中信证券
CITIC SECURITIES

半导体行业专题

ChatGPT对GPU算力的需求测算与相关分析

中信证券研究部 雷俊成/王子源/徐涛/杨泽原

2023年2月16日

核心观点：单个大模型可带来2万GPU销售量，搜索引擎带来成倍空间

核心观点：

1. **短期内GPU增量与市场规模：**参考OpenAI算法，假设每日1亿用户，每人进行10条交互，每个问题的回答长度为50词，算力利用率30%，则单个大语言模型（LLM）的日常需求有望带来2.13万片A100的增量，对应市场规模2.13亿美元。**假设有5家大企业推出此类LLM，则总增量为10.7片A100，对应市场规模10.7亿美元。**
2. **短期服务器增量与市场规模：**单个服务器包含8个GPU，因此单个LLM带来2669台服务器需求，对应市场规模3.39亿美元，**5家大企业共需要13345台，对应市场规模20亿美元。**
3. **长期市场空间：**参考谷歌，若每日搜访问30亿次，**需要106.74万张A100，对应13.3万台服务器DGX A100，带来市场空间200亿美元。**

市场规模相关参数/假设

- A100单卡算力：19.5TFLOPS/s
- 日常算力利用率：30%（依据经验）
- GPU单价：1万美元（A100）
- 每台服务器搭载GPU数量：8
- 服务器单价：15万美元（DGX Station A100）
- 做LLM模型的企业数量：5（BAT、华为、字节）

关键中间变量：GPU与服务器增量

1亿用户所需GPU数量：21348（A100）

= 近期单日交互+训练总算力1.08E+10TFLOPS

÷ A100单卡算力19.5T/s ÷ 算力利用率30%

1亿用户所需服务器数量：2669（DGX A100）

= 一个LLM模型所需GPU数量：21348（A100）

÷ 每台服务器搭载GPU数量：8

5家企业对应10.7万片A100、1.33万台服务器

短期国内GPU/服务器增量市场规模

1亿用户带来国内GPU总市场规模：2.13亿美元

= 一个LLM模型所需GPU数量：21348（A100）

× GPU单价：1万美元（A100）

1亿用户带来国内服务器市场规模：3.39亿美元

= 一个LLM所需服务器数量：2669

× 服务器单价：15万美元（A100）

5家企业对应10.7亿美元GPU、20亿美元服务器

远期GPU增量空间

谷歌+LLM所需GPU数量：1067415（A100）

= 远期总算力需求：5.4 E+11 TFLOPS

÷ A100单卡算力：19.5TFLOPS/s

÷ 算力利用率：30%

谷歌+LLM所需服务器数量：133427（GPU/8）

注：远期由于更高算力的GPU出现或更高效的计算方式，对应市场空间可能变化。

资料来源：Raconteur，OpenAI：Language Models are Few-Shot Learners，NVIDIA官网，Amazon，中信证券研究部

核心观点

- **技术差距：GPGPU的核心壁垒是高精度浮点计算及CUDA生态。从高精度浮点计算能力来看，国内GPU产品与国外产品的计算性能仍或有一代以上差距；在软件和生态层面与英伟达CUDA生态的差距则更为明显。**
  - AI计算GPU领域，国内壁仞科技发布的BR100产品在FP32单精度计算性能上实现超越NVIDIA A100芯片，但是不支持FP64双精度计算；天数智芯推出的天垓100的FP32单精度计算性能实现超越A100芯片，但是在INT8整数计算性能方面却低于A100；海光推出的DCU实现了FP64双精度浮点计算，但是其性能为A100的60%左右，大概相当于其4年前水平。因此，从高精度浮点计算能力来看，国内GPU产品与国外产品的计算性能仍或有一代以上差距。
  - 但是，GPU不仅在硬件上需要提升算力，软件层面对于GPU的应用和生态布局尤其重要，英伟达凭借CUDA构建生态壁垒占领全球GPU市场90%的份额。目前国内企业多采用开源的OpenCL进行自主生态建设，但这需要大量的时间进行布局；我们对比AMD从2013年开始建设GPU生态，近10年时间后用于通用计算的ROCm开放式软件平台才逐步有影响力，且还是在兼容CUDA的基础上。因此我们认为国内厂商在软件和生态层面与英伟达CUDA生态的差距较计算性能更为明显。
  - 虽然目前国内产品的计算性能和软件生态实力与国际厂商还有差距，但是，国内厂商依然在奋起直追，努力实现GPGPU的国产化突破。
- **我们认为长久来看，美国对中国高端GPU的禁售令反而给国产GPGPU和AI芯片厂商带来快速发展的机会。**
  - 短期来看，我们认为对高端通用计算GPU的禁令可能会影响英伟达和AMD的GPU产品在中国的销售，中国AI计算、超级计算和云计算产业进步受到一定的阻碍。可使用英伟达和AMD还没有被禁止的及国产厂商的中高计算性能CPU、GPU、ASIC芯片等替代。
  - 长期来看，国产CPU、GPU、AI芯片厂商受益于庞大的国内市场，叠加国内信创市场带来国产化需求增量，我们预期国内AI芯片的国产化比例将显著提升，借此机会进行产品升级，逐渐达到国际先进水平，突破封锁。
- **对于国内厂商，建议重点关注实现自主创新，打造自主生态体系，打磨产品实现稳定供货的公司。**
  - 重点关注能够实现GPU领域的自主创新，实现架构、计算核、指令集及基础软件栈的全自研的设计公司。
  - 同时，不止成功点亮，要能满足测试、客户适配、稳定供货等一系列要求，成功量产并实现规模应用，实现GPGPU的国产替代。
- **建议关注：**
  - 国内企业：1）芯片：龙芯中科（国内PC CPU龙头，自主研发GPGPU产品）、海光信息（国内服务器CPU龙头，推出深度计算处理器DCU）、景嘉微（国内图形渲染GPU龙头）、寒武纪（国内ASIC芯片龙头）、澜起科技（国内服务器内存接口芯片龙头）；2）PCB：胜宏科技、兴森科技、沪电股份；3）先进封装：通富微电、甬矽电子、长电科技、长川科技等。
  - 海外企业：英伟达（全球GPU龙头）、AMD（全球CPU/GPU领先厂商）、英特尔（全球CPU龙头）、美光（全球存储芯片龙头）。
- **风险因素：用户拓展不及预期风险，AI技术及新产品开发发展不及预期风险，外部制裁加剧风险，宏观经济需求下行风险。**

相关上市公司

ChatGPT相关上市公司及近期涨跌幅（截至2023年2月14日）

分类	公司名	代码	市值（亿元人民币）	ChatGPT 2022年11月30日上线至今涨跌幅	2023年初至今涨跌幅
CPU	龙芯中科	688047.SH	488.78	49%	43%
	海光信息	688041.SH	1,235.85	28%	33%
	中科曙光	603019.SH	425.88	24%	31%
	英特尔	INTC.O	8,049.41	-4%	9%
	AMD	AMD.O	9,134.63	7%	28%
GPU	景嘉微	300474.SZ	381.51	45%	54%
	英伟达	NVDA.O	36,527.90	29%	49%
AI芯片	寒武纪-U	688256.SH	342.62	35%	57%
	澜起科技	688008.SH	713.46	-7%	0%
	Mobileye	MBLY.O	2,343.44	50%	22%
FPGA	紫光国微	002049.SZ	1,032.70	-8%	-8%
	复旦微电	688385.SH	443.24	-9%	-1%
	安路科技-U	688107.SH	283.43	13%	10%
DPU	左江科技	300799.SZ	136.25	4%	2%
IP	芯原股份-U	688521.SH	308.66	30%	41%
AI SoC	瑞芯微	603893.SH	368.62	14%	28%
	晶晨股份	688099.SH	348.41	12%	20%
	富瀚微	300613.SZ	152.06	18%	32%
PCB	兴森科技	002436.SZ	205.11	1%	25%
	胜宏科技	300476.SZ	155.63	28%	39%
	生益电子	688183.SH	94.00	11%	21%
	沪电股份	002463.SZ	294.93	23%	31%
先进封装	长电科技	600584.SH	513.58	15%	25%
	通富微电	002156.SZ	334.58	28%	34%
	甬矽电子	688362.SH	112.76	-1%	27%
	华峰测控	688200.SH	276.32	23%	10%
	长川科技	300604.SZ	289.65	-17%	8%
存储	美光	MU.O	4,470.81	5%	20%

资料来源：Wind，中信证券研究部

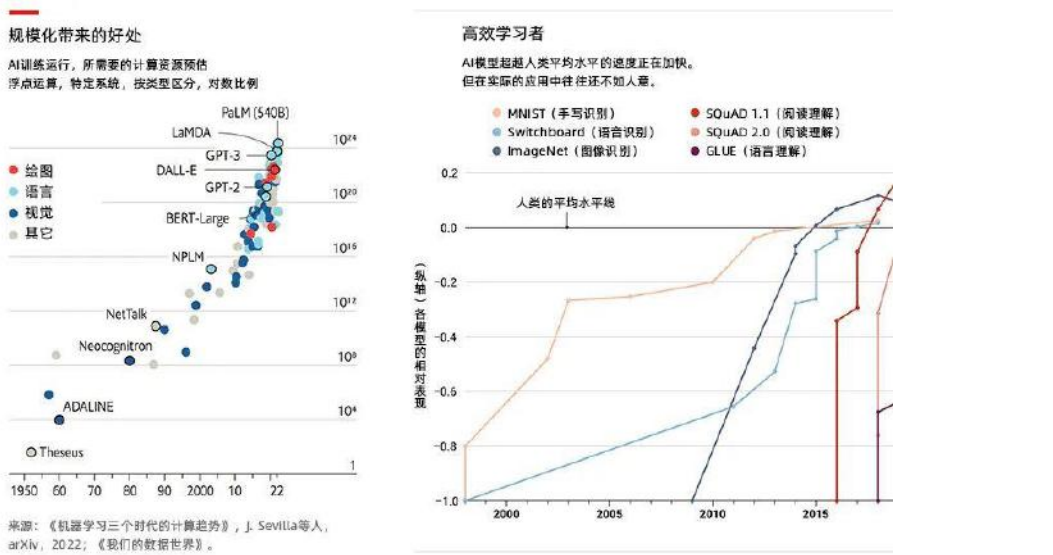
CONTENTS

目录

1.1 生成式AI：实现创造，部分领域的能力超越人类的基准水平

- **不同于分析式AI只能做些分析型或机械式的认知计算，生成式AI可以创造有意义并具备美感的东西，而且在某些情况下，其生成的结果可能比人类手工创造的还要好。**
  - 机器可以分析数据，并针对不同用例需求找到相应的规律，且在不断迭代，变得越来越聪明，这种机器被称为“分析式人工智能”（Analytical AI），或者传统AI。机器并非如之前那样仅分析已有的数据，而是创造了全新的东西，这一新型的AI被称为“生成式人工智能”（Generative AI）。
- **2017年谷歌推出一种用于自然语言理解的新型神经网络架构——Transformers模型，不但能生成质量上乘的语言模型，同时具有更高的可并行性，大大降低了所需的训练时间。这些小样本学习模型，可以更容易地针对特定领域做定制修改。**
  - 2015-2020年，用于训练这些模型的计算量增加了6个数量级，其表现在手写、语音和图像识别、阅读理解和语言理解方面超过了人类的基准水平。

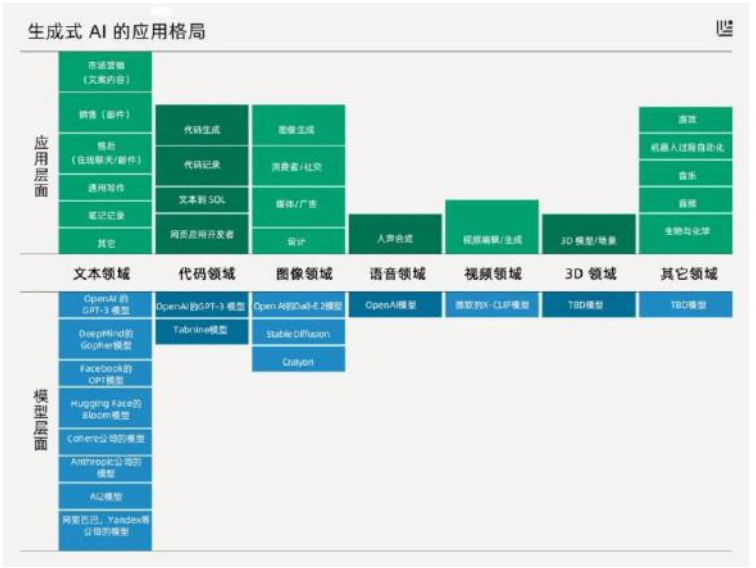
随着AI模型逐渐发展壮大，已经开始超越人类的基准水平



来源：《机器学习三个时代的计算趋势》，J. Sevilla等人，arXiv，2022；《我们的数据世界》。

资料来源：《机器学习三个时代的计算趋势》——Sevilla等人，arXiv，2022，《生成式AI：充满创造力的新世界》——红杉汇内参微信公众号

生成式AI的应用格局

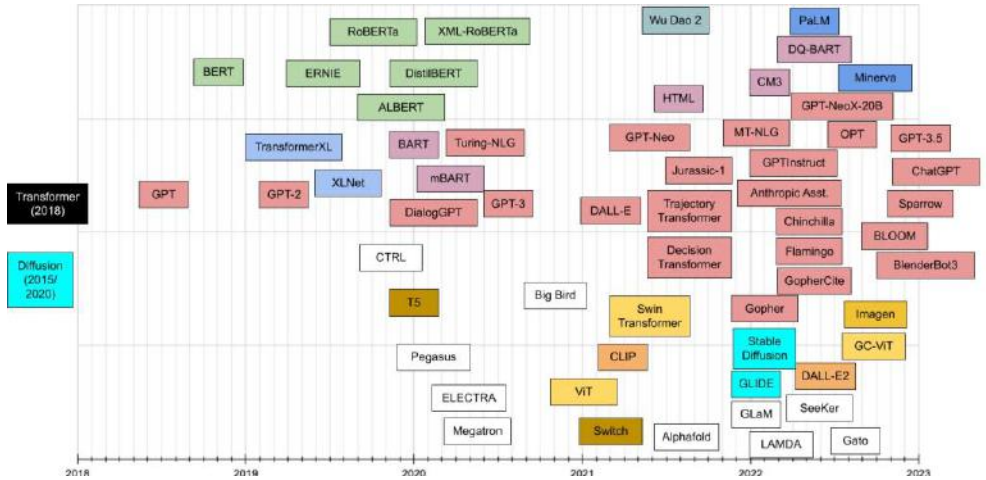


资料来源：《生成式AI：充满创造力的新世界》——红杉汇内参微信公众号

1.2 预训练模型：大模型提高准确率，2018年开始步入快车道

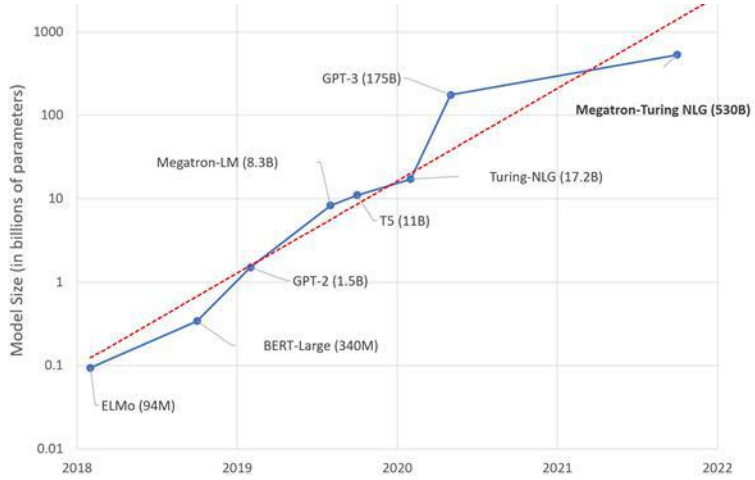
- **预训练模型使得模型的训练可以被复用，大幅降低训练成本，但是前期需要大量的数据进行预训练。**
  - 预训练模型是一种迁移学习的应用，对句子每一个成员的上下文进行相关的表示，通过隐式的方式完成了语法语义知识的学习。预训练模型通过微调的方式具备很强的扩展性，每次扩展到新场景时，只需要针对这个场景的特定标注数据进行定向的学习，便可以快速应用。
- **2018年以来，国内外超大规模预训练模型参数指标不断创出新高，“大模型”已成为行业巨头发力的一个方向。谷歌、百度、微软等国内外科技巨头纷纷投入大量人力、财力，相继推出各自的巨量模型。国外厂商自2021年开始进入“军备竞赛”阶段。**
  - 2018年，谷歌提出3亿参数BERT模型，大规模预训练模型开始逐渐走进人们的视野，成为人工智能领域的一大焦点。
  - 2019年，OpenAI推出15亿参数的GPT-2，能够生成连贯的文本段落，做到初步的阅读理解、机器翻译等。紧接着，英伟达推出83亿参数的Megatron-LM，谷歌推出110亿参数的T5，微软推出170亿参数的图灵Turing-NLG。
  - 2020年，OpenAI以1750亿参数的GPT-3，直接将参数规模提高到千亿级别。
  - 2021 年 1 月，谷歌推出的Switch Transformer模型以高达1.6 万亿的参数量打破了GPT-3作为最大AI模型的统治地位，成为史上首个万亿级语言模型。 2020年10月，微软和英伟达联手发布了5300亿参数的Megatron-Turing自然语言生成模型（MT-NLG）。2021年12月，谷歌还提出了1.2万亿参数的通用稀疏语言模型GLaM，在7项小样本学习领域的性能超过GPT-3。

2018年以来LLM算法（大规模语言算法）成长的时间线



资料来源：Xavier Amatriain，陈巍谈芯@知乎

近年来超大规模预训练模型参数增长趋势

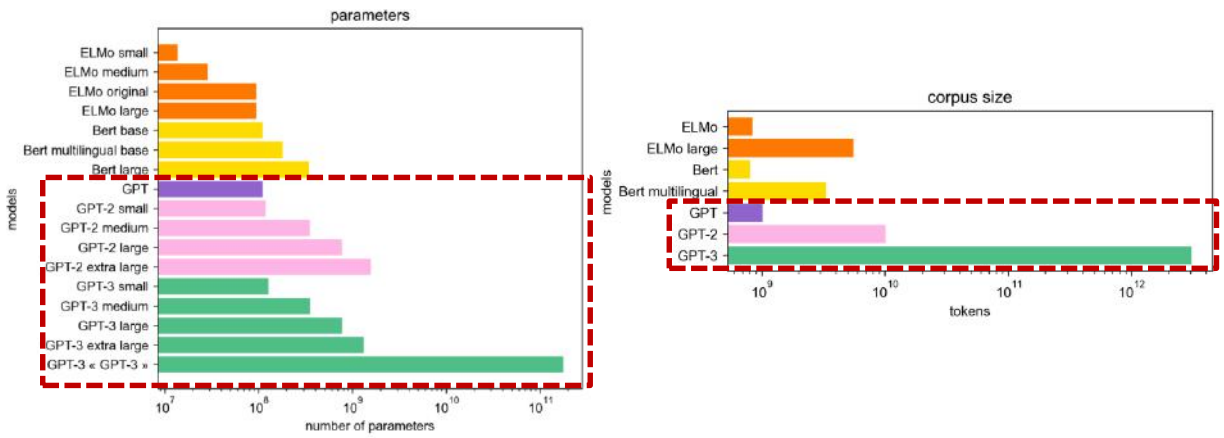


资料来源：《Large Language Models: A New Moore's Law?》——Julien Simon@Hugging Face

1.3 ChatGPT：基于OpenAI推出的深度学习模型GPT打造，成为迄今增长最快的消费应用程序

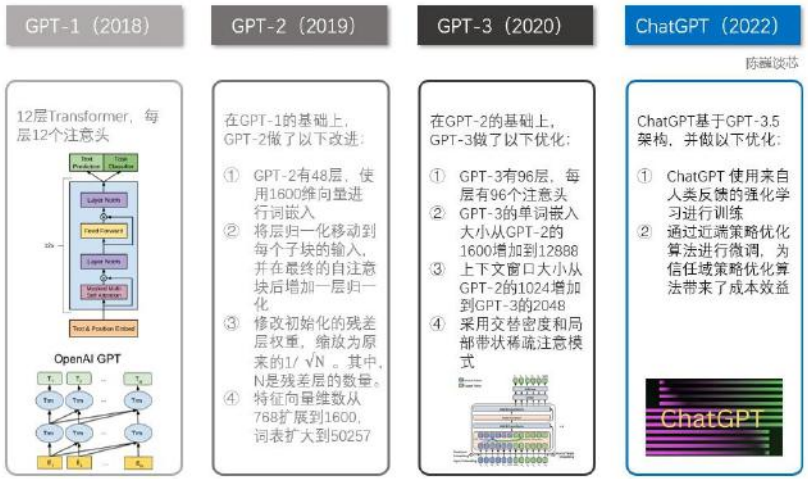
- **ChatGPT（Chat Generative Pre-trained Transformer，聊天生成式预训练器）是OpenAI开发的聊天机器人，于2022年11月推出。它建立在OpenAI开发的GPT-3大型语言模型之上，并使用监督学习和强化学习（人类监督）技术进行了微调。**
  - 虽然聊天机器人的核心功能是模仿人类谈话者，但ChatGPT是多功能的。例如，它可以编写和调试计算机程序，创作音乐、电视剧、童话故事和学生论文；回答测试问题(有时根据测试的不同，答题水平要高于平均水平)；写诗和歌词；模拟Linux系统；模拟整个聊天室等。
- **ChatGPT背后的公司为OpenAI，成立于2015年，由特斯拉CEO埃隆·马斯克、PayPal联合创始人彼得·蒂尔、Linkedin创始人里德·霍夫曼、创业孵化器Y Combinator总裁阿尔特曼（Sam Altman）等人出资10亿美元创立。OpenAI的诞生旨在开发通用人工智能（AGI）并造福人类。**
- **ChatGPT中的GPT（Generative Pre-trained Transformer），是OpenAI推出的深度学习模型。ChatGPT就是基于GPT-3.5版本的聊天机器人。**
  - 截至2022年12月4日，OpenAI估计ChatGPT用户已经超过100万；2023年1月，ChatGPT用户超过1亿，成为迄今增长最快的消费应用程序。
  - 2023年2月，OpenAI开始接受美国客户注册一项名为ChatGPT Plus的高级服务，每月收费20美元；此外，OpenAI正计划推出一个每月42美元的ChatGPT专业计划，当需求较低时可以免费使用。

GPT系列模型的数据集训练规模



资料来源：《The GPT-3 language model, revolution or evolution?》——Hello Future

ChatGPT与GPT 1-3的技术对比



资料来源：《6个问题，用专业视角带你全方位了解ChatGPT》——甲子光年微信公众号

1.4 算力需求：计算资源每3~4个月翻一倍，投入资金指数级增长

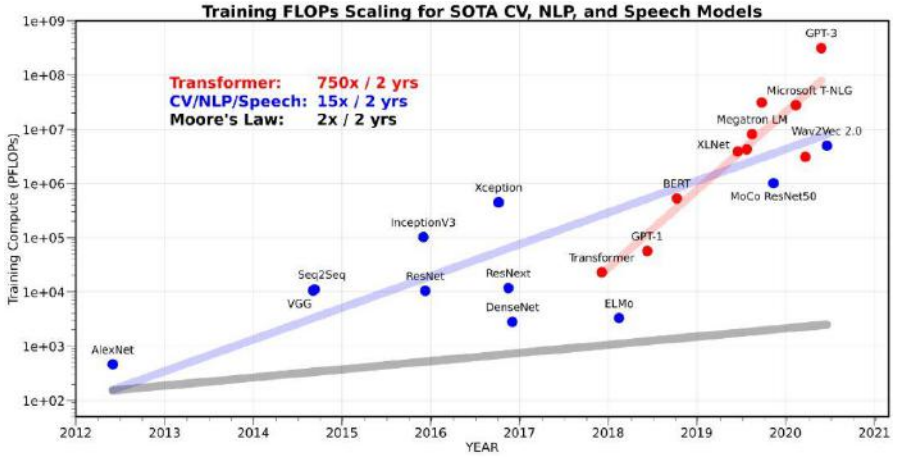
- **OpenAI预计人工智能科学研究要想取得突破，所需要消耗的计算资源每3~4个月就要翻一倍，资金也需要通过指数级增长获得匹配。**
  - 在算力方面，GPT-3.5在微软Azure AI超算基础设施（由V100GPU组成的高带宽集群）上进行训练，总算力消耗约 3640PF-days（即每秒一千万亿次计算，运行3640天）。
  - 在大数据方面，GPT-2用于训练的数据取自于Reddit上高赞的文章，数据集共有约800万篇文章，累计体积约40G；GPT-3模型的神经网络是在超过45TB的文本上进行训练的，数据相当于整个维基百科英文版的160倍。
- **按照量子位给出的数据，将一个大型语言模型（LLM）训练到GPT-3级的成本高达460万美元。**
  - 最新的GPT3.5在训练中使用了微软专门建设的AI计算系统，由1万个英伟达V100 GPU组成的高性能网络集群，总算力消耗约3640PF-days（PD），即假如每秒计算一千万亿（10^{20}）次，需要计算3640天。
  - 采购一片英伟达顶级GPU成本为8万元，GPU服务器成本通常超过40万元。对于ChatGPT而言，支撑其算力基础设施至少需要上万颗英伟达GPU A100，一次模型训练成本超过1200万美元。

预训练模型参数及所需要的算力情况

时间	机构	模型名称	模型规模	数据规模	使用单块V100的训练时间
2018.6	OpenAI	GPT-1	110M	4GB	3天
2018.10	Google	BERT	330M	16GB	50天
2019.2	OpenAI	GPT-2	1.5B	40GB	200天
2019.7	Facebook	RoBERTa	330M	160GB	3年
2019.10	Google	T5	11B	800GB	66年
2020.6	OpenAI	GPT-3	175B	2TB	355年

资料来源：做AI做的事儿微信公众号，《6个问题，用专业视角带你全方位了解ChatGPT》——甲子光年微信公众号

目前 SOTA 模型训练的浮点数运算量（以 FLOPs为衡量单位）

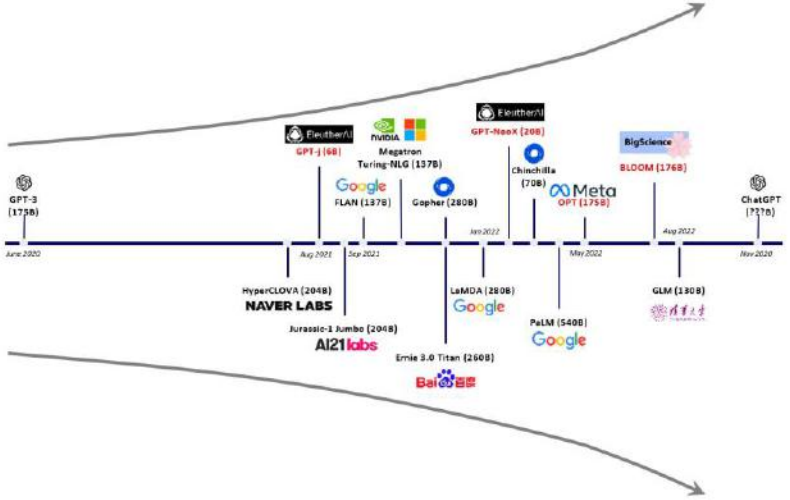


资料来源：《AI算力的阿喀琉斯之踵：内存墙》——Amir Gholami@OneFlow社区 注：蓝线上的是 CV, NLP和语音模型，模型运算量平均每两年翻 15 倍，红线上的是 Transformer 的模型，模型运算量平均每两年翻 750 倍。而灰线则标志摩尔定律下内存硬件大小的增长，平均每两年翻 2 倍。

1.5 产业竞争：训练成本逐渐降低，国内外科技巨头加速布局

- **根据《财富》杂志报道的数据，2022年OpenAI的收入为3000万美元，但净亏损预计为5.445亿美元。公司预测其2023年收入2亿美元，2024年收入预计超过10亿美元。**
  - 投入上：公司CEO阿尔特曼在推特上回答马斯克的问题时表示，在用户与ChatGPT的每次交互中，OpenAI花费的计算成本为“个位数美分”，随着ChatGPT变得流行，每月的计算成本可能达到数百万美元。
  - 创造价值上：ARK认为，AI工具的发展将不断提高生产力，到2030年，人工智能或将知识工作者的生产力提高4倍以上，将软件工程师的效率提高10倍以上，创造约200万亿美元的价值。
- **大模型高昂的训练成本让普通创业公司难以为继，因此参与者基本都是科技巨头。**
  - 在国内科技公司中，阿里巴巴达摩院在2020年推出了M6大模型，百度在2021年推出了文心大模型，腾讯在2022年推出了混元AI大模型。
  - 这些模型不仅在参数量上达到了千亿级别，而且数据集规模也高达TB级别，想要完成这些大模型的训练，就至少需要投入超过1000PetaFlop/s-day的计算资源。

大模型计算布局呈爆发增长态势



资料来源：Xavier Amatriain，陈巍谈芯@知乎

目前全球大模型计算布局情况

企业	大模型	参数	算力	数据量	模型类型
OpenAI	GPT3.5	1750 亿	3640（Pflops-day）/上万块V100 GPU 组成 gao 带宽集群算力	超过万亿单词的人类语言数据集	多模态预训练模型结合人类参与强化学习
清华大学等[1]	“八卦炉”（脑极AI模型）	174万亿（与人脑中突出数量相关）	“海洋之光”超级计算机（国产超算）	中文多模态数据集 M6-Corpus	
阿里	M6	10 万亿	512块 GPU	1.9TB 图像292GB 文本	
腾讯	“混元”HunYuan_tvr	万亿	腾讯太极机器学习平台	五大跨模态视频检索数据集	
华为云	盘古系列大模型	千亿	鹏城云脑II和全场景 AI 计算框架 MindSpore，2048 块 GPU	40TB 训练数据	
瀚舟	孟子	10亿	16 块 GPU	数百 G 级别不同领域的高质量语料	
微软和英伟达	Megatron-Turing	5300亿	280 块 GPU	3390 亿条文本数据	NLP 大模型
百度和鹏程实验室	ERNIW 3.0 Titan	2600 亿	鹏城云脑II（2048 块CPU）和百度飞桨	纯文本和知识图谱的 4TB 语料库	NLP 大模型
浪潮信息	源 1.0	2457 亿	4095（Pflops-day）/2128 张GPU	5000GB 高质量中文数据集	NLP 大模型
商汤科技等[2]	书生(INTERN+)	100亿	商汤AIDC，峰值算力 3740Petaflops3	-	计算器视觉模型
商汤科技	某世界最大规模计算器视觉模型	300亿			计算器视觉模型
中科院自动化所	紫东太初	千亿	昇腾 AI 基础软硬件平台	基于万条小规模数据集	图、文、音三模态

资料来源：《6个问题，用专业视角带你全方位了解ChatGPT》——甲子光年微信公众号 注：1、清华大学和阿里达摩院等合作提出；2、上海人工智能实验室联合商汤科技、香港中文大学、上海交通大学发布；3、Pflops-day 为算力单位，意为一天可以进行约 10^{20}运算。

1.6 ChatGPT带来的算力/GPU需求——测算原理、预训练需求分析

算力消耗测算原理

模型	训练总计算量（PF·日）	训练总计算量(flops)	模型参数量（百万）	训练词数（十亿）	单个词语消耗的总计算次数	计算反向传播后的算力消耗倍数	正向计算时每个词消耗浮点计算次数
BERT-Base	1.89	1.64E+20	109	250	6	3	2
BERT-Large	6.16	5.33E+20	355	250	6	3	2
RoBERTa-Base	17.36	1.50E+21	125	2,000	6	3	2
RoBERTa-Large	49.31	4.26E+21	355	2,000	6	3	2
GPT-3 Small	2.60	2.25E+20	125	300	6	3	2
GPT-3 Medium	7.42	6.41E+20	356	300	6	3	2
GPT-3 Large	15.83	1.37E+21	760	300	6	3	2
GPT-3 XL	27.50	2.38E+21	1,320	300	6	3	2
GPT-3 2.7B	55.21	4.77E+21	2,650	300	6	3	2
GPT-3 6.7B	138.75	1.20E+22	6,660	300	6	3	2
GPT-3 13B	267.71	2.31E+22	12,850	300	6	3	2
GPT-3 175B	3637.50	3.14E+23	174,600	300	6	3	2

核心原理：
每个训练词都会导致模型所有参数的更新，且每个训练词都需要消耗固定的浮点算力。因此：
总算力需求=模型参数量*训练词数*每个词的运算量

测算过程：
表格从右向左计算
1. 最基础的“原子”运算：1个词更新1个参数，需要计算1次乘法和1次加法，共2次浮点运算。
2. 如果是训练，则需要反向传播算法，反向传播需要的运算次数是正向传播2倍，故训练时每个词的运算量是推理情况的3倍，需要消耗6次浮点运算。（2次运算*算力消耗倍数3）
3. 按照核心公式求解，GPT-3的总算力消耗为1.746E+11 * 3E+11 * 6 = 3.14E+23 FLOPS
4. 进行单位换算， 3.14E+23 FLOPS = 3640 PF·日

资料来源：OpenAI：Language Models are Few-Shot Learners：附录D，中信证券研究部，注：为简单起见，本测算方法忽略了Attention计算的算力消耗，该部分占总算力消耗的10%以下

3000亿训练词如何构成

数据集	词数（十亿）	训练轮数	权重占比
网页爬虫	410	0.44	60%
WebText2	19	2.9	22%
Books1	12	1.9	8%
Books2	55	0.43	8%
维基百科	3	3.4	3%

- 不同数据集的数据质量和重要度不一致，因此重要度和质量更高的数据集会进行更多轮次的训练，从而提升其权重占比。
- 将每个数据集的词数乘以训练轮数，加在一起即得到3000亿词的训练数据。

预训练算力消耗及GPU需求测算

- 假设1：ChatGPT使用的数据集与GPT-3 175B模型相同
- 假设2：ChatGPT使用FP32数据格式完成训练

总计算量 = GPU数量 × GPU算力 × 计算用时

GPU数量 = 总计算量 ÷ GPU算力 ÷ 计算用时

GPU数量	总计算量	GPU算力	计算用时
186538/7723/3861/2574	3.14E+23 FLOPS	19.5 TFLOPS/s（A100 FP32）	1天/1个月/2个月/3个月

测算数据来源：
1. 总计算量来自上表OpenAI论文
2. GPU算力来自NVIDIA官网
3. 计算用时取决于语言模型开发者试图在多长时间内完成训练

左侧求得GPU数量与右侧计算用时一一对应，例如一个月完成训练需要7723张A100 GPU

资料来源：OpenAI：Language Models are Few-Shot Learners，NVIDIA官网，中信证券研究部

1.7 ChatGPT带来的算力需求——日常交互、日常训练需求分析

测算核心假设

- **核心假设1-算力需求影响因素：**模型参数量（175B）和单个词计算量（训练6次，推理2次）不变，**算力需求变化主要取决于词数变化。**
- 词数 = 用户访问词数 * 每次访问的提问数量（默认10） * 每个回答包含的词数（默认50），**词数与用户访问数成正比。**

- **核心假设2-算力需求分配：训练采用的数据占当日新生成数据的1%。**
- 假设依据：根据OpenAI论文Language Models are Few-Shot Learners，GPT-3采用的数据集清洗前大小45TB，清洗后大小570GB，清洗前后存在2个数量级的差距，因此可以认为每天新生成的数据有1%用于训练。

阶段1：ChatGPT+bing日常算力需求

交互计算量	=	参数量	×	词数	×	单个词计算量
1.05E+10T		1.75B		300亿		2（推理）

训练计算量	=	参数量	×	词数	×	单个词计算量
3.14E+8T		1.75B		9.06亿		6（训练）

阶段1假设：

- **每日用户访问量1亿**（根据SimilarWeb统计，2023年1月ChatGPT注册用户1亿，单月访问量6.16亿，月底日访问2800万次；bing日访问约4000万次，二者结合后短期有望迅速增长）

阶段2：LLM+Google日常算力需求

交互计算量	=	参数量	×	词数	×	单个词计算量
5.24E+11T		1.75B		15000亿		2（推理）

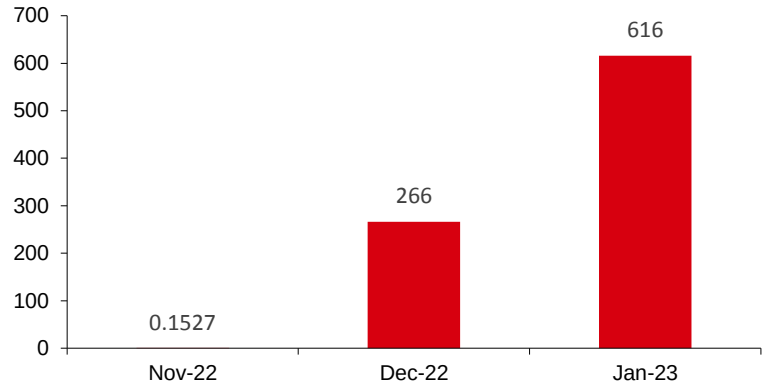
训练计算量	=	参数量	×	词数	×	单个词计算量
1.57E+10T		1.75B		150亿		6（训练）

阶段2假设：

- Bing有望逐渐占据更多市场份额，市场空间参考谷歌，根据SimilarWeb，谷歌月访问量约900亿次，**每日用户访问30亿次。**

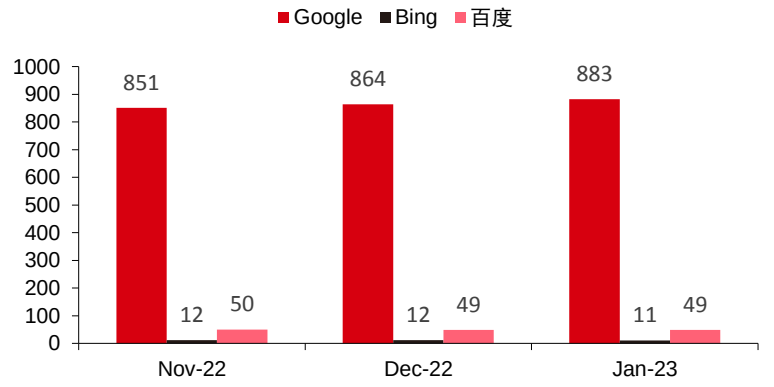
资料来源：SimilarWeb，OpenAI：Language Models are Few-Shot Learners，中信证券研究部

ChatGPT月度访问量（百万次）



资料来源：SimilarWeb，中信证券研究部

Google/Bing/百度月度访问量（亿次）



资料来源：SimilarWeb，中信证券研究部

CONTENTS

目录

2.1 GPU是什么？

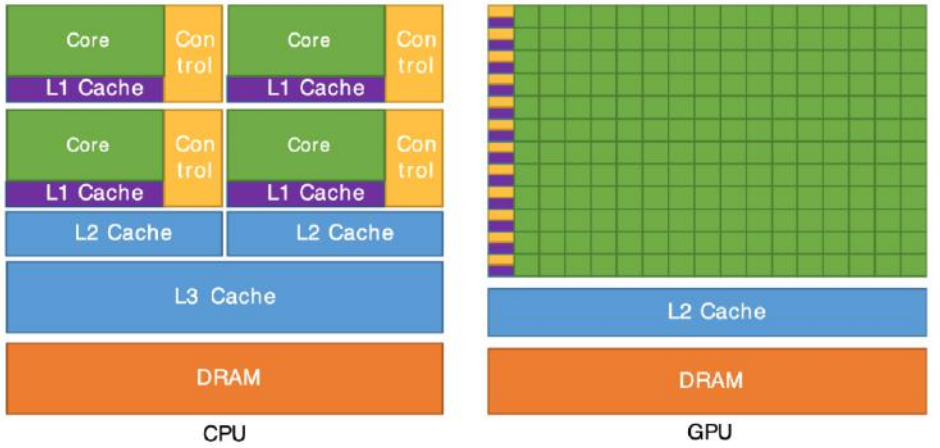
- **GPU（Graphics Processing Unit，图形处理器）：是一种专门在个人电脑、工作站、游戏机和一些移动设备(如平板电脑、智能手机等)上做图像加速和通用计算工作的微处理器。**GPU是英伟达公司在1999年8月发表NVIDIA GeForce 256（GeForce 256）绘图处理芯片时首先提出的概念。
- **GPU应用场景**
  - **图形加速**：此时GPU 内部的顶点渲染、像素渲染以及几何渲染操作都可以通过流处理器完成。
  - **通用计算**：计算通常采用CPU+GPU异构模式，由CPU负责执行复杂逻辑处理和事务处理等不适合数据并行的计算，由GPU负责计算密集型的大规模数据并行计算。
- **GPU 与 CPU 对比**
  - CPU 的逻辑运算单元较少，控制器（Control）和缓存（Cache）占比较大；GPU 的逻辑运算单元小而多，控制器功能简单，缓存也较少。
  - GPU 单个运算单元（ALU）处理能力弱于 CPU，但是数量众多的ALU可以同时工作，当面对高强度并行计算时，其性能要优于 CPU。
  - GPU可以利用多个ALU来做并行计算，而CPU只能按照顺序进行串行计算，同样运行3000次的简单运算，CPU需要3000个时钟周期，而配有3000个ALU的GPU运行只需要1个时钟周期。

GPU的主要分类

类型	应用场景	特点	代表产品
独立GPU	封装在独立的电路板，专用的显存（显示储存器）	性能高，功耗大	NVIDIA Geforce系列 AMD Radeon系列
集成GPU	内嵌到主板上，共享系统内存	性能中等，功耗中等	Intel HD系列 AMD APU系列 苹果M芯片GPU
移动端GPU	嵌在 SoC（System On Chip）中，共享系统内存	性能低，功耗低	Imagination PowerVR系列 高通 Adreon系列 AMD Mali系列 苹果A芯片GPU

资料来源：中信证券研究部

CPU 与 GPU 的芯片资源分布示例



资料来源：《CUDA 编程手册系列第一章：CUDA 简介》——英伟达技术博客

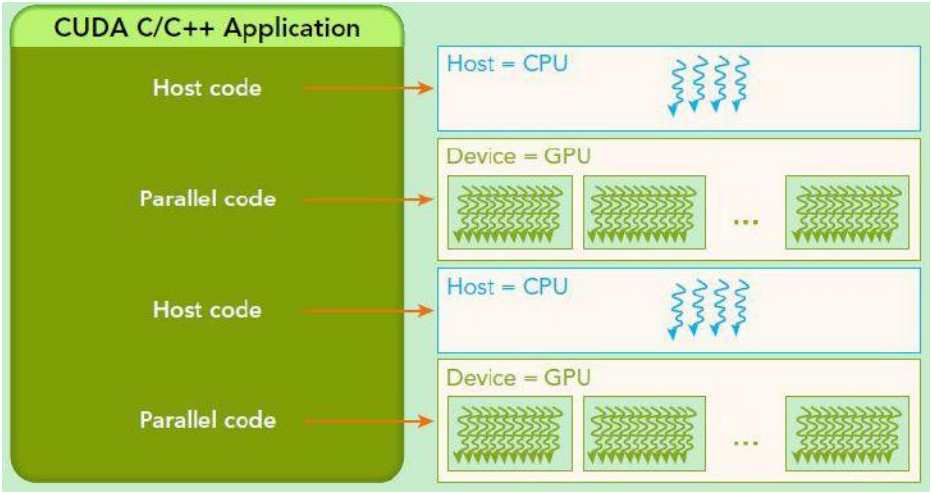
2.2 从GPU到GPGPU的跨越，英伟达CUDA降低开发门槛

- **GPGPU（general-purpose GPU，通用计算图形处理器），利用图形处理器进行非图形渲染的高性能计算。为了进一步专注通用计算，GPGPU去掉或减弱GPU的图形显示部分能力，将其余部分全部投入通用计算，实现处理人工智能、专业计算等加速应用。**
- **2007年6月，NVIDIA推出了CUDA（Computer Unified Device Architecture计算统一设备结构）。**
  - **CUDA是一种将GPU作为数据并行计算设备的软硬件体系。**在CUDA 的架构中，不再像过去GPU架构那样将通用计算映射到图形API中，对于开发者来说，CUDA 的开发门槛大大降低了。
  - CUDA 的编程语言基于标准C，因此任何有C 语言基础的用户都很容易地开发CUDA 的应用程序。由于这些特性，CUDA在推出后迅速发展，被广泛应用于石油勘测、天文计算、流体力学模拟、分子动力学仿真、生物计算、图像处理、音视频编解码等领域。
- **GPU并不是一个独立运行的计算平台，而是需要与CPU协同工作，可以看成是CPU的协处理器。GPU与CPU通过PCIe总线连接在一起来协同工作，因此GPU并行计算实际上指的是基于CPU+GPU的异构计算架构。**

GPGPU的架构与生态

资料来源：《GPU与GPGPU泛淡》—夕阳叹@CSDN

基于CPU+GPU的异构计算应用执行逻辑

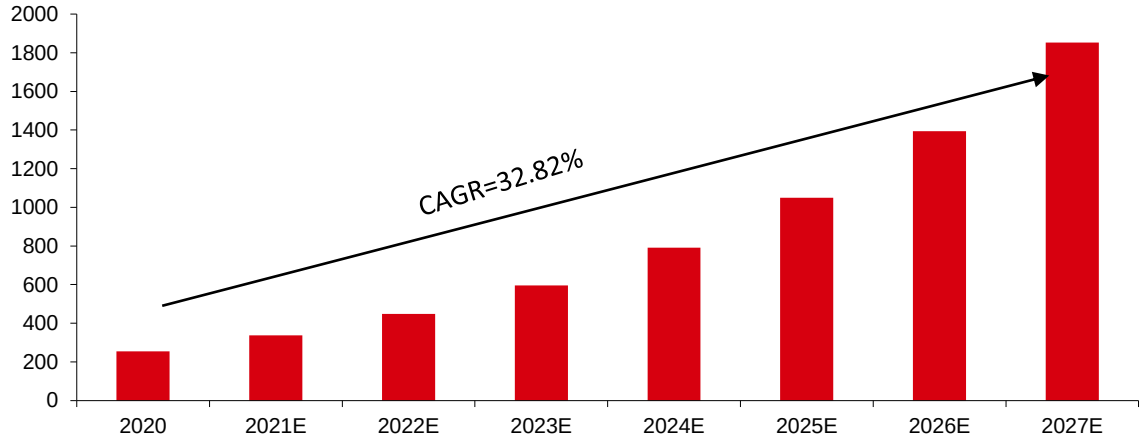


资料来源：Preofessional CUDA® C Programming

2.3 2020年GPU全球市场254亿美元，独显市场英伟达份额约80%

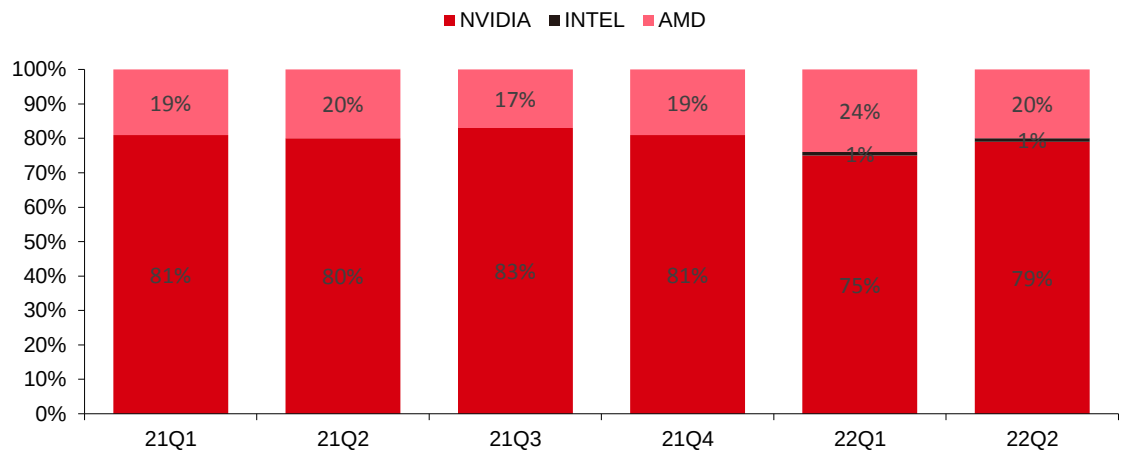
- **根据Verified Market Research数据，2020年，全球GPU市场规模为254.1亿美元（约1717.2亿人民币）。随着需求的不断增长，预计到2028年，这一数据将达到2465.1亿美元（约1.67万亿人民币），年复合增长率为32.82%。**
- **市场研究机构Jon Peddie Research的最新数据显示，2022年二季度，全球独立GPU市场出货量同比增长 2.4% 至 1040 万台，但是较一季度环比则下滑了22.6%。**
  - 从市场份额来看，英伟达**的独立GPU的市场份额从22Q1的75%增加到22Q2的79.6%**，保持了与去年同期相当的份额。AMD和Intel则分别占比20%/1%。
- **据Verified Market Research数据，2020年中国大陆的独立GPU市场规模为47.39亿美元，预计2027年将超过345.57亿美元。**

GPU全球市场规模（亿美元）



资料来源：Verified Market Research（含预测），中信证券研究部

全球独显GPU市场各厂商份额占比

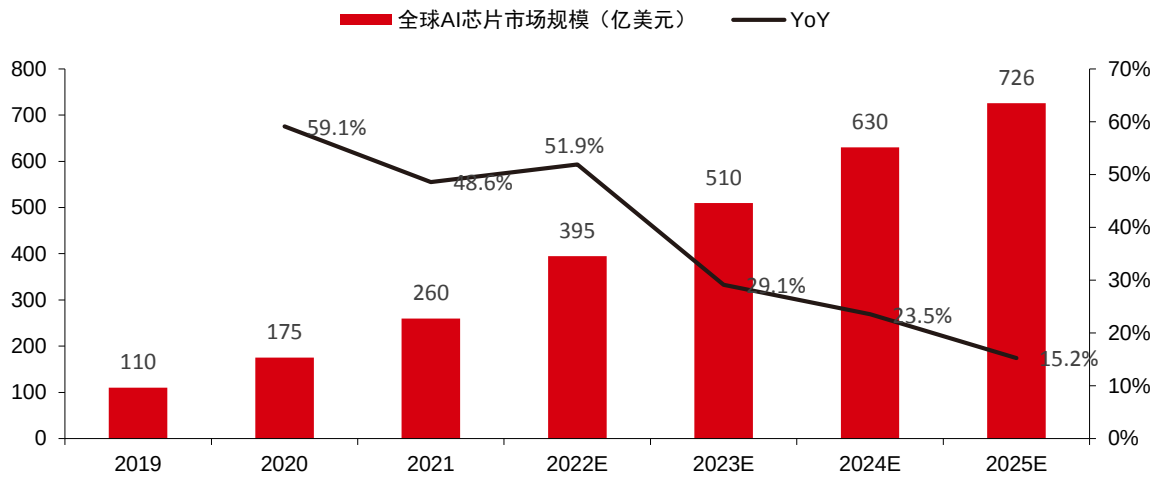


资料来源：Jon Peddie Research，中信证券研究部

2.3 2020年全球AI芯片市场规模约为175亿美元，英伟达份额超80%

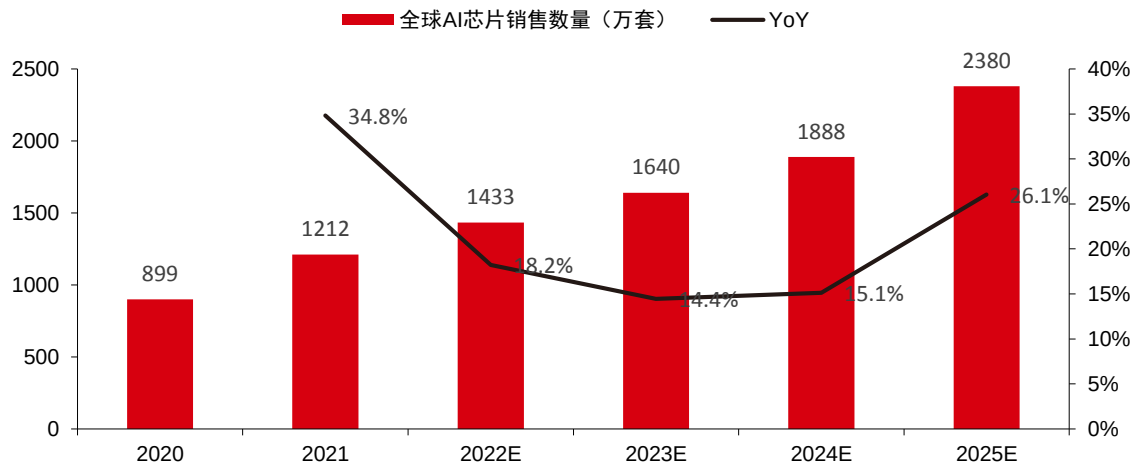
- **伴随着人工智能应用场景的多元化，新算法、新模型不断涌现，模型中的参数数量呈指数级增长，对算力的需求越来越大。OpenAI预估算力需求每3.5个月翻一倍，每年近10倍。**
  - 根据WSTS数据，**2020年全球人工智能芯片市场规模约为175亿美元**。随着人工智能技术日趋成熟，数字化基础设施不断完善，人工智能商业化应用将加落地，推动AI芯片市场高速增长，预计2025年全球人工智能芯片市场规模将达到726亿美元。
  - 未来，随着自动驾驶级别的不断提高，对于AI芯片的需求正不断增长。L2和L3+级汽车都会用AI芯片来取代分立的MCU芯片进行自动驾驶相关的计算工作。WSTS预计AI芯片的数量将从2020年的899万套增长至2025年的2380万套。
  - 据IDC数据，2021年，中国加速卡出货量超过80万片，其中**英伟达占据超过80%市场份额**，此外其他市场参与者还包括AMD、百度、寒武纪、燧原科技、新华三、华为、Intel和赛灵思等。2020年的采购主要集中在搭载**V100、V100S、A100和T4**的加速服务器上，此外英伟达的**A10、A30、A40和Atlas系列**加速卡在部分领域已经开始使用。

全球AI芯片（GPU、FPGA、ASIC等）的市场规模



资料来源：WSTS（含预测），中信证券研究部

全球AI芯片销售数量及预测（万套）

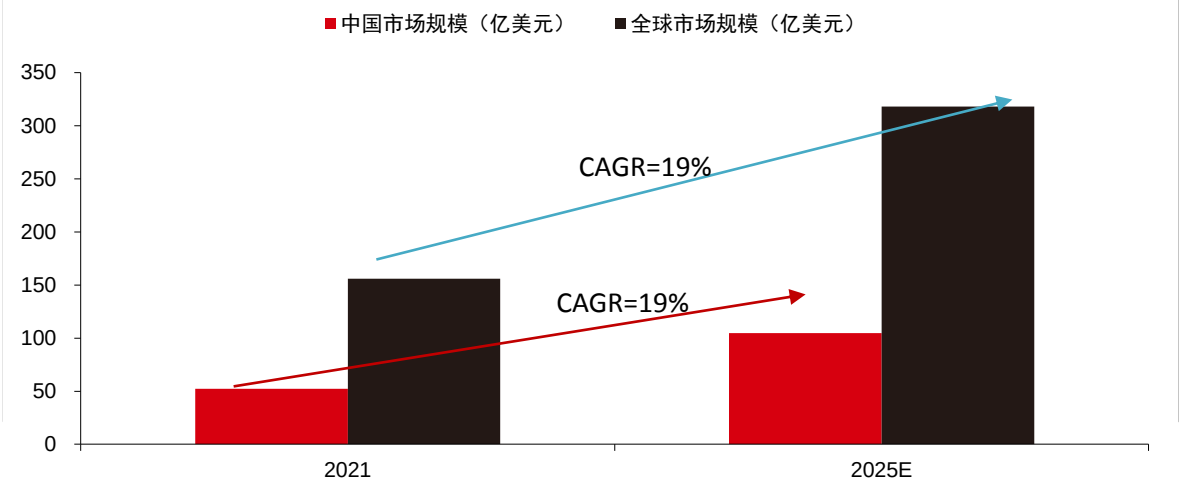


资料来源：WSTS（含预测），中信证券研究部

2.3 中国市场，GPU服务器在AI服务器中占比92%，占主导地位

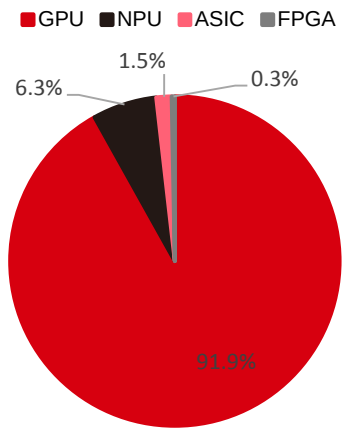
- **据IDC数据，2021年，全球AI服务器市场规模达156亿美元，同比增长39.1%；IDC预测，2025年全球AI服务器市场规模将达317.9亿美元，年复合增长率为19%。**
- **IDC报告显示，2021年中国加速服务器市场规模达到53.9亿美元（约350.3亿人民币），同比+68.6%；预计到2026年将达到103.4亿美元。年复合增长率为19%，占全球整体服务器市场近三成。**
  - 根据IDC数据，2021年，**GPU服务器以91.9%的份额占国内加速服务器市场的主导地位**；NPU、ASIC和FPGA等非GPU加速服务器占比8.1%。IDC预计2024年中国GPU服务器市场规模将达到64亿美元。
  - 从行业的角度看，互联网依然是最大的采购行业，占整体加速服务器市场近60%的份额；**2021年，用于推理工作负载的加速服务器占比已经达到57.6%，预计到2026年将超过60%。**

全球及中国AI服务器市场规模



资料来源：IDC（含预测），中信证券研究部 注：这里统计的AI服务器包括高性能计算

2021年中国AI服务器芯片占比情况

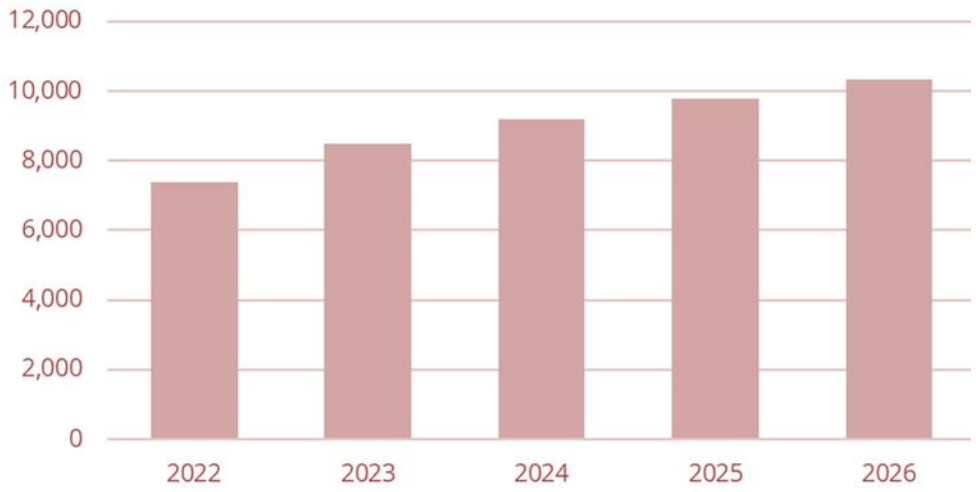


资料来源：IDC，中信证券研究部

2.3 预计2021年中国GPGPU市场规模为149.8亿元，其中AI推理/AI训练/高性能计算分别为93.5/47.1/9.1亿元。

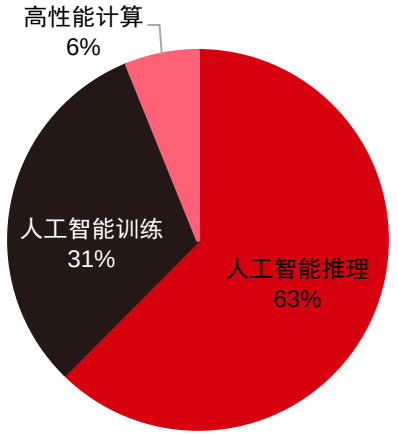
- **市场研究机构Verified Market Research预测，到2025年，中国GPGPU芯片板卡的市场规模将达到458亿元，是2019年86亿元的5倍多，2019-2025年的年复合增长率为32%。其中，**
  - 按行业来分，到2025年，预计互联网及云数据中心需求为228亿元，安防与政府数据中心为142亿元，行业AI应用为37亿元，**高性能计算为28亿元。**
  - 按应用场景来分，到2025年，预计人工智能推理/人工智能训练/**高性能计算**需求分别为286/144/**28亿元，占比分别为62.4%/31.4%/6.1%。**
- **我们预计2021年中国GPGPU市场规模为149.8亿元，其中人工智能推理/人工智能训练/高性能计算分别为93.5/47.1/9.1亿元。**

2022~2026年中国加速计算服务器市场预测（单位：百万美元）



资料来源：IDC预测（2022-2026年均为预测）

GPGPU市场按应用场景拆分



资料来源：Verified Market Research，中信证券研究部

2.4 GPGPU市场英伟达一家独大，全球市场份额约90%

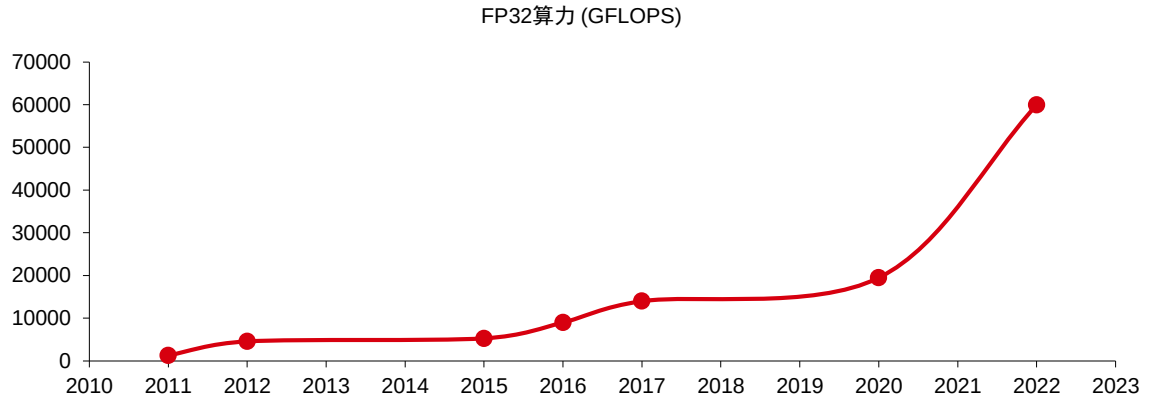
- **GPGPU是一个门槛极高的领域，全球市场基本上被英伟达和AMD两家国际龙头掌控。**
  - 根据 Ark Invest 的数据，2021 年，**英伟达占据了全球数据加速器市场 90% 的份额。**
  - 根据IDC数据，2020年的GPGPU采购主要集中在搭载**V100、V100S、A100和T4**的加速服务器上，此外Nvidia**的 A10、A30、A40和Atlas系列**加速卡在部分领域已经开始使用。2021年，中国加速卡出货量超过80万片，**其中英伟达占据超过80%市场份额。**
  - 根据天数智芯数据，**英伟达在2021年的**中国的**云端AI训练芯片市场份额达到90%**。其中，某一款产品占整个市场的50%，另一款产品占25%。

英伟达历代GPGPU产品的详细信息

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/ST Unit+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108 SMs每个SM 64 个FP32 64 个INT32 32 个FP64 4个Tensor core	H100 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单元拥有更多的逻辑控制电路	NVLink一代，双向互联带宽160GB/s P100有56个SM HBM	Nvlink 2.0 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 MIG 1.0	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 80亿晶体管	16nm 153亿晶体管	12nm 211亿晶体管	12nm 186亿晶体管	7nm 283亿晶体管	4nm 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：智东西

英伟达历代GPGPU产品的FP32算力水平



资料来源：英伟达官网，中信证券研究部

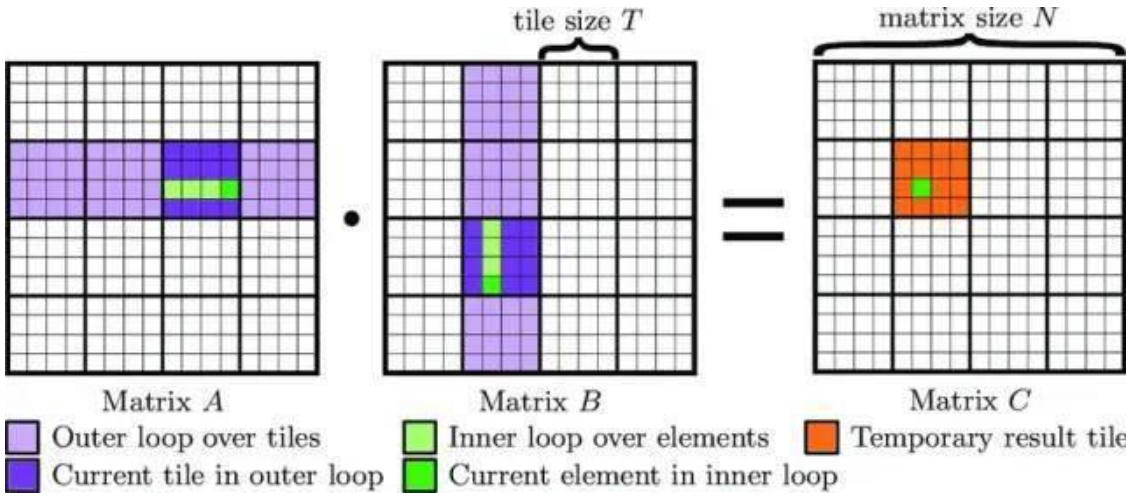
CONTENTS

目录

3.1 壁垒一——高精度浮点计算

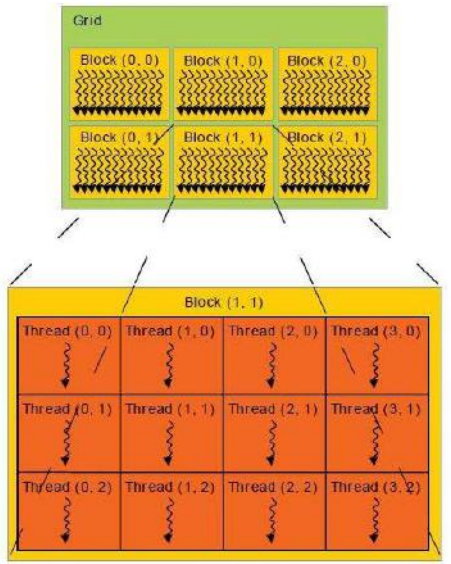
- **CPU是串行处理器，而GPU是并行处理器。**
  - 在机器学习中，绝大多数任务会涉及到耗费时间的大量运算，而且**随着数据集的增加，运算量会越来越大**。解决这个问题的一个方法就是使用多线程并行计算。
  - CUDA 核能够以相对稍慢的速度运行，但是能够通过使用大量运算逻辑单元（ALU）来提供很大的并行度。
- **每个 GPU 核都能运行一个独立的线程，对于矩阵相乘运算来说大大缩短了计算时间。**
  - 对于每个小片的结果可以由一组线程负责，其中每个线程对应小片中的一个元素。**这个线程组将 A 的行小片和 B 的列小片一一载入共享内存，在共享内存上对其做矩阵相乘，然后叠加在原有结果上。**所以对于 2000×2000 的矩阵乘法，只需要 2000 次并行运行。
  - 但是对于CPU来说，因为是串行计算的，所以需要4000000次运行。

矩阵相乘分片算法示意图



资料来源：Matthes, Alexander & Widera, Rene & Zenker, Erik & Worpitz, Benjamin & Huebl, Axel & Bussmann, Michael. (2017). Tuning and optimization for a variety of many-core architectures without changing a single line of implementation code using the Alpaka library.

CUDA线程模型

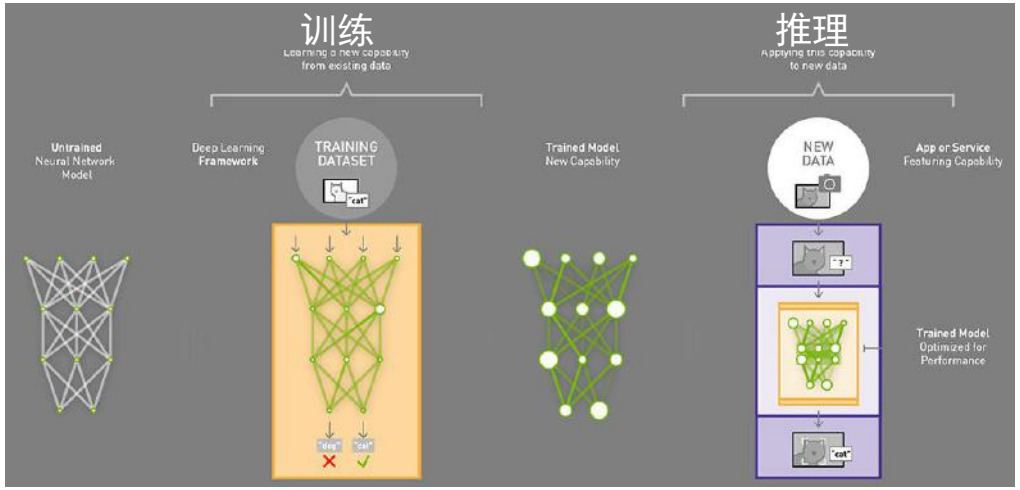


资料来源：《CUDA 轻松入门编程 (一)：CUDA C 编程及 GPU 基本知识》——科技猛兽@极市网站

3.1 人工智能的实现包括两个环节：推理(Inference)和训练(Training)

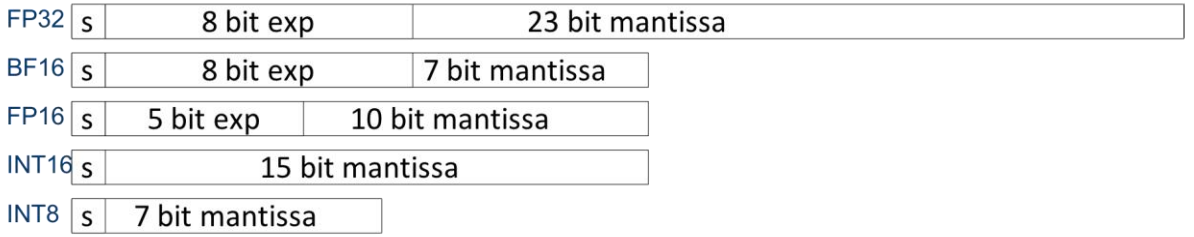
- **训练需要密集的计算得到模型，没有训练，就不可能会有推理。**
  - 训练是指通过大数据训练出一个复杂的神经网络模型，**通过大量标记过的数据来训练相应的系统得到模型，**使其能够适应特定的功能。训练需要较高的计算性能、能够处理海量的数据、具有一定的通用性，以便完成各种各样的学习任务（大数据分析淘宝推荐“你可能感兴趣的产品”模型）。
  - 推理是指利用训练好的模型，使用新数据推理出各种结论。借助神经网络模型进行运算，**利用输入的新数据来一次性获得正确结论的过程**。这也有叫做预测或推断（用户打开手机被推送“可能感兴趣的产品”）。
- **训练需要较高的精度，推理的精度要求较低**
  - 训练的时候因为要保证前后向传播，每次梯度的更新是很微小的，这个时候需要相对较高的精度，一般来说需要float型，如FP32，32位的浮点型来处理数据。
  - 推理对精度的要求没有那么高，可以用低精度，如FP16，也可以用8位的整型（INT8）来做推理，研究结果表明没有特别大的精度损失，但是需要综合考虑功耗、速度等其它问题。

推理是将深度学习训练成果投入使用的过程



资料来源：《NVIDIA DEEP LEARNING INSTITUTE》——英伟达AI Conference

常见的32/16/8位数字格式对比

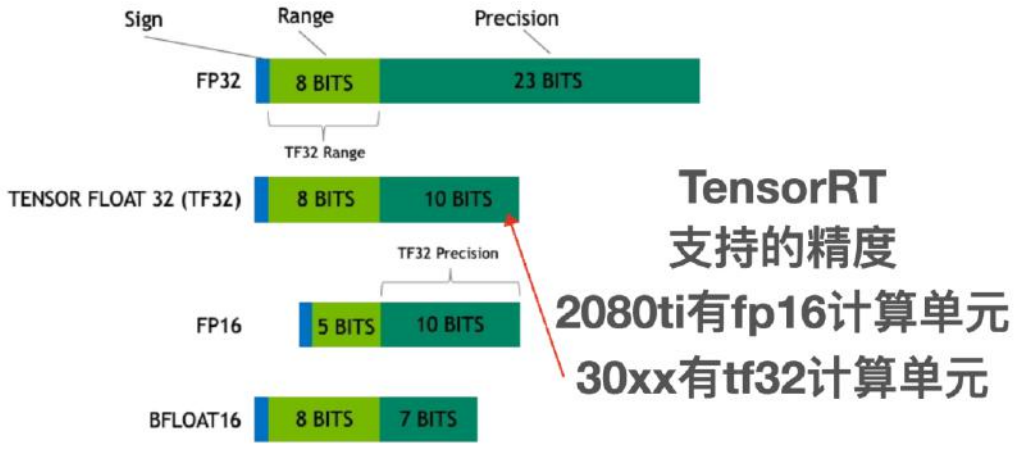


资料来源：《Lower Numerical Precision Deep Learning Inference and Training》——Intel 注：FP32和BF16提供了相同的动态范围，FP32由于更大的尾数提供了更高的精度。

3.1.1 AI训练端：发展目标是精度降低的同时保证模型的准确性

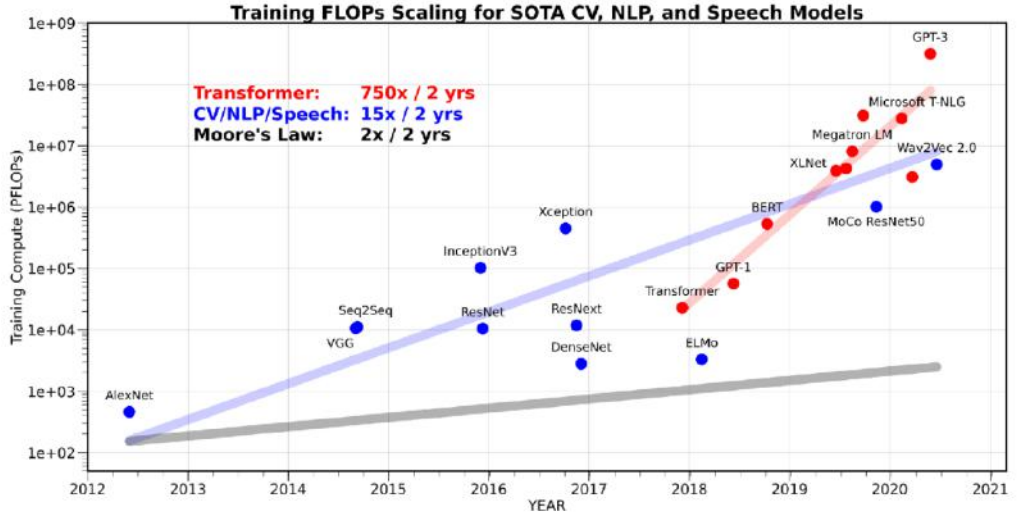
- **浮点计数是利用浮动小数点的方式使用不同长度的二进制来表示一个数字，同样的长度下浮点较整形能表达的数字范围相比定点数更大，结果也更精确**
  - FP64双精度计算：双精度浮点数采用**8个字节也就是64位二进制**来表达一个数字，1位符号，11位指数，52位小数，有效位数为16位。
  - FP32单精度计算：单精度的浮点数中采用**4个字节也就是32位二进制**来表达一个数字，1位符号，8位指数，23位小数，有效位数为7位。
  - FP16半精度计算：半精度浮点数采用**2个字节也就是16位二进制**来表达一个数字， 1位符号、5位指数、10位小数，有效位数为3位。
- **因为采用不同位数的浮点数的表达精度不一样，所以造成的计算误差也不一样。**
  - **对于需要处理的数字范围大而且需要精确计算的科学计算来说，可能需要采用双精度浮点数，**例如：计算化学，分子建模，流体动力学。
  - 对于常见的多媒体和图形处理计算、深度学习、人工智能等领域，32位的单精度浮点计算已经足够了。
  - 对于要求精度更低的**机器学习等一些应用来说，半精度16位浮点数**就可以，甚至8位浮点数就已经够用了。AI计算模型规模的持续扩大，导致模型训练和部署所需求的算力和功耗持续的扩张。**面对算力的挑战，降低精度是一把利器。**

TensorRT支持的计算精度



资料来源：《一起实践神经网络INT8量化系列教程》——老潘的博客@CSDN

AI模型训练算力消耗量与摩尔定律浮点数运算量（单位： FLOPs）

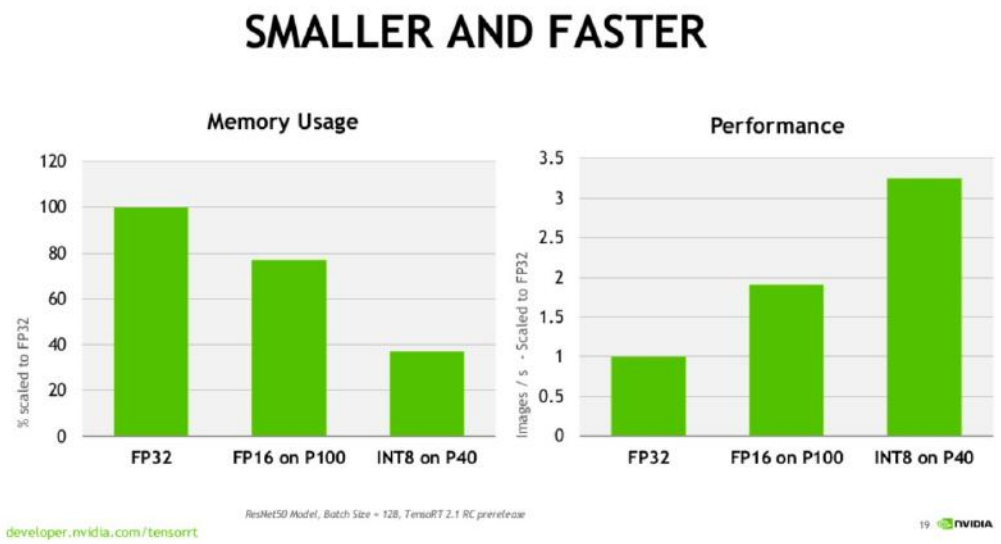


资料来源：《AI算力的阿喀琉斯之踵：内存墙》——Amir Gholami@OneFlow社区 注：蓝线上的是CV，NLP和语音模型，模型运算量平均每两年翻 15 倍，红线上的是 Transformer 的模型，模型运算量平均每两年翻 750 倍。而灰线则标志摩尔定律下内存硬件大小的增长，平均每两年翻 2 倍。

3.1.2 AI推理端：浮点型量化为整形数据，降低算力、加速推理、降低功耗

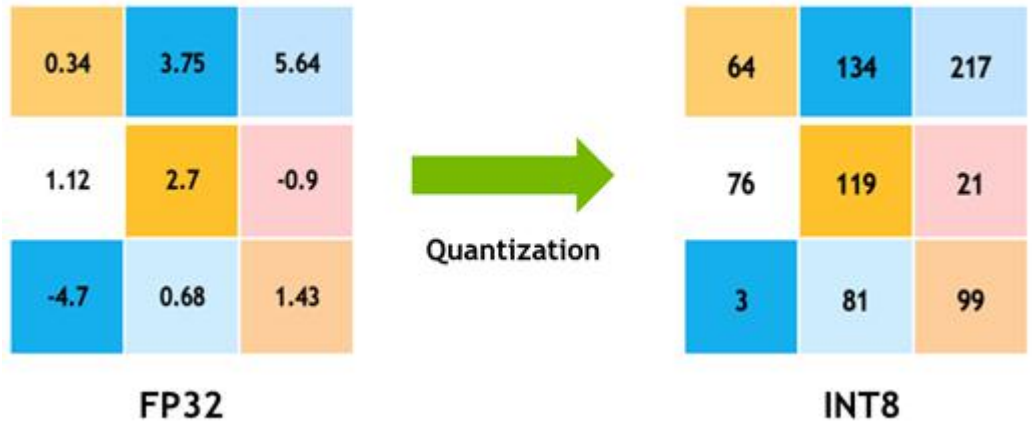
- **量化是通过一组离散符号或整数值去逼近一个连续信号的过程，利用低比特量化(权重或激活)可以在不影响精度的前提下加快推理阶段。随着模型越来越大，需求越来越高，模型的量化自然是少不了的一项技术。**
  - 在低比特表达中（如FP16、INT16、FP8、INT8、INT4等），**INT8因兼顾效率和精度，而被广泛采用**。一方面，INT8的运行速度是FP16/INT16的两倍，并且相比FP8，能被更多的硬件设备支持。另一方面，INT8的量化范围（-128~127）比INT4（-8~7）或其它更低的比特（小于4比特）大，表达能力更强。
- **经过INT8量化后的模型：模型容量变小了，FP32的权重变成INT8，大小直接缩了4倍模型，运行速度可以提升，使用INT8的模型耗电量更少，对于嵌入式侧端设备来说提升巨大。**

INT8有更高的吞吐率、更低的内存要求



资料来源：英伟达—developer.nvidia. com/tensorrt

利用NVIDIA TensorRT 量化感知训练实现INT8 推理的FP32 精度



资料来源：《利用 NVIDIA TensorRT 量化感知训练实现 INT8 推理的 FP32 精度》——英伟达技术博客

3.1.3 GPU中设置各自独立的计算单元，可以针对不同运算优化

- **对于浮点计算来说，CPU可以同时支持不同精度的浮点运算，但在GPU里针对单精度和双精度需要各自独立的计算单元。**
  - 一般在GPU里支持单精度运算的单精度ALU(算术逻辑单元)称之为FP32 core，而把用作双精度运算的双精度ALU称之为DP unit或者FP64 core
  - 在英伟达不同架构不同型号的GPU之间，因为产品定位不同，单精度ALU和双精度ALU的数量的比例差异很大，也决定了产品的定位。

Nvidia Hopper架构中的SMP(流处理块)

资料来源：英伟达，《GPGPU 芯片设计: 原理与实践》——陈巍谈芯@知乎

英伟达不同GPU产品的CUDA计算核数对比

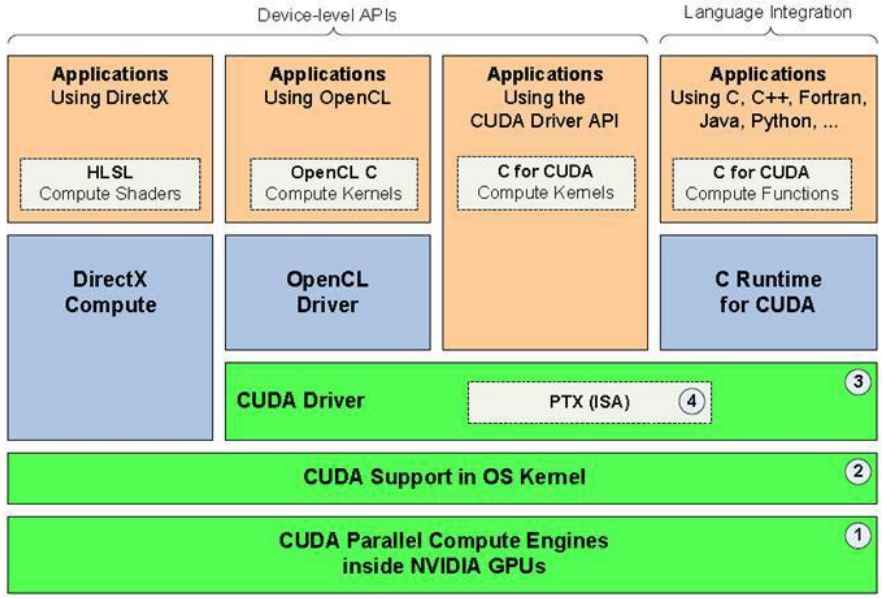
架构代号	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper
中文代号	费米	开普勒	麦克斯韦	帕斯卡	伏特	图灵	安培	赫柏
时间	2010	2012	2014	2016	2017	2018	2020	2022
核心参数	16 个 SM，每个SM包括32 Cuda Cores，共计512 Cuda Cores	15个SMx，每个SMX包括192个单精度+64个双精度的Cuda cores；	16个SMM，每个SM包括4个处理块，每个处理块包括32个CUDA内核+8个LD/ST Unit+8个SFU	Pascal架构有GP100、GP102 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108 SMs每个SM 64 个FP32 64 个INT32 32 个FP64 4个Tensor core	H100 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统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单元拥有更多的逻辑控制电路	NVLink一代，双向互联带宽160GB/s P100有56个SM HBM	Nvlink 2.0 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 MIG 1.0	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 80亿晶体管	16nm 153亿晶体管	12nm 211亿晶体管	12nm 186亿晶体管	7nm 283亿晶体管	4nm 800 亿个晶体管
代表型号	Quadro 7000	K80 K40M	M5000 M4000	P100 GTX 1080 P6000	V100 TiTan V	T4 2080TI RTX 5000	A100、A30 3090	H100

资料来源：智东西微信公众号

3.2 壁垒二——CUDA生态：使 GPU 解决复杂计算问题，基于此开发数千个应用

- **CUDA(Compute Unified Device Architecture，统一计算设备架构)是由 NVIDIA 于2007年推出的通用并行计算架构，专为图形处理单元 (GPU) 上的通用计算开发的并行计算平台和编程模型。借助 CUDA，开发者能够利用 GPU 的强大性能显著加速计算应用。**
  - 它包含了 CUDA 指令集架构（ISA）以及 GPU 内部的并行计算引擎。CUDA 是一个全新的软硬件架构，可以将 GPU 视为一个并行数据计算的设备，**对所进行的计算进行分配和管理，无需将其映射到图形 API（OpenGL和Direct 3D）中运行。**
  - 使用 CUDA 时，开发者使用主流语言（如 C、C++、Fortran、Python 和 MATLAB）进行编程，并通过扩展程序以几个基本关键字的形式来表示并行性。
  - NVIDIA 的 CUDA 工具包提供了开发 GPU 加速应用所需的一切。CUDA 工具包中包含多个 GPU 加速库、一个编译器、多种开发工具以及 CUDA 运行环境。**通过 CUDA 开发的数千个应用已部署到嵌入式系统、工作站、数据中心和云中的 GPU。**

CUDA 架构的组件组成



资料来源：《Nvidia 并行计算架构 CUDA 分析(一)——CUDA 简介》——club111@阿里云

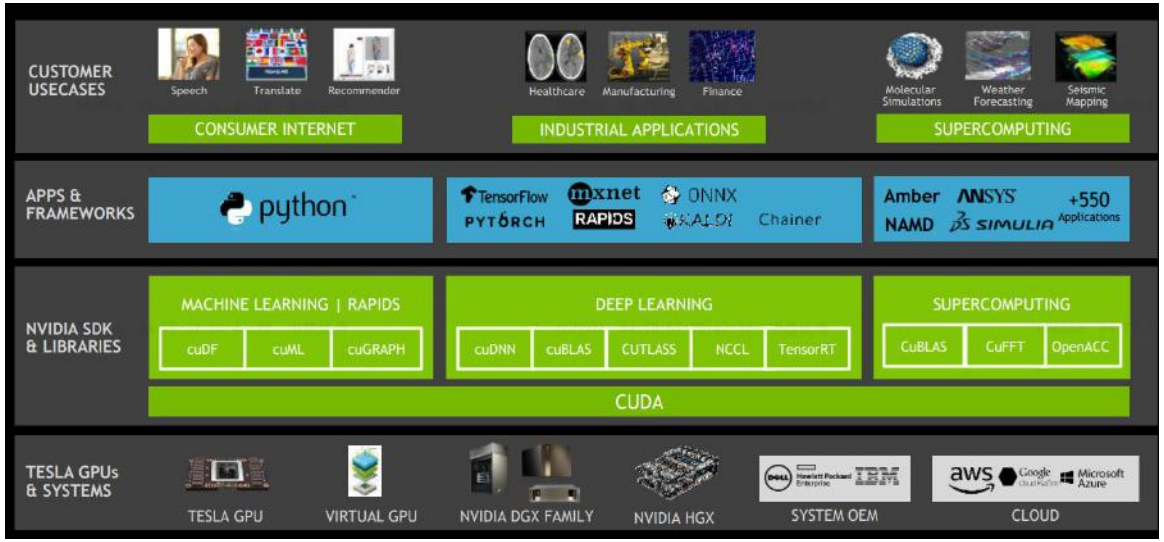
通过 CUDA 开发的部分应用

资料来源：英伟达开发者社区

3.2.1 CUDA：一家独大，助力英伟达GPU生态建设，软硬件深度绑定

- **易于编程和性能飞跃，加上拥有广泛而丰富的生态系统，CUDA让NVIDIA的GPU生态圈迅速成型。**
  - 在2006年问世之初，英伟达就开始对CUDA系统在AI领域进行大力投入和推广。一方面在年营业额只有30亿美元的情况下，每年投入5亿美元的研发经费更新维护；另一方面，为当时美国大学及科研机构免费提供CUDA系统，使其迅速在AI及通用计算领域开花结果。
  - 与任何新平台一样，CUDA的成功依赖于CUDA生态系统可用的工具、库、应用程序和合作伙伴。CUDA 支持 Windows、Linux、MacOS 三种主流操作系统，支持 CUDA C 语言和 OpenCL 及 CUDA Fortran 语言。无论使用何种语言或接口，**指令最终都会被驱动程序转换成 PTX（Parallel Thread Execution，并行线程执行，CUDA架构中的指令集，类似于汇编语言）代码，交由GPU计算。**
- **但是，只有NVIDIA的GPUs才支持CUDA技术，NVIDIA凭借CUDA在科学计算、生物、金融等领域的推广牢牢把握着主流市场。**

CUDA生态支持的主要应用场景



资料来源：《CUDA new features and beyond》——英伟达

CUDA 11 中的平台支撑

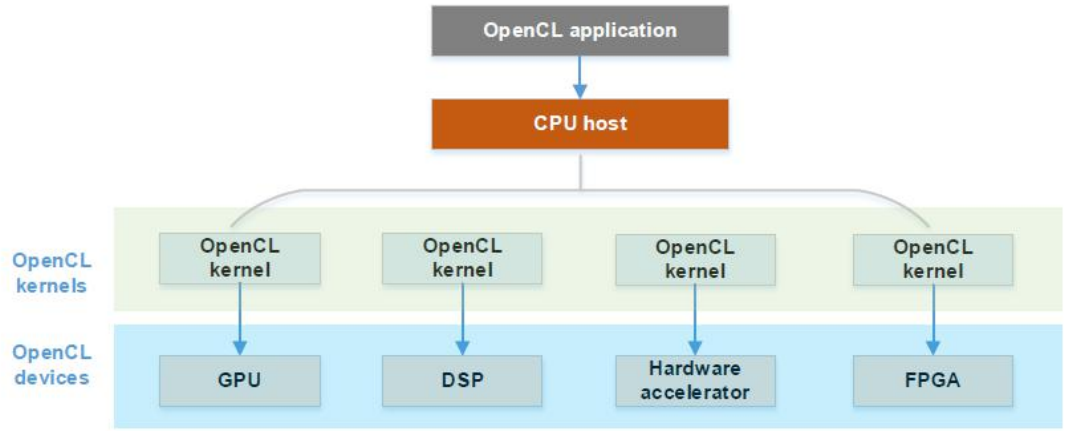
PLATFORM	OS	VERSION	COMPILERS
Linux		18.04.4 LTS 16.04.6 LTS	GCC 9.x PGI 20.x Clang 9.0.x ICC 19.1 XLC 16.1.x (POWER) Arm C/C++ 19.2 (Arm64)
	CentOS	7.7 8.1	
	SUSE	SLES 15.1	
		Leap 15.1	
Windows	Windows Windows Server	10, 2019, 2016	Microsoft Visual Studio 2019 (16.x)

资料来源：《CUDA 11 功能揭晓》——Pramod Ramarao@英伟达社区

3.2.2 OpenCL：开源计算框架，兼容各类硬件设备用于并行计算

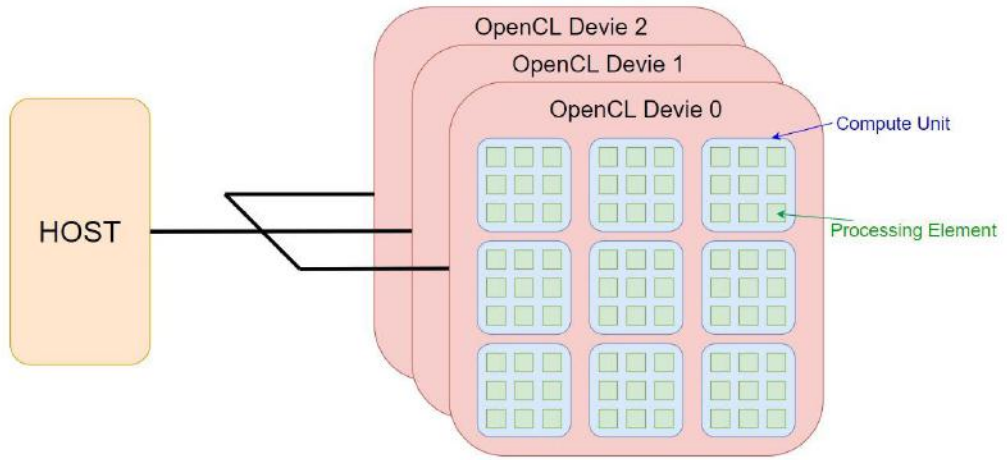
- **由于各个硬件厂家在 GPU 硬件设计上存在着较大差别，为了降低跨平台的开发难度，需要一套能够兼容各类硬件设备的计算框架。**
  - OpenCL 最初由苹果公司开发，拥有其商标权。2008 年，苹果公司向 Khronos Group 提交了一份关于跨平台计算框架 (OpenCL) 的草案，随后与 AMD、IBM、Intel、和 NVIDIA 公司合作逐步完善，其接口大量借鉴了 CUDA。
  - 后续，OpenCL 的管理权移交给了非盈利组织 Khronos Group，且于2008年12月发布了 OpenCL 1.0。最新的OpenCL 3.0 于 2020 年 9 月发布。
- **OpenCL是一个为异构平台（CPU/GPU/DSP/FPGA等）编程设计的框架，是一个面向异构系统通用目的并行编程的开放式、免费标准，也是一个统一的编程环境，便于软件开发人员为高性能计算服务器、桌面计算系统、手持设备编写高效轻便的代码，只要按照标准实现了驱动的硬件，使用OPENCL加速的应用原则上就都能使用，主要用于并行运算。**
- **在 OpenCL 中，首先需要一个主机处理器（Host），一般是 CPU。而其他的硬件处理器（多核CPU/GPU/DSP 等）被抽象成 OpenCL 设备（Device）。每个设备包含多个计算单元（Compute Unit），每个计算单元又包含多个处理单元（Processing Element）。在执行中，主要的流程为 Host 端发送数据和任务给 Device 端，Device 端进行计算，最后在 Host 端进行同步。**

OpenCL - 异构计算框架



资料来源：《移动端算法优化》——高性能计算学院@知乎专栏

OpenCL 平台模型图示

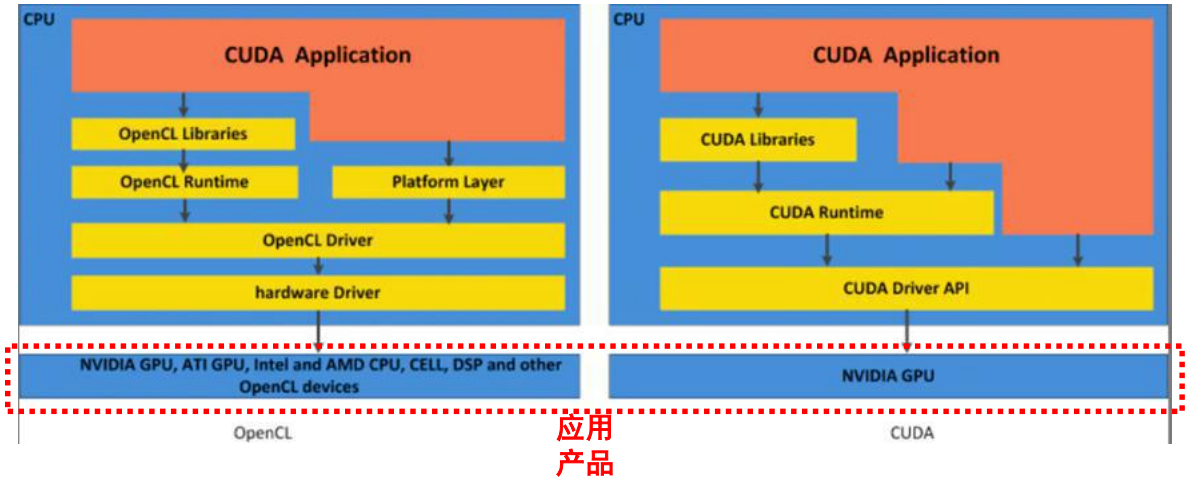


资料来源：《移动端算法优化》——高性能计算学院@知乎专栏

3.2.2 OpenCL对比CUDA：简便性差、市占率低，通用性强

- **OpenCL在应用层次、简便性、市场占有率方面都要弱于CUDA，但是在跨平台和通用性上优于CUDA。**
  - **开发者友好程度**：CUDA在这方面显然受更多开发者青睐。原因在于其统一的开发套件(CUDA Toolkit, NVIDIA GPU Computing SDK以及NSight等等)、丰富的库(cuFFT, cuBLAS, cuSPARSE, cuRAND, NPP, Thrust)以及NVCC(NVIDIA的CUDA编译器)所具备的PTX代码生成、离线编译等更成熟的编译器特性。**相比之下，使用OpenCL进行开发，只有AMD对OpenCL的驱动相对成熟。**
  - **跨平台性和通用性：OpenCL支持包括ATI, NVIDIA, Intel, ARM在内的多类处理器，**CPU、显卡、FPGA、DSP等等都可能可以用OpenCL开发；并能支持运行在CPU的并行代码，同时还独有Task-Parallel Execution Mode，能够更好的支持异构计算。这一点是仅仅支持数据级并行并仅能在NVIDIA众核处理器上运行的CUDA无法做到的。
  - **市场占有率：**作为一个开放标准，**缺少背后公司的推动，OpenCL没有占据通用并行计算的主流市场**。NVIDIA则凭借CUDA在科学计算、生物、金融等领域的推广牢牢把握着主流市场。

OpenCL和CUDA的应用框架



资料来源：《OpenCL与Cuda技术》——吴建明@博客园

OpenCL和CUDA产品对比

	CUDA	OpenCL
技术类型	控制	开源和VIP服务
出现时间	2006年	2008年
SDK企业	NVIDIA	具体根据企业
SDK是否免费	Yes	依赖企业
实现企业	仅NVIDIA	Apple、NVIDIA、AMD、IBM
支持系统	Windows, Linux, Mac OS X; 32 and 64-bit	依赖具体企业
支持设备类型	仅NVIDIA GPU	多种类型
支持嵌入式设备	NO	Yes

资料来源：《CUDA与OpenCL之间的差异》—AI领域@简书网站

3.2.3 其他生态：AMD和Intel都推出自主生态，但都无法摆脱CUDA

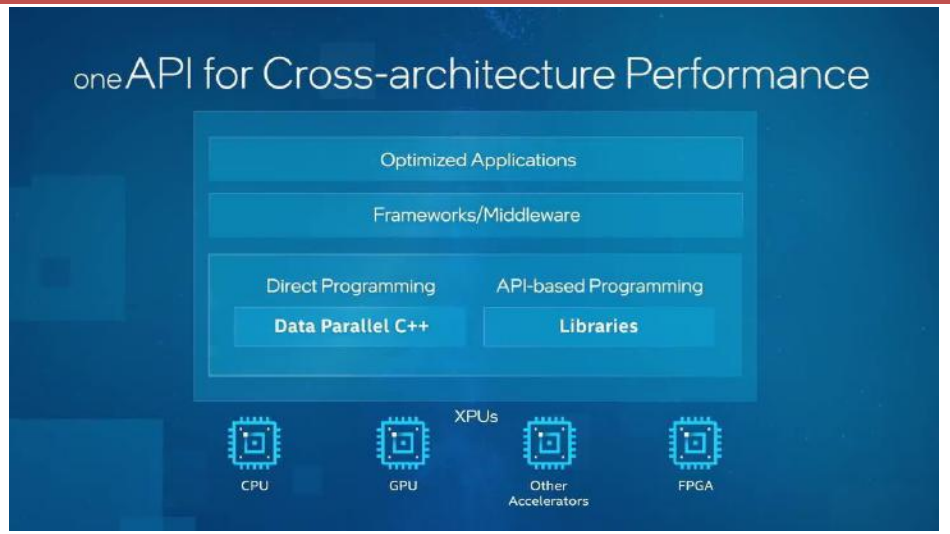
- **AMD推出了ROCm开发环境，目的是建立可替代CUDA的生态，并在源码级别上对CUDA程序支持**
  - A卡上编程模型（硬件生态）使用的是HIP，而运行环境（软件生态）是ROCm，此外AMD发布GPUFORT将CUDA应用转换；N卡上，编程模型是CUDA，运行环境也是CUDA。
  - AMD收购赛灵思后，公司拥有AMD CPU + AMD GPU + FPGA + Xilinx SmartNIC。除了硬件外，AMD的Radeon Open Compute (ROCm)混合CPU-GPU开发环境，再加上赛灵思Vitis，足以对抗英伟达颇受欢迎的CUDA开发平台，以及英特尔力推的oneAPI。
- **英特尔也推出了one API，意在打造跨行业的开放软件生态。**
  - Intel one API是一个跨行业、开放、标准统一、简化的编程模型，旨在促进社区和行业合作、简化跨多架构的开发过程、解决跨体系及供应商代码重用，为跨 CPU、GPU、FPGA、专用加速器的开发者提供统一的开发体验。包括了oneAPI标准组件如直接编程工具、含有一系列性能库的基于API的编程工具，以及先进的分析、调试工具等组件。
- **目前对于AMD和Intel，解决应用问题都是通过工具帮助将 CUDA 代码转换成自己的编程模型，从而能够针对 CUDA 环境的代码编译。**

AMD推出的ROCm与英伟达CUDA的对比

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	

资料来源：《ROCm平台及HIP介绍》——Charles Ren@CSDN

Intel one API

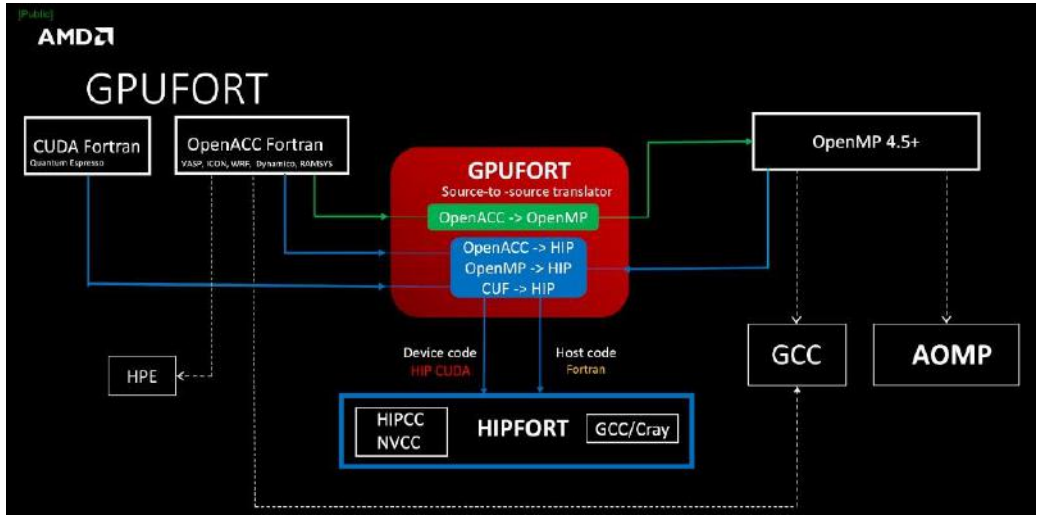


资料来源：Intel官网

3.2.4 国内厂商目前多采用指令翻译兼容CUDA，同时也在构建自主生态

- **国内厂商：多采用指令翻译兼容CUDA及ROCm生态，同时也在构建自主生态。**
  - 壁仞目前兼容主流的GPU生态（CUDA），与客户现有的基础设施做到高度的兼容，方便客户的迁移。也推出了自主的BIRENSUPA软件平台和编程模型，该平台构建在BR100系列产品的底层硬件之上，由驱动层、编程平台、框架层、应用解决方案构成，支持各类应用场景。
  - 沐曦专注研发全兼容CUDA及ROCm生态的国产高性能GPU芯片，满足HPC、数据中心及AI等方面的计算需求。
  - 海光DCU协处理器全面兼容ROCm GPU计算生态，由于ROCm和CUDA在生态、编程环境等方面具有高度的相似性，CUDA用户可以以较低代价快速迁移至ROCm平台。
  - 天数智芯GPGPU计算芯片主要针对云端AI训练+推理和云端通用计算，是国内量产的唯一兼容CUDA等异构计算生态的数据中心高端计算芯片。
- **由于CUDA的闭源特性，以及快速的更新，后来者很难通过指令翻译等方式完美兼容，即使部分兼容也会有较大的性能损失，导致在性价比上持续落后NVIDIA。另一方面，CUDA毕竟是NVIDIA的专属软件栈，包含了许多NVIDIA GPU硬件的专有特性，这部分在其他厂商的芯片上并不能得到体现。因此对于国内厂商来说，还是需要构建自主的软硬件生态。**

AMD ROCm兼容CUDA的方案



资料来源：AMD官网

壁仞BIRENSUPA可实现现有GPU代码平滑迁移

资料来源：壁仞科技发布会

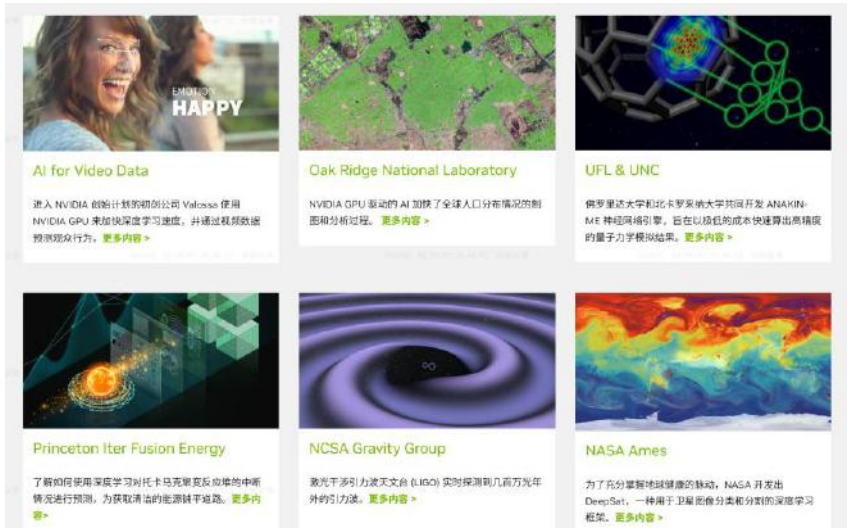
CONTENTS

目录

4.1 GPGPU在计算领域应用：AI计算和高性能计算

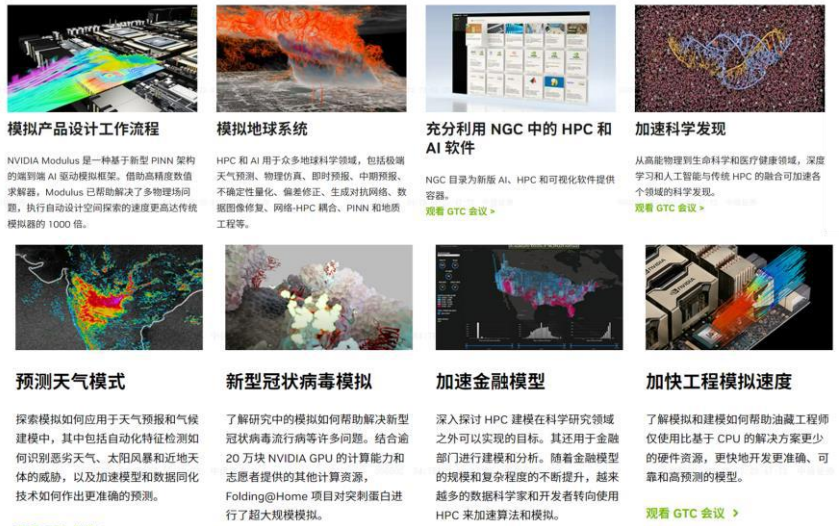
- **GPU在通用计算领域分为两种应用场景，人工智能 (AI) 计算和高性能计算（HPC）**
  - **AI所需的计算力不需要太高精度。**一些AI应用需要处理的对象是语音、图片或视频，运行低精度计算甚至整型计算即可完成推理或训练。
    - 智能计算机是一种专用算力，它们在推理或训练等智能计算方面的确表现出色，**但由于AI推理或训练一般仅用到单精度甚至半精度计算、整型计算**，多数智能计算机并不具备高精度数值计算能力，这也限制其在AI计算之外的应用场景使用。
    - 英伟达新推出的H100芯片搭载Transformer 引擎，使用每层统计分析来确定模型每一层的最佳精度（FP16 或 FP8），在保持模型精度的同时实现最佳性能，相较于上一代产品提供 9 倍的训练吞吐量，性能提升6倍。
  - **高性能计算是一种通用算力，设计目标是提供完备、复杂的计算能力，在高精度计算上能力更强。**应用场景包括行星模拟、分子药物设计等。
    - **超级计算机主要测试的是双精度浮点运算能力(FP64)。**对比单精度(32位，FP32)、半精度(16位，FP16)以及整数类型(如INT8、INT4)等，**数字位数越高，意味着人们可以在更大范围内的数值内体现0/1两个数值的变化，从而实现更精确计算。**

NVIDIA的AI计算应用场景



资料来源：英伟达官网

NVIDIA的高性能计算应用场景

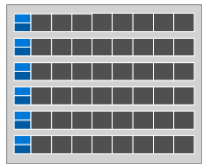


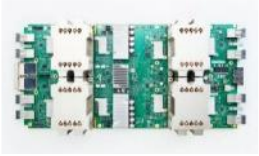
资料来源：英伟达官网

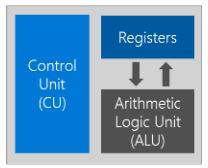
4.1 应用场景一——AI计算

- **根据部署的位置不同，AI芯片可以分为：云端AI芯片、终端AI芯片。**
  - 云端，即数据中心，在深度学习的训练阶段需要极大的数据量和大运算量，**单一处理器无法独立完成，因此训练环节只能在云端实现。**
  - 终端，即手机、安防摄像头、汽车、智能家居设备、各种IoT设备等执行边缘计算的智能设备。**终端的数量庞大，而且需求差异较大。**
- **根据承担任务的不同，AI芯片可以分为：用于构建神经网络模型的训练芯片，利用神经网络模型进行推理预测的推理芯片。**
  - 训练，是指通过大数据训练出一个复杂的神经网络模型，即用大量标记过的数据来“训练”相应的系统，使之可以适应特定的功能。**训练需要极高的计算性能，需要较高的精度，训练芯片受算力约束，一般只在云端部署。**
  - 推理，是指利用训练好的模型，使用新数据推理出各种结论。即**借助现有神经网络模型进行运算，利用新的输入数据来一次性获得正确结论的过程，在云端和终端均有部署。**

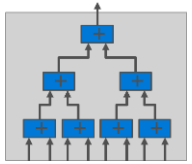
AI芯片的分类和AI应用场景

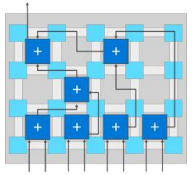






训练端	推理端
GPU：以英伟达为主，AMD为辅，标榜通用性，多维计算及大规模并行计算架构契合深度学习的需要。在深度学习上游训练端（主要用在云计算数据中心里），GPU是当仁不让的第一选择。	**GPU**：英伟达从18年开始通过T4芯片等布局推理端到边缘计算。深度学习下游推理端则更重视低功耗和低延迟，对算力的要求较低，在市场蛋糕变大的同时，逐步形成GPU向推理端渗透，与ASIC和FPGA共同繁荣发展的格局。
ASIC：以谷歌的TPU为代表，包括英特尔、寒武纪、亚马逊、华为等公司均在自行研发。针对特定框架进行深度优化定制。但开发周期较长，通用性较低。比特币挖矿目前使用ASIC专门定制化矿机。	**ASIC**：下游推理端更接近边缘设备，需求也更加细分，英伟达的DLA，寒武纪的NPU、地平线的旭日和征程系列、华为昇腾系列等逐步面市，将依靠特定优化和效能优势，未来在深度学习领域分一杯羹。
CPU：通用性强，但难以适应于人工智能时代大数据并行计算工作。	**FPGA**：依靠可编程性及电路级别的通用性，适用于开发周期较短的IoT产品、传感器数据预处理工作以及小型开发试错升级迭代阶段等。但较成熟的量产设备多采用ASIC。



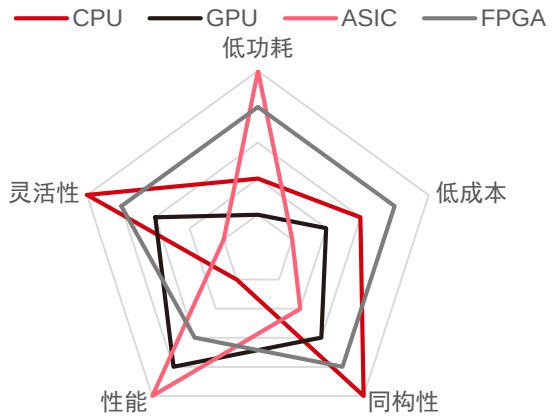


资料来源：英伟达、谷歌官网等，中信证券研究部

4.1.1 AI芯片的三种较为主流的技术路线——GPU、FPGA和ASIC

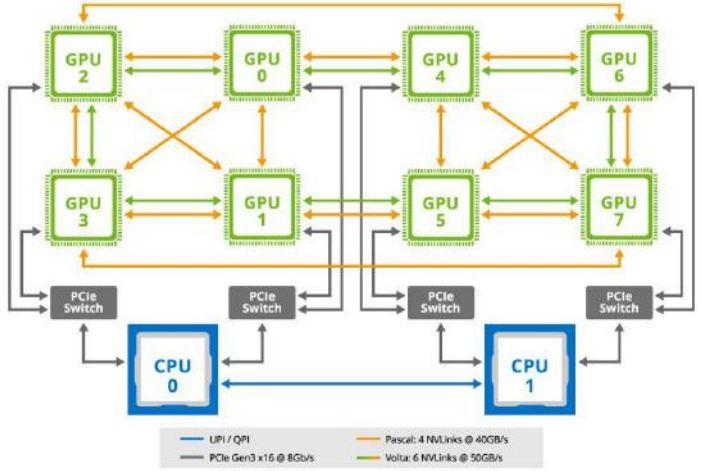
- **AI芯片被称为AI加速器或计算卡，即专门用于加速AI应用中的大量计算任务的模块（其他非计算任务仍由CPU负责），面向AI计算应用的芯片都可以称为AI芯片，包括GPU、FPGA、ASIC等。**
- **因为CPU是图灵完备的，可以自主运行，因此，存在基于多核CPU组成的CPU芯片是同构并行的。但是，GPU、FPGA、DSA、ASIC等处理引擎/芯片是非图灵完备的，都是作为CPU的加速器而存在。因此，其他处理引擎的并行计算系统即为CPU+xPU的异构并行，大体分为三类：**
  - CPU+GPU。**CPU+GPU是目前最流行的异构计算系统，在HPC、图形图像处理以及AI训练/推理等场景得到广泛应用。**
  - CPU+FPGA。目前数据中心流行的FaaS服务，**目前FPGA通常以加速卡的形式配合现有的CPU进行大规模部署**。FPGA的功耗通常为几十瓦，对额外的供电和散热等环节没有特殊要求，因此可以兼容数据中心的现有硬件基础设施。
  - CPU+DSA。谷歌TPU是第一个DSA架构处理器，TPUv1采取独立加速器的方式，实现CPU+DSA（TPU）的方式实现异构并行。
  - 由于ASIC功能固定，缺乏一定的灵活适应能力，因此**不存在CPU+单个ASIC的异构计算。CPU+ASIC形态通常是CPU+多个ASIC组**，或在SOC中，作为一个逻辑上独立的异构子系统存在的，需要与其他子系统协同工作。

CPU、GPU、FPGA和ASIC的特性对比



资料来源：老石谈芯微信公众号，中信证券研究部

典型的用于机器学习场景的GPU服务器主板DGX-1拓扑结构

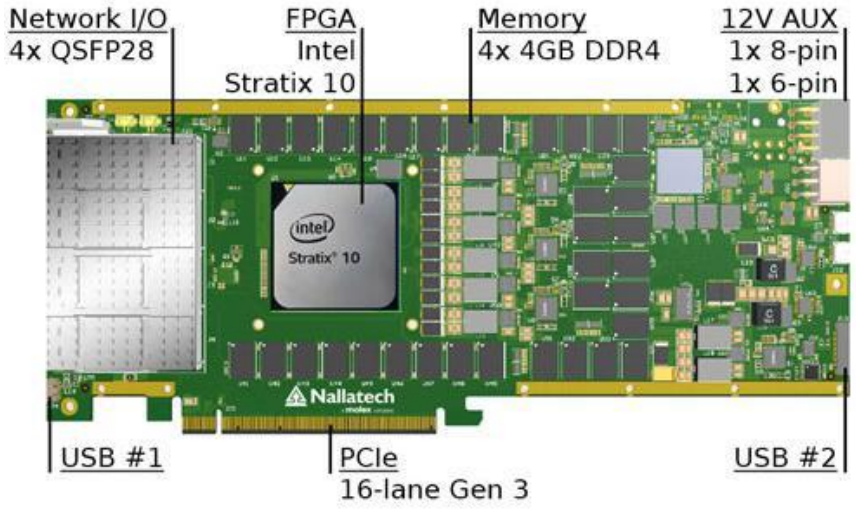


资料来源：英伟达，《DGX系列服务器》——Arcing@博客园

4.1.2 FPGA更适合处理多指令流单数据流，从而适应于推理阶段

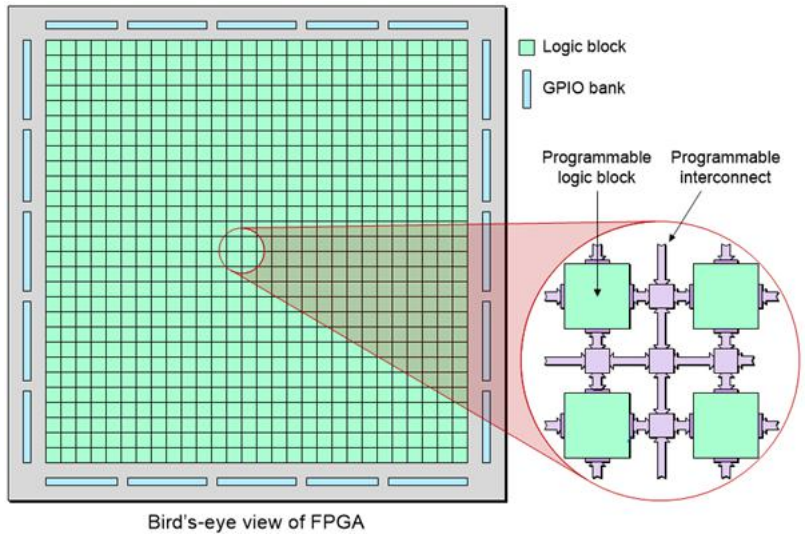
- **FPGA没有极致的性能特点与量产单价高是其未来发展的瓶颈，更适合用于细分、快速变化的垂直行业，应用面上较为狭窄。**
  - **优点：**1. 突破冯诺依曼结构，可直接实现算法，没有指令译码和解读的过程，**功效能耗比是CPU的10倍以上、GPU的3倍**，处理速度和效率要高于GPU。2. 可编译，灵活性很高，开发周期短。FPGA具有可编辑性，用户可以根据自身需求实现芯片功能的转换。基于FPGA灵活编译的特点，**其开发周期较短，上市速度快**。**FPGA更适合处理多指令流单数据流，从而适应于推理阶段。**
  - **缺点：**1. 价格较高，规模量产后的单价更是远高于ASIC。目前FPGA的造价相比GPU更为高昂，如果规模量产后，其不像ASIC可以分摊固定成本，存在单个芯片的编译成本，所以单价远高于ASIC。2. **计算能力和峰值性能不如GPU。** 3. 灵活性占优的同时牺牲了速度与能耗。效率和功耗上劣于专用芯片ASIC。4. FPGA的语言技术门槛较高。目前FPGA的设置要求用户用硬件描述语言对其进行编程，需要专业的硬件知识，具有较高的技术门槛。
  - **FPGA应用于硬件平台加速、数据中心和云端深度学习预测**。FPGA兼具较高的性能和灵活性，加上低能耗的特点，适用于硬件平台的加速。比如微软开发了带有FPGA芯片的主板来提升Bing数据中心的整体性能，**相比于传统CPU在处理Bing的自定义算法时快出40倍。**

Nallatech的 FPGA 加速器方案，采用了英特尔Stratix 10芯片



资料来源：《BittWare Nallatech 520N Network Acceleration Card》——Nallatech公司

FPGA内部结构图



资料来源：《What is FPGA—How Does it Work and its Uses》——Lattice官网

4.1.2 ASIC芯片可以用作人工智能平台训练、推理的芯片

- **ASIC效率高、功耗比佳，但量产前成本高，适用智能终端和AI训练和推理平台。**
  - **优点**：1. **性能上的优势非常明显，具有最高的功效能耗比。**ASIC是专业AI芯片，相比GPU和FPGA没有多余的面积或架构设计，可以实现最快的通信效率与计算速度，实现最低的能耗。2. **下游需求促进人工智能芯片专用化**。随着人工智能的发展和下游智能终端的普及，AI芯片需求大幅上升，而出于对信息隐私保护和云端计算需要联网的考虑，完全依赖云端是不现实的，需要有要有本地的软硬件基础平台支撑，所以专有化的AI芯片有很大的优势。
  - **缺点**：1. 造价昂贵，需要保证量产才能降低成本。2. **一种算法只能应对一种应用；一颗AI芯片只能单一地解决一种问题；**而算法在不断演变，每3~6个月就可能变一次；ASIC芯片或许尚未上市，算法就已经发生进化了。
  - ASIC芯片应用于**人工智能平台和智能终端**。ASIC芯片由于其定制化的特点，具有功能的多样性，应用非常广泛。高性能和低功效使其不再局限于深度学习的训练或推理阶段的其中之一，而是**可以作为支撑人工智能平台全阶段加速的芯片。**

谷歌TPU的发展历史及算力情况

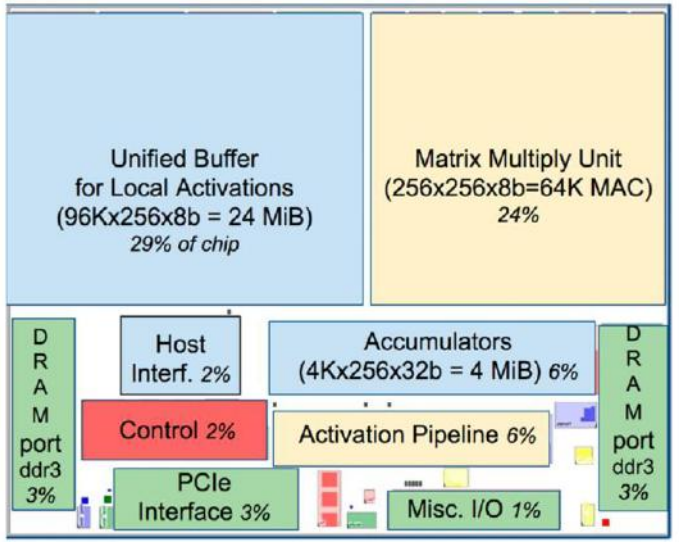
TPU Generations

Year	Inference	Training & Inference	Peak Chip Performance	TDP	Tech. Node	Chips/ Pod	Peak Pod Performance**
2015	TPUv1		92 TOPs/s	75 W	28 nm	--	--
2017		TPUv2	46 TFLOPs/s	280 W	16 nm	256	11 PetaFLOPs/s
2018		TPUv3	123 TFLOPs/s	450 W	16 nm	1024	125 PetaFLOPs/s
2020	TPUv4i (TPUv4 lite)		138 TFLOPs/s	175 W	7 nm	--	--
2021		TPUv4	≥250* TFLOPs/s	--	--	4096	≥1 ExaFLOPs/s

* 1 ExaFLOPs/sec ÷ 4096 TPU v4 chips　　** Bfloat16 FLOPS

资料来源：Jouppi et al., Ten Lessons From Three Generations Shaped Google's TPUv4i, ISCA, 2021

谷歌TPU芯片布局图

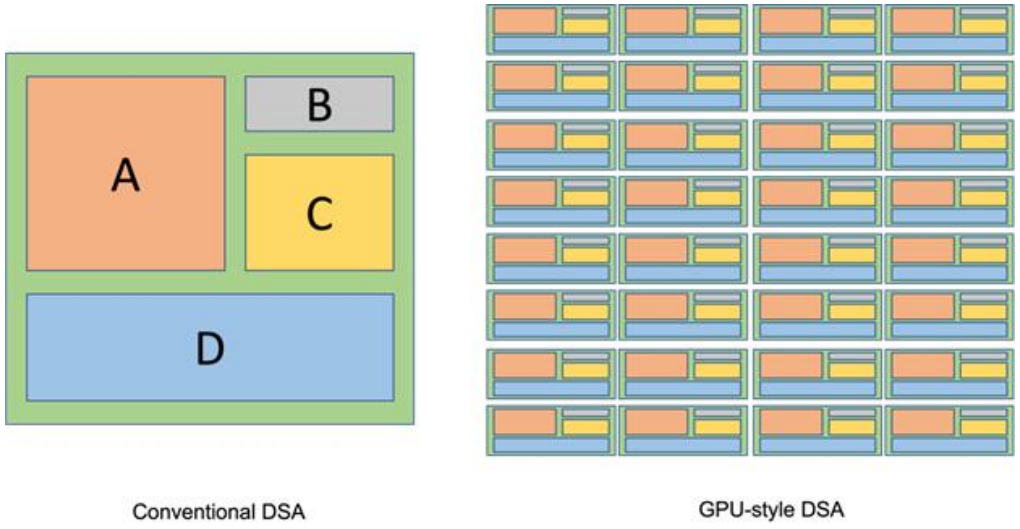


资料来源：谷歌官网

4.1.3 未来趋势：GPU融合DSA，让GPU的通用计算效率接近ASIC的专用效率

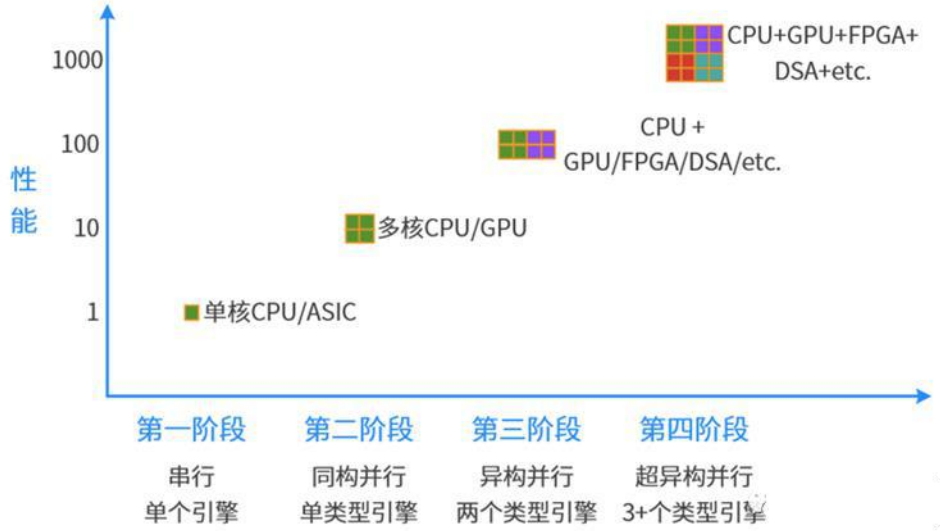
- **DSA通用化：目前的趋势是，GPU融合DSA（Domain Specific Architecture，领域专用架构）于通用架构。但他们是在核心融合，而不是在芯片上层异构化。**
  - 一个DSA设计的硬件资源平均分布到每个运算单元，以特殊指令或是程序呼叫的方式引用，成为各单元通用计算核心的一部分，不在芯片最上层成为一个独立处理器，而是原可编程生态的自然延伸，不影响原先的编程方式。
  - 在提升效能的同时，持续强化通用优势，这使得GPU的通用计算效率处于AI芯片中的领先地位。
- **英伟达H100以“非同步执行”（Asynchronous Execution）提升通用计算效率。因AI算法的多样性及快速演进，非同步执行技术方向的终极目标是要填补通用与专用之间的能效差距，让GPU的通用计算效率接近ASIC的专用效率。**
  - 计算图形化：图形管线是专用管线的代表。虽然中间数个节点已被跑在通用算力池的着色器（Shader）程序取代，它的管线结构依然存在。非同步执行以不浪费时间等待数据传输来接近专用管线的效率。面对后摩尔定律时代的到来，**通用计算借取ASIC风格的专用管线是条必须走下去的路线。**

谷歌的TPU DSA架构对比GPU融合DSA架构



资料来源：《我看英伟达H100 GPU》——吕坚平@天数智芯官网

计算架构从串行到并行，从同构到异构再到超异构

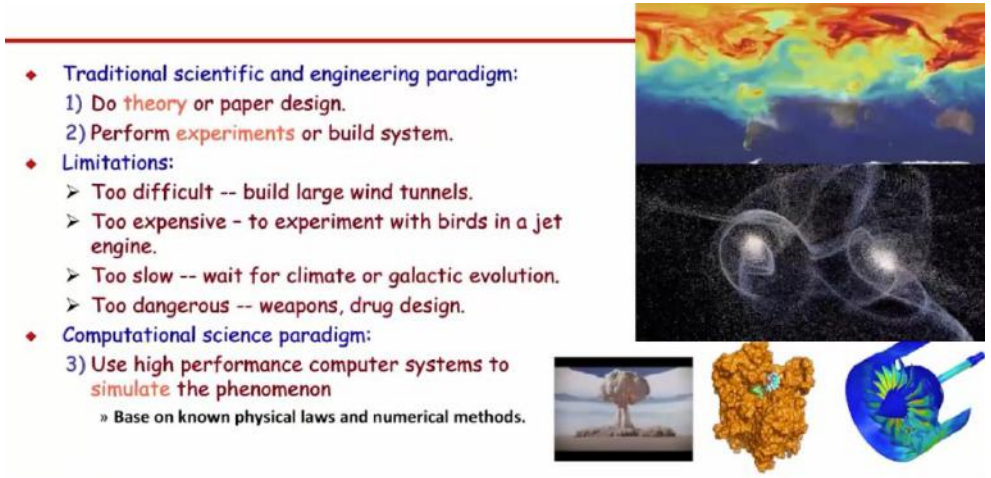


资料来源：软硬件融合微信公众号

4.2 应用场景二——高性能计算（HPC）

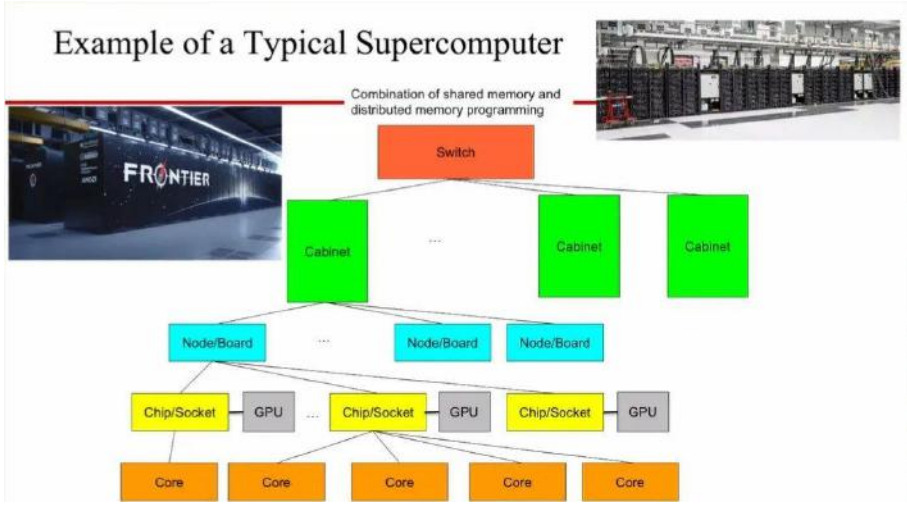
- **高性能计算(HPC) 是指通过聚合计算能力来提供比传统计算机和服务器更强大的计算性能。它能够通过聚合结构，使用多台计算机和存储设备，以极高速度处理大量数据，帮助人们探索科学、工程及商业领域中的一些世界级的重大难题。**
- **GPGPU在图形GPU的基础上进行了优化设计，使之更适合高性能并行计算，加上CUDA多年来建立的完整生态系统，其在性能、易用性和通用性上比图形GPU更加强大。**
  - 基于这种特性，GPGPU将应用领域扩展到图形之外，在自动驾驶、智慧医疗、生命科学、深度学习、云计算、数据处理、金融等方面均得到广泛应用，关于它的科研成果和新应用模式也层出不穷。
  - GPU给计算机提供了强大的数值计算的能力。

高性能计算广泛应用于科研仿真



资料来源：《高性能计算与AI大融合，如何颠覆科学计算》Jack Dongarra@51CTO

高性能计算的载体——超算中心

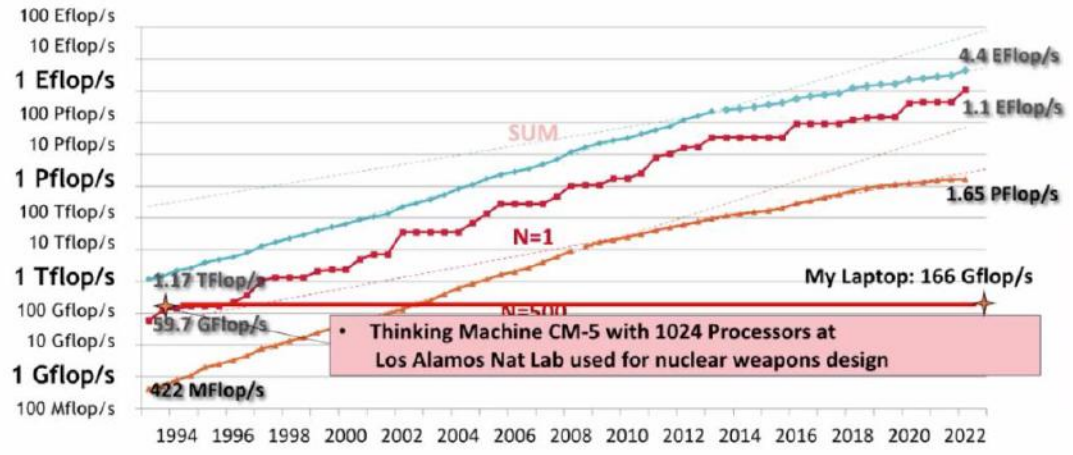


资料来源：《高性能计算与AI大融合，如何颠覆科学计算》Jack Dongarra@51CTO

4.2.1 全球超算算力指数及增长，GPU多采用NVIDIA产品

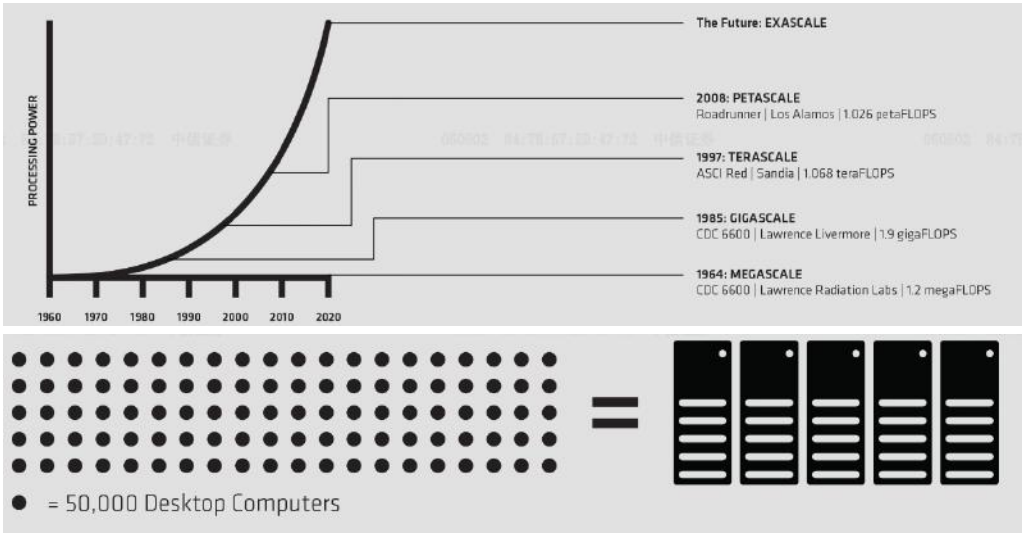
- **HPC主要有CPU和GPU两种类型的处理器，未来或将采用更多不同的单元，比如FPGA、ML加速器和ASIC芯片等等。**
  - 串行处理，由中央处理器 (CPU) 完成。 每个 CPU 核心通常每次只能处理一个任务。
  - 并行处理，可利用多个 CPU 或GPU 完成，GPU 可在数据矩阵（如屏幕像素）中同时执行多种算术运算。
- **超级计算机是计算机中功能最强、运算速度最快、存储容量最大的一类计算机，多用于国家高科技领域和尖端技术研究，是一个国家科研实力的体现。从近三十年间全球超级计算机TOP500的性能变化情况可以发现，超算性能近乎保持着指数级的增长速度。**
  - 如今日常所用的MacBook的性能，比1993年当时世界上最先进的超级计算机的性能还要强大。现在，为了实现 1 百亿亿次级 FLOPS （EFLOPS）的超级计算机处理性能，大概需要 5,000,000 个台式机。
  - 2022年6月的数据显示，全球排名前10的超级计算机当中，有5个来自美国，有2个来自中国（分别位于无锡和广州），其余3个来自芬兰、日本和法国。**2021年，全球最快超级计算机TOP500榜单中，近70%的机器（包括排在前10名中的8台）均采用了NVIDIA技术。**

近三十年间全球超级计算机TOP500的性能变化



资料来源：《高性能计算与AI大融合，如何颠覆科学计算》Jack Dongarra@51CTO

超级计算机算力增长趋势

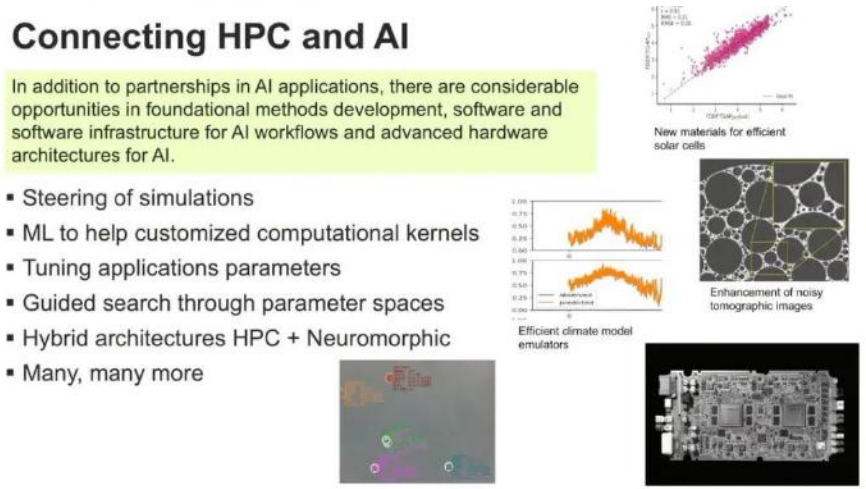


资料来源：AMD官网

4.2.2 HPC和AI计算的关系：数据量不同，未来逐渐融合

- **HPC和AI计算有着即相似又不同的特性。**
  - HPC属于数字计算密集型，通常**输入有限的数据**，经过非常大量的数字计算，**输出大量的数据。**
  - AI计算进行高性能数据处理（HPDA）通常需要**输入大量的数据，输出的却是相对比较少的数据**。
  - 两者使用的数据精度也不同，在科学仿真等高性能计算场景下通常使用64比特浮点数据（FP64），而在AI计算场景下会使用16比特浮点数据（FP16）。
- **高性能计算和AI计算可以非常有效地进行联合，加速计算正在助力研究人员更快取得重大科学突破。**
  - 分析方法得到的模型和其他的模型一起可以被用到计算中去；计算产生的数据和其他来源的数据一起可以被用于AI分析。这样就形成了一个相互促进的良性循环。
  - 在 AI 的助力下，可在更短时间内获得高精度结果，且可与科学模拟结果相媲美。这一结果已推动 AI 在高性能计算中的应用，包括帮助研究人员在实验室开展研究工作，协助工程师解决复杂的技术问题，以及助力金融分析师利用数学算法作出市场预测。

高性能计算和AI计算之间的关系



资料来源：《高性能计算与AI大融合，如何颠覆科学计算》Jack Dongarra@51CTO

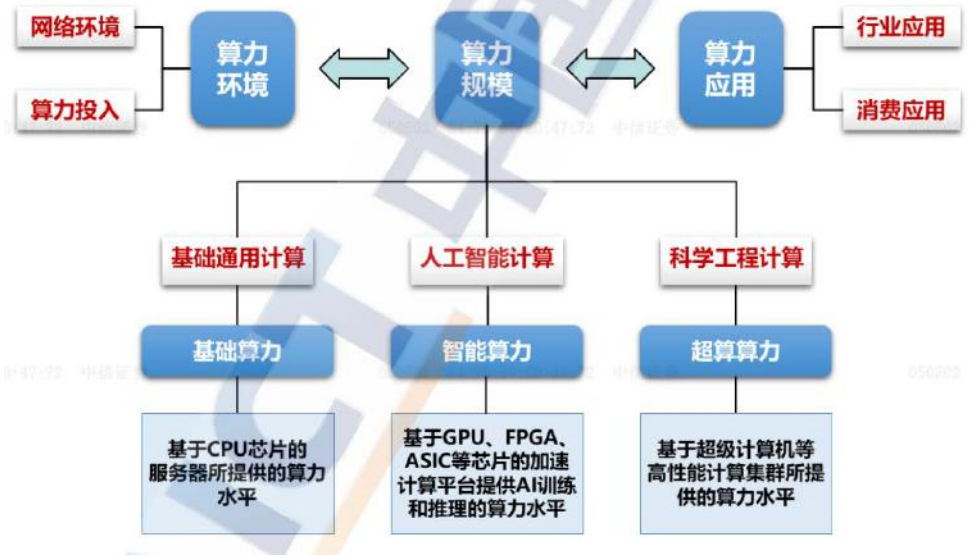
阿里云采用HPC+AI助力新冠药物开发

资料来源：阿里云官网

4.2.3 中国超算事业快速发展，算力核心产业规模达1.5万亿

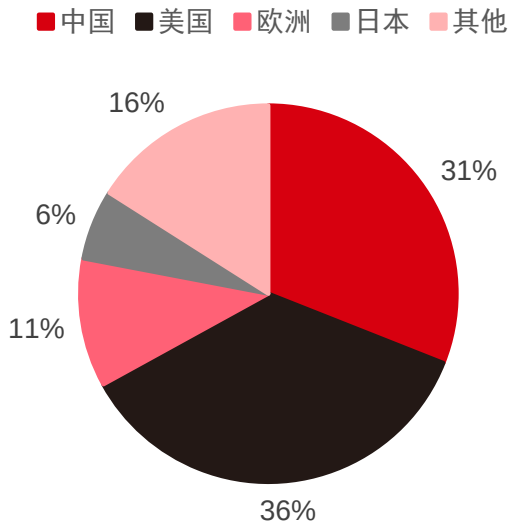
- **现阶段算力规模重点包括基础算力、智能算力和超算算力三部分，分别提供基础通用计算、人工智能计算和科学工程计算**
  - 基础通用算力主要是基于 CPU 芯片的服务器所提供的计算能力；
  - 智能算力主要是基于 GPU、FPGA、ASIC 等芯片的加速计算平台提供人工智能训练和推理的计算能力；
  - 超算算力主要是基于超级计算机等高性能计算集群所提供的计算能力。
- **2021年，我国算力核心产业规模达1.5万亿，关联产业规模超过8万亿。**截至2022年6月底，我国在用数据中心机架总规模超过590万标准机架，服务器规模近2000万台，算力总规模超过150 EFLOPS，位于全球第2。

算力指数计算的构成



数据来源：中国信息通信研究院

全球算力规模分布情况

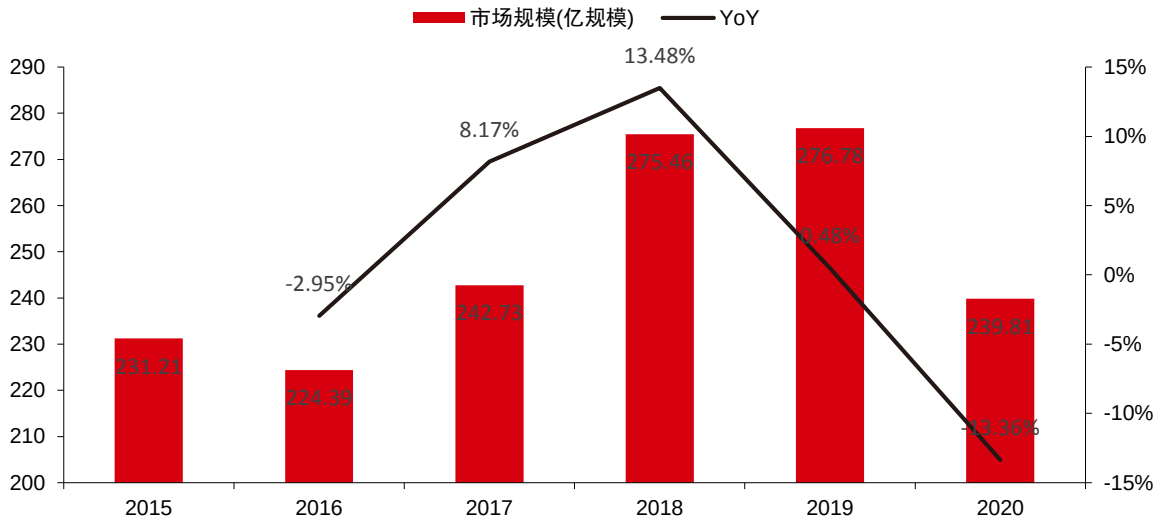


数据来源：中国信息通信研究院，IDC

4.2.4 中国超算GPU市场规模超过19亿美元

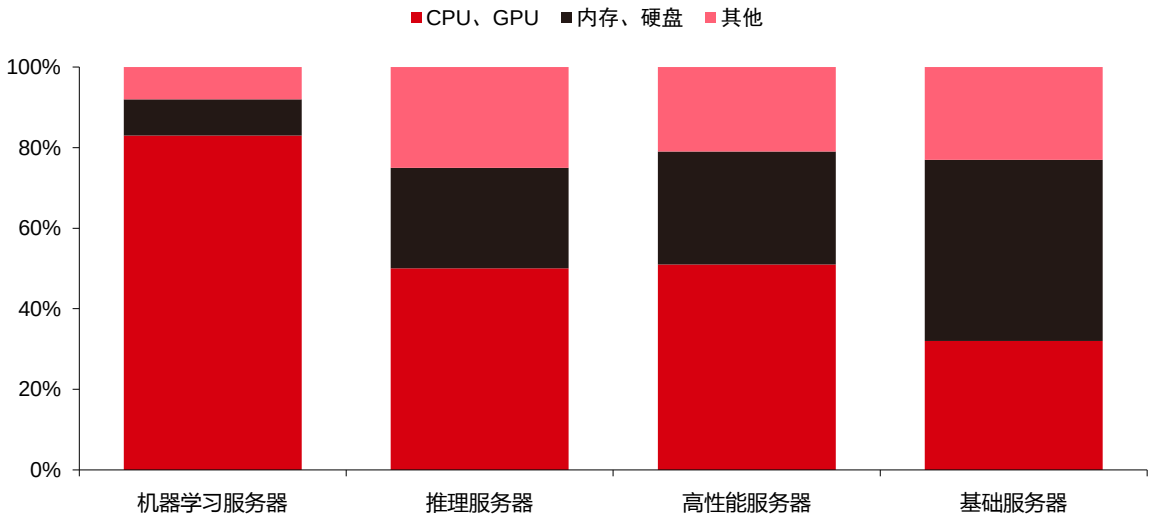
- **根据Hyperion Research报告，IT行业的扩张与虚拟化的进步，以及对混合高性能计算解决方案的需求增长推动着全球超算市场快速发展。**
  - 2019年，全球超算市场规模为276.78亿美元，同比增长0.5%；2020年，受新冠疫情的影响，部分HPC厂商的关闭与延迟产品出货的原因，市场规模下降至239.81亿美元，同比下降13.4%。
  - CPU及GPU为代表的芯片占据主要的成本。在高性能计算服务器中，芯片成本占比高达 51%，按照超算中GPU价值量占比80%计算，**全球超算GPU市场约为96亿美元。**
- **根据信通院发布的《中国算力发展白皮书》，2020年中国超算算力总规模约为2EFlops（换算成FP32），全球占比约为20%，因此我们估算得到，2021年中国超算GPU市场规模约为19.2亿美元。**

2015-2020年全球超算市场规模及增速



数据来源：Hyperion Research，中信证券研究部

各种服务器成本构成



资料来源：芯八哥@芯语，中信证券研究部

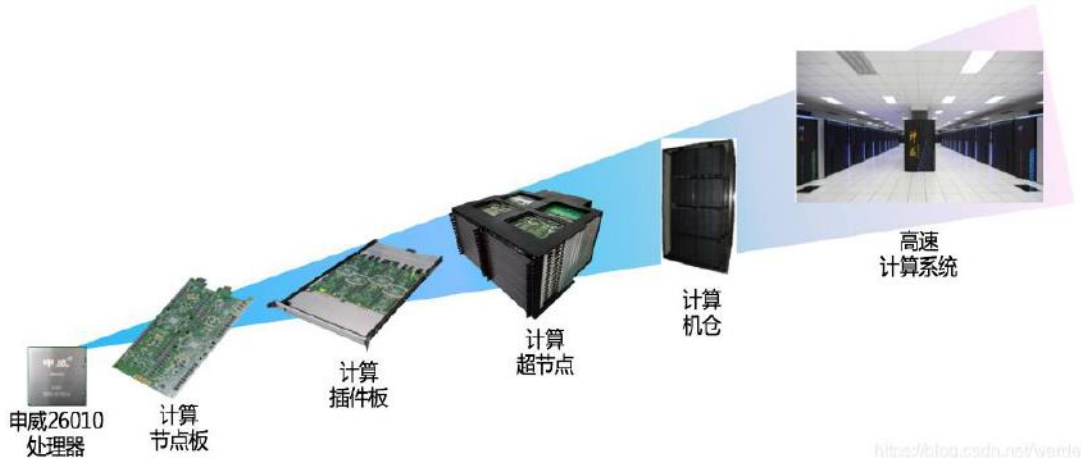
4.2.5 美国对中国超算多次限制，目前国内顶级超算多采用自主设计研发的加速器

- **过往来看，美国已经对中国超算多次限制。**2015年中国“天河二号”项目相关的4家中国机构被美国列入“实体清单”；2019年，海光、中科曙光、无锡江南计算技术研究所等5家进入实体清单；2021年，飞腾、申威等7家超算机构进入实体清单。
- **根据澎湃在2011年，国内就完成了“神威蓝光”超算的研制，这款超算的性能在同时期不突出，新闻，目前国内顶级超算多采用自主设计研发的加速器芯片，实现芯片的国产化。**
  - 但胜在超算芯片完全自主设计。在2016年，采用SW26010的“**神威太湖之光**”正式亮相，成为全球首个100P级超算，并连续4次蝉联TOP500第一名。
  - 另外，天河超算在超算芯片上也使用了自主设计的加速器取代了英特尔的加速器，天河2号使用国产加速器升级之后，性能提升了70%以上。**“天河三号”原型机采用自主的飞腾处理器、天河高速互联通信和麒麟操作系统，实现了芯片的全国产化。**

天河三号超算原型机

资料来源：澎湃网

申威26010处理器组成的超算系统



资料来源：《神威太湖之光简介》——Werdery@CSDN

4.2.6 超级计算——科学研究

- **根据招标采购信息，高校和研究机构，对英伟达标志性的A100芯片依赖度较高。**
  - 清华大学2021年10月斥资超过40万美元购买了两台英伟达AI超级计算机，每台由四颗A100芯片驱动。同月，中国科学院计算技术研究所在A100芯片上花费了约25万美元。今年7月，中科院人工智能学院也在高科技设备上花费了约20万美元，其中包括部分由A100芯片驱动的服务器。
  - 2021年11月，广东暨南大学网络安全学院在英伟达AI超级计算机上花费了超过93000美元，而其智能系统科学与工程学院仅在今年8月就花费了近10万美元购买了8个A100 GPU板卡，单价为8.7万元/片。
  - 招标显示，山东、河南和重庆等省市政府支持的研究所和大学也购买了A100芯片。
- **A100价格昂贵，大部分高校科研机构还是比较追求性价比的，但是通过使用多个中低端芯片来复制高端A100芯片的处理能力，也基本可以满足高性能计算的要求。**

清华大学招标采购A100服务器

采购需求：

包号	名称	数量	是否允许进口产品投标	采购预算（人民币）
01	A100服务器	4台	否	280万元

设备用途介绍：主要用于超大规模预训练模型的训练以及推理，进行模型的多机多卡分布式计算，并能快速完成相关计算过程，可实现基于CUDA的任意神经网络计算。

简要技术指标：AMD EPYC 7002系列处理器（单颗至少16核心）*2；内存不低于 DDR4 32G*16；显卡NVLINK 3.0 A100*4；IB HDR单口网卡*2（200Gbps）；PCI-E 4.0 x16 (FHFL) slots *4。

合同履行期限：合同签订后4个月内

资料来源：中国政府采购网

暨南大学招标采购A100GPU板卡

一、项目编号：ZZ0220310-1（招标文件编号：ZZ0220310-1）
二、项目名称：暨南大学智能科学与工程学院图形处理器GPU板卡采购项目（重招）
三、中标（成交）信息
供应商名称：广州市恒联计算机科技有限公司
供应商地址：广州市天河区天河北路898号3010房
中标（成交）金额：人民币693,000.00元

四、主要标的信息

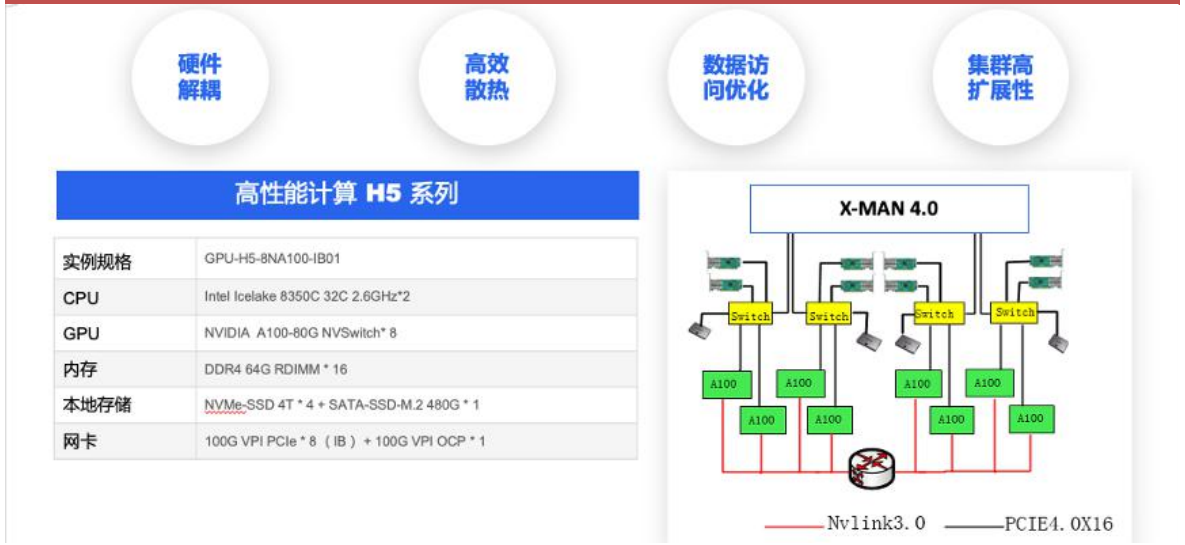
序号	供应商名称	货物名称	货物品牌	货物型号	货物数量	货物单价（元）
1	广州市恒联计算机科技有限公司	图形处理器GPU板卡	英伟达	A100 80GB PCIe	8块	86,625.00

资料来源：中国政府采购网

4.2.7 超级计算——云计算

- **因为云服务需要尽可能提升算力，所以中国云服务提供商采用A100来满足各行各业的多样化计算需求。如果禁售，对云计算厂商的影响较大，但是如果采用前代或中端产品V100、A10、T4等，多个芯片也可以实现相同的计算性能。**
  - 阿里云：基于**NVIDIA A100** 打造的gn7 GPU系列云服务器，该产品主要面向AI训练和**高性能计算**应用，可提供新一代GPU计算实例。相比上一代平台实现最高20倍的AI性能，以及2.5倍的高性能计算速度。
  - 百度智能云：基于**NVIDIA A100**打造的云服务器以及裸金属服务器产品，最高将搭载8块 NVIDIA A100 GPU，主要面向AI训练/推理、**高性能计算**应用、**科学计算**等场景。基于A100 TF32新技术，百度新一代GPU云服务器提供20倍于 V100 FP32云服务器的计算能力。
  - 滴滴云：A100裸金属服务器配置了8块**NVIDIA A100** GPU 、适用于AI、数据分析、**高性能计算**等多种应用场景。
  - 腾讯云：搭载**NVIDIA A100**的GPU云服务器GT4，适用于深度学习训练、推理、**高性能计算**、数据分析、视频分析等领域，可提供更高性能的计算资源，从而进一步降低使用成本，帮助企业、高校及研究人员聚焦模型的优化与创新。

百度智能云搭载的A100产品



资料来源：百度智能云官网

腾讯云搭载的A100产品

型号	GPU (NVIDIA Tesla A100 NVLink 40G)	GPU 显存 (HBM2)	vCPU	内存 (DDR4)	内网带宽	网络收发包 (PPS)	队列数
GT4.4XLARGE96	1颗	1 * 40GB	16核	96GB	5Gbps	120万	4
GT4.8XLARGE192	2颗	2 * 40GB	32核	192GB	10Gbps	235万	8
GT4.20XLARGE474	4颗	4 * 40GB	82核	474GB	25Gbps	600万	16
GT4.41XLARGE948	8颗	8 * 40GB	164核	948GB	50Gbps	1200万	32

资料来源：腾讯云官网

4.3 美国对华禁令如何应对？

美国政府对高端GPGPU芯片封锁

- **根据路透社报道，2022年8月31日，美国政府要求英伟达的A100、H100系列和AMD的MI 250系列及未来的高端GPU产品，是否可以售卖给中国客户，需要获得美国政府的许可。**
  - 这几款芯片均为用于通用计算的高端GPGPU，通常应用在人工智能计算的云端训练和推理场景和超级计算机中，在中国的客户多为云计算厂商及高校和科研院所。
  - 据我们测算，2021年中国GPGPU市场规模为149.8亿元，其中人工智能推理/人工智能训练/高性能计算分别为93.5/47.1/9.1亿元，**本次主要受到影响的是人工智能训练/高性能计算应用，合计约56.2亿元的市场。**

如何应对封锁？

- **在需要大量算力的人工智能的训练端和高性能计算中：**
  - 短期来看，选择英伟达和AMD的还没有被禁止的中低性能GPU芯片。对于云端计算，算力既可以通过产品升级得以提升，也可以通过增加计算卡的数量进行提升，因此**短期内可以通过使用多个算力较低的CPU、GPU和ASIC芯片来复制高端GPU芯片的处理能力，基本可以满足云端训练和高性能计算的要求。**
  - 长期来看，选择国产GPU进行替代。虽然芯片是算力的主要来源和最根本的物质基础，但是算力的生产、聚合、调度和释放是一个完整过程，需要复杂系统的软硬件生态共同配合，才能实现“有效算力”。因此短期内可能会因为无法兼容在人工智能领域广泛使用的CUDA架构而遭遇替换困难，但是**长期来看，国产CPU、通用GPU、AI芯片将获得前所未有的发展机会，通过软硬件技术提升，逐步实现高端GPU领域的国产化替代。**
- **对于不需要太强算力的推理端GPU芯片：**
  - 以往通常采用中低端计算芯片，例如NVIDIA Tesla T4、P4、T40等产品，暂时没有被禁售的风险。但是，长久来看，FPGA和ASIC的优势逐渐凸显，**国产FPGA和ASIC芯片产品目前已经运用到云计算厂商中，未来有望实现推理端替代。**

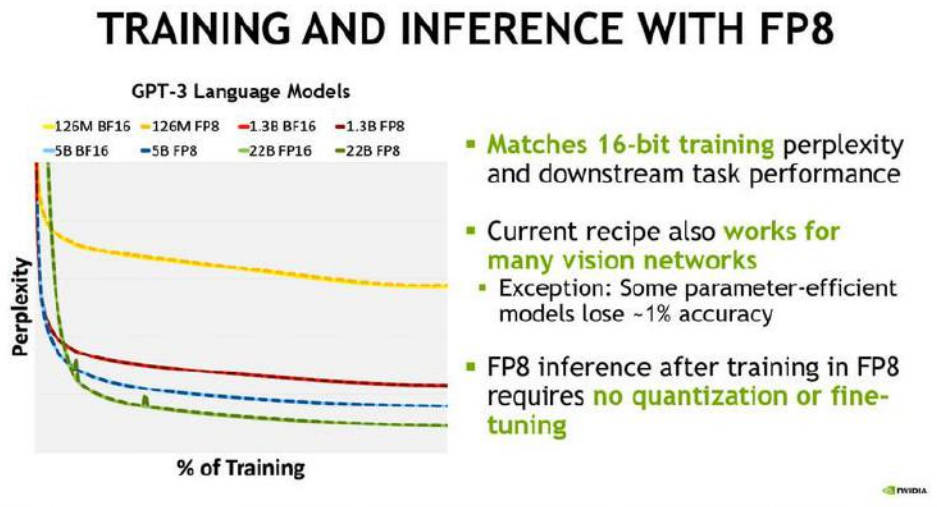
4.3.1 云训练/超算：短期CPU、ASIC、中低端GPU替代，长期国产化替代

- **AI训练端和超级计算需要强大的算力支持，因此计算能力强的GPU仍是第一选择。目前GPU的市场格局以英伟达为主，AMD为辅，预计未来几年GPU仍然是深度学习市场的第一选择。**
- **但是，因为AI训练并不是必须要高精度浮点运算，目前NVIDIA的H100的FP8运算的计算在速度和精度上取得平衡，基本上和FP16/BF16达到一致的精度。随着ASIC芯片的算力逐渐增强，在训练端的应用场景也逐渐增多。**
- **对于禁令，我们预计有两种应对方案：**
  - **短期来看，可用英伟达和AMD的还没有被禁止的、以及国产厂商的中高计算性能CPU、GPU、ASIC芯片。**对于云端计算，算力既可以通过产品升级得以提升，也可以通过增加计算卡的数量进行提升，因此短期内可以通过使用多个中低端芯片来复制高端GPU芯片的处理能力，基本可以满足云端训练要求。
  - **长期来看，改用国产品牌（天数智芯、壁仞科技、寒武纪、燧原科技、沐曦、摩尔线程等）的ASIC、GPU芯片进行国产化替代。**虽然芯片是算力的主要来源和最根本的物质基础，但是算力的生产、聚合、调度和释放是一个完整过程，需要复杂系统的软硬件生态共同配合，才能实现“有效算力”。
- **短期内可能会因为无法兼容在人工智能领域广泛使用的CUDA架构而遭遇替换困难，但是长期来看，随着生态的逐渐补课，国产通用GPU、ASIC芯片将获得前所未有的发展机会，通过软硬件技术提升，逐步实现AI训练芯片的国产化替代。**

阿里云推出的搭载NVIDIA A100的GPU云服务器

资料来源：阿里云，中信证券研究部

FP8 基本上和FP16/BF16 达到 一致的精度



资料来源：英伟达论坛

4.3.2 云推理：ASIC和FPGA加速替代GPU

- **下游推理端更接近终端应用，更关注响应时间而不是吞吐率，考虑的因素更加综合：单位功耗算力、时延、成本等。除了主流的GPU芯片之外，下游推理端可容纳FPGA、ASIC等芯片。**
  - 除了Nvidia、Google、Xilinx（AMD）、Altera（Intel）等传统芯片大厂涉足云端推理芯片以外，Wave computing、Groq 等初创公司也加入竞争；中国公司里，寒武纪、比特大陆、燧原科技等同样积极布局云端芯片业务。
  - 竞争态势中GPU依然占大头，但随着AI的发展，**FPGA的低延迟、低功耗、可编程性（适用于传感器数据预处理工作以及小型开发试错升级迭代阶段）和ASIC的特定优化和效能优势（适用于在确定性执行模型）将凸显出来。**
- **对于推理端GPU芯片来说，因为不需要太强的算力，所以通常采用中低端计算芯片，例如NVIDIA Tesla T4、P4、T40等产品，暂时没有被禁售的风险。但是，长久来看，FPGA和ASIC的优势逐渐凸显，国产FPGA和ASIC芯片产品目前已经运用到云计算厂商中，未来有望实现推理端替代。**

NVIDIA AI计算芯片的主要云服务客户

	M60	P4	P40	P100	T4	RTX 6000/8000	V100	A10	A40	A100	NGC
阿里云		✓		✓	✓		✓			✓	✓
AWS	✓				✓		✓	✓		✓	✓
百度云		✓	✓		✓		✓			✓	
Google Cloud		✓		✓	✓		✓			✓	✓
IBM Cloud	✓			✓			✓				
Microsoft Azure	✓		✓	✓	✓		✓	✓		✓	✓
Oracle Cloud				✓			✓			✓	✓
腾讯云		✓	✓		✓		✓			✓	✓
NPN CSPs	✓			✓	✓	✓	✓		✓		

资料来源：英伟达官网

寒武纪的主要云服务客户及合作伙伴

资料来源：寒武纪官网

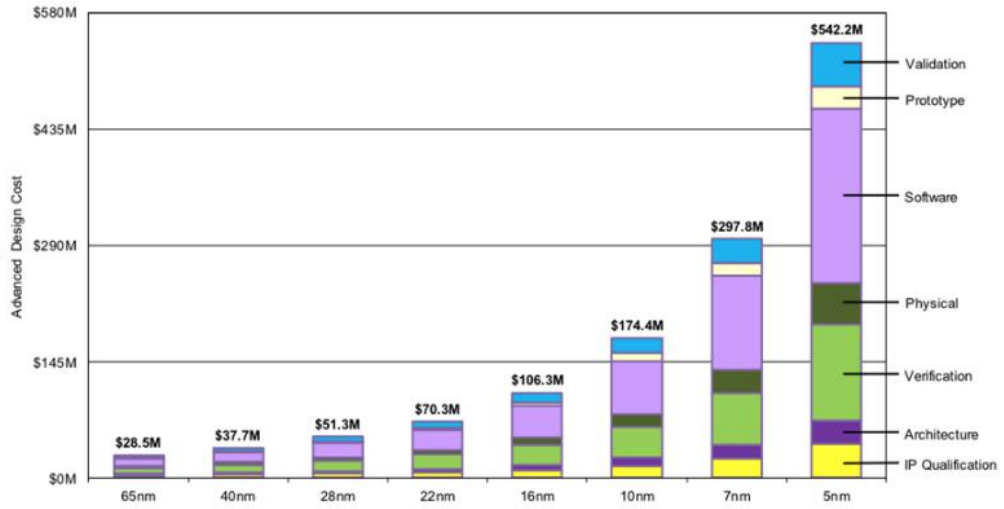
CONTENTS

目录

5.1 制造：目前国内AI芯片先进工艺多集中在7nm，国际大厂已经来到4nm

- **大陆的先进工艺设计（16nm及以下）集中于AI芯片（包含云端及智能驾驶芯片）、交换机芯片、CPU/GPU/DPU、矿机ASIC领域，这些领域各有一些头部企业走在前列，但鲜有企业能够进入个位数先进制程。GPGPU和AI芯片因为去掉了图形渲染功能，功能相对单一，设计起来复杂度及难度较低。**
  - 先进制程芯片的设计成本大幅增加。设计一颗28nm芯片成本约5000万美元，而7nm芯片需要3亿美元，5nm则需要5.42亿美元。
  - 若以麒麟的5nm工艺来对标，除了矿机ASIC中的比特大陆推出了基于最先进的5nm的矿机芯片，平头哥发布了自研5nm服务器芯片倚天710，中兴通讯的7nm芯片已实现商用正在研发5nm芯片之外，
  - 其他领域快的如有些国内自动驾驶芯片公司要量产7nm智能座舱芯片，互联网巨头的一些AI芯片在向5nm迈进，**CPU/GPU/DPU领域大多企业还只是规划向5nm迈进，大多数节点还在16nm或10nm之上，真正实现5nm芯片量产的较少。**
- **目前国内GPGPU芯片的先进制程多集中在7nm，例如已经量产的天数智芯“天垓100”，已经推出的壁仞BR100、沐曦MXN；此外，芯动科技的“风华一号”以及摩尔线程的MTT S2000采用12nm制程。对比已经进入4nm时代的英伟达 H100还有较大差距。**

随着制程节点的提升，开发芯片所需要的资金显著提高



资料来源：International Business Strategies (IBS)

目前采用7nm工艺的国产AI芯片产品

产品型号	产品类型	推出时间	制造工艺	封装工艺
华为 昇腾910	ASIC	2018	7nm	
寒武纪 思元370	ASIC	2021	7nm	Chiplet
天数智芯 天垓100	GPU	2021	7nm	2.5D CoWoS
海光 深算一号	DCU	2021	7nm	
壁仞BR100-OAM	GPU	2022	7nm	2.5D CoWoS
壁仞BR104-300W PCIe	GPU	2022	7nm	2.5D CoWoS

资料来源：各公司官网，中信证券研究部

5.2 生态：GPGPU难点在生态布局，目前市场几乎被CUDA垄断

- **按功能划分，GPU主要分为侧重图形图像的渲染GPU和侧重通用计算的GPGPU。**
  - 目前国内GPGPU公司包括**壁仞、沐曦、天数智芯、红山微电子**等；图形渲染GPU企业包括**景嘉微、芯动科技、摩尔线程、格兰菲**等。
  - 渲染GPU约80%仍是GPGPU部分，20%则是固定渲染部分（fixed function）。图形GPU因为经过了十几年的演化进程，流水线长，实现起来复杂，设计上的挑战更大，同时存在很多专利陷阱，涉及较多的数学公式，**因此比较依靠编译器和驱动等软件能力；GPGPU在硬件和应用层面较为复杂。**
- **GPU生态是除产品外初创企业能否活下来的重要因素。**
  - 渲染GPU在技术层面来相对复杂，但是好处在于有很多业界成熟的标准的API，如OpenGL、OpenGL ES、DirectX、Vulkan等，核心是**打通驱动程序层和编译器生态。**
  - 而**GPGPU领域几乎是被英伟达一手打造的CUDA生态所垄断。**易于编程和性能的巨大飞跃是 CUDA 平台被广泛采用的关键原因之一。CUDA 平台成功的第二大原因是拥有广泛而丰富的生态系统。

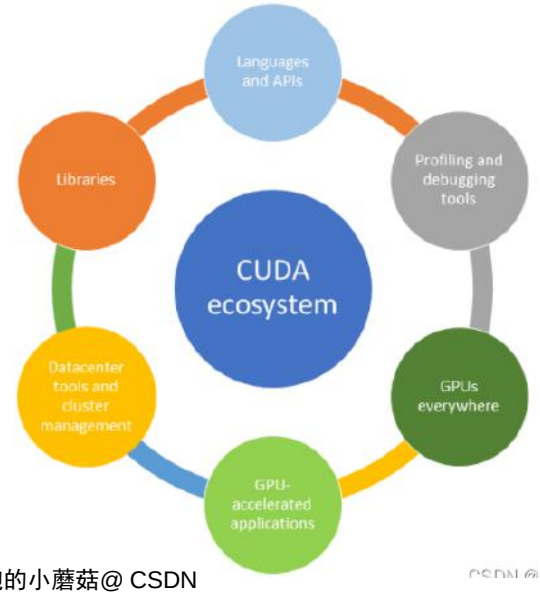
图形渲染GPU和AI芯片的软硬件开发难度对比

技术门槛：GPU的软硬件系统设计复杂度超越AI芯片100倍以上

	GPU	GPGPU&AI
指令集	千条量级	百条以内
算力利用率	**基于硬件层的任务管理和智能调度能力要求高** 让芯片从硬件层即提高算力的利用率	**多依赖于软件层的调度实现，**但： • 增加了软件开发的复杂度 • 降低了硬件算力的利用率 • 减缓了软件栈迭代更新的速度
开发生态	• 生态系统复杂 • 涉及行业应用广泛 • 行业标准成熟(OpenCL, OpenGL, Vulkan, DirectX)	• 新生态面临客户冗长的适配期和巨大的软件投入 • 产品“快”，应用“慢”

资料来源：兴旺投资

GPGPU领域，英伟达的CUDA具有巨大的护城河

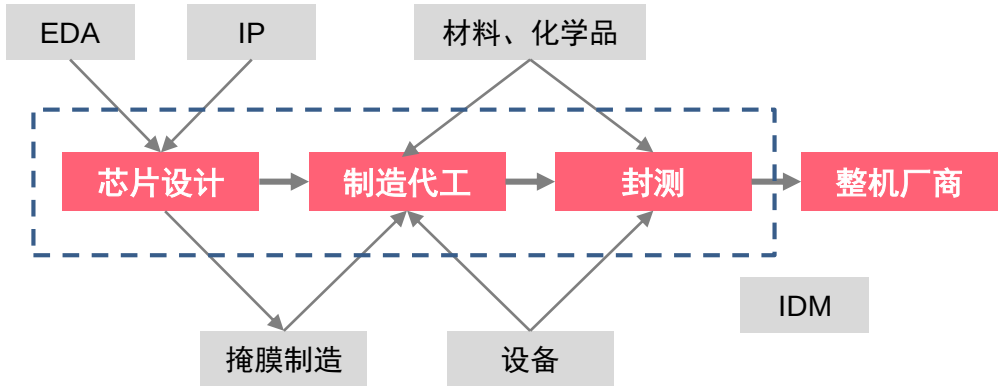


资料来源：英伟达，奔跑的小蘑菇@ CSDN

5.3 IP：国内核心IP厂商，追赶国际厂商

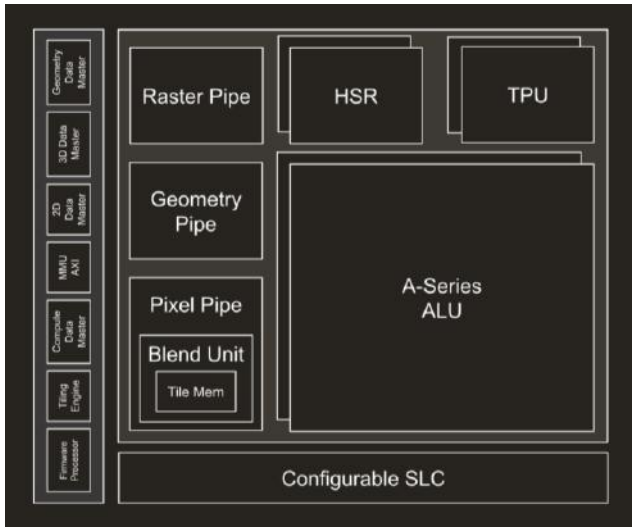
- **自研IP越多，芯片设计上越有把握，产品的差异化更明显。但相对而言，资金、人员、时间上的成本投入也更高。**
  - **GPU IP自研需要36-48个月以及200个工程师，**而采用外购IP的方式，可以减少12-18个月开发周期。
  - 据集微网报道，GPU 的IP主要涉及三大类，一是模拟IP，包括PCIe、Displayport和HDMI等等，这方面国内厂商占有率较低；二是Memory；三是数字IP，包括基于Arm或RISC-V的微控制器IP、编解码芯片IP以及最核心的GPU IP等。
- **核心IP国内有Imagination、芯原、格兰菲等厂商。**
  - 根据集微网报道，相对而言，Imagination认可度较高，芯原是后起之秀，格兰菲则主要面向特定领域用户，目前整体和国际厂商还有较大差距，在此过程需要技术沉淀形成自主IP积累才能具有一定替代性。
  - 我们看来，国内信创和工业市场需求庞大，在当前阶段对于国内IP厂商和GPU厂商来说是一个构建自主生态的机会。

芯片设计和制造流程



资料来源：中信证券研究部绘制

中资控股的Imagination的典型GPGPU IP



资料来源：Imagination官网

5.4 国内GPGPU、AI加速芯片和国外产品对比

- **国内GPU产品梳理以及和国外产品的对比**
  - 加速计算GPU领域，国内壁仞科技发布的BR100产品，在FP32单精度计算性能实现超越NVIDIA A100芯片，但是不支持FP64双精度计算；天数智芯推出的天垓100的FP32单精度计算性能实现超越A100芯片，但是在INT8整数计算性能方面却低于A100；海光推出的DCU Z100实现了FP64双精度浮点计算，但是其性能为A100的60%左右。因此，**从高精度浮点计算能力来看，国内GPU产品与国外产品的计算性能仍或有一代以上差距。**
  - 但是，GPU的表现不仅体现在硬件上，软件层面对于生态的布局尤其重要，目前国内企业多采用OpenCL进行自主生态建设，但这是需要大量的时间进行。对比AMD从2013年开始建设GPU生态近10年时间后才推出用于通用计算的ROCm开放式软件平台，我们认为国内厂商在软件和生态层面与英伟达CUDA生态的差距相较于硬件更为明显。
- **虽然目前国内产品的计算性能和软件生态与国际厂商还有较大差距，但是，国内厂商依然在奋起直追，努力实现GPU国产化突破。**其中包括龙芯中科、海光信息、壁仞科技、寒武纪、天数智芯等厂商均在研发或推出用于AI计算的GPGPU、ASIC等AI芯片，有望实现高端AI芯片的国产替代。

图形渲染GPU产品梳理

产品型号	推出时间	制造工艺	支持API	时钟频率(MHz)	显存带宽(GB/s)	显存容量(GB)	像素填充率(GPixels/s)	浮点性能(FP32)	AI运算性能（INT8）	功耗
景嘉微 JM7200	2018	28nm	OpenGL 2.0	1300	17	4	5.2	500GFlops		＜20W(桌面) ＜10W(嵌入式)
英伟达 GT640	2012	28nm	图形：OpenGL 4.3，DirectX 11； 计算：OpenCL 1.1	900	28.5	2	14.4	692GFlpos		50W
景嘉微 JH920	2021	14nm	图形：OpenGL 4.0，Vulkan 1.1，DirectX 11； 计算：OpenCL 3.0	1500	128	8	32	1.5TFlops		＜30W
格兰菲 Arise-GT-10C0	2022	28nm	图形：OpenGL 4.5，DirectX 11； 计算：OpenCL 1.2	1200		4	48	1.5TFlops		45W
芯动科技 风华2号	2022	-	图形：OpenGL 4.3，Vulkan 1.2，DirectX 11，OpenGLES 3.2； 计算：OpenCL 3.0		102.4	8	48	1.5TFlops	12.5TOPS	4~15W
英伟达 GTX1050	2016	14nm	图形：OpenGL 4.6，Vulkan 1.2，DirectX 12； 计算：OpenCL 3.0，CUDA 6.1	1354	112	2	36.43	1.862TFlops		75W

资料来源：各公司官网，中信证券研究部 注：产品数据均按照最大值进行统计

5.4 国内GPGPU、AI加速芯片和国外产品对比

国内GPGPU、AI加速芯片产品梳理及与国外GPGPU产品对比

产品型号	产品类型	推出时间	制造工艺	封装工艺	浮点算力(TFlops)			INT8定点算力(TOPS)	生态	互联带宽	显存(GB)	接口	功耗
					FP64	FP32	BF16						
华为 昇腾910	ASIC	2018	7nm				320	640	MindSpore			PCIe 4.0	350W
寒武纪 思元370	ASIC	2021	7nm	Chiplet		24		256	Cambricon Neuware	200GB/s	16	PCIE 4.0	250W
天数智芯 天垓100	GPU	2021	7nm	2.5D CoWoS		37	147	295	SIMT	64 GB/s	32	PCIE 4.0	250W
海光 深算一号	DCU	2021	7nm		5.4				兼容 ROCm		32	PCIE 4.0	350W
壁仞BR100-OAM	GPU	2022	7nm	2.5D CoWoS		256	1024	2048	BIRENSUPA	512GB/s	64	PCIe 5.0	550W
壁仞BR104-300W PCIe	GPU	2022	7nm	2.5D CoWoS		128	512	1024	BIRENSUPA	192GB/s	32	PCIe 5.0	300W
英伟达 Tesla V100	GPU	2017	12nm Volta		7.8	15.7	125	62	CUDA	150GB/s	32	PCIe 4.0	300W
英伟达 A100 PCIe	GPU	2020	7nm Ampere		9.7	19.5	312	624	CUDA	600GB/s	80	PCIe 4.0	400W
英伟达 H100 SXM5	GPU	2022	4nm Hopper	2.5D CoWoS	30	500	1000	2000	CUDA	900GB/s	80	SXM5	700W
英伟达 H100 PCIe	GPU	2022	4nm Hopper	2.5D CoWoS	24	48	800	1600	CUDA	900GB/s	80	PCIe 5.0	350W
AMD Instinct MI100	GPU	2020	7nm CNDA 1		11.5	23.1	92.3	184.6	AMD ROCm	276GB/s	32	PCIe 4.0	300W
AMD Instinct MI250	GPU	2021	6nm CNDA 2		47.9	45.3	362	362	AMD ROCm		128	PCIe 4.0	560W
AMD Instinct MI250X	GPU	2021	6nm CNDA 2		47.9	47.9	383	383	AMD ROCm		128	PCIe 4.0	560W

数据来源：各公司官网，中信证券研究部

龙芯中科：国内PC CPU龙头，自主研发GPGPU产品

- **公司2022年上半年完成了第一代龙芯图形处理器架构LG100系列图形处理器核的研制，并随7A2000芯片产品发布。**
  - 基于龙芯最新一代的LG100三维GPU核，完成了GPU驱动、显示需求和系统配套组件的研发，可满足日常桌面办公使用需求，提高产品性价比和商业竞争力。
  - 目前已启动第二代龙芯图形处理器架构LG200系列图形处理器核的研制。
- **公司正在研发新一代完全自主可控的具有高通用性、高可扩展性的 GPGPU 芯片产品及软硬件体系，**
  - 将加速对象从单纯的图形渲染扩展到科学计算领域，提升算力密度同时降低单位算力功耗，并在此基础上有效支持视觉、语音、自然语言及传统机器学习等不同类型的人工智能算法。

搭载龙芯图形处理器架构LG100的龙芯7A2000桥片

资料来源：龙芯中科微信公众号

龙芯中科高性能通用图形处理器芯片及系统研发项目进度安排

项目	2022年				2023年				2024年			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
场地装修改造												
软硬件设备购置												
研发人员招募												
芯片研发												

资料来源：龙芯中科招股说明书

海光信息：基于通用GPGPU架构，推出深度计算处理器DCU

- **公司基于通用的GPGPU架构，设计、发布的适合计算密集型和运算加速领域的一类协处理器，定义为深度计算处理器DCU（Deep-learning Computing Unit，深度计算处理器）**
  - 目前公司的系列产品“深算一号”已经实现商业化应用，主要应用于大数据处理、人工智能、商业计算等应用领域。
  - 海光DCU系列产品已于2021年实现商业化应用。海光DCU兼容“类CUDA”环境，软硬件生态丰富。
- **募投项目：在已有海光DCU产品的基础上，根据大数据处理、人工智能、商业计算等领域具体应用的最新需求，设计新型DCU芯片架构**
  - 增加并行计算单元的数量，优化计算单元的微结构，针对不同领域的特定应用增加专用指令；
  - 扩大高速缓存容量，优化存储子系统的微结构；改进片上网络拓扑结构和路由算法，支持更大的芯片互连规模；设计周期精确的模拟器，支持芯片架构研发和应用性能评估。

海光DCU的基本组成架构

资料来源：海光信息招股说明书

深算一号和国际领先GPU生产商产品对比

项目	海光	NVIDIA	AMD
品牌	深算一号	Ampere 100	MI100
生产工艺	7nm FinFET	7nm FinFET	7nm FinFET
核心数量	4096 (64 CUs)	2560 CUDA processors 640 Tensor processors	120CUs
内核频率	Up to 1.5GHz（FP64） Up to 1.7Ghz（FP32）	Up to 1.53Ghz	Up to 1.5GHz（FP64） Up to 1.7Ghz (FP32)
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存位宽	4096 bit	5120 bit	4096bit
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350 W	400 W	300W
CPU to GPU互联	PCIe Gen4 x 16	PCIe Gen4 x 16	PCIe GEN4 x 16
GPU to GPU互联	xGMI x 2, Up to 184 GB/s	NVLink up to 600 GB/s	Infinity Fabric x 3，up to 276 GB/s

资料来源：海光信息招股说明书

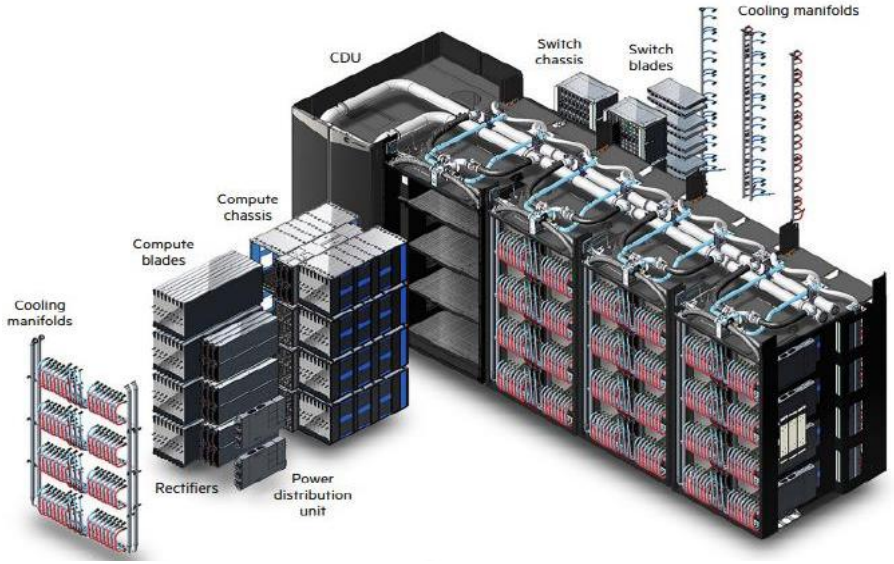
AMD：发力数据中心，产品性能实现代际飞跃

- **AMD 寄希望于 HPC 和 AI 蓬勃发展的未来，它们正被大量部署以支持科学家在气候变化、疫苗等方面的研究工作。AMD 发起了一场与数据中心巨头英特尔和英伟达激烈竞争的运动。AMD还以 490 亿美元收购了赛灵思，完成了历史上最大的芯片收购，扩大了数据中心的机会。**
- **由 AMD EPYC 处理器和 AMD Instinct MI200 加速器提供支持的超级计算机将为大规模模拟和建模以及人工智能和深度学习工作负载提供性能上的代际飞跃。**
- **2022年6月，国际超级计算机大会 (ISC)，发布了最快的超级计算机 TOPP500 榜单，其中一台名为 Frontier 的计算机位居榜首。**
  - 它部署在美国能源部橡树岭国家实验室，是第一台 exascale 机器（每秒 10 18次浮点运算）——**由 AMD Epyc CPU 和 Instinct MI250 GPU 驱动的 HPE-Cray EX 系统。**但是英伟达协处理器可以在 154 台 TOP500 超级计算机中找到；只有七台超级计算机使用 AMD Instinct 卡。

AMD历代GPGPU产品架构的发展

资料来源：AMD官网

由 AMD Epyc CPU 和 Instinct MI250 GPU 驱动的 HPE-Cray EX



资料来源：HPCWire网站

英伟达：H100拥有最强算力，使AI训练速度提升9倍、推理提升30倍

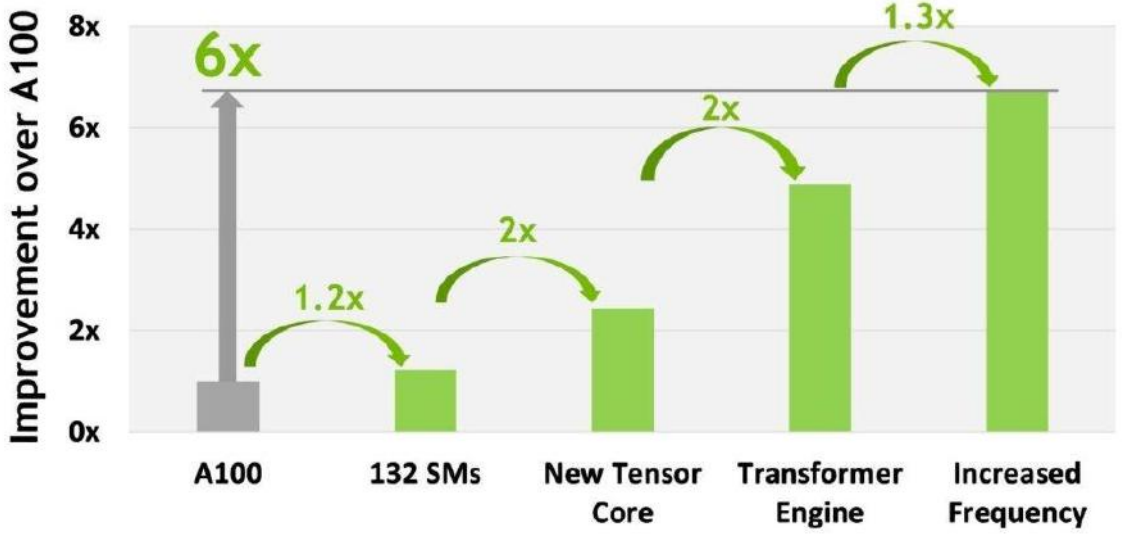
- **NVIDIA A100，作为Ampere架构首发的NVIDIA A100 ，相较Tesla V100 可提高2.5 倍HPC 运算量，单片A100单价约为8万元。**
  - DGX A100系统单节点的峰值性能为：INT8 10 PetaOPS(每秒1亿亿次整数运算)、FP16 5 PFlops(每秒5千万亿次半精度浮点运算)、TF32 2.5 PFlops(每秒2.5千万亿次运算)、FP64 156 TFlops(每秒156万亿次双精度浮点运算)。**相比于高端CPU服务器，它的AI计算性能要高出150倍，内存带宽高出40倍，IO带宽也高出40倍。**
- **H100 所结合的技术创新，可加速大型语言模型速度，比A100快30 倍，提供领先业界的对话式人工智能加速（类似ChatGPT）。**
  - H100于2022年3月22日推出，配备了 80 GB 显存，搭载 NVIDIA Hopper HPC GPU，采用台积电4nm制程，单价超过20万元。
  - H100 配备第四代Tensor核心和具有FP8精确度的Transformer Engine，能够依据动态管理与选择FP8与FP16，并自处理模型每一层FP8与FP16的自动转换，相对现行的A100架构，能使AI训练提升9倍、并使推理能提升30倍，同时不影响精确性。
- **目前华硕、源讯、戴尔、INGRASYS、技嘉、联想与美超微（Supermicro）等NVIDIA的众多合作伙伴推出搭载A100/H100产品，已在 AWS、Google Cloud、Microsoft Azure 及 Oracle Cloud Infrastructure 等各大云端平台上使用。**

英伟达基于CUDA 架构开发的历代GPU微架构

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/ST Unit+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108 SMs每个SM 64 个FP32 64 个INT32 32 个FP64 4个Tensor core	H100 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单元拥有更多的逻辑控制电路	NVLink一代，双向互联带宽160GB/s P100有56个SM HBM	Nvlink 2.0 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 MIG 1.0	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 80亿晶体管	16nm 153亿晶体管	12nm 211亿晶体管	12nm 186亿晶体管	7nm 283亿晶体管	4nm 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：智东西微信公众号

NVIDIA H100相较于A100的6倍性能提升



资料来源：英伟达官网

壁仞科技（未上市）

- **壁仞科技创立于2019年，团队由国内外芯片和云计算领域核心专业人员、研发人员组成，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和独到的行业洞见。**
  - 从发展路径上，壁仞科技将首先聚焦云端通用智能计算，逐步在人工智能训练和推理、图形渲染等多个领域赶超现有解决方案，实现国产高端通用智能计算芯片的突破。
- **公司共计13位核心高管，其中技术研发线的9位核心技术高管中，其中四位来自AMD公司，四位来自英伟达，其他高管也都来自高通、英特尔等国际芯片大厂。**
  - 包括前AMD全球副总裁李新荣；原华为海思GPU 首席架构师、英伟达GPU资深架构师洪洲；曾创建高通公司骁龙GPU团队、领导了5代Adreno GPU架构开发的首席架构师焦国方；原英特尔软件研发负责人、AMD软件工程负责人梁刚；原AMD GPU芯片研发负责人陈文中；原英伟达中国研发中心总经理、台积电设计与技术平台负责人杨超源；原AMD GPU SoC负责人张凌岚；原阿里云AI &GPU负责人、英伟达GPU架构师等。

壁仞科技的发展历程

2019年9月 壁仞科技成立

2020年8月 壁仞科技完成Pre-B轮融资，累计融资近20亿元人民币
2020年7月 壁仞科技签约上海人工智能重点项目
2020年6月 壁仞科技完成A轮融资，总额达11亿元人民币，创近年芯片设计领域新纪录

2021年9月 正式宣布图形GPU产品线
2021年5月 与复旦大学共建联合实验室
2021年4月 与清华大学签约专项研究合作
2021年3月 壁仞科技完成B轮融资，累计融资额超47亿元人民币
2021年3月 壁仞科技与上海交通大学合作成立“智能芯片与生态联合实验室”

2022年8月 发布首款通用GPU芯片，创全球算力新纪录
2022年3月 壁仞科技点亮国内算力最大通用GPU芯片

资料来源：壁仞科技官网，中信证券研究部

壁仞科技的和核心高管及研发负责人

姓名	职务	个人经历
张文	创始人、董事长、CEO	哈佛大学法学博士、哥伦比亚工商管理硕士。美国纽约执照律师，曾担任高级律师和私募基金总经理等要职，曾任职于商汤科技并担任总裁，还是L4自动驾驶方案公司云骥智行的董事长。
洪洲	联合创始人&CTO	北京大学数学学士，清华大学管理硕士与纽约州立大学布法罗分校数学及计算机科学硕士。**原华为海思GPU 首席架构师、英伟达GPU资深架构师。**曾担任海思自研GPU的负责人与主架构师，组建了完整的GPU团队并成功流片了全球领先且拥有自主IP的GPU芯片。
张凌岚	联合创始人、COO	**原AMD GPU SoC负责人，原海光海外GPU部门副总裁**
徐凌杰	联合创始人、总裁	**原阿里云AI &GPU负责人、英伟达GPU架构师**
焦国方	联合创始人，图形GPU产品线总经理	具有超过25年精深的GPU产品架构和研发经验，曾在高通任职11年，**曾创建高通公司骁龙GPU团队**、领导了5代Adreno GPU架构开发的首席架构师。**原华为鸿蒙OS图形图像处理和UI系统框架首席科学家。**
李新荣	联席CEO	**前AMD全球副总裁。**在GPU领域拥有超过30年的丰富经验，加入壁仞科技之前在AMD就职15年，担任全球副总裁、中国研发中心总经理，负责AMD大中华区的研发建设和管理工作。
杨超源	副总裁兼董事长特别助理	毕业于加州大学伯克利分校电子工程专业，在GPU芯片行业拥有超过35年的产品研发与管理经验。**原英伟达中国研发中心总经理、台积电设计与技术平台负责人、英伟达上海总经理**
陈文中	高级副总裁	在GPU行业拥有超过25年的研发与团队管理经验，**此前曾在AMD、S3和Trident等知名GPU企业领导核心产品开发团队。**就职AMD期间，他领导一支规模近500人的技术团队，在8年内实现了9款芯片的流片与量产，其中包括首款采用HBM技术的GPU芯片。
唐杉	研究院执行院长	EDA软件巨头Synopsys（新思科技）前AI Lab负责人

资料来源：公司官网，集微网，凤凰网，中信证券研究部整理

壁仞科技（未上市）

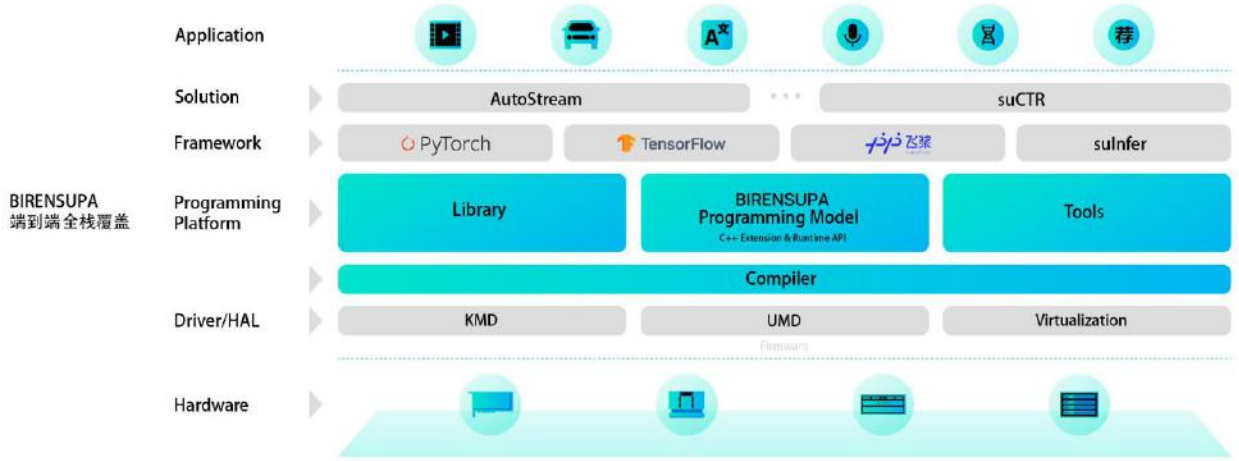
- **主要产品：BR100系列通用GPU芯片，目前BR100系列拥有BR100、BR104两款芯片**
  - 这款芯片采用台积电7nm制程工艺，集成了PCIE Gen5及HBM2e等最新周边IP，采用了3D立体堆叠技术等多种技术组合。
  - BR100对比英伟达在售的旗舰GPU峰值算力在Int8、BF16、TF32/TF32+、FP32数据格式下最少有3.3倍的峰值性能优势，在FP32数据格式下性能优势更是达到了13.1倍。针对人工智能（AI）训练、推理，及科学计算等更广泛的通用计算场景开发，主要部署在大型数据中心。
- **生态方面：目前BR100兼容当前主流软件生态包括CUDA，目的是保证用CUDA写的程序无缝运行在壁仞科技的异构计算开发平台上；终极目标是提供比CUDA更好的自研编程模型。**
  - 公司目前推出自主原创架构“壁立仞”，以及自主研发的BIRENSUPA软件平台，使得BR100在给定的工艺下实现了性能和能效的跨越式进步。

壁仞科技BR100系列通用GPU算力产品规格与NVIDIA 产品对比

		BR100-OAM	BR104-300W Pcle	NVIDIA H100 SXM5	NVIDIA H100 PCIe	NVIDIA A100 PCIe
	产品形态	OAM	双FHFL PCIe	PCIe	Pcle	Pcle
峰值性能	FP32	256 TFLOPS (Tensor Core)	128 TFLOPS (Tensor Core)	60 TFLOPS	48 TFLOPS	19.5 TFLOPS
峰值性能	TF32+(TF32)	512 TFLOPS	256 TFLOPS	500 TFLOPS(TF32)	400 TFLOPS(TF32)	156 TFLOPS(TF32)
峰值性能	BF16	1024 TFLOPS	512 TFLOPS	1000 TFLOPS	800 TFLOPS	312 TFLOPS(FP16)
峰值性能	INT8	2048 TOPS	1024 TOPS	2000 TOPS	1600 TOPS	624 TOPS
显存		64GB HBM2E	32GB HBM2E	80GB HBM3	80GB HBM2E	80GB HBM2E
接口		PCIe 5.0 支持 CXL2.0	PCIe 5.0 支持 CXL2.0	SXM5	PCIe 5.0	PCIe 4.0
互连带宽		512 GB/s	192 GB/s	900 GB/s	600 GB/s	600 GB/s
最大热设计功耗 (TDP)		550W	300W	700w	350W	400W

资料来源：壁仞科技官网，芯东西，中信证券研究部 注：NVIDIA H100仅支持未采用稀疏技术的规格

BIRENSUPA软件开发平台



资料来源：壁仞科技官网

沐曦（未上市）

- **沐曦成立于2020年9月，其研发的高性能GPU芯片可应用于AI推理、AI训练、高性能数据分析、科学计算、数据中心、云游戏、自动驾驶、元宇宙等众多需要高算力的前沿领域。**
- **沐曦的核心技术团队拥有平均20年的高性能GPU产品的设计经验和技术实力，三位创始人均来自AMD公司，曾完整负责10多款世界顶尖高性能GPU产品开发及应用落地，拥有丰富的5nm流片和7nm芯片量产经验。**
  - 其中，公司CEO陈维良曾在AMD任全球GPU SoC设计总负责人，拥有团队管理能力和量产经验；硬件首席架构师彭莉是AMD全球首位华人女科学家（Fellow），曾任AMD首席架构师；软件首席架构师杨建是AMD大中华地区第一位科学家(Fellow)，历任AMD、海思等首席架构师，拥有20年大规模芯片及GPU软硬件设计经验。
- **产品方面，沐曦第一颗高性能通用GPU芯片MXN于2022年1月顺利流片，采用7nm工艺，以AI推理为主，公司预计2023年量产。第二款主要用于科学计算、AI训练、数据中心弹性计算的旗舰GPU芯片MXC的研发也进入收尾阶段，公司计划2023年会进入量产。**

公司主要产品线及介绍

产品	MXN（曦思）	MXC（曦云）	MXG（曦彩）
定位	高性能人工智能推理加速处理器	高性能通用计算加速处理器（GPGPU）	高性能图形加速处理器
应用场景	MXN系列是面向云端数据中心应用的人工智能推理产品，采用先进工艺结合高带宽内存，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系列通用GPU(GPGPU)芯片是针对AI训练和推理及科学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、数据中心以及科学计算、教育和科研等场景。	MXG系列高性能GPU是面向云端图形渲染的GPU产品
应用场景	AI推理、数据中心	AI推理、AI训练、高性能数据分析、科学计算、数据中心	图形渲染
制程	7nm	5nm	
量产时间	2023年	2023年	

资料来源：沐曦官网，中信证券研究部

沐曦发展历程

2020/9/14 沐曦注册成立
2020/10/28 完成天使轮融资，由和利资本、泰达科投联合领投
2020/12/18 临港“十四五”AI重点项目签约
2020/12/23 浙江大学-沐曦联合研发中心成立

2021/1/5 完成Pre-A轮融资，由红杉中国、真格基金联合领投
2021/2/5 完成Pre-A+轮融资，由经纬中国、光速中国联合领投
2021/6/8 完成A轮10亿元融资，由国调基金、中网投联合领投
2021/6/26 获得中国”芯力量”评选双奖
2021/10/26 与清华国际、浪潮共建联合实验室

2022/5/9 获评上海市科技型中小企业
2022/7/5 完成Pre-B 轮10亿元融资
2022/9/2 与UCloud优刻得、脑虎科技和图灵量子，共同发起成立“先进智算联盟”

资料来源：沐曦官网，中信证券研究部

芯动科技（未上市）

- **芯动科技2021年发布了国产显卡GPU——“风华1号”，面向桌面、服务器市场，采用12nm制造**
  - 芯动科技采用Imagination最新推出的IMG B系列BXT高性能多核图形处理器（GPU）IP。
  - 风华1号芯片的**FP32 浮点算力为5T FLOPS**；渲染能力为160G Pixel/s；编解码能力：同时4路4K60帧，16路1080P60帧，32路720P30帧；**AI 计算的算力为25TOPS（INT8）**。
  - 适配方面，风华1号支持Windows、Android、Linux(含国产)等操作系统，支持ARM、MIPS、x86 CPU架构，支持**OpenGL、OpenGL ES、OpenCL、Vulkan、DirectX等主流架构**，支持嵌入式VR/AR/AU、智能座舱、工控机等应用。
- **技术团队：**公司的技术研发团队包括**前AMD图形框架开发领军人物**，现任芯动DX团队负责人张涛；杨喜乐博士在**英国Imagination公司做了25年的架构师**，现担任芯动首席算法科学家，是全球GPU芯片领域从几何物理渲染到计算引擎领域的知名专家，持有GPU 3D计算机图形学核心领域顶级图形专利共计125项，与Imagination、苹果等公司最新的核心GPU产品的设计、优化、迭代相关。

风华1号数据中心GPU介绍

风华1号数据中心GPU计算性能

渲染能力	160 GPixel/s
单精度浮点性能	5TFLOPS
AI性能	25 TOPS （INT8）
系统接口	PCIe 4.0x16,向下兼容PCIe 3.0/2.0
显示接口	HDMI 2.0/DP1.4/VGA多路独立输出
视频编解码	4路4K@60fps，16路1080P@60fps，32路720P@60fps，低延迟硬件编码
图形API支持	OpenGL 4.3，OPENGL ES3.2，VULKAN1.2，DirectX11/12 (2022发布)
显存类型	GDDR6 /GDDR6X （最大速率19Gbps）
显存带宽	最大304GB/s,动态调整
显存容量	4GB / 8GB / 16GB
计算 API	OpenGL1.2/2.1EP/3.0

资料来源：芯动科官网

数据来源：芯动科官网

摩尔线程（未上市）

- **摩尔线程创始人是前英伟达公司全球副总裁、中国区总经理James Zhang，2005年加入英伟达前在惠普、戴尔工作过。公司团队集聚了很多顶尖公司的GPU人才，核心成员主要来自NVIDIA，吸引了Microsoft、Intel、 AMD、ARM、华为、平头哥等各大科技公司的研发人员。**
- **2022年3月，摩尔线程推出MUSA架构**
  - MUSA是摩尔线程产品系列采用的统一系统架构，包括**统一的编程模型、软件运行库、驱动程序框架、指令集架构和芯片架构**。并基于MUSA统一系统架构打造的第一代摩尔线程多功能GPU芯片核心——苏堤。
- **摩尔线程基于MUSA统一系统架构苏堤核心晶片打造的数据中心级多功能GPU产品MTT S2000**
  - MTT S2000采用**12nm**制程，使用**4096个MUSA核心**，最大配置**32GB**显存，FP32单精度算力最高可达到**12TFlops**，支持H.264、H.265、AV1多路高清视频编解码，以及广泛的AI模型算法加速，支持PyTorch、Tensorflow、PaddlePaddle等主流深度学习框架。

MUSA统一系统架构组成

资料来源：摩尔线程官网

MTT S2000结构组成

资料来源：摩尔线程官网

天数智芯（未上市）

- **天数智芯于2020年12月成功点亮国内第一款7nm云端训练通用GPU产品“天垓100”，并于2021年3月正式对外发布。目前已经实现大规模量产和销售，截止至2022年3月底，已实现销售订单近2亿元，并且帮助客户落地了两百多个应用场景。**
  - 天垓100芯片采用**全自研的架构、计算核、指令集及基础软件栈，不受国外IP制约。**
  - 内建FP32/FP16/BF16/INT多种数据类型指令，支持混合精度AI训练。
  - 支持主流的深度学习开发框架，**兼容主流GPU的编程模式**，有效对接现有软件生态，易于扩展支持新的算法与应用领域。
- **天数智芯的第二款产品7nm云边推理芯片“智铠100”在2022年5月成功点亮，目前在开发第二三代AI训练芯片天垓200及300。**
- **天数智芯CTO吕坚平毕业于耶鲁大学并获计算机科学博士学位，曾任三星全球副总裁、联发科资深总监、英伟达全球资深GPU架构师、高级架构经理等职务，拥有近30年芯片研发技术经验；负责芯片设计的副总裁，是一位高端计算芯片设计专家，拥有近30年的处理器、微处理器、GPU研发和管理经验，在近15年内，他主导参与了AMD所有服务器、GPU，APU产品的IP设计，包括最新7nm EPYC和7nm GPU产品。**

天数智芯“天垓100”性能汇总

架构	GPGPU
制程及封装	TSMC 7nm FinFET
	2.5D COWOS 封装
内存规格	32 GB DRAM （4*8GB）HBM2
散热规格	板级功耗250W
	全高全长双槽位主被动式散热
接口规格	PCIe Gen4.0 x 16 lane
	共享 64 GB/s 主控双向带宽
	共享 64 GB/s 片间互联带宽
性能	37 TFLOPS@FP32
	147 TFLOPS@FP16/BF16
	295 TOPS@INT8
	支持 INT32，INT16 计算
	多精度数据类型支持标准/混合训练

资料来源：天数智芯官网，中信证券研究部

天数智芯自主IP特征

资料来源：芯东西微信公众号

登临科技（未上市）

- **登临成立于2017年底，专注于高性能通用计算平台的芯片研发与技术创新，致力于打造云边端一体、软硬件协同、训练推理融合的前沿芯片产品和平台化基础系统软件。**
  - 公司自主创新的GPU+（基于GPGPU的软件定义的片内异构计算架构），在兼容CUDA/OpenCL在内的编程模型和软件生态的基础上，通过架构创新，完美解决了通用性和高效率的双重难题。大量客户产品实测证明，针对AI计算，GPU+相比传统GPU在性能尤其是能效上有显著提升。
  - 登临科技致力于推动国产化AI解决方案在各行各业的发展和落地，通过建立完善的软硬件合作生态体系，全面助力产业数字化转型和智能化升级和改造。
- **作为国内首个实现规模化商业落地的GPU企业，登临首款基于GPU+的创新AI计算加速器-Goldwasser已规模化运用在边缘至云计算的各个应用场景，成功填补了国内高性能GPGPU领域技术、产品及商业方面的空白。**
- **登临第二代产品将于2023年上半年进入市场，其能效比将是第一代产品的2倍，在同等功耗下，峰值算力达到国际主流产品的2倍。**

公司核心技术亮点

资料来源：登临科技官网

公司核心产品性能对比

登临科技 Goldwasser		Goldwasser UL	Goldwasser L	Goldwasser XL
功耗		25-35W	40-70W	200W
性能	INT8	32-64TOPS	128-256TOPS	512TOPS
	FP16	8-16TFLOPS	32-64TFLOPS	128TFLOPS
视频解码(H264/H265)		32-64路1080P@30FPS	128-256路1080P@30FPS	256路1080P@30FPS
内存		2-16GB	16-64GB	16-64GB
PCIe		Gen3x2-4	Gen3x8-16	Gen3x16
应用领域		边缘计算	边缘计算、数据中心	数据中心

资料来源：登临科技@2021年世界人工智能大会（WAIC）

风险因素

- 用户拓展不及预期风险；
- AI技术及新产品开发发展不及预期风险；
- 外部制裁加剧风险；
- 宏观经济需求下行风险。

中信证券
CITIC SECURITIES

感谢您的信任与支持！

THANK YOU

雷俊成（半导体分析师）

执业证书编号： S1010520050003

王子源（半导体分析师）

执业证书编号： S1010521090002

徐涛（科技产业联席首席分析师/电子行业首席分析师）

执业证书编号：S1010517080003

杨泽原（计算机行业首席分析师）

执业证书编号：S1010517080002

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%：澜起科技（688008），对应持股业务类别：自营，持股比例：1.32%；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：0.33%；另类投资子公司，限售持股比例：1.34%/0.25%，限售起始日：2022年08月12日/2022年08月12日，限售期：12个月/24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。