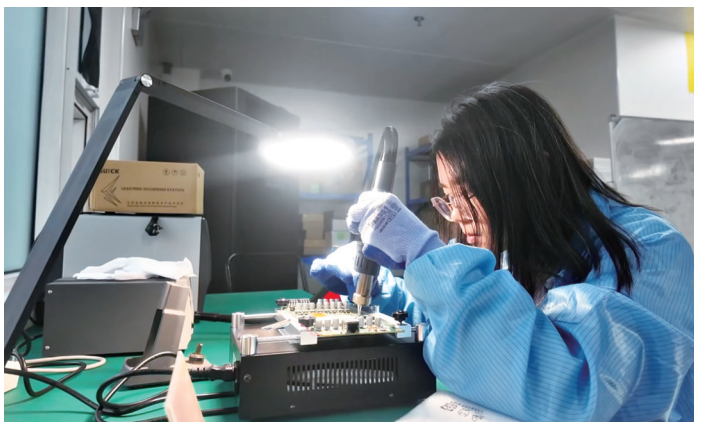
大模型“狂飙”之下 国产算力缺口如何补齐

▶ 本报记者 孙庆阳

2026年以来，国产人工智能（AI）大模型持续迭代，多家知名企业更新甚至“周更”自研大模型，以满足时代发展需求。然而，在大模型竞相发力的背后，算力“荒”逐渐显现。随着热度激增，不少知名国产大模型在发布后发现响应卡顿，用户在使用过程中出现AI回复延迟、回答偏离等问题。一位行业人士向记者透露，相比去年，今年大模型竞速式更新对底层算力需求更为突出。算力一旦没跟上，前端使用就容易产生问题。



左图：汉腾科技生产线

右图：太初元碁实验室工作人员正在做实验。

受访者供图

“在人工智能+制造浪潮下，算力是核心底座，但行业缺少的从来不是单纯的算力硬件，而是能真正适配产业需求、好用又可控的解决方案。”在太初元碁硬件系统研发实验室，太初（无锡）电子科技有限公司（以下简称“太初元碁”）首席产品官、市场副总裁洪源，向记者点出了当下国产AI算力产业发展的痛点。他认为，破解这一行业痛点，需要国内算力企业在技术自主创新与软硬件协同上持续攻坚。

持续优化异构众核架构

2026年1月7日，工业和信息化部联合多部门印发《“人工智能+制造”专项行动实施意见》（以下简称《意见》），提出加快突破训练芯片、异构算力等关键技术，推动人工智能产业高质量发展。2月6日，工业和信息化部发布《关于开展国家算力互联互通节点建设工作的通知》，提出构建“1+M+N”国家算力互联互通节点体系，推动算力资源标准化互联与高效流动，为国产算力产业发展按下加速键。

据中国信息通信研究院预测，到2030年全球算力规模将突破16ZFlops（16泽浮点运算每秒），智能算力占比将从2023年的63%大幅度提升至90%以上。而异构算力正是支撑智能算力提升的关键力量。《意见》明确将异构算力列为关键技术突破方向，要求到2027年实现安全可靠供给。

“异构众核架构不是简单的算力拼凑，而是把标量、向量、矩阵3类算力单元有机融合在单芯片内。”洪源解释称，“这种融合不是静态的，而是动态的，能够根据应用特征灵活调配资源，在软硬件协同下释放极致性能。”

目前，太初元碁正致力于推动高性能计算与人工智能的融合创新。太初元碁硬件系统研发负责人吴志勇介绍说，太初元碁异构众核架构围绕计算、访存、通信三大环节展开系统优化。比如，为应对推理任务的高延迟敏感特性，访存通路径做了专项优化，减少数据搬运开销。此外，融合高速网络与跨数据中心传输的高性能通信机制，保障了大规模分布式计算的数据协同效率。

稳定运行同样关键。通过结合高效液冷散热与动态调频调压技术，太初AI芯片可在不同负载下智能调节性能及功耗，实现更高的系统能效与更好的运行稳定性。吴志勇说，这一特性已落地SuperPod 128等高密液冷智算单元，并支持万卡规模的系统互联。

2025年年初，在DeepSeek-R1系列模型开源后，太初元碁团队依托该架构，仅用2小时就完成了在T100加速卡上的高效适配，快速上线多款大模型服务。2026年春节期间，太初元碁还完成多款国产主流开源大模型的深度适配工作，累计适配国内超40款AI大模型，实现即发即适配，上线即可用。

软硬协同筑根基

要实现技术优势真正落地、实现产业价值，离不开全栈能力的系统性整合。

“一手筑牢硬件根基，一手打磨软件实力。”洪源认为，真正能支撑全行业、全场景的通用AI能力，必须建立在软硬一体的全栈技术体系之上，这也是国产算力走向各领域、服务各场景的关键。

通过自研编译器、跨层转置通信库、适配PyTorch2.7.1的TecoPyTorch框架等一系列工具，太初元碁搭建起从底层指令到智能编程的完整工具链。洪源补充道，“AlphaFold3的复现与商业化落地，让AI能力延伸至生物医药、气象等科研领域。”

不仅如此，全栈协同理念也在更广阔的生态展开实践。以国家超级计算无锡中心为例，基于新一代国产神威超算系统，该中心构建了覆盖“芯片—软件—应用”的全栈国产化高性能计算体系，在生物医药、工业仿真、城市治理等领域孵化出多项创新应用。

从实验室到真实场景，全栈能力正在悄然渗透。政务大模型一体机落地上海宝山“一网通办”，教科研实训一体机走进高校，惠企政策一体机助力政企解读。行业人士表示，软硬协同从来不是“硬件+软件”简单叠加，而是两者在架构层面深度融合、在应用层面彼此成就，最终让AI渗透到每一个行业、每一个场景，推动产业智能化升级。

从“能用”迈向“好用”

当前市面上，算力总量看似充足，实则真正高端、好用的算力极为稀缺。

要解决上述问题，需从产品破局。太初元碁生产的全国产化AI加速卡全程自主可控、稳定量产，筑牢基础算力底座；元碁液冷AI工作站适配国产CPU，开箱即用；而入选“十四五硬核成果”的SuperPod128高密液冷智算集群，支持万卡部署，精准填补高端算力空白。

有了产品之后，如何将零散的硬件部署成高效可靠的系统，才是更大考验。太初元碁高密液冷智算集群方案已助力盐城超级计算中心、河南空港智算中心等地项目规模化落地，还将在今年落地汉腾科技“五个万卡集群”。这摆脱了简单的“算力堆砌”，在提升算力密度的同时大幅度优化了算力能效。

“汉腾科技2026年开工的承德、石家庄、兴化3个项目，全是以国产万卡为起点。”南京汉腾蓝域数据科技有限公司（以下简称“汉腾科技”）董事长王皓霆介绍说，该企业计划2026-2027年做到10万卡规模，覆盖国内核心算力办公需求。

“打造算力产业园，核心是整合芯片端、算力端、应用端企业。”王皓霆总结道，推动相关企业联动研发，前置参与产业设计而非中后期介入，才能让芯片、算力更贴合地域化产业需求，这正是国产算力产业融合的关键。他认为，算力市场必将迎来爆发式增长与深刻变革：一方面是各行各业的产业端算力需求猛增，另一方面产业本身也在升级，国产算力的发展周期已经到来。更重要的是，产业需求重点正在转移，从前看重原材料、硬件设备，如今更渴求算力与算法的深度融合。

挑战亦不容忽视。“内存等有技术壁垒的关键核心材料价格涨幅显著，供应链难以快速跟进市场需求；国产大模型此前多适配英伟达设备，2026年全面向国产算力设备靠拢，成为产业端核心转型难题；芯片供应链仍处于初级阶段，高端芯片流片依赖台积电。但值得期待的是，国产光刻机正朝着这一方向奋力突破。”王皓霆补充道。

由“独生子”变为“双胞胎”

我国单量子点高效双光子源研制成功

本报讯（记者 张伟）北京量子信息科学研究院袁之良团队联合中国科学院半导体研究所，在固态量子光源研究方面取得重要进展，使量子点产生的光源由“独生子”变为“双胞胎”。该研究团队成功制备出一种基于量子点—微柱腔耦合体系的高效率、高纯度双光子发射器。该器件利用创新的暗态双激子激发路径，并结合腔增强的简并双激子—激子级联辐射，大幅度提升了双光子发射效率与纯度。3月2日，相关成果以“基于暗态双激子加载的Purcell增强型量子点双光子发射”为题在线发表于《自然·材料》。

确定性双光子态，是量子计量、量子成像和量子生物医学应用中的关键资源。然而，传统基于非线性参量过程的双光子源本质上服从泊松统计，存在多光子事件概率不可忽略的问题。尽管单量子点可通过双激子级联辐射产生光子对，但如何高效填充双激子态并同时实现高亮度、高纯度的双光子发射，一直是该领域的核心挑战。

经过多年研究，该研究团队提出并实现了一种全新的解决方案。团队将单个In(Ga)As量子点嵌入微柱光学腔中，利用p壳层共振激发技术，选择性地将载流子填充至长寿命的暗激子态，从而绕过亮激子的快速辐射复合通道，实现了对双激子态的高效、确定性加载。与此同时，该体系中的双激子与激子能级近乎简并，使得单一腔模可同时增强级联辐射中的两个光学跃迁，从而在保持高纯度的同时显著提升双光子发射效率。

实验结果显示：在弱连续光激发下，该光源的零延迟二阶关联函数$g^{(2)}(0)$高达3966，表现出极强的双光子聚束效应；在脉冲激发模式下，基于二阶和三阶关联测量的光子数分布重建表明，98.3%的发射光子以光子对形式出现，双光子发射效率在第一个物镜处达到29.9%。这是目前固态量子光源中双光子纯度与效率兼具的领先结果。

进一步的时间分辨关联测量揭示了该体系中的受激辐射机制：由于双激子与激子能级简并，第一个双激子光子的发射会加速后续激子光子的辐射，从而增强双光子的时间关联性。

该研究团队还建立了速率方程模型，成功复现了实验中观测到的功率依赖性和聚束行为，为理解双光子发射动力学提供了理论框架。