ChatGPT带动算力需求 AI服务器需求高涨

本报记者 陈佳岚 广州报道

近日，受益于ChatGPT和AI大模型的火热，AI服务器价格及服务器概念、算力概念受关注度持续升温。

5月17日，算力概念快速反弹，浪潮信息（000977. SZ）大涨超8%。浪潮信息作为英伟达（NVIDIA）在国内最大的分销商之一，目前所生产的AI服务器几乎均牵手英伟达。而作为英伟达算力核心股的鸿博股份（002229.SZ）也涨超4%。与此同时，青云科技（688316.SH）、中科曙光（603019.SH）、寒武纪（688256.SH）、工业富联（601138.SH）等也随之上涨。有报道指出，近期AI服务器价格也大涨，有部分型号的AI服务器不足一年价格涨了近20倍。

不过，《中国经营报》记者注意到，尽管受AI大模型发展热潮影响，市场算力需求大增，来自GPU（图形处理器）、AI服务器等产品的市场需求也在增大，但全年整体的服务器需求或仍将乏力。TrendForce集邦咨询方面预估，2023年全球AI服务器出货量同比增长将逾10%，但全球服务器整机出货量将再下修至1383.5万台，同比减少2.85%。

TrendForce集邦咨询表示，今年服务器市场是否能翻转需观察库存去化速度。依据目前进度来看，短则今年下半年，长则至2024上半年。

GPU难求 AI服务器被传涨价

随着生成式AI的持续火爆，导致AI算力需求骤增，也使得GPU计算卡及高性能AI服务器需求大量增加。

ChatGPT火热出圈后，全球各大科技企业都在积极拥抱AIGC（生成式AI），纷纷发力AI大模型。AI大模型的实现，需要海量数据和强大算力来支撑训练和推理过程，华为预估2030年相比2020年AI爆发带来的算力需求将增长500倍。

服务器，也称伺服器，本质是计算机，核心硬件包括以CPU（中央处理器）、GPU为代表的加速卡、内存、硬盘、网卡、电源、主板等。AI服务器作为算力基础设施单元服务器的一种类型，由于普遍采用CPU、GPU等组合的异构式架构，相较通用服务器具备图形渲染和海量数据的并行运算等优势，能够快速准确地处理大量数据，可以满足大模型所需的强大算力需求，广泛应用于深度学习、高性能计算、搜索引擎、游戏等行业，其价值也逐渐凸显。

随着生成式AI的持续火爆，导致AI算力需求骤增，也使得GPU计算卡及高性能AI服务器需求大量增加，近期A股算力概念股也备受关注。

浪潮信息在互动易平台多次高调回复投资者关于ChatGPT及AIGC等相关问题，宣称公司目前在ChatGPT和AIGC相关方向上已有布局，产品迭代相关的专项研究也在有序推进。

工业富联方面表示，一季度公司AI服务器已应用于ChatGPT，并正加速提升AI服务器及高效运算（HPC）产品的占比和研发速度。此外，工业富联管理层在2022年财报会上表示，在ChatGPT与AI的应用风潮下，公司受益于市场扩容带来的机会，看好下半年AI服务器的销售业绩。

随着算力的需求显著提高，AI服务器核心零部件GPU芯片持续紧缺、GPU价格不断上涨，近期AI服务器价格也被传大涨。

近日，据《证券时报》报道，“有企业透露，其去年6月购买的AI服务器不足一年价格涨了近20倍。同期，GPU价格也不断上涨，例如A100 GPU市场单价已达15万元，两个月前为10万元，涨幅50%。而A800价格涨幅相对较小，价格在9.5万元左右，上月价格为8.9万元左右。”

或受AI服务器涨价消息影响，5月17日，国内多家算力概念上市公司股价上涨。对于股价波动及AI服务器等价格情况，记者多次致电浪潮信息证券部门，不过电话一直未接通。

不过，记者也留意到，目前咨询机构方面向记者透露的A100（A800）GPU市场销售价格范围仍在10万元左右。受美国出台的对华半导体出口限制政策影响，部分高性能GPU无法对华出口，为此英伟达推出了替代型号专供中国市场，A100的替代型号是A800，A800带宽有所限速。

“高端AI服务器的制造总成本大概是一般服务器的15倍到20倍左右，而影响这些成本的就是这些GPU卡片。”TrendForce集邦咨询分析师刘家豪告诉记者，当前，A100（A800）GPU的市场销售价格范围大概是1万～1.5万美元（约合人民币7万～10万元），而H100（800）GPU的市场销售价格范围大概是在2万～3万美元（约合人民币14万～21万元）之间。

在IDC中国企业研究副总裁周震刚看来，“近来AI服务器价格具体上涨幅度尚不确定，不仅与客户所需服务器配置有关，还与大客户采购规模有关。”

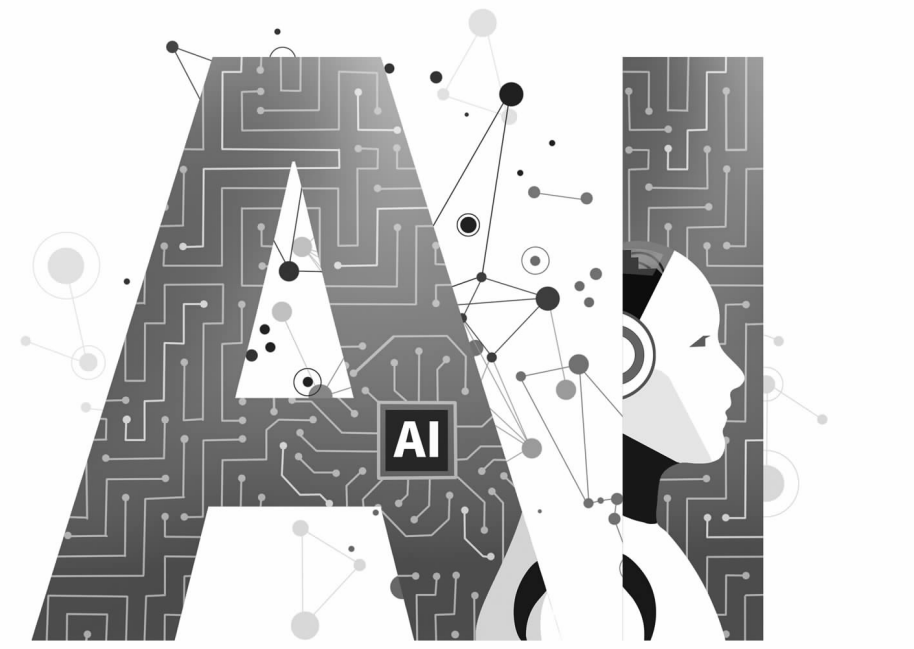
不可否认的是，高端GPU产品供需仍紧张，今年以来，高性能GPU像A100市场价格已经有所上涨。

“当前阶段，高性能的GPU产品市场供应确实紧张，产品价格也在不断上涨。”周震刚对记者补充道，现在最大的问题不是价格因素，而是高端GPU产品货源紧张，由于上游的供货严重不足，就连带宽有所限速的A800 GPU也不好抢。

刘家豪也对记者分析称，AI服务器除了需要搭载GPU等AI加速器外，另外还要考虑不同客户的应用情境，除了考量服务器布局设计外，亦牵涉服务器整体系统适配性，如软硬件布建、调教优化，以及价格考量等，生产设计过程较单就AI芯片考量因素更多。刘家豪也提到，今年AI服务器的出货增长幅度预估会较逊色于GPU等AI芯片。

根据英伟达官网，相较于前代A100，H100综合计算性能提高6倍。海通证券分析师张晓飞在研报中指出，随着GPU向A100、H100升级迭代，AI服务器有望迎来量价齐升。

中信建投则表示，全球GPU市场保持良好成长性，AI服务器成为市场增长的核心支撑。2023年GPU全球市场规模预计为595亿美元，行业保持高速增长，复合增长率为32.9%。



ChatGPT火热，AI服务器需求大增。 视觉中国/图

AI服务器带动增量 全年出货需求或仍不佳

“由于目前AI服务器占全球整体服务器出货比例仍不及一成，故尚无法反转整体服务器疲弱态势。”

抛开AI服务器市场价格上涨的因素，随着AI大模型对算力的需求远超现有供应，高性能GPU需求火爆，也将带动AI服务器行业出货量。

TrendForce集邦咨询表示，今年市场热议的ChatBOT（聊天机器人）确实将带动AI服务器出货量，包含微软、谷歌等云端服务商都积极投入，预估2023年全球AI服务器出货量同比增长将逾10%。

不过，在AI大模型热潮下，被带火的高性能GPU目前给全球服务器市场带来的提振效果仍有限。

“由于目前AI服务器占全球整体服务器出货比例仍不及一成，故尚无法反转整体服务器疲弱态势。”TrendForce集邦咨询预估，由于四大CSP（云服务提供商）陆续下调采购量，Dell（戴尔）及HPE（惠普）等OEM（原始设备制造商）也在2~4月期间下调全年出货量预估，同比分别减少15%及12%，加上国际形势及经济因素影响，服务器需求展望不佳。其预估，今年全球服务器整机出货量将因此再下修至1383.5万台，同比减少2.85%。

而今年上半年整体服务器出货量市场情况也不乐观。

TrendForce集邦咨询数据显示，第一季度受淡季效应与终端库存修正影响，服务器出货量环比减少15.9%；第二季度由于过往产业旺季并未如期发生，环比增长预估仅9.23%。此外，除了OEM调降出货量以及供应链库存调整等持续影响服务器出货量之外，ESG议题使CSP延长服务器使用年限，进而降低采购量，同时顺应企业控制资本支出，OEM提高旧平台的支援年限，也是影响市场情况的原因之一。

浪潮信息披露的今年一季度报显示，当季实现营收94亿元，同比下降45.59%，净利润2.1亿元，同比下降37%。浪潮信息方面表示，营业收入下滑主要系客户需求节奏较上年同期发生变化，本期发货需求减少所致。民生证券指出，主要原因是在全球经济不景气的背景下，美国和中国数据中心运营商的需求减弱（例如：北美云服务商Meta和谷歌搁置新数据中心项目）影响，浪潮服务器发货需求减少所致。不过随着ChatGPT横空出世引爆算力需求，也将催生AI服务器需求高增长。

澜起科技（688008.SH）方面表示，公司第一季度业绩下滑主要受服务器及计算机行业需求下滑及客户去库存影响。

富士康集团董事长刘扬伟也表示，去年富士康仅服务器的营收就达到1.1万亿新台币，其中AI服务器的营收大约占20%，ChatGPT带动的需求虽然强大，富士康会受惠于相关基础建设的需求，但也需要一段时间的沉淀才能成为稳定的业务。

“百模大战”打响 500倍增长的算力何以满足？

本报记者 李玉洋 上海报道

当前，AI（人工智能）大模型已经遍地开花，而AI所需算力的需求也在急剧增长。华为通过有关数据预测，未来10年人工智能算力需求将会增长500倍以上。

GPT（生成式预训练模型）、AIGC（AI生成内容）引发的浪潮，推动着AI 2.0时代的到来。身处这个时代，国内AI算力发展现状如何？能否支撑起“百模大战”甚至“千模大战”？为AI提供算力的芯片，国产AI芯片公司又发展到了哪一步？

对此，国内一家AI芯片企业相关负责人告诉《中国经营报》记者：“AI大模型是目前最火的AI技术应用场景，也是潜在市场最大的方向之一。国内AI算力当前正处于发展方向最明确、发展驱动力最强，同时也是发展挑战最大的阶段。”

而中国移动通信联合会元宇宙产业委执行主任于佳宁也对记者表示：“目前，我国的AI算力发展正在迅速增长，一方面我国政府在政策上积极推动AI算力产业和数据中心的发展，另一方面我国AI算力市场也呈现出多样化和集中化的趋势，使得我国已经成为全球人工智能领域不可忽视的重要力量。”

大模型爆发是远期利好

“无产业不AI，无应用不AI，无芯片不AI。”中国半导体行业协会IC设计分会理事长、清华大学集成电路学院教授魏少军曾这样描述芯片对AI产业的重要性，算力是AI的引擎，AI芯片是AI算力的硬件支撑。

深度科技研究院院长张孝荣告诉记者：“算力主要分三种，按照芯片来划分，以CPU为主的叫基础算力，也叫通用算力；由AI芯片以解决AI计算任务的，叫AI算力；以解决超级计算任务为主的叫超算算力。”

那么，什么样的芯片才是AI算力的最优解？需要指出的是，我们常说的AI芯片并不是一个具体的芯片种类，而是从应用领域来表述，为人工智能提供基础算力的芯片都可以称为AI芯片。

“AI算力芯片包括GPU（图形处理器）、FPGA（现场可编程门阵列）和ASIC（专用集成电路）芯片，其中GPU占垄断地位，市场份额高达八成以上。”张孝荣表示，用于AI计算的GPU叫做GPGPU，相比普通GPU更有计算优势。

于佳宁则指出，英伟达是GPU市场的主导者，其高端GPU占据了AI算法训练市场绝大部分的份额。根据Verified Market Research的数据，2021年全球GPU市场规模335亿元，2028年全球GPU市场规模有望达到4774亿元，英伟达是GPU市场的主导者，全球独立显卡市占率高达80%，其高端GPU如H100、A100和V100等占据了AI算法训练市场绝大部分的份额。

对此，张孝荣也持有类似观点。“AI芯片作为AI算力的支持，其主流产品是英伟达公司GPGPU产品，目前市场上没有挑战者。”张孝荣指出，国内AI芯片发展比较晚、底子薄，目前在低端市场占有一定份额，主要用于解决特定AI任务。

不过，随着AI大模型的爆发，上述AI芯片企业相关负责人指出，这将是国产AI芯片当前最强的发展驱动力，“国内市场对于国产AI芯片的硬需求在逐渐提升，国内应用厂商都在寻找国际巨头以外的第二供货源”。

虽然借助AI大模型火热的东风，国内AI芯片诞生了一个巨大市场，但发展挑战依然存在。“因此，对于国内AI芯片公司而言，AI大模型的爆发是一个远期的利好。”该负责人指出，之所以说是远期，是由于国内的AI芯片产品，特别是在软件生态层面，仍然与国际巨头存在一定差距，客观上需要时间、人才和资源的投入。

谈及未来国内AI芯片的发展趋势，于佳宁认为主要体现在两个方面：“一是AI芯片的规模继续扩大，性能不断提升，支持更为复杂的AI任务；二是AI芯片将进一步向智能家居、智能医疗、智能交通等垂直领域渗透，推动AI技术在各个领域的应用和落地。”

中国在算力领域发展迅速

据了解，人工智能在训练、验证、部署等阶段往往面临应用场景多元化、数据巨量化带来的诸多挑战。于佳宁指出，这就要求算力在支持大规模部署的同时，也要满足高并发、高弹性、高精度等不同计算需求，持续为不同的人工智能负载，高效地提供计算力。

“目前，我国的AI算力发展正在迅速增长，这一趋势符合全球的算力发展趋势。”于佳宁表示，根据IDC数据，2021年智能算力规模为155.2百亿亿次/秒（EFLOPS），2022年智能算力规模达到268百亿亿次/秒，预计2022—2026年中国智能算力规模的年复合增长率将达52.3%，同期通用算力规模复合增速为18.5%。

“一方面在政策上，我国政府在积极推动AI算力产业和数据中心的发展，‘十四五’规划和2035年远景目标纲要中明确提出要‘加快构建全国一体化大数据中心体系，强化算力统筹智能调度，建设若干国家枢纽节点和大数据中心集群，建设E级和10E级超级计算中心’，工信部和国家发改委等先后出台《新型数据中心发展三年行动计划（2021—2023年）》《全国一体化大数据中心协同创新体系算力枢纽实施方案》等重要政策文件，有效规范了我国数据中心产业发展。”于佳宁表示，另一方面我国的AI算力市场也呈现出多样化和集中化的趋势，如腾讯、阿里巴巴等巨头企业正在大力投资、研发并推出了新一代超强算力集群，同时也有很多中小型的AI公司在市场上崭露头角，各自发挥其在特定领域的优势。

于佳宁还指出：“总的来说，中国在算力领域的发展速度非常快，已经成为全球人工智能领域不可忽视的重要力量。未来，中国的算力发展还将持续加速，成为更多应用场景和领域的基础设施，为人工智能的普及和发展提供更加坚实的基础。”

“以大模型为核心的这轮AI浪潮，其算力主要还是由国际巨头厂商所垄断，但国内的AI芯片业取得了从零到一的突破。”上述AI芯片企业相关负责人表示，目前国内的AI芯片厂商都在积极布局AI大模型，并且实现了不少的小规模落地，譬如聚焦大模型开发的智源实验室建立了包含几乎所有国内AI芯片厂商的合作项目。

该AI芯片企业相关负责人还提到，目前国内AI芯片厂商主要还是在做AI相关的软件适配和产品落地工作。然而，在国内掀起的“百模大战”乃至“千模大战”中，企业对于算力的需求爆炸式增长，反映出业界迫切需要“算力供应商”的出现。

所幸的是，国内的算力服务没有拉胯，已在跟进。5月17日，中国电信在业内率先发布算力套餐，为客户提供全系列、标准化、一站购齐、便捷交付的算力服务产品，推动算力赋能千行百业，助力数字经济发展。

据悉，中国电信全新推出“基础算力+算力连接+算法模型+算力安全”一体化算力套餐，为客户提供全栈式算力服务，实现算力触手可及。在基础算力方面，中国电信推出通用算力、智算算力、超算算力三大类型，涵盖训练算力、推理算力等9个系列的产品。在算力连接方面，该公司推出互联网带宽、专线、SD-WAN、云间高速等服务。在算法模型方面，中国电信提供具有百亿参数规模的通用视觉基础模型——星河大模型，同时服务客户大模型训练。在算力安全方面，从平台安全、应用安全、网络安全、数据安全等方面，提供云、网、边、端一体的整体安全防护体系和属地专业安全运营服务。

记者还注意到，上海市服务企业联席会议办公室近日印发的《上海市助力中小微企业稳增长调结构强能力若干措施》中提出，发放“AI算力券”，重点支持租用本市智能算力且用于核心算法创新、模型研发的企业，最高按合同费用的20%进行支持。