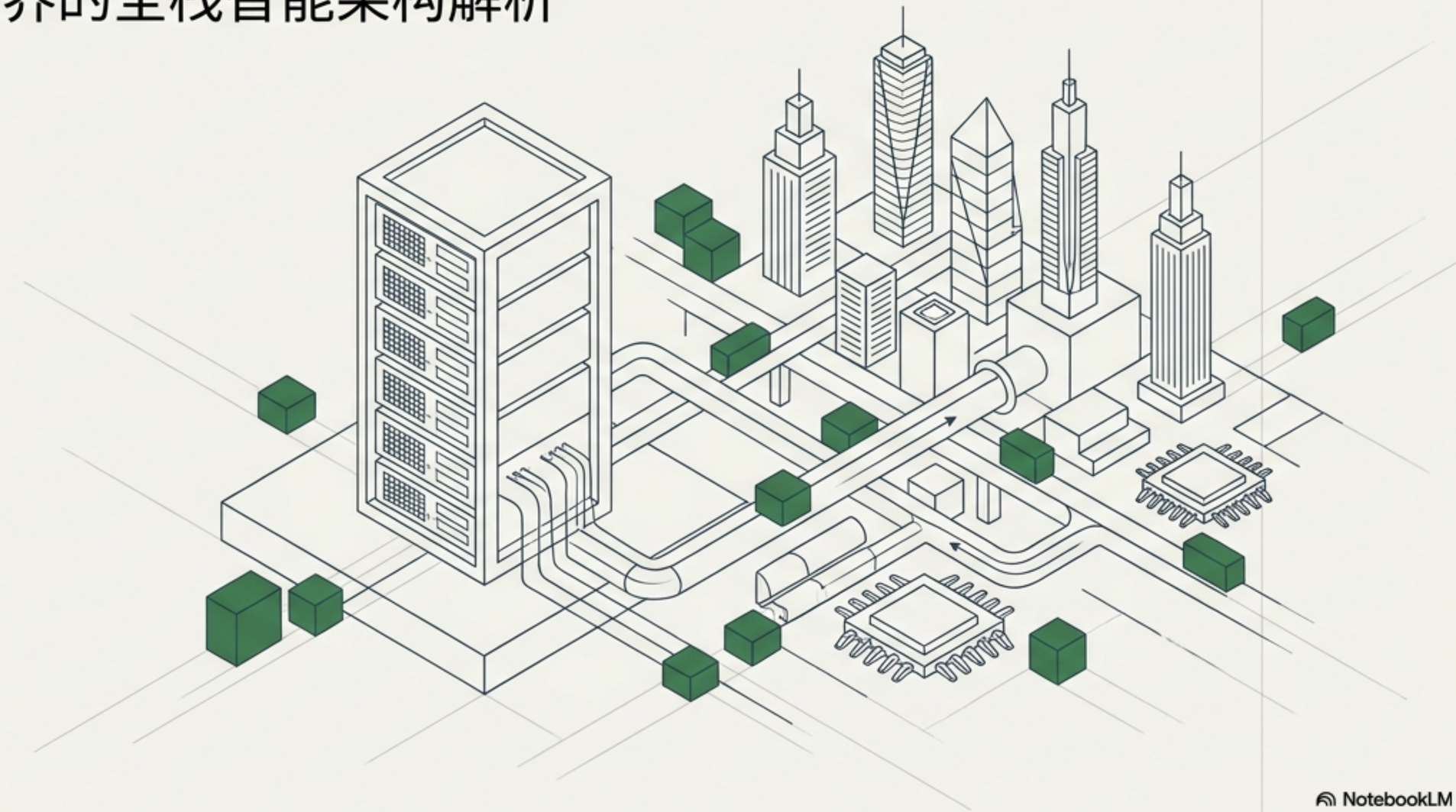
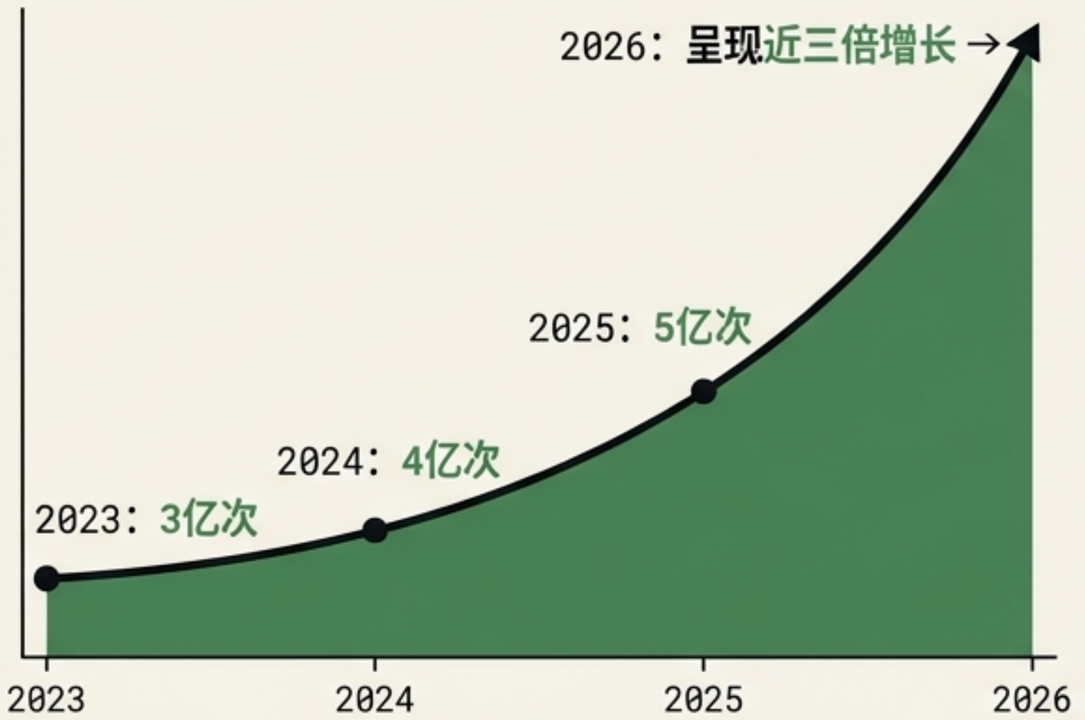
进入代理式AI时代：NVIDIA GTC 核心蓝图

从芯片到物理世界的全栈智能架构解析



算力即收入：指数级生产力革命

GitHub 代码提交量趋势



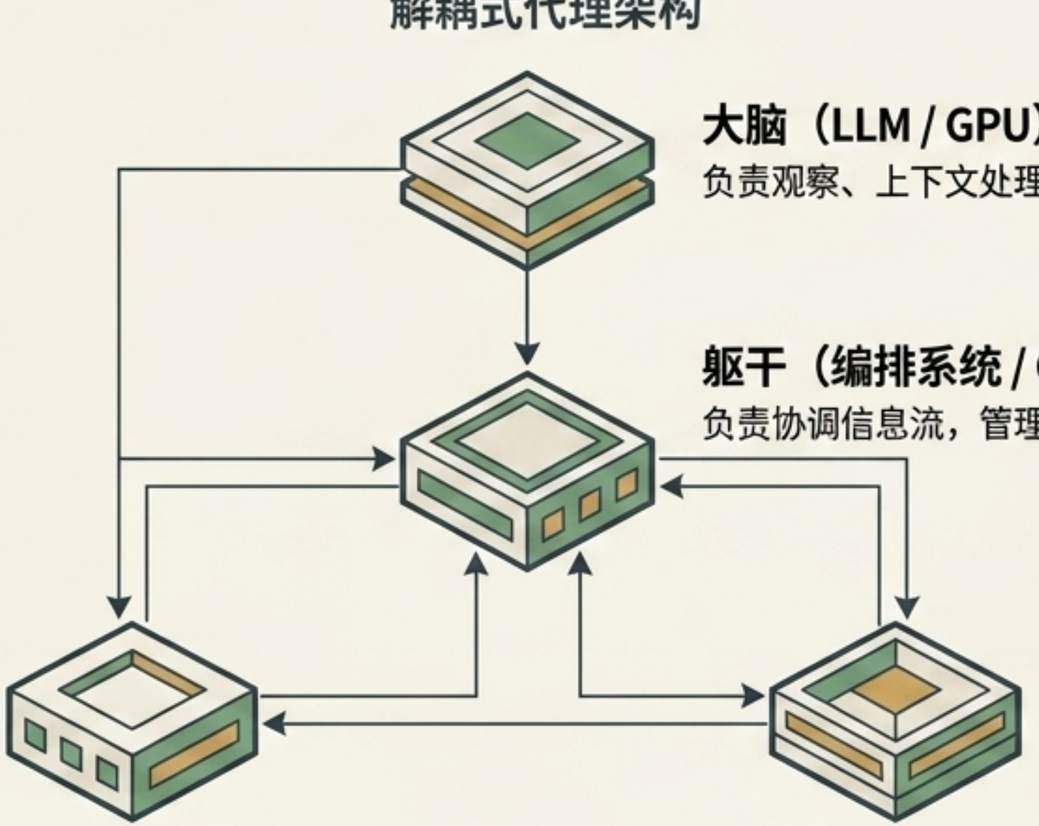
[3,000万开发者 + 3万亿美元薪酬] × [代理式AI] = 9万亿美元经济产出

在实用型AI时代，代币（Tokens）已成为可盈利的收入单元。构建AI工厂是为了制造利润，这就是全球算力需求激增的根本原因。

计算范式的根本性变革

传统PC时代	生成式AI时代	代理式AI时代
交互方式：启动应用，点击与与打字	交互方式：提示词输入	交互方式：表达意图
输出形式：界面反馈	输出形式：文本/图像生成	输出形式：自主调用工具完成工作
核心逻辑：为人类速度设计（以秒计）	核心逻辑：预训练与问答	核心逻辑：观察、推理、规划、行动（以纳秒计）

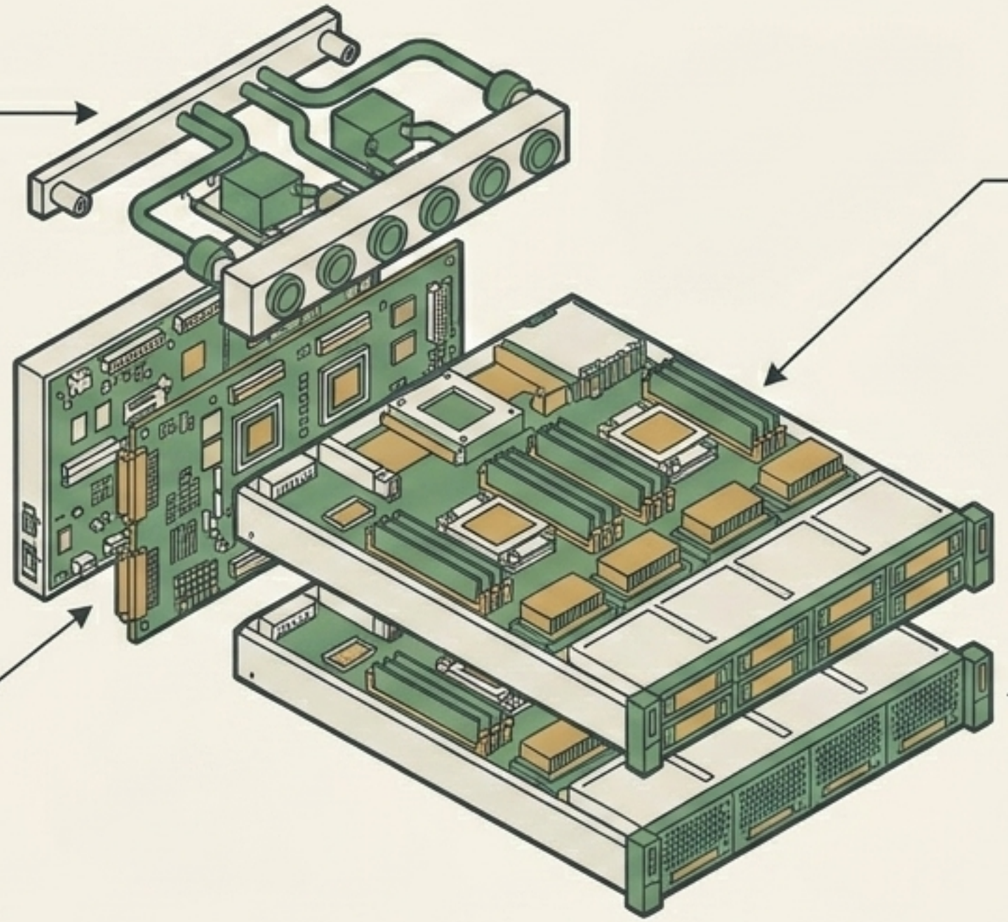
代理的解剖学：解耦式计算集群



这种分布式、异构的计算回路，要求系统在瞬间激活数以万计的节点。传统的单体计算机已无法胜任。

Vera Rubin：专为代理打造的超级计算机

极致协同
基于机架级模块化设计



核心引擎
NVLink 72 提供极高吞吐量的代币生成

无缆化设计
全新PCB中板连接，消除电缆瓶颈，极高提升集群稳定性

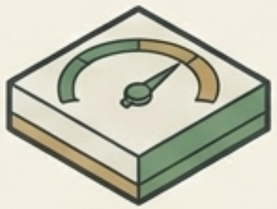
极速记忆
Grace Bluefield 4 DPU 实现网内计算与硅级安全加密

组装时间从2小时缩短至5分钟。目前已全面投入量产。

Vera CPU：为代理而非人类设计

人类级 CPU vs. 代理级 CPU		
用户主体	租户	极度缺乏耐心的AI子代理
时间尺度	秒级响应	纳秒级响应
核心策略	按小时切分出租	88个Olympus核心组成的单一网格

Vera 架构四大支柱



极高IPC

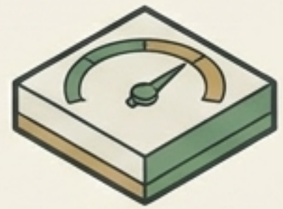
每时钟周期获取/解码/执行10条指令，拥有卓越的单线程性能

极低延迟

率先采用 LPDDR5X，提供 1.2 TB/s 带宽，内存延迟降低40%

海量带宽

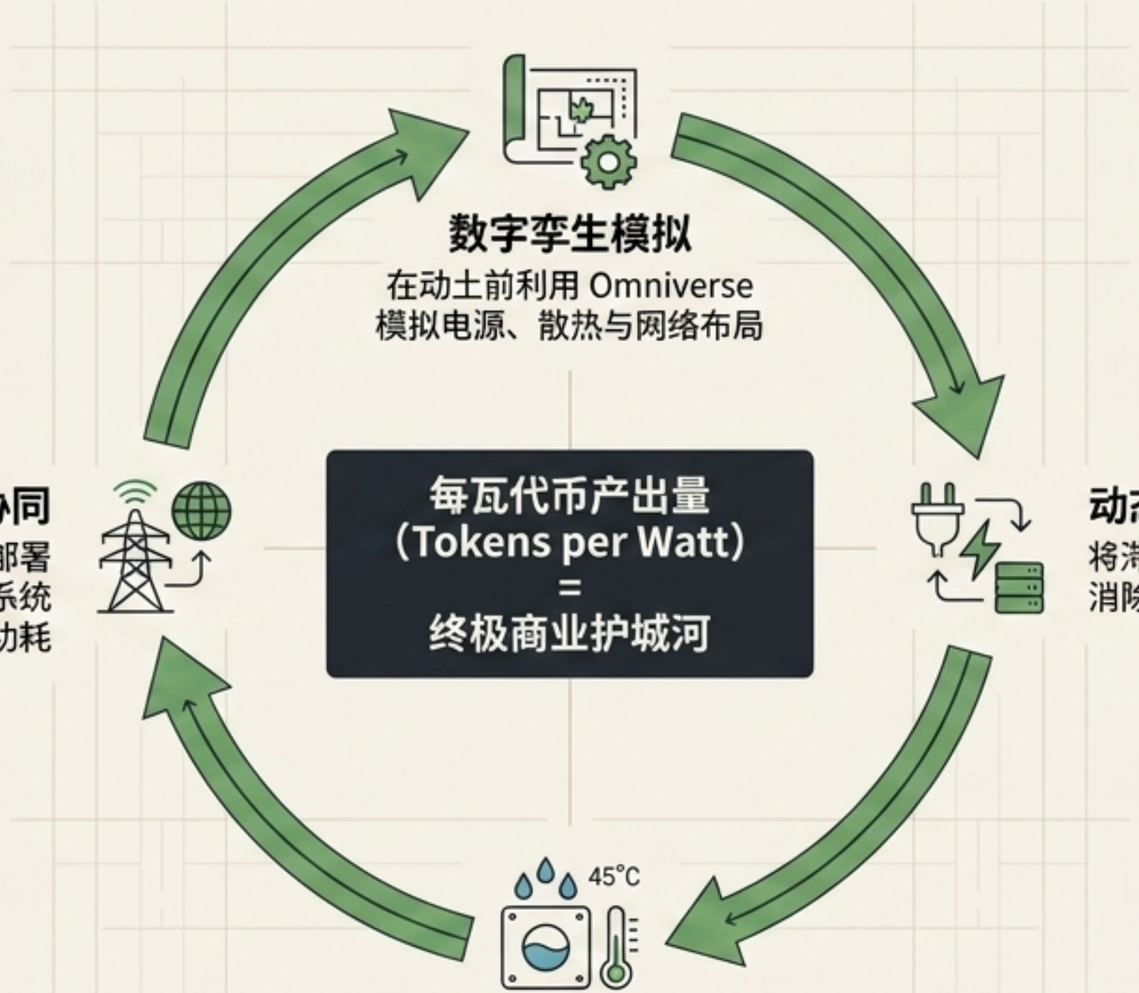
3.6 TB/s 光速互联架构，PCIe Gen 6，无小芯片跨界损耗



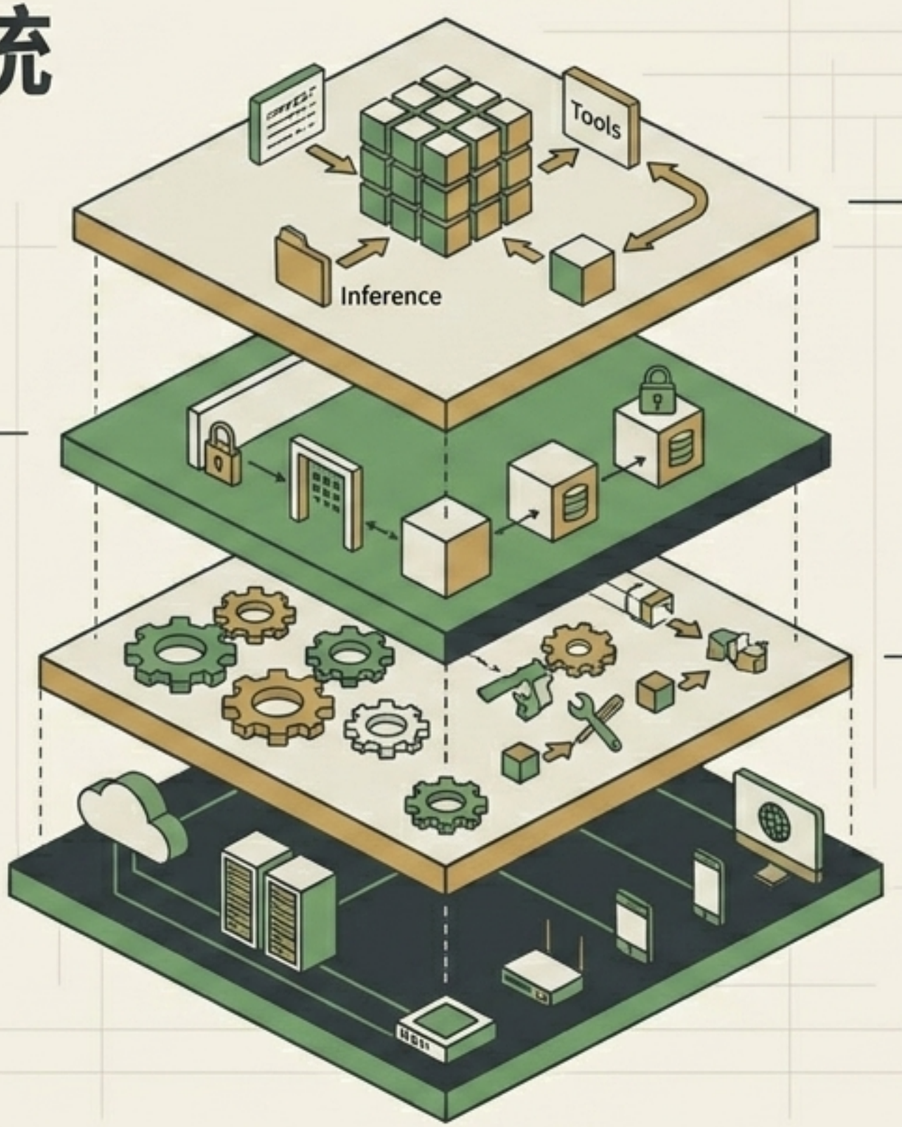
能效极限

节省能耗用于GPU吞吐，最大化工厂代币产出利润

AI 算力工厂蓝图：能源即代币即利润



NVIDIA 企业级代理工具包：AI时代的操作系统



开放模型 (Neotron)
专为复杂推理与工具调用优化的全开源基础模型

编排安全带 (OpenShell)
开源沙盒系统，确保代理权限可控、数据隔离与安全接地

技能与工具 (CUDA X)
数以千计的底层计算库，直接赋予AI解决难题的“双手”

运行时环境
贯穿云、本地到端侧设备的统一运行环境

被全球企业生态全面采纳
（如 Red Hat，Microsoft 等）

开源边疆与超级代理应用

Neotron 3 Ultra 核心优势

- 首创混合架构：结合 SSM 与 MoE
- 性能飞跃：运行速度提升 5倍
- 极致降本：运行成本降低 30%
- 完全开放：提供模型、训练脚本及对应数据集

Cadence 芯片设计超级代理

传统挑战：芯片RTL验证无容错率

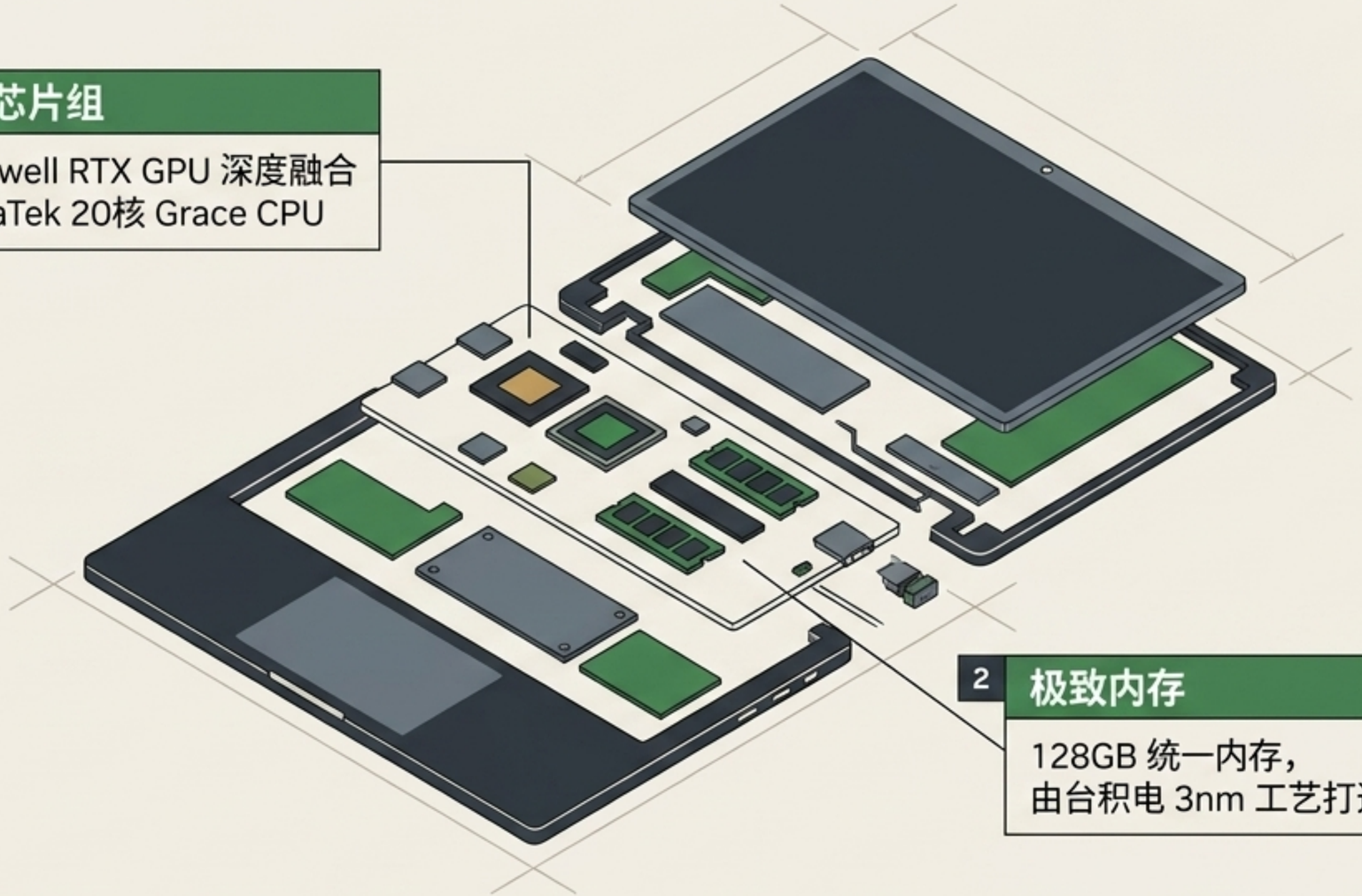
代理介入：驱动专家子代理自动排查漏洞

验证周期实现超 40 倍加速

RTX Spark：PC 40年来的最大重构

1 联合芯片组

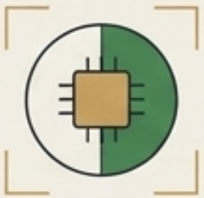
Blackwell RTX GPU 深度融合
MediaTek 20核 Grace CPU



2 极致内存

128GB 统一内存，
由台积电 3nm 工艺打造

1



永远在线的大脑

7x24小时在后台沙盒中运行，无计费焦虑

2

全场景兼容

原生运行由大模型驱动的本地工作流

你的电脑不再是启动软件的工具，而是一个拥有常驻数字助手的AI超级节点。

物理世界 AI 的瓶颈：算力即数据

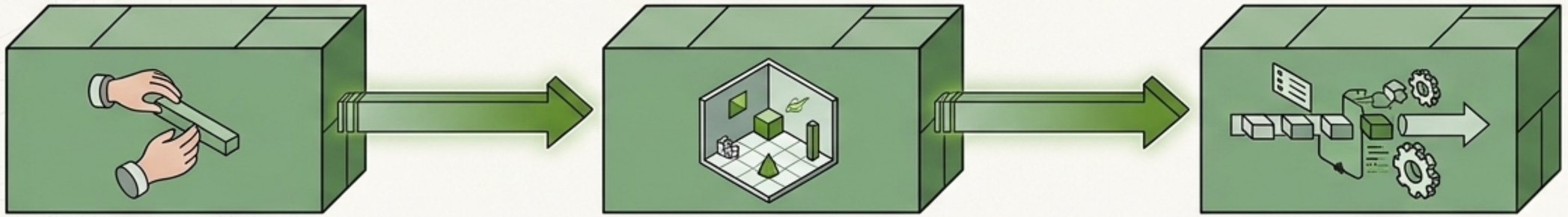
传统数据局限

第三人称视角

互联网上的语言与视频数据为人类视角，机器人无法直接应用

第一人称视角合成

利用算力生成数据

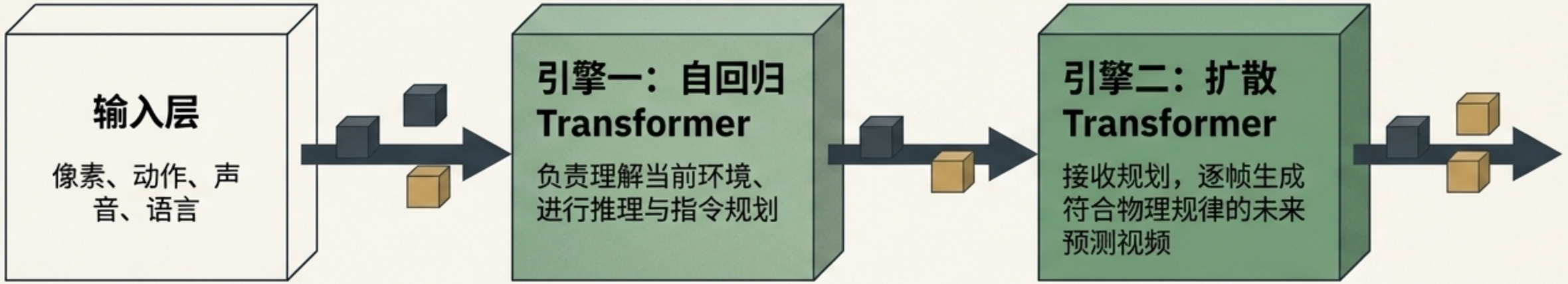


人类演示

Omniverse 仿真

生成物理精准的合成训练数据

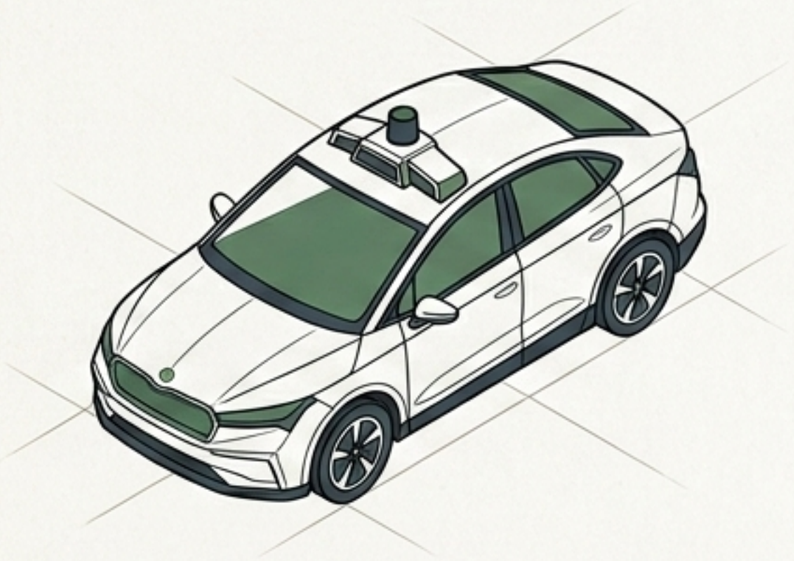
Cosmos 3：物理世界的万物模型



视觉语言模型 （描述环境）	**世界模型** （生成合成视频）
模拟器 （闭环评估策略）	**动作生成器** （控制实体机器人）

从自动驾驶到人形机器人：物理 AI 的载体

Alpha Mayo 2 自动驾驶

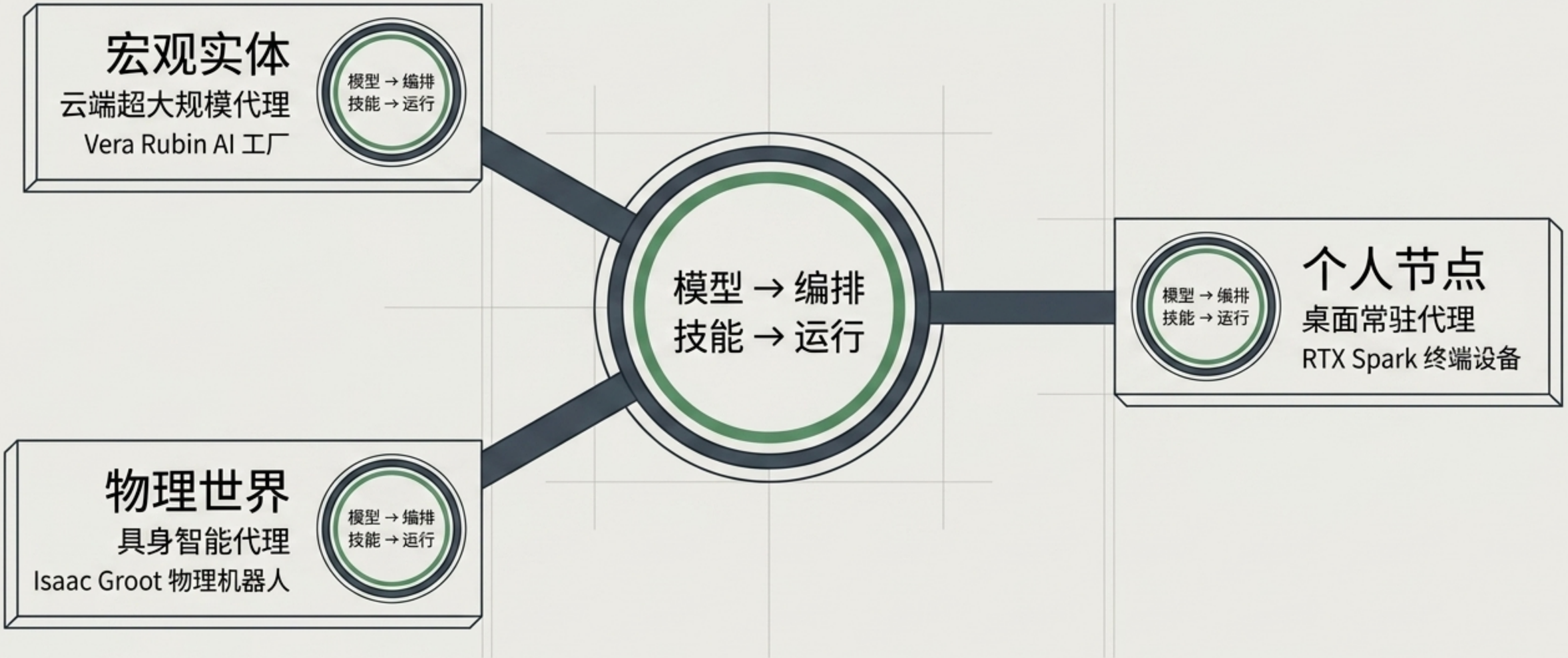


- 首个具备推理能力的自动驾驶基础模型
- 覆盖全球80%汽车制造商与97%的出行服务
- 在复杂路况中自我对话、推理并做出决策

Isaac Groot 人形机器人

- 开箱即用的参考设计平台
- 高 6英尺，重 150磅，全身 31个自由度
- 运行 Jetson Thor，集成 NVIDIA 全套软件栈

统一范式，无处不在：跨越所有尺度的智能回路



不管是在云端、桌面还是机器人体内，计算范式完全一致。这就是时代的统一底层逻辑。

NVIDIA：AI 时代的基建引擎

Point 1	不仅制造GPU，更是全栈式AI工厂架构师
Point 2	从底层的 Vera CPU 到跨设备的 Agent Toolkit，再到 Cosmos 基础模型
Point 3	构建将能源转化为终极通用代币的最高效基础设施

> **下一个十年，每一家公司都将是代理公司，而我们将为之提供完整的运作蓝图。**