主机游戏、3）控制台、4）VR&AR

全球玩家数（亿人）



资料来源：Newzoo Expert（含预测），中信证券研究部

游戏GPU分类

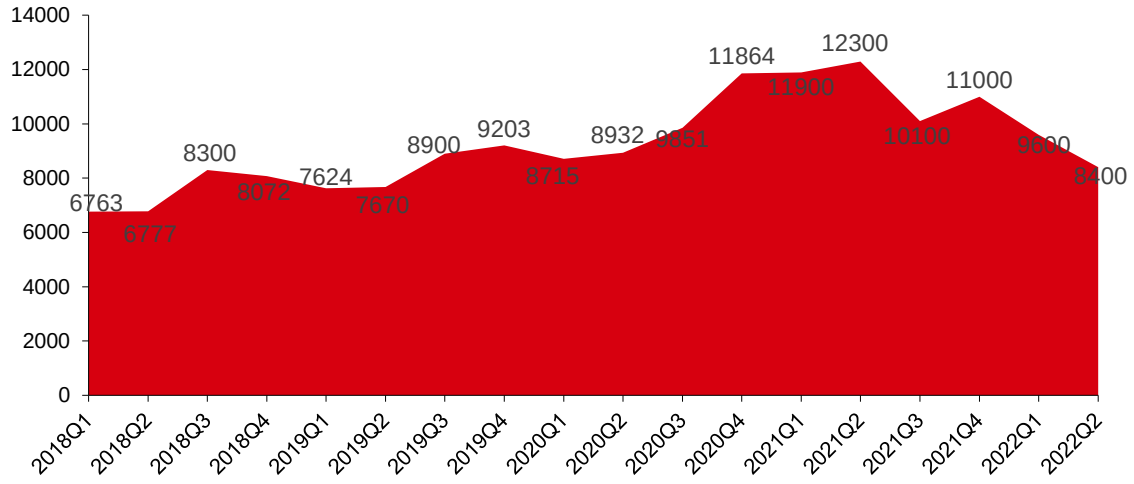
分类	描述
游戏机	主要产品包括Nintendo Switch、Xbox One，同时索尼、微软也相竞推出新产品
PC端主机游戏	随着对于高画质的游戏需求，各厂商均试图超越4k，提升刷新率至120fps
控制台	电子竞技、视频游戏的兴起，增大了对控制台的需求和品控要求，提升了竞争门槛
VR、AR	与各类应用程序逐渐融合，重新定义人机交互、体验模式，提高用户体验感

资料来源：《游戏 GPU 市场 - 增长、趋势、COVID-19 影响和预测》Mordor Intelligence，中信证券研究部

3.3 PC GPU全球出货量稳中有升

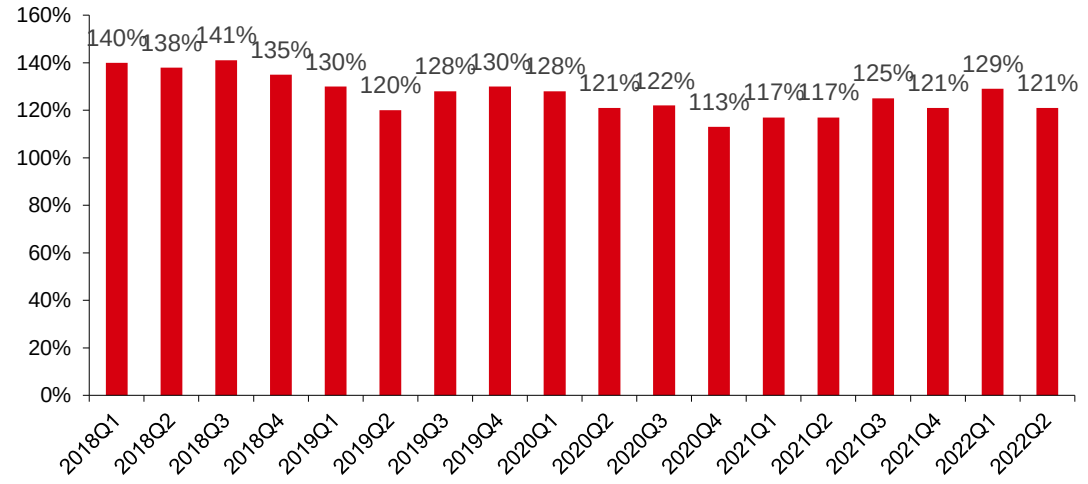
- **PC游戏市场主体**
  - PC指单机电脑、个人电脑。PC端游戏是通过计算机进行相关操作，实现人机交互的游戏方式。
  - 根据Jon Peddie Research统计，2021年Q4全球PC GPU出货量（包括集成和独立显卡）高达11000万片。受到俄乌冲突、天然气供应等冲击性全球事件影响，2022年Q1、Q2，全球PC GPU出货量略有下降。
- **GPU与PC配售比**
  - 配售比指GPU与PC主机的整体采用率，由于PC正常运行必须要求硬件上同时具备CPU和GPU，因此每台PC至少需要一张集成显卡或独立显卡，此外，根据需求可再增购独立显卡。根据JPR统计，全球GPU与PC配售比在2022Q1达到了129%。

2018-2022年Q2全球PC GPU出货量（万片）



资料来源：JPR，华经产业研究院，中信证券研究部

2018-2022年Q2全球GPU与PC配售比



资料来源：JPR，华经产业研究院，中信证券研究部

3.4 国内企业：景嘉微、中船占据军民信创，海光GPGPU领先

国内典型GPU企业列表						
公司	地区	成立时间	估值/市值	创始人/核心人员	技术背景	主要GPU产品
景嘉微电子	长沙	2006	368亿	曾万辉（国防科大微波毫米波硕士）、胡亚华（国防科大通信电子硕士、讲师）、饶先宏（核心技术人员，国防科大运筹学硕士、副教授，公司首席专家）	初始技术来自ATI，核心技术团队来自国防科大	JM5(已在军机广泛使用)、JM7（包含民用信创主力产品）、JM9（2021推出）系列图形GPU
凌久电子	武汉	1983	未上市		初始技术来自ATI，核心团队来自中船重工	GP101(实现了我国通用3D显卡的突破)\GP102
中船重工716研究所	连云港	1965	未上市		核心技术团队来自中船重工	JARIG12(是2018年性能最强的国产通用图形处理器)
航锦科技	辽宁/长沙	1997	229亿		并购长沙韶光	SG6931GPU
芯原微电子	上海	2001	301亿	戴伟民（董事长，UC Berkeley计算机博士，UC圣克鲁兹教授，Ultima创始人，原Celestry董事长/CTO）、戴伟进（副总裁，UC Berkeley计算机硕士，曾任职于惠普、朗讯贝尔实验室、Cadence，美国图芯CEO）	收购图芯美国（Vivante）	Vivante GPU IP
龙芯中科	北京	2008	457亿	胡伟武（董事长、总经理，中科院计算所博士、曾任中科院计算所研究员、博导、总工程师）	自研	7A2000桥片集成显卡
兆芯	上海	2013	未上市	叶峻（上海联和投资总经理，曾任上海华虹董事、上海宏力董事）	架构及IP来源于台湾VIA(VIA收购了原GPU主流厂商S3 Graphics)、美国Centaur IP	C320、C860、C960、C1080集成显卡
海光信息	北京	2014	1176亿	孟宪堂（董事长，原发改委处长、副司长、国科控股副总）、沙超群（总经理，教授级高工，原中科曙光技术副总裁/高级副总裁）、历军（教授级高工，中科曙光总裁）	GPU技术与AMD有合作，结合自主研发，核心团队来自中科曙光	深算一号DCU（海光8100 GPGPU，性能接近V100，核心数达到V100 80%，兼容ROCm）

资料来源：芯榜微信公众号，界面新闻，IT桔子，中信证券研究部；注：市值基于2023年2月10日收盘价；估值分别转引自芯榜、IT桔子

3.4 国内企业：创业公司百花齐放

国内典型GPU企业列表						
公司	地区	成立时间	估值/市值	创始人/核心人员	技术背景	主要GPU产品
芯动科技	珠海	2007	300亿	敖海（创始人/董事长/CEO）	Imagination BXT IP	风华1号服务器GPU，风华2号桌面GPU
天数智芯	上海/南京	2015/12/29	150亿	刁石京（董事长/总经理，曾任紫光集团DRAM事业群董事长、紫光集团联席总裁、紫光国微董事长、紫光展锐执行董事长、长江存储执行董事、工信部电信司司长）、吕坚平（CTO）		天垓100 GPGPU（已量产）,中国第一家通用GPU高端芯片及超级算力提供商
壁仞科技	上海	2019/9/9	170亿	张文（创始人/董事长/CEO，哈佛法学博士，曾任商汤科技总裁）、李新荣（联席CEO，原AMD全球副总裁、AMD中国研发中心总经理）		BR100 GPGPU（理论性能超越英伟达H100）、BR104 GPGPU（AI性能超越英伟达A100）
沐曦集成电路	上海	2020/9/14	150亿	陈维良（创始人/董事长/CEO，原AMD高管，负责GPGPU产品线整体设计管理），彭莉（CTO/首席硬件架构师，AMD首席SoC架构师，系统架构师，AMD全球首位华人女Fellow）、杨建（CTO/首席软件架构师，AMD大中华区首位Fellow）		MXN AI推理芯片，MXC GPGPU，MXG图形渲染GPU
登临科技	上海	2017/11/17	150亿	李建文（创始人/董事长，曾任图芯科技副总裁）、王震宇（联合创始人，曾任职于龙芯、百度美国研究院）、王平（联合创始人，曾任图芯首席架构师）、杨剑（全球运营副总裁，曾任华为全球供应链副总裁、思科全球供应链副总裁）		Goldwasser系列GPU+AI加速卡
摩尔线程	北京	2020/6/11	150亿	张建中（创始人/CEO，前NVIDIA全球副总裁、中国区总经理，曾任职于惠普、戴尔）	Imagination BXT	MTT S80/60/30/10桌面GPU，MTT S3000服务器GPU，国产首颗全功能GPU及PCIe5.0 GPU，成立一年半即推出苏堤架构，2022年底实现第二代架构春晓架构量产
芯瞳半导体	西安	2019/11/20	20亿	黄虎才（董事长，任职于西安邮电大学）		GenBu01 GPU（面向信创市场，40nm制程）
中微电	深圳	2009/4/1		梅思行(前NVIDIA主架构设计工程师，参与设计第一代GeForce、第一个可编程GPU、第一个GPGPU G80，曾任职于IBM、SGI)，周志德(斯坦福计算机博士，MIPS联合创始人/首席工程师，曾任SGI首席架构师，开发Pro64/Open64编译器，主导华为方舟编译器)		南风一号GPU（2022年7月流片成功）、南风二号游戏GPU、南风三号AI GPU

资料来源：芯榜微信公众号，界面新闻，IT桔子，中信证券研究部；注：市值基于2023年2月10日收盘价；估值分别转引自芯榜、IT桔子

3.4 国产GPGPU：算力逐步提升，计算框架力求兼容

中信证券 CITIC SECURITIES

部分国产GPGPU与国际主流产品性能对比														
厂家	产品	推出时间	生态	工艺制程	峰值功耗	核心数	FP32算力	FP32张量算力	FP16/BF16算力	INT8算力	显存类型	显存容量	显存位宽	显存带宽
单位				nm	W		TFLOPs	TFLOPs	TFLOPs	TOPs		GB	bit	GB/s
NVIDIA	V100 SXM2	2017	CUDA	12	300	5120	15.7	125			HBM2	32	4096	900
NVIDIA	A100 SXM	2020	CUDA	7	400	6912	19.5	156		624	HBM2e	80	5120	2039
NVIDIA	H100 SXM	2022	CUDA	4	700	16896	60	500	120	4000	HBM3	80	5120	3072
AMD	MI100	2020	ROCm	7	300	7680	23.1	46.1	184.6	184.6	HBM2	32	4096	1228
AMD	MI210	2022	ROCm	6	300	6656	22.6	45.3	181	181	HBM2e	64	4096	1638
AMD	MI250	2021	ROCm	6	560	13312	45.3	90.5	362.1	362.1	HBM2e	128	8192	3277
AMD	MI250X	2021	ROCm	6	560	14080	47.9	95.7	383	383	HBM2e	128	8192	3277
海光	深算一号	2021	ROCm	7	350	4096					HBM2	32	4096	1024
摩尔线程	MTT S2000	2022	CUDA	12	150	4096	10.6			40	GDDR6	32		
摩尔线程	MTT S3000	2022	CUDA			4096	15.2				GDDR6	32	256	448
天数智芯	BI-V100	2021	CUDA	7	250		18.5	37	37	295	HBM2	32		1228
壁仞科技	BR100	2022	CUDA	7			256	512	1024	2048	HBM2e	64		
壁仞科技	BR104	2022	CUDA	7nm			128	256	512	1024	HBM2e	32		
壁仞科技	BR100P	2022	CUDA	7nm	550		240	480	960	1920	HBM2e	64	4096	1638
壁仞科技	BR104P	2022	CUDA	7nm	300		112	224	448	896	HBM2e	32	2048	819
登临科技	UL32		CUDA		10				8	32				
登临科技	UL64		CUDA		15				16	64				
登临科技	L32		CUDA		25				32	128				
登临科技	L64		CUDA		45				64	256				
登临科技	XL		CUDA		120				128	512		32/64		

资料来源：各公司官网，海光信息招股说明书，中信证券研究部

3.4 国产图形GPU：硬件性能增长，图形API逐步适配

中信证券 CITIC SECURITIES

部分国产图形GPU与国际主流产品性能对比

	发布年份	工艺制程	功耗/W	显存类型	显存容量/GB	显存带宽/G/s	核心频率/MHz	像素填充率/GP/s	FP32算力/TFlops	总线接口	OpenGL	DirectX/Vulkan
ATI M96	2009	55nm		GDDR3	0.5	19.2	-	-	-	PCIe3.0	2.0	DX10
Nvidia GT640	2012	28nm	50	DDR3	2	80	950	7.22	0.69	PCIe3.0	4.1	DX11
Nvidia GTX1050	2016	14nm		GDDR5	2	112	1354	36.4	1.8	PCIe3.0	4.5	DX12
Nvidia GTX1080	2016	16nm		GDDR5	8	320	1607	111	8.9	PCIe3.0*16	4.5	DX12
Nvidia RTX3060	2020	8nm		GDDR6	12	360	1780	85.3	12.8	PCIe4.0*16	4.6	DX12 /Vulkan
景嘉微JM5400	2015	65nm	<6	DDR3	1	9.6	550	2.2	0.16	PCI 2.3	1.3	-
景嘉微JM7201	2018	28nm	5-15	DDR3	4	17	1300	5.2	0.5	PCIe2.0*16	2.0	-
景嘉微JM9100	2022	14nm	5-15	GDDR6	8	256	1500	32	0.5	PCIe4.0*8	4.0	Vulkan1.1
景嘉微JM9200	2022	14nm	15-30	GDDR6	16	512	1800	128	1.2	PCIe4.0*8	4.0	Vulkan1.1
凌久GP101	2018	-	3-7	DDR3	1	10.6	600	2.4		PCIe2.0*4	2.0	-
JARI G12	2018	-	3-7							PCIe3.0	2.0	-
芯瞳GenBu01	2020	40nm	3				533			PCIe2.0*4	4.3	-
芯动科技风华1号	2021	12nm	20/40	GDDR6	16	304		160	5	PCIe4.0*16	4.3	Vulkan1.2
芯动科技风华2号	2022		4-15	LPDDR5X	8	102		48	1.5	PCIe3.0*8	4.3	Vulkan1.2
摩尔线程MTT S60	2022	12nm		LPDDR4X	8			192	6		支持	DX9/DX11(不完全)/Vulkan
摩尔线程MTT S80	2022	7nm	114-250	GDDR6	16	448	1800		14.4	PCIe5.0*16	支持	DX9/DX11(不完全)/Vulkan

资料来源：各公司官网，各公司数据手册，芯参数，中信证券研究部

3.5 海光：CPU+GPU双线发展，本土应用+产业生态双重优势

- **海光CPU：**兼容x86指令集，处理器性能参数与国际同类主流处理器产品相当，支持国内外主流操作系统、数据库、虚拟化平台或云计算平台，能够有效兼容目前存在的数百万款基于x86指令集的系统软件和应用软件，具有优异的生态系统优势。**海光DCU：**兼容“类CUDA”环境，软硬件生态丰富，典型应用场景下性能指标达到国际上同类型高端产品的水平。3）公司主动融入国内外开源社区，积极向开源社区提供适用于海光CPU、海光DCU的适配和优化方案，保证了海光高端处理器在开源生态的兼容性。
- **公司下游服务器厂商开发了多款基于海光处理器的服务器，有效地推动了海光高端处理器的产业化。**目前，海光CPU已经应用到了电信、金融、互联网、教育、交通等行业；海光DCU主要面向大数据处理、商业计算等计算密集型应用领域以及人工智能、泛人工智能应用领域。**公司正持续大力投入研发实现GPU架构创新升级和快速迭代步调，力争赶超国际领先水平；同时加大生态建设力度，打造自主开放的通用计算软件生态体系。**
- **风险因素：**公司核心技术积累不足或研发迭代不及预期的风险；公司的供应商集中度较高且部分供应商替代困难的风险；市场竞争加剧的风险；宏观环境带来的市场不确定性风险；国产化需求节奏放缓的风险。

海光产品生态

资料来源：海光信息官网，中信证券研究部

海光产业链客户

资料来源：海光信息及其他公司官网，中信证券研究部

3.5 海光DCU：基于GPGPU，兼容“类CUDA”环境

- **海光DCU属于GPGPU的一种，兼容通用的“类CUDA”环境。**
  - **海光DCU协处理器全面兼容ROCmGPU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”，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。
  - 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。

海光DCU组成架构示意图

资料来源：海光信息招股说明书，中信证券研究部

- **目前公司将海光DCU产品规划为海光8000系列。**

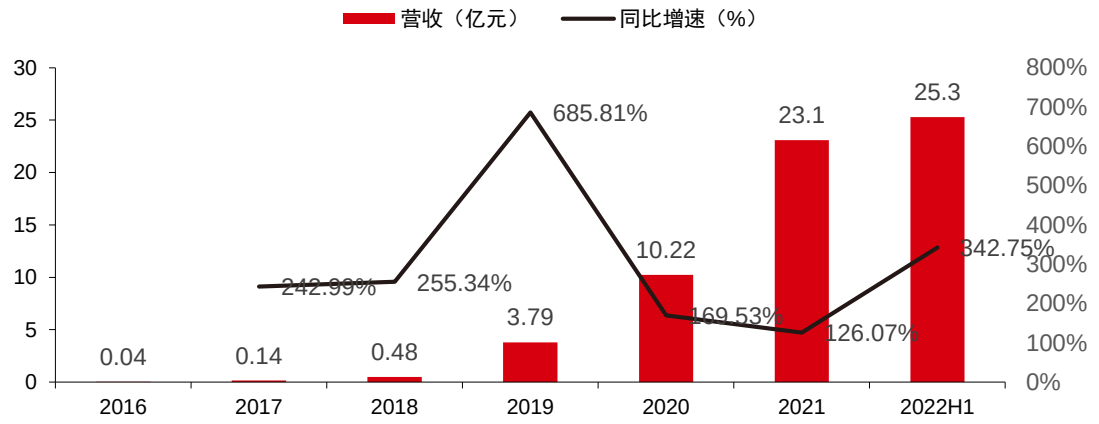
海光8100系列产品主要规格和特点

	海光8100
产品图片	
典型功耗	260-351W
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	①60-64个计算单元（最多4096个计算核心）
	②支持FP64、FP32、FP16、INT8、INT4
内存	①4个HBM2内存通道最高内存带宽为1TB/s
	②最高内存带宽为1TB/s
	③最大内存容量为32GB
I/O	①16LanePCIeGen4
	②DCU芯片之间高速互连

资料来源：海光信息招股说明书，中信证券研究部

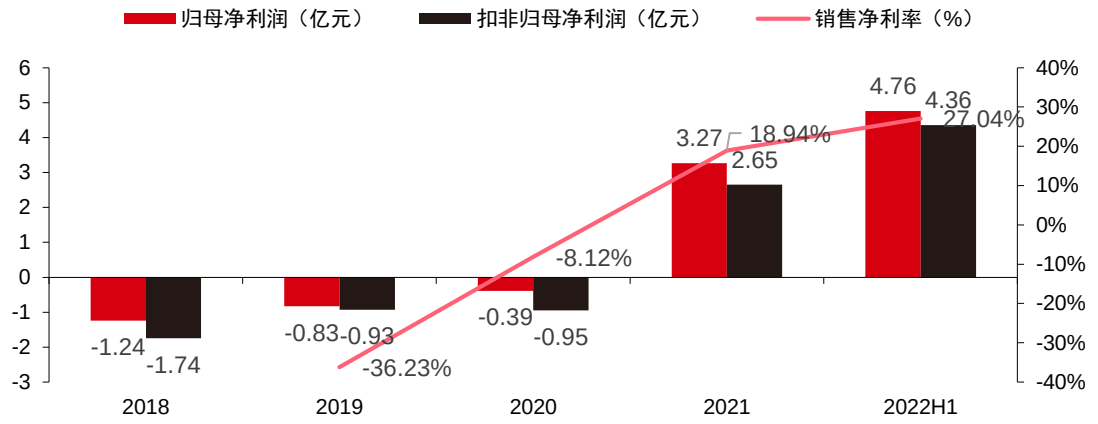
3.5 海光财务分析：收入高增净利扭亏，GPU逐渐起量

海光信息历年营收及增速



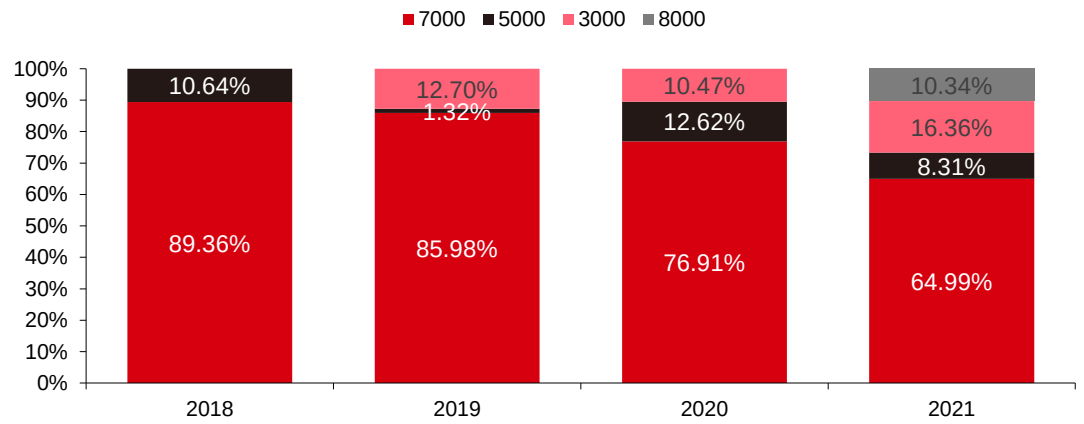
资料来源：海光信息招股说明书，中科曙光年报，中信证券研究部

海光信息历年归母净利润及增速



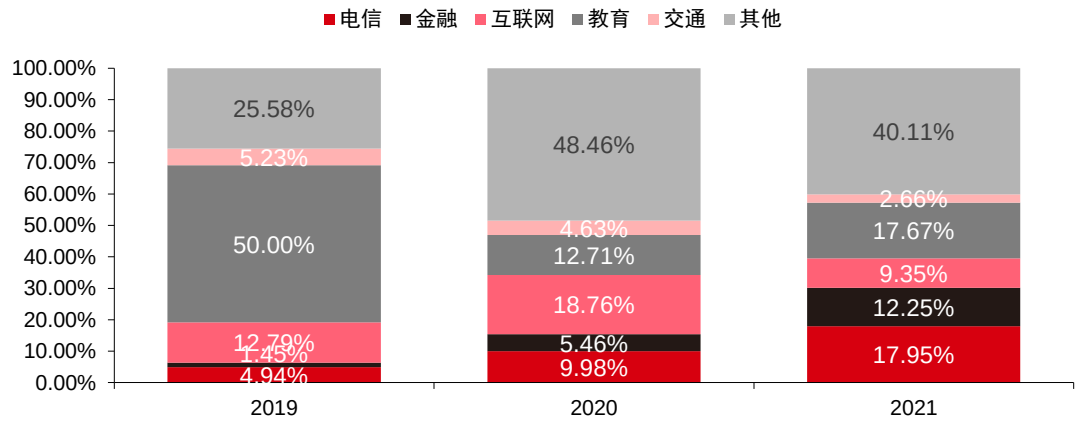
资料来源：海光信息招股说明书，中信证券研究部

海光信息历年分业务营收占比



资料来源：海光信息招股说明书，中信证券研究部

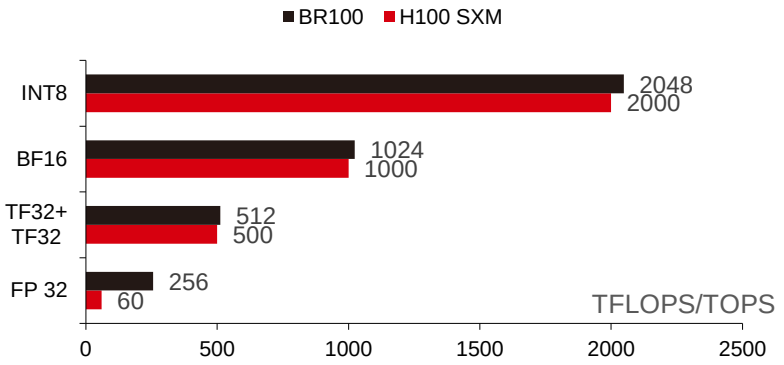
海光信息历年分行业营收占比



资料来源：海光信息招股说明书，中信证券研究部

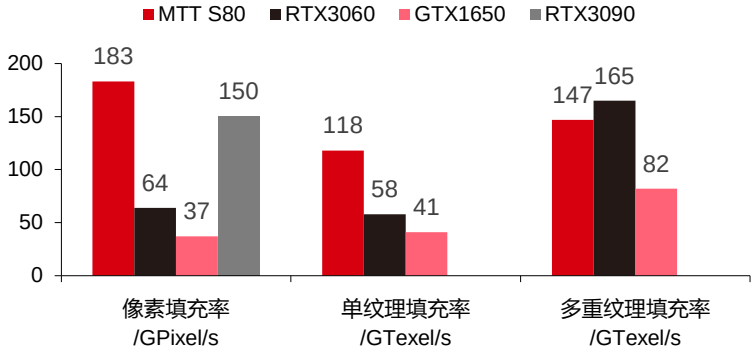
3.5 GPU创业公司：理论性能良好，长期前景可期

壁仞科技BR100与国际领先产品算力对比



资料来源：壁仞科技官网，中信证券研究部

摩尔线程MTT S80与英伟达产品理论性能对比

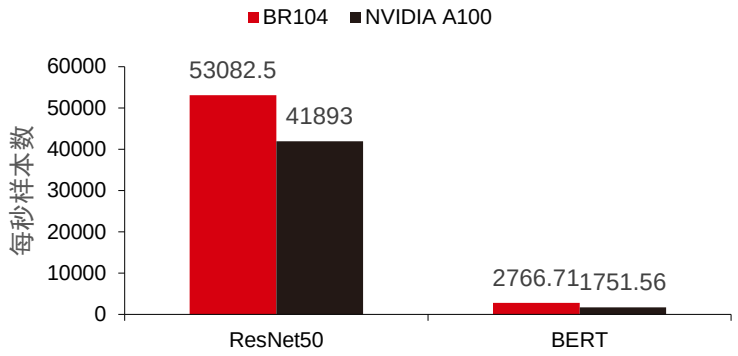


资料来源：FUN科技@bilibili，中信证券研究部

摩尔线程生态合作伙伴

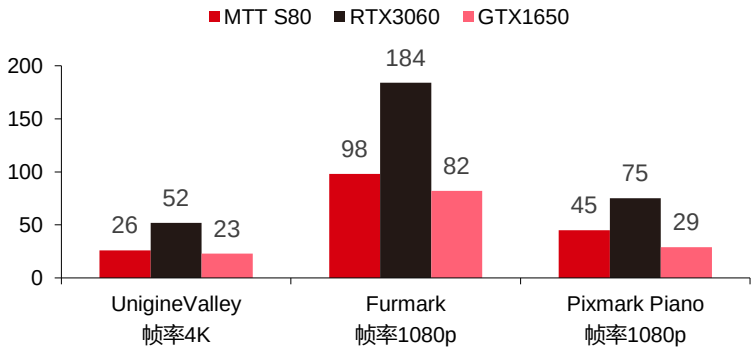
资料来源：摩尔线程2022春季发布会

壁仞科技BR104与英伟达A100单卡AI性能对比



资料来源：壁仞科技官网，中信证券研究部

摩尔线程MTT S80与英伟达产品实测帧率对比



资料来源：FUN科技@bilibili，中信证券研究部

4.风险因素

风险因素

- **1）产业链安全风险：**国外GPU行业起步较早，产业链中诸如EDA工具，芯片制造等重要环节国内与海外发展水平仍有较大差距。介于中美贸易、科技摩擦的背景，国内企业受到制裁导致原有产业链断裂的可能仍存在，对国内企业未来新产品研发进度、产品工艺更新、供应链保障等造成不利影响。

- **2）市场竞争加剧：**全球GPU市场由NVIDIA、AMD主导，国内市场存在部分已经形成规模的政策红利玩家和若干新兴企业。随着国内外厂商技术不断升级和国内企业持续壮大，GPU市场竞争程度或将加剧，对于上市公司的经营能力、技术升级等方面提出更高要求，公司未来业绩或将受影响。

- **3）商业需求不及预期风险：**由于国产GPU在性能和生态建设方面与NVIDIA、AMD等存在差距，在纯商业化领域失去政策驱动，可能因为自身产品竞争力不足，导致需求低于预期。

- **4）产品研发不及预期风险：**产品研发需要持续投入大量资金人才，且研发成果不确定性较高。倘若研发进度不及预期或研发失败，企业将可能面临亏损。

- **5）国产替代进程不及预期风险：**对于技术、政策等因素影响下，对国产替代的需求释放不及预期，将影响公司未来的业绩。

- **6）宏观经济环境风险：**面对宏观经济以及疫情等影响，全球范围的核心零部件供应链或将受到影响。

5.投资建议

投资建议

投资建议：

- 通过对GPU的各类重要参数的研究，我们提出GPU的**核心竞争力在于微架构等因素先进带来性能领先和与之适配的完善软硬件生态**。
- **借鉴这一研究框架并通过复盘NVIDIA/AMD（ATI）的竞争史，再次验证**NVIDIA凭借性能领先和生态完善长期占有GPU市场八成份额。AMD（ATI）也曾凭借Radeon 9800和RDNA架构系列产品实现性能反超。这些经验对国产厂商具有一定的借鉴意义，国产厂商正持续大力投入研发实现GPU架构创新升级和快速迭代步调，力争赶超国际领先水平；同时加大生态建设力度，打造自主开放的通用计算软件生态体系。
- 近年来，GPU行业迎来黄金发展期，游戏、数据中心、汽车市场已爆发大量需求；中长期来看，GPU产业有望逐步走向全面市场驱动。目前供给端国产GPU厂商在性能方面正在加速追赶，已开始具备应对需求爆发式增长的供给能力。加之国际科技制裁带来的发展机遇，**国产GPU厂商有望迎来成长黄金期**。
- 料外部不确定性背景下，国产GPU可控需求加速，伴随国际形势变化、政策大力扶持、游戏&AI、数据中心&汽车领域等行业对GPU需求持续增长、产品性能提升、产业生态完善，国产GPU厂商有望加速崛起。**重点看好GPU领域龙头厂商长期机遇，建议关注国产GPU龙头企业。1）推荐：海光信息（CPU+GPGPU）。建议关注景嘉微、寒武纪（电子覆盖）。2）一级市场（排名不分先后）：壁仞科技、摩尔线程、沐曦、天数智芯、登临科技、燧原科技等。**

中信证券
CITIC SECURITIES

感谢您的信任与支持！

THANK YOU

杨泽原（计算机行业首席分析师）

执业证书编号： S1010517080002

丁奇（云基础设施首席分析师）

执业证书编号： S1010519120003

免责声明

分析师声明

主要负责撰写本研究报告全部或部分内容的分析师在此声明：（i）本研究报告所表述的任何观点均精准地反映了上述每位分析师个人对标的证券和发行人的看法；（ii）该分析师所得报酬的任何组成部分无论是在过去、现在及将来均不会直接或间接地与研究报告所表述的具体建议或观点相联系。

一般性声明

本研究报告由中信证券股份有限公司或其附属机构制作。中信证券股份有限公司及其全球的附属机构、分支机构及联营机构（仅就本研究报告免责条款而言，不含CLSA group of companies），统称为“中信证券”。

本研究报告对于收件人而言属高度机密，只有收件人才能使用。本研究报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。本研究报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。中信证券并不因收件人收到本报告而视其为中信证券的客户。本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断并自行承担投资风险。

本报告所载资料的来源被认为是可靠的，但中信证券不保证其准确性或完整性。中信证券并不对使用本报告或其所包含的内容产生的任何直接或间接损失或与此有关的其他损失承担任何责任。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可跌可升。过往的业绩并不能代表未来的表现。

本报告所载的资料、观点及预测均反映了中信证券在最初发布该报告日期当日分析师的判断，可以在不发出通知的情况下做出更改，亦可因使用不同假设和标准、采用不同观点和分析方法而与中信证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。中信证券并不承担提示本报告的收件人注意该等材料的责任。中信证券通过信息隔离墙控制中信证券内部一个或多个领域的信息向中信证券其他领域、单位、集团及其他附属机构的流动。负责撰写本报告的分析师的薪酬由研究部门管理层和中信证券高级管理层全权决定。分析师的薪酬不是基于中信证券投资银行收入而定，但是，分析师的薪酬可能与投行整体收入有关，其中包括投资银行、销售与交易业务。

若中信证券以外的金融机构发送本报告，则由该金融机构为此发送行为承担全部责任。该机构的客户应联系该机构以交易本报告中提及的证券或要求获悉更详细信息。本报告不构成中信证券向发送本报告金融机构之客户提供的投资建议，中信证券以及中信证券的各个高级职员、董事和员工亦不为（前述金融机构之客户）因使用本报告或报告载明的内容产生的直接或间接损失承担任何责任。

评级说明

投资建议的评级标准		评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即：以报告发布日后的6到12个月内的公司股价（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。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准；韩国市场以科斯达克指数或韩国综合股价指数为基准。	股票评级	买入	相对同期相关证券市场代表性指数涨幅20%以上
		增持	相对同期相关证券市场代表性指数涨幅介于5%～20%之间
		持有	相对同期相关证券市场代表性指数涨幅介于-10%～5%之间
		卖出	相对同期相关证券市场代表性指数跌幅10%以上
	行业评级	强于大市	相对同期相关证券市场代表性指数涨幅10%以上
		中性	相对同期相关证券市场代表性指数涨幅介于-10%～10%之间
		弱于大市	相对同期相关证券市场代表性指数跌幅10%以上

免责声明

特别声明

在法律许可的情况下，中信证券可能（1）与本研究报告所提到的公司建立或保持顾问、投资银行或证券服务关系，（2）参与或投资本报告所提到的公司的金融交易，及/或持有其证券或其衍生品或进行证券或其衍生品交易，因此，投资者应考虑到中信证券可能存在与本研究报告有潜在利益冲突的风险。本研究报告涉及具体公司的披露信息，请访问https://research.citicsinfo.com/disclosure。

截至本报告发布日，中信证券股份有限公司及其另类投资子公司持有下述公司已发行股份的比例达到或超过1%：海光信息（688041），对应持股业务类别：自营，持股比例：0.31%；另类投资子公司，限售持股比例：0.25%/1.34%，限售起始日：2022年08月12日/2022年08月12日，限售期：24个月/12个月。

法律主体声明

本研究报告在中华人民共和国（香港、澳门、台湾除外）由中信证券股份有限公司（受中国证券监督管理委员会监管，经营证券业务许可证编号：Z20374000）分发。本研究报告由下列机构代表中信证券在相应地区分发：在中国香港由CLSA Limited（于中国香港注册成立的有限公司）分发；在中国台湾由CL Securities Taiwan Co., Ltd.分发；在澳大利亚由CLSA Australia Pty Ltd.（商业编号：53 139 992 331/金融服务牌照编号：350159）分发；在美国由CLSA（CLSA Americas, LLC除外）分发；在新加坡由CLSA Singapore Pte Ltd.（公司注册编号：198703750W）分发；在欧洲经济区由CLSA Europe BV分发；在英国由CLSA （UK）分发；在印度由CLSA India Private Limited分发（地址：8/F, Dalamal House, Nariman Point, Mumbai 400021；电话：+91-22-66505050；传真：+91-22-22840271；公司识别号：U67120MH1994PLC083118）；在印度尼西亚由PT CLSA Sekuritas Indonesia分发；在日本由CLSA Securities Japan Co., Ltd.分发；在韩国由CLSA Securities Korea Ltd.分发；在马来西亚由CLSA Securities Malaysia Sdn Bhd分发；在菲律宾由CLSA Philippines Inc.（菲律宾证券交易所及证券投资者保护基金会员）分发；在泰国由CLSA Securities (Thailand) Limited分发。

针对不同司法管辖区的声明

中国大陆：根据中国证券监督管理委员会核发的经营证券业务许可，中信证券股份有限公司的经营范围包括证券投资咨询业务。

中国香港：本研究报告由CLSA Limited分发。 本研究报告在香港仅分发给专业投资者（《证券及期货条例》（香港法例第571 章）及其下颁布的任何规则界定的），不得分发给零售投资者。就分析或报告引起的或与分析或报告有关的任何事宜，CLSA客户应联系CLSA Limited的罗鼎，电话：+852 2600 7233。

美国：本研究报告由中信证券制作。本研究报告在美国由CLSA（CLSA Americas, LLC除外）仅向符合美国《1934年证券交易法》下15a-6规则界定且CLSA Americas, LLC提供服务的“主要美国机构投资者”分发。对身在美国的任何人士发送本研究报告将不被视为对本报告中所评论的证券进行交易的建议或对本报告中所述任何观点的背书。任何从中信证券与CLSA获得本研究报告的接收者如果希望在美国交易本报告中提及的任何证券应当联系CLSA Americas, LLC（在美国证券交易委员会注册的经纪交易商），以及CLSA 的附属公司。

新加坡：本研究报告在新加坡由CLSA Singapore Pte Ltd.，仅向（新加坡《财务顾问规例》界定的）“机构投资者、认可投资者及专业投资者”分发。就分析或报告引起的或与分析或报告有关的任何事宜，新加坡的报告收件人应联系CLSA Singapore Pte Ltd，地址：80 Raffles Place, #18-01, UOB Plaza 1, Singapore 048624，电话：+65 6416 7888。因您作为机构投资者、认可投资者或专业投资者的身份，就CLSA Singapore Pte Ltd.可能向您提供的任何财务顾问服务，CLSA Singapore Pte Ltd豁免遵守《财务顾问法》（第110章）、《财务顾问规例》以及其下的相关通知和指引（CLSA业务条款的新加坡附件中证券交易服务C部分所披露）的某些要求。MCI（P）085/11/2021。

加拿大：本研究报告由中信证券制作。对身在加拿大的任何人士发送本研究报告将不被视为对本报告中所评论的证券进行交易的建议或对本报告中所载任何观点的背书。

英国：本研究报告归属于营销文件，其不是按照旨在提升研究报告独立性的法律要件而撰写，亦不受任何禁止在投资研究报告发布前进行交易的限制。本研究报告在英国由CLSA （UK）分发，且针对由相应本地监管规定所界定的在投资方面具有专业经验的人士。涉及到的任何投资活动仅针对此类人士。若您不具备投资的专业经验，请勿依赖本研究报告。

欧洲经济区：本研究报告由荷兰金融市场管理局授权并管理的CLSA Europe BV 分发。

澳大利亚：CLSA Australia Pty Ltd（“CAPL”）(商业编号：53 139 992 331/金融服务牌照编号：350159) 受澳大利亚证券与投资委员会监管，且为澳大利亚证券交易所及CHI-X的市场参与主体。本研究报告在澳大利亚由CAPL仅向“批发客户”发布及分发。本研究报告未考虑收件人的具体投资目标、财务状况或特定需求。未经CAPL事先书面同意，本研究报告的收件人不得将其分发给任何第三方。本段所称的“批发客户”适用于《公司法（2001）》第761G条的规定。CAPL研究覆盖范围包括研究部门管理层不时认为与投资者相关的ASX All Ordinaries 指数成分股、离岸市场上市证券、未上市发行人及投资产品。CAPL寻求覆盖各个行业中与其国内及国际投资者相关的公司。

印度：CLSA India Private Limited，成立于 1994 年 11 月，为全球机构投资者、养老基金和企业提供股票经纪服务（印度证券交易委员会注册编号：INZ000001735）、研究服务（印度证券交易委员会注册编号：INH000001113）和商人银行服务（印度证券交易委员会注册编号：INM000010619）。CLSA 及其关联方可能持有标的公司的债务。此外，CLSA及其关联方在过去 12 个月内可能已从标的公司收取了非投资银行服务和/或非证券相关服务的报酬。 如需了解CLSA India“关联方”的更多详情，请联系Compliance-India@clsa.com。