

解码词元经济

词元经济“中间层”崛起 AI基础设施进入资本化窗口期

中经记者 许璐 李晖 北京报道

国家数据局数据显示,我国日均词元(Token)调用量由2024年年初的约1000亿增至2025年年底的100万亿,2026年3月进一步突破140万亿,两年增长超过1000倍。

在此背景下,围绕算力调度、推理优化、模型部署和Token供给形成的AI基础设施“中间层”正在进入资本市场视野。2026年以来,基流科技、硅基流动先后向港交所递交上市申请,

“中间层”进入技术红利释放期

中间层的本质作用,是把算力产业从“卖地皮”变成“卖自来水”。

根据沙丘智库发布《2025年中国AI基础设施(AI Infra)市场跟踪报告》,将AI基础设施分为底层的物理基础设施,中间的AI平台和上层的AI服务三层架构。底层包括AI芯片、服务器、存储、网络和数据中心;上层包括基础模型、行业模型、智能体及各类应用;位于二者之间的中间层,则通过算力集群调度、异构芯片适配、推理加速、模型部署、API管理和Token计量,将硬件资源和模型能力转化为可调用的服务。

对于目前国内AI基础设施产业所处的发展阶段,田丰认为,刚完成“从0到1”的需求验证,正进入“从1到100”的结构性扩容期。中间层的本质作用,是把算力产业从“卖地皮”变成“卖自来水”,解决的核心矛盾,是把“芯片的物理算力”翻译成“应用可消费的Token”。

“这一过程类似于将原油加工为不同规格的成品油。中间层企业通过算子重构和模型压缩,提高同等硬件条件下的有效Token产出数量。推理效率的提升空间远大于训练效率——中间层的技术红利才刚刚释放。”田丰说。

目前,AI基础设施中间层形成了由算力集群、系统软件、异构

清程极智、无问芯穹、趋境科技等企业连续完成融资。

快思慢想研究院院长、特邀评论员田丰在接受《中国经营报》记者采访时预计,未来Token产业的健康格局将是云厂商做Token全链路整合,运营商依托分发触点和计费系统;第三方独立中间层企业提供跨平台的调度、适配和效率优化服务,做独立炼化工厂和技术效率提供者;地方国资平台整合本地智算中心资源,对接区域产业需求。

调度、推理平台和词元服务组成的多层次业务结构。靠近底层的是AI算力集群产品与运营。例如基流科技,其招股书显示,公司提供涵盖计算、网络、存储和软件的全栈AI算力集群产品,并提供运维、技术支持和算力服务。

位于中间环节的企业主要提供系统软件和异构算力调度能力。例如,无问芯穹围绕多元异构算力、软硬协同和MaaS平台开展业务。清程极智通过智能算力系统软件连接底层硬件算力与上层模型训练、推理及应用需求。

靠近应用侧是模型推理服务和Token供给平台。以硅基流动为例,公司主要通过无服务器Token服务、专属实例和本地部署向客户提供模型推理能力;趋境科技则以自研ATaaS平台为核心,连接底层资源、模型推理系统和企业业务场景。

模式类似的还有趋境科技。趋境科技方面向记者表示,公司自营模式主要面向模型厂商、互联网平台、AI应用公司和大型企业客户自行生产、销售AIToken;共运营模式面向上市公司、国央企及智算中心等提供Token工厂建设和联合运营等服务,应用场景包括AICoding、智能客服、办公套件和企业Agent。

资本加速入场

资本目前主要观察企业能否获得稳定且成本可控的算力供给,调度和推理优化能否提高算力利用效率、降低单位Token成本。

2026年以来,相关企业的融资进程由一级市场走向IPO申请阶段。

4月29日,基流科技向港交所递交上市申请;6月30日,硅基流动向港交所递交上市申请。两家公司分别从AI算力集群产品及运营、模型推理和词元服务切入中间层市场。

硅基流动成立于2023年8月。招股书显示,公司上市申请前已经完成从天使轮至B+轮的七轮融资,最新一轮融资对应投后估值约77.4亿元。2025年,公司实现收入5533.0万元,同比增长653.2%;年度亏损为3.45亿元。同期,公司算力资源成本为5962.7万元,占收入成本的86.9%。

基流科技2025年实现收入5.20亿元,其中AI算力集群产品收入4.37亿元,AI算力集群运营服务收入8363.7万元。运营服务收入由2024年的1695.5万元增至2025年的8363.7万元,占总收入的比重由5.2%升至16.1%。

IPO申请之外,一级市场融资

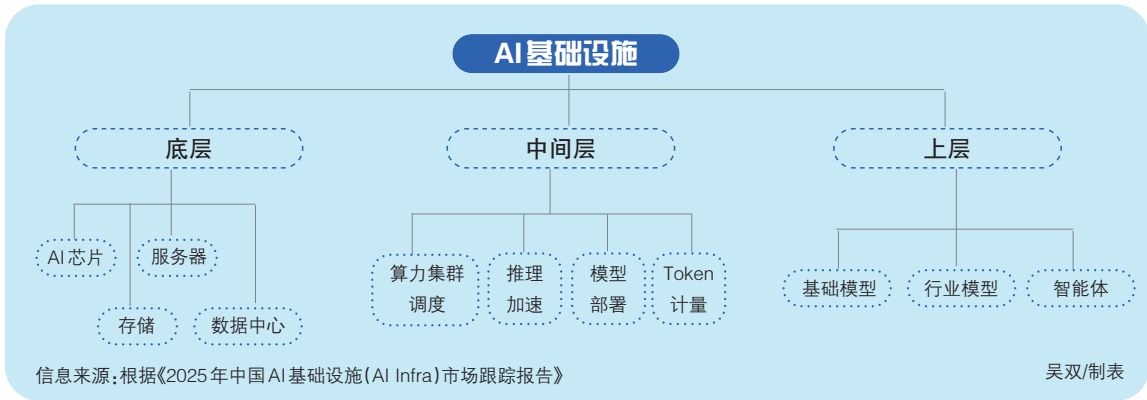
单位产出和有效利用率是关键

客户关注点正由“资源和模型能否使用”,转向“同一批资源能生产多少可销售的Token”。

调用量和营业收入快速增长,并不意味着词元服务已经形成稳定的盈利模式。硅基流动披露的不同交付模式,呈现出较为明显的收入和成本差异。

2025年,硅基流动公有云服务收入2926.1万元,毛利率为-119%;本地部署解决方案收入2606.9万元,毛利率为82.5%。整体毛利率从2023年的83.3%降至2024年的39.4%,2025年由正转负,跌至-24.0%。

从客户需求看,趋境科技称,客户关注点正由“资源和模型能否使用”,转向“同一批资源能生产多少可销售的Token”,并综合考察首Token时延、输出速



仍在继续。2026年5月,无问芯穹宣布完成超过7亿元融资,公开披露的累计融资规模超过22亿元,资金将用于多元异构算力适配、模型服务和AI基础设施建设。

7月13日,趋境科技宣布完成A轮融资,半年内累计融资超过10亿元。本轮融资由河南投资集团全资子公司河南创新投资集团有限公司旗下汇融基金领投,资金主要用于扩大高品质词元产能储备,围绕模型性能、推理效率、资源利用、缓存复用、服务隔离、质量监控和成本

度、并发承载、服务稳定性和单位成本。

对于如何衡量Token生产能力,趋境科技方面称,一是同一批算力能够生产多少有效Token,二是相关算力有多少时间处于有效工作状态。单位产出和有效利用率共同影响Token业务的单位成本和收入。

在田丰看来,视频生成、AI漫剧、多模态Agent每次调用的Token消耗是文本对话的数万倍,这是需求侧最大的乘数因子。Agent则使Token消耗从“人机对话”走向“机器间对话”,调用量级完全不同。

成本方面,官方信息显示,

控制等环节升级ATaaS平台。

趋境科技方面表示,融资还将用于建设面向头部模型、互联网平台和区域产业生态的高品质AIToken工厂,下一阶段,公司将把已经形成的Token工厂能力复制到更多算力中心、芯片组合和实际客户项目中。目前,趋境科技已有日均万亿级产能项目建成投产,并正在筹建日均产能数万亿级的AIToken工厂。

清程极智于2024年3月至2025年7月连续完成三轮融资,投

资方包括地方国资平台、联想创投、并行科技、中科创星及奇绩创坛等。投资方覆盖国有资本、芯片及服务器产业资本、算力运营企业和市场化基金。

天使投资人、资深人工智能专家郭涛表示,资本目前主要观察企业能否获得稳定且成本可控的算力供给,调度和推理优化能否提高算力利用效率、降低单位Token成本,以及模型接入数量、行业客户覆盖和终端付费客户规模能否支撑业务复制。

万亿量级。”

在郭涛看来,在融资审核中,机构会将业务增速、优质大客户储备、单位Token毛利与资金投入规模进行匹配。如果营收和Token消耗量增长能够与算力投入相匹配,政企及产业客户数量持续增加,相关融资需求更容易获得认可。如果企业主要依靠扩大硬件规模和低价获客,而客户增长及资源利用率没有同步改善,后续融资难度可能增加。

“整体看,只有形成技术壁垒或者行业专属服务能力,企业才能在资本循环中找到稳妥的退出路径,行业洗牌速度会在未来两年明显加快。”郭涛表示。

硅基流动拟赴港上市 “Token工厂”是门好生意吗？

中经记者 石健 北京报道

AI迭代加速之下,“卖Token”成为一门生意。

近期,主打“Token工厂”概念的北京硅基流动科技股份有

逐浪Token经济

何为“Token工厂”?

公司创始人袁进辉曾公开表示:它并不直接提供算力,也不试图打造自己的大模型,而是连接模型、芯片和应用,将分散的算力资源和模型能力加工成开发者、企业可以直接调用的Token服务,由此“让模型像水电一样能被按需使用”。

就目前市场来看,Token供应市场主要有全栈生态和独立生态两种模式,像阿里云、腾讯云、百度智云等头部云厂商,即为全栈生态模式,它们依托自有的云资源、算力、模型及应用场景打造“自有算力+自有模型+自有应用”的模式。而独立生态Token供应商则不

烧“代金券”可持续性挑战

除了成本高企的问题,硅基流动营收和毛利率问题也受到市场关注。

招股书显示,硅基流动的核心业务主要分为两部分:一是公有云服务,包括无服务器词元服务(按Token用量计费)和专属实例(预留独享算力,按固定费用或使用量计费),面向个人开发者和中小企业;二是本地部署解决方案,即将推理引擎和算力编排系统部署到客户自有数据中心,收取软件许可费、实施费和维护费,主要面向大型政企客户。

从营收占比来看,上述两项业务目前“五五分”。具体来看,2024

限公司(以下简称“硅基流动”)向港交所主板递交上市申请书。

招股书显示,2025年,硅基流动营收5533万元,净亏损达3.45亿元。公司的毛利率降至-24%。此

绑定任何云提供商、模型或应用场景,通过连接多类型算力、多元模型、不同算力中心及企业客户需求,提供多模型、多算力、跨部署环境的词元供应服务。

“硅基流动是开放独立的Token供应平台,提供高性价比、高可靠的优质Token服务。”硅基流动相关负责人向记者表示。

作为独立第三方,供应Token首先需要算力。招股书显示,硅基流动支持英伟达、AMD等国际GPU及昇腾、沐曦、摩尔线程、海光等国产AI芯片。

业内有形象的比喻——拥有GPU的厂商类似“大房东”,Token工厂作为“二房东”将GPU租过来,

年,硅基流动公有云业务的营收仅为106.9万元,仅1年的时间激增近27倍至2025年的2926.1万元,但毛利率却为-119%。本地部署业务2024年营收为627.7万元,占总营收比重85.4%。2025年,本地部署业务营收增至2606.9万元,1年增长4倍,增幅远不及本地云业务,但该业务仍保持了82.5%的毛利率。

公有云营收的大幅增加,与硅基流动提出的代金券模式密不可分。

招股书显示,2025年,约有5421.3万元专门用于免费发放To-

ken代金券,占当年销售费用的64.7%。结合营收计算,意味着每得到1.02元收入,需发放1元代金券。

硅基流动相关负责人在接受

外,2025年,硅基流动约有5421.3万元专门用于免费发放Token代金券,占当年销售费用的64.7%。结合营收计算,意味着每得到1.02元收入,需发放1元代金券。

不过,该模式之下,招股书显示,2025年营收5533万元,同比增长653.2%;但毛利率从2024年的39.4%降至-24.0%,年内亏损3.45亿元,累计亏损达4.396亿元。

营收高增长,毛利率骤降,问题出在哪?招股书显示,2025年,硅基流动的销售成本从445.2万元飙升至6863.2万元,增幅约15.4倍,远超收入增幅的6.5倍。其中,算力资源成本达5962.7万元,同比上涨22倍,占销售成本的86.9%,这意味着公

64.7%。这意味着每发放1元代金券,可得到1.02元收入。

针对这种“烧钱”模式,招股书提出,随着平台及市场影响力逐步成熟,计划结构性地降低对广泛推广型代金券分发依赖,并预计销售及营销开支占总收入比例将稳步下降,从而带来持续的经营杠杆效应。“公有云服务快速扩张期间,需要提前租赁庞大的底层算力资源来支持激增的词元需求,而算力资源的利用率仍处于爬坡期,导致成本增幅大幅跑赢了收入增幅。”

此外,招股书中也提及硅基流

《中国经营报》记者采访时表示:对于硬科技基础设施企业而言,研发与销售并非相互对立的关系,而是相辅相成。未来将继续扩大在技术与产品、生态层面的优势。

司每产生1元收入,仅算力成本就消耗了1.08元,未能覆盖其他运营成本。

采访中,一些业内人士直言,对于云厂商而言,依托自有模型、云计算可以打包出售,整体成本则可谓摊薄。但对于不涉及模型业务的硅基流动来说,并不能通过该模式来压低Token价格。

从外部看,2026年5月,DeepSeek宣布将V4-Pro模型API永久降价75%,云厂商MaaS上线。这意味着随着基础大模型公司不断完善官方API、持续下调调用价格,并逐步拓展企业端服务,开发者是否转向从模型厂商获得推理能力,值得关注。

动本地部署客户数量从2024年的28家减少至2025年的20家,虽然平均每名客户收益从22.4万元增加至130.3万元,但大单增加的另一面是,其收入依赖少数大客户的情况也开始凸显,是否可持续、可预测,有待拉长周期观察。

对于代金券“降速”之后,是否还能持续注册用户高增长及其他问题,硅基流动相关负责人回应称:长期留存依靠生产级SLA(Service Level Agreement服务级别协议)稳定性,以及提供高性价比、高可靠的优质Token服务。

如何突围？

招股书援引的弗若斯特沙利文数据显示,按2025年的Token年吞吐量计算,硅基流动是中国最大的独立Token供应商,并在中国所有Token供应商中位列第4。

不过硅基流动的市场份额仅为1.5%,排名前3家企业的市场份额分别为42.7%、32.5%、11.8%。

面对市场份额占比问题,硅基流动相关负责人直言,目前的确处于“跑马圈地”阶段。相对于自成生态、垂直整合的云厂商,硅基流动作为独立Token供应平台,对于底层算力厂商、模型厂商以及上层应用都是中立的,满足了企业客户追求“模型自由”与规避“供应商锁定”的刚性需求。

在行业看来,检验“Token工厂”是否是一门好生意,关键在于其能否有成熟的商业化落地模式和商业价值。

记者注意到,类似硅基流动这类的Token经济“中间层”公司,还有北京无问芯穹科技有限公司(以下简称“无问芯穹”)、北京清微智能科技股份有限公司(以下简称“清微智能”)等,技术路线各有侧重,同时通过自建大模型平台或依托算力中心的模式,来减少模型、算力成本的支出——无问芯穹早于2023年年初布局了Agentic MaaS大模型服务平台,而清微智能则采取与联通某地分公司数据中心合作共建“Token工厂”。

对此,一些业内人士直言,“Token工厂”模式价值提升的核心还是在于先降低算力成本,至少让模式支撑下去。换言之,该模式是否有足够的钱来“烧”,仍是关键。

从融资情况看,2023年12月,硅基流动启动天使轮融资,估值达到2.8亿元。截至目前,公司总计完成7轮融资。2026年6月16日,公司官宣完成超20亿元B轮融资,创下2026年以来国内第三方MaaS赛道最大单笔融资纪录。公司投后估值达77.4亿元。

记者注意到,硅基流动此次递交IPO,依据港交所18C章规定,公司为未商业化公司。未商业化公司为无法满足上市规则18C.03(4)条所载收入要求的特专科技公司。

硅基流动相关负责人表示,推理需求大爆发引发Token业务激增,公司通过IPO可以补充长期资金,更好投向技术、研发和产品,进一步服务好客户,并扩大行业领先优势。

另据招股书显示,硅基流动账面现金加定存2.72亿元,按单月现金消耗约1480万元计算,可支撑约18个月。

就“Token工厂”本身来说,闲置的Token是否也会成为影响现金流的原因?对此,硅基流动相关负责人回复称:“从宏观供需来看,随着生成式AI应用与Agent爆发,当前市场整体正处于优质Token供不应求的阶段。”