算力需求暴增，国产芯片如何破局

『用得上、用得起』的算力供给是人工智能产业的核心难题

算力的核心是芯片。当前，算力需求的爆发式增长，正在重塑人工智能产业格局，也给国产算力芯片带来了前所未有的挑战与机遇。

数据显示，中国日均Token（词元）调用量已从2024年初的1000亿跃升至2026年3月份的140万亿，两年间增长超千倍，相比2025年底的100万亿，三个月时间又增长了40%。智能体的兴起更是进一步推高了算力需求，算力不足已成为制约产业发展的核心瓶颈。Token消耗、算力成本优化成为业内人士关注的焦点。

在此背景下，超节点作为破局关键，成为国产芯片厂商角逐的核心赛道。券商机构预计2027年至2028年，超节点渗透率将从当前的10%—20%提升至50%—60%。

算力需求与成本同步走高

算力短缺的困境，已成为行业共识。

“我认为，未来12个月，最关键的问题可能还是算力。”智谱CEO张鹏说，无论是Agent带来的十倍效率提升，还是行业百倍的需求爆发，最终都要落到“用得上、用得起”的算力供给上。面对推理时代需求的指数级爆发，如何解决算力供给的核心难题，是全行业必须共同面对的课题。

无问芯穹联合创始人兼CEO夏立雪透露，无问芯穹从今年1月底开始，每两周Token使用量就会翻一番。当前Token调用量大幅增长，如同当年手机流量从100M时代开启爆发式增长的前夜。

算力焦虑正贯穿近期各类行业论坛，如何破解算力短缺、降低算力成本，是行业关注的焦点。

清微智能研发副总裁李彬受访时坦言，智能体时代，算力需求与成本同步走高，“我国每天调用140万亿Tokens资源，按照我们通常所说的算力，每天有上万台机器在不停运转。”在他看来，降成本，是让智能体真正走入千行百业、千家万户的关键。

而Token消耗作为算力成本的核心组成部分，其消耗逻辑与成本控制也备受关注。360集团创始人周鸿祎受访时强调，Token永远不会像手机流量那样，实现包月无限量的使用模式。“互联网流量的消耗，与使用时长、使用方式呈正相关，固定成本相对稳定，随着用户规模扩大，单个用户边际成本可降至极低水平，但Token的消耗逻辑与之截然不同。”

周鸿祎进一步解释其中的核心逻辑：Token的消耗本质上是智力资源的消耗，任务复杂度越高，所需消耗的智力资源便会同步提升，而智力资源的消耗与成本呈正相关。与此同时，算力的背后是电力支撑，算力消耗的过程本质上也是电力消耗的过程，Token消耗规模越大，电力消耗便越多，这一逻辑遵循信息量与能量守恒定律，不存在以最低成本完成高复杂度工作的可能。“美国曾有企业尝试推出AI服务包月模式，最终因被恶意薅羊毛导致损失巨大，这也印证了该模式不可行。”周鸿祎说。

超节点成破局核心赛道

在算力需求爆发的背景下，超节点成为国产芯片厂商角逐的核心赛道。作为国产算力芯片领域的深耕者，清微智能在2026中关村论坛年会上亮出了自己的创新成果，其联合智源研究院发布“可重构智算超节点”技术，将4096颗可重构计算芯片互联构建成一个超节点。不同于传统超节点依赖交换机的模式，该技术通过可重构网络互联技术，让芯片自身具备智能路由能力，实现无交换机光纤直连组网。“芯片与芯片相连需要经过交换机，交换机越多损耗越多，边际效益就会递减，因为‘会浪费一些时间在路上’。”李彬解释道。

超节点已从概念走向产品，现在进入了实际应用阶段，而这也正是国产芯片实现差异化竞争的重要突破口。“这种无交换机的互联方式，能实现高带宽、低延时的算力输出。目前，该超节点已落地国内多个智算中心。”李彬说。

百度、华为、中科曙光、阿里云等企业也纷纷布局，形成多点开花的竞争格局，共同推动国产算力产业升级。

百度在超节点领域的布局颇具前瞻性，在2025百度世界大会上，百度发布了新一代昆仑芯M100和M300，同步推出天池256超节点与天池512超节点，计划于今年正式上市，其中单个天池512超节点就能完成万亿参数模型训练。2025年，昆仑芯已累计完成数万卡部署，百度已点亮昆仑芯三万卡集群，可同时支撑多个千亿参数大模型训练。

华为此前推出的Atlas 900 A3 SuperPoD（Coud-Matrix 384超节点），已累计部署300余套。3月2日，华为首次在海外展示最新的Atlas 950 SuperPoD，以及TaiShan 950 SuperPoD等多个型号的超节点产品和解决方案。Atlas 950超节点最大支持8192张昇腾950DT卡通过“灵衢”全光互联，这将是昇腾384超节点的20多倍，其中FP8算力达到8E FLOPS，FP4算力达到16E FLOPS，互联带宽达到16PB/s，预计于今年四季度上市。

中科曙光、阿里云等企业也纷纷加快超节点布局。2025世界互联网大会乌镇峰会期间，中科曙光正式发布全球首个单机柜级640卡超节点scaleX640，采用超高速正交架构、超高密度刀片、浸没相变液冷、高压直流供电等技术。

阿里云发布全新一代磐久128超节点AI服务器，由阿里云自主研发设计，可支持多种AI芯片，单柜支持128个AI计算芯片。

此外，上海仪电联合曦智科技、壁仞科技、中兴通讯发布了国内首个光互联光交换GPU超节点——光跃超节点128卡商用版（LightSphere 128），以曦智科技全球首创的硅光OCS光交换芯片为核心，搭载壁仞科技自主原创架构的大算力通用GPU液冷模组壁砺166L，并集成中兴通讯高性能AI服务器及自研软件平台。

算力成本有望持续优化

面对当前算力芯片市场的同质化担忧，李彬并不焦虑。在他看来，算力芯片是整个算力基础设施的最底层，人工智能产业的发展肯定需要一个过程，这个过程中，大家可能会产生对应用或者前景的过度追求，但从中长期来看，算力最终是解决大家生产生活中的实际问题。“今天我们随便问一个大模型App或者相应的智能体，它能帮你直接解决问题，而不是简单回答你的问题，我认为这个趋势不会改变。”

李彬强调，清微智能的核心竞争力，在于从0到1原创芯片底层架构，通过“软件定义硬件”技术，让芯片硬件能根据不同AI任务实时动态重组。这如同赋予了流水线工人自主协作的智慧，天然契合AI算法并行、流式、密集的核心特质。因此，清微智能的可重构芯片也被称为“变形金刚”，可兼顾高效性与灵活性，实现低延迟、低能耗。这种差异化创新，也让企业获得了资本的认可，清微智能于2025年底获得北京国资领投的超20亿元C轮融资，并于今年3月正式开启IPO征程。

除了技术创新与资本支持，算力成本的持续优化也是产业发展的关键。对此，周鸿祎也给出了自己的思考：“未来算力使用成本将逐步下行，主要有两大优化路径：其一，大力研发推理芯片，通过专用芯片的迭代升级，将推理成本降到原来十分之一乃至更低；其二，依托国家电力产业的持续发展，当前我国用电成本已处于相对较低水平，若人类在核聚变领域取得突破性进展，届时电力成本将大幅降低，算力成本也会随之显著下降。”

业内人士表示，随着各类企业在超节点领域的持续深耕，以及技术、资本持续集聚，国产芯片正逐步打破算力瓶颈，未来随着超节点渗透率的提升，国产算力产业将迎来高质量发展，为人工智能产业的持续突破提供坚实支撑。

记者 董婉婉

据2026年4月8日《大众日报》