长江证券 CHANGJIANG SECURITIES

| 证券研究报告 |

行业研究 | 深度报告 | 半导体与半导体生产设备

英伟达：AI 黄金时代中的卖铲人

请阅读最后评级说明和重要声明

报告要点

英伟达（NVIDIA），1993 年由 Jenson Huang(黄仁勋)及来自于 Sun Microsystem 的两位工程师 Chris Malachowsky 和 Curtis Priem 创立，早期专注于图形芯片设计业务，随着技术与业务的发展，已成长为一家提供全栈计算的人工智能公司，产品覆盖 CPU、DPU、GPU 和 AI 软件，应用领域也从游戏拓展至数据中心、专业可视化、自动驾驶等，随着技术与业务的发展。近年来，英伟达已经成长为全球图形加速、AI 算力的龙头企业。

分析师及联系人

杨洋

SAC：S0490517070012

SFC：BUW100

钟智铧

SAC：S0490522060001

| 证券研究报告 |

2024-08-10

行业研究 | 深度报告

投资评级 看好 | 维持

英伟达：AI 黄金时代中的卖铲人

生成式 AI 引爆技术奇点，GPU 行业迎高增机遇

未来，随着人工智能技术的不断进步和应用领域的扩大，AI 服务器预计将成为服务器市场的核心增长点。AI 服务器专为处理复杂的数据密集型任务而设计，它们需要大量的并行计算能力来执行机器学习和深度学习算法，这使得计算芯片在 AI 服务器中占据了更高的成本比例。与传统服务器相比，AI 服务器对计算能力的要求更高，因此对高性能计算芯片的需求也更为迫切。GPU 由于其并行处理能力，在加速这些计算密集型任务中发挥着至关重要的作用，特别是在 AI 训练和推理过程中，GPU 能够提供比传统 CPU 更高的性能和效率。据 IDC 预测，2027 年 AI 服务器硬件市场规模有望达 1000 亿美元，而且其中相比传统服务器占比更高的计算芯片（如 GPU、ASIC、FPGA）有望充分享受快速增长的浪潮。

GPU：并行运算效率领先，诞自游戏而盛于 AI

GPU（Graphics Processing Unit，图形处理器）是一种专门用于处理图像和图形相关运算的微处理器，主要功能是将计算机系统所需的显示信息进行转换驱动，并向显示器提供行扫描信号，从而实现图像的显示。在早期，所有的图形渲染任务都由 CPU 来完成，但随着计算需求的增加，GPU 逐渐成为专门处理图形渲染的硬件。在作为图形显示芯片时 GPU 广泛应用于个人电脑、工作站、游戏机以及一些移动设备中，同时由于 GPU 本身架构非常适合重复冗余的并行数据处理，因此近年来在人工智能、科学计算领域得到了越来越广泛的应用。

英伟达：软硬件大平台铸造核心壁垒，GPU 龙头迎时代浪潮更上一层楼

英伟达（NVIDIA），1993 年由 Jenson Huang(黄仁勋)及来自于 Sun Microsystem 的两位工程师 Chris Malachowsky 和 Curtis Priem 创立，早期专注于图形芯片设计业务，随着技术与业务的发展，已成长为一家提供全栈计算的人工智能公司，产品覆盖 CPU、DPU、GPU 和 AI 软件，应用领域也从游戏拓展至数据中心、专业可视化、自动驾驶等，随着技术与业务的发展。近年来，英伟达已经成长为全球图形加速、AI 算力的龙头企业。

AI 扬帆，巨龙展翅——英伟达踏上宏伟航路

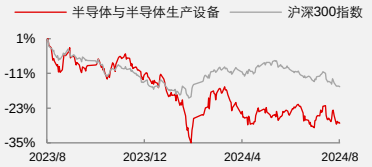
AI 应用的快速爆发&自身不断完善的软硬件体系形成共振，英伟达作为全球 GPU 龙头企业有望踏上高速增长的长期成长通道。在芯片、服务器等硬件设施之上，CUDA、DOCA 等开发套件构成了英伟达软件业务的底层基础框架，在此之上形成 HPC、AI、Omniverse 平台，最终在应用工具&框架层面提供企业 AI、自动驾驶、云游戏、元宇宙、医疗等众多计算服务，三重壁垒联动+螺旋提升打造 AI 全栈体系，系统级 AI 解决方案大平台将成为英伟达长期高增动力。

风险提示

1、下游需求不及预期的风险；

2、全球政治经济动荡影响产品区域性出货的风险。

市场表现对比图(近 12 个月)



资料来源：Wind

相关研究

- 《2024Q1 半导设备及材料综述：收入端加速增长，景气度持续回暖》2024-05-20
- 《景气回暖+Chiplet 加速应用，封测行业多重 β 演绎长期成长逻辑》2023-05-27
- 《AI 重构生产力下的电子行业投资机遇分析》2023-03-27

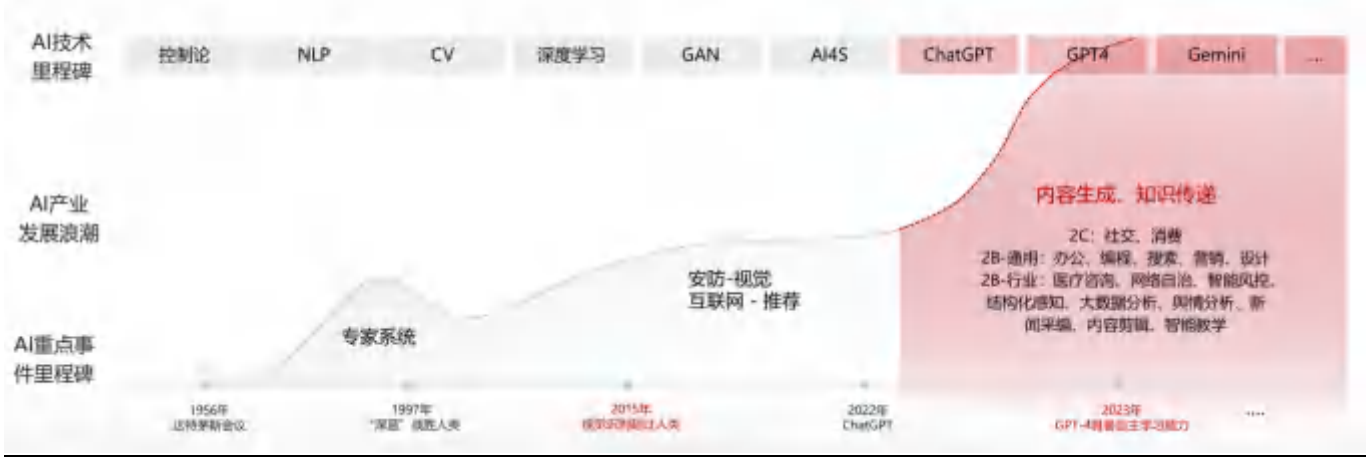
目录

图表目录

生成式 AI 引爆技术奇点，GPU 行业迎高增机遇

AIGC 全称为 AI-Generated Content，指基于生成对抗网络 GAN、大型预训练模型等人工智能技术，通过已有数据寻找规律，并通过适当的泛化能力生成相关内容的技术。与之相类似的概念还包括合成式媒体（Synthetic Media），通过人工智能算法生成、操控与修改数据或媒体，包括文本、代码、图像、语音、视频和 3D 内容等。2020 年，参数量达 1750 亿的 GPT-3 在问答、摘要、翻译、续写等语言类任务上展现出了优秀的通用能力，证明了海量数据、更多参数、多元的数据采集渠道可构成 AI 发展的关键基础。2022 年 12 月，ChatGPT 3.5 令人惊艳的使用体验引爆社会热潮，搜索热度和用户增长都出现了极为明显的提升，目前全球大模型已进入百花齐放的阶段，GPT-4o、阿里通义千问 Qwen2-72B、Llama 3、盘古大模型等多种模型应用层出不穷。

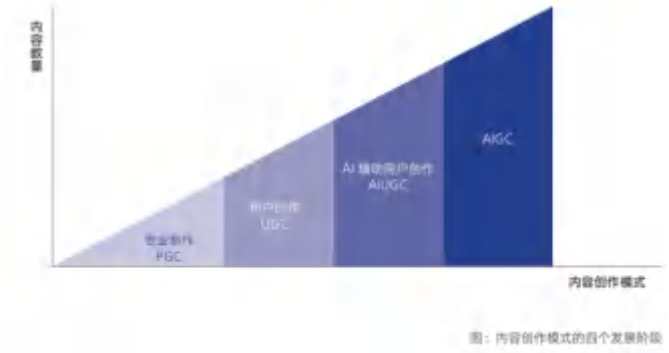
图 1：AI 能力出现拐点，从预测推断走向内容生成



资料来源：《迈向智能世界白皮书 2023》华为，长江证券研究所

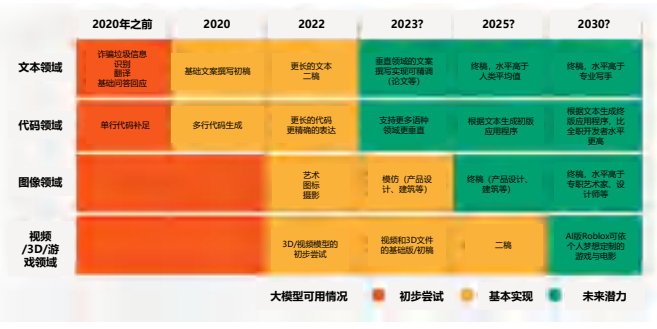
目前，大型文本预训练模型作为底层工具，商业变现能力逐渐清晰。以 GPT-3 为例，其文本生成能力已被直接应用于 Writesonic、Conversion.ai、Snazzy AI、Copysmith、Copy.ai、Headlime 等文本写作/编辑工具中。同时也被作为部分文本内容的提供方，服务于 AI dungeon 等文本具有重要意义的延展应用领域。

图 2：内容创作模式的四个发展阶段



资料来源：《AIGC 发展趋势报告 2023》腾讯研究院，长江证券研究所

图 3：生成式 AI 技术的成熟应用进程时间表



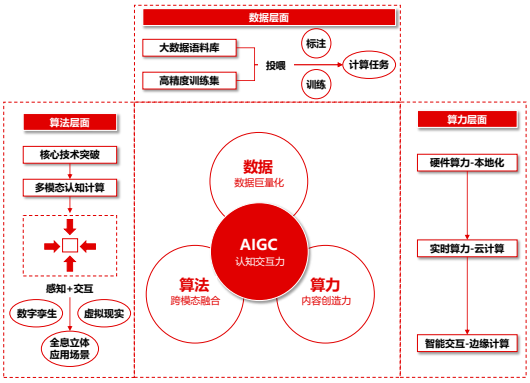
资料来源：《AIGC 发展趋势报告 2023》腾讯研究院，长江证券研究所

AIGC 的本质是内容与场景，其发展需要 AI 与后端基建，算法、算据和算力三要素耦合共振。AIGC 的三大发展阶段是：

- 模型赋智阶段(从现实生成数字)：AIGC 利用 AI 技术构建模拟现实世界的数字孪生模型；
- 认知交互阶段(从数字生成数字)：AI 能够学习并创作更丰富的内容 ；
- 空间赋能阶段(从数字生成现实)：AIGC 基于物联网，多模态技术获取多维信息，实现更加智能的人与机器互动。

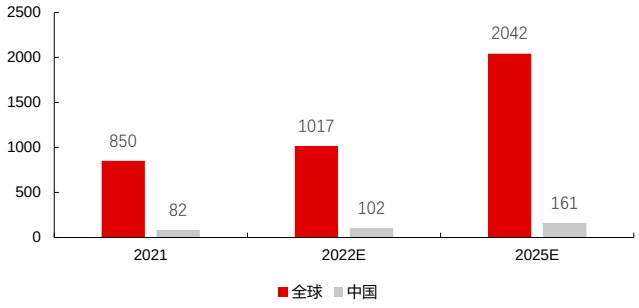
在 AI 的快速发展驱动下，2021 年，全球人工智能市场收支规模(含硬件、软件及服务)达 850 亿美元。IDC 预测，该市场规模将于 2025 年突破 2000 亿美元大关，CAGR 达 24.5%，显示出强劲的产业化增长势头。2021 年，中国人工智能市场收支规模达到 82 亿美元，占全球市场规模的 9.6%，在全球人工智能产业化地区中仅次于美国及欧盟，位居全球第三。IDC 预测，2022 年该市场规模将同比增长约 24%至 102 亿美元，并将于 2025 年突破 160 亿美元。

图 4：人工智能三要素逐步成熟，推动行业进入爆发期



资料来源：甲子光年，长江证券研究所

图 5：全球及我国人工智能市场收支规模及预测（亿美元）

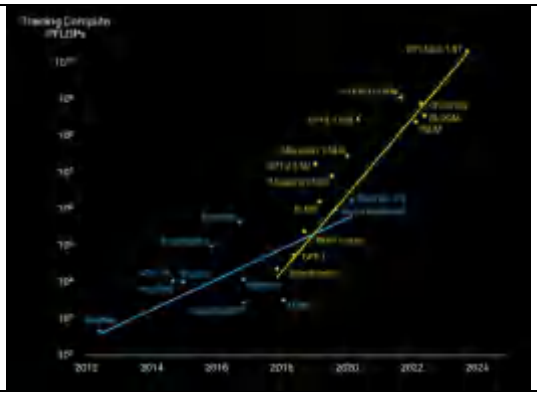


资料来源：IDC Global，IDC China，上海数字大脑研究院，长江证券研究所

在现代人工智能领域，算力扮演着推动创新、实现突破的核心驱动力。为了成功训练大规模的人工智能模型，需要在算力、算法、数据以及系统架构等多个维度进行综合优化。从技术角度来看，预训练、微调和模型推理等环节构成了大模型研发过程中的核心要素和主要的计算需求。

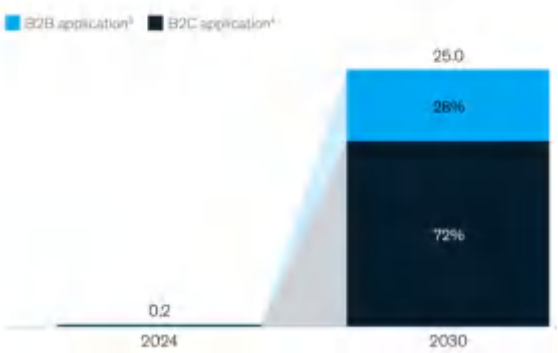
在大模型的核心构成中，除了算法本身，参数设置也至关重要。参数量（Params）是衡量模型规模的一个指标，它与算法中的空间复杂度相似，通常参数量越大，神经网络模型的复杂性越高，对计算资源的需求也越大。一些复杂的神经网络模型的参数量可以达到千亿甚至万亿级别，这与应用级别的模型在参数规模上存在指数级的差异。自 2022 年底以来，随着 ChatGPT 等大规模参数通用大模型的成功推出，这些模型的训练需求推动了智能计算能力的巨大增长。这些模型的训练不仅需要处理千亿甚至万亿级别的参数，还需要处理高达数千 GB 的高质量数据，从而极大地推动了对智能算力的需求增长。

图 6：大模型参数快速提升，对于训练、推理芯片的性能要求越来越高



资料来源：英伟达官网，长江证券研究所

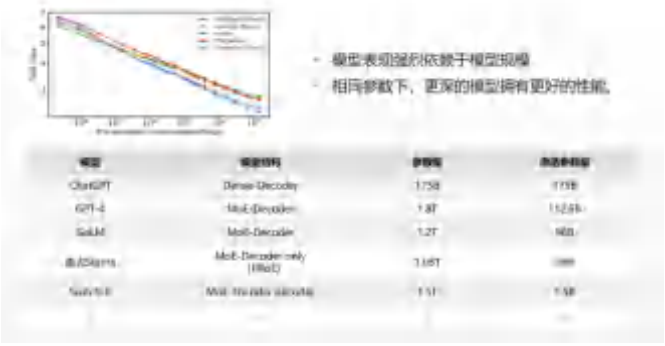
图 7：B2C\B2B 对算力的需求（QFLOPS）



资料来源：McKinsey，长江证券研究所

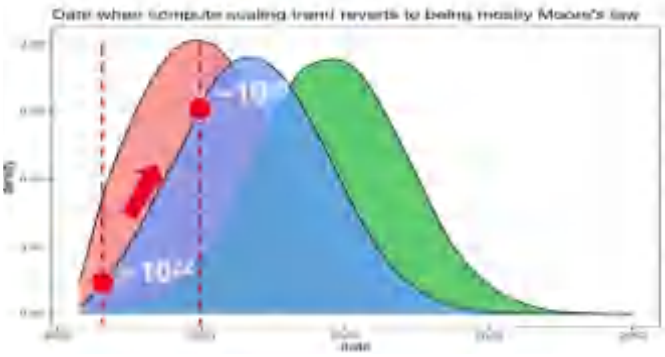
Scaling Law 尚未见顶，目前拥有万亿参数的多模态大型模型已成为研究和应用的新焦点。模型的性能显著地依赖于其规模大小，随着计算量、数据量和参数量的增加而显著提升。在参数数量相同的情况下，更深的神经网络模型往往能够展现出更优越的性能。多模态数据已成为训练这些大型模型的主要数据源，其对计算资源的需求是传统文本数据的数百倍。大型模型的算力需求呈现出每六个月翻倍的趋势，这一趋势预计将至少持续至 2030 年。随着模型规模的不断扩大，对算力的需求也在急剧上升，这不仅推动了计算硬件的发展，也对算法的优化提出了更高的要求。

图 8：Scaling Law 尚未见顶，MOE 万亿参数大模型是新的热点



资料来源：《迈向智能世界白皮书 2023》华为，长江证券研究所

图 9：大模型算力需求 6 个月翻一番的趋势，预计至少持续到 2030 年



资料来源：《迈向智能世界白皮书 2023》华为，长江证券研究所

在人工智能领域，大模型技术正在逐步实现标准化和统一化，生态体系也在向集成化发展，模型设计趋向于更加精简和统一的框架。随着对更高性能 AI 大模型的追求，对算力平台的要求也日益提高，技术门槛随之提升。未来，AI 模型训练的核心将转向集群系统的综合能力。

AI 大模型对算力的需求正以指数级速度增长，推动了 AI 算力平台从单一的单机计算向集群计算的转变。构建超大规模的 AI 集群面临着三大技术挑战：首先是液冷技术的大规模商用，这在工程实施上存在一定的挑战；其次是 AI 集群的建设本身就是一项复杂

的系统工程，需要跨学科、跨领域的协同合作；最后，AI 大模型的训练高度依赖于 AI 集群的高可用性，这要求集群必须具备高度的稳定性和可靠性。

图 10：开发更高性能的 AI 大模型需要更强的算力平台

资料来源：华为，长江证券研究所

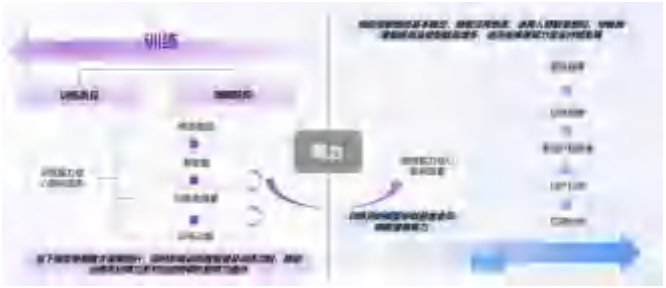
图 11：算力底座技术门槛提高，未来训练核心拼集群系统能力

资料来源：华为，长江证券研究所

AI 技术在实际应用中包括两个环节：训练（Training）和推理（Inference），AIGC 的算力需要考虑训练及推理两个方面。训练是指通过数据开发出 AI 模型，使其能够满足相应的需求，一般为 AI 技术的研发。因此参数量的升级对算力的需求影响大。 推理是指利用训练好的模型进行计算，利用输入的数据获得正确结论的过程，一般为 AI 技术的应用。推理部署的算力主要在于每个应用场景日数据的吞吐量。

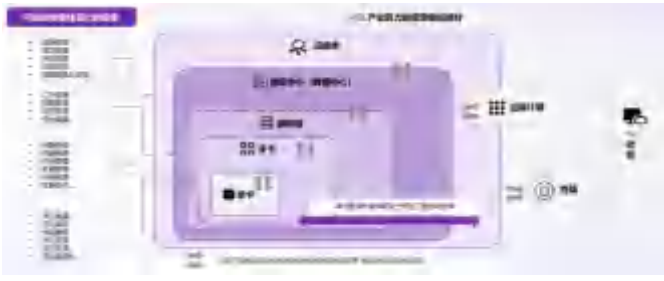
未来大模型的产业化发展是一套复杂的系统工程，构建高效稳定的算力平台是核心要义，成熟的算法、数据产业链，配套工具链及丰富的生态链是关键因素，亟需 以系统的方式寻找最优解。算力设备软硬件兼容性和性能调教上的 Know-How，可以保证 AI 算力的适配性和稳定性，并非单一因素的参数能简单决定。

图 12：训练&推理对算力均带来显著需求



资料来源：甲子光年，长江证券研究所

图 13：AIGC 产业的算力是工程化结果，是从芯片到资源服务的多层次构造

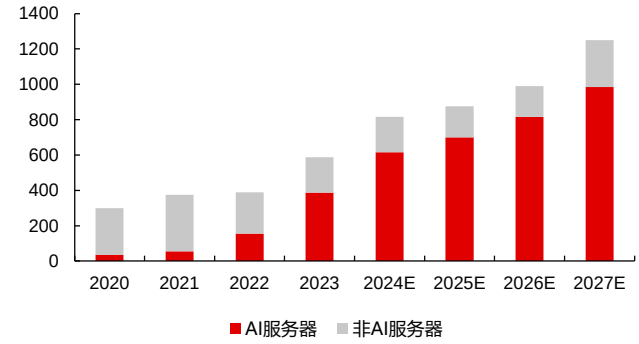


资料来源：甲子光年，长江证券研究所

未来，随着人工智能技术的不断进步和应用领域的扩大，AI 服务器预计将成为服务器市场的核心增长点，而其中的计算芯片又是“灵魂”。AI 服务器专为处理复杂的数据密集型任务而设计，它们需要大量的并行计算能力来执行机器学习和深度学习算法，这使得

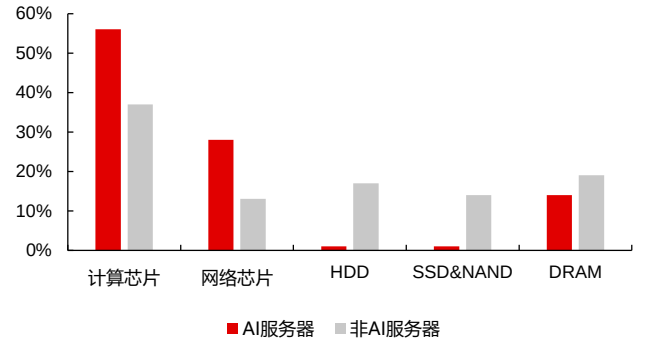
计算芯片在 AI 服务器中占据了更高的成本比例。与传统服务器相比，AI 服务器对计算能力的要求更高，因此对高性能计算芯片的需求也更为迫切。GPU 由于其并行处理能力，在加速这些计算密集型任务中发挥着至关重要的作用，特别是在 AI 训练和推理过程中，GPU 能够提供比传统 CPU 更高的性能和效率。据 IDC 预测，2027 年 AI 服务器硬件市场规模有望达 1000 亿美元，而且其中相比传统服务器占比更高的计算芯片（如 GPU、ASIC、FPGA）有望充分享受快速增长的浪潮。

图 14：AI 服务器将成为服务器的核心增长点（亿美元）



资料来源：IDC，长江证券研究所

图 15：相比传统服务器，AI 服务器整体成本中计算芯片占比更高

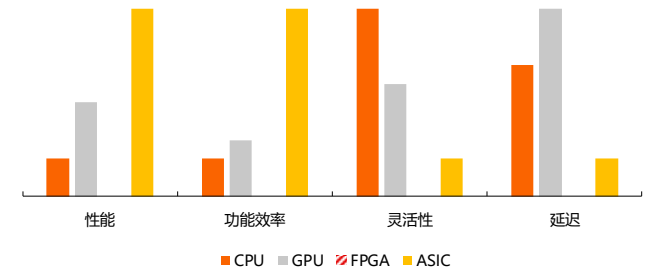


资料来源：IDC，长江证券研究所

以人工智能芯片为例，目前主要有两种发展路径：一种是延续传统计算架构，加速硬件计算能力，主要以 CPU、GPU、FPGA、ASIC 为代表。当前阶段，GPU 配合 CPU 是 AI 芯片的主流，而后随着视觉、语音、深度学习的算法在 ASIC 芯片上的不断优化，此两者也将逐步占有更多的市场份额，从而与 GPU 达成长期共存的局面。

深度神经网络算法是大型多层的网络模型，典型的有循环神经网络和卷积神经网络，模型单次推断通常需要数十亿甚至上百亿次的运算，对芯片的计算力提出了更高要求，同时对器件的体积、功耗还有一定的约束。

图 16：四类逻辑芯片特性比较



资料来源：与非网，长江证券研究所

图 17：主要 AI 芯片的功能特性比较

	GPU	FPGA	ASIC
定制化程度	通用型	半定制化	定制化
灵活性	好	好	不好
成本	高	较高	低
编程语言/架构	CUDA、OpenCL等	Verilog/VHDL等硬件描述语言，OpenCL、HLS	/
功耗	大	较大	小
主要优点	峰值计算能力强、产品成熟	平均性能较高、功耗较低、灵活性强	平均性能很强，功耗很低、体积小
主要缺点	效率不高、不可编辑、功耗高	量产单价高、峰值计算能力较强、编程语言难度大	前期投入成本高、不可编辑、研发时间长、技术风险大
主要应用场景	云端训练、云端推断	云端推理、终端推断	云端训练、云端推断、终端推断
代表企业芯片	英伟达Tesla、高通Adreno等	XilinxVersal、英特尔Arria等	谷歌TPU、寒武纪Cambricon等

资料来源：赛迪智库，长江证券研究所

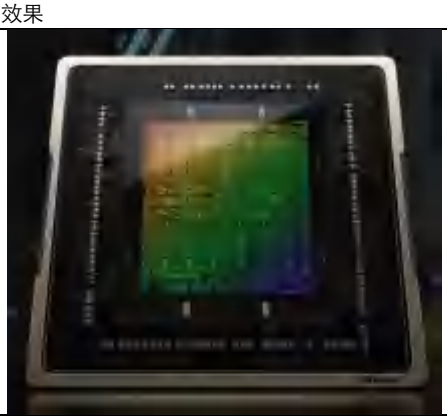
GPU：并行运算效率领先，诞自游戏而盛于 AI

GPU（Graphics Processing Unit，图形处理器）是一种专门用于处理图像和图形相关运算的微处理器，主要功能是将计算机系统所需的显示信息进行转换驱动，并向显示器提供行扫描信号，从而实现图像的显示。在早期，所有的图形渲染任务都由 CPU 来完成，但随着计算需求的增加，GPU 逐渐成为专门处理图形渲染的硬件。在作为图形显示芯片时 GPU 广泛应用于个人电脑、工作站、游戏机以及一些移动设备（如平板电脑、智能手机等）中，但同时由于 GPU 本身架构非常适合重复冗余的并行数据处理，因此近年来在人工智能、科学计算领域得到了越来越广泛的应用。

图 18：1983 年由 TI 推出的第一款 GPU 芯片，用于雅利达游戏机，显存仅为 16kb，分辨率为 256*192

资料来源：MSX，长江证券研究所

图 19：目前最先进的游戏显卡英伟达 RTX40 系列显存达 16GB，可实现光线追踪效果



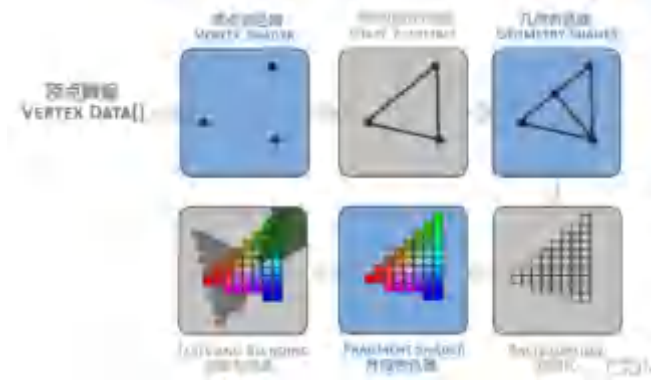
资料来源：英伟达官网，长江证券研究所

GPU 的核心优势在于其拥有大量的核心，这些核心可以同时执行多个任务，从而大幅提高计算速度以提供较强的并行计算能力。这种并行处理能力使得 GPU 能够快速将图形结果计算出来，并在屏幕的所有像素中进行显示。GPU 内部包含多个处理器，如顶点处理器、几何处理器和光栅化处理器等。这些处理器协同工作，分别负责不同的渲染阶段，从而进一步提高渲染效率。**渲染过程通常通过一个称为“渲染管线”的流程进行，该流程包括多个阶段，如顶点处理、几何处理、光栅化和着色等。**在渲染管线中，着色器（shader）起到了至关重要的作用。它们是小型程序，用于定义物体的外观和光照效果。GPU 通过着色器来实现复杂的视觉效果和动态变化。

图 20：GPU 处理图形主要分为几何阶段和光栅化阶段

资料来源：CSDN，长江证券研究所

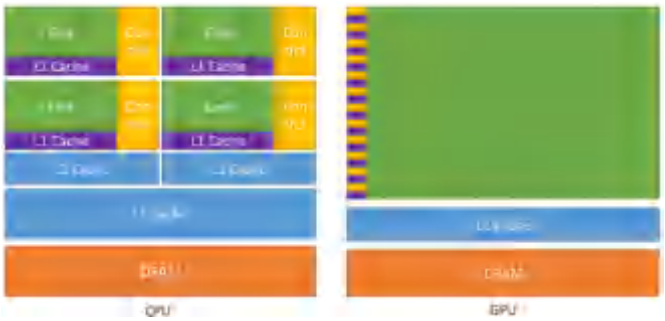
图 21：渲染着色需要大量 shader 串行完成数据处理



资料来源：CSDN，长江证券研究所

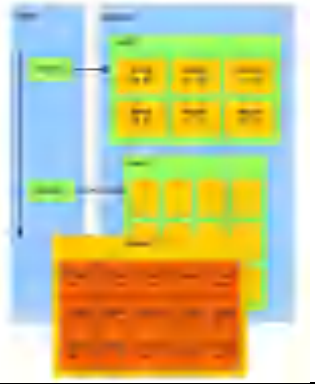
由于存储器的发展慢于处理器，在 CPU 上发展出了多级高速缓存的结构，在 GPU 中，也存在类似的多级高速缓存结构，相比 CPU，GPU 将更多的晶体管用于数值计算，而不是缓存和流控（Flow Control），CPU 的 Cache 和 Control 较多，更为适合处理复杂的逻辑任务，GPU 则有更多的 Core，使其更为适合处理并行线程，相对应的 GPU 在处理复杂逻辑任务的表现相对较弱。

图 22：CPU 与 GPU 的结构差异



资料来源：英伟达官网，长江证券研究所

图 23：核（ire）→线程组（Thread）→线程块（Block）→网格（Grid）的多层级复合堆积结构使得 GPU 更适合处理简单重复的并行运算



资料来源：英伟达官网，长江证券研究所

基于 GPU 在图形显示和并行运算上的优势，GPU 的应用范围逐渐从对图形、游戏的加速图形渲染向电影、电视、医疗影像等领域扩扩展，人工智能、机器学习、科学计算、加密货币挖矿、数据中心和云计算、自动驾驶和机器人等领域也进入百花齐放的阶段，越来越多次世代应用采用了 GPU 为核心的硬件架构，这大大推动了 GPU 市场规模的提升。

图 24：阿凡达中用光线追踪技术制作的镜头，GPU 在图形渲染中的应用范围持续扩大

资料来源：Broadgeek，长江证券研究所

图 25： 以 GPU 为计算核心的服务器集群已成为 AI 发展的关键基础

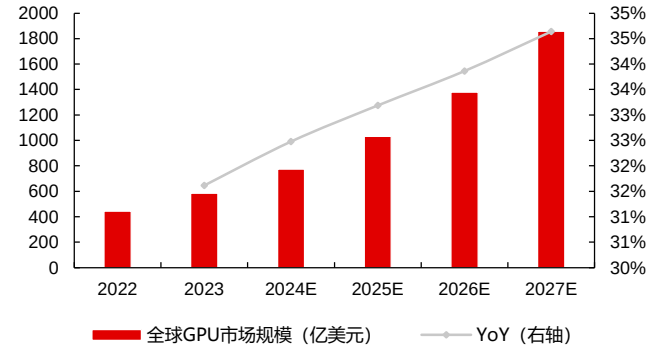
资料来源：英伟达官网，长江证券研究所

在图形渲染、加速运算需求持续爆发的带动下，全球 GPU 市场正迎来一个前所未有的增长期。据 Technavio，2022 年全球 GPU 市场规模已经达到 443.8 亿美元，并在 2023 年进一步增长至 584.1 亿美元，这一上升趋势预计将持续至 2027 年，届时市场规模有望飙升至 1857.5 亿美元，2022~2027 年 GPU 市场规模复合增长率达 33.15%。

目前，移动设备、个人电脑及工作站是 GPU 市场的主要组成部分，据 Modor Intelligence，它们在 2023 年分别占据了全球 GPU 市场规模的 46%和 40%。然而，随着人工智能技术的快速发展，AI 的深度学习和机器学习算法对计算能力的需求日益增长，GPU 因其卓越的并行处理能力而成为这些应用的理想选择，服务器市场预计将成为推动 GPU 增长的新引擎，尤其是在云计算和数据中心的大规模部署中。

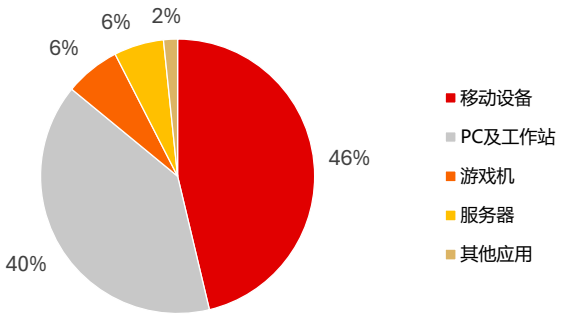
GPU 行业的发展不仅仅局限于 AI 和服务器市场。随着 5G 技术的普及和物联网（IoT）设备的增加，对高性能图形处理的需求也在不断上升，这将进一步推动 GPU 市场的成长。此外，GPU 在游戏、专业图形设计、视频编辑、科学计算以及自动驾驶汽车等领域的应用也在不断扩展，为行业带来新的增长机遇。

图 26：全球 GPU 市场规模 2027 年有望达到 1857.5 亿美元



资料来源：Technavio，长江证券研究所

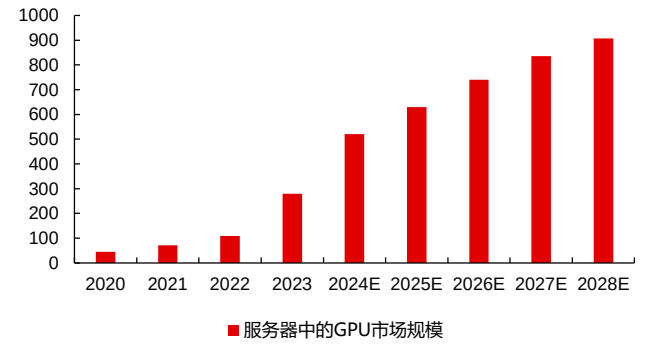
图 27：2023 年全球 GPU 下游应用仍以移动设备、PC 及工作站为主



资料来源：Mordor Intelligence，长江证券研究所

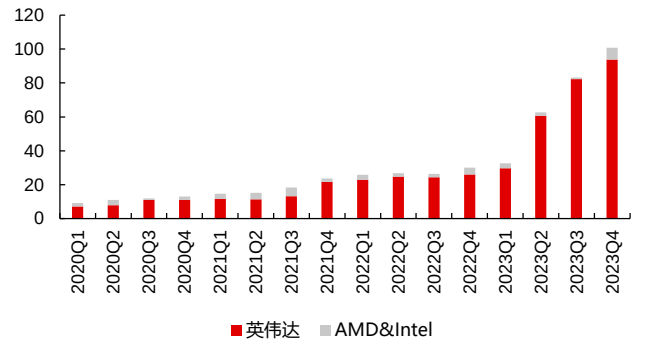
随着人工智能、机器学习、数据分析等技术在各行业的广泛应用，未来服务器中的 GPU 市场规模预计将快速扩大。英伟达作为全球领先的 GPU 制造商，在服务器 GPU 市场中占据核心份额。凭借其强大的产品性能、广泛的软件生态系统以及持续的技术创新，英伟达有望充分受益于这一行业增长趋势。英伟达的数据中心业务已经展现出强劲的增长势头，随着 AI 技术的进一步发展和市场需求的不断扩大，英伟达的 GPU 产品，特别是为 AI 和高性能计算设计的系列产品，预计将在服务器市场中继续保持领先地位，推动公司业务的持续增长。

图 28：服务器中的 GPU 市场规模快速扩大（亿美元）



资料来源：IDC，长江证券研究所

图 29：英伟达在服务器 GPU 中占据核心份额

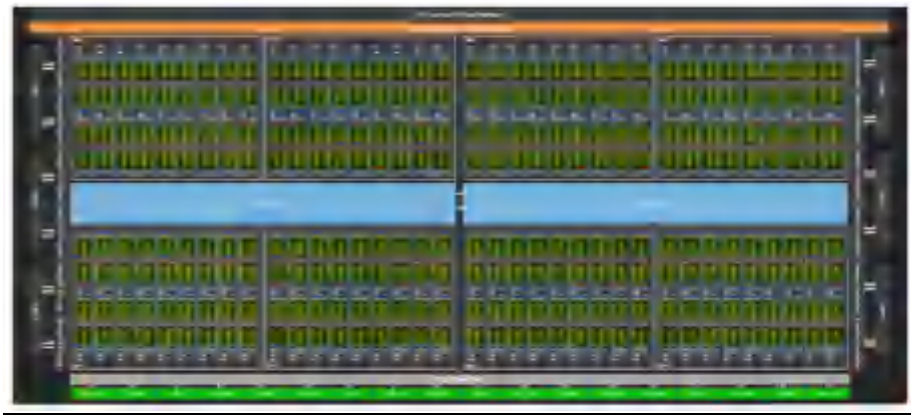


资料来源：IDC，长江证券研究所

GPU 在硬件设计、制造和配套软件配套上有着较高的壁垒。首先，技术复杂性是一个重要壁垒，GPU 设计需要深厚的技术积累和专业知识。其次，研发成本高昂，开发高性能 GPU 需要巨大的资金投入和长期的研发周期。再者，知识产权如专利构成了法律壁垒，现有企业如英伟达和 AMD 拥有大量 GPU 相关专利。此外，市场认可度也是一大壁垒，现有品牌已经建立了强大的市场信任和用户基础。软件生态系统同样关键，强大的软件支持和开发者社区对于 GPU 的成功至关重要。最后，制造工艺也是一个壁垒，先进的半导体制造技术不易获得，需要与顶级的代工厂建立合作关系。这些壁垒共同维护了 GPU 市场的稳定性，同时也限制了新竞争者的进入。

- **技术架构壁垒：**GPU 设计是一项系统工程，其硬件架构复杂，需要高度优化以支持并行处理和高吞吐量计算，这要求精细的工程设计来平衡性能和功耗。设计者必须精通复杂的计算图形学和并行计算理论，确保 GPU 能够有效地执行图形渲染、深度学习和其他计算密集型任务。此外，GPU 架构必须具备高度可扩展性，以适应从移动设备到超级计算机的不同应用场景。散热管理也是设计中的一个挑战，因为 GPU 在运行时会产生大量热量。还需要考虑内存带宽、数据传输效率以及与 CPU 等其他系统组件的协同工作。最后，随着技术的发展，GPU 架构还需要不断创新以支持新兴技术，如光线追踪、AI 加速和虚拟化。这些因素共同构成了 GPU 硬件架构设计的难点。例如，英伟达的 Hopper 架构包含数千个 CUDA Core 和深度学习矩阵运算单元，这些硬件的精密设计构成了 GPU 的硬件壁垒。

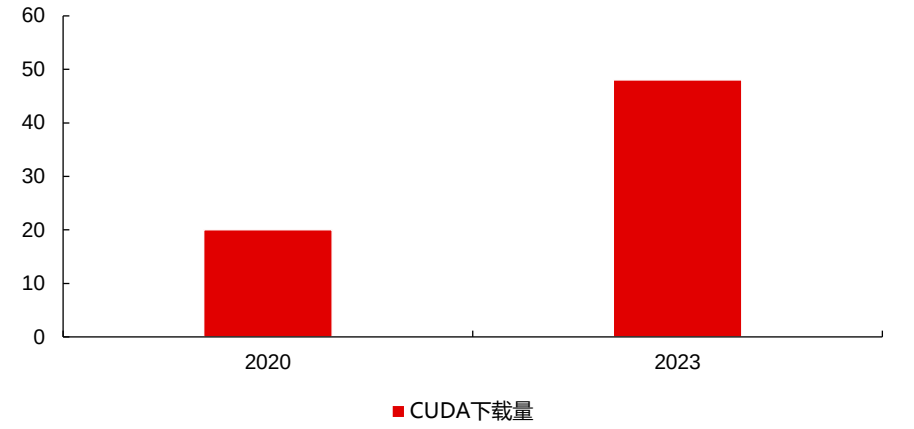
图 30：英伟达 H100 硬件架构示意图，大量 CUDA Core 需要跟片上缓存、管口配合



资料来源：英伟达官网，长江证券研究所

- **算法和软件生态：**GPU 图形渲染需要用到计算图形学，涉及数学、物理等多学科知识。GPU 软件生态的难点在于创建一个支持广泛应用程序、易于开发者使用且能够充分利用 GPU 硬件性能的开发环境。这需要提供强大的编程模型、丰富的 API、高效的运行时库以及优化工具，同时还必须保持与不断演进的硬件架构同步。开发者需要能够轻松地编写、调试和部署在 GPU 上运行的代码，同时软件生态还必须支持多平台、多语言和多种计算框架。此外，构建一个充满活力的开发者社区，提供必要的教育资源和技术支持，也是软件生态成功的关键。此外，软件生态也是 GPU 厂商的重要竞争屏障。例如，英伟达推出的 CUDA 平台形成了开发人员社区生态，增加了竞争对手的进入难度。

图 31：2020 年以来英伟达 CUDA 生态持续扩大（百万次）

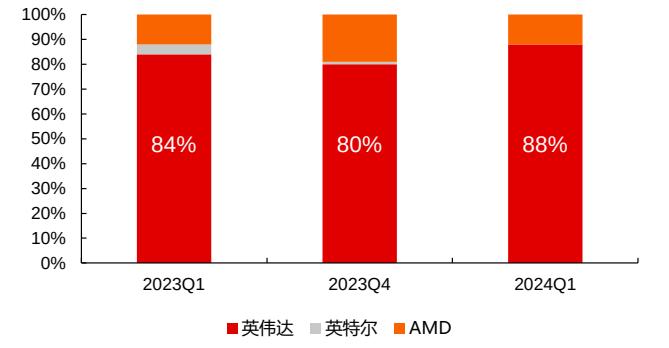


资料来源：英伟达官网，长江证券研究所

- **研发投入和周期：**GPU 的研发需要大量的资金投入和长周期的技术积累。新进入者要自主研发高性能 GPU，需要从零开始，面临较大的难度和风险。
- **供应链和制造能力：**GPU 的生产需要先进的半导体制造工艺，如台积电的 5nm 工艺。新进入者要获得同等水平的制造能力，需要解决供应链和生产上的诸多难题。

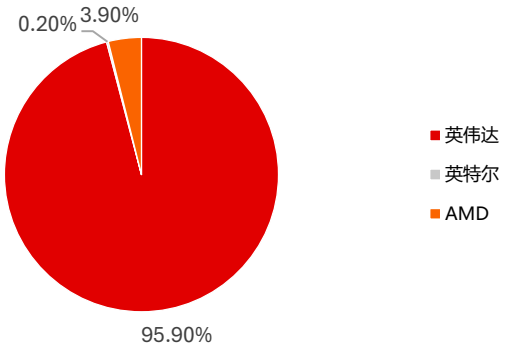
在硬件、软件、生态等多重壁垒的推动下，英伟达已经逐步成为全球 GPU 的龙头企业。英伟达的成功不仅源于其在 GPU 硬件架构上的技术领先，还因为其在软件生态方面的深远布局。通过推出 CUDA 平台，英伟达为开发者提供了强大的工具和库，极大地简化了并行编程的复杂性，吸引了广泛的开发者社区和科研机构的支持。此外，英伟达不断推动技术创新，如实时光线追踪、AI 加速计算等，进一步巩固了其在高性能计算和游戏领域的领导地位。教育和研究领域的合作也扩大了英伟达的影响力，通过学术合作和奖学金项目培养了未来的技术人才。同时，英伟达还积极拓展与行业伙伴的合作，构建了一个强大的产业生态系统，包括 OEM、ISV 和云服务提供商，确保了其技术和产品能够广泛应用于各个领域。这些因素共同作用，使得英伟达在 GPU 市场中占据了难以撼动的地位。目前，据 JPR，英伟达 2023Q4 在全球桌面级 GPU 的市场份额达 80%，据 IDC，英伟达在全球服务器 GPU 中的市场份额高达 95.9%。

图 32：英伟达目前仍为桌面级 GPU 市场的核心龙头，份额持续提升



资料来源：JPR，长江证券研究所

图 33：英伟达在全球服务器 GPU 中的市场份额高达 95.9%



资料来源：IDC，长江证券研究所

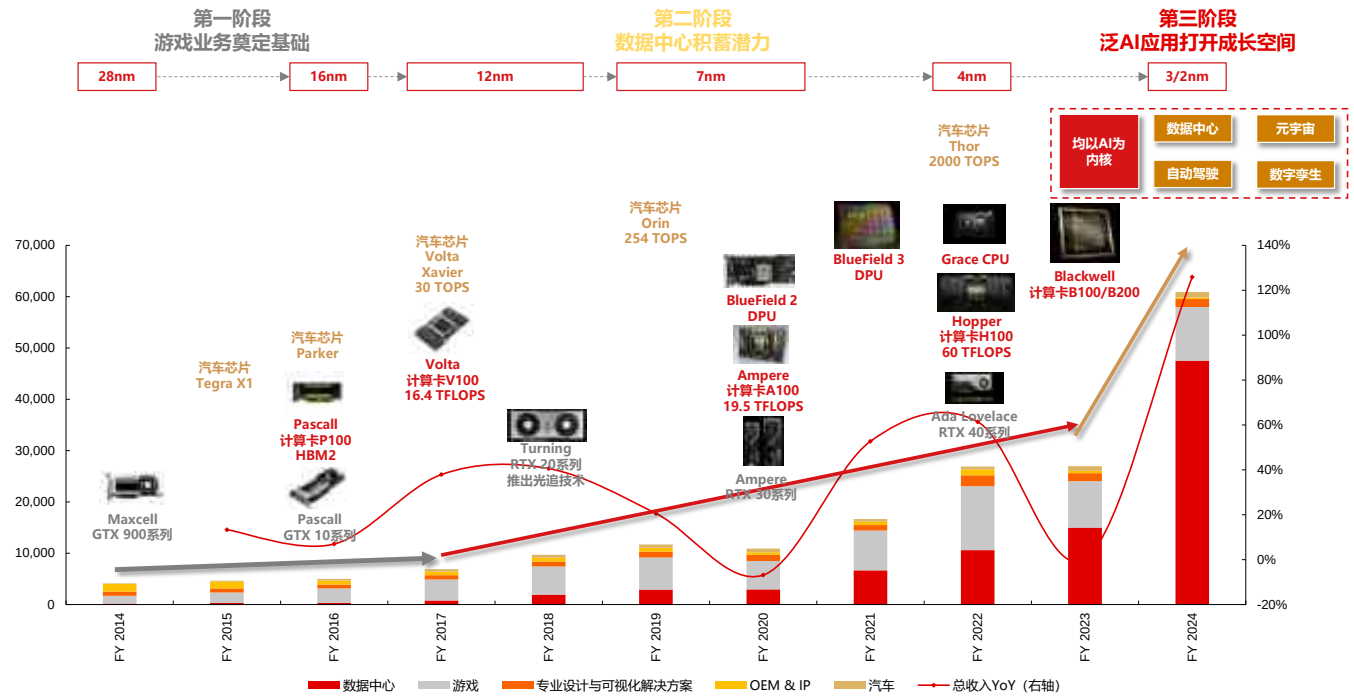
AI 扬帆，巨龙展翅——英伟达踏上宏伟航路

英伟达（NVIDIA），1993 年由 Jenson Huang(黄仁勋)及来自于 Sun Microsystem 的两位工程师 Chris Malachowsky 和 Curtis Priem 创立，早期专注于图形芯片设计业务，随着技术与业务的发展，已成长为一家提供全栈计算的人工智能公司，产品覆盖 CPU、DPU、GPU 和 AI 软件，应用领域也从游戏拓展至数据中心、专业可视化、自动驾驶等，随着技术与业务的发展。近年来，英伟达已经成长为全球图形加速、AI 算力的龙头企业，在硬件端英伟达形成了 CPU+GPU+DPU 的协同布局，其训练和推理芯片性能大幅领先竞争对手，AI 服务器 GPU 份额遥遥领先；在软件端，其 CUDA 架构是目前最适合深度学习和 AI 训练的 GPU 架构之一，已积累 300 个加速库和 400 个 AI 模型主导 AI 训练与推理芯片市场。

GPU 是英伟达的核心产品，围绕 GPU 及其核心应用——图形渲染和加速运算，英伟达持续扩展自身软硬件实力，在硬件的三芯战略+互联网络、软件侧的开发软件+行业应用软件、应用层对各核心下游持续加大投入的三重驱动下，英伟达完成了多个阶段的快速发展，目前已经成为了全球核心 AI 芯片及应用企业：

- **创立初期（1993-2006 年）：**英伟达于 1993 年成立，最初专注于图形处理器（GPU）的研发和生产。在这一阶段，英伟达面临激烈的市场竞争，尤其是在计算机图形芯片市场，当时市场上有 90 个竞争对手。尽管如此，英伟达还是设法在桌面和笔记本电脑的 GPU 市场中占据了一席之地。1999 年，英伟达发布了全球第一款 GPU——GeForce 256，这标志着 GPU 时代的开始。
- **技术创新与扩展（2006-2015 年）：**2006 年，英伟达发布了 CUDA 并行计算平台和编程模型，这一创新极大地推动了后来的人工智能技术发展。此外，英伟达在这一时期还推出了多款重要的 GPU 架构，如 Fermi、Kepler 等。这些技术创新不仅巩固了英伟达在游戏和专业图形市场的地位，也为公司后来进入 AI 领域奠定了基础。
- **AI 时代的崛起（2015-至今）：**从 2015 年开始，英伟达的业绩和估值开始快速增长，股价在 6 年内上涨了 70 倍，市值超过 8000 亿美元，成为全球市值第八大的公司。这一转变主要得益于英伟达在 AI 领域的深入布局和技术创新。2016 年，英伟达推出了 AI 加速器 Tesla P100 和 Volta 架构，进一步加强了其在 AI 计算加速处理器市场的领导地位。此后，英伟达继续推出新一代的 AI 技术和产品，如 Ampere 架构、NVIDIA Xavier 自动驾驶处理器等，并在 AI 算力领域取得了显著成就。

图 34：英伟达增长趋势（单位：百万美元）



资料来源：Anandtech，英伟达官网，Thinkcomputers，快科技，长江证券研究所（注：算力大小均取系列产品中单精度计算性能的较大值）

AI 应用的快速爆发&自身不断完善的软硬件体系形成共振，英伟达作为全球 GPU 龙头企业有望踏上高速增长的长期成长通道。在芯片、服务器等硬件设施之上，CUDA、DOCA 等开发套件构成了英伟达软件业务的底层基础框架，在此之上形成 HPC、AI、Omniverse 平台，最终在应用工具&框架层面提供企业 AI、自动驾驶、云游戏、元宇宙、医疗等众多计算服务，英伟达已从一家 GPU 公司升级成计算平台公司。

产品平台化构建竞争壁垒，应用扩张打造增长动力

硬件、软件、应用：英伟达的三重壁垒

三重壁垒联动+螺旋提升打造 AI 全栈体系，系统级 AI 解决方案大平台是英伟达α内核。英伟达通过其“三芯片四领域”的战略，构筑了一个全面的产品矩阵，涵盖了硬件、软件和生态系统三大方面：

- **硬件：**英伟达的硬件产品线主要包括 GeForce 系列（G 系列）和针对数据中心的 GPU 产品，如 A100、DGX A100 和 InfiniBand 等。这些硬件产品支持高性能的图形处理能力和游戏特性，以及云与数据中心领域的需求。此外，英伟达还涉足了 CPU Grace 等新型处理器的研发，进一步丰富其硬件产品矩阵。

图 35：英伟达应用于 AI 运算的 H100 芯片组

资料来源：英伟达官网，长江证券研究所

图 36：英伟达应用于图形显示的 RTX 系列产品

资料来源：英伟达官网，长江证券研究所

- **软件：**在软件方面，英伟达提供了 CUDA 工具包，这是一个免费、强大的并行计算平台和编程模型，支持开发者创建 GPU 加速的高性能应用。CUDA 工具包包含多个库、多种调试和优化工具、一个编译器以及一个用于部署应用的运行环境库。此外，NVIDIA App 为 PC 游戏玩家和创作者提供了必备的辅助工具，包括驱动程序更新、游戏和应用优化等功能。
- **生态系统：**英伟达构建了一个广泛的生态系统，包括 Omniverse 生态系统、DRIVE Hyperion 自动驾驶汽车平台、量子计算生态系统等。Omniverse 生态系统为开发者、企业和创作者提供了各种新功能和新服务，支持 AR、VR、多 GPU 渲染等功能，并与 Bentley 和 Esri 等公司建立了连接。DRIVE Hyperion 平台开放了访问权限，以推动自动驾驶汽车的发展。量子计算生态系统则与多家合作伙伴建立了合作关系，整合量子云技术到产品中。

图 37：英伟达围绕 GPU 硬件基础打造了 CUDA 生态系统

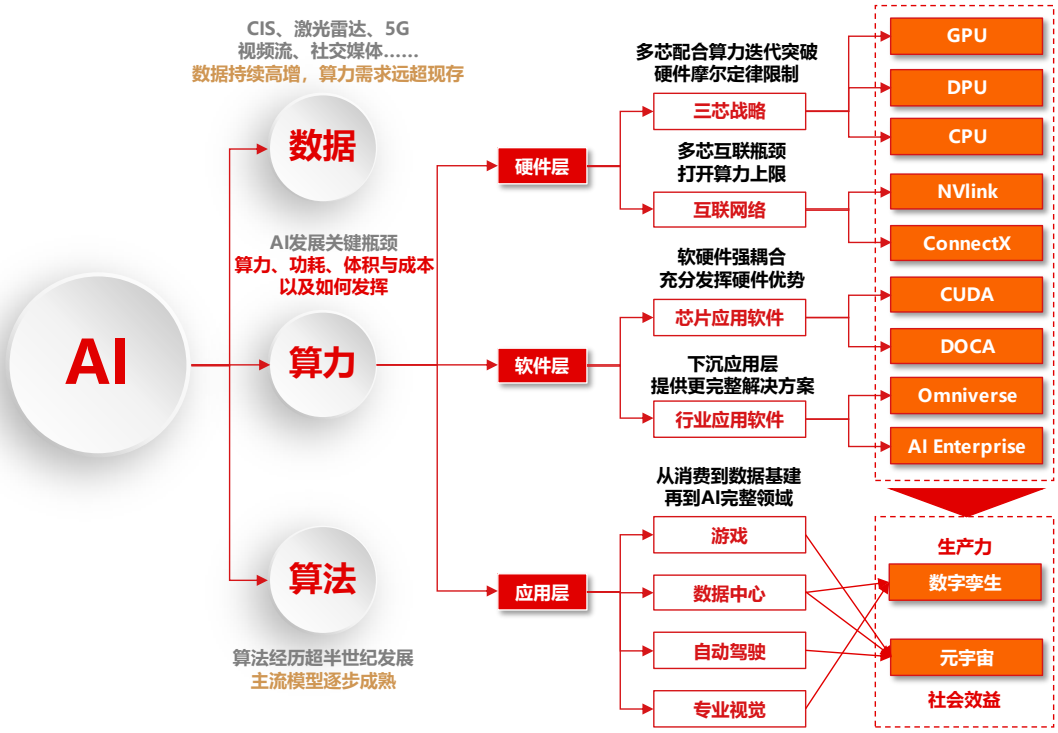
资料来源：英伟达官网，长江证券研究所

图 38：在 CUDA 生态系统至上进一步完善了各类场景应用

资料来源：英伟达官网，长江证券研究所

英伟达通过其硬件产品线、CUDA 等软件工具以及 Omniverse、DRIVE Hyperion 等生态系统，构建了一个全面的产品矩阵，覆盖了从消费级到企业级的不同需求，同时也推动了人工智能、自动驾驶汽车、量子计算等多个领域的技术进步和发展。

图 39：AI 的核心驱动与英伟达的三重壁垒



资料来源：长江证券研究所

硬件层：CPU+GPU+DPU 形成三芯矩阵

GPU：GPU 解决 AI 大规模并行运算痛点：英伟达以 GPU 起家，垄断游戏显卡市场，随后设计 CUDA 平台发挥 GPU 并行运算优势、打造校企研深度绑定的硬件算法生态，目前 90%以上的训练算法依赖 GPU+CUDA，英伟达实际上成为了近十年 AI 发展的底层引擎，最新发布的 Hopper 架构 H100 GPU 在大型 NLP 模型上可提供相比上代 A100 高达 9 倍的 AI 训练速度和 30 倍的 AI 推理速度。由于高性能和良好的通用性，GPU 是 AI 服务器的首选加速方案。超算中心的市场份额超过 70%，AI 加速卡的市场份额超过 90%，且有别于 CPU，**GPU 当前全球并无具备挑战力的 GPU IP 核授权供应商，后发玩家难以快速跟上技术迭代趋势。**

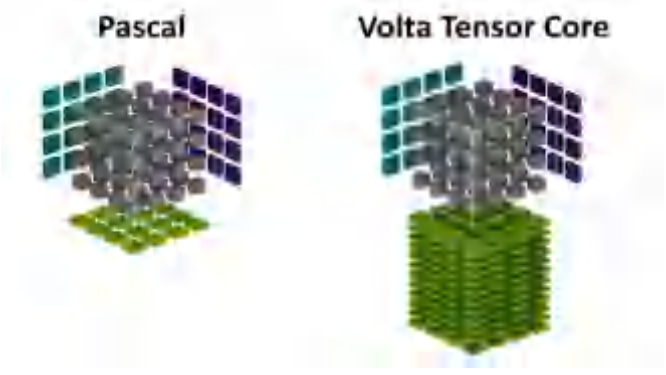
自 1999 年发布第一代 GPU 架构 GeForce 226 以来，英伟达持续推进自身架构的持续升级与发展，其中：1）Tesla 架构首次支持 C 语言编程，开启了 GPU 用于通用计算的新时代；2）Fermi 架构引入了多项技术创新，如 ECC 支持和显著提升的单精度浮点性能；3）Kepler 架构通过动态并行技术和 Hyper-Q 技术显著提升了 GPU 的计算能力和效率；4）Maxwell 架构优化了能效比，通过改进的 SM 设计实现了更高的性能和更低的功耗；5）Pascal 架构引入了 NVLink 技术，增强了多 GPU 系统的互联和扩展性；6）Volta 架构首次引入 Tensor Core，专为深度学习而设计，极大提升了 AI 计算效率；7）Turing 架构进一步推动了光线追踪技术的发展，并引入了第二代 Tensor Core 和 RT Core，为实时渲染和 AI 计算带来革命性进步；8）Ampere 架构作为最新架构，在性能、能效和可扩展性方面都有显著提升，支持更高的显存带宽和计算能力。

表 1：英伟达主要游戏显卡参数

	RTX 40 系列	RTX 30 系列	RTX 20 系列	GTX 16 系列	GTX 10 系列	GTX 900 系列
推出时间	2022	2020	2018	2019	2017	2015
架构名称	Ada Lovelace	Ampere	Turing	Turing	Pascal	Maxwell
制程	5nm	8nm	12nm	12nm	16nm	28nm
流多处理器	2x FP32	2x FP32	1x FP32	1x FP32	1x FP32	1x FP32
RT Core	第 3 代	第 2 代	第 1 代	-	-	-
Tensor Core (AI)	第 4 代	第 3 代	第 2 代	-	-	-
NVIDIA DLSS	DLSS 3.5 超分辨率 DLAA 光线重建帧生成	DLSS 2 超分辨率 DLAA 光线重建	DLSS 2 超分辨率 DLAA 光线重建	-	-	-
PCIe	第 4 代	第 4 代	第 3 代	第 3 代	第 3 代	第 3 代
CUDA 能力	8.9	8.6	7.5	7.5	6.1	5.2
系列旗舰	GeForce RTX 4090 D	GeForce RTX 3090 Ti	GeForce RTX 2080 Ti	GeForce GTX 1660 Ti	GTX 1080 Ti	GTX 980 Ti
NVIDIA CUDA® 核心数量	**14592**	**10752**	**4352**	**1536**	**3584**	**2816**
加速频率 (GHz)	**2.52**	**1.86**	**1.64**	**1.77**	**-**	**1.076**
基础频率 (GHz)	2.28	1.56	1.35	1.5	1.58	1
标准显存配置	24 GB GDDR6X	24 GB GDDR6X	11 GB GDDR6	6GB GDDR6	11 GB GDDR5X	6 GB GDDR5
显存位宽	384 位	384 位	352 位	192 位	352	384
显存带宽	**1.15 TB/s**	**1.01 TB/s**	**616.0 GB/s**	**288.0 GB/s**	**484.4 GB/s**	**336.6 GB/s**
最高 GPU 温度 (℃)	90	92	89	95	91	
显卡总功耗 (W)	425	450	260	120	250	250
要求的系统功率 (W) (8)	850	850	650	450	600	600

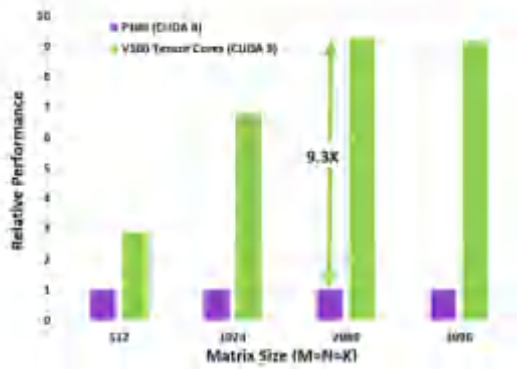
资料来源：英伟达官网，Techhpowerup，长江证券研究所

图 40：Tensor Core 的 4x4 矩阵可大幅提升运算效率



资料来源：英伟达官网，长江证券研究所

图 41：相比无 Tensor Core 的 P100，V100 训练效率大幅提升



资料来源：英伟达官网，长江证券研究所

2022 年，英伟达推出 Hopper 架构，其代表了 NVIDIA 在图形处理和 AI 计算领域的重大进步。Hopper 架构引入了多项先进技术，包括第四代 Tensor Core，这些核心专为 AI 运算而设计，能够显著提升深度学习训练和推理的性能；Hopper 架构还实现了与 NVIDIA Grace CPU 的超级芯片互连，为异构计算提供了强大支持。

表 2：英伟达主要数据中心显卡参数

	P100	V100（SXM2）	A100（80GB SXM）	H100（SXM）
推出时间	2016	2017	2020	2022
制程	16nm	12nm	7nm	4nm
架构	Pascal	Volta	Ampere	Hopper
双精度（TFLOPS）	5.3	7.8	9.7	30
单精度（TFLOPS）	10.6	15.7	19.5	60
显存	16 GB HBM2	32 GB 或 16 GB HBM2	80 GB HBM2e	80 GB
显存带宽（GB/s）	732	900	2039	3000
互联方式	NVLink/PCIe 3.0	NVLink	NVLink/PCIe 4.0	NVLink/PCIe 5.0
互联带宽		300	NVLink 600/PCIe 4.0 64	NVLink 900/PCIe 5.0 128
最大功耗	300W	300	400	700

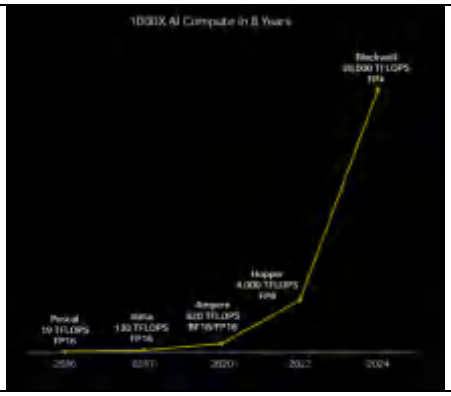
资料来源：英伟达官网，长江证券研究所

目前，英伟达最新的 GPU 架构为 Blackwell，采用 Blackwell 架构的核心产品是 GB200 GPU，它集成了 2080 亿个晶体管，数量是前代 H100 GPU 的两倍多，在 1750 亿参数的 GPT-3 大型语言模型（LLM）基准测试中展现出的性能是 H100 的 7 倍，训练速度则是 H100 的 4 倍。此外，Blackwell 架构的 GPU 支持高达 1.8TB/s 的双向带宽，能够实现多达 576 个 GPU 间的无缝高速通信，显著提升了大规模 AI 系统的性能和效率。Blackwell 架构的优势在于其强大的并行处理能力、超高的内存带宽和容量，以及对 AI 工作负载的优化，使其成为 AI 数据中心和高性能计算领域的理想选择。

图 42：Blackwell 架构下的 GB200 GPU 集成了 2080 亿个晶体管

资料来源：英伟达官网，长江证券研究所

图 43：GB200 的整体运算效率远超英伟达前代产品



资料来源：英伟达官网，长江证券研究所

DPU：DPU 解决 AI 训练推理中设备网络通信与 CPU 负荷问题：英伟达于 2019 年收购 Mellanox，率先推出针对 AI 加速计算的 DPU 数据处理器，构造三芯一体的数据中心新计算架构，BlueField DPU 利用 InfiniBand 技术解决同一系统中不同设备的通信问

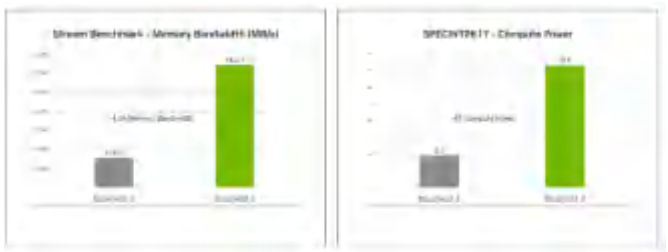
题，可共享 CPU 的网络、存储和安全任务，实际上减轻 CPU 工作负荷。多年配合下游生态构建，对人工智能的算法体系具备深厚理解是 DPU 设计的根本基础，**DPU+DOCA 的定义权与生态圈构建或将复刻 GPU+CODA 的成就。**

多年配合下游生态构建，对人工智能的算法体系具备深厚理解是 DPU 设计的根本基础。英伟达的 DPU 能够承担网络、存储和安全等基础设施任务，从而释放 CPU 和 GPU 资源以专注于更复杂的计算任务。DPU 可以显著提升数据中心的效率，降低能耗，并减少成本。通过集成高性能的多核 CPU、高速网络接口和灵活可编程的加速引擎，实现了对数据中心中网络、存储和安全等基础设施任务的高效处理。此外，DPU 支持先进的 RDMA 技术，提供低延迟和高吞吐量的网络性能，并通过集成的 AI 和机器学习加速器进一步提升数据处理能力。英伟达的 DPU 还具备向后兼容性，支持 DOCA 软件开发平台，使得开发者能够在 DPU 上构建和优化数据中心基础设施应用。这些性能优势共同使得英伟达的 DPU 成为提升数据中心效率、降低运营成本并增强安全性的关键技术。

图 44：NVIDIA BLUEFIELD-3 DPU：可编程片上数据中心基础设施

资料来源：英伟达官网，长江证券研究所

图 45：DPU 可大幅提升通信吞吐量



资料来源：英伟达官网，长江证券研究所

CPU：CPU 填上三芯战略最后一块拼图，GPU 强耦合设计构造完整 AI 解决方案

- 英伟达在 2020 年宣布计划从软银集团手中收购 ARM，交易价值高达 400 亿美元。ARM 是全球领先的半导体知识产权（IP）提供商，其架构广泛应用于移动设备、物联网（IoT）设备和各种嵌入式系统中。英伟达收购 ARM 的目的在于结合 ARM 在能效和设计灵活性方面的优势，以及英伟达在 GPU 和 AI 技术上的领先地位，共同开发新一代计算平台。然而，这笔交易面临了全球反垄断监管机构的严格审查，主要担忧是收购后可能限制 ARM 的中立性，影响整个半导体行业的竞争。最终，在 2022 年，英伟达宣布终止收购 ARM 的交易，原因是监管障碍和市场环境的变化。尽管收购未能实现，**但英伟达与 ARM 继续在技术和产品上保持合作关系。**

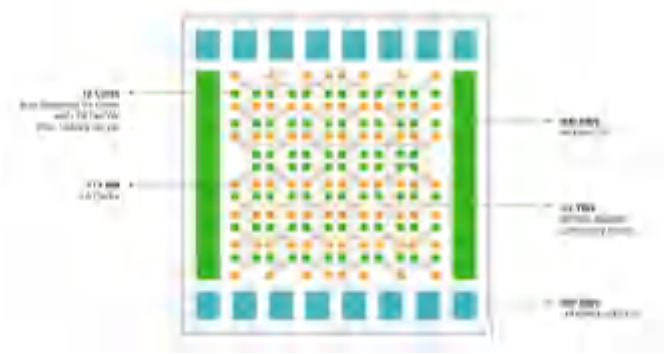
英伟达于 2021 年推出基于 ARM 架构的自研 Grace CPU，面向大型 AI 和 HPC 的高度专业化定制设计主要用于解决 GPU 读取内存数据的带宽瓶颈问题。搭载 Grace CPU 的系统速度相比英伟达前代 DGXM 系统快 10 倍，英伟达 AI 算力提供能力再上大平台。这款 CPU 采用了先进的制程技术和专为数据中心优化的微架构，提供了高吞吐量和低延迟的计算性能。Grace CPU 与英伟达的 Hopper GPU 系列相结合，可以构成 CPU+GPU 的产品形态，通过 NVLink-C2C 互联技术连接，带宽高达 900GB/s，确保了 CPU 和 GPU 之间的高速数据传输，如 GH200 和 GB200，这种结合提供了一致的内存

模型，特别适合于需要处理大量并行任务的高性能计算和 AI 应用，如科学模拟、数据分析和机器学习等，同时 Grace CPU 也是英伟达自动驾驶 SoC 产品线的一部分，集成在 Atlan 和 Orin 芯片中，为智能驾驶和高度互联的汽车提供支持。

图 46：Grace CPU 通过 NVLink 与 GPU 连接，大幅提升吞吐效率

资料来源：CSDN，长江证券研究所

图 47：使用 NVIDIA Scalable Coherency Fabric 扩展内核和带宽



资料来源：CSDN，长江证券研究所

软件层：CUDA+DOCA 构造基础，工具树凝聚生态

英伟达的软件体系可以分为基础架构层和应用工具层，其中基础架构层主要是 CUDA 和 DOCA，RTX 和 Magnum IO 等为辅助

- **CUDA：内存共享的 GPU 硬件调用工具**
- **DOCA：统一部署的 DPU 硬件调用工具**
- RTX：光线追踪和 DLSS 采样插帧
- Magnum IO：存储、网络 IO 配置

应用工具层：两大主体：AI+Omniverse

- Modulus：偏微分物理高保真参数模型
- MonAI：开源医学影像 AI
- Maxine：音视频重编译加速 AI
- NeMo：对话式自然语言模型
- Avatar：AI 虚拟影像
- Drive：自动驾驶
- ISAAC：机器人 AI 训练推理一体开发套件
- Metropolis：视频+传感器融合的开发套件
- Holoscan：医疗设备 AI 开发平台

集群管理层：Kubernetes 云集群管理，以及虚拟 GPU 套件（虚拟服务器、工作站、PC 等）

图 48：英伟达从硬件→软件→应用层的完整结构

资料来源：英伟达官网，长江证券研究所

CUDA：统一计算设备架构（Compute Unified Device Architecture, CUDA），是由英伟达基于 GPU 并行运算特点推出的通用并行计算架构和开发平台。开发者可以将英伟达的 GPU 用于通用的计算处理，而非仅限于图形处理，这使得 GPU 可以直接提供硬件的直接访问接口，而不必像传统方式一样必须依赖图形 API 接口来实现 GPU 的访问，解决的是用更加廉价的设备资源，实现更高效的并行计算。英伟达于 2006 年发布 CUDA 生态系统，投入巨资开发 CUDA 这一软件工具链，让人工智能行业的研究者免费使用该软件来调用 GPU 的计算资源，这使得英伟达成为人工智能中深度学习的训练和推理领域的重要推动者。

CUDA 主要由开发库+运行环境+驱动组成，其中 CUDA 开发库可大幅降低开发者的开发难度；CUDA 运行环境提供了各项开发接口和运行期组件方便开发者调用各类资源接口，进而匹配各种类型计算；CUDA 驱动可理解为 CUDA-Enable 的 GPU 设备抽象层。

表 3：CUDA 主要工作模块及原理

模块	工作原理
编程模型	CUDA 允许开发者使用 C 语言（也可支持 C++和 FORTRAN）编写程序，这些程序可以在支持 CUDA 的处理器上以超高性能运行。
并行处理	CUDA 通过线程并行、数据并行等机制，在 GPU 上实现高效的并行计算。
内存管理	CUDA 提供了灵活的内存管理机制，以优化数据传输和存储。

资料来源：CSDN，长江证券研究所

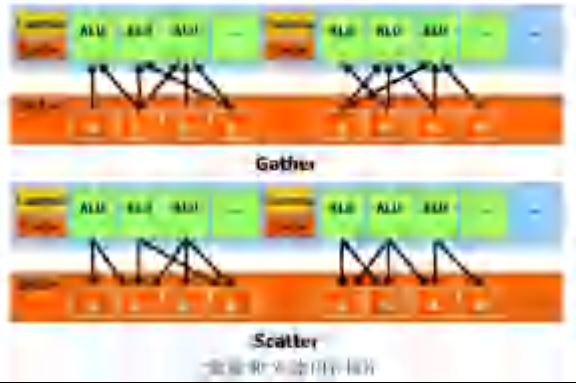
从 CUDA 可扩展的编程模型构成来看，CUDA 通过“线程组层次结构+共享内存+屏障同步”，可帮助程序员将计算问题划分为可以由线程块并行独立解决的粗略子问题，并将每个子问题划分为可以由块内所有线程并行协作解决的更精细的部分。这种分解问题的方法允许线程在解决每个子问题时进行协作来保留语言表达能力，同时实现自动可扩展性。

表 4：CUDA 核心优势

优势	描述
高性能计算	CUDA 能够显著提升计算性能，特别适用于需要大量数值计算和科学计算的任务。
易用性	CUDA 提供了类似于 C 语言的编程接口，使得开发者能够更容易地上手并进行高效的 GPU 编程。
广泛的应用支持	CUDA 已应用于多个 NVIDIA 的 GPU 系列，并在多个领域得到了广泛的应用。
完整的工具链	CUDA 提供了包括性能分析工具、调试器以及样例代码和教程在内的完整工具链，为开发者提供了全面的支持环境。

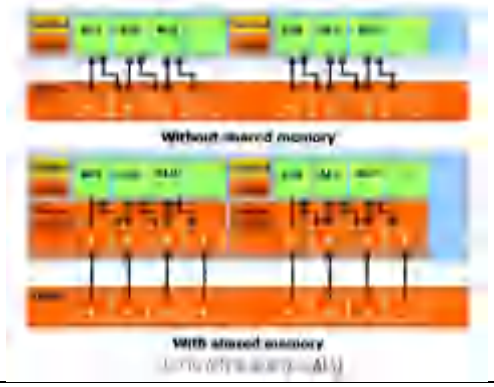
资料来源：CSDN，长江证券研究所

图 49：DRAM 内存寻址：可以在 DRAM 的任何区域进行数据读写



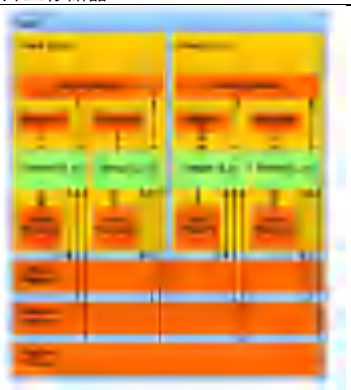
资料来源：英伟达官网，长江证券研究所

图 50：On-chip 内存共享：提升数据读写速度



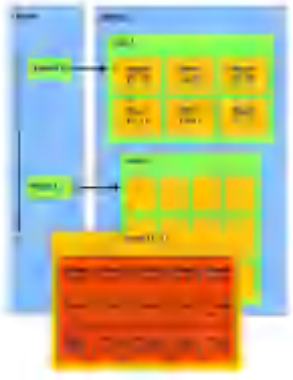
资料来源：英伟达官网，长江证券研究所

图 51：外部内存读取：线程可以通过不同范围的一组内存空间来访问设备的 DRAM 和片上存储器



资料来源：英伟达官网，长江证券研究所

图 52：线程批处理：任务分解



资料来源：英伟达官网，长江证券研究所

依托 CUDA 开发套件向上提炼 CUDA-X AI/CUDA-X HPC 开发套件积极构建 CUDA 软件生态。在 CUDA 软件栈基础上，公司向上抽象和扩展了 CUDA-X，对接不同的行业应用需求。主要包括面向 A 计算的 CUDA-X AI 和面向 HPC 计算的 CUDA-X HPC。此外依托于 CUDA 软件栈进行第三方应用及工具的扩展,形成了广义的 CUDA 生态。从 CUDA 满足易部署(用户开箱即用)、层次灵活的开发接口(OpenCL、OpenGL 类似的一种 API)、满足不同领域开发者编程语言(Fortran,C/C++,Python)、品类齐全的工具

集(GDB、Nsight.Memcheck 等)、第三方工具和软件库(和用户及厂商并肩，构筑软件生态城)。

图 53：CUDA-X AI 开发套件

资料来源：英伟达官网，长江证券研究所

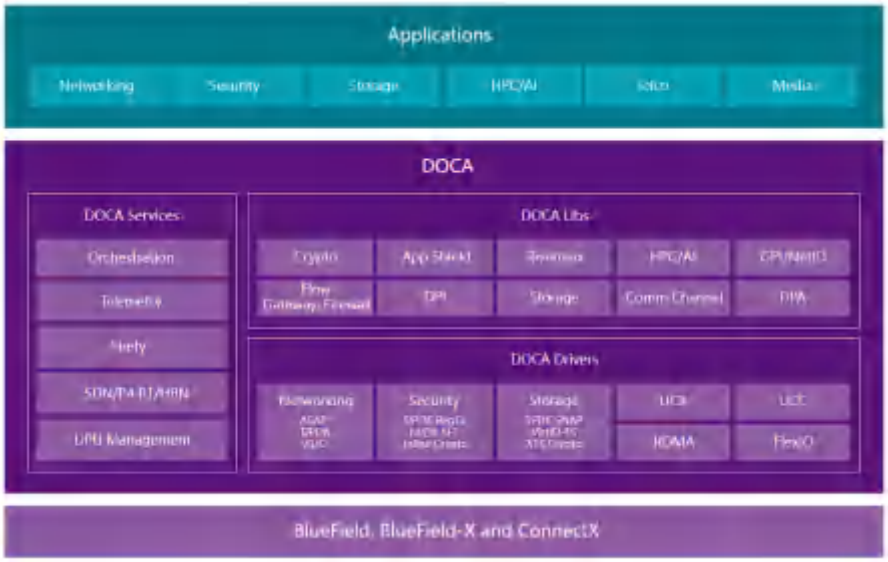
图 54：CUDA-X HPC 开发套件

资料来源：英伟达官网，长江证券研究所

DOCA：**BlueField DPU 是 NVIDIA 推出的一种新型可编程处理器，专注于数据处理，能够满足企业对性能、安全性、可管理等越来越高的需求，英伟达 DOCA 是专为 BlueField DPU 而设计的软件开发套件和加速框架，具备多重优势：**

- 统一访问所有 DPU 功能 – 为开发者节约学习及使用多种不同工具的成本。
- 在 DPU 的底层 API 上提供一个抽象层给上层的库 – 开发者可以更快速、更轻松地进行开发，实现和上层业务的集成，并经优化后提供出色的性能，或者和底层接口合作达到更精细的控制。
- 向前/向后兼容 – 使用 DOCA 开发的应用可在未来版本的 BlueField DPU 上无缝运行，并得到更高的性能和可扩展性。
- 基于容器化服务的 DPU 调配和部署 – DOCA 包含用于简化 DPU 设置、配置和服务编排的工具。

图 55：DOCA 的软硬件结构



资料来源：英伟达官网，长江证券研究所

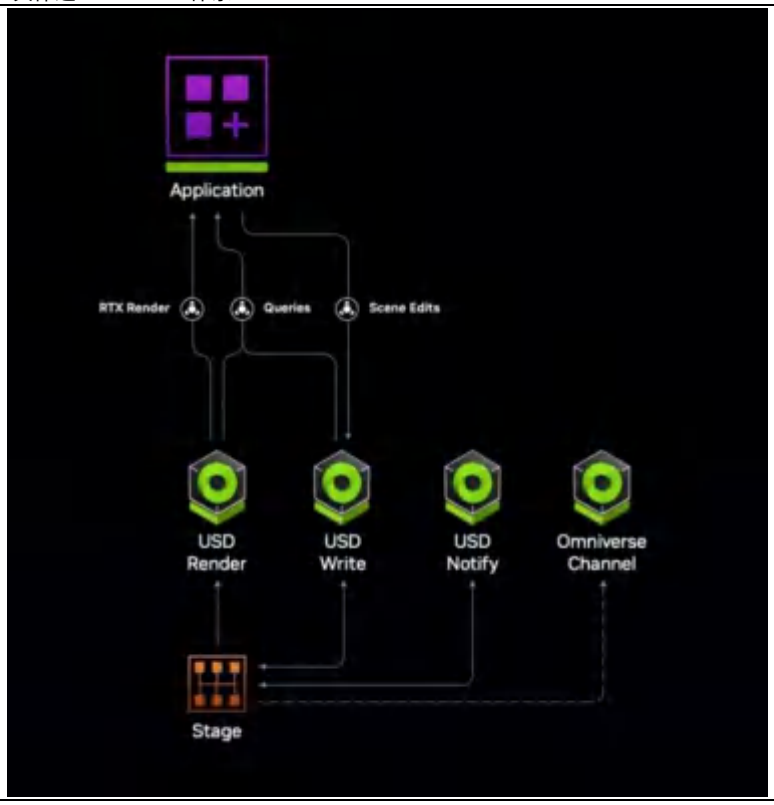
NVIDIA AI Enterprise 加速 AI 模型开发，未来或有望助力实现以 AI 开发 AI：NVIDIA AI Enterprise 是一套端到端的云原生 AI 和数据分析软件套件，使客户能够将 AI 模型的开发时间从 80 周缩短到仅 8 周，并允许客户在 VMware vSphere 上部署和管理高级 AI 应用程序，订阅许可的收费模式进一步拓宽英伟达 SaaS 业务发展空间。

图 56：英伟达 AI Enterprise 应用体系

资料来源：英伟达官网，长江证券研究所

Omniverse 初试工业共享虚拟空间，从硬件→软件→云上社区，在强劲软硬件基础上打造系统级 AI 生态圈：Omniverse 由 Nucleus、Connect、Kit、RTX Render、Simulation 等五大核心部件组成，本质上是一个为设计师、工程师等创造共享虚拟空间，以进行实时协作的云原生技术平台，可以解决数据协同、团队协作、大数据、信息安全等多种痛点。已应用在海内外传媒娱乐，建筑、产品设计、科学运动和仿真、自动驾驶、工业机器人等六大领域。

图 57：英伟达 Omniverse 体系



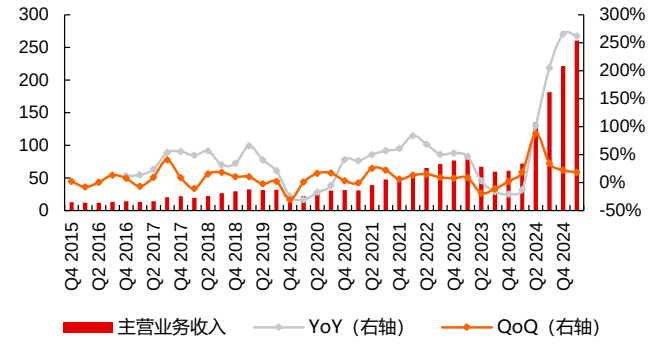
资料来源：英伟达官网，长江证券研究所

收入利润节节高升，长期成长路途清晰

英伟达在游戏、数据中心、专业可视化、自动驾驶等领域的业务发展情况表现出色，最新的财报数据显示了其强劲的增长势头。得益于数据中心业务的强劲表现和 AI 芯片需求的增加，尤其是 A100 和 H100 两款 AI 芯片在下游需求爆发下快速增长的销售规模，英伟达 2025 财年第一季度实现了显著的增长，期间单季度收入达到了 260.44 亿美元，同比增长了 73.8%，其中数据中心营收达到了 225.63 亿美元，环比增长 22.6%，同比增长 426.7%。

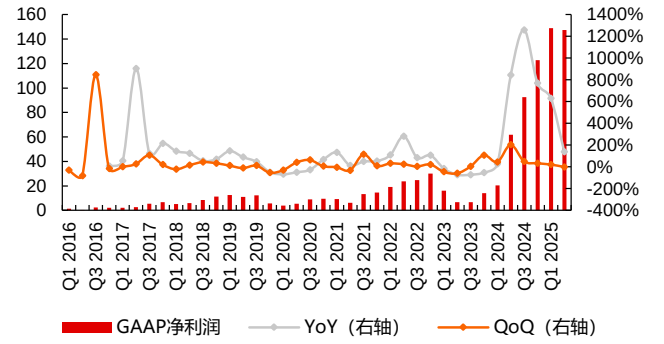
以年度计，在强劲的数据中心收入增长驱动下，英伟达 FY24 收入达 609 亿美元，同比+126%。一方面是高速增长的收入规模，另一方面由于 AI 应用需求的爆发以及 GPU 行业格局的高度集中，英伟达 FY24 的净利润实现了更为惊人的增长，公司 FY24Q4 净利润达 123 亿美元，同比增长了约 7 倍，整个 2024 财年净利润接近 300 亿美元。

图 58：英伟达整体收入及变化（亿美元）



资料来源：Bloomberg，长江证券研究所

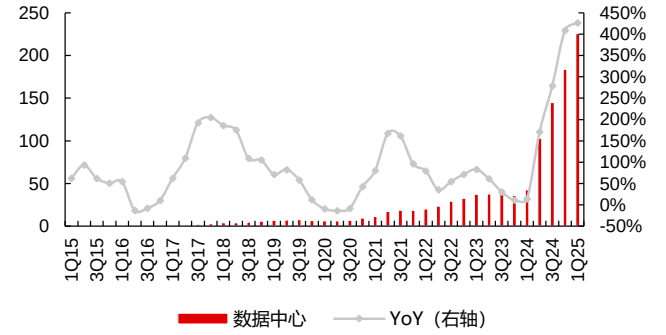
图 59：英伟达归母净利润变化（亿美元）



资料来源：Bloomberg，长江证券研究所

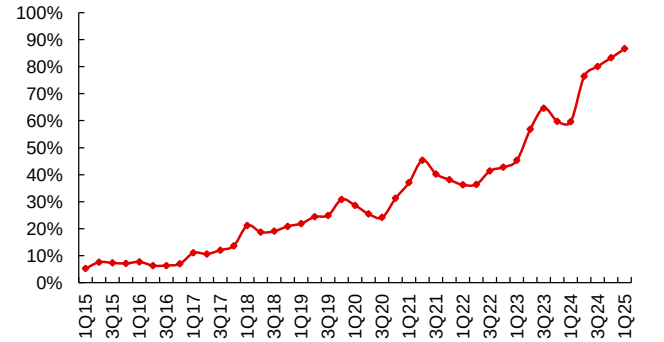
随着 AI 应用的快速发展，数据中心逐步成为英伟达增长的主要动力源泉。在数据中心领域，英伟达的业务收入在 2024 财年达到了 475.25 亿美元，占营业总收入的 78%，同比增长 22.38pct，这一增长主要得益于 AI 的发展及全球云服务提供商的推动。到了 2025 财年第一季度，数据中心营收更是达到了 225.63 亿美元，环比增长 22.6%，同比增长 426.7%，显示出英伟达在数据中心领域的强劲动力和广阔前景。

图 60：英伟达数据中心收入变化（单位：亿美元）



资料来源：Bloomberg，长江证券研究所

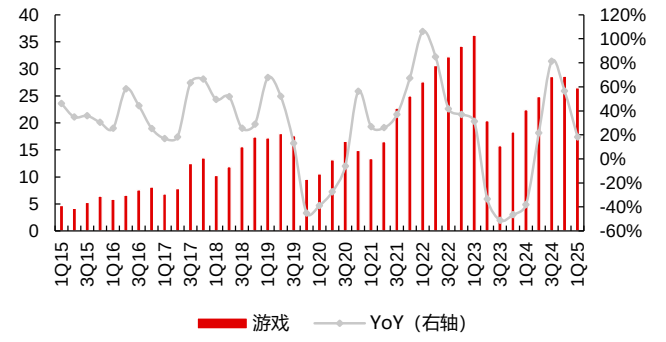
图 61：英伟达数据中心收入占比变化



资料来源：Bloomberg，长江证券研究所

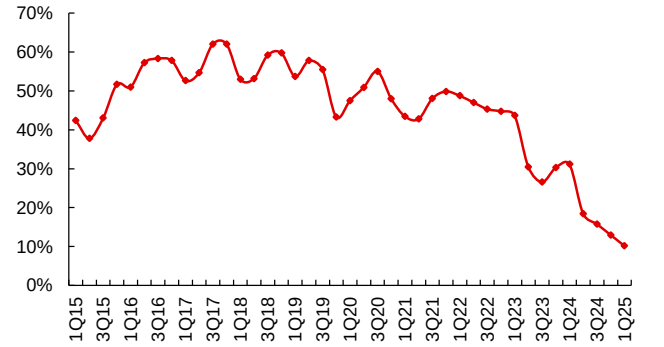
在游戏领域，尽管面临疫情后的复苏挑战，英伟达的游戏业务仍然实现了积极的增长。2025 财年第一季度，游戏相关收入达到 26.47 亿美元，同比+18%，显示出游戏行业逐步回暖的趋势，同时随着 AI 神经渲染能力的新 Ada 架构 GPU 的推出，游戏玩家对高性能 GPU 的需求持续增长。

图 62：英伟达游戏收入变化（单位：亿美元）



资料来源：Bloomberg，长江证券研究所

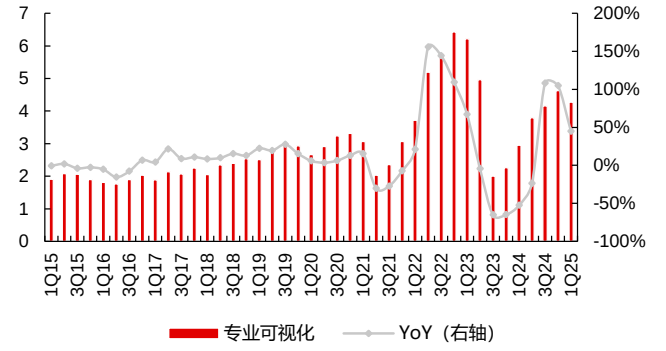
图 63：英伟达游戏收入占比变化



资料来源：Bloomberg，长江证券研究所

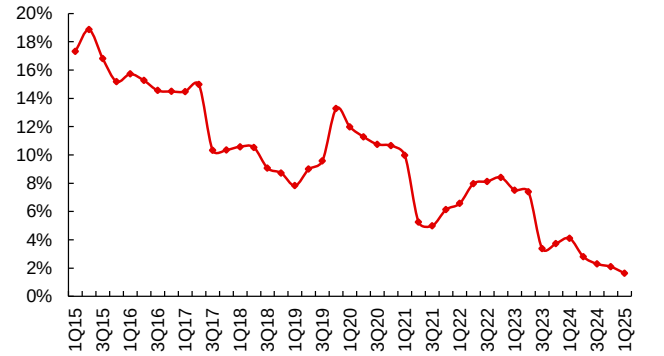
专业可视化方面，英伟达正将其发展成为第三支柱业务。公司不断在软件和硬件解决方案上进行改进，以支持设计和制造领域的需求。虽然专业可视化业务的增长率相对较低，但英伟达仍在积极推动该领域的发展，以期在未来实现更大的增长。

图 64：英伟达专业可视化收入变化（单位：亿美元）



资料来源：Bloomberg，长江证券研究所

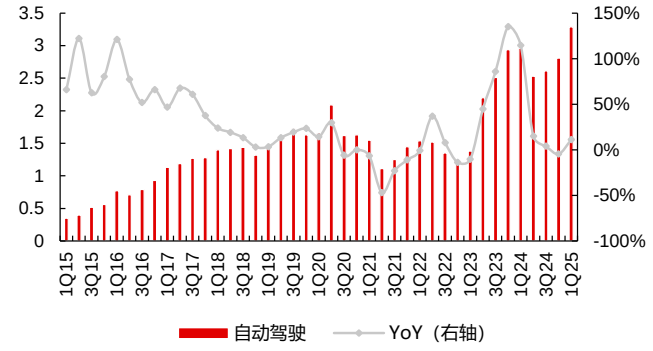
图 65：英伟达专业可视化收入占比变化



资料来源：Bloomberg，长江证券研究所

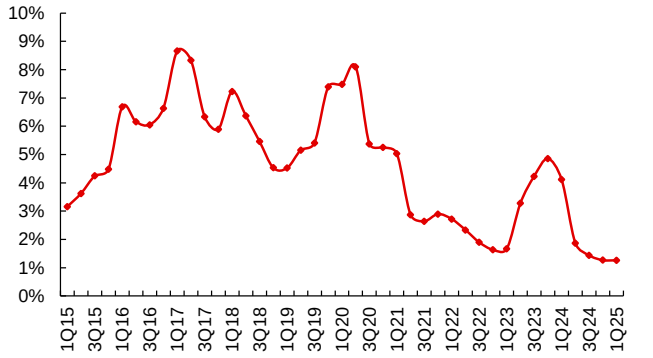
在自动驾驶领域，英伟达提供了全套系统和软件，涵盖数据获取、创建标记、AI 训练等方面，显示出公司在自动驾驶技术方面的深度布局和创新能力。英伟达的自动驾驶平台，如 NVIDIA DRIVE，集成了深度学习、传感器融合和环绕视觉等技术，支持从 L2 到 L5 级别的自动驾驶功能。此外，英伟达的 SoC 产品，例如 Orin 和 Thor，为自动驾驶车辆提供了强大的计算能力。根据英伟达 2025 财年第一季度的财务报告，汽车业务在该季度的收入达到 3.29 亿美元，环比增长 17%，同比也实现了 11%的增长。

图 66：英伟达自动驾驶收入变化（单位：亿美元）



资料来源：Bloomberg，长江证券研究所

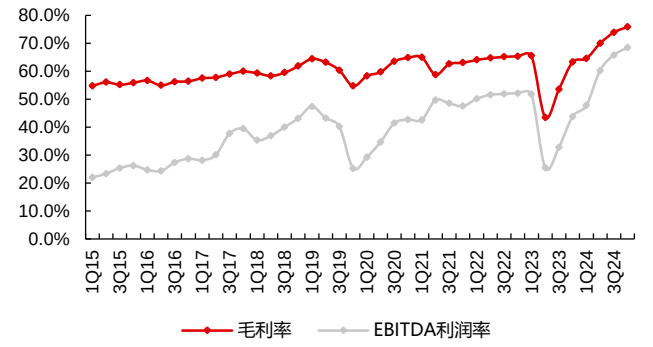
图 67：英伟达自动驾驶收入占比变化



资料来源：Bloomberg，长江证券研究所

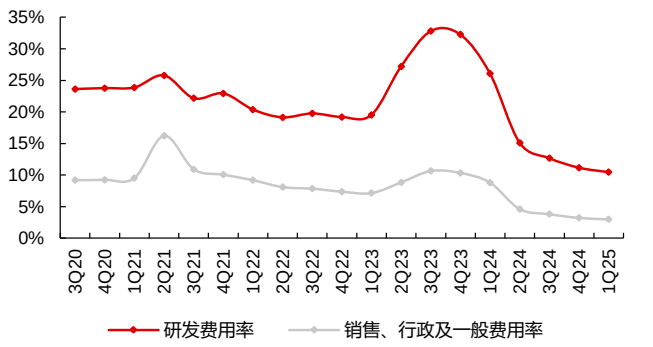
高速增长的 AI 需求+紧缺的供给端带来了较高的单品价值，这为英伟达的盈利能力提升提供的坚实的基础。毛利率方面，英伟达 FY25Q1 的毛利率达到了 78.35%，净利率方面在高毛利率的带动下，英伟达 FY25Q1 净利率达到了 57.23%，同比增长 26.94pct，季度环比增长 1.99pct，这一显著的增长反映了公司在高利润的数据中心业务放量叠加费用率降低的双重影响下利润端的加速回暖。

图 68：英伟达盈利能力



资料来源：Bloomberg，长江证券研究所

图 69：英伟达费用率



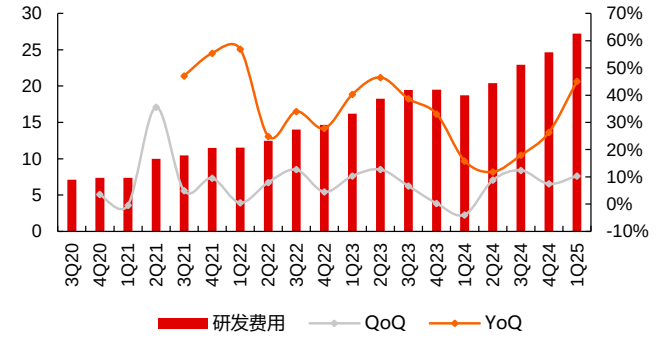
资料来源：Bloomberg，长江证券研究所

英伟达的研发投入和研发效率是其成为半导体行业龙头企业的重要因素之一。英伟达自成立以来便以较高的研发投入和研发效率著称，从 1990 年的 GPU Geforce 256 到如今的 Blackwell 架构，英伟达芯片性能持续提升。2015 年以来其研发投入保持较大规模投入，FY25Q1 研发费用已经超过 27 亿美元。

通过技术进步降低成本和产品价格，英伟达的持续不断推出新的产品吸引更多消费者。例如，CUDA 平台的推出大大降低了利用 GPU 训练神经网络等高算力模型的难度，将 GPU 的应用从 3D 游戏和图像处理拓展到科学计算、大数据处理、机器学习等领域。此外，英伟达还推出了 AI Workbench，旨在为开发大型人工智能项目的公司减少开发时间和成本。在医疗保健领域，英伟达推出的 BioNeMo 平台能够提高研发效率并降低企业运营成本。

伴随着持续不断优化升级的新品体系的完善和半导体需求尤其是 AI 领域的需求回暖，英伟达的存货在近 3 个季度实现同比连续负增长。

图 70：英伟达研发投入（亿美元）



资料来源：Bloomberg，长江证券研究所

图 71：英伟达存货（亿美元）

资料来源：Bloomberg，长江证券研究所

投资建议：

三重壁垒联动+螺旋提升打造 AI 全栈体系，系统级 AI 解决方案大平台将成为英伟达长期增长的关键动力：

1、第一层壁垒：硬件层。GPU 奠定图形渲染和 AI 算力基础，英伟达硬件层的三芯战略已逐步成型。

- GPU 解决 AI 大规模并行运算痛点，且 GPU 当前全球并无具备挑战力的 GPU IP 核授权供应商，后发玩家难以快速跟上技术迭代趋势。
- DPU 解决 AI 训练推理中设备网络通信与 CPU 负荷问题，DPU+DOCA 的定义权与生态圈构建或将复刻 GPU+CODA 的成就。
- CPU 填上三芯战略最后一块拼图，GPU 强耦合设计构造完整 AI 解决方案。
- NVlink+NVSwitch+ConnectX 突破芯片直连和设备网络连接限制。

2、第二层壁垒：软件层。CUDA 释放 GPU 潜力引航 AI 发展，DOCA、Omniverse 等软件层进一步填充生态，增强 AI 行业对英伟达的粘性。

- CUDA 从底层代码出发发挥 GPU 并行运算优势，奠定近十年人工智能发展基础。
- DOCA 为 BlueField DPU 量身定做软件开发平台，复刻 GPU+CUDA 的强耦合成功路径。
- Omniverse 初试工业共享虚拟空间，从硬件→软件→云上社区，在强劲软硬件基础上打造系统级 AI 生态圈。
- NVIDIA AI Enterprise 加速 AI 模型开发，未来或有望助力实现以 AI 开发 AI。

3、第三层壁垒：应用层。游戏显卡、数据中心、自动驾驶、元宇宙先后接力，十年成长曲线浪潮叠加。

- PC/NB 游戏显卡性能稳定提升，AI 图形生成或将成为元宇宙关键底层技术，打开新的成长空间。
- 数据中心将超越游戏业务成为公司支柱业务，Blackwell 架构 GPU 套片进一步拉开竞争差距。
- GPU 完美适配自动驾驶视觉方案，英伟达新成长曲线逐步进入释放期。
- 专业设计业务完成迈向元宇宙的 Ominiverse 平台基建。
- 机器人软硬件平台前瞻布局未来社会生产力构成。

风险提示

1、下游需求不及预期的风险。当前英伟达主要下游领域为 AI、游戏、工业可视化、自动驾驶，整体需求覆盖较为广泛，但近年来主要增长动力来自 AI 应用，若 AI 行业增长不及预期可能会对公司相应业务增速造成负面影响。

2、全球政治经济动荡影响产品区域性出货的风险。当前全球政治经济形势持续动荡，对公司产品制造、销售均可能带来区域性的限制影响。

投资评级说明

行业评级	报告发布日后的 12 个月内行业股票指数的涨跌幅相对同期相关证券市场代表性指数的涨跌幅为基准，投资建议的评级标准为：	
	看　　好：	相对表现优于同期相关证券市场代表性指数
	中　　性：	相对表现与同期相关证券市场代表性指数持平
	看　　淡：	相对表现弱于同期相关证券市场代表性指数
公司评级	报告发布日后的 12 个月内公司的涨跌幅相对同期相关证券市场代表性指数的涨跌幅为基准，投资建议的评级标准为：	
	买　　入：	相对同期相关证券市场代表性指数涨幅大于 10%
	增　　持：	相对同期相关证券市场代表性指数涨幅在 5%～10%之间
	中　　性：	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
	减　　持：	相对同期相关证券市场代表性指数涨幅小于-5%
	无投资评级：	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

相关证券市场代表性指数说明：A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准。

办公地址

上海
Add /虹口区新建路 200 号国华金融中心 B 栋 22、23 层
P.C /（200080）

武汉
Add /武汉市江汉区淮海路 88 号长江证券大厦 37 楼
P.C /（430015）

北京
Add /西城区金融街 33 号通泰大厦 15 层
P.C /（100032）

深圳
Add /深圳市福田区中心四路 1 号嘉里建设广场 3 期 36 楼
P.C /（518048）

分析师声明

本报告署名分析师以勤勉的职业态度，独立、客观地出具本报告。分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点。作者所得报酬的任何部分不曾与，不与，也不将与本报告中的具体推荐意见或观点而有直接或间接联系，特此声明。

法律主体声明

本报告由长江证券股份有限公司及/或其附属机构（以下简称「长江证券」或「本公司」）制作，由长江证券股份有限公司在中华人民共和国大陆地区发行。长江证券股份有限公司具有中国证监会许可的投资咨询业务资格，经营证券业务许可证编号为：10060000。本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格书编号已披露在报告首页的作者姓名旁。

在遵守适用的法律法规情况下，本报告亦可能由长江证券经纪（香港）有限公司在香港地区发行。长江证券经纪（香港）有限公司具有香港证券及期货事务监察委员会核准的“就证券提供意见”业务资格（第四类牌照的受监管活动），中央编号为：AXY608。本报告作者所持香港证监会牌照的中央编号已披露在报告首页的作者姓名旁。

其他声明

本报告并非针对或意图发送、发布给在当地法律或监管规则下不允许该报告发送、发布的人员。本公司不会因接收人收到本报告而视其为客户。本报告的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证，也不保证所包含信息和建议不发生任何变更。本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况，报告接收者应当独立评估本报告所含信息，基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。本公司已力求报告内容的客观、公正，但文中的观点、结论和建议仅供参考，不包含作者对证券价格涨跌或市场走势的确定性判断。报告中的信息或意见并不构成所述证券的买卖出价或征价，投资者据此做出的任何投资决策与本公司和作者无关。本研究报告并不构成本公司对购入、购买或认购证券的邀请或要约。本公司有可能会与本报告涉及的公司进行投资银行业务或投资服务等其他业务(例如:配售代理、牵头经办人、保荐人、承销商或自营投资)。

本报告所包含的观点及建议不适用于所有投资者，且并未考虑个别客户的特殊情况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。投资者不应以本报告取代其独立判断或仅依据本报告做出决策，并在需要时咨询专业意见。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据；在不同时期，本公司可以发出其他与本报告所载信息不一致及有不同结论的报告；本报告所反映研究人员的不同观点、见解及分析方法，并不代表本公司或其他附属机构的立场；本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。本公司及作者在自身所知情范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。