

计算的炼金术

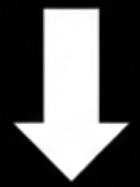
英伟达CES 2026 / PART 1: 物理AI革命

JENSEN HUANG | FOUNDER & CEO

一个历史时刻：两场平台变革同时发生

“每隔十年到十五年，计算行业会迎来新的平台变革。但这次，是两个。”

COMPUTING
计算



ACCELERATED COMPUTING
加速计算

APPLICATIONS
应用



GENERATIVE AI
生成式AI

整个技术栈，全面重构

CODE → **MODEL**
不再编写软件代码，而是训练软件模型

CPU → **GPU**
不再运行在CPU上，而是运行在GPU上

PRE-COMPILED → **GENERATED**
不再是预编译，而是实时生成每个像素与Token

TRADITIONAL → **ACCELERATED**
每次计算都因加速计算被彻底重塑

APPLICATION-FIRST → **AI-FIRST**
不再是构建应用，而是在AI之上构建新应用

$10,000,000,000,000

被现代化的计算价值。

资金来源：研发预算的转移（FROM TRADITIONAL METHODS TO AI METHODS）

2025：飞升之年

测试时扩展（TEST-TIME SCALING）

另一种说法是——实时思考。

代理系统（AGENT SYSTEMS）

具备推理、规划与工具使用能力。

物理AI（PHYSICAL AI）

理解物理定律并与世界交互的AI。

开源崛起（OPEN SOURCE ASCENDANT）

创新在每一家公司、每个行业同时被激活。

我们建造自己的AI超算。
目标只有一个：
锻造世界顶级的开源模型
(FORGE THE WORLD'S BEST OPEN SOURCE MODELS)

Nemo Tron
Neon
OpenFold
ForecastNet
Evl two
Alpha Mayo
...

NVIDIA 开源模型矩阵

领域	模型	关键能力
语言	Nemo Tron 3	混合SSM架构，实现极速推理
生物学	OpenFold 3	理解蛋白质结构与生成
物理世界	ForecastNet	革新天气预测方式

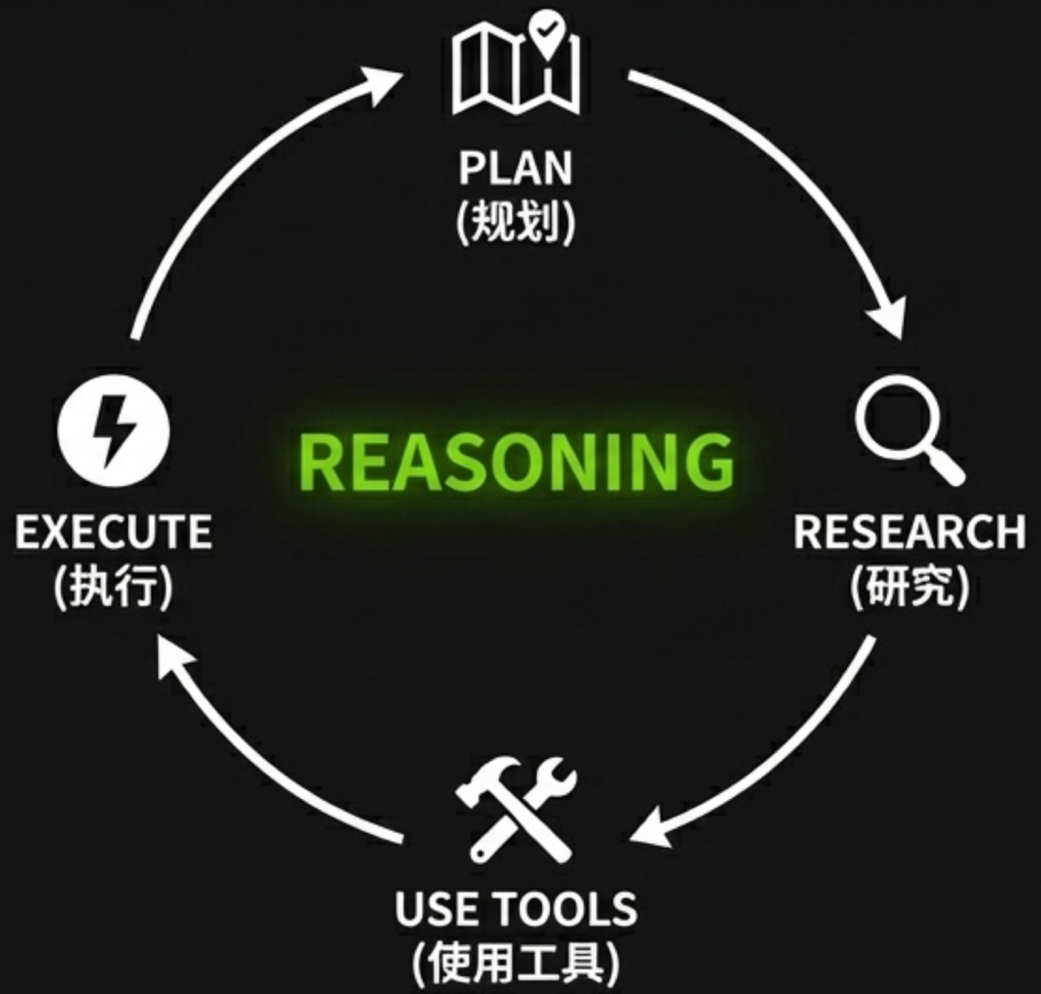
模型 + 数据 = 信任

(MODELS + DATA = TRUST)

新的原子单元：AI代理（THE AI AGENT）

超越记忆。走向推理。

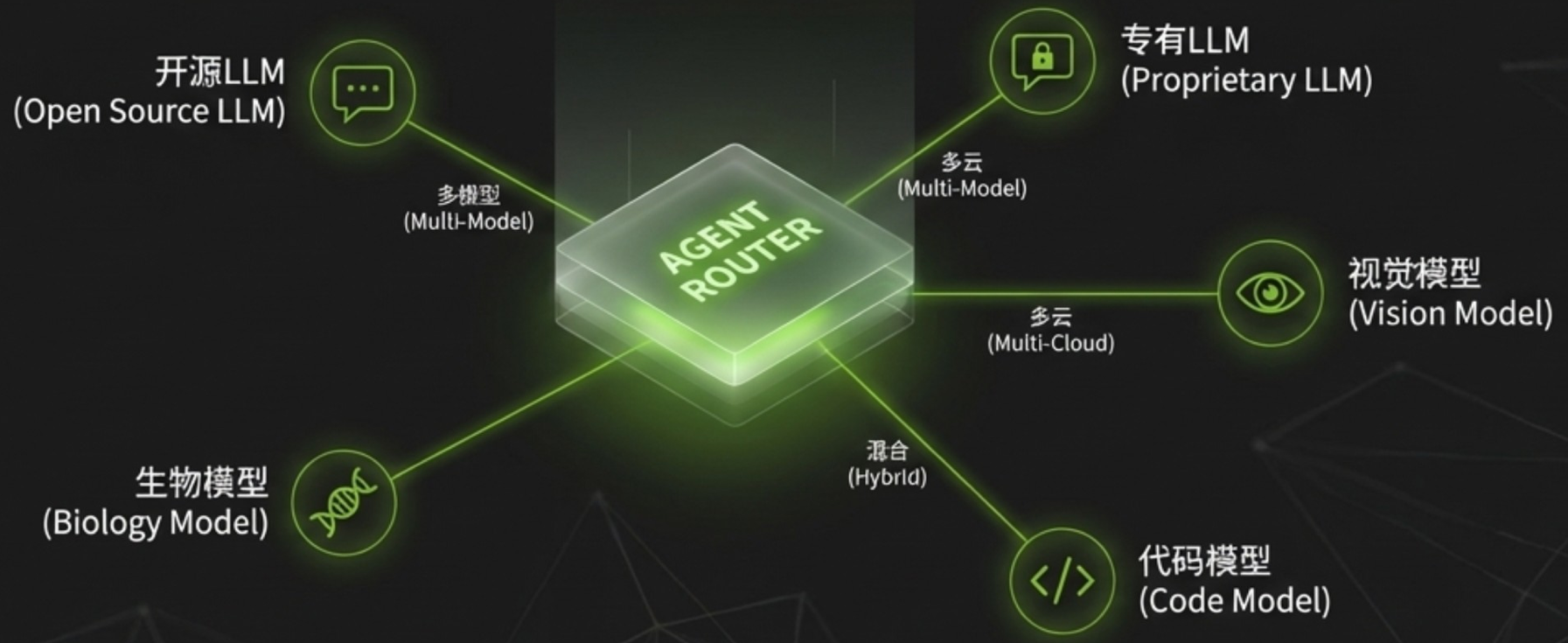
(BEYOND MEMORY. TOWARDS REASONING.)



突破：智能体路由（AGENT ROUTING）

一个模型远远不够。未来是专家的组合。

(ONE MODEL IS NOT ENOUGH. THE FUTURE IS A MIXTURE OF EXPERTS.)



NVIDIA 蓝图：构建你的专属AI



完全可定制
(FULLY CUSTOMIZABLE)

永远最前沿
(ALWAYS STATE-OF-THE-ART)

案例研究：个人AI助手“RICCI”

技术栈（THE STACK）

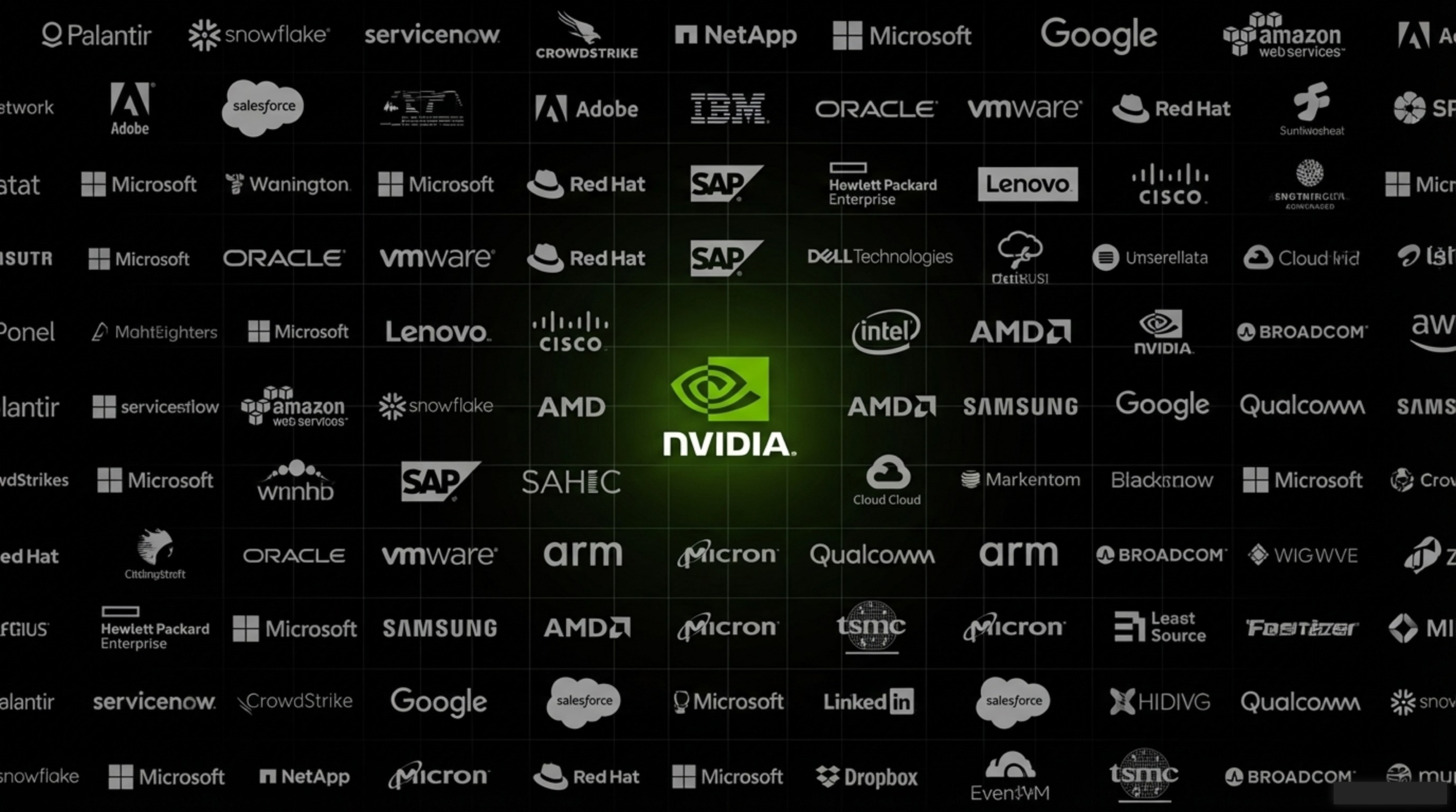
- LOCAL DELL DGX SPARK
- CLOUD APIS
- CLOUD APIS (ELEVENLABS)
- OPEN SOURCE MODELS



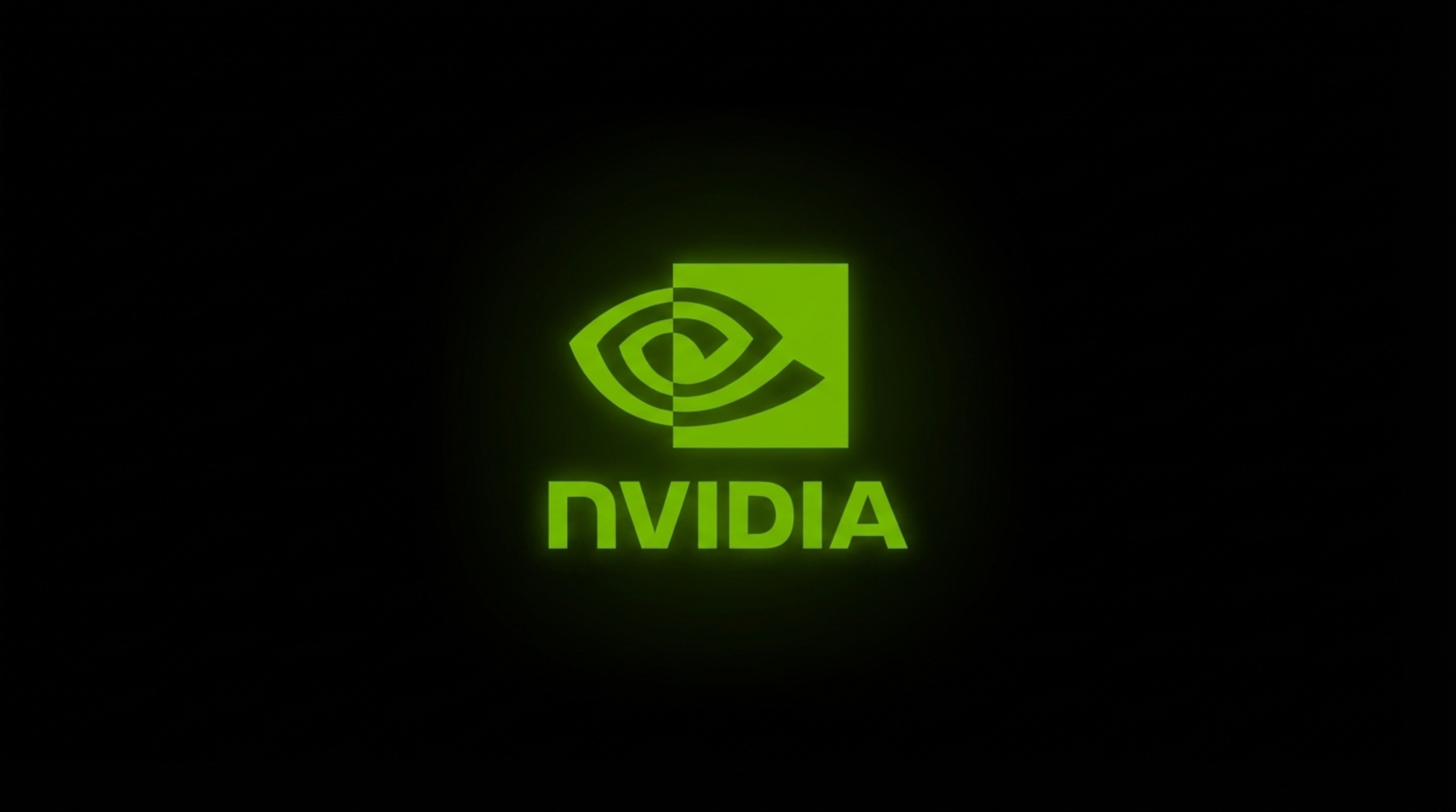
能力（CAPABILITIES）

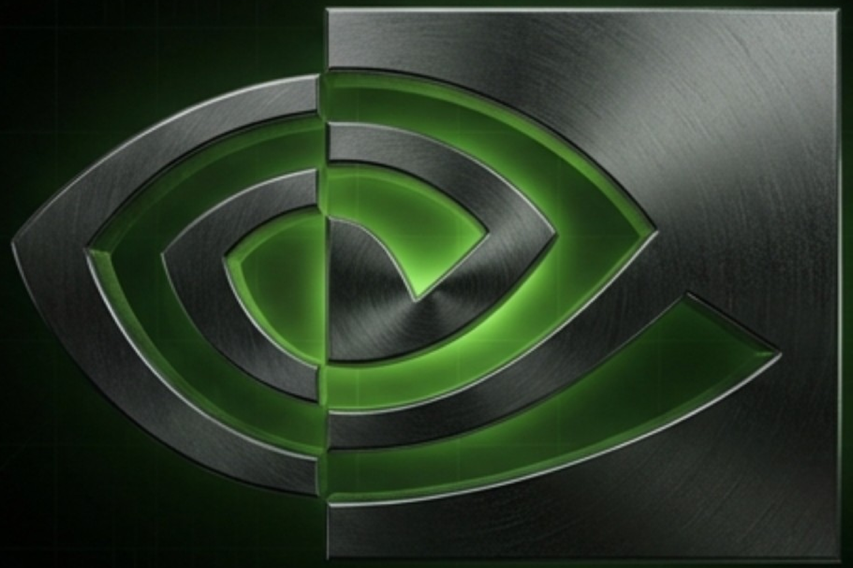
- 管理日程 (MANAGE SCHEDULE)
- 生成内容 (GENERATE CONTENT - from sketch to render)
- 控制机器人 (CONTROL ROBOTICS)

我们是世界领先平台的AI引擎。
(WE ARE THE AI ENGINE FOR THE WORLD'S LEADING PLATFORMS.)


Palantir
snowflake
servicenow
CROWDSTRIKE
NetApp
Microsoft
Google
amazon web services
Adobe
salesforce
IBM
ORACLE
vmware
Red Hat
Suntwosheat
Microsoft
Wanington
Red Hat
SAP
Hewlett Packard Enterprise
Lenovo
CISCO
Microsoft
ORACLE
vmware
Red Hat
SAP
DELL Technologies
Unserellata
Cloud Hid
Microsoft
MahtEighters
Lenovo
CISCO
intel
AMD
NVIDIA
BROADCOM
servicesflow
amazon web services
snowflake
AMD
NVIDIA
AMD
SAMSUNG
Google
Qualcomm
Microsoft
wnnhb
SAP
SAHIC
Cloud Cloud
Markentom
Blacksnow
Microsoft
ORACLE
vmware
arm
Micron
Qualcomm
arm
BROADCOM
WIGWVE
Hewlett Packard Enterprise
Microsoft
SAMSUNG
AMD
Micron
tsmc
Micron
Least Source
servicenow
CrowdStrike
Google
salesforce
Microsoft
LinkedIn
salesforce
HIDIVG
Qualcomm
Microsoft
NetApp
Micron
Red Hat
Microsoft
Dropbox
tsmc
BROADCOM

未来的交互界面
不是菜单或命令行。
系统本身就是交互界面。
(THE SYSTEM IS THE INTERFACE.)


NVIDIA



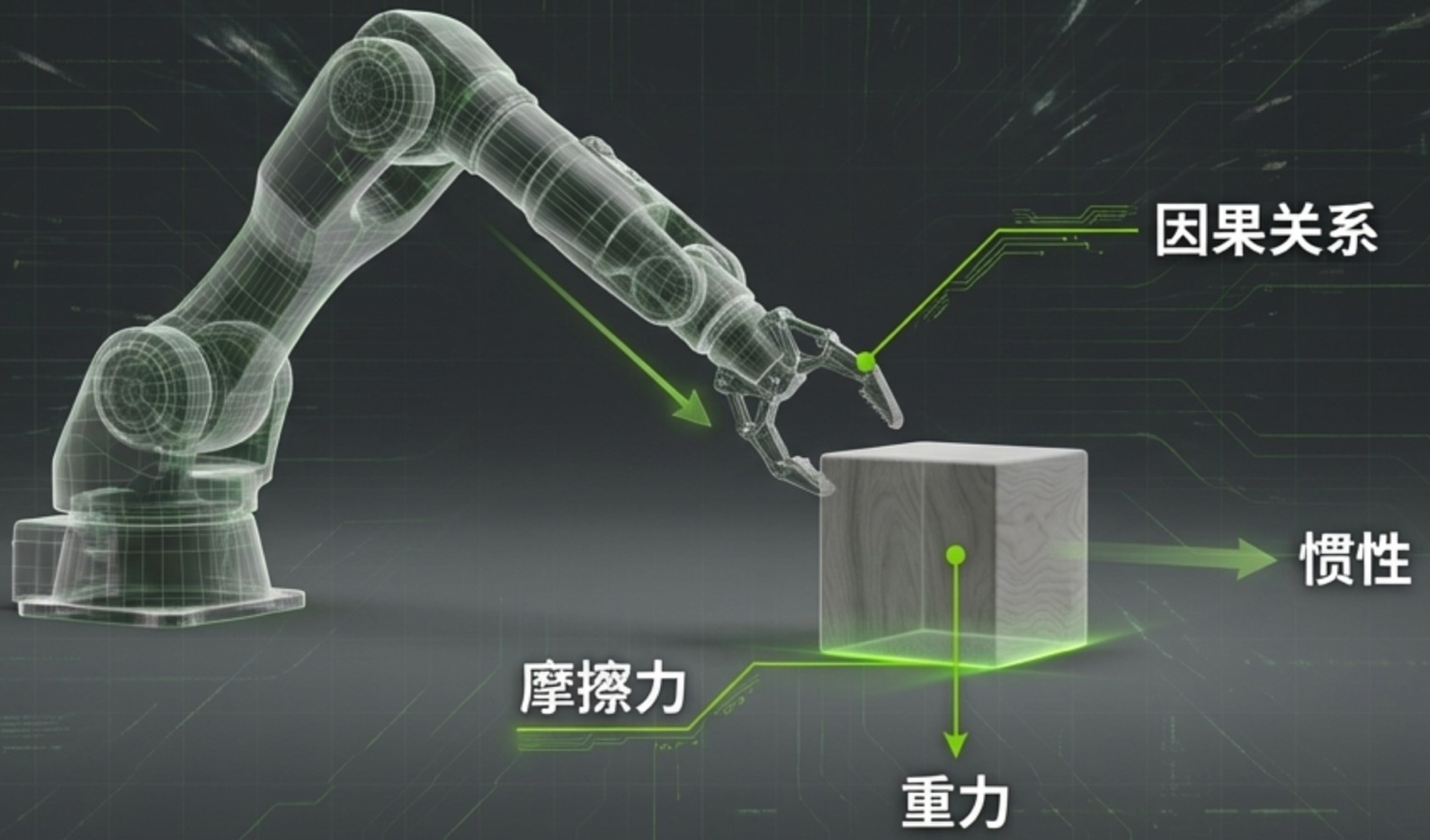
物理AI革命

计算的炼金术

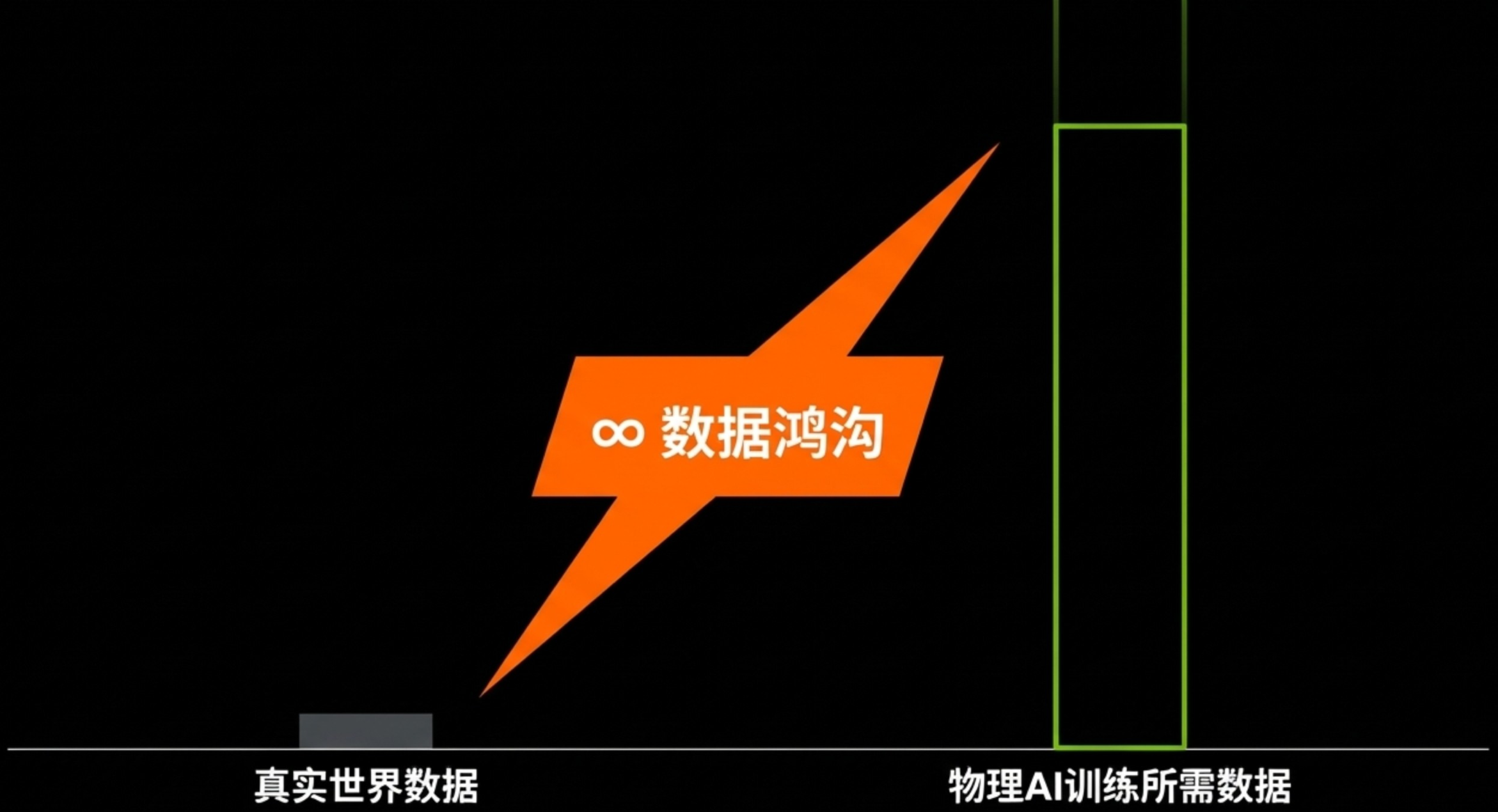
JENSEN HUANG | FOUNDER & CEO

AI的智能
被困于屏幕之内

缺失的物理常识

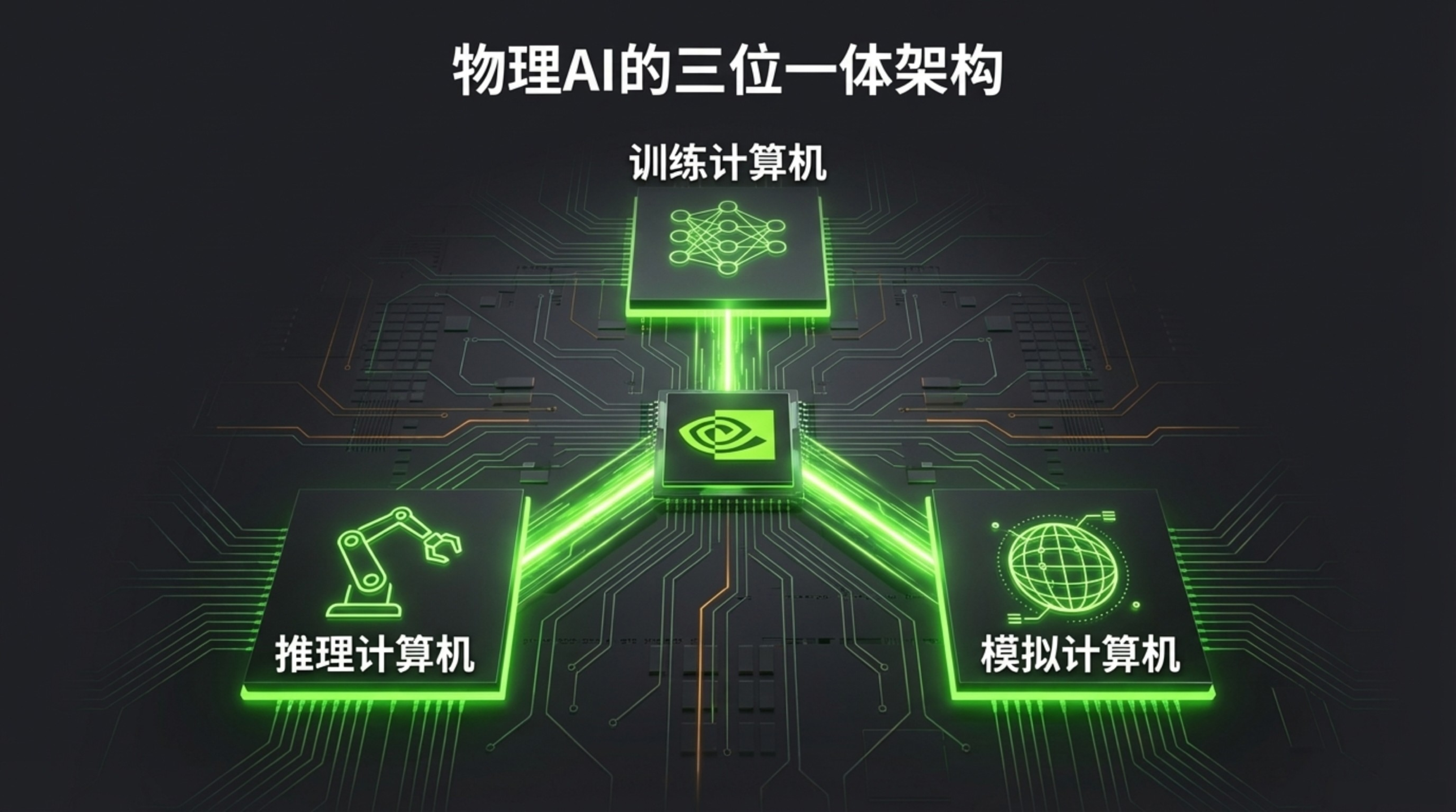


这些常识连幼儿都懂，但对AI来说完全未知。



收集真实训练数据：缓慢、昂贵、永不满足.

物理AI的三位一体架构



一切的基础



“如果没有模拟能力，
无法感知物理世界对动作的反馈。
无法评估效果。”

模拟是物理AI几乎所有工作的基础。



COSMOS: 世界基础模型

物理AI的“CHATGPT”时刻

计算炼成数据



基于物理规律

无限多样性

按需智能生成

COSMOS: 核心能力

- 生成式物理AI技能
- 对齐语言、图像、3D与动作
- 生成照片级、物理连贯视频
- 交互式闭环模拟
- 推理、分析与预测

数万亿英里

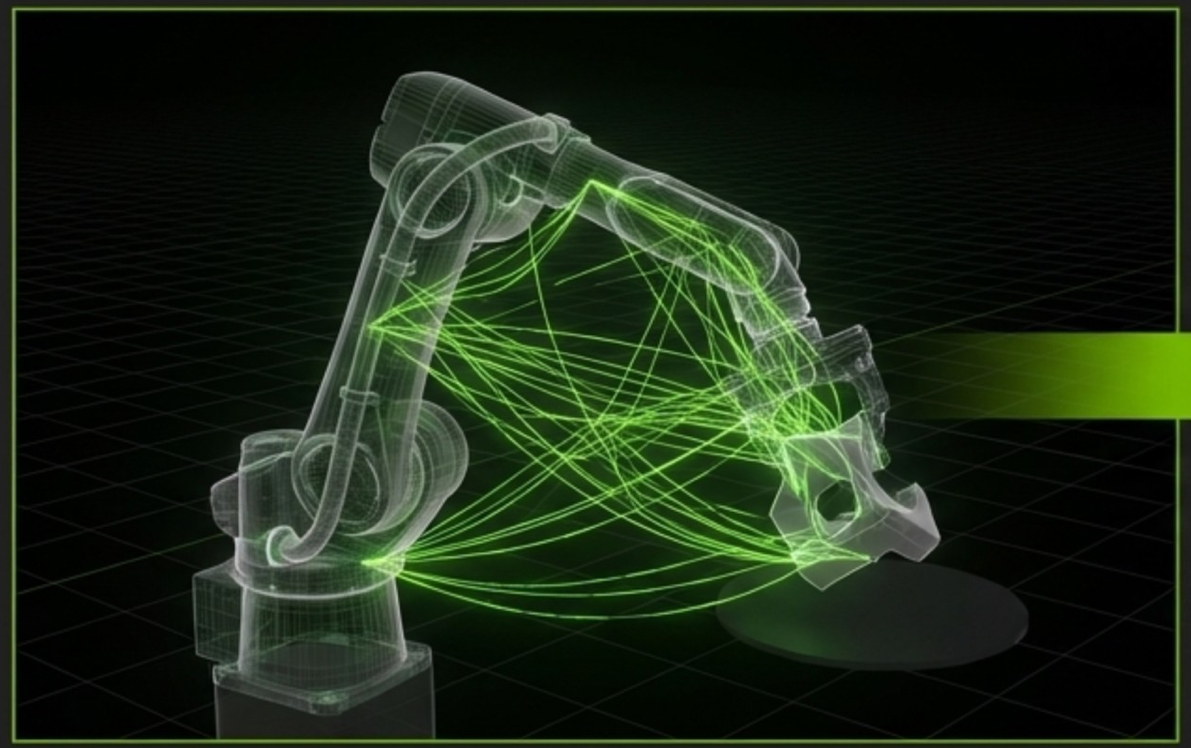
THE MORE YOU COMPUTE,
THE MORE YOU LEARN



物理路测

COSMOS 模拟里程

教会机器人适应每一种场景



模拟中学习



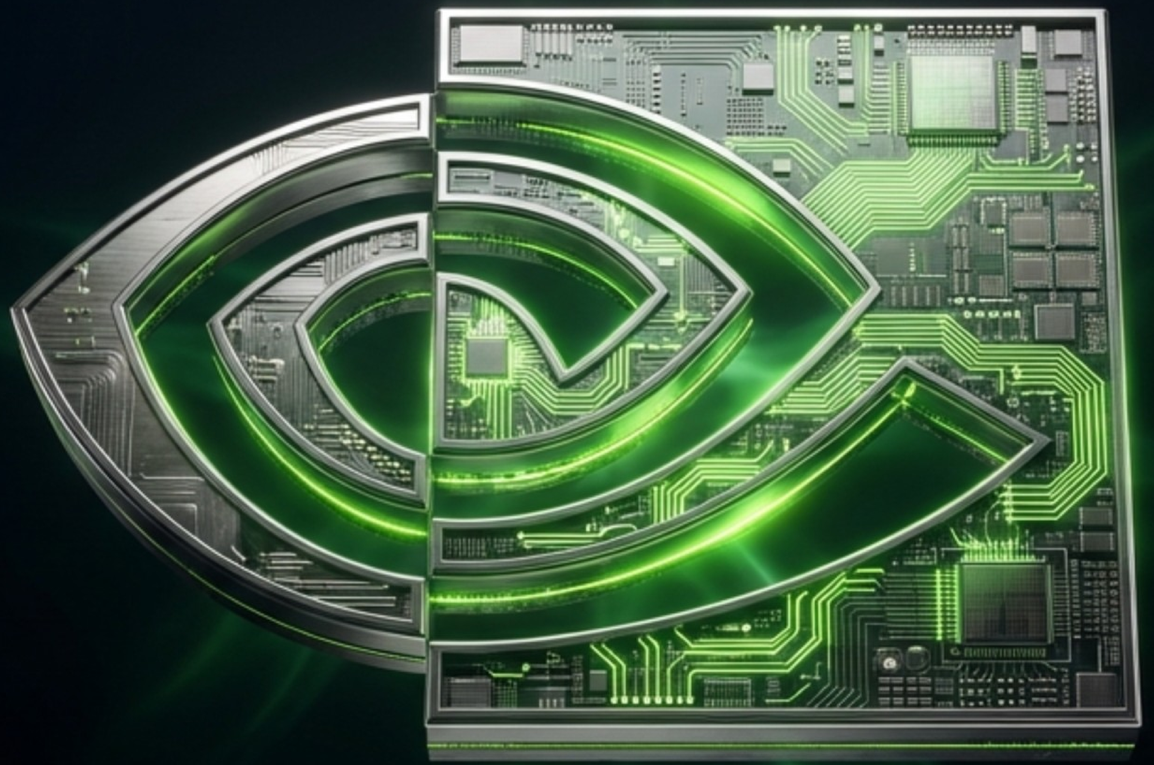
现实中适应

数百万

全球下载与部署


aws
Azure
Google Cloud
TESLA
FANUC
ABB
ASKAWA
YASKAWA
DELL
Technologies
hp
Lenovo
unity
FANUC
YASKAWA
DELL
Technologies
hp
Lenovo
Adobe
AUTODESK
unity
Microsoft
Microsoft
Booking
Microsoft
Dedeal
NVIDIA
ASEISN
Adobe
AUTODESK
unity
NVIDIA
BOOKET
unity
AUTODESK
Dummy
Microsoft

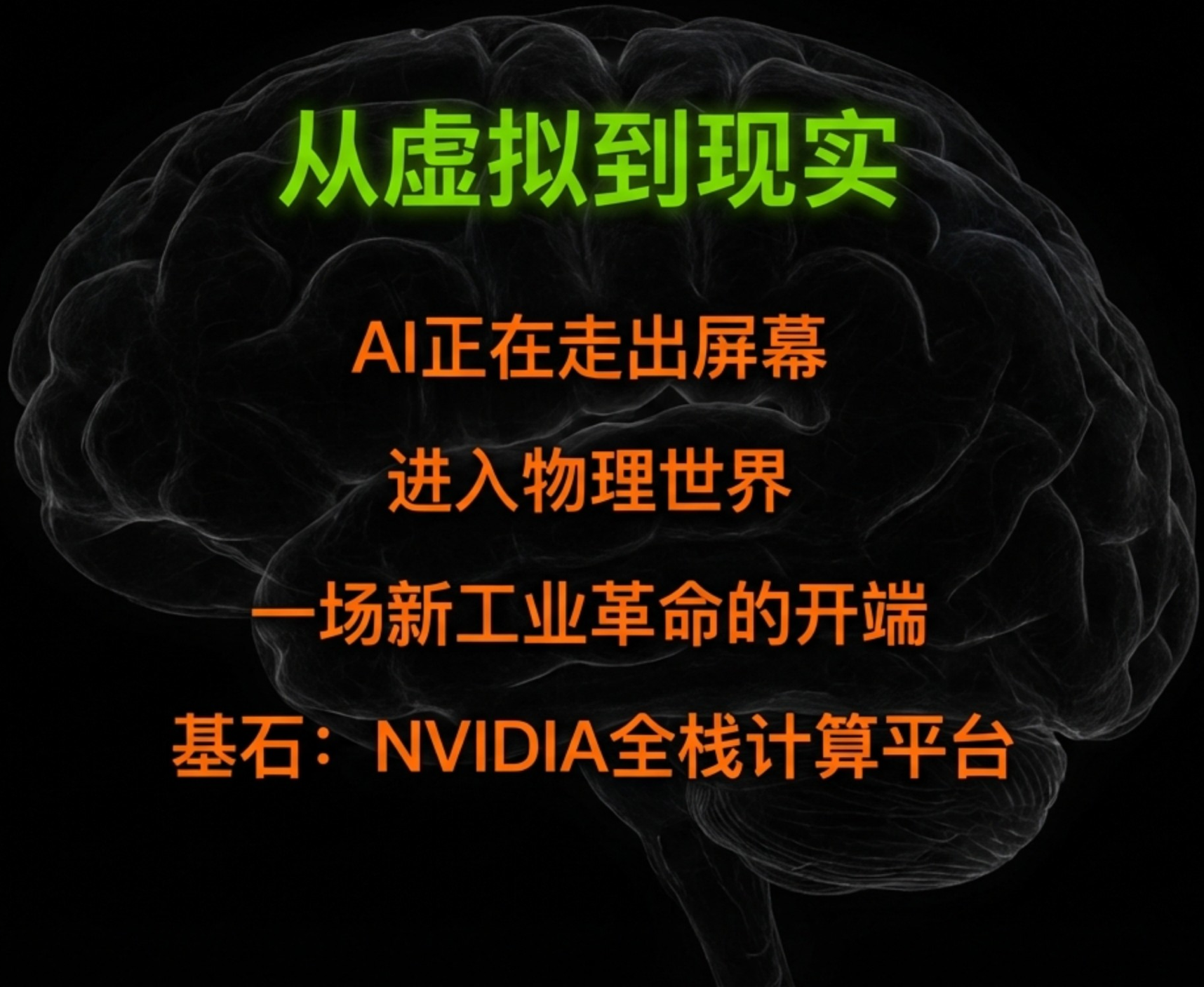

为物理AI新时代
做好准备



物理AI革命

英伟达CES 2026 技术深度解析 | PART 3

JENSEN HUANG | FOUNDER & CEO



从虚拟到现实

AI正在走出屏幕

进入物理世界

一场新工业革命的开端

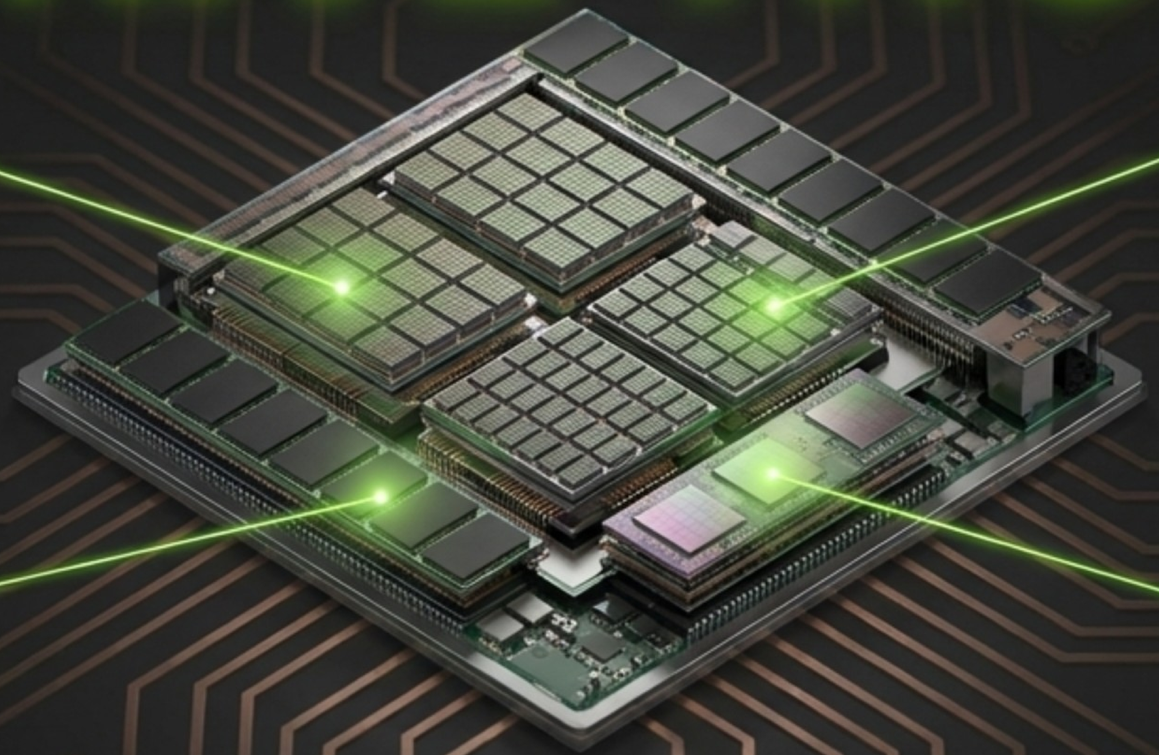
基石：NVIDIA全栈计算平台

地基：为物理AI打造的芯片

DRIVE THOR

双处理器冗余

专为机器人系统设计



最高功能安全

代码级安全认证

大脑：全球首个推理型AI



端到端训练

摄像头输入 -> 行动输出

推理下一步行动及原因

解决长尾场景难题

长尾难题的终结

传统方法

ALPHA MAYO 推理



我们无法收集所有场景
但AI可以推理所有场景

身体：首个大规模物理AI市场



1亿辆

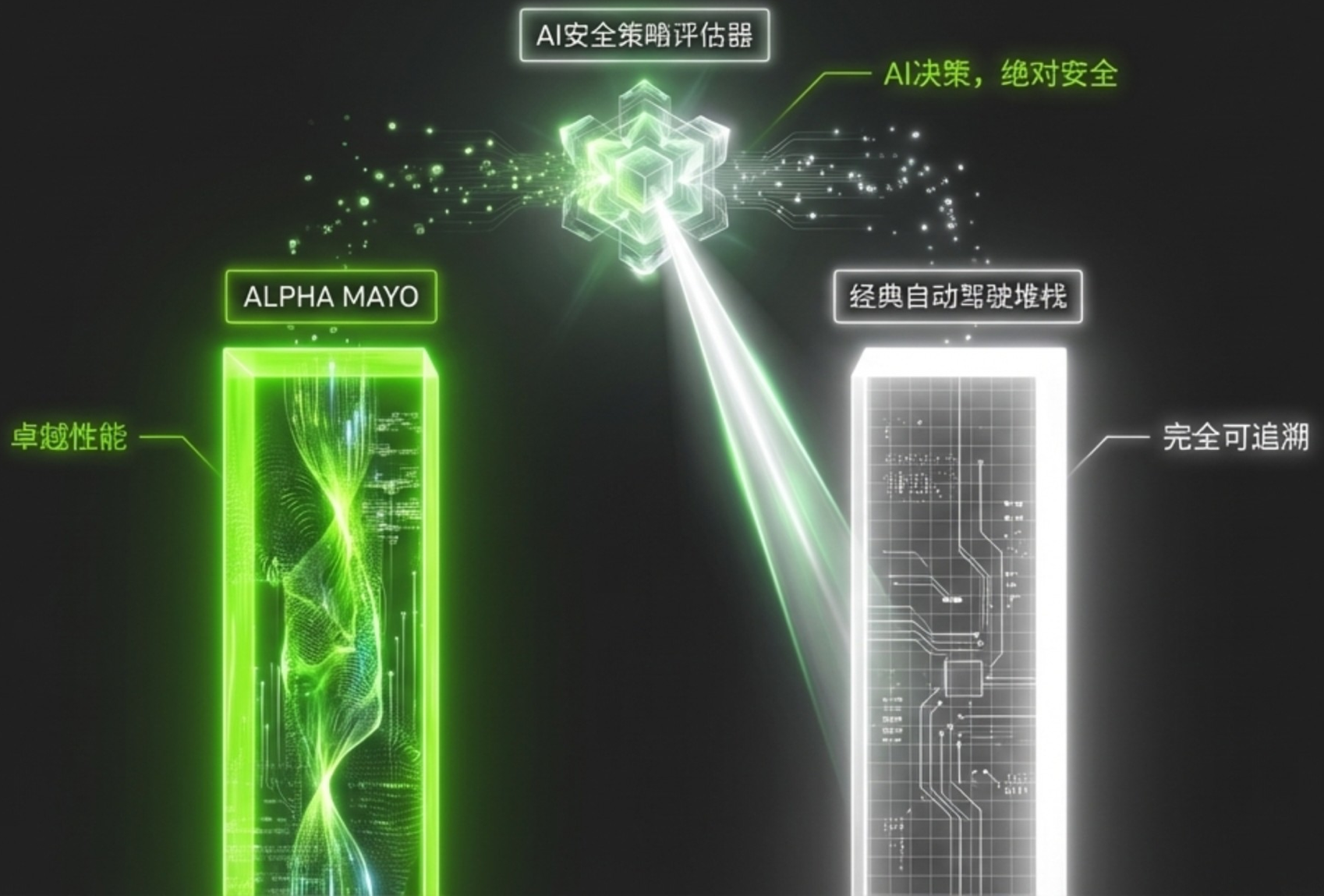
梅赛德斯-奔驰 战略合作

2026 Q1 全球上路

持续OTA更新Alpha Mayo

NVIDIA

守护：双堆栈安全冗余



蓝图：从轮式机器人到万物机器人



自动驾驶的技术
适用于所有机器人系统
机器人产业的黄金时代

宇宙：机器人的创世引擎

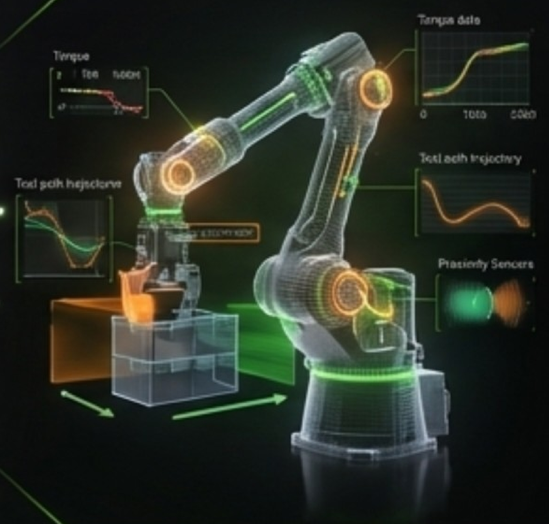
在元宇宙中训练



物理级精准仿真



海量合成数据生成



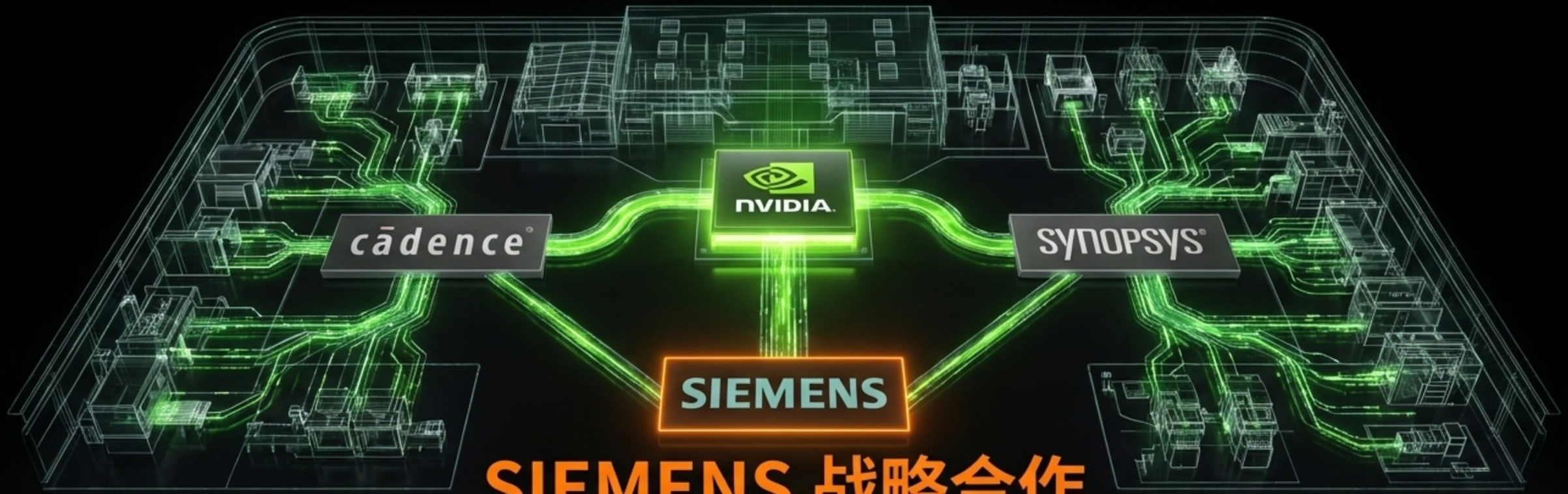
安全 | 高效 | 可扩展

生态：机器人的设计者



物理AI 融入芯片与系统设计
在仿真中设计、验证未来机器人

顶峰：新工业革命



SIEMENS 战略合作

NVIDIA AI 全栈集成

贯穿设计、生产、运营全生命周期

未来工厂 = 巨型机器人

闭环：物理AI的飞轮



设计 -> 仿真 -> 建造 -> 运营

NVIDIA 加速每一个环节

一个由AI驱动的全新工业时代

“我们站在新工业革命的起点。
由英伟达与西门子共建的
物理AI时代。”



计算的炼金术

THE ALCHEMY OF COMPUTATION

英伟达CES 2026 Part 4: VERA RUBIN 平台深度解析

JENSEN HUANG | 创始人 & CEO

AI需求的指数级挑战

10×/年

模型规模增长

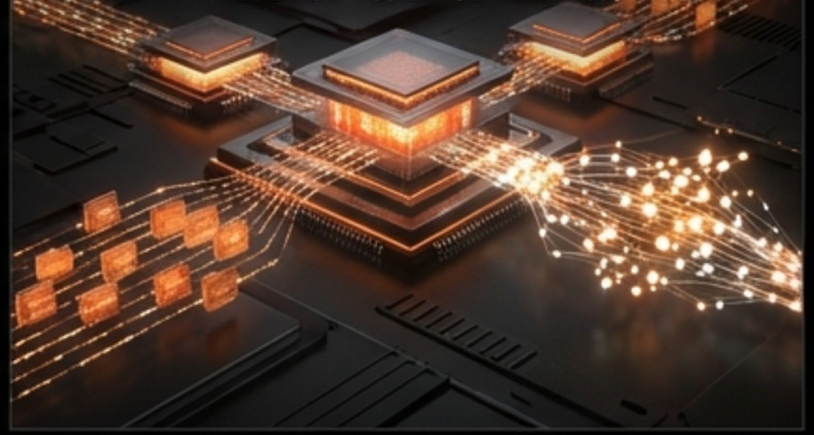
MODEL SCALE GROWTH



5×/年

推理思考令牌生成量

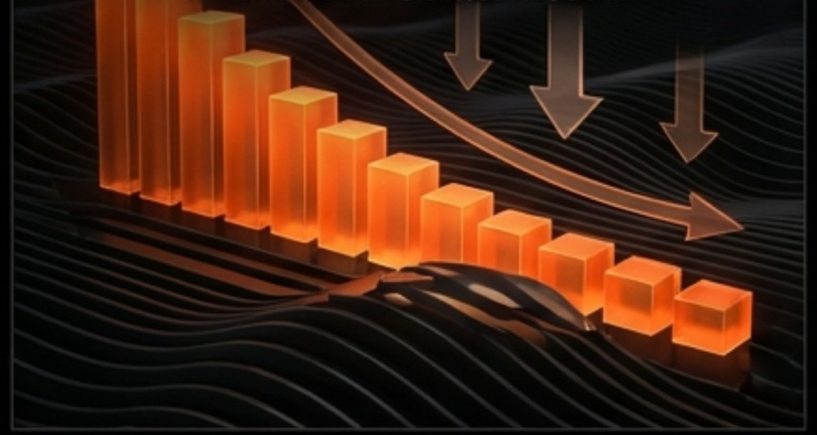
INFERENCE TOKEN GENERATION



10×/年

AI令牌成本下降压力

AI TOKEN COST-DOWN PRESSURE



“强化学习、测试时间扩展和激烈的行业竞争，共同驱动了这场计算风暴。”

传统之路已尽 (THE OLD WAY IS BROKEN)

10×

“如果行业要继续进步，
除非部署激进的极端协同设计。”

1.6×

晶体管年度增长上限
(MAX ANNUAL TRANSISTOR GAIN)

AI性能年度需求
(REQUIRED AI PERFORMANCE GAIN)

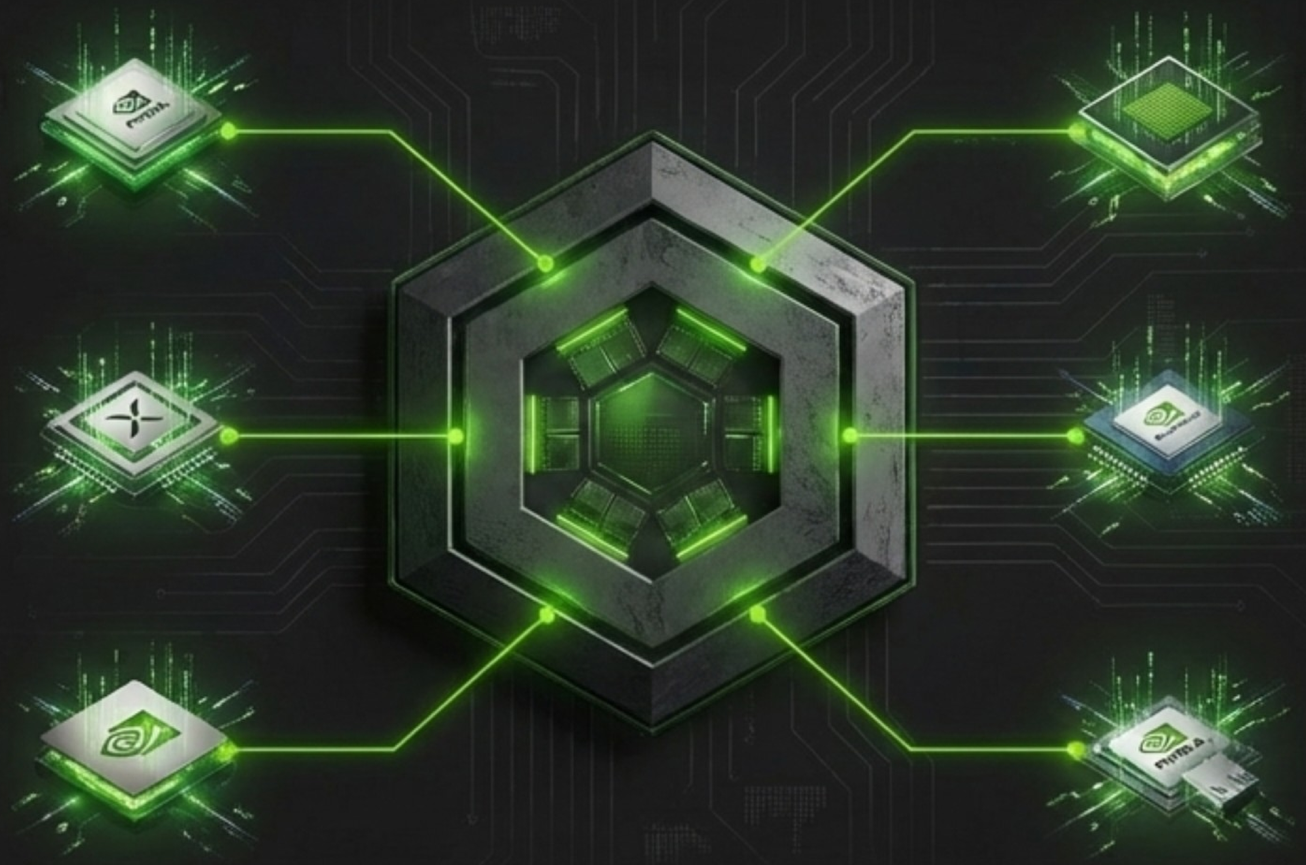
VERA RUBIN



“灵感源自天文学家薇拉·鲁宾，她发现了宇宙中不可见的‘暗物质’。
Vera Rubin平台为解决计算世界中最艰巨的、看不见的挑战而生。”

已全面投产 (NOW IN FULL PRODUCTION)

理念：极致协同设计 (PHILOSOPHY: EXTREME CO-DESIGN)



VERA CPU
（智慧大脑）

RUBIN GPU
（算力引擎）

SPECTRUM-X SWITCH
（外部互联）

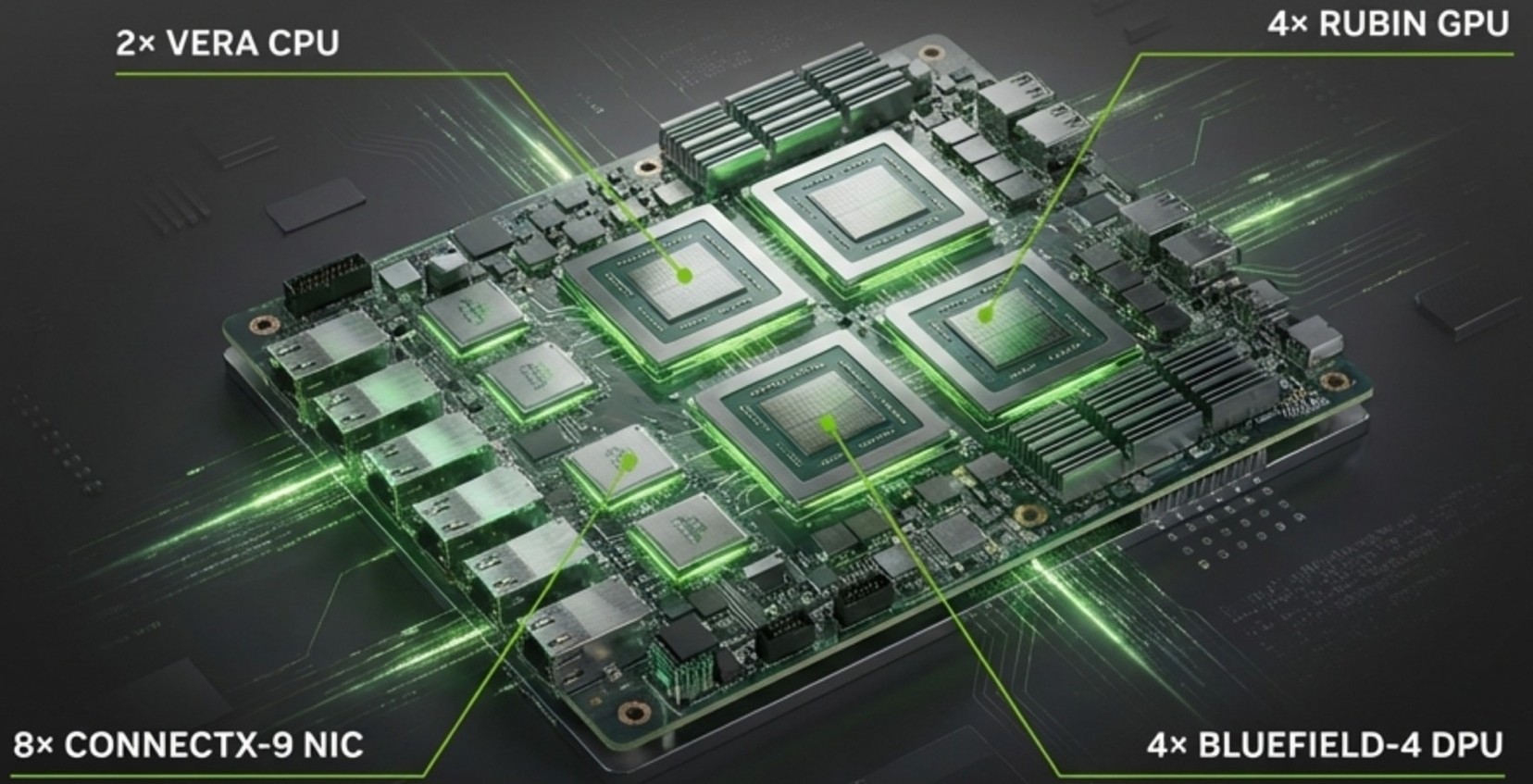
BLUEFIELD-4 DPU
（智能管家）

NVLINK 6 SWITCH
（内部神经）

CONNECTX-9 NIC
（数据接口）

15,000 工程师年投入
(15,000 ENGINEERING-YEARS)

VERA RUBIN 计算模块剖析



100 PFLOPS

AI算力 / 单模块
(AI COMPUTE / PER MODULE)

零线缆设计 (ZERO CABLES. ZERO HOSES. ZERO FANS.)

RUBIN GPU: 超越物理定律的性能飞跃

5×

1.6×



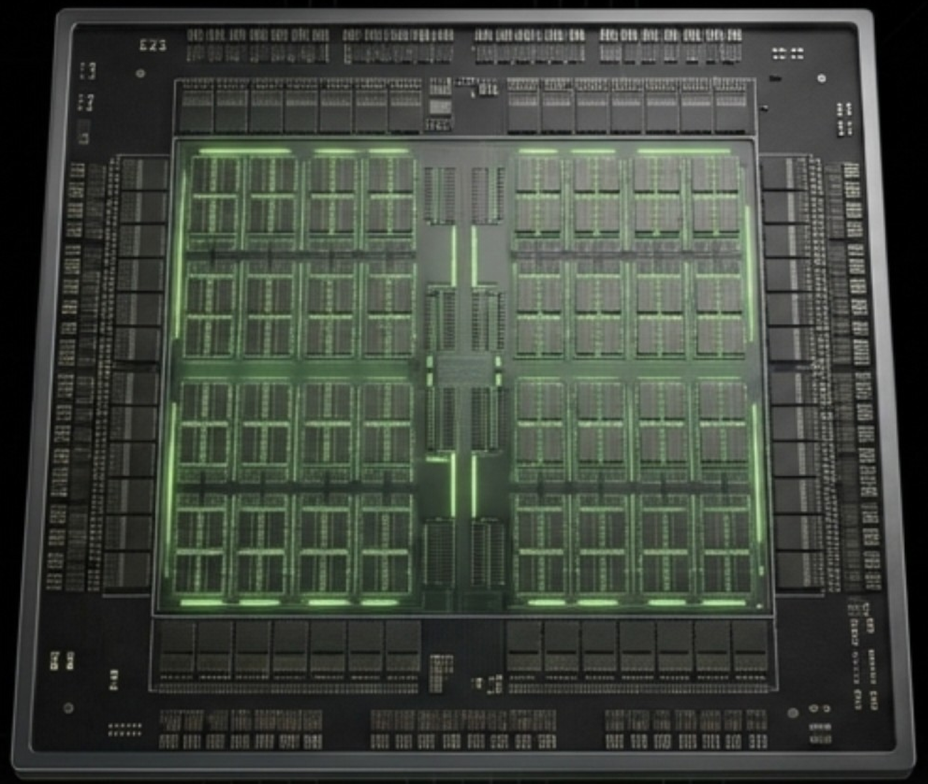
MVF P4 张量引擎
(MVF P4 TENSOR ENGINE)

说明：硬件级自适应精度，无需软件干预，
在处理器内部动态调整精度与结构，
实现吞吐量与精度的最佳平衡。

晶体管数量增长
(VS. BLACKWELL)

浮点性能提升
(VS. BLACKWELL)

VERA CPU：为AI超级计算机定制的大脑



2×

每瓦性能

(PERF/WATT VS. GRACE)

88 核 / 176 线程

空间多线程技术

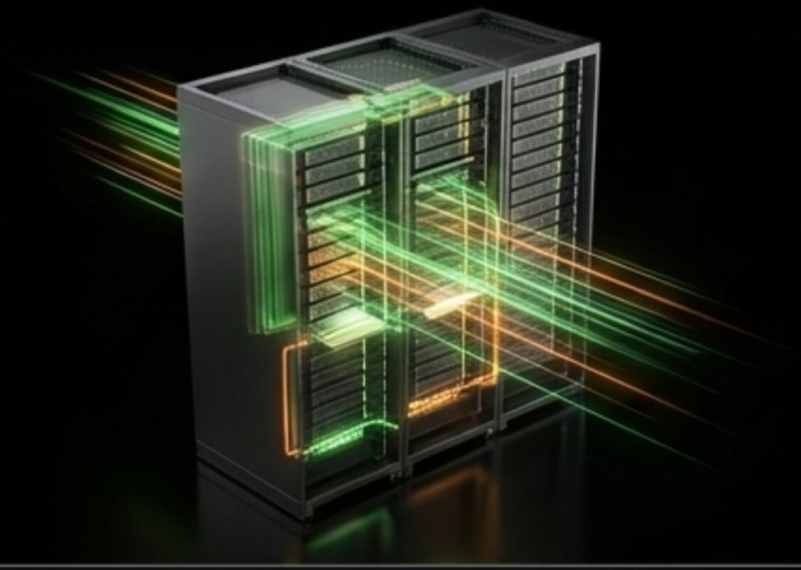
(SPATIAL MULTITHREADING)，每个线程均可获得完整性能

惊人的I/O性能

MONSTROUS I/O

为GPU间的高速数据共享而生

重新定义数据流（REDEFINING DATAFLOW）

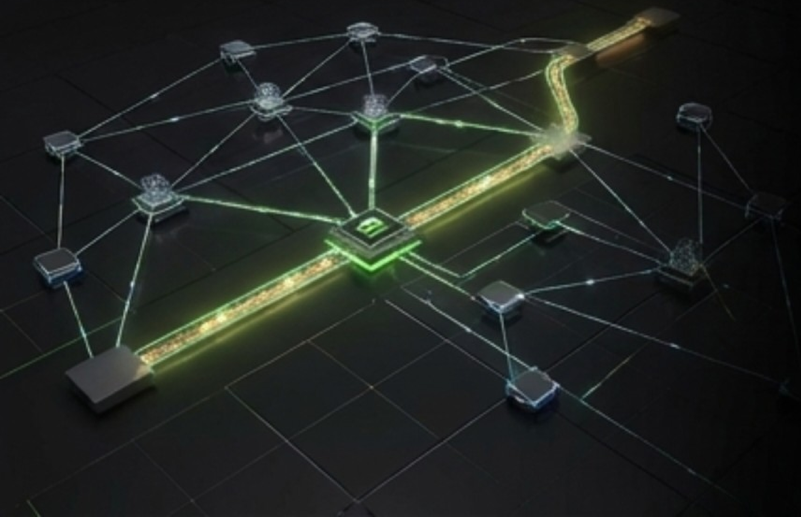


上半部分：机架内神经（INTRA-RACK NERVOUS SYSTEM）

技术：NVLINK 6 SWITCH

240 TB/s

> 2× 全球互联网总带宽



下半部分：AI工厂互联

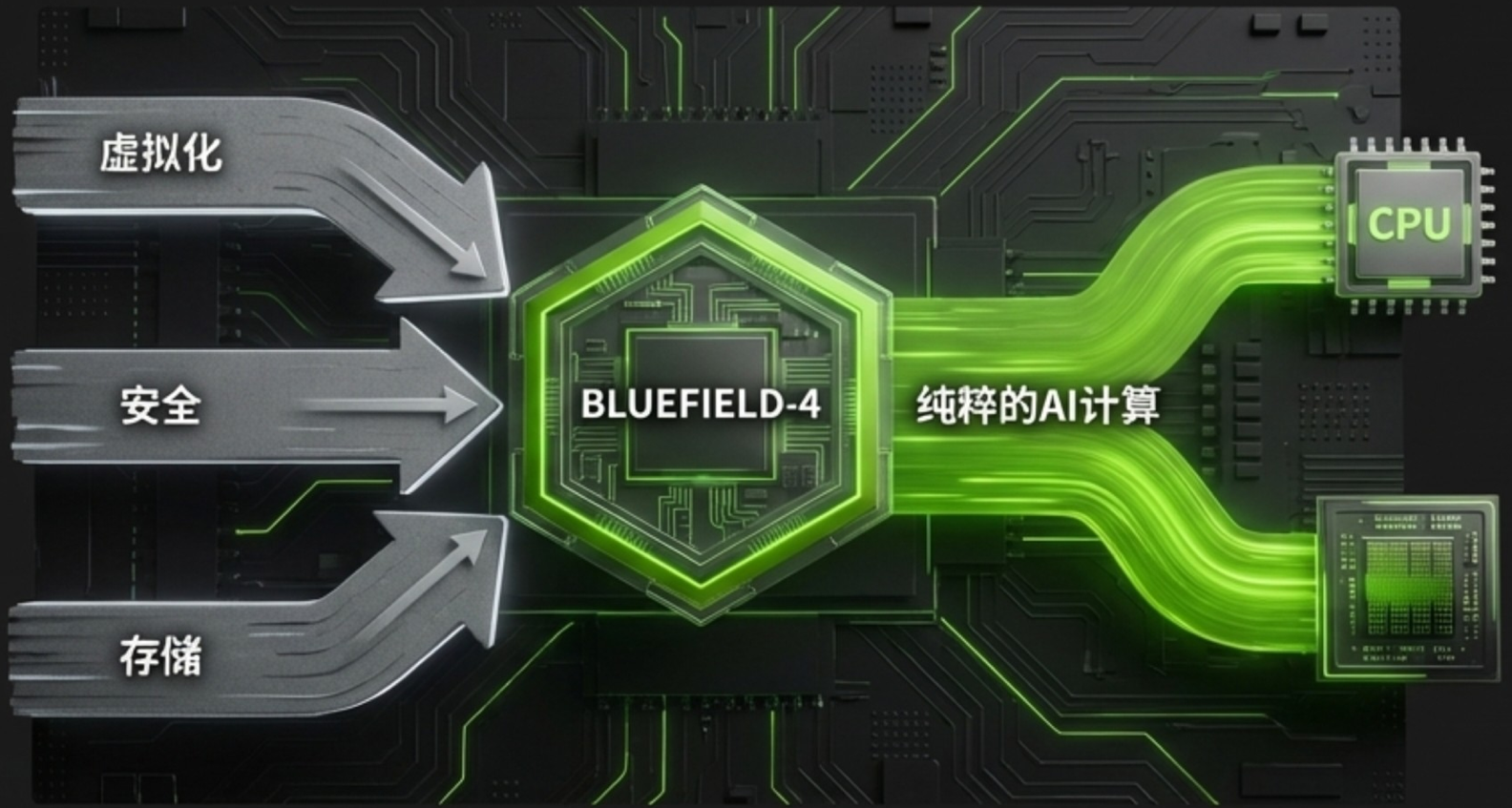
技术：SPECTRUM-X AI ETHERNET

+25% 网络吞吐性能

网络成本趋近于$0 (“THE NETWORK BECOMES FREE”)

- 50亿美元的数据中心，性能提升10%即价值5亿美元。

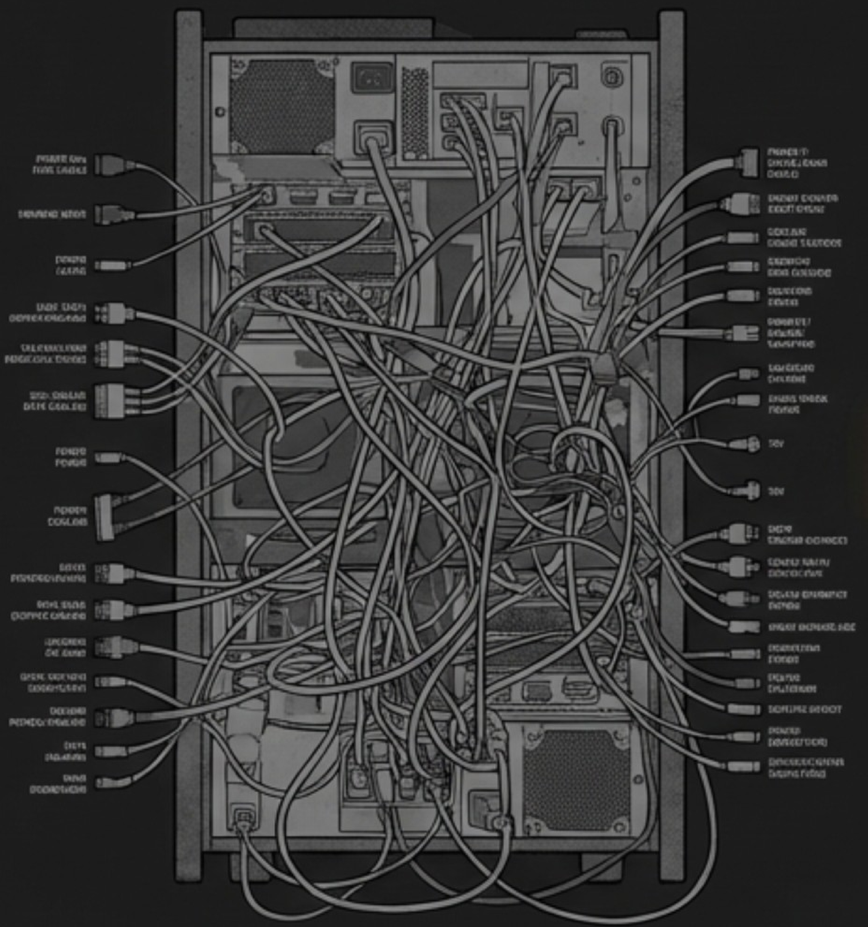
BLUEFIELD-4 DPU：释放算力，专注AI



- 卸载：虚拟化、安全、南北向网络软件
- 优化：东西向流量
- 隔离：支持多租户安全隔离

每个Vera Rubin计算节点的标准配置
（STANDARD ON EVERY COMPUTE NODE）

物理工程的革命：从2小时到5分钟



PHYSICAL SPECS

- 晶体管: 2.2 京
  (2.2 QUADRILLION)
- 重量: 2 吨
  (2 TONS)

- 重量: 2 吨
  (2 TONS)

NVIDIA Vera Rubin
"ZERO CABLE COMPUTE MODULE

PHYSICAL SPECS

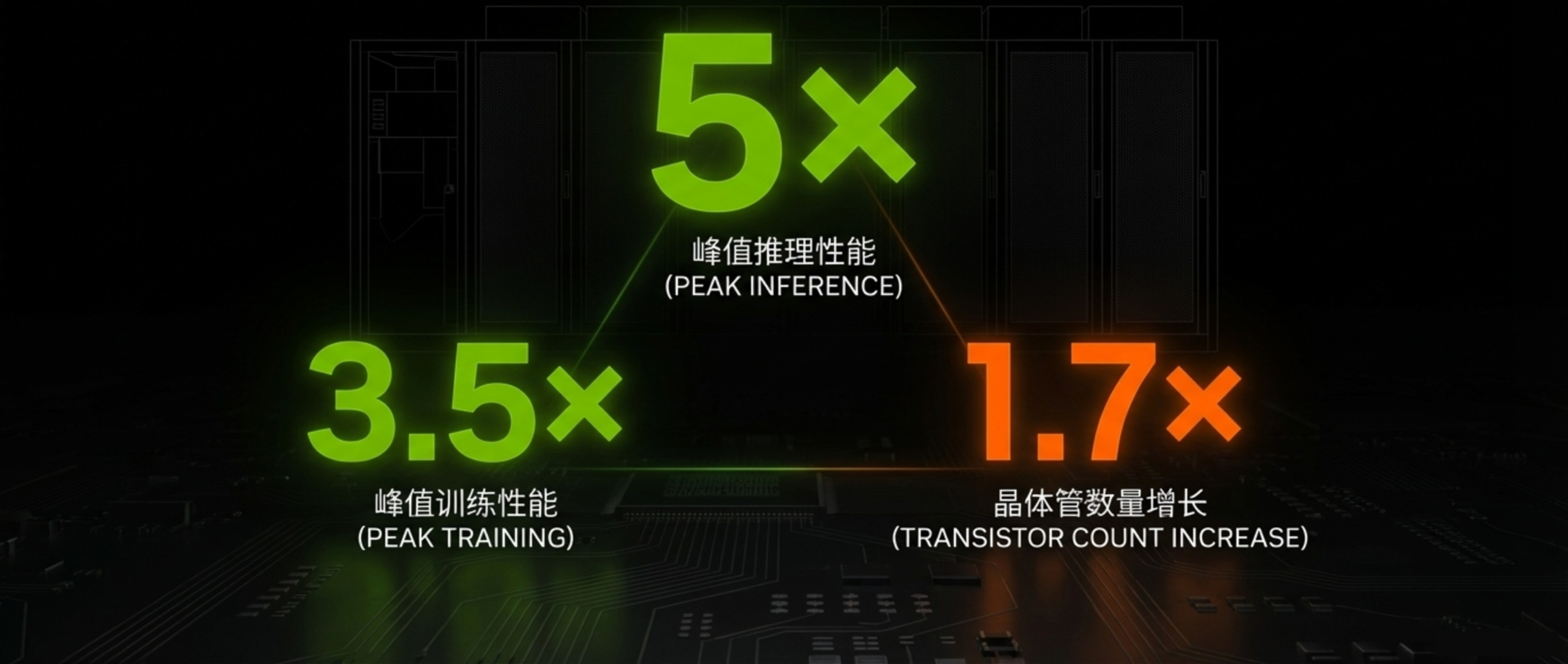
- 重量: 2 吨
  (2 TONS)
- 铜缆: 5000 根 / 2 英里
  (5000 COPPER CABLES / 2 MILES)

ENERGY SPECS

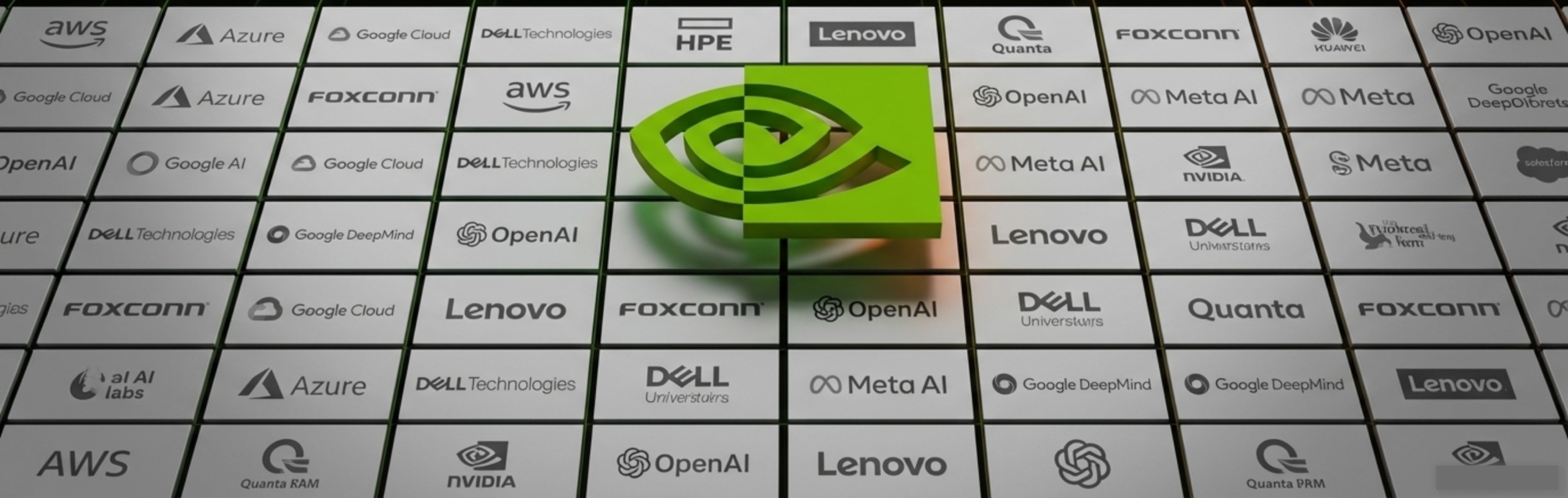
- 冷却: 100% 直接液冷
  (DIRECT-TO-CHIP LIQUID COOLING)
- 水温: 45°C 温水进水
  (WARM WATER INTAKE)

数据中心无需冷水机组
(NO DATACENTER CHILLERS REQUIRED)

VERA RUBIN 平台性能：碾压式飞跃



全球AI工厂的唯一选择
(POWERING THE WORLD'S AI FACTORIES)





迈向AI新边疆 (THE NEXT FRONTIER OF AI)

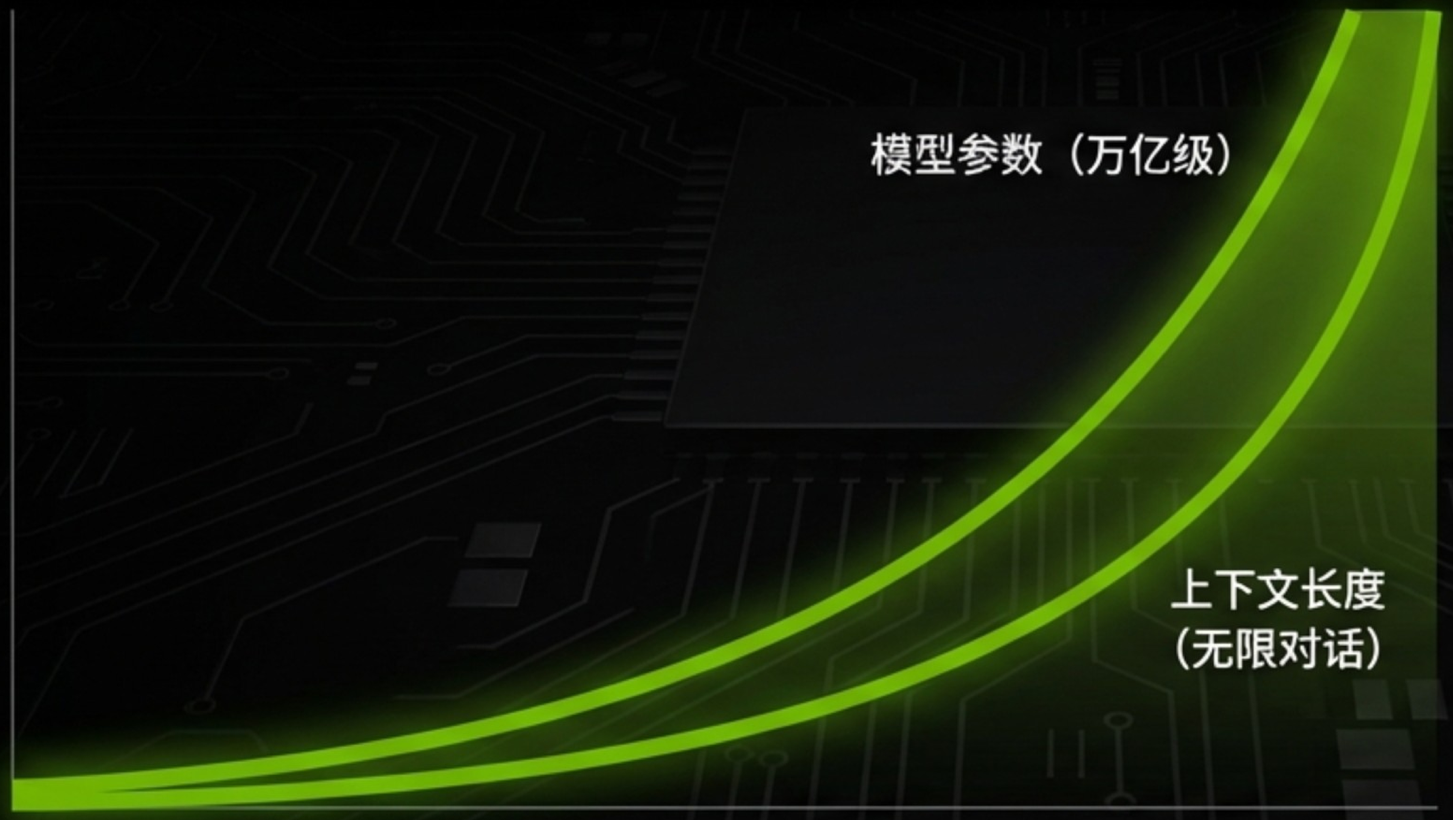
Q&A



VERA RUBIN

计算的炼金术

增长的极限

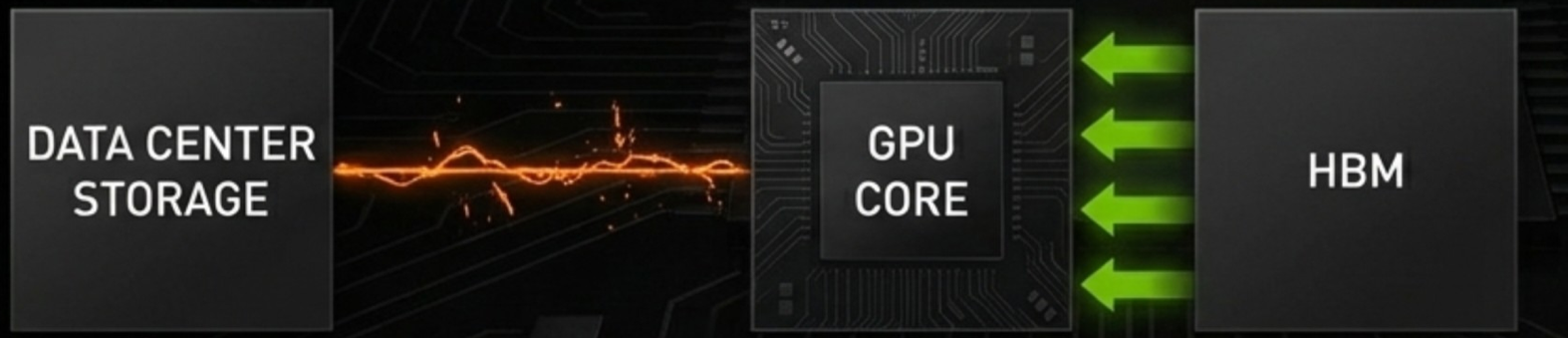


HBM → 容量告急

GRACE高速内存 → 依然告急

现有的一切，即将被需求的巨浪吞噬。

万恶之源：KV 缓存



本质：AI的工作记忆

机制：每生成1个TOKEN = 读取全部工作记忆

后果：网络流量爆炸，南北向网络不堪重负

突破 I - 光之炼金：硅光子连接



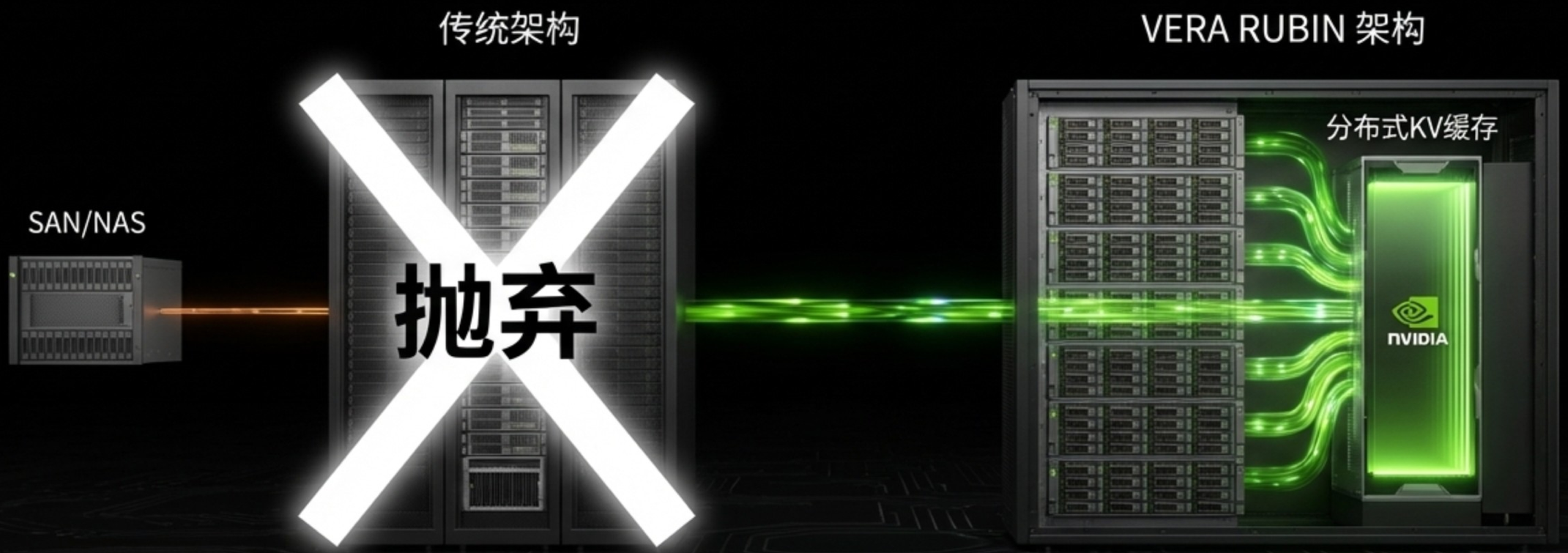
512 x 200 Gbps

工艺**：台积电 (TSMC) Co-op 硅光子集成工艺

集成**：光模块直连芯片

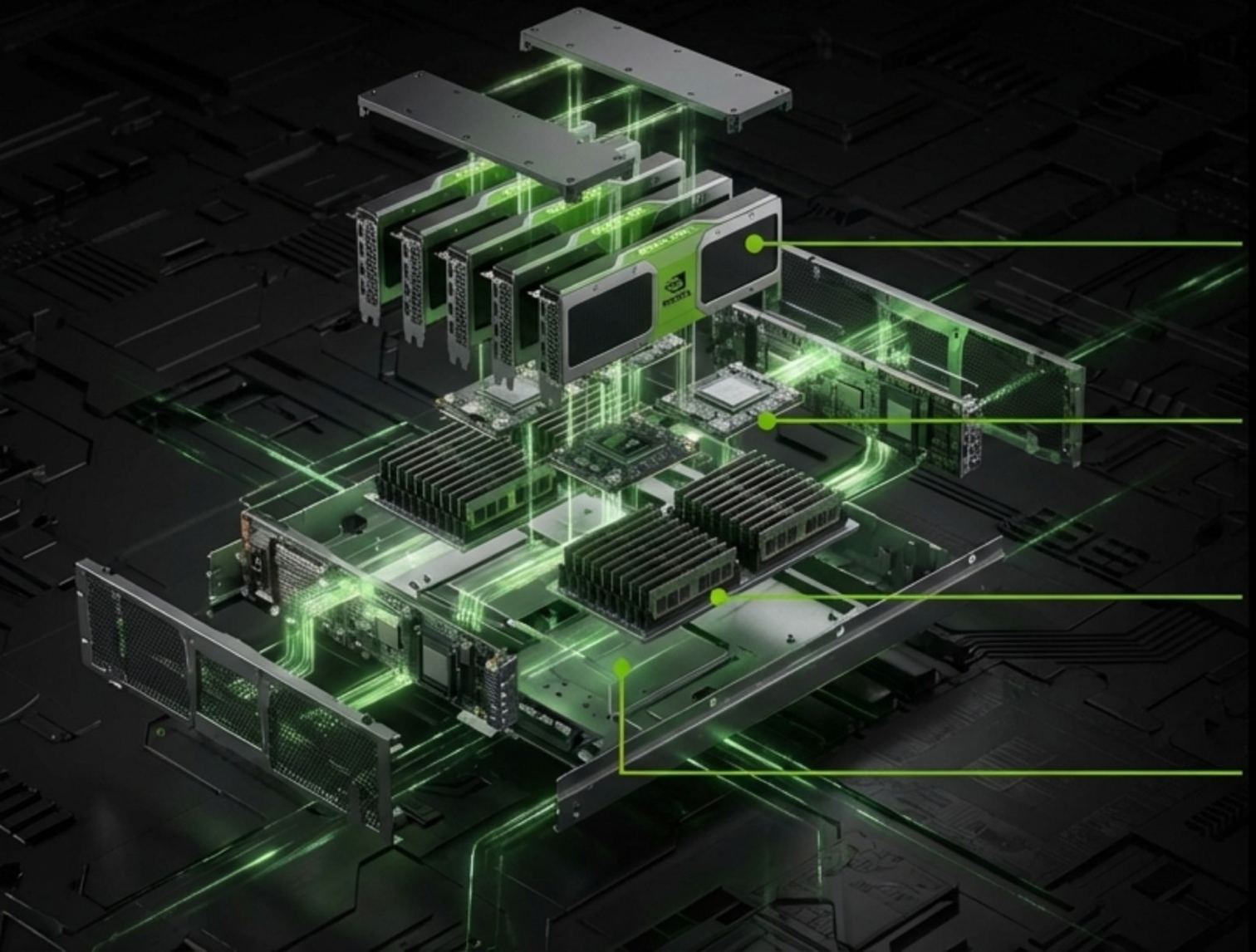
产品**：SPECTRUM-X 新一代AI以太网交换机

突破 II - 内存炼金：上下文内存革命



DIN 2014
- 抛弃传统南北向存储
- 在机架内构建高速东西向流量
- 由BLUEFIELD DPU驱动

新架构解剖：VERA RUBIN 节点



+16 TB / GPU

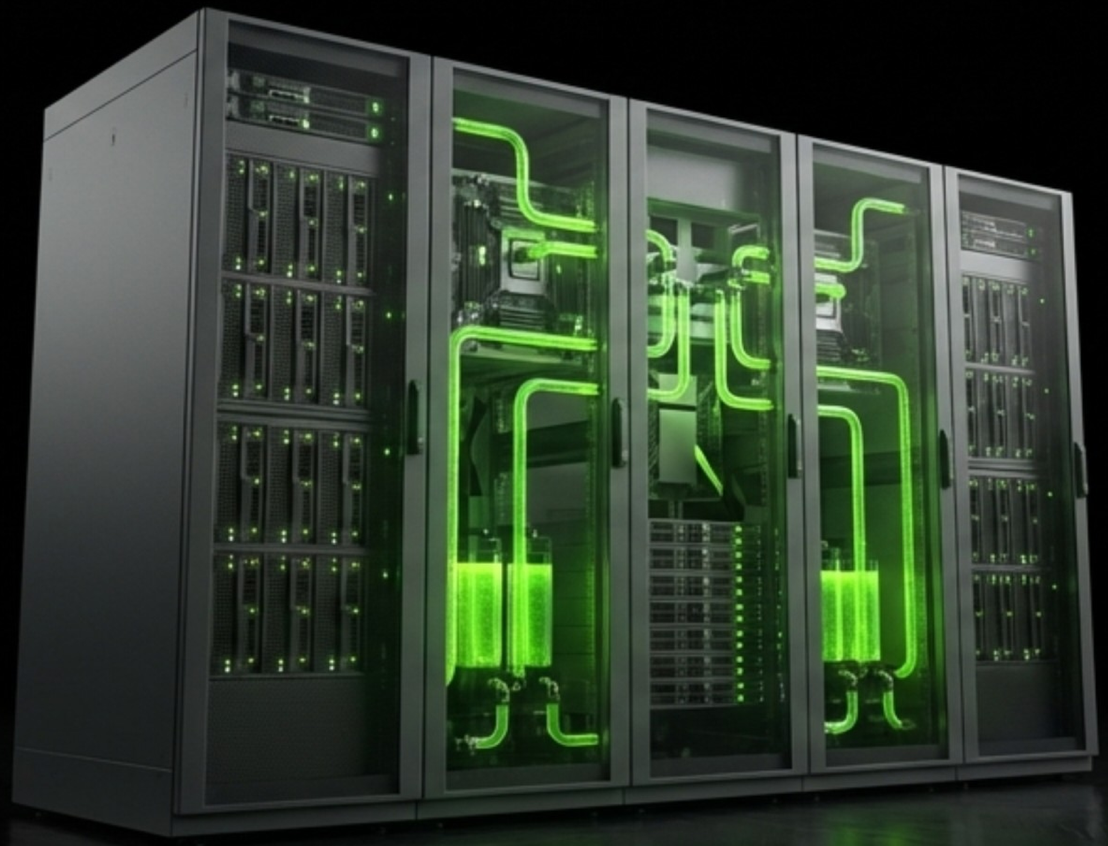
GPU： 144x RUBIN GPUs (NVLINK 72)

DPU： 4x BLUEFIELD DPUs (每内存组)

内存： 150 TB 上下文内存 (每组)

带宽： 200 Gb/s 节点内东西向带宽

VERA RUBIN 系统



全栈重构，为AI而生
不止是芯片，更是数据中心本身

能效 X 安全

能效与散热



能效翻倍，节省约6%数据中心电力



液冷恒温 45°C

安全



全系统可信计算安全

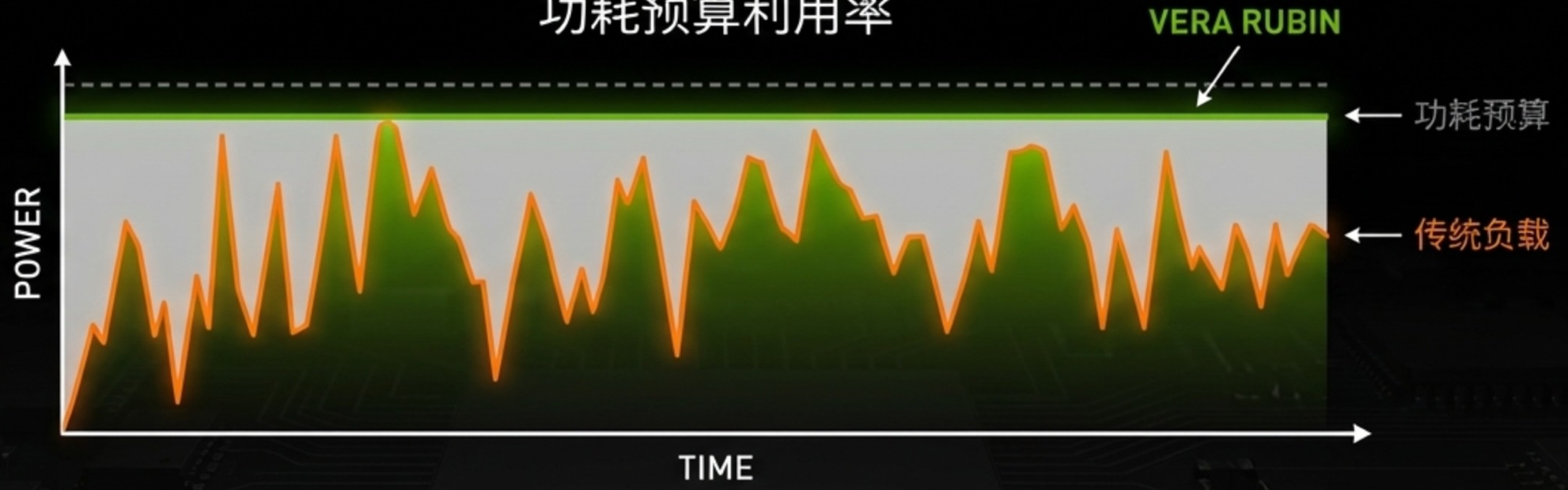
总线 / 连接 / 内存 → 全程加密

(PCIe, NVLink, CPU-GPU, GPU-GPU)

功率平滑：压榨每一瓦

100%

功耗预算利用率



消除 25% 峰值冗余

拒绝浪费

训练速度：抵达未来



时间，是唯一的壁垒。

工厂吞吐量：营收密码



每瓦性能 = 数据中心营收。

TOKEN 成本：重塑经济学

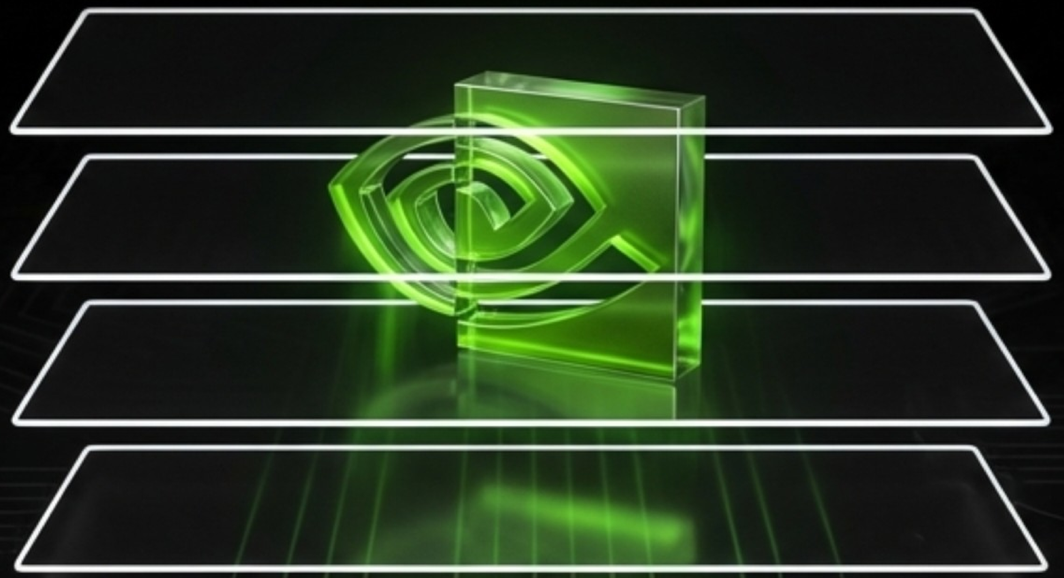


1/10

推理成本 VS. 上一代

“THE MORE YOU BUY, THE MORE YOU SAVE”

全栈AI技术革命



应用（Applications）

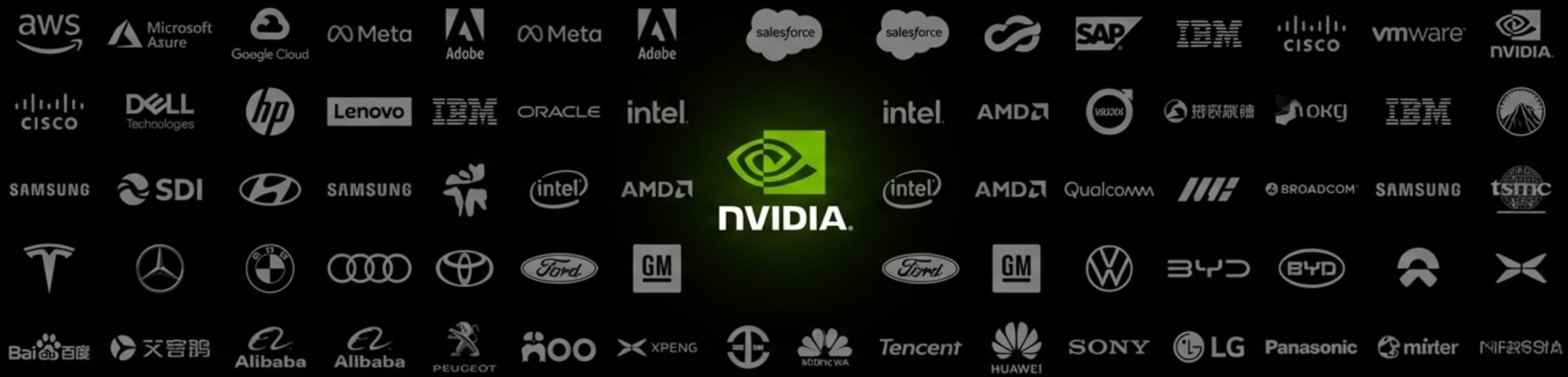
模型（Models）

基础设施（Infrastructure）

芯片（Chips）

我们制造芯片
我们构建系统
我们赋能世界

生态即未来


aws
Microsoft Azure
Google Cloud
Meta
Adobe
Meta
Adobe
salesforce
salesforce
SAP
IBM
cisco
vmware
NVIDIA
cisco
DELL Technologies
hp
Lenovo
IBM
ORACLE
intel
intel
AMD
OKG
IBM
SAMSUNG
SDI
SAMSUNG
intel
AMD
NVIDIA
intel
AMD
Qualcomm
BROADCOM
SAMSUNG
tsmc
Ford
GM
Ford
GM
BYD
BYD
Baidu 百度
Alibaba
Alibaba
PEUGEOT
XPENG
Tencent
HUAWEI
SONY
LG
Panasonic
mirter

