不讲故事不卖情怀 国产AI芯片逐渐成熟

本报记者 李玉洋 李正豪 上海报道

金秋九月，推迟两月的2022世界人工智能大会（WAIC）在上海举行。在美国刚刚对华限售英伟达和AMD高端GPU（图形处理器）的背景下，AI/GPU成为本届大会除元宇宙之外的另一大“流量密码”。

芯片是AI的基石。在“WAIC 2022”评选出来的八大“镇馆之宝”中，上海天数智芯的“智铠100”和壁仞科技的通用GPU芯片BR100系列入选其中，成为业界关注的焦点。

《中国经营报》记者注意到，不只寒武纪、壁仞科技、燧原科技、瀚博半导体等这些AI芯片公司展示了最新的芯片产品，百度、华为等科技大佬也展示出了AI相关的硬件。

另外，在全民关“芯”的背景下，记者在一些AI芯片论坛上注意到，国内AI芯片公司已不止于对外宣扬算力理论峰值和未来愿景了，而把重点更多地放在了算力密度、能源效率、生态合作等话题上。

资深产业分析师黄烨锋表示：“从今年的新品和生态更新中，能看到国产GPU/AI芯片企业在走向成熟。发布POD（智算机）、集群，强调系统和软件生态的重要性，并将其落地转化为生产力，无一不体现着现在的国产AI芯片企业已经脱离了过去讲故事、卖情怀、谈愿景的初期阶段，朝着更具切实意义的方向迈进。”

算力不是看理论峰值

国家与企业两者有着异曲同工的目标，都要求芯片能达成更高的算力效率和算力密度。

据了解，壁仞科技通用GPU芯片BR100采用了台积电7nm制程，单芯片峰值算力可达每秒千万亿次浮点运算，打破了全球通用GPU算力纪录；同样采用7nm工艺的天数智芯首款云端推理通用GPU产品——“智铠100”，于今年5月成功点亮，经后续测试修正后即可量产；成立于2018年的瀚博半导体则展示了国产云端7nm GPU芯片SG100，据悉该芯片是集渲染、AI于一体的全功能GPU，而云游戏、云手机、云桌面、云计算等元宇宙关键性应用场景正是其所要发力的重点领域。

值得一提的是，专注云端算力的人工智能公司燧原科技发布了高性能AI加速集群服务器产品云燧智算机（CloudBlazer POD），里面内置了云端AI训练芯片“邃思1.0”和“邃思2.0”，云燧智算机及集群方案的诞生，也让燧原完成了芯片、板卡、服务器、集群算力中心解决方案的覆盖。

燧原科技创始人兼COO张亚林对记者表示，从前些年开始，AI模型参数规模就以每3个月提高一倍的速度在发展，如今模型参数已经发展到了千亿，甚至万亿级规模了。“大规模集群是AI计算的必需品。”他指出，“算力底座不仅是芯片，还有板卡、软件，更重要的是系统一体化。而这块除了美国友商之外，中国国内能实现的还非常罕见。”

张亚林还指出，如何通过集群和系统的方式使AI大模型达成更高的生产力，已经成为一个关键问题。“我国东数西算工程的落地，不仅对能效、算力密度有要求，还在部署、运维、集成等方面提出了非常高的交钥匙一体化需求。”他说，这是云燧智算机和集群诞生的背景。

“在AI技术、AI芯片发展到一定阶段后，有越来越多的芯片企业开始强调有效算力、算力效率、算力密度之类的概念，且从端到云的不同企业都在谈这些事。”黄烨锋注意到，今年WAIC不止一家企业用PUE（Power Usage Effectiveness，数据中心总能耗/IT设备能耗）来衡量能源效率，这是一个更偏系统层面的指标。

而在单个AI芯片层面，瀚博半导体创始人兼CEO钱军则在人工智能大芯片产业落地论坛上指出“评价算力，不能只看它的绝对值”，并提出了“算力密度”的概念，该概念可用来衡量一家芯片企业的实力。

如何理解算力密度？钱军将其分为两个部分：一是芯片单位面积内可达成的算力，比如一平方毫米芯片的算力如何；二是每瓦性能（Perf/W），即每瓦功耗能够提供多大的算力。而算力密度在具体业务中的性能表现可从最大吞吐率、最大吞吐率下的时延和超低时延下的吞吐率这三个指标的对比中得出。

与算力密度相关的还有“算力网络”。“现在，我国数据中心能耗每年都有10%以上的增长，每年的电费有近3%是服务于数据中心的。”中国移动（上海）产业研究院技术部总经理阴启明指出，“算力网络是将不同的算力孤岛做连接，降低算力成本、提高算力可用性，如将东数西算工程与‘双碳’目标匹配。”

“从企业的角度来看，更低的TCO（总拥有成本）才是追求算力密度的实际目的：以更低的成本获得相同的有效算力，并且散热、电费、运维之类的成本也需要足够低。国家与企业两者有着异曲同工的目标，都要求芯片能达成更高的算力效率和算力密度，这应当是这两年的共识了。”黄烨锋说。

国产AI芯片企业走向成熟

“发布POD、集群，强调系统和软件生态的重要性，并将其落地转化为生产力，无一不体现着现在的国产AI芯片企业正朝着更具切实意义的方向迈进。”

AI芯片密集亮相世界人工智能大会。 视觉中国/图

在钱军看来，芯片及其衍生的产品从来不是“单打独斗”的存在。对此，黄烨锋持有类似观点。“当我们到具体业务中去看算力和效率的时候，就不是拼芯片堆料的事情了，还涉及到系统级硬件、软件框架、库、工具链、生态这种难度显著增大的组成部分。”他说。

在这些方面的建设上，英伟达是座高山，其余AI芯片公司目前只能望其项背，国内同行都对英伟达的生态建设水平感到有些无奈。英伟达CEO黄仁勋曾表示，开发者是英伟达的重要财富，目前英伟达全球开发者近300万，在其CUDA（英伟达推出的通用并行计算架构，该架构使GPU能够解决复杂的计算问题）计算架构平台上有超过50万个开发者，其中包含了百度、腾讯、阿里巴巴等大型跨国企业。

复旦大学芯片与系统前沿技术研究院副研究员陈迟晓则用了通俗易懂的话语阐述了生态对开发者的重要性和凝聚作用，他说学生在使用CUDA时碰到bug，网上一搜就能找到不少人也遇到了相同问题和解决方法，庆幸的是国内AI企业也在重视生态方面的建设了。

为破解硬件性能上的“单打独斗”并不能将芯片功力全部发挥出来的问题，瀚博半导体更新和完善了软件平台VastStream，其不仅能加速各类AI应用的部署，例如计算机视觉、视频处理、自然语言处理、搜索与推荐、算子自定义扩展等，还提供了系统管理等三大管理工具，方便客户部署。同时，VastStream的基础软件栈功能也变得更加丰富。

壁仞科技也发布了类似的BIRENSUPA软件全栈，从驱动、硬件抽象层、编程平台、框架，到具体的解决方案和应用。除了壁仞GPU自身架构特性相关的接口外，BIRENSUPA框架层支持PyTorch、TensorFlow和百度飞桨PaddlePaddle，证明了产品兼容性。

要真正释放集群作为生产力工具的价值，燧原科技认为软件是关键要素。燧池智算平台是燧原科技的软件和生态平台，其结合强大的云燧AI算力集群和先进的大规模算力资源调度，能为用户提供人工智能模型生产及应用发布的全流程服务，能够一站式满足复杂的人工智能业务场景对人工智能服务的需求。

“从底层硬件（芯片到板卡，再到服务器与集群），到中间层的燧池软件平台，以及上层的应用，包括各种网络模型，如视觉模型、语音模型、推荐模型、多模态大模型等。今年不少国产AI芯片企业都开始强调自家的‘一体化方案’，而着墨于系统和软件平台，体现的也是芯片的真正落地。”黄烨锋说。

国内这些AI芯片企业虽然一直都在做软件，但在今年更加注重落地的WAIC上，软件、生态等的重要性更加凸显了出来。黄烨锋认为，软件及各种框架、库、中间件的完善程度才是一家AI芯片/GPU企业是否走向成熟的最直观表现。

芯谋咨询研究总监王笑龙也认为，“（AI芯片）设计得再好再花样多，大家都不用，这搞出来有啥意义？所以关键还是要有合适的应用场景，让大家都用起来。”

“发布POD、集群，强调系统和软件生态的重要性，并将其落地转化为生产力，无一不体现着现在的国产AI芯片企业已经脱离了过去讲故事、卖情怀、谈愿景的初期阶段，朝着更具切实意义的方向迈进。或许对于整个行业而言，这些都是AI芯片从初期步入成熟期的开端。”黄烨锋说。

巨头抢滩“大模型” AI界掀起“新基座战争”

本报记者 秦枭 北京报道

近年来，大模型已经成为整个AI（人工智能）产学界追逐的技术“宠儿”，“炼大模型”如火如荼，包括OpenAI、Google、微软、英伟达、百度、华为、阿里巴巴等企业巨头纷纷参与其中，各式各样参数不一、任务导向不同的“大模型”也陆续面市。一时间，“炼大模型”成为了当下AI产业发展的一个主旋律。

方融科技高级工程师、科技部国家科技专家周迪在接受《中国经营报》记者采访时表示，AI大模型历经了前几年的探索期、突破期，部分技术已经逐渐成熟，现在在一定程度上达到推广期了。各大企业纷纷发布AI大模型，就是抢抓这个时间节点，在这方面先取得入场门票。大模型具有效果好、泛化性强、研发流程标准化程度高等特点，正在成为人工智能技术及应用的新基座。

据中国信息通信研究院测算，2021年，算力核心产业规模超过1.5万亿元，关联产业规模超过8万亿元。其中，云计算市场规模超过3000亿元，IDC（互联网数据中心）服务市场规模超过1500亿元，人工智能核心产业规模超过4000亿元。

AI新基座

在过去，绝大部分人工智能企业和研究机构遵循算法、算力和数据三位一体的研究范式，即以一定的算力和数据为基础，使用开源算法框架训练智能模型。而这也导致了当前大部分人工智能处于“手工作坊式”阶段，面对各类行业的下游应用，AI逐渐展现出碎片化、多样化的特点，也出现了模型通用性不高的缺陷。这不仅是AI技术面临的挑战，也限制了AI的产业化进程。

“从各类电商平台的智能推荐到日常生活中的刷脸支付，现在我们生活的方方面面都离不开AI。为了满足这些需求，我们需要为每种特定场景收集大量的数据，再从中设计出专用于特定任务的模型，”周迪对记者说道，“AI大模型希望做到的就是能够基于这个模型整合各类需求，从而适应多种差异化的业务场景，解决AI在赋能千行百业中面临的碎片化、多样化问题。”

AI大模型提供了一种通用化的解决方案，通过“预训练大模型+下游任务微调”的方式，可以有效地从大量标记和未标记的数据中捕获知识，极大地扩展了模型的泛化能力。

华为昇腾计算业务总裁张迪煊对记者表示，过去在一个单一的AI应用场景里面，其实是由很多模型组成的，通过多模型支撑一个场景来完成多个任务。而大模型是AI发展的趋势，也就是通过大模型能解决人工智能很多的问题，因为大模型具备很好的泛化性，可以通过大模型实现多个任务，原来场景需要多个小模型，现在大模型可以服务多个场景，这是生产效率的提升。现在国家相关部门也在牵头制定大模型的沙盘，避免科研机构、企业重复研发，通过各个领域的大模型与行业场景结合，可以更好地加速人工智能技术产业落地。

阿里巴巴资深副总裁、达摩院副院长周靖人则认为：“大模型模仿了人类构建认知的过程，这是当下我们面临的重要机遇。通过融合AI在语言、语音、视觉等不同模态和领域的知识体系，我们期望多模态大模型能成为下一代人工智能算法的基石，让AI从只能使用‘单一感官’到‘五官全开’，且能调用储备丰富知识的大脑来理解世界和进行思考，最终实现接近人类水平的认知智能。”

巨头角力

事实上，从2020年开始，全球各大公司和研究机构就已经开始了大模型的军备竞赛。2020年夏天，OpenAI推出GPT-3，在自然语言处理方面，GPT-3展示出惊人的能力，它能写文章，做翻译，还能生成代码，甚至可以学习一个人的语言模式，并遵循这个模式与人进行谈话。

GPT-3的面市也使得全球范围内AI大模型迎来大爆发，参与企业越来越多，参数级别越来越大，成为新一轮AI竞赛的赛场。2021年谷歌发布了万亿级模型Switch Transformer，微软和英伟达也推出了包含5300亿个参数的自然语言生成模型。

国内的企业也不甘落后，华为、百度、阿里巴巴、浪潮等企业都相继推出了自己的大模型。

今年9月2日，阿里巴巴达摩院发布了最新“通义”大模型系列。周靖人介绍说，为了让大模型更加“融会贯通”，达摩院在国内率先构建了AI统一底座，在业界首次实现模态表示、任务表示、模型结构的统一。

同日，华为也发布了基于昇腾AI的全球首个三模态大模型“紫东太初”。据悉，“紫东太初”是具备跨模态理解与跨模态生成能力的千亿参数创新模型。除此之外，其首次使“以图生音”和“以音生图”成为现实，是从限定领域的弱人工智能迈向通用人工智能路径的一次重要探索。

据华为方面介绍，自2021年以来，国内产业界仅基于昇腾AI就先后推出了鹏程·盘古、鹏程·神农、紫东·太初、武汉·LuoJia、华为云盘古系列等有影响力的大模型，并陆续在互联网、智慧城市、生物医药、金融、农业等行业孵化出多个解决方案，加速推动AI在各行各业的应用落地。

对此，周迪分析认为，AI大模型历经了前几年的探索和突破，一些技术已经逐渐成熟，现在在一定程度上达到推广期了。各大巨头纷纷发布AI大模型，就是抢抓这个时间节点，先取得入场门票。

挑战仍存

当然，AI大模型的发展也并非一蹴而就。大模型在实现全模态和全任务的通用性上仍存在许多技术难点，同时受算力资源限制，其训练与落地应用颇具挑战性。

清华大学计算机系教授唐杰认为，大模型训练面临着诸多的挑战，训练成本高昂，训练1750亿个参数的GPT-3，用到了上万块英伟达V100 GPU（图形处理器），总成本据悉高达1200万美元。人力投入巨大，谷歌PaLM 530B团队，前期准备29人，训练过程11人，整个作者列表68人。训练过程不稳定，易出现训练不收敛现象（训练过程中的损失值无明显下降趋势甚至上升），且调试困难。

周迪则认为，AI大模型的发展主要面临体量、评价、应用三大瓶颈。一是体量庞大，研发部署困难。AI大模型的参数量和计算量要求给开发、调优、部署等工程化环节带来极大压力，需要加强AI大模型轻量化技术研发。二是评价单一，运用效能难以显现。当前AI大模型的评价以学术榜单为主，在行业场景下的应用效果难以客观有效评价，建议完善AI大模型评估指标体系。三是应用受限，产品形态仍在探索。建议鼓励AI大模型应用服务创新。比如有的企业采用分行业分层体系，逐步进行AI大模型的落地。

上接C1

据GfK扫地机器人零售监测报告称，2022年上半年，全球扫地机器人市场规模23亿美元，与去年同期基本持平。而另据中怡康中国零售数据监测，2022年上半年，中国扫地机器人市场规模达57亿元人民币，同比增长16.2%。在全球经济疲软的大环境下，中国扫地机器人市场持续增长，而iRobot却并没有在中国市场抢食到“红利”。奥维云网的最新报告显示，今年上半年，国内扫地机器人线上销售额的前五位均由本土品牌占据，分别是科沃斯、石头、云鲸、小米、追觅，其中科沃斯以39.8%的份额位居首位，而iRobot的份额仅0.5%，位居第10位。

在iRobot产品迭代步伐放缓之际，中国企业则迅速“攻城拔寨”抢占市场。石头科技在2022年半年业绩报告中披露，该公司重点发展美国、欧洲及东南亚市场，建立全球分销网络，目前已经在美国、日本、荷兰、波兰、德国、韩国等地设立了海外公司。今年上半年，石头科技营收29.23亿元，同比增长24.49%；归属于母公司股东的净利润为6.17亿元，同比减少5.4%。值得关注的是，2021年全年，石头科技收入中约58%来自境外。

科沃斯在半年报中则披露，今年上半年，科沃斯品牌海外业务收入同比增长17.2%，“添可”品牌海外业务收入同比增长15.9%，占各自收入比重分别达到27.1%和27.7%。该公司表示，未来将进一步加大对海外市场的投入，推动海外营收规模的持续快速成长。

石头科技相关负责人告诉记者，目前国产品牌的技术替代性和品牌能力都在增强，这对iRobot形成了冲击。目前，扫地机器人行业正处于结构升级换代期，低客单价产品需求在下降，高客单价需求在上升，未来智能性强、质量优异、产品口碑好的中高端产品会逐步进入快速发展时期，石头科技将持续加快全球化的布局。

丁少将指出，iRobot因具有先发优势仍在全球市场占据领先地位，但中国厂商与iRobot的差距是在不断缩小的，甚至开始超越。这得益于国内产业链的优势，生态化和本地化的运营策略，以及在家居领域的行业积淀，因此iRobot的竞争壁垒被逐步打破。