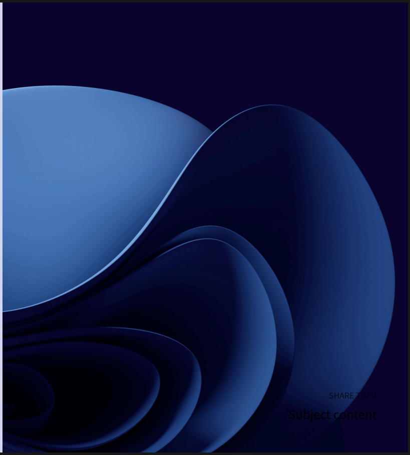
20260119

英伟达

企业价值来源分析



SHARE TOPIC

Subject content

目录

01 英伟达市值

02 AI时代红利

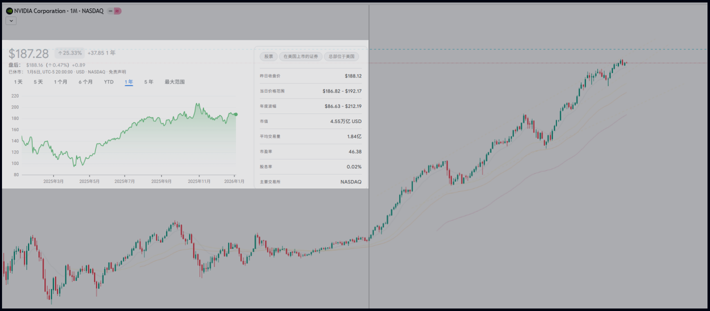
03 英伟达核心价值来源

04 盈利预测

05 自成长模型对英伟达的冲击

01

英伟达市值



NVIDIA Corporation · 1M · NASDAQ

$187.28 ↑25.33% +37.85 1 年

盘后： $188.16 (↑0.47%) +0.89

已休市： 1月6日, UTC-5 20:00:00 · USD · NASDAQ · 免责声明

昨日收盘价	$188.12
当日价格范围	$186.82 - $192.17
年度波幅	$86.63 - $212.19
市值	4.55万亿 USD
平均交易量	1.84亿
市盈率	46.38
股息率	0.02%
主要交易所	NASDAQ

是什么支持了英伟达的市值?

AI时代的GPU需求

淘金热时代，最赚钱的不是淘金者，而是"卖铲人"
AI时代，英伟达就是那个卖铲人——它不做AI应用，但所有AI都需要它的GPU

英伟达的机遇

为什么需要铲?
为什么是英伟达的铲?
这把铲能卖多久?

企业价值来源

时代红利 + 护城河深 = 长期成长

02

AI时代红利——为什么需要“铲”？

生成式 AI 引爆技术奇点，GPU 行业迎高增机遇

1. AI能力出现质变拐点，大模型参数量指数级爆炸（Scaling Law），GPT-3: 1750亿参数 → 万亿参数MoE模型 → 算力需求翻倍

2. AI三要素中，算法相对成熟，数据相对充足，算力严重不足，算力成为核心瓶颈

3. 为什么GPU是最佳解法?
深度学习的本质 = 大规模矩阵运算 = 海量简单重复计算 → GPU完美适配

4. GPU市场迎来爆发式增长

AI能力质变

时间节点	里程碑事件	意义
2020年	GPT-3发布（1750亿参数）	证明“大力出奇迹”的Scaling Law
2022.12	ChatGPT 3.5发布	引爆社会热潮，AI进入大众视野
现在	GPT-4o、Llama 3、通义千问等	大模型百花齐放

2020 年，参数量达 1750 亿的 GPT-3 在问答、摘要、翻译、续写等语言类任务上展现出了优秀的通用能力，证明了海量数据、更多参数、多元的数据采集渠道可构成 AI 发展的关键基础。

→

2022 年 12 月，ChatGPT 3.5 令人惊艳的使用体验引爆社会热潮，搜索热度和用户增长都出现了极为明显的提升。

→

目前全球大模型已进入百花齐放的阶段，DeepSeek、GPT-4o、阿里通义千问 Qwen2-72B、Llama 3、盘古大模型等多种模型应用层出不穷。

Scaling Law

Scaling Law（规模定律）：大模型的性能会随着模型参数量、数据量、计算量的增加而可预测地、持续地提升。

模型性能 $\propto f(N, D, C)$

N：参数量
D：数据量
C：计算量

三者需要协调增长，任何一个成为瓶颈都会拖累整体效果

多模态数据已成为训练这些大型模型的主要数据源，其对计算资源的需求是传统文本数据的数百倍
随着模型规模的不断扩大，对算力的需求也在急剧上升，这不仅推动了计算硬件的发展，也对算法的优化提出了更高的要求

Scaling Law

Scaling Law定律已经出现瓶颈，进入边际收益递减阶段

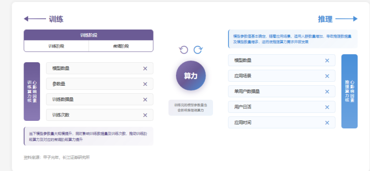
◆ 原有认知
公式：参数↑ + 数据↑ + 算力↑ → 模型能力持续线性提升
推论：引发算力军备竞赛

◆ 当前现实
收益锐减：10倍算力投入 → 可能仅换回20%性能提升
数据枯竭：互联网高质量文本数据接近耗尽
进展放缓：GPT-5等下一代模型飞跃不如预期
范式转变：行业从单纯“放大”模型，转向寻找新架构、新算法等新范式

算力需求

AI 技术在实际应用中包括两个环节：训练（Training）和推理（Inference），AIGC 的算力需要考虑训练及推理两个方面。

1、训练是指通过数据开发出 AI 模型，使其能够满足相应的需求，一般为 AI 技术的研发。因此参数量的升级对算力的需求影响大。
2、推理是指利用训练好的模型进行计算，利用输入的数据获得正确结论的过程，一般为 AI 技术的应用。推理部署的算力主要在于每个应用场景日数据的吞吐量。



当下模型参数量大规模提升，同时影响训练数据量及训练次数，推动训练阶段算力及对应的微调阶段算力提升

训练完的模型参数量也会影响推理端算力

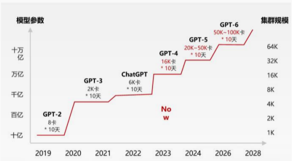
模型参数值基本确定，随着应用场景、适用人群数量增加，导致推理数据量及模型数量增多，进而使推理算力需求井喷发展

资料来源：甲子光年，长江证券研究所

算力需求

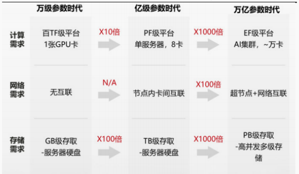
AI 大模型对算力的需求正以指数级速度增长，推动了 AI 算力平台从单一的单机计算向集群计算的转变

开发更高性能的 AI 大模型需要更强的算力平台



资料来源：华为，长江证券研究所

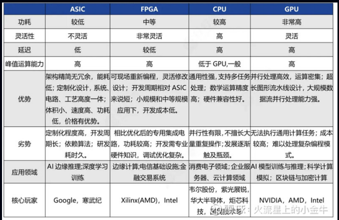
算力底座技术门槛提高，未来训练核心拼集群系统能力



	万级参数时代		亿级参数时代		万亿参数时代
计算需求	百TF级平台 1张GPU卡	X10倍	PF级平台 单服务器，8卡	X1000倍	EF级平台 AI集群，~万卡
网络需求	无互联	N/A	节点内卡间互联	X100倍	超节点+网络互联
存储需求	GB级存取 -服务器硬盘	X100倍	TB级存取 -服务器硬盘	X1000倍	PB级存取 -高并发多级存储

资料来源：华为，长江证券研究所

为什么是GPU



	ASIC	FPGA	CPU	GPU
功耗	较低	中等	较高	非常高
灵活性	不灵活	非常灵活	高	灵活
延迟	低	较低	高	高
峰值运算能力	高	高	低于 GPU,一般	高
优势	架构精简无冗余，能耗低；定制化设计，系统、电路、工艺高度一体；体积小、速度高、功耗低，价格有优势。	可现场重新编程，灵活修改设计；开发周期相对 ASIC 来说短；小规模和中等规模应用下，开发成本低。	通用性强，支持多任务处理；数学运算精度高；硬件兼容性好。	并行处理高效，运算密集；超长图形流水线设计，大规模数据流并行处理能力强。
劣势	定制化程度高，开发周期长；依赖算法；研发耗时久。	相比优化后的专用集成电路，功耗较高；开发需专业硬件知识，调试优化复杂。	并行性有限，不擅长大量重复操作；发展逐渐触及瓶颈。	无法执行通用计算任务；成本较高；难以处理复杂编程模式。
应用领域	AI 边缘推理;深度学习训练	边缘计算;电信基础设施;金融交易系统	消费电子领域；企业服务器、云计算领域	AI 模型训练与推理；科学计算模拟；区块链与加密计算
核心玩家	Google，寒武纪	Xilinx(AMD)，Intel	韦尔股份，紫光展锐，华大半导体，炬芯科技，国民技术等	NVIDIA，AMD，Intel

雪球：火流星上的小金牛

深度学习=大规模矩阵运算=海量简单重复计算

芯片类型	优势	劣势	适用阶段
CPU	通用性强、生态成熟	并行能力弱	传统计算
GPU	并行能力强、生态成熟(CUDA)	功耗大	当前AI主流
FPGA	灵活可编程	开发难度大	边缘推理
ASIC	功耗低、性能强	研发周期长、不可更改	特定场景（如谷歌TPU）

GPU = 性能与灵活性的最佳平衡点 + CUDA生态加持 → 当前AI首选

当前阶段，GPU 配合 CPU 是AI 芯片的主流，而后随着视觉、语音、深度学习的算法在 ASIC 芯片上的不断优化，此两者也将逐步占有更多的市场份额，从而与 GPU 达成长期共存的局面。

03

英伟达核心价值

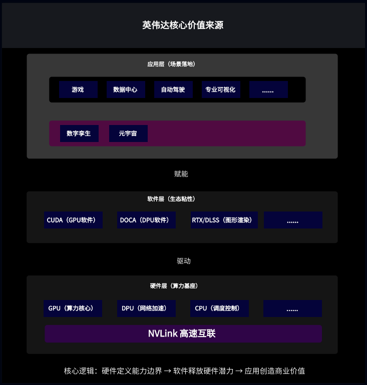
英伟达核心价值来源

GPU 是英伟达的核心产品，围绕 GPU 及其核心应用——图形渲染和加速运算，英伟达持续扩展自身软硬件实力

1、在硬件的三芯战略+互联网络
2、软件侧的开发软件+行业应用软件
3、应用层对各核心下游持续加大投入

英伟达完成了多个阶段的快速发展，目前已经成为了全球核心 AI 芯片及应用企业

更强的芯片 → 更多开发者采用CUDA → 更丰富的软件库
↑ ↓
更多研发投入 ← 更高的市场份额 ← 更好的应用体验



英伟达核心价值来源

核心逻辑：硬件定义能力边界 → 软件释放硬件潜力 → 应用创造商业价值

硬件层——三芯战略

2020 年，在完成对 Mellanox 的收购后，英伟达推出 BlueField-2 DPU，将其定义为继 CPU、GPU 之后“第三颗主力芯片”。。随后在 2021 年的 GTC 大会上，英伟达发布基于 ARM 架构的 CPU——NVIDIA Grace，黄仁勋正式将英伟达产品路线升级为“GPU+CPU+DPU”的“三芯”战略。

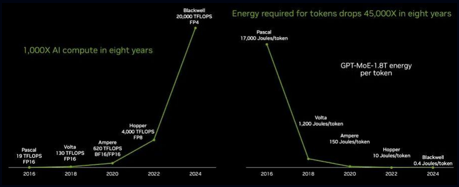
芯片	典型代表	职责	关键场景
CPU	Grace / x86 / ARM	任务调度、通用计算、系统管理	控制、逻辑决策
GPU	Hopper / Blackwell	AI 训练与推理的并行计算引擎	模型训练、矩阵运算
DPU	BlueField	网络、存储、安全的硬件卸载	通信、IO加速、多租户隔离

硬件层——GPU

架构持续迭代，AI 算力的硬通货

英伟达 GPU 架构持续迭代，朝着愈发适宜 AI 计算的方向逐步演进。从 Tesla到Blackwell，公司持续迭代GPU架构，从工业体系上逐层从40nm演进至4nm，CUDA核心数也从最初的 128 个增加至上万个，并添加了 Tensor 张量计算核心、NVLink、RTCore、结构稀疏性矩阵 MIG 等功能，数据计算类型逐步丰富，包含了 FP、INT、TF、BF 等数据类型，计算架构逐步朝更适合 AI 运算的方向演进。

英伟达 GPU 架构算力持续提升，耗能逐步下降

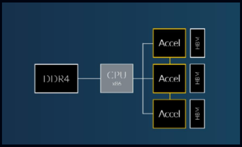


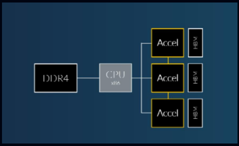
硬件层——CPU

依托 Arm 实现较强内存一致性，NVLink-C2C 保证芯片高宽带互联

在云计算领域，Arm 市场份额逐步提升。实际上，在以云计算为代表的数据基础设施领域，Arm 的份额正逐步提升，根据Arm 公司财报，FYE22-FYE24 年（公历年 2021 年 12 月-2024 年 11 月），在云计算领域，Arm 市场份额从 9%提升之 15%，网络设备领域市场份额从 23%提升至 28%。

1、Arm 架构下，CPU 可以实现较强的内存一致性与定制化，更能适应 AI 数据计算
2、采用 NVLink-C2C 技术，发布基于 Arm 架构的 Grace 系列 CPU





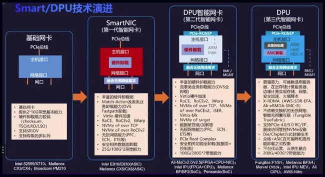
硬件层——DPU

CPU 算力与网络带宽增速剪刀差持续放大，在大型数据中心，网络流量的处理占到了计算的 30% 左右，也催生了市场对于更优网络解决方案的诉求

DPU（数据处理单元）是专门用于处理数据中心网络传输、数据安全和基础设施任务的芯片，旨在减轻 CPU 在数据传输、加密和存储等任务中的负担

英伟达收购 Mellanox，开启 DPU 布局。2019 年英伟达收购 Mellanox，加速了DPU 技术的落地，并在 2020 年发布了 BlueField 系列的 DPU 产品，落地 GPU-direct RDMA 技术，实现了 GPU 对其他主机 GPU 内存的直接访问。

网卡从一个被动的、功能固定的数据传输部件，演变成了一个主动的、可编程的、具备强大算力的独立计算单元



Smart/DPU技术演进

基础网卡

PCIe总线
主机接口
网络接口
网口

- 基础网卡
- 提供2*10G带宽基本能力
- 硬件卸载能力较弱（checksum，TSO/LRO/LSO)
- 支持SRIOV
- 支持有限的多队列

Intel 82599/X710；Mellanox CX3/CX4；Broadcom PM210

SmartNIC
(第一代智能网卡)

PCIe总线
主机接口
硬件卸载
网络接口
融合无损网络需求
网口

- 丰富的硬件卸载能能
- Match-Action流表类业务卸载能力(OVS Fastpath卸载)
- Virtio 硬件加速
- RoCE、RoCEv2、iWarp
- NVMe oF over TCP NVMe oF over RoCEv2
- 无损网络能力(PFC，ECN，ETS等)
- 安全相关数据面卸载
- 25G/100G*2带宽能力

Intel E810/E830(ASIC)
Mellanox CX5/CX6(ASIC)

DPU智能网卡
(第二代智能网卡)

- 丰富的硬件卸载能力
- 流表类业务卸载能力(OVS全卸载)
- Virtio硬件加速
- RoCE、RoCEv2、iWarp
- NVMe oF over TCP，NVMe oF over RoCEv2，iSER，Virtio-blk
- NVMe oF target
- 数据解压缩/加解密
- 无损网络能力(PFC， ECN，ETS等)
- PCIe Root Complex
- 安全相关功能全卸载(数据面+控制面)
- 100G/200G*2带宽能力

Ali-MoCv2.0/v2.5(FPGA+CPU+NIC)；Intel IPU(FPGA+CPU)；Mellanox BF/BF2(SoC)；Pensando(SoC)

DPU
(第三代智能网卡)

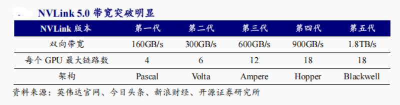
- 更强算力，可替换通用服务器，在近存储计算服务器，边缘计算实现存储、网络、安全加速，大幅降本
- X-RDMA（AWS-SDR-EFA，Ali-eRMDA-SMC-R)
- 不依赖交换机的无损网络，智能无拥塞互联（Fungible TrueFabric)
- 支持PCIe 4.0/5.0 RC/EP，直接访问管控NVMe设备
- Die/Chiplet方式封装技术
- 众核+ASIC在可编程和高性能卸载之间权衡
- 平台化运营，云原生融合
- 200G/400G*2带宽能力

Fungible F1/S1；Mellanox BF3/4；Marvel CN10k；Intel IPU MEV；Ali CIPU；AWS-Nitro

NVLink与Infiniband

NVLink 是英伟达 GPU 与 GPU、GPU 与 CPU 的高速互连技术
Infiniband 是一种为高性能计算和数据中心设计的、速度极快的网络互联技术

1、传统的 GPU通常采用 PCIe 接口与 x86 架构的 CPU 互联，由于记忆系统的差异（GPU 有更快但更小的内存，而 CPU 有较大但较慢的内存），限制了彼此的数据传输能力。
2、2014 年，英伟达联合 IBM 推出 NVLink 高速互联技术，使得 GPU 与 CPU 可以以 5-12 倍的速度分享数据，此外，NVLink 协议在设计时考虑了数据一致性问题，使得不同 GPU 之间的数据访问可以保证一致性。



NVLink 5.0 带宽突破明显

NVLink 版本	第一代	第二代	第三代	第四代	第五代
双向带宽	160GB/s	300GB/s	600GB/s	900GB/s	1.8TB/s
每个 GPU 最大链路数	4	6	12	18	18
架构	Pascal	Volta	Ampere	Hopper	Blackwell

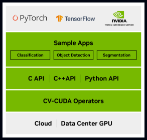
资料来源：英伟达官网、今日头条、新浪财经、开源证券研究所

10 万卡集群时代到来，网络集群能力愈发重要。Infiniband 在高性能计算领域具备优势，英伟达（Mellanox）处于领导地位。
Infiniband 与以太网是数据中心采用的主要网络标准，得益于高传输速率和低延迟的特性，Infiniband 在服务器间的高速通信、存储设备与网络设施之间的高效互联中扮演着至关重要的角色。

软件层

软件层是连接硬件能力与上层应用的桥梁，也是英伟达的护城河之一——不是硬件买不起，而是生态离不开。

组件	对应硬件	核心功能	一句话理解
CUDA	GPU	并行计算开发平台	让GPU不只能画图，还能算AI
DOCA	DPU	复刻CUDA模式，锁定网络/存储生态	为DPU提供类似CUDA的开发模式，聚焦网络与存储
RTX	GPU	光线追踪+DLSS智能插帧	游戏画质的"黑科技"
Magnum IO	全系	存储与网络IO优化	让数据搬运更快



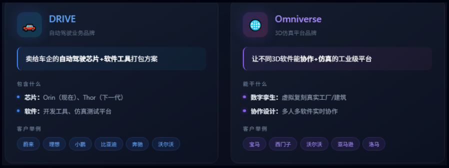
应用工具层（让开发更简单）
基于CUDA向上延伸出300+加速库、400+预训练模型，覆盖AI、科学计算、图形渲染等场景，开发者"拿来即用"。

CUDA作为护城河之一：
17年积累：2006年发布，比竞争对手早跑一个时代
生态锁定：90%以上AI训练依赖CUDA，PyTorch/TensorFlow全面支持
切换成本极高：代码要重写、工程师要重学、优化要从头来
5000万+下载量：开发者社区已形成飞轮效应

更强的芯片 → 更多开发者采用CUDA → 更丰富的软件库
↑ ↓
更多研发投入 ← 更高的市场份额 ← 更好的应用体验

应用层布局

英伟达两大战略布局：DRIVE & Omniverse



DRIVE
自动驾驶业务品牌

卖给车企的**自动驾驶芯片+软件工具**打包方案

包含什么
- **芯片：** Orin（现在）、Thor（下一代）
- **软件：** 开发工具、仿真测试平台

客户举例
蔚来 理想 小鹏 比亚迪 奔驰 沃尔沃

Omniverse
3D仿真平台品牌

让不同3D软件能**协作+仿真**的工业级平台

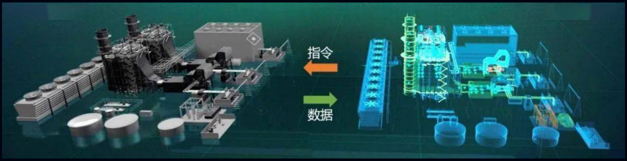
能干什么
- **数字孪生：** 虚拟复刻真实工厂/建筑
- **协作设计：** 多人多软件实时协作

客户举例
宝马 西门子 沃尔沃 亚马逊 洛马

免费/低价提供软件 → 客户粘性 → 购买英伟达芯片

数字孪生和元宇宙

数字孪生和元宇宙都是构建数字虚拟空间的重要概念，它们既有紧密联系，又在目标和范围上存在差异。为了帮你快速把握核心关系，我先用一个表格来梳理它们的对比，然后再深入探讨其中的相互作用。



数字孪生是元宇宙的“基石”与“基础设施”
元宇宙要成为一个逼真且有意义的世界，需要大量真实、有序的内容来填充。数字孪生技术能够将物理世界中的物体、建筑、甚至整个城市高精度地“复制”到数字空间中，为元宇宙提供了坚实的三维内容基底和与现实世界联动的接口。
元宇宙是数字孪生体的“社会化网络”
当无数个独立的数字孪生体（如一个工厂、一栋大楼、一辆车）被创建出来后，它们不会是孤立的。元宇宙作为一个持久存在的共享虚拟空间，为这些数字孪生体提供了相互连接、交互和“社交”的环境。
融合趋势：数实融合的新载体
当前，两项技术的融合已经开始，并催生了新的业态。
产业元宇宙：在工业领域，“工业元宇宙”的概念正是两者融合的体现。通过在元宇宙环境中部署工厂的数字孪生体，工程师和管理者可以沉浸式地进行远程运维、员工培训和生产流程优化，实现“以虚强实”。
消费元宇宙：在消费领域，利用数字孪生技术复刻历史建筑或整个商圈，并将其放入元宇宙，为用户提供虚拟购物、参与线上虚拟时装秀等全新的消费体验。这打破了物理空间的限制，极大地丰富了线上互动。

空间智能

空间智能是AI理解和操作三维物理世界的能力，包括识别3D场景、预测物体运动、掌握物理规律。代表应用：自动驾驶、机器人、AR/VR。

数据需求不同
传统AI：文本、2D图像
空间智能：3D点云、深度图、多视角视频、物理交互数据

训练方法创新
大量依赖仿真环境（虚拟世界生成训练场景）
采用强化学习（通过试错学习操作）
构建世界模型（预测物理规律）
自监督学习（降低标注成本）

从"专用模型"走向"通用空间理解"——训练能理解任意3D场景、执行多样化物理任务的通用AI，这是AI从理解符号世界到操作物理世界的关键跃迁。

04

长期成长——盈利预测

盈利基石

时代红利 + 英伟达多层次的护城河

1、scaling law下大模型涌现出了“智能”，带来了巨大的算力需求，高性能GPU的需求量呈现指数级增长

虽然scaling law，已经见顶，但是只是边际效应递减，十万亿规模的模型还存在需求；
没有算法的突破性创新，在如今的算法体系下，模型的研发依然需要遵循大数据、堆算力的规则；

2、英伟达依靠GPU芯片技术的深厚技术积累，成为全球算力提供的领军者

首先是在硬件架构以及性能上，不断创新突破，推出DPU等新的技术概念与变革；
NVLink和Infiniband技术对高性能算力集群、数据中心的建立提供了支撑；

3、在硬件领域的深厚积累与创新上软件生态（例如CUDA）给予硬件更好的使用体验

目前大模型的研发和商用几乎拥抱围绕英伟达GPU建立起来的软件生态

需求量、需求规模、需求粘性是目前盈利的基础，支撑着现在的高市值

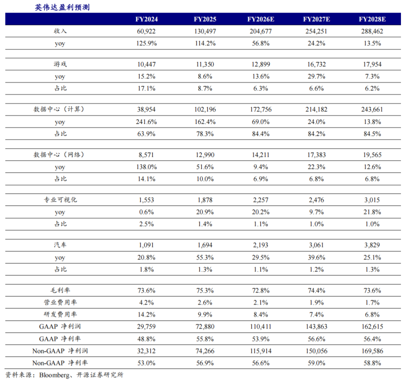
盈利业务

游戏：一方面，美国潜在关税影响或对行业带来负面冲击，另一方面，英伟达于2025年发布GeForce 50系列，有望带动FY2026-2027财年收入增长，且过去几个季度英伟达 AIB 显卡市场份额基本保持提升态势，我们预计英伟达仍有望保持逆势增长；

数据中心：公司最新 Blackwell 架构芯片或存在延迟量产的风险，但 Hopper系列一定程度上可以弥补销售缺口，我们预计 FY2026H1 英伟达 B 系列将进入量产阶段，推动 FY2026-2027 财年数据中心（计算+网络）收入稳健增长；

专业可视化：预计随着英伟达 Blackwell 架构和新一代显卡开始普及，专业显卡亦有望受益，此外英伟达 Omniverse 亦贡献部分增量，考虑到专业显卡市场弹性不如数据中心，我们预计收入仅维持稳健增长；

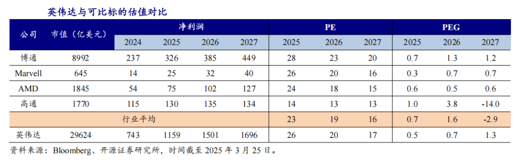
自动驾驶领域：受益于下游智驾行业高资本投入红利、英伟达后续 Thor 系列有望带动增长，基于公司在域控芯片的领先市场地位，我们预计英伟达汽车业务仍将保持 30%-40%左右的同比增长。



英伟达盈利预测

	FY2024	FY2025	FY2026E	FY2027E	FY2028E
收入	60,922	130,497	204,677	254,251	288,462
yoy	125.9%	114.2%	56.8%	24.2%	13.5%
游戏	10,447	11,350	12,899	16,732	17,954
yoy	15.2%	8.6%	13.6%	29.7%	7.3%
占比	17.1%	8.7%	6.3%	6.6%	6.2%
数据中心（计算）	38,954	102,196	172,756	214,182	243,661
yoy	241.6%	162.4%	69.0%	24.0%	13.8%
占比	63.9%	78.3%	84.4%	84.2%	84.5%
数据中心（网络）	8,571	12,990	14,211	17,383	19,565
yoy	138.0%	51.6%	9.4%	22.3%	12.6%
占比	14.1%	10.0%	6.9%	6.8%	6.8%
专业可视化	1,553	1,878	2,257	2,476	3,015
yoy	0.6%	20.9%	20.2%	9.7%	21.8%
占比	2.5%	1.4%	1.1%	1.0%	1.0%
汽车	1,091	1,694	2,193	3,061	3,829
yoy	20.8%	55.3%	29.5%	39.6%	25.1%
占比	1.8%	1.3%	1.1%	1.2%	1.3%
毛利率	73.6%	75.3%	72.8%	74.4%	73.6%
营业费用率	4.2%	2.6%	2.1%	1.9%	1.7%
研发费用率	14.2%	9.9%	8.4%	7.4%	6.8%
GAAP 净利润	29,759	72,880	110,411	143,863	162,615
GAAP 净利率	48.8%	55.8%	53.9%	56.6%	56.4%
Non-GAAP 净利润	32,312	74,266	115,914	150,056	169,586
Non-GAAP 净利率	53.0%	56.9%	56.6%	59.0%	58.8%

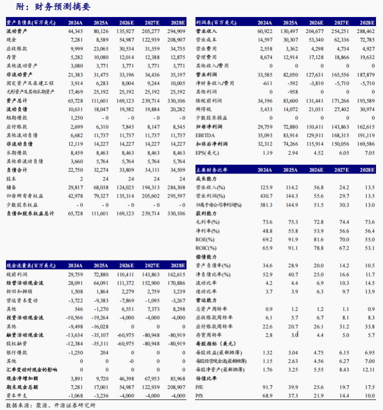
资料来源：Bloomberg、开源证券研究所



英伟达与可比标的估值对比

公司	市值（亿美元）	净利润				PE			PEG		
		2024	2025	2026	2027	2025	2026	2027	2025	2026	2027
博通	8992	237	326	385	449	28	23	20	0.7	1.3	1.2
Marvell	645	14	25	32	40	26	20	16	0.3	0.7	0.7
AMD	1845	54	75	102	127	24	18	15	0.6	0.5	0.6
高通	1770	115	130	135	134	14	13	13	1.0	3.8	-14.0
行业平均						23	19	16	0.7	1.6	-2.9
英伟达	29624	743	1159	1501	1696	26	20	17	0.5	0.7	1.3

资料来源：Bloomberg、开源证券研究所，时间截至 2025 年 3 月 25 日。



附：财务预测摘要

资产负债表(百万美元)	2024A	2025A	2026E	2027E	2028E
流动资产	44,345	80,126	135,927	205,277	294,909
现金	7,281	8,589	54,987	122,939	208,907
应收账款	9,999	23,065	30,534	31,559	34,735
存货	5,282	10,080	12,014	12,388	12,875
其他流动资产	3,080	3,771	3,771	3,771	3,771
非流动资产	21,383	31,475	33,196	34,436	35,197
固定资产及在建工程	3,914	6,283	8,004	9,244	10,005
无形资产及其他长期资产	17,469	25,192	25,192	25,192	25,192
资产总计	65,728	111,601	169,123	239,714	330,106
流动负债	10,631	18,047	19,582	19,884	20,282
短期借款	1,250	- 0	- 0	- 0	- 0
应付账款	2,699	6,310	7,845	8,147	8,545
其他流动负债	6,682	11,737	11,737	11,737	11,737
非流动负债	12,119	14,227	14,227	14,227	14,227
长期借款	8,459	8,463	8,463	8,463	8,463
其他非流动负债	3,660	5,764	5,764	5,764	5,764
负债合计	22,750	32,274	33,809	34,111	34,509
股本	2	24	24	24	24
储备	29,817	68,038	124,025	194,313	284,308
归母所有者权益	42,978	79,327	135,314	205,602	295,597
少数股东权益	- 0	- 0	- 0	- 0	- 0
负债和股东权益总计	65,728	111,601	169,123	239,714	330,106

现金流量表(百万美元)	2024A	2025A	2026E	2027E	2028E
税前利润	29,759	72,880	110,411	143,863	162,615
经营活动现金流	28,091	64,091	111,372	152,900	170,886
折旧和摊销	1,508	1,864	2,279	2,759	3,239
营运资本变动	-3,722	-9,383	-7,869	-1,095	-3,267
其他	546	-1,270	6,551	7,373	8,298
投资活动现金流	-10,566	-19,264	-4,000	-4,000	-4,000
其他	-9,498	-16,028	0	0	0
融资活动现金流	-13,634	-35,107	-60,975	-80,948	-80,919
股权融资	-12,384	-35,311	-60,975	-80,948	-80,919
银行借款	-1,250	204	0	0	0
其他	0	0	-0	0	0
汇率变动对现金的影响	0	0	0	0	0
现金净增加额	3,891	9,720	46,398	67,953	85,968
期末现金总额	7,281	17,001	54,987	122,939	208,907
资本开支	-1,068	-3,236	-4,000	-4,000	-4,000

利润表(百万美元)	2024A	2025A	2026E	2027E	2028E
营业收入	60,922	130,497	204,677	254,251	288,462
营业成本	14,597	30,307	53,340	62,336	72,785
营业费用	2,558	3,362	4,298	4,734	4,927
管理费用	8,674	12,914	17,128	18,866	19,632
其他收入/费用	0	0	0	0	0
营业利润	33,585	82,050	127,631	165,556	187,879
净财务收入/费用	-611	-592	-3,810	-5,710	-5,710
其他利润	0	-958	0	0	0
除税前利润	34,196	83,600	131,441	171,266	193,589
所得税	5,433	14,072	21,031	27,402	30,974
少数股东损益	0	0	0	0	0
归母净利润	29,759	72,880	110,411	143,863	162,615
EBITDA	35,093	83,914	129,911	168,315	191,119
扣非后净利润	32,312	74,266	115,914	150,056	169,586
EPS(美元)	1.19	2.94	4.52	6.05	7.05

主要财务比率	2024A	2025A	2026E	2027E	2028E
成长能力					
营业收入(%)	125.9	114.2	56.8	24.2	13.5
营业利润(%)	430.7	144.3	55.6	29.7	13.5
归属于母公司净利润(%)	581.3	144.9	51.5	30.3	13.0
获利能力					
毛利率(%)	73.6	75.3	72.8	74.4	73.6
净利率(%)	48.8	55.8	53.9	56.6	56.4
ROE(%)	69.2	91.9	81.6	70.0	55.0
ROIC(%)	65.9	91.1	78.8	67.2	53.1
偿债能力					
资产负债率(%)	34.6	28.9	20.0	14.2	10.5
净负债比率(%)	52.9	40.7	25.0	16.6	11.7
流动比率	4.2	4.4	6.9	10.3	14.5
速动比率	3.7	3.9	6.3	9.7	13.9
营运能力					
总资产周转率	0.9	1.2	1.2	1.1	0.9
应收账款周转率	6.1	5.7	6.7	8.1	8.3
应付账款周转率	22.6	20.7	26.1	31.2	33.8
存货周转率	2.8	3.0	4.4	5.0	5.7
每股指标（美元）					
每股收益(最新摊薄)	1.32	3.04	4.75	6.15	6.95
每股经营现金流(最新摊薄)	1.15	2.63	4.56	6.27	7.00
每股净资产(最新摊薄)	1.76	3.25	5.55	8.43	12.11
估值比率					
P/E	91.7	39.9	25.6	19.7	17.5
P/S	68.9	37.3	21.9	14.4	10.0

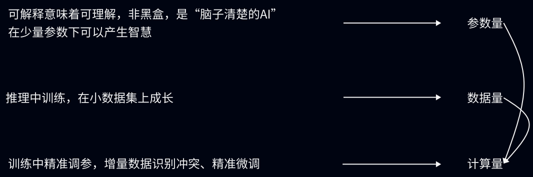
数据来源：聚源、开源证券研究所

05

自成长模型对英伟达的冲击

突破scaling law的限制

自成长模型在算法上进行了突破创新——可解释



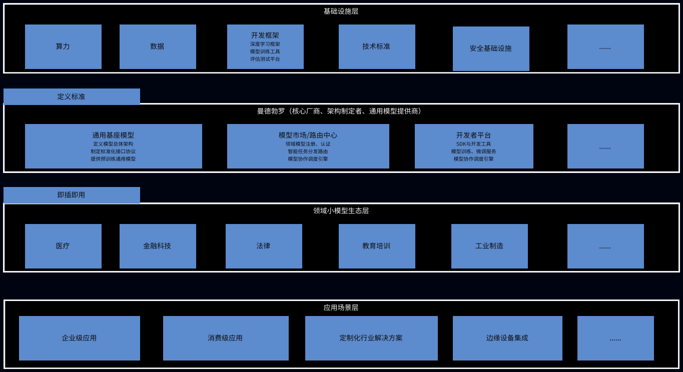
可解释意味着可理解，非黑盒，是“脑子清楚的AI”
在少量参数下可以产生智慧

推理中训练，在小数据集上成长

训练中精准调参，增量数据识别冲突、精准微调

逃出算力黑洞

自成长模块化大模型产业生态链



基础设施层

算力

数据

开发框架
深度学习框架
模型训练工具
评估测试平台

技术标准

安全基础设施

……

定义标准

曼德勃罗（核心厂商、架构制定者、通用模型提供商）

通用基座模型
定义模型总体架构
制定标准化接口协议
提供预训练通用模型

模型市场/路由中心
领域模型注册、认证
智能任务分发路由
模型协作调度引擎

开发者平台
SDK与开发工具
模型训练、微调服务
模型协作调度引擎

……

即插即用

领域小模型生态层

医疗

金融科技

法律

教育培训

工业制造

……

应用场景层

企业级应用

消费级应用

定制化行业解决方案

边缘设备集成

……

不依赖大规模集群

1、参数解耦、参数结构化、行业领域模块化

2、分时分工训练

3、即插即用

不依赖高性能GPU，不依赖大规模算力中心、数据中心

打破CUDA生态

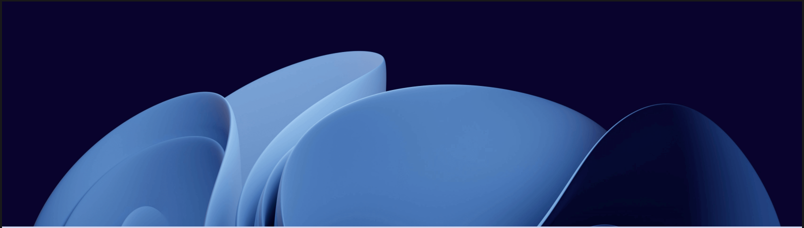
CUDA生态下，是其对硬件依赖性
CUDA技术基本只能在NVIDIA的GPU上运行，无法直接用于AMD或Intel等其他厂商的GPU

自成长模块化大模型产业生态链下

“分工分时训练” 意味着大模型的不同组成部分（比如负责理解语言的模块、负责生成图像的模块）可以独立开发、优化，甚至由不同的团队或公司来负责

“插拔使用”，就像拼乐高或更换电脑内存条一样，开发者可以灵活地组合这些标准化模块，快速构建或定制自己需要的AI大模型，而无需被某一家公司的硬件和软件栈彻底锁死

跨越英伟达核心护城河



致谢