国金证券 SINOLINK SECURITIES

CPU 涨价能持续多久？计算机行业研究

2026 年 01 月 24 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
分析师：陈芷婧（执业 S1130525120008） chenzhijing@gjzq.com.cn
联系人：孙恺祈 sunkaiqi@gjzq.com.cn

CPU 涨价能持续多久？

本周观点

- **Agent 驱动的强化学习(RL)时代，CPU 可能比 GPU 更早成为瓶颈。**与传统单任务 RL 不同，现代 Agent 系统需要同时运行成百上千个独立环境实例，"环境并行化"让 CPU 成为事实上的第一块短板。主要源于三大核心逻辑：1）Multi-Agent 带来 OS 调度压力， Agent 的“推理-执行-反思”循环机制需要操作系统层面频繁调度，且沙盒（Sandbox）的运行高度依赖 CPU 算力。2）为解决长上下文导致的 GPU 显存容量问题，业界采用 KV Cache Offload 技术将数据迁移至 CPU 内存，这不仅要求 CPU 具备大内存，还需其承担繁重的调度与传输任务。3）高并发工具调用： 检索、编码、网页浏览等非模型推理任务上由 CPU 执行，在高并发场景下，多线程/多进程的处理需求推高了 CPU 的负载。
- **DeepSeek Engram 架构或进一步推动以存代算。**DeepSeek 推出 Engram 架构，把大模型里的“计算”和“超大规模记忆”解耦，Transformer 的算子全部在 GPU/加速卡上计算，而 1000 亿参数的 Engram 表存储运行则在 CPU 内存中，仅产生小于 3%的开销。此外 Anthropic 推出的 Claude Cowork，通过知识库为 Claude 设计的一种全新永久记忆方式。我们认为，类 Engram 架构能有效突破 GPU 显存限制，从而推动以存代算需求和 CPU 配比提升。
- **Agent 生态扩张引爆 CPU 性能瓶颈。**全球 Agent 生态正面临指数级跃迁，据 IDC 预测，2025 年至 2030 年间，活跃 Agent 数量、任务执行量及 Token 消耗量将分别以 139%、524% 和 3418% 的年复合增长率飙升。这种增长不仅是数量的堆叠，更伴随着任务复杂度的剧增。英特尔与佐治亚理工学院的研究表明，在 RAG 检索、工具调用等典型 Agent 工作负载中，CPU 承担了大幅度的延迟占比（如 HaystackRAG 任务中 CPU 耗时占 90.6%），成为实际性能的瓶颈。此外随着 Batch Size 增加，CPU 的能耗逼近 GPU，且面临严重的上下文切换压力，证明了 CPU 在 Agent 时代的关键地位。
- **供需失衡全面爆发，算力木桶新短板已现。**英特尔已紧急将产能转向服务器端，导致消费电子端交付受阻；英伟达则因 ARM CPU 瓶颈，计划在下一代 Rubin 架构中大幅提升 CPU 核心数，并开放 NVL72 机柜对 x86 CPU 的支持。市场数据印证了这一趋势，Jon Peddie Research 报告显示，2025 年第二季度全球服务器 CPU 出货量同比大增 22%，客户端 CPU 亦连续两季度增长。英特尔 CFO 表示预计第一季度可用供应将降至最低水平，随后在第二季度及以后有所改善，公司正应对整个行业的供应短缺。Agent 时代算力的“木桶效应”已经显现，目前 CPU 正演变为类似于存储的新短板，补足这一短板将是下阶段算力基础设施建设的重中之重。。

投资建议

- 相关标的：
  CPU：海光信息、中科曙光、禾盛新材、中国长城、龙芯中科、兴森科技、深南电路、宏和科技。
  国内算力：海光信息、寒武纪、东阳光、协创数据、华丰科技、星环科技、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、禾盛新材、润泽科技、浪潮信息、东山精密、亿田智能、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。
  海外算力/存储：中际旭创、新易盛、兆易创新、大普微、中微公司、天孚通信、源杰科技、胜宏科技、景旺电子、英维克等；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。

风险提示

- 行业竞争加剧的风险；技术研发进度不及预期的风险；特定行业下游资本开支周期性波动的风险。

内容目录

图表目录

一、三大逻辑揭示 Agent 对 CPU 的刚性需求

随着大模型的应用从简单的 Chatbot 向能完成复杂任务的 Agent 演进，计算负载的重心正在发生微妙的偏移。Agent 不仅需要 GPU 进行模型推理，更依赖高性能 CPU 来处理复杂的逻辑编排、工具调用和内存管理。以下是我们认为 Agent 驱动 CPU 需求爆发的三大核心逻辑：

① Multi-Agent 架构引发的 OS 调度压力

传统的 LLM 对话是线性的，而 Agent 的工作流则是复杂的闭环。“推理→执行→评估→反思”的循环机制： Agent 需要在生成 Token 之外，执行大量的逻辑判断和状态管理。，模型需要不断在“思考”和“行动”之间切换。导致操作系统层面的上下文切换和进程调度任务大幅增加。

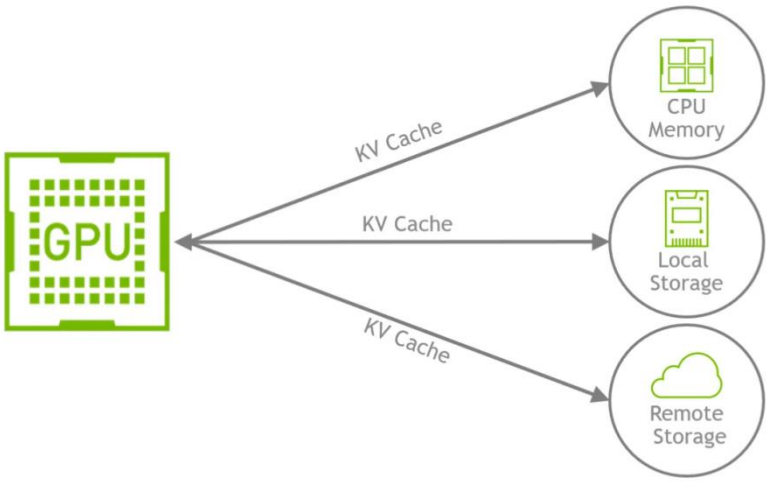
沙盒（Sandbox）需求飙升： Agent 执行代码等操作经常需要在隔离的云端沙盒中运行。这些沙盒环境的启动、运行和销毁依赖 CPU 算力。

② 长上下文场景下的 KV Cache 卸载对 CPU 的挑战

naddod 的技术文章阐述了其原理，键值缓存 (KV Cache) 可以加速 Transformer 推理，但它也会带来一个副作用：消耗大量显存。随着大型语言模型上下文长度的不断增长，这个问题会变得越来越突出。例如当上下文长度达到 8 万个 token 时，仅 KV Cache 本身就可能消耗数十 GB 的显存。更重要的是，GPU 显存不仅要容纳 KV Cache，还要容纳模型权重和中间计算结果。一旦显存耗尽，推理就会崩溃甚至失败。为了解决这一冲突，业界提出了键值缓存卸载（KV Cache Offload）方案。其核心思想是将 GPU 内存中不活跃或暂时未使用的键值数据迁移到其他存储介质例如 CPU 内存或者 SSD。然而 CPU 与 GPU 之间的通信带宽远低于 GPU 内部的 HBM 带宽。而且在进行 KV Cache 传输和管理时，也需要 CPU 进行任务的调度，进一步加剧了 CPU 的负载。

NVIDIA 2025 年 9 月的一篇技术博客《How to Reduce KV Cache Bottlenecks with NVIDIA Dynamo》就专门阐述了在长上下文场景下，利用 NVIDIA Dynamo 等技术将 KV Cache 卸载到 CPU 内存的必要性，并指出这是解决 HBM 瓶颈的关键手段。

图表1：KV Cache 卸载使得 KV Cache 能够从有限的 GPU 内存中传输到更大且性价比更高的存储



来源：Nvidia 官网，国金证券研究所

③ 高并发工具调用带来的 CPU 算力消耗

Agent 的能力不仅在于对话，更在于使用工具，例如检索、写代码、浏览网页。这些非模型推理任务主要由 CPU 承担。前文五大代表性 Agent 工作负载中各项任务的延迟数据证明了这一点。而且在高并发场景下可能有大量 Agent 同时工作，这些任务需要高性能 CPU 进行多线程/多进程处理。

二、Agent 生态扩张引爆 CPU 性能瓶颈

全球 Agent 生态将经历一场指数级的扩张。根据 IDC 最新的预测模型，能够在三个关键维度上看到极具张力的增长趋势：

1.活跃 Agent 数量的激增

IDC 预计，活跃 Agent 的数量将从 2025 年的约 2860 万，快速攀升至 2030 年的 22.16 亿。

这意味着五年后，能够帮助企业执行任务的数字劳动力数量将是今天的近 80 倍，年复合增长率 139%，换言之，平均每年活跃 Agent 数量都将以超过一倍的速度增长。

2.任务执行量的爆炸式增长

与数量相比，Agent 真正干活的频率增长得更快，年执行任务数将从 2025 年的 440 亿次暴涨至 2030 年的 415 万亿次，年复合增长率高达 524%。这意味着，企业将越来越习惯把工作外包给 Agent，从偶尔试用到深度依赖，Agent 将无处不在地嵌入进企业的业务流中。

3.Token 消耗量的数量级跃迁

随着 Agent 处理的任务越来越复杂，所需推理深度与调用链路不断加长，底层 Token 消耗也将呈现数量级的跃迁。预计年度 Token 消耗将从 2025 年的 0.0005 PetaTokens 暴增至 2030 年的 152,667 PetaTokens，年复合增长率高达 3418%。

这三组数字反映的并不只是 Agent 数量的增长，更是任务复杂度与推理深度的指数级提升。

图表2：全球企业活跃 Agent 关键数据预测,2025-2030

来源：IDC，国金证券研究所

Agent 任务对 CPU 负载提出更高要求：

据英特尔与佐治亚理工学院 2025 年 11 月的论文《A CPU-CENTRIC PERSPECTIVE ON AGENTIC AI》对代表性 Agent 任务进行了延迟、吞吐量和能耗指标的分析，揭示 CPU 相对于 GPU 对这些指标的显著影响，结果表明很多情况下 GPU 是 Agent 性能的瓶颈:在五大代表性 Agent 工作负载(HaystackRAG、Toolformer、ChemCrow、LangChain、SWE-Agent)中，.CPU.端的工具处理占延迟的.43.8%~90.6%.(如.ENNS.检索、WolframAlphaAPI.调用、文献搜索).，而.LLM.推理仅占较小部分。如 HaystackRAG 在.Natural.Questions.基准测试中检索耗时.8.0.秒(占总延迟.90.6%)，LLM.推理仅 0.5 秒。2

个

图表3：五大代表性 Agent 工作负载中的任务延迟分布

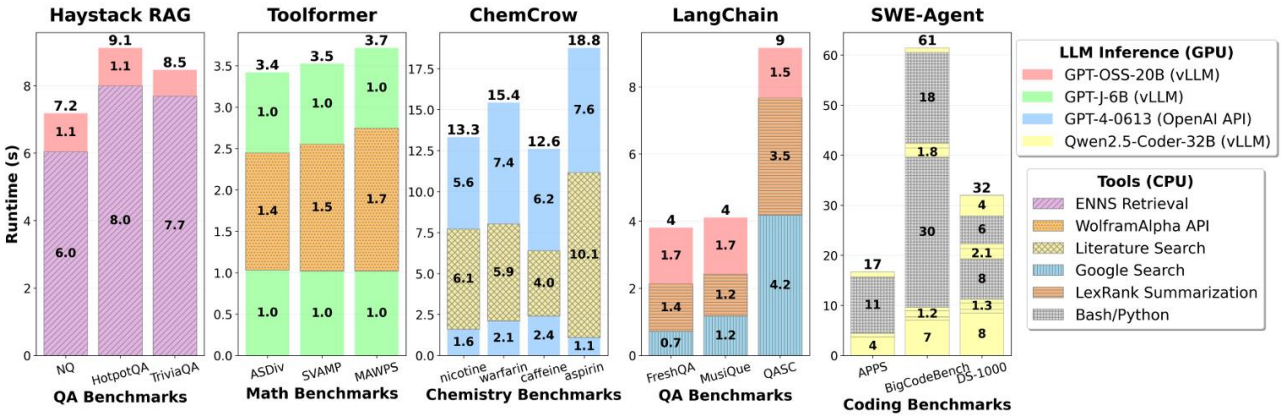


Figure 2. (a) Haystack with ENNS retrieval on QA benchmarks (b) Toolformer with WolframAlpha API on Math benchmarks (c) Chemcrow with literature (Arxiv/Pubmed) search tool on Chemistry benchmarks (d) Langchain with web search and LexRank summarization tools on QA benchmarks (e) Mini-SWE-Agent with bash/Python execution tools on coding benchmarks

来源：《A CPU-Centric Perspective on Agentic AI》，Ritik Raj, Hong Wang, Tushar Krishna，国金证券研究所

在 GPT-OSS-20B 模型的吞吐量测试中，随着 Batch Size 增加，不同 Agentic 工作负载的吞吐量增长逐渐放缓并趋于饱和：以 Langchain 为例，展示了当 Batch Size 达到 128 时，Langchain 基准测试中各组件的平均耗时（数据显示批次大小为 128 时存在严重的 CPU 上下文切换瓶颈）。

图表4：五大代表性 Agent 工作负载中的任务延迟分布

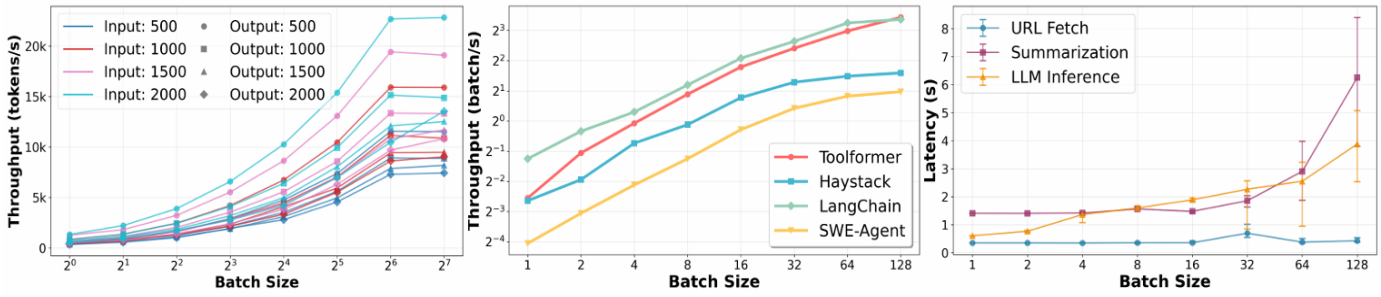


Figure 4. (a) vLLM throughput saturation for GPT-OSS-20B model (b) Throughput saturation for various agentic workloads (c) Average time taken by different components in Langchain benchmark showing a critical CPU context switching bottleneck at batch size 128

来源：《A CPU-Centric Perspective on Agentic AI》，Ritik Raj, Hong Wang, Tushar Krishna，国金证券研究所

在处理 LangChain 工作负载时，AMD Threadripper CPU 和 Nvidia B200 GPU 的动态能耗。关键转折点：在低 Batch Size（如 1-4）时，GPU 能耗显著高于 CPU。但随着 Batch Size 增加到 128 时，CPU 的能耗（1807 Joules）已经非常接近 GPU（2307 Joules）

图表5：处理 LangChain 工作负载时，AMD Threadripper CPU 和 Nvidia B200 GPU 的动态能耗

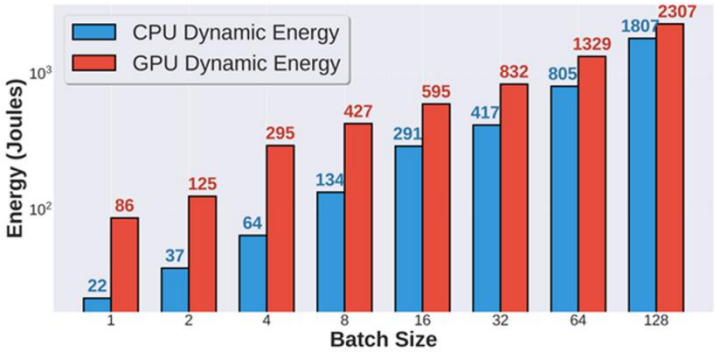


Figure 5. CPU (AMD Threadripper) and GPU (Nvidia B200) dynamic energy consumption for Langchain workload

来源：《A CPU-Centric Perspective on Agentic AI》，Ritik Raj, Hong Wang, Tushar Krishna，国金证券研究所

同时我们认为 DeepSeek Engram 架构或进一步推动以存代算。DeepSeek 推出 Engram 架构，把大模型里的“计算”和“超大规模记忆”解耦，Transformer 的算子全部在 GPU/加速卡上计算，而 1000 亿参数的 Engram 表存储运行则在 CPU 内存中，仅产生可忽略的开销（小于 3%）。

此外，Anthropic 也给 Claude Cowork 重磅升级，通过知识库为 Claude 设计的一种全新永久记忆方式。我们认为，类 Engram 架构能有效突破 GPU 显存限制，从而推动以存代算需求和 CPU 配比提升。

三、供需失衡全面爆发，算力木桶新短板已现

据芯榜 1 月 19 日报道，英特尔将 Intel 3 和 intel 7 产能紧急转向服务器，致使消费电子端交付保证率大幅下滑。

英伟达 Blackwell 架构的 ARM CPU 存在严重瓶颈，因此新一代 Rubin 架构大幅提升 CPU 核心数与超线程；同时英伟达开放英特尔 x86 CPU 用于 NVL72 互联机柜。而 Agent 云端沙盒调用量飙升带动云实例业务增长，进一步加剧了 CPU 供需紧张。

市场研究机构 Jon Peddie Research 2025 年 8 月公布的最新数据显示，全球客户端 CPU 市场已连续两个季度实现增长。2025 年第二季度，客户端 CPU 出货量环比增长 7.9%，同比增长 13%；同期服务器 CPU 出货量同比增长 22%，环比小幅上升 0.6%。

图表6：2025 年 Q2 全球客户端 CPU 市场增长状况以及服务器与客户端 CPU 比例

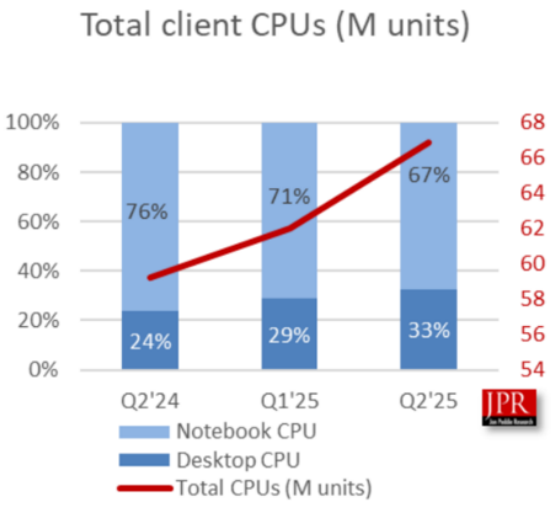


图 1. 今年前两个季度出现了非季节性增长。

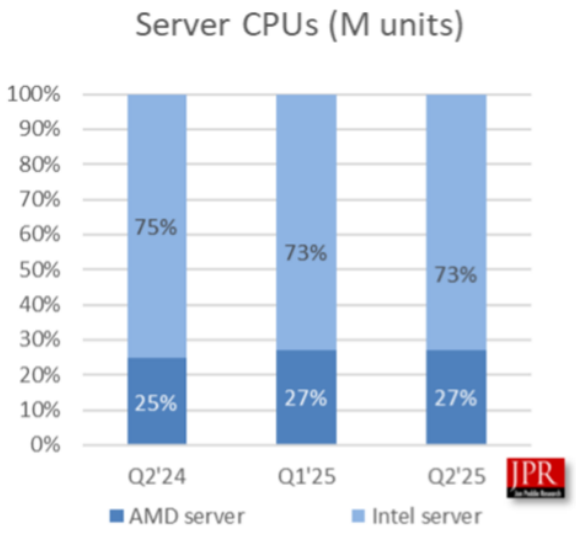


图 2. 服务器在整体市场份额中的比例。

来源：Nvidia 官网，国金证券研究所

2026 年 1 月 22 日，英特尔 CFO 表示预计第一季度可用供应将降至最低水平，随后在第二季度及以后有所改善，公司正应对整个行业的供应短缺。

我们认为，Agent 时代算力的“木桶效应”已经显现，业界从经历了从开始对 GPU 的堆叠，到存储的短缺，目前 CPU 正演变为类似于存储的新短板。补足这一短板将是下一阶段算力基础设施建设的重中之重。

四、相关标的

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

国内算力：海光信息、寒武纪、东阳光、协创数据、华丰科技、星环科技、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、禾盛新材、润泽科技、浪潮信息、东山精密、亿田智能、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。

海外算力/存储：中际旭创、新易盛、兆易创新、大普微、中微公司、天孚通信、源杰科技、胜宏科技、景旺电子、英维克等；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。

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