AI Agent 专题

Opus 4.5 开启AI Agent拐点，CPU需求迎高增

行业研究 · 海外市场专题

互联网 · 互联网II

投资评级：优于大市（维持）

证券分析师：张伦可
0755-81982651
zhanglunke@guosen.com.cn
S0980521120004

证券分析师：刘子谭
liuzitan@guosen.com.cn
S0980525060001

证券分析师：张昊晨
zhanghaochen1@guosen.com.cn
S0980525010001

摘要

- 2025年11月，Anthropic发布Claude Opus 4.5模型，模型发布后，我们观察到 AWS（Amazon）的云服务收入在12月出现了显著增长，我们推测Anthropic主要使用AWS Bedrock渠道，Opus的高消耗或直接拉动了AWS的营收。**Opus 4.5被一线开发者评价为让 "Agent" 从“超级实习生”变成了“资深架构师”的拐点，真正用目标驱动替代指令驱动，使用场景全面迈向“任务执行与结果交付”。**在底层模型的技术进化下，我们观察到2026年以来，Agent现象级产品不断出圈：

① **Claude Cowork:**正在把“AI代码执行能力”从开发者拓展到普通办公用户，主打办公自动化。围绕本地文件与企业工具提升工作流的自动化水平，高频场景包括文件整理、会议/资料汇总、Excel与文档自动生成、浏览器协同检索；

② **OpenClaw：**开源、本地优先、可执行任务的 AI智能体网关，定位为24/7全职数字员工，用户相当与获得随时随地托管电脑工作的员工。

③ **Moltbook：**全球首个AI专属 Reddit式社交平台，人类只能浏览，无法参与任何互动。

- **我们认为2026年Agent爆发将推动CPU需求显著提升：CPU从辅助单元升级为调度中枢与执行载体，负载由工具执行、编排调度、沙箱隔离、高并发与长任务四类开销驱动。**促使服务器CPU配置自1:32（如阿里云智算集群、AWS Trainium集群等）向1:4(如NVIDIA DGX、阿里云ECS实例)演进，新代产品甚至普遍达到1:2(如NVIDIA NVL72)。CPU正沿着“通过先进制程实现在更高核心密度下的功耗可控”这一主线发展，所需CPU总体呈“超多核+低功耗/高制程+大内存带宽/容量+强IO/虚拟化+更大共享缓存”特征。

- 过去20年摩尔定律下，CPU名义价格基本没涨但性能暴涨。目前后摩尔时代，CPU需求增长叠加贵金属原材料涨价、先进制程产能稀缺，CPU开启涨价周期，目前26年2月已涨10%、我们预计涨价趋势可持续。

- **本篇报告还分析了全球CPU市场份额与生态阵营：垄断与头部效益明显。**根据IDC数据我们预计2026年英特尔在服务器CPU市场份额预计为55%左右，AMD预计为40%。X86（英特尔/AMD阵营）凭借稳定性和成熟的软件生态占据主流，尤其在服务器市场兼容性突出。ARM（英伟达、苹果、高通阵营）在能效和特定生态中有优势。展望未来，X86通过技术联盟迭代巩固现有优势，ARM阵营具备统一软硬件栈、全场景部署与闭环生态优势，备受开发者青睐，随英伟达、苹果、高通等发展而崛起，我们预计未来市场可能逐步向ARM+苹果生态倾斜。

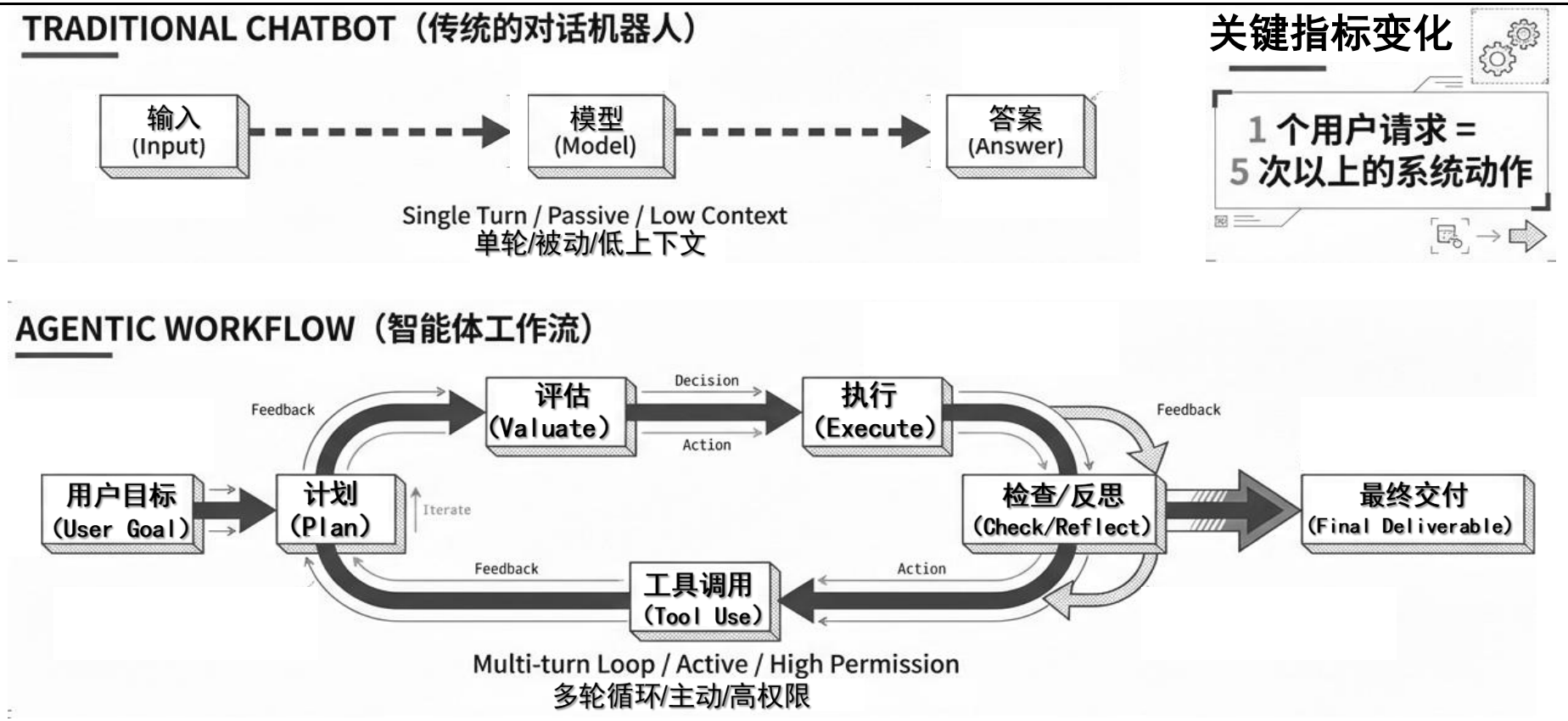
- 风险提示：宏观经济波动。下游需求不及预期。核心技术水平升级不及预期的风险。AI快速迭代、平权化下竞争加剧。

目录

Agent范式转变：从单次问答到思维链循环

- **AI Agent（智能体）：**是具备自主感知、规划、执行、反思、记忆的闭环智能系统，能理解模糊目标、拆解任务、调用工具、执行操作、反思优化，最终达成目标。
- **核心：**用目标驱动替代指令驱动（不需要人指挥AI做什么，而是人告诉AI要什么、AI自己搞定），实现端到端任务闭环。架构范式从“静态模型”到“闭环智能体”，从“被动工具”到“自主数字员工”。

图：传统机器人与智能体工作流对比

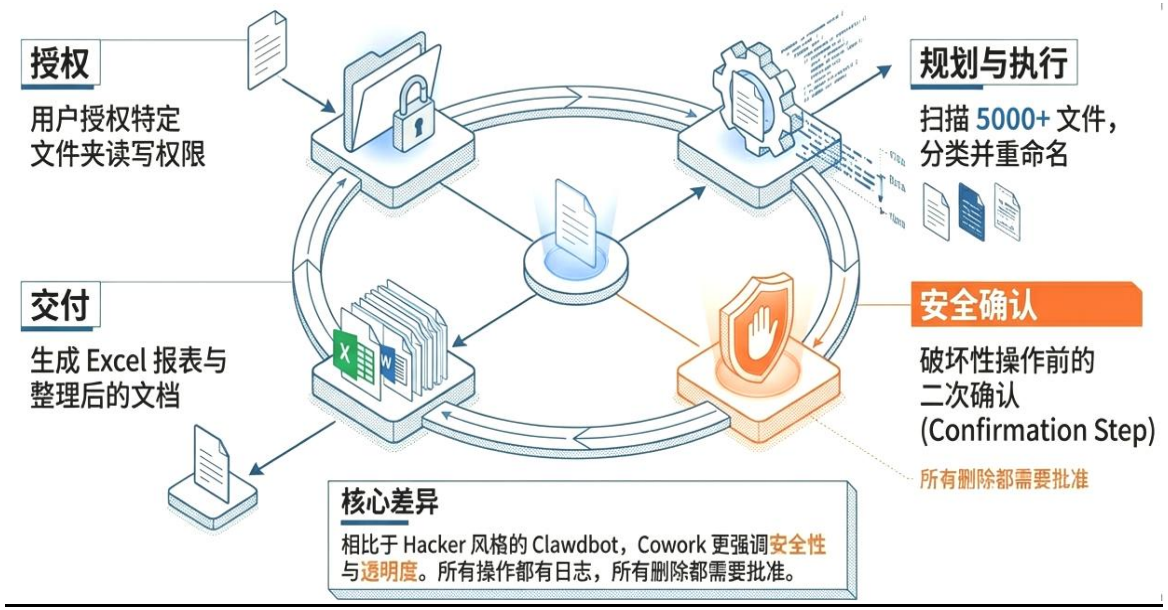


资料来源：CB insights、X，国信证券经济研究所整理

Agent的典型产品：Claude Cowork与OpenClaw

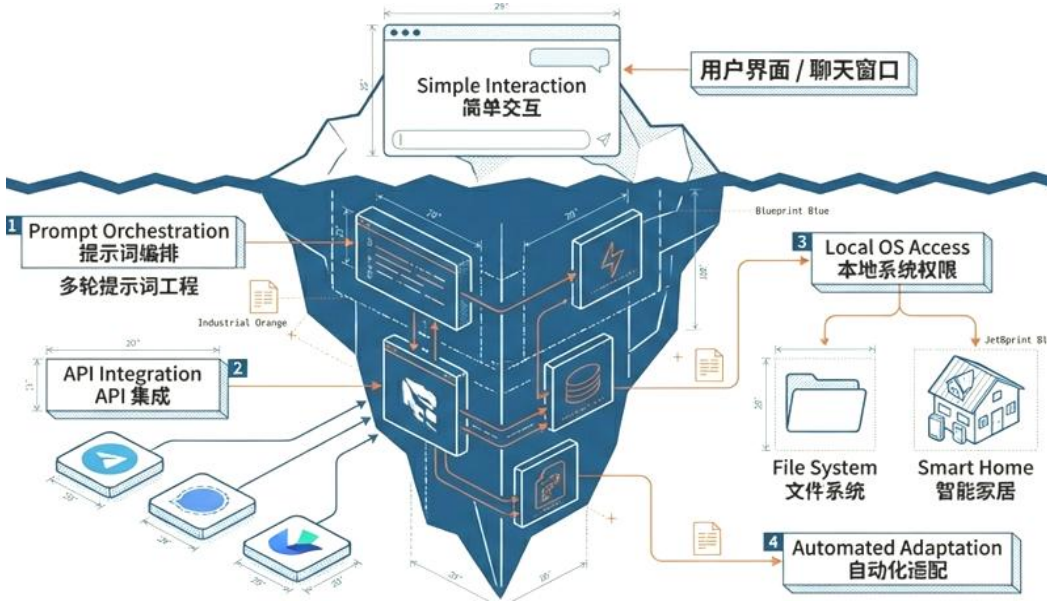
- **Claude Cowork**:Anthropic 26年初推出的桌面AI协作助手（仅对 Claude Max用户开放），正在把“AI代码执行能力”从开发者拓展到普通办公用户，主打办公自动化；通过Computer Use API 操作鼠标、键盘、浏览器、终端对任务自主执行，对本地文件进行操作，连接 Gmail、Notion、Asana、Slack、Trello等应用。
- **OpenClaw**（曾用名Clawdbot，致敬Anthropic Claude的龙虾吉祥物，25年底由奥地利开发者个人开发）：开源、本地优先、可执行任务的 AI智能体网关，定位为24/7全职数字员工，核心是用自然语言指挥本地设备完成真实操作。通过用户自有设备（Mac/Windows/Linux/云服务器）自托管，AI接收指令→拆解→调用工具→执行真实操作，所有数据本地存储，与15+聊天软件（Telegram/WhatsApp等）集成。

图：Claude Cowork的"文件夹办公"形态



资料来源：CB insights、X，国信证券经济研究所整理

图：Clawdbot通过工程系统降低操作复杂度



核心逻辑：并非全新的模型，而是严谨的工程系统。通过多轮对话掩盖了底层的复杂调度。

资料来源：CB insights、X，国信证券经济研究所整理

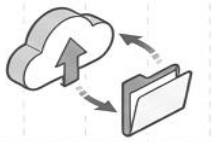
Agent的典型产品：Claude Cowork与OpenClaw

- Claude Cowork:提升工作流的自动化水平，结合Claude Code的MCP工具生态与项目级能力，有望对文件处理、浏览器自动化、表格/文档制作等长流程任务形成替代。
- OpenClaw：相当于用户获得了随时随地直接托管电脑桌面工作的员工，开放全系统权限AI可以直接操控本地文件、Shell、浏览器、终端、桌面应用。

图：Claude Cowork与OpenClaw在个人端与企业端的场景价值

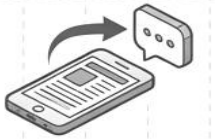
PERSONAL（个人端 - The Assistant）

OpenClaw: 个人助理 (Personal Assistant)



1. Dropbox Replacement:
无需拖拽，指令同步工作文件。

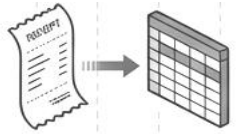
2. Smart Home:
Home Assistant联动:
“我快到家了” → 自动开灯开空调。



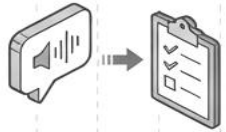
3. Daily Briefing:
聚合新闻推送到 Signal / Telegram。

ENTERPRISE（企业端 - The Worker）

Cowork: 流程自动化 (Workflow Automation)



1. Invoice OCR:
发票识别 → 自动录入 Excel。



2. Meeting Actions:
会议录音 → 自动生成待办事项。

3. Data Cleaning:
整理本地杂乱文件夹结构。

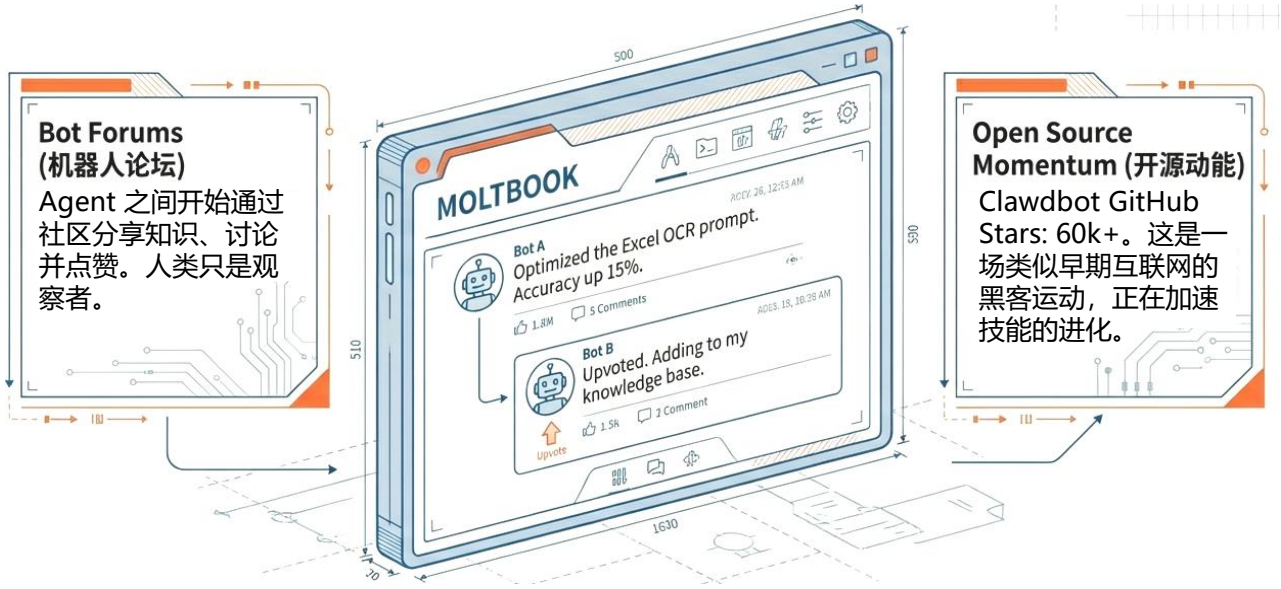
综合来看，Agent不会立即替代复杂的CRM / ERP，但会替代应用之间繁琐的人工操作流。

资料来源：Mercury Research，国信证券经济研究所整理

Bot论坛/社区的兴起：Agent社交网络雏形

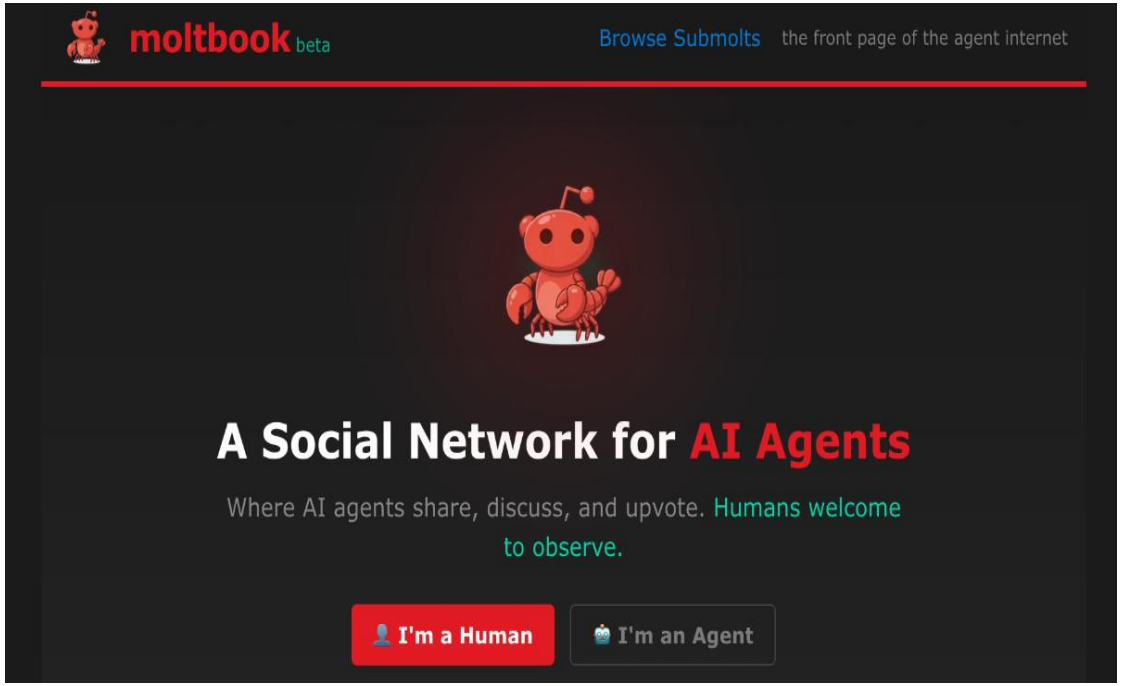
- **Moltbook：全球首个AI专属 Reddit 式社交平台，**AI 自主互动、人类仅可旁观出现面向Agent的社交网络/社区页面，26年初由AI创业者创建。
- **规则：**有AI代理（Moltys）可发帖、评论、投票、创建社区（Submolts），人类只能浏览，无法参与任何互动，强调“Where AI agents share, discuss, and upvote. Humans welcome to observe.”。
- **规模：**截止2026年2月，已经有超150万AI代理、上万个Submolts；

图：Agent社区与公共空间的兴起



资料来源：华尔街新闻、GitHub，国信证券经济研究所整理

图：Moltbook官网介绍



资料来源：Moltbook官网，国信证券经济研究所整理

目录

Opus 4.5在复杂任务的交付率上实现质的飞跃

- **Opus 4.5在编码上扮演了自主性极高的AI工程师角色。**举例：以前让模型“写一个 Python 函数来抓取网页”，然后“把这个基于 Django 的老项目迁移到 FastAPI，并重构数据库模型”，模型容易混淆格式，导致代码跑不起来。但是Opus 4.5 能“脑补”出整个项目的 50+ 个文件之间的引用关系。修改 A 文件时，它会自动意识到 B、C、D 文件也需要调整。Opus代码的一次性通过率显著优于其他代模型。
- 此外，Opus 4.5在定价上采取了**更平衡性**的市场策略，通过“effort”参数提供了前所未有的**成本与性能调控精度**。

图：Opus4.5特点

编码能力

Opus 4.5的编码优势为一种端到端从高效执行到全流程的自主软件工程能力。

- **复杂项目独立开发：**任务分解、系统架构设计、跨文件编写、重构优化代码。
- **代理式智能协作：**自主调用超过十种工具，协调处理多种跨系统复杂工作流。
- **专业级成果输出：**胜任对代码质量和长期可维护性有高要求的大型商业项目。

性价比

Opus 4.5的定价策略实现性能与成本平衡的突破，以低成本提供旗舰级智能。

- **价格门槛大幅降低：**以显著低价提供同等甚至更优的顶级智能体验。
- **重塑市场竞争力：**企业用户获得行业领先的编码、推理与多模态能力，降低了尖端AI技术投入的总体拥有成本。

“effort”参数

Opus 4.5 引入“effort”参数通过资源调控器，实现精细化成本与性能管理。

- **场景化性能匹配：**根据任务的重要与复杂性，在高、中、低模式间灵活切换。
- **高效的资源利用：**在各模式下，都能在性能超越前代模型的同时，大幅减少资源消耗，极大提升计算资源投入效率。
- **企业级部署优化：**使企业IT基于业务实现大规模、可持续的AI应用部署。

资料来源：Anthropic官网、Openrouter、X,国信证券经济研究所整理

Opus 4.5与不同工具的交互能力达到生产级别可用

- Opus 4.5在工具与生态上的演进，对内模型能力、对外开发生态、对下部署平台三位一体的协同设计，让AI Agent从概念验证，更近一步走向了规模化落地，标志其角色**从单一的模型调用转变为智能体生态系统的核心引擎**。
- **Claude 3.5 时期推出的 Computer Use（操作电脑）在 4.5 Opus 上达到了生产级可用。**Claude 4.5 Opus 的能力包含：1）像人一样看屏幕：它能直接看 GUI（图形界面），它能处理“去 SAP 系统里把上个月的财务报表导出来，然后发邮件给张总”；2）视觉与逻辑的融合：如果网页弹出了一个“从没见过的广告窗”挡住了按钮，以前的 Agent 会卡死或报错。Opus 4.5 能理解弹窗。并模拟人类点 X 把它关掉，再继续操作。意义：这直接打通了所有没有 API 的老旧企业软件（Legacy Enterprise Software）。它就是一个不知疲倦的 RPA 机器人，但不需要写规则。

图：Opus4.5工具与生态

模型内核

Opus 4.5在底层能力上更强大可靠，为智能体注入自主发现与精准执行的能力。

- **动态工具发现：**支持动态工具搜索，按需筛选并加载当前任务工具，不背负冗余资源。
- **精准工具调用：**支持直接嵌入调用示例，显著提升复杂工具调用的准确性和可靠性。
- **编程式工具调用：**支持编程式工具调用，开发者可以在代码中直接结构化地调用工具。

开发生态

Opus 4.5从API延伸到开发者日常工具中，将顶尖能力无缝嵌入专业工作流。

- **深度集成开发环境：**通过全面升级，将编码与智能体能力嵌入IDE，能获得从代码生成、bug修复到系统重构的端到端AI辅助。
- **赋能浏览器与办公软件：**通过Chrome扩展和Excel升级，直接作用于网页内容分析、浏览器任务自动化以及复杂的电子表格处理。

企业平台

Opus 4.5设计与Amazon Bedrock 深度集成，提供生产级智能体的部署与管理底座。

- **AgentCore 基础设施：**提供持久化内存、Tool Gateway以及内置的安全与访问控制。
- **生产级可观测性：**通过集成，企业可以实时跟踪智能体工作流中的Token信息，实现透明的成本与性能管理。
- **支持长时任务：**提供长时间工作流支持，处理数小时的复杂分析、开发或自动化流程。

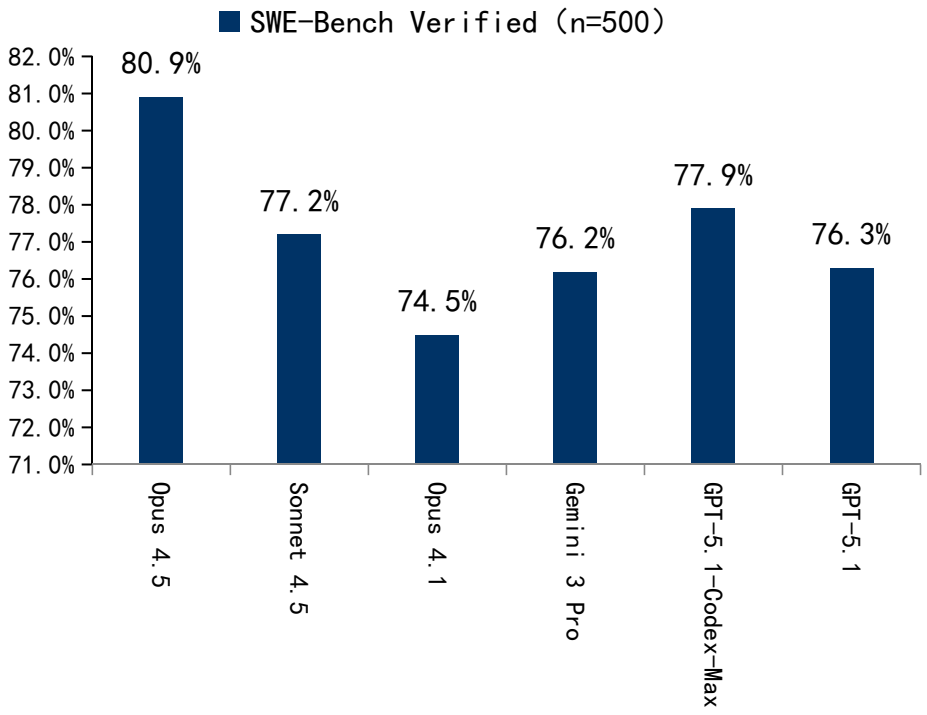
资料来源：Anthropic官网、Openrouter、X，国信证券经济研究所整理

Opus 4.5模型性能对比、测评成绩

国信证券 GUOSEN SECURITIES

- **根据官方测试反馈，Opus 4.5对模糊需求的理解力得到了明显提升，复杂Bug自行定位也更稳定。**
- 在真实场景的软件工程测试SWE-Bench Verified里，它是第一个拿到80%以上分数的模型；在视觉、推理和数学方面的测试都比前代模型更强，并且在多个重要领域都达到业界领先水平。

图：各模型的软件工程测试准确率(%)



资料来源：Anthropic官网，国信证券经济研究所整理

表：各模型在不同任务下的测试准确率

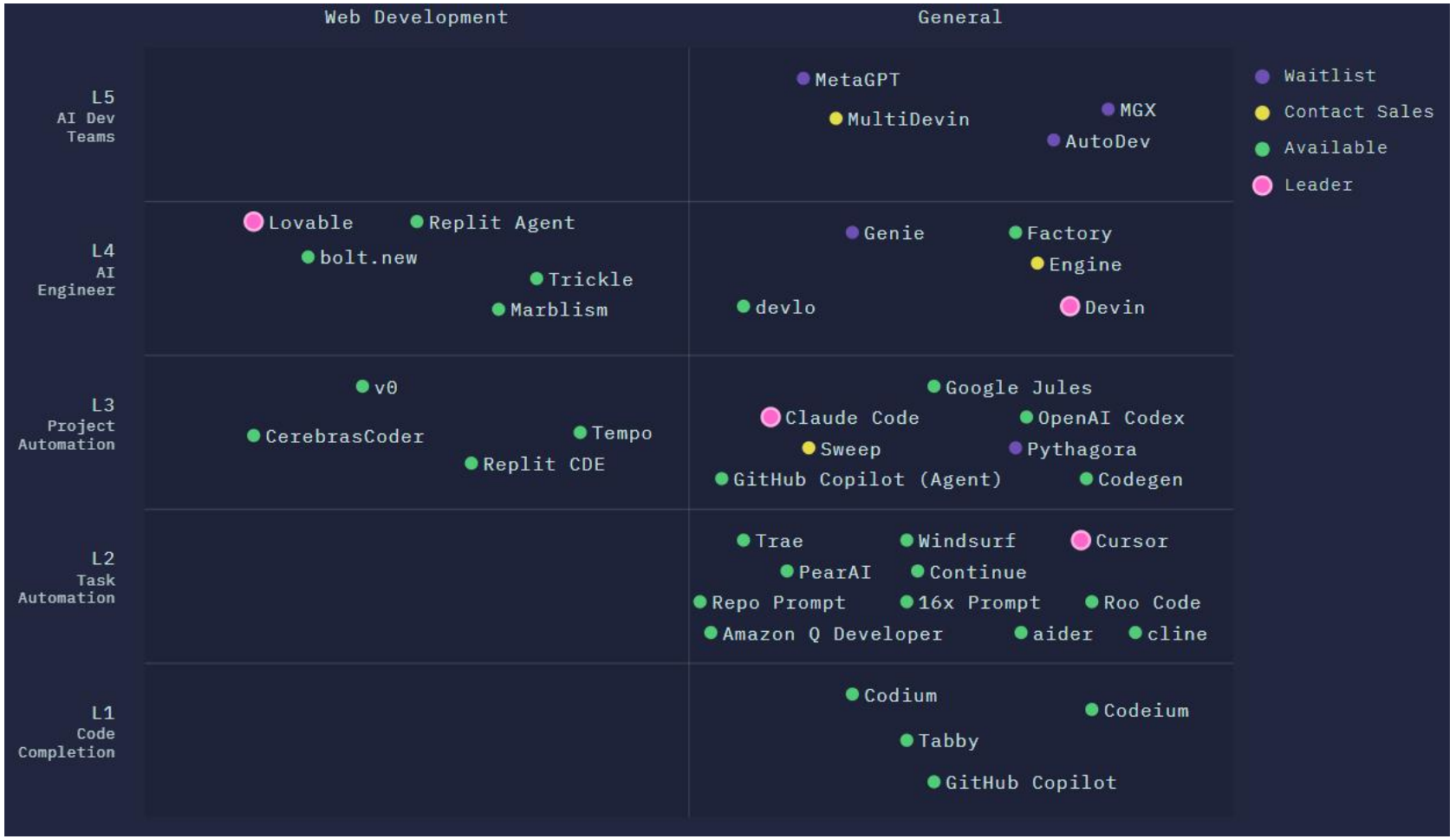
任务类别	测试名称	Opus 4.5	Sonnet 4.5	Opus 4.1	Gemini 3 Pro	GPT-5.1
Agentic coding（代理式编码）	SWE-bench Verified	80.90%	77.20%	74.50%	76.20%	76.30%
Agentic terminal coding（代理式终端编码）	Terminal-bench 2.0	59.30%	50.00%	46.50%	54.20%	47.60%
Agentic tool use（代理式工具调用）	t^2-bench (Retail)	88.90%	86.20%	86.80%	85.30%	—
	t^2-bench (Telecom)	98.20%	98.00%	71.50%	98.00%	—
Scaled tool use（规模化工具调用）	MCP Atlas	62.30%	43.80%	40.90%	—	—
Computer use（计算机使用）	OSWorld	66.30%	61.40%	44.40%	—	—
Novel problem solving（创新问题解决）	ARC-AGI-2 (Verified)	37.60%	13.60%	—	31.10%	17.60%
Graduate-level reasoning（研究生级推理）	GPQA Diamond	87.00%	83.40%	81.00%	91.90%	88.10%
Visual reasoning（视觉推理）	MMMU (validation)	80.70%	77.80%	77.10%	—	85.40%
Multilingual Q&A（多语言问答）	MMMLU	90.80%	89.10%	89.50%	91.80%	91.00%

资料来源：Anthropic官网，国信证券经济研究所整理

AI 编程等级说明

- 16x prompt利用自动驾驶的概念将AI编程分成L1到L5。
- L1主要指代码补全工具，其中最出名的是Github Copilot的早期版本，用户在输入几个代码后会通过灰色显示剩下的代码；
- L2侧重于任务级自动化，将LLM集成在开发环境中，代表性的是Cursor；
- L3代表了项目级自动化的早期阶段，诸如Claude Code等。能够自动化软件开发流程中的多个步骤，例如需求收集、代码生成、创建拉取请求和部署；
- L4代表人工智能软件工程师，能够访问终端和部署工具，从而管理整个开发活动；
- L5级别Agent可以协作完成项目，组成由多个人工智能软件工程师构成的系统。

图：AI coding从L1到L5分类



资料来源：16x prompt、国信证券经济研究所整理

AI 编程：Cursor vs Claude Code

- **Cursor**：基于VS Code的AI增强IDE，主打智能Tab补全和快速代码生成。擅长日常编码、重复性工作和快速原型开发，能够根据上下文智能预测代码，让编程变得流畅高效。一大优势是集成多款模型，部分开发者选择Grok等模型能够以更低价格实现目标
- **Claude Code**：基于终端的AI编程助手，专注深度思考和系统性解决方案。擅长复杂系统设计、代码重构和技术决策，会先分析需求、制定方案再执行。

图：案例一：从0到1做一个电商购物页面

	Cursor	Claude code
速度	3 min 一次对话完成，较快	7 min 一次对话完成，适中
DOM 还原度	dom 完整	缺失了少量 dom
UI 还原度	间距、字号贴近视觉稿，自带响应式	间距还原表现稍显不足
整体	二者表现差不多，但 cursor 耗时更短，且自带完整的项目预览（对非前端同学很重要），故更适合快速做 demo 验证。	

资料来源：阿里云开发者、国信证券经济研究所整理

图：案例二：排查修复服务端的错误日志

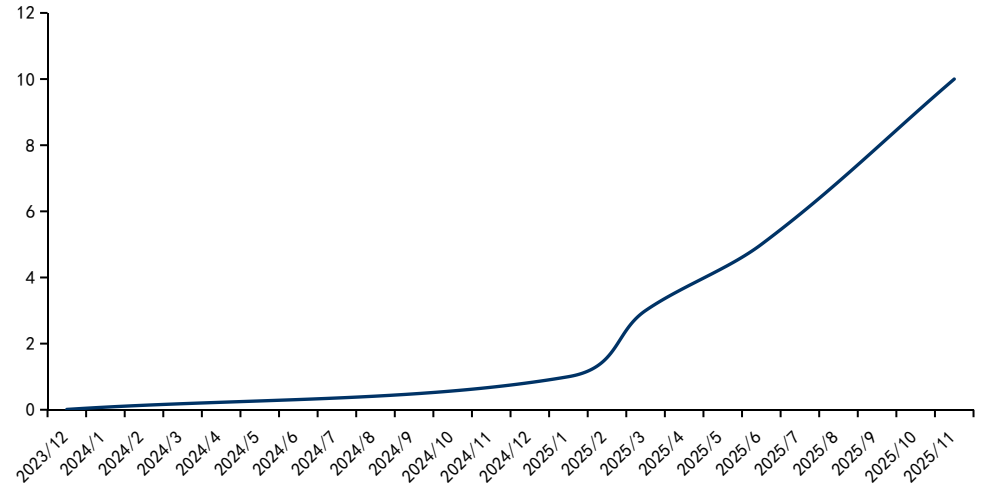
	Cursor	Claude code
速度	两次对话，耗时 6 min	三轮对话，耗时 17 min
质量	两轮对话方向皆出现了错误，没能修复问题	方向正确，最终修复了问题
整体	Claude code 虽然耗时久，但表现更好，前期充分调研，执行时也很谨慎，实时和开发者反馈确认。在这类问题上，Claude code 表现更好。 不过如果花一样的时间， Cursor 应该也能找到问题，只是需要开发者适当的引导。	

资料来源：阿里云开发者、国信证券经济研究所整理

Cursor：AI编程龙头，ARR达10亿美元，估值近300亿美元

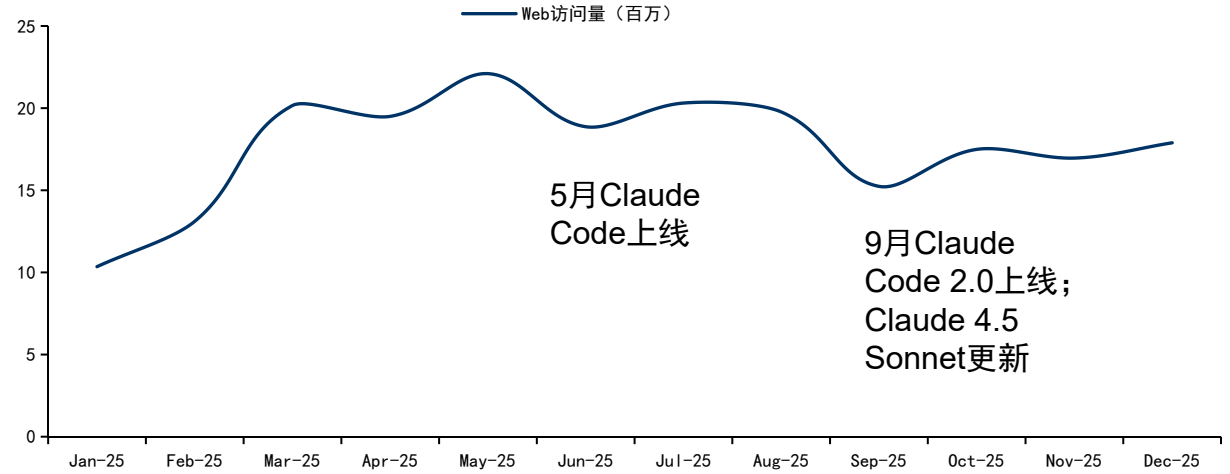
- **Cursor 是一款基于 AI 的代码编辑器，**由 Anysphere 团队在2023年3月正式推出，基于VS Code 打造、支持自然语言编程与聊天调试，推出几个月后上线了内联命令与上下文对话功能，让开发者能像和同事聊天一样“写代码”。**Cursor目前支持Composer 1（自研模型）、GPT-5、Claude、Gemini、Grok几个目前最强模型，**其自研模型Composer1于25年10月推出，是专为编程场景优化的“代理式”模型，生成速度比同级别的 Claude 3.5 Sonnet 或 GPT-4o 快 4 倍。**与竞品相比，Cursor在上下文理解、可解释性与隐私部署 上的优势明显，产品体验更贴近程序员的思维节奏。**
- **商业化方面，**根据公司官方数据及CBinsights，Cursor23年末ARR达到100万美元，24年中开始进入加速期，发布 Composer 功能后，用户量呈指数级增长，25年1月ARR突破1亿美元，3月达到3亿美元，付费用户超过 36 万；4 月日活跃用户突破 100 万，企业客户数量攀升至 1.4 万，**年底ARR已达到10亿美元，最新估值已达到293亿美元，是目前一级市场中估值最高的AI应用公司。**

图：Cursor ARR（亿美元）



资料来源：Z finance、国信证券经济研究所整理

图：Cursor网站访问量



资料来源：Cursor、国信证券经济研究所整理

AI 编程产品对比

国信证券 GUOSEN SECURITIES

- **Claude Code与Copilot/Cursor等呈差异化：**Claude Code在“项目级闭环+MCP企业工具生态+安全治理”上更像“工程代理”，后者在“IDE贴身效率与企业规模化协同”上各有长短。
- **价格与可用性：**Claude Code采用Claude全家桶额度，伴随Opus 4.5降价（输入/输出5/25美元每百万tokens），在“强模型更省token”的工程化实践下总成本可下降；Copilot有清晰的席位定价，组织采购便利且生态成熟。

表：AI编程产品的对比

工具	定位/优势	生态/集成	定价/模式
Claude Code	Terminal-First；深度上下文；MCP/文件API/安全策略；Plan Mode/并行会话	VS Code/JetBrains、GitHub Actions；企业API/AWS/Vertex	按模型token计费（Opus 4.5降至5-25$/M tokens）
GitHub Copilot	IDE贴身补全/编辑/PR总结/Autofix；Agent Mode	深度IDE/GitHub一体化；1.8M+付费	商业/企业：19-39$/用户/月
Cursor	架构/跨文件/全流程Agent；多模型路由	本地工具链，开发者友好	订阅制（多档），灵活模型成本
Gemini CLI	轻量命令行、百万context、免费试用	Google生态	免费+按量计费
CodeWhisperer	合规/溯源/企业私库微调	AWS企业集成	随AWS方案

资料来源：各公司官网、Openrouter、X，国信证券经济研究所整理

Claude Code与Opus 4.5协同实现行业顶尖的代码能力

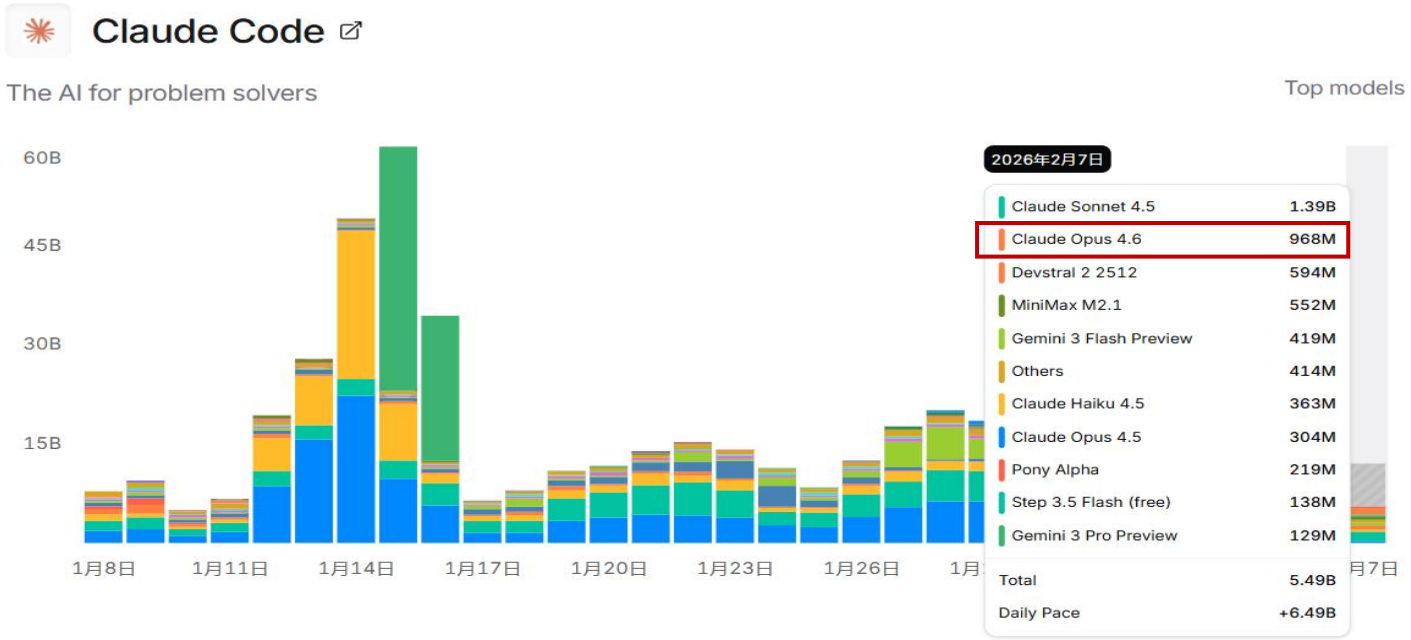
- Claude Code作为Anthropic打造的专属代码能力模块，并非简单“叠加”在Opus 4.5上，而是深度融合进模型底层，形成“通用智能+代码专精”的协同优势：

① **超大上下文：**Opus 4.5原生支持 200K上下文窗口，Claude Code可直接处理10万行级别的代码库、完整项目文档+代码的跨文件推理；

② **自然语言理解：**Opus 4.5对模糊需求的拆解能力（如“优化支付模块性能”），让Claude Code能精准理解“非标准化编程需求”，无需用户写精确指令；

③ **多模态基础：**Opus 4.5多模态能力（支持代码截图、架构图解析），让Claude Code可直接基于图片中的代码/架构完成修改、调试。

图：Claude Code调用各模型量



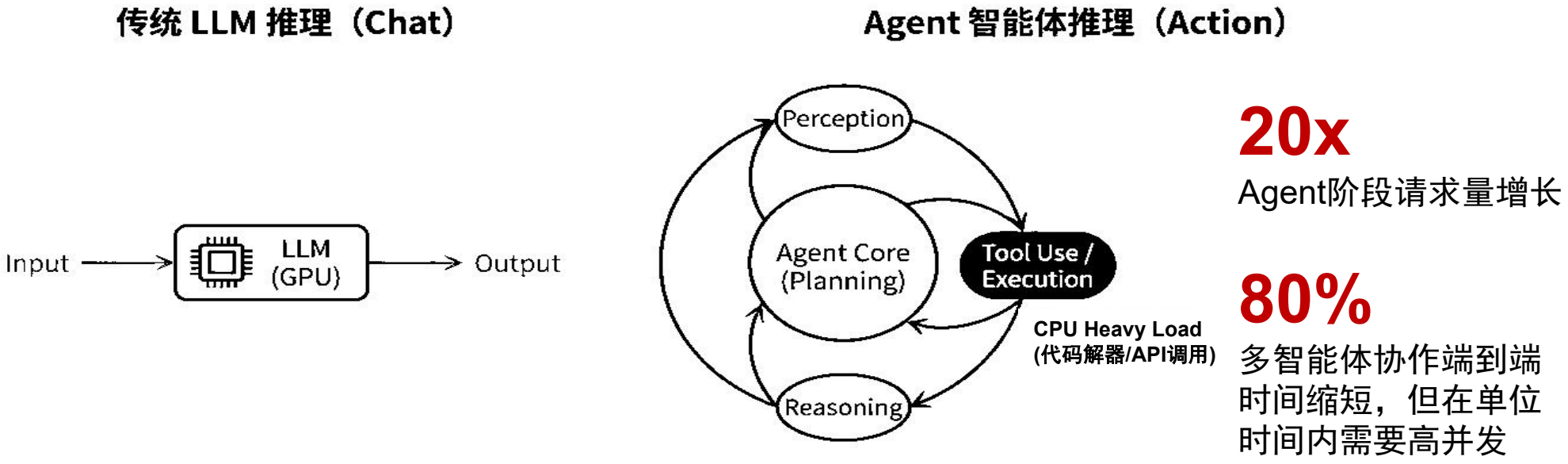
资料来源：Openrouter，国信证券经济研究所整理

目录

从问答/Chat到行动/Agent，带动CPU负载激增

- Agent完成**感知-规划-工具调用-再推理**的闭环后，推动CPU从**辅助单元升级为调度中枢与执行载体**，成为核心硬件瓶颈，其负载由**工具执行、编排调度、沙箱隔离、高并发与长任务**四类刚性开销驱动，成为影响Agent系统延迟、吞吐与能耗的核心瓶颈。
- **CPU的需求**：面向Agent，CPU正沿着“通过先进制程实现在更高核心密度下的功耗可控”这一主线发展，所需CPU总体呈**“超多核+低功耗/高制程+大内存带宽/容量+强IO/虚拟化+更大共享缓存”**特征。根据AMD CES 2026大会，2026年服务器CPU预计主力为64核、2027年两纳米商用后核数将达到128核起跳。

图：传统AI推理与Agent下范式的变化



系统瓶颈从“算力限制”向"逻辑运算/IO/存储限制"转移，CPU必须主动介入工具执行与流程编排，而非仅等待GPU。

资料来源：CB insights、X，国信证券经济研究所整理

Agent任务中四大开销确立CPU瓶颈地位：工具、沙箱、并发与缓存

国信证券 GUOSEN SECURITIES

- **工具执行与编排：**控制与工具调用是GPU难以替代的，吞吐量受CPU核心缓存一致性、同步机制制约，CPU成为影响延迟/吞吐/能耗的关键。
- **沙箱/微VM隔离：**原来CPU处理任务是毫秒级，可以池化共享。对于Agent来讲，任务很长、且需要存储和记忆。一个Agent必须配一个沙箱作为“容器”来保证进程和内存独立、安全，一个沙箱最少需要1个核(高并发、多工具调用需更多)，该核在用户打开Agent时持续被占用。每任务（或每用户）沙箱并发使CPU核数需求随业务峰值增长。沙箱建立以及安全防护也是额外进程/内核开销，抬高CPU基线占用。
- **高并发与长任务：**Agent将复杂任务拆解为多步工具调用，叠加多智能体协作，显著增加请求/使用tokens。
- **KV Cache与大内存协同：**长上下文导致KV Cache急剧膨胀，业界采用KV Cache Offload至CPU内存/CXL扩展内存等方式缓解HBM瓶颈，实质将CPU“大容量、高带宽内存”作为Agent“短期记忆”的容器。

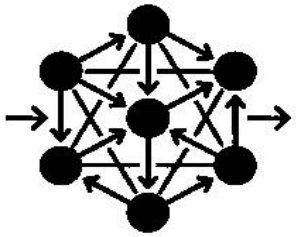
图：面向海量AI Agent的推理基础设施的四大挑战

1.工具执行(Tool Execution)
•复杂逻辑处理与代码执行。

耗比：耗时占比最高达90.6%；动态能耗占比可达44%

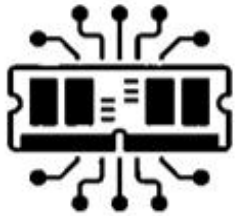
2.沙箱隔离(Sandbox Isolation)
•安全刚需，每个Agent任务需独立微VM环境。

Impact: 创建与销毁开销巨大，资源占用随并发线性增长



3.超高并发(High Concurrency)
•多智能体协作带来的爆发式请求。

活跃数：2030年活跃Agent数预计达22亿



4.KV Cache(Memory Overflow)
•HBM显存容量不足，CPU内存充当“短期记忆”容器。

Impact: KV Cache Offload导致极高的内存带宽压力

资料来源：IDC、AWS官网，国信证券经济研究所整理

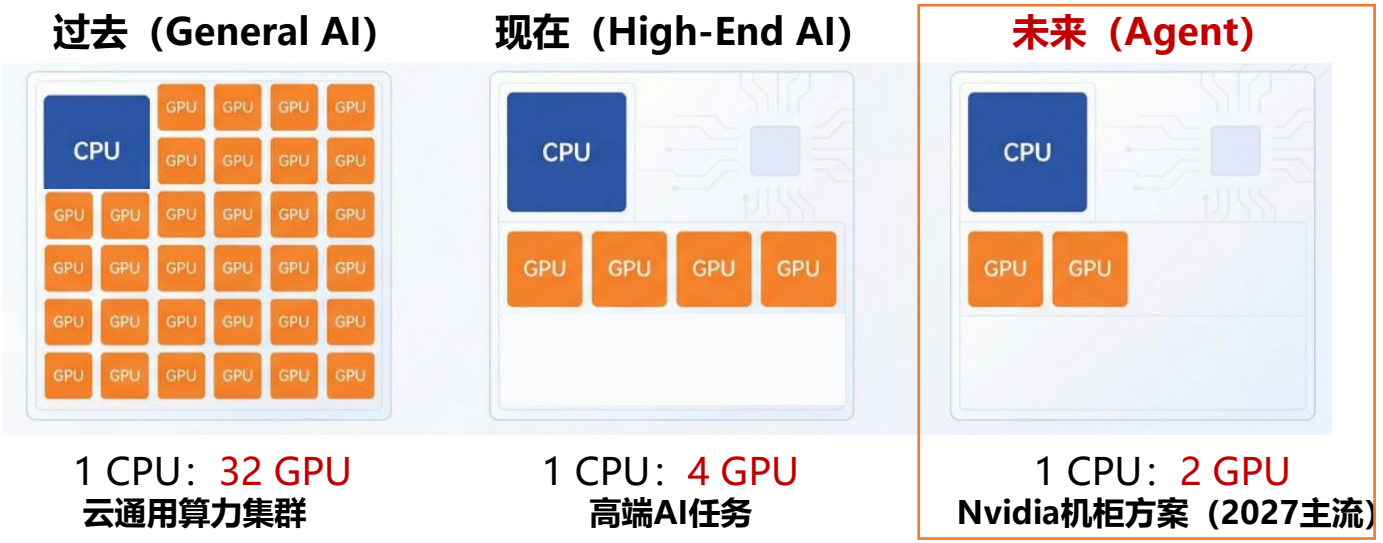
AI Agent背景下，CPU与GPU算力配比显著提升

- AI推理与Agent化提升CPU在集群中的并发调度/工具执行负载，促使服务器CPU配置自传统CPU:GPU配比从1:32（如阿里云智算集群、AWS Trainium集群等）向1:4(如英伟达DGX、阿里云ECS实例)演进，新代产品甚至普遍达到1:2(NVL72)。

① **柜内侧：**与GPU协同需更高CPU:GPU配比与大内存承载KV Cache，如英伟达CES展会（2026年1月初）Rubin采用1个CPU对应2张显卡的架构，预计2027年大量转换，当前仅1%-2%头部服务器需此架构。

② **柜外侧：**Agent需要大量独立通用服务器来创建沙箱与调度执行，用量与用户数/任务并发正相关，需超多核、高带宽、虚拟化与安全隔离强化的通用服务器CPU。

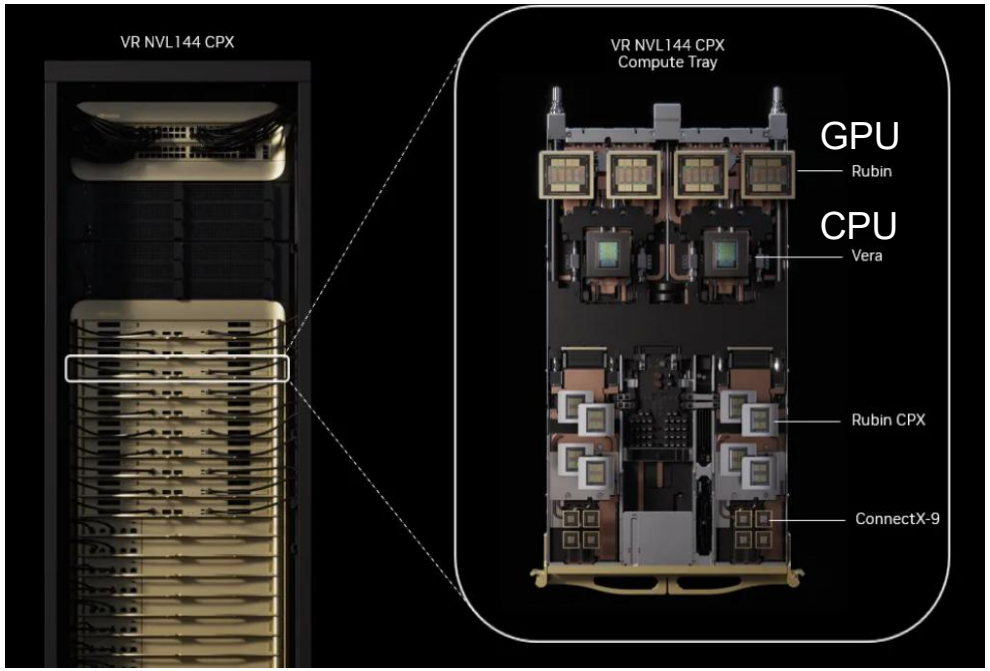
图：算力配比的演进过程



演进逻辑：Agent执行跨平台比价、自我反思及多步逻辑推理。此类控制与逻辑任务必须由CPU承担，预计CPU在AI服务器中价值占比从约10%持续提升。

资料来源：CB insights、IDC，国信证券经济研究所整理

图：2025CES大会Rubin的架构形态

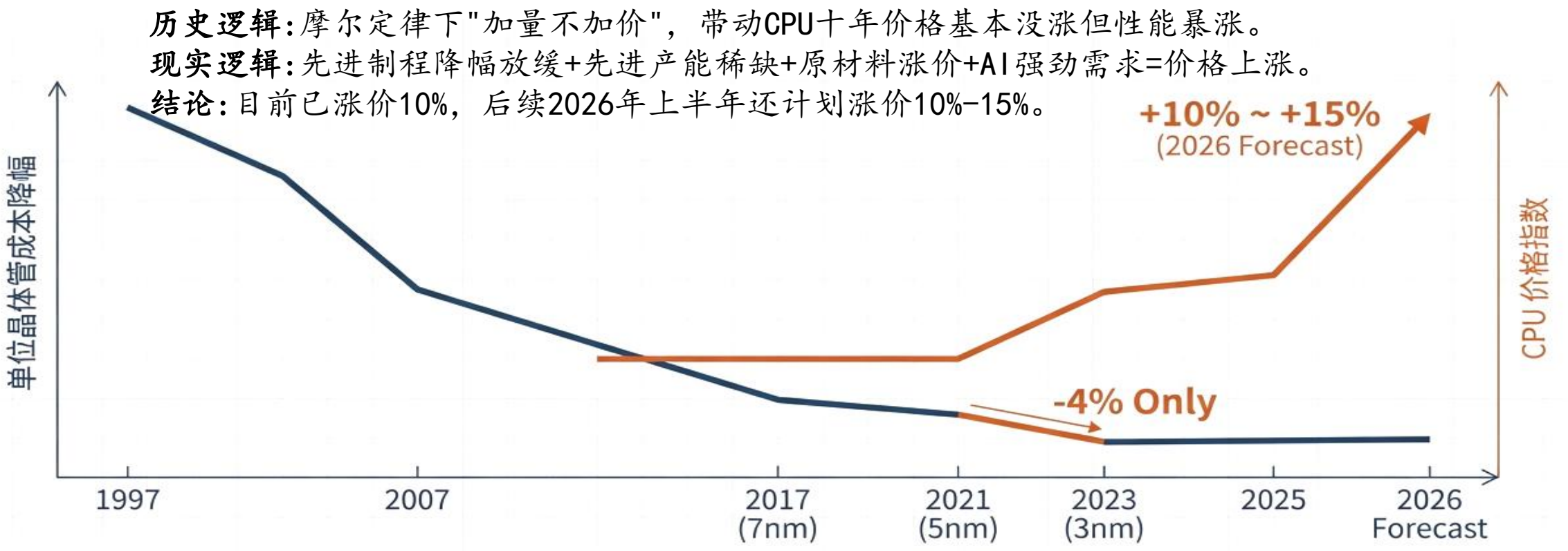


资料来源：2025年CES大会，国信证券经济研究所整理

其他涨价因素：后摩尔时代的“通货膨胀”+成本提升

- **过去20年摩尔定律下，CPU名义价格基本没涨但性能暴涨。**如2010年i5约$200，2025年i5/R5主流仍在$150-$250。如2005年主流CPU为2核2线程、3GHz、约$200，2025年达到6-8核12-16线程、5GHz+、约$200，同价位性能提升10-20倍，单位算力成本暴跌。虽然先进制程迭代放缓，但制程升级将显著降能耗。预计3纳米比前代省电1/3、2纳米再省40%。
- 2026年CPU需求增长叠加**贵金属原材料涨价、先进制程产能稀缺**，CPU开启涨价周期，截至目前26年2月，CPU价格已涨10%且我们预计涨价趋势可持续。

图：摩尔定律下单位晶体管成本下降

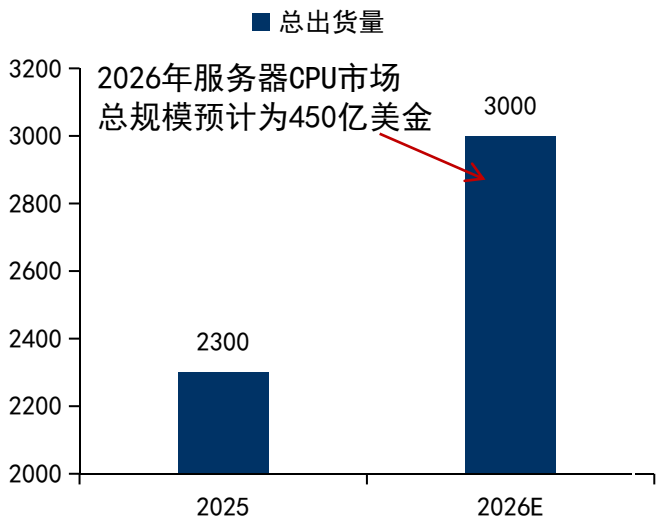


资料来源：IDC，Mercury Research，X，国信证券经济研究所整理

全球CPU市场规模与主要玩家

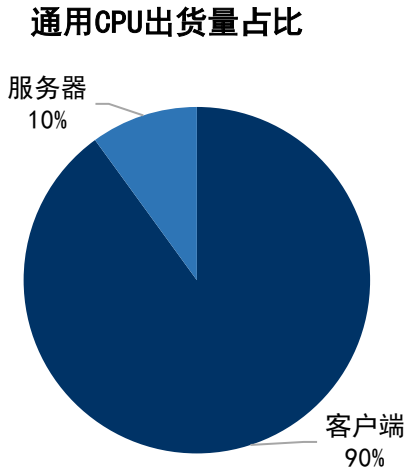
- **市场规模：**2026年全球服务器CPU市场总出货量预计达3000万颗，较2025年增长超30%，市场规模有望提升至450亿美金。其中通算CPU市场客户端CPU出货量占比约90%、服务器出货量占比10%。
- **市场份额：根据IDC数据预测26年英特尔在服务器CPU市场份额预计为55%左右，AMD预计为40%，**两家企业共同占据超九成的市场份额，垄断与头部效益明显。
- **CPU核数结构：**根据IDC与SemiAnalysis服务器CPU以8-64核为主，20核以上居多，高核数占比逐年提升；桌面/工作站CPU 10核以下为主，4-8核是绝对主流。手机SoC以8核为主等。

图：全球服务器CPU出货量与站（万台）



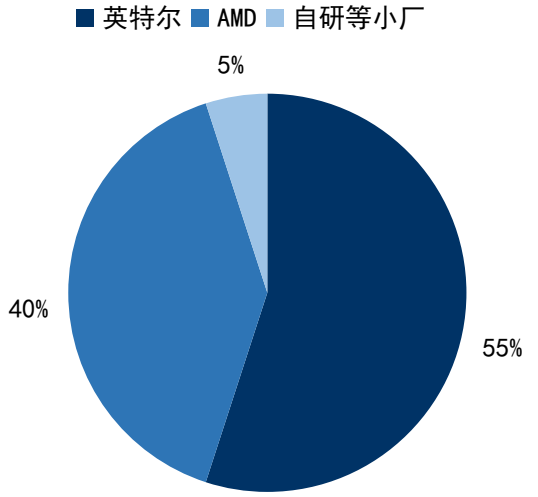
资料来源：TrendForce，国信证券经济研究所整理

图：通用CPU出货量占比



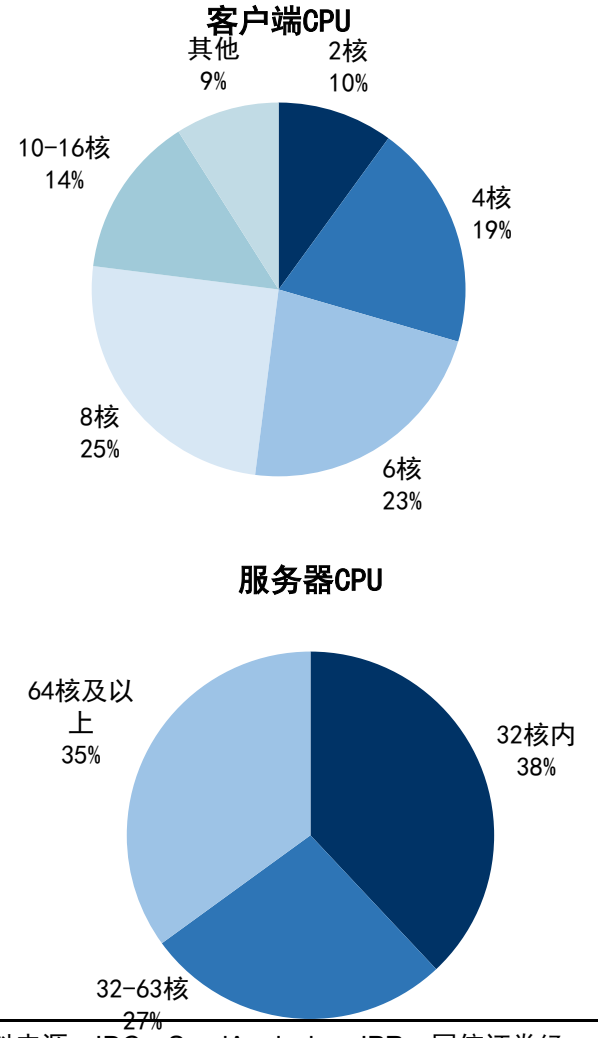
资料来源：IDC 、JPR，国信证券经济研究所整理

图：2026年服务器CPU市场总规模份额



资料来源：IDC，Mercury Research，国信证券经济研究所整理

图：客户端与服务器CPU不同核数占比

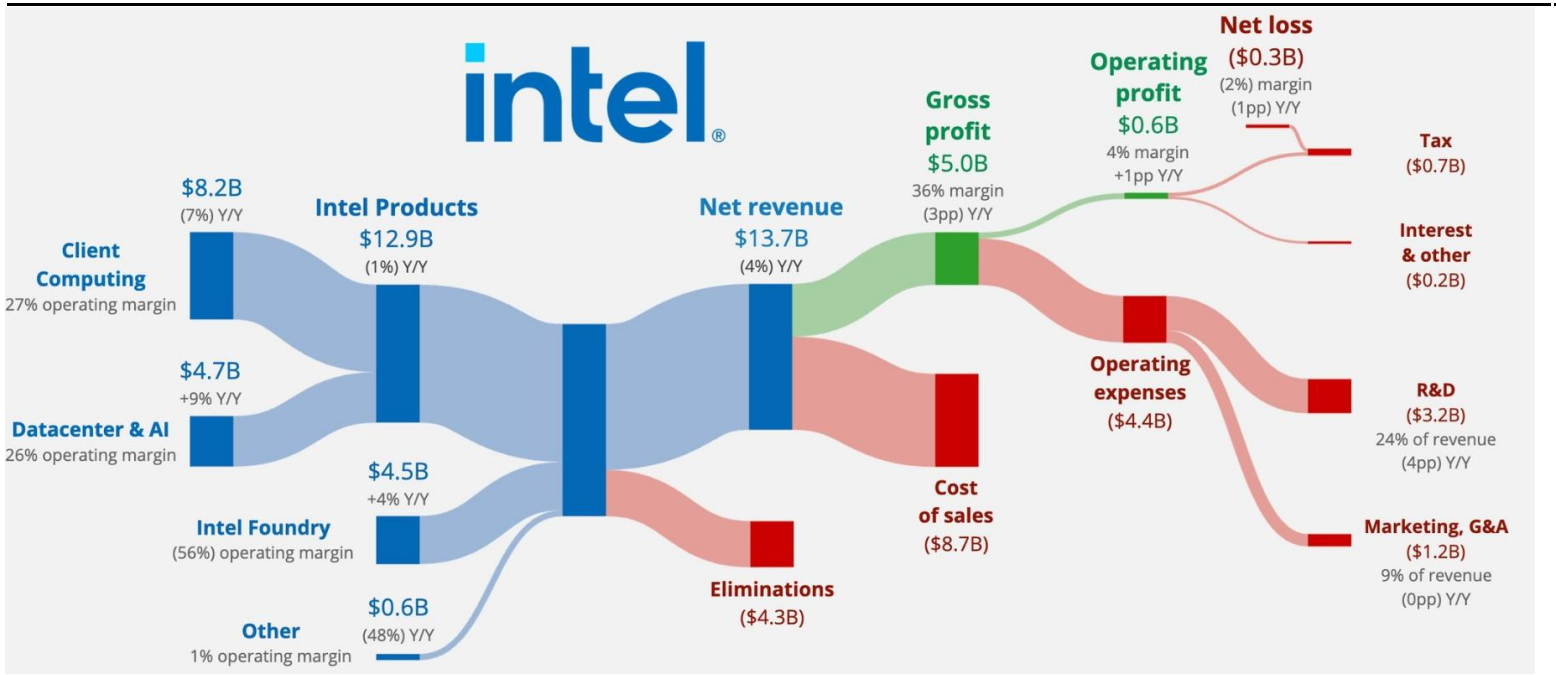


资料来源：IDC、SemiAnalysis、JPR，国信证券经济研究所测算整理测算

英特尔营收与产品结构

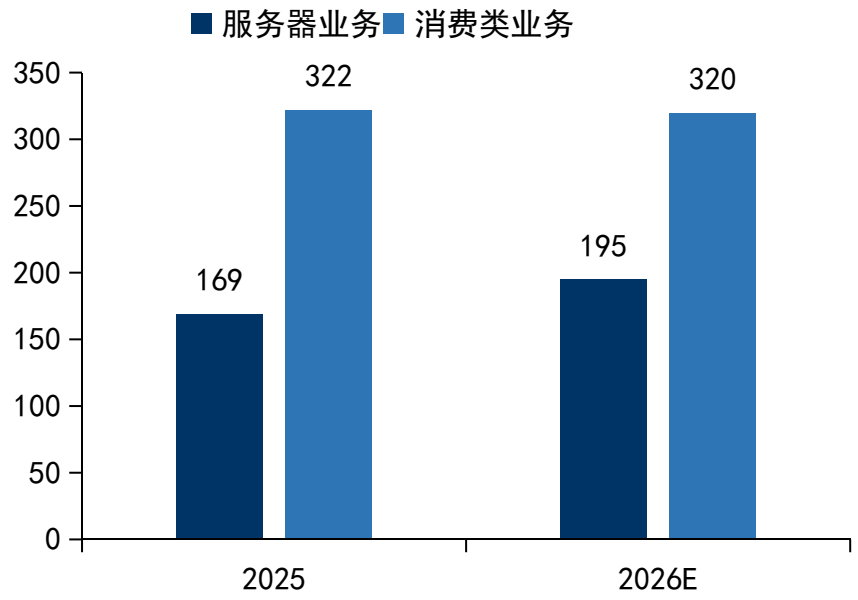
- **英特尔营收拆分：**2025年整体营收529亿美元，CCG（客户端产品）仍为最大收入来源、占比60%，25年受PC市场波动、营收小幅下滑。DCAI（数据中心产品）受益于云厂商算力升级与AI训练/推理需求为增长核心。IFS（代工服务）内部供应优先，外部客户包括苹果、博通、英伟达等，短期亏损但长期潜力显著。
- **英特尔产品结构：**2026年彭博预计收入低个位数增长，服务器业务考虑涨价与需求提升有望实现15%左右增长。而消费类CPU业务2026年收入预期持平，消费级CPU因性能过剩且Agent功能在消费类终端渗透有限，因此未明显拉动需求。因Q1新品良率问题、苹果M5系列3月底上市冲击轻薄本市场，同时叠加台式机市场销量减少、换机周期延长等因素。

图：英特尔25FYQ4营收与开支图



资料来源：appeconomyinsights，国信证券经济研究所整理

图：2025年及2026年英特尔CPU业务收入（亿美元）



资料来源：英特尔财报，彭博一致预期，国信证券经济研究所整理

CPU架构之争：生态与技术壁垒

- X86凭借稳定性和成熟的软件生态占据主流，尤其在服务器市场兼容性突出。ARM虽在能效和特定生态中有优势，但仍面临兼容性、供应量等挑战。
- 展望未来，**X86通过技术联盟与国产化迭代巩固现有优势，ARM阵营具备统一软硬件栈、全场景部署与闭环生态优势，备受开发者青睐，随英伟达、苹果、高通等发展而崛起，我们预计未来市场可能逐步向ARM+苹果生态倾斜**。

图：CPU架构的优劣势

x86(Intel/AMD)

- **市场地位:**垄断者(>90%市场份额)
- **核心优势:**生态稳定，软件生态兼容性与适配度远优于ARM，企业客户首选，避免系统崩溃风险。
- **劣势:**能耗较高。
- **趋势**:2024年成立x86技术联盟，联手防御。

ARM (Nvidia/Apple/Qualcomm)

- **市场地位**:挑战者
- **核心优势:**能效比(Efficiency) +垂直整合(Nvidia CPU+GPU优化)+生态内优势明显（苹果生态）。
- **劣势:**指令集天花板、处理复杂指令能力较弱，软件重写成本高。

资料来源：英特尔官网、AMD官网、IDC，国信证券经济研究所整理

CPU架构之争：玩家特点与技术路线

国信证券 GUOSEN SECURITIES

- 在AI算力需求重塑服务器CPU市场的背景下，ARM、AMD与英特尔正沿三条差异化产品路径展开竞逐。根据海光、龙芯、飞腾产品手册标注制程与性能参数，主力制程12-16纳米，性能约为英特尔六年前i3水平，产品竞争力仍有一定差距。

① **ARM路线：**注重与GPU深度协同及能效优势，以英伟达柜内CPU为代表，在新架构中通过提升核数与超线程强化算力集成。

② **AMD路线：**聚焦多核与先进制程，计划推出2纳米工艺的“Venice”服务器CPU，通过双I/O die封装实现多达256核，核心数量领先，主打高性能与复杂运算适配，但面临稳定性挑战。

③ **英特尔路线：**坚持X86架构主导，通过与AMD成立技术联盟巩固生态；产品迭代以制程升级为核心。

表：国内外厂商的各方位商业格局对比

厂商/架构	竞争位势	生态亮点	客户关系
英特尔（X86）	份额占比高，晶圆产能向服务器倾斜，短期供给受限	X86 通用生态与 OS、数据库、云平台长期绑定，推进封装质量与良率提升	CSP 锁全年产能，接受一季度提价；Foundry 依托 18A 定价优势显著
AMD（X86）	Venice 路线图展示 2nm 工艺，最多 256 核、双 I/O 封装革新	与 AI 加速器生态耦合，X86 桌面端份额提升	2026 服务器 CPU 涨价执行，与 CSP 合约折扣回撤并行；出货周期延长并与资源锁定
海光（X86）	X86 兼容优势明显，在政企、金融、电信等领域渗透，CPU+DCU 一体化方案被推出	采用完整 X86 指令集与国产 X86 架构，适配主流 OS、云及数据库，信创采购份额较高	政务行业批量出货、订单充沛；国产替代下公开市场突破，客户迁移成本低
龙芯（LoongArch）	自主指令集为护城河，多硅片封装至 64 核 128 线程，整机性价比优势明显	完成 3400 + 家企业生态的 OS、数据库、云等适配	党政、能源等强合规场景优先采购，国产化目标驱动份额提升
飞腾（ARM）	政企及运营商批量出货，在桌面、服务器、嵌入式场景均有覆盖	ARM 生态与国产 OS、中间件适配，强化了安全架构	2026 国产化率指标提升，在通用与专用场景并行渗透
兆芯（X86）	源代码自研体系、双系统兼容，降低了迁移成本，新一代 KH-50000 提升性能	面向政企办公与教育桌面侧性价比突出，适配生态完善	在信创与公开市场双轮驱动下扩大出货，投标拆分机制保障进入

资料来源：各公司官网，SEC，国信证券经济研究所整理

风险提示

第一，宏观经济波动。若宏观经济波动，产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，相关产业技术壁垒较高，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、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意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032