



Agent Computer

重新定义计算范式与硅基底座

From Human Computer → Agent Computer

原粒半导体 CEO 方绍峡 | 2026.07

当AI从“回答”走向“执行”

过去三年，行业围绕大模型讨论：

- 模型有多大？
- 跑分有多高？
- 上下文有多长？
- Token 有多快？

VS

但是，

Agent 时代的一个新问题是：

当 AI 开始替人工作，

下一代计算机应该长什么样？

**每一次重大的计算主体切换，
都会诞生全新范式的时代硬件主角。**

AI 的身份变了，计算机也必须变

Chat →→ Copilot →→ Agent

过去

工具

AI is a Tool

搜索引擎、翻译器：人操作，工具辅助

现在

助手

AI is an Assistant

写邮件、写代码、做总结：围绕人的即时交互展开

未来

员工

AI is an Employee

有任务、有权限、有记忆、有 workflow，也有长期目标

如果 AI 只是工具，它可以住在人的电脑里；
如果 AI 是员工，它就需要自己的工作环境、自己的资源、自己的系统、自己的边界。

当 AI 成为员工，它就需要自己的电脑。

历史奖励新的计算平台



每一次计算平台切换，都会重写产业分工，
也会重塑硬件、软件和应用生态之间的关系。

为什么 Agent 不能继续住在传统 AI PC 里？

Agent 的工作负载和人类 PC 的工作负载存在天然冲突。

⚠ 资源冲突

生产力 Agent 会长期占用大容量显存与算力，
不能与主力办公机争抢资源。

⚠ 安全冲突

Agent 需要文件、浏览器、代码库、工具链权限，
必须与主力设备物理隔离。

⚠ 长驻冲突

人类电脑会休眠、移动、断网；
Agent 需要 7×24 小时在线。

⚠ 系统冲突

传统 OS 为人类即时交互优化，
Agent 需要异步任务、状态管理与长期调度。

“AI PC 是在电脑里加入 AI，
Agent Computer 是为 AI 重新发明电脑。”

Agent 的权限越深，数据边界越重要

当 Agent 拥有执行权，
数据安全就不再只是合规问题，
而是**计算架构**问题。

知识权限



- 代码库
- 文档
- 项目资料

行为权限



- 邮件
- 日程
- 协作工具

系统权限



- 浏览器
- 文件系统
- 数据库
- 支付/审批接口

*权限越深，价值越大，风险也越大。如果高价值私域数据完全依赖云端流转，企业和个人都会面临数据边界不可控的问题。

高价值私域数据，需要一个用户可控、物理可信的本地计算边界。

知识在云端，但“记忆”必须在本地

大模型学会了世界知识，Agent 将学习你的知识

🌐 过去：公域知识

- 互联网公开网页
- 学术论文·书籍
- 代码仓库 (GitHub)
- 维基百科·新闻
- 公开论坛·社交媒体

高质量公域数据红利正在递减



🔒 未来：私域数据驱动

个人偏好·习惯·决策记录

个人价值

企业SOP·行业Know-How

行业壁垒

工作流程·工具调用模式

效率资产

项目记忆·沟通风格

关系资产

私人知识库·研究积累

知识资产

私域记忆 = 下一代 Agent 的核心燃料

未来最稀缺的资产，不只是模型，而是私域记忆。
私域记忆必须有一个本地、可信、长期存在的物理承载。

Token 经济学的临界点：从 Chat 到 Work

复杂长任务驱动的云端调用成本快速上升

🗨️ 单次对话 (Chat)

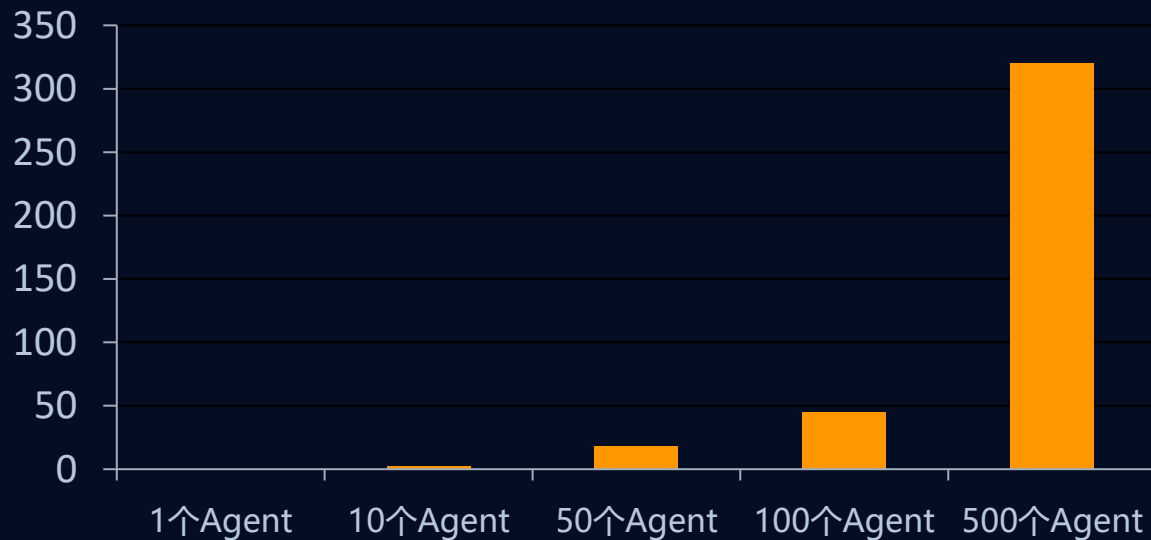
仅消耗约 300 ~ 1,000 级 Token，单次对话成本可控，人类阅读速度天然存在上限，交互频次也有限。

🤖 智能体托管 (Work)

跨模块代码重构或长周期分析，Agent 在后台进行数百次异步推理，单次复杂长任务可消耗 1M ~ 10M+ 的大体量 Token。

注：含规划、工具调用、检索、代码/文档上下文、多轮自我修正等内部 Token 消耗。

相对推理成本指数



注：相对指数为趋势示意，反映 Agent 数量、任务复杂度与上下文消耗共同增长。

**云端推理不会消失，
但Agent长期执行必须拥有本地成本锚点。**

Agent 正在重新定义 “足够大的模型”

Chat 模型：回答问题

- 短上下文
- 单轮输出
- 人类即时阅读
- 错误可人工纠正

Answer → Work

过去，小模型 “够用” 就好。
但进入 Agent 时代，“够用” 的定义正在变化。

Agent 模型：完成工作

- 长上下文
- 多步规划
- 工具调用
- 自我验证
- 长周期执行

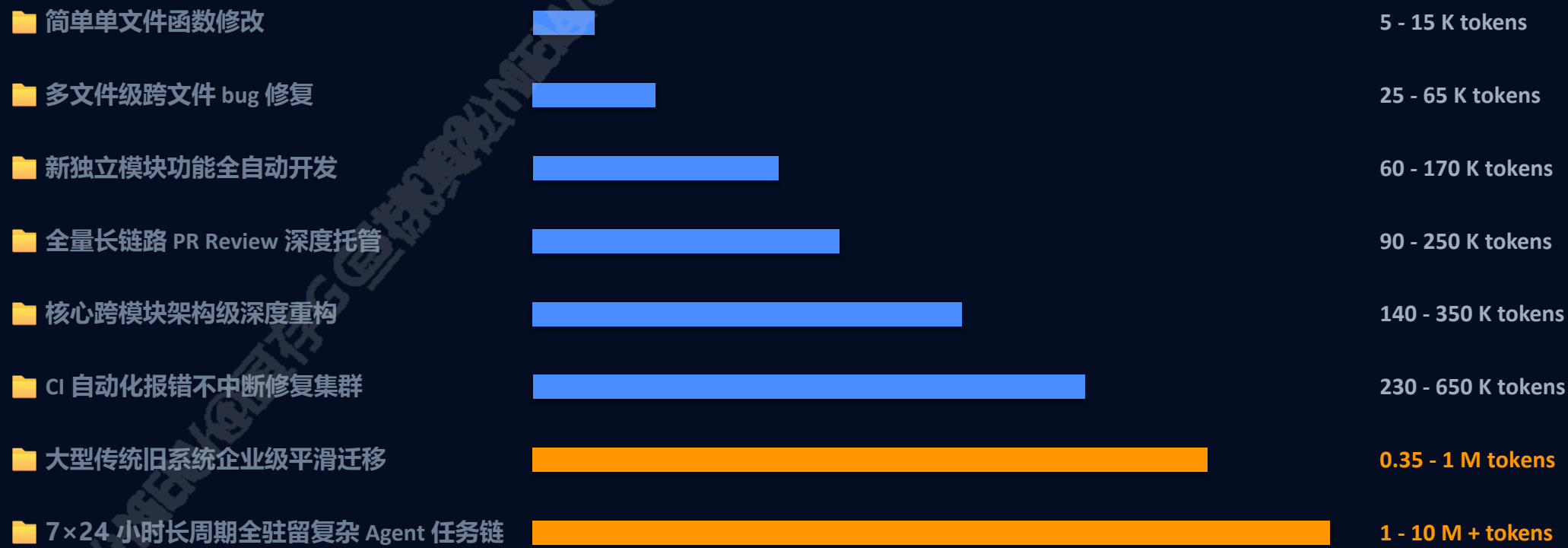
*小模型可以降低交互门槛，但真正决定生产力上限的，往往是更大的模型、更长的上下文、更强的多步推理能力。



小模型降低交互门槛，大模型决定生产力上限。

AI Coding 揭示长任务上下文爆发需求

AI Coding 是观察 Agent 工作负载最典型的场景：



* 典型复杂度估算 * Agent 的上下文不是一次性输入，而是任务状态、代码库、错误日志、工具返回结果和自我修正轨迹的持续累积。
* Agent 时代的生产力底座：更长任务、更大上下文、更稳定状态。

本地设备的核心指标，正在从“峰值算力”转向“长任务承载能力”。

大容量本地模型内存，正在成为端侧生产力分水岭

◆ 24GB 本地模型内存 入门配置

入门体验，适合小模型、短上下文、单轮任务。

◆ 128GB 本地模型内存 工作流常驻基本配置

工作流常驻：支持中等规模模型与多文件任务。

◆ 256GB 本地模型内存 生产力底座

面向长上下文、知识库、代码仓库和复杂 Agent 工作流。

◆ 512GB+ 本地模型内存 旗舰级多 Agent 协同

多 Agent 协同、长期驻留和高并发任务链。

*Agent 时代，本地设备的竞争不再只是峰值算力，更重要的是：本地可用内存有多大，带宽是否足够，长任务状态能否稳定驻留。

Agent 的生产力边界， 越来越取决于本地可用内存与带宽。

Agent 时代比拼的不是 TFLOPS，而是任务完成效率

Agent 算力载体正在从大 Batch GPU 集群，走向推理、内存与调度深度耦合的异构系统。

传统通用服务器架构



面向大 Batch、规则化、高吞吐
矩阵计算优化。

智能体原生异构架构



面向长上下文、动态状态流、多 Agent
并发和工具调用优化。

*最终用户关心的不是这颗芯片峰值有多高，而是一个复杂任务能不能更高效完成

**Agent 工作负载不是单一矩阵乘法，
而是推理、记忆、检索、工具调用与状态调度的复合系统。**

Agent 追求的是任务完成时间，而不是屏幕生成速度

💬 Chat 模式

15–25 token/s
人类流畅阅读交互区间

*过去我们谈 Token 速度，更多是在谈用户体验。

⚡ Agent 模式

100–1000+ token/s
压缩任务完成时间

*Agent时代我们讨论的不是屏幕上的字出现得快不快，而是一个任务链路能否从几个小时压缩到几十分钟，甚至几分钟。

- 在每秒 35 个 token 的极慢速度下，一台 GB200 NVL72 可支持数万名用户运行。
- 但若切换到另一端，每秒 270 个 token，它们只能支持一名用户运行，此时 GPU 成本急剧上升。
- 从大约每秒 100 到 150 个 token 开始，GPU 成本与效率迎来明显转折。

—— Cerebras IPO 路演，2026.5

注：该数据为 Cerebras IPO 路演观点引用，用于说明高速推理下 GPU 经济性变化趋势。Agent时代推理芯片的指标需要被重新定义。

对 Agent 而言，Token 速度不是体验指标，而是生产效率指标。

Agent Computer 的全栈生态正在形成

🏆 顶层：数字劳动力市场
Agent 应用与数字员工服务

🔒 四层：私域安全数据网络
本地加密、授权与敏感数据分发

📁 三层：私域记忆金库
长期记忆、向量检索与本地知识库

⚙️ 二层：Agent OS & Runtime
任务编排、状态管理与工具调用

🌟 底座：硅基原生底座 (Agent-Native Silicon)
面向智能体场景的新型计算体系

* 只有底座足够稳，整个数字劳动力生态才能真正规模化。

原粒不替代上层生态，
而是专注为 Agent Computer 提供高性能、低功耗、可扩展的物理底座。

原粒芯片设计逻辑的底层出发点：Agent-Native Silicon 设计原则



Memory First
记忆优先

大容量本地存储与高带宽访问



Context First
长上下文优先

面向长上下文与长任务链优化



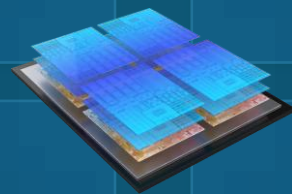
Always-On First
长期驻留优先

7×24 小时常驻、低待机功耗



Energy First
能效优先

以 Token/J 与长任务能效衡量任务效率



**Agent-Native Silicon 的核心，
不是更高峰值算力，而是更低成本地完成更长任务。**

原粒半导体：为 Agent Computer 铸造硅基底座

我们不定义上层应用。

我们不替代整机生态。

我们专注为 Agent Computer 提供高性能、低功耗、可扩展的硅基底座。

我们希望解决的问题：当世界从集中计算走向万物智能，底层芯片应该如何变化？

过去，智能主要集中在云端、集中在少数数据中心。

未来，智能会进入办公室、家庭、设备、生产现场，进入每一个需要数字员工的场景。

**我们的愿景：把一个“集中计算”的世界，
变成一个“万物有脑”的世界。**

PC 让每个人桌上有了一台电脑
智能手机让每个人口袋里有了一台超级计算机
Agent Computer 将让每个人身边拥有一个 7×24 小时在线的数字员工
——这不是一次简单的硬件升级，而是一次计算主体的变化

From Human Computer → Agent Computer Welcome to the Age of Digital Labor.

当 AI 从回答问题走向完成工作，当软件从工具走向员工，当数字劳动力开始成为新生产力...

原粒半导体：为智能体时代重构物理硅基底座。