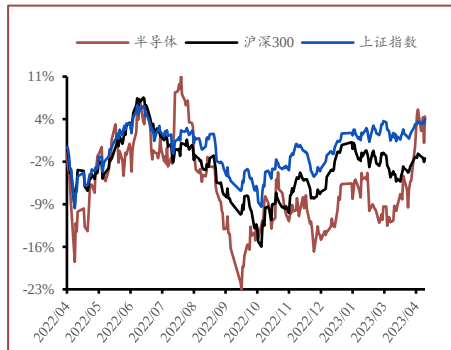


AI 模型乘风起，GPU 掌舵算力大时代

■ 证券研究报告

投资评级:看好(维持)

最近 12 月市场表现



分析师 张益敏

SAC 证书编号: S0160522070002

zhangym02@ctsec.com

相关报告

1. 《行业周期与政策共振，看好自主可控和周期复苏》 2023-03-15

AI 算力行业深度

核心观点

- ❖ **GPU 掌舵 AI 算力大时代，千亿级市场再迎增量：**GPU 因其强大的并行计算能力而广泛应用于人工智能、图像渲染、科学计算等领域。AI、自动驾驶与游戏市场是 GPU 需求增长的主要场景，据 Global Market Insights 数据，全球 GPU 市场预计将以 CAGR 25.9% 持续增长，至 2030 年达到 4000 亿美元规模。其中 AI 领域大语言模型的持续推出以及参数数量的不断增长有望驱动模型训练端、推理端 GPU 需求快速增长。
- ❖ **微架构和平台生态共筑竞争壁垒：**GPU 的微架构设计是决定硬件性能的关键，全球龙头厂商英伟达与 AMD 均以保持架构升级节奏以及制程升级速率来保证产品竞争力。此外，成熟且完善的平台生态形成的强大用户粘性将在长时间内塑造 GPU 厂商的软实力，以英伟达通用计算平台 CUDA 为例，从软件栈的完整度和对硬件性能的高效利用角度出发降低了通用计算 GPU 开发者编译难度，建立起卡位全球的开发生态，从而实现长期竞争壁垒。
- ❖ **兼容主流生态对标行业龙头，国内厂商持续发力：**近年来，国产 GPU 厂商在图形渲染 GPU 和高性能计算 GPGPU 领域上均推出了较为成熟的产品，在性能上不断追赶行业主流产品，在特定领域达到业界一流水平。生态方面国产厂商大多兼容英伟达 CUDA，融入大生态进而实现客户端不断导入。在高端 GPU 芯片进口受限的背景下，国产 GPU 厂商预计将乘政策东风，抓住国产替代契机快速成长。
- ❖ **建议关注：**1) 已上市标的：寒武纪、海光信息、景嘉微、芯原股份、龙芯中科；2) 未上市标的：壁仞科技、摩尔线程、芯动科技、兆芯、天数智芯、沐曦。
- ❖ **风险提示：**技术迭代风险、宏观经济风险、国产替代风险、行业竞争风险。

表 1：重点公司投资评级：

代码	公司	总市值 (亿元)	收盘价 (04.14)	EPS (元)			PE			投资评级
				2021A	2022E	2023E	2021A	2022E	2023E	
688256	寒武纪-U	781.58	195.00	-2.91	-1.80	-1.04	-18.75	-108.30	-187.09	未覆盖
688041	海光信息	1,790.67	77.04	0.38	0.56	0.79	116.27	74.75	52.60	增持
300474	景嘉微	480.70	105.75	0.65	0.93	1.30	163.45	113.48	81.41	未覆盖
688521	芯原股份	474.80	95.39	0.15	0.30	0.54	297.18	319.16	177.45	增持
688047	龙芯中科	635.91	158.58	0.13	0.59	1.09	674.49	269.35	145.94	未覆盖

数据来源：wind 数据，财通证券研究所

注：景嘉微 22 年数据为预测数据，其余公司 22 年数据为年报数据或 22 年业绩快报数据

注：芯原股份、海光信息预测数据来自于财通证券研究所，其余预测数据来自于 一致预期（截止 2023/04/16）。

内容目录

1	行业概况：GPU 掌舵 AI 算力大时代，千亿级市场再迎增量	7
1.1	GPU：提供大规模并行计算解决方案	7
1.2	“AI+汽车+游戏”三驾马车驱动行业发展	8
1.3	大语言模型助推 GPU 算力需求增长	16
2	微架构和平台生态共筑竞争壁垒	19
2.1	微架构：统一计算单元解锁通用计算时代	19
2.2	架构迭代与制程升级是 GPU 性能的生命线	21
2.3	成熟的平台生态是 GPU 厂商的护城河	23
3	国内外发展现状：海外龙头领跑，国产持续发力	30
3.1	海外龙头：深耕多年，技术引领行业	30
3.1.1	英伟达	30
3.1.2	AMD	32
3.1.3	高通	33
3.1.4	Imagination	35
3.1.5	ARM	36
3.2	兼容主流生态对标行业龙头，国内厂商持续发力	37
3.3	高端芯片进口遭限制，国产厂商替代迎契机	39
4	建议关注	40
4.1	寒武纪	40
4.2	海光信息	41
4.3	景嘉微	43
4.4	芯原股份	44
4.5	龙芯中科	46
4.6	壁仞科技（非上市）	47
4.7	摩尔线程（非上市）	48
4.8	芯动科技（非上市）	49
4.9	兆芯（非上市）	50
4.10	天数智芯（非上市）	51
4.11	沐曦（非上市）	52
5	风险提示	53

图表目录

图 1. CPU 架构示意图	7
图 2. GPU 架构示意图	7
图 3. CPU+GPU 的异构计算.....	8
图 4. GPT-3 Transformer 模型结构.....	8
图 5. LLM 基础模型.....	8
图 6. 百度文心大模型.....	9
图 7. 文心大模型性能评测.....	9
图 8. 阿里通义大模型层次示意图.....	10
图 9. 多模态模块化设计.....	10
图 10. 阿里所有产品未来将接入“通义千问”大模型.....	10
图 11. ModelArts 平台架构	11
图 12. 主流 NLP 预训练模型规模.....	12
图 13. 深度学习模型对算力的需求增速.....	12
图 14. 中国 AI 芯片市场份额（按类型）	12
图 15. 全球自动驾驶渗透率.....	13
图 16. 汽车自动驾驶分级以及对算力需求.....	13
图 17. Orin 系统架构	14
图 18. NVIDIA 自动驾驶平台算力升级路线图	14
图 19. 光线追踪算法过程.....	15
图 20. NVIDIA RTX 平台	15
图 21. 英伟达中端 GPU 显卡单位价格性能持续升级	15
图 22. 全球 GPU 市场规模（十亿美元）	16
图 23. 全球独立 GPU 市场占比（按厂商）	16
图 24. Nvidia Tesla 整体架构图.....	19
图 25. Nvidia Tesla 微架构中 TPC 架构图.....	20
图 26. 图像渲染管线相对固定.....	20
图 27. Nvidia Tesla 微架构中 SM 架构图.....	20
图 28. Nvidia Fermi 架构图.....	21
图 29. Nvidia Fermi 微架构中 SM 架构图.....	21
图 30. NVIDIA GPU 架构演进历史.....	22

图 31. AMD GPU 架构演进历史	22
图 32. GPU 在并行计算的应用	24
图 33. CUDA 加速计算解决方案	24
图 34. CUDA 支持 CPU+GPU 的异构计算	24
图 35. CUDA 编程模式示意图	25
图 36. CUDA 在 Host 中的函数库	25
图 37. Kernel 是 GPU 内核函数	26
图 38. GPU 上的 Kernel 执行	26
图 39. CUDA 存储结构	26
图 40. GPU 中的内存层次结构	26
图 41. CUDA 是 GPU 计算生态系统	27
图 42. CUDA 提供强大的开发支持工具	27
图 43. AMD ROCm 5.0	28
图 44. 异构计算框架 OpenCL	29
图 45. 英伟达四大业务	30
图 46. 英伟达下游应用行业	30
图 47. 英伟达营业收入及增速(亿美元)	31
图 48. 英伟达净利润及增速(亿美元)	31
图 49. AMD 业务概览	32
图 50. AMD 核心技术概览	32
图 51. AMD 营业收入及增速(亿美元)	33
图 52. AMD 净利润及增速(亿美元)	33
图 53. 高通骁龙移动平台	34
图 54. 高通营业收入及增速(亿美元)	35
图 55. 高通净利润及增速(亿美元)	35
图 56. IMG B 系列产品	35
图 57. IMG B 系列与 A 系列性能对比	35
图 58. ARM 整体设计解决方案	36
图 59. ARM Immortals-G715 架构	37
图 60. 寒武纪营业收入及增速(亿元)	41
图 61. 寒武纪归母净利润及增速(亿元)	41
图 62. 海光信息 DCU 基本组成架构	42
图 63. 海光深算 DCU 完善软件栈支持	42

图 64. 海光信息营业收入及增速(亿元).....	43
图 65. 海光信息归母净利润及增速(亿元).....	43
图 66. 景嘉微营业收入及增速(亿元).....	44
图 67. 景嘉微归母净利润及增速(亿元).....	44
图 68. 一站式芯片定制服务.....	44
图 69. 芯原 IP 产品阵.....	44
图 70. 芯原股份营业收入及增速(亿元).....	45
图 71. 芯原股份归母净利润及增速(亿元).....	45
图 72. 龙芯中科芯片产品.....	46
图 73. 龙芯中科自主生态技术架构.....	46
图 74. 龙芯中科营业收入及增速(亿元).....	47
图 75. 龙芯中科归母净利润及增速(亿元).....	47
图 76. BR100 系列通用 GPU 芯片.....	47
图 77. BIRENSUPA 软件平台.....	47
图 78. 开发者软件 MT GPU Management Center.....	48
图 79. 第一代 MUSA 架构.....	48
图 80. 芯动科技的定制服务.....	49
图 81. 芯动科技核心产品.....	49
图 82. 基于 GPU 的 TEE 隐私计算解决方案.....	51
图 83. 公司发布的人工智能开源平台 DeepSpark.....	51
 表 1. NVIDIA 架构演进历史.....	 7
表 2. GPU、FPGA、ASIC 指标对比.....	12
表 3. 训练端 GPU 需求增量测算.....	17
表 4. 推理端 GPU 需求增量测算.....	18
表 5. ROCm 与 CUDA 模块对比.....	28
表 6. OpenCL 与 CUDA 对比.....	29
表 7. 英伟达 40 系列显卡产品参数规格.....	30
表 8. 英伟达 A100、H100 系列产品规格参数.....	31
表 9. AMD7000 显卡参数规格.....	32
表 10. AMD Instinct 系列产品规格.....	33
表 11. 高通 Adreno 7 系列产品规格.....	34
表 12. Imagination IMG B 系列产品简介.....	36

表 13. 图形渲染 GPU 产品性能对比	37
表 14. 通用计算 GPU 产品性能对比	38
表 15. 国内 GPU、半导体相关政策（部分）	39
表 16. 寒武纪主要产品性能参数.....	40
表 17. 思元 370 系列产品信息.....	41
表 18. 海光 DCU 产品主要参数	42
表 19. 景嘉微 JM9 产品性能参数.....	43
表 20. 芯原 Vivante®图形处理器 IP 各系列产品参数.....	45
表 21. 壁仞 BR100 系列产品参数	48
表 22. 摩尔线程产品参数.....	49
表 23. 芯动风华系列 GPU 主要参数	50
表 24. 兆芯 Arise-GT10C0 芯片介绍.....	50
表 25. 天数智芯 BI-V100 主要产品参数	51
表 26. 沐曦 GPU 产品矩阵	52

1 行业概况：GPU 掌舵 AI 算力大时代，千亿级市场再迎增量

1.1 GPU：提供大规模并行计算解决方案

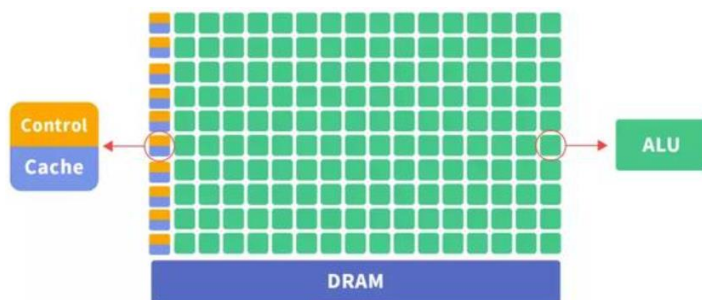
GPU，专注图像处理。GPU（图形处理器）最初是为了解决 CPU 在图形处理领域性能不足的问题而诞生。CPU 作为核心控制计算单元，高速缓冲存储器（Cache）、控制单元（Control）在 CPU 硬件架构设计中所占比例较大，主要为实现低延迟和处理单位内核性能要求较高的工作而存在，而计算单元（ALU）所占比例较小，这使得 CPU 的大规模并行计算表现不佳。GPU 架构内主要为计算单元，采用极简的流水线进行设计，适合处理高度线程化、相对简单的并行计算，在图像渲染等涉及大量重复运算的领域拥有更强运算能力。

图1.CPU 架构示意图



数据来源：cetrend，财通证券研究所

图2.GPU 架构示意图



数据来源：cetrend，财通证券研究所

GPGPU，脱胎于 GPU，通用性提升。GPU 计算单元既可运用于图形渲染领域，也能够进行通用计算。传统 GPU 应用局限于图形渲染计算，而面对非图像显示领域并涉及大量并行运算的领域，比如 AI、加密解密、科学计算等领域则更需要通用计算能力。随着 GPU 可编程性的不断提高，去掉或减弱 GPU 的图形显示部分能力，全部投入通用计算的 GPGPU（通用计算处理器）应运而生。

表1.NVIDIA 架构演进历史

特点：固定流水线 不可编程 • 代表产品：SGI Iris、3dfx Voodoo、NVIDIA TNT2、ATI Rage	特点：可对Shader进行 一定程度的编程 • 代表产品：GeForce FX、Radeon 9000	特点：统一Shader 可灵活编程 • 代表产品：GTX 200、Fermi、Kepler、Maxwell...	特点：无固定流水线 可随意编程 • 代表产品：Intel Larabee、NVIDIA Fermi
ASIC时代	弱编程时代	GPGPU时代	超GPU时代
1990-2000	2000-2005	2006-至今	2008-至今

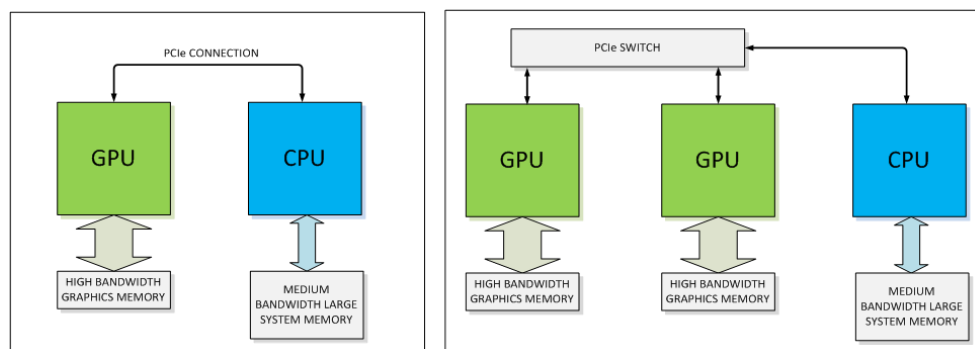
数据来源：CSDN，财通证券研究所

敬请参阅尾页重要声明及财通证券股票和行业评级标准

7

CPU+GPU 异构计算解决多元化计算需求。使用不同的体系架构的计算单元组成混合系统，GPU 作为协处理器负责并行加速计算，CPU 作为控制中心的异构计算面对复杂场景可实现更优性能。

图3.CPU+GPU 的异构计算

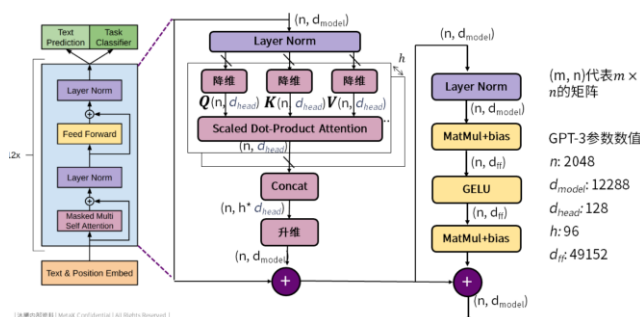


数据来源：NVIDIA，财通证券研究所

1.2 “AI+汽车+游戏”三驾马车驱动行业发展

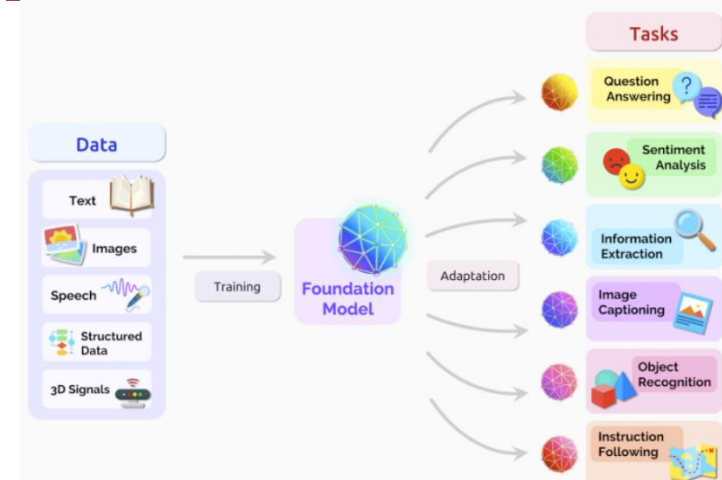
大语言模型开启 AI 元年。2022 年 11 月，OpenAI 推出基于大型语言模型 GPT-3 的 AI 对话机器人 ChatGPT，其可以与用户进行富有逻辑和创造力的自然语言对话。2017 年由 Google 提出的 Transformer 模型是大型语言模型发展的里程碑，Transformer 是一种基于注意力（Attention）机制构建的神经网络模型，克服了传统的递归神经网络（RNN）和卷积神经网络（CNN）在自然语言处理时容易被无关信息干扰的缺点，能够更好的理解长序列和上下文的关系。

图4.GPT-3 Transformer 模型结构



数据来源：MetaX，财通证券研究所

图5.LLM 基础模型



数据来源：《On the Opportunities and Risks of Foundation Models》Rishi Bommasani, i Drew A. Hudson 等，财通证券研究所

国内 AI 巨头持续跟进，大模型产业迎发展契机。腾讯、阿里、百度以及华为等厂商都已布局大模型产业，以“通用大模型+专精小模型”的层次化协同发展模式持续发力。

百度是国内最早进行大模型研发的科技企业之一，立足文心 NLP 大模型推出“文心一言”对话机器人（Ernie Bot）。百度在 2019 年 3 月率先发布中国首个正式开放的预训练模型文心大模型（Ernie）1.0，2021 年 12 月，文心大模型 3.0 参数突破千亿，升级为全球首个知识增强千亿大模型，成为目前为止全球最大的中文单体模型，根据 IDC 发布的《2022 中国大模型发展白皮书》，文心大模型在国内市场格局中处于第一梯队，产品能力、生态能力、应用能力均处于行业领先地位。2023 年 3 月 16 日，百度正式发布“文心一言”对话机器人，拥有文学创作、商业文案创作、数理逻辑推理、中文理解和多模态生成五大能力，表现出对文本语义的深度理解。

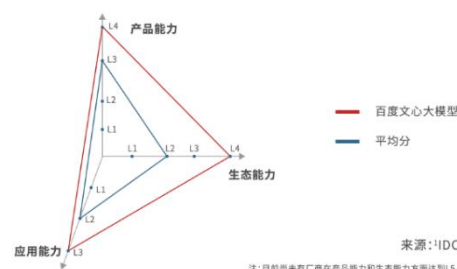
图6.百度文心大模型

产品与社区	文心·悟	文心·悟	杨尚社区		
	AI 艺术创作/短视频平台	大模型生态合作伙伴生态体系	大模型生态与开发者社区		
工具与平台	EasyDL·大模型	BML·大模型	大模型 API		
	智文™ AI 开发平台	智文™ AI 开发平台			
行业大模型	大模型套件				
	数据标注与处理	大模型精调	大模型压缩	高性能部署	
	大模型精调		高性能部署		
	大模型精调		高性能部署		
文心大模型	行业大模型				
	国图·百度·文心				
	通策·百度·文心				
	航天·百度·文心				
	人民同·百度·文心				
	冰城·百度·文心				
	电影翻译·百度·文心				
	深探·百度·文心				
	吉利·百度·文心				
	泰康·百度·文心				
TCL·百度·文心					
群海·百度·文心					
NLP 大模型					
CV 大模型					
跨模态大模型					
生物计算大模型					
医疗 EHRNet·Health	金融 EHRNet·Finance	商品图片标签生成 EHRNet·Image	文图生成 EHRNet·VLCG	文档智能 EHRNet·Layout	化合物表征学习 HoloChem
对话 PLATO	搜索 EHRNet·Search	信息抽取 EHRNet·UE	OCR图像表征学习 EHRNet·DocuText		
跨语言 EHRNet·AI	代码 EHRNet·Code	图网络 EHRNet·Gage	多任务视觉表征学习 EHRNet·LIFT		蛋白质结构预测 HoloProtein
	语言理解与生成		视觉处理	语言	地理
			多任务学习	自监督预训练	表征学习
EHRNet 3.0 1T1Y	EHRNet 3.0 1T1Y	麒麟·百度·文心	EHRNet 3.0 20m		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ESNet	ESNet	ESNet	ESNet		
ES					

图7.文心大模型性能评测



中国大模型市场2022年评估结果—百度文心



数据来源：百度文心大模型，财通证券研究所

数据来源：IDC，财通证券研究所

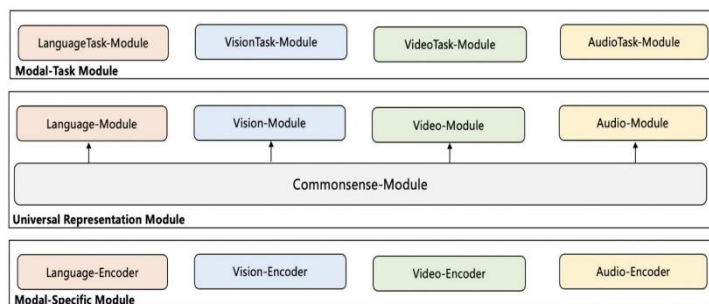
阿里达摩院推通义大模型，打造国内首个 AI 统一底座。2022 年 9 月 2 日，阿里达摩院在世界人工智能大会大规模预训练模型主题论坛上发布了最新的“通义”大模型，其打造了国内首个 AI 统一底座，构建了通用与专业模型协同的层次化人工智能体系，“统一学习范式”是通义大模型的最大亮点，通过多模态统一模型 M6-OFA 完成了架构、模块与任务的三大统一，赋予模型不新增结构即可处理包括图像描述、文档摘要、视觉定位等单模态和跨模态任务的能力。“模块化设计”也是模型特点之一，其借鉴了人脑“能力模块”结构，采用模块化 Transformer Encoder-Decoder 结构，切分出基础层、通用层、任务层、功能性四大模块，每个模块间相互解耦，分工合作。该设计便于对不同板块进行微调与继续训练，以实现大模型的轻量化。

图8.阿里通义大模型层次示意图



数据来源：阿里云发者社区，财通证券研究所

图9.多模态模块化设计



数据来源：阿里云发者社区，财通证券研究所

阿里巴巴集团董事会主席兼 CEO、阿里云智能集团 CEO 张勇在 4 月 11 日阿里云峰会上表示，阿里巴巴所有产品未来将接入“通义千问”大模型，进行全面改造，未来有望重塑产品格局。

图10.阿里所有产品未来将接入“通义千问”大模型

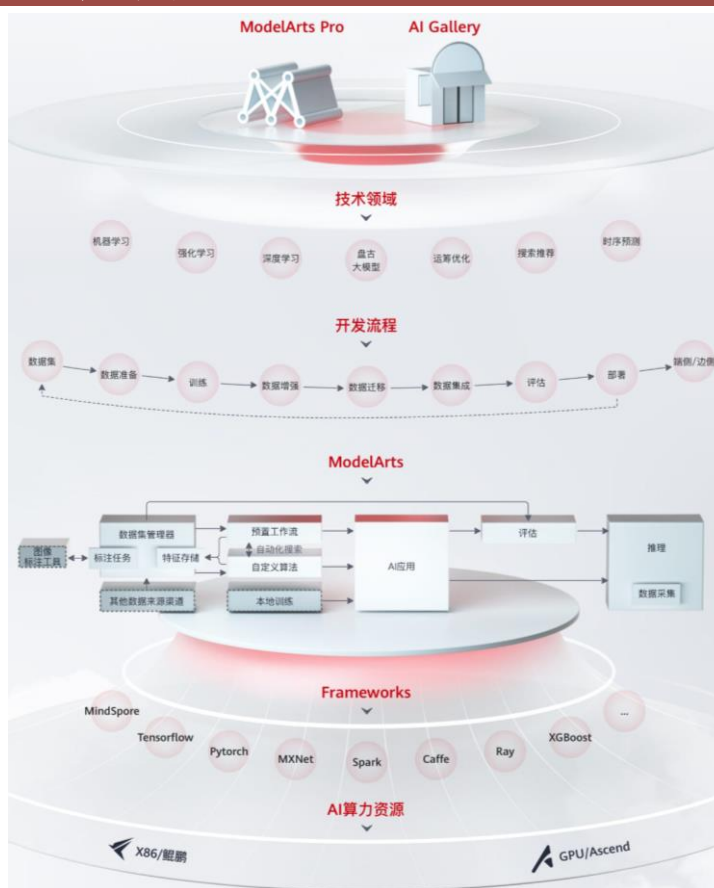


数据来源：2023 阿里云峰会，财通证券研究所

华为盘古大模型基于其 ModelArts 平台开发，模型泛化有望多场景落地。

ModelArts 平台为机器学习与深度学习提供海量数据预处理及交互式智能标注、大规模分布式训练、自动化模型生成，及端-边-云模型按需部署能力。盘古大模型基于 ModelArts 开发，由 NLP 大模型、CV 大模型、多模态大模型、科学计算大模型多个大模型构成，通过模型泛化可在不同部署场景下抽取出不同大小的模型，动态范围可根据需求调整，从特定的小场景到综合性的复杂大场景均能覆盖。目前，盘古大模型已经在能源、零售、金融、工业、医疗、环境、物流等 100 多个行业场景完成验证。

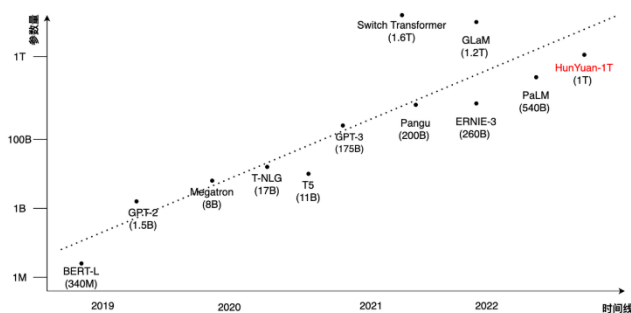
图11.ModelArts 平台架构



数据来源：华为云官网，财通证券研究所

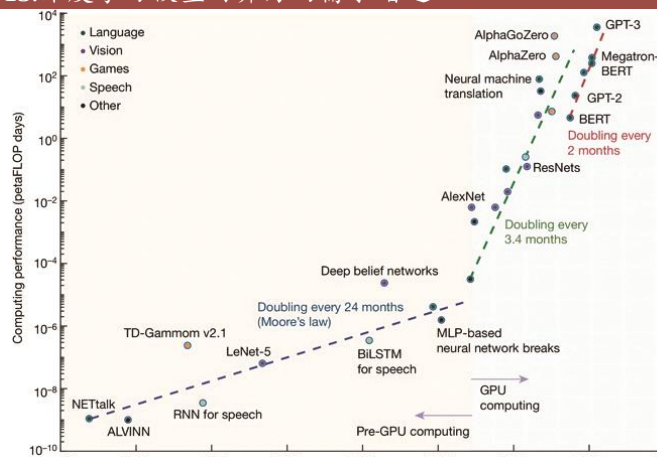
参数数量是决定模型表现的最重要因素。大语言模型的特点是拥有强大的自学习能力，随着训练数据集和模型参数的增加，可以显著提高模型的泛化能力和通用能力，模型规模的扩大已经成为了大语言模型的发展趋势。以 OpenAI 为例，其初代 GPT 模型参数量仅有 15 亿，而基于 GPT-3 的 chatGPT 参数量已经达到了 1750 亿，目前主流 AI 厂商都进入了“千亿参数时代”。模型表现改善的同时，不断增长的参数量对硬件算力提出了更高的要求。据 OpenAI 研究表明，最大的 AI 训练模型所需的算力每 3-4 个月翻倍，而 2012-2018 年间这个指标增长超过 300,000 倍。

图12.主流 NLP 预训练模型规模



数据来源: CSDN, 财通证券研究所

图13.深度学习模型对算力的需求增速



数据来源: 《Intelligent Computing: The Latest Advances, Challenges, and Future》
SHIQIANG ZHU、TING YU 等, 财通证券研究所

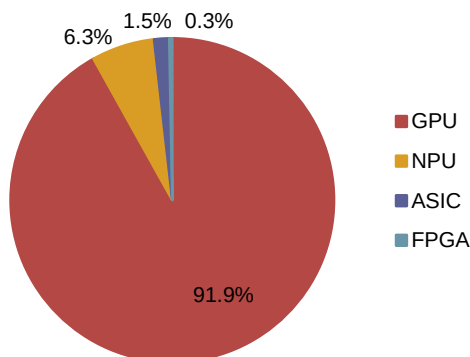
GPU 已成为 AI 加速芯片通用性解决方案, 提供大语言模型推理训练所需的海量算力。为构建有效的 AI 部署方案, CPU 和加速芯片结合的异构计算是经典的计算框架, 目前最常见的 AI 加速芯片主要为 GPU、FPGA 和 ASIC, 而 GPU 凭借其高性能、高灵活度特点成为 AI 加速方案首选。

表2.GPU、FPGA、ASIC 指标对比

	GPU	FPGA	ASIC
灵活性	中	高	低
性能	中	低	高
同构性	高	中	低
功耗	高	中	低
成本	中	低	高

数据来源: CSDN, 财通证券研究所

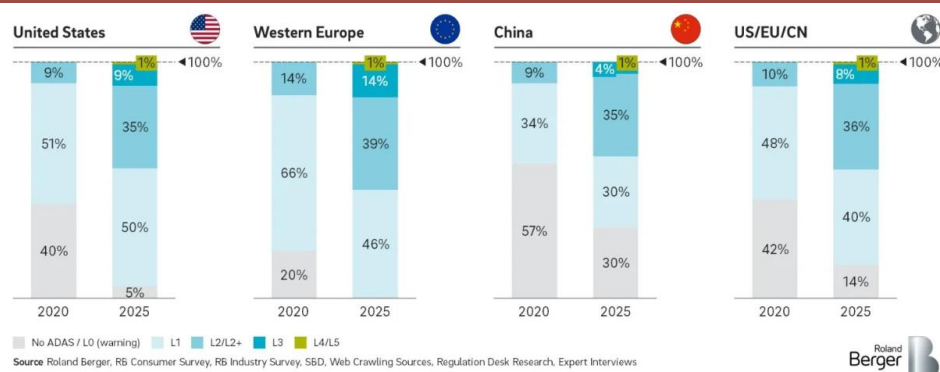
图14.中国 AI 芯片市场份额 (按类型)



数据来源: IDC, 财通证券研究所

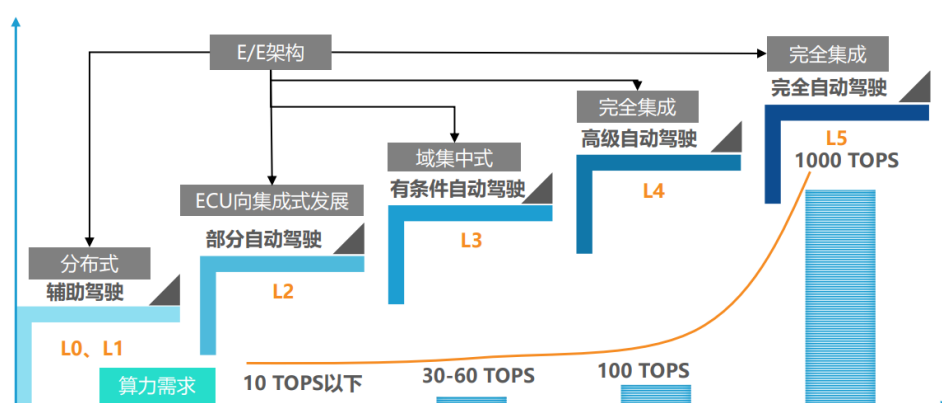
自动驾驶升级推动边缘计算需求增加，GPU 车载领域价值逐步显现。在云计算架构中，数据通过高速网络传输至拥有大规模高性能计算设备的云计算中心进行计算，而边缘计算则将数据计算与储存集中在靠近数据源头的本地设备上，能够更快的响应计算需求。自动驾驶是边缘计算架构最前沿的应用场景之一，目前大多数自动驾驶处于 L2-L3（部分自动驾驶）级别，而要实现 L4-L5 级别高度自动驾驶，则需要人工智能短时、高频地处理大量路况信息并自主完成大部分决策，因此需要 GPU 为汽车芯片提供更多计算能力来处理复杂数据。根据地平线对 OEM 厂商需求情况的分析，更高级别的自动驾驶意味着更高的算力需求，L2 级别需要 2 TOPS、L3 级别需要 24 TOPS、L4 级需要 320 TOPS，L5 级则需要 4000+ TOPS。

图15.全球自动驾驶渗透率



数据来源：Rolandberger，财通证券研究所

图16.汽车自动驾驶分级以及对算力需求



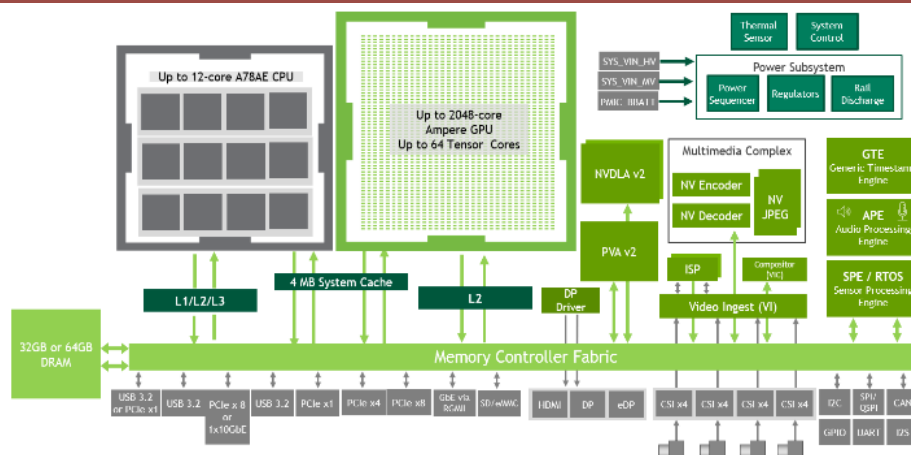
数据来源：亿欧智库《2021 中国车联网行业发展趋势研究报告：渐入佳境，一触即发》，财通证券研究所

GPU 提供核心计算能力，是自动驾驶算力升级趋势关键。目前，市面上主流的自动驾驶芯片采用 NVIDIA 推出的 Orin 系统级芯片（SoC），Orin 集成 NVIDIA

Ampere 架构 GPU 和 Arm Hercules 内核 CPU 以及全新深度学习加速器

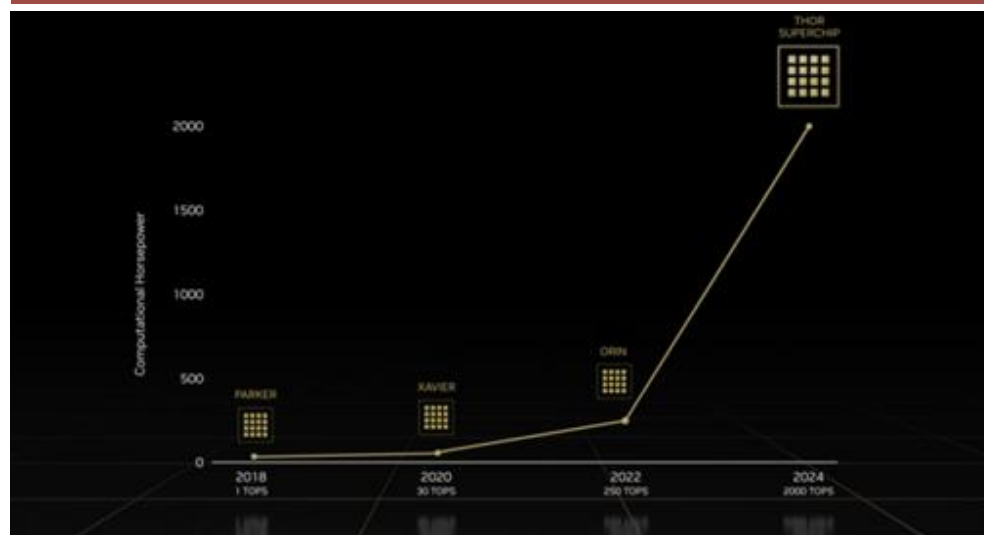
(DLA) 和计算机视觉加速器 (PVA)，可以提供每秒 254TOPS 的计算能力，几乎是 NVIDIA 上一代系统级芯片 Xavier 性能的 7 倍。而根据英伟达公告，其预计在 2024 年发布下一代车载系统级芯片 Thor，通过更新芯片内含的 GPU 架构，Thor 预计可以为自动驾驶汽车提供约 2000 TOPS 的计算能力。

图17.Orin 系统架构



数据来源：NVIDIA，财通证券研究所

图18.NVIDIA 自动驾驶平台算力升级路线图

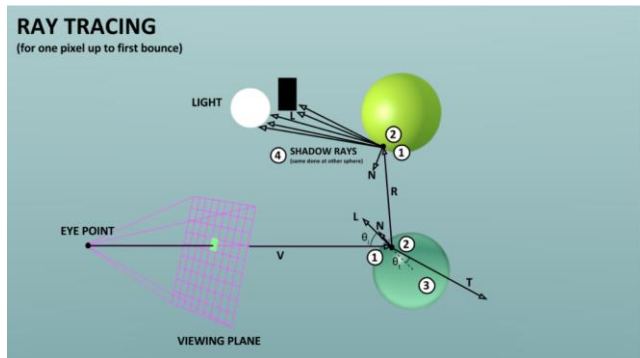


数据来源：NVIDIA，财通证券研究所

游戏市场画质升级驱动 GPU 显卡性能升级需求。GPU 最初作为图形处理器而诞生，在游戏显卡市场伴随玩家对游戏品质的追求不断提升，以光线追踪算法 (Ray Tracing) 为代表的特殊渲染算法更多的应用到游戏显卡以提升显示画质。

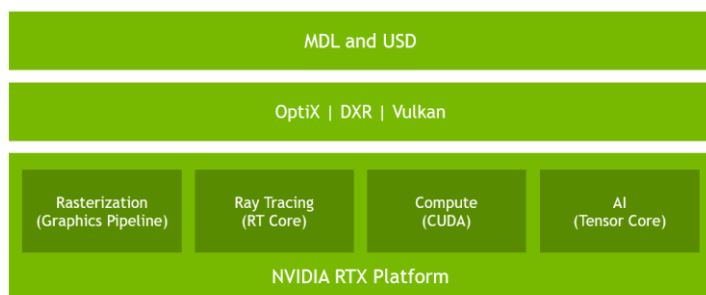
2018 年，NVIDIA 联合 Microsoft 共同发布了 RTX（Ray Tracing X）标准，NVIDIA 也在其同年发布的 Turing 架构 GPU 中引入了加速光线追踪计算的 RT Core，实现了光线追踪的实时化。光追通过在场景中发射光线并跟踪每个像素的光线路径来模拟真实的光传播，在提供更具真实感的画面效果的同时对于计算复杂度以及计算量需求大幅增加，整体游戏市场画质升级将驱动 GPU 显卡性能持续升级

图19.光线追踪算法过程



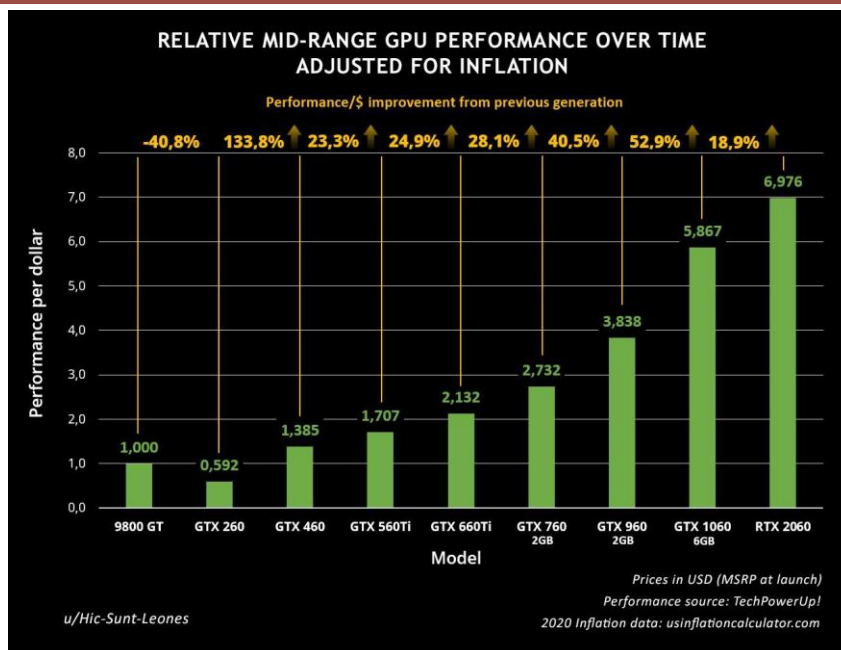
数据来源：CSDN，财通证券研究所

图20.NVIDIA RTX 平台



数据来源：NVIDIA，财通证券研究所

图21.英伟达中端 GPU 显卡单位价格性能持续升级

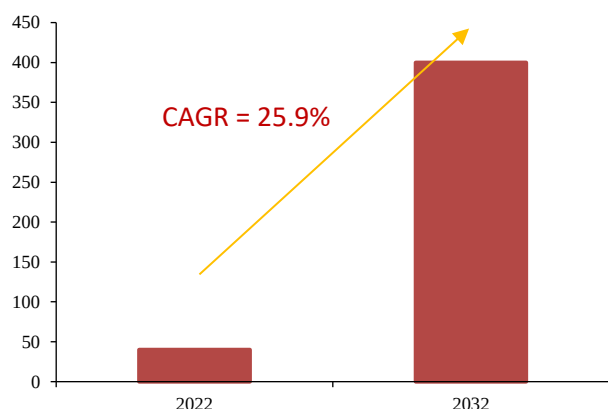


数据来源：Reddit，财通证券研究所

1.3 大语言模型助推 GPU 算力需求增长

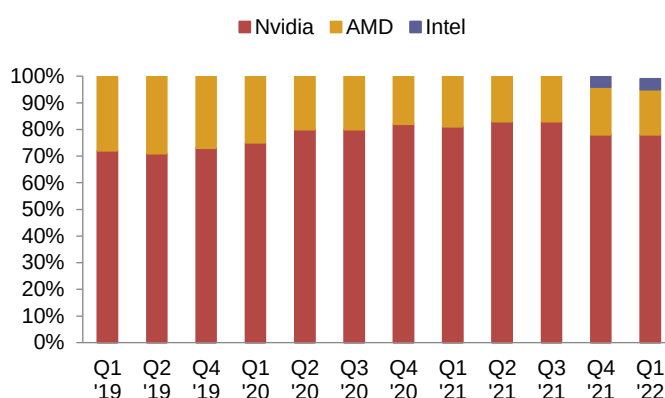
市场对 3D 图像处理和 AI 深度学习计算等需求不断增加，GPU 市场保持高速增长。据 Global Market Insights 的数据，全球 GPU 市场预计将以 CAGR 25.9% 持续增长，至 2030 年达到 4000 亿美元规模。在 GPU 市场中，NVIDIA 依靠在深度学习、人工智能等领域布局的先发优势并凭借其优异产品性能以及成熟的生态平台长期处于领导地位，根据 JPR 数据，2022 年 Q1，NVIDIA 的在独显市场份额约为 78%。

图22.全球 GPU 市场规模（十亿美元）



数据来源：Global Market Insights，财通证券研究所

图23.全球独立 GPU 市场占比（按厂商）



数据来源：STATISTA，JPR，财通证券研究所

大语言模型有望拉动 GPU 需求增量，我们测算 23/24/25 年大模型有望贡献 GPU 市场增量 69.88/166.2/209.95 亿美元。具体假设测算如下：

训练端，近年来各大厂商陆续发布大模型，我们假设 23/24/25 年新增 5/10/15 个大模型，根据 OpenAI 团队于 2020 发表的论文《Scaling Laws for Neural Language Models》提出的计算方法，对于以 Transformer 为基础的模型，假设模型参数量为 N，单 Token 所需的训练算力约为 6N。参考 OpenAI 团队 2020 同年发表的论文《Language Models are Few-Shot Learners》，GPT-3 模型参数量约为 1750 亿个，Token 数量约为 3000 亿个，近年发布的模型均在千亿级参数级别，因此我们中性假设 23 年新增大模型平均参数量约为 2000 亿个，Token 数量约为 3000 亿个，两者后续每年以 20% 增速增加。另外假设单次训练耗时约 30 天，算力效率为 30%，后续伴随算法精进，算力效率预计逐渐提升。以目前主流的训练端 GPU 英伟达 A100 测算，假设 ASP 为 1 万美元，23/24/25 年全球训练端 GPU 需求市场规模预计分别为 0.74/2.00/4.07 亿美元。

表3.训练端 GPU 需求增量测算

	2023E	2024E	2025E	2026E	2027E
全球新增大模型个数 (个)	5	10	15	20	25
大模型平均参数数量 (亿个, N)	2000	2400	2880	3456	4147
YoY		20.00%	20.00%	20.00%	20.00%
单个模型单 Token 训练所需运算次数 (TFLOPS-s, 6N)	1.20	1.44	1.73	2.07	2.49
训练 Tokens 数量 (亿个)	3000	3600	4320	5184	6221
YoY		20.00%	20.00%	20.00%	20.00%
单模型所需算力 (PFLOPs-Days)	4167	6000	8640	12442	17916
假设单次训练所需时间 (天)	30	30	30	30	30
算力效率	30.00%	32.00%	34.00%	36.00%	38.00%
训练端峰值算力需求 (PFLOPs, 单模型所需算力 × 模型数量 / (单次训练时间 × 算力效率))	2315	6250	12706	23040	39289
英伟达 A100 FP16 算力 (TFLOPs)	312	312	312	312	312
训练 GPU 需求量 (以 A100 FP16 算力计算) (万颗)	0.74	2.00	4.07	7.38	12.59
英伟达 A100 单价 (美元)	10000	10000	10000	10000	10000
训练 GPU 需求价值量 (亿美元)	0.74	2.00	4.07	7.38	12.59

数据来源:《Scaling Laws for Neural Language Models》OpenAI,《Language Models are Few-Shot Learners》OpenAI, 英伟达官网, CNBC, 财通证券研究所

推理端, 基于训练端的假设, 根据论文《Scaling Laws for Neural Language Models》, 单 Token 所需的推理算力开销约为 $2N$ 。则对于 GPT-3 模型, 其单 Token 所需的推理算力开销为 3500 亿 FLOPs-S。假设单次最大查询 Tokens 数为 1000(对应汉字约 300-500 字, 英文约 750 词), 每人每天查询 20 次。在并发用户数的估计上, 我们参考国际主流社交媒体日活用户数进行测算, 根据 Dustin Stout 统计, Facebook、WhatsApp、Instagram 全球日活用户数分别为 16 亿、10 亿、6 亿, 考虑到目前(类) GPT 平台仍处于发展早期, 我们预计全球大模型日活用户数在 23/24/25 分别为 2/6/10 亿, 按照所有用户平均分布于 24 小时, 并以 10 倍计算峰值并发数量。以目前英伟达用于推理端计算的 A10 测算, 假设 ASP 为 2800 美元, 23/24/25 年全球推理端 GPU 需求市场规模预计分别为 69.14/164.2/205.88 亿美元。

表4.推理端 GPU 需求增量测算

	2023E	2024E	2025E	2026E	2027E
全球大模型日活用户人数（亿人）	2	6	10	15	20
YoY		200.00%	66.67%	50.00%	33.33%
每人平均每天查询次数（次）	20	20	20	20	20
每人平均每次查询 Tokens 数量（个， 1000Tokens≈750 英文单词≈300-500 中文汉字）	1000	1000	1000	1000	1000
单 Tokens 所需计算次数（TFLOPs-s, 2N）	0.40	0.48	0.58	0.69	0.83
每人每次查询所需计算次数（TFLOPs-s, 2N× Tokens 数量）	400	480	576	691.2	829.44
全天计算次数合计（EFLOPs-s, 每人每次查询 所需计算次数×查询次数×日活人数）	1600000	5760000	11520000	20736000	33177600
算力效率	30.00%	32.00%	34.00%	36.00%	38.00%
平均每 s 所需峰值算力（EFLOPs）	61.73	208.33	392.16	666.67	1010.53
最大并发峰值算力乘数	10	10	10	10	10
最大并发峰值算力（EFLOPs）	617.28	2083.33	3921.57	6666.67	10105.26
峰值算力增量（EFLOPs）	617.28	1466.05	1838.24	2745.10	3438.60
英伟达 A10 FP16 算力（TFLOPs）	250	250	250	250	250
推理 GPU 需求量（以 A10 FP16 算力计算） （万颗）	246.91	586.42	735.29	1,098.04	1,375.44
英伟达 A10 单价（美元）	2800	2800	2800	2800	2800
推理 GPU 需求价值量（亿美元）	69.14	164.20	205.88	307.45	385.12

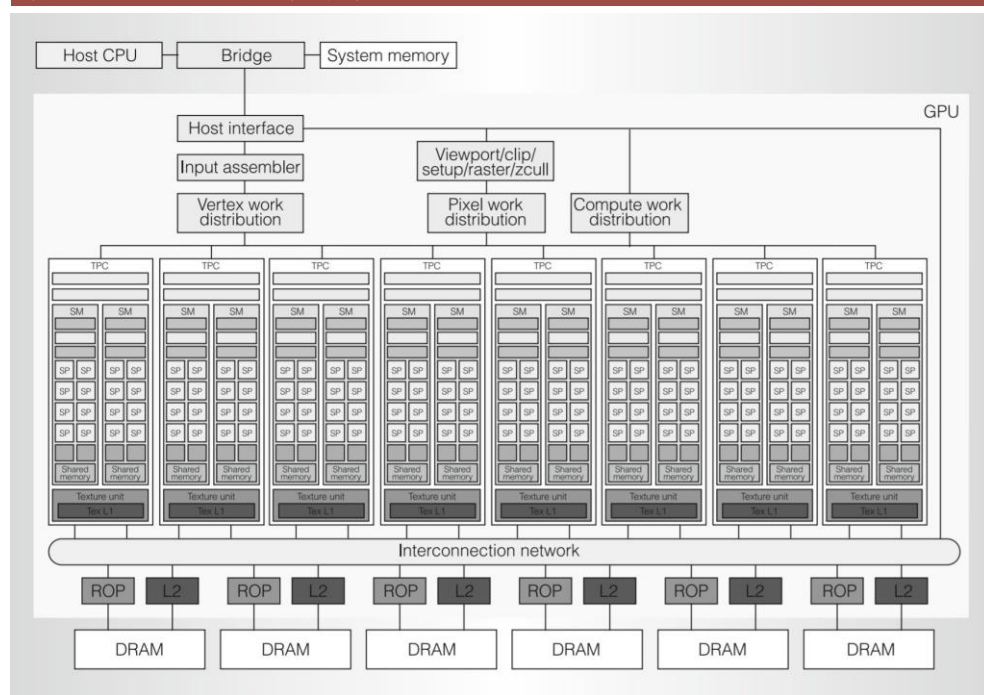
数据来源：《Scaling Laws for Neural Language Models》OpenAI，英伟达官网，Dustin Stout, Dihuni，财通证券研究所

2 微架构和平台生态共筑竞争壁垒

2.1 微架构：统一计算单元解锁通用计算时代

GPU 的微架构是用以实现指令执行的硬件电路结构设计。以 Nvidia 第一个实现统一着色器模型的 Tesla 微架构为例，从顶层 Host Interface 接受来自 CPU 的数据，藉由 Vertex（顶点）、Pixel（片元）、Compute（计算着色器）分发给各 TPC（Texture Processing Clusters 纹理处理集群）进行处理。

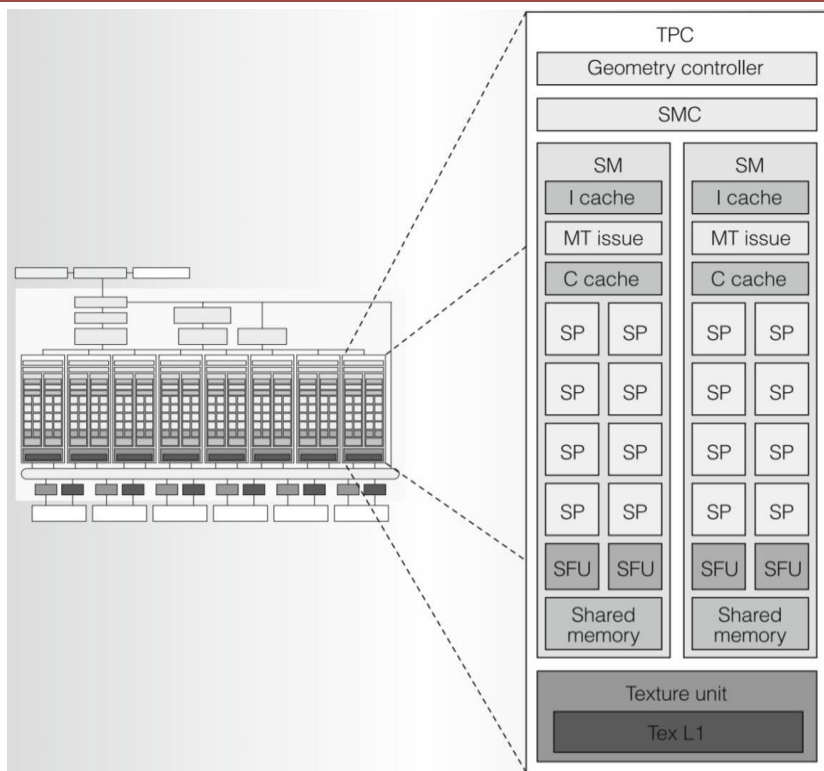
图24.Nvidia Tesla 整体架构图



数据来源：NVIDIA Tesla: A unified graphics and computing architecture，财通证券研究所

流处理器、特殊函数计算单元构成计算核心。在单个 TPC 中主要的运算结构为 SM（Streaming Multiprocessor 流式多处理器），其内在蕴含 I Cache（指令缓存）、C Cache（常量缓存）以及核心的计算单元 SP（Streaming Processor 流处理器）和 SFU（Special Function Unit 特殊函数计算单元），外加 Texture Unit（纹理单元）。

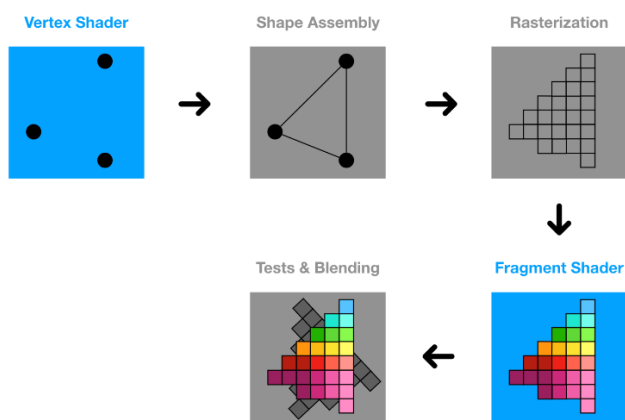
图25.Nvidia Tesla 微架构中 TPC 架构图



数据来源:《NVIDIA Tesla: A unified graphics and computing architecture》 Erik Lindholm、John Nickolls 等, 财通证券研究所

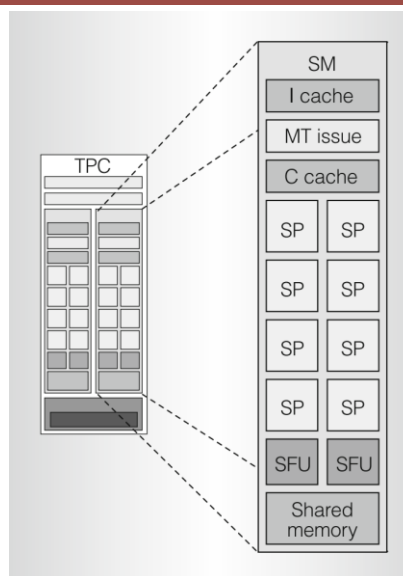
解耦计算单元, 拥抱通用计算。由于图形渲染流管线相对固定, Nvidia 在 Tesla 架构中将部分重要环节剥离并实现可编程, 解耦出 SM 计算单元用于通用计算, 即可实现根据具体任务需要分配相应线程实现通用计算处理。

图26.图像渲染管线相对固定



数据来源: ewind, 财通证券研究所

图27.Nvidia Tesla 微架构中 SM 架构图



数据来源:《NVIDIA Tesla: A unified graphics and computing architecture》 Erik Lindholm、John Nickolls 等, 财通证券研究所

谨请参阅尾页重要声明及财通证券股票和行业评级标准

20

计算核心、纹理单元增加， GPC 功能更加完整， Nvidia Fermi 架构奠定完整 GPU 计算架构基础。在 Tesla 之后， Nvidia 第一个完整的 GPU 计算架构 Fermi 通过制程微缩增加更多计算核心、纹理单元，并且通过增加 PolyMorph Engine（多形体引擎）和 Raster Engine（光栅引擎）使得原来 TPC 升级成为拥有更加完整功能的 GPC（Graphics Processing Clusters 图形处理器集群）。Fermi 架构共包含 4 个 GPC， 16 个 SM， 512 个 CUDA Core。

图28.Nvidia Fermi 架构图



数据来源：A GPU-based discrete element modeling code and its application in die filling，财通证券研究所

图29.Nvidia Fermi 微架构中 SM 架构图



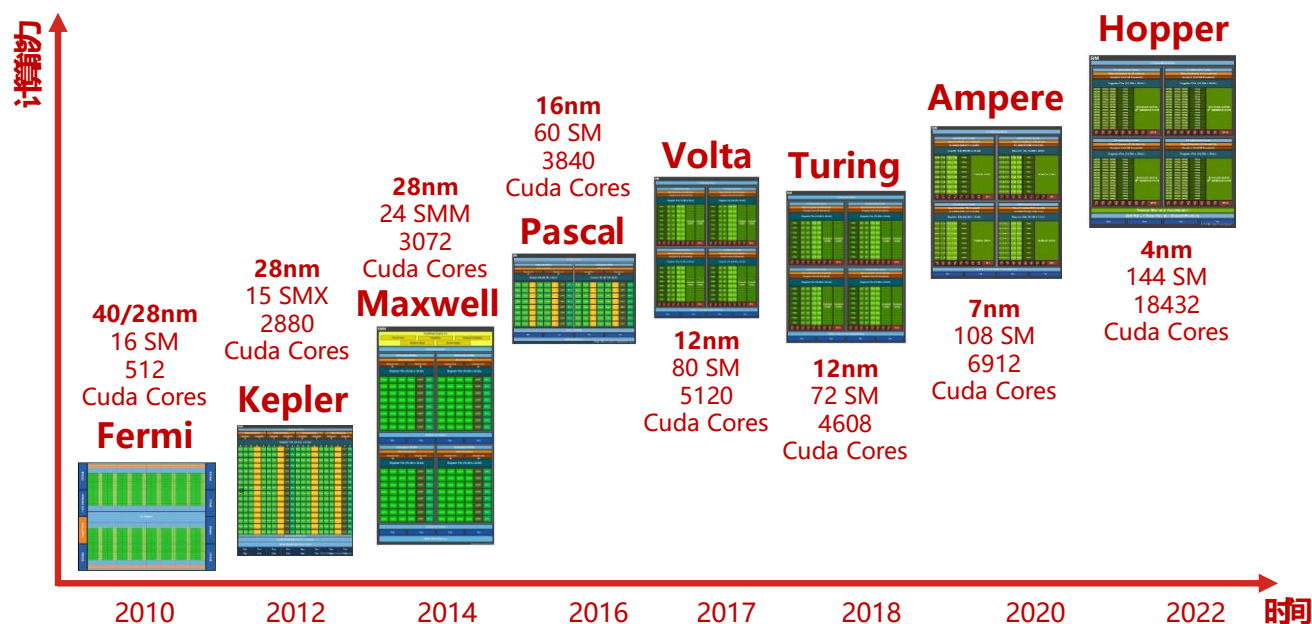
数据来源：NVIDIA，财通证券研究所

2.2 架构迭代与制程升级是 GPU 性能的生命线

不同的微架构设计会对 GPU 的性能产生决定性的影响，因此保持架构升级节奏以及制程升级速率是保证产品竞争力的关键。

英伟达 GPU 架构演进从最初 Fermi 架构到最新的 Ampere 架构和 Hopper 架构。每一阶段都在性能和能效比方面得到提升，引入了新技术，如 CUDA、GPU Boost、RT 核心和 Tensor 核心等，在图形渲染、科学计算和深度学习等领域发挥重要作用。最新一代 Hopper 架构在 2022 年 3 月推出，旨在加速 AI 模型训练，使用 Hopper Tensor Core 进行 FP8 和 FP16 的混合精度计算，以大幅加速 Transformer 模型的 AI 计算。与上一代相比，Hopper 还将 TF32、FP64、FP16 和 INT8 精度的每秒浮点运算(FLOPS)提高了 3 倍。

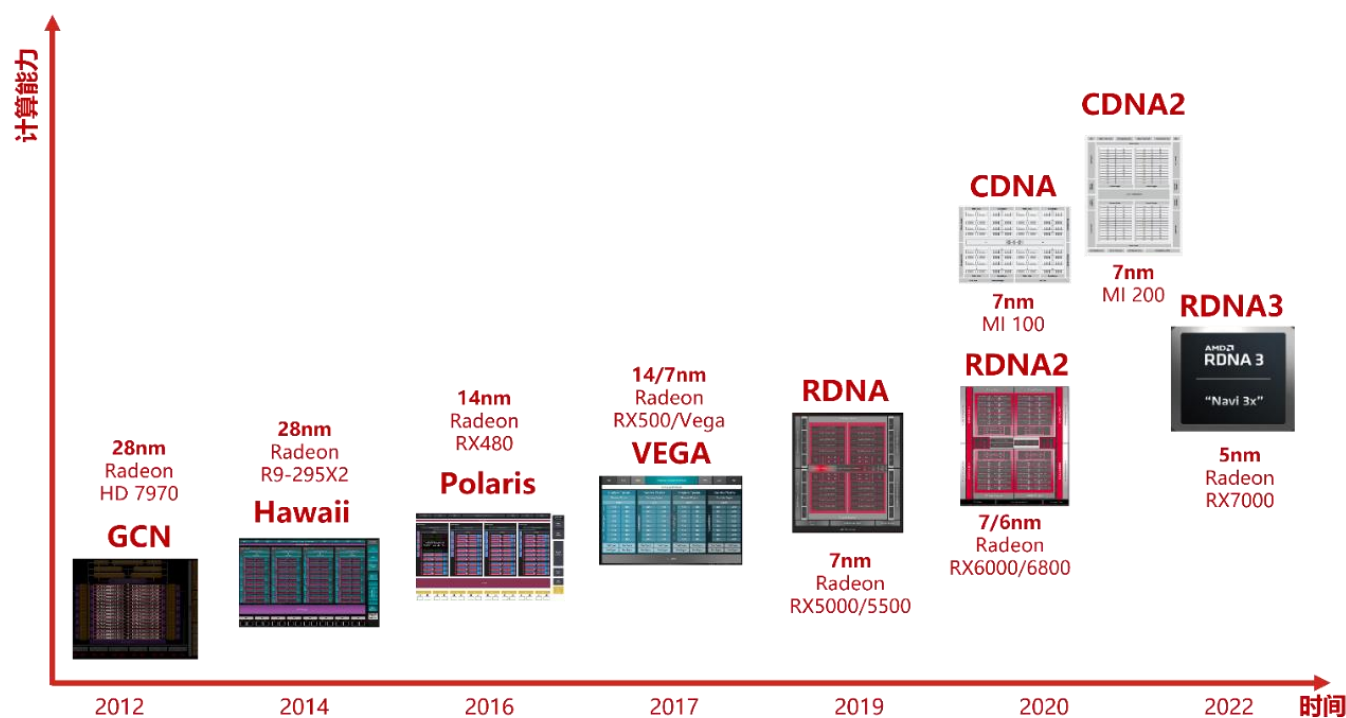
图30.NVIDIA GPU 架构演进历史



数据来源：英伟达官网，财通证券研究所

AMD 作为全球第二大 GPU 厂商，亦通过持续的架构演进保持其市场领先地位。从 2010 年以来，AMD 相继推出：GCN 架构、RDNA 架构、RDNA 2 架构、RDNA 3 架构、CDNA 架构和 CDNA 2 架构。最新一代面向高性能计算和人工智能 CDNA 2 架构于架构采用增强型 Matrix Core 技术，支持更广泛的数据类型和应用，针对高性能计算工作负载带来全速率双精度和全新 FP64 矩阵运算。基于 CDNA2 架构的 AMD Instinct MI250X GPU FP64 双精度运算算力最高可达 95.7 TFLOPs。

图31.AMD GPU 架构演进历史



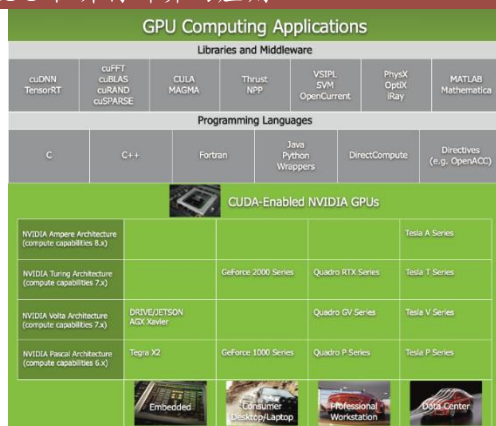
数据来源：AMD 官网，财通证券研究所

2.3 成熟的平台生态是 GPU 厂商的护城河

成熟且完善的平台生态是 GPU 厂商的护城河。相较于持续迭代的微架构带来的技术壁垒硬实力，成熟的软件生态形成的强大用户粘性将在长时间内塑造 GPU 厂商的软实力。以英伟达 CUDA 为例的软硬件设计架构提供了硬件的直接访问接口，不必依赖图形 API 映射，降低 GPGPU 开发者编译难度，以此实现高粘性的开发者生态。目前主流的开发平台还包括 AMD ROCm 以及 OpenCL。

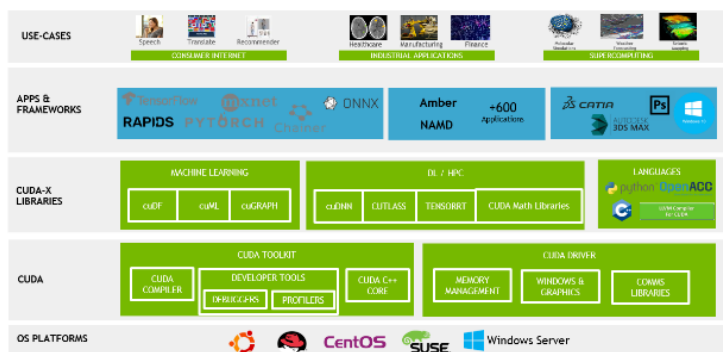
CUDA (Compute Unified Device Architecture)，是 NVIDIA 于 2006 年推出的通用并行计算架构，包含 CUDA 指令集架构 (ISA) 和 GPU 内部的并行计算引擎。该架构允许开发者使用高级编程语言（例如 C 语言）利用 GPU 硬件的并行计算能力并对计算任务进行分配和管理，CUDA 提供了一种比 CPU 更有效的解决大规模数据计算问题的方案，在深度学习训练和推理领域被广泛使用。

图32.GPU 在并行计算的应用



数据来源: NVIDIA, 财通证券研究所

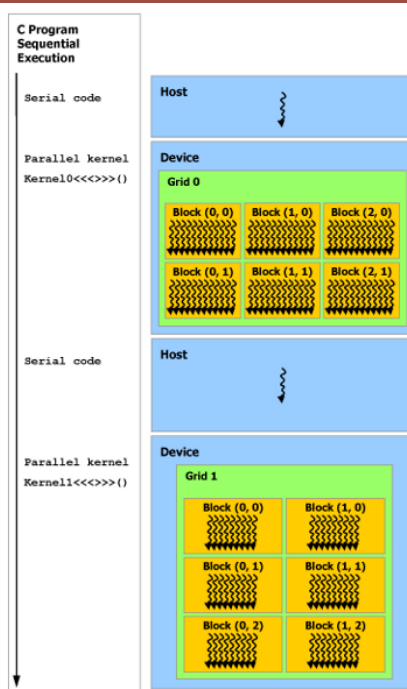
图33.CUDA 加速计算解决方案



数据来源: NVIDIA, 财通证券研究所

CUDA 除了是并行计算架构外, 还是 CPU 和 GPU 协调工作的通用语言。在 CUDA 编程模型中, 主要有 Host (主机) 和 Device (设备) 两个概念, Host 包含 CPU 和主机内存, Device 包含 GPU 和显存, 两者之间通过 PCI Express 总线进行数据传输。在具体的 CUDA 实现中, 程序通常划分为两部分, 在主机上运行的 Host 代码和在设备上运行的 Device 代码。Host 代码负责程序整体的流程控制和数据交换, 而 Device 代码则负责执行具体的计算任务。一个完整的 CUDA 程序是由一系列的端函数并行部分和主机端的串行处理部分共同组成的, 主机和设备通过这种方式可以高效地协同工作, 实现 GPU 的加速计算。

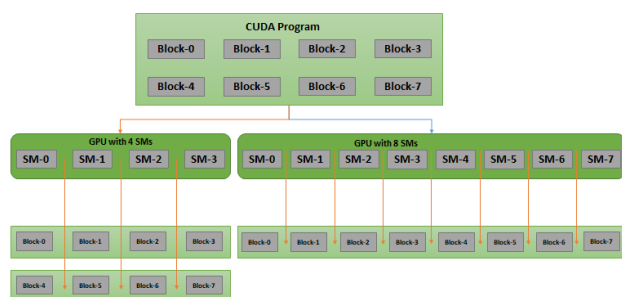
图34.CUDA 支持 CPU+GPU 的异构计算



数据来源: Nvidia 官网, 财通证券研究所

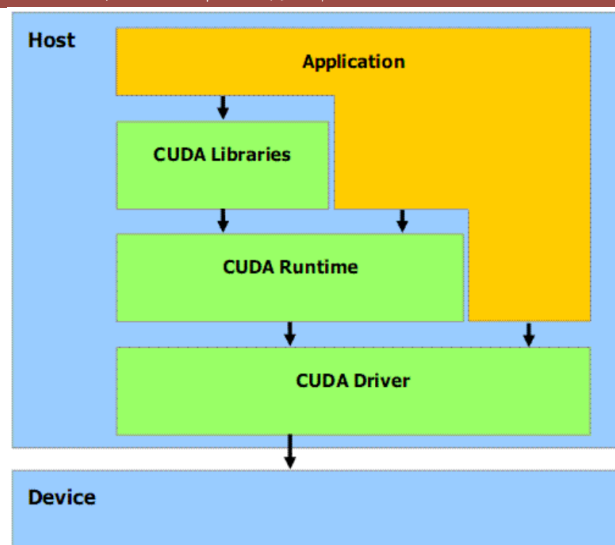
CUDA 在 Host 运行的函数库包括了开发库（Libraries）、运行时（Runtime）和驱动（Driver）三大部分。其中，Libraries 提供了一些常见的数学和科学计算任务运算库，Runtime API 提供了便捷的应用开发接口和运行期组件，开发者可以通过调用 API 自动管理 GPU 资源，而 Driver API 提供了一系列 C 函数库，能更底层、更高效地控制 GPU 资源，但相应的开发者需要手动管理模块编译等复杂任务。

图35.CUDA 编程模式示意图



数据来源：NVIDIA，财通证券研究所

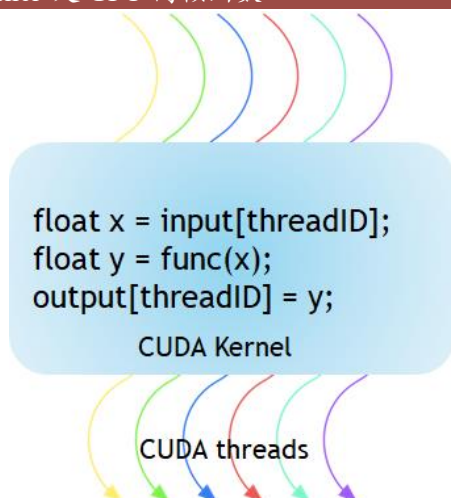
图36.CUDA 在 Host 中的函数库



数据来源：《Accelerating the new SCIARA-fv3 numerical model by different GPGPU strategies》Davide Spataro，财通证券研究所

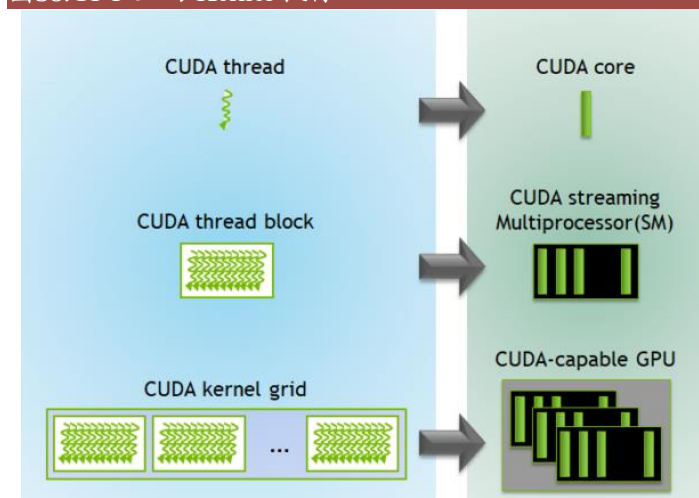
CUDA 在 Device 上执行的函数为内核函数（Kernel）通常用于并行计算和数据处理。在 Kernel 中，并行部分由 K 个不同的 CUDA 线程并行执行 K 次，而有别于普通的 C/C++ 函数只有 1 次。每一个 CUDA 内核都以一个声明指定器开始，程序员通过使用内置变量 `__global__` 为每个线程提供一个唯一的全局 ID。一组线程被称为 CUDA 块（block）。CUDA 块被分组为一个网格（grid），一个内核以线程块的网格形式执行。每个 CUDA 块由一个流式多处理器（SM）执行，不能迁移到 GPU 中的其他 SM，一个 SM 可以运行多个并发的 CUDA 块，取决于 CUDA 块所需的资源，每个内核在一个设备上执行，CUDA 支持在一个设备上同时运行多个内核。

图37. Kernel 是 GPU 内核函数



数据来源：NVIDIA，财通证券研究所

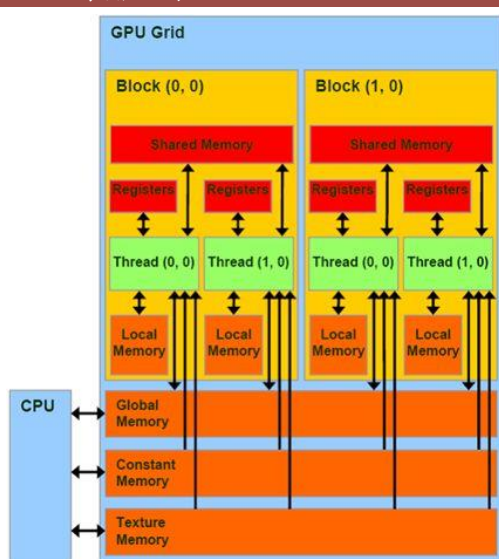
图38. GPU 上的 Kernel 执行



数据来源：NVIDIA，财通证券研究所

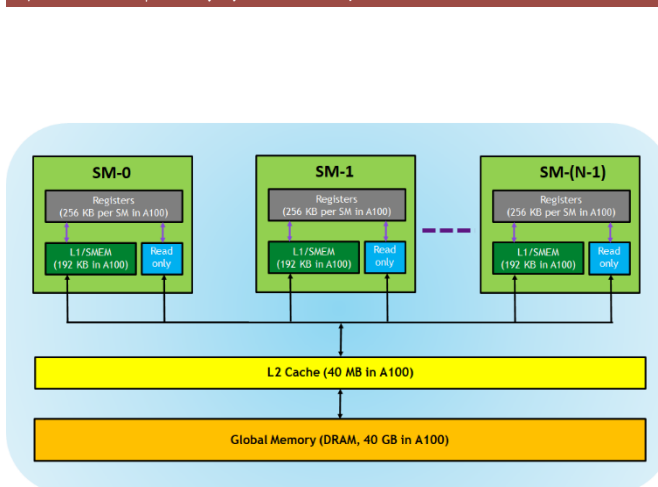
CUDA 的存储结构中，全局内存是所有线程都可以访问的存储区域，共享内存是位于线程块内部，多个线程可以共同访问的存储空间，寄存器是每个线程都有一组用于保存局部变量和中间值的寄存器，而局部内存则是当存储需求超过寄存器和共享内存容量时，分配给当前线程的存储空间。这些存储层次结构的访问速度和容量各不相同，需要在应用时进行合理使用和管理。GPU 的内存层次结构与 CUDA 的存储结构密切相关，比如，在一个 SM 上运行的多个线程块将共享该 SM 的寄存器和共享内存资源，同时也访问全局内存和局部内存资源。这些不同层级的存储在 GPU 中形成了逐层递进的内存架构，使得数据在计算过程中能够以最快的速度流动到被需要的位置，从而实现更高效、更快速的计算任务执行。

图39. CUDA 存储结构



数据来源：腾讯云，财通证券研究所

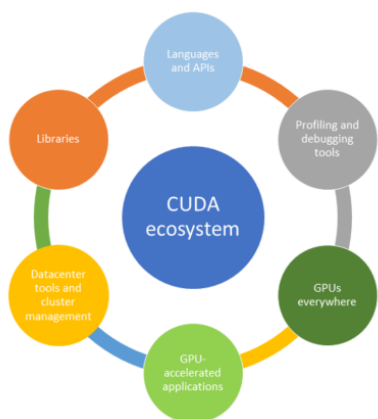
图40. GPU 中的内存层次结构



数据来源：英伟达官网，财通证券研究所

丰富而成熟的软件生态是 CUDA 被广泛使用的关键原因。(1) 编程语言：CUDA 从最初的 1.0 版本仅支持 C 语言编程，到现在的 CUDA 12.0 支持 C、C++、Fortran、Python 等多种编程语言。此外，NVIDIA 还支持了如 PyCUDA、Itimesh Hybridizer、OpenACC 等众多第三方工具链，不断提升开发者的使用体验。(2) 库：NVIDIA 在 CUDA 平台上提供了名为 CUDA-X 的集合层，开发人员可以通过 CUDA-X 快速部署如 cuBLA、NPP、NCCL、cuDNN、TensorRT、OpenCV 等多领域常用库。(3) 其他：NVIDIA 还为 CUDA 开发人员提供了容器部署流程简化以及集群环境扩展应用程序的工具，让应用程序更易加速，使得 CUDA 技术能够适用于更广泛的领域。

图41.CUDA 是 GPU 计算生态系统



数据来源：NVIDIA，财通证券研究所

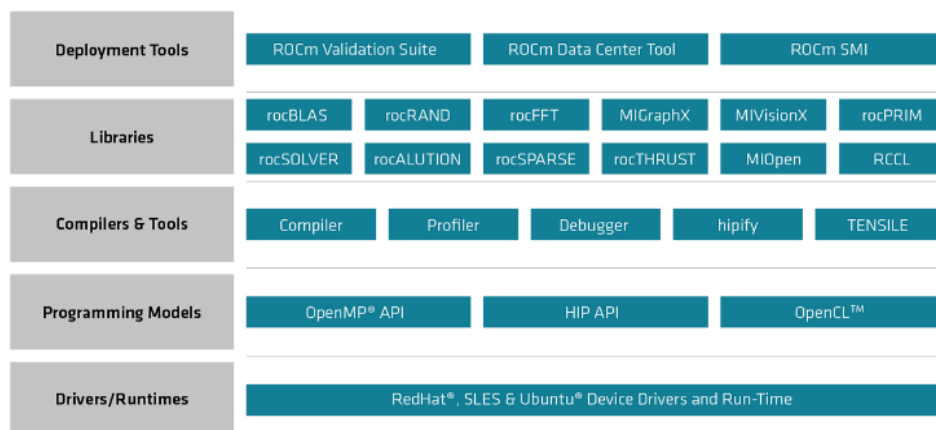
图42.CUDA 提供强大的开发支持工具



数据来源：NVIDIA，财通证券研究所

ROCm（Radeon Open Compute Platform）是 AMD 基于开源项目的 GPU 计算生态系统，类似于 NVIDIA 的 CUDA。ROCm 支持多种编程语言、编译器、库和工具，以加速科学计算、人工智能和机器学习等领域的应用。ROCm 还支持多种加速器厂商和架构，提供了开放的可移植性和互操作性。ROCm 支持 HIP（类 CUDA）和 OpenCL 两种 GPU 编程模型，可实现 CUDA 到 ROCm 的迁移。最新的 ROCm 5.0 支持 AMD Infinity Hub 上的人工智能框架容器，包括 TensorFlow 1.x、PyTorch 1.8、MXNet 等，同时改进了 ROCm 库和工具的性能和稳定性，包括 MIOpen、MIVisionX、rocBLAS、rocFFT、rocRAND 等。

图43.AMD ROCm 5.0



数据来源：AMD 官网，财通证券研究所

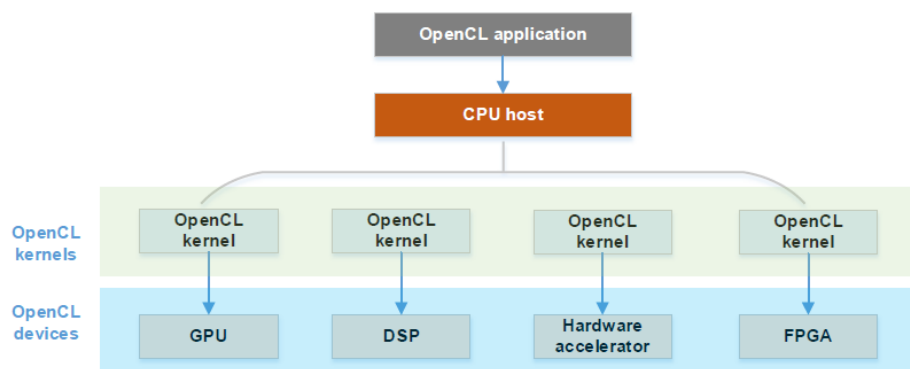
表5.ROCm 与 CUDA 模块对比

CUDA	ROCm	备注
CUDA API	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	
nvidia-smi	rocm-smi	
DirectGPU RDMA	ROCn RDMA	peer2peer
TensorRT	Tensile	张量计算库
CUDA-Docker	ROCm-Docker	

数据来源：CSDN，财通证券研究所

OpenCL (Open Compute Language)，是面向异构系统通用并行编程、可以在多个平台和设备上运行的开放标准。OpenCL 支持多种编程语言和环境，并提供了丰富的工具来帮助开发和调试，可以同时利用 CPU、GPU、DSP 等不同类型的加速器来执行任务，并支持数据传输和同步。此外，OpenCL 支持细粒度和粗粒度并行编程模型，可根据应用需求选择合适模型提高性能和效率。而 OpenCL 可移植性有限，不同平台和设备的功能支持和性能表现存在一定差异，与 CUDA 相比缺少广泛的社区支持和成熟的生态圈。

图44.异构计算框架 OpenCL



数据来源: khronos, 财通证券研究所

表6.OpenCL 与 CUDA 对比

	CUDA	OpenCL
供应商实施	仅由英伟达实施	由大量供应商实施, 包括 AMD, NVIDIA, Intel, Apple, Radeon 等。
可移植性	仅适用于 NVIDIA 硬件	可以移植到各种其他硬件, 只要避免特定于供应商的扩展
操作系统支持	必须使用 NVIDIA 硬件	支持各种操作系统
功能库	拥有广泛的高性能库	拥有大量库, 可用于所有 OpenCL 兼容硬件, 但不如 CUDA 广泛
技术细节	不是一种语言, 而是一种使用 CUDA 关键字实现并行化的平台和编程模型	不支持用 C++编写代码, 但在类似于环境的 C 编程语言中工作

数据来源: incredibuild, 财通证券研究所

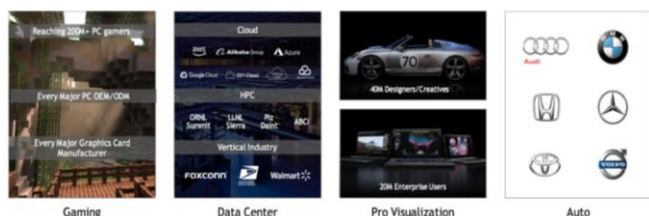
3 国内外发展现状：海外龙头领跑，国产持续发力

3.1 海外龙头：深耕多年，技术引领行业

3.1.1 英伟达

英伟达（NVIDIA）加速计算的先驱者，创立于1993年，公司于1999年发明的GPU推动了PC游戏市场的增长，重新定义了现代计算机显卡，并对并行计算进行了革新。目前，英伟达的产品应用领域包括数据中心和云计算、游戏和创作、高性能计算、自动驾驶汽车、计算机开发和边缘计算等，已逐渐转型为计算机平台公司。长久以来，英伟达是加速计算的先驱者。

图45.英伟达四大业务



数据来源：FourWeekMBA，财通证券研究所

图46.英伟达下游应用行业



数据来源：英伟达官网，财通证券研究所

英伟达 GeForce RTX™ 40 系列 GPU 为游戏玩家和创作者提供了高性能游戏体验。这一系列 GPU 由更高效的 NVIDIA Ada Lovelace 架构提供动力支持，可在性能和 AI 驱动图形领域实现质的飞跃。得益于光线追踪和更高的 FPS 游戏分辨率，玩家和创作者能够以更低的延迟体验栩栩如生的虚拟世界，探索革新的创作方式和远胜以往的工作流程加速技术。

表7.英伟达 40 系列显卡产品参数规格

型号	GeForce RTX 4090	GeForce RTX 4080	GeForce RTX 4070 Ti
GPU 引擎规格：			
NVIDIA CUDA®核心数量	16384	9728	7680
加速频率 (GHz)	2.52	2.51	2.61
基础频率 (GHz)	2.23	2.21	2.31
显存规格：			
标准显存配置	24 GB GDDR6X	16 GB GDDR6X	12 GB GDDR6X
显存位宽	384 位	256 位	192 位

数据来源：英伟达官网，财通证券研究所

NVIDIA A100 Tensor Core GPU 可针对 AI、数据分析和 HPC 应用场景，在不同规模下实现出色的加速，有效助力更高性能的弹性数据中心。A100 采用

NVIDIA Ampere 架构，是 NVIDIA 数据中心平台的引擎，其性能比上一代产品提升高达 20 倍，并可划分为七个 GPU 实例，以根据变化的需求进行动态调整。A100 提供超快速的显存带宽（每秒超过 2 万亿字节 [TB/s]），可处理超大型模型和数据集。

NVIDIA H100 Tensor Core GPU 作为 A100 的迭代产品，可进一步在每个工作负载中实现出色性能、可扩展性和安全性。H100 使用 NVIDIA® NVLink® Switch 系统，可连接多达 256 个 H100 来加速百亿亿级 (Exascale) 工作负载，另外可通过专用的 Transformer 引擎来处理万亿参数语言模型。与 A100 相比，H100 的综合技术创新可以将大型语言模型的速度提高 30 倍，从而提供业界领先的对话式 AI。

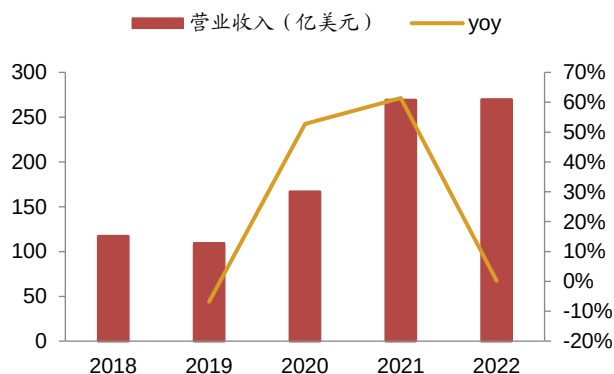
表8.英伟达 A100、H100 系列产品规格参数

外形规格	A100 80GB PCIe	H100 PCIe	H100 NVL2
FP64	9.7 TFLOPS	26 TFLOPS	68 TFLOPs
FP64 Tensor Core	19.5 TFLOPS	51 TFLOPS	134 TFLOPs
FP32	19.5 TFLOPS	51 TFLOPS	134 TFLOPs
Tensor Float 32 (TF32)	156 TFLOPS 312 TFLOPS*	756 TFLOPS	1,979 TFLOPs
BFLOAT16 Tensor Core	312 TFLOPS 624 TFLOPS*	1,513 TFLOPS	3,958 TFLOPs
FP16 Tensor Core	312 TFLOPS 624 TFLOPS*	1,513 TFLOPS	3,958 TFLOPs
INT8 Tensor Core	624 TOPS 1248 TOPS*	3,026 TOPS	7,916 TOPS
GPU 显存	80GB HBM2	80GB	188GB
GPU 显存带宽	1935 GB/s	2TB/s	7.8TB/s
最大热设计功耗 (TDP)	300W	300-350 瓦（可配置）	2x 350-400W（可配置）
多实例 GPU	最大为 7MIG @ 每个 5GB	最多 7 个 MIG @每个 10GB	最多 14 个 MIG @每个 12GB

数据来源：英伟达官网，财通证券研究所

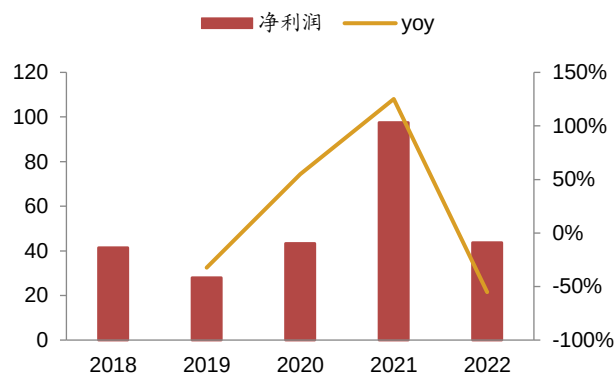
受行业周期下行影响，公司净利润大幅下降。公司 2022 年实现营业总收入 269.74 亿美元，与去年基本持平，净利润为 43.68 亿美元，同比大幅下降 55.21%，主要缘于游戏显卡需求疲软，资产减值损失较大。

图47.英伟达营业收入及增速(亿美元)



数据来源：，财通证券研究所

图48.英伟达净利润及增速(亿美元)



数据来源：，财通证券研究所

谨请参阅尾页重要声明及财通证券股票和行业评级标准

31

3.1.2 AMD

AMD（超微半导体公司），自 1969 年创立以来，专注于处理器及相关技术设计研发。AMD 2009 年将自有晶圆厂拆分为现今的格芯后，从 IDM 厂商转型为 Fabless 公司，目前 AMD 主要产品为 CPU（包括嵌入式平台）、GPU、主板芯片组以及 2022 年收购赛灵思而扩充的 FPGA 业务。AMD 是目前除了英特尔以外，最大的 x86 架构处理器供应商，自 2006 年收购 ATI 后，成为同时拥有 CPU 和 GPU 技术的半导体公司。

图49.AMD 业务概览



数据来源：AMD 企业介绍材料，财通证券研究所

图50.AMD 核心技术概览



数据来源：AMD 年报展示材料，财通证券研究所

AMD 最新于 2022 年推出 AMD Radeon RX 7000 系列显卡，采用 AMD 最新 RDNA 3 计算单元，具有光线追踪和人工智能加速功能。7900 系列创新性地采用了小芯片技术的游戏 GPU，其 AMD Radiance Display 引擎和 DisplayPort™ 2.1 的强强联合可以带来 12 位 HDR 和 REC2020 色彩空间的完全覆盖，最高可达 8K 165Hz。

表9.AMD7000 显卡参数规格

型号	AMD Radeon™ RX 7900 XTX	AMD Radeon™ RX 7900 XT
计算单元	96	84
光线加速器	96	84
游戏频率	2300 MHz	2000 MHz
INFINITY CACHE	96 MB	80 MB
最大显存	24 GB	20 GB

数据来源：AMD 官网，财通证券研究所

AMD 于 2016 年推出 Instinct 计算加速器，旨在加速深度学习、人工神经网络和高性能计算 GPGPU 的应用。AMD Instinct 系列加速器采用创新性的 AMD CDNA 架构、AMD Infinity Fabric 技术以及先进的封装技术。对于高性能计算工

作负载，AMD Instinct MI250X 的 GPU 双精度 (FP64) 结合全新 FP64 Matrix Core 技术更可实现最高达 95.7 TFLOPs 峰值理论性能。

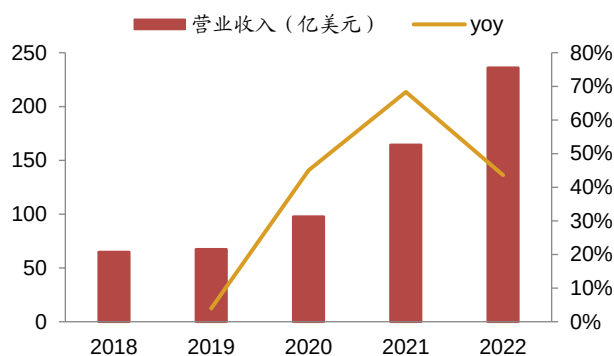
表10.AMD Instinct 系列产品规格

型号	AMD Instinct™ MI250X	AMD Instinct™ MI250	AMD Instinct™ MI210
计算单元	220	208	104
流处理器	14,080	13,312	6656
峰值半精度 (FP16)性能	383 TFLOPS	362.1 TFLOPS	181 TFLOPS
峰值单精度 (FP32)性能	47.9 TFLOPS	45.3 TFLOPS	22.6 TFLOPS
峰值双精度 (FP64)性能	47.9 TFLOPS	45.3 TFLOPS	22.6 TFLOPS
专用显存大小	128 GB	128 GB	64 GB
峰值显存带宽	3276.8 GB/s	3276.8 GB/s	1638.4 GB/s

数据来源：AMD 官网，财通证券研究所

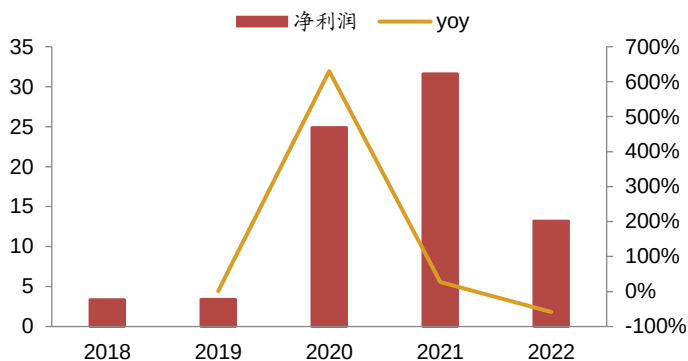
数据中心业务快速增长，推动公司整体营收提升。公司 2022 年实现营业总收入 236.01 亿美元，同比上升 43.61%，净利润为 13.2 亿美元，同比大幅下降 58.25%，主要缘于收购赛灵思后，无形资产摊销数额较大致使净利润下滑。

图51.AMD 营业收入及增速(亿美元)



数据来源：，财通证券研究所

图52.AMD 净利润及增速(亿美元)

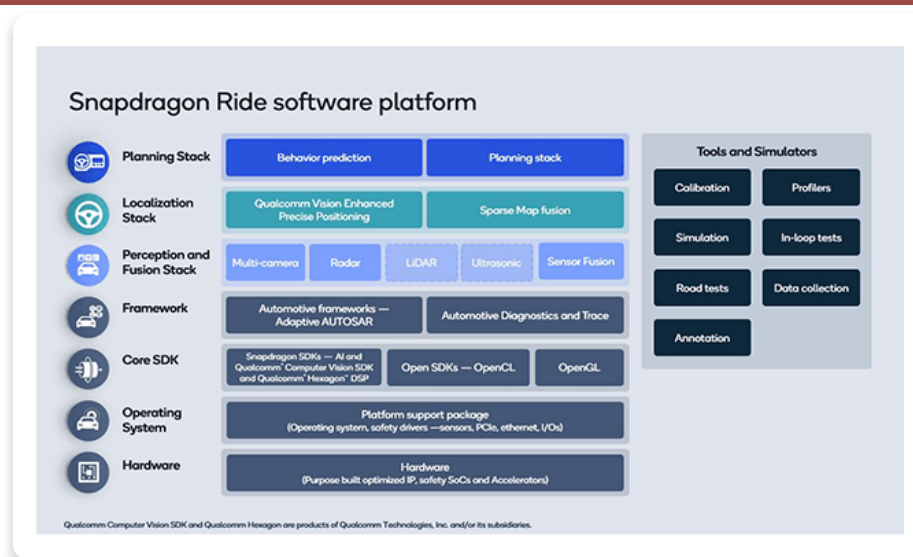


数据来源：，财通证券研究所

3.1.3 高通

高通（Qualcomm）创立于1985年，是全球领先的无线科技创新公司。高通变革了世界连接、计算和沟通的方式，高通的基础科技赋能整体移动生态系统，开启了移动互联时代。2009年，高通收购了AMD的移动GPU Imageon系列，开始发展移动端自研GPU业务。

图53.高通骁龙移动平台



数据来源：高通官网，财通证券研究所

高通 Adreno GPU（原 Imageon）为采用骁龙处理器的移动终端提供游戏机品质的 3D 图形处理能力，为游戏、用户界面和高性能计算任务提供更快的图形处理。作为骁龙异构计算的关键组件，Adreno GPU 为无缝配合骁龙 CPU 和 DSP 而设计，可以帮助支持处理密集型 GPGPU 计算任务。2022 年底，高通已发布全新 4nm 级 GPU Adreno 740。

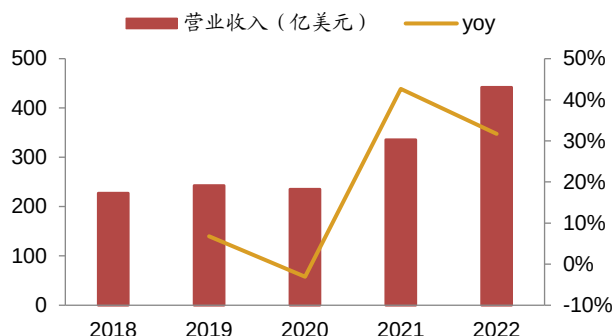
表11.高通 Adreno 7 系列产品规格

型号	微架构		制程 (nm)	时脉 (MHz)	API				
	架构类型	ALU			Vulkan	OpenGL ES	OpenCL	OpenGL	Direct3D
Adreno 730	统一着色器模型、	1024	4	818/900	1.0 and 1.1	3.2	3.0 Full	WIP (freedreno driver)	12.1
Adreno 740	统一内存	1536	4	680/719					

数据来源：Notebookcheck，财通证券研究所

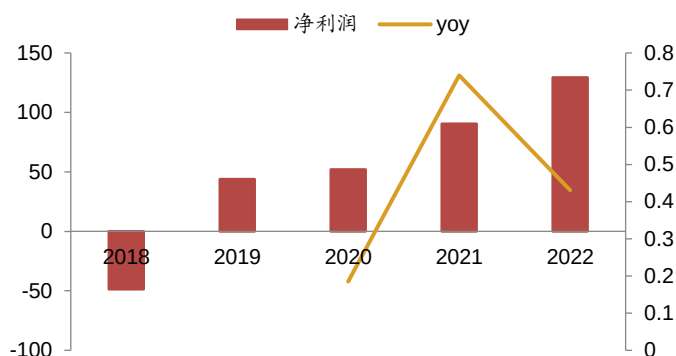
营业收入和盈利能力均稳定增长。公司 2022 年度实现营业收入 442 亿美元，同比上升 31.68%，净利润为 129.36 亿美元，同比上升 43.05%。

图54.高通营业收入及增速(亿美元)



数据来源: , 财通证券研究所

图55.高通净利润及增速(亿美元)



数据来源: , 财通证券研究所

3.1.4 Imagination

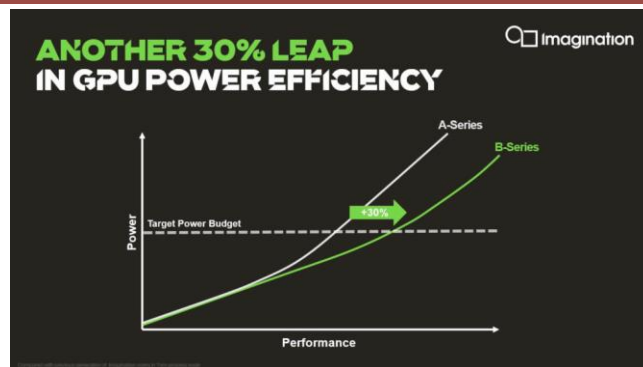
Imagination 成立于 1985 年, 移动端 GPU 设计领域的领军企业。Imagination 业务主要包括设计 PowerVR 移动图形处理器, 网络路由器 (基于 MIPS CPU) 和其他纯消费电子部门。此外还提供无线电基带处理、网络、数字信号处理器、视频和音频硬件、IP 语音软件、云计算以及芯片和系统设计服务。

图56.IMG B 系列产品



数据来源: Imagination 官网, 财通证券研究所

图57.IMG B 系列与 A 系列性能对比



数据来源: Imagination 官网, 财通证券研究所

2020 年 10 月, Imagination 发布 IMG B 系列高性能 GPU IP。此款多核架构 GPU IP 包括 BXE、BXM、BXT、BXS 4 个系列, 分别代表入门级、中端、高端以及汽车安全。其中 BXT 主要应用于移动设备、数据中心, 浮点算力 6TFlops, 每秒可处理 1920 亿像素, AI 算力达 24Tops。

表12.Imagination IMG B 系列产品简介

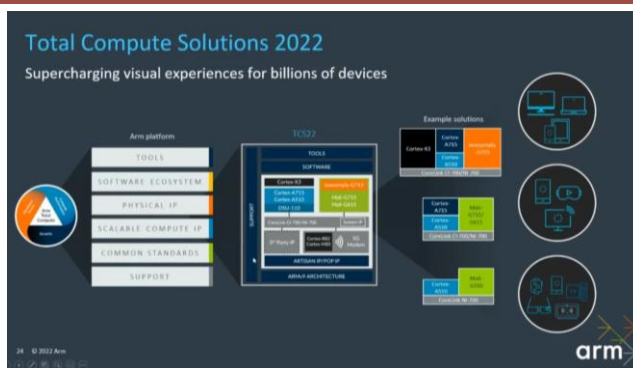
型号	描述
IMG BXE	主要应用于高清显示，每时钟周期可处理 1-16 个像素，支持 720p-8K 分辨率显示。
IMG BXM	主要应用于图形处理，在紧凑的面积、高效的内核上实现了填充率、计算力的最佳平衡，适用于移动游戏、数字电视等复杂 UI 领域。
IMG BXT	主要应用于极致性能，可用于移动设备、数据中心，浮点性能 6TFlops，每秒可处理 1920 亿像素，AI 算力达 24Tops
IMG BXS	主要应用于汽车领域，可广泛用于 HMI 人机界面、UI 显示、信息娱乐系统、数字驾舱、环绕视图、自动驾驶等。

数据来源：Imagination 官网，财通证券研究所

3.1.5 ARM

ARM（安谋控股公司），成立于 1990 年，是全球龙头半导体 IP 供应商。公司主要产品有 CPU、GPU 和 NPU 等处理器 IP。目前，总共有超过 100 家公司与 ARM 公司签订了技术使用许可协议，其中包括 Intel、IBM、LG、NEC、SONY 等。

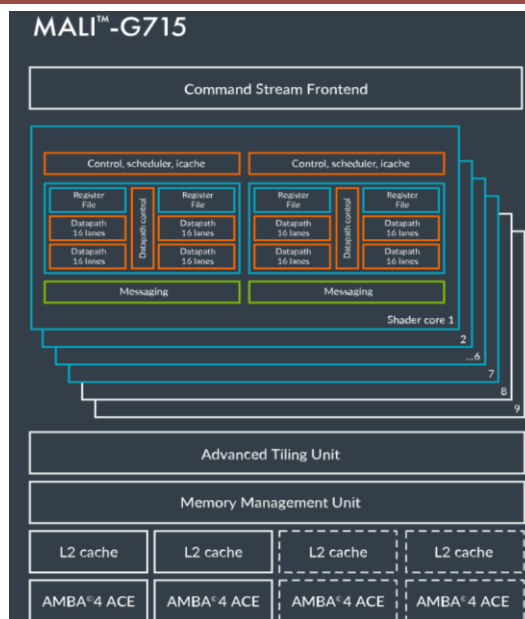
图58.ARM 整体设计解决方案



数据来源：ARM 官网，财通证券研究所

ARM 最新 GPU 产品 Mali-G7 系列中 Immortals-G715 GPU 采用 10 个及以上内核，支持硬件级光线追踪技术。Mali-G715 旨在通过一系列新的图形功能和升级（包括可变速率着色）来满足高端移动市场的需求，适用于移动设备上的复杂 AAA 游戏。

图59.ARM Immortals-G715 架构

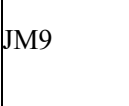


数据来源：ARM 官网，财通证券研究所

3.2 兼容主流生态对标行业龙头，国内厂商持续发力

国产 GPU 持续发力，对标行业龙头缩小差距。GPU 有两条主要的发展路线：分别为传统的 2D/3D 图形渲染 GPU 和专注高性能计算的 GP GPU，近年来，国产 GPU 厂商在图形渲染 GPU 和高性能计算 GPGPU 领域上均推出了较为成熟的产品，在性能上不断追赶行业主流产品，在特定领域达到业界一流水平。生态方面国产厂商大多兼容英伟达 CUDA，融入大生态进而实现客户端导入。

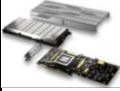
表13.图形渲染 GPU 产品性能对比

	英伟达	芯动科技	摩尔线程	景嘉微	格兰菲
代表产品	GeForce RTX 4090 	风华一号 	MTT S80 	JM9 	Arise-GT10C0 
单精度浮点算力	82.58 TFLOPS	5 TFLOPS	14.4 TFLOPS	512 GFLOPS	1.5TFLOPS
半精度浮点算力	82.58 TFLOPS	NA	NA	NA	NA
像素填充率	443.5 GPixel/s	160 GPixels/s	NA	8 GPixels/s	48 GPixels/s
纹理填充率	1,290 GTexel/s	NA	NA	NA	96 Texels/s
制程	4nm	12nm	NA	NA	28nm

显存容量	24GB	8G/16G	16GB	8GB	NA
显存类型	GDDR6X	GDDR6/GDDR6X	GDDR6	NA	DDR4
总线接口	PCIe 4.0 x16	PCIe4.0 x8 向下兼容	PCIe Gen5 x16	PCIE 4.0 X8	PCIe3.0 x8
生态	CUDA	NA	CUDA	NA	NA

数据来源：各公司官网，财通证券研究所

表14.通用计算 GPU 产品性能对比

	英伟达	英伟达	寒武纪	海光信息	壁仞科技	壁仞科技	摩尔线程	天数智芯
代表产品	H100 	A100 	思元 370 (AI 加速芯片) 	海光 8100 	壁仞 100P 	壁仞 104P 	MTT S3000 	天垓 100 
单精度浮点算力	51.22 TFLOPS	19.49	24 TFLOPS	NA	240 TFLOPS	NA	15.2 TFLOPS	18.5 TFLOPS
半精度浮点算力	204.9 TFLOPS	77.97	96 TFLOPS	NA	NA	NA	NA	147 TFLOPS
整型算力 (INT8)	NA	NA	256 TOPS	NA	1920 TOPS	1024 TOPS	NA	295 TOPS
制程	4 nm	7 nm	7nm	NA	7nm	7nm	NA	7nm
显存容量	80 GB	40 GB/80GB	24GB	NA	64GB	32GB	32GB	32 GB
显存类型	HBM2e	HBM2e	LPDDR5	NA	NA	NA	GDDR6	NA
总线接口	PCIe 5.0 x16	PCIe 4.0 x16	PCIe Gen4 x16	PCIe Gen4.0 x16	PCIe5.0 X16	PCIe5.0 X16	PCIe Gen5 x16	PCIe Gen4.0 x16
生态	CUDA	CUDA	NA	ROCm (兼容 CUDA)	CUDA	CUDA	CUDA	CUDA

数据来源：各公司官网，财通证券研究所

3.3 高端芯片进口遭限制，国产厂商替代迎契机

美国对中国高端芯片出口进行管制。据英伟达于 2022 年 8 月 31 日发布的公告，美国政府通知公司在未来将 A100 和即将推出的 H100 等人工智能芯片出口到中国大陆、中国香港和俄罗斯时须获得许可证。

2023 年 3 月 3 日，美国商务部以“国家安全”和“外交政策利益”为由，将浪潮集团等 28 个中国实体列入所谓的“实体清单”，限制其从美国进口产品和技术。未来在人工智能芯片，特别是 GPU 上对中国的制裁将对中国 AI 产业提出极大的挑战。挑战伴随着机遇，高端 GPU 的限售给予了国产厂商替代空间。

在国产替代的背景下，政策支持推动国产 GPU 行业高速发展。2020 年以来，国家及各省市陆续出台了若干政策，通过税收减免、财政补贴等方式支持半导体与集成电路产业发展。

表15.国内 GPU、半导体相关政策（部分）

时间	政策	相关内容
2020.07.27	新时期促进集成电路产业和软件产业高质量发展的若干政策	集成电路设计、装备、材料、封装、测试企业和软件企业，自获利年度起，第一年至第二年免征企业所得税，第三年至第五年按照 25% 的法定税率减半征收企业所得税。
2022.01.19	新时期促进上海市集成电路产业和软件产业高质量发展若干政策	对于符合条件的设计企业开展有利于促进本市集成电路线宽小于 28 纳米(含)工艺产线应用的流片服务，相关流片费计入项目新增投资，对流片费给予 30% 的支持，支持金额原则上不高于 1 亿元。
2022.10.11	深圳市关于促进半导体与集成电路产业高质量发展的若干措施	重点突破 CPU、GPU、DSP、FPGA 等高端通用芯片的设计，布局人工智能芯片、边缘计算芯片等专用芯片的开发。
2023.01.11	四川省人民政府工作报告	重点发展 CPU、GPU 等高端通用芯片及国产 EDA 工具，支持能源电子、中低轨卫星等新业态产业。

数据来源：中国政府网，上海人民政府网，深圳市发改委官网，四川省生态环境厅官网，财通证券研究所

4 建议关注

4.1 寒武纪

寒武纪自 2016 年成立以来一直专注于人工智能芯片产品研发与技术创新，致力于打造人工智能领域的核心处理器芯片。公司主要提供云端智能芯片及加速卡、训练整机、边缘智能芯片及加速卡、终端智能处理器 IP 及配套基础软件开发平台，产品广泛应用于消费电子、数据中心、云计算等诸多场景。

表16.寒武纪主要产品性能参数

产品类型	主要产品	推出时间	制程	算力 (INT8)	算力 (FP32)	产品实物图
云端智能芯片及加速卡	思元 100 (MLU100) 芯片及云端智能加速卡	2018 年	16nm	32TOPS	NA	
	思元 270 (MLU270) 芯片及云端智能加速卡	2019 年	16nm	128TOPS*	NA	
	思元 290 (MLU290) 芯片及云端智能加速卡	2020 年	7nm	512TOPS	NA	
	思元 370 (MLU370) 芯片及云端智能加速卡	2021 年	7nm	256TOPS	24TFLOPS	
训练整机	玄思 1000 智能加速器	2020 年	NA	支持该精度	支持该精度	
边缘智能芯片及加速卡	思元 220 (MLU220) 芯片及边缘智能加速卡	2019 年	16nm	8TOPS	NA	
终端智能处理器 IP	寒武纪 1A 处理器	2016 年	NA	NA	NA	
	寒武纪 1H 处理器	2017 年	NA	支持 0.5-1TOPS	NA	
	寒武纪 1M 处理器	2018 年	NA	支持 0.5-8TOPS	支持	
基础系统软件平台	寒武纪基础软件开发平台 (适用于公司所有芯片与处理器产品)	持续研发中	NA	NA	NA	
*为 TOPS=1024GOPS 换算，如 INT8 实际算力为 131,072GOPS 或 131TOPS。						

数据来源：寒武纪 2022 半年报，财通证券研究所

2022 年 3 月 21 日，公司正式发布新款训练加速卡 MLU370-X8，搭载双芯片四芯粒思元 370，集成寒武纪 MLU-Link™多芯互联技术，在业界广泛应用于 YOLOv3、Transformer 等训练任务中。

MLU 370-S4、MLU370-X4 和 MLU370-X 均基于思元 370 智能芯片的技术，通过 Chiplet 技术灵活组合产品的特性，可满足更多市场需求。凭借其优异竞争

力，公司已就思元 370 系列与部分头部互联网、银行、服务器厂商实现了深度合作和互利共赢。

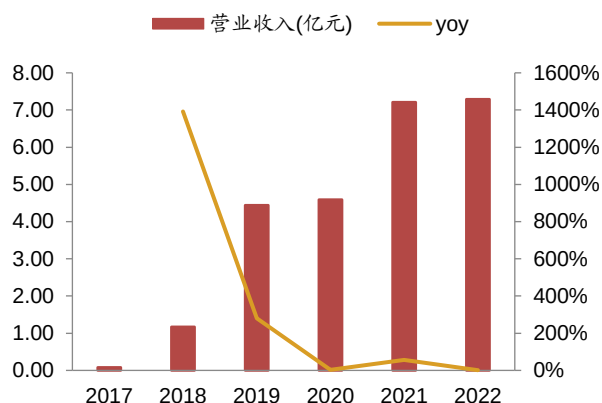
表17.思元 370 系列产品信息

	MLU370-S4	MLU370-X4	MLU370-X8
面向市场	机器视觉/推理任务	互联网行业等推理任务，或训推一体场景	训练任务
市场特点	整机计算密度较高，单卡算力需求适中	单卡算力需求较高	单卡算力需求较高，互联带宽需求高
算力	192TOPS(INT8) 96TOPS(INT16) 72TFLOPS(FP16) 72TFLOPS(BF16) 18TFLOPS(FP32)	256TOPS(INT8) 128TOPS(INT16) 96TFLOPS(FP16) 96TFLOPS(BF16) 24TFLOPS(FP32)	256TOPS(INT8) 128TOPS(INT16) 96TFLOPS(FP16) 96TFLOPS(BF16) 24TFLOPS(FP32)
互联带宽	307.2GB/s	307.2GB/s	614.4GB/s

数据来源：寒武纪 2022 半年报，财通证券研究所

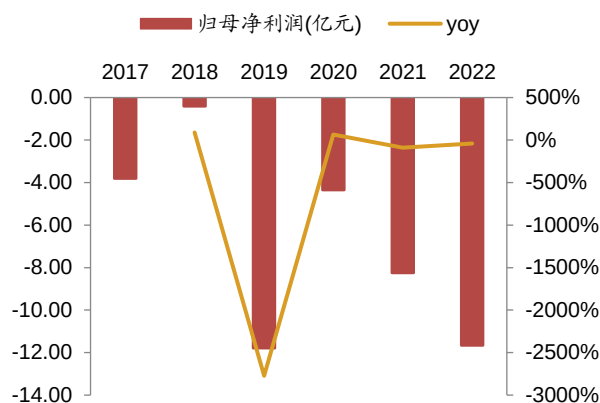
公司克服宏观经济、疫情反复等因素影响，在 2022 年实现度营业总收入为 7.2 亿元，比上年同期增长 1.11%。归属于母公司股东的净利润为-11.66 亿元，较上年同期亏损增加 41.4%，主要系研发费用、资产减值损失、信用减值损失增长所致。

图60.寒武纪营业收入及增速(亿元)



数据来源：choice，财通证券研究所

图61.寒武纪归母净利润及增速(亿元)



数据来源：choice，财通证券研究所

4.2 海光信息

海光信息主要从事高端处理器、加速器等计算芯片产品和系统的研发、设计和销售。公司的产品包括海光通用处理器（CPU）和海光协处理器（DCU），具有

成熟而丰富的应用生态环境，内置专用安全硬件，可满足互联网、金融、能源等行业的广泛应用需求。

图62.海光信息 DCU 基本组成架构



数据来源：海光信息招股书，财通证券研究所

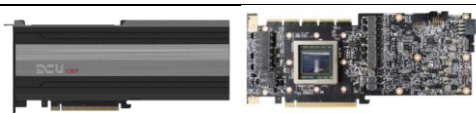
图63.海光深算 DCU 完善软件栈支持



数据来源：海光信息官网，财通证券研究所

公司 DCU 系列产品海光 8100 采用先进的 FinFET 工艺，以 GPGPU 架构为基础，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，可充分挖掘应用的并行性，发挥其大规模并行计算的能力，快速开发高能效的应用程序，在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。

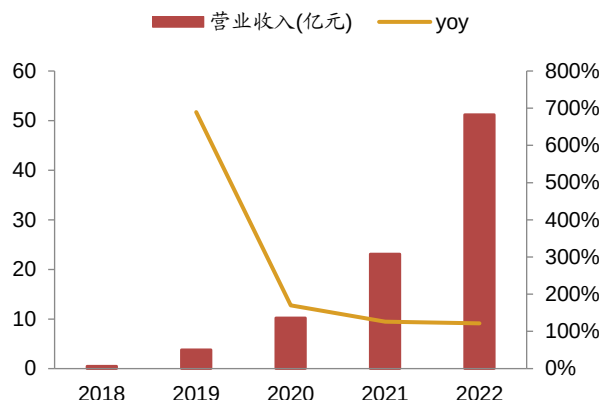
表18.海光 DCU 产品主要参数

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

数据来源：海光招股说明书，财通证券研究所

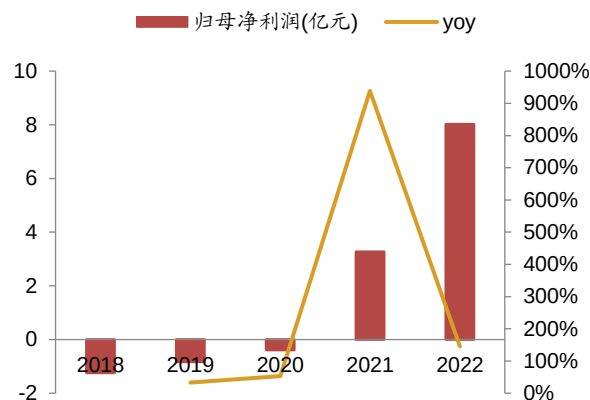
公司营业收入增势可观，2020-2022 年公司分别实现营收 10.22 亿元、23.1 亿元、51.2 亿元，同比增长保持在 120%以上。归母净利润于 2021 年扭亏为盈后持续增长，2022 年达到 8.02 亿元，同比上升 145.3%。

图64.海光信息营业收入及增速(亿元)



数据来源：，财通证券研究所

图65.海光信息归母净利润及增速(亿元)



数据来源：，财通证券研究所

4.3 景嘉微

景嘉微致力于信息探测、处理与传递领域的技术和综合应用。公司产品涵盖集成电路设计、小型雷达系统、无线通信系统、电磁频谱应用系统等方向，广泛应用于有高可靠性要求的航空、航天、航海、车载等专业领域。

公司先后自研制成功 JM5 系列、JM7 系列、JM9 系列高性能 GPU 芯片，其中最新的 JM9 系列两款图形处理芯片皆已完成阶段性测试工作，并进入放量阶段。JM9 系列芯片应用领域广泛，可满足个性化桌面办公、网络安全保护、轨交服务终端、多屏高清显示输出和人机交互等多样化需求。

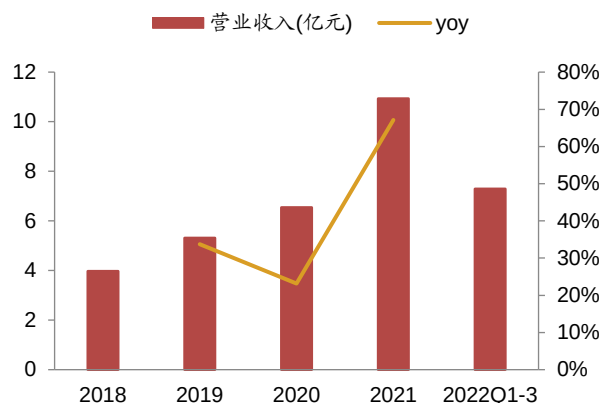
表19.景嘉微 JM9 产品性能参数

产品名称	主要技术指标	
JM9 系列图形处理芯片	2D 图形生成功能	支持 DirectFB 1.7.7; 支持 OpenVG 1.1 矢量图形加速
	3D 图形生成功能	支持 OpenGL4.0, OpenCL3.0, Vulkan1.1, OpenGLES3.2; 像素填充率 8G Pixels/s ; 32 位单精度浮点性能 512GFlops
	内核性能	内核时钟频率 1GHz (支持动态调频)
	总线接口	PCIe 4.0 X8
	显存带宽	25.6GB/s
	显存容量	8GB
	显示接口	支持 2 路独立的图形显示控制器, 支持 2 路 HDMI2.0, 1 路 eDP1.2, 1 路 VGA 输出
	功耗	< 15W

数据来源：公司公告，财通证券研究所

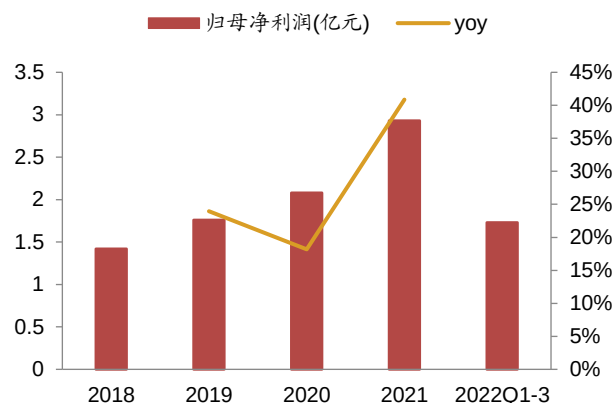
公司营收和归母净利润持续上升，2021 年全年实现营收 10.93 亿，同比增长率达 67.1%，实现归母净利润 2.93 亿元，同比上升 40.9%。

图66.景嘉微营业收入及增速(亿元)



数据来源：，财通证券研究所

图67.景嘉微归母净利润及增速(亿元)



数据来源：，财通证券研究所

4.4 芯原股份

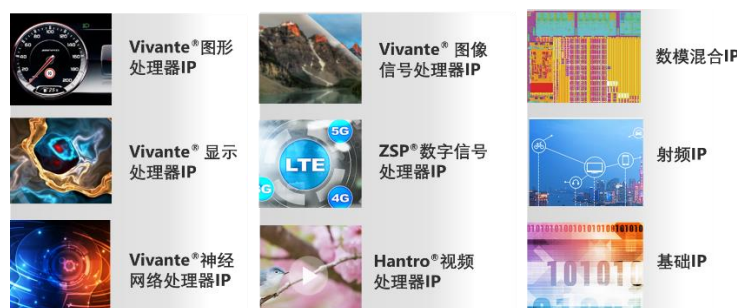
芯原依托自主半导体 IP，为客户提供平台化、全方位、一站式芯片定制服务和半导体 IP 授权服务，拥有独特的“芯片设计平台即服务”经营模式。公司可提供高清视频、物联网连接、数据中心等多种一站式芯片定制解决方案，拥有自主可控的图形处理器 IP、神经网络处理器 IP 等五类处理器 IP 及 1400 多个数模混合 IP 和射频 IP，可快速打造出从定义到测试封装完成的半导体产品，业务范围覆盖消费电子、汽车电子、物联网等多种应用领域。据 IPnest 在 2021 年的统计，芯原的半导体 IP 销售收入排中国大陆第二，全球第七，其中公司的图形处理器 IP 排名全球前三。

图68.一站式芯片定制服务



数据来源：芯原股份官网，财通证券研究所

图69.芯原 IP 产品阵



数据来源：芯原股份官网，财通证券研究所

公司的 GPU IP 已被众多主流和高端的汽车品牌所采用，同时，公司基于约 20 年 Vivante GPU 的研发经验，所推出的 Vivante 3D GPGPU IP 还可提供从低功耗嵌入式设备到高性能服务器的计算能力，满足广泛的人工智能计算需求。

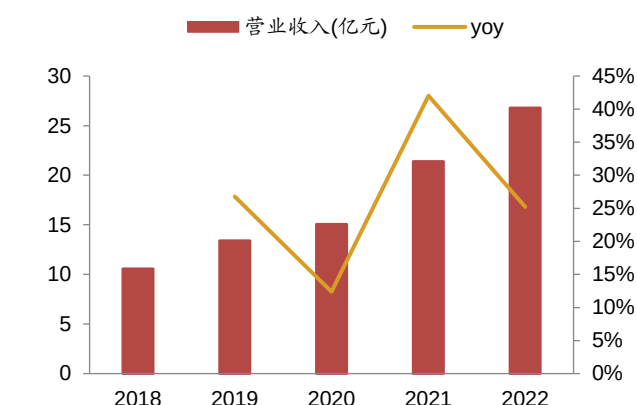
表20.芯原 Vivante®图形处理器 IP 各系列产品参数

	Vivante 3D GPGPU IP	Vivante 3D GPU IP	Vivante 2.5D GPU IP	Vivante 2D GPU IP
简介	提供卓越的计算能力，应用领域覆盖低功耗嵌入式设备到高性能服务器。	具有更高的能效和成本效益，并提高了 3D 图形渲染和计算性能。	针对需要 UI 硬件加速渲染的 MCU 和 MPU 应用而设计。	提高多界面合成性能并有效减少带宽，为移动、嵌入式和消费设备带来全新体验。
Linux Kernel support	-	√	√	×
Android Support	-	√	×	×
Windows Embedded Compact Support	-	√	×	√
Target UI Resolution	-	WVGA~1080p	up to 4K	up to 1080p
Software API Drivers	OpenCL 1.1/1.2/3.0、OpenCV	Vulkan 1.1/1.0、OpenGL ES 2.0/1.1、OpenVG 1.1、OpenCV、OpenCL 1.2/1.1	Vector Graphics、VGLite API	-

数据来源：芯原股份官网，财通证券研究所

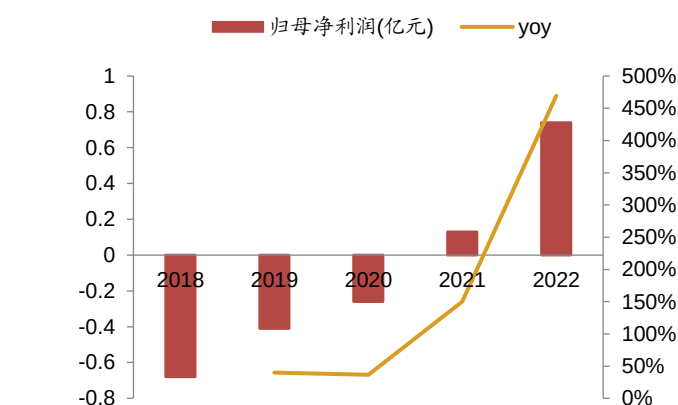
公司营收持续增长，归母净利润保持高增速。2020-2022 年公司营业收入分别为 15.06 亿元、21.39 亿元、26.79 亿元，归母净利润在 2021 年扭亏为盈后持续增长，于 2022 年达 0.74 亿元，同比上升 469.2%。

图70.芯原股份营业收入及增速(亿元)



数据来源：，财通证券研究所

图71.芯原股份归母净利润及增速(亿元)



数据来源：，财通证券研究所

4.5 龙芯中科

龙芯中科主要产品与服务包括处理器及配套芯片产品与基础软硬件解决方案业务。公司基于信息系统和工控系统两条主线，秉承独立自主和开放合作的运营模式，面向网络安全、工控及物联网等领域与合作伙伴保持全面的市场合作，产品广泛应用于电子政务、能源、交通、金融等行业领域，相关软硬件开发人员数万人，已经形成强大的产业链与生态支撑能力。在通用图形处理器及系统研发方面，龙芯中科于 2017 年开始研发 GPU，已掌握 GPU 研发的关键技术，第一款 GPU IP 核已经在龙芯 7A2000 桥片样片中流片成功。

图72.龙芯中科芯片产品



数据来源：龙芯中科官网，财通证券研究所

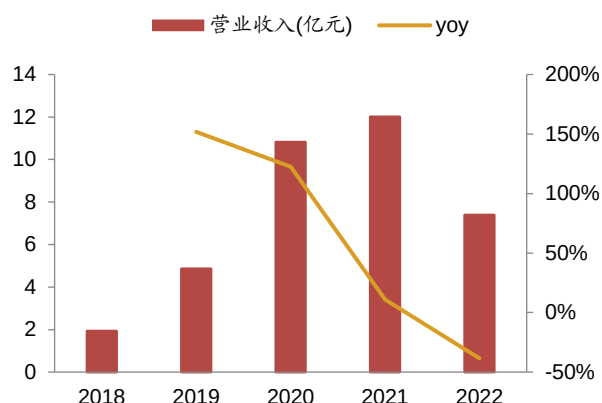
图73.龙芯中科自主生态技术架构



数据来源：龙芯中科官网，财通证券研究所

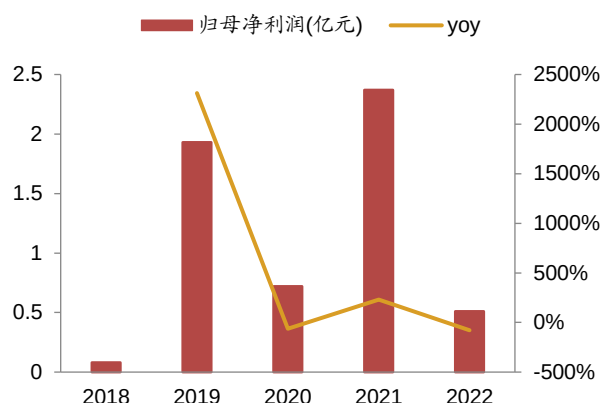
受周期下行和疫情反复影响，公司营收增速持续下降，2022 年全年实现营收 7.39 亿元，同比下跌 38.5%，归母净利润由 2021 年的 2.37 亿下跌至 0.51 亿元，同比下跌 78.5%。

图74.龙芯中科营业收入及增速(亿元)



数据来源：，财通证券研究所

图75.龙芯中科归母净利润及增速(亿元)



数据来源：，财通证券研究所

4.6 壁仞科技（非上市）

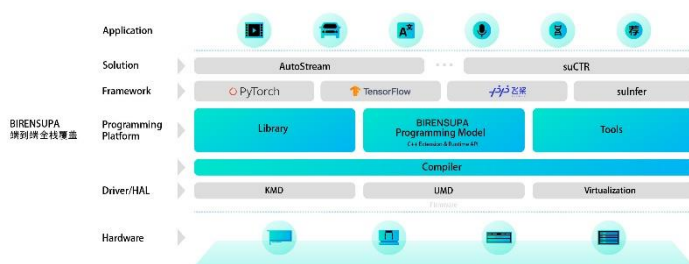
壁仞科技创立于2019年，在GPU、DSA（专用加速器）和计算机体系结构等领域具有深厚的技术积累。公司致力于开发原创性的通用计算体系，建立高效的软硬件平台，同时在智能计算领域提供一体化的解决方案。

图76.BR100 系列通用 GPU 芯片



数据来源：壁仞科技官网，财通证券研究所

图77.BIRENSUPA 软件平台



数据来源：壁仞科技官网，财通证券研究所

2022年8月公司发布的通用GPU芯片BR100创下全球通用GPU算力记录，峰值算力达到国际厂商在售旗舰产品3倍以上。BR100率先采用Chiplet技术、新一代主机接口PCIe 5.0、支持CXL互连协议，确立了公司在国内厂商间的技术领先地位。公司坚持自主研发，同步推出原创架构“壁仞仞”和自研BIRENSUPA软件平台，实现了BR100性能的大幅提升。以壁仞科技于2022年8月发布的首款GP GPU BR100为例，该芯片采用Chiplet技术，16位浮点算力

达到 1000T 以上、8 位定点算力达到 2000T 以上，单芯片峰值算力达到 PFLOPS 级别，是国际厂商在售旗舰产品的 3 倍以上，创造了全球通用 GPU 的算力记录。

表21.壁仞 BR100 系列产品参数

产品	壁仞™100P	壁仞™104P
制程	7nm	7nm
系统接口、带宽、互连协议	PCIe5.0X16, 128GB/s, 支持 CXL	PCIe5.0X16, 支持 CXL
FP32TFLOPS (峰值)	240	NA
TF32+TFLOPS (峰值)	480	256
BF16TFLOPS (峰值)	960	512
INT8TOPS (峰值)	1,920	1024
内存容量、接口位宽、带宽	64GBHBM2E; 4,096bit,1.64TB/s	32GBHBM2E; 2,048bit,819GB/s
互连	448GB/sBLink™, 支持 7 个端口, 最高可实现 8 卡全互连	192GB/sBLink™, 支持 3 个端口, 最高可实现 4 卡全互连
安全虚拟实例	最高 8 份	最高 4 份
视频编解码 (FHD@30fps)	64 路 HEVC/H.264 编码/512 路 HEVC/H.264 解码	32 路 HEVC/H.264 编码、256 路 HEVC/H.264 解码
TDP	450-550W	300W

数据来源：壁仞科技官网，财通证券研究所

4.7 摩尔线程（非上市）

摩尔线程专注于设计高性能通用 GPU 芯片，提供图形计算和 AI 计算的元计算平台的集成电路高科技公司。公司高管团队来自英伟达、AMD、ARM 等知名芯片公司，拥有丰富的 GPU 研究经验，致力于创新面向元计算应用的新一代 GPU，构建融合视觉计算、3D 图形计算、科学计算及人工智能计算的综合计算平台，建立基于云原生 GPU 计算的生态系统。

图78.开发者软件 MT GPU Management Center



数据来源：摩尔线程官网，财通证券研究所

图79.第一代 MUSA 架构



数据来源：摩尔线程官网，财通证券研究所

2022 年 11 月，公司推出基于第二代 MUSA 架构的处理器“春晓”，并基于“春晓”GPU 发布面向消费领域的国产芯片显卡 MTT S80 和面向服务器应用的 MTT S3000 显卡。同时，公司围绕 MUSA 发布了系列 GPU 软件栈与应用工具，包括 MUSA 开发者套件、云原生 sGPU 技术及元宇宙平台 MTVERSE 等。

表22.摩尔线程产品参数

产品	MTT S80	MTT S30	MTT S50	MTT S3000	MTT S2000
流处理单元	4096 MUSA 核心	1024 MUSA 核心	2048 MUSA 核心	4096 MUSA 核心	4096 MUSA 核心
GPU 核心频率	1.8GHz	NA	NA	1.9GHz	NA
FP32 算力	14.4 TFLOPS	2.6 TFLOPS	5 TFLOPS	15.2 TFLOPS	10.6 TFLOPS
显存容量	16GB	4GB	8GB	32GB	32GB
显存位宽	256 bit	128 bit	256 bit	256 bit	256 bit
功耗	255W	40W	75W	250W	150W
最大分辨率	7680 × 4320	3840 × 2160	7680 × 4320	NA	NA

数据来源：摩尔线程官网，财通证券研究所

4.8 芯动科技（非上市）

芯动科技是国内一站式 IP 和芯片定制及 GPU 领军企业，聚焦计算、存储、连接等三大赛道，提供从 55 纳米到 5 纳米全套高速 IP 核以及高性能定制芯片解决方案。公司拥有经验丰富的技术团队，成立 16 年来已赋能全球数百家知名客户，授权逾 80 亿颗高端 SoC 芯片进入规模量产，拥有过十亿颗 FinFET 定制芯片成功量产经验。

图80.芯动科技的定制服务



数据来源：芯动科技官网，财通证券研究所

图81.芯动科技核心产品



数据来源：芯动科技官网，财通证券研究所

公司瞄准商用市场推出芯动风华系列 GPU。该系列 GPU 性能强劲、跑分领先、功耗低、自带智能计算能力，且全面支持国内外 CPU/OS 和生态，包括 Linux、ows 和 Android。

表23.芯动风华系列 GPU 主要参数

产品	风华 1 号	风华 2 号
单精度浮点算力	5TFLOPS	1.5TFLOPS
AI 性能	12.5 TOPS (INT8)	12.5 TOPS (INT8)
像素填充率	160GPixels/sec	48GPixels/sec
接口规格	PCIe4.0 x8 向下兼容	PCIe3.0 x8
显存类型	GDDR6/GDDR6X	LPDDR5X/5/4X/4
显存容量	8G/16G	2GB / 4GB / 8GB
计算 API	OpenCL 1.2/2.1EP/3.0	OpenCL 3.0

数据来源：芯动科技官网，财通证券研究所

4.9 兆芯（非上市）

兆芯成立于 2013 年，提供高效、兼容、安全的自主通用处理器和芯片组等产品，公司掌握自主通用处理器及其系统平台芯片研发设计的核心技术，全面覆盖其微架构与实现技术等关键领域，拥有较为完整的知识产权体系，截至目前已获权约 1300 件专利。

2020 年，兆芯将自身 GPU 业务进行切分独立，建立了格兰菲智能科技有限公司。公司目前已推出 Arise-GT10C0 芯片及 Glenfly Arise-GT-10C0 显卡。芯片内置完全独立自主研发的新一代图形图像处理引擎，兼容银河麒麟 KOS、统信软件 UOS、ows 等主流操作系统，同时可在 X86、ARM、MIPS 等主流硬件平台操作运行，支持多种图形和图像的 API 接口标准。

表24.兆芯 Arise-GT10C0 芯片介绍

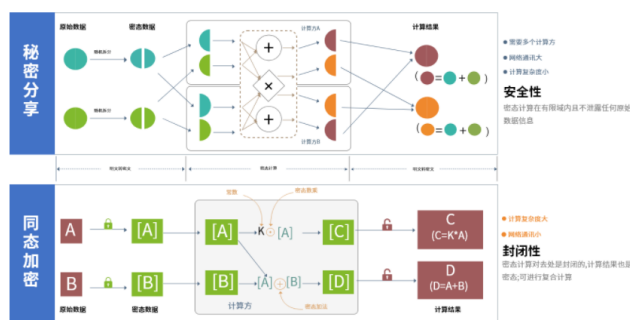
产品	Arise-GT10C0
工艺制程	28nm
单精度浮点算力	1.5TFLOPS
频率	500MHz
像素填充率	48GPixels/sec
纹理填充率	96Texels/sec
显存类型	DDR4
显存频率	1200MH
接口规格	PCIe3.0 x8

数据来源：格兰菲智能官网，财通证券研究所

4.10 天数智芯（非上市）

天数智芯致力于开发自主可控、国际领先的高性能通用 GPU 产品并提供解决方案，是国内头部通用 GPU 高端芯片及超级算力系统提供商。公司以“成为智能社会的赋能者”为使命，立足客户、市场的需求，加速 AI 计算与图形渲染融合，探索通用 GPU 赶超发展道路，产品广泛应用于智算重心、智慧医疗、互联网、智能制造等领域。

图82.基于 GPU 的 TEE 隐私计算解决方案



数据来源：天数智芯官网，财通证券研究所

图83.公司发布的人工智能开源平台 DeepSpark



数据来源：天数智芯官网，财通证券研究所

12月20日，天数智芯推出通用 GPU 推理产品“智铠 100”及其丰富的 AI 应用案例。智铠 100 计算性能高、应用覆盖广、使用成本低，支持 FP32、FP16、INT8 等多精度混合计算，可提供最高 384TFlops@int8、96TFlops@FP16、24TFlops@FP32 的峰值算力，800GB/s 的理论峰值带宽以及 128 路并发的多种视频规格解码能力。

表25.天数智芯 BI-V100 主要产品参数

型号	BI-V100
架构	通用 GPU
频率	1.5GHz
制程及封装	TSMC 7nm FinFET 2.5D COWOS 封装
内存规格	32 GB DRAM (48GB) HBM2 PCIe Gen4.0 x 16 lane
接口规格	共享 64 GB/s 主控双向带宽 共享 64 GB/s 片间互联带宽
半精浮点峰值算力（含 TCU）	147 TFLOPS
整型峰值算力	295 TOPS@INT8 支持 INT32, INT16 计算
内存	32 GB DRAM (4*8GB) HBM2

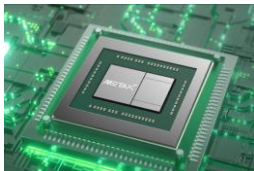
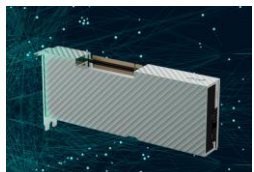
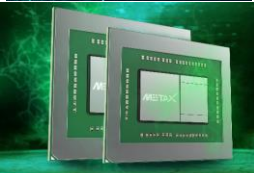
数据来源：天数智芯官网，财通证券研究所

4.11 沐曦（非上市）

沐曦于 2020 年 9 月成立于上海，致力于为异构计算提供全栈 GPU 芯片及解决方案，可广泛应用于人工智能、智慧城市、自动驾驶、数字孪生、元宇宙等前沿领域。公司拥有技术完备、设计和产业化经验丰富的团队，核心成员平均拥有近 20 年高性能 GPU 产品端到端研发经验。

公司拥有完全自主研发的 GPU IP、指令集和架构，以及兼容主流 GPU 生态的完整软件栈（MXMACA），产品具备高能效、高通用性。目前已推出 MXN 系列 GPU（曦思）用于 AI 推理，MXC 系列 GPU（曦云）用于 AI 训练及通用计算，以及 MXG 系列 GPU（曦彩）用于图形渲染，可满足数据中心对高能效和高通用性的算力需求。

表 26. 沐曦 GPU 产品矩阵

产品系列	产品外观	描述
MXN 人工智能推理 GPU		MXN 系列是面向云端数据中心应用的人工智能推理产品，采用高带宽内存，提供强大的 AI 算力和领先的视频编解码能力，可广泛应用于智慧城市、公有云计算、智能视频处理、云游戏等场景。
MXC 通用计算 GPU		MXC 系列通用 GPU(GPGPU)芯片是针对 AI 训练、AI 推理及通用计算的完美解决方案，沐曦自主知识产权架构提供强大高精度及多精度混合算力，可广泛应用于人工智能、数据中心以及通用计算、教育和科研等场景。
MXG 图形渲染 GPU		MXG 系列 GPU 是针对图形渲染加速的解决方案，沐曦自主知识产权架构提供卓越的图形图像渲染与视频处理能力，可广泛应用于元宇宙、云桌面、云游戏、云手机、数字孪生、XR 等场景。

数据来源：沐曦官网，财通证券研究所

5 风险提示

技术迭代风险：

技术迭代与产品研发进度可能不及预期。GPU 行业需要技术的不断迭代以推动新产品的落地，持续的技术研发周期需要长期稳定地投入大量人才资源与资金资源。在研发产出存在较高不确定性的情况下，可能会出现技术研发进度迟滞或新产品周期被动延长，将使企业遭遇亏损。

宏观经济风险：

宏观经济复苏可能不及预期。行业受宏观经济形势影响较大，新冠疫情后全球宏观经济平稳复苏。未来存在国内外宏观经济波动或复苏趋势放缓可能，将不利于企业的经营发展。

国产替代风险：

国产替代可能不及预期。受国内技术发展水平和国际贸易政策等不确定因素的影响，国产替代需求可能无法完全释放，将影响企业的发展前景。

行业竞争风险：

行业竞争可能加剧。国外龙头企业在技术与生态上占据优势地位，国内厂商成熟产品逐步涌现，市场竞争愈发激烈，将对企业的经营与战略规划提出更大挑战。

信息披露**● 分析师承诺**

作者具有中国证券业协会授予的证券投资咨询执业资格，并注册为证券分析师，具备专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解。本报告清晰地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，作者也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。

● 资质声明

财通证券股份有限公司具备中国证券监督管理委员会许可的证券投资咨询业务资格。

● 公司评级

买入：相对同期相关证券市场代表性指数涨幅大于 10%；

增持：相对同期相关证券市场代表性指数涨幅在 5%~10%之间；

中性：相对同期相关证券市场代表性指数涨幅在-5%~5%之间；

减持：相对同期相关证券市场代表性指数涨幅小于-5%；

无评级：由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

● 行业评级

看好：相对表现优于同期相关证券市场代表性指数；

中性：相对表现与同期相关证券市场代表性指数持平；

看淡：相对表现弱于同期相关证券市场代表性指数。

● 免责声明

财通证券股份有限公司的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本报告的信息来源于已公开的资料，本公司不保证该等信息的准确性、完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的邀请或向他人作出邀请。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本公司通过信息隔离墙对可能存在利益冲突的业务部门或关联机构之间的信息流动进行控制。因此，客户应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司不对任何人使用本报告中的任何内容所引致的任何损失负任何责任。

本报告仅作为客户作出投资决策和公司投资顾问为客户提供投资建议的参考。客户应当独立作出投资决策，而基于本报告作出任何投资决定或就本报告要求任何解释前应咨询所在证券机构投资顾问和服务人员的意见；

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。