英伟达人工智能发展战略研究报告

鼎帷咨询战略创新研究院

2024年9月

引言

01 英伟达愿景使命

人工智能大模型的迅猛发展推动算力需求的高速增长，占据数据中心GPU市场98%这一绝对份额的英伟达，已经从曾经游戏显卡之王华丽转身成为人工智能时代“卖铲人”，英伟达的野心绝不仅限于此，英伟达致力于成为人工智能计算领域的领导者，创造下一个工业革命，用算力来驱动智能时代的到来，驱动整个社会的智能化转型，让算力渗透到所有的社会领域，成为人工智能时代基础设施的提供商。

02 英伟达三芯战略

- GPU（图像处理器）作为英伟达的核心产品，特别是在AI训练和推理、高性能计算以及图形渲染等领域，提供了强大的并行处理能力；
- CPU（中央处理器）的加入使得英伟达能够更好地应对需要快速逻辑判断和高度并行处理能力的应用；
- DPU（数据处理器）则专门针对数据中心和网络设备的需求，具有高效处理数据包和协议的能力。

通过集成CPU、GPU和DPU到同一平台上，英伟达的三芯战略为客户提供了一站式的解决方案，能够更好地满足现代数据中心对于多样化计算需求的挑战。

03 英伟达四大业务

- 游戏业务是英伟达的基本盘。英伟达提供GeForce RTX GPU等针对个人电脑的显卡、SoC等针对游戏主机的显卡和GeForce Now为主的云游戏服务。目前游戏业务以“云游戏+AI”为发展主线，不断强化RTX光线追踪和ACE虚拟数字人类生成两大关键技术，预计将在2025年发布AI PC芯片，进入高端笔记本电脑市场。
- 专业可视化业务主要聚焦于为设计和可视化专业人士提供先进解决方案。英伟达面向专业级和企业级市场提供Quadro系列和RTX系列GPU，面向创作者和专业开发者提供6大版本的Omniverse实时协作平台，形成Omniverse生态系统，应用于电影制作、建筑设计、虚拟现实等众多领域。
- 数据中心业务是驱动英伟达市值增长的第一大业务。英伟达提供从边缘计算到云端的全方位产品和解决方案，包括最新架构的CPU、GPU、DPU芯片、DGX系统、高速网络系列产品以及AI Enterprise等软硬件生态系统。英伟达持续强化五层算力体系，布局AI工厂，发布DGX SuperPOD超级计算机，引领下一代AI基础设施。未来，数据中心业务也从面向传统云服务商发展到面向各国家私有云及电信云，推出Jetson平台，大力布局边缘计算领域。
- 汽车业务是英伟达极具未来潜力的关键业务。英伟达提供端到端的解决方案，包括Thor、Atlan汽车芯片、DRIVE软硬件及基础设施，赢得国内外广泛主机厂客户认可。合作伙伴从传统主机厂扩展到上下游汽车零部件及软件服务商。未来英伟达持续布局虚拟工厂及仿真应用，推进自动驾驶的三步走战略，进入AI定义汽车的2.0时代。

04 英伟达应用场景

在医药领域，英伟达提供Clare Holoscan计算平台，支持从医疗设备到边缘服务器的无缝连接，推动医学影像AI分析，与甲骨文、强生等代表性企业进行深度合作。

在汽车领域，英伟达提供DRIVE软硬件及Omniverse平台，赋能端到端的汽车自动驾驶解决方案，与特斯拉、小鹏等代表性企业进行深度合作。

在机器人领域，英伟达提供Issac、Omniverse及Jetson平台，协助开发各类型机器人，并且前瞻布局具身智能，与比亚迪、西门子等代表性企业合作。

05 英伟达竞争策略

- 英伟达采用“三团队-两季度”的创新研发迭代模式，即三个并行开发团队专注于独立的分阶段产品开发，确保公司每6个月推出一次新产品领先市场1-2个研发周期，使得GPU的算力增长始终高于CPU的算力增长而无法被CPU集成，实现了计算机芯片产品品类的重新定义。
- 英伟达通过构筑软件生态、调动开发者、发掘应用场景对计算机形态进行渐进式改造，使得计算机从单纯的“CPU” 形态逐渐演变为“CPU+GPU”形态，塑造计算机形态向着有利于其自身发展的形态演进。

06 英伟达生态壁垒

- CUDA软件生态系统包括了多个层面，从编程语言和API支持到性能分析和调试工具，再到丰富的库和框架，以及对多种应用领域的支持，覆盖AI和HPC领域，CUDA与英伟达的GPU硬件紧密结合，提供了最佳的性能和最优化的体验。
- 这种封闭的集成策略使得CUDA在性能上具有明显优势，构筑了软件覆盖率高、AI框架支持率高、细分行业渗透率高三大生态竞争壁垒，巩固了英伟达在AI和高性能计算市场的领导地位。
- 通过提供全面的软件支持和优化，英伟达的GPU在训练和部署AI模型方面成为行业标准，使得英伟达在AI芯片市场中占据主导地位。

07 供应链主导地位

英伟达充分利用数据中心对人工智能的无限需求，凭借自身的巨大产量和人工智能服务器所有技术和组件的超前领先地位，将网络组件、内存和其他组件都封装（CoWoS）到单个系统中，进而在供应链中占据主导地位。近日黄仁勋还透露，英伟达自主开发了很多技术，必要时，可弃用台积电，能让英伟达把订单转移给替代供应商。

为了方便您的阅读，请先了解以下常见术语

术语	定义	用途
A100	基于Ampere架构的高性能数据中心GPU	支持大规模AI训练和科学计算
Aerial	用于5G和边缘网络的AI平台	提升电信网络的性能和智能化
API	一组预定义的函数或协议，用于构建软件应用程序	支持软件开发和集成
Clara	用于医疗健康领域的AI计算平台	提升医疗设备和应用的性能
CUDA	Compute Unified Device Architecture 的缩写，是 NVIDIA 的并行计算架构和编程模型。	用于开发 GPU 加速的应用程序和算法。
CUDA Cores	CUDA 计算单元，GPU 中用于并行处理任务的基本处理单元。	用于执行并行计算任务，加速计算密集型应用。
CUDA Kernel	CUDA 程序中执行的并行计算函数。	用于在 GPU 上执行并行计算任务。
CUDA Streams	CUDA 的并行执行机制，允许多个任务在不同的流中并行处理。	用于提升并行计算的效率。
CUDA Toolkit	包含编译器、库、开发工具和文档的完整开发工具包，用于 CUDA 编程。	用于开发和优化 CUDA 应用程序。
cuLitho	用于半导体制造的计算光刻技术	提升芯片制造的精度和效率
cuOpt	用于物流和路径优化的AI工具	提高物流效率和降低成本
Deep Learning SDK	NVIDIA 提供的软件开发工具包，包含用于深度学习的库和工具。	用于开发和优化深度学习应用。
DGX	专为数据中心设计的高性能GPU	提供大规模AI训练和推理能力
DGX	NVIDIA 的深度学习超级计算机平台，集成了高性能 GPU 和优化的软件栈。	用于高性能的深度学习训练和推理。
DLSS	AI超分辨率算法，通过较低分辨率输入预测更高分辨率输出	提升游戏帧率和图像质量
DPU	专用于数据处理的处理器	提升数据中心网络和存储性能
DRIVE Sim	用于自动驾驶汽车模拟训练的软件平台	提高自动驾驶系统的安全性和效率
ECC	一种能够检测和修正常见的数据损坏类型内存	提升数据完整性和系统稳定性
EGX	用于边缘计算的GPU平台	实现实时AI推理和分析
FP16 (Half Precision)	16-bit 浮点数精度，用于加速深度学习模型的训练和推理。	用于提高计算效率和减少内存占用。
GeForce	面向游戏娱乐领域的显卡系列	提供高性能游戏体验
GPU	专门用于处理图形和视觉计算任务的处理器	提供图形渲染和视觉计算能力
GPUDirect	NVIDIA 技术，允许直接在 GPU 之间传输数据，从而减少 CPU 和主内存的干预。	用于提升 GPU 之间的数据传输效率。
G-SYNC	同步显示器刷新率与显卡输出，减少画面撕裂	改善游戏和视频的视觉质量
HGX	用于高性能计算的GPU平台	支持科学研究和复杂计算任务
IGX	专为智能边缘设备设计的计算平台	提供安全、高效的边缘AI计算
Isaac	用于机器人开发的软硬件平台	简化机器人的设计、开发和部署

术语	定义	用途
Jetson	专为机器人和边缘设备设计的计算平台	提供AI和计算机视觉能力
Merlin	用于构建推荐系统的AI框架	提高推荐系统的性能和准确性
Metropolis	用于视频分析和监控的AI平台	提升视频分析的智能化水平
MIG	允许单个物理GPU被划分为多个独立实例的技术	提升资源利用率和灵活性
NeMo	用于生成式AI应用开发的框架	支持自然语言处理和语音合成
NVIDIA GFE	提供显卡驱动更新、性能监控和游戏优化的工具	简化显卡管理和游戏设置
NVIDIA Reflex	降低延迟，提升竞技游戏体验	优化FPS游戏中的操作响应
NVLink	用于连接GPU和CPU或其他GPU的高速通信接口	提升数据传输速度和系统性能
NVSwitch	NVIDIA 提供的高带宽交换技术，用于在多个 GPU 之间提供高速连接。	用于实现 GPU 集群中的高效数据交换和并行计算。
Omniverse	用于3D设计和实时协作的平台	提供虚拟世界构建和模拟的能力
Quadro	面向专业设计和可视化的显卡系列	提供精确的颜色再现和高质量的图形渲染
RAPIDS	用于数据分析和机器学习的加速平台	提高数据处理速度和效率
Replicator	用于生成合成数据的AI工具	提高AI模型训练的质量和效率
Riva	用于语音识别和合成的AI平台	提升语音交互的自然度和准确性
RTX	一种用于实时渲染画面，模拟接近现实的光照、阴影等效果的技术	增强游戏的沉浸感，提升画面逼真度
RTX AI PCs	配备RTX显卡的AI增强型个人电脑	支持AI加速的创意和工作流程
SMP	Symmetric Multiprocessing，对称多处理，指多个处理器共享内存和系统资源。	用于提升计算性能，支持多线程和多任务处理。
Studio	为创意专业人士设计的软硬件平台	提供专业级的内容创作工具
SuperNIC	高性能网络接口卡	提升网络通信速度和效率
Tensor Core	NVIDIA GPU 中专为加速深度学习运算而设计的核心。	用于提升深度学习模型的训练和推理性能。
TensorRT	NVIDIA 提供的高性能深度学习推理优化库。	用于优化和加速深度学习模型的推理过程。
Tesla	面向数据中心和高性能计算的GPU系列	提供强大的计算和数据处理能力
TGP	显卡的设定功耗，用于选择合适的电源	电源选择和功耗管理
Tokkio	用于创建虚拟形象和数字人的平台	提供虚拟形象创建和管理的能力
Triton	用于AI模型推理的服务平台	提供高性能的AI推理能力
V100	基于Volta架构的高性能数据中心GPU	提供高性能计算和AI推理能力
vGPU	虚拟化技术，允许多个虚拟机共享物理GPU资源	提升数据中心和云计算环境中的图形处理能力

发展历程

-发展阶段

-市场表现

一、发展阶段。英伟达创立至今经过了创业初期、GPU定义时期、GPU迭代时期、AI萌芽时期、AI辉煌时期五个阶段，并在2024年6月19日成为全球市值最高的公司

1993-1997 创业初期

- 1995：推出公司首款产品NV1
- 1997：推出全球首款128位3D处理器RIVA 128

1998-2006 GPU定义时期

- 1999：纳斯达克上市
- 2000：与台积电签约建立合作伙伴关系
- 2001：携nFORCE进军集成显卡市场
- 2004：推出SLI技术，允许多个GPU连接
- 2006：推出用于通用GPU计算的架构CUDA

2007-2014 GPU迭代时期

- 2007：推出Tesla GPU
- 2009：在第一届GPU技术大会上发布代号为“Fermi”的新一代CUDA GPU架构
- 2011：推出全球首款双核移动处理器Tegra 2
- 2012：发布基于Kepler架构的GPU
- 2013：发布首款完全继承的4GLTE移动处理器Tegra 4i；面向游戏玩家推出GeForce GTX TITAN
- 2014：发布192核超级芯片Tegra K1，推出Maxwell（第10代架构）

2015-2021 AI觉醒时期

- 2015：搭载TEGRA X1的DRIVE问世
- 2016：推出全球首款一体化深度学习超级计算机DGX-1
- 2018：推出开源GPU加速平台RAPIDS，推出Turning GPU架构
- 2019：在HPC、嵌入式、数据中心、自动驾驶汽车和专业图形市场中取得进展
- 2020：推出Ampere GPU架构助力数据中心，产品布局从GPU扩展至兼具DPU
- 2021：发布Grace CPU，专为巨型AI和高性能计算工作负载设计，成为具有CPU、GPU、DPU三种芯片的公司

2022-2024 AI辉煌时期

- 2022：公布最新自动驾驶技术平台Drive Hyperion 9；发布云原生AI模型和服务Omniverse Avatar Cloud Enginr(ACE)
- 2023：推出新一代GH 200 Grace Hopper超级芯片，同时推出基于GH200构建的新服务器设计
- 2024：推出NVIDIA Blackwell平台，以万亿参数规模推动AI计算的新时代；6月，英伟达市值登顶全球第一超越微软和苹果

第一阶段：英伟达从游戏显卡起家，并发布世界上第一款真正意义上的GPU产品，成为游戏显卡市场的领导者和GPU产品的定义者

创业初期 *(1993~1997)*

市场竞争和财务困境
创新精神和战略转型
在**图形芯片**市场取得重要胜利

GPU定义时期 *(1998~2006)*

成为专业可视化领域的领导者
在**GPU技术、游戏机合作、和显卡**市场竞争方面
取得重大进展
CUDA平台的开发和推广为AI铺路

GPU迭代时期 *(2007~ 2014)*

在游戏显卡市场中占据**领导地位**
在高性能计算领域取得进展
推出革命性GPU架构和产品，实现技术突破

第一阶段

1993年，从“狂野西部”通用图形计算起步

1995年，推出第一个产品：针对游戏主机的NV1显卡

1995年，破产危机

1997年，推出关键芯片产品Riva128 (NV3)，首款128位的3D处理器，在图形芯片市场上开始崛起

第二阶段

1999年，借GPU GeForce 256拿下微软X-BOX订单，同年IPO成功

2000年，收购3dfx，市场形成英伟达与ATI的双雄格局

2001年，推出业界首款可编程GPU: GeForce 3，使科学计算开始能够利用GPU完成，为训练AI大模型埋下伏笔

2005年，为索尼PS3开发处理器，收购核心逻辑开发商

2006年，CPU巨头AMD收购ATI后由于资源限制导致显卡份额下滑；英伟达推出通用平行计算平台CUDA，建立CUDA研究中心，后成为深度学习和AI训练的首选GPU架构，助力英伟达取得领先市场地位

第三阶段

2007年，Tesla GPU 问世，标志着NVIDIA在高性能计算领域的重大突破，使得GPU的强大计算能力应用于药物发现、医学成像等领域

2010年，提升CUDA核心至512个

2014年，完全实现Maxwell设计架构；在2010年全球最快超算前五强中借助NVIDIA Tesla GPU占领三个地位，在2012年最快的超级电脑泰坦中有18,688颗基于Kepler的NVIDIA Tesla芯片

第二阶段：2015年英伟达首次发布了面向深度学习处理器，GPU芯片构架向AI方向大步迈进，目前人工智能芯片方向的以市占率80%处于绝对垄断的地位

发展阶段及行业地位	重点领域	重要节点
AI觉醒时期（2015-2021） 硬件供应商 + 深度学习和人工智能时代的先行者和推动者	GPU技术	2015：推出用于训练深度神经网络的最强大的处理器NVIDIA GeForce GTX TITAN X芯片架构开始往深度学习和AI方向前进
	深度学习	2016：发布Volta架构和NVIDIA Tesla® V100 GPU加速器极大提升AI运算效率 推出16nm Pascal架构引入新的总线和通信协议NVLink，用于GPU和CPU，以及多个GPU间的连接
	AI应用等领域	2017：Ampere架构对Tensor Core进行了进一步的升级，增加了对稀疏矩阵计算的支持 推出Turing架构，在Tensor Core上增加计算能力，同时实现游戏中可用的实时光线追踪渲染此结构下的Tesla T4 GPU计算加速器成为经典AI算力卡
AI辉煌时期（2022~今） 人工智能方面处于绝对垄断的地位（市占率80%）	AI	2022：发布4nm工艺的RTX 40系列显卡，性能比上一代提高50%。 发布CPU产品Grace：这款超级芯片结合Grace和Hopper架构，为AI和高性能计算提供了强大的CPU + GPU组合
	数字生物	
	气候科学	
	游戏	2023：发布支持设计、开发、部署和管理的Omniverse Cloud，实现先进工厂数字化 推出GH200U以及游戏AI模型代工服务平台Avatar Cloud Engine (ACE) for Game
	自动驾驶汽车	2024：发布Blackwell GPU B100，芯片尺寸接近EUV光刻技术理论极限（858平方毫米），含有1040亿个晶体管 发布Rubin GPU、Vera CPU等。Omniverse平台提供方案模拟、合成数据生成和高保真可视化等功能，帮助制造商优化工厂布局、测试机器人系统并提高工人安全 公布人形机器人项目GR00T，标志着其在人形机器人领域的重大进展
	机器人等领域	

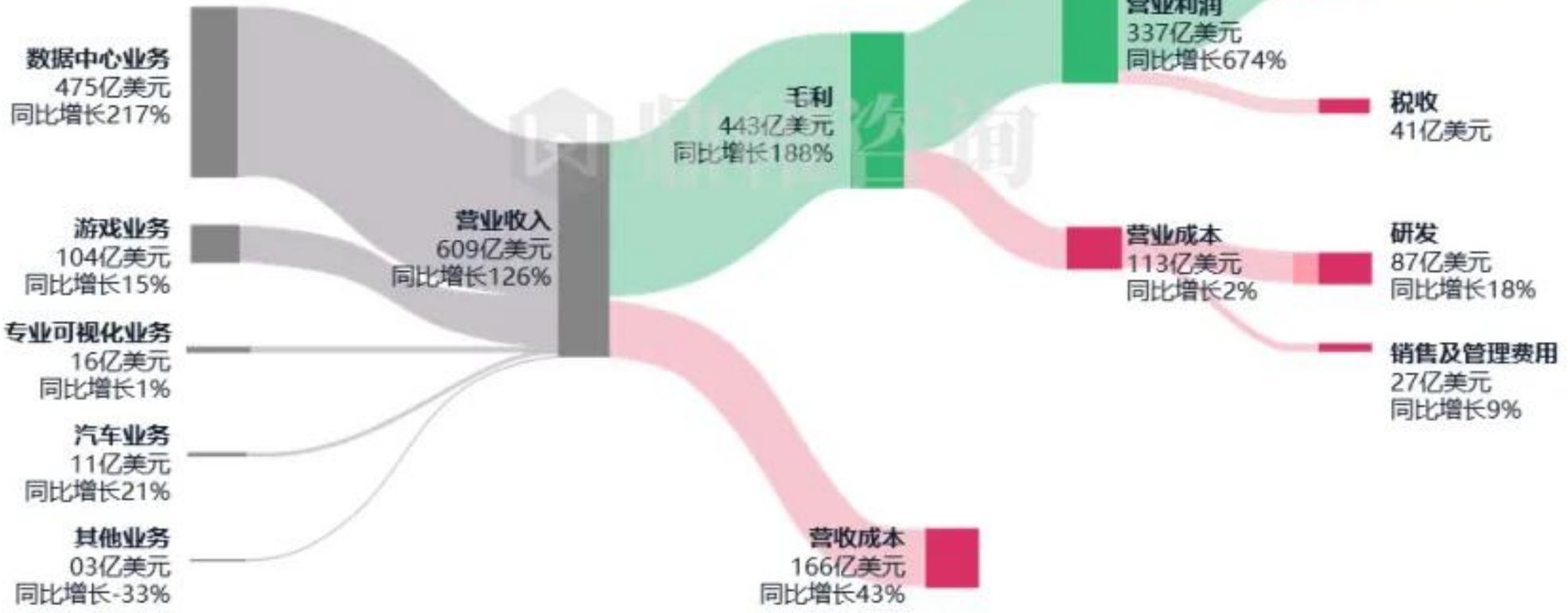
二、市场表现。英伟达从2024财年业务增长迅猛，营收达到609亿美元，净利润达到惊人的298亿美元

NVIDIA

24年财政年度利润表

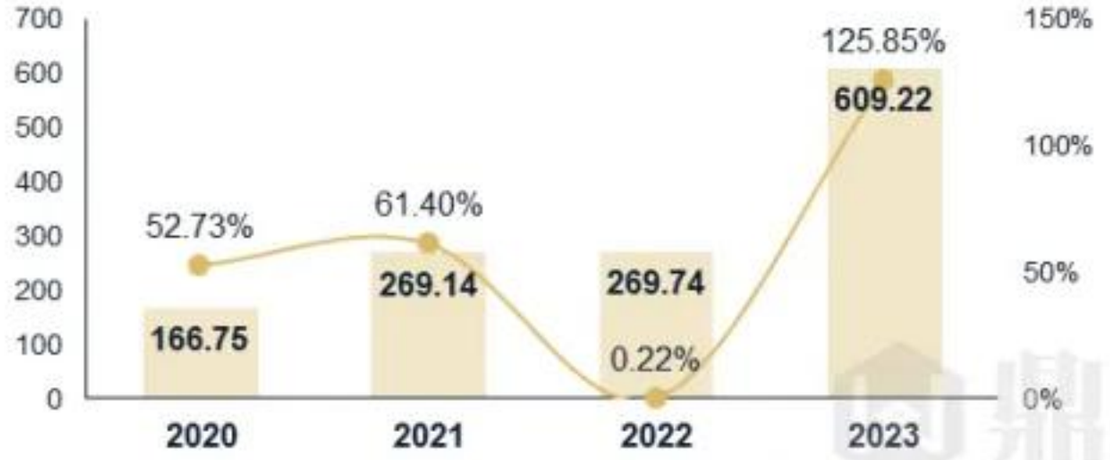
截至2024年一月二十八日

资料来源:NVIDA投资者关系

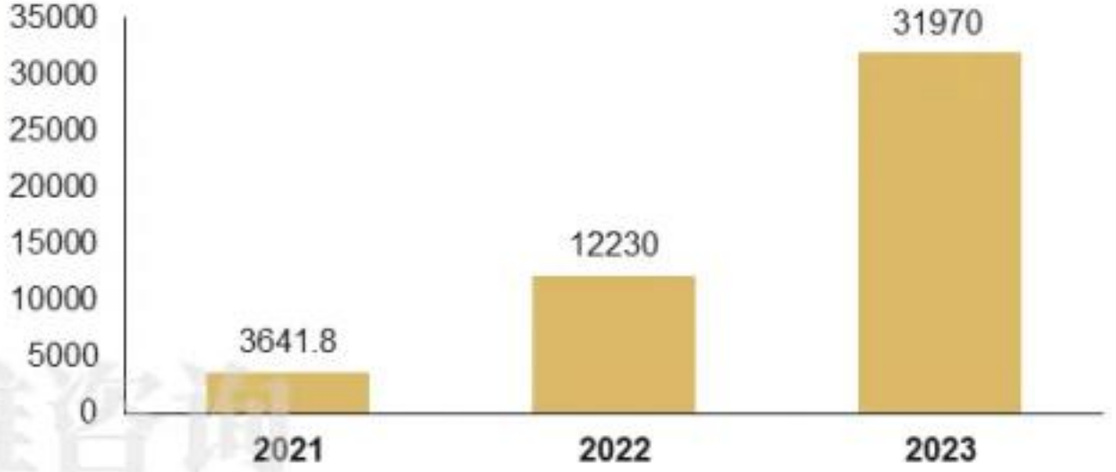


2020年至2023年，公司营业收入和净利润均呈现波动上升趋势，其中2023年营业收入中数据中心业务贡献最大，占比超过70%

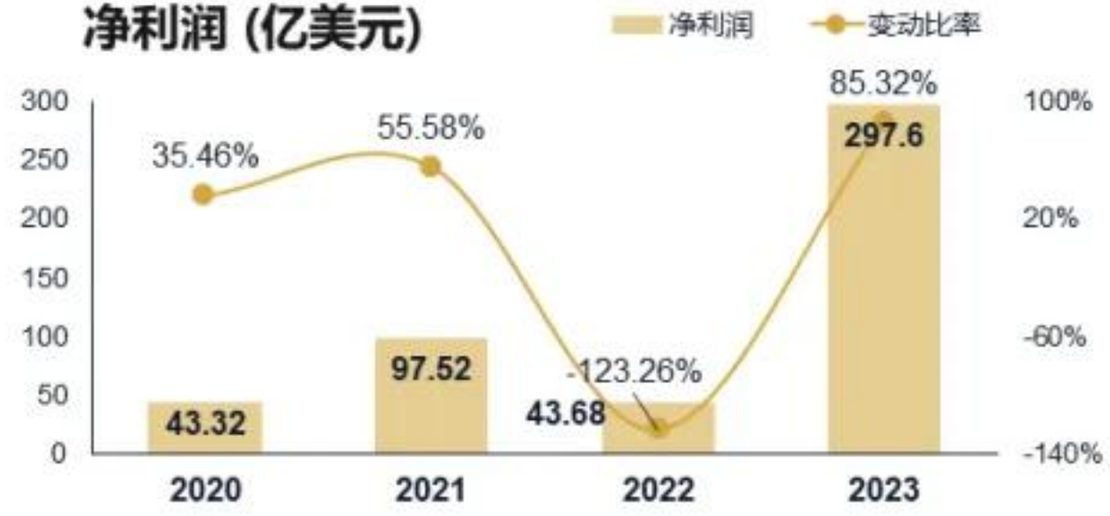
营业收入(亿美元)



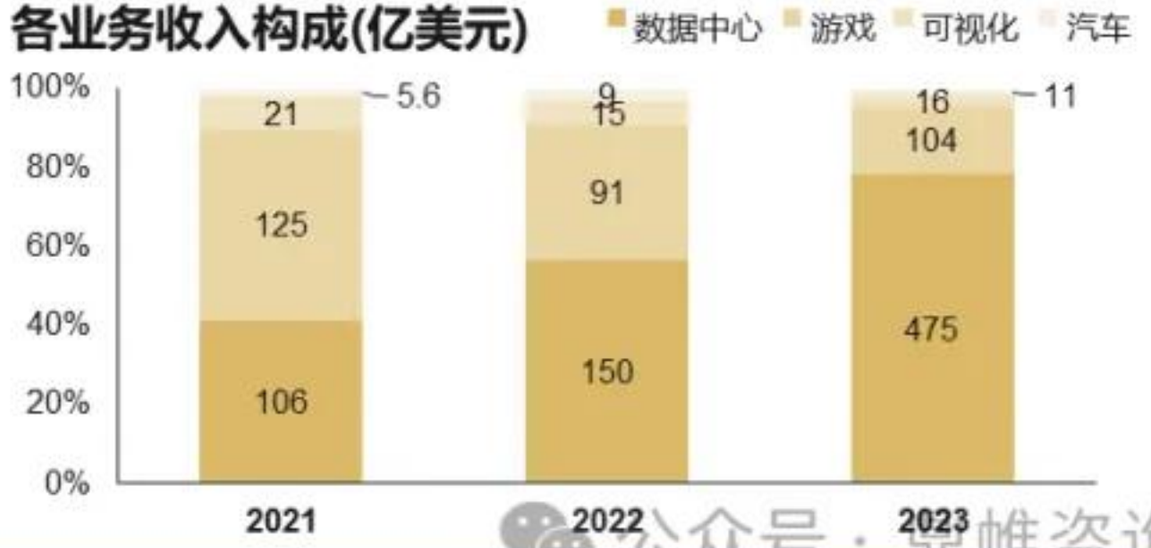
各业务市值(亿美元)



净利润 (亿美元)



各业务收入构成(亿美元)



从2021年到2023年，研发费用和销售与管理费用均呈上升趋势，销售与管理费用率下降比例高于研发费用率，研发费用总投入过去三年增长约150%，销售与管理费用增加约40%

发展战略

-发展战略

-产品策略

-CPU

-GPU

-DPU

一、发展战略。英伟达通过对计算机底层技术的全面更新，以“CPU+GPU+DPU”三大芯片为产品底座，推动英伟达成为人工智能计算领域的领导者

愿景：人工智能计算领域的领导者

使命	创造下一个工业革命，推动人类进入人工智能和图形计算的新纪元
目标	**计算机**：改变计算的工作方式和计算机的功能 **数据中心**：推动全球的数据中心行业现代化 **人工智能**：全球人工智能基础设施的引擎 **机器人**：推动人工智能机器人和工业数字化
路径	**拓宽AI基础算力应用广度，实现全方面覆盖**：加强对垂直行业的服务；研发单一行业适配底层软件；目标市场如医疗、自动驾驶 **提前布局AI工业机器人领域**：提供可供于AI工业机器人训练和操作的GPU，及其对应调试、训练用的底层软件
产品	三芯战略 CPU 中央处理器 GPU 图像处理器 DPU 数据处理器

技术基石：英伟达把计算机整套技术重新发明了一遍

GPU架构	系统互联	系统	软件	网络技术
集成Tensor Cores技术的AI芯片架构 Volta Turing Ampere Hopper Blackwell	NVLink 高速互连技术 NVSwitch 网络交换机 NVLink-C2C 芯片到芯片、裸片到裸片的互连技术	DGX 专为企业级AI设计平台 EGX 加速边缘计算的平台 IGX 工业级边缘AI平台 HGX 灵活定制的AI硬件平台 MGX 模块化灵活的计算平台	cuDNN深度神经网络的GPU加速库 TensorRT 深度学习推理引擎 NCCL 构建多GPU和多节点并行应用程序通信库 NVSHMEM 为跨多个GPU内存数据创建全局地址空间库	InfiniBand 技术 SHARP 技术 ASAP 技术 Spectrum-X 网络平台

英伟达的产品架构包含以“CPU+GPU+DPU”为中心的硬件产品、以CUDA为核心的软件生态、以NVIDIA AI+ NVIDIA Omniverse为主的平台和各个场景下的应用框架

英伟达凭借快速迭代的研发能力、供应链的主导地位、多种营销策略、扁平化的组织、多元投资合作支撑游戏、专业可视化、数据中心、汽车四大业务发展，进一步巩固企业在AI市场领导地位

愿景：人工智能计算领域的领导者

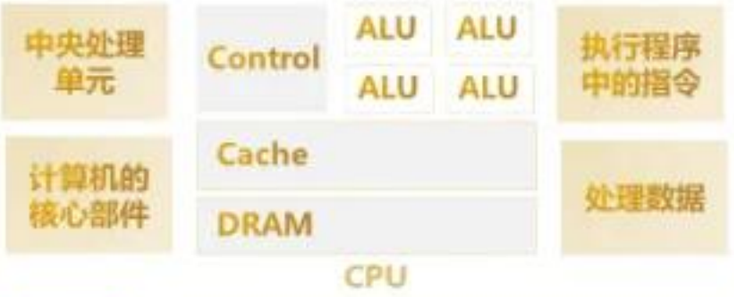
使命	创造下一个工业革命，推动人类进入人工智能和图形计算的新纪元			
目标	**计算机** 改变计算的工作方式和计算机的功能	**数据中心** 推动全球的数据中心行业现代化	**人工智能** 全球人工智能基础设施的引擎	**机器人** 推动人工智能机器人和工业数字化
路径	拓宽AI基础算力应用广度，实现全方面覆盖 加强对垂直行业的服务 / 研发单一行业适配底层软件 / 目标市场如医疗、自动驾驶		提前布局AI工业机器人领域 提供可供于AI工业机器人训练和操作的GPU，及其对应调试、训练用的底层软件	
业务	**游戏业务** GeForce RTX 系列显卡 GeForce NOW 云游戏服务 AI PC	**专业可视化业务** Quadro和RTX系列GPU专业显卡 NVIDIA Omniverse Enterprise平台 专业工作站解决方案 虚拟GPU (vGPU)	**数据中心业务** AI芯片级芯片定制 交换机、网卡等网络产品 超级计算机及AI工厂 国家主权AI、电信云 边缘计算平台jetson 云计算业务（量子计算） 企业 Nvidia AI Enterprise ）	**汽车业务** 自动驾驶芯片 自动驾驶端到端方案 虚拟工厂规划 仿真设计
应用场景	AI+医药 / AI+汽车 / AI+机器人 / AI+军工 / AI+零售 / AI+电信 / AI+娱乐 / AI+高性能计算 / AI+交通 / AI+安全			
保障体系	**研发保障** 三团队-两季度高速迭代创新 超高研发人员占比 研发机构与研发合作伙伴 14研发实验室布局 26大研发领域 英伟达专利墙	**营销保障** 会议营销 社区营销 社交媒体营销 品牌营销 内容营销 全球化营销 社会责任营销	**供应链保障** “胡萝卜加大棒”的管理方式 大订单挤占GPU短缺组件供应 承诺订单不可取消确保巨大供应 提前支付预付款抢占供应 人工智能重塑英伟达产业链地位	**管理保障** 创始人技术专业与销售背景兼具 极致扁平化的组织架构 赋权式管理，不给工作建议 摒弃会议汇报 摒弃传统战略规划 向全体员工透明共享	**投资合作** 投资领域：机器人、人工智能、游戏、软件与硬件、可视化、云计算与数据、网络、自动驾驶、生物科技与医疗等 初创投资项目 企业间的三级合作伙伴 国家政府合作

二、产品策略。未来算力生态以CPU、GPU、DPU为三大核心算力芯片，英伟达、英特尔、AMD及其他代表企业抢占各芯片市场份额

算力芯片	核心功能	代表企业
CPU	▪ 系统管理 ▪ 维持软硬件生态 ▪ 应用程序	intel、AMD、ARM、Qualcomm
GPU	▪ 规则计算 ▪ 科学计算 ▪ 数据集并行运用 ▪ CUDA的核心	NVIDIA、AMD、intel
DPU	▪ 异构计算 ▪ 数据中心基础设施 ▪ “Datacenter Tax” ▪ 卸载网络、存储、安全业务	intel、SYSTeDATA 中科数控、CISCO、BROADCOM

英伟达在通用计算芯片CPU基础之上，开创当前GPU加速芯片时代，并将DPU作为第三颗主力芯片作为构建未来算力的基石和底座

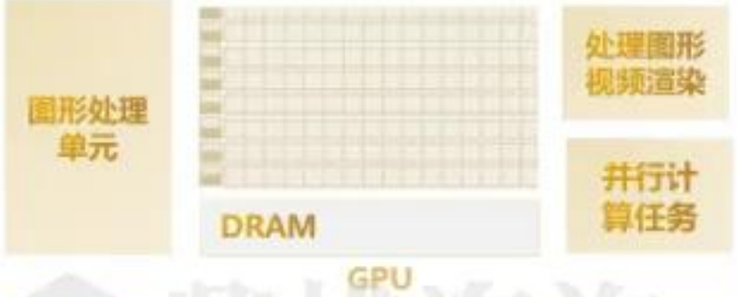
CPU 通用计算



功能	使软件和硬件解耦，实现更高 IPC 和更高频率。
应用	软件基于CPU构建庞大生态，如：x86 架构服务器端还是 ARM 架构移动端
应用领域	PC（一颗CPU） 和服务器（数量不定）
类型	分为一路、双路、四路及以上服务器；以双路服务器为主

计算生态的底座，主力芯片的基石

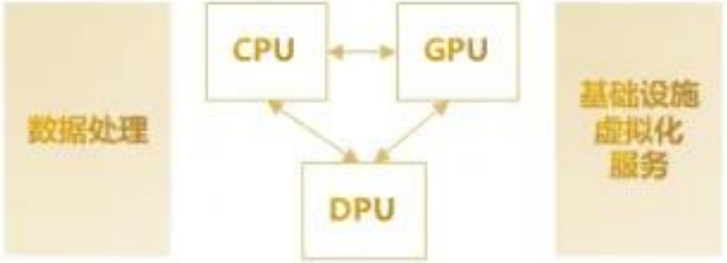
GPU 加速计算



特点	架构复杂度最高的芯片之一
功能	并行计算、浮点以及矩阵运算，是高性能计算最重要的辅助计算单元
英伟达产品	• CUDA使 GPU 处理复杂计算问题，开发者可使用 C 语言来编写程序，降低了用户基于GPU 并行编程门槛 • 对不同场景构建的开发库和中间件，建立了“GPU+CUDA”的强大算力生态

从图形处理到数据处理芯片蜕变

DPU 数据处理



主要应用场景	• 数据中心 • 智能驾驶
英伟达产品	• 预计未来用于数据中心的 DPU 数量将达到和数据中心服务器同等量级 • 英伟达、英特尔等厂商数据处理类芯片 DPU/IPU 大规模量产，全球 DPU 市场将在未来几年迎来爆发式增长

因数据中心而生的“第三颗主力芯片”

基于未来以异构计算为主的算力发展趋势和片上模式为主的数据中心主流形态，英伟达陆续通过自研及并购形成了GPU+CPU+DPU的三芯布局，实现为客户提供更加全面、高效的计算解决方案

异构计算
未来算力需求的重要发展趋势

CPU、GPU、DPU共存的片上模式
未来数据中心主流

“3U”一体（即CPU、GPU、DPU）重塑数据中心算力架构

GPU

提供加速计算的通用算力的基础保障

英伟达自1995年推出首款GPU以来，不断创新发展，成为图形处理行业的领导者。其GPU产品涵盖了从消费级显卡到高性能计算设备的广泛应用。

DPU

针对数据中心和网络设备的需求
具有高效处理数据包和协议的能力

英伟达在2020年宣布收购Mellanox后，开始涉足数据处理单元(DPU)领域。通过整合Mellanox的技术和资源，英伟达在数据中心和网络设备领域加强了布局为DPU市场发展奠定了基础

CPU

能够更好地应对各种计算任务，尤其是那些需要快速逻辑判断和高度并行处理能力的应用

2021年4月，NVIDIA发布首款代号为“Grace”的CPU其专为巨型AI和高性能计算工作负载设计

GPU+CPU+DPU协同互补
在数据中心和边缘端达到高性能与高安全性

“三类芯片、逐年飞跃、一个架构”
为客户提供更加全面、高效的计算解决方案

计算机产业伴随着兼容机的出现，产业结构从过去的纵向一体化转变为横向切片式，这种水平分层的产业形态造就了计算机产业基础层强惯性、高垄断性的特征

旧式纵向型计算机产业

新的横向式计算机产业

(并非按比例展示)　约1995年

销售与分销	零售店	超市	商家	邮购		
应用软件	Word		Word Perfect	其他		
操作系统	DOS及Windows		OS/2	Mac	UNIX	
计算机	康柏	戴尔	Packard Bell	惠普	IBM	其他
芯片	英特尔架构		摩托罗拉	精简指令集计算机		

底层结构的碎片化会带来上层重复工作量和成本的大幅增加 → 软件层面对于底层硬件高度统一的强烈诉求 → 计算机产业底层硬件强惯性、高垄断性特征替代几乎无法发生

英伟达通过对计算机产品的重新定义、计算机形态的演化竞争以及渐进式改造，逐步塑造计算机形态向有利于自身的计算机演化方向

①产品品类的重新定义

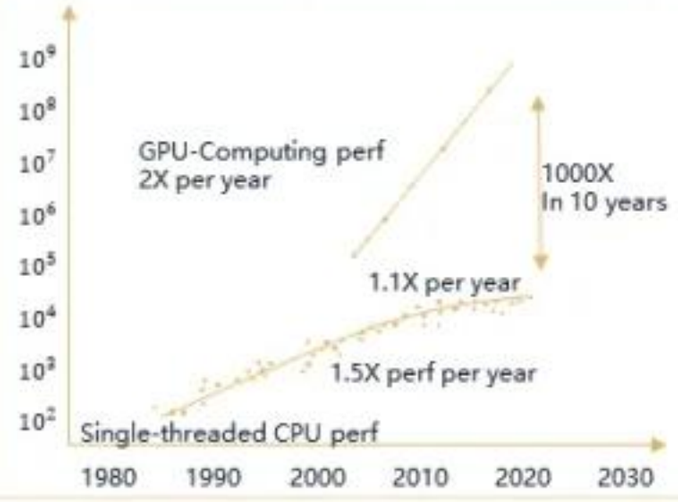
为了在英特尔每18个月更新一代CPU的策略下不显得被动，英伟达每6个月发布一次新品，算力性能始终无法被CPU集成，GPU作为一个单独的品类存活

摩尔定律

集成电路上可容纳的晶体管数目，每隔18至24个月会增加一倍，性能也将提升一倍，而价格下降为原来的一半

采用“三团队-两季度”的创新研发迭代模式

即三个并行开发团队专注于独立的分阶段产品开发，确保公司每6个月推出一次新产品，与行业市场产品周期保持一致，并领先市场1-2个研发周期



②计算机形态的演化竞争

英伟达没有试图从CPU上取代英特尔，而是将其为CPU+GPU，给英特尔塑造计算机系统应该是纯CPU还是CPU+GPU的形态的难题

CPU ✖ 替代 GPU

CPU 逐步演进 CPU+GPU

对英特尔来说，无论选择哪种方案，都是在英伟达设定的框架内竞争

- 如果英特尔选择只做CPU，它必须接受CPU加GPU的体系也能提供改进；
- 如果做GPU，将直接与英伟达竞争，可能导致GPU在体系中的重要性逐渐超过CPU。最终，大家都专注GPU，问题就变成了谁的GPU更强

竞争并不是关于设备的竞争，而是关于利用CPU的竞争，站在巨人的肩膀上，把这个产业带到只有我们能够带到的地方——2009年黄仁勋采访

③渐进式改造

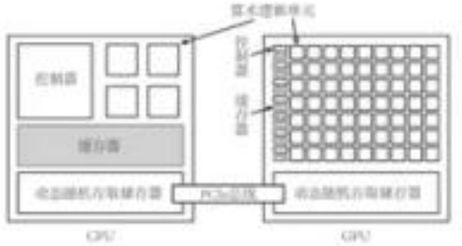
通过推出3D图形加速卡进入市场，吸引众多游戏玩家

逐渐改变计算机的形态，使独立显卡成为PC标配

在数据中心市场，推动在服务器中加入少量显卡，从仅使用CPU转变为结合GPU的形态

随着显卡数量增加，权重增大，加入技术和特性，服务器转变为CPU必须结合GPU的形态

云端服务器的标准配置通常是2U式机架和8个GPU，将英特尔挤到相对边缘的位置



英伟达通过构筑软件生态、调动开发者、发掘应用场景去塑造有利于自身的计算机演化方向

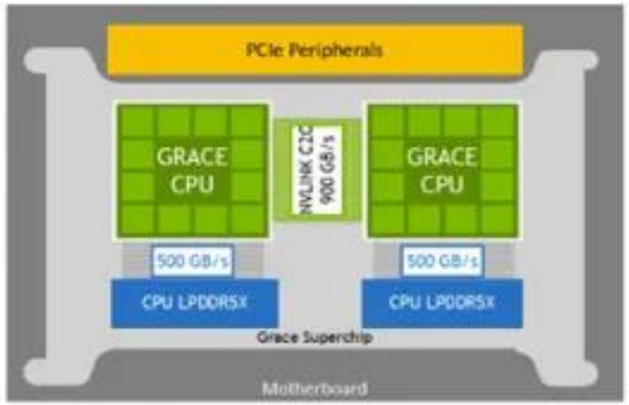
英伟达发布2025-2027年产品规划，AI芯片规划的战略核心是“One Architecture”统一架构，AI芯片从两年一次的更新周期转变为一年一次的更新周期，下一代Rubin架构及其对应核心芯片已提上日程

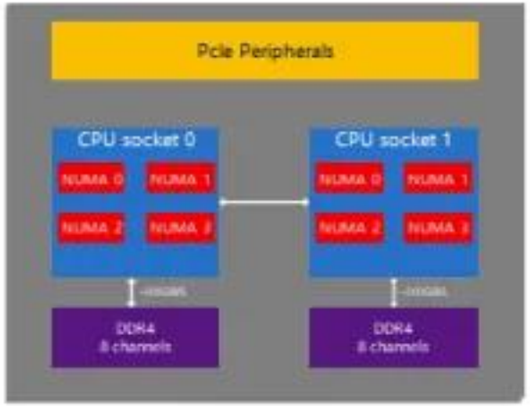
	Hopper 架构	Blackwell 架构	Rubin 架构
GPU	Hopper GPU 6S HBM3 Hopper + GPU 6S HBM3e	Blackwell CPU 8S HBM3e Blackwell Ultra CPU 8S HBM3e 12H	Rubin GPU 8S HBM4 Rubin Ultra GPU 12S HBM4
CPU	Grace CPU		Vera CPU
NVLINK	NVLink Switch 900 GB/sec	NVLINK 5 SWICH 1800 GB/sec	NVLink 6 switch 3600 GB/sec
NIC	CX7 SuperNIC BF3 Super NIC	CXB SuperNIC	CX9 SuperNIC 1600 GB/sec
SWITCH	Quantum-X400 Infiniband Switch	Spectrum-X800 Ethernet Switch Quantum-X800 Switch Spectrum Ultra X800 Ethernet Switch 512-Radix	X1600 IB/Ethernet Switch

2022 2023 2024 2025 2026 2027

数据中心规模 ● 一年节奏 ● 技术限制 ● 一个架构

2.1CPU：英伟达自研Grace CPU，在ARM架构和自身技术优势基础上满足数据中心业务对于CPU的性能需求





背景	市场技术需求	传统的 CPU 架构难以满足AI及高性能计算对计算能力和效率要求
	自身技术优势	NVIDIA 公司凭借其在图形处理和并行计算方面的技术优势，推出 Grace CPU

Grace CPU

- **新一代数据中心的基础**，面向服务器和数据中心生态系统，可采用多种配置满足不同数据中心的需求
- **超强性能与效率**：采用 ARM 架构，拥有 72 个 Arm Neoverse N2 内核，配备 LPDDR5x 内存，具有超强的性能和效率
- **双重使用能效**：可与 GPU 紧密结合以增强加速计算能力，也可作为强大而高效的独立 CPU 进行部署

支持平台	Grace Superchip	Grace Hopper Superchip
	两个Grace CPU，共144个内核	72核Grace CPU + Hopper H200 GPU

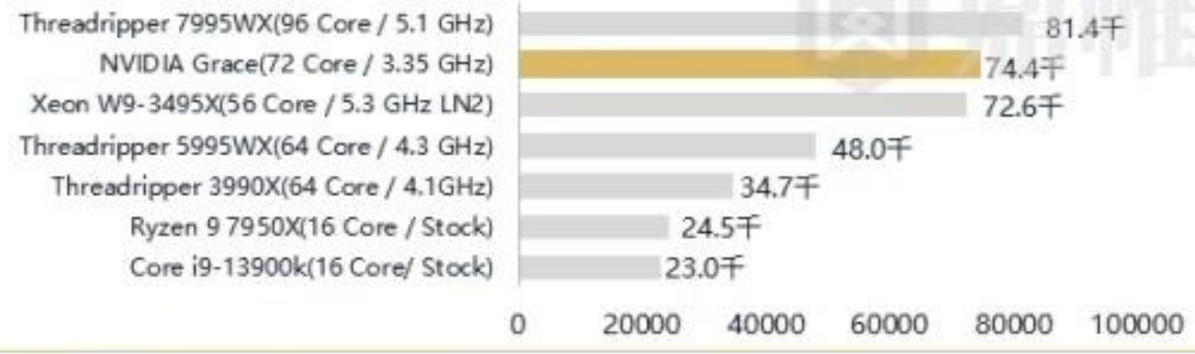
核心架构	采用 Arm V 系列基础架构，CPU 内核中的最新产品—Arm Neoverse V2 CPU 架构，经过优化后提供领先的每线程性能，能效更高
SIMD 指令	实现了两个单指令多数据（SIMD）向量指令集，可加速机器学习、基因组学和密码学等关键 HPC 应用程序。
原子操作	Grace CPU 在 Arm v8.1 中首次引入的大型系统扩展（LSE），提供低成本的原子操作，提高 CPU 到 CPU 通信、锁和互斥锁的系统量
缓存架构	由 NVIDIA 设计的可扩展一致性结构（SCF）是一种网状结构和分布式缓存架构，提供超过 3.2TB/s 的总二分带宽
内存子系统	Grace CPU Superchip 使用高达 960GB 的服务器、级低功耗 DDR5X（LPDDR5X）内存和 ECC，实现带宽、能效、容量和成本的最佳平衡
I/O连接	Grace CPU Superchip 支持 128 条PCIe Gen 5 通道和128GB/s 的双向带宽，可分为 2x8 个以提供额外的连接，并支持各种 PCIe 插槽形状

英伟达目前Grace CPU性能表现出色，专为高性能计算及数据中心打造，并将在2026年推出下一代Vera CPU，计划于未来几年推出消费者级别的CPU产品

NVIDIA Grace CPU 技术亮点

高性能CPU	超级芯片设计	卓越的相干接口	高封装密度	强悍的每瓦性能	极强的兼容性
适用于高性能计算（HPC）和云计算	最多配备 144 个 Arm v9 CPU 内核，首款配备 LPDDR5x	900 GB/s，相比 PCIe Gen 5 快7倍	是 DIMM 解决方案的两倍	当前领先 CPU 的两倍	可运行所有 NVIDIA 软件堆栈和平台

NVIDIA Grace CPU 性能表现 Geekbench5多线程测试比较



英伟达CPUVS英特尔CPU

维度	Grace CPU架构	Intel CPU架构
芯片互连技术	采用了NVIDIA独特的NVLink-C2C互连技术（最大区别）	
指令集架构	Arm 指令集架构	x86 指令集架构
核心数量和线程数量	更多的核心数量和线程数量	核心数量和线程数量相对较少
缓存架构	分布式缓存设计	层次式缓存设计
内存子系统	低功耗的 LPDDR5X 内存	通常采用 DDR4 或 DDR5 内存
集成度	将 CPU、GPU 和其他组件集成在一个芯片上	通常需要与其他芯片配合使用

价值定位

英伟达在异构计算领域的进一步拓展
在高性能计算和服务器市场迈出的重要一步
为AI工厂等新型数据中心提供强大的计算支持

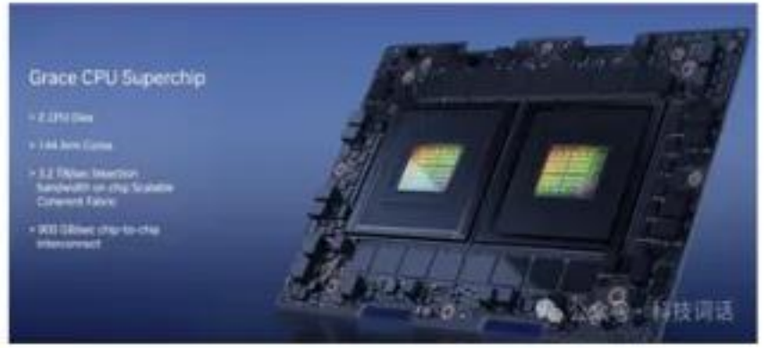
应用领域

超级计算领域： 构建高性能计算系统，为科学研究和工程应用提供强大的计算支持
人工智能领域： 与 NVIDIA 的 GPU 相结合，为训练和推理任务提供高效的计算平台

未来规划

在 2026 年推出其下一代 **Vera CPU**
在未来几年内推出基于 Arm 的**消费者级别CPU**

NVIDIA Grace CPU 软件生态系统将用于 CPU、GPU 和 DPU 的全套 NVIDIA 软件与完整的 Arm 数据中心生态系统相结合



遵循主流 CPU 设计原则

- CPU 符合 Arm 服务器基础系统架构 (SBSA)，以支持符合标准的硬件和软件接口
- 支持 Arm 服务器基本引导要求 (SBBR)，以支持基于 Grace CPU 的系统上启用标准引导流程

提供符合标准的平台

- NVIDIA HPC SDK 和每个 CUDA 组件都有 Arm 原生安装程序和容器。
- 所有主要的 Linux 发行版，都可在 NVIDIA Grace CPU 上完美运行，无需修改
- 编译器、库、工具、分析器、系统管理实用程序以及用于容器化和虚拟化的框架可轻松地在 NVIDIA Grace CPU 上安装和使用
- NVIDIA GPU Cloud (NGC) 还提供深度学习、机器学习和针对 Arm 优化的 HPC 容器

2.2GPU：GPU主要包括核心、显存、流处理器等部件，单核心数量远高于CPU，因此具有高效的并行处理能力，在处理海量数据及加速计算时体现出独特架构优势

- GPU(图形处理单元)是一种专门设计用于快速处理图像和视频渲染的电脑硬件它是现代计算机系统中的关键组件，特别是在处理图形密集型任务时

GPU的构造和主要组件

核心 (Cores)	GPU 的核心类似于 CPU 的核心，但数量通常远多于 CPU，这些核心是 GPU 并行处理能力的基础,使其能够同时处理大量的图形和计算任务。
显存 (Video Memory)	GPU具有自己的专用显存，通常为GDDR(图形双倍数据速率)类型。显存用于存储图形和视频数据，其高速度和大带宽特性对于高效图形处理至关重要。
流处理器 (Streaming Processors)	这些是 GPU 内部的小型处理单元，负责执行图形和某些类型的并行计算操作，流处理器的数量可以从几百到几千不等，这直接影响 GPU 的处理能力。
着色器 (Shaders)	着色器是运行在 GPU 上的小程序，用于渲染图像的不同部分，如顶点着色器、像素着色器和几何着色器。
光栅化单元 (Rasterization Units)	光栅化是将 3D 图像转换为 2D 图像的过程。光栅化单元负责这一转换过程，是3D图形渲染的关键部分。
纹理单元 (Texture Units)	负责处理图像的纹理映射和滤波，这对于生成逼真的图像非常重要。
输出接口 (Output Interfaces)	这些接口，如HDMI或DisplayPort，负责将渲染的图像发送到显示设备。
高级图形特性支持	现代 GPU 支持高级图形特性，如实时光线追踪(RTX 技术)和AI驱动的图像优化技术(如DLSS)

- GPU 的设计重点在于并行处理能力，能够同时处理大量的图形数据，这使得它在视频游戏、图形设计、视频编辑和某些类型的计算密集型任务中表现出色

GPU结构

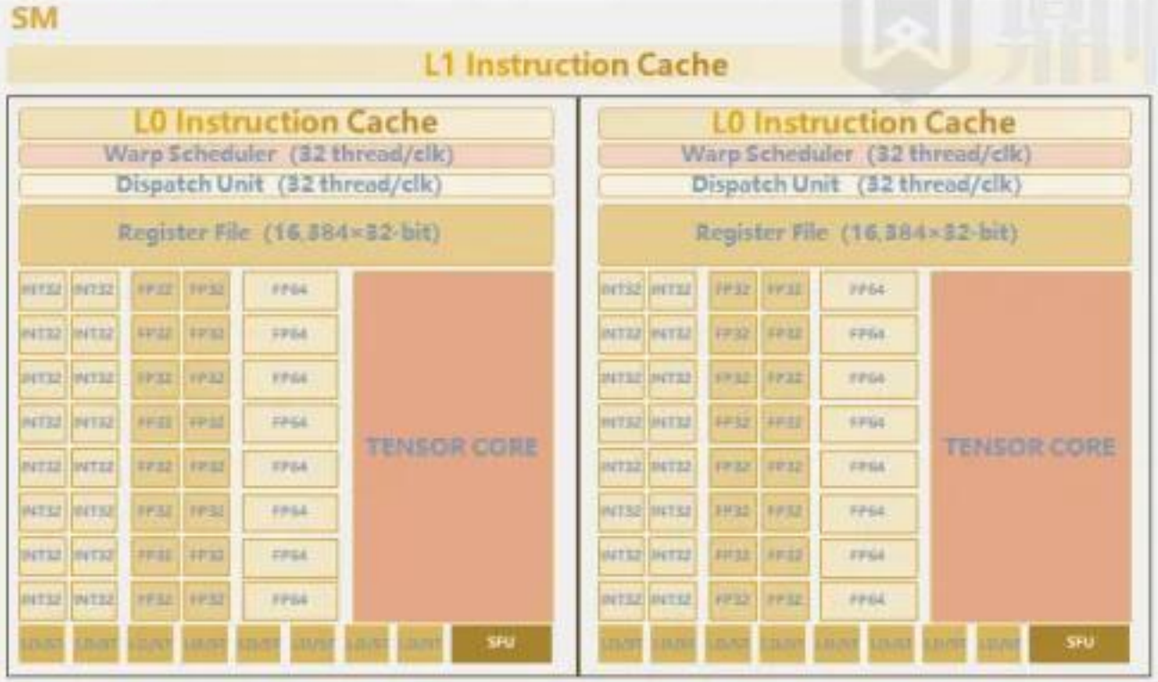
英伟达两年一迭代的速度持续更新GPU架构，费米是世界上第一款真正意义上的GPU架构，帕斯卡是首个加入了支持深度学习功能的架构，伏特是真正意义上的第一款AI芯片

架构代号	Tesla	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper	Blackwell
中文代号	特斯拉	费米	开普勒	麦克斯韦	帕斯卡	伏特	图灵	安培	赫柏	布莱克威尔
时间	2006	2009	2012	2014	2016	2017	2018	2020	2022	2024
制程	-	40nm	28nm	28nm	16nm	12nm	12nm	8nm	4nm	4nm
晶体管	14亿	30亿	71亿	80亿	135亿	211亿	146亿	540亿	800亿	2080亿
规模	484mm²	466mm²	551mm²	601mm²	610mm²	815mm²	754mm²	826mm²	814mm²	814mm²
核数	128个	16个SM*32CUDA Core，共计512个CUDA Core	15个SMX*(192个单清度+64个双精度CUDAcore)	具有3072个CUDA核心	3840个CUDA核心	5120个CUDA核心，新增了640个张量核心	4608个CUDA核心、576个张量核心	具有6912个CUDA核心、432个张量核心	具有18432个FP32CUDA核心，576个张量核心	-
特点	首个通用GPU计算架构，采用全新的CUDA架构，支持使用C语言进行GPU编程，标志着GPU开始从专用图形处理器转变为通用数据并行处理器	引入L1/L2快速缓存、错误修复功能以及GPU Direct技术	首个支持超级计算和双精度计算的GPU架构，计算能力比Fermi架构提高3-4倍，GPU开始成为高性能计算的关注点	在功耗效率、计算密度上获得重大提升，计算密度是Kepler的两倍，标志着GPU的节能计算时代到来	增强了GPU的能效比和计算密度，功耗只有300W，比Maxwell架构提高50%以上，这使得GPU可以进入更广泛的人工智能、汽车等新兴应用市场。	AI计算能力达到112TFLOPS，比Pascal架构提高了近3倍，可以大大加速人工智能和深度学习的训练与推理	新增了RayTracing核心(RTCore)，可硬件加速光线追踪运算	在人工智能、光线追踪和图形渲染等方面性能大幅跃升，功耗却只有400W，能效比显著提高	HopperTransformer引擎可以做到FP16和FP8之间逐层交换，利用NVIDIA提供的启发算法来降低所需精度	包含2080亿个晶体管，采用双倍光刻极限尺寸的裸片，通过10TB/S的片间互联技术连接成一块统一的GPU

Tensor Core是英伟达GPU成为AI芯片的关键技术，加速深度学习模型训练和推理能力，目前迭代至第五代

Tensor Core 加速AI性能

背景	▪ 深度学习技术的飞速发展对计算能力提出了更高要求，传统的CUDA Core在处理复杂的大规模矩阵运算和卷积任务时力不从心
定义	▪ Tensor Cores专门针对深度学习模型训练和推理中的常见操作进行优化，是一种特殊的硬件加速器，被英伟达设计用于GPU
功能	▪ 在保持模型精度的同时大幅提升计算效率



Tensor Core 发展迅速

第一代 (Volta Tensor Core)	▪ GPU 的潜在吞吐量提高了多达 12 倍，与前一代的 Pascal GPU 相比，旗舰V100的640个核心提供高达5倍性能提升速度
第二代 (Turing Tensor Core)	▪ Tensor Core 精度从 FP16 扩展到包括 Int8、Int4 和 Int1，可以将 GPU 的性能吞吐量加速至比 Pascal GPU 高出多达 32 倍，且首次在消费级产品中配备Tensor Core
第三代 (Ampere Tensor Core)	▪ 将计算能力扩展到 FP64、TF32 和 bfloat16 精度，TF32 格式与 FP32 类似，但可实现高达 20 倍的速度提升
第四代 (Hopper Tensor Core)	▪ 随着 Hopper 微架构发布，NVIDIA 声称这将使大型语言模型的训练速度“比上一代快 30 倍”
第五代 (Blackwell Tensor Core)	▪ 2024年NVIDIA发布第五代Blackwell Tensor Core，与上一代相比，可为大型模型提供30倍的加速，且提供更高的准确性和精度

	Blackwell	Hopper
Tensor Core 支持的精度	FP64, TF32, BF16, FP16, FP8, INT8, FP6, FP4	FP64, TF32, BF16, FP16, FP8, INT8
CUDA Core 支持的精度	FP64, FP32, FP16, BF16	FP64, FP32, FP16, BF16, INT8

最新Blackwell架构搭载于GB200、B100、B200，推动数据中心进一步转型，为智能工厂奠定技术基础

结构与特点

- Blackwell架构采用了TSMC的4NP制程技术，特点是搭载了2080亿个晶体管和高达192GB的HBM3e内存，以及8TB/s的内存带宽，显示出其在硅片设计和内存技术方面的先进性。
- Blackwell架构支持新的精度格式和微张量缩放技术，使得AI模型可以在保持高精度的同时实现更高的性能。
- 与CUDA、TensorFlow、PyTorch等主流框架兼容，并通过与各类生态系统合作伙伴协作，增强了应用开发和部署的灵活性。

Blackwell 架构的作用

数据中心转型

Blackwell架构的出现，代表着数据中心转型的实现和工业智能化的进一步推进

强大的计算能力　高效的数据处理性能

转化

数据中心

到

生产和管理AI模型的核心设施

推动发展

基于Blackwell

升级优化AI芯片和服务

构建智能生产工厂

与世界各地的人工智能公司、OEM和国家主权AI合作

建立全新的智能工厂和AI基础设施

代表产品

- 基于Blackwell GPU的几种产品变体包括GB200、B200和B100，涵盖从超级芯片到离散加速器的多种配置，满足不同计算需求和系统兼容性。
- GB200集成了两个Blackwell GPU和一个72核的Grace CPU的超级芯片，在提供一体化解决方案方面实力突出。
- B200 GPU在AI计算领域突出，处理1.8万亿参数的模型，性能提升15倍。
- B100能应对1730亿参数的大语言模型。
- 2000张Blackwell GPU仅需1/4的电力即可完成与8000张Hopper GPU相同的训练任务。

	HGX B200	HGX B100
GPU	HGX B200 8-GPU	HGX B100 8-GPU
构成因素	8x NVIDIA B200 SXM	8x NVIDIA B100 SXM
HPC和AI计算 (FP64/TF32/FP16/FP8/FP4)*	320TF/18PF/36PF/72PF/144PF	240TF/14PF/28PF/56PF/112PF

未来产品规划

2024

启动Blackwell芯片的量产，推出基于x86架构的H200h、L40S、B100和B40芯片。
H200将提升内存至282GB，带宽增加3倍，搭载144GB的HBM3内存。

2025

推出Blackwell Ultra GPU
8颗HBM3e 12hi内存，
并推出降规版B200A
4颗HBM3e 12hi。

2026

推出Rubin GPU
采用8颗HBM4内存，
Vera CPU
作为Blackwell平台的继任者。

2027

推出Rubin Ultra GPU
12颗HBM4内存
Arm的Vera CPU和NVLink 6 Switch（3600GB/s）。
进一步巩固英伟达在高性能计算和AI市场的领先地位，为更复杂计算任务提供支持。

国内主要GPU厂商以7nm制程为主，多数不具备双精度FP64，较海外英伟达、AMD和英特尔有一定距离

地区	企业	代表产品	制程	单精度FP32	双精度FP64	TDP	特点
海外主要品牌	英伟达	H100 PCIe	4nm	48 TFLOPS	9.7 TFLOPS	350W	专为要求最快速计算速度和最高数据吞吐量的应用而设计
		A100 80GB PCIe	7nm	19.5 TFLOPS	24 TFLOPS	300W	具备 54 亿个晶体管和第三代 Tensor Core
	AMD	INSTINCT MI100	7nm	23.1 TFLOPS	11.5 TFLOPS	300W	基于CDNA架构的数据中心系列加速卡
		INSTINCT MI250	6nm	45.3 TFLOPS	45.3 TFLOPS	560W	高性能计算和人工智能加速卡，具备 13,312 个流处理器和 208 个计算单元
	英特尔	锐炫Arc A770	6nm	17.2 TFLOPS	不具备	225W	采用Xe HPG架构，支持DX12 Ultimate、硬件光线追踪、可变刷新率VRS、XeSS超级采样、PCIe 4.0
	高通	Adreno X1	4nm	4.6 TFLOPS	不具备	-	专为Windows on ARM系统设计的骁龙X系列SoC的第一代集成显卡
国内主要品牌	寒武纪	思元370 X4	7nm	24 TFLOPS	不具备	150W	集成了390亿个晶体管，提供云端和边缘智能处理器，满足不同层次的人工智能计算需求
	海光信息	深算一号	7nm	12.2 TFLOPS	10.1 TFLOPS	350W	国内唯一能支持FP64双精度浮点运算，相当于英伟达A100的70%
	摩尔线程	MTT S3000	-	15.2 TFLOPS	不具备	<35W	基于MUSA架构，并搭载了第二颗多功能GPU芯片“春晓”
	壁仞科技	壁砺100P	7nm	2456TFLOPS	不具备	450-550W	像素填充率和AI运算性能接近或部分达到国际市场标准
	天数智芯	天垓100	7nm	37 TFLOPS	不具备	250W	具备自主可控的高性能和通用性，支持AI训练和推理，广泛兼容主流软硬件生态

英伟达为GPU开发的CUDA统一计算架构，利用GPU的并行计算能力加速大规模计算任务，使GPU能够解决复杂的计算问题，促成了GPU+CPU协作的系统生态结构形成

CUDA开发背景

GPU能力开发需要	由于GPU与CPU技术差异的存在，CPU核心有限，难以高效处理大量相同任务，GPU则拥有上千核心，擅长并行处理，但这也使得GPU的通用计算任务开发和移植十分困难	英伟达开发出CUDA（统一计算架构），使GPU不仅能处理图形任务，还能进行通用计算	通过CUDA实现：充分利用GPU并行计算能力加速计算速度；协同CPU与GPU之间的任务分发和并行处理
扩展GPU应用场景	使GPU的并行计算能力能够应用于加速大规模计算任务，从而进入科学计算、机器学习、密码学等领域		
英特尔合作契机	英特尔寻找提高CPU性能、降低功耗和增强并行性的技术方案，为英伟达和英特尔的合作提供了契机	2006年，英伟达与英特尔达成合作，共同开发了基于CUDA的新型GPU（G80 GPU，GeForce 8800 GTX）	

CUDA工作原理

允许程序员编写直接在GPU上运行的代码，并增加了控制GPU并行处理的功能，开发者可以精确分配和控制数据在GPU核心上的处理，实现高效并行计算

CUDA架构	运行内容	开发功能
开发库	提供标准数学运算库，如CUFFT和CUBLAS，解决大规模并行计算问题	开发人员可以基于这些库快速构建自己的计算应用，并在CUDA技术基础上扩展更多的开发库
运行环境	提供应用开发接口和运行期组件，支持程序代码在CPU和GPU上运行	涵盖了数据类型定义、内存管理、设备访问和执行调度等功能，帮助开发人员实现各种计算需求
驱动支持	CUDA应用需要NVIDIA CUDA-enable硬件支持，驱动程序提供了不同版本GPU之间的设备抽象层接口	通过这一层，CUDA可以实现硬件设备的各种功能和计算任务的执行

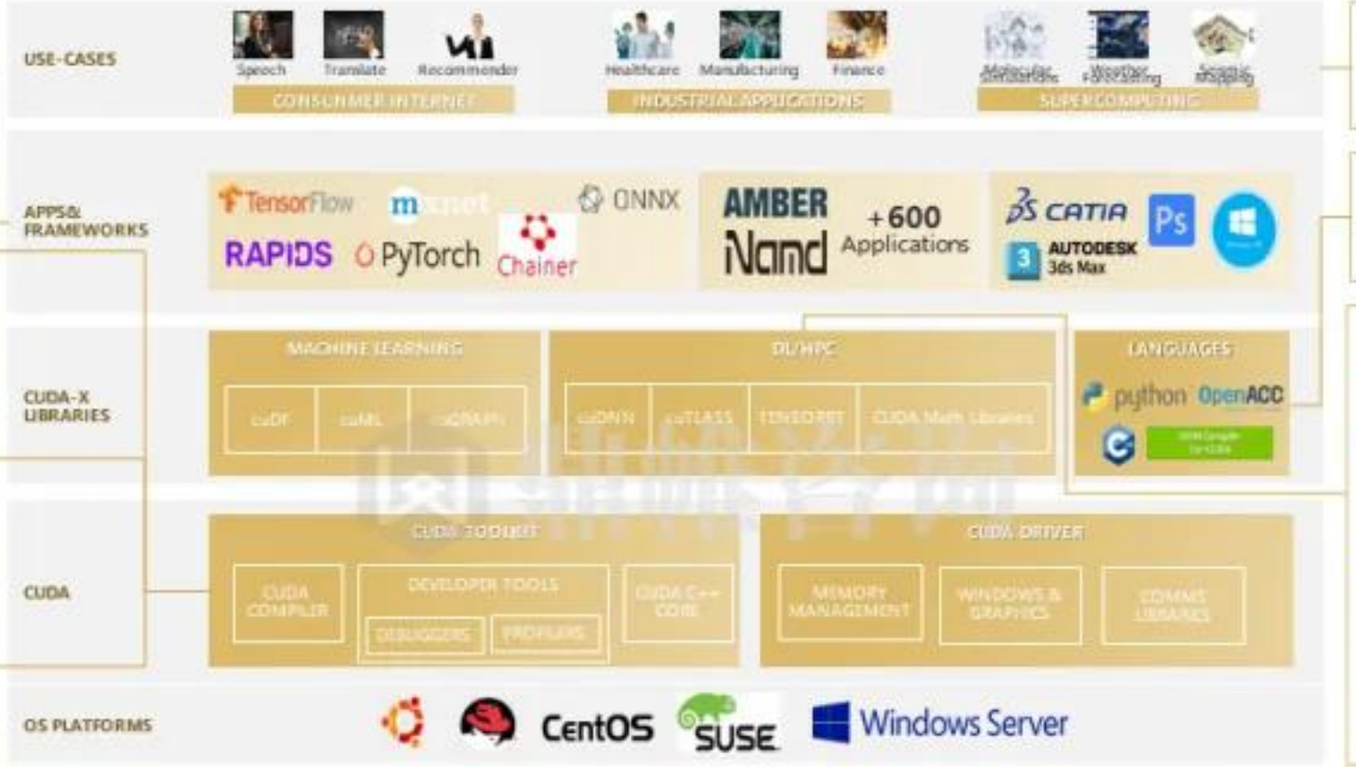
CUDA软件生态包含丰富的组件、编程语言(C、C++、Fortran、Python 和 MATLAB)、API、开发库和调试工具等，覆盖AI和HPC领域，构筑了软件覆盖率高、AI框架支持率高、细分行业渗透率高等竞争壁垒

工具和容器部署支持：支持容器和Kubernetes，提供工具如Nsight、CUDA GDB

1. NVIDIA driver
连接操作系统与显卡，支持与显卡通信，确保显卡正常工作并优化性能。

2. CUDA Toolkit
工具包，提供API和工具，帮助开发者利用GPU加速计算，支持代码编写。

3. CUDA API
编程接口集，允许高性能计算，设备和内存，执行核心功能。



CUDA应用案例：
科学计算、图像处理和深度学习，提升训练和推理速度

编程语言与API：
支持C/C++、Python等语言编写GPU程序，提供CUDA工具包和API接口

CUDA库

- **cuDNN**：深度神经网络库，提供深度学习算法的GPU加速实现
- **cuTLASS**：专为高效矩阵乘法和线性代数操作而设计，广泛应用于深度学习和高性能计算领域
- **TENSORRT**：高性能深度学习推理库，用于优化、加速和部署深度神经网络模型
- **Math Libraries**：专用高性能软件库，旨在高效处理科学计算、深度学习和 AI 应用程序中的复杂数学运算

CUDA竞争壁垒优势

软件覆盖率

- 基础的并行计算软件库
- 细分行业软件库
- 配套辅助软件等

AI框架支持率

- 针对 AI 领域而言，诸多AI框架的完善支持

细分行业渗透率

- 对于细分行业的支持需要较长时间积累

从2006年起，CUDA已完成12次迭代，持续提升整体性能，与英伟达AI生态深度绑定成为最重要的护城河

第一阶段：初期挑战

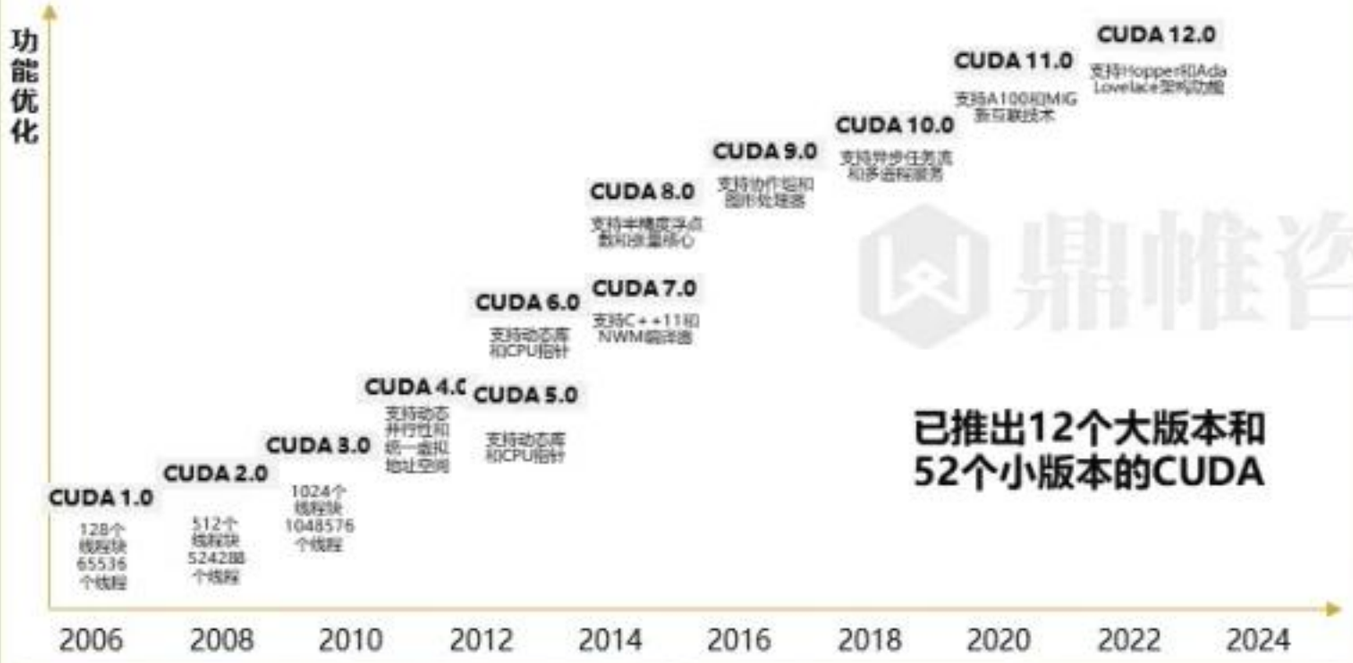
投入大量研发，市场反应冷淡，用户少，市值长期低迷

第二阶段：AI领域渗透

CUDA在科学计算领域持续渗透加速在AI领域的发展

第三阶段：持续迭代

经过多年迭代，推出12.0版本，与HPC与AI生态深度绑定，成为核心工具箱



已推出12个大版本和52个小版本的CUDA

功能增强

每个CUDA新版本都会引入新功能或API，如CUDA 10的任务图功能和CUDA 11的新编程模型和库功能

性能提升

CUDA升级时会持续优化性能，提高运行速度和效率，例如：CUDA 9引入的Cooperative Groups模型

兼容性调整

新版本会对旧版代码的兼容性进行调整，某些旧特性或API行为可能会被修改或不再支持

错误修复与稳定性改进

新版本会修复已知的错误，并提高稳定性和可靠性，使CUDA能更好地处理任务

系统	CUDA 8.0	CUDA 9.x	CUDA 10.x
费米 (Fermi)	GTX580	-	-
开普勒 (Kepler)	GTX680, GTX780Ti, GTX Titan, Titan Z, Tesla K80	GTX680, GTX780Ti, GTX Titan, Titan Z, Tesla K80	GTX680, GTX780Ti, GTX Titan, Titan Z, Tesla K80
麦克斯韦 (Maxwell)	GTX980Ti, Titan X, Tesla M40	GTX980Ti, Titan X, Tesla M40	GTX980Ti, Titan X, Tesla M40
帕斯卡 (Pascal)	GTX1080Ti, Titan Xp, Tesla P100	GTX1080Ti, Titan Xp, Tesla P100	GTX1080Ti, Titan Xp, Tesla P100
伏特 (Volta)	-	Titan V, Tesla V100	Titan V, Tesla V100
图灵 (Turing)	-	-	RTX2080Ti, Titan RTX, Tesla T4

优势特点

- 易部署（用户开箱即用）
- 层次灵活的开发接口
- 满足不同领域开发者编程语言
- 品类齐全的工具集
- 第三方工具和软件库

CUDA生态软硬件协同，赋能深度学习、信号与图像处理、线性代数及并行算法等多个应用场景

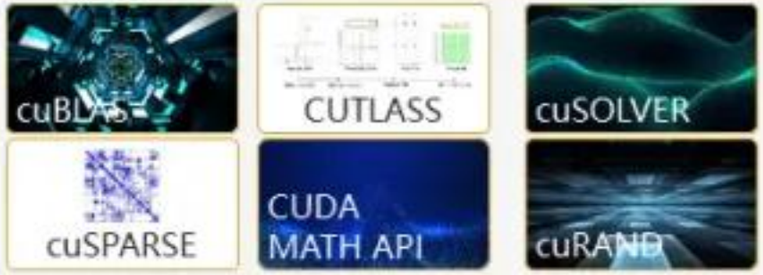
GPU加速库

为您的应用进程提供“嵌入式”加速

深度学习

信号与图像处理

线性代数



并行算法

图形处理

- 常用于视频游戏和电影特效中的大规模图形渲染。
- 例如，电影制作公司利用CUDA加速渲染过程，大幅减少渲染时间，提高效率。（如Blender和Maya）

人工智能

- CUDA加速了深度学习模型的训练和推理
- 处理大量矩阵运算的深度学习框架（如TensorFlow和PyTorch）时，使研究人员能加速实验

科学计算

- CUDA加速了数值模拟和大规模数据处理
- 如物理学中的分子动力学模拟和气象学中的气候预测。通过并行处理，CUDA大幅提高计算速度。

- **GPU和CUDA成为行业标配，CUDA成为深度学习领域最发达、最广泛的生态系统，是推动GPU计算普及的关键力量。**
- **CUDA生态中软硬件协同的设计使英伟达以最小代价保持性能领先，巩固AI界领先地位**

CUDA与主流OpenCL、CTM、Direct Compute对比生态更完善，更好赋能GPU，而中国国内软件生态暂未形成同一架构，缺乏相应核心生态竞争力

对比对象	CUDA 优势	CUDA 劣势
OpenCL	■ 性能 - 与NVIDIA硬件紧密结合，充分发挥GPU性能 ■ 开发体验 - 编程模型简洁易用，提供丰富的开发工具链。 ■ 社区支持 - 社区资源丰富，代码库广泛应用于科学计算与深度学习	■ 兼容性 - 兼容性不如OpenCL广泛，OpenCL适用于更多硬件平台。
ADM CTM	■ 工具链 - 提供完整的开发工具链，包括编译器、调试器和性能分析工具 ■ 应用范围 - 广泛应用于科学计算和深度学习，优化库丰富	■ 硬件控制 - CTM接近硬件，提供更高的硬件控制权。 ■ 应用场景 - 在特定应用场景中有独特优势。
Direct Compute	■ 功能支持 - 功能库齐全，开发工具完善，广泛支持科学计算和深度学习领域 ■ 性能优化 - 针对NVIDIA GPU，性能优化更出色	■ 通用性 - DirectCompute在通用性和多平台支持上更强，适用于微软操作系统及多种硬件。

国产AI芯片软件生态因投入不足和碎片化严重不具备核心竞争力，劣势体现为：

① 指令集不统一，硬件架构分散
② 软件栈不同意。用户学习成本高
③ 算子覆盖度低，用户迁移成本高
④ 企业各自为战，没有足够的生态竞争力

CUDA同时面向个人用户和企业用户市场，以低学习门槛、使用完全免费及开发者生态建设占领市场份额

产品：低学习门槛

- CUDA的推出极大地降低了开发者使用GPU进行通用计算的门槛，使更多的研究人员和工程师能够利用GPU进行AI模型的训练和推理，使其在新一代工程师中迅速普及，为未来进入企业打下基础。

价格：免费使用占领市场

- CUDA采取完全免费的使用策略，不求短期回报，迅速占领市场，并通过持续研发投入和优化，保持CUDA技术的领先地位

服务：开发者生态建设

构建软件生态系统

强化与开发者的联系
提升推广范围

内容服务

- 通过GTC开发者大会、博客、论坛、培训等形式，发布与CUDA相关的新闻、课程、经验方法等，增进用户交流。

工具服务

- 以开源形式打造强大的开发库和工具集，如开源Cub、NCCL、Nvcaffe等，提供基础工具，增强开发者的社区参与度、用户粘性与忠诚度

渠道1：面向个人用户的教育合作

- **机构课程合作**：与全球顶级研究机构和大学合作，在全球350所大学开设相关课程
- **认证计划推广**：在教育领域以认证方式推广CUDA，培养大量熟悉CUDA的工程师，也确保学生进入企业时不再陌生
- **赞助学术研究**：推动CUDA在学术界的应用与普及

渠道2：面向企业用户的赋能支持

- **通过创业支持开发创业公司客户**：投资Inception计划，支持创业公司采用CUDA进行项目开发，吸引企业公司大量采用CUDA，同时搭配GeForce覆盖主流市场
- **通过垂直领域应用场景工具包进入行业顶尖客户**：如推出面向GPU加速数据科学的CUDA-X AI 加速库，为数据科学提供加速平台，得到渣打银行、微软、PayPal、SAS和沃尔玛等顶尖公司所采用，并集成至主流深度学习框架（如TensorFlow、PyTorch、MXNet）和全球主要云服务提供商，实现与顶尖硬件厂商和云计算提供商合作，巩固市场地位

CUDA借助其开发环境成熟度和GPU硬件性能优势，吸引大量平台及生态合作伙伴，形成独特生态圈

CUDA生态圈

成熟度
稳定性

- 开发者借助已有的资源和文档进行开发和部署，为英伟达GPU的开发、优化和部署多种行业应用提供先发竞争优势
- 截至2020年，CUDA开发者数量达到了200万，并于2023年增长到400万，其中包括Adobe等大型企业客户

低门槛
GPU过硬

- 引入了大批开发者建设CUDA生态社区，最终绑定了数百万AI开发者
- 当CUDA几乎与AI画等号的时候，会有大量的社区力量为其助力

CUDA成为英伟达生态基础

应用层			
HPC深度学习框架 公司数据分析可视化			
垂直行业解决方案			
英伟达AI	自动驾驶	智慧城市	健康养护
软件、硬件、库、系统			
深度学习	HPC	数据中心	
TESLA	DGX		
CUDA&GPU计算机构			

英伟达云GPU

CUDA平台部分合作伙伴

American Express, AMP Robotics, Baidu 百度, 字节跳动, CRITEO, Foxconn, GE Healthcare, IBM, JD.COM, John Deere, KLA, NAVER, Nuance, Oracle, Samsung Medison, SAP, serve, Siemens Healthineers, SK telecom, Snapchat, Tencent 腾讯, verizon, WeChat, ZF, ZOOX

CUDA生态部分合作伙伴

Adobe, Ansys, Autodesk, Dassault Systèmes, MathWorks, Microsoft, National Instruments, Wolfram Mathematica

2.3DPU：NVIDIA是全球DPU领域的先行者之一，抢先进入DPU市场，并将其作为数据中心第三颗主力芯片，赋能安全、计算、存储和网络业务

市场机会

CPU计算需求压力大

全球算力需求平均每3.5个月翻一倍，传统“CPU+xPU”多元化异构计算架构在性能上的提升无法满足当前的网络带宽的发展，CPU的计算需求压力不断增大

领先进入

NVIDIA是DPU产业发展的全球先行者

2020 年上半年，NVIDIA 以69 亿美元的对价收购以色列网络芯片公司 Mellanox Technologies，同年推出 BlueField-2 DPU

将 DPU 定义为继 CPU 和 GPU 之后“第三颗主力芯片”，正式拉开 DPU 大发展的序幕

DPU 介绍

概念

- DPU（数据处理器，Data Processing Unit）作为数据中心第三颗主力芯片，与CPU和GPU互补，共同构成高效的算力平台

功能

- 高性能的网络处理能力
- 支持高速网络连接
- 融入通用计算能力
- 可进行安全与存储卸载功能

技术

- 专用的硬件加速器处理各种任务
- ARM或x86架构的CPU核心
- ASIC、FPGA等专用硬件加速引擎

价值与未来

- 将数据中心基础设施操作从CPU卸载到DPU上，提升数据中心能效和性能
- 将成为未来计算的第三大支柱，构建高效计算生态

架构

网络(Network)
据与数据包处理相关的功能，包括协议处理，存储转发，连接管理等

存储(Storage)
与数据存储相关的功能，包括存储读写管理，持久化，去冗，纠错等

计算（Computing)
直接执行应用层算法，或者为应用提供计算平台或统一的计算资源相关的功能

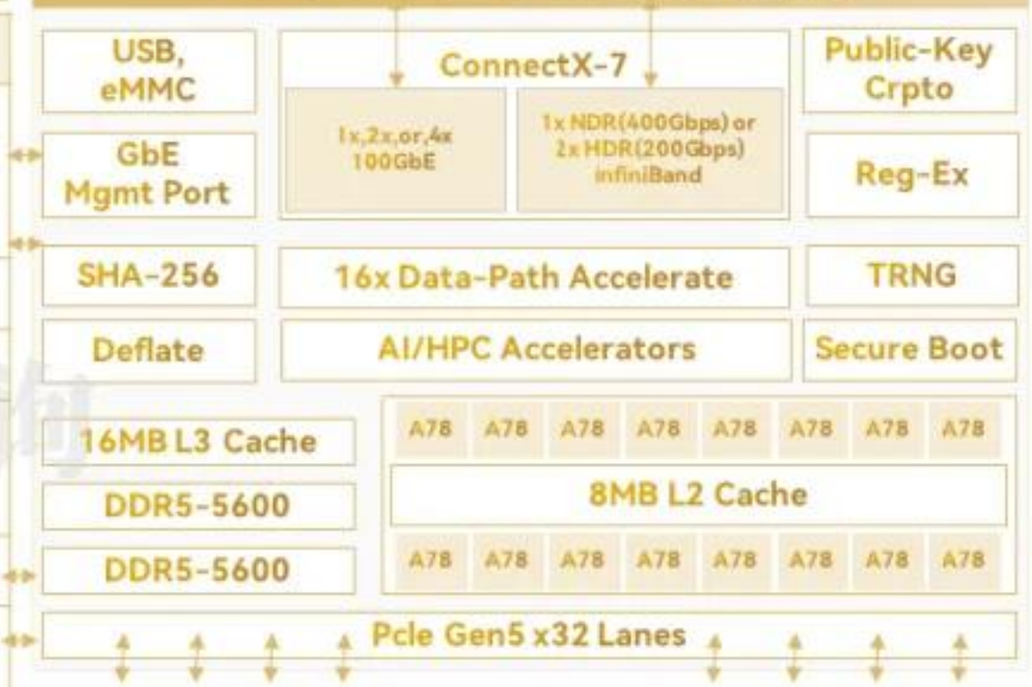
安全（Security)
为不同逻辑层次的安全属性理供直接支撑的功能

英伟达布局DPU产品，每18个月推出新一代Bluefield芯片，目前第四代算力已达1000TOPS（每秒万亿次操作）

推新战略：每 18 个月推出一个新的 DPU

	BlueField-2	BlueField-3	BlueField-4
概述	世界上第一个针对现代企业数据中心优化的芯片架构的数据中心基础设施	是第三代 NVIDIA DPU	预期将GPU进行集成到DPU中，为边缘设备提供低成本、高性能的安全数据处理能力
发布时间	2020年10月	2021年4月	/
晶体管数量	70亿个	220亿个	640亿个
网络端口	10/25/50/100G 双端口 200G单端口	25/50/100G 四端口 400G 以太网 NDR InfiniBand	800G
PCIe端口	8或16道PCIe 4.0	32道 PCIe5.0	
核心处理器	ARM+ASIC+专用加速器	ARM+ASIC+专用加速器	ARM+ASIC+GPU专用加速器
算力	0.7 TOPS	1.5 TOPS	1000 TOPS
SPECint 2017	9	42	160
与上代相比	/	2 倍的网络带宽 4 倍的计算能力 几乎 5 倍的内存带宽 高达8倍的速度 运行工作负载	整数处理性能 提高近4倍

BlueField-3 DPU总体架构



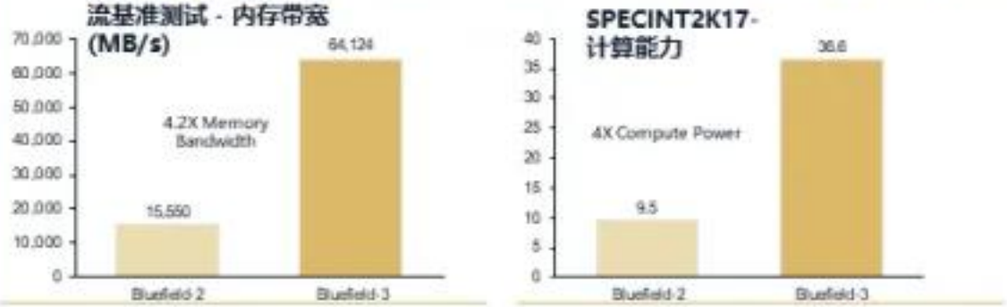


图1. 与NVIDIA BlueField-2 DPU 相比，NVIDIA BlueField-3 DPU 的内存带宽和计算能力提高了四倍

海内外主要AI厂商均通过自研和并购布局DPU，海外聚焦DPU SoC芯片，国内聚焦CPU+FPGA的DPU解决方案，即国外主要做芯片，国内受架构限制主要做基于芯片基础的解决方案

国外布局动作（DPU SoC）

1. AWS收购Annapurna公司开发DPU SoC芯片
2. Intel与Google合作开发Mount Evans系列的DPU SoC
3. AMD于2022年收购了DPU SoC厂商Pensando
4. Nvidia收购Mellanox，推出BlueField系列的DPU SoC

国内研发受阻（架构原因）

基于CPU+FPGA的DPU解决方案

优势

1. 有助于加速和卸载各种基础结构组件
2. 更短的开发时间和快速的迭代，促进功能快速定制

劣势

1. 由于芯片工艺和 FPGA 结构而受到限制
2. 有效控制芯片面积和功耗具挑战性，阻碍发展

公司名	代表产品	核心处理器	技术路线	应用方向	发布时间
NVIDIA	BlueField-2	ARM+ASIC+专用加连器	ARM	数据安全、网络安全、存储卸载等	2020
	BlueField-3	ARM+ASIC+专用加连器			2021
	BlueFieid-4	ARM+ASIC+GPU 专用加速器			2023
Intel	FPGA IPU C5020X	FPGA+X86 SoC	FPGA	面向交换机、路由器芯片	2020
AMD/Xilinx	Alveo U25	FPGA	FPGA	面向网络、存储和计算加速功能	2020
Marvell	OCTEON 10	SOC: ARM+ASIC	ARM	面向集成机器学习推理引擎和内联加密处理器等	2021
Broadcom	Stingray	SOC: ARM+ASIC	ARM	面向交换机、路由器芯片	2018
Pensando	Capri	NP+SoC	软件定义网络处理器	面向P4的SDN	-
Fungible	F1	NP+SoC	MIPS	面向网络、存储、虚拟化	2020
Amazon	Nitro	-	-	为智能网卡数据提供线逸加密和解密	-
Microsoft	Catapultv3	GP+SoC	-	面向深层神经网络加速	2017
阿里巴巴	X-Dragon CIPU	FPGA+ASIC	-	面向虚拟机管理程序	2022
华为	IN300	NP-SOC	Hi1822 /ASIC	面向连接FC 网络的应用，高带宽高性能存储组网方案	2018

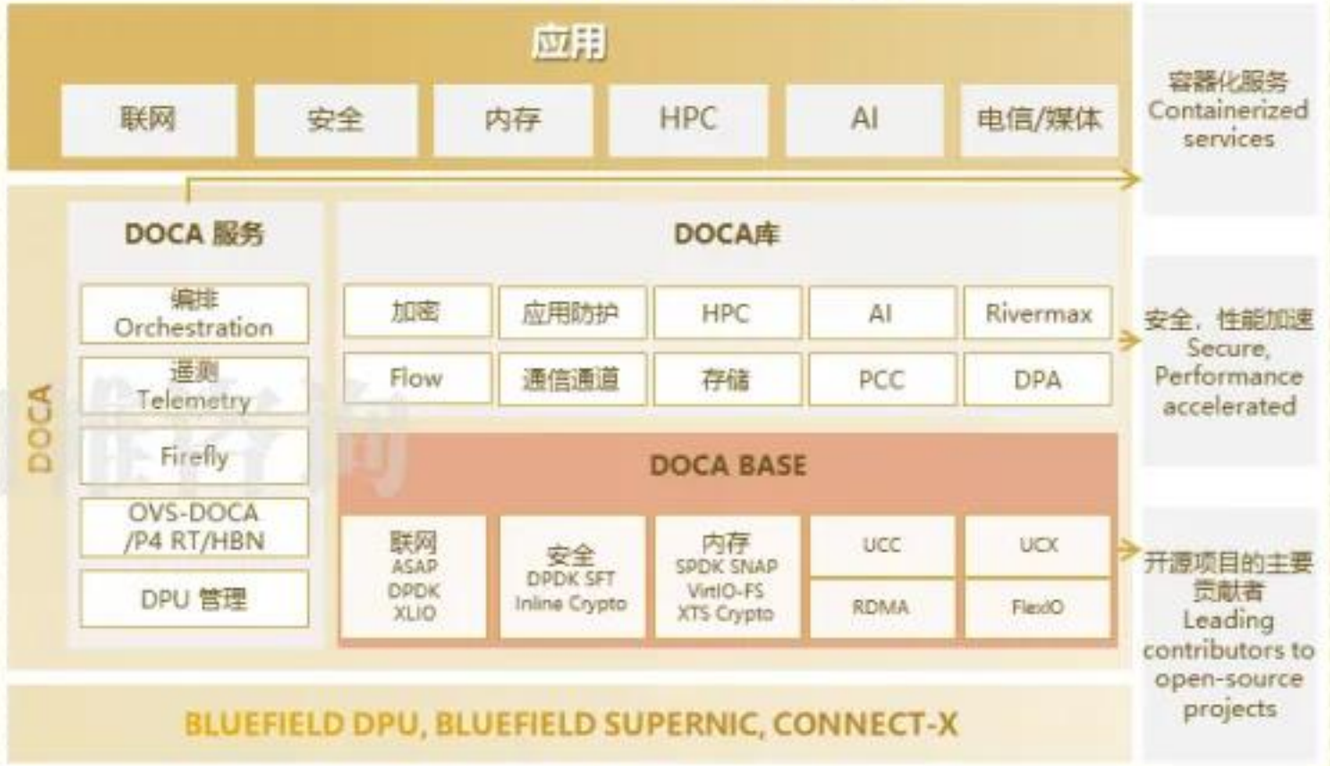
DOCA生态建设：作为DPU量身定做的软件开发平台，围绕其搭建了DOCA的软硬件生态、开发者生态、合作伙伴生态

DOCA软硬件生态建设

NVIDIA DOCA是一个为DPU量身定做的软件开发平台

内容与功能	• 拥有丰富的库、驱动程序和API，全面、开放，让开发者可以快速创建经DPU加速的高性能应用程序和服务 • 也是加速云基础设施服务的关键
应用	• 与VMware、Palo Alt、Juniper Networks共同整合并扩展基于DPU及DOCA软件架构在基础设施、存储、网络安全、5G和边缘计算等应用场景的解决方案
价值	• DOCA是支撑基于DPU的应用程序与服务的核心与灵魂，是释放 DPU 潜力的关键 • DOCA 是英伟达人工智能云服务战略的关键组件，旨在为加速数据中心工作负载和大规模部署人工智能应用程序提供一个灵活而强大的平台

NVIDIA DOCA开发平台架构



DOCA合作伙伴生态建设

- NVIDIA 与领先的平台供应商和合作伙伴合作，整合和扩展 NVIDIA DOCA 在平台基础设施、存储、网络安全和边缘计算应用场景的生态体系，包括 VMware、Red Hat、Palo Alto、ARIA Cybersecurity、极客天成、炎融科技、爱瑞无线等。

DOCA开发者生态建设

- NVIDIA DOCA 中国开发者社区的注册开发者已超过 4,000 人，围绕云计算、高性能计算与人工智能等典型应用场景进行软件定义、硬件加速的数据中心基础设施应用程序和服务开发，并通过积极推出丰富的活动以助力开发者更好地学习。

英伟达目前拓展Oracle、Cisco 、 DDN 、 Dell EMC 、Juniper 、VMWare 等二十多个生态系统合作伙伴使用 BlueField 数据中心加速技术来更高效地运行其软件平台

NVIDIA与其主要合作伙伴的合作内容

ORACLE

- 利用 BlueField -3增强其基础设施，为客户提供无缝的云体验

vmware

- 共同提出 AI-Ready Enterprise Platform，由 NVIDIA 优化、认证和支持，帮助数千位 VMware 客户使用 AI 的强大功能

paloalto NETWORKS

- 共同部署了包括虚拟防火墙在内的 5G 原生安全倡议，致力于满足 5G 云原生环境严格安全需求，为客户提供安全保护

Red Hat

- 在其开放混合云产品组合 RHEL 和 OpenShift 中为 DPU 提供支持

Canonical

- 在 Ubuntu 云平台中支持 BlueField - 2DPU 和 DOCA

CHECK POINT

- 将 BlueField-2 DPU 集成到产品技术中，加速网络安全产品

四大业务

-游戏业务

-专业可视化业务

-数据中心业务

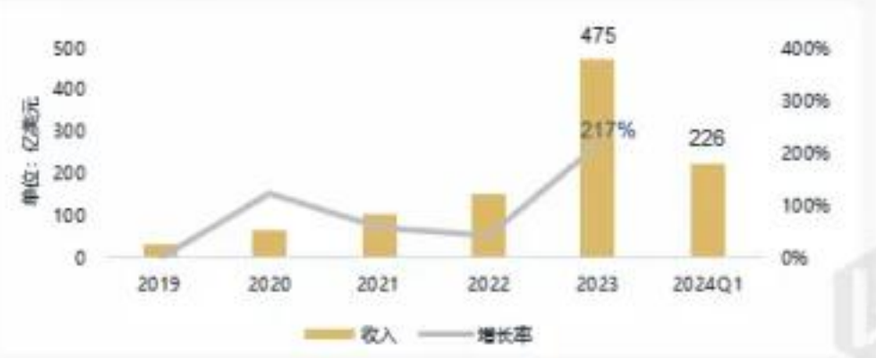
-汽车业务

-应用场景

行业客户层面，英伟达公司布局了游戏、专业可视化、数据中心、汽车市场四大领域，数据中心在2022年超越游戏成为了英伟达的第一大业务

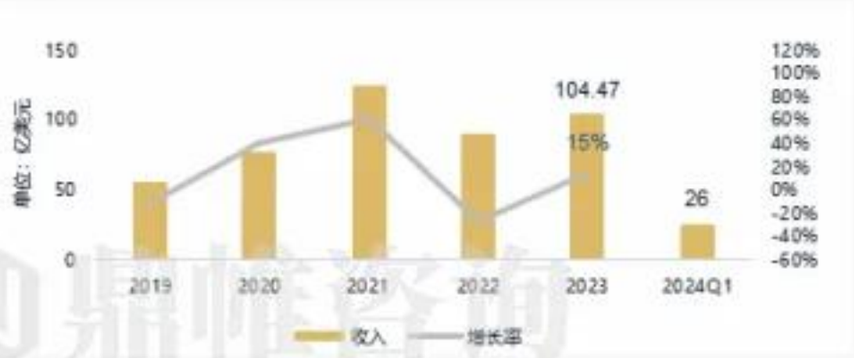
① 数据中心业务

公司已完成 CPU+GPU+DPU 三芯的硬件布局，通过底层硬件架构和 CUDA 生态整合，构建全领域加速计算平台。近几年公司业绩增长主要由数据中心贡献，公司也将加快技术迭代速度，重塑 AI时代的数据中心。



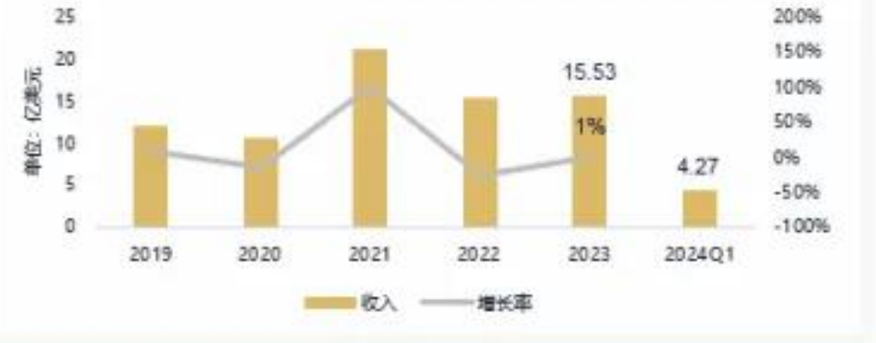
② 游戏业务

公司提供用于 PC的 GeForce RTX和 GeForce GTX 显卡、于其他游戏机的 GeForce NOW 云游戏、用于在电视播放高质量流媒体的 SHIEL并与游戏方合作提供开发服务。



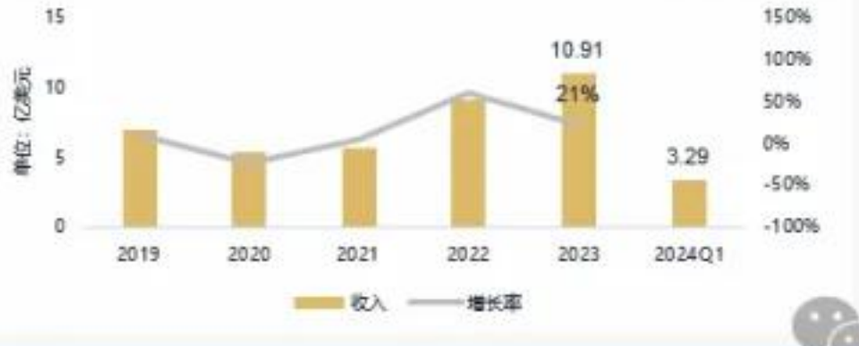
③ 专业可视化业务

NVIDIA RTX GPU 和 EGX 平台为客户提供涵盖专业图形渲染、云端 XR 应用、AI数据科学与大数据研究的专业可视化业务，可应用在汽车、建筑、医疗、影视媒体等多场景。



④ 汽车业务

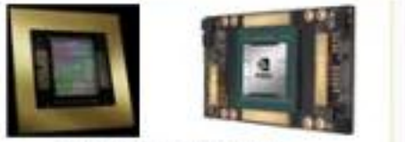
DRIVE Orin SoC 芯片能够为自动驾驶功能、置信度视图、数字仪表盘以及 AI座舱提供强力支持，Hyperion 架构将 AI计算与完整的传感器套件集成整合，能加速自动驾驶的开发、测试和验证过程。



2020-2024Q1 英伟达各业务板块收入占比

英伟达基于GPU等芯片架构上的底层技术积累，围绕各个业务场景设计了系列产品线

数据中心



AI训练加速器：
Blackwell系列
A100 Tensor Core GPU

Grace CPU

DGX
SuperPOD

直连式铜缆

AI推理加速器：
Tesla T4
Jetson Xavier NX

HGX AI
超级计算机

网络：
Bluefield
DPU

光收发器

高性能计算加速器：
Tesla系列

主机网卡

Spectrum-4
以太网
交换机

游戏

游戏GPU：
GeForce系列

游戏笔记本GPU：
RTX 4090 D显卡

RTX系列
笔记本电脑

专业可视化

工作站GPU：
Quadro系列
RTX系列

可视化集群：
Quadro
可视化工作站

EGX专业
可视化平台

汽车

汽车SoC：
DRIVE Thor芯片

汽车SoC：
DRIVE AGX系列

自动驾驶DRIVE
Hperion 架构

英伟达的业务已经触及数以亿计的To C和数以百计的To B客户，庞大且多元的客户群，使得其过去2个财年没有一家大客户超过其总收入的10%

游戏

数据中心

Pro可视化

汽车

一、游戏业务。游戏市场是英伟达的基本盘，以独立显卡为发家产品，同时提供云游戏服务，在游戏芯片领域英伟达是当之无愧的霸主

业务构成

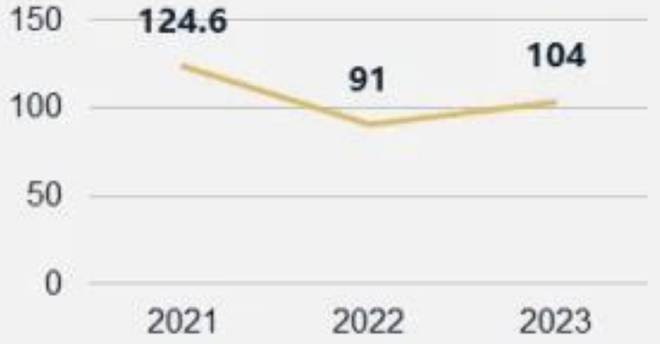
个人电脑	GeForce RTX GPU/GeForce GTX GPU
游戏主机	SoC芯片
云游戏服务	GeForce Now

游戏本GPU

- GTX M：GTX770M与桌面端GTX770性能悬殊
- GTX 900M：与桌面端同型号规格差距缩小，开始采用相同内核设计
- GTX 10："游戏本元年"，除主频降低外，与桌面端同型号规格基本一致
- Max-Q：性能落后于桌面端及非Max-Q版本约15%-20%，使游戏本首次实现轻薄和性能兼备

营收表现

2023年第四季度实现营收28.65亿美元，同比增长56.5%。在经历几个季度的低迷后，游戏业务有所回暖



主要GPU迭代历程

尽管游戏业务已不是英伟达营收的最大来源，但游戏凭借对硬件技术迭代的持续推动，以及作为AI技术的高频应用场景，仍然是英伟达的核心业务，并围绕它确立了“云游戏+AI”发展思路

AI领域的训练场：契合全球科技和业务发展方向

游戏玩家需求推动硬件技术打磨

游戏3A大作是前沿信息技术的集成体现
- 在应用端能带动大量软硬件消费、产值和就业机会
- 科技公司迭代技术、发展自身产业的“顺风车”

游戏技术对前沿科技贡献

前沿科技	贡献率
芯片	14.9%
5G高速网络	46.3%
混合现实产业	71.6%

生成式AI赋能游戏

开发端	美术端	体验端
协助开发人员快速完成基础代码，省去一些耗时、重复的工作，提高开发效率	实现快速出图、建模，节省大量时间，降低游戏开发成本	生成式AI推出的智能NPC能够进行实时对话、自然语言交互，交互体验得到大幅提升

英伟达发展主线：云游戏业务+AI

云游戏业务：GeForce NOW

背景	目前规模	订阅方案
2022年全球云游戏市场规模尚不到整个游戏产业2%，英伟达71%存量用户仍未升级到支持光追的RTX显卡	GeForce Now在全球已拥有超千万的订阅用户，营收则占据“云游”市场的半壁江山	• 免费版 • 49.99美元/半年 • 99美元/年

聚焦客户	市场竞争
• 享受免费PC游戏的玩家（例如《堡垒之夜》） • 或者是已经在主流的PC游戏商店拥有游戏的玩家	• 除去英特尔的集成显卡，目前英伟达和AMD几乎占据了GPU市场100%的份额 • 但目前的主流游戏机以及次时代的主机均采用AMD的解决方案，故英伟达将**云游戏主要精力放到了的PC上**，避免与AMD直接造成冲突

游戏定制AI方案：AI模型代工服务

NeMo	用于构建、定制和部署语言模型
Riva	用于自动语音识别和文本转语音
Omniverse Audio 2 Face	根据语音轨迹，快速生成游戏角色富有表情的面部动画

英伟达通过NVIDIA RTX（光线追踪技术）实现实时电影级渲染，NVIDIA ACE（虚拟数字人类生成技术）将NPC转变为动态交互式角色，极大地提高游戏开发效率和玩家游戏体验

NVIDIA RTX光线追踪技术

发展历程

- 2018年推出，提供广泛应用加速，融入各领域发展，如游戏、内容创作、影音、生产力、开发等

技术水平

- 算力性能超越NPU助力AI PC

核心优势

- 光线追踪技术，可实现实时电影级渲染。长期用于电影行业的特效，是一种计算密集型技术，可模拟光线的物理行为，从而在计算机生成的场景中实现更逼真的效果。

	适用范围	算力水平
NPU	持续性的AI低负载	现阶段10-45 TOPS
RTX	承担任何AI负载	最高可超过1300 TOPS

NVIDIA ACE 虚拟数字人类生成技术

一套可帮助开发者利用生成式 AI 创建栩栩如生的虚拟数字人物的技术。在 ACE 的加持下，普通的非玩家角色 (NPC) 可以摇身一变，成为能够发起对话或引导玩家找到新任务的动态交互式角色。

以微服务架构推出，加速数字人类开发

生成式AI技术套件

- NVIDIA Riva ASR、TTS及NMT（用于自动语音识别、文字转语音转换和翻译）等语音识别转换技术
- NVIDIA Nemotron大型语言模型（用于语言理解和语境脉络回应生成）
- NVIDIA Audio2Face（根据语音音轨制作出面部动画）
- NVIDIA Omniverse RTX（用即时光线追踪技术制作逼真的皮肤和毛发）
- NVIDIA Nemotron-3 4.5B：全新的小型语言模型（SLM），专为低延迟、终端设备的RTX AI PC推理而设计

目前应用

已在RTX AI PC上提供了抢先体验套件，曾CES大会展示与Inworld AI合作开发的Covert Protocol技术内容

极大优化游戏中的NPC角色

在游戏业务领域，英伟达通过与联发科合作积极布局AI PC业务进入高端笔记本电脑市场，预计在25年上半年发布AI PC芯片

发展历程及未来产品规划

提早布局：Tensor Core
于2018年推出RTX技术和首款专为AI打造的消费级GPU芯片GeForce RTX)
AIPC即为搭载专用AI加速硬件Tensor Core的计算机

英特尔首次提出AIPC概念
2023年9月，英特尔CEO帕特·基辛格在硅谷首次提出“AI PC”，伴随着酷睿Ultra系列处理器的推出实现落地

未来规划：与联发科合作
将采用台积电的3nm工艺制造和CoWoS封装，**最终目标是进入高端笔记本电脑市场。**预计25年上半年发布AI PC 芯片

2018 — 2023 — 2025

市场潜力巨大

中国AIPC市场将于2024年开始进入爆发期

预计台数	2028年将达到3300万台
市场占比	整体出货量的73%

AIPC推理优势

- 无需联网
- 低门槛
- 优惠价格
- 更适用于广大消费者市场

两代AI PC对比

对比项目	基础AIPC		高级AIPC
AI加速器	NPU		GPU
高峰AI TOPS	10-45		100-1300+
安装基数	~0m	VS	100m
创造性：视频生成/3D降噪	×		√
生产力：AI视频会议	基础		更高质量
游戏：帧生成/光线重构/AI NPCs	×		√

二、专业可视化业务。英伟达的专业可视化业务主要聚焦于为设计和可视化专业人士提供先进的技术解决方案，凭借着RTX技术和3D设计平台的应用，被广泛应用于建筑和工业设计、高级特效、科学可视化等领域

公司专业可视化业务收入情况(亿美元、%)

50 +应用程序解锁新市场

4500 万设计师和创意人员

21.1/8%

营收情况

22年该板块营收为21.1亿美元，
占英伟达22财年营收的8%，
5年复合年增长率为20%。

图形90%

市场份额

得益于RTX应用、混合工作环境和协作3D设计、AR/VR、人工智能和虚拟世界工作负载、元宇宙企业软件，英伟达在工作站图形市场份额已经超过90%

优化+开拓

业务方向

- **与独立软件供应商合作**，协助优化针对 NVIDIA GPU 的产品；
- **通过GPU计算解决方案提高生产力**，为设计、制造及数字内容创建等领域的关键工作引入新功能

英伟达的Quadro和RTX系列GPU是面向专业级用户和企业级市场的重要产品线，最新发布的RTX 2000 Ada架构 GPU为 AI 加速设计和可视化带来卓越性能

英伟达产品发布时间线

英伟达的QuadroGPU和RTX 显卡系列是产品线的重要组成部分

	Quadro系列	RTX系列
应用场景	图形设计、3D建模、视频编辑等专业领域。 - 如：电影制作公司皮克斯	游戏、内容创作、AI和数据科学、虚拟现实（VR）、工作站
技术规格	拥有更高的显存容量和带宽，以及更多的计算单元，还支持NVIDIA的CUDA和cuDNN加速库	NVIDIA RTX系列显卡采用Ada Lovelace架构，支持第三代光线追踪核心和第四代Tensor Core，具备高显存和先进的AI加速技术，提供卓越的图形和计算性能
市场定位	价格相对较高，定位于专业用户和企业级市场	RTX显卡面向游戏玩家、内容创作者和专业用户，覆盖从主流消费级到高端发烧级市场，为各种图形密集型和计算密集型应用提供解决方案。

英伟达专业可视化产品及其参数

系列名称	产品型号	发布时间	CUDA核心数量	加速频率(GHz)	基础频率(GHz)	显存容量	显存位宽(位)	制程工艺	接口	功耗
Quadro RTX	8000	2018	4608	1.77	1.35	48GB GDDR6	384	12nm	PCIe 3.0x16	295
	6000	2018	4608	1.77	1.35	24GB GDDR6	384	12nm	PCIe 3.0x16	295
	000	2018	3072	1.77	1.65	16GB GDDR6	256	12nm	PCIe 3.0 x16	200
	4000	2018	2304	1.62	1.1	8GB GDDR6	256	12nm	PCIe 3.0x16	160
Quadro P Series	P6000	2016	3840	1.8	1.48	24GB GDDR5X	384	16nm	PCIe 3.0 x16	250
	P5000	2016	2560	1.8	1.6	16GB GDDR5X	256	16nm	PCIe 3.0x16	180
	P4000	2017	1792	1.77	1.47	8GB GDDR5	256	16nm	PCIe 3.0x16	105
	P2000	2017	1024	1.48	1.17	5GB GDDR5	160	16nm	PCIe 3.0x16	75
	P1000	2017	640	1.85	14	4GB GDDR5	128	16nm	PCIe 3.0x16	47
	P620	2018	512	1.8	1.14	2GB GDDR5	128	14nm	PCIe 3.0×16	40
	P400	2017	256	0.98	0.64	2GB GDDR5	64	16nm	PCIe 3.0×16	30
RTX A6000	RTX A6000	2020	10752	1.8	1.41	48GB GDDR6	384	8nm	PCIe 4.0x16	300
RTX A40	RTX A40	2020	10752	1.74	1.3	48GB GDDR6	384	8nm	PCIe 4.0x16	300
RTX 6000	RTX 6000	2022	18176	1.77	1.44	48GB GDDR6	384	7nm	PCIe 3.0x16	300

NVIDIA Omniverse™ 是英伟达基于 USD 构建的实时协作平台，利用NVIDIA光线追踪、人工智能和模拟等先进技术，为用户提供了一个高度逼真和物理上准确的虚拟世界构建和交互的环境

NVIDIA Omniverse™ 介绍

组成	定位	功能	目的
应用 拓展套件 平台	强大、易于拓展的虚拟世界模拟 + 协作的高级仿真能力计算平台	基于NVIDIA RTX GPU和通用场景描述的实时图形和仿真模拟产品	创造者 设计师 研究人员 ↓ 在虚拟空间中实现协作和创新
应用	建筑	设计	娱乐	……

面向客群

创作者

- 以**实时同步**方式创作
- 以前所未有的**速度**打造自订 3D 流程，并模拟大规模虚拟世界
- **多软件互通**可将创意应用程序同步到 Omniverse 和USD，并在统一视图中使用3D 数据

开发者

- **门槛**：通过人工智能增强，几乎不用代码进行开发
- **自定义**：构建自定义扩展、工具和微服务，以加速 3D 工作流、生成合成数据和构建工业元宇宙应用程序。

OMNIVERSER 应用

VIEW | CREATE | AUDIO2FACE | ISAAC SIM | ...

拓展套件

OMNIVERSER 平台





CONNECT	NUCLEUS	KIT	SIMULATION	RTX RENDERER
• 连接SDK • 插件	• 核心服务 • 云 • 本地部署	• 观众 • 编辑 • 框架	• 物理系统 • AI • 动画片 • 行为	• 即时的 • 可拓展 • 准确 • MDL

NVIDIA Omniverse™平台由五大核心组件和两大扩展组件组成，并与一系列第三方数字内容创作工具和其他微服务共同组成Omniverse 生态系统，应用于游戏开发、电影制作、建筑设计等众多领域

组件	组件名	描述
核心组件		
	Omniverse Nucleus 中央数据库	数据库服务器（Omniverse的核心），用于存储、管理和共享3D世界和场景，允许多个用户和应用程序实时访问和协作处理相同的资产和场景。
	Omniverse Connectors	这是连接Omniverse和其他第三方3D创建工具的桥梁，确保数据和场景可以在不同应用程序之间无缝传输
	Omniverse Kit 平台开发工具包	一个开源平台开发工具包，允许开发者自定义和扩展Omniverse功能，包括用于构建应用程序的API和工具。
	Omniverse RTX Renderer	基于NVIDIA RTX技术的实时渲染器，提供逼真的视觉效果，支持多种光线追踪技术，可在不牺牲性能的情况下渲染高质量的图像。
	Simulation	提供仿真服务，通过Omniverse Kit的小插件和微服务使虚拟世界看起来更真实
扩展组件		
	Omniverse Launcher	用于启动和管理Omniverse应用程序和场景
	Omniverse Extentions	提供对各种应用程序和工具的集成支持

英伟达面向不同人群和场景推出了6大版本的Omniverse平台，开放五大云接口，实现Omniverse 的核心技术的直接集成，企业和开发者能够利用 Omniverse 平台的强大功能，加速创新并提升产品的市场竞争力

六种Omniverse版本

版本	用户群体	功能
Omniverse Studio	专业3D艺术家和设计师	提供强大3D创作工具和实时协作功能
Omniverse Kit	开发者	开源平台开发工具包，允许自定义和扩展Omniverse功能
Omniverse Create	/	提供直观界面，用于构建、布置和编辑场景
Omniverse View	非技术用户	用于审阅和批准3D场景的应用程序
Omniverse Enterprise	企业客户	提供如安全性、可扩展性、云支持和高级服务的企业级功能
Omniverse RTX	/	专注于实时渲染，提供逼真的视觉效果

五种Omniverse云接口

接口名称	作用
USD Render	生成OpenUSD数据的全光线追踪RTX渲染
USD Write	用户可修改OpenUSD数据并与之交互
USD Query	支持场景查询和交互式场景
USD Notify	追踪USD变化并提供更新信息
Omniverse Channel	连接用户、工具和世界，实现跨场景协作

应用

借助Cloud API可将Omniverse核心技术集成到软件应用及机器人或自动驾驶汽车等自主机器的仿真工作流中。

案例

目前，微软、西门子和Trimble等全球大型工业软件制造商都在将Omniverse Cloud API加入到其软件组合中。

Omniverser创新点

实时协作提升团队效率

允许全球团队成员在同一虚拟项目上实时工作，能即时看到对方更改和进展

软件兼容性高，可以随意切换不同软件

Omniverse能够连接和兼容多种不同3D设计软件，可在不中断工作流程情况下轻松切换不同软件

统一数据平台，确保信息一致性和时效性

通过Omniverse Nucleus存储、管理和共享3D世界和场景，所有团队成员都能访问到最新数据和场景

高质量渲染，提供逼真的视觉效果

基于NVIDIA RTX技术，提供逼真的视觉效果。它支持多种光线追踪技术，可在不牺牲性能情况下渲染高质量图像

AI驱动的创新，丰富创新过程且提升效率

如Omniverse Audio2Face可将语音转换为3D面部动画的工具，以及Omniverse Replicator可生成高度逼真的3D资产

灵活性和扩展性

Omniverse Kit允许开发者自定义和扩展Omniverse的功能，使其能够适应不断变化的需求和挑战

三、数据中心业务。英伟达数据中心业务提供从边缘计算到云端的全方位产品和解决方案，数据中心业务整体营收飞速增长，成为驱动公司市值不断向上的第一大业务

数据中心业务涵盖多层面

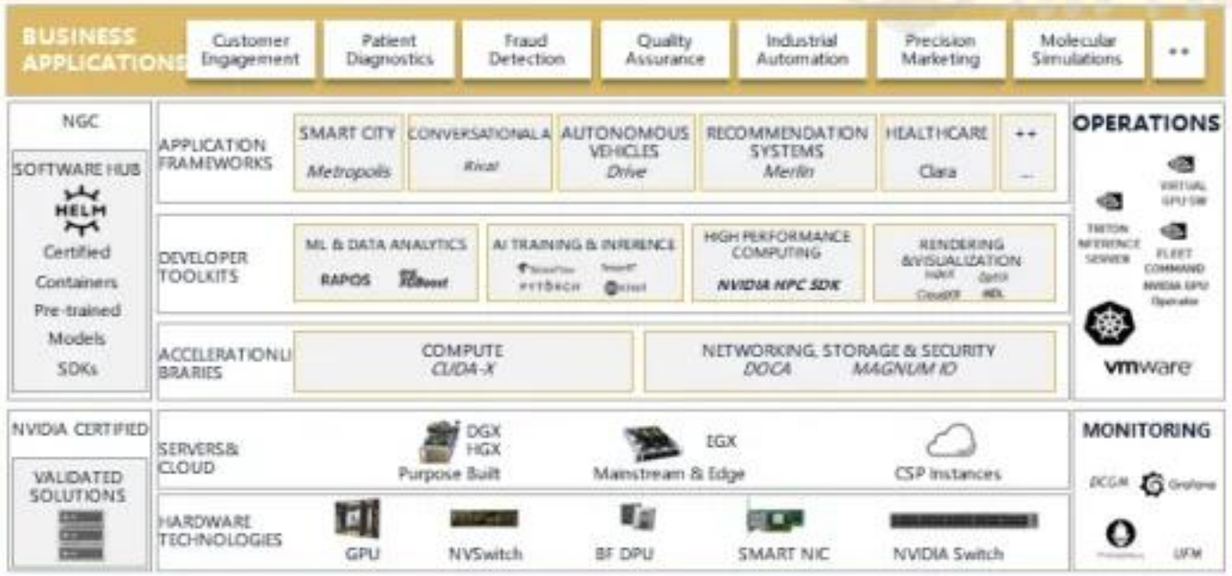
核心产品

- 基于Ampere和Hopper等架构设计的GPU芯片
- NVIDIA DGX系统
- 高速网络技术和解决方案

软硬件生态系统

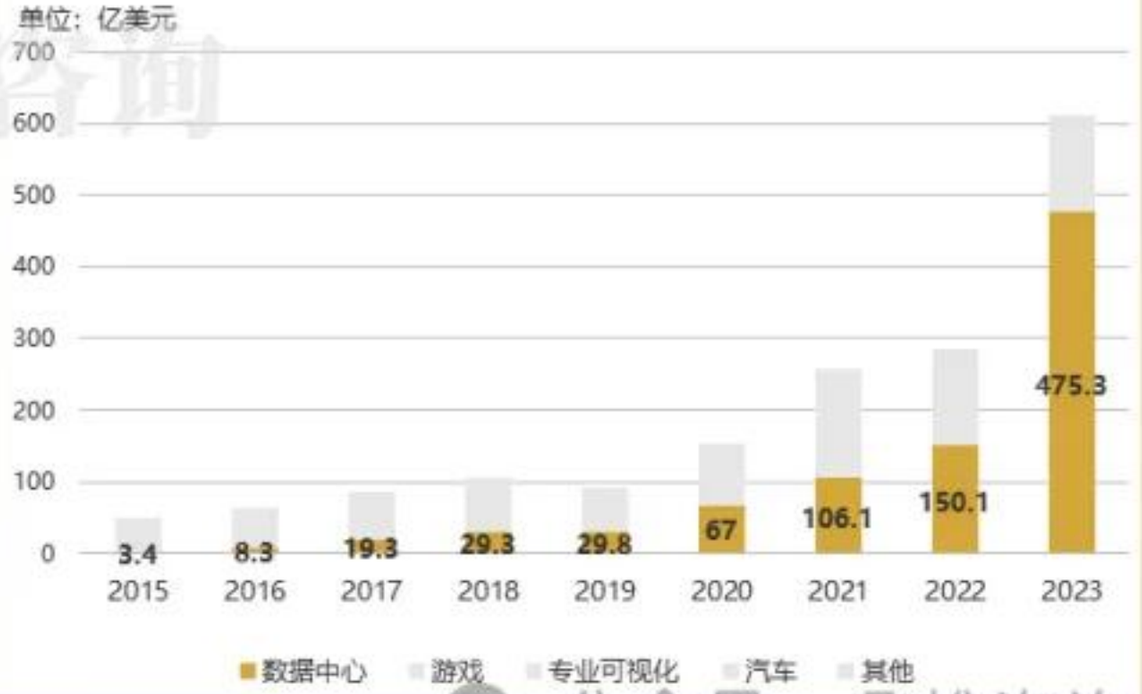
- 包括CUDA编程环境、
- TensorRT推理优化库
- RAPIDS数据分析库在内的软件生态系统，快速响应并提供定制服务，以满足不同行业和规模的企业的复杂计算需求

英伟达数据中心平台



数据中心业务成为公司业绩飞速发展的引擎

- **数据中心业务占英伟达23年营收的78%**，主要来自云服务提供商、专用 GPU 供应商、企业软件公司和消费互联网公司对数据处理、训练和推理能力的激增需求，以及汽车、金融服务和医疗保健等垂直行业的应用也加速业务的扩张
- 到2023年底，英伟达在全球人工智能芯片的**市场份额高达90%**，使AMD和英特尔等竞争对手远远落在后面



根据Open AI论文的测算，基于大模型的训练和推理环节对算力的巨大依赖，AI大模型的井喷式存在极大的算力缺口，这将带来对算力和AI芯片的天量需求

训练阶段

训练阶段算力需求=3×前向传递操作数×模型参数数量×训练集规模，训练所需GPU数量=总算力需求/（每个GPU每秒运算能力×训练时间×有效算力比率），单次训练GPT-4需要约**2.65万张A100**

推理阶段

单次GPT-4推理所需要的算力成本约为0.05美分，按照AIPRM统计，截至2023年12月，ChatGPT拥有约1.8亿用户，平均每月产生17亿次网站浏览量，则平均每天访问次数为567万次，假设每次访问进行10轮推理对话，则平均每秒进行推理次数为17/30*10/3600*10^8≈157407次
对应GPT-4需要**A100为27.7万张**

训练算力需求	GPT-3	GPT-4	SORA
平均参数数量（亿个，N）	1750	2800	100
单Token训练所需运算次数	1.05	1.68	0.06
训练数据	-	-	5亿图片+1000万个视频
图片分辨率*像素数	-	-	9.72E+04
Patch量（个）	-	-	1.75E+16
压缩比例	-	-	20%
Patch到tokens的换算比例	-	-	1.30E-03
单次训练Token数量（亿个）	3000	130000	45689
训练步数（steps）	-	-	20
单次训练所需总算力（TFLOPS）	3.15E+11	2.18E+13	5.48E+12
单次训练所需时间（天）	90	90	90
按上述时间计算，每秒的训练算力需求	4.05E+04	2.18E+06	7.05E+05
A100算力值（非稀疏，TFLOPS）	312	312	312
集群利用率（MFU）	34%	34%	34%
所需卡数	382	26477	6647

训练算力需求	GPT-3	GPT-4	SORA
平均参数数量（亿个，N）	1750	2800	100
单Token训练所需运算次数	1.05	0.56	0.02
训练数据	-	-	60
图片分辨率*像素数	-	-	30
Patch量（个）	-	-	1.94E+05
压缩比例	-	-	3.50E+08
Patch到tokens的换算比例	-	-	1.30E-03
单次训练Token数量（亿个）	1.00E-05	1.00E-05	456E-03
训练步数（steps）	-	-	20
单次训练所需总算力（TFLOPS）	350.00	560.00	182250
单次训练所需时间（天）	3	3	3
按上述时间计算，每秒的训练算力需求	116.67	186.67	60750.00
A100算力值（非稀疏，TFLOPS）	312	312	312
集群利用率（MFU）	34%	34%	34%
所需卡数	1.10	1.76	572.68

英伟达的数据中心业务结构包含了“CPU+GPU+DPU”三类芯片和网络产品的硬件和基于硬件的系统、平台和应用框架等AI基础设施

AI Foundation	Nemo	Picasso	Bionemo
PaaS 平台层	Nvidia AI		Nvidia Omniverse
IaaS 基础设施层	On Prem-本地	Hybrid-混合云	Multi Could-云 (DGX Cloud)
硬件	网络	交换机：Quantum-X800 InfiniBand、Spectrum-X800 Ethernet	DGX GH200 超级计算机
		其他：ConnectX-7 SmartNIC、 NVLink Switch 7.2T	
	三芯片	DPU：BlueField-3 DPU	
		CPU：Grace CPU Superchip、Grace Hopper Superchip CPU Superchip	
		AI训练GPU：H100 NVL、L4 GPU、L40 GPU (Ada Lovelace) AI推理GPU：Blackwell B200 GPU、 Blackwell B100 GPU、 GB200 Superchip	

在网络产品方面，英伟达通过自研及收购实现了网络产品的全覆盖，为客户提供端到端解决方案，提升加速计算性能

英伟达通过自研及收购Mellanox等公司实现交换机、网卡、DPU、互连等网络产品的全覆盖，为客户提供端到端解决方案

交换机	▪ 为企业级数据中心和高性能计算提供卓越的网络能力
网卡	▪ 可为云数据中心、电信运营商等工作负载提供硬件加速，并降低能耗成本
DPU	▪ 为企业在加速计算和AI应用提供安全的硬件加速设施
互连	▪ 可适应各种数据中心的连接速度和距离，通过提供高带宽、低延迟的连接提高计算网络的性能

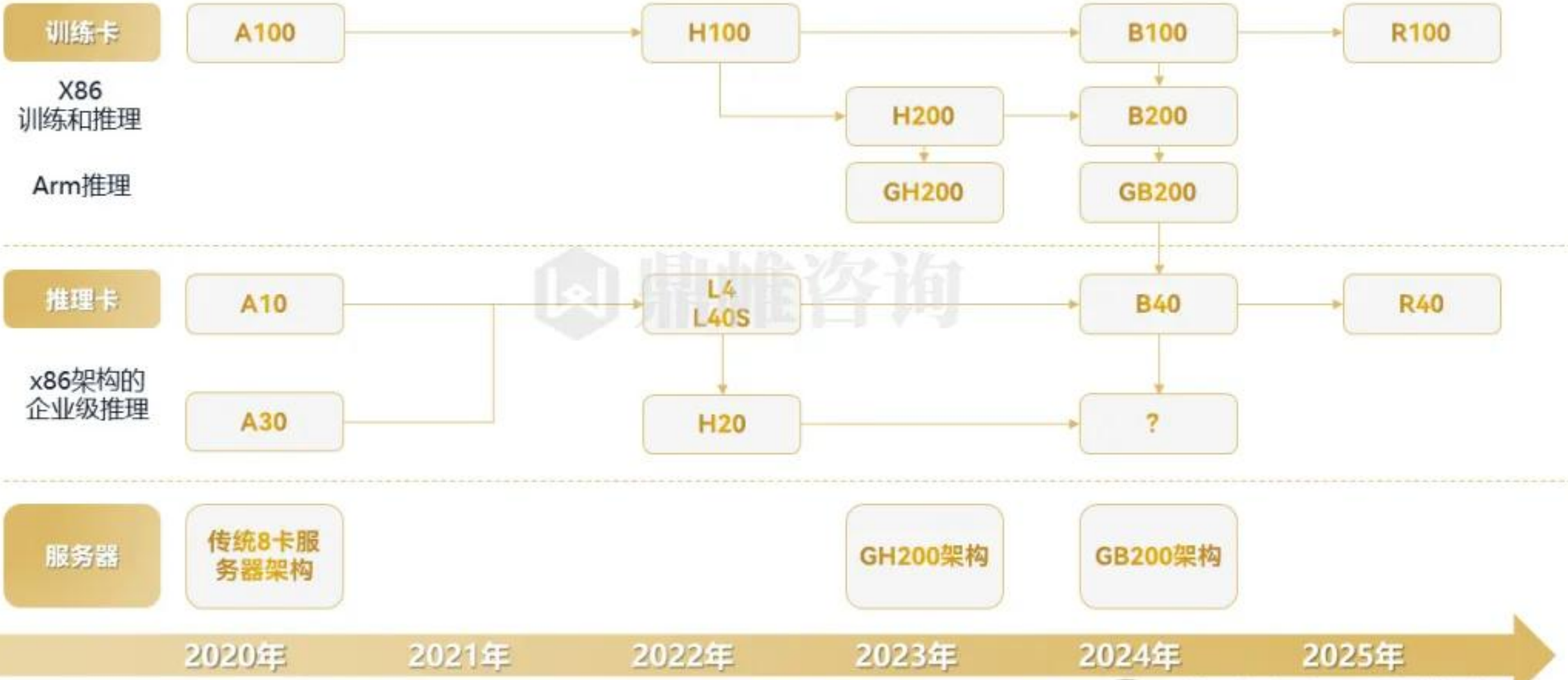
数据中心网络方案及其场景

场景	需求	网络方案
云	• 多租户 • 小规模工作负载	传统以太网
生成式AI云	• 多租户 • 以南北流量需求为主（云和用户间流量）	InfiniBand
AI工厂	• 少数用户 • 大规模模型 • 东西流量需求（数据中心内部流量）	NVLink、InfiniBand

网络

交换机	网卡	DPU	互连
			
Spectrum-4 以太网交换机	ConnectX-7 以太网网卡	BlueField-3 DPU	直连式铜线 (DAC)和分线缆
			
Quantum-X800 InfiniBand	ConnectX-8 InfiniBand网卡		直连式铜线 (DAC)和分线缆

根据数据中心产品路线图，AI芯片布局涵盖了训练和推理两个人工智能关键应用，训练推理融合，并侧重推理，同时支持x86和Arm两种不同硬件生态，预计将于2025年发布基于Rubin架构的R100训练卡和R40推理卡



英伟达全面布局AI芯片市场，从传统的服务各个云服务商向各个国家的私有云（主权AI）及电信云延伸，并积极布局边缘计算领域

利基市场	市场需求	竞争格局	英伟达布局
机器人	依赖多种类型的神经网络：通常需要异构的计算架构，例如用于导航的SLAM（同时定位和映射），用于人机界面的会话AI，用于对象检测的机器视觉，均会在不同程度上使用CPU、GPU和ASIC	目前英伟达、英特尔和高通正在这个领域进行激烈的竞争	打造NVIDIA Jetson™平台，作为专为机器人和嵌入式边缘 AI 应用打造的卓越平台，并通过NVIDIA JetPack™ SDK 提供支持，能够加速软件开发
智能工业应用	涉及制造业、智能建筑、石油和天然气领域	FPGA厂商凭借其遗留设备，也基于FPGA架构的灵活性和适应性，在这一领域表现突出	
“非常边缘”	将超低功耗AI芯片组嵌入WAN网中的传感器和其他小端节点中	由于重点是超低功耗，这个领域主要由FPGA厂商、RISC-V设计和ASIC厂商主导	

数据中心的下游客户涵盖云服务提供商、私人企业、主权国家、电信企业、智能设备制造商等，其产品和服务被广泛应用于大数据分析、人工智能训练、处理高度敏感数据、构建5G基础设施以及边缘计算等多个专业领域

	主要客户	应用场景	应用方向	增长的驱动力
公有云市场	• **云服务提供商：** 亚马逊AWS、微软Azure、谷歌云、甲骨文云、阿里云、腾讯云、华为云等	• 广泛应用于大数据分析、AI 训练	• AI模型训练、推理、实时数据处理等	• 算力和存储方面的需求不断增加
私有云市场	• **主权国家**：如、日本、法国、意大利、新加坡等 • **私人企业客户**：如苹果、谷歌、微软、Meta、特斯拉、京东、阿里、腾讯、字节跳动、小米、OpenAI 等	• 应用于高度敏感数据处理、企业内网应用、行业专属应用如金融、医疗、制造等	• **安全性与隐私**：满足对数据隐私和安全性的高要求，如数据主权保护 • **定制化解决方案：** 根据行业需求定制化 AI 和数据处理服务。	• 对数据的合规和安全需求更加严格
电信云市场	• **电信企业：** 爱立信、诺基亚、新加坡电信（Singtel）	• 提供 5G 基础设施、边缘计算、实时流媒体、低延迟网络等服务	• **网络虚拟化与自动化**：基于 AI 的网络优化和自动化运维 • **高性能通信服务：** 提供 RAN 性能优化、下一代通信服务的创新	• 物联网设备对时延和通讯依赖性有更高的要求
边缘计算市场	• **智能设备制造商**等	• 应用于自动驾驶、智能制造、智能零售、远程医疗等领域，靠近数据生成源头进行处理	• **实时数据处理**：支持在本地节点进行快速处理，减少延迟 • **分布式 AI**：在设备端运行 AI 模型，提供即时响应能力	• 边缘计算的应用场景不断扩大

英伟达将AI算力分为五个层次，每层均达到现有技术条件下的最优性能配置，确保了整体计算架构的高效性与先进性

五层算力概览

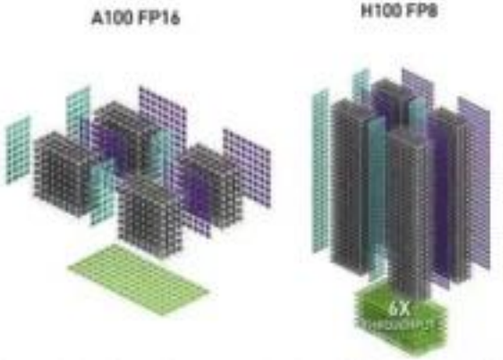
第一层算力 单芯片算力

进化原理： 在同等工艺制程约束下，芯片的面积越大，晶体管的数量就越多

进化方式： 提升光刻工艺、芯片蚀刻、晶圆物理限制

进化历程： 提升Tensor Core架构设计，包括优化Transformer流水线，设计专属CUDA驱动

架构	SM数量	算力FP8	算力FP16
Volta	80	N/A	125 TFLOPS
Turing	40	N/A	65 TFLOPS
Ampere	108	624 TFLOPS	312 TFLOPS
Hopper	132	1.97 PFLOPS	985 TFLOPS
Blackwell	600	10 PFLOPS	5 PFLOPS



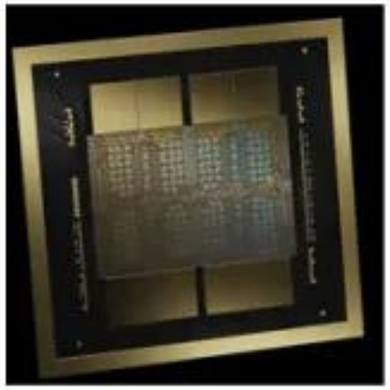
通过Die-to-Die互联技术针对在同一个封装内两个芯片裸晶间提供数据接口，通过NVLinK能够在单个封装中将两个处理器连接成一块超级芯片的板级互连

第二层算力 Die-to-Die互联

进化原理：通过在板内增加贯穿的铜线封装，在PCB版背面添加金属焊球导通不同位置的连线，实现复杂结构的2.5D互联

进化方式：从双Die到多Die互联，提升互联速度，组成一块更大的Chiplet

进化历程：目前由于技术受限，芯片通信互联在限制电路复杂度等外部条件后，单个基底上能够实现的最好的解决方案是双Die互联

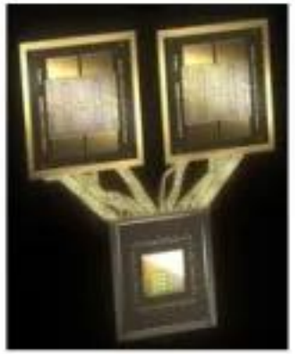


第三层算力 NVLink互联

进化原理：推出GB200 NVL72 将2个Blackwell GPU和1个Grace CPU装在一个主板上，将36块PCB板通过铜缆电信号互联，搭载72个Blackwell GPU核心

进化方式：提升NVLink配置内存，互联更多显卡

进化历程：从P100搭载的NVLink一代进化到Blackwell搭载的NVLink五代



通过GPU分布通信技术NVSwitch实现多个GPU之间实现全互联和机柜间通信，并最终将多个SuperPOD通过光交换机、光线拓扑互联，形成AI超级算力工厂

第四层算力 NVSwitch

进化原理： NVSwitch是一块独立NVLink芯片，利用Scale up超级计算机基本理论，在机柜间实现通信

进化方式： 从单一机柜进化到更多机柜，8台机柜的情况下，将拥有11.52Exa FLOPS

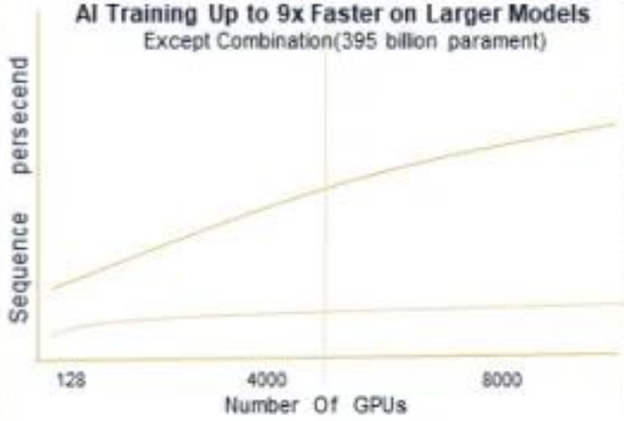
进化历程： NVLink最多连接576个GPU，GB200 Super POD叠加8台GB200-NV72的机柜

第五层算力 AI超级工厂

进化原理： 多个SuperPOD通过光交换机、光线拓扑互联完成

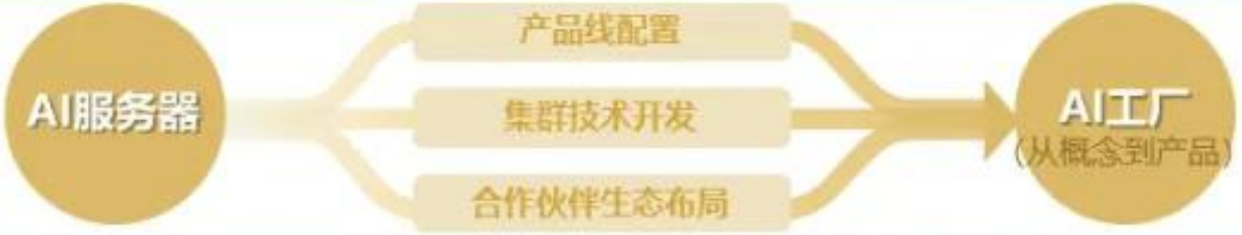
进化方式： 从单一SuperPOD进化到多个SuperPOD，形成AI超级算力工厂

进化历程： 2023年GTC大会提出最高支持8000个GPU进行大语言模型训练



英伟达将持续布局AI工厂，通过其DGX系列AI超级计算机解决方案，引领人工智能发展的下一代可扩展基础设施

AI工厂布局规划



发展计划

- 未来十年，要把深度学习的技术能力再提升100万倍，逐步实现构建超级AI计算工厂的宏图
- 未来要成为世界上唯一一家能吸引大家来共同打造数据中心规模和全栈人工智能超级计算机公司

产品线布局

产品线	产品线定位	主要产品
DGX SuperPOD	相当于NVIDIA自产的顶配和高配AI服务器集群	DGX GB200 SuperPOD DGX B200 SuperPOD
DGX BasePOD	提供给X86服务器厂商、在X86技术体系内的AI服务器集群参考架构，中低配版AI服务器集群	DGX B200 DGX H100

AI工厂

一种新型数据中心

定义	AI工厂也称为AI数据中心，是一种有别于传统数据中心的新型数据中心，提供生产先进AI模型所需的基础设施和专业资源
使命	AI工厂创造收入，也创造智能
意义	为整个行业生产AI，进一步为所有行业带来显著的生产力提升，帮助企业降低成本、提高能源效率，并扩大收入机会

英伟达DGX SuperPOD超级计算解决方案

介绍

- 引领人工智能发展的下一代可扩展基础设施
- DGX SuperPOD是英伟达推出的高度集成的超级计算解决方案，允许用户快速构建和部署大规模的GPU集群，以应对复杂的AI和机器学习挑战

构建单元-DGX GB200

- 36个GB200，共72个Blackwell GPU，36个Grace CPU
- 13.3TB的HBM3e显存、30.2TB的高速内存以及240T高速显存
- 通过第五代NVLink连接，采用水冷技术散热
- FP4精度下提供11.5 EXAFLOPS人工智能超级算力，整个DGX系统在FP4精度下算力达到1440 PFLOS

特性

- 灵活的拓展性，默认DGX SuperPOD由8个DGX GB200系统组成，理论上可配置任意数量的DGX GB200
- 具有高性能的算力，且集成了NVIDIA的软件栈，用户可轻松部署和管理

应用场景

大规模的机器学习以及高性能的计算

数据中心业务面临来自美国出口管制、下游客户自研芯片、安全与隐私限制等多重因素的影响，为业务发展带来潜在风险和不确定性

美国出口管制

- 英伟达在中国的数据中心营收占比从2022年19%降至个位数

下游客户自研芯片

云厂商考虑通过技术手段控制资本支出，与英伟达的增长诉求存在冲突

① **英特尔：芯片厂商**

计划提供制造服务，可能提高行业制造AI芯片的能力

② **OpenAI：大语言模型**

正筹集5-7万亿美元来建设AI芯片等基础设施

③ **亚马逊、谷歌和微软：大型云计算公司**

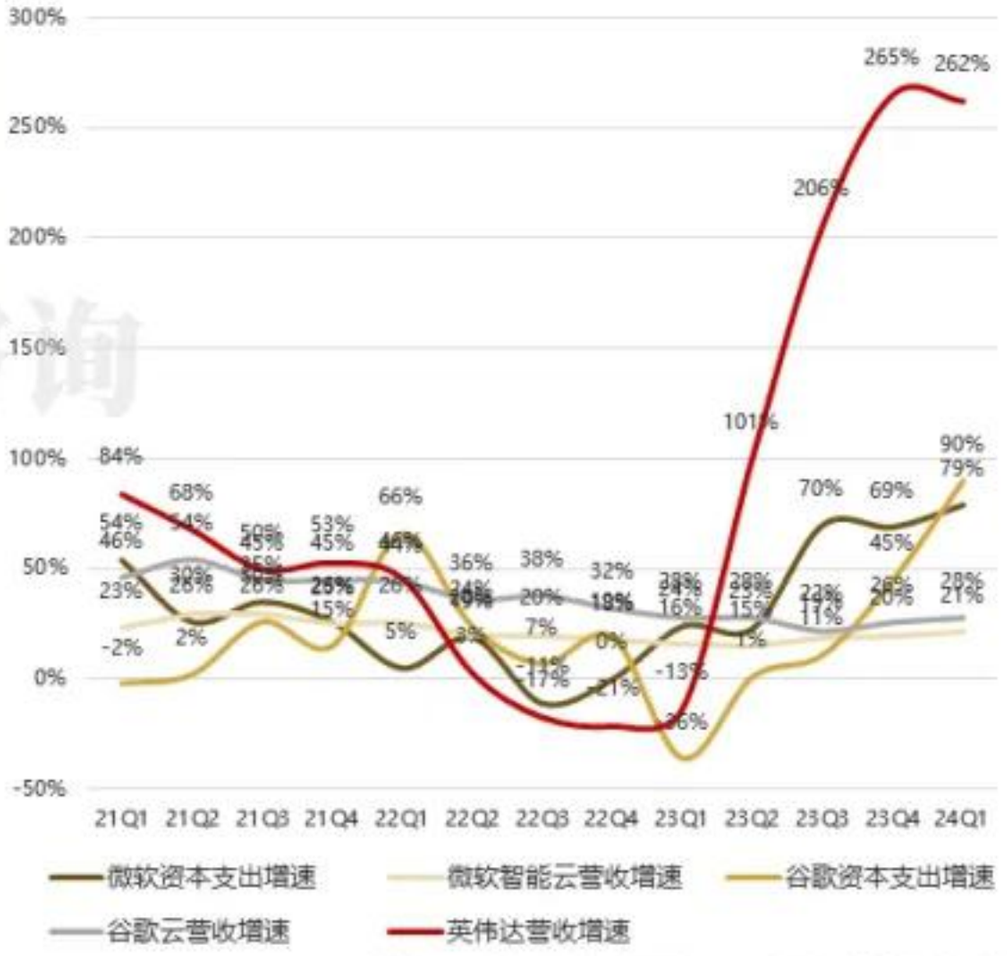
设计自己的人工智能芯片，减少对英伟达芯片的依赖

④ **微软**开发可加速服务器之间数据流的网络设备，与配备其芯片的英伟达网络设备展开竞争

安全与隐私限制

- 由于安全性及隐私，目前云服务提供商提供的人工智能用例要求更加严格

重点科技公司营业收入、资本支出与英伟达营收增速对比



为应对潜在风险与不确定性，英伟达积极拓展新的客户类型，借助各个国家“数据不出境”监管要求，大量出售先进AI芯片为国家和地区赋能建立“主权AI”

主权AI

背景 数据不出境

欧盟、中国等国家或地区数据监管的重要红线， AI主导权在政府手中，能有效管辖本国AI及其产出内容，避免他国AI在本国获取敏感数据

定义 主权AI涵盖物理和数据基础设施。

后者包括自主基础模型（例如大型语言模型），由本地团队开发并在本地数据集上进行训练，以促进特定方言、文化和实践的包容性

常见落地模式

英伟达主权AI业务

目的 向各国政府出售A100/H100/B100等先进AI芯片

营销 **与全球知名软件公司合作：**英伟达提供芯片等硬件，Oracle提供软件服务，联合拓展政府客户
创始人黄仁勋在迪拜世界政府峰会直言，每个国家都应该建立自己的“主权AI”

收入 占比不高，但增长明显。 2024年“主权AI”业务将为英伟达带来近100亿美元营收。而在2023年，该业务收入为0

采购英伟达芯片国家政府的四个共性

积极融入全球AI和半导体产业链，寻找产业升级机会	欢迎海外投资，希望借此强化本国云和AI能力
与国企和公共事业公司合作建设本国AI基础设施	关键领域保留自主权，扶持本国初创AI公司，用本国语言开发符合本国价值观的大模型

中国不会是英伟达“主权AI”的使用者

- 因为中国是极少数能够实现AI全产业链自主可控的国家之一
- 中国的三大电信运营商、各地方城市均在采购国产AI芯片建设智算中心

英伟达主权AI的主要客户

国家	建设方	英伟达的产品服务
美国	Oracle	Oracle用A100、H100芯片搭建美国政府云
新加坡	新加坡电信Singtel	用H100芯片升级国家超级计算机中心，在东南亚建设节能AI工厂
日本	软银集团、KDDI、Sakura Internet	为5G、6G应用和创建生成式AI平台和AI工厂网络，开发日语大模型
印度	塔塔集团、瑞莱斯、信实工业	塔塔集团采购GH200芯片建设超级算力中心，信实工业定制印度语AI大模型
法国	伊利亚特集团旗下云公司Scaleway	127个DGX H100系统、1016个H100芯片，英伟达 AI Enterprise软件
意大利	瑞士电信集团意大利子公司Fastweb	采购英伟达AI芯片建设超算中心，定制开发意大利语大模型
新西兰	TEAM IM、Oracle	Oracle采购英伟达芯片，与TEAM IM建设新西兰本地云服务
阿联酋	阿联酋电信Etisalat、Du	Oracle采购英伟达芯片，[illegible]建设主权云

所有西方的AI大模型在中国没有一张算力卡，这意味着中国所有的AI都是将数据送到国外云端推理后再返回的，这是中美科技竞争不可接受的，因此构建基于国产设备和中文的大模型的主权AI保护数据安全时不我待

国内AI大模型数据现状

AI是中美科技竞争最高的前端

- 所有西方的AI在中国没有一张算力卡
- 中国所有AI的推理都要把数据送到国外去
- 美国去年出台了对中国AI算力发展的限制

华为一直致力于打造全栈、不依赖西方的技术，让中国有第二个选择

8月28日上午
2024中国国际大数据产业博览会开幕式
华为董事、质量流程IT总裁陶景文

未来方向

AI模型和工具链跟数据的工具链深度整合

基础模型 / 企业数据 → 企业垂直应用 → 企业知识库 → 企业应用场景 → AI训练的数据链

华为AI发展板块

①建设“端到端”新型基础设施的各种硬件软件
②把这些硬件软件布局到华为云上，通过服务的方式提供给广大的企业

英伟达一直努力试图进入电信云领域，携手电信行业伙伴，通过其移动基站和边缘计算技术，推动5G网络智能化和运营商的数字化转型

前瞻布局移动基站，剑指边缘计算

目的意义

- 将AI能力和平台扩展到电信业，支持电信运营商内部数字化转型

实施途径

1. 在可信电信基础设施上开发和部署主权AI
2. 通过生成式AI增强电信运营能力
3. 推动RAN的性能（提升）和创新

已实施

电信运营商合作	宣布了与电信运营商和供应商的合作伙伴关系或计划，包括与Telenor Group、新加坡电信（Singtel）、 Indosat Ooredoo Hutchison等的合作
通信设备商合作	与诺基亚合作，改进cloud RAN解决方案以及进行AI-ready RAN的开发
成立AI-RAN联盟	成立AI-RAN联盟，创始成员包括AWS、Arm、DeepSig、爱立信、微软、诺基亚、英伟达、三星电子、软银和T-Mobile等，持续参加3GPP的5G-Advanced AI/ML讨论

英伟达积极部署【英伟达云】服务，通过投资云服务商及其竞争对手、提供定制芯片以应对云服务提供商自研芯片寻求替代的威胁，巩固其在AI和高性能计算领域的市场领导地位，扩大市场份额，并增加营收来源

英伟达云（DGX Cloud）

一项云计算服务，基于DGX系列超级计算机为用户提供高性能计算资源，能够更便捷地访问高性能计算资源快速进行复杂的计算任务

市场优势

巩固领导地位

- DGX Cloud 依托英伟达领先的 GPU 技术，进一步巩固其在 AI 和高性能计算领域的市场领导地位。

扩大市场份额

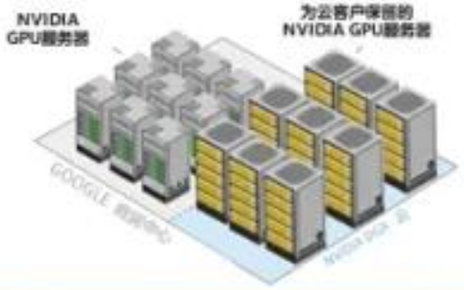
- 通过租用云巨头的服务器再通过自身的优势吸引到云巨头的客户，从而可以拓展其业务覆盖面，进入更多行业和应用场景

增加营收来源

- DGX Cloud 为英伟达提供了一种新的商业模式，即通过订阅和按需计费的方式提供计算服务，公司预计，这代表着3000亿美元的潜在营收机会

Nvidia的特洛伊木马?

Nvidia的新云服务DGX Cloud，实际上是将传统云服务提供商（此处为Google）运行的数据中心划分开来，以便Nvidia能够控制和增强GPU服务器，为其自己的客户服务。



英伟达正在积极推动公司进军云服务领域

背景：由于亚马逊、谷歌和微软开发自有AI芯片，并且云服务商难以建立足够的数据中心容纳英伟达的GPU，英伟达担心其销售将受到影响。

途径

投资主要云服务供应商

计划围绕亚马逊、微软、谷歌和甲骨文等主要云服务供应商投资90亿美元。

通过云厂商部署英伟达云（Nvidia DGX Cloud），销售自己开发的AI软件并赢得市场份额

效果

英伟达的投资为云服务商带来了稳定的收入，但也削弱了它们的影响力，原本云服务提供商采购AI服务的客户可能会倒向英伟达

投资三大云巨头的竞争对手

投资CoreWeave、Lambda Labs两家美国的中小云服务商

向其倾斜分配稀缺的GPU芯片

效果

降低云巨头的市场份额

成立芯片定制部门

英伟达正在建立一个专注于为云计算公司和其他公司设计定制芯片的部门

效果

占领新的市场，保护自己免受替代

英伟达业务从传统适合云端计算的CPU和GPU芯片向数据中心定制芯片延伸，并筹建云计算公司设计定制芯片的业务部门开拓新业务

英伟达持续开拓其他新兴领域和细分赛道的收入

背景

算力成本高 → 云计算公司开始用自研芯片替代英伟达产品 → 单位算力成本下降

ASIC芯片的**训练和推理**效率约为GPU的**10倍**和**100倍**，市场规模有望逐步提升

途径

筹建为云计算公司设计**定制芯片**的业务部门
- 已接触亚马逊、谷歌、微软

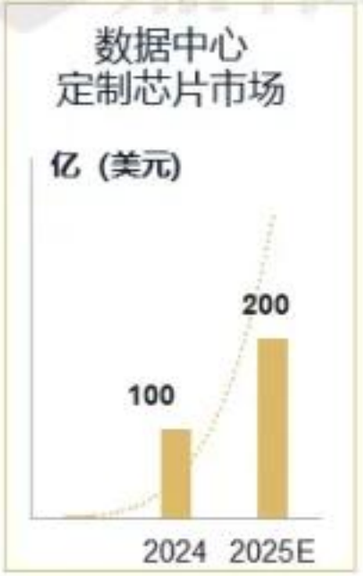
试图**往产业链上游延伸**，布局数据中心定制芯片业务

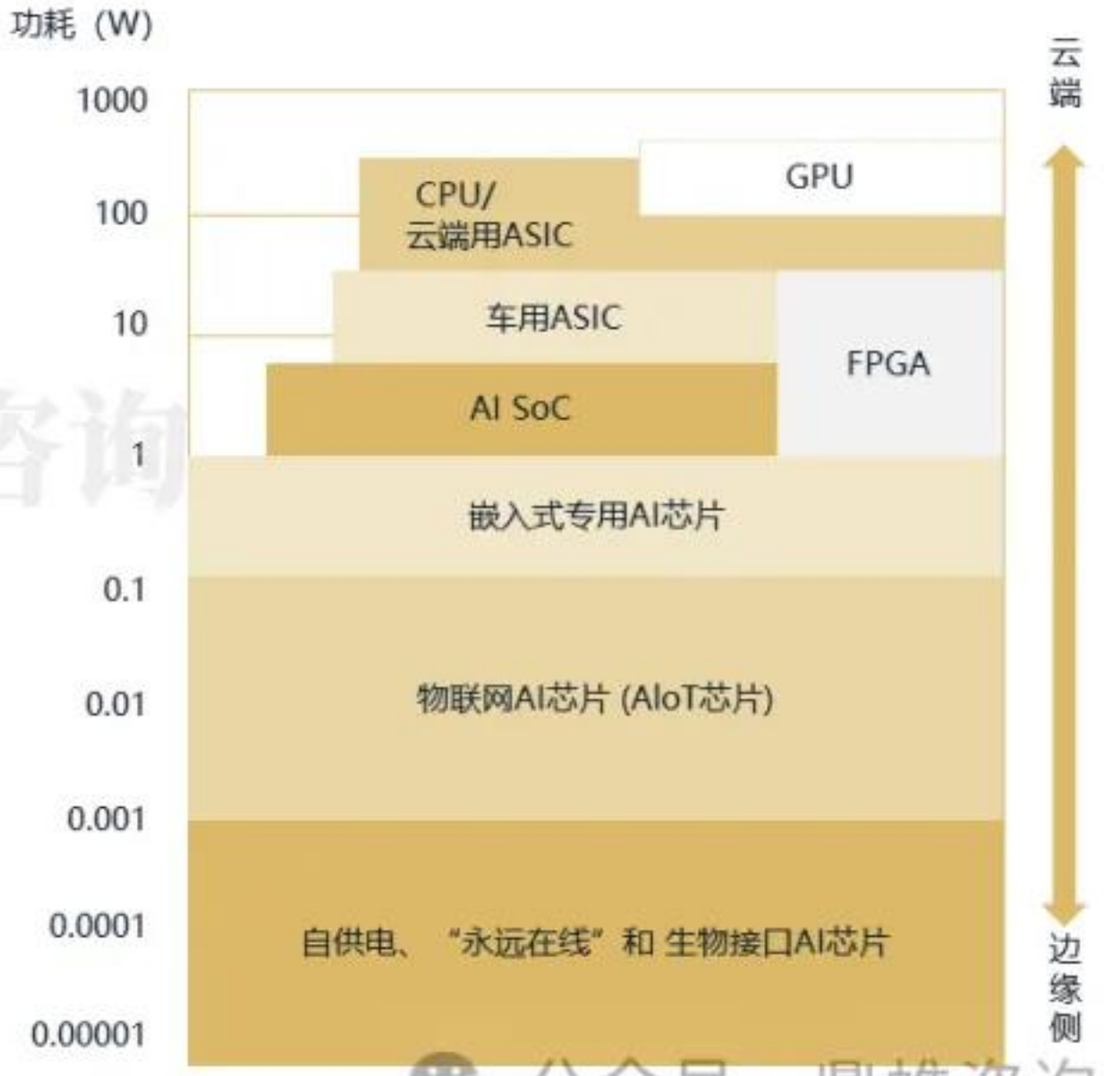
优势

专用的云定制芯片市场大，同时**巩固大客户的关系**

减少客户自研芯片带来的**替代压力**，为长期芯片走向**降本化、定制化**提前做好准备

市场前景





在云计算领域，英伟达积极布局下一代量子计算技术，利用技术研发能力的优势押宝颠覆性技术，通过建立合作伙伴生态、与大学展开合作等方式参与下一代云计算领域的竞争

四大举措

量子云服务	• **目的**：帮助研究人员和开发人员在化学、生物学和材料科学等关键领域推动量子计算的探索。 • **内涵**：一个数据中心，融合人工智能芯片和系统以模拟量子计算机。 • **规划**：英伟达目前没有自己的量子计算机，公司计划未来提供第三方量子计算机的访问服务。
CUDA-Q平台	• **方法**：基于开源CUDA-QTM的量子计算平台，允许在各种量子处理器(QPU)上直接执行混合代码，并通过cuQuantum加速模拟后端，提高量子算法的执行速度。 • **规划**：该平台已被四分之三的量子处理单元(QPU)公司采用，并允许用户在云中构建和测试新的量子算法和应用。
量子模拟器	• **目的**：量子模拟器将集成到Alphabet的谷歌云、微软Azure甲骨文云基础设施中，并被多家量子公司使用。
CUPQC	• **方法**：后量子加密技术(PQC)的软件库，提升加密算法的计算速度，实现NIST正在标准化的KyberPQC算法，使其在单个H100GPU上运行速度比传统CPU实现提高约500倍

竞争格局

英伟达刚刚进入量子计算领域，该领域已有微软、亚马逊、IBM等大公司在竞逐。尽管量子计算机的实际应用案例还不多，但全球对其计算速度的承诺和可能的军事与商业影响激发了广泛的兴趣。

布局生态合作

将其芯片应用于多个**超级计算机项目**，包括由富士通为日本国家先进工业科学技术研究所建造的ABCI-Q超级计算机，这台超级计算机配备超过2000颗Nvidia H100 Tensor Core.GPU，通过英伟达Quantum-2 InfiniBand互联-一个全球独一无二、完全可卸载的网内计算平台，预计明年初投入使用。

英伟达与超过160家**量子技术合作伙伴**建立了合作关系，并与17个量子计算框架中的15个进行合作，包括多伦多大学Classig、QC Ware等，以加速科学探索和量子计算的应用。

Quantum Computing Partners

与超过20所**大学合作**，开发量子计算课程，支持量子计算人才的培养。

随着AI算力需求逐渐向边端侧转移，英伟达积极布局边缘计算，试图在边缘侧复制云端的成功经验，并推出Edge AI Stack全栈方案，满足边缘场景性能需求，并与伙伴共促多领域应用

未来AI算力逐步由云向边、端侧转移

- 依赖云端远程数据传输仍存在时延和安全风险高，边缘侧AI计算则具有更小的时延、更低的通信依赖性以及更好的隐私保护
- 当前云端技术逐渐渗透到边缘侧，未来AI算力增量需求场景将逐步由云向边缘、端侧转移，边缘侧设备的市场规模将远大于云端数据中心

AI处理正在向边缘发展

英伟达发布边缘人工智能堆栈

NVIDIA Edge AI stack英伟达边缘人工智能堆栈

- 英伟达致力于将利用移动通信基站分散 AI 处理的技术实用化，根据 TDIA 预计，2023 年底全球 5G 基站将突破 480 万个，650 Group 的 Weckle 预计电信定制芯片市场每年约为 40 亿至 50 亿美元。

边缘计算包括提供者边缘、企业边缘、工业边缘、嵌入式边缘四大类型，凭借其低时延、减少通信依赖和增强数据安全性等优势，在多样化应用场景中迅速发展

边端计算：新兴行业趋势

边缘计算：分布式计算概念

将智能集成到边缘设备，允许在数据收集源附近实时处理和分析数据，无需上传到云或集中数据处理系统

低延迟 | 降低带宽要求 | 数据隐私 | 改进效率

优势

更小的时延	更低的通信依赖性	更高的数据安全性	× AI推理
提高操作安全性和客户体验	降低传输和存储成本，容纳更多传感器和程序	确保数据主权，保护隐私以及知识产权	从文本、图像、视频等数据中推断信息，以提供见解、做出预测并采取行动

边端计算丰富的应用场景

	定义	应用场景
提供者边缘	由服务提供商部署的服务器或基站等，通过靠近用户来提供高带宽和低延迟的计算资源	网络功能虚拟化（NFV）、云游戏等
企业边缘	企业本地数据中心，处理企业内部数据，减少延迟和成本，提高数据处理效率及安全性	办公自动化、企业内部数据处理和分析、设备管理等
工业边缘	处理工业设备和传感器上传的大量实时数据，对工业操作进行监控和优化，提高生产效率	工业自动化、设备监控、预测性维护、工厂车间优化等
嵌入式边缘	集成在设备或系统内部的小型计算单元，并能处理实时数据，无需依赖中央服务器	物联网（IoT）设备、智能家居、自动驾驶汽车等领域

典型的嵌入式边缘应用

运输和物流	▪ 数字标牌 ▪ 可疑活动监控 ▪ 仓库自主移动机器人 ▪ 交通流量管理	智能零售	▪ 自动结账及库存管理 ▪ 商店流量分析 ▪ 购物者分析 ▪ 社交距离检测	医疗保健	▪ 手术机器人 ▪ 医疗影像助理 ▪ 病人健康监测 ▪ 数字健康系统
工业和制造业	▪ 工业检测及预防维护 ▪ 感知机器人 ▪ 材料处理 ▪ 工厂地面视频分析	智能城市	▪ 交通分析及车辆计数 ▪ 车牌检测 ▪ 监控与公共安全 ▪ 智能停车系统	农业	▪ 人工智能授粉机 ▪ 牲畜健康管理 ▪ 选择性喷洒系统 ▪ 智能农场机器

英伟达针对边缘计算应用场景打造的Jetson平台，凭借其卓越的硬件性能和完备的软件生态，致力于通过提升计算效能和能效比，拓展应用领域，促进智能设备的技术创新

Jetson—世界领先的边缘AI平台

介绍

- NVIDIA Jetson™专为机器人和嵌入式边缘AI应用打造的平台，包括 NVIDIA JetPack™ SDK 、传感器、服务和产品的生态系统以加速软件开发
- 每个NVIDIA Jetson具有完整的系统模组 (SOM)，包括CPU、GPU、内存、电源管理、高速接口等

优势

节能 • 灵活性强 • 占用空间小 • 可定制

产品

Jetson AGX Orin Series
15-60W
32GB/64GB
100mm×87mm

Jetson Orin NX Series
10-25W
8GB/16GB
45mm×70mm

Jetson Orin Nano Series
7-15W
4GB/15GB
45mm×70mm

应用案例

- 宝马集团已采用全新NVIDIA Isaac™机器人平台对其车厂进行优化
- Jetson AGX Xavier用于全电动自动配送机器人，可以携带50磅重的货物，行程高达30英里
- 计算机视觉、机器人与无人机、医疗保健、制造检测、智能自助服务终端、传感器融合、生成式AI...

Jetson软硬件的未来发展方向

硬件 Jetson Thor将够执行更复杂的任务并安全、自然地与人和机器交互，具有针对性能、功耗和尺寸优化的模块化架构。

软件 JetPack 6 将通过微服务和一系列新功能进一步扩展 Jetson 平台的灵活性和可扩展性，可快速集成到工作流程中。

此外，英伟达还通过向企业客户提供软件服务Nvidia AI Enterprise，将智能当做基础资源贩卖给客户增加收入：通过端到端的云原生软件平台来简化企业在构建AI工厂时的工作，使开发人员专注构建和部署AI服务

Nvidia AI Enterprise

定义

端到端的云原生软件平台，旨在**简化企业在构建AI Factory或AI卓越中心时的开发、训练和推理工作**

功能

加速数据科学管道，并简化生产级人工智能的开发和部署

- 从原始数据获取、清洗、转换到模型训练、优化和部署的整个AI工作流程
- 基础设施管理和Base Command Manager
- 支持异构硬件环境，包括CPU和GPU节点
- 能够处理安全扫描和CVE跟踪，确保部署的安全性
- 整合了RAPIDS库、DALI工具、TensorRT和Triton推理服务器等
- 提供微服务以增强这些开源工具的易用性和可靠性
- 提供丰富的示例工作流程，帮助企业快速启动和调整AI项目
- 提供企业支持、安全性保证和服务水平协议，确保项目的稳定性和安全性

作用

使开发人员能够专注于构建和部署新的人工智能服务

提供服务

对象：在专有云和私有云中部署软件的**软件企业公司**

内容：**对企业的软件堆栈进行管理、优化、修补**

合作

对象：**云厂商**的大型团队 ← **并提供CUDA生态**

市场运行

将 Nvidia AI Enterprise 视为操作系统，每年每个 GPU 收取 4500 美元

受支持分支

分支类型	功能	适用对象	发布频率	生命周期
生产分支	确保API稳定性和定期安全更新	需要稳定性、在生产中部署人工智能	每6月一次	9个月
功能分支	树顶端软件更新	想要更快、最新开发环境的人工智能开发人员	每月发布	/
长期支持分支	/	高度监管行业	每2.5年	3年

Nvidia AI Enterprise 整体架构：实现了从底层硬件到顶层应用程序的全栈覆盖，为企业的快速开发与部署提供了强有力的支持

自然语言处理、对话式AI和图像分析等应用场景。

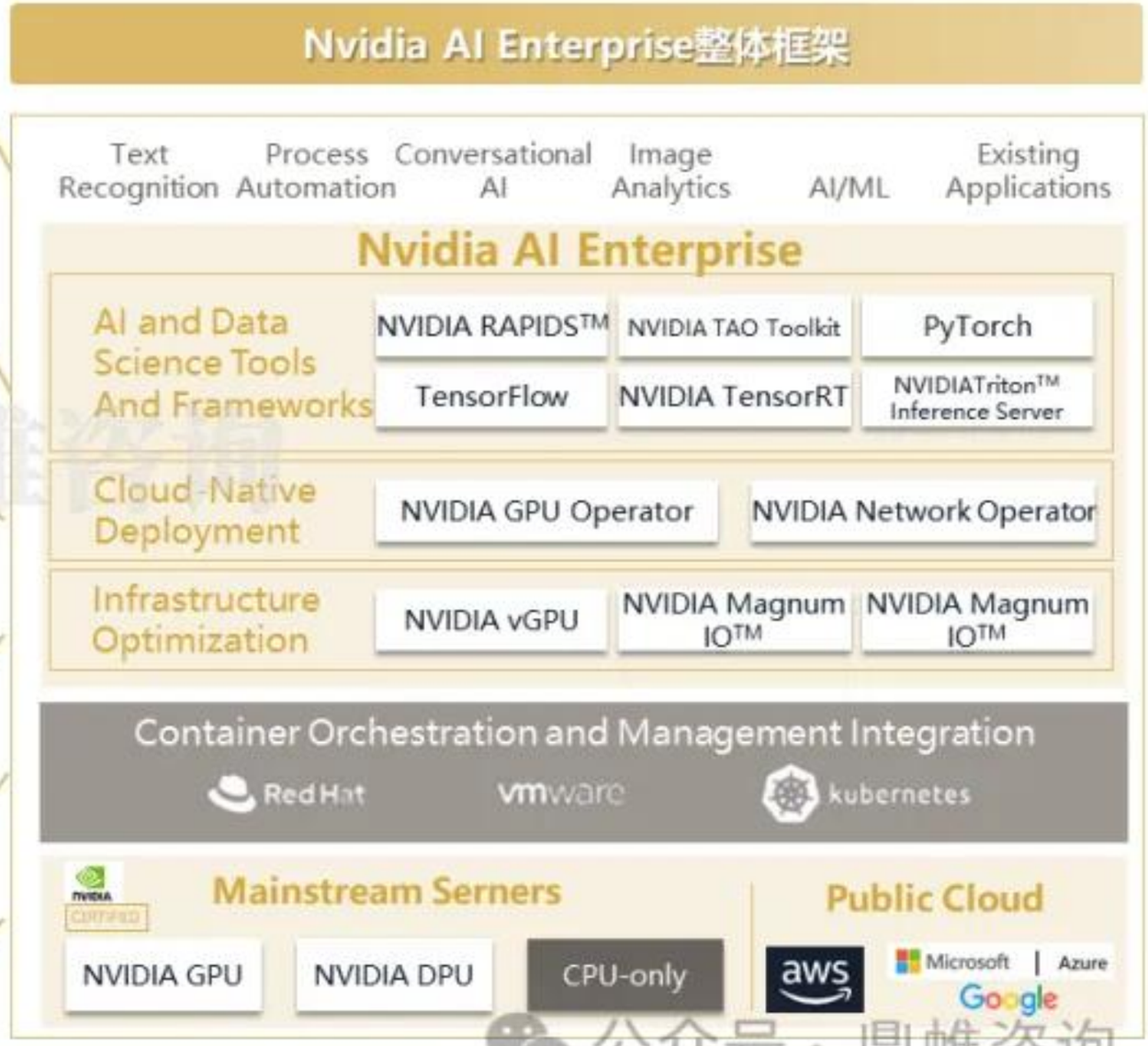
AI开发工具及AI框架，例如大数据加速的NVIDIA RAPIDS™，专注于迁移学习的NVIDIA TAO Toolkit，运用在训练及推理中的PyTorch和TensorFlow等等，并且这一切的工具、SDK及框架，NVIDIA都针对企业环境及用户进行了专门的优化，确保了更高的训练和推理精度，帮助任何想使用AI的企业进行快速开发部署。

云原生开发环境，GPU Operator和Network Operator可以让上层用户更好地在例如kubernetes或容器的环境中，更好挂载及使用GPU及底层网络，尤其是Infinitband资源。

底层架构的优化，包含NVIDIA vGPU，NVIDIA Magnum IO™以及CUDA-X™。

基于虚拟化和容器化的底座，包括Vmware、Red Hat以及Kubernetes。

底层硬件平台，包括搭载GPU、DPU甚至CPU-only的NVIDIA认证服务器



Nvidia AI Enterprise 服务的核心功能NVIDIA NIM：基于微服务架构的软件开发模式，具有高灵活性、可拓展性和可维护性，为企业高效利用AI模型和推理服务提供支持

英伟达 NIM 架构图

概念	微服务架构是一种软件开发模式，将应用程序拆分成一组小型、自治的服务。每个服务独立运行，并通过网络接口进行通信的架构模式称为微服务。
微服务架构优点	**有助于提高灵活性、可扩展性和可维护性**，因为每个服务都可以独立开发、测试、部署和扩展。

核心优势

1 随时随地部署算法模型推理服务

NIM专为可移植性和可控性而构建，可以使得算法模型的部署与调用支持跨各种基础设施。

2 使用行业标准API与优化的推理引擎

NIM针对每个模型和硬件设置利用经过优化的推理引擎，在加速基础设施上提供尽可能好的延迟和吞吐量，从而优化算法调用推理的耗时。

3 体验优化性能

部署于系统内的NIM接口，对调用的延时有着严格的要求，NIM的底层集成了众多获取来自NVIDIA和社区的优化推理引擎，包括TensorRT、TensorRT-LLM、Triton推理服务器等，可提高AI接口服务的应用性能和效率，同时提供更低延迟、更高吞吐量的推理效果。

4 使用自定义AI模型

NIM除提供官网的通用化大模型接口外，更多的提供了容器化部署模型的机制，无论是官网发布的通用类大模型，还是通过LoRA进行微调获得的自定义模型，都可以通过部署模型的接口，以便为特定用例提供模型接口调用，并且其后台使用Docker容器技术，将不同模型进行隔离处理，各个算法模型的调用互不影响。

5 随时随地部署算法模型推理服务

NIM拥有具有专属功能分支和严格验证流程的企业级软件，确保算法应用程序准备就绪，可以进行生产部署，除适用于个人环境下的算法模型接口的部署，更适用于企业级稳定要求的生产环境构建。

四、汽车业务。智能驾驶作为对算力要求极高的应用场景，英伟达10年前就积极布局汽车自动驾驶业务，目前已经推出4代汽车芯片

智能驾驶具体软件应用及需求

智能驾驶层级	L1	L2	L3	L4		L5	
软件应用	• 主动巡航控制 • 车道偏离警告系统	• 停车辅助 • 车道保持辅助	• 自动紧急制动 • 驾驶员监控 • 交通堵塞辅助	• 传感器融合 • 高速无人驾驶辅助		• 随时随地无人驾驶辅助	
硬件需求	-	• 超声波传感器4个 • 长距雷达传感器1个 • 环视摄像头1个	• 超声波传感器4个 • 长距雷达传感器1个 • 短距雷达传感器1个 • 环视摄像头1个	• 超声波传感器10个 • 长距雷达传感器2个 • 短距雷达传感器6个 • 环视摄像头5个	• 长距离摄像镜头2个 • 立体摄像机1个 • Ubolo1个 • 激光雷达1个 • 航位推算1个	• 超声波传感器10个 • 长距雷达传感器2个 • 短距雷达传感器6个 • 环视摄像头5个	• 长距离摄像镜头4个 • 立体摄像机2个 • Ubolo1个 • 激光雷达1个 • 航位推算1个

智能驾驶以及 ADAS 存在着巨大的算力缺口

最高安全可靠性性要求

- L5 级别的汽车会携带的传感器将达到 32 个
- 一辆自动驾驶汽车的数据量将达到 4TB/h
- Intel 测算出的一天数据量将达到 4000GB

英伟达近四代汽车芯片算力性能

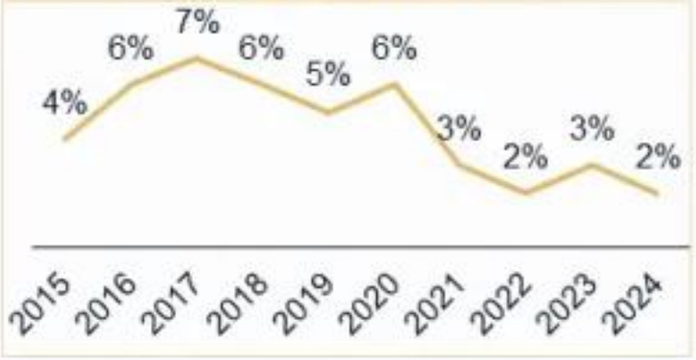
智能驾驶算力需求极高，未来市场空间广阔
是英伟达高算力新能芯片的巨大应用场景机会

英伟达的汽车业务持续增长，涵盖了从硬件到软件的全面解决方案，广泛与行业伙伴合作推动智能驾驶技术发展

汽车业务营收回暖

汽车业务近五年收入情况（百万美元）

汽车业务营收占比



业务涵盖了基础硬件软件转向端到端解决方案

DRIVE硬件

SoC芯片
Drive Orin
Drive Thor
Hyperion

DRIVE软件

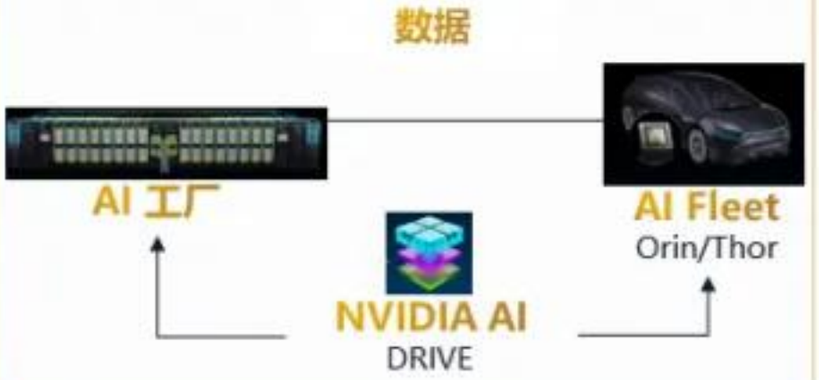
Drive Works/AV/IX/ Map/OS/

DRIVE 基础设施

培训
模拟
资源

端到端解决方案

平台总览



组成部分

DNN训练平台 用于感知的深度神经网络	DRIVE DGX
AV发展平台	DRIVE Hyperion
仿真模拟平台 生成数据集/验证自动驾驶堆栈	DRIVE Sim
模块化软件堆栈	DRIVE SDK

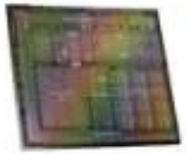
合作伙伴

汽车　卡车　出租车　供应商

模拟　传感器　软件　地图

从2015年开始，英伟达开始涉足自动驾驶业务，Drive Thor是当前汽车业务的最新一代芯片，并陆续与与特斯拉、奔驰、理想和比亚迪等企业在自动驾驶领域进行深度合作

核心产品

2015年	2016年	2018年	2019年	2020年	2021年	2022年	2024年
 DRIVE CX DRIVE PX	 Drive PX2	 Drive AGX Xavier/Pagasus	 Drive AGX Xavier Industrial Drive AGX Orin		 Drive Atlan	 Drive Thor	

合作伙伴

2015年	2016年	2018年	2019年	2020年	2021年	2022年	2024年
• 开始研发自动驾驶 • 达成合作意向	**特斯拉** • 采用Drive PX2芯片			**奔驰** • 合作开发新一代自动驾驶汽车，全方位的合作模式，标志着英伟达从“供应商”转变为“承包商” **小鹏** • P7采用Drive Xavier平台		**蔚来** • ET7采用Drive AGX Orin平台	**理想/BYD** • 下一代采用Drive Thor车载计算平台 **长城/极氪/小米** • 下一代采用Drive Orin车载计算平台

英伟达持续迭代自动驾驶芯片架构，工艺制程和芯片构成，以满足自动驾驶技术算力持续增长的需求

芯片产品	发布时间	SoC芯片	工艺制程	芯片构成
Drive PX	2015年	TegraXI	20nm SoC	2个Tegra X1
Drive PX2	2016年	Parker	16nm FinFET	2个Tegra X2 2个Pascal MWM GPU
Drive Xavier	2018年	Xavier	12nm FinFET	1个Volta
Drive Pegasus	2018年	Xavier	12nm FinFET	2个Volta 2个Turing
Orin	2019年	Orin	7nm FinFET	2个Ampere
Atlan	2021年	Atlan	5nm FinFET	Ampere Arm CPU Bluefield DPU
Thor	2022年	Thor	5nm FinFET	Blackwell Arm AE CPU Bluefield DPU

DRIVE Thor在核心参数及制造工艺等方面具有显著优势，广泛吸引主机厂客户基于Thor构建未来汽车产品

内容	Orin-X	Thor-X	Thor-X-Super
CPU核心数量	12	14	28
CPU架构	ARM Cortex-A78AE	ARM Neoverse V2	ARM Neoverse V2
CPU算力	240KDMIPS	630KDMIPS	1260KDMIPS
GPU架构	Ampere	Blackwell	Blackwell
GPU算力	5.2TFLOPS @FP32	9.2TFLOPS @FP32	18.4TFLOPS @FP32
FP16算力	84TOPS	500TOPS	1000TOPS
INT8/FP8算力	254/0TOPS	1000TOPS	2000TOPS
存储带宽	205GB/s	273GB/s	546GB/s
PCIe	Gen4 24lane	Gen5 16lane	Gen5 32lane
晶体管数量	170亿	770亿	-
制造工艺	7nm	4nm	4nm

核心亮点及市场反馈

专为汽车生成式AI应用计算平台

- 首个Transformer引擎
- ARM Poseidon AE内核
- 搭载Blackwell架构

单个芯片整合多种功能

- 整合数字仪表盘、信息娱乐泊车、辅助驾驶等功能
- 可同时运行 Linux、QNX 和 Android
- 极大提高开发效率和软件更新迭代的速度

获得更多主机厂客户，合作客户认可并采用Thor

理想汽车	比亚迪	广汽埃安	小鹏
基于Thor构建未来汽车产品	基于Thor构建未来汽车产品	基于Thor打造高端豪华品牌昊铂下一代电动汽车，实现 L4 级自动驾驶	将Thor作为其下一代电动汽车的AI大脑

其他主机厂和智能驾驶解决方案公司也在持续布局车规级芯片，英伟达在市场竞争中仍凭借其算力优势占有一席之地

	芯片产品	发布时间	算力TOPS	工艺制程	自动驾驶级别	代表车型	竞争地位
特斯拉	HW1	2014年	<10	基于EyeQ3	L1/L2	Model S/X	特斯拉作为主要竞争对手，新版FSD已实现端到端大模型落地，海量数据对于系统优化助益
	HW2	2016年	10	基于Drive PX2	L2	Model S/X	
	HW3	2019年	144	三星14nm	L2	Model 3/S/X	
	HW4	2022年	216	三星	L3	Model S/X/Y	
Mobileye	EyeQ3	2014年	0.256	40nm	L2	特斯拉	EyeQ系列芯片在出货量上具有优势，累计超过6000万片，但其产品更新速度较慢，算力水平相对较低，预期通过Ultra巩固市场地位
	EyeQ4	2018年	2	28nm	L2+	哈弗H6	
	EyeQ5	2021年	15	7nm	L4	极氪001/宝马iX	
	EyeQ6	2024年	34	7nm	L4	-	
	EyeQ Ultra	2025年	176	5nm	L4	-	
华为	MDC 300F	2019年	64	-	L2+	哈弗H6	华为智驾全栈程度高，涵盖芯片、算法及各类传感器
	MDC610	2020年	200+	7nm	L4+	问界	
	MDC810	2021年	400+	7nm	L4/L5	极狐阿尔法	
地平线	征程3	2020年	5	16nm	L2	荣威RX5/理想ONE	地平线凭借高出货量大幅提升了市场占有率，J6P早于英伟达DRIVE Thor实现量产，先发制人抢占高算力市场份额
	征程5	2021年	128	16nm	L3	BYD汉/理想L8	
	征程6	2024年	560	7nm	L4/L5	BYD/广汽/大众	

汽车制造商倾向于自主研发智能驾驶芯片以解决英伟达智驾芯片的频繁变化和高昂成本以及国际政治经济关系的波动性，这为英伟达汽车业务的发展带来了潜在的风险与挑战

下一代智驾芯片频繁变化的不确定性

芯片规划改变

Altan芯片（Orin后续） → Thor芯片（更高算力级别） → 单芯片算力2000TOPS

芯片规格改变

单芯片组成的2000TOPS算力 → 两颗芯片组成的2000TOPS算力

Thor芯片内核 → Blackwell架构（与数据中心芯片相同）

国际关系变化的不确定性

- 无法确定Thor芯片组能否顺利进入到中国市场
- 英伟达需要尽快完成和奔驰的合作，打造样板工程，与更多车企合作

高昂芯片价格及市场竞争白热化的不确定性

芯片价格高昂 + 市场竞争白热化

英伟达汽车客户的自研倾向

（选择英伟达的客户更多是希望建立技术门槛，愿意投入资金的车企）

- **希望将产品迭代节奏掌握在自己手上**（而非等待英伟达固定周期的性能升级）
- **使用自身大模型并不一定依赖CUDA**（只用AI芯片去做智能驾驶，生态门槛较低）

案例：大客户蔚来的芯片自研

蔚来已经明确表示将会自研智驾芯片，并即将在ET9上首发上车的5nm神玑NX9031

为应对潜在风险与挑战，英伟达汽车业务的合作范围从传统主机厂向汽车零部件供应商再到汽车软件服务商逐步扩大，向汽车产业上下游延伸和渗透，持续提升英伟达在自动驾驶领域的影响力

类型	合作商	合作类型	合作内容
主厂机	比亚迪	从汽车拓展到云端	▪ 在 DRIVE Thor 上构建其下一代电动车车型 ▪ 利用 NVIDIA Isaac™ 和Omniverse™ 平台开发用于虚拟工厂规划和零售配置器的工具 ▪ 在其车辆中提供云端游戏平台
	特斯拉	购买GPU用于训练	▪ 囤积3.5万块H100GPU用于数据中心的AI训练，以支持其自动驾驶算法的开发和优化
	小鹏汽车、极越汽车、广汽、奇瑞汽车	搭载NVIDIA DRIVE	▪ 结合NVIDIA DRIVE 平台作为其下一代电动汽车的“人工智能大脑”
	日产、领克、路特斯、现代、路虎等	多环节合作	▪ 在汽车行业工作流程，如车型研发设计、配置选购等多个环节中应用英伟达AI
汽车供应商	Plus、Waabi、文远知行	基于DRIVE Thor提供不同的L4级解决方案	▪ 向市场提供L4级的人工智能自动驾驶解决方案
	富士康	生产基于DRIVE Orin的电子控制单元	▪ 面向全球汽车市场生产基于 NVIDIA DRIVE Orin™的电子控制单元（ECU） ▪ 富士康生产的电动汽车（EV）将采用 DRIVE Orin ECU 和 DRIVE Hyperion™传感器架构，以实现高度自动化的驾驶功能
软件供应商	联发科	以英伟达GPU芯粒开发汽车SoC	▪ 搭载英伟达 AI 和图形计算 IP，共同为软件定义汽车提供完整的AI智能座舱方案

英伟达从汽车的自动驾驶解决方案入手，并未止步于自动驾驶芯片，从服务汽车最终产品向服务汽车企业的设计、制造等生产环节延伸，提供虚拟工厂规划、仿真应用的模型训练方案

规划虚拟工厂
使用生成式数据补足仿真Corner Case

① 具体做法

- 进行实时算法验证，并将Omniverse数据接口开放给所有的生态圈协作方，从而提高自动驾驶的开发效率
- 多家中国车企通过Omniverse虚拟现实技术实现对汽车制造工厂的工作流程优化

② 实际案例

比亚迪	• 使用英伟达 Isaac, Omniverse 平台来开发用于虚拟工厂规划和零售配置器的工具与应用 • 智能工厂进行合作，利用Omniverse做自主机器的仿真
宝马	• 通过Omniverse 开设全球首家虚拟工厂，作为宝马计算数字化转型和提升效率的一部分

仿真应用
使用像素渲染，毫米波雷达和激光雷达进行数据训练和实时算法验证

① 自动驾驶汽车1.0时代

基于标注图像的训练，并在上面开发和部署深度神经网络的集成，会有40-50个深度神经网络从L2+层级转向更高级的自动驾驶

② 自动驾驶汽车2.0时代

基于视频进行模型的训练，是融合世界的统一模型

其规模将增长13倍，数据存储规模将增长17倍。以GPT4作为基础的话这可能需要上万的服务器节点，即达到超算水平

英伟达提供四大智能驾驶平台和解决方案，涵盖硬件、软件和仿真等多个方面，为自动驾驶汽车的开发、测试和验证提供全面支持，助力自动驾驶技术的发展和应用

四大智能驾驶平台

① NVIDIA DRIVE Infrastructure

完整的工作流平台，可用于数据的提取、管护、标记和训练，还可通过仿真来验证数据。

NVIDIA DGX™ 系统

- 能够为深度神经网络模型的大规模训练与优化提供所需的计算能力

NVIDIA DRIVE Constellation™

- 可在开放的硬件在环平台上实现基于物理性质的仿真，从而在上路之前对自动驾驶汽车进行测试和验证

② NVIDIA DRIVE AGX

具备可扩展和软件定义特性，提供先进的性能，助力自动驾驶汽车处理大量传感器数据，并做出实时驾驶决策。

③ NVIDIA DRIVE Concierge

车辆驾乘人员可以使用基于 NVIDIA DRIVE IX 和 NVIDIA Omniverse™ ACE（Avatar Cloud Engine）的一系列智能服务

NVIDIA DRIVE Chauffeur

- 基于 NVIDIA DRIVE AV SDK 的 AI 辅助驾驶平台，可以实现点到点驾驶

④ NVIDIA DRIVE Hyperion

用于设计自动驾驶汽车的完整开发平台及参考架构，集成基于 NVIDIA Orin™ 的 AI 计算与完整的传感器套件，加速开发、测试和验证

DRIVE AV

- 适用于自动驾驶的完整软件栈

DRIVE IX

- 无线更新驾驶员监控和可视化功能

解决方案

目标

- 让客户快速搭建、验证、部署L2自动驾驶技术，包括传感器套件和计算平台AGX两大部分

DRIVE Chauffeur | DRIVE Mapping | DRIVE Concierge | DRIVE Sim OV | DRIVE Replicator

DRIVEWORKS 加速库

Sensor Abstraction | DNN Framework | Image/Point Cloud Processing | Recorder | Calibration | Ego Motion

DRIVE OS

NVMedia | CUDA | TensorRT | NVStreams | Developer Tools

DRIVE AGX ORIN | DRIVE Hyperion | DGX | DRIVE Constellation

NVIDIA DRIVE全栈自动驾驶平台

英伟达构建从数据采集到验证、感知到规控、需求到交付的端到端自动驾驶解决方案，实现了纵向、横向和纵深的三大闭环

端到端平台	定义	组成部分及具体做法	
DRIVE Hyperion 自动驾驶开发平台	数据采集和验证开发的套件	传感器套件	• 毫米波雷达、激光雷达、摄像头、IMU、GNSS等
		计算平台 AGX	• 基于Orin SoC来构建，提供硬件、软件和应用服务，包括了DRIVE AGX Orin、DRIVE AGX Pegasus（L4/L5自动驾驶系统）和DRIVE Hyperion 8.1（参考架构）开发套件
DRIVE SDK 自动驾驶模块化软件栈	提供基础软件、中间件、应用软件全栈软件	应用软件	• AV：NCAP & Active Safety，自动驾驶，自动泊车 • IX：可视化，AI应用，OEM Cockpit
	汽车生态的关键：模块化设计、软硬件解耦：客户可以根据自身需求选择不同的模块进行开发，关联到英伟达硬件	中间件 工具包	• DriveWorks：中间件框架，由Sensor Abstraction, Image.Point Cloud Processing, Vehicle I/0, DNN Framework, Recorder, Calibration, Egomotion组成
		基础软件栈	• DriveOS：基础软件栈，由NVMedia、NVStreams、CUDA、TensorRT、Developer Tools组成
DRIVE DGX 深度学习训练平台	自动驾驶感知、规划、控制的模型训练和优化	• 通过DGX快速验证和训练大规模神经网络，并实现数据采集、数据标注、数据训练、模拟仿真，自动驾驶道路测试验证，形成数据闭环	
DRIVE Sim 仿真模拟平台	通过虚拟仿真出的模拟数据，与传感器采集到的真实数据对比，对模型和算法进行验证	• 与元宇宙结合，为自动驾驶开发和验证提供天气、道路、车辆、交通、虚拟世界等仿真场景，通过硬件在环（HIL， Hardware in Loop ）的方式测试并校验AI算法	

纵向、横向、纵深 三大闭环

纵向

- 实现了从车端到云端，从应用软件、中间件、基础OS、再到底层硬件计算的闭环

+

横向

- 实现了从感知到规控，从数据采集、标注、训练、仿真、验证的闭环

+

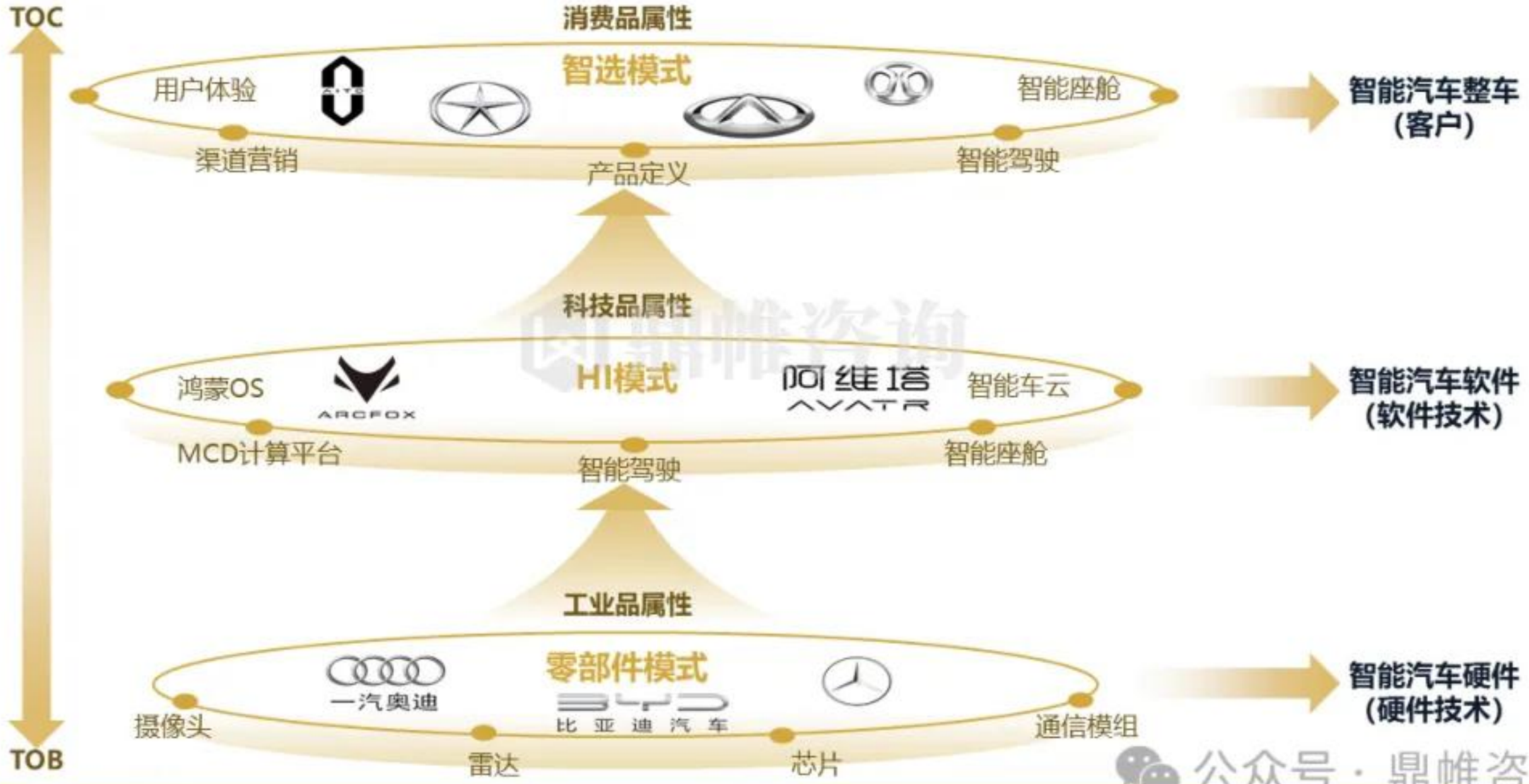
纵深

- 实现从需求到开发到交付再到维护，产品全生命周期的开发流程闭环

英伟达在汽车领域主要布局了四大业务，这四大业务华为目前均有布局，除此之外华为还布局了汽车产品设计、智能增量零部件供应，乃至汽车营销等一站式服务，就汽车业务而言，理论上华为可以完全替代英伟达，且更胜英伟达

英伟达汽车业务布局		华为汽车业务布局			
① DRIVE Hyperion 自动驾驶开发平台	数据采集和验证开发的套件	① 华为MDC计算平台 ADS智能驾驶系统	MDC计算平台：提供丰富的开发工具、API和协同功能 ADS智能驾驶系统：集成数据生成、模型训练等技术完成自动驾驶技术创新	⑤ 产品设计及研发	深度参与产品的设计与研发，基于对用户需求的深刻理解，与车企共同打造符合市场需求的汽车产品
② DRIVE Sim 仿真模拟平台	通过虚拟仿真出的模拟数据，与传感器采集到的真实数据对比，对模型和算法进行验证	② 全栈一体化仿真平台	集成了从基础设施到通信库、编译器、集群管理、专业服务等全栈一体化能力，旨在加速自动驾驶研发仿真场景的应用	⑥ 智能增量部件供应	基于汽车智能化提供全部智能增量部件供应，包括各类车载芯片、电机、雷达、摄像头等
③ DRIVE DGX 深度学习训练平台	自动驾驶感知、规划、控制的模型训练和优化	③ ModelArts 一站式AI开发平台	支持用户从数据处理、模型开发、训练到部署的全流程AI开发。该平台内置多种深度学习框架和工具，为汽车的深度学习训练提供技术支持	⑦ 汽车营销及品牌	华为利用其强大的品牌影响力和渠道，帮助车企进行产品营销和渠道拓展
④ DRIVE SDK 自动驾驶模块化软件栈	提供基础软件、中间件、应用软件全栈软件	④ 华为智能汽车解决方案（华为汽车技术栈）	集成了华为在汽车领域的深厚技术积累，包括芯片、操作系统、中间件、算法等多个层面	⑧ 其他一站式服务	其他包括用户体验、核心制造环节参与、供应链选择与管理、品牌合作与生态构建等新汽车工业时代下的汽车一站式服务

目前，华为服务汽车主要分为三大模式，本质上是完成了供应品-科技品-消费品的全面打通，进而实现了智能汽车硬件、软件、客户的链接，在华为鸿蒙OS下，众多合作车企共用一套系统，强绑定的关系推动华为成长为汽车界的“安卓”



同时，华为未来的服务半径不仅仅限于乘用车领域，而是基于全汽车门类和全场景应用提供智能化增量部件的全栈服务，这意味着华为这一汽车平台瞄准的整个交通领域

华为MDC智能驾驶计算平台产业模式


私人出行


长途运输


Robo Taxi


园区通勤


无人矿卡


快递物流


港口货运

场景应用
应用算法
（客户/伙伴）

华为MDC平台

HUAWEI

摄像头

毫米级雷达

超声波雷达

激光雷达

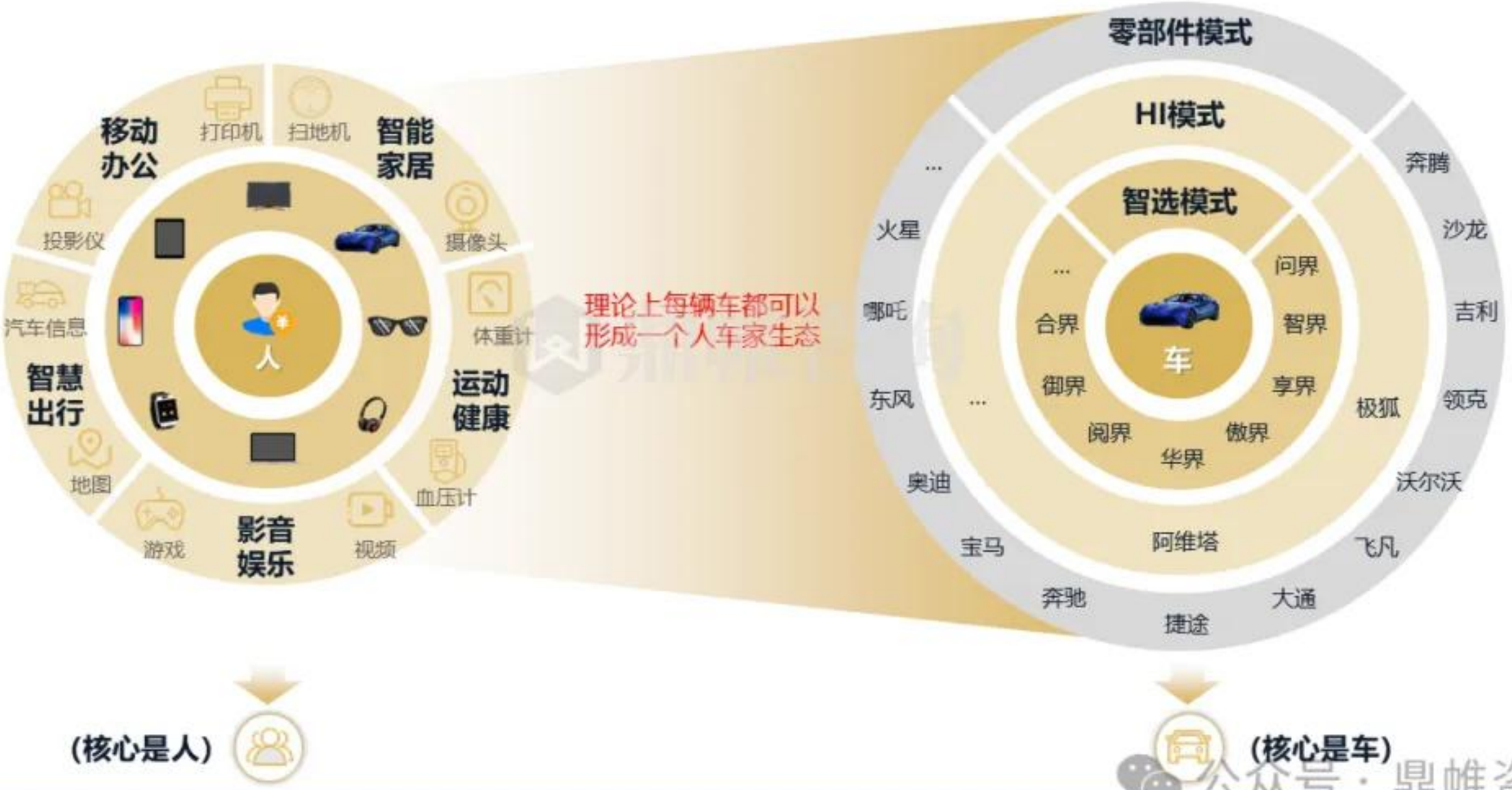
T-Box

组合定位

线控底盘

传感器
执行器
（客户/伙伴）

在三大模式推动下，华为将不断拓展朋友圈，形成以华为为中心的汽车产业生态，相比人车家生态，这一生态更大、更全、兼容度更高。极其庞大的生态+可无限拓展的“安卓”平台属性将极大推高华为汽车成长空间，在战略层面更胜英伟达



英伟达助力汽车厂家从软件定义汽车向AI定义汽车迈进，提出自动驾驶业务三步走战略，逐步实现端到端的自动驾驶系统，实现自动驾驶的真正价值

从软件定义汽车到AI定义汽车

	特点	开发	测试	部署
AI 1.0时代 软件定义汽车	• 需要大量工程师，大量算法开发量，且需要非常好的基础建设实现数据闭环	• 所有组件的编码/工程设计工作量巨大 • 边缘场景识别和融合速度慢	• 涉及海量路测 • 涉及海量大规模测试和运行 • 需要高效的基础设施来支持测试数据闭环-飞轮模式	• OTA更新
AI 2.0时代 AI定义汽车	• 云端训练模型，云端仿真验证，极大减少大规模设备部署和测试依赖性，通过自动数据驱动完成模型自我迭代	• 大部分模型在云端训练 • 完全以数据驱动来解决边缘案例 • 利用FM中的互联网规模数据加强泛化能力	• 通过仿真完成大部分验证（云端）	• OTA=模型更新

三步走战略

1 完善现有的L2和L2+系统

特点	基于规则的自动驾驶系统
目标	• 将英伟达的自动驾驶系统提升至市场领先水平或第一梯队的水平

2 在L2++领域取得新突破（目前阶段）

特点	代用AI逐步取代规则的自动驾驶系统
目标	• 实现端到端的可训练，将上游模型和下游模型打通，通过端到端模型的方式完成开发，并预计在不久的将来展示DEMO

3 计划在2026年量产L3级别的自动驾驶

特点	端到端的自动驾驶系统
目标	• 实现自动驾驶真正价值，正面释放了与竞争对手Mobileye、地平线、华为等争夺市场份额的信号

吴新宙作为英伟达智能驾驶团队灵魂人物，带领团队全面布局中国区智能驾驶，推动智能驾驶技术的发展和落地

智能驾驶团队成员

智能驾驶团队灵魂人物

吴新宙

现任	英伟达智驾软件业务总负责人，直接向黄仁勋汇报
曾任	原高通自动驾驶高级工程总监，小鹏汽车自动驾驶副总裁
过往经历	• 小鹏在智驾上从0到1的奠基人，曾全面主导汽车的自动驾驶技术路径和落地，奠定了小鹏汽车的高阶辅助驾驶领域的地位 • 高速领航辅助驾驶系统（HNGP） • 城市领航辅助驾驶系统（CNGP） • 小鹏最新的第二代智能驾驶辅助系统（XNGP）

团队其他关键人物

罗琦	原百度智能驾驶 L2+业务的车端整体软件架构以及规控和车辆交互技术负责人——英伟达汽车事业部工程总监——负责预测、规划与控制
Patrick 刘兰个川	原小鹏汽车自动驾驶 AI 负责人
Houman Tavakoli	原小鹏汽车软件架构总监
Parixit Aghera	原小鹏汽车北美团队的工程 VP
韩峰	原小鹏汽车多模态感知融合算法总监

团队两大目标

- 迭代量产速度
- 推动英伟达智驾转型

加快英伟达的智能驾驶算法能力落地并进入下一阶段

团队用人要求

- 过硬的专业领域背景
- 强大的自驱力
- 技术产品精益求精的追求

中国区智能驾驶团队全面布局

2023年年末，中国区五个部门开放招聘25个岗位

- 自动驾驶软件组
- 自动驾驶平台组
- 产品组
- 系统集成及测试组
- 地图及仿真组

五、应用场景。英伟达通过丰富的自研产品及针对性的解决方案，推动各个行业应用场景的智能化发展

AI+医药

医药研发

药物发现
锐格医药
Recursion
Genentech
德睿智药

AI模型开发
递归制药
Recursion
英矽智能

基因组研究
甲骨文
东京大学人类基因组中心（HGC）

医疗技术

手术生态系统
强生

IT建设与模型
火山引擎机器学习平台

基因数据分析
序禎达

AI+汽车

自动驾驶机器

边缘人工智能应用的开发
U-blox

人工智能驱动的信息娱乐系统
比亚迪

智能驾驶
极氪
理想
广汽埃安昊铂
阿尔特
长城
小鹏汽车和广汽埃安旗下Hyper
小米

数字孪生技术

渲染与模型优化
宝马

AI+机器人

机器人交互

交互通用场景
Apple

机器人研发与训练

足球机器人与人性机器人
逐际动力

工业机器人
西门子

自主移动机器人
比亚迪

机器人车队
富士康

载物机器人
九号公司

AI+军工

微处理器

无人机
俄军无人机部队

AI+零售

智能全渠道管理

智能供应链
空间虚拟副本

数据管理
沃尔玛

智能商店
AI摄像头

智能快餐店
智能订单处理

零售运营

自主商店平台
Nanostores

质量控制

闭环自监督

包裹分类

边缘服务器与推理服务器
USPS

废料处理

自动废物分析、分类系统
Recycleye

AI+电信

生成式AI

虚拟形象
新加坡电信

实时语音AI
T-Mobile

生成式AI

RAN的性能（提升）和创新
诺基亚

技术支持

开发者社区

其他

5G网络

HEAVY.AI

AI+娱乐

生成式AI
联想

实时通信
Avaya

其他

推荐算法

推荐系统开发
陌陌

排名模型
SNAP

广告推荐
腾讯

模型训练
美团

AI+高能计算

因子挖掘
DolphinDB

数据处理
AT&T

算力一体化
道客

数字孪生技术
施耐德电气

AI+交通

路面管理

交通管理
NoTraffic

交通枢纽管理
Assaia

停车管理

停车管理
DateFromSky

AI+安全

运营管理

医院运营安全
Artisight

场所管理

工作场所安全
Helin Data

校园安全
Icetana

机场安全
Ipsotek

人群安全
印度特兰加纳州当局

AI+医药：通过Clara开发套件及计算平台，提高医药研发及医疗技术水平，推动智能化医药解决方案的发展

合作方向	合作领域		合作企业	内容	技术	价值	特点
AI+医药	医药研发	药物发现	Recursion	与英伟达合作进行AI驱动的药物发现，特别是大规模的生化实验和药物研发流程的加速	AI 驱动的Recursion OS平台	加速药物研发流程，提高实验效率，缩短新药开发时间	依赖英伟达的云服务进行模型分发，通过AI实现药物发现的全流程自动化和优化。
			Genentech	在药物发现中优化和加速专有算法，将BioNeMo AI模型集成到基因泰克的工作流程中	BioNeMo AI模型	提高算法效率和新药发现的成功率，加快药物研发进程	将英伟达AI模型整合进基因泰克的药物发现工作流程，并在云端进行AI模型的授权和分发。
			锐格医药	利用英伟达Clara系列产品加速其药物研发平台的多个关键环节	NVIDIA Clara, CUDA，MIG	显著提高靶点发现、化合物筛选和优化的速度	通过GPU加速技术实现10倍以上的性能提升，尤其是在药物研发的核心环节。
			德睿智药	使用NVIDIA A100 GPU构建高性能分布式计算平台，以处理海量药物化学信息并加快模型训练	A100Tensor Core GPU	大幅提升海量化学信息处理速度，缩短AI模型训练时间	分布式计算能力和GPU加速技术，使得药物研发信息处理的效率提升10倍以上。
		AI模型开发	Recursion	利用英伟达云服务平台训练基于大规模生化数据的AI模型，应用于药物发现	BioNeMo AI模型	提高药物发现模型的精度和实用性，促进新药研发的商业化	通过英伟达云平台管理和分发大重生化数据与模型，并实现AI模型的商业授权。
			递归制药	开发和运营全球最快的超级计算机之一，专注于制药研究领域的超级计算	超级计算机BioHive-2	提供强大的计算能力，加速复杂药物研究任务的处理	BioHive-2在全球TOP500超级计算机排名第35位，是药物研究中性能最强的计算资源之一。
			英矽智能	开发跨领域的大型语言模型Nach0，应用于生化问题的解决	Nach0大型语言模型	提供一站式生化问题解决方案，涵盖多领域问题的处理	NeMo神经网络模块支持跨领域任务，包括自然语言处理、合成路线预测和分子生成。
		基因组研究	HGC	建立新一代基因组学平台，显著加速基因组分析过程，应用于新冠病毒研究	Clara Parabricks 基因组学软件	提高基因组分析效率，缩短分析时间助力应对突发公共卫生事件	基因组分析速度提高40倍，将原本需要20小时的任务缩短至30分钟内。
			甲骨文	推动个性化医疗和精准医学的应用，特别是在医学影像和基因组学领域	NVIDIA Clara AI平台	加快个性化医疗和精准医学的发展，提升患者护理质量	结合甲骨文的企业软件专业知识和英伟达的AI技术，实现个性化治疗的突破。
	医疗技术	手术生态系统	强生	强化强生的数字手术生态系统，实现在手术室中的实时数据分析	IGX边缘计算平台，Clara Holoscan平台	提高手术室内数据分析的实时性和精度，优化全球手术室的操作流程	在全球80%的手术室中应用，支持AI在医疗影像和疾病诊断中的应用。
		IT建设与模型训练	火山引擎机器学习平台	为企业提供基于云的高性能计算解决方案，特别是针对算力需求高的领域	云计算与NVIDIA工具	降低企业IT基础设施投入，提供高性价比的计算资源，企业快速布局市场	云原生的模型训练与推理平台，帮助企业减少前期IT投入并加速产品迭代。
				加速基因组分析、药物发现、自然语言处理和医学影像的AI模型开发	SDKs，容器	加快AI模型开发速度，简化工作流程提升医疗健康领域的研究效率	提供专为医疗健康领域设计的开发工具和容器，支持AI模型的快速启动和部署。
		基因组数据分析	序列达	使用Parabicks GPU加速工具显著提高基因组甲基化数据的分析速度	GPU加速的Parabricks框架	大幅缩短基因组甲基化数据的分析时间，确保结果的高准确性	分析速度提高20倍以上，且与开源工具的结果高度一致，适用于大规模基因组数据分析。
				利用GPU加速单细胞数据的处理，包括加载、转换和分析	RAPIDS数据科学加速库	提升单细胞数据处理的效率，加快分析加载、转换和分析的速度	使用cuDF库提高数据处理效率，实现单细胞分析速度的18-22倍提升。

医学影像AI的应用（1/2）

医学影像AI应用发展的三阶段

发展阶段	产品阶段	规模阶段	经营阶段
阶段价值	产品价值	规模价值	财务价值
参考指标	▪产品价值 ▪市场关注度	▪用户数 ▪总资产周转率	▪盈亏平衡 ▪ROA/ROC/ROE
阶段难点	▪数据 ▪算力 ▪算法	▪三类医疗器械证 ▪二类医疗器械证	▪商业模式 ▪推广路径 ▪生态路局

医学影像AI的应用场景分类

预防　诊断　治疗　健康管理

早筛　诊断　治疗　随访

科研

监管

医学影像AI的应用（2/2）

临床环节的医学影像AI应用场景

乳腺癌的诊断各环节的医学影像AI应用场景

诊断步骤	诊断方式	科室	医疗器械/系统	医学影像类型	医学影像AI应用场景
临床检查	临床表现与触诊	乳腺科	HIS、EMR		
影像学评估	乳腺超声	超声科	彩色多普勒/PACS/RIS	图片 视频	动态捕捉病灶，并对关键帧分析，自动化乳腺容积扫查。辅助检测乳腺病灶位置，并可对肿瘤良恶性进行鉴别，减少假阳性率与漏诊率
影像学评估	乳腺X线检查	放射科	钼靶/PACS/RIS	图片	自动标记钙化或其他可疑病变。基于X线的乳腺癌筛查、良恶性鉴别及分型
影像学评估	乳腺MRI	放射科	MRI/PACS/RIS	图片	基于MR检查结果，输出病灶位置、提及、形态候象、分级。3D重建
临床检查	超声引导下的乳腺活体组织病理学检查	乳腺科	微创活检设备/EMR		
病理诊断	病理学诊断分类、分级	病理科	数字病理扫描仪/LIS	图片WSI	恶性细胞筛查

药物研发AI的应用

现状	挑战	突破关键
起步阶段，AI作用主要集中在类似物识别、构效关系分析、部分物理化学性质预测等方面	在难靶先导化合物发现、多维成药性决策、First-in-class化合物发现和体内功效方面，AI突破有限	选择合适的科学问题和成药指标进行建模

传统药物研发流程

	靶点发现与验证	新化合物实体发现	活性化合物筛选	临床前研究	临床	申请及上市
目的	确定治疗干预的目标	设计合成新化学物质	测试候选化合物的生物活性，初步筛选	评估化合物的安全性和有效性	一期 二期 三期	
AI应用	靶标发现：利用机器学习分析组学数据预测药物靶标，模拟靶蛋白结构，预测小分子结合可能性	虚拟筛选与精准设计：AI快速识别活性化合物，辅助设计高选择性小分子药物 药效和ADMET预测：AI预测新化合物的药效和药代动力学特性，减少实验 合成规划：AI优化合成路径，降低成本和时间		PBPK建模：AI模拟药物在体内的分布和代谢，提升药效预测 毒性预测：AI预测毒性，辅助早期风险评估	生物标志物发现与患者分群：AI分析组学数据，发现药物响应生物标志物，精准分层患者 临床试验模拟与优化：AI模拟不同试验设计，优化样本量、疗效和风险，选择最佳方案 监测与调整：AI实时分析试验数据，监测疗效和安全性，动态调整试验设计，提高效率	

典型应用场景：小分子药物

靶点发现 → 分子生成 → 活性预测 → ADMET预测 → 晶型预测 → 剂型预测

分子生成模型：通过给定分子的各种约束来逆向生成分子
成熟应用领域：nlp、cv等
常见生成模型：AR、VAE、GAN、Flow以及EBM、diffusion model

随着计算机数据挖掘技术的不断发展,机器学习成为了计算机科学领域的一个活跃的研究方法,科学家们应用机器学习方法提高药物活性预测效率。

AI+汽车：通过Drive芯片及Omniverse平台，提升汽车自动驾驶性能及驾驶安全性，推动智能化汽车解决方案的发展

合作方向	合作领域		合作企业	内容	技术	价值	特点
AI+汽车	自动驾驶机器	边缘工智能应用的开发	U-blox	与英伟达合作，推动自动驾驶机器和边缘人工智能应用开发，提供高精度定位解决方案。	GNSS RTK接收器，IMUGNSS传感器	实现厘米级精准定位，支持自主移动机器人、无人机等的开发。	深厚的GNSS RTK解决方案经验与英伟达平台结合，成为自动驾驶开发社区的关键参与者。
		人工智能驱动信息娱乐系统	比亚迪	使用英伟达下一代车载芯片Drive Thor，提升车辆自动驾驶和数字功能，智化工厂和供应链，开发虚拟展厅。	Drive Thor	提高自动驾驶和数字功能的水平，优化工厂和供应链管理。	在车辆中引入云端游戏平台，提升了数字化用户体验。
		智能驾驶	极氪	推出搭载英伟达DRIVE Orin的豪华轿车，支持全栈智能驾驶系统和自动化操作。	DRIVE Orin片上系统	提供高速和城市道路上的智能泊车以及自动驾驶操作，提升驾驶体验和安全性。	提供Lidar +Vision Fusion 和 Pure Vision两种传感器选项，支持告诉领航辅助驾驶系统。
			理想	使用两款DRIVE Orin处理器为辅助驾驶系统AD Max提供动力，实现了高级驾驶辅助和自动泊车的功能。	DRIVE Orin处理器	实现全场景自动驾驶功能，提升驾驶安全性和便利性。	提供508 TOPS的计算能力，支持端到端的AI模型架构。
			广汽埃安昊铂	与英伟达和德赛西威合作，基于英伟达新一代芯片DRIVE Thor，开发L4级别自动驾驶软硬件系统。	DRIVE Thor	推动新一代舱驾一体和中央计算平台的落地，满足高级别自动驾驶需求。	跨国合作推动新一代自动驾驶技术在中国市场的应用和发展。
			阿尔特	与英伟达合作，推动AI赋能的汽车研发和机器人业务发展。	AI赋能技术	推动汽车研发的智能化和自动化水平提高市场竞争力。	深层次合作，结合双方核心优势，推动业务升级
			长城	基于DRIVE Orin平台开发高端智能驾驶系统Coffee Pilot，支持全场景智能导航和辅助驾驶功能。	DRIVE Orin集中计算平台	实现停车、高速和城市道路上的全场景智能驾驶，提升车辆自动化水平。	依托英伟达AI技术，提供城市导航辅助驾驶和跨层记忆泊车等高级智能驾驶功能。
			小鹏、广汽埃安旗下	扩大与英伟达的合作，应用Drive Thor技术提升自动驾驶和车载信息娱乐系统的性能。	Drive Thor	提升自动驾驶性能，增强车载信息娱乐系统的智能化体验。	通过技术合作，推动自动驾驶和车载系统的全面升级。
			小米	推出基于双DRIVE Orin配置的首款电动轿车SU7，具备高速公路驾驶功能，支持中国市场的全国导航。	DRIVE Orin配置，大型语言模型	实现全场景的自动驾驶导航，提高驾驶的安全性和舒适性。	采用先进的语言模型技术，支持无论地区或道路类型的全中国导航，具有多版本续航选择。
	数字孪生技术	-	其他	发布智能驾驶辅助系统，基于英伟达DRIVEOrin芯片，并使用Omniverse平台进行数字孪生应用开发，优化制造过程。	DRIVE Orin芯片，Omniverse平台	提高驾驶辅助系统的性能，优化制造流程，提高生产效率。	结合GPU渲染和实时协作能力，实现虚拟化设计和规划，减少实际生产中的错误和资源浪费。
		-		多家企业利用英伟达技术提升自动驾驶和智能车载系统的性能，包括极星、Z00x等。	DRIVE Orin，DRIVESim，Omniverse平台	加速自动驾驶技术的开发和应用，提升驾驶安全性和智能化体验。	通过英伟达提供的高性能计算平台和开发工具，快速推进自动驾驶汽车的市场化进程。
		渲染与模型优化	宝马	使用Omniverse平台打造虚拟数字化工厂实现实时模拟、协作和优化设计，应用于工厂的规划和运营。	Omniverse平台，GPU架构	优化工厂设计与运营，降低成本，提升生产效率。	在虚拟环境中进行工厂设计和规划，避免物理建造前的潜在问题，实现人机工程的优化。

特斯拉孟菲斯超级计算工厂“Supercluster”开始正常运行

它在一个RDMA结构上有10万个液冷H100，是世界上最强大的人工智能训练集群，用于特斯拉自动驾驶FSD的训练。

加速AI模型训练

- 英伟达的GPU擅长并行处理
- 适合深度学习和神经网络的训练

→ 有助于加速特斯拉相关AI模型的训练过程。

- 英伟达强大的计算资源

→
- 提高特斯拉在AI领域的研发效率；
- 加快新技术的推出和应用。

优化自动驾驶系统

- 配备大量的芯片

→
- 极大提升超算中心的计算能力；
- 满足特斯拉在自动驾驶算法训练、数据分析和模拟仿真等方面的需求。

- 超算中心能处理海量的自动驾驶数据
- 高效的数据分析和挖掘

→
- 为特斯拉的自动驾驶技术提供数据支持；
- 更好地理解驾驶场景和用户需求，不断优化自动驾驶系统。

- 强大的算力支持

→
- 特斯拉能够更快地迭代和优化自动驾驶算法；
- 提升自动驾驶系统的安全性和可靠性。

英伟达在小鹏汽车中的应用

英伟达在小鹏汽车中的技术供应

	应用车型	技术特点	应用效果
① XNGP 智能辅助驾驶系统	小鹏 G9 	• 双DRIVE Orin系统级芯片配置 • 每秒508万亿次算力	• 实现车辆从起点停车位到终点停车位全程使用辅助驾驶
② NVIDIA Omniverse 平台	小鹏 X9 	• 驱动的应用程序改变了复杂的 3D 工作流程 • 构建了统一的工具和数据工作流	• 实现产品设计到团队协作的优化创新 • 模拟大规模、物理精准的虚拟世界
③ NVIDIA DRIVE Orin SoC	小鹏 G6 	• 其为智能辅助驾驶系统的计算平台 • DRIVE Orin 是一款高性能的车规级系统级芯片	• 支持高速环境中 L2+ 级驾驶功能 • 可通过无线更新（OTA）不断升级 AI 功能

AI+机器人：通过Isaac及Omniverse平台，协助企业开发各类型机器人，加速推动机器人数字化转型

合作方向	合作领域		合作企业	内容	技术	价值	特点
AI+机器人	机器人的交互	交互通用场景	Apple	与英伟达合作，通过OmniverseClouds API将3D应用中交互式通用场景串流到Apple Vision Pro混合现实头显中。	Omniverse Clouds API, Graphics Delivery Network(GDN)	提供基于AR/VR的3D场景交互，增强用户体验。	通过云端传输图形数据，实现实时物理渲染，简化开发者在混合现实中的应用开发流程。
	机器人研发与训练	足球机器人与人性机器人	逐际动力	通过与英伟达合作，开发足式机器人和人性机器人，利用大规模仿真训练提升机器人的全地形移动能力和泛化能力。	NVIDIA Isaac Sim, NVIDIA PhysX, Isaac Lab	提升机器人在复杂环境中的稳定性和运动控制能力，加速通用机器人的研发进程。	利用GPU加速多物理引擎进行大规模仿真训练，显著提高机器人强化学习的效果
		工业机器人	西门子	利用英伟达Issac平台为工业机器人系统引入视觉AI，实现物品的自主拾取和包装。	Lsaac Sim, SIMATIC Robot PickAI(PRO)和SIMATIC Robot Pack AI	加速工业机器人数字化转型，减少人工干预，提升自动化水平	AI驱动的工业机器人可自主完成复杂任务，促进工业自动化和智能制造。
		自主移动机器人	比亚迪	与英伟达合作开发自主移动机器人，提出了完整的工厂物流解决方案，应用于智能制造	Isaac Sim,Isaac Perceptor, Jetson Orin	可以提高工厂生产智能化水平，优化运输流程，保障工人安全。	采用多传感器融合技水，实现复杂动态中的高精度建图和定位，提升环境适应能力
		机器人车队	富士康	基于英伟达Omniverse平台开发全仿真自主工厂，应用机器人车队的训练和部署。	Omniverse平台，Isaac平台	能够优化装配线设计流程，减少物理布局错误，提高制造效率。	在虚拟环境中进行工厂流程定义和机器人训练，减少了实际操作中的复杂性和成本。
		载物机器人	九号公司	展示与英伟达联合开发的Isaac Perceptor产品，用于自动驾驶类载物机器人开发。	Isaac Perceptor	能提供可定制的自动驾驶研发平台，支持载物机器人的开发。	适用于自动驾驶类载物机器人，支持不同应用场景下的定制化需求。

英伟达在机器人领域应用：

前瞻布局智能机器人领域

- **2018**：开启布局智能机器人领域
- **2019 2022**：从底层芯片、到计算平台、到训练与验证，开启智能机器人领域的全方位布局
- **2023**：生成式 AI+机器人”成为 NVIDIA 机器人全线产品技术演进和重塑的核心方向
- **2024.2**：成立通用具身智能体研究 GEAR团队，工作围绕多模态基础模型/通用型机器人研究/虚拟世界基础智能体/模拟与合成数据技术四个领域开展
- **2024.3**：GTC 大会上，重磅推出人形机器人通用基础模型 Project GR00T，并发布配套的新型人形机器人计算机 Jetson Thor，以及对 Isaac Lab 进行了重大更新

目标

- 围绕大模型/数据/开发平台三大核心领域展开，打造机器人底层开发生态

合作企业

- 积极对具身智能机器人本体厂商进行投资
- 与领先的人形机器人公司共同开发综合的AI 平台，这些公司包括 1X Technologies、Agility Robotics、Apptronik、波士顿动力公司、Figure AI、傅利叶智能、Sanctuary AI、宇树科技和小鹏鹏行等，加速构建机器人底层开发生态

机器人和边缘运算示例

NVIDIA Jetson™和Isaac™平台提供端对端解决方案，可供在制造、物流、医疗保健、智慧城市和零售等领域开发和部署人工智能驱动的自动化设备和边缘计算应用。

缩小仿真与真实的差距：Spot RL Researcher Kit

- Boston Dynamics、NVIDIA和AI Institute之间的合作通过NVIDIA Isaac Lab将四足机器人从虚拟世界无缝部署到现实世界

利用NVIDIA Issac Sim 4.0加强机器人工作流程

- 最新版本包括英伟达APhysX5.4、Issac Lab、新功能和各种加强

英伟达在机器人领域应用：NVIDIA 侧重给各行业提供开发平台和工具，自上而下入注机器人设计

NVIDIA 侧重给各行业提供开发平台和工具，自上而下入注机器人设计

主要提供主控芯片、开发平台和工具，可帮助开发人员构建、部署和管理机器人

开发平台

- Jetson平台专注于边缘端，AI落地
- ISAAC平台是一套完备的开发套件，包括虚拟化、可视化的工具，进行格式化设计的过程，把不同模组的模块搭在一起
- NVIDIA Jetson™ 和 Isaac™ 平台提供端对端解决方案，可供在多领域开发和部署人工智能驱动的自动化设备和边缘计算应用
- 目前Isaac和Jetson平台等正被超过120万名开发人员和10000名客户以及合作伙伴所使用

GTC大会上发布最强的AI加速卡

Blackwell GB200	Blackwell 架构的 GPU，其 AI性能可达 20 petaflops，为H100的5倍
机器人通用基础模型 GR00T	由 GR00T 驱动的机器人可通过观察人类行为理解自然语言和模仿动作，具备快速学习、协调、灵巧等特点，可以与现实世界互动。
机器人大脑计算芯 Jetson Thor	Jetson Thor是一种新的计算平台，能够执行复杂的任务，并与人和机器安全 自然地交互，旨在为 GR00T 提供动力

比亚迪电子与英伟达宣布合作开发自主移动机器人

比亚迪电子（BYDE）正使用英伟达Isaac Sim和Isaac Perceptor进行全系列自主移动机器人（AMR）的开发，并进一步提出一套完整的工厂物流解决方案

方案

比亚迪电子在AMR开发中，采用了英伟达Jetson Orin高性能人工智能系统级模块，将激光雷达、IMU、轮速计等多传感器融合技术结合，达到可在室内大面积且复杂动态环境中进行实时高精度建图和定位的强大能力。

功能

AMR无需依赖轨道或预定义路线，也无需操作员监督，借助复杂的传感器、人工智能、机器学习技术，AI可自主计算路径规划，从而理解环境并导航。

意义

可保障工人安全、降低生产成本，并优化运输流程，从而提高客户的生产智能化水平

AI+零售：通过边缘设备及开发套件，提高零售企业运营效率，推动智能化零售解决方案的发展

合作方向	合作领域		合作企业	内容	技术	价值	特点
AI+零售	智能全渠道管理	-	-	NVIDIA构建了全面的全渠道管理体系，利用生成式AI、多模态技术、NLP、计算机视觉和深度学习	生成式AI，多模态技术，NLP，Omniverse Enterprise数字孪生技术	提升零售商店布局、劳动力流程优化及客户需求预测的能力	通过模拟商店布局和优化劳动力流程，实现高效的运营管理，并通过预测客户需求来提升市场竞争力。
	智能供应链	空间虚拟副本	-	NVIDIA结合机器人技术与物流管理设备，利用自动化机器人和智能叉车等，提升仓库管理效率。	机器人技术，LMD设备，NLP，数字孪生技术	提高供应链管理效率，减少人力成本，增强供应链的灵活性和可扩展性。	使用智能技术创建仓库虚拟副本进行空间设计和流程模拟，提高包裹生命周期跟踪和最后一英里交付的效率。
		数据管理	沃尔玛	沃尔玛通过NVIDIA RAPIDS提升需求预测准确率，避免新鲜农产品浪费，减少库存过剩和降价促销。	NVIDIA RAPIDS	提高预测准确率，优化库存管理，减少食品浪费和不必要的降价。	每周运行需求预测模型，并处理超过5亿种组合，确保预测的准确性和稳健性，适应异常数据带来的挑战。
	智能商店	AI摄像头	-	NVIDIA的智能商店服务利用AI摄像头、行为分析和自动化工具优化商店运营，增强安全性，优化商品布局和营销策略。	AI摄像头，行为分析技术，热图分析，自助购物和自动结账功能	提升商店安全性，优化顾客购物体验，降低人力成本，增加商店运营效率。	实现资产保护，优化库存管理，提供缺货警报和实时警报系统，引入智能柜和路边取货服务，提供便捷的购物选择。
	智能快餐店	智能订单处理	-	NVIDIA通过智能订单处理系统优化快餐店运营效率，包括自动语音订单接收、车牌识别、自动结账和行为分析	智能订单处理系统，AI算法，车牌识别技术	提高订单处理速度和准确性，优化库存管理，减少食品浪费，增强食品安全。	结合行为分析和监控系统进行资产保护和内部安全管理，实现个性化服务和忠诚度计划，提供路线优化和外送服务。
	零售运营	自主商店平台	Nanos-tores	Nanostores提供无缝多摄像头、实时商品识别和库存跟踪，提供免结账购物体验，同时支持灵活支付方式。	O.A.S.I.S.自主商店平台	提供全年无休的免结账购物体验，降低运营成本，快速部署和重新部署	通过无收银员结账技术提供快速、便捷的购物体验，支持实时跟踪和灵活支付，适应不同环境下的需求。
	质量控制	闭环自监督	-	在铝罐生产线中，NVIDIA合作伙伴开发AI应用，使用GPU加速的AI视觉推理和边缘再训练功能，优化生产线的缺陷检测	GPU加速AI视觉推理，闭环自监督学习算法	提高生产线效率，优化缺陷检测，降低生产成本，提高产品质量。	采用闭环自监督学习算法应对生产线速度和缺陷可变性问题，支持大规模生产环境下的高效质量管理。
	分类包裹	边缘服务器与推理服务器	USPS	USPS使用NVIDIA Metropolis和EGX边缘服务器，加速包裹数据处理和视频分析，提升包裹分拣效率和处理准确度。	NVIDIA Metropolis，EGX边缘服务器，Triton推理服务器	提升包裹分拣站点效率，保证包裹安全，优化物流管理流程。	通过加速边缘服务器和实时视频分析，实现包裹的快速分类和处理，确保在大规模物流操作中的高效性和安全性。
	废料处理	自动废物分析、分类系统	Recycl-eye	Recycleye部署AI自动废物分析系统，利用NVIDIA DeepStream SDK进行视频分析，提高废物分类效率	NVIDIA DeepStream SDK, Jetson Xavier AGX	提高回收处理效率，降低处理成本，优化资源回收流程。	通过计算机视觉技术和迁移学习工具包，实现废物的高效识别、分类和分拣，系统以高帧率运行

AI+电信：通过开发套件及计算平台等，提高客户服务质量，降低运营成本，推动智能化电信解决方案的发展

合作领域		合作企业	内容	技术	价值	特点
技术支持		其他	英伟达与多家电信运营商合作（如Telenor Group、新加坡电信等），推动AI平台在电信行业的企业市场应用，并支持运营商的内部转型。	AI平台	支持电信行业的企业市场发展，推动运营商内部的数字化转型。	通过与多家电信运营商合作，将AI技术应用于电信行业的多个领域，增强行业竞争力和内部运营效率。
生成式AI	虚拟形象	新加坡电信	新加坡电信与微软和英伟达合作，利用5G网络和AI技术，创建生成式AI驱动的虚拟形象，并将AI视频分析应用于多种用例，推动企业创新和生产力提升。	5G网络，AI生成式技术，NVIDIA全栈加速计算平台	提升企业创新能力和生产力，通过5G网络和AI实现跨行业的创新应用。	将AI和5G结合，通过多接入边缘计算服务为客户提供高效的视频分析和虚拟形象解决方案，支持不同应用场景下的快速部署和扩展。
	实时语音AI	T-Mobile	T-Mobile使用NVIDIA Riva SDK支持10,000多个并发的实时语音AI应用，并开发AI客服助手Expert Assist，提升客户服务效率，减少人工客服流失率，增加销售额。	NVIDIA Riva SDK，AI客服助手	提高客户服务质量，增强用户体验，增加企业销售收入。	通过实时语音转录和自动推荐解决方案，实现更快速、更准确的客户服务响应，提升客户满意度和企业竞争力。
RAN的性能（提升）和创新		诺基亚	英伟达与诺基亚合作，改进cloud RAN解决方案，并开发AI-ready RAN，支持5G和未来6G RAN的高效运营和管理。	cloud RAN，AI-ready RAN	提升RAN网络的性能，支持未来6G RAN的发展，增强网络的灵活性和扩展性。	通过cloud RAN解决方案和AI增强功能，实现对RAN网络的高效管理和创新，支持从训练到推理的全流程AI应用。
		开发者社区	NVIDIA与开发者社区合作，使用Aerial AI无线电框架和其他工具开发和测试新的AI/ML算法，增强RAN中的AI应用，支持6G信号处理和未来无线网络的构建。	Aerial AI无线电框架，pyAerial，NVIDIA Sionna	提升RAN网络中的AI/ML应用效果，支持复杂通信系统的快速原型开发。	提供从算法探索到模型训练和推理的全套AI增强功能，支持基于Aerial CUDA加速的RAN网络的AI开发，增强5G和未来6G RAN的商业级和软件定义的云原生特性。
5G网络	数字孪生	HEAVY.AI	HEAVY.AI基于NVIDIA Omniverse平台，构建5G网络的数字孪生，结合GPU加速分析和实时地球物理映射，帮助电信公司更精准地部署和调整基站，降低成本，提高效率	NVIDIA Omniverse，GPU加速分析，HeavyRF模块	优化5G网络部署，降低运营成本，提高基站布局的准确性。	通过数字孪生和实时分析技术，在早期部署5G网络时实现更高效的RF射频传播场景测试，显著减少时间和资金投入，提高网络运营效率。

AI+娱乐：通过Omniverse平台和Merlin推荐算法等，提高用户体验，降低开发成本，推动智能化娱乐创新及解决方案的发展

合作领域		合作企业	内容	技术	价值	特点
生成式AI		联想	英伟达与联想合作推出全新的混合人工智能解决方案，提供量身定制的生成式AI。这是双方在工作站、游戏PC和高性能计算机方面合作的进一步扩展。	混合人工智能解决方案，生成式AI	扩展了AI技术在工作站和游戏PC等领域的应用，提高了产品定制化能力和市场竞争力。	通过进一步合作，双方在生成式AI领域加深了合作，推动AI在不同计算平台上的应用，实现更多的定制化服务。
实时通信		Avaya	NVIDIA Maxine与通信公司Avaya合作，推出支持实时通信的AI解决方案，提高了通信服务的质量和效率。	NVIDIA Maxine，AI实时通信	提升通信服务质量，提供更高效客户通信解决方案。	NVIDIA Maxine提供了实时通信的AI增强功能，与Avaya的合作实现了更高效、更智能的通信服务，特别是在客户服务领域。
推荐算法	推荐系统开发	陌陌	陌陌通过NVIDIA Merlin HugeCTR优化推荐算法的训练性能，采用SOK和HKV库，提升了训练效率，使精排模型训练吞吐提升5倍，整体性能提升12倍。	NVIDIA Merlin HugeCTR，SOK，HKV	提升推荐算法的训练效率，优化模型性能，提高推荐系统的准确性和用户体验。	通过优化训练流程和模型架构，陌陌在推荐算法的训练效率上实现了显著提升，显著减少了训练时间并提高了整体性能，支持更复杂的推荐系统开发。
	排名模型	SNAP	Snap使用NVIDIA Merlin和GPU提升内容排名能力，将机器学习推理的成本效率提高50%，并将服务延迟降低了2倍，支持更复杂和精确的广告和内容排名模型。	NVIDIA Merlin，GPU加速	提升广告和内容排名的准确性，降低推理成本，提高服务响应速度。	通过使用Merlin和GPU技术，Snap优化了广告和内容排名模型，提高了平台的服务效率和广告收入，增强了用户体验和平台的市场竞争力。
	广告推荐	腾讯	腾讯通过NVIDIA Merlin HugeCTR提升广告推荐系统的模型训练性能，加快模型更新频率，提高广告推荐的准确性和收入。	NVIDIA Merlin HugeCTR	提高广告推荐系统的训练速度和模型准确性，增加广告收入。	通过HugeCTR集成到广告推荐系统，腾讯能够训练更多数据，提高推荐系统的准确性，并显著提升了广告业务的整体收入。
	模型训练	美团	美团通过NVIDIA A100 GPU降低模型训练成本，同时提高训练效率，支持更复杂的模型训练，提升业务效果。	NVIDIA A100 GPU	降低模型训练成本，提高训练效率，支持更复杂的模型开发。	使用A100 GPU，美团优化了其模型训练框架，降低了成本并提高了复杂模型的训练效率，支持更高效的推荐和服务优化，提升用户交易频率和业务增长。
其他		Epic Games、Autodesk	英伟达与Epic Games、Autodesk、McNeel & Associates、Trimble Inc等合作，通过Omniverse平台支持跨行业的沉浸式、交互式和协作式体验，扩展3D内容发布能力。	Omniverse平台，3D内容发布，沉浸式体验	提供跨行业的沉浸式虚拟体验，增强3D内容发布能力，支持广泛的协作应用。	合作伙伴包括Epic Games、Autodesk等，通过Omniverse平台提供无缝连接，实现游戏、企业和视觉效果领域的前沿内容开发和协作。

英伟达在《黑神话：悟空》中的应用

英伟达对《黑神话：悟空》的技术贡献

全景光线追踪技术

技术原理

光线追踪原理
- 模拟光线在场景中的传播路径；
- 计算光线与场景中物体的交互效果；
- 从玩家视角反向追踪光线，根据光线与物体相交点的材质属性和光源位置计算出最终颜色。

全景光线追踪（路径追踪）
- 处理全方位的光线交互，包括多次反射、折射等复杂的光学现象；
- 精准模拟光线，确保无论从哪个方向观看，场景中的光线表现都能保持一致和逼真。

应用效果

逼真的的阴影、反射和全局光照
- 准确模拟整个场景中的光线属性，包括光线的传播、反射、折射等；
- 从而在所有物体上形成物理正确的阴影、反射和全局光照。

提升画质与沉浸感
- 通过全景光线追踪技术，《黑神话：悟空》中的场景更加逼真，光影变幻也更加细腻；
- 营造沉浸式视觉效果，使玩家能够更深入地融入游戏世界中，可以享受极高的帧率。

底层支撑

RTX（实时光线追踪）技术
- RTX技术基于物理学的光线追踪原理，通过模拟光线在场景的传播和交互，生成逼真的图像。

Tensor Core
- 用于深度学习，通过加速深度学习算法，提升图像处理效率。

RT Core
- 用于光线追踪计算，能够高效地处理光线追踪任务。

DLSS 3AI渲染技术

技术原理

超分辨率渲染
- 降低游戏渲染的分辨率来减少GPU的负载；
- 对低分辨率图像进行超分辨率渲染，生成图像接近原生分辨率。

帧生成技术
- 分析图像帧，并计算帧与帧之间物体和元素的运动矢量数据；
- 数据输入到卷积神经网络，生成全新的帧，提高了游戏帧率。

Reflex低延迟技术
- 通过优化CPU和GPU之间的通信，降低系统延迟，提高游戏的响应速度。

应用效果

性能提升
- 降低渲染分辨率并用AI算法超分辨率渲染和帧生成，在保持高画质的同时实现更高的帧率；
- 更高分辨率和画质，流畅游戏体验。

兼容性
- DLSS 3技术得到了广泛的硬件和软件支持。
- 几乎所有GeForce RTX 40系列GPU 都能够支持DLSS 3技术。

画质改善
- 通过先进的AI算法和Tensor Core的加速能力，高质量图像接近原生分辨率。

底层支撑

Ada Lovelace架构
- 集成更强大的Tensor Core和新的光流加速器；
- 为DLSS 3提供了强大的底层支持，使得该技术能够在保持高画质的同时显著提升游戏性能。

Tensor Core
- 用于深度学习，通过加速深度学习算法，提升图像处理效率。

AI+高能计算：通过GPU及其加速极端平台等，在服务器提高计算效率，降低开发成本和周期，推动智能化数据管理和数据中心管理的发展

合作领域	合作企业	内容	技术	价值	特点
因子挖掘	DolphinDB	DolphinDB与NVIDIA合作，推出基于NVIDIA RAPIDS的Shark异构计算平台，大幅提升了因子挖掘算法的运行效率，实现计算性能提升2-10倍。	NVIDIA RAPIDS，cuDF，Shark异构计算平台	提高了计算效率，降低开发成本，缩短了开发周期。	利用NVIDIA的GPU算力平台，DolphinDB显著优化了因子挖掘的计算效率，并通过池化显存管理技术有效利用显存资源，降低了存储和计算的开销。
数据处理	AT&T	AT&T使用NVIDIA RAPIDS加速器处理PB级海量数据，通过GPU助力的服务器测试，5小时内处理了2.8万亿行移动数据，速度提升3.3倍，成本降低60%。	NVIDIA RAPIDS，GPU加速器	显著提升了数据处理速度，降低了运营成本。	通过NVIDIA RAPIDS加速器，AT&T大幅提升了海量数据的处理效率，为电信行业的高效数据管理提供了重要支持。
算力一体化	道客	DaoCloud与NVIDIA合作推出d.run算力一体化解决方案，基于NVIDIA GPU构建高性能计算集群，支持AI模型训练和推理，提供高效、稳定的算力支持，加速企业智能化转型。	NVIDIA GPU，MIG技术，云原生调度算法	提供灵活可靠的智算中心解决方案，支持AI创新和应用。	通过硬件与软件的深度整合，d.run实现了高效的资源管理和分配，显著提升了智算中心的构建速度和运营管理效率，为企业带来了持续稳定的高效计算体验。
数字孪生技术	施耐德电气	施耐德电气与NVIDIA合作，优化数据中心基础设施，推进边缘AI和数字孪生技术的变革，推出新的参考设计，提升数据中心的性能、可扩展性与能效。	AI技术，数字孪生，数据中心基础设施	优化数据中心的运营管理，提升可扩展性和能效。	施耐德电气通过与NVIDIA合作，重新定义了数据中心内部的AI部署和运营标准，为行业树立了新的基准，实现了更高效、更智能的数据中心管理。

AI+交通：通过Jetson及EGX平台等，在信号灯管理、机场等拥堵路段、停车场管理等场景，提升交通管理效率和运营效率，推动智能化交通解决方案的发展

合作领域		合作企业	内容	技术	价值	特点
路面管理	交通管理	NoTraffic	NVIDIA Metropolis合作伙伴NoTraffic通过基于AI的交通管理平台优化交通流量，改善交通信号灯效率，减少延迟，提高道路安全性。	NVIDIA Metropolis，Jetson边缘AI设备，AI传感器	提升交通管理效率，减少温室气体排放，节省燃油和维护成本。	通过边缘计算和AI传感器实现实时道路使用者检测和分类，优化交通信号灯，提高城市交通管理的效率，显著减少车辆延误时间和温室气体排放。
	交通枢纽管理	Assaia	Assaia利用图像识别算法和NVIDIA Jetson AGX Xavier模组对机场周转运营进行优化管理，提升机场准点率，增强安全性，降低运营成本	NVIDIA Jetson AGX Xavier，图像识别算法	提升机场运营效率和安全性，减少运营成本，支持可持续发展。	通过AI算法将视频流转化为结构化数据，实时全面了解机场周转活动，提高准点率和运营效率，增强了机场的整体管理能力。
停车管理		DataFromSky	DataFromSky与NVIDIA合作，为停车场提供AI监控解决方案，通过摄像头监控车位占用情况，支持移动支付和导航服务，优化停车管理。	NVIDIA Metropolis，NVIDIA EGX服务器平台，DeepStream SDK	提供高效的停车管理解决方案，减少停车场管理成本，提升用户体验。	使用AI技术和摄像头监控系统，实时分析停车区域的占用情况，简化系统管理，并与移动应用程序连接，为通勤者提供了理想的停车解决方案，同时支持与执法机构对接共享信息。

AI+安全：通过EGX及Jetson平台等，在医院、校园、机场等公共领域，提升运营效率，减少事故发生，推动智能化安全解决方案的发展

合作领域		合作企业	内容	技术	价值	特点
运营管理	医院运营安全	Artisight	Artisight利用AI提升医院手术室生产力，优化工作流程，并通过摄像头网络检测COVID-19患者，提高医院整体运营效率和安全性。	NVIDIA Metropolis，AI热敏摄像头，EGX平台	提升医院运营效率，减少感染风险和人力需求，优化患者护理体验。	通过AI技术和摄像头网络，Artisight在医院环境中实现了高效的患者筛查和工作流程优化，显著提高了医院的生产力和安全性，特别是在疫情期间的应用效果显著。
场所管理	工作场所安全	Helin Data	Helin Data利用AI技术提升海上石油钻塔的安全性和效率，通过边缘计算实时分析工人的位置，并在发生危险时发出警报，避免事故发生。	NVIDIA Metropolis，Jetson边缘AI平台	提高石油钻塔的安全性和工作效率，减少事故发生，优化人力资源配置。	使用边缘计算技术，Helin Data的系统能够实时精确定位并分析工人的位置和行为，迅速发出警报，避免事故发生，提高了高风险工作环境的安全性和效率。
	校园安全	Icetana	Icetana为加拿大皇家山大学提供校园安全解决方案，通过GPU加速的AI视频分析识别异常活动，及时发出警报，防止损失和安全事故。	NVIDIA EGX平台，GPU加速AI视频分析	提升校园安全管理水平，及时处理异常事件，降低安全风险。	通过实时视频分析，Icetana系统能够在校园内快速识别和处理异常活动，显著提升了校园安全性，特别是在疫情期间有效应对了安全管理的挑战。
	机场安全	Ipsotek	Ipsotek在雅加达苏加诺-哈达国际机场部署基于AI的视频分析解决方案，用于监控和管理机场的安全与运营，包括监控无人认领行李和车辆、旅客计数、边界保护等。	NVIDIA EGX平台，AI视频分析，PTZ摄像头	提高机场的安全管理和运营效率，减少事故发生，优化旅客体验。	通过AI技术和视频分析，Ipsotek解决方案能够实时监控机场内的各种活动，提供关键的安全和运营信息，支持机场在高流量环境下实现高效管理和安全运营。
	人群安全	印度特兰加纳州当局	在Medaram Jathara节日期间，印度特兰加纳州当局使用Awiros Crowd Estimation应用通过摄像头实时测量人群密度，预防踩踏事故，保障活动安全顺利进行。	NVIDIA EGX平台，DeepStream SDK	保障大型活动的安全，实时监测人群密度，预防事故发生。	通过AI技术和摄像头网络，实时监测和分析人群密度，提供及时的安全警报，有效防止大规模人群聚集活动中的踩踏和其他安全事故，确保节日活动的顺利进行。

保障体系

-研发保障

-营销保障

-供应链保障

-管理保障

-投资合作

一、研发保障。英伟达建立庞大的研发团队，采用跨学科合作和三团队-两季度的组织模式持续推动技术创新，确保行业领先地位

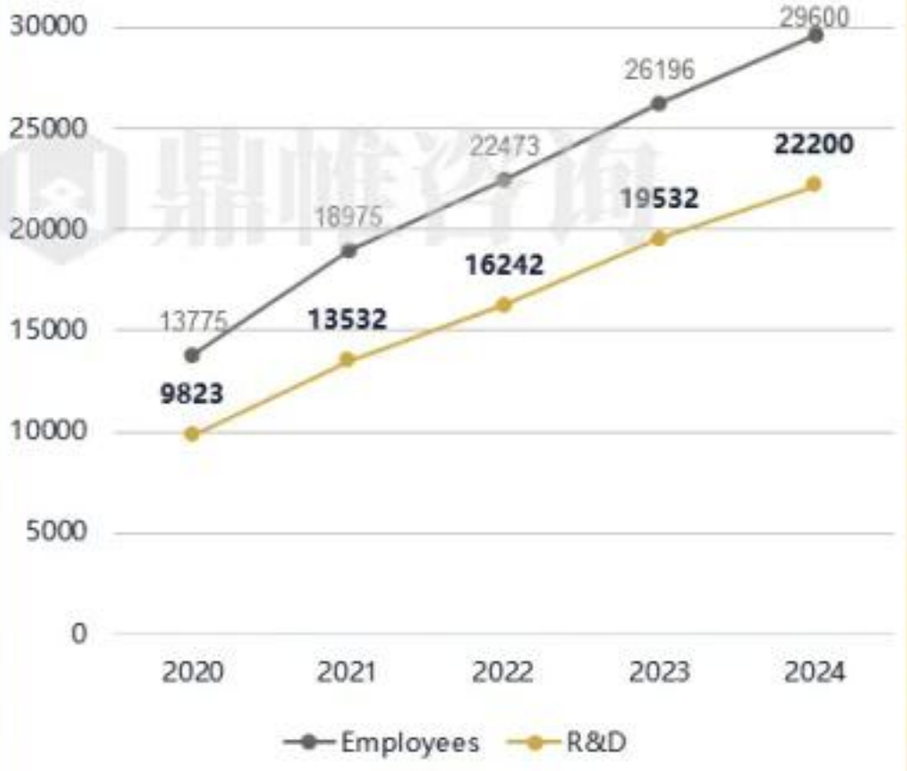
组织模式

形成良好研发生态
保持技术性能双双领先

实现产品快速迭代
在价格高于竞争对手的情况下仍占据绝对的市场份额

研发人员方面，英伟达建立了一支涵盖软件工程、硬件工程、超大规模集成电路工程、工艺工程、架构和算法团队的高质量研发队伍。截至2024年，英伟达研发人员达22200人，占员工总数的75%

英伟达研发人员数量



英伟达研发人员构成

英伟达研发团队与学术界紧密合作，致力于推动AI前沿技术的创新和发展

总体架构

内部研发团队

前沿研发实验室

学术研发中心

1. Aspire Lab UC Berkeley
   渴望实验室加州大学伯克利分校
2. Center for Advanced Electronics Through Machine Learning (CAEML)
   机器学习先进电子学中心（CAEML）
3. Consortium for Vision and Virtual Reality (CV2R)
   视觉与虚拟现实联盟（CV2R）
4. Stanford Center for Image Systems Engineering (SCIEN)
   斯坦福图像系统工程中心（SCIEN）

学术合作伙伴

1. KAYVON FATAHALIAN 凯冯·法塔哈连
   CMU 卡耐基梅隆大学
   研究重点
   计算机密集型应用（例如交互式计算机图形）设计高效并行系统
2. VIVIENNE SZE 施薇薇
   MIT 麻省理工学院
   研究重点
   算法、架构、和电路的联合设计，以建构节能和高性能系统
3. DAVID BROOKS 大卫布鲁克斯
   HARVARD 哈佛大学
   David 在计算机体系结构和 VLSI 设计方法领域做出了贡献
4. GU-YEON WEI 魏具妍
   CMU 卡耐基梅隆大学
   在计算机体系结构和 VLSI 设计方法领域做出了贡献

英伟达成立内部研发团队和前沿研发实验室，共同推动GPU技术在多个前沿学科领域的创新与落地

总体架构

学术研发中心

四位学术合作伙伴

内部研发团队
(内部基础供应、内部应用需求)

基础研发团队
提供底层技术，GPU本身内容

- 电路研究小组 → 用于芯片之间的信号传输
- VLSI小组 → 致力于研发新设计方法 (深度学习的加速器工作等)
- 架构组
- 网络组
- 系统编程组
- GPU存储系统组

应用研发团队
推动需求，开发应用程序，引导客户购买更多GPU

- 图感知学习组
- 机器人技术组
- 人工智能算法组
- 自动驾驶汽车组
- 应用深度学习组
- 以及各地的研究实验室

工作方式
跨组合作

- 300多名博士研发
- 5个不同的人工智能组

前沿研发实验室

研究实验室
NVIDIA Research 通过 26 个学科的项目、论文和活动参与，将开创性的研究变为现实。探索我们的实验室

3D 深度学习	AI 介导的现实和交互	应用研究
自动驾驶汽车	对话式 AI	深邃想象力
动态愿景和学习	通才具身代理	高保真物理
学习和感知	感知、行动和推理	实时图形
机器人	台湾研究实验室	

研究领域
NVIDIA 调查通常涵盖多个研究领域;研究领域列表允许以一种方法来组织我们的出版物、人员和项目

3D 深度学习	人工智能和机器学习	应用研究
算法和数值方法	电路和 VLSI 设计	计算摄影和成像
计算机体系结构	计算机图形学	计算机视觉
电子竞技	高性能计算	人机交互
超大规模图形	应用感知	医疗
感知、行动、推理和深度学习	编程语言、系统和工具	实时渲染...

分布：**美国各大核心城市**
如多伦多、西雅图等

主导：一位教授

研究：针对**前沿重点方向**进行科学研究
出版较多论文及专利

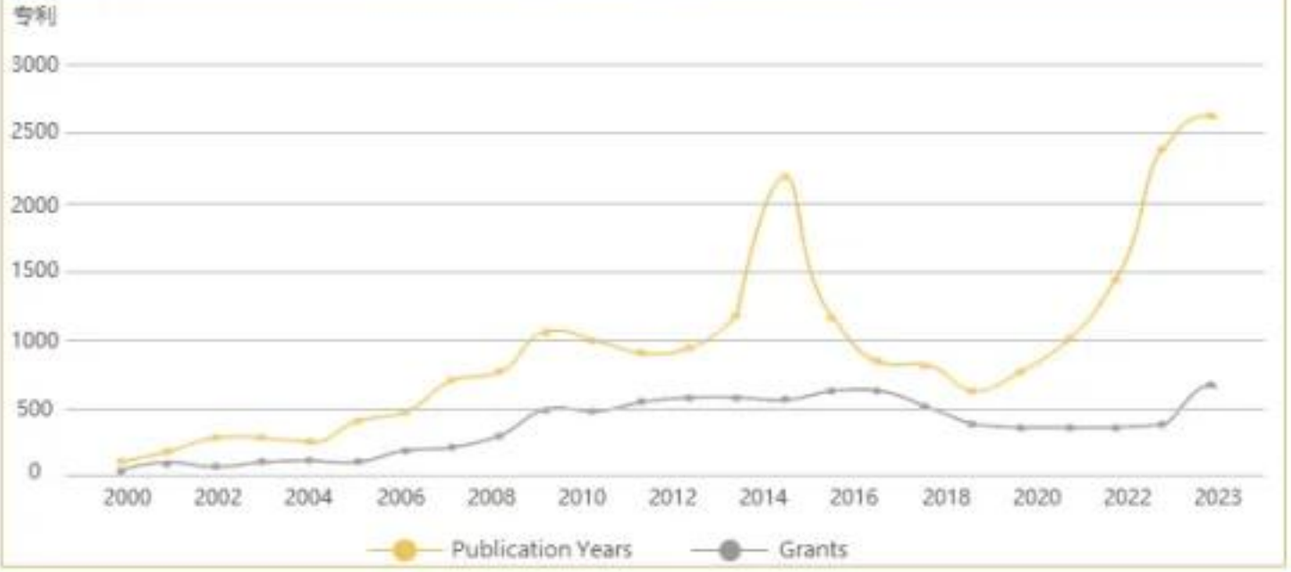
英伟达专利布局遍布全球，以美国和中国为主，涵盖图像显示和数据中心等多个技术领域，且自2000年以来专利数量持续增长

专利布局

英伟达全球各国专利申请量
中国排在除美国外第二位
占14.83%

英伟达中国市场前景可期

专利数量



专利领域

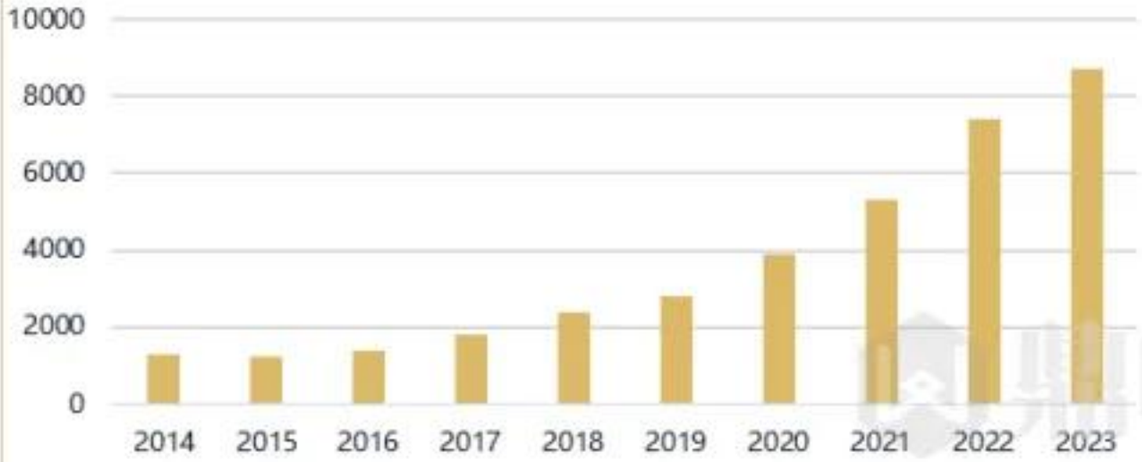
图像显示	数据中心冷却	视频编码
数据追踪	编译器	无线通信系统
微处理器	存储器	SRAM ...

专利关键词

显示图像，超级方案，游戏输出管线	功能电路、测试、管芯、缺陷、灵活	管线 各向异性缓冲器 计算机图形，样本
管线、着色器，制品平台，片段	配置接口、模组、冷却、按钮、数据信号	动态的、无线网络传输功率、视频数据视频编码器
独立管线、着色器寄存器、请求	网络、TCP、安全性协议，网关	有源电路，散热器 半导体管 电路板耗散

英伟达的研发支出持续增长，研发费用率逐步降低，2023年研发费用率为14%

英伟达研发投入情况（单位：百万美元）



注：此处英伟达2023年的研发投入是指2024财年(即2023年1月30日-2024年1月28日)的研发投入，前几年情况以此类推。

英伟达研发费用率

纳斯达克十大公司的研发支出

公司	2023年的研发支出（亿美元）	R&D 收入百分比
amazon	856	15%
Alphabet	391	13%
Meta	385	29%
Apple	299	7%
Microsoft	283	13%
NVIDIA.	86	14%
BROADCOM	51	15%
ASML	44	14%
TESLA	4	4%
PEPSICO	8	1%

*过去 12 个月，截至 2023 年 12 月 31 日。Nvidia 和 Broadcom 的数据截至 2024 年 1 月 29 日。

二、营销保障。会议营销：英伟达建立了以GTC大会为核心，包含系列行业和区域会议、线上和线下会议，以及合作伙伴会议的会议矩阵，塑造整体品牌形象

核心会议

GTC大会GPU Technology Conference
一年一度线上线下面对开发者、研究者、新创企业的大会

定位	技术发布平台	品牌内容传播平台	营销获客平台	技术培训平台
功能	分享技术趋势、发布最新产品、宣传应用场景			

系列会议

行业大会

1 发布最新技术、建立合作关系、推广产品

可视化业务方向

- SIGGRAPH 2024大会面向专家和前沿初创公司推广解决方案、培训英伟达产品，分享图形、研究、通用场景描述 (OpenUSD) 和 AI 领域推动下一代发现的最新突破

游戏业务方向

- CES大会（国际消费类电子产品展览会）发布GeForce RTX 40 Super系列产品、RTX笔记本电脑、游戏应用、机器人应用等

区域大会

2 培训开发者技术
分享应用场景实践

- 杭州2023年云栖大会云计算产业链年度盛会展示交流云计算以及相关领域的产业趋势、技术创新、关键产品及最新应用

AI Summit

3 分享AI应用场景

- 专注于AI话题会议
- 提供超过30场关于生成式AI、产业数字化、机器人等领域的演讲和现场技术展示

线上研讨会

4 面向开发人员培训、学术及技术研讨

- 数据分析类
- 自动驾驶类
- 深度学习类
- 开发者类
- 嵌入式类
- 医疗类
- 专业可视化类
- 电信类
- 机器人类

合作伙伴会议

6 发布合作成果，分享成功案例，推广软硬件产品

Databricks - Data + AI 峰会	SAP Sapphire 大会	ServiceNow - Knowledge 24大会
数据中心业务合作伙伴 Databricks 利用 NVIDIA 全栈解决方案加速生成式 AI 应用，其开源模型 DBRX 作为一项 NIM 微服务托管在API 目录	数据中心及可视化合作伙伴 发布搭载 AI Enterprise 软件的生成式 AI 合作成果，其为 生成式 AI 助手 Joule 提供了 SAP 咨询功能和 ABAP 开发者功能	数据中心业务合作伙伴 宣布合作计划旨在企业运营中合作开发强大的生成式 AI 能力，并结合 NVIDIA ACE 和 NIM 微服务提供高效且具有同理心的支持

GTC大会成为行业标志性会议论坛，并形成行业专家、学者、技术应用者、创业者、赞助商、消费者等众多参与者在内的论坛生态，共同输出行业前瞻思想与科技前沿动态

历史悠久

- 起源于 2009 年加利福尼亚州圣何塞
- 最初关注 GPU 解决计算挑战的潜力
- 目前已举办15年
- 2024年GTC大会于3月18日至31日举办

亮眼数据

	GTC2022	GTC2024
注册人数	22万人	30万人
报道篇数	1.2万篇	2.9万篇
合作商数	50余家	300余家

举办过程

以创始人演讲为重，后续包含产品展览、百场演讲、会谈、研讨及网络交流活动

黄仁勋主题演讲
(开幕日上午)

- 专家演讲
- 技术实验室
- 圆桌会议
- 技术讲座
- 网络研讨会
- 培训课程

演讲者有图灵奖获得者、大学教授、开发者、创业者、技术应用者等多元化群体

赞助商分类

1. Diamond Elite 钻石精英：亚马逊、谷歌、微软、戴尔、甲骨文等
2. Diamond 钻石：AIVRES、DDN、Lambda等12家
3. Platinum 铂：华硕、IBM、三星、Crusoe等11家
4. Gold 金：埃森哲、Ansys、CDW等14家
5. Silver 银：思科、Couldflare、vmware等21家
6. Bronze 青铜：Amphenol、CoolIT、DXC等10家
7. Media 媒体：daily ai、GDN、tryolabs、unite.ai
8. Exhibitor 参展商：七维科技、奔驰、米斯特拉尔AI等众多

个人参与者

参与门票费

- 早鸟票在大会前几个月销售
- 通常1000到2000美元
- 常规票在会前几个星期销售
- 通常2000到3000美元
- 学生票价格在500美元左右

培训报名费

- 全天实战培训：运用云端完全配置的GPU实验室，完成培训测试后，获得英伟达培训证书
- 具体参考价格：
- GTC特惠单人：1099元
- 团购：766元满3人
- 日常价格：3500元/人

品牌营销与内容营销：英伟达以聚焦各细分场景下丰富专业的技术内容，支撑活跃的用户生态系统，形成高度忠诚的消费者与粉丝群体，在此基础上通过英伟达企业品牌与黄仁勋个人品牌的打造形成品牌号召力

NVIDIA

品牌营销

个人品牌

个人魅力

粉丝影响力

企业品牌

品牌视觉符号

消费者辨识度与存在感

NVIDIA GEFORCE RTX

内容营销

个人洞察力
专题采访
主题演讲

博客
白皮书
研究报告

定价策略：英伟达针对企业级、消费级不同产品及其用户利益点，制定不同价格策略，并以捆绑销售方式与其供应商实现双赢

	产品	客户	用户收益	价格
企业级	Quadro和Tesla系列GPU	数据中心、专业设计师、科研人员	企业效率提升和成本节省	在数千到数万美元之间
消费级	GeForce系列	游戏玩家、内容创作者、追求最新技术的消费者	满足游戏和内容创作需求	从几百美元到高端的几千美元不等
捆绑销售	DGX、NIC、交换机、光学器件等（通过多源计划、制造自己的AI芯片计划等）	基础设施提供商	通过获得更多资源分配，短时间内业务实现爆发式增长	搭配产品本身的价格

社区营销策略：英伟达通过开发者计划、Omniverse社区、企业平台顾问计划构建全面创新链生态系统

开发者计划

开发者 + 游戏玩家 + 爱好者

热情倡导者 + 贡献者

忠诚的社区

提供资源：论坛；最新SDK组织或工具；教育材料

获取资讯：产品新闻或更新提示；新产品率先体验通知；相关项目或活动邀请；折扣信息；Bug报告

Omniverse社区（3D创作者和开发者组成的全球性网络）

资源中心（内容）	创作激励（激励对象）	直播（主播与内容）	用户组会议（可吸取内容）
海量文档	使用英伟达产品创作高质量作品	英伟达工程师；合作伙伴	新版本幕后故事
教程	对社区做出贡献	开发者；社区成员	合作伙伴亮点
培训		分享最新技术	新资源
…			部分录像（向重点客户提供）

企业平台顾问计划

顾问团队（丰富技术经验）	英伟达团队（领域专业知识）	加速计算解决方案的未来
顾问加入要求	英伟达提供服务	英伟达平台战略
1 有业内认可的影响力任务	为顾问们提供最新的信息、支持和工具	扩展专业网络 → 获得加速计算知识
2 在英伟达关键技术方面有专业技能和经验	推动英伟达平台和技术的广泛应用	塑造英伟达平台战略
3 经验丰富的专业人士、学者、领导者		

社交媒体营销策略：通过账号矩阵建设与KOL合作，同各行各业广大用户群体形成密切互动与深度关联，打造各平台用户社区

KOL合作与影响者营销

合作对象	合作账号	发布内容	影响方式
游戏玩家（应用端）	• Linus Sebastian • Austin Evans • Jayztwocents	• 发布游戏预览 • 在线回答问题 • 发布播客 • 提供有关其创新如何增强客户体验的见解	分享第一手经验，强调英伟达GPU如何改变其游戏冒险体验
数据科学家（技术端）	• Andrew Ng • Fei-Fei Li • Yann LeCun		扩大科学家影响力，同时通过展现功能与技术提高品牌知名度
主播（通用）	• Shroud • Ninja • TimTheTatman		投放付费广告并推广主播内容，影响其超200万的粉丝

国内社交媒体布局

博客
- 中国官方博客
- 开发者博客

论坛
- NVIDIA开发者
- NVIDIA CSDN社区

微信公众号
- 英伟达
- 英伟达企业解决方案
- 英伟达开发者
- 英伟达网络
- 英伟达GeForce

微博
- 英伟达官方账号
- 英伟达Enterprise
- 英伟达GeForce
- 英伟达网络

视频类
- 英伟达抖音
- 英伟达西瓜视频
- 英伟达哔哩哔哩
- GeForce抖音
- GeForce哔哩哔哩

资讯类
- 英伟达今日头条
- 英伟达知乎
- 英伟达喜马拉雅
- GeForce知乎
- GeForce哔哩哔哩

海外社交媒体布局

博客Blog
The Official NVIDIA Blog
NVIDIA Developer Blog
NVIDIA GeForce News
SHIELD Blog

Facebook
NVIDIA
NVIDIA AI
NVIDIA Data Center
NVIDIA Game Developer
NVIDIA GeForce
NVIDIA GeForce NOW
NVIDIA GTC
NVIDIA Networking
NVIDIA Robotics
NVIDIA SHIELD
NVIDIA Studio

Instagram
NVIDIA
NVIDIA AI
NVIDIA Developer
NVIDIA GeForce
NVIDIA Omniverse
NVIDIA Robotics
NVIDIA Studio
NVIDIA University Recruiting

领英LinkedIn
NVIDIA
NVIDIA AI
NVIDIA Data Center
NVIDIA Design and Visualization
NVIDIA DRIVE
NVIDIA GTC
NVIDIA Healthcare
NVIDIA Networking
NVIDIA Omniverse
NVIDIA Robotics
NVIDIA University Recruiting
NVIDIA Virtual GPU

Twitch
NVIDIA GeForce　NVIDIA Omniverse

X
@NVIDIA – NVIDIA, GPU industry news
@NVIDIAAI – Deep learning, AI news
@NVIDIAAIDev - Developer news about AI
@NVIDIACC – NVIDIA customer support
@NVIDIADC – data center and supercomputing news
@NVIDIADesign – Quadro, pro graphics news
@NVIDIADeveloper – all things developer with NVIDIA
@NVIDIADRIVE – autonomous driving news
@NVIDIAEmbedded – Jetson, embedded solutions
@NVIDIAGameDev – game developer news
@NVIDIAGeForce – GeForce, gaming news
@NVIDIAGFN – NVIDIA GeForce NOW
@NVIDIAGTC – our premier tech event series
@NVIDIAHealth – healthcare industry news
@NVIDIAHPCDev – HPC developer news
@NVIDIANetworking - NVIDIA Networking news
@NVIDIAOmniverse - Omniverse platform news
@NVIDIAPSIRT – Product Security Incident Response Team
@NVIDIASHIELD – SHIELD news and GeForce NOW
@NVIDIAStudio - NVIDIA Studio news
@NVIDIAVirt - NVIDIA vGPU virtualization solutions for enterprise

YouTube
NVIDIA
NVIDIA Developer
NVIDIA GeForce
NVIDIA Omniverse
NVIDIA Studio

其他
NVIDIA on SlideShare
NVIDIA on Threads
NVIDIA AI on Threads
NVIDIA Developer on Threads
NVIDIA Omniverse on Threads
NVIDIA Studio on Threads
NVIDIA GeForce on TikTok

论坛
NVIDIA Forums
NVIDIA GRID Forums
GeForce Graphics Cards
NVIDIA Developer Forums
Customer Support, Known Issues, and FAQs

社会责任营销：通过AI治理、绿色低碳生产，以及参与气候和环境议题，获得市场认可与ESG基金投资

ESG+CSR营销

超过100家ESG基金将其视为投资组合中的优选
市场对其环境治理和长期可持续性策略的认可

企业社会责任

1

降低产品生命周期的环境足迹

确保从设计、生产到产品使用和回收的每一个环节都高效节能，
通过节能产品和回收计划来最大限度地减少对环境的影响

2

优化制造流程，减少废弃物产生，推动绿色供应链管理

Quadro系列显卡：在提升性能的同时，
也在能耗效率上做出了显著改进，
符合许多工作站和数据中心对能效的高要求

3

采用可再生能源

在其数据中心和其他设施中
部署太阳能和风能项目
参与购电协议（PPAs），购
买绿色电力证书，支持清洁
能源的发展

参与更大范围的社会责任

4

参与气候风险评估

公开其气候变化相关数据，遵守科学碳目标倡议（SBTi），
致力于实现与《巴黎协定》相一致的减排目标

5

注重技术创新对环境的正面影响

产品被广泛应用于气候模拟、能源管理等环保领域，
助力科学研究和行业实践更加高效地应对环境问题

技术基座

6

关注安全及隐私保护

发布了AI治理框架，旨在促进透明度、公平性与责任性，减少偏见和误用风险。

英伟达针对不同市场定位实施差异化开发战略，实现AI生态的地域多元化，同时持续降低开发人员使用软件系统的门槛，以此推动全球业务的快速扩张

培育区域 AI 生态系统

- 保持北美主导地位，发力欧亚快速增长，拓展拉美和非洲新兴市场，减少对单一市场地区的依赖，实现地域多元化

全球市场	市场定位	产品布局	开发方式	结果
北美	领先市场	全产品线	投资前沿企业 全渠道发展	55.1%
欧洲	主力市场	通过专业级Quadro和Tesla产品进入	投资初创公司 与欧盟数字主权保持一致	其余地区 15.7%
亚洲	增长市场	定制GPU芯片	投资初创公司 与欧盟数字主权保持一致	中国占比 29.2%
拉丁美洲	潜力市场	基础数据产品	与教育科研合作 建立HPC中心	-
非洲	潜在市场	GPU及数据中心解决方案	创造就业机会 与联合国合作	-

推广全面软件系统

- 以LaunchPad为代表的软件系统降低AI发展的准入门槛，使开发人员和研究人员能够释放英伟达硬件的全部潜力，最终推动全球扩张

LaunchPad 提供端到端解决方案	▪ 提供端到端的AI基础设施软件栈，并成为其他企业AI平台的供应商 ▪ 此计划的服务覆盖范围扩大至全球9个地区，帮助企业快速开发和部署AI应用
GeForce Now 基于云的解决方案	▪ 考虑到某些地区物理基础设施的局限性，允许使用功能较弱的机器的用户获得高性能计算能力，提供云端解决方案助力AI开发
Nvidia Clara Holoviz AI大众化	▪ 通过Clara Holoviz软件使研究人员和科学家能够利用 AI 进行科学可视化，这使得跨各个学科的强大 AI 工具的访问民主化
保证全球品牌一致性	▪ 考虑到文化差异性，全球软硬件产品的推广需要翻译人员对多语言的深入理解，提升全球文件的可访问性和用户采用率

中国市场：英伟达在美国对中国市场的出口管制下，推出特供简配版芯片挽回市场但效果有限，华为积极布局处理器领域成为英伟达在中国市场的最大竞争对手

英伟达推出特供芯片需求疲软

- 政府管控

——美国出口管制，数据中心业务收入大幅下降

- 低需求

——中国特定产品H20，性能与价格不匹配

- 结果

——预计5年内中国云计算公司需求将从80%降至50%-60%

中国AI处理器供应商奋起直追

- 扩容国内市场空间

——产品改进和升级空间变大，缩小美国同行差距

- 本地化优势明显

——“降级”芯片性能优势不及国内替代品

- 华为AI

——2023年，AI处理器市场英伟达占比90%，华为占比6%，未来有望上市

华为——英伟达在中国最大竞争对手

目标

- 建立国产计算基地
- 拥有全栈软硬件AI解决方案——Ascend AI平台

突破

- Ascend 910B：训练大预言模型的效率是A100的80%，其他测试高出A100的20%
- Ascend 910C：超过7万份订单，价值约20亿美元

挑战

- 产能有限：中芯国际的7nm产能有限，短期内无法兼顾智能手机和AI芯片
- 容量有限：910B系列的有源AI内核远少于一代
- 研发速度有限：910B性能仅一代的1.2倍

华为昇腾AI：最强大的人工智能计算平台，支撑企业实现智能转型

行业应用

能源,金融, 公共, 交通, 通信，教育等更多行业应用

应用使能

- ModelArts
- HiAI Service
- 第三方平台
- MindX DL 深度学习使能
- MindX Edge 智能边缘使能
- ModelZoo 优选模型库
- MindX SDK 行业SDK

人工智能框架

- MindSpore：匹配昇腾AI处理器算力的全场景深度学习框架
- 第三方框架 TensorFlow and PyTorch：可基于第三方框架开发的模型进行二次开发、训练与推理

异构计算架构

- CANN：统一异构计算架构，释放昇腾硬件澎湃算力

Atlas系列硬件

设备端 | 边缘端 | 云端

Atlas

全流程开发工具箱 MindStudio

管理运维工具 Fusion Director/SmartKit

昇腾社区 hiascend.com

三、供应链保障。人工智能革命正在改变数据中心的供应链，对更高带宽内存（HBM)的需求、CoWoS（2.5D封装)容量的限制成为AI算力的真正瓶颈

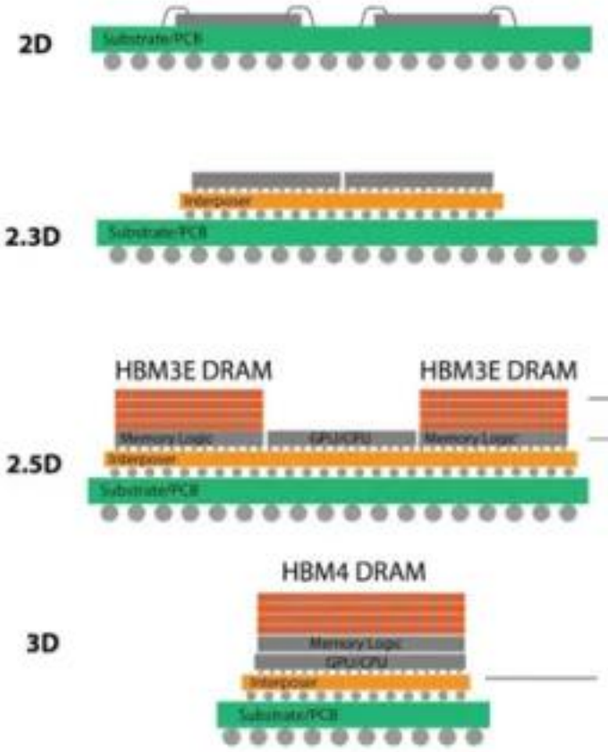
人工智能对更高带宽的需求

- 随着GPU计算能力的提高，需要更快的内存访问速度（存储器更靠近GPU）和更多的内存（更多数量）来提供必要的计算优势

人工智能对供应链数据中心部分的变革

1. 从标准封装 DRAM 过渡到高带宽内存（HBM）
2. 采用新的2.5D封装技术（台积电的CoWoS）

封装技术发展（数据中心处理的GPU封装）



Build-up Substrate

传统2D封装方法：将芯片安装在基板上，用键合线连接焊盘

7nm Silicon and below

2.3D封装方法：通过翻转芯片并将其安装在interposer（通常是硅片）上，使芯片更加接近

Stacked DRAM Die
Memory Logic Silicon

2.5D封装方法：图形处理器两侧是由基本内存逻辑芯片控制的 DRAM 芯片堆叠

NVIDIA.

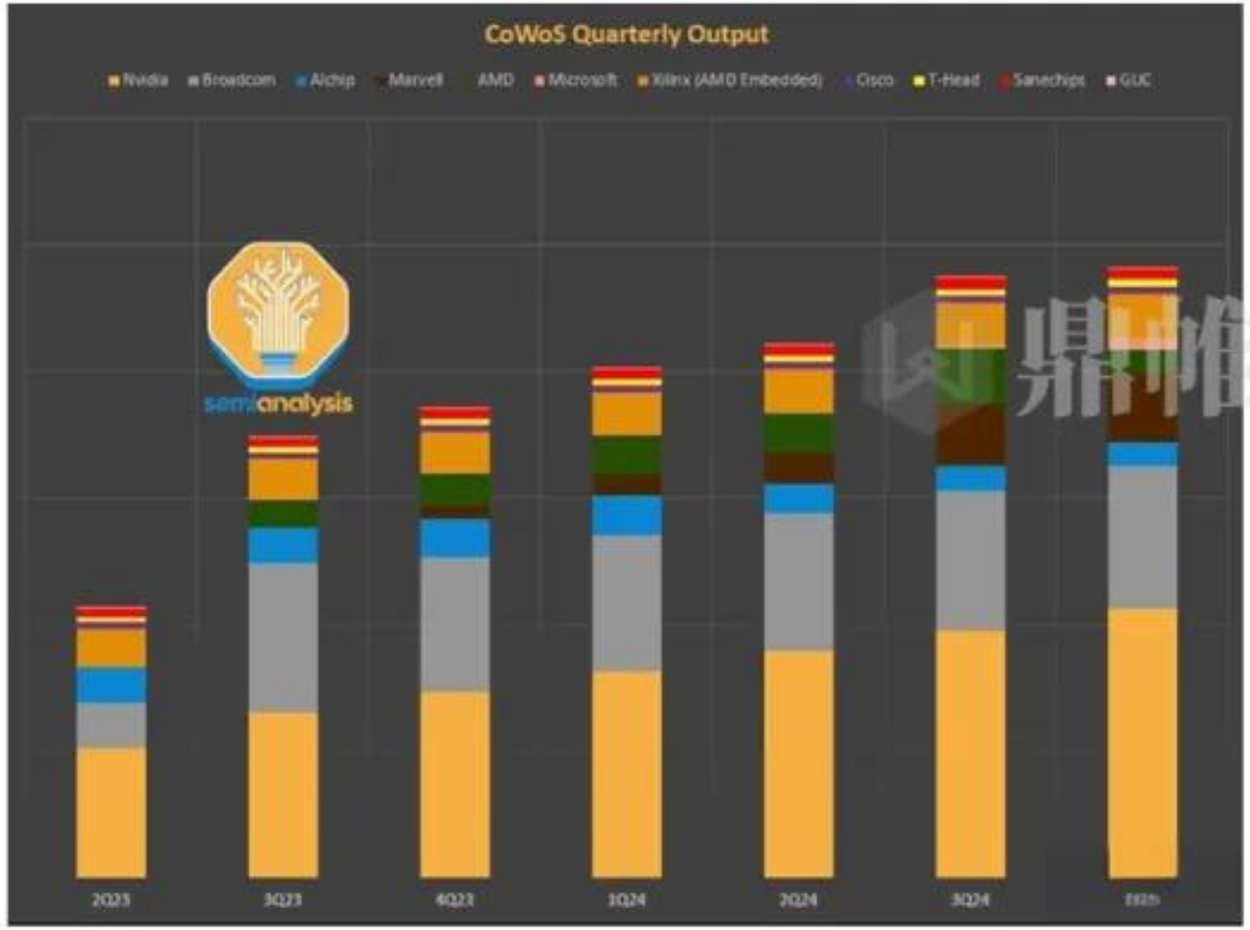
Silicon Interposer
Organic Interposer
RDL Redistribution Layers

3D封装方法：直接使用硅穿孔来连结上下不同芯片的电子讯号，以直接将存储器或其他芯片垂直堆叠在上面

注：CoWoS（2.5D封装）是提供最高互连密度和最大封装尺寸的主流封装技术，成本也相对合理，是目前高性能计算领域唯一选择

新的封装技术正在极大地改变供应链，凭借激进的供应策略，收购了台积电大部分CoWoS供应，同时抢占SK海力士、三星和美光HBM等GPU上游组件的大部分供应，其它竞争对手的供应链被进一步挤占

英伟达对供应链的掌控：已经收购了台积电大部分CoWoS供应



通过愿意承诺不可取消的订单甚至提前支付来确保巨大的供应

- Nvidia有111.5亿美元的采购承诺、产能义务和库存义务
- Nvidia还有额外的38.1亿美元的预付款供应协议。没有其他供应商能够接近这个数字，因此他们无法参与正在发生的疯狂购买潮

大订单抢占GPU上游组件的大部分供应

- 抢占SK海力士、三星和美光HBM等GPU上游组件的大部分供应，向所有 3 家 HBM 供应商下了非常大的订单，并且正在挤占除 Broadcom/Google 之外的其他所有人的供应

对待犹豫不决的供应商“胡萝卜加大棒”的管理方式

- 对英伟达的要求犹豫不决的供应商通常会得到胡萝卜加大棒的对待；一方面，他们可以从英伟达获得看似难以想象的订单，另一方面他们面临着被设计出英伟达现有供应链的问题

英伟达充分利用了数据中心对人工智能的无限需求，通过利用新的封装技术（CoWoS）和提供卓越的AI服务器系统使其在供应链中占据主导地位

传统服务器的供应链模式

- 在传统模式下，服务器卡制造商从供应商那里单独采购所有组件，处理、内存和服务器公司之间的竞争使云公司的定价受到控制

人工智能服务器的供应链模式

- 处理器公司控制了供应链：英伟达™产量巨大，由于其所有技术和组件都处于领先地位，英伟达将网络组件、内存和其他组件封装到单个系统中，从而使商品远离购买者，并能比内存公司获取更多价值。

这种商业模式只有在英伟达的人工智能服务器优于所有其他系统的情况下才能奏效

供应链全景图：在供应链上游HBM是国内企业的关键卡脖子环节

	主要部件	产品示意图	供应商企业	与英伟达关系	市场地位	技术特点
上游	HBM（高带宽存储器）		SK海力士	第一大供应商	HBM市场占有率50%	全球唯一一家能够大规模生产HBM3芯片
			三星	主要供应商	HBM市场占有率30%	最大容量的12堆栈HBM3E
			美光	主要供应商	HBM市场占有率10%	为H200提供8个堆栈的24 GB HBM3E
			华海诚科	间接供应商	国内GMC认证供应商	HBM上游高端材料GMC
			雅克科技	间接供应商	SK海力士和三星供应	为HBM提供逻辑芯片
	PCB（印刷电路板）		沪电股份	认证供应商	1个算力板获得认证	高多层板领先技术
			胜宏科技	核心唯一供应商	深度绑定全球前六大算力巨头	高阶HDI、高频高速PCB等AI服务器产品
			景旺电子	主要供应商	全球前十PCB制造商	高密度互连（HDI）技术成熟
	SSD（固态硬盘）		SK海力士	核心供应商	SSD市场占有率13%+	行业最高性能的基于PCB01的消费端产品
			三星	主要供应商	SSD市场占有率50%	PCIe5.0数据中心专用SSD产品PM9D3a
			美光	主要供应商	SSD市场占有率15%	量产9400SSD
	散热模块		维谛技术	唯一液冷散热供应商	市场占有率第一	与英伟达专家团队历时共研新型制冷方案
			精研科技	大陆唯一供应商	MIM领域龙头企业	专门配合英伟达研发散热方案
			奇鋐	主力供应商	全球领先解决方案商	AI3DVC主力产品
	铜缆		安费诺	核心供应商	全球高速铜缆连接龙头企业	供应GB200铜缆和连接器并负责整个生产

金色字体为非大陆供应商和卡脖子环节

供应链全景图：下游先进制程的芯片制造技术和封装技术是国内企业的卡脖子环节

	主要部件	产品示意图	供应商企业	与英伟达关系	市场地位	技术特点
上游	CPO（光模块）		天孚通信	重要供应商	光纤连接龙头	全球唯一一家能够大规模生产HBM3芯片
			中际旭创	独家设计生产供应商	HBM市场占有率30%	最大容量的12堆栈HBM3E
			罗伯特科	长期供应商	光伏自动化设备龙头	深度合作1.6T光模块工艺研发
			新易盛	间接供应商	国内GMC认证供应商	拥有LPO方案通过认证
	互联接口		自有研发部门	-	-	PCIe Switch、NVLink、NVSwitch实现了极高带宽和能耗的互联性能
	以太网芯片		自有研发部门	-	-	Mellanox ConnectX-5 网卡先进硬件分流技术提升技术设施效率
	交换机		自有研发部门	-	-	收购交换机企业Mellanox和Cumulus自研Spectrum-X800以太网交换机
中游	OEM/ODM服务器代工		工业富联	核心生产供应商	全球AI服务器芯片基板最大供应商	独家设计生产交付HPC/H100/1800
			纬创	A100主要供应商	全球第七大ODM厂商	多品种小批量生产
			超微电脑	AI服务器主要供应商	与英伟达深度绑定	机柜产能5000台/年
	芯片制造		台积电	核心供应商	市场占有率第一	先进的4nm、7nm工艺及CoWoS封装技术
下游	封装		三星	主要供应商	存储市场龙头企业	成立专门团队拿下2.5D封装订单
	测试		博杰股份	主要供应商	光电测试检测龙头	成熟在线功能测试FCT/ICT
			和林微纳	主要供应商	芯片测试探针龙头	产能500万件/月
	分销		神州数码	分销合作伙伴	国内IT分销龙头	深度合作华为、微软等IT厂商，提供定制化数字解决方案
			中电港	国内授权分销商	规模最大元器件分销商	电子元器件供应链资源丰富

金色字体为非大陆供应商和卡脖子环节

四、管理保障。创始人兼CEO黄仁勋是工程专业出身，同时兼有销售岗位经历，至今已执掌英伟达30年以上，是硅谷任职时间最长的CEO，并作为GPU发明者带领英伟达持续突破创新

黄仁勋

深刻的大众印象

- 老顽童
- 硅谷最狂华人
- 皮夹克
- 黑色衬衫
- 深色牛仔裤
- NVIDIA总裁、首席执行官、董事会成员

多面性背景

- 高级领导和运营经验
- 金融/金融社区
- 新兴技术和商业模式
- 监管、法律和风险管理
- 市场营销、传播与品牌管理
- 工业与技术
- 治理与上市公司董事会
- 人力资本管理
- 多样性

履历：

1984 俄勒冈州立大学电气工程学士学位和斯坦福大学电气工程硕士学位

1983-1985 AMD芯片工程师

1985-1993 LSI Logic芯片设计转销售，因业绩出色被提拔部门经理，走上管理岗位

1993年与Sun公司两位年轻工程师共同创立英伟达，任 CEO

获得多种荣誉、称号及奖项

- 《财富》杂志(Fortune) 2017年度最佳企业家
- 《时代周刊》(Time)：2021年度最具影响力的人物之一
- 《经济学人》和Brand Finance：全球最佳CEO
- 美国国家工程院院士
- 台湾国立交通大学、国立台湾大学、俄勒冈州立大学名誉博士学位
- 荣获半导体行业协会最高荣誉——罗伯特·诺伊斯奖
- 荣获IEEE创始人奖章
- 被称为科技行业最受尊敬的高管之一
- 荣获张忠谋博士杰出领导奖

英伟达的汇报关系：向黄仁勋汇报的直接下属超过60人，涵盖各业务条线、各职能部门以及大量的高级技术人员

黄仁勋
创始人，总裁，执行董事　每款产品的首席产品官

管理层

29 位核心人物

- 联合创始人兼CEO黄仁勋
- 22 位直接下属
- 设计用于游戏 PC 和数据中心服务器的 Nvidia 图形处理芯片的高级硬件工程师
- 负责软件产品开发（例如 CUDA 编程语言）的人员

许多工程领导者已经在公司工作了20多年

英伟达汇报关系

Chris Malachowsky
创始人，英伟达研究员

Dwight Diercks
高级软件工程师

Jeff Fisher
Geforce高级副总裁

Colette Kress
执行副总裁，首席财务主管

Jay Puri
全球业务执行副总裁

Debora Shoquist
运营执行副总裁

Tim Teter
执行副总裁，总法律顾问，秘书

Jonah Alben
GPU工程 执行副总裁

Brian Kelleher
硬件工程高级副总裁

Gary Hicok
汽车硬件和系统高级副总裁

Rev Lebaredian
宇宙与模拟技术副总裁

Manuvir Das
企业计算负责人

Ian Buck
副总裁，超大规模和高性能运算总经理

Sonu Nayyar
高级副总裁，首席信息官

Michael Kagan
首席技术官

Ronnie.Vasishta
Telecom副总裁

Deepu Talla
副总裁，嵌入式边缘计算总经理

Bill Dally
首席科学家，研究部高级副总裁

Joe Greco
Advanced Technology Group（ATG）高级副总裁

John Spitzer
开发人员和性能技术副总裁

Shelly Cerio
人力资源高级副总裁

Greg Estes
公司市场和研发项目副总裁

Amit Krig
软件和NIC产品线副总裁

.....

英伟达组织结构：有着极其强大的技术部门、科学研究部门以及基于产品生态建设的运营部门

英伟达（黄仁勋）的管理理念：没有层级、没有汇报、没有计划（1/3）

特点1：组织极致扁平化，直接下属多达60名

发言

“我其实想要一家更小的公司，而不是更大的公司”

“我拥有许多直属下属、不进行一对一谈话，使公司变得扁平，信息迅速传播，员工得到赋权，这使得我不必进行一对一谈话变得可能”

理念

组织极致扁平化、避免等级沟通：他的直接下属多达60名，但避免进行1对1会议，取而代之的是小组讨论，确保从高层开始，每个人（对信息的掌握）都处于同一条线上

特征	金字塔结构	扁平结构
层级数量	多层级，层级分明	层级较少，组织更平坦
决策速度	较慢，需要多级审批	较快，决策过程简化
沟通效率	间接，容易出现信息失真和延误	直接，信息流通快速
员工参与	受限，通常只限于上级指示	高，员工可以直接参与决策
适用场景	需要严格执行命令和控制的环境	创新和快速适应市场变化的环境
企业文化	通常较为封闭和严格	开放和协作，鼓励创意和自主

特点2：赋权式管理—无需上报，自行决定产品进度

发言

“我们在公司的位置应该取决于我们分析复杂问题、引领他人取得成就、鼓舞和赋权他人以及为他人提供支持的能力，管理团队存在的意义就是服务于公司所有其他员工”

理念

推行赋权式管理，将决策权下放给各团队，这种充分授权不仅加速了决策过程，也大大提高了团队的工作热情和创造力

黄仁勋鼓励，英伟达的员工能够在一个无需频繁上报的环境中，更快地响应市场变化和客户需求。

类别	赋权式管理特点
自主权与创新	• 团队能够根据自己的专业判断和市场需求，选择最佳的设计方案和技术路线
专业能力的信任	• 相信团队成员的专业知识和经验足以指导他们做出正确的决策
快速迭代与对应市场	• 由于团队能够自主决定实施路径，这使得产品开发过程更加灵活和迅速
跨功能协作	• 在这种管理模式下，工程团队往往需要与市场营销、客户服务等其他部门紧密合作
结果导向	• 团队的绩效评估基于他们能否成功交付高质量的产品，而不是单纯遵循过程和规定

英伟达（黄仁勋）的管理理念：没有层级、没有汇报、没有计划（2/3）

特点3：随机了解公司状态，摒弃会议汇报

发言

"我不看任何的状态报告，因为这种报告在你得到它的时候就没有价值了，它几乎不再提供任何信息，它们已经被提炼过..."

"如果你发送电子邮件，并将其命名为'最重要的五件事'...无论你的最重要的五件事是什么，是你观察到了什么...这都是重要的信息，将其发给我，我都会阅读"

理念

黄仁勋摒弃了传统的定期状态报告方式，他认为这种方式过于僵化，无法真实反映公司的实际状况，他采用了一种更为动态和随机的方法——鼓励每位员工随时通过电子邮件向他反馈自己当前最关注的五个问题或想法

类别	描述
摒弃传统状态报告	• 认为定期状态报告可能导致信息滞后和过滤，不足以反映公司的即时状况和挑战
鼓励即时反馈	• 鼓励员工通过电子邮件及时向他直报关注的五个最重要的问题或想法
日常的信息回馈（Top5 things）	• 每天清晨浏览约100份员工的电子邮件，直接从基层获得反馈
动态管理和决策制定	• 通过及时信息流通，灵活和时效的做出决策，快速调整策略和资源分配
培养开放和透明的文化	• 通过开放的沟通策略，增强员工的归属感和积极性，建立开放和透明的企业文化
高效的问题解决	• 快速了解和响应公司问题，提高问题解决的效率

特点4：不给工作建议，充分信任团队能力

发言

"我的管理团队中没有一个人会来找我寻求工作建议，他们自己就做得很好，他们不需要我的管理。"

理念

- 黄仁勋领导方式非常独特的是，他从不给团队成员提供具体的工作建议
- 在黄仁勋明确看来，领导者的主要角色不是提供具体的业务建议，而是创造一个能让团队成员自主工作和创新的环境。领导者应该专注于指导方向、设定目标和清除障碍，而非日常的微观管理。这种方式体现了黄仁勋对团队的信任，放手赋权，极大地激发了团队成员的自主性和创造力

英伟达（黄仁勋）的管理理念：没有层级、没有汇报、没有计划（2/3）

特点5：市场是有生命的、会呼吸的，摒弃传统战略规划

发言

“战略不是文字，战略是行动，战略不是我说的，而是他们做的。我理解每个人在做什么真的很重要”

“我们没有定期规划系统，因为世界是一个生动而富有生机的事物。我们只是持续计划，我们没有5年计划，也没有1年计划”

“我们不做这些巨大的五年计划。我认为五年计划对技术来说简直是可怕的，这简直荒谬。”

理念

英伟达遵循的是一体化战略管理逻辑，不制定战略计划即不需要按照计划按部就班执行，当市场的不确定性达到一定程度时，计划越详细越容易妨碍团队创造力的发挥。英伟达的做法就是依托敏捷的组织系统不断根据市场状况随时重新评估已有战略

传统企业战略管理逻辑 VS **一体化战略管理逻辑**

特点6：公司会议向所有员工开放：透明共享

发言

“战略方向和重要议题应当公开讨论，而不是少数人开会讨论决定的。如果公司已经设定了一个战略方向，我们会向公司员工发送邮件，公开给所有员工，而不是只让少数人了解。”

“每次需要发表意见时，总是选择在公众场合进行，经常是即席发言，并鼓励组织内各级人员提供反馈和参与讨论”

理念

在黄仁勋扁平化管理体系下，英伟达成为一家少有的透明沟通的大公司，强调信息的透明度和共享。他消除了许多企业常见的信息壁垒。在他的倡导下，英伟达会议向所有人开放

透明共享机制特点

全员共享战略方向	• 如果有重要的战略方向，应该告知所有相关人员而不是仅仅几个人。这种做法能够确保每个人都在同一页面上，并理解公司的目标和方向 • 通过向全体员工公开战略方向，可以激发员工的参与感和归属感，从而促进团队的协作和创新
反馈和持续改进	• 在公布战略方向之后，会从员工那里收集反馈。这种互动不仅有助于识别可能的问题和缺陷，还能让员工感到他们的意见被重视，从而增加他们对策略的承诺。基于收到的反馈，战略可以进行必要的调整和完善，使其更加有效
充分授权和敏捷性	• 在这种管理模式下，员工被授权访问必要的信息，这不仅提高了工作效率，还使得公司能够灵活应对变化。充分的授权确保员工在没有过多监督的情况下能独立完成工作，这样可以缩短项目时间，提高响应速度

英伟达五大企业价值观：秉承创新精神，凝聚全球英才，以团队力量和灵活快速的行动，不断提升行业标准，勇敢面对挑战，从错误中学习，向着制造客户满意产品的目标不懈努力

创新

梦想远大，从小事做起。敢于冒险，快速学习

- 我们制造让客户满意的产品，提高行业标准。我们鼓励员工在第一原则而非共识的指导下进行创新
- 我们知道，在探索的道路上会有很多错误。我们预测并避免可能出现的错误。我们接受、学习并分享出现的错误。这让我们能够发明出世界甚至不知道自己需要的东西，并以此发明未来

知识诚实

追求真理，从错误中学习，分享经验

- 我们以最高的道德标准开展工作。我们力求准确地认识自己，了解自己的能力——承认自己的弱点，从错误中吸取教训
- 对现实最敏锐的理解能改进我们的工作。找出错误的根源不是为了指责，而是为了学习和不断改进。我们言出必行，并勇于付诸行动

速度与灵活

学习、适应、塑造世界

- 我们时刻保持警惕，不断学习，调整方向，以适应新的现实，从而以惊人的速度创造出突破性的产品
- 没有政治，也没有等级制度妨碍我们创造未来

卓越与决心

保持最高标准

- 我们在全球范围内聘用才华横溢且有决心有所作为的人才。我们挑战自我，努力做到最好
- 我们衡量自己的标准不是与竞争对手比较，而是与完美对比——我们称之为光速测试。如果付出长时间的努力是值得的，我们不会因此而气馁。我们在进行一场漫长的游戏

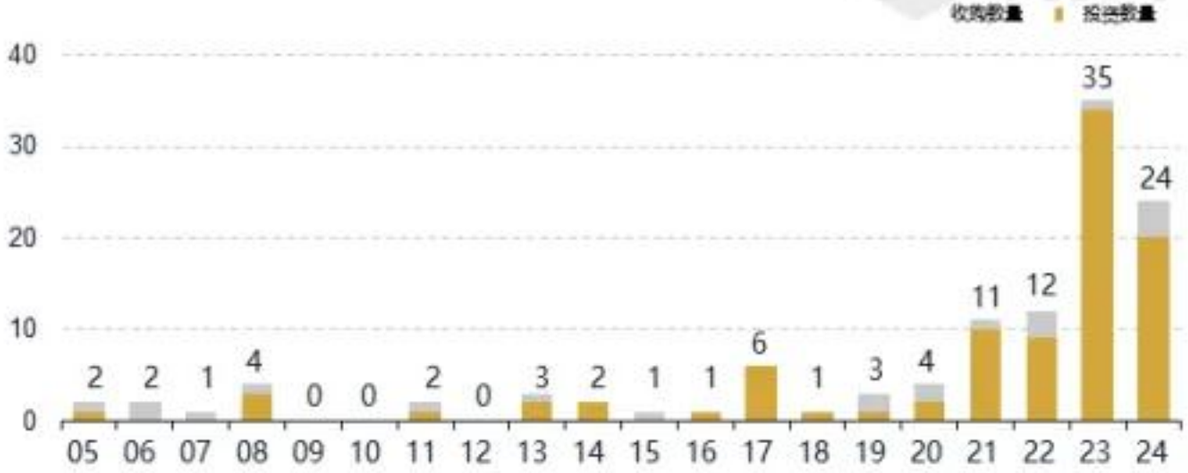
一个团队

做对公司最有利的事

- 我们营造透明、开放和信息共享的环境。这种环境能够激励我们的员工，使他们能够作为一个统一的团队共同工作。
- 我们公开直接地表达不同意见，因为冲突对解决分歧、改善想法和达成共识至关重要。我们注重实质，而非风格。将公司利益置于个人利益之上，我们就能更轻松地实现英伟达的愿景

五、投资合作。英伟达在2023年投资及收购35家公司，聚焦于人工智能、云计算、生物科技等十大投资方向，赋能主营及新兴业务发展

2005年至2024年投资及收购数量



代表性投资

Figure

- 该公司在2024年发布Figure01标志着人形机器人技术的重大突破
- 该公司获得微软、OpenAI、英伟达等知名机构支持，在B轮融资中筹集了6.75亿美元，估值达26亿美元
- 此前，公司获得了7000万美元的首轮融资，由Parkway Venture Capital领投

CoreWeave

- 作为人工智能算力提供商，通过快速创新主导AI技术，提供的计算能力比传统云提供商便宜80%
- 公司采用无限带宽(InfiniBand)技术构建数据中心，以满足AI工作需求

代表性收购

人工智能和机器学习

收购 OmniML、Excelero、Parabricks 和 Oski Technology，引入了更易访问的模型训练软件解决方案、软件定义存储、AI驱动的安全分析及机器人技术

数据中心和高性能计算解决方案

收购获得Mellanox 的互连技术、Cumulus Networks 的网络软件专长和 Bright Computing 的 HPC 管理工具

游戏和图形创新

收购AGEIA 的 PhysX 软件和 Mental Images 的 3D 图形技术，提升在游戏机和 3D 建模行业的产品竞争力

英伟达创立以来在人工智能、生物医药、机器人、云计算与数据等十大行业总计投资91家公司，收购27家公司

1 人工智能 (AI)

生成式AI

Cohere
Fireworks AI
Mistral AI
AI21 Labs
GPU.Net
Runway
Perplexity AI
Luma AI
Inflection AI
Deepgram
SoundHound

AI应用服务

Kore.ai
Essential AI
Imbue
Adept AI
Deepailytics Pvt. Ltd.
ABEJA
Utilidata

AI基础设施

Deci AI
Run:AI
OmniML
Scale AI
Togther AI
Weights & Biases
CentML
BP-FLAC

AI数据分析

Deep Instinct
Fastdata
HEAVY.AI
IFDAQ
Hugging Face
BP-FLAC

2 生物科技与医疗

生物技术

CytoReason
Accure Health
Artisight
Zebra Medical

药物发现

Lambic Therapeutics
Superluminal Medicines
Recursion Pharmaceuticals

临床试验

Inceptive
Activ Surgical

3 机器人

自动化机器人

Bright Machines

人形机器人

Figure

送货机器人

Serve Robotics

制造业机器人

READY Robotics

4 云计算与数据

数据平台

Brev.dev
WEKA
BlazingDB

高性能计算

Parabricks
Bright Computing
PGI
CoreWeave
Rescale
Saturn Cloud

数据平台

Excelero
Storage SwiftStack
Cohesity
Databricks
VAST Data

5 网络

网络管理

Mellanox Technologies
Cumulus Networks
Icera
Aarna Networks
Arrus

网络安全

Cypienta
Deep Instinct

6 软件与硬件

硬件

Enfabrica
Ayar Labs

软件

Shoreline.io
iReady
Vengo AI
Youvize
Rocketick

7 游戏

TransGaming
AGEIA Technologies
MediaQ
3Dfx Interactive
mental images
Ubitus
Real Time Gamers
OUYA

8 可视化

PortalPlayer
Hybrid Graphics
Animatico
Exluna
RayScale
Modviz
ULi Electronics
MotionDSP
Right Hemisphere

9 自动驾驶

Deepmap
Waabi
Wayve
WeRide
Halo
Foretellix
Tusimple
Almotive

10 其他

商业服务

glocali.se

先进材料

3E Nano

气候解决方案

PassiveLogic

保险解决方案

Namic Group Inc.

时尚行业数据分析

IFDAQ

金色字体为收购企业

英伟达通过初创投资项目对初创企业提供技术赋能、市场支持、专业培训、合作对接四大赋能支持，助力初创企业加速成长和创新

初创投资项目
Inception AI Startup Program

整体介绍

项目概述

- 提供GPU等核心技术赋能及风险投资支持
- Inception VC Alliance免费的全球计划于2016 年启动，通过风险投资联盟与风险投资家合作，为风险投资公司网络提供优势，将初创公司与潜在投资者连结

项目目标

- 旨在通过尖端技术、与风险投资家和其他建设者建立联系的机会
- 通过获得英伟达的最新技术资源和专业知识，帮助初创公司更快地建立、发展和扩大规模

发展现状

- 全球已经有超过10,000家创业公司加入该项目
- 中国区已超过1500 家，来自全国 60 余个城市
- 横跨医疗、媒体与娱乐、自主机器、自动驾驶、网络、5G 与电信、制造业、元宇宙、边缘计算、IT 服务行业等超过 30 个行业领域

四大赋能支持

1 技术赋能

大量软硬件支持

- 提供英伟达Jetson系列硬件开发套件，如Jetson Nano和Jetson Xavier NX等
- 提供英伟达JetPack SDK和Isaac ROS GEMs等软件

折扣及信息共享

- 通过英伟达云服务合作伙伴获取GPU云折扣及其他大量产品优惠价格
- 定期发送会员福利信息及技术速递

2 市场支持

- 获得英伟达市场部门提供的宣传支持
- 免费优先参与英伟达初创企业每年展示会
- 英伟达GTC大会免费展位、演讲及海报展示等机会
- 会员成功故事免费宣传，AI博客录制邀请及宣传

3 专业知识

- 英伟达深度学习培训中心DLI课程优惠及AI培训班优惠
- 为初创团队成员提供门户网站访问权限以及社区交流平台
- 无限制访问英伟达开发者论坛

4 合作对接

投融资对接

- 投资需求对接、闭门路演等英伟达创投联盟提供支持
- 联合国内外知名风投机构、创业孵化器、创业加速器、行业合作伙伴及科技创业媒体，打造生态系统

商业拓展

- 获得潜在客户及资本对接机会
- 通过参与各类会议接触广泛合作伙伴和行业资源，助力技术创新和商业拓展

英伟达面向各行业企业开放13个合作伙伴类型以及提供3大合作级别

合作伙伴类型

云合作伙伴	在云或托管服务模型中向利用英伟达产品的最终用户客户提供托管软件和硬件服务的合作伙伴
数据中心提供商	]用于在全球托管英伟达DGX™服务器的高密度数据中心设施、互连基础设施和最先进的冷却技术
经销商	被授权向英伟达产品、基于英伟达的解决方案和英伟达技术的经销商分销英伟达产品的合作伙伴
服务交付合作伙伴 云服务	其核心能力是云业务和技术咨询及服务的合作伙伴，可提供创新解决方案来解决客户的业务挑战
OEM 代加工	在以自己的品牌制造和转售的平台中使用英伟达产品、基于英伟达的解决方案和英伟达技术的合作伙伴
服务交付合作伙伴	教育服务拥有现有教育服务组织并希望正规化或扩大其培训和认证服务的合作伙伴
解决方案提供商 (VAR)	专注于英伟达产品、基于英伟达的解决方案和英伟达技术的增值转售的合作伙伴
服务交付合作伙伴 专业服务	拥有现有专业服务组织并希望正式化或扩大对基于英伟达的解决方案、平台和英伟达技术的使用的合作伙伴
解决方案集成 合作伙伴	专注于英伟达产品、基于英伟达的解决方案和英伟达技术的增值集成和转售的合作伙伴

合作伙伴类型

解决方案顾问	顾问为寻求实施基于英伟达的解决方案或技术的客户提供咨询服务和专家建议的合作伙伴
全球系统集成商	专门从事包含英伟达产品和技术的解决方案的规划、设计、实施和管理的合作伙伴
解决方案顾问 存储合作伙伴	与英伟达一起设计存储解决方案和联合参考架构的合作伙伴，以提供高性能计算和人工智能 (AI) 解决方案
独立软件供应商	开发、营销和销售专为商业和企业组织设计的英伟达加速计算和软件优化应用程序的合作伙伴

合作伙伴级别

Registered 注册级	注册合作伙伴是某些 NPN 合作伙伴类型的入口。这些合作伙伴投资于英伟达销售和技术培训，并有权使用产品、销售和营销工具
Preferred 首选级	首选合作伙伴投资于与英伟达建立更深入的关系。这些合作伙伴致力于实现培训、收入和其他目标
Elite 精英级	精英合作伙伴代表了与英伟达的最深层次的合作伙伴关系，并展现了对合作伙伴关系的最高水平的承诺

英伟达为战略合作伙伴提供长期技术赋能，联合研发项目，同时重点合作公司也是英伟达的大型客户

与AWS的合作内容

1. 云AI 超级计算机
首款具有 NVIDIA Grace Hopper Superchip 和 AWS UltraCluster 的计算机

2. NVIDIA DGX Cloud
首款采用 NVIDIA GH200 NVL32 的 NVIDIA DGX Cloud

3. Project Ceiba
世界上最快的GPU驱动的AI超级计算机，用于 NVIDIA AI 研发和定制模型开发

4. 全新 Amazon EC2 云服务器
由 NVIDIA GH200、H200、L40S 和 L4 GPU 提供支持

5. AWS 上的 NVIDIA 软件
NeMo LLM 框架、NeMo Retriever 和 BioNeMo

与Microsoft的合作内容

英伟达：提供 GPU 加速的 AI
允许 Microsoft 云用户能够访问 Nvidia 基于 GPU 的服务，提供提供 GPU 加速的 AI

+

Microsoft：Azure 云平台
Microsoft Azure作为骨干网，提供强大的基础设施支持

增强云平台的AI功能
Nvidia的GPU集成到 Microsoft 流行的 Azure 云平台中，使企业能够远程访问强大的工作站，而无需高端本地硬件

与Tesla的合作内容

开创人工智能驱动的电动汽车
有效地利用 Nvidia 的硬件功能（Driven PX AI 平台），推动从高级驾驶员辅助服务（ADAS）转变为成熟的自动驾驶解决方案

与IBM的合作内容

英伟达优势：并行计算应用程序
对于同时处理大量数据至关重要

+

IBM优势：响应迅速的服务器系统
创建响应迅速的服务器系统方面的熟练程度

利用Nvidia GPU 计算能力的先进数据中心
前所未有的速度提升+同时保持高水平的准确性

与Google的合作内容

为机器学习社区提供强大的 AI 基础设施
Google Cloud 和 NVIDIA 之间深化的合作伙伴关系侧重于通过提供有效构建、扩展和管理生成式 AI 应用程序所需的技术和基础设施来支持机器学习社区

英伟达扩展另一大重点战略合作伙伴，在全球市场积极寻求政府及国家间合作

美国：技术投资美国国家科学基金会 （NSF）

NVIDIA 为国家 AI 研究资源 （NAIRR） 试点计划做出贡献，通过 3000 万美元的技术投资促进 AI 包容性，该计划旨在为人工智能工具和资源提供更广泛的获取渠道，促进人工智能研究社区的创新和包容性

日本：人工智能伙伴关系计划-推动和筑波大学之间合作

NVIDIA 投资 2500 万美元，支持大学之间的人工智能研究合作，推动下一代人工智能技术的发展。这项耗资 1.1 亿美元的计划涉及华盛顿大学和筑波大学之间的研究合作，强调了国际合作在推进人工智能研究方面的重要性

印尼：投资2亿美元在印尼建人工智能中心（Indonesia AI Nation）

NVIDIA计划投入2亿美元，在印度尼西亚建立一座全新的人工智能中心，并已与印尼电信巨头Indosat Ooredoo Hutchison达成合作意向，共同推进这一项目，Indosat已宣布计划将英伟达的最新芯片技术集成到其基础设施中，旨在推动印尼进入人工智能和技术进步的新纪元

英国：领投2000万美元战略投资Charm Therapeutics

2023年5月15日，CHARM宣布完成由英伟达领投的2000万美元战略投资。此前，CHARM已完成5000万美元A轮融资。CHARM成立于2021年，旨在开创端到端的3D深度学习，以发现和开发针对以前难以成药靶点的转化药物

德国：投资Iambic Therapeutics

Iambic成立于2019年，2021年完成5300万美元的A轮融资，未来计划利用这笔资金推进AI和自动化技术，以及两条管线在明年初进入临床试验阶段，并使用英伟达公司提供的技术，包括AI超级计算平台和NVIDIA BioNeMo云服务以支持公司的药物发现工作。

2024年9月9日，
工信部印发的《首台（套）重大技术装备推广应用指导目录（2024年版）》的通知，
其中集成电路生产装备目录中涉及到两款光刻机型号，
其中：“氟化氩光刻机——照明波长：193nm；分辨率≤65nm；套刻≤8nm”

2024年9月10日，
上海微电子公开了一项名为“极紫外辐射发生装置及光刻设备”的发明专利，
极紫外光刻技术（EUV lithography）是半导体制造领域实现7nm及以下节点芯片制造的关键技术之一。

我们正在快速发展！
中国加油！