AI 时代的算力领军人

2024 年 07 月 08 日

➢ **GPU+DPU+CPU 全面布局，AI 软硬件一体化龙头优势领先。英伟达成立于 1993 年，构建以 CUDA 生态为核心，将底层 GPU+DPU+CPU 进行整合，提供数据中心的全领域加速计算方案，在持续推进的 AI 浪潮中，成为了软硬件一体化的云计算解决方案领导者。**公司业务布局包括数据中心、游戏、专业可视化、汽车四大板块，数据中心业务伴随着 AI 浪潮高速成长，2025 财年第一季度实现营收 225.63 亿美元，同比增长 426.68%，同时公司其他几大业务板块业绩也在稳步提升，为公司整体营收和利润带来正向贡献。2024 财年公司实现营收 609.22 亿美元，同比+125.85%，实现净利润 297.60 亿美元，同比+581.32%；2025 财年第一季度公司业绩进一步提速，实现营收 260.44 亿美元，同比+262.12%，净利润为 148.81 亿美元，同比+628.39%。**AI 浪潮仍处于训练阶段，伴随下游客户需求增长，公司业绩有望持续提升。**

➢ **云商资本开支不断上修，反映出下游对 AI 的需求强劲。**此前市场担忧 AI 训练的需求会随着大模型训练告一段落而放缓，而 **2024 年以来北美四大云商纷纷上修全年资本开支指引**，2024 年一季度四大云商合计资本开支 476 亿美元，同比增长 38.98%，反映出客户对训练算力的需求仍在超预期。当前阶段来看，训练侧的 OpenAI 等公司推出的大模型参数仍在呈指数级增长，而 **Sora 等应用领域的模型推出，以及苹果接入 GPT-4o 等则反映出 AI 应用落地节奏的加速，有望带来新一轮的推理侧算力需求。**目前来看，AI 进展仍处于早期阶段，生成式人工智能的推进将给英伟达带来持续增长的下游需求。

➢ **GB200 等产品大幅加强了英伟达在 AI 算力领域的领先优势。此前英伟达在算力领域的核心优势在于强大的硬件性能以及完善的软件生态壁垒。**2023 年以来，AMD、Intel 等芯片设计公司，以及谷歌、微软、Meta、亚马逊等英伟达的下游客户纷纷增强加速卡领域的布局，例如 AMD 加大 ROCm 生态的投入，并且推出 MI300X 等更高性价比的产品，意图抢占英伟达的市场份额。而英伟达的 H200、Blackwell 平台以及 GB200 等产品则进一步加强了英伟达的领先优势。**英伟达通过 H200 等产品给客户提供更多更具性价比的选择的同时，又通过 GB200 在服务器架构上的优化，解决了 GPU 之间的互联瓶颈，大幅提升了英伟达服务器的推理和训练能力，也改善了英伟达的产业链上下游话语权。**展望未来，我们认为英伟达在算力领域的优势和壁垒有望被不断巩固。

➢ **投资建议：**AI 浪潮正在加速，云厂商在训练侧的需求持续提升，并且后续的推理场景有望带来更大的算力市场空间。英伟达在 AI 领域的布局领先，同时与竞争对手的差距逐步扩大。**在生成式人工智能给英伟达数据中心业务带来持续成长动力的同时，公司的其他三大业务板块也在稳步增长。展望未来，公司业绩有望持续提升，建议积极关注。**

➢ **风险提示：**AI 行业需求波动的风险；行业竞争格局变化的风险；产品研发进度不及预期的风险；宏观经济及下游需求恢复不及预期的风险。

重点公司盈利预测、估值与评级

代码	简称	股价	EPS（美元）			PE（倍）		
		(美元)	FY2024A	FY2025E	FY2026E	FY2024A	FY2025E	FY2026E
NVDA.O	英伟达	125.83	1.20	2.72	3.67	105	46	34

资料来源：Bloomberg，民生证券研究院；（注：股价为 2024 年 7 月 5 日收盘价；公司数据采用 Bloombe 一致预期）

推荐 维持评级

分析师 方竞
执业证书：S0100521120004
邮箱：fangjing@mszq.com

分析师 易永坚
执业证书：S0100523070002
邮箱：yiyongjian@mszq.com
分析师 宋晓东
执业证书：S0100523110001
邮箱：songxiaodong@mszq.com

相关研究

1.电子行业周报：AI 终端新趋势：散热、耳机、AR-2024/07/07
2.半导体行业点评：长鑫金桥扩产，看好存储封测产业链机遇-2024/06/30
3.电子行业点评：全球视角，探讨晶圆厂投资-2024/06/24
4.电子行业 2024 年中期投资策略：AI 产业的新范式-2024/06/18
5.电子行业点评：WWDC 召开，三大视角探索苹果端侧 AI 创新-2024/06/12

目录

1 英伟达：人工智能时代 AI 算力领军者

1.1 业务持续转型升级，全面拥抱人工智能

NVIDIA 成立于 1993 年，通过构建以 CUDA 生态为核心，将底层 GPU+DPU+CPU 进行整合，提供数据中心的全领域加速计算方案，并结合 AI+云计算的时代背景，**成为人工智能时代软硬件一体化 AI+云计算解决方案领导者。**其发展历程主要可分为以下几个阶段：

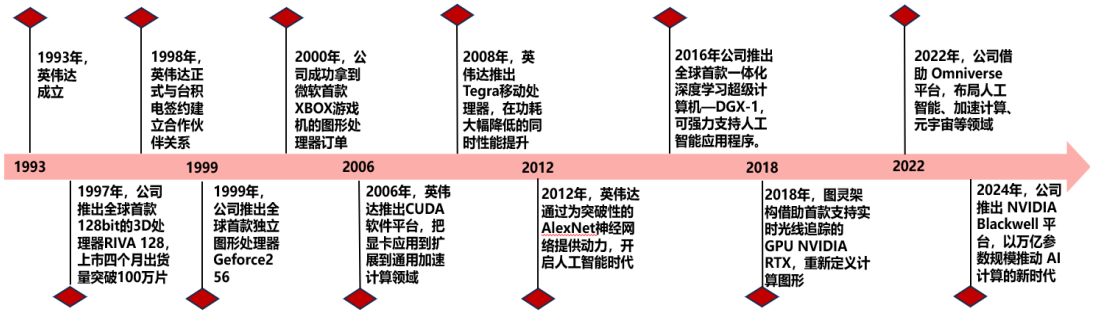
（1993-1998）起步发展阶段：NVIDIA 在成立之初通过生产 3D 图形芯片布局游戏和多媒体市场。1997 年 4 月，公司推出全球首款 128bit 的 3D 处理器 **RIVA 128**，这是第一款支持微软 Direct3D 加速的图形芯片，也是当时市场上唯一真正具有真正 3D 加速能力的 2D+3D AGP 显卡，上市四个月出货量突破 100 万片。1998 年，英伟达正式与台积电签约建立合作伙伴关系。1999 年公司推出的 **RIVA TNT2** 在制程工艺升级的同时核心频率和显存容量皆有极大提升，并以价格优势扩大市场份额，帮助公司成为行业龙头。

（1999-2005）快速成长阶段：英伟达于 1999 年纳斯达克在上市，发行价为 12 美元，发行数量为 350 万股。同年推出 **Geforce256**，这是全球真正意义上第一块 GPU 独立显卡。2000 年 3 月，英伟达成功拿到微软首款 XBOX 游戏机的图形处理器订单，保持图形处理领域的领先优势，游戏产业和英伟达显卡技术进步互相促进。

（2005-2016）转型拓展阶段：2006 年，英伟达推出 **CUDA** 软件平台，展示 GPU 强大的并行处理能力，把显卡的边界从游戏和 3D 图像处理扩展到通用加速计算的领域。在此阶段英伟达全面完善产品线，产品覆盖高中低端下游各应用市场，如 GeForce 系列显卡针对 PC 游戏市场，Tesla 和 Quadro 系列 GPU 被广泛应用于机器学习、数据科学、计算机视觉等领域，DRIVE 高级驾驶辅助系统助力公司布局自动驾驶领域。

（2016-至今）AI 加速阶段： 2016 年公司推出全球首款一体化深度学习超级计算机——DGX-1，搭载 8 块 P100 加速器的超算，可强力支持人工智能应用程序，正式开启人工智能时代。2018 年，英伟达推出“**Turing (图灵)**”GPU 架构，重新定义计算图形，并为全球首款具备实时光线追踪功能的 GPU（即 RTX 8000）提供支持。近年来随着 AI 算力迭代升级，公司借助 **NVIDIA Omniverse** 平台，广泛覆盖人工智能、加速计算、元宇宙、游戏、机器人、自动驾驶等领域，为公司的长远发展拓展更多可能。

图1：英伟达发展历程



资料来源：英伟达官网，民生证券研究院整理

产品线方面，英伟达在数据中心、游戏、专业可视化、汽车四大业务板块均有业务布局。

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、用于其他游戏机的 GeForce NOW 云游戏、用于在电视播放高质量流媒体的 SHIELD，并与游戏方合作提供开发服务。截至 2024 年 3 月，公司在游戏 GPU 领域的市占率超过 80%。

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。

软件方面，英伟达拥有完善的底层基础框架和上层应用工具。CUDA、DOCA 等底层应用框架降低芯片产品的编程难度，更好实现全平台融合，Geforce Now、Omniverse、AI Enterprise、Drive 等应用平台助力公司开拓不同商务场景，英伟达与不同终端客户的深度合作可以让英伟达及时预测技术发展的方向，为公司的可持续发展提供助力。

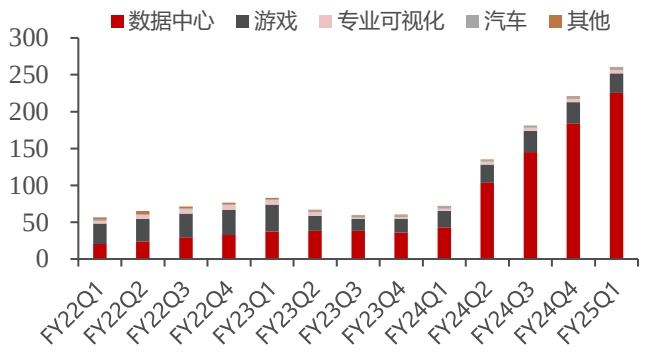
图2：英伟达软硬件产品线

资料来源：英伟达官网，民生证券研究院

1.2 AI 芯片驱动高速增长，软硬件结合构建竞争壁垒

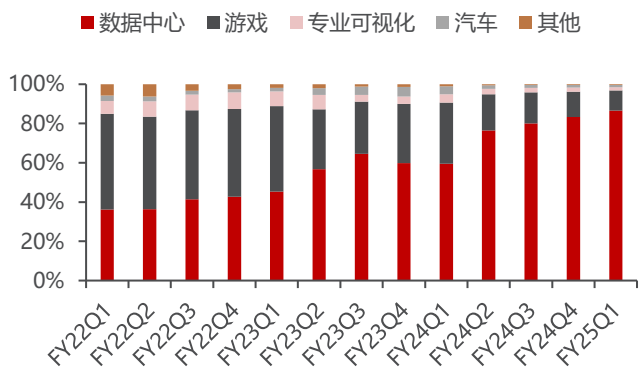
英伟达业务结构可分为数据中心、游戏、专业可视化、汽车和机器人四大板块。数据中心营收占比较高，且近几年公司业绩增长主要由数据中心板块贡献，而游戏和可视化业务保持稳定。数据中心 FY25Q1 营收 225.63 亿美元，占比 86.63%，同比增长 426.68%。游戏业务 FY25Q1 营收 26.47 亿美元，占比 10.16%，同比增长 18.17%。专业可视化业务 FY25Q1 营收 4.27 亿美元，占比 1.64%，同比增长 44.75%。汽车和机器人业务 FY25Q1 营收 3.29 亿美元，占比 1.27%，同比增长 11.15%。从 FY22Q1 到 FY25Q1，数据中心业务营收占比从 36.18%提升到 86.92%，上升 50.74pct；游戏业务营收占比从 48.76%减少到 10.16%，下降 38.60pct。专业可视化业务营收从 FY22Q1 的 3.72 亿美元增加到 4.27 亿美元，汽车业务营收从 FY22Q1 的 1.54 亿美元增加到 3.29 亿美元，**随着 AI 技术的不断发展，数据中心对公司的重要性不断提升，该业务的营收金额和营收占比有望持续走高。**

图3：FY2022-FY2025Q1 公司各业务营业收入（亿美元）



资料来源：Bloomberg，民生证券研究院

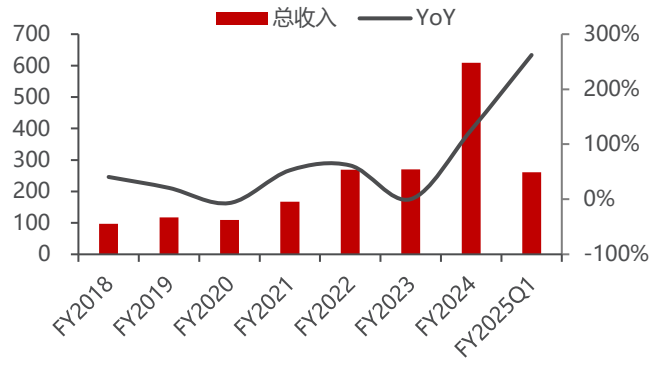
图4：FY2022-FY2025Q1 公司各业务收入占比（%）



资料来源：Bloomberg，民生证券研究院

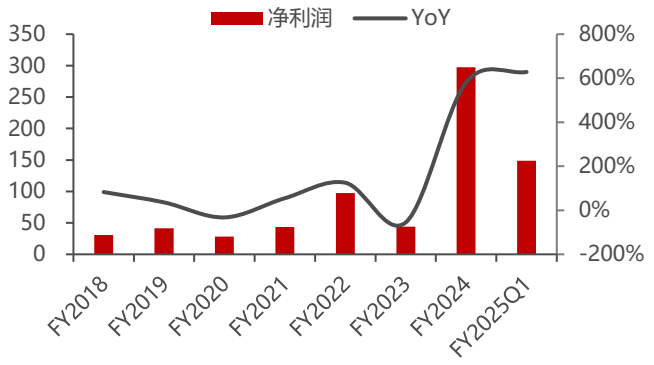
得益于人工智能和高性能计算前所未有的需求水平，英伟达 2024 财年总营收达到 609.22 亿美元，同比增长 125.85%，净利润为 297.60 亿美元，同比增长 581.32%。2024 年一季度公司推出的 Blackwell 平台以万亿参数规模推动 AI 计算的新时代，该季度营收为 260.44 亿美元，同比增长 262.12%，净利润为 148.81 亿美元，同比增长 628.39%。展望 2024 年，随着 Blackwell 在二季度开始量产，并与客户构建数据中心，丰富的软硬件生态系统会加快公司技术更新的速度，业绩有望继续增长。

图5：FY2018-FY2025Q1 公司总营收及同比（亿美元，%）



资料来源：Wind，民生证券研究院

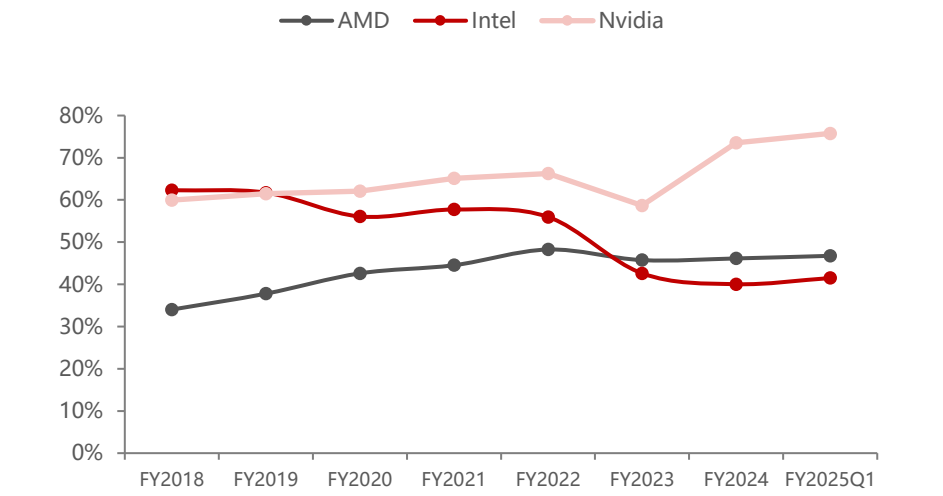
图6：FY2018-FY2025Q1 公司净利润及同比（亿美元，%）



资料来源：Wind，民生证券研究院

受益于产品技术升级，近年来公司毛利率不断改善。近年来公司不断升级芯片架构，并布局加速计算、人工智能等领域，带动毛利率稳健提升并领先可比公司。2023 财年公司毛利率为 58.68%，同比下降 7.56pcts，原因是 PC 需求衰退以及当年全球经济下滑。2024 财年公司毛利率为 73.56%，同比提升 14.88pcts，FY2025Q1 公司毛利率为 75.77%，主要原因为数据中心业务发展带动，公司推出 H100、B100 等高端产品。

图7：FY2018-FY2025Q1 英伟达与可比公司毛利率（%）



资料来源：Bloomberg，民生证券研究院

2 AI 开启大算力时代，加速卡需求井喷

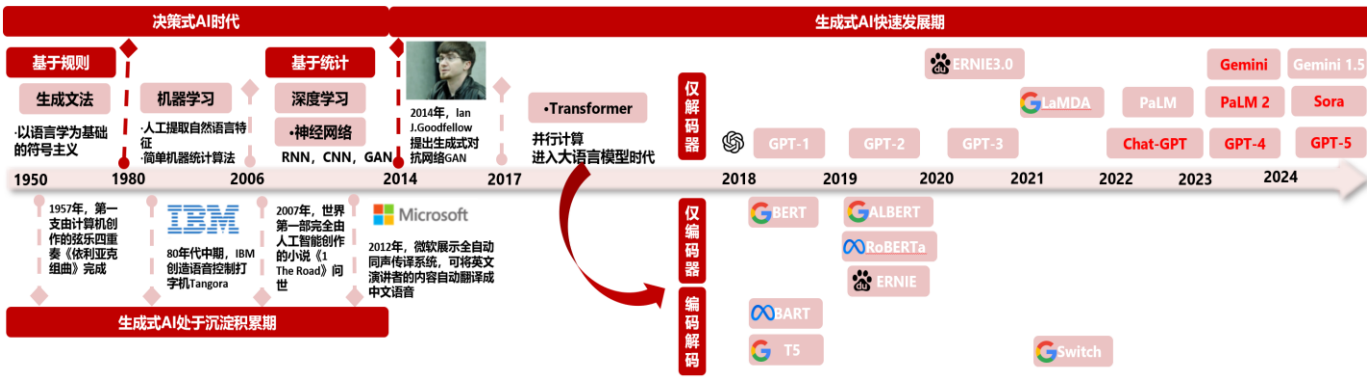
2.1 生成式 AI 带动加速卡需求快速增长

2.1.1 生成式 AI“涌现”，加速卡需求突增

近年来生成式 AI 步入快速发展期。1950 年开始生成式 AI 出现早期萌芽，此后 AIGC 处于漫长的沉淀积累期，决策式 AI 占据主流。随着 2014 年生成式对抗网络等深度学习算法的提出，AIGC 步入快速发展期，生成内容的丰富度和精度都有了较大的提升。英伟达 CEO 黄仁勋在 GTC 2023 大会上将 ChatGPT 比作 AI 的“iPhone“时刻，AI 时代加速来临，推动生成式 AI 加速发展。

多模态大模型有望带动算力需求的进一步增长。伴随着大模型的主要交互方式从文字升级为图片、音频以及视频，大模型对训练和推理的算力需求将进一步提升。谷歌发布的 Gemini 模型开启了大模型的多模态时代，2024 年 2 月 16 日，谷歌发布 Gemini 1.5，模型能力可以支持 100 万 Token 上下文的稳定处理。同一时间，Open AI 发布了 sora 模型，在文生视频领域获得了重要突破，60 秒的视频长度和对真实世界物理引擎的更优理解，有望带动大模型视频生成行业的快速发展。大模型向视频等交互模式的升级有望带动训练侧算力需求的进一步提升，同时这些表现惊人的模型或将加速生成式 AI 在应用侧的落地，加速推理侧算力需求的增长。英伟达在 4Q23 业绩交流会上表示，目前来自推理侧的需求占比已经达到 40%，伴随模型能力的进一步提升，推理侧算力需求的占比有望持续提升。

图8：AI 模型发展历程



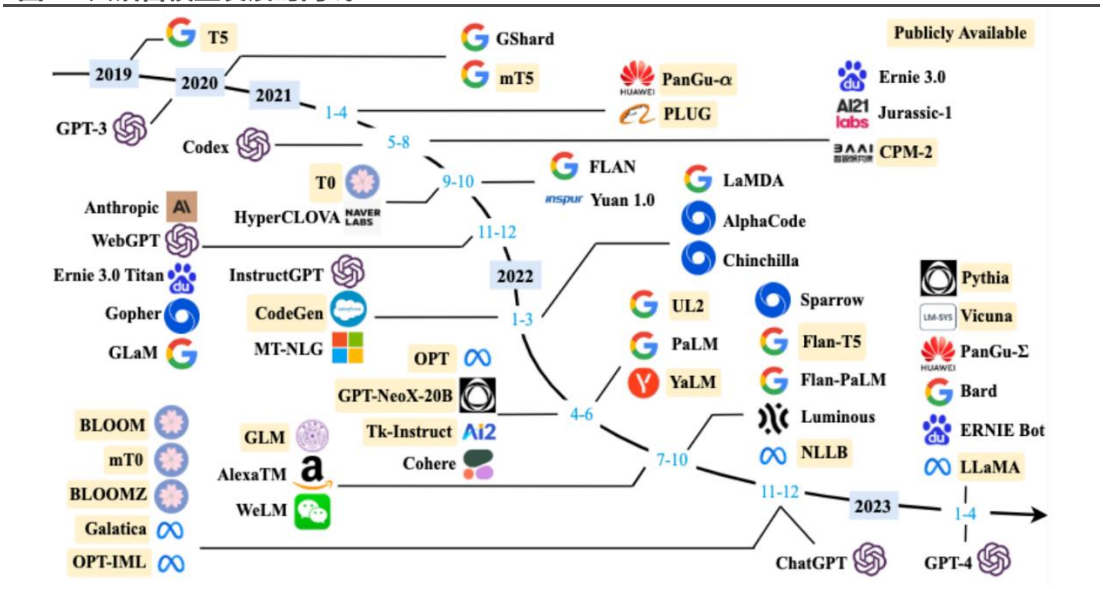
资料来源：中国信通院《人工智能生成内容白皮书》，民生证券研究院

OpenAI 成立于 2015 年，从非盈利组织逐步转变为商业驱动。公司在 2018-2020 三年时间内连续推出了 GPT-1、GPT-2 和 GPT-3 三款产品，后在 2022 年底推出 ChatGPT，面向公众提供生成式 AI 服务，引起全球 AI 浪潮。2023 年 OpenAI 推出万亿参数量级的 GPT-4，能力再上一个台阶。**2024 年 2 月，公司推出的 Sora 文生视频模型相较此前的同类型模型有着全方位的能力提升，引发全球轰动。公司计划后续将推出 GPT-5 大模型，预计参数数量将达到 10 万亿量级，有望进一步提升全球算力需求。**

谷歌早在 2023 年 2 月就推出了对标 GPT 的 Bard 大语言模型，此后在 2023 年 5 月发布了 PaLM 2 模型。**谷歌最重要的大模型产品是在 2022 年 12 月推出的 Gemini 1.0 家族**，该模型是一款具有突破性意义的多模态 AI 大模型，可以处理文本、代码、图像、音频、视频，Gemini 有三个子模型，其中 Nano 被用于谷歌的 Pixel 8 Pro 智能手机中。**2024 年 2 月，谷歌发布了 Gemini 1.5 模型。相较于此前 Gemini 1.0Pro 版本，Gemini 1.5 pro 将上下文处理能提从 3.2 万 Tokens 提升至 100 万以上**，处理能力涵盖包括 1 小时的视频、11 小时的音频、超过 3 万行代码或超过 70 万字的代码库。同月，谷歌又推出了 Gemma 开源大模型，更加轻量化，保持免费且允许商用。

Meta 是全球最重要的开源大模型厂商，主要目的为用开源的方式快速搭建自身的生态，从而在未来更方便地获取数据和推广应用。Meta 最重要的大模型产品是 LLaMa 模型家族，众多大模型厂商在 LLaMa 模型的基础上进行训练和微调，生成自己的大模型。2023 年 7 月，公司推出 LLaMa2 模型，训练数据集达到 2 万亿 token，涵盖 7B、13B 和 70B 三个模型。据 Meta，公司计划在 2024 年 7 月发布 LLaMa3 模型，全球开源大模型能力有望全面提升。

图9：大语言模型发展时间线



资料来源：中国人民大学高瓴人工智能学院，民生证券研究院整理

根据大模型的运算原理，训练和推理所需的算力与模型参数成正比例关系，GPT5 有望带动大模型训练和推理需求的进一步增长。此前市场担忧在大模型在参数指数级提升的情况下，模型能力提升的边际效应是否会减弱，甚至停止，而验证的方法就是看 2024 年将要推出的 GPT5 的能力是否出现质变。Altman 近期在采访中表示，GPT5 的能力相较于 GPT4 将会是一个重大进步，并且他认为目前的大模型能力仍然处于初级阶段，在未来 5-10 年内，模型的能力提升仍将保持一条陡峭的曲线。伴随十万亿参数量级的 GPT5 推出，全球最强的大模型能力和参数再上一个台阶，意味着用于训练大模型的算力需求也将随之提升，同时 H200、B100 等加速卡依次推向市场，或将带动云厂商新一轮的算力军备竞赛。

图10：大模型训练和推理所需算力成本公式

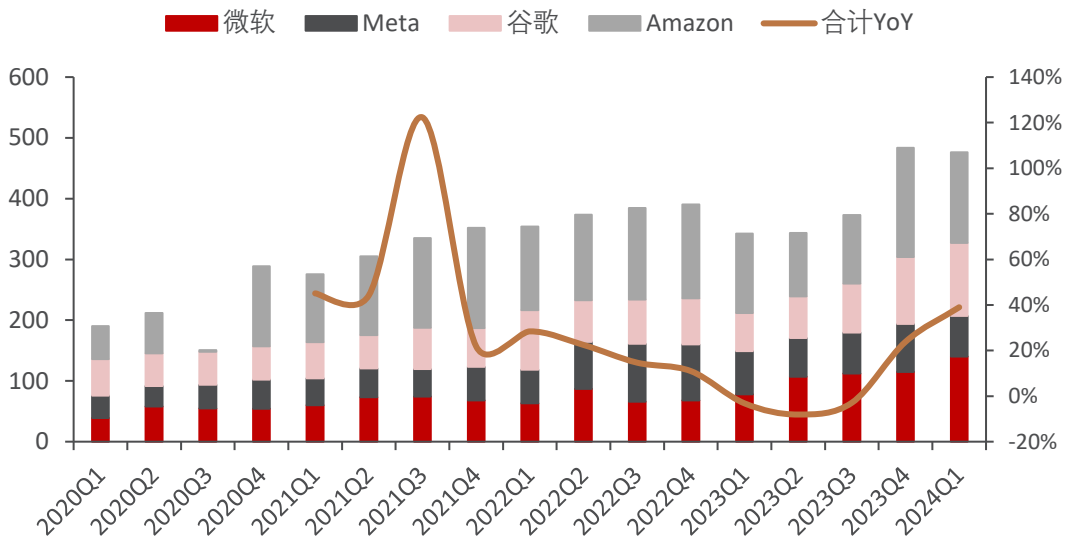
资料来源：民生证券研究院整理

2.1.2 云商算力竞赛加速，资本开支快速提升

我们复盘了北美云厂商过去 4 年的 Capex 走势，虽增速逐年放缓但整体保持逐年增长趋势，但伴随 AI 的强势增长，结合已发布 2024 年第一季度财报四家公司的指引，2024 年北美云商资本开支有望重回高速增长态势。云厂商目前已成为英伟达下游加速卡最大的需求来源，由于全球经济疲软，2023 年云厂商资本开支增速有所放缓，前三季度 Meta、谷歌、亚马逊三家云厂商合计资本开支均为负增长，但受益于 AI 带来加速卡的大量需求，2023 年四季度开始云厂商快速加大资本开支，据 Bloomberg，2023 年四季度北美四大云厂商合计资本开支为 483.67 亿美元，同比增长 23.84%，在 2023 年前三个季度合计资本开支持续同比下滑的情况下大幅转正。

2024 年第一季度北美四大云厂商合计资本开支为 476 亿美元，同比增长 38.98%，展望 2024 年，多家云厂商分别上修全年资本开支指引，云厂商 Capex 有望重回高增。2024 年第一季度微软、谷歌的 Capex 分别同比高增 79%和 91%，并指引 24 年后续季度 Capex 超过 Q1，**则我们预计微软和谷歌 24 年 Capex 分别超过 560 亿美元和 480 亿美元，同比分别增长 36%和 49%；**Meta 2024 年第一季度的 Capex 虽然同比略有下滑 5%，但公司上调全年 Capex 至 350-400 亿美元（前次指引为 300-370 亿美元），**因此 Meta 24 年 Capex 指引中值较 23 年增长 33%；Amazon 指引 2024 年第一季度为 2024 年公司资本开支最低的一个季度，后续几个季度仍将保持增长。**

图11：2020-2024 年第一季度北美云商资本开支（含融资租赁）（亿美元）



资料来源： Bloomberg，民生证券研究院整理

2.2 各环节厂商入局，“一超多强”时代开启

2.2.1 以 AMD 为首的竞争对手发力加速卡

目前英伟达的竞争对手主要有两类，一类是芯片设计公司，包括 AMD、Intel 等，另一类是下游客户，包括谷歌等北美云商，以及中国的互联网公司。目前来看 AMD 等芯片设计公司的加速卡能力在算力、内存、片间互联能力等方面相较于云商和互联网公司的产品具有较大的优势，是英伟达的主要竞争对手。

同样处于图形 GPU 赛道的龙二 AMD，是目前英伟达在加速卡环节最核心的竞争对手。2023 年 12 月，AMD 正式推出 MI300X 和 MI300A 两款芯片，凭借更强的算力，更大的内存容量，以及更优惠的价格，预计获得下游客户青睐。算力方面，MI 300X 的 XCD 加速模块采用 5nm 工艺，共计拥有 1530 亿个晶体管，TF32 浮点运算性能为 653.7TFlops，FP16 和 BF16 运算性能为 1307.4TFlops，FP8 和 INT8 运算性能为 2614.9TFlops，均为英伟达 H100 的 1.3 倍。内存方面，MI 300X 的内存配置是英伟达 H100 的 2.4 倍，峰值存储带宽是其 2.4 倍，在运行 Bloom 时的推理速度是 H100 的 1.6 倍，运行 Llama2 时的推理速度是其 1.4 倍。功耗方面，MI300X 的整体功耗控制在 750W，相较英伟达 H100 更具优势。此外，在价格方面，Lisa Su 表示 MI300 系列芯片购买和运营成本将会低于英伟达。

Intel 的推理卡 Goya 在 2019 年发布，目前公司已经完善了训练卡和推理卡的布局，并开始逐步形成销售。2024 年 4 月 9 日，英特尔推出最新 AI 芯片 Gaudi3。与上一代产品相比，英特尔 Gaudi 3 将带来 4 倍的 BF16 AI 计算能力提升，以及 1.5 倍的内存带宽提升。Gaudi 3 预计可大幅缩短 70 亿和 130 亿参数

Llama2 模型，以及 1750 亿参数 GPT-3 模型的训练时间。此外，在 Llama 7B、70B 和 Falcon 180B 大语言模型（LLM）的推理吞吐量和能效方面也展现了出色性能。Gaudi 3 提供开放的、基于社区的软件和行业标准以太网网络，允许企业灵活地从单个节点扩展到拥有数千个节点的集群、超级集群和超大集群，支持大规模的推理、微调和训练。英特尔在 Vision 2024 大会上宣布，公司的 Gaudi 3 将于 2024 年第二季度面向 OEM 厂商出货。

国内方面，昇腾加速卡处于领先水平。昇腾基于华为自主研发的达芬奇架构，支持全栈全场景的 AI 计算。910B 对标英伟达 A100，蓉和咨询 CEO 吴梓豪预计 2024 年出货量将超过 40 万片。根据集成电路 IC，910C 芯片目前处于芯片级测试阶段，预计在今年第四季度推出样机，对标英伟达 H200 产品。华为也在开发者大会上发布了昇腾 AI 云服务。昇腾 AI 云服务单集群提供 20OOPFlops 算力，千卡训练 30 天长稳率达到 90%，为业界提供稳定可靠的 AI 算力。

表1：AMD 和 Intel 最新一代加速卡的性能参数对比

厂商	型号	发布时间	参数信息		峰值算力 TOPS/TFLOPS					内存信息			互联带宽 GB/s
			制程 nm	功耗 W	INT8	BF16/ FP16	TF32/ FP32	FP32	FP64	类型	容量 GB	带宽 GB/s	
AMD	MI300X	2023	5	750	2614.9	1307.4	654	163	163	HBM3	192	5300	896
Intel	Gaudi 3	2024	5	900	1835	1835	-	-	-	HBM2E	128	3700	600

资料来源：AMD，Intel，民生证券研究院

2.2.2 下游云商客户纷纷推出自研 AI 芯片

谷歌采取自研加速卡为主，同时采购部分英伟达加速卡的策略。谷歌研发 TPU 的时间始于 2013 年，相较于其他云厂商有近 10 年的时间优势。由于谷歌在加速卡领域布局早，产品完善度高，谷歌或为 2024 年北美四大云厂商中采购英伟达加速卡最少的厂商。2023 年 12 月，谷歌推出面向云端的 AI 加速卡 TPU v5p，相较于 TPU V4，TPU v5p 提供了二倍的浮点运算能力和三倍内存带宽提升；集群方面，TPU v5p pod 由 8960 颗芯片组成，使用最高带宽的芯片间连接（每芯片 4,800 Gbps）进行互连；从训练效果来看，相较于上一代产品，TPU v5p 训练大型 LLM 的速度提升了 2.8 倍。

Meta 自 2021 年以来便将企业发展的重点放在元宇宙和 AI 领域，并且修改了公司名称。2023 年，Meta 宣布自研 MTIA v1 芯片。2024 年 4 月，Meta 发布最新版本 MTIA v2 加速卡，新一代 MTIA 加速卡在算力、内存容量、内存带宽等方面更加平衡，采用台积电 5nm 工艺制造，Int 8 稀疏算力可以达到 708TOPS，HBM 内存容量达到 128GB。目前 Meta 仍主要采购英伟达等厂商的加速卡用于 Llama 等模型的训练，后续 Meta 有望采用自研加速卡对大语言模型进行训练。

微软 Azure 的企业数量已经达到 25 万家，是目前采购英伟达加速卡最为激进的云厂商，但考虑到采购英伟达加速卡的高昂成本，微软也宣布了自研加速卡的

计划。微软的 Maia 100 在 2023 年推出，专为 Azure 云服务设计。Maia 100 采用台积电 5nm 工艺，单芯片拥有 1050 亿个晶体管，FP 8 算力可以达到 1600TFLOPS，同时支持 FP 4 运算，算力达到 3200TFLOPS，是目前厂商自研加速卡中算力最强大的产品。除了微软自身以外，OpenAI 也在尝试采用微软的加速卡，强大的下游客户支持有望为微软自研加速卡的进步带来重要动力。

亚马逊同样在自研加速卡方面加大投入，并且已经完善了训练和推理的两方面布局。2023 年，亚马逊推出了用于训练的 Trainium2 加速卡，以及用于推理的 Graviton4 加速卡，补全了亚马逊在训练和推理领域加速卡的布局。亚马逊的 Trainium2 加速卡 Int 8 算力达到 861TOPS，相较上一代产品性能提升 4 倍，在云厂商自研加速卡中表现优秀。同时公司的产品可以在新一代 EC2 UltraClusters 中扩展多达 10 万颗 Trainium2 加速卡，与 Amazon EFA PB 级网络互联，提供高达 65 EFlops 算力，为云端大模型的训练和推理提供强大的算力保障。

表2：英伟达客户在加速卡领域的布局情况

厂商	大类	型号	发布时间	制程 nm	峰值算力 TOPS/TFLOPS			内存信息		互联带宽 GB/s
					INT8/FP8 Dense/Sparse	BF16/FP16 Dense/Sparse	TF32/FP32 Dense/Sparse	类型	容量 GB	
谷歌	训练	TPUv5E	2023	-	394	197	-	HBM2	16	400
		TPUv5P	2023	-	918	459	-	HBM2	95	800
Meta	推理	MTIA v2	2024	5	354/708	177/354	2.76	-	128	-
微软	训练	Maia 100	2023	5	1600	800	-	HBM3	64	1200
亚马逊	训练	Trainium2	2023	4	861	431	215	-	96	-
	推理	Graviton4	2023	-	-	-	-	GDDR5	-	-
TESLA	训练	D1	2021	7	362	362	22.6	-	32	-

资料来源：各公司官网，Semianalysis 等，民生证券研究院

注：未标注的数据为没有在公开渠道披露的信息

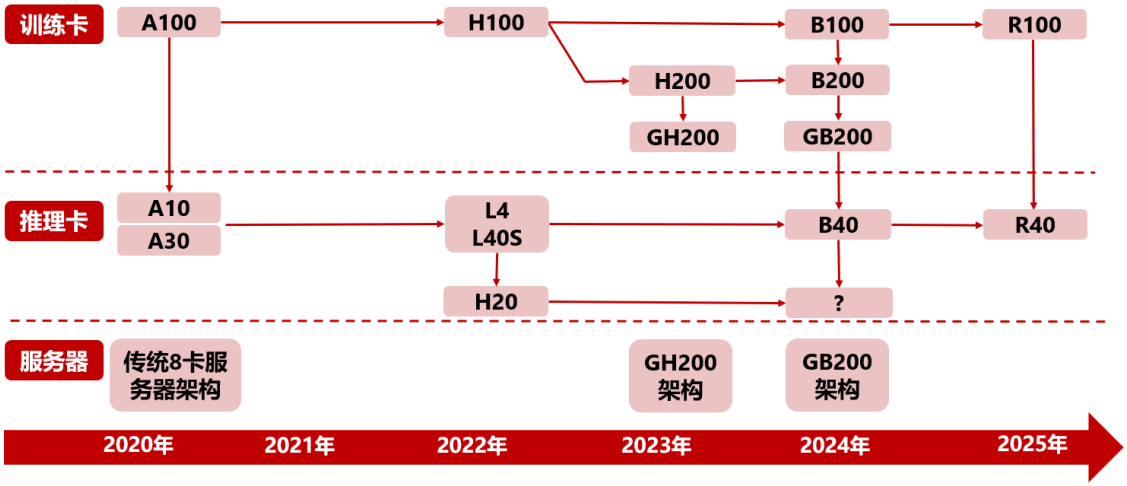
3 软硬件生态领先，加速卡龙头地位稳固

3.1 加速卡全性能领先，GB200 拉开差距

3.1.1 训练卡和推理卡布局完善，产品优势领先

英伟达具有在训练侧和推理侧完善的产品布局，并将加快芯片架构更新速度，从两年一更新加速至一年一更新。训练卡方面，自 H 系列以来，英伟达在 100 系列的基础上，会进一步增加 200 系列产品，从而给客户更多的产品选择和更好的性价比，例如 H200 加速卡在 HBM 代际和容量方面相较于 H100 均有较大提升，但价格方面提升幅度相对较少。**推理卡方面，**2024 年英伟达的产品出货以 L40 和 L40S 为主，并且推出了 L20、L2 等产品供客户选择。针对中国市场，英伟达还推出了 H20 系列加速卡，用于中国市场的推广，H20 在算力方面相对较弱，但是其拥有强大的 HBM 内存，互联带宽和软件生态，并且价格相较于 H100 有明显优惠，在下游客户中仍然具有一定的接受程度。

图12：英伟达历代训练卡和推理卡一览



资料来源：芯语，民生证券研究院整理

3.1.2 H200 和 Blackwell GPU 与竞争对手拉开差距

AMD 等厂商推出的 MI300 等强有力的产品，对英伟达的市场份额产生了一定的争夺，但英伟达很快发布了 H200、Blackwell GPU 等系列的产品，与其竞争对手又拉开了一个身位。

以 AMD 的 MI300X 为例，这颗加速卡在 Int8、FP16、FP32 算力方面均为 H100 的 1.3 倍，互联带宽方面达到了接近于 NV Link4.0 的 896GB/s 双向互联，FP64 算力和 HBM 容量更是达到了 H100 的 2 倍以上，同时 MI300X 在功耗方面相较于 H100 也更具优势，价格相较英伟达的 H100 也有较大的让步。**一系列的**

堆料和价格优惠使得下游云厂商考虑转用一部分 AMD 的产品，微软更是在公开场合表达了对 AMD 加速卡的支持。英伟达推出的 H200 很大程度上是对竞争对手的回应，相较于 H100，H200 将此前的 HBM3 提升为 HBM3E，同时将 HBM 容量从上一代的 80GB 提升至 141GB。在价格方面，H200 相较于 H100 的涨价幅度较小，体现出较高的性价比，该款加速卡现已向全球系统制造商和云服务提供商供货，将推动 AI 技术的发展。

英伟达的下一代 Blackwell GPU 系列产品，在算力、内存和互联带宽的 AI 三要素领域与竞争对手的差距进一步拉开，巩固了英伟达的领先地位。

1）从算力来看，B 系列加速卡使得英伟达与其竞争对手重新拉开了差距。仍然以 AMD 为例，AMD 的 MI300X 共计拥有 1530 亿个晶体管，FP8 和 INT8 运算性能为 2614.9TFlops。而此次 Blackwell GPU 推出了全新的 FP4 算力精度并达到了惊人的 10PFlops，INT8 算力相较 H100 提升了一倍以上。

2）从存储容量来看，Blackwell GPU 单卡内存容量高达 192GB，和 MI300X 的内存带宽一致，并且升级到了 HBM3E，在性价比上全面提升。

3）从互联带宽来看，NVLink 速率升级使得英伟达在片间互联速率相较于其他厂商的竞争优势明显提升。900GB/s 的 NVLink 互联带宽是此前市场上的最高水平，但差距并未明显拉开，AMD 的 Infinity Fabric 带宽也可以达到 896GB/s。而此次 NVLink 升级后，双向互联带宽达到了 1.8TB/s，相较市场上其他的厂商拉开了一倍以上的差距。

表3：B100、B200、H100、H200、MI300X 参数对比

	B100	B200	H100	H200	MI300X
FP4 算力（TFlops）	7000	9000	-	-	-
FP8/INT8 算力（TFlops）	3500	4500	3958	3958	2615
FP16/BF16 算力（TFlops）	1800	2250	1979	1979	1307
TF32 算力（TFlops）	900	1120	989	989	654
FP64 算力（TFlops）	30	40	67	67	163
显存	192GB HBM3E	192GB HBM3E	80GB HBM3	141GB HBM3E	192GB HBM3
互联带宽（GB/s）	1800	1800	900	900	896

资料来源：英伟达，AMD，民生证券研究院

注：表格中均为未经过结构化稀疏的算力

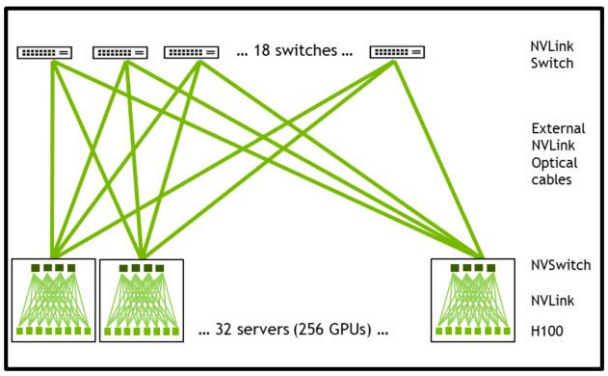
3.1.3 GB200 DGX NVL 72 显著提升集群性价比和片间互联能力

GB200 DGX NVL 72 给英伟达算力性能带来的提升主要来自性价比和片间互联能力方面。从性价比来看，NVL 72 的集群规模增大，一方面节省了除算力芯片以外的系统成本，另一方通过提升产品性能，间接提升了下游客户购买的算力性价比。

从片间互联能力来看，市场上主流的 AI 服务器仍然是传统的 8 卡服务器架构，而伴随 Blackwell 平台推出的最多可以支持 576 卡互联，片间互联数量和带宽的提升极大改善了英伟达平台的推理和训练性能。

H100 支持 8 卡 900GB/s 的 NVLink 带宽无损互联，而 GB200 支持 72 卡 1.8TB/s 的 NVLink 带宽无损互联。在传统 H100 的 8 卡服务器架构中，单台服务器内的八张加速卡的最大互联带宽为 900GB/s，如果一层交换机采用 NVLink 交换机，则最多可以支持 256 卡互联，但由于服务器连接到一层交换机的光模块带宽限制，256 张加速卡之间的互联带宽实际会下降到 100GB/s（按照一台服务器插 8 张 400G 光模块计算），而在 DGX NVL 72 架构中，由于 72 张加速卡之间采用高速铜缆互联，不受光模块带宽限制，可以支持 72 卡的 1.8TB/s 的 NVLink 带宽无损互联，并且未来最多可以支持 8 个 Rack 之间的 576 卡互联。

图13：256 卡的 H100 数据中心



资料来源：sysgen，民生证券研究院

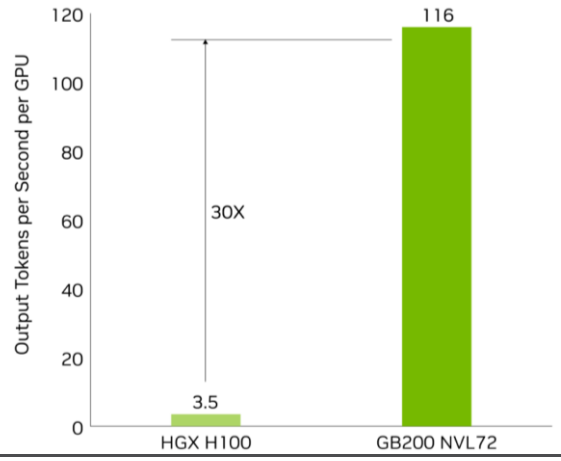
图14：英伟达 DGX NVL 72 参数

组成	36 Grace CPU : 72 Blackwell GPUs
FP4 Tensor 核心²	1,440 PFLOPS
FP8/FP6 Tensor 核心²	720 PFLOPS
INT8 Tensor 核心²	720 POPS
FP16/BF16 Tensor 核心²	360 PFLOPS
TF32 Tensor 核心	180 PFLOPS
FP32	6,480 TFLOPS
FP64	3,240 TFLOPS
FP64 Tensor 核心	3,240 TFLOPS
GPU 内存带宽	最高 13.5 TB HBM3e | 576 TB/s
NVLink 内存带宽	130TB/s
CPU 核心数	2592 个 Arm® Neoverse V2 核心
CPU 内存带宽	最高 17 TB LPDDR5X | Up to 18.4 TB/s

资料来源：英伟达，民生证券研究院

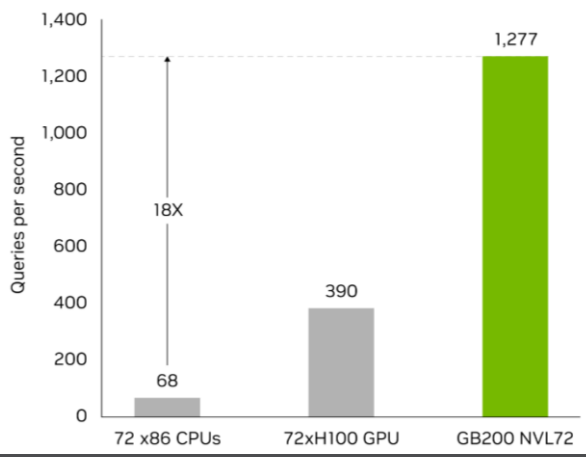
凭借架构优势，GB200 NVL72 的推理性能达到 H100 的 30 倍，相较其他竞争对手领先优势则更为显著。按照每 10 亿个参数的大模型大约需要占用 4GB 的显存容量计算，当前 GPT4 已经达到万亿参数量级，运行时至少需要几十张 GPU 同时进行推理工作，GB200 NVL72 在单个 Rack 内部可以实现万亿参数量级的大模型推理工作，大大降低了大模型在通信环节的算力资源占用，使得 GB200 NVL72 的推理性能达到 H100 的 30 倍。

图15：GB200 的推理实时吞吐量达到 H100 的 30 倍



资料来源：英伟达，民生证券研究院

图16：GB200 NVL72、72*H100、72*x86 CPU 之间的吞吐量对比

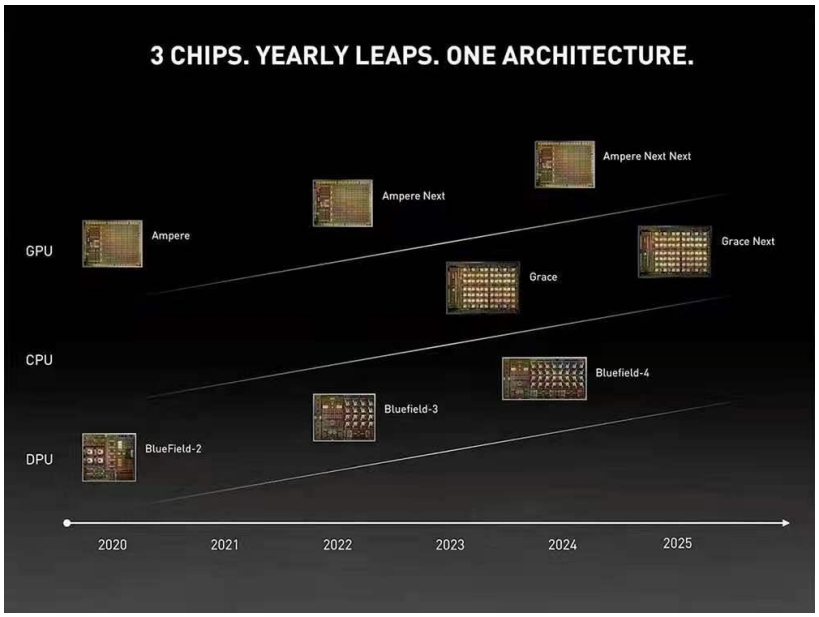


资料来源：英伟达，民生证券研究院

3.2 GPU+CPU+DPU 全布局，从加速卡向系统巨头升级

英伟达在数据中心市场已完成 CPU+GPU+DPU 三大类芯片的硬件布局，并计划加快原计划的每种芯片两年左右的迭代周期，GPU 加速到以一年为周期推出。

图17：英伟达 GPU+DPU+CPU 产品路线图



资料来源：超能网，民生证券研究院整理

3.2.1 推出数据中心专属 CPU，加强互联连接

为搭配高性能 GPU 和满足现代数据中心海量的计算需求，英伟达推出 Grace CPU。**Grace CPU 是首款面向 AI 基础设施和高性能计算的数据中心专属 CPU，由两个 CPU 芯片通过 NVLink-C2C 技术互联组成。**单个 socket 内含 144 个核心并提供 1TB/s 的内存带宽，性能是当今领先服务器芯片内存带宽和能效的两倍。Grace CPU 在数据中心的应用主要有 AI、数据分析、超大规模云应用以及高性能计算（HPC）。

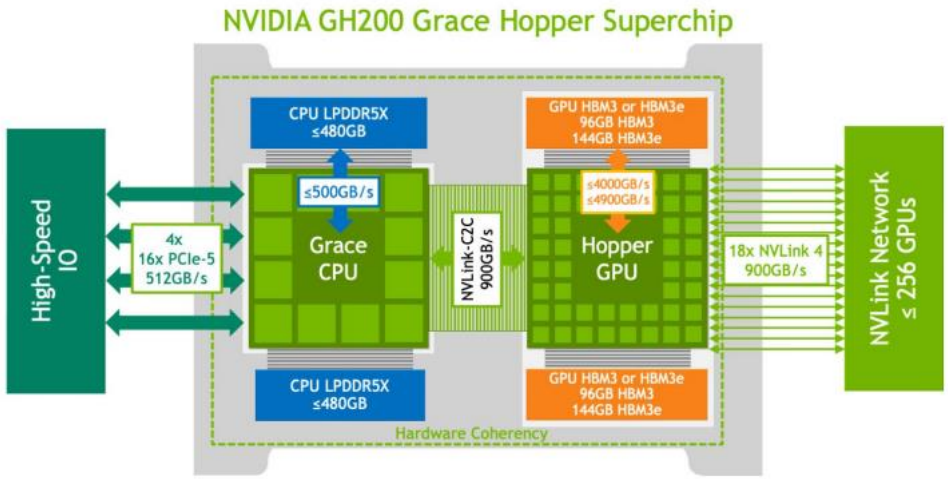
表4：Grace CPU 和 Hopper GPU 参数对比

	Grace CPU	Hopper GPU
GPU HBM 内存 (GB)		96GB HBM3
		144GB HBM3e
Grace CPU cores (数量)		Up to 72 cores
GPU HBM 带宽		4TB/s HBM3 4.9TB/s HBM3e
浮点算力	FP64：peak 7.1 TFLOPS	FP64: 34 TFLOPS FP32: 67 TFLOPS
显存带宽	Up to 1TB/s	Up to 512GB/s(grace) 900 GB/s (hopper)
互连技术	NVLink-C2C bandwidth：900GB/s	NVLink-C2C bandwidth：900 GB/s bidirectional
	PCIe links：Up to 8x PCIe Gen5 xl6 option to bifurcate	PCIe links：Up to 4x PCIe x16 (Gen5)
核心控制 Core count	144 Arm Neoverse V2 Cores with 4x128b SVE2	72 Arm Neoverse V2 cores
低功耗内存 LPDDR5X size	240GB, 480GB and 960GB	Up to 480GB

资料来源：英伟达官网，民生证券研究院

在 GH200 Grace Hopper 芯片中，GPU 与 CPU 通过 NVLink-C2C 互联连接，以优秀的内存和带宽大幅提升加速计算能力。其中 NVLink-C2C 是专门用于超级芯片的高带宽、低延迟互连，通过每秒 900GB/s 或更高的互连带宽提供高达 25 倍的能效，比 PCIe Gen 5 的面积效率高 90 倍。NVLink-C2C 的搭配使外部的应用程序能超额订阅 GH200 的 GPU 内存，并在该带宽下直接使用 Grace CPU 内存。因此 GH200 可以更快部署在标准服务器中，并在加速计算和生成式 AI 方面展现技术优势。

图18：Grace Hopper 超级芯片结构示意图



资料来源：英伟达官网，民生证券研究院整理

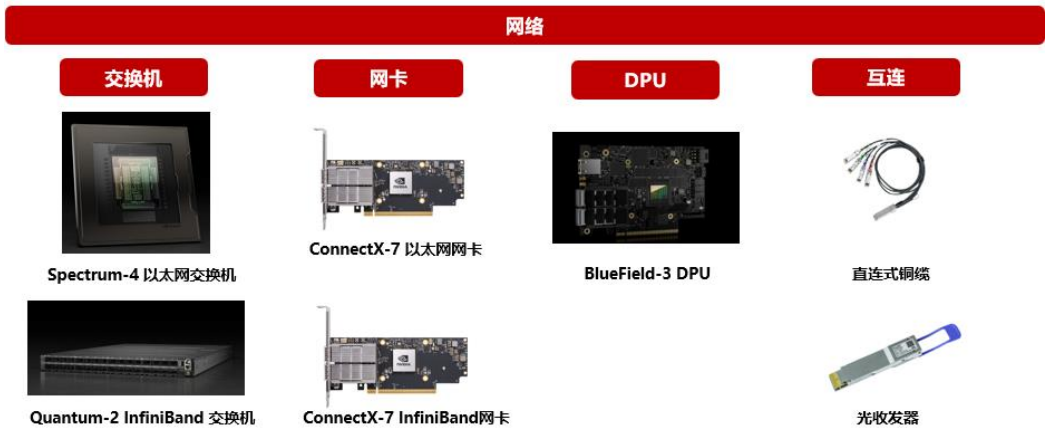
3.2.2 网络产品全覆盖，提升加速计算性能

高性能计算和生成式 AI 时代的到来对服务商提出更高计算能力、效率和可拓展性的需求。因此，**英伟达通过自身研发及收购 Mellanox 等公司实现交换机、网卡、DPU、互连等网络产品的全覆盖，为客户提供基于 InfiniBand 和以太网网络技术的端到端解决方案。**

2020 年 4 月 47 日，英伟达宣布完成对 Mellanox 的收购。Mellanox 是一家以色列面向服务器、存储和超聚合基础设施的端到端以太网和 InfiniBand 智能互联解决方案与服务的领先供应商。作为高性能互联技术的早期创新者，Mellanox 率先推出了 InfiniBand 互联技术，该技术与其高速以太网产品均应用于全球过半的高速超级计算机和许多领先的超大规模数据中心。**NVIDIA 和 Mellanox 强强联合，将能够优化整体计算、网络和存储堆栈的数据中心级工作负载，从而助力客户实现更高的性能和利用率，并降低运营成本。**

目前英伟达在网络的产品矩阵有交换机、网卡、DPU 和互连等。交换机包括 InfiniBand 和以太网两大类交换机，可无缝连接主机上的 GPU 设备，为企业级数据中心和高性能计算提供卓越的网络能力。Connect-X 系列智能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。BlueField-3 DPU 为企业在加速计算和 AI 应用提供安全的硬件加速设施。直连式铜缆和光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。

图19：英伟达网络产品线



资料来源：英伟达官网，民生证券研究院

3.3 软件生态至关重要，CUDA 优势长期存在

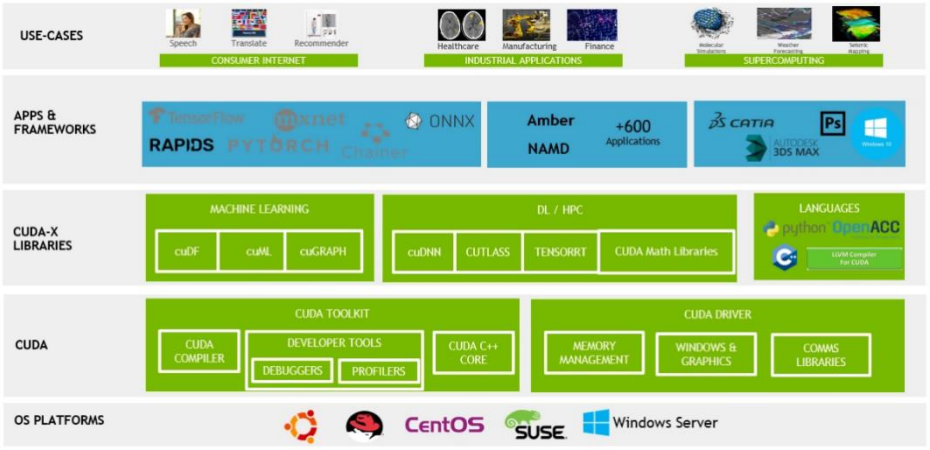
GPU 的软件生态主要包括一些开源或专有的平台和工具，它们允许开发者利用 GPU 进行高效的并行计算。在 GPU 市场中，成熟的软件生态已成为各厂商之间难以逾越的护城河。相较于不断迭代的微架构技术，生态所带来的用户粘性在长

期竞争中显得更为关键。

目前英伟达的 CUDA、AMD 的 ROCm、微软的 DirectX 以及 OpenCL、OpenGL、Vulkan 等已成为主流的开发平台，为开发者提供高效、便捷的 GPU 计算解决方案，随着技术的不断发展和应用场景的不断拓宽，这些平台将继续发挥重要作用。

CUDA 是英伟达于 2006 年推出的一种异构计算平台，开发人员能够通过 CUDA 对 GPU 进行通用计算（GPGPU）的部署。在 CUDA 编程模型中，Host 代表主机部分，主要由 CPU 和主机内存组成；而 Device 代表设备部分，主要由 GPU 和显存构成。Host 与 Device 之间通过 PCI Express 总线进行高效的数据传输和通信。这种架构使得 CUDA 能够充分利用 GPU 的并行计算能力，实现高性能的并行计算任务。

图20：CUDA 生态示意图

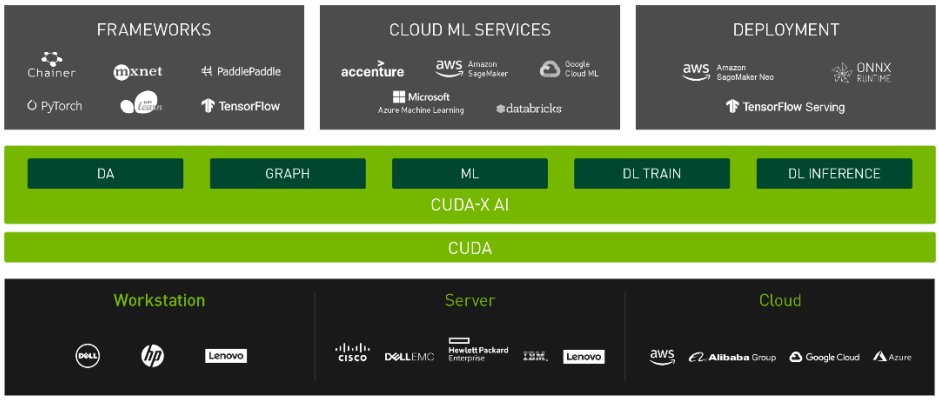


资料来源：搜狐，民生证券研究院

CUDA 在 Host 运行的函数库包括了 Libraries、Runtime 和 Driver 三大部分。其中，Libraries 通常是一些成熟的高效函数库，Runtime API 则简化了应用开发过程，提供了便捷的接口和组件，让开发者能够轻松地调用并自动管理 GPU 资源。应用程序可以通过调用 CUDA Libraries 或者 CUDA Runtime API 来实现所需功能的控制 GPU 资源的能力。当涉及到 Device 端的计算任务时，CUDA 使用内核函数进行并行计算和数据处理，从而充分发挥 GPU 的性能优势。

CUDA 生态支持多种编程语言。目前的 CUDA 12.0 支持 C、C++、Fortran、Python 等多种编程语言，还支持众多第三方工具链。此外，**英伟达在 CUDA 平台上提供了 CUDA-X，CUDA-X 是一个功能强大且灵活的软件加速库集合**，开发人员可以通过 CUDA-X 快速将这些库部署到多种设备内的 NVIDIA GPU 上，包括台式机、工作站、服务器、云计算和物联网（IoT）设备。CUDA 平台对开发者友好程度高，其提供的一系列容器部署流程简化以及集群环境扩展应用程序的工具使得 CUDA 技术能够适用于更广泛的领域。

图21：CUDA-X 平台



资料来源：NVIDIA，民生证券研究院

凭借先发优势和长期技术积累，**CUDA 生态圈已经具有更高的成熟度和稳定性。**这使得开发者能够借助已有的资源和文档进行开发和部署，减少学习曲线和风险，并为英伟达 GPU 的开发、优化和部署多种行业应用提供了独特的先发竞争优势。全球范围内，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。较高的需求粘性也使得 CUDA 的使用者更倾向于使用熟悉的、更兼容的软件，因此更多开发者选择或持续使用 CUDA。

4 其他业务板块稳健增长

4.1 游戏业务：收入保持稳定，技术业内领先

英伟达的游戏显卡产品主要有 GeForce RTX 30 系列和 GeForce RTX 40 系列等。

GEFORCE RTX 40 系列 GPU 由更高效的 NVIDIA Ada Lovelace 架构提供动力支持。公司主推的 GeForce RTX 4090 D 拥有 24 GB 的 G6X 显存，搭配的新型 SM 多单元流处理器可将性能功耗比最高提升至 2 倍。第四代 Tensor Core 与 DLSS 3 结合将图像渲染性能提升至 4 倍，第三代 RT Core 将光线追踪性能提升至 2 倍。综上，该显卡在性能、效率和 AI 驱动的图形效果方面均比上代实现大幅增长。

GEFORCE RTX 30 系列采用第 2 代 NVIDIA Ampere 架构，配置有第 2 代 RT Core、第 3 代 Tensor Core 和 SM 多单元流处理器，为游戏用户带来逼真的光线追踪效果和先进的 AI 性能，NVIDIA Reflex 的搭配降低系统延迟，将带来超凡的游戏体验。

图22：GEFORCE RTX 40 系列的规格

	GeForce RTX 4090 D	GeForce RTX 4080 SUPER	GeForce RTX 4080	GeForce RTX 4070 TI SUPER	GeForce RTX 4070 TI	GeForce RTX 4070 SUPER	GeForce RTX 4070	GeForce RTX 4060 TI	GeForce RTX 4060
GPU 引擎规格：									
NVIDIA CUDA® 核心数量	14592	10240	9728	8448	7680	7168	5888	4352	3072
Shader Core	Ada Lovelace 74 TFLOPS	Ada Lovelace 52 TFLOPS	Ada Lovelace 49 TFLOPS	Ada Lovelace 44 TFLOPS	Ada Lovelace 40 TFLOPS	Ada Lovelace 36 TFLOPS	Ada Lovelace 29 TFLOPS	Ada Lovelace 22 TFLOPS	Ada Lovelace 15 TFLOPS
RT Core	第 3 代 170 TFLOPS	第 3 代 121 TFLOPS	第 3 代 113 TFLOPS	第 3 代 102 TFLOPS	第 3 代 93 TFLOPS	第 3 代 82 TFLOPS	第 3 代 67 TFLOPS	第 3 代 51 TFLOPS	第 3 代 35 TFLOPS
Tensor Core (AI)	第 4 代 1177 AI TOPS	第 4 代 836 AI TOPS	第 4 代 780 AI TOPS	第 4 代 706 AI TOPS	第 4 代 641 AI TOPS	第 4 代 568 AI TOPS	第 4 代 466 AI TOPS	第 4 代 353 AI TOPS	第 4 代 242 AI TOPS
加速频率 (GHz)	2.52	2.55	2.51	2.61	2.61	2.48	2.48	2.54	2.46
基础频率 (GHz)	2.28	2.29	2.21	2.34	2.31	1.98	1.92	2.31	1.83

资料来源：英伟达官网，民生证券研究院

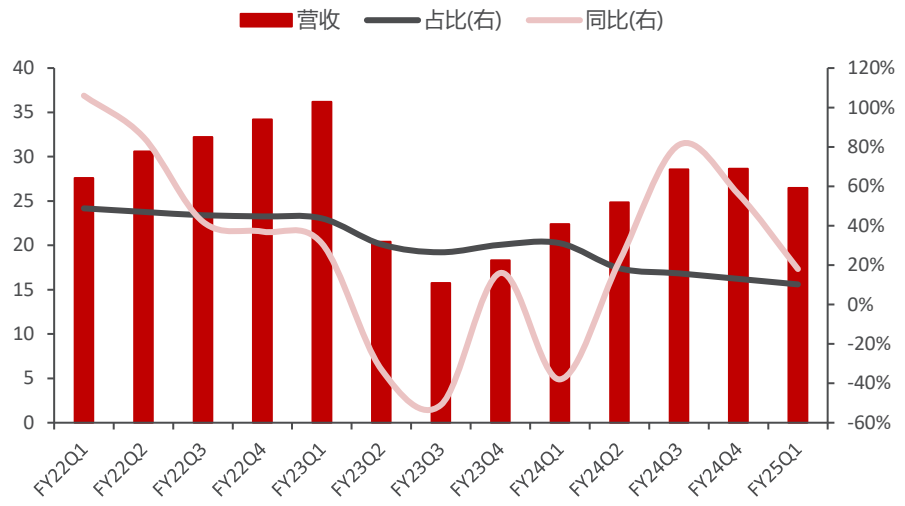
图23：GEFORCE RTX 30 系列的规格

	GeForce RTX 3090 Ti	GeForce RTX 3090	GeForce RTX 3080 Ti	GeForce RTX 3080	GeForce RTX 3070 Ti	GeForce RTX 3070	GeForce RTX 3060 Ti	GeForce RTX 3060	GeForce RTX 3050 (8 GB)	GeForce RTX 3050 (6 GB)
GPU 引擎规格：										
NVIDIA CUDA® 核心数量	10752	10496	10240	8960 / 8704	6144	5888	4864	3584	2560 *	2304
加速频率 (GHz)	1.86	1.70	1.67	1.71	1.77	1.73	1.67	1.78	1.78 *	1.47
基础频率 (GHz)	1.56	1.40	1.37	1.26 / 1.44	1.58	1.50	1.41	1.32	1.55 *	1.04
显存规格：										
标准显存配置	24 GB GDDR6X	24 GB GDDR6X	12 GB GDDR6X	12 GB GDDR6X / 10 GB GDDR6X	8 GB GDDR6X	8 GB GDDR6	8 GB GDDR6 X / 8 GB GDDR6X	12 GB GDDR6 / 8 GB GDDR6	8 GB GDDR6	6 GB GDDR6
显存位宽	384 位	384 位	384 位	384 位 / 320 位	256 位	256 位	256 位	192 位 / 128 位	128 位	96 位

资料来源：英伟达官网，民生证券研究院

FY1Q25 英伟达游戏业务实现营收 26.47 亿美元，同比增长 18.17%，环比下降 7.61%。在 2024 年第一季度，英伟达推出适用于 Windows 的全新 AI 性能优化和集成，可在 AI PC 和工作上提供最佳性能。公司还宣布“星球大战：亡命之徒 （Star Wars Outlaws）”和“黑神话：悟空 （Black Myth Wukong）”等采用 RTX 技术的重磅游戏。

图24：英伟达游戏业务收入、同比增速及占比（亿美元，%）

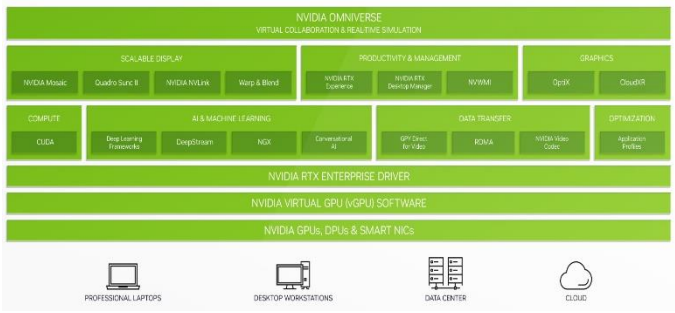


资料来源：Bloomberg，民生证券研究院

4.2 专业可视化业务：加强 AI 导入，应用场景广阔

英伟达通过 RTX 服务器、数据中心工作站和虚拟 GPU 等系列产品将 AI 与专业可视化技术结合，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等专业视觉解决方案，并可应用在建筑、医疗、影视媒体等多种应用场景。公司新推出的 RTX 2000 Ada GPU 搭配第三代 RT Core 和第四代 Tensor Core，在提升光线追踪性能的同时为 AI 加速工具的导入提供接口。16GB 现存可为关键任务提供更高的计算精度与可靠性。该芯片可助力客户在建筑设计、工业设计、内容设计、医疗数据处理等场景提高效率，提升用户体验。

图25：英伟达加速视觉计算平台



资料来源：英伟达官网，民生证券研究院

图26：RTX 2000 Ada GPU 产品的规格

NVIDIA RTX 2000 Ada Generation

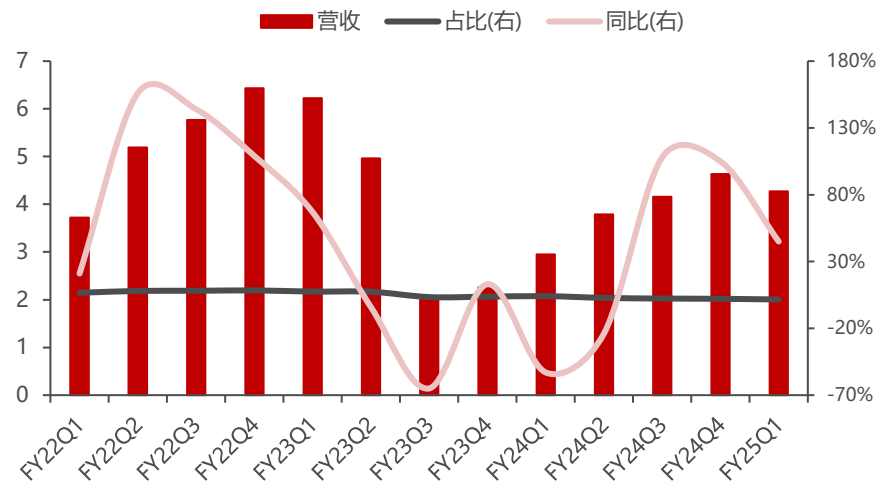
GPU 内存	16GB GDDR6，支持错误修正码（ECC）
显示器端口	4 个 mini DisplayPort 1.4a
最大功耗	70 W
图形总线	PCIe Gen 4 x 8
尺寸规格	2.7 吋（高） x 6.6 吋（长），双插槽
散热	主动式
VR Ready	有

资料来源：英伟达官网，民生证券研究院

FY1Q25 英伟达专业可视化业务实现营收 4.27 亿美元，同比增长 44.75%，环比下降 7.61%，同比回归正增长态势。在 2024 年第一季度，公司推出 NVIDIA RTX 500 和 1000 GPU 加强 AI 应用，并通过 NVIDIA Omniverse Cloud API 来支持工业数字孪生软件工具，如扩大与西门子的合作伙伴关系，以及 Apple

Vision Pro 的新框架。

图27：英伟达专业可视化业务收入、同比增速及占比（亿美元，%）

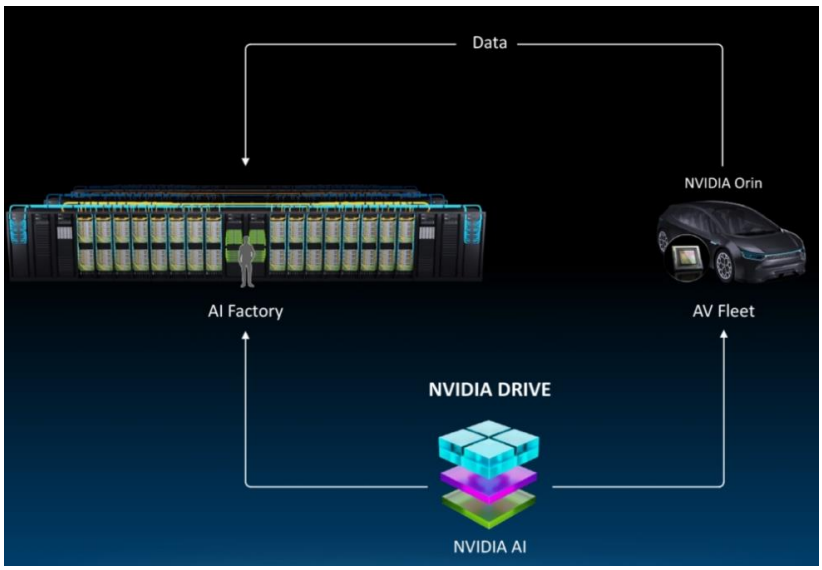


资料来源：Bloomberg，民生证券研究院

4.3 汽车：拥有完整解决方案，与客户深度合作

英伟达可提供完整的自动驾驶汽车硬件，软件和基础设施。其中自动驾驶 DRIVE 平台提供从汽车系统到数据中心的端到端解决方案，其架构包含传感器、DRIVE AGX 计算平台、以及实现强大的自动驾驶和智能驾舱功能所需的软件工具。在数据中心层面，英伟达提供必要的硬件和软件支持包括用于感知的深度神经网络的 NVIDIA DGX，以及用于生成数据集和验证自动驾驶堆栈的 DRIVE Sim。

图28：英伟达端到端自动驾驶汽车开发平台



资料来源：英伟达官网，民生证券研究院

在 DRIVE 平台中，**NVIDIA DRIVE Orin 系统级芯片（SoC）起到核心计算**

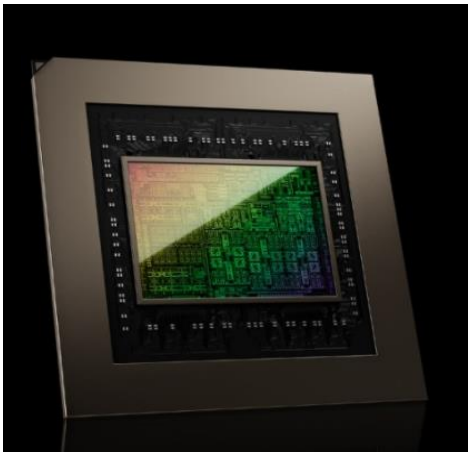
模拟作用。该芯片可提供每秒 254 万亿次的运算性能，并能够为自动驾驶功能、置信度视图、数字仪表盘以及 AI 座舱提供强力支持。借助 Orin 芯片的构建和扩展功能，开发者利用一次投资便可从 L2+级系统顺利升级至 L5 级全自动驾驶汽车系统。

DRIVE Thor 作为新一代车载计算平台，是公司在智能驾驶领域的主推产品。该平台在单个安全可靠的系统中整合高级驾驶员辅助功能和车载信息娱乐功能。借助全新的 CPU 和 GPU 技术，计算系统可在降低系统成本的同时提供高达 2000 万亿次的浮点运算性能。Thor 平台计划于 2025 年量产，会为公司带来汽车业务新的业绩增长点。

图29：NVIDIA DRIVE Orin 系统级芯片

资料来源：英伟达官网，民生证券研究院

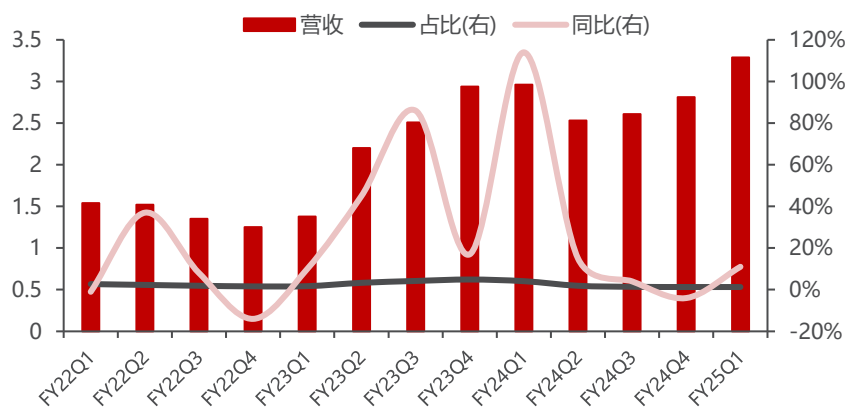
图30：NVIDIA DRIVE Thor 平台



资料来源：英伟达官网，民生证券研究院

FY1Q25 英伟达汽车业务实现营收 3.29 亿美元，环比增长 17.08%，同比增长 11.15%，增长势头明显。在 2024 年第一季度，公司与比亚迪、小鹏汽车、广汽等车企合作，为全新的移动出行提供最佳解决方案。美国和中国的电动汽车制造商也在研究将 Orin 芯片应用于欧洲市场的车型。

图31：英伟达汽车业务收入、同比增速及占比（亿美元，%）



资料来源：Bloomberg，民生证券研究院

5 投资建议

AI 浪潮不仅没有放缓，反而在持续加速。生成式 AI 出现后，人工智能获得快速发展，以 ChatGPT 为分水岭，此后 AI 大模型的成长节奏显著提速。2023 年的 GPT 4、LLaMA、Gemini 等重要模型出现，大模型的能力已经有了质变，此前市场所担忧的是 2024 年以后，算力的需求是否会伴随着各家厂商大模型的逐步成型而放缓，然而我们从云厂商的资本开支来看，下游客户对算力的需求不仅没有放缓，反而在持续加速。我们认为这背后有两大原因：1）首先大模型产业倾向于成为“赢者通吃”的行业，持续扩大的模型参数和训练样本量使得云和大模型厂商不得不持续加大算力投入；2）其次展望未来，目前算力需求的核心下游仍然是训练，伴随着后续 AI 应用的落地，推理侧的算力需求有望带来加速卡更大的市场空间。我们认为，算力的强劲需求或许并不会在大模型训练告一段落以后就开始放缓，反而可能伴随大模型发展加速，以及应用侧的落地而进一步提速。

需要重视英伟达在 AI 算力领域的话语权提升。根据半导体研究机构 TechInsights 数据，2023 年全球数据中心 GPU 总出货量达到了 385 万颗，英伟达的市场份额为 98%，并通过 CUDA 生态构建了强大的软件生态壁垒。尽管在 2023 年以来，AMD、Intel 等芯片设计公司，以及谷歌、微软、Meta、亚马逊等英伟达的下游客户也在增强 AI 加速卡领域的布局，但我们看到的现象是，英伟达的产品节奏迭代速度在加快，一方面通过 H200、B200 等系列的加速卡给客户提供了更多更高性价比的选择，拉开了和竞争对手的性能差距，另一方面通过 GB200 NVL 36/72 的 Rack 架构，从两个维度提升了英伟达的竞争力：1）GB200 NVL72 的推理性能达到 H100 的 30 倍，大幅提升了英伟达的产品在推理领域的竞争优势，有望使得英伟达在推理领域的份额持续提升；2）GB200 Rack 形态的服务器增强了英伟达在服务器产业链的设计话语权，在硬件领域进一步提升了公司的壁垒。我们看到，与上一代的 GH200 不同，GB200 从推出以后就受到了下游客户的推崇，各大云厂商均下单了大量服务器，反映出英伟达在该产品定义中的前瞻性。

伴随着 AI 下游需求的持续增长，以及 PC 等市场逐步复苏，公司业绩有望持续增长。数据中心业务来看，FY1Q24 英伟达实现营收 225.63 亿元，同比增长 426.68%，从当前下游客户的需求来看，后续该板块业务有望维持高速增长。其他业务来看，游戏业务 FY1Q24 营收 26.47 亿美元，同比增长 18.17%；专业可视化业务 FY1Q24 营收 4.27 亿美元，同比增长 44.75%；汽车和机器人业务 FY1Q24 营收 3.29 亿美元，同比增长 11.15%，均有不同程度的同比增长。公司下游 PC 等市场有望持续复苏，据 IDC 预计，2024 年中国 PC 出货量同比增长 3.8%，重回增长节奏。

英伟达在数据中心领域竞争优势明显，充分受益于 AI 浪潮，其他业务板块业绩稳步增长。我们看好英伟达在 AI 领域的前瞻布局和市场地位，建议积极关注。

表5：行业重点关注个股

股票代码	公司简称	收盘价（美元）	EPS（美元）			PE（倍）		
			FY2023A	FY2024E	FY2025E	FY2023A	FY2024E	FY2025E
NVDA.O	英伟达	125.83	1.20	2.72	3.67	105	46	34

资料来源：Bloomberg，民生证券研究院预测；（注：股价为 2024 年 7 月 5 日收盘价；公司数据采用 Bloomberg 一致预期）

6 风险提示

1）AI 行业需求波动的风险。目前英伟达主要的收入来源来自 AI 带来的数据中心业务收入，当前云厂商仍处于 AI 快速投入的阶段，下游需求十分旺盛。如果生成式人工智能大模型的参数量增速开始放缓，或云厂商大模型的训练告一段落，则可能放缓加速卡的下游需求。目前来看，AI 应用仍处于早期阶段，推理侧需求量较难判断，如果 AI 训练对算力的需求放缓，而推理侧需求没有及时起量，则加速卡下游需求可能出现一段时间的空窗期，从而对英伟达业绩产生不利影响。

2）行业竞争格局变化的风险。当前英伟达在 AI 加速卡领域处于显著领先的地位，拥有绝大多数的市占率，而目前来看，芯片设计公司如 AMD、Intel 等纷纷加强在加速卡领域的投入，英伟达的下游云厂商客户也在持续研发 AI 芯片，英伟达在加速卡领域的竞争格局从此前的一家独大逐步转变为多强竞争，可能对公司的市占率产生一定的不利影响。另一方面，英伟达在 AI 训练领域的话语权强大，而进入到 AI 推理领域，公司的市占率相较于 AI 训练领域略低，如果后续 AI 训练对算力的需求增速放缓，而 AI 推理伴随着 AI 应用落地需求快速增长，则可能对公司的市占率产生一定的不利影响。

3）产品研发进度不及预期的风险。当前，生成式 AI 技术正处于行业发展的初级阶段，单一及多模态大模型的持续涌现，使得 AI 模型在智能化、高效化方面取得了显著进步，能够胜任更加复杂且多样化的任务。而当前行业需求快速增长的同事，竞争也尤为激烈，这对整个 AI 加速卡的厂商提出了更高的要求。目前来看，AMD 等厂商的产品也在快速追赶英伟达的产品性能和迭代节奏，如果公司的硬件产品升级节奏放缓，或软件生态建设节奏不及预期，则可能对公司的业绩产生不利影响。

4）宏观经济及下游需求恢复不及预期的风险。生成式人工智能的核心下游仍然需要落地到具体的应用场景，如果宏观经济不及预期，或下游宏观需求恢复不佳，则可能直接影响企业、个人用户在新领域、新技术方面的投资热情，从未对整体 AI 下游应用的需求产生不利影响。宏观经济的变动除了可能影响英伟达在数据中心业务的需求以外，还可能会影响公司其他三大板块的业务需求。

插图目录

表格目录

分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

民生证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，客户应当充分考虑自身特定状况，不应单纯依靠本报告所载的内容而取代个人的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

民生证券研究院：

上海：上海市浦东新区浦明路 8 号财富金融广场 1 幢 5F； 200120

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层； 100005

深圳：广东省深圳市福田区益田路 6001 号太平金融大厦 32 层 05 单元； 518026