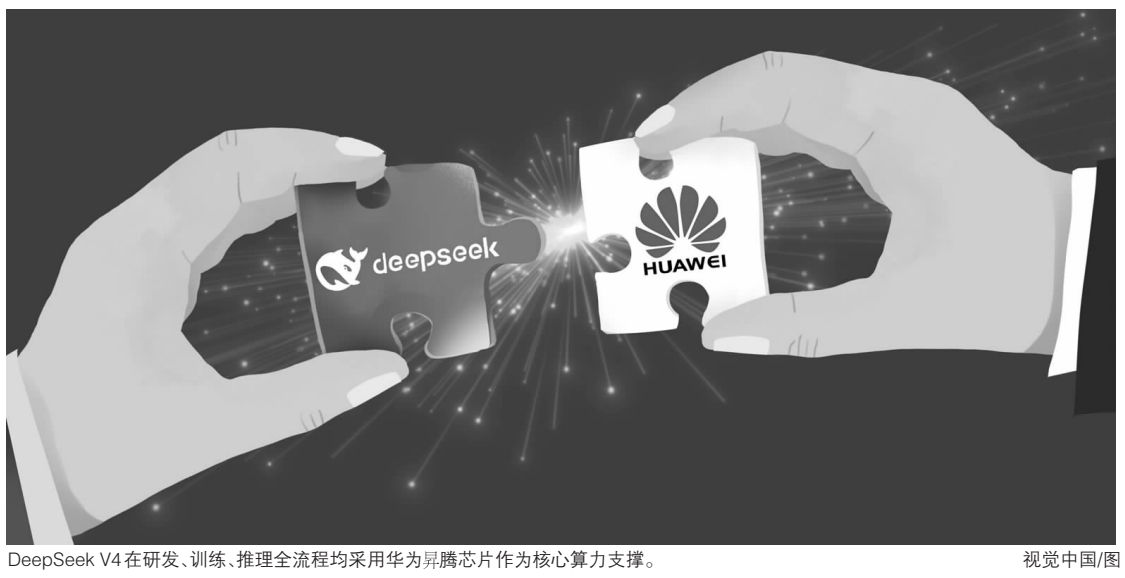
DeepSeek+华为昇腾 全国产AI生态突围

中经记者 秦枭 北京报道

时隔一年多，DeepSeek终于迎来重大更新。4月24日，国产大模型企业DeepSeek正式对外发布新一代大模型DeepSeek V4预览版，包含Pro与Flash双版本，并同步开放技术报告及模型权重开源权限。除了版本更迭外，更让行业关注的是，在该模型发布的同时，华为方面即宣布昇腾超节点全系列产品及华为云已实现对DeepSeek-V4的全面支持。此次联动虽在业界预期之内，却仍引发广泛关注，不仅印证了英伟达首席执行官黄仁勋此前的警示，更标志着中国人工智能产业在降低对CUDA生态依赖方面取得重要进展。

“DeepSeek如果选择在华为芯片上完成首发，对美国来说将是一场灾难。”英伟达CEO黄仁勋在4月中旬的一次访谈中坦言。

多位业内人士在接受《中国经营报》记者采访时表示，DeepSeek V4基于华为昇腾实现全栈适配，标志着国产大模型和国产算力芯片已经打通了从训练到部署的全流程，验证了万亿参数模型在国产算力架构下落地的可行性，打破了此前行业对于“高端AI训练只能依赖英伟达”的固有认知。



DeepSeek V4在研发、训练、推理全流程均采用华为昇腾芯片作为核心算力支撑。 视觉中国/图

“沉默”145天的爆发

4月24日，DeepSeek举办线上发布会，正式推出新一代大模型DeepSeek V4，DeepSeek V4采用双版本布局，兼顾高性能与高性价比。

DeepSeek上一次出现在公众视野中还是145天前。2025年12月，DeepSeek发布V3.2版本后，其研发团队便进入“静默期”。在此后的145天内，国产大模型领域相继涌现春节AI技术竞争、智能体（Agent）发展热潮，以及OpenAI发布GPT-5、Anthropic推出Claude Opus 4.7、Kimi发布K2.6等重要行业动态——对于上述绝大多数关键节点，DeepSeek均未参与。

145天后的2026年4月24日，DeepSeek举办线上发布会，正式推出新一代大模型DeepSeek V4，DeepSeek V4采用双版本布局，兼顾高性能与高性价比。

其中，V4-Pro版本拥有1.6T总参数、49B激活参数，性能对标GPT-5.5，在编程、推理、多模态处理等核心能力上表现突出。

DeepSeek-V4发布后，主流评测平台进行了能力测试和排名。Artificial Analysis对DeepSeek-V4进行了推理能力专项测评。结果显示，V4-Pro在人工分析智能指数中斩获52分，相较V3.2版本的42分实现10分跃升，成为仅次于Kimi K2.6的全球第二大开源推理模型。

除Pro版本外，V4-Flash版本则主打轻量化与低成本，284B总参数、13B激活参数，推理性能接近Pro版本，可满足中小企业及轻量化应用场景的需求。两大版本均标配百万Token（词元）上下文，能高效处理长文本、复杂推理等任务，且通过技术优化，大幅降低了算力消耗，为后续商业化落地奠定了基础。

V4-Flash在评测中的得分为47分，性能弱于V4-Pro，但显著超越DeepSeek-V3.2，综合智能水平对标Claude Sonnet 4.6（全力版），介于顶尖闭源模型与主流中端模型之间。

DeepSeek也坦言：V4与GPT-5.4存在3至6个月差距。

不仅是性能得到提升，DeepSeek-V4在发布后仅两天便启动大幅降价策略。4月25日晚间，DeepSeek宣布对V4-Pro模型API实施限时2.5折优惠。

仅一天后，26日晚间，该公司再次发布公告，将V4全系列API服务的输入缓存命中价格下调至原价的十分之一，其中Pro模型在本年度5月5日前可叠加2.5折限时优惠。调价后，DeepSeek-V4-Flash的输入缓存命中价格为每百万Token 0.02元，DeepSeek-V4-Pro则为每百万Token 0.025元。此价格不仅较国外大模型具有显著优势，同时也低于国内其他同类大模型。

在官宣降价的次日，DeepSeek-V4-Flash的调用量达814亿Token，较前一日环比增长62.2%；DeepSeek-V4-Pro的调用量则为96亿Token。

不仅如此，DeepSeek多模态研发团队的核心研究员陈小康还在社交平台X上公开发文，明确预告“新版DeepSeek V4”即将推出。结合当前语境，这一“新版”毫无悬念地指向了外界翘首以盼的多模态版本。

国产算力的“换芯”

DeepSeek V4发布能够引起关注的原因，是其完成了从英伟达CUDA生态向华为CANN框架的全栈重构。

相较于性能的提升，价格的下降，DeepSeek V4不同于以往国产大模型优先适配英伟达GPU的行业惯例，其在研发、训练、推理全流程方面均采用华为昇腾芯片作为核心算力支撑，华为昇腾同步官宣，昇腾系列芯片（A2、A3、950）已全面完成V4模型适配，其中昇腾950PR芯片成为该模型的主力推理硬件。

华为方面表示，基于DeepSeek V4-Pro模型，在8K输入场景，昇腾950超节点可实现TPOT约20ms。DeepSeek V4-Flash模型，8K输入场景下，TPOT约10ms时单卡Decode吞吐1600TPS，TOPT约20ms时单卡Decode吞吐4700TPS。

除华为昇腾外，在发布会当天，寒武纪（688256.SH）、海光信息、摩尔线程、沐曦股份、百度昆仑芯、阿里平头哥真武、天数智芯等国产AI芯片宣布均已适配DeepSeek-V4。

在DeepSeek V4发布之前，大多数模型围绕CUDA体系开发，并没有摆脱英伟达生态的引力。

国内一家智算中心的负责人告诉记者，长期以来，英伟达凭借GPU的性能优势及CUDA生态的垄断地位，成为全球AI大模型研发的“标配”算力供应商，国内头部大模型企业大多依赖英伟达H100、H20等芯片开展研发与部署。而DeepSeek V4的发布，首次证明了顶级万亿参数大模型可完全脱离英伟达生态，在国产算力平台上实现稳定运行，打破了国产算力无法支撑顶级大模型的行业偏见。

CUDA是英伟达推出的并行计算平台与编程模型，经过多年的发展，已形成完善的软件生态，涵盖算子库、开发工具、应用场景等多个层面，全球绝大多数AI模型的研发与部署都基于CUDA框架。而国产算力芯片及框架起步较晚，无论是生态成熟度还是软件适配性，都与CUDA存在较大差距，这也是长期以来国产大模型依赖英伟达算力的原因之一。

而这正是DeepSeek V4发布能够引起关注的原因，其完成了从英伟达CUDA生态向华为CANN框架的全栈重构，这一过程并非简单的技术迁移，而是一场涉及底层架构、核心算子、精度优化的全方位技术革新，其难度被行业内形容为“万米高空换发动机”，也正是这一重构，奠定了国产算力支撑顶级大模型的技术基础。

路透社称，据知情人士透露，DeepSeek发布V4之前，没有向美国芯片公司英伟达和超微半导体（AMD）提供模型早期访问权限，而是让中国企业华为提前数周开展软件适配优化工作。

路透社在报道中用了一个表述：“breaking from standard industry practice（打破行业惯例）。”

北京社科院副研究员王鹏表示，这一跨越标志着我国AI产业正式摆脱了对外部单一技术路径的依赖。通过全链路的自主实践，不仅在物理层面实现了软硬一体的闭环，更在逻辑层面瓦解了由先发优势构建的生态壁垒。这意味着国产算力不再是应急的替代品，而是具备自我演进能力的独立体系，保障了国家级智能演进的安全边界与技术主权。

天使投资人、人工智能专家郭涛表示，从长期行业发展来看，这一成果将有望逐步打破海外GPU及配套框架长期垄断的市场格局。此次DeepSeek V4发布之际，多款主流国产芯片同步完成原生适配，实现了模型与芯片的高效协同适配，彻底扭转了此前国产AI产业“有芯无模、有模无芯”的割裂局面，真正构建起完善的国产AI自主生态。未来开发者无须再依赖海外单一技术框架，依托国产自研技术体系就能高效完成模型开发与优化工作，随着国产生态持续完善、开发者群体不断壮大，海外技术垄断生态的市场份额与行业影响力将被持续挤压，推动全球AI算力生态走向多元化发展。

在一系列利好催化下，算力板块表现强劲。4月24日发布会当天，A股国产算力相关板块集体走强，其中海光信息（688041.SH）涨幅超8%，寒武纪、中芯国际（688981.SH）等国产芯片企业股价全线飘红，截至4月29日收盘，寒武纪累计涨幅达7.91%。

中信证券认为，DeepSeek V4对国产算力的影响体现在三个维度：一是强化了国产AI芯片使用场景的确定性；二是改变了行业需求结构，市场关注点从训练卡向推理卡、超节点、互联、液冷及软件栈全面延伸；三是提高了国产算力的商业化天花板，Agent、Coding、长上下文等能力进入低成本可用阶段，企业级AI需求有望增加。

王鹏表示，在资本与产业双重维度下，算力板块的走强反映了市场对“自主底座+原生应用”模式长期价值的认可。这种联动效应将资金、人才与需求高度聚集在自主链条上，加速了技术迭代与应用落地之间的正向循环。从长远来看，这将推动我国从算力消耗大国向算力标准输出国转变，在全球数字经济版图中占据更有利的位置。

摩尔线程增长强劲 国产GPU加速商业化

中经记者 李玉洋 上海报道

“国内GPU第一股”摊牌了，带着亮点和弱点。

近日，摩尔线程（688795.SH）发布了上市后的首份年报和一季报：2025年实现营收15.05亿元，同比增长243.37%；2026年第一季度营收同比增长155.35%，跃升至7.38亿元。

然而，在核心财务数据亮眼表现的背后，也存在着这样的“暗面”：经营活动产生的现金流量净额持续为负、预付款和存货价值飙升，以及应收账款的持续膨胀等。

“一家芯片企业的初期本来就需要大笔资金来运维。”行业机构Omdia人工智能首席分析师苏廉节告诉《中国经营报》记者，从财报上看，摩尔线程增长得不错，但弱点也非常明显——头部客户过度集中，“市场竞争依旧激烈，还需要进行融资和政府的辅助来持续运营”。

在他看来，国产AI芯片/GPU在国内仍有良好的发展前景。“AI应用在持续多元化，像摩尔线程这种做全方位的GPU芯片厂商，能从不同的增长赛道中获益，如AI训推以外的3D渲染、具身智能、大数据分析等。”苏廉节表示。

值得注意的是，在2025年年度业绩说明会上，针对投资者关注的供应端是否趋紧的问题，摩尔线程董事长、总经理张建中表示，公司始终与上游核心供应商保持紧密的战略协同，根据市场情况积极进行战略备货。

业内人士认为，摩尔线程上市后的首份财报，其价值不仅在于使外界得以了解该芯片企业的经营状况，更在于首次较为完整地向市场揭示，国产GPU公司在实现产品“做出来”到“卖出去”将面临哪些实际问题。

市场愿意买单

根据财报数据，分季度看，摩尔线程从2025年第一季度到2026年第一季度，当季营收分别为2.89亿元、4.13亿元、0.83亿元、7.21亿元和7.38亿元，其中最近的两个季度营收最为亮眼，可看到市场为国产GPU买单的意愿。

今年3月底，摩尔线程还披露了一份6.6亿元的夸娥（KUAE）智算集群大单，仅此一单就接近公司一季度总营收的九成。要知道，这个订单不是边缘或消费端的零散放量，而是高度集中在云端产品——智算卡、服务器和智算集群，是一份AIDC（AI数据中心）基础设施订单。

摩尔线程表示，公司是市场中为数不多的真正实现千卡级、万卡级大规模集群商业化应用落地的GPU供应商，充分说明了其大规模智算集群的交付能力和市场竞争力。据悉，公司的智算卡已在多家智算中心及云服务平台部署。

在回复投资者提问时，摩尔线程将2025年收入高增长归因于两个因素：一方面，受制于美国对高端GPU芯片的出口政策，国产AI芯片迎来历史性发展机遇；另一方面，公司全功能GPU具有通用性强、支持全计算精度等特点，对于当前多模态、融合计算场景具有较高的支撑性。

经梳理，4月以来摩尔线程基于旗舰AI训推一体GPU MTT S5000 Day-0极速适配了智谱GLM-5.1、MiniMax M2.7、DeepSeek-V4-Flash以及中国移动九天35B这些国产大模型的最新版本。

不过，摩尔线程的业绩增长也存在些许隐患。公司在2025年第三季度实现营收8283万元，第四季度则骤增至约7亿元；同期，前五大客户贡献了全年销售额的91.36%。这表明该公司的增长主要依托于少数核心客户及重大项目。而在实际运营中，客户采购周期、项目验收时点以及预算拨付进度等任一环节的变动，均可能对公司单季度业绩产生显著影响。

“国内大企业还是会持续部署国产芯片以跟上算力需求，自产自研这个政策在短时间内也不会有太大的改变。”苏廉节依旧看好国产AI芯片在国内市场的前景。

从“做出来”到“卖出去”

财报显示，摩尔线程2025年经营活动产生的现金流量净额为-29.56亿元，相较于2024年的-19.55亿元大幅减少；2026年第一季度，该公司经营活动产生的现金流量净额为-14.87亿元，上年同期为-7.52亿元。

有关经营性现金流缺口规模扩大的原因，摩尔线程公司董事会秘书及财务负责人薛岩松在上证路演平台回复投资者提问时称：一是公司根据市场情况积极进行战略备货，为应对下游市场需求的持续增长，并保障供应链稳定，结合市场情况，公司主动增加了原材料采购和战略备货力度；二是持续保持高强度研发投入，短期内也会对经营性现金流形成压力。

财报还显示，截至2025年年末，摩尔线程的存货账面价值为13.32亿元，较期初增长105.87%，主要为应对在手订单与客户需求提前备货，包括芯片、板卡、服务器部件等；同期，公司预付款项17.82亿元，同比增长214.24%，主要为向上游供应商预付晶圆、封测等产能款项，以保障在手订单按时交付。

也就是说，为了保供应、保交付、保项目落地，摩尔线程把大量资金用在备货（库存增加）、预付款和产能里。这也使得该公司虽属无晶圆厂模式，但业务模式日趋重资产化。

如果说“做出来”和“卖出去”是国产GPU公司商业化的两个阶段，那么在第一阶段，竞争焦点集中于技术突破与产品可用性；第二阶段，竞争维度则进一步扩展，不仅仅局限于芯片自身性能，更涵盖了交付能力、客户结构、现金流质量、供应链稳定性，以及能否将高额投入有效转化为可持续的运营闭环。

在张建中看来，公司当前仍处于高强度的研发投入期，主要目标是为了实现快速产品迭代、增强技术实力和市场竞争力。持续的研发是实现长期盈利的关键。

“为了平衡这两者，公司采取了以下策略：明确研发方向，聚焦于高市场潜力和高门槛技术领域，以确保研发成果能够迅速转化为商业价值，并构建长期竞争力；持续优化资源配置，提高研发效率，确保每个项目的投入都能产生最大效益；构建了云边端产品线，充分发挥全功能GPU的能力，通过多元化产品线和服务，加速实现收入增长。”张建中在上证路演平台如此回复投资者。

针对记者在上证路演平台“相比国际巨头和国内同行，公司在‘花港’架构、软件生态和大规模集群调度上的核心差异化优势是什么”的提问，张建中表示：“公司新一代全功能GPU架构‘花港’在算力密度上提升50%，能效比实现了10倍的跨越式增长。‘花港’架构不仅支持FP4到FP64的全精度计算，更在单芯片上融合了AI计算加速、图形渲染与物理仿真等功能。这使得公司GPU产品能够满足从大模型训练到具身智能、数字孪生等复合型场景的复杂需求。”

张建中还指出，软件是GPU的护城河。“摩尔线程MUSA架构的核心优势在于其高度兼容CUDA生态，深度兼容PyTorch、vLLM、SGLang、Triton、TileLang等主流生态，通过MUSAC++、Triton-MUSA、TileLang-MUSA等抽象层实现新算子较低成本迁移，确保前沿模型发布当日完成极速适配。”在集群调度上，公司的夸娥万卡智算集群已实现大规模部署并上线运行，通过公司自研的MT-Link高速互联技术，“花港”架构支持十万卡以上的超大规模智算集群扩展，并能保持极高的算力利用率（MFU）；通过“端云结合”的解决方案，公司直接赋能包括具身智能在内的前沿产业的仿真训练。

“面对快速变化的外部环境，摩尔线程坚持技术与生态双轮驱动战略。”张建中表示，通过MUSA全栈软件生态与硬件的深度协同，解决国产GPU生态黏性难题，并积极推动开发者社区发展，目前已汇聚45万名开发者。