锚定AI新型基础设施 算力网攻坚进行时

——全国一体化算力网建设观察

深秋的内蒙古乌兰察布，辽阔大地上一片拔地而起的新建筑群里，施工声此起彼伏，数台塔机来回吊运建材，屋顶施工队伍忙碌的身影，被远处山坡映衬得格外渺小。

千里之外的江苏无锡，一座国产智算中心内，模块化液冷算力机柜有序排列，2000张加速卡几乎满载运行，算力正源源不断输送给长三角的模型和AI应用团队。

一北一南两处现场，勾勒出中国算力产业当下的鲜明图景：算力部署的节奏加快，但供给远未追上需求的步伐。同时，我们感知算力的方式也在发生变化，过去算力更多是模型训练的专属“燃料”，如今算力更多是Token的形态呈现。

这背后是怎样的算力流动？2026年3月，“十五五”规划纲要提出深入推进“东数西算”工程，构建多层次算力设施体系和全国一体化算力网。从“堆卡”到“织网”，据机构测算，“十五五”期间算力网新增直接投资约4万亿元，预计将为智能经济铺就更加坚实的底座，托起千行百业的智能应用。

● 本报记者 杨洁 郑萃颖



中国电信云计算贵州信息园的东数西算（贵州）智算中心 本报记者 杨洁 摄

需求换挡：“碎片化”呼唤“一张网”

远景乌兰察布星河基地东侧，三期项目工地仍在紧张施工，北侧的一期项目已经向客户完成交付。远景智算中心测试负责人朱振国向记者介绍，该基地建成了建筑面积约12万平方米的AI算力超级单体，整个园区建成后可承载百万P算力，支持百万卡运行，规划电力负荷容量最高达3GW。

朱振国表示，在AI算力超级单体内，大规模计算资源集中在同一建筑空间，缩短了设备之间的物理连接距离，降低了跨楼宇通信复杂性，电力、制冷、网络等基础设施也得以集中优化。“同等面积下，算力输出规模最高可达到传统数据中心的10倍。”朱振国说。目前，基地一期客户为头部互联网企业，算力服务需求旺盛，订单已经排到后年；二期至四期预计在2027年至2028年完成交付。

东部的算力供给也同样紧俏。走进无锡（惠山）国产智算中心，2000张摩尔线程MTT S5000几乎满载运行，这个国产GPU集群从今年6月点亮到满载，只用了不到三个月。项目运营方无锡市惠山招商（集团）有限公司副总经理张涛告诉记者：“截至目前，92%以上的集群算力已在使用中，付费用户占比超过60%。整体上算力供不应求，不少客户正在等待集群算力腾出来后进一步做测试。”

供不应求背后，算力的需求结构也正在发生拐点性变化。百度创始人李彦宏近日判断，过去AI算力需求以训练为主，如今已是训练和推理并重，尤其是智能体应用规模化带动推理需求大爆发，推理将成为AI芯片长期增长的主要推动力。百度方面披露，相较去年，百度日活智能体数增长6倍，有效任务完成数增长11倍；云沙箱总量增长315倍，并发量增长60倍。

推理算力需求的本质不同于训练。在硅基流动创始人、CEO袁进辉看来，推理更分散、更潮汐化、更贴近实时业务，天然要求算力“有弹性”“可调度”，不同模型、不同场景的需求差异极大，单一厂商的算力池难以覆盖全部推理负载。当需求从集中式、可预约的训练，转向碎片化、潮汐化的推理，把散落各地、不同架构的算力“织”成一张网，实现供需更好地对接，就成为绕不开的命题。

需求碎片化，算力的供给同样是碎片化的。让算力的供需双方更好地实现匹配，是算力网必须直面的起点和难点。

袁进辉认为，算力网真正的难点，不是把算力中心简单地连起来，而是把碎片化的供给变成可匹配、可交付的服务。他打了一个“盖房子”的比方，当前我国算力投资建设的主体比较多元，房子很多，但标准不一，毛坯房居多，不便于零售。需求端也同样分散，有人要长期稳定的专属算力，有人只要按Token调用，“问题在于能不能把合适的房子在合适的时间交给合适的人，其中既有技术需要解决的问题，也存在机制上可以改善的空间”。

打通调度：从“物理连通”到“逻辑可用”

把卡连起来，是工程；让卡“可调度”，是体系。业内人士告诉记者，算力网要依托计算、通信、协议和调度技术，把云、边、端等分散算力资源整合为统一调度的“全国一张网”，而调度不是“把卡连起来”那么简单，它涉及流量预测、弹性扩缩容、异构芯片混合部署、模型加载速度等系统级工程能力，这是算力网从“物理连通”走向“逻辑可用”的关键一跃。

国家超算互联网平台提供了一个样本。截至2026年9月，平台注册用户突破170万，累计调度作业2.7亿条，日均调度超过30万条，接入全国17个省区市30余个算力中心，汇聚超25万张GPU卡，其中95%为国产算力。

在河南郑州，全国首个全国产十万卡AI超集群曙光8000已作为超算互联网核心节点上线。中科曙光创新方案部经理邹强利认为，算力的调度网络应该是智能的，一方面平台的算力供给多元丰富，另一方面，调度平台要能智能分析用户需求并自动匹配算力，让供需更好地对接。

但跑通一个平台，不等于跑通一张全国性的网。同时，算力也无法孤立“织”网，必须与能源网、通信网节奏对齐。

能源方面，作为“电老虎”，算力中心提高绿电占比、搭配储能保障供电稳定性，已成为业内认为的主流方案。记者了解到，远景乌兰察布星河基地在规划之初就采用能源与算力一体化的思路，其基地100多公里外的商都县，风机转动发出电能，通过风场的升压站升压，再通过高压专线输送至园区，随后由园区内的变电站，从220千伏降到10千伏，送入变电站后方的机房；在其基地的西北侧，远景还在规划建设新的风电场，匹配未来几期算力项目的用电需求。

“对于GW级算力中心，电力已不只是运营成本，更成为决定算力规模和长期竞争力的核心要素。”朱振国对记者表示，远景在建设算力设施的同时统筹布局新能源电站、绿电直连专线、储能系统，前期投入虽多，但既能降低长期用电成本和电价波动风险，也能通过创新供配电架构提高能源利用效率，最终影响每个Token的生产成本。

通信底座的同步升级，决定着算力流动的效率，推进“算网融合”是另一条并行的路线。中国信通院报告显示，截至2026年二季度末，我国8大算力枢纽集群之间的光层互联时延全部在20ms内，与2025年同期相比，约70%集群地市时延波动小于0.1ms；长三角苏州、青浦、嘉兴集群间可以实现1ms光层互连。

下一代光通信的技术也在突破。记者从中国电信了解到，近日，中国电信研究院联合长飞公司、中兴通讯完成全球首个千公里空芯光纤双向实时传输验证，在1007.48公里链路上实现单波800Gb/s单纤双向传输，总容量达128Tb/s，为跨区域算力调度和大模型分布式训练提供了新选择。

资本加码：算力采购开启放量周期

织“网”的步伐，最终落在真金白银的投资上。在远景乌兰察布星河基地东侧，世纪互联乌兰察布云智算数据中心二号基地也正在紧锣密鼓施工，中金数据、华为、快手等也在当地布局了新的算力基地。这座“草原云谷”正加速向“智算之城”升级。据乌兰察布市委副书记、市长刘海泉介绍，目前全市算力运行规模达到22万P，智算占比超90%；已签约算力中心项目109个，总投资超7000亿元，累计完成投资550亿元。

视野转向全国，算力基建节奏也普遍提速。在呼和浩特和林格尔县，中国移动智算中心与火山引擎联合打造的机密词元工厂今年8月落地运行；华为投资750亿元建设的华为云和林格尔数据中心项目，云谷西区（一期）第二批次工程已于今年9月完成竣工验收备案。在宁夏中卫，截至2026年9月底全市标准机架、算力规模累计达50.4万架、29万P，今年新签约项目概算总投资超2000亿元……据国家发展改革委援引机构测算数据，“十五五”时期我国算力网建设将新增直接投资约4万亿元；2026年“六张网”及相关重点领域投资规模超过7万亿元。

企业端，三大运营商率先加码。11月6日，中国移动宣布2027年计划算力投资超1000亿元，中国移动董事长陈忠岳表示，算力是中国移动投资的最大增量，未来两年智算能力将连续翻番。

投资的加速已经体现为采购端的放量。今年10月，中国移动新一代人工智能超节点集采落地，20480卡及配套448台RoCE交换机、82套存储，五家厂商投标报价均在116亿元上下，是运营商在智算基础设施领域单笔规模最大的招标之一。不到一个月，中国电信于11月16日发布智算超节点（N系列）集群设备集采，最高投标限价257.88亿元，预估采购智算服务器5248台、高性能存储110台、RoCE交换机374台，规模进一步抬升。

值得留意的不只是单笔金额，更是采购重心的迁移：运营商正从“盖机房”转向“买超节点、买高速互联”。在运营商的规划里，算力网也并非孤军突进。陈忠岳提出，一体建强通信网、算力网、智能网，通信网贯通流量路由，连接平台与终端，让字节高效传递；算力网贯通模型路由，连接算力与数据，让Token高效流动；智能网贯通任务路由，连接模型与场景，让意图高效执行。

业内将算力网建设概括为资源汇聚、资源可调度、任务可协同三步，最终实现“像用电一样用算力”。值得提醒的是，虽然当前资源汇聚正在加速，调度已在局部跑通，但全国性市场机制、异构算力标准化接口、算电协同价格信号，仍处于试验和磨合阶段。

回看乌兰察布拔地而起的超级单体，和无锡满载运转的国产机柜，它们之间，也许有一张横跨东西、贯通南北的算力大网正在经纬日密，为发展新质生产力、建设数字中国提供坚实底座。



远景乌兰察布星河基地 本报记者 郑萃颖 摄

推理算力是奔涌不息的河流

硅基流动创始人袁进辉：AI基础设施迎来“生逢其时”窗口

● 本报记者 杨洁

“如果说，训练大模型对算力的需求是个泳池，那么做模型推理服务所需要的算力则是条河流，它没有明确‘边界’，奔涌不息。”北京硅基流动科技股份有限公司创始人、CEO袁进辉在接受中国证券报记者专访时这样形容眼前的形势：智能体的规模化应用，正在改变算力需求的增长动力，模型一旦进入日常使用，Token（词元）日复一日消耗，推理算力的需求就停不下来，用于训练的算力反而是阶段性的。在他看来，推理算力的供需曲线至今仍然陡峭，“供不应求的态势大概率还要维持两三年”。

这正是硅基流动所站的位置——做芯片、模型和应用之间的AI基础设施，把分散、异构的算力适配、编排成标准化Token服务，让用户更好地把模型用起来。目前其平台单日峰值Token吞吐已进入万亿量级，注册用户超过1000万，服务了超10000家企业客户。2026年6月，成立仅34个月的硅基流动向港交所递交上市申请，冲刺“AI Token工厂第一股”。

专注AI基础设施

袁进辉师从我国人工智能领域奠基人之一的张钹院士，是AI领域的“老兵”。在清华大学计算机系读博期间，袁进辉主要做算法相关研究，而在微软亚洲研究院的那些年，也正是全球算法研究领域风起云涌、明星辈出的阶段。但他后来的两次创业，却完全转向了AI Infra（人工智能基础设施）系统层方向。这个转向，既基于对自我优势的清醒认知，也潜藏着实现行业“影响力”的追求。

“做算法，灵感和运气非常重要；做系统软件，更重要的是耐心，要想看到最终效果，就得靠日积月累写代码。”袁进辉对自己的判断很直白，他觉得自己在算法上冲到最顶尖把握不大，于是选了攻坚系统层。同时，他认为，两者形成壁垒的方式也有差别，系统层有无穷的工程细节，这种积累在商业上更容易形成护城河。

这是他十年前第一次创业做OneFlow的出发点。袁进辉的判断是，模型会越做越大，走向超大规模，现有框架支撑不住，必须重构底层架构。然而2016年，主流模型都还停留在千万级参数规模，国内很少有人相信大模型有一天会变得巨大。OneFlow创立后最早实现“让分布式多机多卡训练大模型像使用单卡一样简单”，这套理念后来被PyTorch等主流框架跟进。但OneFlow并没有走完商业闭环，2023年先被王慧文的光年之外收购，随后又并入了美团。

一年之中经历两次并购，不少企业向袁进辉抛出橄榄枝，但一丝不甘最终还是把他推回创业之路。袁进辉的遗憾在于：“虽然OneFlow被并购也算对投资人有所交代，但并没有走完独立的商业闭环，相当于我们很早就看见模型进化的趋势，也为此‘磨了刀’，但刀却没有直接砍向果实。”

2023年8月硅基流动成立，第二次出发，袁进辉瞄准了推理赛道，同时也吸取了上一段创业的经验：“第一次创业，是先从idea出发，做出技术产品，再去找市场；第二次创业，我们决定先从市场需求出发，再反推我们的优势。”他算了一道数学题认为，训练的算力用量有阶段性、有边界，但推理的用量没有边界。同时，他看到国内“百模大战”正酣，已经不缺一家大模型公司，缺的是让模型被用起来的中间层，“硅基流动要解决的，是怎么让终端场景和用户更好地把模型用起来的问题”。

复盘两次创业，采访中，袁进辉多次谈到战略判断的时机和节奏。OneFlow的方向没错，但早了行业需求好几年，可能“生不逢时”；这一轮硅基流动选择做推理基础设施，恰好“生逢其时”，需求的潮水恰好在身后涨了起来——模型落地、应用铺开，推理算力的需求跟着爆发。国家数据局数据显示，今年3月，我国日均Token调用量突破140万亿，较2024年初增长超千倍。前一次磨过的“刀”，这一次真正“切”到了市场推进的节奏上。

在芯片、模型、应用之间架桥

硅基流动被全行业看见，是在2025年春节。DeepSeek V3开源后迅速发酵，全网算力告急。硅基流动联手华为，把DeepSeek V3大规模部署在昇腾910B上。“我们春节没放假，初三就把服务上线，大量用户一下子涌了进来，这是全行业第一个在国产芯片上这么大规模地跑大模型的案例。”袁进辉说。

外界看到的是硅基流动一炮打响，但看不见的是在机会来临前的研判和投入。袁进辉回忆，2024年DeepSeek发过V2和V2.5两版模型，当时并未引发太多关注，只是有不少开发者反馈编程方面不错。“我们觉得值得研究，花了好几个月来回沟通把架构搞通。正是当时的适配过程，让硅基流动团队能够在2025年DeepSeek V3发布后快速完成适配。”

袁进辉表示，华为当时还有很多算力闲置，也非常希望把国产卡更好地用起来，因此双方在关键窗口期“一拍即合”。在此之前，国产算力还没有在推理场景大规模跑通。硅基流动率先上线了基于昇腾云CloudMatrix超节点服务的DeepSeek-R1推理服务，一下子把几十万卡给激活了，使用大规模专家并行方案，通过多专家负载均衡和通信优化，实现高吞吐及更高性能。

袁进辉表示，国产算力芯片的主要差距体现在生产制造和软件生态两个层面，“生产制造决定的是物理表现天花板，就像马路修多宽，是四车道还是八车道，决定了能并排开几辆车；而马路上的车流量怎么样、大家是不是开得又快又稳又安全，要借助交通指挥软件的协同指挥。”他坦言，当前国产算力芯片在访存速度上的确存在差距，会影响“吐字速度”。目前，硅基流动和业界合作伙伴联合通过异构混布的方式，尽可能实现使用效率和用户体验的提升，“通过系统软件把不同性能的芯片放到其最合适的位置，用系统问题解决单颗芯片的短板，实现整体效能最大化。”

把单颗芯片用好之后，下一层问题是解决算力碎片化的挑战。在袁进辉看来，算力网，就是AI时代的“电网时刻”。“电网之所以伟大，不仅仅有主网架，还因为有无数的‘变压器和插座’。‘变压器和插座’主要解决三个痛点：其一，‘插座’不统一，电有220伏标准电压，但算力没有，英伟达、昇腾、沐曦、摩尔线程各有各的软件栈和生态；其二，供需错配，最隐蔽的是架构错配，账面算力够了、集群也点亮了，但真跑目标模型时可能才发现跑不动；其三，计量有了、标准还缺，Token已是公认计量单位，但它只计量数量，品质标准和成本口径都还没有。这几个痛点，需要中间层AI Infra来解决。”

中间层的意义还体现于在芯片和快速迭代的模型之间进行“弥合”，让算力利用率最大化。袁进辉表示，芯片迭代的节奏比较慢，从设计到量产通常需要一两年，硬件基本定型，但模型和应用的演化非常快，其间差异全靠中间层软件来搞定，“打个比方，芯片就像是一个盒子，形状是固定的，但模型就像是橡皮泥，没有形状，通过软件就能把橡皮泥捏成适合盒子的形状，放进去，把盒子用满。”

算力供求曲线依然陡峭

袁进辉把训练的算力需求比作泳池，把推理的算力需求比作河流。根据袁进辉的观察，目前显然处于这条河流的“汛期”。他判断明确，今年整体都供不应求，每家产能都不足。产能建设落地需要时间，最近两三年都会是供需关系主导Token市场价格的情况。

这样的态势明确体现在硅基流动的招股书上，其平台规模快速扩容：截至2026年4月，平台注册用户突破1028万，日均Token吞吐量达到5785亿，单日峰值超过万亿级别；公有云付费开发者达到71.6万家，专属实例企业客户从2024年的7家增长至2025年的49家，单客户平均收入提升248%。

“长期来看，伴随模型技术进步，算力产能供给充分，Token价格一定会降下来，让更多人使用，进入规模经济时期，和过去的电、水以及流量资费一样。”袁进辉认为，硅基流动赚取的并非成本的绝对值，而是效率提升的部分，同样的投入产生更多的Token，价值就不随价格波动而转移。

开放独立是硅基流动从成立起的定位。袁进辉强调，不同于云厂商或模型厂的“全家桶”策略，他们不给客户限制某款芯片、某个模型，算力来源也比较多元化，专注给用户提供“开箱即用”的Token。

当更多开发者和中小企业更容易地调用模型，当国产芯片的可用性又往前迈了一步，这些真实而又具体的价值，正是袁进辉创业之初想要实现的影响力。人工智能方兴未艾，算力的河流仍在奔涌，他坦言，深处这个行业“没有人不焦虑”，但看长周期、看趋势，焦虑会少一些。没有人能提前画好终点，但“做长期难而正确的事情”，至少提供了方向和锚点。