大模型和AI芯片联合突围的中国叙事

文/李玉洋

2025年年初，中国AI大模型初创公司深度求索因发布开源推理大模型DeepSeek-R1而在全世界一鸣惊人，犹如一颗深水炸弹，打破了由OpenAI、Anthropic、谷歌等国外豪门所把持的全球顶级大模型俱乐部的平静。此后的一整年，DeepSeek的每一个动作，都成为全球大模型市场的焦点。

不过，对于中国大模型生态圈而言，深度求索的“鲇鱼效应”更体现在其让FP8（8位浮点数）这种低精度格式获得全行业认可，一方面搅动了顶级大模型俱乐部的世界格局，另一方面也力挺了国内AI芯片的发展。

“2025年国内AI芯片行业呈政策驱动、需求爆发、国产替代深化的态势，市场规模与国产化率双升。”复盘2025年，刚刚在香港联合交易所通过聆讯的国内GPU厂商天数智芯相关人士对《中国经营报》记者如此表示。

需要指出的是，2025年不断更新的国内大模型成为国产AI芯片发展的一股助力，尤其是开源模型。“以DeepSeek为代表的国产大模型的优秀表现加快了AI技术在行业的落地速度和广度，在算力国产化大背景下，国内GPU厂商的市场空间被极大拓展。”2025年12月5日已在科创板正式上市的摩尔线程（688795.SH）方面对记者表示，“2025年国内大模型的发展对国产AI芯片行业来说是强催化剂，影响显著且持续。”

天数智芯上述人士也有类似感受：“在生态端，模型厂商与芯片企业联合优化成常态，开源模型降低适配门槛，加速‘芯片—模型—应用’闭环生态形成。”

诚然，虽然国内AI芯片企业与国际龙头之间仍有差距，但回望国产AI芯片叙事脉络，2025年一定会是个重要节点，因为这一年国产GPU行业已经迈入商业化与规模化的关键时期。

2025“年度字词”的揭晓，凸显了中国AI芯片和大模型在2025年的韧性突破。 视觉中国/图

降低算力需求 通过软件提升效能

芯片，是AI所需算力的硬件基础，而开源模型使得AI的部署和使用成本大大降低。在AI推理方面，国内各类AI芯片百花齐放。

据不完全统计，包括华为昇腾、昆仑芯、海光信息、沐曦股份（688802.SH）、摩尔线程、天数智芯等多家AI芯片厂商相继宣布了对DeepSeek模型的快速部署和训练，使得国产AI芯片能够在推理任务中与英伟达GPU竞争，甚至在某些场景中表现更好。

“DeepSeek的成功表明，通过模型压缩、稀疏计算、混合精度训练等技术手段降低算力需求，可以在一定程度上弥补硬件性能不足。”摩尔线程方面表示，这为国内芯片提供了软硬件协同设计的新思路，“证明了在硬件性能短期内难以赶超的情况下，通过软件层面的创新仍可提升整体计算效能”。

同时，深度求索还触发了下一代AI芯片的“底层代码”——FP8低精度格式，该格式对计算精度要求相对较低，一定程度上降低了对晶体管密度的依赖，意味着不再那么迫切需求世界上最先进的芯片制造工艺。

而深度求索在成功训练出世界首个使用FP8精度的开源大模型DeepSeek-V3后，又在DeepSeek-V3.1中使用了UE8M0 FP8 Scale的参数精度。

摩尔线程方面指出，DeepSeek在混合精度训练方面的成功，展示了低精度计算在AI训练中的潜力，“国内芯片厂商可以借鉴这种模式，优化芯片的计算单元，支持更灵活的精度配置”。

根据公开资料整理，截至2025年12月，国内已有包括摩尔线程、燧原科技、沐曦股份、壁仞科技、天数智芯、砺算科技、海光信息（688041.SH）、寒武纪（688256.SH）、中昊芯英、芯原股份（688521.SH）、昆仑芯等10余家厂商推出支持FP8精度的AI芯片，其中摩尔线程、燧原科技、沐曦股份、寒武纪处于领先地位，均已实现原生FP8支持并完成与DeepSeek这些主流大模型的适配。

专注于推理应用 平替4090没问题

随着国产FP8标准的推广，国产AI芯片在大模型部署方面正逐步缩小与英伟达的差距。

“需求端，基座大模型与行业定制化模型催生高端训推芯片及高性价比推理芯片需求，扩大国产芯片替代空间。技术端，UE8M0 FP8量化这些模型创新倒逼芯片适配，推动硬件架构与算子库升级。”天数智芯方面表示。

“2025年，我们一直在加紧对国产AI芯片的配合和小规模测试。这一过程中最直观的感受是，国产AI芯片的迭代速度在不断攀升，无论是适配性、兼容性还是稳定性，都有明显的提升。我们发现，目前很多过去常用的国外通用芯片的型号，已经完全能用国产GPU进行平替。”国内一家音乐大模型初创公司表示。

该公司还表示，经过相关测试，在推理上，国产GPU平替英伟达4090是没问题的。同时，在政策支持下，使用国内AI芯片的成本也有一定优势。

根据摩根士丹利2025年12月在亚洲市场的实地调研，中国AI加速器市场正出现新的路径分化。部分中国本土AI芯片设计公司开始主动调整技术路线，通过开发相对中端规格的产品，以专注于推理应用。

该调研还指出，中国AI计算的侧重点正在从训练端加速向推理端迁移。

而内存供应链信息同样印证了推理需求的兴起。摩根士丹利指出，新一代国产AI推理芯片正加速采用LPDDR作为主要配套存储方案，以替代供应紧张且成本高昂的HBM。

如何从“可用”到“好用”

知名股权研究机构Bernstein Research在2025年12月发布的一份报告中预测，华为与英伟达在中国AI芯片市场各占40%份额，形成双雄并立格局。到2026年，华为份额将进一步上涨至50%，英伟达萎缩至8%，而寒武纪将取得9%的市场份额。

值得注意的是，2025年8月底，作为A股较为稀缺的AI芯片标的，寒武纪股价一度最高达1595.88元/股，一举超过贵州茅台，成为A股股价最高的个股。

而GPU作为AI芯片的主要技术路线，国内GPU行业在2025年12月迎来上市热潮：12月5日，摩尔线程科创板上市，创2025年科创板募资规模最大的IPO（募资规模约80亿元）；12月17日，沐曦股份登陆科创板，首日股价涨幅692.95%，单签盈利36.26万元刷新A股打新纪录。

除科创板外，港交所则成为国产GPU厂商抢滩资本市场的另一个选择。近期，壁仞科技、天数智芯都通过港交所上市聆讯，“港股GPU第一股”也就要呼之欲出。

然而，这些体现的是国产GPU行业在资本层面的繁荣。业界多认为，2025年国产AI芯片已实现“可用”，但性能、能效仍落后国际竞品。

“如果要说（国产AI芯片）还有哪部分有所欠缺，目前比较明确的需求是希望有更大显存的版本，来满足我们模型训练的需求。”前述国内音乐大模型初创公司表示。

那么，国产AI芯片从“可用”到“好用”的关键突破点是什么？

对此，摩尔线程方面表示，接下来的市场竞争焦点正从“能用”全面转向“好用”，决胜于综合产品力与生态生命力。

首先在产品层面，竞争超越了单纯的算力指标，进入了“综合效率”的比拼。“客户不仅要求高算力，更追求极致的能效比、更高的算力利用率、更优的软件开发体验，其根本目的是降低总体拥有成本（TCO）并获取明确的投资回报。因此，提供高性价比的全栈解决方案，而非单一硬件，将成为关键。”摩尔线程方面表示。

其次在生态层面，挑战更为根本。“在算力供应快速向国产化转移的过程中，国产GPU面临两大考验：一是能否无缝兼容现有的主流软件与开发习惯，确保平滑迁移，这是入场的基础；二是在此之上，能否展现出持续的自主创新能力，通过提供独特的价值吸引开发者，从‘生态兼容者’逐步成长为‘生态定义者’。这要求厂商必须具备强大的战略定力，坚持长期主义，持续投入资源耐心打磨生态，做难而正确的事。”摩尔线程方面如此指出。

还是在2025年12月，美国政府已批准英伟达向中国出售H200芯片，在国内的AI算力生态中激起涟漪。近日还有市场消息称，英伟达已告知中国客户，计划于2026年2月中旬向他们交付H200。

面对英伟达仍占据全球超70%的市场份额，且在高端训练芯片领域仍无替代方案的现状，2026年若海外高端芯片供应变化，国产AI芯片的“换道超车”是继续在推理端深耕，还是必须在训练端实现技术突破？

对此，天数智芯表示，2026年国产AI芯片“换道超车”，推理端要聚焦爆发场景，推出高性价比、低功耗芯片，构建“芯片—模型—解决方案”生态闭环；训练端则需渐进式突破，先攻克中大规模训练，再联合模型厂商优化框架，同时加速核心IP自主与供应链自主。

“关键在于平衡短期市场份额与长期竞争力，推理端是快速扩大份额的主阵地，训练端是构建核心竞争力的关键，二者并行不可或缺。”天数智芯表示。

2025年算力股翻倍频出 2026年或存估值消化压力

文/顾梦轩

2025年A股市场，AI算力可谓“最耀眼的星”。Wind（万得）数据显示，截至12月25日，万得AI算力指数年内上涨44.53%。

2025年，资本市场对AI算力板块的追捧，AI算力领域翻倍股频出，市场热点组合“易中天”〔由新易盛（300502.SZ）、中际旭创（300308.SZ）、天孚通信（300394.SZ）三家公司组成〕、“纪连海”〔由寒武纪（688256.SH）、工业富联（601138.SH）、海光信息（688041.SH）三家公司组成〕应运而生。此外，开普云（688228.SH）和芯原股份（688521.SH）也有亮眼表现，均在年内翻倍。

Wind数据显示，截至12月25日，AI算力板块54只概念股中，有8只翻倍股，其中有2只“四倍股”。股价涨幅最高者是新易盛，年内股价涨幅为444.84%；股价涨幅排名第二的是中际旭创，年内股价涨幅为420.63%；第三名是开普云，年内股价涨幅为378.68%。

《中国经营报》记者注意到，除海光信息外，“易中天”和“纪连海”中剩余5只个股均在2025年翻倍股之列。

国泰基金有关人士在接受记者采访时指出，从供需两端与产业环境看，2025年AI算力赛道表现强势由“技术突破+需求爆发+业绩落地”三重逻辑共振驱动，具体可拆解为三个维度：从供给端来看，DeepSeek以美国十分之一的算力实现等效大模型性能，打破了海外对中国AI技术“存在代差”的认知，不仅引发国内技术突破热潮，也倒逼海外大厂加速升级，产品不断迭代，持续为赛道注入热度；从需求端来看，AI大模型参数量从万亿级向十万亿级突破，催生“万卡集群”需求，直接拉动算力产业链上游爆发；从业绩端来看，因上游订单明确，头部企业实现订单、营收双增长。

需求持续爆发

2025年，全球大模型训练和推理对算力的需求持续爆发。英伟达、谷歌、Meta等科技巨头加速部署新一代GPU/TPU（图形处理单元/张量处理单元）集群，推动数据中心内部互联带宽需求指数级增长。在此背景下，800G光模块（CPO）进入大规模放量阶段，1.6T产品也开始批量交付，成为支撑AI集群高速通信的关键硬件。

数据显示，2025年，全球AI服务器市场规模同比增长约27%，800G光模块出货量突破1800万—2000万只，1.6T产品进入商用元年，直接拉动上游硬件产业链业绩爆发。同时，政策层面“新质生产力”“东数西算”等战略持续加码，为算力基础设施建设提供长期支撑。

在此背景下，“易中天”组合应运而生，该组合三家上市公司均属光模块赛道。该组合在2025年A股市场表现抢眼。Wind数据显示，截至2025年12月25日，新易盛和中际旭创年内股价涨幅均超400%，天孚通信涨超200%。

跟“易中天”组合相区别，“纪连海”组合的三家上市公司分属于AI领域不同细分领域。其中寒武纪是AI芯片专业厂商，工业富联属于工业互联网领域企业，海光信息是数字芯片企业。“纪连海”组合的三家公司在2025年股价表现亦强势，Wind数据显示，截至2025年12月25日，寒武纪股价年内上涨100.52%，工业富联上涨216.41%，海光信息上涨47.35%。

爆发背后：赛道热度+企业实力

受访人士皆认为，“易中天”及“纪连海”等翻倍股的爆发，既有赛道因素也有企业自身因素。

对于“易中天”组合的爆发，南开大学金融发展研究院院长田利辉向记者指出，“易中天”组合的爆发，是赛道正确性与企业业绩优异表现完美结合的典范。光模块领域是解决AI算力“带宽瓶颈”的关键。随着AI模型对数据传输速率的要求呈指数级增长，光模块正经历从400G到800G再到1.6T的快速迭代，带来了显著的“量价齐升”效应。这三家公司不仅是行业龙头，其800G高端产品已开始大规模出货，业绩高增长得到验证，是“基本面驱动”的典型代表。

济安研究院研究员万力向记者指出，赛道决定了“有没有风”，公司决定了“飞得多高、能飞多久”。对于“易中天”和“纪连海”的区别，万力向记者指出，“易中天”更像“全球AI资本开支的映射”，受海外云厂商投入节奏、互联升级路径影响更大；而“纪连海”更像“国内算力供给侧的映射”，更依赖国内生态成熟度、国产替代推进速度，以及政策与供应链环境的变化。

记者注意到，不论是“易中天”还是“纪连海”，这六家企业在2025年的业绩表现均亮眼。根据财报数据，以新易盛为例，公司2025年前三季度营业收入165.05亿元，同比增长221.7%，净利润63.27亿元，同比增长284.37%。

剩余五家公司中，有四家公司2025年前三季度营业收入和净利润增幅均可观，只有海光信息净利润增速稍慢，但也保持盈利且实现增长。财报数据显示，2025年前三季度，海光信息净利润19.61亿元，同比增长28.56%。

不过也有反例，记者注意到，如果从2025年翻倍股的业绩表现来看，年内股价涨幅378.68%的开普云和涨幅143.47%的芯原股份业绩都不甚理想。这两家公司2025年前三季度均处于亏损状态，不过相对于2024年净利润均有不同程度增长。其中，开普云增长70.51%，芯原股份增长12.42%。

对此，张思远指出，芯原股份体现了“订单先行、业绩滞后”的特征。尽管2025年前三季度净利润尚未显著释放，但其AI ASIC业务新签订单创历史新高，单季收入同比翻倍。

此外，板块情绪与资金偏好也起到重要作用，苏商银行特约研究员张思远指出，AI算力作为2025年资本市场核心主线，资金对具备“AI备案”、“数据要素”和“东数西算”等概念标签的标的关注度较高，开普云的“数字政府+AI大模型”、芯原股份的“半导体自主可控”属性，使其在短期业绩“真空期”内仍能获得资金青睐。

2026年警惕估值消化压力

AI行情自2023年爆发以来已经持续三年，经过三年上涨，AI算力板块在2026年将会有怎样的表现？是否有回调风险？

路博迈基金方面表示，AI这轮周期具备持续性。与过去“主题驱动”的板块轮动不同，当前AI投资建立在可验证的需求增长与清晰的产业瓶颈之上，这为其持续上涨提供了基本面支撑。

路博迈基金方面主要看好以下领域：先进半导体、存储芯片尤其是HBM（高带宽内存）、高速网络以及电源与散热。核心逻辑在于，AI基础设施建设是多年期、多层次的资本开支周期，不同于过去单一技术路线的“一波流”行情。只要AI应用端持续验证ROI（投资回报率）、Token（令牌/词元）使用量保持指数级增长，上游硬件需求就具备可持续性。

金鹰基金方面指出，AI算力赛道或仍旧是科技领域最为坚定的方向。以中国光模块行业龙头公司为代表的海外算力景气度依旧，“无论是光模块、PCB（印刷电路板）还是液冷，2026年甚至2027年可能都将持续高增长，或可逢低配置。”该基金人士表示。

在细分领域上，万力认为2026年可能出现两个变化：一是从“拼训练规模”逐步过渡到更重视“推理与落地效率”，因此围绕数据中心内部互联、服务器交付、能耗与散热等“效率环节”的重要性会更突出；二是资金会更关注ROI与现金流质量，算力链条内部可能更“看业绩说话”，而不是简单按题材整体抬估值。

风险方面，万力表示，回调风险客观存在，而且并不意味着产业趋势结束，更多是估值与兑现节奏的再平衡。当市场把未来几年增长提前计价后，只要出现资本开支节奏波动、交付不及预期、价格竞争提前或外部环境变化，股价就可能先于基本面调整。

张思远指出，2026年，AI算力行业将进入“高景气与高波动并存”阶段。硬件端（光模块、存储）仍是确定性最高的方向，但需警惕估值消化压力。投资者应理性看待短期波动，聚焦具备技术壁垒与盈利韧性的企业，在产业趋势与估值安全边际间寻找平衡。

对投资者，万力建议抓住三条“朴素但有效”的原则：第一，只跟踪能验证的东西：订单可见度、交付进展、毛利与费用趋势，而不是只看概念热度。第二，警惕“高预期下的小失误”：越是高估值阶段，市场越容易因为小幅不达预期而重新定价。第三，重视结构性风险：客户集中度、海外依赖度、技术路线切换带来的产品淘汰风险——这些往往决定回撤深度。