AMD vs 英伟达：AI成决定性战场

本报记者 秦枭 北京报道

AMD与英伟达的显卡之争已经持续了20余年，双方势均力敌，交错领先。但在最近以AI大模型为战场的芯片之争中，英伟达优势尽显。据统计机构Jon Peddie Research的数据统计报告显示，截至2022年四季度，英伟达（NVIDIA）占据了独立GPU84%的市场份额。其中，AI芯片当居首功。8月8日，英伟达再向市场展示了新一代英伟达超级AI芯片GH200。其创始人兼CEO黄仁勋说：“GH200为世界上最快的内存。”有业内人士预测，英伟达与AMD之间的差距将进一步拉大。

不过，《中国经营报》记者了解到，AMD正计划在2023年第四季度扩产其AI芯片MI300系列，并保证在2024年供应充足，以此来应对英伟达的“一家独大”，其CEO苏姿丰预测，新芯片的发布将为其带来强劲的业绩增长，预计2023年包括MI300芯片在内的销售额将超过2022年的60.4亿美元。

英伟达强势“称雄”

在上一代还未量产上市的背景下，英伟达又在8月8日世界计算机图形学会议SIGGRAPH上，由黄仁勋发布了H100的升级版GH200。

“2018年是一个‘孤注一掷’的时刻，要求我们重新发明硬件、软件、算法。而当我们用AI重新发明CG（计算机图形学）时，我们也在重新发明GPU（图形处理器）用于AI。”黄仁勋5年前的豪赌正在迎来收获期。

ChatGPT的横空出世，让AI产业迎来“iPhone时刻”，科技巨头争相谋局落子，继微软、谷歌之后，国内企业百度、阿里巴巴等也先后发布大模型，并进行用户测试和企业应用接入。汹涌的人工智能浪潮导致创建高级人工智能程序所需的芯片极为短缺。

为了训练ChatGPT，OpenAI构建了由近3万张英伟达V100显卡组成的庞大算力集群，GPT-4更是达到了100万亿的参数规模，其对应的算力需求同比大幅增加。TrendForce分析认为，要处理近1800亿参数的GPT-3.5大型模型，需要2万颗GPU芯片，而大模型商业化的GPT需要超过3万颗，GPT-4则需要更多。

GPU Utils 8月4日公布的一组数据显示：OpenAI的GPT-4可能在1万到2.5万张A100 GPU芯片上进行训练；Meta拥有约21000个英伟达A100；特斯拉拥有约7000个A100；Stability AI拥有约5000个A100；Falcon-40B模型（400亿参数）在384个A100上进行训练。根据马斯克的说法，GPT-5可能需要3万~5万张H100显卡。

即便台积电开足马力为英伟达生产，也仍存在巨大缺口。根据GPU Utils的测算，AI芯片H100在2023年8月的市场总需求可能在43.2万张左右，而目前一张H100芯片在eBay上的价格甚至炒到了4.5万美元，折合人民币超过了30万元。

近几年，与英伟达类似，AMD的战略重心也在向AI转移。 视觉中国/图

凭此，英伟达今年以来的股价上涨超200%，市值突破了1万亿美元，成为全球最有价值的芯片公司。

然而，英伟达并未满足于此，开始频频推出新款GPU用来提升AI训练能力。今年3月，英伟达发布了H100 NVL GPU、L4 Tensor Core GPU、L40 GPU、NVIDIA Grace Hopper四款AI芯片，以满足生成式AI日益增长的算力需求。

在上一代还未量产上市的背景下，英伟达又在8月8日世界计算机图形学会议SIGGRAPH上，由黄仁勋发布了H100的升级版GH200。

据了解，GH200全球首发采用HBM3e高带宽内存，与英伟达目前最高端的AI芯片H100使用同样的GPU，但不同之处在于，GH200将同时配备高达141GB的内存和72核ARM中央处理器，每秒5TB带宽。和现有Grace Hopper型号相比，最新版本的GH200超级芯片能够提供3.5倍以上的内存容量和3倍以上的带宽。和H100相比，GH200超级芯片的内存增加了1.7倍，带宽增加了1.5倍。全新一代的GH200预计明年二季度开始生产。

黄仁勋表示，一台服务器可以同时装载两个GH200超级芯片，大型语言模型的推理成本将会大幅降低。按照黄仁勋的介绍，在相同的成本（1亿美元）下，2500块GH200组成的计算中心，在AI计算的能效上，要比传统的CPU计算中心高20倍。

东方证券研报指出，英伟达仍旧牢牢占据AI基础设施领域的主导地位。自ChatGPT引领生成式AI浪潮以来，NVIDIA GPU已成为支持生成式AI和大模型训练的大算力AI芯片首选。随着此次GH200超级AI芯片的升级以及多款GPU、服务器产品的推出，英伟达展现了在AI基础设施领域的绝对主导地位。

AMD背水一战

AMD执行副总裁Forrest Norrod此前曾坦承，英伟达在GPU运算加速卡方面建构了丰富的软件生态系统，几乎覆盖了多数市场需求，其护城河之深，让AMD如今不可能来得及复制英伟达走过的这条路，需另辟蹊径。

《福布斯》杂志评论称：“如果业界还有英伟达潜在的对手，那一定包括苏姿丰和她掌管的AMD。”

英伟达发布GH200被看做是对于近期动作频频的AMD的“反击”。在近日举行的AMD第二季度业绩说明会上，苏姿丰表示，到2027年，数据中心的人工智能加速器市场可能会超过1500亿美元。个人电脑是推动半导体处理器销量的传统产品，但随着个人电脑销量下滑，人工智能芯片成为半导体行业的新亮点之一。

与此同时，苏姿丰指出，计划在2023年第四季度扩产MI300系列芯片，包括CPU-GPU混合型MI300A和纯GPU型MI300X。相关样品已经送达客户进行测试，预计在2024年批量销售。

MI300X是一款专门面向生成式AI的加速芯片，拥有1530亿个晶体管。其HBM（高带宽存储器）容量及显存带宽，分别是英伟达H100的2.4倍及1.6倍，由于HBM容量大幅提升，单颗MI300X芯片可以运行800亿参数模型。

实际上，近几年，与英伟达类似，AMD的战略重心也在向AI转移。苏姿丰也不止一次地表明自己对于AI的态度。在今年早些时候举办的CES 2023科技大会上，苏姿丰发表了“AI is the defining megatrend in technology（AI是未来科技的决定性趋势）”的主题演讲。她说道：“AI已是AMD当前的第一战略重点，我们正积极与所有客户合作，将联合解决方案推向市场。”

除了硬件端之外，AMD在软件端也欲与英伟达试比高。其在不断加大软件生态的投入，推出了用于数据中心加速、一套完整软件栈工具AMD ROCm系统，包括为PyTorch 2.0提供即时“零日”支持，AI模型“开箱即用”等。ROCm软件栈可与模型、库、框架和工具的开放生态系统配合使用，ensorFlow和Caffe深度学习框架也已加入第五代ROCm。

不仅如此，AMD也在尝试“重启”中国市场，AMD正在认真考虑采用类似策略以将MI300和旧版MI250芯片产品出口中国。苏姿丰说道：“当然，我们的计划是完全遵守美国的出口管制。但我们确实相信有机会为我们正在寻找人工智能解决方案的中国客户开发产品，我们将继续朝着这个方向努力。”

即便如此，在半导体分析师王志伟看来，在AI芯片领域，英伟达的地位不可撼动。其成熟的芯片已经得到市场认可，并被广泛使用，而AMD相关产品仍处量产初期，其效果也需要长时间的市场验证。

值得注意的是，AMD执行副总裁Forrest Norrod此前曾坦承，英伟达在GPU运算加速卡方面建构了丰富的软件生态系统，几乎覆盖了多数市场需求，其护城河之深，让AMD如今不可能来得及复制英伟达走过的这条路，需另辟蹊径。

全球半导体市场缓步回暖 全面复苏或待2024年

本报记者 谭伦 北京报道

在经历2022年的萧条后，全球半导体市场显示出回暖迹象。

日前，半导体产业协会（SIA）发布报告称，2023年第二季度全球半导体销售总额为1245亿美元，比2023年第一季度增长4.7%，但比2022年第二季度下降17.3%。2023年6月全球销售额为415亿美元，同比增长1.7%，这是全球芯片销售额连续第四个月实现小幅上升。

与此同时，报告显示，我国6月半导体销售额也实现了环比3.2%的增长。对此，SIA总裁John Neuffer表示，这为下半年全球半导体市场的继续反弹提供了乐观预期。

此前受半导体产能短缺转为过剩的态势影响，全球半导体市场在连续增长8个季度后，于2022年第二季度首次出现收入下滑，第三季度更是延续颓势，下降了7%。Gartner预测，全球半导体市场的低迷将持续到2023年。

对此，CINNO Research首席分析师周华向《中国经营报》记者分析道，目前，全球半导体市场销量在中国市场销量的带动下有所回升，并且随着全球资本市场预期回升，由此带来环比增长。

华福证券研报则指出，在2021年12月全球销售额增速达到峰值后，全球半导体市场开始进入下行周期，此轮景气度下沉已持续较长时间，半导体行业基本面“筑底”已基本完成，本次连续数月的稳定环比增长或将为半导体行业触底回升带来一缕曙光。

有复苏但低于预期

在业内看来，此轮复苏的主要动力来自消费电子产业的缓步回升，以及AI芯片的剧增需求。华福证券电子行业分析师杨钟表示，从需求侧来看，一方面，2023年第二季度全球PC出货量创2022年第一季度之后首次环比增长，2023年5月中国手机出货量实现同比增长25.2%，消费电子需求出现回暖迹象。另一方面，杨钟指出，高性能运算、AIoT、汽车智能化、XR等新兴领域成为半导体行业长期发展新的源头活水。McKinsey预测数据显示，2030年，全球半导体销售规模有望达到1万亿美元，其中汽车领域2021~2030年CAGR达到13%~15%，是增长最快的领域。

此外，据TrendForce预测，2023年AI服务器出货量将接近120万台，年增38.4%，AI相关应用拉动的计算、存储、数据互联芯片需求有望赋能半导体行业发展。

而世界半导体贸易统计组织（WSTS）今年6月发布的研报则更明确地指出，此轮增长主要由中国市场带动。该组织援引区域数据显示，在今年4月欧美环比下跌的情况下，中国市场的芯片销售额环比增长2.9%。

与此同时，半导体相关产业的热度也侧证了中国芯片市场的回温。国际半导体产业协会（SEMI）在其6月发布的《半导体材料市场报告》显示，2022年全球半导体材料营收约727亿美元，同比增长8.9%，创历史新高。其中中国内地半导体材料市场规模达129.7亿美元，在半导体材料市场排名中位居第二，同比增长7.3%。

不过，虽然小幅增长释放了利好信号，但在同比数据上，全球半导体的销售额仍旧大幅落后于2022年。SIA数据显示，与2022年第二季度相比，全球半导体市场销售额下降了17.3%，其中，中国市场的销售额下降了24.4%。

周华援引CINNO Research发布的调研数据向记者表示，2023年第一季度国内代表性IC设计厂商库存周转天数约为310天，较去年同期增加约150天，从去库存的角度观察，我国半导体复苏不及预期。

中金公司研报也指出，今年上半年，A股半导体板块呈现“先扬后抑”态势，而4月以来半导体板块的下跌已经充分计入，需求复苏弱于年初预期。受到消费类重点需求市场回暖暂不明显影响，投资人预期，半导体行业或将进入“温和复苏”阶段。

整体形势仍旧低迷

经历了过去一年的需求侧冲击后，复苏中的全球半导体市场仍旧于近日显示出稍许疲态。其中，芯片巨头们的中期业绩成为此轮半导体产业态势的缩影与注脚。

全球市场方面，三大芯片巨头英特尔、AMD、高通近日相继发布了2023年第二季度财报。其中，英特尔该季总营收为129亿美元，相比去年同期下降15%；AMD该季营业额则为53.59亿美元，同比下降18%，净利润更是仅为2700万美元，较去年同期的4.47亿美元下降了94%；同样命运的还有高通，其最新季度营收84.51亿美元，同比下滑22.7%，净利润为18.03亿美元，较去年同期下滑51.7%。

手机和PC两大消费终端出货的走弱被认为是高通与AMD的主要困因所在。高通方面表示，由于对某些主要原始设备制造商的芯片组出货量减少了19亿美元，即受手机消费疲软等负面影响。其管理层在财报会上预计，手机和其他电子产品的零部件支出将持续下降，并延续至今年年底。

英特尔方面虽然同样营收下滑，但在净利方面成功扭转了第一季度的亏损情况。不过，英特尔CEO基辛格仍然表示，目前整个公司在所有细分市场的表现都很疲软，整个行业的复苏时间比预期更长。

制造方面，台积电也未能避免颓势，其6月营收较5月减少194.7亿新台币，环比下滑11.4%，而这也是台积电连续第4个月同比下滑，这意味着台积电今年上半年的营收，较去年同期也将下滑。台积电官方表示，预计今年全年按美元计算的营收将比去年下降10%左右。

而在中国市场，A股半导体企业中报业绩预告也于日前密集出炉，设备市场呈现出整体强劲的姿态。其中，中微公司和北方华创均实现上半年业绩走强，且净利润均实现翻番。中微公司预计营收约25.27亿元，同比增长约28.13%，归母净利润为9.80亿元至10.30亿元，同比增长109.49%至120.18%；北方华创则预计实现营收78.20亿元至89.50亿元，同比增长43.65%至64.41%，归母净利润为16.70亿元至19.30亿元，同比增长121.30%至155.76%。

值得注意的是，在芯片设计、封测、制造以及分立器件等产业链相关企业方面，中报业绩则相对惨淡。在已披露业绩预告的15家芯片设计公司中，13家均预计业绩下滑。对此，上海贝岭公开表示，上半年集成电路行业整体尚处于低位，市场需求持续疲软。大为股份也表示，业绩下滑系受到了半导体存储行业及智能终端行业需求疲软及细分市场竞争加剧等因素影响。

对此，半导体分析师季维向记者表示，业绩数据的整体低迷更多是与去年同期比较，除反映全球半导体整体走势提振不及预期外，与产业已经进入复苏期并不矛盾。

2024年有望全面复苏

缓步复苏的信号释放后，市场更加关心的是，全面的复苏何时到来？对此，产业上下游及分析人士都将时间节点放在了2024年。

“目前国内内需不足，仍不能对全面复苏起到有力支撑，但同时，存储、AI类芯片市场已经出现拐点，预计2024年，全球半导体产业或将出现‘由点带面’形式的复苏。”周华向记者表示。

记者注意到，日前国内某存储头部企业已率先传出涨价通知，将针对企业级客户调升NAND价格3%~5%。之后，三星、SK海力士跟进涨价。而在国内，存储龙头兆易创新近日发布的半年报预告也显示，公司今年第二季度单季实现净利润约1.9亿元，环比增长约26.55%。因此，业内预计，存储板块或将在2023年下半年迎来拐点。

同时，受大模型训练的拉动，AI芯片的需求则更为突出。在7月下旬举行的2023年世界半导体大会期间，华为公司董事、首席供应官应为民便透露，国内AI芯片需求与年初相比，在半年时间里增长了10倍以上。中金公司研报也做出预计，在半年维度上，看好半导体行业周期复苏及AIGC新应用催生的新需求。

此外，工银瑞信基金观点指出，得益于近年来官方对半导体制造业的支持，我国晶圆制造能力持续提升，中国内地半导体材料市场规模持续快速增长，2020~2022年电子材料企业收入规模将实现翻倍。

而在全球方面，标准普尔在其最新供应链报告中，引述苹果公司等8家全球领先企业的前瞻性评论指出，到今年第三季度或年底，半导体产业的情况或出现好转。

SIA则对这一预期的时间节点更为保守。其预计，全球半导体市场在2024将增长11.8%，达到5760亿美元，而这一扩张将主要由内存产业推动，届时其全球收入规模将恢复到1200亿美元，同比年增长率有望超过40%，而其他包括分立、传感器、模拟、逻辑和MCU领域，预计都将呈现个位数增长。