# 八巨头抱团挑战英伟达 AI芯片市场风云再起

**本报记者 李玉洋 上海报道**

若问最好的生成式AI算力供应商是谁，英伟达会是一个版本答案。它用以拳头产品GPU为代表的硬件产品和以CUDA为基础的软件解决方案，筑起了牢固的AI系统生态。

于是，“卖铲人”英伟达赚得盆满钵满。当地时间6月5日，英伟达（Nasdaq:NVDA）收涨5.16%，连续三个交易日刷新股价新高，总市值已达3.01万亿美元，在苹果、微软之后，成为全球第三个市值超过3万亿美元的公司，当前则是市值仅次于微软的全球第二大上市公司。

在此背景之下，友商英特尔、AMD、微软、Meta、OpenAI等公司在过去两年争相开始自研AI芯片也就可以理解了。如今，在对抗英伟达AI芯片“霸权”的道路上，巨头们又有了新动作。

近日，谷歌、Meta、微软、AMD、英特尔、博通、思科、惠普八大科技巨头宣布成立新联盟，旨在推动一个叫作Ultra Accelerator Link（UALink，超级加速链路）的行业标准。据悉，该标准支持多个AI加速器互联、内存结构新标准以及超以太网覆盖，大有合谋对抗英伟达之意。

《中国经营报》注意到，UALink不只是在名字上对标英伟达的NVLink，在技术路线上同样也是。能够实现GPU到GPU高速通信的NVLink、用于扩展pod之外的Infiniband以及用于连接到更广泛基础设施的以太网，这些技术让众多客户在过去数年间越来越倾向于购买英伟达的GPU。

最新消息显示，英伟达创始人兼CEO黄仁勋在2024年中国台北国际电脑展上公开了未来几年的GPU发展路线图，包括2025年的Blackwell Ultra、2026年的新架构Rubin以及2027年的Rubin Ultra。这意味着英伟达打破了以往两年更新一代的节奏，驶入一年一更新的快车道。

需要指出的是，英伟达的NVLink并不向行业开放，且NVLink已成为英伟达人工智能数据中心系统的标配。“UALink联盟企业正在努力创建一个开放、高性能和可扩展的加速器结构，这对于AI的未来至关重要。”AMD数据中心解决方案事业部执行副总裁兼总经理Forrest Norrod表示。

据了解，UALink专家组将制定用于管理数据中心中不同GPU之间连接的标准，并预计于2024年第三季度将这些标准提供给加入UALink联盟的公司。而博通据传已开始生产UALink交换机。“谷歌、Meta、微软、AMD等对英伟达NVLink的垄断已经隐忍很久。”研究机构Omdia AI行业首席分析师苏廉节对记者表示，UALink的出现会减少业界对英伟达通信协议的需求，只不过对英伟达的龙头地位影响不大，但确实对业界是很重要的。


6月2日，英伟达CEO黄仁勋在中国台北国际电脑展上，发表“开启产业革命的全新时代”演讲。 视觉中国/图

## NVLink：将系统扩展为超算

NVLink、交换以太网结合在一起让英伟达把200多块GPU连接起来，成为一个AI性能“爆炸”的超算系统。

简单而言，NVLink是英伟达开发的一种总线及通信协议，采用点对点结构、串列传输，既可用于连接中央处理器（CPU）与图形处理器（GPU），也可用于多个GPU之间相互连接的技术标准。

相比于传统的PCIe，该技术可实现多个GPU之间的高速数据传输和协同工作。2018年，英伟达首次向公众推出NVLink技术，其架构包括NVLink桥接器和NVLink交换机。

记者通过英伟达中国官网了解到，其NVLink技术构成多元，既包含软件协议，又有芯片这样的硬件。

根据英伟达中国的定义，NVLink由一个强大的软件协议组成，一般通过印在电路板上的多对导线实现，可以让处理器以闪电般的速度收发共享内存池中的数据。据了解，NVLink最初作为NVIDIA P100 GPU的互联通道推出，之后便与每一代新的NVIDIA GPU架构同步发展。

相比于传统x86服务器的互联通道PCIe，NVLink主打的就是速度快、能效低。比如，第四代NVLink连接主机和加速处理器的速度高达每秒900GB，是PCIe 5.0带宽的7倍多，而每传输1字节数据仅消耗1.3皮焦，NVLink的能效是PCIe 5.0的5倍。

另外，NVLink所包含的NVIDIA NVLink-C2C则是一种板级互联技术，它能在单个封装中将两个处理器连接成一块超级芯片，Grace Hopper超级芯片就是NVIDIA NVLink-C2C将Grace CPU和Hopper GPU连接而成。

对于NVLink的作用，英伟达中国用了一个形象比喻：“NVLink就像是乐高积木的凸粒和凹槽。”在2024年的GTC大会上，英伟达已对外公布了第五代NVLink，其总带宽达到1.8兆字节/秒（TB/s），是上一代产品的2倍。

英伟达中国指出，NVLink是一项关键的技术，它可以让用户将模块化的NVIDIA DGX系统扩展成为一个AI超级计算机。

利用DGX内部的NVLink网络与两者之间的NVIDIA Quantum-2 InfiniBand交换以太网，用户就可以将32个DGX系统模块连接成一台AI超算。例如，一台NVIDIA DGX H100 SuperPOD包含256个H100 GPU，可提供最高1 EXAFLOP的峰值AI性能。

总结来看，NVLink、交换以太网结合在一起让英伟达把200多块GPU连接起来，成为一个AI性能“爆炸”的超算系统。为什么英伟达要去做GPU之外的硬件产品？其实这里面还有一个趣事。早在10多年前，英伟达首席科学家Bill Dally找黄仁勋谈面向HPC开发networking技术的问题，黄仁勋问他：“我们为什么要做networking？我们不是一家开发GPU的公司吗？”

黄仁勋虽然有疑问，但后来还是全力支持该技术的开发。“但他当时的质疑是合情合理的。这个问题延伸来可能还囊括了英伟达为什么要收购Mellenox？为什么要做DPU（数据处理器）？为什么要做交换芯片和交换机？为什么要研究封装之间的光通信技术这些问题。”资深产业观察人士黄烨锋表示。

在今年GTC大会上，英伟达宣称Blackwell架构的GPU的推理能力相比于前代Hopper，有了30倍的提升。“这里的30倍当然不是芯片层面的，摩尔定律、超越摩尔或任何摩尔都做不到隔代30倍性能提升；特定数据格式的支持强化，以及更重要的NVLink互联技术升级、NVSwitch芯片引入，才是GB200 NVL72整个系统在多模态模型推理上达成30倍性能提升的关键。”黄烨锋认为，这些都是“底层硬件”，只不过是扩展到了系统层面，互联、存储、解热都是其中关键。

“解决跨节点通信的瓶颈，显然是生成式AI时代最关键的组成部分之一。这些东西对英伟达来说都属于生态，也是很多在PPT上吊打英伟达的竞争对手难以逾越的障碍。”黄烨锋表示。

而UALink的亦步亦趋，说明了微软、谷歌等巨头们从造芯之后，追赶英伟达的步伐又往前迈了一步。这也反过来证明，英伟达从做GPU开始向外拓展开发互联技术、以太网等的技术路径是符合行业发展趋势的。

## UALink：对业界很重要

该联盟虽然可能会对英伟达NVLink构成一定威胁，但是否能够取得成功，还需要时间来观察。

根据UALink的计划，首个UALink 1.0版本将允许AMD的Instinct GPU或英特尔的Gaudi等专用处理器之间的直接数据传输，从而提高AI计算的性能和效率。

“目前如果要用英伟达的GPU，就必须用其NVLink。”苏廉节表示，谷歌、Meta、微软是云大厂，长期大规模化地部署GPU，AMD、英特尔则是AI芯片供应商，而博通、思科和惠普负责数据中心连接，“他们对英伟达NVLink的垄断已经忍很久了”。

值得一提的是，在2023年7月，诸多云服务提供商、芯片制造商、系统供应商就联合组建了超级以太网联盟UEC，期望构建基于以太网的完整通讯栈架构，用于高性能网络，主要为了适配AI和HPC。

以太网或InfiniBand的主要作用，是连接包含GPU的服务器。同样，英伟达也没有加入超级以太网联盟。2019年3月收购Mellanox后，英伟达基本独占了高性能InfiniBand互联市场。

UALink联盟的8家发起厂商指出，成立一个开放行业标准机构来制定相关技术规范，以促进新使用模式所需的突破性性能，同时支持数据中心加速器用开放生态系统的发展。“行业规范对于建立下一代AI数据中心标准化以及实施AI、机器学习、HPC（高性能计算）和云应用程序的接口至关重要。”他们在一份声明中如此表示。

据悉，UALink联盟的核心于2023年12月就已经建立。UALink联盟成员表示，系统制造商将创建使用UALink的机器，并允许客户将来自许多参与者的加速器放入系统中。比如，用户可以把来自AMD、英特尔或其他第三方的GPU（AI加速器）连接在一起。UALink1.0版规范预计将于2024年第三季度推出，并向参加超级加速器链联盟的公司开放。

“UALink作为一个行业开放标准，将有助于推动人工智能数据中心的发展。通过实现更高效的数据传输和通信，它将改变GPU网络的整体性能。”深度科技研究院院长张孝荣表示，该联盟虽然可能会对英伟达NVLink构成一定威胁，但是否能够取得成功，还存在很多未知因素，比如能否达到预计的效果、市场接受程度等都需要时间来观察。

苏廉节也表示，虽然这个联盟对于英伟达几乎没影响，但对业界很重要。

事实上，对标英伟达的NVLink，由于其是软硬件一体的，那UALink联盟成员会放弃已有的软硬件产品开发而按照该标准开发新的产品吗？

对此，电子创新网CEO张国斌表示：“一般就控制风险而言，大家会两手抓。目前来看，这些公司更有可能采取一种混合策略，而不是完全放弃已有的软硬件产品开发。这意味着他们将继续支持和发展现有的产品，同时逐步适应并整合新的开放标准。”

而根据UALink团队的对外发声，将UALink标准落地成产品2024年还太早，2026年将是一个快速实施的时间点。

# 价格战未熄火 大模型下一个战场在哪儿

**本报记者 曲忠芳 北京报道**

国产AI大模型在今年5月掀起的价格战火一直烧到了本月。6月5日，智谱AI在开放日活动上宣布对旗下全模型矩阵进行降价，这是该公司在不到一个月时间里做出的第二次价格下调动作。

《中国经营报》记者首先以智谱AI旗下GLM-3-Turbo作为考察目标，这是一款于2023年10月发布的大模型。今年5月11日，智谱AI宣布将GLM-3-Turbo模型每千tokens（token是大模型文本处理的最小单位）的价格从0.005元降低至0.001元，换算即可得出每百万tokens的价格为1元；到6月5日，GLM-3-Turbo的最新价格是每百万tokens费用为0.6元，不难看出，在不到一个月的时间里，GLM-3-Turbo的价格从5元/百万tokens历经两次降价达到0.6元/百万tokens。

针对近期的大模型厂商集体降价潮，智谱AI首席执行官张鹏回应称，大模型商业化策略“并不是简单的价格战”。智谱AI“切实通过模型核心技术的迭代创新和效率的提升，实现了应用成本的持续降低，以及客户价值的持续升级”。

大模型的价格战远未熄火，市场竞争日益白热化，如何将大模型转化为真正的生产力，如何在同质化严重的AI体（AI Agent）生态战中拔得头筹，打造出“杀手级”应用，已成为大模型厂商需要回答的新考题。

## 将大模型转化成切实生产力

智谱AI首席运营官张帆指出：“去年年初GLM大模型还是每千tokens为0.5元，到6月5日这一年多里完成了大幅降价。”在他看来，“价格（降低）是快速推动大模型API（应用程序编程接口）的必要路径，让AI更加普惠，从而使每个企业都能够非常容易地使用全系列模型服务”。

小米集团小爱团队总经理王刚在智谱AI开放日活动上指出，覆盖上亿用户的产品每日tokens接近2000亿至3000亿，如果想要覆盖所有用户，这一规模量级所需要的成本相对低端或大众的机器而言还存在一定的压力。大模型的降价，使产品运营方能够有机会将大模型能力覆盖全终端设备。同时他也强调，模型的性能效果同样重要，价格下降了，效果别跟着下降。

智谱AI方面介绍，除了GLM-3-Turbo之外，智谱AI的多模态图生文模型GLM-4V每百万tokens的价格从100元降至50元，而文生图CogView-3模型的价格则从0.25元/张下降60%至0.1元/张。更为重要的是，智谱AI的MaaS（模型即服务）开放平台进行了一系列升级，尤其是最新开源的GLM-4-9B模型、GLM-4-Air模型。GLM-4-Air主打高性价比，每百万tokens花费1元，推理速度提升300%的同时，每秒可以输出71个tokens，超过人眼的阅读速度。

“2024年是AGI（通用人工智能）元年，AI技术增长进入了一个全新阶段。”张鹏指出，“技术曲线的陡峭与让更多人感受到AI强大生产力之间还有非常大的差距，如何把大模型能力真正用来帮助人们的工作、学习和生活，其中还有很多工作要做。”

据介绍，截至6月5日，智谱AI开放平台上的企业级用户数量超过30万，每日模型调用量已超过400亿tokens。最近的6个月里，大模型API每日消费量呈现出50倍以上的增长，这些数据表明越来越多的企业真正把模型应用到了日常的工作中，初步完成了生产力的转换。

基于业务感知与洞察，张帆总结了企业客户对于大模型的四项集中需求：一是模型的性能强、速度快；二是服务成本更低；三是对于构建私有模型的需求；四是将模型转化为业务价值。针对这一需求趋势，智谱AI全面升级模型服务，帮助企业客户仅需三步可以完成私有模型的训练，即准备数据、创建微调任务、部署训练完成的模型。企业可以选择LoRA微调、全参微调两种模式，前者主打高性价比，例如GLM-4每千tokens仅需0.4元，而GLM-4-Air只需要0.03元；后者则相当于探索模型微调的极限。

记者观察到，近一个多月，无论是OpenAI这些海外企业，还是国产大模型厂商，AI大模型的性能迭代、成本降低成为主流趋势，大模型朝着易用的方向加速前进。银河证券研报中指出，大模型领域价格战进入白热化阶段，推理成本的下降将持续推动AI应用加速落地。

## 厂商纷纷加码AI体

模型迭代升级、价格持续下探，让企业应用AI大模型的门槛不断降低。更为重要的是，在众多的大模型服务平台，谁能率先构建成应用生态，甚至打造出一款“杀手级”的产品，更是摆在大模型厂商面前的考验。

AI Agent，即AI智能体，或简称“智能体”“AI体”，是指由生成式AI生成的各类工具或助手等应用，目前已成为市面上主流大模型的标配功能，背后实质也是大模型厂商构建模型生态的重要一环。

记者观察到，腾讯元宝“发现”页中已有数十款应用上线，覆盖外语学习、招聘、营销、绘画、美食、社交等各类工作生活场景。尤其值得一提的是，近期热播的电视剧《庆余年2》中的主要人物IP也已在元宝平台上线；阿里云的通义平台既提供了听课开会、办公提效、学习工具三大类“工具”，又在“百宝袋”中提供了趣味生活、创意文案、办公助理、学习助手等不同场景的多款垂直应用；另一家大模型应用Kimi的Kimi+平台提供的AI智能体应用同样覆盖办公提效、辅助写作、社交娱乐、生活实用等类别的不同智能体。其他大模型平台的各类智能体也呈现相似的状态。

张鹏分享了智谱AI旗下智谱清言的最新进展，目前已有超过30万个智能体活跃在清言App上，包括诸如思维导图、文档助手、日程安排等生产力工具。“它们不仅是个人的得力助手，也是每个人的助理天团，基于GLM模型的能力及开发者的想象力，越来越高效和精准地帮助用户解决问题。”张鹏如是说道。

值得注意的是，智谱清言还推出了一项新功能“清流”，支持在同一个对话内调用不同智能体协同工作。智谱AI还为《三体》作者、科幻作家刘慈欣，带货主播罗永浩等知名IP设计了智能体，并开放供C端用户使用。刘慈欣认为：“在未来，当人工智能拥有超过人类的智力时，想象力可能是我们相对AI的唯一优势，即使在面对宇宙级别的挑战，人类的创想和智慧，也是我们宝贵的财富。”

中信证券在5月底发布的研报分析指出，AI智能体作为当前语言模型应用落地的最佳形式，有望将迎来技术转折。展望后续AI Agents应用的发展路径，成本优化将是焦点，目前的技术方案提供了多种针对成本问题的优化方案，但仍需要时间来进行实践，预测距离AI Agents应用落地还有6—12个月时间。因此，究竟哪家企业能在生成式AI时代率先打造出真正主流的“杀手级”应用，显然还有待时间的验证。