英伟达（NVDA.O）

从硬件 GPU 设计到软件 CUDA+Omniverse 开发，建立人工智能和元宇宙生态系统

买入（首次）

盈利预测与估值	FY2022A	FY2023E	FY2024E	FY2025E
营业收入（百万美元）	26914	36208	49837	70454
同比（%）	61%	35%	38%	41%
归母净利润（百万美元）	9752	11109	17708	27028
同比（%）	125%	14%	59%	53%
每股收益（美元/股）	3.89	4.43	7.06	10.77
P/E（倍）	71.27	62.57	39.25	25.72

投资要点

- **英伟达高瞻远瞩，积极发展软件工具链，试图在未来拥有广阔市场空间的人工智能和元宇宙领域形成生态系统，实现对硬件芯片的绑定，筑牢行业壁垒。**与英特尔和 AMD（超威半导体）等以芯片设计为主的公司相比，我们认为英伟达之所以享有高估值，主要因为其已经不仅仅是芯片设计公司，而是软硬件并重的、形成生态系统的公司。英伟达对于扩展 GPU 的使用场景非常重视，以人工智能行业为例，2006 年在英伟达意识到 GPU 并行计算的优势后，就开始投入巨资开发 CUDA 这一软件工具链，让人工智能行业的研究者免费使用该软件来调用 GPU 的计算资源，这使得英伟达成为人工智能中深度学习的训练和推理领域的重要推动者，因此从人工智能行业发展开始，人工智能产业人员就在使用英伟达的软硬件套装，这些领域很难有动力使用英伟达之外的产品；英伟达也将这种软硬件并重的模式推向元宇宙（Omniverse）等新领域。**我们认为，在以云计算、自动驾驶等人工智能和以数字孪生、虚拟人为代表的元宇宙领域，英伟达有可能成为通用硬件平台+软件工具生态的供应商，类似于智能手机中的高通芯片+安卓操作系统的地位，行业壁垒很高。**

- **由于 GPU 架构的使用场景越来越丰富，我们认为英伟达潜在的市场空间（TAM）是现有传统业务的 4 倍。**我们认为其传统的消费者（游戏）业务的 TAM 为 1000 亿美元左右，目前正在快速扩张的数据中心领域 TAM 在 1500 亿美元左右，这是英伟达目前主要的营收领域。展望未来 10 年，**以下将是英伟达发挥巨大空间的领域：**在逐渐显露出竞争优势的汽车芯片 TAM 为 3000 亿美元左右；英伟达计划组成 GPU+DPU+CPU 的产品矩阵，未来的云服务器领域 TAM 在 3000 亿美元左右；在元宇宙时代可能大放异彩的 Omniverse 平台，其 TAM 在 1500 亿美元左右。

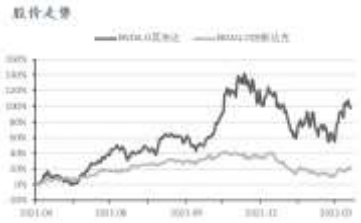
- **盈利预测与投资评级：**我们预计，随着英伟达在数据中心（三年复合增速 52%）、汽车(三年复合增速 86%)以及专业显示领域（三年复合增速 51%）营收快速增长，且 To B 利润率会略高于 To C 业务，其高 PE 会得到快速消化。我们考虑到公司在新兴领域的龙头地位和稀缺性，给予公司 FY2023 年 90 倍 PE，估值为 9188 亿美元，对应当前目标价为 366 美元，首次覆盖，给予“买入”评级。

- **风险提示：**国家政策风险；法律风险；自身技术风险；竞争者风险。

2022 年 04 月 01 日

证券分析师 张良卫
执业证书：S0600516070001
021-60199793
zhanglw@dwzq.com.cn
证券分析师 王紫敬
执业证书：S0600521080005
021-60199781
wangzj@dwzq.com.cn
研究助理 刘睿哲
执业证书：S0600121070038
liurz@dwzq.com.cn

股价走势



市场数据

收盘价(美元)	272.86
一年最低/最高价	206.50/307.11
市净率(倍)	25.74
流通股市值(百万美元)	684,879

基础数据

每股净资产(美元)	10.62
资本负债率(%)	39.77
总股本(百万股)	2510
流通股(百万股)	2510

内容目录

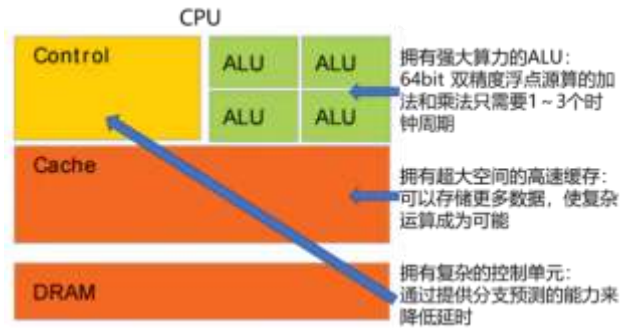
图表目录

1. 公司历史及业务简介

1.1. GPU 简介

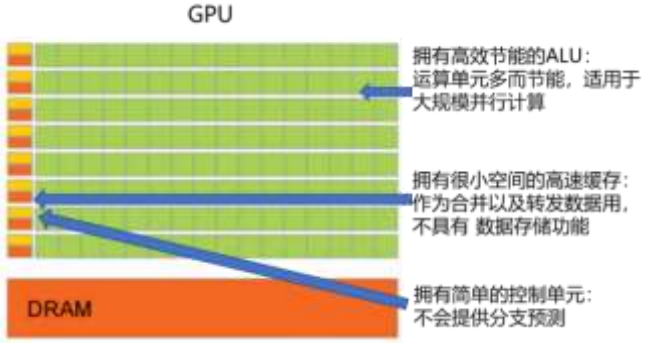
多核心的并行结构 GPU 比少核心串行结构的 CPU 更适合处理图形图像（矩阵结构）信息。CPU（Central Processing Unit，中央处理器）的功能主要是解释计算机指令以及处理计算机软件中的数据，是计算机的核心大脑，可以处理计算机遇到的所有指令。GPU（Graphics Processing Unit，图形处理器）是图形计算的重要元件， 主要用来处理与与图形图像相关的数据，在高端 PC 中通常会有独立 GPU，以获得更好的视觉体验。他们二者的区别主要是，CPU 通常有 4 个、8 个或 16 个强力 ALU 核心（arithmetic logic unit，算术逻辑单元），适合做复杂的通用串行任务；而 GPU 可能有数千个简单 ALU 核心，适合做简单特定的并行任务。我们通过以下的例子来说明 CPU 和 GPU 的差异：CPU 就像一个大学生，可以进行微积分等复杂计算，但若要在短时间内完成几万道加减算数问题，也是很难办得到的；而 GPU 就像几百个小学生，虽然都不会微积分等复杂计算的能力，但人数多，可以在很短时间内完成几万道加减算数问题。也有例子把 CPU 比作跑车，GPU 比作大卡车，对于将少量货物从 A 运到 B 来说，是作为跑车的 CPU 更快；但如果货物非常多，那么作为跑车的 CPU 需要往返的次数远远多于作为货车的 GPU，作为货车的 GPU 虽然完成一次任务较慢，但是可以携带更多的货物，其效率会高于 CPU。总而言之，对于复杂的单个计算任务来说， CPU 的执行效率更高，通用性更强；而对于图形图像这种矩阵式多像素点的简单计算，更适合用 GPU 来处理，但通用性较弱。

图1：CPU 的基本结构及原理



数据来源：简书网，东吴证券研究所绘制

图2：GPU 的基本结构及原理



数据来源：简书网，东吴证券研究所绘制

GPU 按接入方式分为独立 GPU 和集成 GPU；按照应用端划分为移动 GPU、服务器 GPU 和 PC GPU。GPU 是图形处理单元，在 PC（个人电脑）早期，图形数据较为简单，主要都是由 CPU 来进行图形处理。随着图形显示规模的增加，CPU 已经很难分出更多精力来处理图形信息，而且 CPU 的架构决定了其处理图形信息的效率是偏低的，因此逐渐发展出了专门处理图形信息的 GPU。英伟达专做 GPU，开发了独立于 CPU 的 GPU；英特尔作为 CPU 的霸主，开发了寄生于 CPU 芯片上的 GPU 单元，被称为集成 GPU。通常来讲，独立 GPU 的性能都要优于集成 GPU，在对图形实时处理要求不高的

日常办公领域，使用普通的集成 GPU 即可；在对图形实时处理能力要求很高的游戏及设计领域，一般都需要使用独立 GPU。随着移动设备的发展，GPU 也从 PC 端扩展到了移动端，高通骁龙以及苹果的 A 系列芯片都开发了相应的 GPU 芯片模块。

图3：GPU 的分类

接入方式	独立GPU	AMD （Radeon系列）
		NVIDIA （Geforce系列）
	集成GPU	英特尔 HD系列
		AMD APU系列
应用端	移动GPU	高通Adreno、ARM Mali、苹果GPU
	服务器GPU	英伟达Tesla、AMD Fire Stream
	PC GPU	英特尔、AMD、英伟达

数据来源：东吴证券研究所整理

随着 AI 以及云计算的兴起，具有并行计算架构的 GPU 具有更高的效率，这也使得 GPU 被应用到 AI 及云计算等数据处理之中。这是一个全新的领域，拥有巨大的成长空间。值得一提的是，市场上还存在着比 GPU 专用程度更高的芯片，包括 FPGA（Field-programmable gate array，可编程逻辑阵列）和针对某一类 AI 计算的 ASIC（Application-specific integrated circuit，特定场景芯片），包括谷歌推出的 TPU（张量计算单元）和特斯拉推出的 NPU（神经网络计算单元），虽然在某些特定计算上效率更高，但目前这些芯片的使用场景比较单一，市场规模还较小。

图4：常见芯片特点总结

专用性越来越强，效率越来越高

芯片种类	CPU	GPU	FPGA	ASIC
芯片架构	计算单元和高速存储单元占用的晶体管数量相当，适合串行计算	晶体管大部分构建计算单元，运算复杂度低，适合大规模并行计算	可编程逻辑，计算效率高，更接近底层IO，通过冗余晶体管和连线实现逻辑可编程	晶体管根据算法定制，不会有冗余，功耗低，计算性能高，计算效率高
擅长领域	通用计算领域	图像处理以及并行计算	算法更新频繁的专用领域，通常作为ASIC的试验版本	市场需求量大的专用领域
优点	通用性强	并行运算能力强	计算效率比CPU和GPU更高	体积小、功耗低、计算性能高、计算效率高、芯片出货量越大成本越低
缺点	针对某些特定领域效率很低	价格贵、功耗散热高	编程门槛高、峰值性能不如ASIC、量产成本高	算法固定、开发周期长、上市速度慢、一次性成本高、风险大

数据来源：东吴证券研究所整理

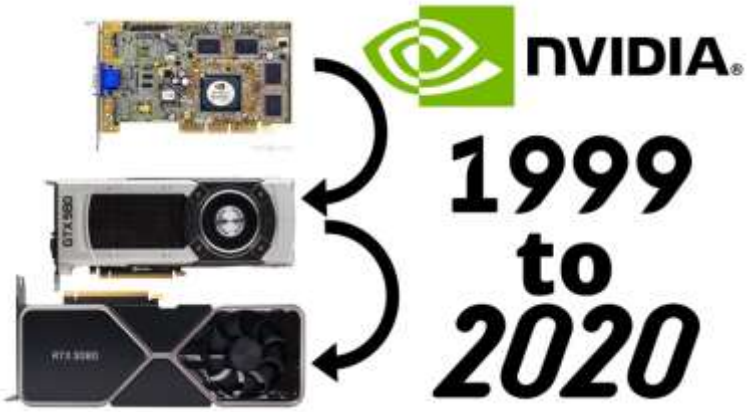
1.2. 英伟达发展历史

英伟达（NVIDIA）是一家以 GPU（Graphics Process Unit，图形处理单元）芯片设计起家的人工智能计算公司。公司创立于 1993 年，总部位于美国加利福尼亚州圣克拉拉市。美籍华人 Jensen Huang（黄仁勋）是创始人兼 CEO。1999 年，NVIDIA 定义了 GPU，GPU 的出现被业界视为现代计算机图形技术的开端。英伟达于 1999 年 1 月在纳斯达克挂牌上市，在 2000 年它收购了曾经在 90 年代称霸图形显示市场的 3dfx 公司的知识产权，逐渐占据图形显示市场的优势地位。到 2021 年为止，在消费 PC 领域，能够量产 GPU 的公司只有英伟达、AMD 和英特尔，其中英特尔主要是以集成 GPU 为主，

AMD 既有集成 GPU 也有独立 GPU，英伟达主要是独立 GPU。在独立 GPU 领域，英伟达 2021Q1 占据 81%的市场份额，处于绝对的领先地位。

公司 20 多年来始终引领 GPU 行业的发展，将 GPU 的主要应用场景从游戏以及画图等图像显示扩展到了以 AI、云计算等大数据相关的并行计算领域。英伟达保持着两年升级一次 GPU 架构的步伐，不断提高 GPU 的性能。在英伟达 GTC 2020 主题演讲中，NVIDIA 宣布推出安培(Ampere)架构，这是 NVIDIA 发布的第八代 GPU 架构，包含超过 540 亿个晶体管，性能相较于前代提升了高达 20 倍，也是 NVIDIA 8 代 GPU 历史上最大的一次性能飞跃。安培架构的最新一代 RTX30 系列游戏 GPU 和 AI 计算 GPU A100 作为各自领域的代表产品，继续推动着相关领域的发展。

图5：英伟达 GPU 发展历史

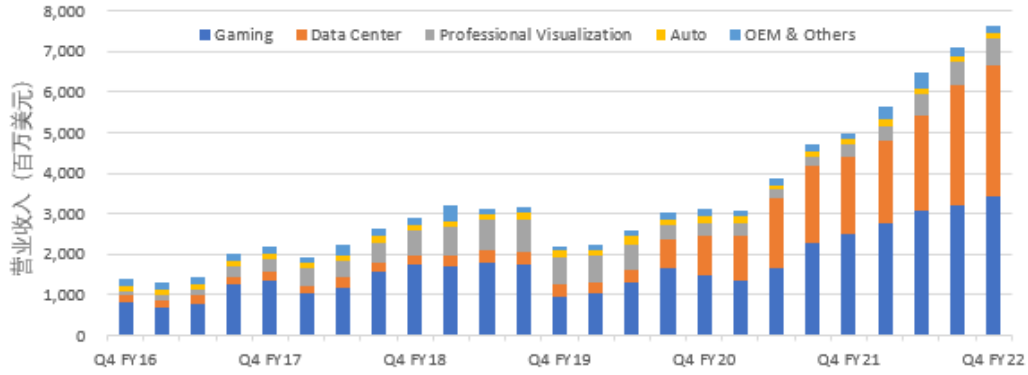


数据来源：公司官网，东吴证券研究所

1.3. 英伟达业务简介

按照 FY2022（对应公历 2021.1~2022.1）的年报分法，英伟达有消费者（游戏）业务 Gaming、数据中心业务 Data Center、汽车业务 Auto、专业解决方案业务 Professional Visualization 以及 OEM 和其他业务 OEM&Others，英伟达主要为这些领域提供 GPU 芯片及相应的软件工具链。从消费者行为来看，在 PC 端购买独立 GPU 的主要目的是为了体验高性能游戏，因此英伟达将 PC 端 GPU 的销售业务称之为游戏业务。游戏业务一直是英伟达的主营业务，在各板块中营收排名领先；随着 AI 和云计算的不断兴起，英伟达逐渐将 GPU 打造为 AI 和云计算提供算力的底层芯片，这部分与 AI 和云计算相关的业务被称为数据中心业务。英伟达数据中心业务营收从 2020 年以来迎来迅速增长，我们预计在 2025 年左右将成为营收规模最大的板块；汽车智能化对算力需求的提高，英伟达也将 GPU 芯片装入车辆中为其提供高算力。随着汽车智能化的不断提速，我们预计英伟达汽车业务营收也会快速增长，成为公司的一个重要板块。

图6：英伟达分业务的历史营收变化



数据来源：历年公司财报，Wind，东吴证券研究所

2. 传统业务：消费者（游戏）相关业务保持稳定增长

2.1. 英伟达 GPU 五年来持续占据 PC 独显六成以上市场

由于疫情导致的居家时间延长，公司 GPU 量价齐升，FY2022Q2 游戏相关营收同比大增 85%，单季收入首次超过 30 亿美元。英伟达的 GPU 在 PC 端是以独立显卡的形式存在，通过独立显卡可以实现高帧率高分辨率 3A 游戏、专业绘图等应用。独显领域是一个壁垒极高的市场，经过 20 多年的充分竞争后，目前仅有英伟达、AMD 可以推出相关产品，而英伟达占据绝对的领先优势。据研究机构 Jon Peddie Research 报道，2021Q1 英伟达在 PC 独显市场占据 81%的市场份额（2020 全年为 77%）。

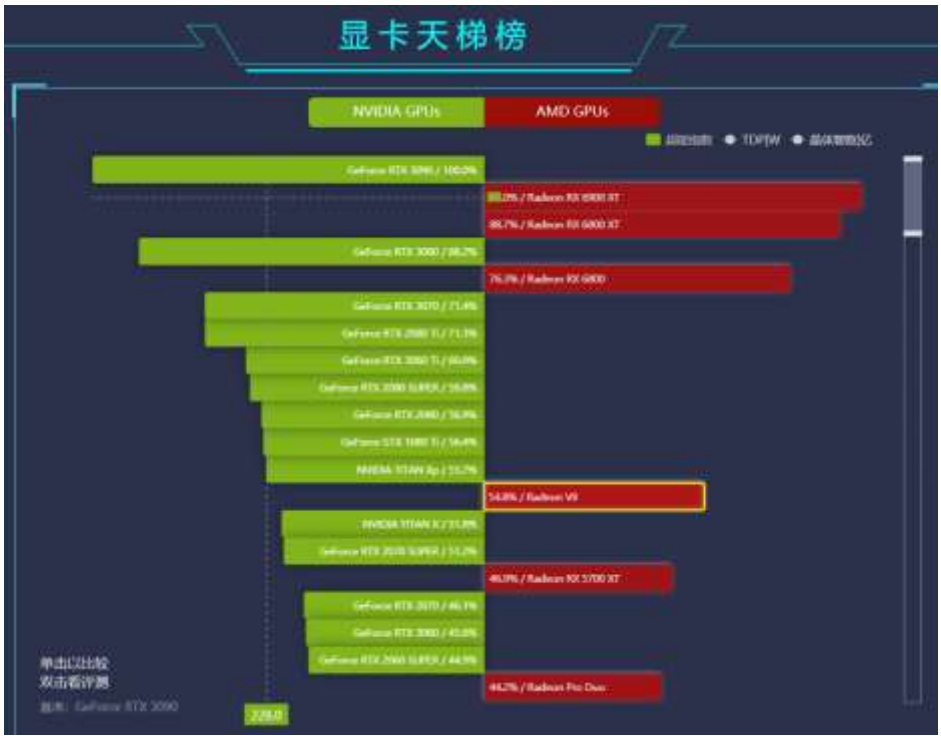
图7：全球台式机 GPU 市场份额变化（单位：百万片）

数据来源：Jon Peddie Research，东吴证券研究所

以每两年更新一次架构、每半年性能翻倍的速度，持续引领消费级 GPU 市场。2020

年 9 月 2 日，英伟达发布了新一代显卡 RTX30 系列，与前一代 RTX20 系列相比，采用了全新的安培架构，在核心数、显存、频率等性能都有了大幅度提升。RTX30 的高算力加上英伟达的 DLSS（Deep Learning Super Sampling，深度学习超采样）技术，大大提高实际场景的运算力（在算力不变的情况下提高帧率），使得英伟达显卡深受游戏玩家的喜爱。在中国，RTX30 系列中的 RTX3080 由发售价的 5499 元人民币被一路炒高至 18000 元左右，足见其火爆程度（虽然部分原因是受到数字货币“挖矿”抢货的影响）。英伟达以半年性能提升一倍的“黄氏定律”牢牢占据 GPU 的领导者地位。截止到 2021 年 3 月，英伟达的各系列 GPU 在性能排行的前 20 名中占据了包括第一名在内的 14 个席位，可以看出英伟达在 GPU 领域的霸主地位。

图8：截止到 2021 年 3 月的 GPU 排行榜



数据来源：henglong，东吴证券研究所

图9：英伟达 DLSS 技术展示

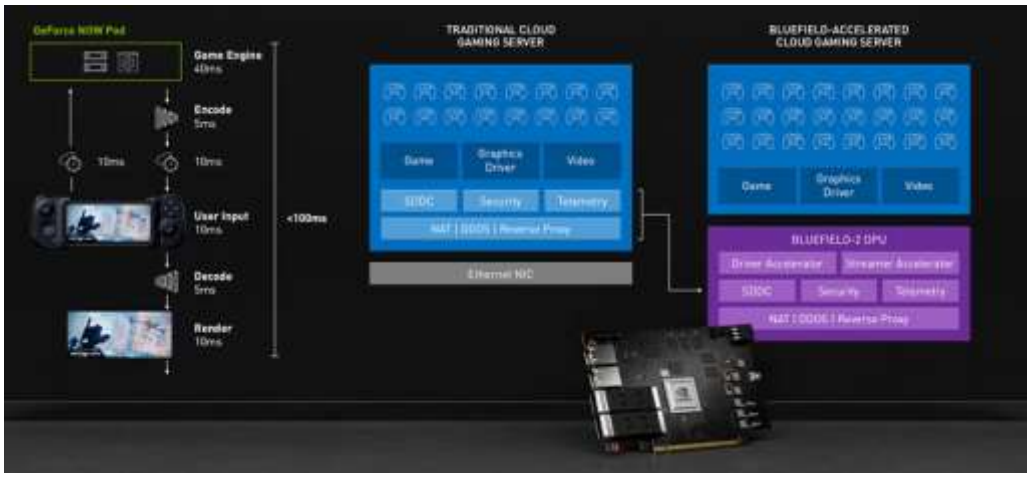
数据来源：公司官网，东吴证券研究所

2.2. 借助 Bluefield 能力，发力云游戏 Geforce Now 业务

云游戏是以云计算为基础的游戏方式，在云游戏的运行模式下，所有游戏都在服务器端运行，并将渲染完毕后的游戏画面压缩后通过网络传送给用户。在客户端，用户的游戏设备不需要任何高端处理器和显卡，只需要基本的视频解压能力就可以，因此其市场潜力很大，据 Newzoo 2021 年 3 月发布的报告预测，2023 年全球云游戏市场收入可能达到 51 亿美元。但目前主要受限于网络延迟以及服务器延迟等方面，市场尚处于初期阶段。除英伟达外，目前还有微软、谷歌、索尼、腾讯以及网易等也在拓展云游戏业务。

英伟达云游戏平台 Geforce Now 采用 Bluefield 架构，解决云游戏服务器的延迟问题。对于云游戏来说，延迟是最亟待解决的问题。而控制延迟的关键，不仅需要良好的通信网络能力，更为重要的是对云端服务器的数据处理特别是图形相关的处理速度。英伟达利用其在数据中心的经验，优化了服务器架构，推出了英伟达云游戏平台 Geforce Now，采用 RTX 服务器来实现更低延迟（整体延迟小于 100ms），使云游戏体验得到了优化。由于目前云游戏仍受限于网络延迟，整个市场尚不成熟，但随着基础设施的不断发展，此项业务将为英伟达带来未来全新增长空间。

图10：英伟达云游戏 Geforce Now 采用 Bluefield 架构来减小延迟

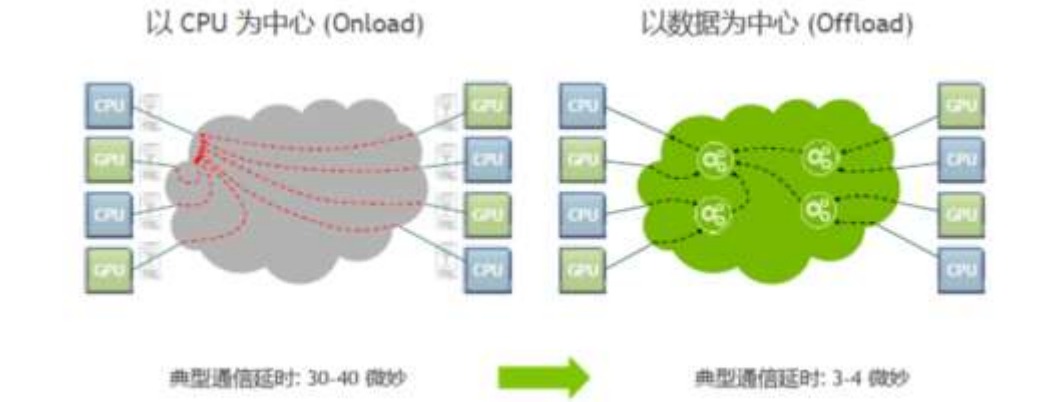


数据来源：公司官网，东吴证券研究所

3. 成长业务：数据中心成为云和 AI 领域基础设施，营收迅速扩大

英伟达成为云计算和 AI 这个未来“金矿”行业的芯片及服务器等“铲子”工具——GPU 的主要供应商，2020 年以来以 AI 和云计算为主要服务对象的数据中心业务营收规模已经和游戏业务相当。英伟达创始人、CEO 黄仁勋于 2021 年 6 月份在接受第一财经的采访中表示，“数据中心规模计算的时代已经来临。我们想成为一家数据中心企业，数据中心正在占据我们业务越来越重要的地位。”他说道，“而各种新兴技术的汇聚，比如云计算、人工智能、加速计算、工业 5G 等，将会成为解决计算时代重要问题的最后几块拼图。”英伟达在数据中心上布局很早，利用在 GPU 中积累的芯片设计经验，推广到了数据中心业务。从英伟达近一年的财报中也可以看出，英伟达在数据中心的业务收入已经和游戏业务比肩，且有超越游戏业务的潜力。从 2021 年 6 月举办的国际超级计算大会 ISC 上公布的超级计算榜单可以看出，TOP10 中有 8 台使用英伟达的技术，TOP500 中有 342 台使用英伟达的技术，可见英伟达在数据中心业务的优势。英伟达在数据中心领域的成功离不开硬件（A100、DGX A100、InfiniBand）以及相关软件（CUDA）等的支持，英伟达在云与数据中心领域形成了一整套完整的生态系统，成为云和 AI 领域基础算力及算法工具链等基础工具的供应商，在 AI 的布局中拥有不可替代的位置。

图11：数据中心逐渐成为互联网架构的核心



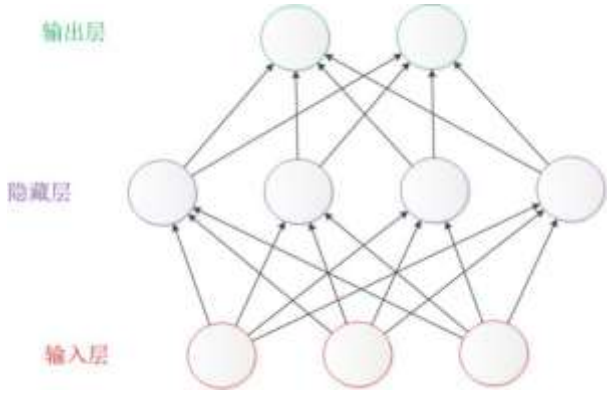
数据来源：Ithome，东吴证券研究所

3.1. 采用并行计算的 GPU 天生适合 AI 领域的运算

AI 算法多为并行结构。AI 领域中用于图像识别的深度学习、用于决策和推理的机器学习以及超级计算都需要大规模的并行计算，更适合采用 GPU 架构。我们以深度学习中的神经网络算法来举例说明 GPU 架构的优势。

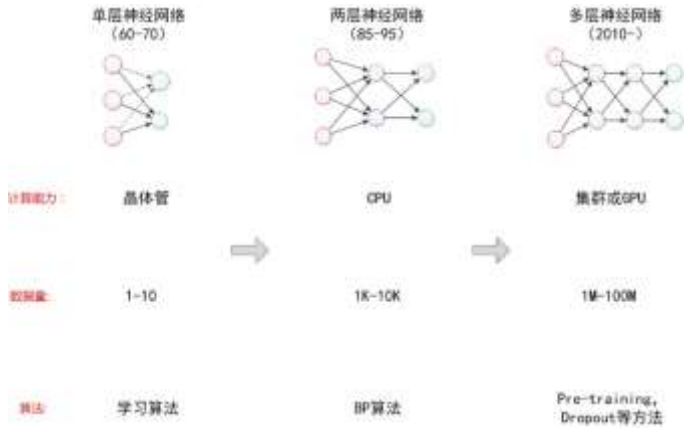
神经网络是一种模拟人脑的以期能够实现人工智能的机器学习技术，适合采用并行计算的 GPU 架构。一个经典的神经网络分为输入层、隐藏层和输出层，通常隐藏层的数量越多，神经网络模拟的结果越精确，但相应的计算量会呈指数的增长。最初人们使用 CPU 来模拟多层神经网络需要很长时间；随后科学家认为，输入层到输出层的计算关系是矩阵形式，与 GPU 对图像像素处理的架构类似，都是并行计算为主，因此产生了使用 GPU 来进行神经网络计算的想法。2010 年时，Google 负责人工智能的吴恩达为了训练神经网络来识别猫，最初使用了 16000 台计算机的 CPU 完成了训练，但为了搭建庞大的 CPU 耗费巨大；随后他与英伟达公司探讨了这件事情，英伟达仅采用 12 个 GPU 就完成了训练，使人们看到了 GPU 对神经网络的优势。随着神经网络的复杂程度逐渐提高，用 GPU 来训练神经网络成为了更优的选择。

图12：具有并行结构的神经网络



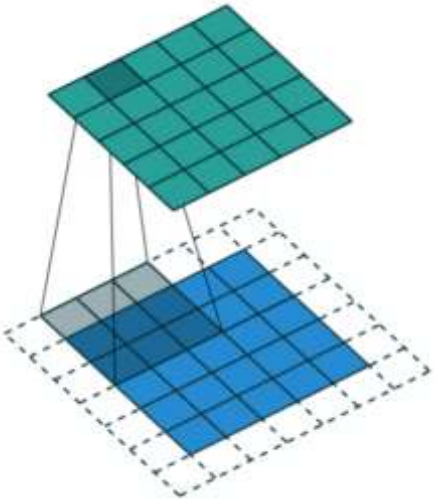
数据来源：CNblogs，东吴证券研究所绘制

图13：神经网络的发展历程



数据来源：CNblogs，东吴证券研究所绘制

图14：卷积算法示意图



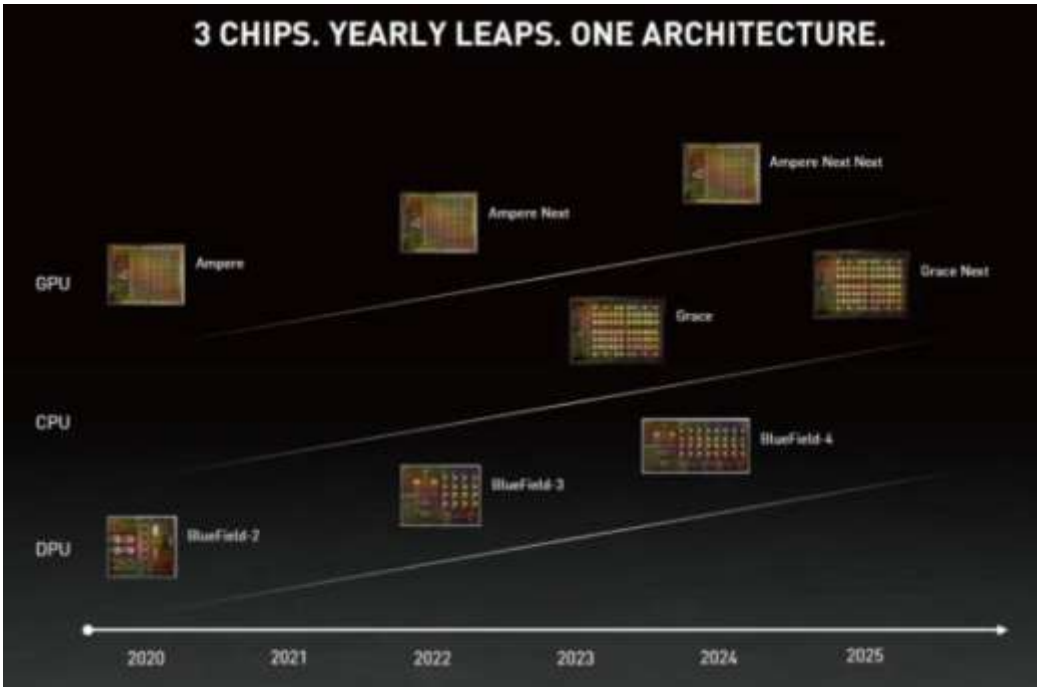
数据来源：Lebooj 知乎公众号，东吴证券研究所

如上图所示，我们在底部有一个蓝色的输入通道。在输入通道上滑动的底部有一个阴影的卷积滤波器，还有一个绿色的输出通道。卷积算法流程如下：蓝色（底部）→ 输入通道 →阴影（覆盖在蓝色上）→ 3x3 的卷积过滤器 →绿色（顶部）→输出通道。对于蓝色输入通道上的每个位置，3x3 过滤器进行计算，将蓝色输入通道的阴影部分映射到绿色输出通道的相应阴影部分。每个计算都是独立于其他计算的，这意味着任何计算都不依赖于任何其他计算的结果，所有这些独立的计算都可以在 GPU 上并行进行，虽然单个卷积计算要比 CPU 慢，但是对于整个任务来说，CPU 要逐个依次完成，速度要大大慢于 GPU。因此，卷积运算可以通过使用并行编程方法和 GPU 来加速。

3.2. 英伟达全面布局数据中心硬件市场

CPU+GPU+DPU 形成产品矩阵，全面发力数据中心市场。自从 2021 年 GTC 大会上英伟达宣布推出第一款 CPU Grace 以来，英伟达已经涉足了与 AI 和云计算相关的数据中心市场的大部分领域。利用 GPU 在 AI 领域的先天优势，英伟达借此切入数据中心市场。针对芯片内部带宽以及系统级互联等诸多问题，英伟达推出了 Bluefield DPU 和 Grace CPU，提升了整体硬件性能。在 2021 年 GTC 大会上，英伟达公布了 GPU、CPU 和 DPU 的发展规划，每年都会有新产品问世；英伟达在数据中心硬件市场的不断升级，推动了数据中心以及 AI 整个产业的发展步伐。

图15：英伟达硬件的升级规划路线



数据来源：公司官网，东吴证券研究所

3.2.1. 基于安培架构的 A100 系列，为数据中心打造高性能算力基础

作为安培架构的代表，A100 GPU 在在深度学习、数据分析、能效方面都获得了前所未有的优化，被广泛应用于自然语言识别、大数据分析、科学计算领域。在 GTC2020 大会上，英伟达推出了安培架构的首款超算 GPU——A100。A100 引入了有着里程碑式意义的 Tensor Cores 双精度计算技术，这使得 A100 的算力比前一代 V100 提高了 175%。NVIDIA A100 Tensor Core GPU 针对 AI、数据分析和 HPC（high performance computing，高性能计算）等应用上，实现了更强的加速，针对极其严峻的计算挑战上有了更大作为。作为 A100 GPU 系列中的最新力作，在架构特性上有如下特点：

- 采用第三代 Tensor Core 核心。通过全新 TF32，将上一代 Volta 架构的 AI 吞吐量提高多达 20 倍。通过 FP64，将 HPC 性能提升了 2.5 倍。通过 INT8，将 AI 推理性能提高多达 20 倍，并且支持 BF16 数据格式。
- 采用更大、更快的 HBM2e GPU 内存。从而使内存容量增加一倍，在业内率先实

现 2TB/s 以上的内存带宽。

- 采用 MIG（Multi-Instance GPU，多实例 GPU）技术，将单个独立实例的内存增加一倍，可最多提供七个 MIG，每个实例具备 10GB 内存。
- 采用结构化稀疏技术，将推理稀疏模型的速度提高两倍。
- 第三代 NVLink 和 NVSwitch，相较于上一代互连技术，可使 GPU 之间的带宽增加至原来的两倍，将数据密集型工作负载的 GPU 数据传输速度提高至 600 GB/s。

图16：英伟达 GPU 架构升级带来的性能提升

数据来源：公司官网，东吴证券研究所

图17：A100 成为世界上最强性能的 AI 计算 GPU

数据来源：公司官网，OFweek，东吴证券研究所

A100 被广泛应用于大数据分析、天气预报、量子化学以及材料模拟等领域，推动了相关领域的发展。基于以上算力、内存以及数据交互上的优化，A100 在自然语言识别、大数据分析、科学计算领域提供了更强的硬件实力。对于如 RNN－T 等自动语言识别模型的 AI 推理，单个 A100 MIG 实例可处理更大规模的批量数据，将生产中的推理吞吐量提高 1.25 倍。在 TB 级零售大数据分析基准上，A100 将其性能提高了 2 倍，使其成为可对最大规模数据集进行快速分析的理想平台。随着数据的动态更新，企业可以实时做出关键决策。对于科学应用，A100 可为天气预报和量子化学等领域提供巨大的加速。材料模拟软件 Quantum Espresso 采用单节点 A100 实现了近 2 倍的吞吐量提升。

除了高性能的 A100 系列外，英伟达还在在功耗、性能上做了优化与调整，推出了 A10、A30 等产品，旨在面向中小型客户。英伟达的一系列产品满足了不同用户的需求。

3.2.2. DGX A100 数据中心及 DGX SuperPOD 解决方案，使英伟达保持超算领域优势

一体式 AI 数据中心 DGX Station A100，使 AI 超算中心的搭建更为方便。以 A100 GPU 为核心的数据中心 DGX Station A100，AI 性能可以达到 2.5 Petaflops，通过 NVIDIA NVLink 完全互连，实现多个 NVIDIA A100 GPU 融合在一起的工作组服务器，目前有 320GB/640GB 等不同版本可供选择。借助 MIG，单一 DGX Station A100 最多可提供 28 个独立 GPU 实例以运行并行任务，并可在不影响系统性能的前提下支持多用户应用。作为服务器级的系统，DGX Station A100 无需配备数据中心级电源或散热系统，用户可以极为方便地部署 AI 超算中心；与前代相比，如果要搭建同样算力的数据中心，成本

会降低 90%，耗电量会减少 95%（如图 18 所示数据），大大降低了数据中心的使用门槛，客观上推动了 AI 领域的蓬勃发展。

图18：DGX Station A100 与上一代价格、场地占用和耗电量对比图

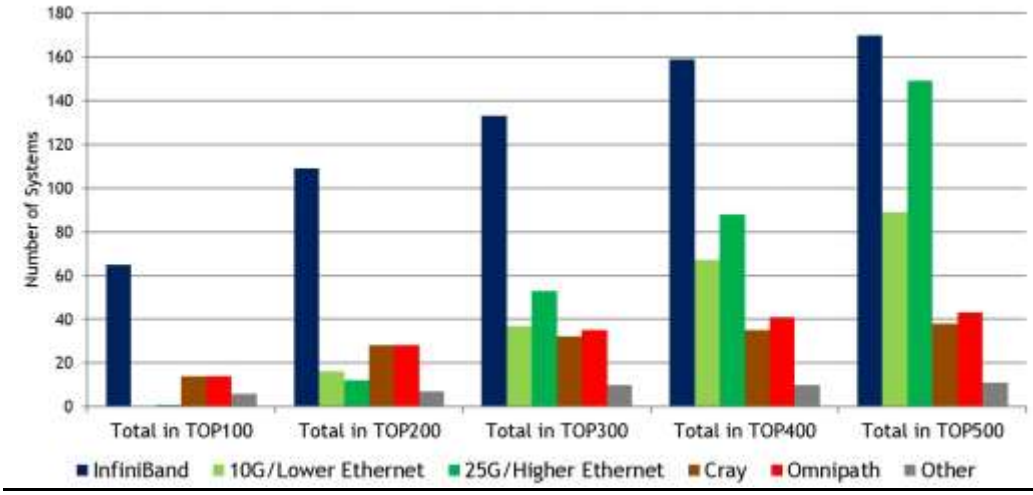
数据来源：nextplatform，东吴证券研究所

NVIDIA DGX SuperPOD 解决方案，促进了 AI 超算行业的发展。全新 DGX A100 640GB 系统也将集成到企业版 NVIDIA DGX SuperPOD 解决方案，使机构能基于以 20 个 DGX A100 系统为单位的一站式 AI 超级计算机，实现大规模 AI 模型的构建、训练和部署。配备 A100 80GB GPU 的 NVIDIA DGX SuperPOD 系统将率先安装于英国的 Cambridge-1 超级计算机，以加速推进医疗健康领域研究；佛罗里达大学的全新 HiPerGator AI 超级计算机，将开展 AI 赋能的科学发现。新一代 DGX Station A100 和 DGX A100 640GB 移动数据中心的出现，将给 AI 超级计算机的行业格局带来一次新的震动。同时 AI 超算上的创新也将因为 DGX Station A100 而再次迎来新的发展，对 AI 超算的行业应用普及带来了更大的发展潜力与空间。

3.2.3. 战略眼光独到，收购 Mellanox，提高数据交互性能

英伟达并购 Mellanox 后，充分挖掘了其掌握的 InfiniBand 技术，使网络交换速度得到保证。2019 年，英伟达以 69 亿美元并购了 Mellanox，后者以 InfiniBand 技术见长。InfiniBand 和以太网是超算领域较常用的互联和协议，以太网设计的初衷是解决各种各样设备之间的连接问题，其核心是通用性强；而 InfiniBand 的设计初衷是解决同一个系统中不同设备之间的连接问题，其核心是为了让通讯更快。举例来说，以太网像是快递中转站，它需要尽可能识别所有的包裹并将其送到各种各样的目的地，其主要精力需要放在数据处理上，信息的传递效率相对较低；而 InfiniBand 更像是地铁系统，轨道都是确定好的且目标车站数量有限，因此不同站点间信息获取速度就会很快。对于高性能超级计算机来说，为了提高数据交换速度，一般会采用 InfiniBand 技术。英伟达在得到 InfiniBand 技术后，开发出了 NVIDIA Mellanox InfiniBand 交换器系统，每个端口的速度可达 400Gb/s（以太网的速度通常在 0.1~25 Gb/s），这让运算丛集和聚合数据中心能在任何规模中运作，并同时降低营运成本和基础架构的复杂性。

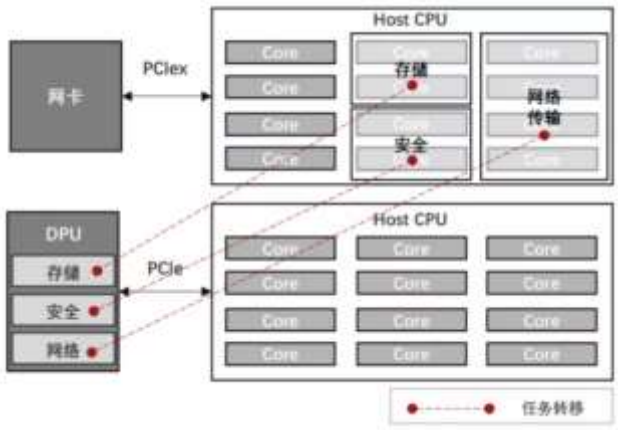
图19：100 个机器增量将孔径从 Top100 扩大到 Top500 的互联分布



数据来源：nextplatform，东吴证券研究所

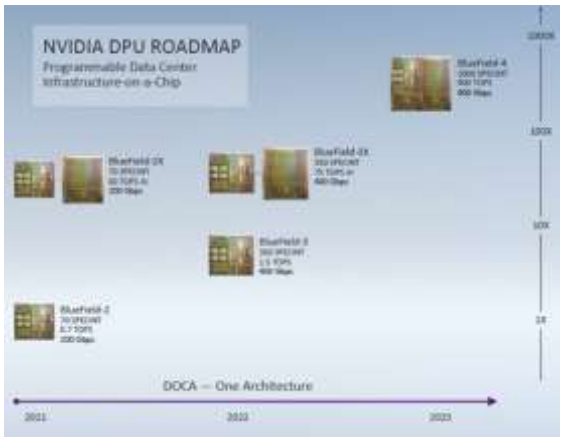
Bluefield 芯片可分担 CPU 的网络、存储和安全等任务，可以大大减少 CPU 的工作量的同时提高数据交互性能。Mellanox 的主要产品就是名为 Bluefield 的芯片，英伟达也将其称为 DPU(Data Processing Unit，数据处理单元)，其实际上是一个高级的网卡。基于 DPU 的智能网卡将成为云数据中心设备中的核心网络部件，逐渐承担原先需要 CPU 来执行的网络数据处理、分发的重任，从而从根本上实现软件定义网络（SDN）和网络功能虚拟化（NFV）的诸多优势，有效降低云计算的性能损失，释放 CPU 算力，降低功耗的同时大大减少云数据中心的运营成本。按照英伟达的说法，一个 DPU 顶 125 个 CPU 的网络处理能力。英伟达计划在 2022~2023 年推出第 3 代与第 4 代 Bluefield，在保持 400Gb/s 的数据传输速度下，其 AI 算力会从 75TOPS 提高到 400TOPS，进一步满足高性能数据交互的要求。

图20：英伟达 GPU 架构升级带来的性能提升



数据来源：半导体行业观察，东吴证券研究所

图21：英伟达 DPU 的升级规划



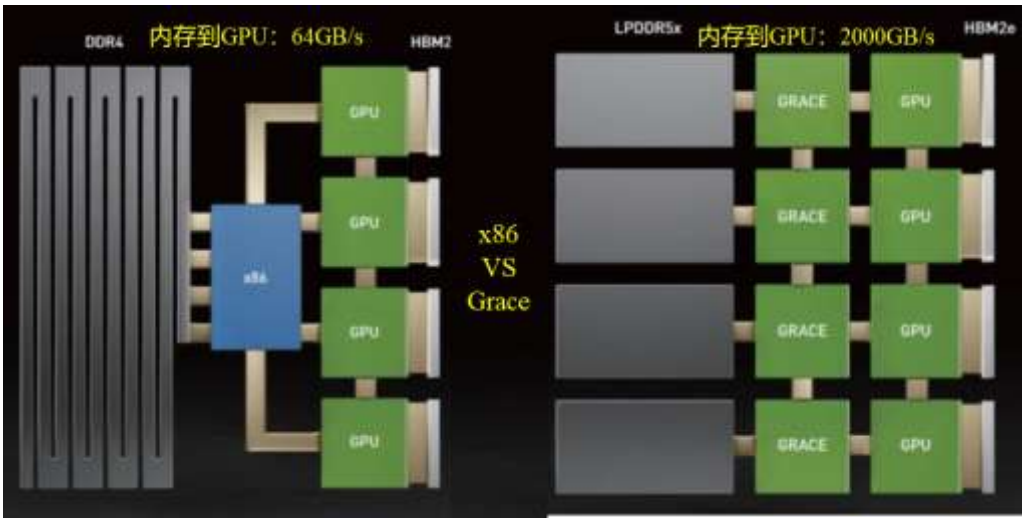
数据来源：nextplatform，东吴证券研究所

3.2.4. 推出英伟达自研 CPU，补齐数据中心短板

推出自研 CPU Grace，实现英伟达在数据中心、HPC 以及计算设备上的的全自研。在 2021GTC 大会上，英伟达推出了 Grace CPU 并计划在 2023 年量产。这款 CPU 是英伟达第一次推出的 CPU 产品，采用了 ARM v9 指令集，该指令集主要是增强面向矢量、机器学习和数字信号处理器的相关内容，与数据中心所需要处理的事物息息相关，因此这款 CPU 的主要应用场景将是在数据中心领域。据英伟达宣称，Grace CPU 是高度专业化的、面向巨型人工智能和 HPC 的产品，可以训练拥有超过一万亿个参数的 NLP 模型。

自研 CPU 的主要目的是为了解决 GPU 读取内存数据的带宽瓶颈问题。英特尔的 x86 CPU 的优势是灵活的扩展性和对各类设备的支持，因此 x86 依然是目前 HPC 和服务器应用场合的重点，但 x86 架构存在带宽不足的缺点。目前 x86 CPU 通过内存控制器连接 DDR4 内存，最新的英特尔至强处理器可以实现 8 通道 DDR4 内存连接，其带宽大约为 200GB/s，GPU 本地内存（显存）的带宽在使用 HBM2 的情况下大约可以达到 2000GB/s；CPU 和 GPU 自身的连接带宽都是足够的，但是 CPU 和 GPU 连接的带宽只能依靠 PCIe 4.0 x16，带宽大约只有 16GB/s，如果考虑典型的一个 x86 CPU 带 4 个 GPU 的情形，则将一个待处理文件从内存（Memory）经过 CPU 到 GPU 的最大带宽就只有 64GB/s，这就是带宽瓶颈的由来。英伟达拥有 NVlink 技术，其带宽可达 500GB/s，但 x86 并不支持其协议，因此英伟达决定自研 CPU，来解决带宽瓶颈问题。英特尔回应称其 PCIe 总线技术也会逐步升级，但据推测在 2023 年也就是 Grace 推出的当年，PCIe 的带宽也只能达到 32GB/s（或者更进一步升级到 64GB/s），这也比 NVLink 的带宽要小一个数量级。因此，英伟达可能会重塑数据中心的底层硬件市场，进一步获取数据中心领域的优势地位。

图22：英伟达 Grace 与 GPU 配合可解决读取内存的带宽瓶颈问题



数据来源：公司官网，东吴证券研究所绘制

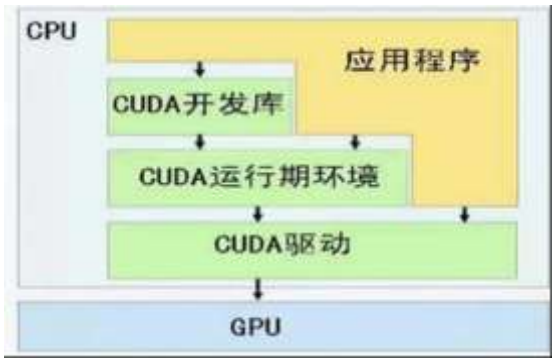
3.3. CUDA 软件生态助力 GPU 硬件，打造软硬件生态系统，形成行业壁垒

CUDA 系统助力英伟达 GPU 方便且高效地发挥其并行计算能力，使 GPU 的使用范围不仅限于显卡，而成为了 GPGPU（General-Purpose Graphics Processing Unit，图形处理器通用计算）。GPU 的微架构天生适合矩阵类并行计算，其能力不仅限于显卡领域，于是从 21 世纪早期就有专业的计算人员想要使用 GPU 做一些 AI 领域相关的并行计算。但在 CUDA 问世之前，想要调用 GPU 的计算能力必须编写大量的底层语言代码，这是主要使用高级语言为主的程序员不折不扣的噩梦。英伟达公司的 David Kirk 慧眼识珠，在他的主导下，英伟达推出了 CUDA 系统。CUDA（Compute Unified Device Architecture，统一计算架构）是一个基于英伟达 GPU 平台上面定制的特殊计算体系/算法，一般只能在英伟达的 GPU 系统上使用。CUDA 是一种类 C 语言，本身也兼容 C 语言，所以其虽然是一种独立语言，但 CUDA 本身和 C 差距不算很大，适合普通开发者使用且能够最大化 GPU 的计算效率，这使得 GPU 的使用范围不仅仅局限在显卡，而是扩展到所有适合并行计算的领域，GPU 也逐渐成为了 GPGPU。我们通过一个例子来说明 CPU、GPU 以及拥有 CUDA 的 GPU 的运算能力：比如，我们要算 100 次从 1 加到 100 的加法，如果利用一个 4 线程 CPU，需要 100/4*100=2500 次，而用 GPU（假定它是 1000 个线程），性能相同的情况下，AMD 公司的 GPU 要算 100/1000*100=10 次。如果使用 CUDA 优化的英伟达的 GPU 来计算的话，它能提供优化算法的“1+100，2+99 的这种利用首尾相加再除以 2”的方法来简化计算，那么使用 CUDA 后的英伟达显卡可能只需要计算 100/1000*100/5=2 次，可见效率提高了很多。所以，即便竞争对手的 GPU 在硬件参数上比肩英伟达的 GPU，但缺少 CUDA 的优化，其计算效率还是无法达到英伟达 GPU 的水平。

CUDA 初期采用免费推广策略，不求短期回报，使英伟达迅速占领 AI 市场。英伟达的 CEO 黄仁勋高瞻远瞩，对 GPU 的扩展应用十分重视，早在 2006 年就大力支持 CUDA 系统在 AI 领域的开发与推广，在当时每年投入 5 亿美元的研发经费（年营业额只有 30 亿美元）对 CUDA 进行不断更新与维护，并让当时美国大学及科研机构免费使用 CUDA 系统，使 CUDA 系统迅速在 AI 以及通用计算领域开花结果。

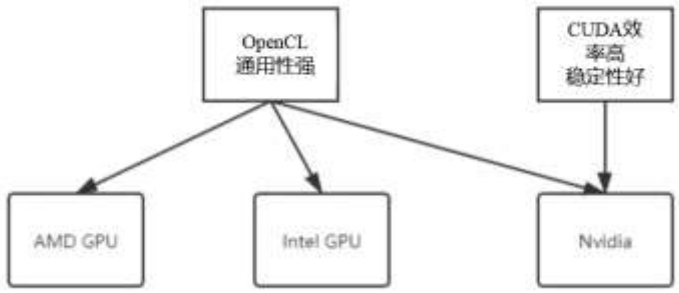
CUDA 经过多年优化，形成独特软硬件配合的生态系统，业界独此一家，产品壁垒极高。为了让广大程序员以及科研人员方便使用 GPU 的算力，英伟达不断优化 CUDA 的开发库及驱动系统。操作系统的多任务机制可以同时管理 CUDA 访问 GPU 和图形程序的运行库，其计算特性支持利用 CUDA 直观地编写 GPU 核心程序。CUDA 在软件方面组成有：一个 CUDA 开发库、一个应用驱动及其运行环境(Runtime)、两个较高级别的通用数学库，即 CUFFT 和 CUBLAS。CUDA 改进了 DRAM 的读写灵活性，使得 GPU 与 CPU 的机制相吻合。另一方面，CUDA 提供了片上（on-chip）共享内存，使得线程之间可以共享数据。应用程序可以利用共享内存来减少 DRAM 的数据传送，更少的依赖 DRAM 的内存带宽。除 CUDA 外，目前还有 OpenCL 也可以实现对 GPU 计算能力的调用，但由于其通用性较强，整体优化效果不如 CUDA，在大规模计算中劣势很大。

图23：CUDA 架构示意图



数据来源：科创实验室公众号，东吴证券研究所

图24：OpenCL 和 CUDA 的比较



数据来源：东吴证券研究所绘制

CUDA 成为连接 AI 的中心节点，CUDA+GPU 系统极大推动了 AI 领域的发展。搭载英伟达 GPU 硬件的工作站（Workstation）、服务器（Server）和云（Cloud）通过 CUDA 软件系统以及开发的 CUDA-X AI 库，为 AI 领域的机器学习（Machine Learning）、深度学习（Deep Learing）中的训练（Train）和推理（Inference）提供软件工具链，来服务众多的框架、云服务等等，推动了 AI 领域的迅速发展。因此，英伟达也被称作 AI 时代最大的推动力量。英伟达 CEO 黄仁勋 2020 年在接受 Barron 周刊的采访时也不断强调，“我们是一家拥有高性能计算的 AI 公司，视频游戏只是我们一个极为成功的应用”；“ Nvidia 不是游戏公司，它将推动下一个人工智能大爆炸”。

图25：CUDA 成为支持 AI 发展的重要力量

数据来源：公司官网，东吴证券研究所绘制

3.4. AI 的普及助力数据中心业务蓬勃发展

3.4.1. GPU 在 AI 应用领域的硬件占比逐渐增加

随着 AI 的不断普及，GPU 在云计算、工业、金融及医疗领域的硬件结构的占比会越来越多。在云计算刚刚兴起的时候，人们沿用计算时期的惯性，首先选择 CPU 来进行底层的搭建。随着 AI 等并行计算越来越流行，人们发现在 AI 等特定领域中 CPU 的

效率远不如 GPU，因此随着 AI 的不断发展，以 GPU 架构为主的硬件系统占比会不断增加。据 Yole 预测，AI 计算领域的硬件营收中，GPU 的占比会从 2019 年的 12%上升到 2025 年的 16%；而作为 GPU 的主要供应商，英伟达将会从这个趋势中受益。目前，英伟达的硬件系统已经广泛使用在金融防诈骗系统、石油开采预测系统、医疗影像识别以及云计算领域中。

3.4.2. 全球云服务提供商采用英伟达的硬件系统为其用户赋能

全球顶级云服务商采用英伟达硬件系统为其用户赋能。鉴于英伟达 GPU 在并行计算中的良好表现，亚马逊 AWS、微软 Azure、谷歌、甲骨文都纷纷采用英伟达的 GPU 进行硬件架构的搭建。英伟达的数据中心收入也快速增加，目前其营收已经可以与游戏显卡业务比肩，FY2021Q2 营收还一度超过游戏业务，成为英伟达所有业务板块中收入最高的项目，可见数据中心业务的发展势头。

英伟达积极开拓中国市场，推动中国云服务业务发展。除美国客户外，英伟达还积极拓展中国的客户。据英伟达官网报道，在 GTC China 2020 大会上，英伟达宣布阿里云、百度智能云、滴滴云、腾讯云的大型数据中心正在迁移至基于英伟达安培架构的平台，以充分利用 A100 在图像识别、语音识别以及包括计算流体动力学、计算金融学、分子动力学在内的推理和训练方面提供的速度与可扩展性。A100 不仅可以满足全球云服务提供商用户对性能的要求，而且还可以为全球用户提供强大的可靠性支持。中国云服务提供商正在采用 A100 来满足各行各业的多样化需求:

- 阿里云已经发布了基于 NVIDIA A100 打造的 gn7 GPU 系列云服务器，该产品主要面向 AI 训练和高性能计算应用，可提供新一代 GPU 计算实例。云服务器中的 8 块 NVIDIA A100 GPU 可通过 NVIDIA NVLinkTM 和 NVSwitchTM 技术实现先进的多 GPU 通信。这些 NVIDIA 技术可支持阿里巴巴 gn7 云服务器相比上一代平台实现最高 20 倍的 AI 性能，以及 2.5 倍的高性能计算速度。

- 百度智能云即将推出基于 NVIDIA A100 打造的 lgn3 GPU 系列云服务器、vGPU 云服务器以及百度太行裸金属服务器产品，该系列产品最高将搭载 8 块 NVIDIA A100 GPU，8T NVMe SSD 磁盘以及百 G 带宽，主要面向 AI 训练/推理、高性能计算应用、科学计算等场景。基于 A100 TF32 新技术，百度新一代 GPU 云服务器提供 20 倍于 V100 FP32 云服务器的计算能力。

- 滴滴云 A100 裸金属服务器配置了 8 块 NVIDIA A100 GPU 、2 颗 AMD EPYC 7302 CPU 处理器、1024GB 内存、2 个 240GB SATA SSD，以及 2 个 2T NVME SSD 磁盘，适用于 AI、数据分析、高性能计算等多种应用场景。

- 腾讯云已推出首款搭载 NVIDIA A100 的 GPU 云服务器 GT4，其搭配 AMD ROME CPU 平台，支持 PCIe 4.0 技术以及最高 180 核的 vGPU 配置。适用于深度学习训练、推理、高性能计算、数据分析、视频分析等领域，可提供更高性能的计算资源，

从而进一步降低使用成本，帮助企业、高校及研究人员聚焦模型的优化与创新。

中国 OEM 厂商致力于满足全球对搭载 A100 的 NVIDIA 认证系统不断增长的需求，包括新华三、浪潮、联想、宁畅等在内的中国领先系统制造商也在以前所未有的速度将 NVIDIA A100 GPU 引入到它们的产品中，并推出了多款针对不同应用场景的系列产品，为超大型数据中心提供基于 NVIDIA 安培架构的加速系统，进而为用户提供兼具超强性能与灵活性的 AI 计算平台。

4. 未来业务：布局自动驾驶平台化芯片，抢占智能汽车市场份额

4.1. 自动驾驶介绍

自动驾驶主要指自动驾驶汽车，也即无人车（driverless car），是一种无须人工干预而能够完成出行需求的车辆。它利用了包括雷达、超声波、GPS、计算机视觉等多种技术来感知其周边环境，通过先进的计算和控制系统，来识别障碍物和各种标识牌，规划合适的路径来控制车辆行驶。

4.1.1. 自动驾驶历史

科技巨头、独角兽公司以及整车厂纷纷开展自动驾驶研究，自动驾驶迎来快速发展的时期。自动驾驶的研究历史非常悠久，早在 1977 年时日本就有基于摄像头的自动驾驶汽车问世。但限于软硬件能力及成本的束缚，自动驾驶的发展较为缓慢。直到 2004 年美国国防部推出的 DARPA 项目，很大程度上推动了自动驾驶的复兴。现代意义上的自动驾驶技术在 DARPA 挑战赛上已经成型，参赛车辆上已经配备了激光雷达、摄像头以及分析决策系统。2005 年的 DARPA 挑战赛中，有五支队伍的参赛车辆已经可以完成限定场景的无人驾驶。目前的自动驾驶技术都是在这个基础上进行的不断升级，主要在成本优化和车规级适配性等实用性方面进行完善，不仅有各种科技巨头领导相关研究，科技独角兽公司以及整车厂也都纷纷加入这个领域，自动驾驶全面商业化的时代就要到来。

4.1.2. 自动驾驶等级分类及技术路线

目前较为通用的一种自动驾驶等级分类如下表所示：

表1：自动驾驶等级分类

自动驾驶分级	名称	定义	驾驶操作	周边监控	接管	应用场景
L0	人工驾驶	由人类驾驶员全权驾驶车辆	人类驾驶员	人类驾驶员	人类驾驶员	无
L1	辅助驾驶	车辆对方向盘和加减速中的一项操作提供驾驶，人类驾驶员负责其余的驾驶动作	人类驾驶员和车辆	人类驾驶员	人类驾驶员	限定场景
L2	部分自动驾驶	车辆对方向盘和加减速中的多项操作提供驾驶，人类驾驶员负责其余的驾驶动作	车辆	人类驾驶员	人类驾驶员	
L3	条件自动驾驶	由车辆完成绝大部分驾驶操作，人类驾驶员需保持注意力集中以备不时之需	车辆	车辆	人类驾驶员	
L4	高度自动驾驶	由车辆完成所有驾驶操作，人类驾驶员无需保持注意力集中，但限定道路和环境条件	车辆	车辆	车辆	
L5	完全自动驾驶	由车辆完成所有驾驶操作，人类驾驶员无需保持注意力集中	车辆	车辆	车辆	所有场景

数据来源：维基百科自动驾驶词条，东吴证券研究所

目前有两种自动驾驶研发思路。一种是可称之为**自上而下**的不考虑成本的研究 L4+级完全自动驾驶，代表企业有谷歌的 Waymo、通用的 Cruise、百度的 Apollo 等，目前其实现自动驾驶的系统成本在数十万到百万元人民币以上；另一种主要是车企，他们要考虑成本因素，所以一般是**自下而上**的，由低级别的自动驾驶开始逐渐提升水平，目前商业化的汽车基本上可以达到 L2 级的水平，代表企业有特斯拉、奥迪、蔚来、小鹏等。值得一提的是，本文所提到的分类级别是从法律意义上已经实现的级别而不是能力上的分类级别，也即如果是 L3 级以上的话，自动驾驶公司将为车辆的事故负责。因此本文所谓的 L3 及以上级别主要是由 Robotaxi 组成的。

4.2. 自动驾驶细分领域的市场规模

汽车市场正在经历快速的变革期，电动化是汽车升级的上半场，智能化是汽车升级的下半场。 智能化将会迎来快速发展期，主要源于以下几个方面：

- 半导体技术的提升与成本的下降：随着半导体制造商向汽车领域逐渐发展，规模化生产有利于成本的降低，从而推动销量扩大形成正反馈，汽车半导体有望复制手机半导体领域的发展规模和速度；
- 电动化的不断普及加速了智能化：电动车的电机电控特性，相较于燃油车更有助于智能化的控制系统发展；
- 对安全性便捷性和高效出行的要求：为了提升车辆差异化的竞争力，汽车厂商将继续增加在驾驶辅助系统 ADAS 方面的投入，提升自动避险刹车、自动泊车、道路领航等能力，以提升车辆的安全性与便捷性；随着自动驾驶能力的不断提高，自动驾驶将有效缓解交通拥堵，大大提高出行的效率。

我们预计，拥有智能化功能的车辆将从 2020 年的 2773 万辆增长到 2025 年的 6332 万辆。据 IDC 报告，2020 年售出的汽车中，拥有自动驾驶（辅助）功能的汽车数量（包

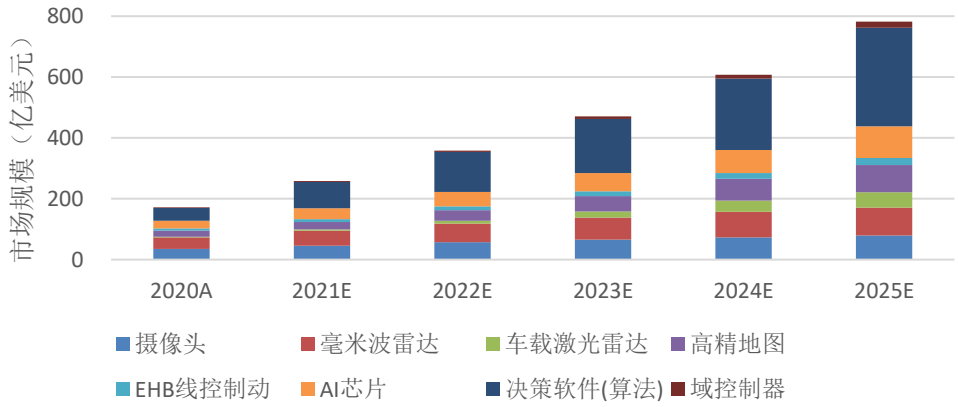
含 L1~5 级）为 2773.2 万辆，其中 L1 为 1874 万辆，L2 为 896 万辆，L3+为 3.2 万辆。我们根据市场智能化趋势以及前几年的增速为基础进行测算，到 2025 年，拥有自动驾驶（辅助）功能的汽车数量（包含 L1~5 级）为 6332 万辆，其中 L1 为 3900 万辆，L2 为 2320 万辆，L3+为 112 万辆；2020~2025 的 CAGR 为 17.8%。

图26：自动驾驶车辆年出货量预测（万辆）

数据来源：Yole，东吴证券研究所测算

到 2025 年时，与单车自动驾驶相关的革新性部件，其市场总额可达 781 亿美元，2020~2025CAGR 可达 35.8%。巨大的市场增量使得相关公司都希望能够乘着智能化升级的东风扩大公司业务，占领市场空间。我们将与自动驾驶有关的市场进行拆分，主要有八个模块，其中与人工智能息息相关的决策软件、**AI 芯片**以及传感器（摄像头、激光雷达、高精地图、毫米波雷达）的发展空间更大。

图27：自动驾驶的细分领域市场规模测算



数据来源：Yole，东吴证券研究所测算

4.3. 积极入局汽车芯片领域，成为平台化芯片的领导者

4.3.1. 从移动业务起家，逐渐扩大应用市场

在智能手机兴起的 2008 年时，英伟达试图进入移动芯片市场。为此，公司开发了 Tegra 系列芯片，采用了 ARM 的 CPU 架构，并集成了自家的 GPU 芯片，组成了一套 SOC（system on a chip）系统。早期的 Tegra 芯片注重功耗及效率的表现，主要用在微软

的一款 MP3 和 Kin 手机上;后期则更专注于提供高性能,其典型产品是任天堂的 Switch,英伟达的 Tegra X1 给任天堂 Switch 带来了极高的画面体验。由于自动驾驶中对于画面的实时处理要求很高,因此后续的 Xavier 以及 Orin 系列也开发了相应的车规级芯片。从移动芯片的发展轨迹来看,英伟达的 CUDA 核心数量也快速增长,RAM 的容量和带宽也迅速提高,移动芯片的性能始终保持竞争优势。

表2:英伟达移动芯片发展历程

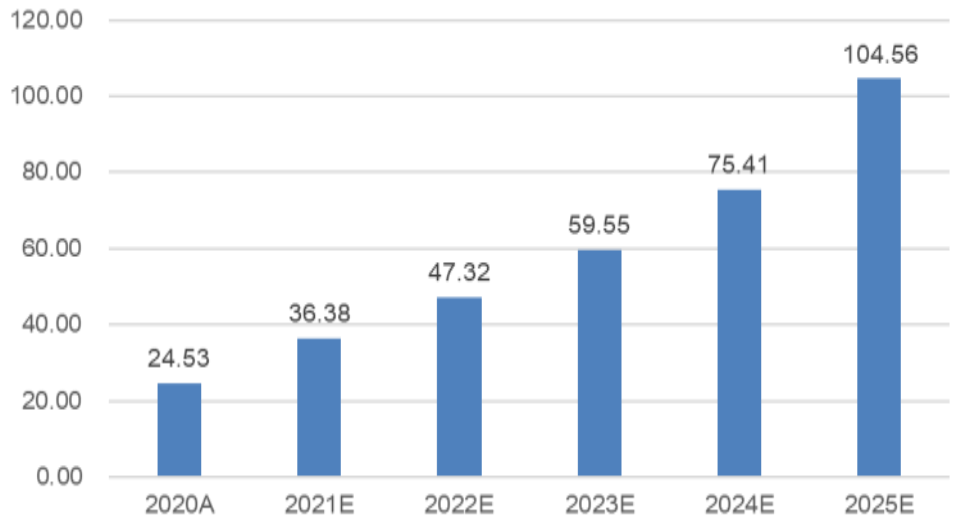
芯片名称		Tegra 2	Tegra 3	Tegra 4	Tegra 4i	Tegra K1	Tegra X1	Tegra X2	Xavier	Orin	Atlan
CPU	指令集	ARMv7-A (32 bit)				ARMv8-A (64 bit)			ARMv8.2-A (64 bit)		ARMv9 (64 bit)
	内核	2 A9	4+1 A9	4+1 A15	4+1 A9	4+1 A15 2 Denver	4 A53 + 4 A57	2 Denver + 4 A57	8 Nvidia Carmel	12 Arm Cortex-A78AE	Nvidia Grace-Next
	L1级高速缓存	32/32 KB				128/ 64 KB	32/ 32 KB + 64 / 32 KB	128/64 KB + 48 / 32 KB	64/64 KB	/	/
	L2 级高速缓存	1 MB		2 MB			128 KB + 2 MB	2 MB + 2 MB	8 MB	/	/
	L3 级高速缓存	NA							4 MB	/	/
GPU	架构	Vec4				Kepler	Maxwell	Pascal	Volta	Ampere	Ampere-Next
	CUDA核心数	4+4	8+4	48+24	48+12	192	256		512	2048	/
RAM缓存	协议	DDR2	DDR3/DDR2	DDR3			LPDDR3/LPDDR4	LPDDR4/LPDDR4X		LPDDR5	LPDDR5X
	容量	1 GB	2 GB	4 GB	4 GB	8 GB	8 GB	8 GB	32 GB	/	/
	带宽	2.7 GB/s	6.4 GB/s		7.5 GB/s	14.88 GB/s	25.6 GB/s	59.7 GB/s	136.5 GB/s	200 GB/s	/
制程		40 nm		28 nm HPL	28 nm HPM		20 nm SOC	16 nm FF	12 nm FFN	/	/
上市SOP时间		2010		2013		2014	2015		2016	2022E	/
代表产品		2009年微软Zune、Kin手机;2012款奥迪车载影音;2012版特斯拉Model S车载系统					任天堂Switch;2016版特斯拉		汽车自动驾驶平台		

数据来源:公司官网,维基百科英伟达词条,东吴证券研究所整理

4.3.2. AI 芯片逐渐专业化,平台化芯片发展空间更广

AI 芯片为自动驾驶提供算力保障。随着图像/视频和雷达等传感器接受的数据量越来越大,对视觉芯片的实时算力要求也越来越高,据估计满足安全冗余的 L2 级的算力要求至少需要 10TOPS(INT 8)以上,传统的 MCU(Microcontroller Unit,微控制单元,也称为单片机)算力最多只能达到 GOPS(比 TOPS 小一千倍),完全不能满足图像识别的算力要求。为了满足自动驾驶的需求,多家芯片厂商开发出了针对车载市场优化的 AI 芯片。当前主流的车载 AI 芯片按架构主要分为三类:GPU、FPGA、ASIC。其中 GPU 通用性较强因而场景适应性强,但功耗相对较高。FPGA 运算速度快,通用性弱于 GPU 但功耗优于 GPU,因其易修改,主要用途是做 ASIC 的验证版本。ASIC 属于为 AI 特定场景定制的芯片,通用性低但针对特定场景的每瓦功耗以及安全性更好,属于最终阶段的产品,开发成本较高。我们测算,汽车领域的 AI 芯片市场规模将从 2020 年的 25 亿美元增长到 2025 年的 105 亿美元,2020~2025 的 CAGR 达 44%。

图28：车载 AI 芯片的市场规模预测（亿美元）



数据来源：Yole，东吴证券研究所测算

以英伟达为代表的平台化芯片的发展空间更为广阔。目前芯片的解决方案主要有为提供软硬件整套解决方案、传统汽车电子厂商转型、平台化芯片以及整车厂自研四种模式。特斯拉可类比为手机界的苹果，核心的 AI 芯片以及相应的算法均自研，但由于芯片设计等要求非常高，不仅需要投入大量时间、资金，还需要有相应的技术人才支撑，对于大多数整车厂来说很难实现；Mobileye 以自研算法起家，早期与意法半导体合作研发芯片，后来被 Intel 收购后，形成了软硬件一体化的能力，因此 Mobileye 是以整套解决方案的模式向整车厂兜售，其优点是可靠性强且整车厂使用方便，但缺点是整车厂获得的是封闭的算法系统，无法自研算法，因此被特斯拉、小鹏、蔚来等希望掌握算法能力的整车厂所弃用；平台化芯片以英伟达为代表，目前市场上还有高通、地平线、华为和黑芝麻等厂商，这个方案的思路是提供平台化芯片以及算法开发工具链（包括示例算法），整车厂可根据自身软件研发能力自行选择从哪个层面开始进行软件/算法的研发，自由度较大，因此受到了以小鹏、蔚来、理想、百度、小马智行以及 AutoX 为代表的整车厂和科技公司的欢迎，类比来看，平台化芯片类似于智能手机领域的高通和联发科，市场空间较一体化自研的苹果大；除平台化芯片外，市场上还有传统的汽车电子厂商瑞萨、恩智浦以及德州仪器等，也开始纷纷布局高算力的车载 AI 芯片，但以目前推出的产品来看，其芯片算力相对较低，且单瓦功耗也比较大，客户主要是传统的 Tier 1 厂商以及部分科技公司。

表3：车规级 AI 芯片的解决方案分类

技术路线	厂家	芯片	架构	车辆配备/合作商	INT8算力(TOPS)	算法支持	每瓦功耗(W)	制程(nm)	SOP时间	优劣势
提供整套解决方案	Mobileye（英特尔）	EyeQ3	CPU+ASIC	奥迪A8/沃尔沃/凯迪拉克	0.256	自带算法，算法一般是封闭的。目前声称提供修改工具，客户可进行部分优化	10	40	2014	**视觉算法领先**，2020年市占率70%，车载市场经验丰富，产品即装即用，提供全套解决方案，稳定性高；但是较为封闭的算法令车企担忧，因为车企逐渐意识到**算法为自动驾驶的核心能力，有部分车企因此转向别家芯片**；行车**数据的归属权**不明，容易引发争议
		EyeQ4		蔚来/理想/大众/宝马/福特/日产/广汽/长城等主力在售车型	2.5		1.2	28	2018	
		EyeQ5		宝马iNext/极氪001	24		0.416	7	2021E	
传统汽车电子厂商转型	瑞萨	V3H	CPU+ASIC	博世/海拉	4	提供硬件平台，提供算法相关支持	2.5	16	2019	具有成本优势，产品可靠性高，主要合作方是Tier1厂商或是第三方科技公司；产品面向的是量产车辆的驾驶辅助，算力不高，架构优化不足，每瓦功耗较高
	恩智浦	S32V		RTI（软件公司）	4		1.5	16	2022E	
	德州仪器	TDA4VM		百度（威马W6）	8		1.5	16	2020	
平台化硬件	英伟达	Xavier	CPU+GPU+ASIC	小鹏P7/P5	30	提供工具链和软件算法参考模型，客户自定义算法	1	12	2020	视觉芯片架构与英伟达GPU类似，所以英伟达很适合转型做视觉芯片。芯片的算力目前属于开放平台中最高的，下一代的客户越来越多，预计增速会越来越快
		Orin		蔚来ET7/下一代小鹏/下一代理想/沃尔沃XC90/上汽RES33/奔驰	200		0.225	7	2022E	
		Atlan		奔驰	1000		--	5(E)	2023E	
	地平线	征程2	CPU+ASIC	长安UNI-T/奇瑞蚂蚁/上汽智己	4		0.125	28	2019	2016年成立，以AI芯片研发起家，灵活性高，在某些算法方面有独特优势，深耕中国市场，可以为中国企业提供**更快更便捷**的服务支持
		征程3		江淮/理想One	5		0.5	16	2020	
		征程5/5P		长城	96/128		0.195	7	2022E	
	华为	昇腾310	CPU+ASIC	/	16		0.5	12	2018	软硬件实力一流，硬件调教能力出众，可做到高算力、高可靠、高能效和低延时。计划做万物互联的生态，可扩展性很强。芯片制程受限，未来产量无法确定。
		昇腾910		北汽/长城	640		0.48	/	2022E	
	黑芝麻	A1000	CPU+ASIC	一汽红旗、上汽	40		0.2	16	2020	2017年成立，企业创新活力强。拥有较高的芯片算力能力，车载经验丰富，深耕中国市场，可以更好服务中国的车企
		A1000 Pro		东风	106		0.24	16	2022E	
全栈自研	特斯拉	FSD	CPU+GPU+ASIC	Model3/S/X/Y	72	自研	1	14	2019	特斯拉**全栈自研**，硬件软件适配性强，算力可充分发挥；算法可根据真实行车数据进行快速迭代，拥有其他公司无可替代的上下游一体化优势
		升级版FSD		Model3/S/X/Y	210		--	5/7	2021E	

数据来源：各公司官网，佐思汽研，东吴证券研究所整理

4.3.3. 整合移动芯片的车载 AI 芯片平台，成为平台化芯片的代表

产品自由度高，客户可根据需求选择合适的芯片平台方案。在 Tegra 系列芯片的基础上，英伟达集成了一些特殊功能的 GPU 以及辅助芯片，推出了英伟达 Drive 系列车载 AI 芯片平台。早期的车载 AI 芯片平台与单个移动芯片差别不大，但随着车载系统的要求不断多样化，英伟达 Drive 系统也增加了很多选择。例如 Drive PX Xavier 仅配备了一块 Xavier 芯片，其算力为 30 TOPS，功耗仅为 30W，适合用在 L2 级的量产车型中，例如小鹏 P7 就采用了此款车载芯片平台；对于 L4 级车辆的车载 AI 芯片平台，仅仅一个 Xavier 芯片算力不够，因此采用了两个 Xavier 芯片加上两个图灵架构的 GPU，使算力达到了 320TOPS，其功耗也增加到了 500W；蔚来希望打造自己的计算平台，因此从英伟达这里选购的是独立的 Orin 芯片。不同的客户可以依照不同的使用场景选择适合的产品，这极大地增加了英伟达车载 AI 芯片的使用场景。

表4：英伟达车载 AI 芯片平台发展历程

车载系列名称	Drive CX	Drive PX	Drive PX 2 (Auto Cruise)	Drive PX 2 (Tesla)	Drive PX 2 (Auto Chauffeur)	Drive PX 2 (Tesla 2.5)	Drive PX Xavier	Drive PX Pegasus	Drive AGX Orin	
发布时间	2015年1月		2016年9月	2016年10月	2016年1月	2017年8月	2017年1月	2017年10月	2019年12月	
芯片构成	1* Tegra X1	2* Tegra X1	1*Tegra X2 (Parker) + 1*Pascal GPU		2* Tegra X2 (Parker) + 2* Pascal GPU	2* Tegra X2 (Parker) + 1x Pascal GPU	1* Tegra Xavier]	2* Tegra Xavier + 2* Turing GPU	2* Tegra Orin	2* Tegra Orin + 2* Ampere GPU
算力	/	/	4 FP32 TOPS	4 FP32 TOPS	8 FP32 TOPS	4 FP32 TOPS	30 INT8 TOPS	320 INT8 TOPS	400 INT8 TOPS	2000 INT8 TOPS
功耗	/	20W	40W	40W	80W	60W	30W	500W	130W	750W
代表产品	/	/	/	2016款特斯拉	英伟达自动驾驶训练	/	2020款小鹏 P7	英伟达自动驾驶训练	2022款蔚来ET7	英伟达自动驾驶训练

数据来源：公司官网，维基百科英伟达 Drive 词条，东吴证券研究所整理

4.3.4. 软件安全性高，易于上手且生态丰富，助力 AI 芯片占领市场

不仅算力领先，英伟达易于上手的软件工具链极大地了方便了芯片使用者的开发过程。同数据中心基础芯片类似，英伟达十分重视对软件工具链的开发。英伟达不仅花费了大量的研发资金，成立了测试小组专门改装了车辆以提高英伟达的芯片及相关软件工具链的安全性与稳定性，还积极听取客户的意见并对相关要求作出回应。在不断的测试中，软件工具链的可用性也不断提高。安全、可靠且易用的软件工具链不仅可以让软件开发人员快速上手并熟练掌握芯片的调用技巧，还可以保证软件的不会在汽车这个安全性要求极高的领域出现差错，这也是整车厂采用英伟达方案的主要原因之一。英伟达的软件还有一个特点是其软件开放性高。有丰富软件开发能力的客户可以从底层操作系统开始自行研发，而初入此领域的客户可以从较上层的应用软件开始研发，底层使用英伟达搭建的通用系统。英伟达灵活的使用方案适配性强，潜在客户数量巨大。

图29：英伟达汽车软件相关的支持模块

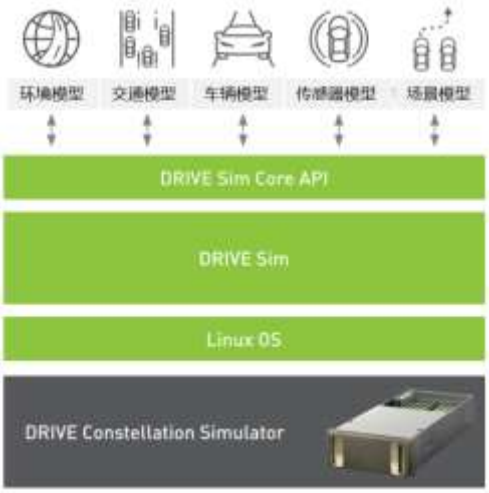
数据来源：东吴证券研究所绘制

4.3.5. 开拓自动驾驶虚拟测试平台，降低自动驾驶设计门槛

除平台化芯片外，英伟达也积极推广虚拟测试平台 Constellation。NVIDIA DRIVE Constellation 是数据中心解决方案，集成了功能强大的 GPU 和 DRIVE AGX Pegasus。

在 GPU 上运行的高级可视化软件模拟输入到 Pegasus 的摄像机、普通雷达和激光雷达数据，而 Pegasus 对这些数据进行处理，就好像它真的在路上行驶一样。这个可扩展系统能够生成数十亿英里的不同自动驾驶汽车测试场景，用于在部署之前对“硬件在回路”和“软件在回路”进行验证，极大减轻自动驾驶的初期开发成本。这个虚拟平台包括环境测试、车流测试、车辆测试、传感器测试以及超车模型测试等等。当然，目前的模拟测试还无法替代真实路况测试，但随着英伟达模拟功能的不断完善，其测试能力也会逐渐提高。

图30：英伟达自动驾驶虚拟平台系统示意图

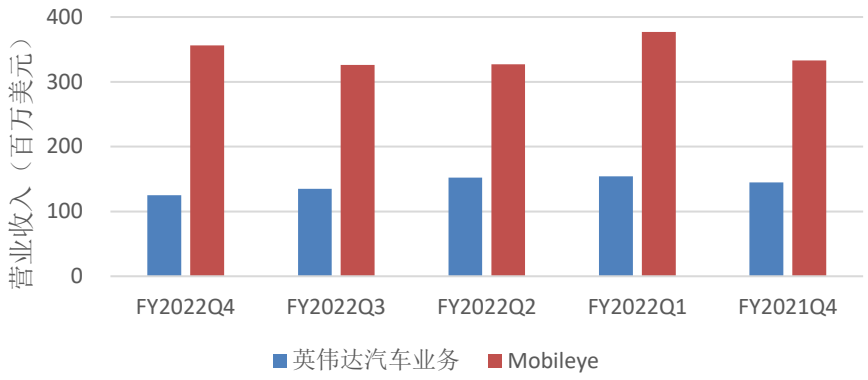


数据来源：东吴证券研究所绘制

4.4. 汽车业务营收稳定增长，平台化芯片市场空间更大

英伟达汽车业务目前仍落后于 Mobileye。 截至到 2021 年 7 月，在车载汽车芯片市场，按营收来看，只有 Mobileye 和英伟达两大巨头，两者的营业收入可占整个市场收入的 90%。Mobileye 以整体解决方案的形式较早占领市场，英伟达以平台化芯片形式逐渐获得了希望自研自动驾驶算法的客户的青睐。从营收来看，英伟达在 FY2021 Q3（公历 2020 年 8~10 月）以来收入增长较慢，与 Mobileye 差距拉大，主要与英伟达客户多为新势力车厂，受新冠疫情影响其需求波动性较大导致。但从长期来看，希望算法自研的整车厂会越来越多，英伟达的平台化模式的优势会逐渐显露出来。

图31：英伟达汽车业务与 Mobileye 的营收对比



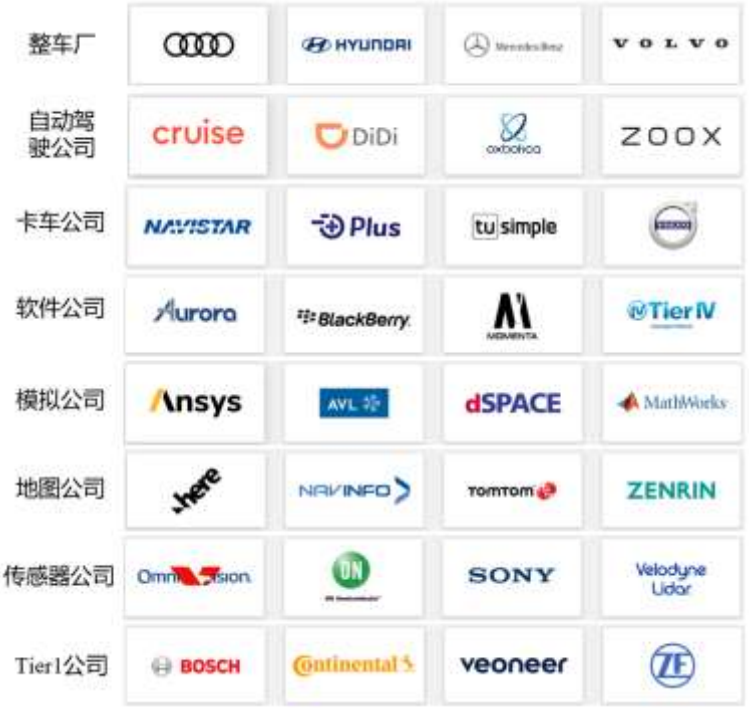
数据来源：公司财报，statista，东吴证券研究所

注：英伟达 Q1 为 2~4 月，FY2022 对应公历 2021.1~2022.1，Mobileye Q1 为 1~3 月。

自动驾驶的决策算法是自动驾驶的核心竞争力，能够进行算法自研的英伟达平台化芯片更受欢迎。在家用车领域，沃尔沃、奔驰以及中国的造车新势力蔚来、小鹏与理想和传统整车厂上汽集团都选择和英伟达合作，推动汽车的智能化。值得一提的是，造车新势力早期版本的车辆都采用了 Mobileye 的芯片，但由于无法自研算法，于是都转向了英伟达。小鹏的 P7 是中国最早的搭载 Drive Xavier 车载芯片的量产车型，于 2020 年 7 月问世；由于英伟达车载芯片的良好编程平台基础，小鹏 P7 得以在短时间内数次 OTA 升级，向用户推出了高速领航辅助驾驶 NGP（Navigation Guided Pilot）以及不依赖停车场改造的自主泊车功能，使车辆用户不断体验到最新的功能，也促进了汽车的销量。在商用车领域，英伟达也收获了新的合作。专注于无人出租车 Robotaxi 的 AutoX 公司使用英伟达的车载芯片系统实现了 L4 级功能，专注于卡车领域的智加科技也宣布，即将交付给亚马逊物流的 1000 辆自动驾驶卡车也将采用英伟达的车载芯片系统。据英伟达在 GTC2021 大会透露，**其自动驾驶在手订单达 80 亿美元，**可见其芯片平台的受欢迎程度。

英伟达非常重视汽车业务的发展，积极布局汽车领域上下游合作。2017 年，黄仁勋曾把英伟达的未来押注在 AI 上，把它称之为一家 AI 公司。事实证明，他的判断成就了现在的英伟达。如今，黄仁勋又把眼光看向了汽车行业和自动驾驶。2021 年 1 月，他在与欧洲《汽车新闻》交谈中曾预测，到 2030 年将有 20% 的汽车实现高级自动驾驶，其中大部分会采用英伟达技术。黄仁勋说：我们不是一家汽车制造商，我们是科技创造者。他表示，有的客户想只购买英伟达的计算解决方案，软件完全完全自己开发，而有的客户希望英伟达能提供完全的堆栈。对于这两种合作方式，英伟达都欢迎。目前，汽车业务营收在英伟达总营收中占比还非常小，但黄仁勋看得足够长远：“我们的优势是很有耐心，这需要公司有很大的决心、持久力，以及具有影响力的核心技术，我们着眼于长远发展。”随着英伟达在汽车领域布局的不断深入，其合作伙伴已经深入到整车厂和上游的软硬件公司，英伟达在汽车领域的影响力在不断的扩大。

图32：英伟达汽车业务的合作伙伴



数据来源：公司官网，东吴证券研究所绘制

5. 未来业务：Omniverse—制定通用标准，打通不同设计平台，成为元宇宙平台级应用

NVIDIA Omniverse 是一个易于扩展的开放式平台，专为虚拟协作和物理级准确的实时模拟打造。创作者、设计师、研究人员和工程师可以连接主要设计工具、资产和项目，从而在共享的虚拟空间中协作和迭代。开发者和软件提供商还可以在 Omniverse 的模块化平台上轻松地构建和销售扩展程序、应用、连接器和微服务，以扩展其功能。Omniverse 易于扩展并支持多 GPU，基于 Pixar 的 Universal Scene Description (USD) 并由 NVIDIA RTX 技术提供支持，能够简化和加速复杂的 3D 工作流程。与传统三维设计制作流程不同，Omniverse 并不需要多个流程分开贴图、渲染最后出图，它提供实时协同工作的平台，可以让建模师、灯光师、特效师等每个部门专业人员无缝协同工作。

我们认为，英伟达希望用 Omniverse 来复制 CUDA 的成功经验，使英伟达成为未来元宇宙时代（虚拟世界）软硬件一体化的基石性公司，给英伟达带来新的收入增长点。英伟达早在 2006 年就前瞻性地投入资金开发了 CUDA 系统，可以方便开发者调用 GPU 资源来进行 AI 模型训练以及科学计算等领域，英伟达的 GPU 也成为 AI 领域不可或缺的基础，其以 CUDA 为基础的数据中心业务营收在 2019-2021 年近三年已经占公司营收的 40%以上。而 Omniverse 将虚拟世界的设计门槛大大降低，这有助于 UGC（User

Generated Content，用户生成内容）的形成和生态系统的建立，且将和英伟达的数据中心业务和云游戏业务产生联动，使英伟达介入更多软件层面的业务，形成软硬件联动，筑牢虚拟世界的壁垒。

5.1. Omniverse 迭代历史

2020 年 12 月，英伟达发布 Omniverse 公测版并发布多个 APP，包括专为建筑、工程和施工专业人员设计的 Omniverse View；专为媒体、娱乐和制造/产品设计行业的设计师、创作者和专家设计的 Omniverse Create；以及专为 3D 深度学习研究人员设计的 Omniverse Kaolin。

2021 年 2 月，Autodesk 3ds Max Connector 在 NVIDIA Omniverse 上推出，将 Autodesk 3dx Max 中的 3D 建模、可视化等实时同步到 Omniverse 中。

2021 年 4 月，英伟达推出面向企业的 Omniverse 设计协作和模拟平台。包含 NVIDIA Omniverse Nucleus 服务器和 NVIDIA Omniverse Connectors，两个终端用户应用：Omniverse Create，可加速场景构成，用户可通过实时互动来装配、点亮、模拟和渲染场景；Omniverse View，支持无缝设计协作，并能通过逼真的渲染技术实现建筑和工程项目的可视化。同时包含 vWs 软件，让协作者在任何地方自由运行各类图形密集型 3D 应用。

2021 年 8 月，英伟达更新 Omniverse，三管齐下，增加 Blender USD 支持、Adobe Substance 3D 插件和 GANverse3D 扩展工具。推出基于 Omnviserse 的新的 DRIVE Sim。

2021 年 11 月，英伟达 Omniverse Enterprise 正式发布，推出 Omniverse Replicator，一种生成具有正确标注的合成数据的引擎，用于训练 AI 网络。缩小仿真到真实的域差距。推出 NVIDIA Omniverse Avatar 创建 AI 虚拟形象的平台。推出其他新功能 NVIDIA CloudXR，Ominverse VR，Omniverse Remote，Omniverse Farm，Omniverse Showroom。

5.2. Omniverse 的组成

整体来看，Omniverse 有三块内容，第一块是 Connect 和 Nucleus，通过这两个模块可以实现数据在不同平台的同步设计和实时渲染。第二块是 Simulation 和 Renderer，这两个模块是英伟达技术的集成，通过这两个模块实现虚拟世界的物理级仿真和渲染，是实现数字孪生的底层工具。最后一块是 KIT，它是搭建 Omniverse 的基础包，Omniverse 即是通过 KIT 搭建而成，而 KIT 的底层代码是用 Python 来写的，这使得开发者可以更方便地开发个性化的 Omniverse 模块。

图33：Omniverse 的组成

数据来源：公司官网，东吴证券研究所

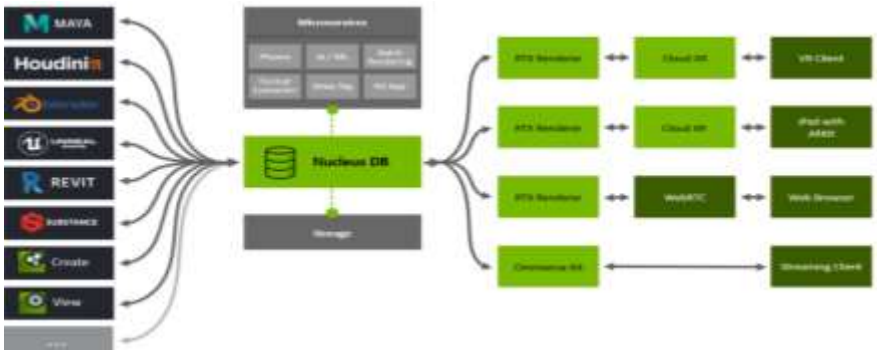
5.2.1. Omniverse Connect，以插件分布连接 Nucleus

Omniverse Connect 库以插件的形式分布，使客户端应用能够连接到 Nucleus。完成必要的同步后，DCC 插件将使用 Omniverse Connect 库应用从外部接收的更新，并在必要时发布内部生成的更改。NVIDIA 已经在开源的 USD 发行版上构建了扩展和附加的软件层，使得 DCC 工具和计算服务可以通过 Omniverse Nucleus DB 轻松地相互通信。这些扩展和添加以及利用它们的应用程序插件统称为 NVIDIA Omniverse Connect。

5.2.2. Omniverse Nucleus，数据库与协作引擎链接多名用户

Omniverse Nucleus 是 Omniverse 的数据库和协作引擎。通过 Omniverse Nucleus，团队可以将多个用户连接在一起，同时使用多个应用程序。这允许 人们使用他们最舒适和最快的应用程序，并为快速迭代打开了许多大门。为此，Omniverse Nucleus 提供了一组基本服务，这些服务允许各种客户端应 用程序、渲染器和微服务共享和修改虚拟世界的表示。

图34：Nucleus 实现用户实时协作



数据来源：公司官网，东吴证券研究所

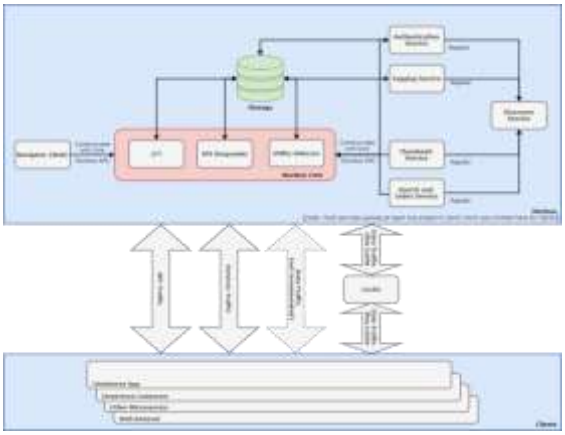
Nvidia Omniverse Nucleus 提供的相关服务将允许任何应用或者渲染器的使用者同步地对虚拟时间和相关渲染文件进行编辑和修改。比如通过 Connector 将 Unreal Engine 和 3DS MAX 文件同步到 Omniverse Nucleus 后，再将这两个平台的文件以 Omniverse View 或者 Omniverse Create 的方式进行打开和编辑。

Nvidia Omniverse Nucleus 具有两个的两个特点：

1. Nucleus 具有应用程序之间的高效协作和实时同步、用户和群组管理、保留所有更改历史、控制访问列表和权限管理、安全的传输协议、边缘缓 存与备份等功能，具备数据库的基础和高级功能。

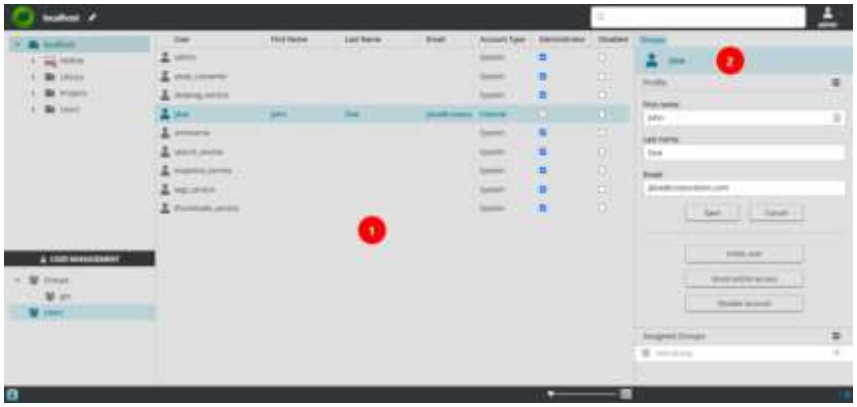
2. Nucleus 是一组服务的集合，利用其复杂的架构，同时支持 Windows 和 Linux，可以在网络上使用，并允许 Client 层应用程序连接到这些服务。在 Nucleus 中，每个组件都可以与多个其他组件相互通信。

图35：Nucleus 内部架构



数据来源：公司官网，东吴证券研究所

图36：Nucleus 用户权限管理

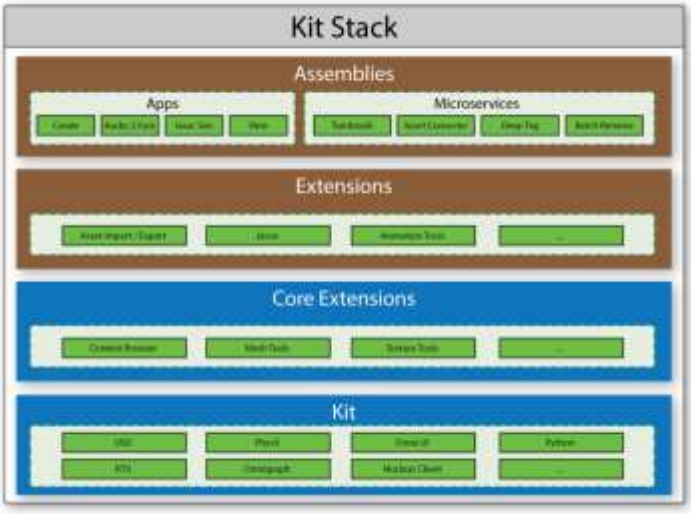


数据来源：公司官网，东吴证券研究所

5.2.3. Omniverse Kit，基于 USD 构建的工具包

Omniverse Kit 是基于 USD 构建的工具包，集合了英伟达所有技术，用于创建能够解决实际问题的应用程序。它还可以用于开发用户自己的 Omniverse 应用程序。使用 Python 编写脚本，以 C++为核心提高效率，同时随附完整源代码，开源进行远程控制、可帮助开发者轻松搭建各种模拟组件和应用程序。 Kit 核心是高度模块化。Kit 中包含许多用于执行 PhysX、渲染、计算机图形等 SDK，用于协调本机存储的 USD 数据对其进行操控。Kit 中应用程序由扩展程序组成，扩展程序是 Kit 构建块，代码完全使用 python 编写，扩展程序还包括图标、图形、配置等。Kit 可以独立使用 Nucleus 文件路径字段、文件网络、树状视图、添加搜索，连接到 Nucleus AI 索引服务的单独扩展程序，基本组件主要由 USD/Hydra、Omniverse Client Libaray、Carbonite、Omniverse RTX Render、Scriping、UI 构成。 n Extensions 是基于 Omniverse Kit 的应用程序的核心构建块。Extensions 分为核心扩展和由核心扩展构成的其他扩展，核心扩展包括 Viewport、Stage、Layers、 Content Browser、Version Control、Details Panel、Extensions Manager、Movie Capture、Audio、Saved Layouts 等程序，实现了窗口界面、加载卸载扩展程序、声 音支持等多种程序基本功能。 n 在 Core Extensions 支持下，实现了 Simulation Extensions、Animation Extensions、Design Extensions、Omniverse Remote Extensions、AI ToyBox Extensions 、Utility Extensions 等多种扩展，构建了 Isaac、Create 等多种应用。

图37：Omniverse Kit 构成



数据来源：公司官网，东吴证券研究所

表5：Omniverse Kit 主要组成

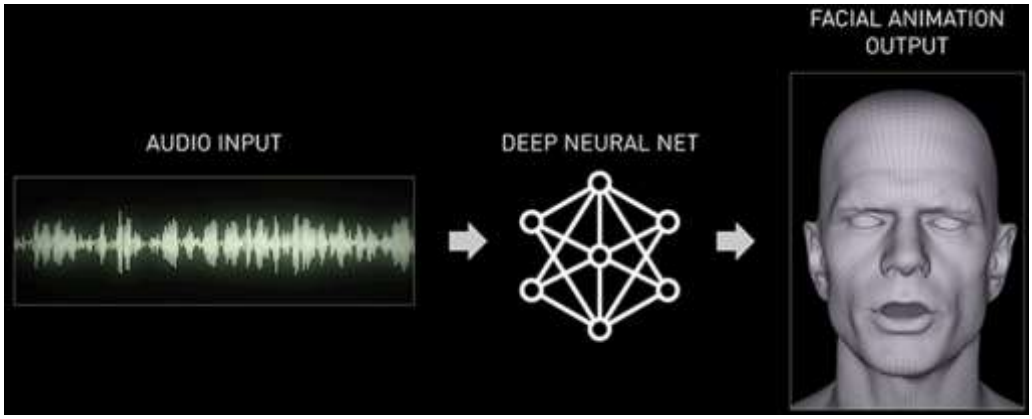
USD/Hydra	Omniverse Client Library	Carbonite	Omniverse RTX Render	Scriping	UI
USD 是 Kit 使用的主要场景描述，可以通过外部共享库直接访问。Pixar 的 Hydra 是渲染工具。**任何兼容 Hydra 的渲染器都可以连接到 Omniverse Kit 渲染窗口。**	**Omniverse 客户端(如 Kit)在加载和保存 Assets(如 USD, MDL 和纹理)时使用的与 Omniverse 服务器和本地文件系统通信的库。**如用于在 Omniverse Nucleus 服务器上读/写/复制数据/文件和文件系统类查询的 API。	**Omniverse Kit 基于 Carbonite 基础框架而构建**，Carbonite 插件都是使用 C 接口编写的，以实现持久的 ABI 兼容性。该框架可通过一组轻量级扩展程序提供插件管理、输入、文件访问、持久设置管理、音频、资产加载和管理、线程和任务管理、图像加载、本地化、同步和基本窗口等各类功能。	**Pixar 的 Hydra 被用来连接 USD 和 RTX。**因为 Kit 需要支持大量的渲染器，多个自定义场景委托，多个 Hydra 引擎和一系列其他需求，**在 Kit 应用程序中提供一个带有 Gizmos 和其他控件的 Viewport，所有渲染在高帧率下异步。**	**Kit 附带一个 python 版本。可以在基于 Kit 的应用程序中运行任意的 python 脚本。**	**Omniverse Kit 带有一个默认的 UI，允许它作为一个 USD 检查器，编辑器，布局工具和查看器。**

数据来源：东吴证券研究所

5.2.4. Audio2Face：基于 Omniverse Kit 的面部动画生成技术

Omniverse Audio2Face 是一款由 AI 支持的应用程序，仅从一个音频来源即可生成和音频同步的对唇面部表情动画（由 AI 提供技术支持）。可简化 3D 角色的动画制作，以与任何配音音轨匹配，方便游戏电影和实时数字助理制作动画角色。Audio2Face 还提供了一个完整的字符传输管道，为用户提供了一个简化的工作流，使他们能够使用 Audio2Face 技术驱动自己的字符。数据转换选项卡还提供了各种输出格式——包括连接自定义的 Blendshape 网格。 Audio2Face 工作原理：Omniverse Audio2Face 应用程序基于原始 NVIDIA 研究论文。Audio2Face 预装有“Digital Mark”。这是一个 3D 角色模型，可以使用用户的音轨进行动画处理，因此入门步骤非常简单。只需选择音频并将其上传到应用程序即可。该技术将音频输入到预先训练的深度神经网络中，而网络输出会实时生成角色的面部动画。用户可以通过编辑各种后处理参数来编辑角色的性能。然后，网络输出会生成角色的 3D 顶点网格，以创建面部动画。用户在此页面上看到的结果主要为 Audio2Face 的原始输出，很少或没有编辑过后处理参数。 Audio2Face 支持音频输入、角色转换、实例多样、数据转换等多种丰富的功能，为用户带来多样体验。

图38：Audio2Face 功能示意图



数据来源：公司官网，东吴证券研究所

5.2.5. Isaac Sim：基于 Omniverse Kit 的 AI 机器人模拟仿真平台

NVIDIA Isaac Sim 由 Omniverse 提供支持，是一款可扩展的机器人仿真应用程序和合成数据生成工具，为逼真、精确的虚拟物理环境提供支持，以开发、测试和管理 AI 机器人，拥有生成合成数据、模拟操作、模拟导航、机器人/资产导入四大功能。

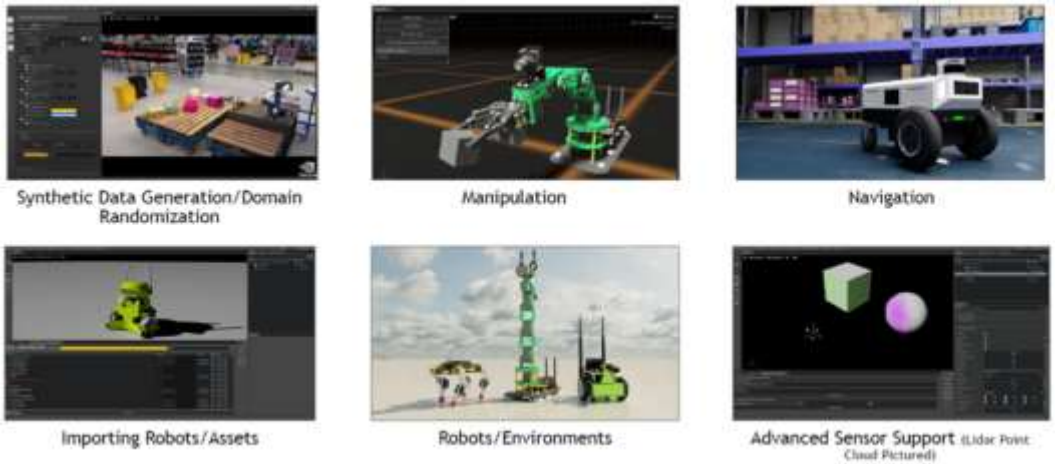
1. 生成合成数据：训练感知模型需要大量多样化的数据集。对这些数据集进行合成非常耗时。通过在 Isaac Sim 中使用 Omniverse Replicator，开发者可以引导训练任务。在项目的早期阶段，合成数据可以加速概念验证或验证 ML 工作流。在开发周期的后期阶段，可以用合成数据扩充真实数据，以减少训练生产模型的时间。Isaac Sim 内置了对领域随机化的支持，允许在纹理、颜色、光照和位置上进行改变。

2. 模拟操作：机器人的关键应用领域之一是能够识别、拾取和移动物体的机械手。在现代工厂或仓库环境中，机械手可以大大提高处理和分拣物料的效率和吞吐量。Issac Sim 内置了一些常见任务的例子，比如填充垃圾箱和堆叠垃圾箱。这些基于 python 的示例可以修改为适用于用户自定义的任务。

3. 模拟导航：自主移动机器人必须能够在其环境中从 A 点移动到 b 点。这种能力是由导航堆栈实现的。Isaac Sim 支持开发和测试机器人导航能力。Isaac Sim 提供了一个完整的例子。

4. 机器人/资产导入：将机器人模型和其他资产导入机器人模拟器至关重要，并且在设置培训或测试场景时是一项重大挑战。利用 Omniverse 中内置的强大连接器功能，Isaac Sim 内置了对流行产品设计格式的支持。 高级 URDF 导入器已在机器人模型上进行了测试。 此外，CAD 文件可以直接从 Onshape 和 STEP 文件导入，只需少数后处理即可。

图39：Isaac SIM 可以完成的物理模拟场景



数据来源：enterpriseai，东吴证券研究所

5.2.6. Omniverse Create，基于 Kit 加速高级场景合成

Omniverse Create: 是一款基于 Omniverse Kit 构建的应用程序，它能够加速高级场景的合成，同时允许使用者以交互方式实时组装、模拟和渲染 Pixar USD 中的场景。通过 Create，用户可以将自己的任意 assets 转译进 Create 中进行编辑，也可以直接从 Create's Libraries 中调用已有的 assets，这些 assets 包括：树木，家具，道路，人物等。Create 同时也配备了 NVIDIA RTX™ 渲染器，该渲染器支持多 GPU 的系统，可以高速可视化真实照片场景。再加上基于 MDL 的 NVIDIA vMaterials，使得场景更加真实且具备在视觉上的交互性。对于高速播放和创作，Create 还包括由 RTX 支持的实时光线追踪。通过使用 Omniverse Connector，设计师还可以从行业领先的渲染工具（如 Unreal Engine 或 Houdini）导入景观。

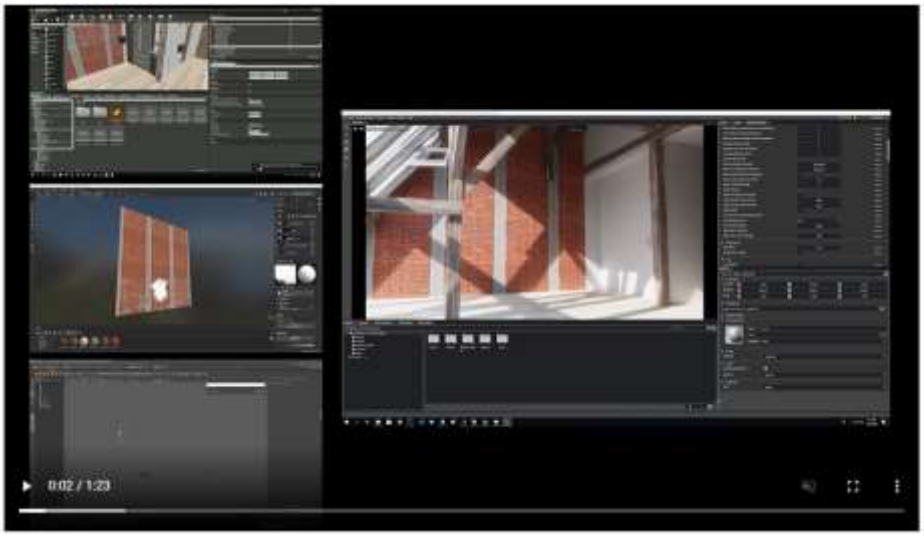
5.3. Omniverse 特点与行业应用场景

5.3.1. Omniverse 特点突出，优势定位明晰，与传统软件比更易上手

英伟达能够实现 Omniverse，与其软硬件生态系统布局息息相关。英伟达的构想是，流媒体数据通过云传递到以 GPU 为主的 RTX 处理单元，通过 Nucleus DB，一方面背靠存储和各个算法模块，另一方面连接所有的相关设计平台，来实现通用平台实时渲染能力。

Omniverse 的优点十分突出，在技术上提供了更加便捷的编辑方式，提供协同编辑的平台，也让上手难度降低。与传统三维设计制作流程不同，Omniverse 并不需要多个流程分开贴图、渲染最后出图，它提供实时协同工作的平台，可以让建模师、灯光师、特效师等每个部门专业人员无缝协同工作。Omniverse 可以让使用者一边查看渲染结果，一边进行编辑。现在效果图的渲染主要使用的是 3d max 和 maya 等软件，这类软件在每次编辑后需要生成渲染效果图，时间长度根据渲染的复杂程度不同而不同，这使得在开发者们在制作渲染时，很难直观的修改一些自己不满意的渲染效果，每次都需要退回原来的开发文件进行修改后再次生成渲染效果。Omniverse 则允许 3d max 和 maya 文件进行实时渲染，可以一边查看渲染结果一边进行编辑。由于算力限制，目前 Omniverse 还无法做到像 3d max 和 maya 那种细腻的渲染画面，但英伟达也在不断升级算法能力，以期带来更好的渲染效果。同时，Omniverse 内置了很多纹理和材质的预设值，可以更加便捷的对渲染效果图进行一键编辑。通过 connector 软件，使用者可以使得 3d max 或 maya 与 omniverse 同步运行，在一款软件上进行编辑时另一款软件也会同步反馈结果。这意味着开发者或者好设计师可以在 3d max 或者 maya 中修改纹理或者材质的同时在 Omniverse 中立即看到效果，极大提高多人协作的效率。

图40：Omniverse 实现多人协同设计及渲染



数据来源：公司官网，东吴证券研究所

5.3.2. 应用场景革新，改变行业流程

我们认为，NVIDIA Omniverse 的物理级模拟能力和强大的渲染能力，能够转变行业工作流程，通过物理级准确的模拟，不仅可以更好的监控工业流程，甚至可以进行流程的预测和效率改进。

1. 建筑、工程和施工：可以实现无缝协作，即使使用不同的软件应用程序，项目团队也能在单个交互式平台上进行整合；可以一键实现光线追踪设计；可以缩短审批时间，借助导出逼真模型的能力，团队、客户和承包商可以随时随地在不同设备上查看高保真模型。

2. 制造业：从草图和表面模型到物理属性准确的渲染，团队能通过无缝的交互开发产品概念和确定工程定义；真正连接起来的供应链组合各种应用程序的数据，以确保建立一致且最新的数字核心，从而连接整个产品设计和制造流程，帮助实时获取对业务至关重要的见解。

3. 媒体和娱乐业：NVIDIA Omniverse 让内容创作者能够利用各种创意应用程序制作、反复修改和协同创作内容，以提供实时结果。

4. 游戏开发：NVIDIA Omniverse 通过将美术师、他们的资产和软件工具统一在一个强大的平台上，帮助游戏开发者以创纪录的速度构建逼真的、物理上精确的游戏。

5. 科学可视化：用户可以快速构建受广泛支持的开源 USD 格式的文件。借助该解决方案，团队成员能够灵活地在其惯用应用中对数据进行可视化处理。

图41：革新建筑、工程和施工

数据来源：公司官网，东吴证券研究所

图42：革新制造业

数据来源：公司官网，东吴证券研究所

图43：革新媒体和娱乐业

数据来源：公司官网，东吴证券研究所

6. 盈利预测与估值

6.1. 盈利预测

我们对英伟达的各项业务收入做如下预测：

表6：英伟达各业务营收预测

营业收入（亿美元）	FY 2021A	FY 2022A	FY 2023E	FY 2024E	FY 2025E
汽车业务营收	5.36	5.66	12.46	22.37	36.20
yoy		6%	120%	80%	62%
消费者（游戏）业务	77.59	124.62	143.31	166.24	199.49
yoy		61%	15%	16%	20%
数据中心业务	66.96	106.13	165.03	251.67	376.23
yoy		58%	55%	52%	49%
专业解决方案	10.53	21.11	28.50	42.75	72.67
yoy		100%	35%	50%	70%
其他业务	6.31	11.62	12.78	15.34	19.94
yoy		84%	10%	20%	30%
营业总收入	166.75	269.14	362.08	498.37	704.54
yoy		61%	35%	38%	41%

数据来源：公司财报，东吴证券研究所

6.1.1. 消费级显卡业务

消费级（游戏）业务主要是个人 PC 上的以提升游戏和图像显示能力的独立显卡业务。由于新冠疫情影响，人们居家时间较多，游戏的需求明显增加，再加上许多 3A 游戏需要通过全新一代英伟达显卡才能达到体验效果，英伟达全新一代 RTX 系列显卡一直处于供不应求的状态；虽然 AMD 推出的显卡也获得了好评，且拥有一定的性价比优势，不过由于（1）英伟达在高端显卡优势明显，且针对英伟达显卡进行优化的游戏数量远超过 AMD，因此英伟达显卡仍是消费者首选；（2）整个市场仍有很大的空间，全球 80%以上的 PC 尚未升级新一代显卡，与此同时全球 PC 游戏玩家数量大幅增长，因此我们预计英伟达游戏业务的收入仍将保持强劲势头。FY2022 财报显示，游戏业务营收达到 124.62 亿美元，同比增长 61%。我们预计，随着 RTX 系列产能不断释放，叠加 RTX 带来均价的上涨，FY2025 游戏显卡业务（包括云游戏相关）营收可达 199.49 亿美元，FY2023~2025 的 CAGR 为 17%。

6.1.2. 数据中心业务

数据中心业务主要指向云服务商提供的云服务器硬件芯片及系统。鉴于 AI 以及云业务的快速兴起，GPU 因并行架构优势将获得更多的市场份额，英伟达在云业务的营收迎来了快速的增长，FY2022 的营收达 106.13 亿美元，同比增长 58%。我们预计，随着 AI 以及云计算的快速普及带来的硬件配置需求，英伟达数据中心业务将会延续快速增长的势头。Yole 预计全球云服务的市场规模，从 2019 年到 2025 年的 CAGR 为 13%，

其中 GPU 的需求远大于 CPU。我们认为英伟达作为行业龙头，FY2025 数据中心业务营收可达 376.23 亿美元，FY2023~2025 的 CAGR 为 52%。

6.1.3. 汽车业务

我们通过芯片单价乘以芯片出货量来估计汽车业务收入。通过对竞争对手 Mobileye2020 年的财报分析，2020 年 Mobileye 出货量为 1930 万片，营收为 9.67 亿美元，得到其平均售价为 50 美元；但考虑到英伟达芯片主要是面对 L2+级车辆，其单价水平较高，英伟达芯片定价能力很强，我们假定英伟达单芯片平均价格在 100 美元左右（价格包含相应的软件服务）。随着技术的进步，芯片更新换代能力变强的同时成本也会有所下降，这两个因素使得芯片价格会在 2021~2025 年内维持在 100 美元。FY2020 英伟达汽车业务收入为 7.0 亿美元，但由于疫情因素，以及新冠疫情对汽车销售的影响，FY2021 汽车业务收入下降到 5.36 亿美元。 FY2022 财报显示，汽车业务收入是 5.66 亿美元，同比仅增长 6%。但随着 Orin 芯片放量在即，疫情逐渐恢复后汽车整体销量恢复叠加智能化汽车开始放量等因素，英伟达汽车业务 FY2025 营收可达 36.20 亿美元，FY2023~2025 的 CAGR 为 86%。

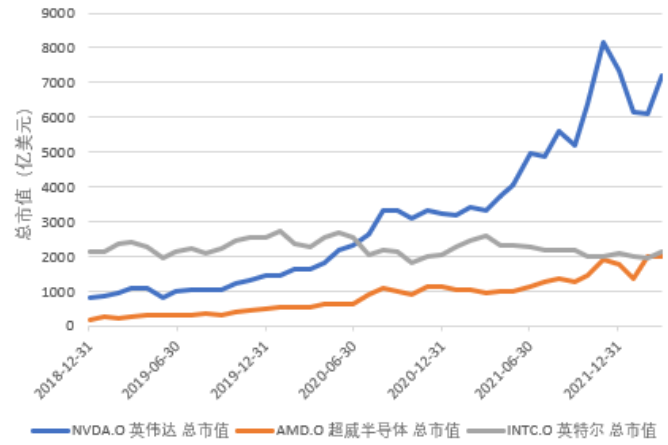
6.1.4. 专业解决方案业务

英伟达还拥有专业图像显示业务，专业图像显示主要是利用英伟达的 Quadro 技术，将高性能 GPU 嵌入戴尔、联想等 PC 中打造专业图形工作站的业务。此项业务客户主要为图形图像相关的建筑设计、医疗影像、影音等公司，FY2019~2021 营收为 11.30/12.12/10.53 亿美元，整体营收较为稳定。值得注意的是，英伟达目前将全新推出的 Omniverse 归入此项业务之中，而 Omniverse 的变现能力会随着社会对元宇宙认识程度及技术发展而快速变化，因此我们难以准确的预估数年后的收入情况。我们相信 Omniverse 在未来可能成为英伟达营收的第二增长曲线，但近三年可能还无法非常快速的成长，我们预计 FY2025 专业解决方案业务营收可达 72.67 亿美元，FY2023~2025 的 CAGR 为 51%。

6.2. 估值预测

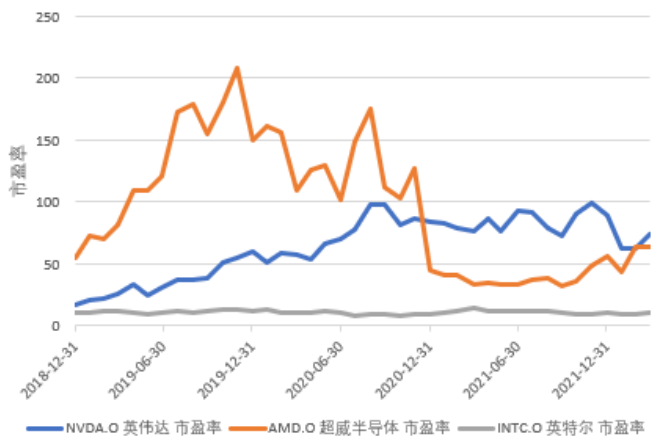
英伟达近些年来市值屡创新高，自 2020 年 7 月首次超过英特尔后，截止到 2022 年 3 月 29 日，其市值已经是英特尔的近 3.5 倍，成为全球半导体企业中市值最高的几家公司之一。英伟达 2022 年 3 月 29 日的 PE 为 72 倍，处于历史较高水平，与可比公司英特尔的 10 倍 PE 和 AMD 的 58 倍 PE 相比也较高，与费城半导体指数的平均 31 倍的 PE 也高出不少，这都可以看出资本市场对英伟达未来的预期很高。

图44：英伟达与英特尔、AMD 市值对比



数据来源：Wind，东吴证券研究所绘制

图45：英伟达与英特尔、AMD 市盈率对比



数据来源：Wind，东吴证券研究所绘制

我们认为，英伟达的强势表现主要源于两点，一是由于 GPU 的架构使用场景越来越丰富，使得英伟达的潜在市场空间得到了扩张；另一点是源自英伟达对于未来科技的高瞻远瞩，使其在人工智能以及元宇宙领域提前做了大量研究储备而获得了很强的先发优势。以人工智能行业为例，英伟达在 2006 年就开始投入巨资开发 CUDA 工具链，让人工智能行业的研究者方便调用，甚至成为某些领域的推动者（人工智能领域中深度学习的训练和推理）并逐渐形成生态，使这些领域很难有动力使用英伟达之外的产品。

我们预计，随着英伟达在数据中心（三年复合增速 52%）、汽车(三年复合增速 67%)以及专业显示领域（三年复合增速 51%）营收快速增长，且 To B 利润率会略高于 To C 业务，其高 PE 会得到快速消化。我们考虑到公司在新兴领域的龙头地位和稀缺性，给予公司 FY2023 年 90 倍 PE，估值为 9188 亿美元，对应当前目标价为 366 美元，首次覆盖，给予“买入”评级。

7. 风险提示

国家政策风险：政府政策导致游戏市场、云计算市场或自动驾驶市场增速放缓。政府的行动，包括美国和外国政府机构的贸易保护和国家安全政策，比如关税、进出口条例、贸易和经济制裁、其他贸易壁垒和限制可能会影响公司的运营。尤其在世界范围内地缘政治紧张局势和冲突的背景下，中国大陆、台湾、香港、以色列和韩国，这些地区集中了公司产品组件的制造和产品的最终组装，各国贸易政策和出口管制的变化都可能影响公司的运营策略、产品需求、进入全球市场的机会、招聘情况和盈利能力。

法律风险：自动驾驶相关领域、人工智能相关领域法律趋严，且涉及法律面较广，导致商业化项目迟迟无法落地。公司在国内和世界范围内都受到法律法规的约束，影响因素包括但不限于：知识产权所有权和侵权；税收；进出口要求和关税；反腐败；商业收购；外汇管制和现金返还限制；数据隐私要求；竞争和反托拉斯；广告；就业；产品法规；网络安全；环境，健康，安全要求；负责任地使用人工智能；气候变化；加密货币；消费者法律。遵守这些要求可能是繁重而昂贵的，进而影响竞争地位并对业务运营产生负面影响。比如人工智能带来了新出现的伦理问题，如果公司无法制定有效的内部政策和框架，向市场负责任地推出 AI 模型和系统，就可能会遭遇品牌或声誉损害，竞争损害或法律责任，削弱公众对人工智能的信心。

自身技术风险：GPU 升级路线、AI 芯片架构选择不符合市场需求。英伟达的 GPU 统治了 AI 芯片市场，但快速发展的人工智能领域需要新的架构。行业需要更高效的硬件来处理更多参数和更多数据以提高准确性，同时还要防止人工智能成为环境灾难，这就需要更多、更好的人工智能芯片。

竞争者风险：GPU 受到 AMD 和英特尔的持续挑战；AI 芯片竞争者众多。比如英特尔和 AMD 提供的高端分立 GPU，在价格和参数上存在公司不具备的优势，可能会导致公司产品的销售价格低于预期。另外，虽然英伟达的早期工作奠定了领先优势，但挑战者正在竞相迎头赶上。谷歌于 2015 年开始制造自己的芯片；亚马逊在 2016 年收购 Annapurna Labs 后，2020 年开始将 Alexa 的大脑转移到自己的 Inferentia 芯片上；百度拥有昆仑，估值约 20 亿美元；高通拥有 Cloud AI 100；IBM 正在致力研发一种节能设计。AMD 收购了 Xilinx 用于 AI 数据中心产品，英特尔在 2019 年为其 Xeon 数据中心 CPU 添加了 AI 加速模块；它还收购了两家初创公司：Nervana 和 Habana Labs。在过去几年中，Graphcore、SambaNova、Cerebras、Mythic AI、Blaize 和 TensTorrent 等公司也都发布或展示了自己的 AI 芯片。这些众多竞争者的入局，对公司的长期发展提出了挑战。

英伟达三大财务预测表

资产负债表（百万美元）	FY2022A	FY2023E	FY2024E	FY2025E
流动资产	28,829	41,123	62,361	91,613
现金及现金等价物	1,990	9,911	24,297	44,782
应收款项及票据	5,016	6,276	8,721	12,415
存货	2,605	3,441	4,358	5,776
其他流动资产	19,218	21,495	24,987	28,640
非流动资产	15,358	14,726	14,170	13,678
固定资产	2,778	2,681	2,617	2,578
商誉及无形资产	6,688	6,153	5,661	5,208
长期投资	0	0	0	0
其他长期投资	266	266	266	266
其他非流动资产	5,626	5,626	5,626	5,626
资产总计	44,187	55,848	76,531	105,291
流动负债	4,335	5,887	7,862	10,794
短期借款	258	1,311	2,626	4,609
应付款项	1,783	2,282	2,941	3,890
其他流动负债	2,294	2,294	2,294	2,294
非流动负债	17,575	12,240	13,240	12,040
长期借款	10,946	9,946	10,946	9,746
其他非流动负债	2,294	2,294	2,294	2,294
负债合计	17,575	18,127	21,102	22,834
少数股东权益	0	0	0	0
归属母公司股东权益	26,612	37,721	55,429	82,457
负债和股东权益	44,187	55,848	76,531	105,291

现金流量表（百万美元）	FY2022A	FY2023E	FY2024E	FY2025E
经营活动现金流	9,108	11,887	17,537	25,942
投资活动现金流	-9,830	-2,467	-3,707	-3,889
筹资活动现金流	1,865	-1,499	555	-1,568
现金净增加额	1,143	7,921	14,385	20,485
折旧和摊销	1,174	822	771	726
资本开支	-976	-190	-215	-235
营运资本变动	-3,363	-543	-1,387	-2,180

利润表（百万美元）	FY2022A	FY2023E	FY2024E	FY2025E
营业总收入	26,914	36,208	49,837	70,454
营业成本	9,439	12,546	15,699	20,911
销售费用	2,166	3,461	4,685	6,502
管理费用	0	0	0	0
研发费用	5,268	8,330	10,983	15,175
其他费用	0	0	0	0
经营利润	10,041	11,870	18,470	27,866
利息收入	29	23	78	150
利息支出	236	499	445	368
其他收益	107	0	0	0
利润总额	9,941	11,395	18,103	27,648
所得税	189	286	395	620
净利润	9,752	11,109	17,708	27,028
少数股东损益	0	0	0	0
归属母公司净利润	9,752	11,109	17,708	27,028
EPS	3.89	4.43	7.06	10.77
EBITDA	11,215	12,693	19,241	28,592

重要财务与估值指标	FY2022A	FY2023E	FY2024E	FY2025E
每股收益（美元）	3.89	4.43	7.06	10.77
每股净资产(美元)	10.60	15.03	22.08	32.85
发行在外股份（百万股）	2510	2510	2510	2510
ROIC(%)	23.23%	21.60%	24.55%	26.86%
ROE (%)	36.65%	29.45%	31.95%	32.78%
毛利率（%）	64.93%	65.35%	68.50%	70.32%
销售净利率(%)	36.23%	30.68%	35.53%	38.36%
资产负债率(%)	39.77%	32.46%	27.57%	21.69%
收入增长率(%)	61.40%	34.53%	37.64%	41.37%
净利润增长率(%)	125.12%	13.91%	59.41%	52.63%
P/E	71.27	62.57	39.25	25.72
P/S	25.83	19.20	13.95	9.87
EV/EBITDA	62.78	54.76	35.43	23.08

数据来源：Wind,东吴证券研究所

注：若无特别说明，表中货币单位均为美元；英伟达财年从 1 月 31 日开始计算，FY2022 对应公历年份 2021.1.31~2022.1.30。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（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客户。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，本公司不对任何人因使用本报告中的内容所导致的损失负任何责任。在法律许可的情况下，东吴证券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并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，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。

东吴证券投资评级标准:

公司投资评级:

买入：预期未来 6 个月个股涨跌幅相对大盘在 15%以上;

增持：预期未来 6 个月个股涨跌幅相对大盘介于 5%与 15%之间;

中性：预期未来 6 个月个股涨跌幅相对大盘介于-5%与 5%之间;

减持：预期未来 6 个月个股涨跌幅相对大盘介于-15%与-5%之间;

卖出：预期未来 6 个月个股涨跌幅相对大盘在-15%以下。

行业投资评级:

增持： 预期未来 6 个月内，行业指数相对强于大盘 5%以上;

中性： 预期未来 6 个月内，行业指数相对大盘-5%与 5%;

减持： 预期未来 6 个月内，行业指数相对弱于大盘 5%以上。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn