

人工智能行业点评

英伟达 GPU、DGX 云、AI 工厂三驾马车发布，AI 算力和应用再迎跃迁

◆ 行业研究 · 行业快评

◆ 计算机

◆ 投资评级：超配（维持评级）

证券分析师： 熊莉 021-61761067
证券分析师： 库宏尧 021-60875168

xiongli1@guosen.com.cn
kuhongyao@guosen.com.cn

执证编码：S0980519030002
执证编码：S0980520010001

事项：

英伟达本周举办 GTC 大会，围绕 AI、量子计算、芯片等前沿科技发布了一系列产品：如用于量子计算、光刻等领域的加速库；面向 AI、数据中心等领域的 CPU、DPU 加速芯片；面向 AI 超级计算集群出租的 DGX 云；用于生成式 AI 的大模型“代工厂”；面向 AI 视频、ChatGPT 等应用的推理芯片平台；面向智能制造的 Omniverse 虚拟工厂。对于当前发展火热的 AI 产业，英伟达发布会成为市场关注焦点。

国信计算机观点：1) ChatGPT 在 AI 问答上的优秀表现引爆网络，GPT-4 能力更加卓越，OpenAI 面临的访问流量呈现快速增长态势。最新的 GPT-4 访问价格已上涨数倍，其算力消耗和成本必然也有较大提升。AI 大规模的应用和普及，算力是瓶颈之一，英伟达发布产品有望助力 AI 产业跃迁；2) 芯片端，英伟达 H100 NVL 专门针对生成式 AI 语言模型设计，可以将大型语言模型的处理成本降低一个数量级；3) 云端，公司 DGX 云服务更方便快捷的出租 AI 超级计算集群算力，对于 AI 创业公司将大幅降低投入成本，有望繁荣 AI 生态；4) 应用端，公司已推出语言、视觉、和生物学三大模型，AI 模型“代工厂”的角色，让各类行业客户更加容易训练并使用针对自身的 AI 应用模型，将进一步普及 AI 应用；5) 英伟达为 AI 产业，从算力到应用均带来跃迁，将加速 AI 的应用和落地。参考国内产业对比，以百度文心大模型发布来看，国内也会逐步复制国外 AI 产业的发展路径，且当前已具备一定的应用案例，我们持续看好国内 AI 产业的发展，重点关注 AI 算力、算法、应用相关产业和标的；5) 风险提示：AI 等应用发展和落地情况不及预期；国内 AI 产业发展不及预期；国内算力、芯片等基础环节进展缓慢等。

评论：

◆ 英伟达发布 ChatGPT 专用 GPU，算力基础大幅提升

英伟达再发重磅 GPU 产品，AI 的“iPhone 时刻”来临。英伟达本次发布会展示了多项产品和技术，其中 GPU 性能的持续提升一直是大会焦点。英伟达发布了全新的 GPU 推理平台，该平台包括 4 种不同配置，针对不同工作负载进行优化：

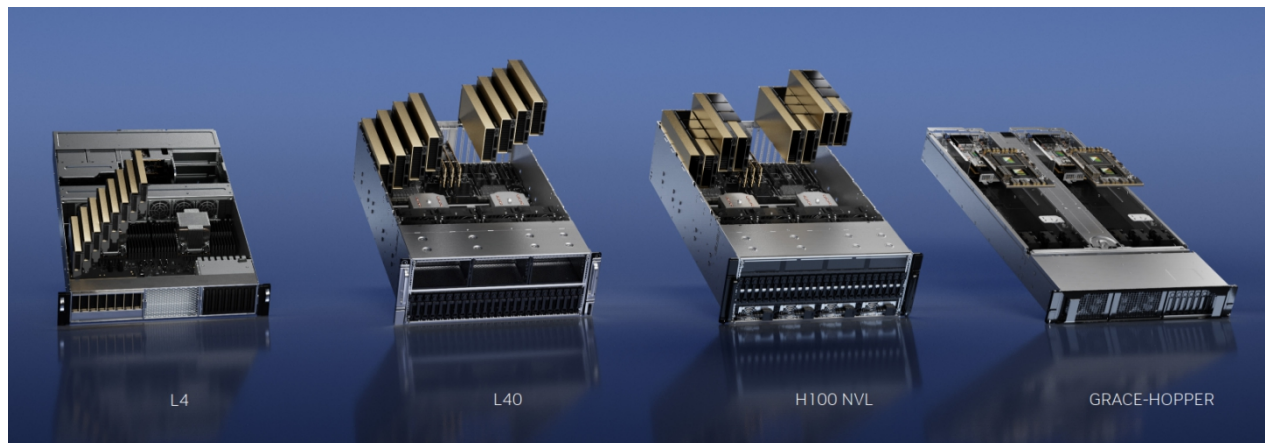
L4：针对 AI 视频工作负载。L4 在视频解码和编码、视频内容审核、视频通话功能等领域做了较大优化，其在 AI 视频加速方面比 CPU 高 120 倍的性能。一台 8-GPU L4 服务器将取代一百多台用于处理 AI 视频的双插槽 CPU 服务器。Google 云已与其合作推出了新的云硬件产品 L4 平台。

L40：针对 Omniverse、图形渲染等生成式 AI。L40 推理性能是英伟达的云推理 GPU T4 的 10 倍；L40 是英伟达元宇宙 Omniverse 的引擎。当前火爆的 AI 图片生成平台 Midjourney、Stable Diffusion 等未来均是 L40 的潜在的可能应用场景。

H100 NVL：针对 ChatGPT 等 LLM 大语言模型的 Hopper GPU。该产品是基于去年已经发布的 H100 的改进版本：配备双 GPU NVLink，将两张拥有 94GB HBM3 显存的 PCIe H100 GPU 拼接在一起。H100 可处理拥有 1750 亿参数的 GPT-3，还可支持商业 PCIe 服务器扩展。英伟达宣称 H100 NVL 提供的 GPT3-175B 推理吞吐量是上一代 HGX A100 的 12 倍（8 个 H100 NVL 对比 8 个 A100）。H100 NVL 卡将于今年下半年开始发货，以满足当前对大语言模型训练的强烈需求。

Grace Hopper：针对推荐系统和大型语言模型 AI 数据库设计的超级芯片。该芯片可以通过 900GB/s 的高速一致性芯片到芯片接口连接英伟达 Grace CPU 和 Hopper GPU，非常适合处理大型数据集。

图1: NVIDIA 推理平台



资料来源：英伟达官网，国信证券经济研究所整理

算力瓶颈制约 AI 大规模应用和普及，英伟达助力算力跃迁。近期 OpenAI 对于 Plus 付费用户的 GPT-4 访问连续下降阈值，已由最初的 150 msg/ 4 hr 下降至 25msg/3hr；其理由主要是为了控制成本。由于 ChatGPT 在 AI 问答上的优秀表现引爆网络，GPT-4 能力更加卓越，OpenAI 面临的访问流量呈现快速增长。根据最新的 GPT-4 访问价格，其每 1000 个 token 的价格数倍于原 GPT-3 和 GPT-3.5-turbo，因此其算力消耗和成本必然也有较大提升。英伟达最新产品即是解决算力瓶颈的关键，H100 NVL 专门针对生成式 AI 语言模型设计，可以将大型语言模型的处理成本降低一个数量级。根据公司演示，训练 Transformer 模型从 5 天缩短到了 19 小时，而在结合了英伟达多项尖端软硬件条件下，整个模型训练从 7 天缩短到了 20 小时，效率足足提升 9 倍。算力性价比的提升，将帮助 AI 在各种应用下迅速普及。

◆ 英伟达推出 DGX 云，AI 即服务再次降低行业门槛

DGX Cloud 云服务问世，让 AI 使用更简单。英伟达 DGX Cloud 是未来公司重点计划，公司提供了专用的 NVIDIA DGX AI 超级计算集群，搭配 NVIDIA AI 软件，向每家企业出租 AI 算力。英伟达的 AI 计算集群强大而昂贵，企业以云租赁的方式采购将更快、更方便的实现自身的 AI 训练和应用。目前英伟达已经和微软云、谷歌云、Oracle OCI 开展合作，其首个 NVIDIA DGX Cloud 是与 Oracle 合作的云服务，每个实例每月 36,999 美元起。每个实例包括八个 Nvidia H100 或 A100 80 GB GPU，每个 GPU 节点内存高达 640 GB。计算资源是专用的，不与云中的其他租户共享，网络也在租户之间隔离。

参考 IaaS 到 SaaS 的发展逻辑，DGX 云有望推动 AI 生态更加繁荣。以 OpenAI 大语言模型 GPT3 为例，其 1750 亿个参数，需要庞大成本投入和算力支持，一般企业难以涉足。目前在 AI 大模型领域持续投入的主要是互联网大厂，如谷歌、百度、阿里等；OpenAI 的迅速发展也离不开微软等巨头的持续投资。因此，类似于 2000 年后云计算的兴起，IaaS 让各类企业创业和发展更加容易，催生 B 端 SaaS 和 C 端互联网生态。可以预期，英伟达 DGX 云的推出，让 AI 创业更为方便，不再只是巨头游戏，未来 AI 应用和生态有望快速发展。

图2: 英伟达推出 DGX Cloud 云服务



资料来源: 英伟达官网, 国信证券经济研究所整理

◆ 英伟达 AI 代工厂再向前端迈进一步, 模型应用值得期待

生成式 AI 不断取得突破, 英伟达进一步推出模型制作服务。 ChatGPT、Midjourney 等生成式 AI 应用让整个世界为之振奋, 也彰显了大模型极强的应用能力和价值。英伟达发布 AI Foundations, 旨在为用户提供定制化的 LLM 和生成式 AI 解决方案。英伟达希望扮演 AI 模型“代工厂”的角色, 帮助客户构建、改进和操作自定义大型语言模型, 这些模型和生成 AI 使用其专有数据和特定领域的任务进行训练。英伟达 AI 大模型包含语言、视觉和生物学三大模型:

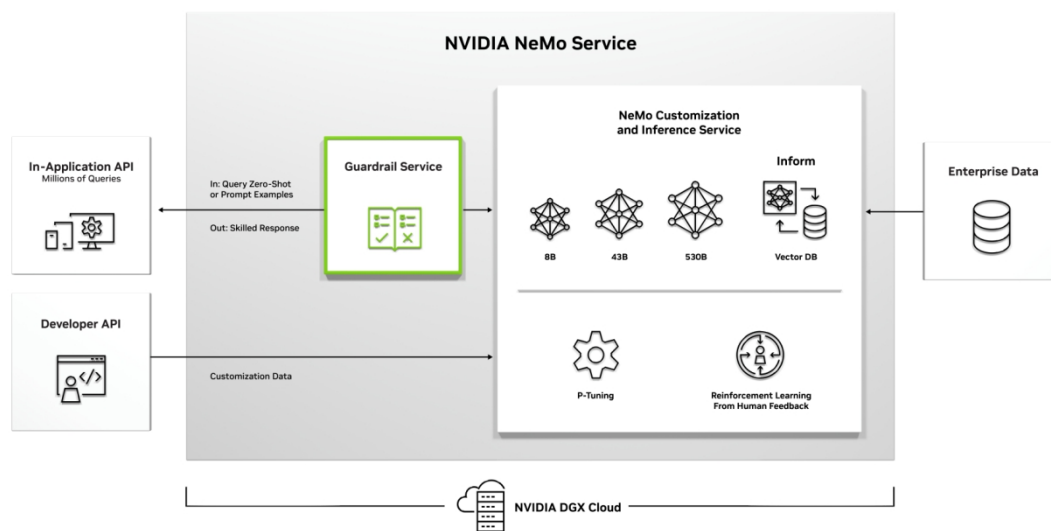
NEMO: 语言模型。 NEMO 是用于自然语言文本的生成式模型, 可以提供 80 亿、430 亿、5300 亿参数的模型, 且会定期更新额外的训练数据, 英伟达 AI 专家也可以提供全程指导合作。根据 NEMO 模型, 企业可以生成自己的大模型, 实现客服、聊天机器人、撰写电子邮件等功能。NEMO 模型是一个基于 PyTorch 的开源工具包, 允许开发者快速构建、训练和微调会话式人工智能模型。NeMo 由 NeMo Core 和 NeMo Collection 组成, 在 NeMo 的 Speech Collection 中, 可以找到用于语音识别、命令识别、说话人识别、说话人验证和语音活动检测的模型和各种构建模块。NeMo 的 NLP Collection (nemo_nlp) 包含了诸如问题回答、标点符号、命名实体识别等任务的模型。本次发布会, 公司还发布了 NVIDIA TAO Toolkit 5.0, 借助 NVIDIA TAO (训练、适应、优化) 工具包, 开发人员可以利用迁移学习的强大功能和效率, 通过适应和优化在创纪录的时间内实现最先进的准确性和生产级吞吐量, 可以促进对话 AI 模型的培训和微调。

Picasso: 视觉模型。 Picasso 用于训练包括图像、视频和 3D 应用的生成式 AI 模型。公司宣布 gettyimages、Shutterstock 将采用英伟达 Picasso 服务, Adobe 将与英伟达共同打造新一代的生成式 AI 模型。以 Shutterstock 为例, 其是一家面向创新品牌和媒体公司的领先创作平台。Shutterstock 使用 NVIDIA Picasso 生成式 AI 云服务, 使这些模型能够根据简单的文字提示创建生成式 3D 资产, 同时将训练的 3D 模型转换为高保真的 3D 内容, 创作时间可以从数小时缩短至数分钟。

Bionemo: 生物学模型。 Bionemo 专门为用户提供创建、微调、提供自定义模型的平台, 可用于蛋白质结构预测、蛋白质生成、蛋白质性质预测、分子生成、分子对接等各个流程。在数百家新的 AI 制药初创公司之中, 英伟达已经积累了至少 7 家公司, 作为 Bionemo 大模型的早期用户。

英伟达 AI 模型向产业前端再迈进一步，形成全产业链 AI 能力输出。英伟达再提供芯片、AI 算力云服务后，也向产业提供自己的大模型能力，让客户能最简化定制自身需要的 AI 能力。AI “代工厂”可以类比为云计算中 PaaS 及 SaaS 层服务，实现 AI 全生命周期能力输出，更进一步推动 AI 的应用普及。目前英伟达推出了三大模型，且已在业内有所应用，未来有望进一步推出更多模型。除此之外，英伟达“工业元宇宙” Omniverse Cloud 也是 PaaS 平台，旨在建设虚拟工厂，让各个行业在投入实际生产之前先以数字化方式设计、建造、运营和优化物理产品和工厂。

图3：英伟达语言模型 NEMO



资料来源：英伟达官网，国信证券经济研究所整理

◆ 投资建议：

以 ChatGPT、Midjourney 为代表的生成式 AI 应用席卷全球，对 AI 算力、算法、应用的需求已形成快速增长趋势。即使单个 GPT-4 的应用，算力和成本已呈现暴涨之势。英伟达近期发布的一系列产品对于 AI 产业形成较大的利好：芯片端，H100 NVL 专门针对生成式 AI 语言模型设计，可以将大型语言模型的处理成本降低一个数量级。云端，公司 DGX 云服务更方便快捷的出租 AI 超级计算集群算力，对于 AI 创业公司将大幅降低投入成本，有望复制云计算催生 B 端 SaaS 和 C 端互联网生态，同样繁荣 AI 生态。应用端，公司已推出语言、视觉、和生物学三大模型，AI 模型“代工厂”的角色，让各类行业客户更加容易训练并使用针对自身的 AI 应用模型，将进一步普及 AI 应用。因此我们认为英伟达为 AI 产业，从算力到应用均带来跃迁，将加速 AI 的应用和落地。参考国内产业对比，以百度文心大模型发布来看，国内也会逐步复制国外 AI 产业的发展路径，且当前已具备一定的应用案例，我们持续看好国内 AI 产业的发展：

算力端：重点关注中科曙光、海光信息等。

算法端：重点关注科大讯飞、三六零等。

应用端：重点关注金山办公、同花顺、凌志软件、彩讯股份等。

◆ 风险提示：

AI 等应用发展和落地情况不及预期；国内 AI 产业发展不及预期；国内算力、芯片等基础环节进展缓慢等。

相关研究报告：

《人工智能行业点评-OpenAI 访问限流，GPT-4 算力测算》——2023-03-20

《人工智能行业点评-Microsoft 365 Copilot 发布，国内外 AI 应用有望加速落地》——2023-03-19

《国央企 ERP 专题报告：数字化转型下的新机遇》——2023-03-10

《计算机行业 2023 年 3 月投资策略暨年报前瞻-2022 行业业绩承压，关注 ChatGPT 引发的 AI+应用表现》
——2023-03-06

《人工智能行业点评-ChatGPT 开放 API 接口，应用侧有望迎来全面爆发》——2023-03-05

免责声明

分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

国信证券投资评级

类别	级别	说明
股票 投资评级	买入	股价表现优于市场指数 20%以上
	增持	股价表现优于市场指数 10%-20%之间
	中性	股价表现介于市场指数 $\pm 10\%$ 之间
	卖出	股价表现弱于市场指数 10%以上
行业 投资评级	超配	行业指数表现优于市场指数 10%以上
	中性	行业指数表现介于市场指数 $\pm 10\%$ 之间
	低配	行业指数表现弱于市场指数 10%以上

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所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或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所

深圳

深圳市福田区福华一路 125 号国信金融大厦 36 层

邮编：518046 总机：0755-82130833

上海

上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层

邮编：200135

北京

北京西城区金融大街兴盛街 6 号国信证券 9 层

邮编：100032