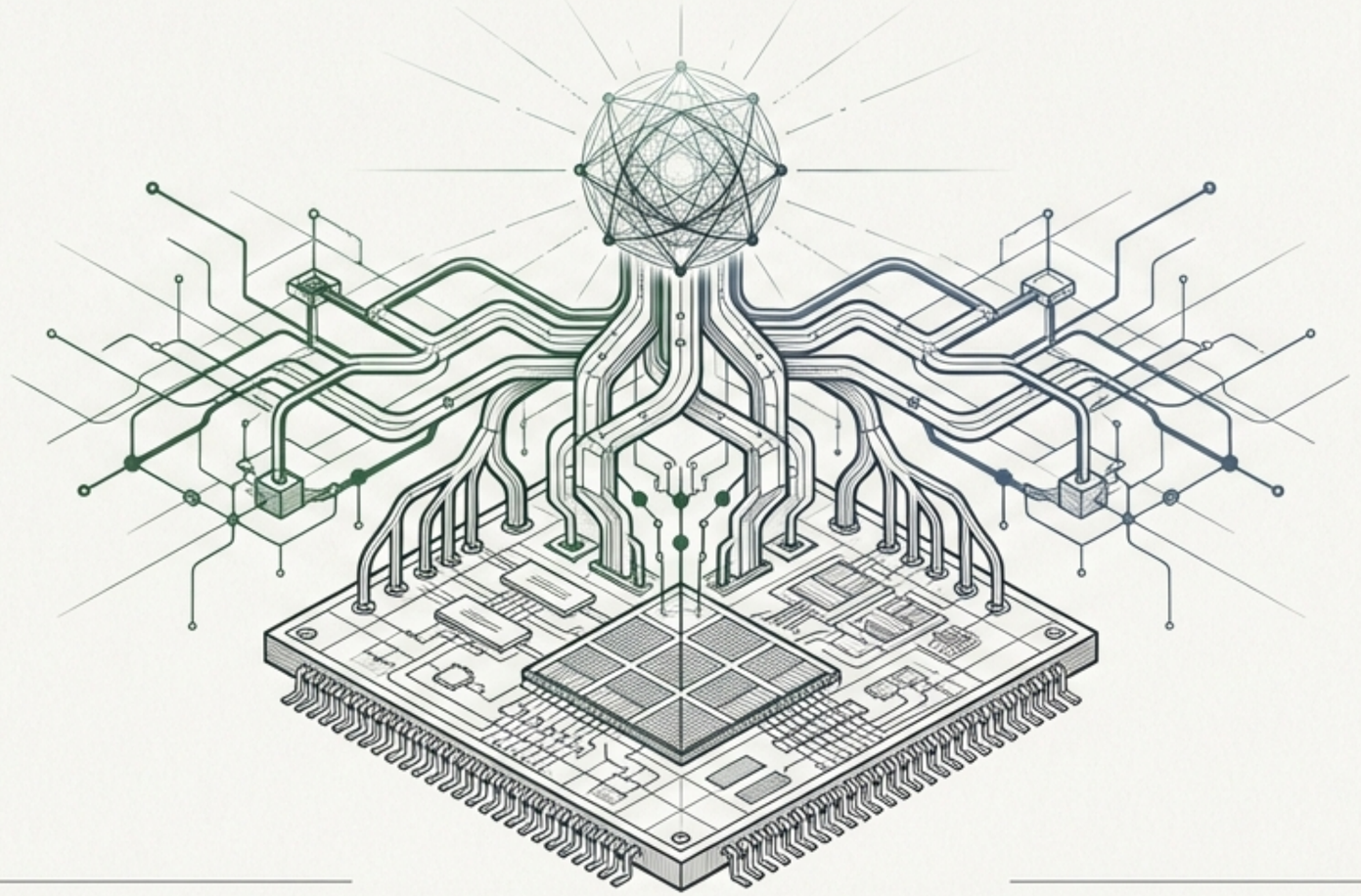
所有 AI 问题，拆到底都是算力问题

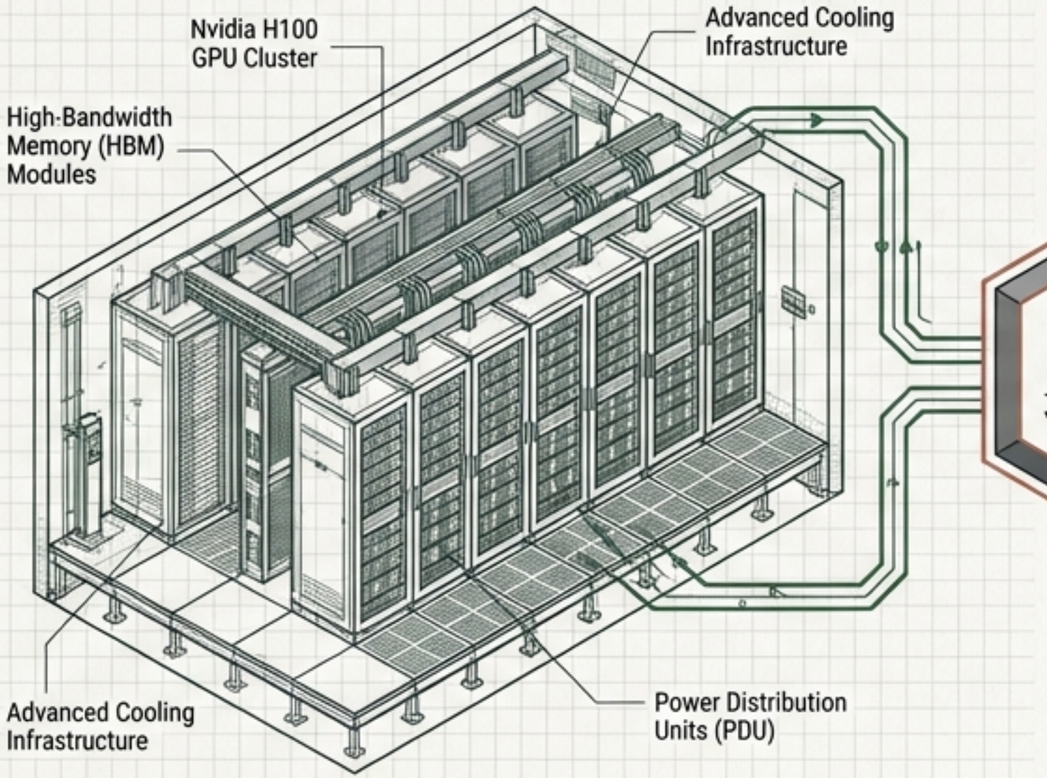
从第一性原理看 AI 发展轨迹与底层经济学



宏观基建与微观产品，撞上同一堵墙

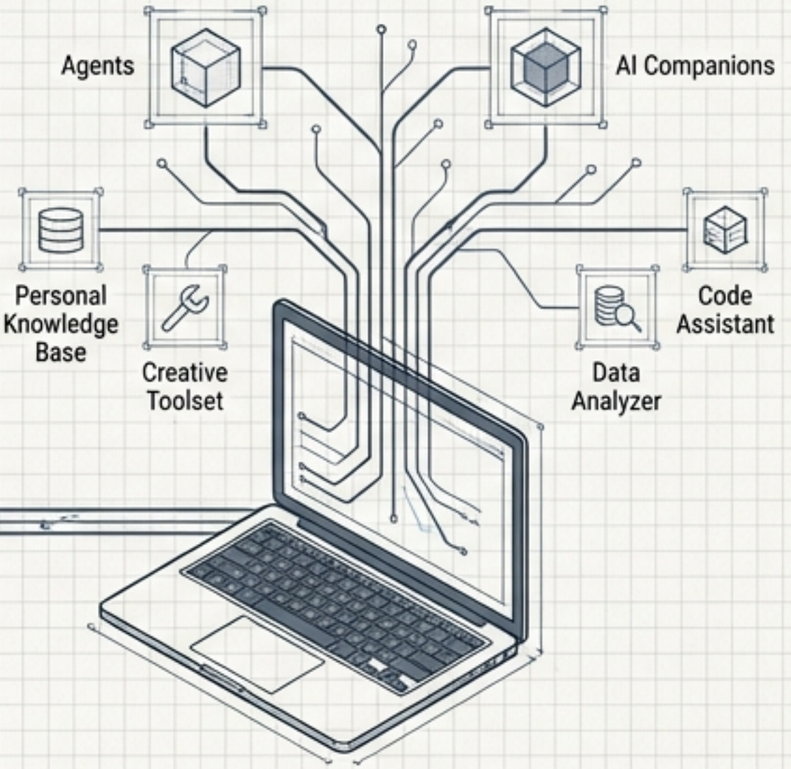
无论是搭建千万级别的 GPU 集群，还是打磨一个纯粹出于好奇心的 AI 产品，当产品走到极限时，挡路的永远是同一个物理约束。

Macro side



同一个
最终瓶颈

Micro side



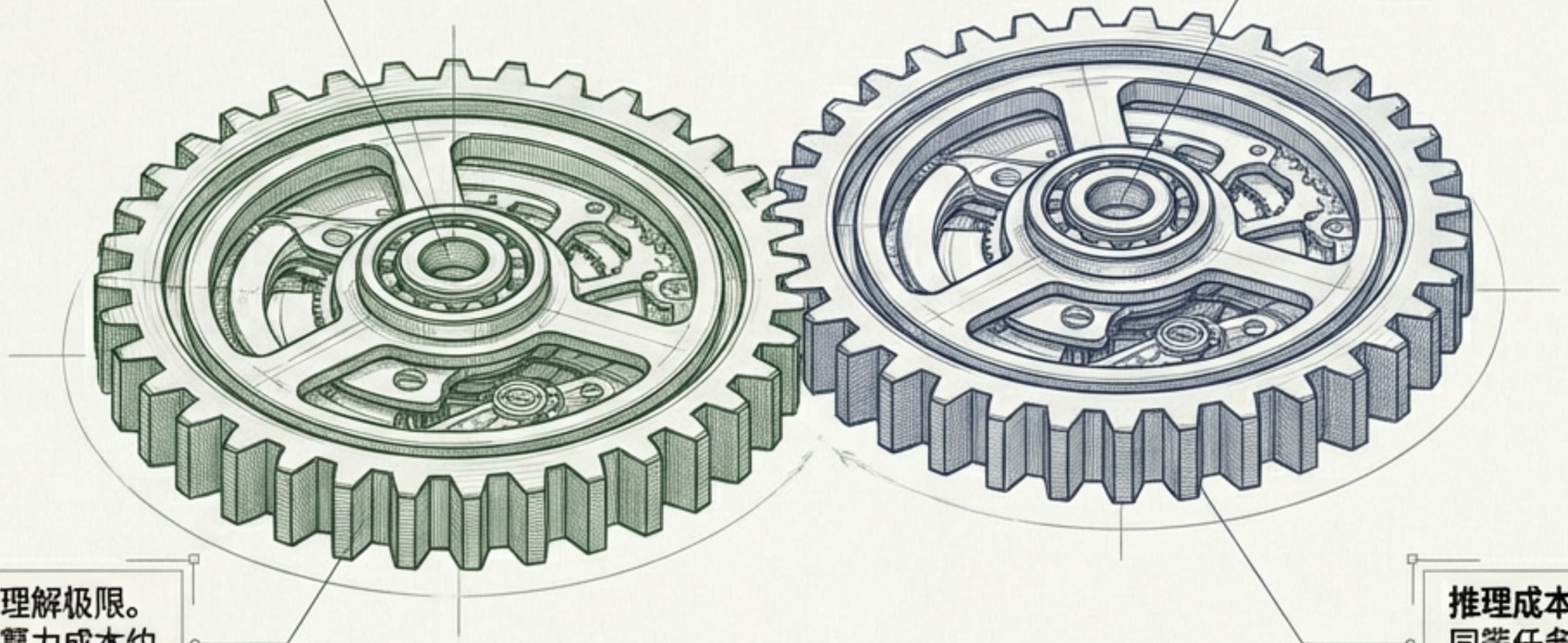
算力金融化与基础设施建设
(GPU 集群采购 / 数据中心)

独立 AI 产品研发
(两个月六个 AI 产品)

智能不是魔法，是算力堆叠出的计算工程

训练算力 (Training Compute)
解决“不够聪明”的问题。

推理算力 (Inference Compute)
解决“太贵”的问题。

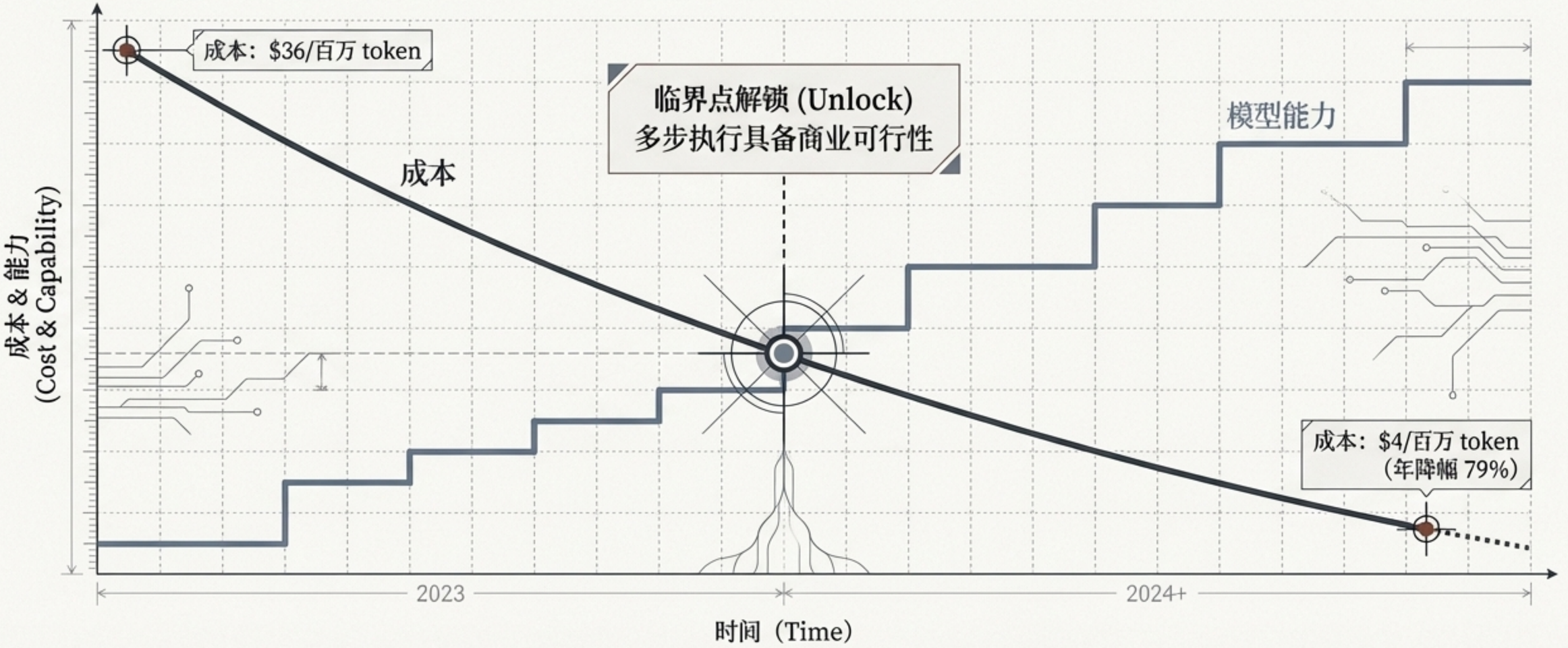


训练成本：突破理解极限。
前沿模型训练纯算力成本约 7800 万美元。

推理成本：扩大使用边界。
同等任务成本两年内下降一到两个数量级。

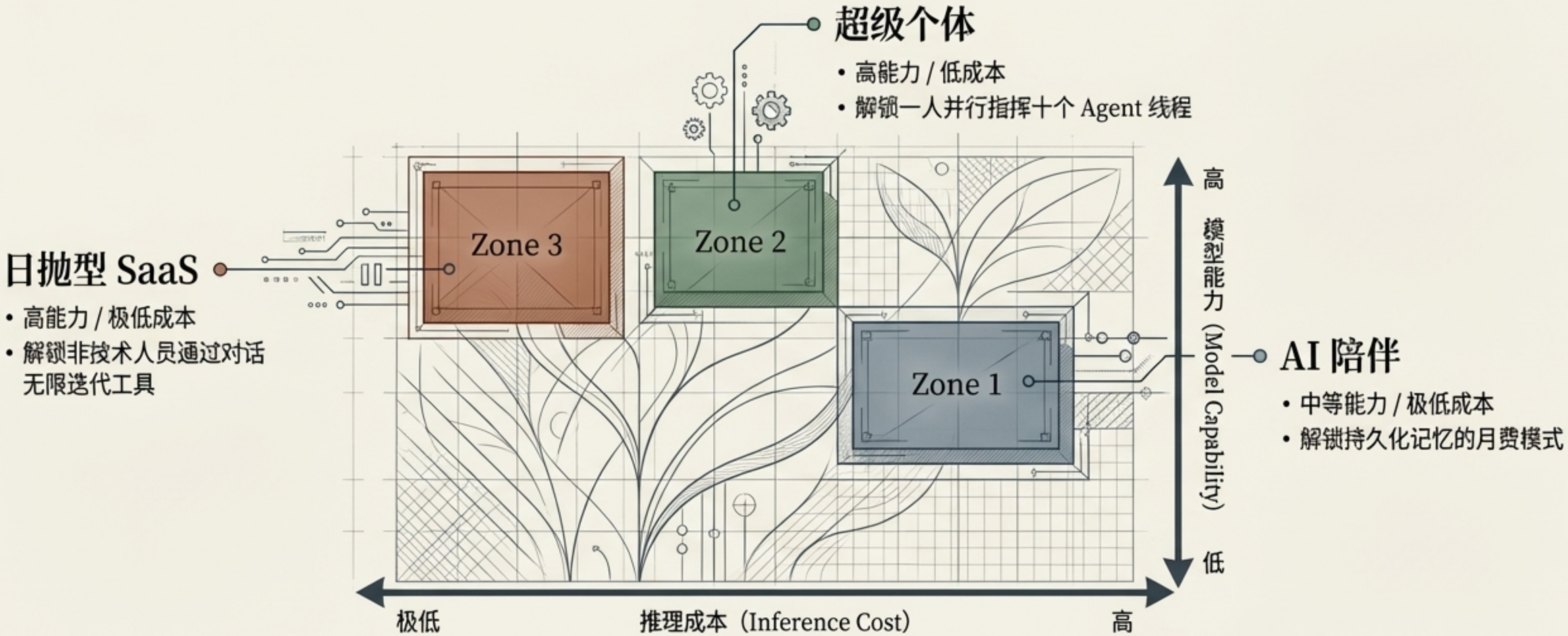
Agent 的爆发六并非偶然，是跨越了成本阈值

2023 年的模型无法胜任多步规划，且跑一次任务成本极高。只有当“模型能力”与“推理降本”同时跨过临界点，Agent 经济才具备商业可行性。



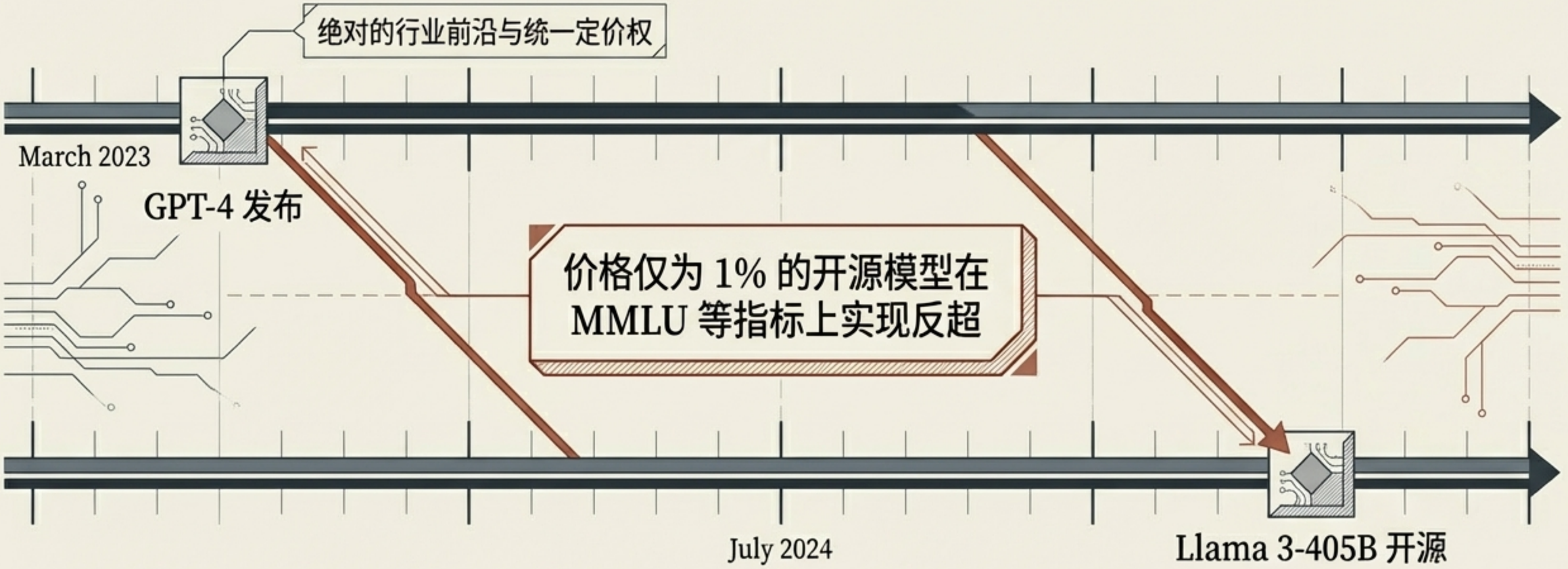
成本决定可能性的边界

每一个 AI 应用赛道的爆发，本质上都是算力在特定维度上跨越了经济阈值。不看发布会，看算力曲线。



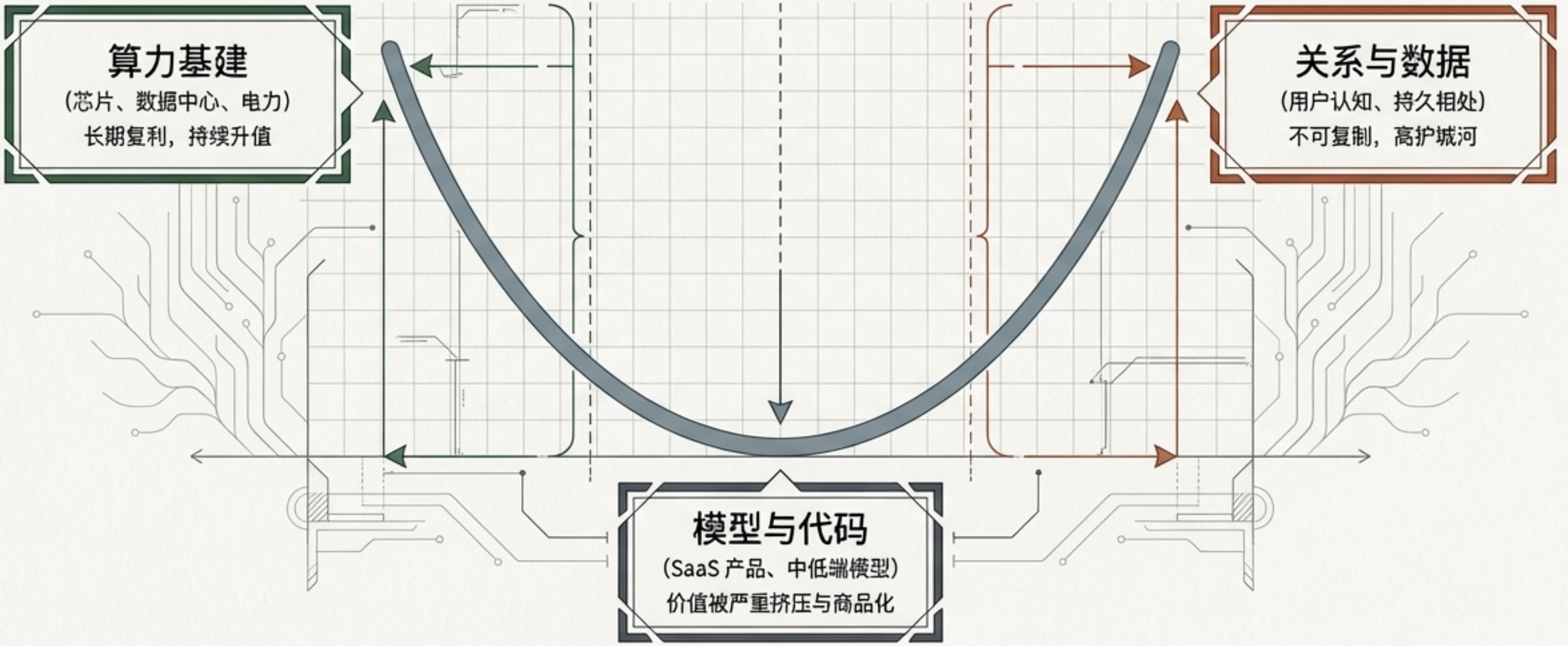
模型具有极短的半衰期

前沿模型的同等性能成本每年下降 5-10 倍。完全商品化的性能层级下降 40 到 900 倍不等。今天的前沿，18 个月后就是商品。



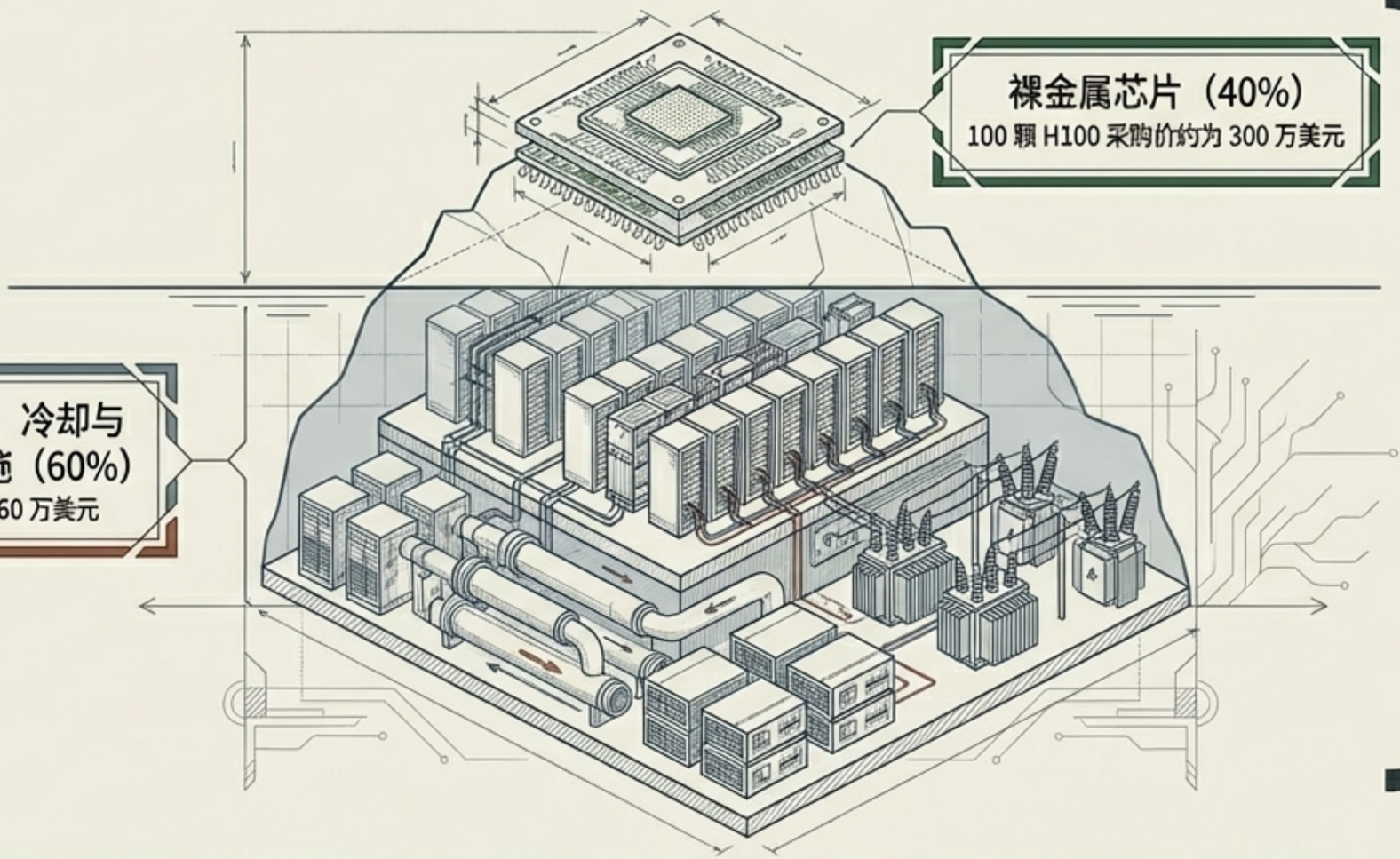
价值的重新分配：两端升值，中间贬值

硬件是平台，模型是应用。赌算力基建，如同 1998 年赌互联网骨干网。
模型在贬值，关系在升值，算力在两端都是刚需。



算力不仅是芯片，更是重资产的物理基建

行业分析显示，GPU 仅占大型集群总成本的约 40%。今天建的数据中心，明天跑新一代模型，基础设施的生命周期与价值远超远超单代芯片。



裸金属芯片（40%）

100 颗 H100 采购价约为 300 万美元

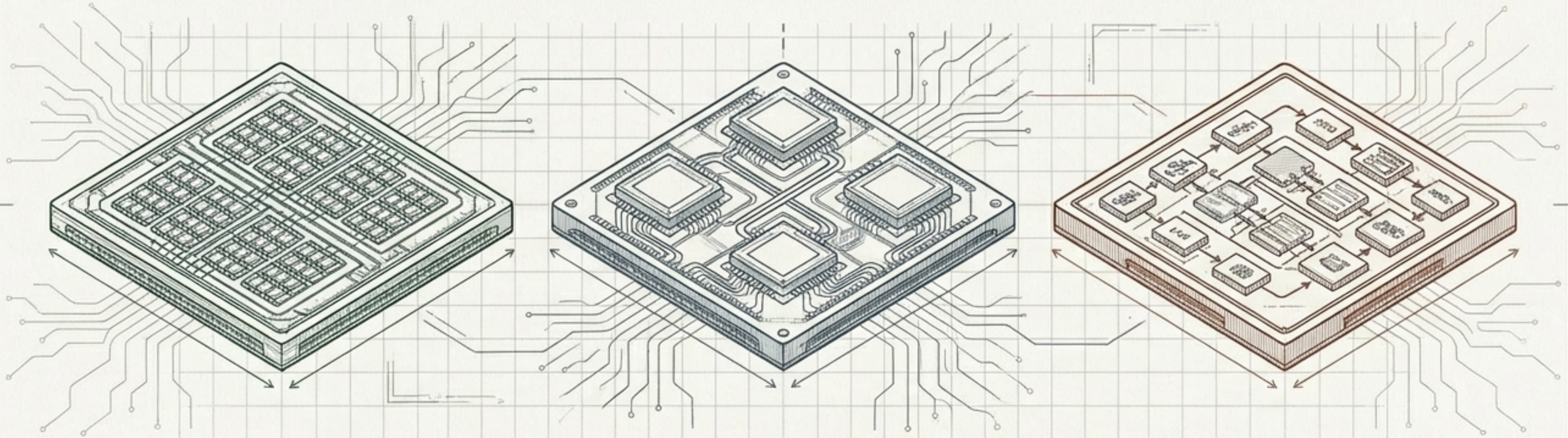
网络、电力、冷却与数据中心设施（60%）

隐性支出高达 560 万美元

5 年总拥有成本（TCO）= 860 万美元

芯片战争的本质，是世界观的赌注

架构之争并非纯粹的技术指标比拼，而是对“智能终极形态”的底层哲学预判。
每一块芯片，都编码了一种对未来的假设。



GPU
(Nvidia)

TPU
(Google)

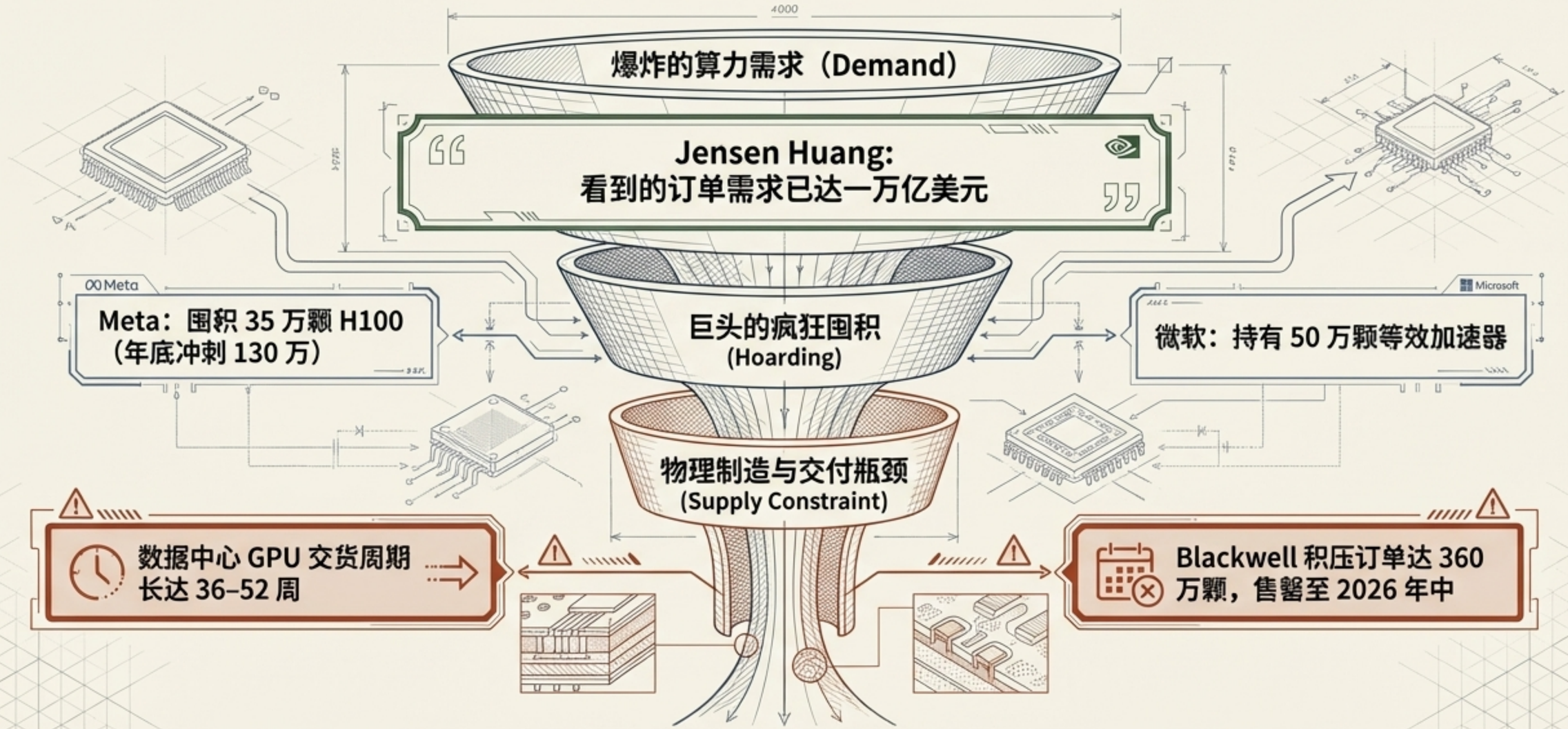
ASIC
(Groq)

架构之争背启之争背后的底层哲学

	GPU (Nvidia)	TPU (Google)	ASIC (Groq)
核心假设 (Worldview)	智能是发散的，未来极不确定	智能是收敛的，矩阵运算是核心	推理是主战场，计算模式已稳定
架构策略 (Strategy)	极致的灵活性与通用性	定制互联，超大规模系统级协同	摒弃冗余，专注纯推理速度
护城河 (Moat)	深不可测的 CUDA 生态锁定	闭环生态与极优的内部成本	针对特定场景的极致性价比
风险评级 (Risk)	极低（胜负在生态，不在实验室）	中（若 Transformer 统治延续则持续受益）	极高（极易受范式转移打击而报废）

万亿美元需求挤过现实的物理缝隙

芯片比电力更先成为瓶颈。谁能拿到算力、谁能高效使用算力，就是最实在的竞争优势。

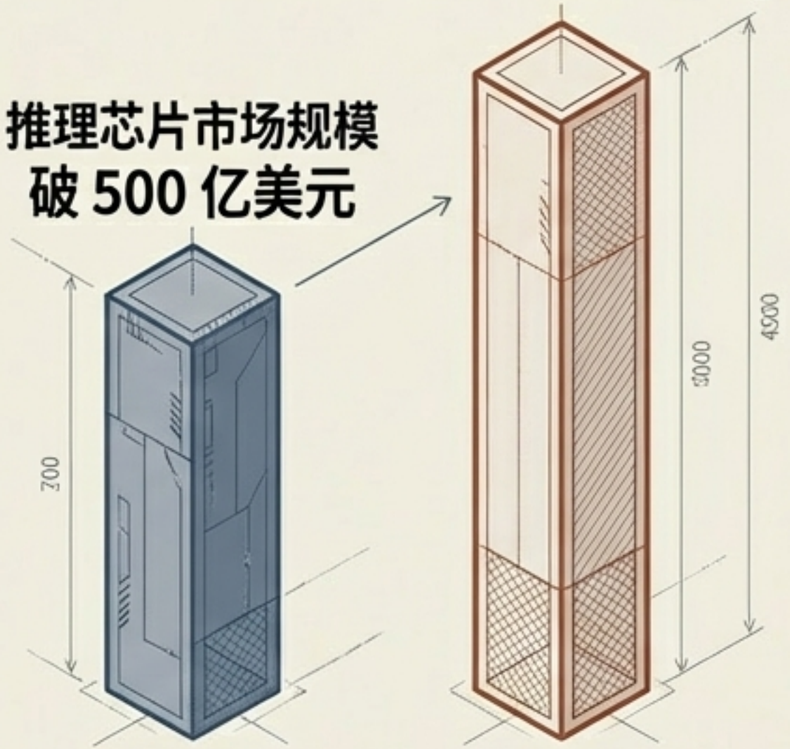


终局前瞻：推理主导与基建狂飙

随着模型继续变强与变便宜，纯 AI 在越来越多领域追平人机协作。资资本开支将以千亿级别持续注入底层物理基建。

全球 AI 数据中心资本开支达
4000-4500 亿美元

推理芯片市场规模
破 500 亿美元

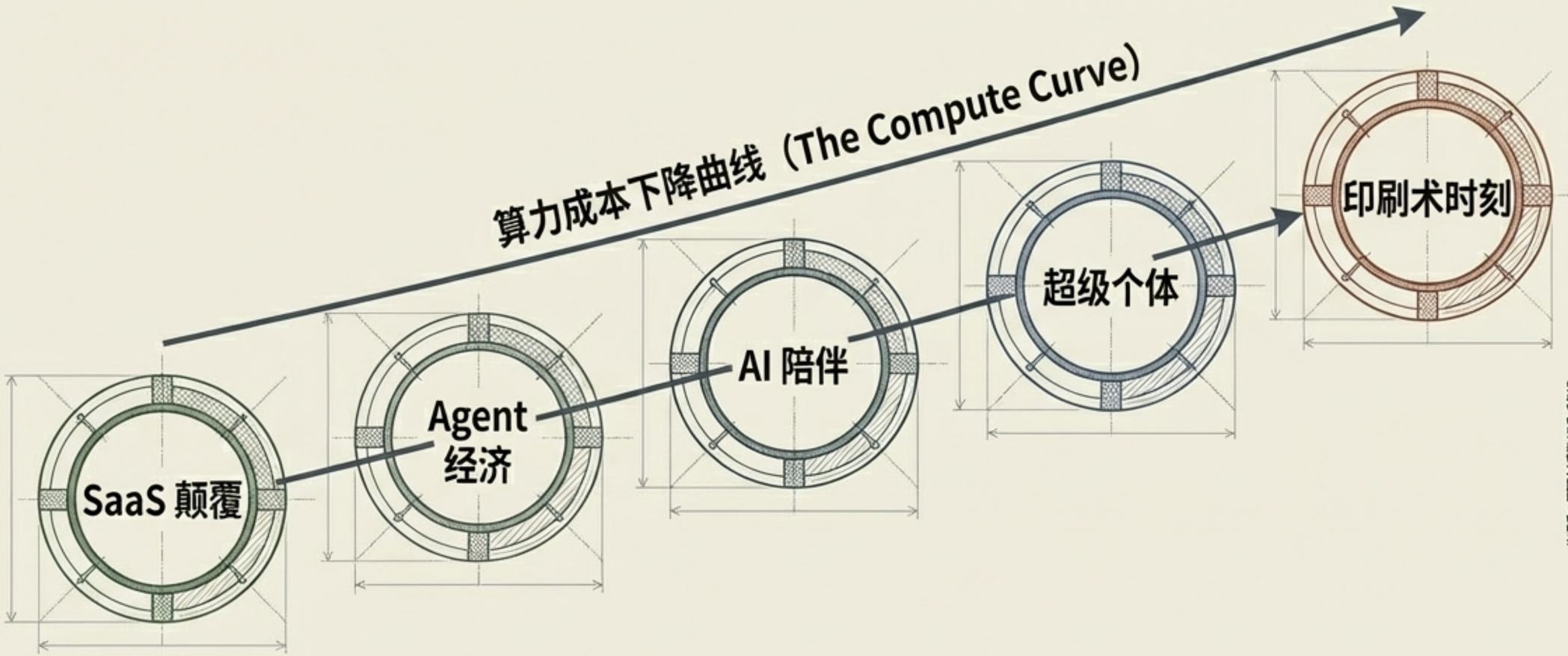


推理芯片市场规模　全球 AI 数据中心资本开支

Projections based on Deloitte data for 2026

万物归基：单一根变量的因果链

要预测 AI 的未来，你不需要猜测下一个爆款应用。如果理解算力曲线分别走到了什么位置，就能精确预判哪些应用即将爆发。



算力是根，应用是叶。

所有关于 AI 未来的争论——模型谁最强、赛道哪个火——拆到底，
答案都在同一个地方：算力够不够多，够不够好。
根往下扎多深，叶子就能伸多远。