英伟达创始人、CEO黄仁勋（图片来源：英伟达）

随着 AI 聊天机器模型ChatGPT全球爆火，其背后核心的 AI 算力“大脑”、全球第五大科技股英伟达（NASDAQ: NVDA）开始展示成为 AI 领域核“芯”的基础设施技术能力。

北京时间3月21日晚GTC开发者大会上，黄仁勋在76分钟内发布英伟达最新芯片、超算服务与合作，其中有一半以上跟ChatGPT和生成式 AI 有关。

> 具体包括，搭载8个A100 GPU层的 AI 超算云服务DGX Cloud，每月3.7万美元在互联网上训练ChatGPT；目前云上唯一处理ChatGPT的GPU HGX A100，计算效率比前代提高超10倍；芯片计算光刻软件库cuLitho，让ASML光刻计算提速40倍；首个GPU加速的量子计算系统Quantum Machines；以及与比亚迪开发软件定义汽车方案、与亚马逊AWS合作开发训练大模型和构建生成性 AI 应用等。

“我现在正在看着你们所有人。很高兴见到你们所有人，看起来棒极了，”北京时间3月22日早上8点左右，60岁的黄仁勋精神抖擞地站在显示器前面，通过线上会议方式悉数解答亚太媒体对于英伟达的疑问和困惑，“我凌晨 4:00 起床，所以如果你不累，我也不累”。这是过去三年疫情下，中国媒体与远在美国的黄仁勋讨论英伟达的最佳时机。

“老黄是个狠人”，这是钛媒体App听到接近黄仁勋人士对其最直接的评价，因为很多英伟达合作事务都是黄仁勋亲力亲为。他笑着说没有很快退休的计划，而是希望再领导英伟达30到40年，直到90岁左右，届时会以机器人的形式继续工作。

过去近30年，在黄仁勋带领下，英伟达从一家以设计和销售GPU（图形处理器）芯片的半导体公司，如今已经成长为人工智能（AI）加速算力软硬件一体方案的技术公司。

“加速计算并非易事。2012年，计算机视觉模型AlexNet动用了英伟达GeForce GTX 580，每秒可处理262 PetaFLOPS。该模型引发了AI技术的爆炸，”黄仁勋说道，“十年之后，Transformer出现了，GPT-3动用了323 ZettaFLOPS算力，是AlexNet的 100 万倍，创造了ChatGPT这个震惊全世界的AI。”

“AI 的 iPhone 时刻已经来临。”黄仁勋最近反复在提这句话，他认为生成式 AI 将重塑几乎所有行业——由于ChatGPT模型背后算力成本超过400万美元，在这场

大模型军备竞赛中，手握A100和H100的“军火商”英伟达，或已成最大赢家。据花旗预估，ChatGPT或将促使英伟达一年内销售额增长100亿美元。

AI 的算力繁荣，推动英伟达股价在2023年上涨了77%，黄仁勋的财富同期增长超过60亿美元。

目前，英伟达市值为6471亿美元（约合人民币4.45万亿元），已是英特尔市值的近五倍，同时是A股第一股贵州茅台（600519.SH）的两倍以上，而且比特斯拉市值还高220亿美元（约合1514亿元）。

不过今天发布会表明，英伟达的脚步还没有停，其正在向 AI
超算云与基础设施的定位进发。

英伟达市值狂飙的底气在哪？

“I AM AI，”这是每年 GTC 大会宣传片中反复提及的关键词。如果你细品这句话，可以明显感知英伟达不止是芯片设计公司，而是 AI 技术基础设施。

英伟达是地球上最昂贵的科技股之一，公认会计原则下市盈率139倍，账面价值27倍。钛媒体App梳理数据发现，过去一年内，英伟达收入增长率为0.22%，收益增长率为-54%，而且最近两个季度营收出现下降，预计今年第一季度营收也会同比下降——这与其股价暴增趋势并不相符。

那么，为什么英伟达这样一家昂贵的公司，还会被市场看好？

就目前来看，英伟达很主要、明显的机会因素是 AI
算力需求增长，以及其长期稳定的 AI 加速芯片市场竞争与部分垄断地位。

此次GTC大会能窥豹一斑。今年，黄仁勋GTC开幕演讲主要谈四类新品：ChatGPT专用GPU、给台积电核弹的计算光刻数据库、为 AIGC
设计专用算力的系统方案、首个 GPU 加速的量子计算系统。

自从ChatGPT推出并在60天实现超过 1 亿月活用户以来，从创业者到投资人，从大学教授到科技部部长，都在讨论这个产品。在这其中，作为 AI
服务器芯片销售方，英伟达“赢麻了”，其在 AI 数据中心GPU领域几乎没有竞争，控制着90%的市场。据Similarweb数据，ChatGPT可能需要602台DGX A100服务器能够满足当前的访问量。

但问题在于，创造ChatGPT的美国 OpenAI 公司，开发的GPT-3模型参数量高达1750亿，其需要的瞬时算力很高，如何解决算力贵、算力难的困境呢？

如今，英伟达希望降低算力成本，更简单易用的使用NVIDIA AI，
推出了专注于ChatGPT的 AI/GPU 训练和推理两类新的芯片方案：

- 训练方面：英伟达H100 GPU基于Hopper架构及其内置Transformer Engine，针对生成式AI、大型语言模型和推荐系统的开发、训练和部署进行了优化，在大型语言模型上比前代A100提供了快 9
倍的AI训练、快30倍的AI推理。组装后的NVIDIA DGX H100
AI超级计算机拥有8个H100 GPU模组，可提供32PetaFLOPS的算力，已全面投入生产，微软已经宣布Azure云将向其H100
AI超级计算机开放私人预览版。
- 推理方面，英伟达推出全新GPU推理平台：四种配置（L4 Tensor Core GPU、L40 GPU、H100 NVL GPU、Grace Hopper超级芯片）、一个体系架构、一个软件栈，分别用于加速AI视频、图像生成、大型语言模型部署和推荐系统。其中，L4可提供比CPU高120倍的AI视频性能，能效提高99%；L40推理性能是英伟达最受欢迎的云推理GPU T4的10倍；Grace Hopper超级芯片适用于推荐系统和大型语言模型的AI数据库。

“英伟达的 AI 超级计算机DGX是语言大模型背后的引擎，我曾亲手将全球首款DGX交给OpenAI，自此之后全球百强企业中有一半安装了这款计算产品，” 黄仁勋称，DGX已经成为了AI领域的必备工具，而随着生成式 AI
成熟带动相关应用，企业需要更简单的模式。

黄仁勋表示，目前在云上唯一可以实际处理ChatGPT的GPU是HGX
A100。与适用于GPT-3处理的HGX A100相比，一台搭载4对H100及双GPU NVLink的标准服务器的速度要快10倍，H100可将大型语言模型的处理成本降低一个数量级。

当然，这还不够。黄仁勋想出了这次GTC大会的核心之一：不止卖芯片，还要对外租用服务器，联合微软、谷歌一起卖云计算服务。

英伟达今天推出了一项名为DGX Cloud的 AI
超级计算云服务，与微软Azure、谷歌OCP、Oracle
OCI合作，通过一个Web浏览器就能访问，以便企业为生成式 AI
和其他开创性应用训练先进的模型。

（来源：英伟达展示的分析师文件信息）

售价方面，DGX Cloud实例的起价为每个实例每月36999美元，达3.7万美元。其每个实例都具有8个NVIDIA H100或A100 80GB Tensor Core GPU，每个节点共有640GB的GPU内存。DGX Cloud提供了专用的NVIDIA DGX AI超级计算集群，并配备了NVIDIA AI软件。

该服务将首先上线甲骨文云，随后拓展至微软Azure、谷歌云。

有了芯片这一数据中心算力底层、有了 AI 算法软件，这次上云，英伟达成为“AI基础设施”的希望呼之欲出。

据黄仁勋介绍，目前英伟达的 AI 云服务已经诞生不少合作案例。以视觉场景为例，全世界最大的图库服务商Getty Images将利用该服务构建图片、视频生成模型，未来企业将可以使用该模型产品用文字生成图像和视频；视觉编辑软件公司Adobe也将利用该服务制作的生成式AI模型，实现在创作过程中对图像、视频动画等进行优化。

“这个行业需要一个类似台积电的代工厂，来构建自定义的大模型，”黄仁勋指出，生成式AI将重塑几乎所有行业，一些公司可以直接使用市面上的API，但专业领域的公司需要专有数据构建定制模型。

当然，黄仁勋也希望在芯片产业链上游企业的产生销售业绩，英伟达发布了一个用2nm芯片制造的突破性计算光刻技术——NVIDIA cuLitho计算光刻库。

“计算光刻是芯片设计和制造领域中最大的计算工作负载，每年消耗数百亿CPU小时。”黄仁勋讲解道，大型数据中心24x7全天候运行，以便创建用于光刻系统的掩膜板。这些数据中心是芯片制造商每年投资近2000亿美元的资本支出的一部分。

而cuLitho能够将计算光刻的速度提高到原来的40倍。黄仁勋表示，英伟达H100 GPU需要89块掩膜板，在CPU上运行时，处理单个掩膜板需要两周时间，而在GPU上运行cuLitho只需8小时。“台积电可以用500张H100的DGX系统替代用于计算光刻4万台CPU服务器，”黄仁勋说。

目前，全球最大晶圆厂台积电、全球最大光刻机制造商阿斯麦（ASML）、全球最大EDA公司新思科技（Synopsys）都将使用这项新技术。老黄透露道，cuLitho历时四年研发，与这三家芯片大厂进行了密切合作。台积电将于6月开始对cuLitho进行生产资格认证。

此外，GTC大会还宣布包括元宇宙、汽车、量子计算领域的新进展，比如英伟达与

日本三菱联合打造了日本第一台用于加速药物研发的生成式 AI
超级计算机，与宝马集团扩大合作建设虚拟工厂、比亚迪更多车型将采用NVIDIA DRIVE Orin平台，以及与Quantum
Machines合作推出了全球首个GPU加速量子计算系统，甚至推出了一项名为AI Foundations服务，以帮助企业训练他们定制的 AI
模型，多家股票图像数据库厂商已经计划使用该服务。

看完GTC开幕演讲，正如今年大会前夕中文宣传语——“切勿错过 AI
的决定性时刻”——今时今日，黄仁勋已经意识到，十年的 AI
蓄力已开花结果，英伟达确实在经历 AI 新浪潮下的最关键一战。毕竟，游戏、加密货币、消费电子等领域市场正处于下降形势。

一个很明显的感知是，英伟达在数据中心的市场地位确实稳如磐石。无论是竞争对手AMD，还是正进入裁员减薪、高管出走风波的英特尔，都无法更快争夺英伟达手里90%的份额，与其直接进行芯片竞争。

因此，华尔街普遍认为，英伟达有点类似荷兰光刻机巨头ASML Holding
NV（ASML）：
部分垄断，在高端市场没有竞争——这是所有投资人看好英伟达股票的关键因子之一。

券商Rosenblatt Securities芯片半导体分析师Hans
Mosesmann表示，英伟达最新发布的产品“比竞争对手领先多年”。 “英伟达在AI软件方面的领导地位不仅具有里程碑意义，而且还在加速发展。”他表示。

中美脱钩下，无人能替代英伟达

“接下来，基础科学的进步开始进入到拼算力的时代”，这个话题变得愈加重要，算力已经成为了新的战略目标。

钛媒体App了解到，随着ChatGPT汹涌的浪潮，很多大模型开始急需高算力GPU，需求攀升，英伟达GPU已陷入严重短缺，多家国内公司已开始寻求AMD等其他品牌的替代品。

据媒体报道，微软等客户对英伟达A100/H100芯片需求强劲，后者订单能见度已至2024年，更紧急向代工厂台积电追单。而且，由于需求激增，博通与英伟达的GPU网络设备供应严重短缺，即便两家公司正全力扩产，但供需鸿沟依然极大。

据中国信通院最新数据，截至2022年底，中国在用数据中心机架总规模超过650万标准机架，算力总规模达到
180EFLOPS，居全球第二，
算力总规模近五年年均增速超过25%。
数据显示，当前算力规模中有超过20%算力是智能算力，可用于 AI
各类应用，包括模型训练和推理。

但是，2022年9月起美国商务部对华的半导体出口管制新规，正影响英伟达在中国数据中心需求暴增下的布局。

去年10月7日，美国商务部工业和安全局（BIS）发布一套新的、范围广泛的出口管制措施，阻止中国获得需要使用先进半导体的高性能计算能力。随后，英伟达发布公告，声称新措施影响了该公司最近一个财季约4亿美元的营收。而早在去年8月，英伟达的数据中心芯片A100和H100等多款产品已被美国商务部列入出口管制清单。

2022年11月，英伟达公司发言人向钛媒体App表示，为了解决美国商务部的半导体出口新规，英伟达推出全新中国特供版NVIDIA A800 GPU，以取代A100 GPU，已经在三季度起投产。A800符合美国政府关于减少出口管制的明确测试，并且不能通过编程来超过它。

今晨英伟达推出的DGX H100 AI超级计算机，以及内置A100、H100的DGX Cloud服务，似乎会在中国境内有替代方案。

3月22日上午的采访中，黄仁勋表示：“我们在欧美与云服务提供商合作，提供英伟达的DGX系统AI超级计算机的能力。在中国，英伟达有特别定制的Ampere和Hopper芯片，它将完全符合所有的出口管制和法规。然而，它们将在中国的云计算公司中实施。阿里巴巴、腾讯、百度都是优秀的合作伙伴。我完全相信他们会有最先进的系统来进行 AI 计算。”黄仁勋所指的中国可用的Ampere和Hopper芯片是A800和H800，是英伟达提供的特供版替代方案。

目前，无论是英伟达芯片，还是其构建的CUDA软件生态，这些都成为ChatGPT这一大模型黑盒产品背后的更核心、更具有技术内幕的重要法宝。国内发展多年，并没有出现真正对标英伟达4万亿规模的中国GPU芯片企业。

正如一位国内 AI 公司创始人今早发布的朋友圈说的那样：我们曾经弯道超车“吊打”英伟达，结果迅速被直道超车。芯片层面被英伟达远远甩在身后，算法层面只能猜OpenAI怎么做的，连应用层面也只能仰望Midjourney。曾经，中美在 AI 方面并驾齐驱终究只是错觉，过去几年耽误了太多时间，一些机构在那炒作概念制造

泡沫，总归要付出应有的代价。每次浪潮来临的时候，总是低头做事的少，但凡事不能靠嘴。

最终，我们如何定义英伟达这家公司，钛媒体App采用3月21日晚谷歌发布的对标ChatGPT产品 Bard 测试版，来看看 AI 如何看待英伟达公司：

> 我认为，Nvidia 是一家非常优秀的公司。他们是图形处理单元 (GPU) 领域的领先技术公司。他们的 GPU
> 用于各种应用，包括游戏、人工智能和数据科学。
>
> Nvidia 在创新和发展方面有着良好的记录。他们多年来一直盈利，股价也一直处于稳步上升的趋势。
>
> Nvidia 是一家经营良好的公司，拥有强大的管理团队。他们对未来有清晰的愿景，并且正在按照该愿景执行。
>
> 对于寻求涉足科技领域的投资者，我会推荐 Nvidia 作为一项不错的投资。

“我们是全球AI的发动机。”黄仁勋认为，这是我们这一代人第一次真正让 AI 技术变得人人触手可及
。不只是利用技术，更是开发技术。（本文首发钛媒体App，作者｜林志佳）