大飞机报
China Commercial Aircraft News
2026/5/6 星期三
责任编辑丨郑小芳 投稿邮箱:zhengxiaofang@comac.cc

翻墙：DeepSeek-V4的解题思路

□ 记者 郑小芳

工程技术跟科学研究的不同之处或许在于，当正门不通时，我们可以“翻墙”进去。

——题记

近日，在经历了爆火后的沉寂，深度求索全新系列模型DeepSeek-V4预览版正式上线并同步开源，在全球AI圈引起广泛关注。

此次DeepSeek-V4之所以成为行业焦点，在于它实现了技术与产业层面的双重突破。技术层面，V4标配了过去只在海外闭源模型高端版本中才具备的100万Token（词元）超长上下文能力，相当于《三体》三部曲总字数，可一次性承载海量文档并完成全局推理，彻底告别长文本处理的切片割裂痛点。产业层面，V4实现了与华为昇腾、寒武纪等国产算力芯片的深度耦合，为国产AI生态的自主构建按下了加速键。

在DeepSeek-V4发布前，国产芯片一直面临一个尴尬的境地：用户少，生态建不起来。这一困境的根源在于，传统Transformer架构在处理长上下文时，显存占用随文本长度呈平方级增长。国产芯片受制于制程工艺，虽在算力峰值上奋力追赶，但显存带宽与容量方面仍与英伟达H100/Rubin存在代际差距。在传统架构下用国产芯片处理长文本几乎是“不可能的任务”。

但DeepSeek-V4给出了一条全新的解题思路：既然硬件条件暂时追不上，那就从架构层面重构“生产关系”。

V4引入了CSA（压缩稀疏注意力）与HCA（高度压缩注意力）相结合的混合架构。通俗地讲，它不再逐字逐句比对百万字文档，而是学会了“抓重点”和“做摘要”。通过对token进行智能压缩，V4大幅降低了对计算和显存的需求，在提升性能的同时实现了资源高效利用，破解了大模型“高性能必高成本”的行业难题。在百万Token上下文设定下，V4-Pro的单token推理算力仅为上一代V3.2的27%，KV缓存只需10%；更经济的V4-Flash版本这两项指标分别仅为10%和7%。

如此一来，原本需要80GB显存才能胜任的长上下文推理任务，现在可能只需8GB显存就能运行。英伟达的高带宽显存稀缺性，在这套新架构面前被大幅削弱，国产芯片的显存短板便不再成为瓶颈。

此次V4没有按行业惯例给予英伟达早期测试权限，而是将提前适配的机会独家开放给华为昇腾和寒武纪，目标直指从CUDA生态到华为CANN框架的整体迁移。可以预见，当一款开源模型的架构创新能够系统性降低显存门槛，跑通国产芯片并实现推理成本大幅下降，英伟达凭借显存优势筑起的护城河便不再是铁板一块。

这正是DeepSeek“曲线救国”的深层逻辑：不执着于单卡性能上的硬碰硬，而是凭借系统级优化、软硬深度协同和架构创新，重新定义了竞争维度。

豆包收费引发热议

AI正在加速变现

5月4日，字节跳动旗下的AI助手豆包在苹果应用商店悄然挂出一则服务声明——标准版68元/月，加强版200元/月，专业版500元/月，这个坐拥3.45亿月活的超级AI应用，正计划推开付费订阅的大门。显然，国内大模型即将告别那个“靠免费换规模”的粗放时代，硬生生闯入需要“真金白银”说话的下半场。

□ 沈毅斌

当每个用户都是一笔“算力账单”

豆包收费背后，首先是一笔经济账。

截至今年3月，豆包大模型日均Token消耗量从2024年5月发布时的1200亿，飙升至120万亿，增长约1000倍。火山引擎披露，累计Token调用量超一万亿的企业客户从100家增长到140家。

Token时代，新的经济规律已初现端倪：边际成本效应消失，用户规模越大，算力成本越高。

一份流传于中文IT技术社区（CSDN）的成本拆解模型显示，豆包单次推理成本中，硬件折旧占58%、电力占29%。简言之，每一次用户提问、每一轮对话生成，都直接对应实打实的算力消耗，成本并没有随着规模扩大而下降。

与此同时，收入端还没有打通。过去，中国互联网的变现逻辑是：用户越多，广告价值和商业变现空间越大。AI时代，这条路刚刚开始试水。今年5月，OpenAI推出自助式广告平台，目标收入是2026年25亿美元、2030年1000亿美元。但质疑声甚嚣尘上，毕竟人们因AI而放弃搜索引擎，原本是为了得到一个有效结果，而不是带着“金钱味道”的推荐广告。

豆包也在探索广告模式，但至少在目前，它还很难彻底跨过GEO（生成式引擎优化）和“大模型投毒”的界限。QuestMobile数据显示，截至3月，豆包月活用户接近3.45亿，日活用户约1.4亿，月人均使用次数达54.8次。

一面是居高不下的成本，一面是还未成熟的商业模式，即便是字节跳动这样公认的“家有余粮公司”，也越来越烧不起Token了。有媒体报道，2025年字节跳动总营收预计接近2000亿美元，但净利润同比下滑超70%，核心原因正是对AI业务的投入。

尽管抖音集团副总裁李亮在微博发文澄清，实际经营利润率降幅“远没有那么大”，但他也承认，抖音电商增速放缓和新兴业务相关投入增大，导致下半年经营利润率有小幅下滑。有人戏称，“抖音赚的钱，正在被豆包烧光”。

国内AI大模型是否跟进？

豆包收费消息一出，所有人的目光都投向了腾讯元宝、阿里千问等其他面向C端的大模型。它们跟还是不跟？截至目前，腾讯元宝和阿里千问并未传出要收费的消息，甚至有业内人士猜测，元宝和千问很可能会借此机会，抢夺豆包用户。

3月以来，腾讯云连续两次涨价：3月11日大模型API输入价格单轮暴涨463%，4月9日又宣布5月9日起AI算力相关产品上调5%；阿里云也宣布从4月18日起，平头哥真武810E等算力卡上调5%～34%，CPFS（智算版）上调30%，5月15日起，阿里云百炼平台部分模型单元服务涨幅从2%～7%不等。

目前来看，两家友商的节奏是B端先动、C端观望，以生态吸引更多用户。

只是，元宝和千问能承接多少豆包流失的用户？能坚持免费策略多久？同样要看能否找到用户数量和Token成本的平衡点。

豆包能“物有所值”吗？

毕竟，全球大模型厂商都玩不起打折竞赛了。

进入2026年后，全球AI市场涨声一片。随着AI逐渐成为有效生产力工具，B端用户的付费意愿逐渐提升。

Claude Code的官网显示，在企业级部署中，平均成本约为每位开发人员每天13美元、每月150～250美元，90%的用户每天成本低于30美元，按上限测算，年成本约为1.1万美元，而数据显示，一个硅谷软件工程师的平均年薪约为14万美元。

智谱AI年内三度提价，API调用定价累计上调83%，而Token消耗量反而增长400%。智谱CEO张鹏表示，智能上限决定定价权，Token消耗规模决定价值体量。大模型正在从“谁便宜买谁”转向“谁好用谁值钱”。

对于已经积累足够用户黏性和产品口碑的头部玩家来说，率先尝试付费分层，其实是主动探索可持续的商业闭环。一旦付费模式跑通，就能给整个赛道吃下一颗“定心丸”，也让市场对AI应用的商业化有着更清晰的预期。

豆包的付费探索，正是这一转向在C端的具象化体现。与全球付费体系相比，豆包的定价不算出格。

目前ChatGPT已建立完整的四层定价体系：Go版约8美元/月、Plus版20美元/月、Pro版100美元/月及200美元/月。其中Pro档位直接对标Anthropic Claude Max，面向重度编程用户。豆包标准版的68元人民币（约合9.9美元）与ChatGPT Go/Plus（8～20美元）处于相近区间，专业版500元（约73美元）则大致位于Pro版和中国市场消费水平交叉点上。

凯基证券的分析报告认为，豆包付费标志着中国大模型行业从“野蛮增长”转向“价值兑现”，竞争重心将从“模型参数”转向“商业ROI”。

但豆包需要解决的问题是，能否提供与其收费水平匹配的能力，让具有付费能力和意愿的用户觉得“物有所值”。

豆包能走通这条“基础功能免费+高算力功能付费”道路吗？它所保留的免费功能，足以让它留住大部分活跃用户吗？分层订阅能满足高阶用户需求吗？

在这个新开启的AI时代，人们只能等待答案自己浮现。

据《IT时报》

人物

13年的『光速突围』

新晋『股王』源杰科技创始人张欣刚

□ 牧龙

最近，A股市场上演了一场惊心动魄的“股王”争夺战。短短两周多的时间里，两家科技公司——源杰科技与寒武纪，先后超越贵州茅台，登顶“股王”宝座。截至5月6日收盘，寒武纪以1825元/股排在第一位，源杰科技以1531.01元/股排在第二位，贵州茅台则以1375元/股位列第三。而去年同期（2025年5月7日收盘），源杰科技的股价不过132.17元/股，短短一年涨幅超1000%。

受益于全球AI产业爆发、算力需求暴涨，光模块、GPU、存储芯片、光芯片、云计算等领域备受关注，作为光芯片赛道龙头企业的源杰科技顺势而起，成为了AI风口上的“股王”，其创始人张欣刚也因此被推到了聚光灯下。

► 选择最难的路

2013年的一天，陕西咸阳某酒店的停车场内，一位日本供应商望着眼前一辆老旧的奥拓，眉头紧锁，迟迟不敢上车。

他一度认为车里的人是骗子。毕竟，对方口口声声说要造高端光芯片，却开着这样一辆破旧的小车带他参观所谓的“先进工厂”。这巨大的反差，让他实在无法把眼前的场景与芯片研发联系起来。车里的人，就是张欣刚。

那位日本供应商绝不会想到，眼前这个人会在之后十几年的时间里，将企业打造成中国光芯片领域的龙头，在全球光芯片市场挤占一席之地。

1970年出生的张欣刚，本科毕业于清华大学材料系，后赴美深造，先后取得南加州大学材料科学硕士、博士学位。2001年，博士毕业的张欣刚进入光通信公司Luminent工作，他从基层研发员起步，逐步晋升至研发经理，深度参与光芯片从研发设计到制造封测的全流程工作。2007年，Luminent与另一家光通信设备商Fiberxon整合，成立索尔思光电，张欣刚出任研发总监。

索尔思光电在全球光模块行业中占据重要地位，不少业内大佬都是从索尔思走出来的，张欣刚就是其中之一。

张欣刚于2010年回国创业，在北京成立了一家光电技术公司，但因合伙人挪用资金，导致公司停摆，首次创业失败。但张欣刚没有止步。2013年，他决定再次创业。当时，陕西咸阳正好推出一些科创扶持政策，张欣刚便将落脚点定在咸阳，成立陕西源杰半导体技术有限公司。

那个时候，国内高端光芯片几乎被海外厂商全盘垄断。为了打破这一局面，张欣刚选择了一条最难的路——IDM模式，即从芯片设计、外延生长、晶圆制造到封装测试，每一步都自己掌握。这种模式前期投入大、回报周期长，但张欣刚坚信，只有把全流程攥在自己手里，才能真正突破卡脖子限制，做出性能稳定、可靠的高端芯片。

► 布局AI算力赛道

创业初期，一切都很艰难。张欣刚面临的是国内产业基础薄弱、国际技术封锁、公司资金匮乏等一系列问题。他一方面筹措资金，另一方面扎进实验室，带领团队死磕技术、反复打磨工艺。

从2013年到2016年，源杰科技逐步推出2.5G与10G DFB激光器芯片，并实现小批量出货，为企业的后续发展积累了宝贵的市场经验和技术基础。

真正的拐点出现在2018年。这一年，源杰科技自主研发的25G DFB激光器芯片成功通过客户验证，一举打破国外垄断，成为国内率先攻克该技术并实现规模化商用的企业。凭借这一突破，源杰科技顺利切入5G建设赛道，将华为、中兴等通信行业头部企业纳入客户名单。

此后，张欣刚带领团队，乘势推出面向5G通信与数据中心的25G DFB CWDM/LWDM系列产品，进一步丰富高端产品矩阵。

到2020年，源杰科技已在磷化铟光芯片领域实现国内收入第一，10G、25G光芯片出货量双双登顶行业榜首。

2022年，在源杰科技成立的第10个年头，张欣刚带领源杰科技登陆上交所科创板，首日市值60.39亿元。源杰科技就此步入发展的快车道。

在资本的加持下，张欣刚大力投入研发。他颇具前瞻性地布局AI算力赛道，将战略重心转向数据中心高速光芯片。这一决策在2025年开始显现成果。

2025年，随着AI产业的爆发，数据中心对高速光芯片的需求量呈指数级增长。

与传统电芯片相比，光芯片传输速率更快、功耗更低、抗干扰能力更强，更重要的是，光芯片对光刻工艺依赖度较低，可利用国内较成熟的原材料和设备实现生产，很大程度上绕开了高端光刻机的“卡脖子”限制。这也是市场普遍看好源杰科技、股价大幅上涨的重要原因。

从开奥拓接待客户的咸阳小厂，到市值超1300亿元的硬科技巨头，张欣刚和团队用13年的坚持，实现了中国芯片的“光速突围”。

资讯

6G技术试验频率获工信部批复

本报讯 为进一步推动我国6G技术研发、标准研制与产业化进程，工业和信息化部近日向IMT-2030（6G）推进组批复6GHz频段6G试验频率使用许可，支持其在部分地区开展6G技术试验，面向国际电信联盟确定的6G典型场景与关键性能指标，开展技术研发攻关和测试验证。

上海发布全球科技伙伴计划

本报讯 上海市政府近日印发《上海全球科技伙伴计划》，提出国际联合研究、人才交流访问、国际科技组织引育、开放科学、产业创新国际化、创新生态优化六大行动，支持海外科研机构、企业、组织和科技创新创业人才等成为上海科技伙伴，建设具有全球竞争力的开放创新生态，提升上海科技创新的全球影响力和引领力，助力共建全球科技共同体。

英伟达高端AI芯片在华市占归零

本报讯 近日，英伟达首席执行官黄仁勋表示，该公司在中国AI加速器市场的份额目前已降至零。据统计，2025年中国市场的AI加速卡出货量约为410.6万块，其中英伟达以55%的市场份额排名第一，这与该公司2021年在华95%的市场份额相比，已大幅下滑。

本报地址：上海市浦东新区世博大道1919号 邮编：200126 电话：021-20887106 传真：021-20887151 每逢周三出版 本期售价：1元 印刷：商务印书馆上海印刷有限公司