AI“大行其道” 英伟达“坐享其成”

本报记者 秦枭 北京报道

ChatGPT在全球的关注度持续火热，再次将AI产业推到聚光灯下，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。随着AI产业迎来“iPhone时刻”，算力需求正在持续释放，以AI服务器核心零部件GPU（图像处理器、加速芯片）为代表的供给端走俏，其价格也在不断上涨，而在AI芯片GPU市场占据绝对优势的英伟达也赚得盆满钵满。

多位业内人士在接受《中国经营报》记者采访时表示，大型模型通常需要庞大的算力和存储资源来进行训练，GPU已成为AI加速芯片通用性解决方案，越来越多的企业和个人开始使用GPU来训练大型深度学习模型。这种需求的增加可能导致GPU的价格上涨，从而导致显卡价格的上涨。此外，由于供应链问题、半导体短缺等因素的影响，显卡价格的波动也可能受到一定程度的干扰。

英伟达大秀肌肉

目前主流AI厂商都进入了“千亿参数时代”，多采用了英伟达的GPU。

AI场景需要多核、高并发、高带宽AI芯片。AI芯片，也被称为AI加速器或计算卡，即专门用于处理人工智能应用中的大量计算任务的模块。当前，AI芯片主要分为GPU、FPGA，及以TPU、VPU为代表的ASIC芯片，而GPU凭借其高性能、高灵活度特点成为AI加速方案首选。据IDC数据，预计到2025年，GPU仍将占据AI芯片80%市场份额。

资料显示，2018年OpenAI开发的GPT-1的预训练大模型参数为1.1亿，2019年发布的GPT-2提高至15亿，2020年GPT-3的预训练大模型参数已经提高至1750亿。而为了训练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。

不仅如此，目前主流AI厂商都进入了“千亿参数时代”，多采用了英伟达的GPU。以科大讯飞星火认知大模型为例，其使用了英伟达的T4 Tensor Core GPU进行训练。除了T4 Tensor Core GPU，科大讯飞还使用了其他类型的GPU和其他硬件设备来支持其深度学习平台的开发和应用。这些硬件设备包括英伟达的PaddlePaddle、NVIDIA Tesla V100、AMD EPYC等，以及多种CPU、内存、网络设备等。

昆仑万维集团CEO方汉也表示：“超过千亿级别的大模型，它的训练大概需要1000—2000张A100的卡，没有2000张A100的卡，实验都做不了。”

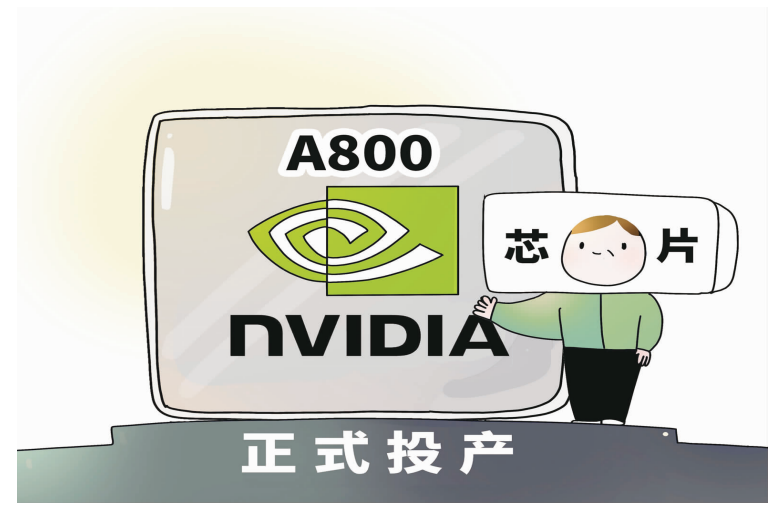
招商证券指出，从通用服务器到AI服务器，一个最显著的变化就是GPU取代了CPU成为整机最核心的运算单元以及价值量占比最大的部分，传统服务器通常至多配备4个CPU+相应内存和硬盘，在AI服务器中，通常是2颗CPU+8颗GPU，部分高配4U服务器中甚至可以搭配16颗GPU，预计AI服务器中GPU+CPU+存储的价值量占比有望达到80%以上的水平。

据统计，英伟达当前在售的用于大模型训练的GPU卡至少有9款型号，其中高性能的有4款，分别是V100、A800、A100及H100。而此轮AI“军备竞赛”也让用于上述显卡的价格一路高涨。其中，A100此前售价在1.5万美元（约合人民币10.35万元），但目前在一些平台上，此款显卡价格上涨到15万元左右。

英伟达也借机赚足了“钱包”。TrendForce数据显示，如果以英伟达A100显卡的处理能力计算，GPT-3.5大模型需要2万块GPU来处理训练数据。目前英伟达A100显卡的售价在10000~15000美元之间，预估英伟达可以赚3亿美元（约20多亿元人民币）。

值得注意的是，英伟达还在源源不断地为这场军备竞赛输送弹药。在此前GTC开发者大会上，英伟达推出了新的Hoppper CPU——配有双GPU NVLink的H100 NVL，这是专门针对算力需求巨大的ChatGPT而设计的显卡，拥有夸张的188GB HBM3内存（每张卡94GB）。

不过，即使价格上涨，目前市面上几乎“一卡难求”。一位业内人士对记者表示，客户对英伟达A100/H100芯片需求强劲，后者订单能见度已至2024年，更紧急向代工厂台积电追单。



英伟达也借这场“AI军备竞赛”之机赚足了“钱包”。 视觉中国/图

国产厂商的机遇

目前中国GPU开发者大多使用国外厂家提供的IP，自主性不高，不过经过多年沉淀是能够实现国产替代的。

虽然国内外的大模型项目接连落地，但除了百度、阿里巴巴等企业采用自研芯片外，国内大多数企业仍难求高端GPU。据透露，国内可用于训练AI大模型的A100大约有4万—5万个。

英伟达在去年收到美国政府的通知，通知称：“若对中国（含中国香港）和俄罗斯的客户出口两款高端GPU芯片——A100和H100，需要新的出口许可。”不仅如此，该许可证要求还包括未来所有的英伟达高端集成电路，只要其峰值性能和芯片间I/O性能均大于或等于A100的阈值，以及包括这些高端电路的任何系统，都被纳入许可证限制范围。

不过，英伟达针对中国客户推出了替代型号A800，与原有的A100系列计算卡相比，A800系列的规格基本相同，比较大的区别在于NVLink互连总线的连接速率，前者为600GB/s，后者限制在了400GB/s。综合使用效率只有A100的70%左右。前不久英伟达还发布了特供版的H800，作为其旗舰芯片H100的替代版。

4月14日，腾讯云正式发布新一代HCC（High-Performance Computing Cluster）高性能计算集群。据悉，该集群采用腾讯云星星海自研服务器，搭载英伟达最新代次H800 GPU，H800基于Hopper架构，对跑深度推荐系统、大型AI语言模型、基因组学、复杂数字孪生等任务的效率提升非常明显。与A800相比，H800的性能提升了3倍，在显存带宽上也有明显的提高，达到3TB/s。

伴随着近期宏观经济回暖以及国内互联网企业纷纷加大AI算力布局，PC和服务器的需求上升有望为国内GPU市场带来整体拉动效应。

目前，国内已涌现出一批优秀的GPU设计和制造厂商。

其中，海光信息目前已经成功掌握高端协处理器微结构设计等核心技术，并以此为基础推出了性能优异的DCU产品。其深算一号产品和英伟达A100及AMD高端GPU产品（MI100）进行对比，单芯片产品基本能达到其70%的性能水平。

值得注意的是，上述业内人士表示，虽然国内的GPU厂商取得了一些成绩，但是由于GPU研发难度大、开发周期长，例如A100，英伟达只用了三个月的时间便研发出替代方案，而国内却并不多见。而且目前中国GPU开发者大多使用国外厂家提供的IP，自主性不高，不过经过多年沉淀是能够实现国产替代的。

三星大变局：以AI重构商业版图

本报记者 吴清 北京报道

在全球消费电子和半导体产业持续低迷的背景下，老牌巨头三星全线出击，正在酝酿一场全球业务的大变局。

5月15日，据日媒报道，韩国三星电子计划将投资300亿日元（约合人民币15.4亿元）在日本横滨建设半导体后段封装测试产线，预计将在2025年完工量产。而就在前一日，据韩媒报道，三星电子与Naver计划联合开发一款用于企业的生成式AI（人工智能）工具，对标ChatGPT。三星和Naver还计划在今年下半年推出AI芯片。

而更值得关注的是，三星掌舵者、三星电子会长李在镕刚刚结束为期22天的访美行程。在此期间，他分别会见了特斯拉CEO马斯克、英伟达CEO黄仁勋、微软CEO纳德拉、谷歌CEO皮查伊等20多位科技公司的高管，遍及互联网、AI、通信、生物等行业。

《中国经营报》记者注意到，尽管此次访美行程紧凑，但李在镕花了很多时间与当地AI、半导体产业的“领袖”互动交流，顺应近期AIGC（AI生成内容）的热潮，AI和半导体芯片成为李在镕此行的重点。一位三星内部人士告诉记者：“在全球信息与通信技术不景气的情况下，三星需要培育增长型业务作为公司的新支柱。”

今年4月末，三星电子公布的业绩报告显示，公司第一季度营业利润为6402亿韩元（约合人民币33.2亿元），同比下降95.5%，为14年来的最低季度营业利润。上述内部人士补充道：“在这个关键时刻，李在镕会长决定亲自激活全球网络，探索新的业务战略，并取得了突破。”

而就在此前，李在镕还和苹果CEO库克等一道到访中国并低调出席了一系列活动，去年以来三星手机也通过线下开店、线上广告等形式“重返”中国市场。

布局半导体封测 应对台积电竞争

据前述日媒报道，此次三星电子投巨资在日本横滨建立半导体后段封装测试产线，主要是因为当地有三星的研发机构，这样可以相互配合发展，同时将会获得日本政府“芯片法案”的配套补贴。

从日本“芯片法案”条款来看，三星此次投资将可获得约100亿日元（约合人民币5.12亿元）的税收减免和资金补助，这是日本政府为在当地投资半导体芯片制造领域企业提供的激励政策。

目前尚不清楚该项目的具体投资计划，不过根据已公布的信息，该项目接下来会招聘数百个工作岗位以进行产线上的工作。

过去多年来，三星一直专注在半导体前端的先进制程技术发展。不过在当前摩尔定律“失效”，先进制程推进越发困难的背景下，三星近年来也开始积极发展先进封装技术，以更先进的封装技术来弥补制程技术提升的放缓。

有意思的是，三星在半导体领域的长期竞争对手台积电早在2021年宣布在日本熊本与SONY、DENSO合作建立半导体晶圆厂。因此有业内人士认为，三星在日本的投资，其实也是针对竞争对手台积电而来。

不过，三星此举可能还有更为现实的考量。“日本一直是半导体设备及材料强国，前两年因为半导体材料被‘卡脖子’曾让三星陷入被动，此次，三星希望借投资深化与上游日本材料和设备供应商的合作，寻求半导体制造更大的突破。”一位半导体行业人士对记者表示。

22天旋风访美强化AI、芯片领袖合作

而就在此前，李在镕刚刚结束为期22天的访美之行，在这并不长的时间内，李在镕“旋风”般地与20多位跨国企业的高管进行了会面。

据悉，李在镕于今年4月底前往美国，5月12日返回韩国，在此期间，李在镕从美国东海岸一路飞到西海岸，是其自2014年以来最久的访美行程。在此行中，李在镕与谷歌、微软、英伟达、特斯拉、辉达、百健、欧嘉隆等公司高层会晤，而这些公司都是全球AI、车用芯片、通讯芯片及生物领域的领导者。

比如，5月10日李在镕和黄仁勋在硅谷一家餐厅一对一会面，讨论了未来在AI芯片技术和晶圆代工合作的计划；辉达则是全球首屈一指的生成式AI绘图处理器（GPU）设计业者，这场会面提高了辉达下单三星高频宽半导体芯片（HBM）的市场预期。

备受各方关注的是，李在镕在硅谷的三星电子半导体研究中心与马斯克的会晤，这也是两人第一次私人会谈。消息人士说，三星和特斯拉就共同开发自动驾驶汽车芯片、新一代信息通信技术进行了交流，这场会面将为双方扩大车用芯片合作铺路。

车用芯片被认为是半导体行业未来发展的重中之重，前景广阔。市场研究机构Strategic Analytics和Research & Markets的数据显示，全球汽车芯片市场预计将在明年升至4000亿美元（约合人民币2.79万亿元），到2028年，这个数字将进一步增至7000亿美元（约合人民币4.87万亿元）。

今年4月，三星刚获得了知名自动驾驶科技公司Mobileye的高性能芯片订单。业内认为，三星将基于特斯拉全新一代自动驾驶芯片生产经验，进一步增强其在汽车芯片市场的影响力。

作为全球最大的DRAM和NAND Flash生产商之一，也是全球最大的半导体代工厂商之一，半导体业务一直是三星的营收和利润的主力军。不过受累于低迷的行业形势，今年一季度该业务却亏损4.58万亿韩元（约合人民币236.79亿元）。三星方面还预计，由于需求大幅下滑，今年全球芯片市场规模将萎缩6%至5630亿美元（约合人民币3.92万亿元），困难状况可能持续全年。“在此背景下，三星急需稳固原有合作关系的同时，激活更多市场需求。”上述半导体行业人士表示。

此次行程中，李在镕也和纳德拉、皮查伊等企业领袖举行一连串会议，并积极与全球知名的AI学者建立人脉，讨论AI的多元应用和未来的合作，显示出三星想重新连接受新冠疫情影响的国际合作关系，为AI技术未来成长引擎的蓝图奠定基础。三星目前已视AI为未来公司成长的关键引擎之一。

抢滩AIGC 全线出击

作为近年来AI领域最重大的突破，近期全球掀起的AIGC和ChatGPT浪潮，三星也未错过。

5月14日，也就是李在镕结束访美回国后的第三天，据韩媒报道，三星电子与韩国最大在线和搜索引擎运营商Naver计划联合开发一款用于企业的生成式AI工具，对标ChatGPT。

两家公司计划最早于今年10月发布该工具，并将首先用于三星的设备解决方案（DS）部门，该部门包括半导体业务。三星还将在测试后，将其应用范围扩大到三星的其他业务，包括负责智能手机和家电业务的设备体验（DX）部门。

“安全”或是三星下场自研AIGC工具的重要考量因素之一。上述韩媒报道指出，通过与Naver合作，三星可在提高公司生产力的同时规避使用其他平台带来的商业机密泄露风险。此前出于安全考虑，三星电子已禁止员工使用ChatGPT、谷歌Bard和微软必应等流行的生成式AI工具。

同时，三星和Naver还计划在今年下半年推出AI芯片。两公司已为AI芯片业务筹备数月。自三星和Naver于2022年12月正式宣布合作开发AI芯片，短短5个月时间内，已完成了相关AI芯片的技术验证，计划最快8月做成可编程芯片并正式测试。

在强化上游AI芯片等领域合作的同时，三星在下游传统消费电子领域的动作也没落下。从去年开始，中国的消费者在电视和线下门店中看到了更多三星手机的身影。

市场分析机构Canalys公布的2023年第一季度全球智能手机厂商排名显示，三星以6030万部出货量重回市场第一，苹果公司以5800万部的出货量位居第二；小米、OPPO和vivo分别以3050万部、2660万部和2090万部位列第三、四、五名。2023年第一季度全球智能手机出货量同比下降13%，降至2.7亿部。

正是在全球手机市场下滑背景下，以及面对来势汹汹的苹果及中国手机品牌，“重返”中国这个全球最大消费电子市场成为三星虽然艰难但必须完成的动作。三星手机曾一度占据中国约30%的手机市场份额，但此前受各种因素影响已降至1%左右，与其全球第一的地位形成鲜明对比。

据悉，2021年12月，三星电子副会长兼CEO韩钟熙挂帅的“中国业务创新团队”开始实施线下流通渠道扩张战略。此前，一位三星电子内部人士也向记者证实，折叠屏手机是三星布局中国市场的重点。

此外，三星近期组织业务调整频频。分析人士认为，李在镕可能调整三星电子的组织架构，使集团文化更为灵活，以适应最新的产业潮流所带来的快速变化。

“三星拥有雄厚的资金、技术、人才积淀，全产业链布局优势以及良好的上下游合作伙伴关系，业绩压力之下，三星多路出击、全线布局。”上述半导体行业人士对记者表示，不能低估一个少壮派领导重振三星的意志和三星作为老牌跨国巨头复苏的潜力。