

通用算力相对过剩 智能算力相对短缺 中国算力市场的成长烦恼

●记者 郭成林 郑维汉

4月以来,城地香江、平治信息等多家上市公司披露中标数据中心项目或签署算力服务合同的公告。但此前,也有飞利信、莲花控股、锦鸡股份等多家公司的算力业务未能开展、服务器未能供货等原因相继终止算力服务合同。

这种“接单”与“撤单”并存的现象,折射出我国算力市场正面临某种“成长的烦恼”。

上海证券报记者广泛采访产业链公司获悉,目前部分老旧的算力中心因与市场需求脱节而被闲置,而符合时代需求的智能算力供给仍有缺口。同时,部分买卖双方的专业性与成熟度仍有待提升,卖方对于建设和运营智算中心的认知有限,搭建了计算和服务能力有限的通用算力;部分买方对于自身算力需求的认知也不充分,不具备使用数据中心的能力,导致数据中心的利用率不甚理想。

可以说,这些因素共同导致了当下中国算力市场“通用算力相对过剩,智能算力相对短缺”并存的结构失衡。

展望未来,我国算力市场发展空间仍然巨大,行业正在算力调度、算力硬件与配套软件等三个层面持续优化迭代,加速实现从算力大国到算力强国的跨越。

“一半过剩,一半短缺”

我国数据中心市场正处于“一半过剩,一半短缺”的结构性矛盾中。

从上市公司4月披露的算力中心建设目标信息中可知,招标人主要集中在中国移动、中国电信等央企国企。比如,城地香江多次作为联合体之一成为位于安徽、宁夏等地的数据中心项目中标人或中标候选人,中标金额在5.95亿元至24.33亿元之间。同期,平治信息公告预中标陕西电信的GPU算力建设项目及成为浙江移动数智科技有限公司杭州滨江区智能算力服务三期项目中标候选人,中标金额分别为2.46亿元、2.14亿元。

其背后,是三大运营商大踏步加码算力投资。

公开信息显示,中国移动2025年算力预算为373亿元,占总资本开支的25%,计划将智算规模提升至34EFLOPS(每秒10的18次方次浮点运算),并强调对推理资源的投资“不设上限”。中国联通算力投资2025年计划增长28%,并为人工智能重点基础设施和重大工程专项作了特别预算安排。中国电信则明确2025年算力投资同比增长22%,公司董事长柯瑞文在业绩说明会上表示:“算力方面,我们初步安排是百分之二十几的增长,但不设限。”

同时,亦有上市公司公告签订了算力租赁合同。如宏景科技将向Y公司提供总金额约为2.35亿元的算力组网集成服务;中贝通信将向万界数据提供智算中心的算力服务,金额为4.41亿元。

但是,近期市场上“撤单”现象也日渐增多。

飞利信、莲花控股、锦鸡股份、鸿博股份等多家公司的算力服务终止、服务器未能如期开展、服务器未能按期供货等。而

“接单”表明我国算力市场的需求仍在不断涌现,市场规模持续扩大。IDC(国际数据公司)预测,2025年中国智能算力规模将达到1037.3EFLOPS,并将在2028年达到2781.9EFLOPS。

“撤单”也为行业泼上冷水:算力市场存在结构性失衡的问题。

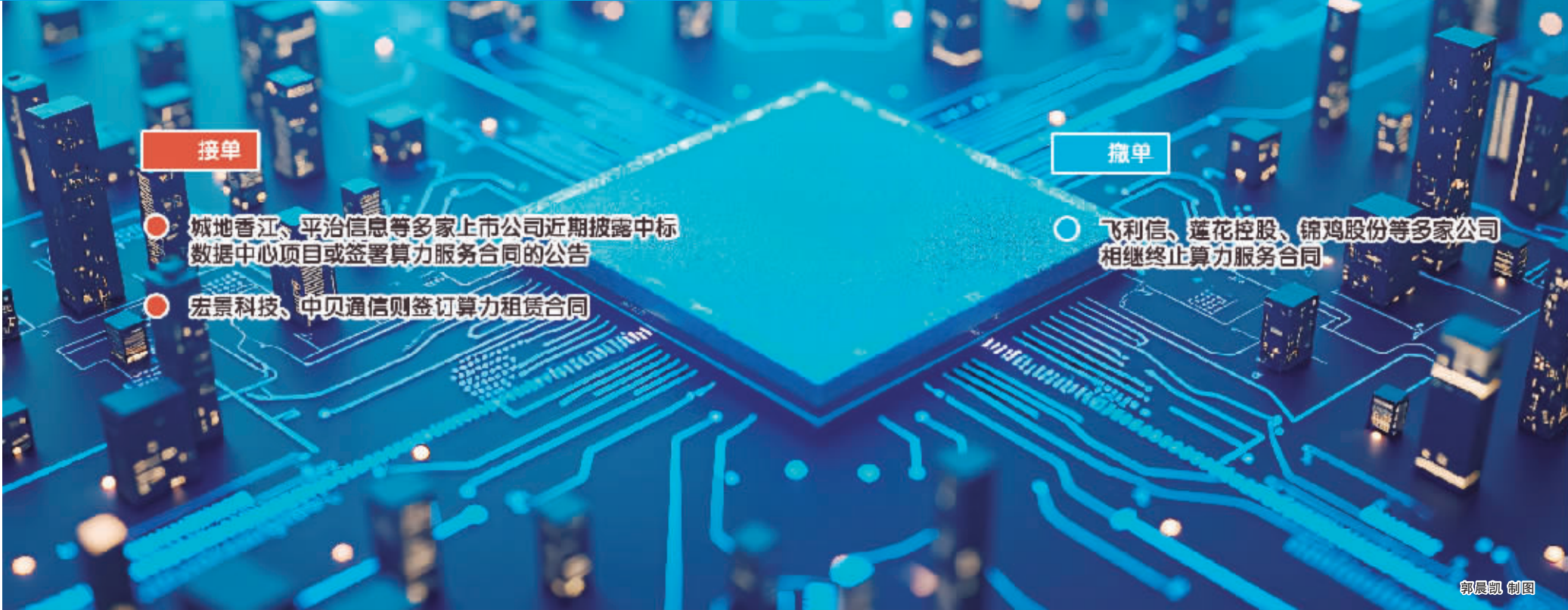
科智咨询此前发布报告显示,我国IDC机柜平均上架率在58%左右,且这一数据在几年内将保持平稳。有业内人士向记者表示,我国部分地区IDC的PUE(电源使用效率)甚至超过2.5,这表明我国市场仍存在闲置算力。

“过剩”与“短缺”并存,呈现出较为明显的按算力质量与地域分布的情况:通用算力相对过剩而智能算力相对短缺,西部算力相对过剩而东部算力相对短缺。

从算力质量上看,计算性能强、配套软件服务好的智能算力相对稀缺。

记者了解到,云服务商优刻得在内蒙古乌兰察布的数据中心一柜难求。“乌兰察布离北京非常近,网络延迟只有5毫秒。”优刻得创始人、董事长兼首席执行官李昕华表示,“而且显卡不是拿过来就能用的,需要做很多的性能优化,以及提供配套的故障处理等服务,优刻得在这方面的技术相对成熟,因此我们的算力比较紧俏。”

“从地域上看,算力需求可以基本按照胡焕庸线划分。胡焕庸线东部人口相对密集,经济相对发达,算力尤其是智能算力基本满载满租,西部地区的供需关系则相对宽松。”快思慢想研究院院长、商汤智能产业研究院原创始院长田丰表示。



结构性失衡的三大主因

为何我国算力市场会出现“通用算力相对过剩,智能算力相对短缺”的结构性失衡?

首先,算力需求高速增长与迭代,底层算力硬件正处于“新老交替”之际,使市场出现供需不匹配。

纵观国内GPU算力集群市场的发展历程,早期的需求主要用于动漫及游戏的视频渲染。其后就进入“百模大战”的时代,这一时期的算力需求主要为大语言模型的训练。再之后,智能体、AI硬件终端市场出现对模型推理任务的新增需求,而云端推理则对数据中心的性价比和使用率提出了更高的要求。

市场对算力需求的快速增长与迭代不仅令部分与市场需求脱节的算力中心被闲置,也造成智能算力的供给缺口。

“一些此前建造的,老款芯片,算力没有那么强的通用AI算力正面临淘汰和转型,而能够进行大规模训练的高端智能算力不足。”有业内人士表示,“曾有测算显示,2025年中国需要300万张GPU,但是我们的产能还不太能满足。所以从中短期来看,国内算力市场确实面临着智能算力的短缺。”

其次,“通用算力过剩而智能算力短缺”与交易双方的市场认知相关。

部分数据中心的卖方对于建设和运营智算中心的认知有限,搭建了一些计算和服务能力有限的通用算力。

“一些非专业人士或企业看到数据中心行业比较火,就想进入这个行业。但是他们对行业基本情况的了解不够深入,缺乏技术和经验积累,导致他们建设的数据中心性能有限。”李昕华表示,“比如数据中心会存在5%至10%的掉卡率,当某张卡掉线之后需要调度另一张卡进行断点续算,其中的计算数据需要做自动化迁移。但部分公司不具备相应的服务能力,这会影响其算力服务的市场竞争力。”

从买方来看,部分公司对于自身算力需求的认知不充分,同时也不具备使用数据中心的能力,导致数据中心的利用率不甚理想。

如某上市公司服务器产品代理商对记者表示,搭建数据中心需要严谨的规划论证和专业的人才支撑。当智算中心交付后,需要客户结合自身的应用环境做API的对连以及之后的相应优化,部分情况下还需要客户对模型进行训练才能够达到预期的使用效果。

“有时候我们向客户交付了产品,但他们不会用。甚至在有些时候,意向客户的规划和预算都很‘理想’,但其后经过公司内部层层审批和论证,发现自身对于数据中心的需求和运维能力都没有那么强,计划就不了了之。”该代理商补充道。

最后,部分企业看重在西部地区建设算力中心的低成本,而忽略了客观条件限制,这也导致算力资源在地域层面的错配。

实际上,有些地区并不适合建设数据中心。“算力中心不能只考虑算力,还要考虑网络。某些地区到上海的网络延时可能高达100毫秒,这样的延时损耗对于算力输出而言影响很大。因此,相关的算力应用只能部署在本地,但当地往往不存在太多的算力需求,致使数据中心的上架率偏低、PUE偏高。”有业内人士认为。

从算力大国到算力强国的跨越

如何解决我国算力市场“成长的烦恼”?

业内人士普遍认为,行业正在算力调度、算力硬件与配套软件等三个层面进行优化迭代。

“在调度层面,随着市场开放、技术成熟,一定会有运营公司把闲置算力整合起来,赋予这些算力应有的市场价值。”云服务商天宽科技相关负责人告诉记者。

目前,针对部分算力中心因配套软件服务能力欠缺而闲置的情况,市场上已出现“算力合伙人”模式。

“优刻得正在运用自身的软件服务能力优势,管理一些算力中心的空闲硬件资源,合伙将闲置算力利用起来。”李昕华表示。

同时,西部地区在土地、电力等资源禀赋优势也使“东数西存”“东数西训”“东数西渲”成为趋势。

具体来看,“东数西存”是指东部的数据中心更多地存储需要频繁访问与在线运算的数据,而将大量的灾备数据、异步分析数据存储至西部数据中心;“东数西训”是指东部的数据中心较多地负责对通信延迟要求较高的实时推理任务,而将高耗能但对延迟要求较低的训练任务部署于西部的数据中心;“东数西渲”则是指将东部数据运送至西部进行游戏、电影视频渲染,将东部地区算力中心的负载转移至西部。

“其实,对通信延迟要求较低的计算任务基本上可以部署至西部地区的数据中心。”田丰表示,针对搭建时期较早的老旧通用算力,可作为本地的推理中心,开放给学校、科研院所及社会大众创新使用,进一步减少算力闲置。

在硬件层面,国产高端算力硬件正在加速发展。

据TrendForce,华为正加速推出昇腾910C芯片。据悉,该芯片的性能接近英伟达H100的80%。此外,华为昇腾920芯片也或将于2025年下半年进入量产,该芯片可能会在中国市场取代英伟达H20。

天宽科技相关负责人称,国产算力的可用性问题已经解决,接下来只需要逐步提升单卡算力、能耗与总体产量。“国内高端算力硬件市场存在供给缺口,谁先推出此类产品,谁就能拿下份额,国产高端算力硬件发展势不可挡!”亦有业内人士向记者表示。

在软件层面,国内硬件与算法间的中间层软件也在不断发展。

“这一层软件非常重要。”田丰表示,“英伟达耗费了十余年才打造出CUDA软件层这条‘护城河’,使得英伟达显卡在流体力学、纳米科研算法等细分领域的算法加速表现优异。当前,国产算力硬件多种多样,视觉算法、语言模型算法、多模态算法、语音算法等新算法也层出不穷,而这些算法均需要软件层才能更好地和算力硬件适配。”

目前,国内市场的模型厂商正在和算力硬件厂商共同开发中间层软件,如商汤已在和华为昇腾、寒武纪合作研发算法加速的基础算力库,补上“国产CUDA”这一层。“模型厂商和芯片厂商共同进行软硬件适配工作是一种趋势。未来,国内软件厂商的算法将在国产的新算力硬件上越跑越快。”田丰称。

“算力过剩”与“算力短缺”并非矛盾,而是同一生态系统不同层面的结构性错位现象。

未来,随着算力调度、算力硬件和配套软件的不断优化与迭代,“沉睡的机房”将变成数字经济的持续动能,我国也将逐步实现从算力大国到算力强国的跨越。

■董事长专访

并行科技陈健:以超算基因撬动算力革命



●记者 胡嘉树

DeepSeek的横空出世,让并行科技意外迎来“高光时刻”。然而,鲜为人知的是,这家从冷门赛道起步的企业,已历经十八载春秋——从最初只有数人的创业团队,到如今成为北交所上市公司,其发展轨迹可谓厚积薄发。

“未来,公司将努力在提供更高质量、更好产品体验的同时降低生产成本。”并行科技董事长陈健在接受上海证券报记者专访时表示。

“并行”由小众走向大众

“DeepSeek的底层技术是并行算法。”陈健介绍,“并行”这一计算机术语不仅是技术的关键,也成为公司名称的灵感来源。

2007年,IT行业迎来了重大技术变革——CPU(中央处理器)由单核迈入多核时代。这一变革使得大量计算机程序需要从串行(按顺序逐项处理)转向并行(同时处理多项任务)。

在此之前,并行计算主要应用于超级计算机等尖端领域。陈健,正是这一领域的深度使用者。作为清华大学工程力学系的毕业生,陈健的研究方向是流体力学。陈健解释道:“超级计算机的用户群体多样,其中计算流体力学研究对算力的需求最为严苛,因此我对高性能计算有切身的体验。”这一对口的履历背景,不仅赋予陈健扎实的超算知识,更让他能从用户视角出发,精准把握需求。

CPU的技术革新使得并行计算从超级计算机的“象牙塔”走向大众市场,陈健敏锐地察觉到这一变革背后蕴藏的划时代机遇,毅然选择创业,成立了并行科技。

“很幸运,公司成立时成功注册了‘并行’商标。”陈健回忆道。多年来,并行科技始终专注于提供科学计算和工程计算服务,客户主要为国家顶级科研机构和头部研发企业。

“确实离大众有点遥远。”陈健表示,“自创立以来,我们专注于完成科技任务、服务用户,认真做好自己的事情。”

记者获悉,目前,并行科技智算云平台已部署DeepSeek模型,支持多种产品形态,能够实现开箱即用。除此之外,并行科技智算云平台还适配Llama 3、CogView2等模型,同时支持自行部署训练和推理模型;昇腾910集群也已在部署中,性价比有望提升200%。

让计算变得简单

在长时期面临盈利挑战的超算行业,并行科技交出了一份亮眼的成绩单。据业绩快报,2024年,公司实现营业收入6.55亿元,同比增长32.07%;实现净利润1152.87万元,实现扭亏为盈。

“作为科技型企业,上市为并行科技带来了资金活水,对公司业绩增长起到很大的助力作用。”陈健说,公司业绩变动也受到国内智能算力市场持续呈现高景气度,公司业务规模扩大、

经营效率稳步提升,AI云头部客户需求强劲以及较高的客户复购率等因素的影响。

记者获悉,并行科技算力资源主要有基于CPU服务器的超算算力资源和基于GPU(图形处理器)服务器的AI云算力资源两种,主要资源提供方为各类算力中心、运营商、算力租赁商等。2024年前三季度,公司CPU算力资源利用率为55%至60%,GPU算力资源利用率为80%至90%。

公开信息显示,并行科技已搭建起65万核CPU、超万卡GPU的算力体系。并行科技还在持续拓展自身算力资源池,2024年前三季度,公司CPU、GPU算力设备采购金额累计超4亿元。今年以来,公司接连发布购买资产公告,拟斥资不超过2.62亿元,向多家公司采购AI算力服务器等。

陈健向记者介绍,当前,我国大模型产业在训练环节对海外算力依赖度较高,而并行科技致力于为大模型在行业的落地应用提供多方位算力服务支持,助力构建完整的国产人工智能技术生态。并行科技也已构建完成“陆海联动”算力网络新格局,通过运营南海海底智算中心项目,依托深海自然冷源和绿色能源优势,打造低碳算力标杆。

在研发投入方面,并行科技也持续加码。公司在分布式集群、算力资源网络、AI算力调优等核心领域掌握领先技术,使得计算和通信成本得以降低,计算资源利用率得到大幅提升。

“并行科技从成立初期就是一家深耕算力服务的公司。”陈健说,“我们把时间精力聚焦在真正能服务好用户的核心关键点上,不做华而不实的、炫技的东西。无论是硬件还是研发上的投入,都是为了给用户提供更好的资源和服务,让计算变得更简单。”

用专注力打造硬科技

“经营企业,要带着团队打一场又一场胜仗。打输了,人心就会散了。”陈健认为,科技型企业需要经历漫长的成长周期,而每个细分领域最终能上市的公司都凤毛麟角。“唯有在行业中将产品做到极致、精益求精,完美满足客户的需求,才能赢得资本市场的认可。”

创业路上,并行科技也有“打输的仗”——早期推出私有云服务时,公司过于专注于技术研发而忽视了服务,这一教训让团队深刻认识到客户需求的重要性。吸取经验教训后,公司调整了战略,于2018年开始提供公有云服务。“当我们真正把精力放在打磨产品、围绕客户需求提升服务上时,反而获得了客户长期认可,节省了大量营销成本。”陈健称。

从抵押房产给员工发工资,到如今带领公司成功上市,陈健将企业家精神凝练总结为“专注、永不放弃”。“选定一个方向,扎扎实实把事情做好,一直干下去。”陈健说。

这种专注力在当下“硬科技”兴起的时代,显得更为有价值。陈健认为:“随着中国科技实力的提升,市场对原创技术的需求日益增长。有专业背景、长期在一个行业里深耕的人将带来更多机遇,踏实做事而非仅靠营销的人和企业将有更大的发展空间。”

展望未来,陈健坚信,人工智能将引发人类社会的“第四次工业革命”。“社会的生产效率将得到颠覆性提升。未来,可能只需少数人维持生产效率极高的人工智能工业生产体系,大多数人将转向服务业或消费领域。”陈健说。

在这一变革浪潮中,并行科技将持续深耕国产大模型及AI产业。“我们的使命很明确:作为企业家,努力提升产品质量,为用户带来更好的使用体验,同时不断优化成本结构。”陈健表示。