开源证券

英伟达（NVDA.O）

2025 年 03 月 25 日

“三芯”齐驱，高速互联，再战 10 万卡集群

——美股公司首次覆盖报告

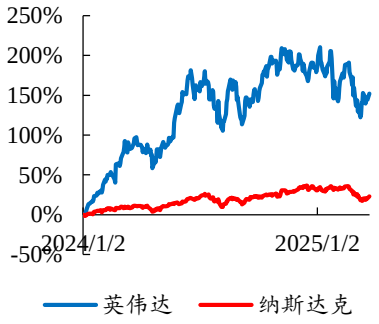
投资评级：买入（首次）

吴柳燕（分析师）
wuliuyan@kysec.cn
证书编号：S0790521110001

杨哲（分析师）
yangzhe@kysec.cn
证书编号：S0790524100001

日期	2025/3/25
当前股价(元)	121.41
一年最高最低(元)	153.12/75.58
总市值(亿元)	29624
流通市值(亿元)	29624
总股本(亿股)	244.00
流通股本(亿股)	244.00
近 3 个月换手率(%)	66.52%

股价走势图



数据来源：聚源

● **全球 AI 算力龙头，B 系列放量在即，给予“买入”评级**

依托 CUDA 体系构建的护城河，英伟达逐步发展为高性能计算行业的领军者，在整体 GPU 领域市场份额达到 80%，在数据中心 GPU 更是达到 98%的市场份额，公司下一代 GPU 产品 B 系列放量在即，有望驱动后续业绩增长，预计 FY2026-2028 年 GAAP 净利润为 1104/1439/1626 亿美元，对应 EPS 分别为 4.75/6.15/6.95 美元，同比增长 52%/30%/13%，当前股价对应 FY2026-2028 年的 PE 估值为 25.6/19.7/17.5 倍。随着架构持续升级，英伟达 GPU 仍有望成为高算力集群时代的首要选择，“三芯战略”、10 万卡网络互联平台、汽车及机器人等领域存在想象空间。首次覆盖，给予“买入”评级。

● **打造“三芯”战略，实现数据摩尔定律**

当前数据中心已成为英伟达核心的业绩驱动，公司以 GPU 为核心，实行“GPU+CPU+DPU”三位一体的产品战略，提供基于 CUDA 的行业领先 GPU 设备，并可通过组件形式（HGX、DGX、NVL72 等）提供加速计算解决方案。(1) GPU：架构持续迭代，在最新的 Blackwell 架构中，GPU 有望达到 20000 TFLOPS FP4 算力，较以往代际的架构有本质的提升；(2) CPU：依托 Arm 实现较强内存一致性，NVLink-C2C 保证芯片高宽带互联，更能适应 AI 数据计算；(3) DPU：收购 Mellanox，加速了 DPU 技术的落地，以实现数据摩尔定律。

● **深化网络互联技术布局，静待 10 万卡集群时代**

10 万卡时代到来，网络集群能力将愈发重要。在 GPU 互联上，英伟达 NVLink 技术可实现 GPU 数据直连，NVSwitch 提升 GPU 链路上限，用于 Blackwell 架构的 NVLink5.0，整体双向带宽将达到 1.8TB/s，是 PCIe 带宽的 14 倍，相比上代突破较大；在算力集群上，英伟达充分布局 Infiniband 和以太网，Spectrum 有望在推理场景中充分放量。

● **风险提示：**产能爬坡低于预期、行业需求低于预期、行业竞争加剧。

财务摘要和估值指标

指标	2024A	2025A	2026E	2027E	2028E
营业收入(百万美元)	60,922	130,497	204,677	254,251	288,462
YOY(%)	125.9	114.2	56.8	24.2	13.5
净利润(百万美元)	29,759	72,880	110,411	143,863	162,615
YOY(%)	581.3	144.9	51.5	30.3	13.0
毛利率(%)	73.6	75.3	72.8	74.4	73.6
净利率(%)	48.8	55.8	53.9	56.6	56.4
ROE(%)	69.2	91.9	81.6	70.0	55.0
EPS(摊薄/美元)	1.32	3.04	4.75	6.15	6.95
P/E(倍)	91.7	39.9	25.6	19.7	17.5
P/B(倍)	68.9	37.3	21.9	14.4	10.0

数据来源：聚源、开源证券研究所

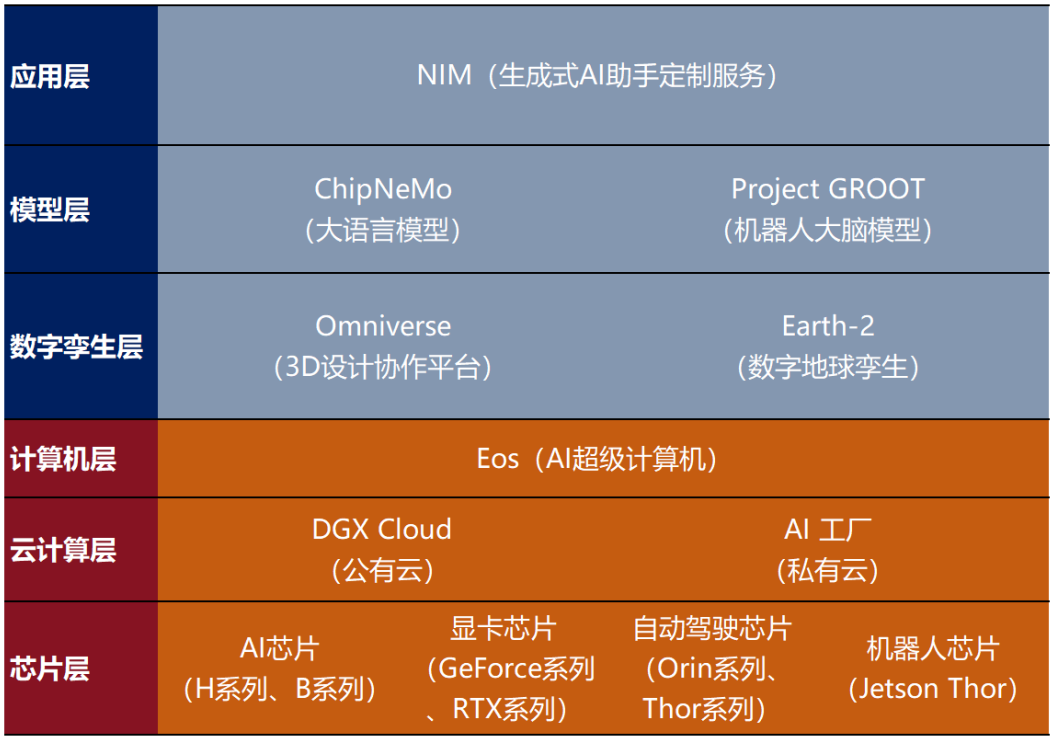
目 录

图表目录

1、 英伟达：全球算力领军者，全方位布局 AI 产业

全球 GPU 龙头，充分布局 Gen-AI。英伟达业务起始于图形处理器，打造了通用计算体系 CUDA 架构，由此开启了加速计算的新纪元，逐步发展为高性能计算行业的领军者，在当下火热的 Gen-AI 行情中占据关键位置。依托 CUDA 体系构建的护城河，英伟达在整体 GPU 领域市场份额达到 80%，在数据中心 GPU 市场更是达到 98%的份额。产品上，英伟达充分布局，在数据中心业务持续创收的同时，发掘多条成长曲线，实现了从芯片层、云计算层到软件应用层的全方位布局，为未来持续发展奠定基础。

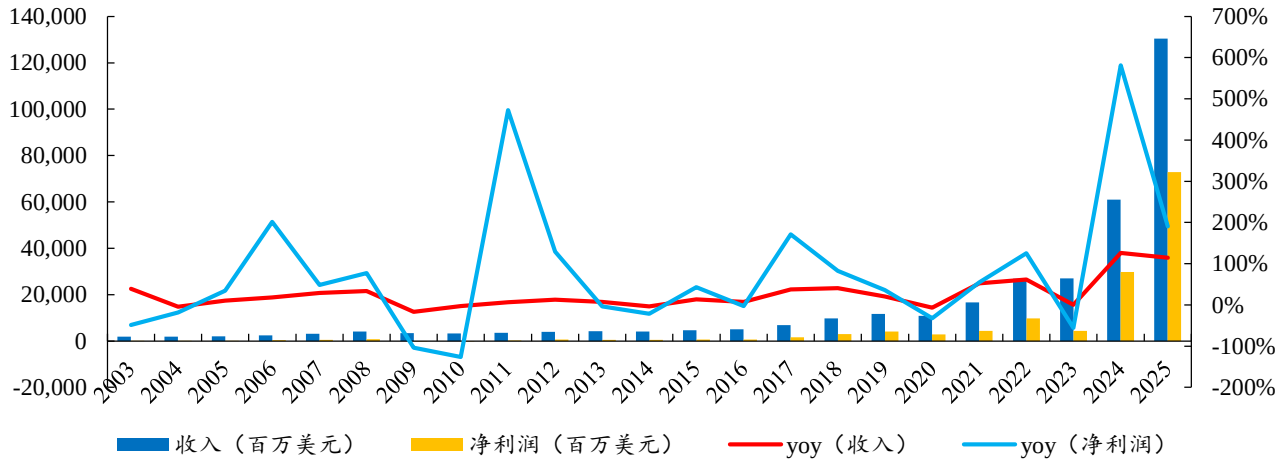
图1：英伟达自下而上布局了从芯片到应用的几乎所有层级



资料来源： 公司官网、开源证券研究所

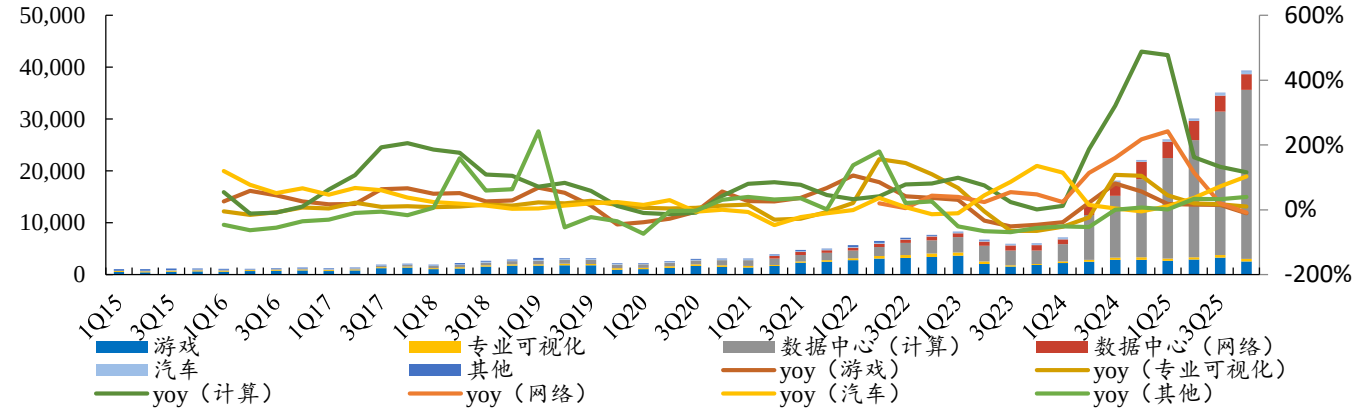
受益于生成式 AI 带来的行业变革，数据中心业务成为核心增长引擎。在 2023 年（对应英伟达 2024 财年）之前，尽管英伟达在数据中心已有充分布局，但收入整体仍然受到游戏行业周期及 GeForce 更新迭代影响。2022 年 Q4，基于 Transformer 架构的 ChatGPT 诞生，带动科技行业加大 GPU 数据中心投入，作为核心“卖水人”的英伟达，数据中心业务迎来快速增长，FY2025Q4，英伟达游戏/专业可视化/计算/网络/汽车收入占比为 6%/1%/83%/8%/1%。

图2：FY2024 以来英伟达收入提速明显



资料来源： Bloomberg、开源证券研究所

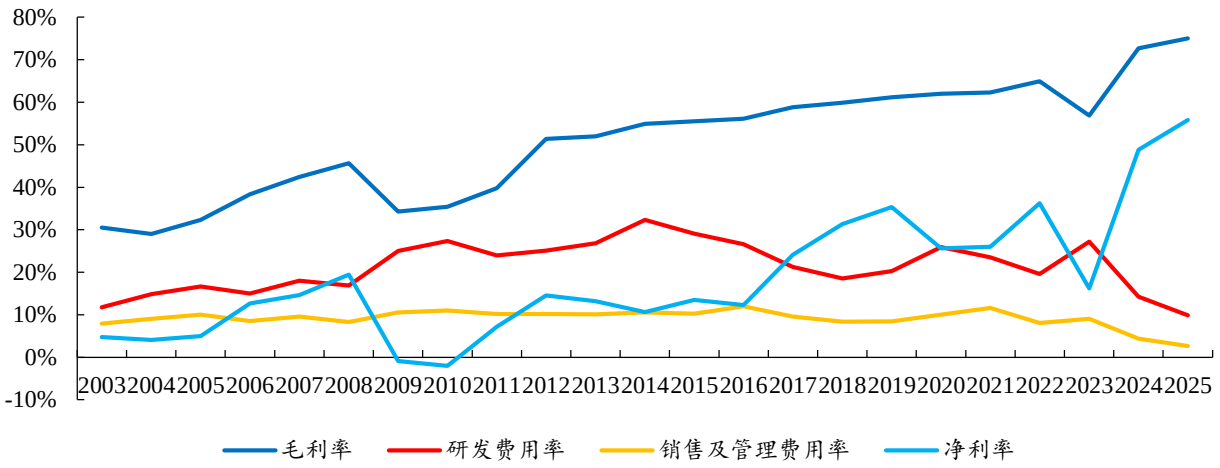
图3：数据中心业务成为英伟达的核心增长来源（营收单位：百万美元）



资料来源： Bloomberg、开源证券研究所

英伟达 GPU 地位稳固，稳健升级带动毛利率提升。从较长的时间周期上看，得益于英伟达产品稳固的市场地位，GPU 架构按照 2-3 年的速度持续更新，带动毛利率稳步提升，FY2014-FY2018 年，数据中心业务快速起量，规模效应下费率摊薄明显，至 FY2019 年净利率达到 35%。此后在生成式 AI 带动下，高盈利水平的数据中心业务占比持续提升，至 FY2025 净利率达到 56%。

图4：随着产品持续迭代，中长期看英伟达毛利率稳步提升，带动净利率上行



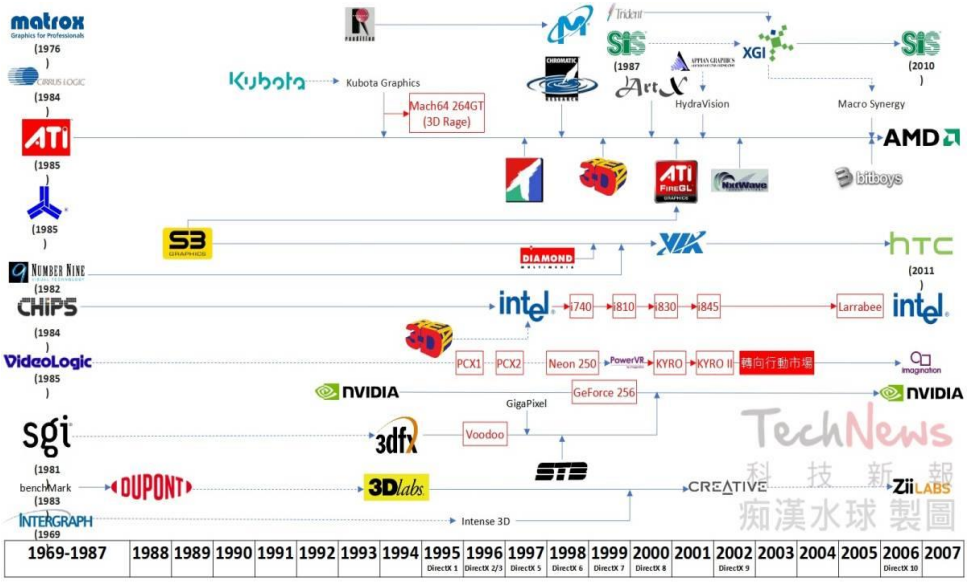
资料来源： Bloomberg、开源证券研究所

2、 发展历程：三十年历经沉浮，终成算力王者

2.1、 1993-2004 年（3D 加速卡时代）：背靠微软掌握标准，显卡龙头地位初显

公司早期聚焦图形芯片，依靠游戏主机厂世嘉赚取第一桶金。1993 年 4 月，从集成电路生产商 LSI Logic 出来的黄仁勋，联合 Sun 公司两位年轻工程师——Chris Malachowsky 和 Curtis Priem 共同创立了英伟达。初期，公司旨在通过生产 3D 图形芯片布局游戏和多媒体市场。彼时 3D 游戏及 3D 渲染仍然处于早期，业内并无统一标准，企业鱼龙混杂，既包括索尼、东芝、IBM 等大厂，也有很多如英伟达一般的创业者，这其中，1994 年成立的 3dfx 凭借 Voodoo 显卡，成为 PC 端 3D 游戏的领袖。1995 年英伟达推出公司首款面向游戏主机的多媒体加速器——NV1，集成了声卡和手柄控制单元。尽管该产品相较 Voodoo 性能不高，兼容性差，但 NV1 仍被运用于世嘉第六代游戏主机“土星”，为公司赚得了第一桶金（游戏机不需要考虑兼容性问题）。

图5：90 年代消费型 3D 显卡市场参与者较多



资料来源： 科技新报

公司濒临破产，绑定微软重获新生。1996 年，微软发布了 Direct 3D 标准（只支持“三角形绘图”），而英伟达因坚持“四边形绘图”的研发路线，NV1 很快便无人问津，同时，为世嘉研发的 NV2 以失败告终，而对手 Voodoo 则顺应规律获得 80%的市场份额，英伟达走到破产边缘。基于此，英伟达做出如下应对：

（1） **人事方面：**任命主机游戏厂商水晶动力的首席技术官 David Kirk 作为英伟达的“首席科学家”；

（2） **研发方面：**确定了为期六个月的内部周期目标，产品更新迭代较快，更快满足下游需求的变化，同时即便某一产品失败，也不会威胁到公司的生存；

（3） **拓客方面：**绑定 PC 大客户微软，1997 年推出全球首款 128bit 的 3D 处理器 RIVA128（NV3），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 3D 加速能力的 2D+3DAGP 显卡，上市四个月出货量突破 100 万片。至 1997 年底，英伟达的 3D 显卡市场份额为 24%，排名第二（仅次于 3Dfx Interactive）。随后，英伟达进一步发布的 RIVA 128ZX 支持 OpenGL，在雷神之锤中表现不错，而雷神之锤不支持 GLIDE 标准，使得 Voodoo 的优势有所弱化。

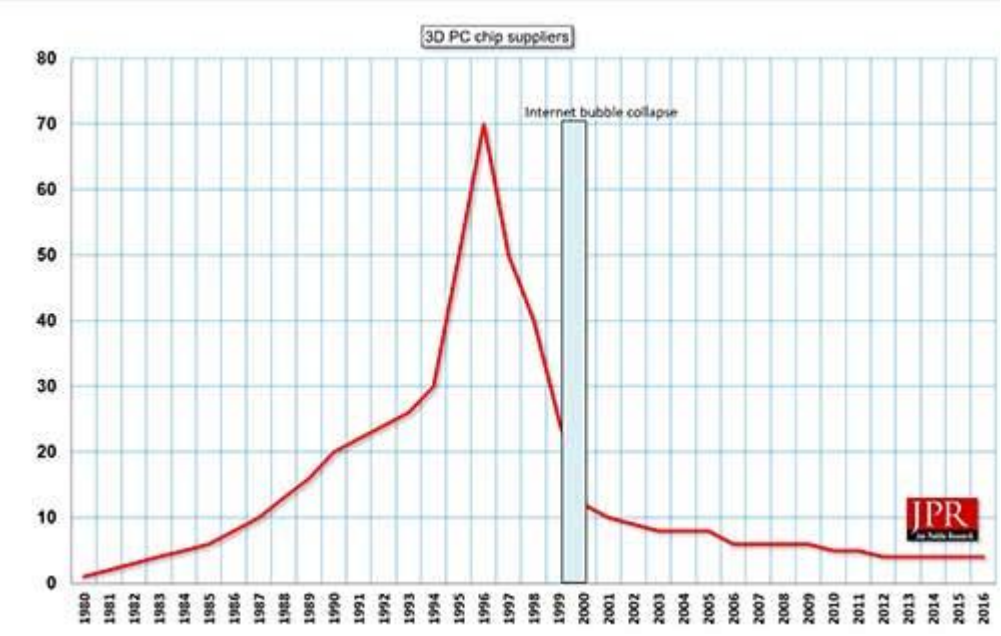
表1：90 年代主要 3D 显卡芯片有着多种显示标准

厂商	NVIDIA	3dfx	ATI	S3
产品	RIVA 128	Voodoo	Rage Pro	Virge GX
显示标准	direct 3D	GLIDE	3DCIF	S3D
发布时间	1997 年 8 月	1996 年 10 月	1997 年 3 月	1996 年 10 月
亮点	尽管在图像质量上不敌 3dfx 的 Voodoo，但优势在于 100M/秒的像素填充率、对 Open GL 的兼容性以及价格较低	早期显卡领袖企业，在 DirectX 崛起前，其招牌 Glide API 受到广泛的软件厂商支持	在 4MB 格式下几乎与 Voodoo Graphics 的性能相匹敌，在 8MB 和 AGP 接口使用时，性能超过了 3Dfx 卡	尽管 3D 速度不尽如人意，但依靠 S3 的品牌声望亦收到一些 S3D 增强游戏

资料来源：维基百科、科技新报、搜狐新闻、开源证券研究所

随着 90 年代计算机的普及和 Windows 的崛起，图形芯片主流市场逐步从主机转向 PC，也使得英伟达在微软的助力下快速起势。1999 年 1 月，英伟达全年营收突破 1.5 亿美元，并在纳斯达克挂牌上市。同年 5 月，其图形处理器销量超过 1000 万。8 月，英伟达推出第一款以 GeForce 命名的显示核心——GeForce 256，并首次提出 GPU 概念，而后戴尔、Gateway、康柏、NEC、IBM 等纷纷宣布预装英伟达的 GPU，与此同时，传统 3D 加速卡市场也进入了快速洗牌阶段，2000 年底英伟达以 7000 万美元现金、100 万股公司股票，将 3Dfx 收入囊中，正式成为行业老大，彼时市场仍具备竞争力的厂商主要为 ATI。在这一过程中，英伟达绑定微软持续推进业务，DirectX 7.0 推出 T&L 技术（极大解放了 CPU 的算力，也是显卡从 3D 处理器转称为 GPU 图形处理器的核心原因）、DirectX 8.0 实现了称为显卡革命的动态观影效果，而 GeForce 亦成为这些 DX（DX 即 DirectX 缩写，下同）系列的代表性显卡。

图6：1996 年起，3D 芯片厂商在经历过蛮荒增长后进入行业洗牌期



数据来源：科技新报

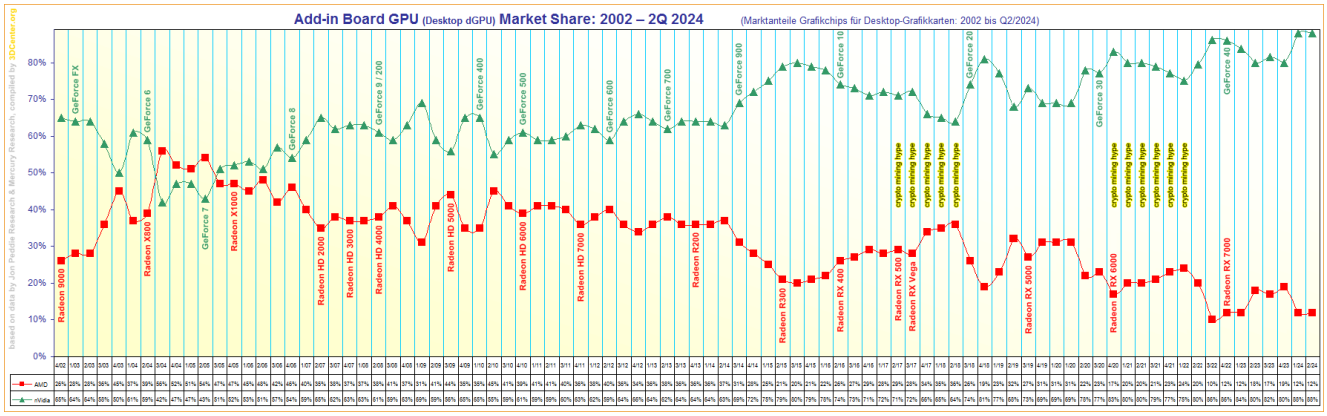
成也微软，败也微软，Xbox 首发失利引发英伟达与微软矛盾。英伟达 GPU 的畅销加速了 DirectX 的普及，微软与英伟达相辅相成，由此微软不仅让英伟达参与到 DirectX 标准的制定中，亦在 2000 年将初代 Xbox 订单交于英伟达，这成为当时英伟达创办以来最大的订单。但由于研发时间短，期间出现电源供应 Bug、数据库功能不足等一系列问题，最终 Xbox 错过先机败给了 PS2。为了与 PS2 竞争，微软计划降低 Xbox 二代产品主机售价，并同时要求英伟达降低芯片价格，但受到黄仁勋拒绝，叠加各种品控问题，最终双方矛盾激化。

微软扶持 ATI，最终带来 N 卡与 A 卡长期拉锯战。GPU 行业更新迭代迅速，上一世代的赢家并不必定能锁定下一时代的胜局，而在 DX9 之前，英伟达产品持续领先 ATI，核心在于跟紧 DX 标准更新，通过抢先发布支持新显示标准的产品来抢占市场。然而，由于英伟达与微软的嫌隙，微软转而重视 ATI 的扶植，使得英伟达错过了微软 DX9 规格确立的重要消息，直接导致当年推出的 GeForce FX 由于兼容性问题败给 ATI 的 Radeon 9700，此后 Intel 也开始扶持 ATI，进一步强化了 ATI 的生命力，尽管之后英伟达与微软达成和解，亦拿下索尼 PS3 的订单，但英伟达龙头地位已经开始动摇，至 2004 年三季度，在独立显卡市场，ATI 市场占有率达到 59%，英伟达只有 37%。

2.2、 2005-2016 年（CUDA 通用计算时代）：打造 CUDA 通用计算体系，埋下时代伏笔

2006 年英伟达推出 CUDA 通用计算平台，为 AI 时代埋下伏笔。2004-2007 年，英伟达业务发展相对平稳，在这其间，AMD 于 2006 年收购 ATI，但整合过程困难，并让 AMD 背上承重的负债，致使 ATI 在与英伟达的竞争中落伍。当此之时，英伟达开始思考更为长远的问题，彼时英特尔的 CPU 可以通过多线程技术被所有计算机应用分享，但 GPU 还只能通过 OpenGL/DirectX 等接口与用户交互，如果能够在 GPU 中提供合适的编程模型，依托 GPU 的并行计算能力，每台 PC 都可以变成一座超大规模高性能计算机。基于此，2006 年，英伟达发布 CUDA 平台，并运用于 2007 年发售的 Tesla 系列，标志着 GPU 不再是图形处理器，而成为通用计算平台。尽管在较长的时间里，CUDA 带来的高投入低回报并未得到市场的充分认可，前谷歌 CEO Eric Schmidt 称“CUDA 不过是 NVIDIA 为推广其 GPU 产品而推出的一项‘多余’的技术”。但随着 AI 时代到来，CUDA 即成为维护英伟达深厚护城河的重要力量。

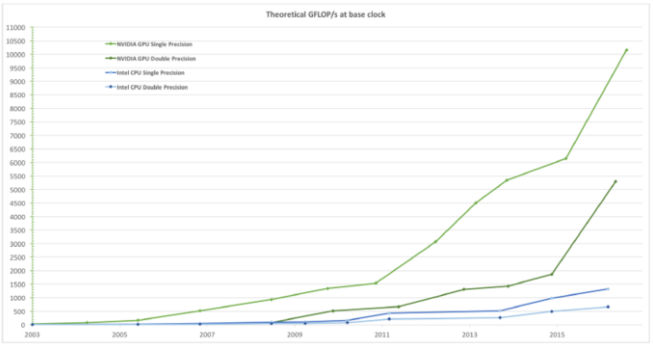
图7：2002-2004 年 ATI 市场份额逐步攀升，短暂超越英伟达后持续向下



资料来源： PConline

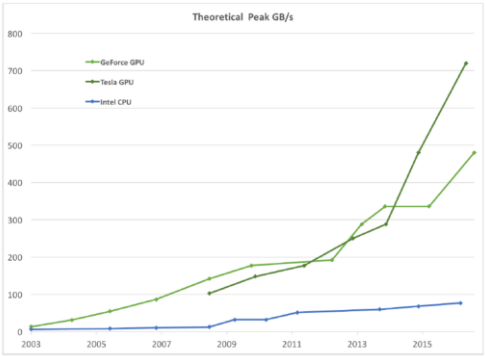
相比 CPU，GPU 拥有更多的数据处理单元、更高的算力与内存带宽，使得其更适合大规模并行运算。从运行效果上看， GPU 体现出远高于 CPU 的运算能力及内存带宽，从运行逻辑上看，CPU 适合复杂、灵活的逻辑运算，GPU 适合简单、大规模的并行运算，在底层硬件上，CPU 的控制单元、缓存单元占有较大比重，而 GPU 则以并行的数据处理单元为主。

图8：英伟达 GPU 浮点运算数远高于 Intel 的 CPU



资料来源：CSDN

图9：英伟达 GPU 的内存带宽远高于 Intel 的 CPU



资料来源：CSDN

图10：CPU 与 GPU 架构对比，GPU 拥有更多的数据处理单元

资料来源：《NVIDIA CUDA Programming Guide（2007）》

英伟达通过 GPU 实现加速计算的核心在于 2 个技术：SIMT（Single-Instruction, Multiple-Thread）和 Hardware Multithreading。

SIMT：即单指令，多线程。所有线程共享同一指令流，这种设计使得 GPU 能够在大量数据上同时进行相同或几乎一致的计算；

Hardware Multithreading：将进程的运行上下文一直保存在硬件上，因而不存在运行上下文切换带来开销的问题（传统 CPU 的多进程是将进程运行上下文保存在内存中，进程切换时涉及到内存的读取，因而开销较大）。

CUDA 体系由 3 部分构成：

1、**指令集架构：**CUDA 定义了一种针对 GPU 特性的指令集，允许程序员直接编写针对 GPU 硬件的代码。这些指令专为大规模并行处理而设计，能够高效地驱动 GPU 上的数千个并行处理单元（如 CUDA 核心或流处理器）同时工作。

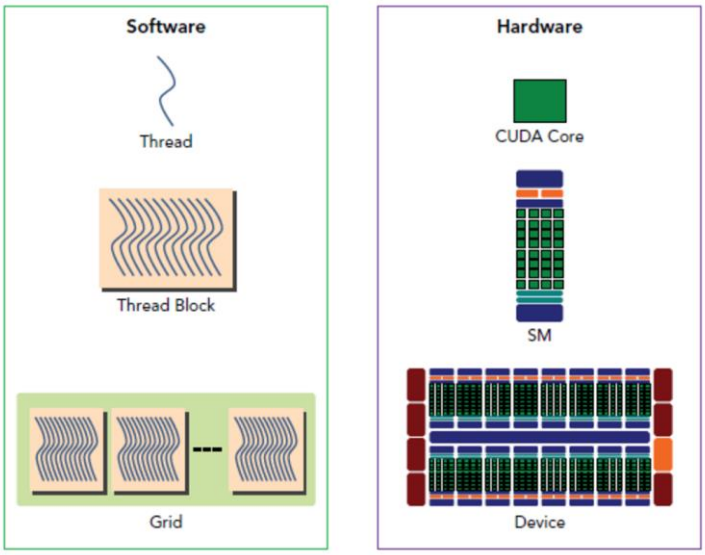
2、**硬件：**即英伟达 GPU 内部的 CUDA Core，这种高度并行的硬件设计使得 GPU 在处理大量数据时能显著提高计算效率，尤其适合于处理诸如矩阵运算、图像处理、物理仿真、机器学习等需要大规模并行计算的任务。

3、**软件：**包括如编程语言与 API、内存模型与管理、并行编程模型、广泛的开发工具链等。

CUDA 硬件和数据架构的对应关系：（1）从硬件的构成关系上，CUDA Core 是英伟达 GPU 最小的计算单元，多个 CUDA Core 叠加 warp scheduler，register，shared memory 等构成一个 SM（streaming multiprocessor），多个 SM 再构成整个 GPU；（2）从数据架构上看，一个 CUDA Core 一次可以执行一个 Thread（线程），数个 Threads 组成一个 Block，同一个 Block 中的 Threads 可以同步，也可以通过 shared memory 通信，最后，多个 Blocks 则会再构成 Grid。此外，英伟达通常将 32 个 Thread 组合

成一个 Warp，作为调度和运行的基本数据单元。

图11：英伟达 CUDA 硬件及数据处理架构有着对应关系



资料来源：CSDN

CUDA 的诞生标志着 GPU 正式从传统的图像处理进阶到通用计算领域，并在如物理仿真、机器学习等需要大规模并行计算的任务中表现出色。CUDA 与英伟达 GPU 强绑定，推出至今已更新至 12.0 版本，在英伟达常年的运营下，拥有极为丰富且成熟的软件生态，使得用户在选择 GPU 时倾向于继续使用英伟达的产品，形成较高的用户粘性和迁移成本，成为英伟达的重要护城河。

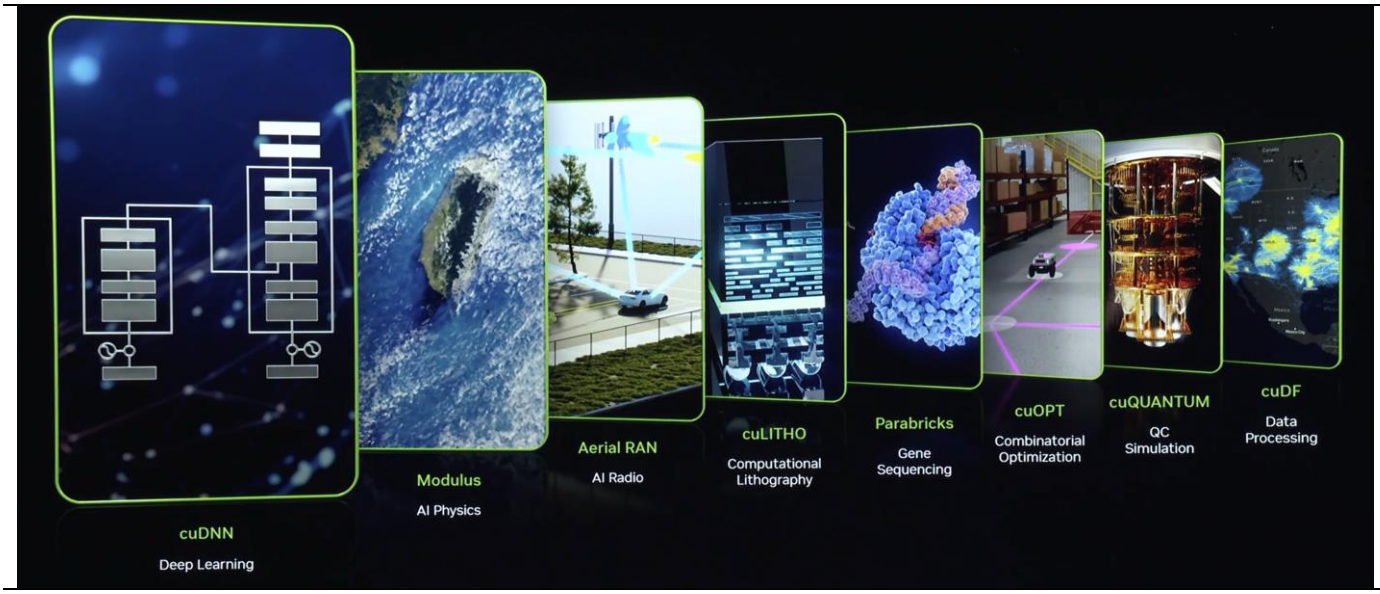
表2：CUDA 按照 1-2 年的频率持续更新

时间	CUDA 版本	更新情况
2006 年	推出 CUDA	这是一个革命性的步骤，因为使得 GPU 不仅仅用于图形处理，还可以用于更广泛的计算任务
2007 年	CUDA 1.0	支持 C 语言编程，为开发者提供了一个新的计算平台
2008 年	CUDA 2.0	引入了对 C++的部分支持，并增加了更多的库和工具
2010 年	CUDA 3.0	引入了 Fermi 架构的 GPU，提供了更好的双精度性能和更多的内存带宽
2011 年	CUDA 4.0	引入了 Kepler 架构的 GPU，提供了更高的能效和更好的并行处理能力
2012 年	CUDA 5.0	引入了 Maxwell 架构的 GPU，进一步提升了能效和性能
2014 年	CUDA 6.0	采用的统一内存方案，支持 Tegra K1 芯片
2015 年	CUDA 7.0	引入新的独立默认流选项，避免旧默认流的序列化，实现异构计算流处理的简化并发
2016 年	CUDA 8.0	引入了 Pascal 架构的 GPU，支持更多的并行计算优化
2017 年	CUDA 9.0	引入了 Volta 架构的 GPU，提供了更好的深度学习性能
2018 年	CUDA 10.0	引入了 Turing 架构的 GPU，提供了更好的图形和计算性能
2020 年	CUDA 11.0	引入了 Ampere 架构的 GPU，提供了更好的 AI 和 HPC（高性能计算）性能
2022 年	CUDA 12.0	引入 Hopper 和 Ada Lovelace 架构的 GPU，正式支持 JIT LTO、改善和引入新的 API 等等

资料来源：英伟达官网、CSDN、IT 之家、至顶网、开源证券研究所

拥有超 400 个 CUDA 函式库，构筑牢固生态壁垒。自 CUDA 诞生以来，英伟达持续在优化及简化 CUDA 的运用市场，并推出超过 400 个函式库，包括专注于处理神经网络的深度学习库 cuDNN、可用于流体动力学等物理定律的 Modulus、专注 5G 无线网络的 Aerial RAN、计算光刻平台 cuLITHO（运用于台积电）等等。CUDA 函式库为细分领域与英伟达架构提供了有效结合，以 cuDNN 为例，因为 CUDA 与 TensorFlow、Pytorch 中的深度学习算法差异较大，CUDA 本身不能被深度学习科学家直接使用，而 cuDNN 为开发者提供了与 GPU 便捷交互的桥梁。如此数百个高性能计算场景的叠加，共同维护了英伟达广泛且丰富的生态护城河，成为英伟达 GPU 在加速计算领域处于垄断地位的核心原因。

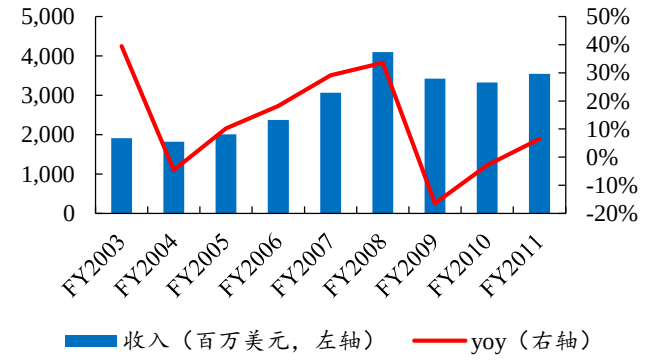
图12：英伟达 CUDA 函式库开拓新市场



资料来源：公司官网

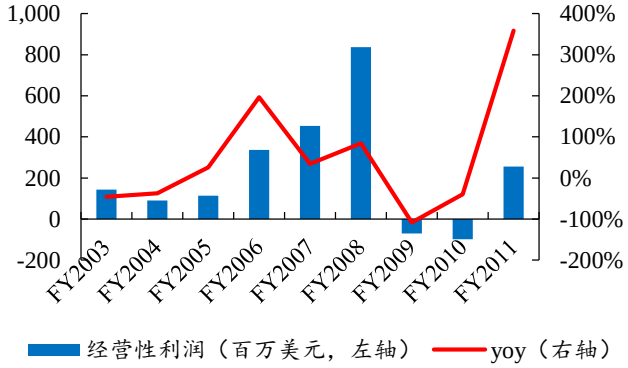
然而 CUDA 在推出早期诟病颇多，核心归结于 2 点：（1）对于 CUDA 的研发每年需花费约 5 亿美元的研发费用，而彼时 GPU 的高性能通用计算或主要用于科学计算中，市场空间有限；（2）CUDA 对散热的更高需求导致了芯片瑕疵，市场推测这或许导致了 2008 年诸多 PC 品牌的屏幕异常问题（显卡门事件）。因此早期资本市场对 CUDA 认可度低，2009-2010 财年在次贷危机下，高研发投入也导致英伟达出现亏损。

图13：FY2009 次贷危机期间英伟达收入出现负增长



资料来源：Bloomberg、开源证券研究所

图14：次贷危机叠加 CUDA 高投入造成英伟达 FY2009-2010 的亏损



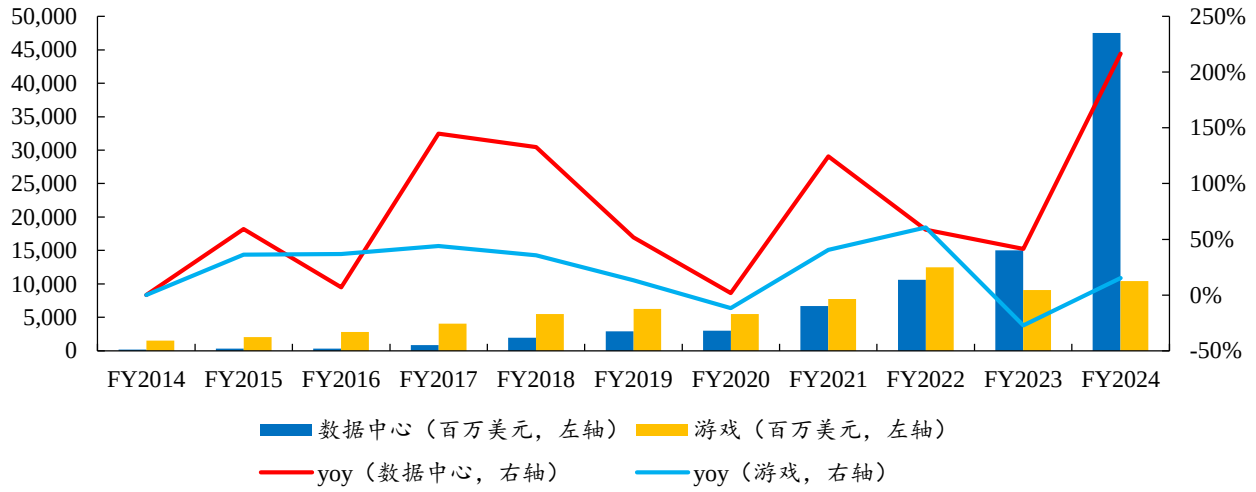
资料来源：Bloomberg、开源证券研究所

AI 驱动初见端倪，CUDA 前期重投入成效初显。转折出现在 2012 年，后来被称为“深度学习之父”的 Jeffery Hinton 教授使用英伟达的 GPU 卡参加全球最为权威的计算机视觉大赛 ImageNet 大赛，其设计的深度卷积神经网络 AlexNet 一举夺冠，成为 AI 历史上的重大突破，也成为英伟达在加速计算上的重要发展方向。2016 年，英伟达发布 Pascal 架构，推出 DGX-1，采用 NVLink 互连架构，首次将 8 个 Tesla P100 GPU 连在一起，并将第一台 DGX 交付给刚成立的 OpenAI。2016 年也成为公司加速计算的财务拐点，FY2017 公司数据中心收入同比增长 145%至 8.3 亿美元，CUDA 前期的重投入初见成效。

2.3、 2017 年-至今（全面 AI 时代）：生成式 AI 崛起，英伟达成为万亿“卖水人”

2017 年，对 AI 行业与英伟达均是具有里程碑式意义的一年。这年 6 月，谷歌大脑团队发表论文《Attention Is All You Need》，提出自注意力模型 Transformer 架构，成为当下生成式 AI 的基石。而早在 1 个月前的 2017 GTC 大会上，英伟达 CEO 黄仁勋开展了围绕 AI 与深度学习的主体演讲，并发布了 Volta V100 与 Tensor Core，标志着英伟达将重点投入 AI 领域，其高性能 GPU 迅速在数据中心取得垄断性地位。与此同时，得益于云计算行业进入成长期、疫情加速线上办公渗透等因素，英伟达数据中心业务保持快速增长。

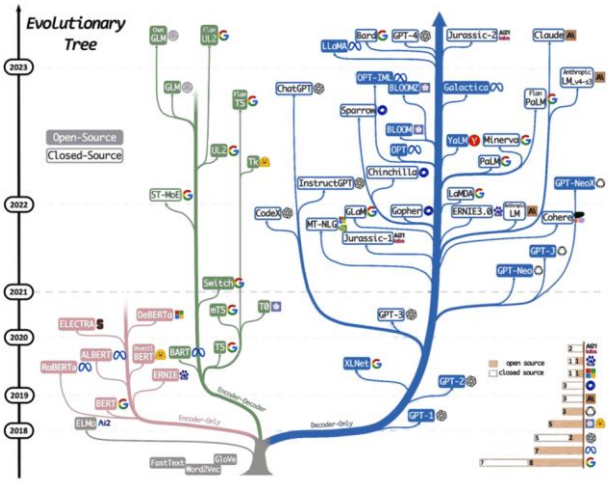
图15：FY2017 年之后，英伟达数据中心收入进入加速成长态势



资料来源：Bloomberg、开源证券研究所

2022 年末 OpenAI 发布 ChatGPT，正式开启生成式 AI 浪潮。ChatGPT 并非最早开始采用 Transformer 的大语言模型，如谷歌早在 2018 年便发布了 BERT，但参数量仅有 1.09 亿个，ChatGPT 的成功得益于千亿级的参数规模，以及其背后使用的 few-shots（小样本）和用户反馈技术，证明了大模型中存在的涌现效应和 scaling law，前者意味着当模型的规模和训练参数达到一定的阈值时，模型的性能和泛化能力会突然出现显著提升；后者即指参数规模越大，模型性能越优秀。此后科技龙头围绕大语言模型 LLM 逐步延伸产品体系，包括文生图、文生视频、多模态等方案陆续推出，英伟达作为核心 GPU 厂商充分受益。在当前市场阶段，训练仍为 GPU 主要运用场景，但随着商业化进程推进，推理占用的工作负载有望从 40%提升至 70%。

图16：大语言模型衍生的产品及技术方向愈发丰富



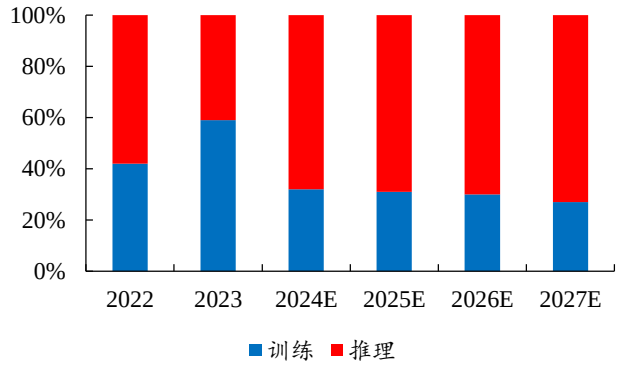
资料来源：腾讯云

图17：提高算力可明显减少大模型训练时长

模型	消耗卡的数量	训练时长
LLaMA - 65B 模型	2048 张 A100 80GB	21 天
GPT3 - 175B 模型	1024 张 A100 40GB	34 天
GLM - 130B 模型	768 张 A100 40GB	60 天
Falcon - 40B 模型	384 张 A100 40GB	60 天
Inflexion - 2 模型	5000 张 H100	-
Megatron - Turing 模型	4480 张 A100	-
GPT4 - 1800B 模型	2.5 万张 A100	90 - 100 天

资料来源：《2024 中国智能算力行业白皮书》、开源证券研究所

图18：AI 服务器推理工作负载占比有望逐步提升

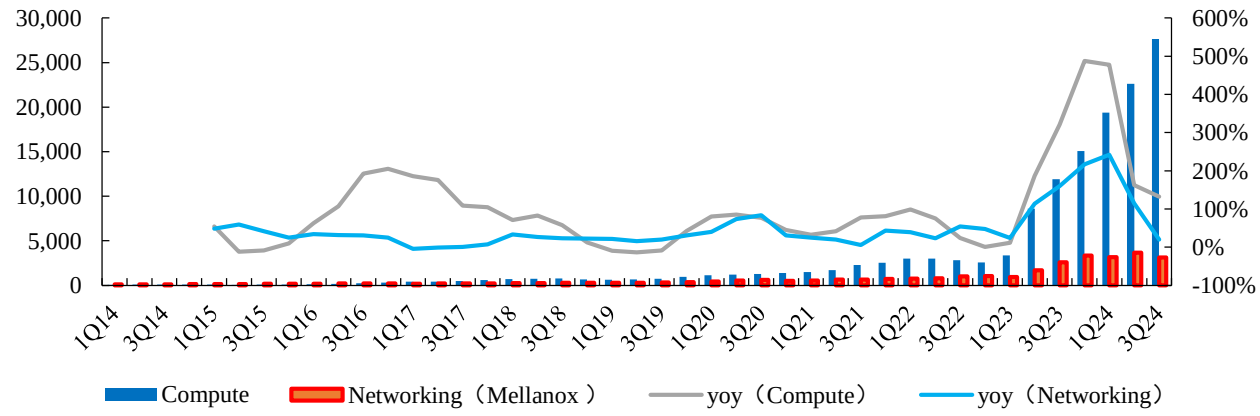


资料来源：《2024 中国智能算力行业白皮书》、开源证券研究所

3、 数据中心：立足 GPU 领先优势，打造“三芯”战略

英伟达提供完善的加速计算解决方案，数据中心成为增长最大驱动力。自 CUDA 诞生以来，从质疑到理解，英伟达数据中心业务已超越游戏业务，成为本轮行情的核心驱动。硬件方面，英伟达实行“GPU+CPU+DPU”三位一体的产品战略，提供基于 CUDA 的 GPU 设备，并可通过组件形式（HGX、DGX、NVL72 等）提供加速计算解决方案；软件方面，英伟达还提供包括丰富的加速软件库、NVIDIA AI Enterprise、DGX 云服务、API、SDK、特定领域应用程序等软件，使得公司数据中心业务成为全栈技术平台。客户包括云厂商（CSP）、消费互联网企业、智算中心、超算中心等部门，2019 年英伟达以 69 亿美元收购 Infiniband 互联技术龙头企业 Mellanox，完善了英伟达在高速互联领域的布局，结合 Mellanox 的优势，NVIDIA 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。

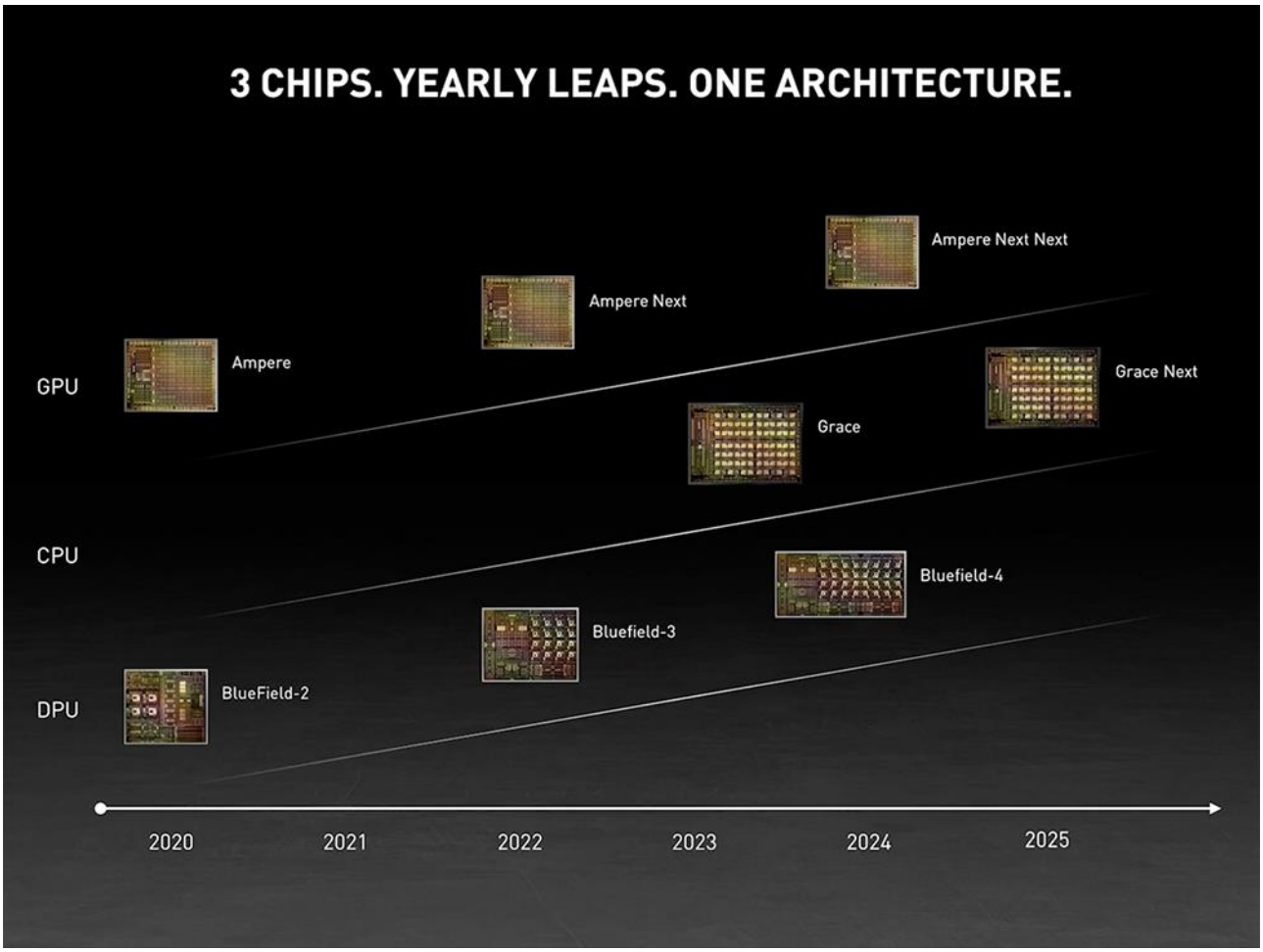
图19：CY3Q23 数据中心业务开始加速启动（营收单位：百万美元）



资料来源：Bloomberg、开源证券研究所，注英伟达收购 Mellanox 并于 CY1Q20 实现并表，图中 CY1Q20 之前 Mellanox 的收入数据并不归于英伟达，仅作为 Mellanox 过往收入增长的呈现。

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

图20：英伟达“CPU+GPU+DPU”三芯战略演变图



资料来源：公司官网

3.1、 GPU： 架构持续迭代，AI 算力的硬通货

英伟达 GPU 架构持续迭代，朝着愈发适宜 AI 计算的方向逐步演进。从 Tesla 到 Blackwell，公司持续迭代 GPU 架构，从工业体系上逐层从 40nm 演进至 4nm，CUDA 核心数也从最初的 128 个增加至上万个，并添加了 Tensor 张量计算核心、NVLink、RTCore、结构稀疏性矩阵 MIG 等功能，数据计算类型逐步丰富，包含了 FP、INT、TF、BF 等数据类型，计算架构逐步朝更适合 AI 运算的方向演进。而在最新的 Blackwell 架构中，GPU 有望达到 20000 TFLOPS FP4 算力，较以往代际的架构有本

质的提升，每 token 的耗能也在持续下降，部分性能是通过降低浮点精度来实现的（从 Pascal 的 FP16 降至 Blackwell 的 FP4），但在数据格式、软件处理和硬件的配合演进下，对 LLM 性能带来的影响并不大。

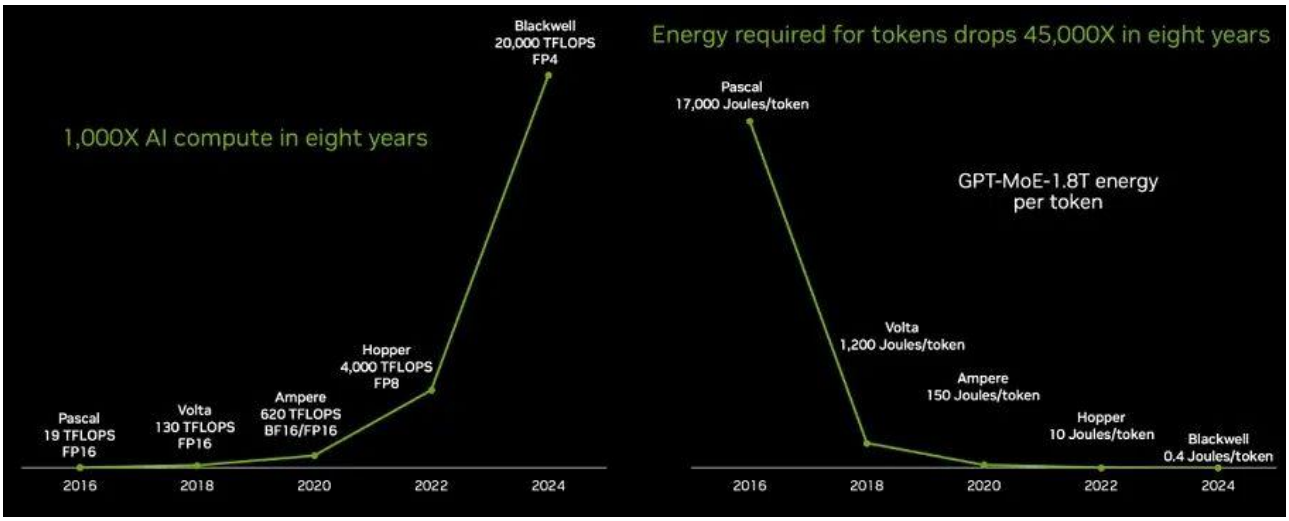
表3：英伟达最新 GPU 架构 Blackwell 可实现 20000 TFLOPS FP4 算力

架构	推出时间	制程	晶体管	核心参数	CUDA 核心数（FP32）	Tensor 核心数	NVLink 版本	特点	备注
Tesla	2006	-	14 亿	-	128	-	-	首个通用 GPU 计算架构，标志 GPU 从专用图形处理器转变为通用数据并行处理器	-
Fermi	2009	40 nm	30 亿	16 个 SM*32 CUDA Core	512	-	-	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	CUDA Core 的数量并不固定，还与具体型号相关
Kepler	2012	28nm	71 亿	15 个 SMX*（192 个单精度+64 个双精度 CUDA Core）	2880	-	-	游戏性能大幅提升，首次支持 GPU Direct 技术，首个支持超级计算机和双精度计算的 GPU 架构	将 SM 改改为了 SMX
Maxwell	2014	28nm	80 亿	16 个 SMM*4 个处理块*（32 个 CUDA Core+8 LD/ST Unit+8 SFU）	3072	-	-	在功耗效率、计算密度上有重大提升，计算密度是 Kepler 的两倍，标志着 GPU 节能计算时代到来	将 SM 改为了 SMM
Pascal	2016	16nm	135 亿	GP100 有 60 个 SM*（64 个 CUDA Core+32 个 DP Core）	3840	-	NVLink1.0	是第一个考虑 DeepLearning 的架构，引入了第一代 NVLink，双向带宽 160GB/s，P100 拥有 56 个 SM HBM	减少了 FP32 但增加了 FP64 的 CUDA Core
Volta	2017	12nm	211 亿	80 个 SM*（32FP64+64Int32 +64FP32+8Tensor Core）	5120	640	NVLink2.0	推出第一代 Tensor Core，支持 AI 运算	将原本的 CUDA Core 拆分为了 FP32 CUDA Core 和 INT32 CUDA Core，这意味着可以同时执行 FP32 和 INT32 的操作
Turing	2018	12nm	146 亿	92 个 SM*（64Int32+64Int32 +64FP32+8Tensor Core）	4608	576	NVLink2.0	新增了 RTCore，Tensor Core2.0	去除了对 FP64 的计算支持，但是增加了对 INT8/INT4/Binary 的支持
Ampere	2020	8nm	540 亿	108 个 SM*（64FP32 +64INT32+32FP64 +4Tensor Core）	6912	432	NVLink3.0	Tensor Core3.0、RT Core2.0、结构稀疏性矩阵 MIG1.0	升级了 Tensor Core，除了在 Volta 的 FP16 以及在 Turing 中的 INT8/INT4/Binary，还新

架构	推出时间	制程	晶体管	核心参数	CUDA 核心数（FP32）	Tensor 核心数	NVLink 版本	特点	备注
									加入了 TF32，BF16，并重新支持了 FP64
Hopper	2022	4nm	800 亿	132 个 SM*（128FP32+64INT32+64FP64+4Tensor Core）	16896	576	NVLink4.0	Tensor Core4.0、结构稀疏性矩阵 MIG1.0	-
Blackewell	2024	4nm	2080 亿	-	-	-	NVLink5.0	多芯片模块（MCM）设计、第二代 Transformer Engine	引入了 FP4 和 FP6 的精度，新的精度计算的引入将进一步减少模型的计算和存储

资料来源：CSDN、稀土掘金、腾讯云、开源证券研究所

图21：英伟达 GPU 架构算力持续提升，耗能逐步下降



资料来源：CSDN

多形态 GPU 组合销售，英伟达更好满足不同客户需求，更好将“三芯”战略与网络技术相结合。英伟达亦通过模组将 GPU、CPU、网络连接技术等组合到一起，形成 AI 计算平台进行销售，代表产品有 HGX 系列、DGX 系列等，不同规格的产品适用于不同客户、不同场景。例如，HGX 仅提供 8 个 GPU 集成的模组，方便 OEM 厂商集成，注重灵活性与定制型，可以根据客户的特定需求来调整和优化系统配置；而 DGX 包含了完整的 GPU、CPU、存储和网络，尤其包含了与英伟达 GPU 适配的 NVLink、以太网/InfiniBand 网络技术，是标准化产品，强调简易性和便捷性，可以快速部署和运行，适合需要即用型解决方案的大型企业。

表4：英伟达可提供多种 GPU 组合形态

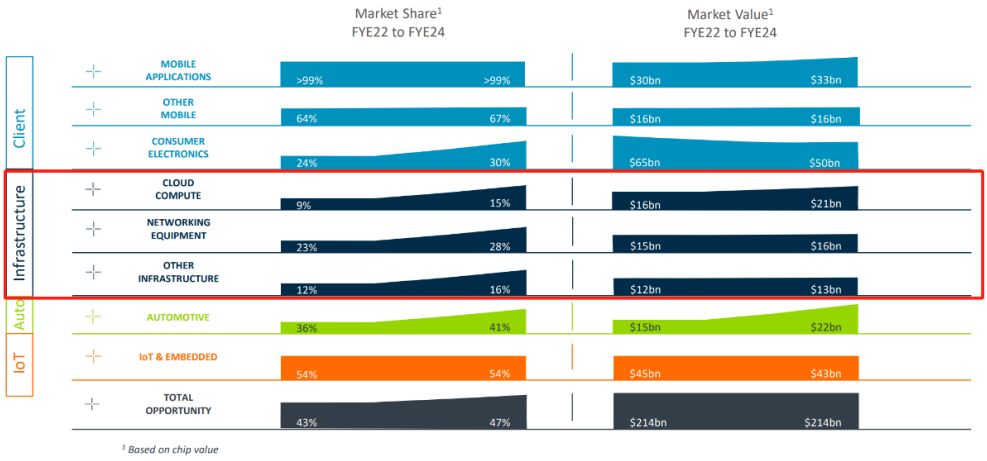
类型	推出时间	运用方向	概述	适用场景
HGX	2017 年	深度学习	代表英伟达 AI 硬件解决方案中的灵活性和定制性，允许客户自由选择和调整 CPU、RAM、存储和网络配置	多场景运用
MGX	2023 年（发布服务器规范）	深度学习	可以针对不同用例打造量身定制的解决方案，支持不同的 GPU、CPU 和 DPU 配置，可作为 HGX 与 DGX 中间的过渡形态	多场景运用
DGX	2016 年	深度学习	英伟达官方整机，除了涵盖 HGX GPU，还有服务器的其他部件，机箱、主板、CPU、内存、硬盘等，不支持定制	多场景运用
EGX	2019 年	边缘计算	将经过加速的 AI 计算从数据中心扩展到边缘，使企业能够在边缘设备上实时进行低延迟的 AI 计算	5G 基站、仓库、零售店和工厂等
IGX	2022 年	工业级边缘计算	英伟达的工业级、边缘 AI 平台，专为工业和医疗环境设计，适用于任务关键型应用	工业、医疗等
OVX	2023 年	数字孪生	用于驱动基于 NVIDIA Omniverse Enterprise 所构建的大型数字孪生，第三代 OVX 基于 4 颗 Ada Lovelace 架构的 L40 GPU	数字工厂、机器人开发、科学模拟等

资料来源：公司官网、TechWeb、CSDN、开源证券研究所

3.2、 CPU： 依托 Arm 实现较强内存一致性，NVLink-C2C 保证芯片高宽带互联

在云计算领域，Arm 市场份额逐步提升。实际上，在以云计算为代表的数据基础设施领域，Arm 的份额正逐步提升，根据 Arm 公司财报，FYE22-FYE24 年（公历年 2021 年 12 月-2024 年 11 月），在云计算领域，Arm 市场份额从 9%提升之 15%，网络设备领域市场份额从 23%提升至 28%，尽管其中或许包含了中国市场为应对 x86 架构供给限制而增加对 Arm 的运用，但英伟达、微软、AWS 等企业相继开发基于 Arm 的 CPU，也表明相比 x86 架构，Arm 在数据中心领域亦有其发展优势。

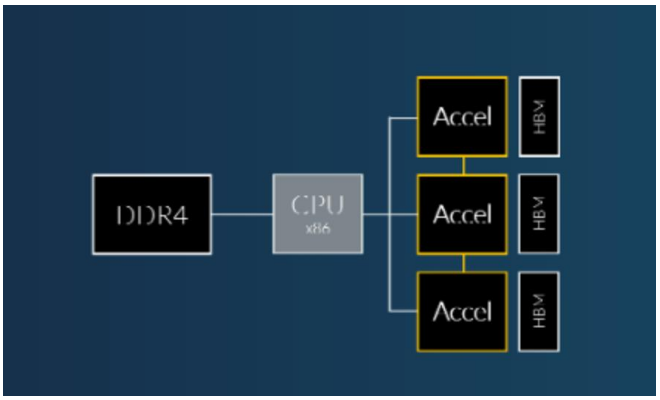
图22：Arm 在数据基础设施领域市场份额持续提升



资料来源：Arm 公司官网

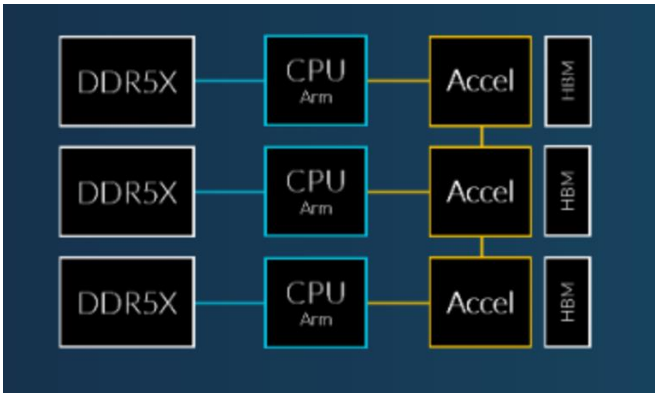
Arm 架构下，CPU 可以实现较强的内存一致性与定制化，更能适应 AI 数据计算。传统的 x86 服务器系统架构，内存通过 PCIe 连接一个通用现成的 CPU，但 CPU 以及加速器之间的接口限制了产品最终的性能水平。因为所有的加速器都必须通过该 CPU 访问额外内存，无法达到内存的一致性。而在 Arm 架构下，每一个 CPU 都单独和一个加速器相连，实现较强的内存一致性，能够更好支持 AI 计算。此外，由于 x86 提供的是标准化芯片，而 Arm 可以根据需求提供定制化 CPU，是 Arm 攫取市场份额的另一重要原因，英伟达能够开发出 Grace CPU 的前提也在于 Arm 的可定制性。

图23：传统 x86 服务器系统架构，加速器需共用一个 CPU 访问内存



资料来源：中微创芯官网

图24：Arm 服务器系统架构，每一个 CPU 都单独和一个加速器相连



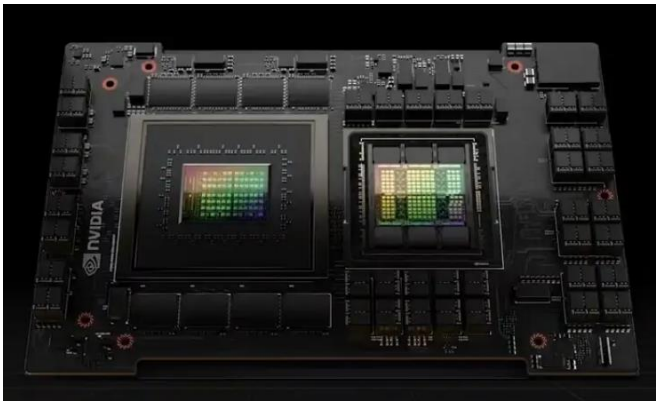
资料来源：中微创芯官网

采用 NVLink-C2C 技术，发布基于 Arm 架构的 Grace 系列 CPU。传统的 CPU 框架难以满足 AI 高性能计算对计算能力和效率的要求，基于此，2021 年英伟达发布数据中心 CPU——Grace，并于 2022 年 3 月在 GTC 大会上正式宣布推出 Grace Hopper 和 Grace CPU 超级芯片，采用 Arm Neoverse V2 核心，具体来讲：

Grace Hopper：以 CPU+GPU 的设计专为应对巨型 AI 和 HPC 挑战，能使用 NVLink-C2C 技术，并且有达到了 900 GB/s 速率的全新一致性接口。

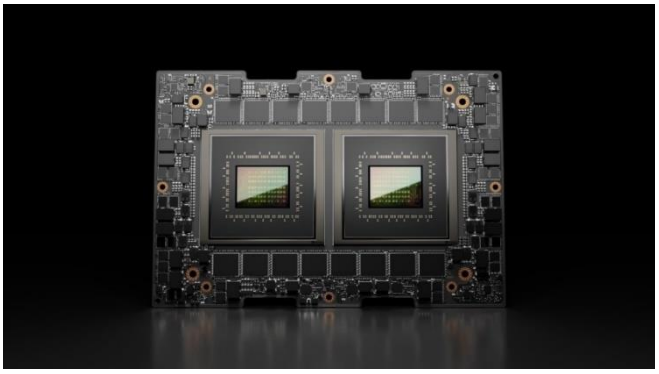
Grace CPU 超级芯片：由两个 CPU 芯片组成，通过 NVLink-C2C 互连技术连接，CPU 内核达到 144 个核心，能对 LPDDR5X ECC 内存进行支持，带宽达到 1TB/s。

图25：英伟达 Grace Hopper 产品形态



资料来源：公司官网

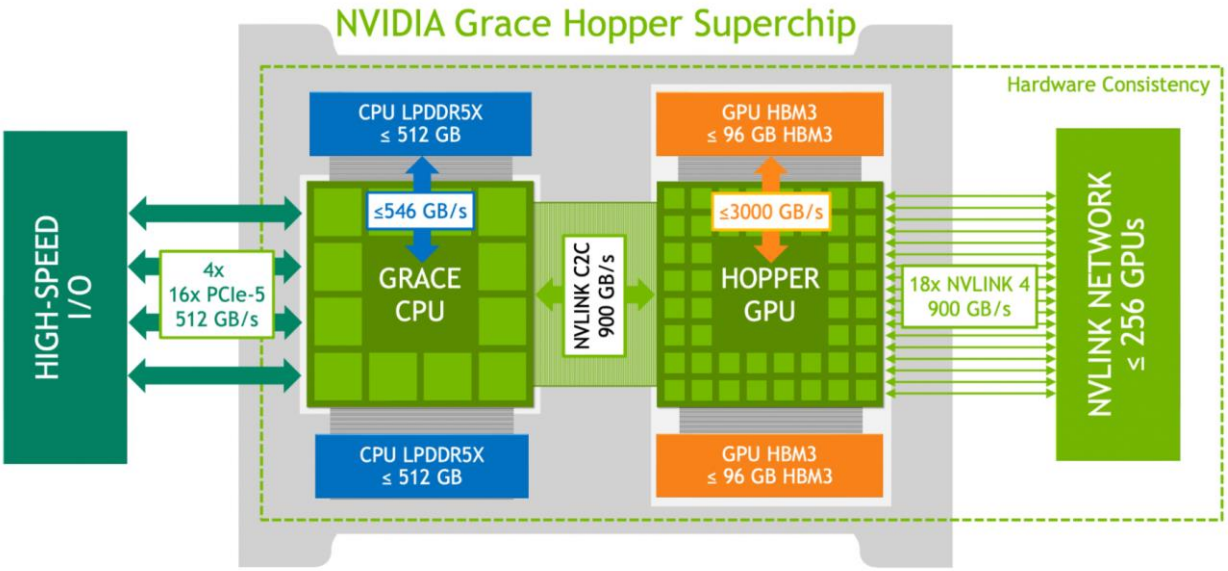
图26：英伟达 Grace CPU 超级芯片产品形态



资料来源：公司官网

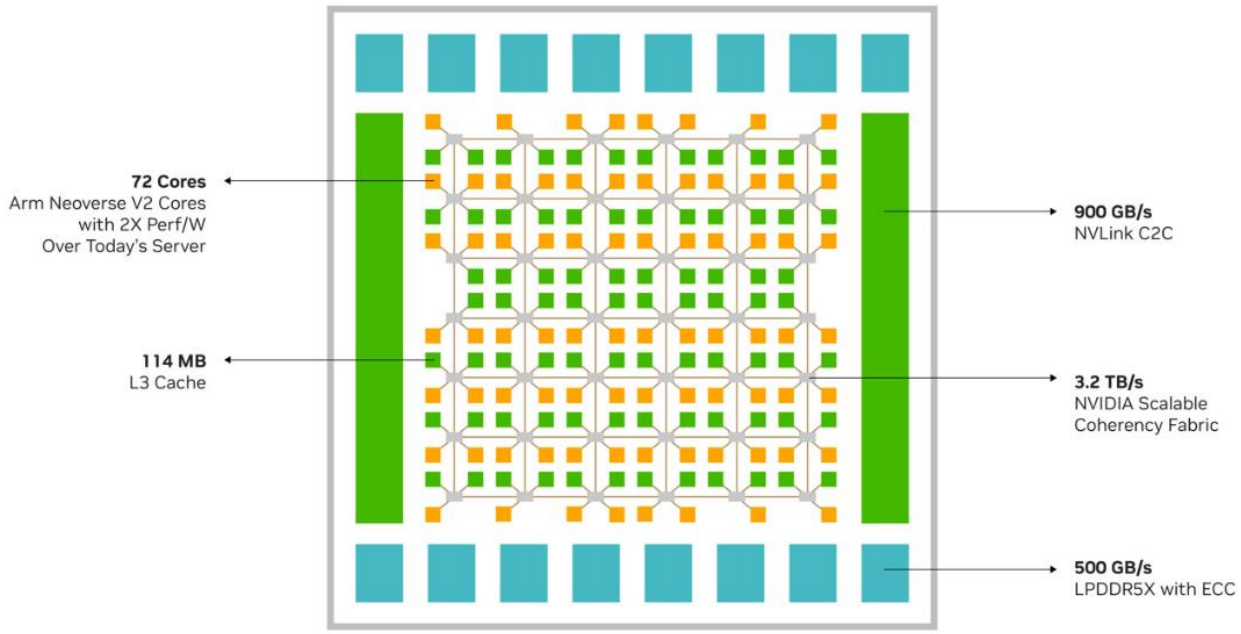
Grace Hopper 超级芯片的核心在于 NVLink-C2C 技术及内存一致性：NVLink-C2C 是一种内存连贯、高带宽和低延迟超级芯片互连，是 Grace Hopper 超级芯片的核心，提供高达 900 GB / s 的总带宽，比通常用于加速系统的 x16 PCIe Gen5 通道带宽高 7 倍。在 Arm 架构下，Grace 可以实现 CPU 核心和缓存的分布式架构，保障了内存一致性及高速的总对分宽带，使得 CPU 和 GPU 线程可以同时透明地访问 CPU 和 GPU 驻留内存，让开发者专注于算法而非显示内存管理。

图27：英伟达 Grace Hopper 超级芯片逻辑概述



资料来源：公司官网

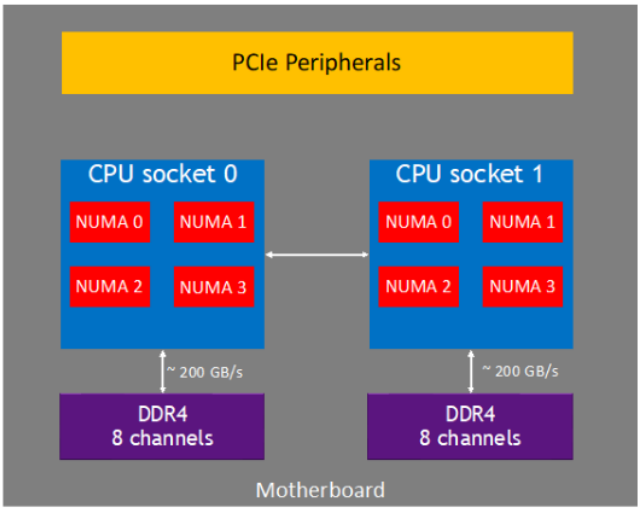
图28：英伟达 Grace CPU 分布式 CPU Core 和缓存



资料来源：公司官网

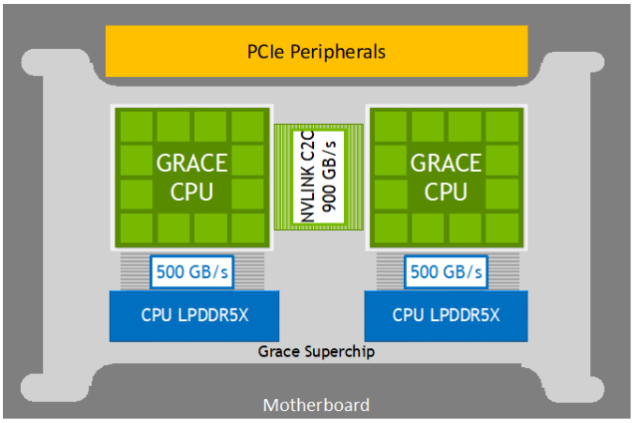
此外，在服务器 CPU 架构上，通常采用 NUMA（非一致性内存访问）来减少内存访问延迟的问题，与传统的多个 NUMA 节点的架构不同，英伟达 Grace CPU 简化为仅有 2 个节点，进一步缓解 NUMA 应用程序开发人员的瓶颈。

图29：传统服务器 CPU 采用多节点的 NUMA 架构



资料来源：公司官网

图30：英伟达 Grace 简化为仅有 2 个 NUMA 节点

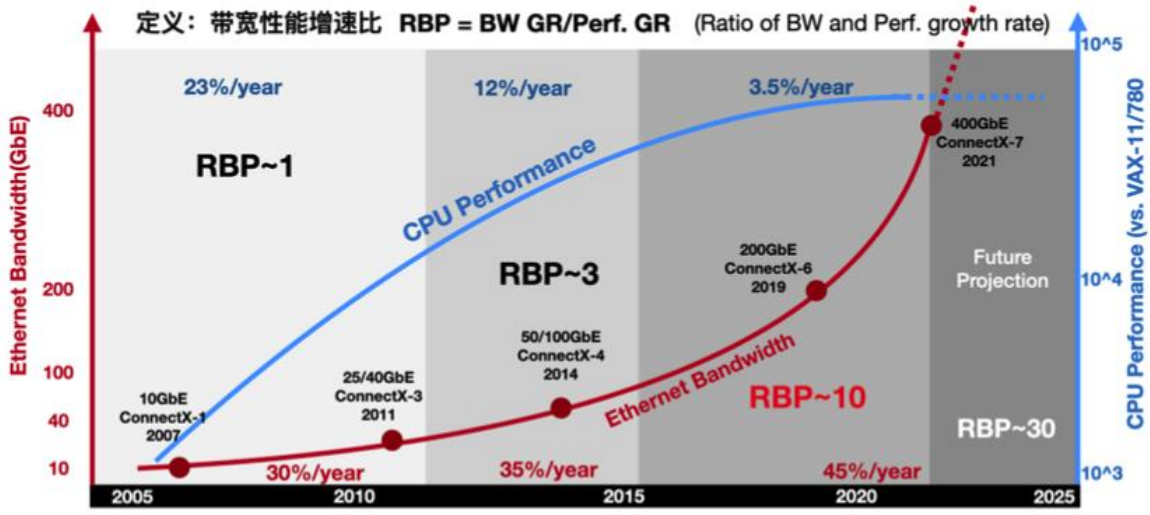


资料来源：公司官网

3.3、 DPU：收购 Mellanox，实现数据摩尔定律

摩尔定律放缓与带宽加速成长的矛盾，催生对高效网络的需求。制程上的摩尔定律逐步失效，但“数据摩尔定律”却持续存在。2010 年前，网络的带宽年化增长大约是 30%， 2015 年增长到 35%，然后在近年达到 45%。相对应的，CPU 的性能增长从 10 年多前的 23%逐步下降到近几年的 3.5%。RBP 指标在 2010-2015 年达到 3 左右，并预计在未来几年达到 30。CPU 算力与网络带宽增速剪刀差持续放大，根据 Fungible 和 AWS 的统计，在大型数据中心，网络流量的处理占到了计算的 30% 左右，也催生了市场对于更优网络解决方案的诉求。

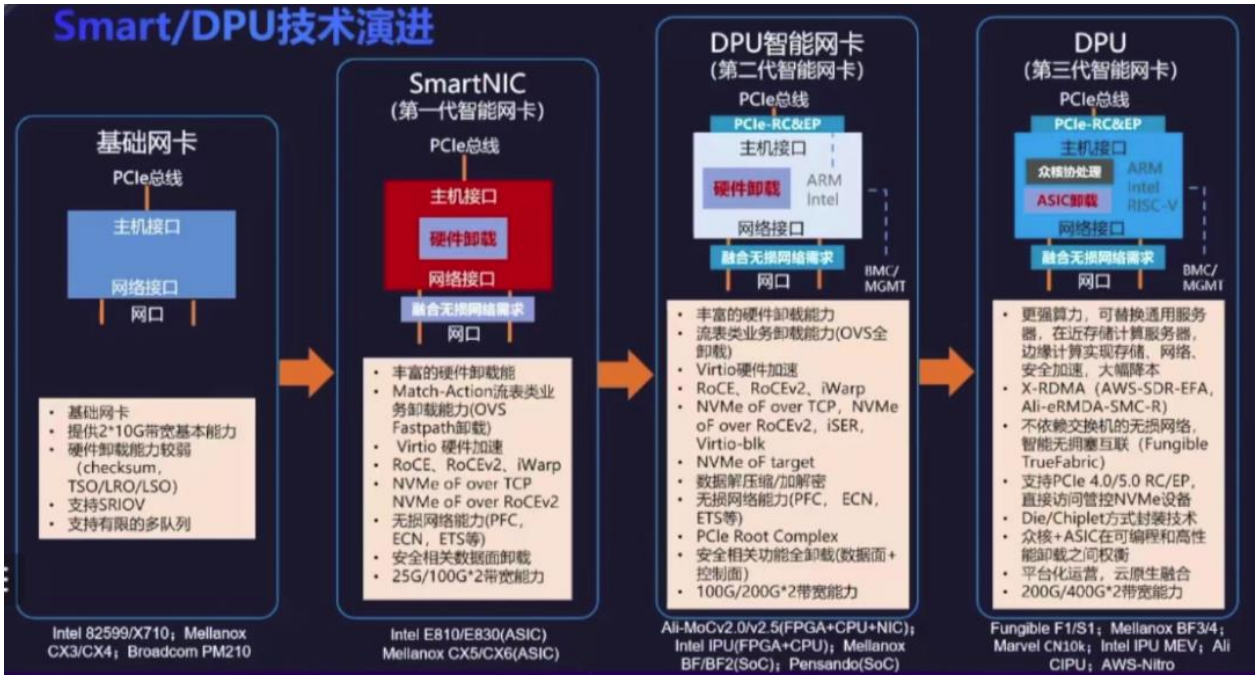
图31：网络宽带增速远高于 CPU 算力增速



资料来源：腾讯云

DPU（数据处理单元）是专门用于处理数据中心网络传输、数据安全和基础设施任务的芯片，旨在减轻 CPU 在数据传输、加密和存储等任务中的负担。DPU 由 NIC（网卡）逐步演进而来，基础的 NIC 是一个 PCIe 设备，它仅实现了与以太网的连接，即实现了网络层次中的 L1-L2 层，此后的智能网卡（SmartNIC）普遍实现了部分 L3-L4 层逻辑的卸载，可处理包括校验和计算、传输层分片重组、云化网络转发功能等工作。而到了 DPU 时代，可进一步实现安全相关功能全卸载、虚拟化、I/O 优化等问题。

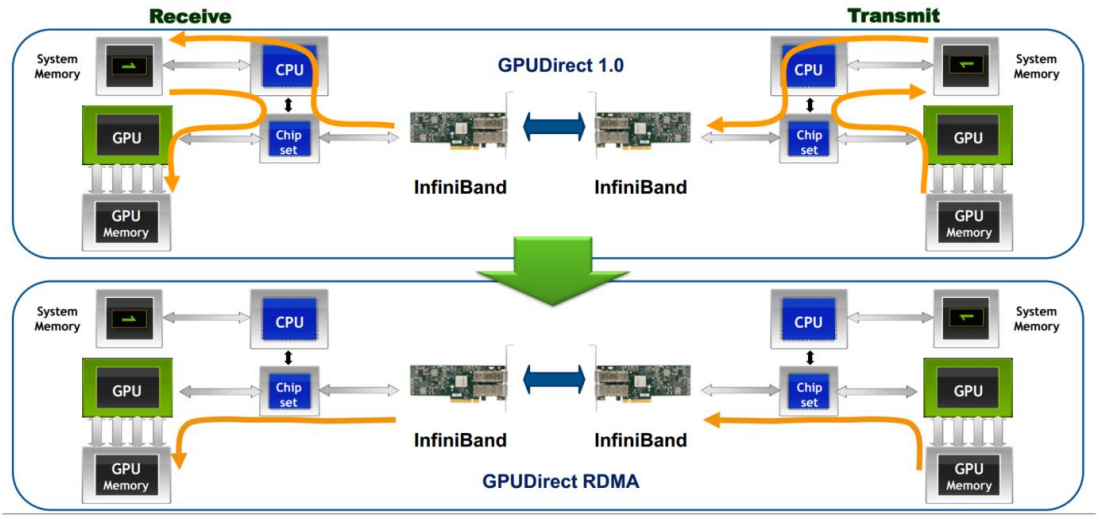
图32：SmartNIC 逐步演化至 DPU



资料来源：腾讯云

英伟达收购 Mellanox，开启 DPU 布局。2019 年英伟达收购 Mellanox，加速了 DPU 技术的落地，并在 2020 年发布了 BlueField 系列的 DPU 产品，落地 GPU-direct RDMA 技术，实现了 GPU 对其他主机 GPU 内存的直接访问。此后，英伟达围绕 DPU 持续完善 BlueField 产品布局，目前英伟达已发布 BlueField-3 DPU 及 SuperNIC，并利用 DOCA 软件开发套件为 BlueField DPU 快速创建应用程序和服务。

图33：GPU Direct RDMA 可以实现高效的远程直接内存访问



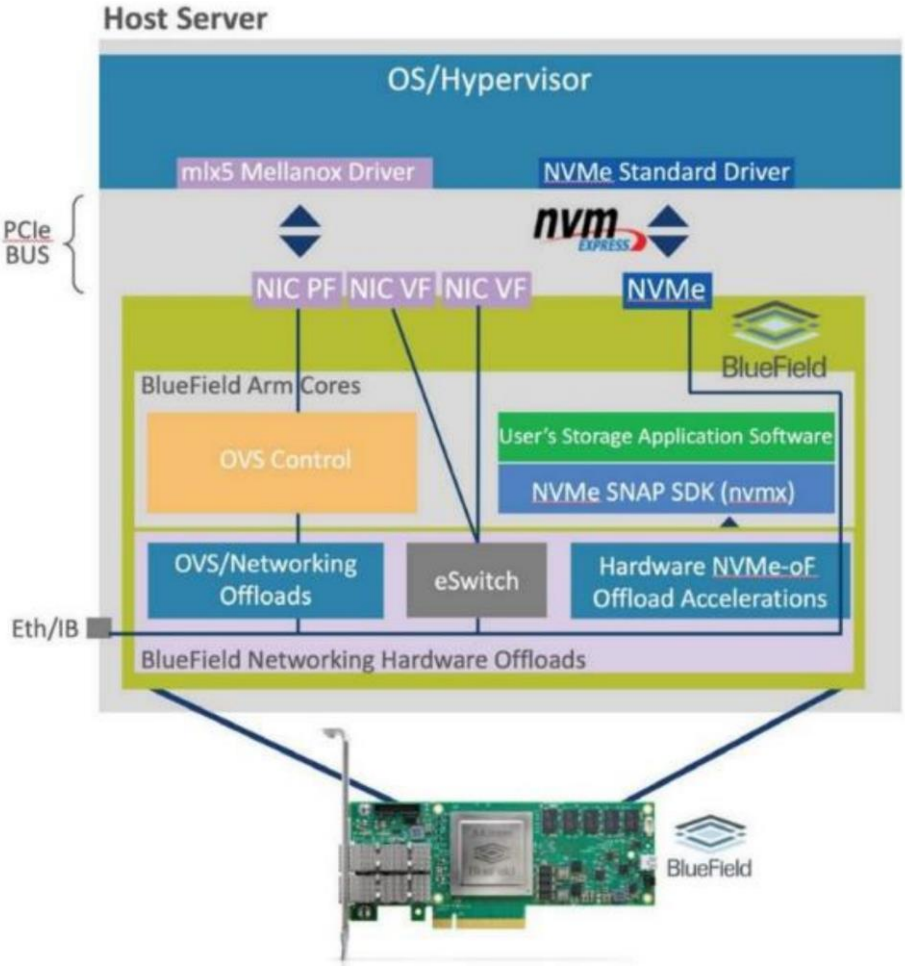
资料来源： CSDN

除了 GPU-Direct DRMA，Mellanox 为英伟达提供了更为关键的两个技术：ASAP2 和 NVMe SNAP 技术。

ASAP2：即加速交换及数据包处理技术，针对服务器虚拟化场景 OVS 存在的 IO 性能不佳、高 CPU 开销的问题，ASAP2 可将虚拟交换数据路径完全的卸载到 NIC 中的嵌入式交换机（eSwitch）中，几乎所有进出服务器的流量都可以由 eSwitch 快速处理，大大释放 CPU 性能；

NVMe SNAP：针对 NVMe 存储虚拟化的加速处理技术。NVMe SNAP 使得远程存储看起来像本地 NVMe SSD，消除了本地存储的低效性，同时满足了对云计算和存储解耦以及可组合性的日益增长的需求。

图34：英伟达 NVMe SNAP 使得远程存储看起来像本地 NVMe SSD



资料来源：极术社区

3.4、 NVLink 技术：实现 GPU 数据直连，NVSwitch 提升 GPU 链路上限

NVLink 是英伟达 GPU 与 GPU、GPU 与 CPU 的高速互连技术。传统的 GPU 通常采用 PCIe 接口与 x86 架构的 CPU 互联，由于记忆系统的差异（GPU 有更快但更小的内存，而 CPU 有较大但较慢的内存），限制了彼此的数据传输能力。2014 年，英伟达联合 IBM 推出 NVLink 高速互联技术，使得 GPU 与 CPU 可以以 5-12 倍的速度分享数据，此外，NVLink 协议在设计时考虑了数据一致性问题，使得不同 GPU 之间的数据访问可以保证一致性。此后英伟达 NVLink 持续迭代，至 NVLink4.0 版本，带宽速度已达到 900GB/s，是 PCIe 5.0 的 5 倍。在 2024 年的 Hotpoint 大会上，英伟达介绍了用于 Blackwell 架构的 NVLink5.0，整体双向带宽将达到 1.8TB/s，是 PCIe 带宽的 14 倍，相较上一代，可以说 NVLink5.0 有着明显的突破。

表5：NVLink 5.0 带宽突破明显

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

资料来源：英伟达官网、今日头条、新浪财经、开源证券研究所

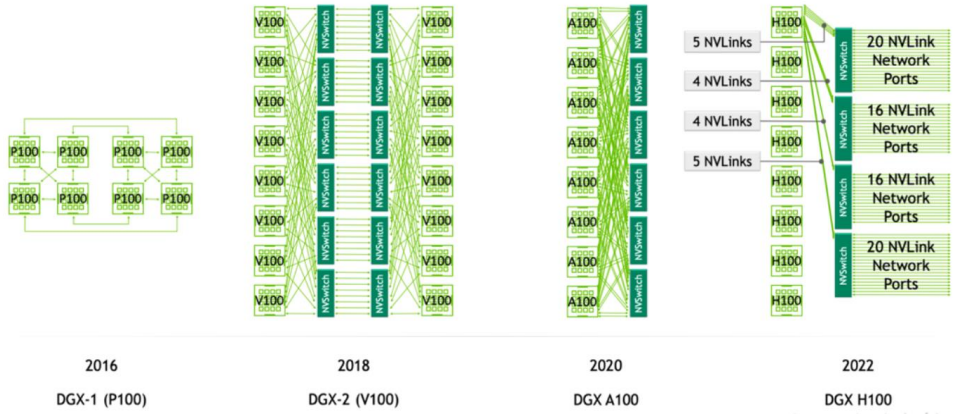
NVSwitch 进一步放大了 NVLink 的优势，带动 NVLink 带宽数倍放大。在 NVLink 协议的基础上，英伟达在 2018 年的 GTC 大会上进一步推出 NVSwitch。在仅有 NVLink 技术的模式下，尽管 GPU 实现了数据的直连，但采用的是点对点的方式，假设在一个 8 卡 H200 的服务器中，该方式下每个 GPU 必须将带宽（900GB/s）拆分为 7 个点对点的专用连接，则每个连接的带宽为 900 /7=128GB/s，而系统的总带宽取决于正在通信的 GPU 数量。NVSwitch 的引入取消了点对点直连的方式，能够将 GPU 带宽持续维持在 900GB/s 的水平。也正是这一技术特征，NVLink 能够持续提升链路数。

表6：以 8 卡 H200 的服务器为例，在 NVSwitch 模式下带宽不受 GPU 数量影响

GPU 数量	点对点带宽	NVSwitch 带宽
2	128GB/s	900GB/s
4	3x129GB/s	900GB/s
8	7x130GB/s	900GB/s

资料来源：英伟达官网、开源证券研究所

图35：NVSwitch 扩大了英伟达 GPU 互联的潜力



资料来源：CSDN

在 Blackwell 架构下，NVLink 域内直连 GPU 数量大幅提升，带动聚合总带宽达到 1PB/s。

表7：NVSwitch 随着架构更新持续升级

NVSwitch 版本	第一代	第二代	第三代	第四代
一个 NVLink 域内直连 GPU 的数量	最多 8 个	最多 8 个	最多 8 个	最多 576 个
NVSwitch GPU 之间带宽	300GB/s	600GB/s	900GB/s	1800GB/s
聚合总带宽	2.4 TB/s	4.8TB/s	7.2TB/s	1PB/s
架构	Volta	Ampere	Hopper	Blackwell

资料来源：英伟达官网、开源证券研究所

3.5、 网络解决平台：充分布局 Infiniband 与以太网，期待 Spectrum 后续突破

10 万卡集群时代到来，网络集群能力愈发重要。随着大模型的深化及对算力的持续追求，10 万卡集群已成为新的追求目标，2024 年 7 月 23 日，马斯克在社交媒体 X 上宣布，xAI 的孟菲斯超级集群拥有 10 万台液冷 H100 GPU，开启了鲶鱼效应，国内头部云计算公司陆续发布 10 万卡集群方案，随后 11 月 Meta 亦称 Llama 4 模型正在 10 万片 H100 的集群上训练。可以预见，10 万卡集群将成为头部大模型难以回避的发展方向，与之相关的网络集群能力也愈发重要。

表8：各企业陆续开展 10 万卡集群计算

时间	企业	集群信息
2024 年 7 月 23 日	xAI	马斯克在社交媒体 X 上宣布，xAI 的孟菲斯超级集群，拥有 10 万台液冷 H100 GPU，在一个单一的 RDMA 架构上运行
2024 年 12 月 5 日	xAI	马斯克计划将 xAI 新建的 Colossus AI 超级计算机系统规模扩增 10 倍、整合百万以上的 GPU
2024 年 7 月 1 日	腾讯云	发布了支持十万卡集群的星脉网络 2.0
2024 年 9 月 25 日	百度智能云	发布了为部署十万卡大规模集群而设计的百舸 4.0
2024 年 9 月 19 日	阿里云	宣布其灵骏单网络集群已拓展至 10 万卡级别
2024 年 11 月 1 日	Meta	Llama 4 模型正在一个由 10 万片 H100 GPU 组成的集群上进行训练，并预计在明年首次推出

资料来源：IT 之家、36 氪、格隆汇、观察者网等、开源证券研究所

Infiniband 在高性能计算领域具备优势，英伟达（Mellanox）处于领导地位。Infiniband 与以太网是数据中心采用的主要网络标准，得益于高传输速率和低延迟的特性，Infiniband 在服务器间的高速通信、存储设备与网络设施之间的高效互联中扮演着至关重要的角色。根据 2022 年 6 月公布的数据，超级计算机 TOP500/TOP100 榜单中，有 38%/59%的系统采用了 InfiniBand 作为关键的互连技术手段，其中英伟达 Mellanox HDR Quantum QM87xx 交换机和 BlueField DPU，在超过三分之二的超级计算机中占据了主导互连的地位，因此在 Infiniband 交换机领域，英伟达已经有明显优势。

表9：在高性能计算领域，Infiniband 较以太网更有优势

属性对比	Infiniband	以太网
推出组织	InfiniBand 贸易协会（IBTA）	施乐公司、英特尔和 DEC
适用场景	用于数据中心内部服务器、通讯基础设施设备、存储解决方案以及嵌入式系统之间互连	用于连接各种局域网（LAN）或广域网（WAN）
网络带宽	发展速度较快，尤其是对高性能计算场景的高度优化和降低 CPU 处理负载的能力上	在高带宽需求层面不如 InfiniBand 迫切
网络延迟	凭借明确的第 1 层至第 4 层协议格式设计以及端到端流控制机制，确保了无损网络通信	缺乏类似的基于调度的流控制机制，依赖于芯片更大的缓存区域临时存储消息，增加了成本和功耗
网络管理	每个第二层网络内部都配备了一个子网管理器，用于配置节点并智能计算转发路径信息，网络架构更为简洁	以太网需要依赖 MAC 地址条目、IP 协议以及 ARP 协议等多个层次实现网络互联，从而增加了网络管理的复杂性

资料来源：电子发烧友官网、开源证券研究所

表10：Infiniband 各版本参数规格

版本	发型时间	Line 编码		每通道传输率 (Gbit/s)	吞吐量 (Gbit/s)				Adapter latency (μs)
					1x	4x	8x	12x	
SDR	2001, 2003	NRZ	8b/10b	2.5	2	8	16	24	5
DDR	2005			5	4	16	32	48	2.5
QDR	2007			10	8	32	64	96	1.3
FDR10	2011		64b/66b	10.3125	10	40	80	120	0.7
FDR	2011			14.0625	13.64	54.54	109.08	163.64	0.7
EDR	2014			25.78125	25	100	200	300	0.5
HDR	2018	PAM4		53.125	50	200	400	600	<0.6
NDR	2022		256b/257b	106.25	100	400	800	1200	
XDR	2024			200	200	800	1600	2400	
GDR	2025（预计）			400	400	1600	3200	4800	

资料来源：维基百科、CSDN、开源证券研究所

表11：PCIe 各版本参数规格

版本	发行时间	Line 编码		每通道传输率（GT/s）	吞吐量			
					x1	x4	x8	x16
PCIe 1.0	2003	NRZ	8b/10b	2.5	250 MB/s	1GB/s	2GB/s	4GB/s
PCIe 2.0	2007			5.0	500 MB/s	2GB/s	4GB/s	8GB/s
PCIe 3.0	2010		128b/130b	8.0	984.6 MB/s	3.938GB/s	7.877GB/s	15.75 GB/s
PCIe 4.0	2017			16.0	1969 MB/s	7.877GB/s	15.754GB/s	31.51 GB/s
PCIe 5.0	2019			32.0	3938 MB/s	15.8GB/s	31.5GB/s	63.02 GB/s
PCIe 6.0	2021	PAM-4 FEC	1b/1b 242B/256B FLIT	64.0	7.563 GB/s	30.25 GB/s	60.5 GB/s	121GB/s
PCIe 7.0	2025（预计）			128.0	15.125 GB/s	60.5GB/s	121 GB/s	242GB/s

资料来源：维基百科、开源证券研究所

为进一步实现超大型数据集的网络效率，英伟达推出 Quantum 及 Spectrum 网络平台。在英伟达长远的愿景中，数据中心将取代单个芯片，成为计算系统的基本单元，因此除了 DPU、NVLink，整体网络加速以及实现万卡甚至十万卡集群的能力亦是发展重点。2024 年 3 月，英伟达推出 Quantum-X800 InfiniBand 和 Spectrum-X800 以太网平台，是全球首款能够实现端到端 800Gb/s 吞吐量的网络平台，被 Microsoft Azure 和 Oracle Cloud 采用。从运用场景上看，Quantum 得益于 Infiniband 的高吞吐、低延时，可用于对大模型训练有极致需求的场景（AI 工厂），而 Spectrum 可用于追

求性价比、与以太网兼容的场景（AI 云）。此外，全球首个 10 万卡集群的 xAI 亦采用了英伟达的 Spectrum-X 以太网平台。

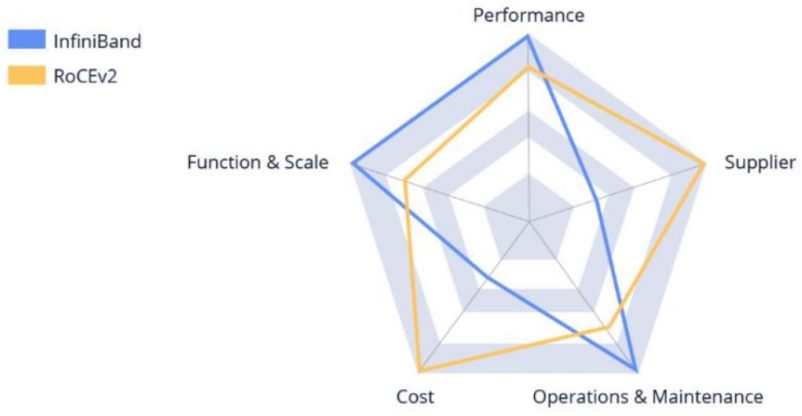
表12：英伟达充分布局 IB 及以太网领域

网络平台	Quantum-X800 平台	Spectrum-X800 平台
网络属性	Infiniband	以太网
运用场景	AI 工厂	AI 云
交换机	NVIDIA Quantum Q3400	Spectrum SN5600 800Gb/s
智能网卡	NVIDIA ConnectX ® -8 SuperNIC™	NVIDIA BlueField ® -3 SuperNIC
特点	实现了业界领先的 800Gb/s 端到端吞吐量，借助 SHARPv4，宽带容量提高了 5 倍，网络内计算能力提高了 9 倍，达到 14.4Tflops	提供了对多租户生成 AI 云和大型企业至关重要的高级功能集

资料来源：英伟达官网、开源证券研究所

随着推理场景占比加重，Spectrum 以太网解决方案或愈发重要。尽管 Infiniband 在高宽带、低延迟上具备优势，但以太网与 PCIe 持续更新，与 Infiniband 并未拉开较大差距，因此从性价比以及英伟达一家独大的规避上，以太网解决方案的生态愈发具备生命力。2023 年 7 月，AMD、微软等 9 家硅谷大厂联手成立了超以太网联盟（UEC），对以太网进行了三项重要改进（数据包喷洒、访问灵活排序、网络拥塞管理），以强化与 Infiniband 的竞争；2024 年根据《The Information》报道，微软和 OpenAI 正在共建一个大型数据中心“星际之门”（Stargate），在网络基础设施方面倾向于使用开放以太网协议而非 InfiniBand。此外，随着推理场景的计算逐步起量，出于对性价比、端侧计算、兼容性等方面考虑，以太网网络方案也逐渐成为大模型厂商的考虑方向，英伟达的 Spectrum 业务也将愈发重要。

图36：相比 Infiniband，RoCEv2 性价比更高



资料来源：CSDN

主流企业以太网交换芯片企业主要企业以太网方案各有侧重，看好英伟达 Gen-AI 网络开发能力 。当前全球已发布 51.2Tbps 以太网交换芯片的共有 Broadcom、Marvell、NVIDIA、 Cisco 与华为五家，其中华为与 Cisco 主要以自用为主。头部企业所推出的交换机产品基本都能提供拥塞管理、数据包喷射、链路故障转移等核心功能，不同企业着重点略有不同，如英伟达强调与 AI 推训的适配、博通强调功耗、

Marvell 强调低延迟、Cisco 强调高 SerDes 配置基数。然而随着技术更新，企业彼时的优势也很快被对手赶超，如当前主要企业均实现 512x112 Gbit/s 的 SerDes 带宽，Cisco Silicon ONE G200 的优势相对弱化。而就英伟达而言，尽管当前公司 SerDes 带宽较竞品略低，但我们认为其优势在于 GPU 端到端整体优化能力，基于 NCCL 无缝支持 RDMA 接口，可大大降低 AI 应用从 TCP 转向 RDMA 框架的开发难度。目前英伟达 Spectrum-X 方案已经落地 xAI 的 10 万卡计算机集群，2025 年公司或将进一步推出 Spectrum Ultra X800，英伟达有望在以太网网络成功卡位，进一步放大自身优势。

表13：51.2Tbps 主要企业的以太网交换机方案各有侧重

	NVIDIA	Broadcom	Marvell	Cisco
芯片推出时间	2024 年 3 月	2022 年 8 月	2023 年 3 月	2023 年 6 月
交换机	Spectrum-X SN5600	Tomahawk 5 Bailly	/	Cisco 8501
交换机芯片	NVIDIA Spectrum ASIC	Tomahawk 5	Teralynx 10	Silicon ONE G200
单端口速率	400G	400G	400G	400G
端口数量	128	128	128	128
制程	4nm	5nm	5nm	5nm
功耗（每 100Gbps）	/	<1W	1W	/
SerDes 带宽	512x100 Gb/s	512x112 Gbit/s	512x112 Gbit/s	512x112 Gbit/s
亮点	支持拥塞控制和 RDMA 技术，支持 AI Cloud，支持推训一体	Tomahawk 5 Bailly 为业界首款每秒 51.2 Tbps 的联合封装光学（CPO）以太网交换机，光学互连的功耗降低 70%，硅面积效率提高了 8 倍	延迟低至 500 纳秒，所有数据包大小的延迟均低于 600 纳秒	与对手的关键区别在于支持 512x 基数配置，可提供更小、更紧密的交换结构，使用的交换机数量减少了 40%
运用企业	xAI	字节跳动、Meta	亚马逊（潜在客户）	Meta

资料来源：各公司官网、凤凰网、CSDN 等、开源证券研究所

4、 游戏&专业可视化：公司传统优势业务，推陈出新挖掘增量

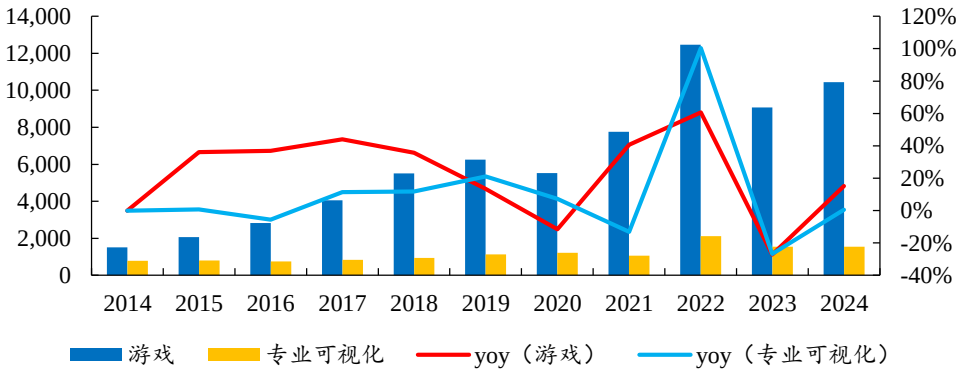
游戏与专业可视化是英伟达 GPU 作为图形处理器的重要方向，也是公司的传统优势业务，持续处于行业垄断地位：

1、 **游戏：**1999 年，英伟达推出 GeForce 系列，首次定义 GPU，2018 年发布 GeForce 20 系列，通过搭载 RT Core 实现了实时光追，同时 Turing 架构的 Tensor Core 可实现 DLSS 技术，进一步放大光追效果。经过 20 余年迭代，GeForce 系列已更新至 GeForce 40 系列（2022 年 9 月发布），采用 Ada Lovelace 微架构，支持第三代光追功能，GeForce 50 系列有望在 2025 年发布，根据往年数据，有望带动销售增长。英伟达提供的软硬件产品和服务包括：（1）用于桌面端的 GTX 和 RTX 系列 GPU。（2）用于移动端笔电 GTX 和 RTX 系列 GPU。（3）用于显示器的 G-SYNC 处理器。（4）Geforce Now 云游戏平台。

2、 **专业可视化：**专业显卡是图形工作站的主要组成部分，与消费类显卡相比，3D 专业显卡主要面对的是 3D 动画（如 3DS Max、Maya、Softimage|3D）、

渲染（如 LightScape、3DS VIZ）、CAD（如 AutoCAD、Pro/Engineer、Unigraphics、SolidWorks）、模型设计（如 Rhino）以及部分科学应用等专业 OpenGL 应用市场。工作站对显卡的速度、稳定性尤其是软件的兼容性要求更高。目前全球主要的工作站显卡厂商是英伟达和 AMD，虽然专业显卡和消费显卡在终端要求有着明显的不同，但是近年来英伟达和 AMD 都逐渐将旗下娱乐级显卡和专业级显卡统一到相同的核心架构下，甚至是完全相同的芯片，由外围电路和软件控制决定是消费类显卡还是专业类显卡。

图37：英伟达游戏与专业可视化业务（百万美元）增速基本趋同

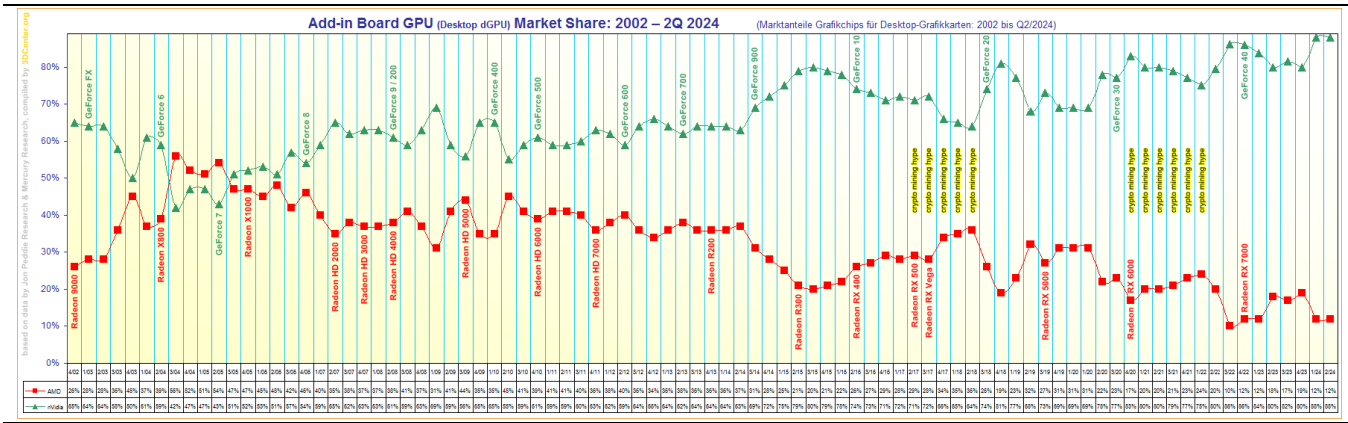


资料来源：Wind、开源证券研究所

4.1、 游戏：龙头地位稳固，关注 AI PC 驱动机会

GeForce 市场份额领先，主打高端市场。早期因英伟达与微软矛盾激化、英特尔扶持 ATI 等因素，Radeon 系列在 2004 年市场份额曾短暂超越英伟达，而随着英伟达与微软和解、拿下索尼订单，业务恢复正常化，重回领先地位，但 2005-2013 年英伟达与 AMD（2006 年收购 ATI）整体上处于来回拉锯的阶段。后续因 AMD 对 ATI 收购的整合效果较差，负债提升、逐步对 GPU 部门造成拖累，彼时 AMD 的产品在内存、带宽等性能上可以短暂性优于英伟达，但能耗表现却远不如同期英伟达的 Maxwell 架构。2014 年后，二者份额差距持续拉大，目前英伟达 GeForce 系列主打高端市场，而 AMD 主要聚焦中低端市场。

图38：英伟达 GeForce 系列显卡龙头地位稳固



资料来源： PConline

英伟达 GeForce 旗舰产品性能优于竞品，主打中高端市场。对比当下英伟达（GeForce RTX 4090）及 AMD（Radeon RX 7900 XTX）的旗舰产品，英伟达在核心性能参数上明显优于 AMD，由此，在售价上英伟达聚焦中高端，AMD 主打中低端，英伟达售价高出 AMD 60%。此外，由于 AMD 的显卡没有 Tensor Core，因而无法实现 DLSS（深度学习超级采样）功能，AMD 主要通过 FSR（FidelityFX 超级分辨率）来升级图像，但画质较英伟达 DLSS 仍有差距。

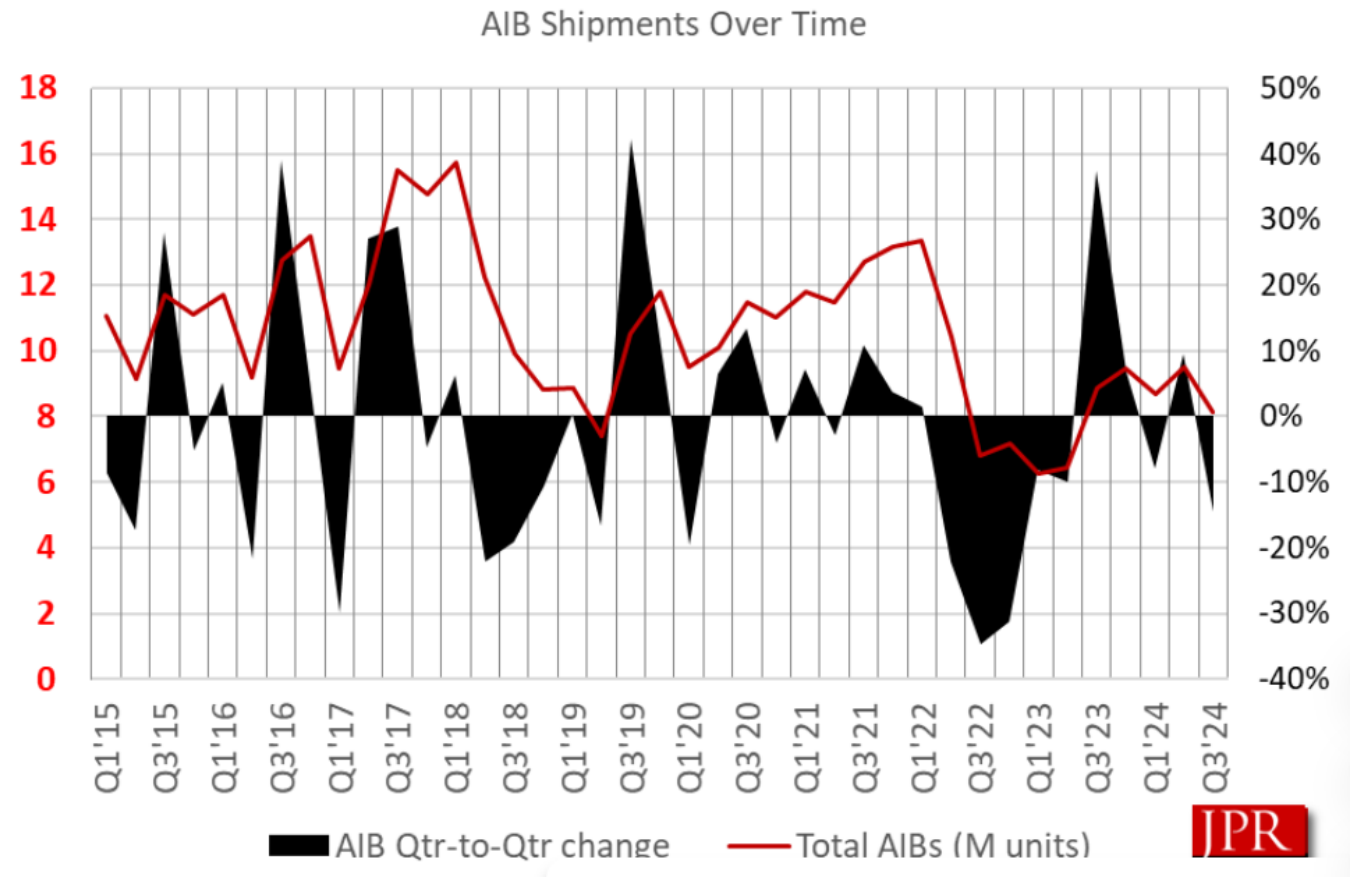
表14：英伟达聚焦中高端，AMD 主打中低端

	英伟达	AMD		英伟达	AMD
基础信息			渲染规格		
产品	GeForce RTX 4090	Radeon RX 7900 XTX	流处理器数量	16384	6144
发布时间	2022 年 9 月	2022 年 11 月	纹理单元	512	384
发行价	1599 美元	999 美元	光栅单元	176	192
架构	Ada Lovelace	Navi lll/RDNA 3	张量核心	512	/
工艺制程	台积电 4N 定制工艺	台积电 N5+N6 混合工艺	光线追踪核心	128	96
晶体管数量	763 亿	577 亿	一级缓存	128KB/SM	256KB/Array
尺寸	609 mm²	529 mm²	二级缓存	72MB	6MB
晶体管密度	1.25 亿/mm²	1.09 亿/mm²			
时钟速度			理论性能		
基础频率	2235MHz	1929MHz	像素填充率	443.5 Gpixel/s	479.6 Gpixel/s
加速频率	2520MHz	2498MHz	纹理填充率	1290 Gpixel/s	479.7 Gpixel/s
显存频率	1313MHz	2500MHz	FP16 性能	82.58 TFLOPS	122.8 TFLOPS
显存			FP32 性能	82.58 TFLOPS	61.39 TFLOPS
显存类型	GDDR6X	GDDR6	FP64 性能	1.29 TFLOPS	1.918 TFLOPS
显存容量	24GB	24GB			
显存位宽	384bit	384bit			
显存带宽	1008GB/s	960GB/s			

资料来源：techpowerup.com、游侠网、开源证券研究所

英伟达显卡市场份额持续提升，行业或面临衰退风险。2022-23 年因为疫情、加密市场退潮，导致 GPU 需求减弱，行业进入一段时期的库存消化中，并于 2023 年下半年开始逐步修复，根据 JPR 数据，3Q24 全球 AIB 显卡市场出货量 810 万片，同比下降 7.9%，英伟达/AMD 在 AIB 显卡市场份额为 90%/10%，英伟达市场份额同比提升 8 pcts（与之对应的是 AMD 市场份额的下降），或因为 AMD 主要主机客户（微软、索尼）调整库存导致半定制收入下降。展望未来，根据 JPR 预测，美国关税政策或将大幅提升终端用户价格，进而抑制消费，预计 2024-2028 年 AIB 显卡出货量 CAGR 为-6%。

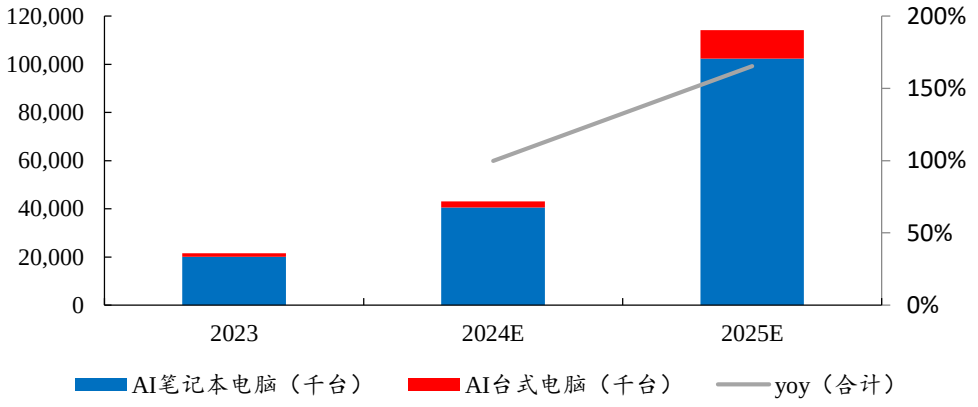
图39：AIB（附加板）显卡全球出货量或进入下行通道



资料来源： JPR

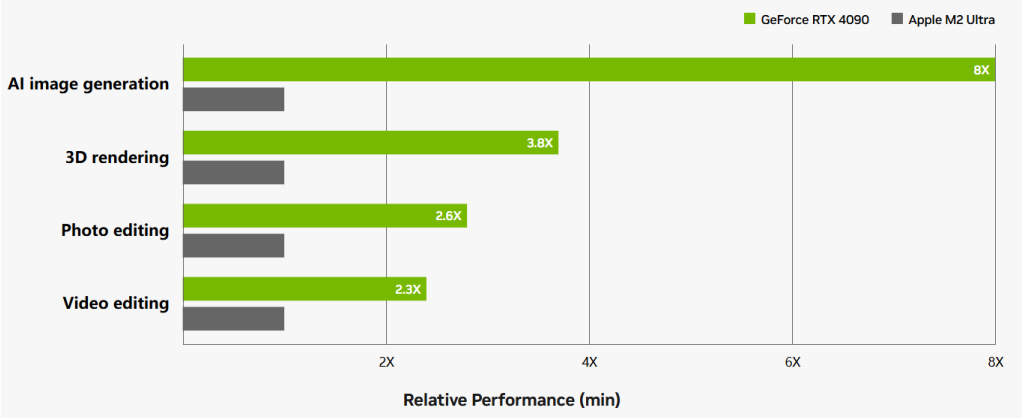
以 AI PC 主导的换机潮获将进入加速阶段，英伟达显卡有望从中受益。尽管行业景气度有待改善，但英伟达 GeForce 持续更新版本，2025 年 RTX50 系列发布，性能进一步提升；另一方面，我们认为本轮 AI PC 替换浪潮有望为英伟达显卡提供增长机遇。根据 Gartner 预测，2024/25 年 AI PC 出货量预计达到 4303/11422 万台，同比增长 100%/165%，2025 年 AI PC 出货量在 PC 中占比将从 2024 年的 17%增长至 43%，2024 年高通 Snapdragon X 系列、AMD Ryzen AI 300 系列、英特尔 Lunar Lake 系列相继发布，为 Copilot+ PC 做好铺垫。落脚到英伟达，基于公司在 AI 领域的积淀，有望联合 PC 厂商推出基于 AI PC 的显卡产品，根据英伟达 FY2025Q3 业绩交流，公司已开始出货华硕和 MSI 的新款 GeForce RTX AI PC，最高配备 321 AI TOPS，利用 RTX 光线追踪和 AI 技术的力量来增强游戏、照片和视频编辑、图像生成和编码。

图40：AI PC 出货量增长较快



资料来源：Gartner（2024 年 9 月预测）、开源证券研究所

图41：英伟达 GeForce RTX 4090 相对 Apple M2 Ultra 在内容创造上性能领先

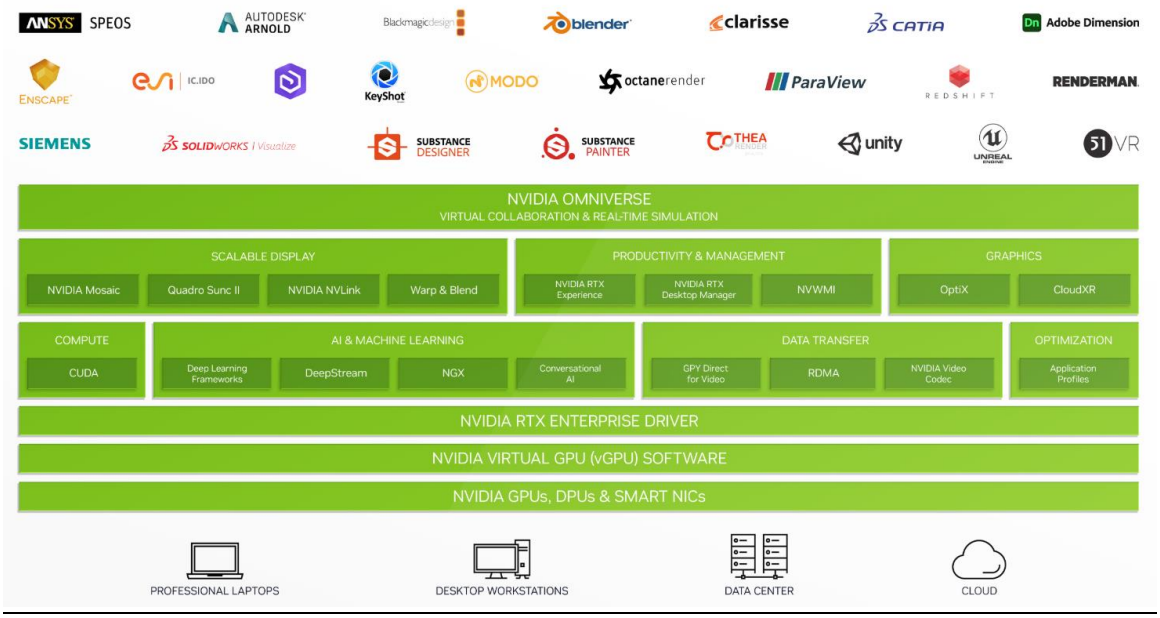


资料来源：英伟达官网

4.2、 专业可视化：构建丰富生态，打造 Omniverse 平台布局未来

打造生态平台，赋能专业领域新发展。在专业可视化领域，英伟达于 2018 年在 GPU 品牌 Quadro 中引入 RTX 技术，并在后续逐渐以 RTX 替代传统的 Quadro 命名方式。专注游戏场景的 GeForce 强调高性能，而用于专业绘图场景的 RTX 追求稳定性、正确性。英伟达围绕 NVIDIA RTX 开发了一个完整的生态系统，包括硬件、高级软件和工具、跨行业平台以及丰富的第三方应用程序网络，以此提供解决方案助力设计师、艺术家、科学家和研究人员以更快的速度解决问题，运用场景包括专业笔记本电脑、工作站、虚拟化、嵌入式场景等。

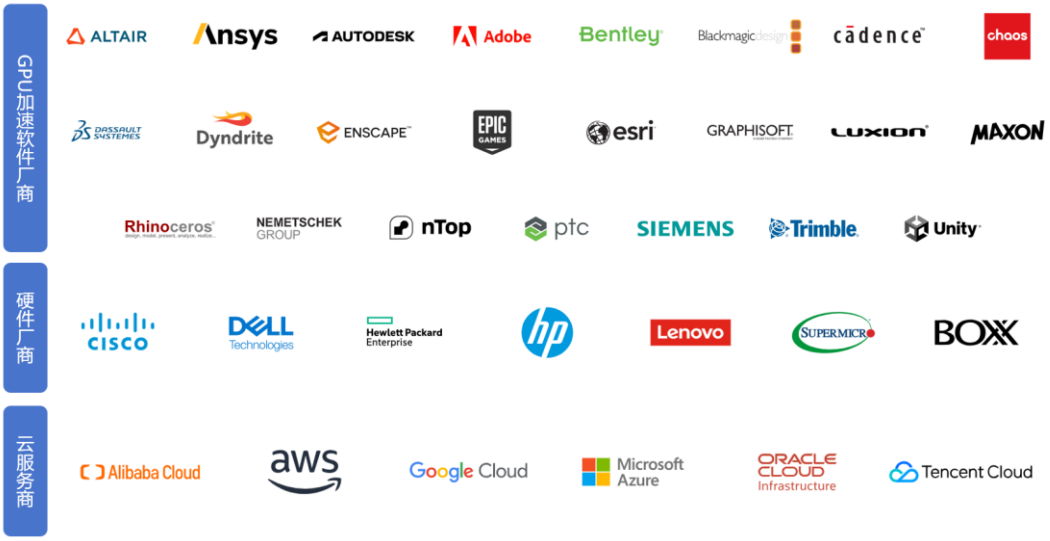
图42：英伟达专业视觉平台生态丰富



资料来源：公司官网

从软硬件到云服务上，英伟达专业显卡已经有较好渗透。超过 20 家主流创作软件厂商的产品针对 RTX 和 QUARDO RTX 进行加速优化；Dell、HP 和联想（3 大品牌工作站市占率超过 90%）是英伟达的核心合作伙伴；亚马逊、阿里等全球领先的云服务商为英伟达提供稳定的云服务支持。

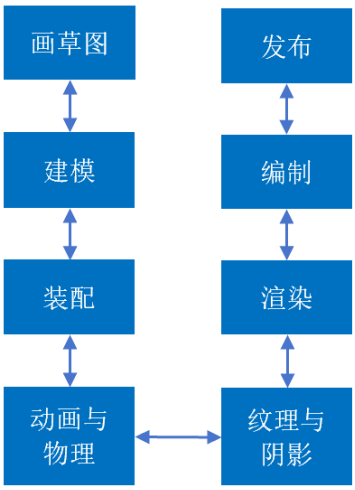
图43：英伟达专业可视化有诸多合作伙伴



资料来源：公司官网、开源证券研究所

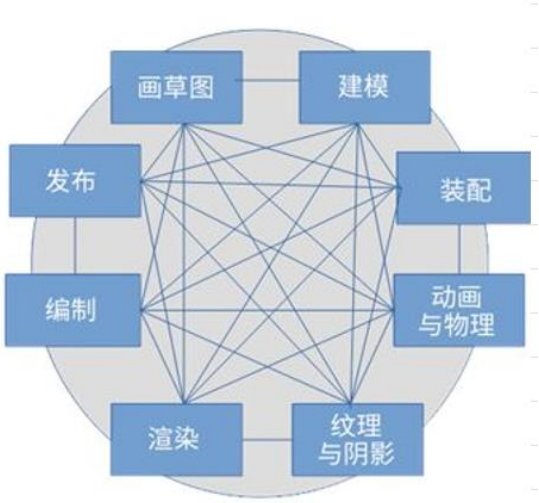
英伟达依托 Omniverse 平台，改变创作工作流程。NVIDIA Omniverse 是由英伟达开发的一个易扩展开放式平台，专为虚拟协作和实时逼真模拟打造。可以让各行业设计者能够通过云在软件之间、在本地或世界各地无缝地实时工作。传统的内容创作工作流程是线性的，需要逐步进行，且无法多个流程同时进行操作。Omniverse 将工作流程网络化，一个程序中的修改会立即反映到所有相关程序中，制作流程整合到一个统一的查看和修改环境中。Omniverse 被行业采用的关键是大型团队能够在共享的 3D 场景中跨多个软件应用程序同时工作，工程师可以同时处理模拟图像的相同部分。

图44：传统创作流程为线性模式



资料来源：pcidv.com、开源证券研究所

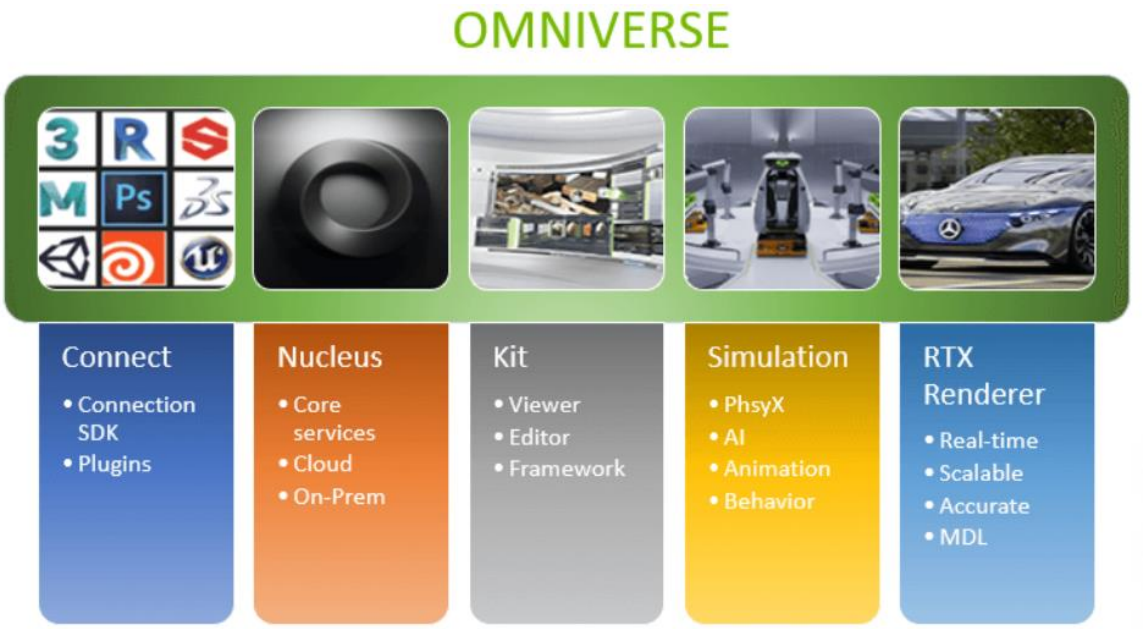
图45：Omniverse 创作流程可多流程实时同步



资料来源：pcidv.com

Omniverse 生态系统由 5 个组件组成：Nucleus，Connect，套件，仿真和 RTX。管理基于 USD 的 Omniverse Nucleus 服务器、用于先进设计应用程序的插件 Omniverse Connectors，最终用户应用程序 Omniverse Create 和 Omniverse View，以及 RTX 虚拟工作站工具。

图46：英伟达 Omniverse 平台核心组件



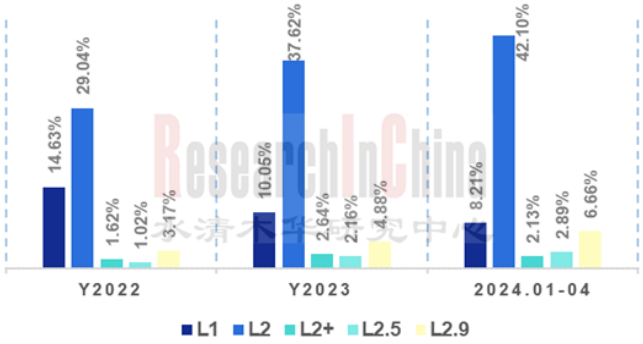
资料来源：公司官网

Omniverse 市场前景广阔，有望成为拉动专业可视化业务的重要力量。Omniverse 基于世界顶尖动画制作工作室 Pixar 被广泛采用的开源动画工具 USD（通用场景描述），将数十种设计者熟悉的开发平台兼容于一体，省去了设计师对于新开发环境的适应过程，简化应用间繁琐的导入/导出，实现了简洁高效的协作，以满足来自不同行业的多元需求。Omniverse 已将其覆盖范围从工程师扩大到几乎任何可以使用 Blender 的用户（主流 3D 创作软件），被称作是“工程师的元宇宙”，目前已被 700 多家公司和 7 万多名个人创作者采用，而全球有超过 4000 万使用高性能 PC 进行内容创作的创作者和工作室，未来可拓展市场空间较为广阔。

5、 汽车业务：域控芯片份额领先，期待 Thor 发布巩固地位

L2 及以上 ADAS 系统装配率快速提升，智能驾驶市场正处在加速渗透的窗口期。随着软件算法的不断迭代以及算力芯片和传感器等硬件成本的降低，智能驾驶已进入 L2+时代，国内乘用车 ADAS 系统功能(L1-L2.9)装配率稳步提升，且进入 2024 年，新上市乘用车 L2.9 装配率提升明显，这与国内车企和 Tier 1 将重点集中在高阶辅助驾驶、大规模落地行泊一体及 NOA 方案的趋势一致。

图47：国内乘用车 L2 及以上 ADAS 功能（分级别）装配率持续提升



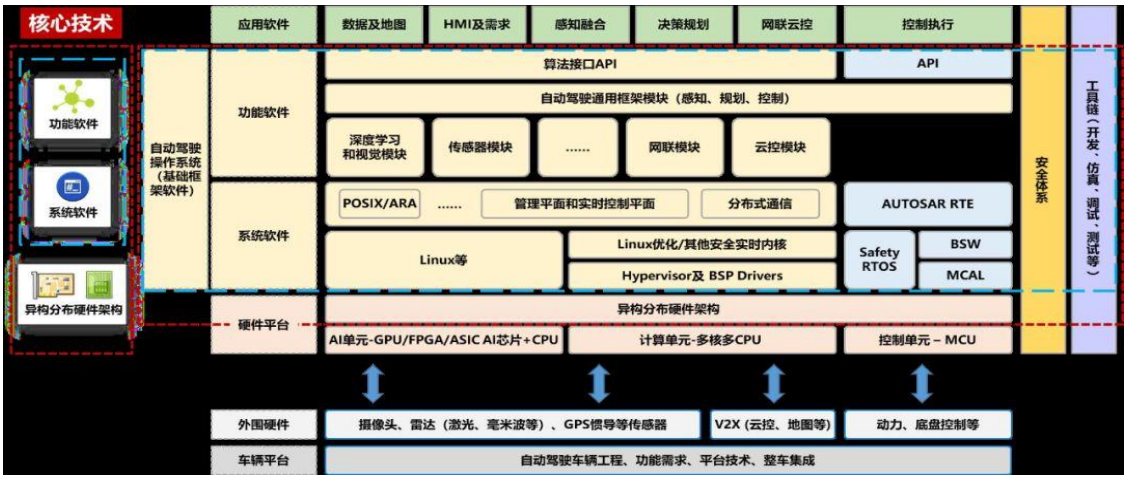
资料来源：佐思汽研官网

图48：2024 年后国内新上市乘用车 L2.9 ADAS 功能装配率提升明显

资料来源：佐思汽研官网

GPU 承担汽车 AI 能力主要角色：现阶段的 ADAS(高级驾驶辅助系统)功能较为独立，每个功能的前期预处理、数据融合、控制指令输出均有单独的芯片处理。随着芯片算力的迅速提升，软件算法的持续优化，大量计算将由一颗主芯片来承担。传统 CPU 存在算力不足和难以处理非结构化数据的缺陷，而 GPU 既可同时处理大量简单任务又可完成图像运算的特点，使其成为实现汽车高等级自动驾驶的主流方案。

图49：GPU 可适应汽车 AI 异构分布硬件架构的特征



资料来源：中国软件评测中心

英伟达构建了 DRIVE AGX 软硬件平台，整合了高性能的 GPU 计算能力、丰富的传感器接口以及高度优化的软件算法，为智能驾驶的训练和模拟提供了全方位的支持：

硬件上：2018 年英伟达发布 DRIVE Orin 芯片（Ampere 架构），2022 年继续发布 DRIVE Thor（Hopper 架构），算力达到 2000TOPS，相当于 Orin 的 8 倍，2024 年 DRIVE Thor 超级芯片进一步升级至 Blackwell 架构，并将于 2025 年量产，理想、极氪、比亚迪、广汽埃安昊铂、小鹏加入到 Thor 芯片的合作中。Thor 可以实现多域计

算整合车辆功能，而不是依赖分布式 ECU；

软件上：英伟达提供 DriveOS 操作系统，可用于 CUDA 库和 TensorRT，同时在 DriveOS 上提供 DriveWorks 中间件。

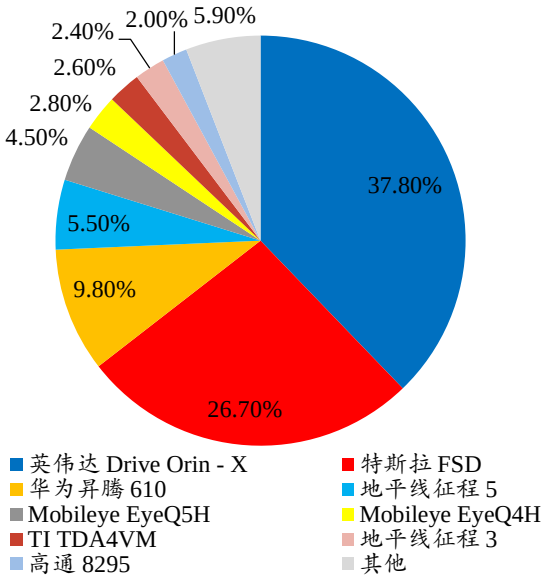
表15：英伟达 Thor 核心参数相较 Orin 有明显提升

	Orin	Thor
GPU 架构	Ampere	Blackwell
GPC/TPC 数量	2/8	3/11
算力	87TOPS(INT8)(DLA)+ 167TOPS(INT8)	1000INT8 TOPS/1000FP8 TFLOPS /500FP16 TFLOPS
最大时钟频率	1275MHz	1512MHz
L2 缓存	4MB	32MB

资料来源：智东西、开源证券研究所

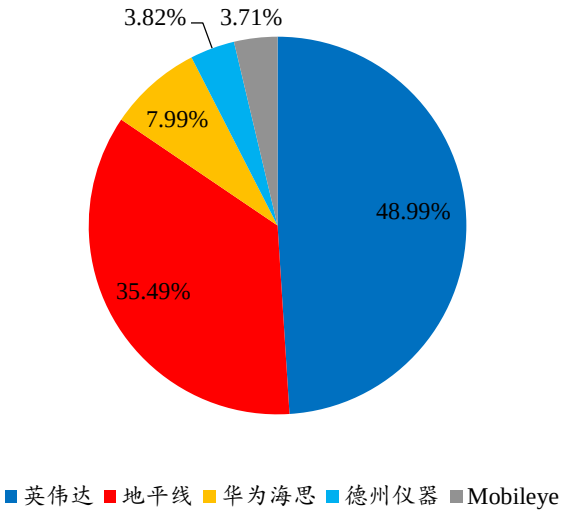
除了自研体系的特斯拉，英伟达在智驾域控芯片领域处于垄断地位。根据盖世汽车数据，2024 年 1-9 月英伟达中国智驾域控芯片装机量市场份额达到 37.8%，大幅领先除特斯拉外的其他厂商，2023 年 NOA 计算方案市场份额亦达到 48.99%，当前市场已经形成英伟达主导高端，地平线、黑芝麻智能等品牌主导中低端的市场格局。

图50：英伟达 Orin 系列在智驾域控芯片装机量市场份额领先



资料来源：盖世汽车官网、开源证券研究所（2024 年 1-9 月数据）

图51：2023 年英伟达在中国乘用车前装标配 NOA 高阶智驾计算方案市场份额领先



资料来源：高工智能汽车官网、开源证券研究所

英伟达算力、能效比领先，成为车企首选。在当前主流的智驾芯片方案中，英伟达算力明显领先于其余国内外厂商，同时保持了最高的能效比，此外英伟达采用模态化设计，为未来扩展到 L3-L5 留有空间，依托完善的软硬件工具链及更强的定

制化能力，英伟达成为众多智能汽车品牌的首选。待 2025 年 Thor 发布，在 Blackwell 框架下英伟达智驾芯片性能有望继续提升，市场地位或将持续巩固。

表16：英伟达智驾芯片算力、能效比领先

	NVIDIA	Mobileye	Tesla	高通	华为	地平线
芯片产品	Drive Orin	EyeQ5H	FSD	骁龙 Ride 8650	昇腾 610	征程 5
工艺	7nm	7nm	7nm	4nm	7nm	16nm
单颗算力	254TOPS	24TOPS	122TOPS	100TOPS	160TOPS	128TOPS
功耗	45W	10W	200W（双芯）	25-40W	65W	30W
能效比 TOPS/W	5.64	2.40	1.22	3.08	2.46	4.27
平台运用车型	理想、小鹏、蔚来、上汽智己	广汽、吉利、宝马等	Model 3/X/S	丰田、现代、宝马、长城	华为 HI 模式和鸿蒙智行模式合作车型、AION LX 和昊铂 GT	上汽、长城、比亚迪等
特点	可提供 Level2+级别的高级辅助驾驶功能，升级到双片 Orin 后可提供 Level4 级别方案，采用模块化设计，可对 L2-L5 灵活配置	黑盒，较难定制开发，算力和功耗较低，主要用于辅助驾驶	专用性能高，能耗低，全栈自研	走高性价比路线	可受益于华为汽车生态	强调开放生态，与英伟达路线类似

资料来源：佐思汽研官网、盖世汽车官网等、开源证券研究所

6、 盈利预测及投资建议：

英伟达收入预测：

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

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

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

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

英伟达盈利预测：

（1） **毛利率：**因复杂度更高的 B 系列 GPU 处于前期投入阶段，会对毛利率有所拖累，因此预计 FY2026 年毛利率将有所下滑，但随着 B 系列进入量产阶段，毛利率有望逐步回升至 74%-75%；

（2） **净利润：**得益于数据中心收入快速放量带来的规模效应，营业费率自 FY2024 财年以来持续降低，因此尽管 FY2026 财年毛利率或有所承压，我们预计公司净利率仍将有所改善。

表17：英伟达盈利预测

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

资料来源：Bloomberg、开源证券研究所

英伟达估值及投资建议：

我们选取具有算力芯片开发业务的可比公司与英伟达进行比较，从 PE 估值上看，英伟达略高于行业平均水平，从 PEG 的角度上看，若不考虑高通 2026-2027 年异常值影响，英伟达估值与行业平均水平相当，2025-2026 年 PEG 估值低于博通，总体上看，当前公司估值仍然合理。我们认为，当前 AI 算力芯片市场仍然供不应求，依托 CUDA 体系、持续推动芯片架构及互联技术的升级，英伟达 GPU 仍有望成为高算力集群时代的首要选择，“三芯战略”、10 万卡网络互联平台、汽车及机器人等领域存在想象空间。首次覆盖，给予“买入”评级。

表18：英伟达与可比标的估值对比

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

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

7、 风险提示

产能爬坡低于预期：若遇到技术瓶颈导致后续 GPU 产品系列开发及产能爬坡不及预期，或影响公司销售预期及估值。

行业需求低于预期：若下游大模型及 AI 运用没有较大突破，行业对 AI 算力的需求减弱，获将影响公司 GPU 销售。

行业竞争加剧：当前如博通、Marvell 等企业开发的 ASIC 芯片在推理市场上有所运用，或会对英伟达未来算力业务造成竞争。

附：财务预测摘要

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

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

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

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

数据来源：聚源、开源证券研究所

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已于2017年7月1日起正式实施。根据上述规定，开源证券评定此研报的风险等级为R4（中高风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为专业投资者及风险承受能力为C4、C5的普通投资者。若您并非专业投资者及风险承受能力为C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。
因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师承诺

负责准备本报告以及撰写本报告的所有研究分析师或工作人员在此保证，本研究报告中关于任何发行商或证券所发表的观点均如实反映分析人员的个人观点。负责准备本报告的分析师获取报酬的评判因素包括研究的质量和准确性、客户的反馈、竞争性因素以及开源证券股份有限公司的整体收益。所有研究分析师或工作人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（outperform）	预计相对强于市场表现 5%～20%；
	中性（Neutral）	预计相对市场表现在－5%～＋5%之间波动；
	减持（underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（underperform）	预计行业弱于整体市场表现。
备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普 500 或纳斯达克综合指数。我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。		

分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（以下简称“本公司”）的机构或个人客户（以下简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用，如接收人并非开源证券客户，请及时退回并删除。

本报告是基于本公司认为可靠的已公开信息，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、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。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、价值及投资收入可能会波动。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。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。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。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。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所引致的任何损失负任何责任。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。

本报告可能附带其它网站的地址或超级链接，对于可能涉及的开源证券网站以外的地址或超级链接，开源证券不对其内容负责。本报告提供这些地址或超级链接的目的纯粹是为了客户使用方便，链接网站的内容不构成本报告的任何部分，客户需自行承担浏览这些网站的费用或风险。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

开源证券研究所

上海
地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

深圳
地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

北京
地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

西安
地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn