

WAIC 2026“超节点”密集登场 国产AI基础设施迈向系统级竞争时代

走进2026世界人工智能大会（WAIC）H2算力主题展区，一整排像墙一样的黑色机柜立即映入眼帘，这是华为首次线下展示的昇腾950超节点真机，1024张NPU卡高速互联，像一台计算机一样工作，能够提供1 EFLOPS FP8、2 EFLOPS FP4的澎湃算力。

这似乎宣告了一个系统级的算力时代到来。展馆里几乎所有算力厂商，都推出了卡数各异的超节点产品——中科曙光展示了全国产十万卡AI超集群曙光8000（登峰）、中兴通讯发布了OEX超节点、沐曦股份展示了曦景S600超节点……整个WAIC 2026几乎被这些“大家伙”占领了C位。从单卡的性能比拼到超节点系统的整体较量，国产算力正经历一场深刻的竞争范式转移。

● 本报记者 杨洁

算力供不应求

“算力处于供不应求状态，整个人工智能的产业规模比互联网产业规模要大得多。”摩尔线程高级副总裁董飞龙在接受中国证券报记者采访时说，“人工智能时代才刚刚开始，我们投入打造的算力底座，实际上是为未来十年、二十年搭建的新型数字基础设施。就像当年大规模建铁路、高速公路一样，水泥等基建材料肯定供不应求，因为需求太大了，而今天就是这样的场景。”

供不应求并非空话。他解释称，模型参数从千亿奔向数万亿，Agent应用将上下文窗口从256K推至1M

甚至更高，推理和训练任务动辄横跨成百上千甚至上万张卡，超节点解决的正是让这些卡能更高效、更稳定地分工协作的问题。“这好比是一个交响乐队，256张卡就像256个乐手，每一个乐手自己都拉得很好。当整个乐队要合奏一首曲子时，大家就需要协同，这里蕴含着挑战。”

壁仞科技AI框架架构副总裁丁云帆告诉记者，AI的发展正面临“超大参数、超长文本、超大集群”三大核心挑战，仅凭单张GPU的算力，已无法独立应对，超节点正是在此背景下成为市场的焦点。

系统级大考

超节点并不是简单地把芯片塞进机柜，其核心思路是放大纵向扩展（Scale-up）范围，将更多GPU通过高带宽、低时延的互联方式紧密耦合，使它们如同一台超级计算机那样协同高效工作。

首先需要打通的是通信瓶颈——超节点内GPU间需要高速交换设备互联，这也是超节点厂商着力攻克的技术难题。壁仞科技联合创始人、CTO洪洲表示，AI超节点正面临规模扩展、带宽提升和通信延迟等多维挑战，公司自研的BLink™2.0超节点互联协议，以“超节点、在网计算、拥塞控制、链路修复”四大核心能力构建了开放、自主可控的算力底座，目标就是实现“时延一带宽一可靠”端到端系统级协同。百度智能云混合云部总经理杜海告诉记者，百度智能云自研XPU-Link互联协议与可编程交换技术，把超节点内大量加速卡组织为高带宽、低时延统一通信域，重构Scale-up全互联通信域，支持从32卡扩展到最高1024卡。

接着还有物理工程难题需要解决。董龙飞介绍，超节点的功耗非常大，一个超节点可能要达到400KW，一台高密度机柜就能达到

48KW，而两三年前国内的普通机柜仅6KW，供电、散热、承重、布线，每一项都是考验。

更棘手的还有可靠性与运维问题。一个超节点价值动辄超亿元，涉及数百甚至上千张芯片运行。沐曦股份联合创始人、CTO彭利表示，厂商需要监控每张芯片的工作状态，在通信异常后保存故障现场、恢复任务，并在软件层面将工作负载拆分至不同芯片，“它更多考验的是系统化的能力。”董龙飞也坦言，“以前一台服务器是8张卡，一张卡对整体影响不大；现在超节点里坏一张卡，价值上亿元的设备就可能停摆，哪怕只是停用一天，也会损失很多钱，所以保证超节点运行的可靠性、解耦性非常关键。”

伴随Token时代全面到来，成本和服务优化不断受到关注。燧原科技创始人、董事长、CEO赵立东表示，Token正成为AI产业的价值载体与计价单位，行业的关注重点从“算力规模”转向“Token价值”，客户不再单纯为硬件成本付费，而是以Token生成效率与交付质量作为核心评判标准，“这要求我们必须从系统级、架构级层面优化算力综合成本，更好支撑Token经济发展”。

“上天入地”算力演进不息

超节点并不是算力进化的终点。在本届大会的各个角落，更具前瞻性的探索已浮出水面。

光互连被普遍视为突破电互连物理极限的关键。大会期间，壁仞科技推出下一代NPO（近封装光学）光互连、分布式解耦架构超节点，支持单超节点1024卡Scale-up扩展。丁云帆解释称，现有铜缆电信号在3米内严重衰减，而光信号传输可达数百米，天然适合大规模GPU互联。壁仞选择NPO路线，将光引擎直接与GPU模组融合，省去高功耗的DSP芯片，在性能、功耗与工程落地间寻求平衡。这意味着未来的超节点或告别定制化的“大铁盒子”，采用标准机箱、光纤互联，像搭积木一样可以实现灵活扩展。

在芯片硬件底层，先进封装、新材料相关技术正持续迭代。在本届大会期间，东方芯晟展出全球首颗基于国内成熟工艺的“软件定义+3D堆叠近存计算”AI芯片DF1000，试图以14nm+工艺对标4nm芯片性能；燧原科技宣布，联合先封科技发布CoPoS（面板级先进封装）样品，探索玻璃基板在封装

中的迭代应用。这些微观层面的技术聚合，正在为下一代超节点乃至更大规模算力集群铺路。

值得注意的是，算力布局边界正跳出地面机房的局限。在本届大会期间，上海发布“星枢计划”，旨在构建“天数天算、地数天算”双向协同的天地一体化智能算力网络。清微智能与东方星链联合宣布，基于国产可重构芯片的全球首台太空AI智能体已完成研制，将于2026年底入轨。燧原科技联合星际天算发布天基算力应用场景，致力于将复杂天基算力解构为标准化的计算Token。

系统化能力提升之后还有一站——生态。正如东方算芯创始人、董事长兼CEO魏少军所言：“算力产业的竞争，最终是体系与生态的竞争。”摩尔线程高级副总裁高林丽告诉记者，作为芯片厂商，不能让开发者今天适配这家、明天适配那家，工具链五花八门，这样开发的成本太高，所以业界正在推动生态整合，在技术上实现“写一套代码，适配所有国产硬件”。只有解决了这一难题，整个国产算力行业才能拧成一股绳，实现蓬勃发展。



俯瞰WAIC 2026“登峰时代·伙伴之城”展区
本报记者 杨洁 摄

启明创投周志峰：算力需求远超预期 将做更积极布局

● 本报记者 杨洁

“国内某科技大厂去年算力预算最终落地500多亿元，而今年的预算是去年的6倍。”启明创投主管合伙人周志峰近日在接受中国证券报记者采访时

表示，算力需求增速远超预期，启明创投今年在算力板块的投资布局会比去年更积极，从新一代AI芯片、光通信芯片，一直延伸到新型数据中心散热、电源供电等配套环节，都在周志峰团队考量的算力布局范畴。

但硬币的另一面是，AI应用的打开方式与预想不同。周志峰认为，目前市场还未跑出一款比肩微信、抖音的国民级AI应用。受Token线性算力成本制约，C端AI应用无法依靠规模摊薄成本，商业模式难以维系。因此他把应用侧投资聚焦垂直场景与高付费用户。身处算力、应用、资本三重火热的AI周期中，他反复强调两大原则：布局节奏要比行业共识快半步，投资范围守住自身能力圈。

算力积极扩张

谈及算力，周志峰一连用了多个“超预期”来描述自己的感受。海内外

科技巨头纷纷上调AI算力投资预算，他据此判断，未来五年全球新增算力投入总额将达5万亿美元，并称这是当前商业市场里确定性更强、落地更快的核心刚需，赛道逻辑清晰。

重要的还有算力需求结构的变化。前几年算力主要用于大模型训练，但今年推理算力需求已超越大模型训练需求，且增速惊人，这对于国内近年来雨后春笋般成长起来的GPU芯片企业来说是重要机遇。“两三年后，像字节跳动这样的科技大厂，每年的推理算力需求都会达到百万张卡级别，这是真实且巨大的需求。”周志峰算了一笔账，国内GPU芯片企业哪怕只分到1%的市场份额，都意味着百亿元级营收规模。

而芯片设计企业能否锁定供应链产能、保障供货稳定，是行业洗牌中决定企业能否站稳脚跟的核心关键。“市场向来都是先分散，再集中，五年或十年之后国内还能剩下几家GPU芯片厂商？我相信行业一定会有收敛，不过当前的爆发阶段确实能容纳很多企业，毕竟整体需求实在太大了。”周志峰说。

周志峰透露，启明创投今年布局算力赛道的力度将较去年明显加大。除国产GPU芯片外，团队正密集挖掘算力赛

道下各类新兴细分技术，涵盖新一代云端AI推理芯片、热管理方案等；即便是太空算力这类尚未形成行业共识的赛道，团队也已着手布局产业链上游市场化程度较高的环节。谈及热管理赛道，周志峰举例：“未来仅一家头部科技企业每年芯片采购量可达数百万片，若热管理方案能降低1%的能耗，一年即可节省亿元级电费，产业经济效益十分可观。”

应用仍在探索

与算力的火热形成鲜明对比的是，AI应用端发展让周志峰感到“未达预期”。他去年曾判断2025年是AI应用爆发元年，但现实却是另一番景象。

“互联网和移动互联网时代，产品是核心，上游是流量和用户，下游是商业化，路径清晰，且边际成本持续递减。但AI时代完全不同，To C端应用的商业化模式并没有发生质的改变，流量成本与移动互联网时期持平甚至更高，服务1万用户和1亿用户，Token成本几乎是线性放大的，边际成本没有下降。”周志峰认为，这是当下C端AI应用最难解决的痛点。

他举了一个例子：某AI玩具企业卖出几十万台产品，但九成用户并未长期启用AI交互功能。“这家企业告诉我们，这反而是好事——如果用户高频使用AI、持续消耗Token，这些成本公司根本扛不住。”

在B端应用领域，周志峰同样强调要算好Token这本账。“如果产品体现不了产业价值，营收覆盖不了Token成本，我们就不会看好。”他更看好诸如AI教育、AI金融这些赛道，因为商业闭环逻辑更清晰——AI虚拟老师能为每位用户提供高质量的定制化教学内容。对比传统的真人教师授课模式，其价值直观突出，成本也显著低于购买真人教师线上或线下课程。而金融机构早已养成为彭博、万得这类金融数据终端付费的习惯。

警惕热钱涌入风险

“头部大模型公司市值能够冲击万亿元级规模，同时市面上涌现出大量新的大模型创业公司，创始人基本都是95后、00后，这类项目天使轮估值都能达到二三十亿元。从业这么多年，我从未见过如此火热的行情。大模型技术的发展及资本市场对大模型企业的追捧热度，都远超我的预判。”周志峰坦言。

这种高估值也是一把双刃剑，周志峰表达了担忧：“有些企业刚完成融资，本身并不缺钱，一批鲜有耳闻的机构主动上门寻求增资，开口就要把估值抬升50%甚至一倍，要立刻追加一轮投资。这批热钱的涌入对行业冲击很大，企业拿到超出自身需求的资金，很可能会干扰自身战略判断和日常经营。”

此外，大量热钱也会助长行业内卷。周志峰表示，当下市场存在不少非理性行为，“算力成本水涨船高，原本300万元一台的服务器直接炒到上千万元，全行业算力成本被动抬高；人才被恶性争抢，行业薪资飙升；客户层面无序竞争，各家都没有成熟落地的商业场景，只能扎堆抢夺大客户，单纯比拼营收规模。”这些现象长期来看不利于行业发展。

面对市场狂热以及投机行为增多，周志峰反复强调“守住能力圈”。“很多行为都是被市场热度与FOMO（错失恐惧症）驱动。我们团队更强调坚持‘快半步’的方法论——在市场共识形成前一年进行布局，只做能力圈内的投资。”他认为在市场热度空前高涨的环境下，遵守投资纪律是一件特别重要的事。

“行业发展越是狂热，越要吸取历史教训。”周志峰称，“互联网、移动互联网周期里都有前车之鉴。不能创造实质价值的企业，哪怕受资本追捧两三年，终究会归于平淡。我时常提醒团队，杂音充斥的环境下，哲科思维至关重要，务必把业务逻辑想明白、研究透彻。”

商汤科技则在本届大会期间发布算电协同Agent，基于算力、电力、Agent三位一体体系，使商汤大装置实现单位电量Token产出提升80%、平均电费单价低于同地区数据中心的约10%。

“Token工厂”优化算力使用

除了基础设施升级与算电协同，“Token工厂”正通过优化算力调度、算力运营来提升算力使用效率。

百度智能云将原有的“MaaS模型服务”升级为“Token工厂”，尽可能减少Token重复计算，推理生成速度较市场水平提升约25%，同时支持多款国产主流模型的调用，提高算力服务性价比。

杨新安介绍，新华三集团通过优化服务机制，充分发挥算力效能。例如，根据用户付费水平分配不同等级的模型服务，根据问题复杂度匹配相应模型，在保障输出质量的同时减少Token消耗。企业或工厂可通过后台界面查看每个Token的具体使用路径、成本与计费详情，管理日常运营中的算力成本。



WAIC 2026摩尔线程展台
本报记者 杨洁 摄

技术升级、算电协同推动算力普惠

● 本报记者 郑萃颖

随着AI算力需求不断增长，高质量、低成本算力供给成为市场所需。如今，词元（Token）成为衡量算力消耗的最小单位，“Token工厂”则是将AI基础设施、模型和电力能源相结合，提供算力服务。在2026世界人工智能大会上，行业企业通过基础设施能力提升、算电协同、算力调度优化等方式，推动算力普惠。

技术升级降低算力成本

合肥中科类脑智能技术有限公司副总经理、算力Token运营事业部总经理常峰告诉记者，具备低成本、规模化的AI算力供给能力，将是未来全球AI基础设施领域最重要的竞争壁垒。

中科类脑研究院院长丁海松介绍，通过提升国产芯片性能，依托高效KV Cache（键值缓存）管理、动态PD分离（预填充与解码分离）节点优化等前沿技术，可最大化提升设备的Token吞吐效率。

新华三集团AI解决方案专家杨新

安表示，网络设备、存储机制是影响算力成本的关键因素，“800G网络较400G网络可显著提升相同算力下的Token吞吐量。”在存储机制上，通过分级缓存可以降低算力成本。

算力架构同样影响算力成本。在本届大会上，商汤科技展出的商汤大装置通过异构混合推理，使整体推理性价比提升，同样成本下算力产出规模扩大2.5倍。太初（杭州）集成电路有限公司（简称“太初元基”）采用HCU异构融合计算架构，CPU与AI算力核通过统一高速总线互联，配合共享分级存储架构，提升整体效率。

太初元基CEO杨晋喆认为，当前算力产业已从单卡性能比拼进入系统级竞争阶段，模型参数突破万亿级、“Token工厂”向吉瓦级演进，降低单位算力成本成为核心命题。

算电协同提升能效

降低算力成本的另一条路径是降低用能成本，其核心在于算电协同。

中科类脑算电一体化平台可基于

分时电价、风光发电实时波动，灵活迁移算力任务负载，主动错峰使用低价清洁能源，从源头压缩算力用电成本。中科类脑董事长兼总经理刘海峰介绍，一个1000P（PFLOPS）标准算力集群年度总耗电量约2000万度，依托平台优化可节约20%用电量，每年预计可辅助算力集群减少400万度电力开支。

润建股份通过RunDoBidder虚拟电厂和RunDoGrid智能微电网管理系统实现算电协同。公司相关负责人介绍，微电网汇聚分布式光伏、风电、储能等资源，通过在线数据监测实现统一管理，根据对算力中心的任务分析与负荷预测，数据传输至虚拟电厂，虚拟电厂结合电价预测生成调度策略，再下发调度优化指令。湖南某园区引入微电网平台后，每年因能源利用率、运维效率提升，产生效益约100万元。

中兴通讯在本届大会期间展出了算电协同、源网荷储一体化AIDC整体解决方案，通过电力、算力、热力融合管理，实现全域统一调度，提升单位电力与算力产出，使每一度电支撑更多AI Token计算。