英伟达（NVDA）

英伟达：重塑计算，世界 AI 的引擎

——英伟达首次覆盖报告

	李奇(分析师)	梁昭晋(分析师)
	0755-23976888	0755-23976666
	liqi028295@gtjas.com	liangzhaojin027677@gtjas.com
证书编号	S0880523060001	S0880523010002

本报告导读：

以超异构创新重塑大规模 AI 计算，占 GPU 市场近 80%份额，数据中心业务高速增长，成为世界 AI 的增长引擎。

摘要：

- **首次覆盖，给予“增持”评级。**英伟达作为行业龙头当仁不让，考虑到其 1QFY2024 营收的出色表现，包括数据中心收入创下 42.8 亿美元的纪录，以及英伟达自身对于 2QFY2024 的收入展望达 110.0 亿美元的乐观预期，我们预计公司 FY2024E/FY2025E/FY2026E 营业收入分别为 400.0/516.26/620.0 亿美元，同增 48.29%/29.07%/20.09%，FY2024E/FY2025E/FY2026E 经调整净利润分别为 151.96/ 223.07/ 285.79 亿美元，同增 247.89%/46.80%/28.12%。
- **英伟达以超异构创新构建面向大规模 AI 计算的系统性竞争优势。**英伟达面向 AI 时代大规模并行计算，进行了全栈系统的优化。英伟达芯片互联通信技术 NVLink 性能快速迭代，GPU + Bluefield DPU + Grace CPU 的结合开创性地实现了芯片系统间的高速通信互联。同时 CUDA 充当通用平台，引入英伟达软件服务和全生态系统。我们认为，芯片和系统耦合的实现使得英伟达真正实现了超异构创新。
- **GH200 超级芯片是英伟达产品与技术的集大成者。**我们认为，GH200 集合了最先进的 Grace Hopper 架构，并应用第四代 Tensor Core 提升计算性能、进行模型优化，NVLink 实现了高速的传输，尤其是 NVLink 改变了传统 PCIe 复杂的传输过程，满足了在每个 GPU 之间实现无缝高速通信的需求，构建起了芯片间的高速互联系统，将进一步形成英伟达的竞争壁垒。
- **英伟达作为龙头企业将大比例享受 AI 芯片行业整体需求高增带来的红利。**IDTechEx 预测 2033 年全球 AI 芯片市场将增长至 2576 亿美元；JPR 预测 2022-2026 年全球 GPU 销量复合增速将保持在 6.3%水平。英伟达作为业内有目共睹的头部公司，产品生态具备显著的稀缺性，将在算力领域充分受益，享受市场爆发带来的客户需求高增。
- **风险提示**：AI 应用发展不及预期；公司研发进度不及预期；地缘政治冲突影响产品销售。

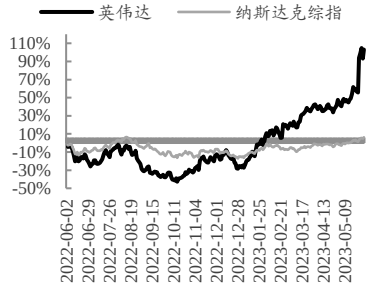
财务摘要（百万美元）	FY2020A	FY2021A	FY2022A	FY2023A	FY2024E	FY2025E	FY2026E
营业收入	10,918	16,675	26,914	26,974	40,000	51,626	62,000
(+/-)%	-6.81%	52.73%	61.40%	0.22%	48.29%	29.07%	20.09%
毛利润	6,768	10,396	17,475	15,356	27,500	35,926	43,000
净利润	2,796	4,332	9,752	4,368	15,196	22,307	28,579
(+/-)%	-32.48%	54.94%	125.12%	-55.21%	247.89%	46.80%	28.12%
PE	37.02	74.36	59.96	114.97	61.50	41.89	32.70
PS	14.04	19.32	21.72	18.62	23.36	18.10	15.07

评级：	增持
当前价格（美元）：	392.35
	2023.06.05

交易数据

52 周内股价区间（美元）	112.27-401.11
当前股本（百万股）	2,470
当前市值（百万美元）	972,600

52周内股价走势图



感谢秦和平对本报告的贡献

相关报告

目 录

1. 一台不断学习进化的机器，三十年打造生态帝国

1.1. 图形芯片时代开端，帝国之路就此开启

英伟达成立于 1993 年，怀揣打造图形芯片时代愿景。英伟达（NVIDIA）总部位于美国加利福尼亚州圣克拉拉市，依托硅谷作为全球电子工业基地的地缘优势，1993 年，黄仁勋、克里斯（Chris A.Malachowsky）与普雷艾姆（Curtis Priem）怀着 PC 有朝一日会成为畅享游戏和多媒体的消费级设备的信念，共同创立了英伟达。

图 1 英伟达初代商标

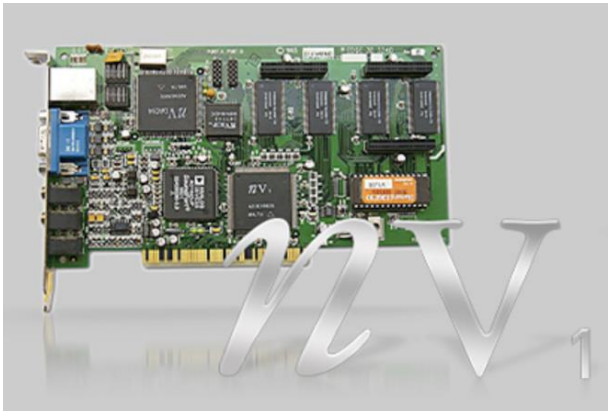
数据来源：英伟达官网

1.2. 多方求索重塑行业，重新定义现代图形

1.2.1. 1993 年-1998 年:萌芽期

图形芯片市场竞争日益激烈，英伟达多方探索寻求突破。英伟达成立之初，市场上仅有 20 余家图形芯片公司。1994 年，英伟达与 SGS-THOMPSON 首次开展战略合作；1995 年，英伟达推出其首款显卡产品 NV1，配备了基于正交纹理映射的 2D/3D 图形核心，支持 2D、3D 处理能力的同时还拥有音频处理能力;1996 年,英伟达推出首款支持 Direct3D 的 Microsoft DirectX 驱动程序；1997 年，英伟达发布全球首款 128 位 3D 处理器 RIVA 128，发布后四个月内销量超 100 万台，但此时，图形芯片这一市场的竞争者已飙升至 70 家，英伟达深陷财务泥淖，最终决定将研发和生产重心放在 2D/3D 的 PC 专用融合显卡领域；1998 年，英伟达与台积电签订多年战略合作伙伴关系，台积电开始协助制造英伟达产品。

图 2 英伟达首款显卡产品 NV1



数据来源：英伟达官网

图 3 英伟达首款驱动 MICROSOFT DIRECTX

数据来源：英伟达官网

1.2.2. 1999 年-2005 年:成长期

1999 年发明 GPU，行业重塑之路就此开启。GeForce 256 是由英伟达发布的全球首款 GPU，英伟达将 GPU 定义为“具有集成变换、照明、三角设置/裁剪和渲染引擎的单芯片处理器，每秒可处理至少 1000 万个多边形”。同年，英伟达推出适用于专业图形的 Quadro GPU，并宣布以每股 12 美元的价格首次公开募股。2000 年，显卡先驱 3dfx 因先前拒绝使用微软 Direct3D 通用 API 标准而导致其显卡通用性降低，并因其市场战略的失误，最终被英伟达低价收购；2003 年，英伟达收购无线领域图形和多媒体技术领导者 MEDIA Q，2004 年，NVIDIA SLI 问世，大大提升了单台 PC 的图形处理能力。

图 4 英伟达 GeForce 256 DDR

数据来源：techpowerup

图 5 显卡先驱 3dfx

数据来源：英伟达官网

1.2.3. 2006 年-2014 年:成熟期

CUDA 打造 GPU 计算的开发环境，硬件+软件生态帝国初现。2006 年，英伟达推出基于通用 GPU 计算的 CUDA 架构，借助 CUDA 和 GPU 的并行处理能力，英伟达收获了开发者庞大的用户群；2007 年，英伟达推出 Tesla GPU,让此前只能在超级计算机中提供的计算能力被更广泛的应用；2008 年，Tegra 移动处理器问世，其能耗约为一般的 PC 笔记本的三十分之一；2013 年，四核移动处理器 Tegra 4 发布；2014 年，英伟达推出 192 核超级芯片 Tegra K1 和平板电脑 SHIELD tablet。至此，英伟

达的几大产线均逐步成熟，应用行业逐步扩张，产品生态逐步健全。

图 6 CUDA 的生态系统

数据来源：英伟达官网

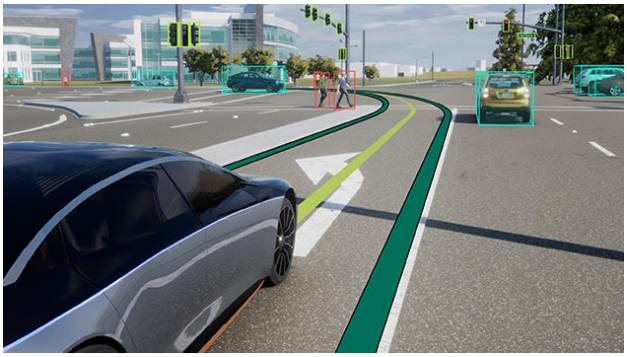
图 7 平板电脑 SHIELD tablet

数据来源：英伟达官网

1.2.4. 2015 年至今:转型期

深度学习需求催化英伟达产品转型，为 AI 革命注入强劲动力。2015 年，搭载 256 核移动超级芯片的 Tegra X1 的 NVIDIA DRIVE 问世，其可用于驾驶辅助系统，为自动驾驶汽车技术发展铺平了道路，也标志着英伟达正式投身深度学习领域；2016 年，英伟达推出第 11 代 GPU 架构 PASCAL、首款一体化深度学习超级计算机 DGX-1 和人工智能车辆计算平台 DRIVE PX 2，相较 CPU 而言，DGX-1 可将深度学习训练速度提高 96 倍；2017 年，更适合超算的 Volta 架构发布；在随后的几年里，Turing、Ampere 等架构陆续发布，持续助力 AI 革命。

图 8 DRIVE SDK 平台



数据来源：英伟达官网

图 9 DGX-1 可大幅提升深度学习训练速度

数据来源：英伟达官网

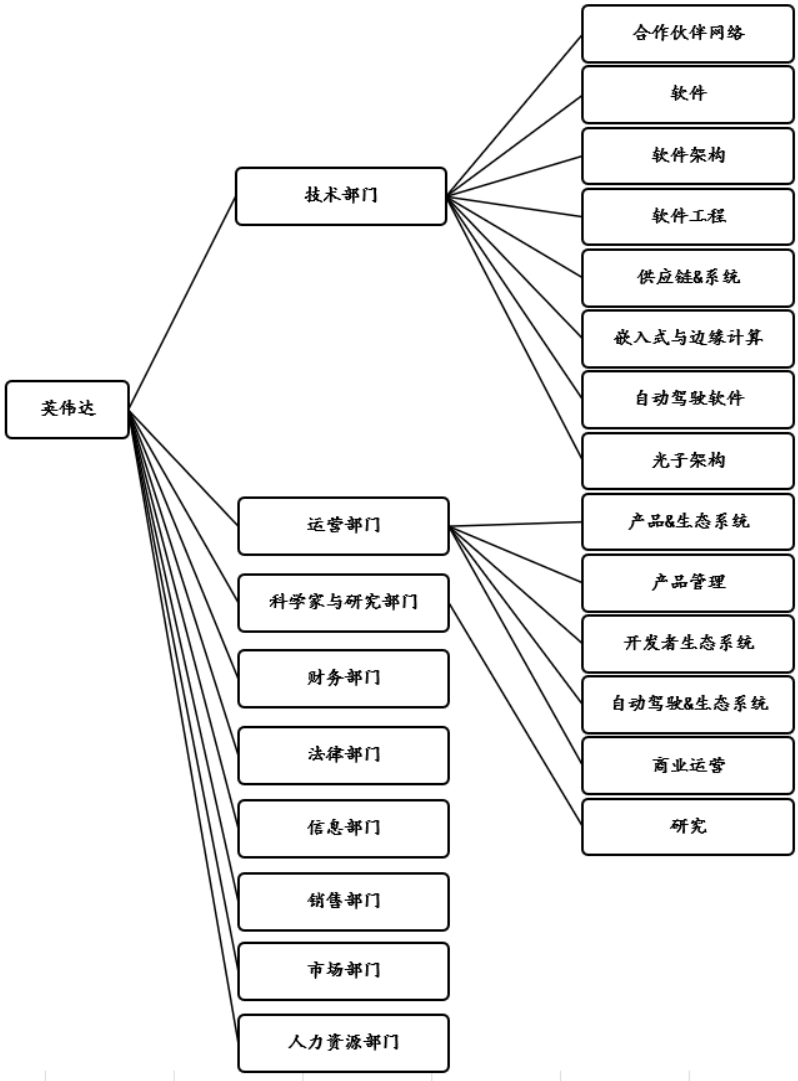
1.3. 组织架构明晰，管理团队专业

组织架构服务产品业务条线，管理团队权责清晰。据 theofficialboard，英伟达的组织架构清晰，技术和运营部门较为庞大，各大核心业务条线均有团队专门负责。英伟达官网招聘信息显示，英伟达定义的其核心业务部门包括 AI、研究和硬件三大类。我们认为，公司组织架构设置平行于产品业务，有助于发挥研究者的专项技术才能，并强调研究的前瞻性和突破性。同时，以黄仁勋为首的管理团队具有专业的业务背景与管理才能，公司管理层与董事会均由经验丰富的人士担任。

图 10 英伟达主要管理团队

数据来源：英伟达官网

图 11 英伟达主要组织架构



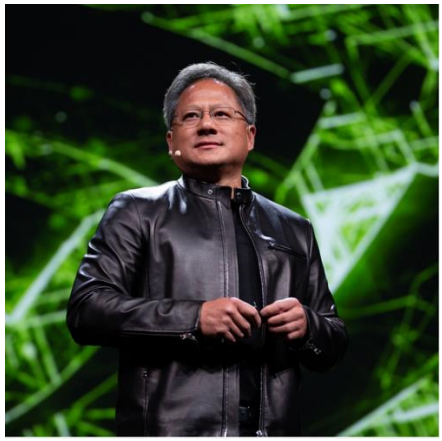
数据来源：theofficialboard，国泰君安证券研究

1.4. 黄仁勋：不止是 CEO，更是精神领袖

作为创始人、CEO 与精神领袖，黄仁勋带领英伟达创造 AI 龙头奇迹。黄仁勋，1963 年出生于中国台北，美籍华人。作为公司创始人，黄仁勋历经 30 载依旧任英伟达的总裁兼首席执行官。他曾被《哈佛商业评论》和 Glassdoor 评为全球最佳 CEO 和受雇员评价最高的 CEO。2021 年 9 月，黄仁勋登上《时代》杂志封面，成为《时代》杂志 2021 年世界最具影响力的百位人物之一。

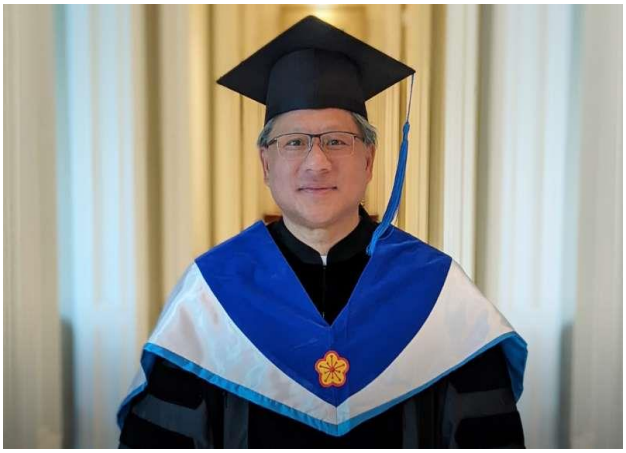
兼具技术与业务背景，葆有实干与远见特质。黄仁勋 1984 年于俄勒冈州立大学取得学士学位，1990 年获得斯坦福大学硕士学位，1983-1985 年间，其担任 AMD 芯片工程师，而后跳槽至 LSI Logic 继续从事芯片设计，在 LSI Logic 任职期间，黄仁勋转岗销售部门，因其出色的表现很快晋升为部门经理，从此踏上管理岗位。在 1993 年英伟达筹建之初，因其出色的技术和业务背景，克里斯与普雷艾姆推举黄仁勋担任英伟达总裁兼 CEO。2020 年，黄仁勋获颁台湾大学名誉博士学位，以表彰其在人工智能与高效能计算领域的伟大贡献。

图 12 英伟达 CEO 黄仁勋



数据来源：英伟达官网

图 13 黄仁勋获台湾大学名誉博士学位



数据来源：英伟达官网

2. 技术与产品高筑壁垒，让 AI 照进现实

细分英伟达的产品线，我们可将其划分为硬件产品、软件平台、应用框架三个维度。同时英伟达基于“硬件+软件”的技术优势，同时依托面向行业打造的应用框架，提供了对于细分行业定制的行业解决方案。

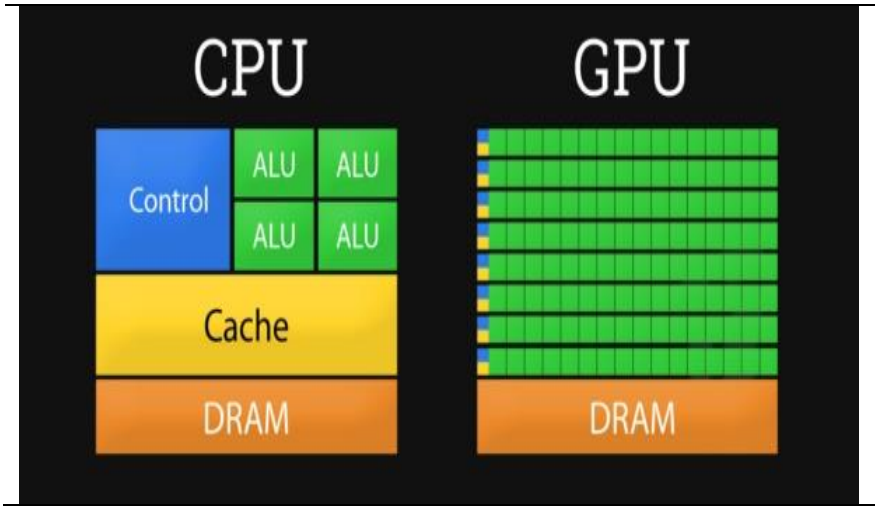
图 14 英伟达产品架构图

数据来源：英伟达官网，国泰君安证券研究

2.1. 硬件产品始于 GPU，但不止 GPU

英伟达首创 GPU 产品，推动处理器中逻辑运算单元数量增长。CPU 是电脑的中央处理器，同时也是电脑的控制和运算核心，能够解释计算机发出的指令。而 GPU 是电脑的图形处理器，最初主要用于进行图像运算工作。英伟达研发世界上首款 GPU GeForce 256，开 GPU 之先河，令 GPU 逐渐演化为普遍使用的并行处理器。整体而言，GPU 和 CPU 同为基于芯片的微处理器，是重要的计算引擎。CPU 拥有更大的逻辑运算单元和控制单元，同时拥有更大的缓存空间，但 GPU 却拥有更多的逻辑运算单元数量。

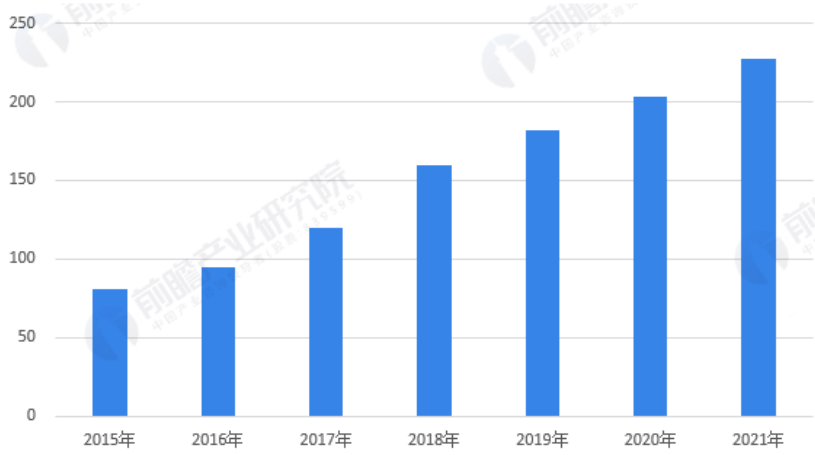
图 15 CPU 与 GPU 的结构区别



数据来源：Github

需求激增催化 GPU 市场规模爆发式增长。IC Insights 数据显示，2015 年至 2021 年间，全球 GPU 芯片市场规模年均增速超 20%，2021 年，全球 GPU 芯片市场规模已超过 220 亿美元，全年出货总量超过 4.6 亿片。我们认为，目前 GPU 仍占全球 AI 芯片的主导地位。

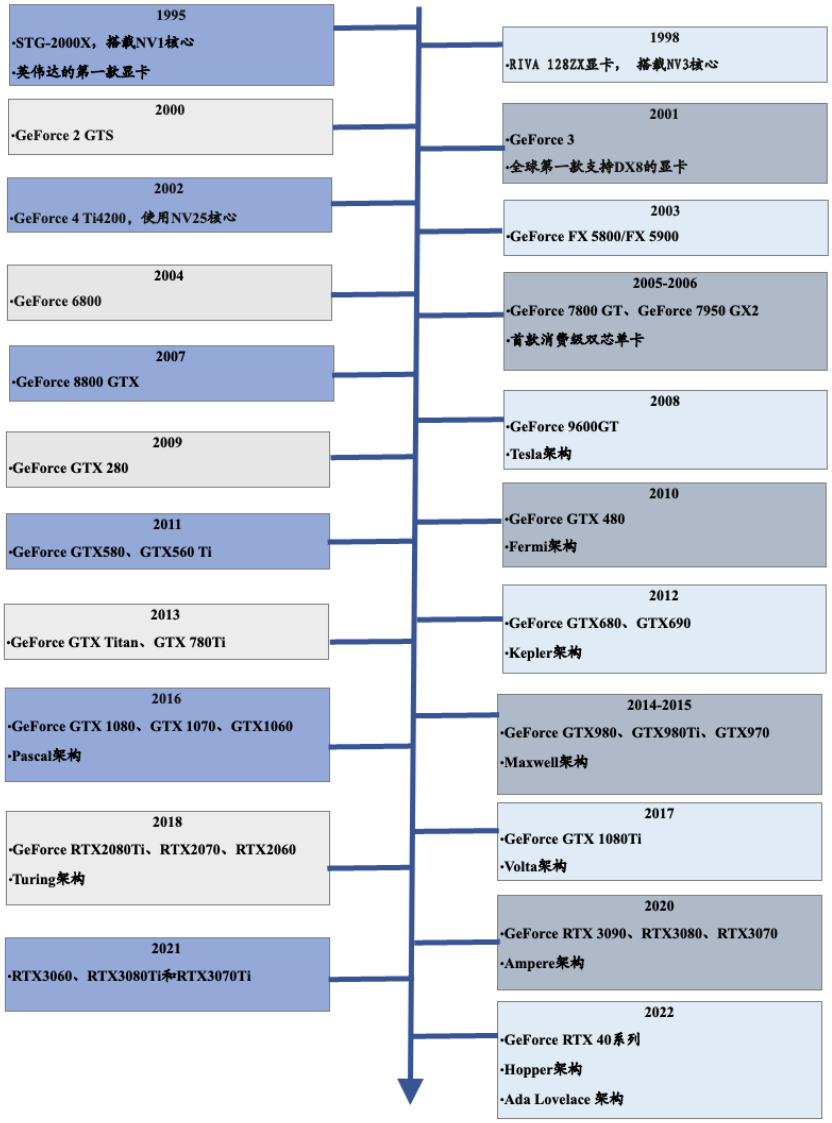
图 16 2015-2021 年全球 GPU 芯片行业市场规模（亿美元）



数据来源：IC Insights，前瞻产业研究院

英伟达深耕 GPU 业务，主要显卡产品更迭迅速。英伟达主要显卡产品以 GeForce 为前缀命名，自 2000 年发布 GeForce 2 GTS 起，GeForce 系列划分出多种型号，直至目前，英伟达在售的主要显卡产品包括 GeForce16、GeForce20、GeForce30、GeForce40 等。从 GPU 架构角度，自 2008 年发布 Tesla 架构后，英伟达依次发布了 Fermi、Kepler、Maxwell、Pascal、Volta、Turing、Ampere、Hopper、AdaLovelace 等 GPU 微架构，近年来 GPU 架构的更新速度显著加快。

图 17 英伟达主要显卡和架构发展史



数据来源：英伟达官网，CSDN，知乎，国泰君安证券研究

表 1 2010 年至今英伟达 GPU 架构发展

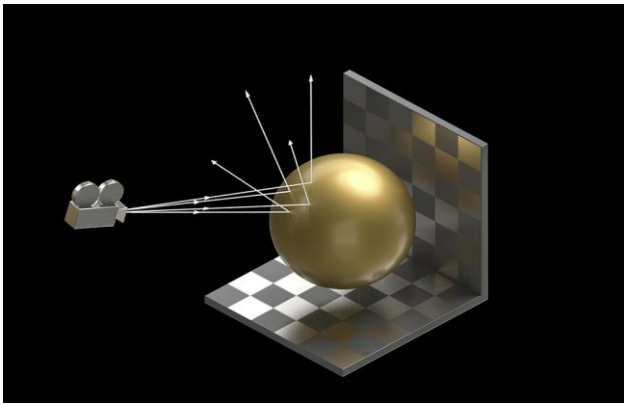
架构	中文名称	发布时间	核心参数	特点&优势	纳米制程	代表型号
Fermi	费米	2010	16 个 SM，每个 SM 包含 32 个 CUDA Cores，一共 512 CUDA Cores	首个完整 GPU 计算架构，支持与共享存储结合的 Cache 层次 GPU 架构，支持 ECC GPU 架构	40/28nm 30 亿晶体管	Quadro 7000
Kepler	开普勒	2012	15 个 SMX，每个 SMX 包括 192 个 FP32+64 个 FP64CUDA Cores	游戏性能大幅提升，首次支持 GPU Direct 技术	28nm 71 亿晶体管	K80K40M
Maxwell	麦克斯韦	2014	16 个 SM，每个 SM 包括 4 个处理块，每个处理块包括 32 个 CUDA Cores+8 个 LD/ST Unit+8 SFU	每组 SM 单元从 192 个减少到每组 128 个，每个 SMM 单元拥有更多逻辑控制电路	28nm 80 亿晶体管	M5000 M4000GTX 9XX 系列
Pascal	帕斯卡	2016	GP100 有 60 个 SM，每个 SM 包括 64 个 CUDA Cores，32	NVLink 第一代双向互联带宽	16nm 153 亿晶体管	P100 P6000 TTX1080

			个 DP Cores	160GB/s，P100 拥有 56 个 SM HBM		
Volta	伏特	2017	80 个 SM，每个 SM 包括 32 个 FP64+64 Int32+64 FP32+8 个 Tensor Cores	NVLink2.0，Tensor Cores 第一代支持 AI 运算	12nm 211 亿晶体管	V100 TiTan V
Turing	图灵	2018	102 核心 92 个 SM，SM 重新设计，每个 SM 包含 64 个 Int32+64 个 FP32+8 个 Tensor Cores	Tensor Core2.0，RT Core 1.0	12nm 186 亿晶体管	T4，2080TI RTX 5000
Ampere	安培	2020	108 个 SM，每个 SM 包含 64 个 FP32+64 个 INT32+32 个 FP64+4 个 Tensor Cores	Tensor Core3.0RT Core2.0，NVLink 3.0，结构稀疏性矩阵 MIG1.0	7nm 283 亿晶体管	A100 A30 系列
Hopper	赫柏	2022	132 个 SM，每个 SM 包含 128 个 FP32+64 个 INT32+64 个 FP64+4 个 Tensor Cores	TensorCore4.0，NVlink4.0，结构稀疏性矩阵 MIG2.0	4nm 800 亿晶体管	H100
Ada Lovelace	爱达·洛芙莱斯	2022	144 个 SM，每个 SM 包含 128 CUDA Cores，1 个第三代 RT Core，4 个第四代 Tensor Core，四个纹理单元、一个 256 KB 的寄存器文件和 128 KB 的 L1/共享内存	TensorCore4.0，RT Core 3.0	4nm 763 亿晶体管	RTX 40 系列

数据来源：哔哩哔哩，英伟达官网，国泰君安证券研究

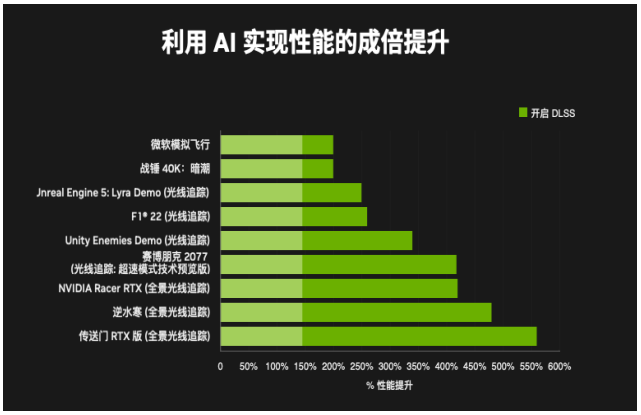
Ada Lovelace 架构为英伟达 GeForce RTX 40 系列显卡提供动力支持。Ada Lovelace 架构主要用于游戏显卡的生产，其采用的第四代 Tensor Core 使用首次推出的全新 FP8 Transformer 引擎，能够提升四倍吞吐量；其中的第三代 RT Core 配备全新 Opacity Micromap 和 Displaced Micro-Mesh 引擎，可大幅提升进行光线追踪的速度，所占用的显存只有之前的二十分之一；并且，Ada Lovelace 架构可使用 DLSS 3（深度学习超采样）算法，可对多个分辨率较低的图像进行采样，并使用先前帧的运动数据和反馈来重建原生质量图像，从而创建更多高质量帧，显著提升 FPS（Frames per second），目前已应用于 200 多款游戏和应用。

图 18 英伟达第三代 RT Core 实现实时光线追踪



数据来源：英伟达官网

图 19 DLSS 利用 AI 实现 FPS 性能的成倍提升



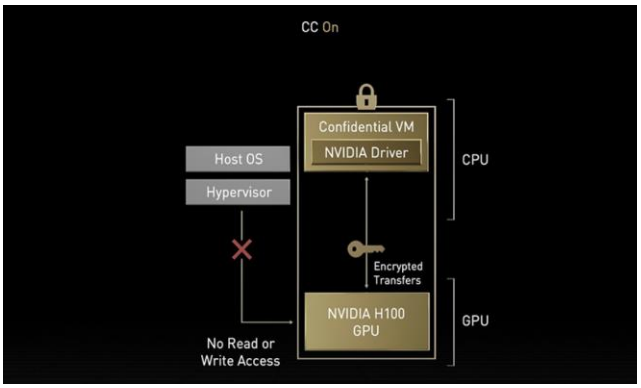
数据来源：英伟达官网

Hopper 架构为加速计算实现新的巨大飞跃。与 Ada Lovelace 架构不同，Hopper 架构主要用以打造加速计算平台。Hopper 架构以 Transformer 为加速引擎，其中的 Hopper Tensor Core 能够大幅加速 Transformer 模型的 AI 计算。Hopper 架构同时搭载 NVLink Switch 系统，NVLink 作为一种纵向扩展互联技术，与新的外部 NVLink 交换机结合使用时，系统可以跨多个服务器以每个 GPU 900 GB/s 的双向带宽扩展多 GPU IO，能够满足每个在 GPU 之间实现无缝高速通信的多节点、多 GPU 系统的需求。同时，Hopper 架构还采用了具有机密计算功能的加速计算平台 CCX，以保障数据处理期间的 GPU 使用安全。

图 20 NVLink 技术可提升系统吞吐量

数据来源：英伟达官网

图 21 CCX 在 AI 模型和应用的各阶段保障数据安全



数据来源：英伟达官网

GeForce RTX 40 显卡基于 Ada Lovelace 架构打造。英伟达最新的显卡为 GeForce RTX 40 系列，GeForce RTX 40 搭载英伟达最先进的 GPU，其采用新型 SM 多单元流处理器将性能功耗比提升 2 倍，并应用第四代 Tensor Core 提升计算性能，达到 1.4 Tensor-petaFLOPS，同时，搭载的第三代 RT Core 实现了光线追踪性能的两倍提升，可模拟真实世界中的光线特性，能够显著提升玩家游戏体验。

图 22 英伟达 RTX 40 显卡性能优越

数据来源：英伟达官网

Tensor Core 是自 Volta 架构以来英伟达的核心技术，为 HPC 和 AI

实现大规模加速。Tensor Core 可实现混合精度计算，动态调整算力，从而在保持准确性的同时提高吞吐量，Tensor Core 提供了一整套精度（TF32、Bfloat16 浮点运算性能、FP16、FP8 和 INT8 等），确保实现出色的通用性和性能。目前，Tensor Core 已广泛用于 AI 训练和推理。

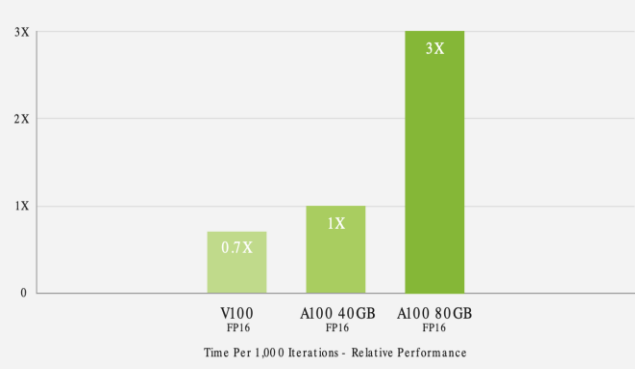
表 2 英伟达 Tensor Core 的迭代

版本	架构	支持精度	性能
第一代	Volta	FP16	Volta 配备 640 个 Tensor Core，通过 FP16 和 FP32 下的混合精度矩阵乘法提供了突破性的性能。与 Pascal 相比，用于训练的峰值 TFLOPS 性能提升高达 12 倍，用于推理的峰值 TFLOPS 性能提升高达 6 倍。
第二代	Turing	FP16、INT8、INT4、INT1	每秒可提供高达 500 万亿次的张量运算，能够极大加速 AI 增强功能，如去噪、分辨率缩放和视频调速，并有助构建具有全新超强功能的应用程序。
第三代	Ampere	FP64、TF32、bfloat16、FP16、INT8、INT4、INT1	通过使用新的精度（TF32 和 FP64）来加速和简化 AI 使用，并将 Tensor Core 的强大功能扩展至 HPC，可为 AI 训练和推理创建高度通用的加速器。
第四代	Hopper	FP64、TF32、Bfloat16、FP16、FP8、INT8	使用新的 8 位浮点精度（FP8），可为万亿参数模型训练提供比 FP16 高 6 倍的性能，使用 TF32、FP64、FP16 和 INT8 精度，将性能提升 3 倍，能够加速处理各种工作负载。

数据来源：英伟达官网，国泰君安证券研究

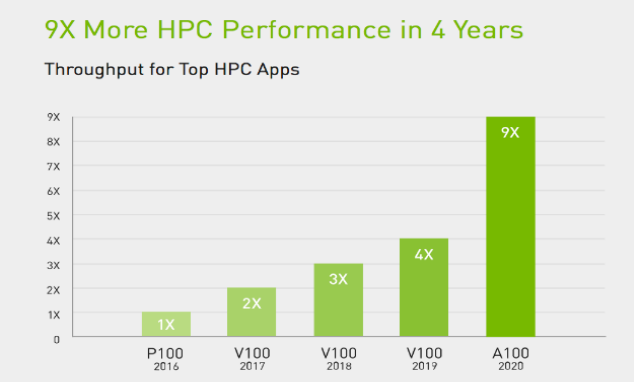
从 A100 到 H100 为 AI 训练和推理带来历史性变革，成就加速计算的数量级飞跃。H100 的上一代产品，2020 年推出的 A100，较 2016 年的 P100 已在四年间将高性能计算的运行速度提升至 9 倍，但 H100 真正实现了数量级的飞跃。H100 基于 Hopper 架构的卓越优势，配备第四代 Tensor Core 和 Transformer 引擎，使双精度 Tensor Core 的每秒浮点运算量提升 3 倍。与 A100 相比，H100 可为多专家模型（MoE）提供高九倍的训练速度。推理端，H100 表现同样优越，H100 可将推理速度提高至 A100 的 30 倍，并提供超低的延迟，在减少内存占用和提高计算性能的同时，大语言模型的准确度仍旧得到保持。

图 23 A100 提升深度学习训练速度



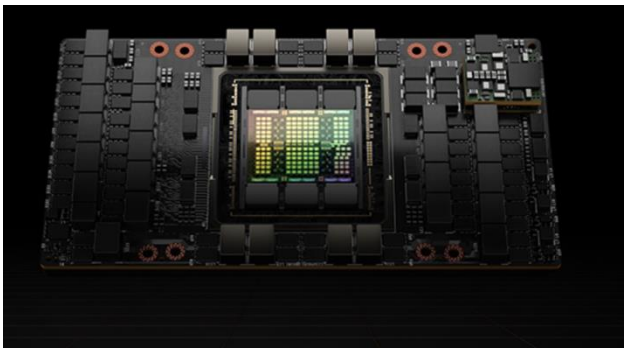
数据来源：英伟达官网

图 24 A100 将 HPC 的运行速度提升至 9 倍



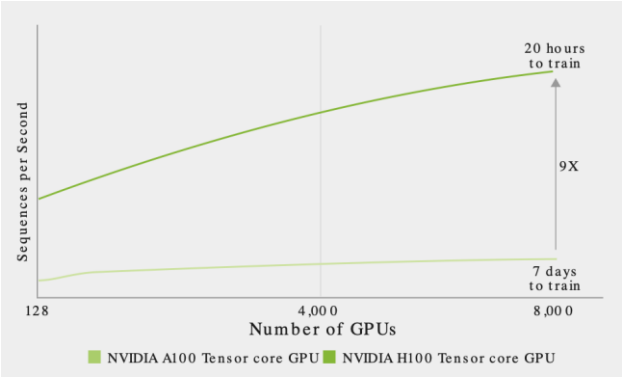
数据来源：英伟达官网

图 25 英伟达 H100 Tensor Core GPU



数据来源：英伟达官网

图 26 H100 提供高达 9 倍的 AI 训练速度

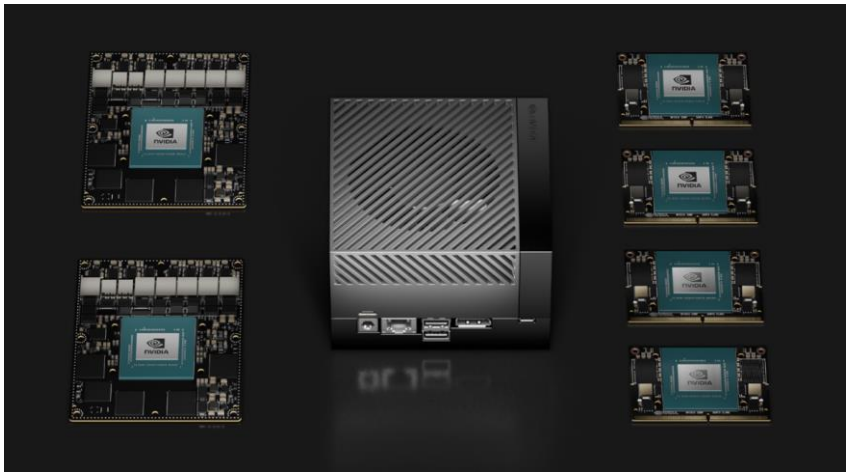


数据来源：英伟达官网

Jetson 嵌入式系统打造灵活且可拓展的嵌入式硬件解决方案。Jetson 是用于自主机器和其他嵌入式应用的先进平台，该平台包括 Jetson 模组、用于加速软件的 JetPack SDK，以及包含传感器、SDK、服务和产品的生态系统。其中，每一个 Jetson 均包含了 CPU、GPU、内存、电源管理和高速接口，是一个完整的系统模组，并且所有 Jetson 模组均由同一软件堆栈提供支持，意味着企业只需一次开发即可在任意地方部署。目前英伟达在售的 Jetson 主要包括 Jetson Orin 系列、Jetson Xavier 系列、Jetson TX2 系列和 Jetson Nano，能够在数据中心和云部署的技术基础上为 AI 应用提供端到端加速。

以 Jetson Orin 为例，Jetson Orin 模组可实现每秒 275 万亿次浮点运算(TOPS)的算力，性能是上一代产品的 8 倍，可适用于多个并发 AI 推理，此外它还可以通过高速接口为多个传感器提供支持，这使得 Jetson Orin 成为机器人开发新时代的理想解决方案。量产级 Jetson Orin 模组能够为企业提供在边缘构建自主机器所需的性能和能效，以帮助企业更快地进入市场。并且英伟达提供 Jetson AGX Orin 开发者套件，可实现对整个 Jetson Orin 模组系列进行模拟。

图 27 Jetson Orin 模组

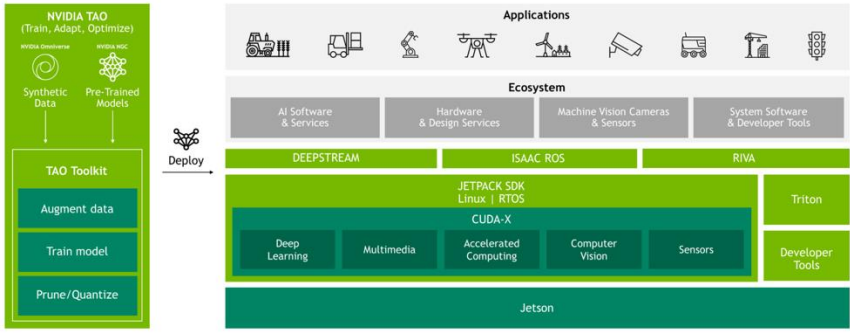


数据来源：英伟达官网

Jetson 与 VIMA 将有望与具身智能相结合，直面 AI 的下一波浪潮。具身智能是能理解、推理、并与物理世界互动的智能系统。ITF World 2023

半导体大会上，黄仁勋表示，人工智能下一个浪潮将是"具身智能"，同时英伟达也公布了 Nvidia VIMA，VIMA 是一个多模态具身人工智能系统，能够在视觉文本提示的指导下执行复杂的任务。我们认为，伴随着 Jetson 和 VIMA 的系统逐步研发完善，英伟达将成为推动具身智能发展的引领者。

图 28 Jetson 嵌入式软件堆栈



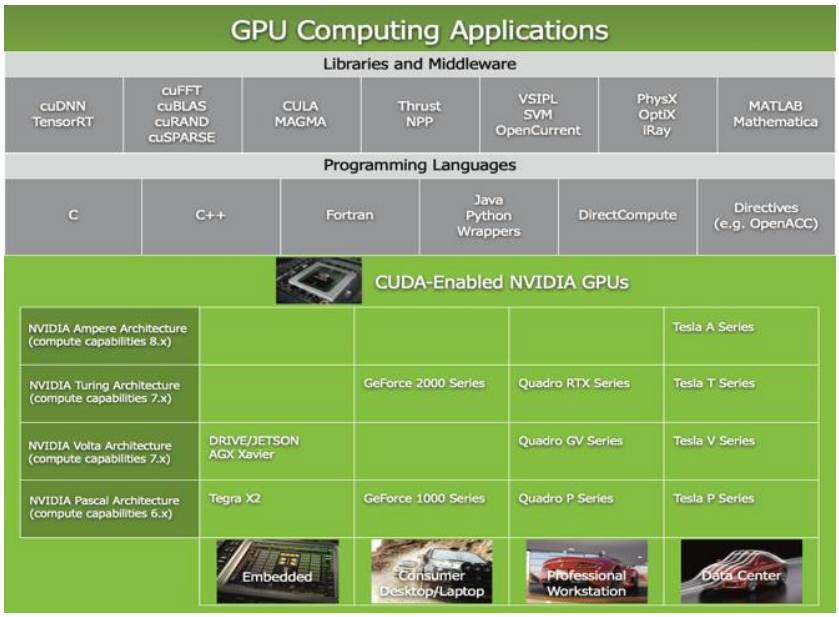
数据来源：英伟达官网

整体而言，英伟达在边缘的优势能够为扩大市场提供更多可能性。通过使用 Jetson，企业可以自由开发和部署 AI 赋能的机器人、无人机、IVA 应用和其他可以自我思考的自主机器。中小企业和初创企业能够承担 Jetson 的部署开销，以此开发自主机器和其他嵌入式应用，且英伟达在嵌入式技术领域同时具有领先优势，我们对其市场积极看好。

2.2. 软件平台带来更多可能，奠定生态帝国基石

CUDA 构筑软件业务底层框架基石，打造对接行业解决方案的开发平台。英伟达于 2006 年发布 CUDA，成为首款 GPU 通用计算解决方案。借助 CUDA 工具包，开发者可以在 GPU 加速的嵌入式系统、桌面工作站、企业数据中心、基于云的平台和 HPC 超级计算机上开发、优化和部署应用程序。CUDA 工具包主要包括 GPU 加速库、调试和优化工具、C/C++ 编译器以及用于部署应用程序的运行环境库。不论是图像处理、计算科学亦或是深度学习，基于 CUDA 开发的应用都已部署到无数个 GPU 中。

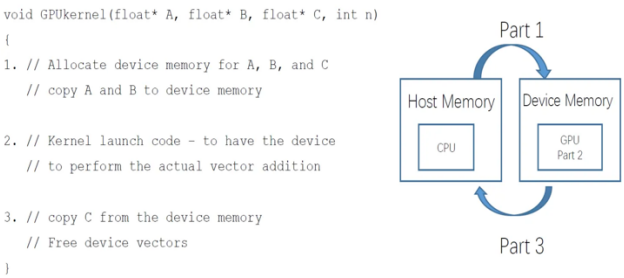
图 29 CUDA 附带的软件环境



数据来源：英伟达官网

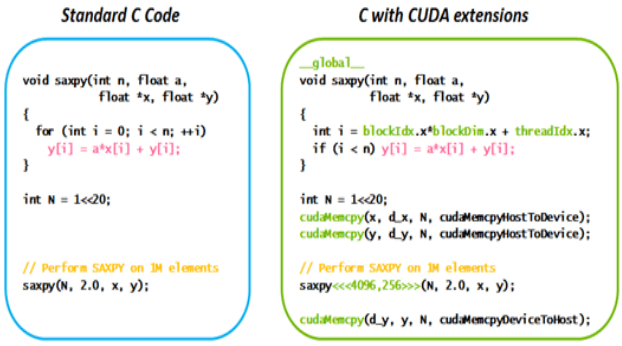
开发者从此不再需要通过写大量的底层语言代码对 GPU 进行调用。CUDA 与 C 语言的框架较为接近，作为一种类 C 语言，CUDA 对于开发者而言上手难度较小，且同时也支持 Python、Java 等主流编程语言。此外，一个 CUDA 程序可分为三个部分：第一，从主机端申请调用 GPU，把要拷贝的内容从主机内存拷贝到 GPU 内；第二，GPU 中的核函数对拷贝内容进行运算；第三，把运算结果从 GPU 拷贝到申请的主机端，并释放 GPU 的显存和内存，整个过程较为清晰且易操作。可以说，CUDA 是搭建了一个帮助开发者通过高级编程语言使用 GPT 完成特定行业需求功能的平台，英伟达也因此打造了一个“硬件+软件平台”的生态帝国。

图 30 一个 CUDA 程序可分为三部分



数据来源：知乎

图 31 标准 C 语言代码和应用 CUDA 代码的区别

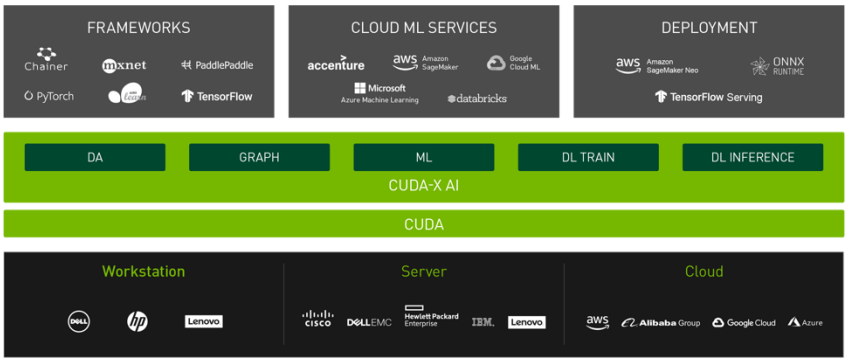


数据来源：英伟达官网

打造软件加速库的集合 CUDA-X AI，帮助现代 AI 应用程序加速运行。 CUDA-X AI 作为软件加速库集合，建立在 CUDA 之上，它的软件加速库集成到所有深度学习框架和常用的数据科学软件中，为深度学习、机器学习和高性能计算提供优化功能。库包括 cuDNN（用于加速深度学习基元）、cuML（用于加速数据科学工作流程和机器学习算法）、TensorRT（用于优化受训模型的推理性能）、cuDF（用于访问 pandas 等数据科学

API）、cuGraph（用于在图形上执行高性能分析），以及超过 13 个的其他库。CUDA-X AI 已成为领先的云平台，包括 AWS、Microsoft Azure 和 Google Cloud 在内的一部分，而且可以通过 NGC 网站逐个地或作为容器化的软件栈免费下载。

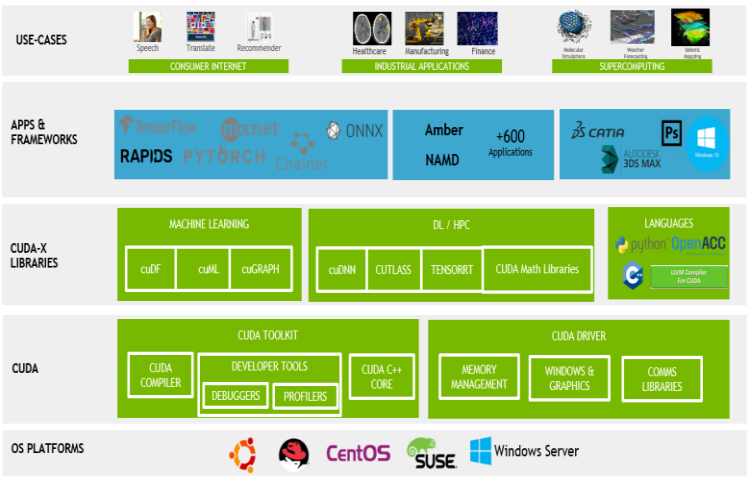
图 32 CUDA-X AI 帮助现代 AI 应用程序加速运行



数据来源：英伟达官网

CUDA 打造高兼容性的 GPU 通用平台，推动 GPU 应用场景持续扩展。CUDA 可以充当英伟达各 GPU 系列的通用平台，因此开发者可以跨 GPU 配置部署并扩展应用。CUDA 最初用于辅助 GeForce 提升游戏开发效率，但随着 CUDA 的高兼容性优势彰显，英伟达将 GPU 的应用领域拓展至计算科学和深度学习领域。因此，通过 CUDA 开发的数千个应用目前已部署到嵌入式系统、工作站、数据中心和云中的 GPU。同时，CUDA 打造了开发者社区，提供开发者自由分享经验的途径，并提供大量代码库资源。我们认为，目前 CUDA 已形成极高的准入壁垒，也成为了英伟达持续扩展人工智能领域市场的品牌影响力来源。

图 33 CUDA 架构

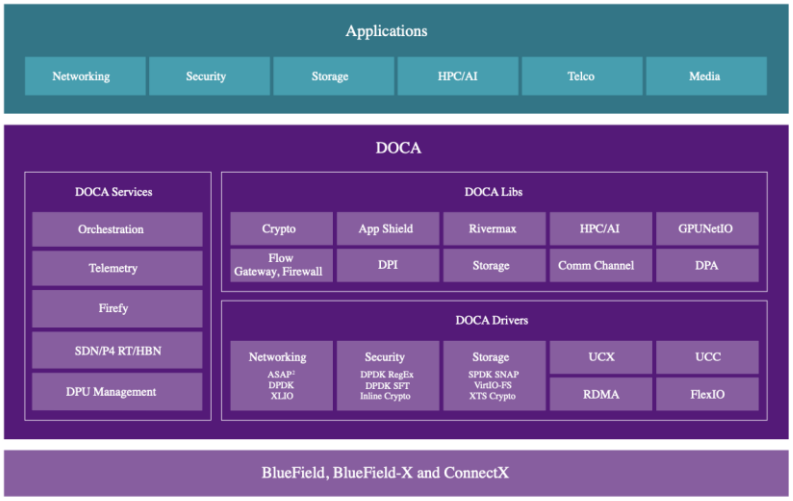


数据来源：英伟达官网

DOCA 与 DPU 结合打造开发平台，成为激发 DPU 潜力的关键。借助 DOCA，开发者可通过创建软件定义、云原生、DPU 加速的服务来对未来的数据中心基础设施进行编程。具体而言，DOCA 软件由软件开发套件（SDK）和运行时（Runtime）环境组成，SDK 中包含了系统的软件框架，Runtime 则包括用于在整个数据中心的成百上千个 DPU 上配置、

部署和编排容器化服务的工具。DOCA 与 DPU 的结合能够开发具备突破性的网络、安全和存储性能的应用，有效满足现代数据中心日益增长的性能和安全需求。

图 34 DOCA 2.0 软件框架

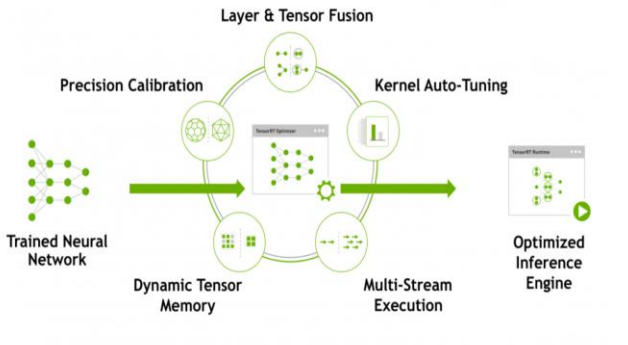


数据来源：英伟达官网

打造深度学习推理优化器 TensorRT，显著提高了 GPU 上的深度学习推理性能。 TensorRT 是英伟达一款高性能推理平台，此 SDK 包含深度学习推理优化器和运行时环境，可为深度学习推理应用提供低延迟和高吞吐量。与仅使用 CPU 的平台相比，TensorRT 可使吞吐量提升高达 40 倍。借助 TensorRT，开发者可以在所有主要框架中优化训练的神经网络模型，提升模型激活精度校准，并最终将模型部署到超大规模数据中心、嵌入式或汽车产品平台中。

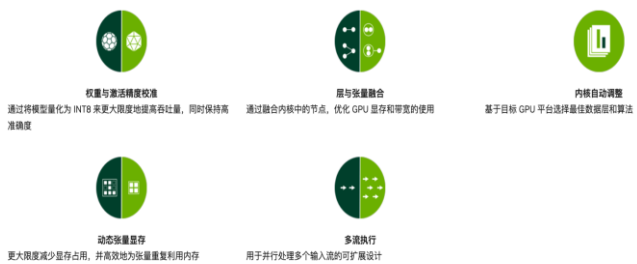
TensorRT 以 CUDA 为基础构建，同时与开发框架紧密集成。 TensorRT 以 CUDA 为基础，可帮助开发者利用 CUDA-X 中的库、开发工具和技术，针对人工智能、自主机器、高性能计算和图形优化所有深度学习框架中的推理。通过 TensorRT 的使用，可以对训练的神经网络模型进行 INT8 和 FP16 优化，例如视频流式传输、语音识别、推荐算法和自然语言处理，并将优化后的模型部署至应用平台。同时 TensorRT 也与 Tensorflow、MATLAB 的深度学习框架集成，可以将预训练的模型导入至 TensorRT 进行推理，具备较高的兼容性。

图 35 利用 TensorRT 进行模型优化



数据来源：英伟达官网

图 36 TensorRT 的优化方法



数据来源：英伟达官网

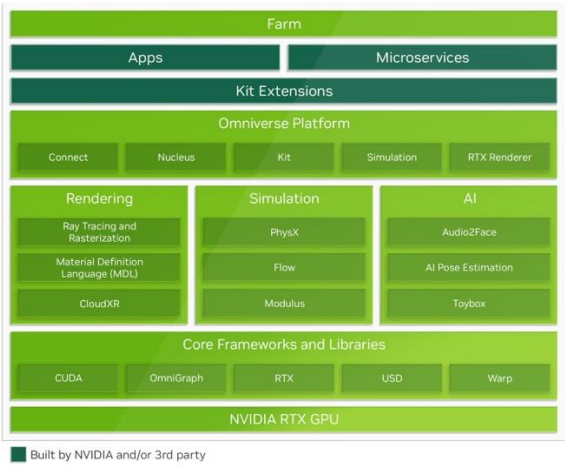
2.3. 应用框架构筑封装 SDK，打造标准行业场景

SDK 助力标准行业场景搭建，大幅提升开发效率和性能。SDK 全称 Software Development Kit，指为特定的硬件平台、软件框架、操作系统等建立应用程序时所使用的开发工具的集合。英伟达基于自身丰富的“软件+硬件”一体化优势，将其进行优化并封装为 SDK，形成了自身完备的应用框架体系，为行业中突出问题的解决打造了标准行业场景。完备的 SDK 体系有助于更大程度提高开发者的工作效率，相关应用框架的性能和可移植性也将因此得到显著提升。

2.3.1. 元宇宙应用-Omniverse

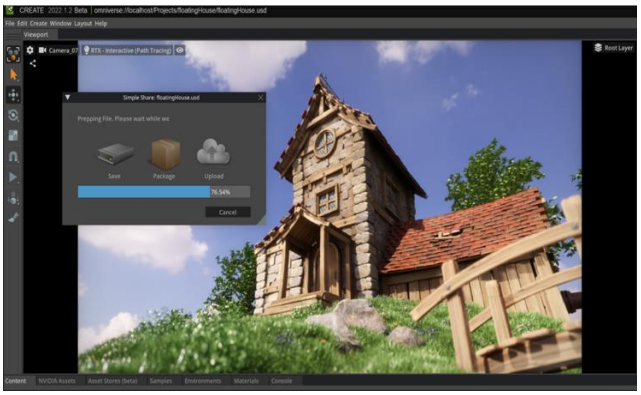
开创元宇宙模拟平台 Omniverse，共同设计运行虚拟世界和数字孪生。Omniverse 是一个基于 USD（Universal Scene Description）的可扩展平台，在 Omniverse 中，艺术家可以使用 3D 工具创作具备全设计保真度的实时虚拟世界，企业可以通过数字孪生模型在产品投产前实时设计、仿真和优化他们的产品、设备或流程。目前，Omniverse 拥有 15 万余名个人用户和 300 余家企业用户。此外，英伟达也推出了 LaaS 产品 Omniverse Cloud，可连接在云端、边缘设备或本地运行的 Omniverse 应用，实现在任何位置设计、发布和体验元宇宙应用，例如，借助 Omniverse Cloud Simple Share 服务，只需单击即可在线打包和共享 Omniverse 场景。

图 37 Omniverse 平台架构



数据来源：英伟达官网

图 38 Omniverse Cloud 可共享 Omniverse 场景



数据来源：英伟达官网

2.3.2. 云端 AI 视频流-Maxine

Maxine 提供 GPU 加速的 AI SDK 和云原生服务，可用于部署实时增强音频、视频和增强现实效果的 AI 功能。Maxine 使用最先进的模型创造出可以使用标准麦克风和摄像头设备实现的高质量效果。其中，Audio Effects SDK 提供基于 AI 的音频质量增强算法，提高窄带、宽带和超宽带音频的端到端对话质量，包括提供去噪、回声消除、音频超分辨率等效果，而 Video Effects SDK 提供虚拟背景、放大器、减少伪影和眼神接触等 AI 的 GPU 加速视频效果。Maxine 可以部署在本地、云端或边缘，微服务也可以在应用程序中独立管理和部署，从而加快开发时间。

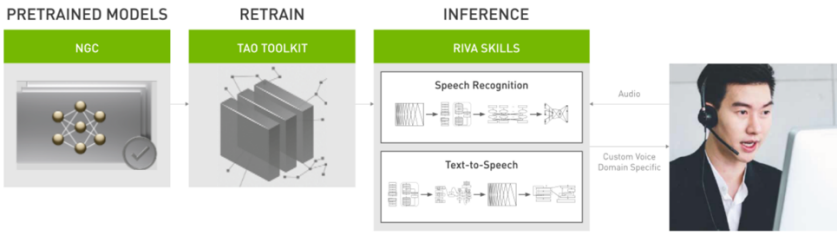
图 39 英伟达为诸多通信服务商提供 Maxine 服务

数据来源：英伟达官网

2.3.3. 语音 AI-Riva

Riva 构建定制实时语音 AI 应用，形成端到端语音工作流程。随着基于语音的应用在全球的需求激增，这要求了语音 AI 应用需识别行业特定术语，并跨多种语言作出自然的实时响应。Riva 包含先进的实时自动语音识别(ASR)和文字转语音 (TTS)功能。用户可选择预训练的语音模型，在自定义数据集中使用 TAO 工具套件对模型进行微调，能将特定领域模型的开发速度提升 10 倍。Riva 的高性能推理依赖于 TensorRT，并已完全容器化，可以轻松扩展到数千个并行流。

图 40 Riva 训练和部署端到端对话式 AI 制作流程

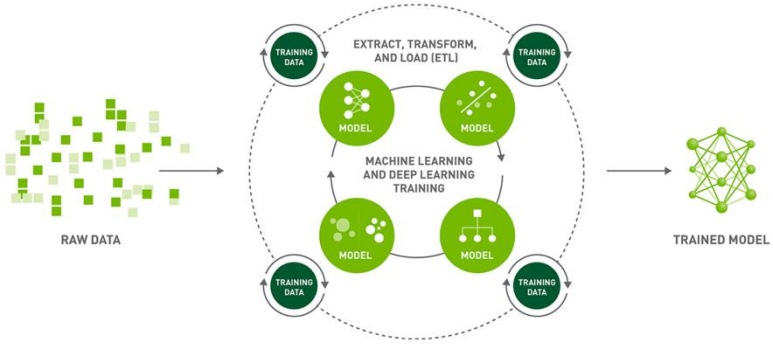


数据来源：英伟达官网

2.3.4. 数据分析-RAPIDS

RAPIDS 为全新高性能数据科学生态系统奠定了基础，并通过互操作性降低了新库的准入门槛。英伟达打造了由一系列开源软件库和 API 组成的 PAPIDS 系统，支持从数据读取和预处理、模型训练直到可视化的全数据科学工作流程。通过集成领先的数据科学框架（如 Apache Spark、cuPY、Dask 和 Numba）以及众多深度学习框架（如 PyTorch、TensorFlow 和 Apache MxNet），RAPIDS 可帮助扩大采用范围并支持集成其他内容。整体而言，RAPIDS 以 CUDA-X AI 为基础，融合了英伟达在显卡、机器学习、深度学习、高性能计算(HPC)等领域多年来的发展成果。

图 41 RAPIDS 用以模型训练、评估、迭代和再训练

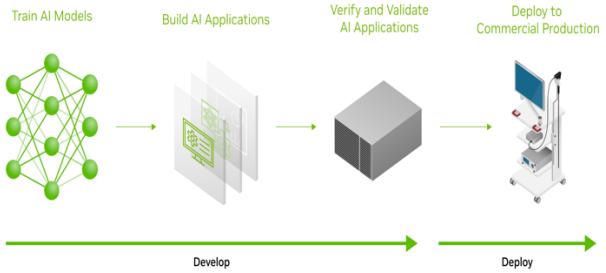


数据来源：英伟达官网

2.3.5. 医疗健康-Clara

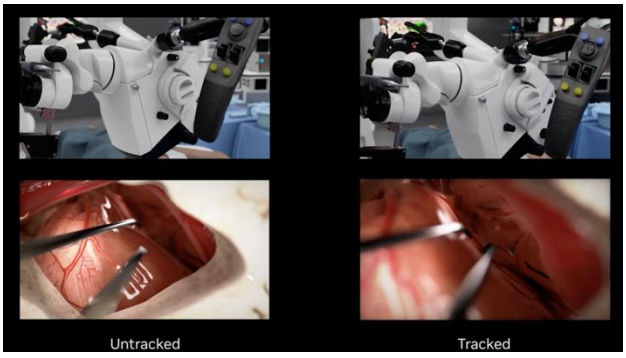
打造 AI 助力的医疗健康平台 Clara，助力新一代医疗设备和生物医学研究。Clara 主要包含 Holoscan、Parabricks、Discovery 和 Guaradian 四大应用，分别用于医疗影像和医疗设备、基因组学、生物制药和智慧医院建设。以 Holoscan 为例，开发者可以构建设备并将 AI 应用直接部署到临床环境中，使用准确的数字孪生模拟手术环境有助于提高手术效率并缩短患者留在手术室内的时间。其中，MONAI 是专用的开源医疗 AI 框架，目标是通过构建一个强大的软件框架来加快创新和临床转化的步伐。

图 42 Holoscan 简化了医疗设备的开发和部署



数据来源：英伟达官网

图 43 Holoscan 可使用数字孪生模拟手术环境

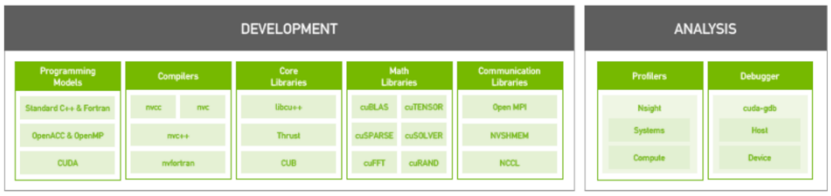


数据来源：英伟达官网

2.3.6. 高性能计算

HPC 软件开发套件助力高性能计算。HPC SDK C、C++和 Fortran 编译器支持使用标准 C++和 Fortran、OpenACC 指令和 CUDA 对 HPC 建模和模拟应用程序进行GPU加速。GPU加速的数学库提高了常见HPC 算法的性能，而优化的通信库支持基于标准的多 GPU 和可扩展系统编程。性能分析和调试工具可简化 HPC 应用程序的移植和优化，而容器化工具可在本地或云端轻松部署。

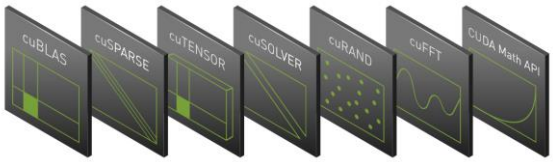
图 44 HPC SDK 框架



数据来源：英伟达官网

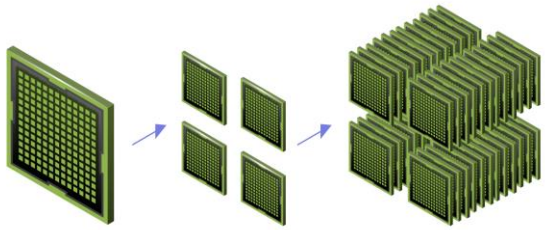
HPC SDK 的主要功能包括 GPU 数学库、Tensor Core 优化、CPU 优化、多 GPU 编程、可拓展系统编程、Nsight 性能分析等。其中，GPU 加速的数学库适用于计算密集型应用，cuBLAS 和 cuSOLVER 库可提供来自 LAPACK 的各种 BLAS 例程以及核心例程的多 GPU 的实施，并尽可能自动使用 GPU Tensor Core。集合通信库 (NCCL) 能够实现多 GPU 编程，使用 MPI 兼容的 all-gather、all-reduce、broadcast、reduce 和 reduce-scatter 例程实现高度优化的多 GPU 和多节点集合通信基元，以利用 HPC 服务器节点内和跨 HPC 服务器节点的所有可用 GPU。

图 45 HPC SDK 的 GPU 数学库



数据来源：英伟达官网

图 46 NCCL 可实现多 GPU 编程

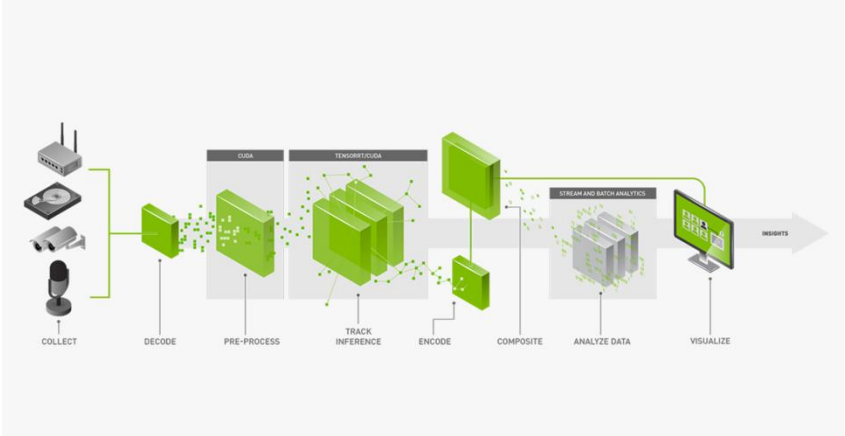


数据来源：英伟达官网

2.3.7. 智能视频分析-Metropolis

Metropolics 将像素转化为见解，致力打造全方位智能视频分析应用框架。Metropolics 将可视化数据和 AI 整合，处理数万亿传感器生成的海量数据，提高众多行业的运营效率和安全性，企业可以创建、部署和扩展从边缘到云端的 AI 和物联网应用。DeepStream SDK 是由 AI 驱动的实时视频分析 SDK，可以显著提高性能和吞吐量；TAO 工具包借助计算机视觉特定的预训练模型和功能，加速深度学习训练；TensorRT 将高性能计算机视觉推理应用程序从 Jetson Nano 部署到边缘的 T4 服务器上。目前，Metropolics 已广泛用于智慧城市建设、零售物流、医疗健康、工业和制造业等。

图 47 DeepStream SDK 实现从云端到边缘的无缝开发

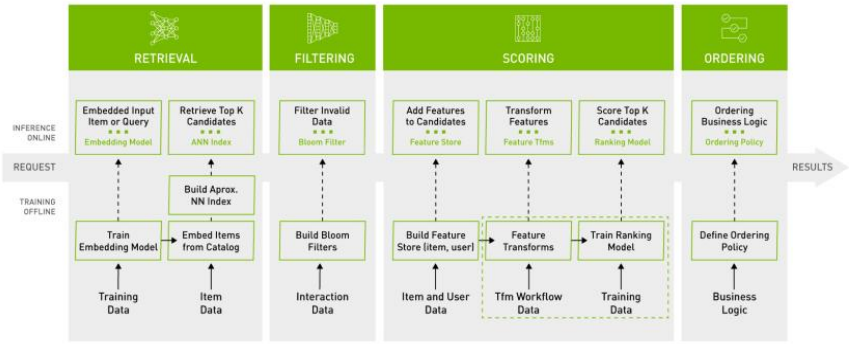


数据来源：英伟达官网

2.3.8. 推荐系统-Merlin

英伟达提供用于大规模构建高性能推荐系统的开源框架 Merlin。Merlin 使数据科学家、机器学习工程师和其他研究人员能够大规模构建高性能的推荐器。Merlin 框架包括库、方法和工具，通过实现常见的预处理、特征工程、训练、推理和生产部署，简化了推荐算法的构建。Merlin 组件和功能经过优化，可支持数百 TB 数据的检索、过滤、评分和排序，并可以通过易于使用的 API 访问。

图 48 Merlin 推荐算法流程

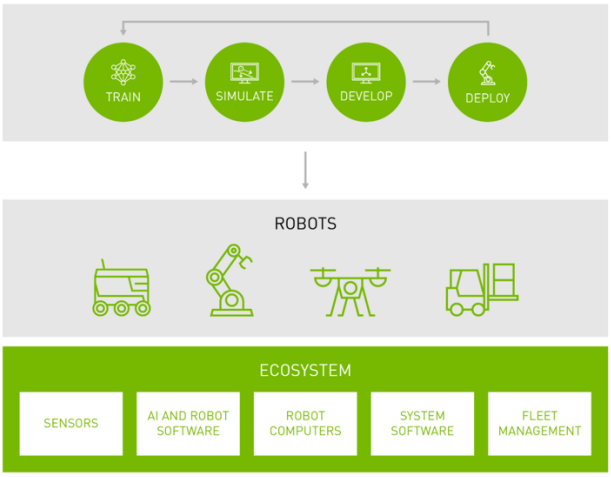


数据来源：英伟达官网

2.3.9. 机器人-Isaac

从开发、仿真到部署，Isaac 平台加速并优化机器人开发。工业和商用机器人的开发过程相当复杂，在许多场景中，缺乏结构化的环境为开发提供支持。Isaac 机器人开发平台为解决这些挑战，打造了端到端解决方案可帮助降低成本、简化开发流程并加速产品上市。其中，本地和云端提供的 Isaac Sim 能够创建精准的逼真环境，为机器人产品提供仿真测试环境; EGX Fleet Command 和 Isaac for AMR（包括 Metropolis、CuOpt 和 DeepMap）能够管理机器人编队以进行部署。

图 49 Isaac 平台

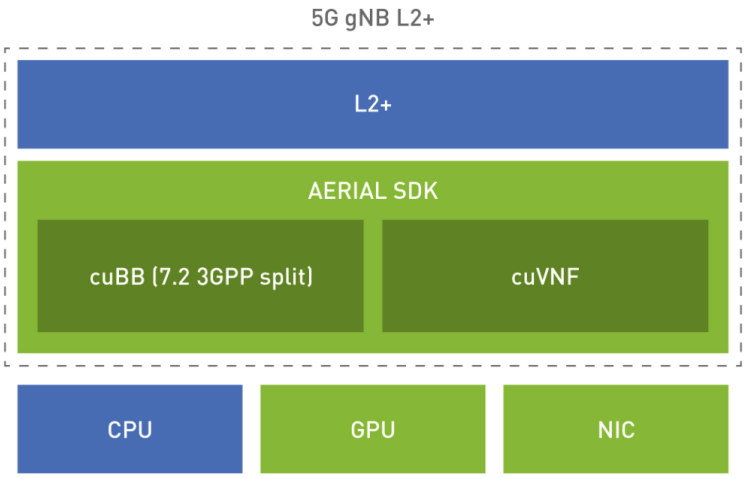


数据来源：英伟达官网

2.3.10. 电信-Aerial

Aerial 是用于构建高性能、软件定义、云原生的 5G 应用框架。Aerial 旨在构建和部署 GPU 加速的 5G 虚拟无线接入网。Aerial SDK 是一个可高度编程的物理层，能够支持 L2 及以上的函数，借助 GPU 加速，复杂计算的运行速度超过现有的 L1 处理解决方案。Aerial SDK 支持 CUDA Baseband(cuBB)和 CUDA 虚拟网络函数(cuVNF)，将构建可编程且可扩展的软件定义 5G 无线接入网的过程变得更为简单。

图 50 Aerial SDK 堆栈



数据来源：英伟达官网

2.4. 行业解决方案全覆盖，助推行业生态迭代

2.4.1. 人工智能与机器学习技术

AI Foundations 打造面向企业的生成式 AI，MaaS（模型即服务）帮助企业开发自己的人工智能模型。英伟达 AI Foundations 是专为 AI 打造的行业解决方案。如今，生成式 AI 正在扩展到全球的企业中，黄仁勋指出，AI Enterprise 将如 Red Hat 之于 Linux 一般，为英伟达的所有库提供维护和管理服务，未来它还被整合至全球范围的机器学习操作渠道内。整体而言，英伟达正在通过一系列云服务套件、预训练的基础模型、尖

端框架、优化推理引擎，和 API 一同为生成式 AI 提供支持。AI Foundations 通过搭载在 DGX Cloud - AI 超级计算机上的 NeMo、Picasso 和 BioNeMo 云服务发挥潜能，可以提供文本生成、图像生成、聊天机器人、总结和翻译等生成式 AI 开发服务。

图 51 AI Foundations 利用生成式 AI 解锁机遇

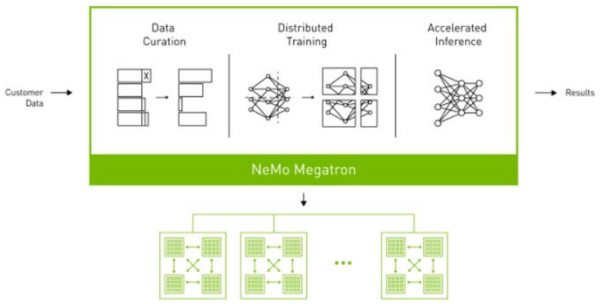
数据来源：英伟达官网

提供 NeMo LLM 服务，致力大型语言模型的开发与维护。英伟达 NeMo LLM 服务令用户可以自定义和使用在多个框架上训练的 LLM，并可在云上使用 NeMo LLM 服务部署企业级 AI 应用。NeMo LLM 降低了大模型开发与维护的难度，实现了文本生成、摘要、图像生成、聊天机器人、编码和翻译等功能。同时 NeMo LLM 将 Megatron 530B 模型作为一款云 API 公开，作为一种端到端框架，Megatron 530B 可用于部署最高数万亿参数的 LLM。

图 52 NeMo LLM 服务具备优势

数据来源：英伟达官网

图 53 利用 NeMo Megatron 进行大模型训练

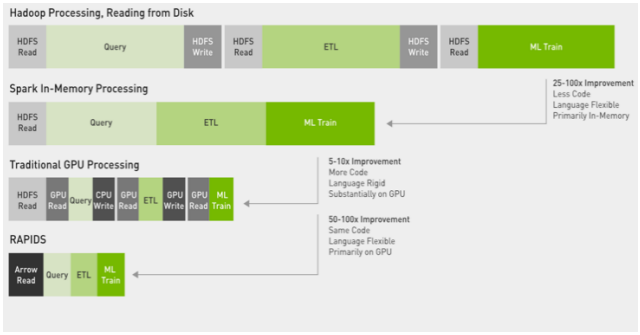


数据来源：英伟达官网

加速机器学习训练时间，打造高性能的数据科学解决方案。除上述的 Maas 外，英伟达也为 AI 提供训练和推理的计算机平台。从机器学习角度，英伟达借助高速 GPU 计算运行整个数据科学工作流程。APIDS 应用框架的使用令原本需要花费几天的流程现在只需几分钟即可完成，因此用户可以更加轻松、快速地构建和部署价值生成模型。基于英伟达的解决方案，仅使用约 16 台 DGX A100 即可达到 350 台基于 CPU 的服务器的性能。减少机器学习中的由于算力限制而被迫产生的缩减取样、

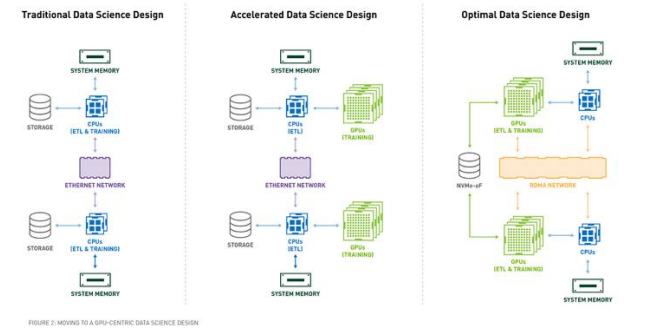
限制模型迭代次数等对企业实际业务决策产生的负面影响，加速模型投入生产的周期。

图 54 数据处理方式的演进



数据来源：英伟达官网

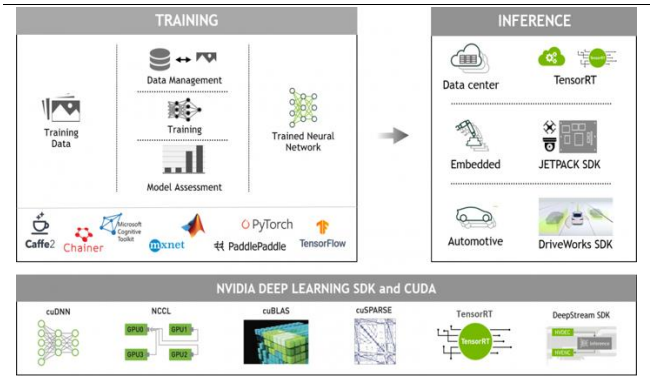
图 55 GPU 改变机器学习训练方式



数据来源：英伟达官网

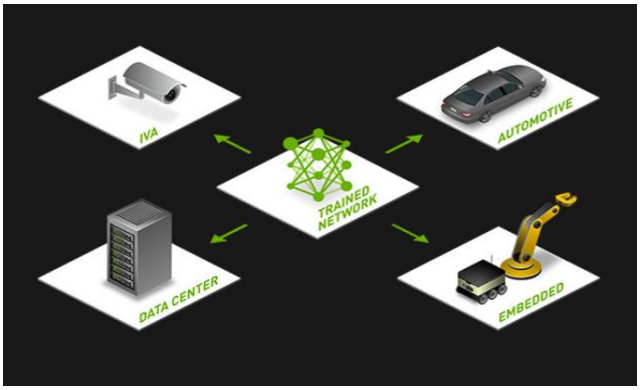
打造完整深度学习训练和深度学习推理平台，持续扩大深度学习领导地位。深度学习领域，从训练平台角度，用户可选择本地工作站、数据中心、云端作为训练平台，借助 SDK 中的软件和框架库进行深度学习训练，也可从英伟达 GPU Cloud 免费访问所有所需的深度学习训练软件。从推理平台角度，用户可借助 TensorRT 平台以及 Triton 推理服务器进行模型推理和部署，Triton 服务器允许团队通过 TensorFlow、PyTorch、TensorRT Plan、Caffe、MXNet 或其他自定义框架，在任何基于 GPU 或 CPU 的基础设施上，从本地存储、Google 云端平台或 AWS S3 部署经训练的模型。

图 56 英伟达深度学习框架



数据来源：英伟达官网

图 57 经 TensorRT 训练优化的模型具备拓展性



数据来源：英伟达官网

AI Enterprise 提供 AI 工作流解决方案。AI Enterprise 是英伟达打造的端到端的云原生 AI 软件套件，它可以加速数据科学流程，简化预测性 AI 模型的开发和部署。AI Enterprise 将 AI 框架、预训练模型和各种资源（例如 Helm 图表、Jupyter Notebook 和文档）封装组合，可缩短开发时间、降低成本、提高准确性和性能。

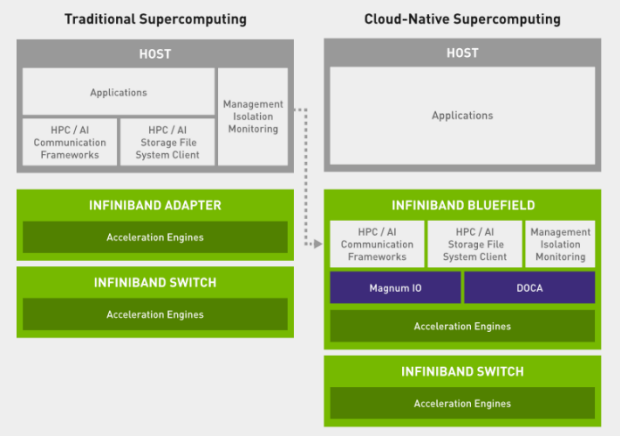
图 58 AI Enterprise

数据来源：英伟达官网

2.4.2. 数据中心与云计算解决方案

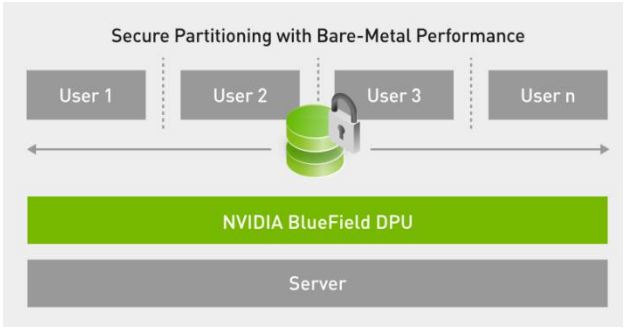
云计算解决方案优势充分释放，为全球创新者提供巨大算力。英伟达的云合作伙伴包括阿里云、谷歌云、腾讯云、AWS、IBM Cloud 和 Microsoft Azure 等，用户可以通过云合作伙伴使用英伟达服务。此外，英伟达基于 BlueField DPU 架构和 Quantum InfiniBand 网络搭建了云原生超级计算平台。DPU 能够为主机处理器卸载和管理数据中心基础设施，实现超级计算机的安全与编排；并且云原生超级计算机实现在多租户环境中的零信任架构，最大程度保障了安全性。同时，英伟达也具备强大的边缘计算服务，形成“云计算+边缘计算”的服务体系。

图 59 云原生超级计算平台



数据来源：英伟达官网

图 60 平台具备零信任架构



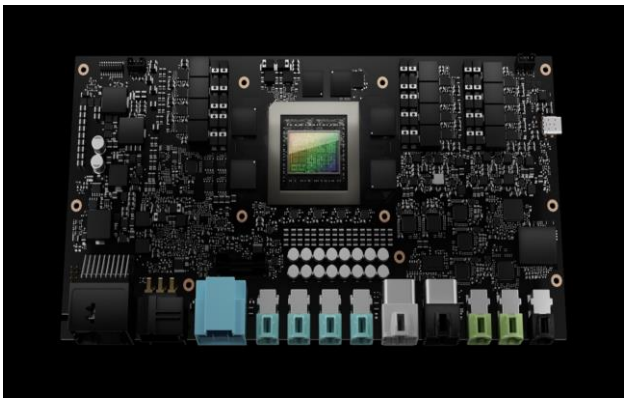
数据来源：英伟达官网

cuLitho 计算光刻技术软件库引入加速计算，加速半导体行业芯片设计和制造速度。英伟达 cuLitho 的推出以及与半导体行业领导者 TSMC、ASML 和 Synopsys 的合作，使晶圆厂能够提高产量、减少碳足迹并为 2 纳米及更高工艺奠定基础。cuLitho 在 GPU 上运行，其性能比当前光刻技术工艺提高了 40 倍，能够为目前每年消耗数百亿 CPU 小时的大规模计算工作负载提供加速，仅需 500 个 DGX H100 系统即可完成原本需要 4 万个 CPU 系统才能完成的工作。在短期内，使用 cuLitho 的晶圆厂每天的光掩模（芯片设计模板）产量可增加 3-5 倍，而耗电量可以比当前配置降低 9 倍。

2.4.3. 汽车行业解决方案

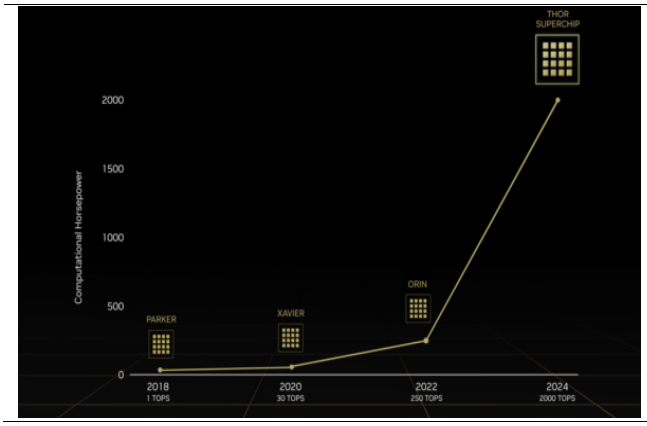
英伟达自研 NVIDIA DRIVE，形成适合自动驾驶汽车的硬件+软件+架构有机统一。硬件端，DRIVE Hyperion 是用于量产自动驾驶汽车的平台，具备用于自动驾驶的完整软件栈，以及驾驶员监控和可视化功能。DRIVE Hyperion 搭载 DRIVE Orin SoC（系统级芯片），可提供每秒 254 万亿次运算的算力负荷。同时，英伟达 2022 年 9 月借助最新 GPU 和 CPU 打造了新一代 SoC 芯片 DRIVE Thor，其可提供 2000 万亿次浮点运算性能，计划 2025 年 DRIVE Thor 能够得到量产。

图 61 新一代 SoC 芯片 DRIVE Thor



数据来源：英伟达官网

图 62 DRIVE Thor 较 DRIVE Orin 有较大算力提升



数据来源：英伟达官网

DRIVE SDK 令开发者高效部署自动驾驶应用程序成为可能，造就未来出行体验。DRIVE SDK 为开发者提供适应自动驾驶的构建块和算法堆栈，开发者可以构建和部署包括感知、定位、驾驶员控制和自然语言处理的一系列应用程序。

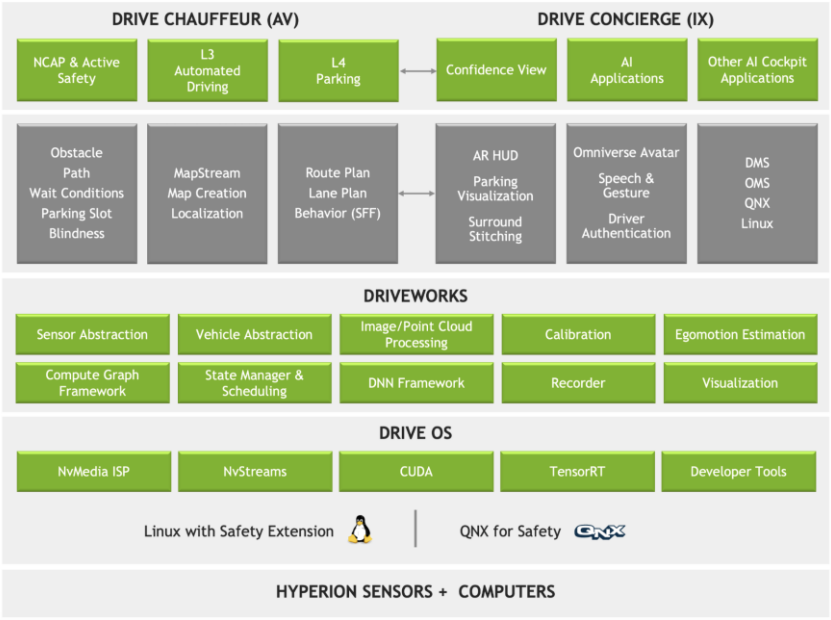
表 3 英伟达自动驾驶汽车主要软件

名称	主要功能和结构
DRIVE OS	针对车载加速计算率先推出的安全操作系统，包括用于传感器输入处理的 NvMedia、用于实现高效并行计算的 NVIDIA CUDA 库、用于实时 AI 推理的 NVIDIA TensorRT，以及可访问硬件引擎的其他开发者工具和模块。
DriveWorks	在 DRIVE OS 之上提供对自动驾驶汽车开发至关重要的中间件功能，包括传感器抽象层 (SAL) 与传感器插件、数据记录器、车辆 I/O 支持和深度神经网络 (DNN)框架。
DRIVE AV	可以用于自动驾驶和地图构建，DRIVE AV 软件栈包含感知、地图构建和规划层，以及各种经过高质量真实驾驶数据训练的深度神经网络 (DNN)。
DRIVE Chauffeur	基于 NVIDIA DRIVE AV SDK 的 AI 辅助驾驶平台，使用高性能计算参考架构和 NVIDIA DRIVE Hyperion 8 传感器集，可以在高速公路和城市之间自由穿梭，并确保极高安全性。
DRIVE IX	是一个开放软件平台，可为 AI 驾驶舱创新解决方案提供舱内感知，提供用于访问各项功能的感知应用程序，还可提供 DNN 以实现高级驾驶员和乘客监控功能、AR/VR 可视化以及车辆与乘客之间的自然语言交互。
DRIVE Concierge	基于 NVIDIA DRIVE IX 和 NVIDIA Omniverse Avatar 构建，可实现实时对话式 AI，助力车辆驾乘人员时刻享受新的智能服务，能够充当每个人的数字助手，帮助他们提出建议、进行预定、拨打电

	话、控制车辆，并使用自然语言发出提醒。
DRIVE Map	使用准确的真值地图和可扩展的车队来源地图来创建和更新自动驾驶汽车地图，利用数百万消费者车队的集体记忆，以及来自 DRIVE Hyperion 车辆的丰富传感器数据。

数据来源：英伟达官网，国泰君安证券研究

图 63 英伟达自动驾驶软件生态框架



数据来源：英伟达官网

DRIVE 基础架构包括开发自动驾驶技术全流程所需的数据中心硬件、软件和工作流。英伟达提供高效节能的 AI 计算加速训练，有助于 AI 收集大量真实行驶数据作为训练集；在 DRIVE Sim 中，可以通过模拟驾驶在虚拟世界中进行测试，得到各种罕见和危险驾驶情形下的驾驶数据。目前，英伟达开发的 AI 赋能自动驾驶汽车已经应用至各大主流汽车制造商，成为自动驾驶汽车开发的首要工具。

图 64 英伟达基于驾驶数据生成 3D 模拟环境

数据来源：英伟达官网

图 65 英伟达助力自动驾驶汽车革新

数据来源：英伟达官网

2.4.4. VR 与游戏产业产品

英伟达 GPU 为 VR 头盔和 GeForce Game Ready 驱动提供即插即

用的兼容性。VR 成像是否连贯将极大影响头显的使用体验，舒适的 VR 体验要求显示器有效分辨率至少为 4K 且最低刷新率为 90Hz，这就需要 GPU 为其提供支持。GeForce RTX GPU 兼容目前市场上主流 VR 头盔，通用性较强。从性能上看，GeForce RTX GPU 依托其 DLSS、光线追踪和 PhysX 三大成像技术为用户模拟如真实世界般的 VR 体验。

图 66 GeForce RTX GPU 兼容主流 VR 头显

数据来源：英伟达官网

图 67 PhysX 基于物理模拟使 VR 生动逼真

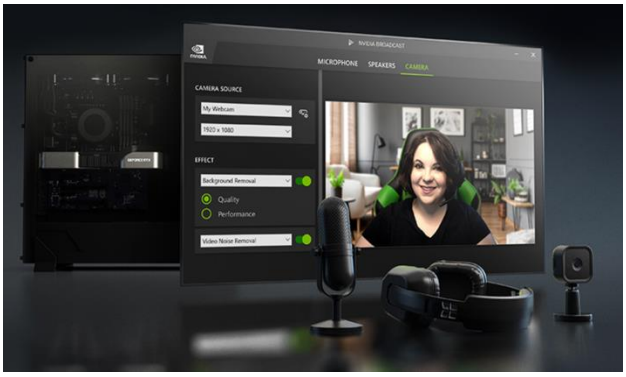
数据来源：英伟达官网

全方位覆盖游戏娱乐体验，打造专业游戏环境。目前有超 2 亿游戏玩家和创作者使用 GeForce GPU，针对这一客户群体，英伟达打造了一系列专业游戏服务: GeForce Experience 可以截取并与好友分享截图、视频和直播；Game Ready 驱动程序可实现一键优化游戏设置；Broadcast App 提供专业化直播服务，如只需点击一个按钮即可消除噪音或添加虚拟背景; Omniverse Machinima 可以实现对虚拟世界中的角色及其环境进行操作处理并实现动画化。

图 68 GeForce Experience 在守望先锋中优化设置

数据来源：英伟达官网

图 69 Broadcast App 让居家环境秒变 AI 工作室



数据来源：英伟达官网

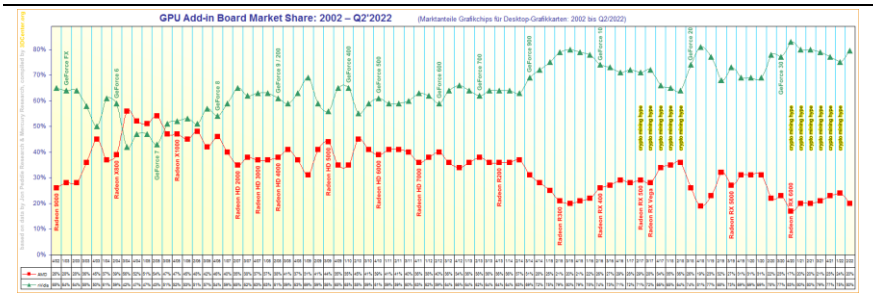
3. 重新定义市场，助推 AI 发展

3.1. 长期稳居显卡市场龙头，市场份额保持高位

英伟达独显市场份额长期稳居高位，与 AMD 呈此消彼长关系。据 3DCenter，2022Q2 全球独立显卡共计出货约 1040 万张，总销售额约 55 亿美元，与 2021 年存在较大差距，其中显卡平均售价从 2021Q2 的 1029

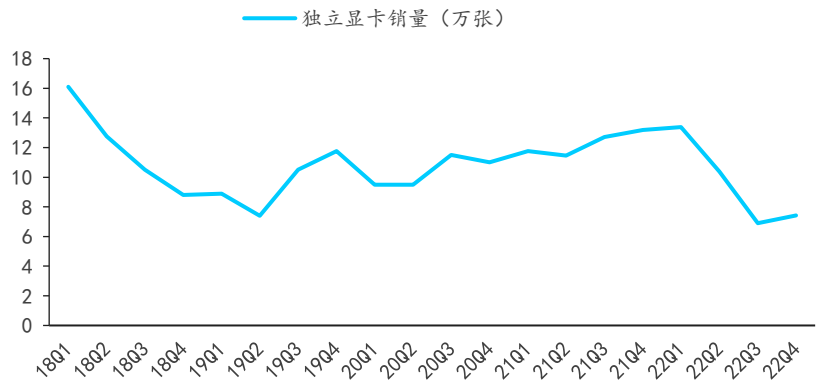
美元大幅跌落至 2022Q2 的 529 美元。据 JPR 测算，22Q2 英伟达出货占全球独立显卡市场份额 79%，同比增长 4pct，环比降低 1pct。此外，AMD（超威半导体）囊括了 20%的市场份额，作为新入局者英特尔（Intel），其市场份额仅 1%，可见英伟达在独立显卡领域长期耕耘的市场优势显著，尤其是高端显卡市场。而后，22Q3 全球独立显卡销量同降 33.7%至 690 万张，22Q4 同增 7.8%至 743 万张。

图 70 3DCenter 测算的独立显卡市场份额



数据来源：3DCenter

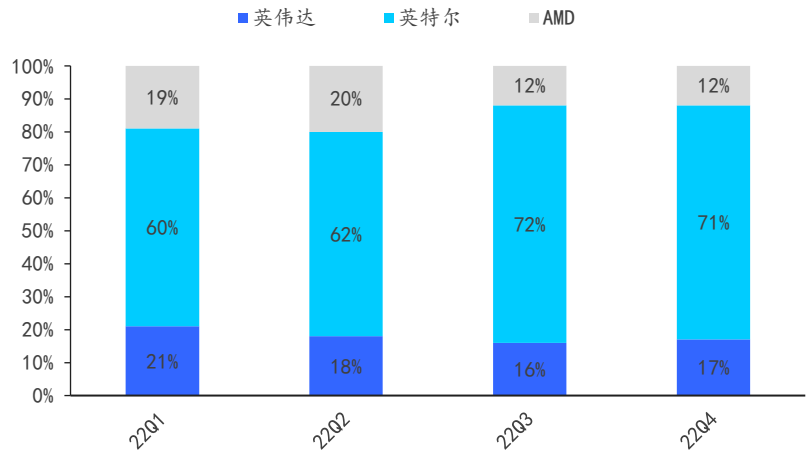
图 71 全球独立显卡销量



数据来源：JPR，国泰君安证券研究

2022 年全球 GPU 市场低迷，英特尔保持全球最大 PC 端 GPU 供应商地位。据 JPR, 22Q4 全球共出货 6420 万块独立 GPU 和集成 GPU，同比-38%，环比-15.4%，整体降幅明显，彰显市场需求低迷情绪，尤其是集成显卡制造商采购意愿下滑严重。从市场份额角度，以 22Q4 为例，英特尔 PC 端 GPU 销售额占 71%，英伟达和 AMD 分别占 17%和 12%。整体来看，集成显卡市场库存过剩和需求减弱的供需矛盾仍暂未缓解，出货量或将继续维持低位。

图 72 JPR 测算的 PC 端 GPU 市场份额



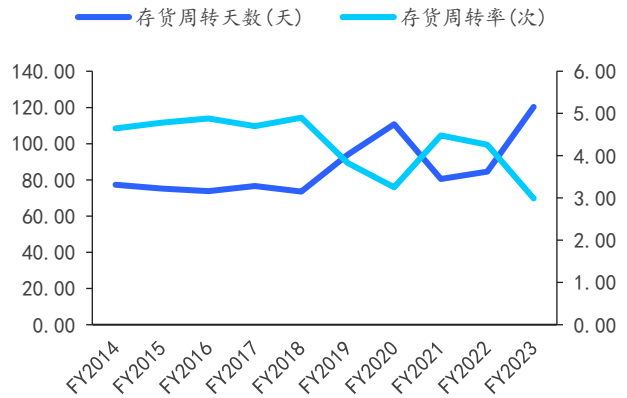
数据来源：JPR，国泰君安证券研究

3.2. 合作伙伴网络庞大，AI 市场持续开拓

英伟达主要客户群体覆盖顶尖科技公司，未来将持续向人工智能市场开拓。英伟达处半导体产业链上游研发设计环节，半导体细分领域几大头部厂商垄断力较强，其主要客户包括华硕、联想、惠普、Facebook、IBM、慧与、三星等。下游需求严重影响英伟达的存货与生产计划，从存货角度分析，FY2020 存货周转天数上涨主要由原材料价格上涨提前追加采购所致，FY2023 存货周转天数再度高涨则由于需求疲软造成的库存积压。但随着 AI 算力需求提高重振英伟达销售预期，我们认为英伟达存货周转有望重返合理区间，同时其 AI 研发的持续投入也将有望吸引更多 AI 公司使用英伟达芯片产品。

英伟达基于庞大合作伙伴网络，共同推动视觉计算未来。英伟达作为行业领导者，率先推出了视觉计算解决方案，并在近 30 年来通过合作伙伴网络（NPN）将产品投入市场。合作伙伴包括增值经销商、解决方案集成、设计或制造系统、托管服务、咨询以及为英伟达产品和解决方案提供维护服务的公司。同时，英伟达积极通过 GTC 大会吸引更多的全球合作伙伴，2023 年 GTC 大会钻石合作商就包括微软、谷歌云、阿里云、戴尔科技等国内外大厂，黄仁勋指出，目前全球英伟达生态已有 400 万名开发者、4 万家公司和英伟达初创加速计划中的 1.4 万家初创企业。

图 73 英伟达存货周转天数与存货周转率



数据来源：英伟达财报，国泰君安证券研究

图 74 英伟达 GTC 大会钻石合作商

数据来源：英伟达官网

表 5 英伟达国内主要合作公司整理

公司名称	合作主要内容
联想集团	首家采用英伟达新的自动驾驶域控制器的一级制造商，基于新一代 NVIDIA（英伟达） DRIVE Thor 系统级芯片（SoC），自主研发最新一代车载域控制器平台。
胜宏科技	英伟达是胜宏科技前五大客户之一，供应英伟达 A100+H100 的板卡。作为英伟达算力板国内第一供应商，占英伟达显卡的全球市场份额 50%，在 AI 服务器方向，胜宏科技有近 10 个算力板型号正在通过英伟达的认证
沪电股份	英伟达服务器印制电路板（PCB）的供应商，1 个算力板型号正在通过英伟达的认证
工业富联	英伟达新的 GPU HPC 平台供应商
天孚通信	与英伟达深度合作研发光引擎技术，光引擎二期已开工
龙迅股份	与英伟达深度合作，并生产 GPU 显卡 pcle 接口
长电科技	负责提供英伟达 GPU 芯片的封装测试业务
和林微纳	GPU 芯片探针核心供应商
千方科技	子公司宇视科技与英伟达联合开发服务器
鸿博股份	全资子公司英博数科为英伟达提供算力出租
铂科新材	英伟达芯片电感软磁粉芯独家供应商
先进数通	英伟达 Elite 级合作伙伴
中际旭创	作为英伟达重要供应商，受益于 AI-GC 持续催化和头部企业对于 800G 光模块的需求，800G 光模块先发优势显著
精研科技	英伟达芯片散热供应商
顺网科技	英伟达大陆早期合作伙伴，在国内智能算力场景存在合作机遇
中电港	英伟达芯片国内授权分销商，同时与英伟达有全面合作关系
华工科技	全球最大的光模块生产厂商之一，跟英伟达、微软、AR-I-S-TA、Me-ta、In-t-el 等进行了良好对接
博杰股份	工业自动化设备供应商，与英伟达有板卡检测设备合作订单
浪潮信息	目前国内唯一一家使用英伟达 GPU 服务器的厂商，其 AI 服务器中国大陆市占率高达 52.4%，居第一位
海康威视	与英伟达合作开发智能视频分析系统
步长制药	与英伟达合作推进 AI 人工智能加速药物研发
鼎阳科技	独家提供英伟达电子测试测量仪器，应用于其半导体、AI 人工智能领域
万润科技	子公司长江存储与英伟达合作推出 AI 芯片
启明星辰	与英伟达合作开发 AI 自动驾驶技术
景旺电子	英伟达合格供应商，目前已实现批量供货，针对 AI 服务器的产品正参与客户开发和验证流程
中兴通讯	与英伟达合作开发 5G 网络、云游戏

奥拓电子	作为行业内最早推出 XR 虚拟拍摄解决方案的公司之一服务英伟达
华大基因	与英伟达合作推进 AI 基因药物研发
海能达	与英伟达合作推进智慧城市建设
山子股份	与英伟达合作开发人工智能技术
京东方 A	与英伟达合作开发 AI 自动驾驶系统
世运电路	通过客户与英伟达产业供应链进行合作，发展 AI 服务器、数据中心等相关业务
中富通	英伟达是控股子公司英博达的上游合作方之一，英博达系一家专业从事边缘计算、智能化终端的研发设计以及 AI 智能视觉检测技术方案服务商
恒信东方	发布《人工智能算力中心平台建设及运营项目可行性研究报告》，变更募集资金用途，开工建设采用英伟达 AI 训练、AI 推理核心服务器的大数据中心
中科创达	提供包括多模态模型等人工智能技术，中国首家获得英伟达画质调优授权的公司
同方股份	公司具备英伟达和昇腾两条技术路线的人工智能服务器产品线，具备大批量出货能力，用于建设内蒙古超算中心、宁波智算中心、北京昇腾智算中心等项目。此外，公司和部分人工智能头部厂商保持密切技术合作，其中部分产品已经得到应用。
移远通信	2022 年 5 月，公司 5G 模组与英伟达 Je-t-s-on AGX Orin 超级人工智能平台成功完成联调，实现 5G 通信+AI 边缘计算能力。

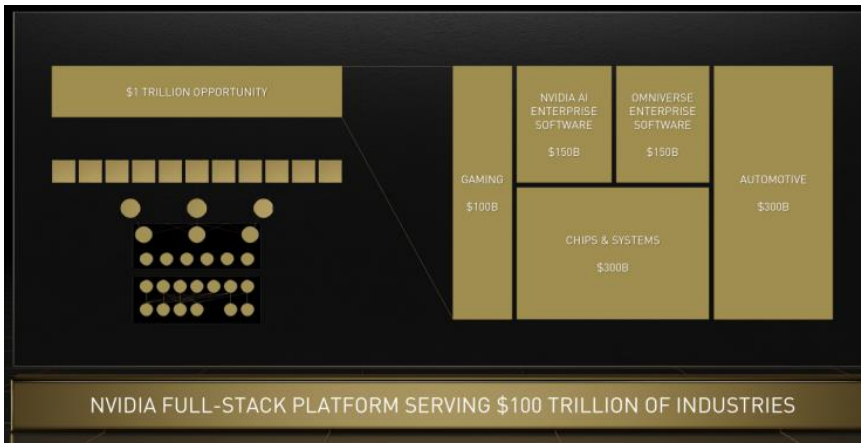
数据来源：雪球网，英伟达官网，国泰君安证券研究

3.3. AI 市场持续高增，周期布局价值彰显

AI 芯片市场成为新的增长极，周期布局价值渐显。云计算、人工智能、工业 5G 和加速计算等业务增长将成为解决计算时代症结的最后几块拼图。硬件+软件的完整生态系统将有助英伟达在 AI 的极速发展中稳定其头部供应商地位。据 IDTechEx 发布的报告《人工智能芯片 2023-2033》预测，到 2033 年，全球 AI 芯片市场将增长至 2576 亿美元。JPR 也曾预测，2022-2026 年全球 GPU 销量复合增速将保持在 6.3%水平，2027 年全球 GPU 市场规模有望超 320 亿美元。目前 Open AI 模型主要由英伟达 GPU 进行训练，我们看好 AI 芯片市场激增对英伟达投资价值的催化作用。

英伟达预测自身总潜在市场为万亿美元量级，对各业务线持整体乐观预期。在 2022 年 3 月投资者的活动中，英伟达指出其业务领域的总潜在市场 (TAM) 为 1 万亿美元，其中游戏业务约 1000 亿美元，人工智能企业软件 1500 亿美元，Omniverse 业务 1500 亿美元，硬件与系统 3000 亿美元，以及自动驾驶业务市场 3000 亿美元。即便英伟达并未清晰给出其计划实现这一目标的具体时间，但仍从一定程度上反映了英伟达对其各业务条线市场份额权重的合理预期。

图 75 英伟达测算的其各业务板块远期市场份额

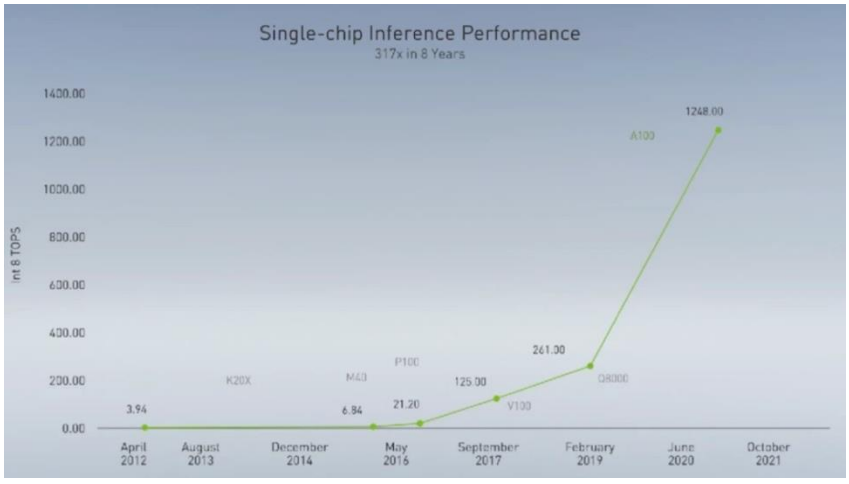


数据来源：英伟达官网

3.4. 重塑摩尔定律，AI iPhone 时刻提供新机遇

摩尔定律逐渐失效，“黄氏定律”重塑行业生态正当时。摩尔定律指在价格不变的前提下，集成电路上可容纳的晶体管的数目，约每隔约 18 个月便会增加一倍，半世纪以来，摩尔定律指引着芯片市场迈向繁荣。但随着传统半导体晶体管结构已进入纳米级别，摩尔定律也逐渐在高成本的驱动下逐渐失效。但如今，大模型对于算力激增的需求已远大于摩尔定律所预估。黄仁勋对 AI 性能的提升作出预测，指出 GPU 将推动 AI 性能实现每 1 年翻 1 倍，也就是每 10 年 GPU 性能将增长超 1000 倍。这一论断也被称之为“黄氏定律”。英伟达首席科学家兼研究院副总裁 Bill Dally 表示，目前单芯片推理性能的提升主要原因在于 Tensor Core 的改进、更优化的电路设计和架构，而非制程技术的进步。因此，在摩尔定律消失之后，黄氏定律将不断催生计算性能的进步。

图 76 黄氏定律表现在 GPU 助推 AI 推理性能每年提升一倍以上

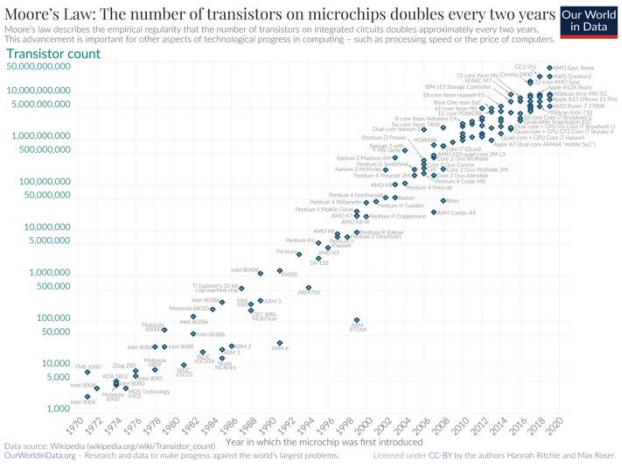


数据来源：英伟达官网

ChatGPT 成为 AI 的 iPhone 时刻。无论是率先发明 GPU 并保持约两年一次架构更新速度，亦或是成为首个打造硬件+软件生态的公司，英伟达都为行业生态系统创造了新的发展机遇。而当下以 ChatGPT 为代表的人工智能对社会的影响正如当年 Apple 通过 iPhone 打开全球智能手机市场一般。而英伟达的远见即在于提前布局 AI 业务，早在 2016 年，英

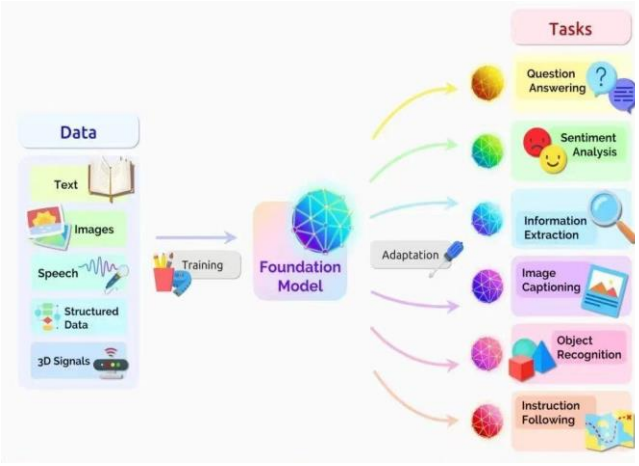
伟达就向 OpenAI 交付了英伟达 DGX AI 超级计算机，成为支持 ChatGPT 的大语言模型突破的引擎，可以说 DGX 超级计算器是现代“AI 工厂”。

图 77 半世纪来摩尔定律指引芯片市场增长



数据来源：维基百科

图 78 基础模型的用途日益广泛



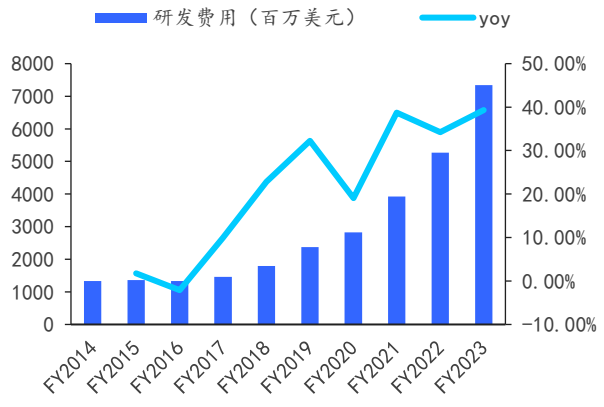
数据来源：英伟达官网

4. 研发创新贯穿公司历史，迭代公司增长曲线

4.1. 研发投入持续高增，研发团队规模日益壮大

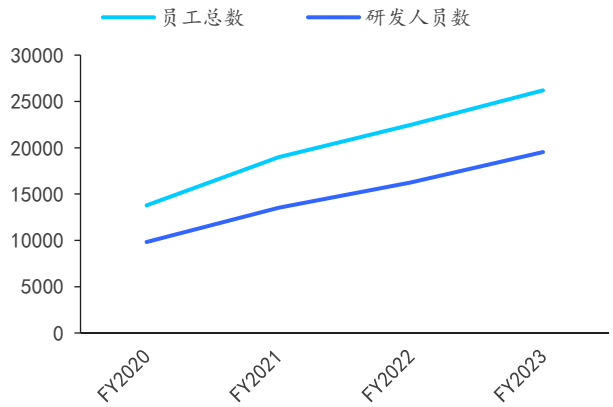
英伟达持续加大研发投入，注重创新能力培育。FY2023 年英伟达研发费用达 73.39 亿美元，同增 39.31%，近年来英伟达研发费用增速明显，在 FY2021-FY2023 已连续三年呈现超 30%的同比增长率。据 FourWeekMBA 统计，截至 2023 年 1 月，英伟达全球员工总数共 26196 人，其中研发人员 19532 人，研发人员占比约 75%。四年间英伟达研发人员数量近乎翻倍，研发人员的高占比反应了公司对于研发创新这一企业生命线的重视。

图 79 英伟达研发投入持续高增



数据来源：iFinD，国泰君安证券研究

图 80 英伟达员工总数与研发人员数



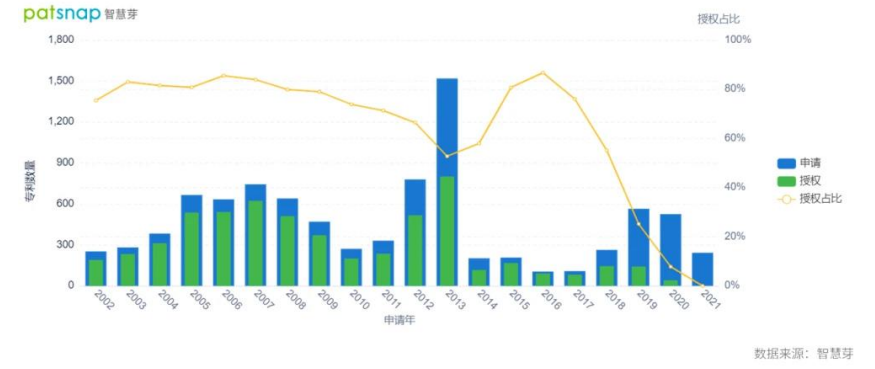
数据来源：FourWeekMBA，国泰君安证券研究

4.2. AI 拐点时刻，大型语言模型形成新技术重心

专利申请数处行业前列，神经网络领域成为研究和专利申请重心。据智慧芽数据，截止 2021 年，英伟达及其关联公司共计申请超 9700 件

专利，集中在 GPU 相关硬件领域。其中 2013 年达到专利申请与授权最高值。自 2014 年起专利申请与授权较前值显著降低，授权占比亦呈现下滑趋势。出现这种转变的原因主要在于研发重心转移带来的产出成果更迭。对比 1993-2013 年和 2014-2021 年专利关键词云，“处理器“、”存储器“、“计算机程序单元”的比重相对降低，取而代之的首位关键词为“神经网络”，反映了神经网络相关技术成为英伟达研发的首要方向。

图 81 2002-2021 年英伟达专利申请与授权数



数据来源：智慧芽

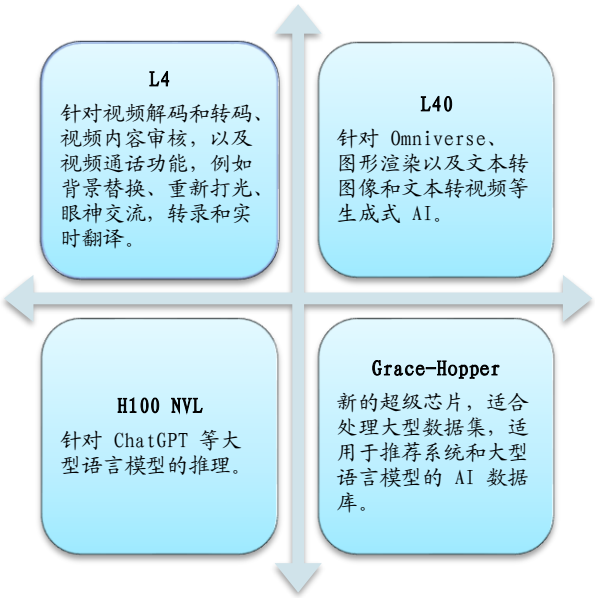
注：受专利授权时间滞后影响，2021 年授权数可能存在缺失

图 82 英伟达专利关键词云迭代明显

数据来源：智慧芽

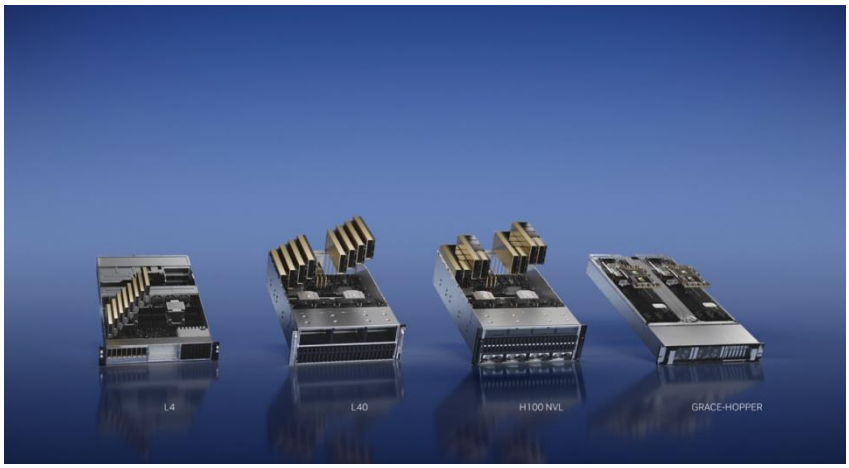
大型语言模型业务成为未来技术发展重心，发布四大新计算技术平台。在 GTC 2023 上，英伟达加快生成式 AI 应用的部署，推出四个计算技术平台，分别是用于 AI 视频的英伟达 L4，针对 Omniverse、图形渲染以及文本转图像和文本转视频等生成式 AI 的英伟达 L40，用于大型语言模型推理的 H100 NVL 以及适用于推荐系统和大型语言模型数据库的 Grace Hopper。黄仁勋表示：“AI 正处于一个拐点，为每个行业的广泛采用做准备。从初创企业到大型企业，我们看到人们对生成式 AI 的多功能性和能力越来越感兴趣。”而大型语言模型业务也将因此成为英伟达技术发展的重心。

图 83 四个计算技术平台的主要应用



数据来源：英伟达官网，国泰君安证券研究

图 84 英伟达发布四个新计算技术平台



数据来源：英伟达官网

4.3. 区位优势突出，持续强化产学研深度合作

英伟达充分利用硅谷的区位优势，与学术界保持着长期的合作关系，提供不竭的创新动力。英伟达除了与专业的研究团队开展合作外，也将顶尖高校的优秀毕业生作为重点人才储备，持续强化产学研深度合作。主要合作学术研究项目包括与加州大学伯克利分校的 ASPIRE 项目、与北卡罗来纳州立大学等多所高校联合的 CAEML 项目和 CV2R 项目、以及与斯坦福工程学院的 SCIEN 项目等，涵盖机器学习、虚拟现实等领域，覆盖软硬件市场。

图 85 硅谷附近具备多所顶尖大学

数据来源：搜狐网

5. 打造多元文化，勇担社会责任

5.1. 坚持可持续发展，践行 ESG 目标

英伟达注重可再生能源与生产效率，助力践行 ESG 目标。英伟达在每年度均计划购买或生产大量的可再生能源，以全面满足全球对电力的使用需求。此外，英伟达的 GPU 通过算力提升降低了能源消耗，其生产的 GPU 对于某些 AI 和 HPC 工作负载，其能效通常比 CPU 高 20 倍。2022 年 5 月，英伟达推出液冷 GPU，据 Equinix 和英伟达单独测试，采用液冷技术的数据中心工作负载可与风冷设施持平，同时消耗的能源减少约 30%。值得一提的是，Green500 排行是衡量超级计算机的能效的重要指标，在 2022 年 6 月的 Green500 榜单里排名前 30 的超级计算机中，有 23 台由英伟达的 GPU 提供支持。

图 86 英伟达推出液冷 GPU

数据来源：英伟达官网

5.2. 承担社会责任，投身公益活动

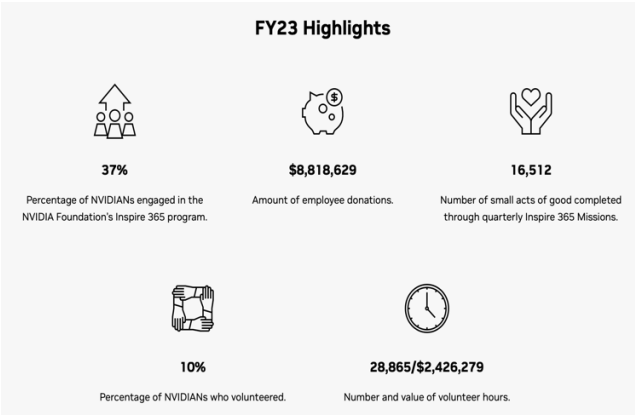
员工致力于构建推动人类进步的技术，并为其工作和生活的社区提供支持。英伟达表示，作为积极承担社会责任的优秀公司，他们的员工

古道热肠，向全球数百家慈善组织提供捐助。同时英伟达建立了专项基金会，37%的员工在 FY2023 参与了基金会 Inspire 365 计划，共计捐赠超 880 万美元，提供了约 29000 小时的志愿服务时间，较 FY2022 同增 74%。加上以公司名义的捐赠，总捐赠额共计 2250 万美元，覆盖了 55 个国家或地区的 5800 多家非营利组织。

图 87 英伟达积极参与公益活动

数据来源：英伟达官网

图 88 2023 财年英伟达基金会公益表现

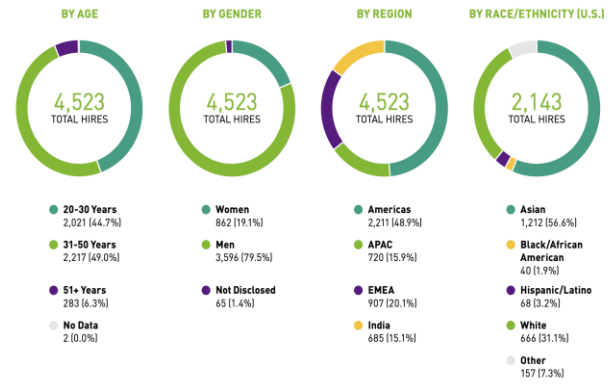


数据来源：英伟达官网

5.3. 强调以人为本，深耕企业文化

英伟达注重打造多元企业文化，提升员工福祉。Glassdoor 的评选结果显示，英伟达的员工将公司评为全美排名第 1 的工作场所。《财富》杂志亦将其评为“最佳雇主 100 强”。并且，英伟达致力于创造更加多元化的文化，构建“残障平等指数”、“企业平等性指数”和“性别平等指数”等指标，彰显企业以员工为本的理念，提供包容性的工作场所，并始终坚持履行其对同工同酬的承诺。

图 89 英伟达雇员组成结构



数据来源：英伟达官网

图 90 GTC 大会邀请科技领域杰出女性演讲

数据来源：英伟达官网

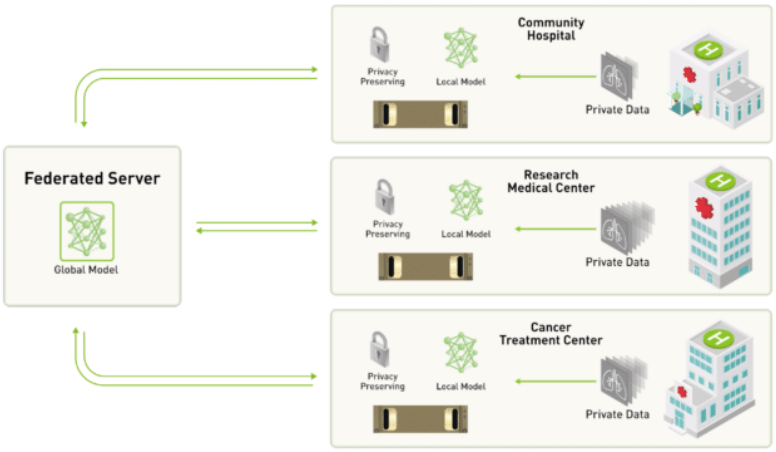
5.4. 关注客户隐私，持续提升产品安全

注重 AI 时代下数据安全问题，建立专业风险响应团队。英伟达打造了全球产品安全事件响应团队（PSIRT），通过及时的信息传递处理产品和服务相关的安全漏洞，并将 NIST 网络安全框架的元素和控件集成到其安全程序中。同时参与 MITRE 这一全球网络安全组织，扩展 AI 的

MITRE ATT&CK 框架，以更好响应 AI 时代新的威胁。

打造注重隐私保护的联合学习系统，产品安全整体可控。以医疗行业为例，英伟达推出的医学影像分析的联合学习系统（Federated Learning），可以通过构建全局模型避免患者的信息被无条件共享。医院、研究中心和疾控中心能够各自根据其既有数据于本地训练模型，并间隔一定时间将数据提交给全局参数服务器，该服务器可以通过整合各节点模型信息并生成新的模型，最后将模型重新反馈回各节点。该系统在隐私保护基础上最大程度保障了模型性能，合理利用了各方数据信息。

图 91 英伟达联合学习系统



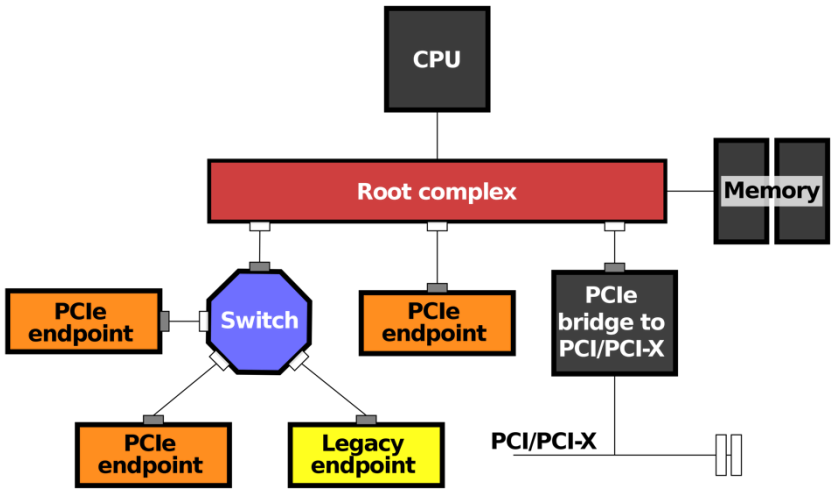
数据来源：英伟达官网

6. 以超异构创新重塑大规模 AI 计算，发动世界 AI 引擎

6.1. CPU 难以支撑 AI 算力需求，市场亟需更强算力

CPU 主要以串行计算，基于 CPU 和 PCIe 的数据中心吞吐量严重不足。串行计算指的是多个程序在同一个处理器上被执行，只有在当前的程序执行结束后，下一个程序才能开始执行，CPU 的运行主要以串行计算的方式进行。同时，据 CSDN，以 PCIe 最新版本 5.0 为例，其传输速率仅有 32 GT/s 或 25GT/s，PCIe 吞吐量的计算方法为：吞吐量=传输速率*编码方案，因此传输速率的不足直接导致了 CPU 基于 PCIe 的吞吐量较小，也就意味着其带宽较小。并且，在此过程中 CPU 产生的功耗和延时均较高，会产生较高的计算成本。因此，基于 CPU 串行计算的特点和较小的带宽，已无法适应如今数据中心的算力要求。

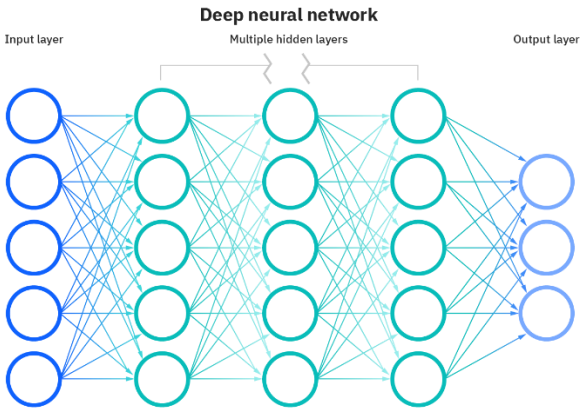
图 92 典型的 PCIe 系统框图



数据来源：维基百科

CPU 无法适应深度学习高并发、并行计算和矩阵处理等算力要求。以神经网络模型为例，其包含输入层、输出层和中间层（亦称隐藏层）。近年来，深度学习应用需求的激增倒逼开发者实现更强的函数模拟能力，这需要通过提升模型的复杂度来实现，这直接导致神经网络中间层数量的大增，最终使得神经网络参数数量的飙升。由于神经网络是高度并行的，使用神经网络做的许多计算都需要分解成更小的计算，尤其是利用卷积神经网络进行图像识别时，卷积和池化等过程需进行大量矩阵运算，而 CPU 内部计算单元有限，在执行此类任务时将极大的消耗模型训练的时间。基于多层神经网络的复杂运算亟需更强算力的现实需求。

图 93 CPU 无法满足训练神经网络模型的要求



数据来源：IBM

6.2. GPU 生逢其时，英伟达异军突起

6.2.1. 技术日新月异，AI 芯片应时代需求而生

GPU 解决算力限制顽疾，高带宽适应模型训练需要。与 CPU 相比，使用 GPU 进行大规模并行计算的优势得到了充分彰显，以 H100 Tensor Core GPU 为例，其支持多达 18 个 NVLink 连接，总吞吐量为 900 GB/s，是 PCIe 5.0 带宽的 7 倍，进而实现超快速的深度学习训练。对于神经网络模型的训练，GPU 逻辑运算单元较多的优势能够得到充分的发挥，能够满足 GPU 无法实现的深度学习高并发、并行计算和矩阵处理的算力

要求，因此 GPU 无疑成为了深度学习的硬件选择。

AI 迭代飞速催生芯片技术创新，DPU、FPGA、ASIC 等 AI 芯片应时代需求而生。AI 时代呼唤新架构的产生，即便 GPU 相较 CPU 存在显著的算力优势，但市场可能需要比 GPU 性能更加优越的专用芯片，目前已并不仅只有 GPU 能适用以深度学习模型训练。近年来 AI 芯片技术爆发式增长，各类 AI 芯片上新迅速，我们参考《科学观察》杂志论文《AI 芯片专利技术研发态势》，将 AI 芯片技术体系划分为如下 11 个分支领域。

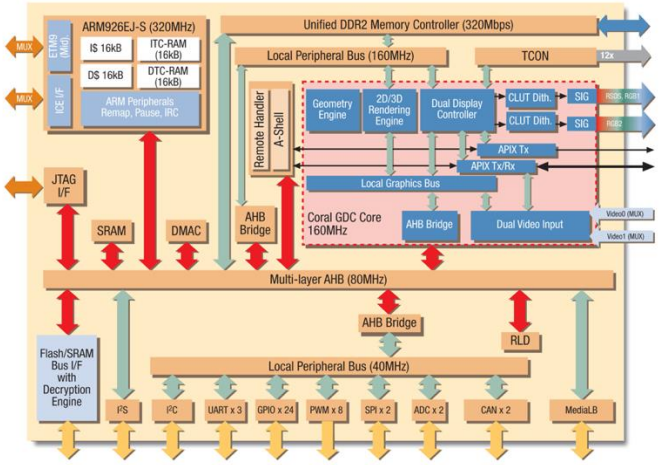
表 6 AI 芯片技术领域

分支领域	相关术语
类脑芯片	类脑芯片、大脑处理器
GPU	图像处理器、图形处理器、显示核心、视觉处理器、显示芯片、图像处理单元、图形处理单元
VPU	视觉处理器、视觉处理单元
NPU	神经网络处理器、神经网络芯片
IPU	智能处理单元
TPU	张量处理器、张量处理单元
FPGA	现场可编程门阵列
ASIC	专用集成电路

数据来源：《科学观察》，国泰君安证券研究

ASIC 适应定制化高需求使用场景，计算能力和效率可根据算法需要进行定制。专用集成电路（ASIC）指根据用户特定的要求和特定电子系统的需要而制造的集成电路，设计完成后集成电路的结构即固定。ASIC 适用于对于芯片高需求且定制化程度较高的应用场景，如先前的矿机芯片和如今火热的自动驾驶芯片。Frost & Sullivan 数据统计，全球 ASIC 市场规模从 2018 年的 299 亿美元增长至 2023 年的 674 亿美元，复合增速达到 17.7%。ASIC 的发展有望一定程度上满足 AI 对算力激增的需求，但短期内难以打破英伟达 GPU 在市场份额的领先优势。

图 94 ASIC 复杂的设计流程

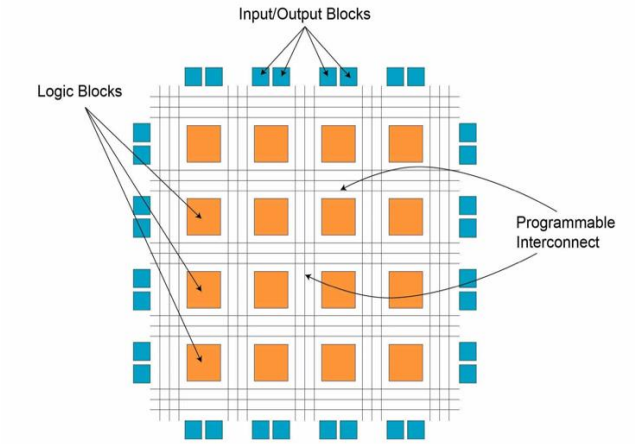


数据来源：CSDN

FPGA 作为 ASIC 中半定制电路，“先购买再设计”，与 AI 相互成

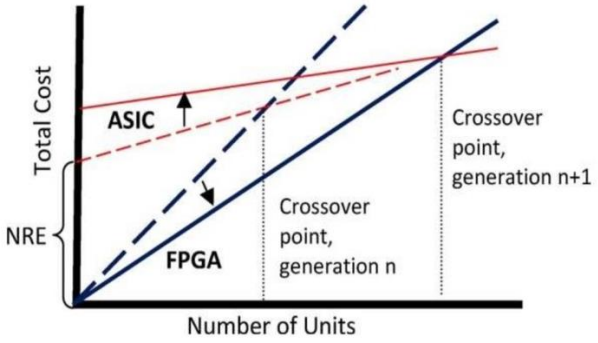
就。现场可编程门阵列（FPGA）指在硅片上预先设计，同时具有可编程特性的集成电路，开发者能够根据产品需求进行设计配置。相较原有的 ASIC 而言，FPGA 具备了后期可编程性，适合需求量相对较小的定制化场景，具备更高的灵活性。FPGA 技术目前具备较高的技术壁垒，但受益于 AI 技术持续扩展，行业需求具备明显确定性，将有望吸引更多竞争者入局，也将会对 GPU 的潜在市场产生冲击。

图 95 FPGA 的基本结构



数据来源：logic-fruit

图 96 FPGA 适合需求较小的定制化场景

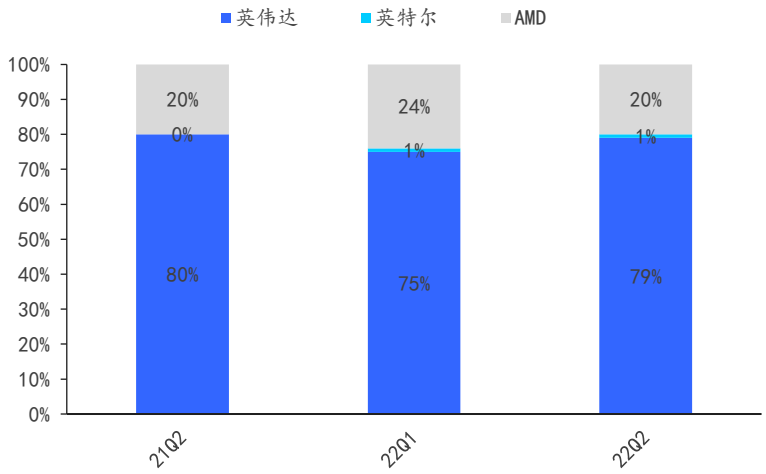


数据来源：ResearchGate

6.2.2. 激战 AMD、英特尔及互联网巨头

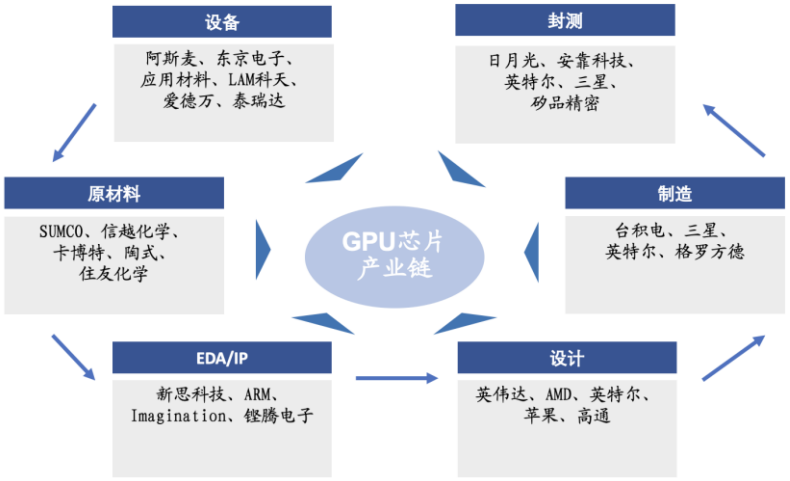
英伟达、英特尔、AMD 为 GPU 领域行业巨头，苹果、高通等破局者不断涌入带来涟漪。据 JPR 测算,英伟达长期占全球独立显卡的市场份额近 80%，其余市场份额几乎均被 AMD 抢占。因此 GPU 芯片市场英伟达和 AMD 共同主导。而英特尔为主要 CPU 制造商，同时也在 PC 端 GPU 具备领先份额。英伟达的主要竞争对手集中在 GPU 产业链的设计环节。但同时，苹果、高通等破局者也在进入 GPU 市场企图实现自研 GPU 以降低对外技术依赖的需求。

图 97 JPR 测算的独立显卡市场份额



数据来源：JPR，国泰君安证券研究

图 98 全球 GPU 芯片产业链各代表环节企业



数据来源：前瞻产业研究院，国泰君安证券研究

AMD 是高性能与自适应计算领域的领先企业，处在半导体行业前沿。AMD 作为英伟达在独立 GPU 领域的主要竞争对手，提供从处理器、显卡、软件和应用等全方位的产品服务，CPU+GPU+DPU+FPGA 的产品线已全面布局。AMD 在汽车、超级计算和高性能计算、网络电信、机器人领域自适应计算等也都提出了自己的全套解决方案。

作为 AMD 最可能对标英伟达 GH200 的产品 MI300 年内将发布。Instinct MI300 具备开创新的适应数据中心设计，共包含 13 个小芯片，其中许多是 3D 堆叠的，以创建一个具有 24 个 Zen 4 CPU 内核并融合了 CDNA 3 GPU 和 128G HBM3 显存的超级芯片，集成了 5nm 和 6nm IP。总体而言，该芯片拥有 1460 亿个晶体管，是 AMD 投入生产的最大芯片。我们认为，MI300 不仅距离实现量产还有较长时间，且其算力相较于英伟达已量产的产品线依旧较低，与英伟达 GPU 研发和生产的整体差距约两年，目前对于英伟达 GH200 产生的竞争压力较小。

表 7 英伟达 GH200 和 AMD MI300 对比

对比项	GH200	MI300
算力	提供 1 exaflop 算力	与先前的 MI250X 相比，MI300 在 47.9 TFLOPS 的基础上预计提升至 8 倍，能耗将降低至原有的 1/5
内存	使用 NVLink 将 GPU 和 CPU 之间的带宽提高了 7 倍，将互连功耗减少了 5 倍以上，同时应用 NVLink Switch、Quantum-2 InfiniBand，DGX GH200 提供 144 TB 共享内存	使用最新统一内存架构（Unified Memory）减少 GPU 和 CPU 间数据传输的时间，实现系统内存的一部分与集成显卡控制器共享
软件	可基于 CUDA 进行高性能计算，充分利用 CUDA 的 GPU 加速库、调试和优化工具、C/C++ 编译器以及用于部署应用程序的运行环境库	基于 ROCm 的软件生态，支持 TensorFlow 和 PyTorch 等主要机器学习框架，以帮助用户加速人工智能工作负载，提供优化的 MIOpen 库到全面的 MIVisionX 计算机视觉和机器智能库、实用程序和应用程序，帮助扩大加速计算所适用的工作负载

价格	价格相对较高	计划延续先前产品的高性价比价格，预计价格相对较低
架构	基于英伟达最新 Hopper 架构，以 Transformer 为加速引擎，Tensor Core 能够大幅加速 Transformer 模型的 AI 计算	CPU+GPU+内存三者结合，内置 13 个小芯片，包括 24 个 Zen4 CPU 内核，同时融合了 CDNA 3 和 8 个 HBM3 显存堆栈，集成了 5nm 和 6nm IP，总共包含 128GB HBM3 显存和 1460 亿晶体管

数据来源：英伟达官网，AMD 官网，IT 之家，国泰君安证券研究

英特尔依托其在集成 GPU 市场的主导地位，提供具有卓越性能的图形解决方案。英特尔与英伟达和 AMD 不同，其在 GPU 领域更加专注集成显卡业务。英特尔的 GPU 家族包括锐炫显卡、锐炬 Xe 显卡和 Data Center GPU 等。英特尔研发了 Xe-HPG 微架构，Xe-HPG GPU 中的每个 Xe 内核都配置了一组 256 位矢量引擎，可实现加速传统图形和计算工作负载，而新的 1024 位矩阵引擎或 Xe 矩阵扩展则旨在加速人工智能工作负载。英特尔也形成了覆盖云计算、人工智能、5G、物联网、边缘计算和商用电脑的业务解决方案，并且其业务也覆盖了 GPU 的制造和封测环节，在台式机和笔记本电脑等领域也具备较客观的市场份额。但整体而言，英特尔的收入增速相对缓慢，受 PC 端出货量负面影响使得其在 GPU 这一核心业务增长动力不足。

高通等破局者投身 GPU 研发制造。以高通发布的第二代骁龙 8 旗舰移动平台（骁龙 8 Gen 2）为例，其采用的新一代 Adreno GPU 相比上一代性能提升 25%、功耗减少了 45%，CPU 的性能也提升了 35%、功耗减少了 40%，反映出了高通在 GPU 芯片设计领域已具备较快的迭代能力，包括华硕、荣耀、OPPO、小米、夏普、索尼、vivo 等企业都将推出搭载骁龙 8 Gen 2 的产品。

图 99 诸多企业将推出搭载骁龙 8 Gen 2 的产品

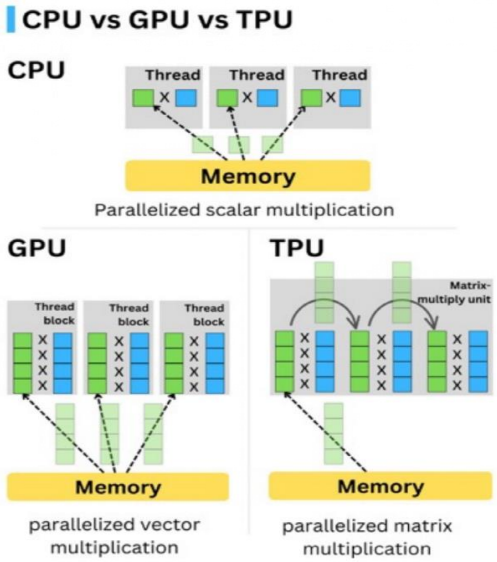
数据来源：中关村在线

头部大厂加速 AI 芯片布局，英伟达的潜在竞争对手或许是互联网头部厂商。我们发现，英伟达的竞争对手或许并不是目前正在研发 GPU 的专业厂商。互联网市场中的头部大厂，包括 Google、阿里、微软、亚马逊和 IBM 等均在进行 AI 芯片研究。微软同时也在着手其 AI 芯片 Athena 的研发，为其 OpenAI 提供硬件支持。整体而言，如 TPU、NPU

的发展，同样适用于人工智能，因此英伟达的潜在竞争风险仍存，并不仅局限于 GPU 设计领域。

Google 推出 TPU，云端服务器提升深度学习计算效能。2014 年起，Google 开始自主研发 AI 专用芯片，并于 2016 年 AlphaGo 战胜李世石之后推出 TPU (Tensor Processing Unit)，TPU 也成为近年来最火热的 ASIC。TPU 使用矩阵乘法阵列进行矩阵运算，在训练复杂神经网络过程中无须像 GPU 多次访问存储单元，并可以通过云 TPU 服务器进行跨设备操作。因此，TPU 实现了将模型参数保存至同一高带宽存储器中，将调用的芯片的空间用以模型运算，降低了能耗并有效提升运行速度。直至 2021 年，Google 已经推出了 TPUv4，一定程度上阻滞了英伟达的市场需求增长。

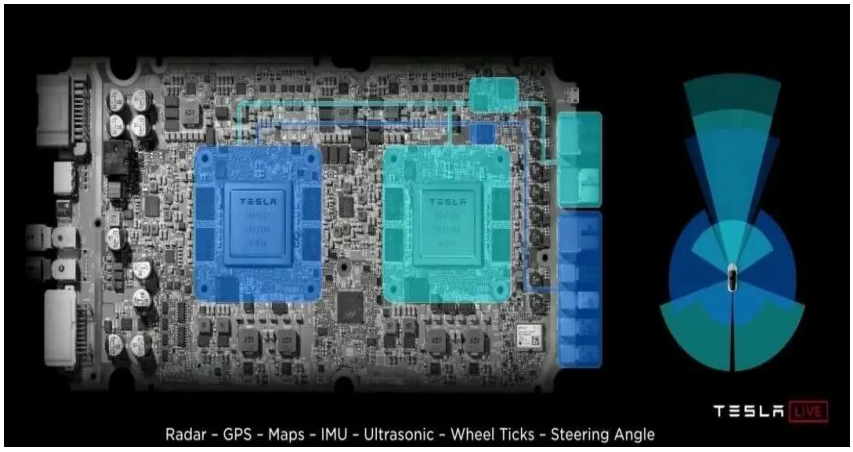
图 100 CPU、GPU、TPU 内存访问次数对比



数据来源：远川研究所

客户向竞争对手转变，特斯拉先后推出以 NPU 为基础的 FSD 车载芯片和 D1 芯片。NPU（Neural Network Processing Unit）在训练神经网络模型时相较 GPU 能耗和成本更低，并更适配嵌入环境，可减少神经网络运算过程的时间。2019 年英伟达的重要客户特斯拉发布其自研 FSD 平台（Full Self-Driving Computer），搭载两块车载芯片，其中的最大组件 NPU 由特斯拉硬件团队定制设计，每个 FSD 芯片内均包含两个相同的 NPU，一块 GPU 和一块 CPU。2021 年特斯拉发布 D1 芯片，并用其打造了 AI 超级计算机 ExaPOD，对比英伟达对特斯拉的既有方案预算，拥有 4 倍的性能、1.3 倍的能效比和仅 1/5 的体积。我们认为，FSD 车载芯片和 D1 芯片的推出，标志着特斯拉对英伟达的芯片依赖度开始下降。

图 101 特斯拉 FSD 平台



数据来源：懂车帝

基于 GPU 相对低的成本和繁荣的生态，仍旧是超算的首位选择，短期内市场地位不会改变。以史为鉴，2017 年 Google 推出 Transformer 模型，成为了 OpenAI 开发 GPT-1 的基础。此后英伟达迅速抓住全球算力需求爆发时机，推出搭载 Transformer 加速引擎的 Hopper 架构，同时推出 H100 Tensor Core GPU，满足了超算的算力要求。整体而言，GPU 的制造成本相比 ASIC 等 AI 芯片最低，生态也最繁荣。同时，由于目前模型正处在不断变化的飞速增长期，基于其较快的迭代速度，ASIC 的定制化设计需要同时根据模型变化的新需求迭代，难以实现稳定的生产。因此 GPU 仍是解决 AI 算力的不二选择，短时间内其市场地位不会改变。

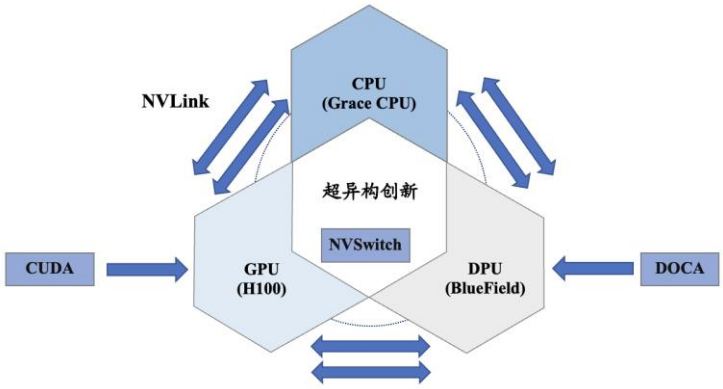
6.3. 以超异构创新构建面向大规模 AI 计算的系统性竞争优势

6.3.1. 超异构创新总览

以超异构创新构建面向大规模 AI 计算的超级计算机。异构计算指是通过调用性能、结构各异的计算单元（包括 CPU、GPU 和各类专用 AI 芯片等）以满足不同的计算需求，实现计算最优化。我们认为，英伟达的核心竞争优势在于，构建了 AI 时代面向大规模并行计算而设的全栈异构的数据中心。英伟达 NVLink 性能快速迭代，同时 NVSwitch 可连接多个 NVLink，在单节点内和节点间实现以 NVLink 能够达到的最高速度进行多对多 GPU 通信，满足了在每个 GPU 之间、GPU 和 CPU 间实现无缝高速通信的需求，同时基于 DOCA 加速数据中心工作负载的潜力，实现 DPU 的效能提升，GPU +Bluefield DPU+Grace CPU 的结合开创性地实现了芯片间的高速互联。同时 CUDA 充当通用平台，引入英伟达软件服务和全生态系统。我们认为，芯片和系统耦合的实现使得英伟达真正实现了超异构创新。

.

图 102 超异构创新整体框架

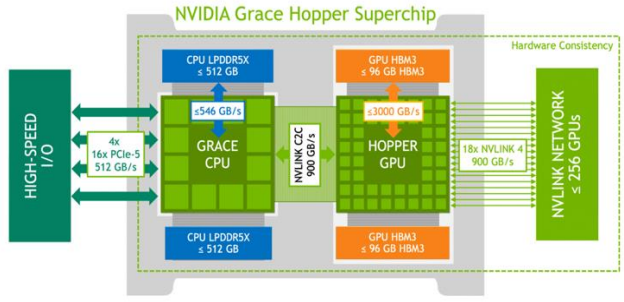


数据来源：英伟达官网，国泰君安证券研究

6.3.2. NVLink

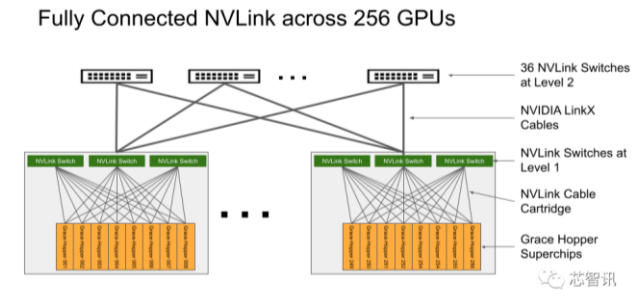
首先，NVLink 改变了传统 PCIe 复杂的传输过程，实现 GPU 与 CPU 的直接连接。以 GH200 超级芯片为例，其使用 NVLink-C2C 芯片互连，将基于 Arm 的 Grace CPU 与 H100 Tensor Core GPU 整合，从而不再需要传统的 CPU 至 GPU PCIe 连接。传统的 PCIe 需要经历由 CPU 到内存，再到主板，最后经过显存到达至 GPU 的过程。因此 NVLink 与传统的 PCIe 技术相比，将 GPU 和 CPU 之间的带宽提高了 7 倍，将互连功耗减少了 5 倍以上，并为 DGX GH200 超级计算机提供了一个 600GB 的 Hopper 架构 GPU 构建模块。

图 103 英伟达 Grace Hopper 超级芯片逻辑概述



数据来源：英伟达官网

图 104 DGX GH200 与 NVLink 系统的拓扑结构



数据来源：芯智讯

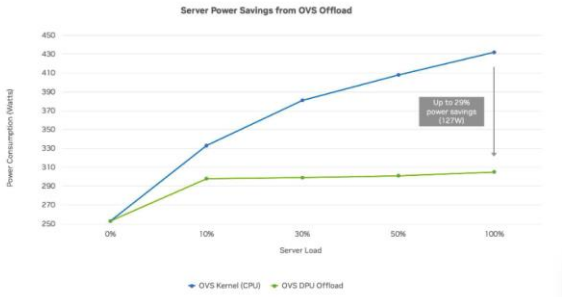
6.3.3. DPU

DPU 大幅降低 CPU 的负荷，为现代数据中心带来前所未有的性能提升。2020 年，英伟达发布 BlueField-2 DPU,将 ConnectX-6 Dx 的强大功能与可编程的 Arm 核心以及其他硬件卸载功能相结合，用于软件定义存储、网络、安全和管理工作负载。之后发布的 BlueField-3 DPU 更为强大，作为一款 400Gb/s 基础设施计算平台，其计算速度高达每秒 400 Gb，计算能力和加密加速均较 BlueField-2 DPU 提高 4 倍，存储处理速度提高 2 倍，内存带宽也提高了 4 倍。同时，BlueField 系列 DPU 有助于降低能耗，在 OVS 平台上进行的一项测试中，在服务器最大荷载时，DPU 能耗较 CPU 低 29%。英伟达亦推出了融合加速器产品，结合其 Ampere GPU 架构和 BlueField DPU 的安全和网络增强功能。

图 105 英伟达 BlueField-3 DPU

数据来源：英伟达官网

图 106 基于 OVS 的测试中 DPU 显著降低了能耗



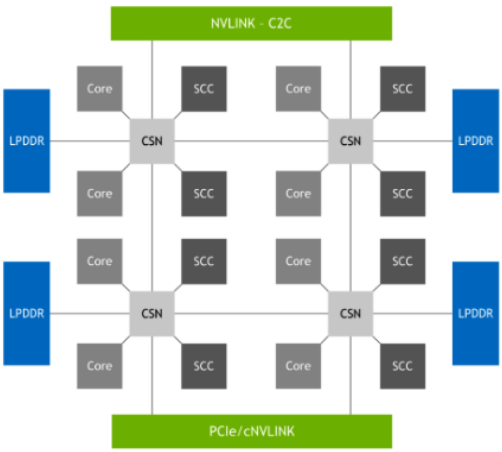
数据来源：英伟达官网

最新 Spectrum-X 网络平台集英伟达 Spectrum-4、BlueField-3 DPU 和加速软件于一身。Spectrum-X 是基于网络创新的新成果而构建，将 Spectrum-4 以太网交换机与英伟达 BlueField-3 DPU 紧密结合，网络平台具有高度的通用性，可用于各种 AI 应用，它采用完全标准的以太网，并与现有以太网的堆栈实现互通，全球头部云服务提供商都可采用该平台来横向扩展其生成式 AI 服务。我们认为，Spectrum-X 的上市将进一步提升英伟达以太网 AI 云的性能与效率，成为英伟达为 AI 工作负载扫清障碍的关键一环。

6.3.4. CPU

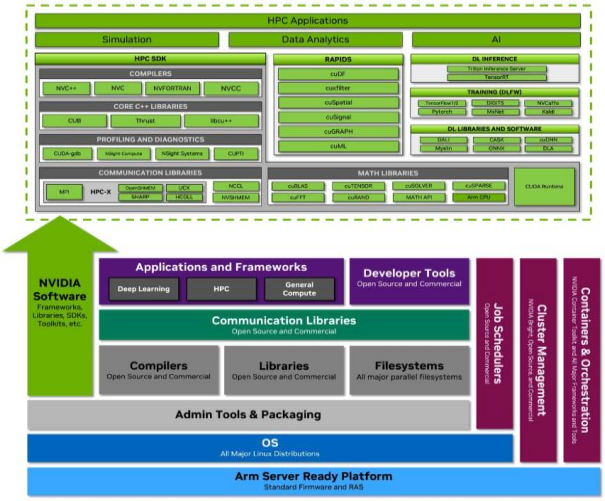
英伟达自研 Grace CPU 超级芯片，为 AI 数据中心而生。不同于传统的 CPU，英伟达 Grace CPU 采用 NVLink C2C 技术，是一款专为数据中心而设计的 CPU，其可运行包括 AI、高性能计算、数据分析、数字孪生和云应用在内的工作负载。Grace CPU 可提供 144 个 Arm Neoverse V2 核心和 1 TB/s 的内存带宽，并引入了可扩展一致性结构 (SCF)，SCF 可用以确保 NVLink-C2C、CPU 内核、内存和系统 IO 之间的数据流量流动。从软件角度，英伟达 Grace CPU 软件生态系统将用于 CPU、GPU 和 DPU 的全套英伟达软件，与完整的 Arm 数据中心生态系统相结合。

图 107 Grace GPU 引入 SCF



数据来源：英伟达官网

图 108 Grace CPU 软件生态系统

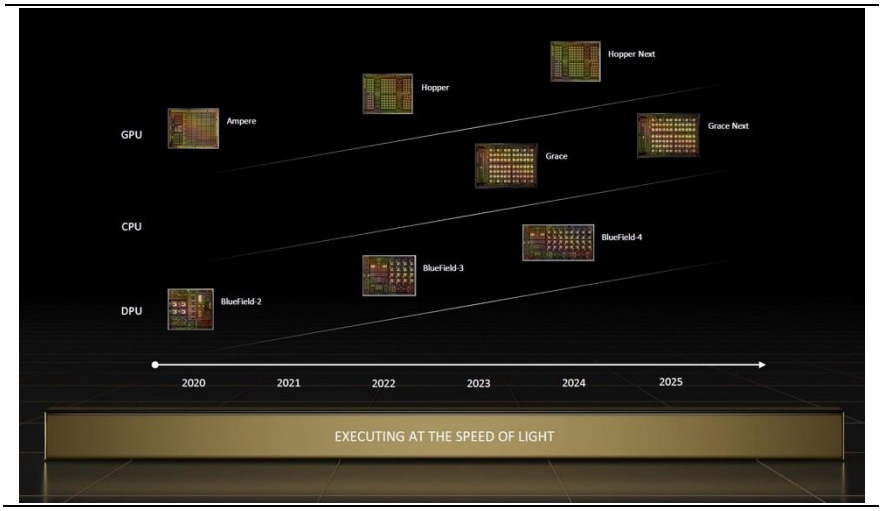


数据来源：英伟达官网

6.3.5. “GPU+DPU+CPU”的三芯战略

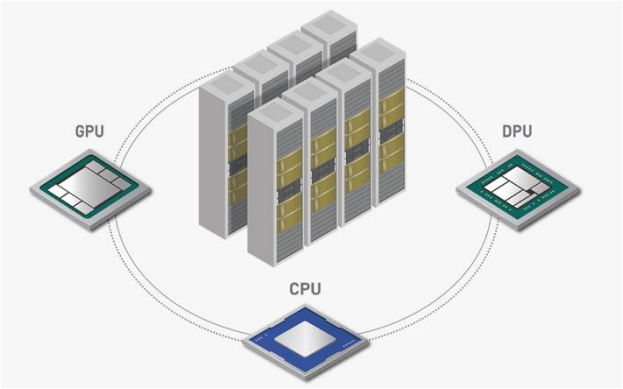
综上，英伟达基于“GPU+DPU+CPU”的三芯战略已初步实现，软件和硬件相互支持，成为 AI 发展的技术标杆。我们认为，英伟达的商业模式正在由销售“硬件+软件”的制造商向大规模 AI 计算的平台公司持续转型，持续通过基于异构计算的硬件迭代加软件服务的整体生态更新提升运算速度，降低运算成本。英伟达通过“GPU+DPU+CPU”构建英伟达加速计算平台，和传统服务器的计算系统相比，加速计算系统新增添了 GPU 和 DPU，为包括 AI 和可视化等现代业务应用提供计算加速器支持。英伟达亚太区开发技术部总经理李曦指出，目前世界上只有 5%的计算任务被加速，而未来十年所有的计算任务都将被加速，还会诞生十倍于现阶段的新计算任务，这将为加速计算市场带来超 100 倍的增长空间。

图 109 英伟达的“三芯”战略



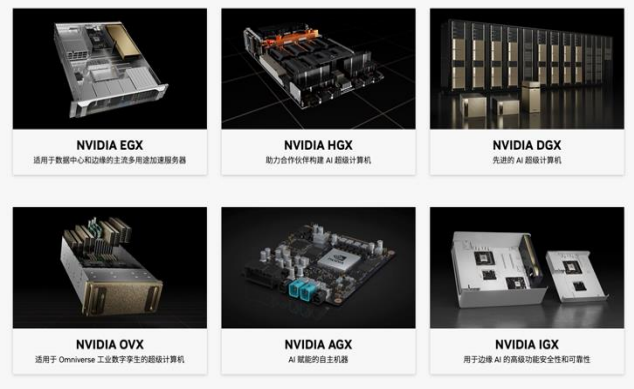
数据来源：英伟达官网

图 110 GPU+DPU+CPU 构建加速计算平台



数据来源：英伟达官网

图 111 英伟达打造的一系列加速计算平台



数据来源：英伟达官网

6.3.6. CUDA 和 DOCA

CUDA 和 DOCA 打造软件生态，进而与硬件组成全栈系统优势。如前所述，CUDA 可以充当英伟达各 GPU 系列的通用平台，因此开发者可以跨 GPU 配置部署并扩展应用。借助于 CUDA 的高兼容性，英伟达成功将 GPU 的应用领域拓展至计算科学和深度学习领域。而 DOCA 的最主要功能为加速、卸载并将数据中心基础架构 DPU 隔离，真正充分

发挥了人工智能的潜力，推动数据中心转向加速计算，以满足日益增长的计算需求。

6.3.7. GH200

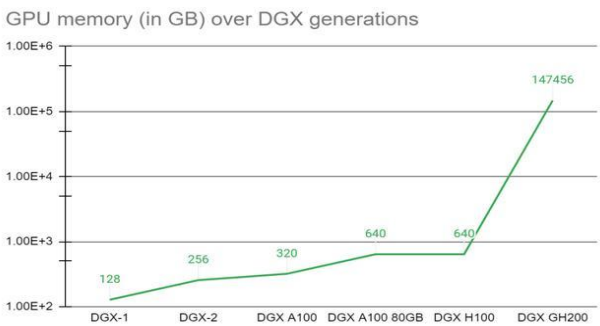
基于超异构创新，英伟达发布能提供超强 AI 性能的 DGX GH200 大内存 AI 超级计算机。DGX 系统利用全堆栈解决方案和企业级支持，为企业 AI 基础架构设定标杆，是应用于 TOP500 中多台超级计算机的核心基础模组。DGX GH200 作为最新产品，整合了 Grace CPU 和 H100 GPU，拥有近 2000 亿个晶体管，通过定制的 NVLink Switch System 将 256 个 GH200 超级芯片和高达 144TB 的共享内存连接成一个单元，使 DGX GH200 系统中的 256 个 H100 GPU 作为一个整体协同运行。DGX GH200 提供 1 exaflop 性能与 144 TB 共享内存，比单个 DGX A100 320GB 系统高出近 500 倍。这让开发者可以构建用于生成式 AI 聊天机器人的大型语言模型、用于推荐系统的复杂算法，以及用于欺诈检测和数据分析的图形神经网络。

图 112 DGX GH200 超级计算机

数据来源：英伟达官网

GH200 超级芯片是英伟达系统性竞争优势的集大成者。我们认为，GH200 超级芯片集合了最先进的 Grace Hopper 架构，并应用第四代 Tensor Core 提升计算性能、进行模型优化，NVLink 实现了高速的传输，这都将进一步形成英伟达的竞争壁垒。随着 Grace Hopper 超级芯片的全面投产，全球的制造商很快将会提供企业使用专有数据构建和部署生成式 AI 应用所需的加速基础设施。谷歌云、Meta 和微软是首批有望接入 DGX GH200 的企业。

图 113 DGX GH200 的 GPU 内存大幅增长



数据来源：芯智讯

总的来说，英伟达作为龙头企业将大比例享受 AI 芯片行业整体需求高增带来的红利。如本报告先前所述，IDTechEx 预测 2033 年全球 AI 芯片市场将增长至 2576 亿美元。JPR 预测 2022-2026 年全球 GPU 销量复合增速将保持在 6.3%水平。摩根大通的预测认为，英伟达将在 2023 年的人工智能产品市场中获得 60%的份额，主要来自于 GPU 和网络互连产品。因此，英伟达作为人工智能产业的上游龙头供应商，我们看好市场需求的激增对于英伟达产品的爆发式需求增长。以超异构创新研发能力优势和业内领先的生态，以及对于以生成式 AI 为代表的人工智能迅速带来业务变革的准确把握，其依旧具备市场领先的地位，短时间内其龙头地位不会改变。

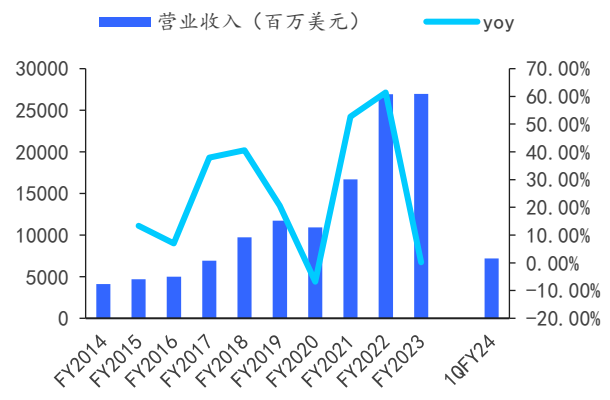
7. 数据中心助推营收超预期，市值突破开创新高点

7.1. 营收指标增势明显，盈利能力优势充分彰显

营收及利润波动较大，盈利能力增长可期。英伟达 FY2022/FY2023/1QFY24 营业收入分别为 269.14/269.74/71.92 亿美元，同比+61.40%/+0.22%/-13.22%；FY2022/FY2023/1QFY24 销售成本为 94.39/116.18/25.44 亿美元，同比+50.33%/+23.09%/-10.96%；FY2022/FY2023/1QFY24 净利润为 97.52/43.68/20.43 亿美元，同比+125.12%/-55.21%/+26.27%。营业收入和净利润近年来整体呈波动上升趋势，呈现较大波动特征，尤其 FY2023 净利润出现大幅下跌，不及 FY2022 一半。但 1QFY24 营收增长超预期明显，未来盈利能力有望持续高增。

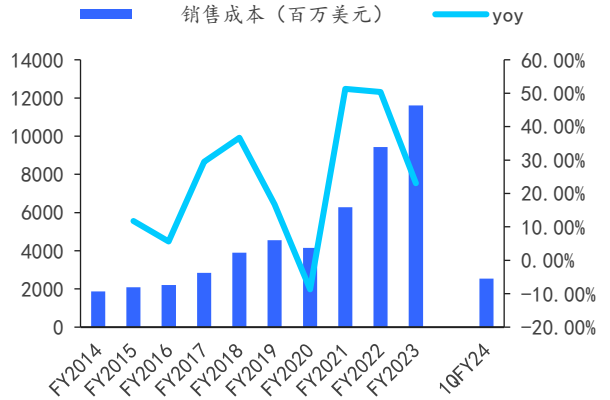
销售毛利率和净利率增势明显，但 2023 财年出现小幅下跌。公司 FY2022/FY2023/1QFY24 销售净利率分别为 36.23%/16.19%/28.41%，同比+10.25pct/-20.04pct/+8.89pct，销售毛利率分别 64.93%/56.93%/64.63%，同比+2.59pct/-8.00pct/+0.90pct，整体保持积极增速，但 FY2023 呈现一定跌幅。1QFY24，销售毛利率和净利率再度回升。

图 114 英伟达 FY2014-1QFY24 营业收入波动上升



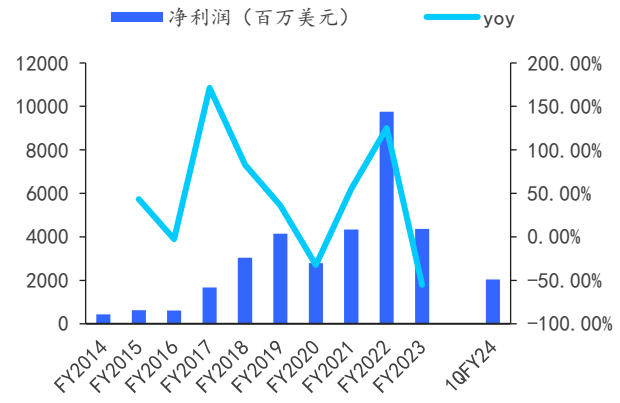
数据来源：iFinD，国泰君安证券研究

图 115 英伟达 FY2014-1QFY24 销售成本逐步上升



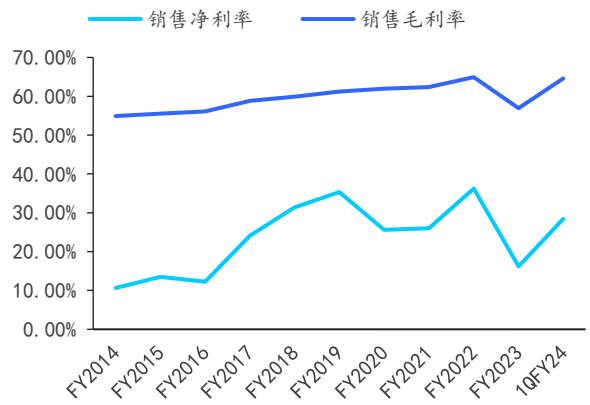
数据来源：iFinD，国泰君安证券研究

图 116 英伟达 FY2014-1QFY24 净利润波动上升



数据来源：iFinD，国泰君安证券研究

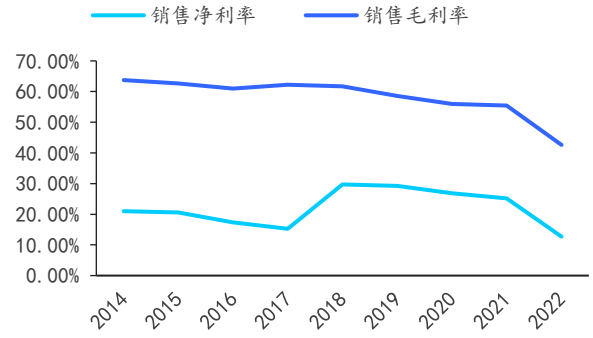
图 117 英伟达销售毛利率、净利率呈上升态势



数据来源：iFinD，国泰君安证券研究

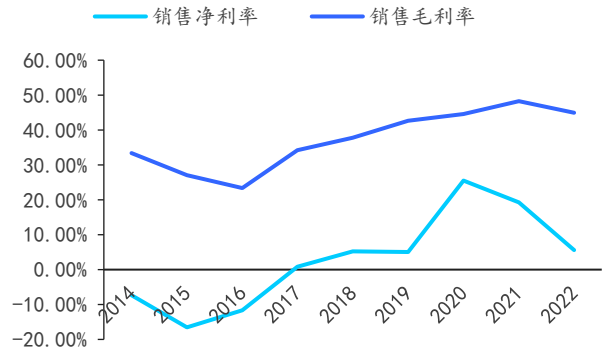
英伟达销售毛利率和净利率显著高于英特尔和 AMD，彰显盈利能力优势。对比公司两大竞争对手英特尔和 AMD：英特尔 2022 年销售净利率 12.71%，销售毛利率 42.61%；AMD 销售净利率 5.59%，销售毛利率 44.93%，二者均低于英伟达在 FY2023 的表现，反映英伟达相比主要竞争对手具备更高的盈利能力。

图 118 英特尔销售净利率与毛利率呈下跌趋势



数据来源：iFinD，国泰君安证券研究

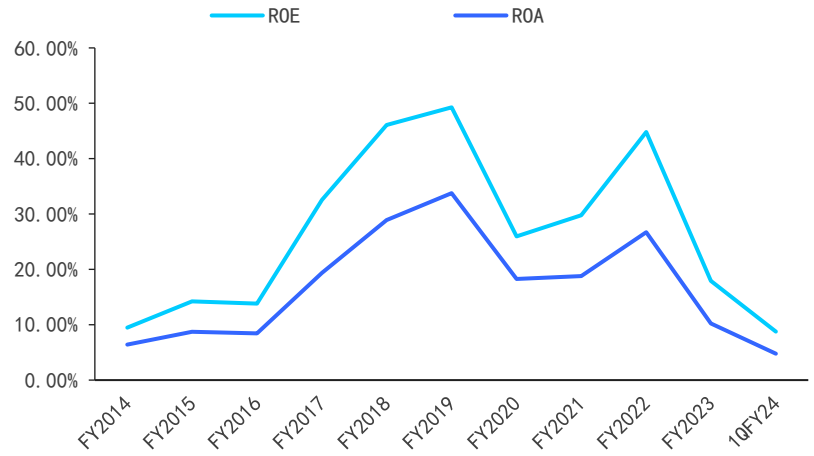
图 119 AMD 销售净利率与毛利率不及英伟达



数据来源：iFinD，国泰君安证券研究

英伟达 FY2023 营收与利润下跌因素逐步化解，看好公司长期盈利能力。FY2022/FY2023/1QFY24 英伟达 ROE 分别为 44.83%/17.93%/8.76%，ROA 分别为 26.73%/10.23%/4.77%，公司 FY2023 盈利能力层面逆风。我们认为，英伟达 FY2023 营收不及预期主要由游戏收入下降导致，2020 年受全球疫情影响，显卡市场炒作情绪狂热，显卡价格一路飙升，而随着疫情影响逐步减弱，显卡市场需求导向转向疲弱。同时黄仁勋指出，中国市场业务受阻也极大影响了英伟达营收表现，但随着宏观逆风因素逐步消散，以及 2022 年末 GPT 席卷行业带来的需求激增，我们认为英伟达在 2024 财年营收有望得到持续改善。

图 120 英伟达 FY2014-1QFY24 ROE 与 ROA 增长可期

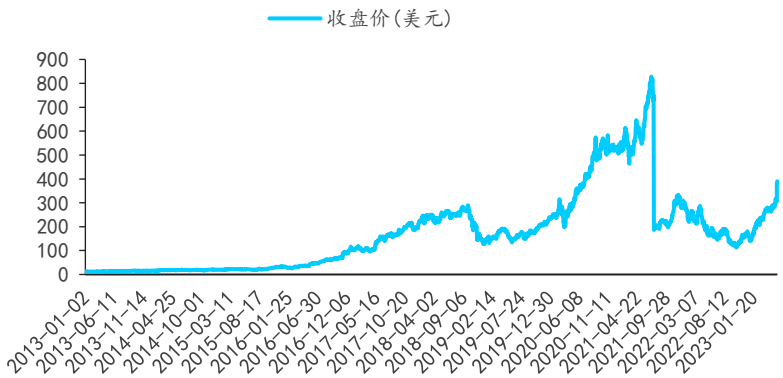


数据来源：iFinD，国泰君安证券研究

7.2. GPT 带动市值高增，股价转向上升通道

股价重返上升通道，盈利能力持续释放。英伟达股价 2013 年 1 月 2 日仅 12.72 美元，2016 年起一路高增，2018 年末回调后自 2019 年年终起再度踏入上升通道（注：图中收盘价在 2021 年 7 月 20 日直线下跌是由于英伟达当日以 1：4 的比例拆分股票所致）。2022 年初，受业绩预期放缓影响，英伟达股价呈较明显下跌趋势，自 2023 年年初起，市场逐步对英伟达投资价值形成一致预期，伴随着价值挖掘深入，潜在盈利能力有望持续释放。2023 年 5 月 25 日，受一季报营收超预期和 2QFY24 应用收入展望达 110 亿美元影响，英伟达股价迅速高增至 379.8 美元。

图 121 英伟达股价重返上升通道



数据来源：iFinD，国泰君安证券研究

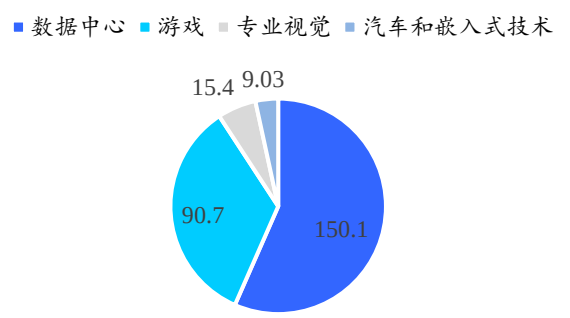
公司市值受 GPT 带动一路高升。伴随着公司股价高涨，英伟达股票市值爆发式抬升。截至 2023 年 5 月 26 日，英伟达市值约 9630 亿美元，而同日英特尔市值约 1230 亿美元、AMD 市值约 1700 亿美元，英伟达市值处行业龙头水平，已远超英特尔与 AMD 市值之和。

7.3. 数据中心成为盈利主要驱动，成就营收高增奇迹

数据中心业务营收占比过半，成为营收增长的主要驱动因素。据英

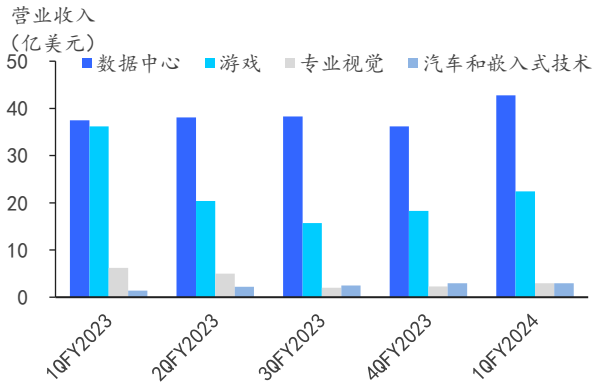
伟达财报，英伟达将其主营业务分为四大领域，分别是数据中心、游戏、专业视觉、汽车和嵌入式技术。FY2023 上述四大业务营收分别为 150.1/90.7/15.4/9.03 亿美元，同比+41%/-27%/-27%/+60%。1QFY2024 四大业务营收分别为 42.8/22.4/2.95/2.96 亿美元，同比+14%/-38%/-53%/+114%，数据中心和游戏业务为英伟达营业收入的最主要来源。其中，FY2023Q2 起游戏业务大幅下跌，此后的三季度依旧低位徘徊，对全年营收造成较大负面影响。但整体而言，数据中心业务高增速推动了营收的高增量，部分缓解了游戏业务低迷对营收增长的阻滞。

图 122 FY2023 英伟达营收中数据中心占比过半



数据来源：英伟达财报，国泰君安证券研究

图 123 数据中心业务高增速推动了营收的高增量



数据来源：英伟达财报，国泰君安证券研究

大模型训练催生算力需求，英伟达当下在模型训练和推理中的地位短期不会改变。对于以 ChatGPT 为代表的 AI 产业，英伟达已形成 CPU+GPU+DPU 的硬件组合，并已 CUDA 软件平台为基石打造应用生态。1QFY24 中英伟达推出的四款推理平台，这些平台将英伟达的全栈推理软件与最新的 NVIDIA Ada、NVIDIA Hopper 和 NVIDIA Grace Hopper 处理器结合在一起，更加稳固了英伟达在模型训练和推理中的地位。英伟达表示，云服务商对公司的基础架构十分感兴趣，英伟达直接与全球近一万家人工智能初创公司合作，同时随着经济好转，宏观逆风逐渐消散，企业上云的进程将会恢复。我们认为，其数据中心业务未来盈利可期。

8. 投资建议

行业龙头当仁不让，英伟达盈利能力可期。考虑到英伟达 1QFY2024 营收的出色表现，包括数据中心收入创下 42.8 亿美元的纪录，以及英伟达自身对于 2QFY2024 的收入展望达 110.0 亿美元的乐观预期，我们预计公司 FY2024E/FY2025E/FY2026E 营业收入分别为 400.0/516.26/620 亿美元，同增 48.29%/29.07%/20.09%，FY2024E/FY2025E/FY2026E 经调整净利润分别为 151.96/223.07/285.79 亿美元，同增 247.89%/46.80%/28.12%。

图 124 英伟达盈利预测

资产负债表 单位：美元(百万)

项目/报告期	2021A	2022A	2023A	2024E	2025E	2026E
流动资产	16,055	28,829	23,073	39,351	63,565	95,346
货币资金	847	1,990	3,389	17,293	37,991	67,224
应收账款	2,668	5,016	4,618	6,901	9,230	10,615
预付及其他流动资产	12,540	21,823	15,066	15,157	16,344	17,507
非流动资产	12,736	15,358	18,109	17,124	16,237	15,438
长期投资	0	0	0	0	0	0
固定资产	2,149	2,778	3,807	3,426	3,084	2,775
无形资产净值	6,930	6,688	6,048	5,443	4,899	4,409
其他非流动资产	3,657	5,892	8,254	8,254	8,254	8,254
资产总计	28,791	44,187	41,182	56,475	79,801	110,784
流动负债	3,925	4,335	6,563	7,840	10,039	13,623
应付账款	1,201	1,783	1,193	1,910	2,399	2,903
应交税金	61	0	467	280	447	1,073
交易性金融负债	0	0	0	0	0	0
借贷到期部分	999	0	1,250	1,250	1,250	1,250
其他流动负债	1,664	2,552	3,653	4,400	5,943	8,396
非流动负债	7,973	13,240	12,518	12,518	12,518	12,518
负债合计	11,898	17,575	19,081	20,358	22,557	26,141
普通股	1	3	2	2	2	2
库存股	10,756	0	0	0	0	0
储备	27,629	26,620	22,142	36,158	57,285	84,684
其他综合收益	19	-11	-43	-43	-43	-43
归属母公司股东权益	16,893	26,612	22,101	36,117	57,244	84,643
少数股东权益	0	0	0	0	0	0
负债和股东权益	28,791	44,187	41,182	56,475	79,801	110,784

现金流量表 单位：美元(百万)

项目/报告期	2021A	2022A	2023A	2024E	2025E	2026E
经营活动现金流	5,822	9,108	5,641	15,599	22,373	30,961
净利润	4,332	9,752	4,368	15,196	22,307	28,579
折旧摊销	1,098	1,174	1,544	986	887	798
营运资金变动	-703	-3,363	-2,207	-1,097	-1,317	1,036
其它	1,095	1,545	1,936	515	496	547
投资活动现金流	-19,675	-9,830	7,375	32	52	1
资本支出	-1,128	-976	-1,833	0	0	0
投资变动	-10,023	-8,591	9,257	0	0	0
其他	-8,524	-263	-49	32	52	1
筹资活动现金流	3,804	1,865	-11,617	-1,728	-1,728	-1,728
银行借款	4,968	3,977	0	0	0	0
股本增加	0	0	-10,039	0	0	0
支付的利息和股利	-395	-399	-398	-548	-548	-548
其他	-769	-1,713	-1,180	-1,180	-1,180	-1,180
现金净增加额	-10,049	1,143	1,399	13,904	20,697	29,234
期初现金余额	10,896	847	1,990	3,389	17,293	37,991
期末现金余额	847	1,990	3,389	17,293	37,991	67,224

利润表 单位：美元(百万)

项目/报告期	2021A	2022A	2023A	2024E	2025E	2026E
营业收入	16,675	26,914	26,974	40,000	51,626	62,000
营业成本	6,279	9,439	11,618	12,500	15,700	19,000
毛利	10,396	17,475	15,356	27,500	35,926	43,000
其他收入	0	0	0	0	0	0
销售费用	0	0	0	2,800	3,356	3,906
研发费用	3,924	5,268	7,339	9,200	10,841	12,400
营业利润	4,532	10,041	5,577	15,500	21,729	26,694
利息收入	57	29	267	169	865	1,900
利息支出	184	236	262	548	548	548
权益性投资损益	0	0	0	0	0	0
其他非经营性损益	4	107	-48	32	52	1
其他损益	0	0	-1,353	0	0	0
除税前利润	4,409	9,941	4,181	15,154	22,098	28,047
所得税	77	189	-187	-42	-210	-533
净利润（含少数股东损益）	4,332	9,752	4,368	15,196	22,307	28,579
少数股东损益	0	0	0	0	0	0
净利润	4,332	9,752	4,368	15,196	22,307	28,579
优先股利及其他调整项	0	0	0	0	0	0
归属普通股东净利润	4,332	9,752	4,368	15,196	22,307	28,579

主要财务比率

项目/报告期	2021A	2022A	2023A	2024E	2025E	2026E
成长能力(%)						
营业收入增长	52.7%	61.4%	0.2%	48.3%	29.1%	20.1%
营业利润增长	59.2%	121.6%	-44.5%	177.9%	40.2%	22.9%
归母净利润增长	54.9%	125.1%	-55.2%	247.9%	46.8%	28.1%
获利能力（%）						
毛利率	62.3%	64.9%	56.9%	68.8%	69.6%	69.4%
净利率	26.0%	36.2%	16.2%	38.0%	43.2%	46.1%
ROE	25.6%	36.6%	19.8%	42.1%	39.0%	33.8%
偿债能力						
资产负债率(%)	41.3%	39.8%	46.3%	36.0%	28.3%	23.6%
流动比率	4.09	6.65	3.52	5.02	6.33	7.00
速动比率	0.90	1.62	1.22	3.09	4.70	5.71
营运能力						
总资产周转天数	497.70	488.07	569.68	439.46	475.14	553.31
应收账款周转天数	48.38	51.39	64.29	51.84	56.24	57.61
存货周转天数	80.41	84.50	120.29	149.89	133.99	132.98
每股指标（元）						
每股收益	6.99	3.81	1.77	6.15	9.03	11.57
每股经营现金流	9.39	3.56	2.29	6.32	9.06	12.53
每股营业收入	26.90	10.51	10.94	16.19	20.90	25.10
每股净资产	27.25	10.40	8.96	14.62	23.18	34.27
估值比率						
P/S	19.32	21.72	18.62	23.36	18.10	15.07
P/E	74.36	59.96	114.97	61.50	41.89	32.70
EV/EBITDA	58.26	52.43	89.12	56.19	40.03	31.94

数据来源：Bloomberg，wind，国泰君安证券研究预测

估值方面，我们选取全球半导体市场的头部企业作为英伟达的可比公司。结合彭博的一致预测，可比公司 2023E 平均 PE 46.1X。英伟达作为业内有目共睹的头部公司，在图形处理领域拥有超群的技术实力和领导地位，产品生态具备显著的稀缺性。同时，在此次人工智能的大浪潮中，英伟达将在算力领域充分受益，客户需求递增，强大的生态系统使得其他竞争对手难以复制。因此我们给予其超出行业平均的 PE 70.0X，首次覆盖，并给予“增持”评级。

表 8 可比公司估值表

代码	名称	市值（百万美元）	股价（美元）	2022：NON-GAAP 净利润（百万美元）	2023：PE	2023：PS	2022：EPS（美元）	2022：净利润（百万美元）
平均值		205715.4	194.2	4258	46.1	11.0	5.7	4912
NVDA.US	英伟达	990741.7	401.1	8366	122.9	35.7	1.8	4368
AMD.US	AMD	201730.0	125.3	727.0	44.7	8.9	0.8	1320.0
INTC.US	英特尔	125088.3	30.0	-2893.4	116.5	2.4	1.9	8014.0
AVGO.US	博通股份	334931.5	803.3	12927.6	19.4	9.5	26.5	11495.0
QCOM.US	高通公司	129224.0	116.0	11287.6	14.0	3.4	11.4	12936.0
LSCC.US	莱迪斯半导体	11621.7	84.4	204.9	40.7	15.0	1.3	178.9
MRVL.US	迈威尔科技	54524.0	63.4	1822.0	29.8	9.5	-0.2	-164.0
MCHP.US	微芯科技	41912.8	76.8	3353.1	12.8	5.1	4.1	2237.7

RMBS.US	Rambus	7133.7	65.5	86.7	37.1	12.1	-0.1	-14.3
TXN.US	德州仪器	160246.3	176.6	8264.1	23.5	8.9	9.4	8749.0

数据来源：Bloomberg，iFind，国泰君安证券研究（注：数据截至 2023/5/31,各公司数据以其最新财年年报计算）

9. 风险提示

AI 应用发展不及预期；公司研发进度不及预期；地缘政治冲突影响产品销售。

国泰君安海外科技团队介绍

深耕全球互联网，辐射海外大科技，全面覆盖社交、游戏、电商、互联网金融、互联网服务、AI 及硬科技、美股等领域，致力于结合产业视角与买方视角做差异化研究。

秦和平

执业证书编号：S0880123010042

海外科技领域负责人、首席研究员

梁昭晋

执业证书编号：S0880523010002

海外科技分析师

李奇

执业证书编号：S0880523060001

海外科技分析师

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，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，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

本报告仅供国泰君安证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。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，本公司可发出与本报告所载资料、意见及推测不一致的报告。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。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

评级说明

投资建议的比较标准		评级	说明
投资评级分为股票评级和行业评级。 以报告发布后的 12 个月内的市场表现为比较标准，报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的当地市场指数涨跌幅为基准。	股票投资评级	增持	相对当地市场指数涨幅 15%以上
		谨慎增持	相对当地市场指数涨幅介于 5%～15%之间
		中性	相对当地市场指数涨幅介于-5%～5%
		减持	相对当地市场指数下跌 5%以上
	行业投资评级	增持	明显强于当地市场指数
		中性	基本与当地市场指数持平
		减持	明显弱于当地市场指数

国泰君安证券研究所

	上海	深圳	北京
地址	上海市静安区新闸路 669 号博华广场 20 层	深圳市福田区益田路 6003 号荣超商务中心 B 栋 27 层	北京市西城区金融大街甲 9 号 金融街中心南楼 18 层
邮编	200041	518026	100032
电话	（021）38676666	（0755）23976888	（010）83939888

E-mail：gtjaresearch@gtjas.com

附：海外当地市场指数

亚洲指数名称	美洲指数名称	欧洲指数名称	澳洲指数名称
沪深 300	标普 500	希腊雅典 ASE	澳大利亚标普 200
恒生指数	加拿大 S&P/TSX	奥地利 ATX	新西兰 50
日经 225	墨西哥 BOLSA	冰岛 ICEX	
韩国 KOSPI	巴西 BOVESPA	挪威 OSEBX	
富时新加坡海峡时报		布拉格指数	
台湾加权		西班牙 IBEX35	
印度孟买 SENSEX		俄罗斯 RTS	
印尼雅加达综合		富时意大利 MIB	
越南胡志明		波兰 WIG	
富时马来西亚 KLCI		比利时 BFX	
泰国 SET		英国富时 100	
巴基斯坦卡拉奇		德国 DAX30	
斯里兰卡科伦坡		葡萄牙 PSI20	
		芬兰赫尔辛基	
		瑞士 SMI	
		法国 CAC40	
		英国富时 250	
		欧洲斯托克 50	
		OMX 哥本哈根 20	
		瑞典 OMXSPI	
		爱尔兰综合	
		荷兰 AEX	
		富时 AIM 全股	