算力网建设序幕拉开

万亿投资“织网” 撑起中国智能经济新版图

人工智能产业高速迭代之下，全国一体化算力网建设驶入规模化推进的快车道。作为“十五五”时期国家重点布局的“六张网”之一，算力网与新型电网、新一代通信网等协同布局，成为我国扩大有效投资、筑牢数字经济底座、赋能智能产业升级的核心抓手。

截至今年3月底，我国智能算力规模已达188.2万P（PetaFLOPS），是去年同期的2.5倍；围绕算力枢纽建成干线光缆145条，8大国家算力枢纽间互联以及全国各省份至8大枢纽的网络传输时延全面压至20毫秒以内，算力基础设施建设已取得阶段性亮眼成果。

算力网的加速推进，直面的正是过往算力中心建设过程中“成长的烦恼”——很多算力中心资源利用率偏低，算力“供不应求”与“闲置浪费”并存的结构性矛盾日益突出，制约了算力生产力的释放。与此同时，算力网系统架构复杂、技术壁垒较高，与新型电网、新一代通信网的深度融合仍有提升空间。但行业共识明确，打破算力资源碎片化格局，从“算力孤岛”到“全国一盘棋”势在必行，而这张超级算力网的铺就，有望撬动万亿级社会投资，为智能经济发展注入澎湃动能。

● 本报记者 杨洁 王婧涵

上海世界移动通信大会期间，企业展示的算力相关产品 本报记者 杨洁 摄

重构智能经济底层服务

在无问芯穹联合创始人兼CEO夏立雪看来，算力网的本质，是一张把分散在全国各地的异构算力设施统一连接、统一调度、统一服务的网络。

“算力网调度和流动的不是抽象的‘算力’，而是计算任务、所需要的数据以及计算结果。”夏立雪告诉中国证券报记者，算力网的核心目标是让算力从“稀缺商品”变为“公共资源”，实现“一点接入、全网随用”。

需求侧的爆发式增长，以及当前算力市场突出的“供需错配”结构性问题，让这张网的建设变得刻不容缓。国家数据局披露，到今年3月，中国日均Token调用量已突破140万亿，相比2024年初增长1000多倍。高盛今年5月发布的报告《解码智能体经济》预测，到2030年，全球Token消耗量将较2026年增长24倍。无问芯穹自身统计数据显示，从2025年底到今年5月，Agentic MaaS平台日均Token调用量增速超过20倍，其中95%以上都是智能体场景。

夏立雪介绍，团队在创业时便关注到“算力资源碎片化”的痛点——全国智算中心平均利用率不到60%，一边是算力不够用，另一边是算力被闲置。与此同时，十余种异构芯片在算力中心运行，彼此之间软件栈、编译框架、通信协议并不统一；不同模型、不同任务对算力需求差异极大，传统的“卖卡时”模式无法精准匹配供需。

九章云极董事长方磊也向中国证券报记者坦言：“除非是基础模型厂商，不然绝大部分客户在单一业务场景里GPU的使用率很难超过20%。”

把碎片化的异构算力变成可用、好用、便宜的Token，让用户不需要关心底层用了什么芯片，只需为Token付费，成为无问芯穹创业的初心。方磊也表示，Token正在成为智能经济时代的基础计量单元，算力基础设施的服务方式和评价体系需要发生根本性转变。

来自运营商的专家向记者表示，全国一体化算力网本质上就是希望通过全国一盘棋优化原本孤立的算力设施布局，提升算力资源利用率，让算力如水、电等公共资源一样，成为全社会可靠的、易用的基础性公共服务，推动全社会的智能化转型。

把散落算力“织”成一张网

多位受访行业专家表示，算力基础设施体系投资体量或超万亿规模。天风证券研究称，当前算力网建设仍存在算力资源错配、供给侧余缺并存、中小需求侧应用门槛偏高等痛点，未来的重点建设方向包括建设新一代超算、通算、智算设施体系，积极发展公有云服务，建设算力监测调度平台，制定完善算力资源池化、并网、检测、运营、调度等标准规范。

从政策蓝图到产业落地，算力网的“织就”需要打通技术、机制、标准多重关卡。

从算力到Token的转化效率，考验的是基建能力、能源利用以及成本控制等多种底层硬实力。夏立雪认为，当下最紧迫的任务，是把所有能用的算力资源整合起来，以最高效率生产Token。“核心指标是稳定性、能效、对不同类型资源的兼容性，需要全栈软硬协同设计与端到端优化”，他介绍，在Token经济爆发期，算力规模成为智能上限的核心瓶颈，无问芯穹以“多元异构”的核心技术，在多种模型算法与多种芯片硬件之间，实现跨层次的优化，大幅提升算力的可用规模，破解当前算力短缺困局。

算力网不仅仅涉及算力本身，存力、运力同样关键。中科曙光高级副总裁李斌表示：“未来算力产业的发展趋势，不只是单点硬件突破，而是从芯片、系统、平台到应用的全链路协同创新。”日前，德国ISC2026大会期间发布最新一期IO500全球存储性能榜单显示，中科曙光ParaStor F9000分布式全闪存储系统包揽生产型总榜与10节点挑战榜两项冠军，并刷新世界纪录。

在机制层面，日前召开的国家发改委6月例行新闻发布会明确提出，算力网具体建设过程中，“市场力量将起决定性作用”。“十五五”时期将更加注重供需适配，在“硬投资”方面探索更多行之有效的算电协同模式，做到以电强算、以算促电；加强算网融合创新，适度推动国家枢纽间直连线路扩容，进一步降低网络传输时延。在“软建设”方面，强化算力资源监测与市场化调度。

在标准与生态建设方面，联想万全异构平台相关负责人表示，当前算力节点虽多但彼此缺乏协作，标准化正是破局关键。其介绍，国内包括中国信通院等机构正加快牵头制定相关标准，“这些标准一旦确定，大家都按标准来，就很容易拉通。”

向光而行 产业押注“光互连”

高速光网络是算力网的物理传输底座，也是实现算力跨区域高效协同的核心支撑。国家发改委等部门在《国家数据基础设施建设指引》中明确提出，推动国家枢纽节点和需求地之间400G/800G高带宽全光连接，引导电信运营商等提升“公共传输通道”效能，推进算网深度融合。

方磊介绍，国家算力枢纽节点之间干线光缆建设，让远距离算力协同的成本大幅下降，让跨区域调度算力从技术概念变成了可落地的商业场景，“两年前，济南到北京之间100G带宽的年租赁费用可能超过百万元，如今已降至二三十万元的区间。”

算力节点之间的光缆网络将有望越织越密、越织越宽。记者从鹏城实验室获悉，实验室依托“中国算力网”科研专项任务，联合运营商及光纤光缆头部企业，着力突破超大带宽、低时延、稳定可靠的长距离新型光纤直连关键技术。近日，其长距离光纤传输系统示范工程已通过专家组评审论证，重点攻关算网大容量光直连、异芯混合成缆、低损快速接续等技术难点，为国家算力网络基础设施建设筑牢大容量光传输底座。

正在召开的2026年MWC上海展会上，通算一体、算网融合的趋势更加清晰。华为算力网全光直连方案的展台人头攒动，据介绍，其首创的无损DC-OTN方案，拉远算效达98%以上，有助于实现5-20ms的全国一体化算力网技术愿景。

光互连技术也正在从远距离传输走向机柜内部，中国移动展台展示了DORA（可重构光互连）架构的NPO交换机样机，印证了“以光代电”的产业变革趋势。展台专家表示，面对超高带宽场景下传统电互连的功耗高、距离受限等瓶颈，NPO（近封装光学）、CPO（共封装光学）、OIO（光学输入输出）等光互连技术方案，能够大幅降低传输功耗、打破物理距离束缚，为异构智能超节点集群的高效部署、弹性扩容提供核心支撑。

从硬件基建迭代、技术壁垒突破，到机制标准完善、产业生态聚合，专家表示，全国一体化算力网有望牵引算力基础设施体系全方位蜕变，重塑智能经济的底层服务逻辑与价值体系。随着光互联、异构协同等核心技术持续落地，市场化机制与行业标准不断健全，算力将真正成为普惠全民、赋能千行百业的核心生产力，持续激活智能经济创新活力。

九章云极董事长方磊：

从一度电到“一度算力” 寻找智能时代的计量标尺

● 本报记者 王婧涵 杨洁

一边是高性能算力供给持续紧张，另一边是普通客户GPU利用率不足20%、存在资源闲置，九章云极董事长方磊在接受中国证券报记者专访时形象地将当前算力产业链所需的转变形容为从打造高端跑车到造出平价电车。

他表示，当前人工智能从训练进入推理与应用的下半场，行业竞争点不再仅限于训练性能更强的单点大模型，搭建现代化的工厂体系，锻造标准化、规模化、低成本、高稳定交付的智能能力，显得更为重要和迫切。近日，九章云极发布“AI工厂”战略，提出以规模化工业级的AI底座重新定义全球智算云，向成为智能时代基础设施运营平台的目标迈出重要一步。

降低用户使用门槛

“除非是基础模型厂商，不然绝大部分客户在单一业务场景里GPU的使用率很难超过20%。”方磊表示，“这就像租了一辆车整月待命，实际只开两三天。”

方磊坦言，解决这种矛盾的供需现状，成为九章云极提出“一度算力”概念的初衷。

类比电力行业“一度电”的逻辑，九章云极将“一度算力”定义为312 TFLOPS×1小时。这种算法让用户不是以时间而是按实际消耗的算力量付费，不用为闲置硬件买单。

方磊进一步解释称，从价值链条看，“一度算力”体现了算力产业的三层附加值递进：最底层是一度电的能源成本，向上叠加芯片折旧、调度系统、网络互联等软硬件技术后形成一度算力，再叠加不同模型的能力，转化为可直接交付的Token服务。

这套按算量付费的模式，拉低了算力的使用门槛，也是九章云极“普惠算力”的核心支撑。

“我们的客户还有不少是大学生、研究生。九章云极一度算力的定价是18元，他们花几百元钱就可以在平台上进行AI实操，能以很低的价格了解AI、大模型的全流程。”方磊说。

他进一步举例表示，公司有一些中小企业客户，其用户访问是偶发性的，如果按照传统方式租用算力，起步就要签几千万元；但按“度”购买后，一年算下来只花了约200万元。这样的模式对打算“试水”AI的大型企业也十分友好，方磊介绍，有制造业领域大型客户第一年在平台上花费600万元来做AI创新尝试，第二年就增长到3000万元。

实现千倍级综合降本

“随着大模型对AI芯片的需求更多从训练走向推理，靠堆高性能顶尖芯片所获得的边际效益持续下降，就像人不应该一直开着高端跑车去送外卖，我们需要平价好用的电车。”方磊表示，AI产业的竞争内核已经发生了根本性变化，行业比拼的不再是单一模型的性能高低，而是能否搭建一套规模化、低成本、高稳定的智能能力标准化生产体系。

近日，九章云极在战略发布会上披露了三大战略目标：计划建成10万P智能算力集群，实现单日10万亿Token的流转承载力，依托全栈自研技术实现千倍级综合降本。

“千倍降本绝非低价内卷的价格战，本质是底层工程体系效率之战；极致效能是大规模稳定运营的根基，也是长期成本优势的核心来源。公司的业务本质不是售卖智算硬件，而是运营一套将算力投入转化为专业Token智能产出的完整工业化交付体系。”方磊表示。

在技术路径上，九章云极提出系统架构、计算调度、能效架构三大范式重构。“现在AI训练环节依赖高强度计算，而推理环节需要在计算之后反复调用内存，两者对硬件的要求不同。这实际上为国产芯片提供了机会窗口。”方磊表示，“行业都在积极探索，在训练环节继续使用顶尖GPU芯片，推理环节采用其他芯片配合，不仅节省成本，还能根据不同类型的模型，形成更适配的组合。”

据悉，当前九章云极已在山东、安徽、宁夏、浙江、青海、云南、湖北、广东等多个区域完成智算中心布局，海外也已在印度尼西亚实现节点运营，并在全球多个国家和地区积极推进布局。

算力组网 聚沙成塔

当前，全国算力网建设正如火如荼地展开。国家发展改革委等部门在《国家数据基础设施建设指引》中明确提出，推动国家枢纽节点和需求地之间400G/800G高带宽全光连接，引导电信运营商等提升“公共传输通道”效能，推进算网深度融合。

方磊坦言，与全球顶尖芯片比，国产芯片受限于单卡性能等因素，在单点规模突破上存在客观瓶颈，“单点突破不好走，那就把分散的节点连起来，聚沙成塔，一样能支撑大规模训练。这也是算网融合、以网强算的核心逻辑。”

他表示，国家枢纽节点的建设落地已让远距离算力协同的成本大幅下降。“两年前，济南到北京之间100G带宽的年租赁费用可能超过百万元，如今已降至二三十万元的区间。这让跨区域调度算力从技术概念变成了可落地的商业场景。”

不仅如此，有了高速专网打底，企业也不用将算力投资押注在单一区域，可以多点布局，相当于用一张网托住了所有分散的投资，不用再只看单点的成败。

“比如做基础模型的公司，可以买一些九章的算力，再买一些其他公司的，因为有高速专网，他们就可以把两地的算力一起使用。”方磊强调，这种统筹和组合应该由使用者而非算力提供者实现，“最终用户自己会做一套系统来决定如何调配不同云厂商的算力，因为他们是最关心使用成本的主体。通过市场主体之间的充分竞争，才能带来持续的技术创新和成本下降。国家搭建的算力基础设施，恰好为这种市场化交易提供了可运行的底座。”

腾讯汤道生：

算力网系统远比想象复杂 关键是要把算力资源用好

● 本报记者 杨洁

日前，腾讯集团高级执行副总裁、云与智慧产业事业群CEO汤道生在接受中国证券报记者采访时谈及算力资源紧缺之下算力网的意义和挑战。“匹配不同类型的工作流对于算力的需求，是一个非常复杂的优化问题。”汤道生表示，算力网绝非简单的算力堆砌，数据与算力的地理距离、不同类型工作流在不同阶段需要什么样的算力存力等资源调度与匹配问题，共同构成了一道复杂的系统难题，但解题的方式也很简单，就是“不断提升资源利用率”，“把自己的算力资源用好”。

算力不够用

“我们一直以来在算力基础设施方面的确是处于一个不太够的状态。”采访中，汤道生直言不讳谈到算力资源的稀缺并表示，Token调用已经出现爆发式增长，但仍然受限于算力的供给。

算力不够用，也直接体现在腾讯的资本开支曲线上。据披露，2025年全年，腾讯资本开支达792亿元，同比增长仅3%，低于原定目标，腾讯总裁刘炽平在财报电话会上解释称，主要受GPU供应限制——“我们也想买卡，但在很长一段时间内都面临买不到的情况”。

为支持模型迭代和AI基础设施建设，2026年一季度腾讯资本开支达到319.4亿元，同比增长16%。汤道生表示：“我们非常期待下半年有更多国产算力可以支持到我们的云业务，可以把一些推理场景服务得更好。”

在记者问及腾讯是否会下场自研算力芯片时，汤道生直言：“自己去做芯片设计并不解决产能问题。我们跟很多芯片厂商、合作伙伴都有打交道，我相信没有一家有足够的产能去满足今天市场的需求。”他强调，腾讯选择开放生态的战略，“可以跟更多芯片厂商合作，也让国产算力芯片厂商愿意拥抱腾讯，作为它们算力应用的一个展现标杆。”

算力网系统复杂度高

在算力资源稀缺的情况下，汤道生认为算力网建设的重要性不言而喻：“算力调度在资源缺乏的时候肯定很重要，让资源效率能够做到最高。”但他也同时指出了其中的技术复杂性——“毕竟不同类型的工作流对于算力的需求，在流水线什么阶段需要算力、什么阶段需要闪存，是一个非常复杂的优化问题。”他强调，算力网系统远比外界想象得复杂，需要极强的技术能力。

汤道生还指出，异构芯片的算力集群做一体化调度难题尤为突出。此外还要考虑数据局部性的问题，因为数据与算力之间的距离也非常重要。“算力也许在北方，但是如果要解决问题的数据在南方，长距离也会带来很大的资源损耗。”汤道生介绍，国家有大的集群和算力网，腾讯内部也在不断优化不同业务形态对算力的需求——有多模态模型训练的需求、推理的需求、标注的需求、音视频处理需求等，工作流很复杂，但争取把自己的算力资源利用率提高，把算力资源用好。

Token爆发重塑商业模式

Token调用量的爆发式增长正在改写AI行业的成本公式和商业模式，也带来了新的焦虑。汤道生说，已经听到很多客户甚至同事们在紧盯着积分消耗或者Token消耗，这方面的成本不容小觑。

每一个Token生成背后都是GPU的一次次运算以及数据中心的一度度电。汤道生指出，移动互联网的边际服务成本相对较低，可以通过广告或眼球经济带动某些交易行为，从而建立商业模式。但今天AI的原生服务，运营成本、推理成本还非常高，在这种情况下，很难单纯通过移动互联网时代广告模式来涵盖成本，并且Token的成本与任务的复杂度强关联，成本消耗差异很大，也很难通过过去的商业模式变现。

刘炽平此前在财报电话会上也表达了类似观点：在互联网场景下，信息交付的可变成本非常低，主要是带宽成本，而且算力大多是在用户设备完成的，因此互联网产品几乎可以实现无限扩展；但在AI场景下，每一次向用户交付智能服务，其实都会产生相当可观的成本，“我们不能简单地把互联网时代的逻辑直接套用到AI上。找到高价值场景，与单纯盲目获取大量日活跃用户和用户时间相比，同样重要，甚至更重要。”

汤道生表示，商业化不是团队目前做产品的重点，主要还是把产品打磨好，而正因为当前算力资源有限，也构成了一种无形的产品筛选机制——哪些功能是用户最有需要、最认可其价值，并值得为之付费来获得算力，这是AI Agent产品发展过程中需要考虑的地方。