计算机行业研究

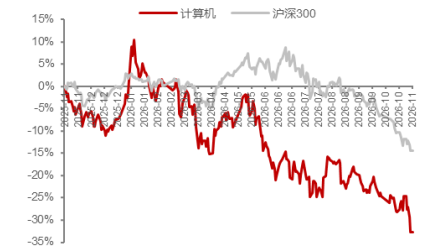
2026 年 11 月 2 日

行业研究报告

—AI 系列专题（一）：Agent 沙箱打开 CPU 增量空间，四大场景演绎需求弹性

评级：强于大市（维持）

行业指数相对沪深 300 表现



相关报告

《Jev 模型：给 Agent 装上"小脑"，专用决策模型开启效率新范式》

2026-09-22

《Meta 发布个人 AI 智能体 Muse，Agent 安全有望成为网络安全新需求》

2026-09-11

《GPT-6 Astra 发布：Agent 能力迎来跃迁，AI 应用落地有望加速》

2026-09-04

《工信部部署 AI 应用服务商培育专项行动，把握两条核心受益主线》

2026-09-01

《FDE 加速 AI 应用范式迁移：从模型供给走向业务价值交付》

2026-08-17

分析师：王鑫旸
执业证书编号:S0370526030001
邮箱：wangxy4@jyzq.cn

投资要点：

- **AI 需求由内容生成转向自主执行演进，沙箱有望成为服务器 CPU 的重要增量来源。**随着大模型推理及任务执行能力持续提升，AI 需求逐步从对话问答向 Agent 实际任务执行延伸。与主要由 GPU 承担的模型推理不同，沙箱作为 Agent 的执行环境，其计算负载主要由 CPU 承担。Agent 应用范式的变化推动 AI 算力需求由单纯模型推理向“推理+执行”拓展，CPU 有望从 GPU 的配套资源进一步发展为独立的算力需求来源。
- **Agent 规模化应用有望带来千万核级 CPU 需求，编程 Agent 为当前主要拉动力量。**沙箱 CPU 需求主要由并发规模及单位任务资源消耗决定。我们将 Agent 拆分为个人、编程、浏览器及企业级四类场景，分别以 1 亿日活、人均每日 1 个任务为统一基准，综合考虑沙箱运行时长、CPU 利用率及资源冗余后，**四类场景物理核需求加总约 2022 万核，折合约 31.6 万颗 64 核 CPU，其中编程 Agent 贡献近七成。**随着 Agent 应用规模扩大，沙箱有望带来可观的服务器 CPU 需求增量。
- **用户规模扩张与任务复杂度提升形成乘数效应，编程 Agent 兑现最快，企业级 Agent 长期弹性最大。**沙箱 CPU 需求增长不仅来自 Agent 渗透率提升，更取决于单用户任务频次、任务执行时长及并发规模的持续增加。编程 Agent 受益于 Vibe Coding，用户边界扩大与复杂工程任务并行化同步推进，需求兑现确定性较强；个人 Agent 由按需调用向常驻化发展，浏览器 Agent 随任务成功率提升加速渗透；企业 Agent 则逐步由轻量交互向复杂业务流程延伸，任务执行深度及并发规模具备更大的长期增长空间。
- **GPU 配套 CPU 配置提升叠加 Agent 独立执行需求，服务器 CPU 有望迎来新一轮增长机遇。**一方面，随着 GPU 计算能力持续提升，数据处理、任务调度及内存管理对 CPU 的性能和资源配置提出更高要求，提升 CPU 配置以改善 GPU 利用效率具备较强经济性，为服务器 CPU 需求提供基础支撑。另一方面，Agent 将 CPU 需求从 GPU 配套进一步拓展至独立任务执行，增长驱动由 GPU 部署规模延伸至 Agent 用户数量及任务执行密度。我们认为，两条需求主线有望形成共振，推动 CPU 在 AI 算力基础设施中的价值量持续提升，**建议关注服务器 CPU 产业链的长期增长机会。**

相关公司（以下公司均未评级）：

美股：Intel、AMD

A 股：海光信息、禾盛新材

- **风险提示：**Agent 商业化与渗透率不及预期；沙箱资源复用效率提升；Agent 执行向本地端迁移。

目录

一、Agent 用量高增，沙箱成为 CPU 需求的新增量 5
1. Agent 的 Token 消耗超越人类，成为大模型应用主要形式 5
2. 沙箱是 Agent 执行层基础设施，负载主要落在 CPU 6
3. 厂商布局：主流大模型平台集中推出沙箱与托管 Agent 服务 7
二、四类场景测算：基准需求达千万核量级 9
1. 场景拆解：个人、编程、浏览器、企业级四类沙箱需求 9
2. 测算口径及基准参数 9
3. 测算结果：1 亿日活基准下合计需求 2022 万物理核，编程 Agent 贡献近七成 10
4. 弹性判断：编程兑现最快，企业级理论弹性最大 10
4.1 个人 Agent：常驻化与渗透率提升共同放大需求 11
4.2 编程 Agent：用户边界扩大与任务复杂度提升正同步发生 12
4.3 浏览器 Agent：单沙箱 CPU 强度最高，成功率提升驱动放量 13
4.4 企业 Agent：渗透率与任务复杂度同步抬升，理论弹性最大 15
三、投资建议：建议关注服务器 CPU 17
四、风险提示 19

图表目录

图表 1：Agent Tokens 消耗量增速远超人类 5
图表 2：前沿模型可完成的任务时长呈指数增长 6
图表 3：沙箱对 CPU 及内存的需求 6
图表 4：沙箱对 CPU 及内存的需求 7
图表 5：大模型厂商 Agent 产品及沙箱概况 8
图表 6：Agent 沙箱场景分类 9
图表 7：四类场景基准参数 10
图表 8：物理核测算结果及对应 CPU 需求 10
图表 9：四类场景增长因子的上行空间 11
图表 10：任务型与常驻型 Agent 的沙箱占用对比 11
图表 11：全球和中国 AI 月活 12
图表 12：头部 AI 大模型用户规模趋势（截至 202606） 12
图表 13：Lovable 4/5 用户从事非技术性工作 12
图表 14：美国 AI 用户想要创作的应用类型 12
图表 15：Coding 单轮自主运行时长提升至 45 分钟 13
图表 16：任务难度提升，Coding Agent 并发规模扩大 13
图表 17：WebArena：Agent 任务成功率逼近人类 14
图表 18：OSWorld：Agent 计算机操作成功率快速提升 14
图表 19：浏览器 Agent 请求量快速增长 15
图表 20：企业应用集成任务型 Agent 的比例快速提升 15
图表 21：APEX-Agents 榜单 16

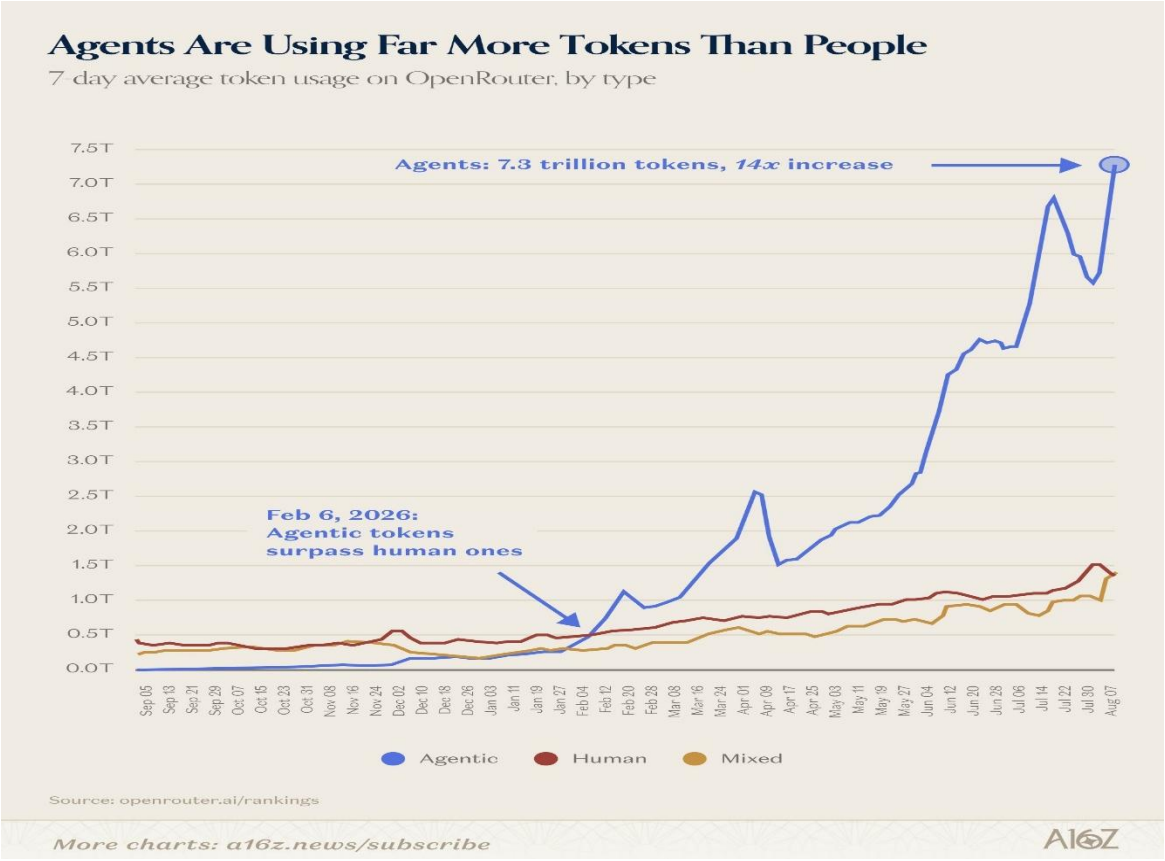
图表 22：企业 tokens 用量中 Agent 占比快速提升 16
图表 23：领先企业 token 使用差距已拉大到 8.3 倍 16
图表 24：相关公司 CPU 产品线简介 18

一、Agent 用量高增，沙箱成为 CPU 需求的新增量

1. Agent 的 Token 消耗超越人类，成为大模型应用主要形式

Agent 已成为大模型应用的主要形式。2025 年以来，大模型应用的主要形式由对话式问答逐步转向能够自主规划、调用工具并交付成果的 Agent。用户不再满足于获得一段回答，而是要求模型直接产出可运行的代码、整理好的表格、完成网页操作或生成一份报告。任务形态的变化首先体现为 Token 消耗的高增：根据 a16z 统计，2026 年 2 月 6 日起 Agent 消耗的 Token 量超过人类，此后增长斜率更为陡峭；截至 2026 年 8 月，Agent 的 Token 消耗（7 日均值）已达 7.3 万亿，为人类的 5 倍以上。

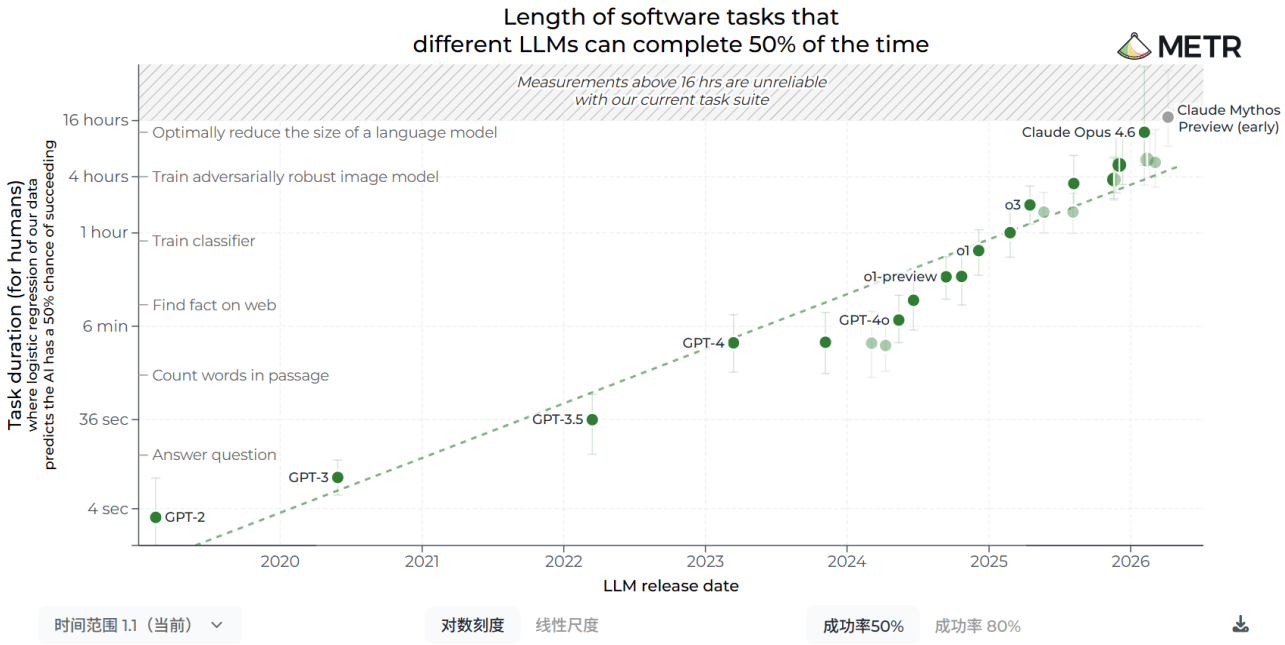
图表 1：Agent Tokens 消耗量增速远超人类



数据来源： a16z，金元证券研究所整理

模型可连续完成的任务时长呈指数级拉长，Agent 的可执行上限随之抬升，负载需求快速增长。据 METR 测算，前沿模型以 50%成功率可完成的任务时长在 2019—2025 年间约每 7 个月翻倍，2024 年以来缩短至 3.5—4 个月：从 GPT-2 的 2 秒、o1 的 39 分钟、o3 的 110 分钟，到 2026 年初 Claude Opus 4.6 的约 14.5 小时。底层模型更长的连续工作时间使 Agent 能够承接更复杂的任务，带来更多的工具调用与代码执行需求。

图表 2：前沿模型可完成的任务时长呈指数增长



数据来源：METR，金元证券研究所整理

2. 沙箱是 Agent 执行层基础设施，负载主要落在 CPU

沙箱是 Agent 执行层的基础设施，Agent 的并发数量与执行时长增长将直接转化为服务器 CPU 需求。Agent 的典型流程为“模型推理与任务规划—工具调用—沙箱执行—结果回传—再次推理”。沙箱（Sandbox）是为 Agent 提供安全隔离、运行环境管理和资源调度的执行容器：模型在 GPU 上完成规划与生成后，由编排层创建或分配沙箱，在其中运行代码、安装依赖、操作浏览器或处理文件，再将结果回填给模型。与 GPU 推理不同，沙箱内的工作负载以通用计算为主，资源消耗落在服务器 CPU 上，成为 GPU 配套之外的 CPU 增量重要来源。

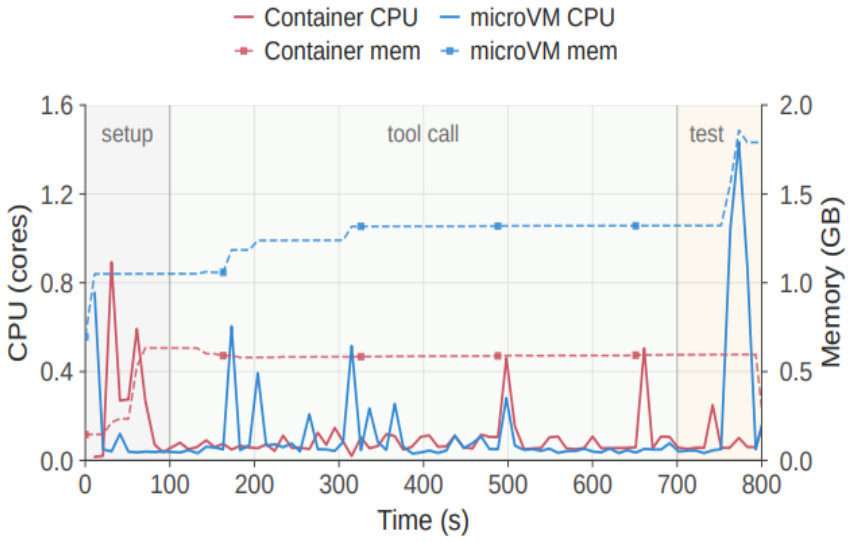
图表 3：沙箱对 CPU 及内存的需求



数据来源：金元证券研究所整理

沙箱 CPU 负载呈现“低均值、高峰值”特征，物理核需求取决于峰值并发规模及资源复用效率。据 DSec 论文，Agent 任务在环境初始化阶段出现约 0.9 核的容器 CPU 尖峰，在测试阶段 microVM 的 CPU 尖峰达到约 1.4 核，而工具调用阶段多数时间占用不足 0.1 核。此外，约 90%的容器/microVM 实际 CPU 使用量不超过申请量的 5%。**因此，沙箱 CPU 需求需同时考虑高峰时段的并发规模与实际 CPU 利用率，并为突发负载预留一定冗余。**

图表 4：沙箱对 CPU 及内存的需求



数据来源：《DSec: A Sandbox Infrastructure for Effective Agentic Training at Scale》，金元证券研究所整理

3. 厂商布局：主流大模型平台集中推出沙箱与托管 Agent 服务

主流大模型平台已集中推出沙箱与托管 Agent 服务，单沙箱公开规格多在 1—4 个 vCPU。OpenAI、Anthropic、Google 等海外厂商已提供云端沙箱或代码执行能力，阿里云百炼、火山方舟等国内平台亦推出沙箱及托管 Agent 服务，覆盖代码运行、浏览器操作与文件处理。从公开资源配置看，Anthropic API 代码执行环境为 1 CPU，Google Gemini 托管沙箱默认 2 个 vCPU，阿里云百炼提供 1 核 2GB 与 4 核 8GB 两档规格；Meta 于 2026 年 9 月发布的 Muse 根据三方实测，每位用户分配一台 2 vCPU、7.7GiB 内存的专属虚拟机。

图表 5：大模型厂商 Agent 产品及沙箱概况

厂商	Agent 产品/平台	沙箱部署形式	公开资源配置
OpenAI	ChatGPT、Codex、API Code Interpreter	官方云端隔离容器；Python、Shell 执行、文件处理；支持创建和复用容器	内存支持 1/4/16/64GB 等档位
Anthropic	Claude Code、Claude API Code Execution	API 提供云端隔离容器，支持 Python、Bash 及文件操作；Claude Code 另支持本地沙箱执行	API 代码执行环境：1 CPU、5GiB 内存、5GiB 工作区
Google	Gemini Managed Agents、Antigravity	官方托管 Linux 沙箱；支持代码执行、文件操作、环境复用与网络访问控制	默认配置为 2 个 vCPU 和 1.5 GiB 的 RAM
Meta	Muse、Muse Secure VM	每用户 1 台专属虚拟机；支持后台工作与多子 Agent 协同，浏览器位于独立环境	三方实测配置为 2 vCPU、7.7GiB 内存
Xai	Grok API Code Execution	官方托管 Python 隔离执行环境；主要用于计算、数据分析和代码验证，执行上下文不跨请求持久化	/
阿里	百炼 Managed Agents	官方云端容器沙箱；支持代码解释器、浏览器、综合运行环境，支持暂停和恢复	1 核 2GB 或 4 核 8GB
字节	火山方舟 Managed Agents、Computer Use Agent	官方托管沙箱，支持 Shell、文件处理；另提供 Windows/Linux 远程执行环境	/
Kimi	Kimi Agent、Kimi Claw、Agent Swarm	云端 Agent 执行；网页与应用构建涉及沙箱预览；Kimi Claw 支持云端持续运行	/

数据来源： OpenAI、Claude、Gemini、Google Cloud、Xai、阿里云、火山引擎、Meta、Starkinsider、Kimi，金元证券研究所整理

二、四类场景测算：基准需求达千万核量级

1. 场景拆解：个人、编程、浏览器、企业级四类沙箱需求

不同 Agent 场景的任务规模、沙箱运行方式和 CPU 资源消耗差异较大。我们将 Agent 分为个人 Agent、编程 Agent、浏览器 Agent、企业级 Agent 四类，覆盖当前 Agent 的主要使用场景。

图表 6：Agent 沙箱场景分类

场景	典型任务	沙箱特点	代表产品
个人 Agent	长期目标跟踪、比价监控、代下单、日程与邮件处理、文件处理、搜索、报告/PPT 生成	用户 1 台常驻 VM+按需临时沙箱，关闭 App 后仍运行；多步骤执行，复杂任务调用多个并行子 Agent	Meta Muse、Gemini Spark、Manus
编程 Agent	写代码、运行测试、编译、安装依赖、终端操作	多版本并行尝试，长时自主任务可达数小时，CPU 强度高	Codex、Claude Code、Cursor
浏览器 Agent	网页操作、数据抓取、表单填写、GUI 自动化	单次会话较短、任务量大，渲染负载重	各类 Browser Use 类产品
企业级 Agent	轻量：检索、审批流； 复杂：财务对账、批量文档审核、代码迁移	轻量调用频繁；复杂任务单次时长可达数十分钟至数小时	企业流程自动化 Agent

数据来源：金元证券研究所整理

2. 测算口径及基准参数

四类场景物理核测算口径：

$$物理核需求 = 日活用户 * 人均任务数 * 单任务沙箱数 * \frac{日均沙箱运行时长}{24} * 峰均比 * \boldsymbol{单沙箱物理核} * (1 + 编排开销)$$

其中：

$$单沙箱物理核 = vCPU \times 0.5 \times \frac{vCPU\ 平均利用率}{宿主机vCPU\ 利用率}$$

注：考虑到超线程，每个 vCPU 占用物理核数为 0.5

我们以 1 亿日活用户、人均每日 1 个任务作为跨场景对比的统一口径，四类场景基准参数如下：

图表 7：四类场景基准参数

各场景基准	个人 Agent	编程 Agent	浏览器 Agent	企业级 Agent
日活用户（万人）	10000	10000	10000	10000
人均任务数	1	1	1	1
单任务沙箱数	2	2	1	1
日均沙箱运行时长（小时）	0.25	1	1/6	1/3
峰均比	3	3	3	2
vCPU/沙箱	2	4	2	2
vCPU 平均利用率①	20%	12%	50%	25%
宿主机 vCPU 利用率②	50%	50%	50%	50%
编排开销	12.5%	12.5%	12.5%	12.5%

数据来源：金元证券研究所整理

注：

① vCPU 平均利用率：编程 Agent 取 12%考虑到依赖安装、编译构建等重负载环节；浏览器 Agent 因页面渲染负载重取 50%；个人、企业 Agent 分别取 20%、25%。

② 宿主机 vCPU 利用率：统一取 50%，高峰时段保留一半余量，用于吸收突发负载、容灾冗余等需求。

3. 测算结果：1 亿日活基准下合计需求 2022 万物理核，编程 Agent 贡献近七成

综合来看，在 1 亿日活用户、人均 1 个任务的基准下，个人、编程、浏览器、企业级 Agent 的物理核需求分别为 281.3 万核、1350 万核、234.4 万核、156.3 万核，合计约 2021.9 万核；其中编程 Agent 占比约 67%，对物理核的拉动最大：其日均运行 1 小时、单任务调用 2 个沙箱，峰值并发沙箱达 2500 万个，为四类之首。按 32 核、64 核、88 核 CPU 折算，四类场景合计分别需要 63.2 万颗、31.6 万颗、23.0 万颗 CPU。

图表 8：物理核测算结果及对应 CPU 需求

场景	峰值并发沙箱（万个）	单沙箱物理核（核）	物理核需求（万核）	32 核 CPU（万颗）	64 核 CPU（万颗）	88 核 CPU（万颗）
个人 Agent	625	0.4	281.3	8.8	4.4	3.2
编程 Agent	2,500	0.48	1350	42.2	21.1	15.3
浏览器 Agent	208.3	1	234.4	7.3	3.7	2.7
企业级 Agent	277.8	0.5	156.3	4.9	2.4	1.8
合计			2,021.9	63.2	31.6	23.0

数据来源：金元证券研究所整理

4. 弹性判断：编程兑现最快，企业级理论弹性最大

四类场景中，编程 Agent 兑现最快，企业级 Agent 理论弹性最大。沙箱的 CPU 需求由用户基数、人均任务、任务并行度与日均运行时长四

大因素共同决定，未来的弹性来自四个因素的持续上行：渗透率提升带动用户基数扩大；Agent 进入更多使用场景，人均任务增加；任务并行度与子 Agent 数量增加，单任务沙箱数上升；任务复杂度提升与常驻化拉长运行时间，同时抬高单沙箱的 vCPU 规格与利用率，放大单任务的 CPU 调用强度。

图表 9：四类场景增长因子的上行空间

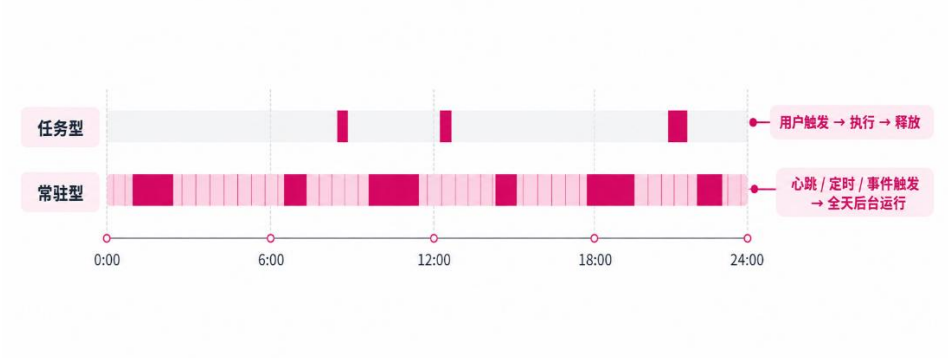
	个人	编程	浏览器	企业级
① 用户基数	高	高	中	高
② 人均任务	中	中	中	高
③ 任务并行度	中	高	低	高
④ 日均运行时长	高	高	低	高
综合弹性判断	较大	大·确定性最强	中	理论最大

数据来源：金元证券研究所整理

4.1 个人 Agent：常驻化与渗透率提升共同放大需求

弹性一：由任务型转向常驻型，沙箱运行频率由“按次”变为“全天候”。任务型 Agent 在用户发起请求后短时运行、完成即释放；常驻型 Agent 按心跳、定时与事件触发持续运行，监控、提醒、整理、跟踪等长期任务会产生持续性的沙箱占用，进而持续消耗 CPU 资源。如开源个人 Agent OpenClaw：其 Gateway 常驻后台，默认每 30 分钟巡检一次任务清单，无需用户触发即可执行；项目 GitHub 星标在 2026 年 1 月仍接近零，数周内升至 24.7 万，截至 10 月已接近 40 万。

图表 10：任务型与常驻型 Agent 的沙箱占用对比

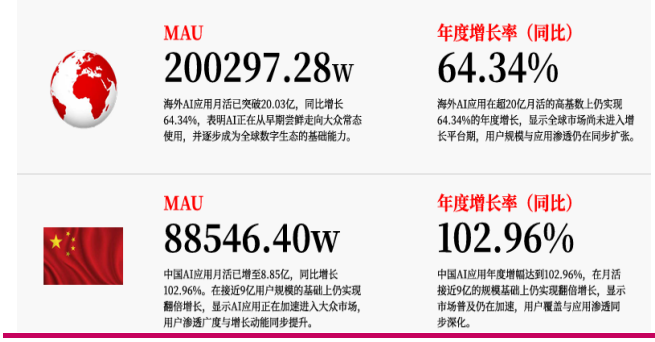


数据来源：Openclaw，金元证券研究所整理

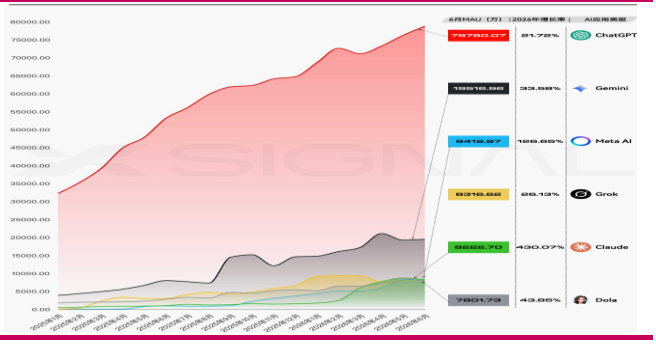
弹性二：渗透率提升带动日活扩大，大模型用户均为潜在的 Agent 用户。根据《Xsignal 全球 AI 应用行业半年报告 2026》，全球 AI 用户规模加速扩张，截至 2026 上半年全球 AI 应用 MAU 近 29 亿，其中海外 AI 应用 MAU 达 20 亿，同比增长 64.3%；国内 AI 应用 MAU 达 8.85

亿，同比增长 102.96%。此外，头部大模型用户基数仍保持稳健增长。截至 2026 年 6 月，ChatGPT APP MAU 达到 7.88 亿，上半年增长 21.72%；Claude APP MAU 达到 0.82 亿，上半年增长 430.07%。**庞大的大模型用户池为 Agent 提供了天然的转化基数：当“对话”升级为“代办”，每一位活跃用户都可能成为 Agent 的用户，进而产生大量的 CPU 资源消耗。**综合两条路径，个人 Agent 在运行时长与用户规模两个维度同时具备上行空间，弹性较大。

图表 11：全球和中国 AI 月活



图表 12：头部 AI 大模型用户规模趋势（截至 202606）

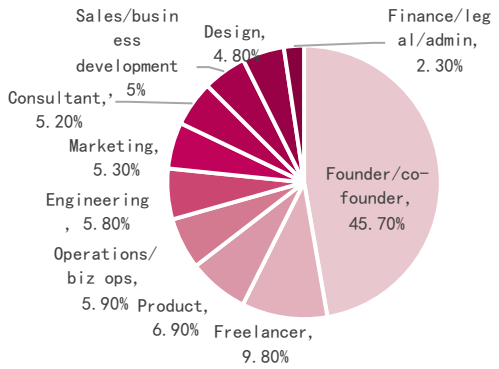


数据来源：《Xsignal 全球 AI 应用行业半年报告 2026》，金元证券研究所整理

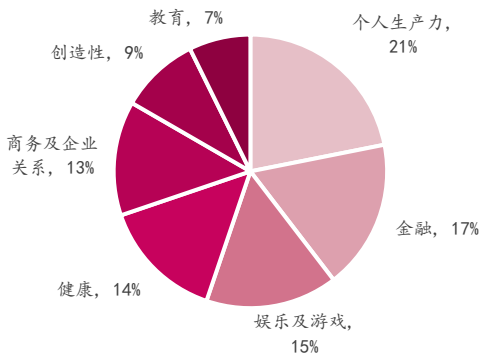
4.2 编程 Agent：用户边界扩大与任务复杂度提升正同步发生

弹性一：用户边界扩大：Vibe Coding 场景应用使“写代码的人”不再局限于专业开发者。根据 Lovable《The Build Economy》于 2026 年的调查数据统计，80%的用户从事非技术性工作。当前 Lovable 日均新建项目数超 10 万个，年项目量达 2500 万个，非技术性用户带来重要增量。而根据 All About Cookies 做的一份美国 AI 调研中，近 60%的美国人具备开发应用、工具或网站的想法，当技术不再成为障碍时，其中 80%的用户愿意去开发对应的工具或产品。AI 能力提升降低 Coding 的开发门槛，个人用户的定制化编程需求将带动编程 Agent 用户规模快速提升。

图表 13：Lovable 4/5 用户从事非技术性工作



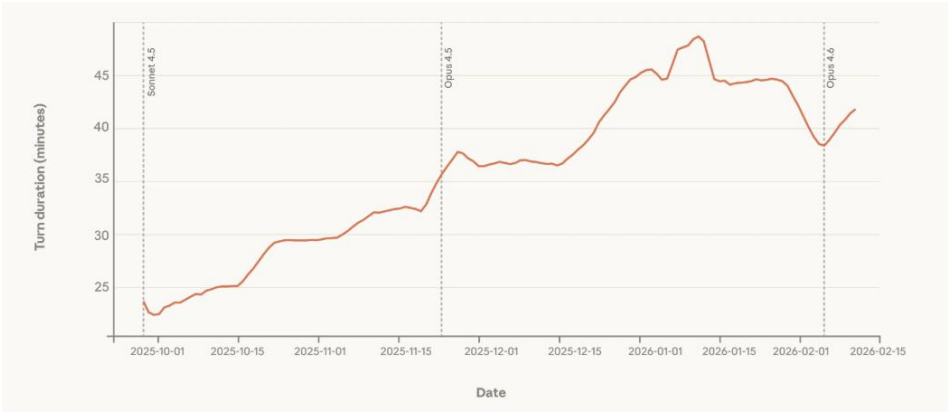
图表 14：美国 AI 用户想要创作的应用类型



数据来源：《Lovable： The Build Economy》，All About Cookies，金元证券研究所整理

弹性二：任务复杂度提升：编程任务由单文件补全走向仓库级工程交付，单任务的运行时间与资源强度同步抬升。Anthropic 披露，Claude Code 最长一档会话的单轮自主运行时长，在 2025 年 10 月至 2026 年 1 月间由不足 25 分钟升至 45 分钟以上，内部最难任务的成功率提高一倍。任务越复杂，沙箱内的依赖安装、编译构建与回归测试越密集，这类负载以 CPU 为主，单任务的 CPU 占用与沙箱存续时间随之上升。

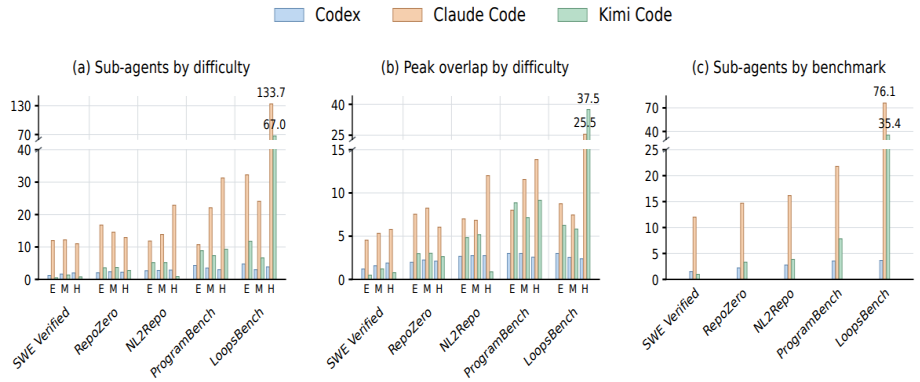
图表 15：Coding 单轮自主运行时长提升至 45 分钟



数据来源：Anthropic，金元证券研究所整理

弹性三：并行项目增加：复杂的编程任务将被拆解给多个 Agent 同时推进，单任务并发沙箱数成倍增加。根据相关测试数据，当任务难度从简单提升至困难，单任务平均创建的子 Agent 数量由 6.69 个提升至 15.84 个，峰值并发子 Agent 由 3.66 个提升至 6.7 个，困难任务峰值并发规模较简单任务提升约 83%。对应单任务所需的并发沙箱数量相应上升。综合来看，编程 Agent 有望实现用户边界、任务复杂度与并行度同步上行，兑现确定性最强。

图表 16：任务难度提升，Coding Agent 并发规模扩大



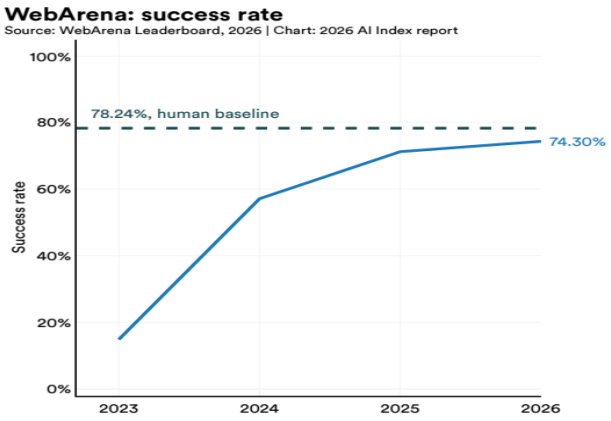
数据来源：《When Sub-Agents Work in Parallel: The Promises and Pitfalls of Dynamic Concurrency in Long-Horizon Coding Tasks》，金元证券研究所整理

4.3 浏览器 Agent：单沙箱 CPU 强度最高，成功率提升驱动放量

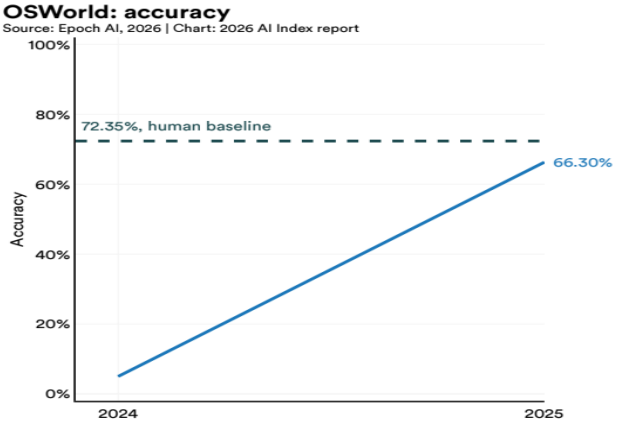
浏览器 Agent 的需求弹性主要来自任务执行能力提升，以及由此带来

的使用场景扩张。浏览器 Agent 需要通过网页解析、JavaScript 执行、DOM 渲染及页面交互完成复杂任务，其执行过程主要依赖 CPU 等通用计算资源。此前，网页 Agent 的任务执行成功率较低，限制了其在实际业务流程中的应用。随着模型推理及计算机操作能力提升，浏览器 Agent 的可用性明显改善。根据 Stanford HAI《2026 AI Index》，WebArena 基准中领先 Agent 的任务成功率由 2023 年的约 15%提升至 2026 年初的 74.3%，接近 78.24%的人类参考水平；OSWorld 基准中，领先 Agent 的计算机操作成功率亦由早期约 5%提升至 2025 年的 66.3%，接近 72.35%的人类参考水平。随着任务执行成功率提升，网页检索、表单填写、跨站信息处理等操作有望逐步成为个人及企业 Agent 的通用执行能力。

图表 17：WebArena：Agent 任务成功率逼近人类



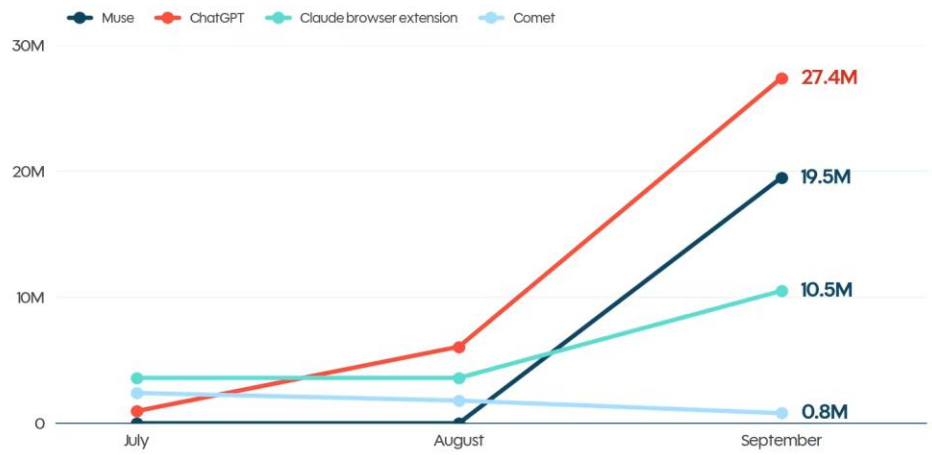
图表 18：OSWorld：Agent 计算机操作成功率快速提升



数据来源：Stanford HAI《2026 AI Index Report》，金元证券研究所整理

实际使用量增长进一步验证浏览器 Agent 需求释放，云端执行有望打开 CPU 增量空间。根据 DataDome，其监测网络中 Agent 浏览器请求量由 2026 年 8 月的 1250 万次增长至 9 月的 5820 万次，环比提升至约 4.7 倍；即使剔除 Meta Muse 影响，相关请求量仍增长至约 3.1 倍。其中 ChatGPT 云端浏览器与 Muse（运行于专属云端虚拟机）合计占 9 月请求量的约 81%，请求量的增长主要由云端执行承接。我们认为，随着浏览器 Agent 由信息查询向表单操作、跨系统业务处理等复杂任务延伸，云端浏览器实例数量及运行需求有望进一步增长，带动服务器 CPU 资源需求提升。

图表 19：浏览器 Agent 请求量快速增长

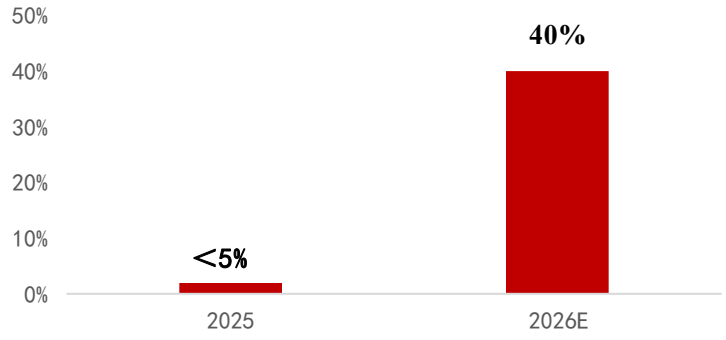


数据来源：Datadome，金元证券研究所整理

4.4 企业 Agent：渗透率与任务复杂度同步抬升，理论弹性最大

企业 Agent 加速渗透，规模化部署与应用场景拓展有望打开长期增长空间。随着 Agent 技术逐步成熟，企业级 Agent 正由单点试用向规模化部署延伸，应用场景及业务流程覆盖范围持续扩大。根据 IDC 于 2026 年发布的企业调查，95.3%的受访企业已在生产环境中运行至少一项 Agent 工作流，平均部署约 11 项，Agent 正逐步融入企业实际业务流程。从应用渗透来看，Gartner 预计，到 2026 年底，约 40%的企业应用将集成特定任务型 AI Agent，较 2025 年不足 5%的水平显著提升。随着 Agent 在财务、客服、IT 运维及企业管理等场景中加速落地，企业自动化流程数量及任务执行频次有望持续增长，并推动多任务并发执行需求提升，为沙箱及服务器 CPU 带来增量需求。

图表 20：企业应用集成任务型 Agent 的比例快速提升

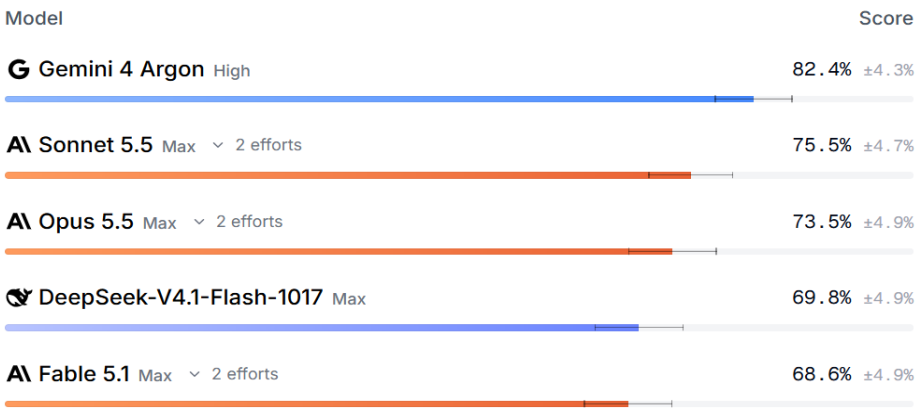


数据来源：Gartner，金元证券研究所整理

企业 Agent 已初步具备处理复杂专业工作的能力。随着大模型推理及工具调用能力持续提升，企业 Agent 正由信息检索、问答及审批提醒等轻量任务，逐步向财务分析、批量文档处理、跨系统业务操作等复杂场景拓展。2026 年 9 月，Mercor 发布 APEX-Agents 1.1 评测，涵盖投资银行、管理咨询及企业法律三大专业领域的 240 项真实工作任

务，要求 Agent 在邮件、文档、表格等多个应用之间获取信息并完成长流程任务。截至 2026 年 10 月，领先模型在该评测中的单次任务完成率（Pass@1）已达 82.4%，Agent 处理复杂专业工作流程的能力正持续提升。

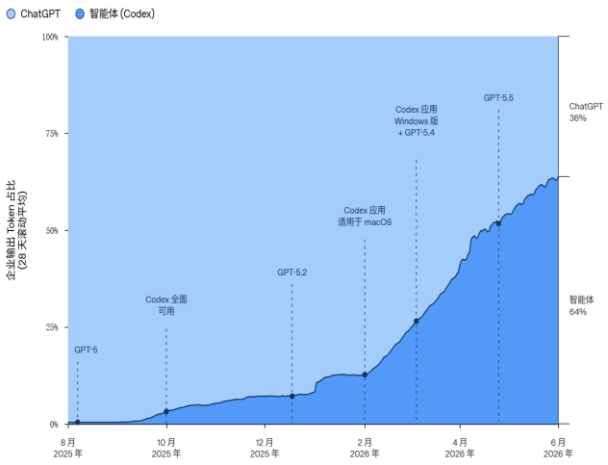
图表 21：APEX-Agents 榜单



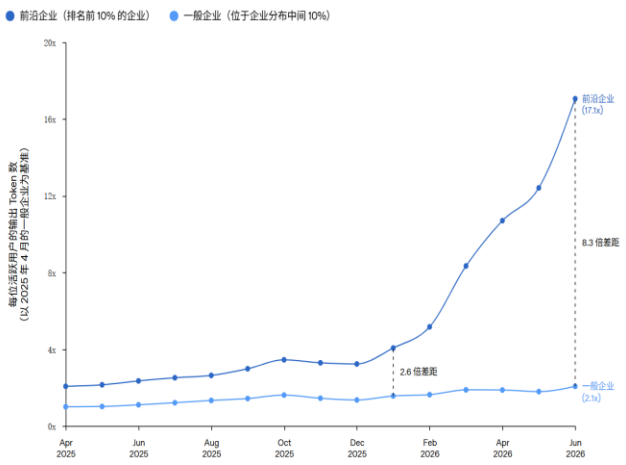
数据来源：Mercor，金元证券研究所整理

领先企业正逐步将 Agent 融入企业核心工作流。OpenAI 于 2026 年 8 月发布的《Enterprise Signals》显示，截至 2026 年 6 月，企业客户 Agent 的输出 Token 已占合计输出 Token 的 64%；高使用强度企业与普通企业的人均输出 Token 倍数由年初的 2.6 倍提升至 8.3 倍，侧面反映领先企业正加大 AI 在复杂任务中的使用深度。综合来看，企业级 Agent 正处于应用渗透、并行工作流与单任务时长的快速抬升期，理论弹性最大。

图表 22：企业 tokens 用量中 Agent 占比快速提升



图表 23：领先企业 token 使用差距已拉大到 8.3 倍



数据来源：OpenAI，金元证券研究所整理

三、投资建议：建议关注服务器 CPU

投资线索一：GPU 配套 CPU 核数持续提升，为服务器 CPU 需求提供确定性增量。随着 AI 算力需求由训练向推理拓展，数据处理、任务调度及内存管理等环节对 CPU 的要求持续提高，增加 CPU 配置有助于提升 GPU 利用效率。从 NVIDIA 硬件架构看，DGX H100 采用 8 颗 GPU 搭配 2 颗 56 核 CPU，单 GPU 对应 14 个 CPU 核心；GB300 NVL72 采用 72 颗 GPU 搭配 36 颗 72 核 CPU，单 GPU 对应 36 个 CPU 核心；Vera Rubin NVL72 进一步采用 72 颗 GPU 搭配 36 颗 88 核 Vera CPU，单 GPU 对应 44 个 CPU 核心。考虑到 GPU 价值量较高，通过增加 CPU 投入改善 GPU 利用效率具备较强经济性，我们认为配套 CPU 的核心数、带宽及性能要求有望同步提高，GPU 配套 CPU 需求将成为行业增长的重要支撑。

投资线索二：Agent 沙箱打开独立于 GPU 配套的 CPU 增量空间，提供需求弹性。Agent 将模型能力由内容生成拓展至实际任务执行，代码运行、工具调用及浏览器操作等环节均产生独立的 CPU 计算负载。在 1 亿日活的统一基准下，四类场景合计需要约 2022 万物理核，折合 31.6 万颗 64 核 CPU。弹性排序上，编程 Agent 兑现最快，企业级 Agent 理论弹性最大，个人 Agent 的常驻化与浏览器 Agent 的成功率提升亦将带来增量。产业端已开始布局独立的 CPU 执行资源：NVIDIA 推出集成 256 颗 Vera CPU 的独立 CPU 机架，可支持超过 2.25 万个并发 CPU 执行环境，面向 Agent 工具调用、沙箱运行及任务编排等场景。

综合判断：GPU 配套提供确定性，Agent 沙箱提供弹性，建议关注服务器 CPU。两条线索叠加，CPU 需求有望由依附 GPU 配置进一步向独立任务执行扩展，为服务器 CPU 打开新的长期增长空间。我们建议关注服务器 CPU 及相关产业链的长期增长机会。

相关公司（以下公司均未评级）：
美股：Intel、AMD 等
A 股：海光信息、禾盛新材等

图表 24：相关公司 CPU 产品线简介

公司	系列/产品	架构/工艺	核心规格	AI/Agent 相关定位	发布与供货进度
AMD	EPYC 9575F	x86；TSMC 3-4nm	64 核；3.3/5.0GHz；L3 256MB；TDP 400W；12 通道 DDR5；PCIe 5.0；512 位 AVX-512	专为 GPU AI 方案打造的 AI 主机节点 CPU	2024 年 10 月发布并供货，售价 11791 美元
	EPYC 9006 SP7	x86；TSMC 2nm	最高 256 核/512 线程；16 通道 DDR5-8000/MRDIMM-12800；PCIe Gen6	面向高密度服务器及大型智能 AI 部署，每秒处理 tokens 数量为上一代的 1.8 倍	2026 年 5 月量产爬坡；SP7 平台 2026Q4 供货，顶配 EPYC 9996 售价 14904 美元
Intel	Xeon 6776P	x86	64 核/128 线程；2.3/3.9GHz；336MB 缓存；8 通道 DDR5-6400/MRDIMM-8800；88 条 PCIe 5.0；AMX	英伟达 B300 集成 Xeon 6776P，基于 PCT 技术，能够动态对核心进行优先级排序，更好的处理 AI 负载	2025Q2 发布
	Xeon 6+	x86；Intel 18A	最高 288 核；12 通道 DDR5；96 条 PCIe 5.0；TDP 最高 450W	较 Sierra Forest 性能+126%、能效比+55%	2026 年 6 月 Computex 推出
海光信息	新一代 C86	x86 兼容架构	据媒体转述：最高 128 核/512 线程（SMT4）；IPC 较上代提升超 15%；104 条 PCIe 5.0；AVX512	原生支持 INT8 和 BF16 的 AI 加速	/
	7000、5000、3000 系列	x86 兼容架构	7000 面向数据中心等高端通用服务器；5000 面向通用服务器；3000 面向工作站与边缘计算	/	已量产销售
禾盛新材（熠知电子）	熠知电子 TF9000 系列	Arm V9；异构合封	较前代核心性能提升 30%、成本下降 30%，内存带宽、PCIe 5.0 带宽及内存总容量分别提升 200%、100%、300%。	性能全面对标 NVIDIA Grace 系列	2026 年 1 月发布

数据来源：AMD 投资者关系，Phoronix，Intel，Digitalcitizen，Hothardware，熠知电子，金元证券研究所整理

四、风险提示

（1）Agent 商业化与渗透率不及预期。若 Agent 在任务执行能力、可靠性及使用成本等方面的改善不及预期，可能影响个人及企业用户的使用意愿和应用场景拓展，进而制约沙箱部署规模及 CPU 需求释放。

（2）沙箱资源复用效率提升。随着沙箱调度、休眠唤醒及资源复用等技术持续优化，单位沙箱的实际物理 CPU 需求可能下降，部分抵消 Agent 并发规模及任务执行量增长带来的硬件增量需求。

（3）Agent 执行向本地端迁移。随着终端设备算力提升及本地 Agent 技术成熟，部分代码执行、文件处理及浏览器操作等任务可能由云端转向本地，降低云端沙箱部署需求，进而影响沙箱对于 CPU 增量需求。

金元证券股票投资评级标准：

买入：预期未来 6 个月内股价收益率超越沪深 300 指数的涨跌幅 15%以上；

增持：预期未来 6 个月内股价收益率超越沪深 300 指数的涨跌幅 5%～15%；

中性：预期未来 6 个月内股价收益率相对沪深 300 指数涨跌幅差异在-5%～+5%之间；

减持：预期未来 6 个月内股价收益率弱于沪深 300 指数的涨跌幅 5%以上。

金元证券行业投资评级标准：

强于大市：预期行业相关指数在未来 6 个月内超越沪深 300 指数表现；

中性：预期行业相关指数在未来 6 个月内基本与沪深 300 指数表现持平；

弱于大市：预期行业相关指数在未来 6 个月内明显弱于沪深 300 指数表现。

金元证券机构销售负责人：詹宝强
电子邮件：zhanbq@jyzq.cn

免责声明

本报告由金元证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格）制作。本报告所载资料的来源及观点的出处皆被金元证券认为可靠，但金元证券不保证其准确性或完整性。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专业财务顾问的意见。对依据或者使用本报告所造成的一切后果，金元证券及/或其关联人员均不承担任何法律责任。投资者需自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

本报告所载的信息、材料或分析工具仅提供给阁下作参考用，不是也不应被视为出售、购买或认购证券或其他金融工具的要约或要约邀请。该等信息、材料及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，金元证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

金元证券的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。金元证券没有将此意见及建议向报告所有接收者进行更新的义务。金元证券的自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

在法律许可的情况下，金元证券可能会持有本报告中提及公司所发行的证券头寸并进行交易，也可能为这些公司提供或争取提供投资银行业务服务。因此，投资者应当考虑到金元证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。

本报告的版权仅为金元证券所有，未经书面许可任何机构和个人不得以任何形式转发、翻版、复制、刊登、发表或引用。