证券研究报告

行业报告：行业专题研究

2024年08月07日

计算机

算力知识普惠系列一：AI芯片的基础关键参数

作者：

分析师 缪欣君 SAC执业证书编号：S1110517080003

联系人 刘鉴

天风证券 TF SECURITIES

行业评级：强于大市（维持评级）

上次评级：强于大市

摘要

算力是衡量计算机处理信息能力的重要指标，其中AI算力专注于AI应用，常见单位为TOPS和TFLOPS，通过GPU、ASIC、FPGA等专用芯片提供算法模型训练和推理。算力精度作为衡量算力水平的一种方式，其中FP16、FP32应用于模型训练，FP16、INT8应用于模型推理。

AI芯片通常采用GPU和ASIC架构。GPU因其在运算和并行任务处理上的优势成为AI计算中的关键组件，它的算力和显存、带宽决定了GPU的运算能力。GPU的核心可分为Cuda Core、Tensor Core等；**Tensor Core是增强AI计算的核心，**相较于并行计算表现卓越的Cuda Core，它更专注于深度学习领域，通过优化矩阵运算来加速AI深度学习的训练和推理任务，其中Nvidia Volta Tensor Core架构较Pascal架构（Cuda Core）的AI吞吐量增加了12倍。**此外，TPU作为ASIC的一种专为机器学习设计的AI芯片，**相比于CPU、GPU，其在机器学习任务中的高能效脱颖而出，其中TPU v1在神经网络性能上最大可达同时期CPU的71倍、GPU的2.7倍。

建议关注：

1）四小龙：寒武纪、海光信息、神州数码、中科曙光

2）华为： 软通动力、烽火通信、广电运通、拓维信息

风险提示：AI算力景气度下降的风险、AI芯片竞争加剧的风险、技术发展风险

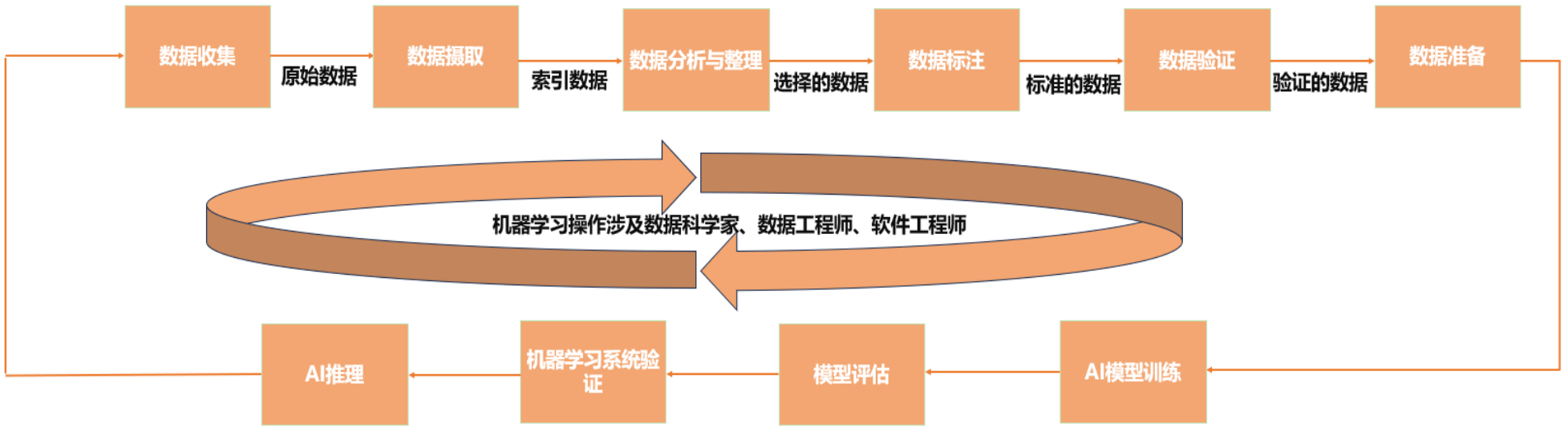
1
算力基础

1、AI计算的生命周期

AI计算是一种计算机器学习算法的数学密集型流程，通过加速系统和软件，从大量数据集中提取新的见解并在此过程中学习新能力。

AI计算的三个主要过程包括：1）提取/转换/加载数据（ETL）：数据科学家需要整理和准备数据集。2）**选择或设计AI模型：**数据科学家选择或设计最适合其应用的AI模型，一些公司会从一开始就设计并训练自己的模型，另一些公司可能采用预训练模型并根据需求进行自定义。3）**AI推理：**企业通过模型对数据进行筛选，AI在此过程中提供可行的洞察与见解。

图：AI计算的生命周期



1、算力及AI算力主要芯片的分类

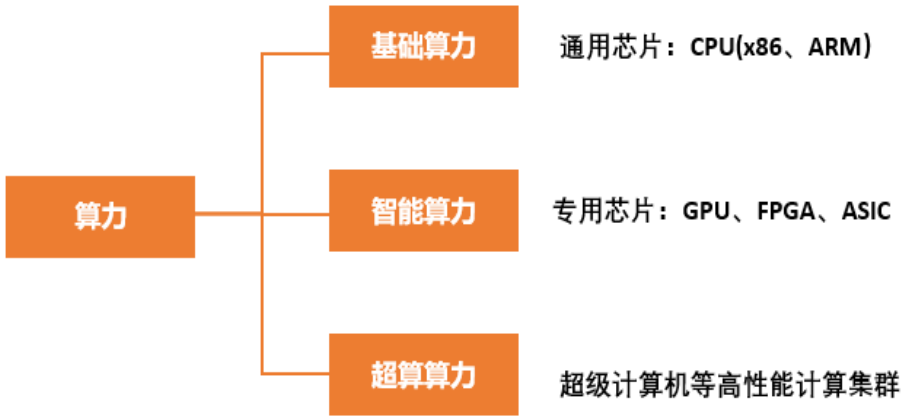
算力通常是指计算机处理信息的能力，特别是在进行数学运算、数据处理和执行程序时的速度和效率。根据使用设备和提供算力强度的不同，算力可分为：基础算力、智能算力、超算算力。智能算力即AI算力，是面向AI应用，提供AI算法模型训练与模型运行服务的计算机系统能力，其算力芯片通常包括GPU、ASIC、FPGA、NPU等各类专用芯片。

1) **基础算力：**由基于CPU芯片的服务器所提供的算力，主要用于基础通用计算，如移动计算和物联网等。日常提到的云计算、边缘计算等均属于基础算力。

2) **智能算力：**基于GPU（图像处理器）、FPGA（现场可编程逻辑门阵列）、ASIC（专用集成电路）等AI芯片的加速计算平台提供的算力，主要用于AI的训练和推理计算，比如语音、图像和视频的处理。

3) **超算算力：**由超级计算机等高性能计算集群所提供的算力，主要用于尖端科学领域的计算，比如行星模拟、药物分子设计、基因分析等。

图：算力的主要分类



图：AI算力芯片的主要分类

AI算力芯片主要分类	定义
GPU	图形处理单元，是一款专门的图形处理芯片，做图形渲染、数值分析、金融分析、密码破解，以及其他数学计算与几何运算。
FPGA	现场可编程阵列，根据用户的需要，在制造后进行无限次数的重复编程来实现数字逻辑功能，是可以重构的芯片。
ASIC	应特定用户要求或特定电子系统的需求，专门设计、制造的集成电路，是一种专用于特定任务的芯片。

资料来源：中国科学院半导体研究所公众号、《2023年智能算力发展白皮书》、珠海组工公众号等，天风证券研究所

1、算力的常见单位

在计算机领域，常用算力的衡量指标包括FLOPS（每秒浮点运算次数）、OPS（每秒运算次数）。FLOPS特别适用于评估超级计算机、高性能计算服务器和GPU等设备的计算性能。

在计算性能的度量中，常见单位包括Kilo/Mega/Giga/Tera/Peta/Exa，算力通常以 PetaFlOPS（每秒千万亿次浮点运算）单位来衡量。

AI 算力常见单位分为TOPS和TFLOPS。推理算力，即通常用设备处理实时任务的能力，通常以TOPS（每秒万亿次操作）为单位来衡量。而训练算力，即设备的学习能力和数据处理能力，常用TFLOPS（每秒万亿次浮点操作）来衡量。TFLOPS数值越高，反映了模型在训练时的效率越高。

图：算力的通常计量单位

资料来源：联想智库公众号、石河子市政府服务和大数据局公众号，天风证券研究所

1、不同场景对应算力精度表示不同

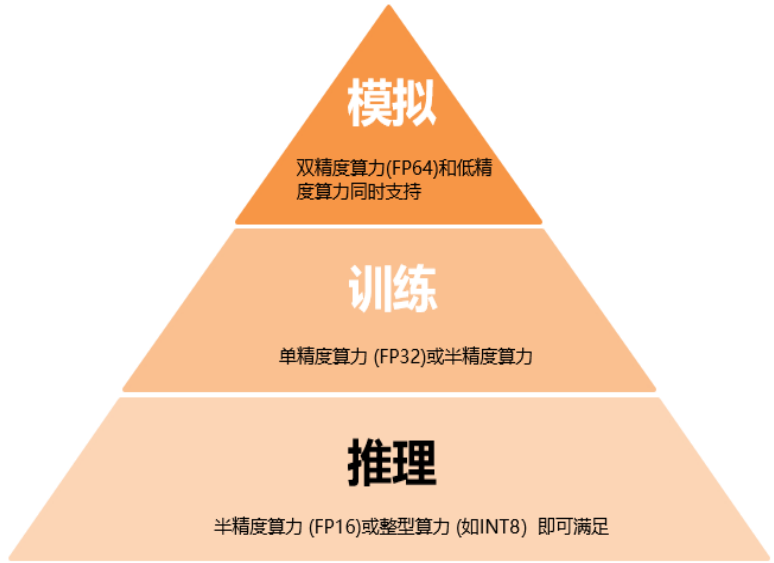
算力精度作为可以衡量算力水平的一种方式，可分为浮点计算和整型计算。其中浮点计算可细分为半精度（2Bytes，FP16）、单精度（4Bytes，FP32）和双精度（8Bytes，FP64）浮点计算，加上整型精度（1Byte，INT8）。

不同场景对应算力精度表示不同。FP64主要用于对精度要求很高的科学计算，如制造产品设计、机械模拟和Ansys应用中的流体动力学，AI训练场景下支持FP32和FP16，模型推理阶段支持FP16和INT8。

表：常见浮点/整型规格及定义

常见浮点/整型规格	定义
FP64	被称为双精度，它是使用64位(8 Bytes)来表示一个浮点数，精度较高，常用于科学计算和对精度要求较高的场景。
FP32	被称为单精度，使用32位(4 Bytes)来表示一个浮点数，精度略逊于FP64，仍然足够用于大多数AI训练任务。
FP16	被称为半精度，它的精度极低，但占用存储空间和计算资源较少，对精度要求低，通常运用在节能场景下。
INT8	是一种低精度、高效率的数值表示方式。在推理阶段(即模型训练完成用于实际应用的阶端),使用它可以显著提高运算速度，降低能耗。

图：不同精度可执行任务对比



资料来源：网络安全和信息化公众号、九维数智公众号，天风证券研究所

1、稀疏算力和稠密算力

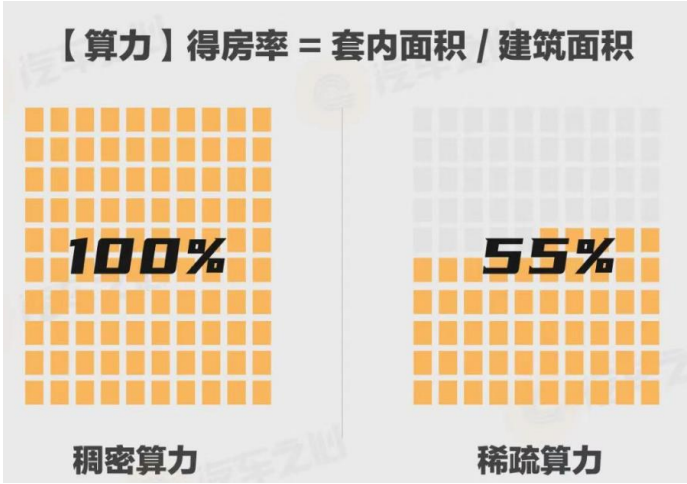
稀疏算力和稠密算力用于描述计算资源的利用程度。在实际场景中，稀疏算力和稠密算力存在互补关系与转换关系。

- **稠密算力：**指的是在计算过程中，数据点之间的管理都较高，需要处理大量连续的数据。通常用于需要密集型计算的任务，如图像处理、视频编码、大规模数值模拟等
- **稀疏算力：**指在计算过程中，数据点之间的关联度较低，数据分布稀疏。这种算力常用于处理稀疏矩阵或者稀疏数据集，如社交网络分析、推荐系统、基因序列分析等。

表：稠密算力与稀疏算力特性对比

特性	稠密算力	稀疏算力
数据关联度	高	低
数据存储	连续存储	稀疏存储、使用特殊数据结构
计算模式	并行计算	稀疏优化算法，可能包含并行计算
应用场景	图像处理、科学计算	社交网络分析、推荐系统
能耗	相对较高	相对较低
算法优化	针对大规模数据处理优化	针对稀疏数据处理优化
硬件需求	高性能CPU/GPU、大容量内存	优化的存储解决方案、可能需要特定硬件支持

图：稠密算力与稀疏算力结构对比



资料来源：九维数智公众号、汽车之心公众号，天风证券研究所

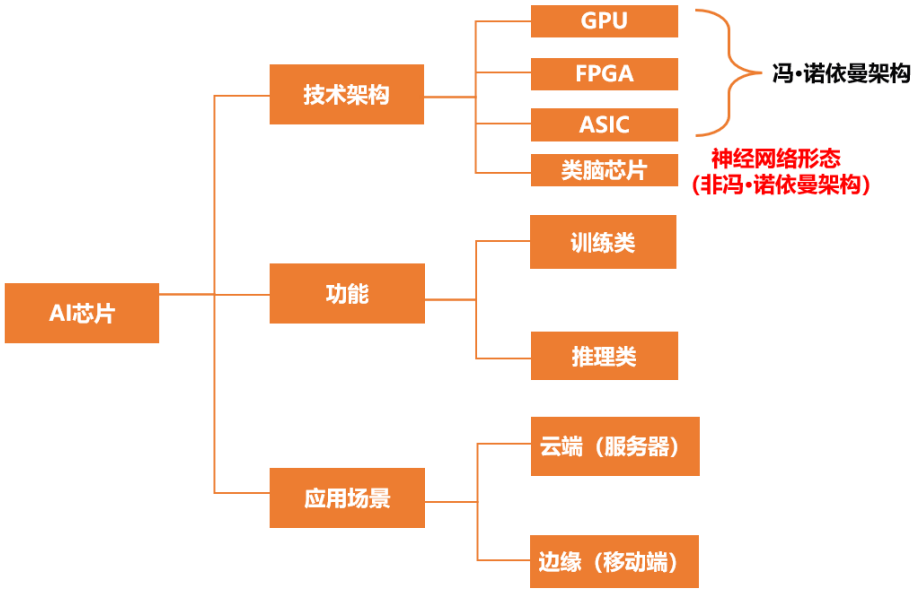
2
AI芯片架构与参数

2、AI芯片通常采用GPU与ASIC架构

目前通用的CPU、GPU、FPGA等都能执行AI算法，只是执行效率差异较大。但狭义上讲一般将AI芯片定义为“专门针对AI算法做了特殊加速设计的芯片”。AI芯片可以分为GPU、FPGA和ASIC架构，根据场景可以分为云端和端侧。和其他芯片相比，AI芯片重点增强了运行AI算法的能力。

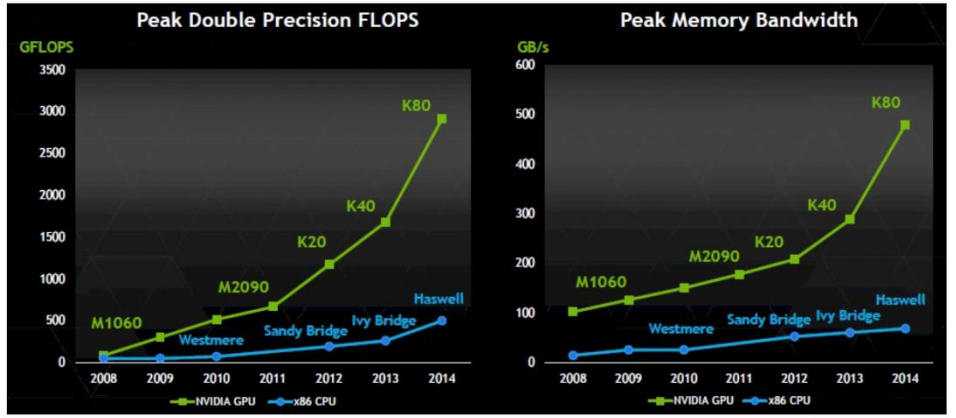
目前主流AI芯片为GPU和ASIC。国际上，Nvidia的H200 Tensor Core GPU以其卓越的计算性能和能效比领先市场，而Google的第六代TPU Trillium ASIC芯片则以其专为机器学习优化的设计提供高速数据处理。在国内，寒武纪的思元370芯片(ASIC)凭借其先进的计算处理能力在智能计算领域占据重要地位，已与主流互联网厂商开展深入适配; 海光信息的DCU系列基于GPGPU架构，以其类“CUDA”通用并行计算架构较好地适配、适应国际主流商业计算软件和AI软件。

图： AI芯片的分类



图：AI芯片在模拟场景和模型运行中具有显著的计算优势

(对比英伟达GPU和AMD x86 CPU)



资料来源：迪普微科技公众号，The NextPlatform官网，天风证券研究所

2、Tensor Core是增强AI计算的核心，能更好的处理矩阵乘运算

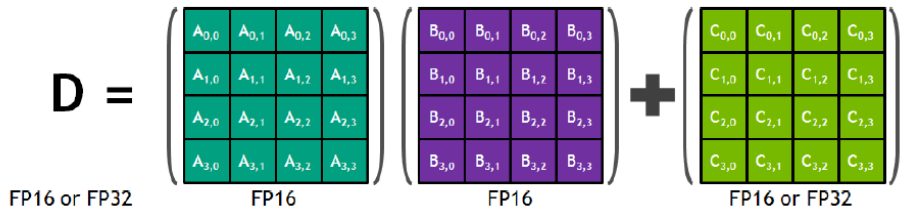
图： Nvidia 初代Tensor Core的Volta GV100 SM架构

Tensor Core是用于加速深度学习计算的关键技术，其主要功能是执行深度神经网络中的矩阵乘法和卷积计算。

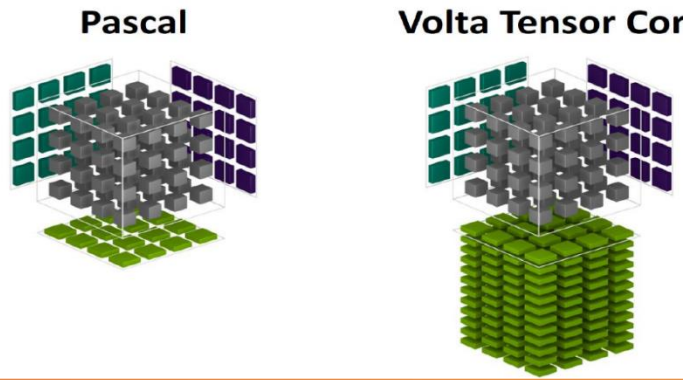
与传统CUDA Core相比，Tensor Core在每个时钟周期能执行多达4x4x4的GEMM运算，相当于同时进行64个浮点乘法累加（FMA）运算。

其计算原理是采用半精度（FP16）作为输入和输出（**矩阵Ax矩阵B**），并利用全精度（**矩阵C**）进行存储中间结果计算，以确保计算精度的同时最大限度地提高计算效率。

图：Tensor Core计算原理



图：Volta Tensor Core较Pascal架构的AI吞吐量增加了12倍



资料来源：《Nvidia Tesla V100 GPU Architecture》白皮书、摩崖芯公众号，天风证券研究所

2、AI芯片的硬件重点性能指标

AI芯片指标	定义
计算能力	GPU执行浮点运算的能力，通常以TFLOPs(每秒浮点操作次数)为单位衡量。高计算能力对科学计算、模拟和深度学习等计算密集型任务至关重要。它能加速模型训练、数据分析以及复杂模拟的处理速度。
显存	是GPU用于存储数据和纹理的专用内存，与系统内存(RAM)不同，显存具有更高的带宽和更快的访问速度。显存的大小和性能直接影响GPU处理大规模数据的能力。
功耗	即功率损耗，指单位时间内的能量消耗，反应消耗能量的速率，单位是瓦特(W)。
卡间互联	NVIDIA® NVLink™ 是世界首项高速 GPU 互连技术，与传统的 PCIe 系统解决方案相比，能为多 GPU 系统提供更快速的替代方案。NVLink 技术通过连接两块 NVIDIA® 显卡，能够实现显存和性能扩展，从而满足最大视觉计算工作负载的需求。
显存带宽	作为GPU与显存之间数据传输的桥梁；显存带宽=显存位宽x 显存频率

图：Nvidia H200 Tensor Core GPU规格

计算能力

显存

显存带宽

功耗

卡间互联

Technical Specifications	H200 SXM[1]	H200 NVL[1]
FP64	34 TFLOPS	34 TFLOPS
FP64 Tensor Core	67 TFLOPS	67 TFLOPS
FP32	67 TFLOPS	67 TFLOPS
TF32 Tensor Core[2]	989 TFLOPS	989 TFLOPS
BFLOAT16 Tensor Core[2]	1,979 TFLOPS	1,979 TFLOPS
FP16 Tensor Core[2]	1,979 TFLOPS	1,979 TFLOPS
FP8 Tensor Core[2]	3,958 TFLOPS	3,958 TFLOPS
INT8 Tensor Core[2]	3,958 TFLOPS	3,958 TFLOPS
GPU Memory	141GB	141GB
GPU Memory Bandwidth	4.8TB/s	4.8TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Confidential Computing	Supported	Supported
Max Thermal Design Power (TDP)	Up to 700W (configurable)	Up to 600W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @16.5GB each	Up to 7 MIGs @16.5GB each
Form Factor	SXM	PCIe
Interconnect	NVIDIA NVLink™: 900GB/s PCIe Gen5: 128GB/s	2- or 4-way NVIDIA NVLink bridge: 900GB/s PCIe Gen5: 128GB/s
Server Options	NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs	NVIDIA MGX™ H200 NVL partner and NVIDIA-Certified Systems with up to 8 GPUs
NVIDIA AI Enterprise	Add-on	Included

1. Preliminary specifications. May be subject to change.
2. With sparsity.

资料来源：英伟达官网、GPU算力公众号等，天风证券研究所

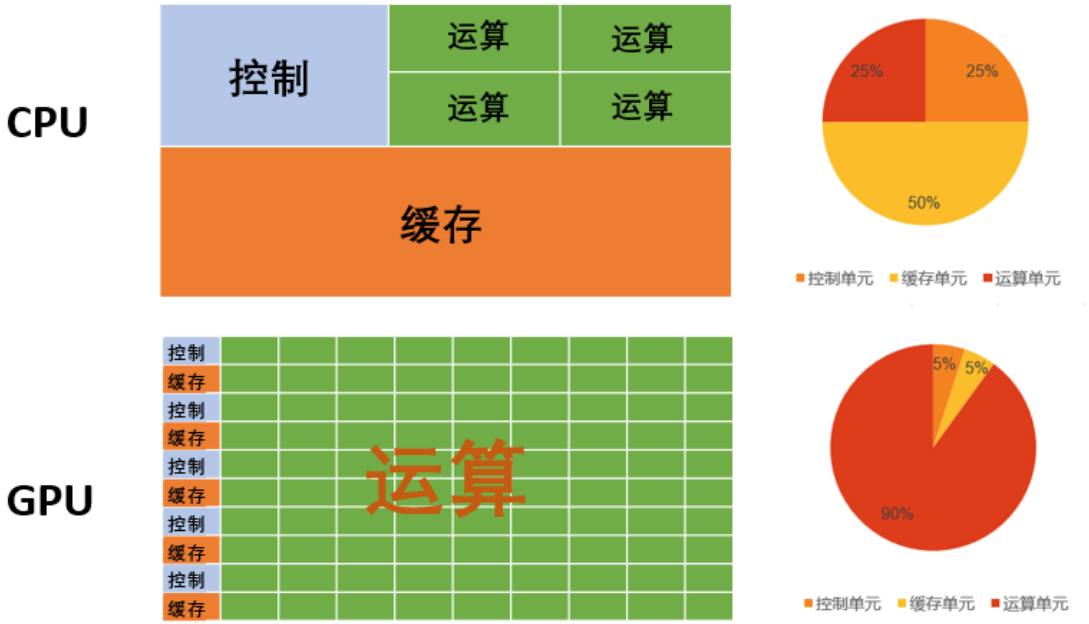
2、GPU在运算及并行任务处理能力上具有显著优势

图片处理器GPU又称显示核心、视觉处理器、显示芯片，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像运算工作的微处理器，是显卡或GPU卡的“心脏”。

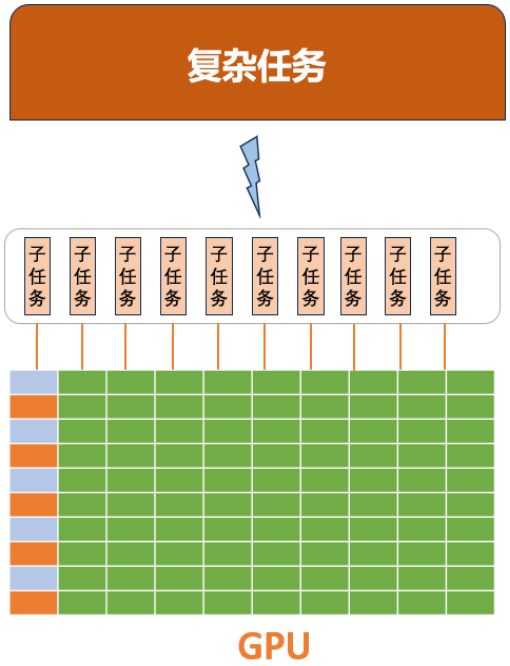
CPU和GPU在架构组成上都包括3个部分：运算单元（ALU）、控制单元（Control）、缓存单元（Cache）。从结构上看，在CPU中，缓存单元占50%，控制单元占25%，运算单元占25%；然而在GPU中，运算单元占90%比重，缓存、控制各占5%；由此可见，**CPU运算能力更加均衡，GPU更适合做大量运算。**

GPU通过将复杂的数学任务拆解成简单的小任务，并利用其多流处理器来并行处理，从而高效地执行图形渲染、数值分析和AI推理。

图：CPU与GPU基本组成单元对比



图：GPU将极为复杂的任务进行拆解并行处理



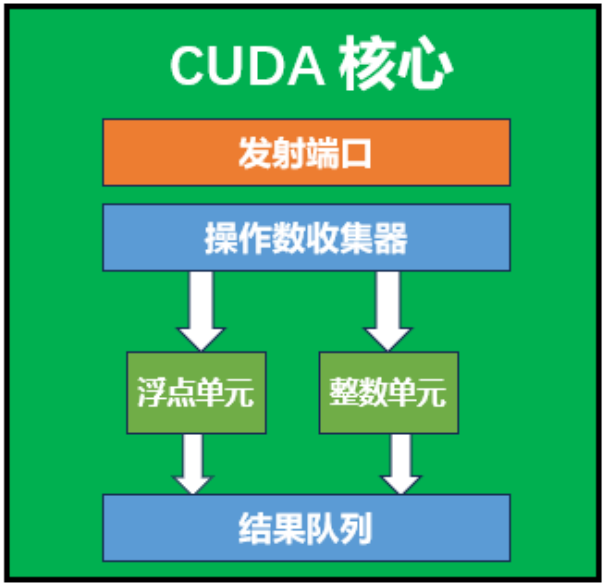
资料来源：书圈公众号，天风证券研究所

2、GPU核心分类及CUDA Core结构特点

通常GPU核心可分为三种：CUDA Core、Tensor Core、RT Core。

每个CUDA核心含有一个ALU(整数单元)和一个浮点单元，并且提供了对于单精度和双精度浮点数的FMA指令。

图：Cuda核心结构



表：通用GPU核心类型

GPU 核心类型	定义
CUDA Core	核心用于通用并行计算任务，可以执行FP32和FP64运算，以及整数运算，在处理广泛并行计算任务方面非常高效。
Tensor Core	针对深度学习和AI工作负载而设计的专用核心，擅长处理FP16和FP32的矩阵乘法和累加，在加速深度学习训练和推理中发挥重要作用。
RT Core	专门用于光线追踪处理的核心，能够高效进行光线和声音的渲染，对图形渲染和光线追踪等任务具有重要意义。

如果将GPU处理器比作玩具工厂，CUDA核心就是其中的流水线。流水线越多，生产的玩具就越多，虽然“玩具工厂”的性能可能会越好，但也受限于每个流水线的生产效率、生产设备的架构、生产存储资源能力等。**反应在GPU上，还需考虑显卡架构、时钟速度、内存带宽、内存速度、VRAM等因素。**

图：CUDA的核心数量并不能直接反映不同代GPU性能的好坏

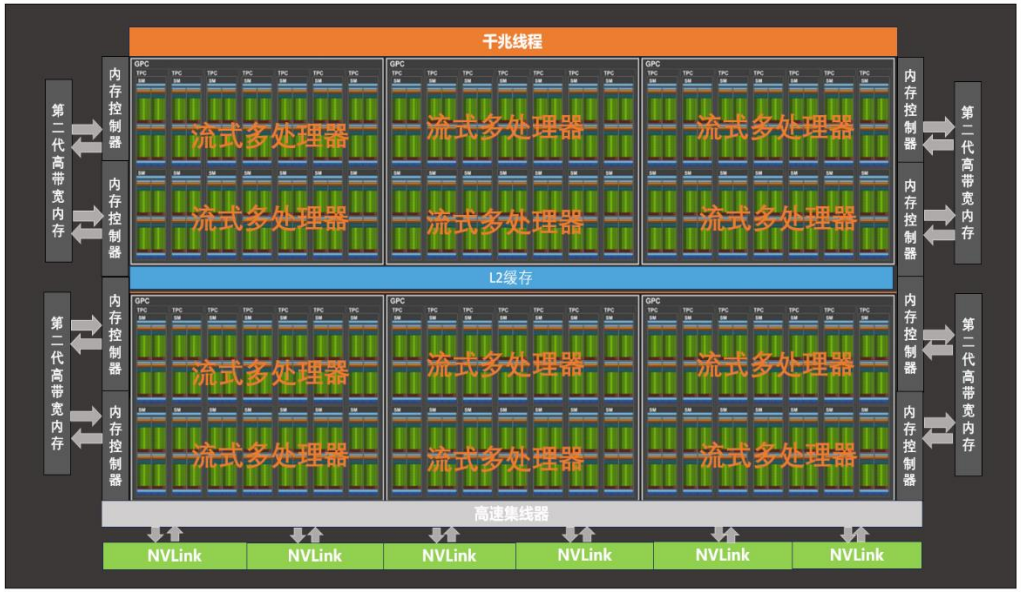
	Nvidia Geforce GTX 980 Ti	GTX GeForce GTX 1080
晶体管数量	8,100,000,000	7,200,000,000
CUDA 核心数量	2816	2560
晶体管数量/核心	2,876,420	2,812,500
时钟速度	1500兆赫	2000兆赫

2、GPU的架构及流式多处理器的结构组成

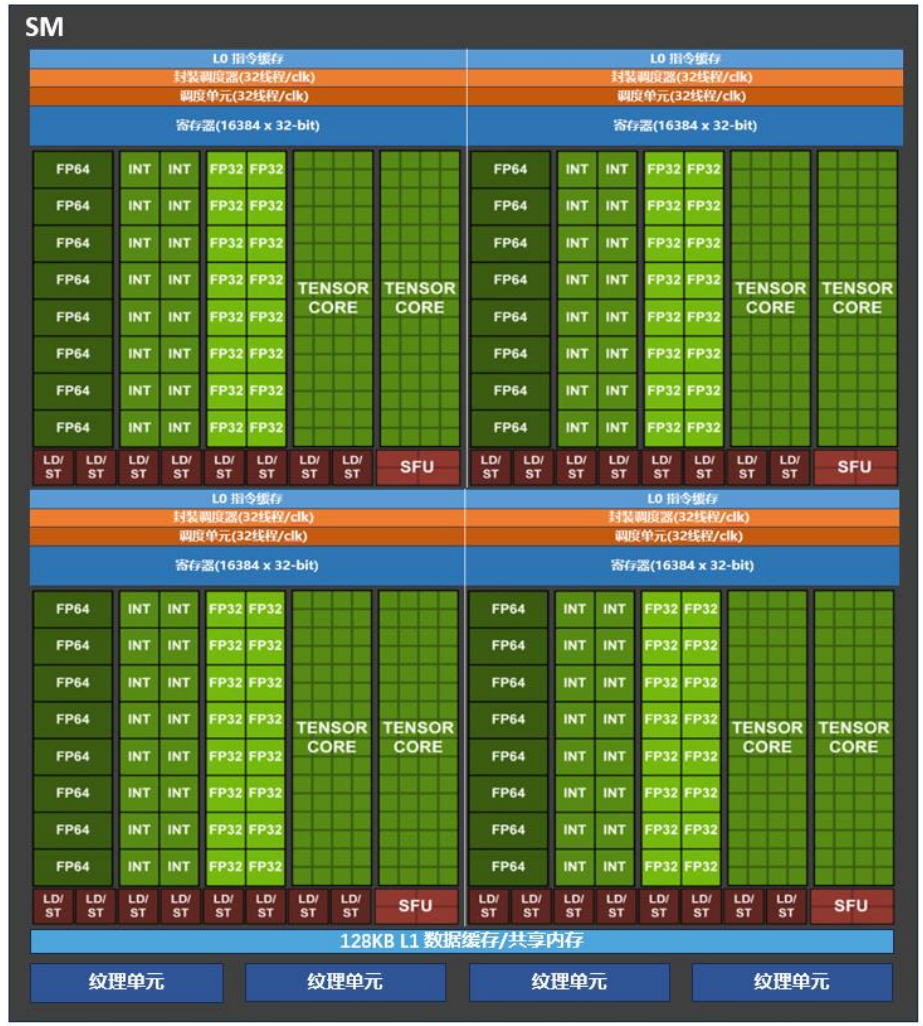
以Nvidia Volta架构的GV100为例，其主要组成部分可分为：

1）6个GPC（图像处理集群）：每个包含7个纹理处理集群（TPCs），每个TPC包括两个SM，共14个SM；**2）84个Volta SM（流式多处理器，见右图）：**每个包含8个Tensor Core、 64个FP32核心、64个INT32核心、32个FP64核心、4个纹理单元；**3）8个512位内存控制器（总共4096位）。**

图：GPU架构组成（以Nvidia GPU架构Volta GV100为参考）



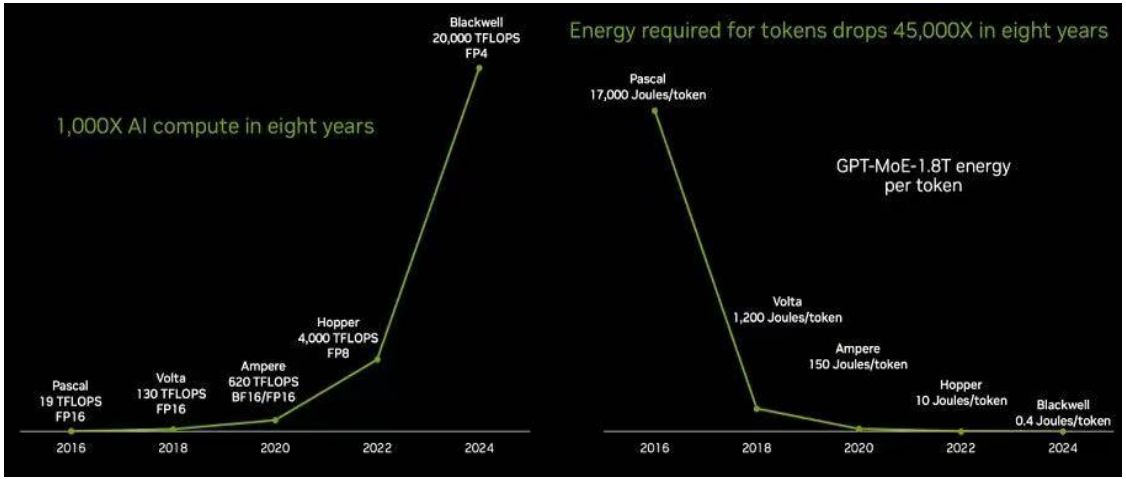
图：GPU的流式多处理器结构



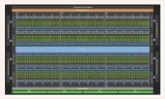
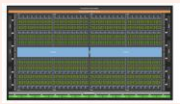
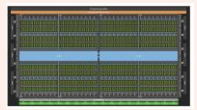
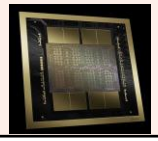
2、Nvidia AI芯片的技术演进

Nvidia的AI芯片在过去八年中实现了显著的技术进步。从“Pascal” P100 GPU到“Blackwell” B100 GPU，性能提升了1053倍。通过降低浮点精度（从FP16到FP4），实现了更高效的计算，同时每单位能耗显著下降，从P100的17000焦耳/token降低到B100的0.4焦耳/token。尽管GPU价格上涨了约7.5倍，但性能的大幅提升使得其在十天内训练1.8万亿参数的大模型成为可能。

图：B100 GPU较P100处理Token的能耗减少了45000倍

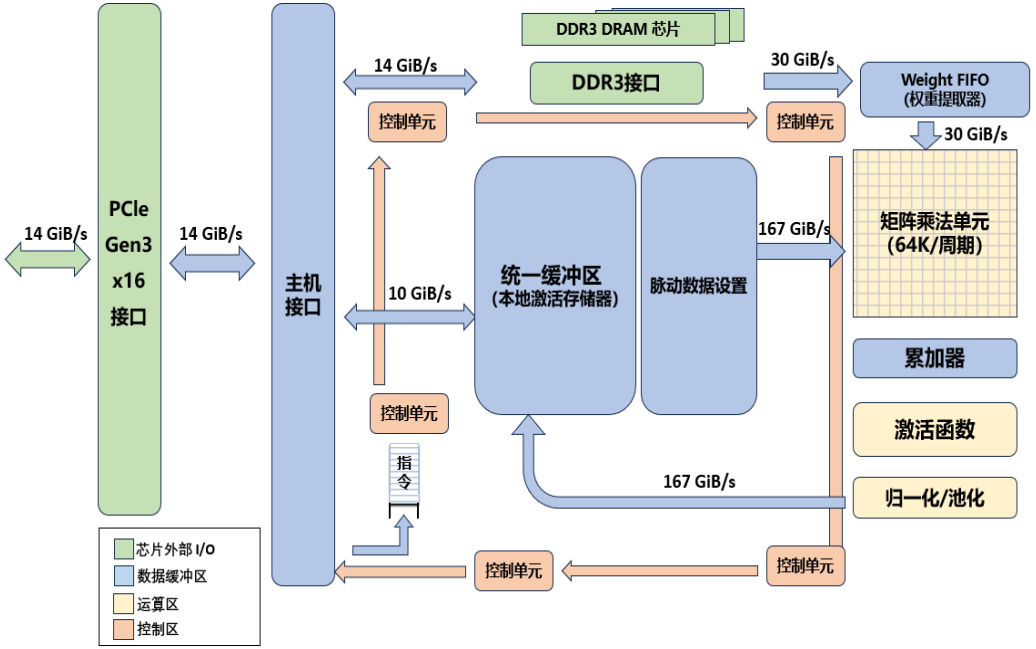


图：P100到B100各项指标参数(以GPT-4 1.8T MOE 10天基准测试为参考)

年份	2016	2018	2020	2022	2024
GPU	“Pascal” P100	“Volta” V100	“Ampere ” A100	“Hopper” H100	“Blackwell ” B100
峰值(TFLOPs)	19	130	620	4000	20000
训练和推理精度	FP16	FP16	FP16	FP8	FP4
推理每token焦耳数	17000	1200	150	10	0.4
训练所需电力(GW/小时)	1000	140	40	13	3
GPU 架构图					

2、ASIC-AI芯片：TPU架构基础

图：TPU架构（以Google TPUv1为参考）



TPU的运算资源包括：

- **矩阵乘法单元(MXU):**65536个8位乘法和加法单元，运行矩阵计算。
- **统一缓冲(UB):** 作为寄存器工作的24MB容量 SRAM。
- **激活单元(AU):** 硬件连接的激活函数。

TPU（张量处理单元）属于ASIC的一种，是谷歌专门为加速深层神经网络运算能力而研发的一款芯片，为机器学习领域而定制。

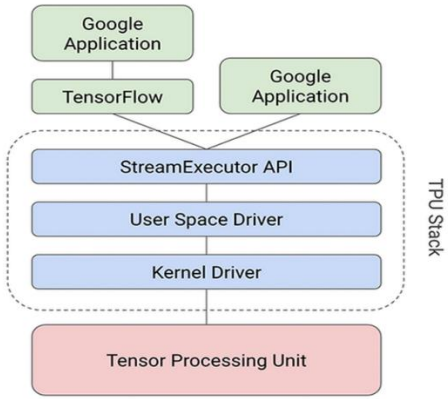
TPUv1依赖于通过PCIe(高速串行总线)接口与主机进行通信；它还可以直接访问自己的DDR3存储。

矩阵乘法单元：256 x 256大小的矩阵乘法单元，顶部输入256个权重值，左侧是256个input值。

DDR3 DRAM/Weight FIFO：权重存储通过DDR3-2133接口连接到TPUv1的DDR3 RAM芯片中，权重通过PCIe从主机的内存预加载，然后传输到权重FIFO存储器中，供矩阵乘法单元使用。

统一缓存区/脉动数据设置：应用激活函数的结果存储在统一缓冲区存储器中，然后作为输入反馈矩阵乘法单元，以计算下一层所需的值。

图：从TensorFlow到TPU：软件堆栈



TPU的设计封装了神经网络计算的本质，可以针对各种神经网络模型进行编程。此外，Google创建了编译器和软件栈，可以将来自TensorFlow的图像的API调用转换成TPU指令。

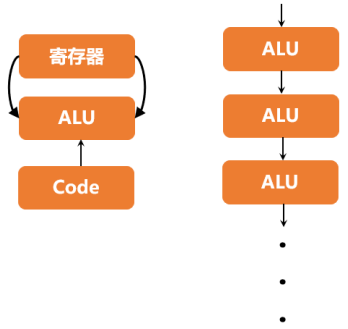
资料来源：Google Cloud官网、新智元公众号、量子位公众号，天风证券研究所

2、ASIC-AI芯片：TPU布局及性能对比

与传统CPU、GPU架构不同，TPU的MXU设计采用了脉动阵列(systolic array)架构，数据流动呈现出周期性的脉冲模式，类似于心脏跳动的供血方式。

如右图所示，CPU与GPU在每次运算中需要从多个寄存器中进行存取；而TPU的脉动阵列将多个运算逻辑单元(ALU)串联在一起，复用从一个寄存器中读取的结果。

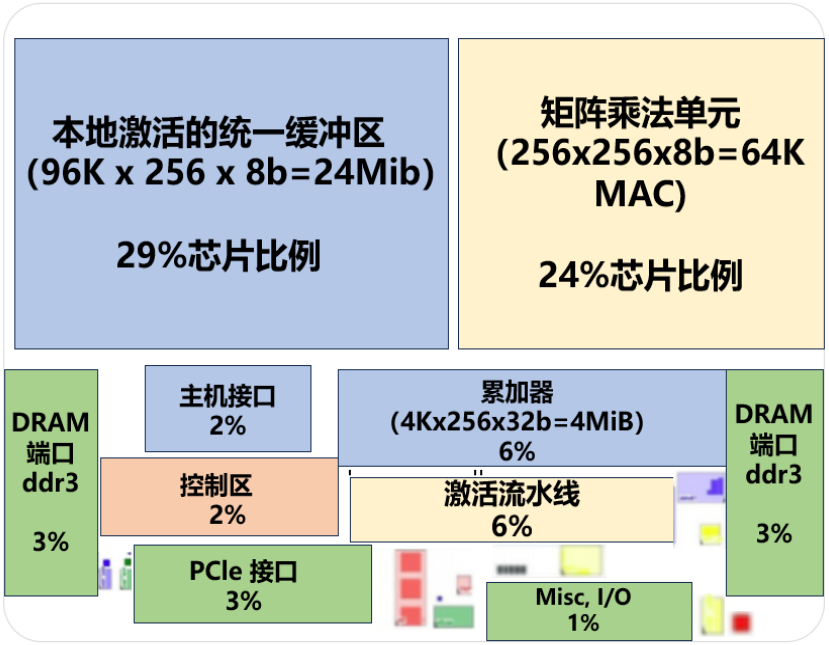
图：TPU与CPU/GPU运算方式对比



表：各芯片每周期算术运算量对比

芯片	每周运算量
CPU	数个
CPU(向量扩展)	数十
GPU	数万
TPU	数十万

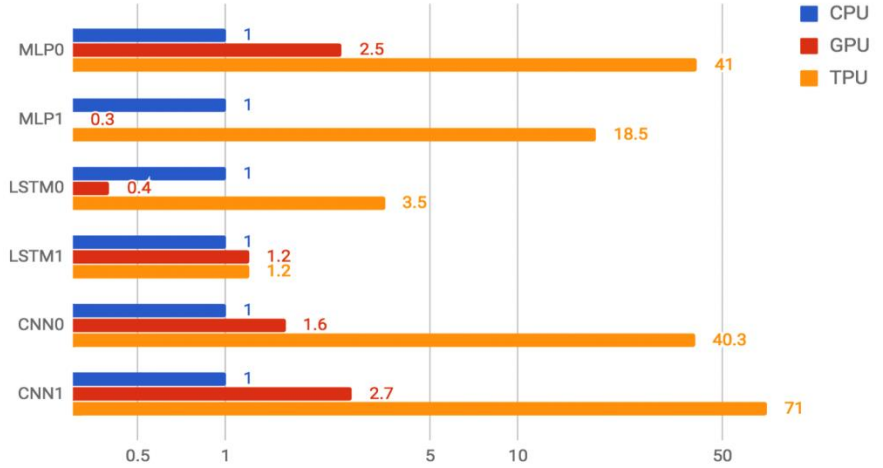
图：TPU芯片布局（以Google TPUv1为参考）



左图TPU芯片布局中，黄色代表运算单元（占比30%），蓝色代表数据单元（占比37%），绿色代表I/O（占比10%），橙色代表控制逻辑单元（仅占比2%）。

与CPU和GPU相比，TPU控制单元更小，给予存储器和运算单元留下了更大的空间。

图：TPU在神经网络上性能最大可达CPU的71倍



资料来源：Google Cloud官网、量子位公众号，天风证券研究所

3
风险提示

3、风险提示

1.AI算力景气度下降的风险：算力支出与下游应用息息相关，若AI应用需要更长期才能突破，则算力支出的高景气可能不可持续。

2.AI芯片竞争加剧的风险：AI芯片领域有较多参与者，未来市场竞争可能加剧。

3.技术发展风险：AI芯片及相关技术快速发展，技术迭代可能导致现有产品迅速过时，投资者应密切关注技术发展趋势，评估相关企业的创新能力和市场适应性。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

本报告是机密的，仅供我们的客户使用，天风证券不因收件人收到本报告而视其为天风证券的客户。本报告中的信息均来源于我们认为可靠的已公开资料，但天风证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，天风证券及/或其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，天风证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

天风证券的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。天风证券没有将此意见及建议向报告所有接收者进行更新的义务。天风证券的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

特别声明

在法律许可的情况下，天风证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。因此，投资者应当考虑到天风证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突，投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的6个月内，相对同期沪深300指数的涨跌幅	买入	预期股价相对收益20%以上
		增持	预期股价相对收益10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的6个月内，相对同期沪深300指数的涨跌幅	强于大市	预期行业指数涨幅5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

THANKS