国金证券 SINOLINK SECURITIES

算力再次加速的底层逻辑计算机行业研究

2026 年 03 月 21 日

买入（维持评级）

行业点评

证券研究报告

计算机组
分析师：刘高畅（执业 S1130525120005）
liugaochang@gjzq.com.cn

算力再次加速的底层逻辑

计算需求范式跃迁：从 Prompt 到长 Agent

1)人工智能的交互范式正在经历根本性转变,AI 系统已从单次问答工具演进为能够推理、规划、持续运行的自主 Agent，趋势已获明确印证：OpenRouter 平台数据显示多步骤推理和链式工具调用正在快速取代传统单轮交互 Agent 开源框架 OpenClaw 发布仅四个多月便以超过 24.8 万 GitHub 星标登顶全球开源项目榜首，标志着长运行 Agent 从实验阶段全面进入生产部署。2)Agent 任务对 Token 的消耗已远超传统问答场景：Anthropic 实测数据显示，单 Agent 消耗约为对话模式的 4 倍，多 Agent 系统则高达 15 倍。NVIDIA 在其 2026 年 1 月技术博客中亦明确指出，下一代 AI 工厂必须具备处理数十万输入 Token 的能力，以支撑 Agentic 推理所需的长上下文。范式跃迁已经发生，算力需求的新增长逻辑形成。

长 Agent 驱动算力需求非线性提升

长 Agent 对算力需求的拉动有几个核心原因： 1）技术机制：首先大模型自注意力机制的计算成本与上下文长度呈二次方增长，其次推理 Decode 阶段天然受制于内存带宽，随着 KV Cache 随上下文线性膨胀，GPU 利用率持续下降，吞吐瓶颈日益突出，主流厂商的定价结构就是物理成本的体现：谷歌 Gemini 3.1 Pro 和阿里云 Qwen 均采用按上下文长度分档的阶梯定价。2）多 Agent 协作架构的兴起引入了额外的通信开销维度。Gartner 数据显示，2024 年 Q1 至 2025 年 Q2 企业对多 Agent 系统的询盘量暴增 1445%；而 Google DeepMind 研究指出，并行 Agent 之间的全局上下文压缩传递会产生不可避免的"协调税"，通信成本随 Agent 数量非线性扩大。3）杰文斯悖论进一步放大了上述效应，微软 CEO 纳德拉预判，模型推理效率的提升带来成本下降反而刺激使用量以更快速度增长。综合分析，我们认为 Agent 运行时长的增加是技术趋势的必然，在可见的未来，对内存带宽、互联吞吐与智能计算密度的需求，将持续以非线性速率扩张。

投资建议

相关标的：

海外算力/存储：中际旭创、东山精密、胜宏科技、天孚通信、新易盛、工业富联、兆易创新、大普微、源杰科技、景旺电子、英维克等；Lumentum、闪迪、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

国内算力：寒武纪、东阳光、海光信息、协创数据、豫能控股、华丰科技、亿田智能、星环科技、网宿科技、首都在线、神州数码、百度集团、大位科技、润建股份、中芯国际、华虹半导体、中科曙光、润泽科技、浪潮信息、东山精密、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、欧陆通、杰创智能。

CPU：海光信息、中科曙光、澜起科技、禾盛新材、中国长城、龙芯中科、兴森科技、深南电路、宏和科技、广合科技。

AI 应用：1）超级入口：腾讯控股、Minimax、智谱、阿里巴巴、科大讯飞。2）星环科技、德才股份、美年健康、中控技术、卓易信息、昆仑万维等 AI INFRA&高增长&高壁垒。其他：空天时代、具身智能等。

风险提示

行业竞争加剧的风险；技术迭代不及预期的风险；特定行业下游资本开支周期性波动的风险。

一、计算需求范式跃迁：从 Prompt 到长 Agent

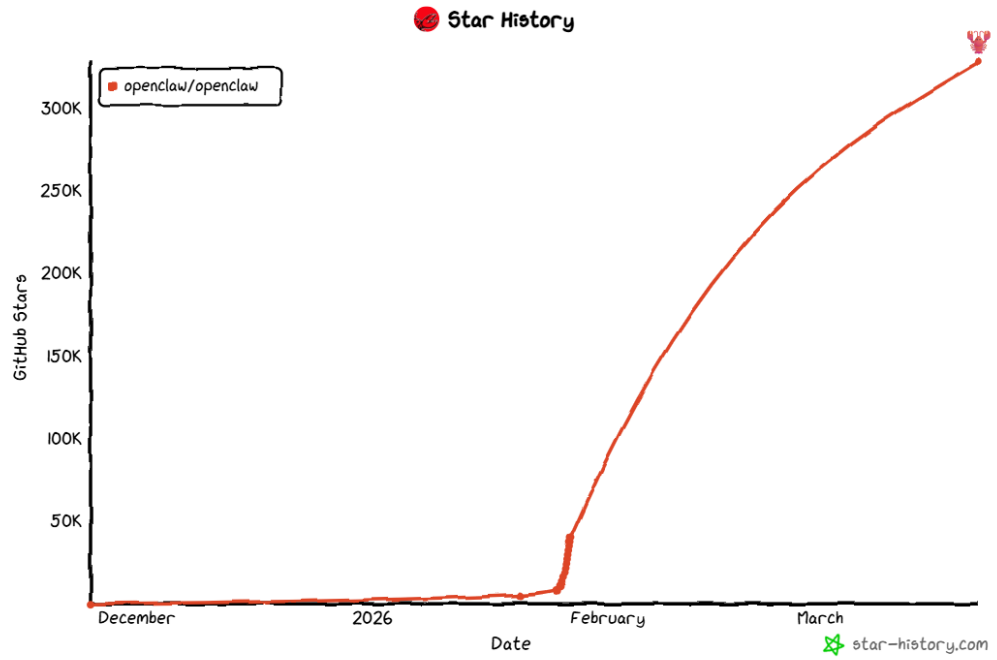
1.1 从 Prompt 到长 Agent 的变迁

据英伟达 GTC 2026 大会博客，人工智能正从简单的、基于 Prompt 的工具发展成为能够推理、规划和行动的智能、长期运行的系统。这些自主 Agent 不仅能生成文本，还能编写代码、调用工具、分析数据、模拟结果并持续改进。

大模型聚合平台 OpenRouter 的报告也提到 LLM 的使用正从单回合交互转向智能推理，模型需要进行规划、推理和执行，并跨越多个步骤。它们不再生成一次性响应，而是协调工具调用、访问外部数据，并迭代优化输出以达成目标。早期证据表明，多步骤查询和链式工具使用正在增加。随着这种范式的扩展，评估标准将从语言质量转向任务完成度和效率。下一个竞争前沿是模型执行持续推理的有效性，这一转变最终可能会重新定义大规模智能推理在实践中的意义。

科创板日报 2026 年 3 月 3 日报道，发布仅四个多月的开源智能体项目 OpenClaw 创造了历史——以超过 24.8 万的 GitHub 星标数正式登顶星标榜，超越 Linux 成为 GitHub 平台上最受欢迎的开源项目。OpenClaw 的爆炸性扩散，标志着长运行 Agent 从实验阶段进入大规模生产部署。

图表1：OpenClaw 的 GitHub 星标增长趋势



来源：star-history，国金证券研究所

1.2 Agent 上下文长度的结构性增长

Agent 任务中模型所需处理的 Token 数量往往远超传统问答场景。

Anthropic 发表的测试数据表示智能体通常比聊天交互消耗的令牌多约 4 倍，而多智能体系统比聊天消耗的令牌多约 15 倍。

英伟达 2026 年 1 月的技术博客对下一代 AI 工厂的计算需求做出了明确定性：为了大规模地提供这些功能，下一代 AI 工厂必须处理数十万个输入标记，以提供智能推理、复杂工作流程和多模态管道所需的长期上下文，同时在功率、可靠性、安全性、部署速度和成本的限制下维持实时推理。

二、为何长 Agent 驱动算力需求非线性提升

2.1：大模型架构的天然机制

首先广为人知的点在于大模型的自注意力机制的计算成本与上下文长度呈二次方关系。

其次，另一个瓶颈在于内存：大型语言模型的推理分为两个阶段：Prefill（预填充）阶段与 Decode（解码）阶段。前者对输入 Token 并行处理，计算效率较高；后者逐 Token 串行生成，天然受制于内存带宽。

行业研究机构 Clarifai 明确指出，解码发生在预填充阶段之后，每次生成一个 token；每个 token 的计算都依赖于之前所有 token，因此该阶段是顺序执行且受限于内存。模型会从之前的步骤中检索缓存的键值对，并为每个 token 添加新的键值对，这意味着限制吞吐量的是内存带宽，而非计算能力。由于模型无法跨 token 并行处理，GPU 核心经常在等待内存读取时处于空闲状态，导致利用率不足。随着上下文窗口增长到 8K、16K 甚至更大，键值缓存（KV Cache）变得非常庞大，进一步加剧了这一瓶颈。

图表2：KV Cache 工作机制示意图

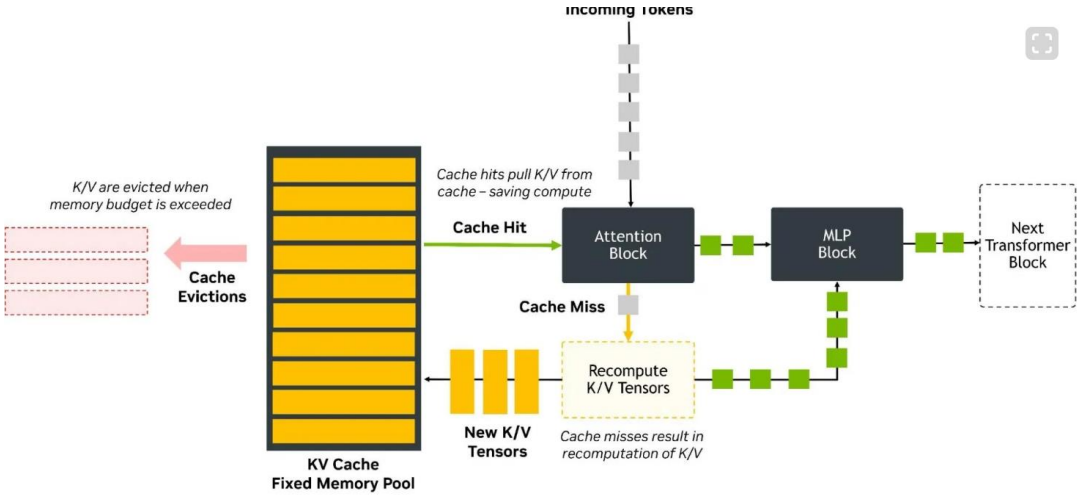


图 2. 传入的令牌查询固定的 K/V 张量内存池（KV 缓存）；缓存命中会重用存储的值以减少计算，而缓存未命中会触发 K/V 重新计算，并在达到内存限制时触发可能的驱逐。

来源：英伟达，国金证券研究所

从成本角度看，Token 阶梯定价正是这种算力瓶颈的货币化表现。在当前许多大模型厂商都按上下文长度阶梯定价。

例如 2026 年初，谷歌发布的 Gemini 3.1 Pro 定价：

输入 token：每百万 2.00 美元（输入量 <= 200K），每百万 4.00 美元（输入量 > 200K）。

输出 token：每百万 12.00 美元（输入 <= 200K），每百万 18.00 美元（> 200K）。

国内阿里云则更是对 Qwen 模型根据上下文长度分了 0-32k，32-128k 和 128k 以上三档定价。

图表3：Qwen 模型阶梯定价

模型名称	模式	单次请求的输入 Token 数	输入单价（每百万 Token）	输出单价（每百万 Token） 思维链+回答
qwen3-max Batch 调用半价 上下文缓存享有折扣	非思考和思考模式	0<Token≤32K	2.5 元	10 元
		32K<Token≤128K	4 元	16 元
		128K<Token≤252K	7 元	28 元
qwen3-max-2026-01-23	非思考和思考模式	0<Token≤32K	2.5 元	10 元
		32K<Token≤128K	4 元	16 元
		128K<Token≤252K	7 元	28 元
qwen3-max-2025-09-23	仅非思考模式	0<Token≤32K	6 元	24 元
		32K<Token≤128K	10 元	40 元
		128K<Token≤252K	15 元	60 元
qwen3-max-preview 上下文缓存享有折扣	非思考和思考模式	0<Token≤32K	6 元	24 元
		32K<Token≤128K	10 元	40 元
		128K<Token≤252K	15 元	60 元

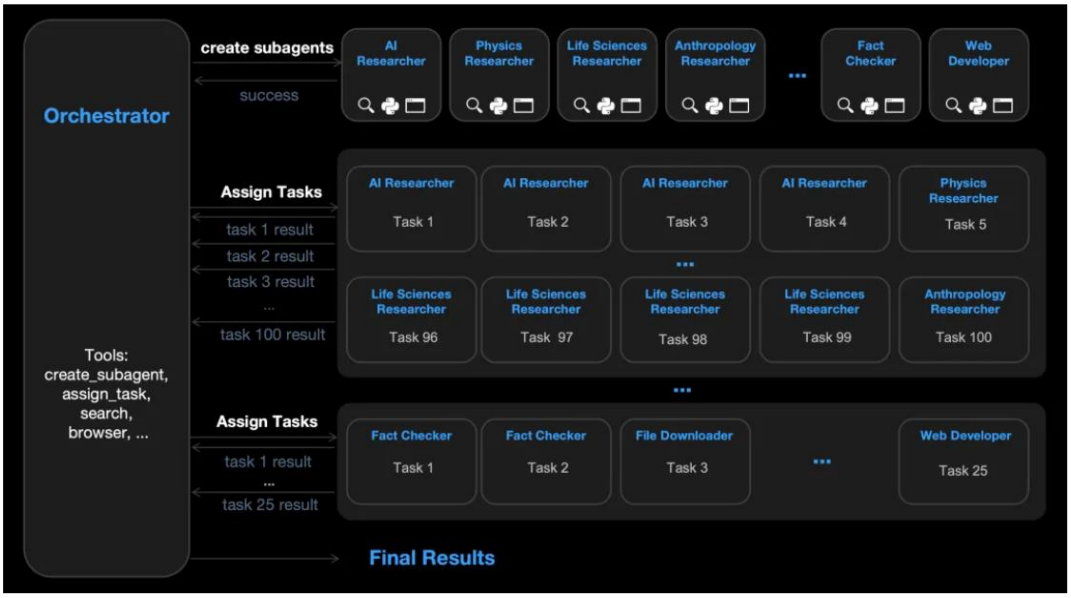
来源：阿里云，国金证券研究所

2.2 多 Agent 通信：计算量指数级扩张的另一原因

Agentic AI 的另一个主流架构趋势是多 Agent 协作（Multi-Agent Systems）。Gartner 报告显示，从 2024 年 Q1 到 2025 年 Q2，多 Agent 系统的企业询盘量暴增 1445%。

月之暗面的 Kimi k2.5 模型就重点推出了 Agent 集群的能力，它能根据任务需求，现场调度多达 100 个分身，并行处理 1500 个步骤。所有的角色分配与任务拆解，无需预设，全由 K2.5 现场决策。给 Kimi Agent 集群投喂 40 篇关于心理学和 AI 的论文。Kimi 先是通过多次调用工具，按顺序把这 40 篇论文通读一遍，以此确保上下文里完整保留了所有必要信息。紧接着，它衍生出几个子 agent，本质上是 Kimi 的「分身」，分别负责不同章节撰写。最后，主 agent 负责把关验收，将所有内容汇总生成了一份长达几十页的专业 PDF 综述。

图表4：kimi k2.5 的 Agent 集群协作示意图



来源：月之暗面 kimi，国金证券研究所

这一架构在提升任务复杂度处理能力的同时，引入了新的算力消耗维度：Agent 间通信。Google DeepMind 研究团队在 2025 年 12 月发表的《Towards a Science of Scaling Agent Systems》中指出，多 Agent 系统存在固有的通信瓶颈：并行 Agent 之间必须将全局上下文压缩为 Agent 间消息传递，产生不可避免的协调税（coordination tax）。

2.3 杰文斯悖论：效率提升反而提高算力消耗

AI 领域已出现明显的杰文斯悖论——随着模型推理效率的提升，反而刺激使用量以更快速度上升，最终推高总算力消耗。

微软 CEO 纳德拉就表示"随着人工智能变得更加高效和普及，我们将看到它的使用量呈爆炸式增长，使其成为一种我们永远都无法满足的商品。

综合分析，我们认为 Agent 运行时长的增加是技术趋势的必然，在可见的未来，对内存带宽、互联吞吐与智能计算密度的需求，将持续以非线性速率扩张。轮算力基建的加速，是由一条清晰的因果链驱动的结构性通胀：

Agent 运行时长的增加（Long-Running Agent）➞ KV Cache 线性膨胀与解码阶段吞吐下降 ➞ 单位 Token 成本上升 ➞多 Agent 协作的通信矩阵进一步放大需求规模 ➞ 杰文斯悖论使单位推理成本下降反而提高总消耗增长 ➞ 宏观算力资本开支将不断提高。

三、投资建议

相关标的：

海外算力/存储：中际旭创、东山精密、胜宏科技、天孚通信、新易盛、工业富联、兆易创新、大普微、源杰科技、景旺电子、英维克等；Lumentum、闪迪、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

国内算力：寒武纪、东阳光、海光信息、协创数据、豫能控股、华丰科技、亿田智能、星环科技、网宿科技、首都在线、神州数码、百度集团、大位科技、润建股份、中芯国际、华虹半导体、中科曙光、润泽科技、浪潮信息、东山精密、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、欧陆通、杰创智能。

CPU：海光信息、中科曙光、澜起科技、禾盛新材、中国长城、龙芯中科、兴森科技、深南电路、宏和科技、广合科技。

AI 应用：1）超级入口：腾讯控股、Minimax、智谱、阿里巴巴、科大讯飞。2）星环科技、德才股份、美年健康、中控技术、卓易信息、昆仑万维等 AI INFRA&高增长&高壁垒。

其他：空天时代、具身智能等。

风险提示

行业竞争加剧的风险： 在信创等政策持续加码支持计算机行业发展的背景下，众多新兴玩家参与到市场竞争之中， 若市场竞争进一步加剧，竞争优势偏弱的企业或面临出清，某些中低端品类的毛利率或受 到一定程度影响。
技术研发进度不及预期的风险： 计算机行业技术开发需投入大量资源，如果相关厂商新品研发进程不及预期，表观层面将 呈现出投入产出在较长时期的滞后特征。
特定行业下游资本开支周期性波动的风险： 部分计算机公司系顺周期行业，下游资本开支波动与行业周期性相关性较强，或在个别年 份对于上游软件厂商的营收表现产生扰动。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究