万亿美元蜂拥AI芯片 多方“围剿”英伟达

本报记者 秦枭 北京报道

过去一年里，由ChatGPT引爆的AI浪潮愈演愈烈，英伟达（NASDAQ：NVDA）作为这场浪潮背后最大的“卖铲人”，其GPU芯片价格被炒高数倍，但市场上仍一卡难求，英伟达因此赚得盆满钵满。

重金造芯

一颗芯片从立项到量产至少两年的时间，而一家晶圆代工厂的投产需要3—5年的时间。

近日，OpenAI CEO山姆·奥特曼计划筹资高达7万亿美元建立一个新的芯片帝国的消息不胫而走，其欲通过建立一个新的芯片生态系统，将制造商、供应商和用户聚集在一起，以满足全球人工智能需求。

实际上，在过去很长一段时间里，奥特曼公开抱怨英伟达GPU显卡稀缺已久。OpenAI也一直重点讨论AI芯片的供需问题。根据花旗研究分析师Christopher Danely在2023年7月发布的一份报告，英伟达在AI训练领域占据了“至少90%”的市场份额，这也给OpenAI造成了价格和供应两方面的制约。2023年10月，有消息曝出，OpenAI计划自研AI芯片已有一段时间，甚至已经开始评估潜在的收购目标。

虽然大多数人对于OpenAI自研AI芯片早有预期，但7万亿美元的融资数额还是令人咋舌，因为这一金额相当于2023年全球半导体行业总营收的14倍。根据Gartner预测，2023年全球半导体行业的总收入是5330亿美元。

因此，也有业内人士调侃，有7万亿美元为何不直接买下英伟达。截至2024年2月22日，英伟达市值约为1.7万亿美元。

对此，英伟达创始人兼CEO黄仁勋略带嘲讽地表示：“（7万亿美元）显然能买下所有的GPU。但是，计算机架构其实在不断地进步。”

与奥特曼“浮夸”的7万亿美元投资计划相比，孙正义1000亿美元的造芯计划显得实际许多。

据悉，该计划项目代号为“伊邪那岐”（Izanagi），孙正义计划向该项目投入300亿美元，而额外的700亿美元可能来自中东机构。孙正义希望新成立的公司能与软银旗下半导体设计公司Arm的业务互补，并打造一个最新的AI芯片巨头，以抗衡英伟达。不过，软银集团对于项目资金来源和具体用处并未透露。但值得注意的是，孙正义曾与奥特曼讨论过合作建立半导体业务和筹募资金等相关事项。

在天使投资人、人工智能专家郭涛看来，无论是1000亿美元还是7万亿美元，要在短时间内实现AI芯片供应“自由”，还需要解决以下问题：首先，AI芯片的研发需要大量的资金和人力投入，而且技术难度很大，需要相当长的研发周期；其次，除了芯片本身之外，还需要考虑与之配套的软件、硬件等产业链的建设，这也是一个长期过程；最后，即使成功研发出高性能的AI芯片，也需要市场的接受和认可，这需要时间和市场的教育。

据业内人士透露，一颗芯片从立项到量产至少两年的时间，而一家晶圆代工厂的投产需要3—5年的时间。

“英伟达不会永远在大规模训练和推理芯片市场占据垄断地位。”特斯拉CEO马斯克曾表示。英特尔CEO基辛格也曾透露：“整个行业都被推动来减少CUDA（英伟达推出的运算平台）的市场。”

随着人工智能热潮的不断升温，传统豪门与行业新贵向英伟达筑下的AI芯片“护城河”发起了进攻。一方面，英特尔、AMD等传统豪门正在推进新一轮AI芯片研发计划；另一方面，以Groq为代表的初创公司亦在积极推动自主研发芯片，还有日本软银集团创始人孙正义以及OpenAI欲携千亿乃至万亿美元资金入局，AI芯片的战场霎时间硝烟四起。

虽然英伟达面临豪门和新贵的前后围堵，但行业真正摆脱英伟达并非易事。 视觉中国/图

新贵走红

不只Groq，还有其他AI芯片新贵也在对英伟达虎视眈眈。

就在业内还在争论孙正义的1000亿美元的“脚踏实地”和奥特曼7万亿美元的“仰望星空”之时，人工智能芯片公司Groq已一夜走红，其推出的大模型推理芯片LPU，推理速度较英伟达GPU提高10倍，成本只有其1/10；运行的大模型生成速度接近每秒500 tokens，碾压ChatGPT-3.5大约40 tokens/秒的速度。

据悉，Groq成立于2016年，定位为一家人工智能解决方案公司。值得注意的是，在Groq的创始团队中，有8人来自谷歌早期仅有10人的TPU核心设计团队。例如，Groq创始人兼CEO Jonathan Ross设计并实现了TPU原始芯片的核心元件，TPU的研发工作中有20%都由他完成。

根据Groq官网的介绍，LPU是一种专为AI推理所设计的芯片。驱动主流大模型的GPU，是一种为图形渲染而设计的并行处理器，有数百个内核，而LPU架构则与GPU使用的SIMD（单指令，多数据）不同，这种设计可以让芯片更有效地利用每个时钟周期，确保一致的延迟和吞吐量，也降低了复杂调度硬件的需求。

郭涛解释道，LPU是一种专为线性代数运算优化的处理单元，而线性代数是深度学习和AI模型中的核心计算任务。与传统的GPU相比，LPU可能在架构上进行了特定的优化，以更高效地执行矩阵运算和向量计算，这些是大语言模型（LLM）和其他深度学习模型的关键操作。GPU最初设计用于处理图形和图像，但它们在并行处理大量数据方面表现出色，这使得它们非常适合深度学习任务。GPU拥有大量的核心，可以同时处理多个任务，但它们可能在执行特定类型的数学运算时不如专门为这些运算设计的ASIC芯片高效。

Groq创始人兼首席执行官Jonathan Ross曾表示，在大模型推理场景，Groq LPU芯片的速度比英伟达GPU快10倍，但价格和耗电量都仅为后者的十分之一。而且他还强调，Groq的芯片，由于技术路径不同，在供应方面比英伟达更充足，不会被台积电或者SK海力士等供应商卡脖子。

不过，并不是所有人都认可Groq。Facebook原人工智能科学家、阿里技术原副总裁贾扬清在推特上算了一笔账，因为Groq内存容量只有230MB，在运行Llama-270b模型时，需要305张Groq卡才足够，而用H100则只需要8张卡。从目前的价格来看，这意味着在同等吞吐量下，Groq的硬件成本是H100的40倍，能耗成本是10倍。

不只Groq，还有其他AI芯片新贵也在对英伟达虎视眈眈。据The Information统计，全球有超过18家用于AI大模型训练和推理的芯片设计初创公司，包括Cerebras、Graphcore、壁仞科技、摩尔线程、d-Matrix等，融资总额已超过60亿美元，企业整体估值共计超过250亿美元（约合1792.95亿元人民币）。

稳坐钓鱼台

针对围追堵截是否会影响到英伟达市场份额的问题，黄仁勋不以为然。

“山雨欲来风满楼”，英伟达依旧“稳坐钓鱼台”。虽然英伟达面临豪门和新贵的前后围堵，但行业真正摆脱英伟达并非易事。上述业内人士对记者表示：“目前，国内的英伟达高端AI芯片依旧非常紧缺，就连之前大家嗤之以鼻的阉割版H20系列也极为抢手。”

这在英伟达最新的财报中可见一斑。2月22日，英伟达公布了截至2024年1月28日的2024财年业绩报告，英伟达全年的营收创历史新高，为609亿美元，增长126%。其中，在2024财年第四财季，公司的营收也创纪录地达到了221亿美元，同比增长265%，数据中心业务收入为184亿美元，同比增长409%，环比增长27%。

英伟达方面表示，在2024财年第四财季，数据中心的增长是由跨越不同行业、用例和地区的生成式AI和大型语言模型的训练和推理推动的。数据中心平台的多功能性和领先性能可为许多用例带来高投资回报，包括AI训练和推理、数据处理和广泛的CUDA加速工作负载。“我们估计，去年数据中心大约40%的收入来自AI推理。”

英伟达CFO科莱特·克雷斯则在财报电话会议时表示，公司下一代产品的市场需求远超过供给水平，尤其是该公司预计今年晚些时候发货的新一代芯片B100。他表示，“构建和部署AI解决方案已经触及几乎每一个行业”，预计数据中心基础设施规模将在五年内翻番。

黄仁勋也表示：“加速计算和生成式AI已经达到了引爆点。全球各地的公司、行业和国家的需求正在激增。”

针对围追堵截是否会影响到英伟达市场份额的问题，黄仁勋不以为然，他认为：“从根本上来说，我们认为2025年及以后的持续增长条件仍会非常好。由于生成式人工智能以及整个行业从CPU转向GPU，英伟达GPU的需求仍将保持较高水平。”

不过，黄仁勋也表示，供应状况虽然在改善，但仍面临短缺，供应受限状况将在全年时间内持续下去。由于生成式AI以及整个行业的计算硬件需求从CPU向英伟达制造的加速器转移，市场对公司GPU的需求将保持高涨。