ASIC，AI 芯片霸权的终结

原创 来觅研究院 RimeData 来觅数据

撰稿 李沛瑶 2024-12-25

导读：12 月 13 日，博通公布新一季度财报，截至目前，股价大涨超 30%。与此同时，英伟达开始了漫长的阴跌，较高点跌幅近 10%，市值蒸发约 3000 亿美元。二者泾渭分明的关键正是博通新推出的 ASIC 芯片服务。博通如何撼动英伟达在算力领域的霸权？ASIC 有何过人之处？国内投融现状如何？本文尝试分析和探讨。

被寄予厚望的博通

博通是一家专注于半导体和基础设施软件解决方案的多元化公司，在无线连接、网络处理器和定制 AI 芯片领域取得了领先优势。12 月 13 日，公司 CEO Hock Tan 在业绩说明会上表示，公司 2027 年超大规模客户的 AI 收入将达到 600-900 亿美元，几乎每年翻倍。上述言论一举燃爆市场热情，博通股价当日上涨超 24%，成为人类历史上第九家市值突破 1 万亿美元的公司。

博通说的 AI 收入究竟是什么？除了其优势的交换机业务外，最大的部分在于 ASIC 芯片服务。ASIC 是一种为某种特定任务设计的芯片，一般会被应用于特定设计和制造的设备中，执行必要的功能。在 AI 芯片中，ASIC 被用来处理特定的任务，且相比 GPU 而言，拥有更高的处理速度和更低的能耗。

图表 1：AI 芯片的分类

AI 芯片	定义	优势	典型厂商
GPU	通用图形处理器	高并行结构，生态体系成熟，跨平台支持，易于编程，成为主流的并行数据处理加速器	英伟达、AMD、海光信息
ASIC	专用集成电路	专门为深度学习计算定制的芯片，如神经网络处理器 NPU，张量处理器 TPU，效率高，功耗低，体积小	博通、Marvell、寒武纪
FPGA	现场可编程逻辑阵列	高度并行的结构和低延迟，可编程和灵活性强，能够适应模型算法迭代	赛灵思、Altera、紫光国微

资料来源：公开资料、来觅数据整理

按功能分类，AI 芯片可以分为训练卡和推理卡两个类型。训练卡也叫大卡，通常拥有更高的计算能力和内存带宽，以支持训练过程中的大量计算和数据处理；推理卡也称小卡，其参数较低，只需满足推理需求。一般情况下，训练卡可以作为推理卡使用，但推理卡不能作为训练卡使用。简单来说，大模型的训练需要大量的训练卡形成显卡集群，而在应用上，则需要推理卡运行 AI 模型进行计算。

在 2023 年以来的 AI 大规模基建中，“百模大战”推升了算力需求。对大模型的预训练中，训练卡是焦点，而 GPU 由于高适配性、性能强大成为了训练卡的标配。英伟达也几乎垄断了所有算力市场，其在 AI 芯片市场占有率超过 90%。

英伟达为什么这么强大？原因在于英伟达卖出的不仅仅是算力芯片，而是一整套生态系统。英伟达生态系统中手握三张王牌，包括领先的 GPU、十年磨一剑的 CUDA 以及网络传输 NvLink。英伟达自 2010 年以来发力 AI 算力，尤其是近年来推出的 H100、H200、GB200 等，单卡算力稳坐第一梯队。CUDA 是一套芯片编程模型，为开发者提供了利用 GPU 进行高效并计算的全方位支持。NvLink 采用点对点结构，通过串列传输实现高速数据运输，传输速率是传统 PCIE 的 7 倍。

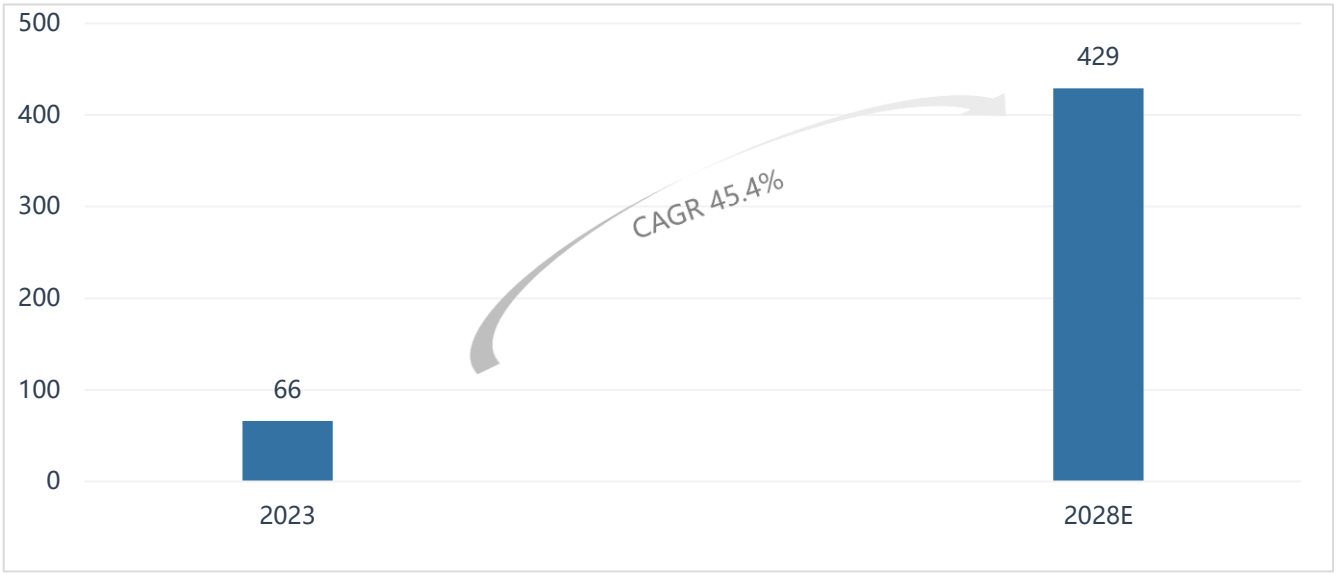
不过，目前局面正在改变，博通可能是最接近英伟达的挑战者。一方面博通在网络芯片、交换机和路由器

方面有强大的积累，其推出的千兆以太网方案，其传输速率不弱于 Nvlink。另一方面，博通在半导体领域浸润多年，可帮助客户完成从设计到封装所有的环节。在生态方面，由于执行特定的任务，一般不需要复杂的生态环境。简而言之，博通打出的组合拳约等于以自己在交换芯片方面的积累和定制芯片加上客户自己的软件硬撼英伟达。

前景很美好，但现实很骨感。就目前而言，英伟达其三季度在 AI 方面收入 308 亿美元，而博通最近一个季度 AI 收入仅为 37 亿美元，体量差距还相当大。而市场对博通的追捧可能仅仅意味着反抗英伟达霸权的开始。

据 Marvell 预测，2023 年 ASIC 占数据中心加速计算芯片的 16%，规模约为 66 亿美元；随着 AI 计算需求的增长，ASIC 占比有望提升至 25%，预计 2028 年数据中心 ASIC 市场规模将提升至 429 亿美元，CAGR 为 45.4%。就整体规模而言，AI ASIC 增速相对较快，但赶上 GPU 尚需时间。

图表 2：2023-2028E 全球 ASIC 市场规模（亿美元）



资料来源：Marvell，来觅数据整理

焦头烂额的英伟达

英伟达最近不是很太平。从官宣的 GB200 AI 服务器出货延迟，到被美国、中国、欧洲等全球多地政府联合调查反垄断，英伟达似乎已经陷入多事之秋。然而，这并不能证明英伟达的虚弱。英伟达的麻烦似乎仅仅只是他强大的副产品。

英伟达傲人的市场份额证明他在 AI 芯片领域几乎已无敌手。就目前而言，各大云厂商选择英伟达的产品时，要么选择捆绑销售，要么买多者先得，要么只能原地等上好几个月。正如 AWS CEO 表示：“目前 GPU 市场只存在一种选择，那就是英伟达，若是市场上有更多选择，我们相信客户会欢迎。”

云厂商并非没有准备，它们正着手开发 ASIC 以减轻对英伟达的依赖。如微软已推出了首款用于内部数据业务的数据处理器 Azure Boost DPU，亚马逊宣布将推出 Trainium2 芯片。最激进的当属谷歌，其推出的 Trillium TPU，已用于大模型 Gemini 2.0 的训练中，而帮助谷歌完成这一宏伟设计的，正是博通。

博通与谷歌的合作开始于 2016 年，迄今为止已经迭代至第七代产品。在最新一期的财报电话会议中，博通表示谷歌、Meta、亚马逊都是公司 AI 定制芯片的大客户，此外，还有两家大型客户正在要求博通对其产品进行深入开发。博通 CEO Hock Tan 表示，在未来，50%的算力都会是 ASIC，至于超大规模的云计算厂商，他们将 100%使用 ASIC。

不过，花旗的多位分析师对此表达了不一样的看法，他们认为到 2028 年，GPU 至少会占据 AI 加速器总市场规模的 75%，而博通为代表的 ASIC 阵营将至少占据 25%。不管怎样，GPU 市场的一部分空缺将会被 ASIC 吃下，而这背后代表的，正是 AI 算力生态的变化。

前 OpenAI 创始人 Ilya Sutskever 在不久前的 NeurIPS 2024 大会上陈述了一个观点：“由于数据是有限

的，因此模型的预训练时代即将结束，AI 的重心将由训练转向推理。”过去的一年里是属于大模型追赶者的：行业领导者 OpenAI 的 GPT-5 频繁跳票，追赶者的距离越拉越小。与此同时，AI 应用开始风起云涌，AI Agent、端侧 AI 开始频繁出现在 C 端用户视野里。

在预训练时代，AI 芯片主要是训练卡。而 GPU 芯片由于灵活且兼容，占据了大部分的市场规模。不过事情正在悄悄变化，对推理领域的倾斜将会导致 AI 芯片的格局发生改变。而由于 ASIC 芯片更加“专一”，更快的处理速度和更低的能耗下，它也被广泛认为更适合推理端。

不过，英伟达在训练卡上的壁垒仍然牢不可破。目前云厂商对博通的追捧并非完全出于性能要求，而更多的是对英伟达的替代考虑。而只要大模型还在继续迭代，算力需求还在增长，英伟达的优势就依然存在。市场对 ASIC 的追捧也许会让英伟达头疼，但却不会大伤元气。

中国 AI 芯片投融动态

由于全球算力的不平衡，目前国内与海外相比存在一定的差异。目前国内 AI 芯片公司多以 ASIC 为主，如知名的昇腾、寒武纪等都属于这一品类。在近两年的全球大模型竞赛中，中国企业并未落后太多，而在未来百花齐放的应用时代，ASIC 将不再成为软肋，也将随着 AI 芯片的发展大放异彩。

AI 芯片市场近年来呈现出强劲的增长势头。2024 年全球 AI 芯片市场规模预计将达到 712.52 亿美元，同比增长 33%，并有望在 2025 年进一步增长至 919.55 亿美元。在中国市场，2023 年 AI 芯片市场规模达到 1206 亿元，同比增长 41.9%，预计 2024 年将增长至 1412 亿元。据来觅数据显示，AI 芯片亦是今年最为活跃的赛道之一，融资轮次仍偏向早期，但部分明星项目已得到市场认可，资本正不断加码。感兴趣的读者，可以登录 Rime PEVC 平台获取 AI 芯片赛道全量融资案例、被投项目及深度数据分析。

图表 3：AI 芯片 2024 年以来投融事件

融资方	来觅赛道	融资时间	融资轮次	融资金额	投资方
太初元碁	AI 芯片	2024-11-29	A3 轮	未披露	宝山国投集团、青创投等
紫光展锐	AI 芯片	2024-11-29	E+轮	近 20 亿人民币	元禾璞华
光羽芯辰	AI 芯片	2024-11-21	天使轮	未披露	讯飞科创
长沙韶光	AI 芯片	2024-11-19	战略融资	3 以人民币	工银资本
行云	AI 芯片	2024-11-18	天使+轮	数亿人民币	同创伟业
紫光同创	AI 芯片	2024-11-14	战略融资	未披露	交银资产
瀚博半导体	AI 芯片	2024-09-25	C+轮	未披露	中网投、经纬创投等
天数智芯	AI 芯片	2024-09-18	战略融资	111 万人民币	鼎祥资本
中昊芯英	AI 芯片	2024-09-14	战略融资	2.5 亿人民币	艾布鲁
知合计算	AI 芯片	2024-09-09	A1 轮	数亿人民币	源码资本、临港科创投等
图影视讯	AI 芯片	2024-09-07	A 轮	数千万人民币	华西银峰
沐曦	AI 芯片	2024-08-23	股权融资	未披露	浦东资本、启夏私募投资等
太初元碁	AI 芯片	2024-08-09	A2 轮	数亿人民币	霜叶创投、金蚂投资
寒序科技	AI 芯片	2024-08-05	种子轮	未披露	零以创投
元启半导体	AI 芯片	2024-07-15	A 轮	数亿人民币	奥飞数据、闻名投资
昆仑芯	AI 芯片	2024-06-12	C 轮	未披露	北京国管、君联资本
紫光展锐	AI 芯片	2024-06-03	E 轮	超 40 亿人民币	中信建投资本、交银金融资本等
后摩智能	AI 芯片	2024-05-21	A++轮	数亿人民币	中移资本
亿铸科技	AI 芯片	2024-05-11	B 轮	超 1 亿人民币	盛视科技、行至资本等
灵汐科技	AI 芯片	2024-04-16	B+轮	未披露	国鼎资本
润芯微科技	AI 芯片	2024-04-02	B 轮	未披露	工商银行、核聚资本等
摩尔线程	AI 芯片	2024-03-04	C 轮	未披露	华通车业
太初元碁	AI 芯片	2024-02-07	战略融资	数亿人民币	龙芯资本
知合计算	AI 芯片	2024-01-10	Pre-A 轮	数亿人民币	联新资本、临港科创投等

资料来源：来觅数据